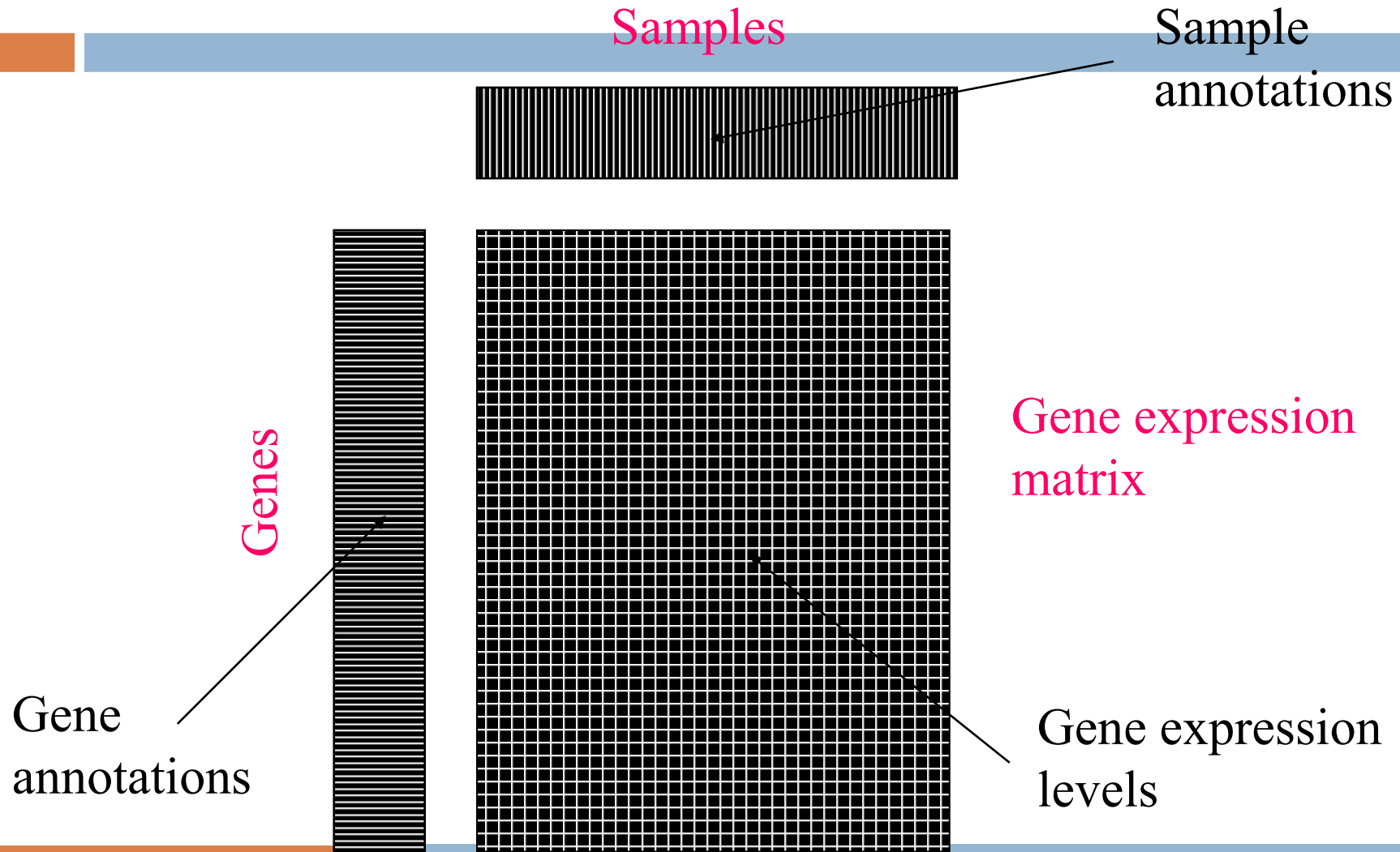


Bioinformatika

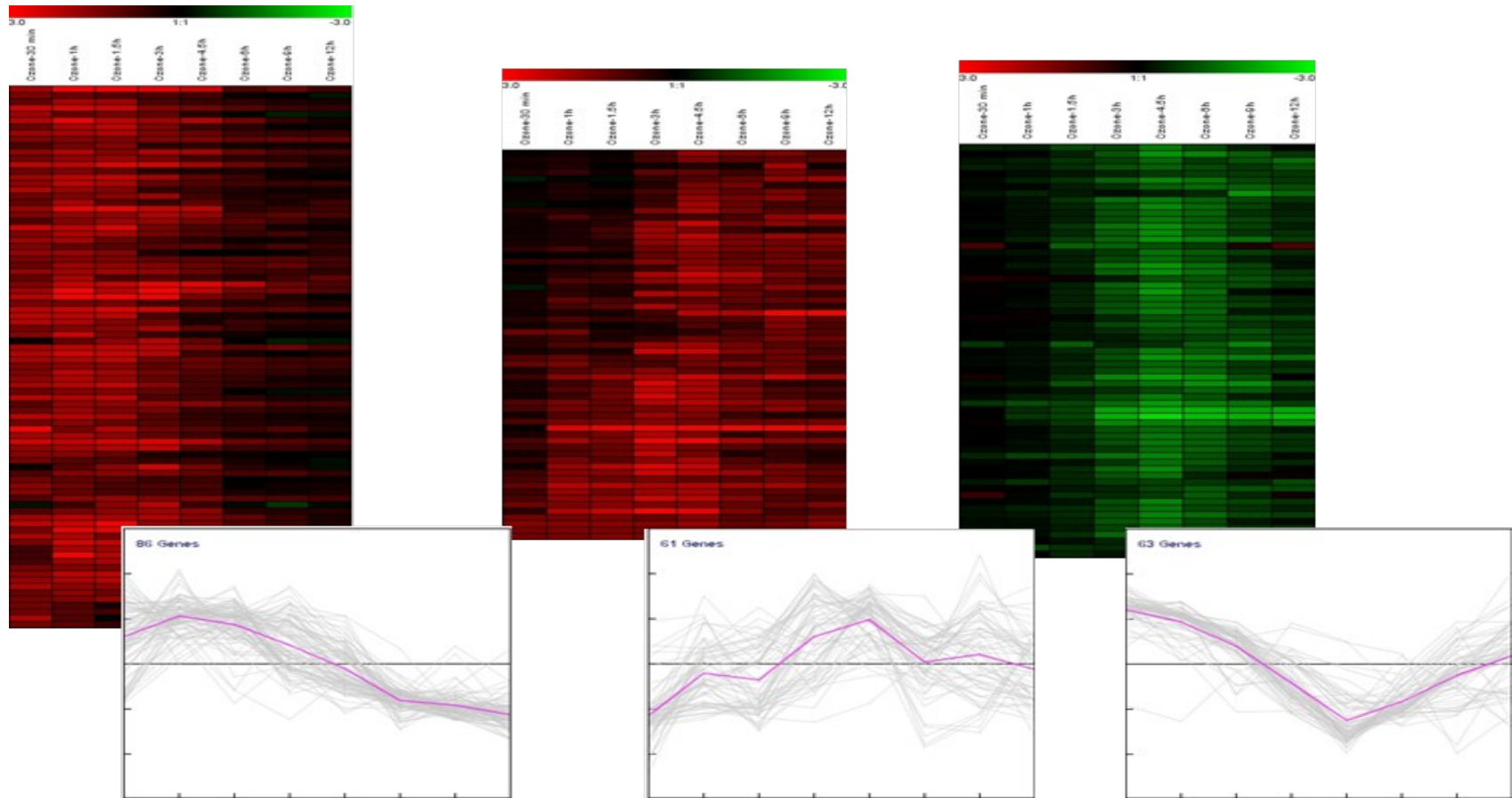
Gene expression data analysis

Michael Anděl
(some slides are courtesy of Mark Craven, U. of Wisconsin)

