Some methods for longitudinal and cross-sectional visualization with further applications in the context of heat-maps

Shankar Srinivasan, Lihua Yue and Rick Soong

Statistics and Programming, BDOMWSP, Celgene

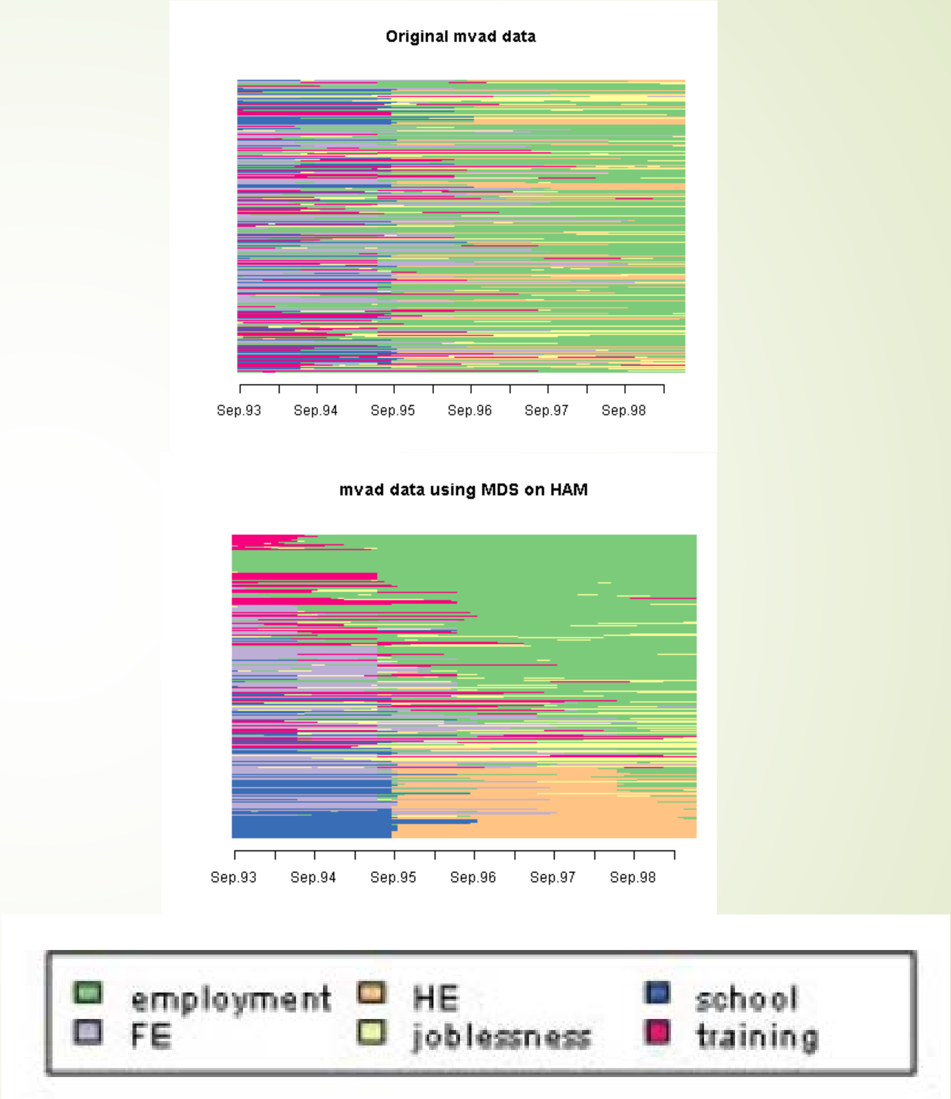
BASS Conference, Savannah, GA, 24th October 2017.

Motivation – Uncovering Latent Informative Images

Each horizontal strip in the plots to the right, represents for each subject, a sequence of states over time in the context of employment and education. The top graphic plots the unsorted raw data. The bottom graphic orders the data to bring out more clearly longitudinal as well as cross-sectional patterns in the data.

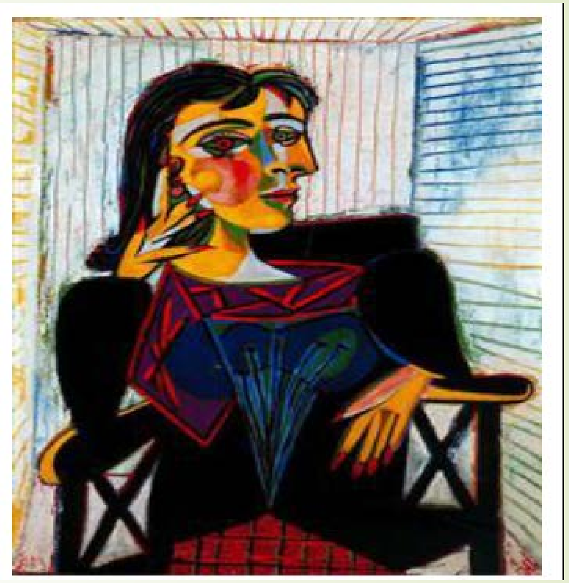
We seek to improve on the bottom graphic through an improved heuristic.

Clinical applications include subject transitions between responder categories over time while on cancer therapy and in gene expression heat-maps.



Gabadinho et al (2011). Mining sequence data in R with the TraMineR package: A user's guide.

Some Iconic Informative Images

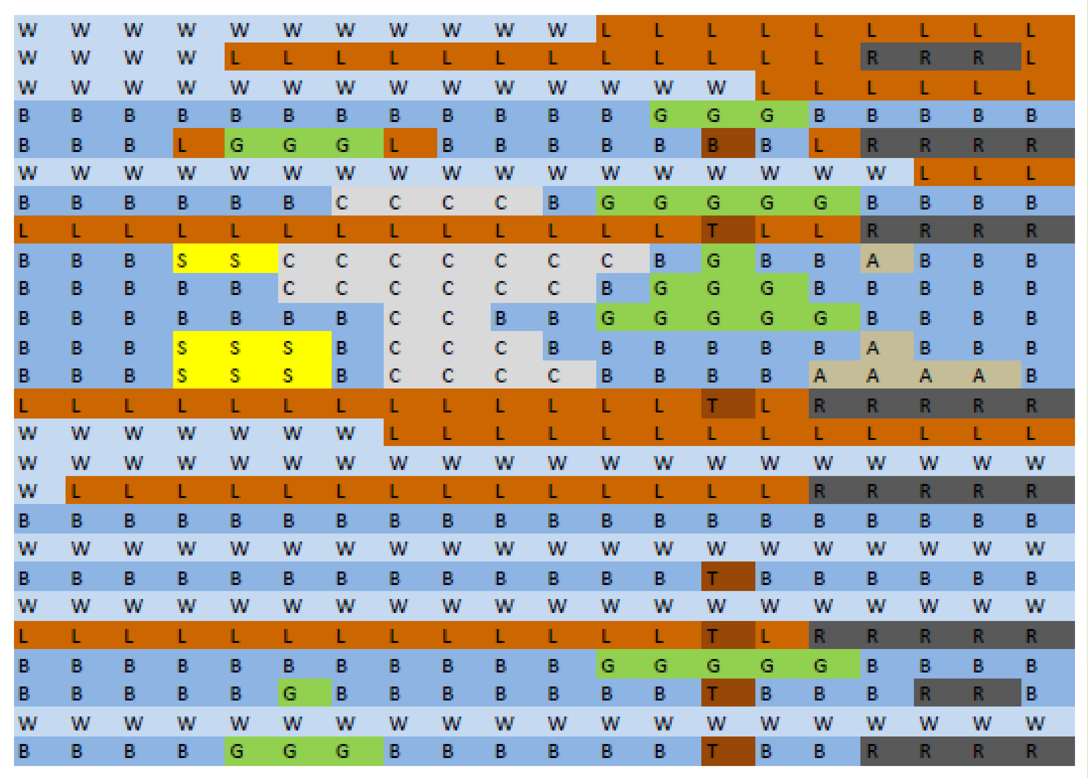


Useful Analogy – Rearranged 'TV Lines' Data

I have taken a picture and randomly rearranged all the horizontal rows of pixels in the graphic to the right.

Can the original picture be recovered from this random re-arrangement?

Can you see the original picture?

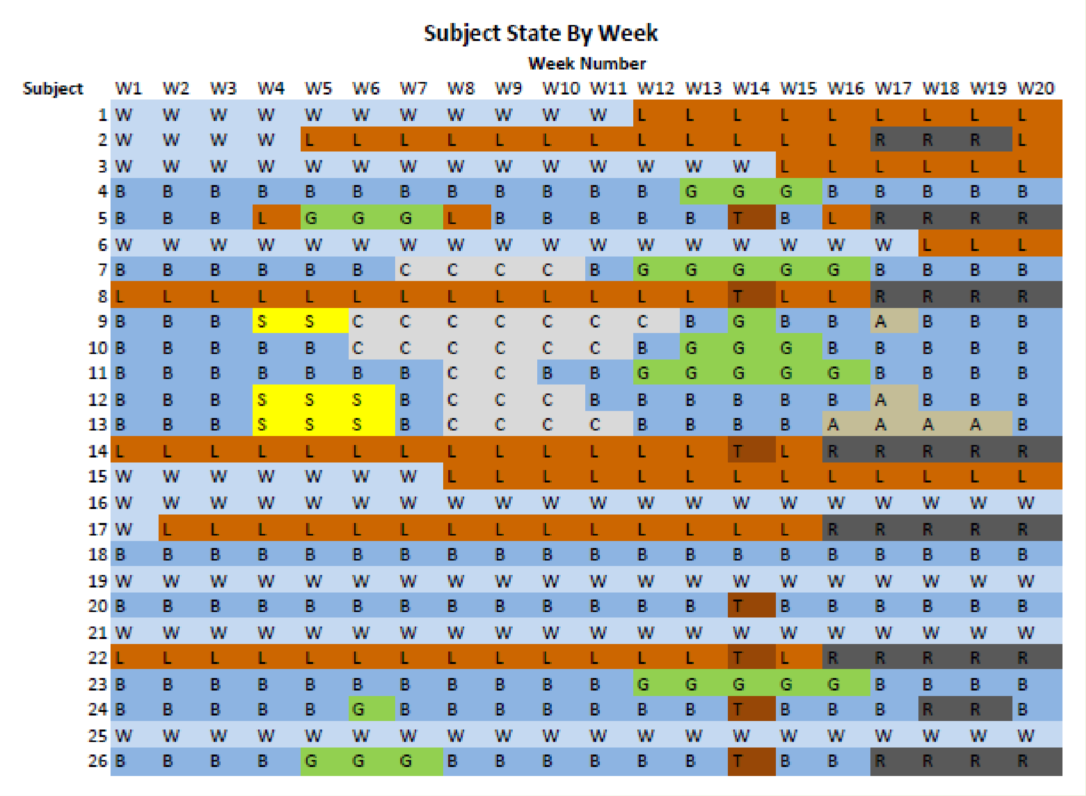


W	W	W	W	W	W	W	W	W	W	W	L	L	L	L	L	L	L	L	L
W	W	W	W	L	L	L	L	L	L	L	L	L	L	L	L	R	R	R	L
W	W	W	W	W	W	W	W	W	W	W	W	W	W	L	L	L	L	L	L
B	B	B	B	B	B	B	B	B	B	B	B	G	G	G	B	B	B	B	B
B	B	B	L	G	G	G	L	B	B	B	B	B	B	B	L	R	R	R	R
W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	L	L	L
B	B	B	B	B	B	C	C	C	C	B	G	G	G	G	G	B	B	B	B
L	L	L	L	L	L	L	L	L	L	L	L	L	T	L	L	R	R	R	R
B	B	B	S	S	C	C	C	C	C	C	C	B	G	B	B	A	B	B	B
B	B	B	B	B	C	C	C	C	C	C	B	G	G	G	B	B	B	B	B
B	B	B	B	B	B	B	C	C	B	B	G	G	G	G	G	B	B	B	B
B	B	B	S	S	S	B	C	C	C	B	B	B	B	B	B	A	B	B	B
B	B	B	S	S	S	B	C	C	C	C	B	B	B	B	A	A	A	A	B
L	L	L	L	L	L	L	L	L	L	L	L	L	T	L	R	R	R	R	R
W	W	W	W	W	W	W	L	L	L	L	L	L	L	L	L	L	L	L	L
W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W
W	L	L	L	L	L	L	L	L	L	L	L	L	L	L	R	R	R	R	R
B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B
W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W
B	B	B	B	B	B	B	B	B	B	B	B	B	T	B	B	B	B	B	B
W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W
L	L	L	L	L	L	L	L	L	L	L	L	L	T	L	R	R	R	R	R
B	B	B	B	B	B	B	B	B	B	B	G	G	G	G	G	B	B	B	B
B	B	B	B	B	G	B	B	B	B	B	B	B	T	B	B	B	R	R	B
W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W
B	B	B	B	G	G	G	B	B	B	B	B	B	T	B	B	R	R	R	R

Useful Analogy – Rearranged TV Lines Data

The randomly rearranged horizontal rows are similar to raw data on transitions between states over time for individual subjects. We can see that when we add row headers with subject numbers and column headers with week numbers.

If the original picture can be recovered from the random rearrangement, then is there a latent image hidden in raw transition data that can be recovered?



Subject State By Week

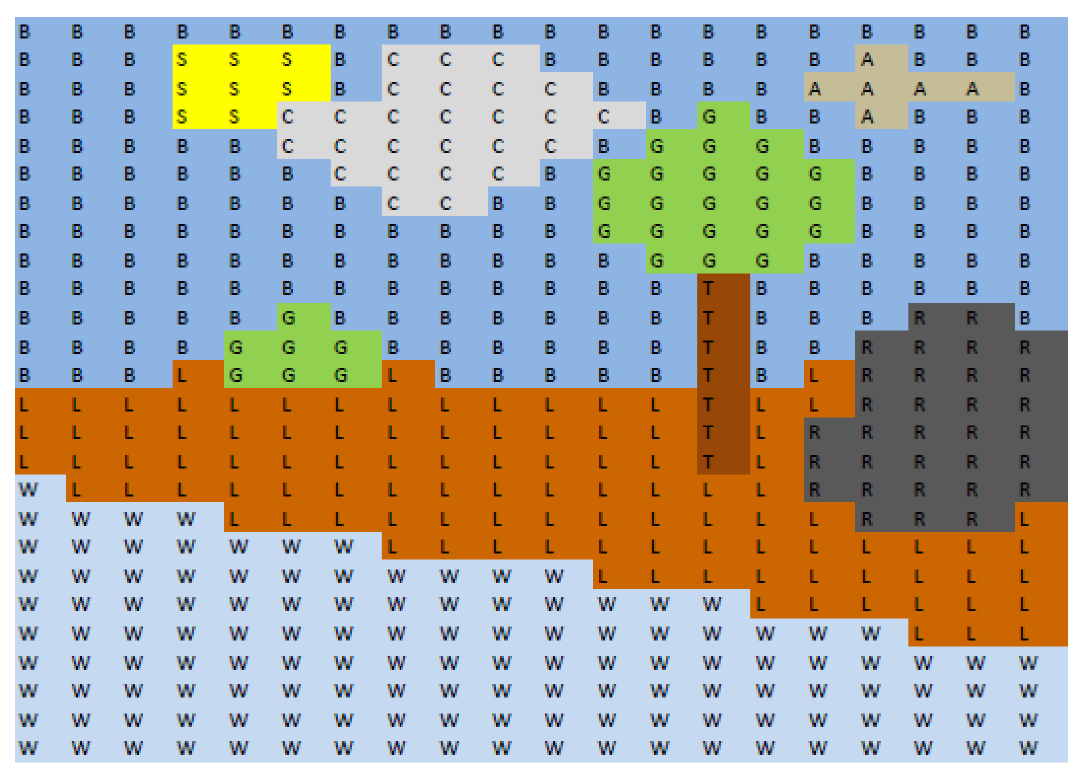
Week Number

Subject	W1	W2	W3	W4	W5	W6	W7	W8	W9	W10	W11	W12	W13	W14	W15	W16	W17	W18	W19	W20
1	W	W	W	W	W	W	W	W	W	W	W	L	L	L	L	L	L	L	L	L
2	W	W	W	W	L	L	L	L	L	L	L	L	L	L	L	L	R	R	R	L
3	W	W	W	W	W	W	W	W	W	W	W	W	W	W	L	L	L	L	L	L
4	B	B	B	B	B	B	B	B	B	B	B	B	G	G	G	B	B	B	B	B
5	B	B	B	L	G	G	G	L	B	B	B	B	B	T	B	L	R	R	R	R
6	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	L	L	L
7	B	B	B	B	B	B	C	C	C	C	B	G	G	G	G	G	B	B	B	B
8	L	L	L	L	L	L	L	L	L	L	L	L	L	T	L	L	R	R	R	R
9	B	B	B	S	S	C	C	C	C	C	C	C	B	G	B	B	A	B	B	B
10	B	B	B	B	B	C	C	C	C	C	C	B	G	G	G	B	B	B	B	B
11	B	B	B	B	B	B	B	C	C	B	B	G	G	G	G	G	B	B	B	B
12	B	B	B	S	S	S	B	C	C	C	B	B	B	B	B	B	A	B	B	B
13	B	B	B	S	S	S	B	C	C	C	C	B	B	B	B	A	A	A	A	B
14	L	L	L	L	L	L	L	L	L	L	L	L	L	T	L	R	R	R	R	R
15	W	W	W	W	W	W	W	L	L	L	L	L	L	L	L	L	L	L	L	L
16	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W
17	W	L	L	L	L	L	L	L	L	L	L	L	L	L	L	R	R	R	R	R
18	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B
19	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W
20	B	B	B	B	B	B	B	B	B	B	B	B	B	T	B	B	B	B	B	B
21	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W
22	L	L	L	L	L	L	L	L	L	L	L	L	L	T	L	R	R	R	R	R
23	B	B	B	B	B	B	B	B	B	B	B	G	G	G	G	G	B	B	B	B
24	B	B	B	B	B	G	B	B	B	B	B	B	B	T	B	B	B	R	R	B
25	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W
26	B	B	B	B	G	G	G	B	B	B	B	B	B	T	B	B	R	R	R	R

The Original Picture –
TV Lines Data

This is the original picture with the sky coded "B", the sun coded "S", the clouds coded "C", foliage coded "G", an airplane coded "A", a rock coded "R", land coded "L", trunk is coded "T" and a river coded "W".

