ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Model validation in clustering by information theory

Joachim M. Buhmann

Computer Science Department, ETH Zurich

Clustering: Science or Art?

- What is science? **Hypothesis driven reasoning!**
- **Hypothesis**: Validation of clustering models is the fundamental challenge, characterizing given or invented models is of secondary nature!

$\Rightarrow$Algorithmic search for clustering models requires garbage collection; noise tolerant and expressive models are preferred over brittle, simplistic ones!

$\Rightarrow$Information theory allows us to measure the information content of clusterings!

Clustering is the partitioning of objects into groups!

- 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.,

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

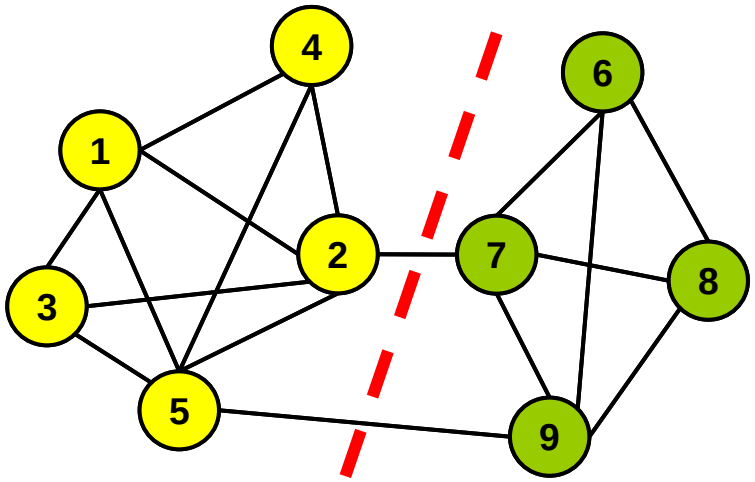
- Hypothesis class $c \in \mathcal{C} \equiv \{\text{partitions of data}\}$

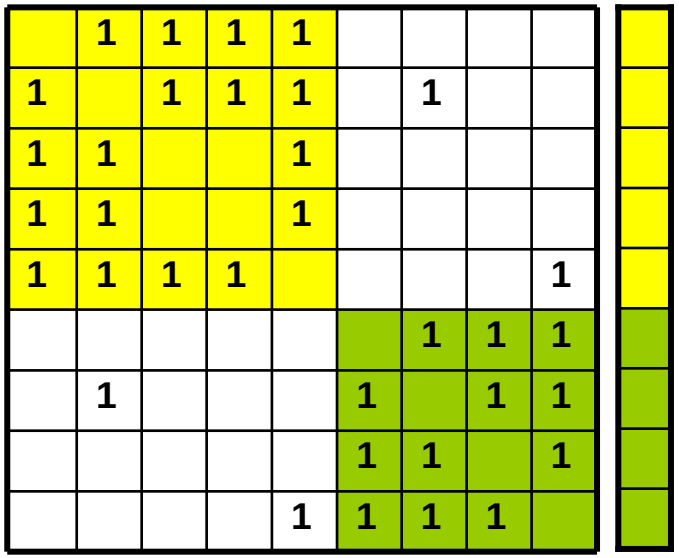
Design decisions for cluster analysis

- Topology of solution space $\mathcal{O}$
- Metric of measurement space $\mathcal{X}$
- Criterion to measure the quality of clusterings
- Algorithm α to search for “good” clusterings
- Validation method to test clusterings

Ex.: Graph Cut - Clustering in two groups

- **Graph** representation: **vertices** denote **objects**
  edges express **similarities**
- **Hypothesis class:** all **cuts** of a graph





Generative vs. discriminative learning - complexity consideration for graph cut

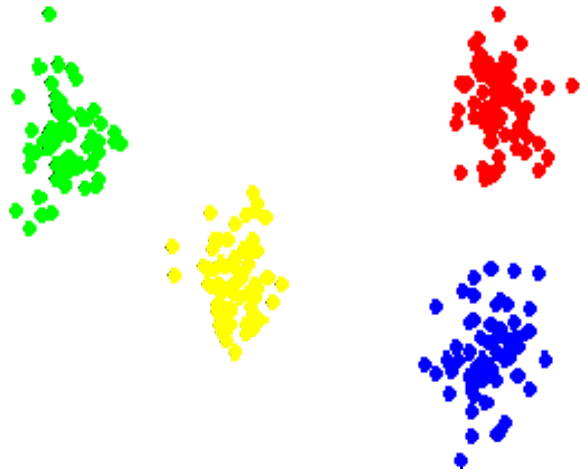
- Cardinality of hypothesis class for graph cut is 2^n
- The problem space is represented by the set of all adjacency matrices; its cardinality is 2^{n^2}

=> Identifying the problem, i.e., estimating 2^{n^2} different probabilities requires much more information than finding a set of approximate solutions.

What is the “right” clustering model?

