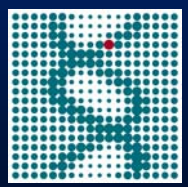
Clustering of Gene Expression Time Courses:
Methods and Validity of Solutions

Ivan G. Costa Filho
Alexander Schliep



Computational Biology Department
Max-Planck-Institute for Molecular Genetics, Berlin

Goal

(1) Clustering of gene expression time-courses

(2) *Validating clusterings* of genes

(3) Methods for clustering *heterogeneous data*

gene expression **+** functional annotation, regulatory region information, protein-protein interactions ...

(1) Clustering Method

Time-course models

Biological Truism

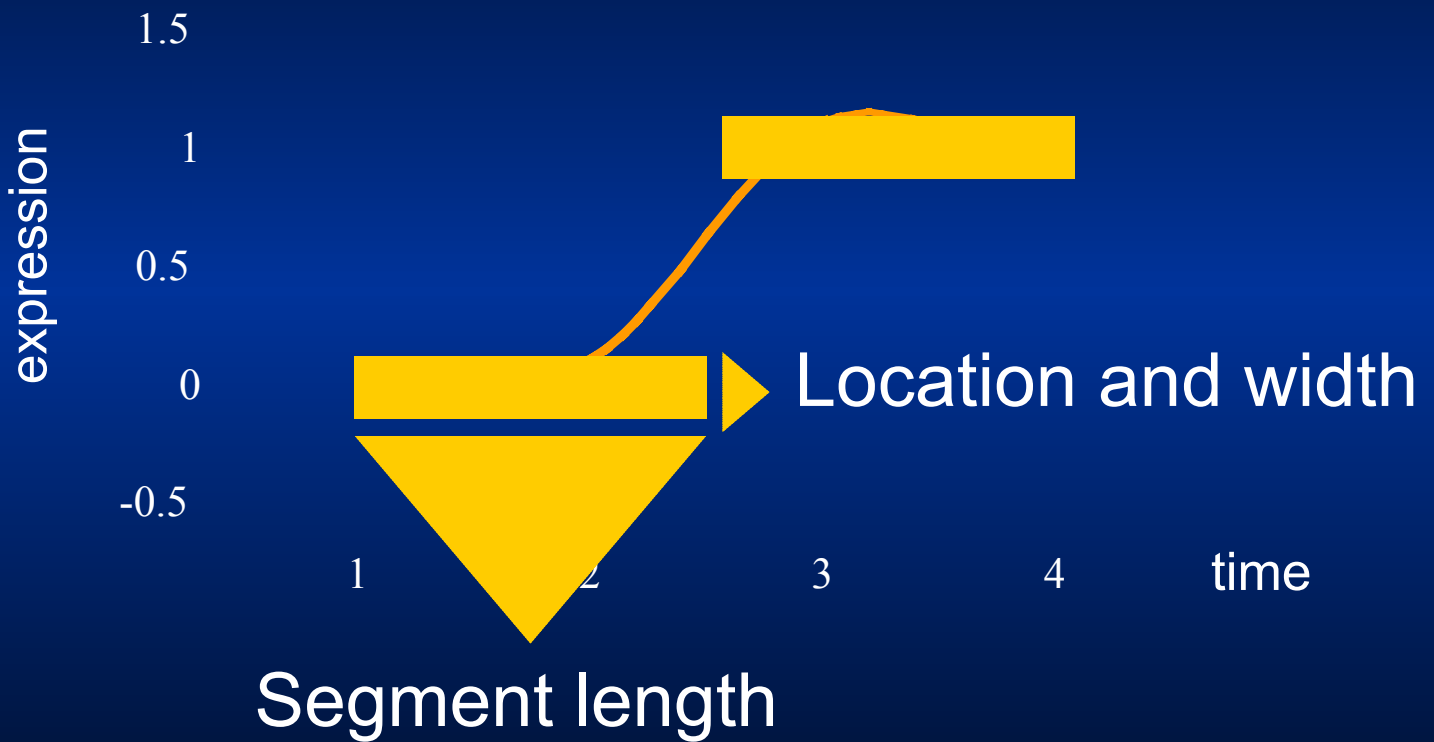
- Many genes have
  - multiple functions
  - are involved in several regulatory networks
- *Unique assignment to groups dubious*

Method Outline

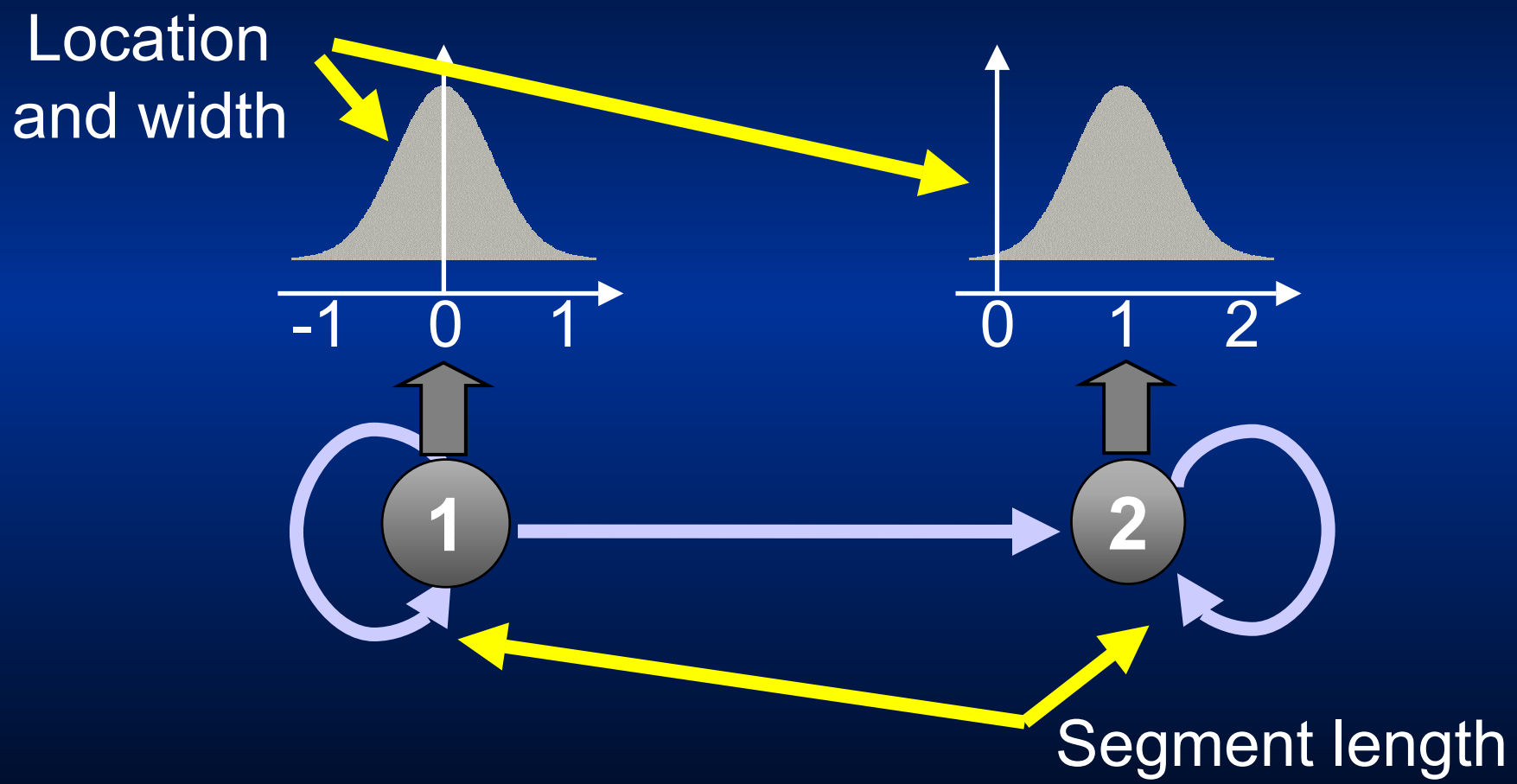
1. Define a class of statistical models for time-courses
2. Combine them in a mixture model
3. Decode the mixture to infer groups

Time-course models

Example: Up-regulation

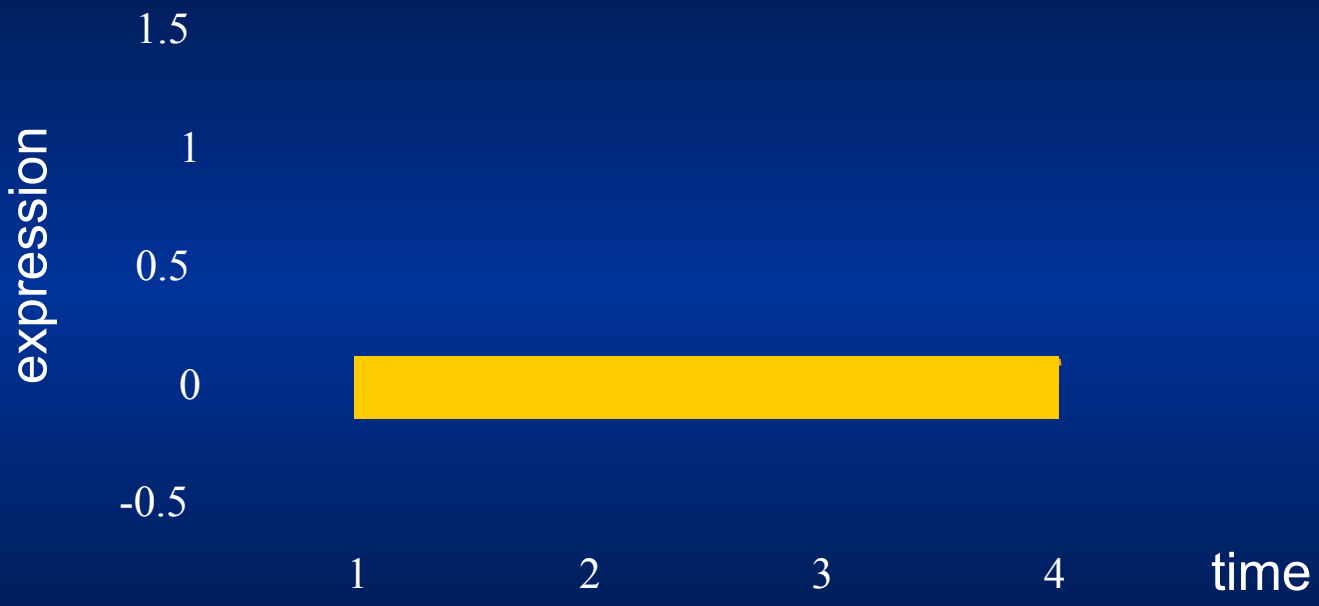


Prototype: Up-regulation

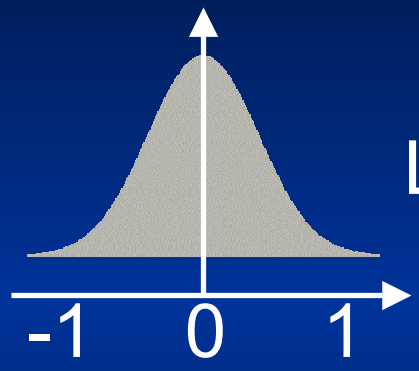


Hidden Markov Model (HMM)

