Adaptive Feature Selection in Image Segmentation

Volker Roth, Tilman Lange

ETH Zurich, Institute of Computational Science,
CH-8092 Zurich, Switzerland
{vroth, langet}@inf.ethz.ch

Outline

1. Image segmentation as a **clustering problem**

2. Feature extraction and **data fusion**
   ⇝ "stacked" feature vectors

3. **Adaptive feature selection** in clustering

4. Choosing #(segments) and #(features)
   ⇝ **model section** by resampling

5. **Segmentation Results**.

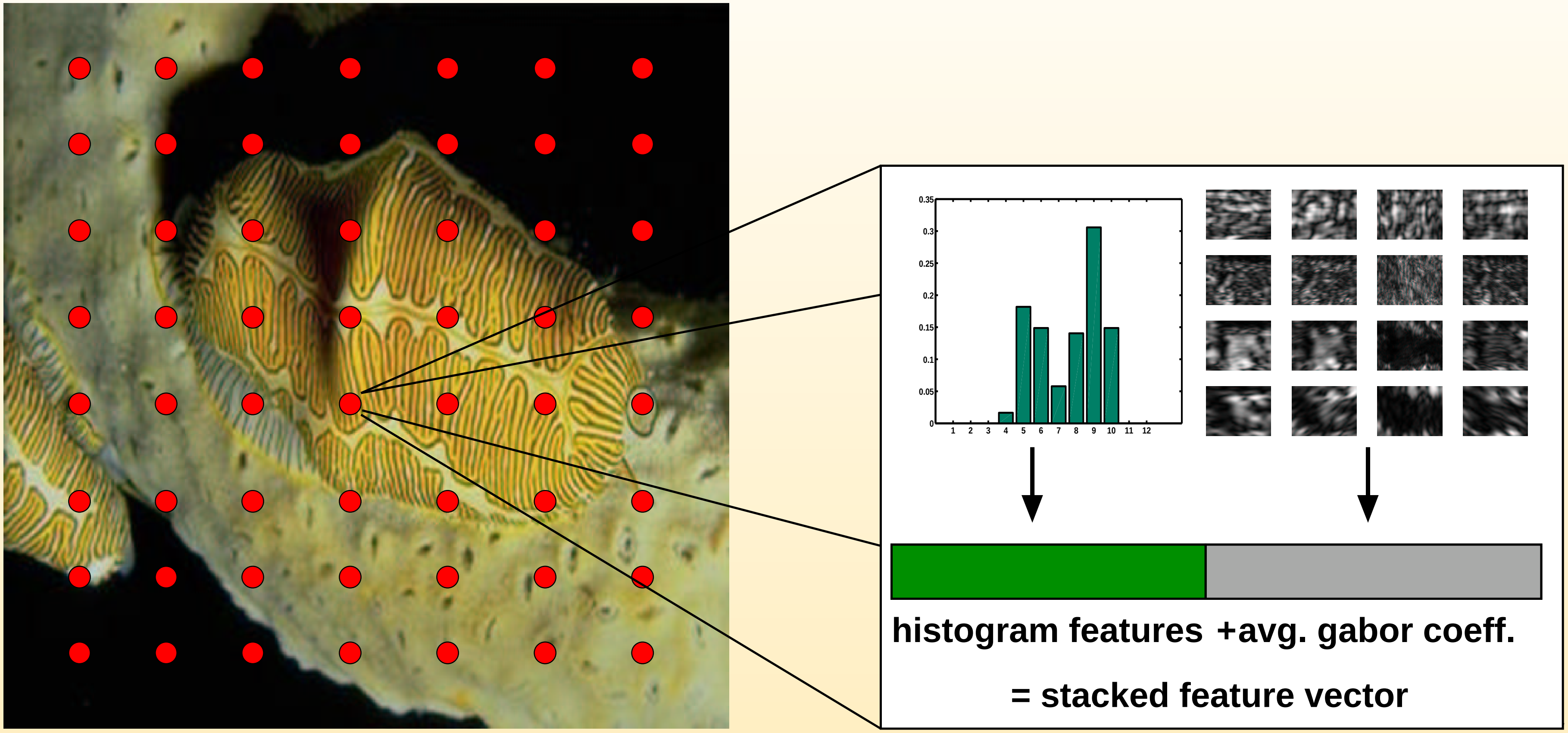
Image segmentation

- **Goal:** divide image into connected regions
  $\rightsquigarrow$ semantic equivalence classes.
- Semantic classes modeled **indirectly** by similarity criterion based on **low-level features.**

