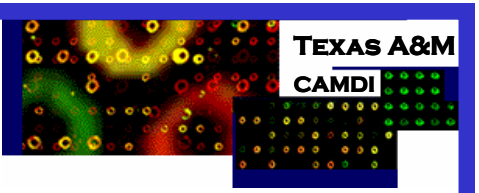


Clustering Algorithms:

Can anything be Concluded?

Edward R. Dougherty, Seungchan Kim

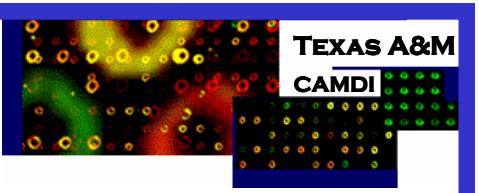
Texas A&M University

Junior Barrera, Marcel Brun, Roberto Marcondes

Universidade de Sao Paulo

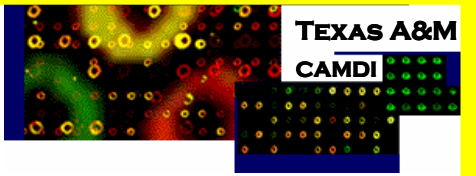
Yidong Chen, Michael Bittner, Jeffrey Trent

National Human Genome Research Institute, National Institutes of Health

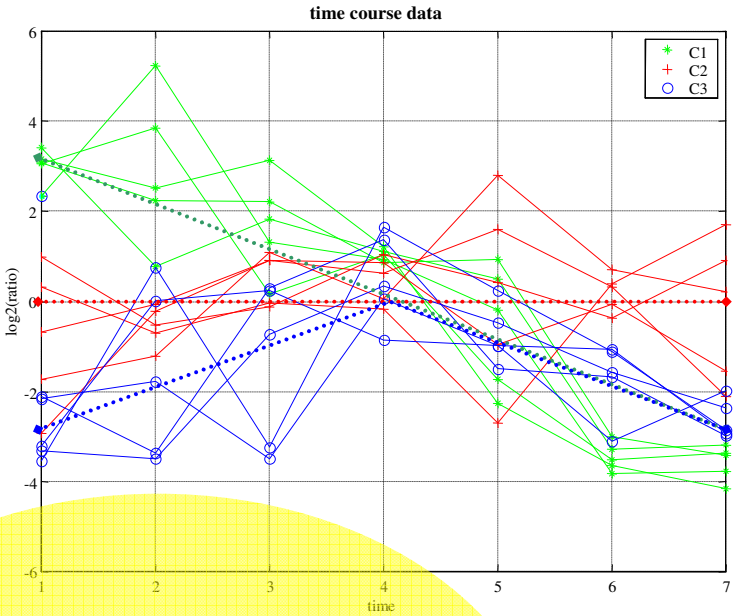


Objectives

- Examine the precision of sample-based clustering relative to population inference
- Study the effects of the number of replicates of microarray experiments
- Comparison between the various clustering methods



