Training Structural SVMs when Exact Inference is Intractable

Thomas Finley, Thorsten Joachims
Cornell University

Talk Outline

- Structured Prediction
- Structural SVMs (SSVMs)
- Approximate Inference in SSVMs
  - Theoretical Analysis
  - Empirical Analysis

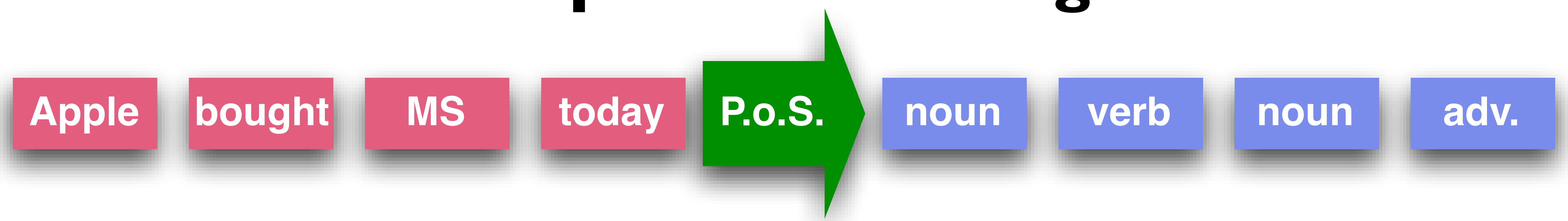
Structured Learning

Learning **functions** mapping **inputs** to **complex structured outputs**

Structured Learning

Learning **functions** mapping **inputs** to **complex structured outputs**

Sequence Labeling

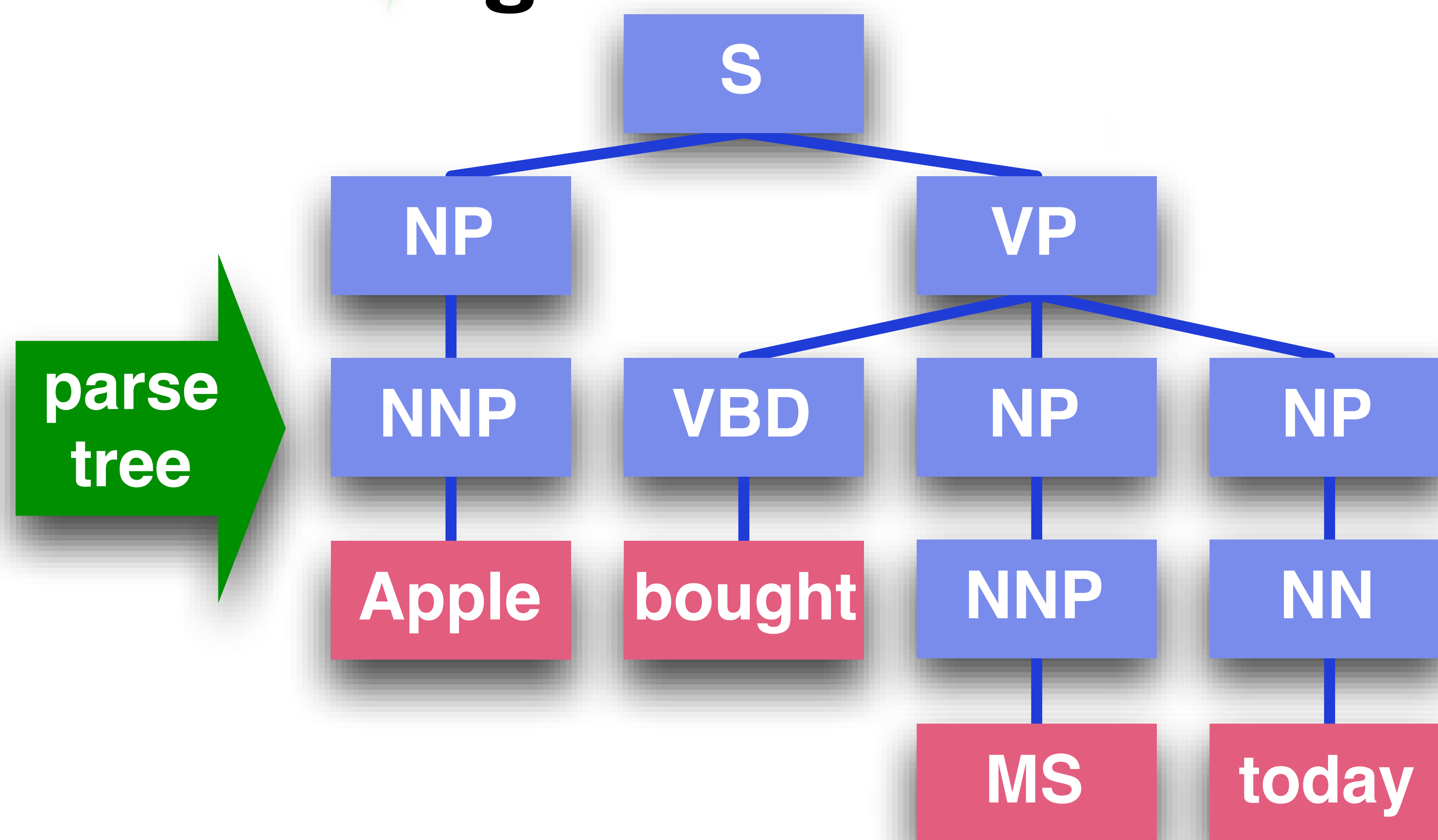


Structured Learning

Learning **functions** mapping **inputs** to **complex structured outputs**

Parsing

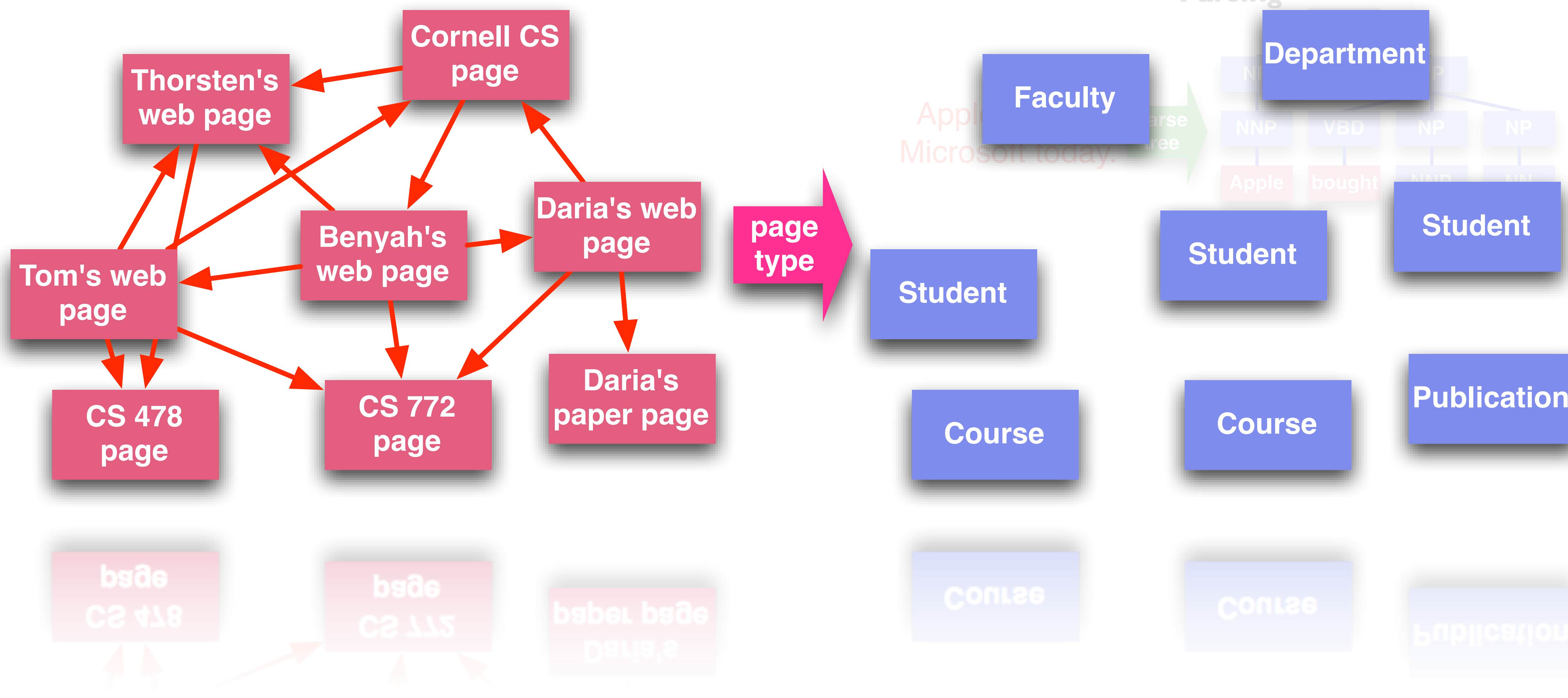
Apple bought Microsoft today.



Structured Learning

Learning **functions** mapping **inputs** to **complex structured outputs**

Collective Classification



Structured Learning

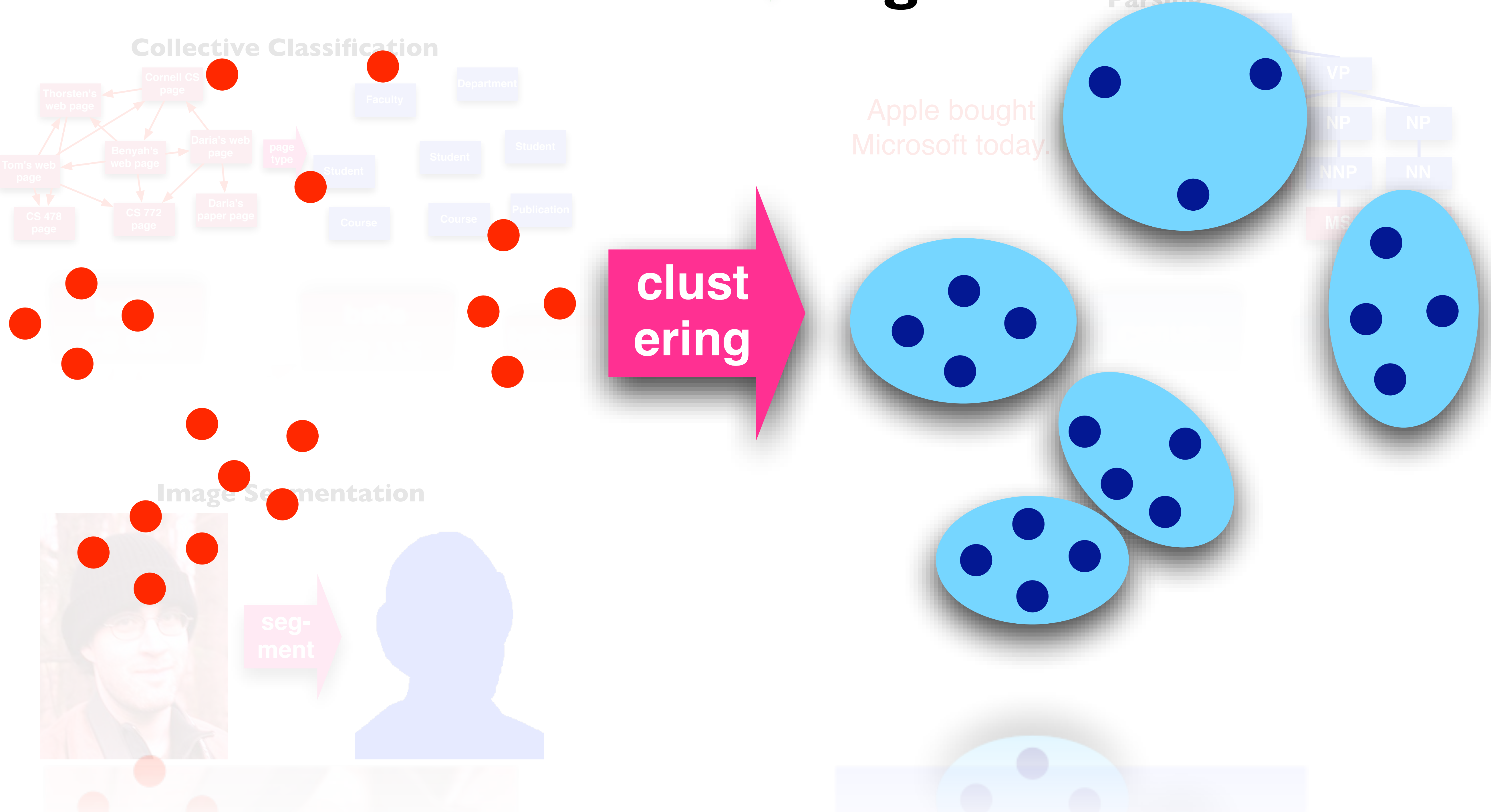
Learning **functions** mapping **inputs** to **complex structured outputs**

Image Segmentation

Structured Learning

Learning **functions** mapping **inputs** to **complex structured outputs**

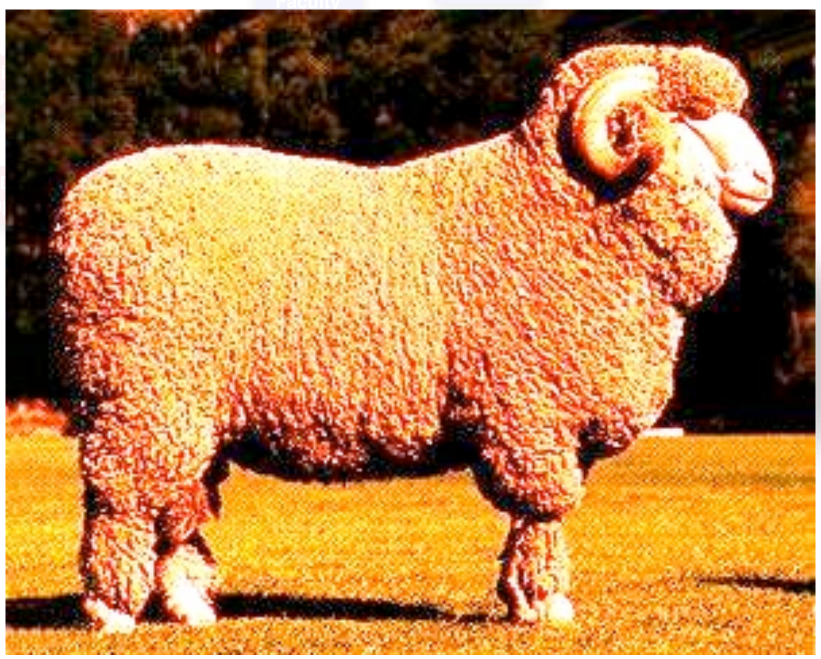
Clustering



Structured Learning

Learning **functions** mapping **inputs** to **complex structured outputs**

...even **Binary Classification**

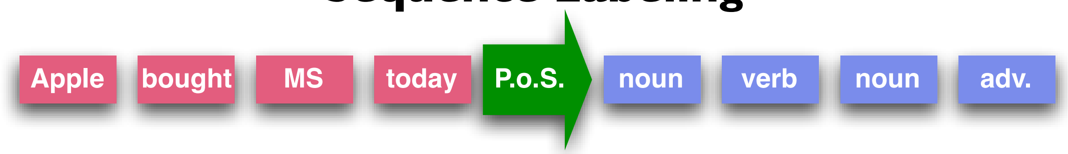


is merino? → yes

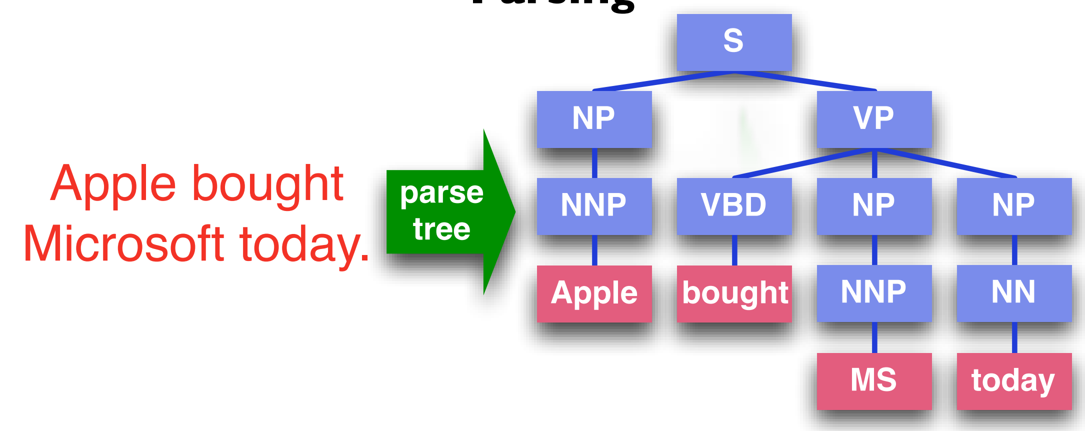
Structured Learning

Learning **functions** mapping **inputs** to **complex structured outputs**

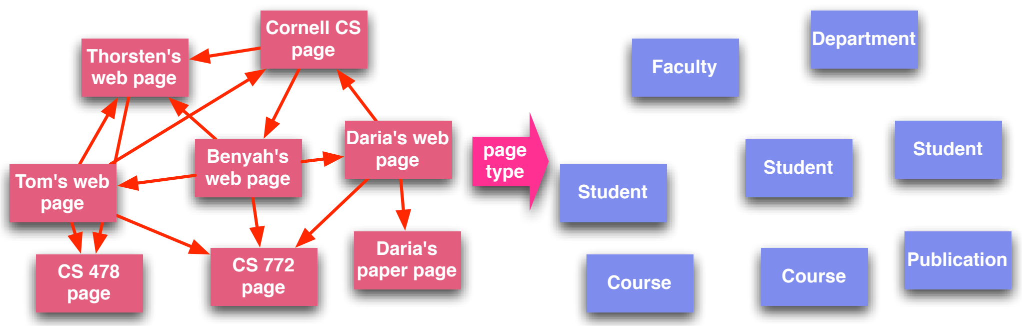
Sequence Labeling



Parsing



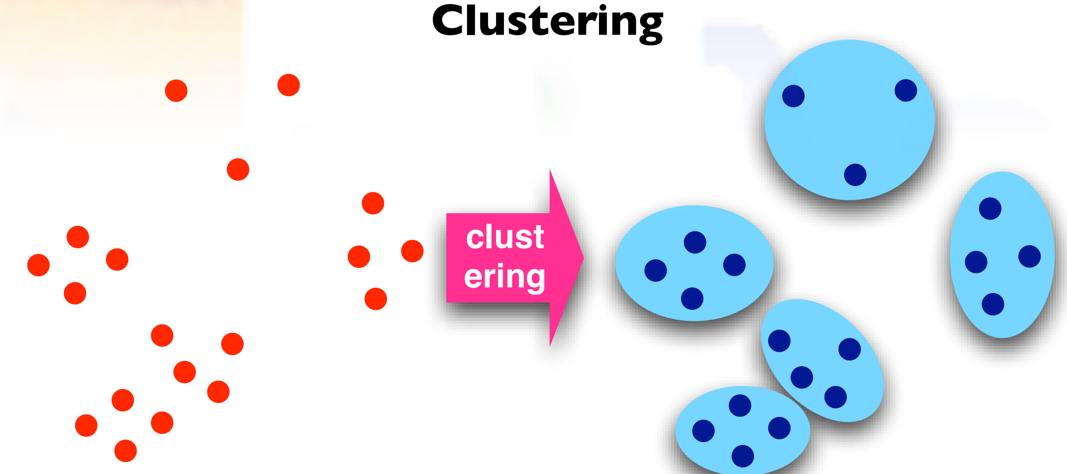
Collective Classification



...even **Binary Classification**

Image Segmentation

Clustering



Parameters for Structured Predictors

Parameters for Structured Predictors

- **Prediction Functions**: Output to maximize discriminant function.

$$h(\mathbf{x}) = \operatorname*{argmax}_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})$$

Parameters for Structured Predictors

- **Prediction Functions**: Output to maximize discriminant function.

$$h(\mathbf{x}) = \operatorname*{argmax}_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})$$

- **Discriminant Function *f* Form**: Product of model **w**, combined feature function Ψ.

$$h(\mathbf{x}) = \operatorname*{argmax}_{\mathbf{y}} \langle \mathbf{w}, \Psi(\mathbf{x}, \mathbf{y}) \rangle$$

Parameters for Structured Predictors

- **Prediction Functions**: Output to maximize discriminant function.
- **Discriminant Function *f* Form**: Product of model **w**, combined feature function Ψ.
- **Learning a Model**: Given (**x**,**y**) in-out pairs, find model **w**.

$$h(\mathbf{x}) = \operatorname*{argmax}_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})$$

$$h(\mathbf{x}) = \operatorname*{argmax}_{\mathbf{y}} \langle \mathbf{w}, \Psi(\mathbf{x}, \mathbf{y}) \rangle$$

Parameters for Structured Predictors

- **Prediction Functions**: Output to maximize discriminant function.

$$h(\mathbf{x}) = \operatorname*{argmax}_{\mathbf{y}} f(\mathbf{x}, \mathbf{y})$$

- **Discriminant Function *f* Form**: Product of model **w**, combined feature function Ψ.

$$h(\mathbf{x}) = \operatorname*{argmax}_{\mathbf{y}} \langle \mathbf{w}, \Psi(\mathbf{x}, \mathbf{y}) \rangle$$

- **Learning a Model**: Given (**x**,**y**) in-out pairs, find model **w**.
- **Learning methods**: CRF, M^3N, Structural SVM, Structural Perceptrons (Tsochantaridis et al. '04, Lafferty et al. '01, Taskar et al. '03, Collins et al., Altun et al. '03). All common in this way! Differ how they pick **w** given (**x**,**y**) sample.

Some tasks have intractable exact $\text{argmax}_{\mathbf{y}}\, f(\mathbf{x},\mathbf{y})$......

Some tasks have intractable exact $\text{argmax}_{\mathbf{y}}\ f(\mathbf{x},\mathbf{y})$......

Image segmentation...

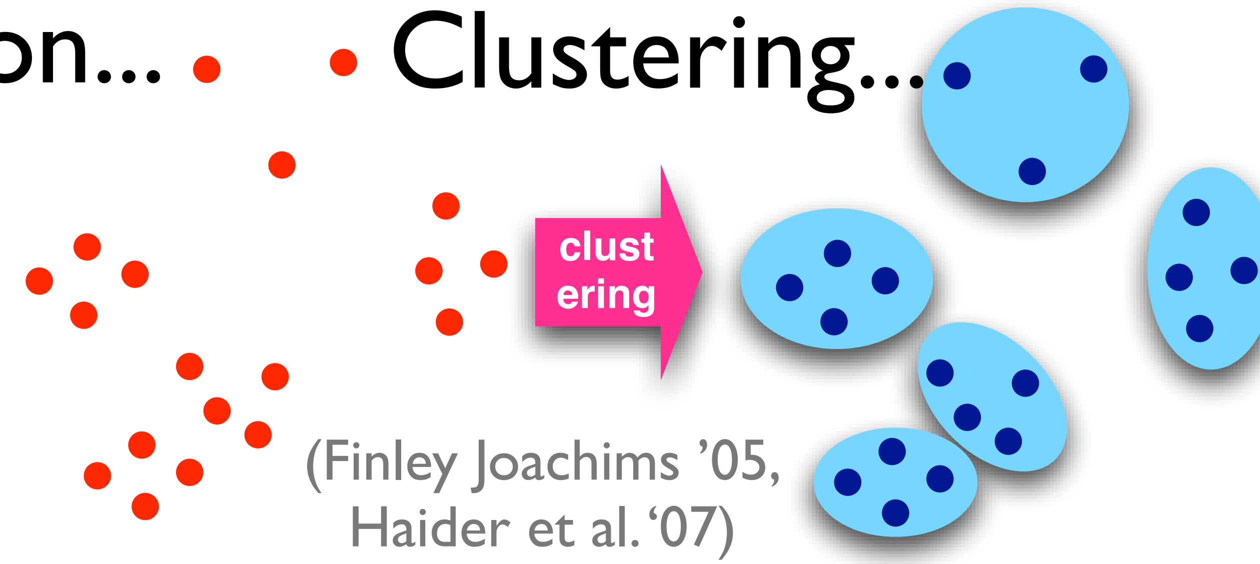
(Anguelov et al. '05, Cinque et al. '00, He et al. '04, Kumar et al. '03)

Some tasks have intractable exact $\text{argmax}_{\mathbf{y}}\, f(\mathbf{x},\mathbf{y})$......

Image segmentation...

(Anguelov et al. '05, Cinque et al. '00, He et al. '04, Kumar et al. '03)

Clustering...



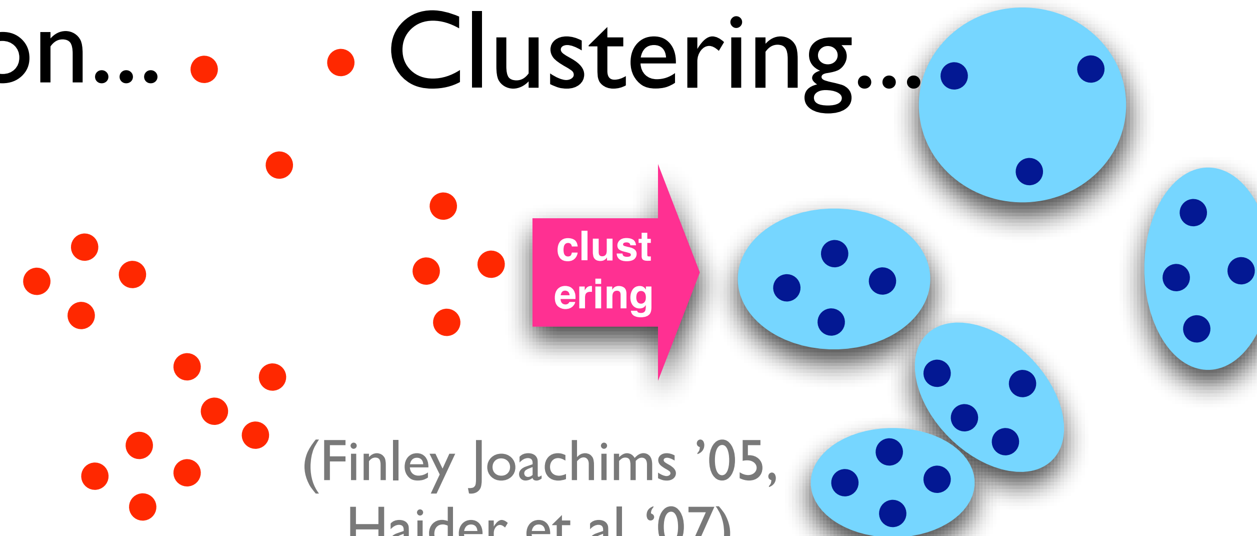
(Finley Joachims '05, Haider et al. '07)

Some tasks have intractable exact $\text{argmax}_{\mathbf{y}}\, f(\mathbf{x},\mathbf{y})$......

Image segmentation...

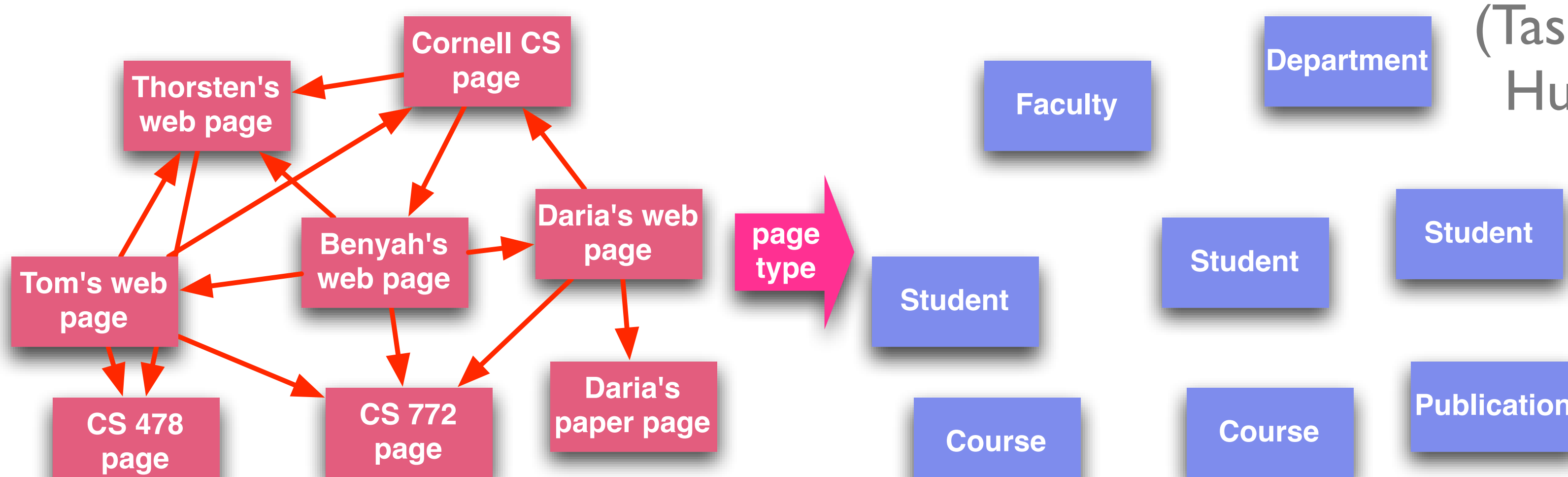
(Anguelov et al. '05, Cinque et al. '00, He et al. '04, Kumar et al. '03)

Clustering...



