Cluster Analysis

Organize observations into similar groups

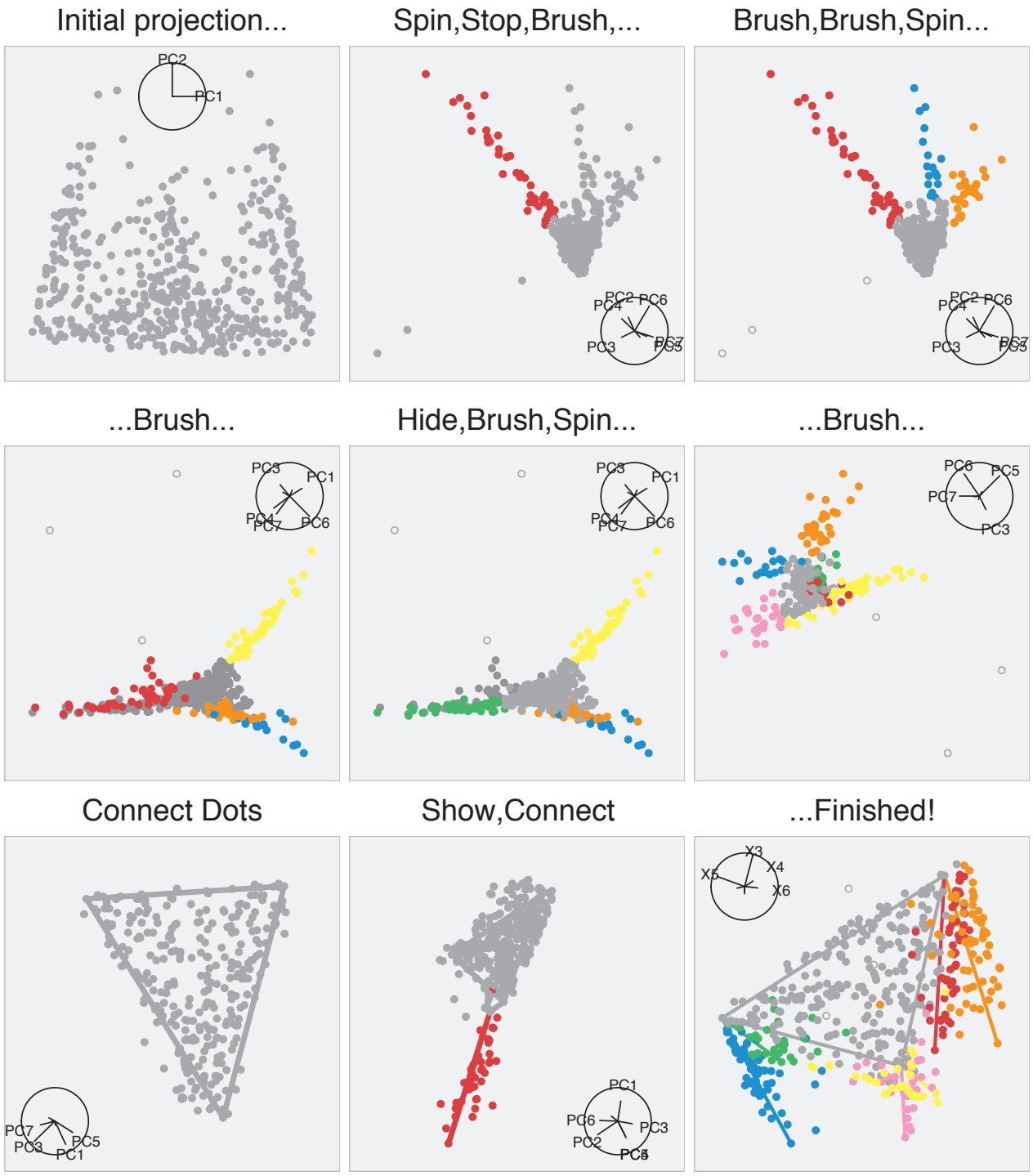
Outline

- Visual clustering
- Algorithmic clustering
  - Hierarchical clustering
  - Self-organising maps

Graphical clustering

How many clusters are cluster-unknown.csv?

Use brush and spin to identify them



Spin and brush

Tour, paint a cluster, continue touring, until no more clusters are revealed.

What is similar?

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \mathbf{X}_3 \end{bmatrix} = \begin{bmatrix} 7.3 & 7.6 & 7.7 & 8.0 \\ 7.4 & 7.2 & 7.3 & 7.2 \\ 4.1 & 4.6 & 4.6 & 4.8 \end{bmatrix}$$

$$d_{\text{Euc}}(\mathbf{X}_i, \mathbf{X}_j) = ||\mathbf{X}_i - \mathbf{X}_j|| \qquad i, j = 1, \ldots, n,$$