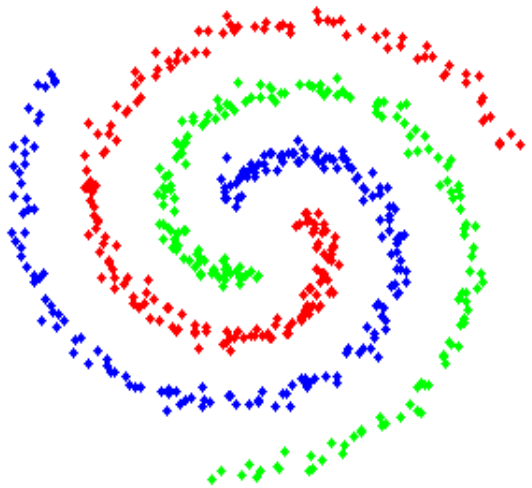
Compactness criterion

- *k*-Means, *k*-Median Clustering
- Pairwise Clustering, AvAssoc
- Correlation Clustering
- Max-Cut, Average Cut
- Normalized Cut



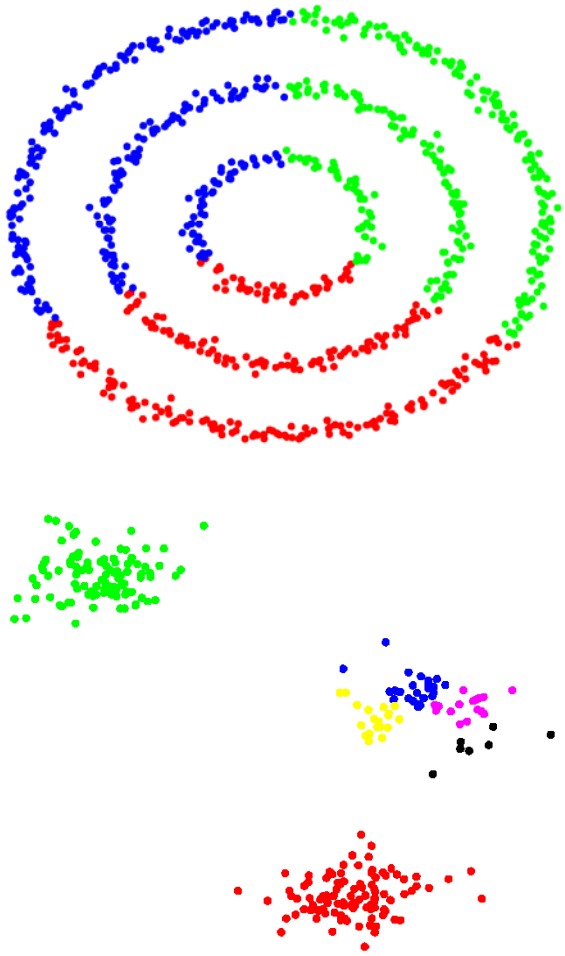
Connectedness criterion

- Single Linkage
- Path Based Clustering



Design problems in clustering: validation

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



Approximate optimization and information theory

- **Problem**: Noise in data renders solutions of optimization problems unstable.

=> robustness = generalization is required

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

Order relation of clusterings

- **algorithm** α selects "statistically optimal" clusterings

$$\begin{aligned} \alpha : \mathcal{O} \times \mathcal{X} \times \mathbb{R}_{\geq 0} &\rightarrow \mathfrak{P}(\mathcal{C}) \\ (o, x, \gamma) &\mapsto \alpha(o, \mathbf{X}, \gamma) = \mathcal{C}_\gamma \subset \mathcal{C} \end{aligned}$$

- **"Stratification"**: α returns the global minimizer for γ=0 and all clusterings $\mathcal{C}_\infty = \mathcal{C}$ for $\gamma = \infty$ with

$$\forall \gamma \leq \tilde{\gamma} \Rightarrow \mathcal{C}_\gamma \subset \mathcal{C}_{\tilde{\gamma}}$$

