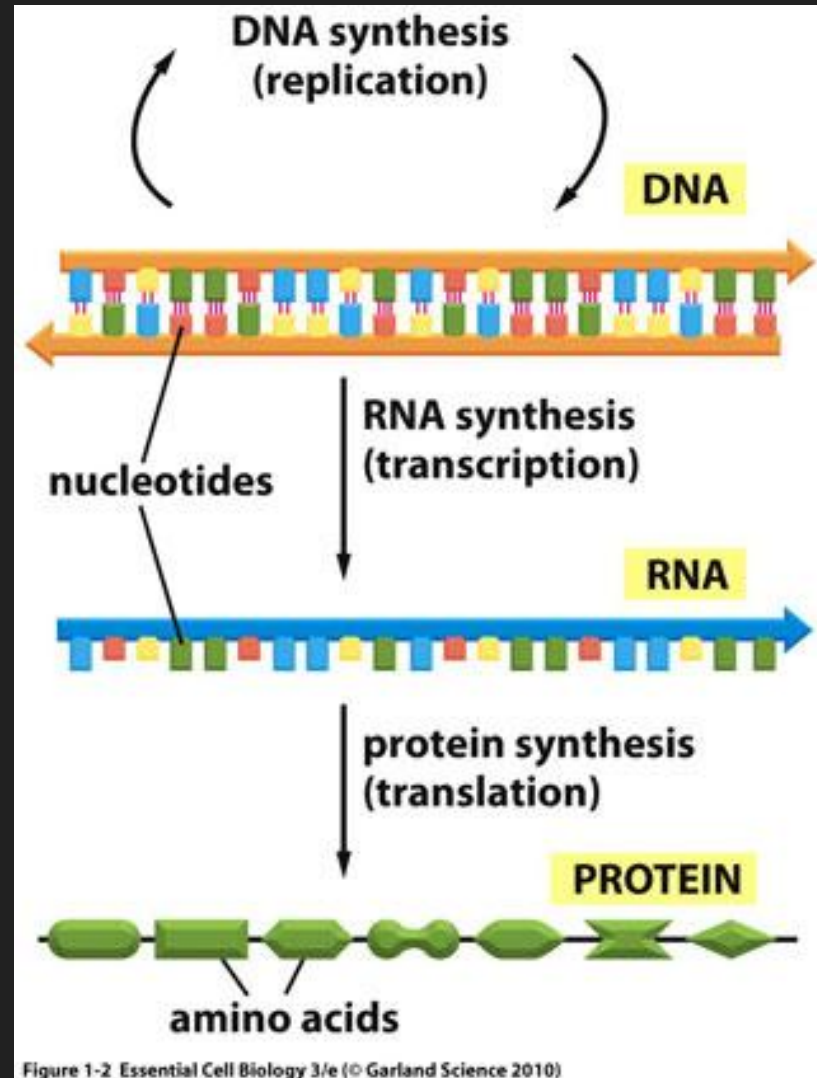


# Non-negative matrix factorization for gene expression analysis

Stevan Radanović 2015/3398  
stevan.radanovic@gmail.com