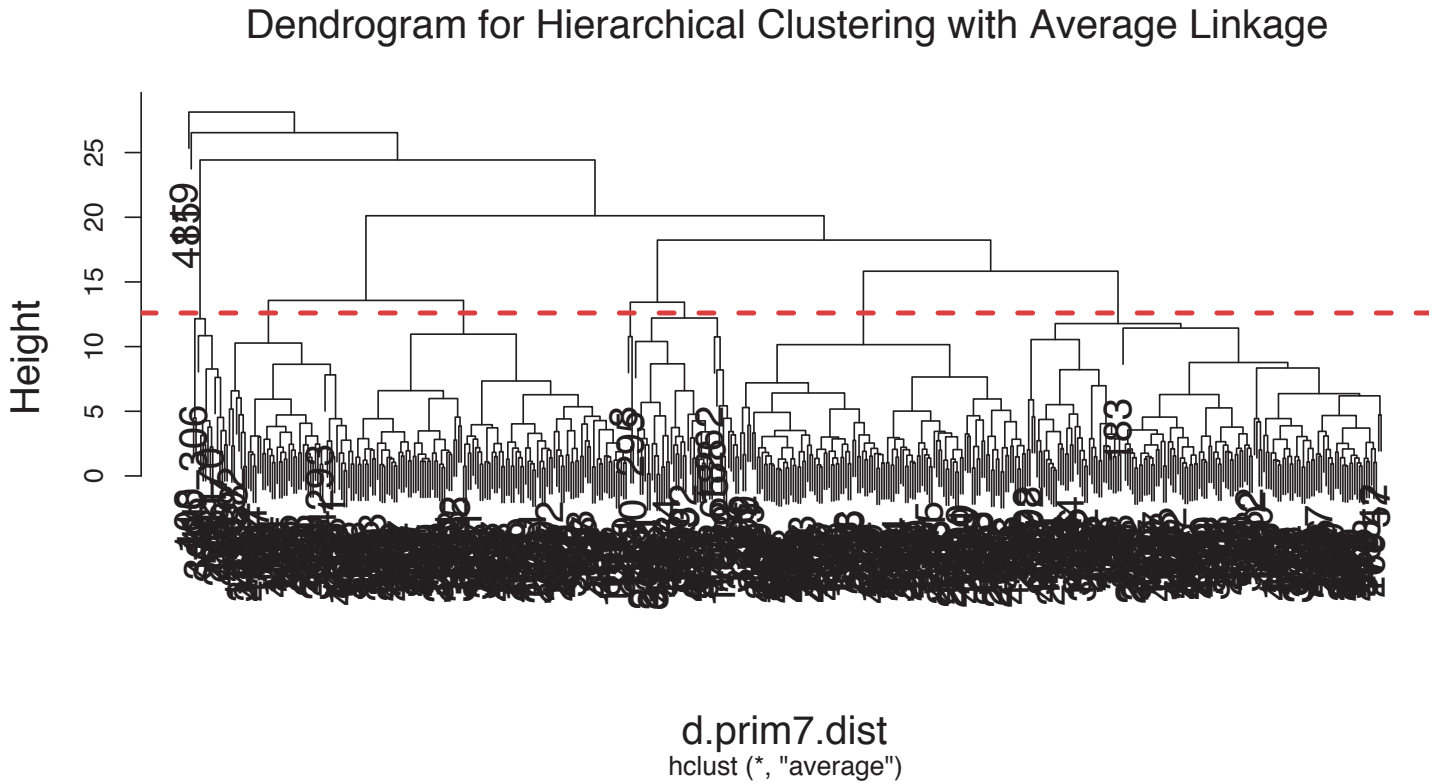
where $||\mathbf{X}_i|| = \sqrt{X_{i1}^2 + X_{i2}^2 + \ldots + X_{ip}^2}$. For example, the Euclidean distance between cases 1 and 2 in the above data, is

