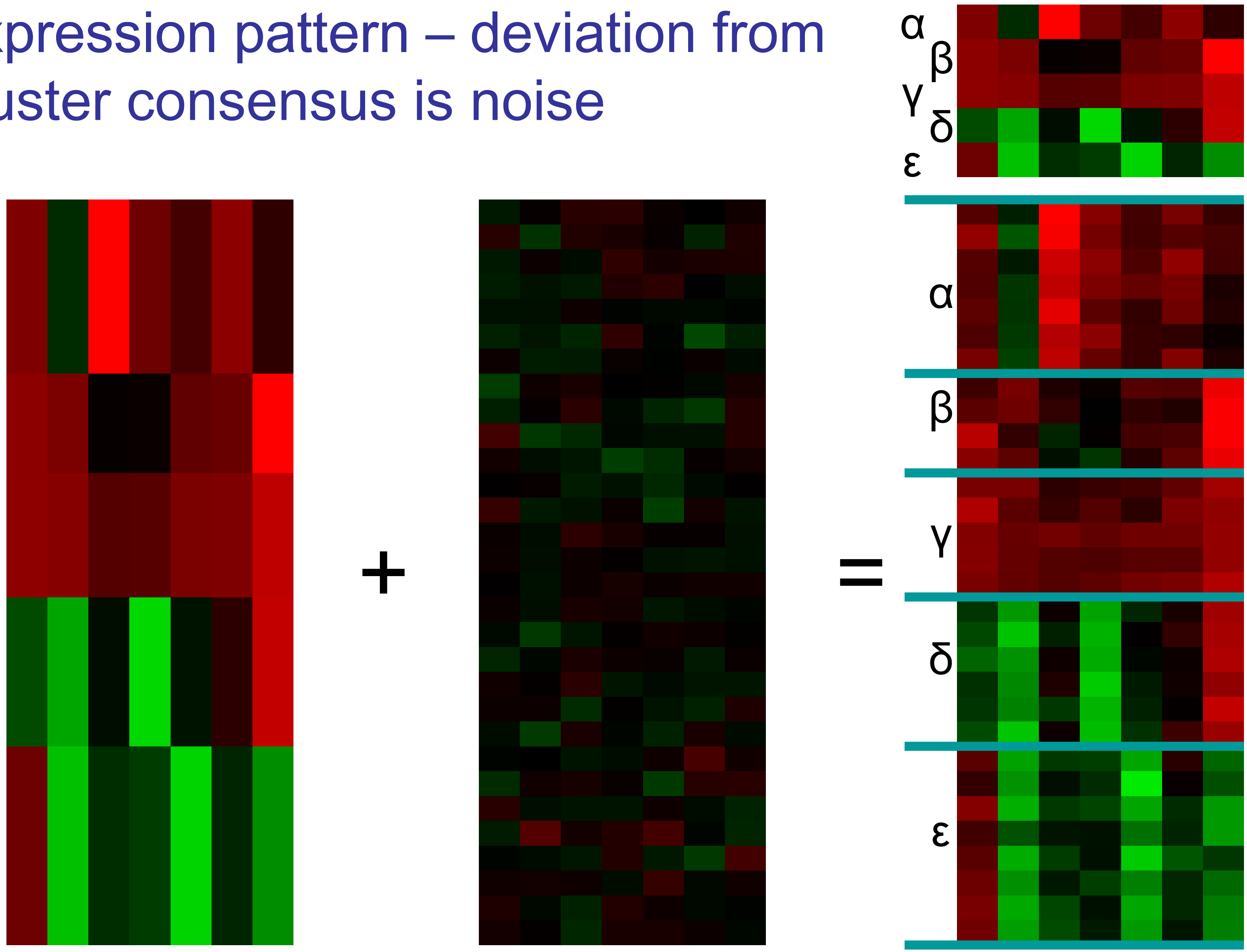
Sparse Matrix Factorization for Gene Expression Analysis

(Work in Progress)

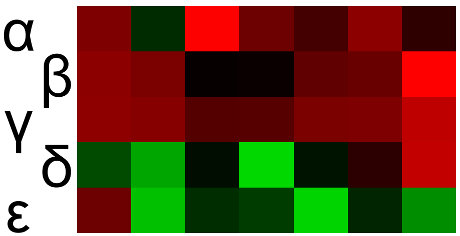
Nati Srebro and Tommi Jaakkola

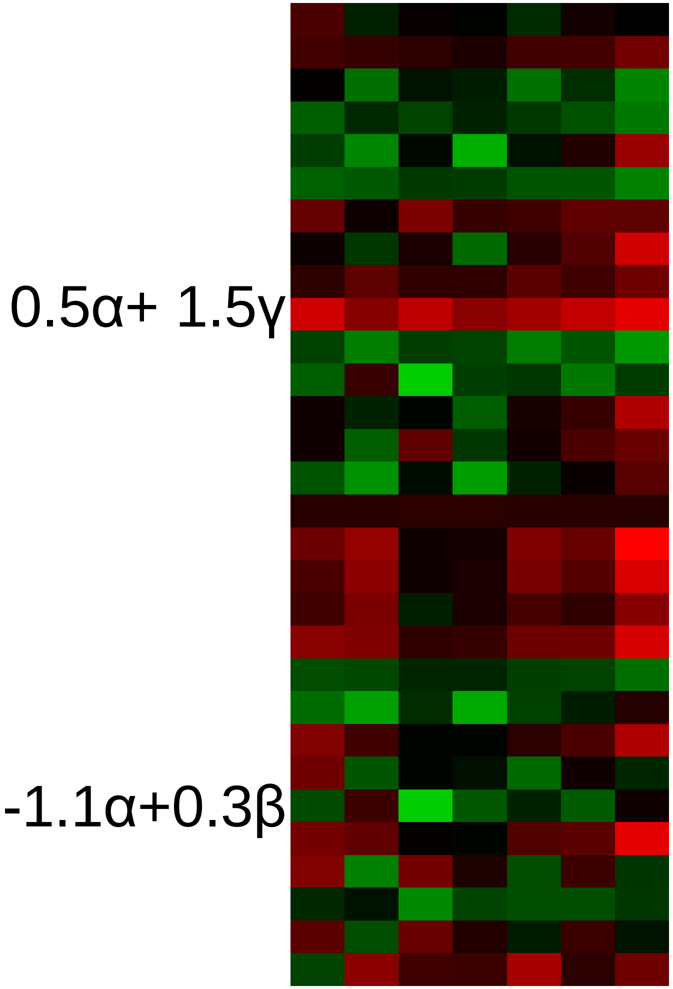
MIT EECS

Genes within cluster follow same expression pattern – deviation from cluster consensus is noise

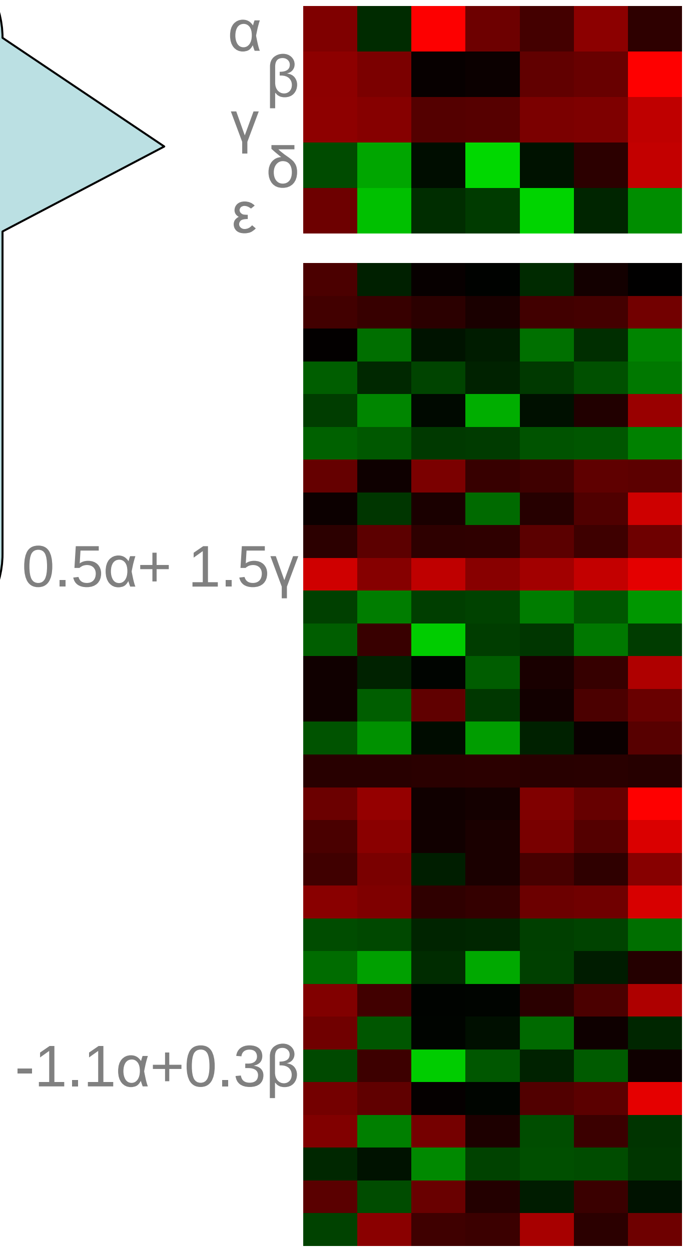


Genes within cluster follow same expression pattern – deviation from cluster consensus is noise



