Semi-supervised Learning with Random Walks on Graphs

Michal Valko (University of Pittsburgh)
Branislav Kveton (IRSC), Matthai Philipose (IRS)

Týždeň absolventov Matfyzu 2009, 18. decembra 2009 12:30

Main Ideas

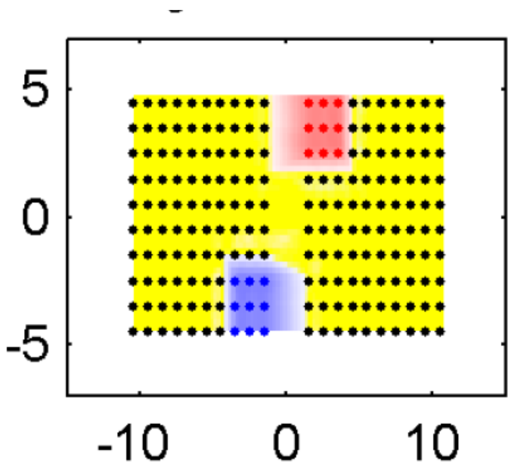
- **Goal**: Adaptation to (structured) patterns with minimal human feedback (labels)
  - Most of data around is unlabeled
  - Labeling is expensive
- **Solution**: Semi – Supervised learning
  - Labeled examples are provided in the beginning
    - Provide initial bias
  - Unlabeled examples come as available
- **Approach**: Graph – based inference with max-margin learning

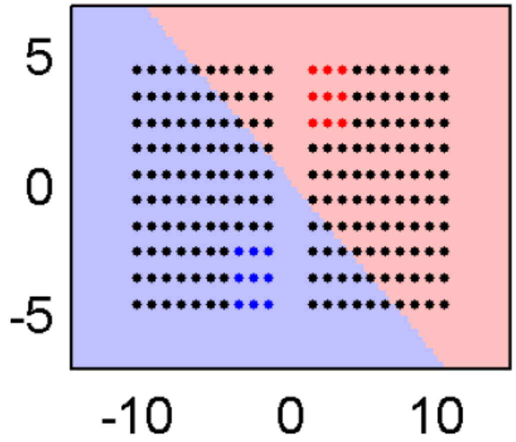
Semi-supervised learning

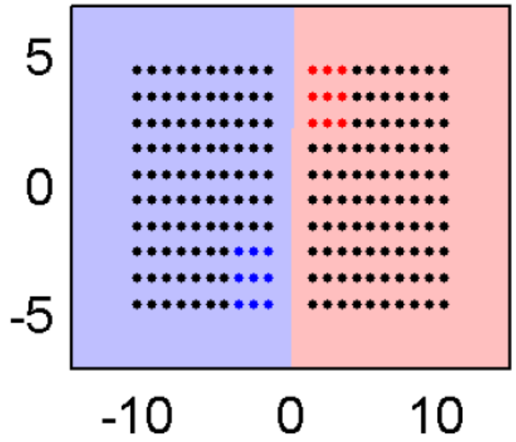
Data

Supervised

Semi-Supervised

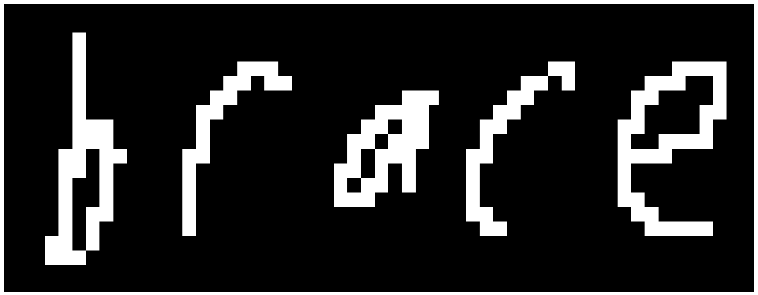






Structured data

INPUT **x**



LABEL **y**

brace

With sequences we have dependencies between $\mathbf{y}_i$ and $\mathbf{y}_{i+1}$

