Data Visualization
Discover, Explore and be Skeptical

Di Cook
Statistics, Iowa State University
soon to be Business Analytics, Monash University

Les Diablerets, Feb 1-4, 2015

Seminar 2
Plotting many dimensions

- Tours
- Parallel coordinate plots
- Scatterplot matrices
- Multiple linked plots
- Using these together to explore data
- What we can learn about tennis!

Notation

Data

Crab	species	sex	frontal lobe	rear width	carapace length	carapace width	body depth
1	blue	male	8.1	6.7	16.1	19.0	7.0
2	blue	male	8.8	7.7	18.1	20.8	7.4
3	blue	male	9.2	7.8	19.0	22.4	7.7
4	blue	male	9.6	7.9	20.1	23.1	8.2
51	blue	female	7.2	6.5	14.7	17.1	6.1
52	blue	female	9.0	8.5	19.3	22.7	7.7
53	blue	female	9.1	8.1	18.5	21.6	7.7
101	orange	male	9.1	6.9	16.7	18.6	7.4
102	orange	male	10.2	8.2	20.2	22.2	9.0
151	orange	female	10.7	9.7	21.4	24.0	9.8
152	orange	female	11.4	9.2	21.7	24.1	9.7
153	orange	female	12.5	10.0	24.1	27.0	10.9

$$\mathbf{X} = [\mathbf{X}_1 \ \mathbf{X}_2 \ \dots \ \mathbf{X}_p] = \begin{bmatrix} X_{11} & X_{12} & \dots & X_{1p} \\ X_{21} & X_{22} & \dots & X_{2p} \\ \vdots & \vdots & & \vdots \\ X_{n1} & X_{n2} & \dots & X_{np} \end{bmatrix}_{n \times p}$$

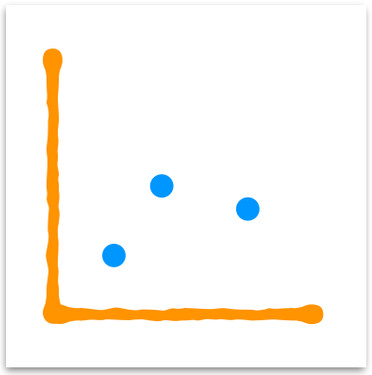
Notation

Data

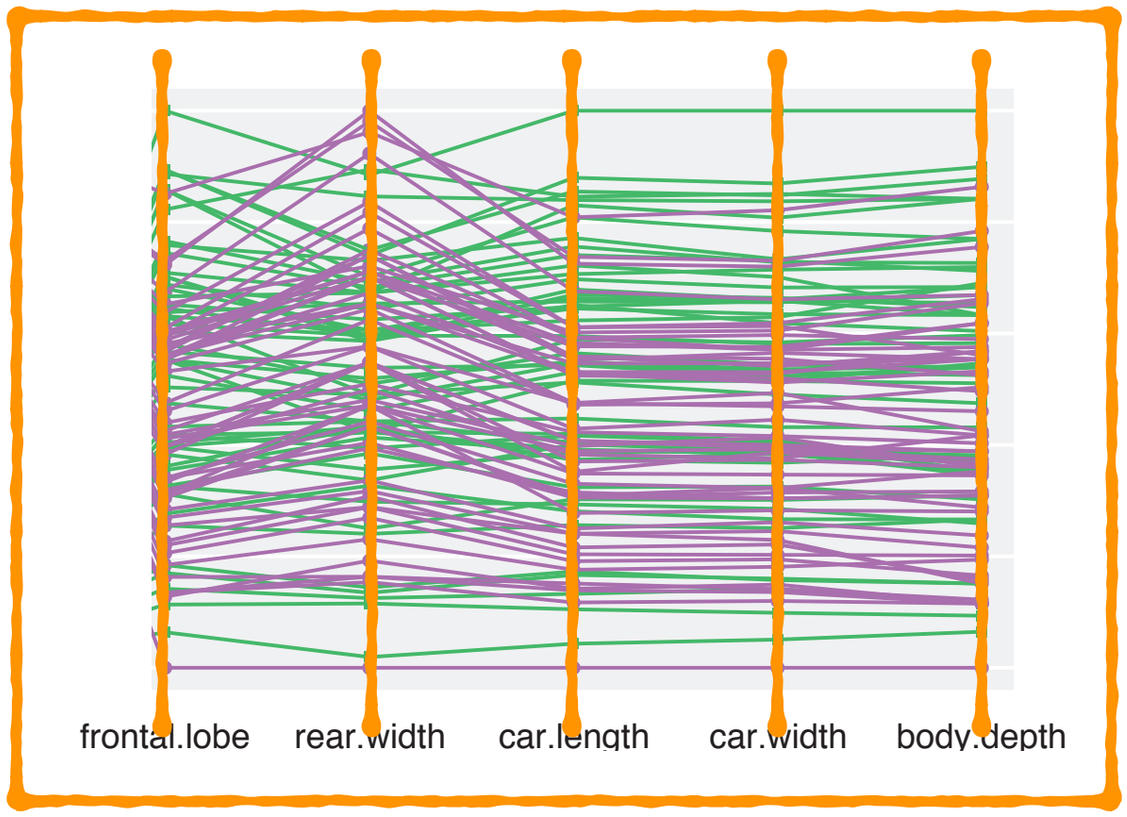
Crab	species	sex	frontal lobe	rear width	carapace length	carapace width	body depth
1	blue	male	8.1	6.7	16.1	19.0	7.0
2	blue	male	8.8	7.7	18.1	20.8	7.4
3	blue	male	9.2	7.8	19.0	22.4	7.7
4	blue	male	9.6	7.9	20.1	23.1	8.2
51	blue	female	7.2	6.5	14.7	17.1	6.1
52	blue	female	9.0	8.5	19.3	22.7	7.7
53	blue	female	9.1	8.1	18.5	21.6	7.7
101	orange	male	9.1	6.9	16.7	18.6	7.4
102	orange	male	10.2	8.2	20.2	22.2	9.0
151	orange	female	10.7	9.7	21.4	24.0	9.8
152	orange	female	11.4	9.2	21.7	24.1	9.7
153	orange	female	12.5	10.0	24.1	27.0	10.9

$$\mathbf{X} = [\mathbf{X}_1\ \mathbf{X}_2\ \ldots\ \mathbf{X}_p] = \begin{bmatrix} X_{11} & X_{12} & \ldots & X_{1p} \\ X_{21} & X_{22} & \ldots & X_{2p} \\ \vdots & \vdots & & \vdots \\ X_{n1} & X_{n2} & \ldots & X_{np} \end{bmatrix}_{n \times p}$$

