Hierarchical Active Transfer Learning

Dave Kale[1] Marjan Ghazvininejad[1] Anil Ramakrishna[1]
Jingrui He[2] Yan Liu[1]

[1]University of Southern California
[2]Arizona State University

May 1, 2015

Learning with few or no labels

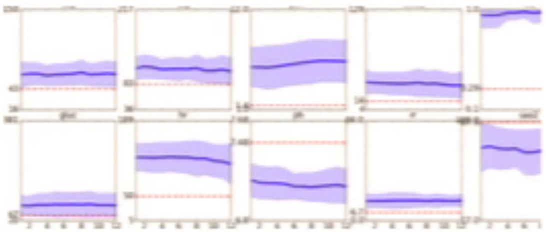
A significant challenge in many domains – even in 21st century!

$\Longleftarrow$ Cost of acquiring observations $\ll$ cost of acquiring labels

Learning with few or no labels

A significant challenge in many domains – even in 21st century!

$\Longleftarrow$ Cost of acquiring observations $\ll$ cost of acquiring labels



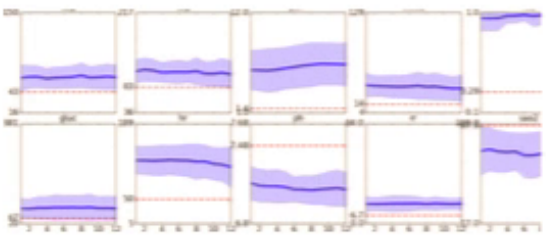
Health

- Finding patients with specific diseases in EHRs (*phenotyping* [8])
- Identifying cancer in radiologic images [9]

Learning with few or no labels

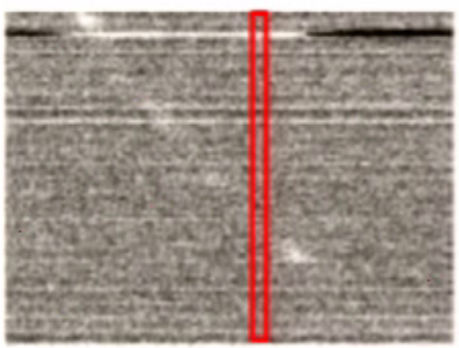
A significant challenge in many domains – even in 21st century!

$\Longleftarrow$ Cost of acquiring observations $\ll$ cost of acquiring labels



Health

- Finding patients with specific diseases in EHRs (*phenotyping* [8])
- Identifying cancer in radiologic images [9]



Science

- Detecting rare, transient events in astronomical sensor data [10]
- Virtual screening in drug discovery [11]

Personalization

- Cold starts in recommender systems [12]
- Activity recognition in wearables [13]

Learning with few or no labels: *solutions*

Active Learning: ask *oracle* to label only most informative examples

- Substantially reduce amount of labeled data needed
- *Example:* doctor labels CT-scans near SVM decision boundary
- Cold start problem: start w/0 labels, must use random sampling [5]

Learning with few or no labels: *solutions*

Active Learning: ask *oracle* to label only most informative examples

- Substantially reduce amount of labeled data needed
- *Example:* doctor labels CT-scans near SVM decision boundary
- Cold start problem: start w/0 labels, must use random sampling [5]

Transfer Learning: apply knowledge from similar problems to new one
(related: domain adaptation, covariate shift, multi-task learning, etc.)

- *Source* provides a useful initial bias for *Target*, can then be adapted
- *Example:* adapt a diagnostic model for pediatric patients to adults
- Negative transfer problem: tasks too different or adaptation fails
- *Also:* Can be difficult to define task/domain similarity appropriately

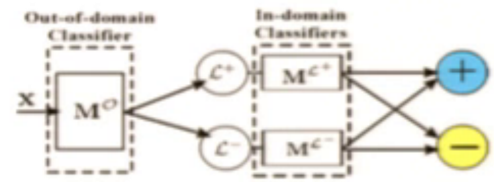
Complementary strengths...*why not combine?* Active+Transfer Learning!

Active+Transfer Learning: related work

Shi, Fan, and Ren, ECML 2008 [2] use *Transfer* classifier (TC) to choose queries

Partitioning Target data using Source model, train/combine separate classifiers

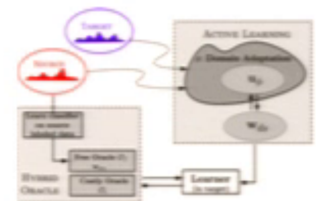
Query label when TC disagrees with classifier trained on $\mathcal{L}^{\text{all}}$



Saha, et al., ECML 2011 [3]: two-step TL, then AL

Reweight Source points to match Source, Target $P(X)$'s

Use Source classifier as free "oracle" to label Target points



Chattopadhay, et al., ICML 2013 [4]: Joint Optimization for TL and AL (JOTAL)

Jointly reweight source points, query target labels

Objective: match $P(X)$'s for labeled (Source+Target), unlabeled data

$$\left\| \frac{1}{n_s+n_l+b}\left(\sum_{i\in S}\beta_i\Phi(x_i)+\sum_{j\in L}\Phi(x_j)+\sum_{i\in U}\alpha_i\Phi(x_i)\right) - \frac{1}{n_u-b}\sum_{i\in U}(1-\alpha_i)\Phi(x_i)\right\|_{\mathcal{H}}^2,$$

$$\text{s.t.}\quad \alpha_i\in\{0,1\},\beta_i\in[0,1],\alpha^T\mathbf{1}=b.$$

Kale and Liu, ICDM 2013 [5]: Transfer Importance-weighted Consistent AL (TIWCAL)

TL (convex combination of losses) + AL (IWCAL)

Intuitive upper bound on Target generalization error

$$\epsilon_T(\hat{h}_t) \le \epsilon_T(h_T^*) + \alpha\left(\sqrt{\frac{2C_0\log(t+1)}{t}}+\frac{2C_0\log(t+1)}{t}\right) + 2(1-\alpha)\left(\sqrt{\frac{C_0\log 2}{2m}}+\frac{1}{2}d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S,\mathcal{D}_T)+\epsilon^*_{ST}\right)$$

Active+Transfer Learning: opportunities for innovation

- AL algorithms with theoretical guarantees (e.g., consistency, no bias)
- TL frameworks with fewer or no assumptions (vs., e.g., MMD)
- Transfer by adapting $P(Y|X)$ as well as $P(X)$
- General learning framework for AL, TL, semi-supervised learning, etc.