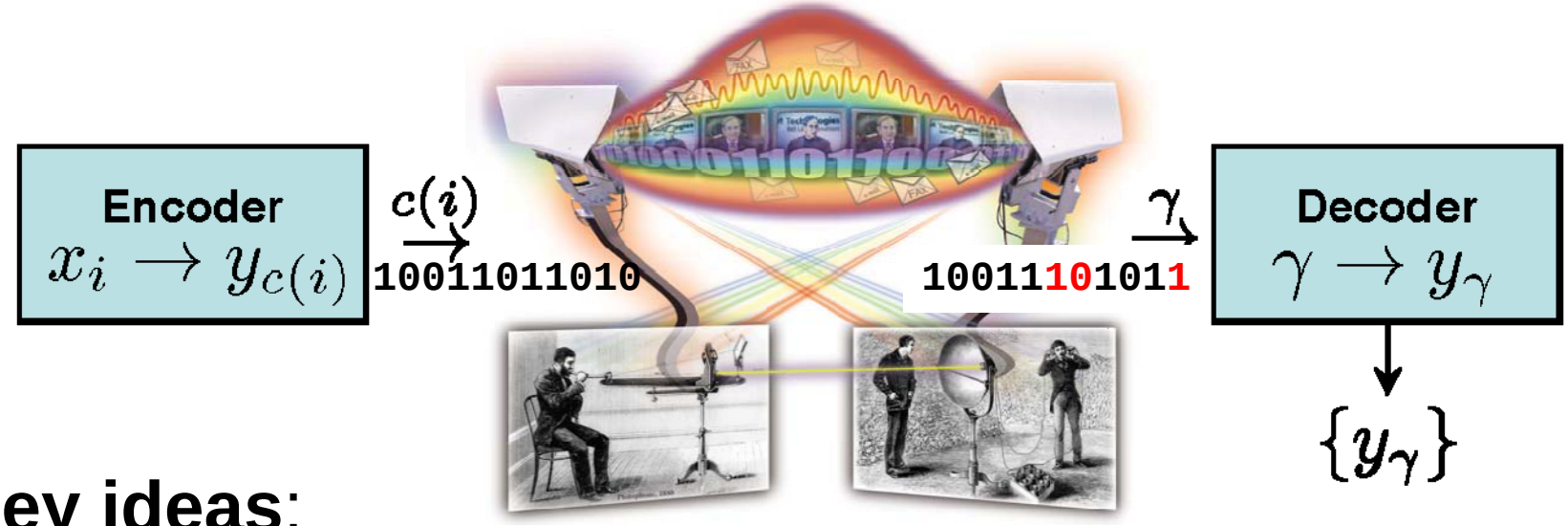
Empirical Risk Approximation

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

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

- 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**

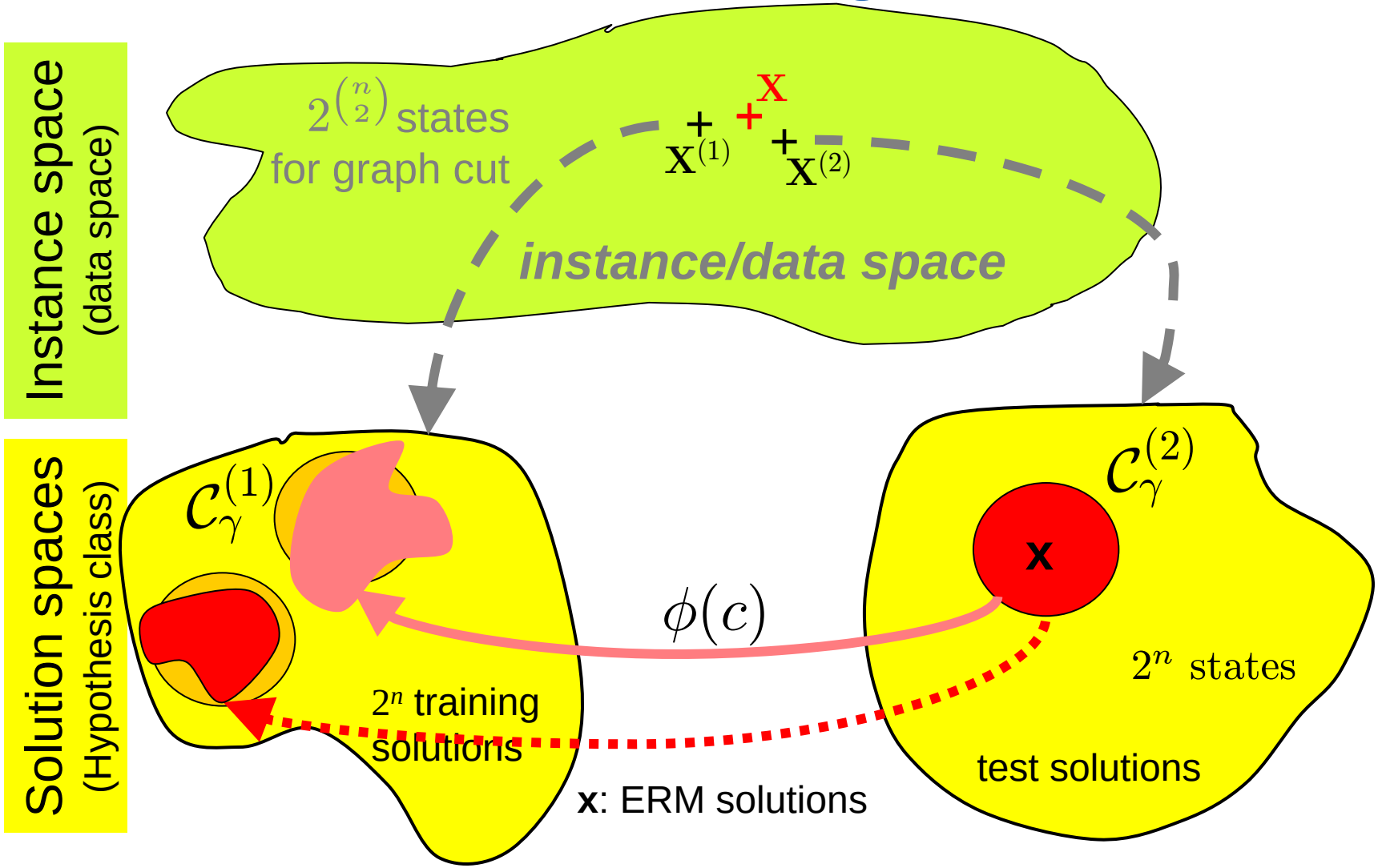
Shannon's information theory



Key ideas:

1. **Quantization**: Coding by random code vectors
2. **Minimize Hamming distance** for decoding
3. **Asymptotic error rate** of a channel => **mutual information**