Image segmentation

- **Goal:** divide image into connected regions
  $\rightsquigarrow$ semantic equivalence classes.
- Semantic classes modeled **indirectly** by similarity criterion based on **low-level features.**
- **Problem:** informative information spread over different types of features $\rightsquigarrow$ **how to combine them?**
- **Naive solution:** simply stack all different features into a high-dimensional vector.

Stacked feature vectors



Goal: assign image sites to clusters $\rightsquigarrow$ segmentation.

Stacked feature vectors: problems

- Grouping algorithms become **instable in high dimensions**: increasing number of local minima.

Stacked feature vectors: problems

- Grouping algorithms become **instable in high dimensions**: increasing number of local minima.
- **Special structure of stacked vectors**
  - only few features **relevant** for actual task.
  - most features contain merely “noise”

Stacked feature vectors: problems

- Grouping algorithms become **instable in high dimensions**: increasing number of local minima.

- **Special structure of stacked vectors**
  - only few features **relevant** for actual task.
  - most features contain merely “noise”

- Relevance determination by **adaptive feature selection** mechanism.

Feature selection in image segmentation

- **Goal:** simultaneously discover groups of image sites and subsets of **relevant** features.

Feature selection in image segmentation

- **Goal:** simultaneously discover groups of image sites and subsets of **relevant** features.

- **Segmentation** is an **unsupervised** problem: feature selection complicated due to missing labels.

Feature selection in image segmentation

- **Goal:** simultaneously discover groups of image sites and subsets of **relevant** features.

- **Segmentation** is an **unsupervised** problem: feature selection complicated due to missing labels.

- **Common problems:**
  - Lack of **model selection** strategy $\rightsquigarrow$ ambiguities.

Feature selection in image segmentation

- **Goal:** simultaneously discover groups of image sites and subsets of **relevant** features.

- **Segmentation** is an **unsupervised** problem: feature selection complicated due to missing labels.

- **Common problems:**
  - Lack of **model selection** strategy $\rightsquigarrow$ ambiguities.
  - Stepwise approach: iterated clustering and relevance determination steps
    $\rightsquigarrow$ **different objective functions**.

Wish-list & Solutions

- Overcome inherent ambiguities.

Model selection: observing **stability** of cluster solutions.

Wish-list & Solutions

- Overcome inherent ambiguities.

Model selection: observing **stability** of cluster solutions.

- Clustering and feature selection should optimize the same objective function.

incorporating **Automatic Relevance Determination** principle into a Gaussian mixture model: feature selection and clustering maximize **same likelihood function**.

Gaussian mixtures and relevance determination

- **Data:** collection of N image sites
- **Clustering of samples:** Gaussian mixture model with k components and identical covariance Σ.

Gaussian mixtures and relevance determination

- **Data:** collection of N image sites
- **Clustering of samples:** Gaussian mixture model with k components and identical covariance Σ.
- Maximize likelihood by **EM-algorithm**:
  - **E-step:** estimate cluster-membership probabilities
  - **M-step:** find model parameters $(\boldsymbol{\mu}_k, \Sigma)$ that maximize the likelihood.

M-step as indicator regression

- M-step restated as **linear regression** of data X against current class-membership probabilities $\boldsymbol{z}$:

$$\text{minimize } \|\boldsymbol{z} - X\boldsymbol{\beta}\|_2^2.$$

$\rightsquigarrow$ Re-parameterization: $(\boldsymbol{\mu}, \Sigma) \to \boldsymbol{\beta}$.

M-step as indicator regression

- M-step restated as **linear regression** of data X against current class-membership probabilities $\boldsymbol{z}$:

$$\text{minimize } \|\boldsymbol{z} - X\boldsymbol{\beta}\|_2^2.$$

$\rightsquigarrow$ Re-parameterization: $(\boldsymbol{\mu}, \Sigma) \to \boldsymbol{\beta}$.