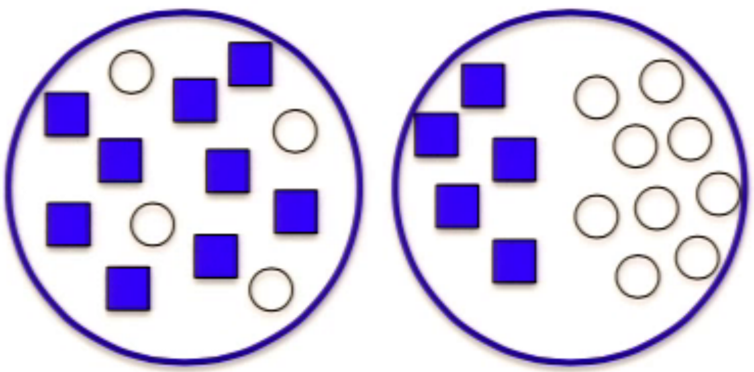
Active+Transfer Learning: opportunities for innovation

- AL algorithms with theoretical guarantees (e.g., consistency, no bias)
  Utilizes theoretically sound HSAL for AL [1]
- TL frameworks with fewer or no assumptions (vs., e.g., MMD)
  Clustering to capture similarities, differences in $P(X)$
- Transfer by adapting $P(Y|X)$ as well as $P(X)$
  TL by relabeling both Source, Target points
- General learning framework for AL, TL, semi-supervised learning, etc.
  Can be used for AL, TL, ATL, semi-supervised learning (SSL)

Hierarchical Active Transfer Learning: a step in this direction!

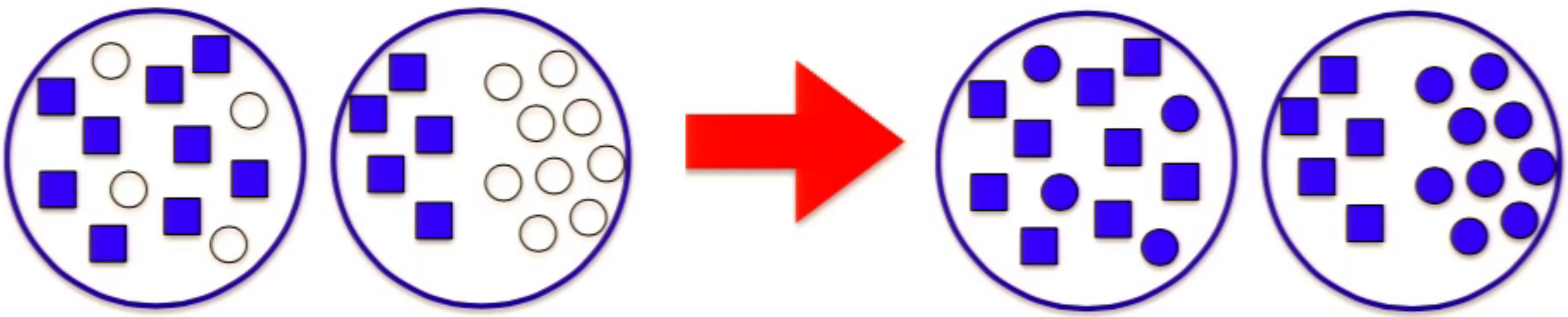
HATL *core intuition:* leverage cluster structure for TL,AL

2 clusters, source labels only



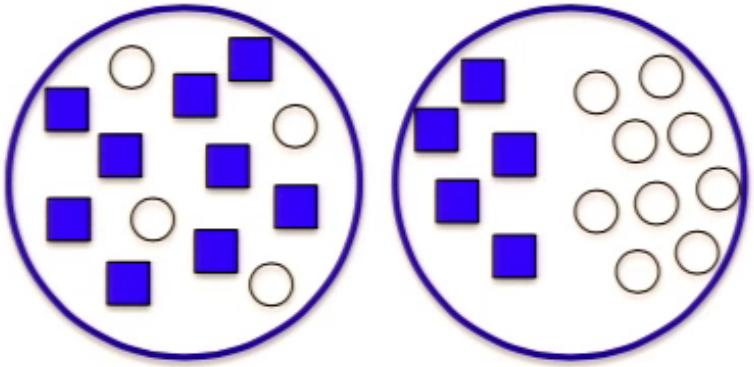
HATL *core intuition:* leverage cluster structure for TL,AL

2 clusters, source labels only | Impute cluster labels (*TL*)

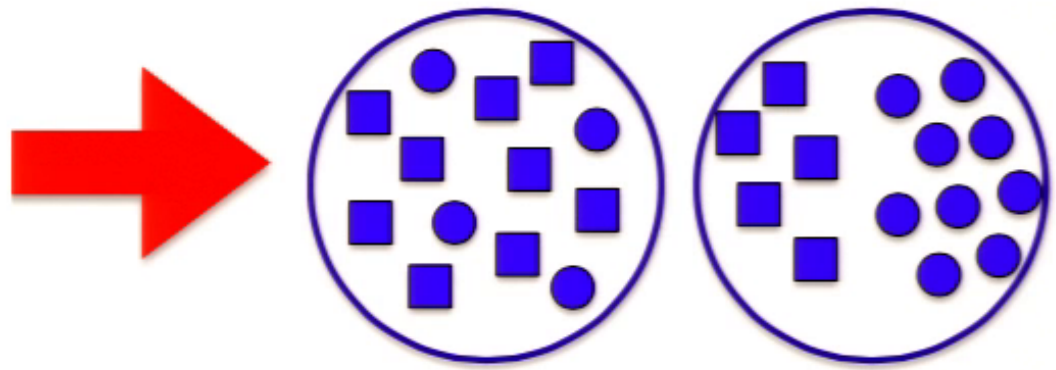


HATL *core intuition:* leverage cluster structure for TL,AL

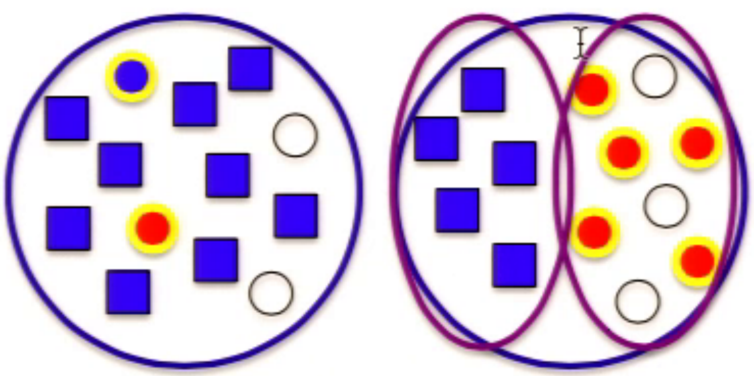
2 clusters, source labels only



Impute cluster labels (*TL*)