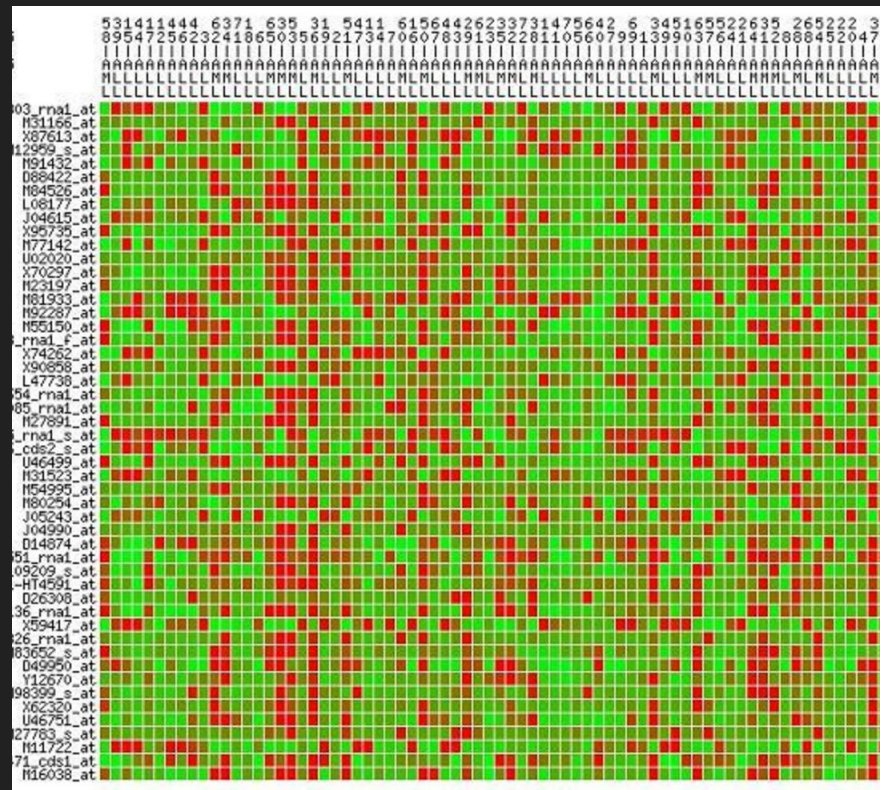
# Gene expression [01 | 13]

- Central dogma
- Proteins - workers
- RNA-Seq



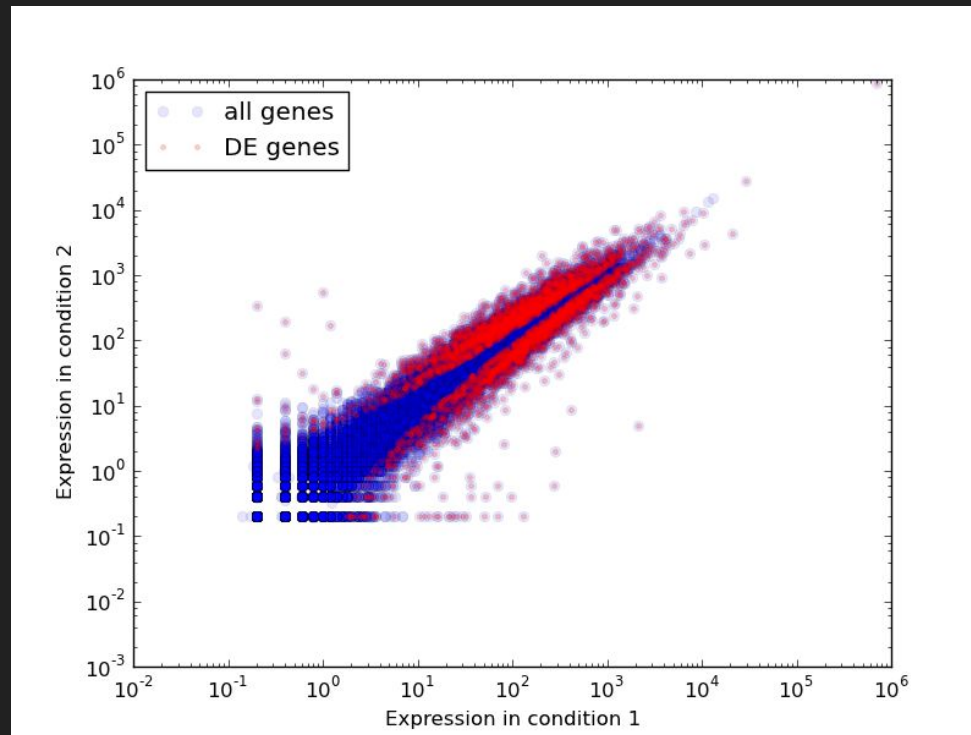
# Gene expression [02 | 13]