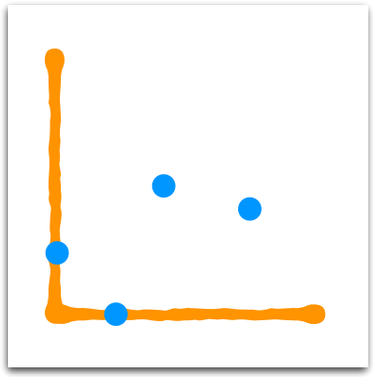
Parallel coordinate plot



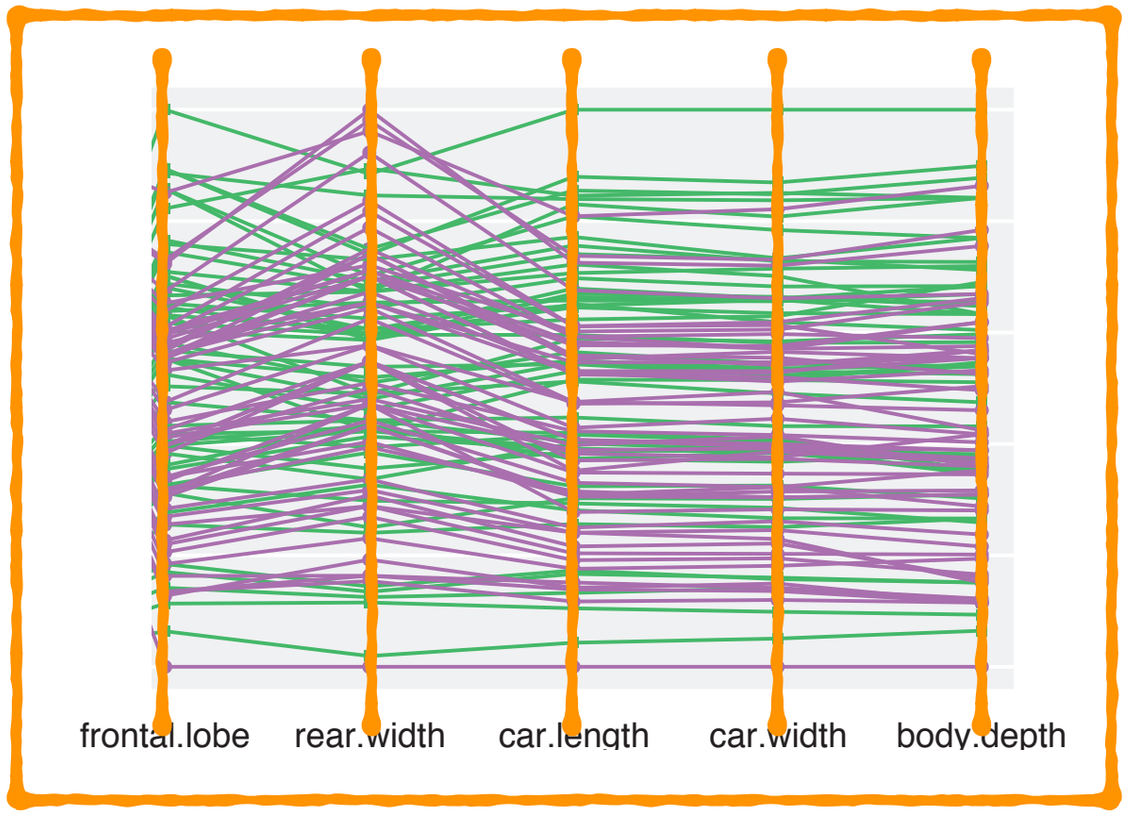
Cartesian to
parallel coords



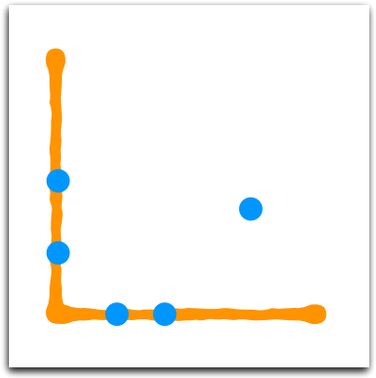
Parallel coordinate plot



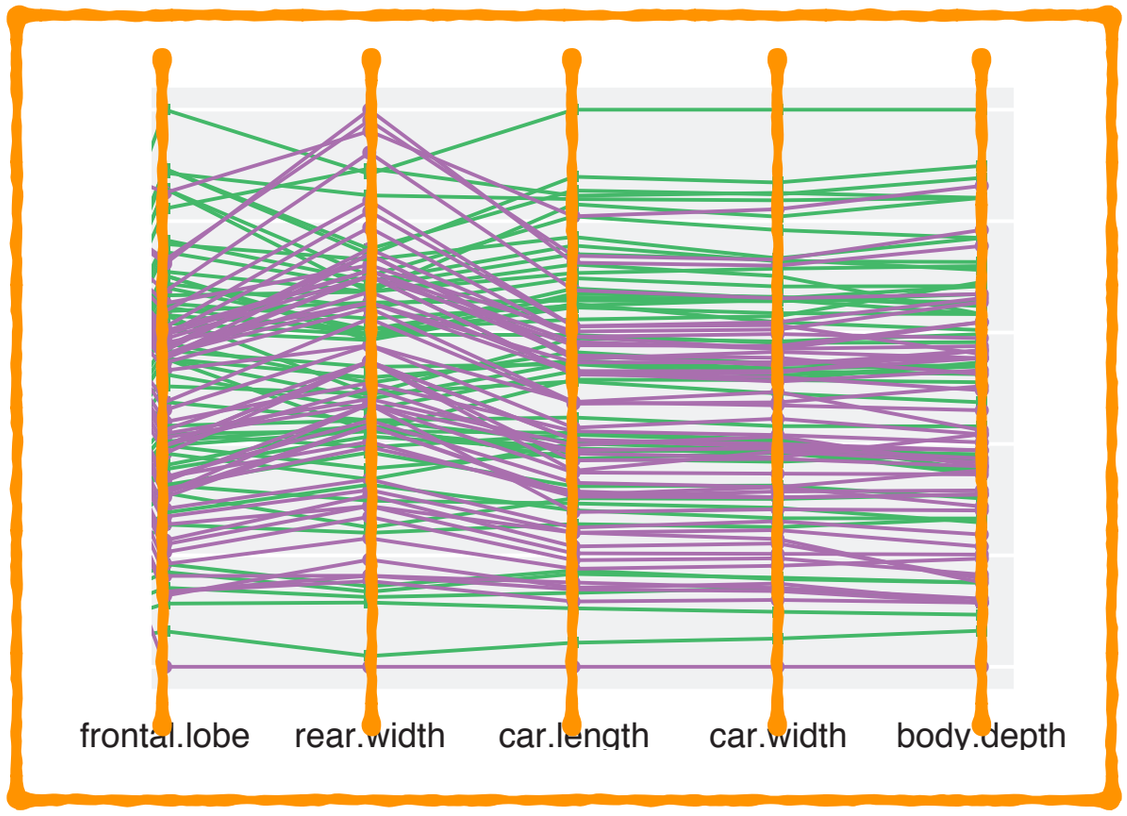
Cartesian to parallel coords



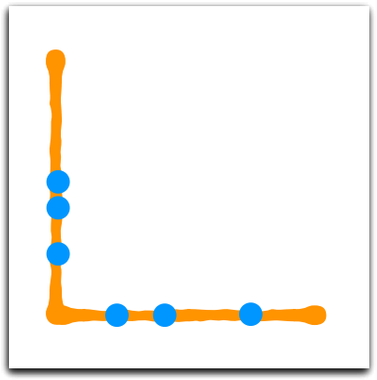
Parallel coordinate plot



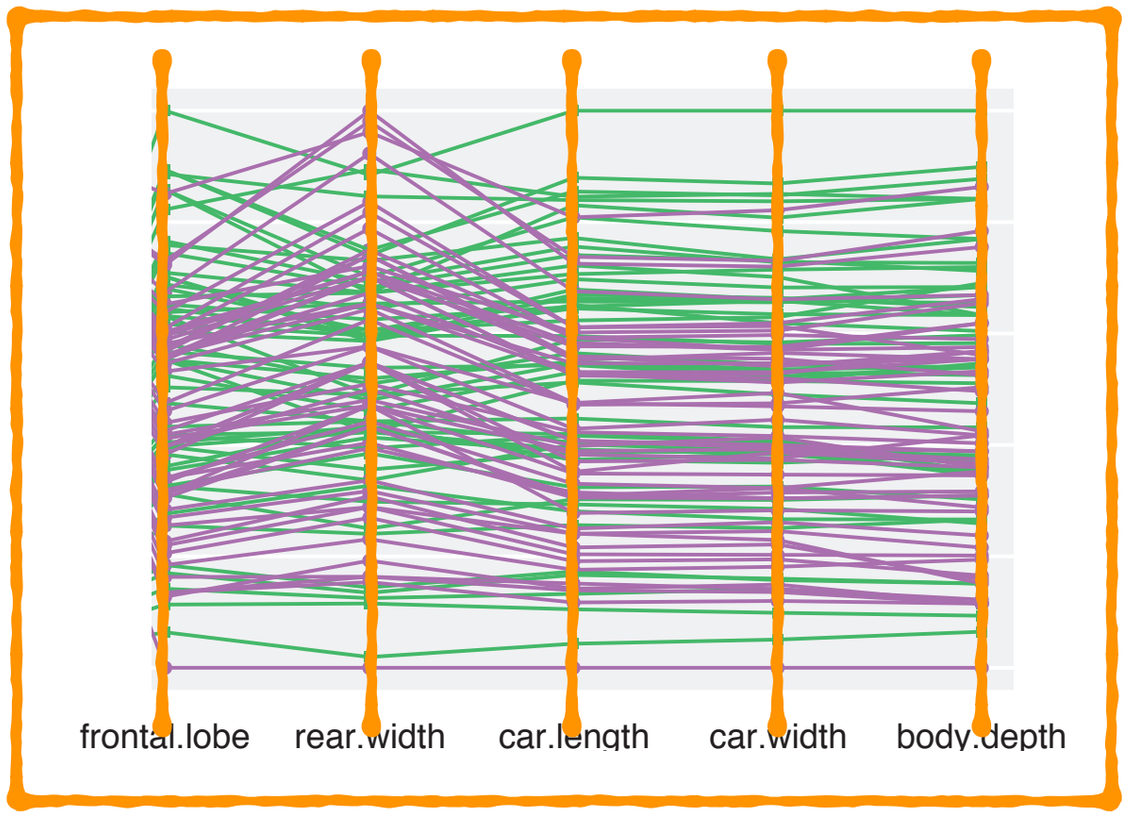
Cartesian to parallel coords



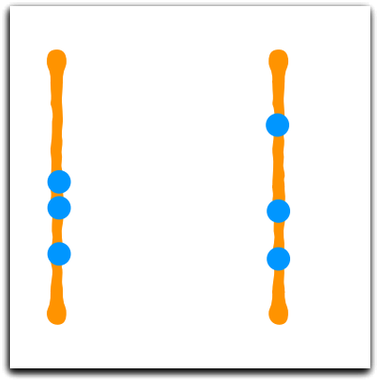
Parallel coordinate plot



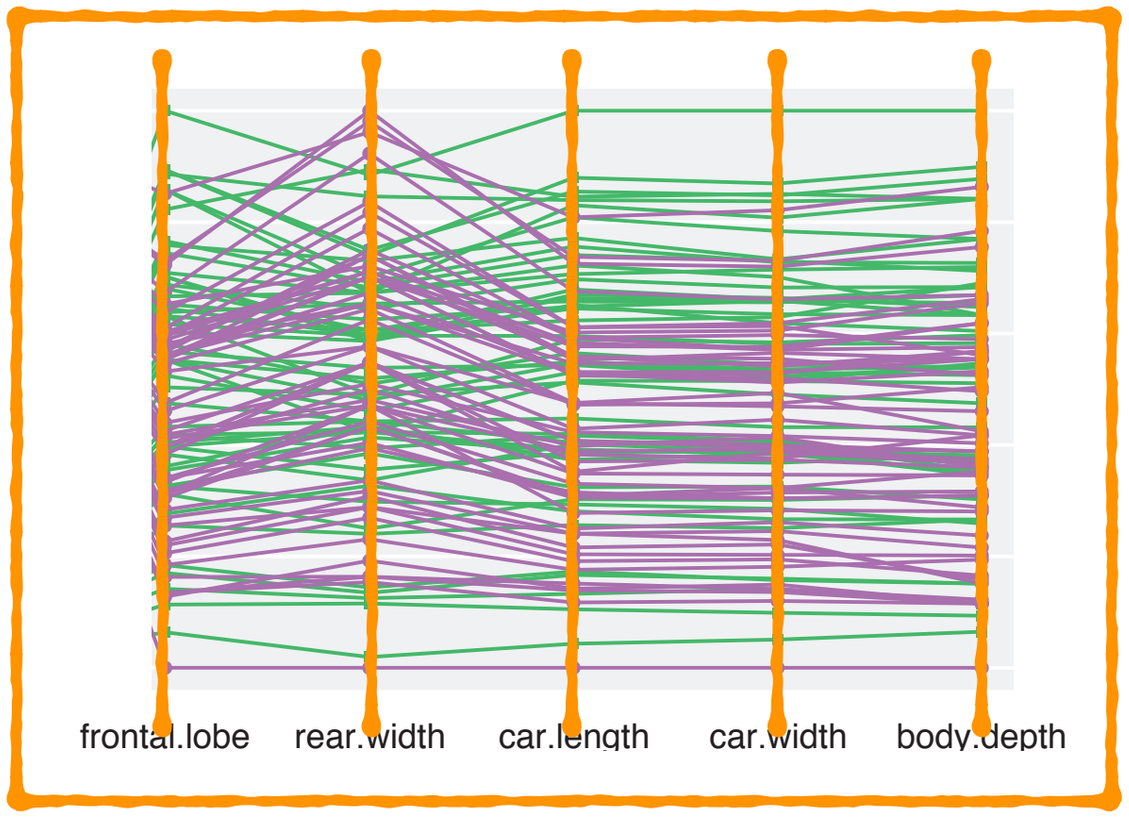
Cartesian to parallel coords



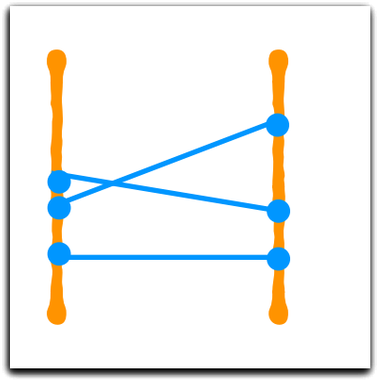
Parallel coordinate plot



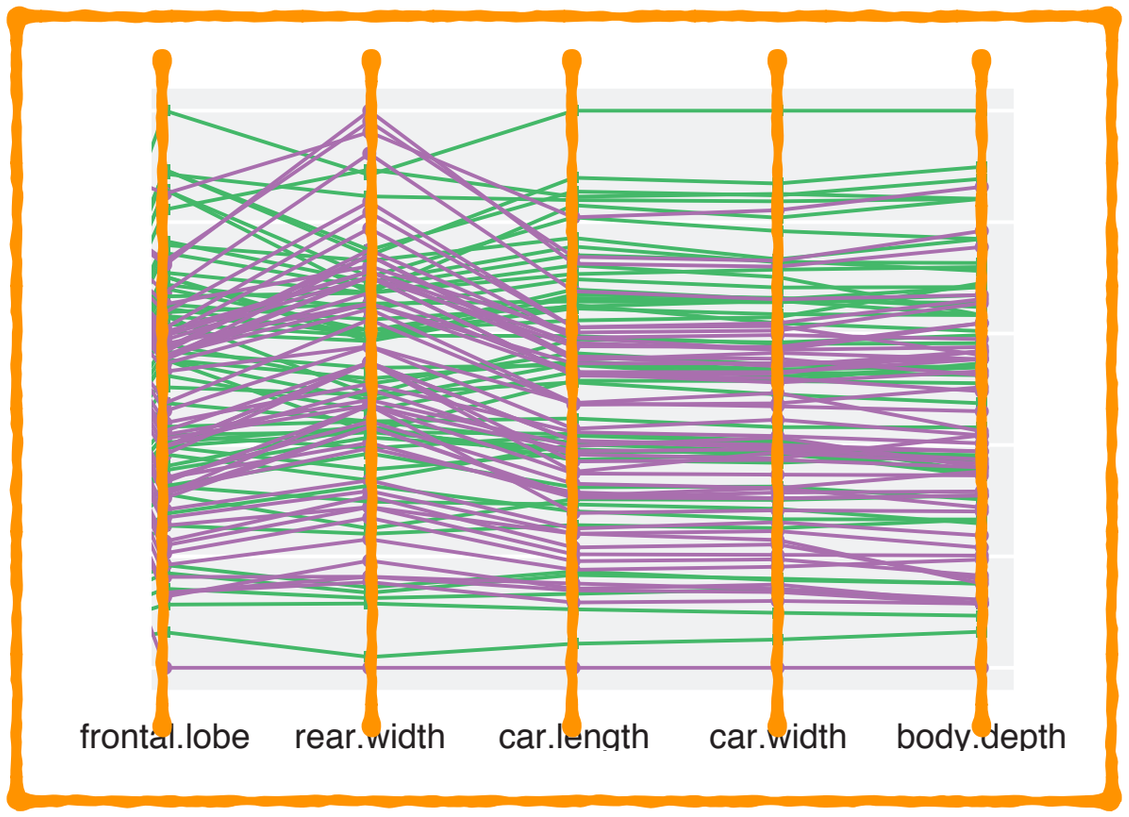
Cartesian to parallel coords



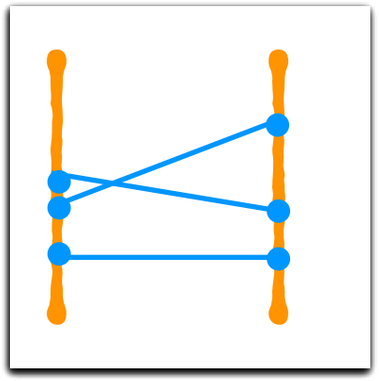
Parallel coordinate plot



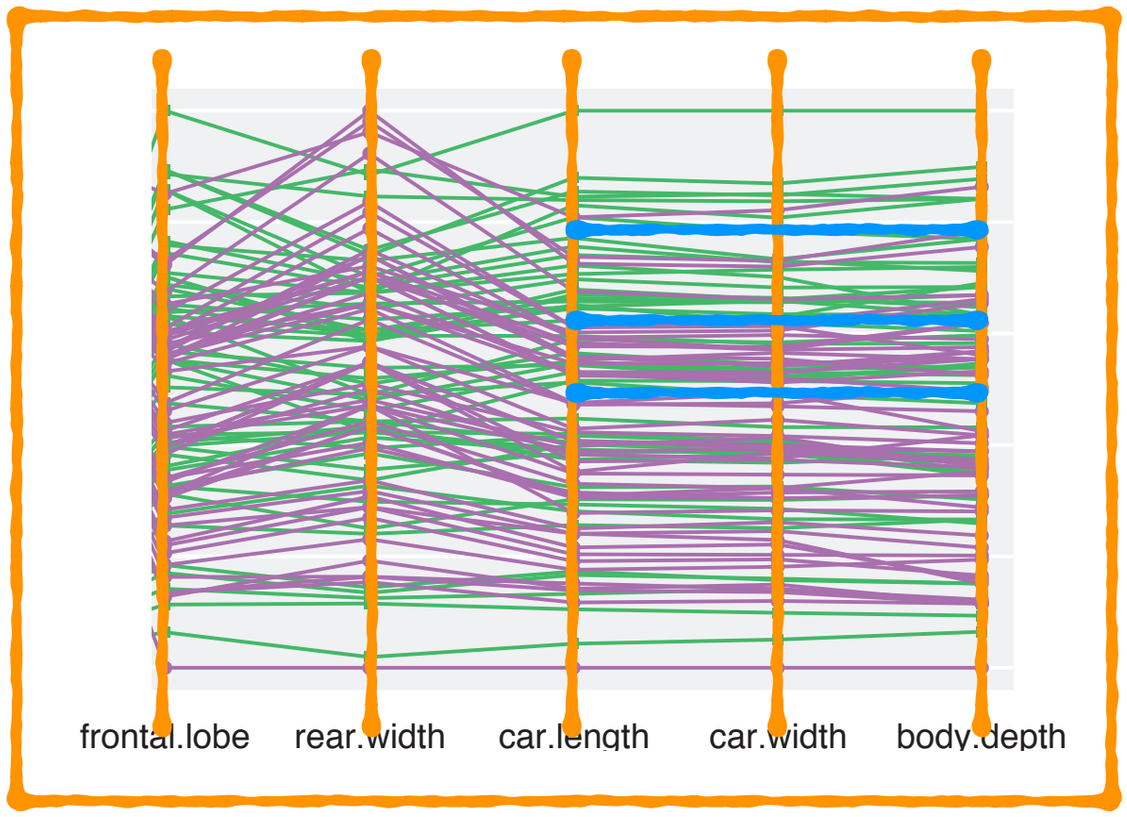
Cartesian to parallel coords



Parallel coordinate plot

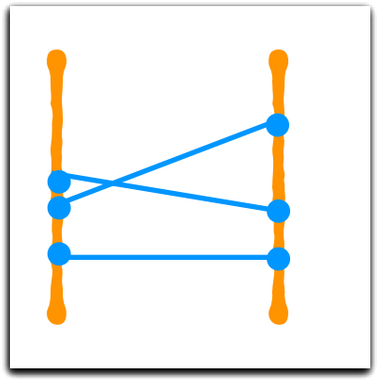


Cartesian to parallel coords

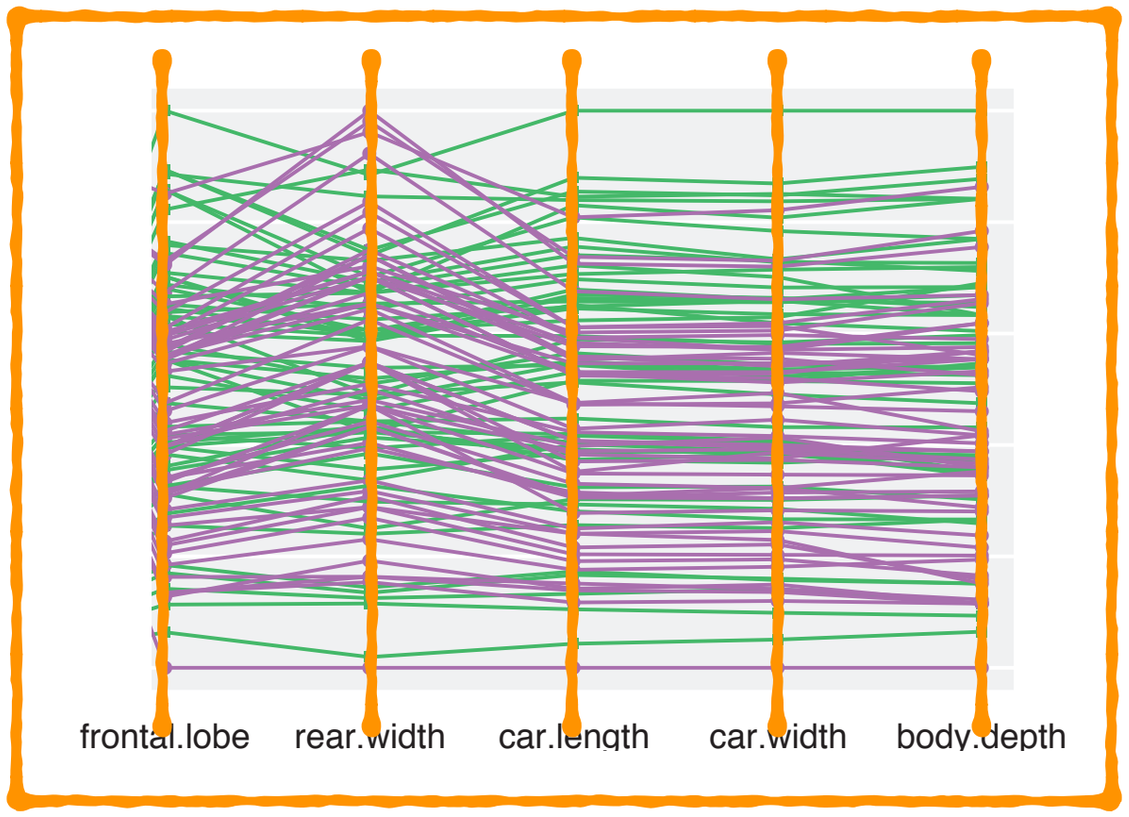


flat lines means positive association

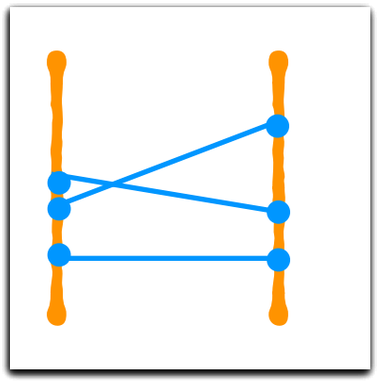
Parallel coordinate plot



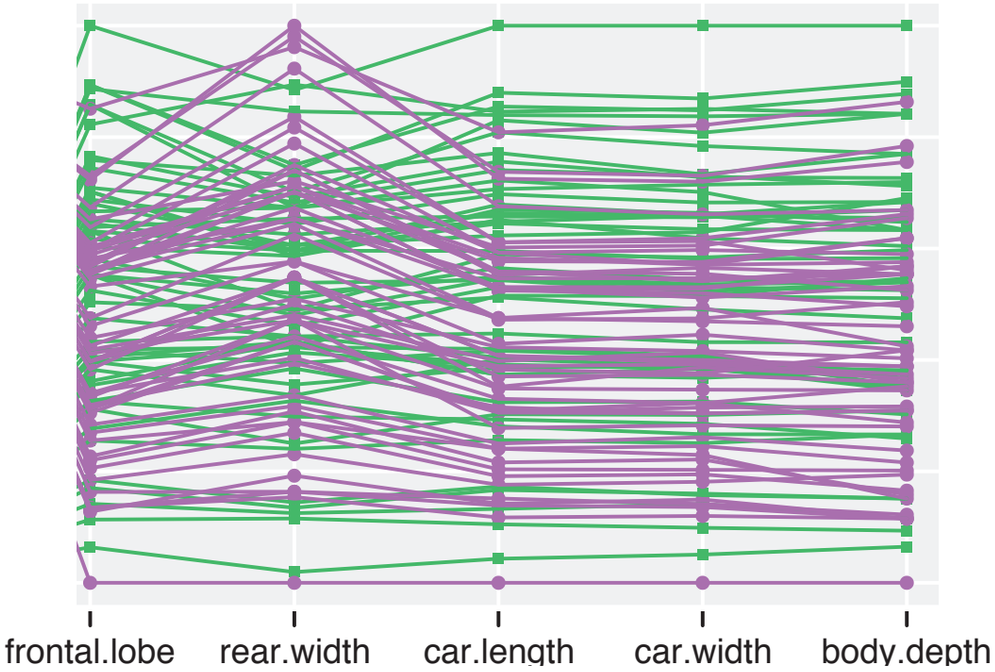
Cartesian to parallel coords



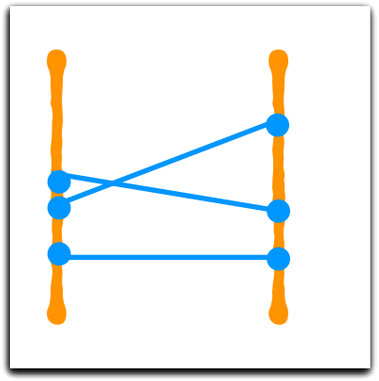
Parallel coordinate plot



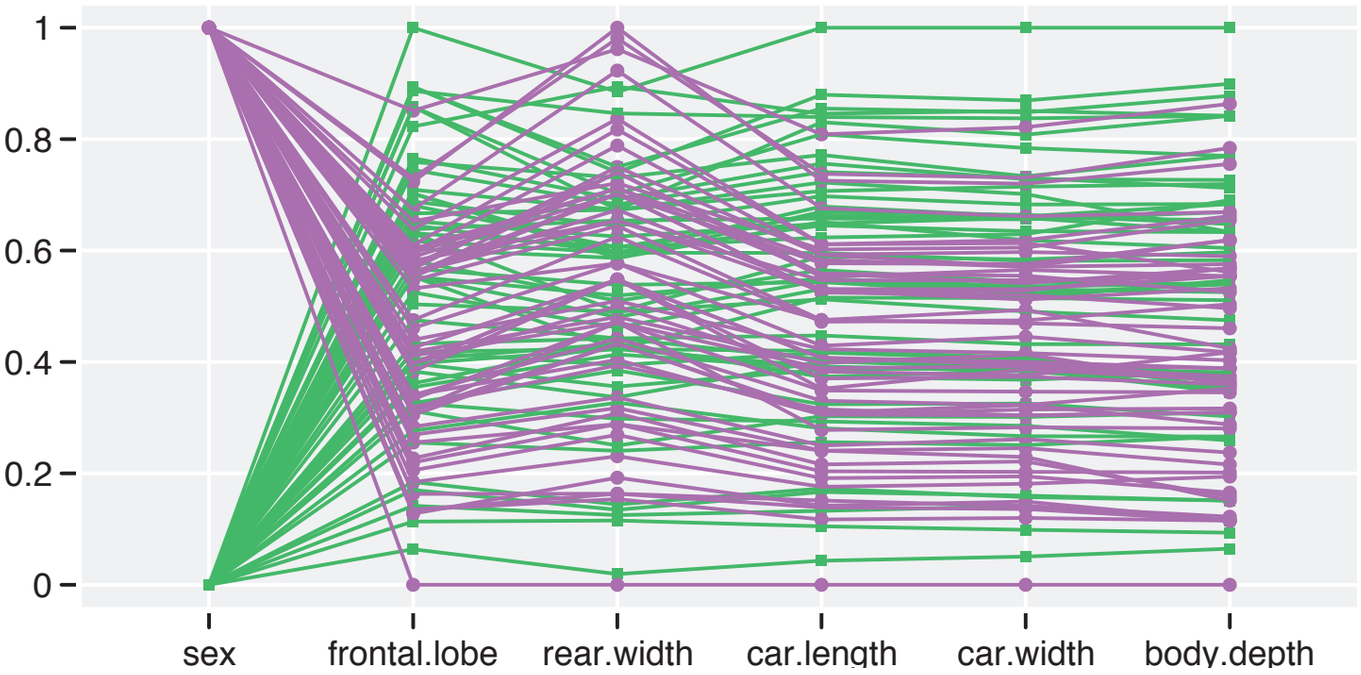
Cartesian to parallel coords



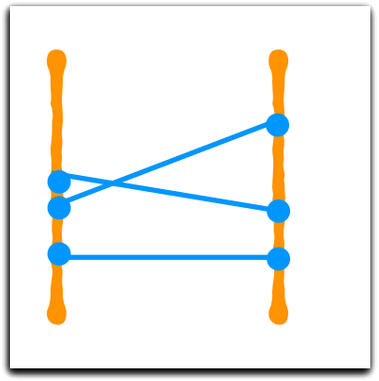
Parallel coordinate plot



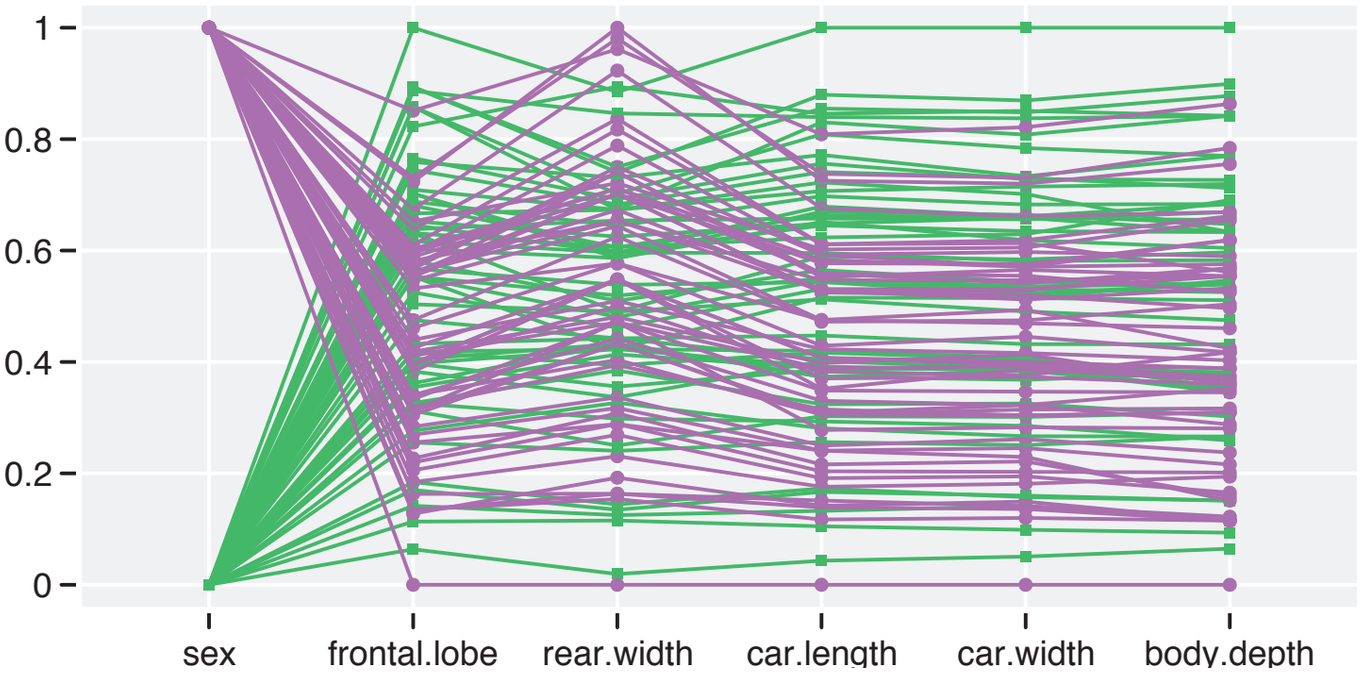
Cartesian to parallel coords



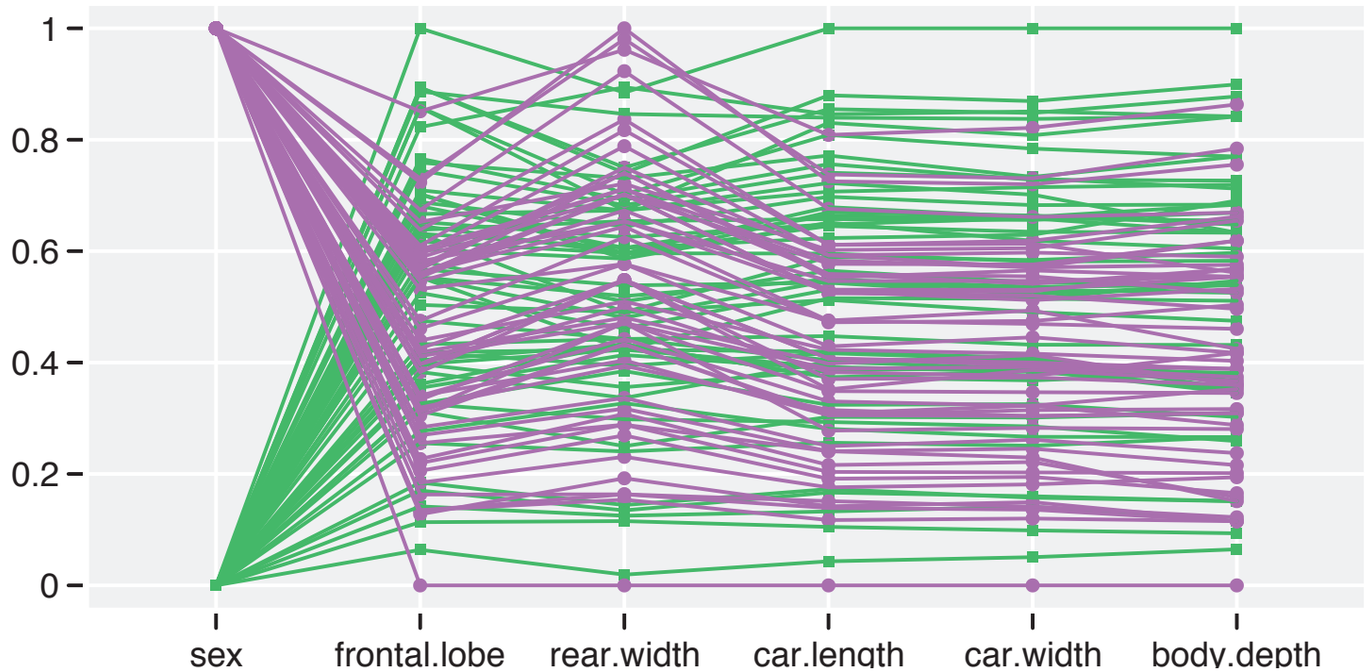
Parallel coordinate plot



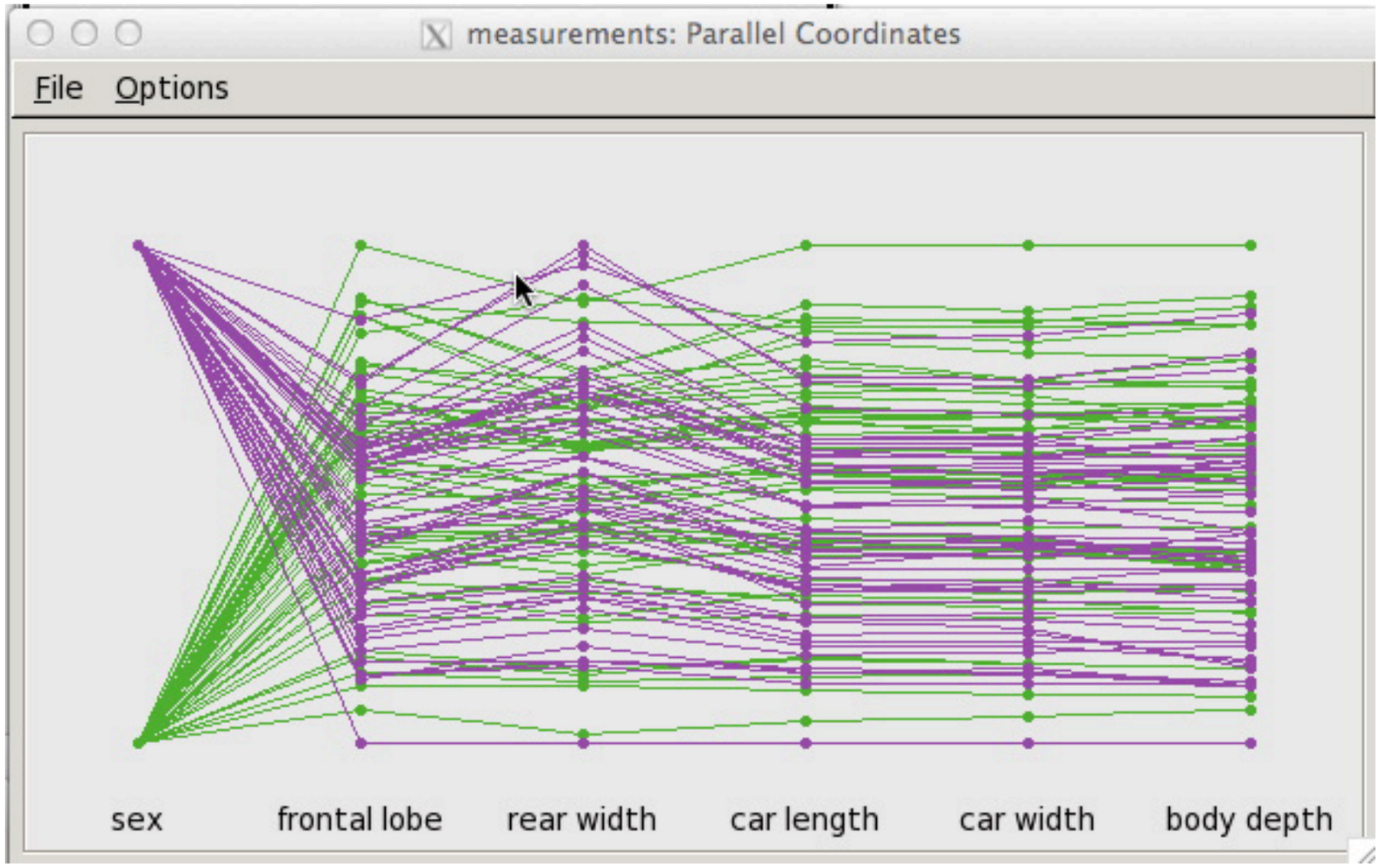
Cartesian to parallel coords

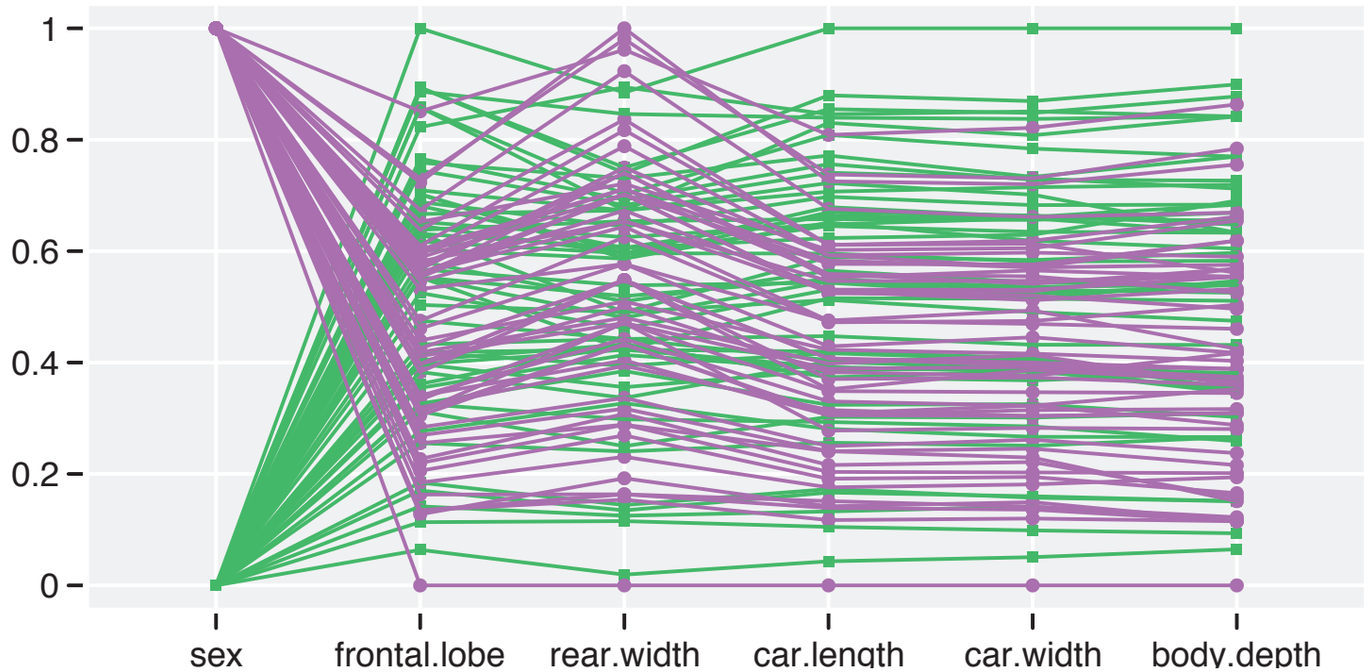


Here, each variable is scaled individually by min/max

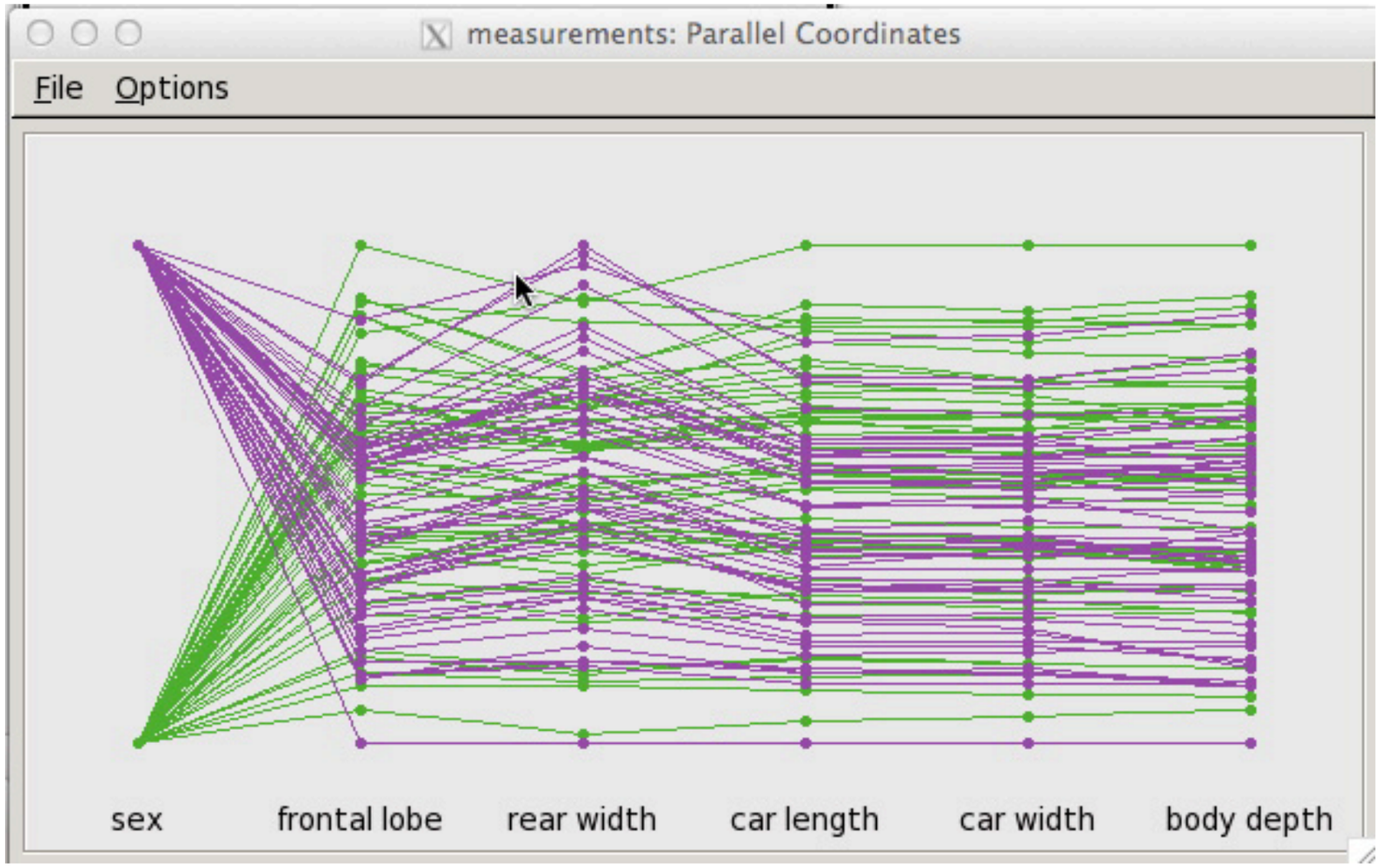


Order of variables is important



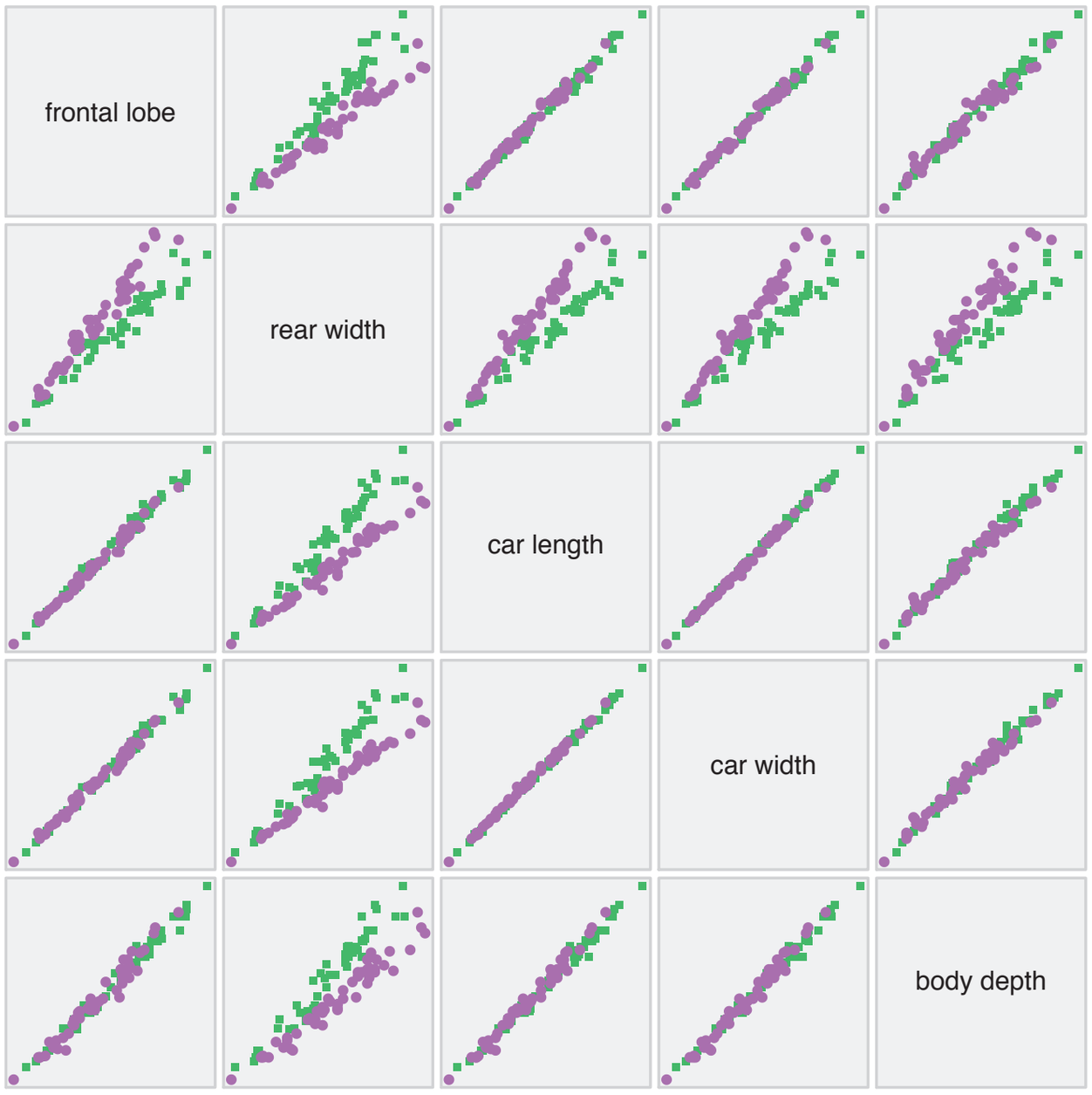


Order of variables is important



Scatterplot matrix

- Same data: all pairs of the 5 variables displayed
- Strong association between all pairs.
- Difference between males and females on "rear.width"



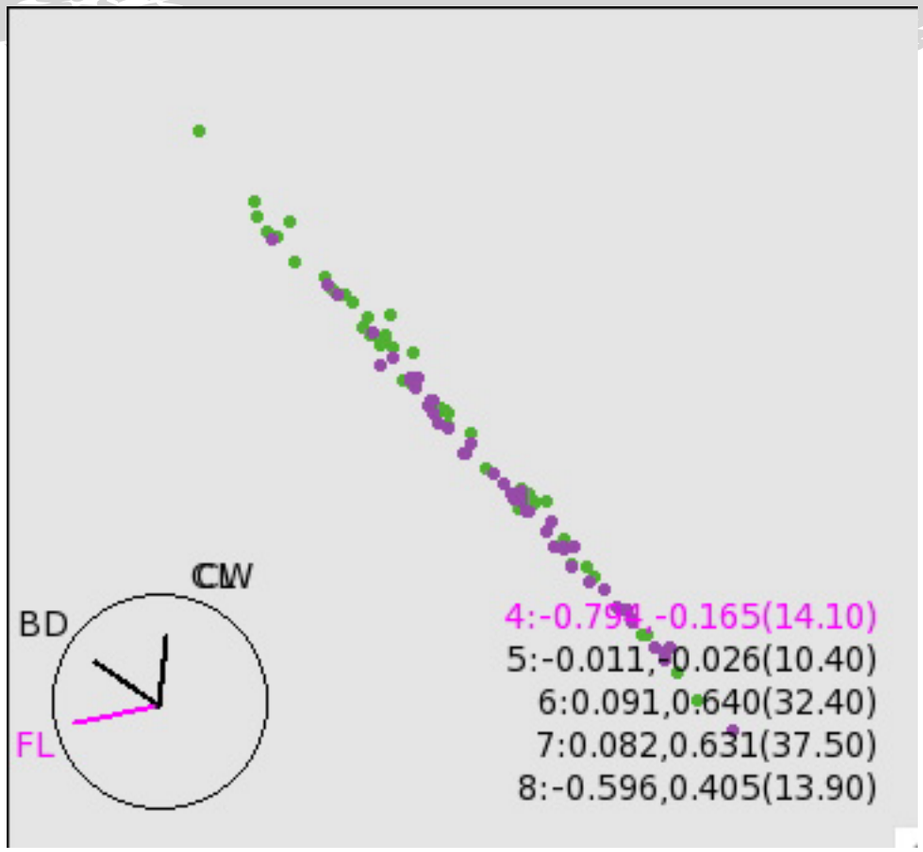
Tours

Motion graphic designed to study the joint distribution of multivariate data (Asimov 1985), in search of relationships that may involve several variables. It is created by generating a sequence of low-dimensional projections of high-dimensional data; these projections are typically 1D or 2D.

Tours

Motion graphic designed to study the joint distribution of multivariate data (Asimov 1985), in search of relationships that may involve several variables. It is created by generating a sequence of low-dimensional projections of high-dimensional data; these projections are typically 1D or 2D.

Tours



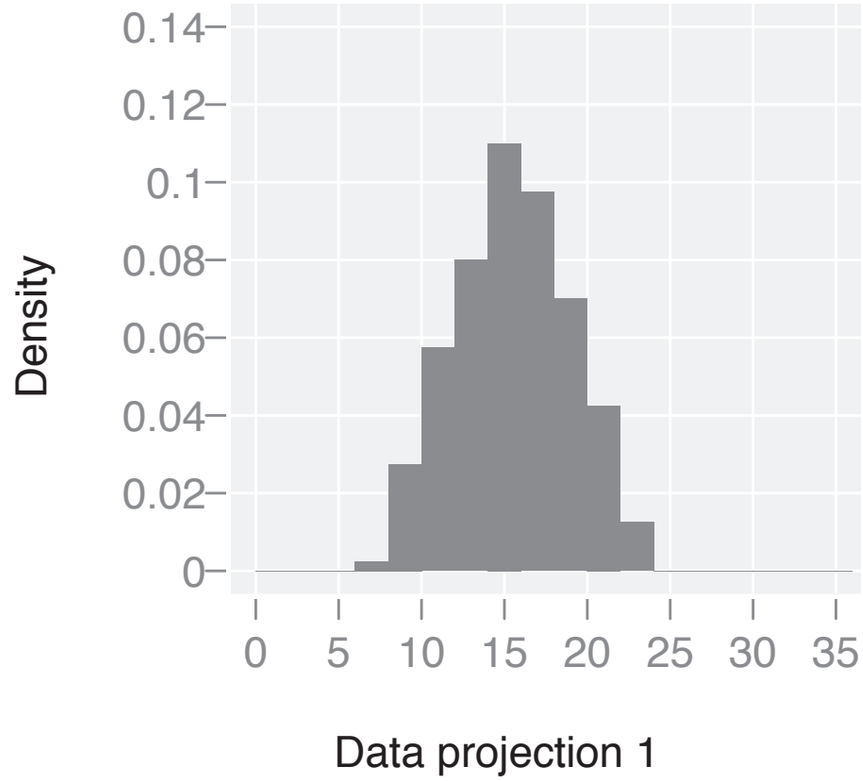
Motion graphic designed to study the joint distribution of multivariate data (Asimov 1985), in search of relationships that may involve several variables. It is created by generating a sequence of low-dimensional projections of high-dimensional data; these projections are typically 1D or 2D.

Constructing a tour

$\mathbf{X} =$

frontal lobe	rear width	carapace length	carapace width	body depth
8.1	6.7	16.1	19.0	7.0
8.8	7.7	18.1	20.8	7.4
9.2	7.8	19.0	22.4	7.7
9.6	7.9	20.1	23.1	8.2
7.2	6.5	14.7	17.1	6.1
9.0	8.5	19.3	22.7	7.7
9.1	8.1	18.5	21.6	7.7
9.1	6.9	16.7	18.6	7.4
10.2	8.2	20.2	22.2	9.0
10.7	9.7	21.4	24.0	9.8
11.4	9.2	21.7	24.1	9.7
12.5	10.0	24.1	27.0	10.9

$$\mathbf{A}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \text{then the data projection is} \quad \mathbf{X}\mathbf{A}_1 = \begin{bmatrix} 8.1 \\ 8.8 \\ 9.2 \\ 9.6 \\ 7.2 \\ \vdots \end{bmatrix}$$

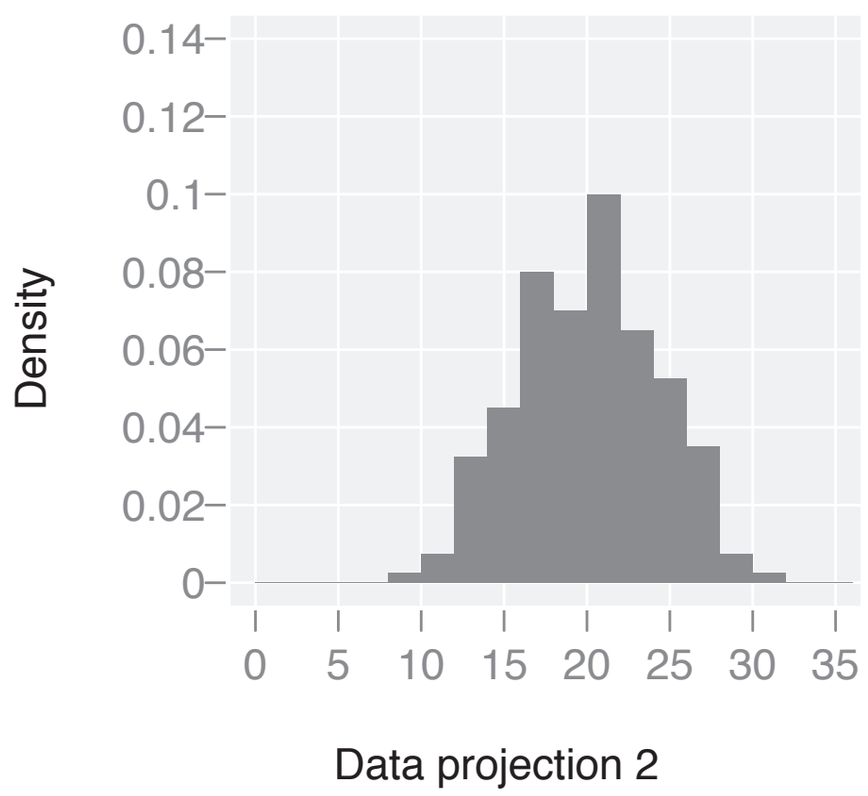


Constructing a tour

$\mathbf{X} =$

frontal lobe	rear width	carapace length	carapace width	body depth
8.1	6.7	16.1	19.0	7.0
8.8	7.7	18.1	20.8	7.4
9.2	7.8	19.0	22.4	7.7
9.6	7.9	20.1	23.1	8.2
7.2	6.5	14.7	17.1	6.1
9.0	8.5	19.3	22.7	7.7
9.1	8.1	18.5	21.6	7.7
9.1	6.9	16.7	18.6	7.4
10.2	8.2	20.2	22.2	9.0
10.7	9.7	21.4	24.0	9.8
11.4	9.2	21.7	24.1	9.7
12.5	10.0	24.1	27.0	10.9

$$\mathbf{A}_2 = \begin{bmatrix} 0.707 \\ 0.707 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \text{ then } \mathbf{X}\mathbf{A}_2 = \begin{bmatrix} 0.707 \times 8.1 + 0.707 \times 6.7 = 10.5 \\ 0.707 \times 8.8 + 0.707 \times 7.7 = 11.7 \\ 0.707 \times 9.2 + 0.707 \times 7.8 = 12.0 \\ 0.707 \times 9.6 + 0.707 \times 7.9 = 12.4 \\ 0.707 \times 7.2 + 0.707 \times 6.5 = 9.7 \\ \vdots \end{bmatrix}$$

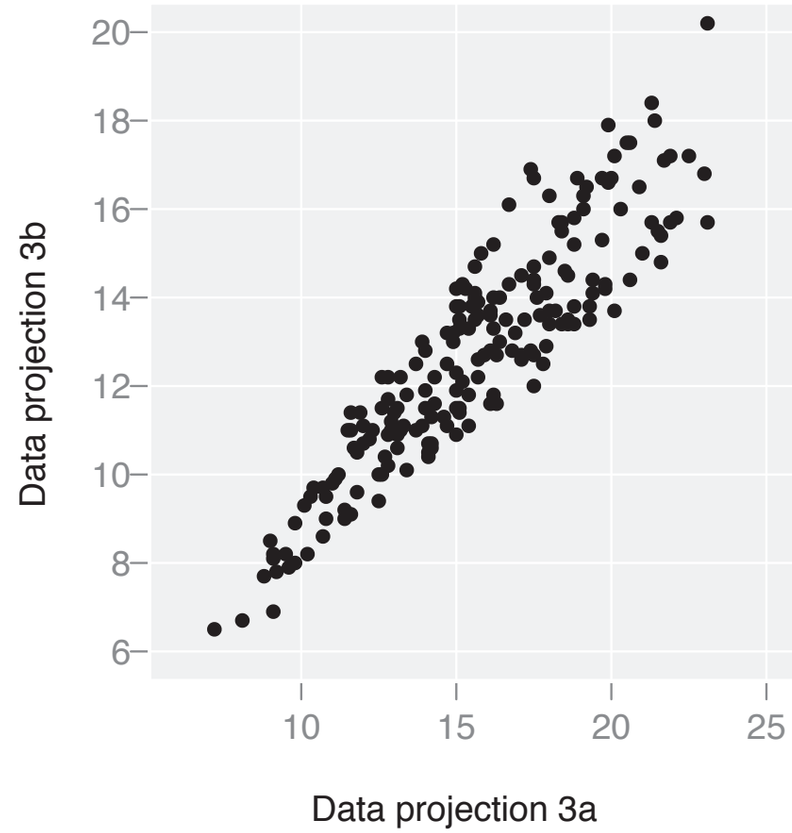


Constructing a tour

$\mathbf{X} =$

frontal lobe	rear width	carapace length	carapace width	body depth
8.1	6.7	16.1	19.0	7.0
8.8	7.7	18.1	20.8	7.4
9.2	7.8	19.0	22.4	7.7
9.6	7.9	20.1	23.1	8.2
7.2	6.5	14.7	17.1	6.1
9.0	8.5	19.3	22.7	7.7
9.1	8.1	18.5	21.6	7.7
9.1	6.9	16.7	18.6	7.4
10.2	8.2	20.2	22.2	9.0
10.7	9.7	21.4	24.0	9.8
11.4	9.2	21.7	24.1	9.7
12.5	10.0	24.1	27.0	10.9

$$\mathbf{A}_3 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}, \quad \text{then the data projection is} \quad \mathbf{X}\mathbf{A}_3 = \begin{bmatrix} 8.1 & 6.7 \\ 8.8 & 7.7 \\ 9.2 & 7.8 \\ 9.6 & 7.9 \\ 7.2 & 6.5 \\ \vdots & \end{bmatrix}$$



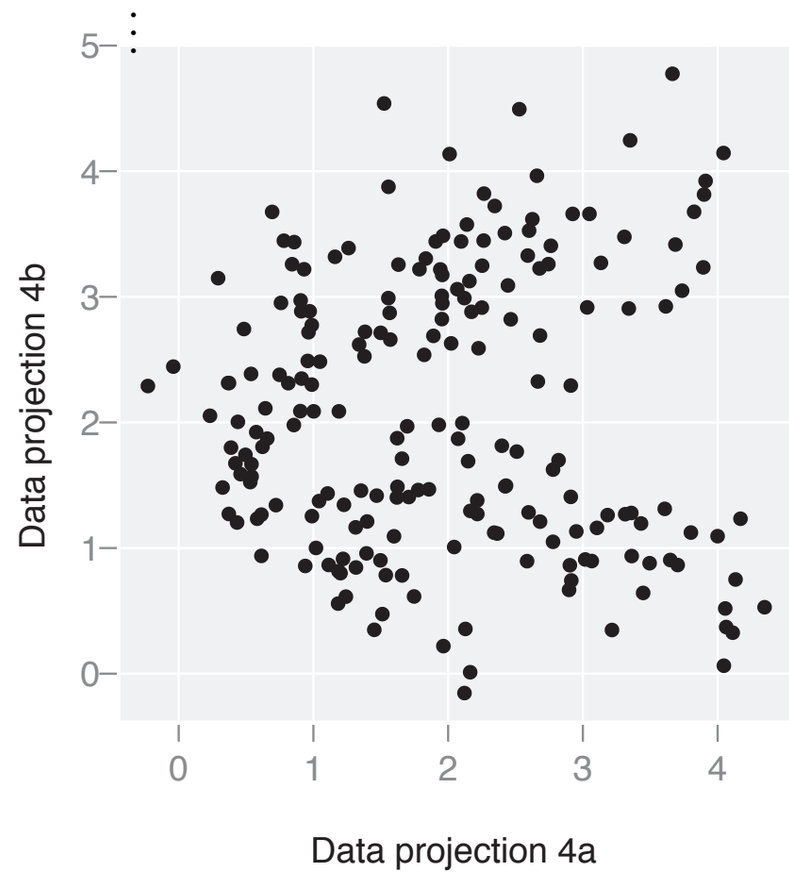
Constructing a tour

$\mathbf{X} =$

frontal lobe	rear width	carapace length	carapace width	body depth
8.1	6.7	16.1	19.0	7.0
8.8	7.7	18.1	20.8	7.4
9.2	7.8	19.0	22.4	7.7
9.6	7.9	20.1	23.1	8.2
7.2	6.5	14.7	17.1	6.1
9.0	8.5	19.3	22.7	7.7
9.1	8.1	18.5	21.6	7.7
9.1	6.9	16.7	18.6	7.4
10.2	8.2	20.2	22.2	9.0
10.7	9.7	21.4	24.0	9.8
11.4	9.2	21.7	24.1	9.7
12.5	10.0	24.1	27.0	10.9

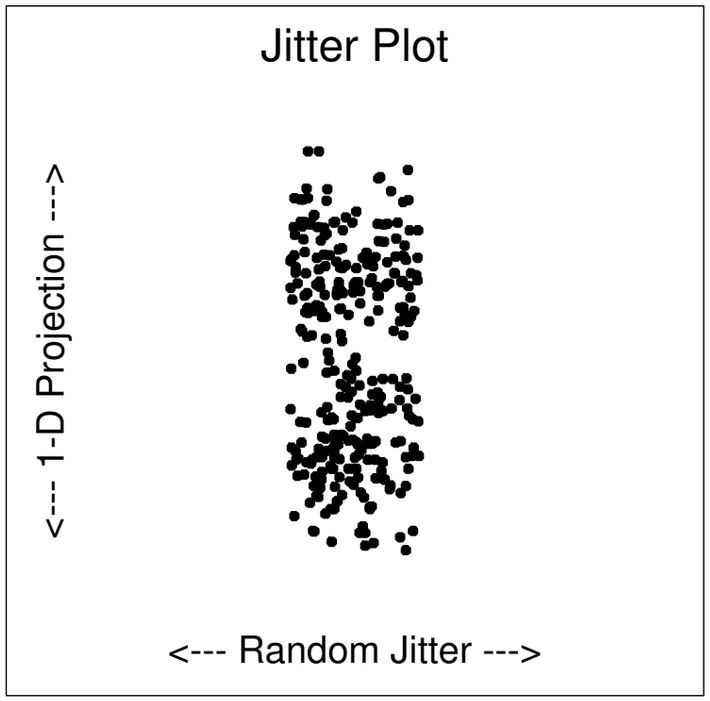
$$\mathbf{A}_4 = \begin{bmatrix} 0 & 0 \\ 0 & 0.950 \\ 0 & -0.312 \\ -0.312 & 0 \\ 0.950 & 0 \end{bmatrix} \quad \text{then} \quad \mathbf{XA}_4 =$$

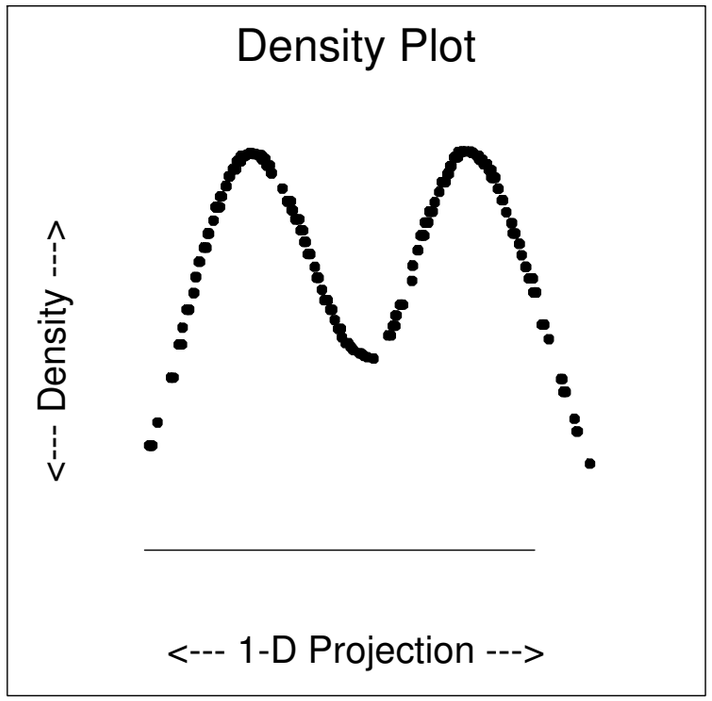
$$\begin{bmatrix} -0.312 \times 19.0 + 0.950 \times 7.0 = 0.72 & 0.950 \times 6.7 - 0.312 \times 16.1 = 1.34 \\ -0.312 \times 20.8 + 0.950 \times 7.4 = 0.54 & 0.950 \times 7.7 - 0.312 \times 18.1 = 1.67 \\ -0.312 \times 22.4 + 0.950 \times 7.7 = 0.33 & 0.950 \times 7.8 - 0.312 \times 19.0 = 1.48 \\ -0.312 \times 23.1 + 0.950 \times 8.2 = 0.58 & 0.950 \times 7.9 - 0.312 \times 20.1 = 1.23 \\ -0.312 \times 17.1 + 0.950 \times 6.1 = 0.46 & 0.950 \times 6.5 - 0.312 \times 14.7 = 1.59 \\ \vdots & \end{bmatrix}$$

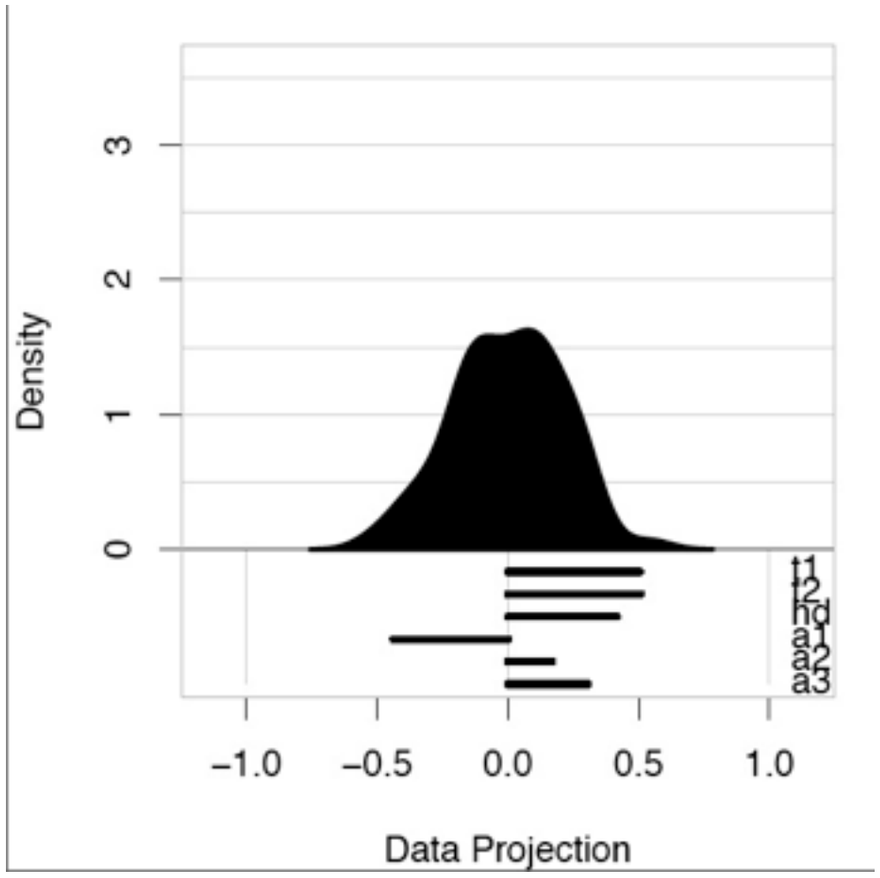


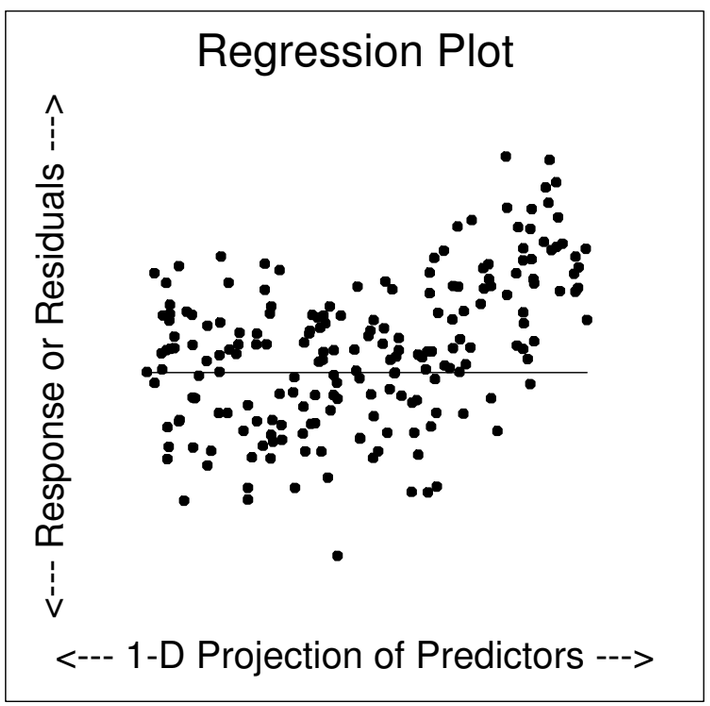
Sequencing the projections

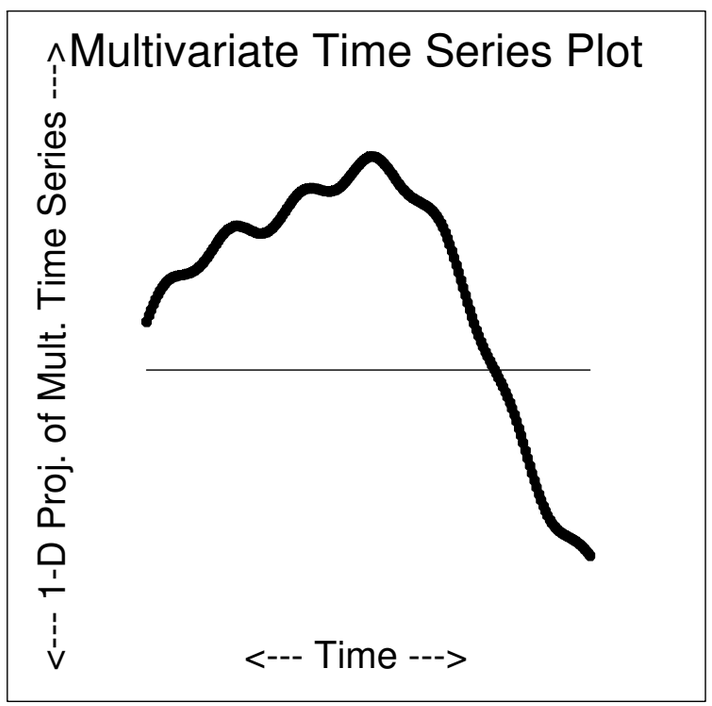
- Tours to display strictly real-valued multivariate data
- $View_i(t) = F(t)^T \boldsymbol{x}_i$, $F(t) = (\boldsymbol{f}_1(t), \ldots, \boldsymbol{f}_d(t))$
- Render the view, and navigation info
- Method for choosing $F(t)$
- Interpolate between consecutive $F(t)$

