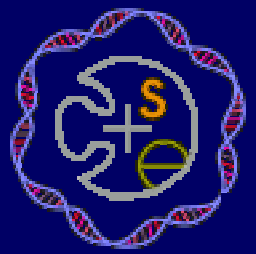
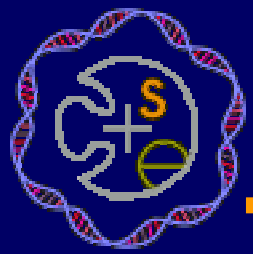


Improved gene selection in microarrays by combining clustering and statistical techniques

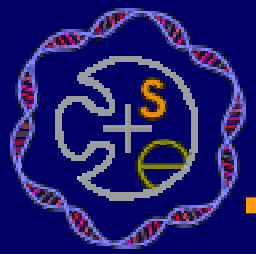
Jochen Jäger
University of Washington
Department of Computer Science

Advisors:
Larry Ruzzo
Rimli Sengupta



Motivation

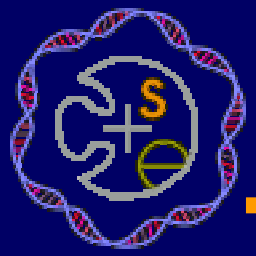
- Think of a complicated question:
- Will it be sunny tomorrow?
- How can you answer it correctly if you DO NOT know the answer?
- Ask around or better, make a poll



Majority vote

- Student: I heard it is supposed to be sunny
- Weather.com: partly cloudy with scattered showers
- Yourself: Considering the past few days and looking outside I would guess it will rain
- TV: partly sunny

- Result: 2 (sunny) : 2 (not sunny)
- Better: Use weights
- Idea: remove redundant answers as well



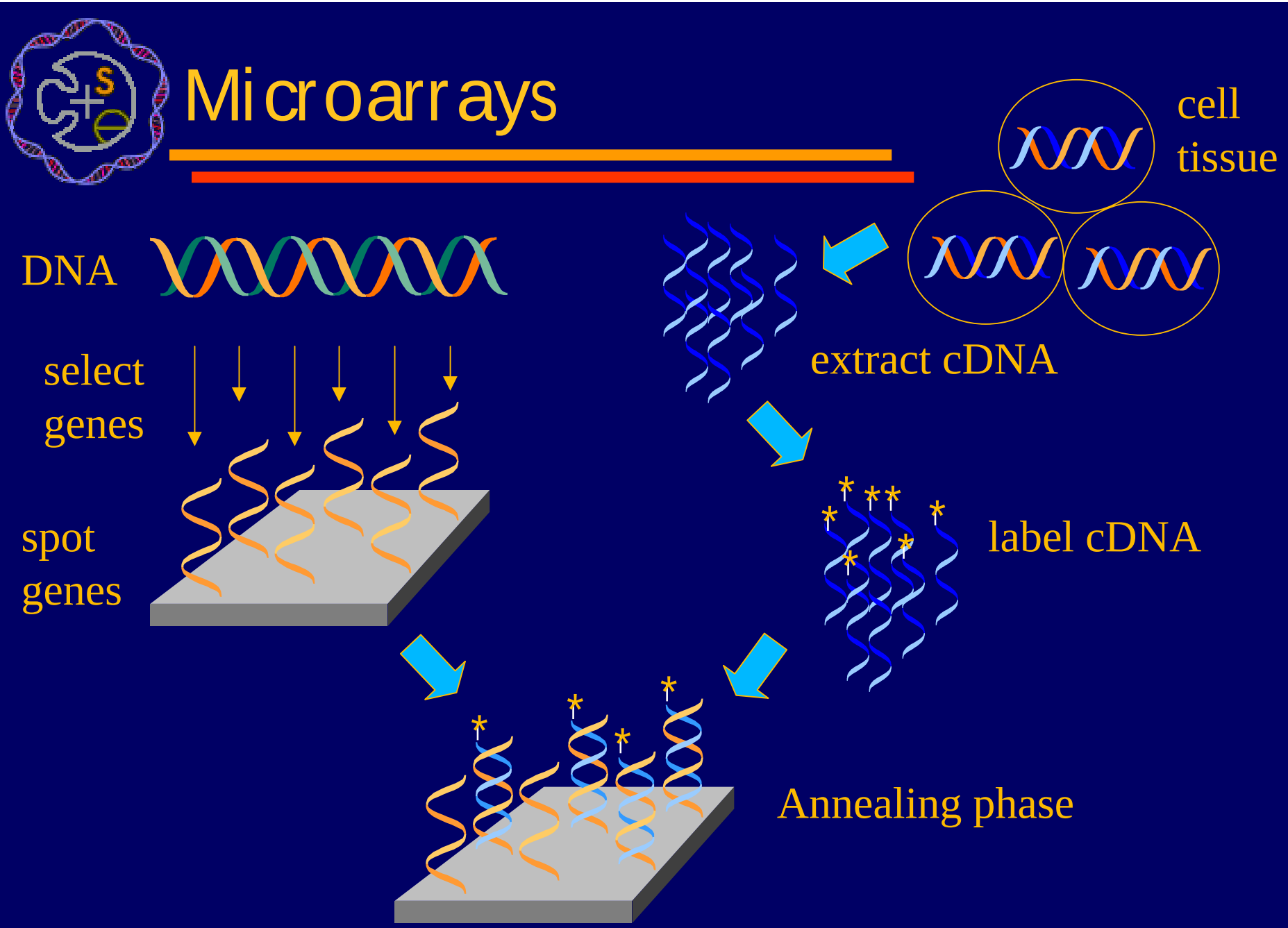
Outline

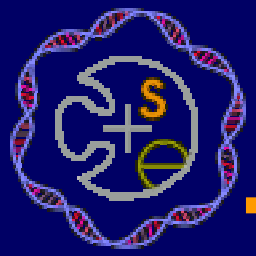
- Motivating example
- Biological background
- Problem statement
- Current solution
- Proposed attack
- Results
- Future work

Biological task

- Find informative genes
- (e.g. genes which can discriminate between cancer and normal)
- Use series of microarrays
- Compare results from different tissues

Microarrays



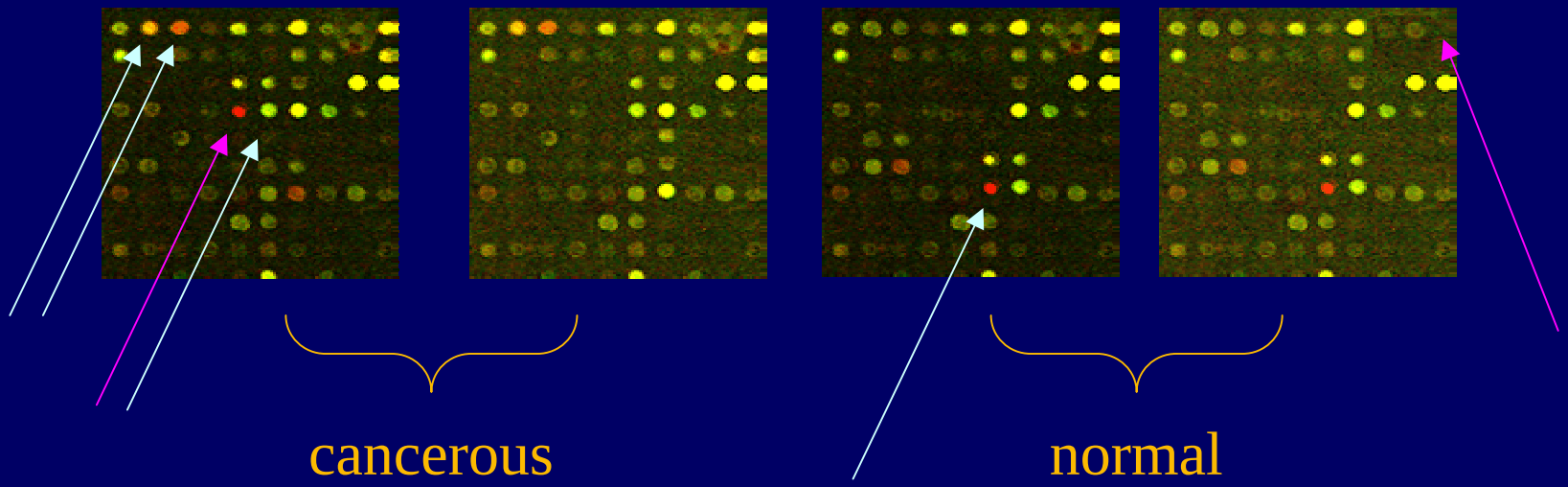


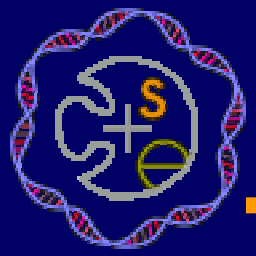
Outline

- Motivating example
- Biological background
- Problem statement
- Current solution
- Proposed attack
- Results
- Future work

Finding informative genes

- Microarrays from different tissues





Outline

- Motivating example
- Biological background
- Problem statement
- Current solution
- Proposed attack
- Results
- Future work

Current solution

- Use a test statistic on all genes

$$t = \frac{\bar{x}_{\text{tumor}} - \bar{x}_{\text{normal}}}{\sqrt{\frac{\sigma^2_{\text{tumor}}}{n_{\text{tumor}}} + \frac{\sigma^2_{\text{normal}}}{n_{\text{normal}}}}}$$

Gene	Tumor 1	Tumor 2	Tumor 3	Normal 1	Normal 2	Normal 3	t-test P-value
A	80	72	85	50	44	15	0.0448836
B	80	72	85	50	44	51	0.0048027
C	71	53	62	57	64	70	0.8024078