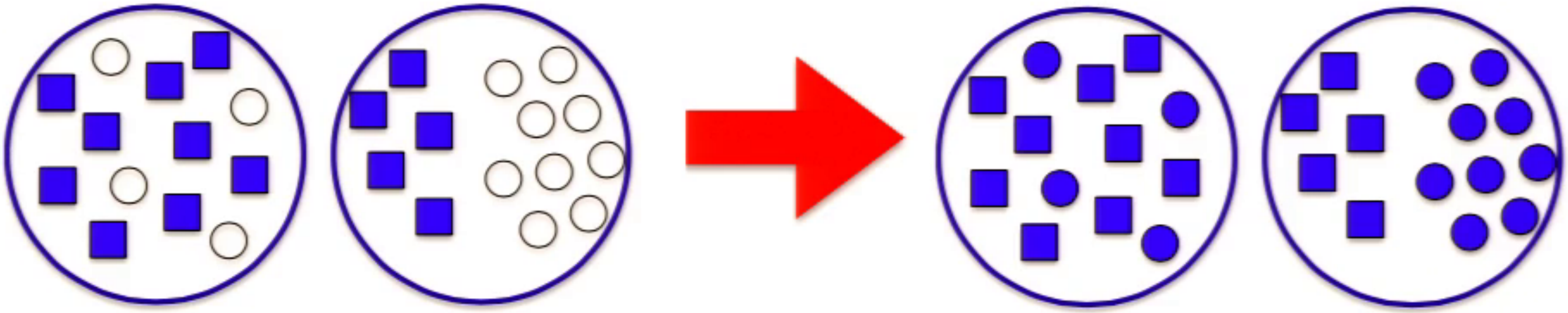
Query $\propto$ cluster size (*AL*)



HATL *core intuition:* leverage cluster structure for TL,AL

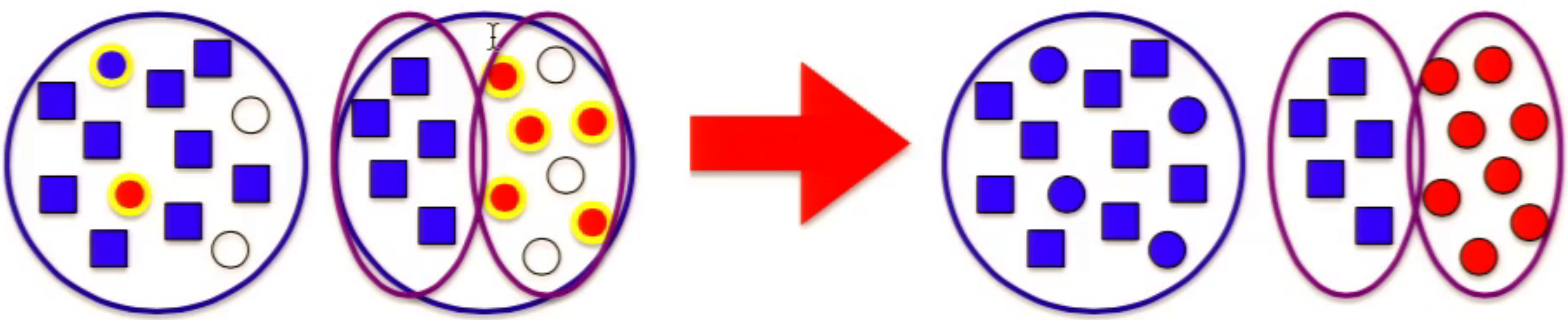
2 clusters, source labels only

Impute cluster labels (*TL*)



Query $\propto$ cluster size (*AL*)

Split cluster, impute (*TL*)



HATL Overview

Inspired by *Hierarchical Sampling for Active Learning* (HSAL) [1]

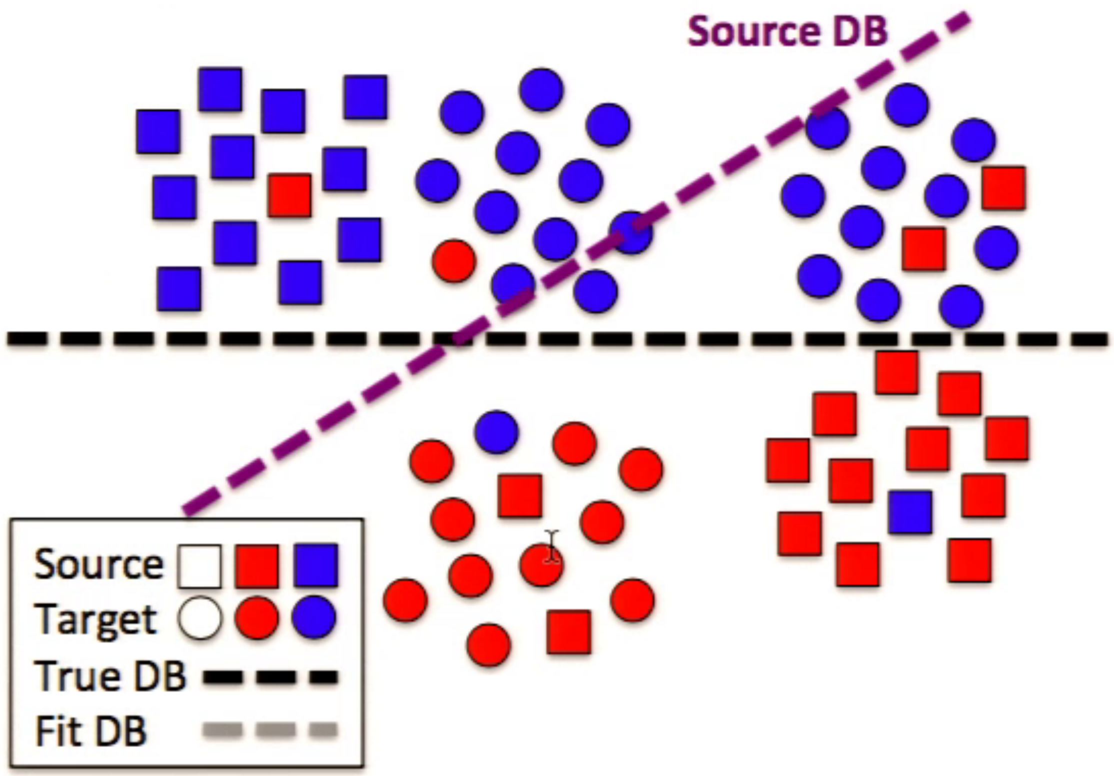
Inputs: Source $\mathcal{X}_{\mathrm{S}}$, Target $\mathcal{X}_{\mathrm{T}}$, cluster tree T, budget B
Initialize pruning $P = 0$ (i.e., root), root label $L_0 = 0$

For each cluster $v \in T$*, label* ℓ*: estimate CI for counts:* $[C^l_{v,\ell}, C^u_{v,\ell}]$

- UpdateLabelCounts($\mathcal{X}_{\mathrm{S}}$)
- $P \leftarrow$ UpdatePruning(P)
- Run *HSAL* algorithm for B queries
  - $(v, x, y) \leftarrow$ GetNextQueryAndLabel(P)
  - UpdateLabelCounts($\{(x, y)\}$)
  - $P \leftarrow$ UpdatePruning(P)
- $\hat{y}(x) \leftarrow L_v$ for all $x \in v$, for each $v \in P$

