ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

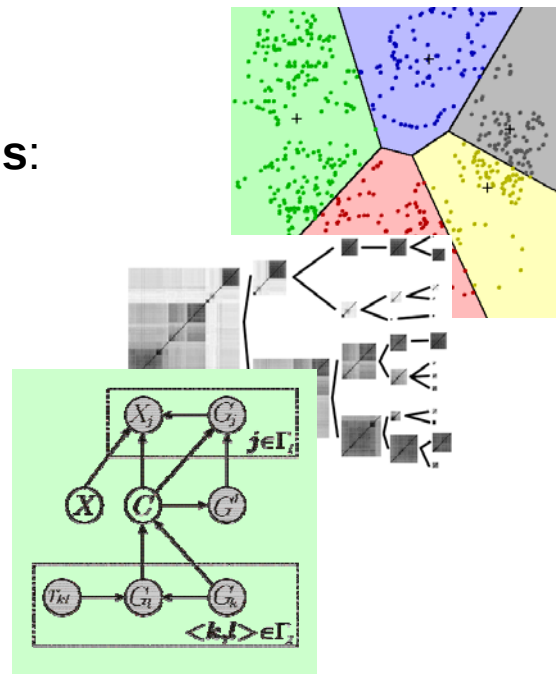
Structure validation in clustering by stability analysis

Joachim M. Buhmann

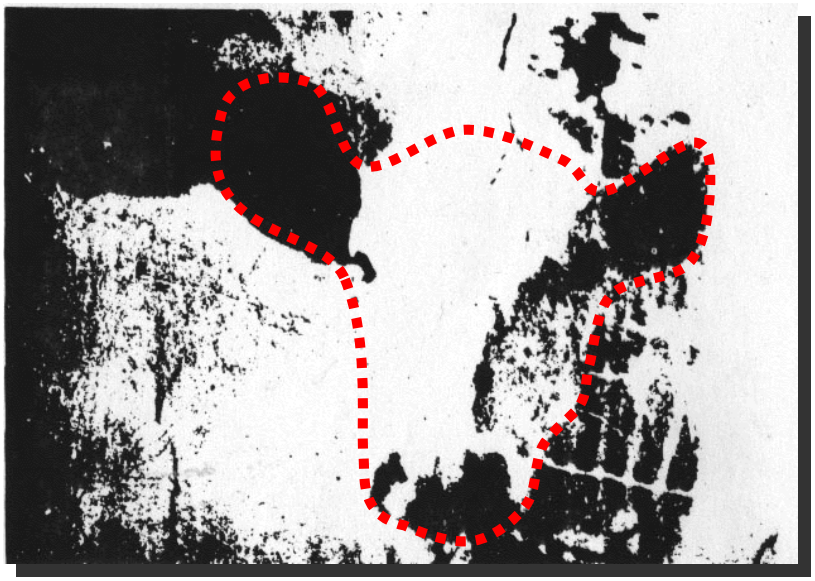
Computer Science Department, ETH Zurich

Learning discrete structures under uncertainty

- **Partitions or clusterings**:
  Which model?
  How many clusters?
- **Trees or dendrograms**:
  What depth of the tree?
  How many leaves?
- **Graphical models**:
  How many edges?



Do you recognize the object ?



What is information?

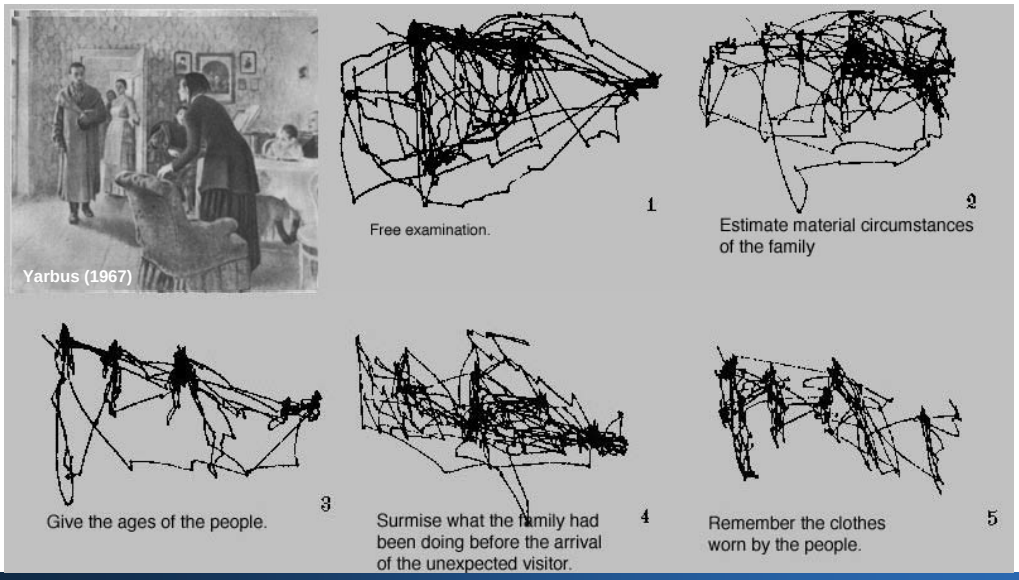
Is Shannon information useful

as the technical measure of information?

620.393 Bytes **842.414 Bytes**

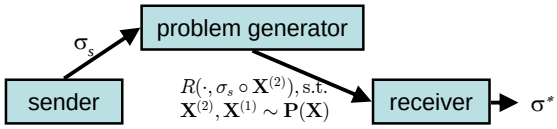
Relevant information in image – where can we find it ?

The relevance of **information** is determined by the task !

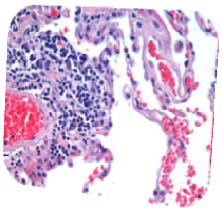


Machine learning & statistical modelling

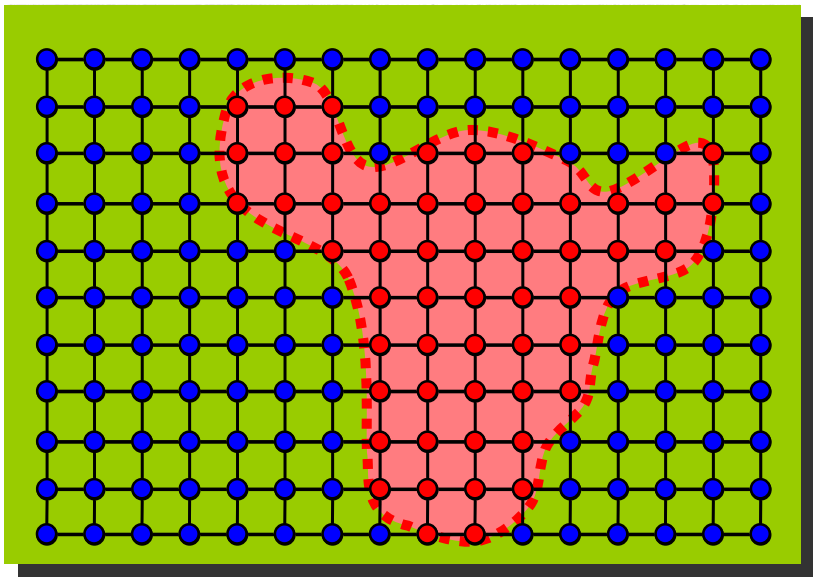
- **Approximate optimization, stability** and information theory
  - optimization as coding
  - statistical models



- ***Medical image processing*** of **tissue data** for high throughput cancer diagnosis
  - **computational pathology**: *Tissue Microarrays* for clear cell renal carcinoma



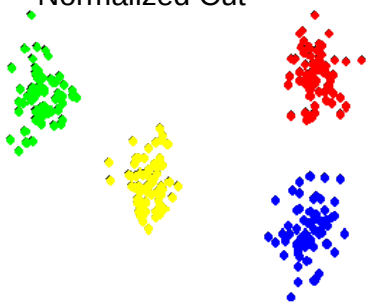
What is the „right“ model for figure ground segmentation?



Clustering models for figure ground segmentation?

