



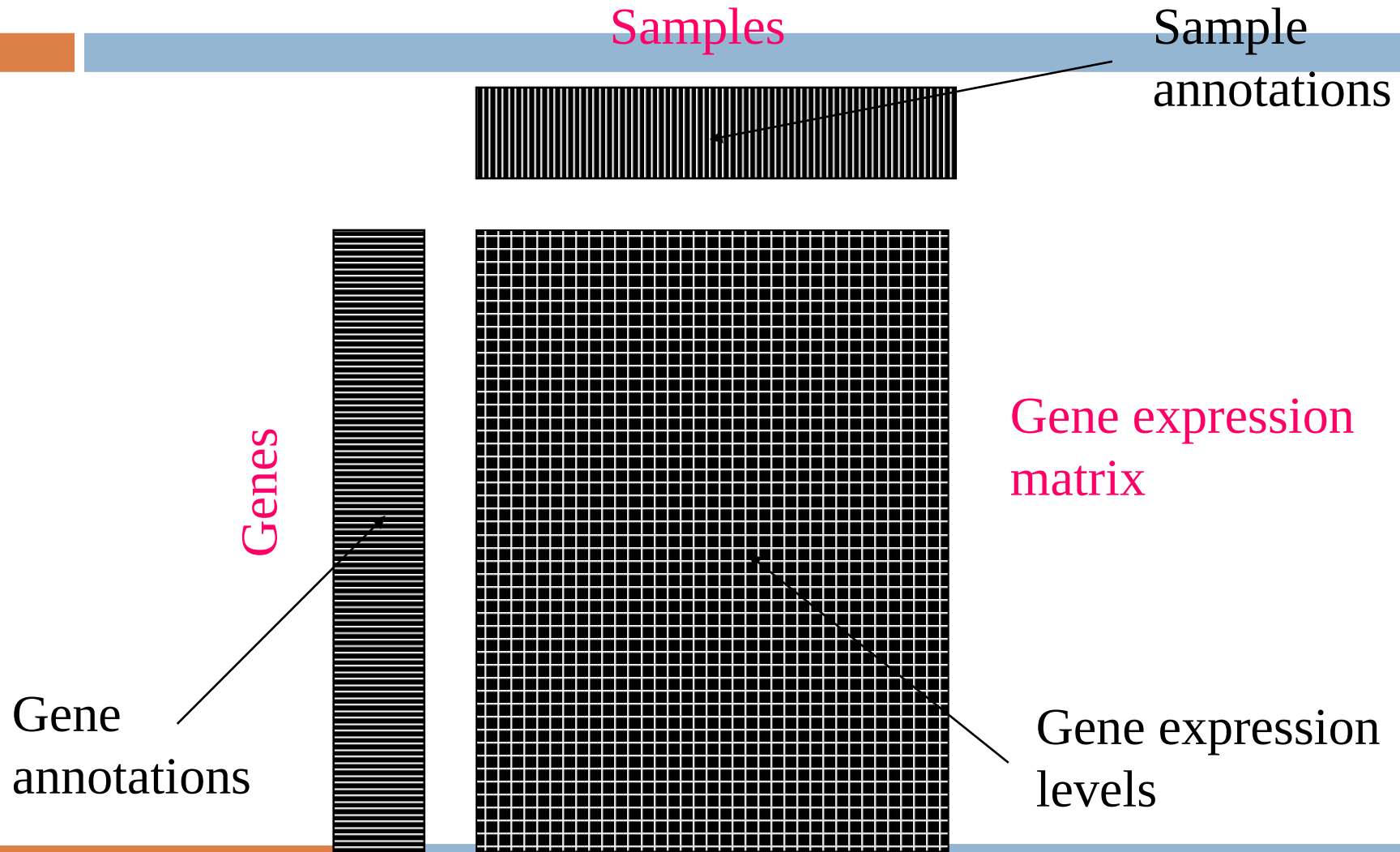
Bioinformatika

Gene expression data analysis

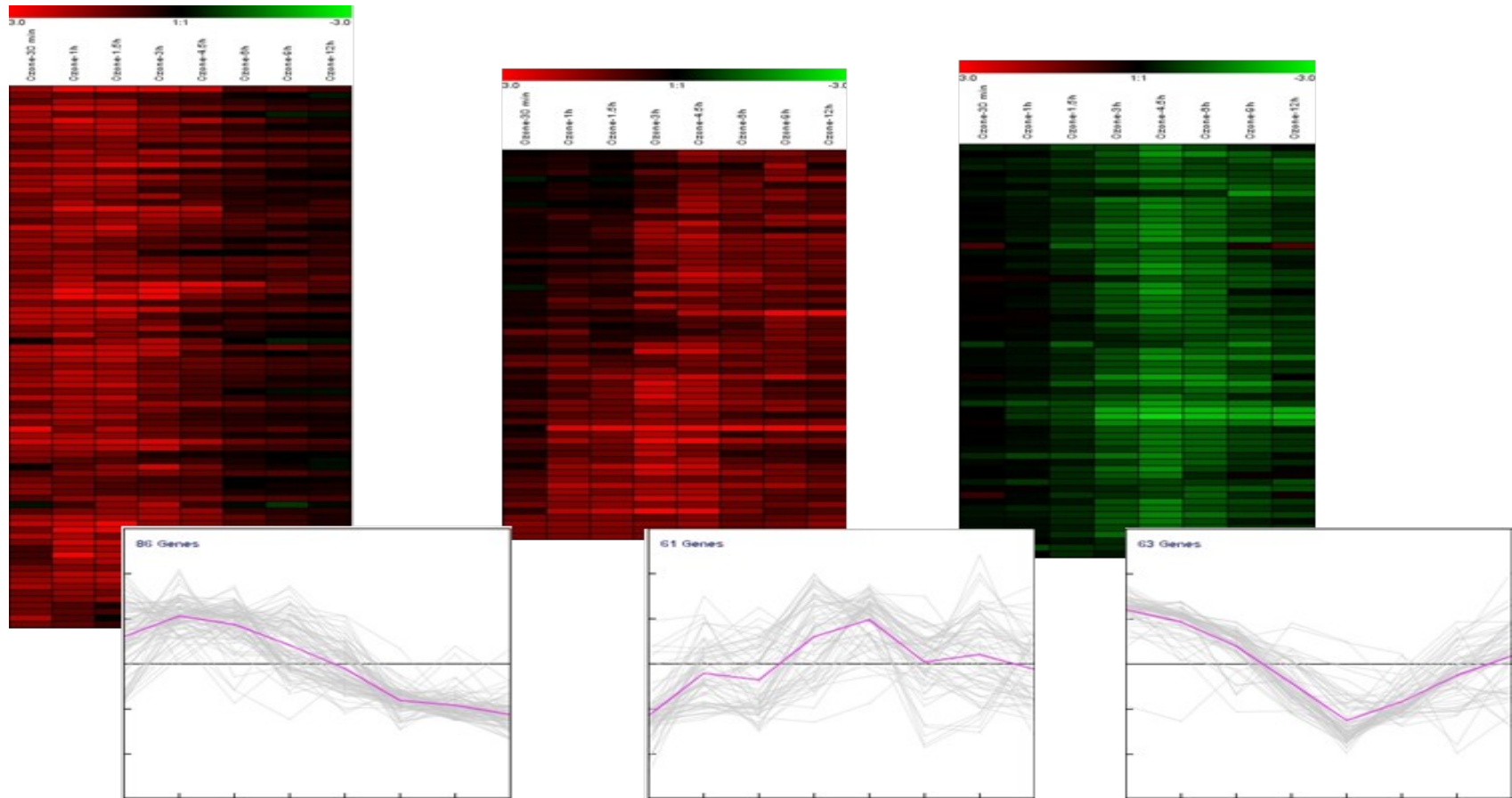


Michael Anděl

GE data - conceptual view