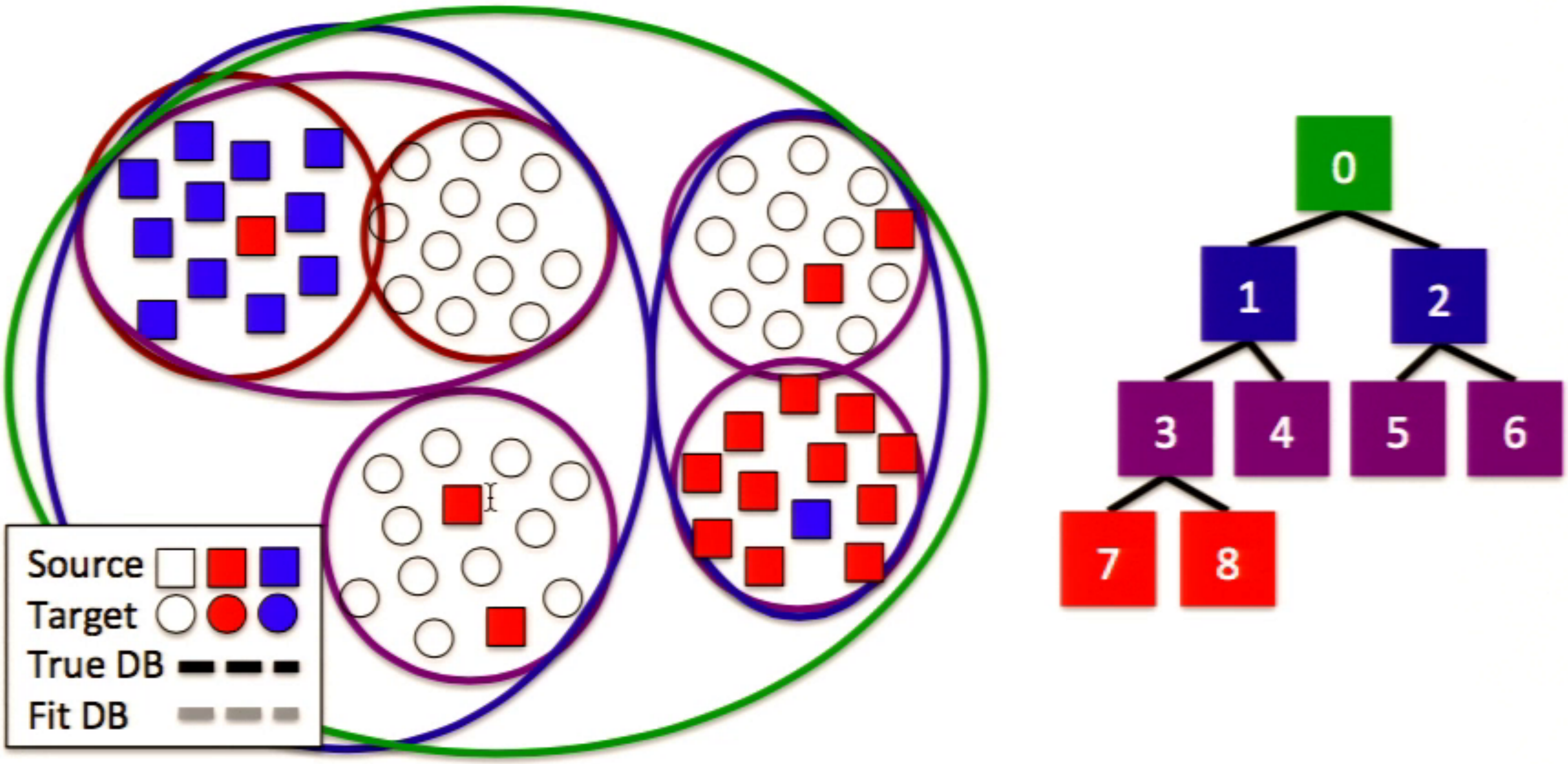
HATL in action

Assume these data, labeled w/dashed decision boundaries (w/some noise).



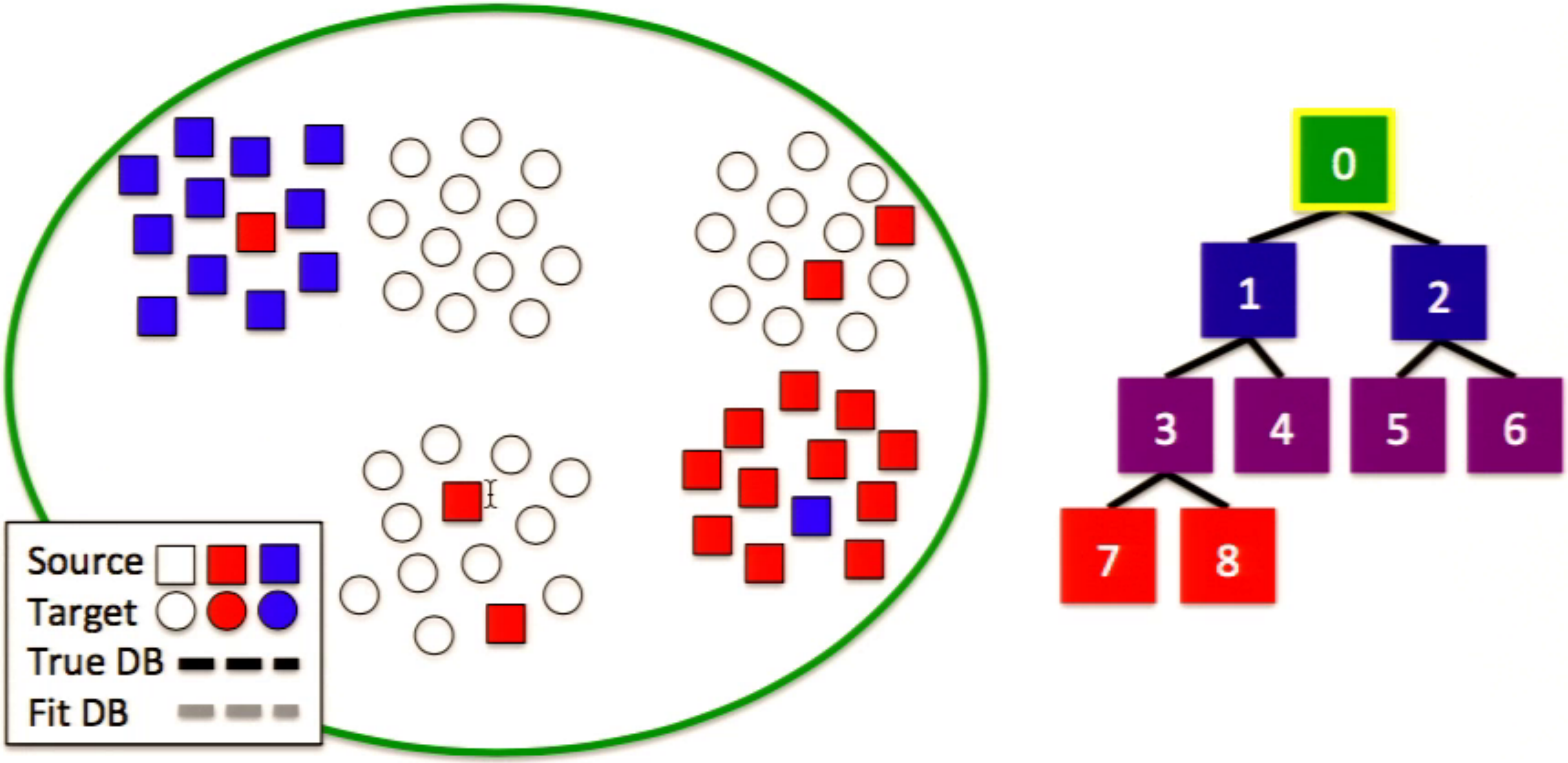
HATL in action

Start with *only* Source labels. Construct a hierarchical clustering.