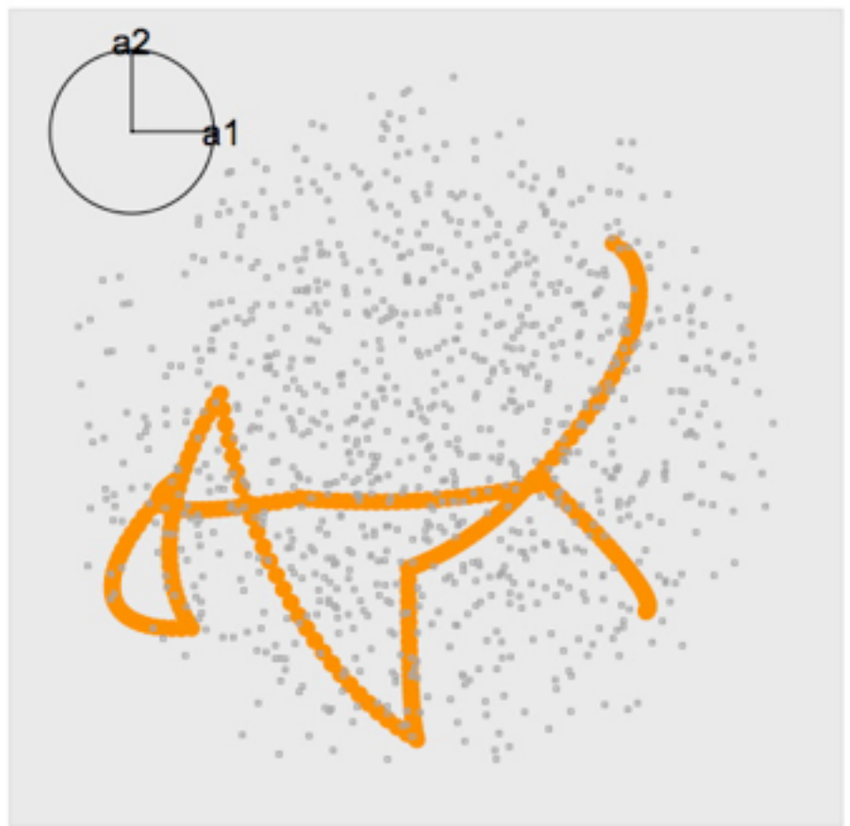

Jitter Plot
<--- 1-D Projection --->
<--- Random Jitter --->

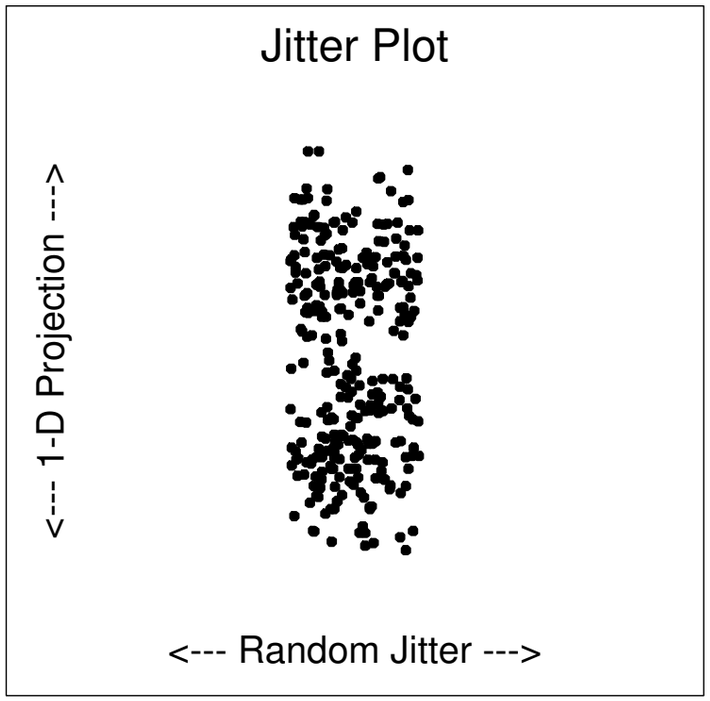

Density Plot
<--- Density --->
<--- 1-D Projection --->

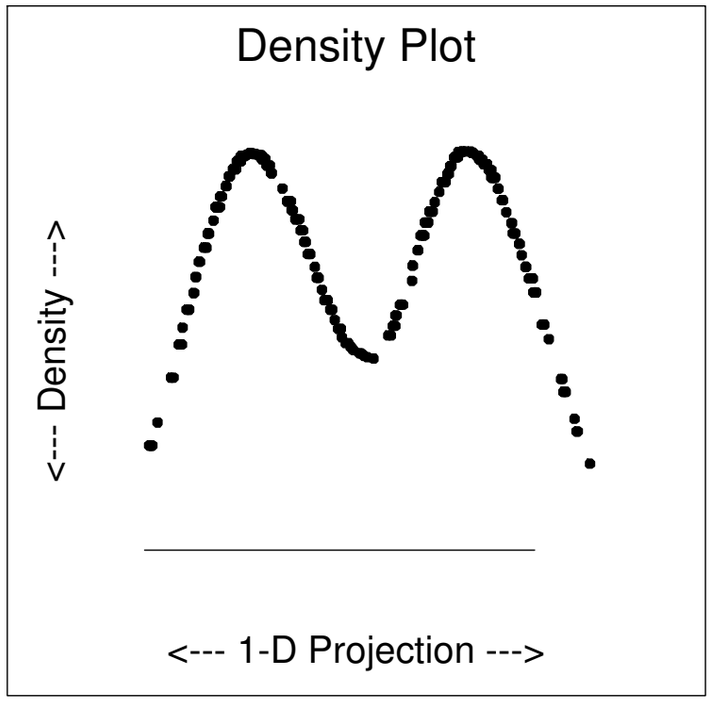

Density
0
1
2
3
−1.0
−0.5
0.0
0.5
1.0
Data Projection
t1
t2
h0
a1
a2
a3

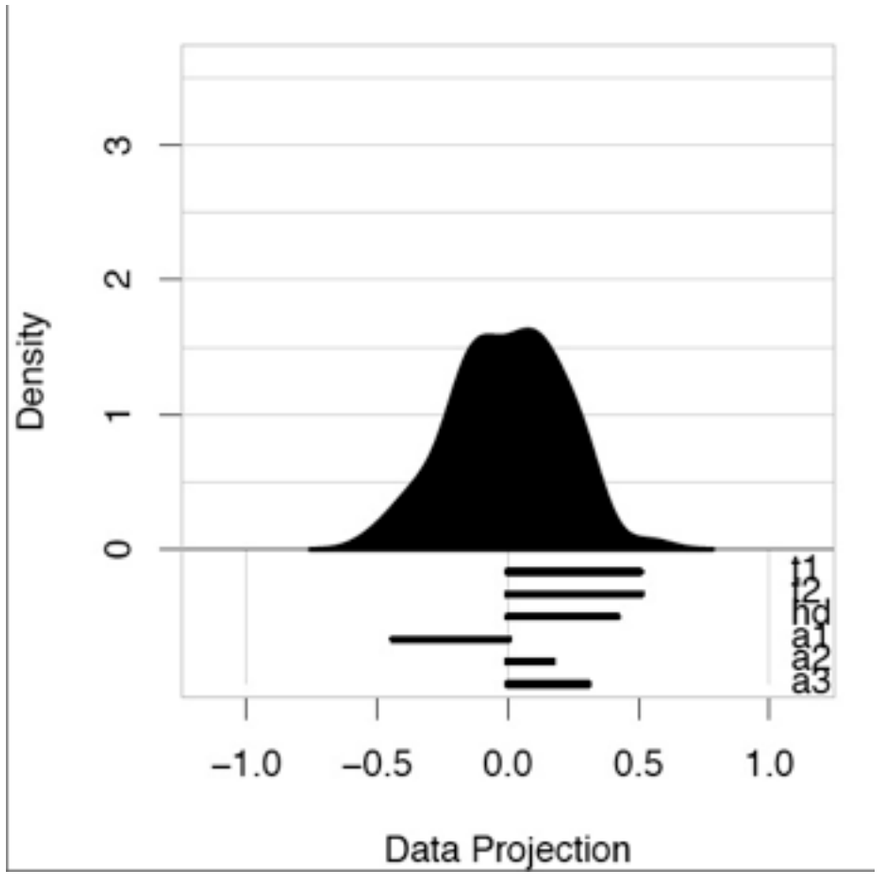

Regression Plot
<--- Response or Residuals --->
<--- 1-D Projection of Predictors --->

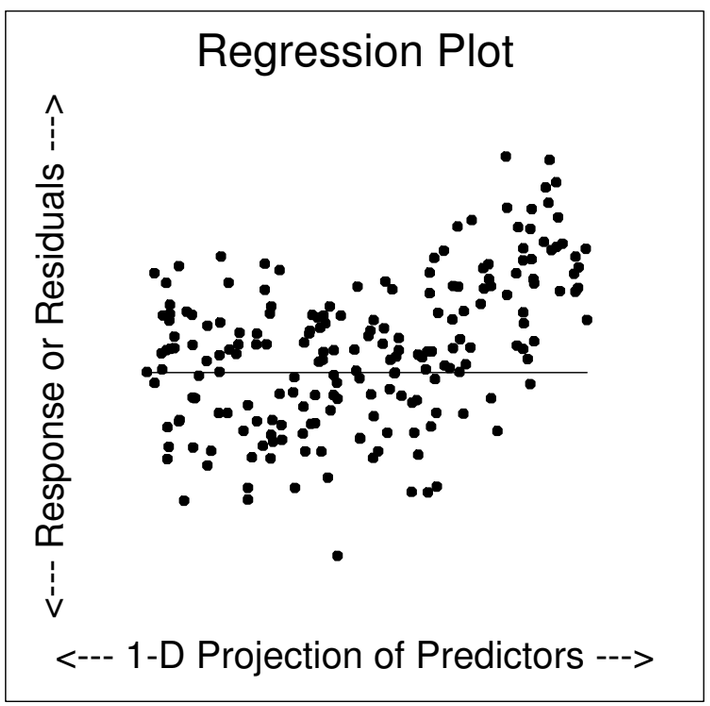

Multivariate Time Series Plot
<--- 1-D Proj. of Mult. Time Series --->
<--- Time --->

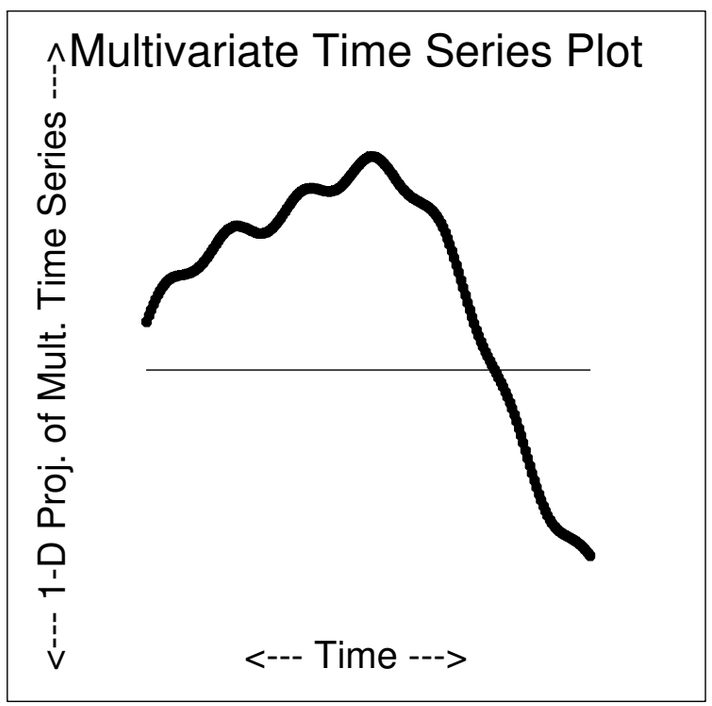

a2
a1

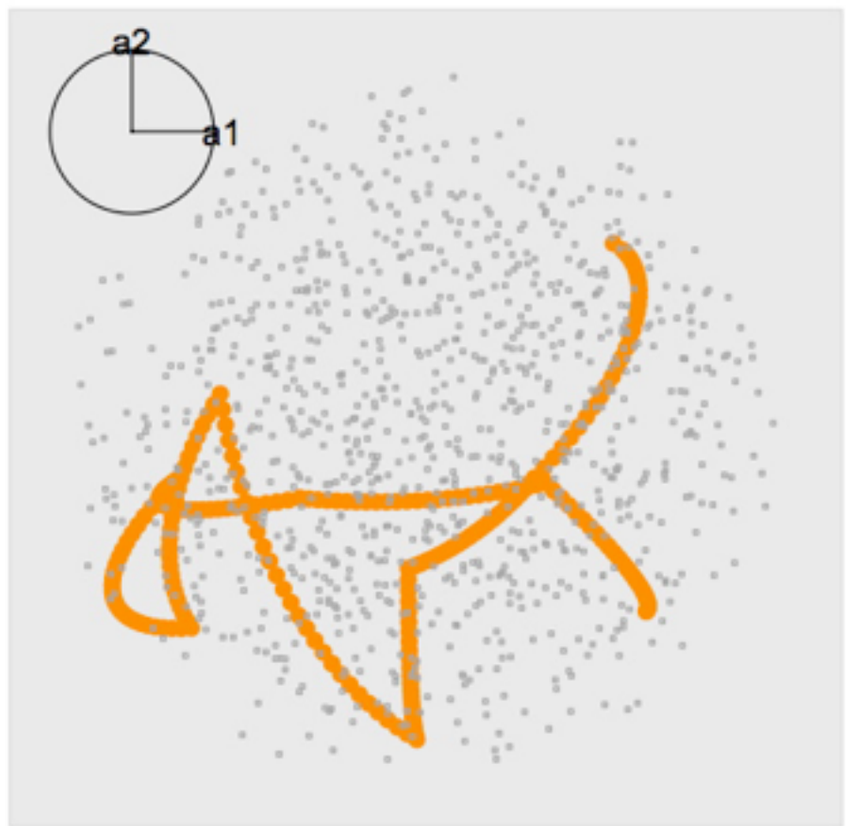

Jitter Plot
<--- 1-D Projection --->
<--- Random Jitter --->

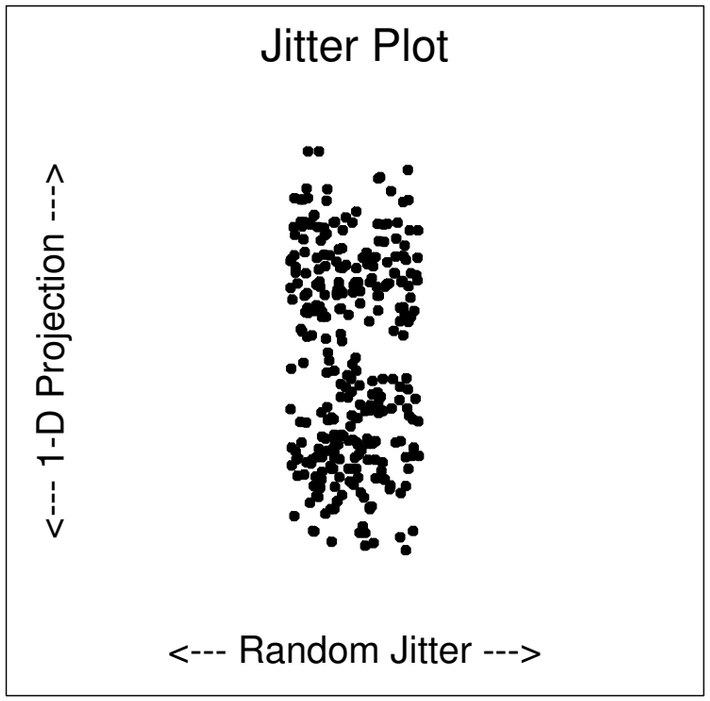

Density Plot
<--- Density --->
<--- 1-D Projection --->

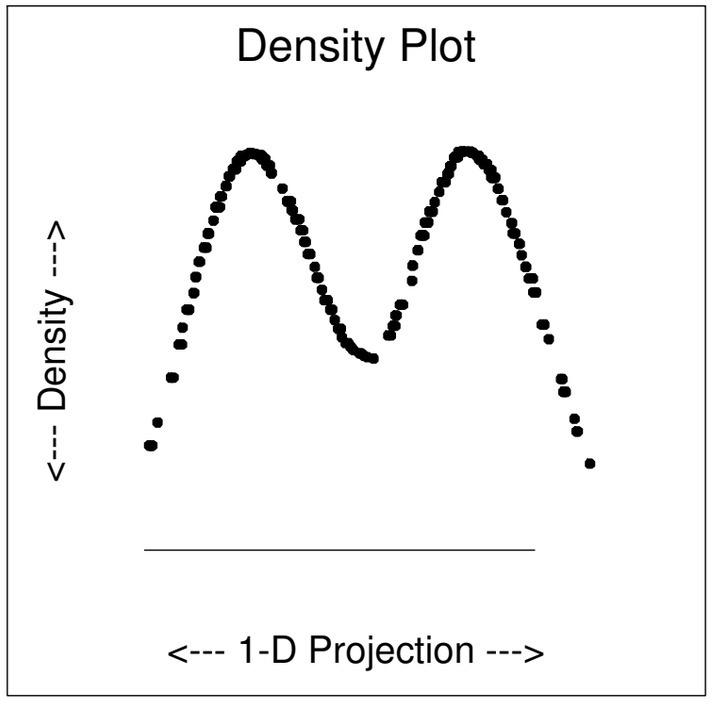

Density
0
1
2
3
−1.0
−0.5
0.0
0.5
1.0
Data Projection
t1
t2
h0
a1
a2
a3

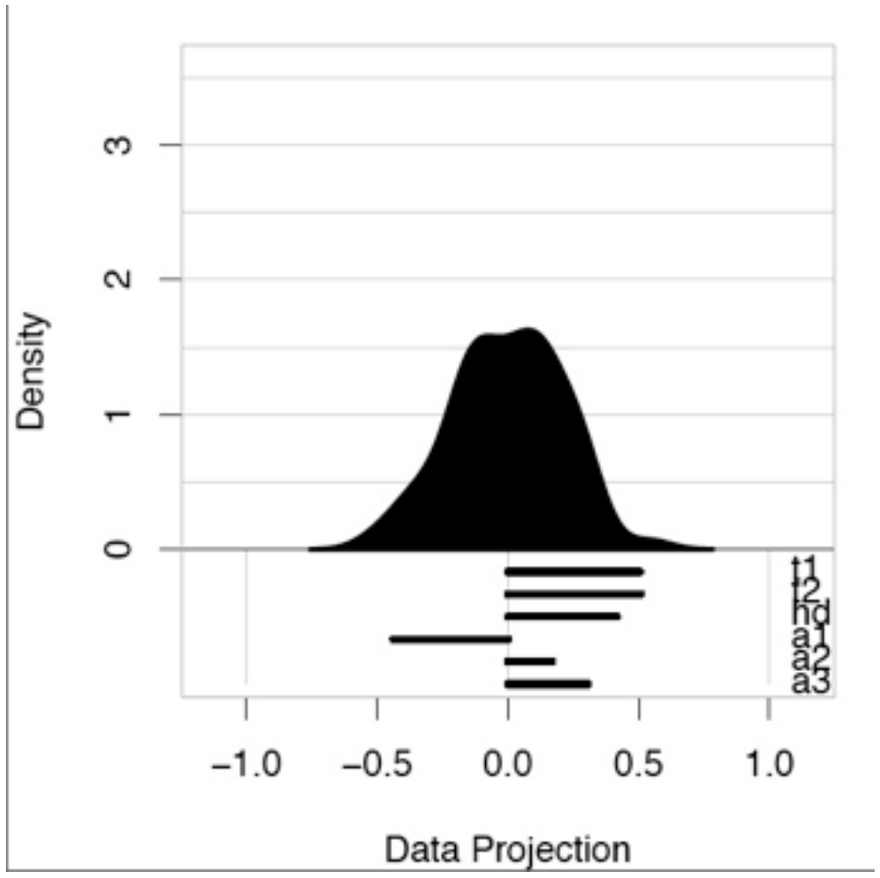

Regression Plot
<--- Response or Residuals --->
<--- 1-D Projection of Predictors --->

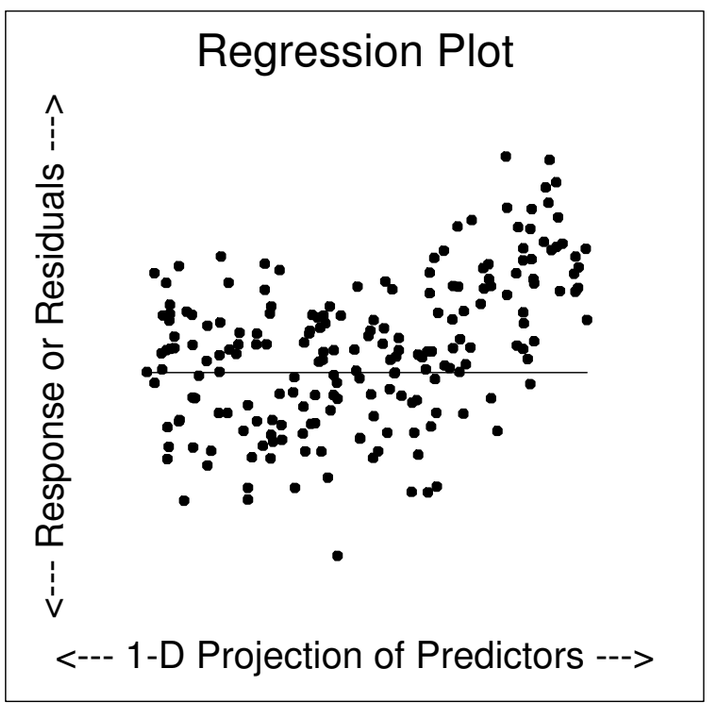

Multivariate Time Series Plot
<--- 1-D Proj. of Mult. Time Series --->
<--- Time --->

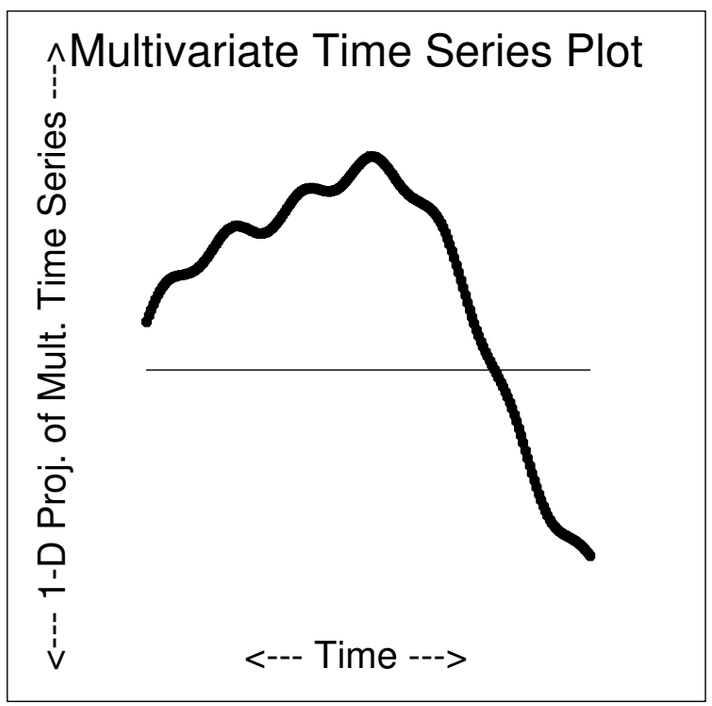

a2
a1

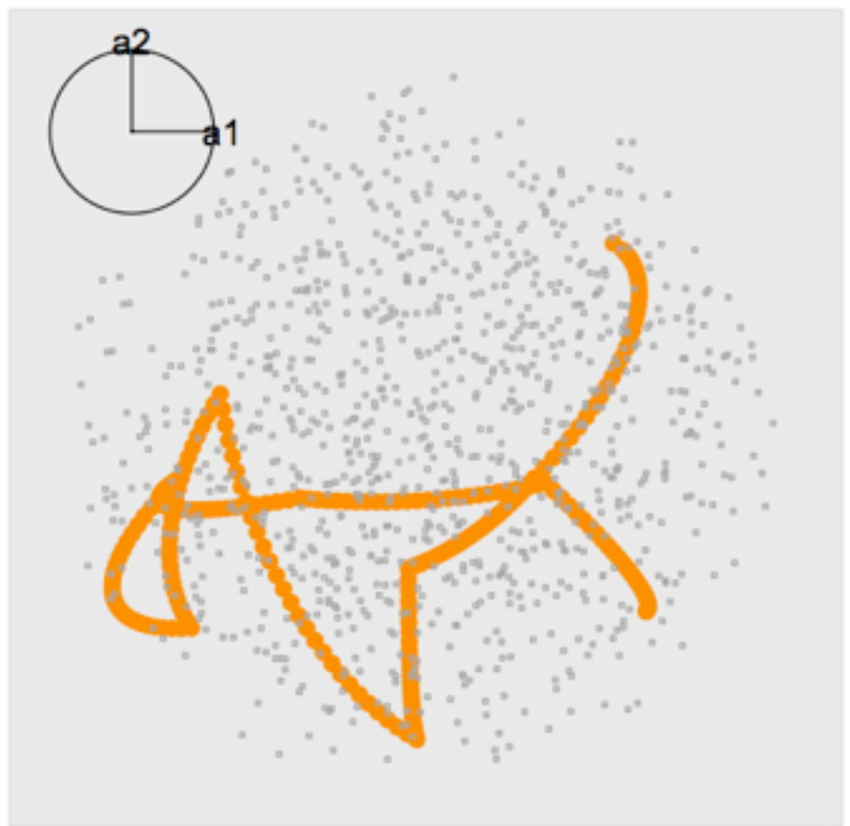

Jitter Plot
<--- 1-D Projection --->
<--- Random Jitter --->


Density Plot
<--- Density --->
<--- 1-D Projection --->


Density
0
1
2
3
−1.0
−0.5
0.0
0.5
1.0
Data Projection
t1
t2
h0
a1
a2
a3


Regression Plot
<--- Response or Residuals --->
<--- 1-D Projection of Predictors --->


Multivariate Time Series Plot
<--- 1-D Proj. of Mult. Time Series --->
<--- Time --->


a2
a1

Choosing F

- Grand tour: sample from a uniform on a sphere
- Guided: Optimize a projection pursuit index, eg

$$I_{LDA}(F) = 1 - \frac{|F^T W F|}{|F^T (W+B) F|}$$

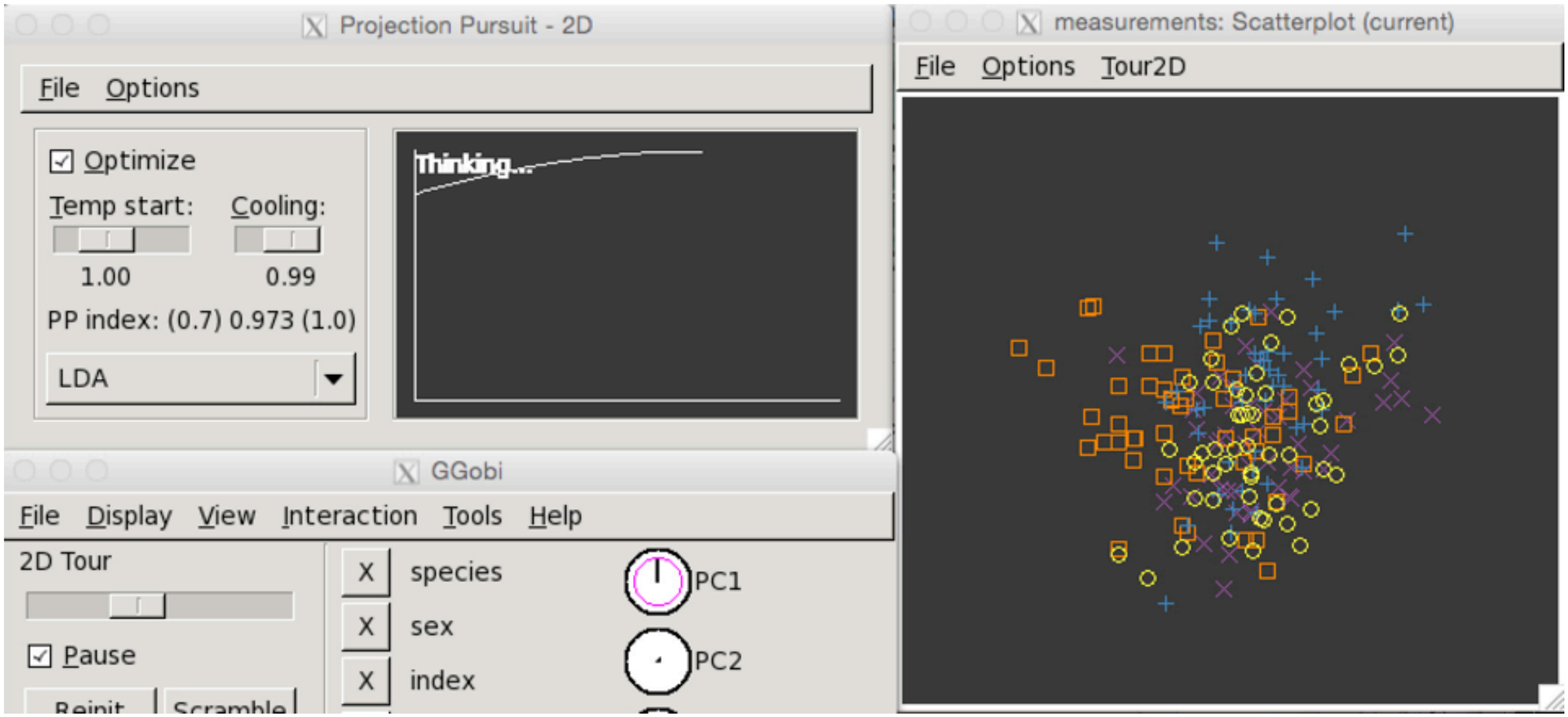
$$y = F^T x$$

$$B = \sum_{i=1}^{g} n_i (\bar{y}_{i.} - \bar{y}_{..})(\bar{y}_{i.} - \bar{y}_{..})^T$$

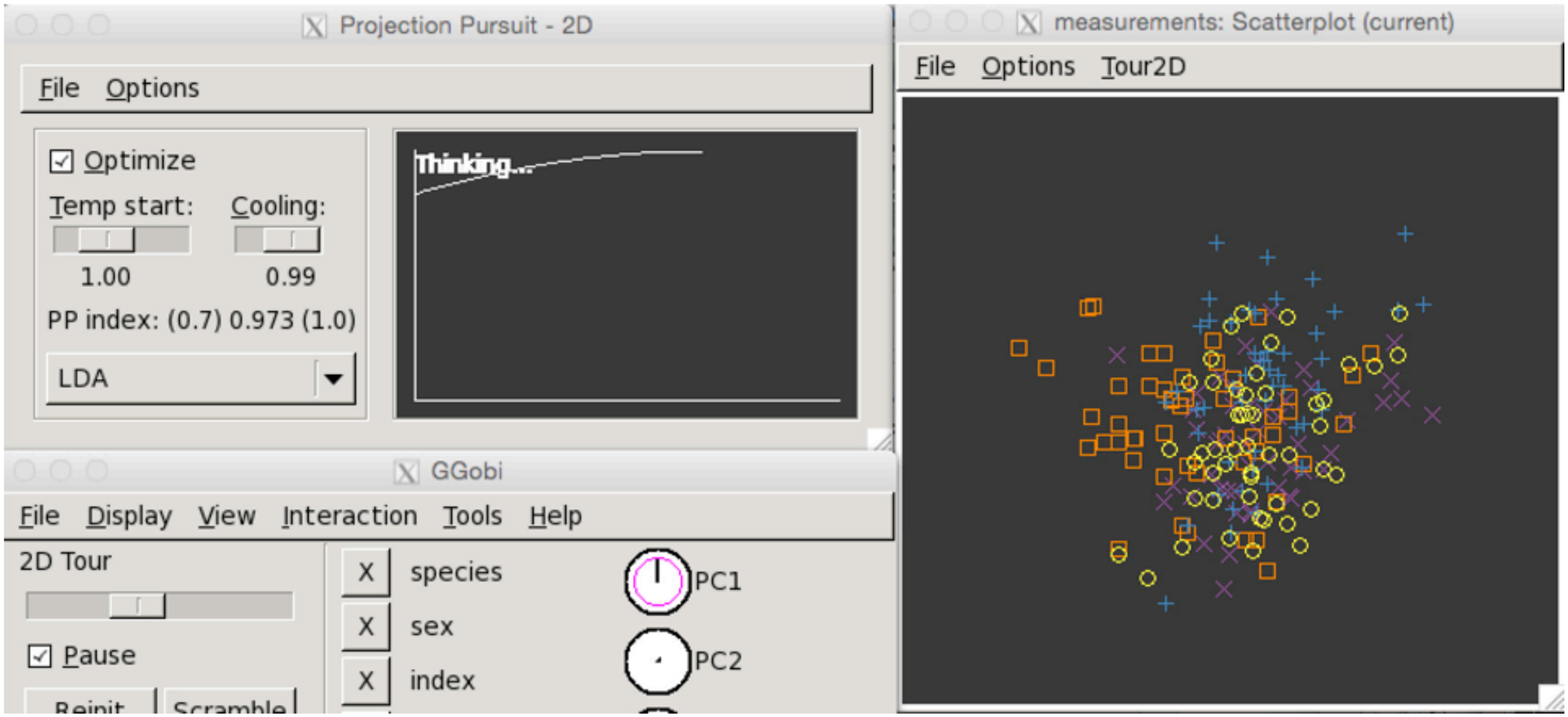
$$W = \sum_{i=1}^{g} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})(y_{ij} - \bar{y}_{i.})^T$$

- Manual: Choose a variable to control, allow user to interactively control coefficient, ranging between -1, 1, constrained on all other variables

Guided tour: four groups of crabs, optimizing PP index



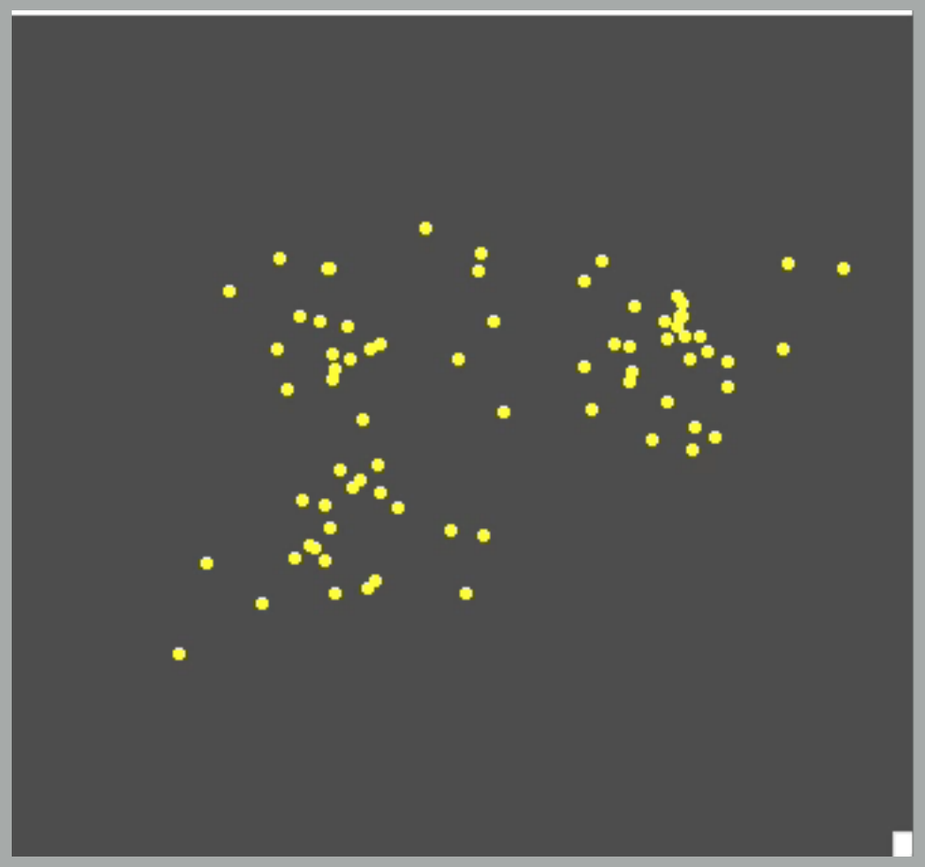
Guided tour: four groups of crabs, optimizing PP index



Why?

- Learn how several variables jointly vary
- Examine the multivariate distribution
- Check model fit
- Explore deviations from distribution: outliers, clusters, nonlinear relationships

Your Turn

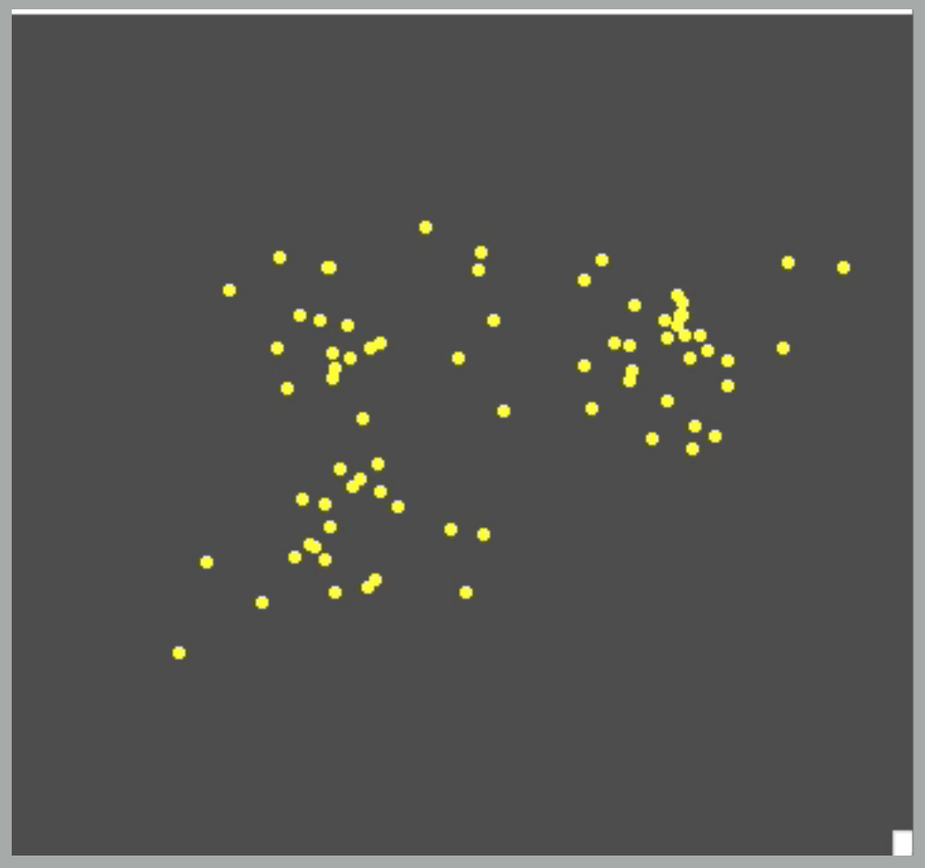


How many clusters?

Anything else you see?

Your Turn

For each of the following videos answer these questions

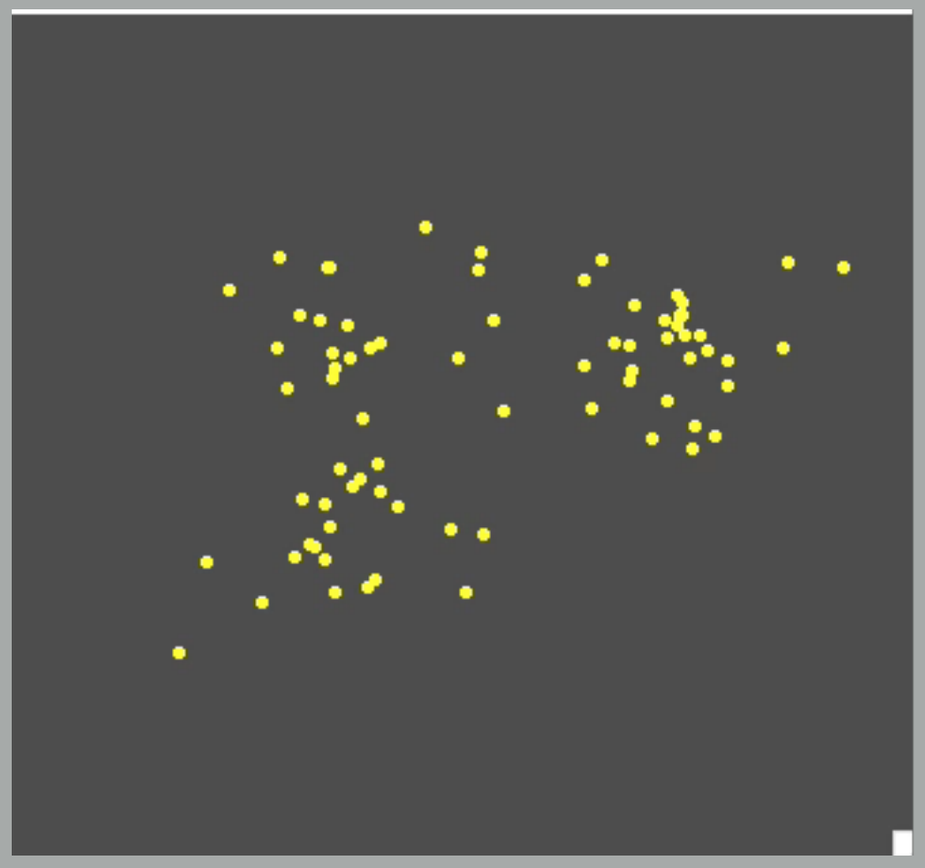


How many clusters?

Anything else you see?

Your Turn

For each of the following videos answer these questions

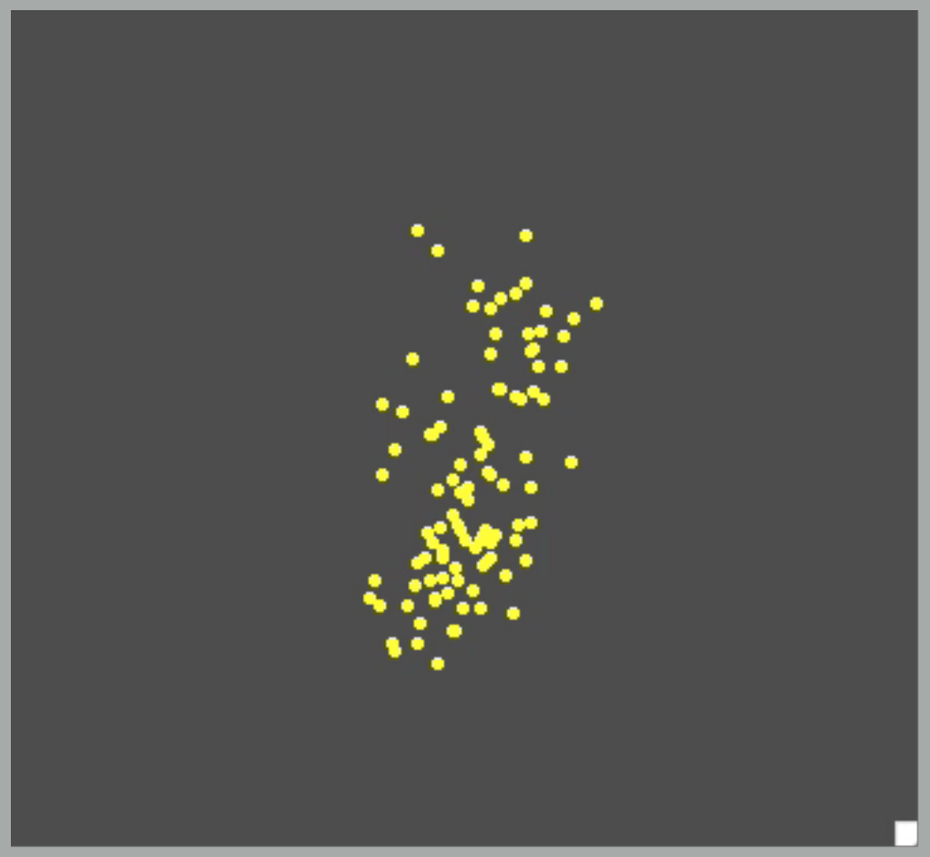


How many clusters?