(Finley Joachims '05, Haider et al. '07)

Some classification tasks...



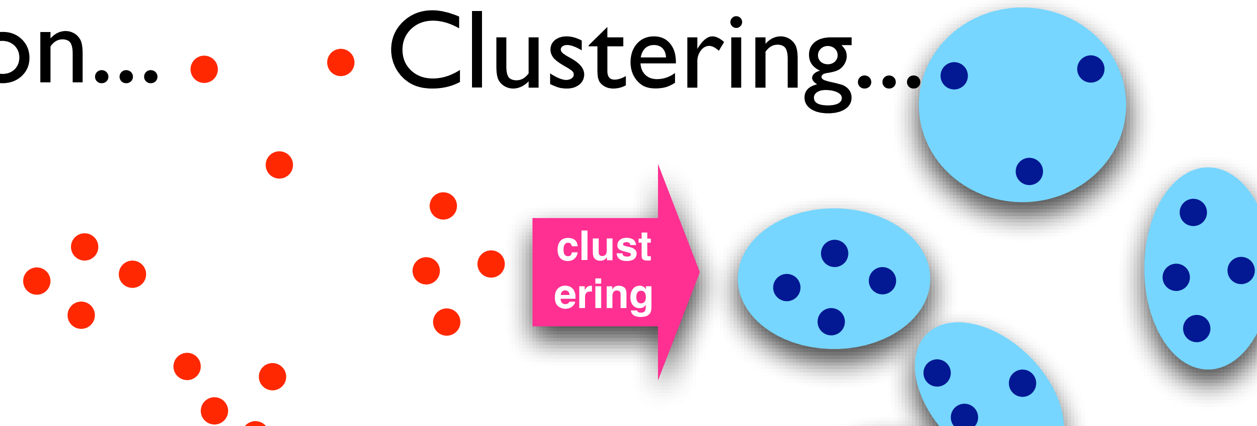
(Taskar et al. '03, Lan Huttenlocher '05)

Some tasks have intractable exact $\text{argmax}_{\mathbf{y}}\, f(\mathbf{x},\mathbf{y})$......

Image segmentation...

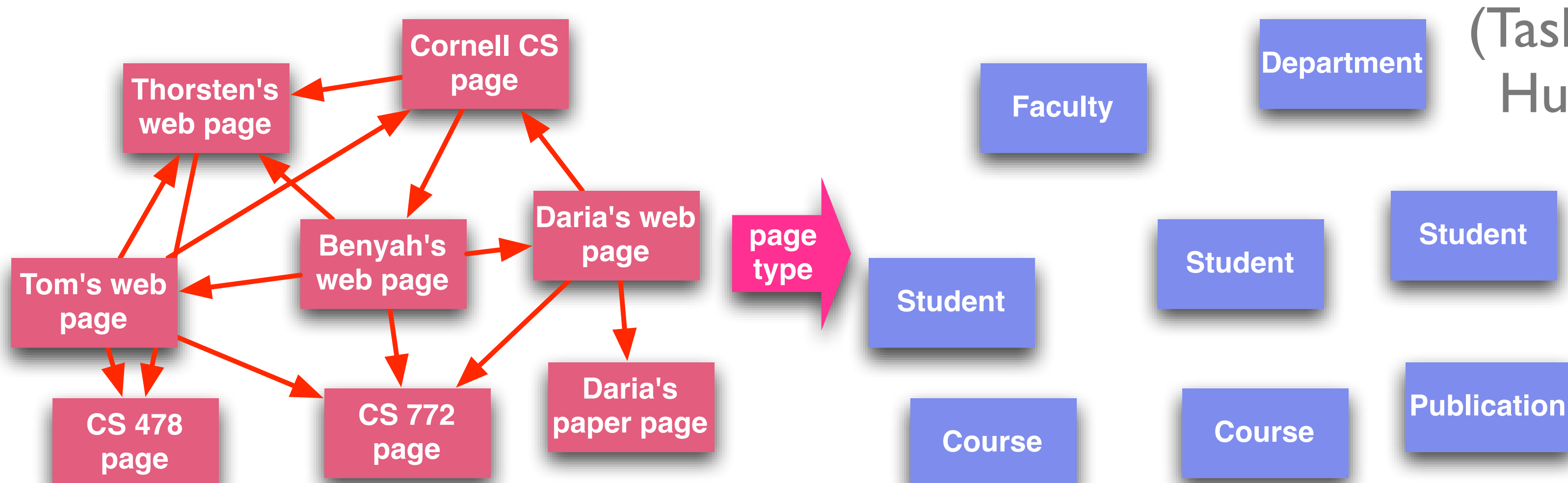
(Anguelov et al. '05, Cinque et al. '00, He et al. '04, Kumar et al. '03)

Clustering...



(Finley Joachims '05, Haider et al. '07)

Some classification tasks...



(Taskar et al. '03, Lan Huttenlocher '05)

When one must approximate argmax, learning **w** faces new challenges.

Talk Outline

- Structured Prediction
- Structural SVMs (SSVMs)
- Approximate Inference in SSVMs
  - Theoretical Analysis
  - Empirical Analysis

Linear Constraint

$$\forall i, \forall \mathbf{y} \in \mathcal{Y} : \langle \mathbf{w}, \Psi(\mathbf{x}_i, \mathbf{y}_i) \rangle - \langle \mathbf{w}, \Psi(\mathbf{x}_i, \mathbf{y}) \rangle \geq \Delta(\mathbf{y}_i, \mathbf{y}) - \xi_i$$

Linear Constraint

- For **all training examples ($\mathbf{x}_i$, $\mathbf{y}_i$)**...

$$\boxed{\forall i}, \forall \mathbf{y} \in \mathcal{Y} : \langle \mathbf{w}, \Psi(\mathbf{x}_i, \mathbf{y}_i) \rangle - \langle \mathbf{w}, \Psi(\mathbf{x}_i, \mathbf{y}) \rangle \geq \Delta(\mathbf{y}_i, \mathbf{y}) - \xi_i$$

Linear Constraint

- For **all training examples ($\mathbf{x}_i$, $\mathbf{y}_i$)**...
- ...and **any possible wrong output y**...

$$\boxed{\forall i,} \boxed{\forall \mathbf{y} \in \mathcal{Y}} : \langle \mathbf{w}, \Psi(\mathbf{x}_i, \mathbf{y}_i) \rangle - \langle \mathbf{w}, \Psi(\mathbf{x}_i, \mathbf{y}) \rangle \geq \Delta(\mathbf{y}_i, \mathbf{y}) - \xi_i$$

Linear Constraint

- For **all training examples ($\mathbf{x}_i$, $\mathbf{y}_i$)**...
- ...and **any possible wrong output y**...
- ...have the **discriminant function for the correct output**...

$$\boxed{\forall i,} \boxed{\forall \mathbf{y} \in \mathcal{Y}} : \boxed{\langle \mathbf{w}, \Psi(\mathbf{x}_i, \mathbf{y}_i) \rangle} - \langle \mathbf{w}, \Psi(\mathbf{x}_i, \mathbf{y}) \rangle \geq \Delta(\mathbf{y}_i, \mathbf{y}) - \xi_i$$

Linear Constraint

- For **all training examples ($\mathbf{x}_i$, $\mathbf{y}_i$)**...
- ...and **any possible wrong output y**...
- ...have the **discriminant function for the correct output**...
- ...greater than the **discriminant function for the incorrect output**...

$$\boxed{\forall i,} \boxed{\forall \mathbf{y} \in \mathcal{Y}} : \boxed{\langle \mathbf{w}, \Psi(\mathbf{x}_i, \mathbf{y}_i) \rangle} - \boxed{\langle \mathbf{w}, \Psi(\mathbf{x}_i, \mathbf{y}) \rangle} \geq \Delta(\mathbf{y}_i, \mathbf{y}) - \xi_i$$

Linear Constraint

- For **all training examples ($\mathbf{x}_i$, $\mathbf{y}_i$)**...
- ...and **any possible wrong output y**...
- ...have the **discriminant function for the correct output**...
- ...greater than the **discriminant function for the incorrect output**...
- ...by at least the **loss between the correct and incorrect output**.

$$\boxed{\forall i,} \boxed{\forall \mathbf{y} \in \mathcal{Y}} : \boxed{\langle \mathbf{w}, \Psi(\mathbf{x}_i, \mathbf{y}_i) \rangle} - \boxed{\langle \mathbf{w}, \Psi(\mathbf{x}_i, \mathbf{y}) \rangle} \geq \boxed{\Delta(\mathbf{y}_i, \mathbf{y})} - \xi_i$$

Linear Constraint

- For **all training examples ($\mathbf{x}_i$, $\mathbf{y}_i$)**...
- ...and **any possible wrong output y**...
- ...have the **discriminant function for the correct output**...
- ...greater than the **discriminant function for the incorrect output**...
- ...by at least the **loss between the correct and incorrect output**.
- **Slack** serves as a bound on empirical risk.

$$\boxed{\forall i,} \boxed{\forall \mathbf{y} \in \mathcal{Y}} : \boxed{\langle \mathbf{w}, \Psi(\mathbf{x}_i, \mathbf{y}_i) \rangle} - \boxed{\langle \mathbf{w}, \Psi(\mathbf{x}_i, \mathbf{y}) \rangle} \geq \boxed{\Delta(\mathbf{y}_i, \mathbf{y})} - \boxed{\xi_i}$$

Linear Constraint

$$\forall i, \forall \mathbf{y} \in \mathcal{Y} : \langle \mathbf{w}, \Psi(\mathbf{x}_i, \mathbf{y}_i) \rangle - \langle \mathbf{w}, \Psi(\mathbf{x}_i, \mathbf{y}) \rangle \geq \Delta(\mathbf{y}_i, \mathbf{y}) - \xi_i$$

Quadratic Program Formulation

$$\min_{\mathbf{w},\xi} \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{n}\sum_{i=1}^{n}\xi_i$$

$$\begin{aligned} \text{s.t.} \quad & \forall i : \xi_i \geq 0 \\ & \forall i, \forall \mathbf{y} \in \mathcal{Y} : \langle \mathbf{w}, \Psi(\mathbf{x}_i, \mathbf{y}_i)\rangle - \langle \mathbf{w}, \Psi(\mathbf{x}_i, \mathbf{y})\rangle \geq \Delta(\mathbf{y}_i, \mathbf{y}) - \xi_i \end{aligned}$$

- **Empirical Risk:** each ξ_i upper bounds training error, so ξ term overall upper bound on empirical risk.

Quadratic Program Formulation

$$\min_{\mathbf{w},\xi} \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{n}\sum_{i=1}^{n}\xi_i$$

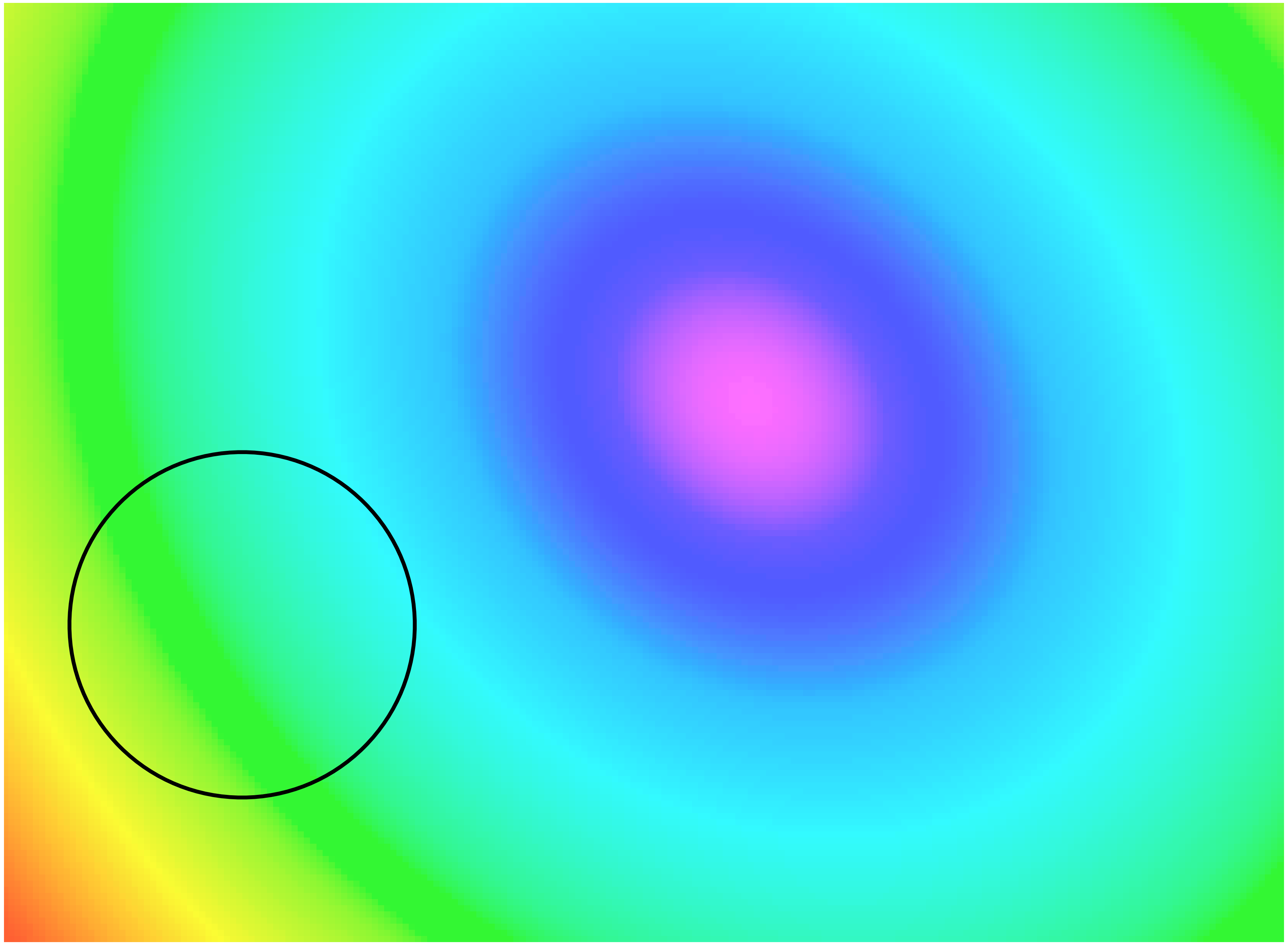
$$\text{s.t.} \quad \forall i : \xi_i \geq 0$$

$$\forall i, \forall \mathbf{y} \in \mathcal{Y} : \langle \mathbf{w}, \Psi(\mathbf{x}_i, \mathbf{y}_i)\rangle - \langle \mathbf{w}, \Psi(\mathbf{x}_i, \mathbf{y})\rangle \geq \Delta(\mathbf{y}_i, \mathbf{y}) - \xi_i$$

So many constraints!

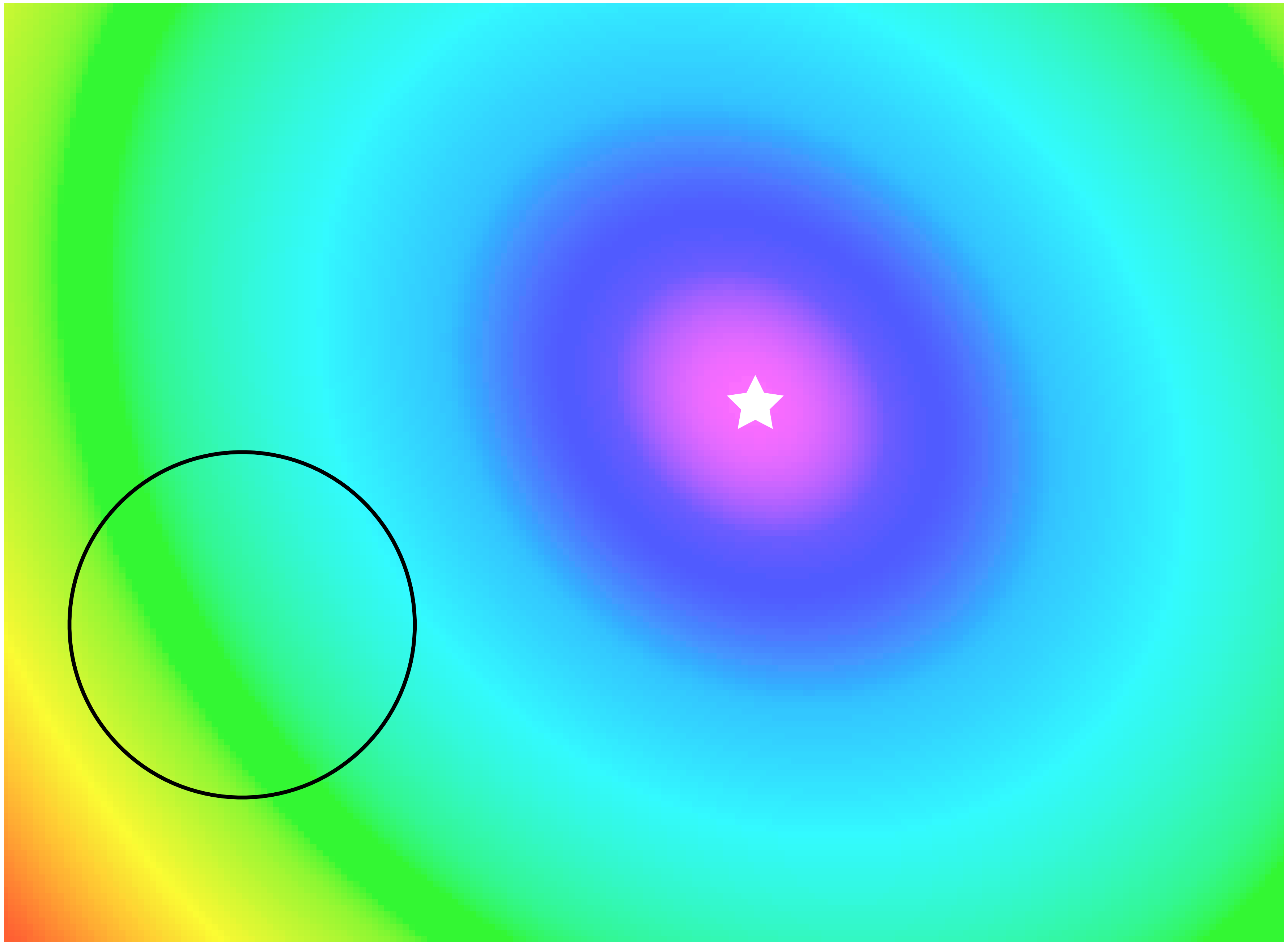
- **Empirical Risk:** each ξ_i upper bounds training error, so ξ term overall upper bound on empirical risk.

Cutting Plane Example



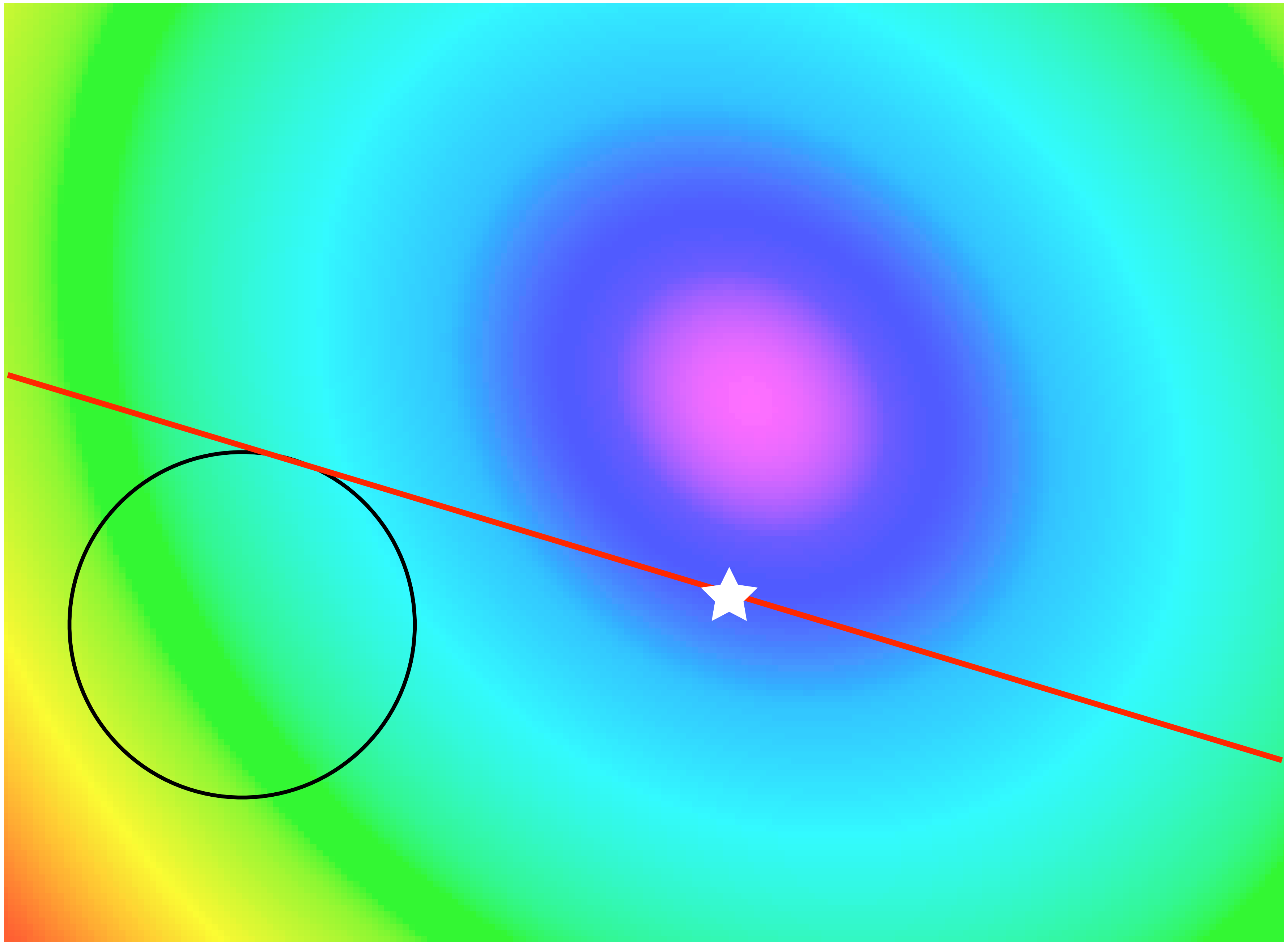
- Use column generation!
- Start with unconstrained problem.
- Optimize, find most violated constraint, introduce, and reoptimize.
- Repeat until no constraint in full problem violated by more than some tolerance!

Cutting Plane Example



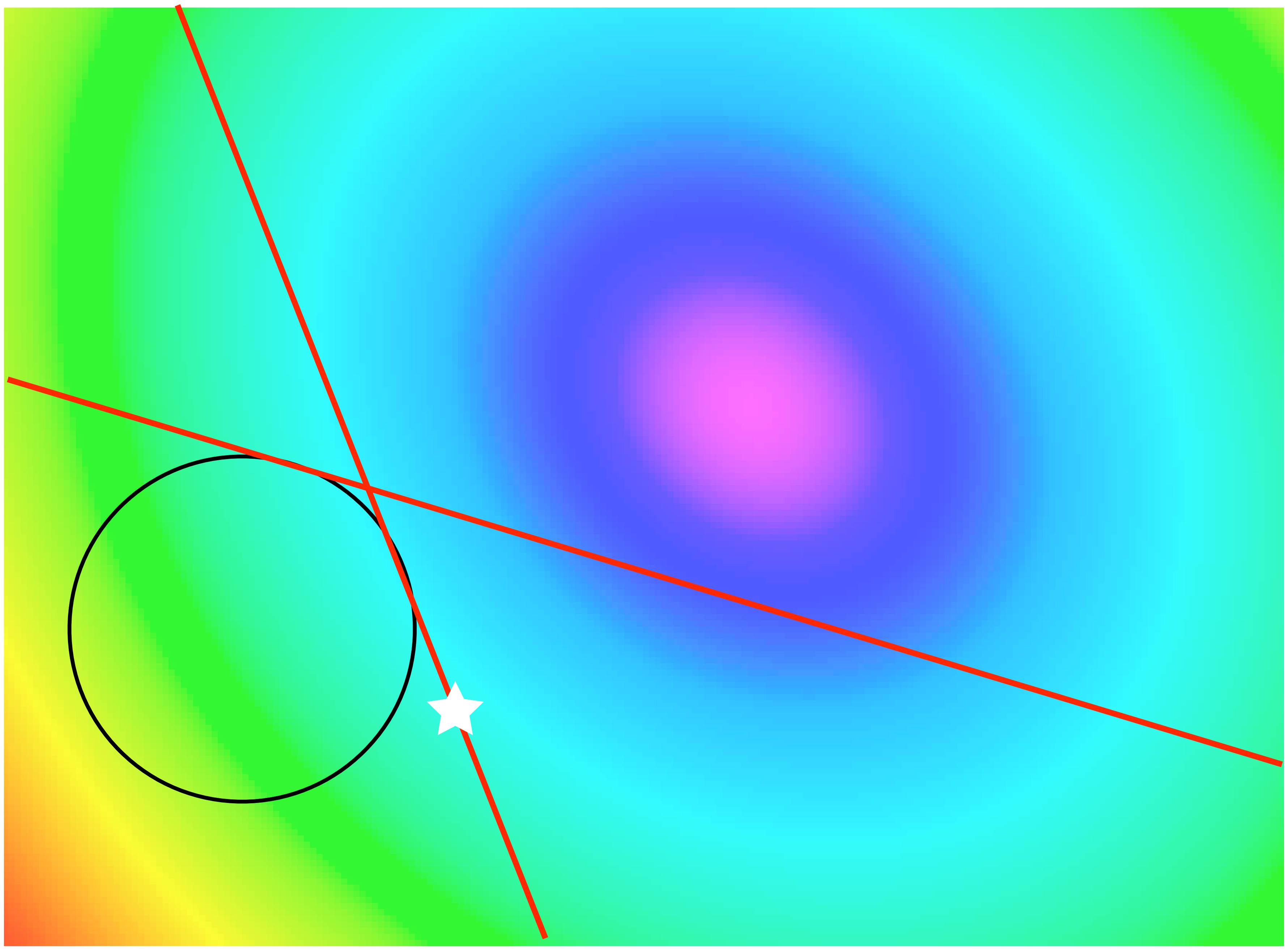
- Use column generation!
- Start with unconstrained problem.
- Optimize, find most violated constraint, introduce, and reoptimize.
- Repeat until no constraint in full problem violated by more than some tolerance!

Cutting Plane Example



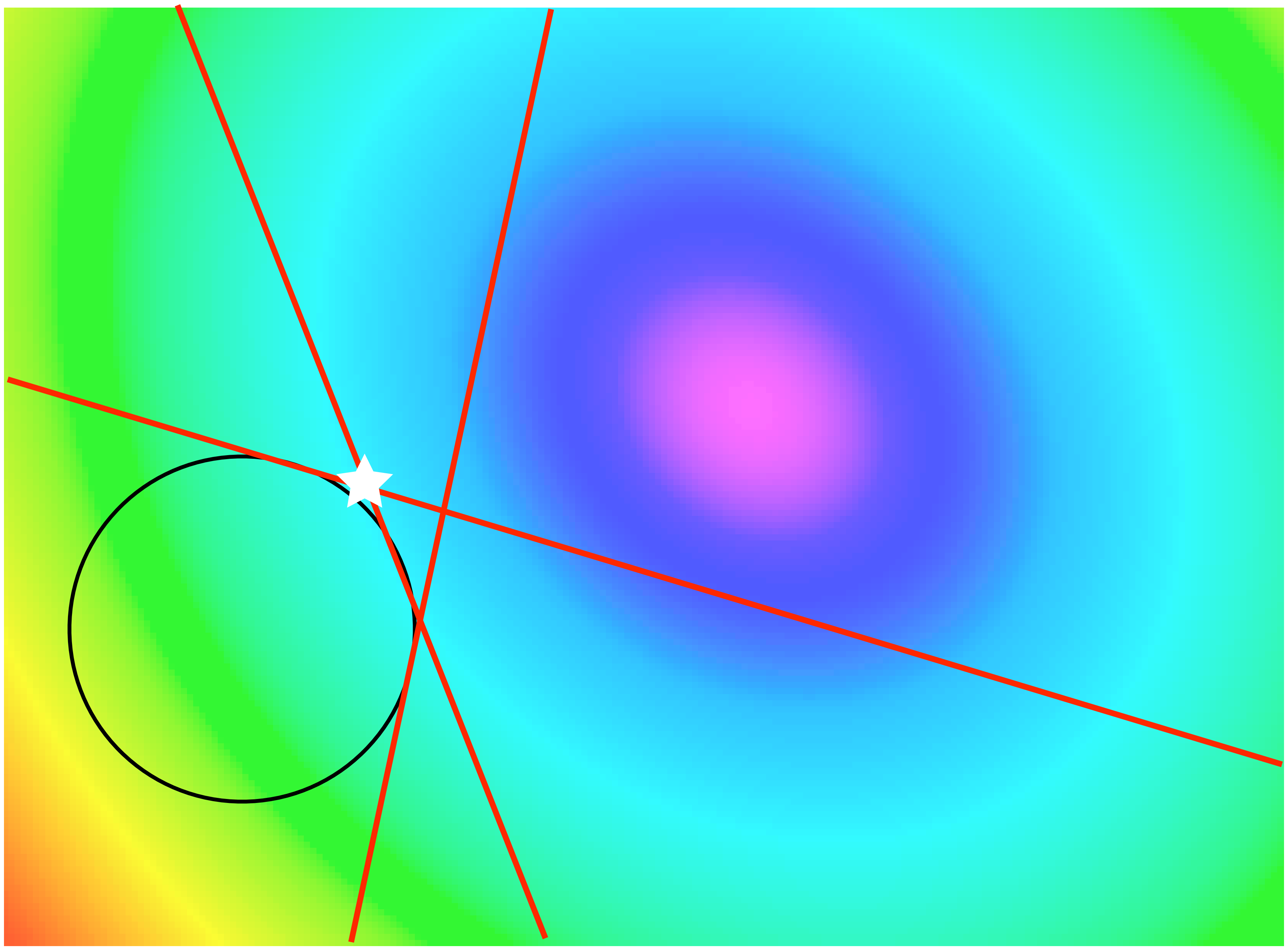
- Use column generation!
- Start with unconstrained problem.
- Optimize, find most violated constraint, introduce, and reoptimize.
- Repeat until no constraint in full problem violated by more than some tolerance!