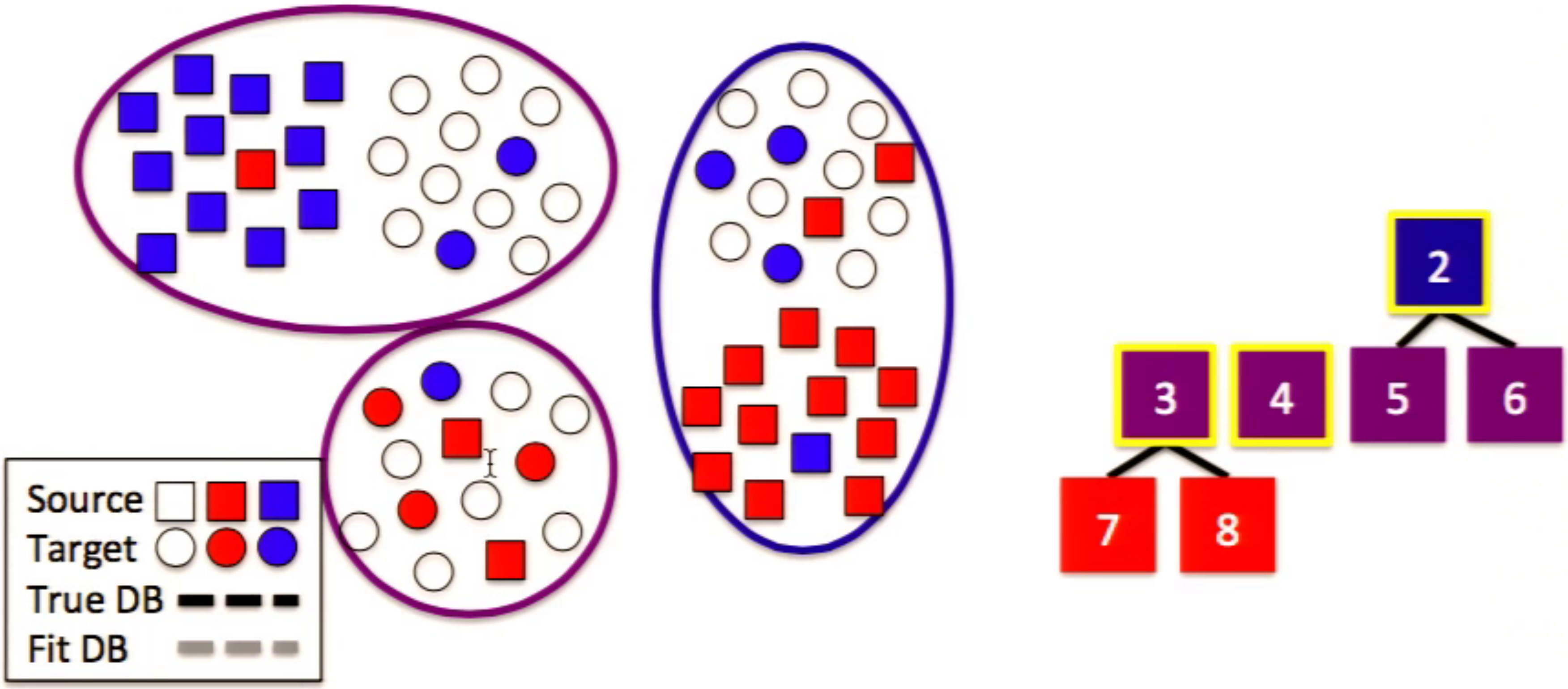
HATL in action

Initialize $P = \{0\}$ (1 cluster). UpdateLabelCounts($\mathcal{X}_S$).



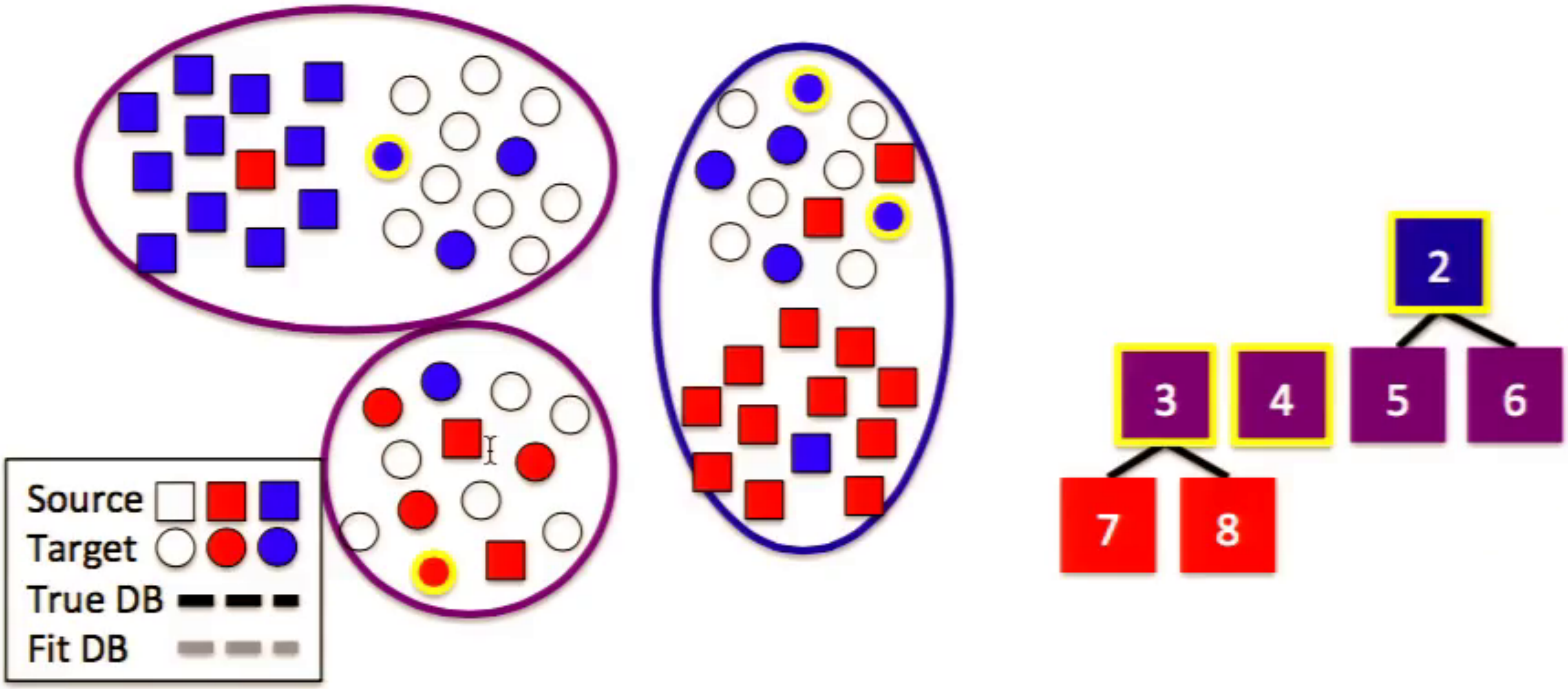
HATL in action

Iteratively query labels, update counts, refine pruning: $P = \{2, 3, 4\}$.