- **Automatic relevance determination** (ARD) priors:

$$p(\boldsymbol{\beta}|\,\boldsymbol{\vartheta}) = \prod_i \mathcal{N}(0, \vartheta_i^{-1}).$$

$\vartheta_i \to \infty$: i-th feature removed $\rightsquigarrow$ relevance variable

M-step as indicator regression (cont'd)

- Bayesian view: instead of **selecting** relevance variables ($\rightsquigarrow$ **hard combinatorial problem!**)...

M-step as indicator regression (cont'd)

- Bayesian view: instead of **selecting** relevance variables ($\rightsquigarrow$ **hard combinatorial problem!**)...
  choose hyperpriors $p(\vartheta_i) = \frac{\gamma}{2} \exp\{-\frac{\gamma\vartheta_i}{2}\}$

M-step as indicator regression (cont'd)

- Bayesian view: instead of **selecting** relevance variables ($\rightsquigarrow$ **hard combinatorial problem!**)...
  choose hyperpriors $p(\vartheta_i) = \frac{\gamma}{2} \exp\{-\frac{\gamma \vartheta_i}{2}\}$
  and **average**:

$$p(\beta_i) = \int_0^\infty p(\beta_i|\vartheta_i) p(\vartheta_i)\, d\vartheta_i = \frac{\gamma}{2} \exp\{-\sqrt{\gamma}|\beta_i|\}.$$

M-step as indicator regression (cont'd)

- Bayesian view: instead of **selecting** relevance variables ($\leadsto$ **hard combinatorial problem!**)...
  choose hyperpriors $p(\vartheta_i) = \frac{\gamma}{2}\exp\{-\frac{\gamma\vartheta_i}{2}\}$
  and **average**:

$$p(\beta_i) = \int_0^\infty p(\beta_i|\vartheta_i)p(\vartheta_i)\,d\vartheta_i = \frac{\gamma}{2}\exp\{-\sqrt{\gamma}|\beta_i|\}.$$

- ℓ_1-constrained regression in log-space $\Rightarrow$ LASSO:

$$\text{minimize } \|\boldsymbol{z} - X\boldsymbol{\beta}\|_2^2 + \lambda'\|\boldsymbol{\beta}\|_1.$$

Combined clustering/selection model

E-step	**M-step**	
$\downarrow$	M-step $\downarrow$	by linear regression
$\downarrow$	ARD priors $\downarrow$	$p(\boldsymbol{\beta}\vert\,\boldsymbol{\vartheta})=\prod_i \mathcal{N}(0,\vartheta_i^{-1})$
$\downarrow$	Hyperpriors $\downarrow$	$p(\vartheta_i)=\frac{\gamma}{2}\exp\{-\frac{\gamma\vartheta_i}{2}\}$
$\downarrow$	Marginalization $\downarrow$	$p(\beta_i)=\int_0^\infty p(\beta_i\vert\vartheta_i)p(\vartheta_i)\,d\vartheta$
Prob$(\boldsymbol{x}_i \in \text{class } \eta)$	**Optimize LASSO**	

Combined clustering/selection model

E-step	**M-step**
$\downarrow$	M-step $\downarrow$ by linear regression
$\downarrow$	ARD priors $\downarrow$ $p(\boldsymbol{\beta}\mid \boldsymbol{\vartheta}) = \prod_i \mathcal{N}(0, \vartheta_i^{-1})$
$\downarrow$	Hyperpriors $\downarrow$ $p(\vartheta_i) = \frac{\gamma}{2} \exp\{-\frac{\gamma \vartheta_i}{2}\}$
$\downarrow$	Marginalization $\downarrow$ $p(\beta_i) = \int_0^\infty p(\beta_i \mid \vartheta_i) p(\vartheta_i)\, d\vartheta$
Prob$(\boldsymbol{x}_i \in \text{class } \eta)$	**Optimize LASSO**

Clustering and feature selection optimize the same objective function $\rightsquigarrow$ constrained likelihood

Model selection by resampling: Idea

- **Sampling:** take two random samples from the same data source

Model selection by resampling: Idea

- **Sampling:** take two random samples from the same data source
- **Inference:** learn mixture models from both samples