Cutting Plane Example



- Use column generation!
- Start with unconstrained problem.
- Optimize, find most violated constraint, introduce, and reoptimize.
- Repeat until no constraint in full problem violated by more than some tolerance!

Cutting Plane Example



- Use column generation!
- Start with unconstrained problem.
- Optimize, find most violated constraint, introduce, and reoptimize.
- Repeat until no constraint in full problem violated by more than some tolerance!

Structural SVM Learner

Input: $(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_n, \mathbf{y}_n)$, C, ϵ
$S_i \leftarrow \emptyset$ for all $i = 1, \ldots, n$
repeat
    for i=1, …, n **do**
        set up a cost function
        $H(\mathbf{y}) = \Delta(\mathbf{y}_i, \mathbf{y}) + \langle \mathbf{w}, \Psi(\mathbf{x}_i, \mathbf{y}) \rangle - \langle \mathbf{w}, \Psi(\mathbf{x}_i, \mathbf{y}_i) \rangle$
        compute $\hat{\mathbf{y}} = \operatorname{argmax}_{\mathbf{y} \in \mathcal{Y}} H(\mathbf{y})$
        compute $\xi_i = \max\{0, \max_{\mathbf{y} \in S_i} H(\mathbf{y})\}$
        if $H(\hat{\mathbf{y}}) > \xi_i + \epsilon$ **then**
            $S_i \leftarrow S_i \cup \{\hat{\mathbf{y}}\}$
            $\mathbf{w} \leftarrow$ solution to Q.P. with constraints for $\bigcup_i S_i$
        end if
    end for
until no S_i has changed during iteration

Structural SVM Learner

- Starts with no constraints for any of the *n* examples.

Input: $(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_n, \mathbf{y}_n)$, C, ϵ
$S_i \leftarrow \emptyset$ for all $i = 1, \ldots, n$
repeat
  for i=1, …, n **do**
    set up a cost function
    $H(\mathbf{y}) = \Delta(\mathbf{y}_i, \mathbf{y}) + \langle \mathbf{w}, \Psi(\mathbf{x}_i, \mathbf{y}) \rangle - \langle \mathbf{w}, \Psi(\mathbf{x}_i, \mathbf{y}_i) \rangle$
    compute $\hat{\mathbf{y}} = \mathrm{argmax}_{\mathbf{y} \in \mathcal{Y}} H(\mathbf{y})$
    compute $\xi_i = \max\{0, \max_{\mathbf{y} \in S_i} H(\mathbf{y})\}$
    if $H(\hat{\mathbf{y}}) > \xi_i + \epsilon$ **then**
      $S_i \leftarrow S_i \cup \{\hat{\mathbf{y}}\}$
      $\mathbf{w} \leftarrow$ solution to Q.P. with constraints for $\bigcup_i S_i$
    end if
  end for
until no S_i has changed during iteration

Structural SVM Learner

- Starts with no constraints for any of the *n* examples.
- Repeatedly pass through examples.

Input: $(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_n, \mathbf{y}_n)$, C, ϵ
$S_i \leftarrow \emptyset$ for all $i = 1, \ldots, n$
repeat
  for i=1, . . . , n **do**
    set up a cost function
    $H(\mathbf{y}) = \Delta(\mathbf{y}_i, \mathbf{y}) + \langle \mathbf{w}, \Psi(\mathbf{x}_i, \mathbf{y}) \rangle - \langle \mathbf{w}, \Psi(\mathbf{x}_i, \mathbf{y}_i) \rangle$
    compute $\hat{\mathbf{y}} = \operatorname{argmax}_{\mathbf{y} \in \mathcal{Y}} H(\mathbf{y})$
    compute $\xi_i = \max\{0, \max_{\mathbf{y} \in S_i} H(\mathbf{y})\}$
    if $H(\hat{\mathbf{y}}) > \xi_i + \epsilon$ **then**
      $S_i \leftarrow S_i \cup \{\hat{\mathbf{y}}\}$
      $\mathbf{w} \leftarrow$ solution to Q.P. with constraints for $\bigcup_i S_i$
    end if
  end for
until no S_i has changed during iteration

Structural SVM Learner

- Starts with no constraints for any of the *n* examples.
- Repeatedly pass through examples.
- Find output $\hat{y}$ associated with most violated constraint! (**Separation Oracle** / **Cutting Plane**)

Input: $(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_n, \mathbf{y}_n)$, C, ϵ
$S_i \leftarrow \emptyset$ for all $i = 1, \ldots, n$
repeat
  for i=1, . . . , n **do**
    set up a cost function
    $H(\mathbf{y}) = \Delta(\mathbf{y}_i, \mathbf{y}) + \langle \mathbf{w}, \Psi(\mathbf{x}_i, \mathbf{y}) \rangle - \langle \mathbf{w}, \Psi(\mathbf{x}_i, \mathbf{y}_i) \rangle$
    compute $\hat{\mathbf{y}} = \mathrm{argmax}_{\mathbf{y} \in \mathcal{Y}} H(\mathbf{y})$
    compute $\xi_i = \max\{0, \max_{\mathbf{y} \in S_i} H(\mathbf{y})\}$
    if $H(\hat{\mathbf{y}}) > \xi_i + \epsilon$ **then**
      $S_i \leftarrow S_i \cup \{\hat{\mathbf{y}}\}$
      $\mathbf{w} \leftarrow$ solution to Q.P. with constraints for $\bigcup_i S_i$
    end if
  end for
until no S_i has changed during iteration

Structural SVM Learner

- Starts with no constraints for any of the *n* examples.
- Repeatedly pass through examples.
- Find output ŷ associated with most violated constraint! (**Separation Oracle / Cutting Plane**)
- If the constraint is violated more than ϵ, introduce the constraint and reoptimize.

Input: $(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_n, \mathbf{y}_n)$, C, ϵ

$S_i \leftarrow \emptyset$ for all $i = 1, \ldots, n$

repeat

  for i=1, . . . , n **do**

    set up a cost function

    $H(\mathbf{y}) = \Delta(\mathbf{y}_i, \mathbf{y}) + \langle \mathbf{w}, \Psi(\mathbf{x}_i, \mathbf{y}) \rangle - \langle \mathbf{w}, \Psi(\mathbf{x}_i, \mathbf{y}_i) \rangle$

    compute $\hat{\mathbf{y}} = \mathrm{argmax}_{\mathbf{y} \in \mathcal{Y}} H(\mathbf{y})$

    compute $\xi_i = \max\{0, \max_{\mathbf{y} \in S_i} H(\mathbf{y})\}$

    if $H(\hat{\mathbf{y}}) > \xi_i + \epsilon$ **then**

      $S_i \leftarrow S_i \cup \{\hat{\mathbf{y}}\}$

      $\mathbf{w} \leftarrow$ solution to Q.P. with constraints for $\bigcup_i S_i$

    end if

  end for

until no S_i has changed during iteration

Structural SVM Learner

- Starts with no constraints for any of the *n* examples.
- Repeatedly pass through examples.
- Find output $\hat{y}$ associated with most violated constraint! (**Separation Oracle** / **Cutting Plane**)
- If the constraint is violated more than ϵ, introduce the constraint and reoptimize.
- Stops when no constraints introduced in a pass.

Input: $(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_n, \mathbf{y}_n)$, C, ϵ

$S_i \leftarrow \emptyset$ for all $i = 1, \ldots, n$

repeat

    for i=1, ..., n **do**

        set up a cost function

        $H(\mathbf{y}) = \Delta(\mathbf{y}_i, \mathbf{y}) + \langle \mathbf{w}, \Psi(\mathbf{x}_i, \mathbf{y}) \rangle - \langle \mathbf{w}, \Psi(\mathbf{x}_i, \mathbf{y}_i) \rangle$

        compute $\hat{\mathbf{y}} = \operatorname{argmax}_{\mathbf{y} \in \mathcal{Y}} H(\mathbf{y})$

        compute $\xi_i = \max\{0, \max_{\mathbf{y} \in S_i} H(\mathbf{y})\}$

        if $H(\hat{\mathbf{y}}) > \xi_i + \epsilon$ **then**

            $S_i \leftarrow S_i \cup \{\hat{\mathbf{y}}\}$

            $\mathbf{w} \leftarrow$ solution to Q.P. with constraints for $\bigcup_i S_i$

        end if

    end for

until no S_i has changed during iteration

Important Theoretical Properties

- **Polynomial Time Termination**: Terminates in polynomial number of iterations.
- **Correctness**: Returns solution to full QP accurate to desired ε.
- **Empirical Risk Bound**: Slack term upper bounds empirical risk.

$$\min_{\mathbf{w},\xi} \frac{1}{2}\|\mathbf{w}\|^2 + \frac{C}{n}\sum_{i=1}^{n}\xi_i$$

$$\text{s.t.} \quad \forall i : \xi_i \geq 0$$

$$\forall i, \forall \mathbf{y} \in \mathcal{Y} : \langle \mathbf{w}, \Psi(\mathbf{x}_i, \mathbf{y}_i)\rangle - \langle \mathbf{w}, \Psi(\mathbf{x}_i, \mathbf{y})\rangle \geq \Delta(\mathbf{y}_i, \mathbf{y}) - \xi_i$$

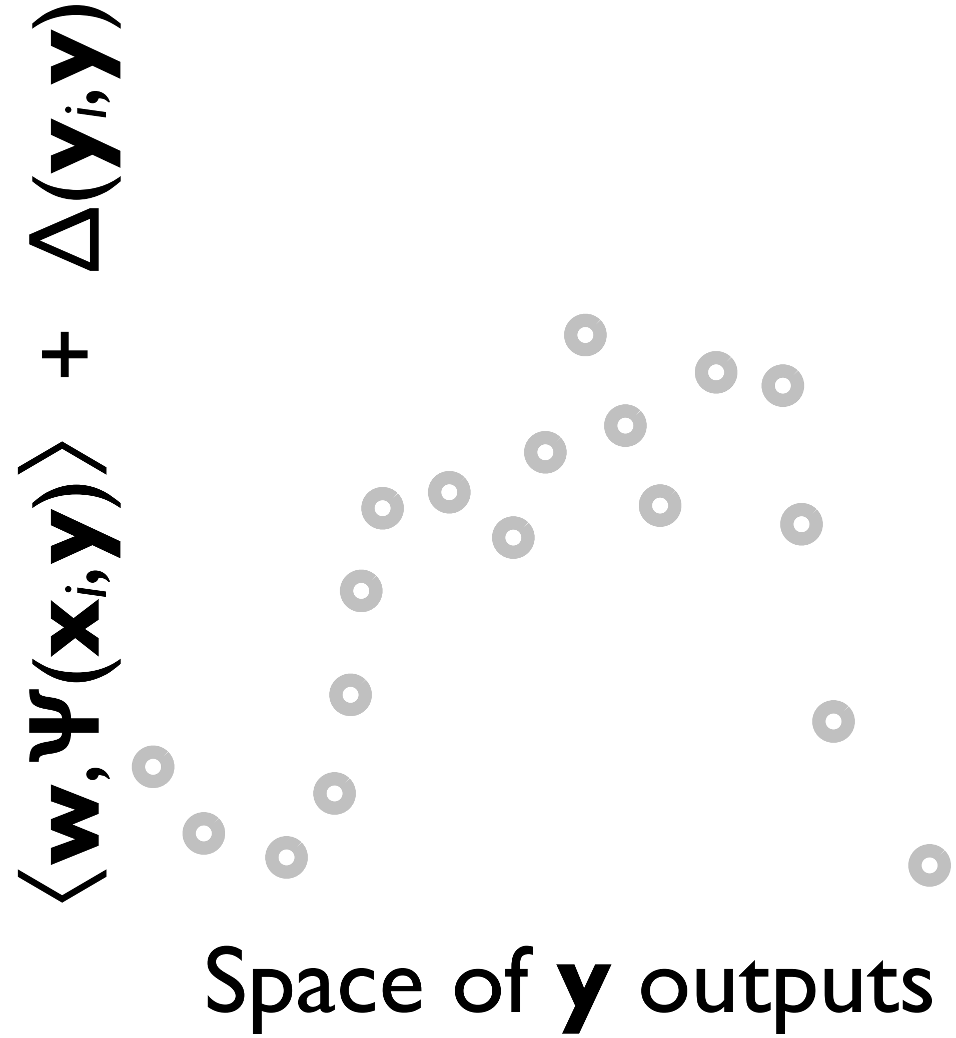
Input: $(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_n, \mathbf{y}_n)$, C, ϵ
$S_i \leftarrow \emptyset$ for all $i = 1, \ldots, n$
repeat
  for i=1, …, n **do**
    set up a cost function
    $H(\mathbf{y}) = \Delta(\mathbf{y}_i, \mathbf{y}) + \langle \mathbf{w}, \Psi(\mathbf{x}_i, \mathbf{y})\rangle - \langle \mathbf{w}, \Psi(\mathbf{x}_i, \mathbf{y}_i)\rangle$
    compute $\hat{\mathbf{y}} = \operatorname{argmax}_{\mathbf{y}\in\mathcal{Y}} H(\mathbf{y})$
    compute $\xi_i = \max\{0, \max_{\mathbf{y}\in S_i} H(\mathbf{y})\}$
    if $H(\hat{\mathbf{y}}) > \xi_i + \epsilon$ **then**
      $S_i \leftarrow S_i \cup \{\hat{\mathbf{y}}\}$
      $\mathbf{w} \leftarrow$ solution to Q.P. with constraints for $\bigcup_i S_i$
    end if
  end for
until no S_i has changed during iteration

Talk Outline

- Structured Prediction
- Structural SVMs (SSVMs)
- Approximate Inference in SSVMs
    - Theoretical Analysis
    - Empirical Analysis

Approximations

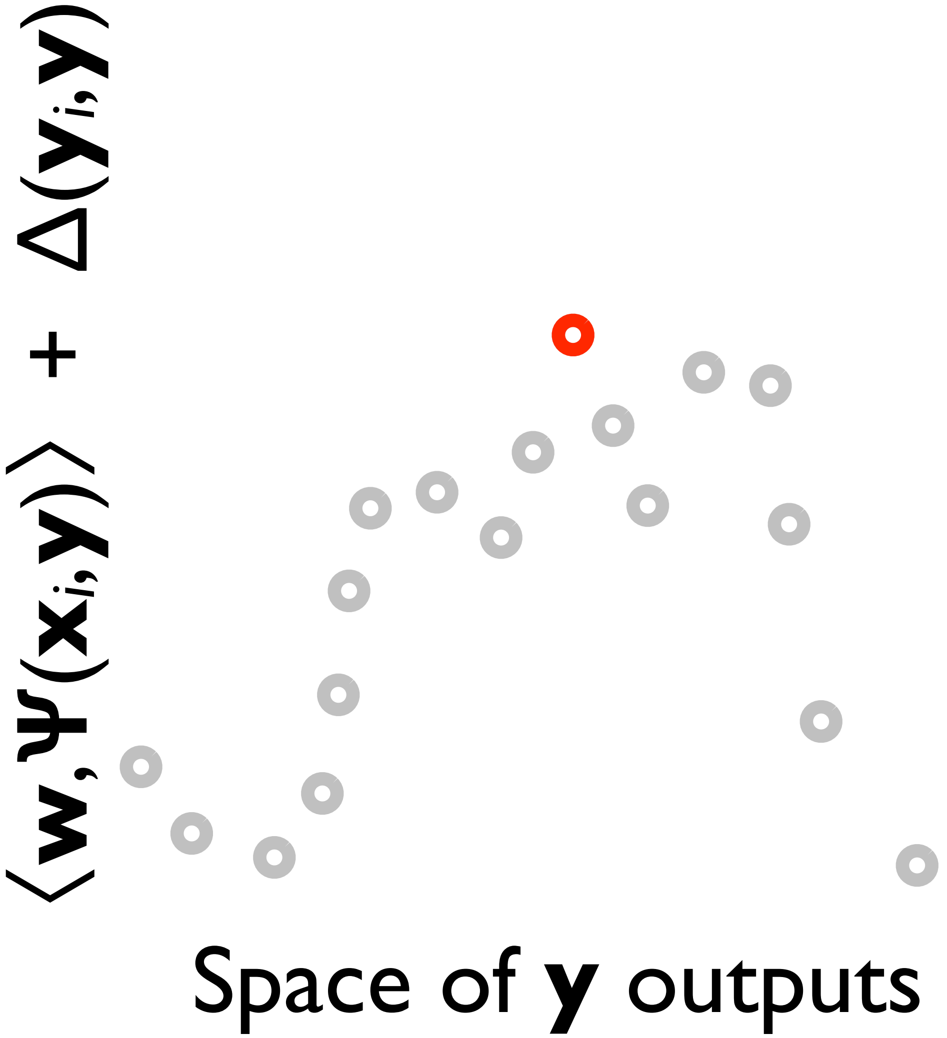
$$\hat{\mathbf{y}} = \operatorname*{argmax}_{\mathbf{y} \in \mathcal{Y}} \langle \mathbf{w}, \Psi(\mathbf{x}_i, \mathbf{y}) \rangle + \Delta(\mathbf{y}_i, \mathbf{y})$$



Approximations

$$\hat{\mathbf{y}} = \operatorname*{argmax}_{\mathbf{y} \in \mathcal{Y}} \langle \mathbf{w}, \Psi(\mathbf{x}_i, \mathbf{y}) \rangle + \Delta(\mathbf{y}_i, \mathbf{y})$$

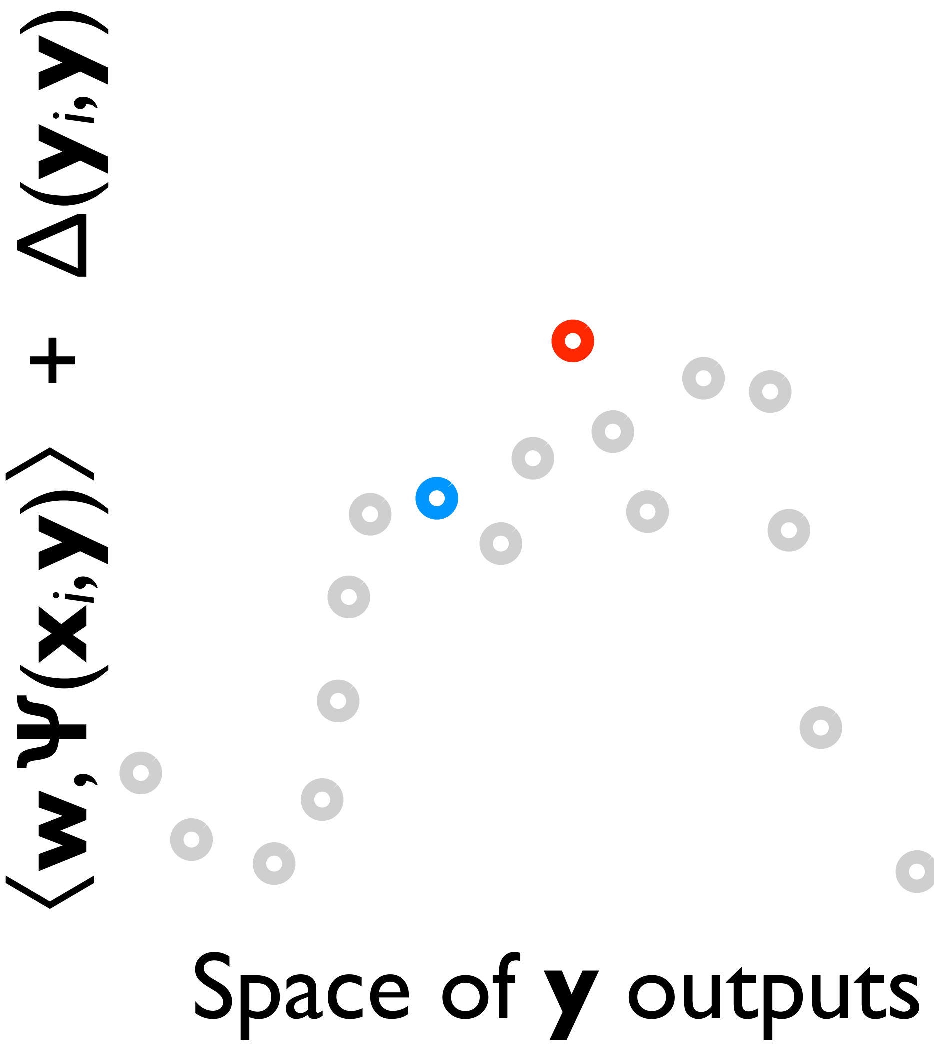
- **Exact**: Finds actual maximizing $\hat{\mathbf{y}}$.



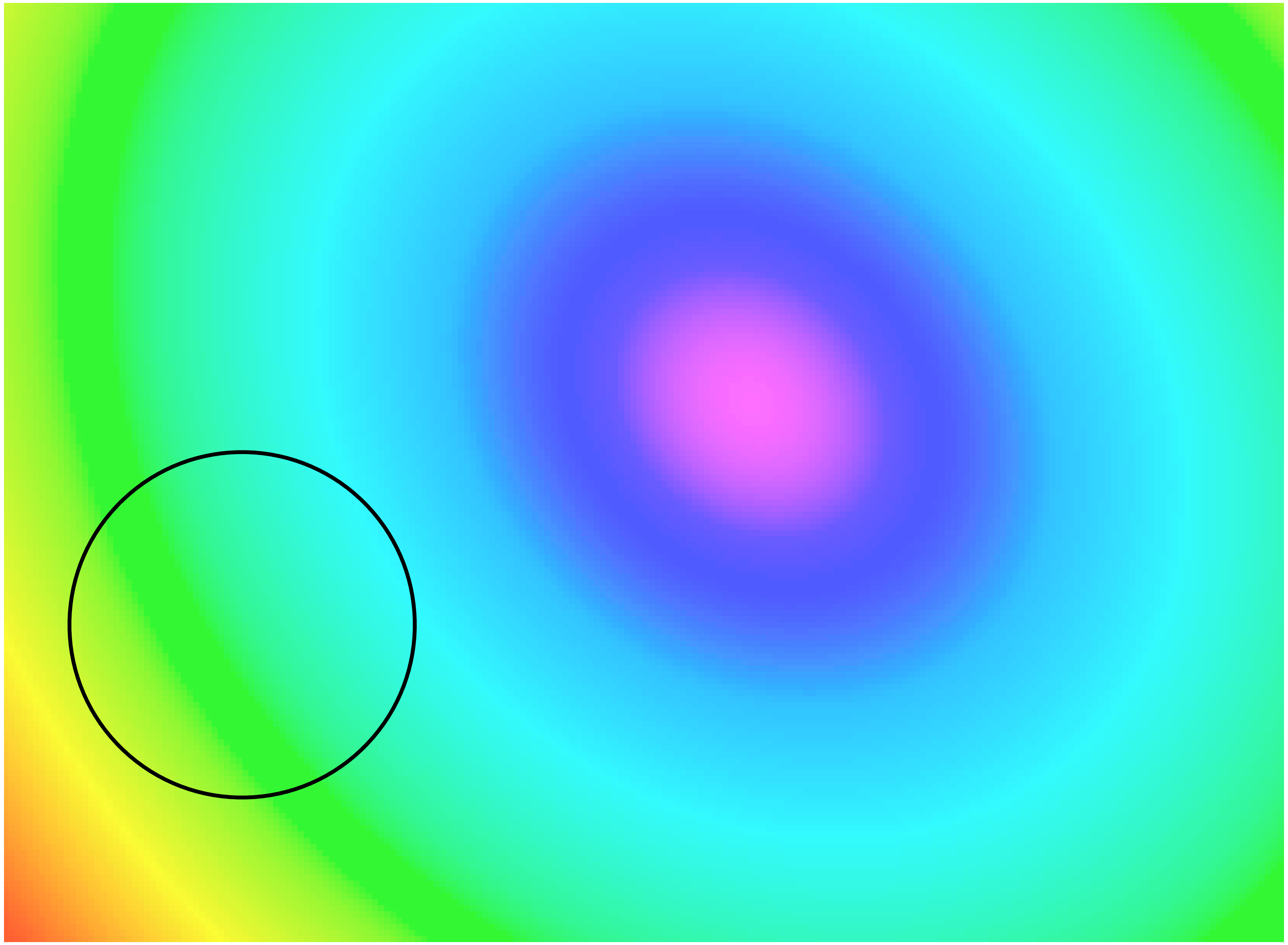
Approximations

$$\hat{\mathbf{y}} = \underset{\mathbf{y}\in\mathcal{Y}}{\operatorname{argmax}} \langle \mathbf{w}, \Psi(\mathbf{x}_i, \mathbf{y})\rangle + \Delta(\mathbf{y}_i, \mathbf{y})$$

- **Exact**: Finds actual maximizing $\hat{\mathbf{y}}$.
- **Undergenerating Approximations**: Finds possibly suboptimal $\hat{\mathbf{y}}$ from search space, i.e., some form of local search.

