

# Machine Learning

## Unsupervised Methods

### Part 1

Sepp Hochreiter

Institute of Bioinformatics  
Johannes Kepler University, Linz, Austria

---

# Course



3 ECTS 2 SWS VO (class)

1.5 ECTS 1 SWS UE (exercise)

Basic Course of Master Bioinformatics

Basic Course of Master Computer Science: Computational Engineering / Int. Syst.

Class: Mo 15:30-17:00 (HS 19)

Exercise: Mon 13:45-14:30 (MT 226) – group 2

Mon 14:30-15:15 (MT 226) – group 1+ group 3

EXAMS:

VO: 3 written intermediate exams

UE: weekly homework (evaluated)

# Other Courses



|         | Lecture   |    | Lecturer     |                        |
|---------|---|----|--------------|------------------------|
| 365,077 | <b>Machine Learning: Unsupervised Techniques</b>  | VL | Hochreiter   | Mon 15:30-17:00/HS 19  |
| 365,078 | <b>Machine Learning: Unsupervised Techniques – G2</b>   | UE | Kofler       | Mon 13:45-14:30/MT 226 |
| 365,095 | <b>Machine Learning: Unsupervised Techniques – G1+G3</b>  | UE | Brandstetter | Mon 14:30-15:15/MT 226 |
| 365,041 | <b>Theoretical Concepts of Machine Learning</b>   | VL | Nessler      | Thu 15:30-17:00/S2 048 |
| 365,042 | <b>Theoretical Concepts of Machine Learning</b>   | UE | Kofler       | Thu 14:30-15:15/S2 053 |
| 365,081 | <b>Genome Analysis &amp; Transcriptomics</b>  | KV | Regl         | Fri 8:30-11:00/S2 053  |
| 365,082 | <b>Structural Bioinformatics</b>  | KV | Regl         | Tue 8:30-11:00/S3 057  |
| 365,093 | <b>Deep Learning and Neural Networks</b>  | KV | Unterthiner  | Thu 10:15-11:45/MT 226 |
| 365,090 | <b>Special Topics on Bioinformatics (B/C/P/M):<br/>Population genetics</b>                      | KV | Klambauer    | Fri 9:15-11:00/S2 219  |
| 365,096 | <b>Special Topics on Bioinformatics (B/C/P/M):<br/>Artificial Intelligence in Life Sciences</b> | KV | Klambauer    | Fri 12:45-14:30/S2 046 |
| 365,101 | <b>Special Topics: Deep Reinforcement Learning</b>  | VL | Arjona       | Mon 12:45-13:30/K 224B |

- 1 Introduction
- 2 Basic Terms and Concepts
- 3 Principal Component Analysis
- 4 Independent Component Analysis
- 5 Factor Analysis
- 6 Scaling and Projection Methods
- 7 Clustering
- 8 Biclustering
- 9 Hidden Markov Models
- 10 Boltzmann Machines

## **1 Introduction**

1.1 Machine Learning Introduction

1.2 Course Specific Introduction

1.3 Generative vs. Descriptive Models

## **2 Basic Terms and Concepts**

2.1 Unsupervised Learning in Bioinformatics

2.2 Unsupervised Learning Categories

2.3 Quality of Parameter Estimation

2.4 Maximum Likelihood Estimator

2.5 Expectation Maximization

2.6 Maximum Entropy

## **3 Principal Component Analysis**

- 3.1 The Method
- 3.2 Variance Maximization
- 3.3 Uniqueness
- 3.4 Properties of PCA
- 3.5 Examples
- 3.6 Kernel Principal Component Analysis

## **4 Independent Component Analysis**

- 4.1 Identifiability and Uniqueness
- 4.2 Measuring Independence
- 4.3 Whitening and Rotation Algorithms
- 4.4 INFOMAX Algorithm
- 4.5 EASI Algorithm
- 4.6 FastICA Algorithm
- 4.7 ICA Extensions
- 4.8 ICA vs. PCA
- 4.9 Artificial ICA Examples
  - 4.9.1 Whitening and Rotation
- 4.10 Real World ICA Examples
- 4.11 Kurtosis Maximization Results in Independent Components

## **5 Factor Analysis**

- 5.1 The Factor Analysis Model
- 5.2 Maximum Likelihood Factor Analysis
- 5.3 Factor Analysis vs. PCA and ICA
- 5.4 Artificial Factor Analysis Examples
- 5.5 Real World Factor Analysis Examples

## **6 Scaling and Projection Methods**

- 6.1 Projection Pursuit
- 6.2 Multidimensional Scaling
- 6.3 Non-negative Matrix Factorization
- 6.4 Locally Linear Embedding
- 6.5 Isomap
- 6.6 The Generative Topographic Mapping
- 6.7 t-Distributed Stochastic Neighbor Embedding
- 6.8 Self-Organizing Maps

## **7 Clustering**

- 7.1 Mixture Models
- 7.2 k-Means Clustering
- 7.3 Hierarchical Clustering
- 7.4 Similarity-Based Clustering

## **8 Biclustering**

- 8.1 Types of Biclusters
- 8.2 Overview of Biclustering Methods
- 8.3 FABIA Biclustering
- 8.4 Examples

## **9 Hidden Markov Models**

- 9.1 Hidden Markov Models in Bioinformatics
- 9.2 Hidden Markov Model Basics
- 9.3 Expectation Maximization for HMM: Baum-Welch Algorithm
- 9.4 Viterby Algorithm
- 9.5 Input Output Hidden Markov Models
- 9.6 Factorial Hidden Markov Models
- 9.7 Memory Input Output Factorial Hidden Markov Models
- 9.8 Tricks of the Trade
- 9.9 Profile Hidden Markov Models

## **10 Boltzmann Machines**

- 10.1 The Boltzmann Machine
- 10.2 Learning in the Boltzmann Machine
- 10.3 The Restricted Boltzmann Machine



- ML**: Duda, Hart, Stork; Pattern Classification; Wiley & Sons, 2001
- ML**: C. M. Bishop; Neural Networks for Pattern Recognition, Oxford Univ. Press, 1995
- ML**: T. M. Mitchell; Machine Learning, Mc Graw Hill, 1997
- Statistics**: S. M. Kay; Fundamentals of Statistical Signal Processing, Prent. Hall, 1993
- Belief Nets**: M. I. Jordan; Learning in Graphical Models, MIT Press, 1998
- Data Analysis**: R. Peck, C. Olsen and J. L. Devore; Introduction to Statistics and Data Analysis, 3rd edition, ISBN: 9780495118732, Brooks/Cole, Belmont, USA, 2009
- Statistical Data Analysis**: B. Shahbaba; Biostatistics with R: An Introduction to Statistics Through Biological Data; Springer, series UseR!, New York, 2012
- Statistical Data Analysis**: C. T. Ekstrom and H. Sorensen; Introduction to Statistical Data Analysis for the Life Sciences; CRC Press, Taylor & Francis Group, USA, 2011
- Clustering**: L. Kaufman and P. J. Rousseeuw; Finding Groups in Data. An Introduction to Cluster Analysis, Wiley, 1990

# Chapter 1

## Introduction

## 1 Introduction

### 1.1 Machine Learning

#### Introduction

#### 1.2 Course Specific Introduction

#### 1.3 Generative vs. Descriptive Models

### 2 Basic Terms and Concepts

#### 2.1 Unsupervised Learning in Bioinformatics

#### 2.2 Unsupervised Learning Categories

#### 2.3 Quality of Parameter Estimation

#### 2.4 Maximum Likelihood Estimator

#### 2.5 Expectation Maximization

#### 2.6 Maximum Entropy

- part of curriculum “master of science in bioinformatics”
- part of curriculum “computer science” (major CE, major int. sys.)
- Machine learning major research topic: Google, Microsoft, Amazon, Facebook, AltaVista, Zalando, and many more
- Applications: computer vision (image recognition), speech recognition, recommender systems, analysis of Big Data, information retrieval
- Mining the web: search engines, social networks, videos, music
- Machine learning applications in biology and medicine:
  - microarrays, sequencing
  - alternative splicing, nucleosome positions, gene regulation
  - single nucleotide polymorphisms / variants (SNPs, SNVs)
  - copy number variations (CNVs)
  - diseases: Alzheimer, Parkinson, cancer, multiples sclerosis, schizophrenia or alcohol dependence

- 1 Introduction
- 1.1 Machine Learning Introduction
- 1.2 Course Specific Introduction
- 1.3 Generative vs. Descriptive Models
- 2 Basic Terms and Concepts
- 2.1 Unsupervised Learning in Bioinformatics
- 2.2 Unsupervised Learning Categories
- 2.3 Quality of Parameter Estimation
- 2.4 Maximum Likelihood Estimator
- 2.5 Expectation Maximization
- 2.6 Maximum Entropy

This course introduces **unsupervised** machine learning methods:

- output is not given
- objective: cumulative output on all samples

Objectives:

- information content
- orthogonal
- statistical independence
- variation explained
- entropy
- likelihood: probability that model produces observed data
- distances between and within clusters

Used for analyze data:

- explore
- find structure
- visualize
- compress

Understand and explore the data and generate new knowledge

- 1 Introduction
  - 1.1 Machine Learning Introduction
  - 1.2 Course Specific Introduction
  - 1.3 Generative vs. Descriptive Models
- 2 Basic Terms and Concepts
  - 2.1 Unsupervised Learning in Bioinformatics
  - 2.2 Unsupervised Learning Categories
  - 2.3 Quality of Parameter Estimation
  - 2.4 Maximum Likelihood Estimator
  - 2.5 Expectation Maximization
  - 2.6 Maximum Entropy

concepts of unsupervised learning:

- maximum likelihood
- maximum a posteriori
- maximum entropy
- expectation maximization
- maximal variance
- independence
- non-Gaussianity
- sub- and super-Gaussian distributions
- sparse and population codes

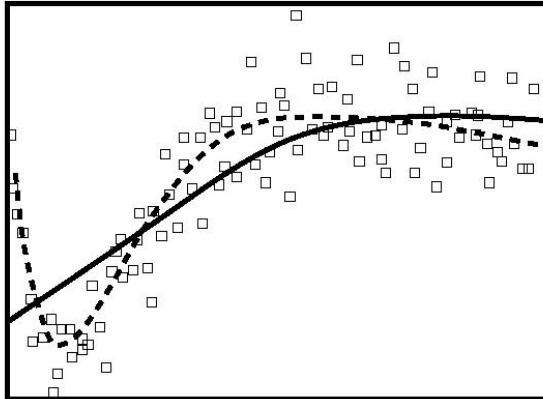
- 1 Introduction
  - 1.1 Machine Learning Introduction
  - 1.2 Course Specific Introduction
  - 1.3 Generative vs. Descriptive Models
- 2 Basic Terms and Concepts
  - 2.1 Unsupervised Learning in Bioinformatics
  - 2.2 Unsupervised Learning Categories
  - 2.3 Quality of Parameter Estimation
  - 2.4 Maximum Likelihood Estimator
  - 2.5 Expectation Maximization
  - 2.6 Maximum Entropy

- Goal: select model
  - with highest generalization performance, that is
  - with the best performance on future data,
  - from the model class
- **model selection** is **training** is **learning**
- model which best explains or approximates the training set
- “**overfitting**”: model is fitted (adapted) to special training characteristics
  - noisy measurements
  - outliers
  - labeling errors

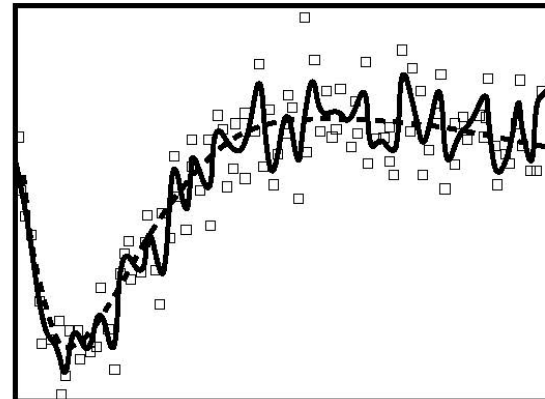
# Introduction

- 1 Introduction
- 1.1 Machine Learning Introduction
- 1.2 Course Specific Introduction
- 1.3 Generative vs. Descriptive Models
- 2 Basic Terms and Concepts
- 2.1 Unsupervised Learning in Bioinformatics
- 2.2 Unsupervised Learning Categories
- 2.3 Quality of Parameter Estimation
- 2.4 Maximum Likelihood Estimator
- 2.5 Expectation Maximization
- 2.6 Maximum Entropy

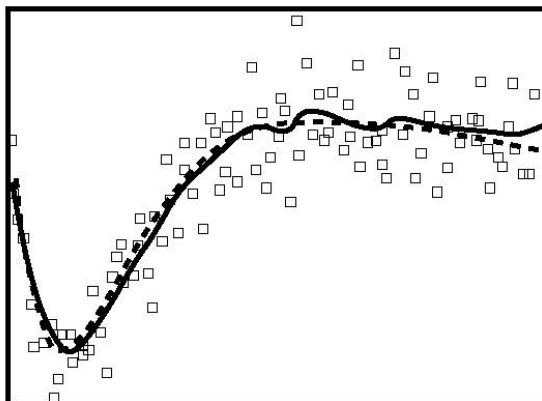
(a) large underfitting error



(b) large overfitting error



(c) best trade-off between over- and underfitting error



□ training examples (with noise)