Anything else you see?

Your Turn

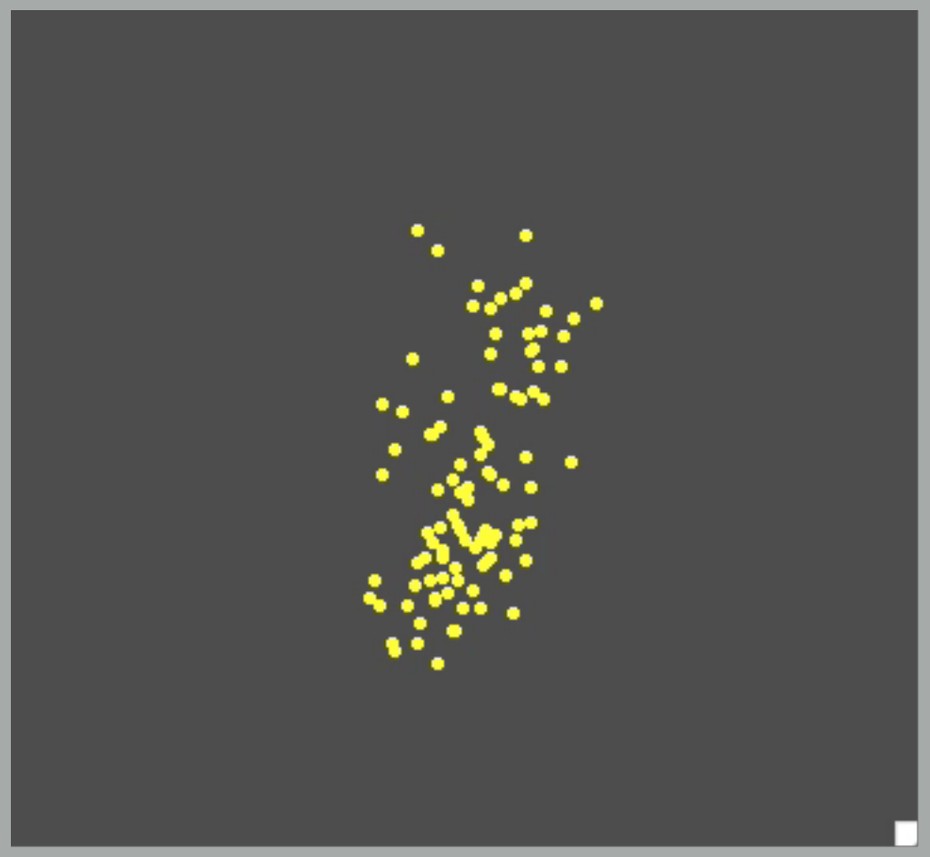
For each of the following videos answer these questions



Linear dependence, or
nonlinear dependence?

Your Turn

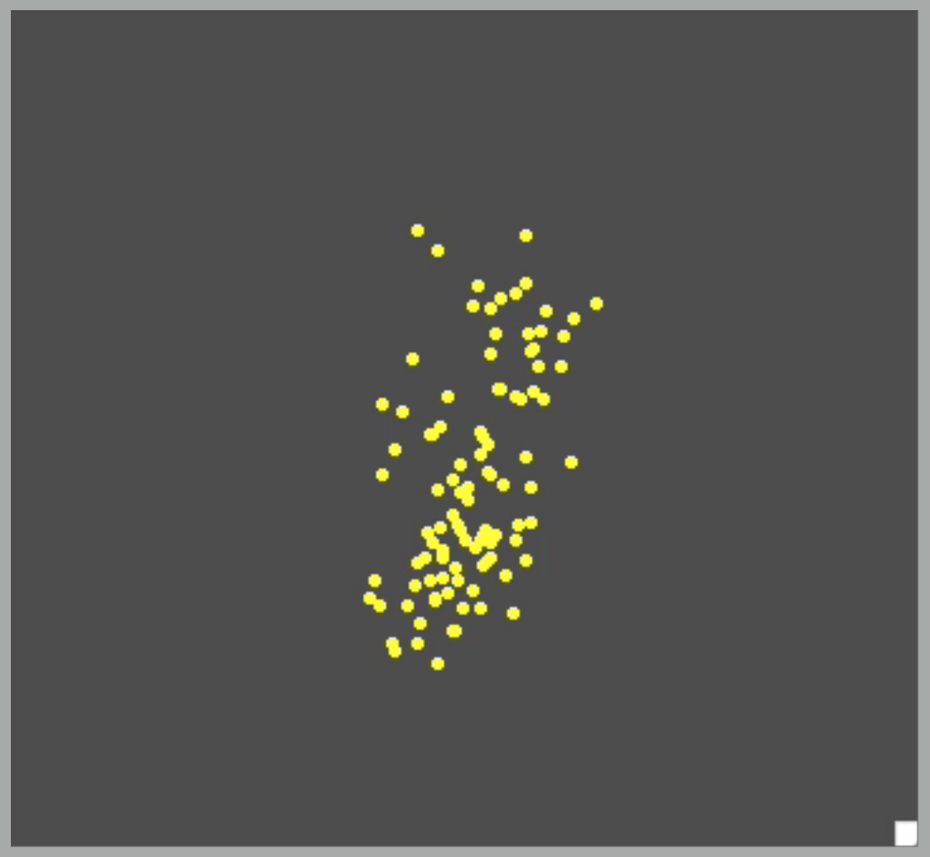
For each of the following videos answer these questions



Linear dependence, or
nonlinear dependence?

Your Turn

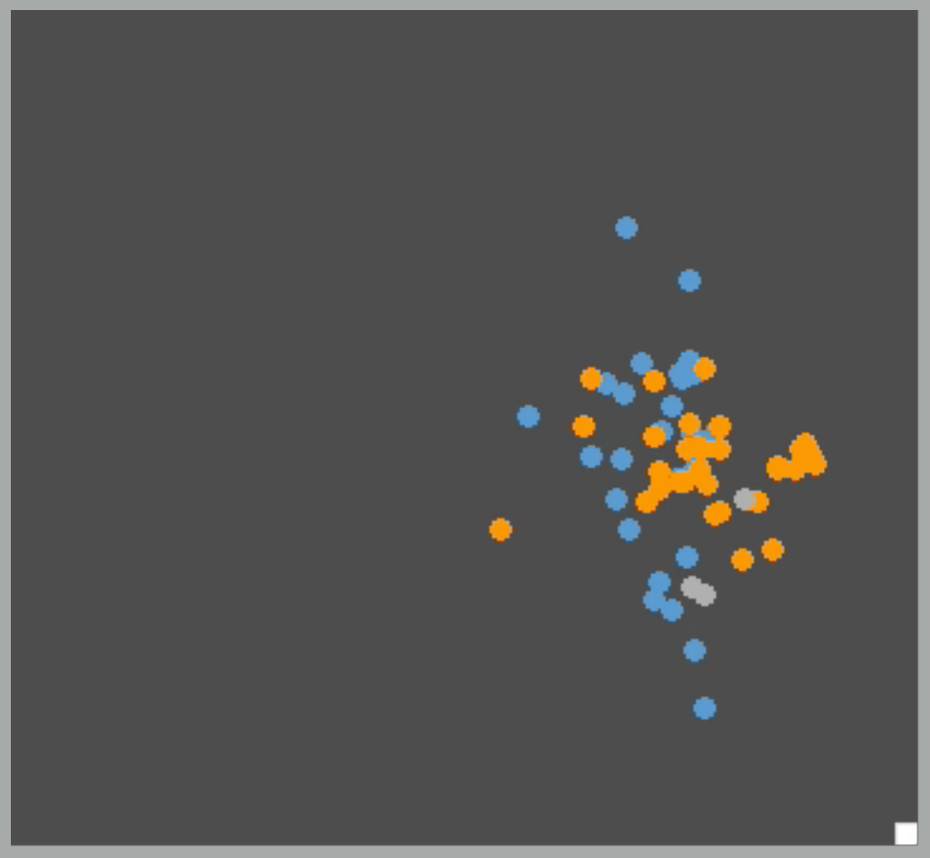
For each of the following videos answer these questions



Linear dependence, or
nonlinear dependence?

Your Turn

For each of the following videos answer these questions



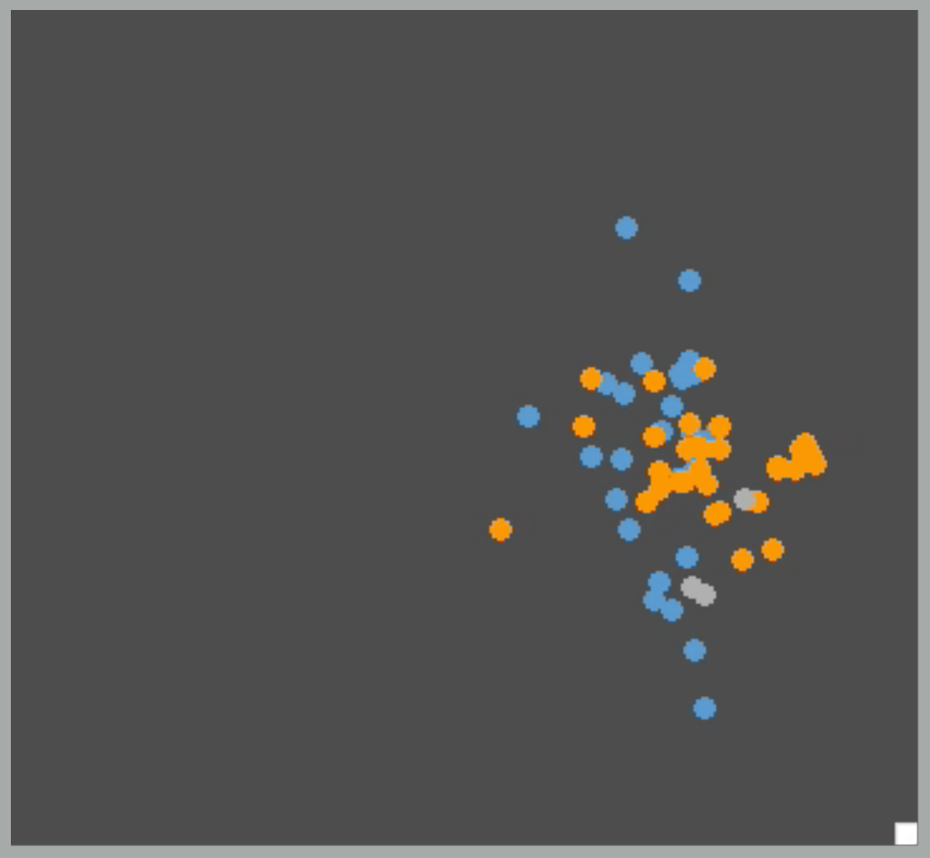
Are the two groups different from each other?

Do you see any outliers?

Any small clusters?

Your Turn

For each of the following videos answer these questions



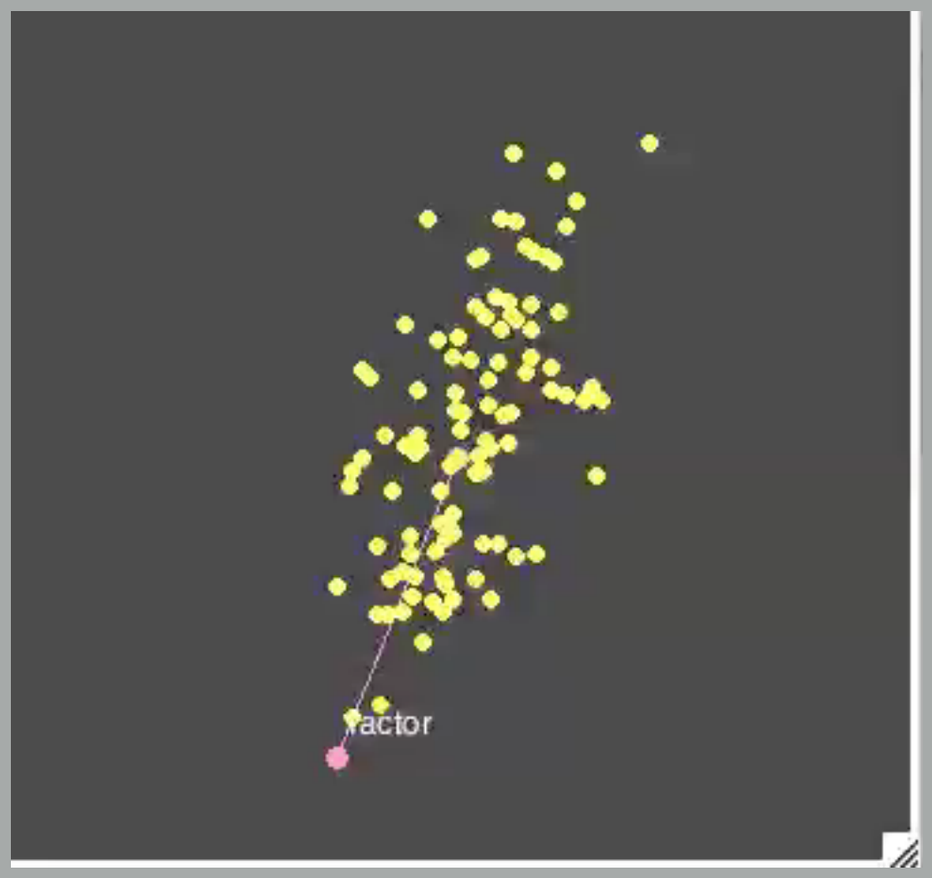
Are the two groups different from each other?

Do you see any outliers?

Any small clusters?

Your Turn

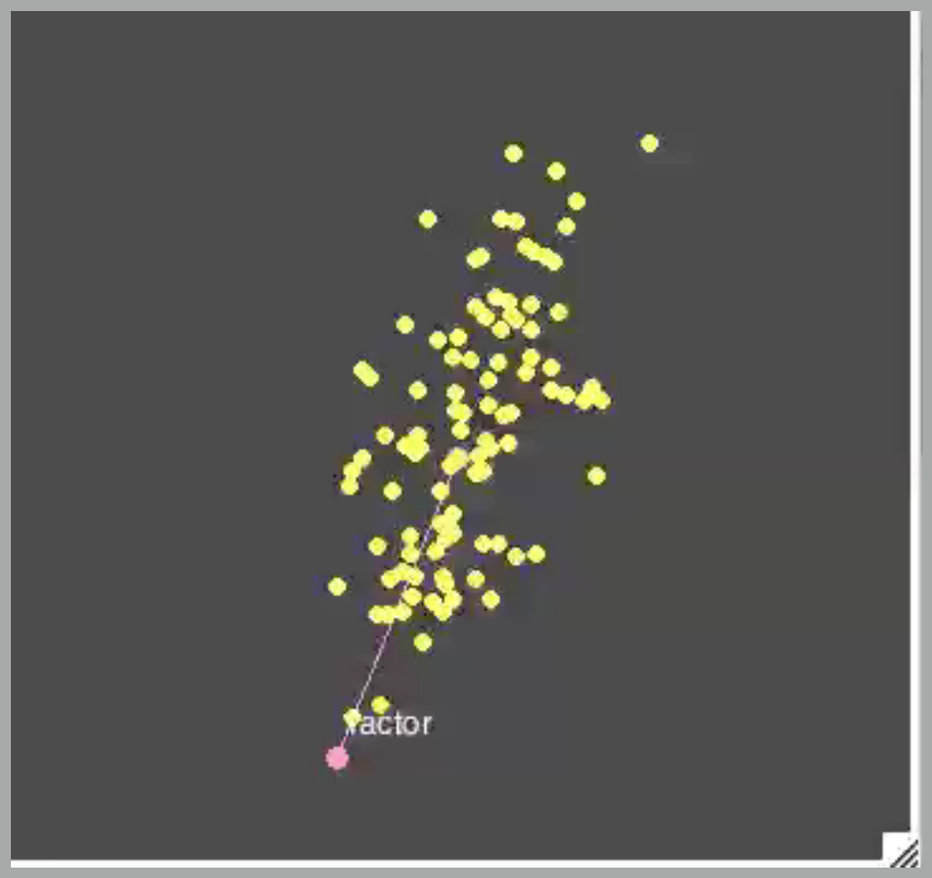
For each of the following videos answer these questions



One- or two-dimensional?

Your Turn

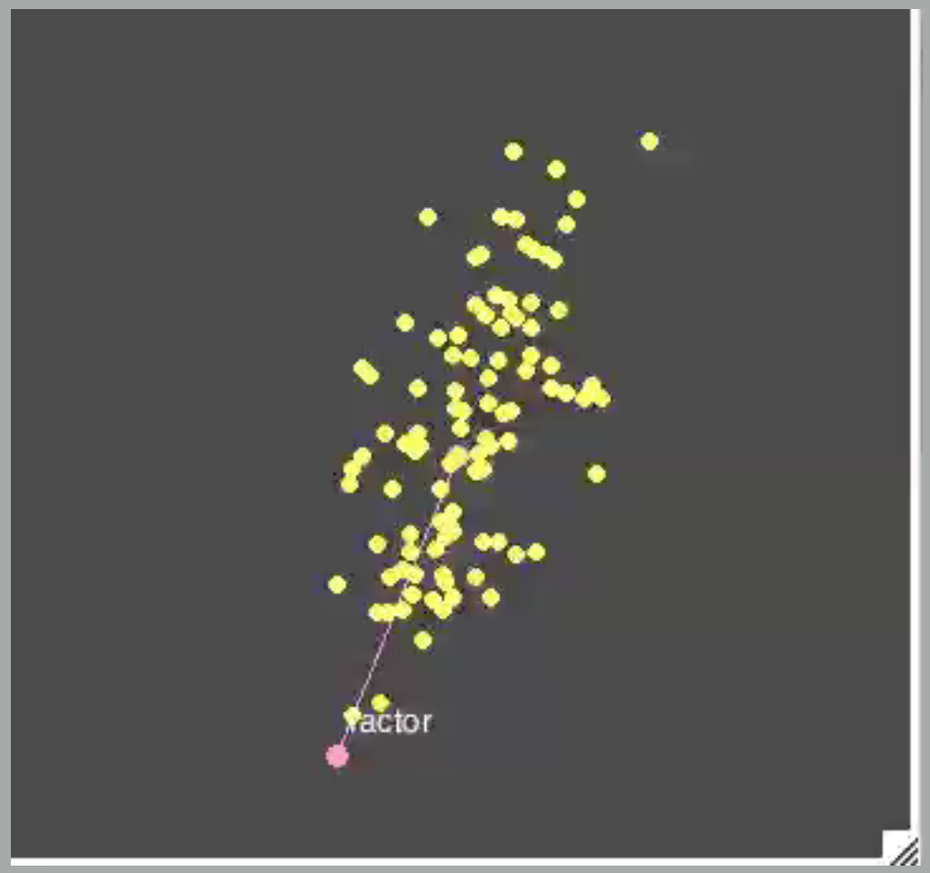
For each of the following videos answer these questions



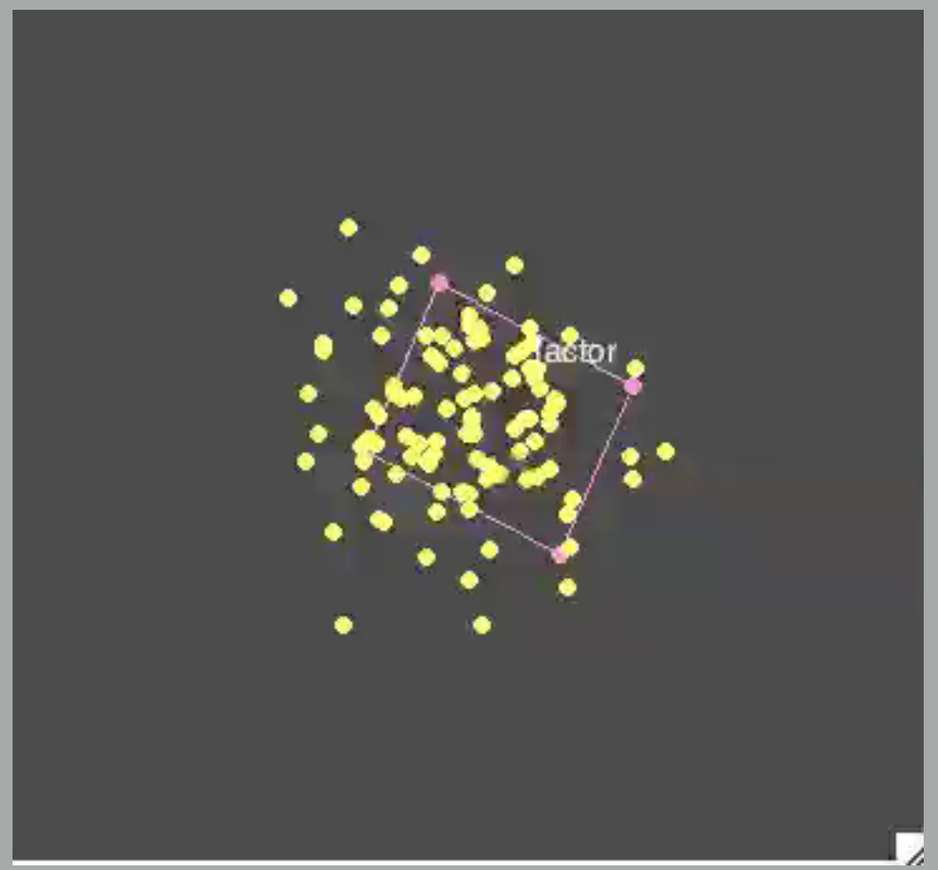
One- or two-dimensional?

Your Turn

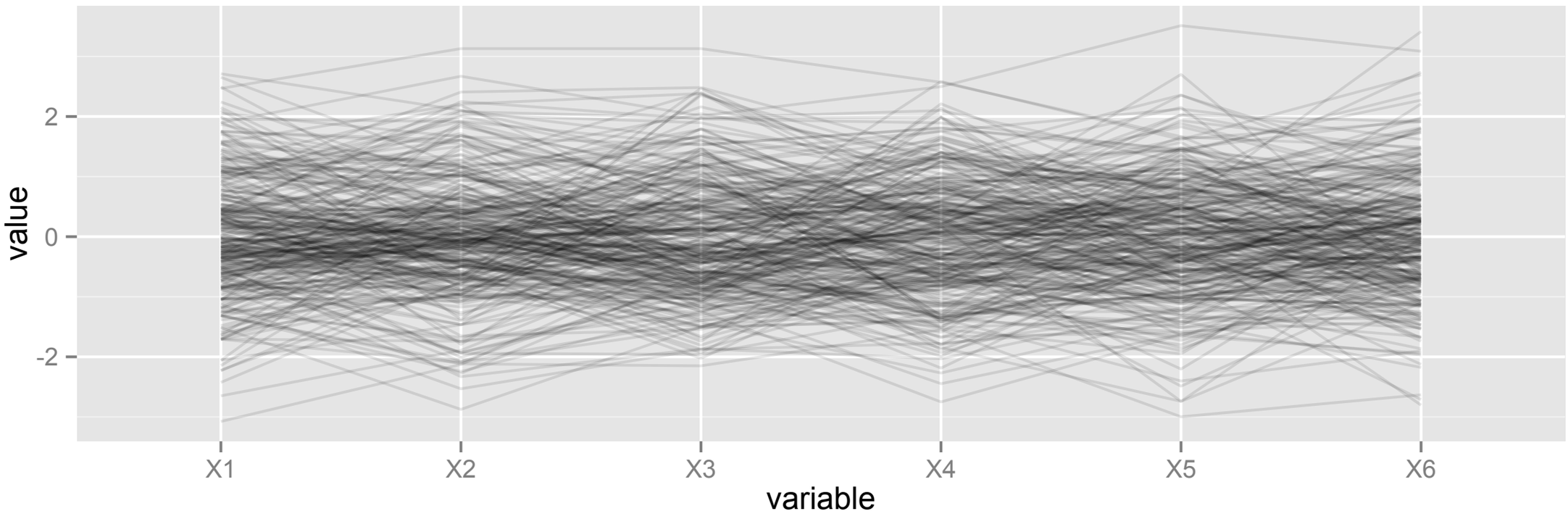
For each of the following videos answer these questions



One- or two-dimensional?



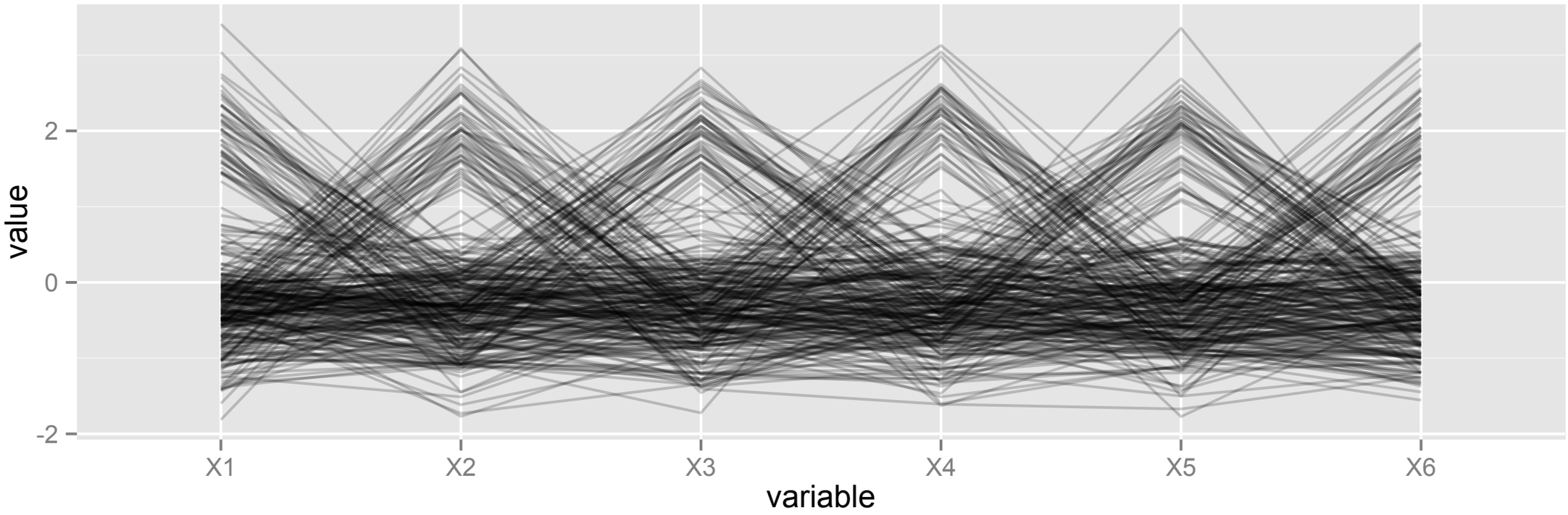
Normal data



Nothing interesting! All a little moderate correlation.

Modelling is going to be easy!

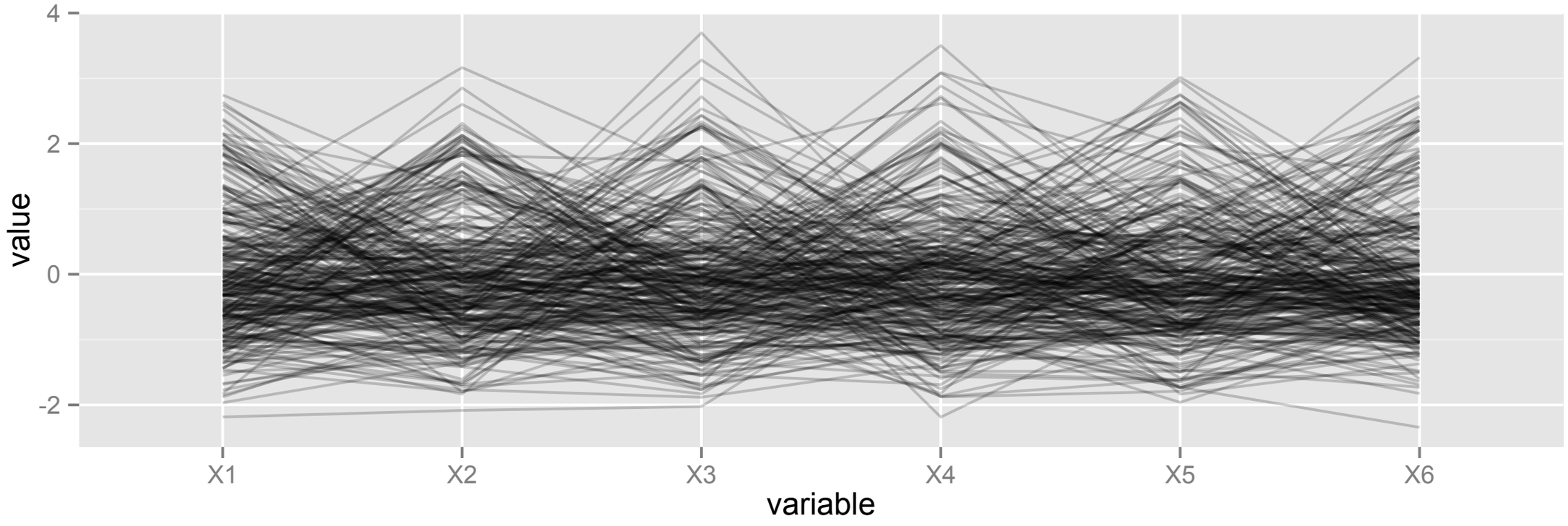
Clustered data



See the criss-crossing, gaps between lines.

Will need to extract the clusters before doing any other modeling, otherwise pretty regular data

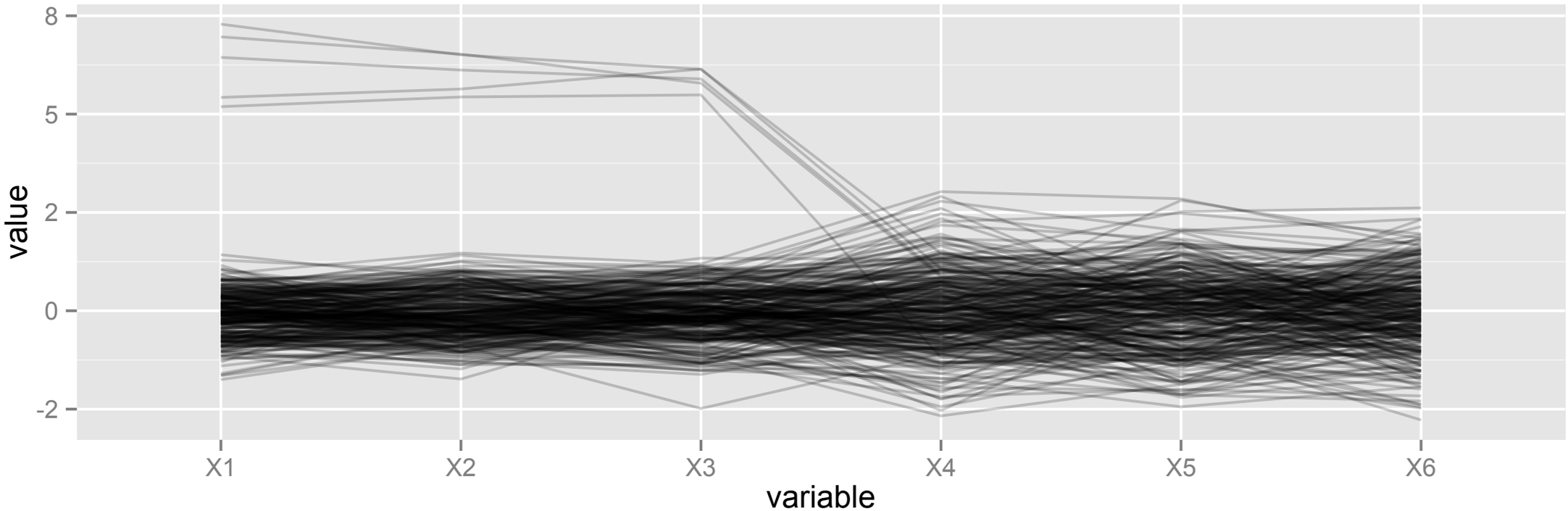
(Less) Clustered data



Still see the criss-crossing, gaps between lines, but less prominent.

Will need to deal with the multi-modality

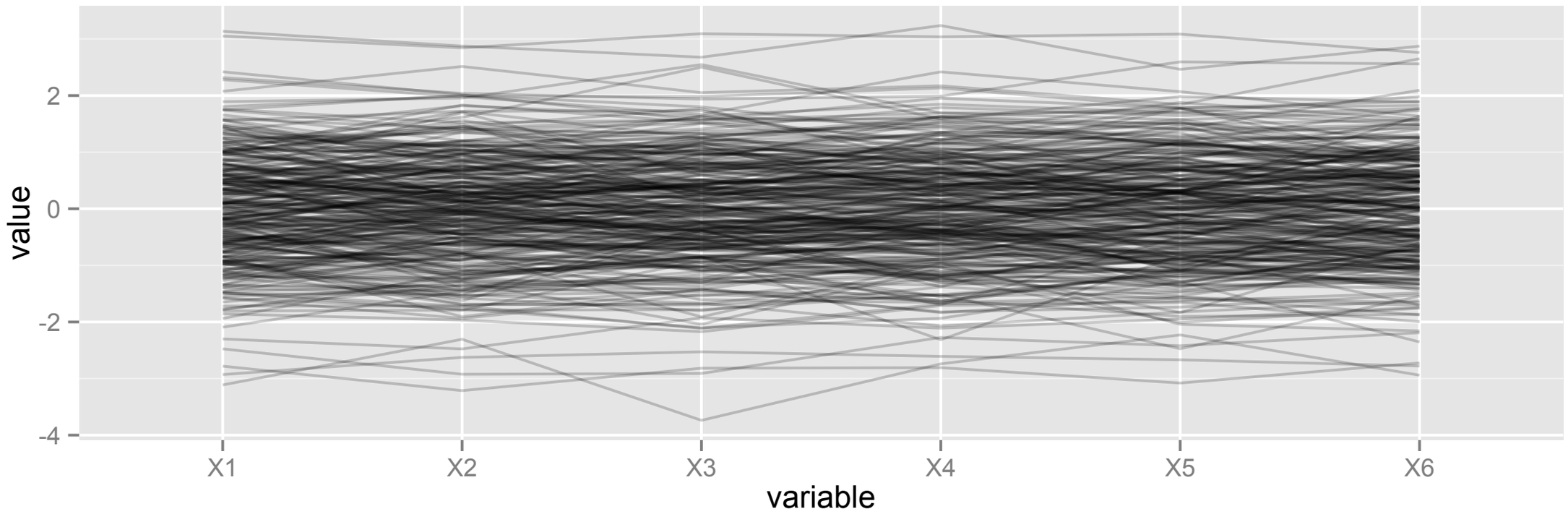
Outliers in the data



Small group of observations that are outliers on X1-X3.

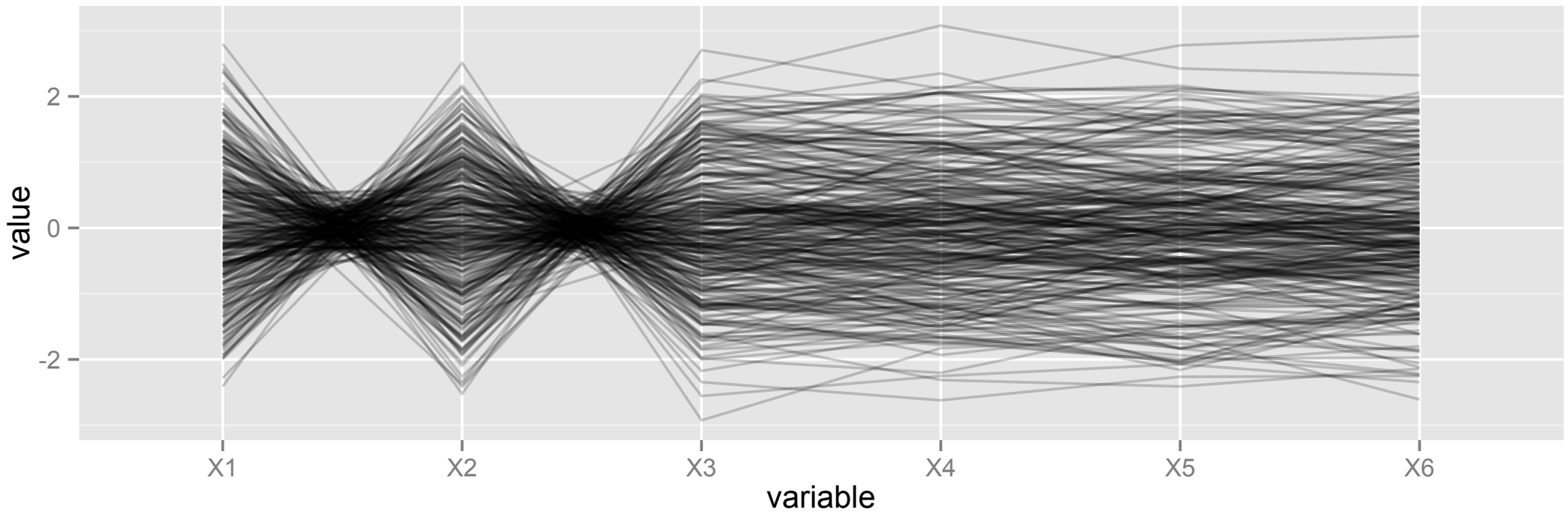
Need to do something with these cases, remove with justification, or fix

Strong association



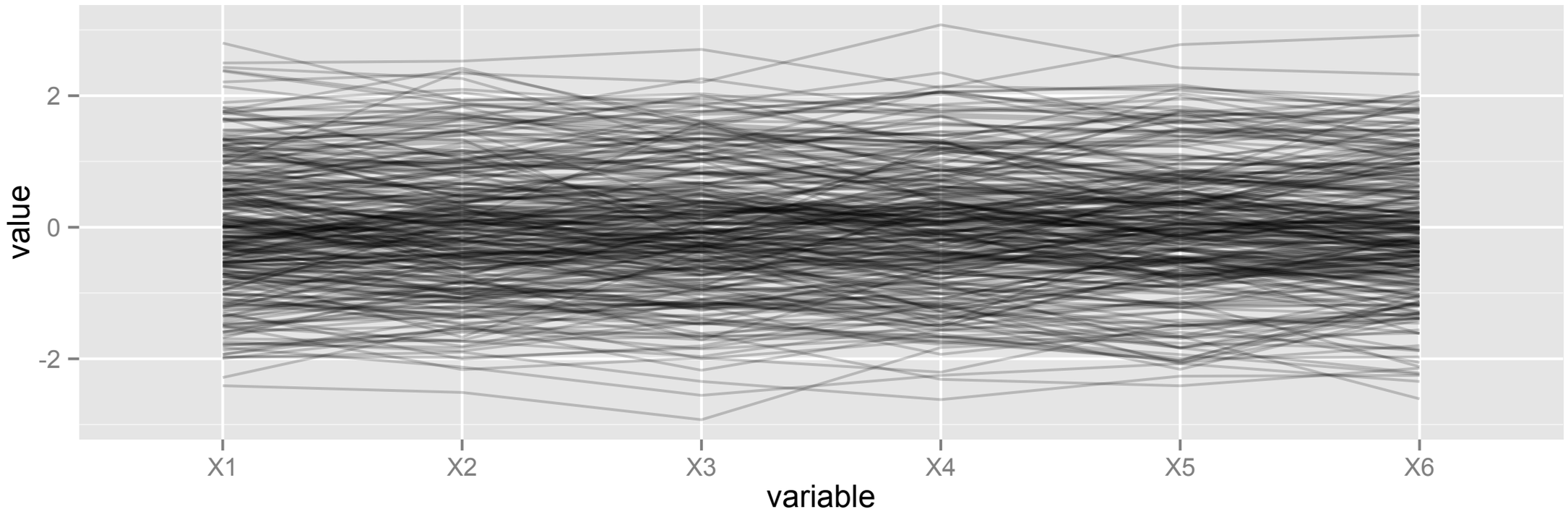
Very flat lines indicate strong positive association between all variables.

Strong negative association



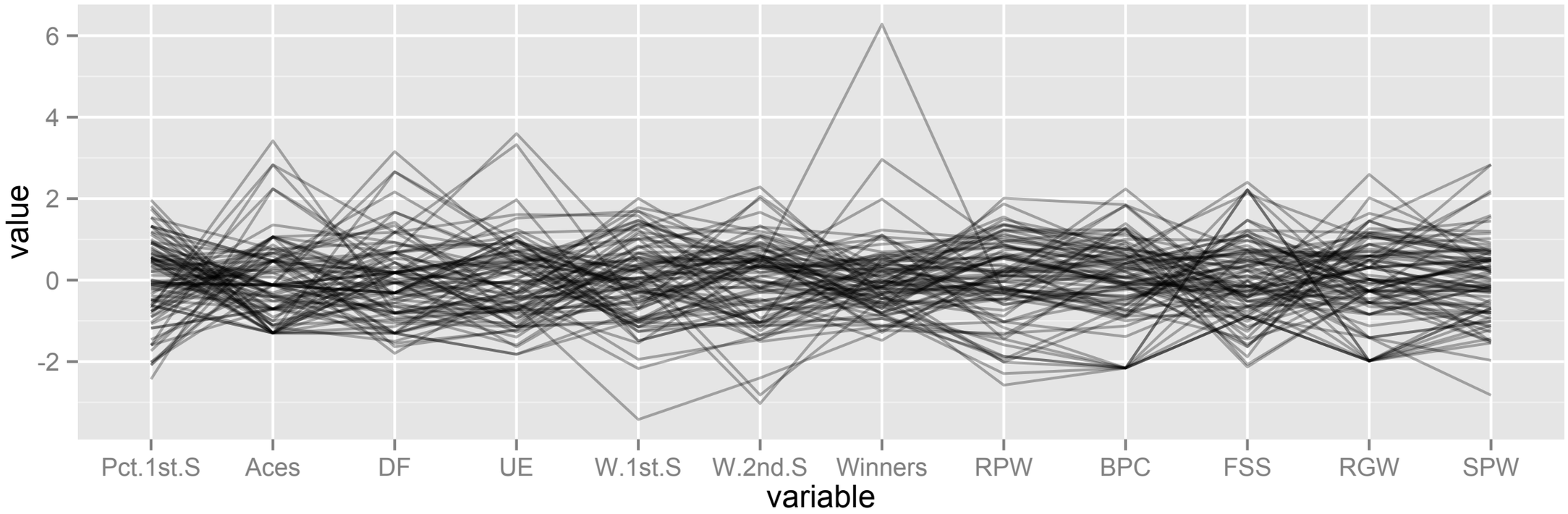
Crossed lines in first three variables indicate X2 is strongly negatively correlated with other vars.