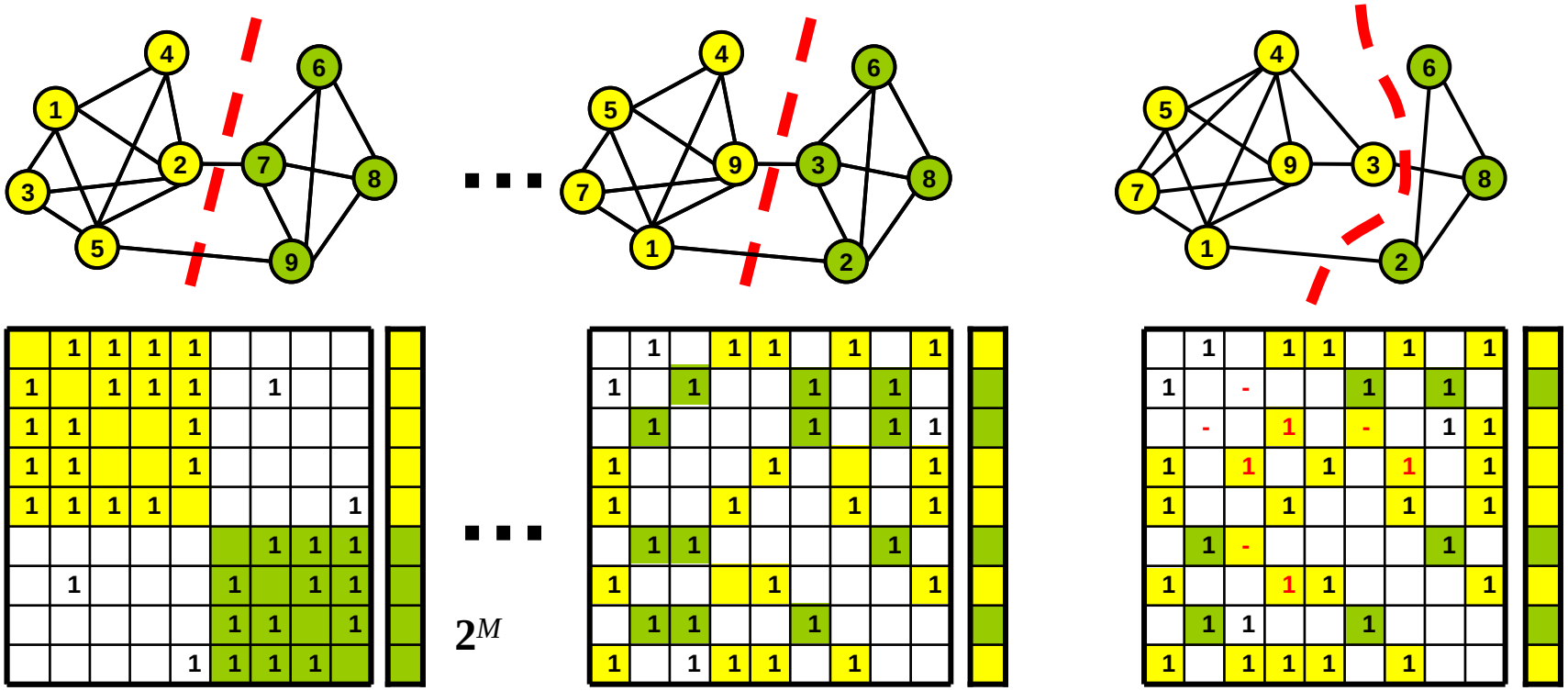
Generalization in clustering



Code problem generation for Graph Cut

graph cut code problems

graph cut test problem



Coding with Graph Cut approximation sets

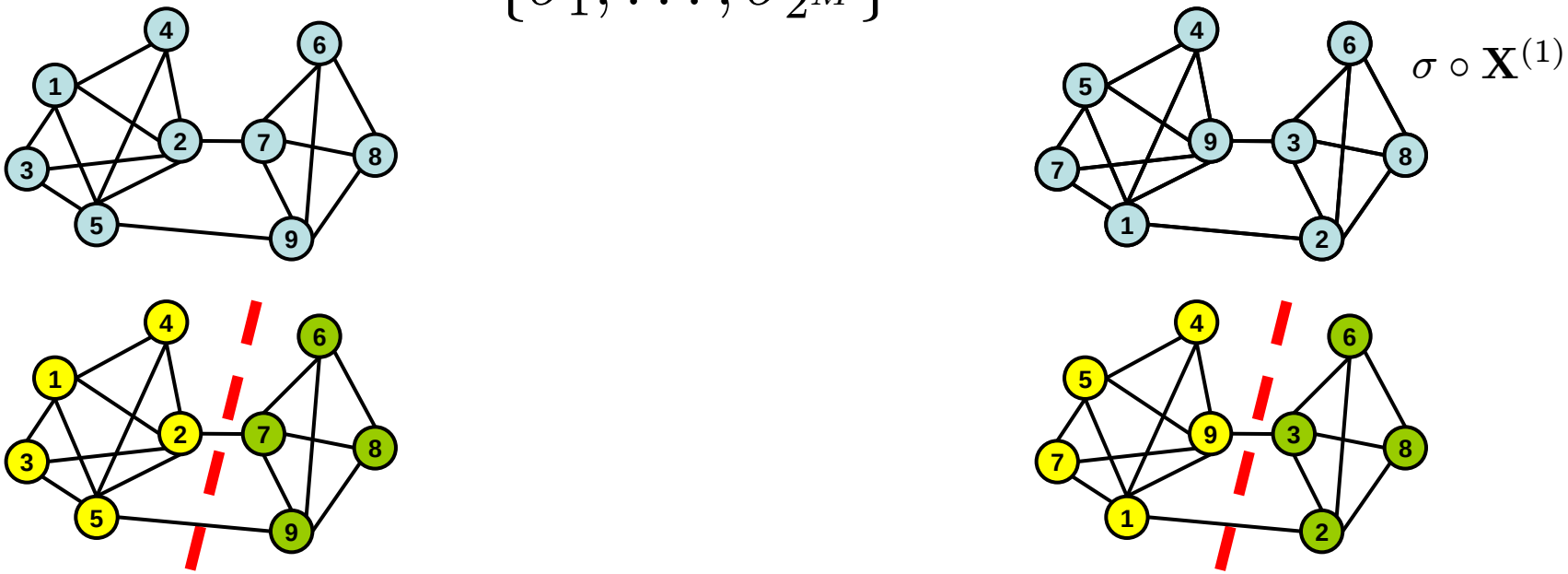
define a set of code problems

problem generator PG

$R(\cdot, \mathbf{X}^{(1)})$ $R(\cdot, \mathbf{X}^{(1)})$

sender → receiver

$\{\sigma_1, \ldots, \sigma_{2^M}\}$



Communication by approximation sets

estimate the coding error

problem generator

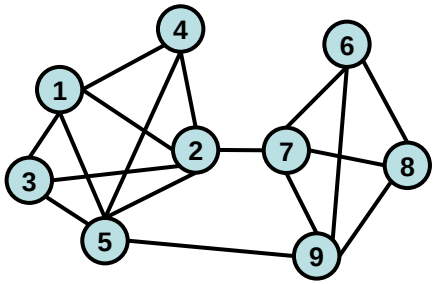
σ_s

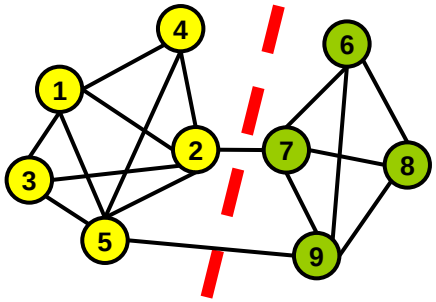
sender

$R(\cdot, \sigma_s \circ \mathbf{X}^{(2)}), \text{s.t.}$
$\mathbf{X}^{(1)}, \mathbf{X}^{(2)} \sim P(\mathbf{X})$

receiver

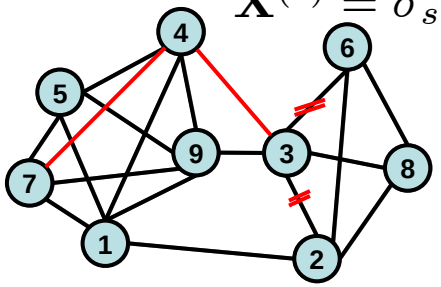
σ^*

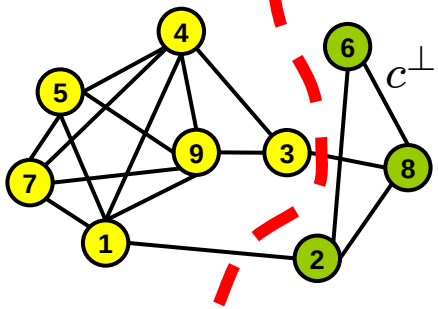




1. Sender sends a permutation index σ_s to problem generator.
2. Problem generator sends a new problem with permuted indices to receiver without revealing σ_s.
3. Receiver identifies the permutation σ^* by comparing approximation sets.

$\tilde{\mathbf{X}}^{(2)} = \sigma_s \circ \mathbf{X}^{(2)}$





$c^{\perp}(\tilde{\mathbf{X}}^{(2)})$