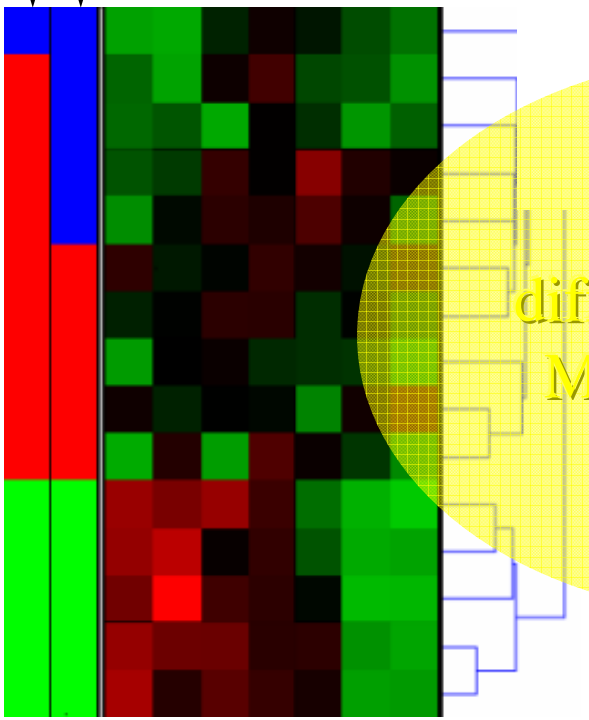
Time course data



Clustered by dendrogram

Original clusters

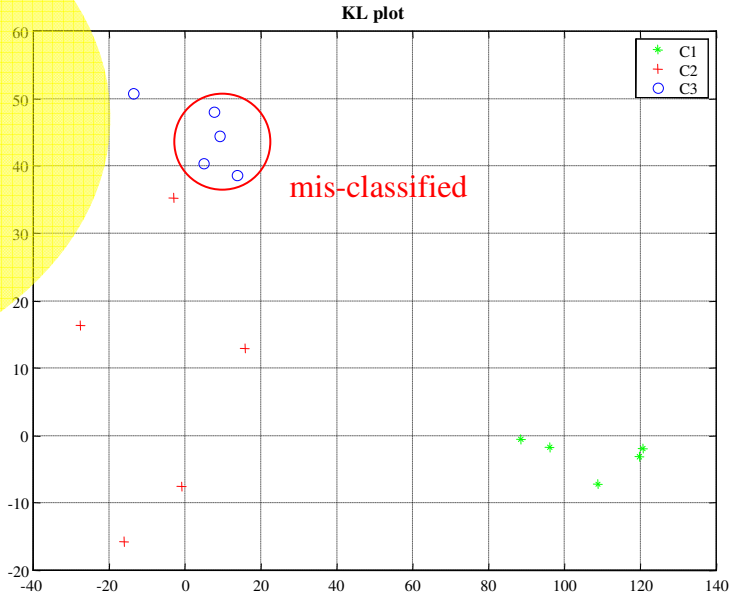
Dendrogram



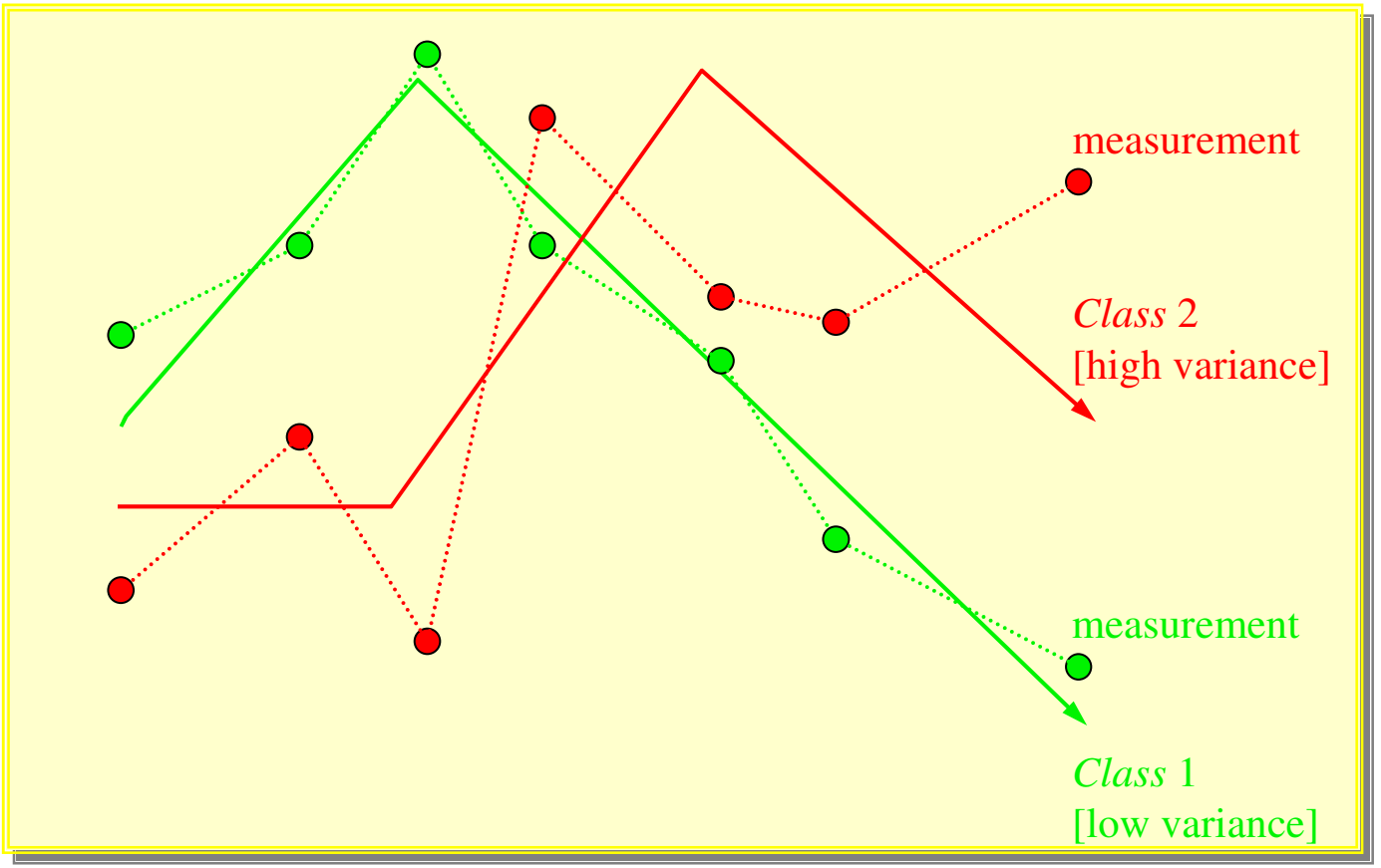
different views of Microarray data

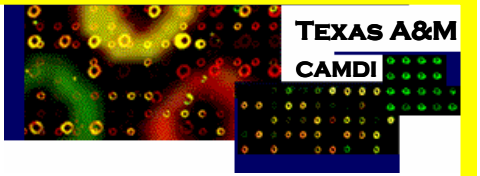
PCA plot

multidimensional space



Time Course Model





Clustering algorithms

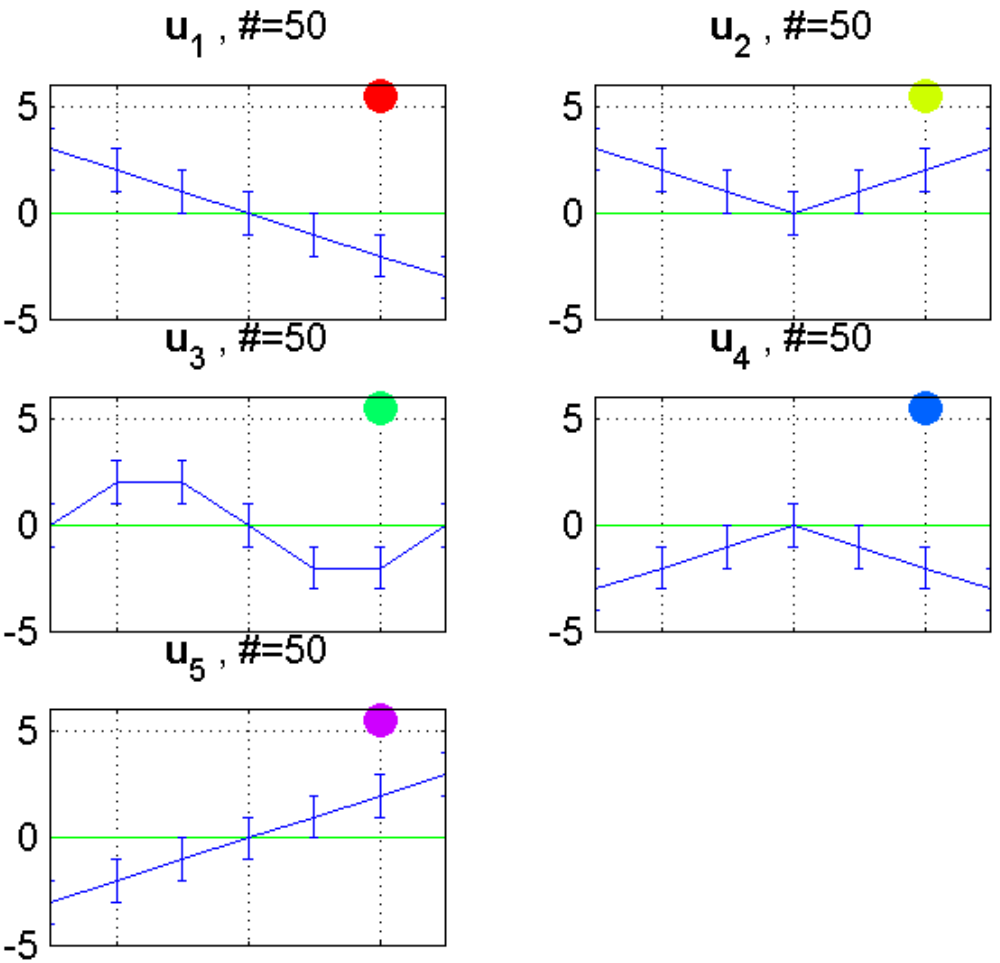
- K-means
- Fuzzy c-means
- Self Organizing Map
- Hierarchical clustering with Euclidean
- Hierarchical clustering with correlation

Clustering errors

- How clustering errors change as the number of replicates increases?
- How differently each clustering algorithm perform?