We will go through a series of transformations of this picture to get the random row re-arranged picture in the last two slides.

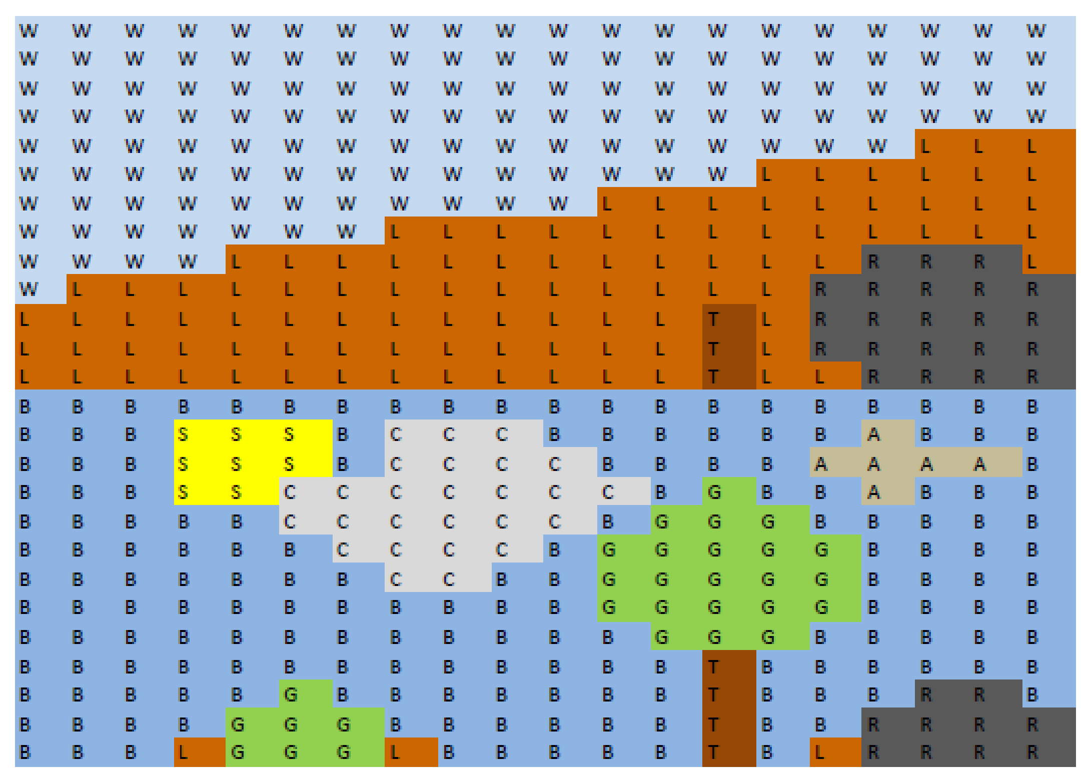


B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B
B	B	B	S	S	S	B	C	C	C	B	B	B	B	B	B	A	B	B	B
B	B	B	S	S	S	B	C	C	C	C	B	B	B	B	A	A	A	A	B
B	B	B	S	S	C	C	C	C	C	C	C	B	G	B	B	A	B	B	B
B	B	B	B	B	C	C	C	C	C	C	B	G	G	G	B	B	B	B	B
B	B	B	B	B	B	C	C	C	C	B	G	G	G	G	G	B	B	B	B
B	B	B	B	B	B	B	C	C	B	B	G	G	G	G	G	B	B	B	B
B	B	B	B	B	B	B	B	B	B	B	G	G	G	G	G	B	B	B	B
B	B	B	B	B	B	B	B	B	B	B	B	G	G	G	B	B	B	B	B
B	B	B	B	B	B	B	B	B	B	B	B	B	T	B	B	B	B	B	B
B	B	B	B	B	G	B	B	B	B	B	B	B	T	B	B	B	R	R	B
B	B	B	B	G	G	G	B	B	B	B	B	B	T	B	B	R	R	R	R
B	B	B	L	G	G	G	L	B	B	B	B	B	T	B	L	R	R	R	R
L	L	L	L	L	L	L	L	L	L	L	L	L	T	L	L	R	R	R	R
L	L	L	L	L	L	L	L	L	L	L	L	L	T	L	R	R	R	R	R
L	L	L	L	L	L	L	L	L	L	L	L	L	T	L	R	R	R	R	R
W	L	L	L	L	L	L	L	L	L	L	L	L	L	L	R	R	R	R	R
W	W	W	W	L	L	L	L	L	L	L	L	L	L	L	L	R	R	R	L
W	W	W	W	W	W	W	L	L	L	L	L	L	L	L	L	L	L	L	L
W	W	W	W	W	W	W	W	W	W	W	L	L	L	L	L	L	L	L	L
W	W	W	W	W	W	W	W	W	W	W	W	W	W	L	L	L	L	L	L
W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	L	L	L
W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W
W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W
W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W
W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W

Initial re-arrangement

Here the picture is still recognizable.

The top half of the original picture has been shifted down to the bottom and bottom half has been moved to the top and inverted.

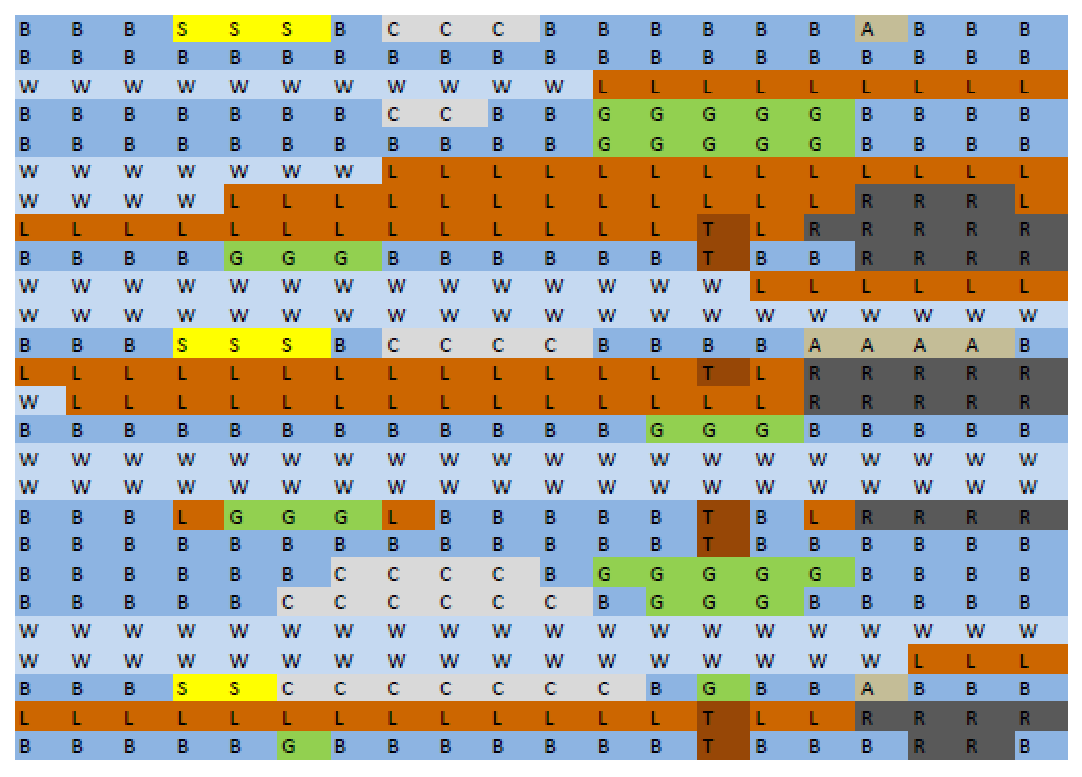


W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W
W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W
W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W
W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W
W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	L	L	L
W	W	W	W	W	W	W	W	W	W	W	W	W	W	L	L	L	L	L	L
W	W	W	W	W	W	W	W	W	W	W	L	L	L	L	L	L	L	L	L
W	W	W	W	W	W	W	L	L	L	L	L	L	L	L	L	L	L	L	L
W	W	W	W	L	L	L	L	L	L	L	L	L	L	L	L	R	R	R	L
W	L	L	L	L	L	L	L	L	L	L	L	L	L	L	R	R	R	R	R
L	L	L	L	L	L	L	L	L	L	L	L	L	T	L	R	R	R	R	R
L	L	L	L	L	L	L	L	L	L	L	L	L	T	L	R	R	R	R	R
L	L	L	L	L	L	L	L	L	L	L	L	L	T	L	L	R	R	R	R
B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B
B	B	B	S	S	S	B	C	C	C	B	B	B	B	B	B	A	B	B	B
B	B	B	S	S	S	B	C	C	C	C	B	B	B	B	A	A	A	A	B
B	B	B	S	S	C	C	C	C	C	C	C	B	G	B	B	A	B	B	B
B	B	B	B	B	C	C	C	C	C	C	B	G	G	G	B	B	B	B	B
B	B	B	B	B	B	C	C	C	C	B	G	G	G	G	G	B	B	B	B
B	B	B	B	B	B	B	C	C	B	B	G	G	G	G	G	B	B	B	B
B	B	B	B	B	B	B	B	B	B	B	G	G	G	G	G	B	B	B	B
B	B	B	B	B	B	B	B	B	B	B	B	G	G	G	B	B	B	B	B
B	B	B	B	B	B	B	B	B	B	B	B	B	T	B	B	B	B	B	B
B	B	B	B	B	G	B	B	B	B	B	B	B	T	B	B	B	R	R	B
B	B	B	B	G	G	G	B	B	B	B	B	B	T	B	B	R	R	R	R
B	B	B	L	G	G	G	L	B	B	B	B	B	T	B	L	R	R	R	R

Further splitting, re-shuffling and inverting.

The picture is becoming difficult to recognize.

We used the steps of splitting, shuffling and inverting to go from the original picture to this one which seems like a random re-arrangement we might see in the raw subject state sequence data.

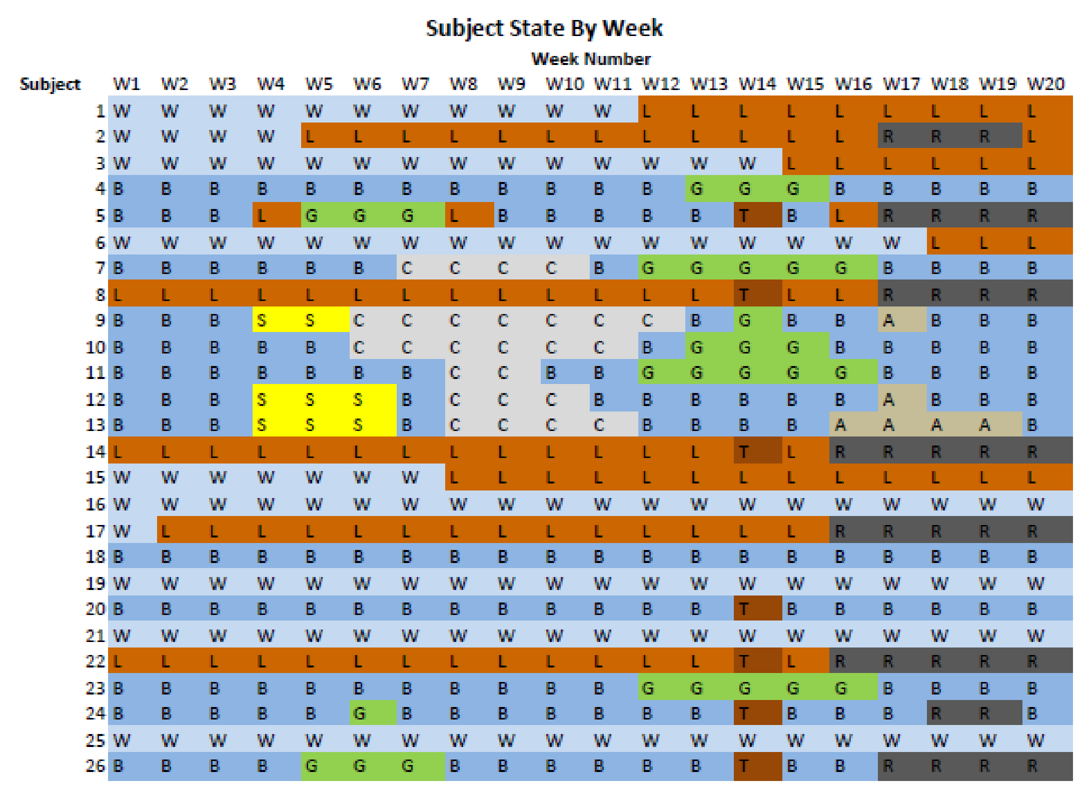


B	B	B	S	S	S	B	C	C	C	B	B	B	B	B	B	A	B	B	B
B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B
W	W	W	W	W	W	W	W	W	W	W	L	L	L	L	L	L	L	L	L
B	B	B	B	B	B	B	C	C	B	B	G	G	G	G	G	B	B	B	B
B	B	B	B	B	B	B	B	B	B	B	G	G	G	G	G	B	B	B	B
W	W	W	W	W	W	W	L	L	L	L	L	L	L	L	L	L	L	L	L
W	W	W	W	L	L	L	L	L	L	L	L	L	L	L	L	R	R	R	L
L	L	L	L	L	L	L	L	L	L	L	L	L	T	L	R	R	R	R	R
B	B	B	B	G	G	G	B	B	B	B	B	B	T	B	B	R	R	R	R
W	W	W	W	W	W	W	W	W	W	W	W	W	W	L	L	L	L	L	L
W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W
B	B	B	S	S	S	B	C	C	C	C	B	B	B	B	A	A	A	A	B
L	L	L	L	L	L	L	L	L	L	L	L	L	T	L	R	R	R	R	R
W	L	L	L	L	L	L	L	L	L	L	L	L	L	L	R	R	R	R	R
B	B	B	B	B	B	B	B	B	B	B	B	G	G	G	B	B	B	B	B
W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W
W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W
B	B	B	L	G	G	G	L	B	B	B	B	B	T	B	L	R	R	R	R
B	B	B	B	B	B	B	B	B	B	B	B	B	T	B	B	B	B	B	B
B	B	B	B	B	B	C	C	C	C	B	G	G	G	G	G	B	B	B	B
B	B	B	B	B	C	C	C	C	C	C	B	G	G	G	B	B	B	B	B
W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W
W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	L	L	L
B	B	B	S	S	C	C	C	C	C	C	C	B	G	B	B	A	B	B	B
L	L	L	L	L	L	L	L	L	L	L	L	L	T	L	L	R	R	R	R
B	B	B	B	B	G	B	B	B	B	B	B	B	T	B	B	B	R	R	B

Final re-arrangement - Rearranged TV Lines Data

Reversing what we have been doing, note that to obtain the original picture we look at a match of each subject's horizontal panel with all other panels as the first step. Find the closest matching pair and join them. Then this pair of panels (or remaining singlets) should be joined with another panel with the best match, We proceed similarly to combine the horizontal panels.

Matches should consider the best joining surface between existing composite or singlet panels. We continue agglomerating panels till we combine all horizontal panels into one picture. **We call this ordering process edge clustering.**

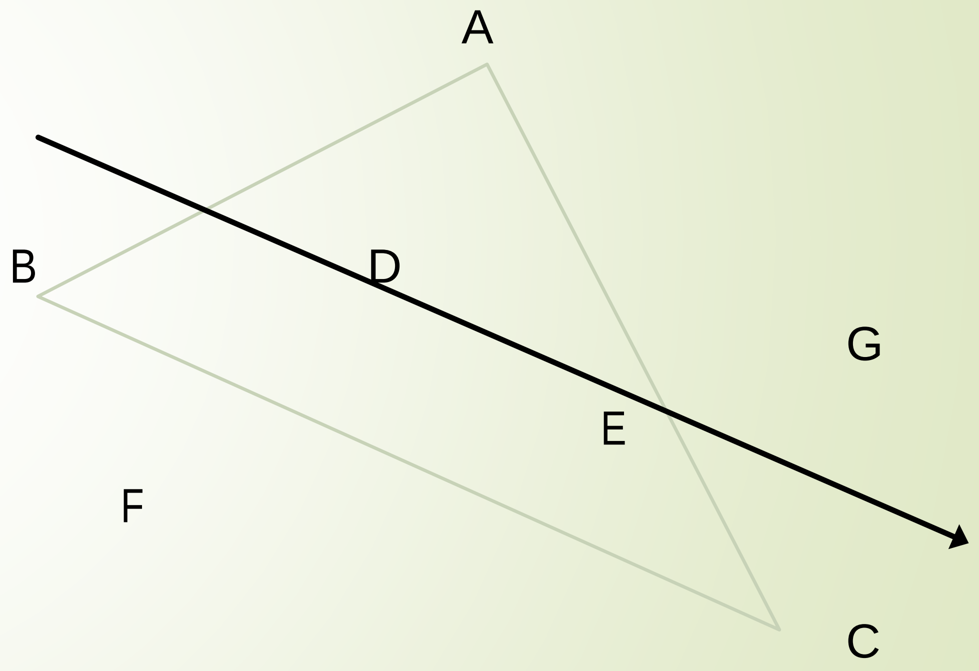


Subject State By Week

	Week Number																			
Subject	W1	W2	W3	W4	W5	W6	W7	W8	W9	W10	W11	W12	W13	W14	W15	W16	W17	W18	W19	W20
1	W	W	W	W	W	W	W	W	W	W	W	L	L	L	L	L	L	L	L	L
2	W	W	W	W	L	L	L	L	L	L	L	L	L	L	L	L	R	R	R	L
3	W	W	W	W	W	W	W	W	W	W	W	W	W	W	L	L	L	L	L	L
4	B	B	B	B	B	B	B	B	B	B	B	B	G	G	G	B	B	B	B	B
5	B	B	B	L	G	G	G	L	B	B	B	B	B	T	B	L	R	R	R	R
6	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	L	L	L
7	B	B	B	B	B	B	C	C	C	C	B	G	G	G	G	G	B	B	B	B
8	L	L	L	L	L	L	L	L	L	L	L	L	L	T	L	L	R	R	R	R
9	B	B	B	S	S	C	C	C	C	C	C	C	B	G	B	B	A	B	B	B
10	B	B	B	B	B	C	C	C	C	C	C	B	G	G	G	B	B	B	B	B
11	B	B	B	B	B	B	B	C	C	B	B	G	G	G	G	G	B	B	B	B
12	B	B	B	S	S	S	B	C	C	C	B	B	B	B	B	B	A	B	B	B
13	B	B	B	S	S	S	B	C	C	C	C	B	B	B	B	A	A	A	A	B
14	L	L	L	L	L	L	L	L	L	L	L	L	L	T	L	R	R	R	R	R
15	W	W	W	W	W	W	W	L	L	L	L	L	L	L	L	L	L	L	L	L
16	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W
17	W	L	L	L	L	L	L	L	L	L	L	L	L	L	L	R	R	R	R	R
18	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B
19	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W
20	B	B	B	B	B	B	B	B	B	B	B	B	B	T	B	B	B	B	B	B
21	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W
22	L	L	L	L	L	L	L	L	L	L	L	L	L	T	L	R	R	R	R	R
23	B	B	B	B	B	B	B	B	B	B	B	G	G	G	G	G	B	B	B	B
24	B	B	B	B	B	G	B	B	B	B	B	B	B	T	B	B	B	R	R	B
25	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W
26	B	B	B	B	G	G	G	B	B	B	B	B	B	T	B	B	R	R	R	R