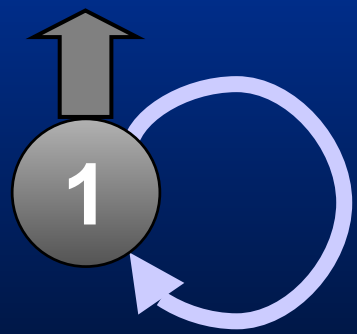
Example: Constant



Prototype: Constant



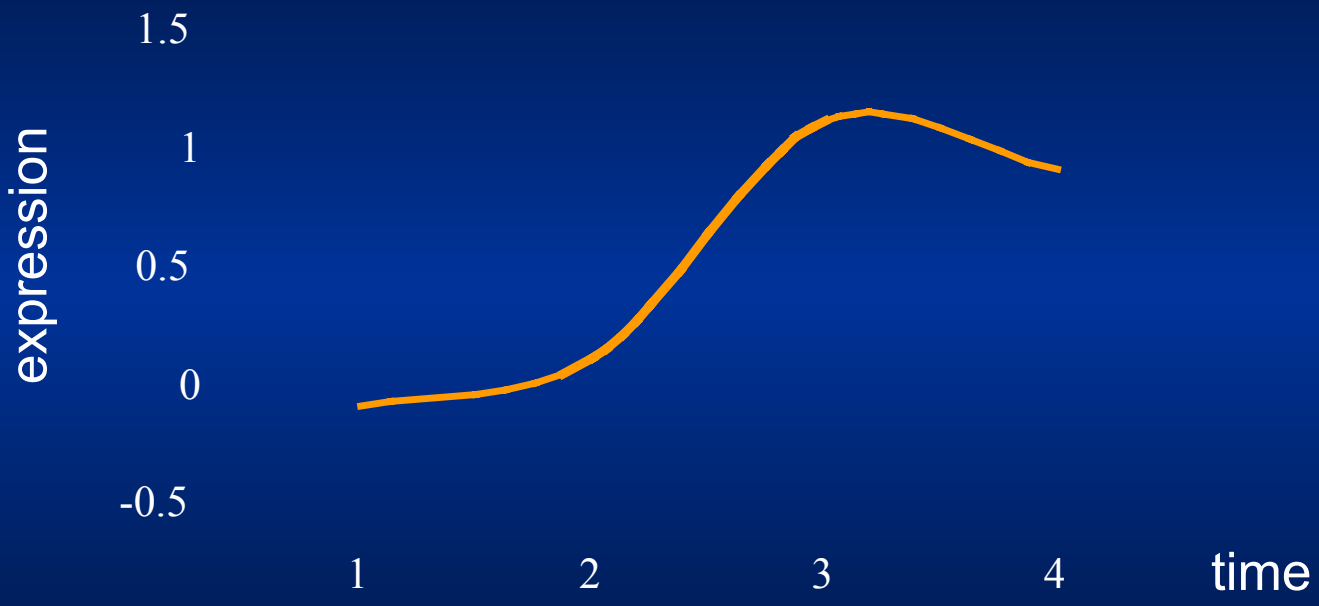
Location and width



Segment length

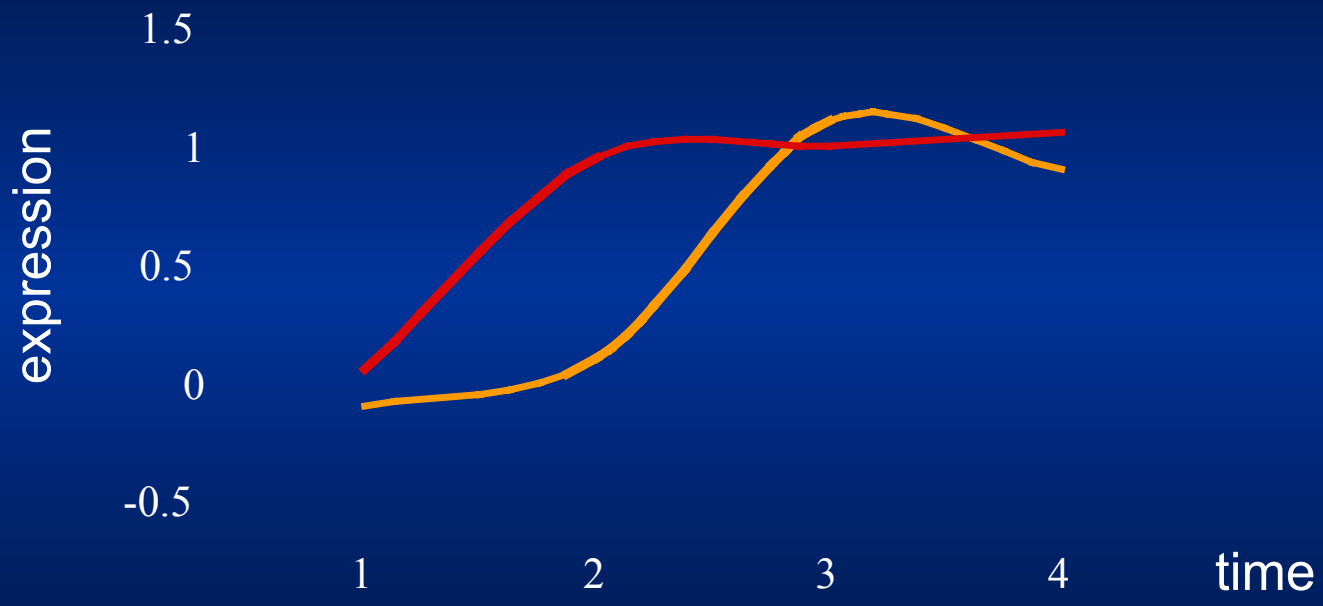
Hidden Markov Model (HMM)

Viterbi Path: Up-regulation



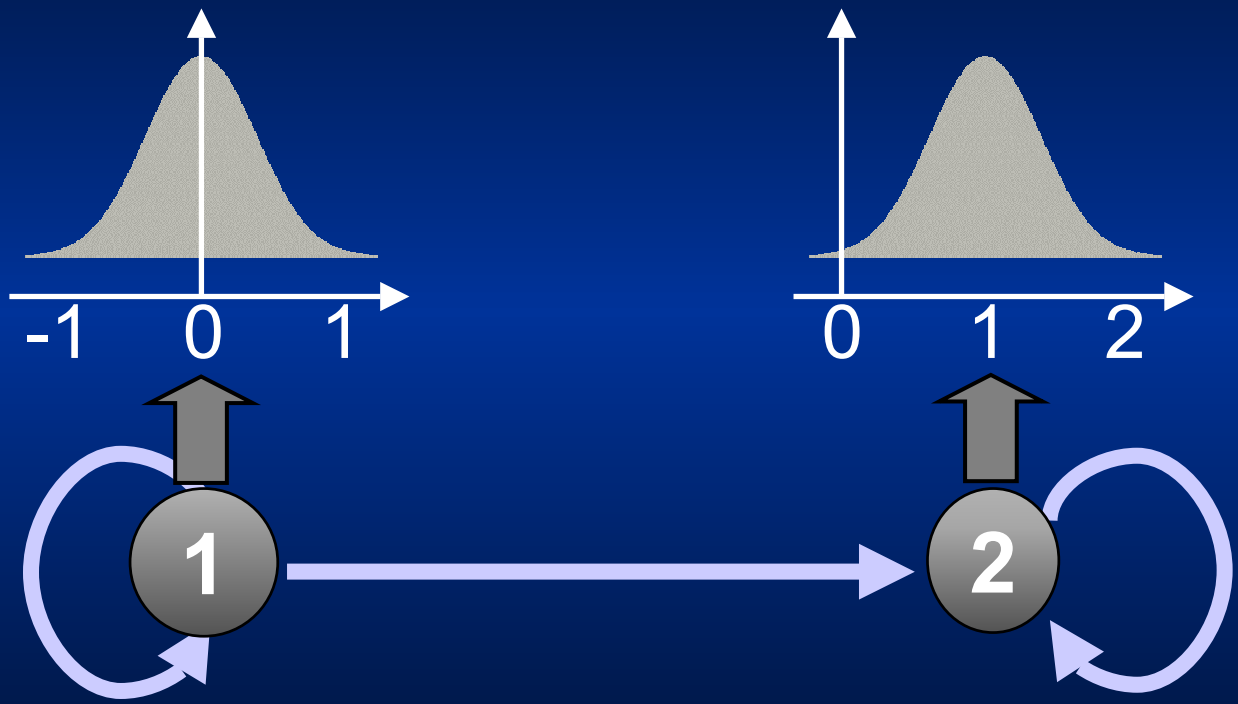
State **1** **1** **2** **2**

Viterbi Path: Up-regulation

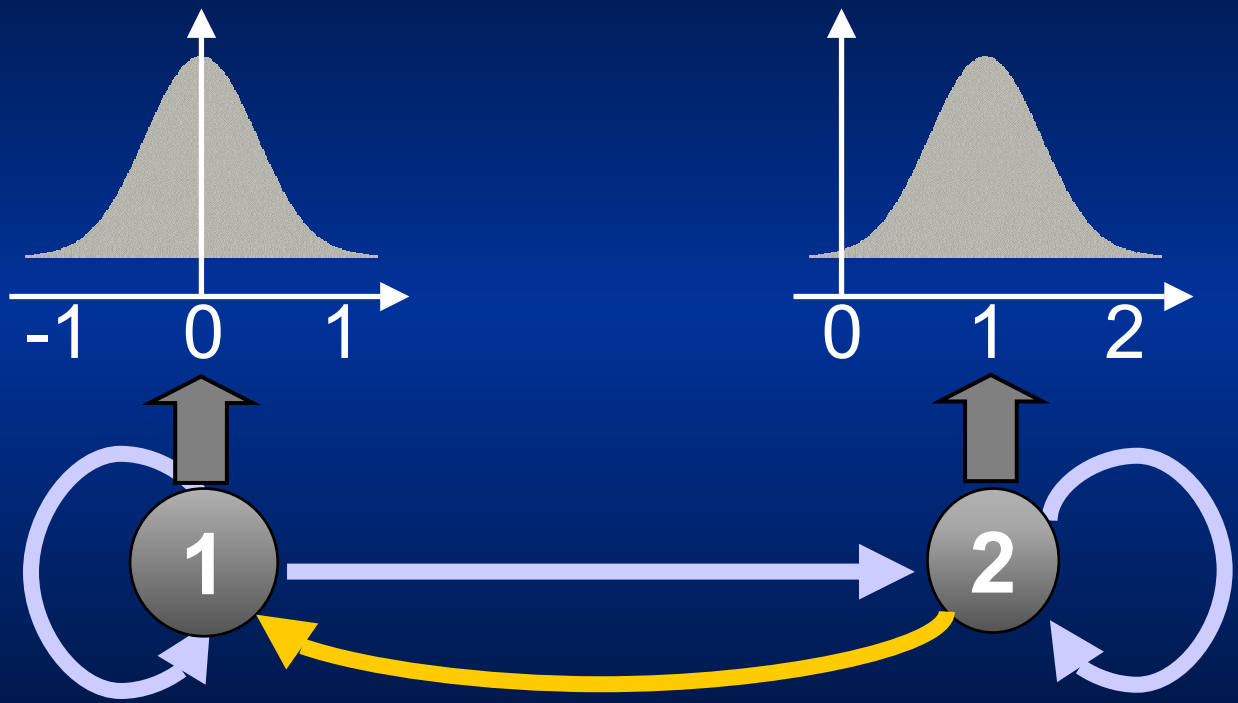


State	**1**	**1**	**2**	**2**
State	**1**	**2**	**2**	**2**

Prototype: Cyclic



Prototype: Cyclic



Perspective

- #states = #time-points:
  - linear HMM ″ multi-variate Gaussian
  - covariance matrix diag(σ_1,…, σ_t)
- Typically #states << #time-points

Mixture of HMMs

Mixture models

- Mixture components: HMMs $\lambda_1, \lambda_2, \ldots, \lambda_k$
- Mixture model ″ weighted sum of λ_i

P[gene | mixture] =

$$\alpha_1 \, P[\text{gene} \mid \lambda_1] + \alpha_2 \, P[\text{gene} \mid \lambda_2] + \ldots$$

$\alpha_i \geq 0$, add to unity

HMM-based 'Clustering'

- Input:
  - genes profiles g_i
  - collection of k **HMMs**
- Initialization:
  - Assign the probability that a data-point belongs to each k **HMMs** randomly
- Iteration (until convergence of assignment):
  - Compute the **new HMM parameters (B-Welch)**
  - Re-assign g_i to a **HMM** proportionally to P[gene | λ_1]

Inference of groups

From Mixtures to Groups

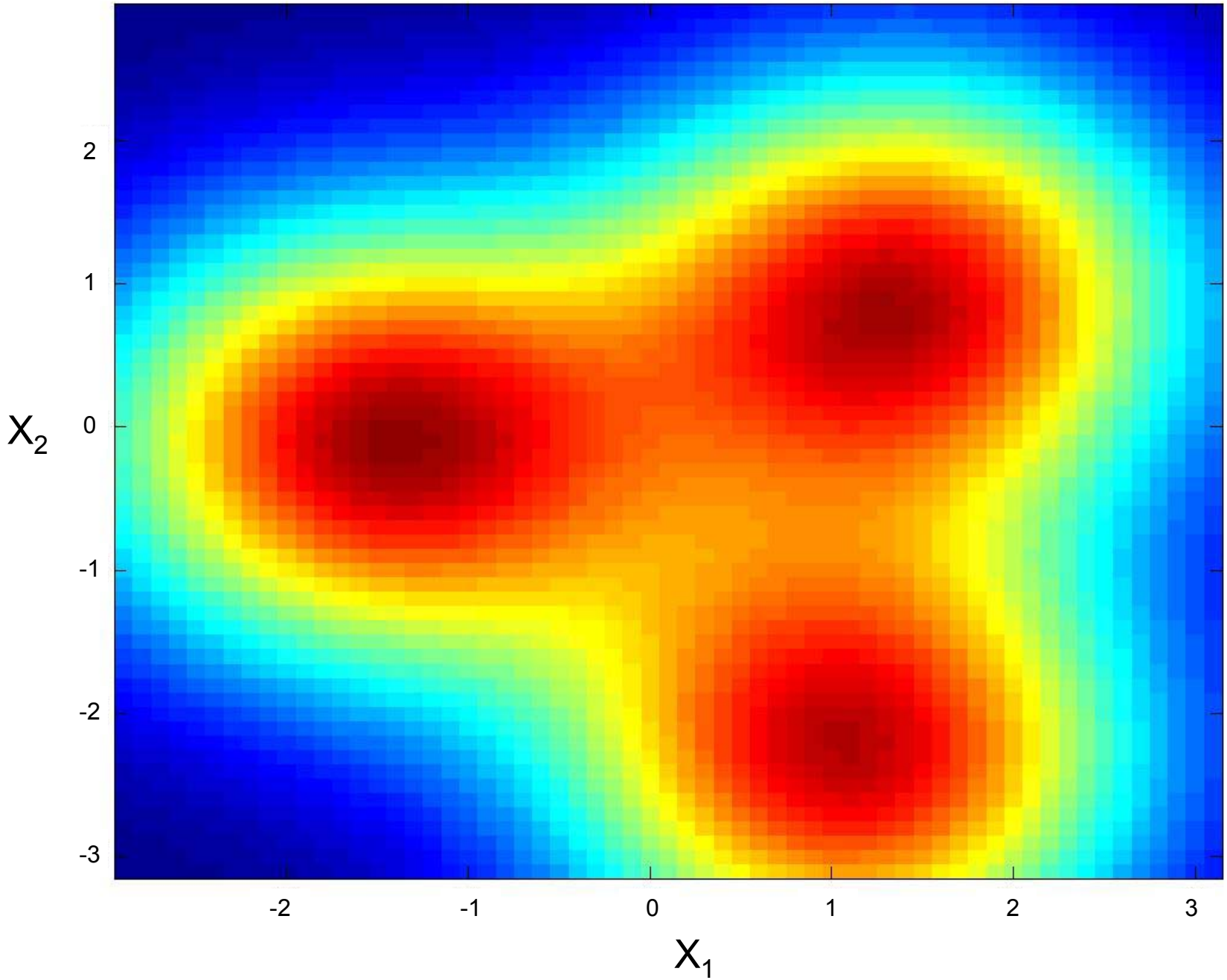
- Posterior probability of mixture component λ_i