- Rank them
- Select top k

Problem with current solution

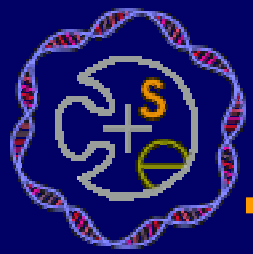
- Each gene independently scored
- Top k ranking genes might be very similar and therefore no additional information gain
- Reason: genes in similar pathways probably all have very similar score
- What happens if several pathways involved in perturbation but one has main influence
- Possible to describe this pathway with fewer genes

Problem of redundancy

Accession Number	Adenoma 1	Adenoma 2	Adenoma 3	Adenoma 4	Normal 1	Normal 2	Normal 3	Normal 4	t-test P-value
AF001548	54.55	43.93	55.69	28.47	1354.36	1565.42	1459.48	1612.85	0.00012
M12125	35.9	46.64	35.73	35.27	642.46	577.81	580.5	707.35	0.00028
X13839	46.16	47.72	26.79	17	652.66	653.14	546.12	720.43	0.0003
X15882	13.52	15.73	27.32	16.15	209.3	209.64	221.24	267.43	0.0004
AB002533	659.25	958.82	812.77	786.24	407.91	558.33	529.68	379.84	0.00557
M93651	40.1	54.77	39.93	40.37	8.74	21.07	14.45	32.94	0.01038

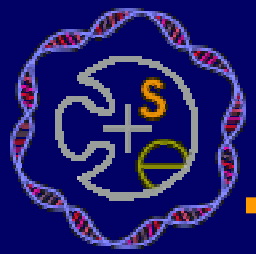
Top 3 genes highly correlated!

	AF001548	M12125	X13839	X15882	AB002533	M93651
AF001548	1					
M12125	0.99	1				
X13839	0.991	0.996	1			
X15882	0.992	0.995	0.988	1		
AB002533	-0.87	-0.898	-0.891	-0.888	1	
M93651	-0.8	-0.802	-0.789	-0.776	0.808	1



Outline

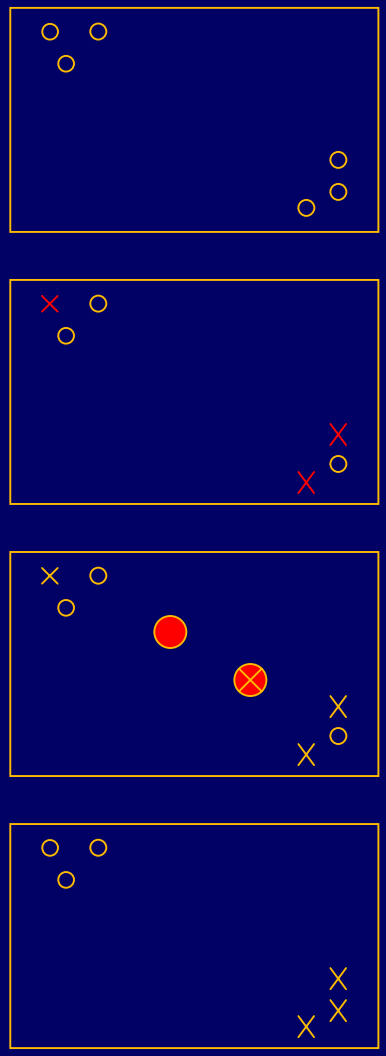
- Motivating example
- Biological background
- Problem statement
- Current solution
- Proposed attack
- Results
- Future work



Proposed solution

- Several possible approaches
  - next neighbors
  - correlation
  - euclidean distance
- Approach: instead use clustering
- Advantages using clustering techniques
  - natural embedding
  - many different distance functions possible
  - different shapes, models possible

Hard clustering – k-means



Randomly assign cluster to each point

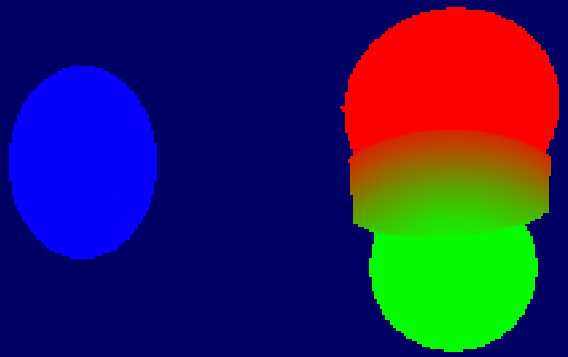
Find centroids

Reassign points to nearest center

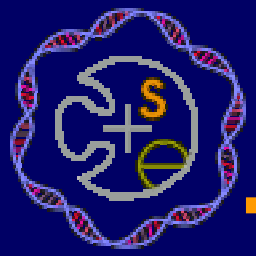
Iterate until convergence

Soft - Fuzzy Clustering

instead of hard assignment,
probability for each cluster



Very similar to k-means but fuzzy softness factor m (between 1 and infinity) determines how hard the assignment has to be



Fuzzy examples

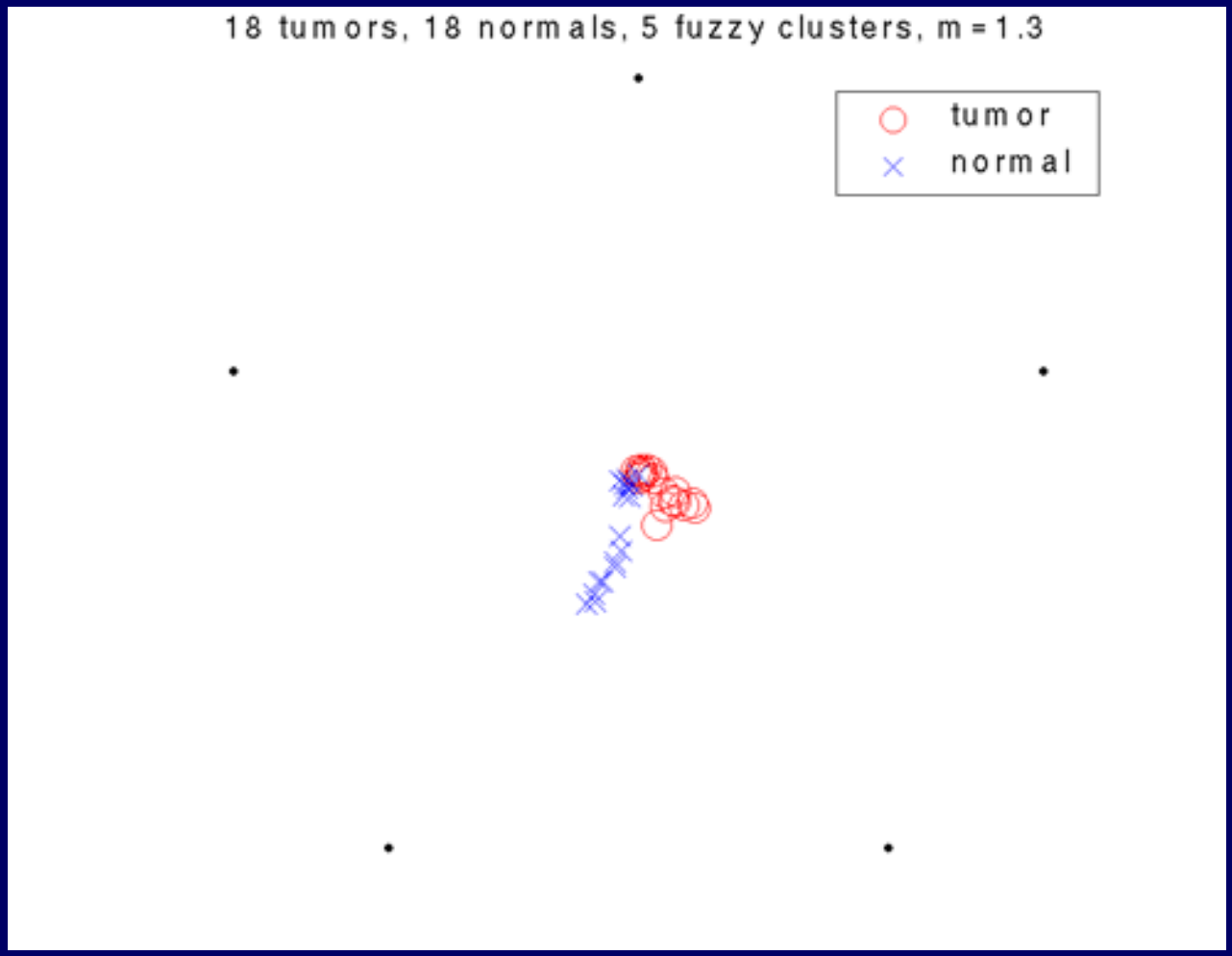
Nottermans carcinoma dataset:

18 colon adenocarcinoma and 18 normal tissues

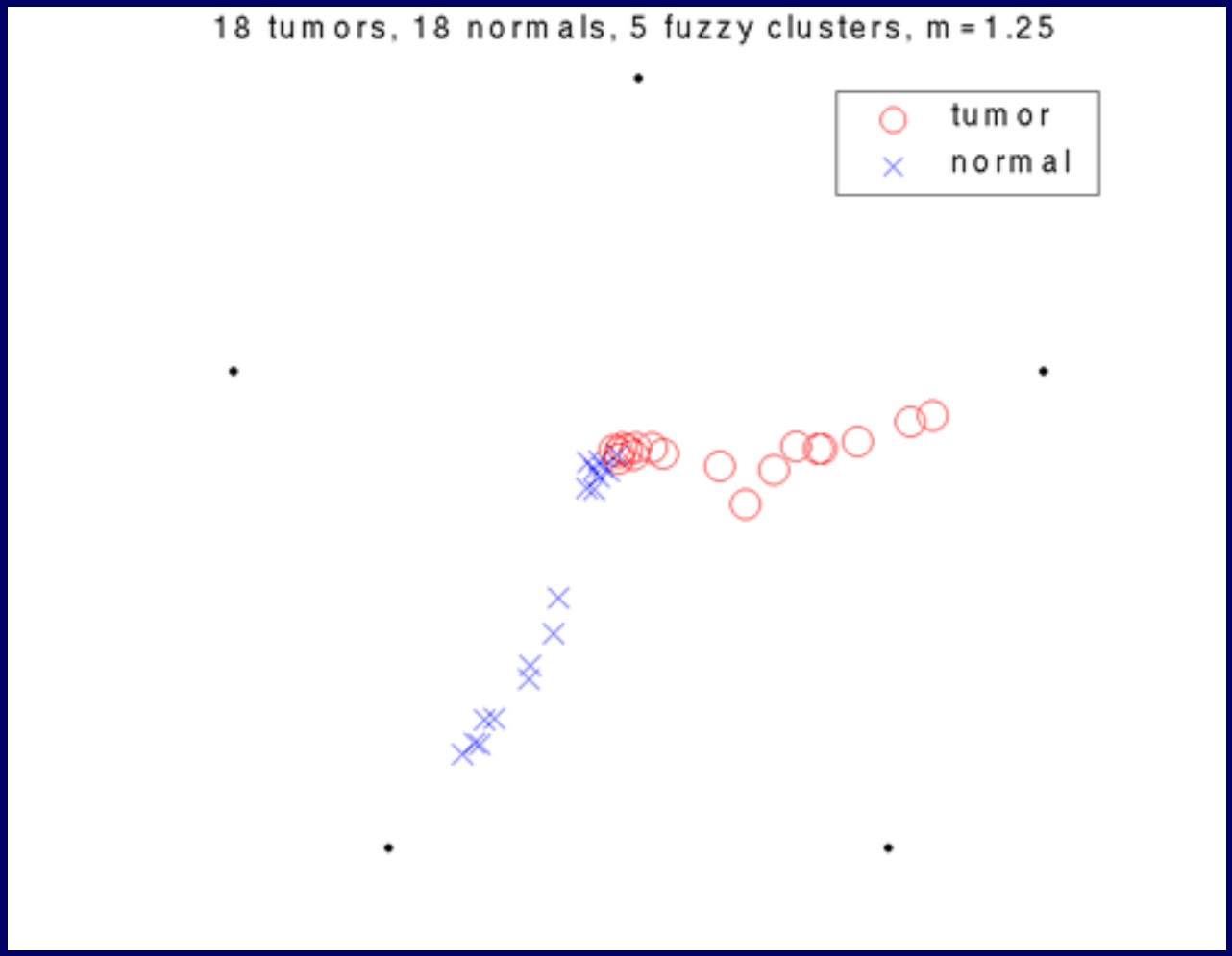
data from 7457 genes and ESTs

cluster all 36 tissues

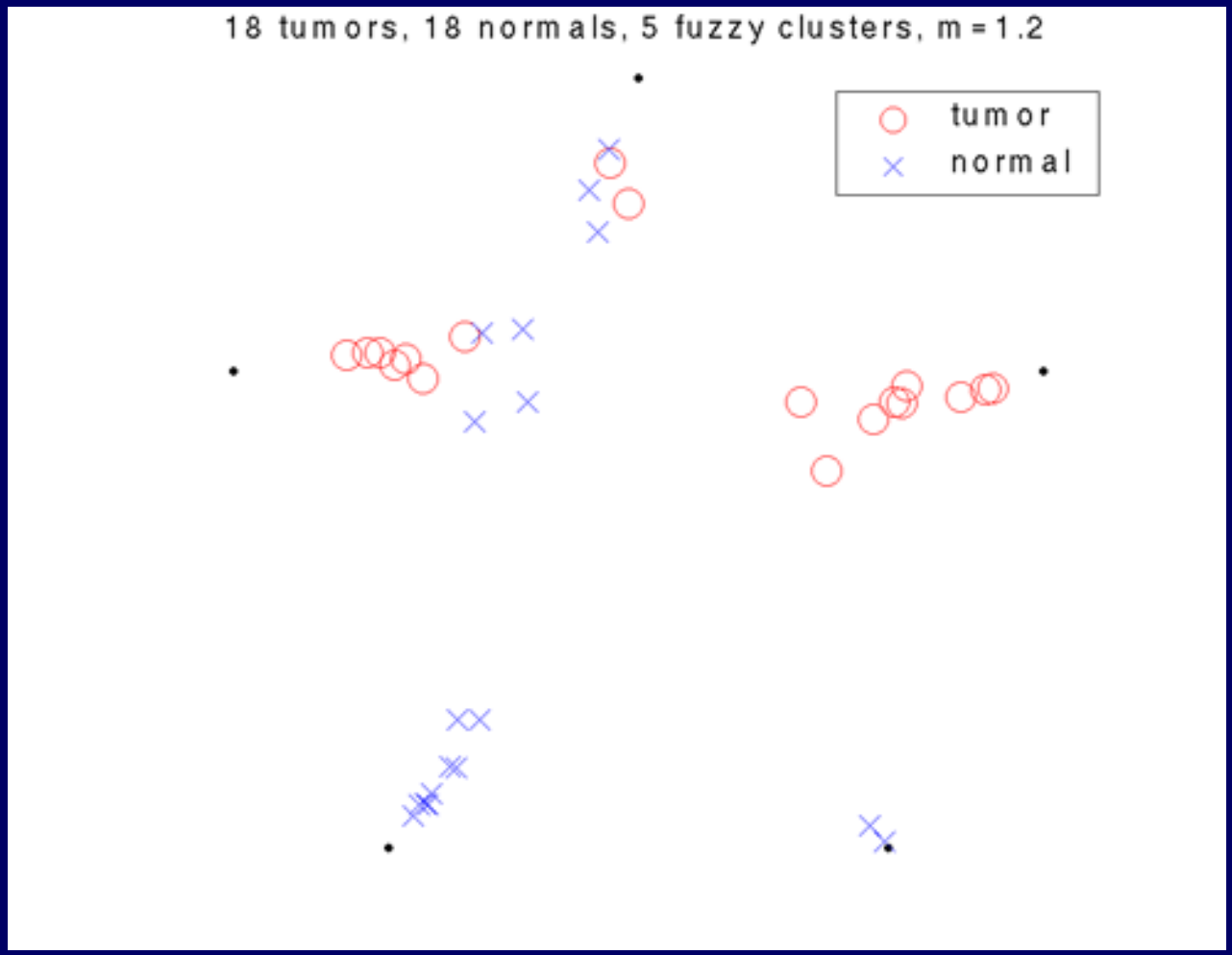
Fuzzy softness 1.3



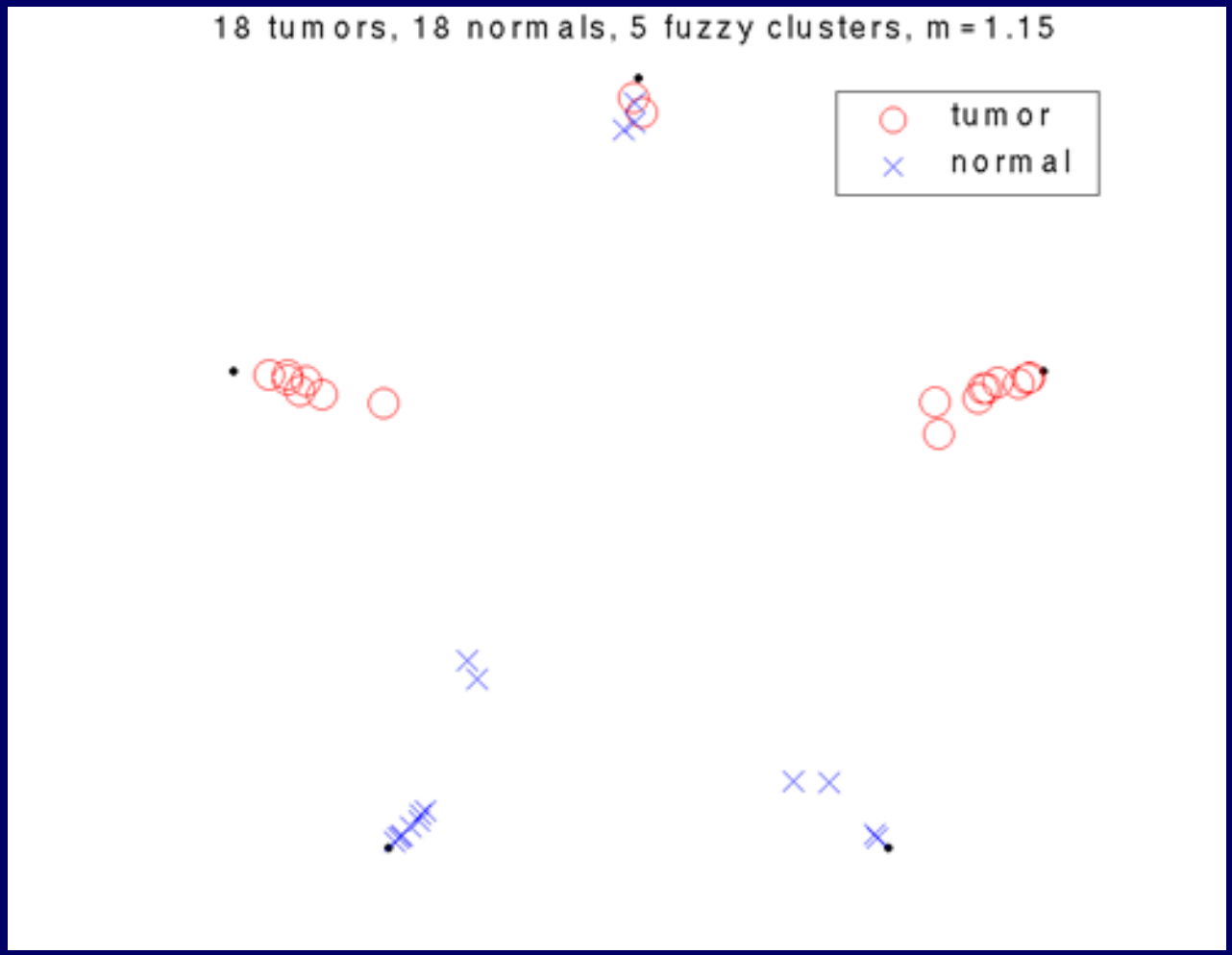
Fuzzy softness 1.25



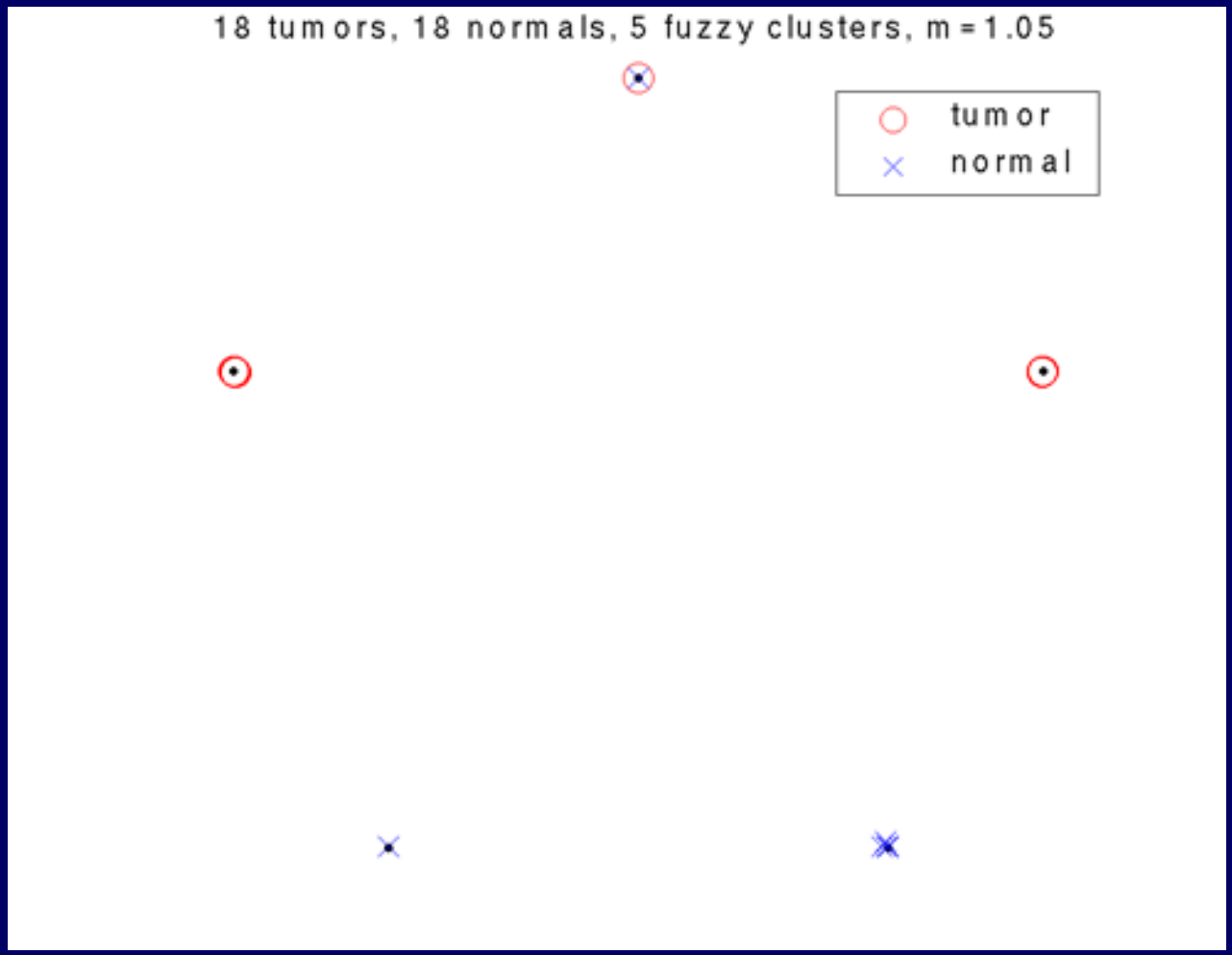
Fuzzy softness 1.2

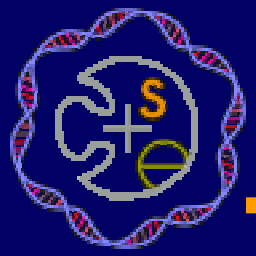


Fuzzy softness 1.15



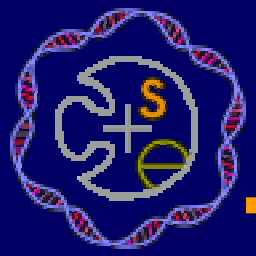
Fuzzy softness 1.05





Selecting genes from clusters

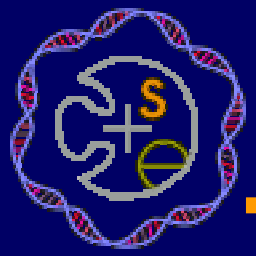
- Two way filter: exclude redundant genes, select informative genes