- RNA-Seq measures concentrations of RNAs
- Non-negative values



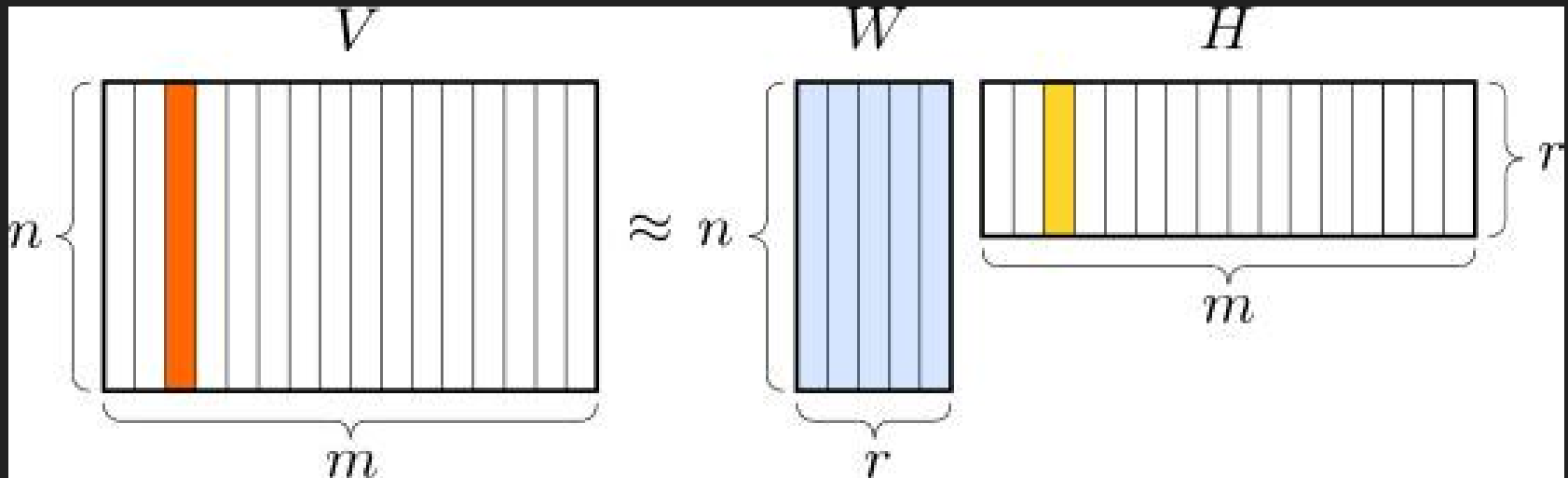
# Gene expression [03 | 13]

- Current experiments - labeled data
- Future - diagnostic testing?