INPUT **x**

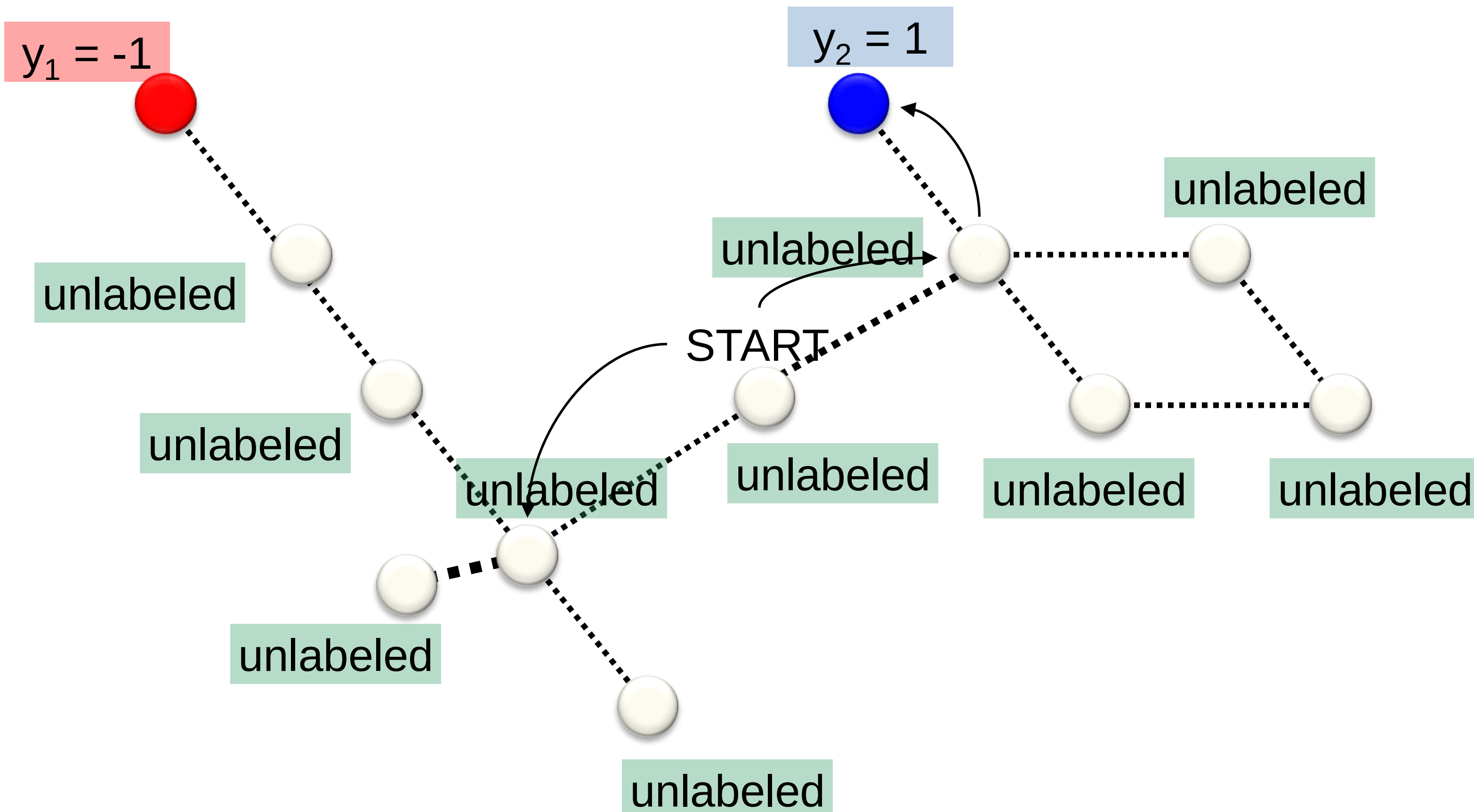
LABEL **y**

matryoshka, matryoshka … matryoshka matryoshka tylenol, tylenol, tylenol

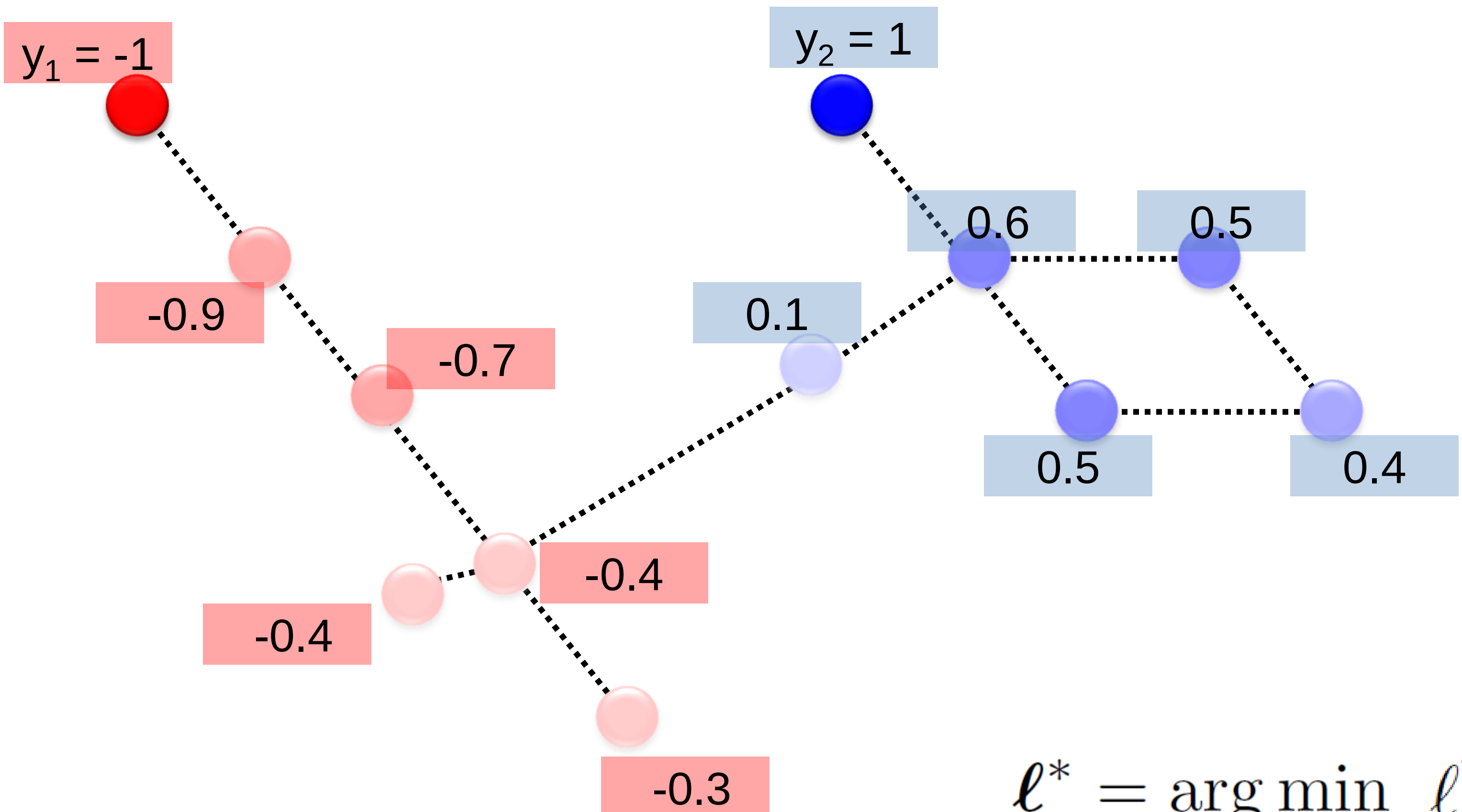
Overview

- **Graph – based inference**
- Offline Learning
- Online Learning
  - Face Recognition
- Max – Margin graph cuts
- Structured Learning
  - Handwriting Recognition
- Online Learning
  - Object Recognition

Graph-based Semi-Supervised Learning



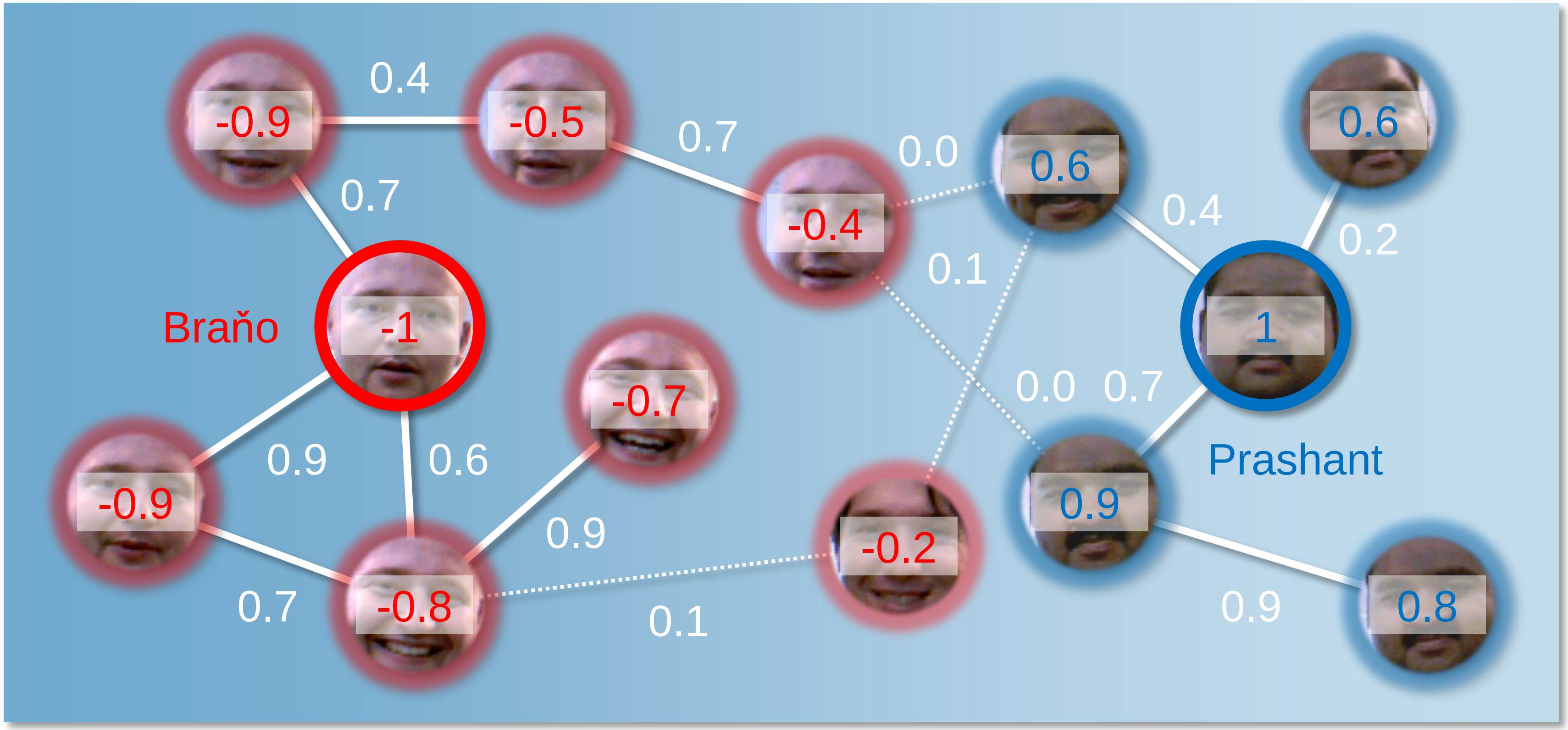
Graph-based Semi-Supervised Learning



$$\boldsymbol{\ell}^* = \arg\min_{\boldsymbol{\ell}} \ \ell^{\mathsf{T}} L \ell$$

Harmonic Function Solution (HFS)