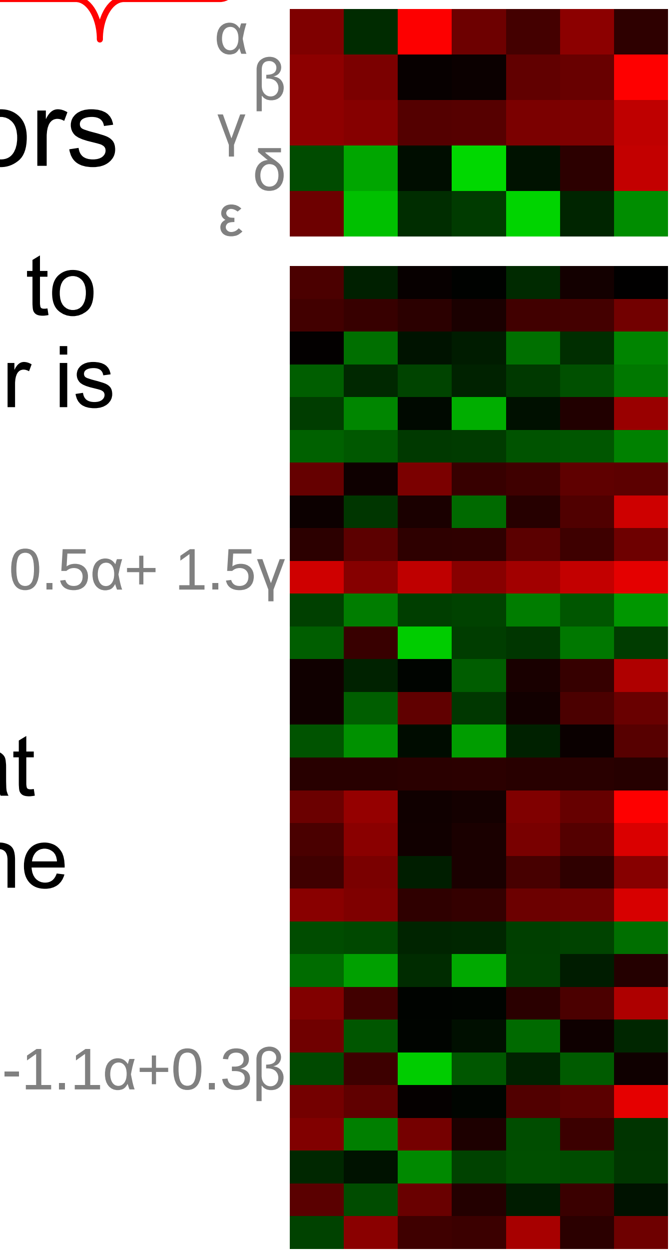


- Transcriptional factors
- Regulatory cascades
- Responses / stimuli
- Processes
  - Protein complexes
  - Pathways
  - Cell activites

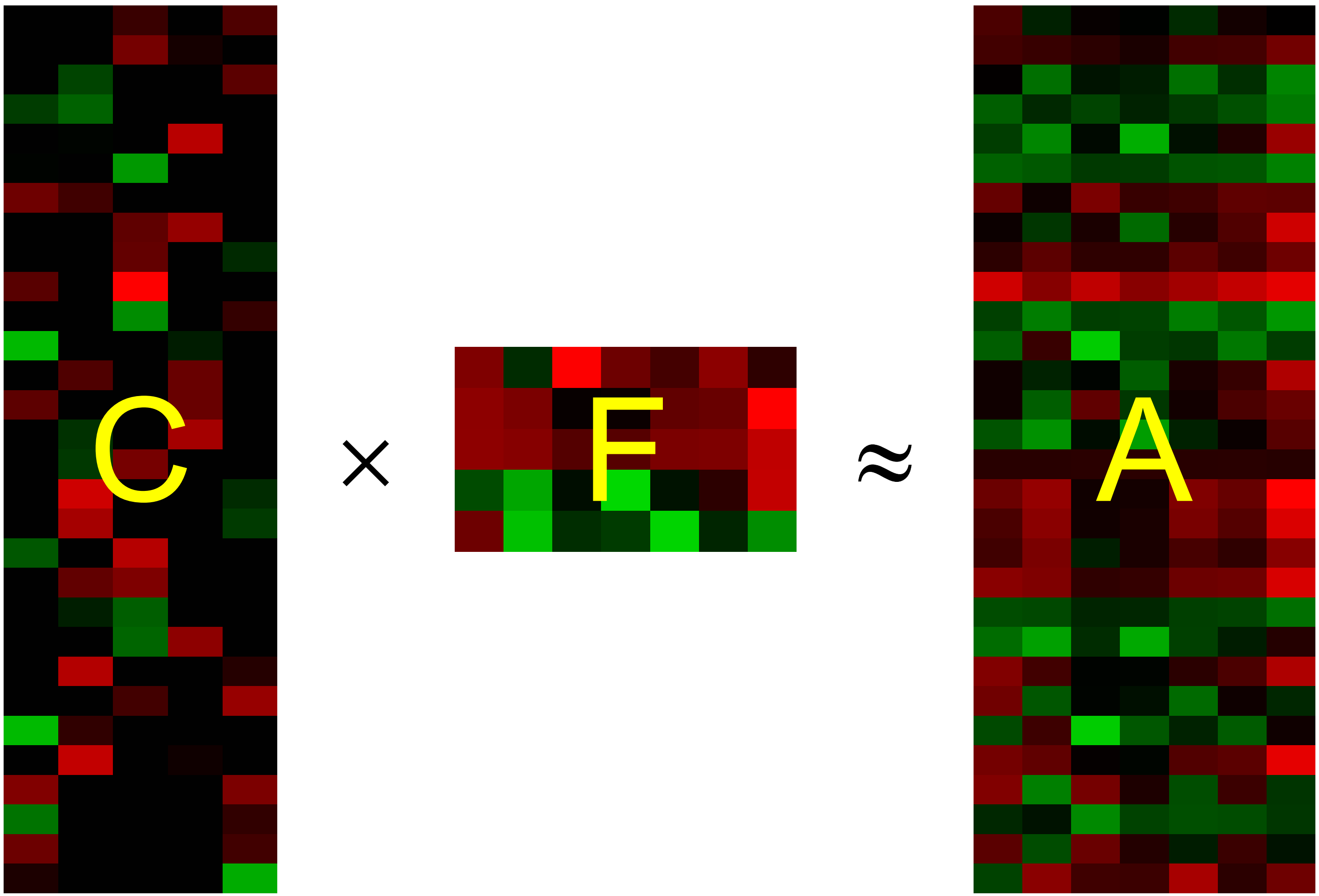


Modeling Data as Linear Combinations of Factors

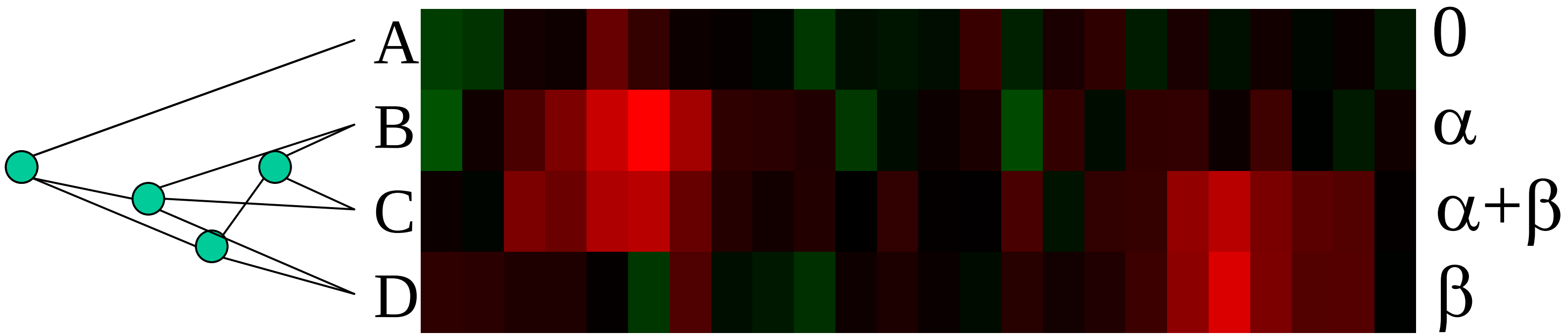
- Instead of being assigned to a cluster, each data vector is a linear combination of ‘factors’.
- ‘Factors’ represent basic structural components that are combined to get the the data vectors