Strong negative association - fixed



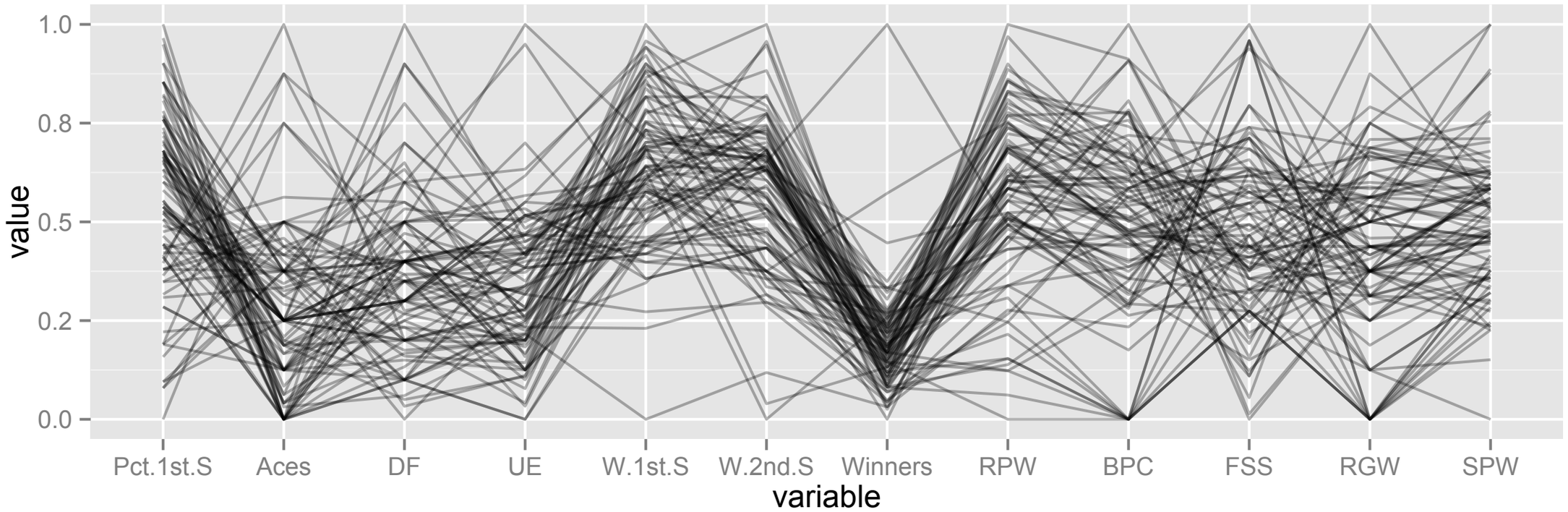
X2 is multiplied by -1, then it is positively associated with other variables.

Tennis statistics



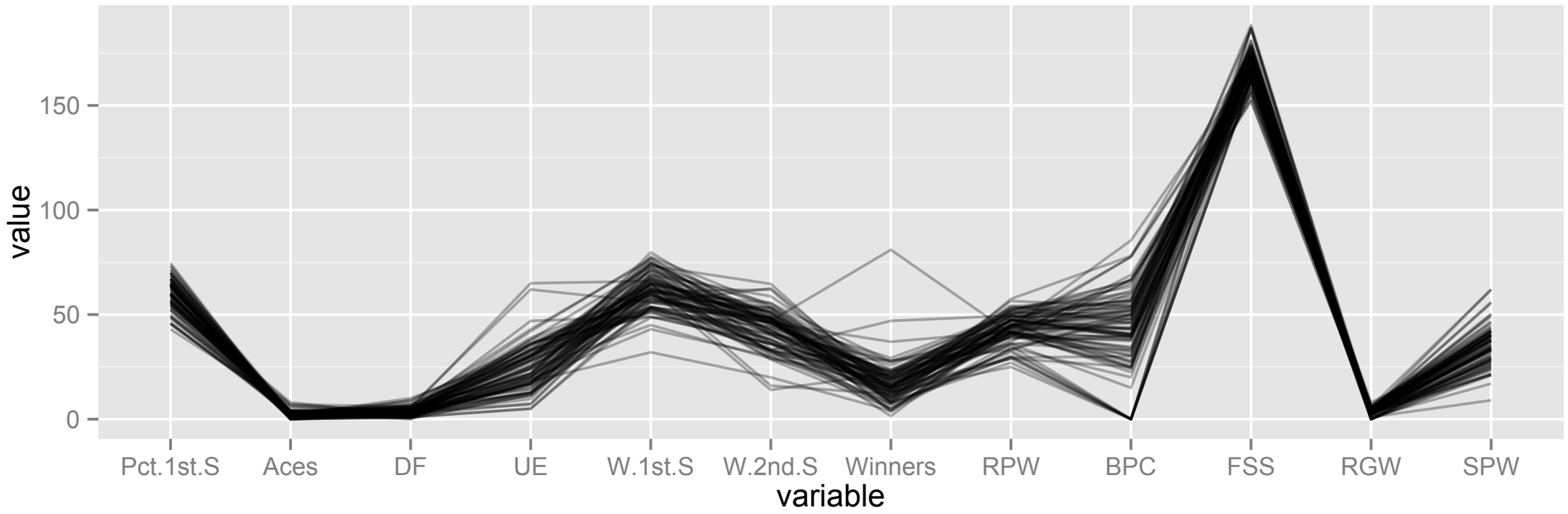
Scale matters: mean/sd, 0/1, individual/global

Enables correlation to be seen better, and outliers

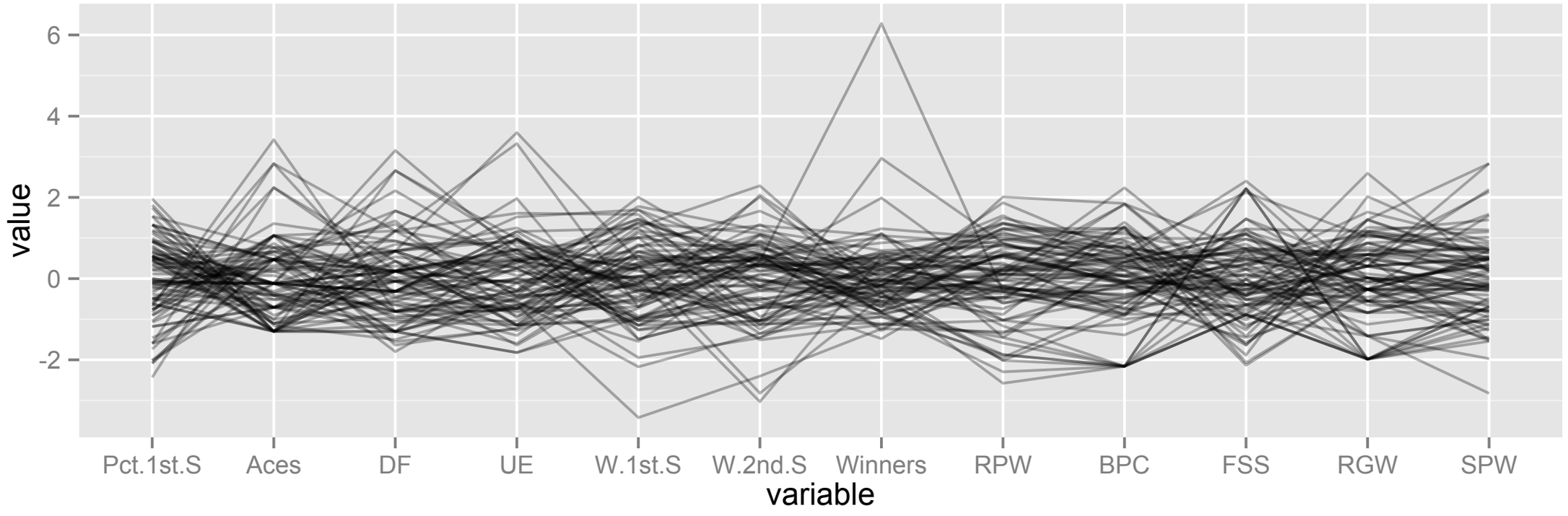


Scale matters: mean/sd, 0/1, individual/global

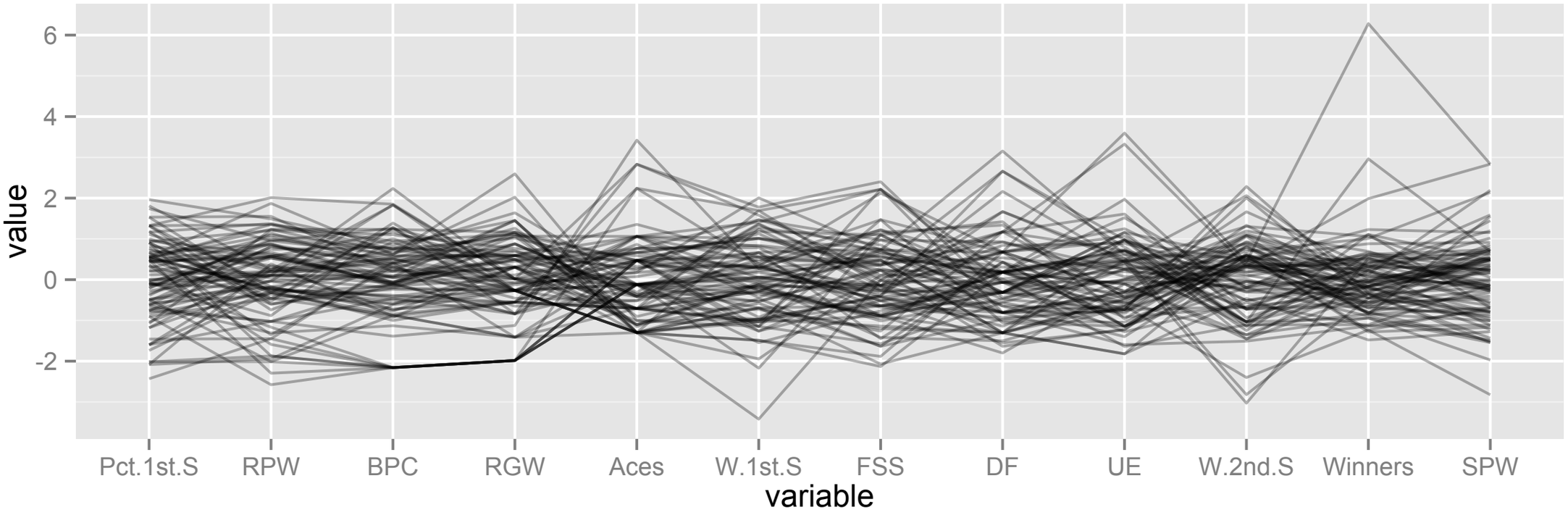
Emphasizes the univariate distributions



Scale matters: mean/sd, 0/1, individual/global



Order matters: place variables that are highly correlated close to each other

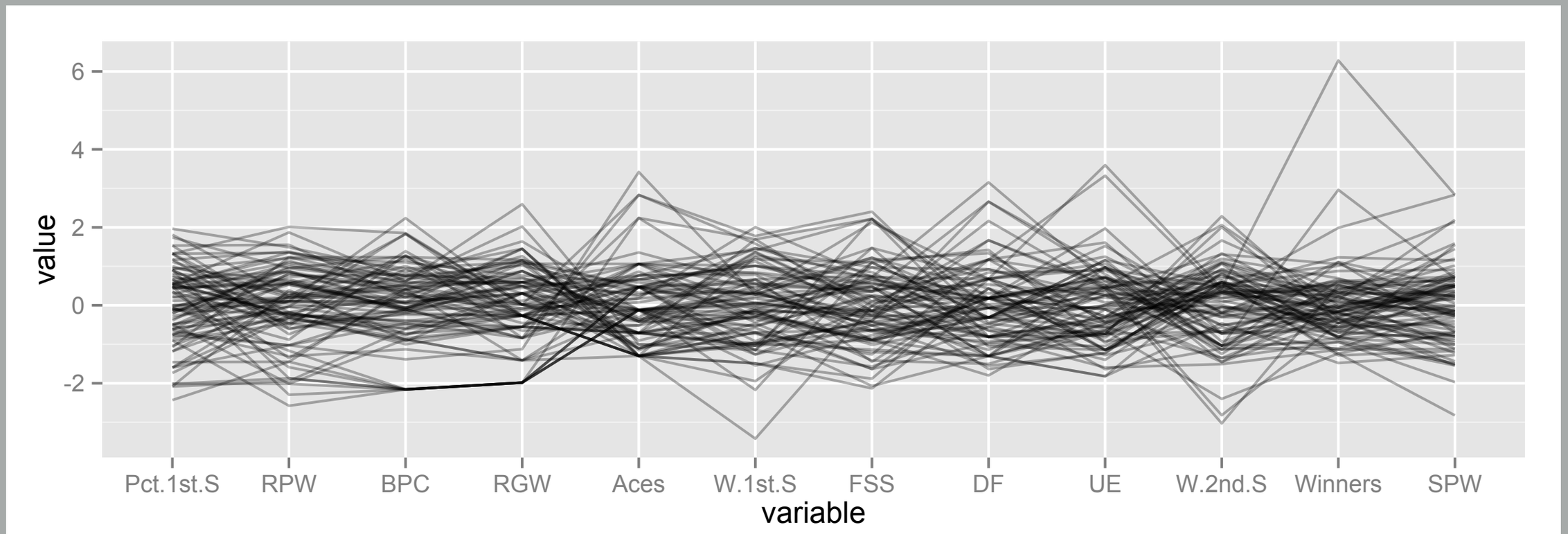


Order matters: place variables that are highly correlated close to each other

Less line crossing, easier to digest positive correlation, and then negative correlation

Your Turn

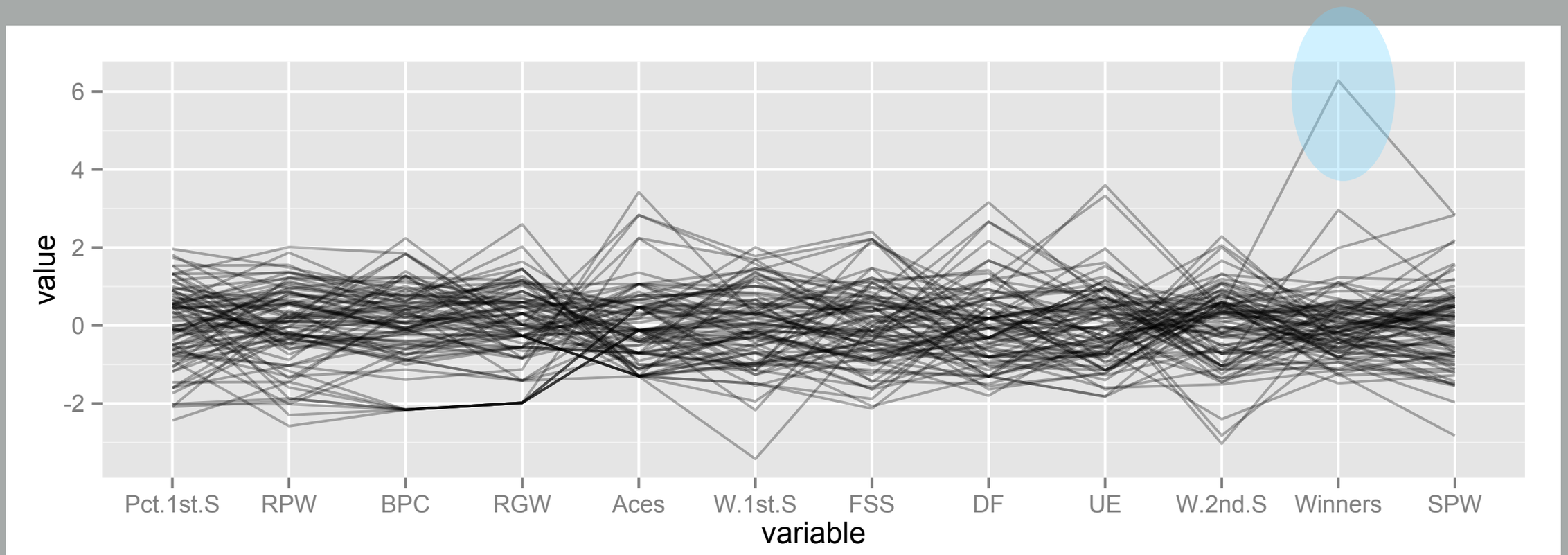
Take two minutes and discuss with your neighbor what you see



Clusters? Outliers? (In one variable or multiple variables?) Association?

Your Turn

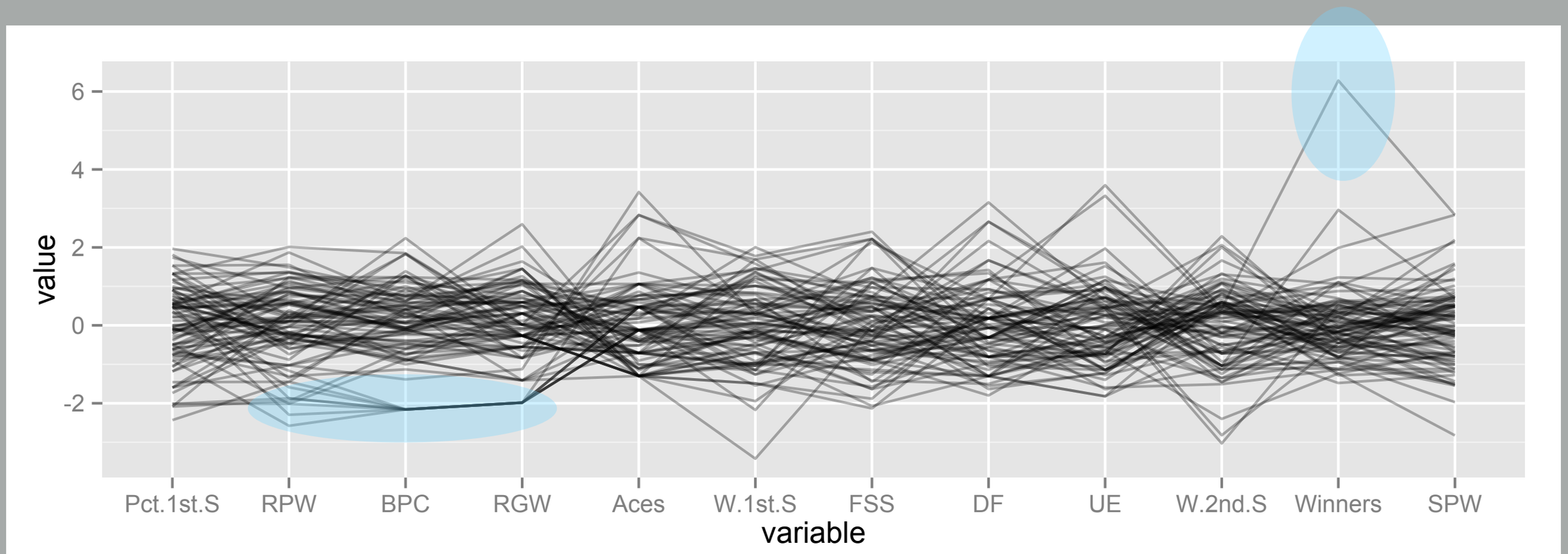
Take two minutes and discuss with your neighbor what you see



Clusters? Outliers? (In one variable or multiple variables?) Association?

Your Turn

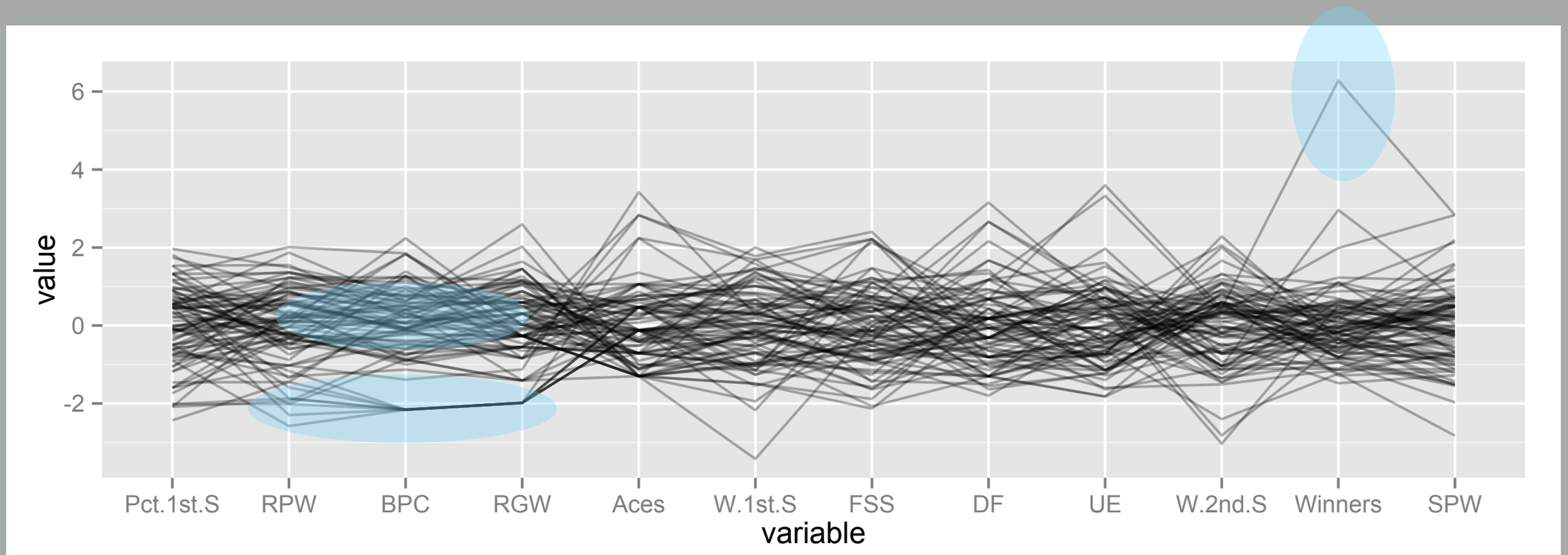
Take two minutes and discuss with your neighbor what you see



Clusters? Outliers? (In one variable or multiple variables?) Association?

Your Turn

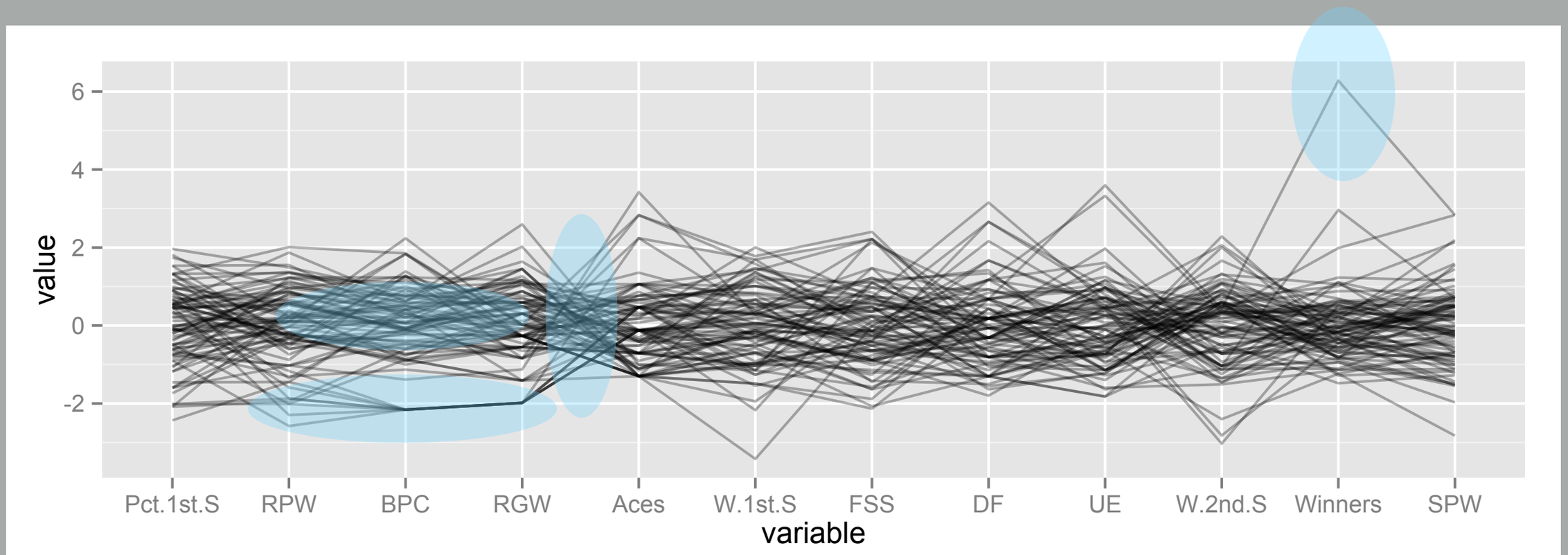
Take two minutes and discuss with your neighbor what you see



Clusters? Outliers? (In one variable or multiple variables?) Association?

Your Turn

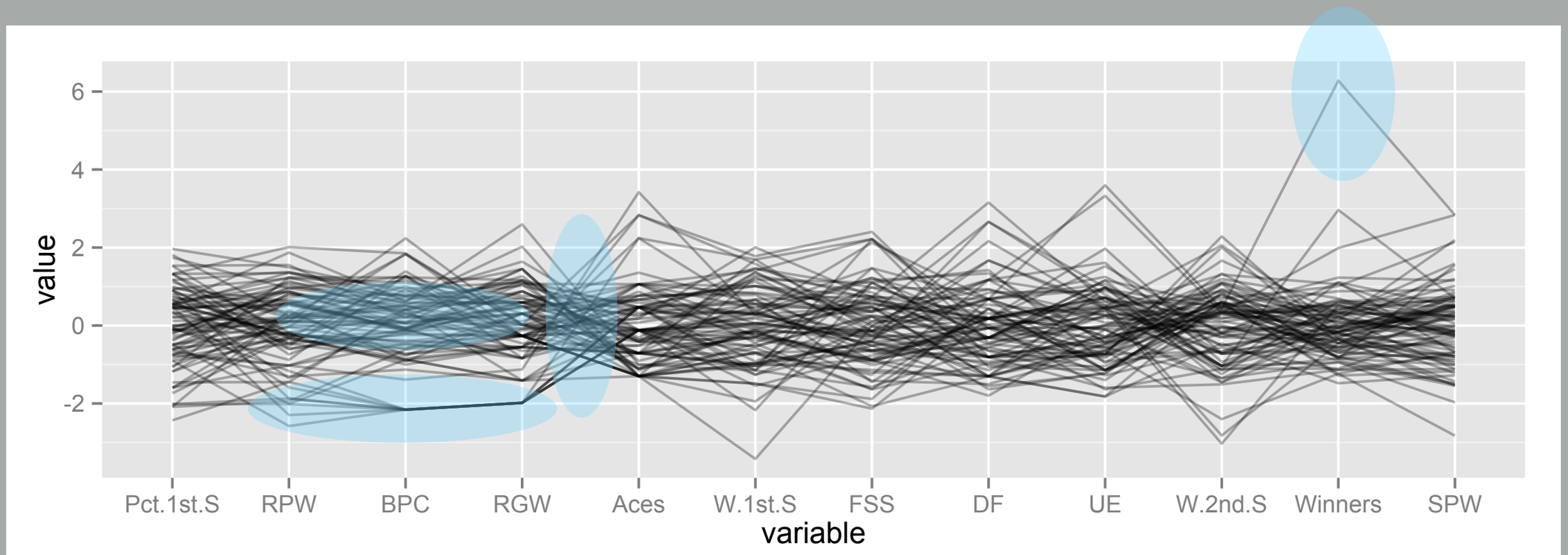
Take two minutes and discuss with your neighbor what you see



Clusters? Outliers? (In one variable or multiple variables?) Association?

Your Turn

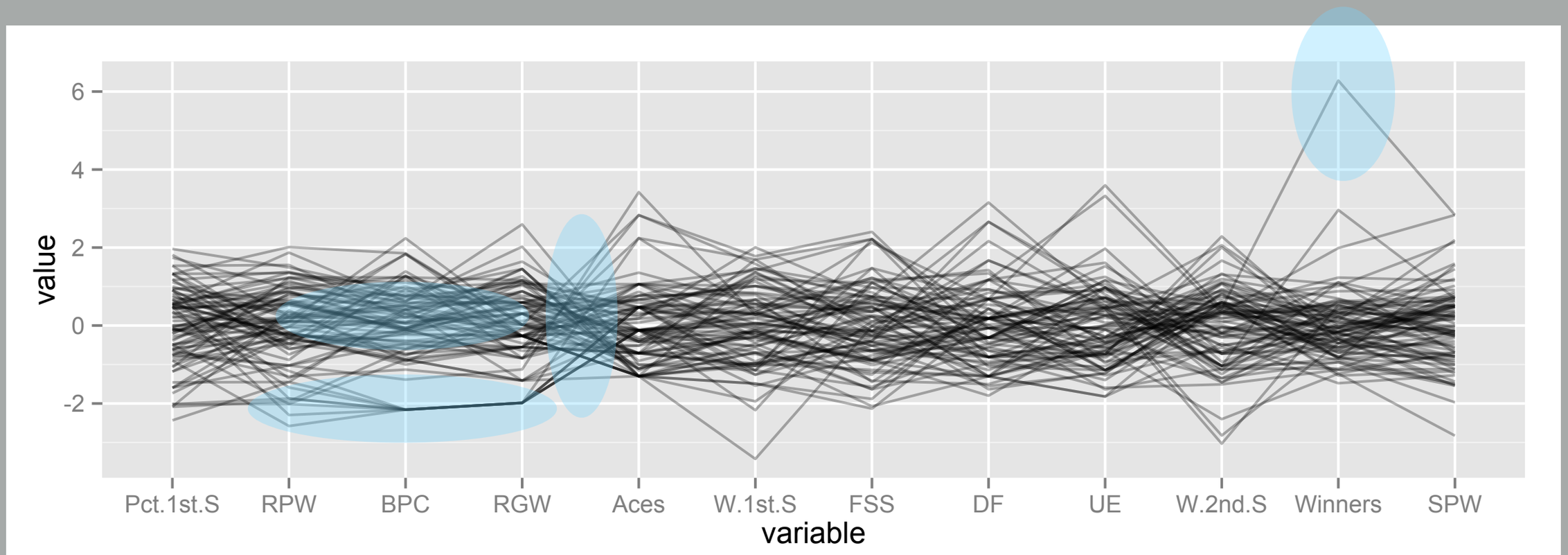
Take two minutes and discuss with your neighbor what you see



Clusters? Outliers? (In one variable or multiple variables?) Association?

Your Turn

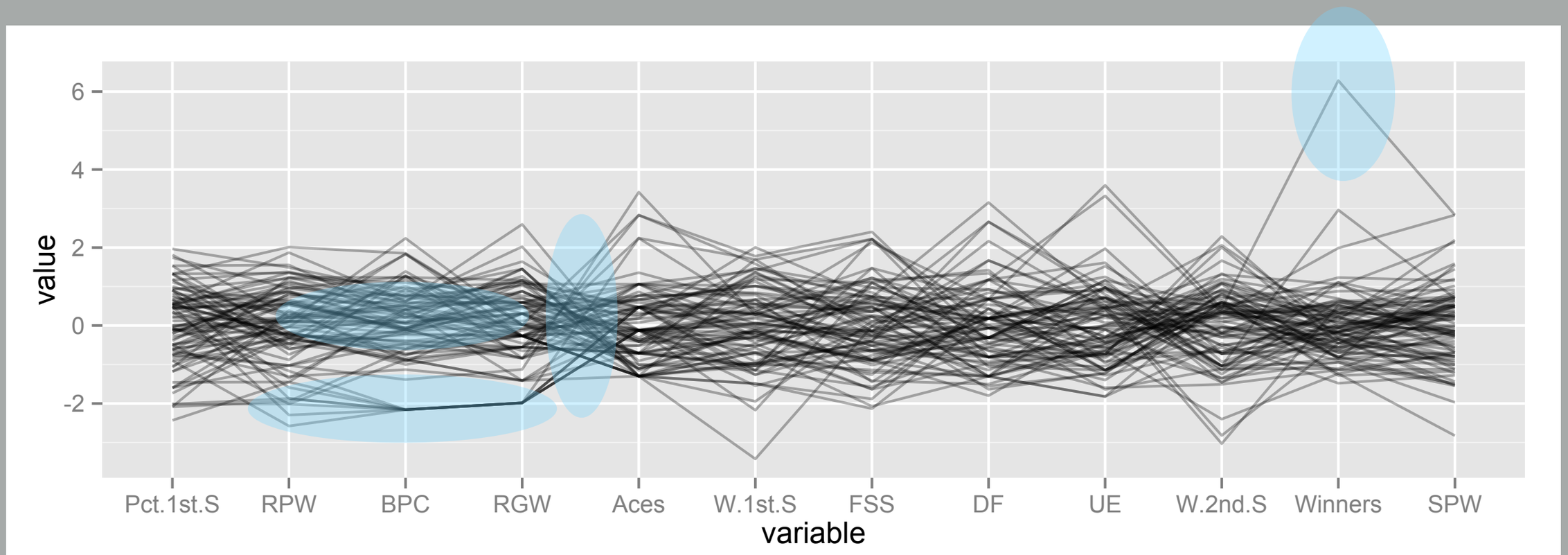
Take two minutes and discuss with your neighbor what you see



Clusters? Outliers? (In one variable or multiple variables?) Association?

Your Turn

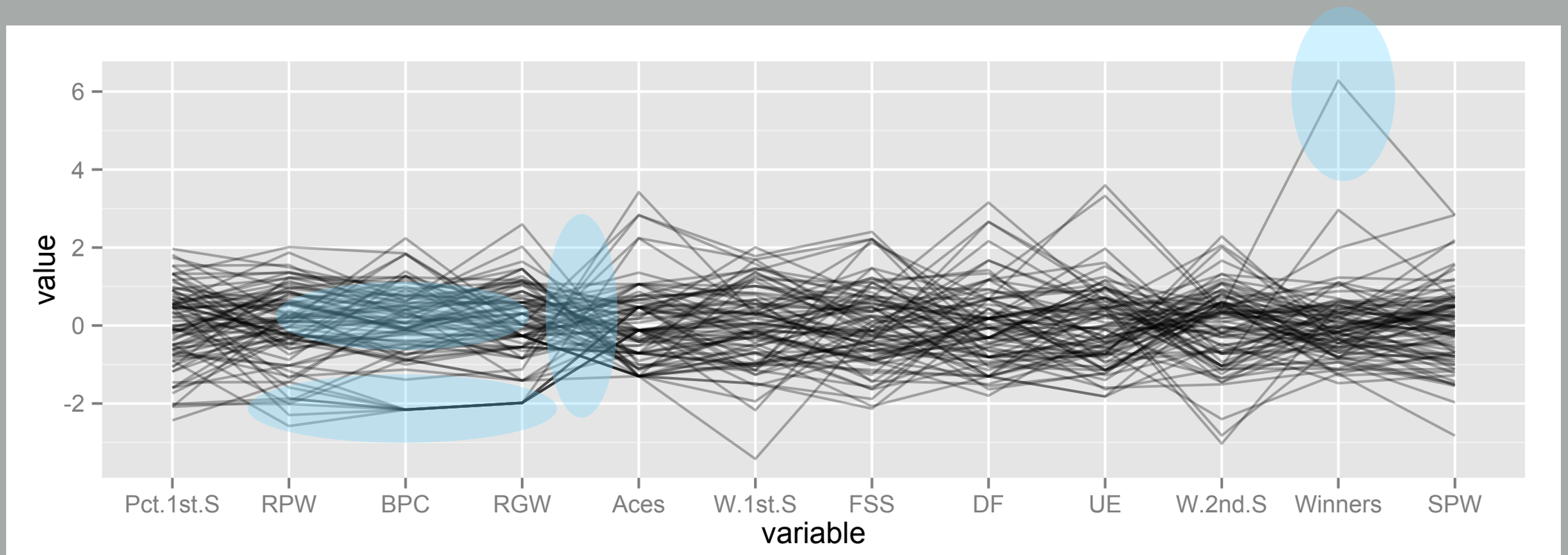
Take two minutes and discuss with your neighbor what you see



Clusters? Outliers? (In one variable or multiple variables?) Association?

Your Turn

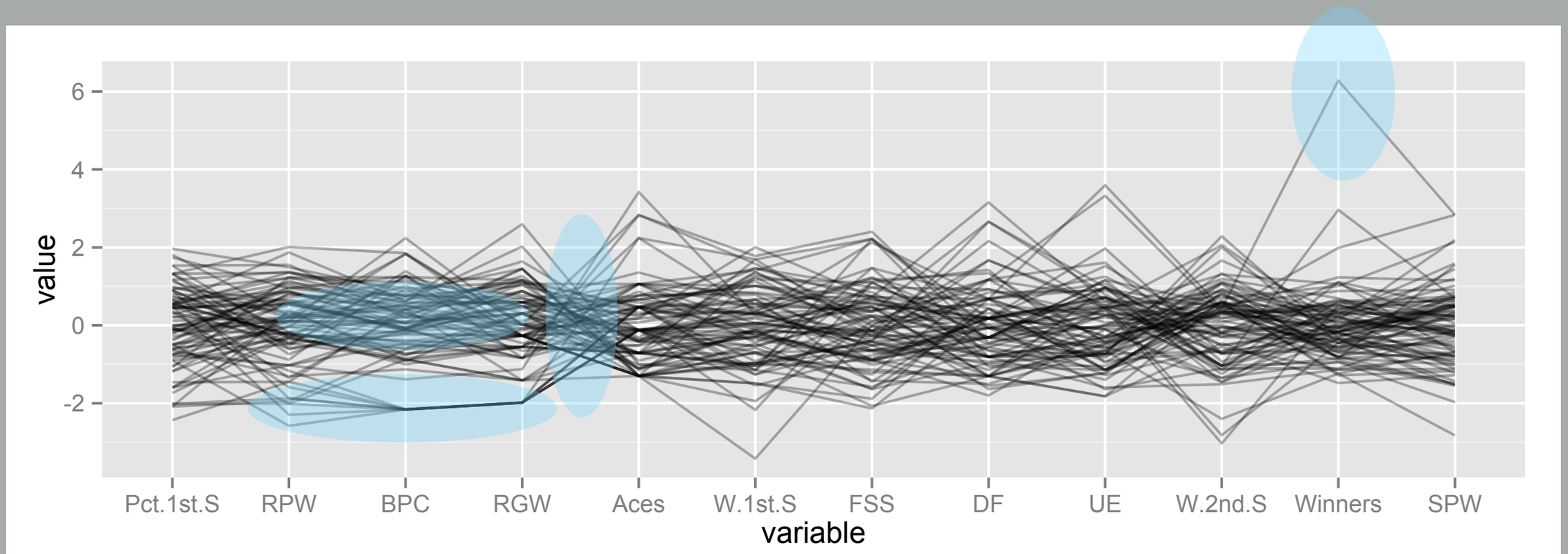
Take two minutes and discuss with your neighbor what you see



Clusters? Outliers? (In one variable or multiple variables?) Association?

Your Turn

Take two minutes and discuss with your neighbor what you see



✘ ✔ ✔ ✔

Clusters? Outliers? (In one variable or multiple variables?) Association?