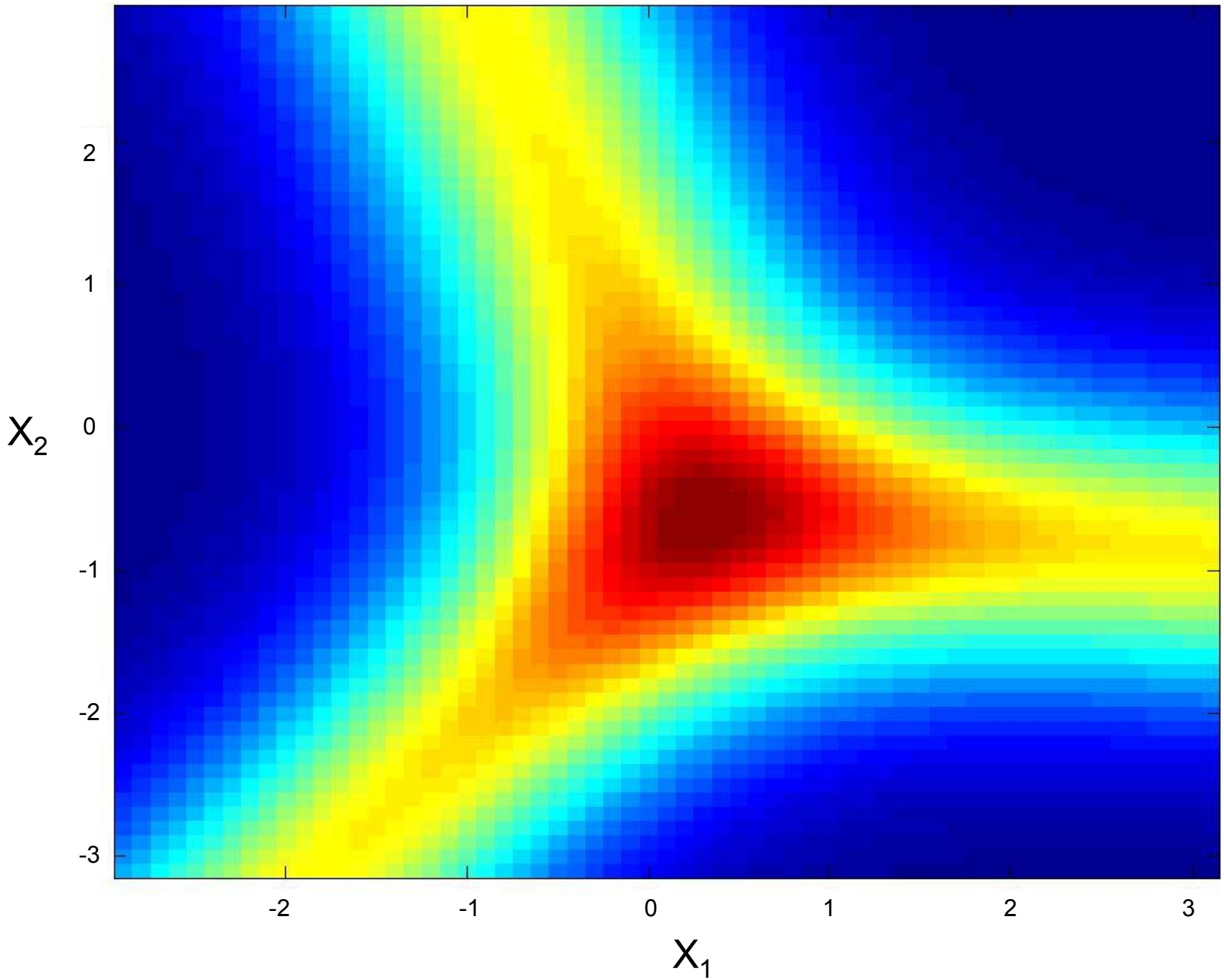
$$P[\ \lambda_i \mid \textit{gene}\ ]$$

- Shannon-entropy

$$\mathbf{H}(\ \{\ P[\ \lambda_i \mid \text{gene}\ ]\ \}_{1 \le i \le k}\ )$$

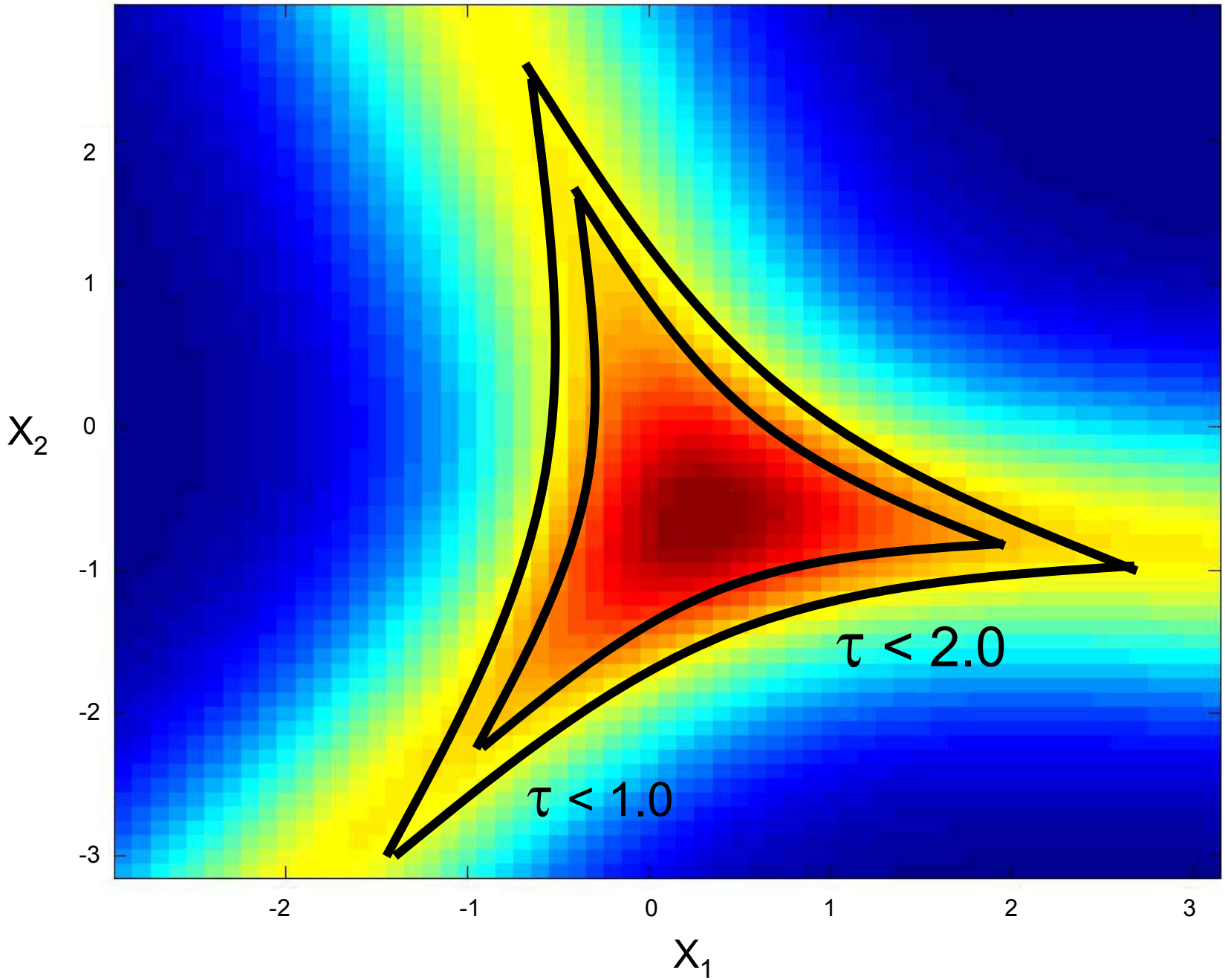
quantifies level of ambiguity in assignment


2
1
0
-1
-2
-3
X_2
-2
-1
0
1
2
3
X_1


2
1
0
-1
-2
-3
X_2
-2
-1
0
1
2
3
X_1

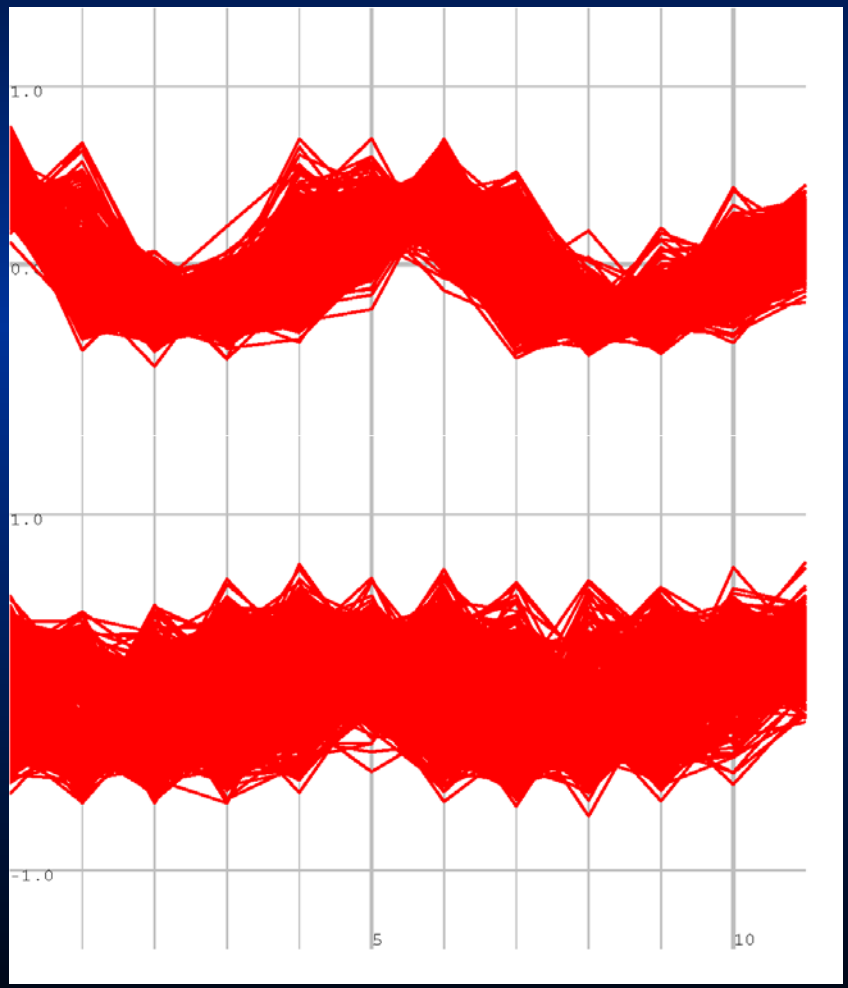
From Mixtures to Groups

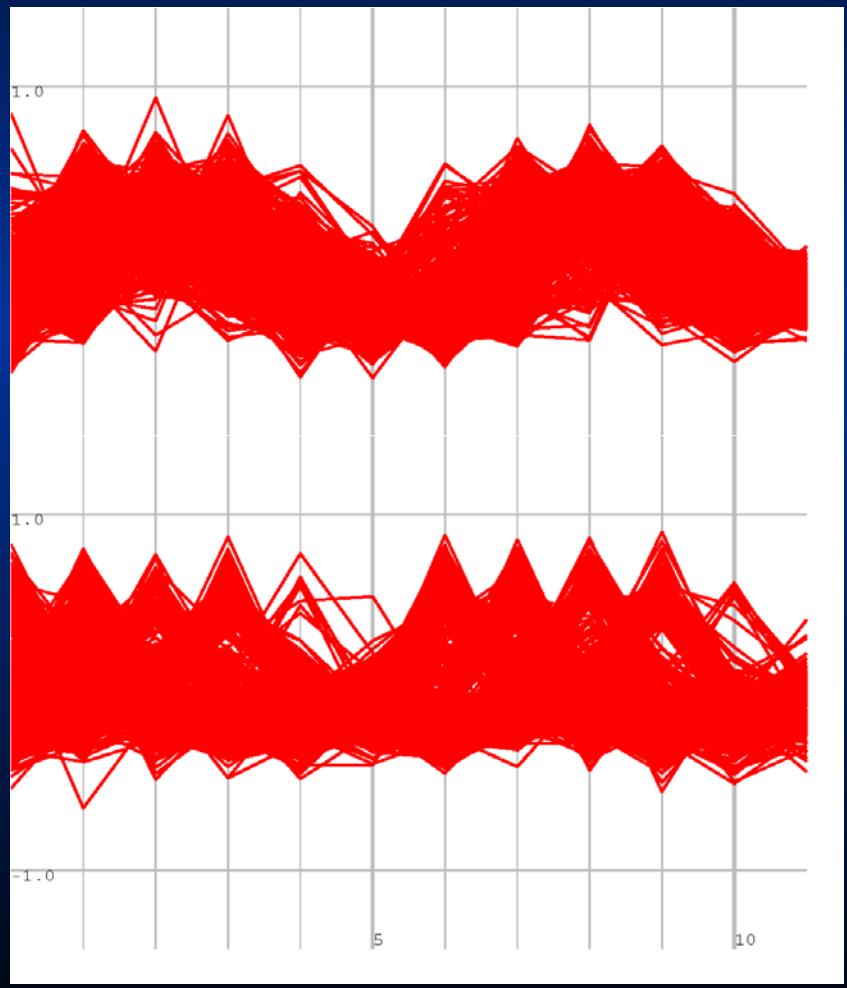
- Choose entropy threshold τ
- If entropy of posterior is below τ
  - Assign gene to group i of maximal posterior
- Else:
  - leave gene unassigned