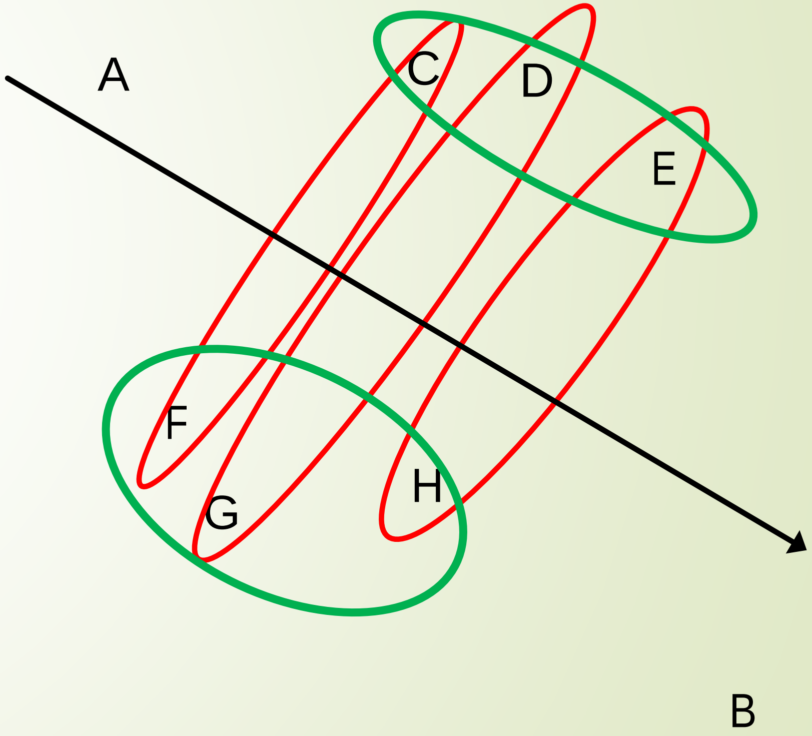
Ordering Methods: Multidimensional Scaling using First Extracted Dimension (MDS, k =1)

- The Ordering of the sequences starts by ordering the sequences based on distance or similarity measures between pairs of sequences.
- MDS resolves distances by adding dimensions.
- For example if the distance between sequence A and B is 3 units, A to C is 4 units and B to C is 5 units, then these distances cannot be resolved in one dimension.
- IF D, E, F, G are additional points, then
- .. one of way of ordering the data is projections onto the black line. We refer to this as MDS with k = 1.



Clustering versus MDS with k=1

- MDS with k= 1 places points based on their projections on the most informative first dimension.
- So in the MDS (k=1) ordering, C pairs with F, D with G and E with H- see red ovals.
- Clustering joins C, D and E and F,G and H – see green ovals.



Ordering Methods: Hierarchical Clustering

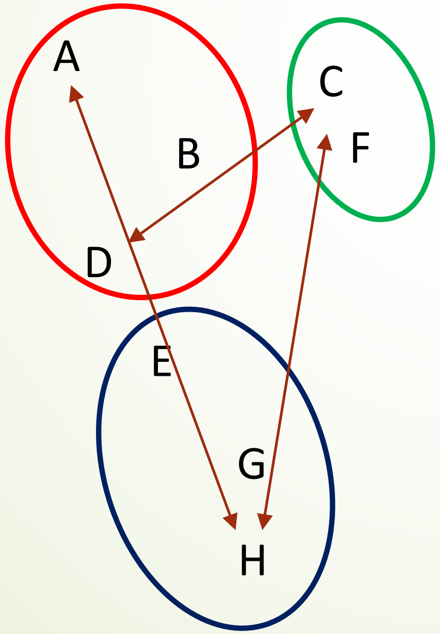
- Hierarchical clustering uses a distance measure and a clustering rule to form clusters.
- Starts with each single element as a cluster. The two closest elements form a cluster. All distances are reassessed between elements, doublet or remaining singlets, and the closest distance leads to a new cluster and the process continues till we combine all the data into one cluster.
- Popular types of hierarchical clustering – single linkage, average linkage and complete linkage.
- We hope to add a new hierarchical clustering tool called edge clustering to the lexicon.

Step 1 (All Singlets)	Step 2 (1 Doublet, Rest Singlets)	Step 3 (2 Doublets, Rest Singlets)	Step 4 (1 Triplet, 1 Doublets, Rest Singlets)
S1	D1{S1&S3}	D1{S1&S3}	D1{S1&S3}
S2	S2	D2{S2&S7}	T1{S2&S7&S6}
S3			
S4	S4	S4	S4
S5	S5	S5	S5
S6	S6	S6	
S7	S7		

Hierarchical Clustering Methods – Cluster Agglomerating Rules.

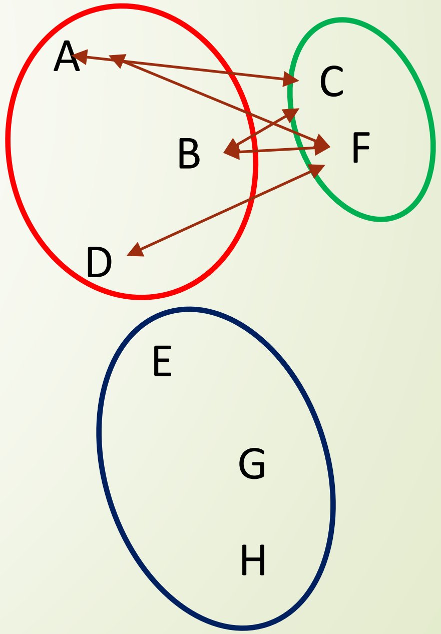
Complete Linkage – Based on the minimum maximum distance.

$$\max\{d(a,b) : a \in A, b \in B\}$$



Average Linkage –based on distances between all pairs of elements.

$$(1/n_A n_B)\sum_{a\in A}\sum_{b\in B} d(a,b)$$

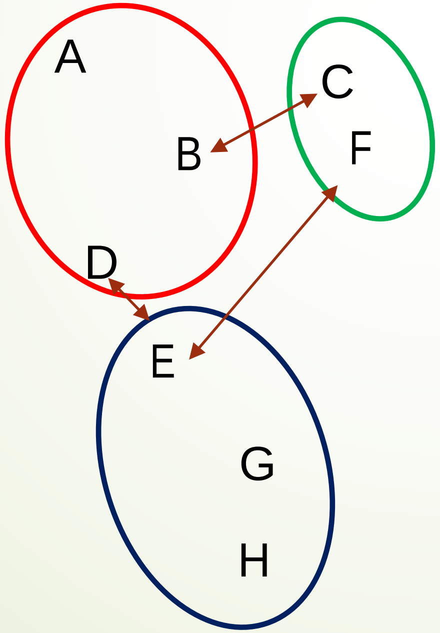


Ward Method - Ward clustering minimizes variances to obtain tight spherical clusters.

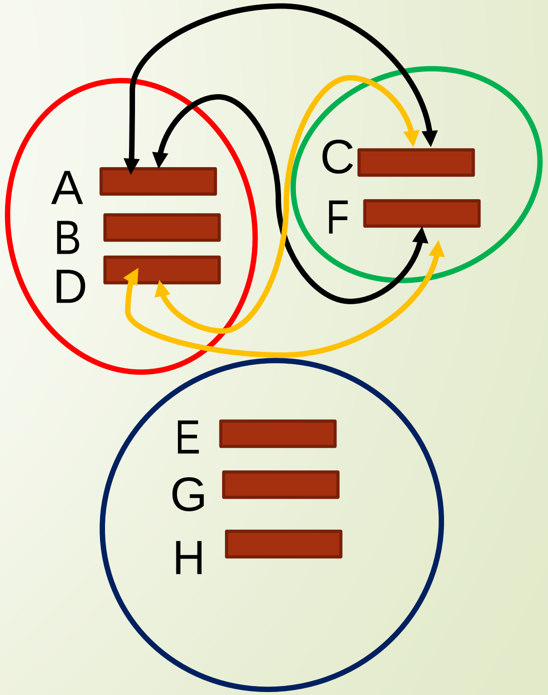
Hierarchical Clustering Methods – Cluster Agglomerating Rules.

Single Linkage – Based on a minimum distance.

$$\min\{d(a,b) : a \in A, b \in B\}$$

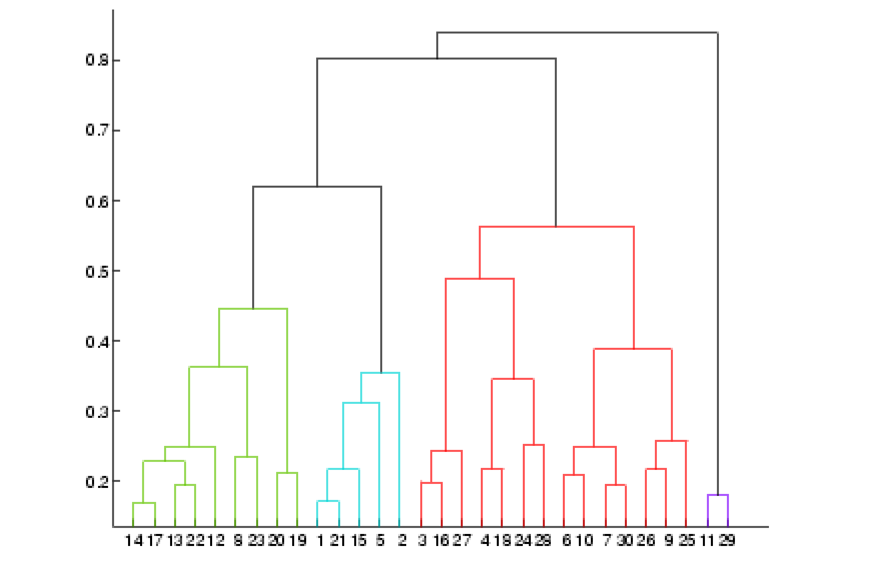


Edge Clustering –based on comparisons of elements at edges of ordered clusters.



A Tweak to Hierarchical Clustering Methods

- The single, complete and average linkage methods, as conventionally presented, do not provide a unique ordering of clusters or of elements within clusters. The emphasis is on classification. Any combination of the staples in the dendrogram to the right could be reversed to yield a new ordering.
- Sakai et. al (2014) obtain an ordering by sorting, as a separate step, the dendrogram obtained from classificatory hierarchical clustering methods.



Sakai et al. dendsort: modular leaf ordering methods for dendrograms representations in R. F1000Research 2014, 3:177

Distance and Similarity Measures

- A number of similarity measures are derived based on the number of operations which transform one sequence into another.
- Distance measure can be obtained from similarity measures by using any monotonic decreasing function.
- The operations considered are substitution, insertion and deletion and typically have associated costs. Popular measures are:
    - The Hamming (HAM) measure involves substitutions alone.
    - The Optimal Matching (OM) measure involves substitution as Insertions and Deletions.

S1	A	A	C	B	C		
S2	A	C	B	B	B		
3 Substitutions							
S1	A	A	C	B	C		
		S	S		S		
S2	A	C	B	B	B		
2 Insertions and 2 Deletions							
S1	A	A	C	B	C		
	D				D	I	I
S2?	-	A	C	B	-	B	B

Examples from: Gabadinho et al. Workshop on sequence analysis. New York, October 11, 2013.

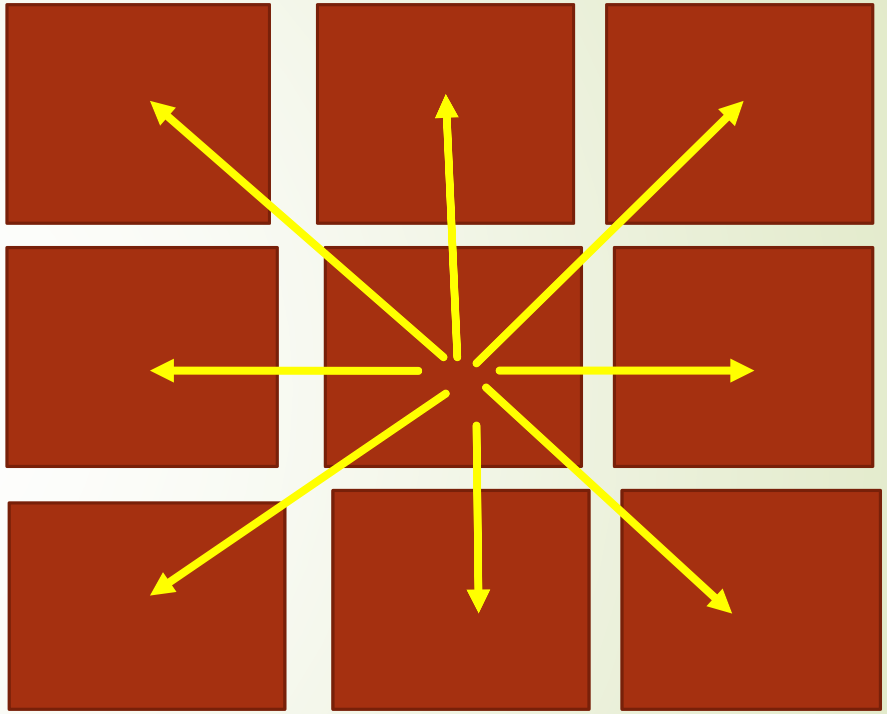
A New Measures we tried: Cell to Cell Matching

We derive a match score based on the premise that every pixel should match with as many of the 8 pixels that surround it as is possible.

We will be comparing a pair of rows at a time and so we could consider the arrows and the pixels in say the bottom two rows.

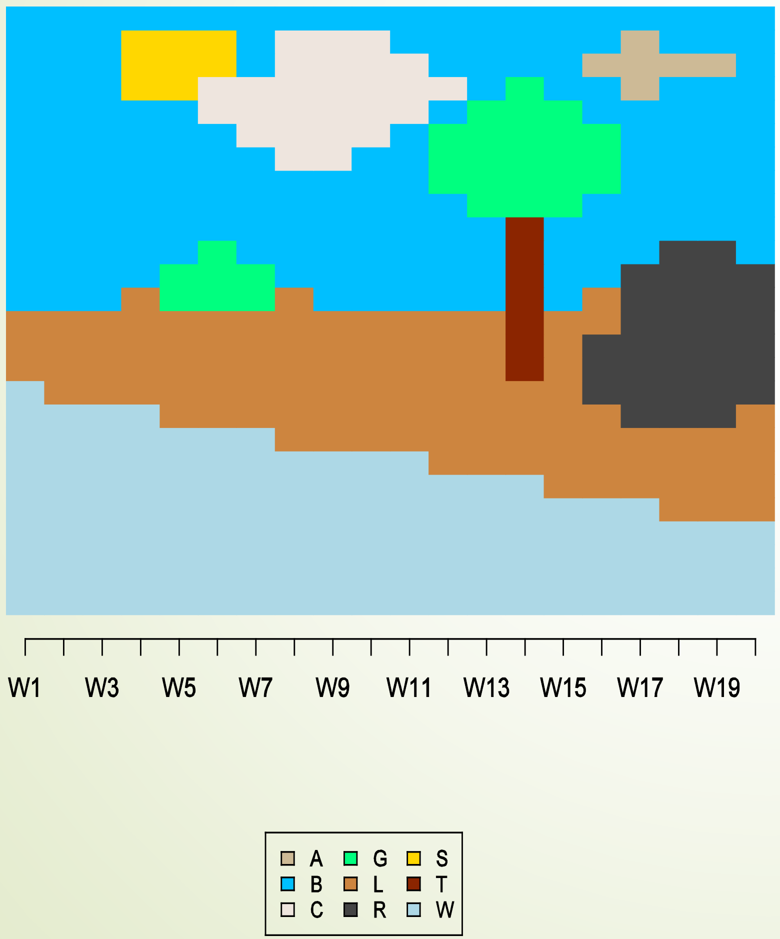
The horizontal matches will contribute equally to the match score of a pixel in matches with any other row compared pair-wise. Hence can be ignored.