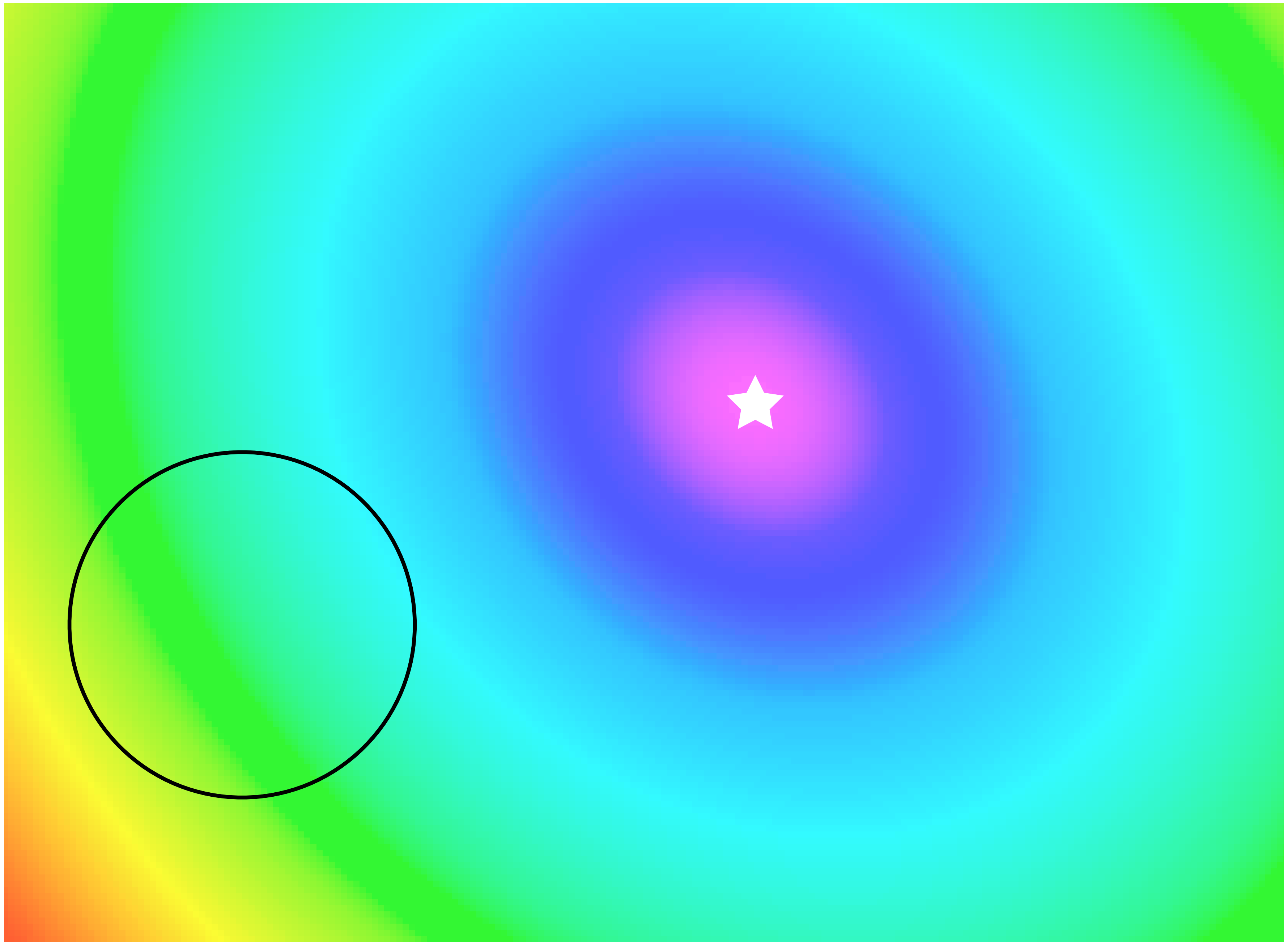


Cutting Plane Example



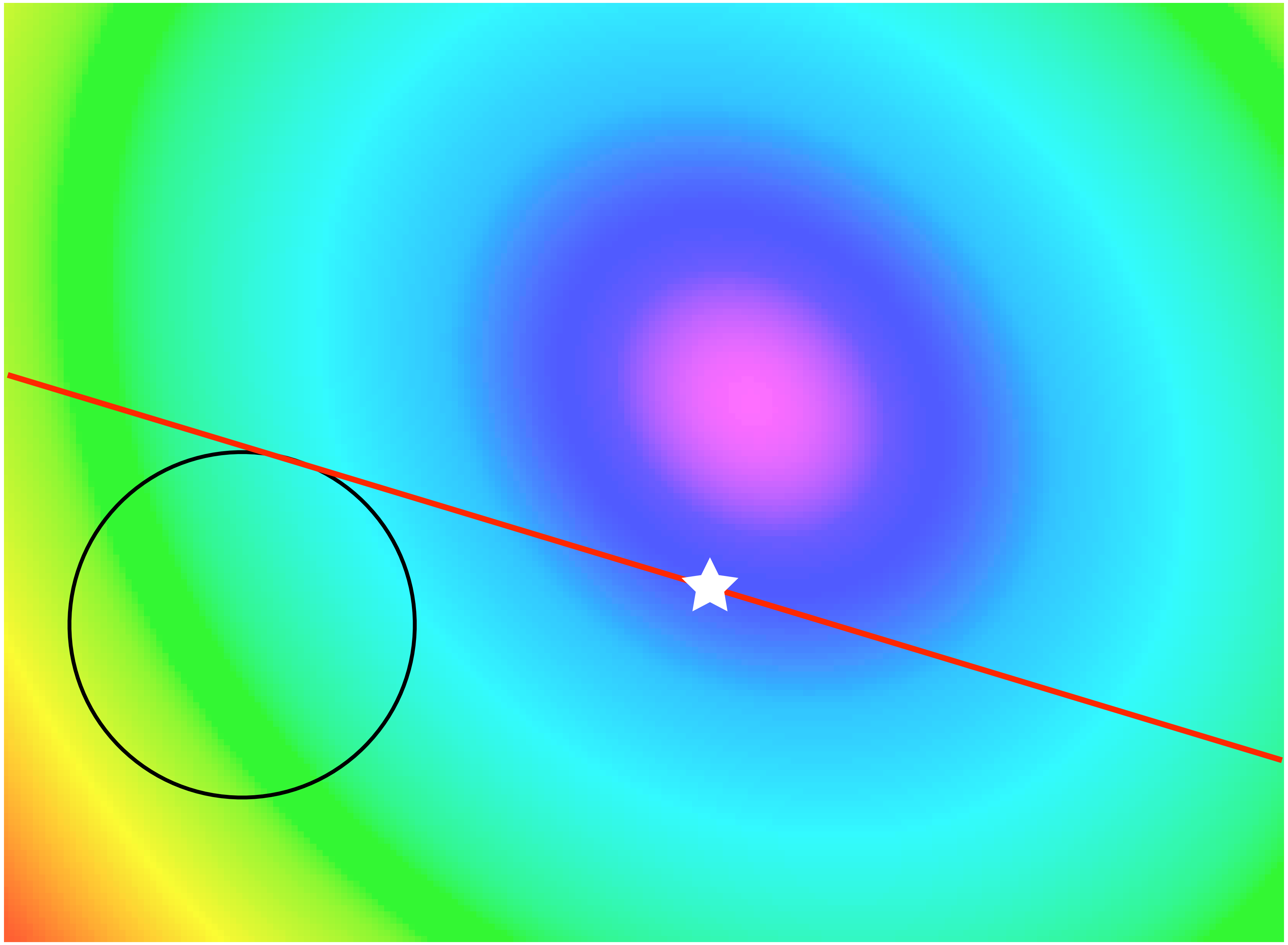
- Suppose you cannot find the most violated constraint.
- Theory depends upon finding *the* most violated constraint.
- Ability to find feasible point compromised.

Cutting Plane Example



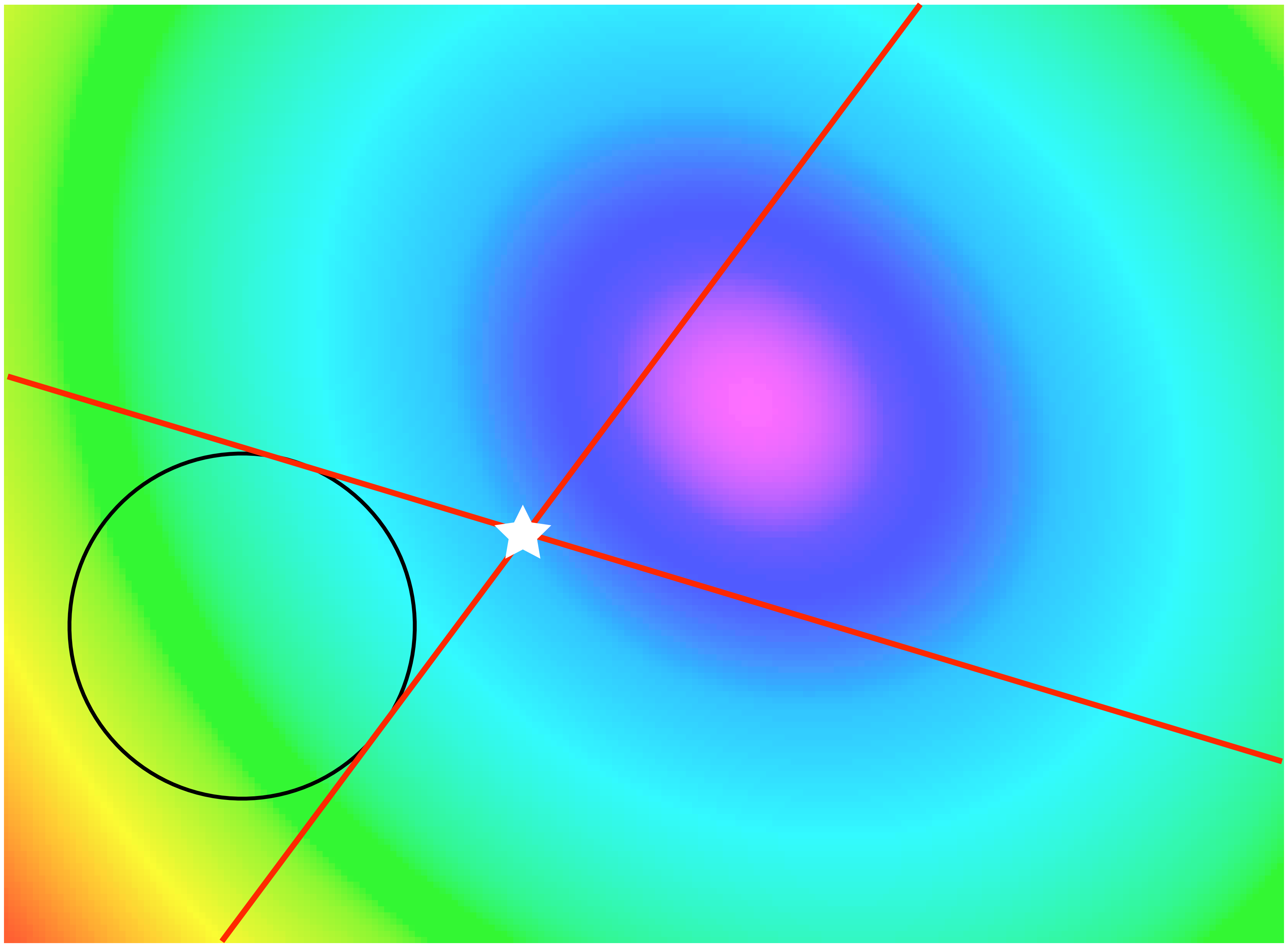
- Suppose you cannot find the most violated constraint.
- Theory depends upon finding *the* most violated constraint.
- Ability to find feasible point compromised.

Cutting Plane Example



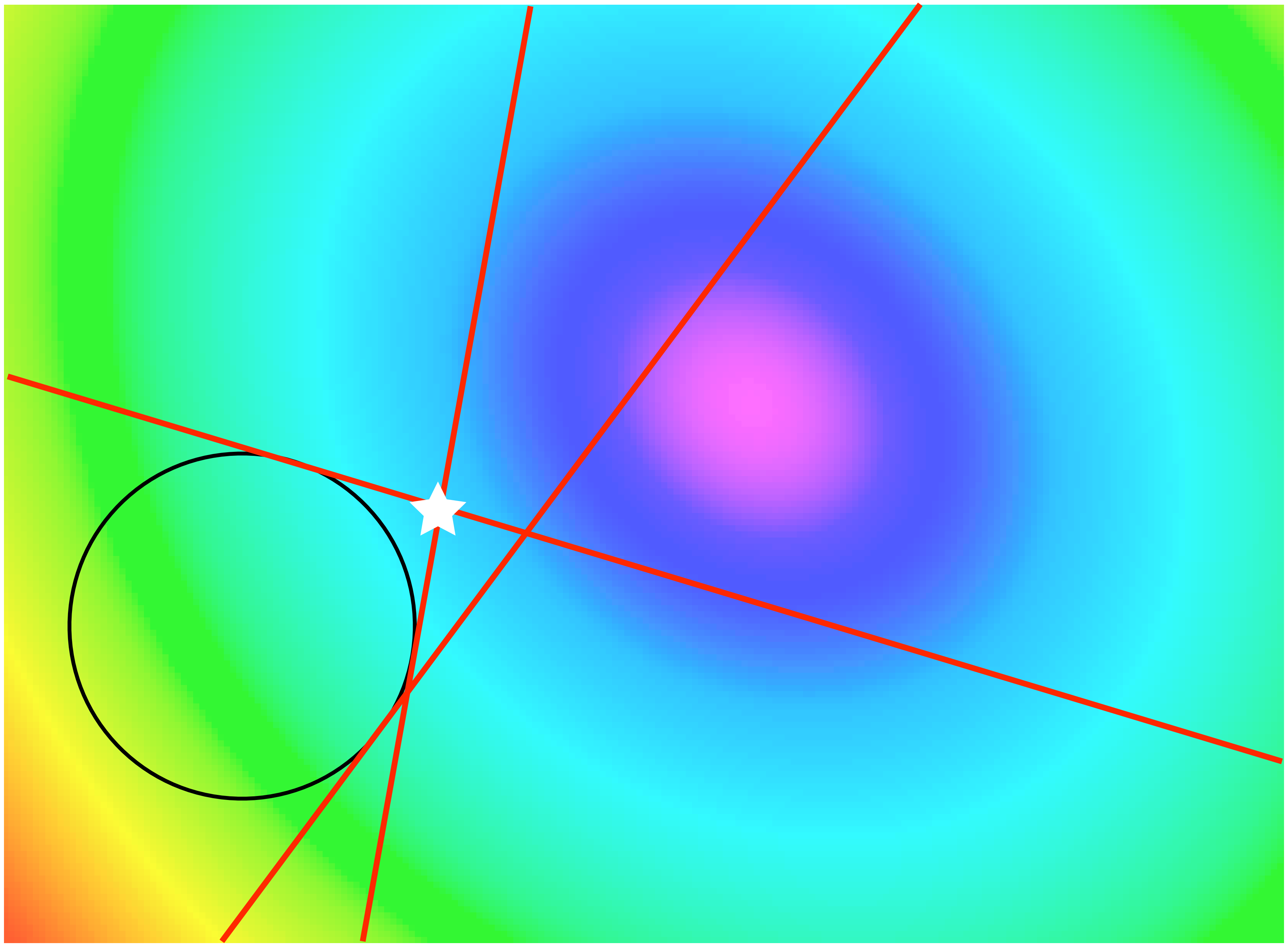
- Suppose you cannot find the most violated constraint.
- Theory depends upon finding *the* most violated constraint.
- Ability to find feasible point compromised.

Cutting Plane Example



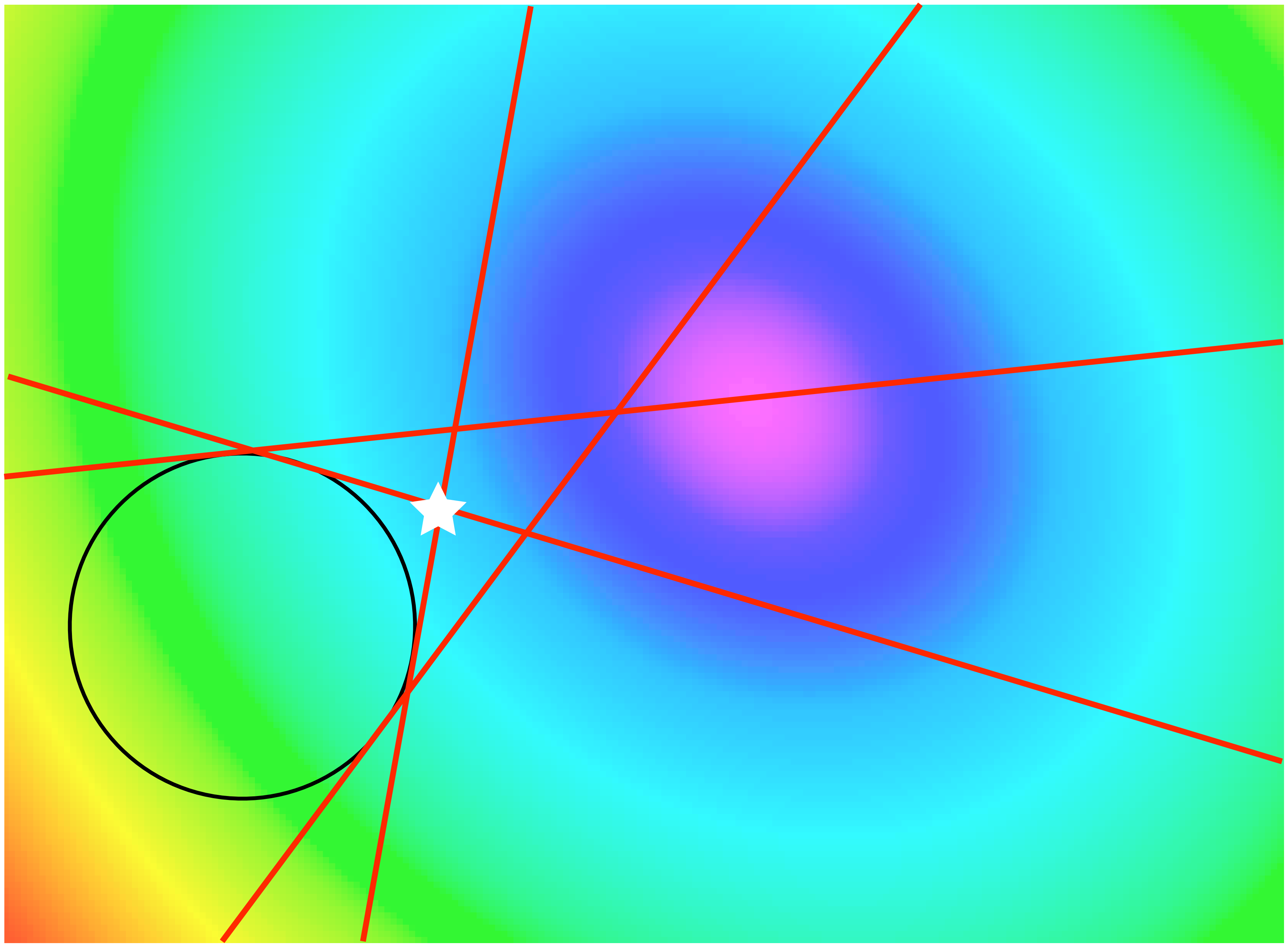
- Suppose you cannot find the most violated constraint.
- Theory depends upon finding *the* most violated constraint.
- Ability to find feasible point compromised.

Cutting Plane Example



- Suppose you cannot find the most violated constraint.
- Theory depends upon finding *the* most violated constraint.
- Ability to find feasible point compromised.

Cutting Plane Example



- Suppose you cannot find the most violated constraint.
- Theory depends upon finding *the* most violated constraint.
- Ability to find feasible point compromised.

Undergenerating Approximations

- **Polynomial Time Termination**: Yes, bound indifferent to quality of approximation.
- **Correctness**: No, some constraints in full QP may remain unfound.
- **Empirical Risk Bound**: No, same reason.

Undergenerating ρ-Approximations

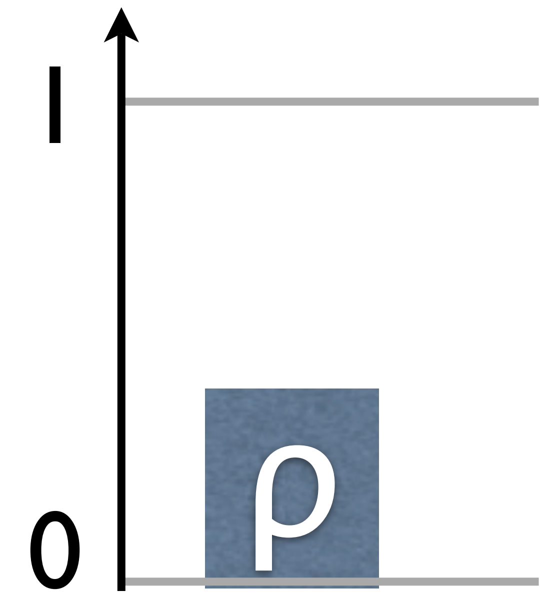
- Restrict attention to make theoretical statements
- ρ-Approximation finds $\hat{\mathbf{y}}$ such that $\hat{f} \geq \rho f^*$

  where $\hat{f} = \langle \mathbf{w}, \boldsymbol{\Psi}(\mathbf{x}_i, \hat{\mathbf{y}}) \rangle + \Delta(\mathbf{y}_i, \hat{\mathbf{y}})$

  where $f^* = \langle \mathbf{w}, \boldsymbol{\Psi}(\mathbf{x}_i, \mathbf{y}^*) \rangle + \Delta(\mathbf{y}_i, \mathbf{y}^*)$

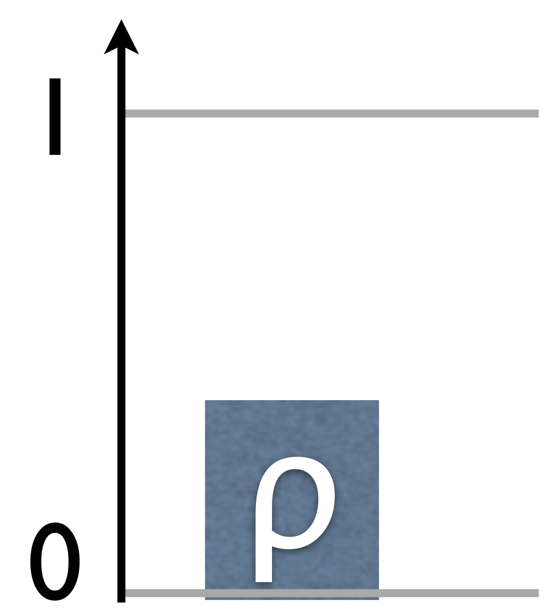
- Smaller ρ means worse approximation
- ρ=1 equivalent to exact inference

Undergenerating ρ-Approx Theorems



Undergenerating ρ-Approx Theorems

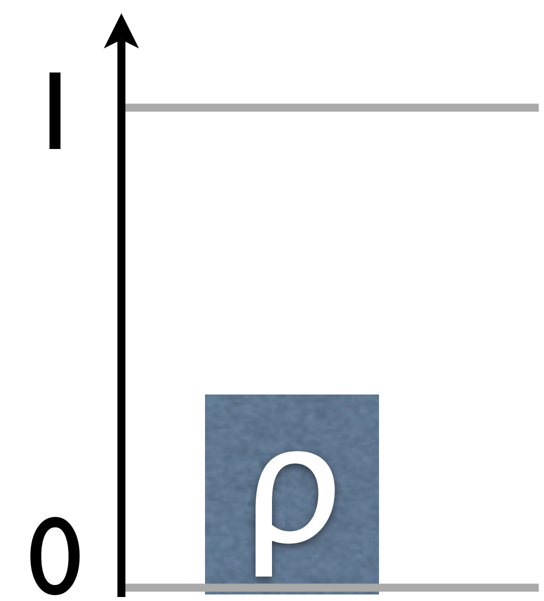
- **Three theorems**:



Undergenerating ρ-Approx Theorems

$\hat{\xi}$

- **Three theorems**:
  - “Required” slack $\hat{\xi}$ in iteration.

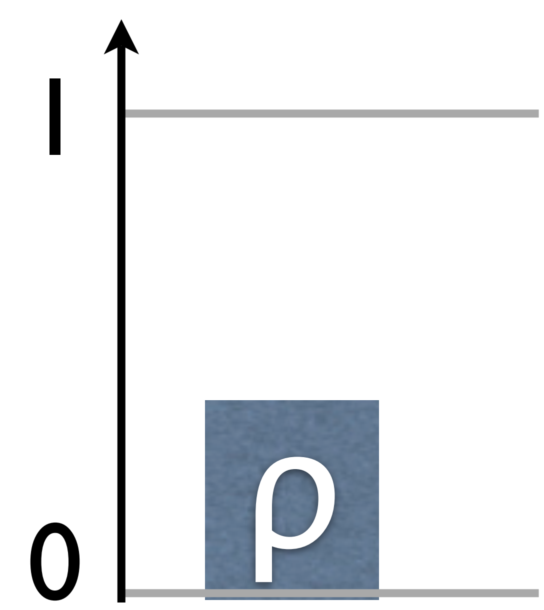


Undergenerating ρ-Approx Theorems

$\hat{\xi}$

$\frac{1}{2}\|\mathbf{w}\|^2 + C\xi$

- **Three theorems**:
  - "Required" slack $\hat{\xi}$ in iteration.
  - The objective $\frac{1}{2}\|\mathbf{w}\|^2 + C\xi$.



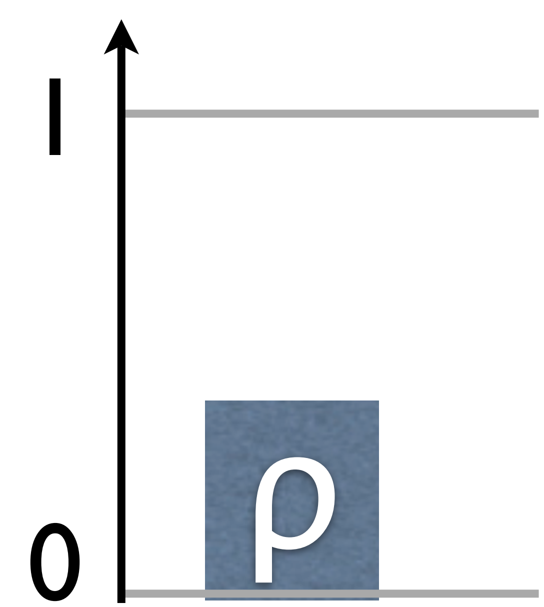
Undergenerating ρ-Approx Theorems

$$\hat{\xi}$$

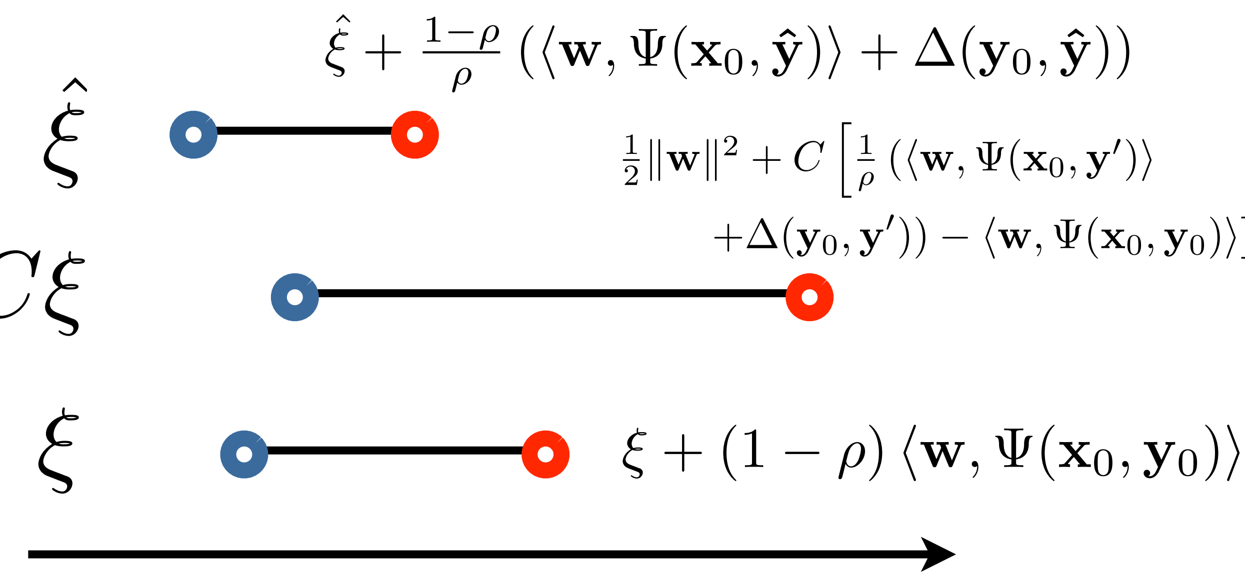
$$\frac{1}{2}\|\mathbf{w}\|^2 + C\xi$$

$$\xi$$

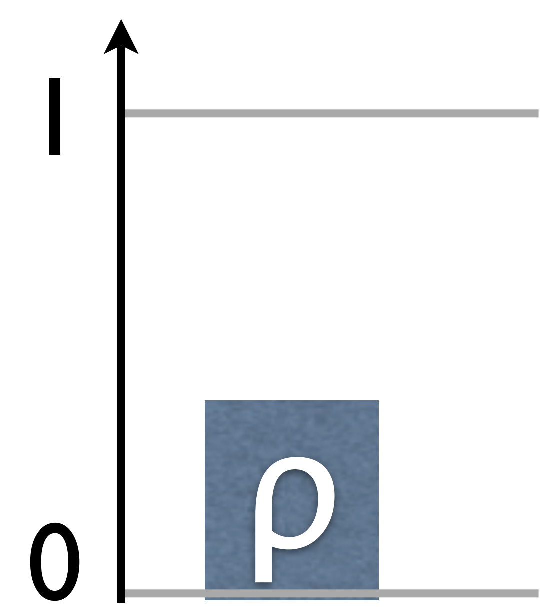
- **Three theorems**:
  - "Required" slack $\hat{\xi}$ in iteration.
  - The objective $\frac{1}{2}\|\mathbf{w}\|^2 + C\xi$.
  - Empirical risk bound ξ.