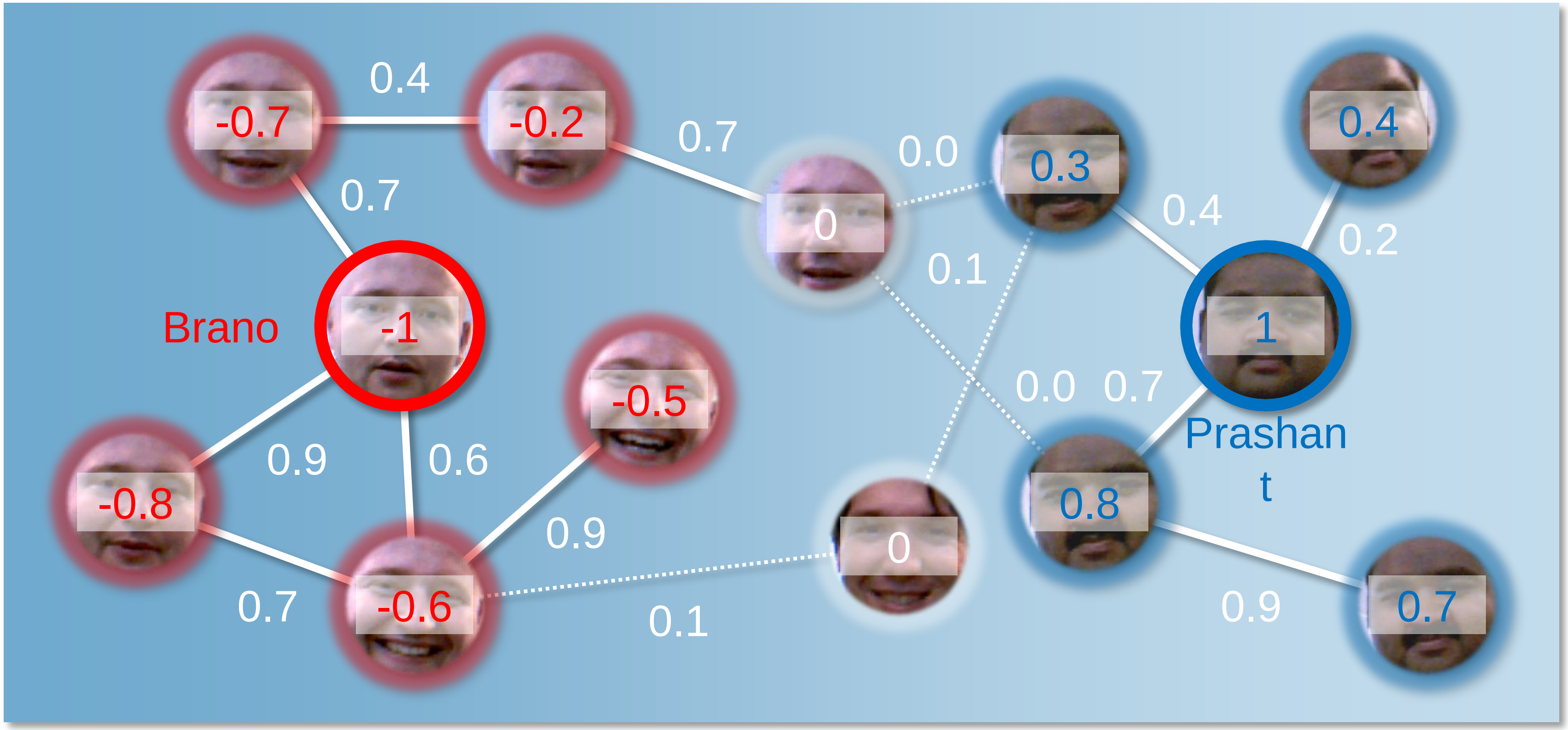
- Labels of unlabeled vertices are inferred using the harmonic function solution



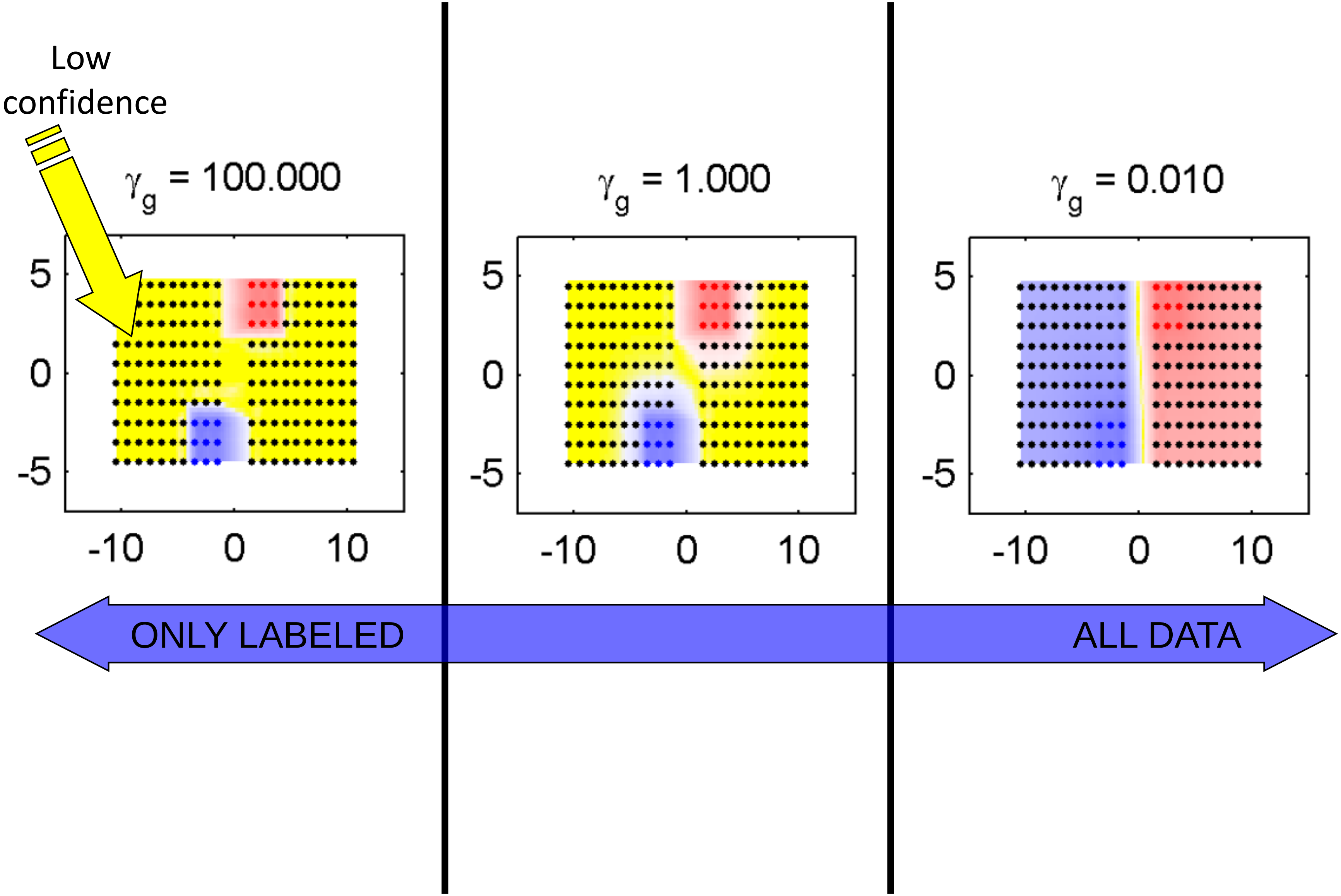
Regularized HFS

$$\boldsymbol{\ell}^* = \arg\min_{\boldsymbol{\ell}} \boldsymbol{\ell}^\mathsf{T}(\boxed{\gamma_g} I + L)\boldsymbol{\ell}$$

$$\text{s.t. } \ell_i = y_i \text{ for all } i \in l;$$



Regularization



Online HFS

Inputs: an example x_t, a data adjacency graph W

What is wrong with this algorithm?