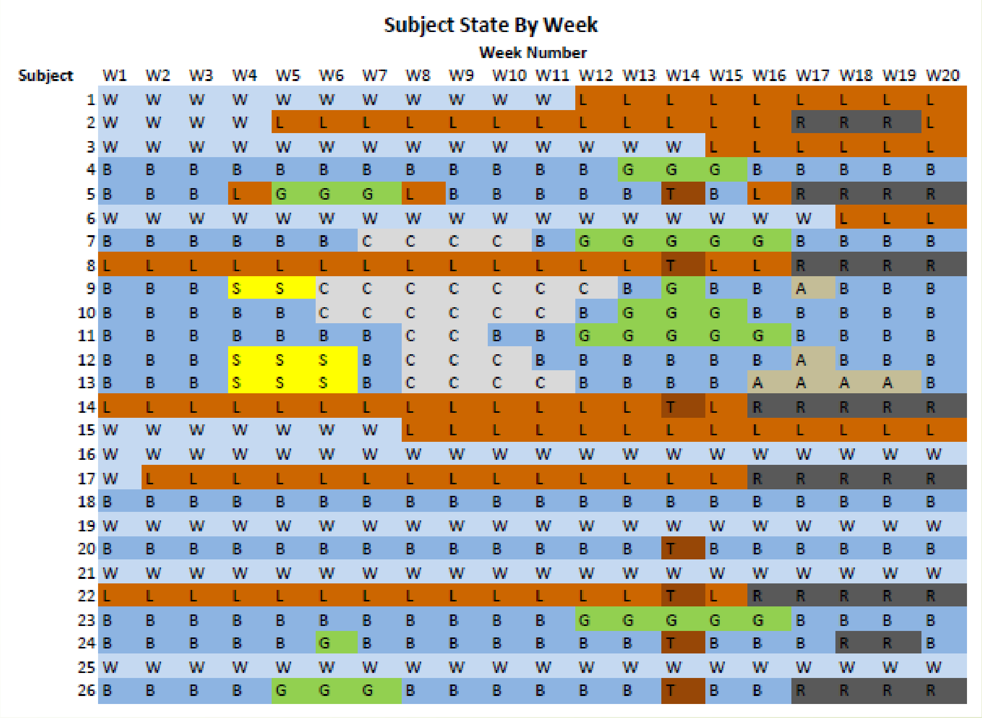
The only matches that need to be considered are those represented by the down arrows from the target cell or pixel.



Non Ordinal Data: Original “TV lines” data (left) and re-arranged data (right)

TV Line Original Data





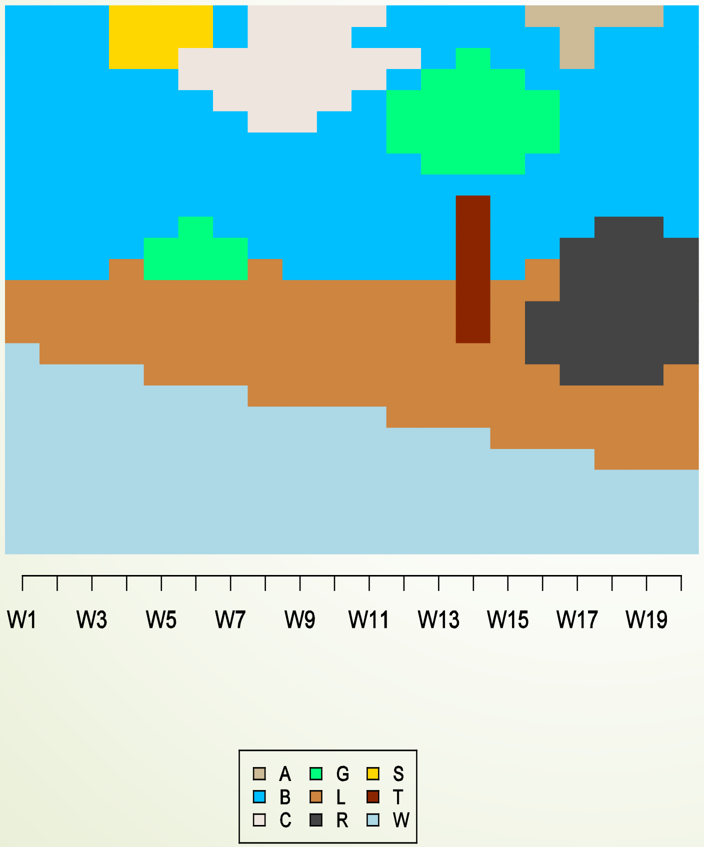
Subject State By Week

Week Number

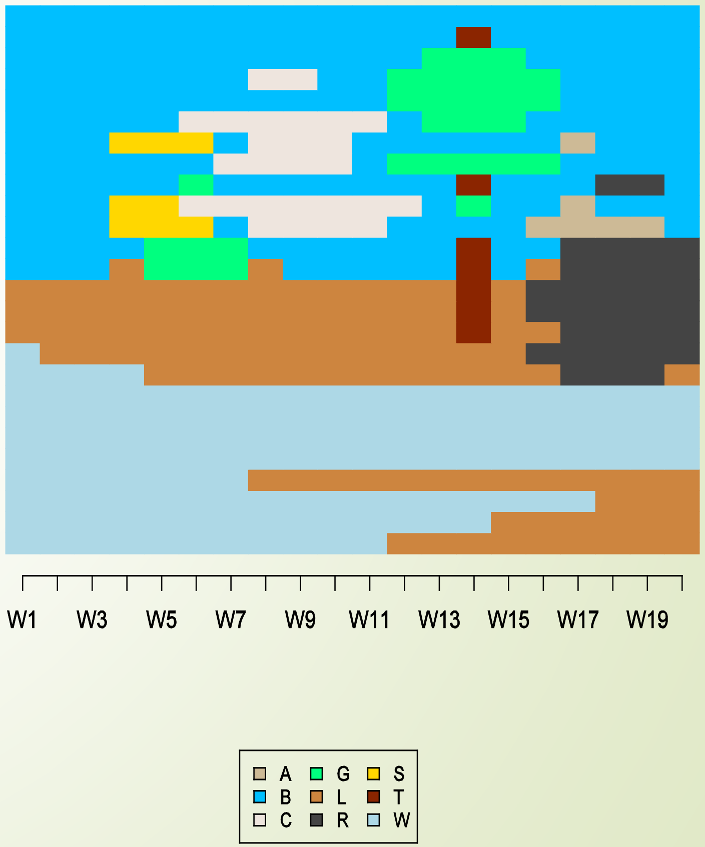
Subject	W1	W2	W3	W4	W5	W6	W7	W8	W9	W10	W11	W12	W13	W14	W15	W16	W17	W18	W19	W20
1	W	W	W	W	W	W	W	W	W	W	W	L	L	L	L	L	L	L	L	L
2	W	W	W	W	L	L	L	L	L	L	L	L	L	L	L	L	R	R	R	L
3	W	W	W	W	W	W	W	W	W	W	W	W	W	W	L	L	L	L	L	L
4	B	B	B	B	B	B	B	B	B	B	B	B	G	G	G	B	B	B	B	B
5	B	B	B	L	G	G	G	L	B	B	B	B	B	T	B	L	R	R	R	R
6	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	L	L	L
7	B	B	B	B	B	B	C	C	C	C	B	G	G	G	G	G	B	B	B	B
8	L	L	L	L	L	L	L	L	L	L	L	L	L	T	L	L	R	R	R	R
9	B	B	B	S	S	C	C	C	C	C	C	C	B	G	B	B	A	B	B	B
10	B	B	B	B	B	C	C	C	C	C	C	B	G	G	G	B	B	B	B	B
11	B	B	B	B	B	B	B	C	C	B	B	G	G	G	G	G	B	B	B	B
12	B	B	B	S	S	S	B	C	C	C	B	B	B	B	B	B	A	B	B	B
13	B	B	B	S	S	S	B	C	C	C	C	B	B	B	B	A	A	A	A	B
14	L	L	L	L	L	L	L	L	L	L	L	L	L	T	L	R	R	R	R	R
15	W	W	W	W	W	W	W	L	L	L	L	L	L	L	L	L	L	L	L	L
16	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W
17	W	L	L	L	L	L	L	L	L	L	L	L	L	L	L	R	R	R	R	R
18	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B	B
19	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W
20	B	B	B	B	B	B	B	B	B	B	B	B	B	T	B	B	B	B	B	B
21	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W
22	L	L	L	L	L	L	L	L	L	L	L	L	L	T	L	R	R	R	R	R
23	B	B	B	B	B	B	B	B	B	B	B	G	G	G	G	G	B	B	B	B
24	B	B	B	B	B	G	B	B	B	B	B	B	B	T	B	B	B	R	R	B
25	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W	W
26	B	B	B	B	G	G	G	B	B	B	B	B	B	T	B	B	R	R	R	R

Recovered Images: Using Edge Clustering (left) and using Multi-Dimensional Scaling (right)

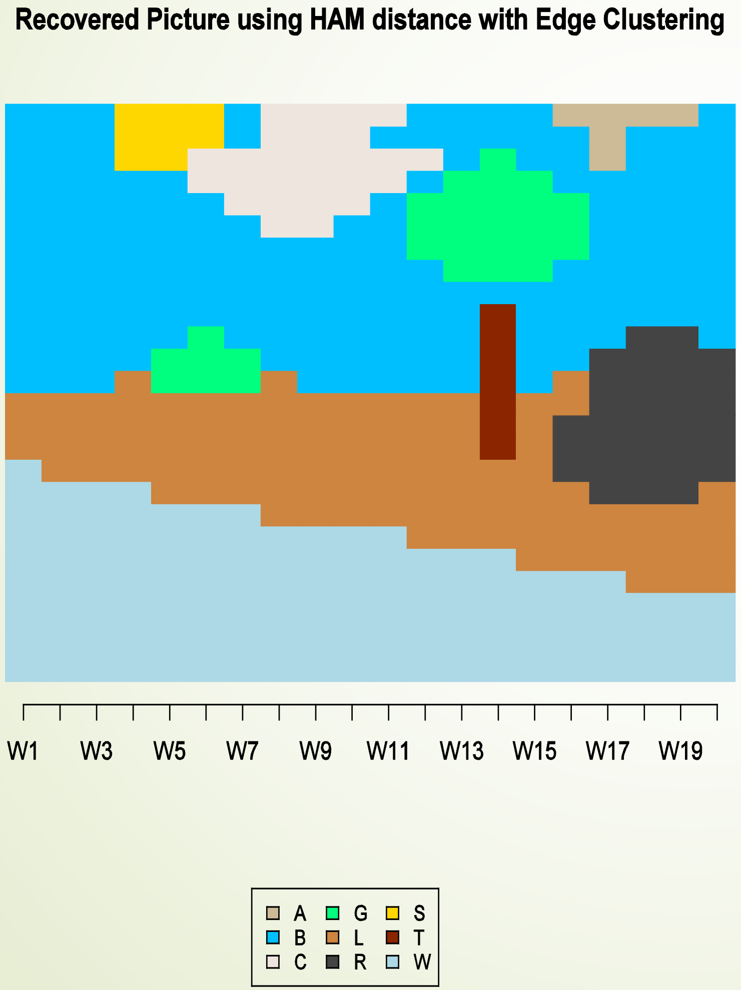
Recovered Picture using HAM distance with Edge Clustering

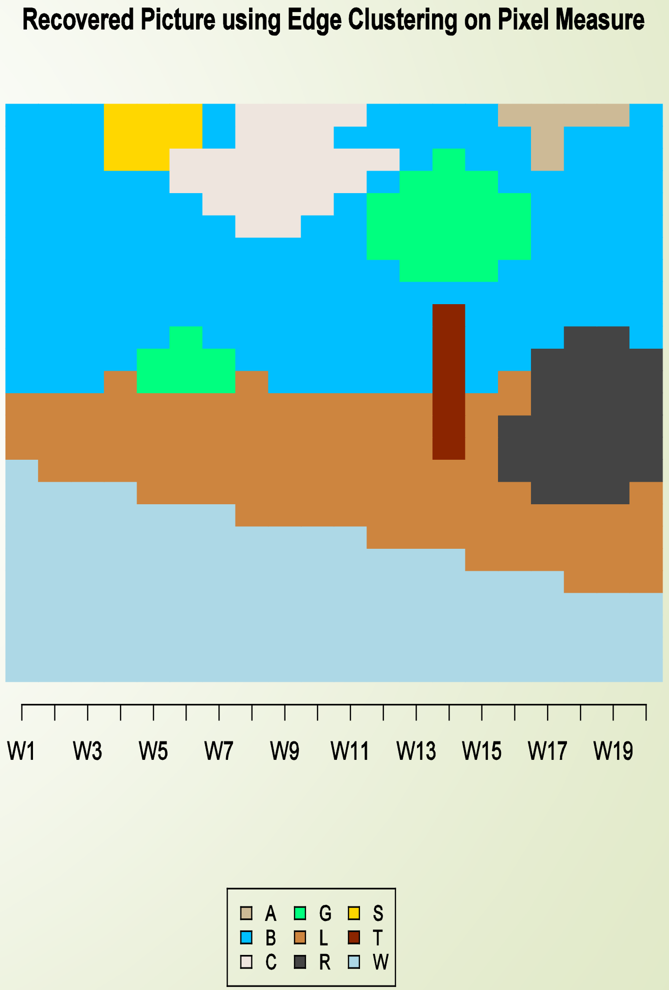


Recovered Picture using MDS on HAM



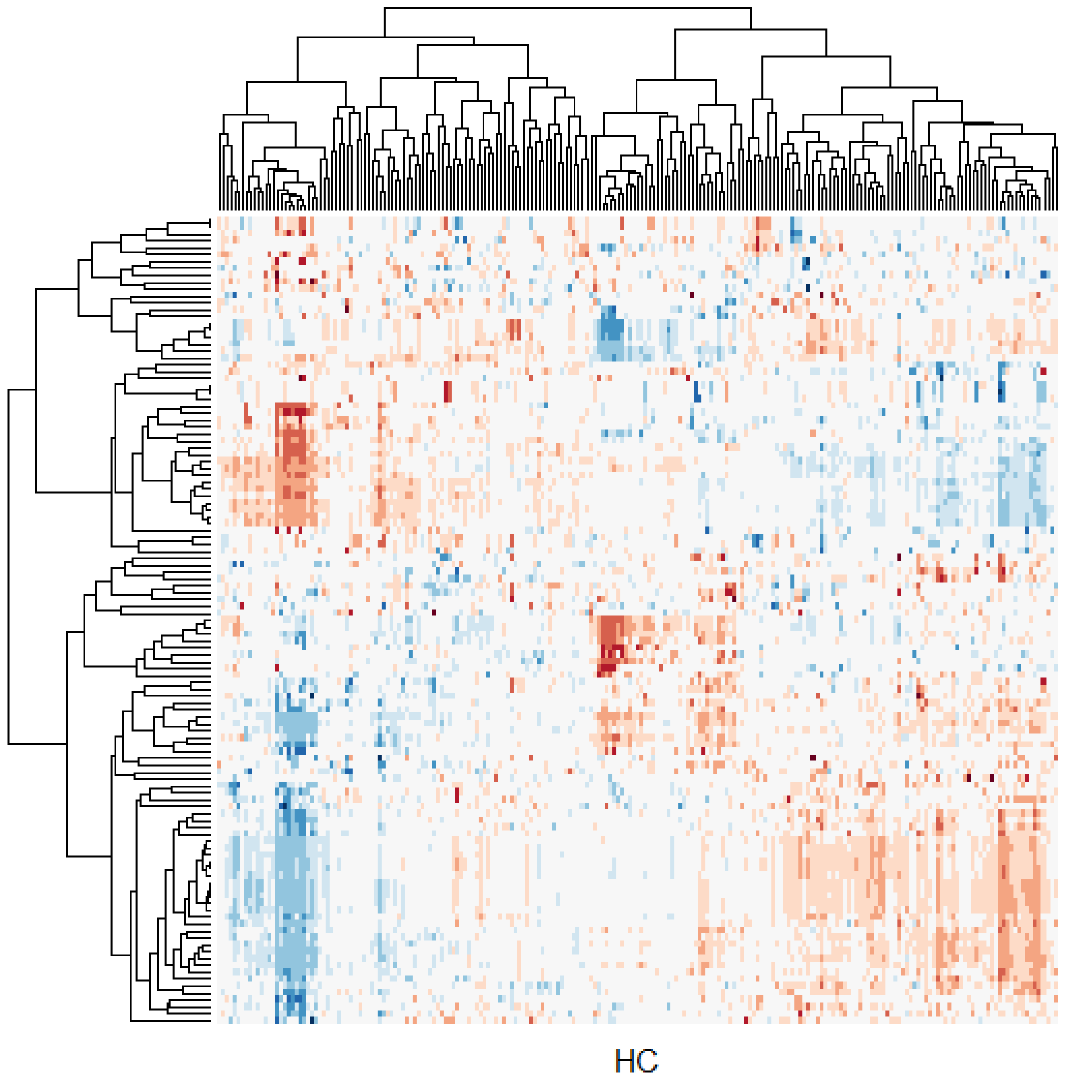
Edge Clustering with different similarity measures: HAM Distance (left) and Pixel Distance (right) – similarity measures may have less impact then the row ordering heuristic





Additional Edge Clustering Applications: two way clustering

- Compute distance measures between every pair of rows (samples) in a dataset.
- Use the row to row distance measures to order the rows using edge clustering or other sorted hierarchical clustering.
- Similarly order the columns and plot the row/column ordered data.
- To left is a Complete (unsorted) heatmap of TCGA data (courtesy Sakia et al 2104). Scaled association between genes (columns) and samples (rows) are plotted.