- Get as many pathways as possible
- Consider cluster size and quality as well as discriminative power



How many genes per cluster?

- Constraints:
  - minimum one gene per cluster
  - maximum as many as possible
- Take genes proportionally to cluster quality and size of cluster
- Take more genes from bad clusters
- Smaller quality value indicates tighter cluster

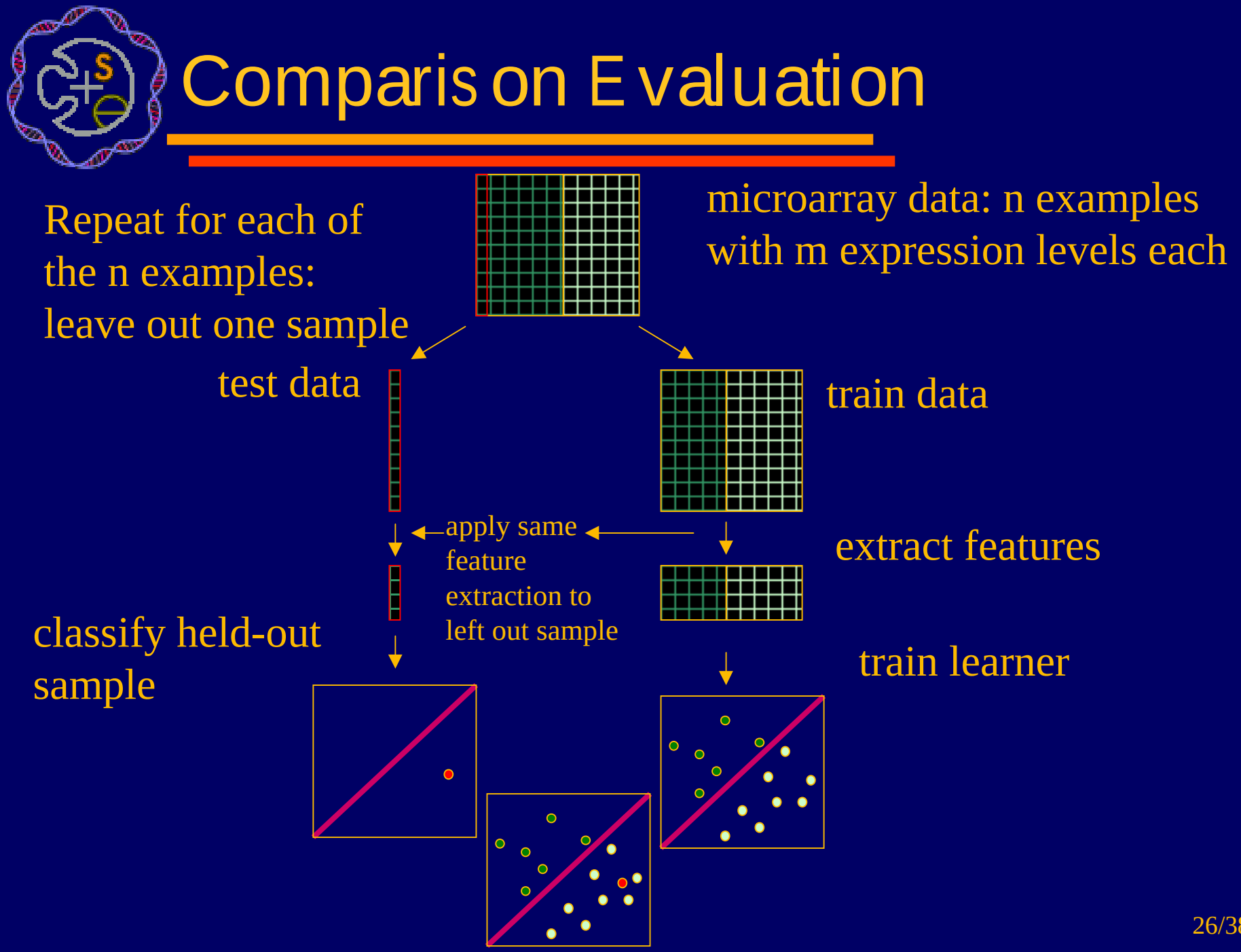
- Quality for k-means: sum of intra cluster distance
- Quality for fuzzy c-means: avg cluster membership probability



Which genes to pick?