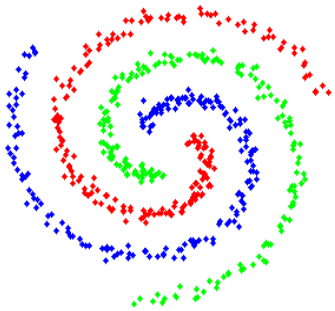
Compactness criterion

- K-Means Clustering
- Pairwise Clustering, AvAssoc
- Correlation Clustering
- Max-Cut, Average Cut
- Normalized Cut



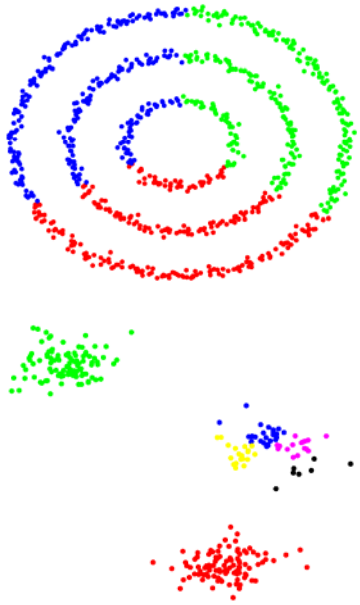
Connectedness criterion

- Single Linkage
- Path Based Clustering

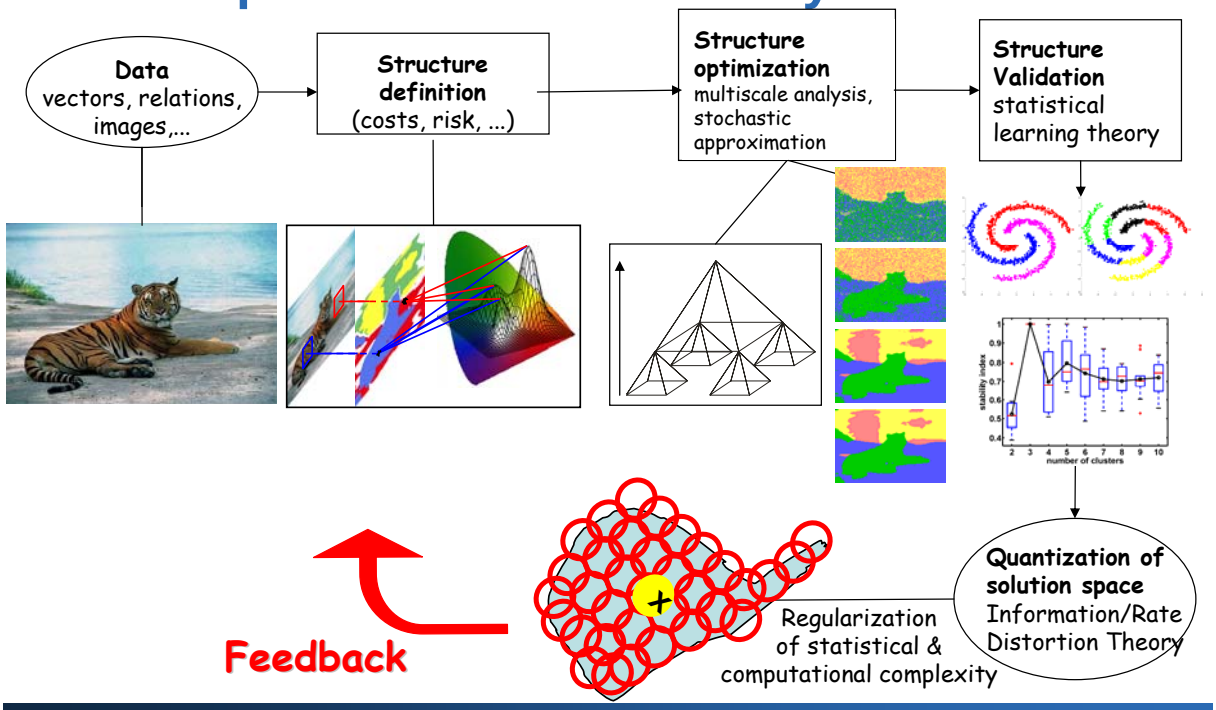


Design problems in clustering: validation

- **Modeling problem**: Does the cluster model "describe" the data? Selection of the costs/hypothesis class!
- **Model order selection problem**: Is the number of clusters and/or features correct?



Concept for robust data analysis



Mathematical formalization of clustering

- Given: **object space** $\mathcal{O}$ with **objects** $o \in \mathcal{O}$.
- Given: **measurement space** $\mathcal{X}$
- **data** are relations $(o, \mathbf{X}) \in \mathcal{O} \times \mathcal{X}$
- Clusterings **partition** objects into groups, i.e.,

$$\begin{aligned} c \ : \ \mathcal{O} \times \mathcal{X} \ &\rightarrow \ \{1, \ldots, k\} \\ (o, \mathbf{X}) \ &\mapsto \ c(o, \mathbf{X}) \end{aligned}$$

- Hypothesis class $c \in \mathcal{C} \equiv \{\text{partitions of data}\}$

Order relation of clusterings

- **algorithm** α selects "statistically optimal" clusterings $\alpha : \mathcal{O} \times \mathcal{X} \rightarrow \mathcal{C}_\gamma \subset \mathcal{C}$
- Remark: α could minimize a cost function where $\mathcal{C}_\gamma$ is a γ close approximation of the minimum.
- **Model selection problem**: Which properties should a good clustering algorithm α possess?
- **Stability! Small changes of data should yield similar clusterings.**

Why risk approximation?

- **Data often contain noise**! Very frequently **data** are best modeled as **random variables**.
- An *empirically optimal* clustering often is statistically indistinguishable from other *equally plausible* data partitionings!
- **Data noise reduces resolution in data space!**

=> This quantization induces a **quantization of the hypothesis class** for structures.

Empirical Risk Approximation

- **Learning**: sample typical solutions of an approximation set $\mathcal{C}_\gamma^{(1)} \equiv \mathcal{C}_\gamma(\mathbf{X}^{(1)})$ given data $\mathbf{X}^{(1)}$

$$c \in \mathcal{C}_\gamma^{(1)} \equiv \left\{ c : d\left(c(\mathbf{X}^{(1)}), c^\perp(\mathbf{X}^{(1)})\right) \leq \gamma \right\}$$

- Algorithm: Gibbs sampling of clusterings with temperature $T(\gamma)$ explores approximation set.
- **Interpretation**: $T(\gamma)$ controls the resolution of the hypothesis class, i.e., the **minimal similarity** of **statistically indistinguishable structures**

Approximate optimization and information theory

- **Problem**: Data processing in Machine Learning is often formulated as an optimization question.

=> noisy data require **robustness = generalization**

- Use **approximate optimization results** as **code**
- **"Communication"** is achieved via **approximate optimization of instances** since test instances are considered to be perturbed training instances.

Two instance setting for learning

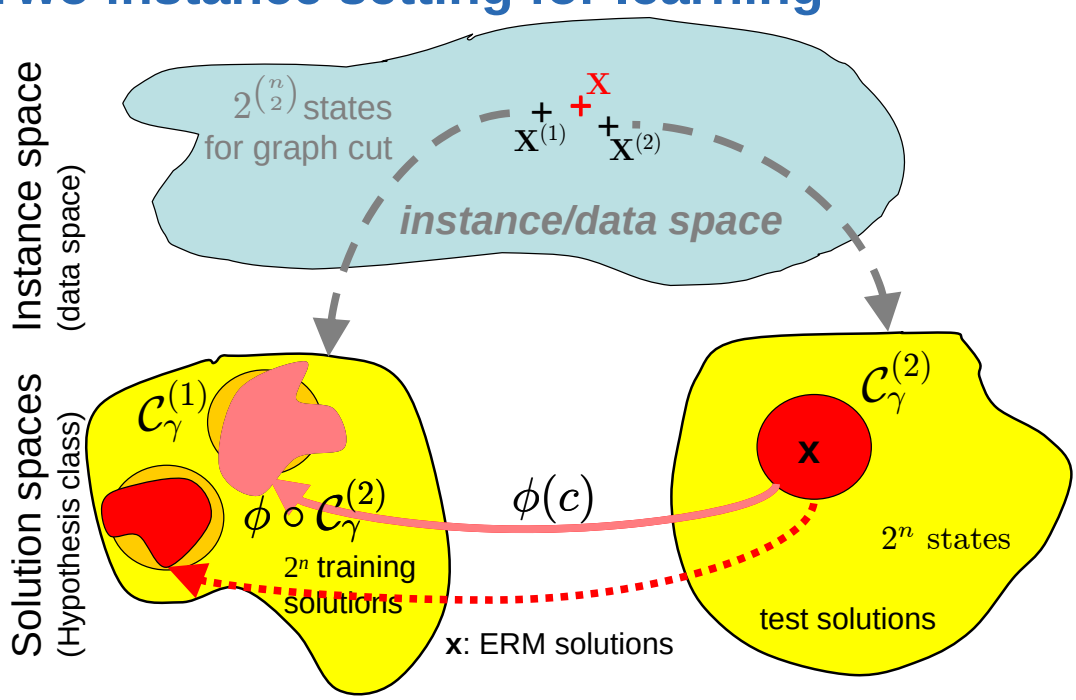
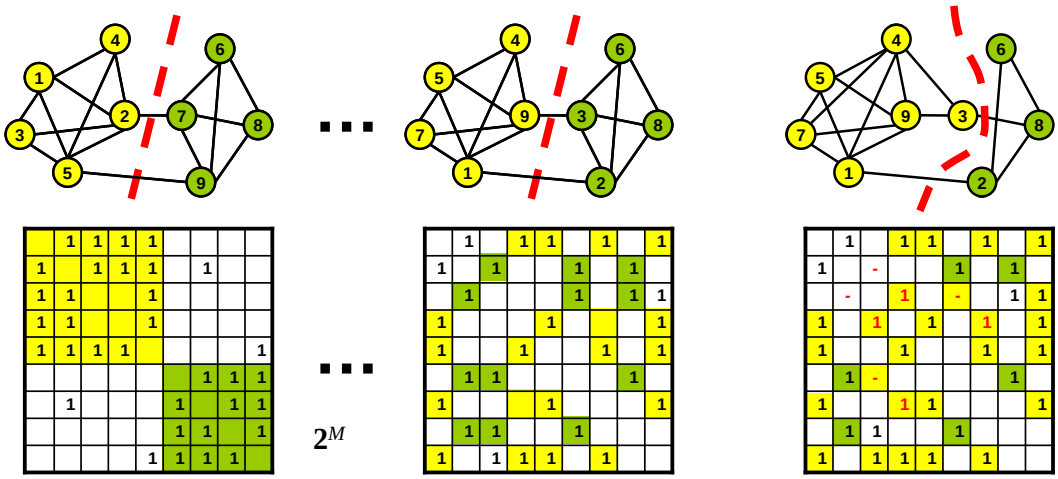