Modeling Data as Linear Combinations of Factors



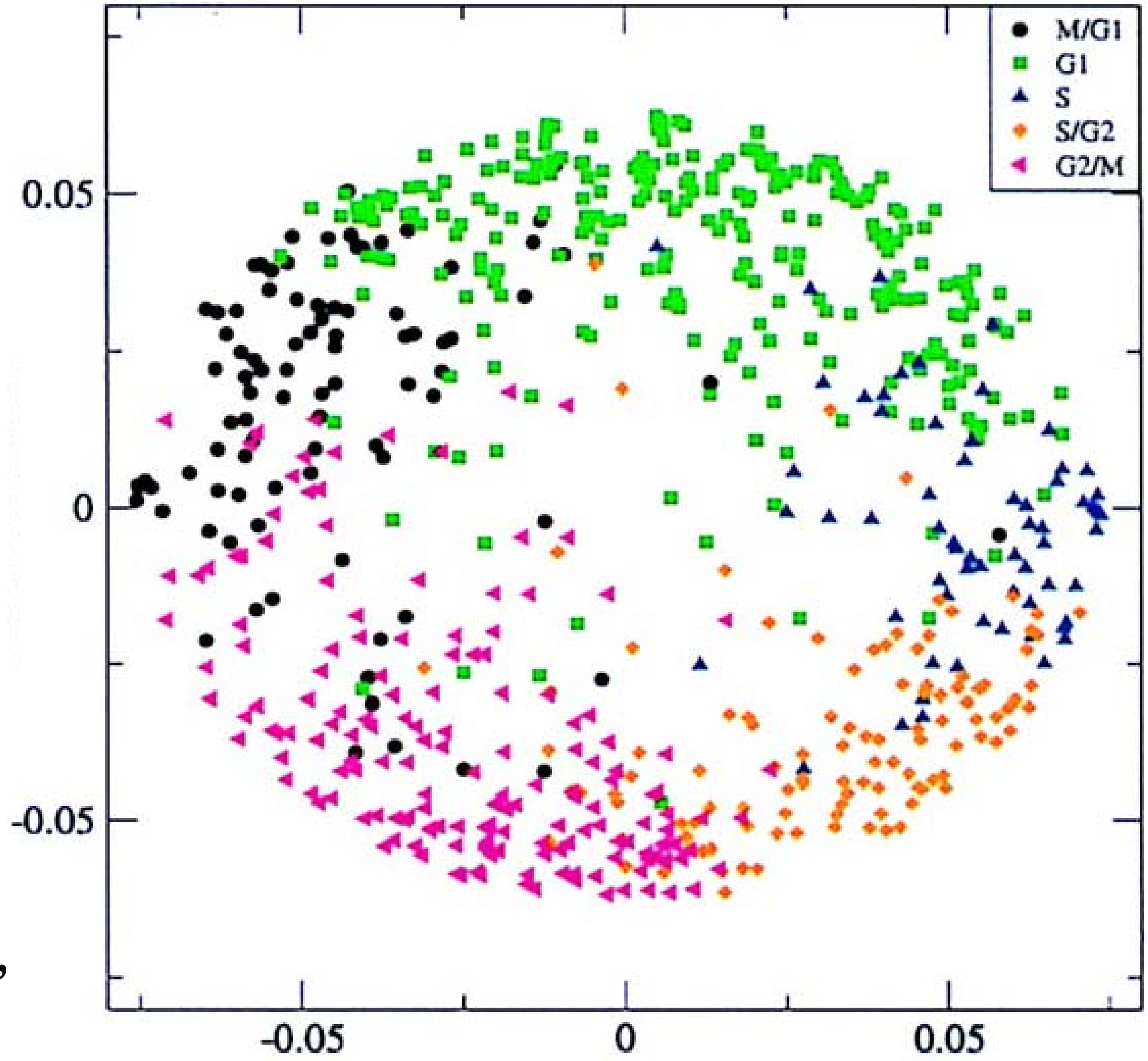
Limitations of hierarchical clustering



SVD Analysis of Gene Expression Patterns

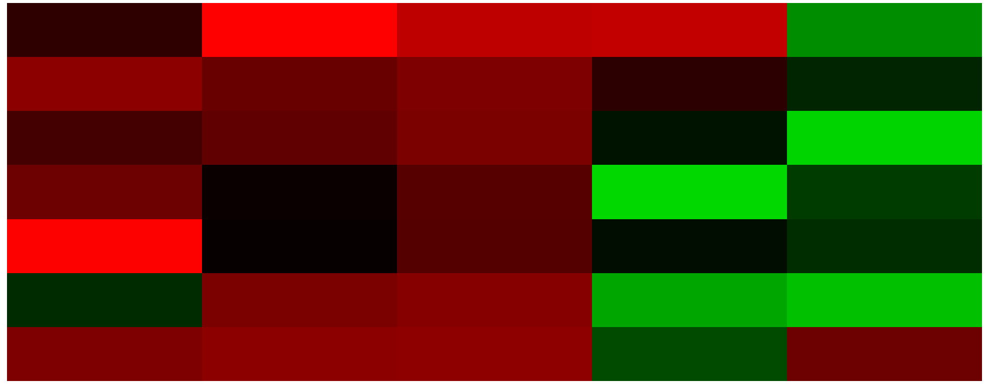
- Alter, Brown, Botstein: PNAS 2000
  - Yeast cell-cycle
- Raychaudhuri, Stuart, Altman: PSB 2000
  - Yeast cell-cycle and sporulation; serum-treated human fibroblast
- Holter et al: PNAS 2000
  - Yeast cell-cycle

Expression of cell-cycle genes projected to leading two eigenfactors:

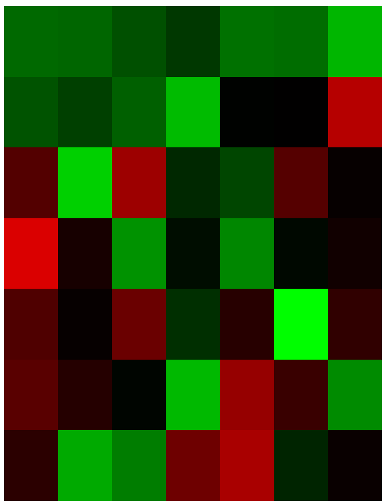


Holter, et al,
PNAS 2000

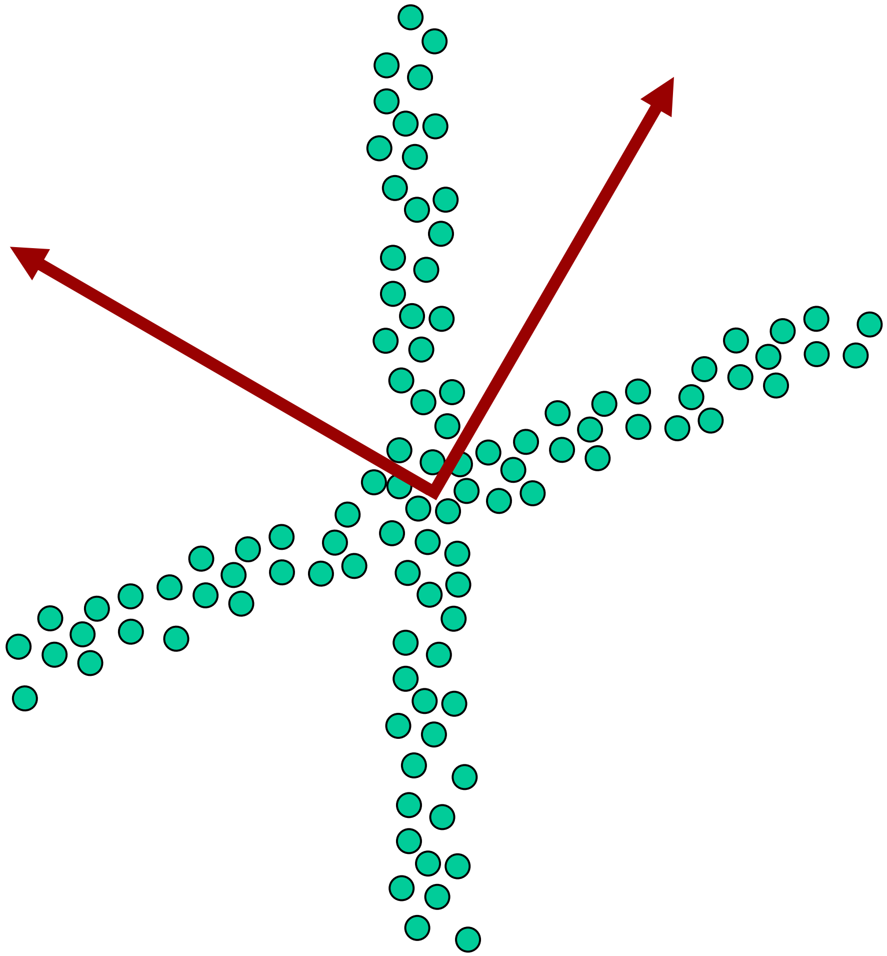
True (Planted) Factors



Eigenfactors



-0.6431	-0.9408	-0.9932	-0.0034	0.6628
-0.4274	0.2364	-0.0917	0.9128	-0.0681
0.6019	-0.0967	-0.0332	0.3965	0.5372
-0.2036	0.2226	0.0622	-0.0977	0.5172
0.0067	0.0144	-0.0130	-0.0069	0.0017
-0.0012	-0.0025	0.0003	0.0006	-0.0037
-0.0033	0.0019	0.0003	-0.0001	0.0048