Algorithm:

Add x_t to the graph W and compute the Laplacian L

Infer labels on the graph:

$$\min_{\lambda \in \Re^N} \lambda^{\mathrm{T}} (L + \gamma_g I) \lambda \quad \text{s.t. } \lambda_i = y_i \text{ for all } i \in l$$

Predict $\hat{y}_t = \lambda_t$

O(t)

O(t^3)

Outputs: a prediction $\hat{y}_t$, an updated data adjacency graph W

Online HFS

Inputs: an example x_t, a data adjacency graph W

Algorithm:

If the graph W has more than M vertices, quantize it

Add x_t to the graph W and compute the Laplacian L

Infer labels on the graph:

$$\min_{\lambda \in \Re^N} \lambda^{\mathrm{T}} (L + \gamma_g I)\lambda \quad \text{s.t. } \lambda_i = y_i \text{ for all } i \in l$$

Predict $\hat{y}_t = \lambda_t$

O(M)

O(M^3)

Outputs: a prediction $\hat{y}_t$, an updated data adjacency graph W

Quantizing Data Adjacency Graphs

- Preferably a strategy that minimizes the error:

$$\left\|\lambda_u - \lambda'_u\right\| = \left\|(L_{uu} + \gamma_g I)^{-1} W_{ul} \lambda_l - (L'_{uu} + \gamma_g I)^{-1} W'_{ul} \lambda_l\right\|$$

  where W and W' are quantized and complete data adjacency graphs, respectively, and L and L' are the corresponding graph Laplacians
- We merge the two most similar vertices in the graph W and increase the multiplicity of the new vertex
- The harmonic function solution on the quantized graph can be computed in $O(M^3)$ instead of $O(t^3)$

Theoretical Guarantees

- We seek a regret bound of the form:

$$\frac{1}{N}\sum_{t}(\hat{y}_t - y_t)^2 \le \frac{1}{N}\sum_{t}(y_t^* - y_t)^2$$

Online learning risk

Offline learning error

$$+ \frac{1}{N}\sum_{t}(y_t' - y_t^*)^2$$

Online learning error

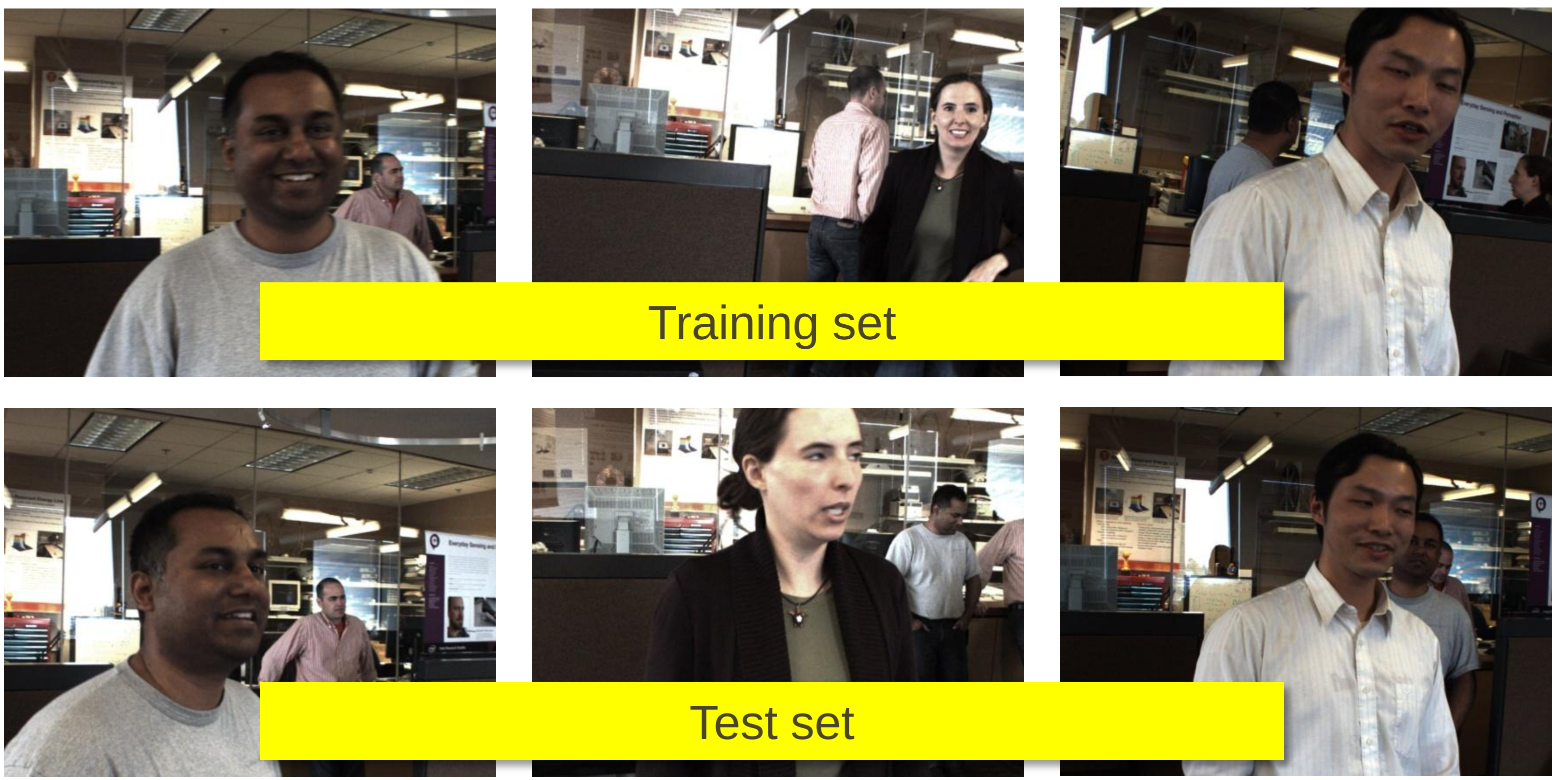
$$+ \frac{1}{N}\sum_{t}(\hat{y}_t - y_t')^2$$

Quantization error

- The errors should be bounded on the order of $\mathrm{O}\left(\sqrt{N}\right)$

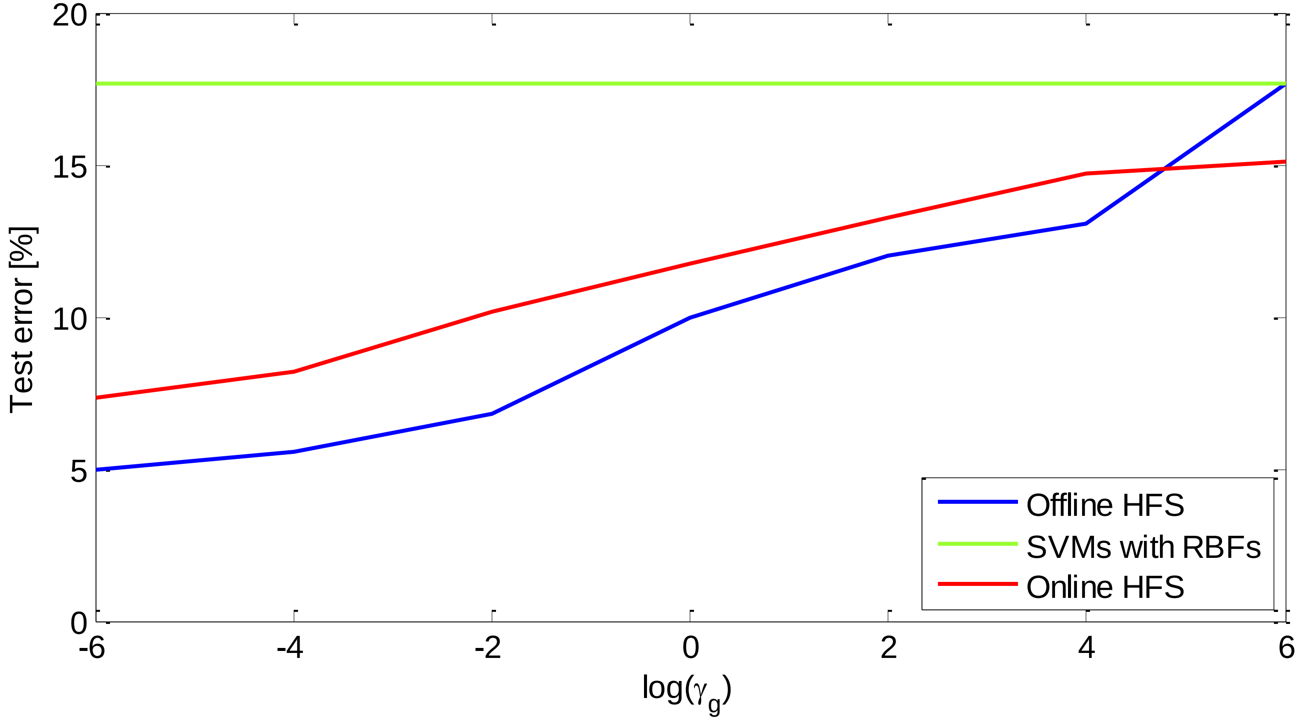
Experiments

- Face recognition of 3 people (roughly 1,500 faces) on a 60-second video from ILS Open House 2008



Experimental Results

- SVMs with RBFs misclassify 18 percent of faces
- Online HFS reduces the error to 7 percent