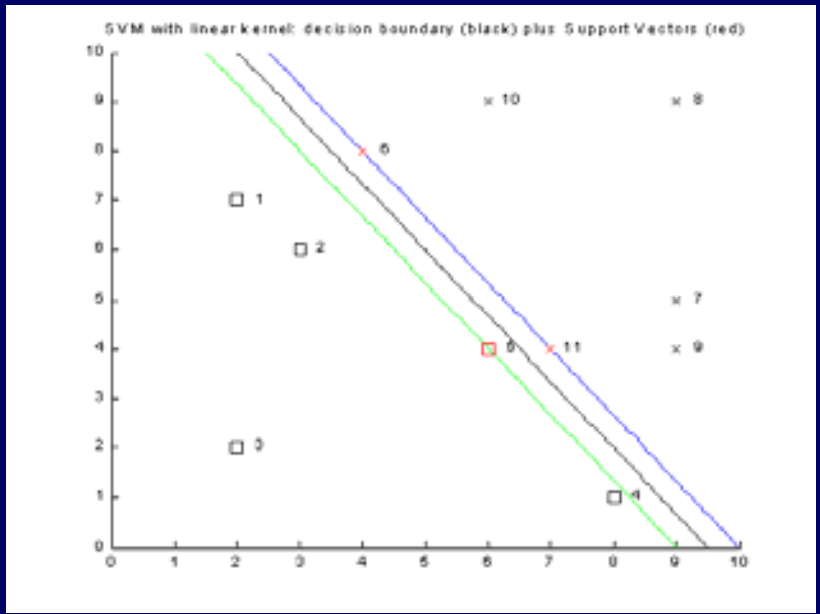
- Choices:
  - Genes closest to center
  - Genes farthest away
  - Sample according to probability function
  - Genes with best discriminative power

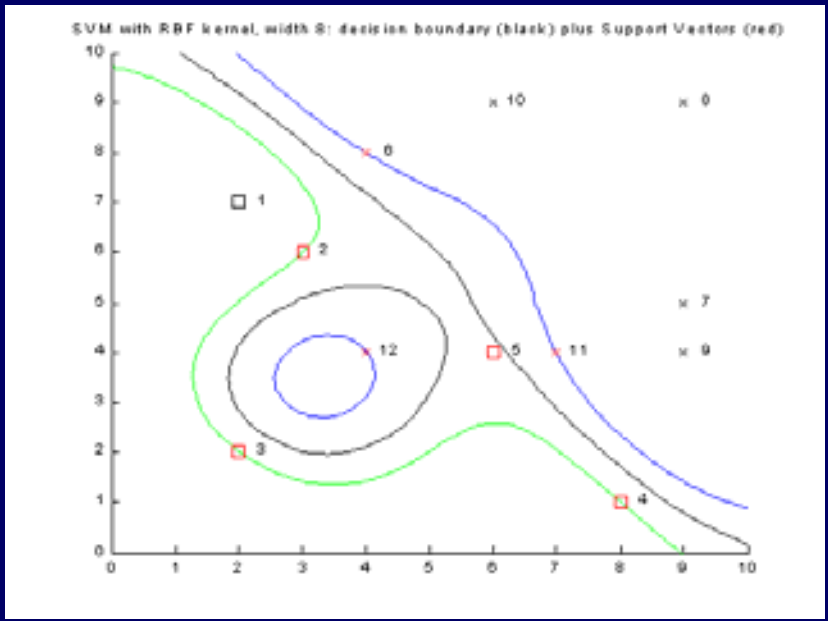
Comparison Evaluation



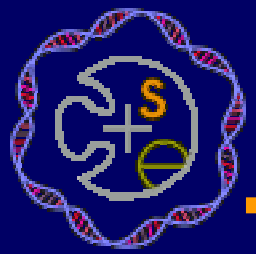
Support Vector machines

- Finds separating hyperplane with maximal distance to closest training example



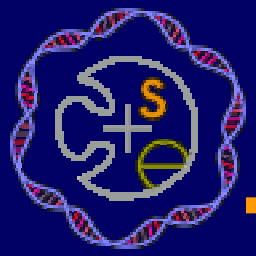


- Advantages:
  - avoids overfitting
  - can handle higher order interactions and noise using kernel functions and soft margin



Outline

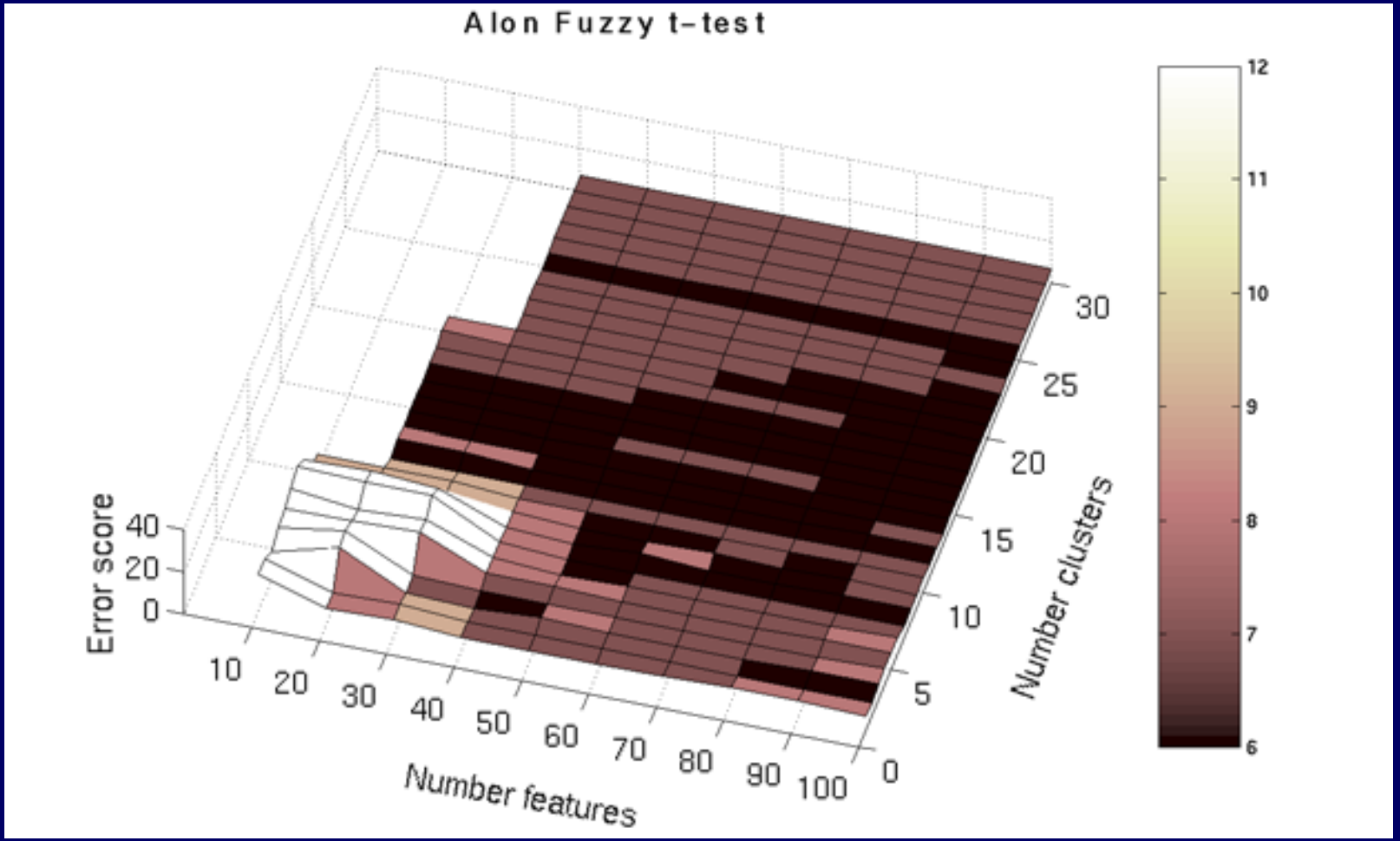
- Motivating example
- Biological background
- Problem statement
- Current solution
- Proposed attack
- Results
- Future work