Figure Ground Segmentation by Graph Cut

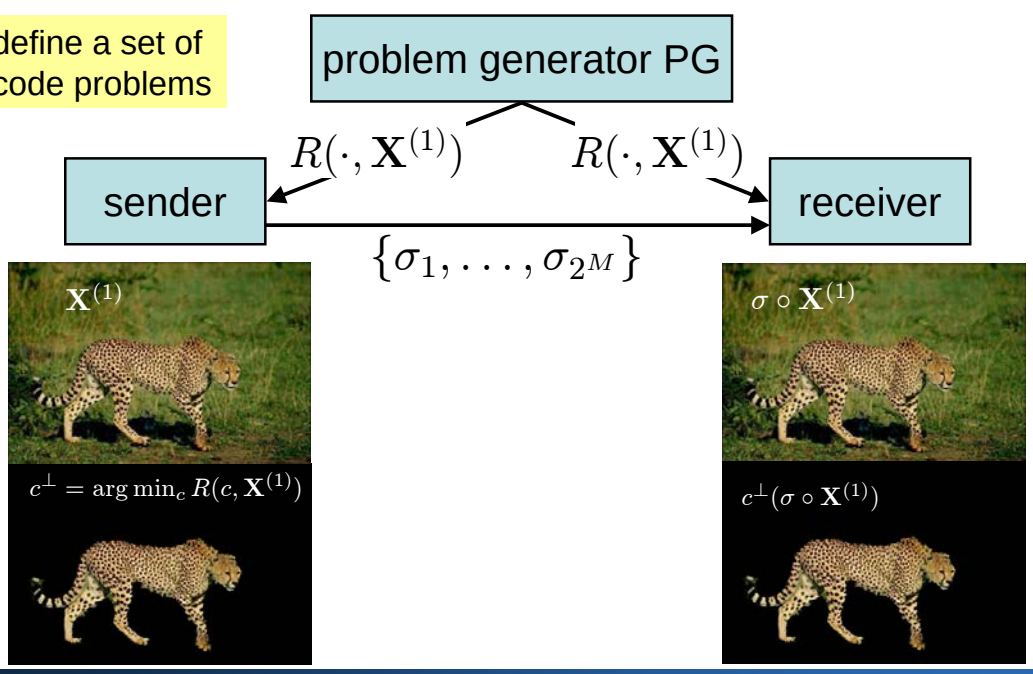
graph cut code problems

graph cut test problem

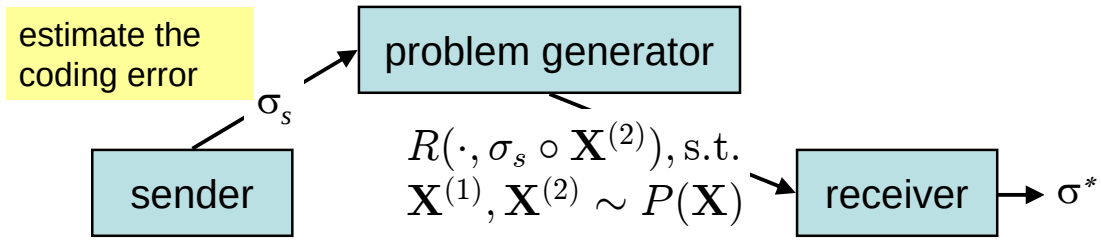


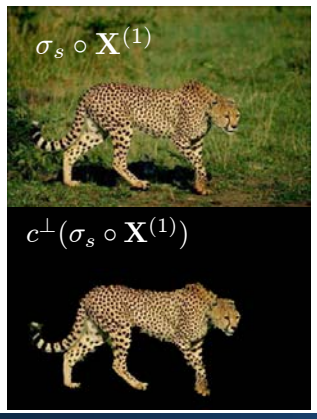
Robust Approximations by Stability

define a set of code problems

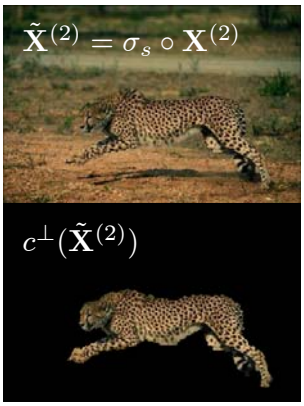


Robust Approximations by Stability





1. Sender sends a permutation index σ_s to problem generator.
2. Problem generator sends a new problem with permuted indices to receiver without revealing σ_s.
3. Receiver identifies the permutation σ^* by comparing approximation sets.



Communication Process

- Receiver has to **compare sets of clusterings** $\mathcal{C}_\gamma(\mathbf{X}^{(1)})$ of training instance (code problem) with approximate clusterings $\mathcal{C}_\gamma(\mathbf{X}^{(2)})$ of the test data.
- Define a mapping $\phi : \mathcal{C}(\mathbf{X}^{(2)}) \to \mathcal{C}(\mathbf{X}^{(1)})$
- **Decoding**

$$\sigma^* = \arg\max_{\sigma} \left| \mathcal{C}_\gamma(\sigma \circ \mathbf{X}^{(1)}) \cap \phi\left(\mathcal{C}_\gamma(\tilde{\mathbf{X}}^{(2)})\right) \right|$$

$$\text{if} \quad \frac{|\mathcal{C}_\gamma(\sigma^* \circ \mathbf{X}^{(1)}) \cap \phi\left(\mathcal{C}_\gamma(\tilde{\mathbf{X}}^{(2)})\right)|}{|\mathcal{C}_\gamma(\sigma^* \circ \mathbf{X}^{(1)})|} \geq 1 - \epsilon$$

Error Analysis and Approximation Capacity

- Approximations of sender and receiver have little in common! => **Irreproducibility**
  This condition determines approximation precision
- **Approximations** of test problem has a large overlap with approximations of wrong training problem! => **Confusion**

=> **Select model** with maximal information capacity, i.e., high precision and high noise robustness

Approximation Capacity

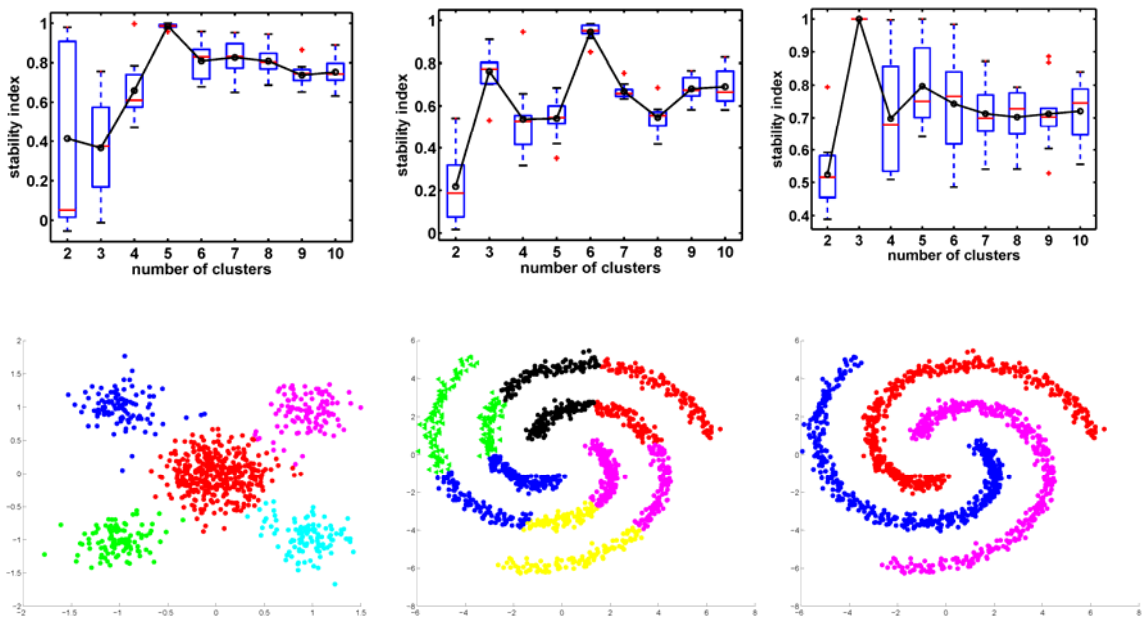
- Condition of **vanishing total error**

$$\begin{aligned} M \log 2 \quad &< \quad \log |\phi \circ \mathcal{C}_\gamma(\sigma_s \circ X^{(2)}) \cap \mathcal{C}_\gamma(\sigma_j \circ X^{(1)})| \\ &\quad - \log |\mathcal{C}_\gamma(\sigma_s \circ X^{(2)})| - H(\sigma_s) \\ &\equiv \quad \mathcal{I}\big(\sigma_s, \mathcal{C}_\gamma(\sigma_s \circ X^{(2)})\big) \end{aligned}$$

mutual information

- **Model selection**: Maximize the mutual information w.r.t. **topology** of solution space, **metric** of solution space, **cost** function, transfer function ϕ and approximation **precision** γ.