2
1
0
-1
-2
-3
X2
-2
-1
0
1
2
3
X1
τ < 2.0
τ < 1.0

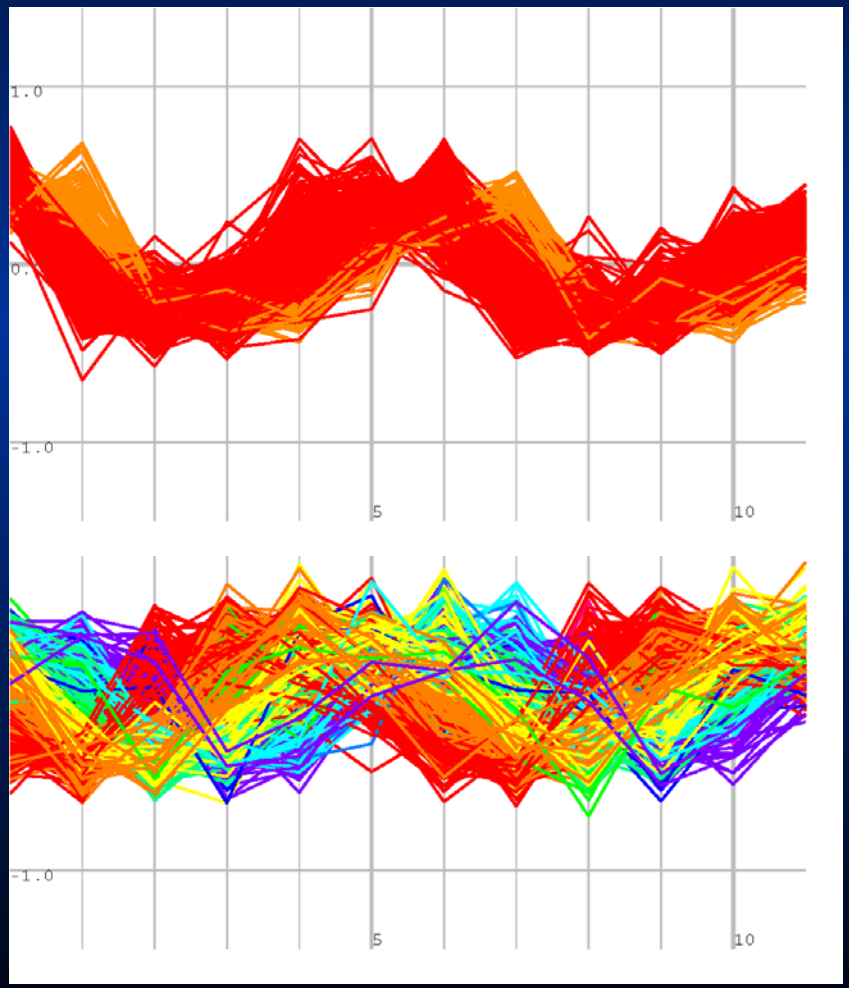
Results

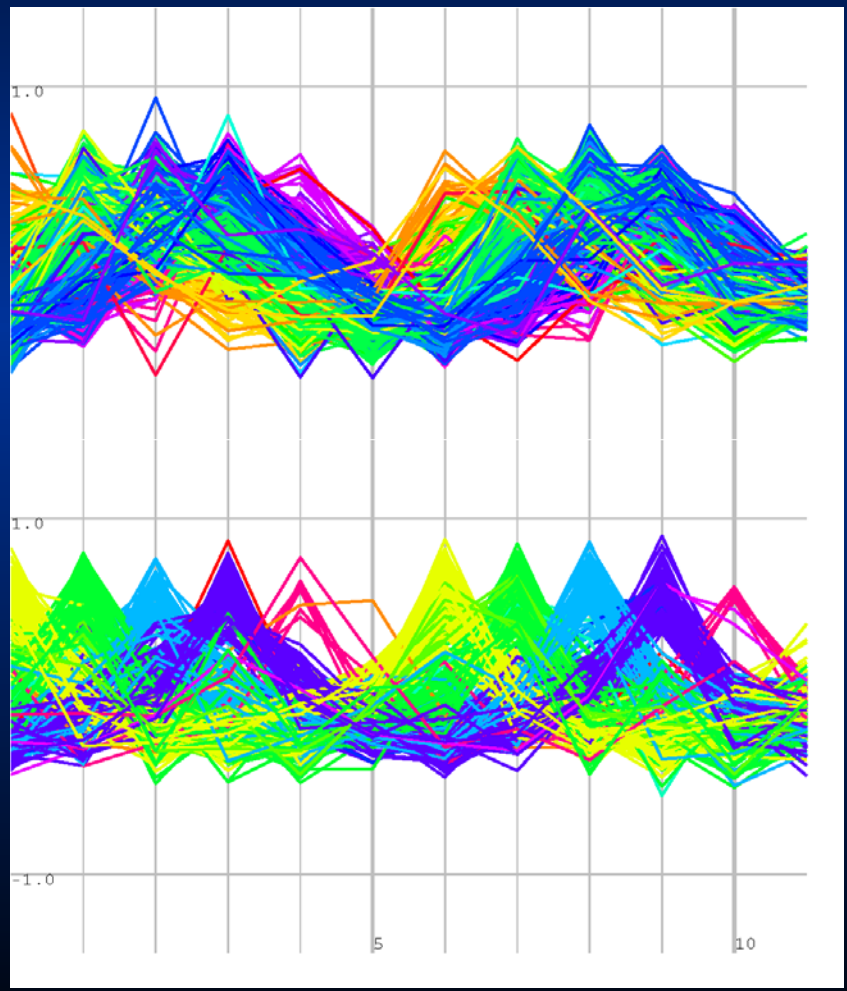
Diurnal Time Courses





Results
Diurnal Time Courses





(2) Cluster Validation

Motivation

- Which clustering method to use?
- How good is a clustering?
- GO enrichment gives no comparative basis!

Use gene annotation (GO) as external data to validate mixtures (or clusterings) from gene expression

External Indices

In clustering, external indices look for th
number of genes pairs that

	Same categ.	*Distinct categ.*
Same cluster	**True Positive**	**False Positive**
Distinct clusters	**False Negative**	**True Negative**

External Indices

$$corr.Rand = \frac{\#TP + \#TN - n_c}{\#Pairs - n_c}$$

External Indices for Mixture Model

Given mixture models *U* and *V*, consider the posteriors of the mixture components :

$$\{P[u_i \mid g]\}_{1<i<C} \qquad \{P[v_j \mid g]\}_{1<j<R}$$

$g_k \equiv g_l \leftrightarrow$ the event of co-occurrence of g_k and g_l

$$P[g_k \equiv g_l \text{ given } U] = \sum_{j=1}^{C} P[u_j \mid g_k] \cdot P[u_j \mid g_l]$$

External Indices for Mixture Model

$$TP = \sum_{k=1}^{N} \sum_{l=k+1}^{N} P[g_k \equiv g_l \text{ given } U] \cdot P[g_k \equiv g_l \text{ given } V]$$

$$TN = \sum_{k=1}^{N} \sum_{l=k+1}^{N} P[g_k \equiv g_l \text{ given } U]^C \cdot P[g_k \equiv g_l \text{ given } V]^C$$

Biological Data Experiments

- Gene expression data:
  - Yeast during sporulation (7 time points)
  - 1027 genes after 2 fold filtering
- 'Clustering' Methods:
  - Hierarchical clustering (Pearson correlation)
  - K-means (Pearson correlation)
  - Mixture of HMMs
  - Mixture of Multivariate Normals (full covariance)

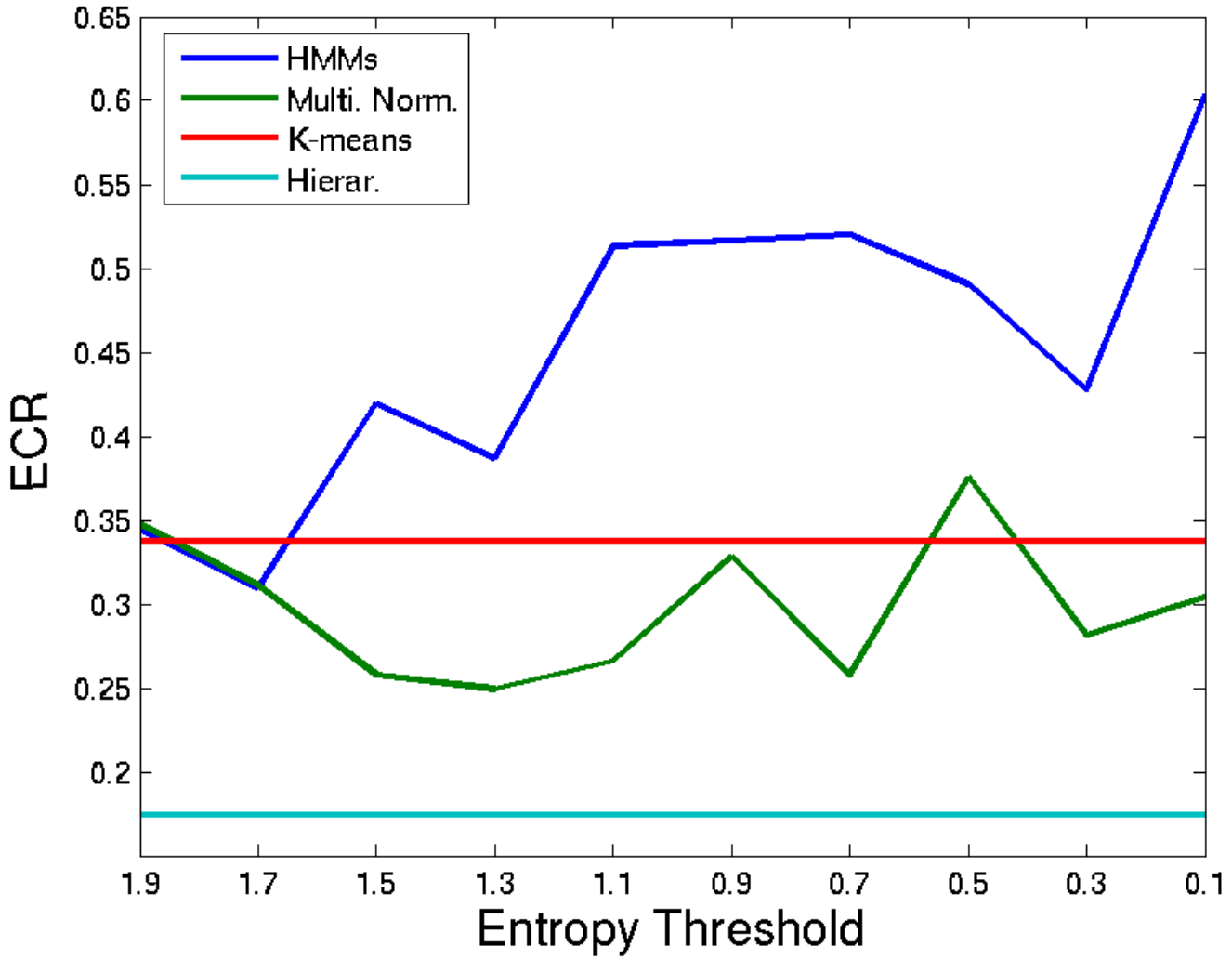
Biological Data Representation

- Mixture Model representation
  - Use each GO term as a component in the mixture
  - Maximum likelihood estimator of a multinomial distribution

$$P[t_i \mid g] = \begin{cases} 1/\#\{j \mid g \in t_j\}, & \text{if } g \in t_i \\ 0, & \text{otherwise} \end{cases}$$

Results

Methods x ‘All’ GO Terms

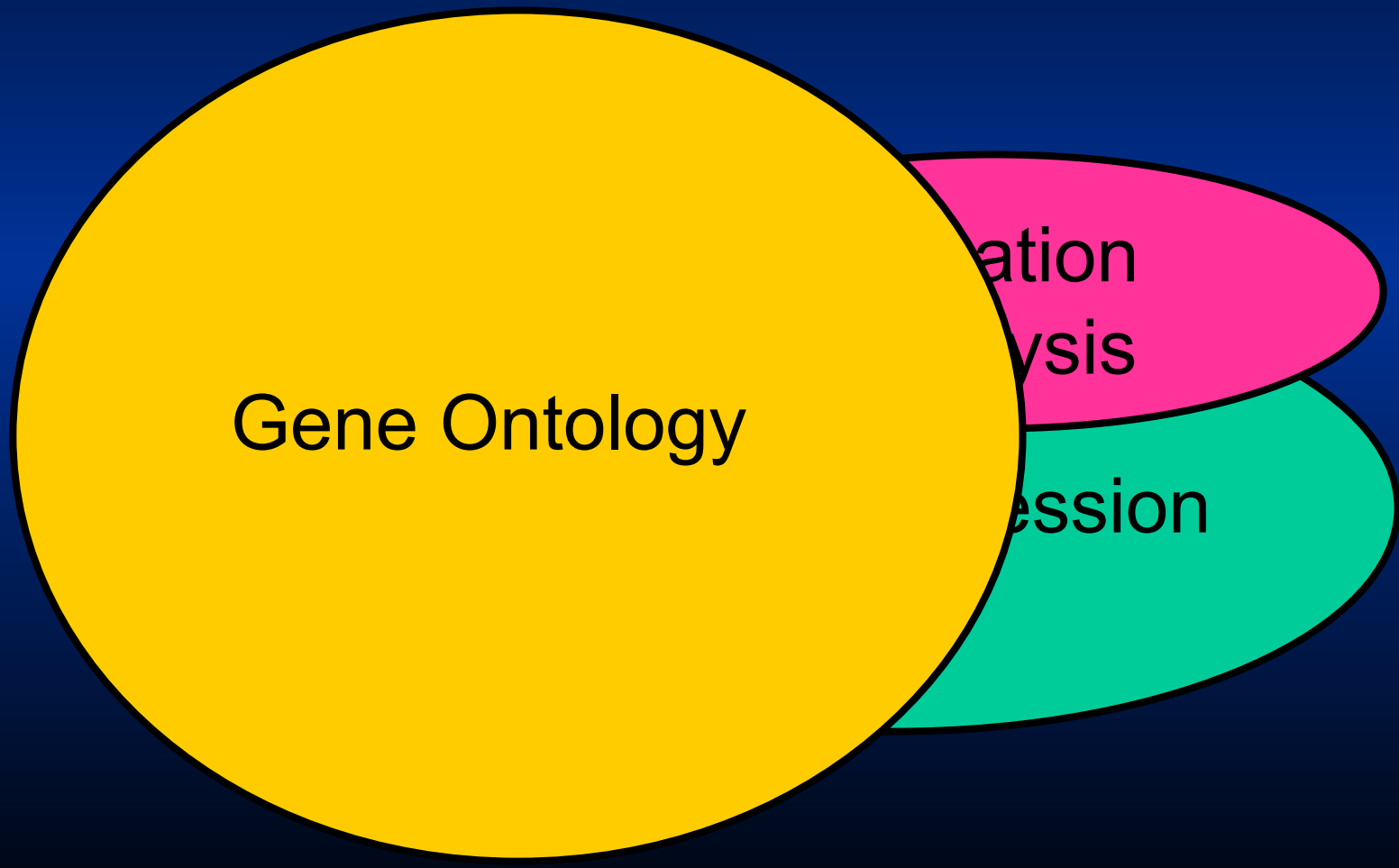


(3) Clustering of Heterogeneous data

Motivation

Use additional large scale biological data to improve clustering of gene expression time-courses