GE data - image view



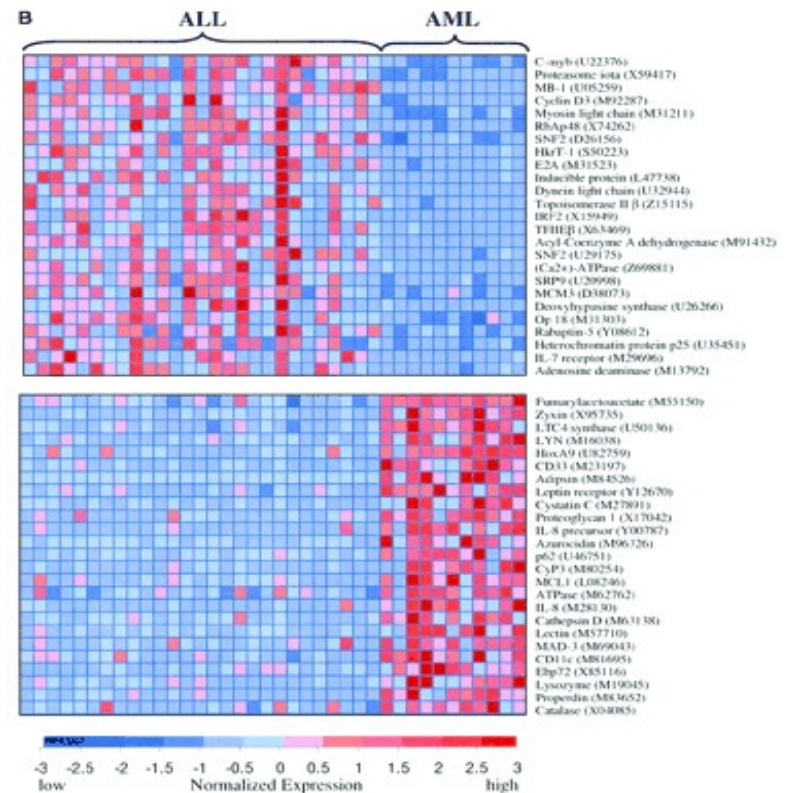
Classification task

Challenge:

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

What to do?