Some Distance Measures for Continuous Data

- Euclidean Distance. The distance between two strings of numeric data is given by

$$\sqrt{\sum_i (a_i - b_i)^2}$$

- Manhattan Distance

$$\sum_i |a_i - b_i|$$

- Maximum Distance

$$\max_i |a_i - b_i|$$

- Mahalanobis Distance

$$\sqrt{(a-b)^T S^{-1} (a-b)}$$

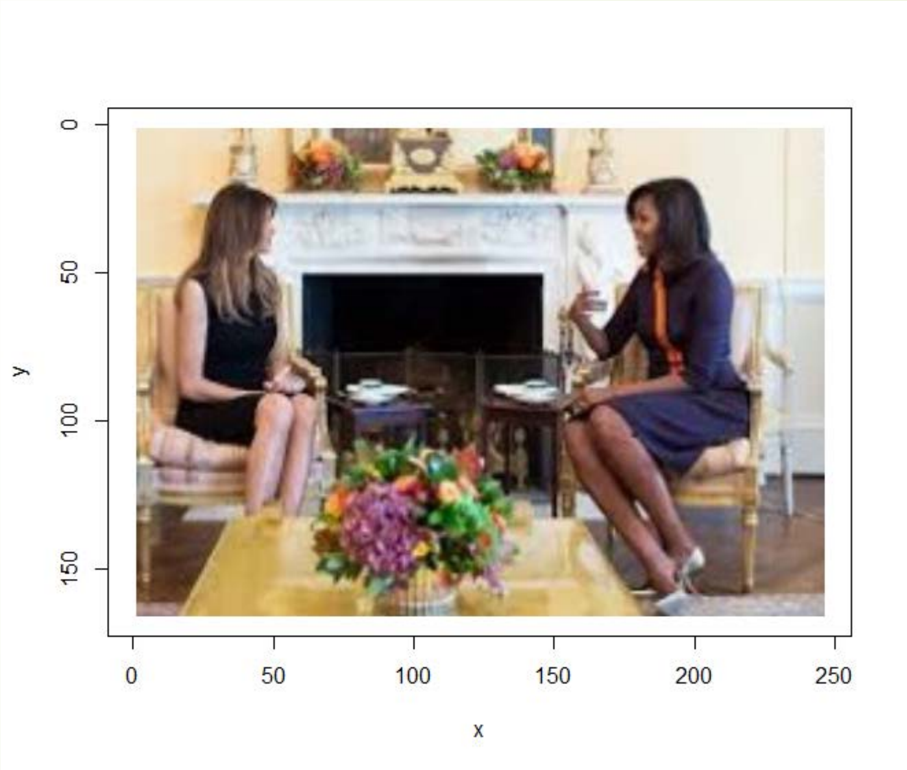
Testing Our Ordering Heuristics

- What is the latent image in a gene expression heat-map? How should a heat map look? -do we know the 'parameter' we are estimating?
- Subjective assessments of the rightness of an assessed heat map could be inaccurate.
- As we did before we need contexts where we know the 'parameter' (our latent informative image).
- For such a known image we would randomly permute the row strings of pixel values and then the column vectors of pixel values. This would preserve all information while removing all ordering. Then we can check to see how well our row and column ordering heuristics help us uncover our known latent image.
- In such a random permutation in gene expression data, we might for example permute Sample D in row 8 and Gene Y in column 26 in what might be a 'right' heat-map to say row 61 and column 17. The normalized gene expression value corresponding to Sample D and Gene Y of say 1.73 would now be in cell {61,17} instead of cell {8,26} in what is likely a very non-informative heat-map.

A latent image we could use

- To the right is an informative image.
- It contains data – the x pixel co-ordinate, the y pixel co-ordinate and three numeric values for the intensity of the red, blue and green colors at that co-ordinate.
- Here is R-Code to extract this information:

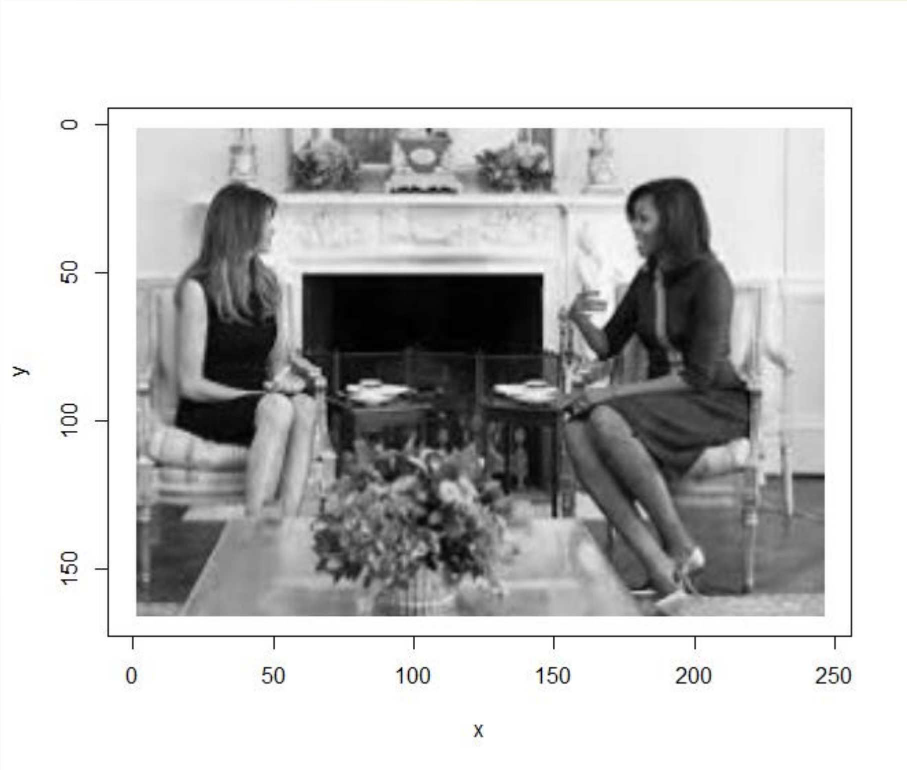
```
library(imager)
color <- load.image("H:/Stat articles/two way cluster/First_ladies.jpg")
```



Is it like gene expression data?

- Gene expression data has rows with samples, columns with genes and a numeric gene expression.
- Our color image contained an x pixel co-ordinate (like genes), the y pixel co-ordinate (like samples)and three numeric values for the intensity of the red, blue and green colors at that co-ordinate.
- This can be converted to monochrome and one intensity value (like the gene expression) using

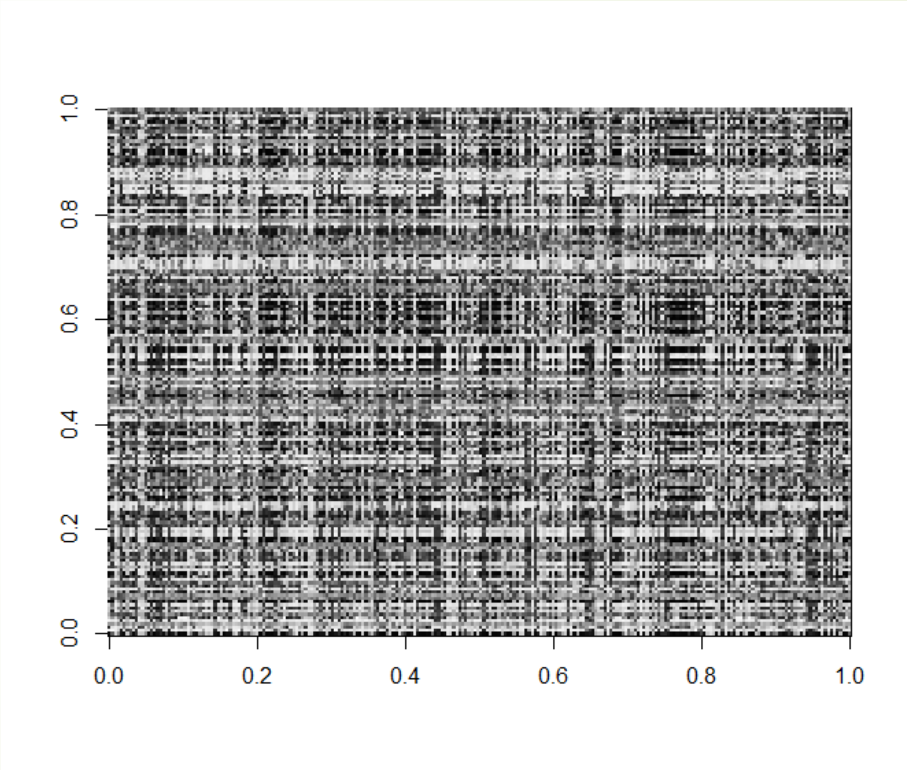
```
bw<- grayscale(color)
bw_data <- as.data.frame(bw);
```



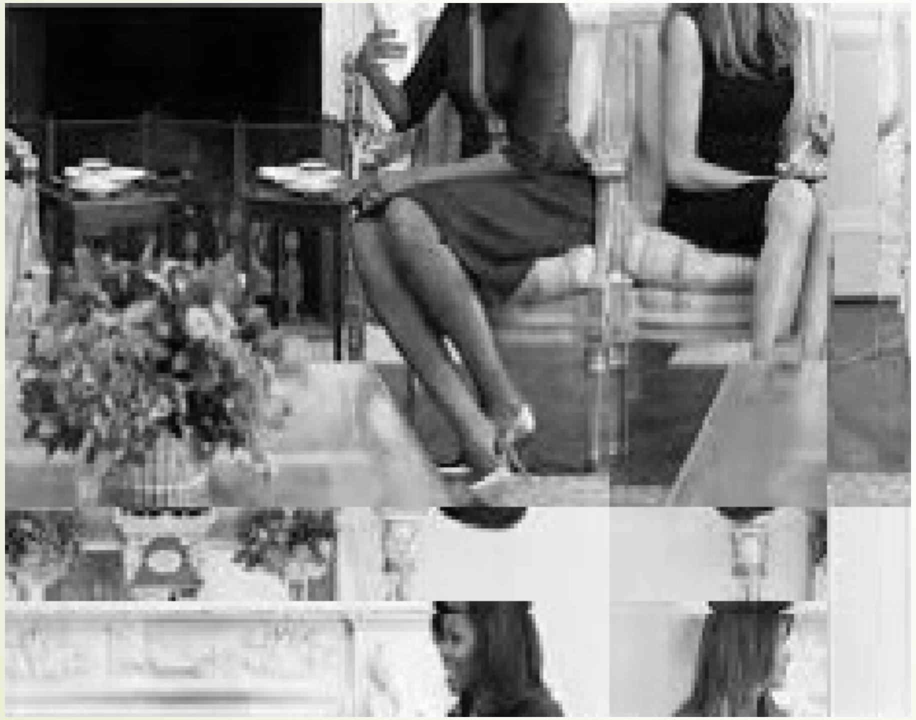
Permuted columns and rows

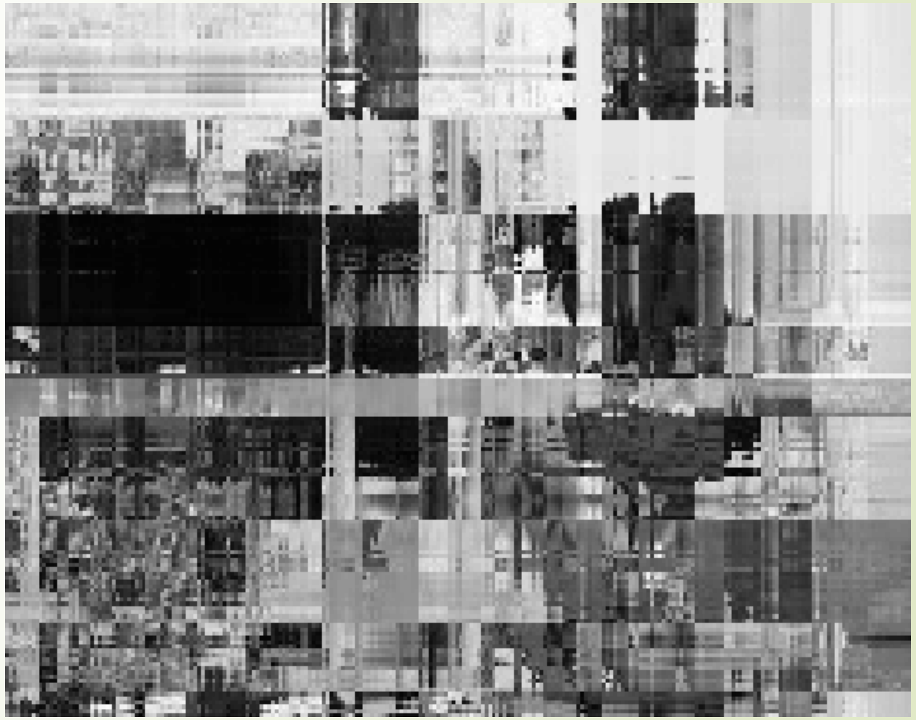
- We can permute our dataset using the following R code. Note that the image has 246X166 cells.

```
VEC <- bw_data$value
VECt <- matrix(VEC,246,166)

set.seed(1234567)
ind <- 1:166
Rind <-
sample(ind,length(ind),replace=FALSE,prob=NULL)
rVECt <-VECt[Rind,]
Rind

set.seed(145967)
ind <- 1:246
Cind <-
sample(ind,length(ind),replace=FALSE,prob=NULL)
rcVECt <-rVECt[,Cind]
Cind


image(t(rcVECt),col=paste("gray",1:99,sep=""))
```

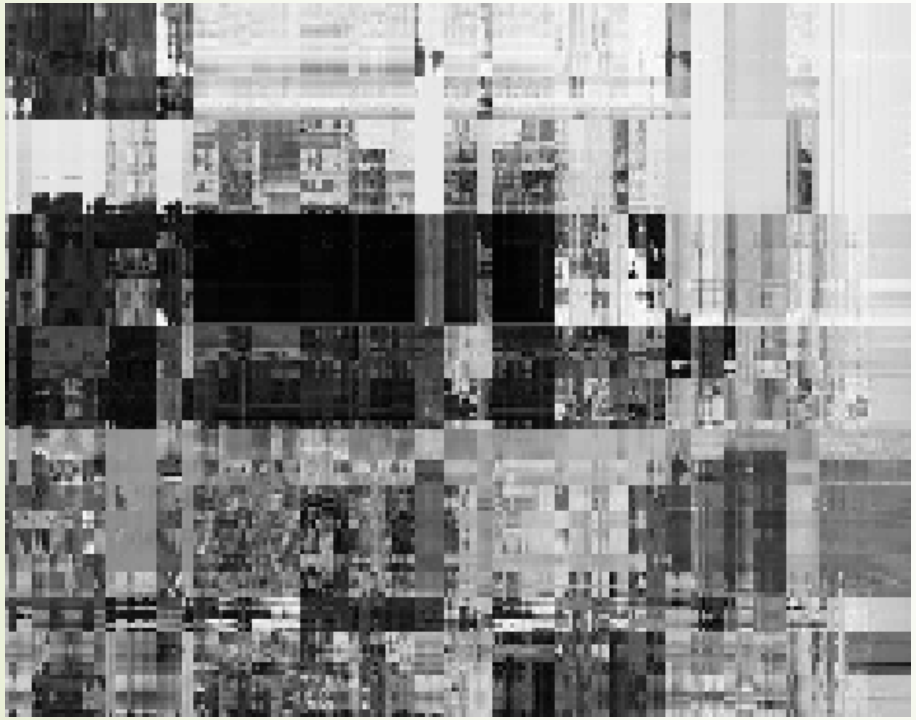


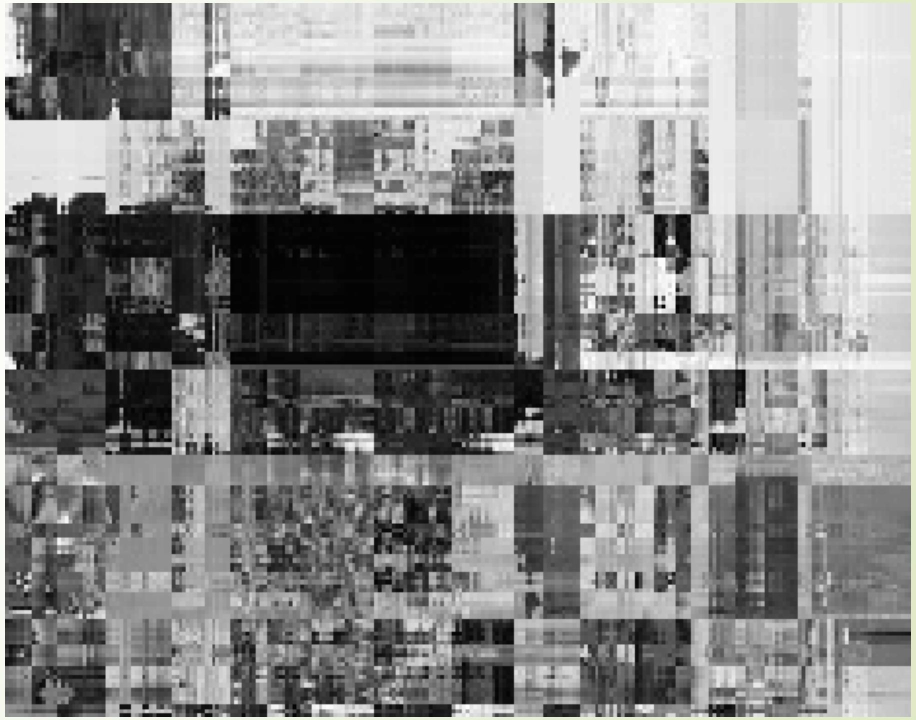
Edge Clustering (left) and Sorted Single Linkage Clustering (right)