Results on Toy Data



Clustering of Microarray Data

(dataset from Golub *et al.*, Science, Oct. 1999, pp.531-537)

Task: Find groups of different Leukemia tumour samples

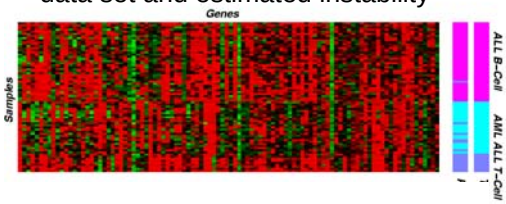
(two- and three class classifications are known).

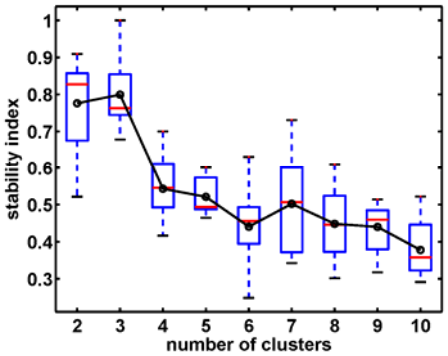
Problem: Number of groups is unknown a priori.

Via Stability with *k*-means: Estimated number of groups is 3.

Result: 3-means solution recovers 91% of known sample classifications.

3-means grouping of Golub *et al.* data set and estimated instability

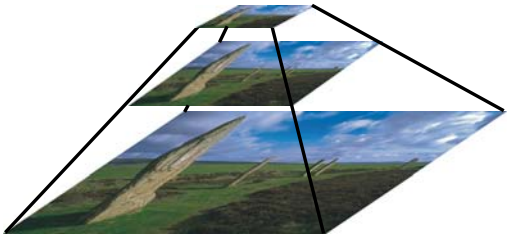




Scales in Data Analysis and Vision

Refinement of Variable Space

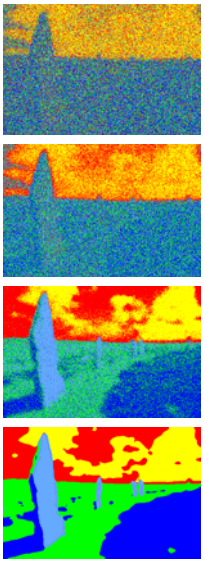
coarse



fine

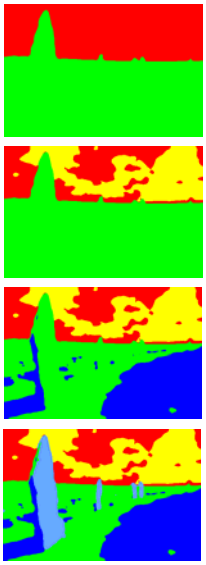
Increment Level of Resolution Pyramid

Refinement of Optimization Criterion



Increase Regularization

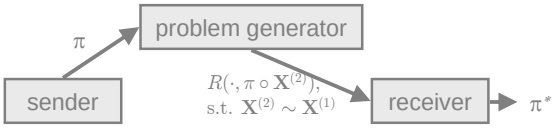
Refinement of Model Order



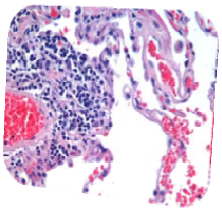
Increase # of Segments

Machine learning & statistical modelling

- **Approximate optimization, stability** and information theory
  - optimization as coding
  - statistical models



- ***Medical image processing*** of **tissue data** for high throughput cancer diagnosis (MICCAI '08)
  - **computational pathology**: *Tissue Microarrays* for clear cell renal carcinoma



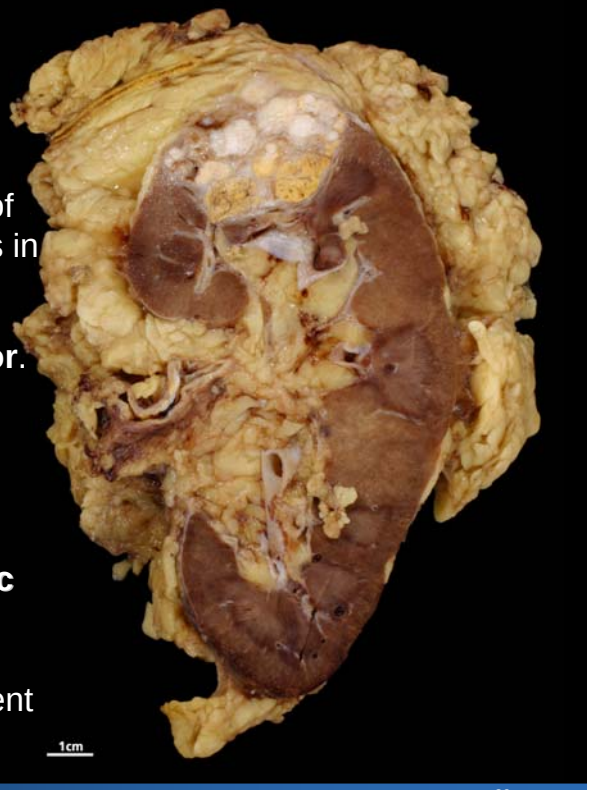
Renal (Clear) Cell Carcinoma

Renal cell carcinoma (RCC) is one of the ten most frequent malignancies in Western societies.

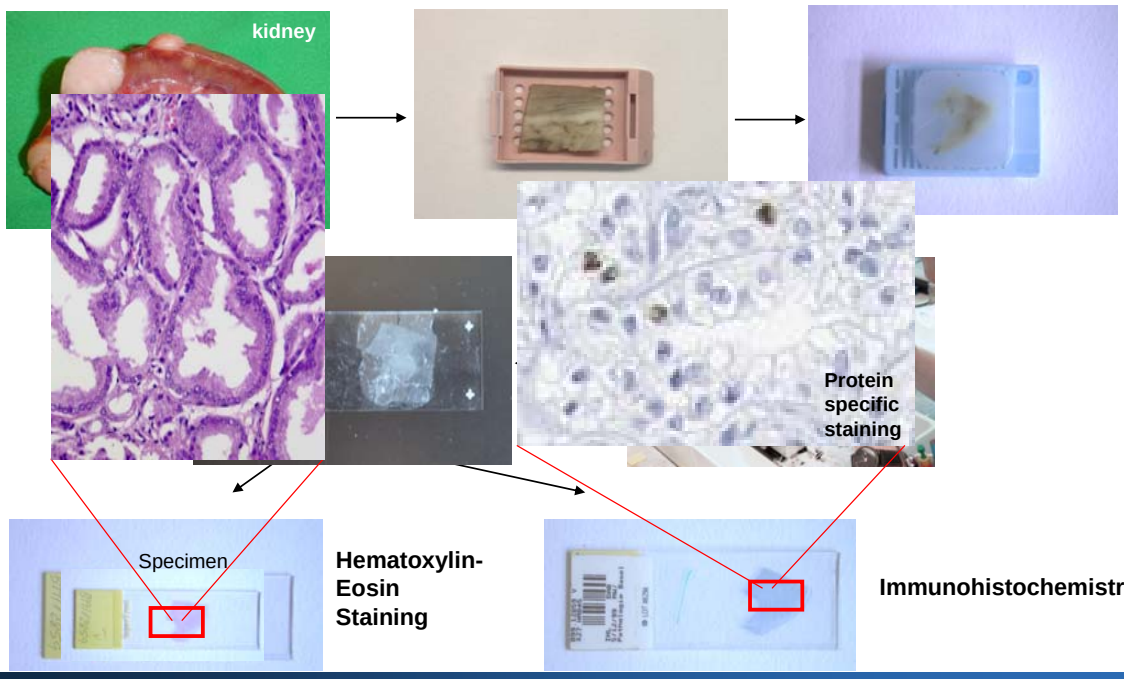
The **prognosis** of renal cancer is **poor**.

Many patients suffer already from metastases at first diagnosis.