SVD recovers subspaces – eigenfactors describe them

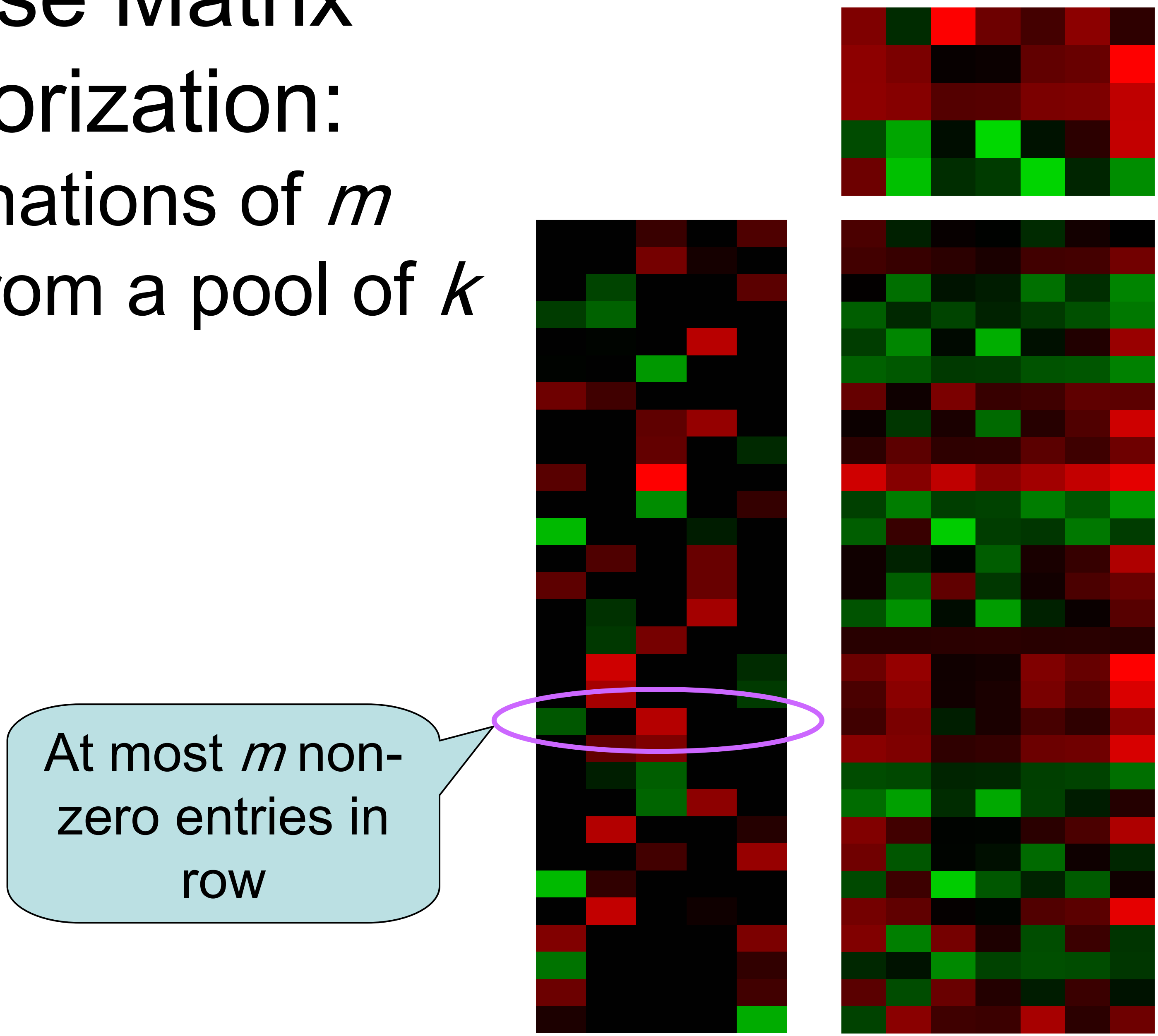
Are eigenfactors interpretable ?

- Degrees of freedom in choosing factors
- Is orthogonality desired ?
- Can only reconstruct a few factors (<<dimension)
- Additional eigenfactors used to refine non-linear interactions, instead of corresponding to new factors

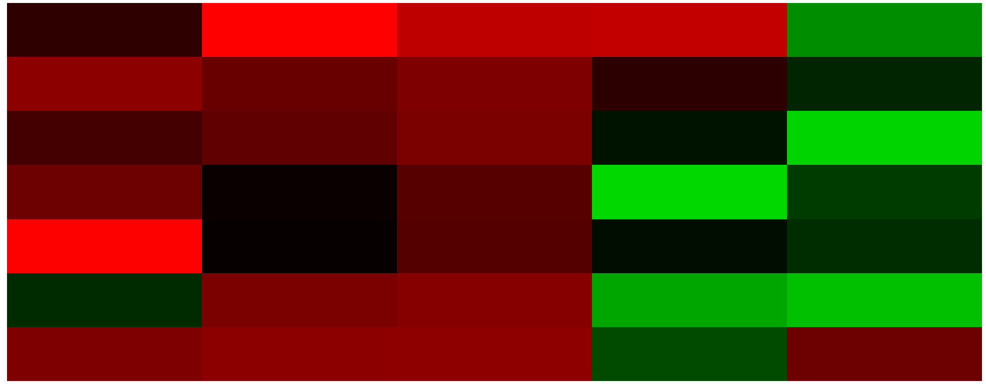
Does each data vector really depend on *all* factors ?

Sparse Matrix Factorization:

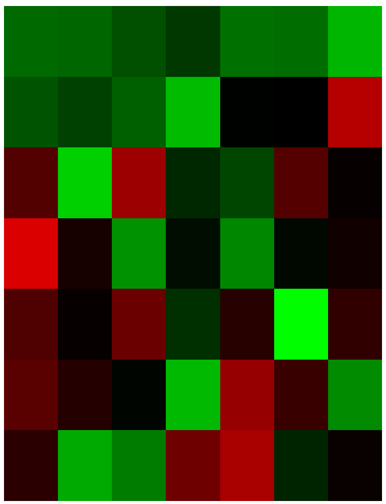
combinations of *m* factors, from a pool of *k*



True (Planted) Factors

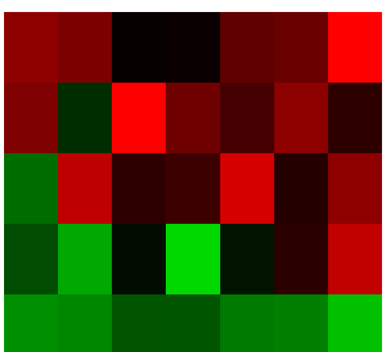


Eigenfactors



-0.6431	**-0.9408**	**-0.9932**	-0.0034	**0.6628**
-0.4274	0.2364	-0.0917	**0.9128**	-0.0681
0.6019	-0.0967	-0.0332	0.3965	0.5372
-0.2036	0.2226	0.0622	-0.0977	0.5172
0.0067	0.0144	-0.0130	-0.0069	0.0017
-0.0012	-0.0025	0.0003	0.0006	-0.0037
-0.0033	0.0019	0.0003	-0.0001	0.0048