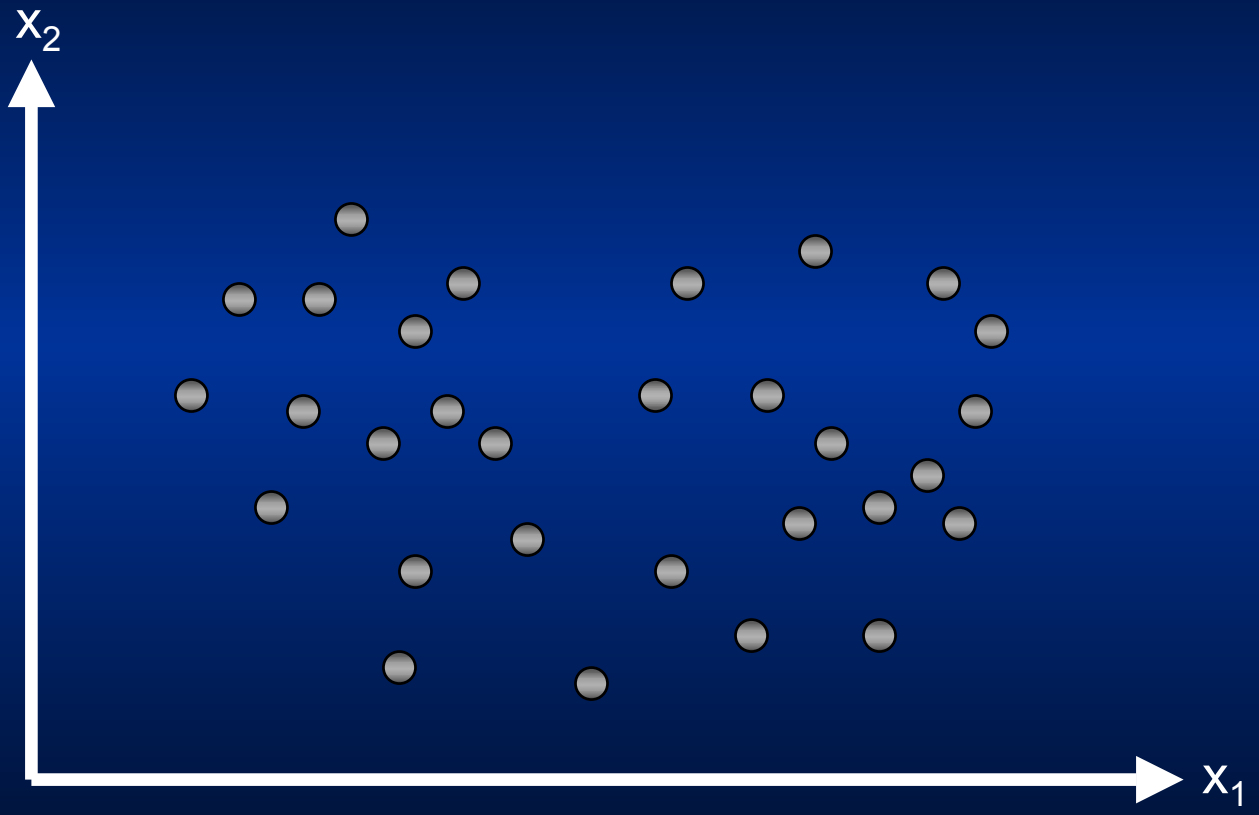
Challenges of Heterogeneous Biological Data



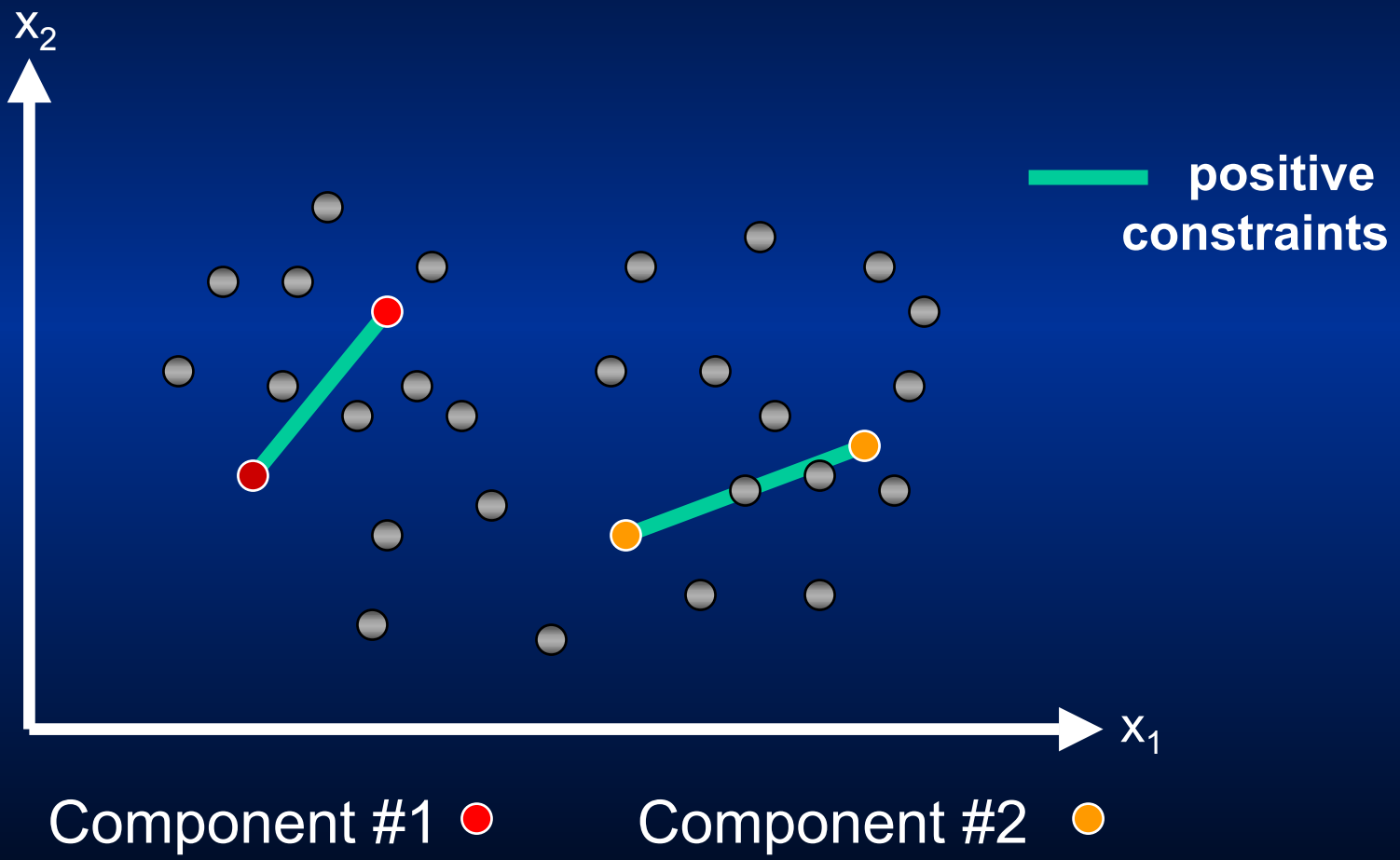
Our Approach

- Semi-supervised learning
  - Encode location analysis as soft pairwise constraints
  - Mixture estimation with constraints (Lange *et al.*, 2005, Lu and Leen, 2005)