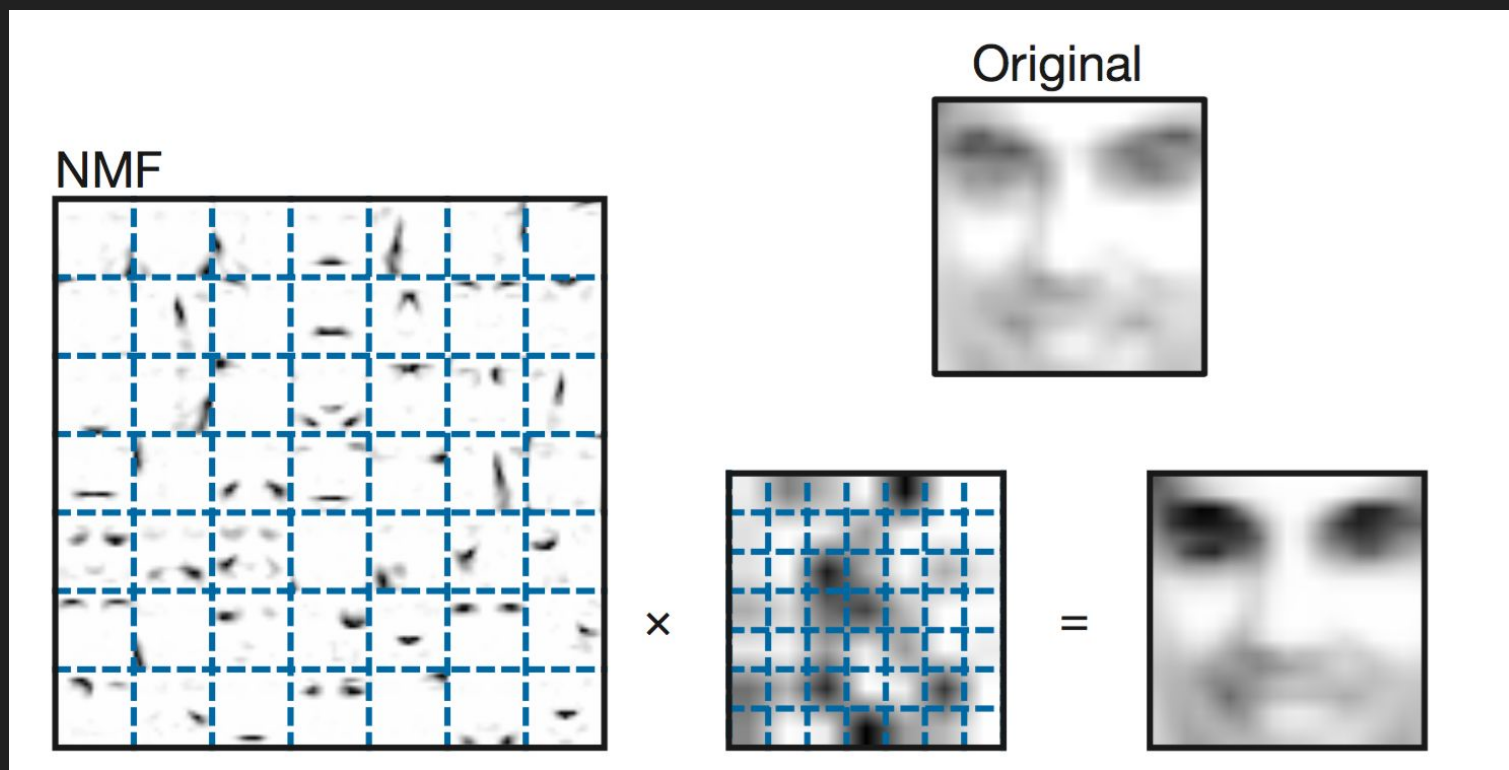
# NMF [04 | 13]