Synthetic example

- 5 synthetic templates
- simulated data from the templates
  - different variances
  - different replicates
- 5 different clustering methods

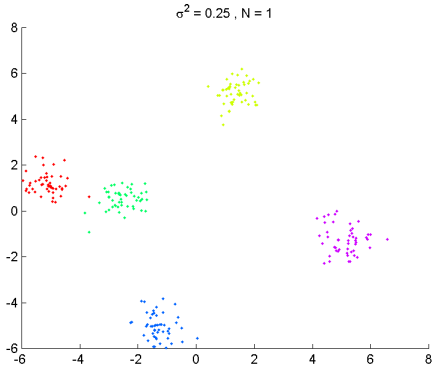


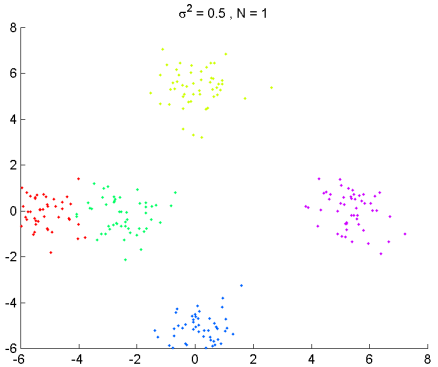
Variances of Data vs. # of Replicates

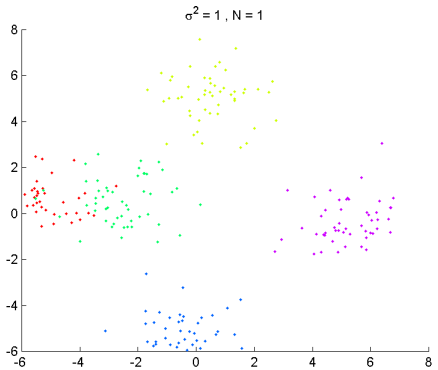
- The number of replicates required to get a reasonable clustering result varies, depending on the variance of gene expression levels
- Clustering algorithm must also be chosen correspondingly to get the best clustering algorithm. No universal clustering algorithm!