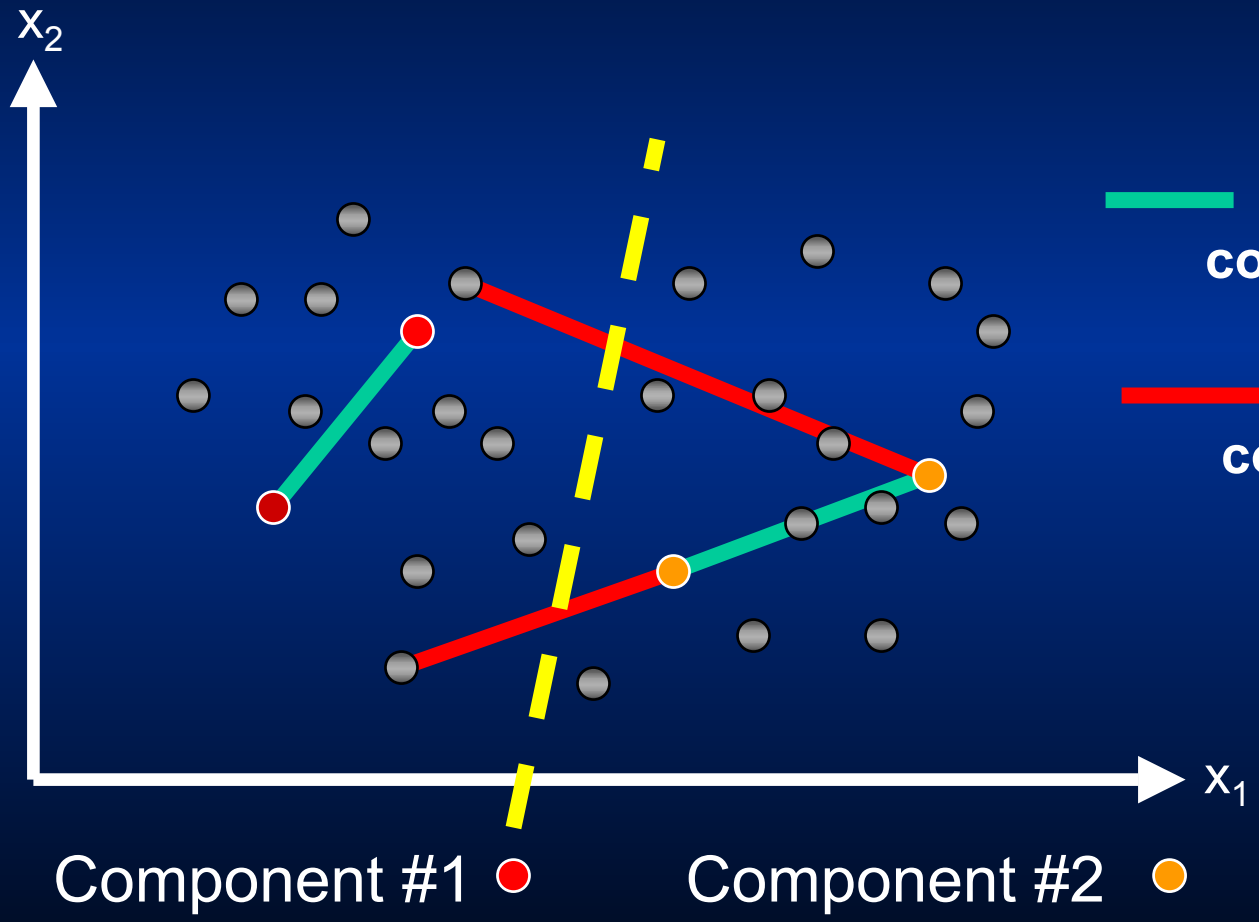
Semi-Supervised Learning



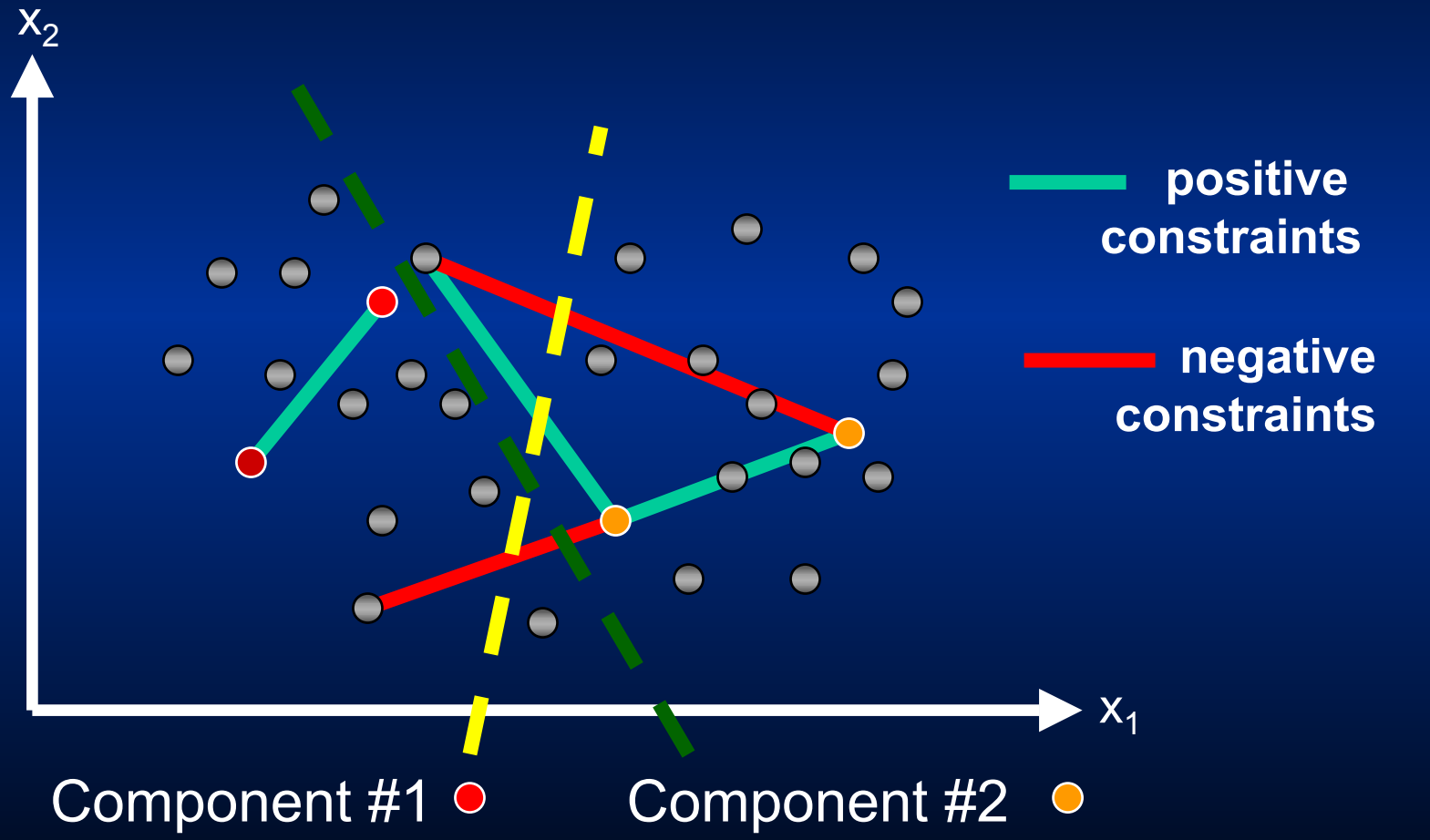
Semi-Supervised Learning



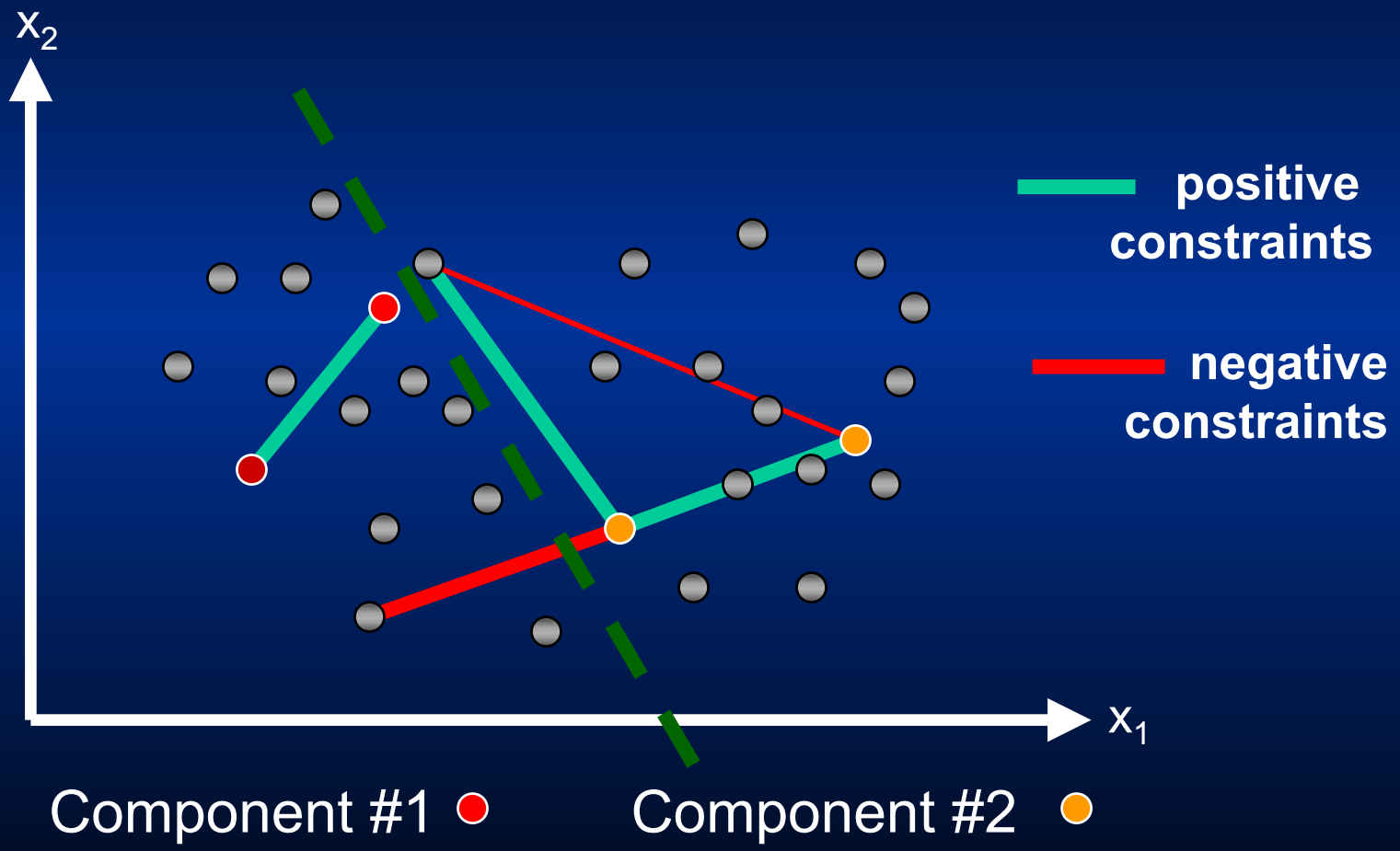
Semi-Supervised Learning



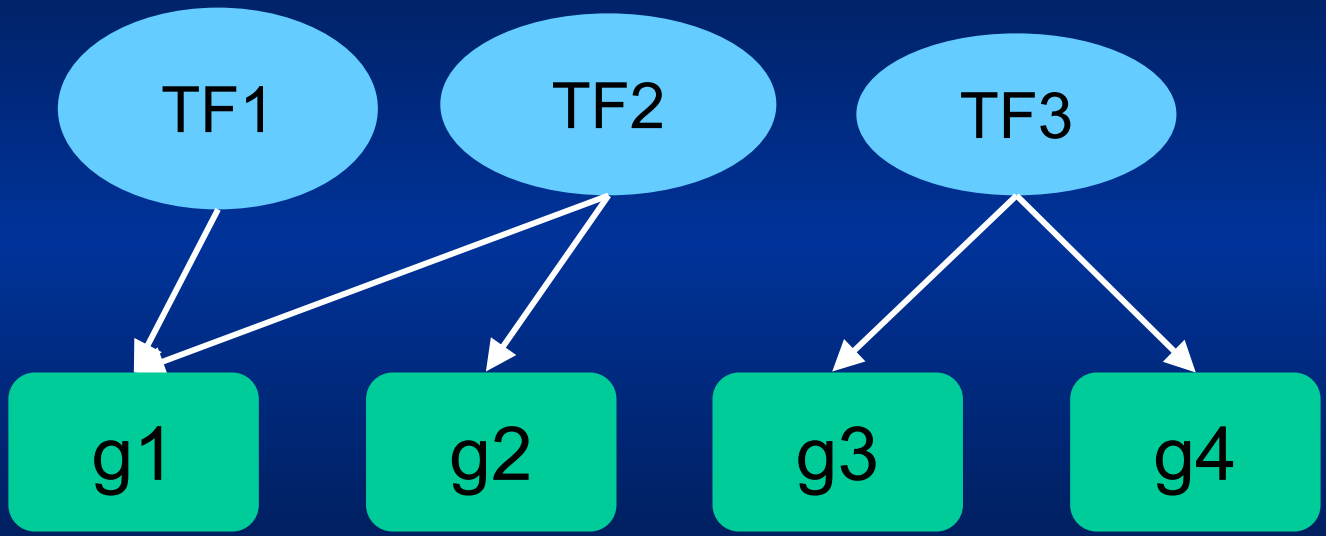
Semi-Supervised Learning



Semi-Supervised Learning



Pairwise Constraints Location Analysis



$w^{+}_{(1,2)} = 0.5$ $w^{-}_{(1,2)} = 0.5$

$w^{+}_{(2,3)} = 0.0$ $w^{-}_{(2,3)} = 1.0$

$w^{+}_{(2,3)} = 1.0$ $w^{-}_{(2,3)} = 0.0$

Mixture Estimation with Constraints (1)

Maximize the complete likelihood:

$$P[X, Y|W, \Theta] = P[X|Y, \Theta]\, P[Y|W, \Theta]$$

where $\boldsymbol{X}$ is the observable data, $\boldsymbol{Y}$ the hidden data, Θ the model parameters, $W = \{W^+, W^-\}$ the pairwise constraints

The prior can be decomposed at:

$$P[Y|\Theta, W] = P[Y|\Theta]\, P[W^+|Y, \Theta]\, P[W^-|Y, \Theta]$$

Mixture Estimation with Constraints (2)

$$\mathrm{P}[W^{+} \mid Y, \Theta] \approx \exp \sum_{i} \sum_{j \neq i} -w_{ij}^{+} 1\{y_i \neq y_j\} \lambda^{+}$$

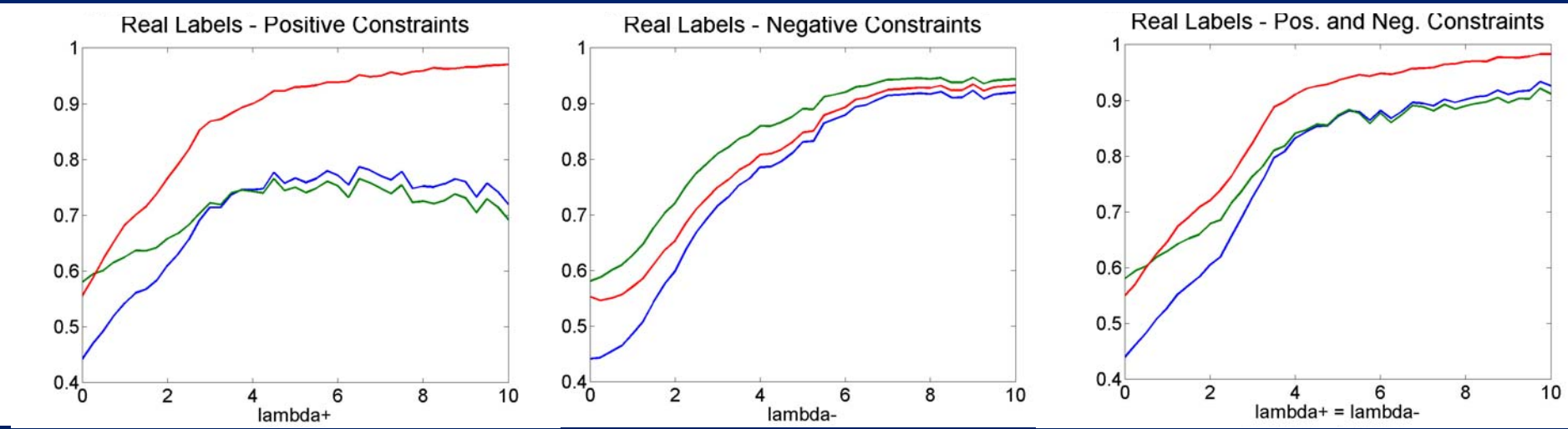
The posterior assignments are approximated by means of Gibbs sampling (Lu and Leen, 2005)

Data

- Gene expression data
  - time-courses of 384 genes during mitotic cell division in Yeast (Cho, 1998)
  - expert classification into ‘five’ cell-cycle phases
- Constraints
  - transcription factor location analysis (Lee, 2002)
  - true labels