$$\sqrt{(7.3 - 7.4)^2 + (7.6 - 7.2)^2 + (7.7 - 7.3)^2 + (8.0 - 7.2)^2} = 1.0.$$

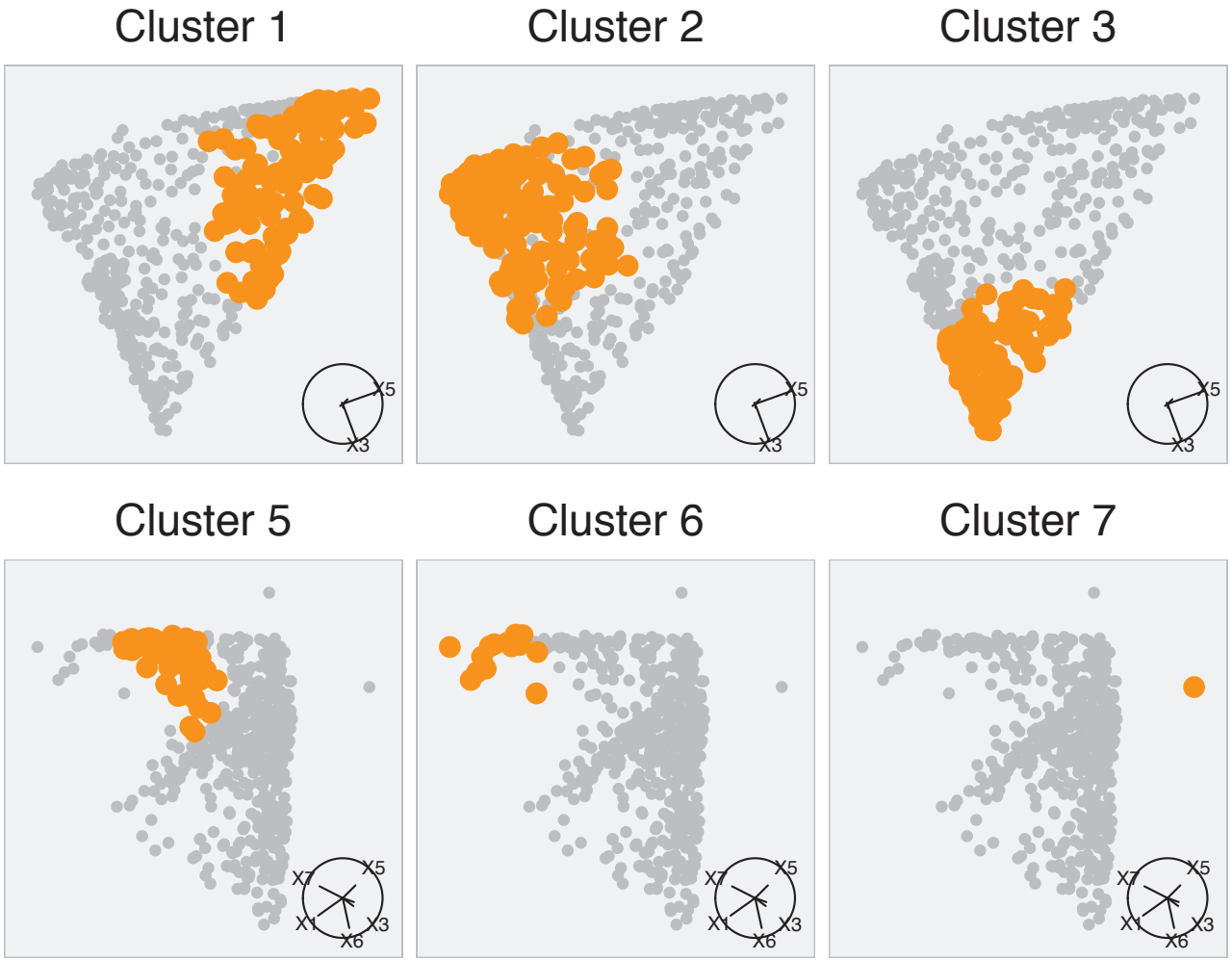
$$d_{\text{Euc}} = \begin{bmatrix} 0.0 & & \\ 1.0 & 0.0 & \\ 6.3 & 5.5 & 0.0 \end{bmatrix} \begin{matrix} \mathbf{X}_1 \\ \mathbf{X}_2 \\ \mathbf{X}_3 \end{matrix}$$

Hierarchical clustering

```
> library(rggobi)
> d.prim7 <- read.csv("prim7.csv")
> d.prim7.dist <- dist(d.prim7)
> d.prim7.dend <- hclust(d.prim7.dist, method="average")
> plot(d.prim7.dend)
```



```
> gd <- ggobi(d.prim7)[1]
> clust9 <- cutree(d.prim7.dend, k=9)
> glyph_color(gd)[clust9==1] <- 9 # highlight triangle
> glyph_color(gd)[clust9==1] <- 1 # reset color
> glyph_color(gd)[clust9==2] <- 9 # highlight cluster 2
```



Clusters 1, 2, 3 carve up the base triangle.

Clusters 5, 6 divide an arm.

Cluster 7 is a singleton cluster containing an outlier.

Self-organizing maps

- A constrained k-means algorithm.
- A 1D or 2D net is stretched through the data. The knots in the net form the cluster centres, and the points closest to the knot are considered to belong to that cluster.
- Net can be unwrapped and laid out flat to give a visual representation of the clusters.

Music data

- Descriptive statistics for 62 songs:
  Variance, Mean, Max, Energy & Frequency
- 32 rock (Abba, Beatles, Eels)
  27 classical (Beethoven, Mozart, Vivaldi)
  3 new wave (Enya)
- Can a computer identify the different types?

Can you hear the difference?

- One
- Two
- Three

Can you hear the difference?

- One
- Two
- Three

Can you hear the difference?

- One "Rock" - Abba
- Two
- Three

Can you hear the difference?

- One "Rock" - Abba
- Two
- Three

Can you hear the difference?

- One "Rock" - Abba
- Two New wave - Enya
- Three

Can you hear the difference?

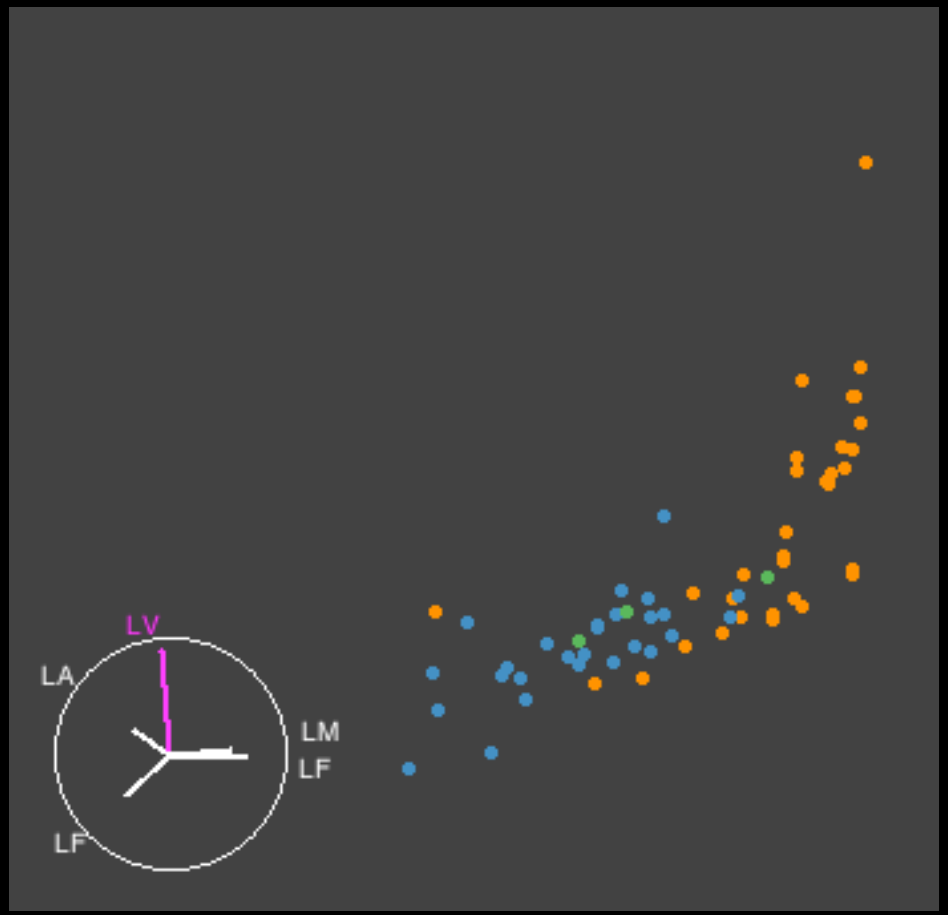
- One "Rock" - Abba
- Two New wave - Enya
- Three