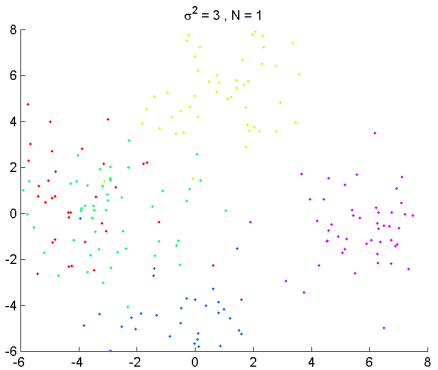
Simulated Data

different variances

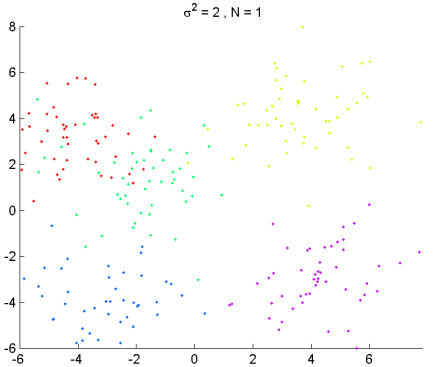


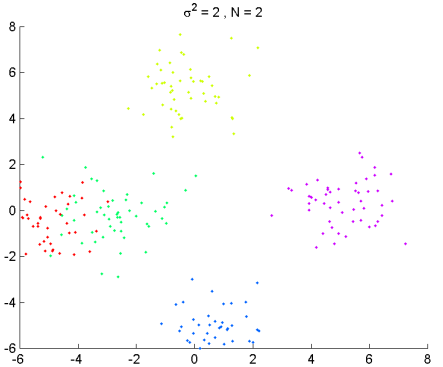


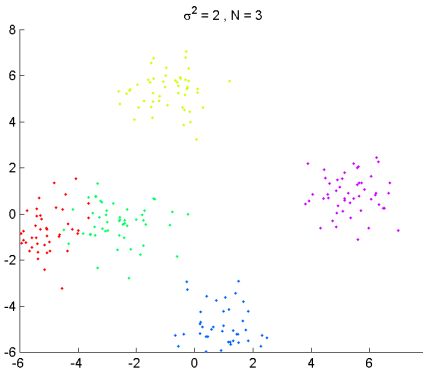


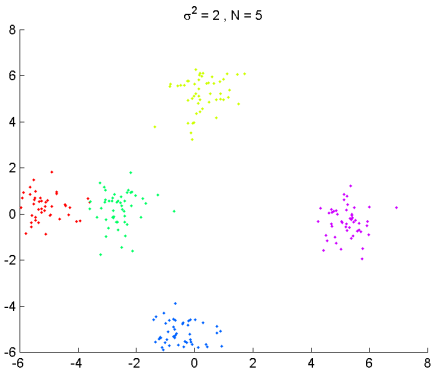


different replicates







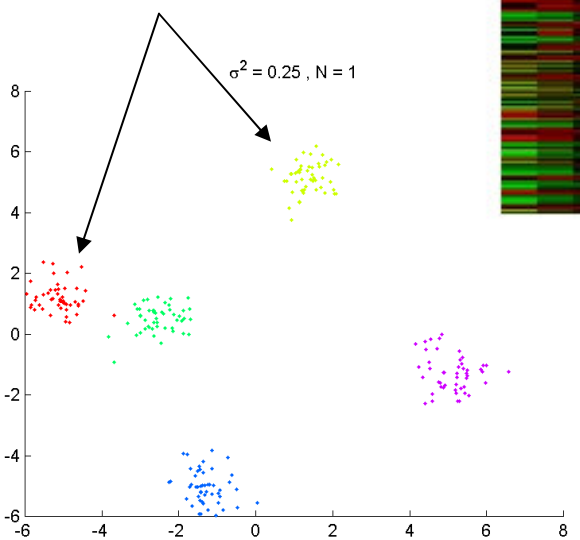


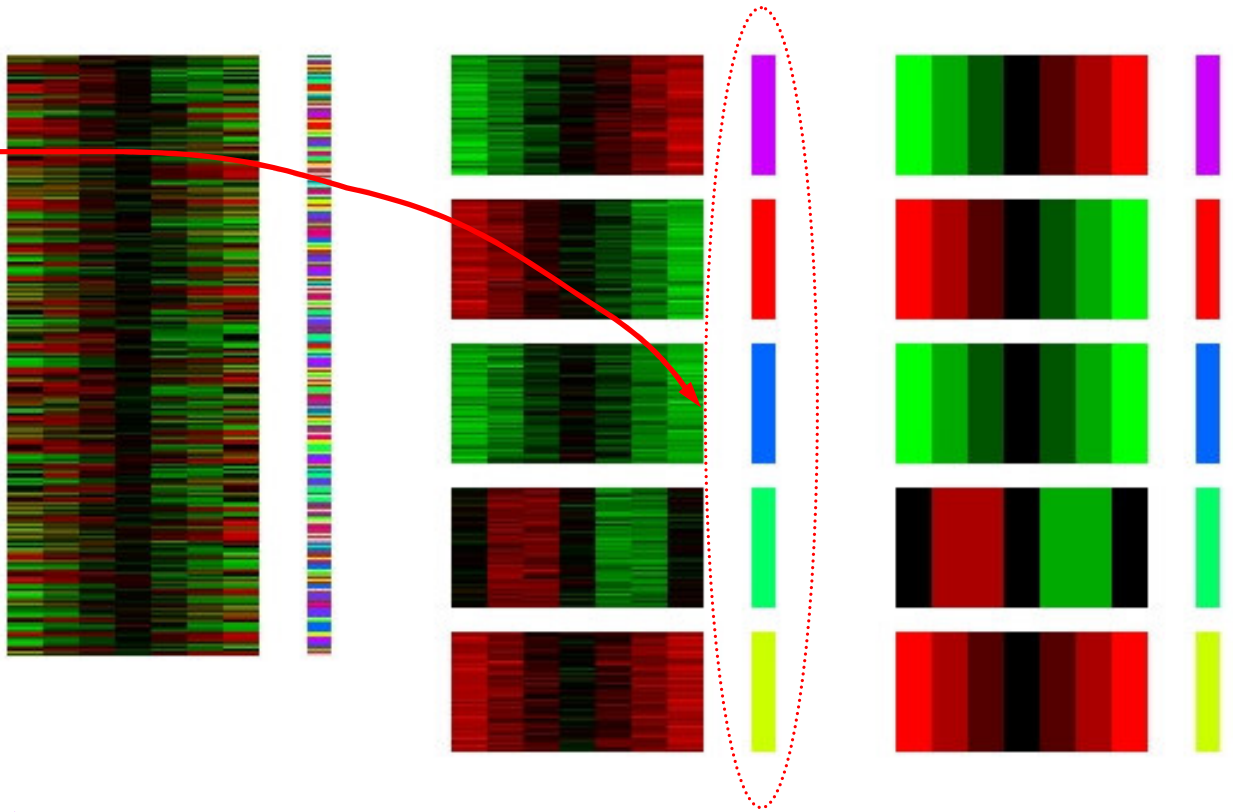
Single experiment

$\sigma^2 = 0.25$

No error!