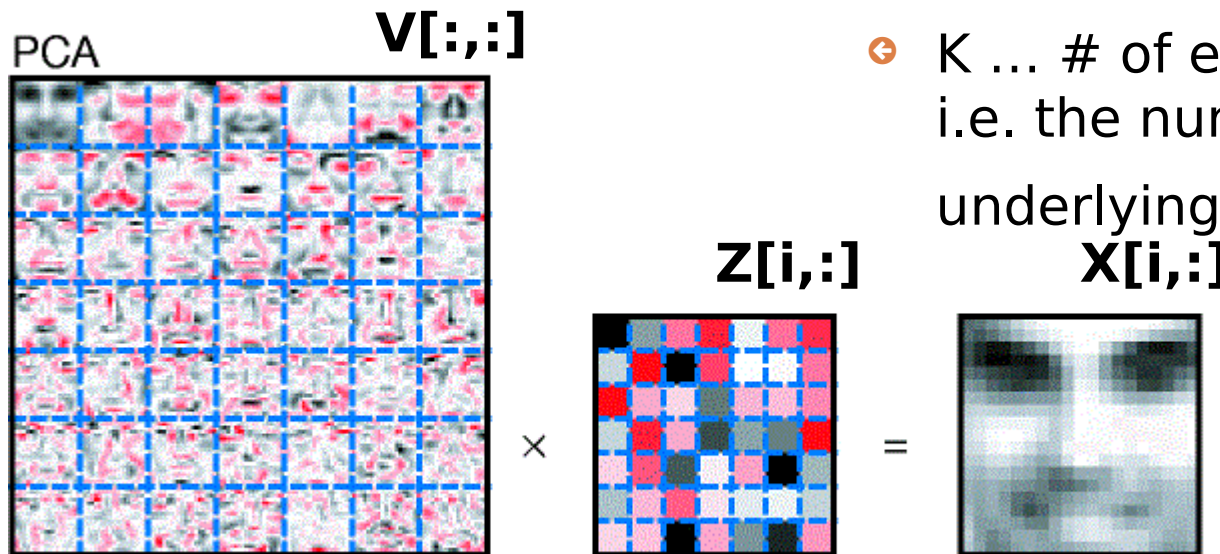
- ↻ Decrease number of hypotheses
- ↻ Analyze more abstract entities than genes, eg. principal components



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

PCA - motivation

- ↻ M ... # genes
- ↻ N ... # samples
- ↻ \mathbf{X} ($N \times M$) ... GE data in the space of *genes*
- ↻ \mathbf{V} ($M \times K$) ... transformation basis, eigengenes
- ↻ \mathbf{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

Data:

- 7,129 GE profiles of 72 patients
- 25 samples: acute myeloid leucaemia (AML)
- 47 samples: acute lymphoblastic leucaemia (ALL)

Golub, T., et al. *Molecular classification of cancer: class discovery and class prediction by gene expression monitoring*. Science 286.5439 (1999): 531-537.

Task:

- Construct decision model to differentiate these types of tumours. **Just complete the code in attached script `ge_cv.m`**

Assignment

Workflow I:

1. Learn a decision tree on subjected data. Use Matlab class `ClassificationTree` and its method `fit`.
2. Show the tree (method `view`) and enumerate its *training* accuracy. How would you interpret this model? Which gene is crucial for decision? Compare it with the original article. Have this gene been previously reported as Leucaemia associated? Use e.g. [phenopedia](#).
3. Estimate its *real* accuracy by the means of crossvalidation. Compare the two accuracies. Why are they so different?

Assignment

Workflow II:

1. Learn a basis matrix V of the data. Use attached function `pca.m`.
2. For $K = \{5, 10, 20, 50\}$, project the original data X to the top K components V . The result is reduced data $Z = XV_{1:K, :}^T$.
3. On these 4 transformations, learn decision trees, show them and enumerate their *training* accuracy.
4. Choose optimal decomposition and corresponding tree-model (use training accuracy and Occam razor). Estimate *real* accuracy of the optimal model by crossvalidation.

Assignment

Understanding the model:

1. Use attached function `mineGenes` to extract the genes, frequently appearing in the components presented in your tree-model
2. Resulting gene-sets, related to each component in the tree, interpret in overrepresented GO terms. Use e.g. [amigo](#).
3. Create a story!