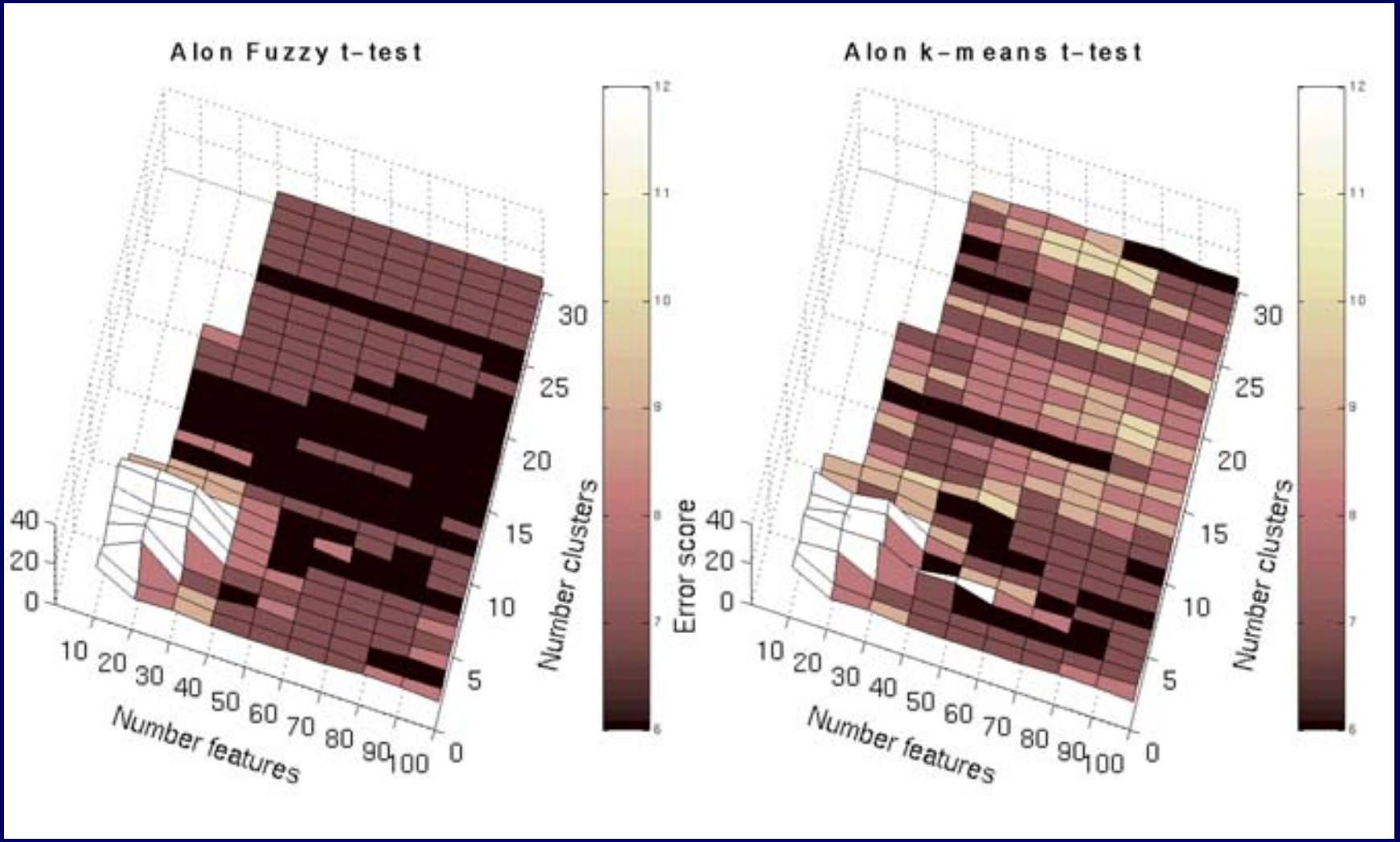
Experimental setup

- Datasets:
  - Alons Colon (40 tumor and 22 normal colon adenocarcinoma tissue samples)
  - Golubs Leukemia (47 ALL, 25 AML)
  - Nottermans Carcinoma and Adenoma (18 adenocarcinoma, 4 adenomas and paired normal tissue)
- Experimental setup:
  - calculate LOOCV using SVM on feature subsets
  - do this for feature size 10-100 (in steps of 10) and 1-30 clusters

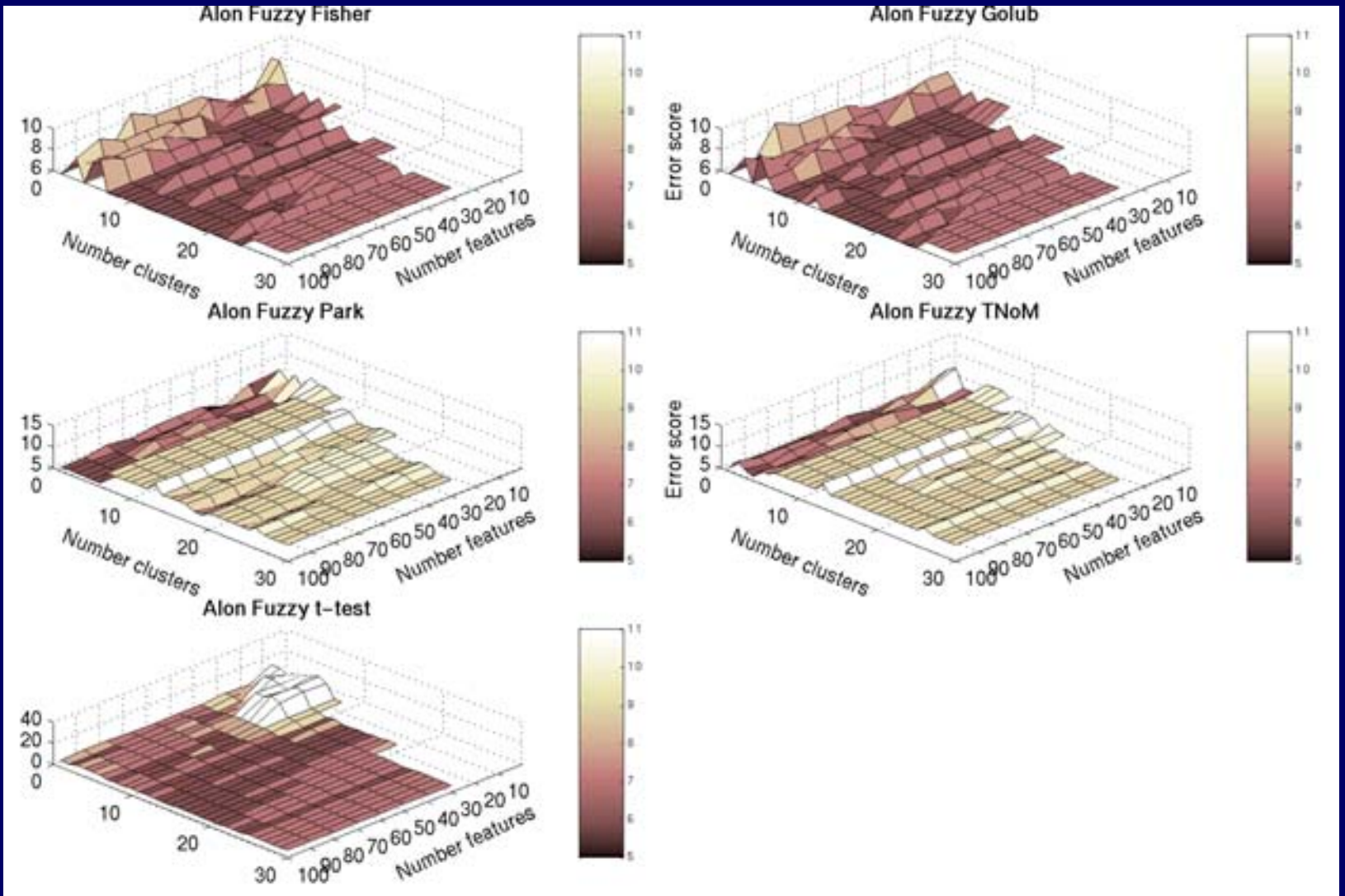
Results


Alon Fuzzy t-test
Error score
40
20
0
10
20
30
40
50
60
70
80
90
100
Number features
0
5
10
15
20
25
30
Number clusters
12
11
10
9
8
7
6