Results

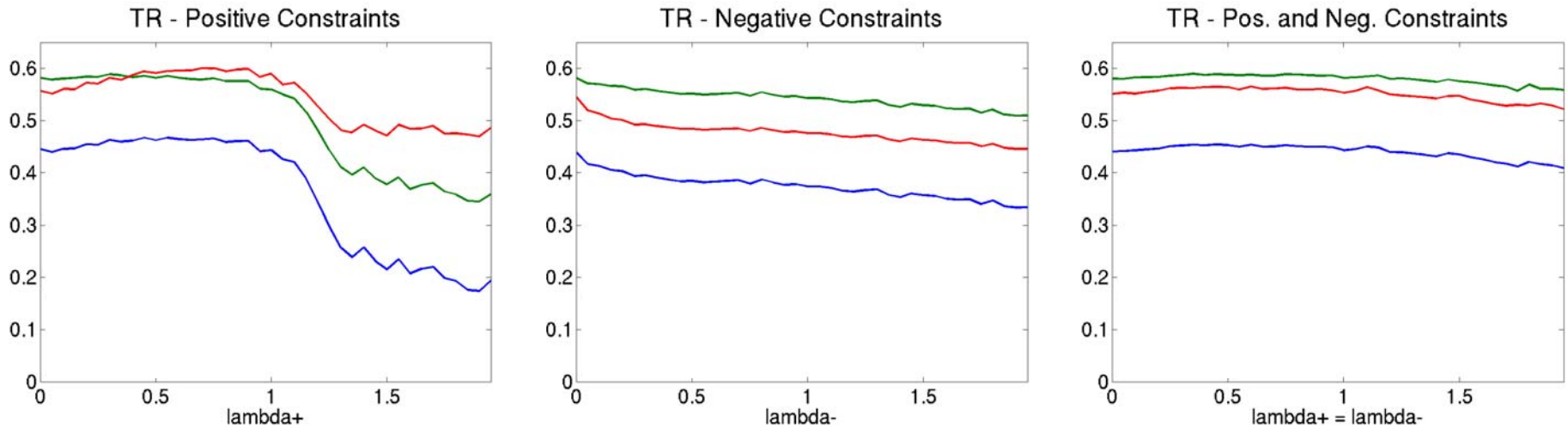
Constraints from True Labels



corrected Rand
specificity
sensitivity

5% of gene pairs constrained

Constraints from Location Analysis



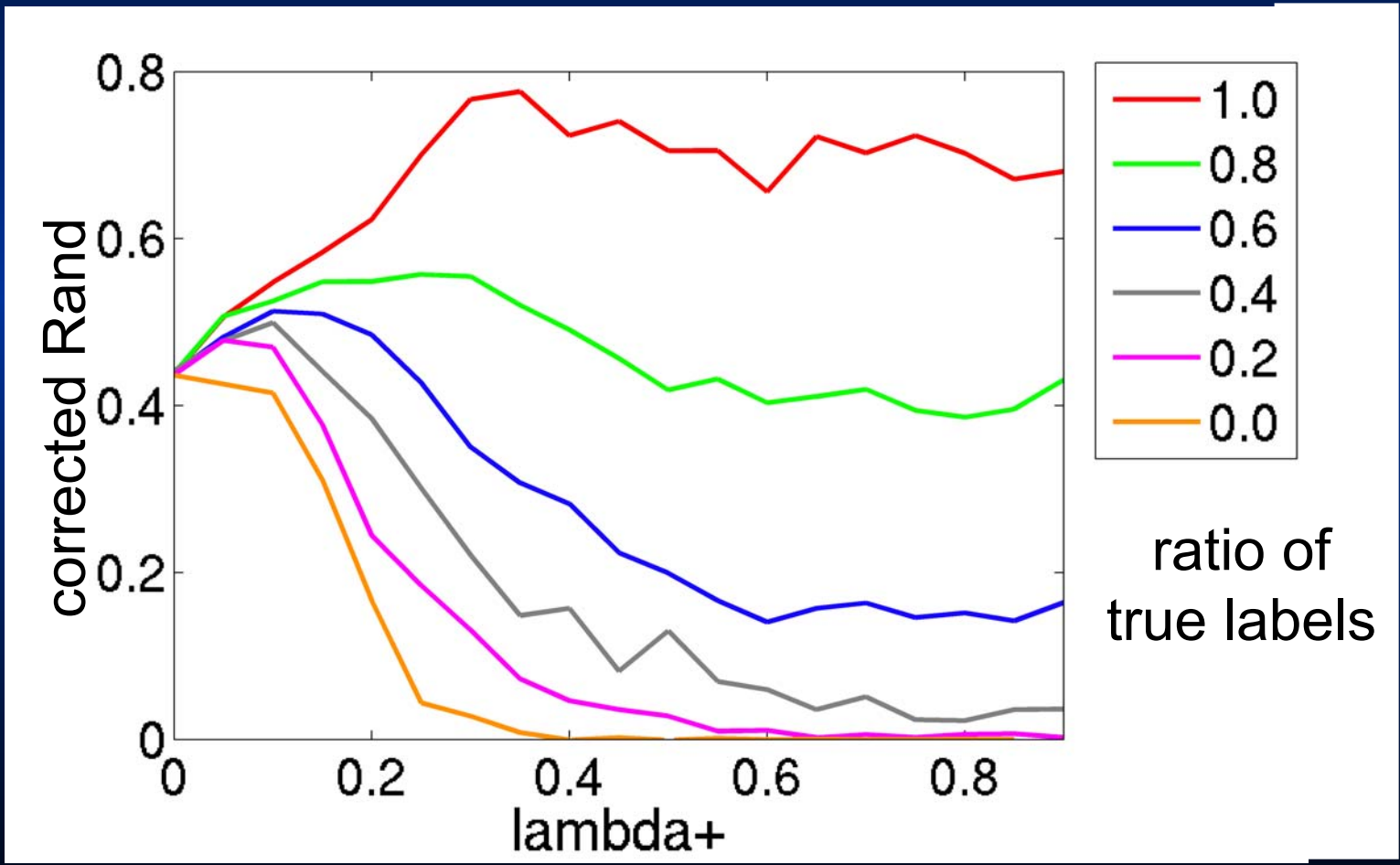
corrected Rand
specificity
sensitivity

40% of gene pairs constrained

Possible Explanations

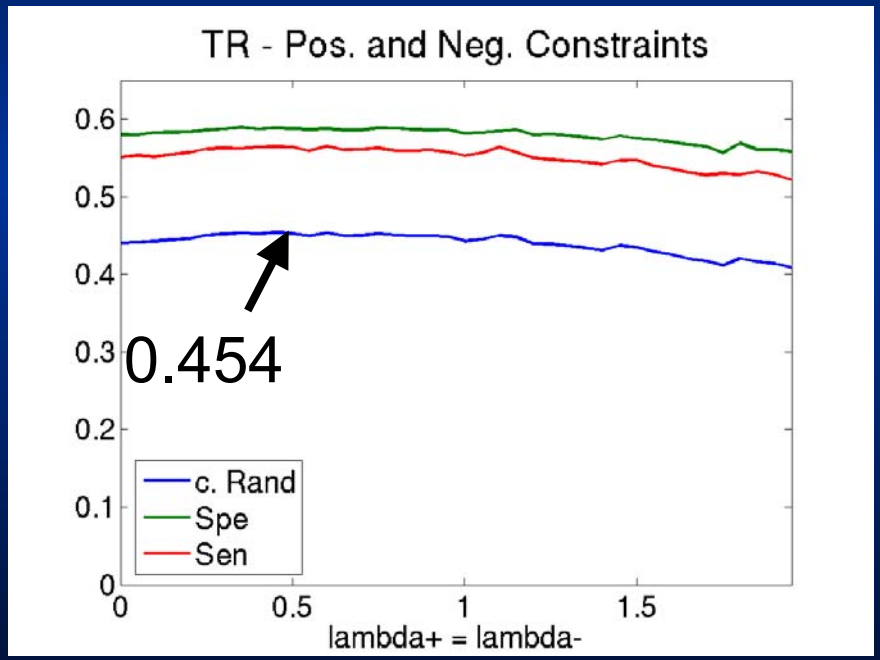
- Non-specific information content
- Noise in the data
- …

Constraints from True and Random Labels

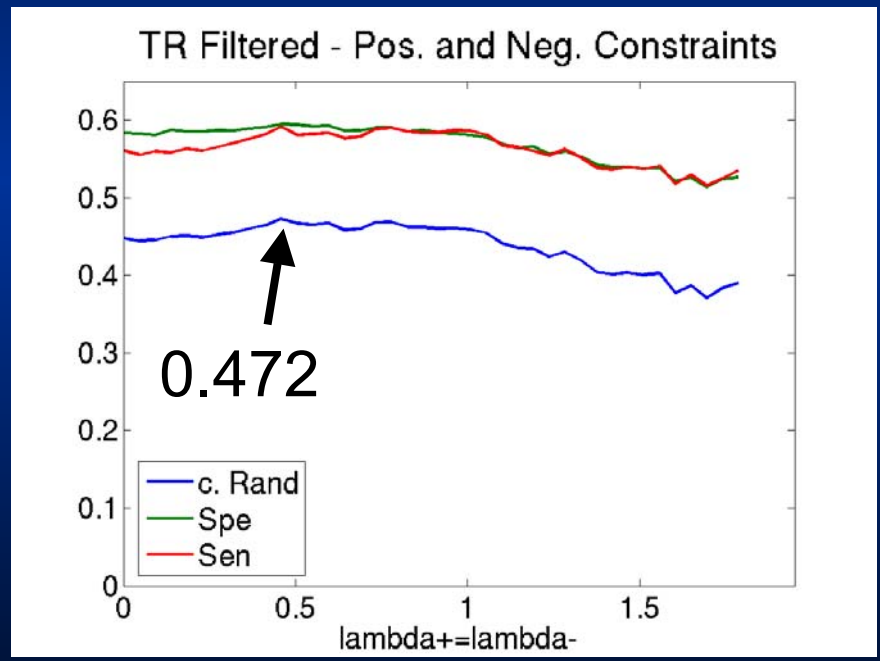


Filtered Constraints from Location Data

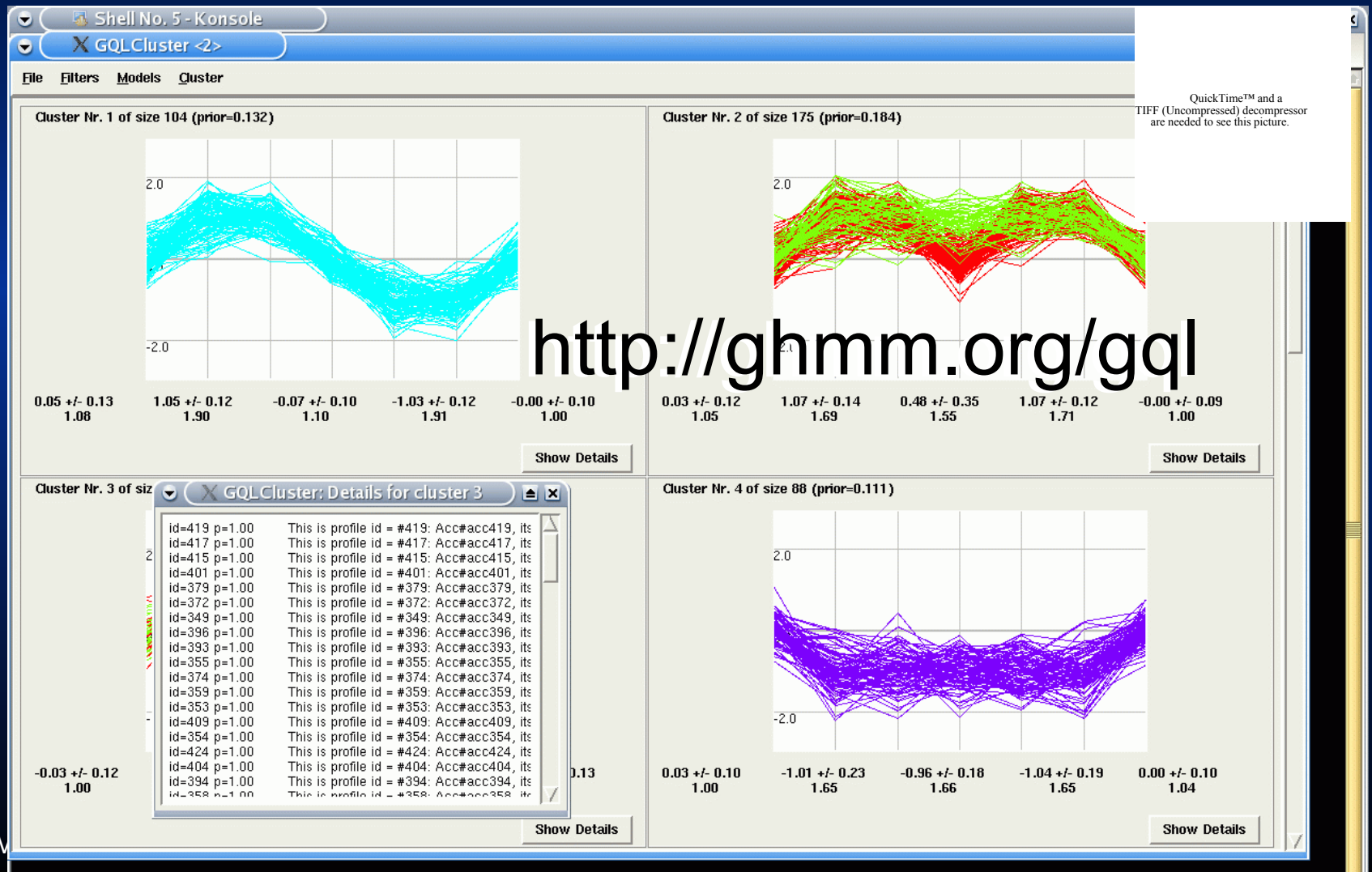
Non-filtered

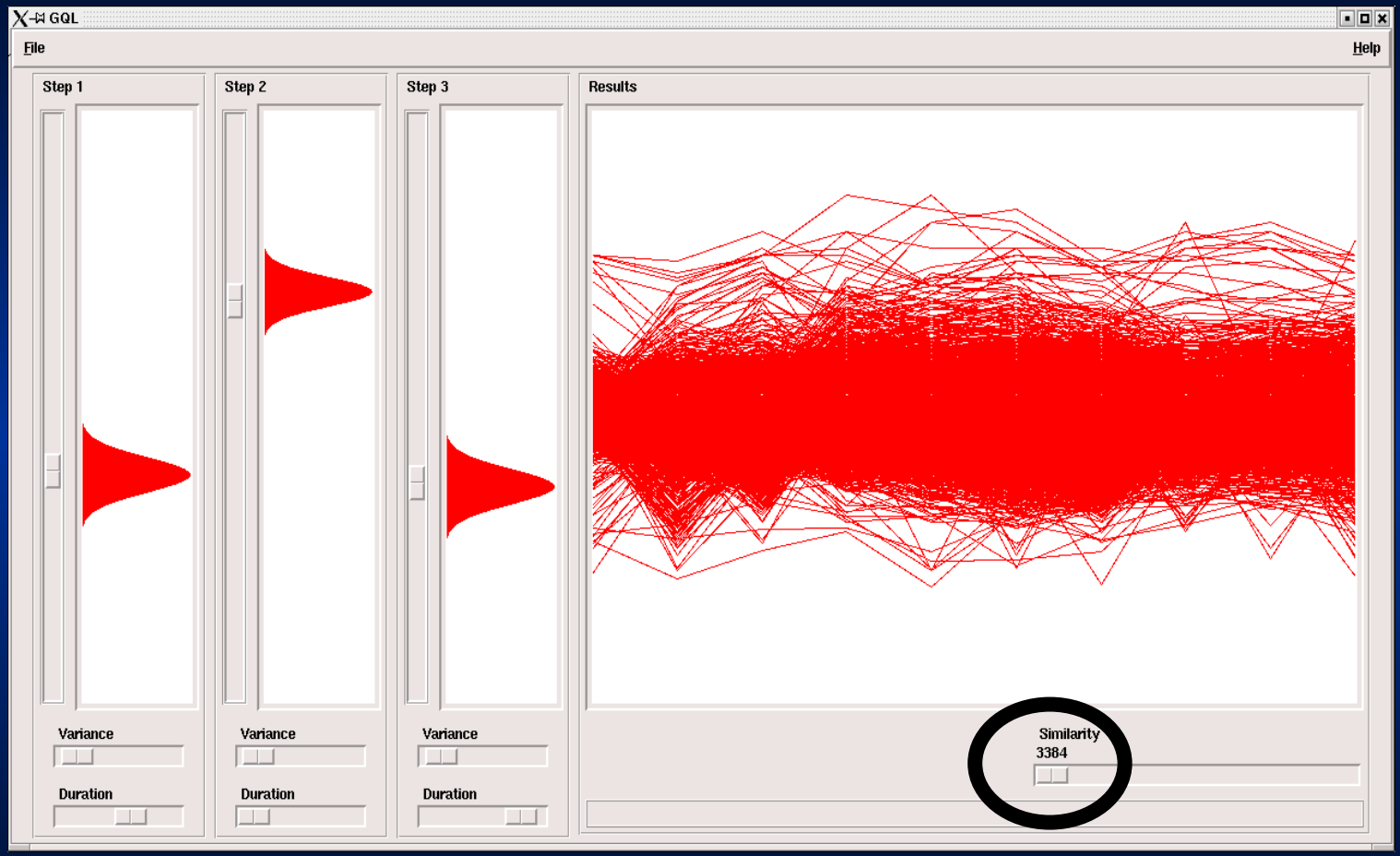


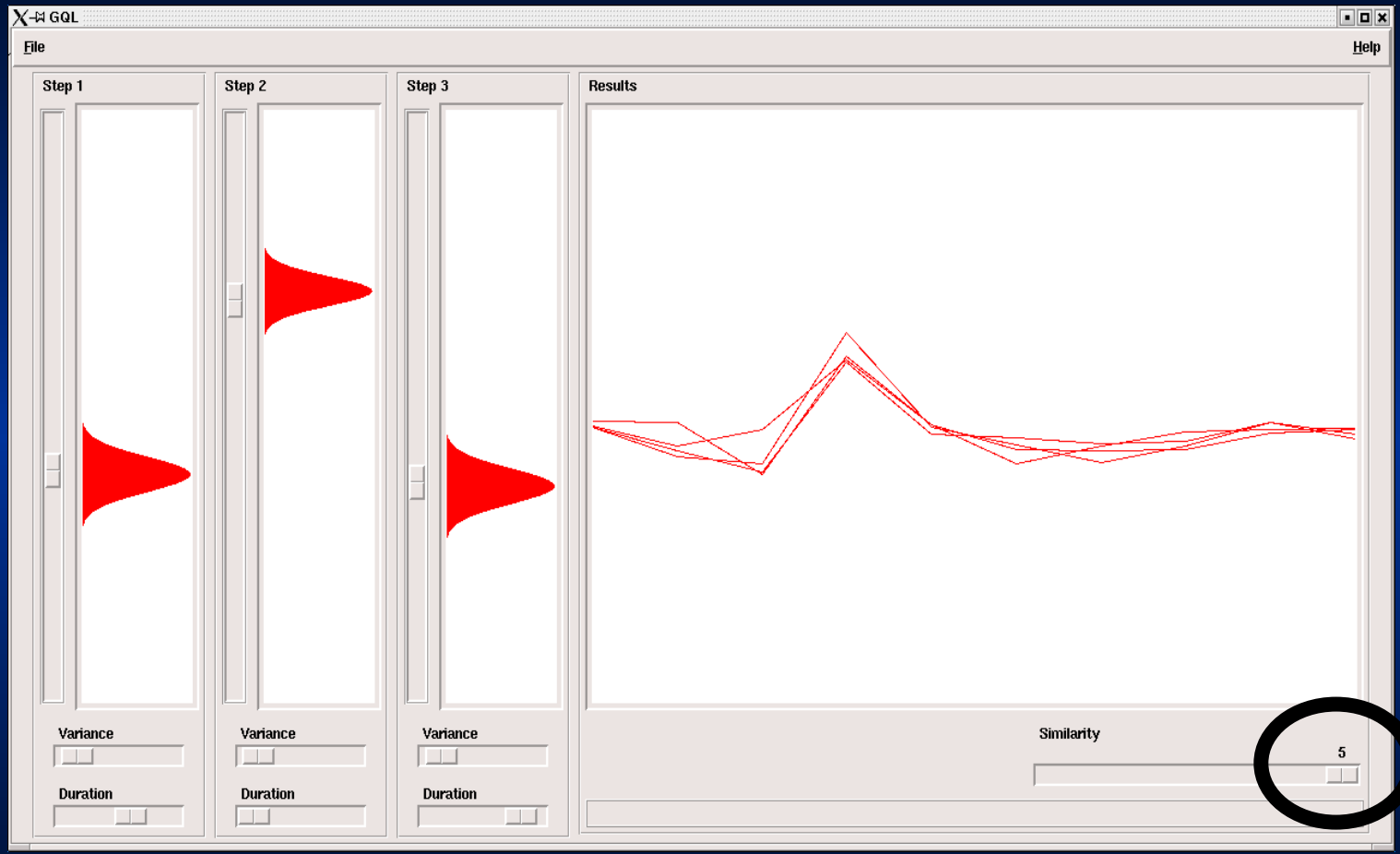
Filtered



Software GQL




GQL
File
Help
Step 1
Step 2
Step 3
Results
Variance
Duration
Variance
Duration
Variance
Duration
Similarity
3384