Tighter clusters due to small variance





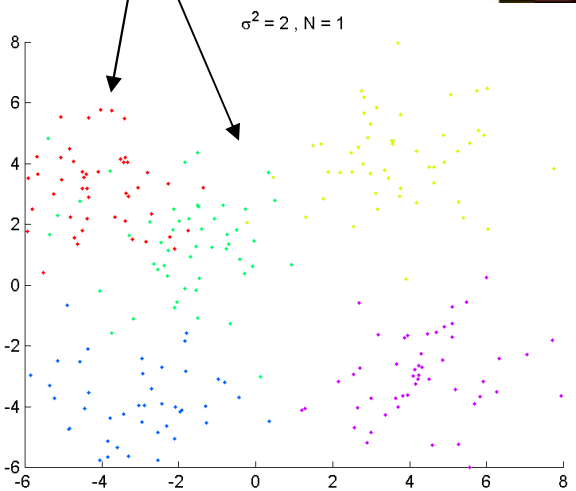
Results from Fuzzy c-means

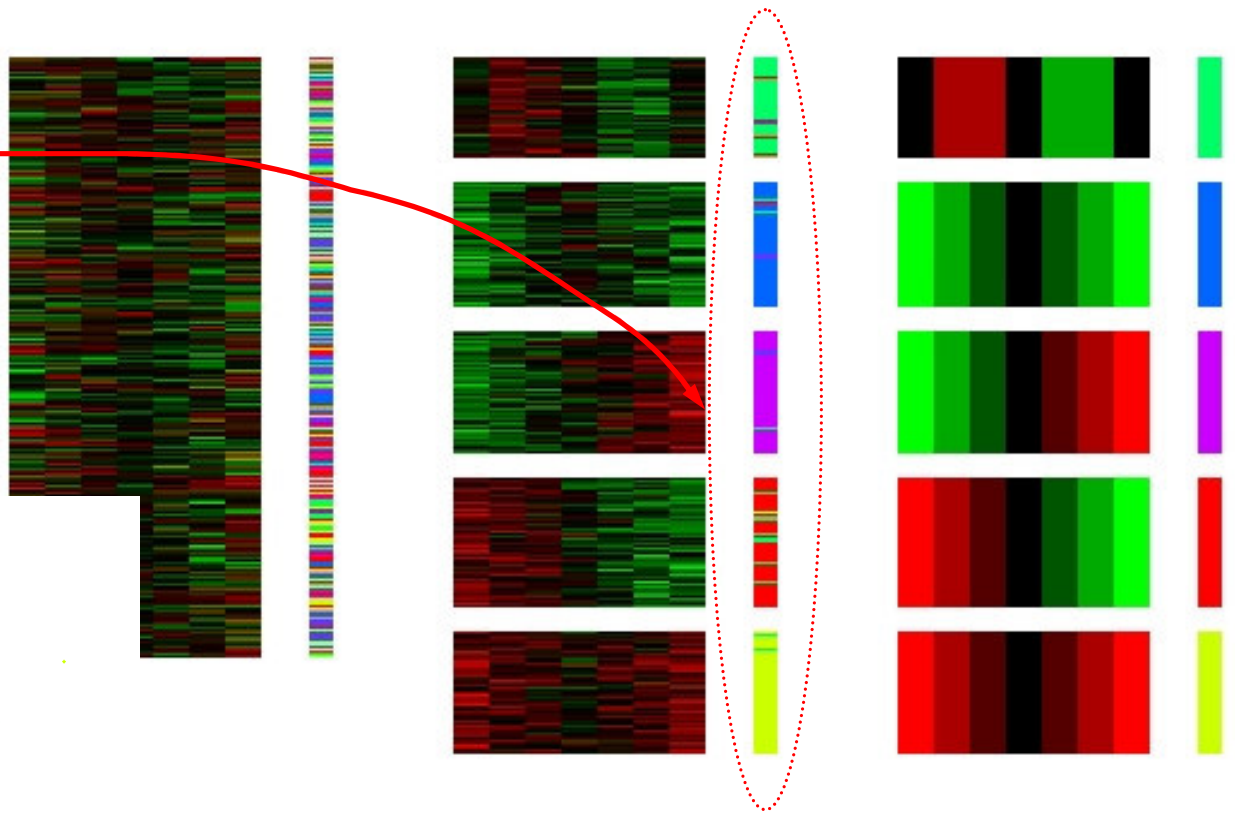
Single experiment

$\sigma^2 = 3.0$

many misclassifications

clusters start mixing



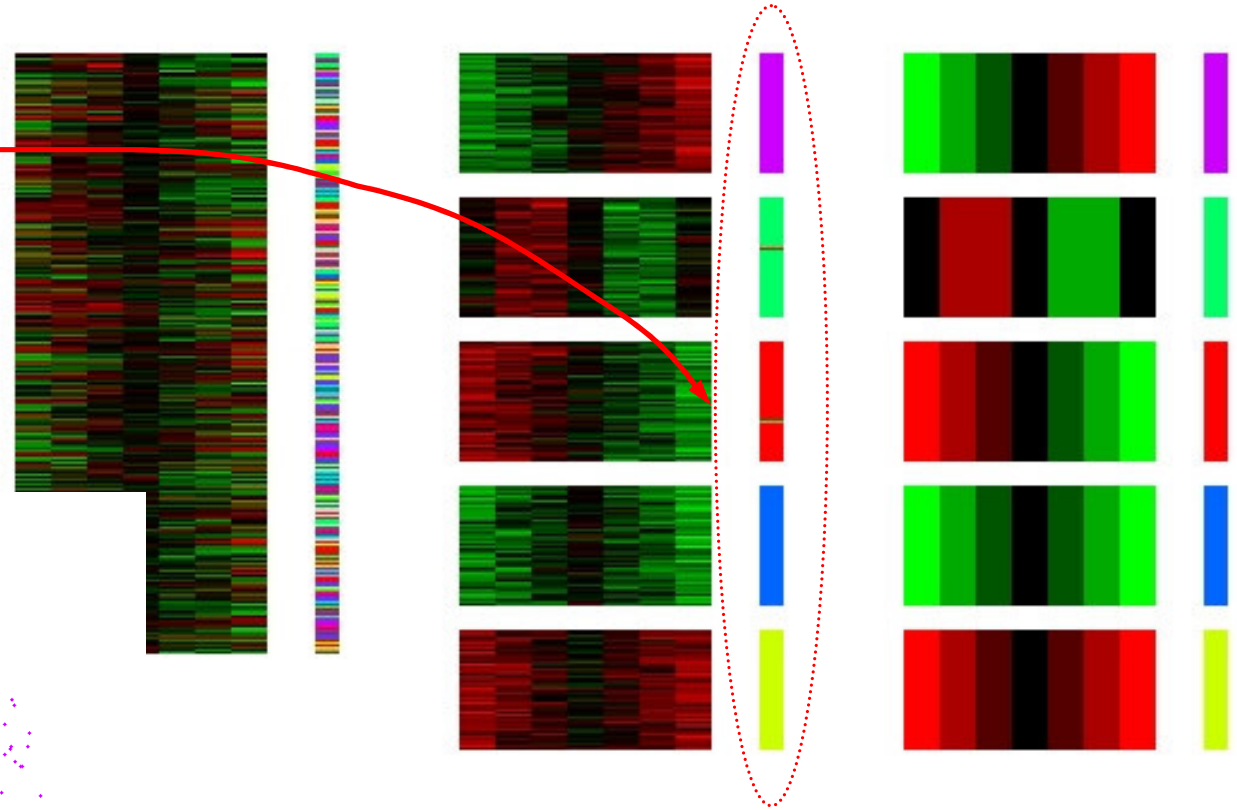


22 misclassifications (8.8%)

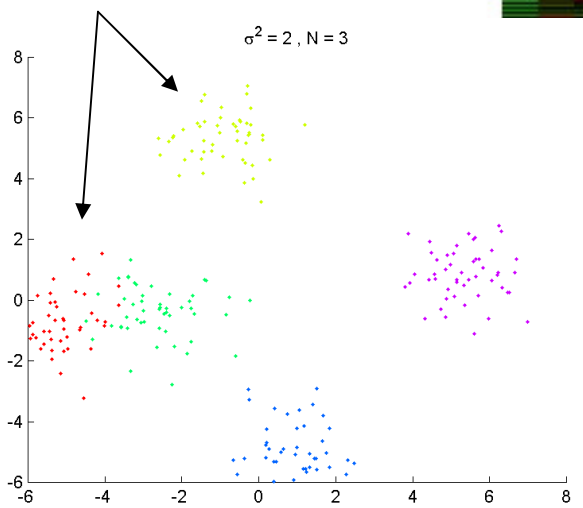
Replicated experiment

$\sigma^2 = 3.0, N = 3$

very few misclassifications

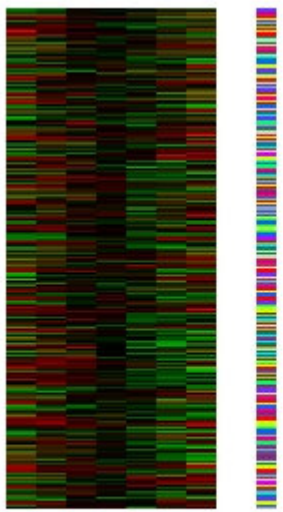


Clusters well separated due to the replication

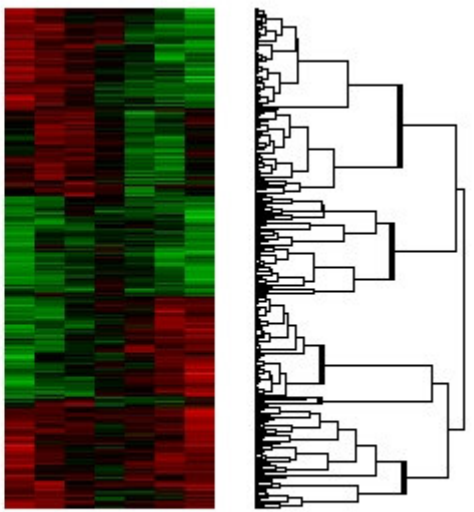


2 misclassifications (0.8%)