---- target curve without noise

— approximated curve

- 1 Introduction
  - 1.1 Machine Learning Introduction
  - 1.2 Course Specific Introduction
  - 1.3 Generative vs. Descriptive Models
- 2 Basic Terms and Concepts
  - 2.1 Unsupervised Learning in Bioinformatics
  - 2.2 Unsupervised Learning Categories
  - 2.3 Quality of Parameter Estimation
  - 2.4 Maximum Likelihood Estimator
  - 2.5 Expectation Maximization
  - 2.6 Maximum Entropy

## Unsupervised Methods:

- principal component analysis
- independent component analysis
- factor analysis
- projection pursuit
- k-means clustering
- hierarchical clustering
- mixture models: Gaussian mixtures
- self-organizing maps
- kernel density estimation
- hidden Markov models
- Markov networks (Markov random fields)
- restricted Boltzmann machines
- neural network: auto-associators, unsupervised deep nets



- 1 Introduction
  - 1.1 Machine Learning Introduction
  - 1.2 Course Specific Introduction
  - 1.3 Generative vs. Descriptive Models
- 2 Basic Terms and Concepts
  - 2.1 Unsupervised Learning in Bioinformatics
  - 2.2 Unsupervised Learning Categories
  - 2.3 Quality of Parameter Estimation
  - 2.4 Maximum Likelihood Estimator
  - 2.5 Expectation Maximization
  - 2.6 Maximum Entropy

## Projection methods:

- new representation of objects
- down-projection into lower-dimensional space: keeps the neighborhoods
- finding structure in the data

## Generative models:

- build a model of the observed data
- match the observed data density

- 1 Introduction
- 1.1 Machine Learning Introduction
- 1.2 Course Specific Introduction
- 1.3 Generative vs. Descriptive Models
- 2 Basic Terms and Concepts
- 2.1 Unsupervised Learning in Bioinformatics
- 2.2 Unsupervised Learning Categories
- 2.3 Quality of Parameter Estimation
- 2.4 Maximum Likelihood Estimator
- 2.5 Expectation Maximization
- 2.6 Maximum Entropy

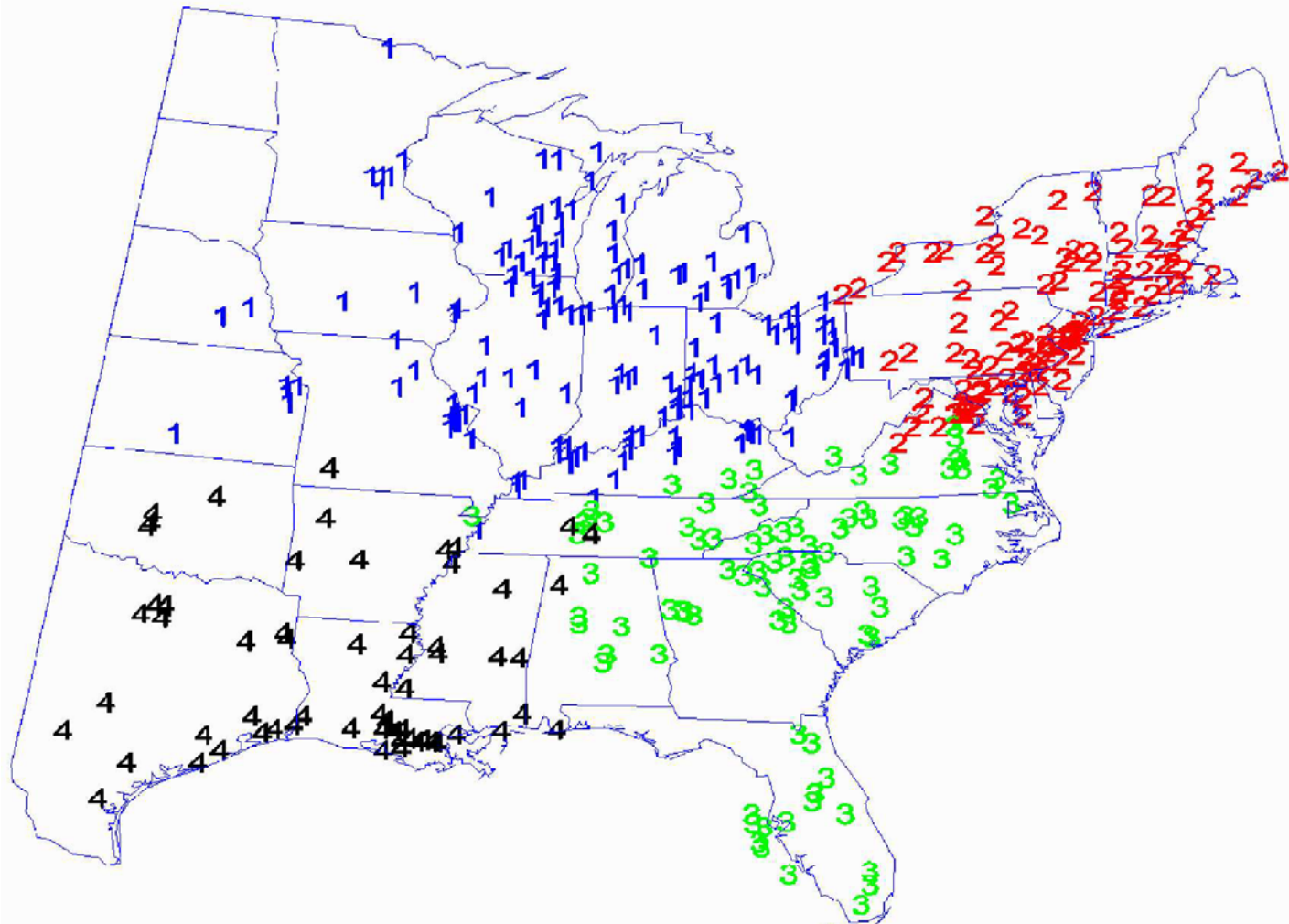
- projection: representation of objects, down-project feature vectors ,  
PCA: orthogonal maximal data variation components,  
ICA: statistically mutual independent components,  
factor analysis: PCA with noise
- density estimation: density model of observed data
- clustering: extract clusters – regions data accumulation (typical data)

## Goals of this course:

- how to chose appropriate methods from a given pool
- understand and evaluate the different approaches
- where to obtain and how to use them
- adapt and modify standard algorithms

# Introduction

FIGURE 1 OZONE 8-HR AVGS (4 CLUSTERS)  
1991-1995



- 1 Introduction
- 1.1 Machine Learning Introduction
- 1.2 Course Specific Introduction
- 1.3 Generative vs. Descriptive Models
- 2 Basic Terms and Concepts
- 2.1 Unsupervised Learning in Bioinformatics
- 2.2 Unsupervised Learning Categories
- 2.3 Quality of Parameter Estimation
- 2.4 Maximum Likelihood Estimator
- 2.5 Expectation Maximization
- 2.6 Maximum Entropy

# Introduction

## 1 Introduction

### 1.1 Machine Learning

#### Introduction

### 1.2 Course Specific Introduction

### 1.3 Generative vs. Descriptive Models

## 2 Basic Terms and Concepts

### 2.1 Unsupervised Learning in Bioinformatics

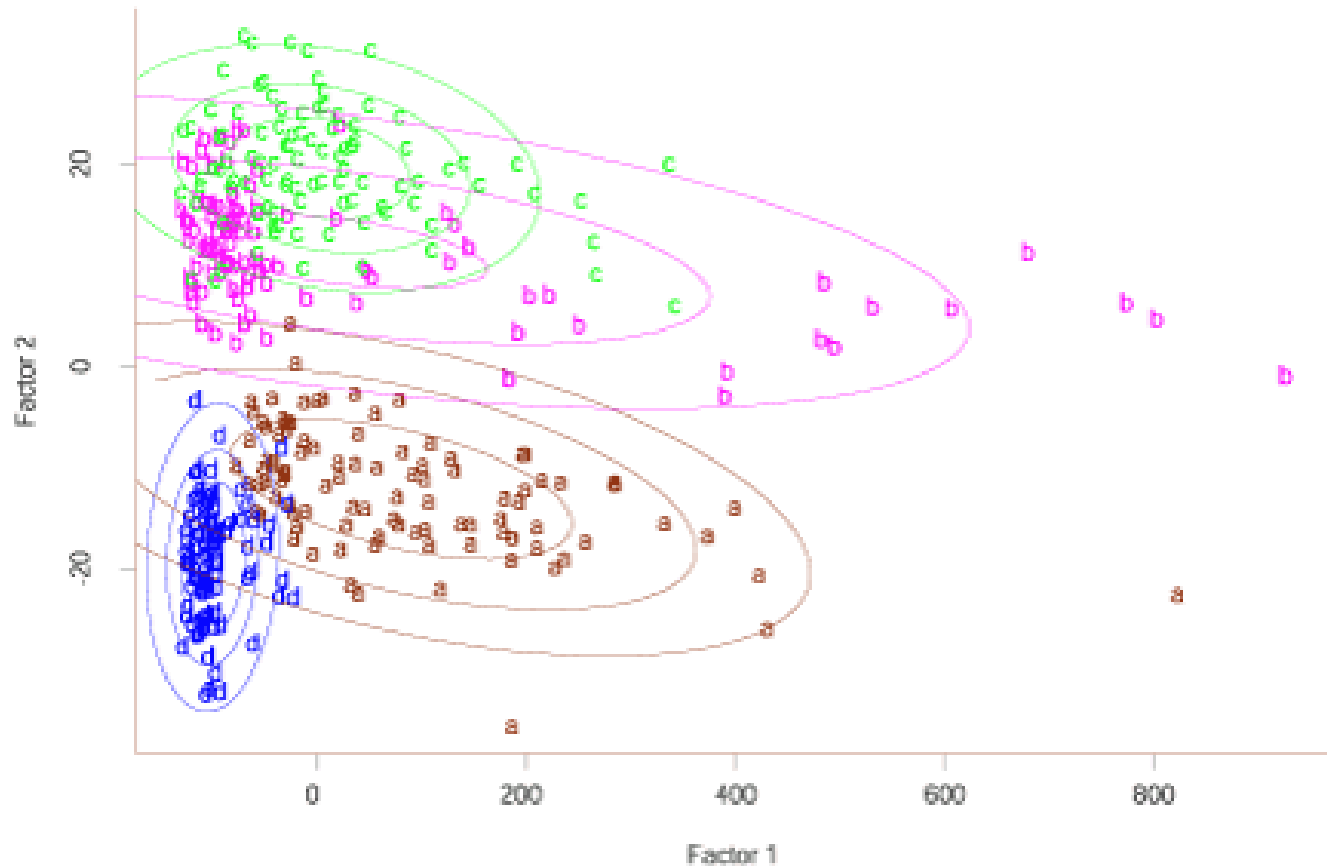
### 2.2 Unsupervised Learning Categories