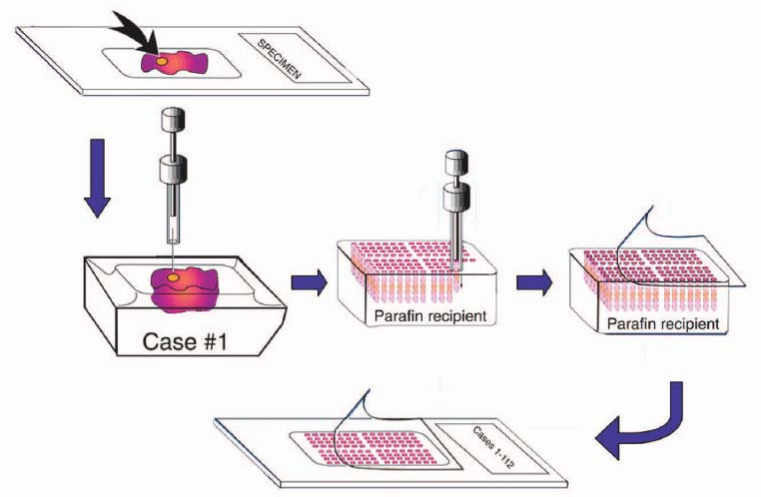
The identification of biomarkers for prediction of prognosis (**prognostic marker**) or response to therapy (**predictive marker**) is therefore of utmost importance to improve patient prognosis.