Model selection by resampling: Idea

- **Sampling:** take two random samples from the same data source
- **Inference:** learn mixture models from both samples
- **Prediction:** use **first** model to predict cluster membership of data in **second** sample

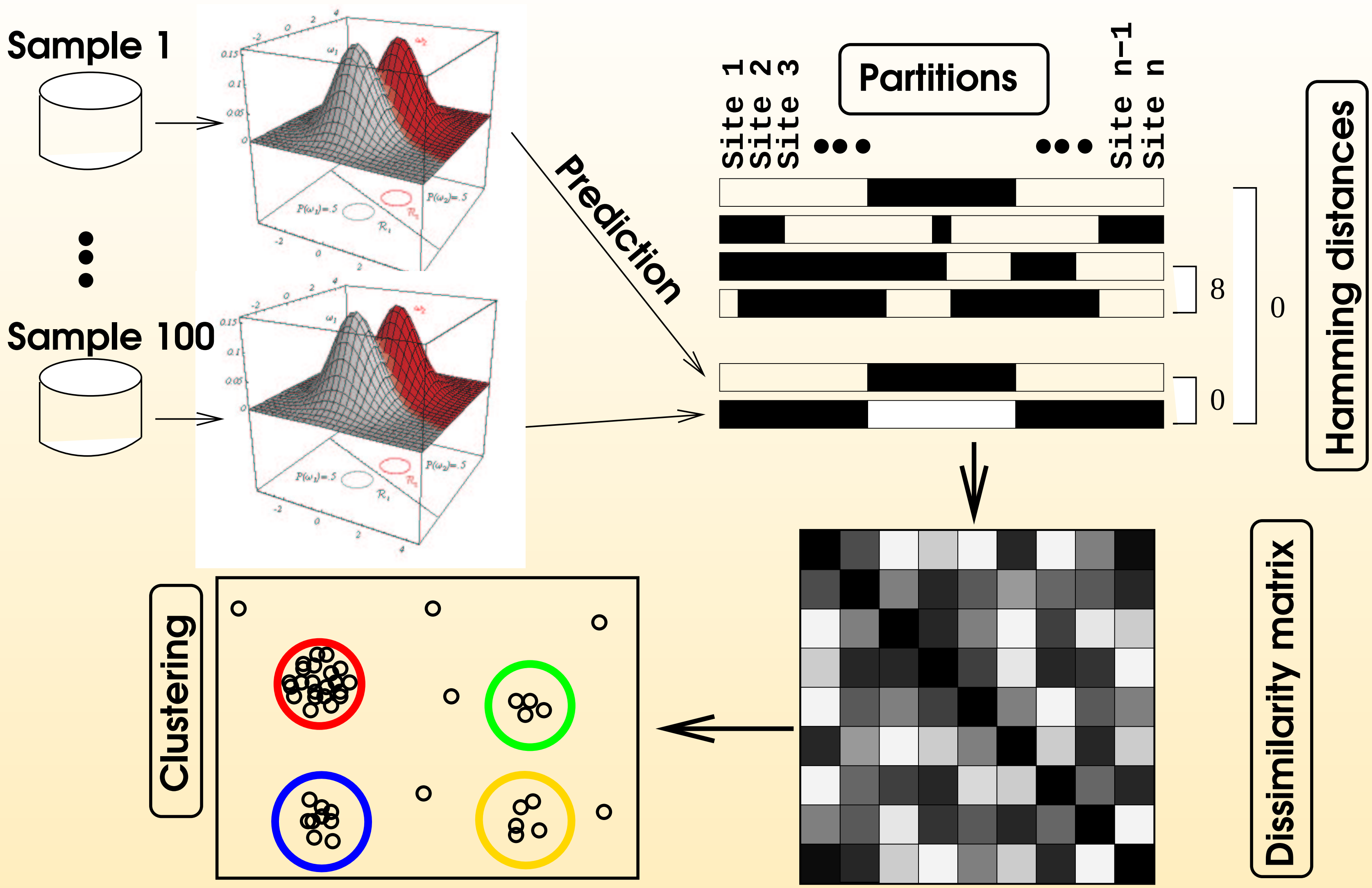
Model selection by resampling: Idea