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 $\phi : \mathcal{C}(\mathbf{X}^{(2)}) \rightarrow \mathcal{C}(\mathbf{X}^{(1)})$
- **Decoding**

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

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

Error Analysis and the Tradeoff in Approximation

- 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** that maximizes information transfer, i.e., high precision and high noise robustness

Error Events and Approximation Capacity

- Sender selects approximation set $\mathcal{C}_\gamma(\sigma_s \circ \mathbf{X}^{(1)})$
- **Error sets**

$$\overline{\mathcal{E}_s} = \left(\mathcal{C}(\mathbf{X}^{(1)}) \setminus \mathcal{C}_\gamma(\sigma_s \circ \mathbf{X}^{(1)})\right) \cap \left(\phi \circ \mathcal{C}_\gamma(\sigma_s \circ \mathbf{X}^{(2)})\right)$$

$$\mathcal{E}_j = \mathcal{C}_\gamma(\sigma_j \circ \mathbf{X}^{(1)}) \cap \left(\phi \circ \mathcal{C}_\gamma(\sigma_s \circ \mathbf{X}^{(2)})\right)$$

$$\text{for } 1 \leq j \leq 2^M,\ j \neq s$$

- **Error events** :

$$|\overline{\mathcal{E}_s}| \geq |\mathcal{E}_s| \text{ or } |\mathcal{E}_j| \geq |\mathcal{E}_s| \text{ for } 1 \leq j \leq 2^M,\ j \neq s$$