Can you hear the difference?

- One "Rock" - Abba
- Two New wave - Enya
- Three Classical - Vivaldi

Can you see the difference?

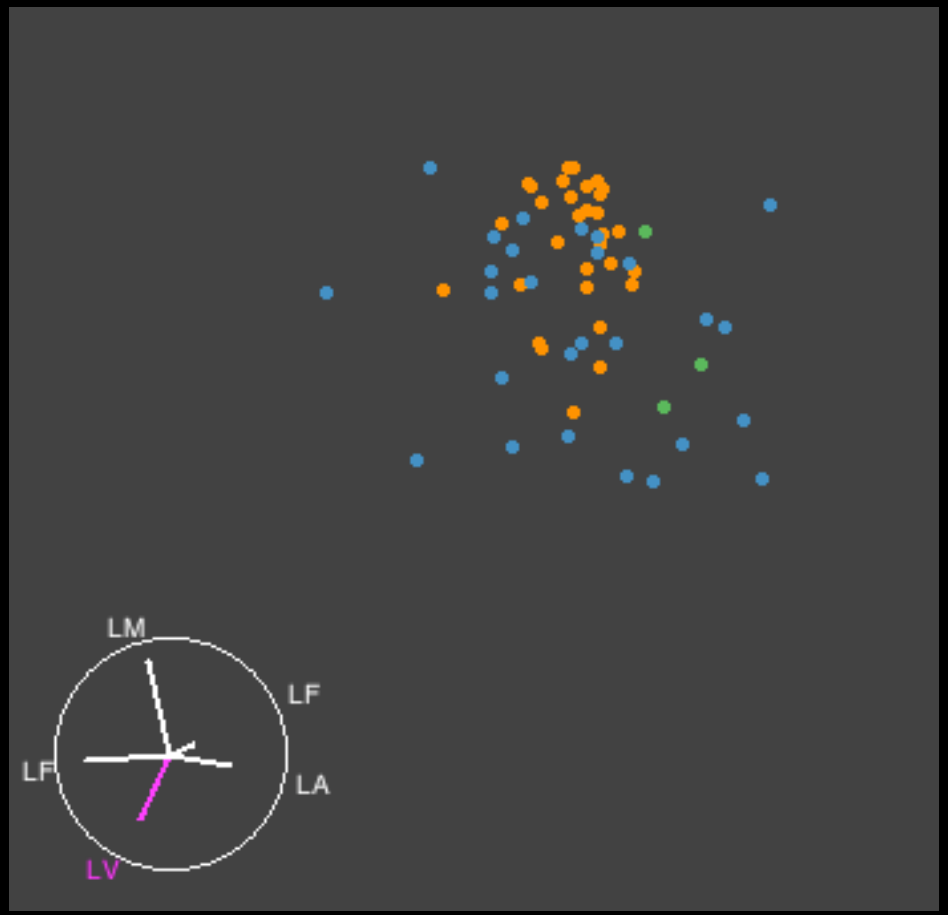
Is rock different from classical?



Can you see the difference?

Is rock different from classical?