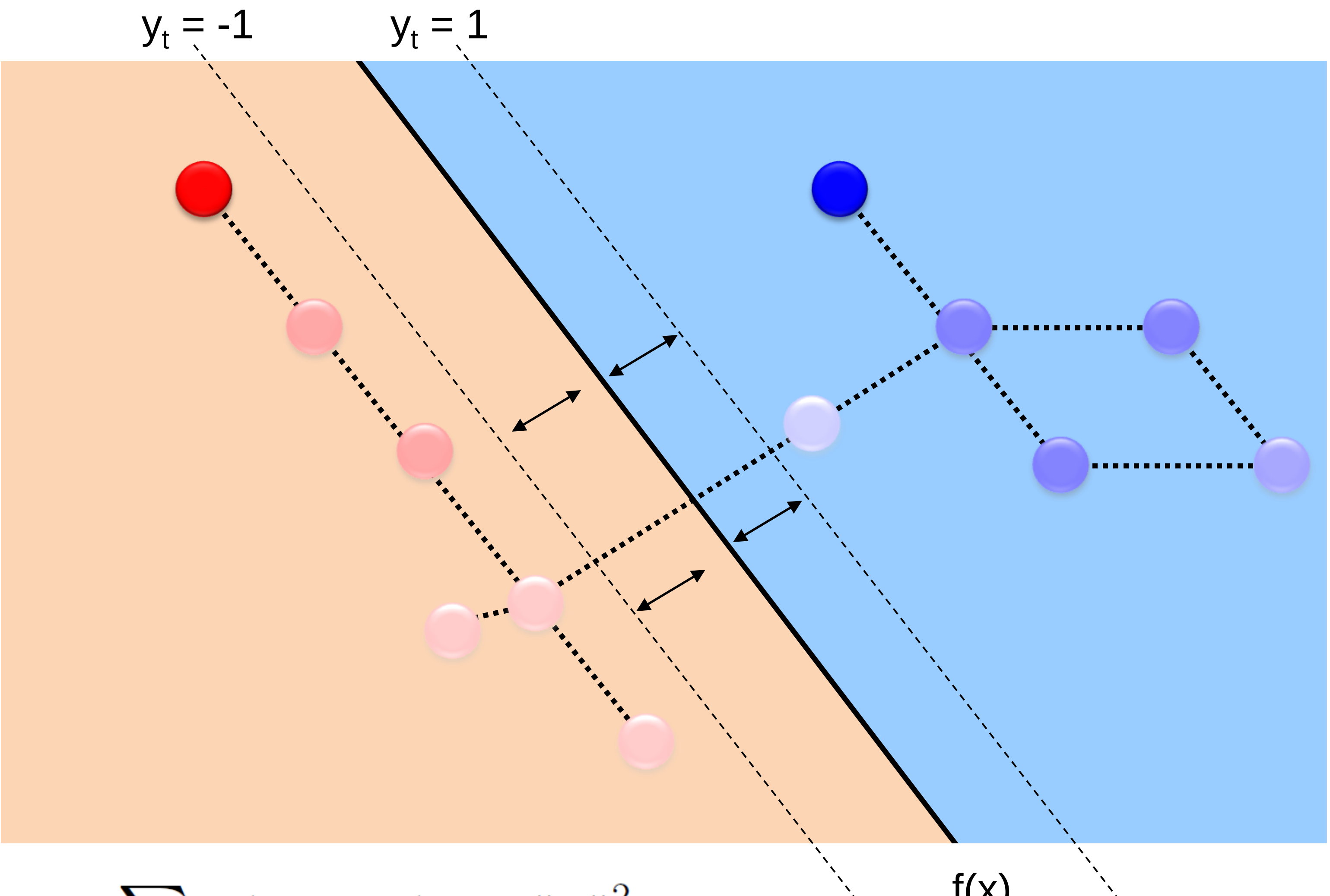
Video(s)

- [Go to Structured learning?](#)

Overview

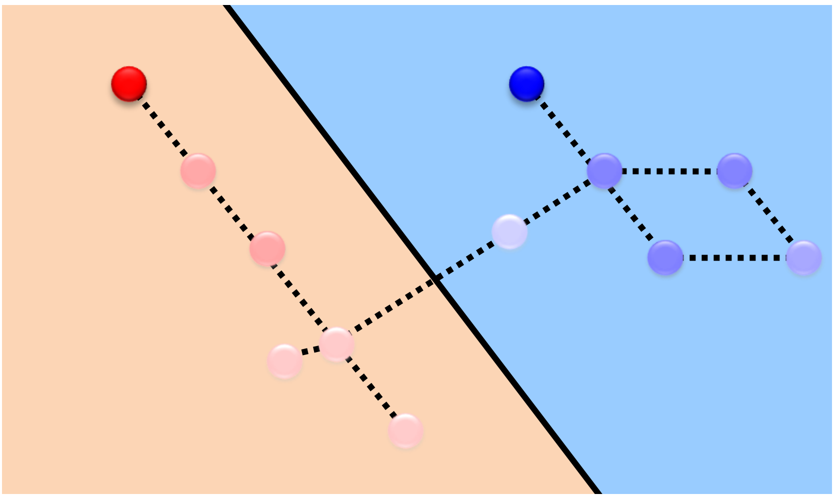
- Graph – based inference
- Offline Learning
- Online Learning
  - Face Recognition
- **Max – Margin graph cuts**
- Structured Learning
  - Handwriting Recognition

Max Margin Graph Cuts



$$\min_f \sum_{i \in l} V(f, \mathbf{x}_i, y_i) + \gamma \|f\|_K^2$$

Max Margin Graph Cuts

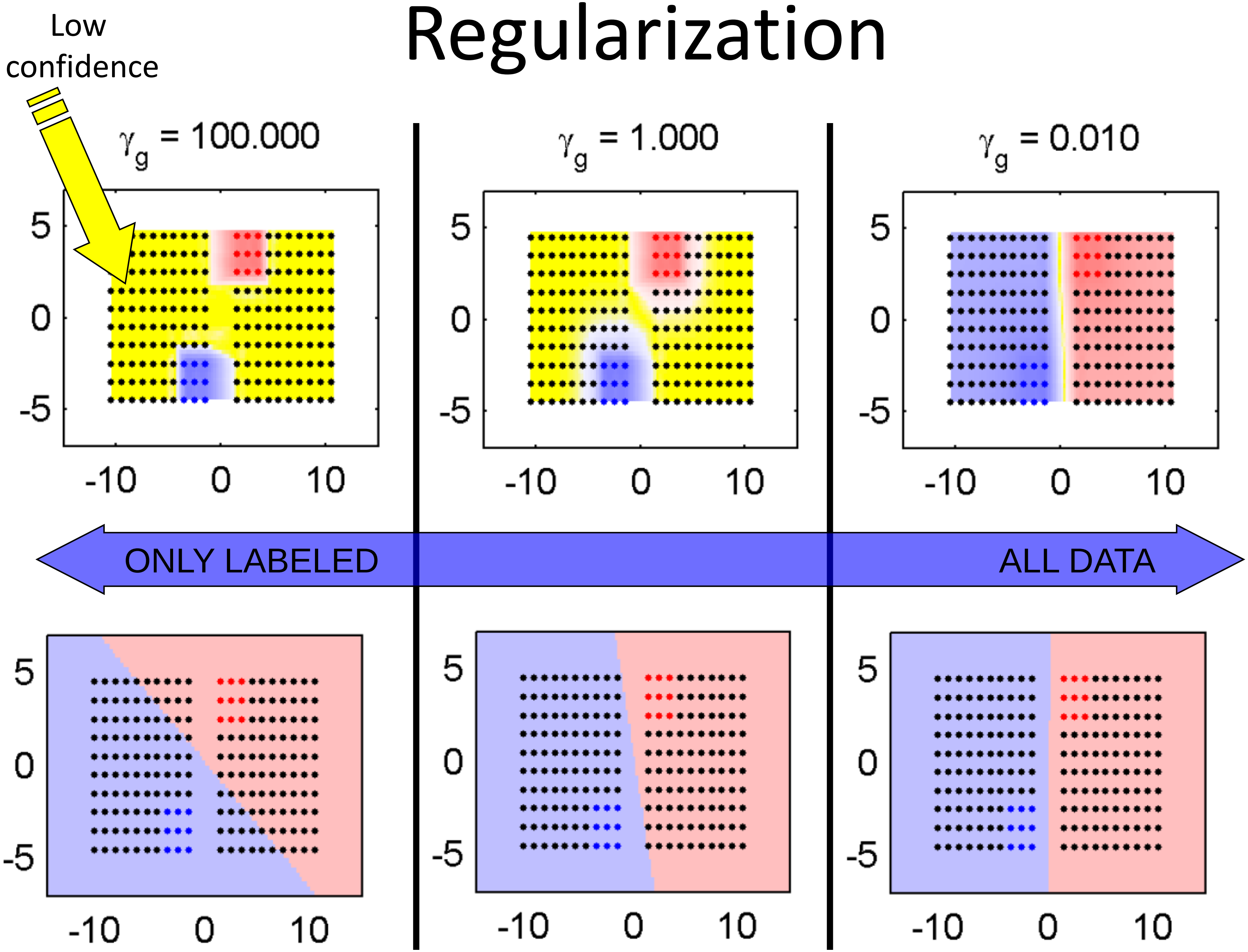


$$\min_f \sum_{i:|\ell_i^*| \geq \varepsilon} V(f, \mathbf{x}_i, \mathrm{sgn}(\ell_i^*)) + \gamma \|f\|_K^2$$