- **Sampling:** take two random samples from the same data source

- **Inference:** learn mixture models from both samples

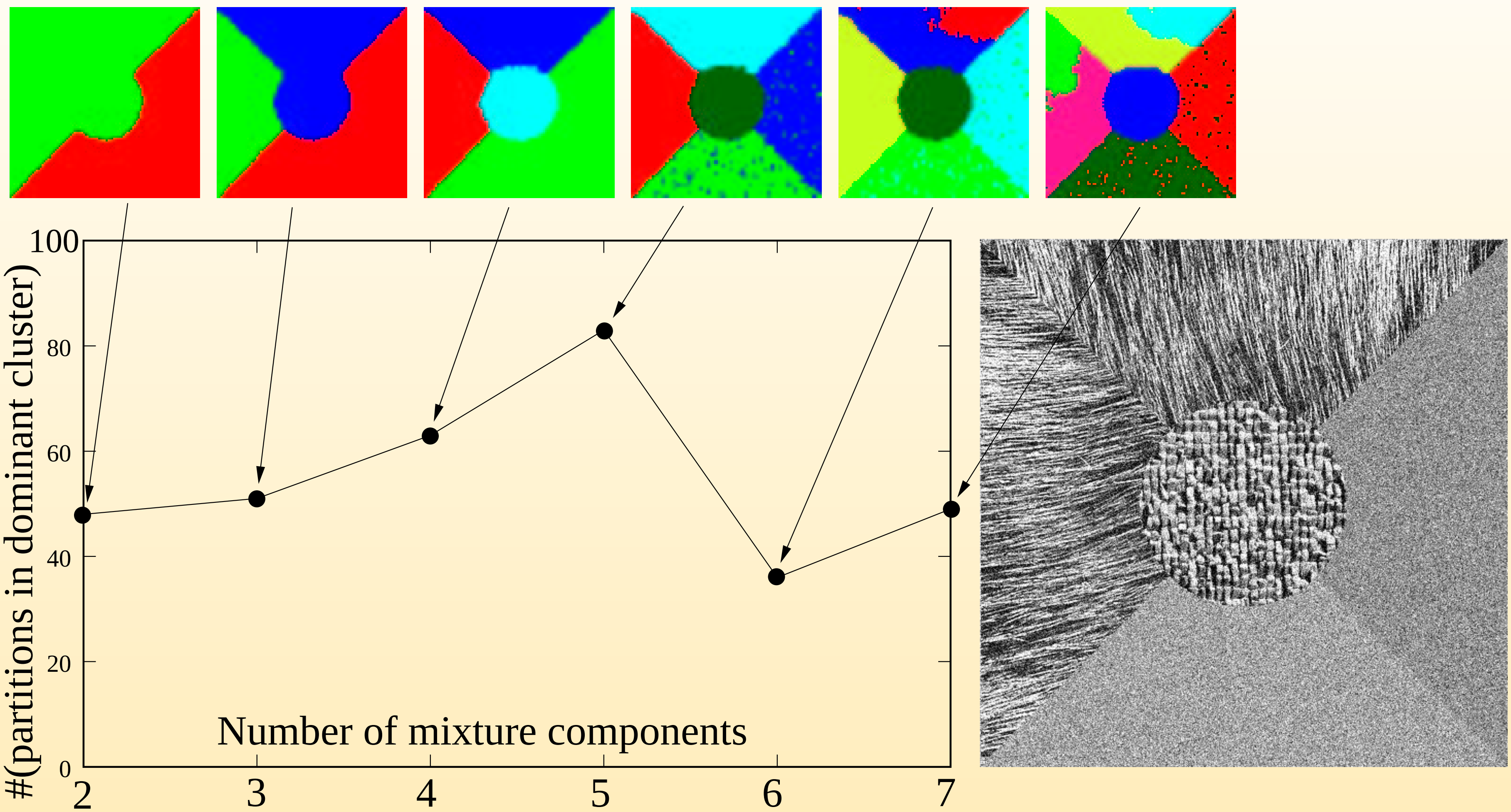
- **Prediction:** use **first** model to predict cluster membership of data in **second** sample

- **Comparison:** for **second** sample, compare cluster labels inferred from **second** model with **predicted** labels.