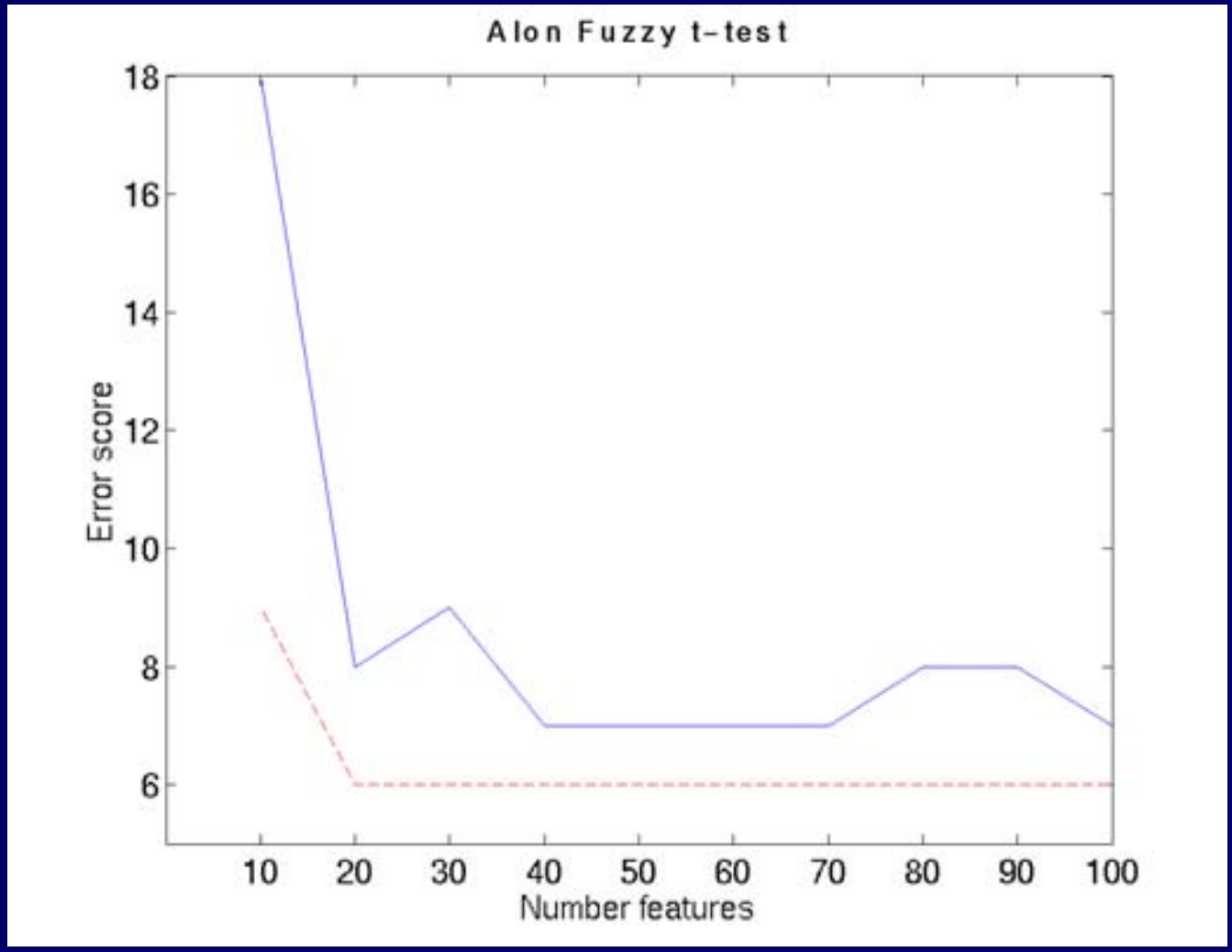
fuzzy c-means vs k-means



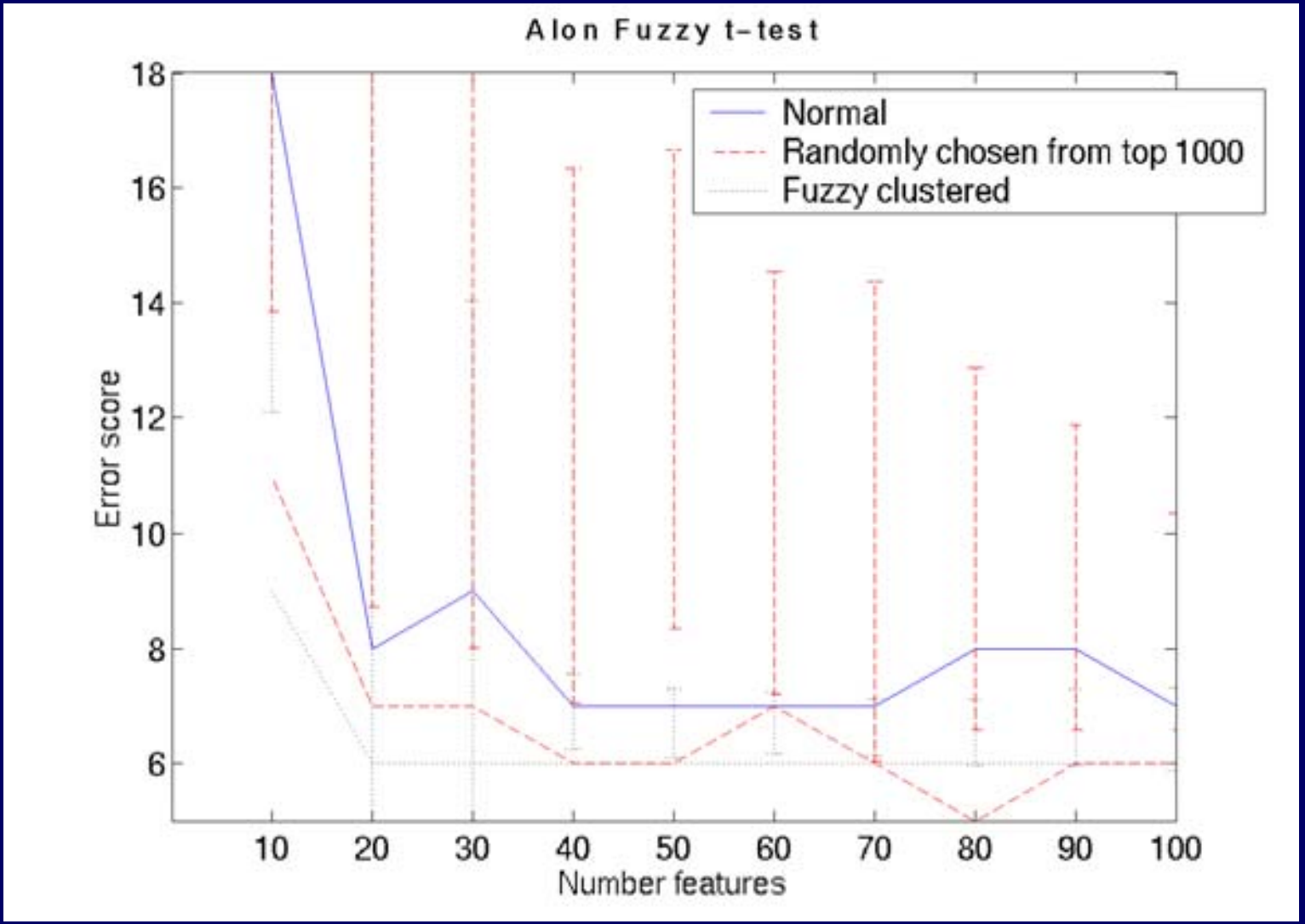
Different test-statistics

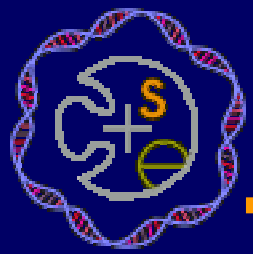


Comparing best results



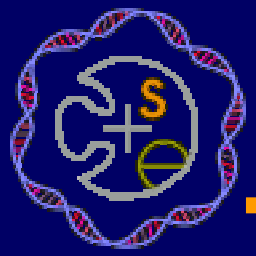
How about randomly choosing?





Related work

- Tusher, Tibshirani and Chu (2001): Significance analysis of microarrays applied to the ionizing radiation response, PNAS 2001 98: 5116-5121
- Ben-Dor, A., L. Bruhn, N. Friedman, I. Nachman, M. Schummer, and Z. Yakhini (2000). Tissue classification with gene expression profiles. In Proceeding of the fourth annual international conference on computational molecular biology, pp. 54-64
- Park, P.J., Pagano, M., Bonetti, M.: A nonparametric scoring algorithm for identifying informative genes from microarray data. Pac Symp Biocomput :52-63, 2001.
- Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh M, Downing JR, Caligiuri MA, Bloomfield CD, and Lander 18 ES. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 286: 531-537, 1999.
- J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, T. Poggio, and V. Vapnik. Feature selection for SVMs . In Sara A Solla, Todd K Leen, and Klaus-Robert Muller, editors, Advances in Neural Information Processing Systems 13. MIT Press, 2001. 11

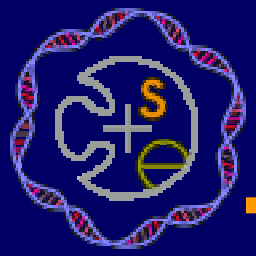


Outline

- Motivating example
- Biological background
- Problem statement
- Current solution
- Proposed attack
- Results
- Future work

Future work

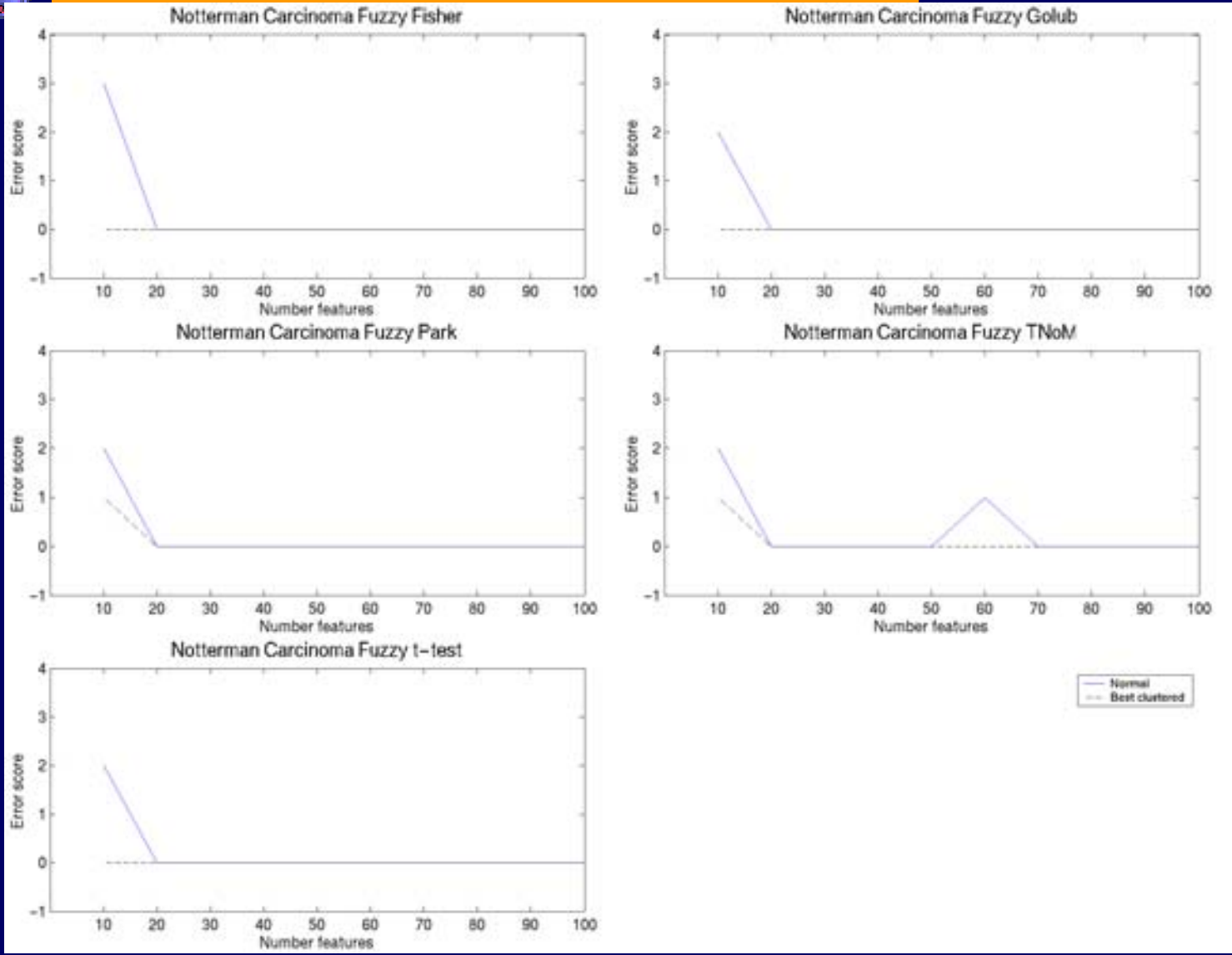
- Problem how to find best parameters (model selection, model based clustering, BIC)
- Combine good solutions
- Incorporate overall cluster discriminative power into quality score
- Use of non integer error score
- ROC analysis