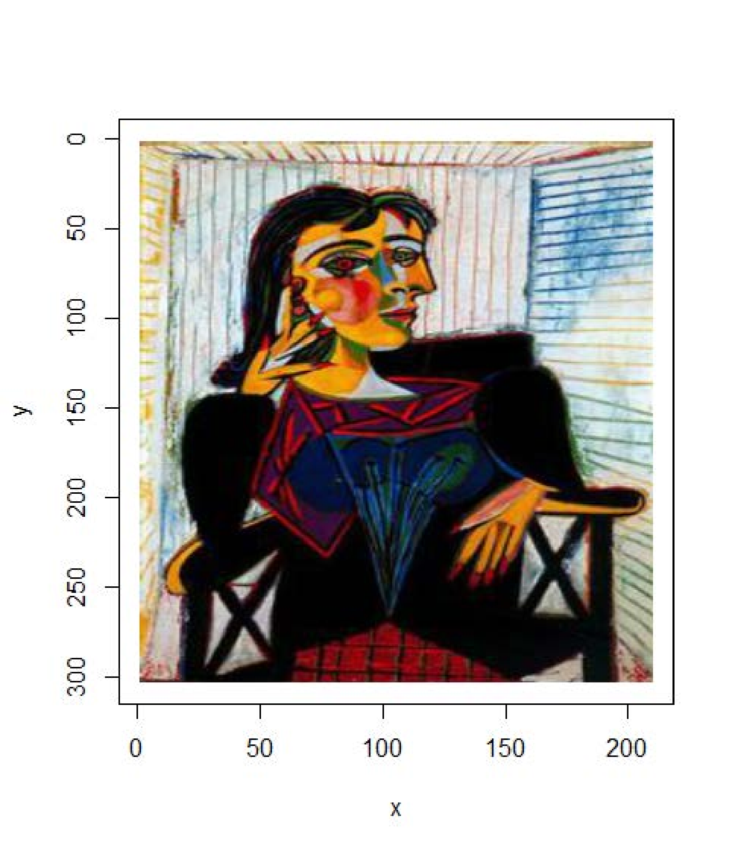


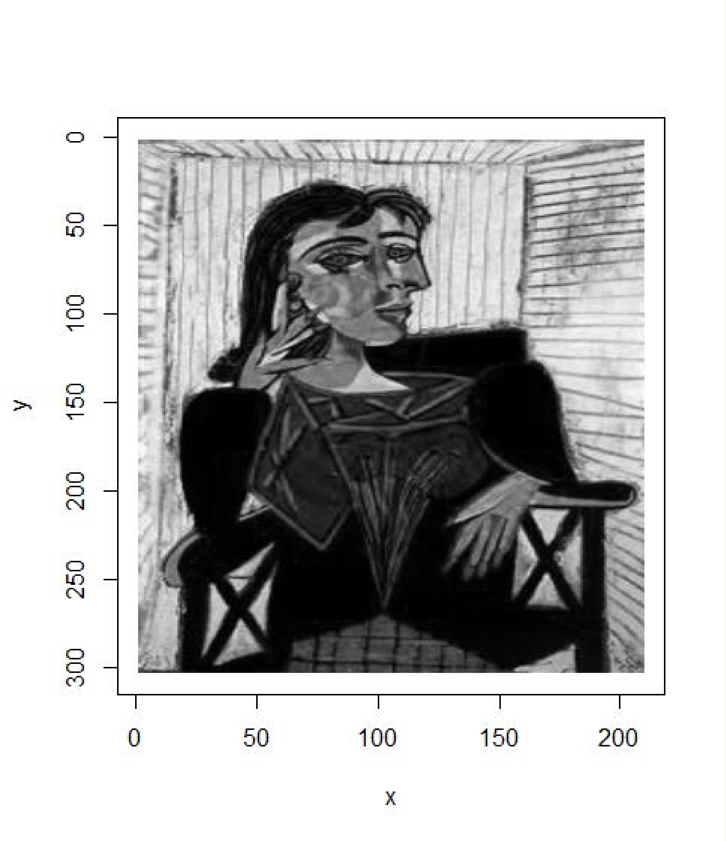
Sorted Complete Linkage (left) and Sorted Average Linkage Ordering (right)





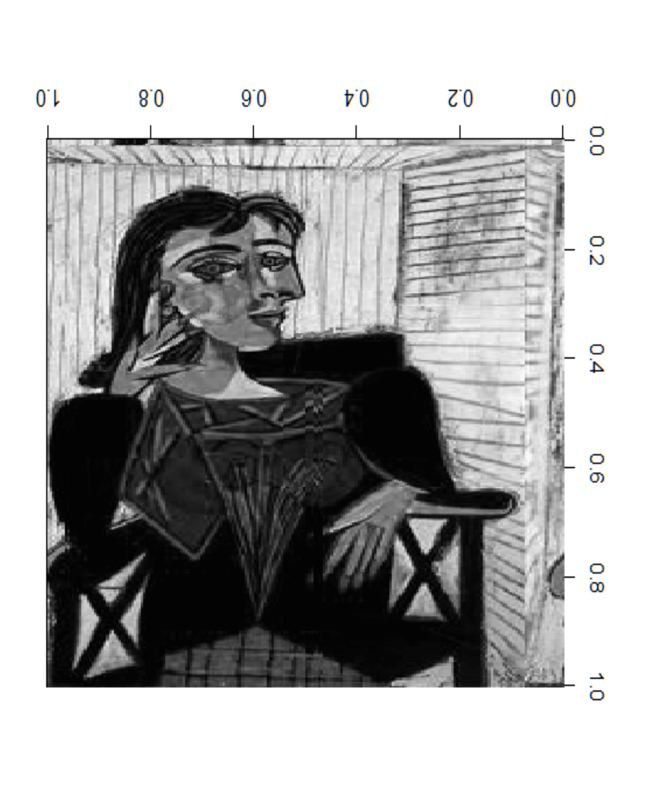
Picasso Portrait of Dora Maar-1937. Original (Left) and Grey Scaled (Right)

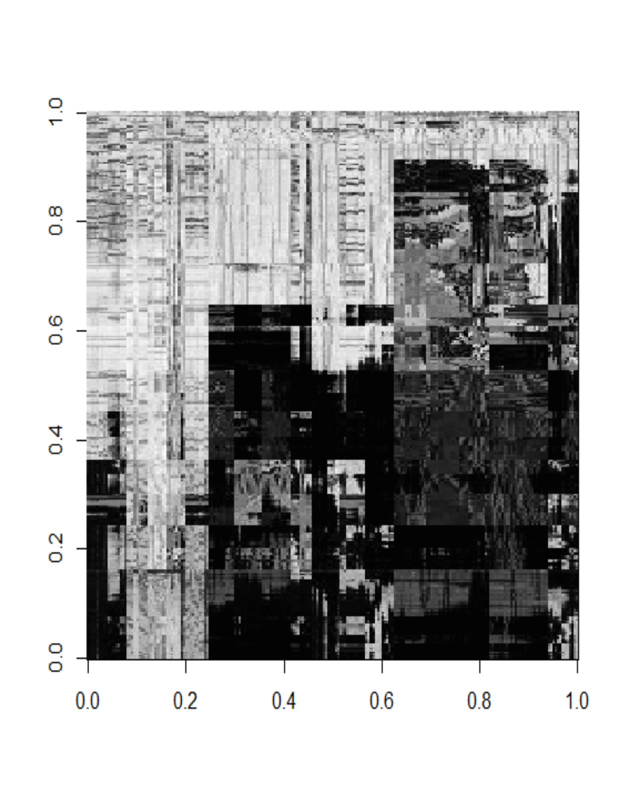




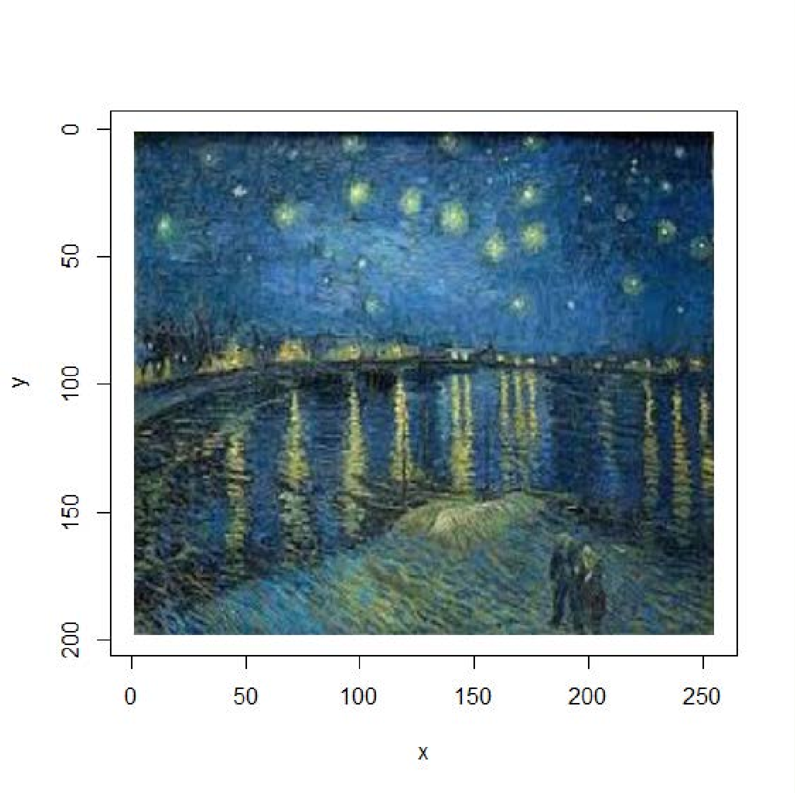
A look at Picasso inspired by the discombobulated first ladies!!

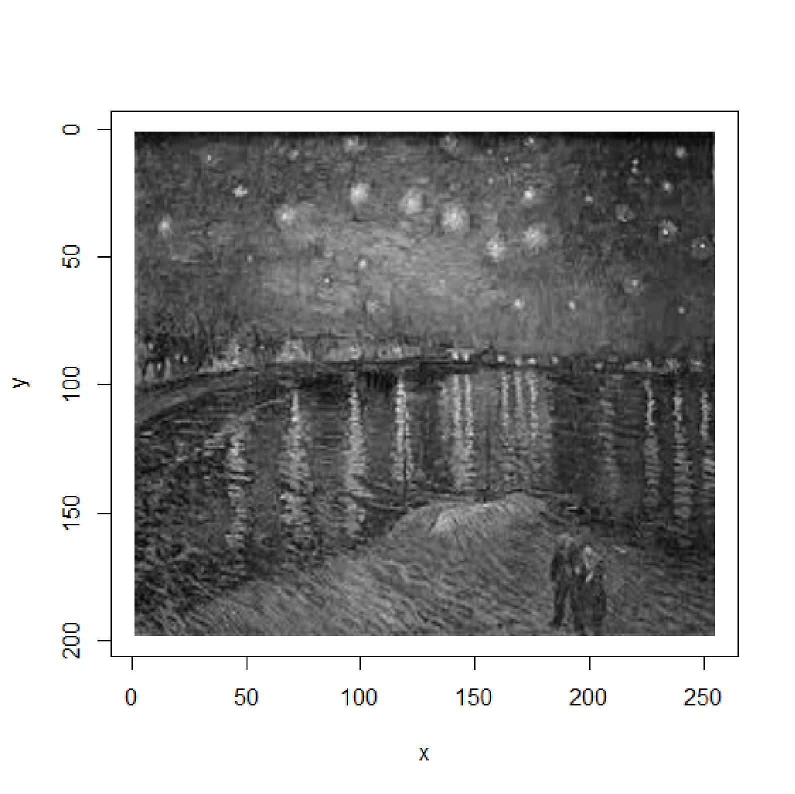
Picasso recovered – Edge cluster (Left) and Sorted Ward (Right –new improved Picasso on sale by author at Sotheby’s for $10 million!)



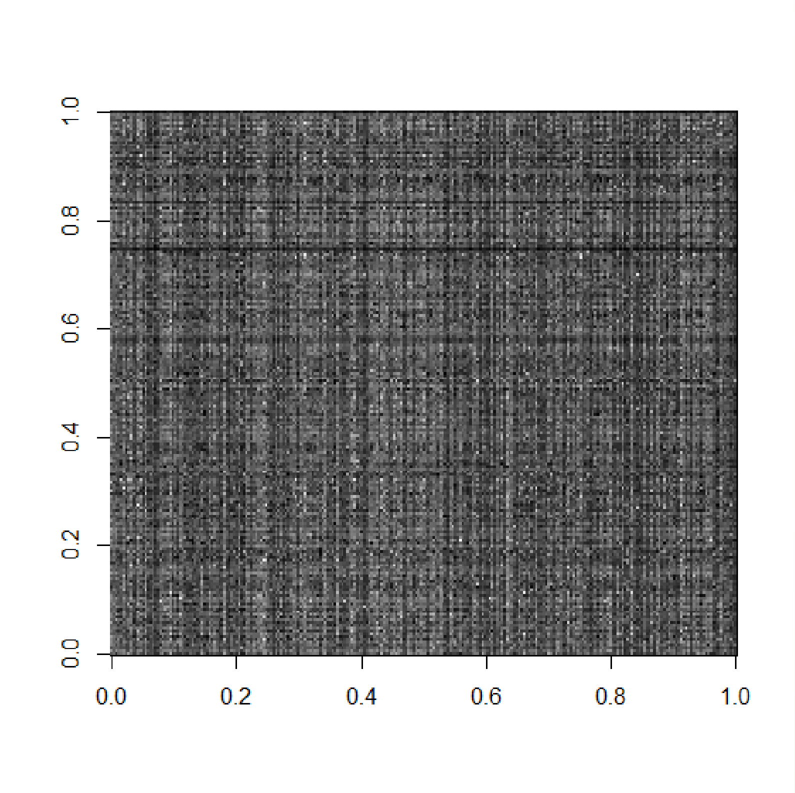


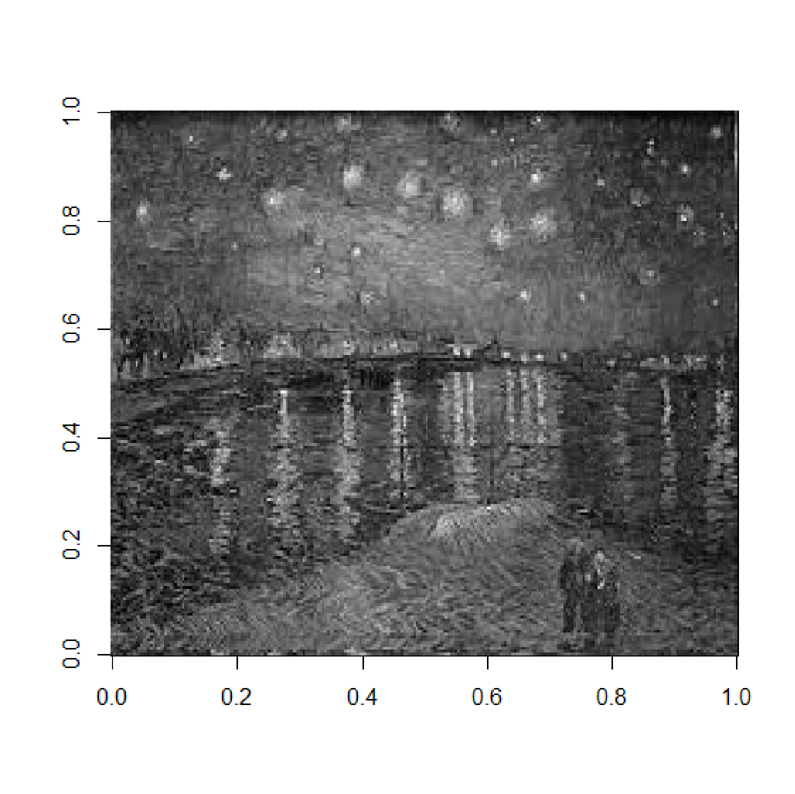
Van Gogh Starry Night over Rhone – Color (Left) and Grey Scaled (Right)



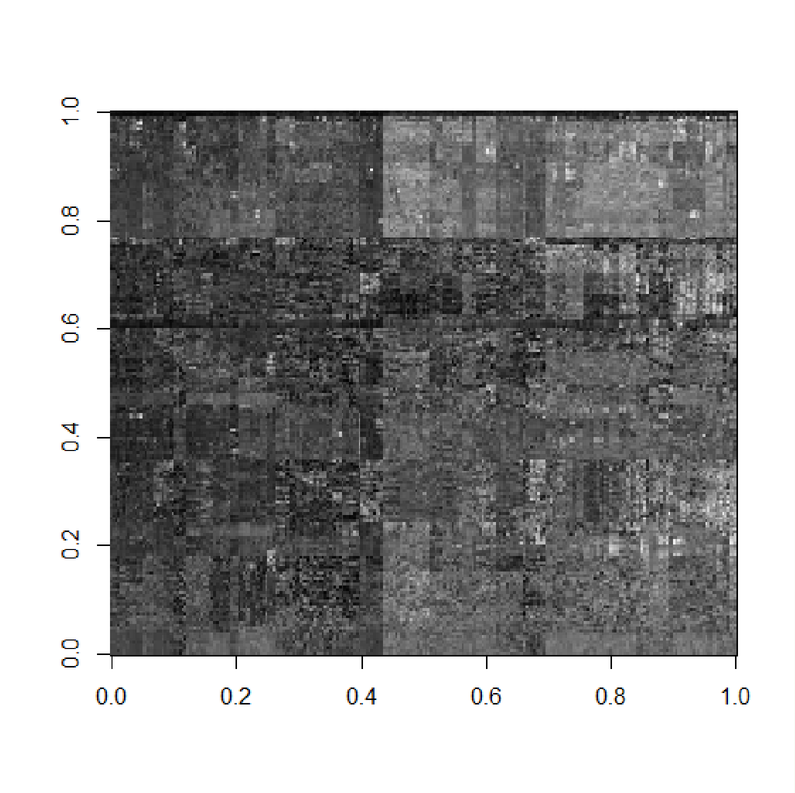


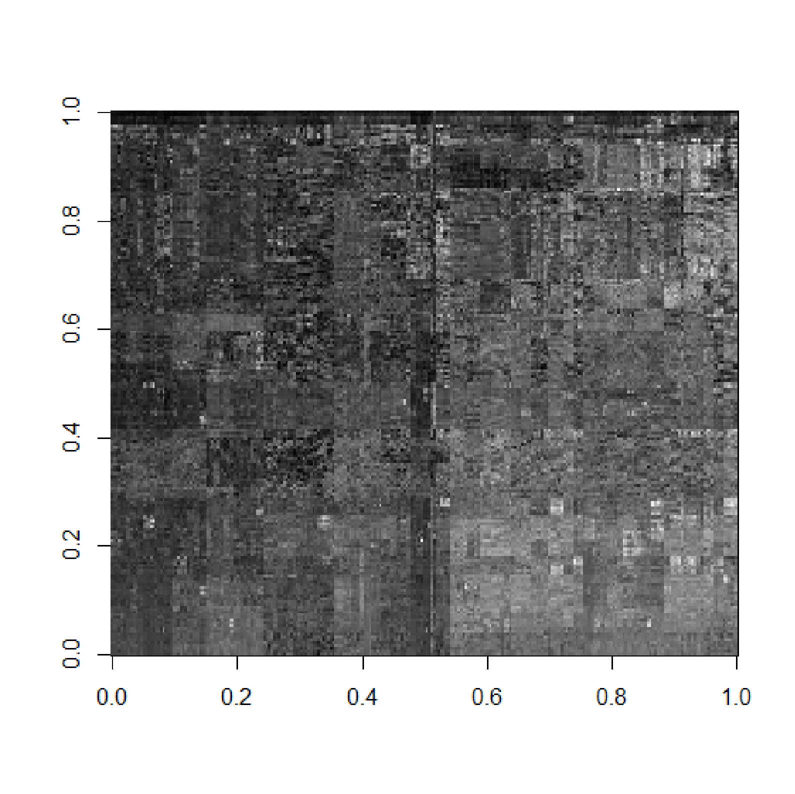
Van Gogh – randomly permuted data (left) and edge ordered (right)