SMF Factors



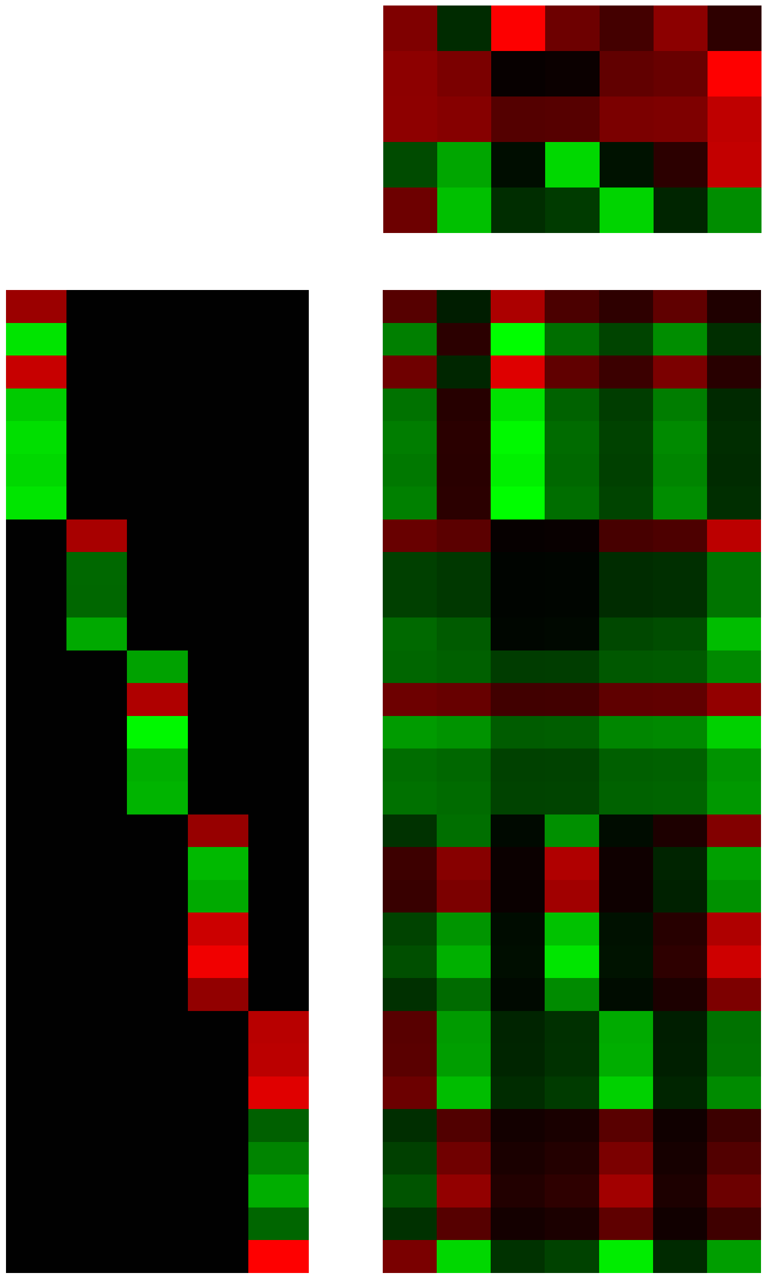
0.4006	**1.0000**	0.9303	0.1553	-0.5772
0.9999	0.3976	0.6425	-0.1286	-0.1773
0.1814	0.5768	0.6388	-0.0995	**-0.9999**
-0.1335	0.1487	-0.1094	**0.9999**	0.1056
-0.6435	-0.9306	**-1.0000**	0.0973	0.6378

Sparse Matrix Factorization

Sparsity (*m*)

←*sparse* *dense*→

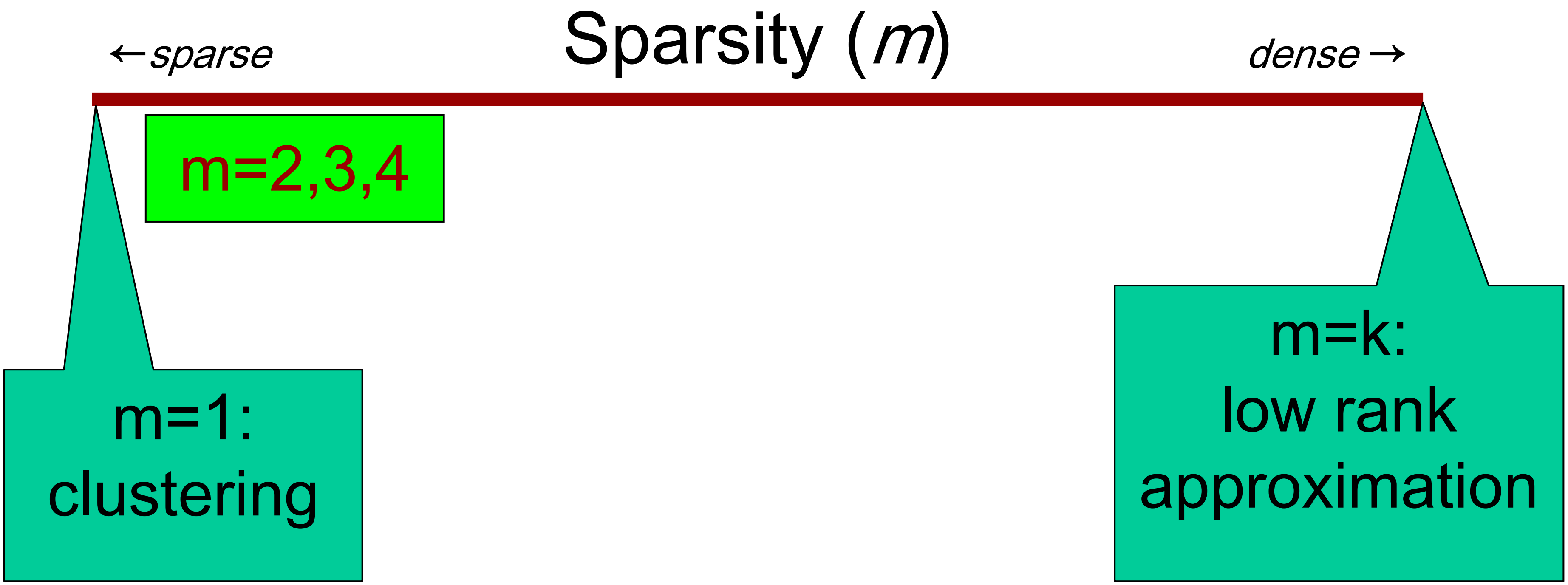
m=k:
low rank
approximation



If m=1, and coefficients are 0/1, matrix decomposition is equivalent to k-means clustering.

For general coefficients with m=1, matrix decomposition is equivalent to clustering with a correlation distance measure.

Sparse Matrix Factorization



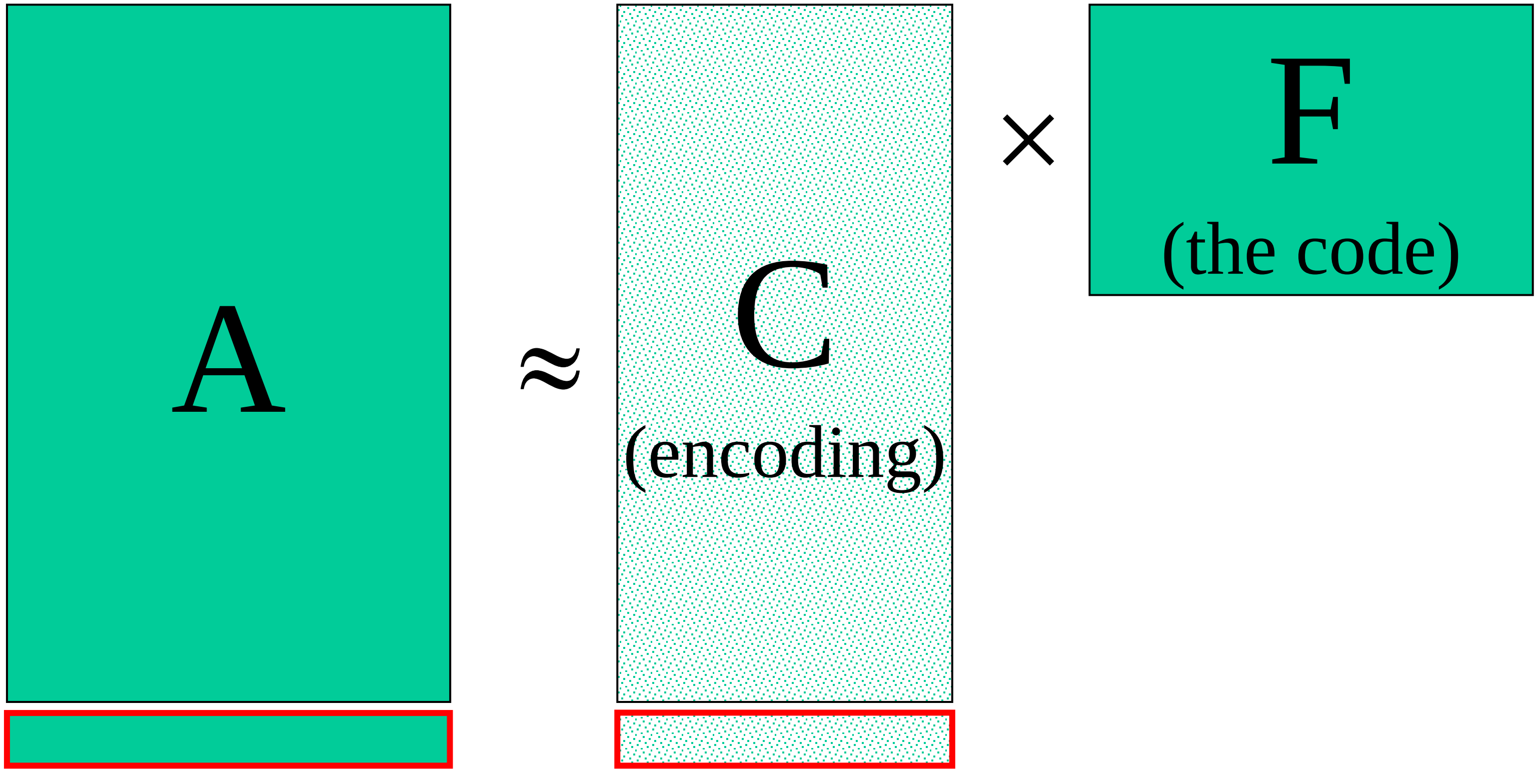
Sparse Matrix Factorization
(m>1, but small)

- Model limited interactions
- Recovery even with large number of factors (beyond dimensionality of data / width of data matrix).
- No* degrees of freedom in recovery.