Error Probability

- Conditional error

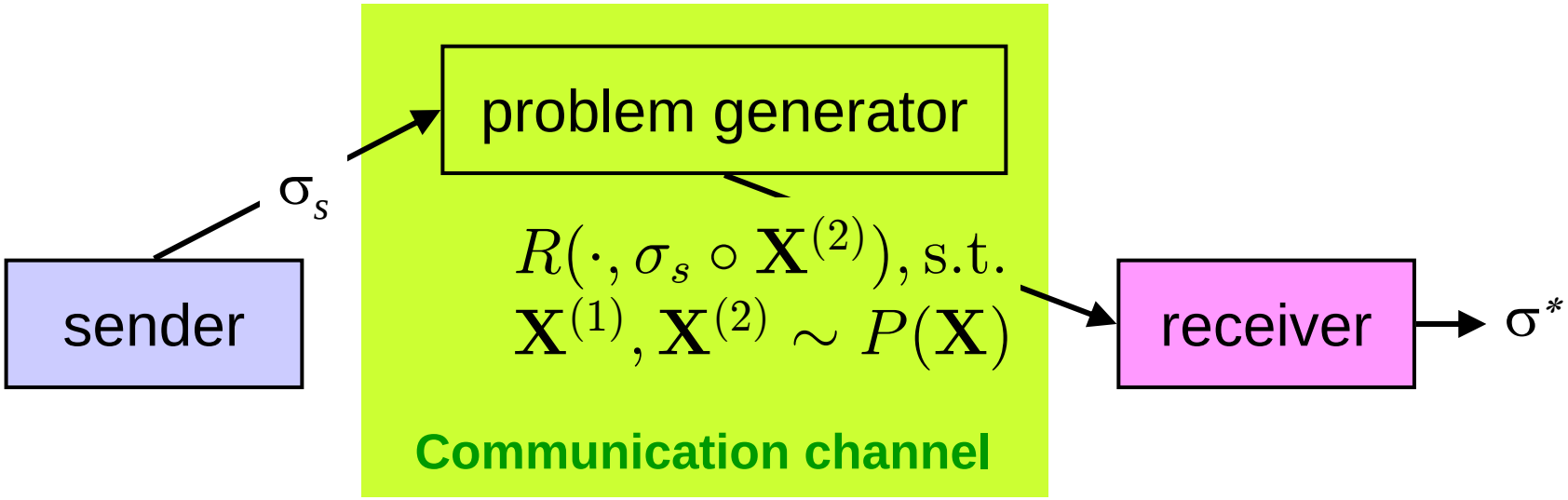
$$
\begin{aligned}
P(\text{error}|\sigma_s) &= P(\max_{i\neq s}|\mathcal{E}_j| > |\mathcal{E}_s| \vee |\overline{\mathcal{E}_s}| > |\mathcal{E}_s|\big|\sigma_s) \\
&\leq P\left(|\overline{\mathcal{E}_s}| > |\mathcal{E}_s|\big|\sigma_s\right) + \sum_{j\neq s} P\left(|\mathcal{E}_j| > |\mathcal{E}_s|\big|\sigma_s\right) \\
&\leq \epsilon + 2^M P\left(|\mathcal{E}_{\neq s}| > |\mathcal{E}_s|\big|\sigma_s\right)
\end{aligned}
$$

- Condition of **vanishing total error**

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

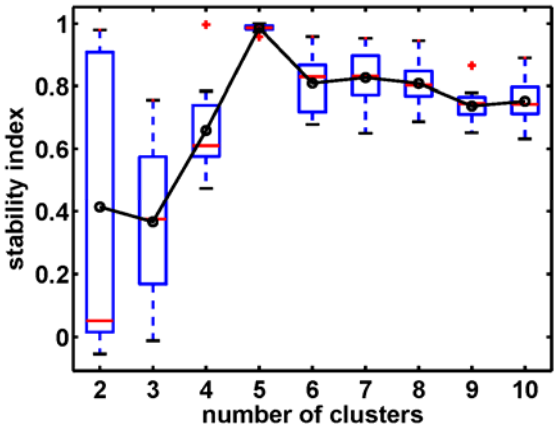
mutual information

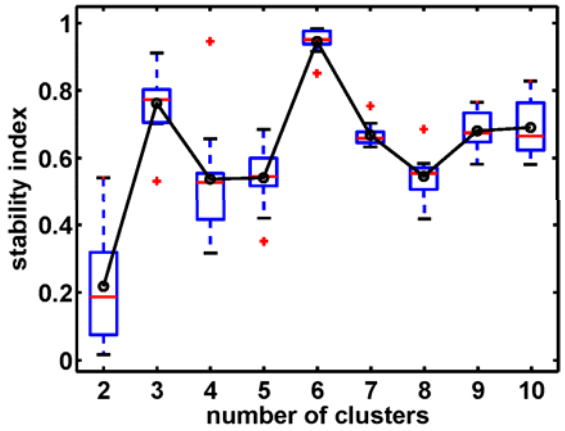
Model Selection by Maximization of Approximation Capacity

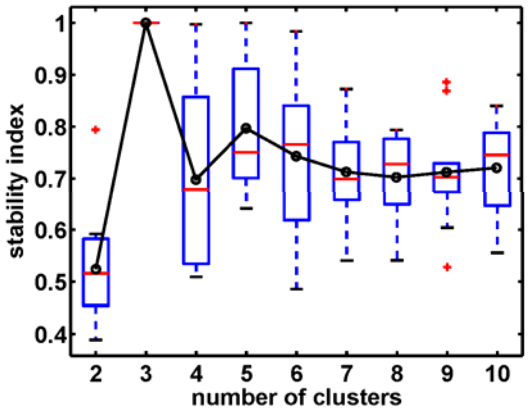


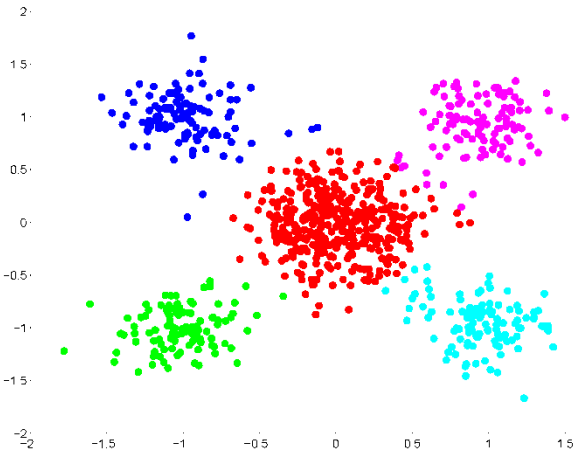
- Optimize the communication channel w.r.t. approximation quality γ, topology and metric of solution space, cost function $R(.,.)$, transfer function ϕ