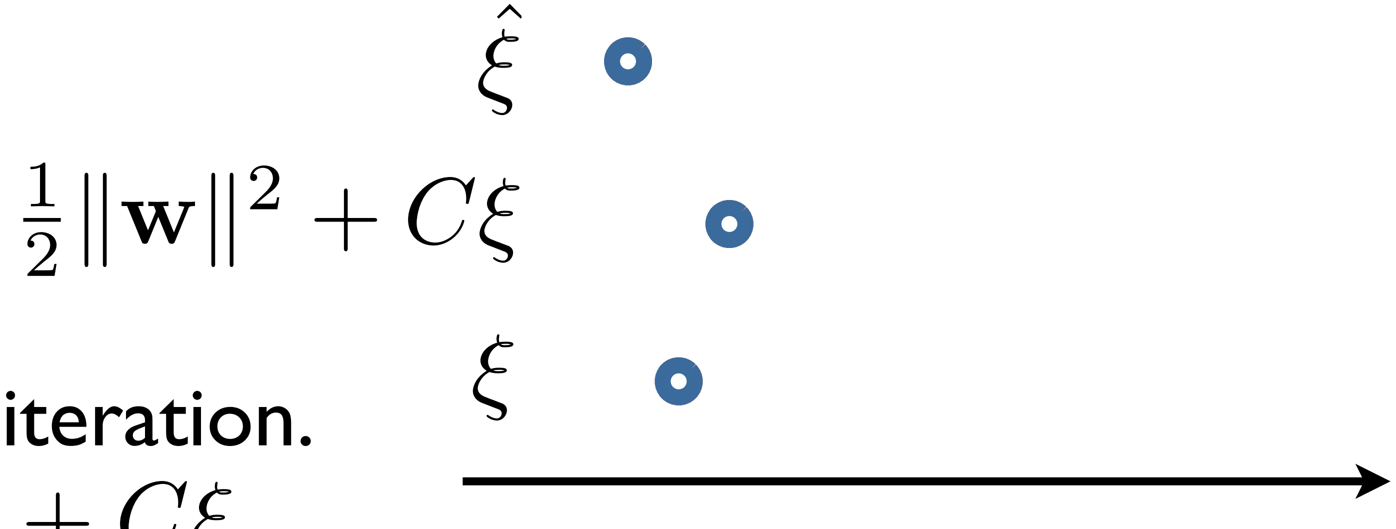
Undergenerating ρ-Approx Theorems



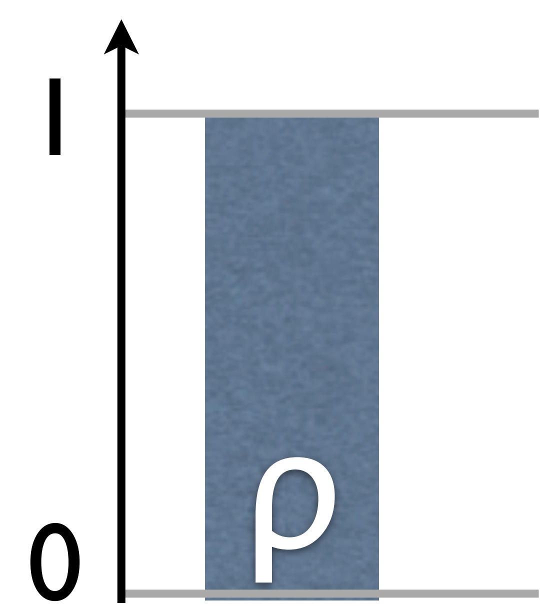
- **Three theorems**:
  - "Required" slack $\hat{\xi}$ in iteration.
  - The objective $\frac{1}{2}\|\mathbf{w}\|^2 + C\xi$.
  - Empirical risk bound ξ.
- True value for these quantities lies in interval between **found value**, and an **upper bound** depending on ρ.



Undergenerating ρ-Approx Theorems



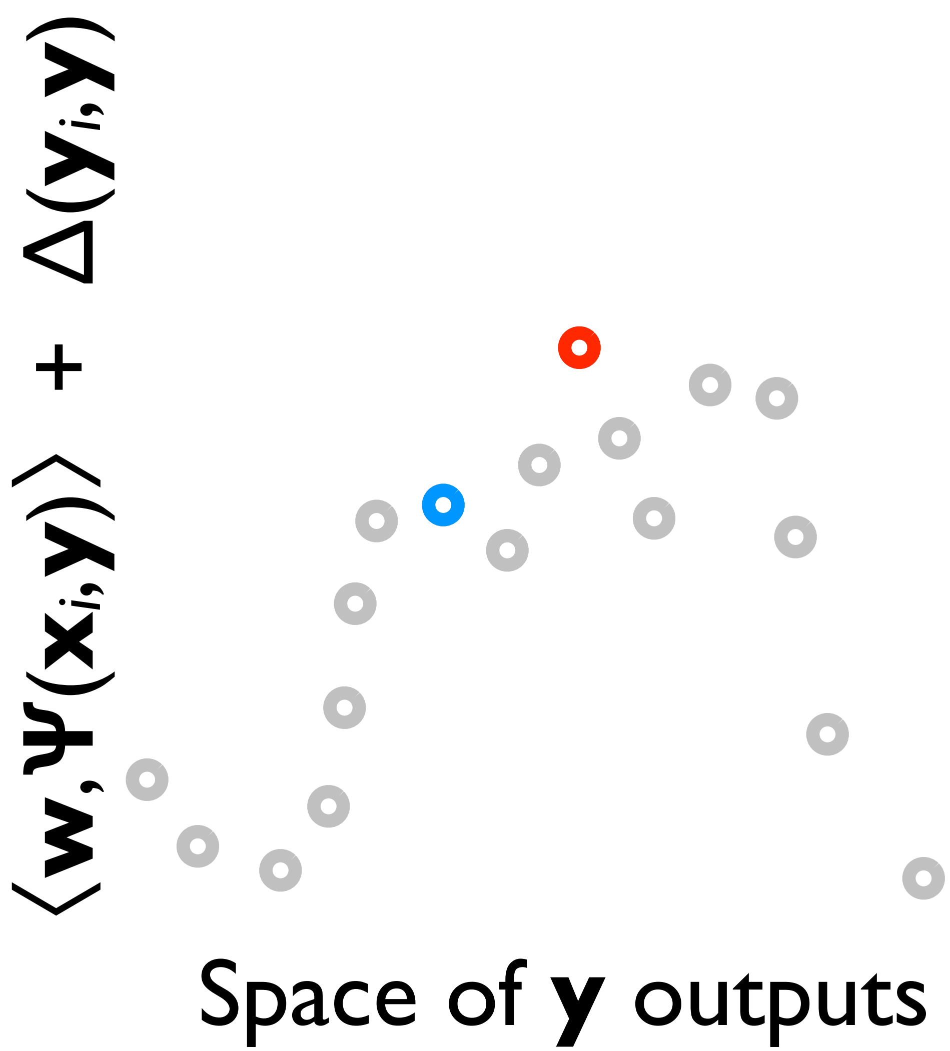
- **Three theorems**:
  - "Required" slack $\hat{\xi}$ in iteration.
  - The objective $\frac{1}{2}\|\mathbf{w}\|^2 + C\xi$.
  - Empirical risk bound ξ.
- True value for these quantities lies in interval between **found value**, and an **upper bound** depending on ρ.
- As ρ→1, interval is of size 0.



Approximations

$$\hat{\mathbf{y}} = \operatorname*{argmax}_{\mathbf{y} \in \mathcal{Y}} \langle \mathbf{w}, \Psi(\mathbf{x}_i, \mathbf{y}) \rangle + \Delta(\mathbf{y}_i, \mathbf{y})$$

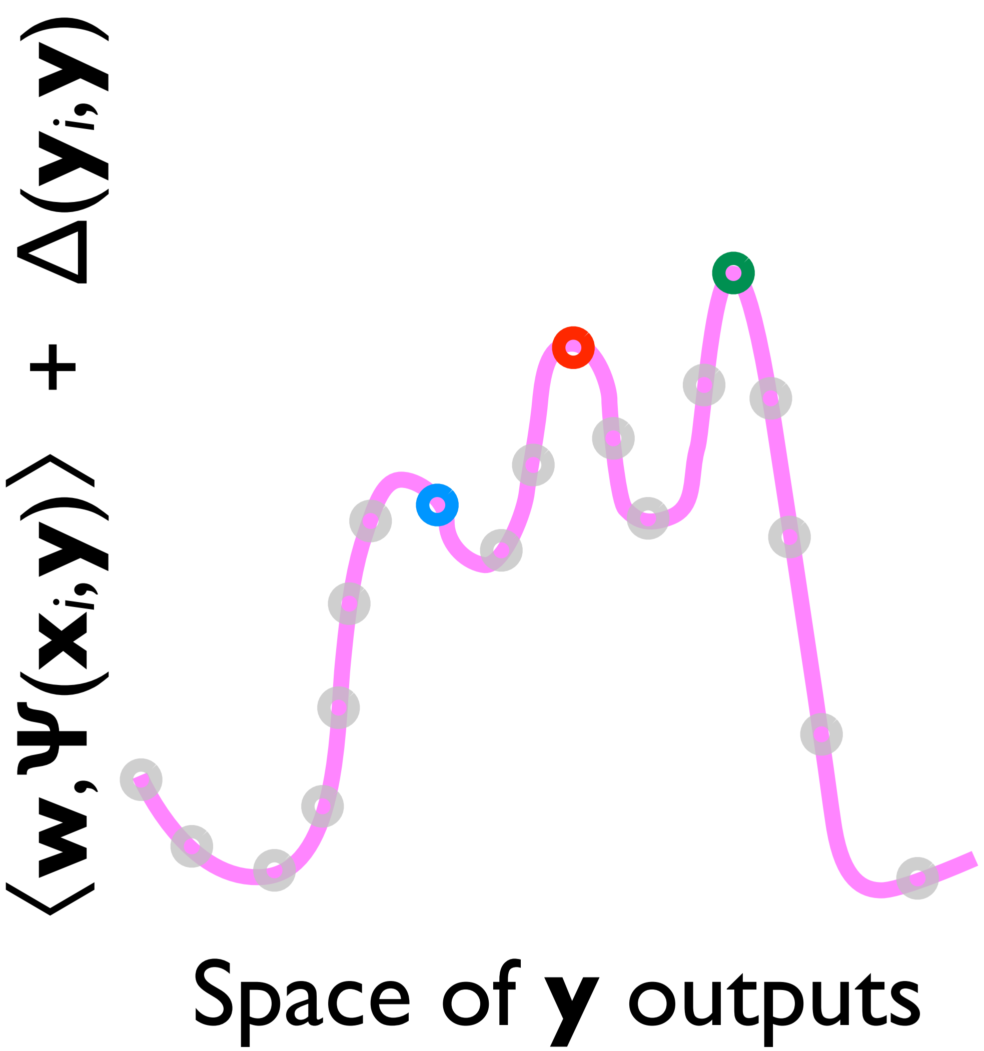
- **Exact**: Finds actual maximizing $\hat{\mathbf{y}}$.
- **Undergenerating Approximations**: Finds possibly suboptimal $\hat{\mathbf{y}}$ from search space, i.e., some form of local search.



Approximations

$$\hat{\mathbf{y}} = \operatorname*{argmax}_{\mathbf{y}\in\mathcal{Y}} \langle \mathbf{w}, \Psi(\mathbf{x}_i, \mathbf{y})\rangle + \Delta(\mathbf{y}_i, \mathbf{y})$$

- **Exact**: Finds actual maximizing $\hat{\mathbf{y}}$.
- **Undergenerating Approximations**: Finds possibly suboptimal $\hat{\mathbf{y}}$ from search space, i.e., some form of local search.
- **Overgenerating Approximations**: Finds optimal $\hat{\mathbf{y}}$, but only by virtue of expanding the search space so original search space is a subset, e.g., relaxations.



Overgenerating Approx Theory in a Nutshell

- **Polynomial Time Termination**: Yes, assuming Ψ lengths and Δ remain bounded.
- **Correctness**: Yes, the solution that is found is feasible in the full QP. (Though not necessarily optimal.)
- **Empirical Risk Bound**: Yes, since all constraints in full QP respected. (Though the bound may be weaker.)

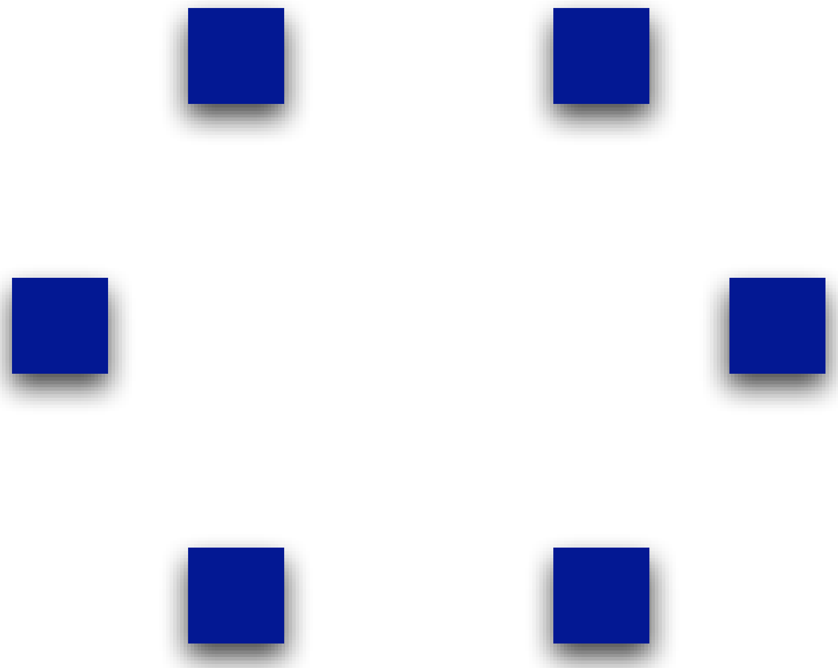
Talk Outline

- Structured Prediction
- Structural SVMs (SSVMs)
- Approximate Inference in SSVMs
  - Theoretical Analysis
  - Empirical Analysis

Our Testbed:
Binary Pairwise MRFs

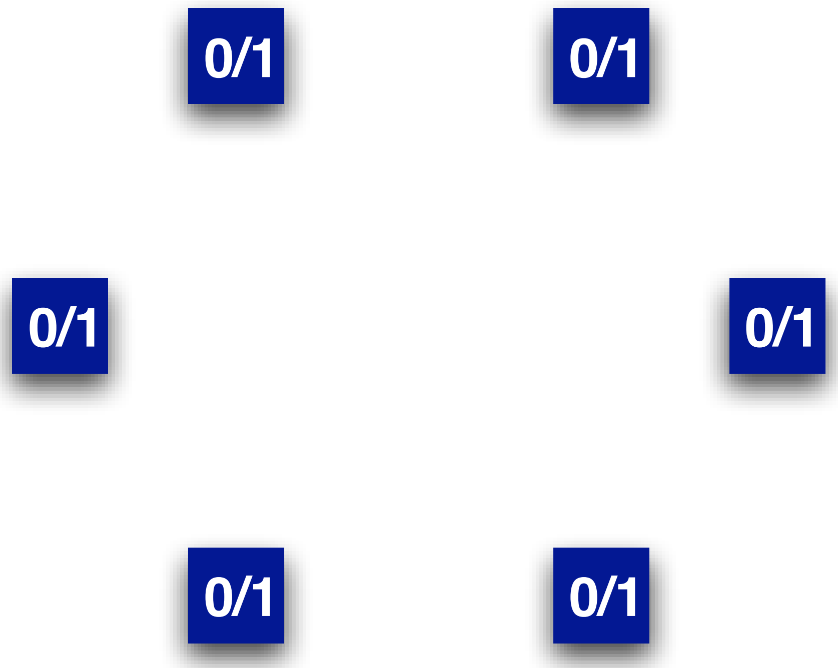
Our Testbed: Binary Pairwise MRFs

- Markov random field.



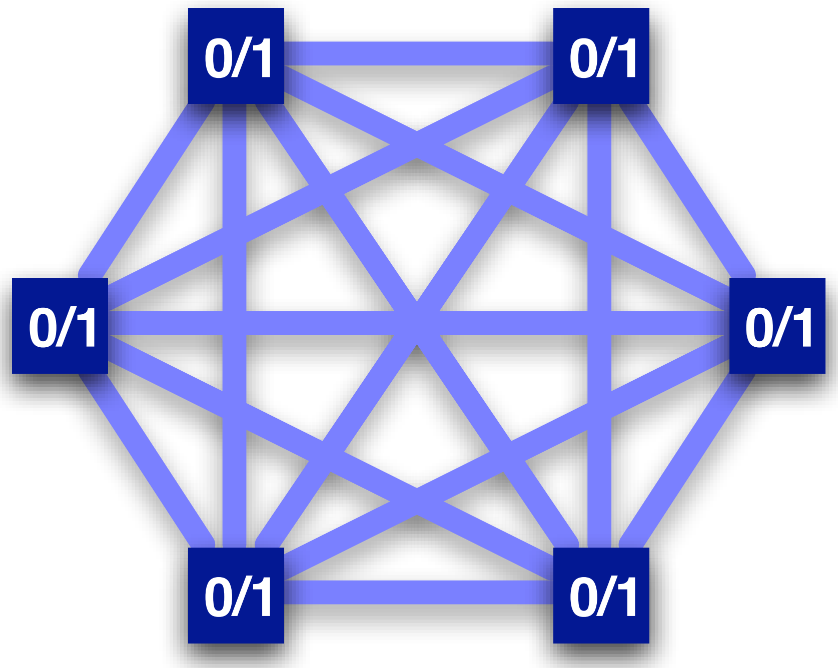
Our Testbed: Binary Pairwise MRFs

- Markov random field.
- Node variables may take binary values (0,1).



Our Testbed: Binary Pairwise MRFs

- Markov random field.
- Node variables may take binary values (0,1).
- Completely connected.



Application: Multilabel Classification

- **Task**: For input **x**, output set of relevant labels **y** from finite set of labels.
- **MRF**: Nodes represent labels. If has 1 value, label is on.
  - **Node potentials**: Input **x**'s tendency to have label.
  - **Edge potentials**: Two labels' tendency to co-occur.
- **Model**: One hyperplane within **w** for each label. A single value within **w** for each pair of labels.
- **Loss**: $\Delta(\mathbf{y},\bar{\mathbf{y}})$ counts proportion of different labels.

Training/Predictive Inference

- **Prediction**: MAP inference on the MRF inferred from example **x** and model **w**.

$$h(\mathbf{x}) = \operatorname*{argmax}_{\mathbf{y} \in \mathcal{Y}} \langle \mathbf{w}, \Psi(\mathbf{x}, \mathbf{y}) \rangle$$

- **Training**: Finding most violated constraint for ($\mathbf{x}_i$,$\mathbf{y}_i$) very similar, except with modified node potentials to incorporate loss.

$$\hat{\mathbf{y}} = \operatorname*{argmax}_{\mathbf{y} \in \mathcal{Y}} \langle \mathbf{w}, \Psi(\mathbf{x}_i, \mathbf{y}) \rangle + \Delta(\mathbf{y}_i, \mathbf{y})$$

- Both can utilize same inference techniques.

Datasets

Dataset	Labels	Train	Test	Feats.	**w** Size
Scene	6	1211	1196	294	1779
Yeast	14	1500	917	103	1533
Mediamill	10	29415	12168	120	1245
Reuters	10	2916	2914	47236	472405
Synth1	6	471	5045	6000	36015
Synth2	10	1000	10000	40	445

- Real data from LIBSVM multilabel dataset page: **Scene**, **Yeast**, **Reuters**, **Mediamill**.
  - **Reuters** and **Mediamill**: Selected 10 most frequent labels.
- Two synthetic datasets:
  - **Synth1**: Pairwise potentials unneeded to learn underlying concept (but could make learning easier if exploited).
  - **Synth2**: Pairwise potentials are needed.

Undergenerating Approximations

- **Greedy**: Makes single value assignment by what most increases discriminant function.
- **LBP**: Loopy belief propagation.
- **Combine**: Run greedy and LBP, return best.

Overgenerating Approximations

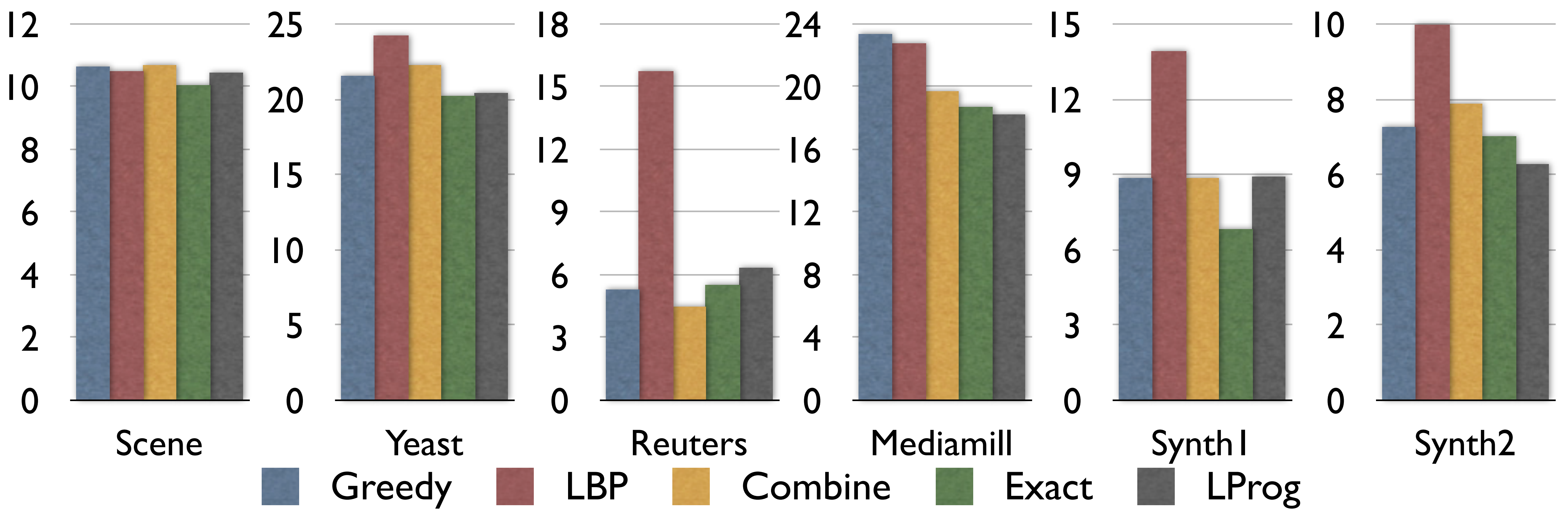
- **LProg**: Based on ILP encoding of MAP inference, subsequently relaxed.
- **Cuts**: Relaxation based on graph cut inference.
  - Both really equivalent -- cuts much faster.

Third Algorithm Class, for Comparison Only

- **Exact**: Constrained our problems so exact inference through exhaustive enumeration was reasonable. (“Best” one could do)

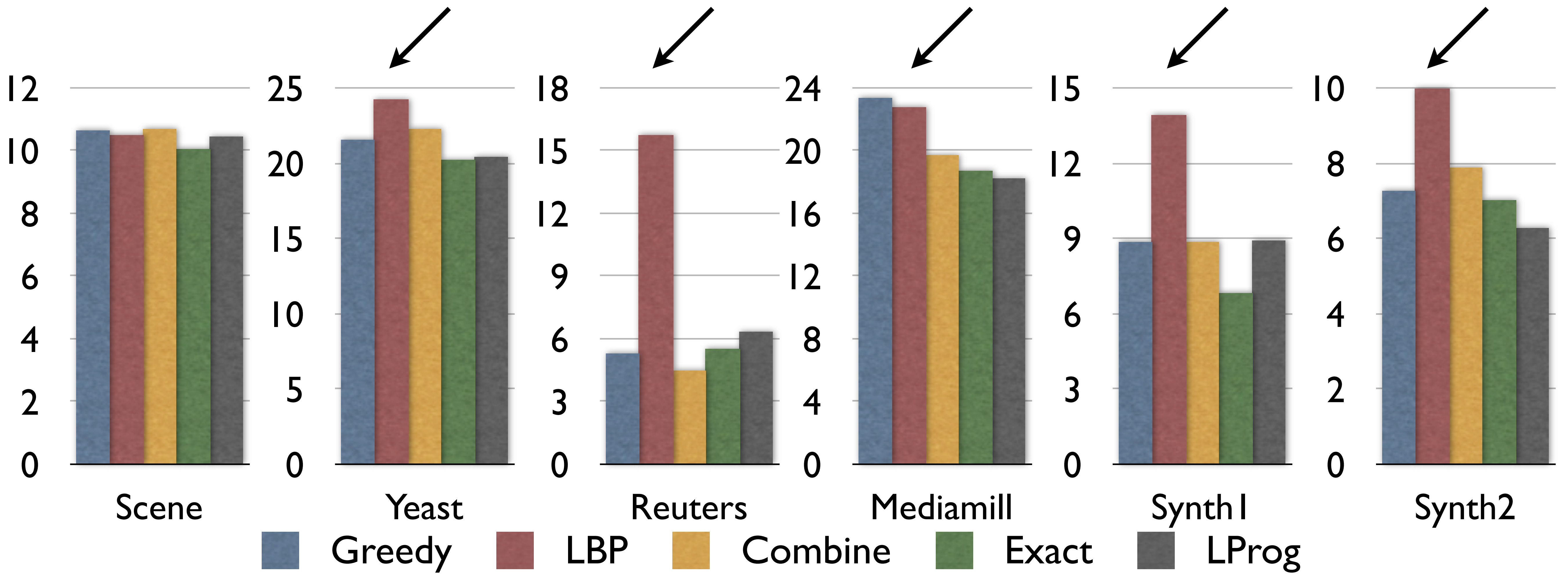
The Sorry State of LBP

- Losses on the six datasets (lower is better).
- Five inference methods used to train and evaluate models.
- **LBP** seems to do pretty poorly!



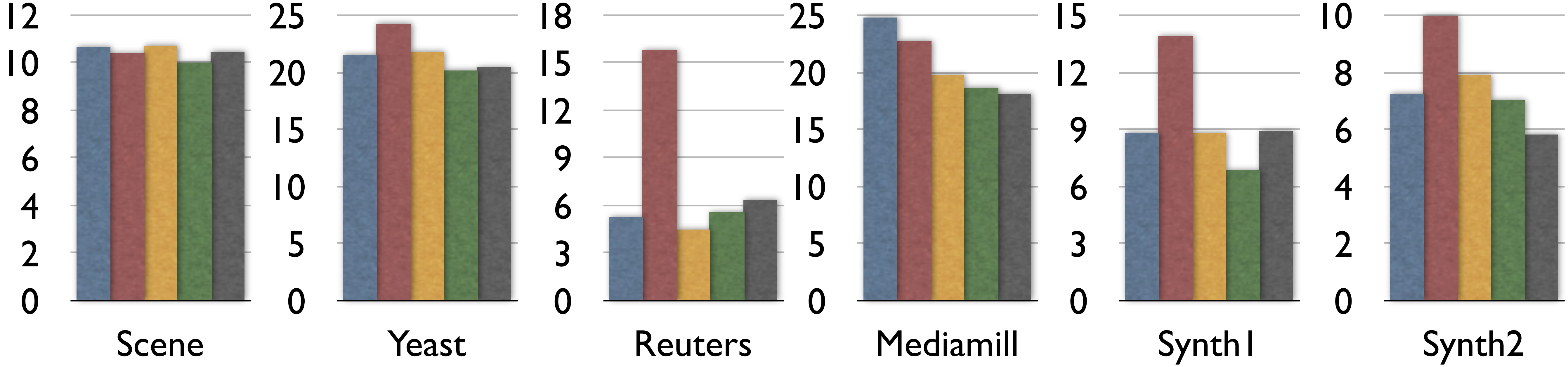
The Sorry State of LBP

- Losses on the six datasets (lower is better).
- Five inference methods used to train and evaluate models.
- **LBP** seems to do pretty poorly!

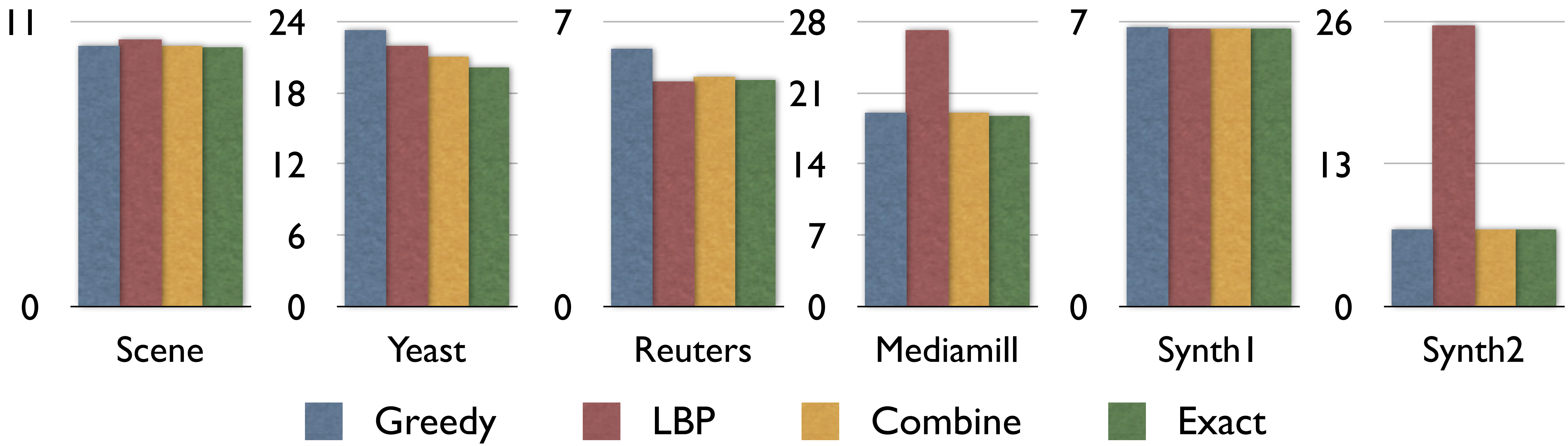


The Sorry State of LBP

Bad as a training method (all predicted with **Exact**)...