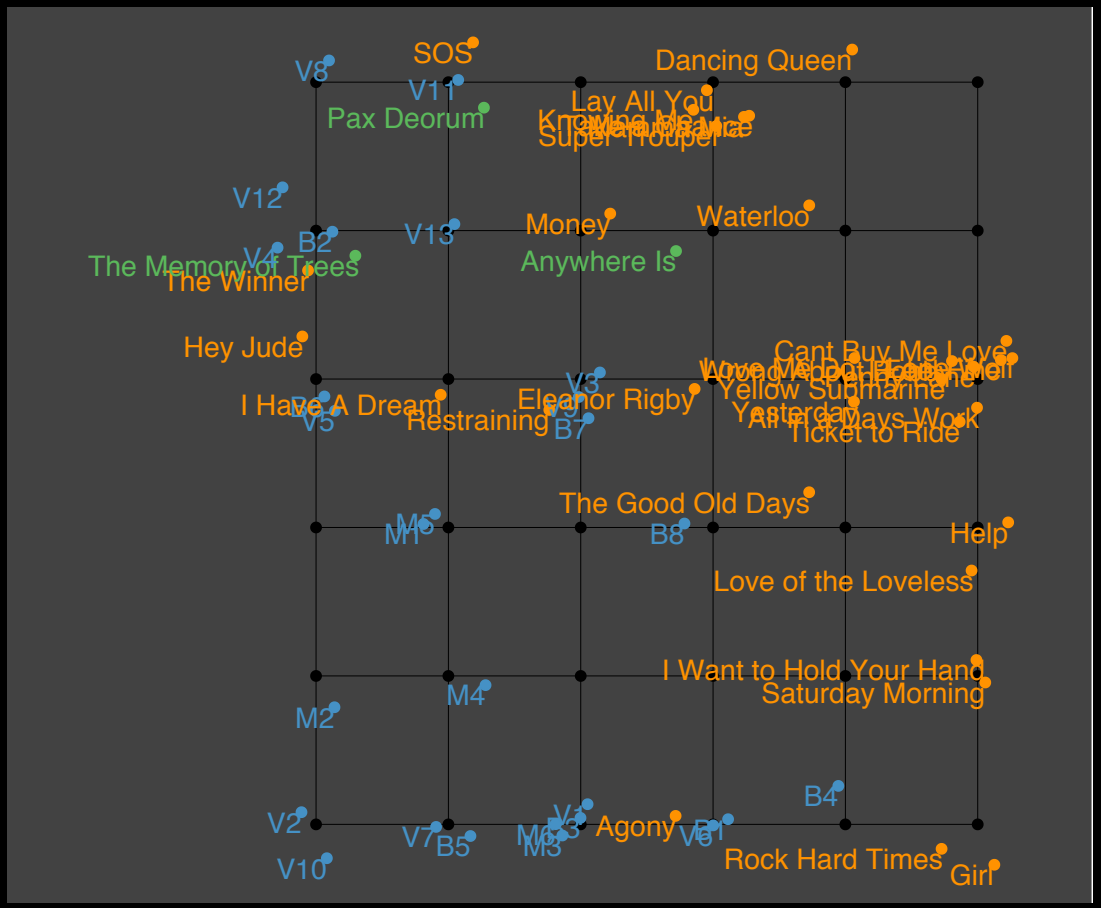
Can also see the shape in 5d: outliers and non-linearity



Model & data

data in the model space

SOM

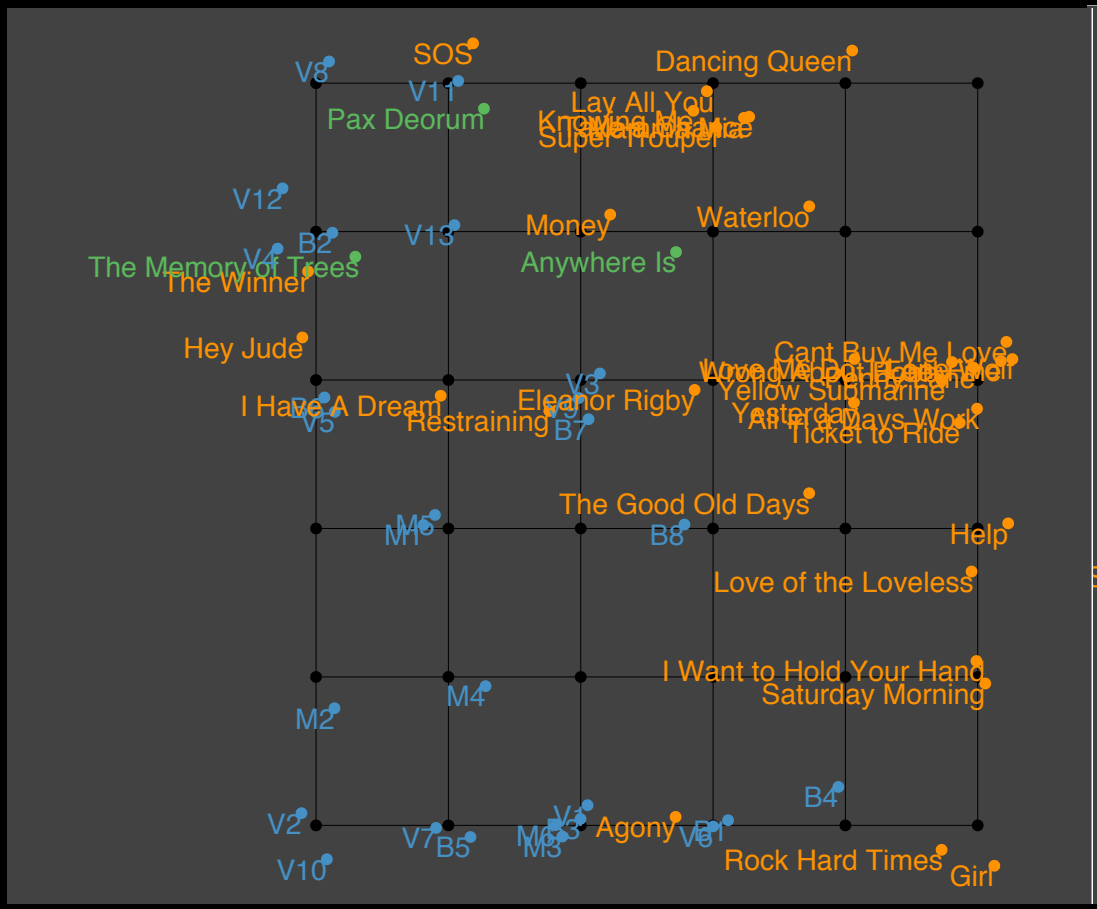


How well does the model fit the data?