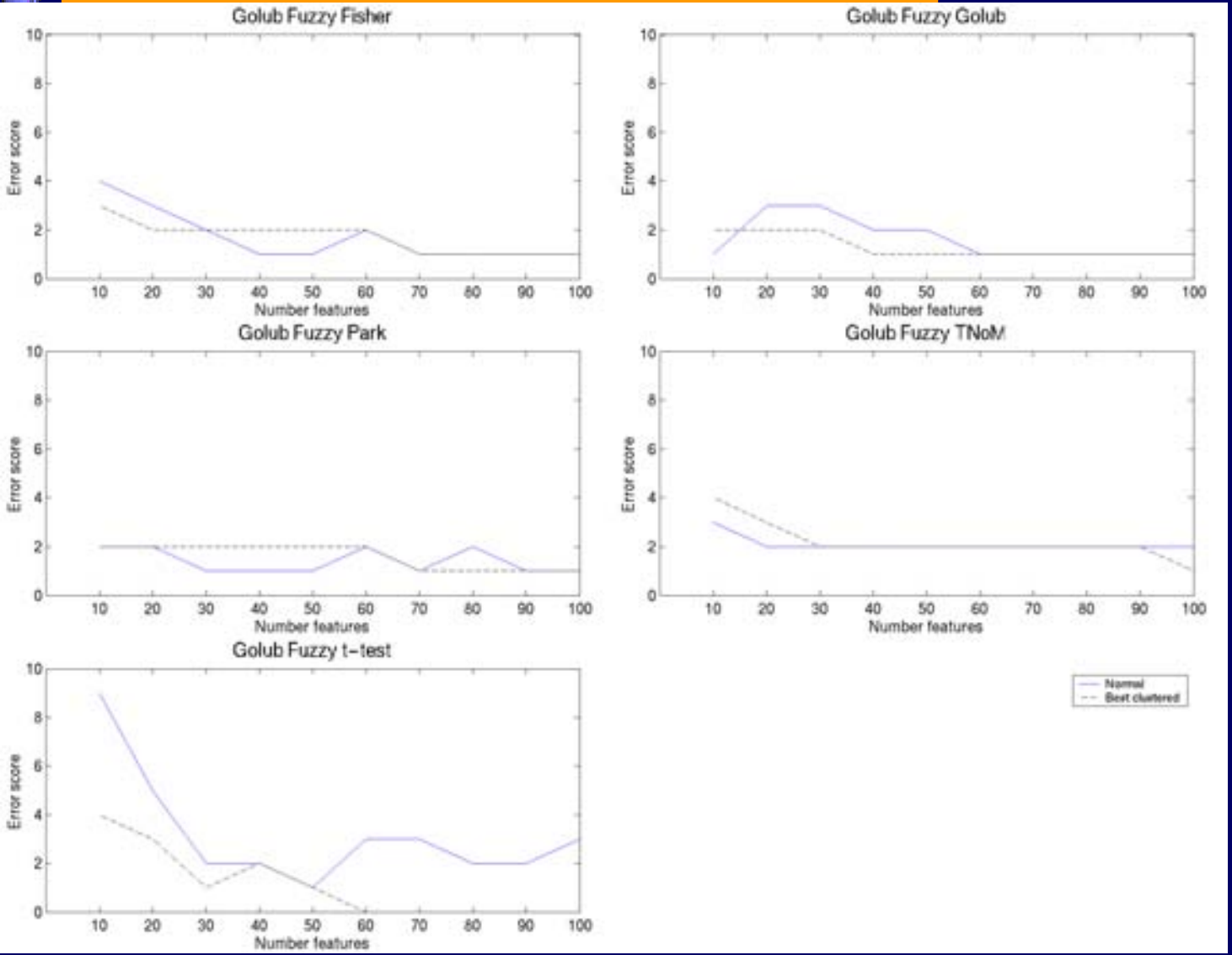
Summary

- Used clustering as a pre-filter for feature selection in order to get rid of redundant data
- Defined a quality measurement for clustering techniques
- Incorporated cluster quality, size and statistical property into feature selection
- Improved LOOCV error for almost all feature sizes and different related tests

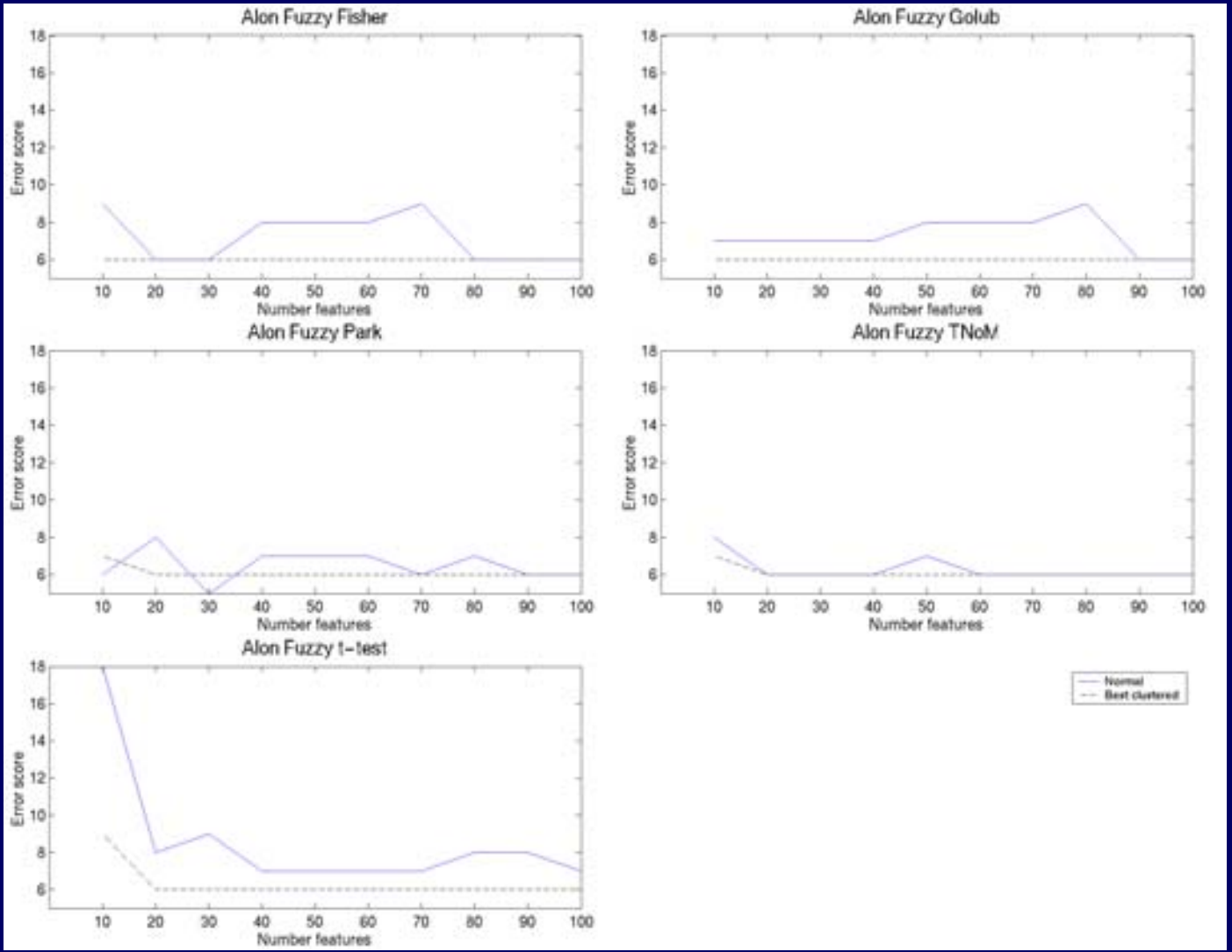
Result Notterman



Result Golub



Result Alon


Alon Fuzzy Fisher
Alon Fuzzy Golub
Alon Fuzzy Park
Alon Fuzzy TNoM
Alon Fuzzy t-test
Error score
Number features
Normal
Best clustered

Result Alon 2