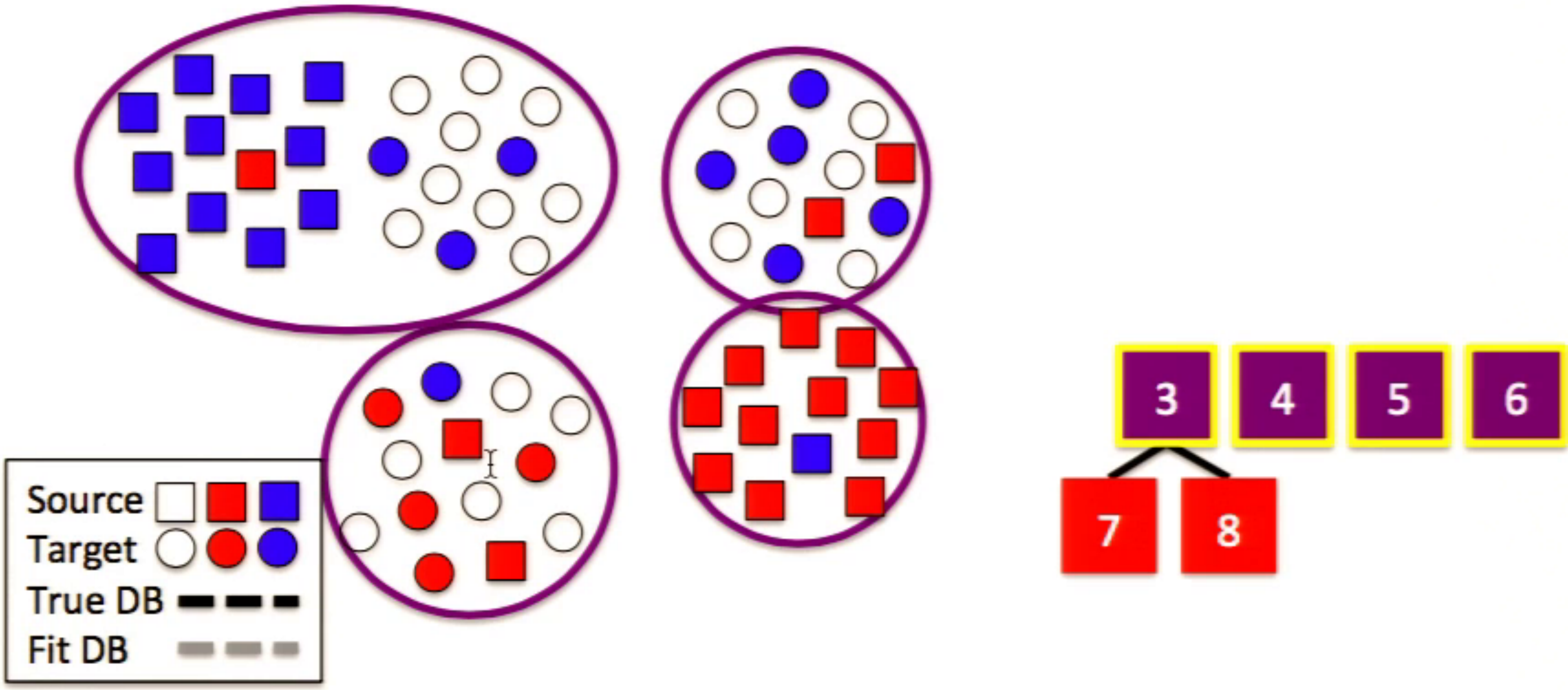
HATL in action

$(2, x, \text{Blue}) \leftarrow$ GetNextQueryAndLabel$(\{2, 3, 4\})$
UpdateLabelCounts$(\{(x, y)\})$