### 2.3 Quality of Parameter Estimation

### 2.4 Maximum Likelihood Estimator

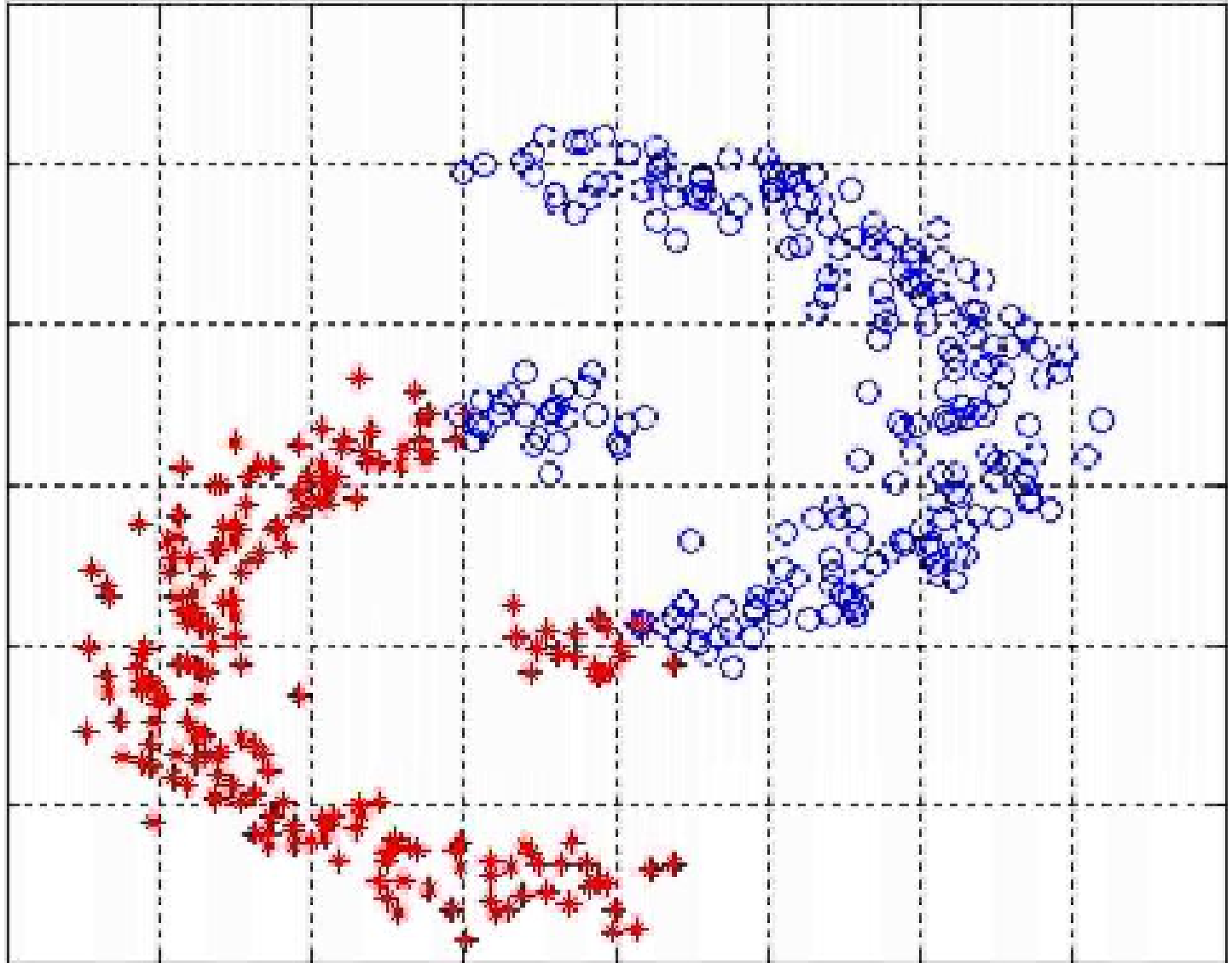
### 2.5 Expectation Maximization

### 2.6 Maximum Entropy



# Introduction

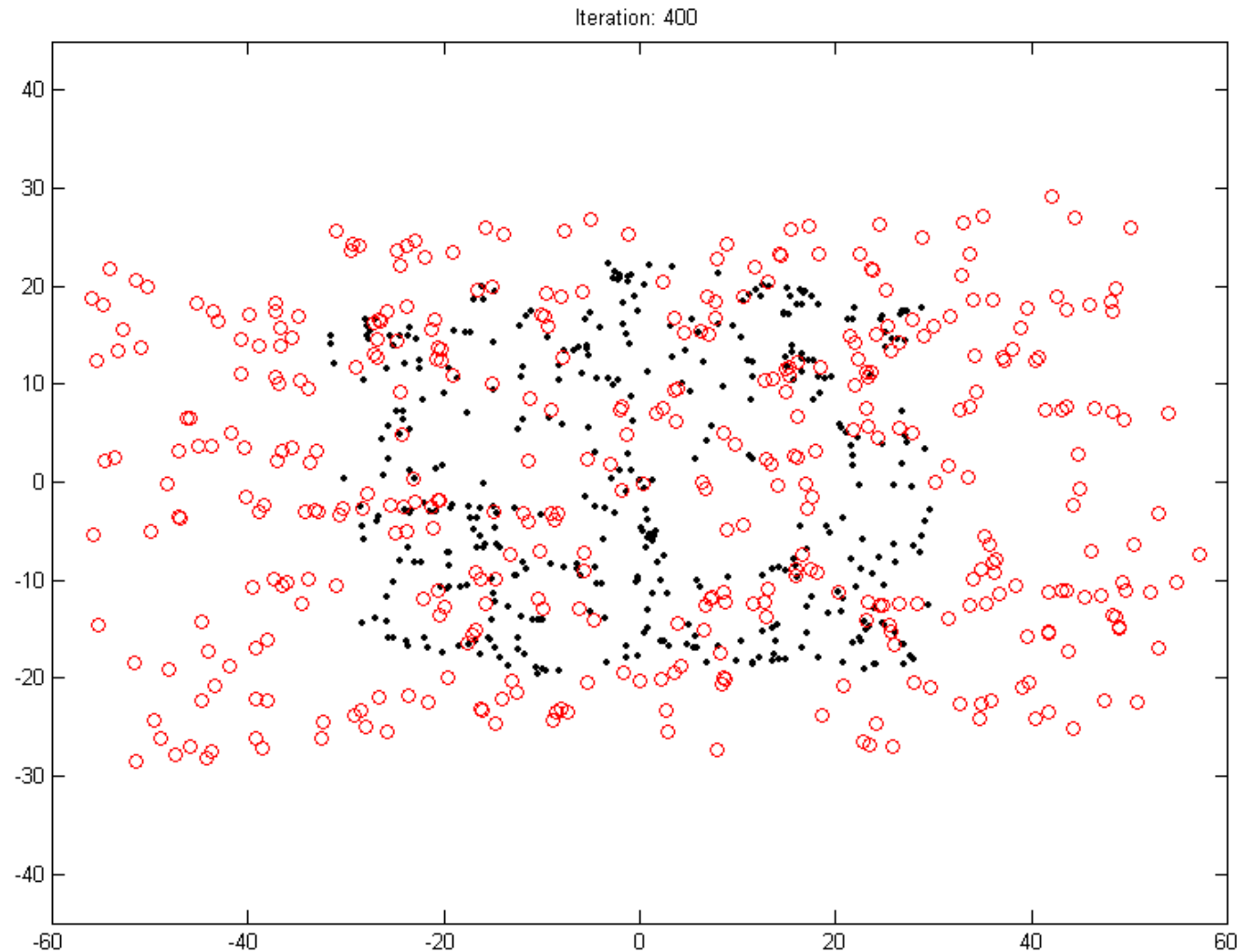
- 1 Introduction
  - 1.1 Machine Learning
    - Introduction
    - 1.2 Course Specific Introduction
    - 1.3 Generative vs. Descriptive Models
  - 2 Basic Terms and Concepts
    - 2.1 Unsupervised Learning in Bioinformatics
    - 2.2 Unsupervised Learning Categories
    - 2.3 Quality of Parameter Estimation
    - 2.4 Maximum Likelihood Estimator
    - 2.5 Expectation Maximization
    - 2.6 Maximum Entropy



# Introduction

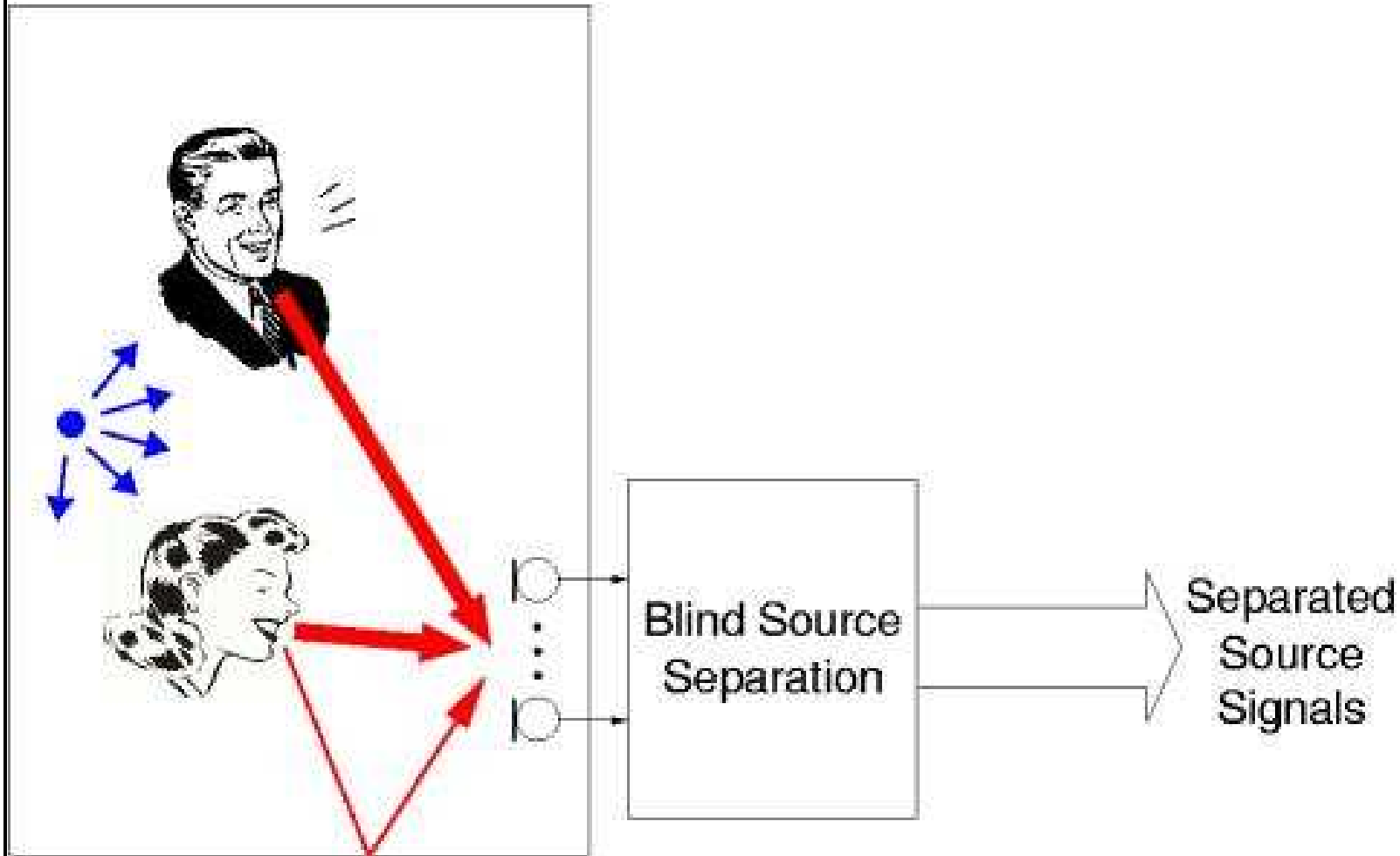
- 1 Introduction
- 1.1 Machine Learning Introduction
- 1.2 Course Specific Introduction
- 1.3 Generative vs. Descriptive Models
- 2 Basic Terms and Concepts
- 2.1 Unsupervised Learning in Bioinformatics
- 2.2 Unsupervised Learning Categories
- 2.3 Quality of Parameter Estimation
- 2.4 Maximum Likelihood Estimator
- 2.5 Expectation Maximization
- 2.6 Maximum Entropy

## Isomap: method for down-projecting data



# Introduction

- 1 Introduction
  - 1.1 Machine Learning Introduction
  - 1.2 Course Specific Introduction
  - 1.3 Generative vs. Descriptive Models
- 2 Basic Terms and Concepts
  - 2.1 Unsupervised Learning in Bioinformatics
  - 2.2 Unsupervised Learning Categories
  - 2.3 Quality of Parameter Estimation
  - 2.4 Maximum Likelihood Estimator
  - 2.5 Expectation Maximization
  - 2.6 Maximum Entropy



# Introduction

- 1 Introduction
  - 1.1 Machine Learning Introduction
  - 1.2 Course Specific Introduction
  - 1.3 Generative vs. Descriptive Models
- 2 Basic Terms and Concepts
  - 2.1 Unsupervised Learning in Bioinformatics
  - 2.2 Unsupervised Learning Categories
  - 2.3 Quality of Parameter Estimation
  - 2.4 Maximum Likelihood Estimator
  - 2.5 Expectation Maximization
  - 2.6 Maximum Entropy

Original:



Mixtures:



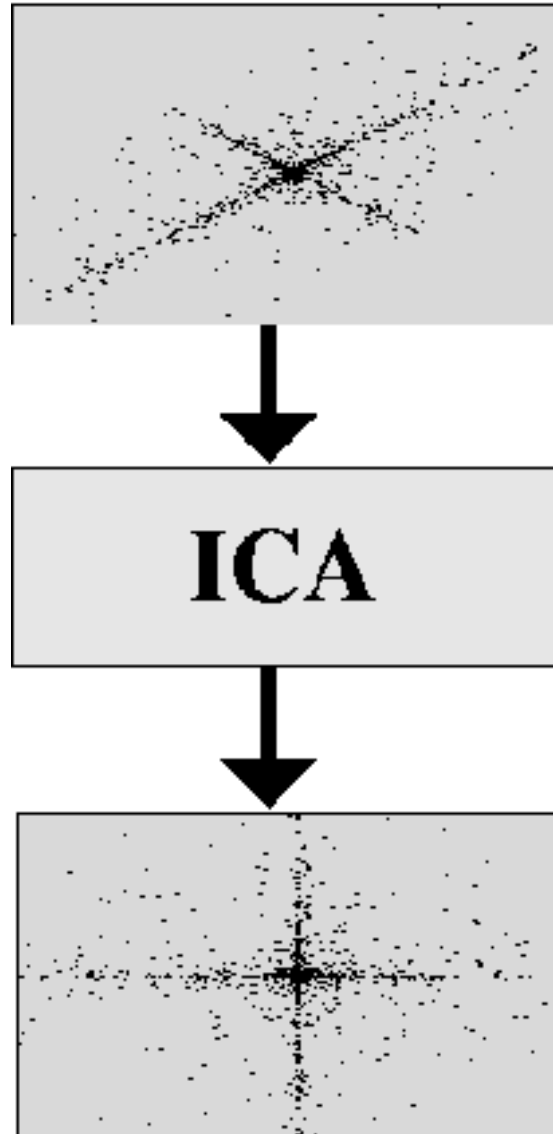
Demixed by ICA:





# Introduction

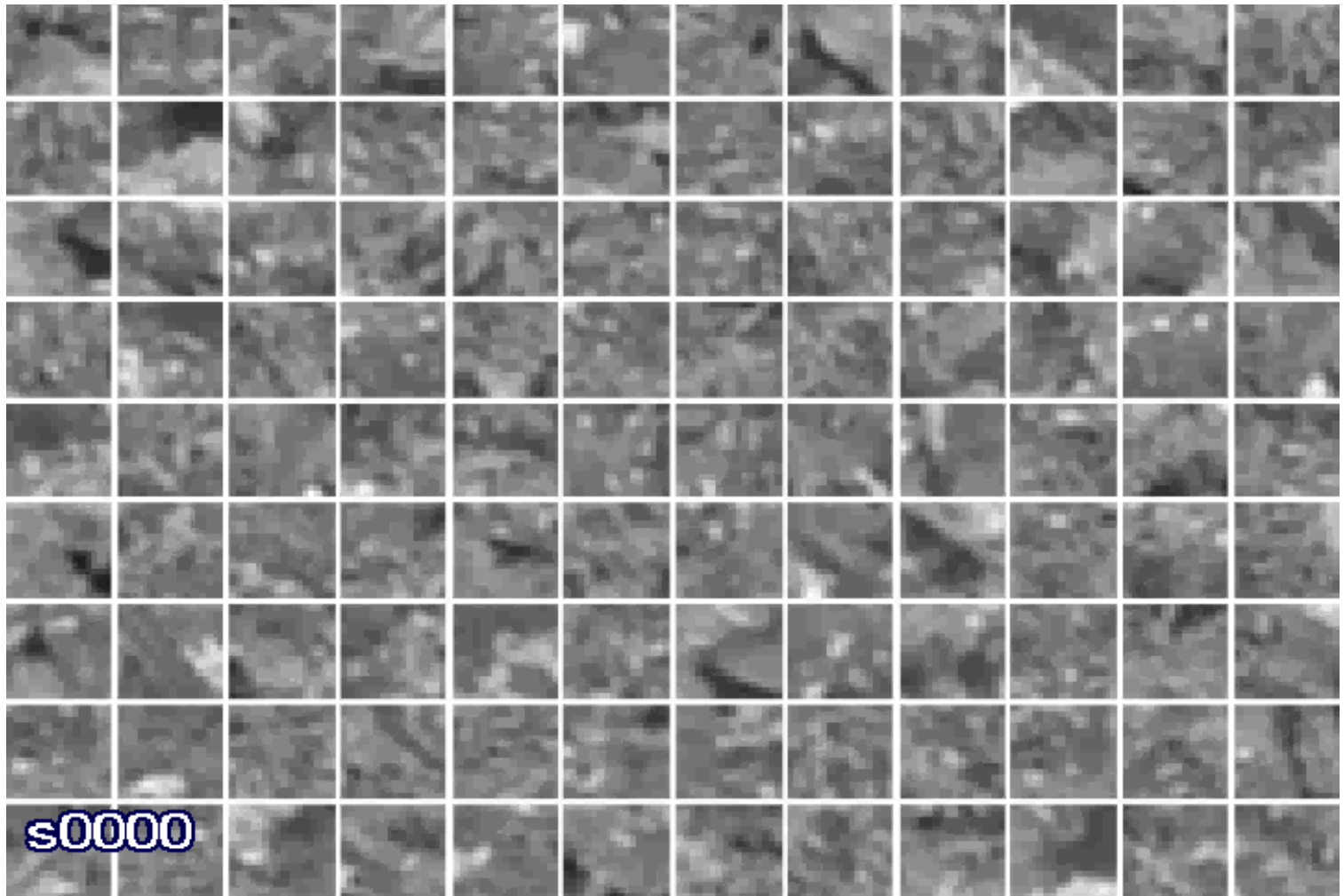
- 1 Introduction
  - 1.1 Machine Learning Introduction
  - 1.2 Course Specific Introduction
  - 1.3 Generative vs. Descriptive Models
- 2 Basic Terms and Concepts
  - 2.1 Unsupervised Learning in Bioinformatics
  - 2.2 Unsupervised Learning Categories
  - 2.3 Quality of Parameter Estimation
  - 2.4 Maximum Likelihood Estimator
  - 2.5 Expectation Maximization
  - 2.6 Maximum Entropy



# Introduction

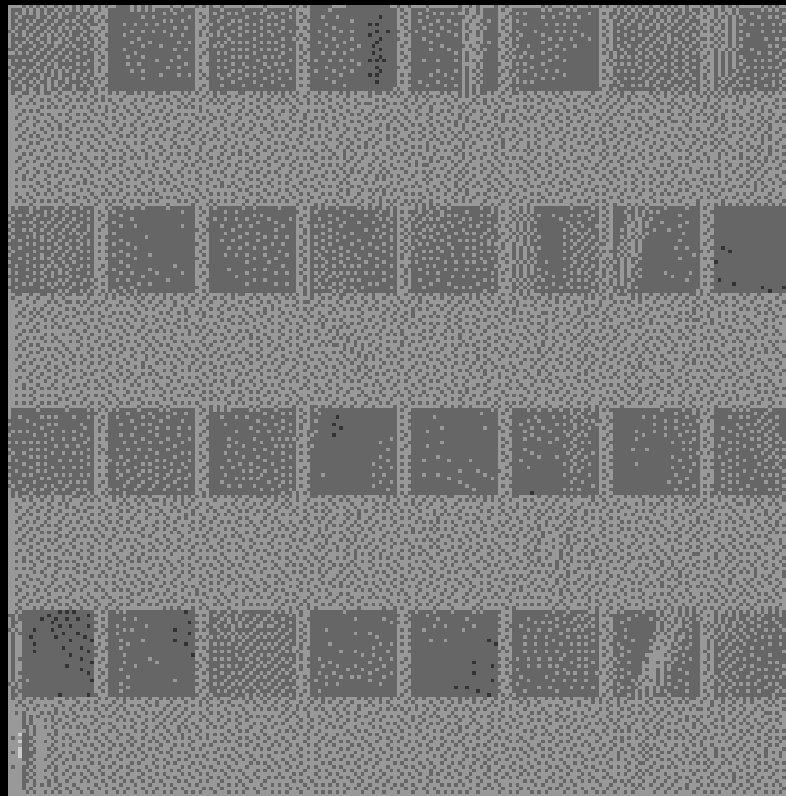
- 1 Introduction
  - 1.1 Machine Learning Introduction
  - 1.2 Course Specific Introduction
  - 1.3 Generative vs. Descriptive Models
- 2 Basic Terms and Concepts
  - 2.1 Unsupervised Learning in Bioinformatics
  - 2.2 Unsupervised Learning Categories
  - 2.3 Quality of Parameter Estimation
  - 2.4 Maximum Likelihood Estimator
  - 2.5 Expectation Maximization
  - 2.6 Maximum Entropy

## ICA: on images



- 1 Introduction
  - 1.1 Machine Learning Introduction
  - 1.2 Course Specific Introduction
  - 1.3 Generative vs. Descriptive Models
- 2 Basic Terms and Concepts
  - 2.1 Unsupervised Learning in Bioinformatics
  - 2.2 Unsupervised Learning Categories
  - 2.3 Quality of Parameter Estimation
  - 2.4 Maximum Likelihood Estimator
  - 2.5 Expectation Maximization
  - 2.6 Maximum Entropy

## ICA: on video components



# Parametric vs. Non-Parametric Models

- 1 Introduction
  - 1.1 Machine Learning Introduction
  - 1.2 Course Specific Introduction
  - 1.3 Generative vs. Descriptive Models
- 2 Basic Terms and Concepts
  - 2.1 Unsupervised Learning in Bioinformatics
  - 2.2 Unsupervised Learning Categories
  - 2.3 Quality of Parameter Estimation
  - 2.4 Maximum Likelihood Estimator
  - 2.5 Expectation Maximization
  - 2.6 Maximum Entropy

important step in machine learning is to select a **model class**

## parametric models:

- each parameter vector represents a model
- examples:
  - neural networks (synaptic weights) or SVMs (vector  $w$ )
  - factor analysis
- learning: paths through the parameter space
- disadvantages:
  - different parameterizations of the same function
  - model complexity and class via the parameters

## nonparametric models:

- model is locally constant / superimpositions
- Examples:
  - $k$ -nearest-neighbor ( $k$  is hyperparameter – not adjusted)
  - kernel density estimation
  - decision tree
- constant models (rules) must be a priori selected that is hyperparameters must be fixed ( $k$ , kernel width, splitting rules)

# Generative vs. descriptive Models



- 1 Introduction
- 1.1 Machine Learning Introduction
- 1.2 Course Specific Introduction
- 1.3 Generative vs. Descriptive Models
- 2 Basic Terms and Concepts
- 2.1 Unsupervised Learning in Bioinformatics
- 2.2 Unsupervised Learning Categories
- 2.3 Quality of Parameter Estimation
- 2.4 Maximum Likelihood Estimator
- 2.5 Expectation Maximization
- 2.6 Maximum Entropy

## descriptive model:

- additional description or another representation of the data
- projection methods (PCA, ICA)

## generative model:

- model should produce the distribution observed for the real world data points
- describing or representing random components which drive the process
- prior knowledge about the world or desired model
- predict new states of the data generation process (brain, cell)

# **Chapter 2**

## **Basic Terms and Concepts**

---

# Unsupervised Learning in Bioinformatics

BIOINF

## 1 Introduction

### 1.1 Machine Learning

#### Introduction

### 1.2 Course Specific Introduction

#### 1.3 Generative vs. Descriptive Models

## 2 Basic Terms and Concepts

### 2.1 Unsupervised Learning in Bioinformatics

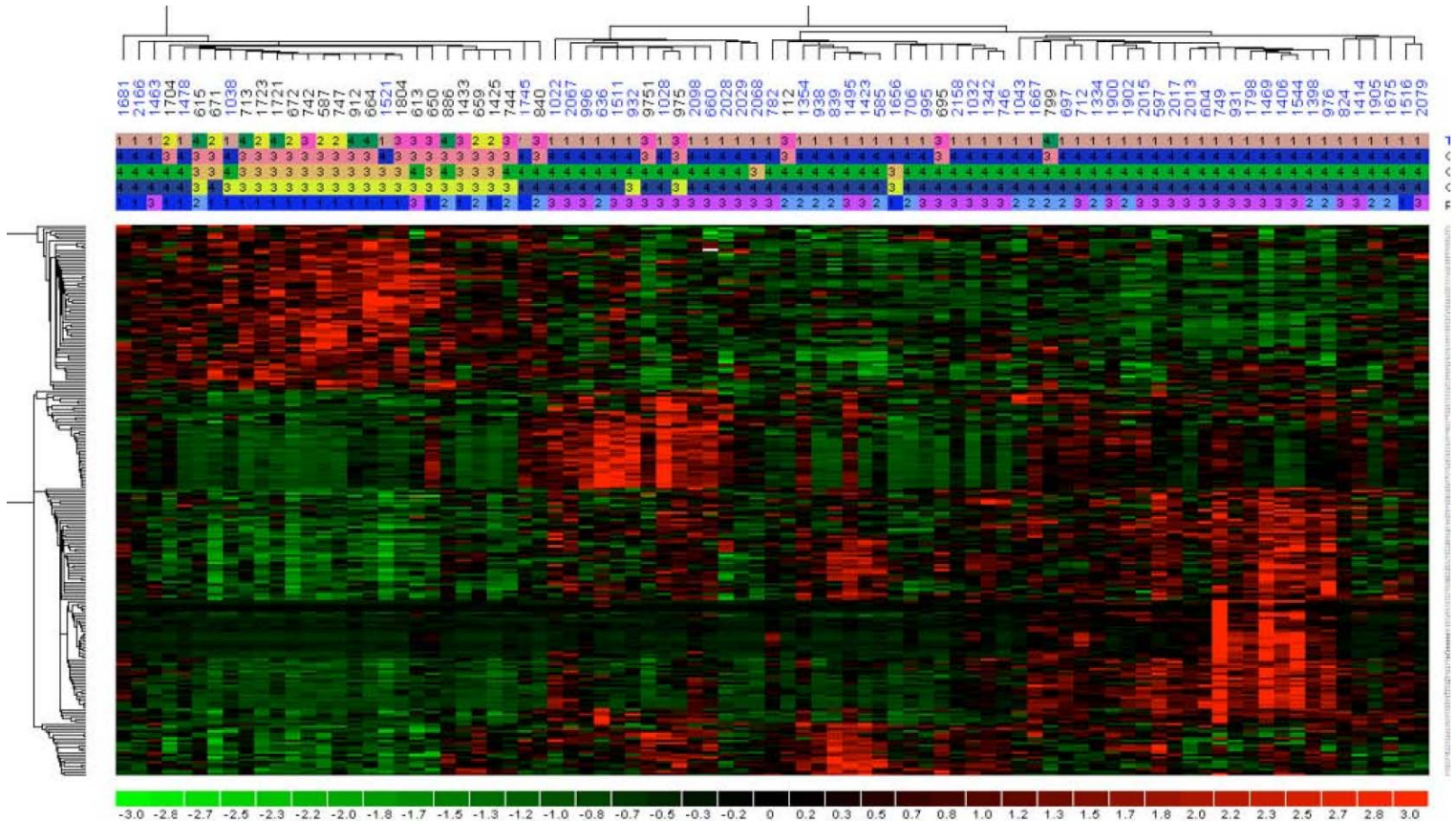
### 2.2 Unsupervised Learning Categories