Model & data

data in the model space

SOM



PCA

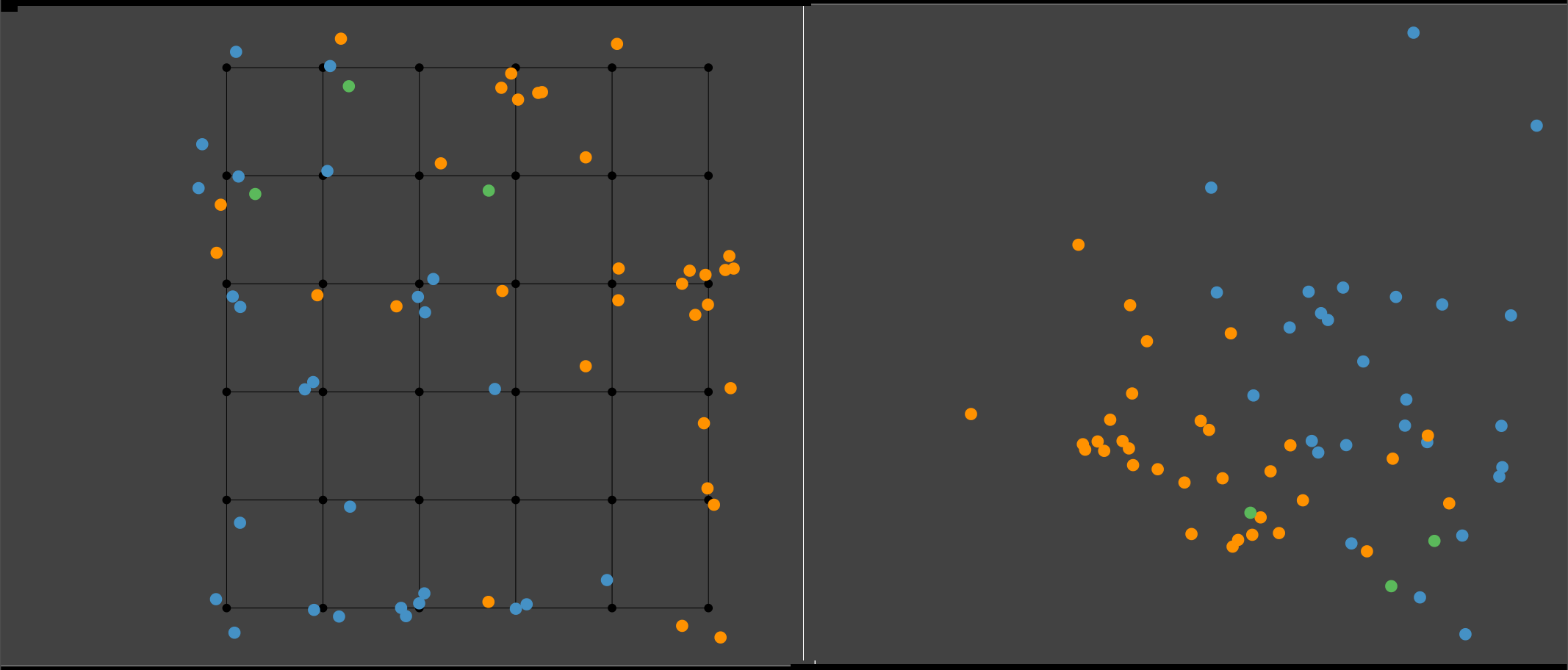
How well does the model fit the data?