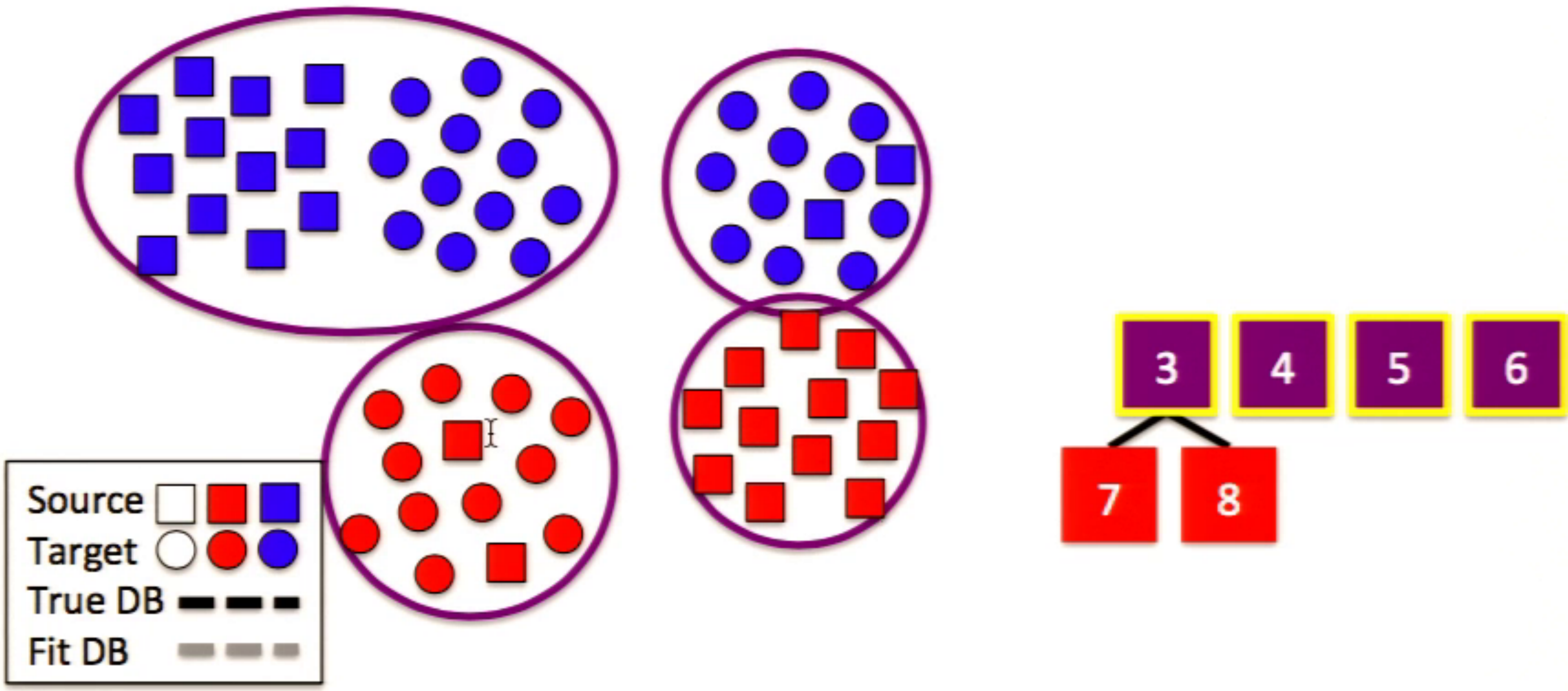
HATL in action

$\{3, 4, 5, 6\} \leftarrow \text{UpdatePruning}(\{2, 3, 4\})$



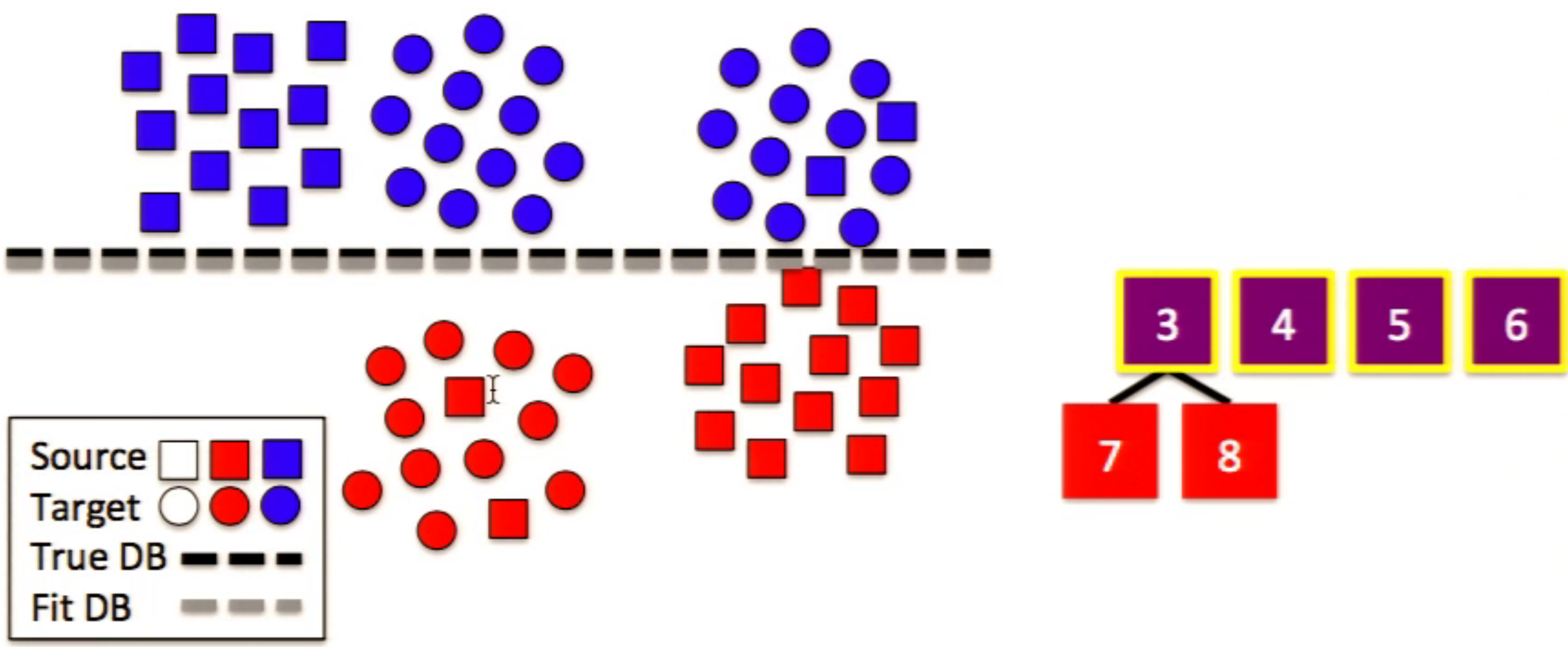
HATL in action

Label imputation after 13 Target queries.



HATL in action

Final large margin classifier.



Expected *label imputation error* for Target data

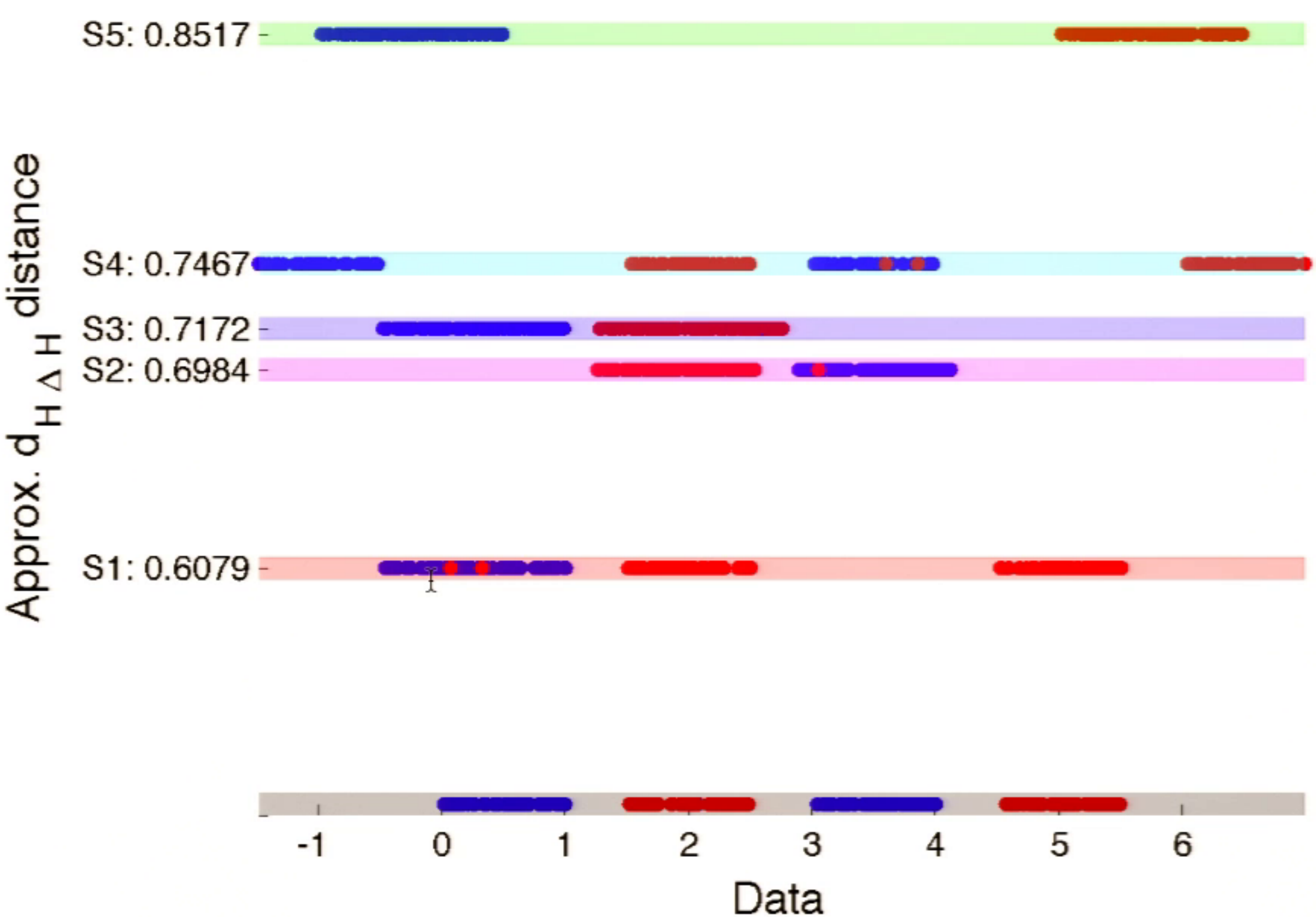
Theorem (Upper bound on target label imputation error)