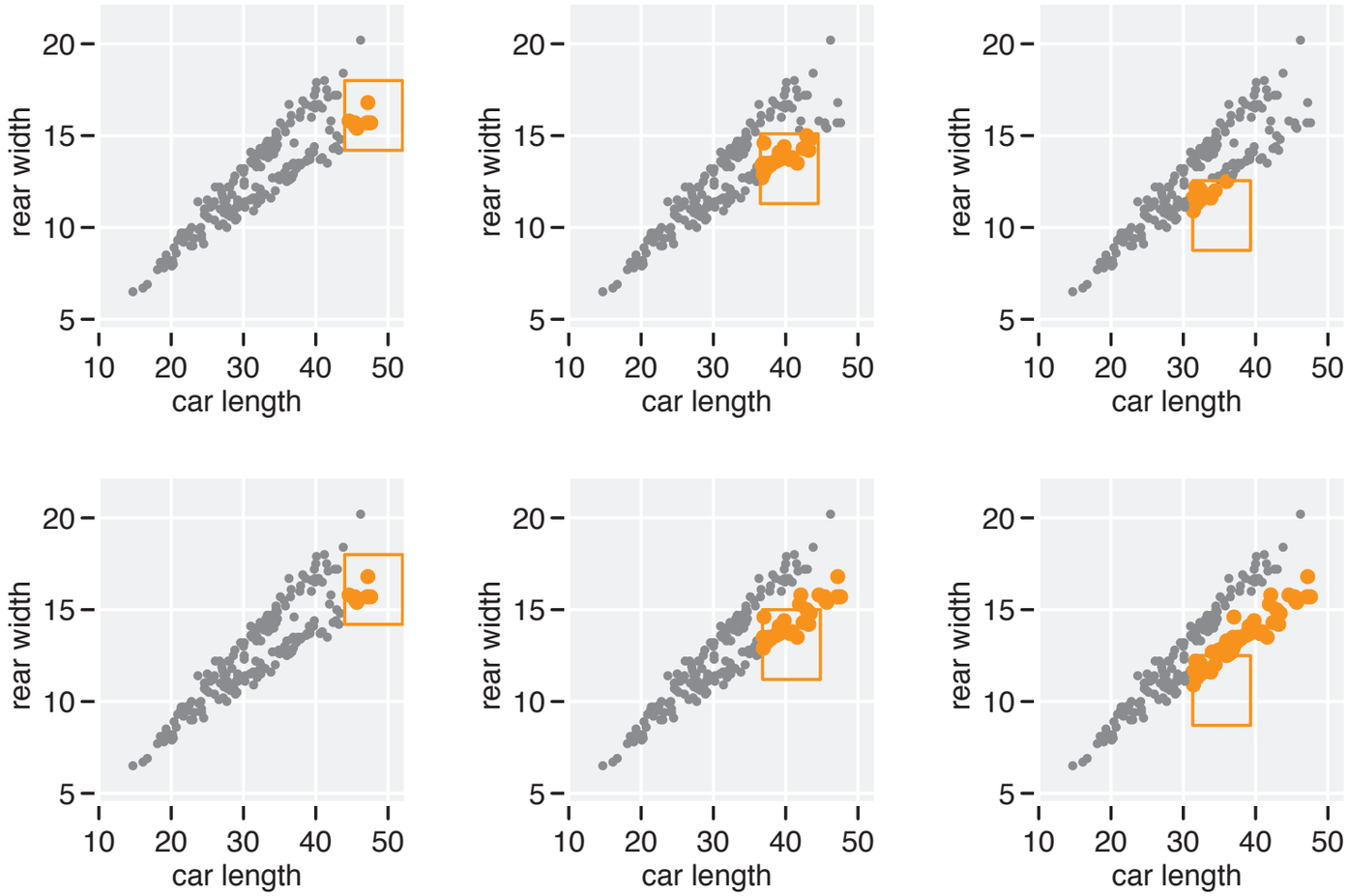
✔

Interactivity

- Brushing and linking between multiple plots
- Identifying
- Scaling

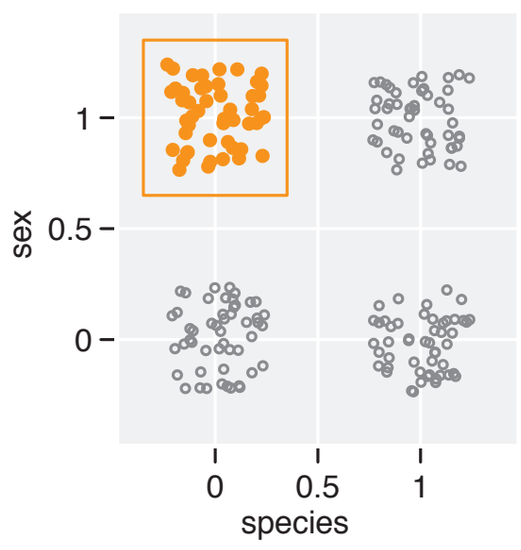
Brushing

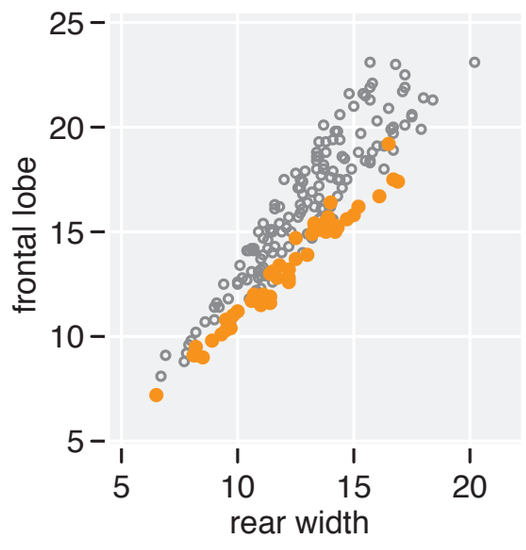
Persistent painting vs transient brushing



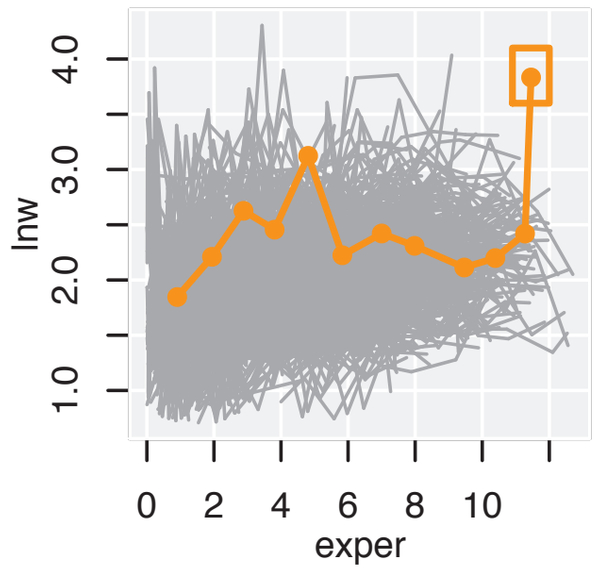
Linking

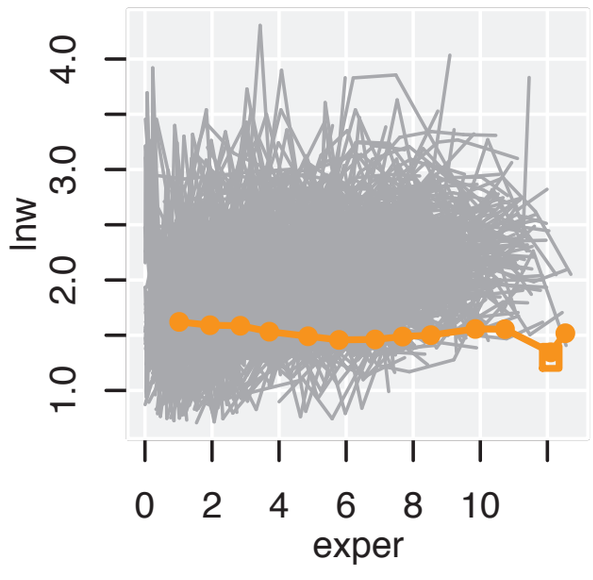
- One-to-one





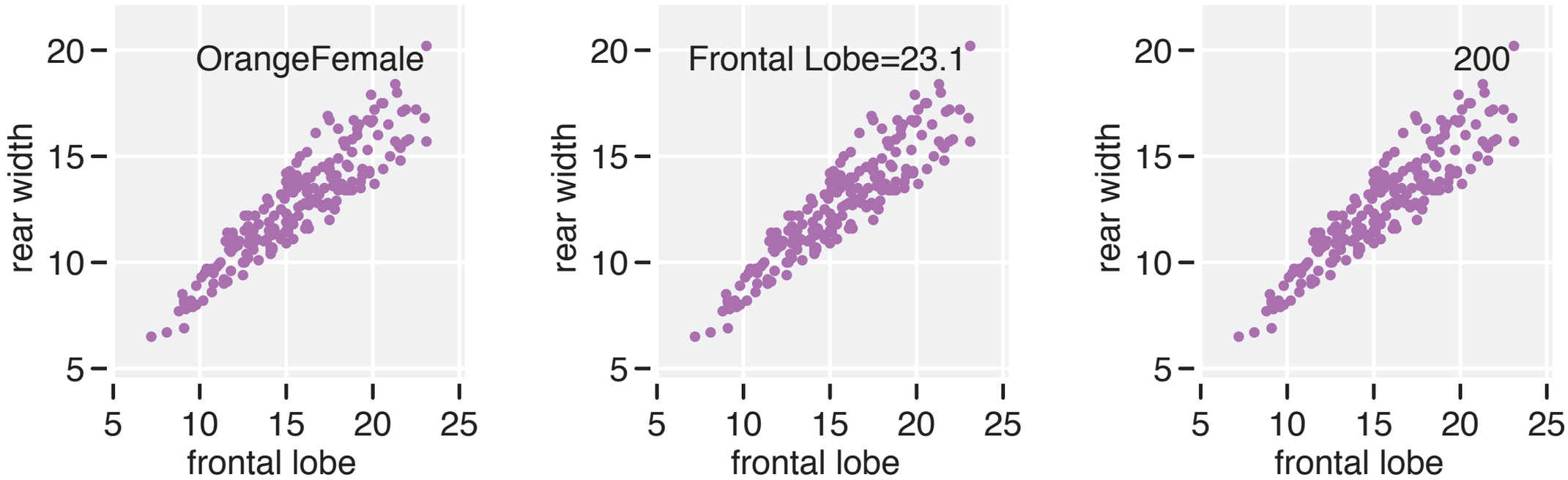
- Using a categorical variable



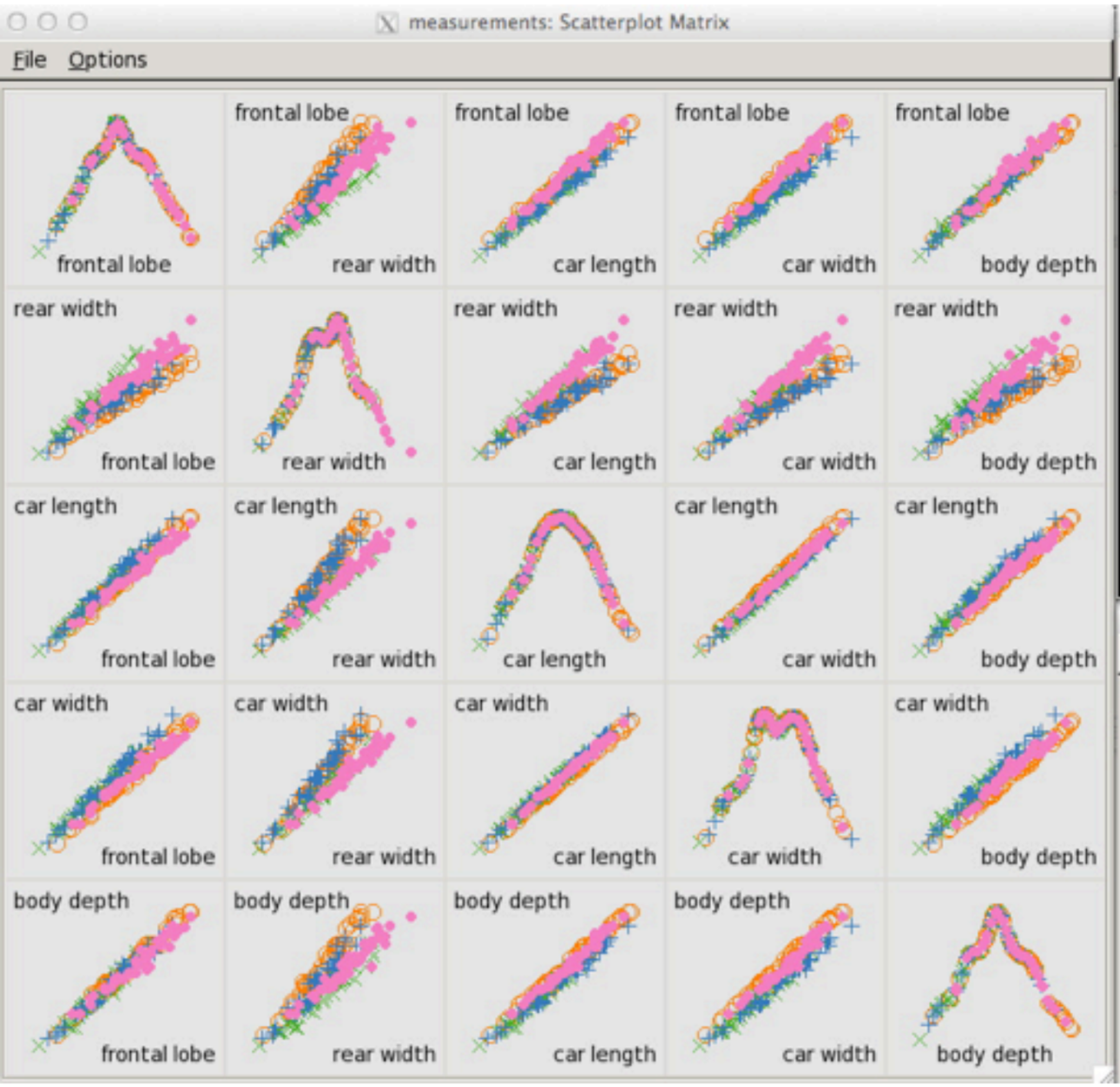


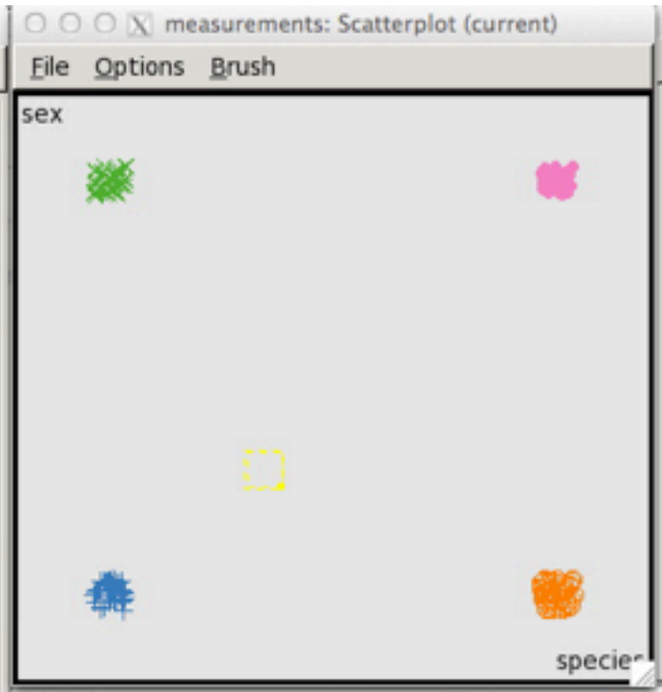
Identification

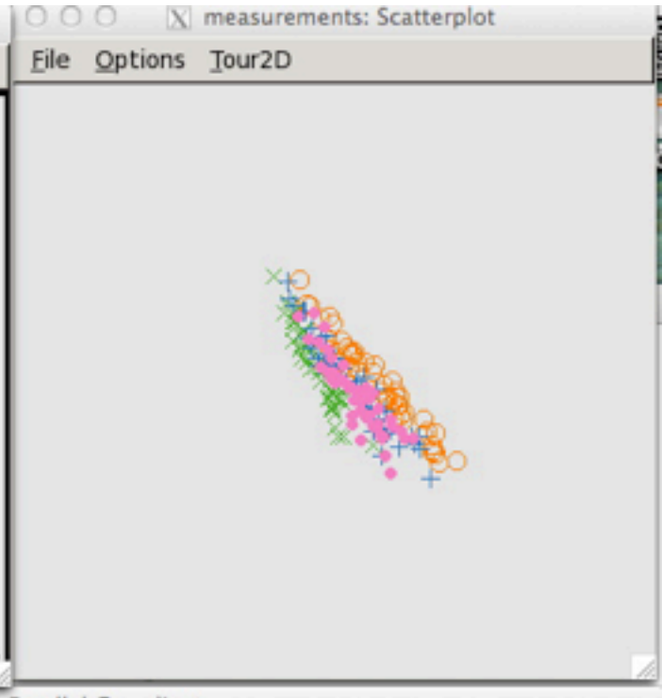
Lookup label information on data element

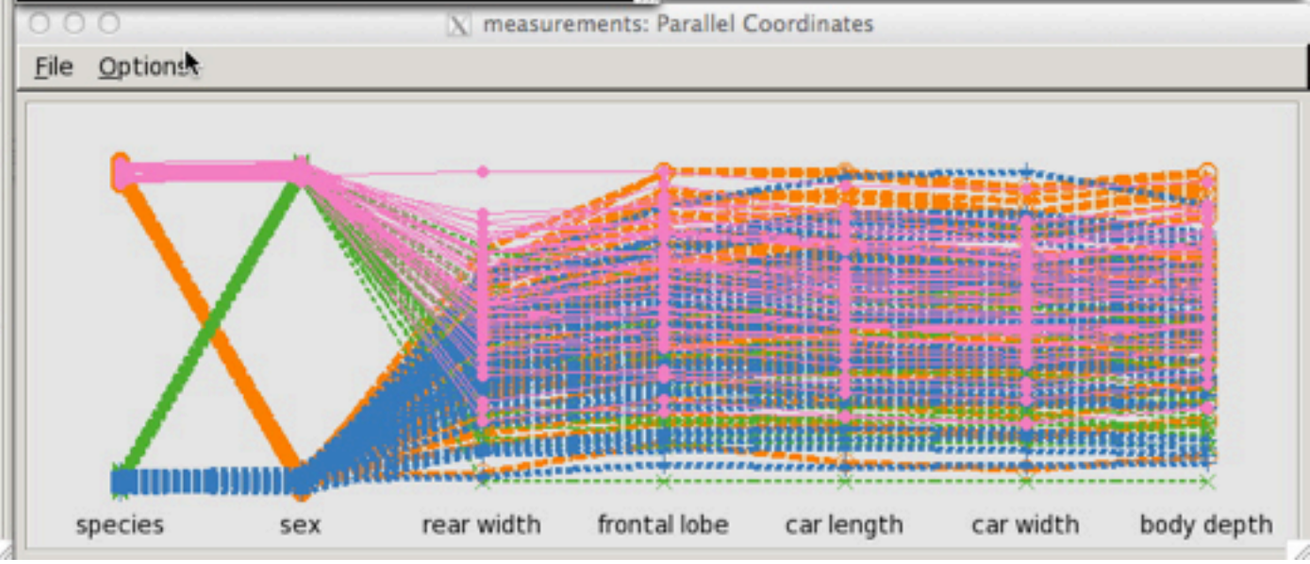


Putting it together

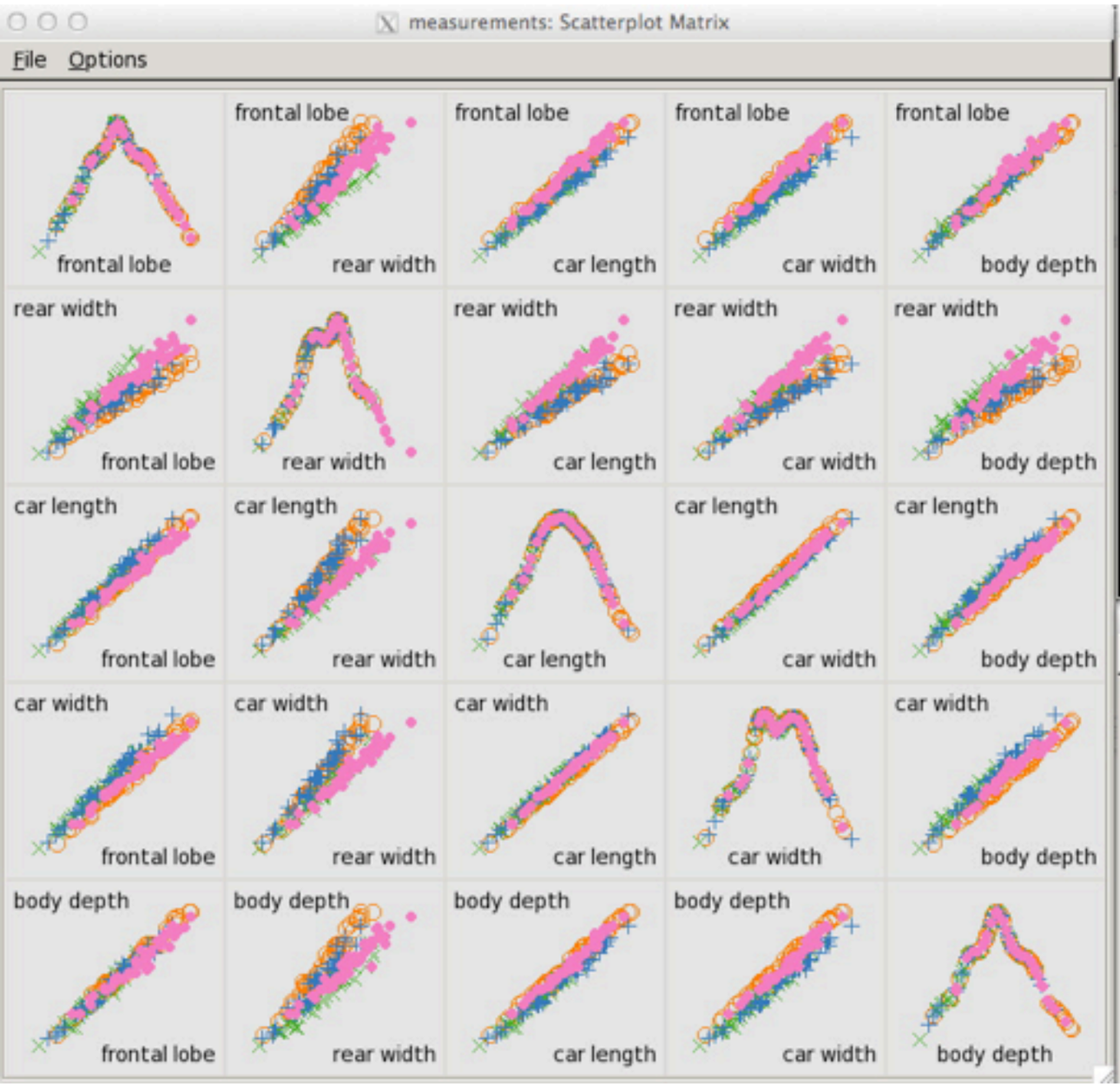


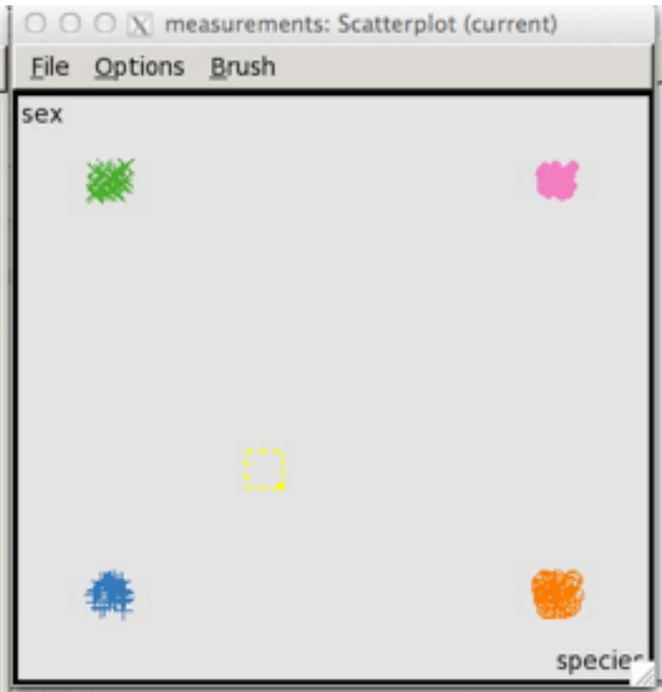


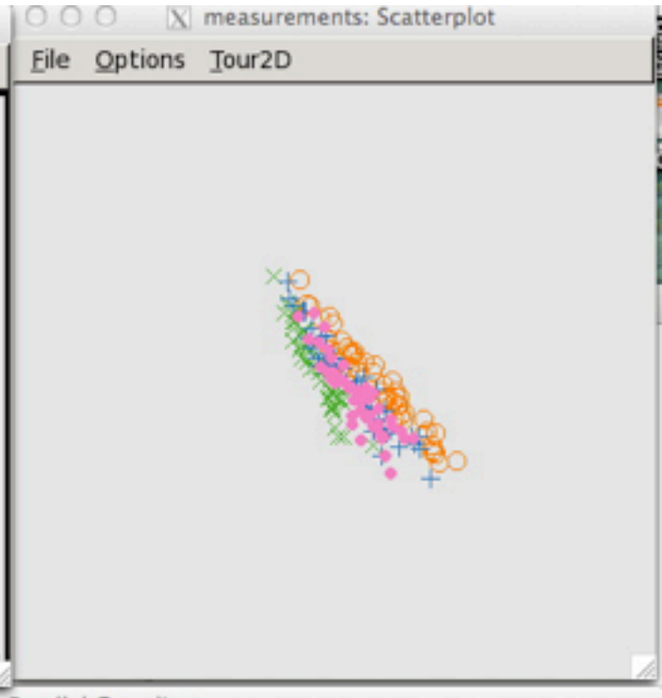


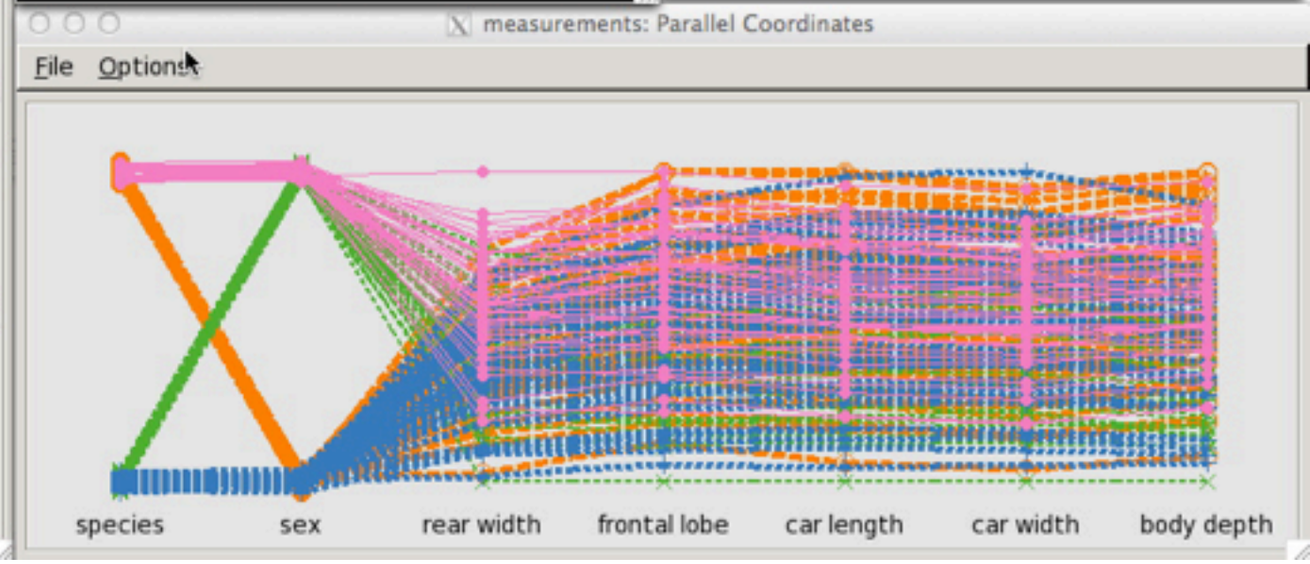


Putting it together









Exploring tennis statistics

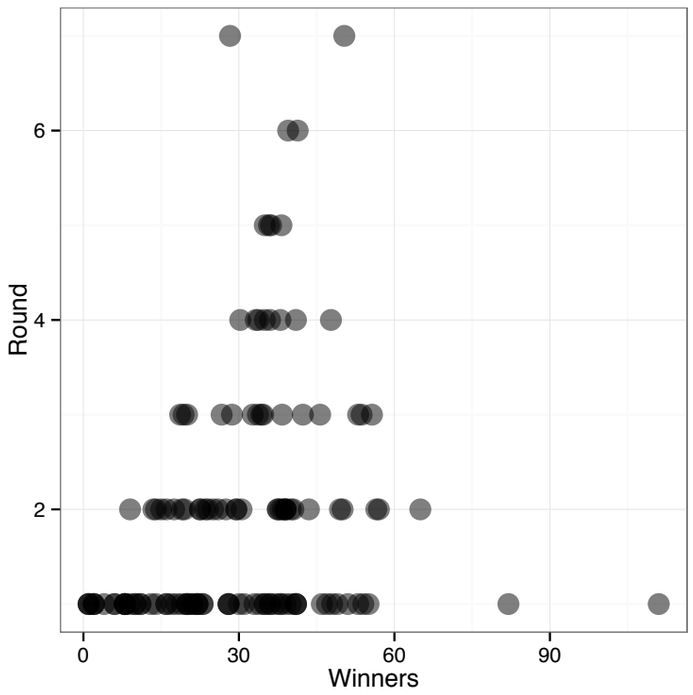
- 2014 was a great year for Swiss tennis
- Stan Wawrinka surprised everyone and defeated Nadal in the final, after defeating Berdych and Djokovic to get there
- Switzerland won its first ever Davis Cup

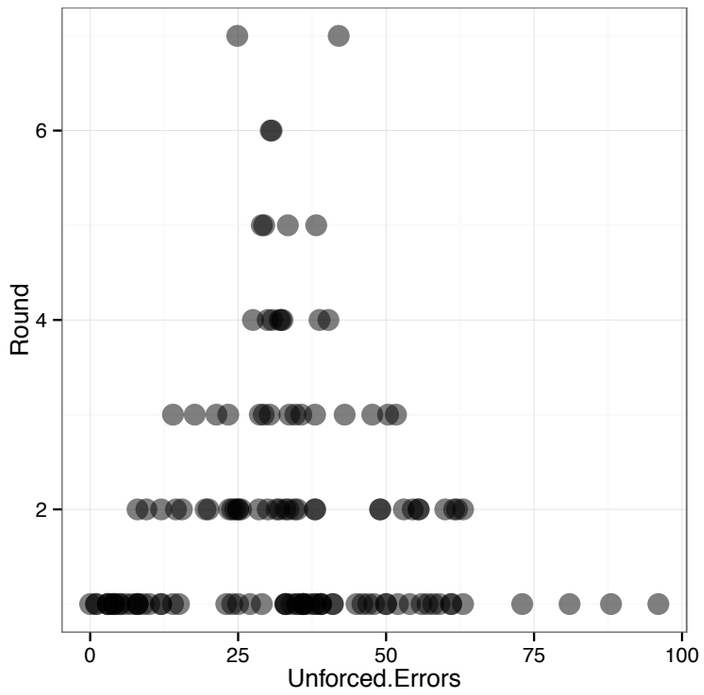
Statistics for 2014 pulled from http://www.ausopen.com/en_AU/players/overview/atpw367.html

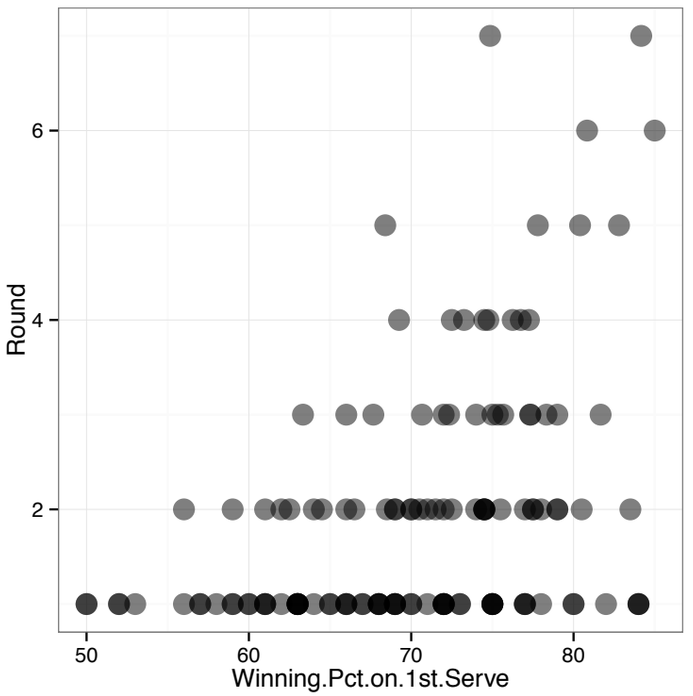
Questions

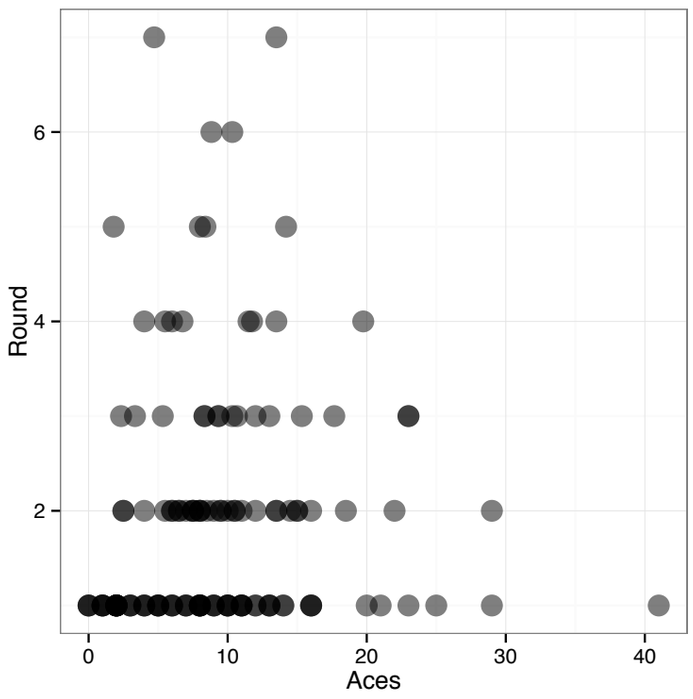
- What performance statistics suggest advancing in the tournament?
- How did Stan Wawrinka win?

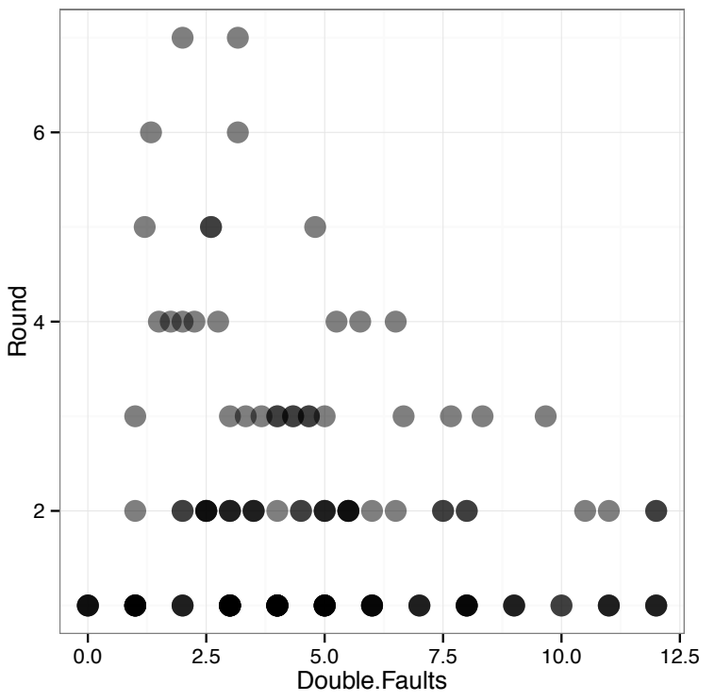
Performance

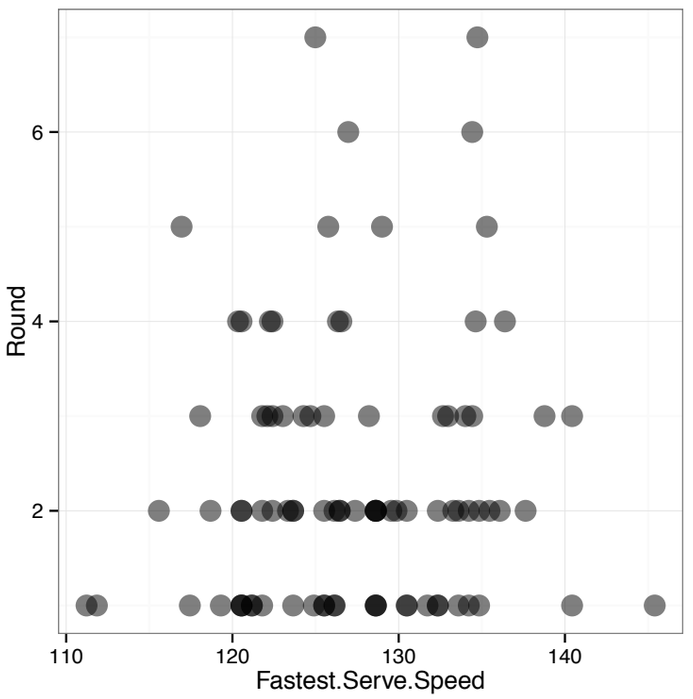




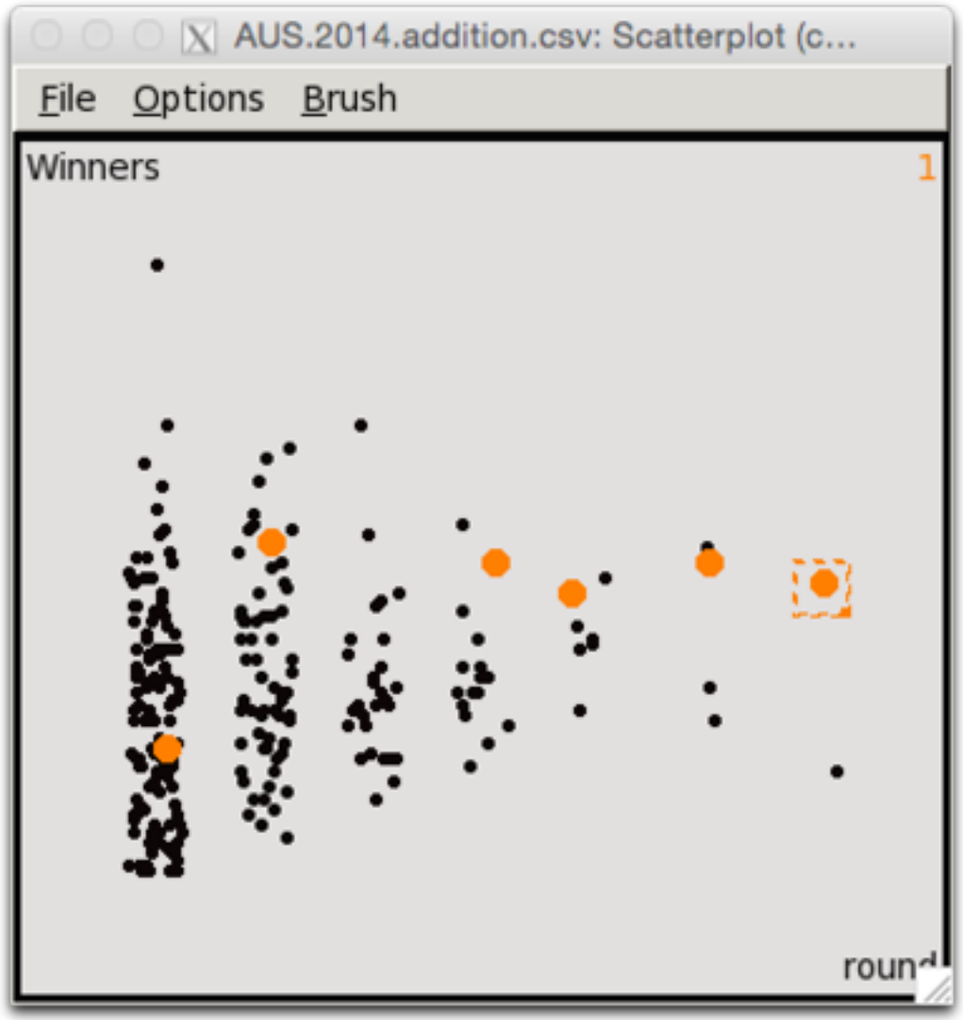


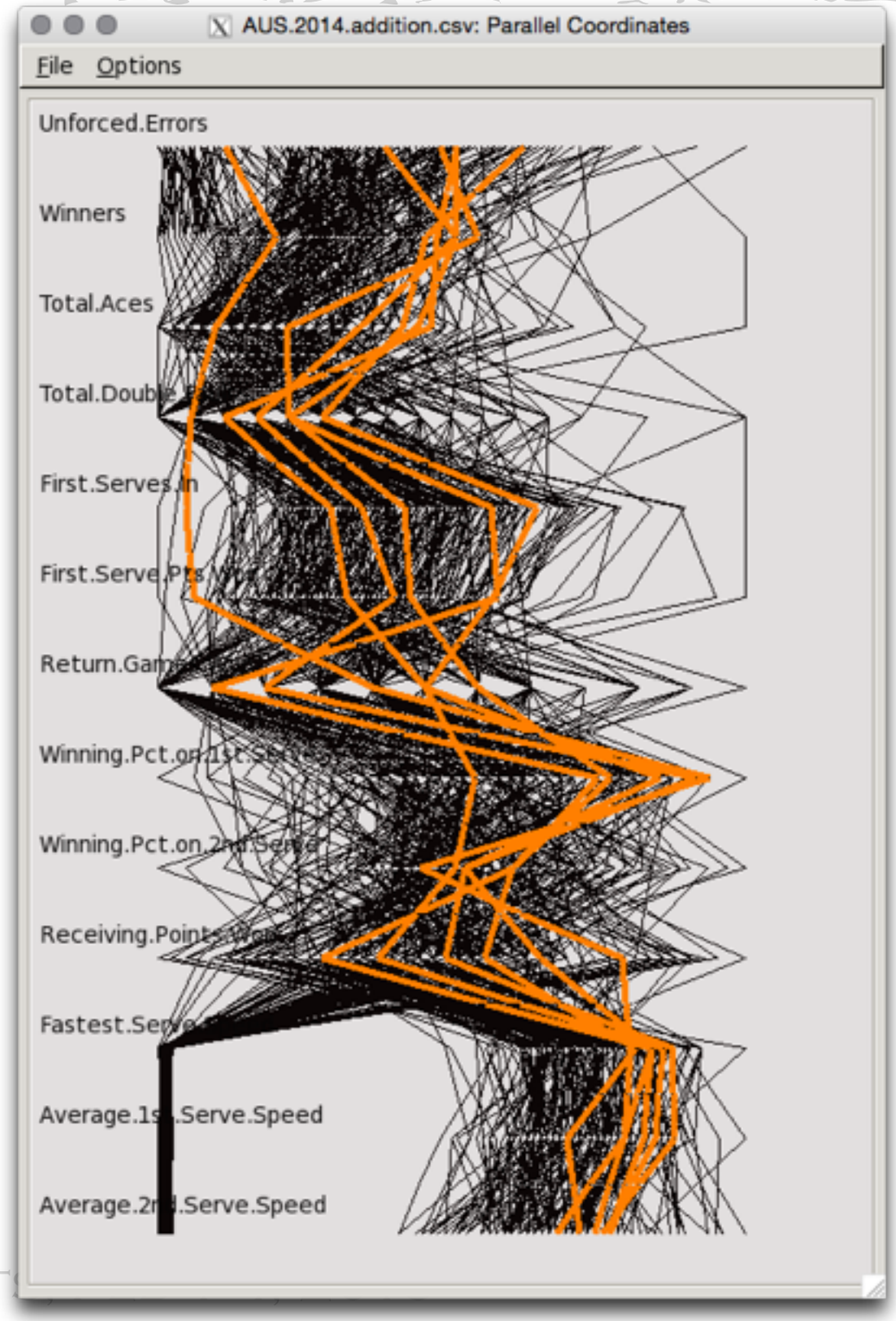




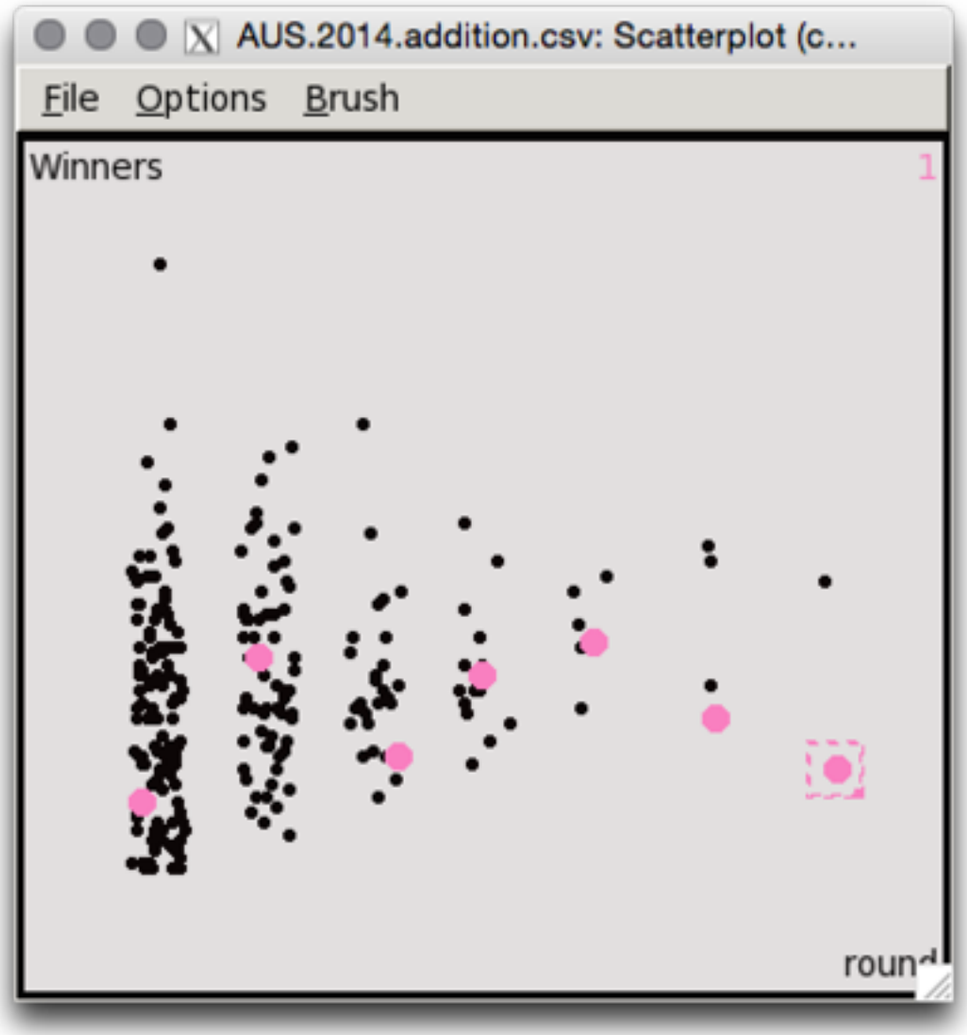


Wawrinka vs Nadal





Wawrinka vs Nadal



What we learn

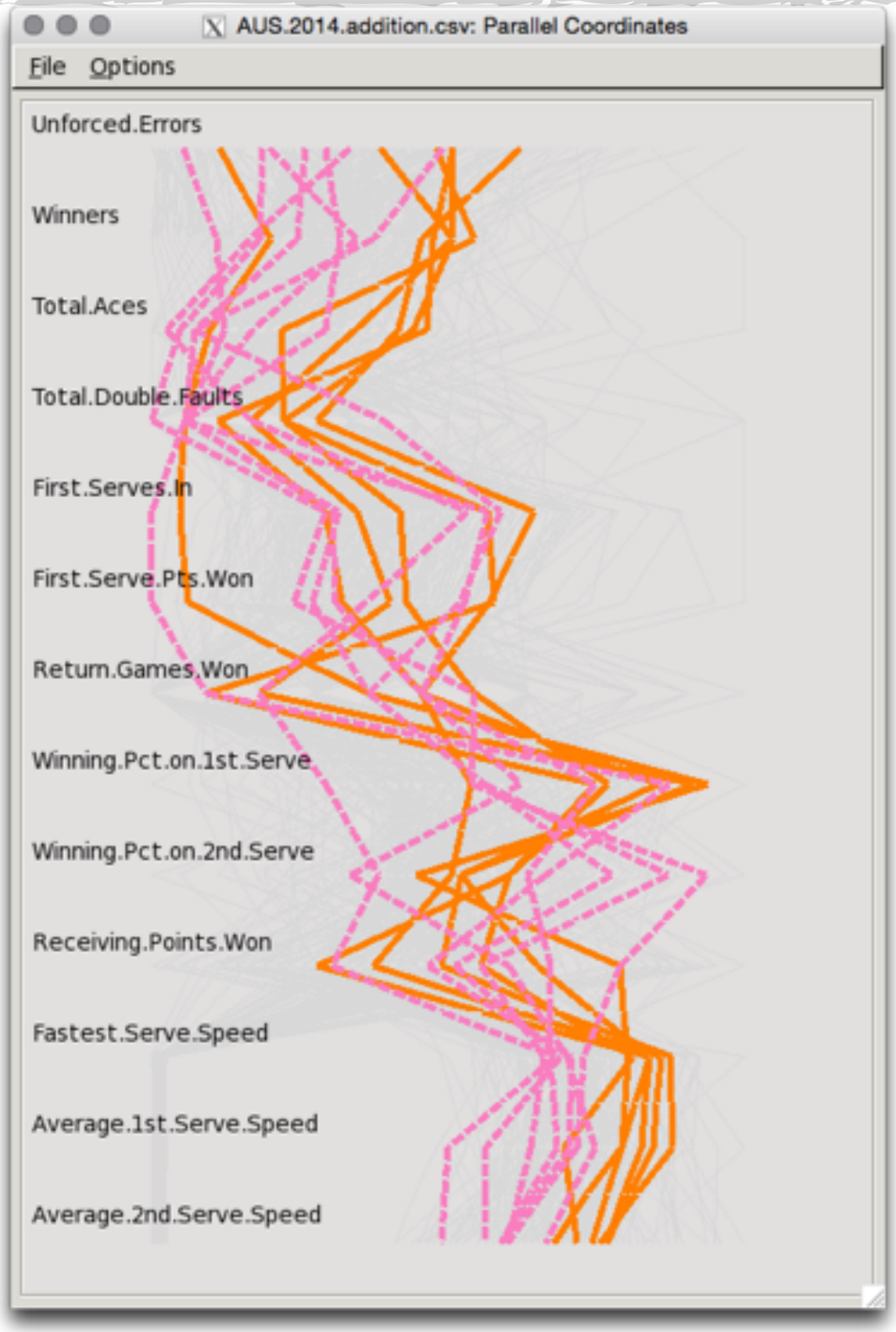
- It is important to keep control in the game, not too many winners, not too few, and errors same.
- Serve speed is not that important.
- Winning your serve is important.
- Stan beat Rafa on winners in final!
- Nadal serves slowly, and not so much slower in final.