$$\text{s.t.} \quad \boldsymbol{\ell}^* = \arg\min_{\boldsymbol{\ell}} \boldsymbol{\ell}^\mathsf{T} (\gamma_g I + L) \boldsymbol{\ell}$$

$$\text{s.t. } \ell_i = y_i \text{ for all } i \in l;$$

Regularization



Theory

- With enough correctly inferred labels we can generalize well.

RISK

EMPIRICAL RISK WRT INFERRED LABELS

$$R_P(f) \leq \frac{1}{N} \sum_{i:|\ell_i^*| \geq \varepsilon} \mathcal{L}(f(\mathbf{x}_i), \mathrm{sgn}(\ell_i^*)) + \frac{\varepsilon N_\varepsilon}{N} +$$

$$\sqrt{\widehat{R}_T(\boldsymbol{\ell}^*)} + \sqrt{\Delta_T(\beta, N_l, \delta)} + \Delta_I(h, N, \eta)$$

GRAPH RISK

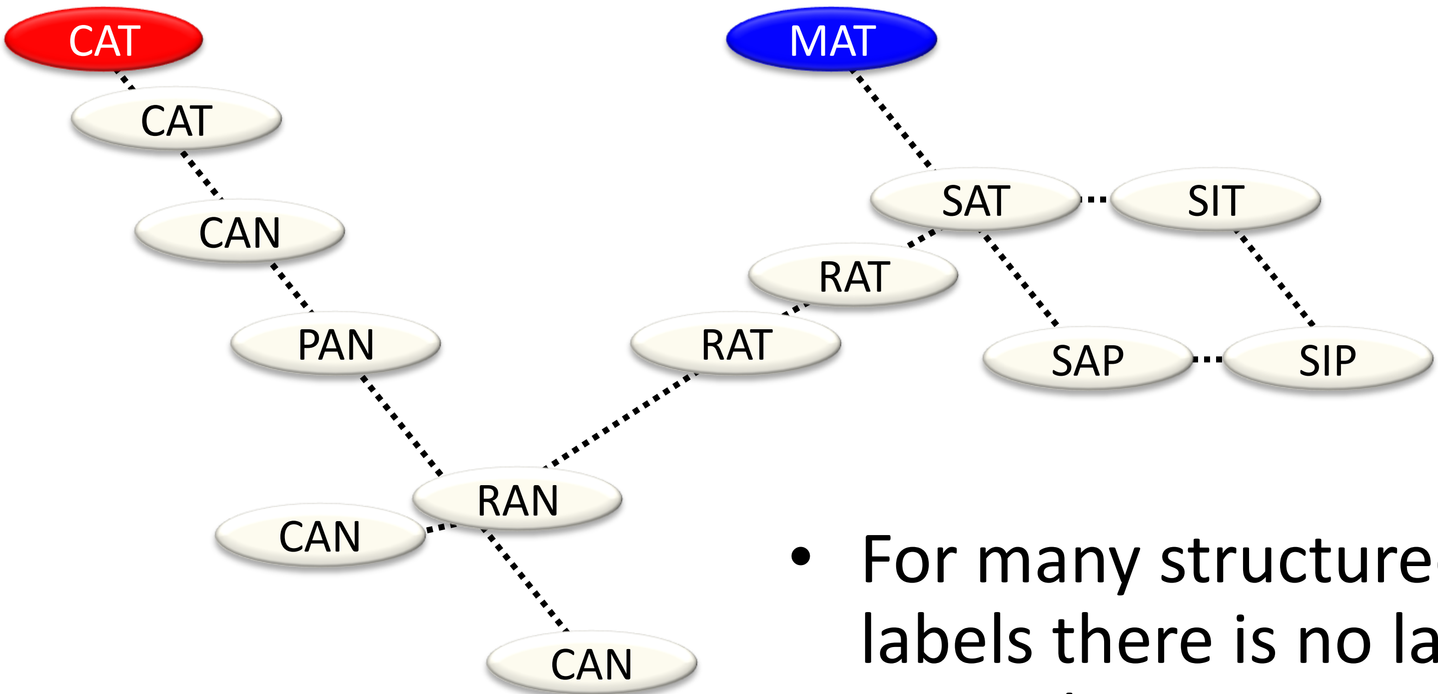
GRAPH COMPLEXITY

COMPLEXITY

Overview

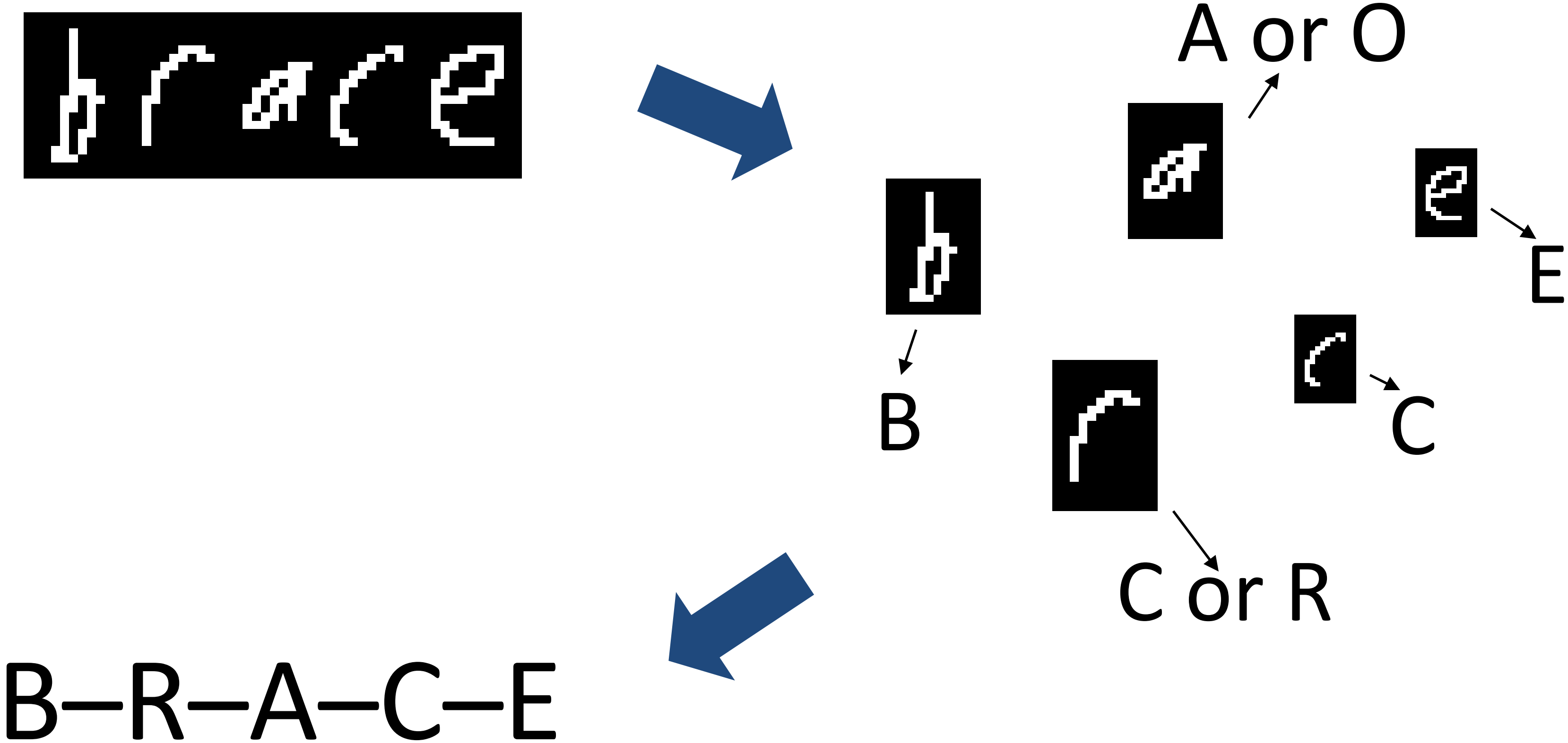
- Graph – based inference
- Offline Learning
- Online Learning
  - Face Recognition
- Max – Margin graph cuts
- **Structured Learning**
  - Handwriting Recognition