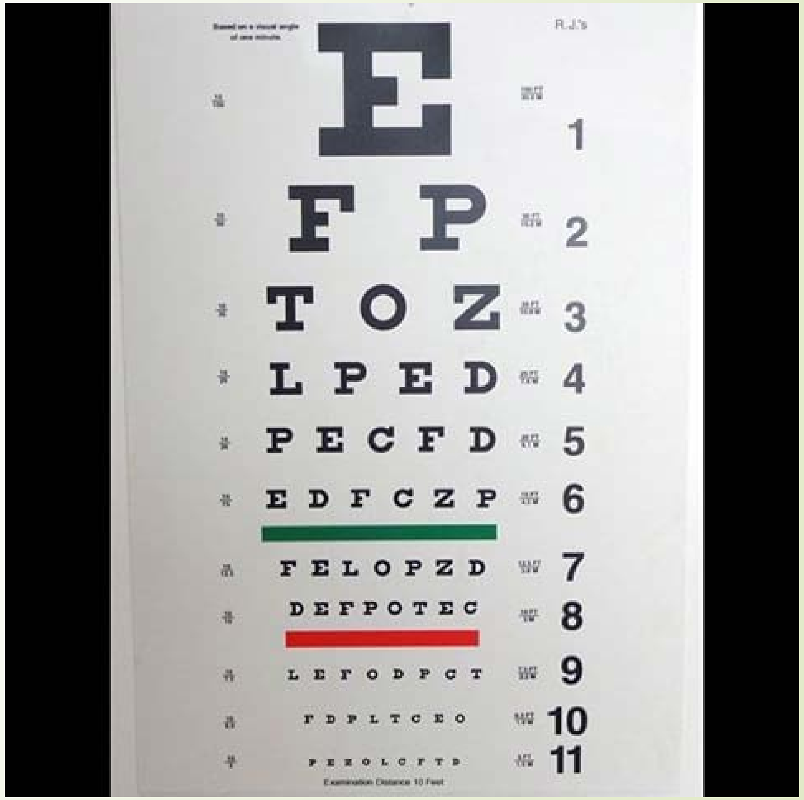
Van Gogh – sorted complete linkage (left) and sorted average linkage (right)



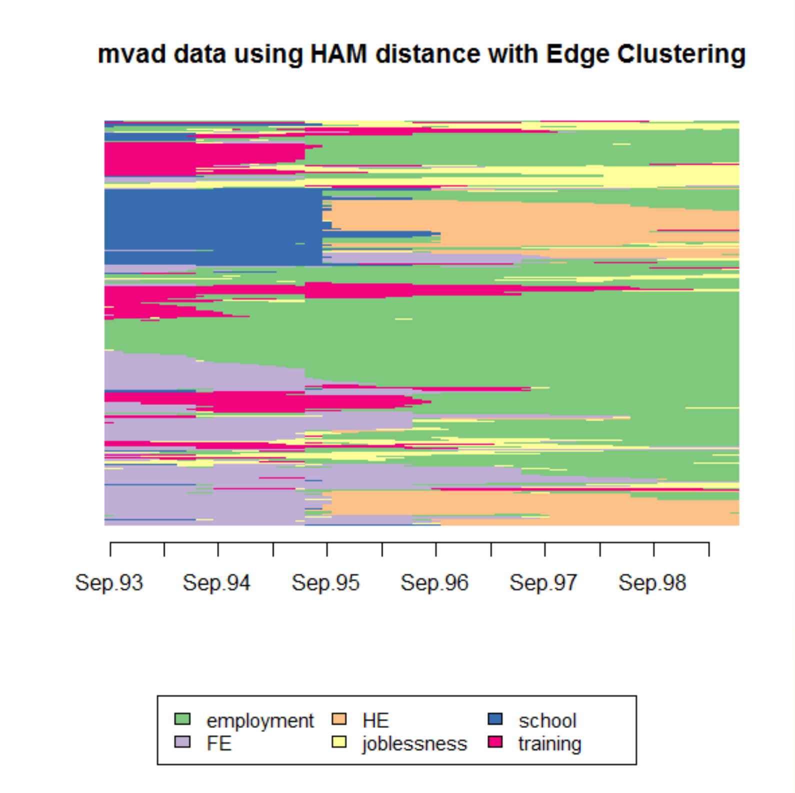


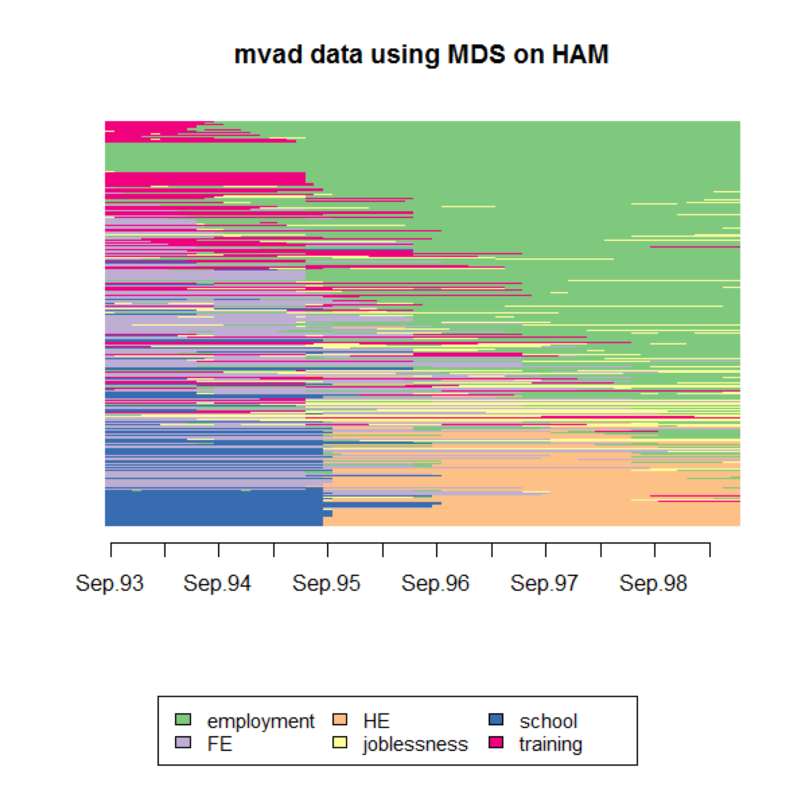
Next few slides will be like getting fitted for glasses: Which is better? 1.. or 2..? This .. or that…

- 100 trials, 50 successes. Confidence Intervals for the proportion. Can you tell which one is better?
- (0.4038, 0.5962) agresti-coull
- (0.4020, 0.5980) asymptotic
- (0.3983, 0.6017) exact
- (0.4038, 0.5962) wilson
- Sometimes you can tell from the estimates. Often you can't. We have to go back to when it was tested with 'known' parameters.

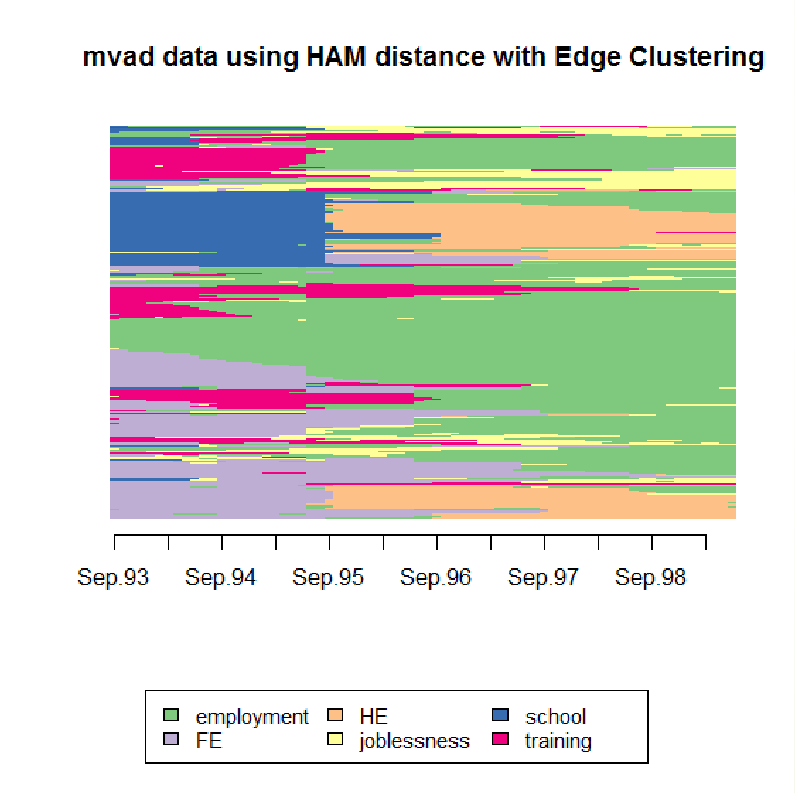


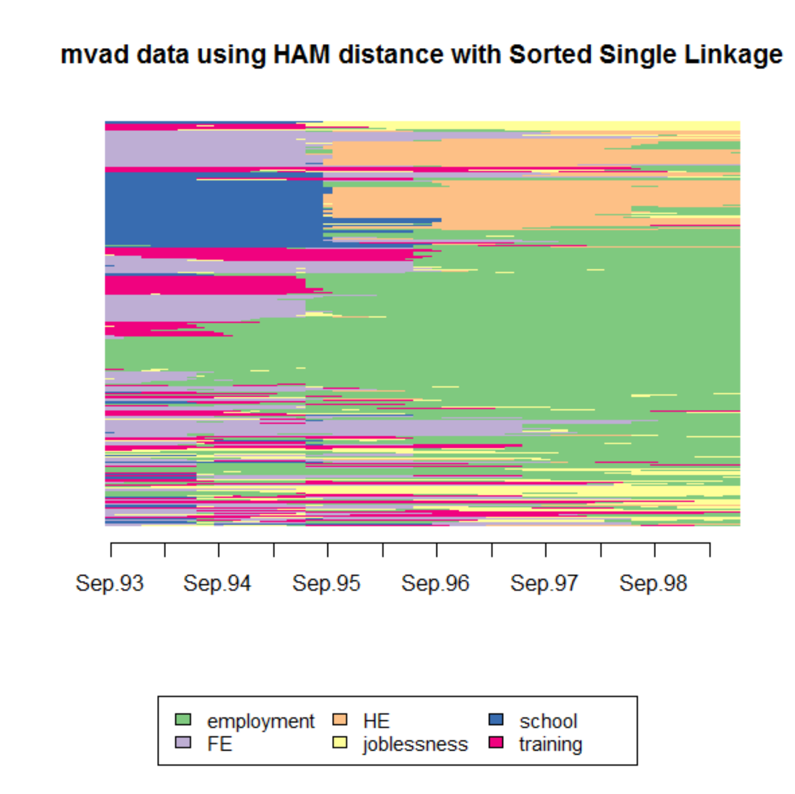
Back to Sequence Data - ‘Estimates’ of the Latent Image: Multi-dimensional scaling ordering (left) versus Edge Clustering Ordering (right) – can you pick the better estimate?



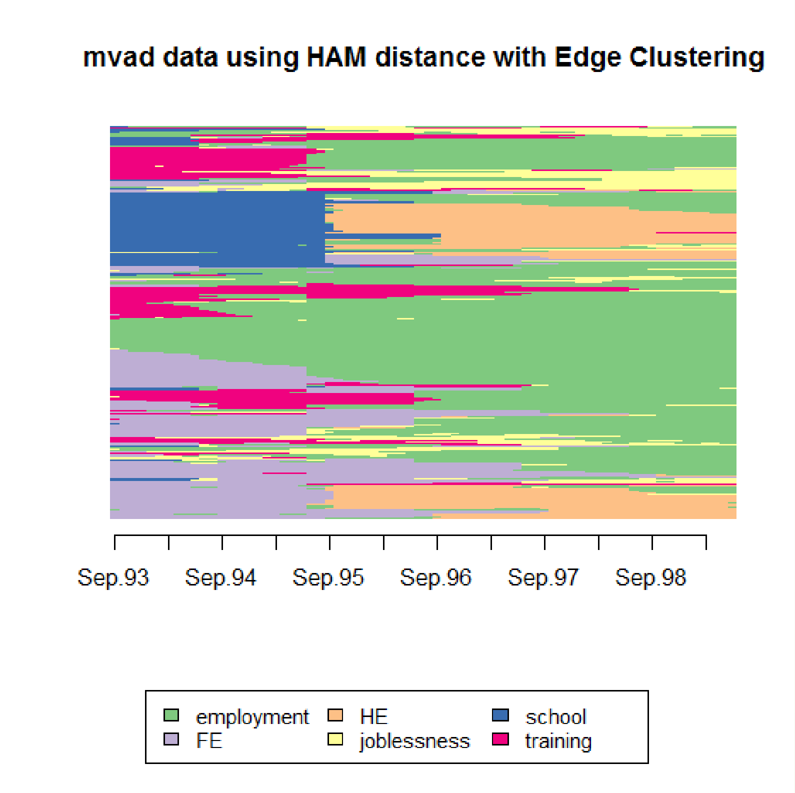


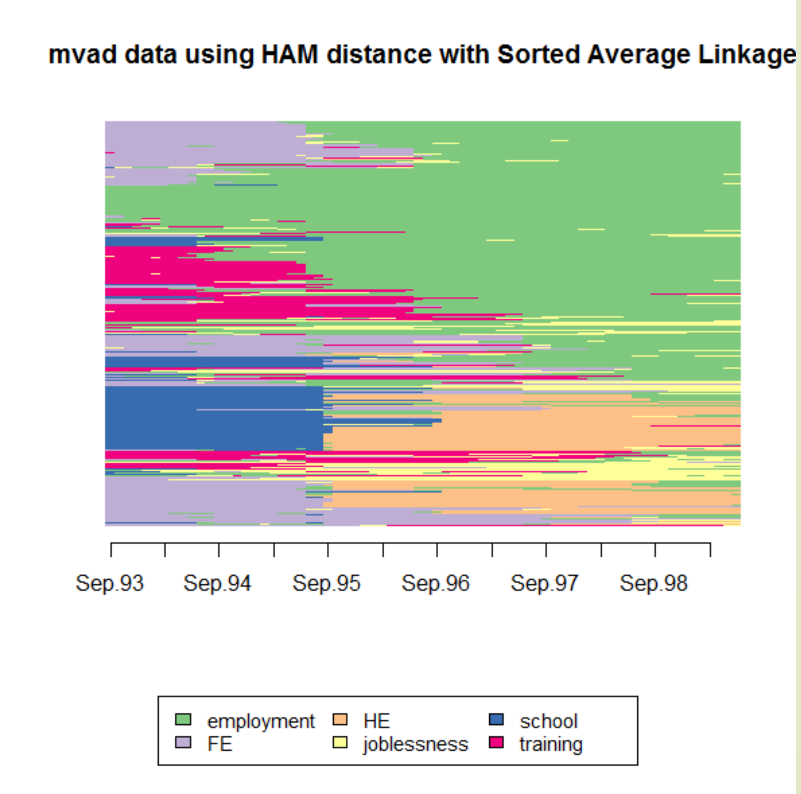
‘Estimates’ of the Latent Image in the MVAD Data: Edge Clustering Ordering (left) versus Sorted Single Linkage (right) – can you pick the better estimate?