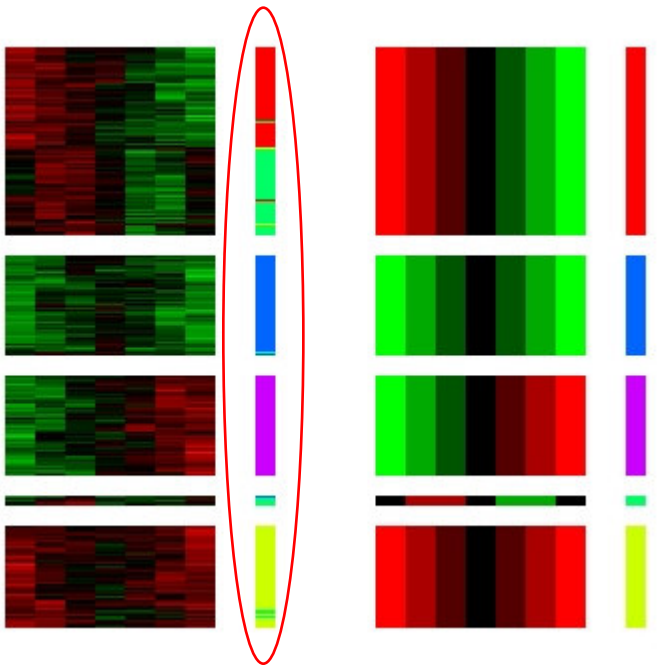
Hierarchical clustering error!!!



Before clustering



After clustering
with a NICE dendrogram



24.5% Error!!

Algorithm: *Hierarchical clustering with correlation measure*

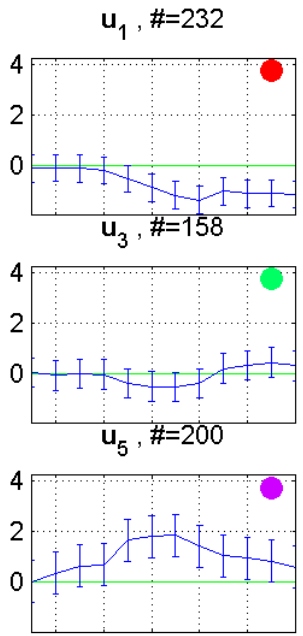
Real Data Applications

- Initial clustering to generate templates
  - means
  - variance (individual or pooled)
- Simulate time course data based on the templates generated by initial clustering
  - different # of replicates
- Apply various clustering methods
  - expected clustering error for each method

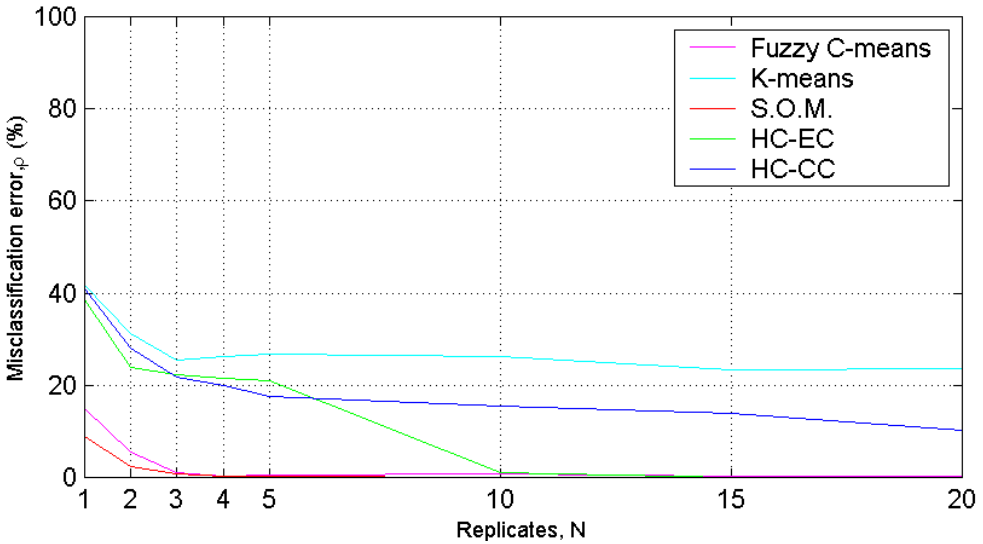
Sample data and Initial clustering

- Data from V.R. et al. (1999) Science 283: 83-87
  - The Transcriptional programs in the Response of Human Fibroblasts to Serum
  - http://genome-www.stanford.edu/serum/
- Initial Clustering
  - Five clustering methods
    - Fuzzy c-means
    - K-means
    - S.O.M.
    - Hierarchical clustering (correlation and Euclidean distance)
  - # of clusters
    - 5 or 9 clusters