#### 2.3 Quality of Parameter Estimation

#### 2.4 Maximum Likelihood Estimator

#### 2.5 Expectation Maximization

#### 2.6 Maximum Entropy

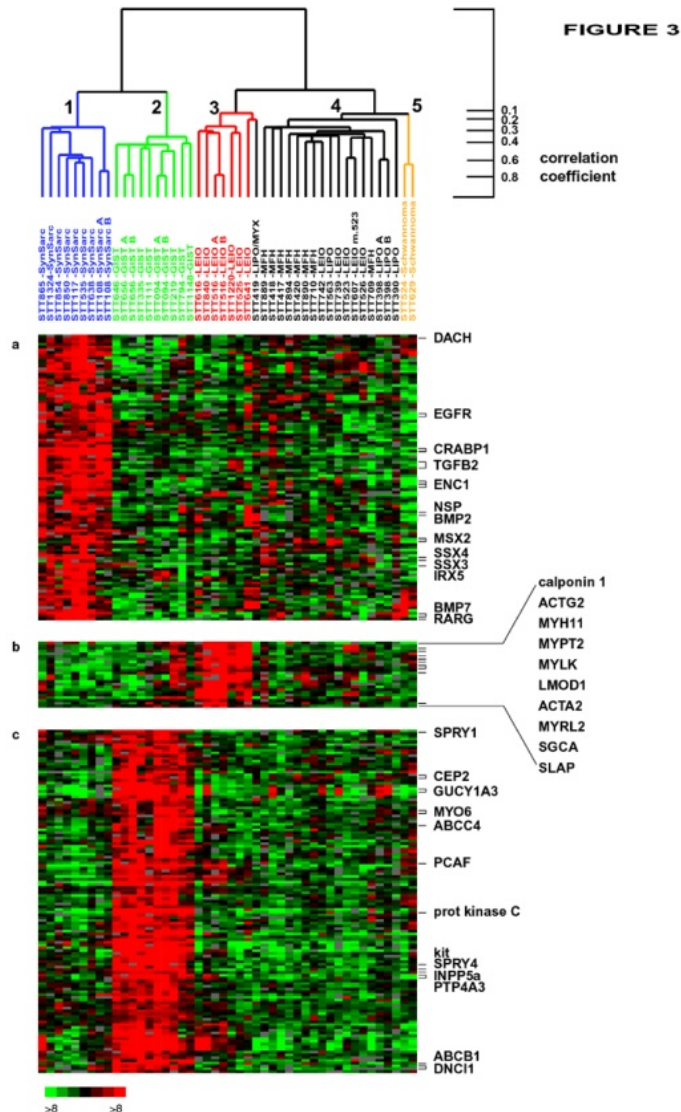


Clustering of microarray data



# Unsupervised Learning in Bioinformatics

- 1 Introduction
  - 1.1 Machine Learning Introduction
  - 1.2 Course Specific Introduction
  - 1.3 Generative vs. Descriptive Models
- 2 Basic Terms and Concepts
  - 2.1 Unsupervised Learning in Bioinformatics
  - 2.2 Unsupervised Learning Categories
  - 2.3 Quality of Parameter Estimation
  - 2.4 Maximum Likelihood Estimator
  - 2.5 Expectation Maximization
  - 2.6 Maximum Entropy



Clustering of microarray data.

Representative portions of the tumor specific gene clusters. The spectrum of green to red spots represents the relative centered expression for each gene.

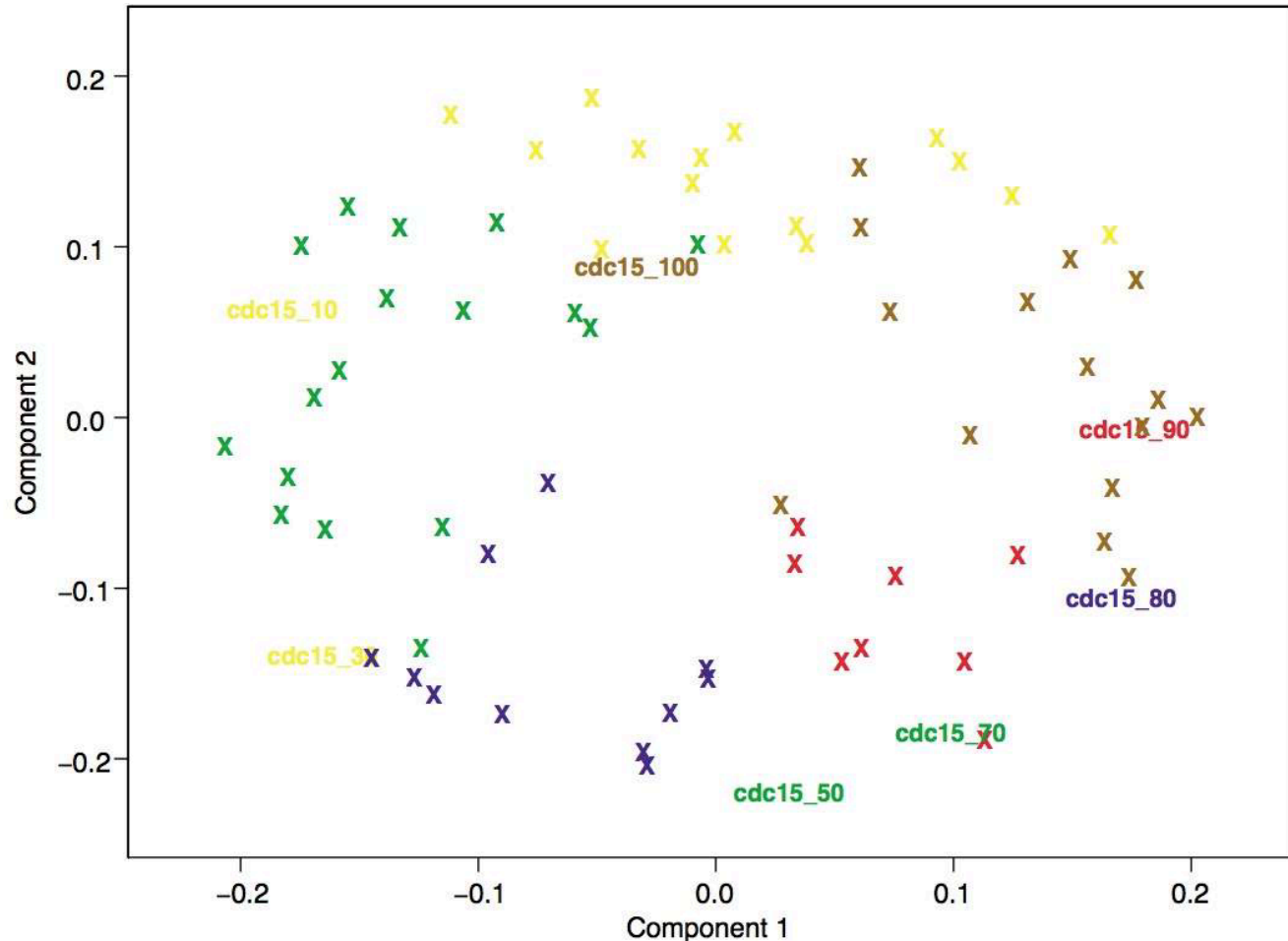
Correlation coefficient bar shown to the right side of the dendrogram indicates the degree of relatedness between branches of the dendrogram.



# Unsupervised Learning in Bioinformatics

- 1 Introduction
  - 1.1 Machine Learning Introduction
  - 1.2 Course Specific Introduction
  - 1.3 Generative vs. Descriptive Models
- 2 Basic Terms and Concepts
  - 2.1 Unsupervised Learning in Bioinformatics
  - 2.2 Unsupervised Learning Categories
  - 2.3 Quality of Parameter Estimation
  - 2.4 Maximum Likelihood Estimator
  - 2.5 Expectation Maximization
  - 2.6 Maximum Entropy

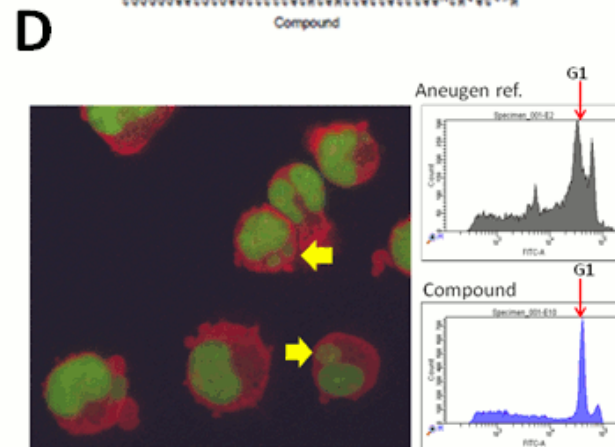
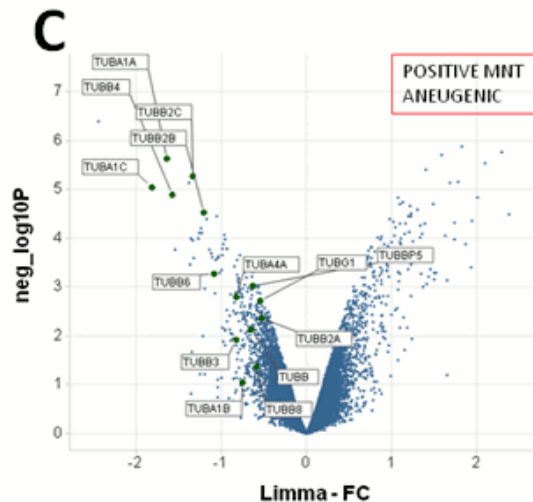
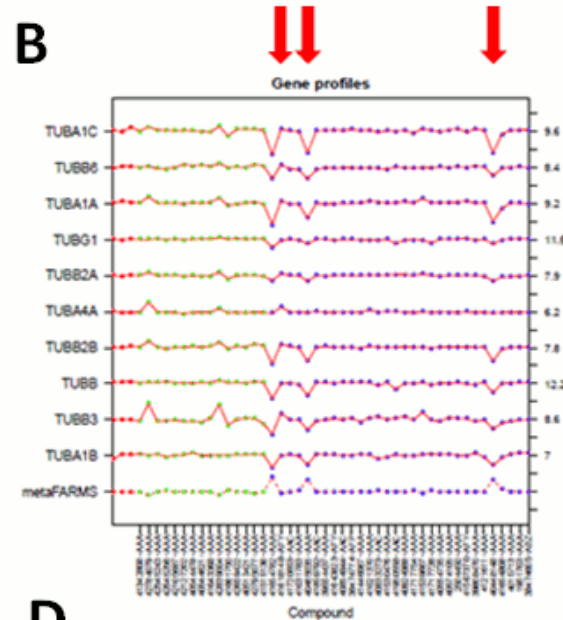
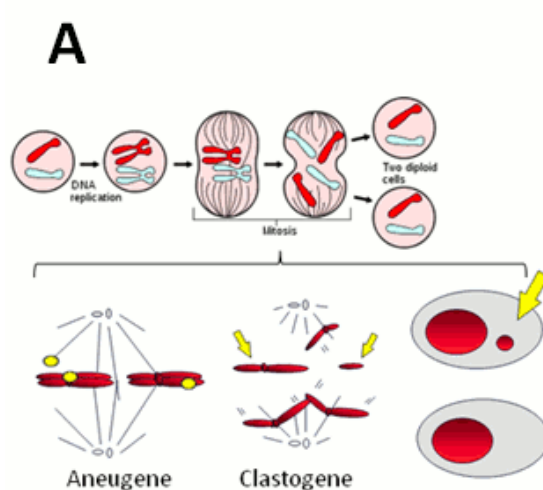
unsupervised methods: visualize dependencies and clusters



Spellmans cell-cycle data: first principal components

# Unsupervised Learning in Bioinformatics

- 1 Introduction
- 1.1 Machine Learning Introduction
- 1.2 Course Specific Introduction
- 1.3 Generative vs. Descriptive Models
- 2 Basic Terms and Concepts
- 2.1 Unsupervised Learning in Bioinformatics
- 2.2 Unsupervised Learning Categories
- 2.3 Quality of Parameter Estimation
- 2.4 Maximum Likelihood Estimator
- 2.5 Expectation Maximization
- 2.6 Maximum Entropy



An example for a signature of a rare event (micronuclei formation).