Stan vs Rafa

Stan won the game on winners, unforced errors, aces, serving % and speed.

It looks like he played aggressively, he was in the zone, and the gamble paid off.

Stan vs Rafa



Generally, throughout tournament, Stan had more winners, errors, first serve %, serve speed.

He had a better tournament performance.

Your Turn

Take two minutes to come up with some more question

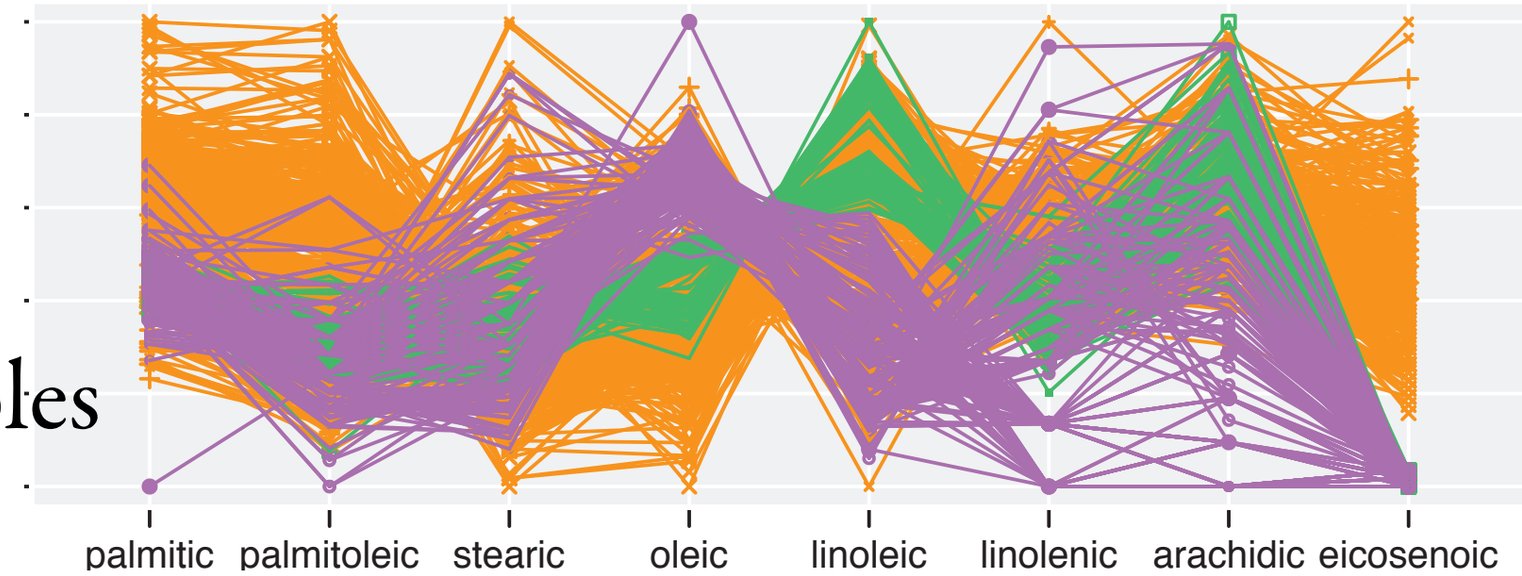
- What other things would you like to investigate in the game of tennis?
- What calculations, tables, plots would you make to tackle these questions?

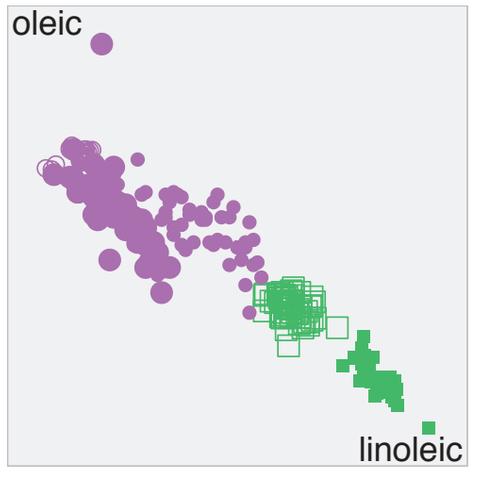
Exploring model fits

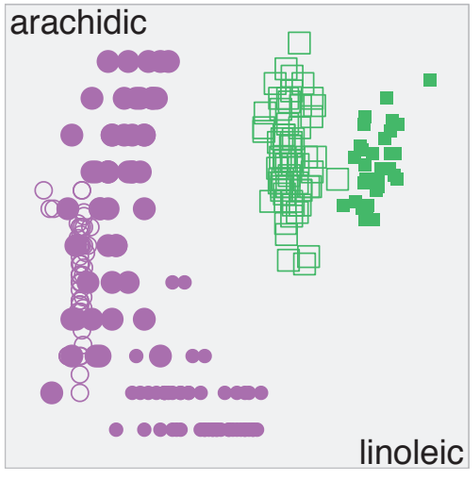
- Classification, examining boundary rules and misfits
- Clustering, exploring self-organizing maps

Italian Olive oils

- Ancient statistics textbook data
- 572 samples
- 8 fatty acid composition variables
- 9 classes
- Related to food quality and pricing
- Fatty acid signature associated with growing region

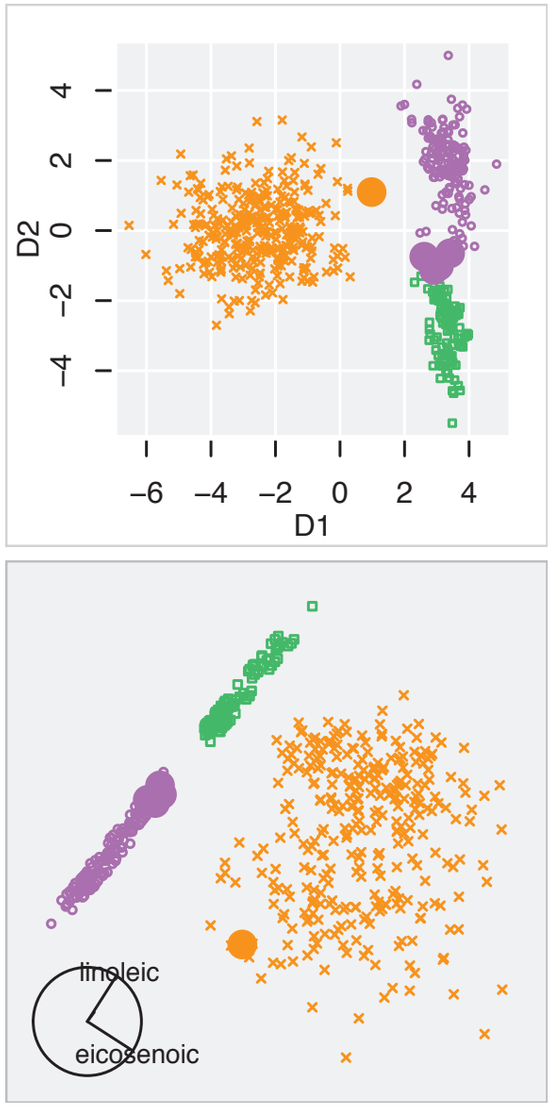


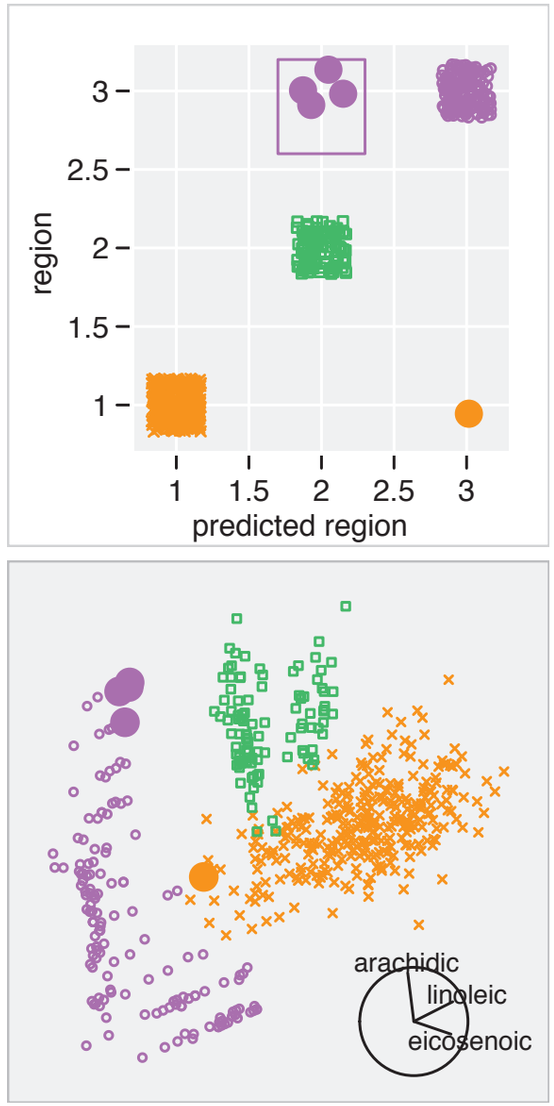


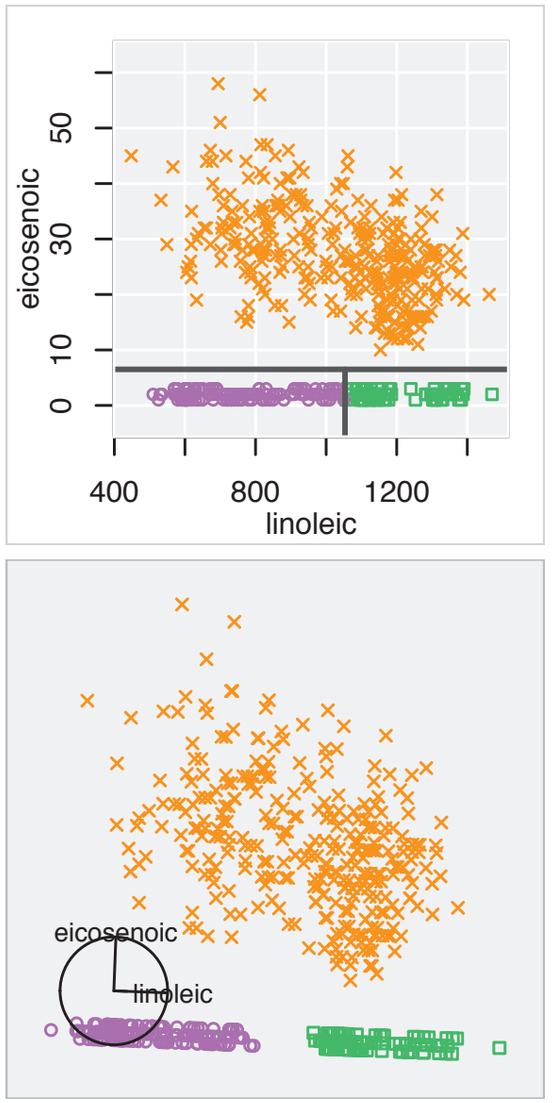


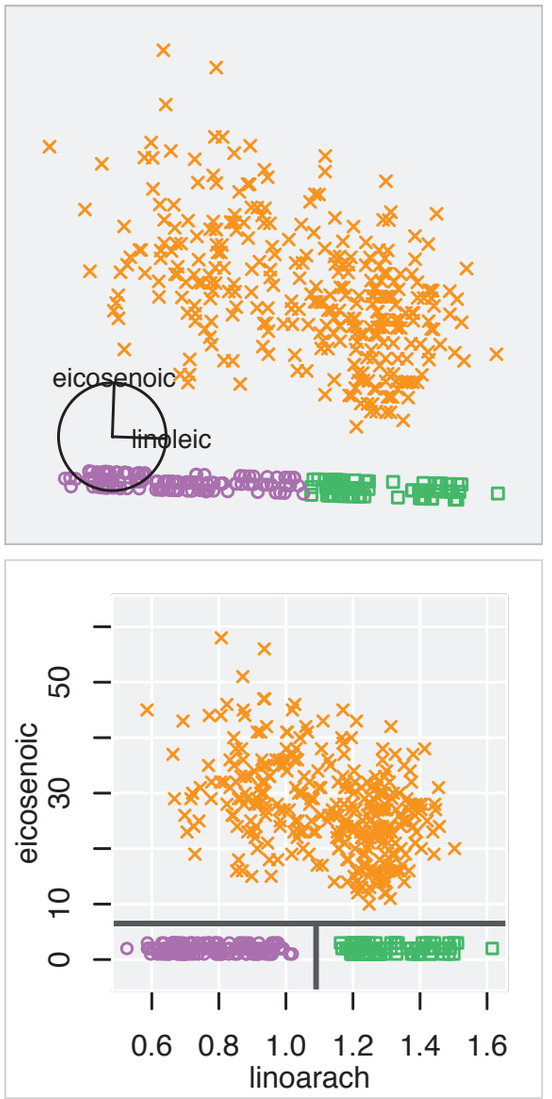
LDA

Trees

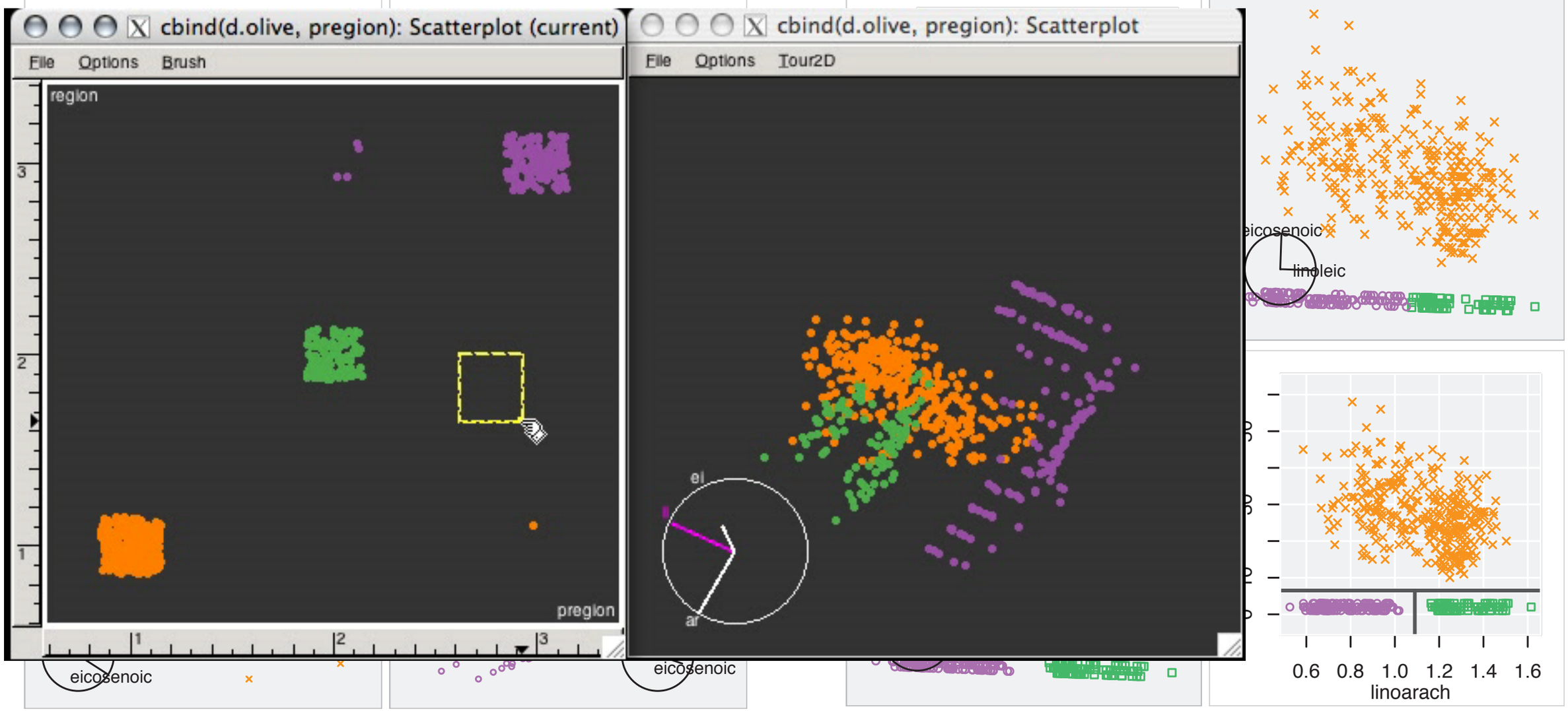

D2
D1
−6 −4 −2 0 2 4
−4 −2 0 2 4
linoleic
eicosenoic

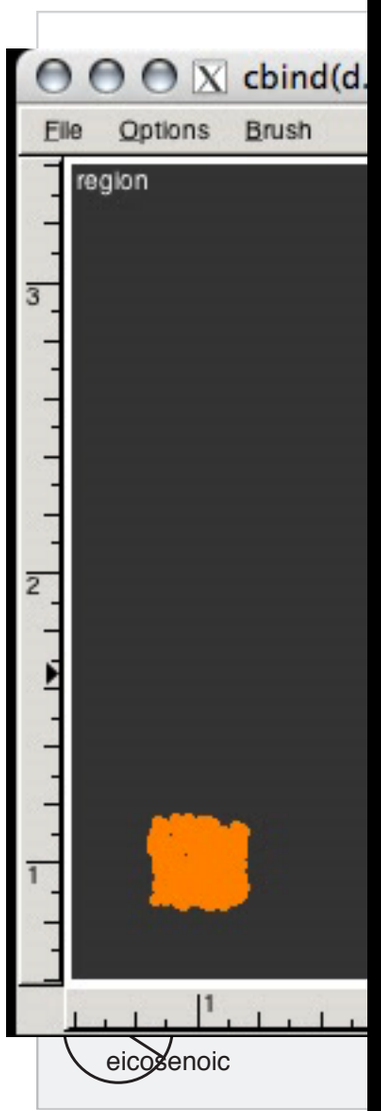

region
predicted region
1 1.5 2 2.5 3
arachidic
linoleic
eicosenoic


eicosenoic
linoleic
0 10 30 50
400 800 1200
eicosenoic
linoleic


eicosenoic
linoleic
eicosenoic
0 10 30 50
0.6 0.8 1.0 1.2 1.4 1.6
linoarach

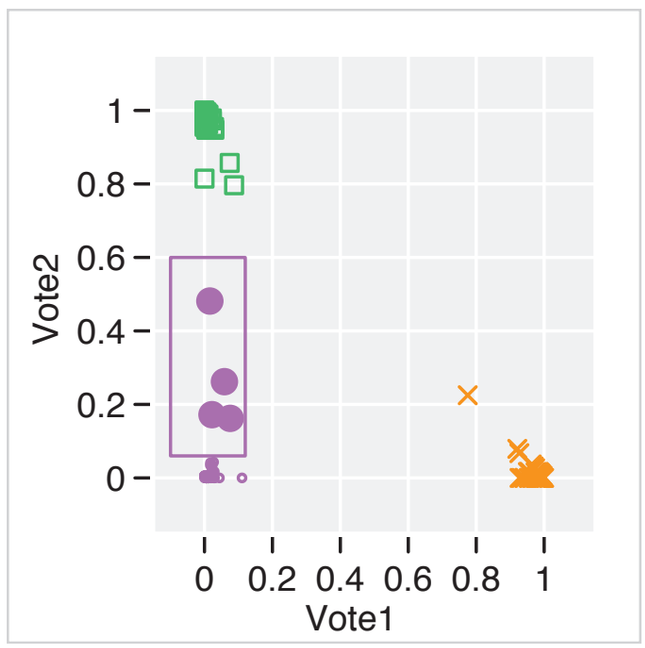
LDA Trees

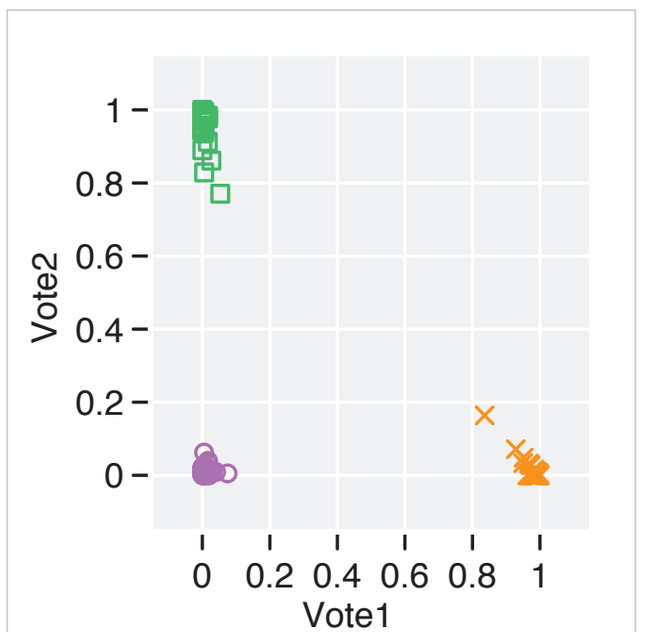

cbind(d.olive, pregion): Scatterplot (current)
File Options Brush
region
pregion
cbind(d.olive, pregion): Scatterplot
File Options Tour2D
ei
ar
eicosenoic
linoleic
0.6 0.8 1.0 1.2 1.4 1.6
linoarach



Trees

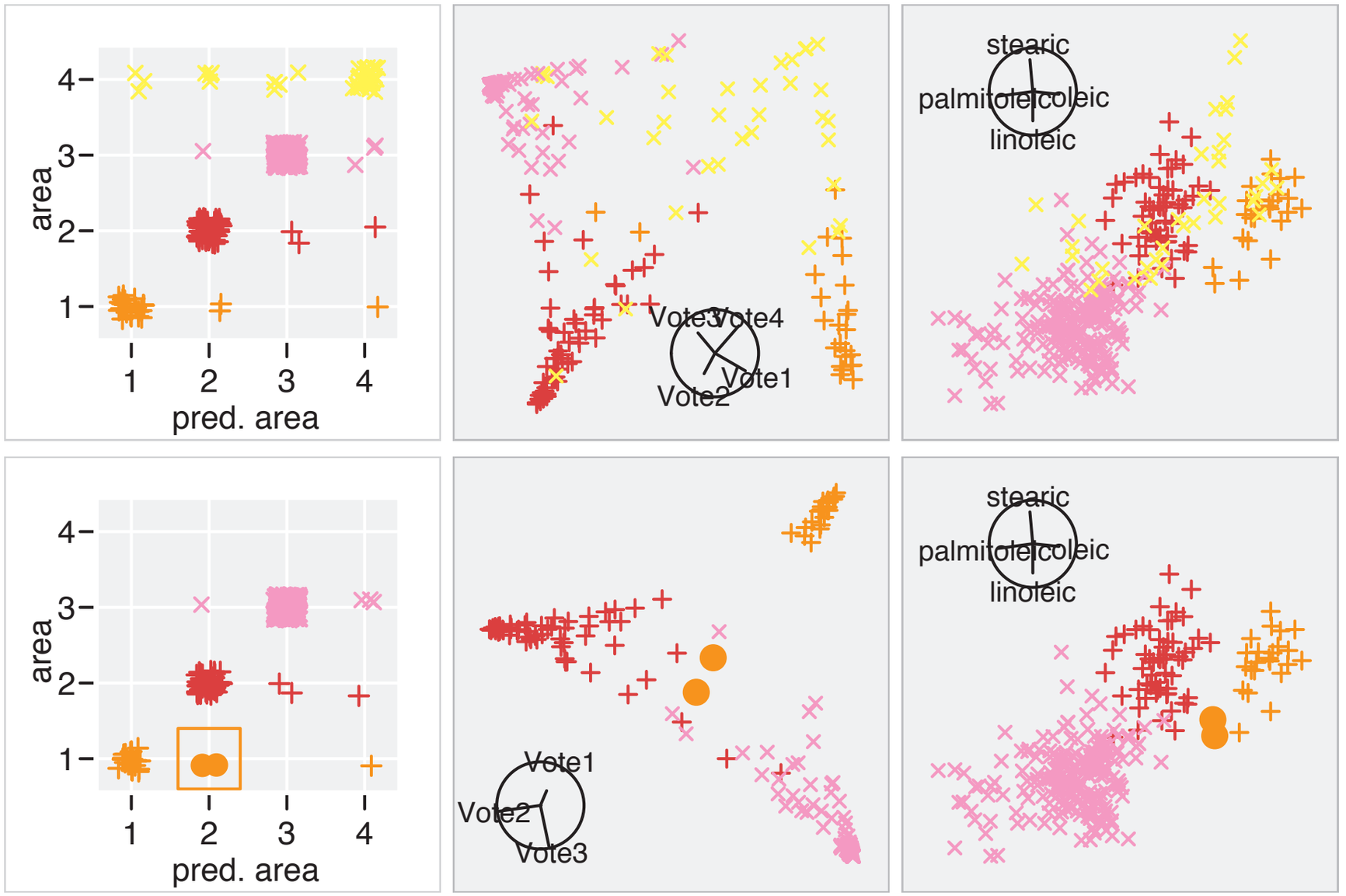
Random forests





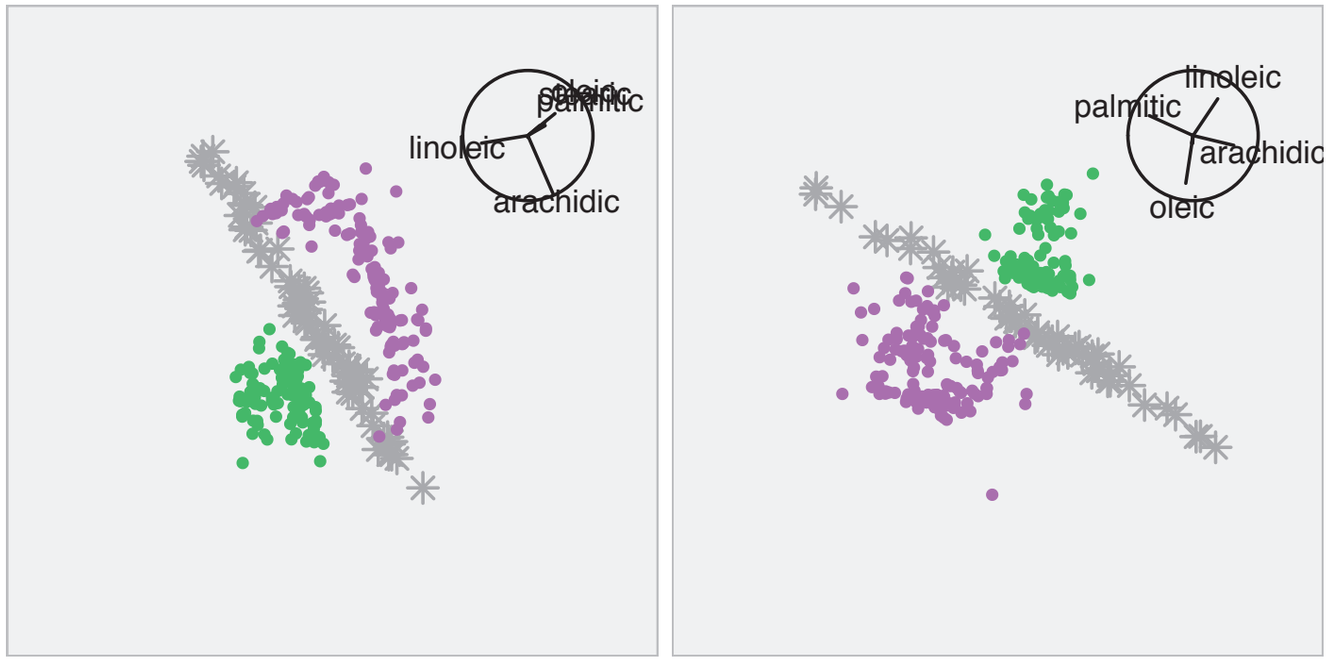
Random Forests

Random forests

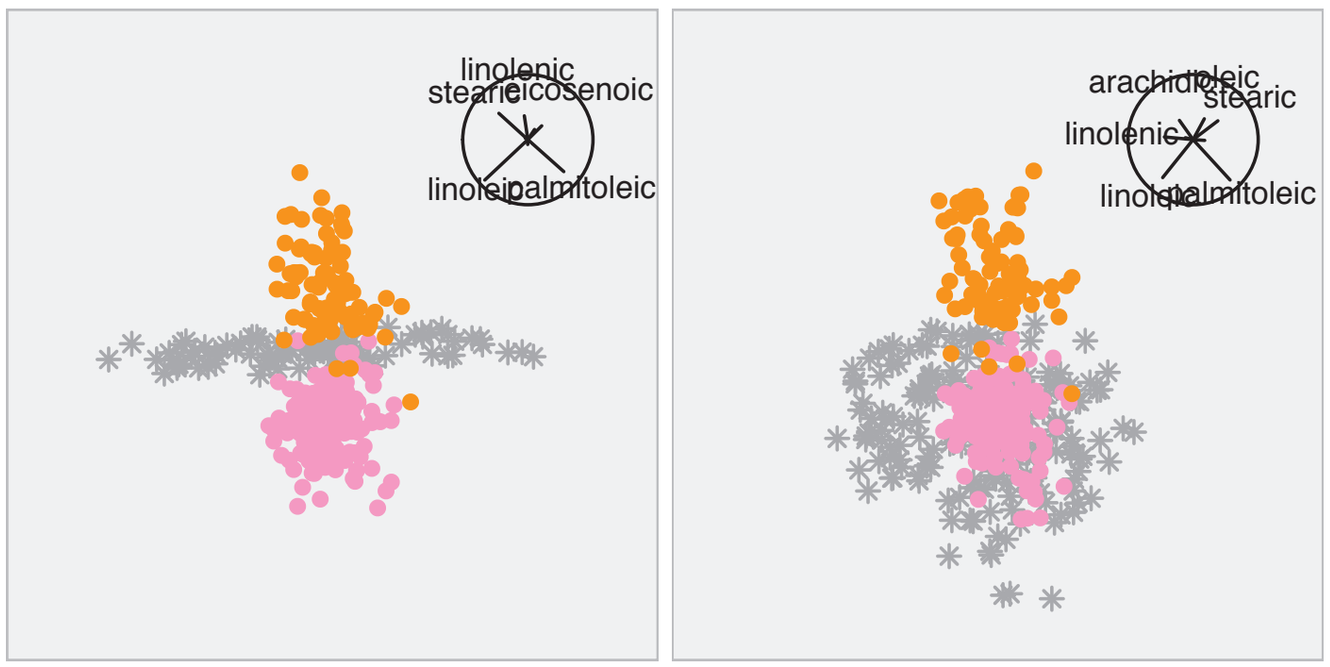


Support vector machines

LDA



Linear kernel



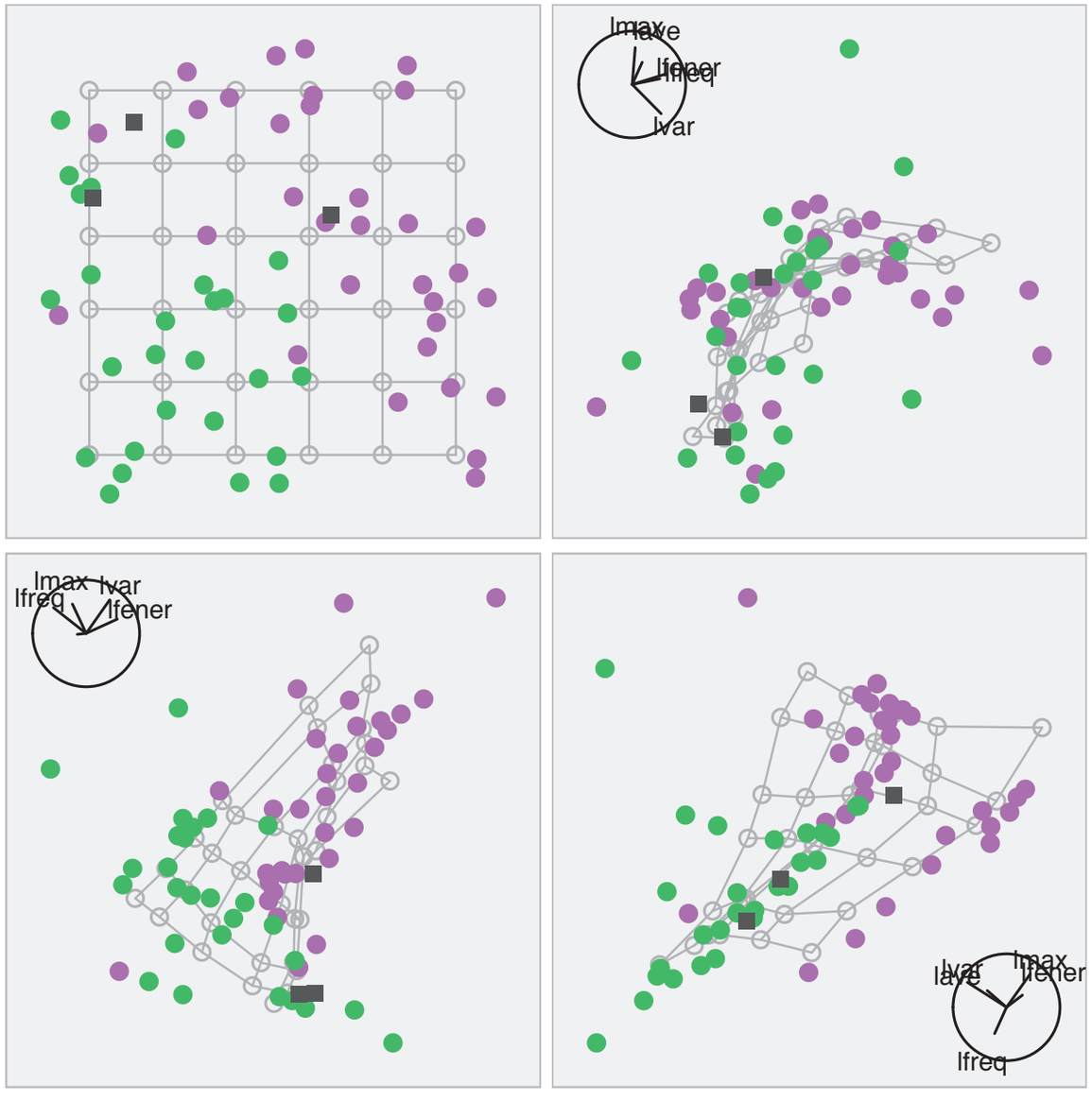
Linear kernel

Radial kernel

Examining boundaries

classifly

Self-organizing maps



- Model is fit by warping a sheet through high-d
- Lay out sheet to see proximities
- Instead, use the tour to examine how sheet fits the data

Self-organizing maps

rping
gh-d
ce
ur to
t fits

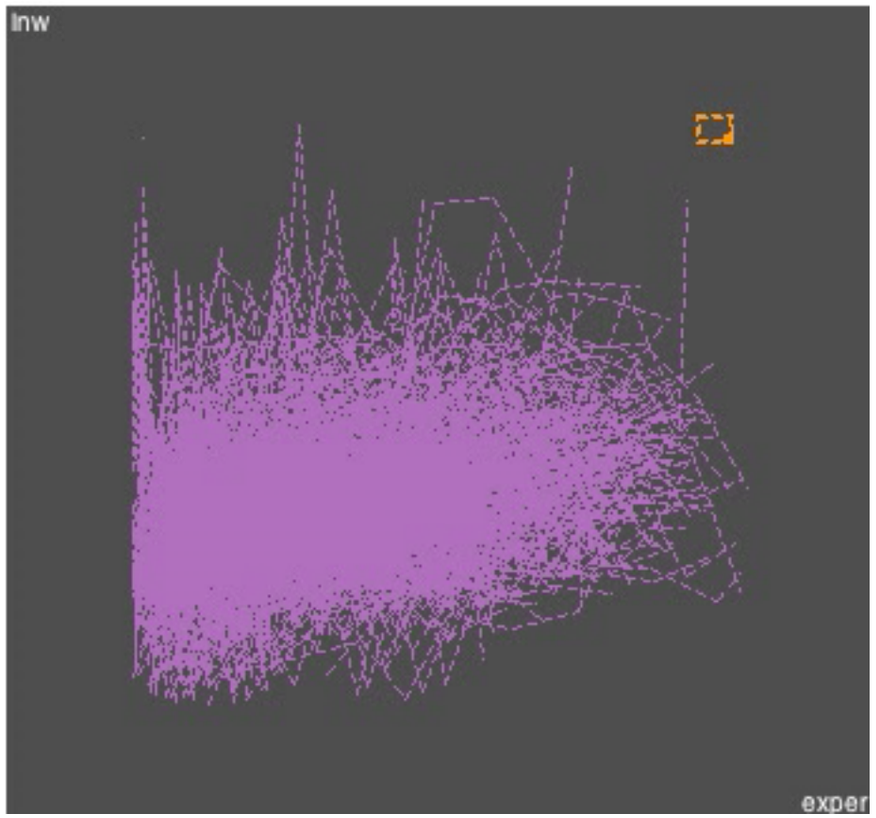
Additional topics

- Longitudinal data
- Ecological data (extend to large text datasets)
- Large biological data sets

Longitudinal data

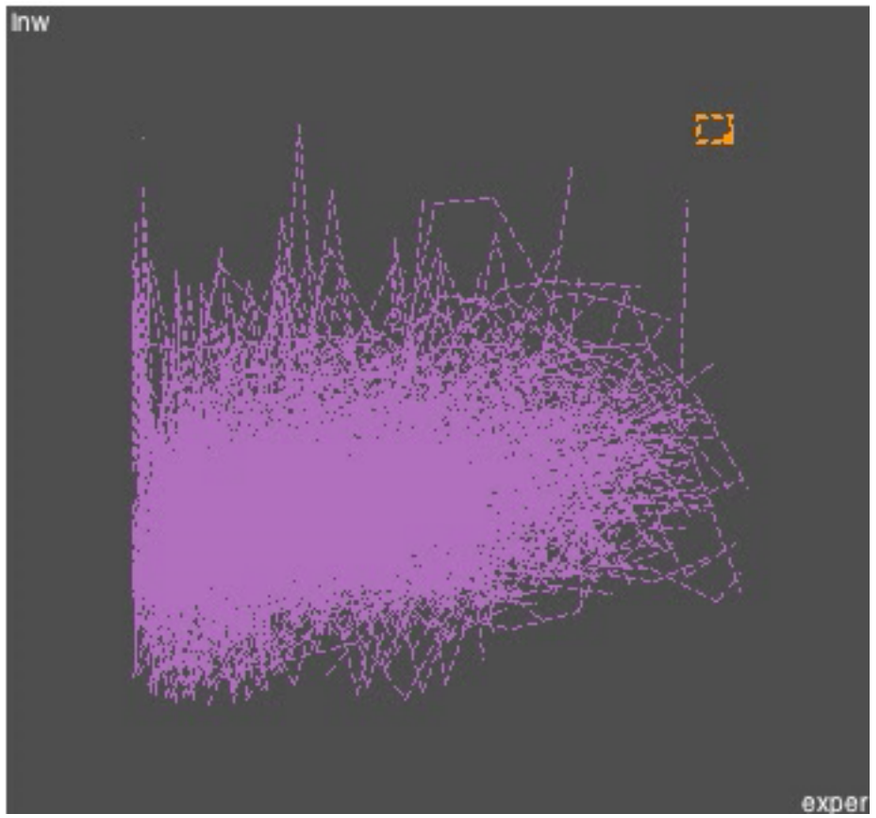
- Extensive use of categorical variables linking
- Explore fit of a linear mixed effect model
- Computing "cognostics" to help detect structure
- Principal components analysis/multidimensional scaling to explore space of cognostics

12 years of wages data for 888 subjects



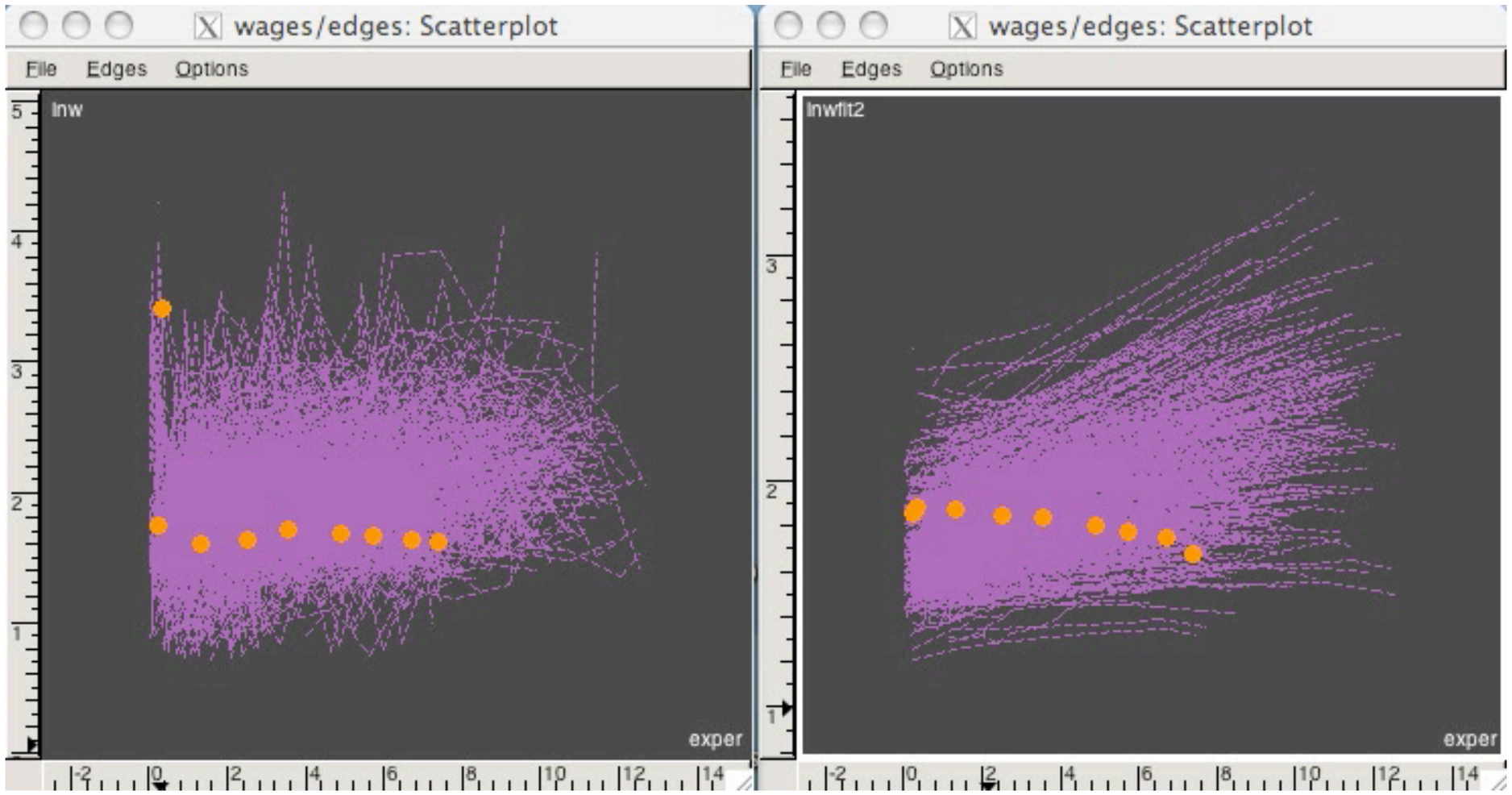
Overview individual wage histories. On average wages increase, and there are different increases depending on race and educational background.