- Non-negative matrix factorization
- Feature extraction

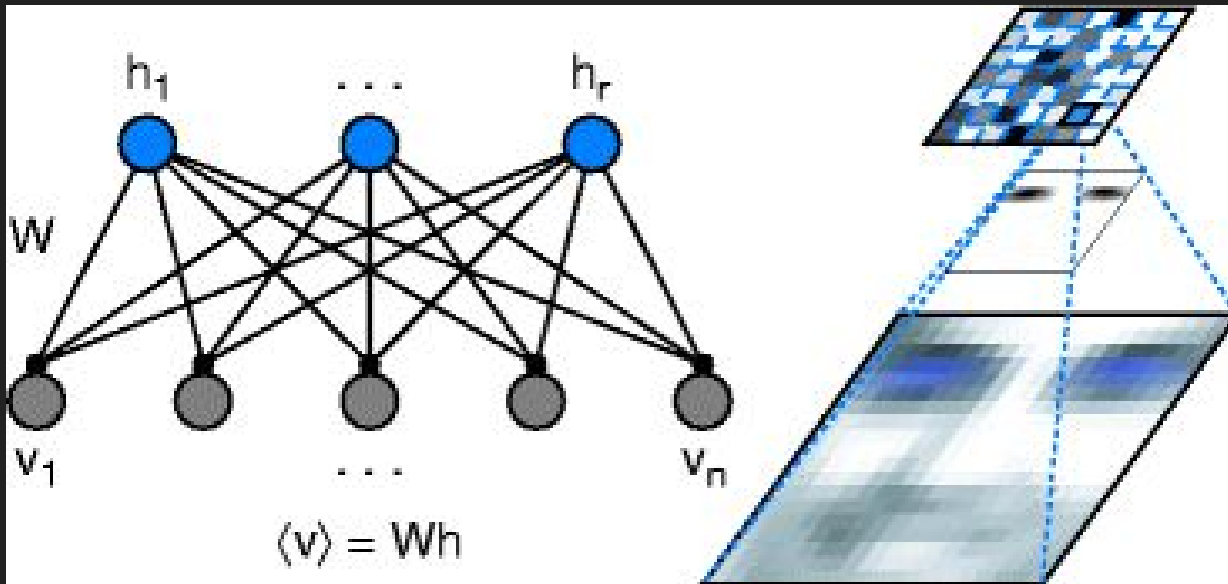
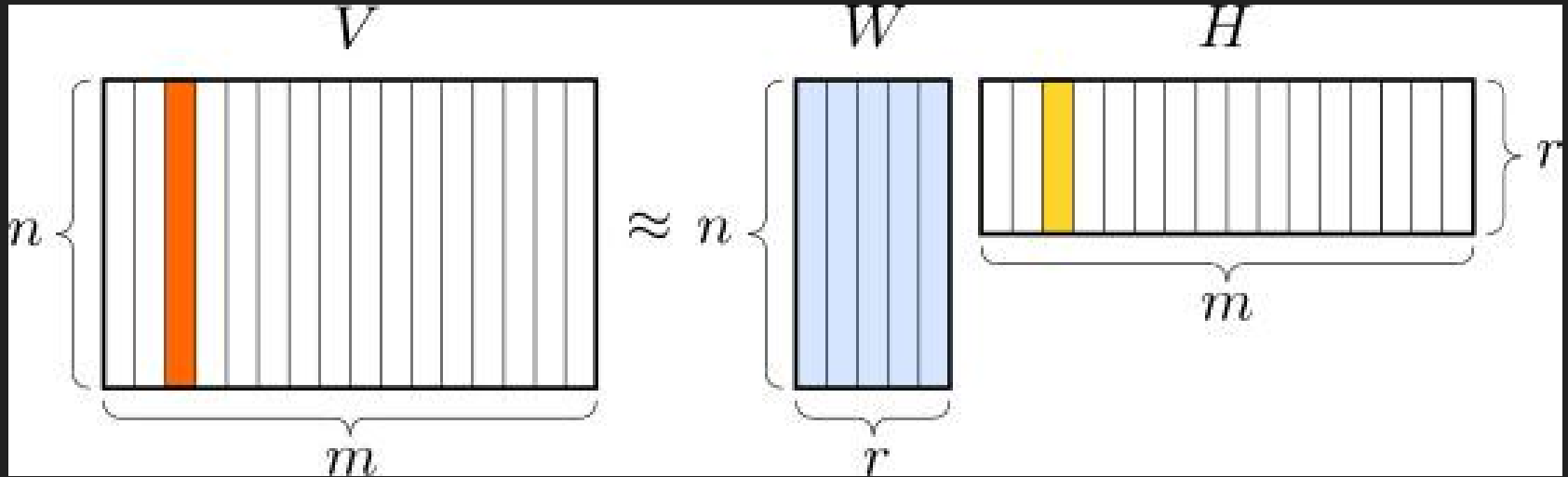