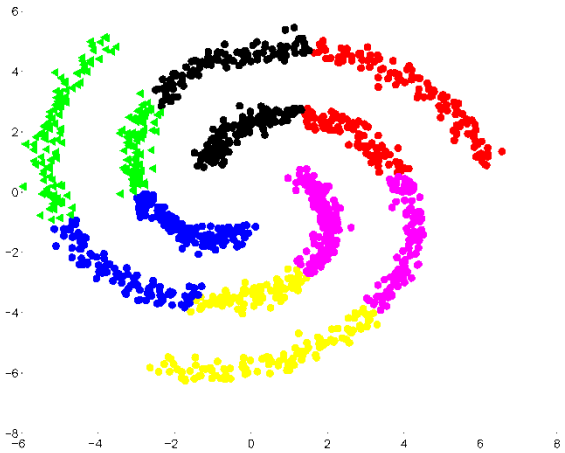
Results on Toy Data

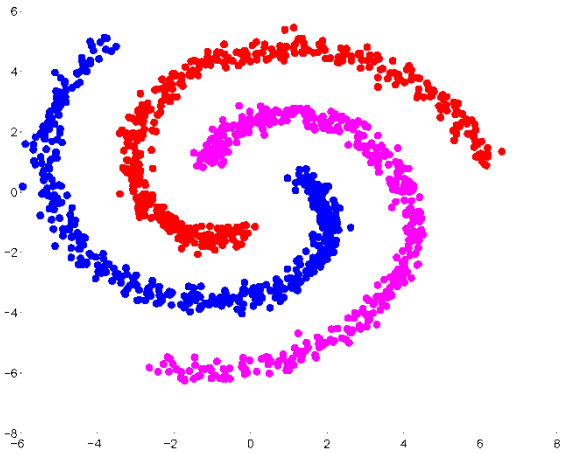












Clustering of Microarray Data

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

Task: Find groups of different Leukemia tumour samples

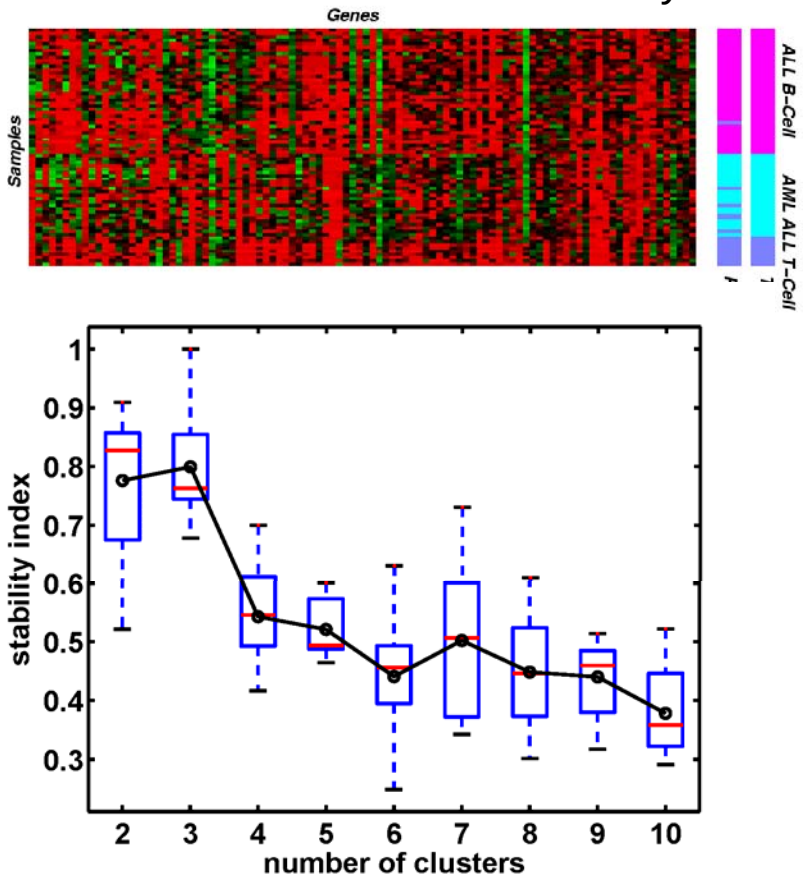
(two- and three class classifications are known).

Problem: Number of groups is unknown a priori.