Individually the experiences are different to the average.

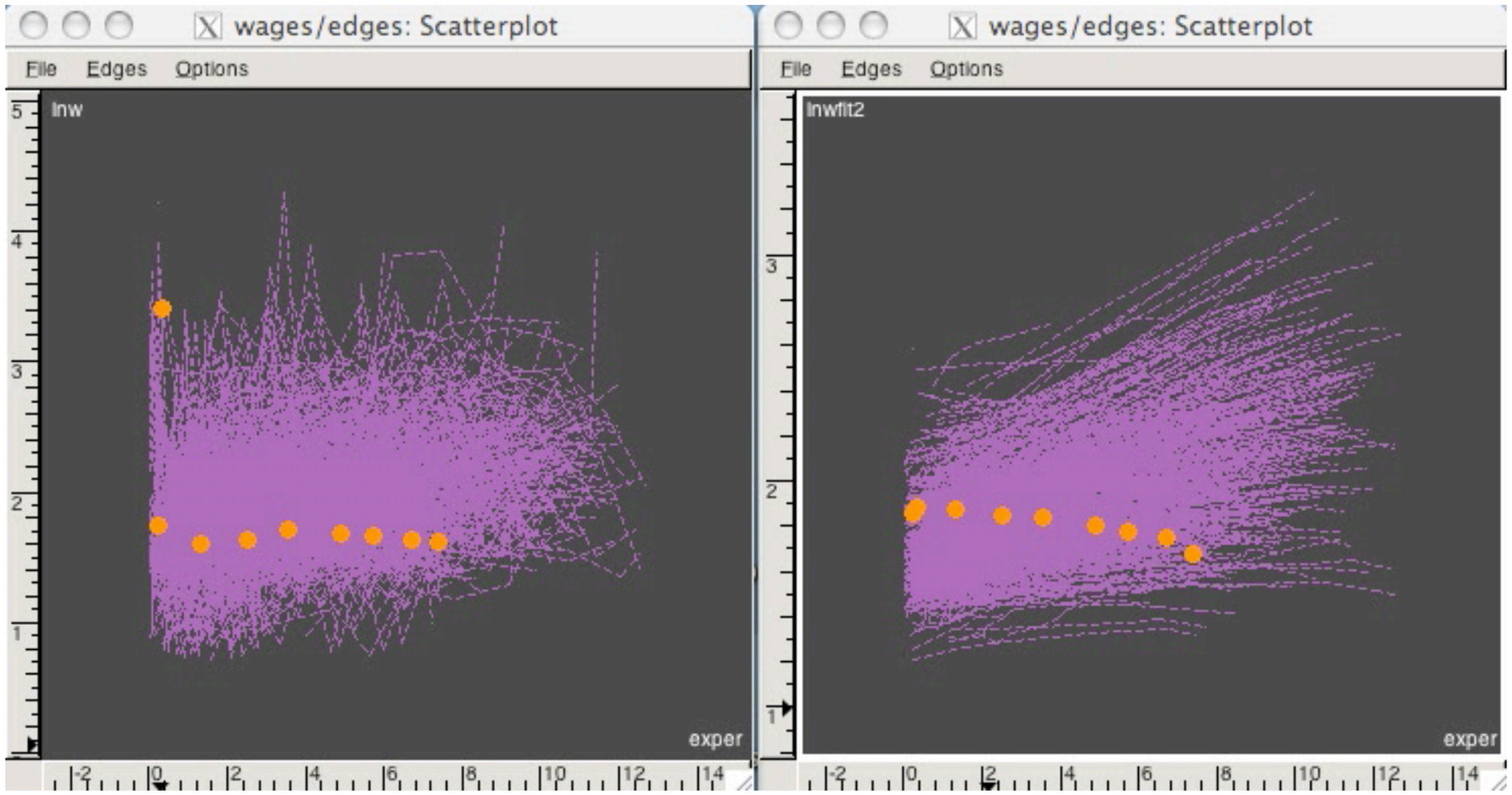


Overview individual wage histories. On average wages increase, and there are different increases depending on race and educational background.

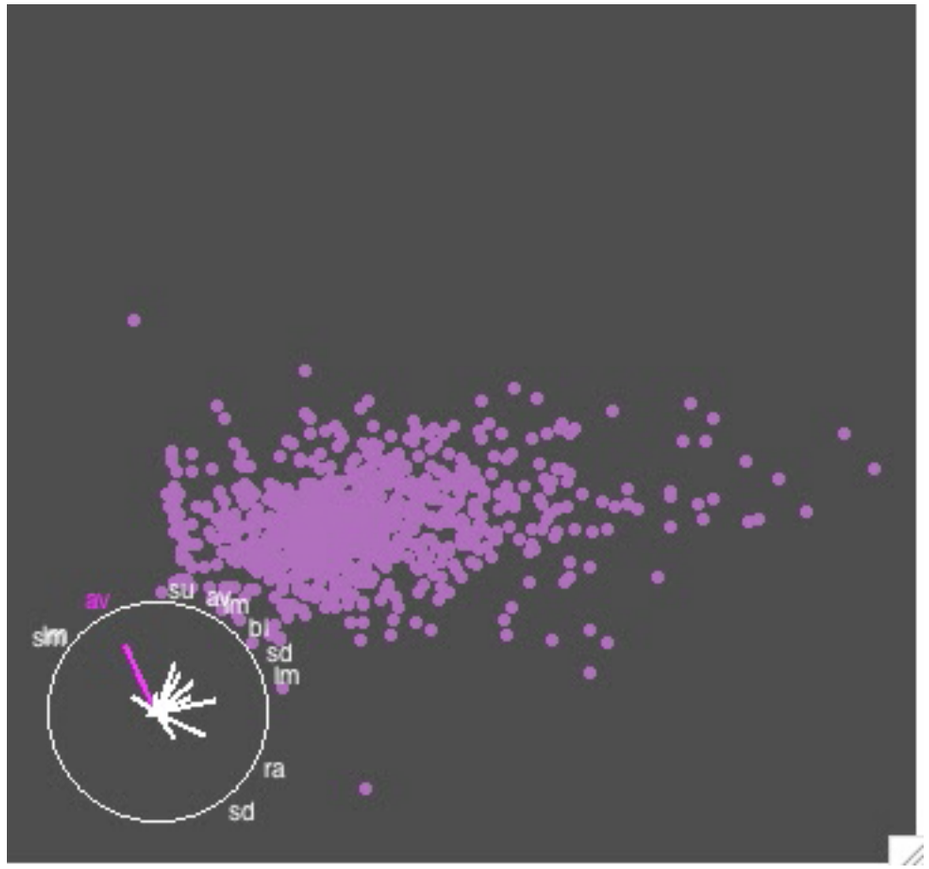
Individually the experiences are different to the average.



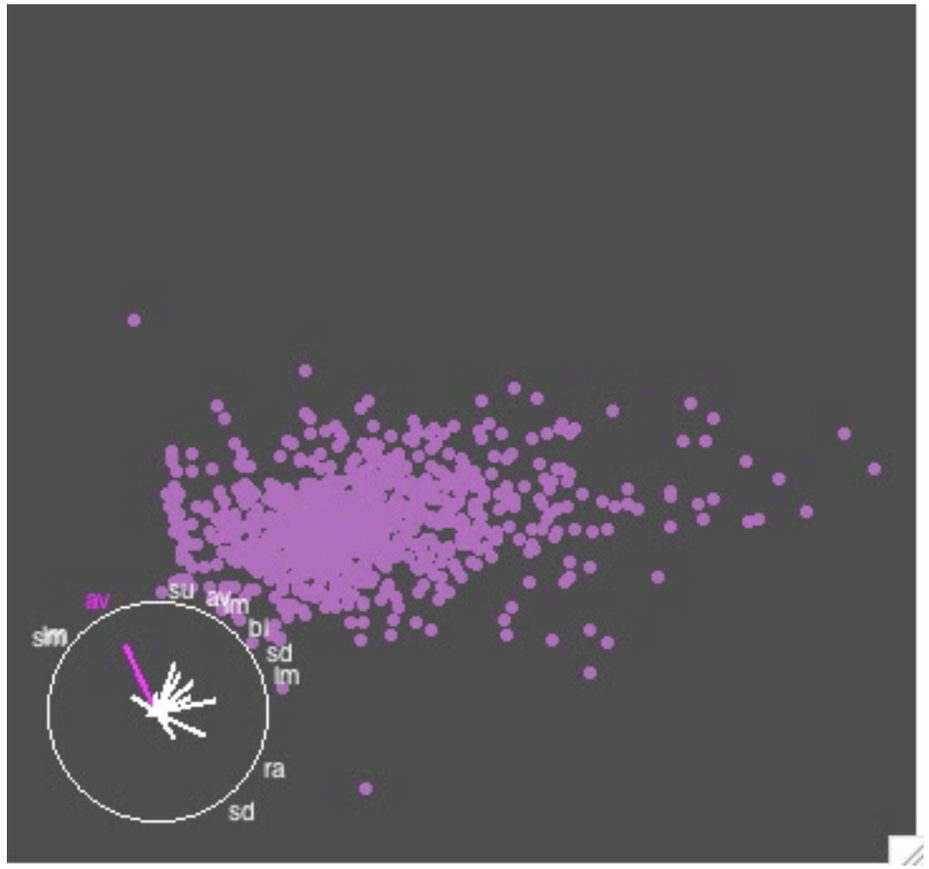
Fit a linear mixed effects model. Compare observed and fitted in separate plots. (Not ideal, but interesting problems with fit are evident.)



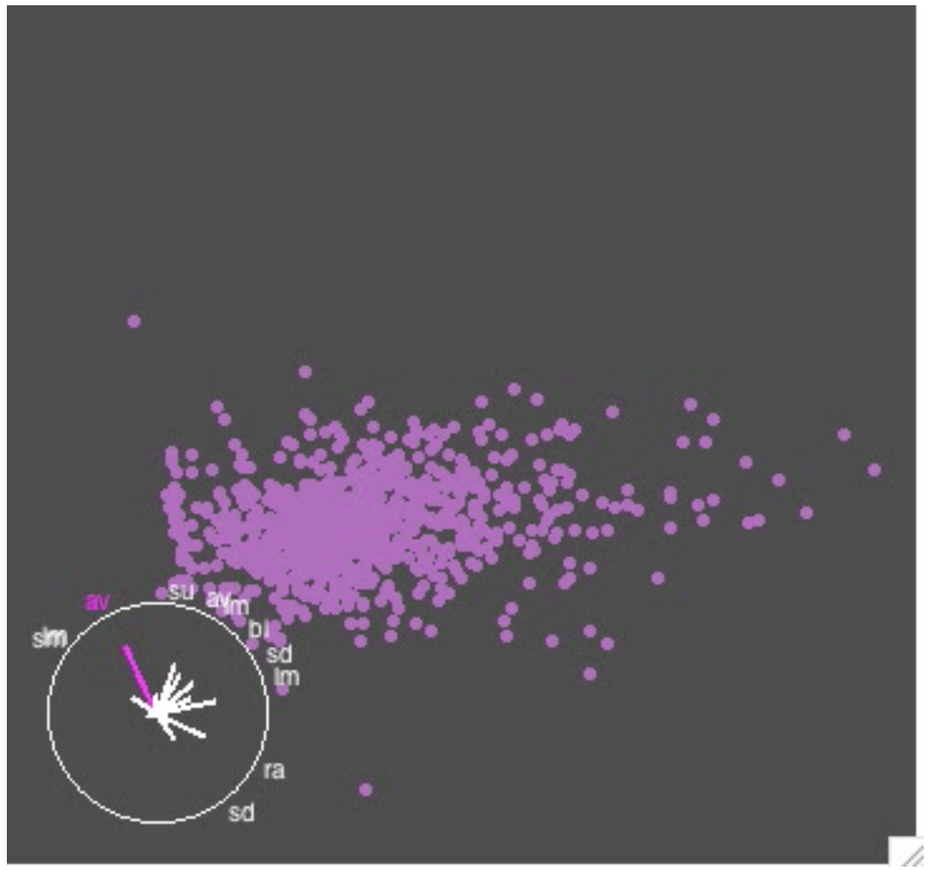
Fit a linear mixed effects model. Compare observed and fitted in separate plots. (Not ideal, but interesting problems with fit are evident.)



About 20 cognostics calculated to describe linear trend, variability, big jumps, ...

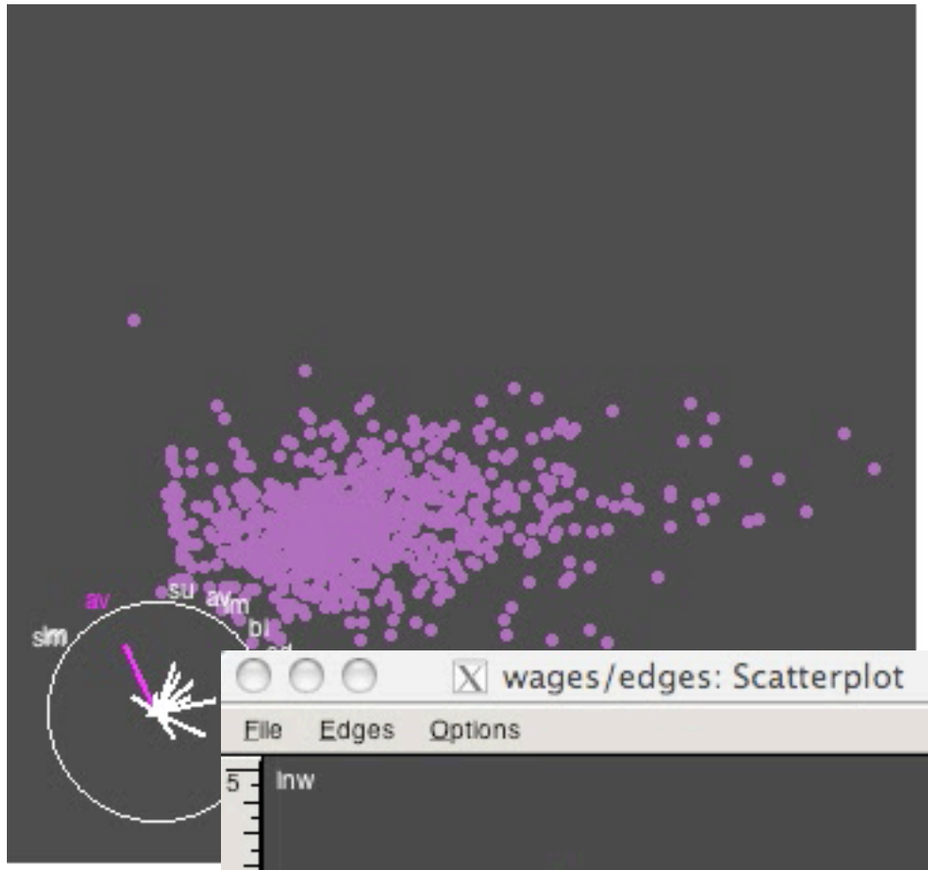


About 20 cognostics calculated to describe linear trend, variability, big jumps, ...



About 20 cognostics calculated to describe linear trend, variability, big jumps, ...

What do we see: clump of points (lots of people with similar values), a few outliers (a few people have different characteristics)



About 20 cognostics calculated to describe linear trend, variability, big jumps, ...

What do we see: clump of

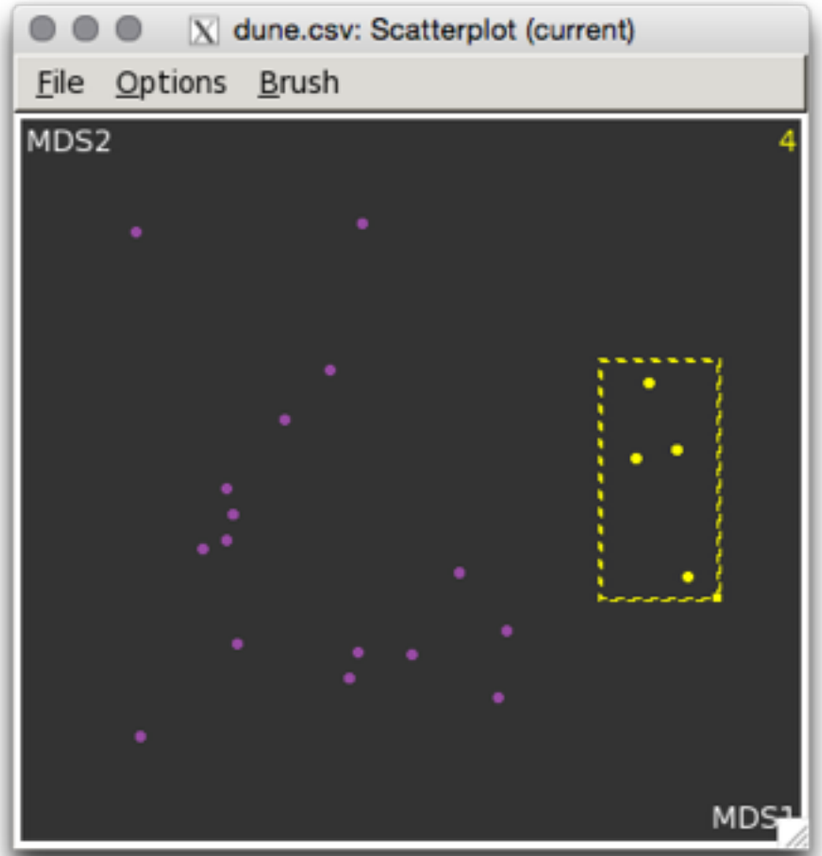
h

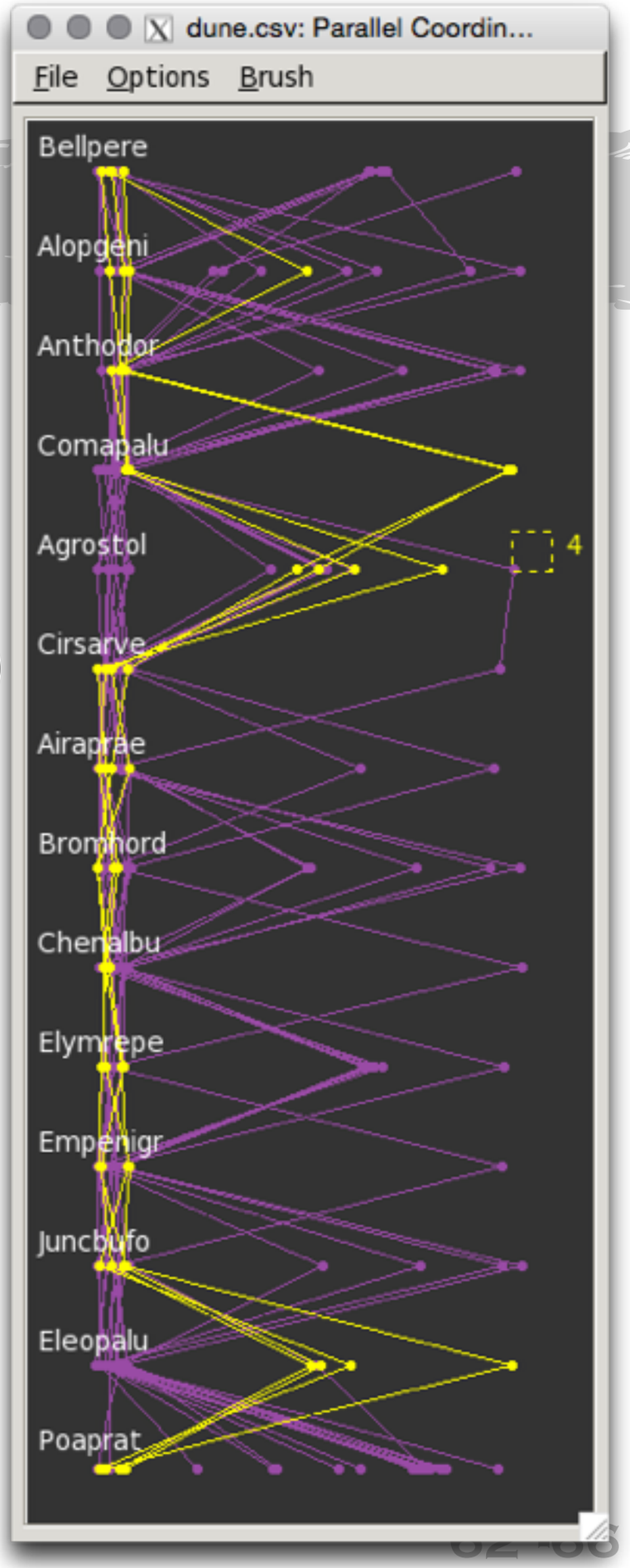
iers

ent

Ecological data

- **Sort** (species from highest to lowest count)
- Dimension reduction (MDS)
- Link MDS plot to par coords (use jittering)

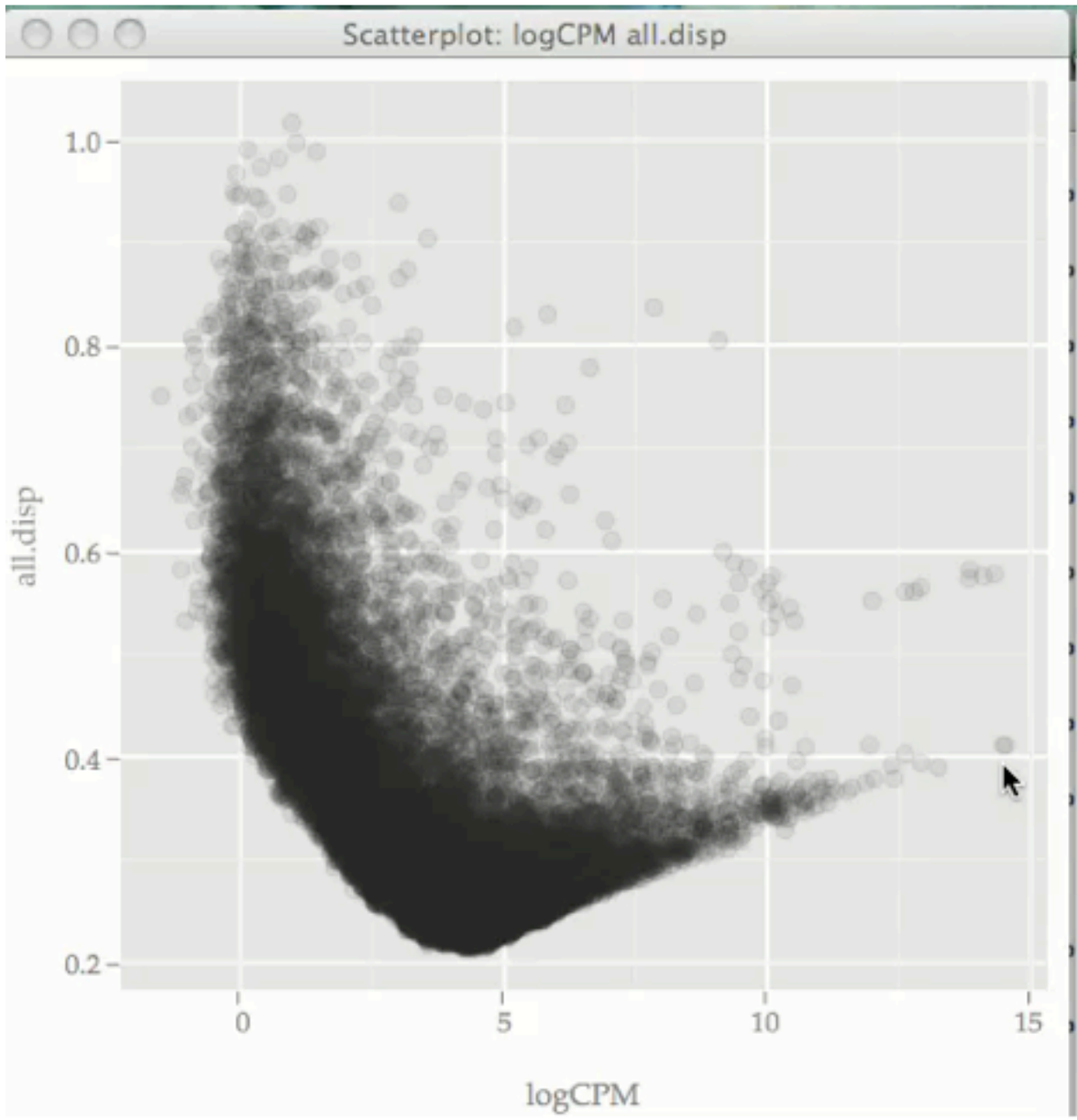


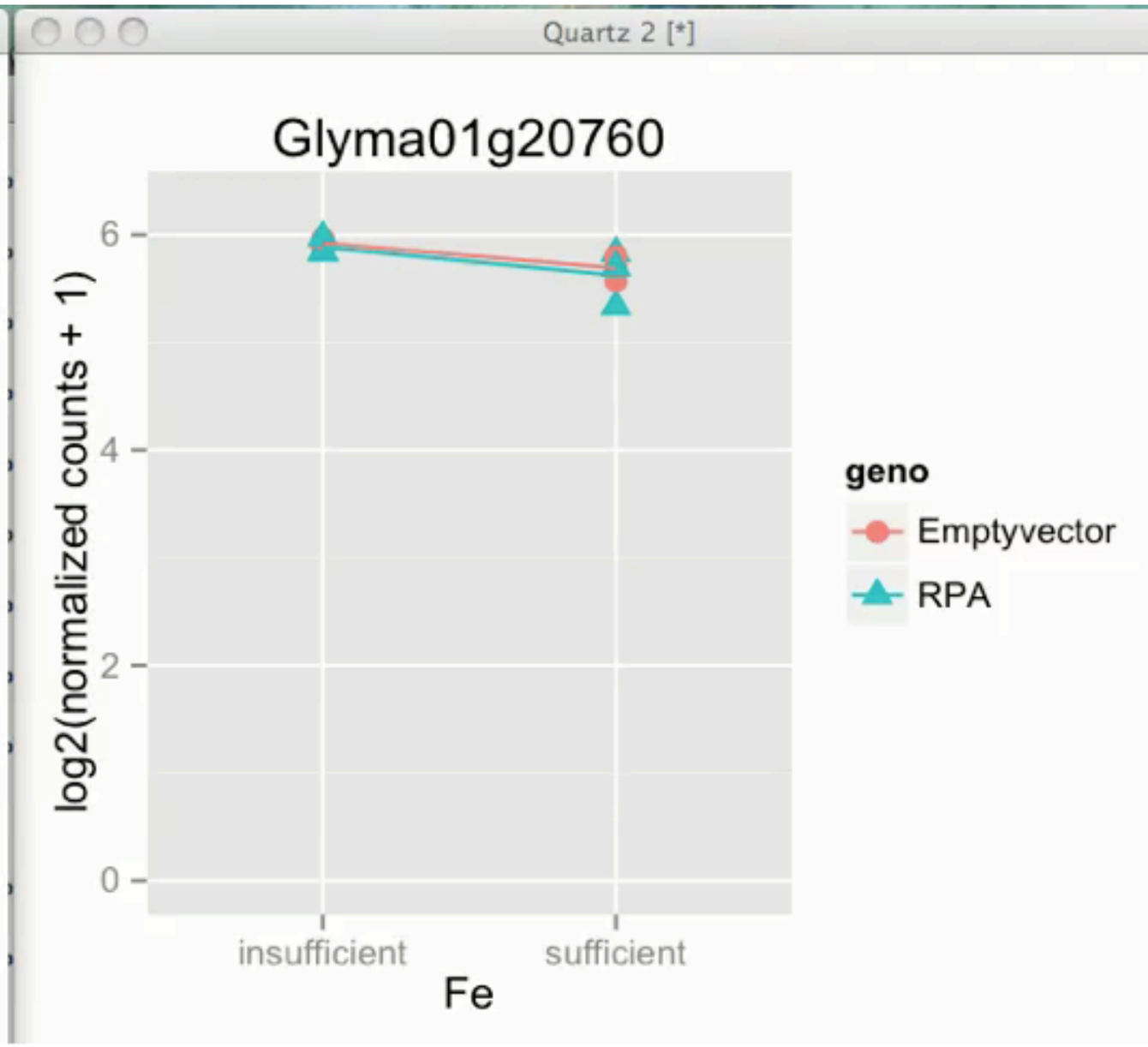


Biological data

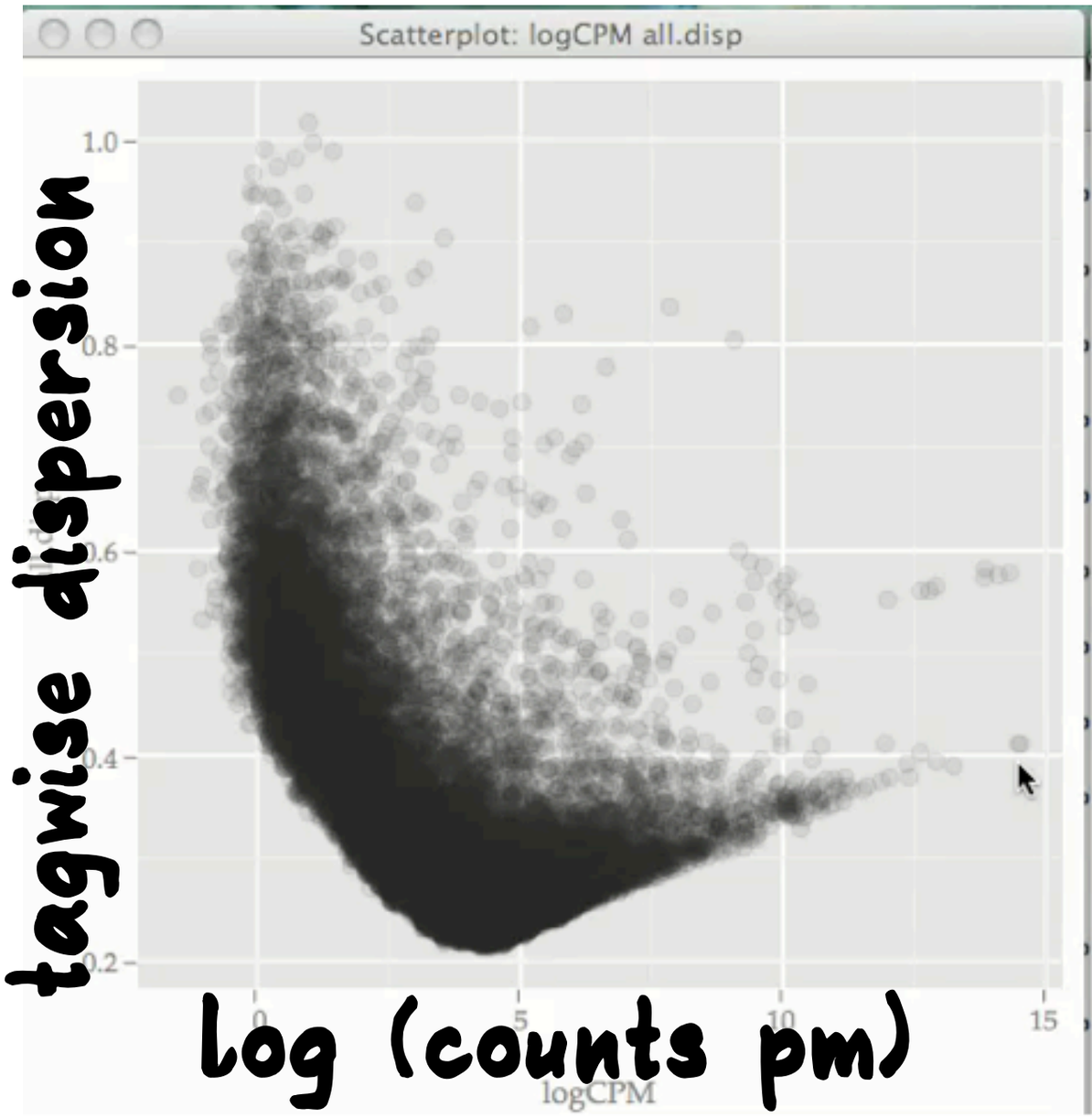
- ~30,000 genes tested in 2x2x3 design
- Estimate dispersion by “borrowing” from other genes
- Link model diagnostics to old-fashioned interaction plots

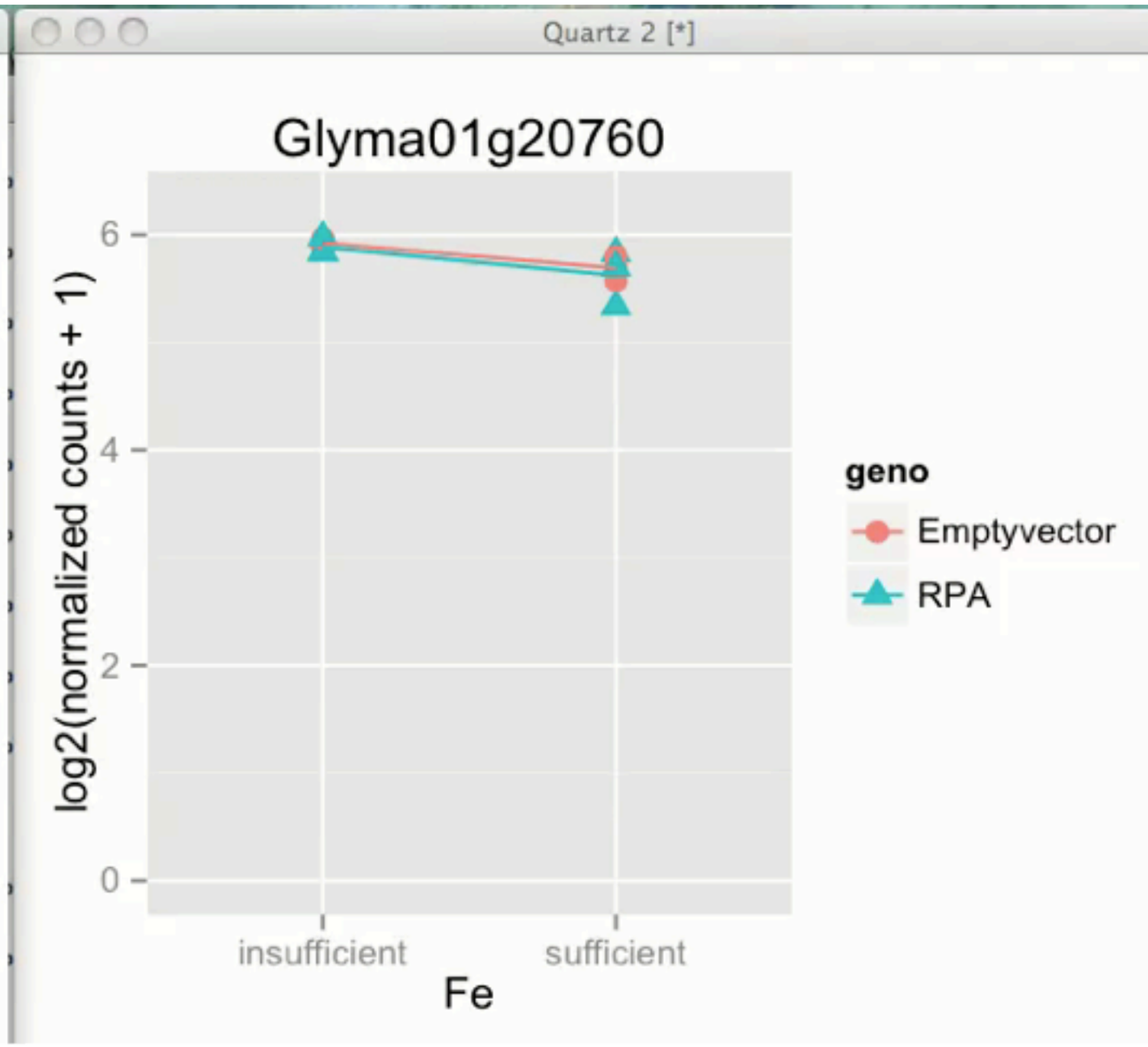
Biological data





Biological data





Summary

- We can see beyond 3D, with a combination of dynamic graphics and linking between multiple plots.
- Statistical graphics explores abstract relationships between variables, and enables building a conceptual map of structure in data
- It is important to examine the model fit IN THE DATA SPACE.

Acknowledgements & Resources

- Used R packages ggplot2, cranvas (http://cranvas.org), and also ggobi (http://www.ggobi.org), knitr/rmarkdown
- R packages: tourr (LDA/PDA indexes), PPTree, classifly, clusterfly
- Web scraping: XML, scrapeR, rvest
- New tools: ggvis, shiny, animint