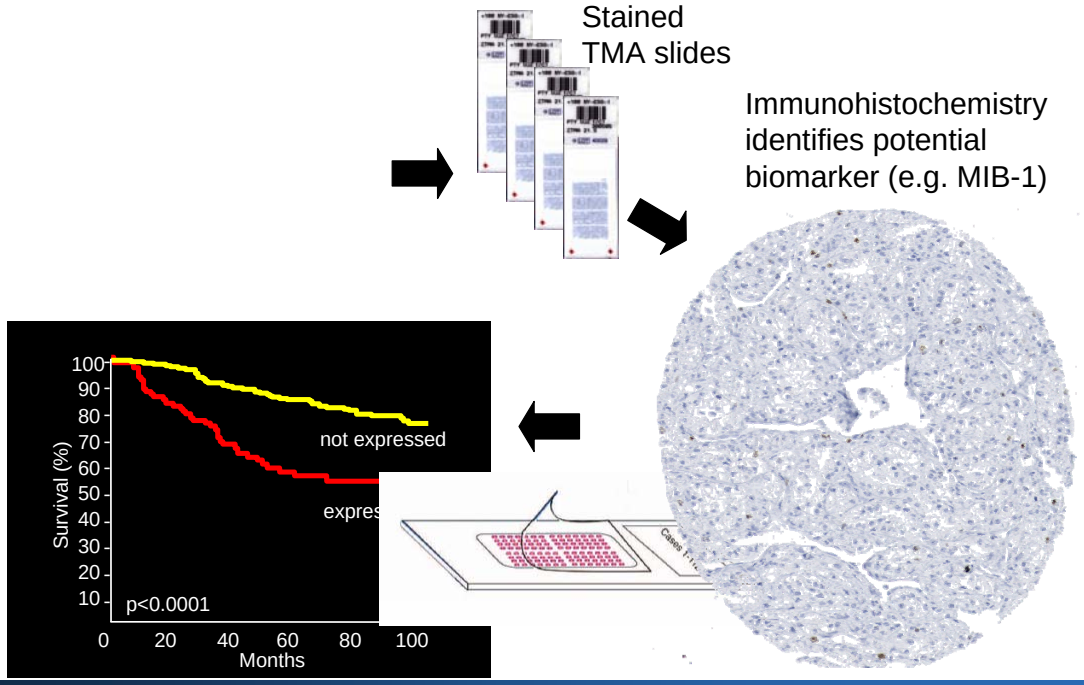
Tissue Microarray Preparation



Tissue Microarray preparation



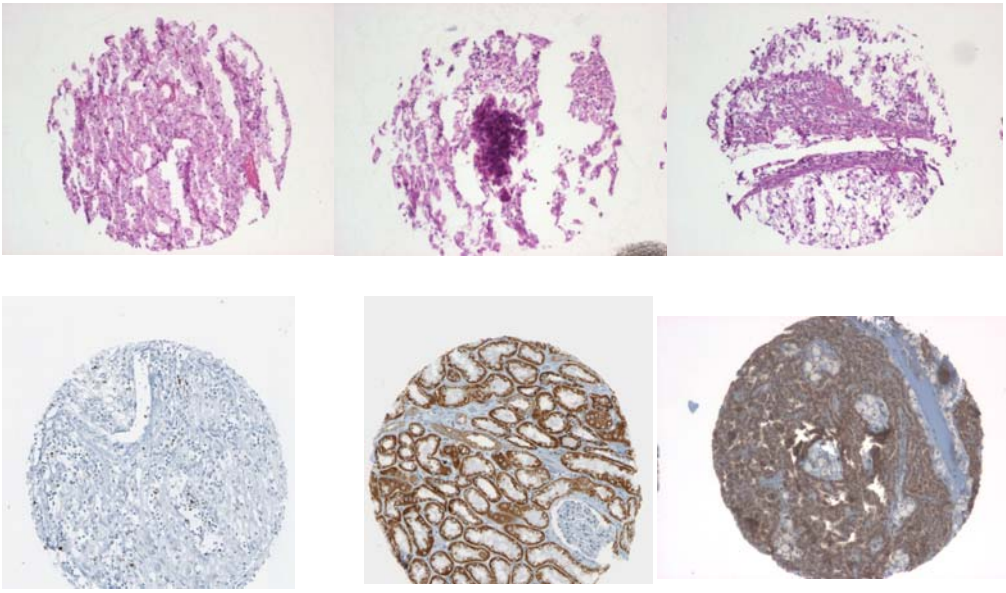
Tissue Microarray Analysis



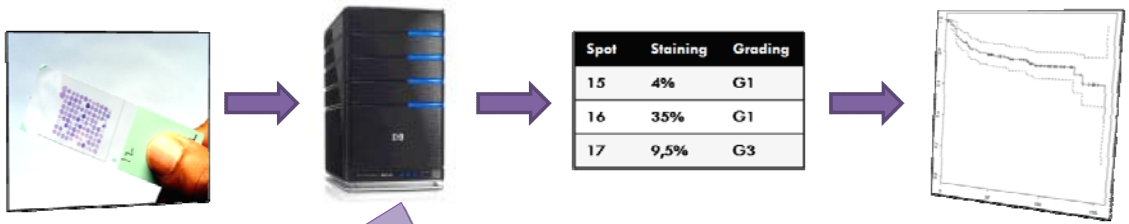
Data Analysis Problem: Variability

H & E

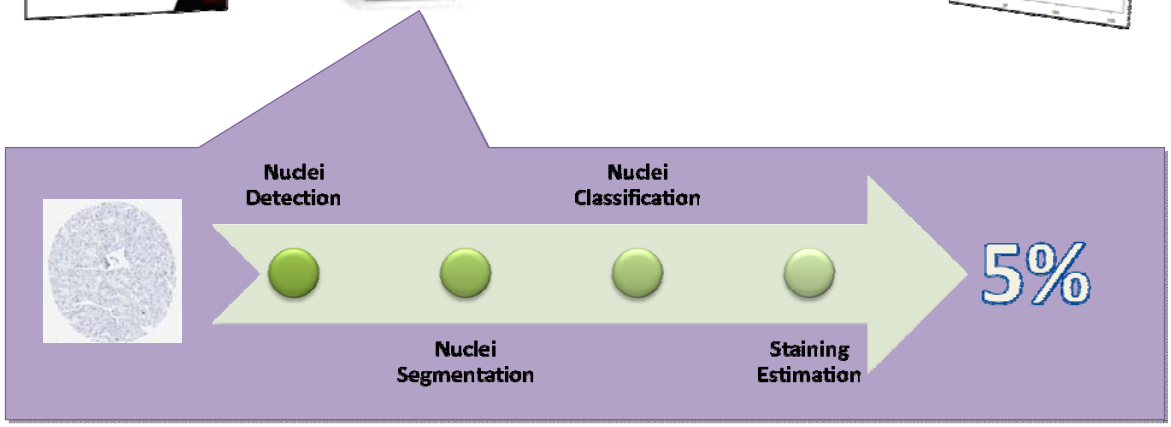
MIB-1

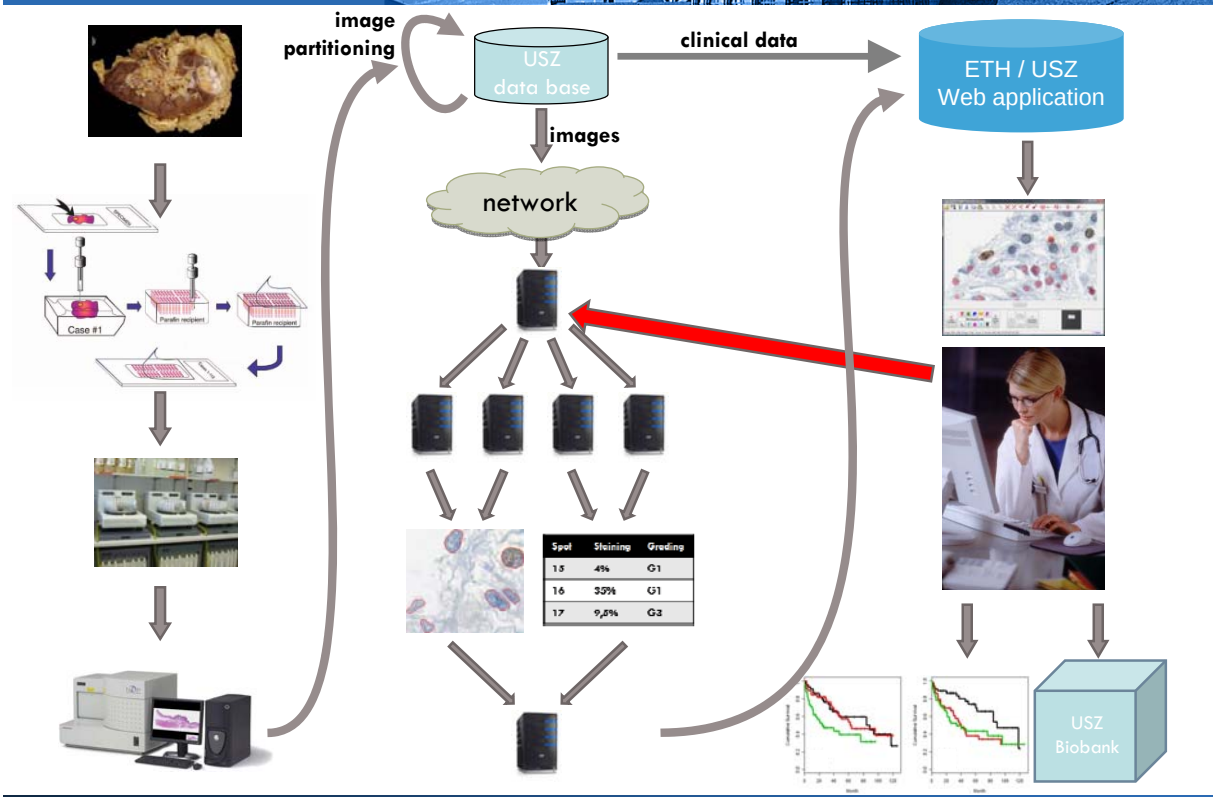


Tissue Microarray Analysis



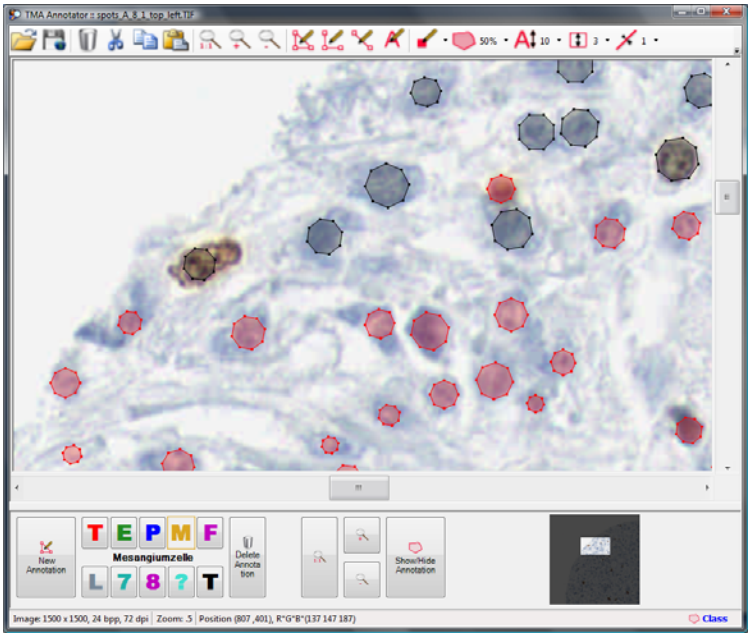
Spot	Staining	Grading
15	4%	G1
16	35%	G1
17	9,5%	G3





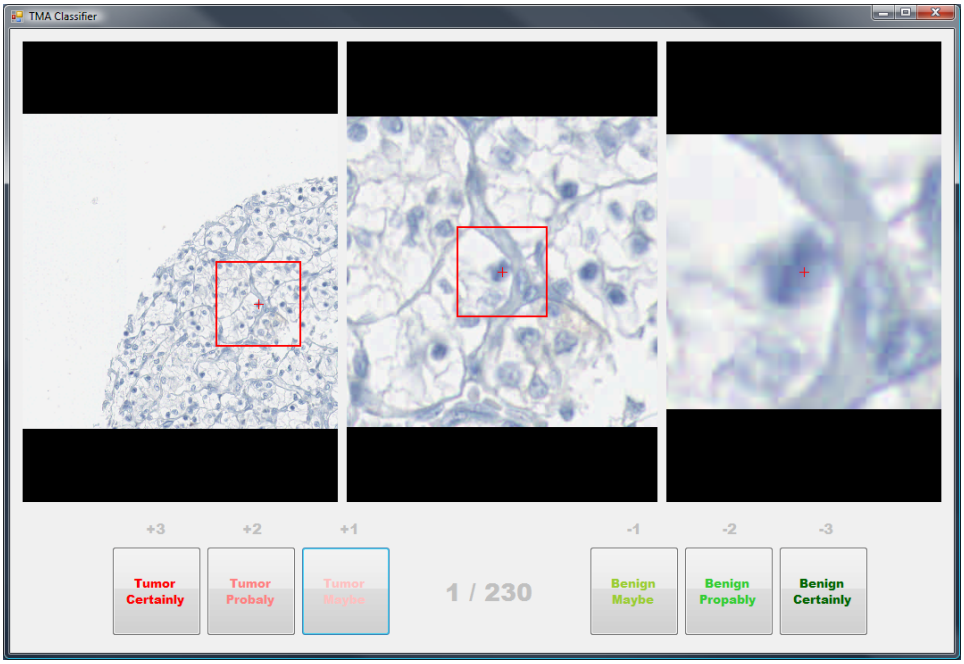
TMA Annotator

Gold-Standard by expert classifications

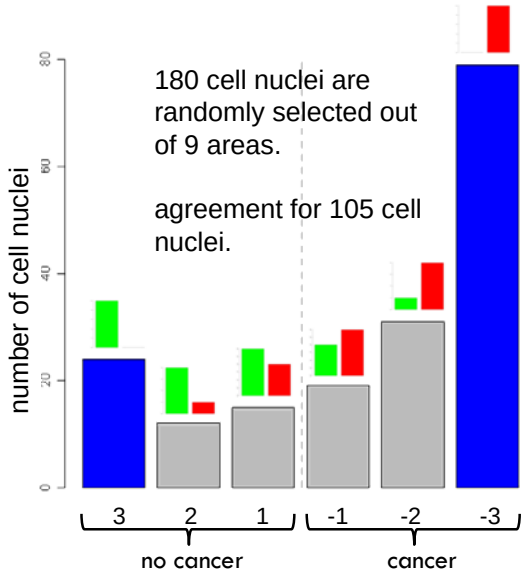


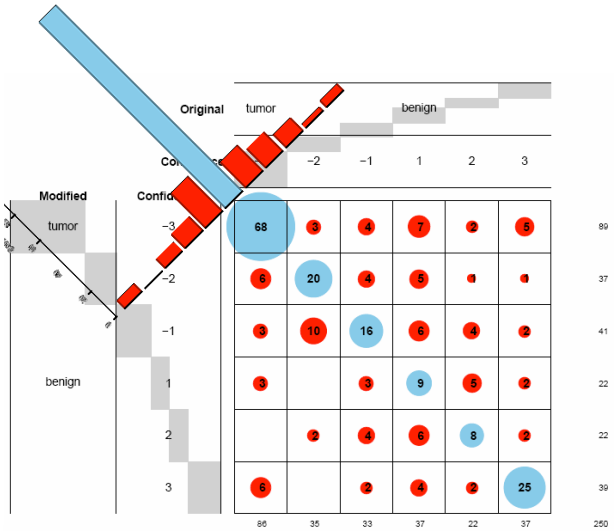
- 2 pathologists
- ~2500 nuclei
- 15% detection mismatch