Structured Graph Cuts

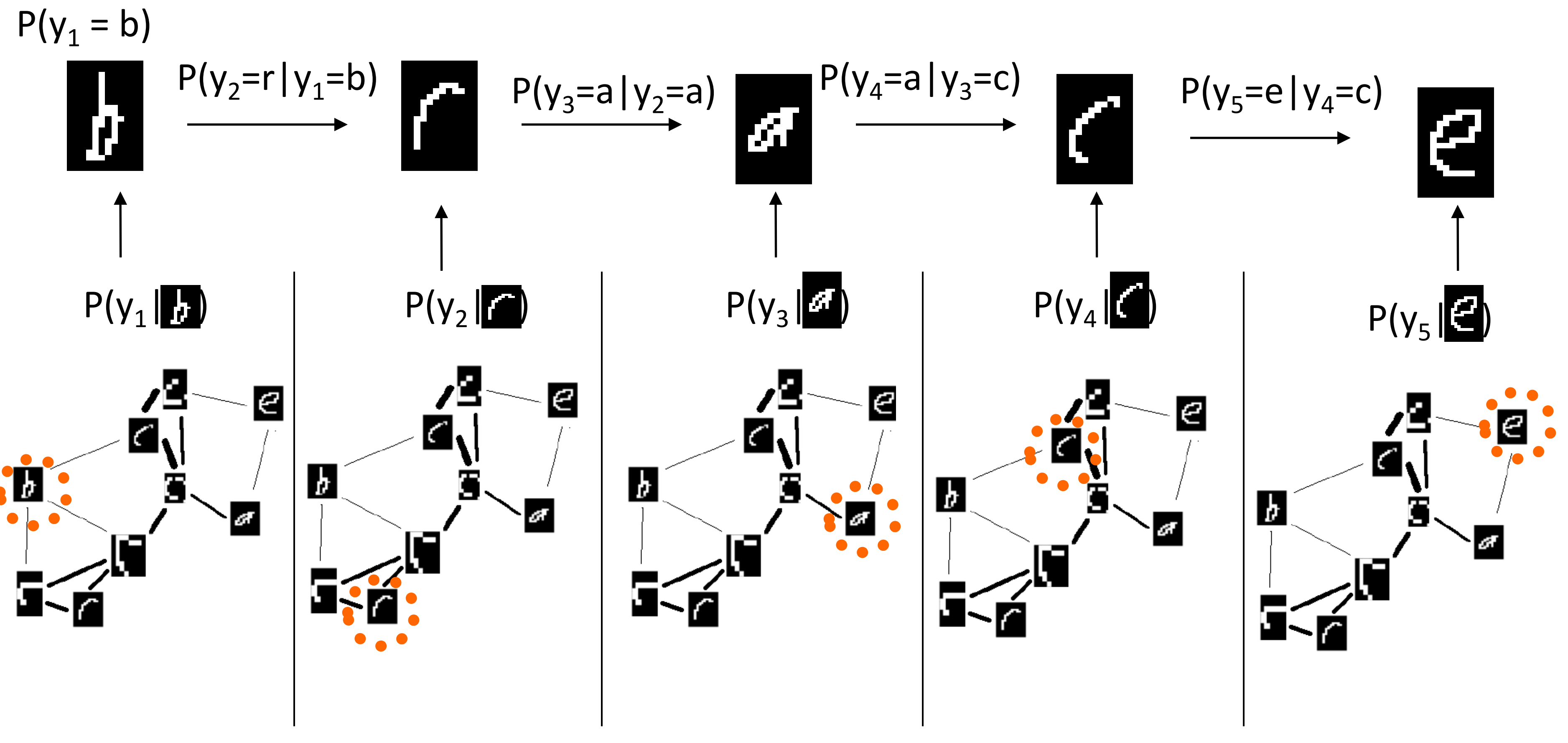


- For many structured labels there is no labeled example
- Huge number of possible structured labels

BREAK – INFER – SYNCHRONIZE



BREAK – INFER – SYNCHRONIZE

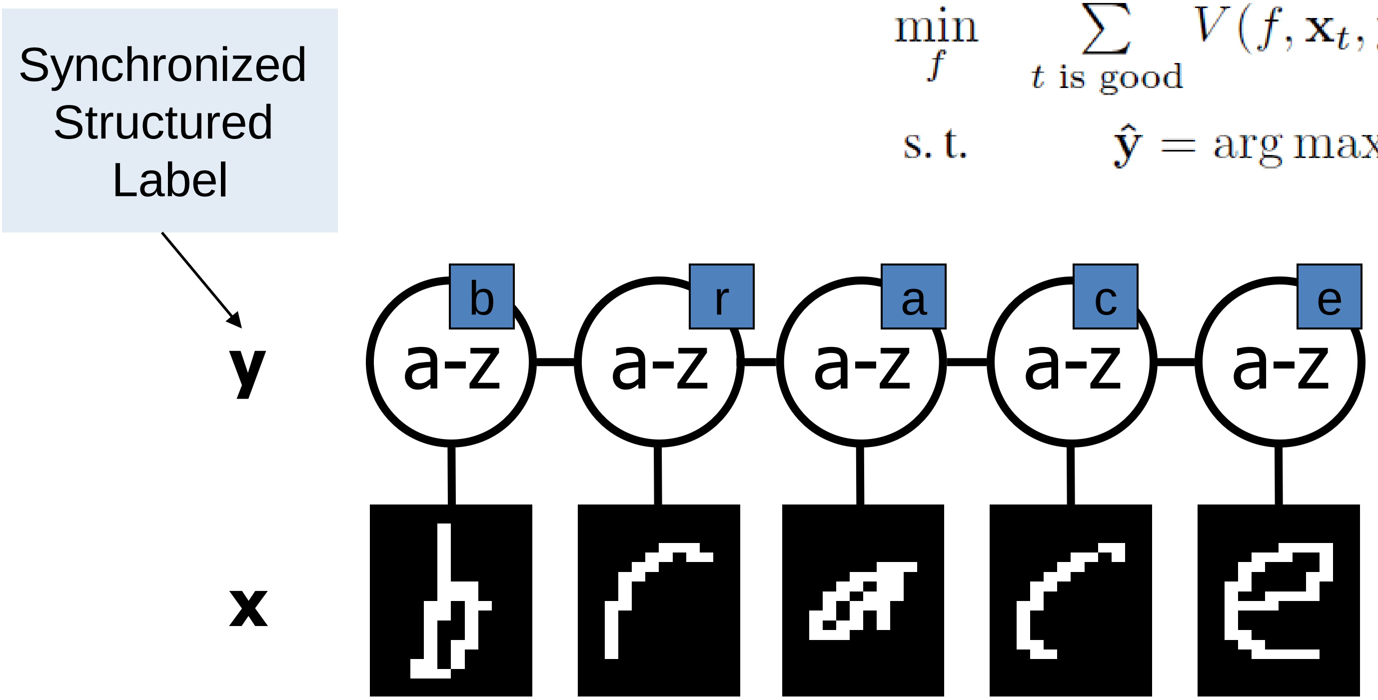


- Synchronization for sequences is done using Viterbi algorithm

Max Margin Markov Networks

- Augment M^{3}N learning with unlabeled structured data

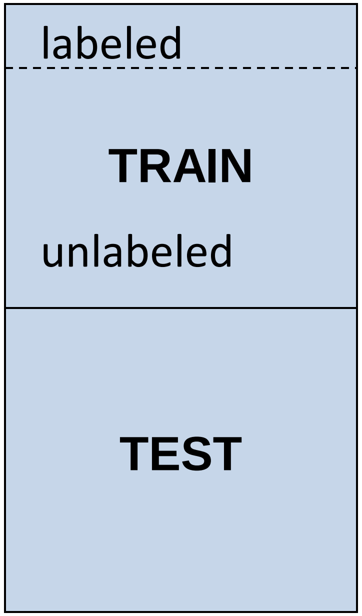
$$\min_f \sum_{t \text{ is good}} V(f, \mathbf{x}_t, \hat{\mathbf{y}}_t) + \gamma ||f||_K^2$$
$$\text{s.t.} \quad \hat{\mathbf{y}} = \arg\max_{\mathbf{y}} P(\mathbf{y}|\mathbf{x})$$