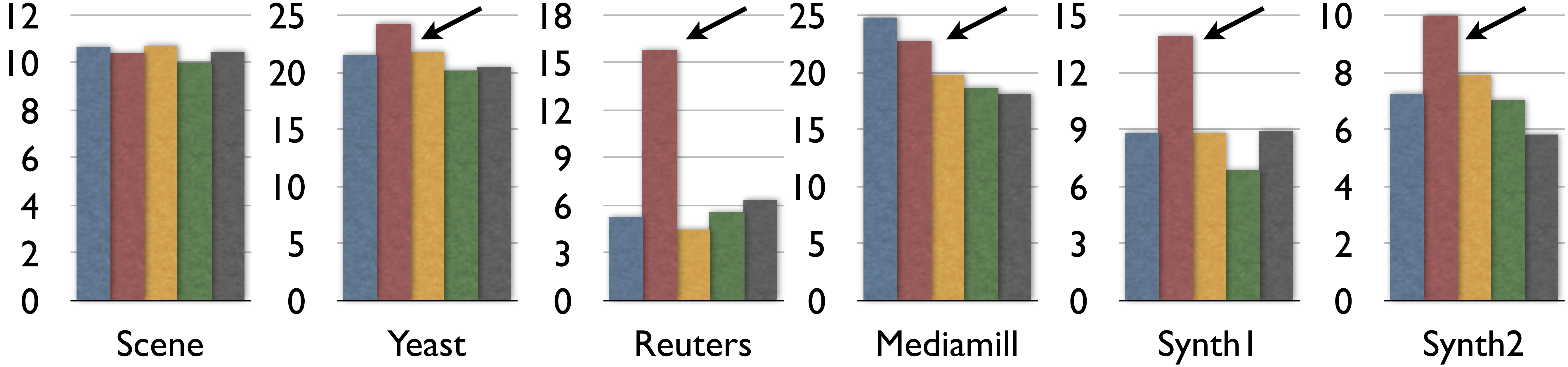


Bad as a prediction method (all trained with **Exact**)...

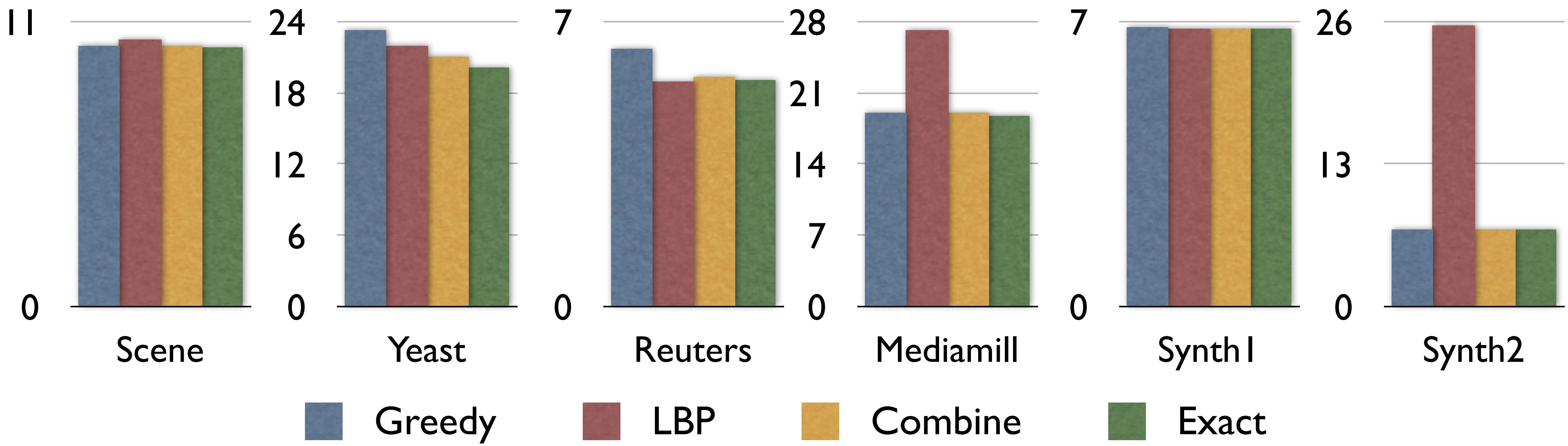


The Sorry State of LBP

Bad as a training method (all predicted with **Exact**)...

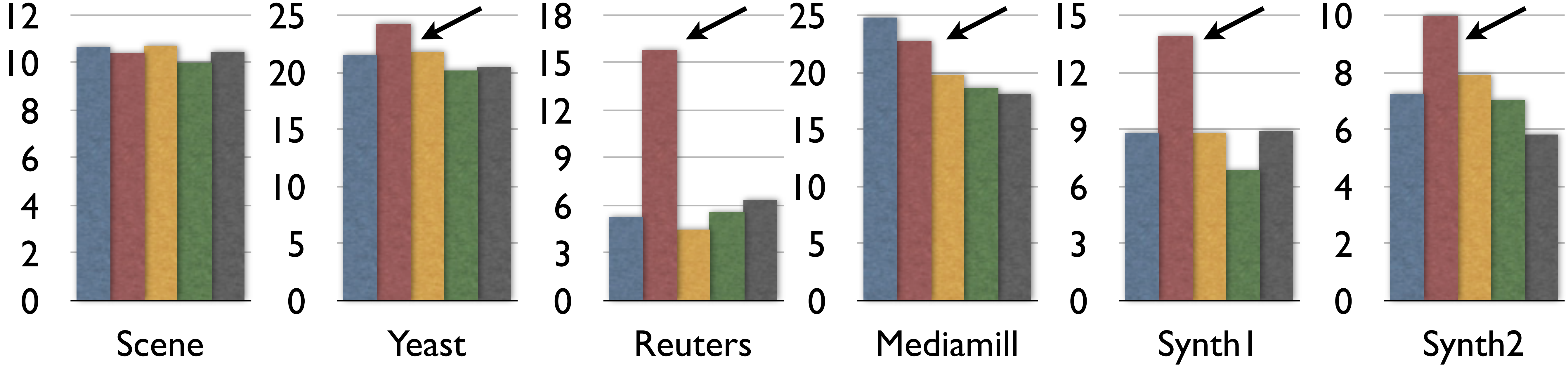


Bad as a prediction method (all trained with **Exact**)...

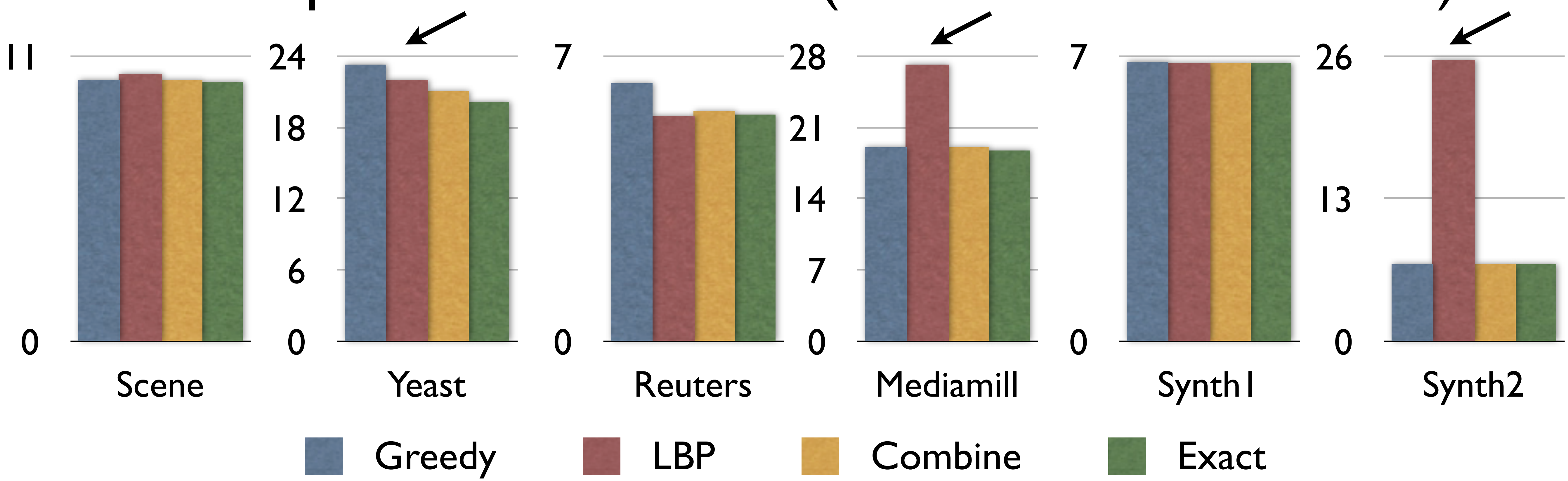


The Sorry State of LBP

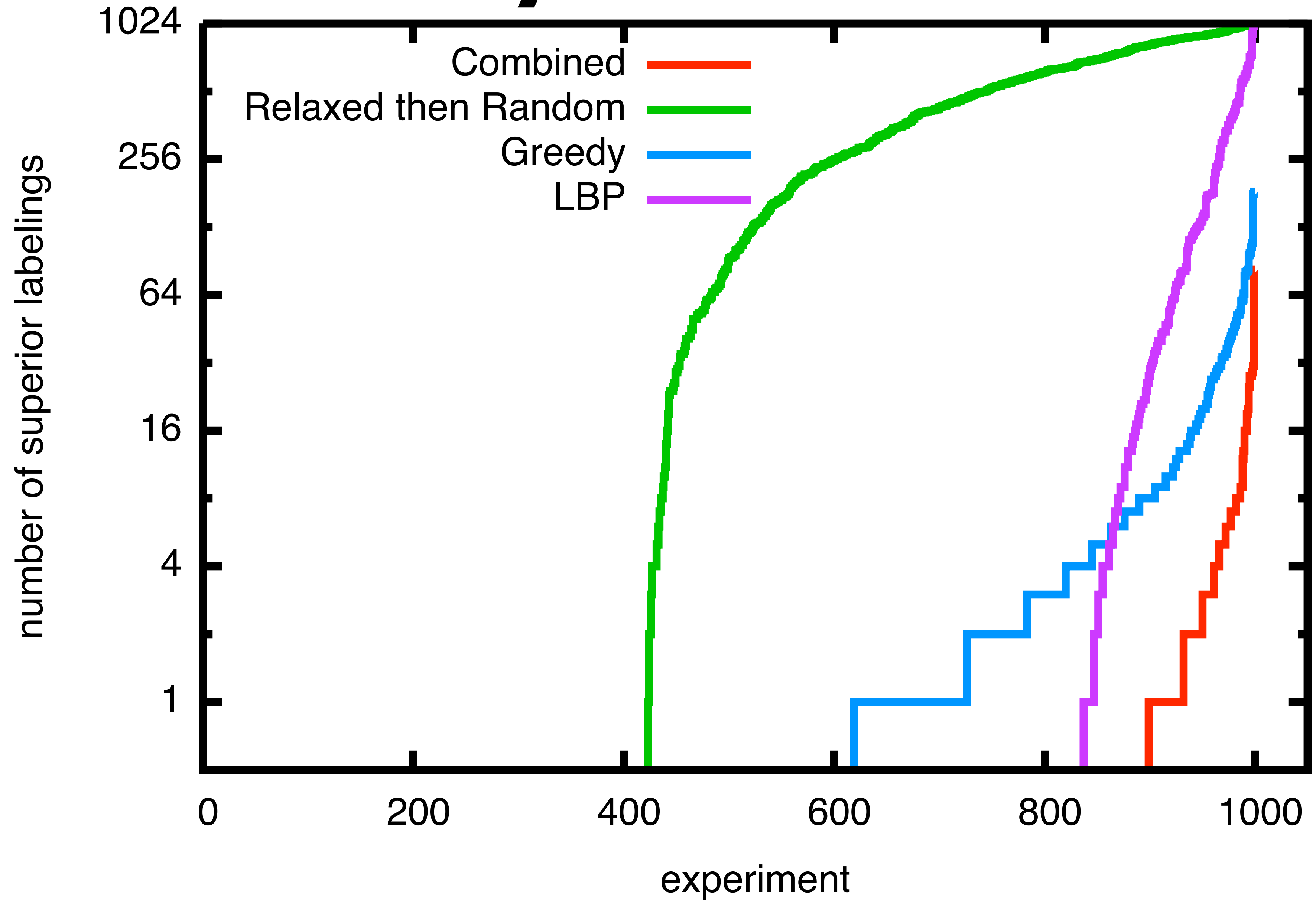
Bad as a training method (all predicted with **Exact**)...



Bad as a prediction method (all trained with **Exact**)...



The Sorry State of LBP

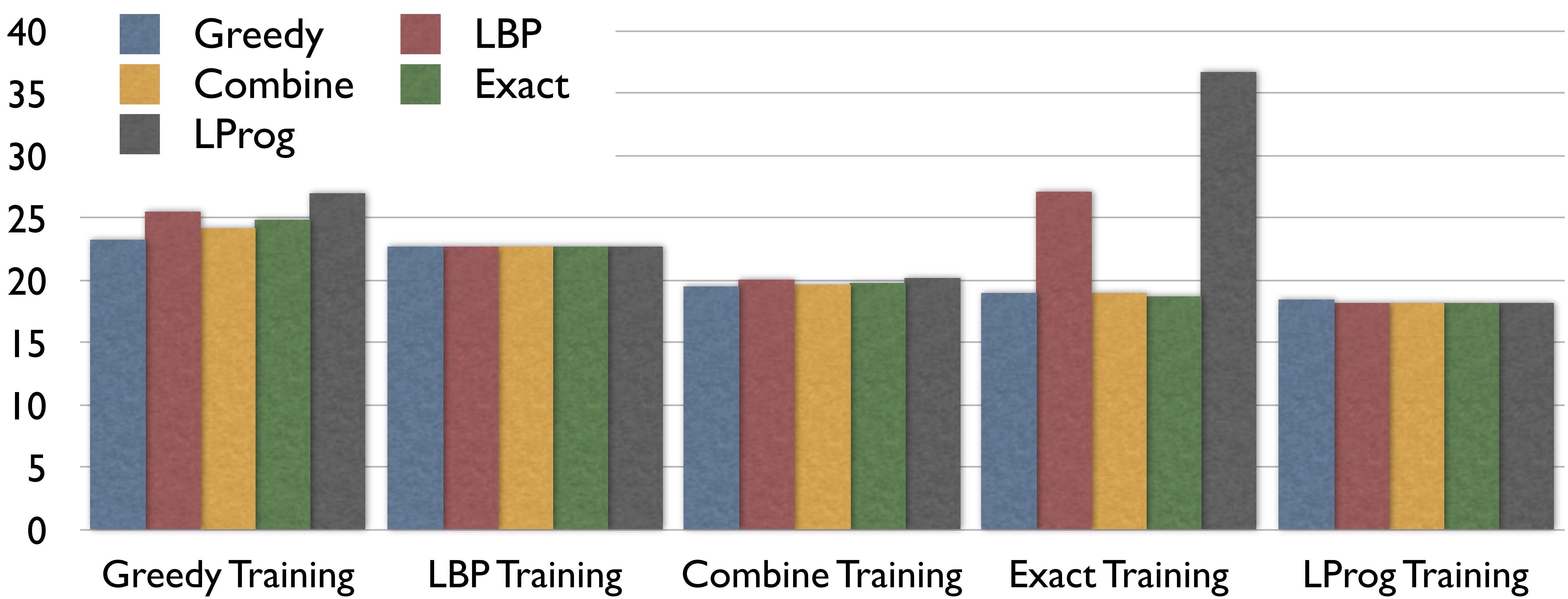


- 1000 MRFs with random [-1,1] node/edge potentials on 10 nodes.
- Vertical axis has (for each MRF) # of labelings better than returned by each inference method.
- **LBP** returns optimal labelings more often than **Greedy**. However, when it does poorly, it does very poorly.

Relaxation

- Results for Mediamill!
- Notice **predictor consistency** with relaxed LProg trained models.
- Notice occasional **very poor** performance of LProg as a classifier.
- Presence of fractional constraints in LProg trained models leads to “smoothed” **easier** space.
- Lack of fractional constraints in other models **hurts** relaxed LProg predictor.

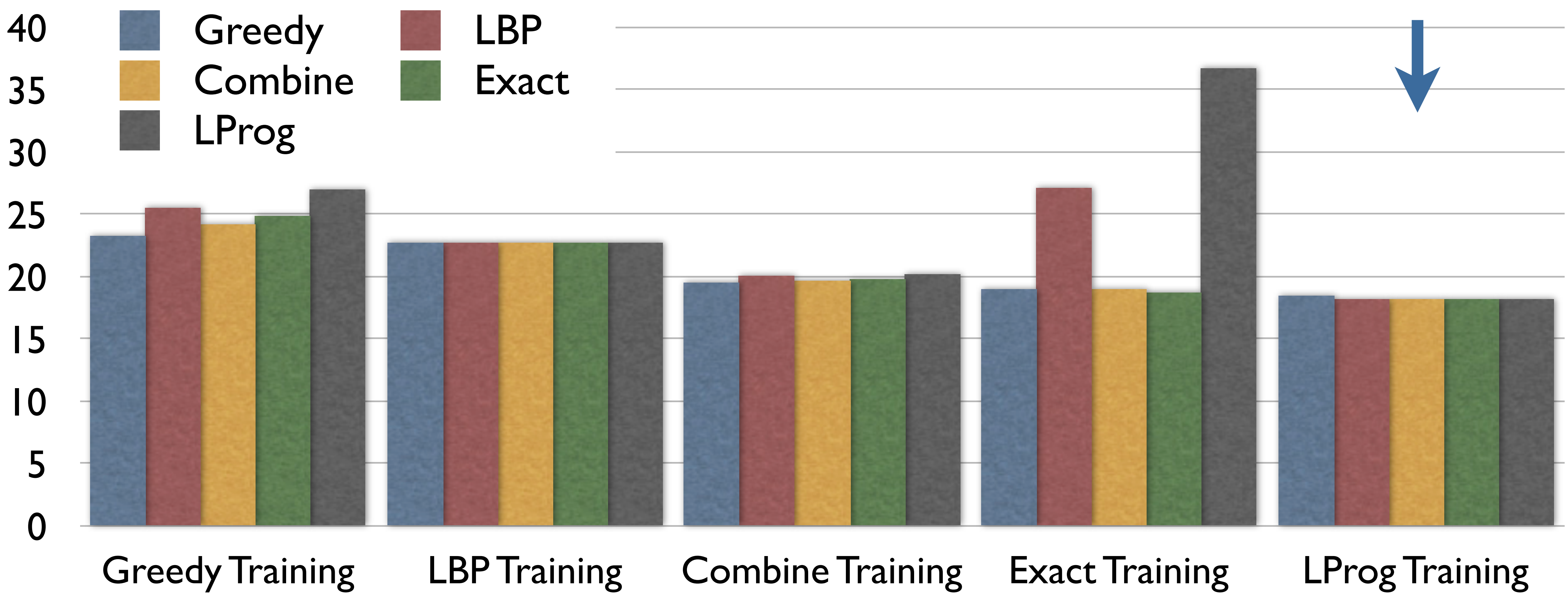
Losses per Dataset. Inference method used during training and prediction.



Relaxation

- Results for Mediamill!
- Notice **predictor consistency** with relaxed LProg trained models.
- Notice occasional **very poor** performance of LProg as a classifier.
- Presence of fractional constraints in LProg trained models leads to "smoothed" **easier** space.
- Lack of fractional constraints in other models **hurts** relaxed LProg predictor.

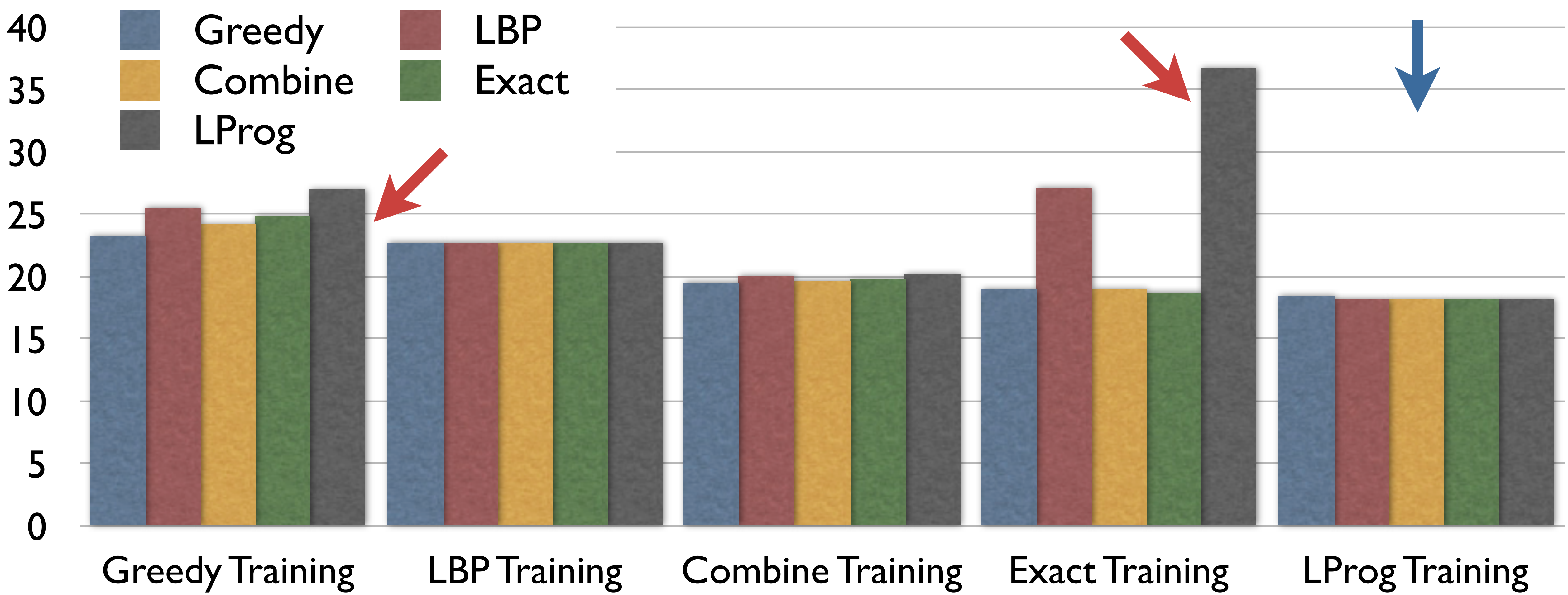
Losses per Dataset. Inference method used during training and prediction.



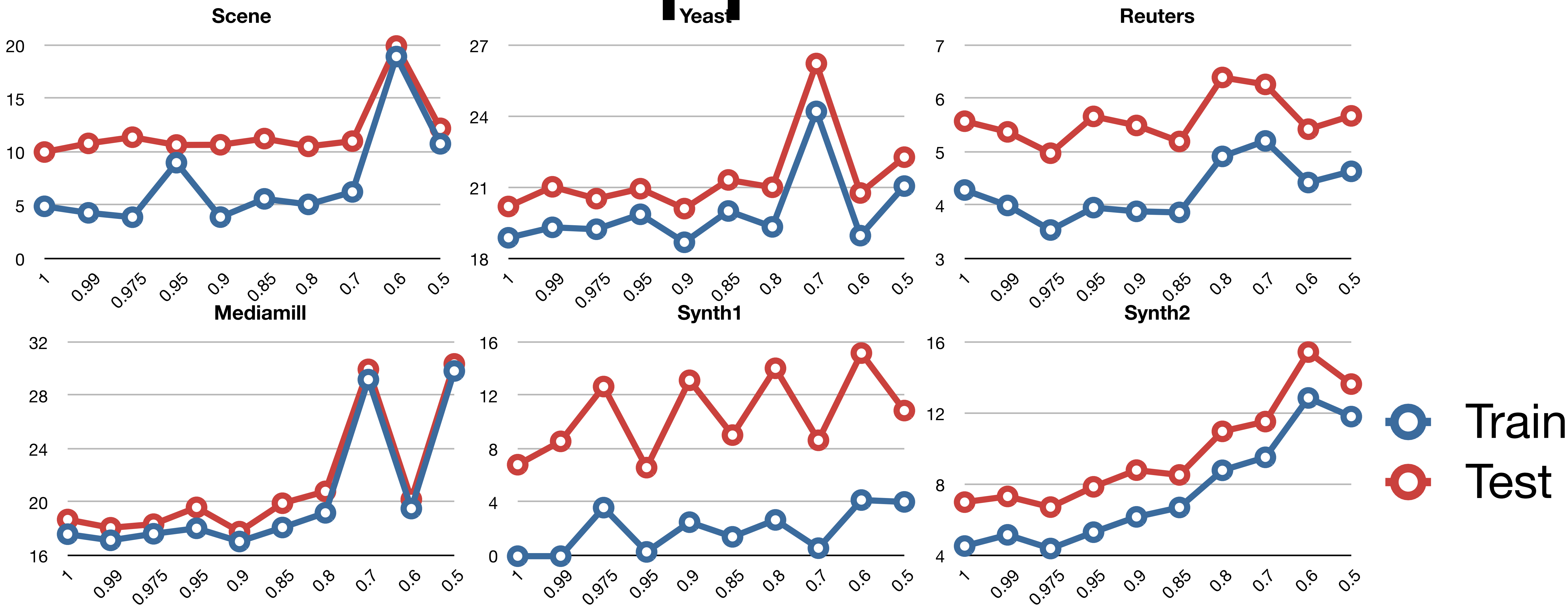
Relaxation

- Results for Mediamill!
- Notice **predictor consistency** with relaxed LProg trained models.
- Notice occasional **very poor** performance of LProg as a classifier.
- Presence of fractional constraints in LProg trained models leads to “smoothed” **easier** space.
- Lack of fractional constraints in other models **hurts** relaxed LProg predictor.

Losses per Dataset. Inference method used during training and prediction.



Known Approximations



- Do training with artificial ρ-approximate inference methods.
- Testing uses exact inference.
- Lower ρ means worse method.
- **Train** and **test** set losses reported.

- **Encouraging**: Learning seems at least partially tolerant to inexact inference methods.
- **Discouraging**: Not a smooth climbdown in test error!

Summary

- Reviewed structural SVMs.
- Explained the consequences of inexact inference.
- Theoretically and empirically analyzed two approximation families.
  - **Undergenerating** (i.e., local)
  - **Overgenerating** (i.e., relaxations)
- Completely connected binary pairwise MRFs applied to multilabel classification serves as example application.
- Overgenerating methods:
  - Preserve key theoretical SSVM properties.
  - Learn robust "stable" predictive models.

Software

- **SVM$^{\textbf{\textit{python}}}$**: SVM^{struct}, but API functions in Python, not C. Obviates annoying details (IO of model structures, memory management). http://www.cs.cornell.edu/~tomf/svmpython2/
- **PyGLPK**: GNU Linear Programming Kit (Andrew Makhorin) as a Pythonic extension module. http://www.cs.cornell.edu/~tomf/pyglpk/
- **PyGraphcut**: Graphcut based energy optimization framework (Boykov and Kolmogorov) as a Pythonic extension module. http://www.cs.cornell.edu/~tomf/pygraphcut/

Thank you

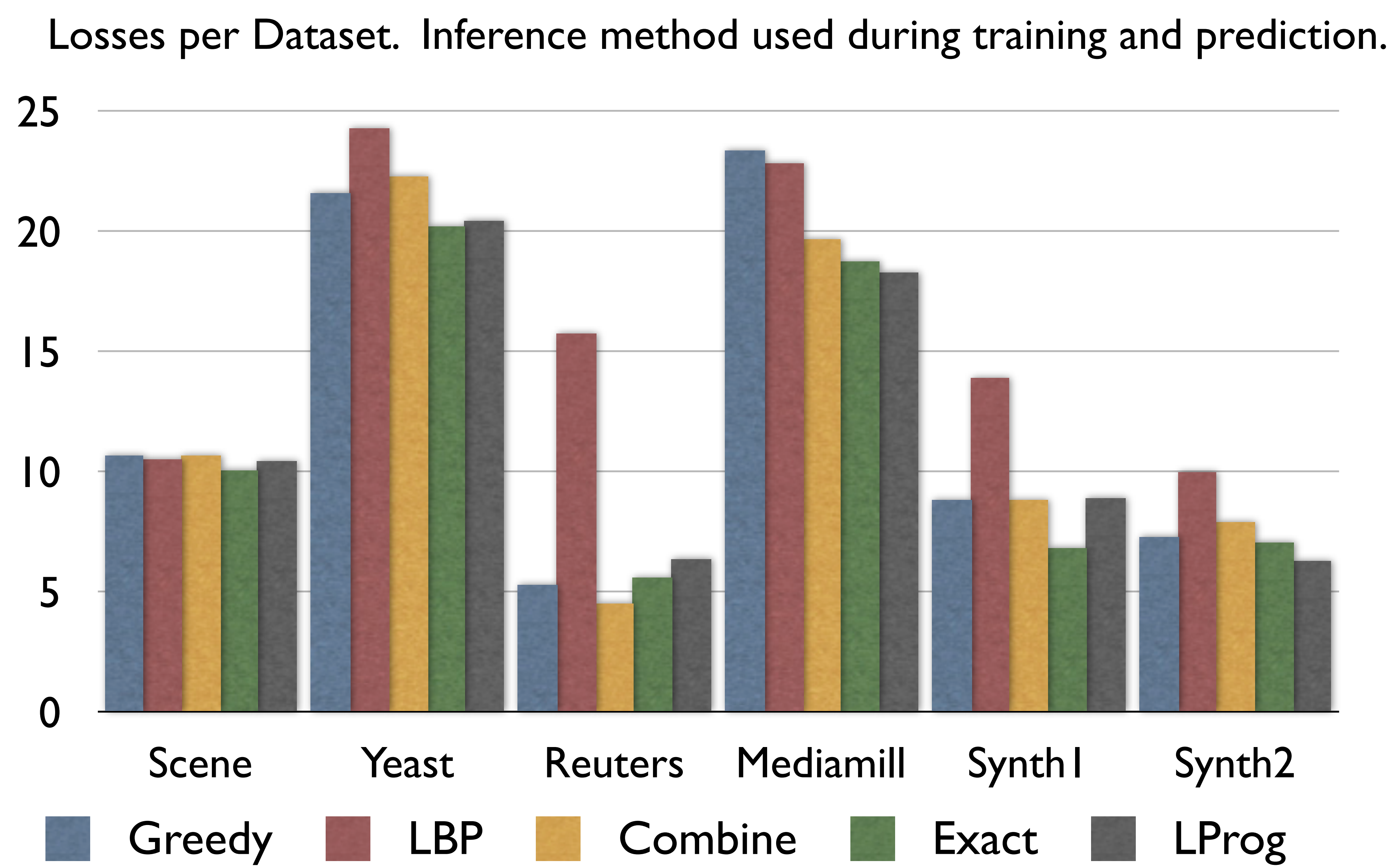
Questions?