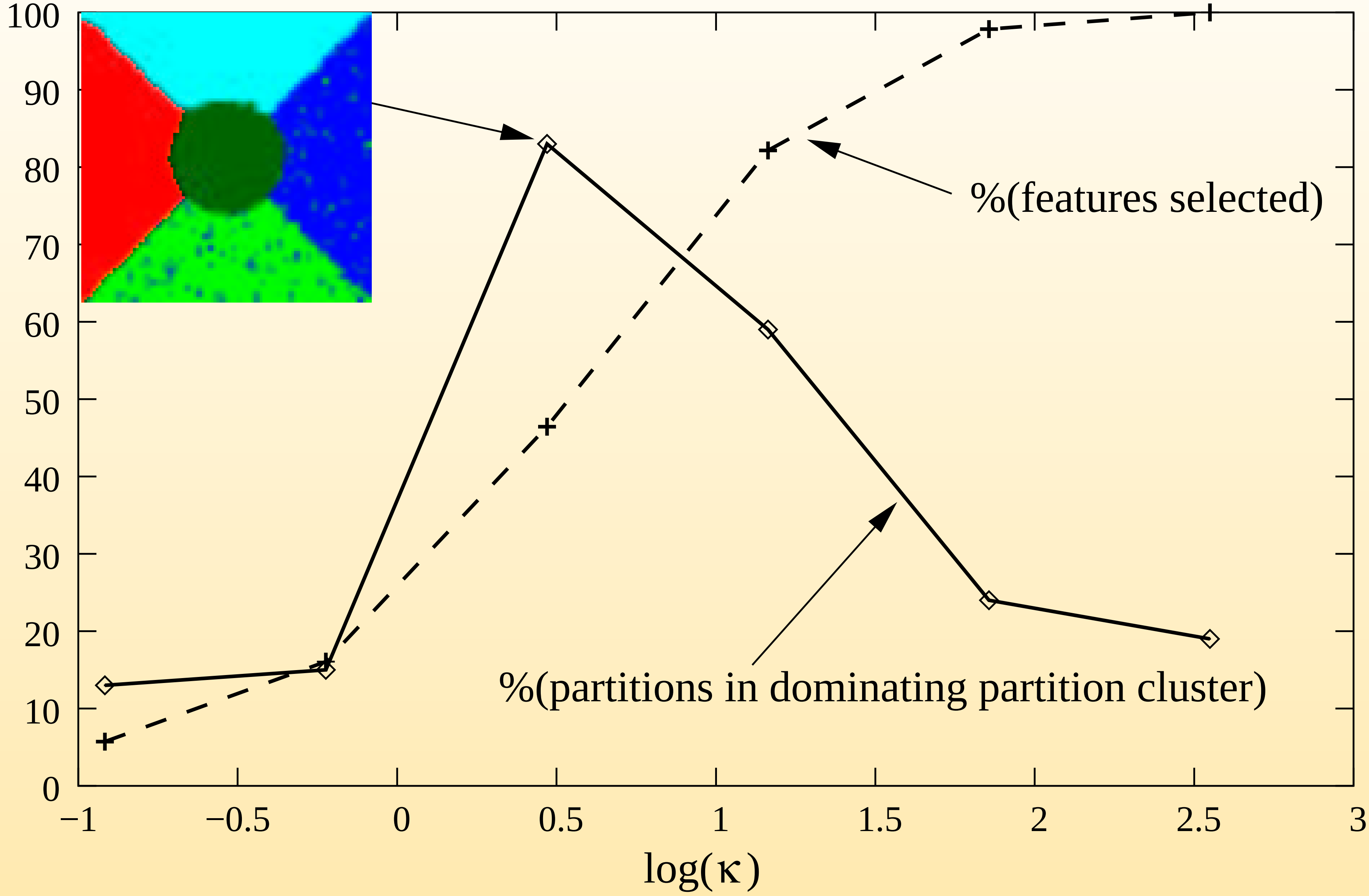


Sample 1
Sample 100
Prediction
Partitions
Site 1
Site 2
Site 3
Site n−1
Site n
8
0
0
Hamming distances
Dissimilarity matrix
Clustering