GQL
File
Help
Step 1
Step 2
Step 3
Results
Variance
Duration
Variance
Duration
Variance
Duration
Similarity
5

Summary

- Clustering of Time-Series
  - Flexible: *cyclic & transient time-courses*
  - Interactive & robust
- Cluster Validation
  - Methodology for evaluating clustering given functional annotation
- Heterogeneous Analysis
  - Successful integration of location analysis

Outlook

- Cluster Validation
  - Perform a ‘extensive’ evaluation of clustering methods
- Heterogeneous Analysis
  - learning of relevant constraints
  - *In-situ hybridization, protein-protein interactions, …*
- Clustering of Development Trees
  - In progress …

Acknowledgements

- ***A. Schönhuth*** and ***C. Steinhoff***
- GHMM (http://ghmm.org): Wasinee Rungsarityotin, Benjamin Georgi, Ben Rich, Matthias Heinig, Alexander Riemer, Janne Grunau
- *T. Beissbarth* for help with *GO*

Thanks.

ghmm.org/gql

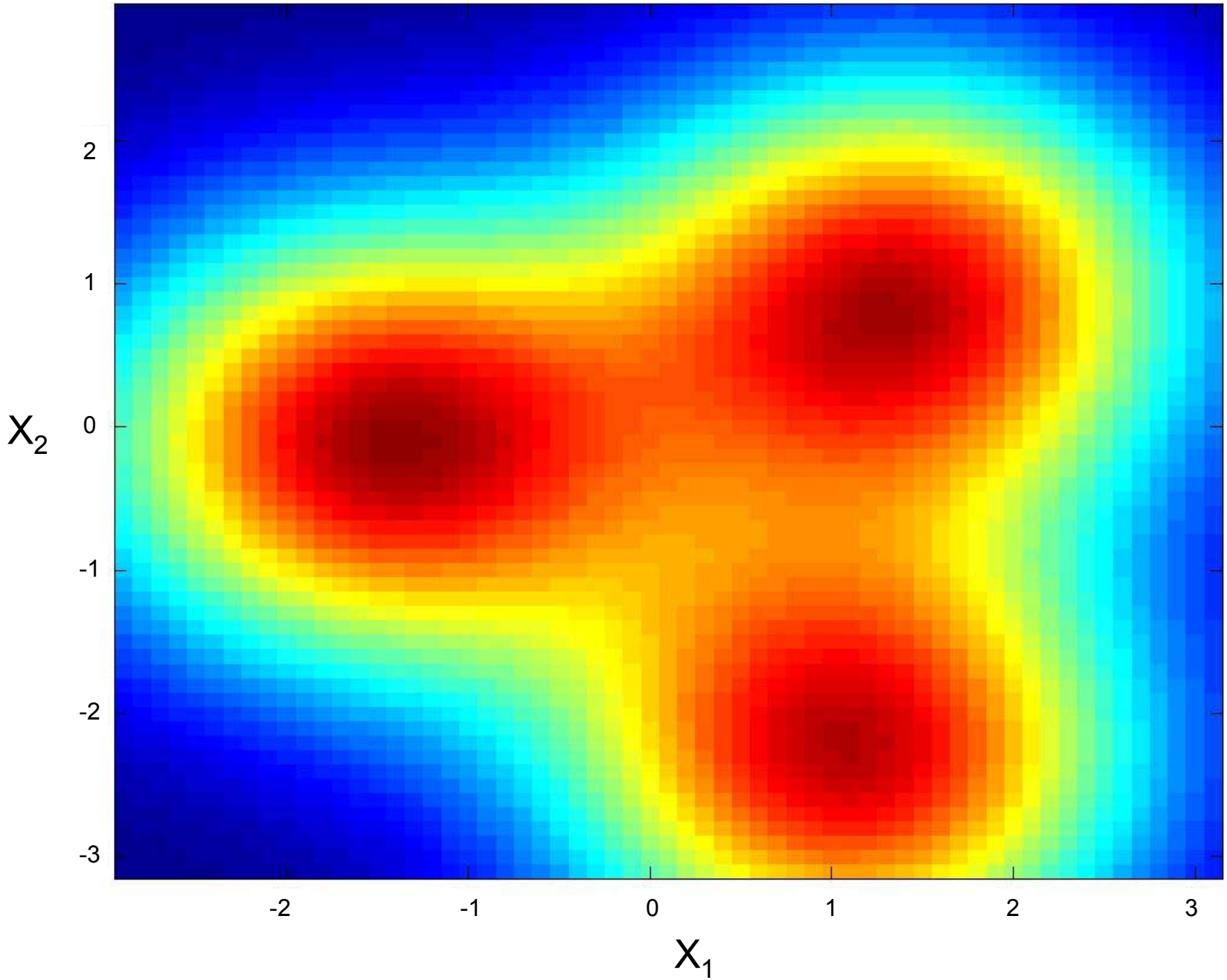
Motivational Buzzphrases

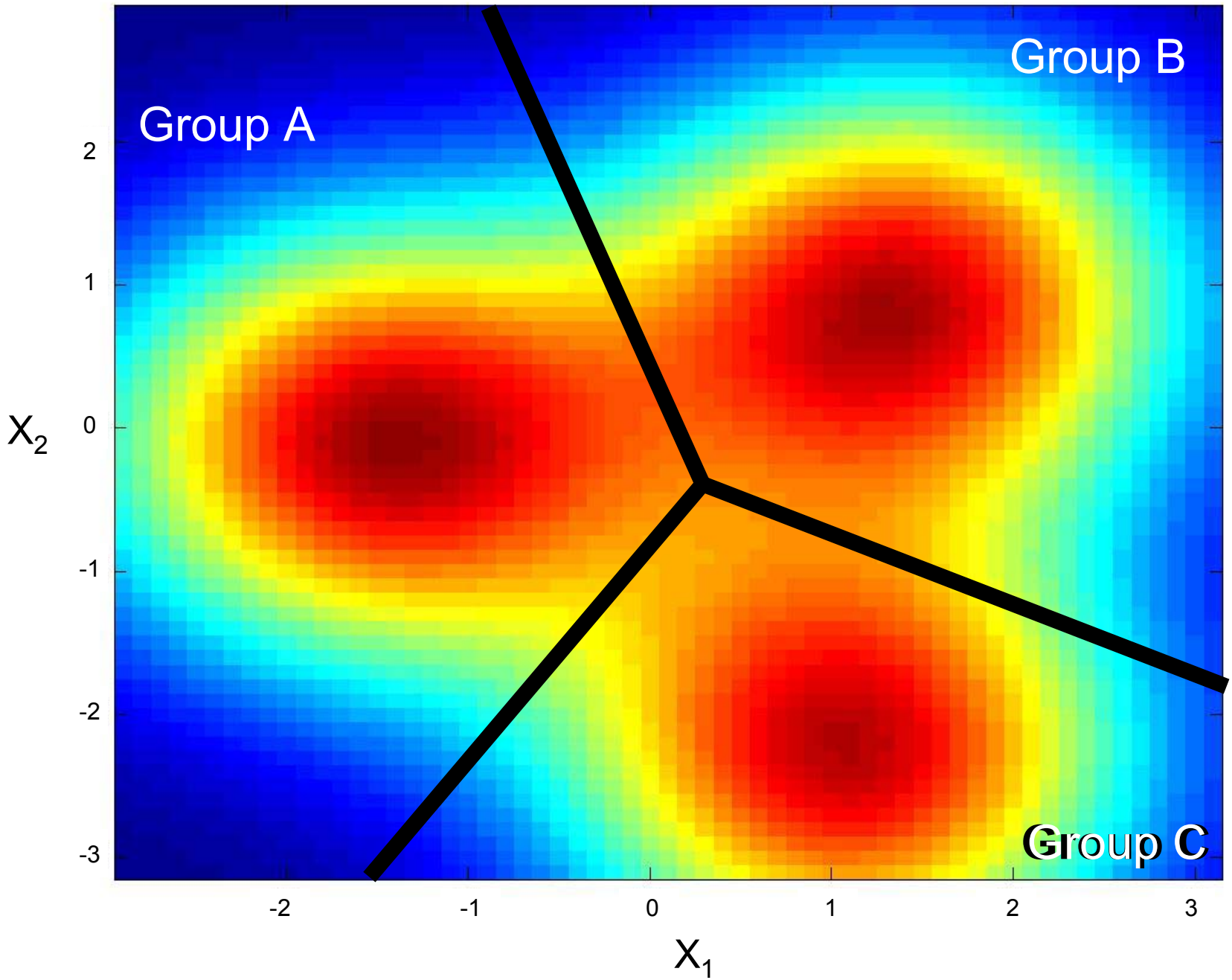
Model qualitatively

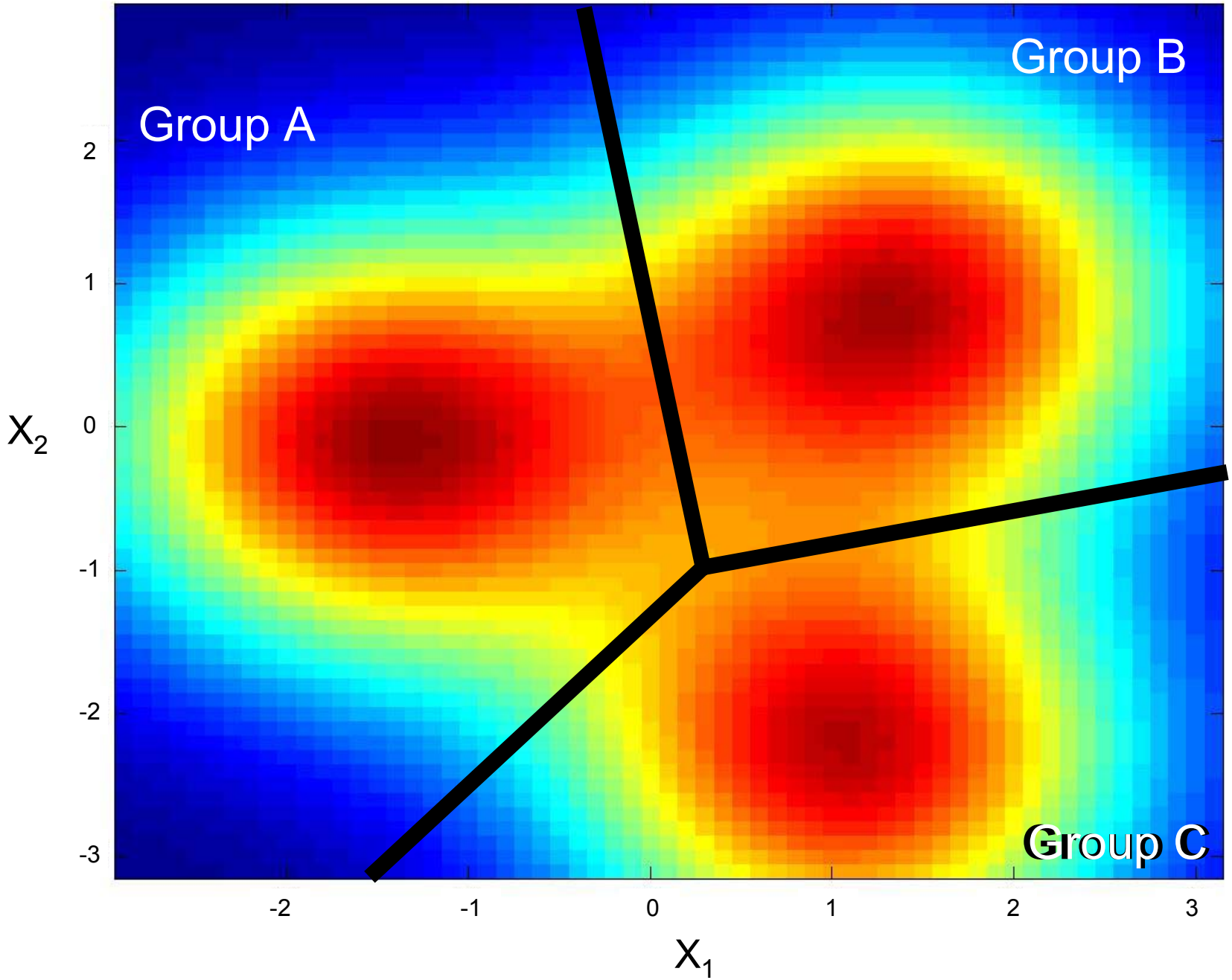
Embrace ambiguity

Assure robustness

Don’t be ignorant


2
1
0
-1
-2
-3
X_2
-2
-1
0
1
2
3
X_1


Group A
Group B
Group C
2
1
0
-1
-2
-3
X_2
-2
-1
0
1
2
3
X_1


Group A
Group B
Group C
2
1
0
-1
-2
-3
X_2
-2
-1
0
1
2
3
X_1

Yeast Cell Cycle