More Slides

- The detailed tables.

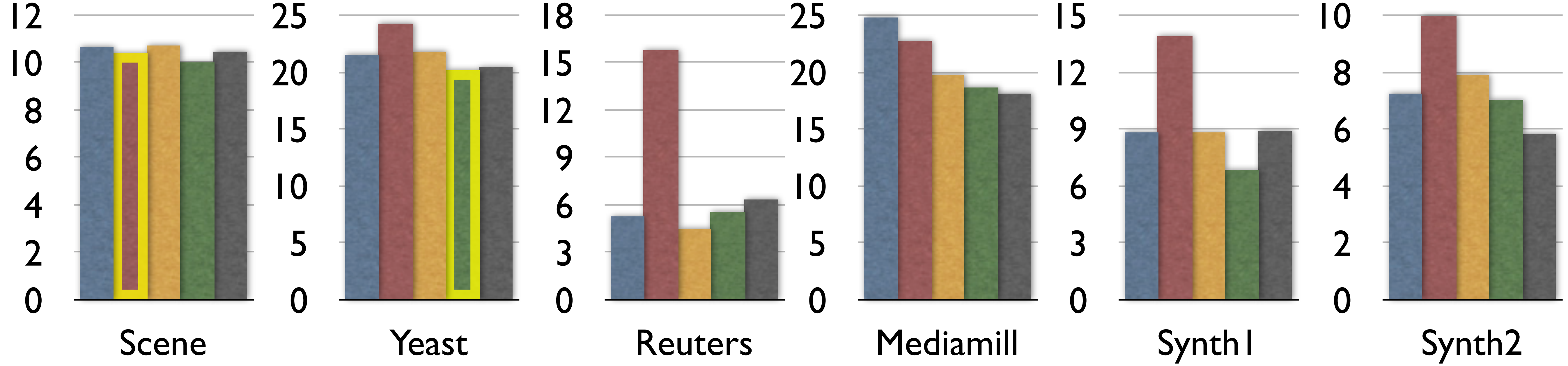
The Sorry State of LBP

- Lower is better

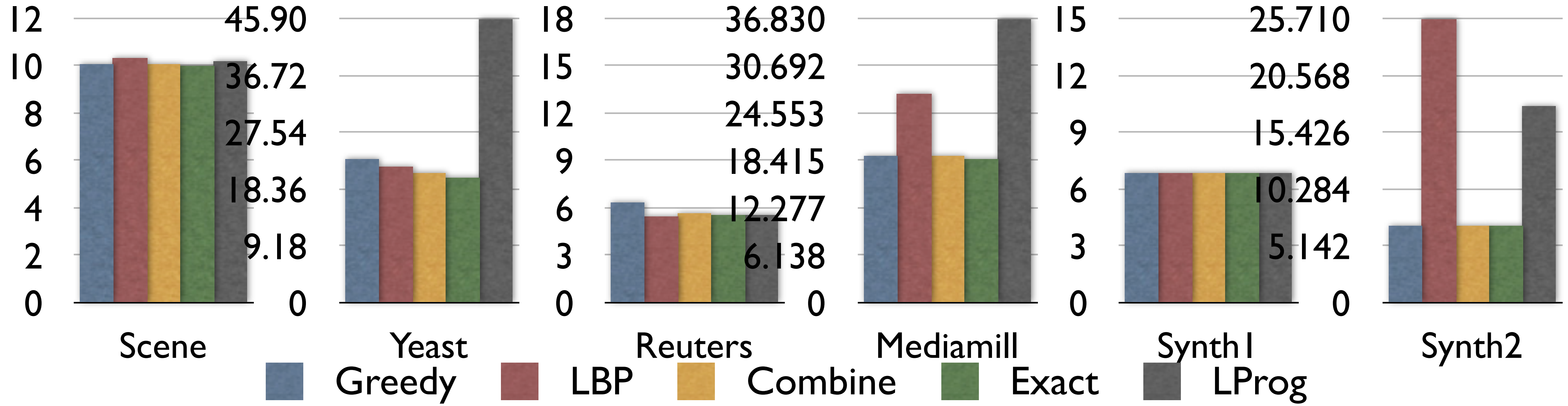


The Sorry State of LBP

Bad as a training method (all predicted with **Exact**)...



Bad as a prediction method (all trained with **Exact**)...



Great Big Table

	Greedy	LBP	Combine	Exact	Relaxed	Greedy	LBP	Combine	Exact	Relaxed
	Scene Dataset			11.43±.29	18.10	Mediamill Dataset			18.60±.14	25.37
Greedy	10.67±.28	10.74±.28	10.67±.28	10.67±.28	10.67±.28	23.39±.16	25.66±.17	24.32±.17	24.92±.17	27.05±.18
LBP	10.45±.27	10.54±.27	10.45±.27	10.42±.27	10.49±.27	22.83±.16	22.83±.16	22.83±.16	22.83±.16	22.83±.16
Combine	10.72±.28	11.78±.30	10.72±.28	10.77±.28	11.20±.29	19.56±.14	20.12±.15	19.72±.14	19.82±.14	20.23±.15
Exact	10.08±.26	10.33±.27	10.08±.26	10.06±.26	10.20±.26	19.07±.14	27.23±.18	19.08±.14	18.75±.14	36.83±.21
Relaxed	10.55±.27	10.49±.27	10.49±.27	10.49±.27	10.49±.27	18.50±.14	18.26±.14	18.26±.14	18.21±.14	18.29±.14
	Yeast Dataset			20.91±.55	25.09	Synth1 Dataset			8.99±.08	16.34
Greedy	21.62±.56	21.77±.56	21.58±.56	21.62±.56	24.42±.61	8.86±.08	8.86±.08	8.86±.08	8.86±.08	8.86±.08
LBP	24.32±.61	24.32±.61	24.32±.61	24.32±.61	24.32±.61	13.94±.12	13.94±.12	13.94±.12	13.94±.12	13.94±.12
Combine	22.33±.57	37.24±.77	22.32±.57	21.82±.56	42.72±.81	8.86±.08	8.86±.08	8.86±.08	8.86±.08	8.86±.08
Exact	23.38±.59	21.99±.57	21.06±.55	20.23±.53	45.90±.82	6.89±.06	6.86±.06	6.86±.06	6.86±.06	6.86±.06
Relaxed	20.47±.54	20.45±.54	20.47±.54	20.48±.54	20.49±.54	8.94±.08	8.94±.08	8.94±.08	8.94±.08	8.94±.08
	Reuters Dataset			4.96±.09	15.80	Synth2 Dataset			9.80±.09	10.00
Greedy	5.32±.09	13.38±.21	5.06±.09	5.42±.09	16.98±.26	7.27±.07	27.92±.20	7.27±.07	7.28±.07	19.03±.15
LBP	15.80±.25	15.80±.25	15.80±.25	15.80±.25	15.80±.25	10.00±.09	10.00±.09	10.00±.09	10.00±.09	10.00±.09
Combine	4.90±.09	4.57±.08	4.53±.08	4.49±.08	4.55±.08	7.90±.07	26.39±.19	7.90±.07	7.90±.07	18.11±.15
Exact	6.36±.11	5.54±.10	5.67±.10	5.59±.10	5.62±.10	7.04±.07	25.71±.19	7.04±.07	7.04±.07	17.80±.15
Relaxed	6.73±.12	6.41±.11	6.38±.11	6.38±.11	6.38±.11	5.83±.05	6.63±.06	5.83±.05	5.83±.05	6.29±.06

Great Big Table

	Greedy	LBP	Combine	Exact	Relaxed	Greedy	LBP	Combine	Exact	Relaxed
	Scene Dataset			11.43±.29	18.10	Mediamill Dataset			18.60±.14	25.37
Greedy	10.67±.28	10.74±.28	10.67±.28	10.67±.28	10.67±.28	23.39±.16	25.66±.17	24.32±.17	24.92±.17	27.05±.18
LBP	10.45±.27	10.54±.27	10.45±.27	10.42±.27	10.49±.27	22.83±.16	22.83±.16	22.83±.16	22.83±.16	22.83±.16
Combine	10.72±.28	11.78±.30	10.72±.28	10.77±.28	11.20±.29	19.56±.14	20.12±.15	19.72±.14	19.82±.14	20.23±.15
Exact	10.08±.26	10.33±.27	10.08±.26	10.06±.26	10.20±.26	19.07±.14	27.23±.18	19.08±.14	18.75±.14	36.83±.21
Relaxed	10.55±.27	10.49±.27	10.49±.27	10.49±.27	10.49±.27	18.50±.14	18.26±.14	18.26±.14	18.21±.14	18.29±.14
	Yeast Dataset			20.91±.55	25.09	Synth1 Dataset			8.99±.08	16.34
Greedy	21.62±.56	21.77±.56	21.58±.56	21.62±.56	24.42±.61	8.86±.08	8.86±.08	8.86±.08	8.86±.08	8.86±.08
LBP	24.32±.61	24.32±.61	24.32±.61	24.32±.61	24.32±.61	13.94±.12	13.94±.12	13.94±.12	13.94±.12	13.94±.12
Combine	22.33±.57	37.24±.77	22.32±.57	21.82±.56	42.72±.81	8.86±.08	8.86±.08	8.86±.08	8.86±.08	8.86±.08
Exact	23.38±.59	21.99±.57	21.06±.55	20.23±.53	45.90±.82	6.89±.06	6.86±.06	6.86±.06	6.86±.06	6.86±.06
Relaxed	20.47±.54	20.45±.54	20.47±.54	20.48±.54	20.49±.54	8.94±.08	8.94±.08	8.94±.08	8.94±.08	8.94±.08
	Reuters Dataset			4.96±.09	15.80	Synth2 Dataset			9.80±.09	10.00
Greedy	5.32±.09	13.38±.21	5.06±.09	5.42±.09	16.98±.26	7.27±.07	27.92±.20	7.27±.07	7.28±.07	19.03±.15
LBP	15.80±.25	15.80±.25	15.80±.25	15.80±.25	15.80±.25	10.00±.09	10.00±.09	10.00±.09	10.00±.09	10.00±.09
Combine	4.90±.09	4.57±.08	4.53±.08	4.49±.08	4.55±.08	7.90±.07	26.39±.19	7.90±.07	7.90±.07	18.11±.15
Exact	6.36±.11	5.54±.10	5.67±.10	5.59±.10	5.62±.10	7.04±.07	25.71±.19	7.04±.07	7.04±.07	17.80±.15
Relaxed	6.73±.12	6.41±.11	6.38±.11	6.38±.11	6.38±.11	5.83±.05	6.63±.06	5.83±.05	5.83±.05	6.29±.06

- Results per dataset in **blocks**.

Great Big Table

	Greedy	LBP	Combine	Exact	Relaxed	Greedy	LBP	Combine	Exact	Relaxed
	Scene Dataset			11.43±.29	18.10	Mediamill Dataset			18.60±.14	25.37
Greedy	10.67±.28	10.74±.28	10.67±.28	10.67±.28	10.67±.28	23.39±.16	25.66±.17	24.32±.17	24.92±.17	27.05±.18
LBP	10.45±.27	10.54±.27	10.45±.27	10.42±.27	10.49±.27	22.83±.16	22.83±.16	22.83±.16	22.83±.16	22.83±.16
Combine	10.72±.28	11.78±.30	10.72±.28	10.77±.28	11.20±.29	19.56±.14	20.12±.15	19.72±.14	19.82±.14	20.23±.15
Exact	10.08±.26	10.33±.27	10.08±.26	10.06±.26	10.20±.26	19.07±.14	27.23±.18	19.08±.14	18.75±.14	36.83±.21
Relaxed	10.55±.27	10.49±.27	10.49±.27	10.49±.27	10.49±.27	18.50±.14	18.26±.14	18.26±.14	18.21±.14	18.29±.14
	Yeast Dataset			20.91±.55	25.09	Synth1 Dataset			8.99±.08	16.34
Greedy	21.62±.56	21.77±.56	21.58±.56	21.62±.56	24.42±.61	8.86±.08	8.86±.08	8.86±.08	8.86±.08	8.86±.08
LBP	24.32±.61	24.32±.61	24.32±.61	24.32±.61	24.32±.61	13.94±.12	13.94±.12	13.94±.12	13.94±.12	13.94±.12
Combine	22.33±.57	37.24±.77	22.32±.57	21.82±.56	42.72±.81	8.86±.08	8.86±.08	8.86±.08	8.86±.08	8.86±.08
Exact	23.38±.59	21.99±.57	21.06±.55	20.23±.53	45.90±.82	6.89±.06	6.86±.06	6.86±.06	6.86±.06	6.86±.06
Relaxed	20.47±.54	20.45±.54	20.47±.54	20.48±.54	20.49±.54	8.94±.08	8.94±.08	8.94±.08	8.94±.08	8.94±.08
	Reuters Dataset			4.96±.09	15.80	Synth2 Dataset			9.80±.09	10.00
Greedy	5.32±.09	13.38±.21	5.06±.09	5.42±.09	16.98±.26	7.27±.07	27.92±.20	7.27±.07	7.28±.07	19.03±.15
LBP	15.80±.25	15.80±.25	15.80±.25	15.80±.25	15.80±.25	10.00±.09	10.00±.09	10.00±.09	10.00±.09	10.00±.09
Combine	4.90±.09	4.57±.08	4.53±.08	4.49±.08	4.55±.08	7.90±.07	26.39±.19	7.90±.07	7.90±.07	18.11±.15
Exact	6.36±.11	5.54±.10	5.67±.10	5.59±.10	5.62±.10	7.04±.07	25.71±.19	7.04±.07	7.04±.07	17.80±.15
Relaxed	6.73±.12	6.41±.11	6.38±.11	6.38±.11	6.38±.11	5.83±.05	6.63±.06	5.83±.05	5.83±.05	6.29±.06

- Results per dataset in **blocks**.
- **Rows** indicate training inference method (separation oracle).

Great Big Table

	Greedy	LBP	Combine	Exact	Relaxed	Greedy	LBP	Combine	Exact	Relaxed
	Scene Dataset			11.43±.29	18.10	Mediamill Dataset			18.60±.14	25.37
Greedy	10.67±.28	10.74±.28	10.67±.28	10.67±.28	10.67±.28	23.39±.16	25.66±.17	24.32±.17	24.92±.17	27.05±.18
LBP	10.45±.27	10.54±.27	10.45±.27	10.42±.27	10.49±.27	22.83±.16	22.83±.16	22.83±.16	22.83±.16	22.83±.16
Combine	10.72±.28	11.78±.30	10.72±.28	10.77±.28	11.20±.29	19.56±.14	20.12±.15	19.72±.14	19.82±.14	20.23±.15
Exact	10.08±.26	10.33±.27	10.08±.26	10.06±.26	10.20±.26	19.07±.14	27.23±.18	19.08±.14	18.75±.14	36.83±.21
Relaxed	10.55±.27	10.49±.27	10.49±.27	10.49±.27	10.49±.27	18.50±.14	18.26±.14	18.26±.14	18.21±.14	18.29±.14
	Yeast Dataset			20.91±.55	25.09	Synth1 Dataset			8.99±.08	16.34
Greedy	21.62±.56	21.77±.56	21.58±.56	21.62±.56	24.42±.61	8.86±.08	8.86±.08	8.86±.08	8.86±.08	8.86±.08
LBP	24.32±.61	24.32±.61	24.32±.61	24.32±.61	24.32±.61	13.94±.12	13.94±.12	13.94±.12	13.94±.12	13.94±.12
Combine	22.33±.57	37.24±.77	22.32±.57	21.82±.56	42.72±.81	8.86±.08	8.86±.08	8.86±.08	8.86±.08	8.86±.08
Exact	23.38±.59	21.99±.57	21.06±.55	20.23±.53	45.90±.82	6.89±.06	6.86±.06	6.86±.06	6.86±.06	6.86±.06
Relaxed	20.47±.54	20.45±.54	20.47±.54	20.48±.54	20.49±.54	8.94±.08	8.94±.08	8.94±.08	8.94±.08	8.94±.08
	Reuters Dataset			4.96±.09	15.80	Synth2 Dataset			9.80±.09	10.00
Greedy	5.32±.09	13.38±.21	5.06±.09	5.42±.09	16.98±.26	7.27±.07	27.92±.20	7.27±.07	7.28±.07	19.03±.15
LBP	15.80±.25	15.80±.25	15.80±.25	15.80±.25	15.80±.25	10.00±.09	10.00±.09	10.00±.09	10.00±.09	10.00±.09
Combine	4.90±.09	4.57±.08	4.53±.08	4.49±.08	4.55±.08	7.90±.07	26.39±.19	7.90±.07	7.90±.07	18.11±.15
Exact	6.36±.11	5.54±.10	5.67±.10	5.59±.10	5.62±.10	7.04±.07	25.71±.19	7.04±.07	7.04±.07	17.80±.15
Relaxed	6.73±.12	6.41±.11	6.38±.11	6.38±.11	6.38±.11	5.83±.05	6.63±.06	5.83±.05	5.83±.05	6.29±.06

- Results per dataset in **blocks**.
- **Rows** indicate training inference method (separation oracle).
- **Columns** indicate prediction inference method.

Great Big Table

	Greedy	LBP	Combine	Exact	Relaxed	Greedy	LBP	Combine	Exact	Relaxed
	Scene Dataset			11.43±.29	18.10	Mediamill Dataset			18.60±.14	25.37
Greedy	10.67±.28	10.74±.28	10.67±.28	10.67±.28	10.67±.28	23.39±.16	25.66±.17	24.32±.17	24.92±.17	27.05±.18
LBP	10.45±.27	10.54±.27	10.45±.27	10.42±.27	10.49±.27	22.83±.16	22.83±.16	22.83±.16	22.83±.16	22.83±.16
Combine	10.72±.28	11.78±.30	10.72±.28	10.77±.28	11.20±.29	19.56±.14	20.12±.15	19.72±.14	19.82±.14	20.23±.15
Exact	10.08±.26	10.33±.27	10.08±.26	10.06±.26	10.20±.26	19.07±.14	27.23±.18	19.08±.14	18.75±.14	36.83±.21
Relaxed	10.55±.27	10.49±.27	10.49±.27	10.49±.27	10.49±.27	18.50±.14	18.26±.14	18.26±.14	18.21±.14	18.29±.14
	Yeast Dataset			20.91±.55	25.09	Synth1 Dataset			8.99±.08	16.34
Greedy	21.62±.56	21.77±.56	21.58±.56	21.62±.56	24.42±.61	8.86±.08	8.86±.08	8.86±.08	8.86±.08	8.86±.08
LBP	24.32±.61	24.32±.61	24.32±.61	24.32±.61	24.32±.61	13.94±.12	13.94±.12	13.94±.12	13.94±.12	13.94±.12
Combine	22.33±.57	37.24±.77	22.32±.57	21.82±.56	42.72±.81	8.86±.08	8.86±.08	8.86±.08	8.86±.08	8.86±.08
Exact	23.38±.59	21.99±.57	21.06±.55	20.23±.53	45.90±.82	6.89±.06	6.86±.06	6.86±.06	6.86±.06	6.86±.06
Relaxed	20.47±.54	20.45±.54	20.47±.54	20.48±.54	20.49±.54	8.94±.08	8.94±.08	8.94±.08	8.94±.08	8.94±.08
	Reuters Dataset			4.96±.09	15.80	Synth2 Dataset			9.80±.09	10.00
Greedy	5.32±.09	13.38±.21	5.06±.09	5.42±.09	16.98±.26	7.27±.07	27.92±.20	7.27±.07	7.28±.07	19.03±.15
LBP	15.80±.25	15.80±.25	15.80±.25	15.80±.25	15.80±.25	10.00±.09	10.00±.09	10.00±.09	10.00±.09	10.00±.09
Combine	4.90±.09	4.57±.08	4.53±.08	4.49±.08	4.55±.08	7.90±.07	26.39±.19	7.90±.07	7.90±.07	18.11±.15
Exact	6.36±.11	5.54±.10	5.67±.10	5.59±.10	5.62±.10	7.04±.07	25.71±.19	7.04±.07	7.04±.07	17.80±.15
Relaxed	6.73±.12	6.41±.11	6.38±.11	6.38±.11	6.38±.11	5.83±.05	6.63±.06	5.83±.05	5.83±.05	6.29±.06

- Results per dataset in **blocks**.
- **Rows** indicate training inference method (separation oracle).
- **Columns** indicate prediction inference method.
- **Numbers** are Hamming loss percentage, ± **standard error** (with a twist).

Great Big Table

	Greedy	LBP	Combine	Exact	Relaxed	Greedy	LBP	Combine	Exact	Relaxed
	Scene Dataset			11.43±.29	18.10	Mediamill Dataset			18.60±.14	25.37
Greedy	10.67±.28	10.74±.28	10.67±.28	10.67±.28	10.67±.28	23.39±.16	25.66±.17	24.32±.17	24.92±.17	27.05±.18
LBP	10.45±.27	10.54±.27	10.45±.27	10.42±.27	10.49±.27	22.83±.16	22.83±.16	22.83±.16	22.83±.16	22.83±.16
Combine	10.72±.28	11.78±.30	10.72±.28	10.77±.28	11.20±.29	19.56±.14	20.12±.15	19.72±.14	19.82±.14	20.23±.15
Exact	10.08±.26	10.33±.27	10.08±.26	10.06±.26	10.20±.26	19.07±.14	27.23±.18	19.08±.14	18.75±.14	36.83±.21
Relaxed	10.55±.27	10.49±.27	10.49±.27	10.49±.27	10.49±.27	18.50±.14	18.26±.14	18.26±.14	18.21±.14	18.29±.14
	Yeast Dataset			20.91±.55	25.09	Synth1 Dataset			8.99±.08	16.34
Greedy	21.62±.56	21.77±.56	21.58±.56	21.62±.56	24.42±.61	8.86±.08	8.86±.08	8.86±.08	8.86±.08	8.86±.08
LBP	24.32±.61	24.32±.61	24.32±.61	24.32±.61	24.32±.61	13.94±.12	13.94±.12	13.94±.12	13.94±.12	13.94±.12
Combine	22.33±.57	37.24±.77	22.32±.57	21.82±.56	42.72±.81	8.86±.08	8.86±.08	8.86±.08	8.86±.08	8.86±.08
Exact	23.38±.59	21.99±.57	21.06±.55	20.23±.53	45.90±.82	6.89±.06	6.86±.06	6.86±.06	6.86±.06	6.86±.06
Relaxed	20.47±.54	20.45±.54	20.47±.54	20.48±.54	20.49±.54	8.94±.08	8.94±.08	8.94±.08	8.94±.08	8.94±.08
	Reuters Dataset			4.96±.09	15.80	Synth2 Dataset			9.80±.09	10.00
Greedy	5.32±.09	13.38±.21	5.06±.09	5.42±.09	16.98±.26	7.27±.07	27.92±.20	7.27±.07	7.28±.07	19.03±.15
LBP	15.80±.25	15.80±.25	15.80±.25	15.80±.25	15.80±.25	10.00±.09	10.00±.09	10.00±.09	10.00±.09	10.00±.09
Combine	4.90±.09	4.57±.08	4.53±.08	4.49±.08	4.55±.08	7.90±.07	26.39±.19	7.90±.07	7.90±.07	18.11±.15
Exact	6.36±.11	5.54±.10	5.67±.10	5.59±.10	5.62±.10	7.04±.07	25.71±.19	7.04±.07	7.04±.07	17.80±.15
Relaxed	6.73±.12	6.41±.11	6.38±.11	6.38±.11	6.38±.11	5.83±.05	6.63±.06	5.83±.05	5.83±.05	6.29±.06

- Results per dataset in **blocks**.
- **Rows** indicate training inference method (separation oracle).
- **Columns** indicate prediction inference method.
- **Numbers** are Hamming loss percentage, ± **standard error** (with a twist).
- **Edgeless loss** next to name.

Great Big Table

	Greedy	LBP	Combine	Exact	Relaxed	Greedy	LBP	Combine	Exact	Relaxed
	Scene Dataset			11.43±.29	18.10	Mediamill Dataset			18.60±.14	25.37
Greedy	10.67±.28	10.74±.28	10.67±.28	10.67±.28	10.67±.28	23.39±.16	25.66±.17	24.32±.17	24.92±.17	27.05±.18
LBP	10.45±.27	10.54±.27	10.45±.27	10.42±.27	10.49±.27	22.83±.16	22.83±.16	22.83±.16	22.83±.16	22.83±.16
Combine	10.72±.28	11.78±.30	10.72±.28	10.77±.28	11.20±.29	19.56±.14	20.12±.15	19.72±.14	19.82±.14	20.23±.15
Exact	10.08±.26	10.33±.27	10.08±.26	10.06±.26	10.20±.26	19.07±.14	27.23±.18	19.08±.14	18.75±.14	36.83±.21
Relaxed	10.55±.27	10.49±.27	10.49±.27	10.49±.27	10.49±.27	18.50±.14	18.26±.14	18.26±.14	18.21±.14	18.29±.14
	Yeast Dataset			20.91±.55	25.09	Synth1 Dataset			8.99±.08	16.34
Greedy	21.62±.56	21.77±.56	21.58±.56	21.62±.56	24.42±.61	8.86±.08	8.86±.08	8.86±.08	8.86±.08	8.86±.08
LBP	24.32±.61	24.32±.61	24.32±.61	24.32±.61	24.32±.61	13.94±.12	13.94±.12	13.94±.12	13.94±.12	13.94±.12
Combine	22.33±.57	37.24±.77	22.32±.57	21.82±.56	42.72±.81	8.86±.08	8.86±.08	8.86±.08	8.86±.08	8.86±.08
Exact	23.38±.59	21.99±.57	21.06±.55	20.23±.53	45.90±.82	6.89±.06	6.86±.06	6.86±.06	6.86±.06	6.86±.06
Relaxed	20.47±.54	20.45±.54	20.47±.54	20.48±.54	20.49±.54	8.94±.08	8.94±.08	8.94±.08	8.94±.08	8.94±.08
	Reuters Dataset			4.96±.09	15.80	Synth2 Dataset			9.80±.09	10.00
Greedy	5.32±.09	13.38±.21	5.06±.09	5.42±.09	16.98±.26	7.27±.07	27.92±.20	7.27±.07	7.28±.07	19.03±.15
LBP	15.80±.25	15.80±.25	15.80±.25	15.80±.25	15.80±.25	10.00±.09	10.00±.09	10.00±.09	10.00±.09	10.00±.09
Combine	4.90±.09	4.57±.08	4.53±.08	4.49±.08	4.55±.08	7.90±.07	26.39±.19	7.90±.07	7.90±.07	18.11±.15
Exact	6.36±.11	5.54±.10	5.67±.10	5.59±.10	5.62±.10	7.04±.07	25.71±.19	7.04±.07	7.04±.07	17.80±.15
Relaxed	6.73±.12	6.41±.11	6.38±.11	6.38±.11	6.38±.11	5.83±.05	6.63±.06	5.83±.05	5.83±.05	6.29±.06

- Results per dataset in **blocks**.
- **Rows** indicate training inference method (separation oracle).
- **Columns** indicate prediction inference method.
- **Numbers** are Hamming loss percentage, ± **standard error** (with a twist).
- **Edgeless loss** next to name.
- **Default loss** next to that.