TMA Classifier – Tumor / Normal Labels



Agreement among pathologists

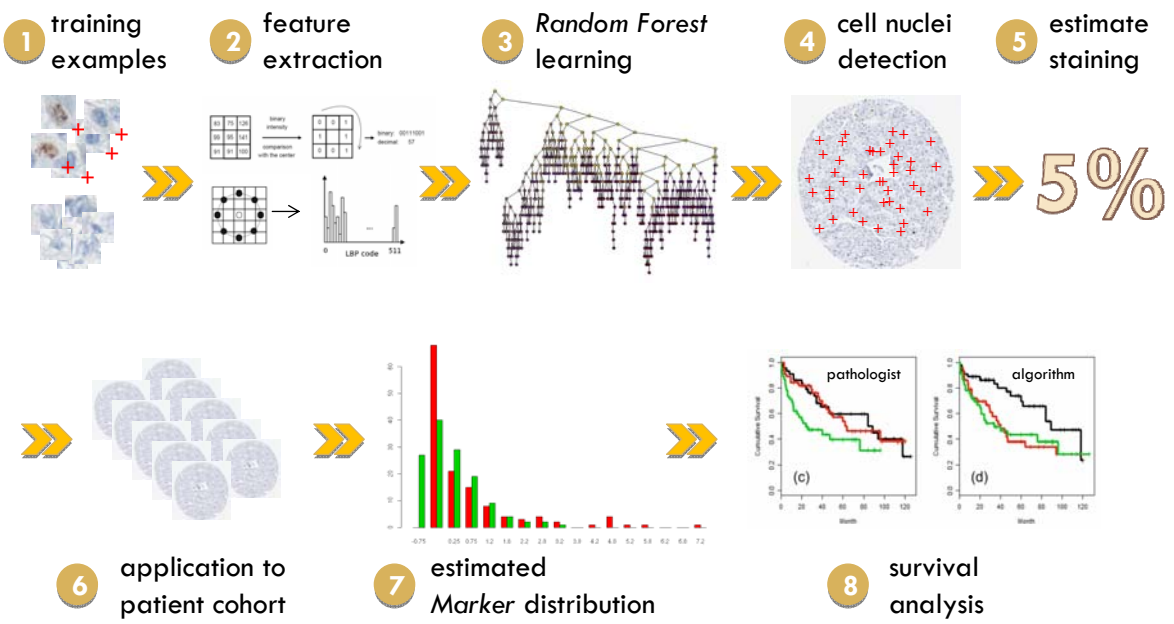




consistency among pathologists

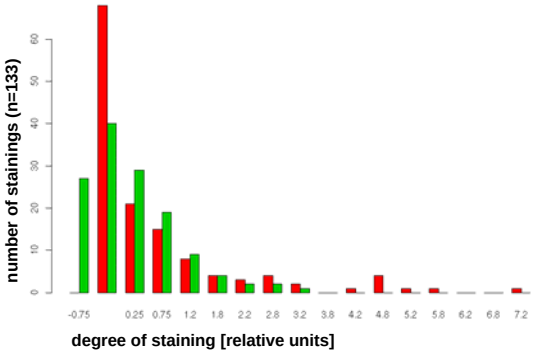
self confusion matrix of pathologists

Workflow of computational pathology

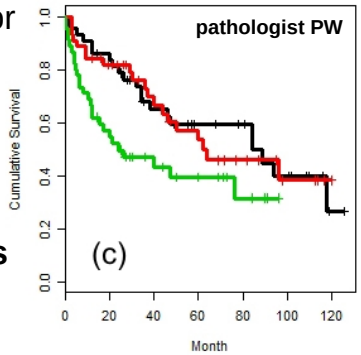


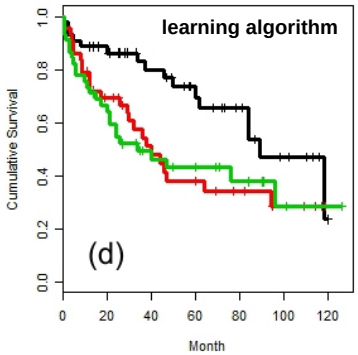
Prediction of survival

- The **learning algorithm** estimates the number of stained cancer cell nuclei more reliably than a trained pathologists for patients with good prognosis



- **Kaplan-Meier curve** for
  - 33% low risk patients
  - 33% intermediate risk
  - 33% high risk patients
- **reliability** of **diagnosis** will be increased!





Improving the quality of training data

- **Problem**: Labeling information by pathologists is often inconsistent! **No ground truth**
- **Solution**: Filter out samples which are hard to classify, i.e., denoising
- **Strategy**
  1. Compute how similar the samples appear to the trees of the random forest.
  2. Cluster the samples into groups of high similarity.
  3. Analyze label inconsistency in these groups.

Wishart-Dirichlet Cluster Process …

- … is a sequence of inner product matrices of growing size, and a random partition B of objects into k blocks.
- **Dirichlet-Multinomial prior** over partitions

$$P_n(B|\lambda,k) = \frac{k!}{(k-k_B)!}\frac{\Gamma(\lambda)\prod_{b\in B}\Gamma(n_b+\lambda/k)}{\Gamma(n+\lambda)[\Gamma(\lambda/k)]^{k_B}}.$$

- Similarity matrices are Wishart distributed

$$\mathcal{S}|B \sim \mathcal{W}_d(\Sigma_B) \text{ with } \Sigma_B = I_n \otimes \Sigma_0 + B \otimes \Sigma_1.$$

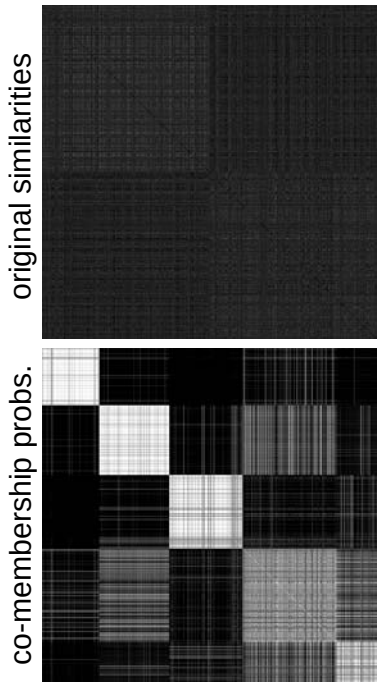
Application: Automated TMA Analysis

Dataset: 500 cancerous and 500 normal nuclei from TMA spots of renal clear cell carcinoma patients.

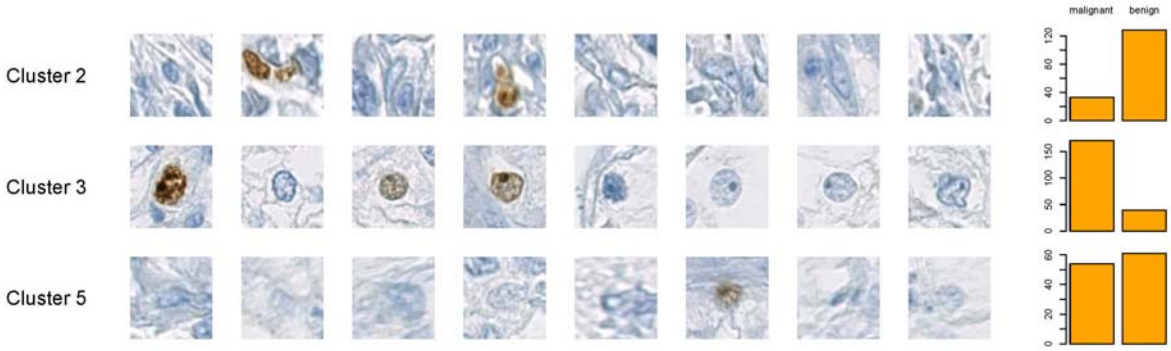
Malignant/benign classification by Random Forest yields **36% error**.

Enhancement: search for subgroups of highly discriminative nuclei. Similarity matrix is defined by proximity of nuclei in tree ensemble.

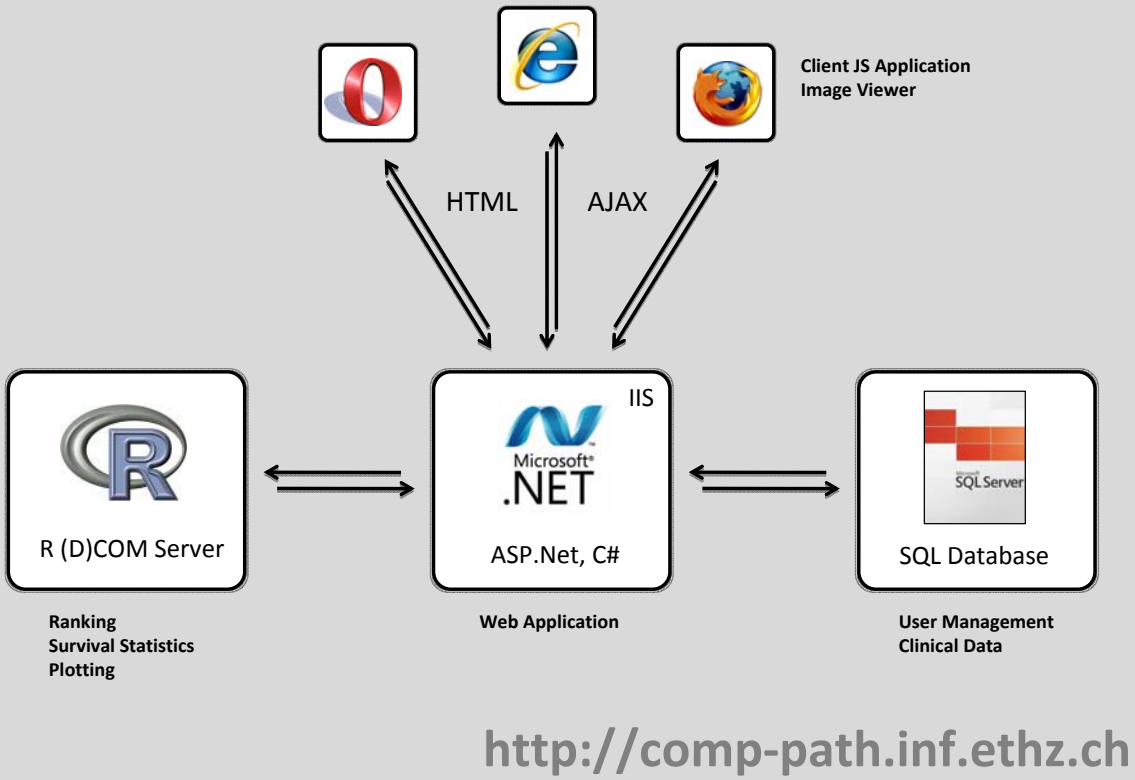
Some **negative eigenvalues** of similarities require to use a **shift invariance** method.