# NMF [05 | 13]

- Lee and Seung (1999), Nature
- Parts-based representation of faces

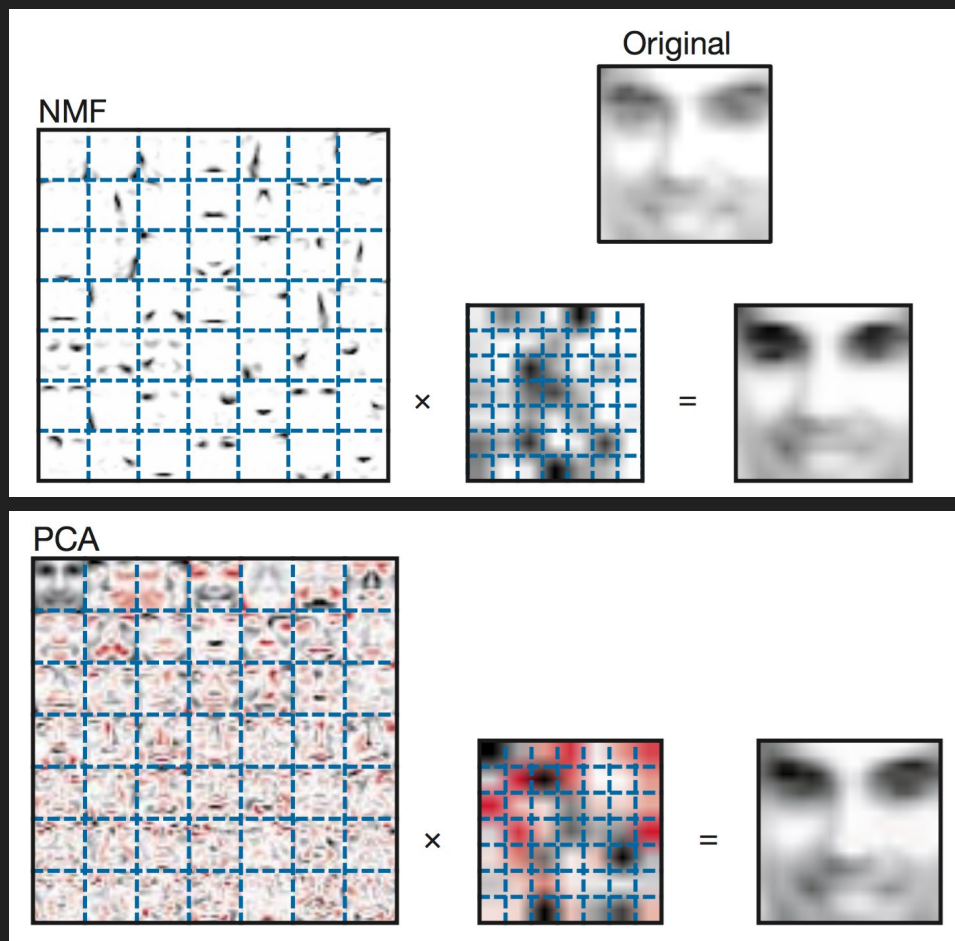


# NMF [06 | 13]



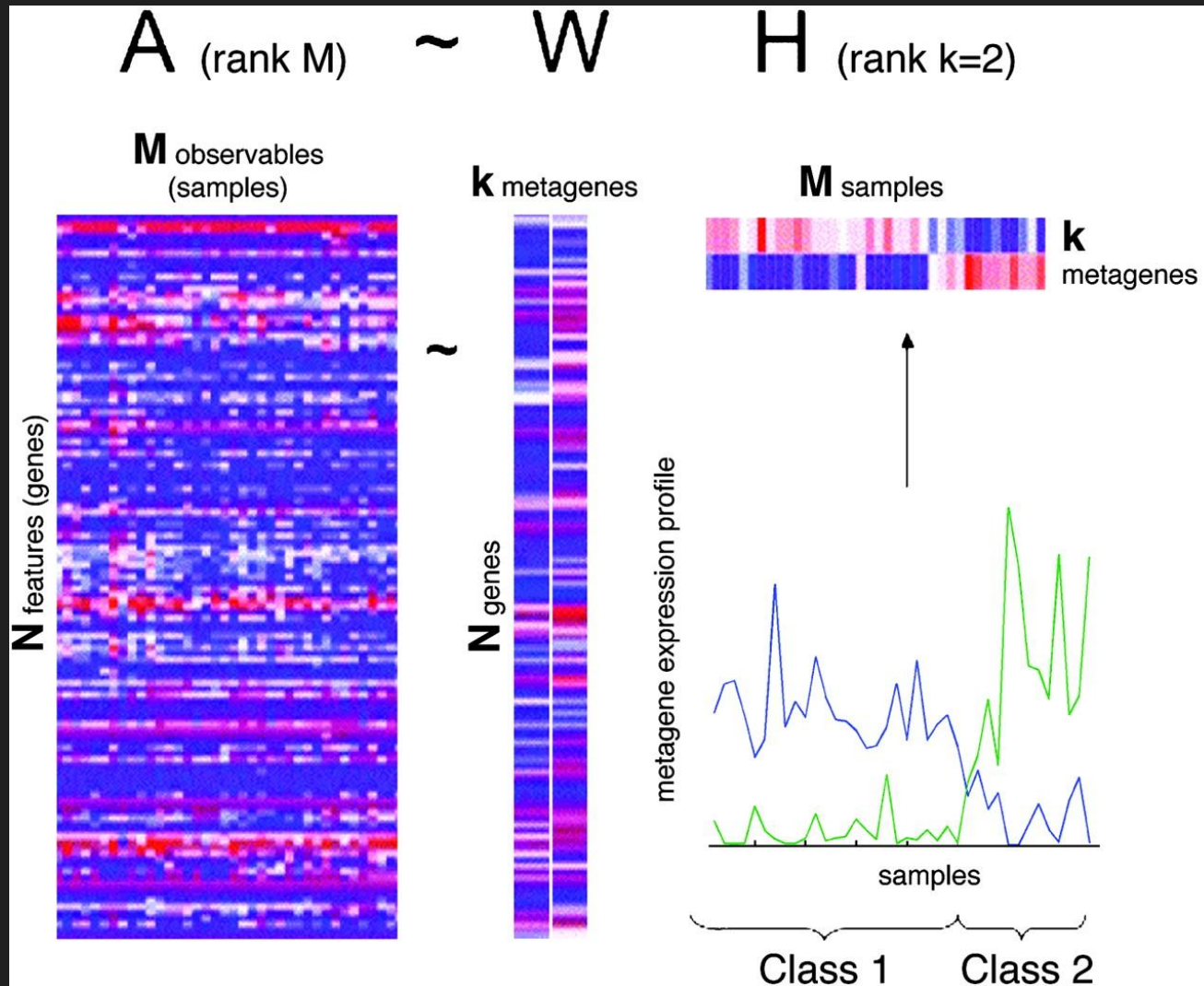