Model & data

data in the model space

SOM

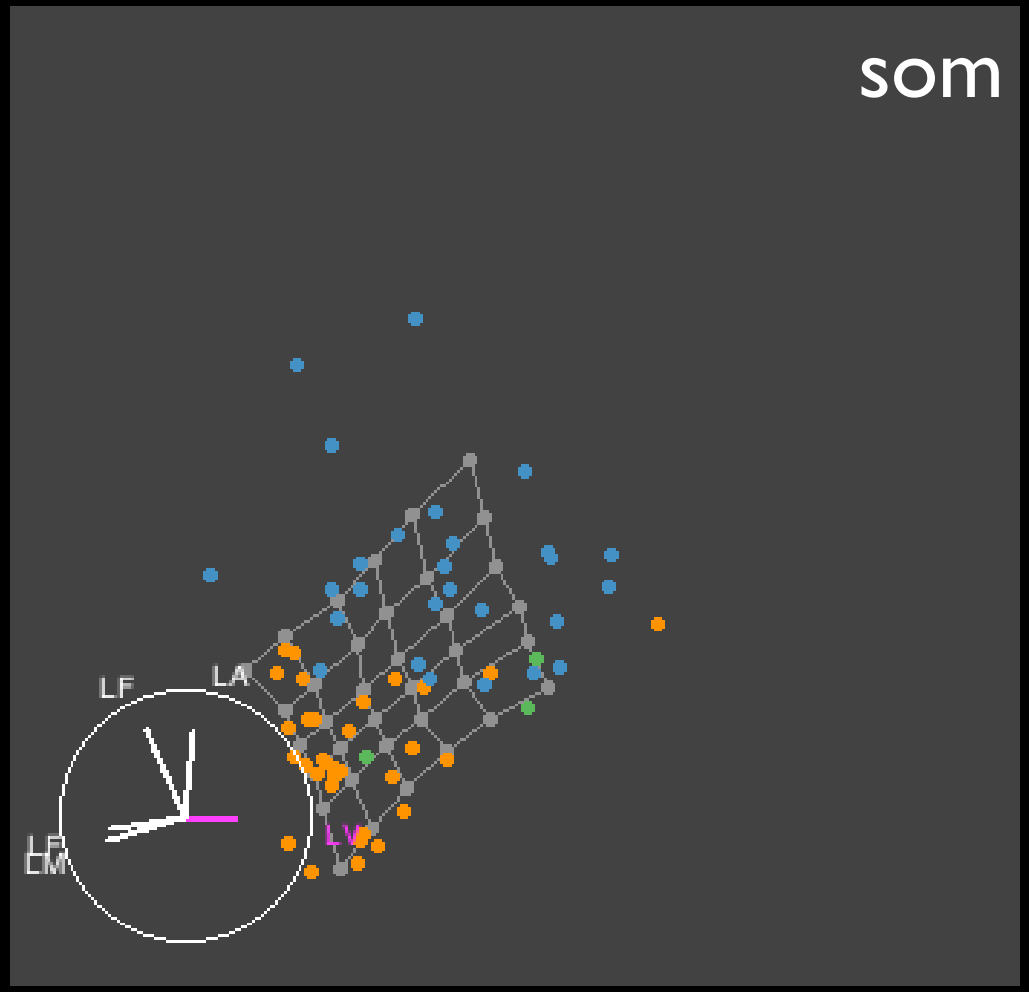
PCA

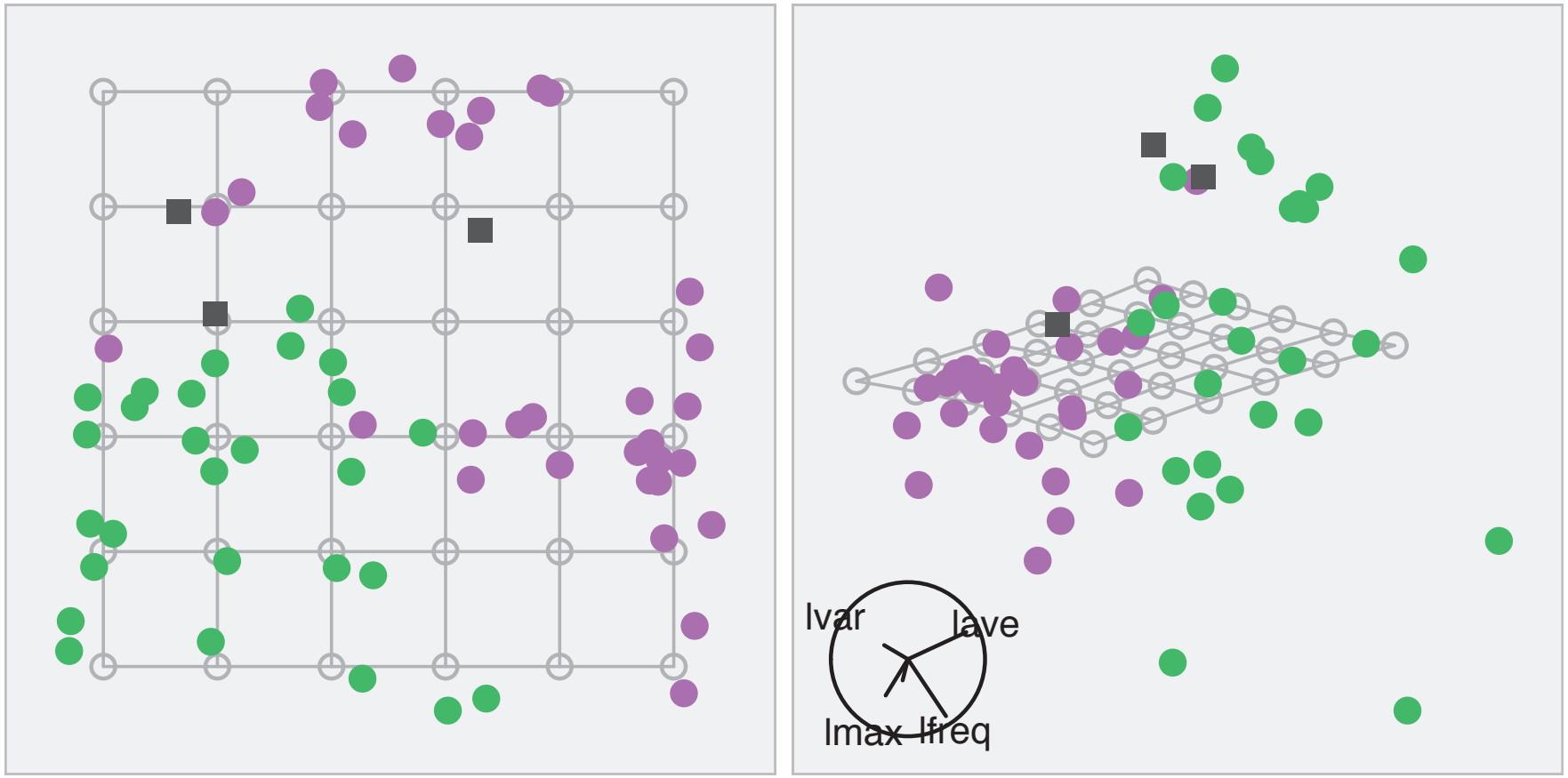


How well does the model fit the data?