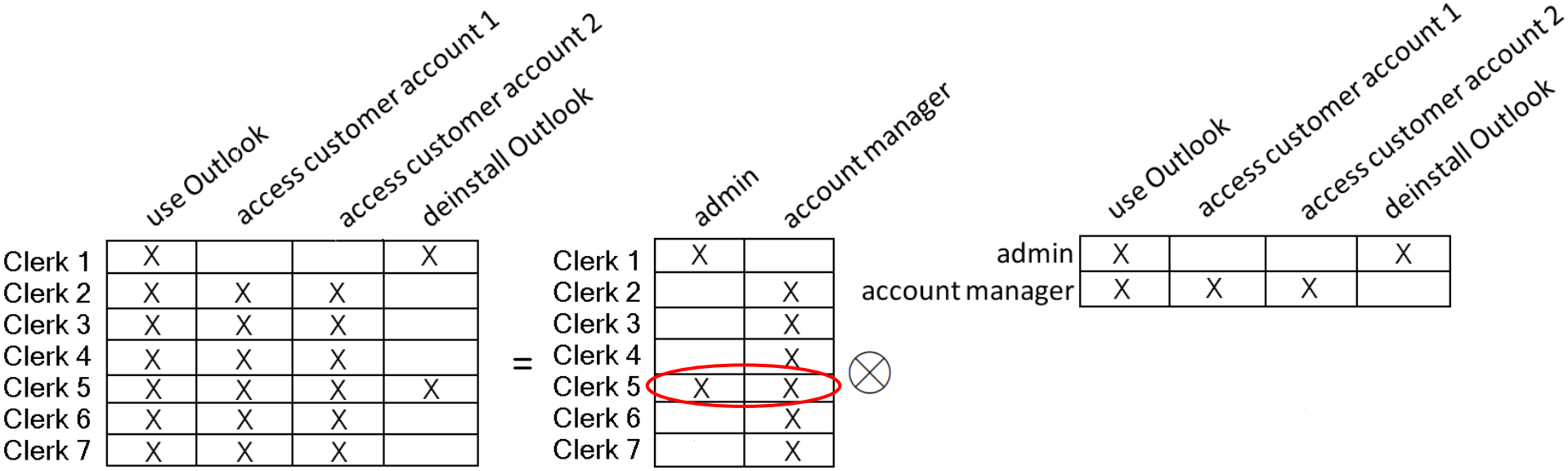
Via Stability with *k*-means: Estimated number of groups is 3.

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

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



The Role Mining Problem



Direct user permission assignment

Role-Based Acces Control (RBAC)

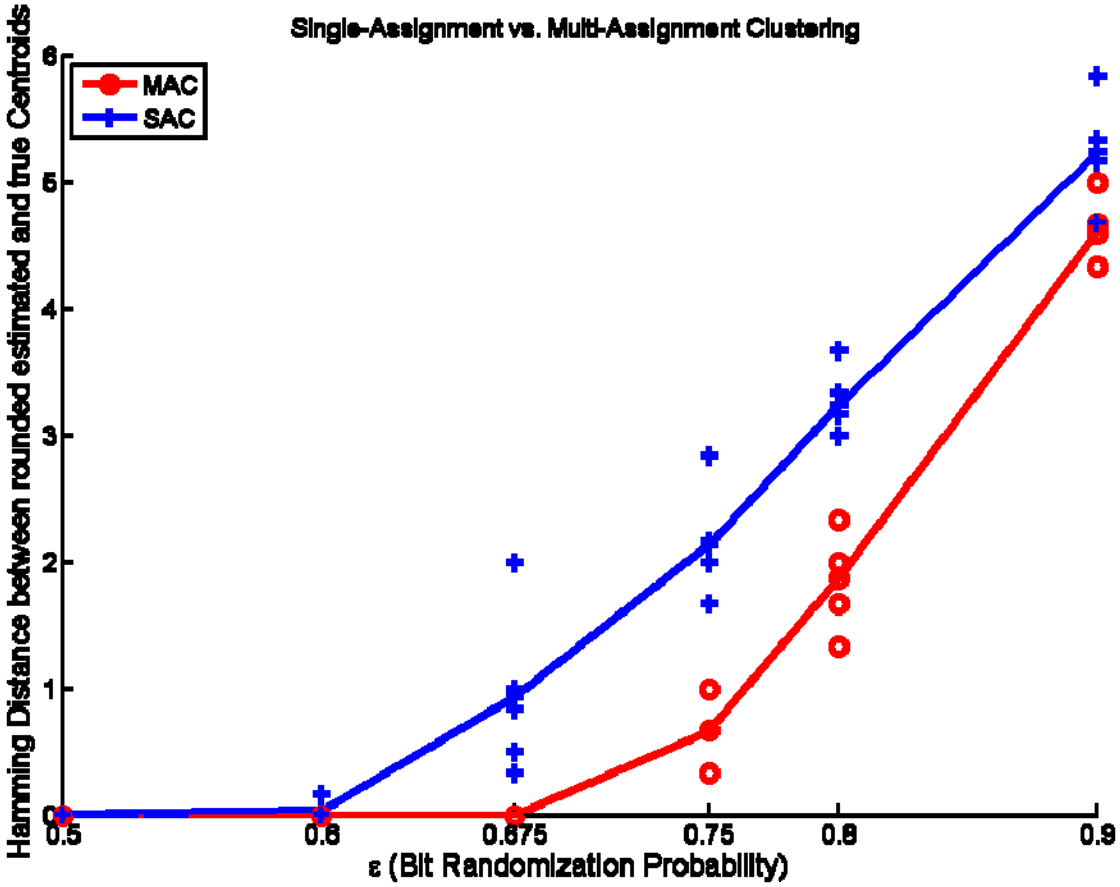
Direct user permission assignment

Role Mining

Role-Based Access Control (RBAC)

Stability of multi-label clustering

- Two roles: n=12000, K=3
- 2/3 of users have one role, 1/3 has two roles.
- SAC is clearly inferior to MAC



Conclusion

- **Quantization**: Noise quantizes solution spaces that yields symbols.
- These symbols can be used for **coding**!
- Optimal error free coding scheme determines **approximation capacity** of a model class.

$\Rightarrow$bounds for robust optimization.

$\Rightarrow$**Quantization** of hypothesis class measures **structure specific information** in data.

Philosophical speculations

- We experience a **paradigm shift from model driven reasoning to algorithm dominated reasoning** (Bernard Chazelle "The Algorithm: Idiom of Modern Science")

=> **model validation** more essential than modeling since modeling can be algorithmically formulated as exploration of model space.

- *Ceterum censeo*: The coupling of **statistical complexity** and **algorithmic complexity** should be reconsidered in the light of information theory.