GE data – conceptual view



GE data – image view



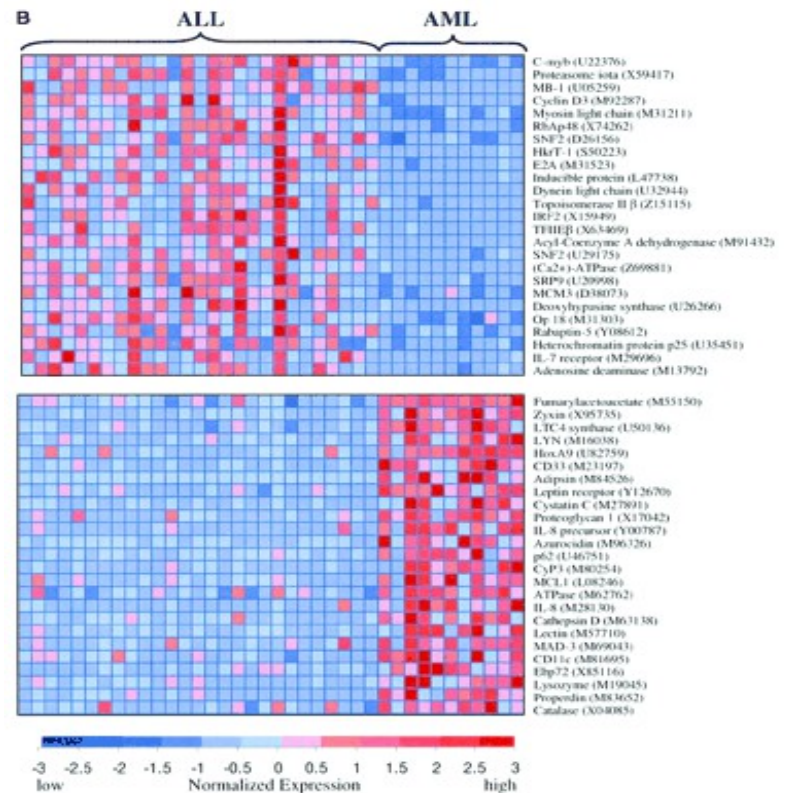
Classification task

Challenge:

- samples (10^1) x features (10^3)
- False hypotheses, overfitting
- Interpretability: are the expressed genes the causal ones?

What to do?

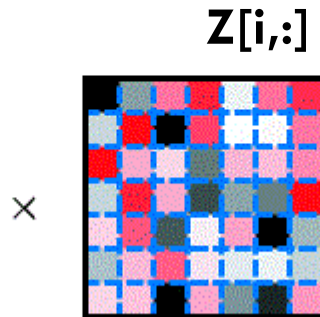
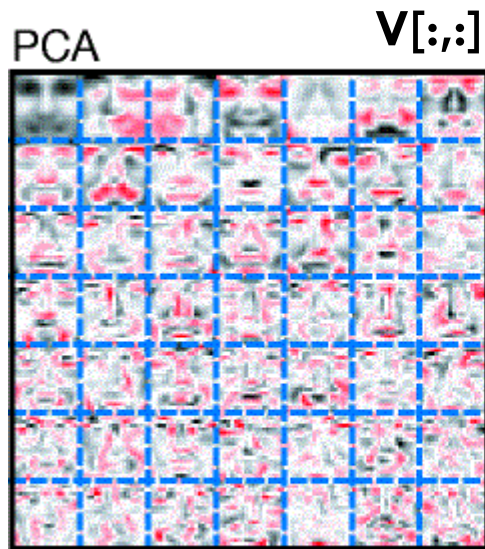
- Decrease number of hypotheses
- Analyze more abstract entities than genes, eg. principal components



Golub et al.: *Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring.* Science, 1999

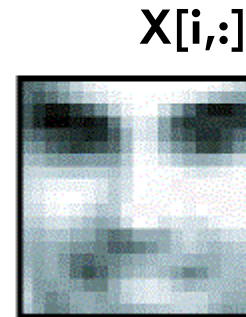
PCA – motivation

- M ... # genes
- N ... # samples
- X ($N \times M$) ... GE data in the space of genes
- V ($M \times K$) ... transformation basis, eigengenes
- Z ($N \times K$) ... transformed GE data in the space of eigengenes
- K ... # of eigengenes, i.e. the number of underlying concepts



×

=



Lee et al.: *Learning the parts of objects by non-negative matrix factorization*. Science, 1999

Assignment

Workflow

- 1) Learn a decision tree on subjected data. Use Matlab class **ClassificationTree**, for learning use method **fit()**.
- 2) Visualize resulting model (function **view()**) and enumerate the training error (function **predict()**)
- 3) Estimate the real error of the model by the means of crossvalidation. Use script **cv11.m**.
- 4) Compare the training error and real error estimation. Why are so different?

Assignment

Analyzing in terms of PCAs

Transform the data to the reduced space of eigengenes. Use attached function **pca.m**:

- 1) Do the transformation for the number of components $K=5:5:64$, i.e. $\mathbf{Z} = \mathbf{X} \mathbf{V}(1:K,:)$.
- 2) For each of these transformed data \mathbf{Z} learn the decision tree.
- 3) Choose an appropriate model according to its training error and model complexity.
- 4) Evaluate the chosen model by the crossvalidation
- 5) Compare the error estimation with the one of simple tree.

Assignment

Interpreting the results

- 1) Have the genes from your model been already reported in previous studies as associated with leukemia in terms of mutations or polymorphisms? Use web tool Phenopedia HuGE Navigator.
- 2) Try to find some gene from most important PCAs of the tree under the disease terms 'Leukemia, Myelocytic, Acute' and 'Leukemia, Lymphocytic, Acute'.
- 3) To mine the genes from the important components of the tree use attached function **mineGenes.m** and **geneNames.mat**.
- 4) Similarly find some genes from the **component tree**.