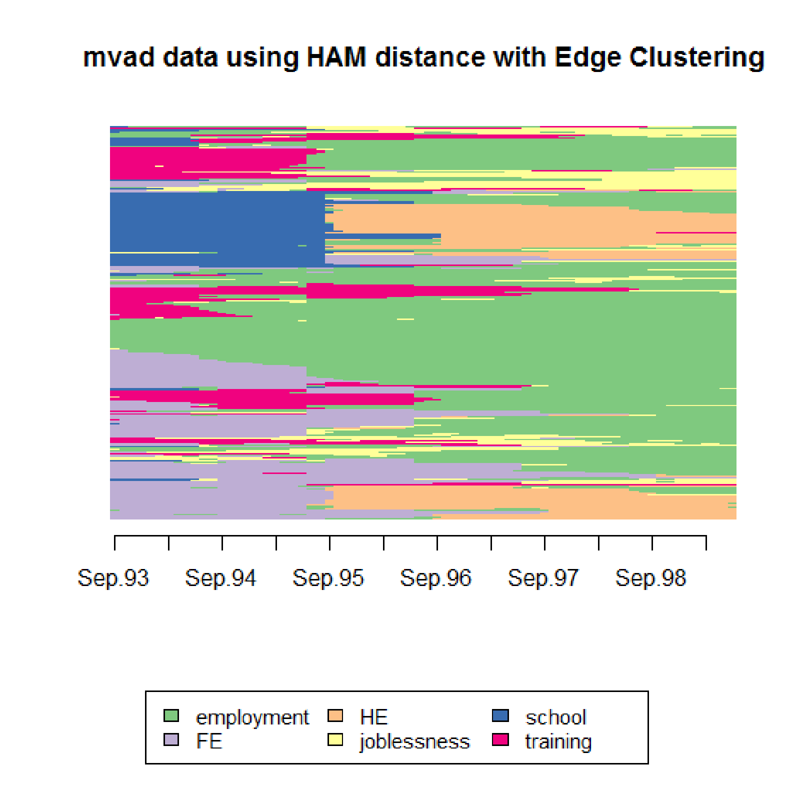


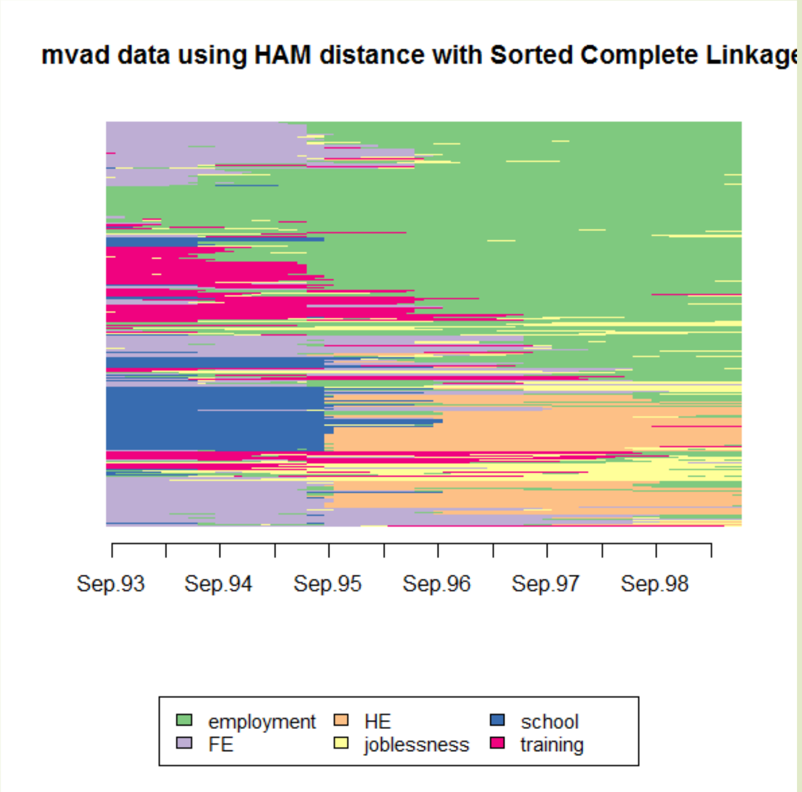
‘Estimates’ of the Latent Image in the MVAD Data: Edge Clustering Ordering (left) versus Sorted Average Linkage (right) – can you pick the better estimate? –pretty similar.





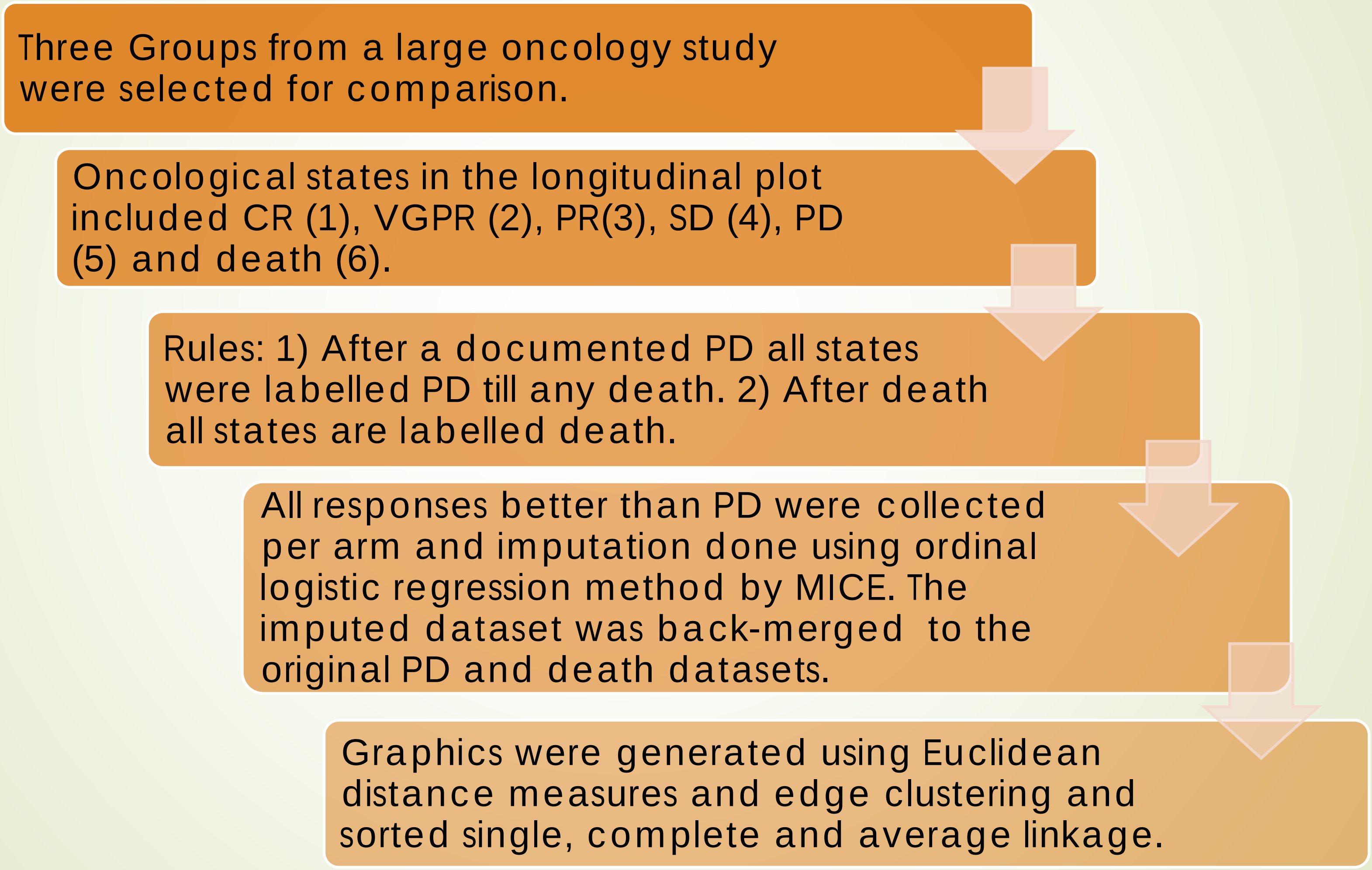
‘Estimates’ of the Latent Image in the MVAD Data: Edge Clustering Ordering (left) versus Sorted Complete Linkage (right) – pretty similar on most gross features.





Similarity in gross features validates edge clustering. It has been developed within the hierarchical clustering and statistical evaluation framework and is not a wild soul-less data crunching heuristic.

Oncology Longitudinal Graphics Example



Three Groups from a large oncology study were selected for comparison.

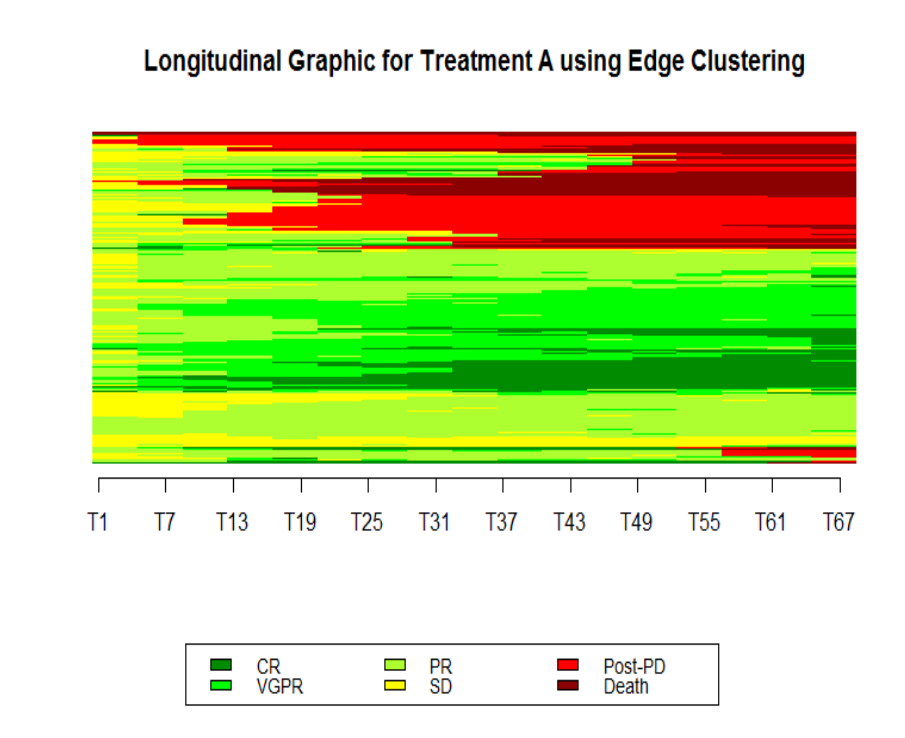
Oncological states in the longitudinal plot included CR (1), VGPR (2), PR(3), SD (4), PD (5) and death (6).

Rules: 1) After a documented PD all states were labelled PD till any death. 2) After death all states are labelled death.

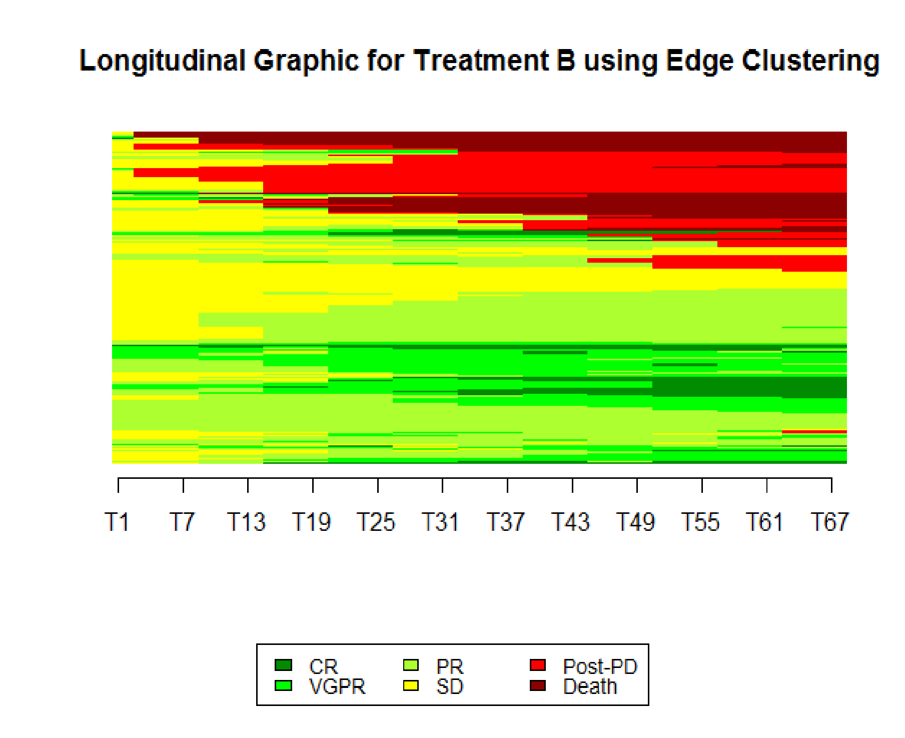
All responses better than PD were collected per arm and imputation done using ordinal logistic regression method by MICE. The imputed dataset was back-merged to the original PD and death datasets.

Graphics were generated using Euclidean distance measures and edge clustering and sorted single, complete and average linkage.

Oncology State Sequences sorted using edge clustering – TRT A (left), TRT B (right)

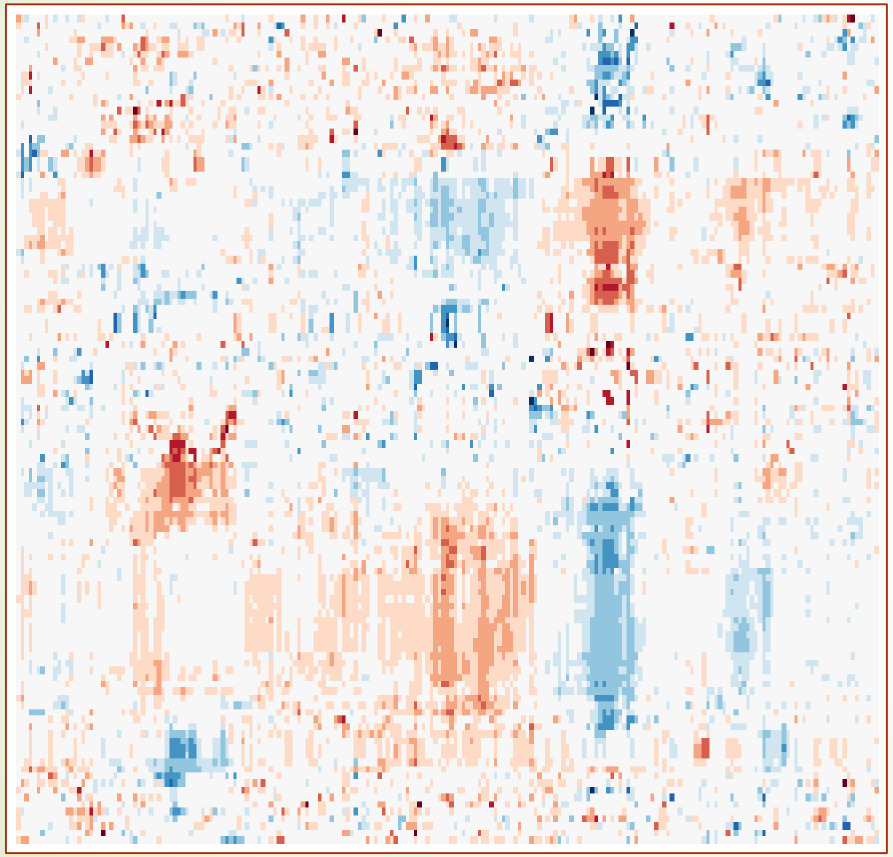


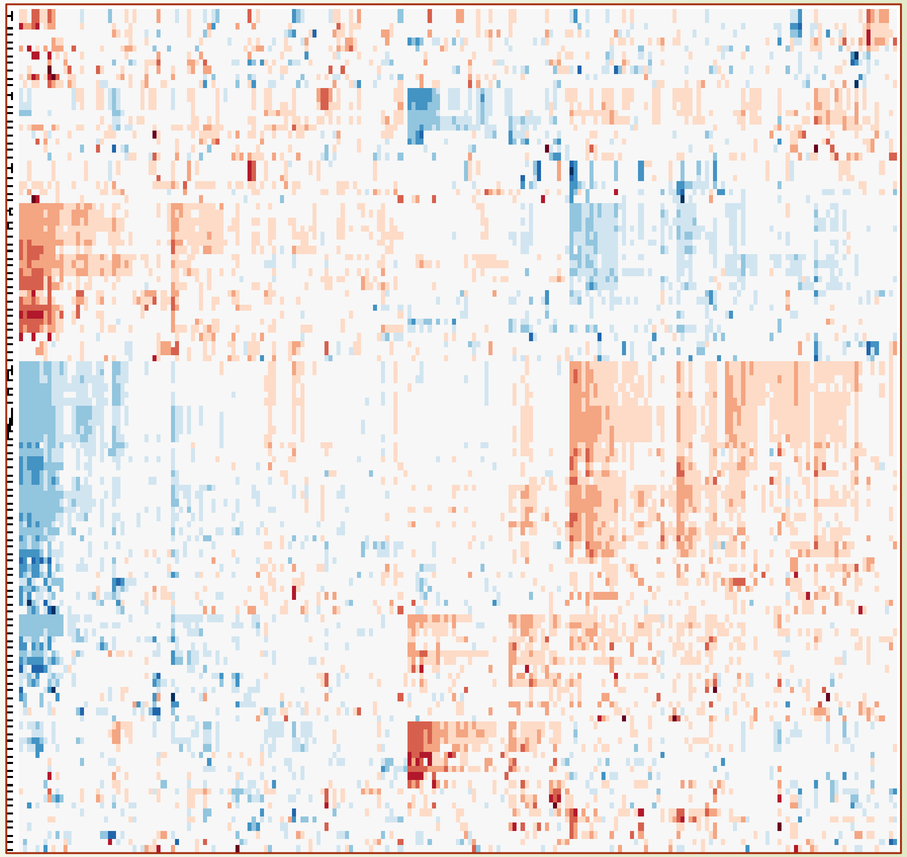
Longitudinal Graphic for Treatment A using Edge Clustering



Longitudinal Graphic for Treatment B using Edge Clustering

Back to gene - sample association data:
1.. or 2..? This .. or that…
Edge (left) and Sorted Complete (right)





Code for the sorted complete linkage at https://rdrr.io/cran/dendsort/f/vignettes/example_figures.Rmd

Final Notes

- The ‘parameter’ here is an informative image. We are looking at contexts where the latent informative image is unknown and unknowable.
- Are the contexts where we somehow know that an image is right and informative (the first ladies photograph, art work) very different from the ones where we want to find the latent unknowable image. By analogy, are we testing estimators under say, a mixture of normal distributions, when our real data will never meet such assumptions.
- The sorted hierarchical clustering methods work well as does the edge clustering with some evidence that the edge method improves on these methods.
- Clopper-Pearson, Wilson, Exact, Normal approximation? Who cares? The methods may be close enough - one may bring out a feature in a heat-map better than another. There can be many images. Aesthetic images may be different from scientifically informative images.
- **If you or your translational scientist try edge clustering and see Elvis in a heat-map please do let us know!!**

Questions:
Right: Hats of Different Sizes and Colors!

Acknowledgements:

The authors acknowledge the support of Stanley Kotey and Mia He for developing the sequence plots for oncological states and Sibabrata Banerjee for help with the application and image processing for the two-way heat-maps. We thank statisticians at Celgene for their helpful comments during a dry-run of this presentation. Management support by Arlene Swern and Janis Grechko is much appreciated.