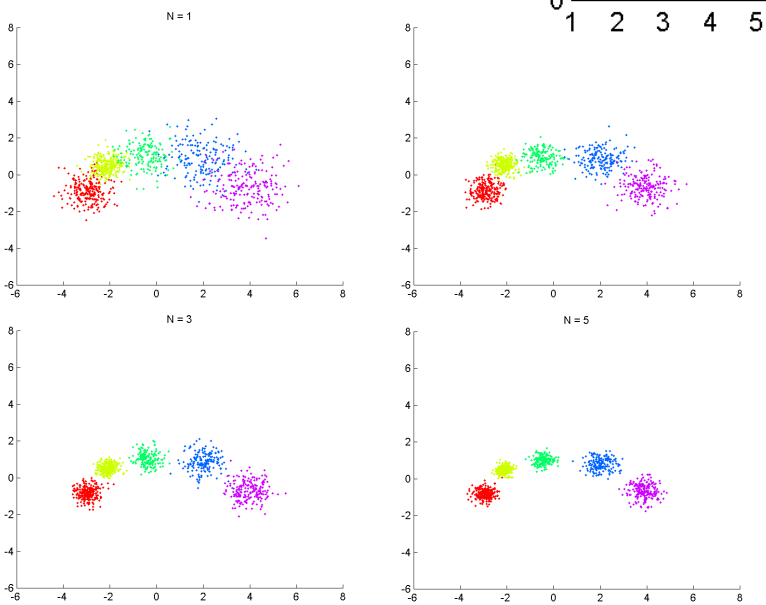
S.O.M. with 5 templates



templates

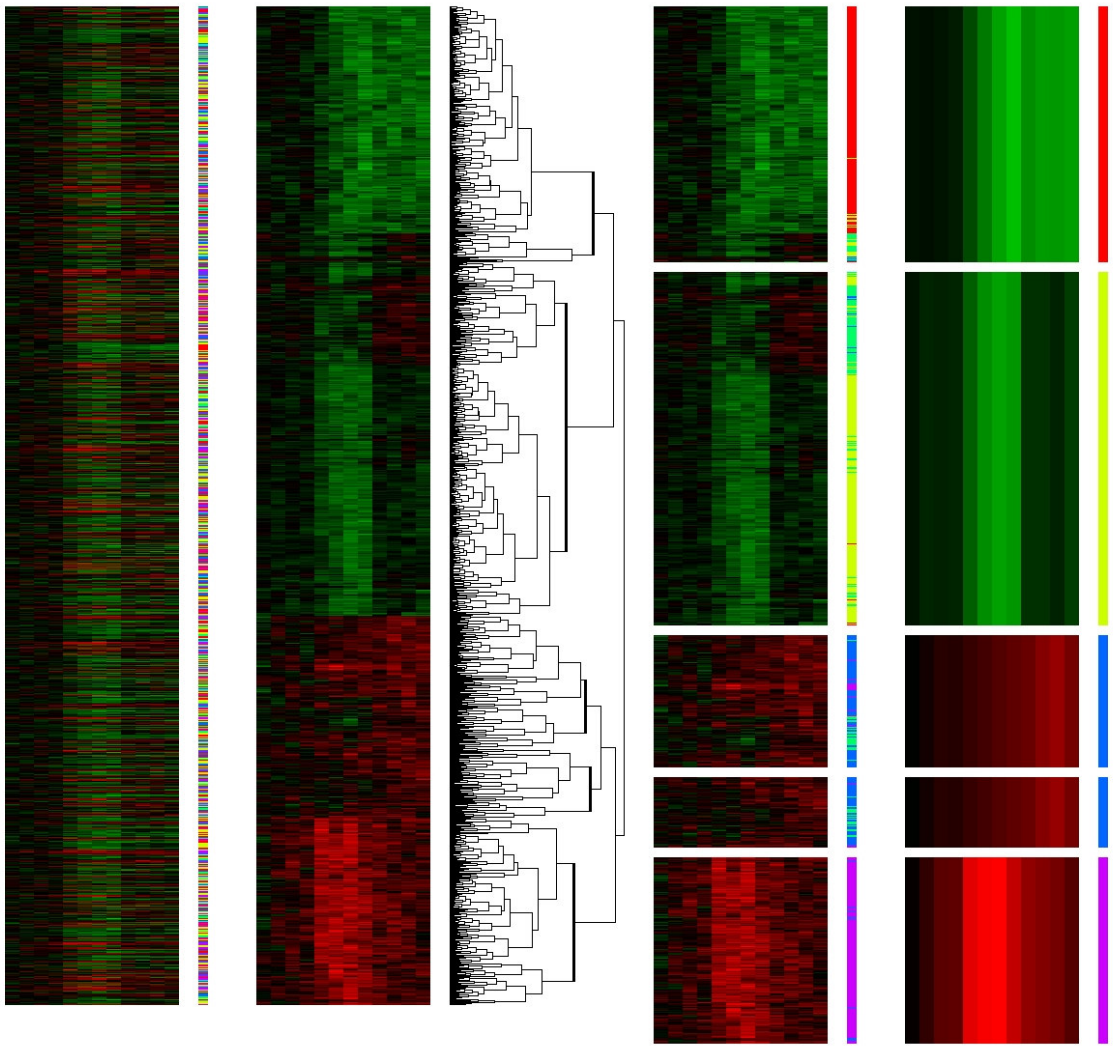


misclassification error



simulated data (PCA plot)

Hierarchical clustering map

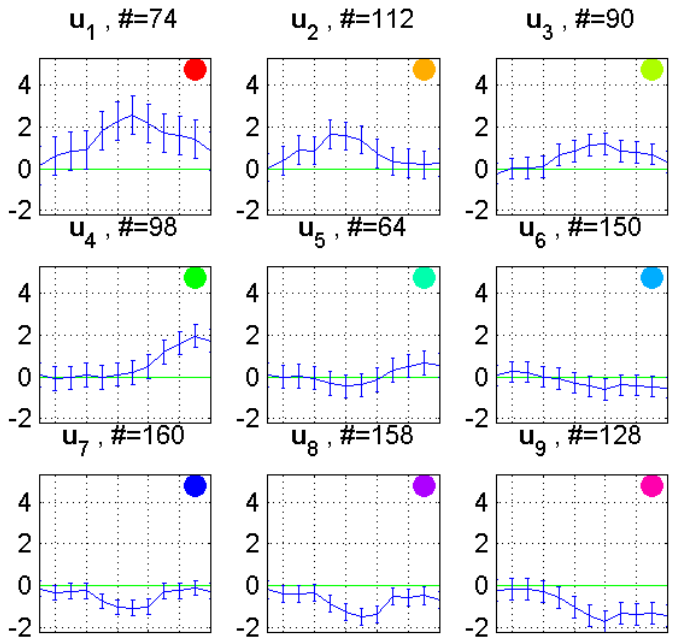


- Correlation based
- N = 3
- 5 classes vs. 4 clusters
- Error = 215 (20.8%)

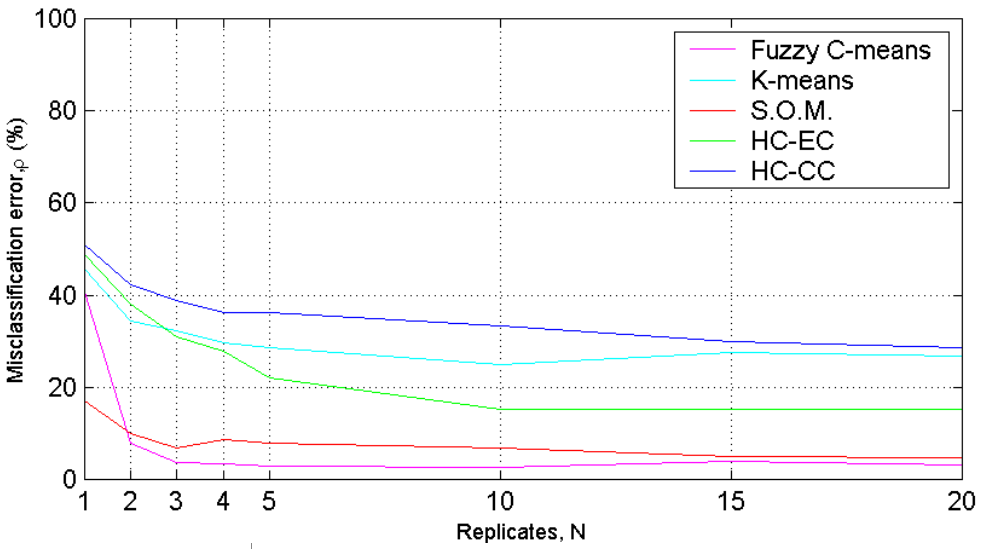
- N = 1
- Error = 352 (34.0%)