# NMF [07|13]

- NMF vs PCA





# NMF for gene expression [08 | 13]



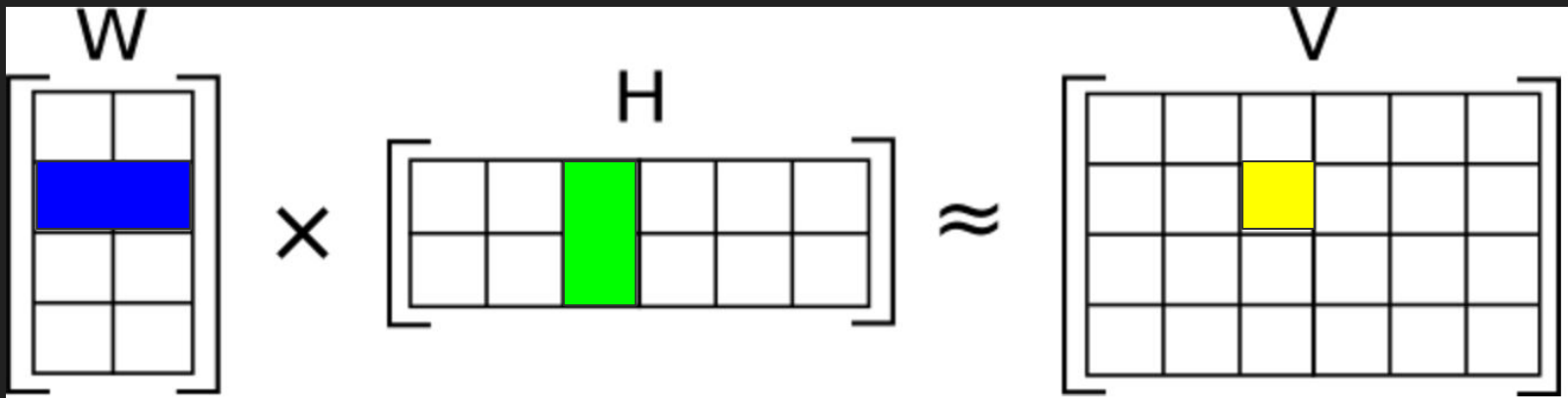
# NMF for gene expression [09 | 13]