  *except scaling and permutation

- *More interpretable factors ?*

An Encoding of the Data



Constraints reduce description length

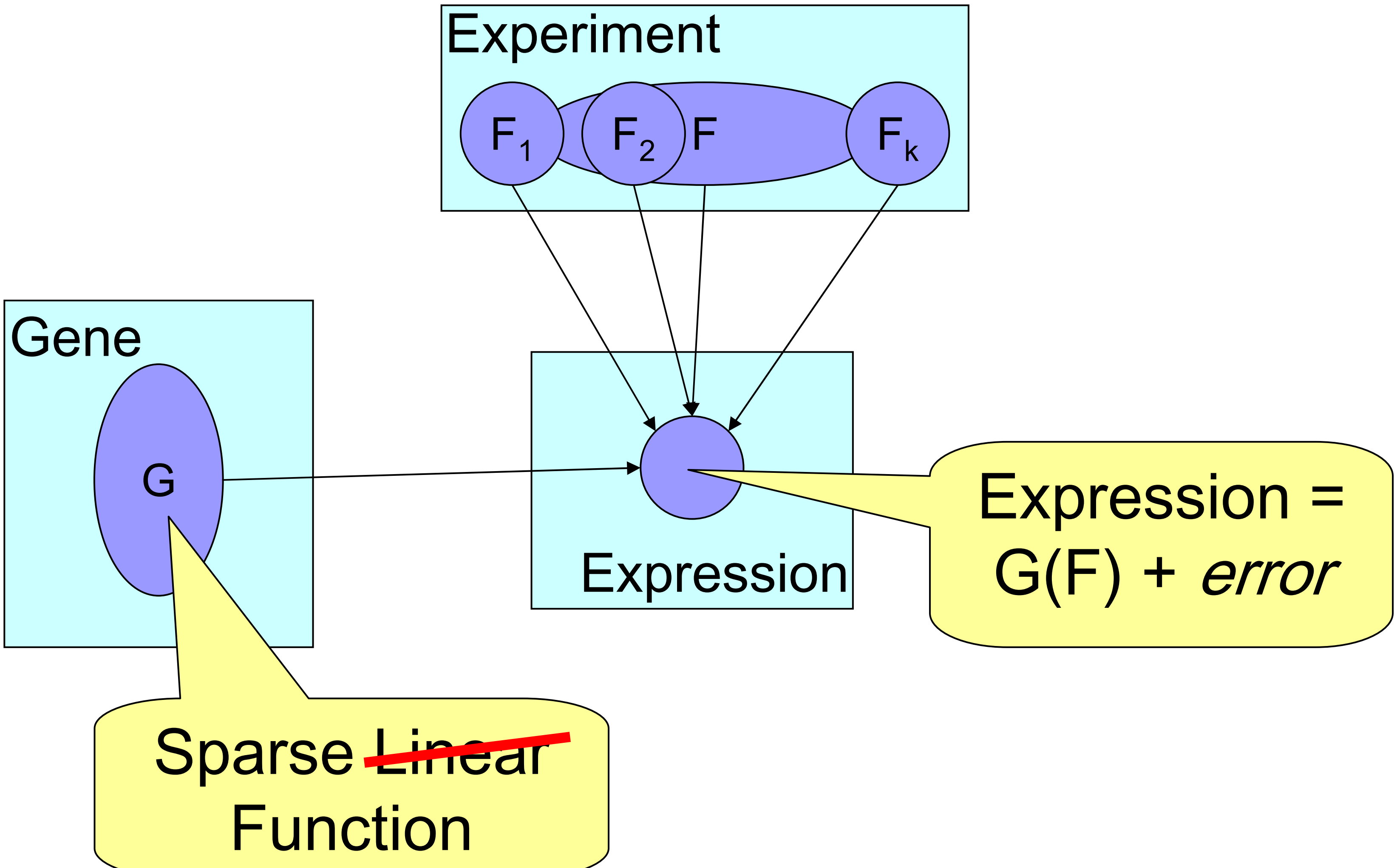
Constrained Matrix Factorization

Lee & Seung, NIPS 97, Nature 99, NIPS 00

- Conic (non-negative coefficients)
- Convex (stochastic coefficients)
- Non-negative coefficients AND factors

Non-negativity appropriate for gene expression?