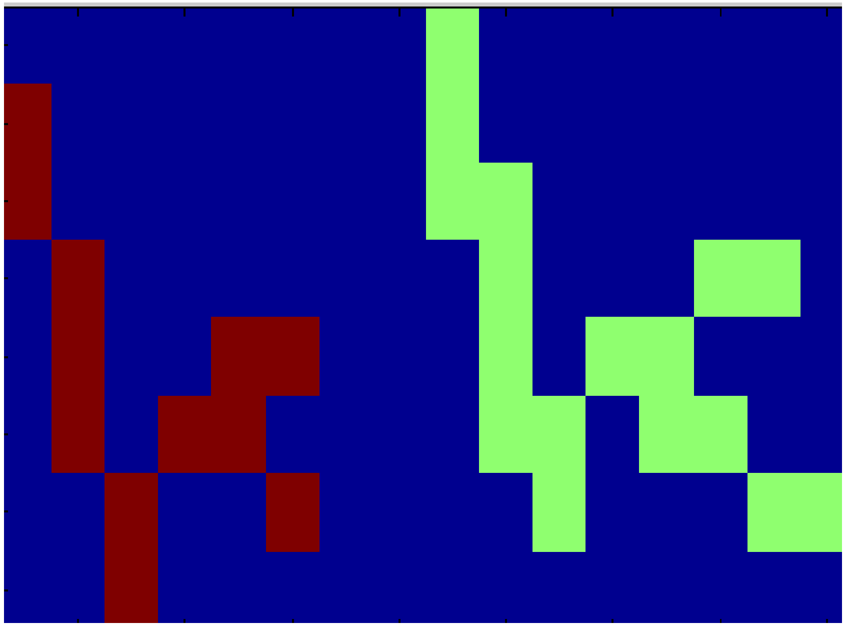
Maximum Margin Markov Networks (Taskar '03)

Offline experiments

- Handwriting Recognition – 26 way classification
- Letter: 16x8 pixels
- 7K words, 50K letters



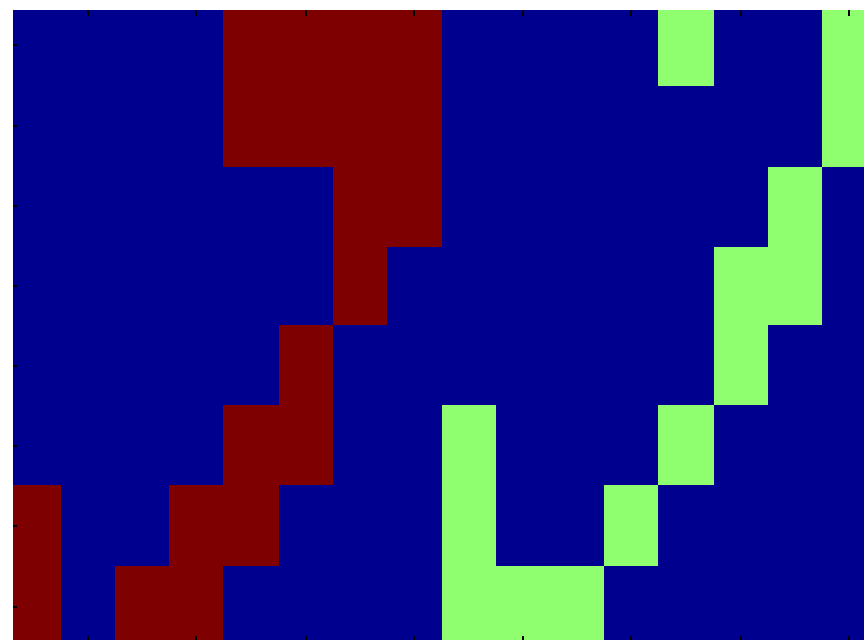
Letter K



‘Synchronizing’ structured labels

- *skiing*
- *s, k, i, i, n, g*
- S, K, I, I, N, Y
- S–K–I–I–N–G

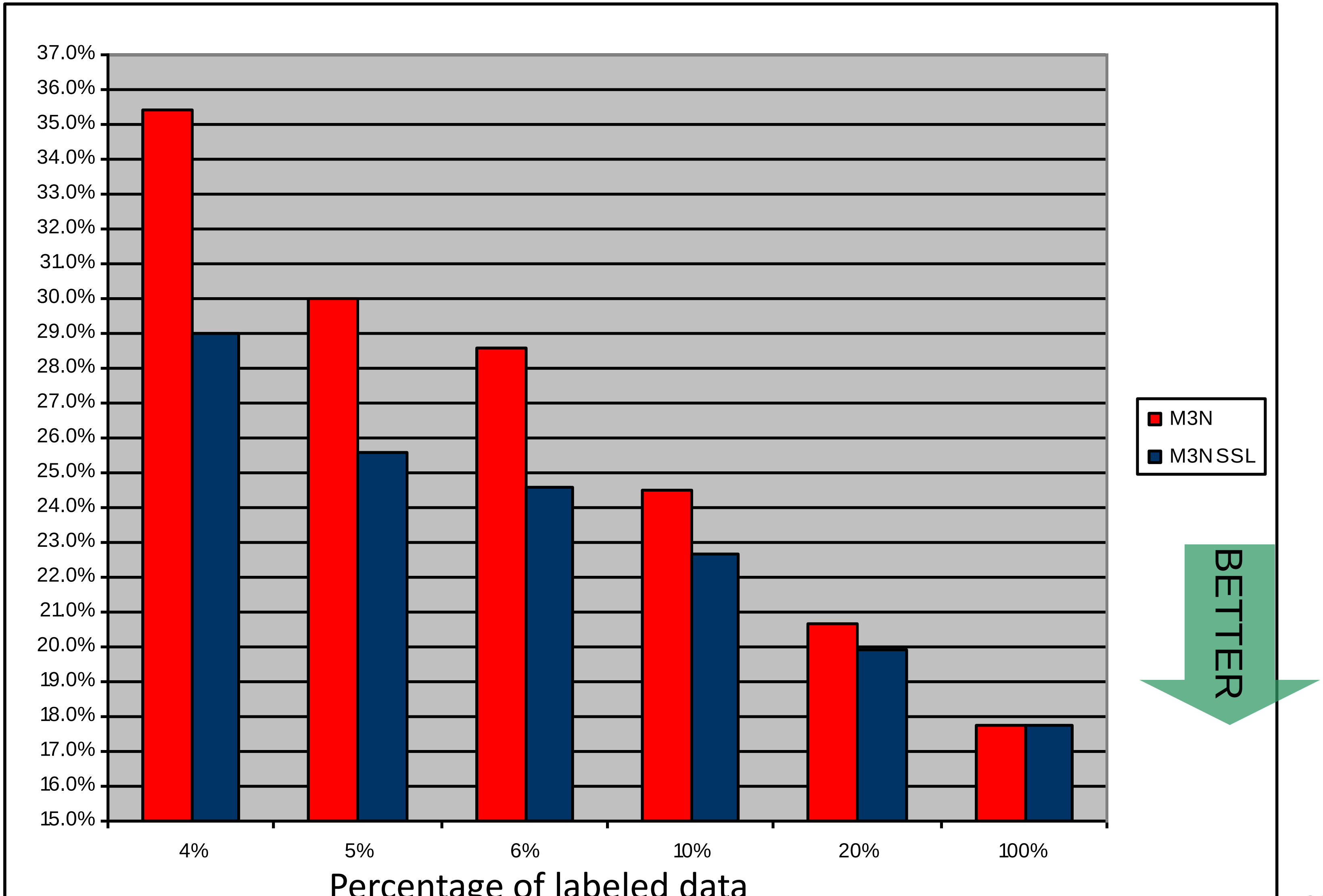
$P_Y = 0.28$ $P_G = 0.13$



Last letter of the word SKIING

Results

(supervised vs. semi-supervised error rates)



Results

(structured vs. unstructured error rates)