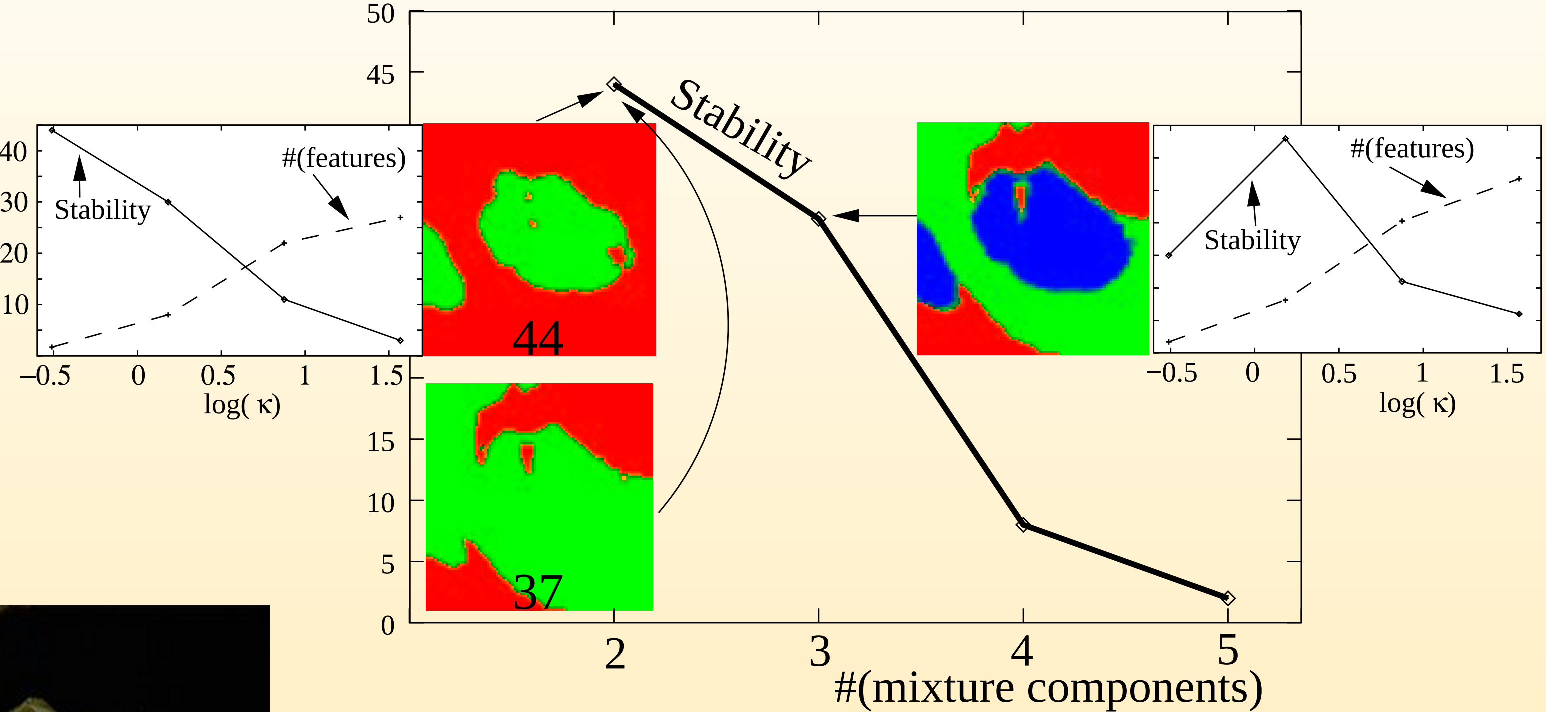
Model selection: model order

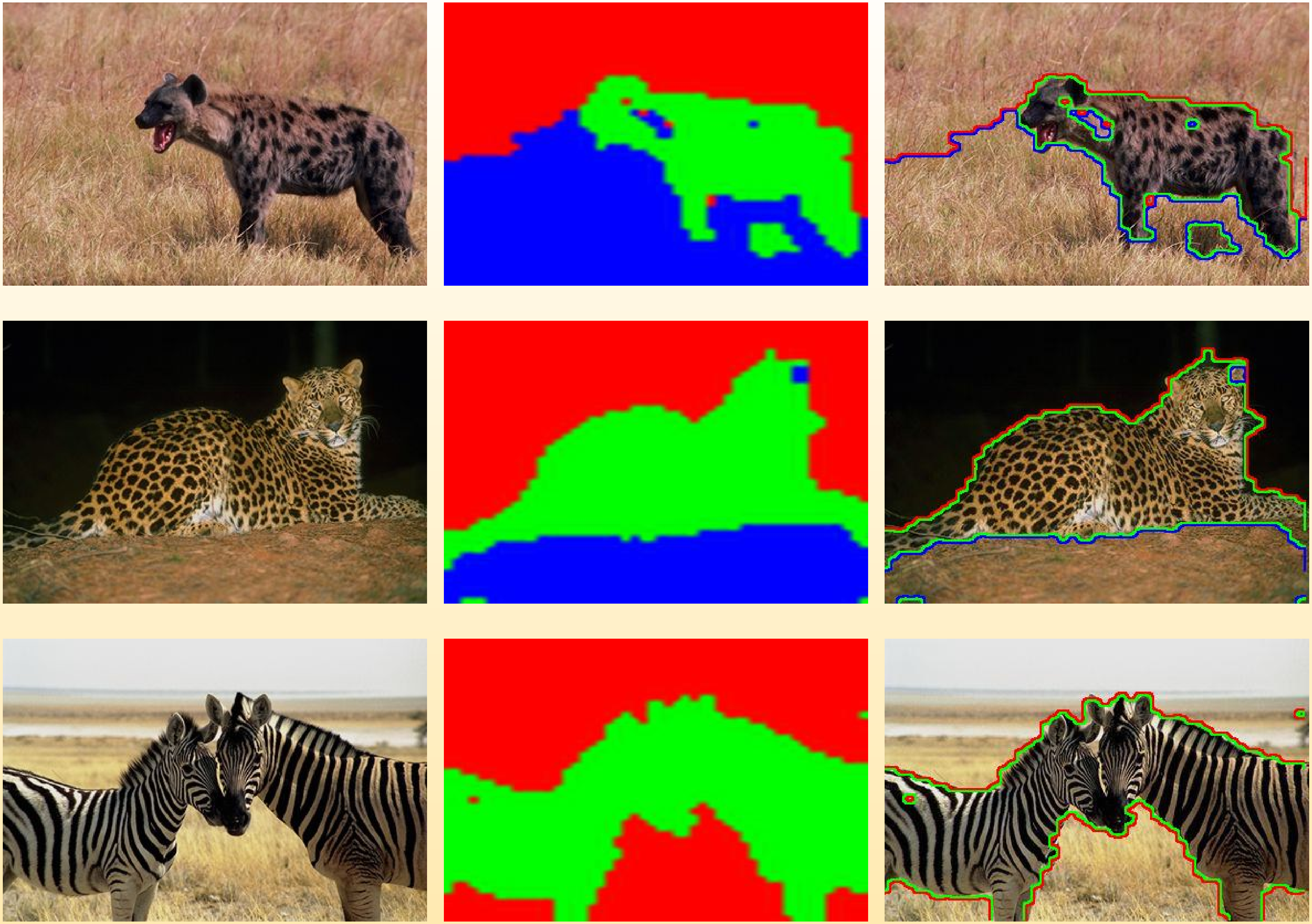


Model selection: optimal #(features)



Model selection: real world images





Summing up

- Informative information often spread over different types of features $\rightsquigarrow$ **data fusion** $\rightsquigarrow$ **stacked vectors**

- Most entries of stacked vectors are **irrelevant** for actual task $\rightsquigarrow$ need for **feature selection**

- Feature selection **difficult** due to missing labels.
  - Overcoming instabilities & ambiguities $\rightsquigarrow$ **stability-based model selection.**
  - Grouping and selection should optimize same objective function $\rightsquigarrow$ **Gaussian mixture model with built-in relevance determination.**