- Dataset: EBI Expression Atlas
  - RNA-Seq of coding RNA
  - 122 individuals
  - 32 tissues
  - 17,995 genes
- Hardware: Amazon Web Services
  - g2.2xlarge
  - NVIDIA Kepler G104
  - Intel Xeon E5-2670

# NMF for gene expression [10 | 13]

- Gradient descent
- Error for each element in  $V$ :

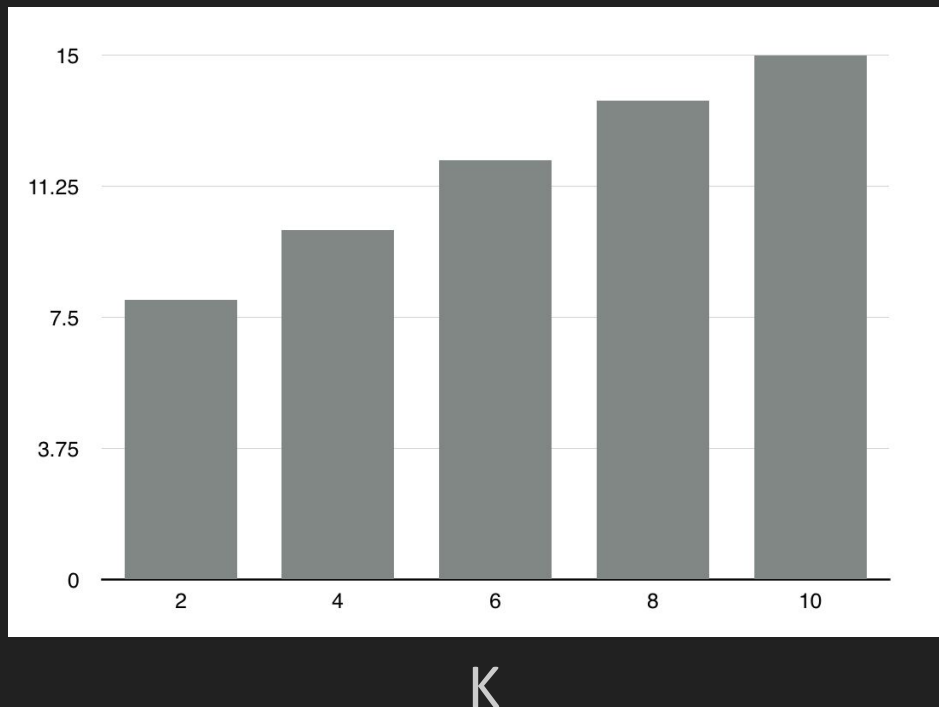
$$e_{ij}^2 = (v_{ij} - \hat{v}_{ij})^2 = (v_{ij} - \sum_{k=1}^K w_{ik}h_{kj})^2$$



$$w'_{ik} = w_{ik} + \alpha \frac{\partial}{\partial w_{ik}} e_{ij}^2 = w_{ik} + 2\alpha e_{ij} h_{kj}$$

# NMF for gene expression [11 | 13]

- GPU vs single CPU:
  - Difference in results - negligible
  - Speedup vs number of metagenes



# Conclusion [12 | 13]

- Positives:
  - Pathways and metagenes - domain knowledge
  - Quick
  - (At least) good results
- Neutrals:
  - To be seen whether it is the best
- Negatives:
  - Local minimum
  - Multiple CUDA-threads reading same data
  - GPU memory potential problem for large datasets

# References [13 | 13]

1. [Lee and Seung \(1999\), Nature](#)
2. [Devarajan \(2008\), PLoS Comp. Biology](#)
3. [Mejia-Roa et al \(2015\), BMC Bioinformatics](#)