**A.** Genotoxic compounds can cause chromosomal breaks (aneugene) or affect the formation of the mitotic spindle or microtubuli (clastogene).

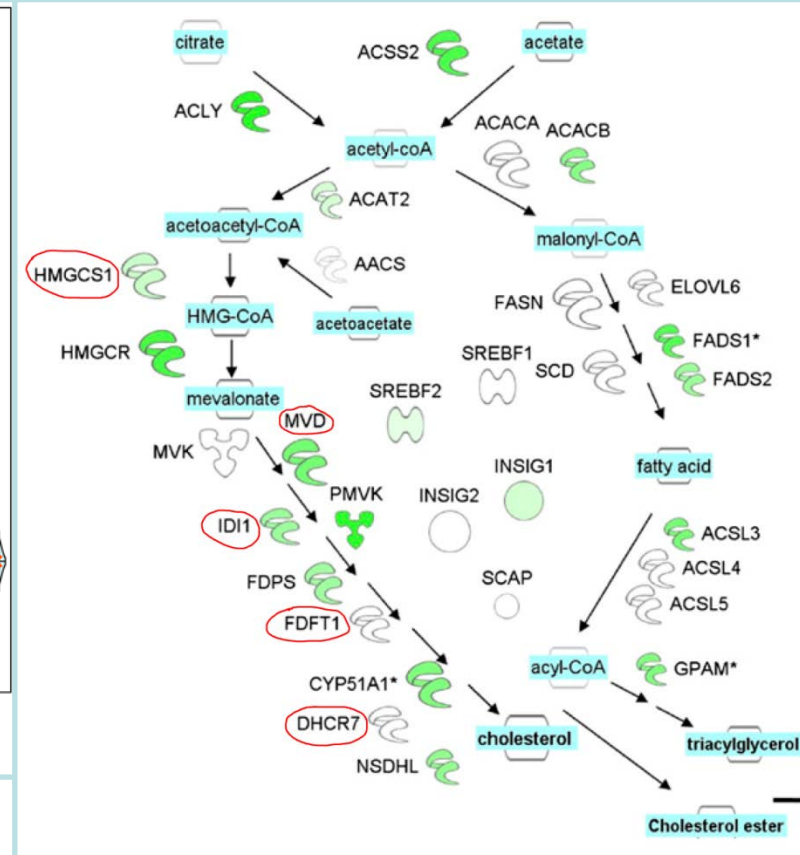
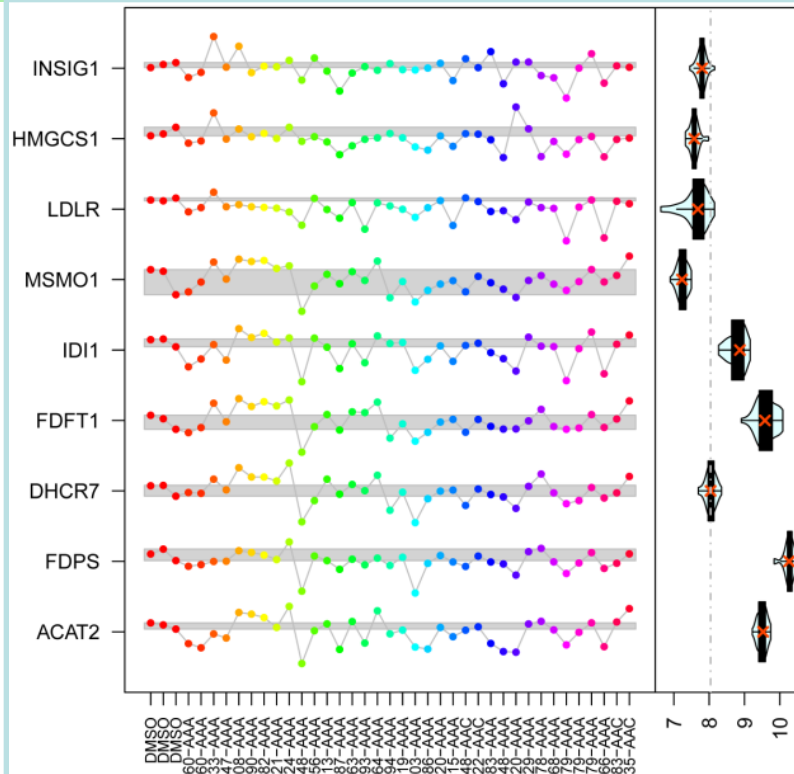
**B.** The gene expression signature of only three compounds (red arrows) show down-regulation of several tubulin-genes.

**C.** Volcano plot of one compound showing a down-regulation of tubulin genes.

**D.** Microscopic and FACS analysis confirmed micronuclei formation (yellow arrows) and G1-cell cycle arrest indicating microtubuli-based chromosome segregation.

# Unsupervised Learning in Bioinformatics

- 1 Introduction
- 1.1 Machine Learning Introduction
- 1.2 Course Specific Introduction
- 1.3 Generative vs. Descriptive Models
- 2 Basic Terms and Concepts
- 2.1 Unsupervised Learning in Bioinformatics
- 2.2 Unsupervised Learning Categories
- 2.3 Quality of Parameter Estimation
- 2.4 Maximum Likelihood Estimator
- 2.5 Expectation Maximization
- 2.6 Maximum Entropy

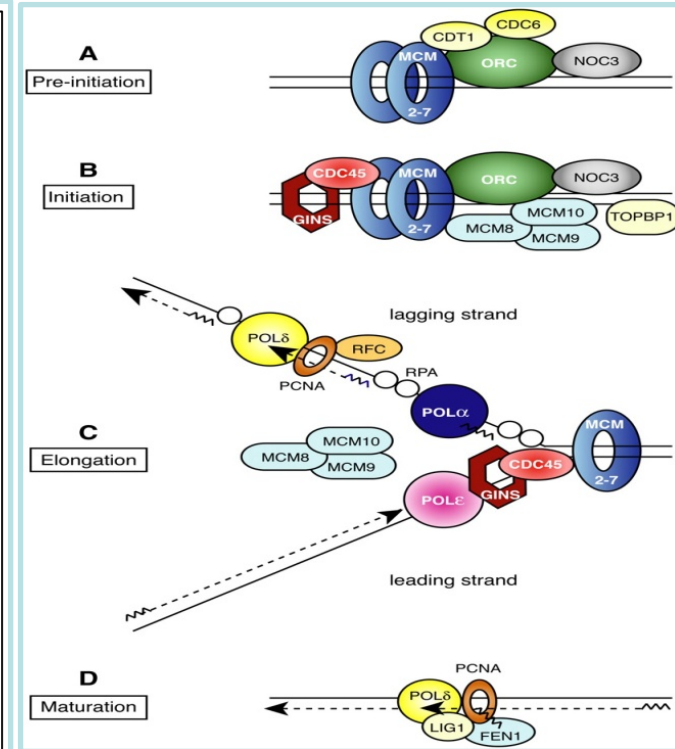
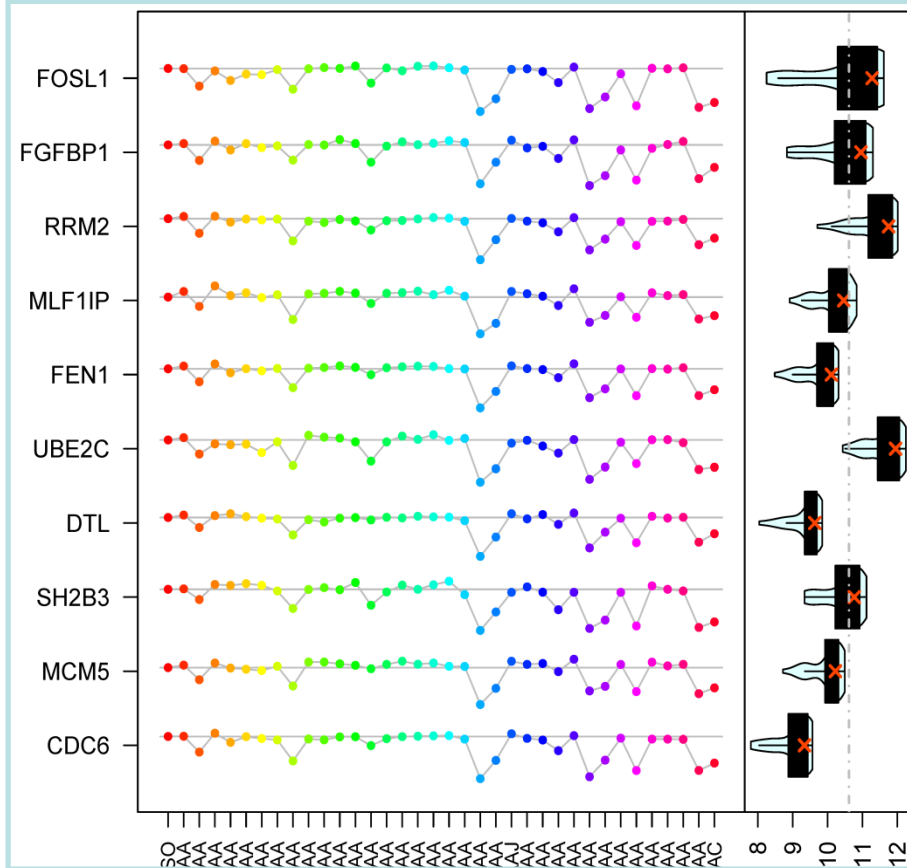


**left panel:** Biclustering results of gene expression data from a cell line where a compound was added that affects metabolic pathways. **right panel:** The genes HMGCS1, IDI1, FDFT1, DHCR7 of the bicluster code for proteins that belong to the SREBP cholesterol metabolism pathway. FABIA was capable to identify this bicluster of 9 genes activated by few compounds in a data set of tens of thousands of genes.

# Unsupervised Learning in Bioinformatics

BIOINF

- 1 Introduction
- 1.1 Machine Learning
- 1.2 Course Specific Introduction
- 1.3 Generative vs. Descriptive Models
- 2 Basic Terms and Concepts
- 2.1 Unsupervised Learning in Bioinformatics
- 2.2 Unsupervised Learning Categories
- 2.3 Quality of Parameter Estimation
- 2.4 Maximum Likelihood Estimator
- 2.5 Expectation Maximization
- 2.6 Maximum Entropy



**left panel:** Biclustering result of gene expression of a cancer cell line to which a compound has been added. **right panel:** The genes CDC6, MCM5, FEN1 are coding for proteins that participate at DNA replication complex. The other bicluster genes code for proteins that initiate or are involved DNA replication (MLF1IP  $\rightarrow$  chromosome segregation; RRM2  $\rightarrow$  DNA synthesis; DTL  $\rightarrow$  regulation of DNA replication).

# Unsupervised Learning Categories



- 1 Introduction
  - 1.1 Machine Learning Introduction
  - 1.2 Course Specific Introduction
  - 1.3 Generative vs. Descriptive Models
- 2 Basic Terms and Concepts
  - 2.1 Unsupervised Learning in Bioinformatics
  - 2.2 Unsupervised Learning Categories
  - 2.3 Quality of Parameter Estimation
  - 2.4 Maximum Likelihood Estimator
  - 2.5 Expectation Maximization
  - 2.6 Maximum Entropy

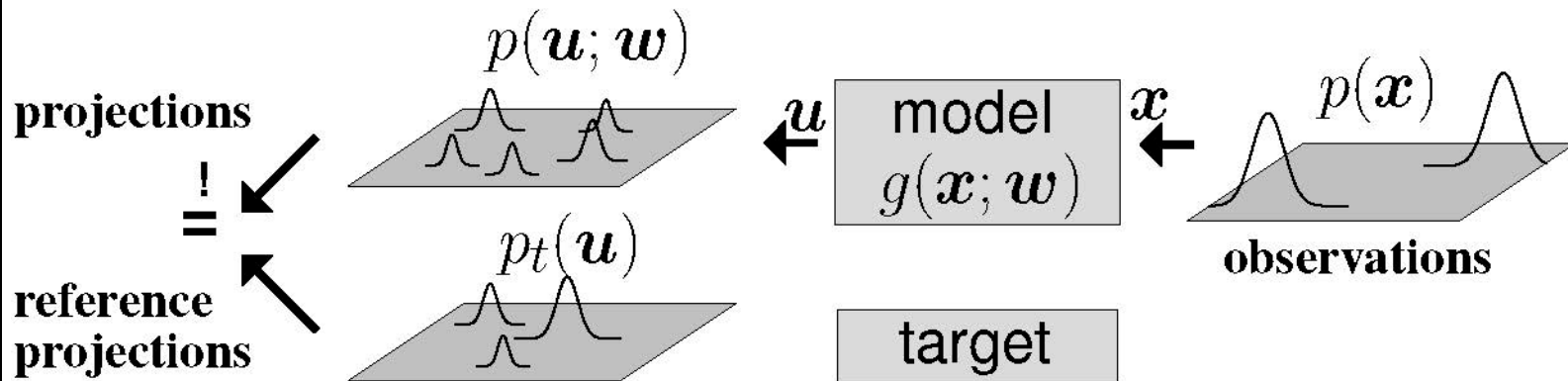
unsupervised categories:

- **generative framework**: density estimation, hidden Markov models  
→ objectives are maximum likelihood or maximum a posteriori
- **recoding or descriptive framework**: projection methods, PCA, ICA  
→ objectives are maximal variance, orthogonality, independence, maximum entropy

# Projection Methods

- 1 Introduction
  - 1.1 Machine Learning Introduction
  - 1.2 Course Specific Introduction
  - 1.3 Generative vs. Descriptive Models
- 2 Basic Terms and Concepts
  - 2.1 Unsupervised Learning in Bioinformatics
  - 2.2 Unsupervised Learning Categories
  - 2.3 Quality of Parameter Estimation
  - 2.4 Maximum Likelihood Estimator
  - 2.5 Expectation Maximization
  - 2.6 Maximum Entropy

projection methods project the data into a space with desired properties





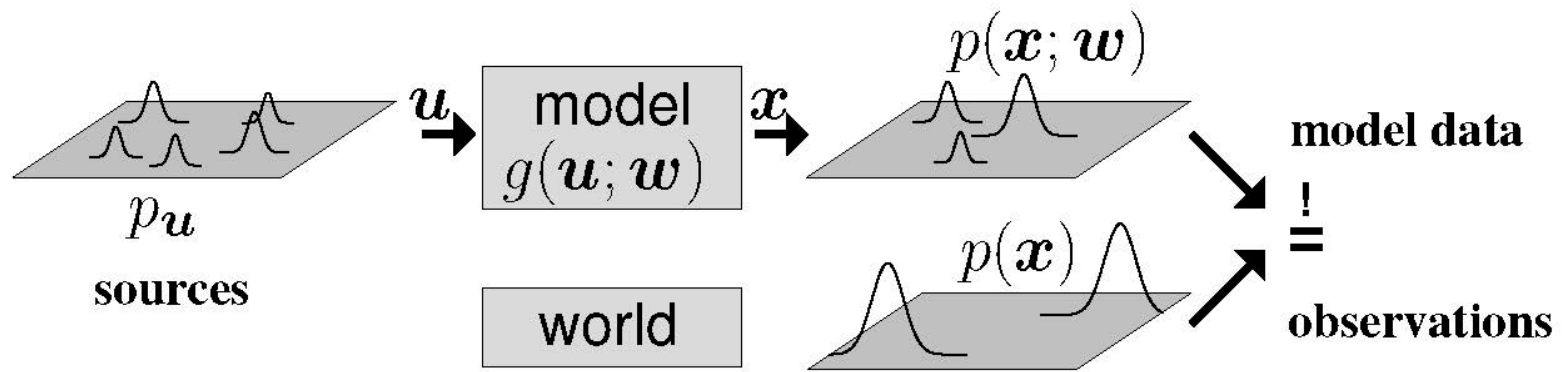
- 1 Introduction
  - 1.1 Machine Learning Introduction
  - 1.2 Course Specific Introduction
  - 1.3 Generative vs. Descriptive Models
- 2 Basic Terms and Concepts
  - 2.1 Unsupervised Learning in Bioinformatics
  - 2.2 Unsupervised Learning Categories
  - 2.3 Quality of Parameter Estimation
  - 2.4 Maximum Likelihood Estimator
  - 2.5 Expectation Maximization
  - 2.6 Maximum Entropy

- Principal Component Analysis (PCA): projection to a low dimensional space under maximal information conservation
- Independent Component Analysis (ICA): projection into a space with statistically independent components (factorial code)
  - often characteristics of a factorial distribution are optimized:
    - maximal entropy (given variance)
    - cummulants
  - or prototype distributions should be matched:
    - product of special super-Gaussians
- Projection Pursuit: components are maximally non-Gaussian

# Generative Models

- 1 Introduction
- 1.1 Machine Learning Introduction
- 1.2 Course Specific Introduction
- 1.3 Generative vs. Descriptive Models
- 2 Basic Terms and Concepts
- 2.1 Unsupervised Learning in Bioinformatics
- 2.2 Unsupervised Learning Categories
- 2.3 Quality of Parameter Estimation
- 2.4 Maximum Likelihood Estimator
- 2.5 Expectation Maximization
- 2.6 Maximum Entropy

a generative model simulates the world and produces the same data





- 1 Introduction
  - 1.1 Machine Learning Introduction
  - 1.2 Course Specific Introduction
  - 1.3 Generative vs. Descriptive Models
- 2 Basic Terms and Concepts
  - 2.1 Unsupervised Learning in Bioinformatics
  - 2.2 Unsupervised Learning Categories
  - 2.3 Quality of Parameter Estimation
  - 2.4 Maximum Likelihood Estimator
  - 2.5 Expectation Maximization
  - 2.6 Maximum Entropy

- data generation process is probabilistic: underlying distribution
- generative model attempts at approximation this distribution
- loss function the distance between model output distribution and the distribution of the data generation process
- examples: factor analysis, latent variable models, Boltzmann machines, hidden Markov models

- 1 Introduction
  - 1.1 Machine Learning Introduction
  - 1.2 Course Specific Introduction
  - 1.3 Generative vs. Descriptive Models
- 2 Basic Terms and Concepts
  - 2.1 Unsupervised Learning in Bioinformatics
  - 2.2 Unsupervised Learning Categories
  - 2.3 Quality of Parameter Estimation
  - 2.4 Maximum Likelihood Estimator
  - 2.5 Expectation Maximization
  - 2.6 Maximum Entropy

Generative models **estimate the true parameter** given a parametrized model class

Data are generated from a model of the class: **find this model**

- model class known
- task: estimate actual (true) parameters
- loss: difference between true and estimated parameter
- evaluate estimator: expected loss

# Mean Squared Error, Bias, and Variance



- 1 Introduction
  - 1.1 Machine Learning Introduction
  - 1.2 Course Specific Introduction
  - 1.3 Generative vs. Descriptive Models
- 2 Basic Terms and Concepts
  - 2.1 Unsupervised Learning in Bioinformatics
  - 2.2 Unsupervised Learning Categories
  - 2.3 Quality of Parameter Estimation
  - 2.4 Maximum Likelihood Estimator
  - 2.5 Expectation Maximization
  - 2.6 Maximum Entropy

## Theoretical concepts of parameter estimation

- training data:  $\{\mathbf{x}\} = \{\mathbf{x}^1, \dots, \mathbf{x}^l\}$

simply  $\mathbf{X} = (\mathbf{x}^1, \dots, \mathbf{x}^l)^T$  (the matrix of training data)

- true parameter vector:  $\mathbf{w}$
- estimate of  $\mathbf{w}$ :  $\hat{\mathbf{w}}$

# Mean Squared Error, Bias, and Variance



- 1 Introduction
  - 1.1 Machine Learning Introduction
  - 1.2 Course Specific Introduction
  - 1.3 Generative vs. Descriptive Models
- 2 Basic Terms and Concepts
  - 2.1 Unsupervised Learning in Bioinformatics
  - 2.2 Unsupervised Learning Categories
  - 2.3 Quality of Parameter Estimation
  - 2.4 Maximum Likelihood Estimator
  - 2.5 Expectation Maximization
  - 2.6 Maximum Entropy

- **unbiased** estimator:  $E_{\mathbf{X}} \hat{\mathbf{w}} = \mathbf{w}$

on average (over training set) the true parameter is obtained

- **bias**:  $b(\hat{\mathbf{w}}) = E_{\mathbf{X}} \hat{\mathbf{w}} - \mathbf{w}$

- **variance**:  $\text{var}(\hat{\mathbf{w}}) = E_{\mathbf{X}} \left( (\hat{\mathbf{w}} - E_{\mathbf{X}}(\hat{\mathbf{w}}))^T (\hat{\mathbf{w}} - E_{\mathbf{X}}(\hat{\mathbf{w}})) \right)$

- **mean squared error** (MSE, different to supervised loss):

$$\text{mse}(\hat{\mathbf{w}}) = E_{\mathbf{X}} \left( (\hat{\mathbf{w}} - \mathbf{w})^T (\hat{\mathbf{w}} - \mathbf{w}) \right)$$

expected squared error between the estimated and true parameter

Objective: **minimize MSE!**

# Mean Squared Error, Bias, and Variance



- 1 Introduction
- 1.1 Machine Learning Introduction
- 1.2 Course Specific Introduction
- 1.3 Generative vs. Descriptive Models
- 2 Basic Terms and Concepts
- 2.1 Unsupervised Learning in Bioinformatics
- 2.2 Unsupervised Learning Categories
- 2.3 Quality of Parameter Estimation
- 2.4 Maximum Likelihood Estimator
- 2.5 Expectation Maximization
- 2.6 Maximum Entropy

$$\text{mse}(\hat{\mathbf{w}}) = \mathbb{E}_{\mathbf{X}} \left( (\hat{\mathbf{w}} - \mathbf{w})^T (\hat{\mathbf{w}} - \mathbf{w}) \right) =$$

$$\mathbb{E}_{\mathbf{X}} \left( \underbrace{((\hat{\mathbf{w}} - \mathbb{E}_{\mathbf{X}}(\hat{\mathbf{w}})) + (\mathbb{E}_{\mathbf{X}}(\hat{\mathbf{w}}) - \mathbf{w}))^T}_{\text{means zero}} \right)$$

means zero

$$((\hat{\mathbf{w}} - \mathbb{E}_{\mathbf{X}}(\hat{\mathbf{w}})) + (\mathbb{E}_{\mathbf{X}}(\hat{\mathbf{w}}) - \mathbf{w})) =$$

$$\mathbb{E}_{\mathbf{X}} \left( (\hat{\mathbf{w}} - \mathbb{E}_{\mathbf{X}}(\hat{\mathbf{w}}))^T (\hat{\mathbf{w}} - \mathbb{E}_{\mathbf{X}}(\hat{\mathbf{w}})) - \right.$$

$$2 (\hat{\mathbf{w}} - \mathbb{E}_{\mathbf{X}}(\hat{\mathbf{w}}))^T (\mathbb{E}_{\mathbf{X}}(\hat{\mathbf{w}}) - \mathbf{w}) +$$

$$\left. (\mathbb{E}_{\mathbf{X}}(\hat{\mathbf{w}}) - \mathbf{w})^T (\mathbb{E}_{\mathbf{X}}(\hat{\mathbf{w}}) - \mathbf{w}) \right) =$$

Only  $\hat{\mathbf{w}}$  depends on  $\mathbf{X}$

$$\mathbb{E}_{\mathbf{X}} \left( (\hat{\mathbf{w}} - \mathbb{E}_{\mathbf{X}}(\hat{\mathbf{w}}))^T (\hat{\mathbf{w}} - \mathbb{E}_{\mathbf{X}}(\hat{\mathbf{w}})) \right) +$$

$$(\mathbb{E}_{\mathbf{X}}(\hat{\mathbf{w}}) - \mathbf{w})^T (\mathbb{E}_{\mathbf{X}}(\hat{\mathbf{w}}) - \mathbf{w}) =$$

$$\text{var}(\hat{\mathbf{w}}) + b^2(\hat{\mathbf{w}})$$

$$\mathbb{E}_{\mathbf{X}} \left( (\hat{\mathbf{w}} - \mathbb{E}_{\mathbf{X}}(\hat{\mathbf{w}}))^T (\mathbb{E}_{\mathbf{X}}(\hat{\mathbf{w}}) - \mathbf{w}) \right) =$$

$$(\mathbb{E}_{\mathbf{X}}(\hat{\mathbf{w}}) - \mathbb{E}_{\mathbf{X}}(\hat{\mathbf{w}}))^T (\mathbb{E}_{\mathbf{X}}(\hat{\mathbf{w}}) - \mathbf{w}) = 0$$

# Maximum Likelihood



1 Introduction  
1.1 Machine Learning Introduction  
1.2 Course Specific Introduction  
1.3 Generative vs. Descriptive Models  
2 Basic Terms and Concepts  
2.1 Unsupervised Learning in Bioinformatics  
2.2 Unsupervised Learning Categories  
2.3 Quality of Parameter Estimation  
2.4 Maximum Likelihood Estimator  
2.5 Expectation Maximization  
2.6 Maximum Entropy

- ML is one of the major objectives in unsupervised learning
- ML is **asymptotically efficient** and **unbiased**
- ML does everything right and this efficiently (enough data)

# Maximum Likelihood Estimator

- 1 Introduction
- 1.1 Machine Learning Introduction
- 1.2 Course Specific Introduction
- 1.3 Generative vs. Descriptive Models
- 2 Basic Terms and Concepts
- 2.1 Unsupervised Learning in Bioinformatics
- 2.2 Unsupervised Learning Categories
- 2.3 Quality of Parameter Estimation
- 2.4 Maximum Likelihood Estimator
- 2.5 Expectation Maximization
- 2.6 Maximum Entropy

The likelihood  $\mathcal{L}$  of the data set  $\{\mathbf{x}\} = \{\mathbf{x}^1, \dots, \mathbf{x}^l\}$ :

$$\mathcal{L}(\{\mathbf{x}\}; \mathbf{w}) = p(\{\mathbf{x}\}; \mathbf{w})$$

probability of the model  $p(\mathbf{x}; \mathbf{w})$  to produce the data

iid (independent identical distributed) data:

$$\mathcal{L}(\{\mathbf{x}\}; \mathbf{w}) = p(\{\mathbf{x}\}; \mathbf{w}) = \prod_{i=1}^l p(\mathbf{x}^i; \mathbf{w})$$