- N = 2
- Error = 236 (22.8%)

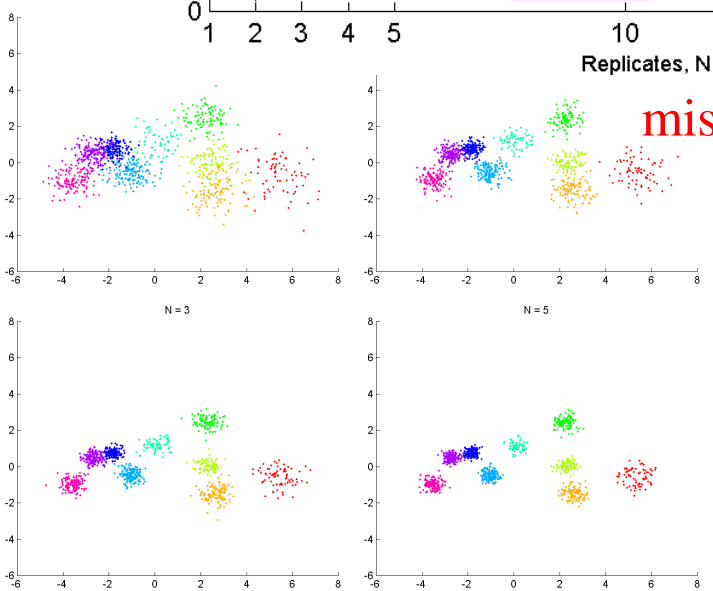
S.O.M. with 9 templates



templates

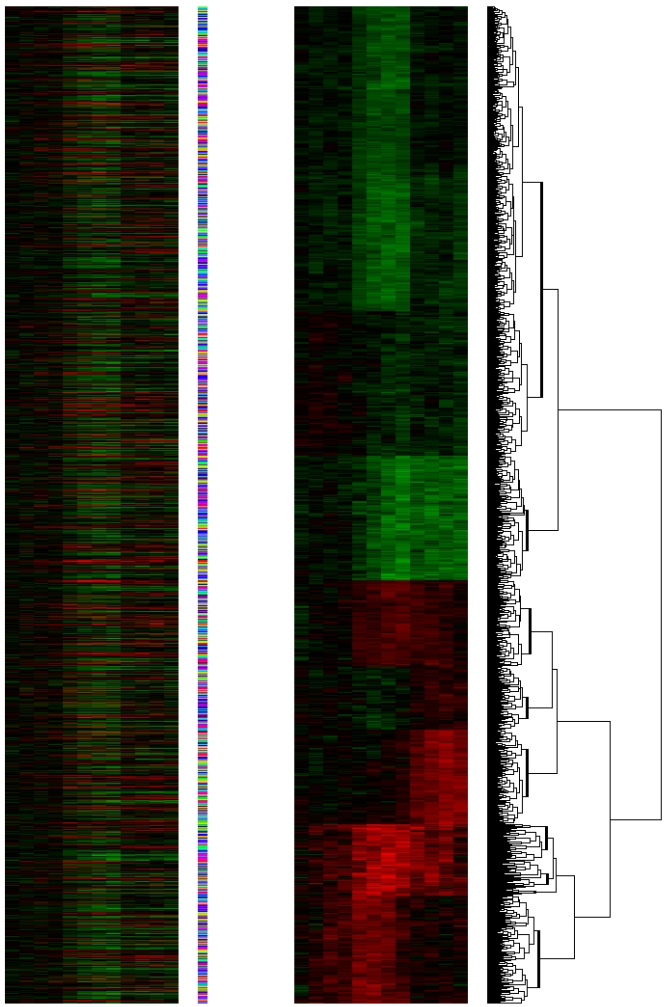


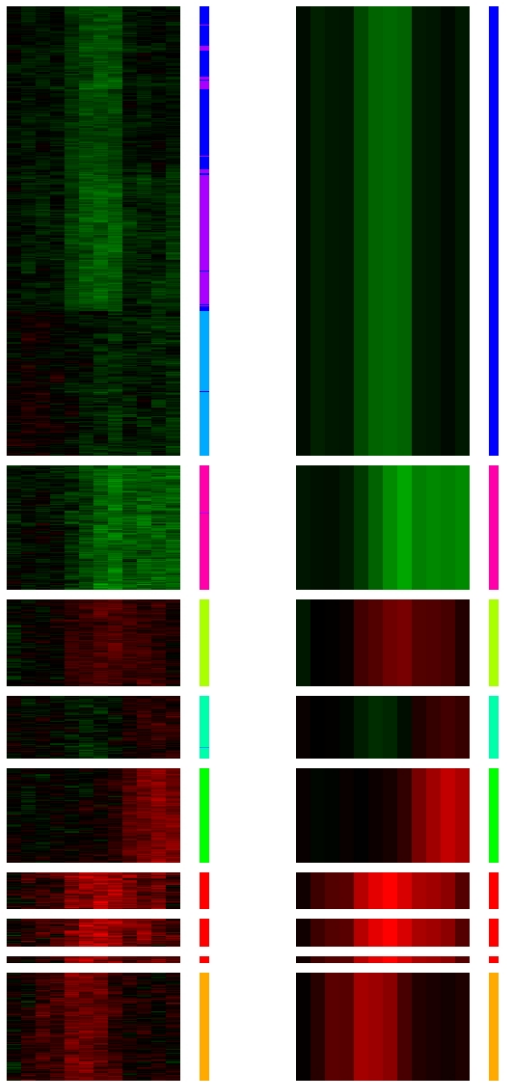
misclassification error



simulated data (PCA plot)

Hierarchical clustering map





- Euclidean based
- N = 3
- 9 classes vs. 7 clusters
- Error = 308 (29.8%)

- N = 1
- Error = 484 (46.7%)

- N=2
- Error = 312 (30.2%)

Closure

- Website for full scale analysis
  - http://gspsnap.tamu.edu/clustering/
    - Username: clustering
    - Password: clustering
  - All five clustering methods analyzed
  - Extensive study on
    - the variance and the replicates
    - lots of graphs and images
    - lots of error measures including confusion matrix