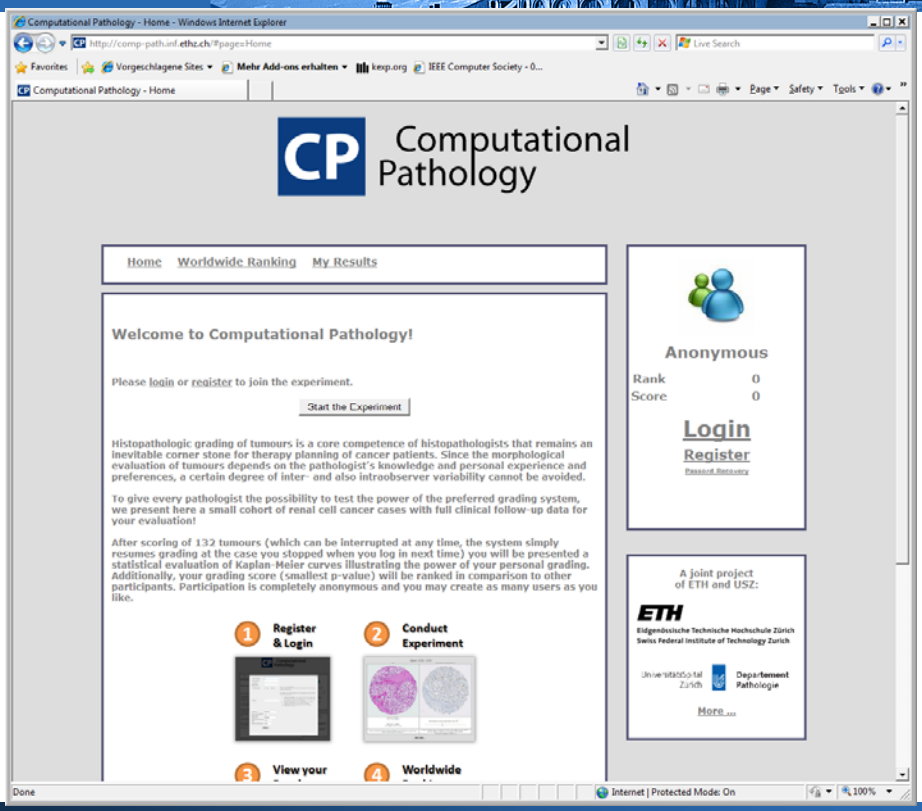
Classification Results



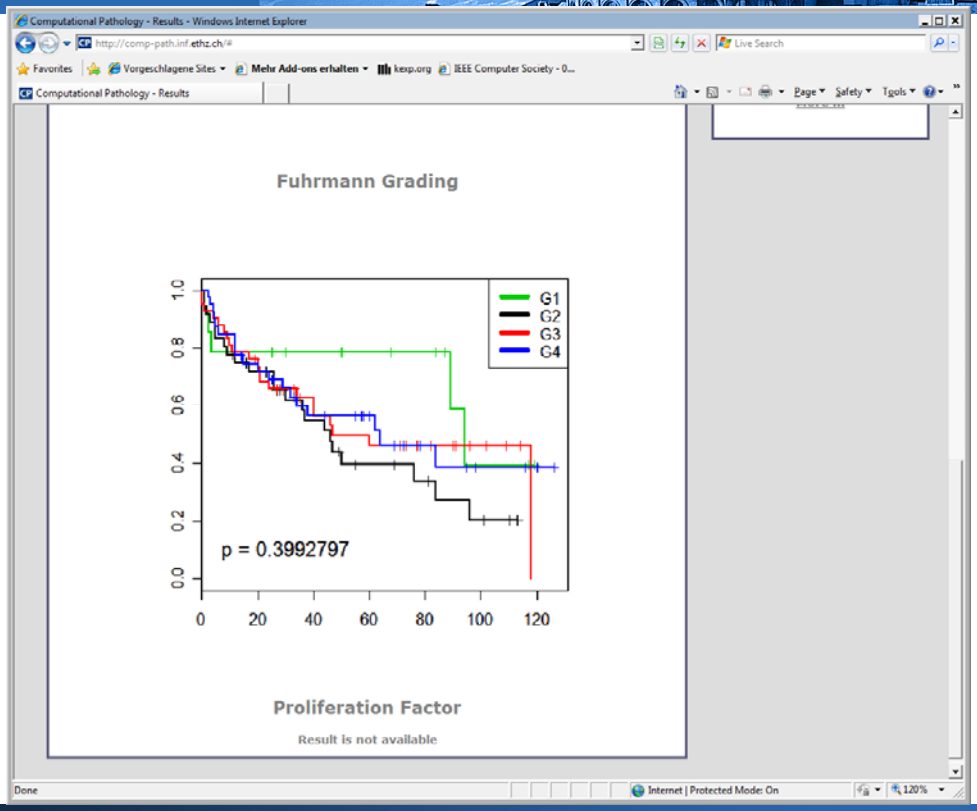
- Random Forest trained on subset of nuclei from cluster 2 and 3: **19.4% test error** => Importance of quality assessment prior to classification!
- **This strategy** overcomes a severe problems in the design of computational TMA analysis tools.

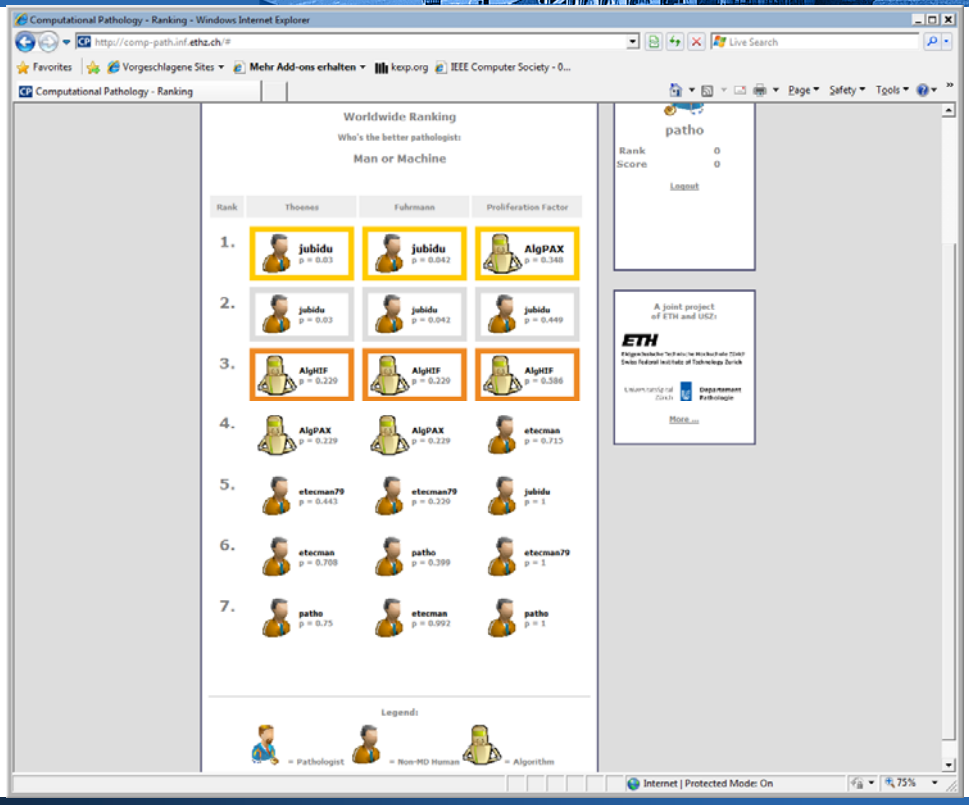


http://comp-path.inf.ethz.ch




Computational Pathology - Exercise - Windows Internet Explorer
Back
Spot 133 / 133
H&E
Proliferation Factor (MIB-1)
Thoenes Grading:
G1 G2 G3
Fuhrmann Grading:
G1 G2 G3 G4
Percentage of stained abnormal nuclei: 14
Please grade the tissue based on Thoenes and Fuhrmann respectively. Click on the image to enlarge it.
Please estimate the percentage of tumor cells which express MIB-1. Use the slider or enter the percentage in the textbox. Click on the image to enlarge it.
Next Spot
Back


Computational Pathology - Results - Windows Internet Explorer
Fuhrmann Grading
G1
G2
G3
G4
p = 0.3992797
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
100
120
Proliferation Factor
Result is not available



Future challenges in Computer Science

- **Machine Learning** has matured to an essential **method of computer science**, to generate complex statistical models for **Computational Biology**, **Visual Computing** but also other core areas of informatics.
- **Statistical modeling** and **algorithmics** as core problems of computer science can be excellently studied in the area of ***Pattern Recognition***.
- We will provide society with the **intelligent technology of the 21st century!**