Negative log-likelihood:

$$- \ln \mathcal{L}(\{\mathbf{x}\}; \mathbf{w}) = - \sum_{i=1}^l \ln p(\mathbf{x}^i; \mathbf{w})$$

# Properties of Maximum Likelihood Estimator

- 1 Introduction
  - 1.1 Machine Learning Introduction
  - 1.2 Course Specific Introduction
  - 1.3 Generative vs. Descriptive Models
- 2 Basic Terms and Concepts
  - 2.1 Unsupervised Learning in Bioinformatics
  - 2.2 Unsupervised Learning Categories
  - 2.3 Quality of Parameter Estimation
  - 2.4 Maximum Likelihood Estimator
  - 2.5 Expectation Maximization
  - 2.6 Maximum Entropy

MLE:

- invariant under parameter change
- asymptotically unbiased and efficient  $\rightarrow$  asymptotically optimal
- asymptotically consistent

consistent:  $\hat{w} \xrightarrow{l \rightarrow \infty} w$

for large training sets the estimator approaches the true value  
(difference to unbiased  $\rightarrow$  variance decreases)



# Expectation Maximization

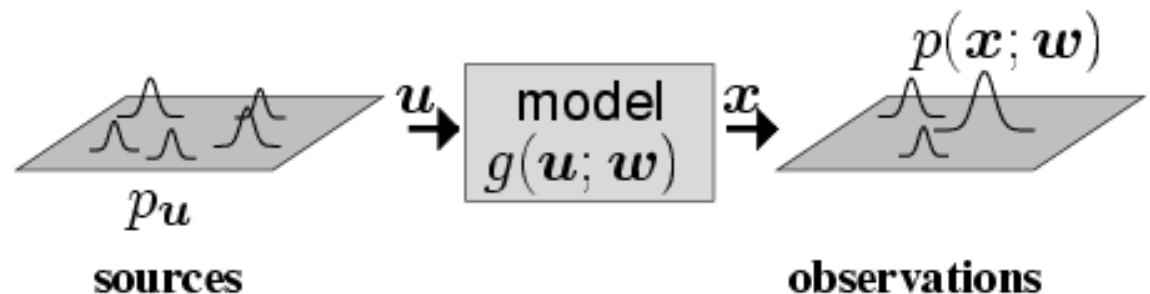
- 1 Introduction
- 1.1 Machine Learning Introduction
- 1.2 Course Specific Introduction
- 1.3 Generative vs. Descriptive Models
- 2 Basic Terms and Concepts
- 2.1 Unsupervised Learning in Bioinformatics
- 2.2 Unsupervised Learning Categories
- 2.3 Quality of Parameter Estimation
- 2.4 Maximum Likelihood Estimator
- 2.5 Expectation Maximization
- 2.6 Maximum Entropy

likelihood can be optimized by gradient descent methods

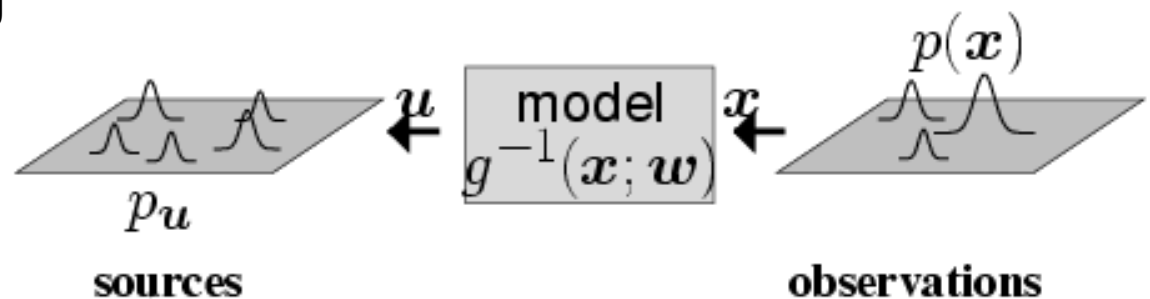
likelihood **cannot** be computed analytically:

- hidden states
- many-to-one output mapping
- non-linearities

## Generative Model



## Likelihood



$$p(x; w) = \int_U p_u(u) \delta(x = g(u; w)) du$$

# Expectation Maximization

- 1 Introduction
- 1.1 Machine Learning Introduction
- 1.2 Course Specific Introduction
- 1.3 Generative vs. Descriptive Models
- 2 Basic Terms and Concepts
- 2.1 Unsupervised Learning in Bioinformatics
- 2.2 Unsupervised Learning Categories
- 2.3 Quality of Parameter Estimation
- 2.4 Maximum Likelihood Estimator
- 2.5 Expectation Maximization
- 2.6 Maximum Entropy

- **hidden variables, latent variables, unobserved variables**  $u$
- likelihood is determined by all  $u$  mapped to  $x$

$$p(\boldsymbol{x}; \boldsymbol{w}) = \int_U p_u(\boldsymbol{u}) \delta(\boldsymbol{x} = g(\boldsymbol{u}; \boldsymbol{w})) d\boldsymbol{u}$$

# Expectation Maximization

## 1 Introduction

### 1.1 Machine Learning Introduction

### 1.2 Course Specific Introduction

### 1.3 Generative vs. Descriptive Models

## 2 Basic Terms and Concepts

### 2.1 Unsupervised Learning in Bioinformatics

### 2.2 Unsupervised Learning Categories

### 2.3 Quality of Parameter Estimation

### 2.4 Maximum Likelihood Estimator

### 2.5 Expectation Maximization

### 2.6 Maximum Entropy

Expectation Maximization (EM) algorithm:

- joint probability  $p(\mathbf{x}, \mathbf{u}; \mathbf{w})$  is easier to compute than likelihood
- estimate  $p(\mathbf{u} | \mathbf{x}; \mathbf{w})$  by  $Q(\mathbf{u} | \mathbf{x})$

$$\ln \mathcal{L}(\{\mathbf{x}\}; \mathbf{w}) = \ln p(\{\mathbf{x}\}; \mathbf{w}) = \ln \int_U p(\{\mathbf{x}\}, \mathbf{u}; \mathbf{w}) d\mathbf{u} =$$

$$\ln \int_U \frac{Q(\mathbf{u} | \{\mathbf{x}\})}{Q(\mathbf{u} | \{\mathbf{x}\})} p(\{\mathbf{x}\}, \mathbf{u}; \mathbf{w}) d\mathbf{u} \geq \quad \text{Jensen's inequality}$$

$$\int_U Q(\mathbf{u} | \{\mathbf{x}\}) \ln \frac{p(\{\mathbf{x}\}, \mathbf{u}; \mathbf{w})}{Q(\mathbf{u} | \{\mathbf{x}\})} d\mathbf{u} =$$

$$\int_U Q(\mathbf{u} | \{\mathbf{x}\}) \ln p(\{\mathbf{x}\}, \mathbf{u}; \mathbf{w}) d\mathbf{u} -$$

$$\int_U Q(\mathbf{u} | \{\mathbf{x}\}) \ln Q(\mathbf{u} | \{\mathbf{x}\}) d\mathbf{u} =$$

$$\mathcal{F}(Q, \mathbf{w})$$

Expectation of log joint probability is easy for exponential family

# Expectation Maximization

- 1 Introduction
- 1.1 Machine Learning Introduction
- 1.2 Course Specific Introduction
- 1.3 Generative vs. Descriptive Models
- 2 Basic Terms and Concepts
- 2.1 Unsupervised Learning in Bioinformatics
- 2.2 Unsupervised Learning Categories
- 2.3 Quality of Parameter Estimation
- 2.4 Maximum Likelihood Estimator
- 2.5 Expectation Maximization
- 2.6 Maximum Entropy

EM algorithm is an iteration between E-step and M-step:

**E-step:**

$$Q_{k+1} = \arg \max_Q \mathcal{F}(Q, w_k)$$

**M-step:**

$$w_{k+1} = \arg \max_w \mathcal{F}(Q_{k+1}, w)$$

# Expectation Maximization

- 1 Introduction
  - 1.1 Machine Learning Introduction
  - 1.2 Course Specific Introduction
  - 1.3 Generative vs. Descriptive Models
- 2 Basic Terms and Concepts
  - 2.1 Unsupervised Learning in Bioinformatics
  - 2.2 Unsupervised Learning Categories
  - 2.3 Quality of Parameter Estimation
  - 2.4 Maximum Likelihood Estimator
  - 2.5 Expectation Maximization
  - 2.6 Maximum Entropy

EM increases the lower bound in both steps

Beginning of the M-step:  $\mathcal{F}(Q_{k+1}, \mathbf{w}_k) = \ln \mathcal{L}(\{\mathbf{x}\}; \mathbf{w}_k)$

E-step does not change the parameters

$$\begin{aligned} \ln \mathcal{L}(\{\mathbf{x}\}; \mathbf{w}_k) &= \mathcal{F}(Q_{k+1}, \mathbf{w}_k) \leq \\ \mathcal{F}(Q_{k+1}, \mathbf{w}_{k+1}) &\leq \mathcal{F}(Q_{k+2}, \mathbf{w}_{k+1}) = \ln \mathcal{L}(\{\mathbf{x}\}; \mathbf{w}_{k+1}) \end{aligned}$$

EM algorithm:

- hidden Markov models
- mixture of Gaussians
- factor analysis
- independent component analysis

- 1 Introduction
  - 1.1 Machine Learning Introduction
  - 1.2 Course Specific Introduction
  - 1.3 Generative vs. Descriptive Models
- 2 Basic Terms and Concepts
  - 2.1 Unsupervised Learning in Bioinformatics
  - 2.2 Unsupervised Learning Categories
  - 2.3 Quality of Parameter Estimation
  - 2.4 Maximum Likelihood Estimator
  - 2.5 Expectation Maximization
  - 2.6 Maximum Entropy

## maximum entropy probability distribution:

- maximal entropy given a class of distributions
- minimal prior assumptions
- physical systems converge to maximal entropy configurations
- most likely observed solution
- connection: statistical mechanics and information theory

**principle of maximum entropy** first expounded by E.T. Jaynes in 1957

- 1 Introduction
  - 1.1 Machine Learning Introduction
  - 1.2 Course Specific Introduction
  - 1.3 Generative vs. Descriptive Models
- 2 Basic Terms and Concepts
  - 2.1 Unsupervised Learning in Bioinformatics
  - 2.2 Unsupervised Learning Categories
  - 2.3 Quality of Parameter Estimation
  - 2.4 Maximum Likelihood Estimator
  - 2.5 Expectation Maximization
  - 2.6 Maximum Entropy

$$\text{Entropy } H = - \sum_{k \geq 1} p_k \log p_k$$

$$p_k \log p_k = 0 \text{ for } p_k = 0$$

Examples:

- normal distribution: given mean and standard deviation
- uniform distribution: supported in the interval  $[a, b]$
- exponential distribution: given mean in  $[0, \infty]$

- 1 Introduction
  - 1.1 Machine Learning Introduction
  - 1.2 Course Specific Introduction
  - 1.3 Generative vs. Descriptive Models
- 2 Basic Terms and Concepts
  - 2.1 Unsupervised Learning in Bioinformatics
  - 2.2 Unsupervised Learning Categories
  - 2.3 Quality of Parameter Estimation
  - 2.4 Maximum Likelihood Estimator
  - 2.5 Expectation Maximization
  - 2.6 Maximum Entropy

Not all classes of distributions contain a maximum entropy distribution:

- arbitrarily large entropy: distributions with mean
- entropies of a class are bounded from above but not attained: distributions with mean zero, second moment one, and third moment one



# Maximum Entropy Solution

- 1 Introduction
- 1.1 Machine Learning Introduction
- 1.2 Course Specific Introduction
- 1.3 Generative vs. Descriptive Models
- 2 Basic Terms and Concepts
- 2.1 Unsupervised Learning in Bioinformatics
- 2.2 Unsupervised Learning Categories
- 2.3 Quality of Parameter Estimation
- 2.4 Maximum Likelihood Estimator
- 2.5 Expectation Maximization
- 2.6 Maximum Entropy

Constraints: 
$$\sum_{i=1}^n p(x_i) f_k(x_i) = F_k \quad k = 1, \dots, m$$

$$\sum_{i=1}^n p(x_i) = 1$$

Solution, the **Gibbs distribution**

$$p(x_i) = \frac{1}{Z(\lambda_1, \dots, \lambda_m)} \exp(\lambda_1 f_1(x_i) + \dots + \lambda_m f_m(x_i))$$

with **partition function**

$$Z(\lambda_1, \dots, \lambda_m) = \sum_{i=1}^n \exp(\lambda_1 f_1(x_i) + \dots + \lambda_m f_m(x_i))$$

The **Lagrange multipliers** are determined by the equation system

$$F_k = \frac{\partial}{\partial \lambda_k} \log Z(\lambda_1, \dots, \lambda_m)$$