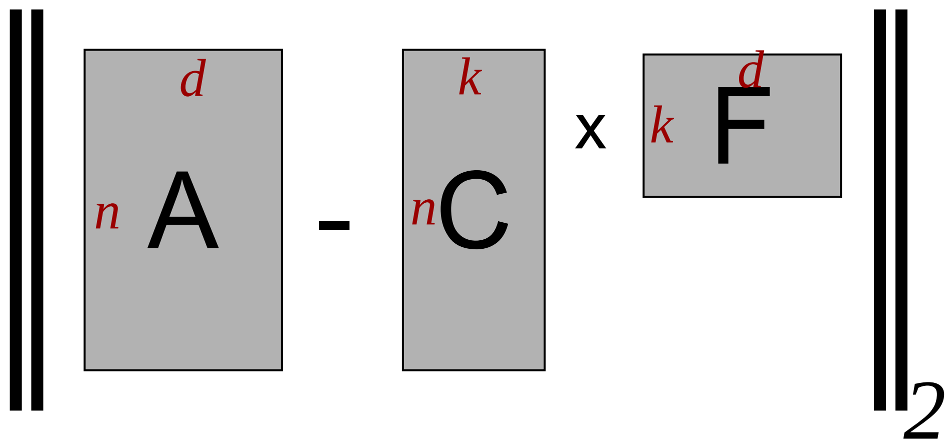
Viewed as PRMs



Reconstructing a SMF from (noisy) Data: An Optimization Problem

Finding SMFs

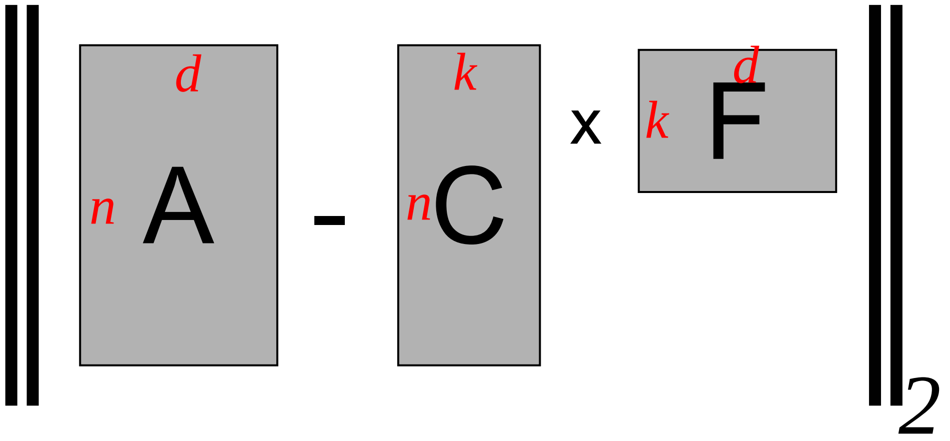
Given *A*, find *C*,*F* that minimize



Subject to: at most m non-zero entries in each row of *C*

Iterative Alternate Optimization

Optimize *F* given *C*, and *C* given *F*



Generalization of k-means clustering

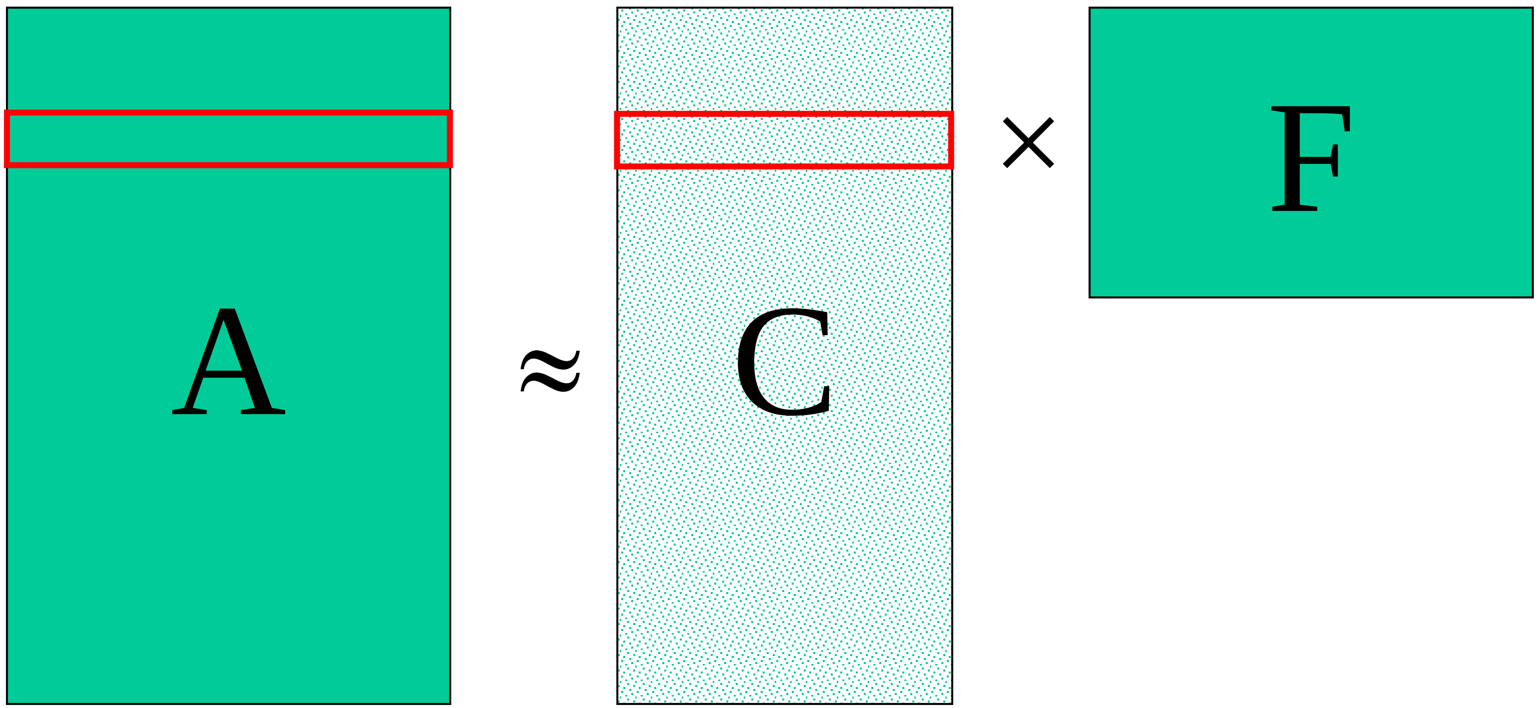
Iterative optimization

- For fixed *C*, finding optimal *F* is easy:

$$A \approx CF \Rightarrow F = pinv(C)A$$

- For fixed *F:* each row of *A* should be projected to a subspace spanned by *m* of the rows of *F*

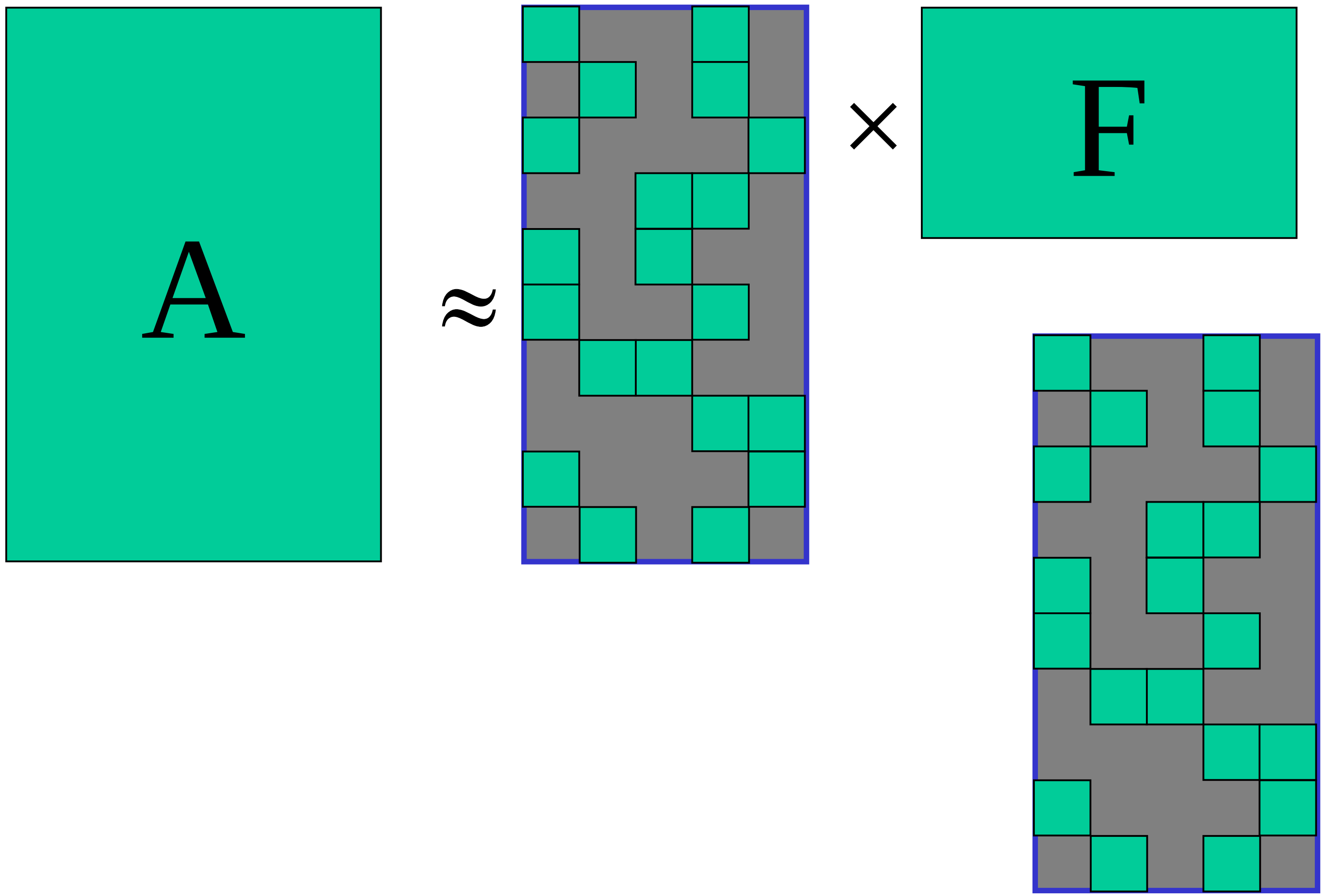
Optimizing *C* for fixed *F* *(decoding)*



Optimizing *C* for fixed *F* *(decoding)*

- For each row, find best projection to subspace spanned by *m* of the rows of *F.*
  - need $\binom{k}{m} n$ projections
  - Perhaps with geometric data structure $\binom{k}{m} + n$
- Heuristic approach: change one coefficient at a time
  - With other coefficients fixed (simple projection)
  - With only coefficient *mask* fixed

Optimizing *C*,*F* for fixed mask



Initializing the Factors *F*

- Where do we start our alternate-maximization search ?
- In k-means: start with random rows of *A*
  - Problematic for SMF: to close to local minima with factors resembling cluster centers.

Jumping out of local minima

- Instead of restarting from scratch, keep the useful factors, replace the less-used factors.
- Can measure the effect of each factor on reducing the error.
- Back to a familiar problem: how do we pick new factors to replace those removed?