Model & data

model in the data space





First attempt at model fitting doesn't work. Model appears to be unfinished, flat in the data space rather than wrapped into the points.

Next attempt looks better.

Problem was solved by increasing the number of iterations used.

Deciding on a solution

- One part in coming to a decision about the best solution is to compare the results from different methods.
- This can be done using a confusion table.

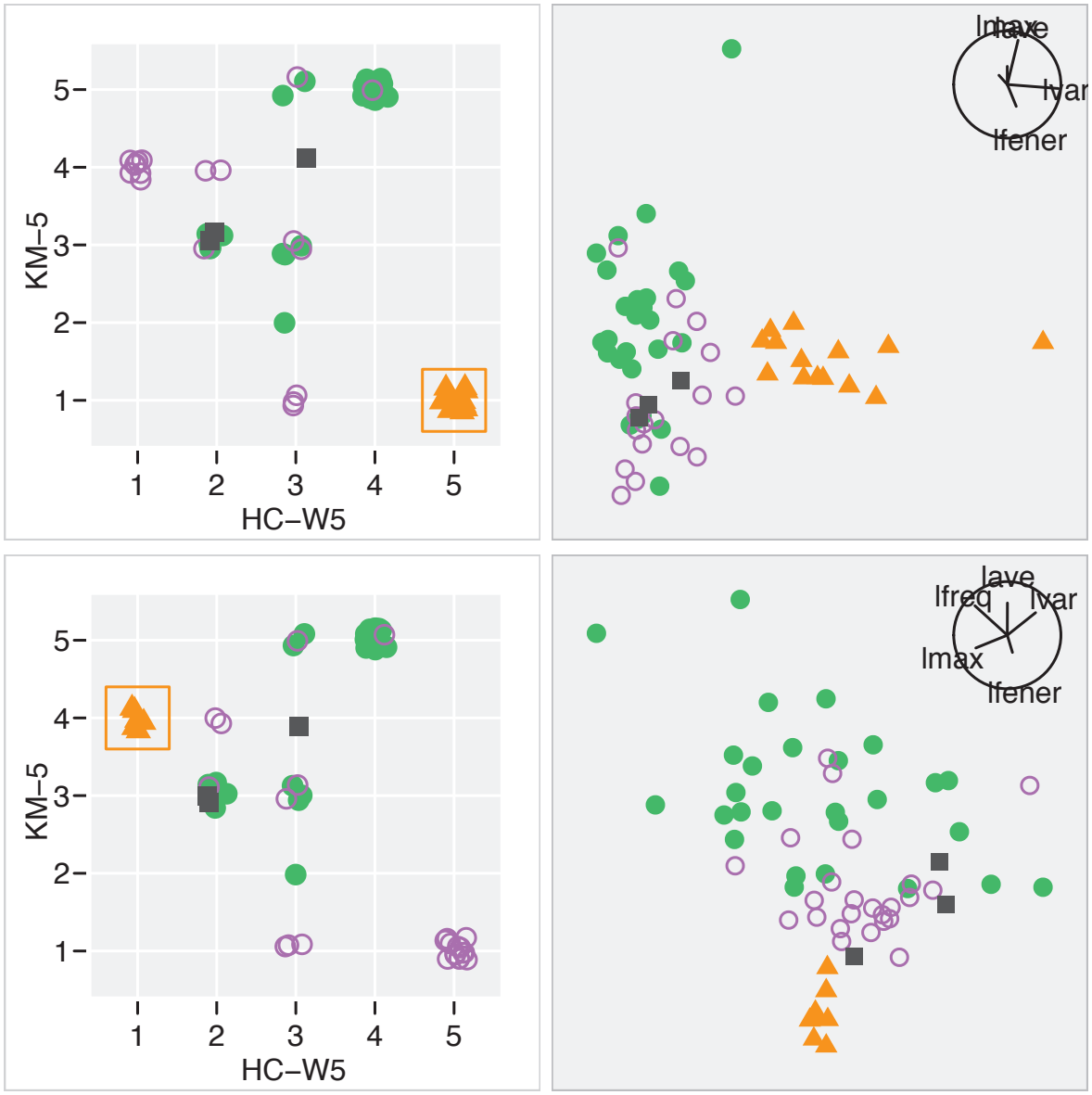
	Ward's				
k-means	1	2	3	4	5
1	0	0	3	0	14
2	0	0	1	0	0
3	0	9	5	0	0
4	8	2	1	0	0
5	0	0	3	16	0

Rearrange rows $\Rightarrow$

	Ward's				
k-means	1	2	3	4	5
4	8	2	1	0	0
3	0	9	5	0	0
2	0	0	1	0	0
5	0	0	3	16	0
1	0	0	3	0	14

Agree on 48/62 cases = 77%

...and graphically



Two different areas of agreement have been highlighted, and the tour projections show the tightness of each cluster where the methods agree.

Your turn

Run hierarchical clustering with average linkage on the Flea Beetles data (excluding species).

- Cut the tree at three clusters and append a cluster id to the dataset.
- How well do the clusters correspond to the species? (Plot cluster id vs species} and use jittering if necessary.)
- Using brushing in a plot of cluster id linked to a tour plot of the six variables, examine the beetles that are misclassified.
- Now cut the tree at four clusters, and repeat the last part.
- Which is the better solution, three or four clusters? Why?

Timeline

20 mins	Toolbox
30 mins	Missing values
45 mins	Supervised Classification
Break	
45 mins	Unsupervised Classification
30 mins	Inference