The Sorry State of LBP

	Greedy	LBP	Combine	Exact	Relaxed	Greedy	LBP	Combine	Exact	Relaxed
	Scene Dataset			11.43±.29	18.10	Mediamill Dataset			18.60±.14	25.37
Greedy	10.67±.28	10.74±.28	10.67±.28	10.67±.28	10.67±.28	23.39±.16	25.66±.17	24.32±.17	24.92±.17	27.05±.18
LBP	10.45±.27	10.54±.27	10.45±.27	10.42±.27	10.49±.27	22.83±.16	22.83±.16	22.83±.16	22.83±.16	22.83±.16
Combine	10.72±.28	11.78±.30	10.72±.28	10.77±.28	11.20±.29	19.56±.14	20.12±.15	19.72±.14	19.82±.14	20.23±.15
Exact	10.08±.26	10.33±.27	10.08±.26	10.06±.26	10.20±.26	19.07±.14	27.23±.18	19.08±.14	18.75±.14	36.83±.21
Relaxed	10.55±.27	10.49±.27	10.49±.27	10.49±.27	10.49±.27	18.50±.14	18.26±.14	18.26±.14	18.21±.14	18.29±.14
	Yeast Dataset			20.91±.55	25.09	Synth1 Dataset			8.99±.08	16.34
Greedy	21.62±.56	21.77±.56	21.58±.56	21.62±.56	24.42±.61	8.86±.08	8.86±.08	8.86±.08	8.86±.08	8.86±.08
LBP	24.32±.61	24.32±.61	24.32±.61	24.32±.61	24.32±.61	13.94±.12	13.94±.12	13.94±.12	13.94±.12	13.94±.12
Combine	22.33±.57	37.24±.77	22.32±.57	21.82±.56	42.72±.81	8.86±.08	8.86±.08	8.86±.08	8.86±.08	8.86±.08
Exact	23.38±.59	21.99±.57	21.06±.55	20.23±.53	45.90±.82	6.89±.06	6.86±.06	6.86±.06	6.86±.06	6.86±.06
Relaxed	20.47±.54	20.45±.54	20.47±.54	20.48±.54	20.49±.54	8.94±.08	8.94±.08	8.94±.08	8.94±.08	8.94±.08
	Reuters Dataset			4.96±.09	15.80	Synth2 Dataset			9.80±.09	10.00
Greedy	5.32±.09	13.38±.21	5.06±.09	5.42±.09	16.98±.26	7.27±.07	27.92±.20	7.27±.07	7.28±.07	19.03±.15
LBP	15.80±.25	15.80±.25	15.80±.25	15.80±.25	15.80±.25	10.00±.09	10.00±.09	10.00±.09	10.00±.09	10.00±.09
Combine	4.90±.09	4.57±.08	4.53±.08	4.49±.08	4.55±.08	7.90±.07	26.39±.19	7.90±.07	7.90±.07	18.11±.15
Exact	6.36±.11	5.54±.10	5.67±.10	5.59±.10	5.62±.10	7.04±.07	25.71±.19	7.04±.07	7.04±.07	17.80±.15
Relaxed	6.73±.12	6.41±.11	6.38±.11	6.38±.11	6.38±.11	5.83±.05	6.63±.06	5.83±.05	5.83±.05	6.29±.06

The Sorry State of LBP

	Greedy	LBP	Combine	Exact	Relaxed	Greedy	LBP	Combine	Exact	Relaxed
	Scene Dataset			11.43±.29	18.10	Mediamill Dataset			18.60±.14	25.37
Greedy	10.67±.28	10.74±.28	10.67±.28	10.67±.28	10.67±.28	23.39±.16	25.66±.17	24.32±.17	24.92±.17	27.05±.18
LBP	10.45±.27	10.54±.27	10.45±.27	10.42±.27	10.49±.27	22.83±.16	22.83±.16	22.83±.16	22.83±.16	22.83±.16
Combine	10.72±.28	11.78±.30	10.72±.28	10.77±.28	11.20±.29	19.56±.14	20.12±.15	19.72±.14	19.82±.14	20.23±.15
Exact	10.08±.26	10.33±.27	10.08±.26	10.06±.26	10.20±.26	19.07±.14	27.23±.18	19.08±.14	18.75±.14	36.83±.21
Relaxed	10.55±.27	10.49±.27	10.49±.27	10.49±.27	10.49±.27	18.50±.14	18.26±.14	18.26±.14	18.21±.14	18.29±.14
	Yeast Dataset			20.91±.55	25.09	Synth1 Dataset			8.99±.08	16.34
Greedy	21.62±.56	21.77±.56	21.58±.56	21.62±.56	24.42±.61	8.86±.08	8.86±.08	8.86±.08	8.86±.08	8.86±.08
LBP	24.32±.61	24.32±.61	24.32±.61	24.32±.61	24.32±.61	13.94±.12	13.94±.12	13.94±.12	13.94±.12	13.94±.12
Combine	22.33±.57	37.24±.77	22.32±.57	21.82±.56	42.72±.81	8.86±.08	8.86±.08	8.86±.08	8.86±.08	8.86±.08
Exact	23.38±.59	21.99±.57	21.06±.55	20.23±.53	45.90±.82	6.89±.06	6.86±.06	6.86±.06	6.86±.06	6.86±.06
Relaxed	20.47±.54	20.45±.54	20.47±.54	20.48±.54	20.49±.54	8.94±.08	8.94±.08	8.94±.08	8.94±.08	8.94±.08
	Reuters Dataset			4.96±.09	15.80	Synth2 Dataset			9.80±.09	10.00
Greedy	5.32±.09	13.38±.21	5.06±.09	5.42±.09	16.98±.26	7.27±.07	27.92±.20	7.27±.07	7.28±.07	19.03±.15
LBP	15.80±.25	15.80±.25	15.80±.25	15.80±.25	15.80±.25	10.00±.09	10.00±.09	10.00±.09	10.00±.09	10.00±.09
Combine	4.90±.09	4.57±.08	4.53±.08	4.49±.08	4.55±.08	7.90±.07	26.39±.19	7.90±.07	7.90±.07	18.11±.15
Exact	6.36±.11	5.54±.10	5.67±.10	5.59±.10	5.62±.10	7.04±.07	25.71±.19	7.04±.07	7.04±.07	17.80±.15
Relaxed	6.73±.12	6.41±.11	6.38±.11	6.38±.11	6.38±.11	5.83±.05	6.63±.06	5.83±.05	5.83±.05	6.29±.06

- Models trained with LBP often have **terrible performance**.

The Sorry State of LBP

	Greedy	LBP	Combine	Exact	Relaxed	Greedy	LBP	Combine	Exact	Relaxed
	Scene Dataset			11.43±.29	18.10	Mediamill Dataset			18.60±.14	25.37
Greedy	10.67±.28	10.74±.28	10.67±.28	10.67±.28	10.67±.28	23.39±.16	25.66±.17	24.32±.17	24.92±.17	27.05±.18
LBP	10.45±.27	10.54±.27	10.45±.27	10.42±.27	10.49±.27	22.83±.16	22.83±.16	22.83±.16	22.83±.16	22.83±.16
Combine	10.72±.28	11.78±.30	10.72±.28	10.77±.28	11.20±.29	19.56±.14	20.12±.15	19.72±.14	19.82±.14	20.23±.15
Exact	10.08±.26	10.33±.27	10.08±.26	10.06±.26	10.20±.26	19.07±.14	27.23±.18	19.08±.14	18.75±.14	36.83±.21
Relaxed	10.55±.27	10.49±.27	10.49±.27	10.49±.27	10.49±.27	18.50±.14	18.26±.14	18.26±.14	18.21±.14	18.29±.14
	Yeast Dataset			20.91±.55	25.09	Synth1 Dataset			8.99±.08	16.34
Greedy	21.62±.56	21.77±.56	21.58±.56	21.62±.56	24.42±.61	8.86±.08	8.86±.08	8.86±.08	8.86±.08	8.86±.08
LBP	24.32±.61	24.32±.61	24.32±.61	24.32±.61	24.32±.61	13.94±.12	13.94±.12	13.94±.12	13.94±.12	13.94±.12
Combine	22.33±.57	37.24±.77	22.32±.57	21.82±.56	42.72±.81	8.86±.08	8.86±.08	8.86±.08	8.86±.08	8.86±.08
Exact	23.38±.59	21.99±.57	21.06±.55	20.23±.53	45.90±.82	6.89±.06	6.86±.06	6.86±.06	6.86±.06	6.86±.06
Relaxed	20.47±.54	20.45±.54	20.47±.54	20.48±.54	20.49±.54	8.94±.08	8.94±.08	8.94±.08	8.94±.08	8.94±.08
	Reuters Dataset			4.96±.09	15.80	Synth2 Dataset			9.80±.09	10.00
Greedy	5.32±.09	13.38±.21	5.06±.09	5.42±.09	16.98±.26	7.27±.07	27.92±.20	7.27±.07	7.28±.07	19.03±.15
LBP	15.80±.25	15.80±.25	15.80±.25	15.80±.25	15.80±.25	10.00±.09	10.00±.09	10.00±.09	10.00±.09	10.00±.09
Combine	4.90±.09	4.57±.08	4.53±.08	4.49±.08	4.55±.08	7.90±.07	26.39±.19	7.90±.07	7.90±.07	18.11±.15
Exact	6.36±.11	5.54±.10	5.67±.10	5.59±.10	5.62±.10	7.04±.07	25.71±.19	7.04±.07	7.04±.07	17.80±.15
Relaxed	6.73±.12	6.41±.11	6.38±.11	6.38±.11	6.38±.11	5.83±.05	6.63±.06	5.83±.05	5.83±.05	6.29±.06

- Models trained with LBP often have **terrible performance**.
- Predictions made with LBP also are often **quite poor**.

The Sorry State of LBP

	Greedy	LBP	Combine	Exact	Relaxed	Greedy	LBP	Combine	Exact	Relaxed
	Scene Dataset			11.43±.29	18.10	Mediamill Dataset			18.60±.14	25.37
Greedy	10.67±.28	10.74±.28	10.67±.28	10.67±.28	10.67±.28	23.39±.16	25.66±.17	24.32±.17	24.92±.17	27.05±.18
LBP	10.45±.27	10.54±.27	10.45±.27	10.42±.27	10.49±.27	22.83±.16	22.83±.16	22.83±.16	22.83±.16	22.83±.16
Combine	10.72±.28	11.78±.30	10.72±.28	10.77±.28	11.20±.29	19.56±.14	20.12±.15	19.72±.14	19.82±.14	20.23±.15
Exact	10.08±.26	10.33±.27	10.08±.26	10.06±.26	10.20±.26	19.07±.14	27.23±.18	19.08±.14	18.75±.14	36.83±.21
Relaxed	10.55±.27	10.49±.27	10.49±.27	10.49±.27	10.49±.27	18.50±.14	18.26±.14	18.26±.14	18.21±.14	18.29±.14
	Yeast Dataset			20.91±.55	25.09	Synth1 Dataset			8.99±.08	16.34
Greedy	21.62±.56	21.77±.56	21.58±.56	21.62±.56	24.42±.61	8.86±.08	8.86±.08	8.86±.08	8.86±.08	8.86±.08
LBP	24.32±.61	24.32±.61	24.32±.61	24.32±.61	24.32±.61	13.94±.12	13.94±.12	13.94±.12	13.94±.12	13.94±.12
Combine	22.33±.57	37.24±.77	22.32±.57	21.82±.56	42.72±.81	8.86±.08	8.86±.08	8.86±.08	8.86±.08	8.86±.08
Exact	23.38±.59	21.99±.57	21.06±.55	20.23±.53	45.90±.82	6.89±.06	6.86±.06	6.86±.06	6.86±.06	6.86±.06
Relaxed	20.47±.54	20.45±.54	20.47±.54	20.48±.54	20.49±.54	8.94±.08	8.94±.08	8.94±.08	8.94±.08	8.94±.08
	Reuters Dataset			4.96±.09	15.80	Synth2 Dataset			9.80±.09	10.00
Greedy	5.32±.09	13.38±.21	5.06±.09	5.42±.09	16.98±.26	7.27±.07	27.92±.20	7.27±.07	7.28±.07	19.03±.15
LBP	15.80±.25	15.80±.25	15.80±.25	15.80±.25	15.80±.25	10.00±.09	10.00±.09	10.00±.09	10.00±.09	10.00±.09
Combine	4.90±.09	4.57±.08	4.53±.08	4.49±.08	4.55±.08	7.90±.07	26.39±.19	7.90±.07	7.90±.07	18.11±.15
Exact	6.36±.11	5.54±.10	5.67±.10	5.59±.10	5.62±.10	7.04±.07	25.71±.19	7.04±.07	7.04±.07	17.80±.15
Relaxed	6.73±.12	6.41±.11	6.38±.11	6.38±.11	6.38±.11	5.83±.05	6.63±.06	5.83±.05	5.83±.05	6.29±.06

- Models trained with LBP often have **terrible performance**.
- Predictions made with LBP also are often **quite poor**.
- Likely explanation?

Relaxation

	Greedy	LBP	Combine	Exact	Relaxed	Greedy	LBP	Combine	Exact	Relaxed
	Scene Dataset			11.43±.29	18.10	Mediamill Dataset			18.60±.14	25.37
Greedy	10.67±.28	10.74±.28	10.67±.28	10.67±.28	10.67±.28	23.39±.16	25.66±.17	24.32±.17	24.92±.17	27.05±.18
LBP	10.45±.27	10.54±.27	10.45±.27	10.42±.27	10.49±.27	22.83±.16	22.83±.16	22.83±.16	22.83±.16	22.83±.16
Combine	10.72±.28	11.78±.30	10.72±.28	10.77±.28	11.20±.29	19.56±.14	20.12±.15	19.72±.14	19.82±.14	20.23±.15
Exact	10.08±.26	10.33±.27	10.08±.26	10.06±.26	10.20±.26	19.07±.14	27.23±.18	19.08±.14	18.75±.14	36.83±.21
Relaxed	10.55±.27	10.49±.27	10.49±.27	10.49±.27	10.49±.27	18.50±.14	18.26±.14	18.26±.14	18.21±.14	18.29±.14
	Yeast Dataset			20.91±.55	25.09	Synth1 Dataset			8.99±.08	16.34
Greedy	21.62±.56	21.77±.56	21.58±.56	21.62±.56	24.42±.61	8.86±.08	8.86±.08	8.86±.08	8.86±.08	8.86±.08
LBP	24.32±.61	24.32±.61	24.32±.61	24.32±.61	24.32±.61	13.94±.12	13.94±.12	13.94±.12	13.94±.12	13.94±.12
Combine	22.33±.57	37.24±.77	22.32±.57	21.82±.56	42.72±.81	8.86±.08	8.86±.08	8.86±.08	8.86±.08	8.86±.08
Exact	23.38±.59	21.99±.57	21.06±.55	20.23±.53	45.90±.82	6.89±.06	6.86±.06	6.86±.06	6.86±.06	6.86±.06
Relaxed	20.47±.54	20.45±.54	20.47±.54	20.48±.54	20.49±.54	8.94±.08	8.94±.08	8.94±.08	8.94±.08	8.94±.08
	Reuters Dataset			4.96±.09	15.80	Synth2 Dataset			9.80±.09	10.00
Greedy	5.32±.09	13.38±.21	5.06±.09	5.42±.09	16.98±.26	7.27±.07	27.92±.20	7.27±.07	7.28±.07	19.03±.15
LBP	15.80±.25	15.80±.25	15.80±.25	15.80±.25	15.80±.25	10.00±.09	10.00±.09	10.00±.09	10.00±.09	10.00±.09
Combine	4.90±.09	4.57±.08	4.53±.08	4.49±.08	4.55±.08	7.90±.07	26.39±.19	7.90±.07	7.90±.07	18.11±.15
Exact	6.36±.11	5.54±.10	5.67±.10	5.59±.10	5.62±.10	7.04±.07	25.71±.19	7.04±.07	7.04±.07	17.80±.15
Relaxed	6.73±.12	6.41±.11	6.38±.11	6.38±.11	6.38±.11	5.83±.05	6.63±.06	5.83±.05	5.83±.05	6.29±.06

Relaxation

	Greedy	LBP	Combine	Exact	Relaxed
	Scene Dataset			11.43±.29	18.10
Greedy	10.67±.28	10.74±.28	10.67±.28	10.67±.28	10.67±.28
LBP	10.45±.27	10.54±.27	10.45±.27	10.42±.27	10.49±.27
Combine	10.72±.28	11.78±.30	10.72±.28	10.77±.28	11.20±.29
Exact	10.08±.26	10.33±.27	10.08±.26	10.06±.26	10.20±.26
Relaxed	10.55±.27	10.49±.27	10.49±.27	10.49±.27	10.49±.27
	Yeast Dataset			20.91±.55	25.09
Greedy	21.62±.56	21.77±.56	21.58±.56	21.62±.56	24.42±.61
LBP	24.32±.61	24.32±.61	24.32±.61	24.32±.61	24.32±.61
Combine	22.33±.57	37.24±.77	22.32±.57	21.82±.56	42.72±.81
Exact	23.38±.59	21.99±.57	21.06±.55	20.23±.53	45.90±.82
Relaxed	20.47±.54	20.45±.54	20.47±.54	20.48±.54	20.49±.54
	Reuters Dataset			4.96±.09	15.80
Greedy	5.32±.09	13.38±.21	5.06±.09	5.42±.09	16.98±.26
LBP	15.80±.25	15.80±.25	15.80±.25	15.80±.25	15.80±.25
Combine	4.90±.09	4.57±.08	4.53±.08	4.49±.08	4.55±.08
Exact	6.36±.11	5.54±.10	5.67±.10	5.59±.10	5.62±.10
Relaxed	6.73±.12	6.41±.11	6.38±.11	6.38±.11	6.38±.11

	Greedy	LBP	Combine	Exact	Relaxed
	Mediamill Dataset			18.60±.14	25.37
Greedy	23.39±.16	25.66±.17	24.32±.17	24.92±.17	27.05±.18
LBP	22.83±.16	22.83±.16	22.83±.16	22.83±.16	22.83±.16
Combine	19.56±.14	20.12±.15	19.72±.14	19.82±.14	20.23±.15
Exact	19.07±.14	27.23±.18	19.08±.14	18.75±.14	36.83±.21
Relaxed	18.50±.14	18.26±.14	18.26±.14	18.21±.14	18.29±.14
	Synth1 Dataset			8.99±.08	16.34
Greedy	8.86±.08	8.86±.08	8.86±.08	8.86±.08	8.86±.08
LBP	13.94±.12	13.94±.12	13.94±.12	13.94±.12	13.94±.12
Combine	8.86±.08	8.86±.08	8.86±.08	8.86±.08	8.86±.08
Exact	6.89±.06	6.86±.06	6.86±.06	6.86±.06	6.86±.06
Relaxed	8.94±.08	8.94±.08	8.94±.08	8.94±.08	8.94±.08
	Synth2 Dataset			9.80±.09	10.00
Greedy	7.27±.07	27.92±.20	7.27±.07	7.28±.07	19.03±.15
LBP	10.00±.09	10.00±.09	10.00±.09	10.00±.09	10.00±.09
Combine	7.90±.07	26.39±.19	7.90±.07	7.90±.07	18.11±.15
Exact	7.04±.07	25.71±.19	7.04±.07	7.04±.07	17.80±.15
Relaxed	5.83±.05	6.63±.06	5.83±.05	5.83±.05	6.29±.06

- Notice predictor **consistency** with relaxed trained models.

Relaxation

	Greedy	LBP	Combine	Exact	Relaxed	Greedy	LBP	Combine	Exact	Relaxed
	Scene Dataset			11.43±.29	18.10	Mediamill Dataset			18.60±.14	25.37
Greedy	10.67±.28	10.74±.28	10.67±.28	10.67±.28	10.67±.28	23.39±.16	25.66±.17	24.32±.17	24.92±.17	27.05±.18
LBP	10.45±.27	10.54±.27	10.45±.27	10.42±.27	10.49±.27	22.83±.16	22.83±.16	22.83±.16	22.83±.16	22.83±.16
Combine	10.72±.28	11.78±.30	10.72±.28	10.77±.28	11.20±.29	19.56±.14	20.12±.15	19.72±.14	19.82±.14	20.23±.15
Exact	10.08±.26	10.33±.27	10.08±.26	10.06±.26	10.20±.26	19.07±.14	27.23±.18	19.08±.14	18.75±.14	36.83±.21
Relaxed	10.55±.27	10.49±.27	10.49±.27	10.49±.27	10.49±.27	18.50±.14	18.26±.14	18.26±.14	18.21±.14	18.29±.14
	Yeast Dataset			20.91±.55	25.09	Synth1 Dataset			8.99±.08	16.34
Greedy	21.62±.56	21.77±.56	21.58±.56	21.62±.56	24.42±.61	8.86±.08	8.86±.08	8.86±.08	8.86±.08	8.86±.08
LBP	24.32±.61	24.32±.61	24.32±.61	24.32±.61	24.32±.61	13.94±.12	13.94±.12	13.94±.12	13.94±.12	13.94±.12
Combine	22.33±.57	37.24±.77	22.32±.57	21.82±.56	42.72±.81	8.86±.08	8.86±.08	8.86±.08	8.86±.08	8.86±.08
Exact	23.38±.59	21.99±.57	21.06±.55	20.23±.53	45.90±.82	6.89±.06	6.86±.06	6.86±.06	6.86±.06	6.86±.06
Relaxed	20.47±.54	20.45±.54	20.47±.54	20.48±.54	20.49±.54	8.94±.08	8.94±.08	8.94±.08	8.94±.08	8.94±.08
	Reuters Dataset			4.96±.09	15.80	Synth2 Dataset			9.80±.09	10.00
Greedy	5.32±.09	13.38±.21	5.06±.09	5.42±.09	16.98±.26	7.27±.07	27.92±.20	7.27±.07	7.28±.07	19.03±.15
LBP	15.80±.25	15.80±.25	15.80±.25	15.80±.25	15.80±.25	10.00±.09	10.00±.09	10.00±.09	10.00±.09	10.00±.09
Combine	4.90±.09	4.57±.08	4.53±.08	4.49±.08	4.55±.08	7.90±.07	26.39±.19	7.90±.07	7.90±.07	18.11±.15
Exact	6.36±.11	5.54±.10	5.67±.10	5.59±.10	5.62±.10	7.04±.07	25.71±.19	7.04±.07	7.04±.07	17.80±.15
Relaxed	6.73±.12	6.41±.11	6.38±.11	6.38±.11	6.38±.11	5.83±.05	6.63±.06	5.83±.05	5.83±.05	6.29±.06

- Notice predictor **consistency** with relaxed trained models.
- Notice occasional **ludicrously poor** performance of relaxation as a classifier.

Relaxation

	Greedy	LBP	Combine	Exact	Relaxed	Greedy	LBP	Combine	Exact	Relaxed
	Scene Dataset			11.43±.29	18.10	Mediamill Dataset			18.60±.14	25.37
Greedy	10.67±.28	10.74±.28	10.67±.28	10.67±.28	10.67±.28	23.39±.16	25.66±.17	24.32±.17	24.92±.17	27.05±.18
LBP	10.45±.27	10.54±.27	10.45±.27	10.42±.27	10.49±.27	22.83±.16	22.83±.16	22.83±.16	22.83±.16	22.83±.16
Combine	10.72±.28	11.78±.30	10.72±.28	10.77±.28	11.20±.29	19.56±.14	20.12±.15	19.72±.14	19.82±.14	20.23±.15
Exact	10.08±.26	10.33±.27	10.08±.26	10.06±.26	10.20±.26	19.07±.14	27.23±.18	19.08±.14	18.75±.14	36.83±.21
Relaxed	10.55±.27	10.49±.27	10.49±.27	10.49±.27	10.49±.27	18.50±.14	18.26±.14	18.26±.14	18.21±.14	18.29±.14
	Yeast Dataset			20.91±.55	25.09	Synth1 Dataset			8.99±.08	16.34
Greedy	21.62±.56	21.77±.56	21.58±.56	21.62±.56	24.42±.61	8.86±.08	8.86±.08	8.86±.08	8.86±.08	8.86±.08
LBP	24.32±.61	24.32±.61	24.32±.61	24.32±.61	24.32±.61	13.94±.12	13.94±.12	13.94±.12	13.94±.12	13.94±.12
Combine	22.33±.57	37.24±.77	22.32±.57	21.82±.56	42.72±.81	8.86±.08	8.86±.08	8.86±.08	8.86±.08	8.86±.08
Exact	23.38±.59	21.99±.57	21.06±.55	20.23±.53	45.90±.82	6.89±.06	6.86±.06	6.86±.06	6.86±.06	6.86±.06
Relaxed	20.47±.54	20.45±.54	20.47±.54	20.48±.54	20.49±.54	8.94±.08	8.94±.08	8.94±.08	8.94±.08	8.94±.08
	Reuters Dataset			4.96±.09	15.80	Synth2 Dataset			9.80±.09	10.00
Greedy	5.32±.09	13.38±.21	5.06±.09	5.42±.09	16.98±.26	7.27±.07	27.92±.20	7.27±.07	7.28±.07	19.03±.15
LBP	15.80±.25	15.80±.25	15.80±.25	15.80±.25	15.80±.25	10.00±.09	10.00±.09	10.00±.09	10.00±.09	10.00±.09
Combine	4.90±.09	4.57±.08	4.53±.08	4.49±.08	4.55±.08	7.90±.07	26.39±.19	7.90±.07	7.90±.07	18.11±.15
Exact	6.36±.11	5.54±.10	5.67±.10	5.59±.10	5.62±.10	7.04±.07	25.71±.19	7.04±.07	7.04±.07	17.80±.15
Relaxed	6.73±.12	6.41±.11	6.38±.11	6.38±.11	6.38±.11	5.83±.05	6.63±.06	5.83±.05	5.83±.05	6.29±.06

- Notice predictor **consistency** with relaxed trained models.
- Notice occasional **ludicrously poor** performance of relaxation as a classifier.
- Presence of fractional constraints leads to "smoothed" **easier** space.

Relaxation

	Greedy	LBP	Combine	Exact	Relaxed	Greedy	LBP	Combine	Exact	Relaxed
	Scene Dataset			11.43±.29	18.10	Mediamill Dataset			18.60±.14	25.37
Greedy	10.67±.28	10.74±.28	10.67±.28	10.67±.28	10.67±.28	23.39±.16	25.66±.17	24.32±.17	24.92±.17	27.05±.18
LBP	10.45±.27	10.54±.27	10.45±.27	10.42±.27	10.49±.27	22.83±.16	22.83±.16	22.83±.16	22.83±.16	22.83±.16
Combine	10.72±.28	11.78±.30	10.72±.28	10.77±.28	11.20±.29	19.56±.14	20.12±.15	19.72±.14	19.82±.14	20.23±.15
Exact	10.08±.26	10.33±.27	10.08±.26	10.06±.26	10.20±.26	19.07±.14	27.23±.18	19.08±.14	18.75±.14	36.83±.21
Relaxed	10.55±.27	10.49±.27	10.49±.27	10.49±.27	10.49±.27	18.50±.14	18.26±.14	18.26±.14	18.21±.14	18.29±.14
	Yeast Dataset			20.91±.55	25.09	Synth1 Dataset			8.99±.08	16.34
Greedy	21.62±.56	21.77±.56	21.58±.56	21.62±.56	24.42±.61	8.86±.08	8.86±.08	8.86±.08	8.86±.08	8.86±.08
LBP	24.32±.61	24.32±.61	24.32±.61	24.32±.61	24.32±.61	13.94±.12	13.94±.12	13.94±.12	13.94±.12	13.94±.12
Combine	22.33±.57	37.24±.77	22.32±.57	21.82±.56	42.72±.81	8.86±.08	8.86±.08	8.86±.08	8.86±.08	8.86±.08
Exact	23.38±.59	21.99±.57	21.06±.55	20.23±.53	45.90±.82	6.89±.06	6.86±.06	6.86±.06	6.86±.06	6.86±.06
Relaxed	20.47±.54	20.45±.54	20.47±.54	20.48±.54	20.49±.54	8.94±.08	8.94±.08	8.94±.08	8.94±.08	8.94±.08
	Reuters Dataset			4.96±.09	15.80	Synth2 Dataset			9.80±.09	10.00
Greedy	5.32±.09	13.38±.21	5.06±.09	5.42±.09	16.98±.26	7.27±.07	27.92±.20	7.27±.07	7.28±.07	19.03±.15
LBP	15.80±.25	15.80±.25	15.80±.25	15.80±.25	15.80±.25	10.00±.09	10.00±.09	10.00±.09	10.00±.09	10.00±.09
Combine	4.90±.09	4.57±.08	4.53±.08	4.49±.08	4.55±.08	7.90±.07	26.39±.19	7.90±.07	7.90±.07	18.11±.15
Exact	6.36±.11	5.54±.10	5.67±.10	5.59±.10	5.62±.10	7.04±.07	25.71±.19	7.04±.07	7.04±.07	17.80±.15
Relaxed	6.73±.12	6.41±.11	6.38±.11	6.38±.11	6.38±.11	5.83±.05	6.63±.06	5.83±.05	5.83±.05	6.29±.06

- Notice predictor **consistency** with relaxed trained models.
- Notice occasional **ludicrously poor** performance of relaxation as a classifier.
- Presence of fractional constraints leads to "smoothed" **easier** space.
- Lack of fractional constraints in other models **hurts** relaxed predictor.