- Regularization penalty promoting sparseness (instead of hard constraint)

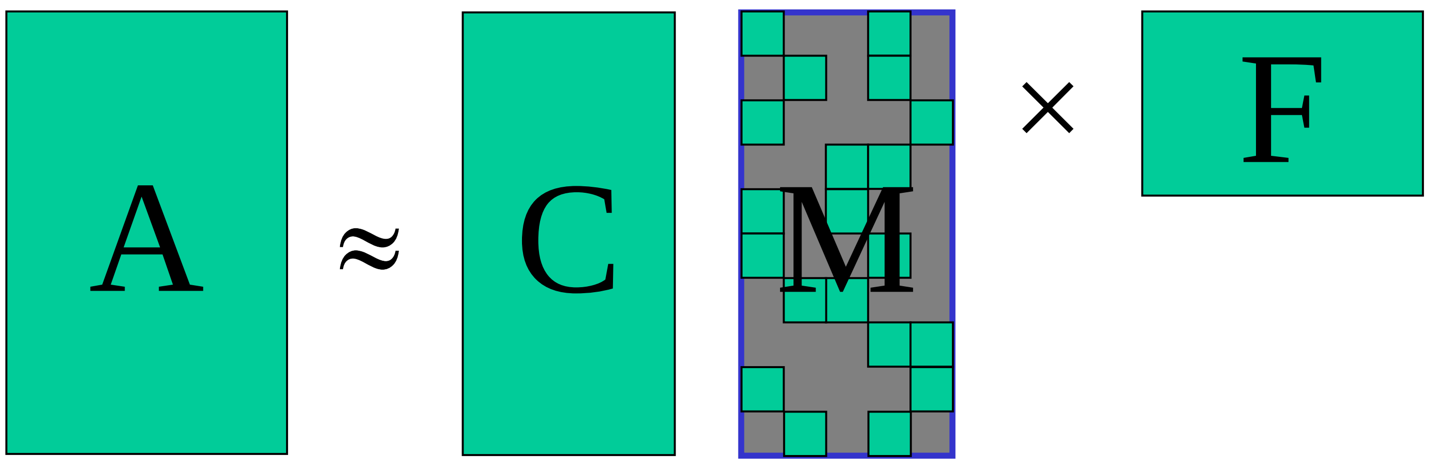
- EM instead of MM:
  - Search for distribution over C
  - Optimize F for C=E[C|F]

Maximum Entropy Setting

$$\min D\left(Q \| P_0\right) \quad \text{s.t.} \quad \left\| A - \mathrm{E}_Q\left[(C.M)\right]F \right\|_2 < R$$

$$P_0(M_{i,j}) \sim Bernoulli(q) \qquad P_0(C_{i,j}) \sim N(0,\sigma^2)$$

$$P_0(C,M) = P_0(C)P_0(M)$$



SMF with partially known *C,F*

- Some factors are known:
  - How well can they combine to explain data?
  - Find additional factors beyond known ones

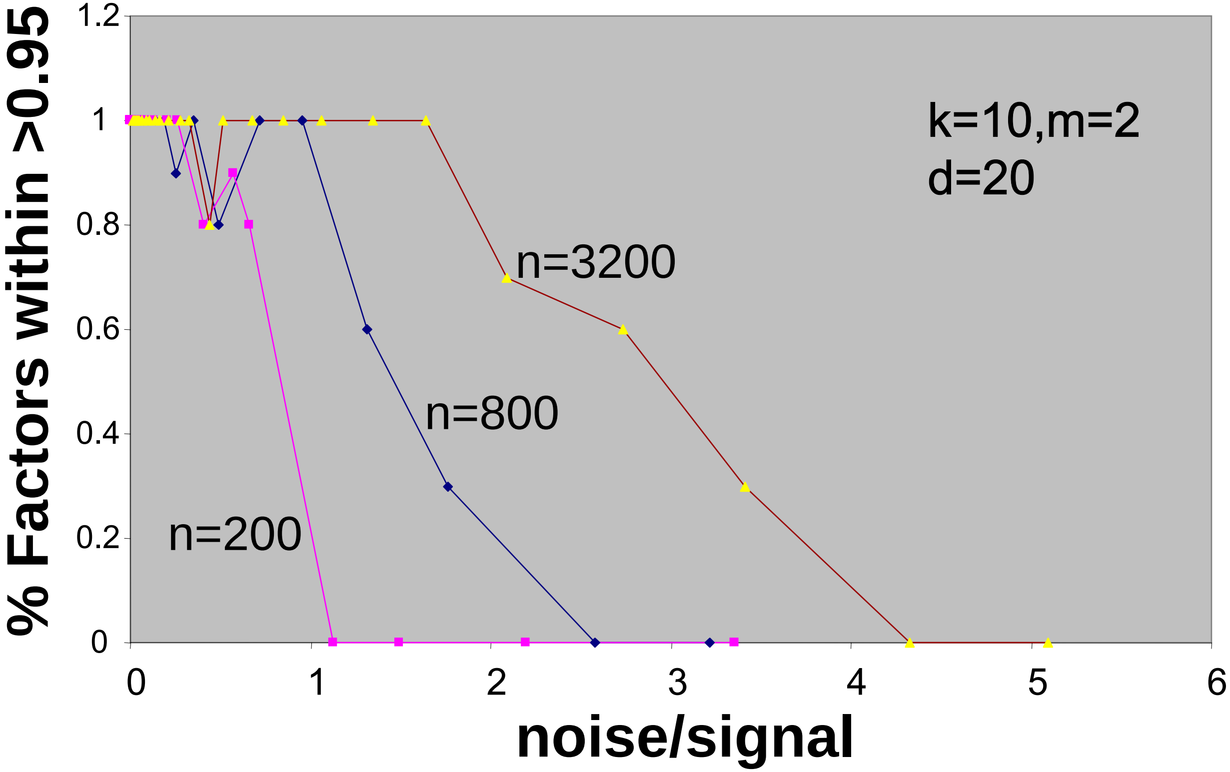
- Combined with factor localization data: partial knowledge about coeficients

Reconstructing a SMF from (noisy) Data: A Statistical Problem

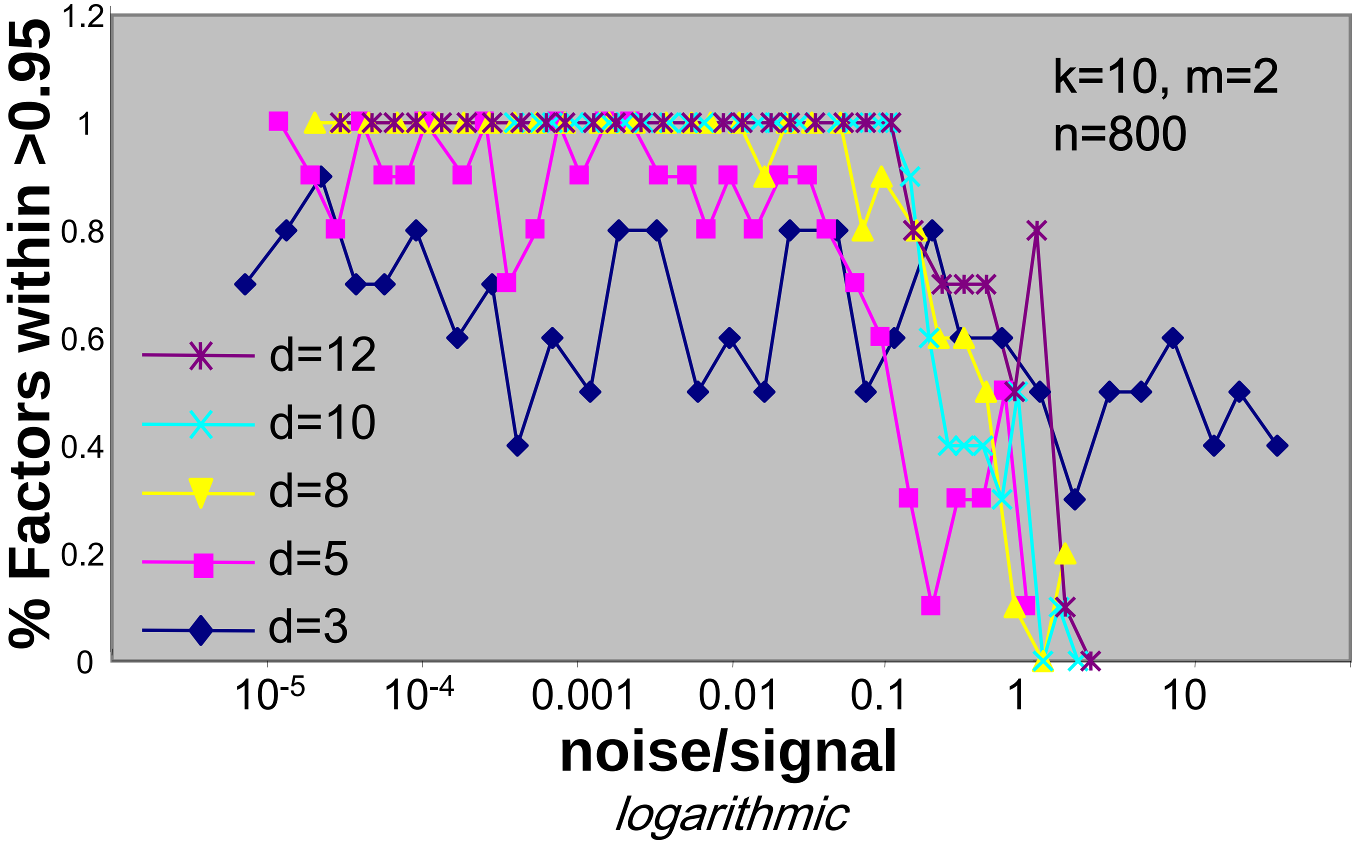
For $A = C \times F + E$, up to what level of noise is $C \times F$ the optimal factorization ?

Measure: correlation of reconstructed F to true F, as a function of $Var(E)/Var(C \times F)$

Reconstruction in the Presence of Noise



Reconstruction in the Presence of Noise – low dimension



Current directions

- Better optimization methods
- Investigating the SMF of expression data (cell cycle, stress response)
- Model selection: choosing k,m