Choose $\delta, \eta > 0$*. We can find an "optimal" pruning* P^* *of* T *with overall (Source+Target) label imputation error* $\epsilon(P^*) \leq \eta$*. Suppose HATL discovers pruning* P *after* B *queries. With probability at least* $1 - \delta$*, the Target label imputation error* $\epsilon_{\mathrm{T}}(P)$ *is*

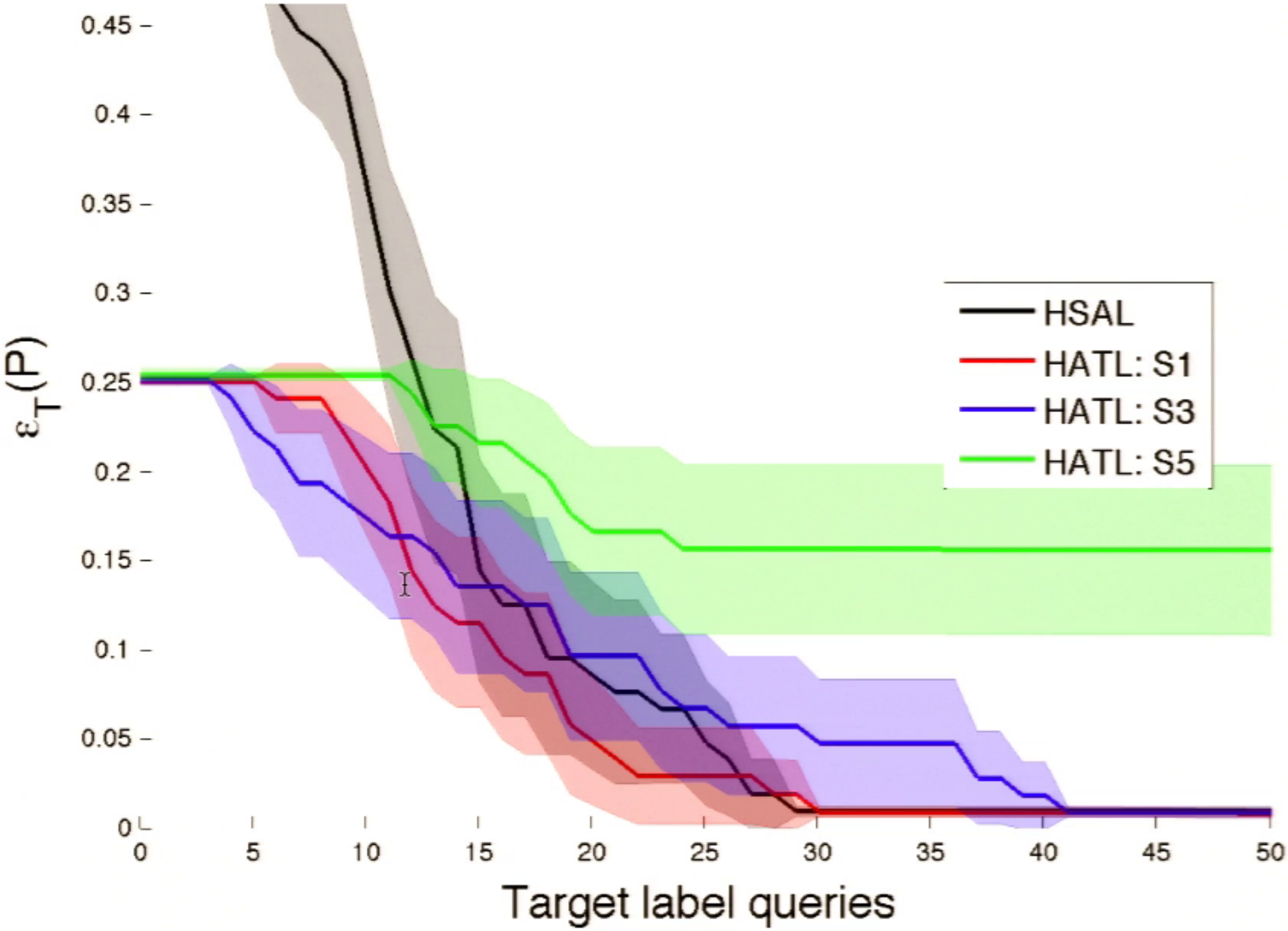
$$\epsilon_{\mathrm{T}}(P) \leq \tilde{O}\left(\epsilon(P^*) + \eta + \frac{1-\alpha}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_{\mathrm{S}}, \mathcal{D}_{\mathrm{T}})\right)$$

- Optimal B is function of size, depth of P^*, η, δ (defined as in HSAL)
- $\alpha = N_{\mathrm{T}}/(N_{\mathrm{S}} + N_{\mathrm{T}})$: fraction of data from Target domain
- $d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_{\mathrm{S}}, \mathcal{D}_{\mathrm{T}})$ [6]: distance between Source, Target distributions
  - hypothesis-class dependent similarity for distributions
  - can approximate from samples using *domain separator* classifer

Experiments with synthetic 1D data sets



Smaller $d_{H\Delta H}$ distance $\Longrightarrow$ target error decreases faster



Sentiment classification experiments

Ran experiments with sentiment classification data set from [7]

- Benchmark data set for transfer learning, domain adaptation
- *Task:* classify sentiment of Amazon product reviews
- 4 categories, 2000 reviews per category

Sentiment classification experiments

Ran experiments with sentiment classification data set from [7]

Benchmark data set for transfer learning, domain adaptation

Task: classify sentiment of Amazon product reviews

4 categories, 2000 reviews per category

Experimental setup:

Target: 1800 unlabeled examples, e.g., kitchen

Source: 400 labeled, 1600 unlabeled examples from, e.g., dvd

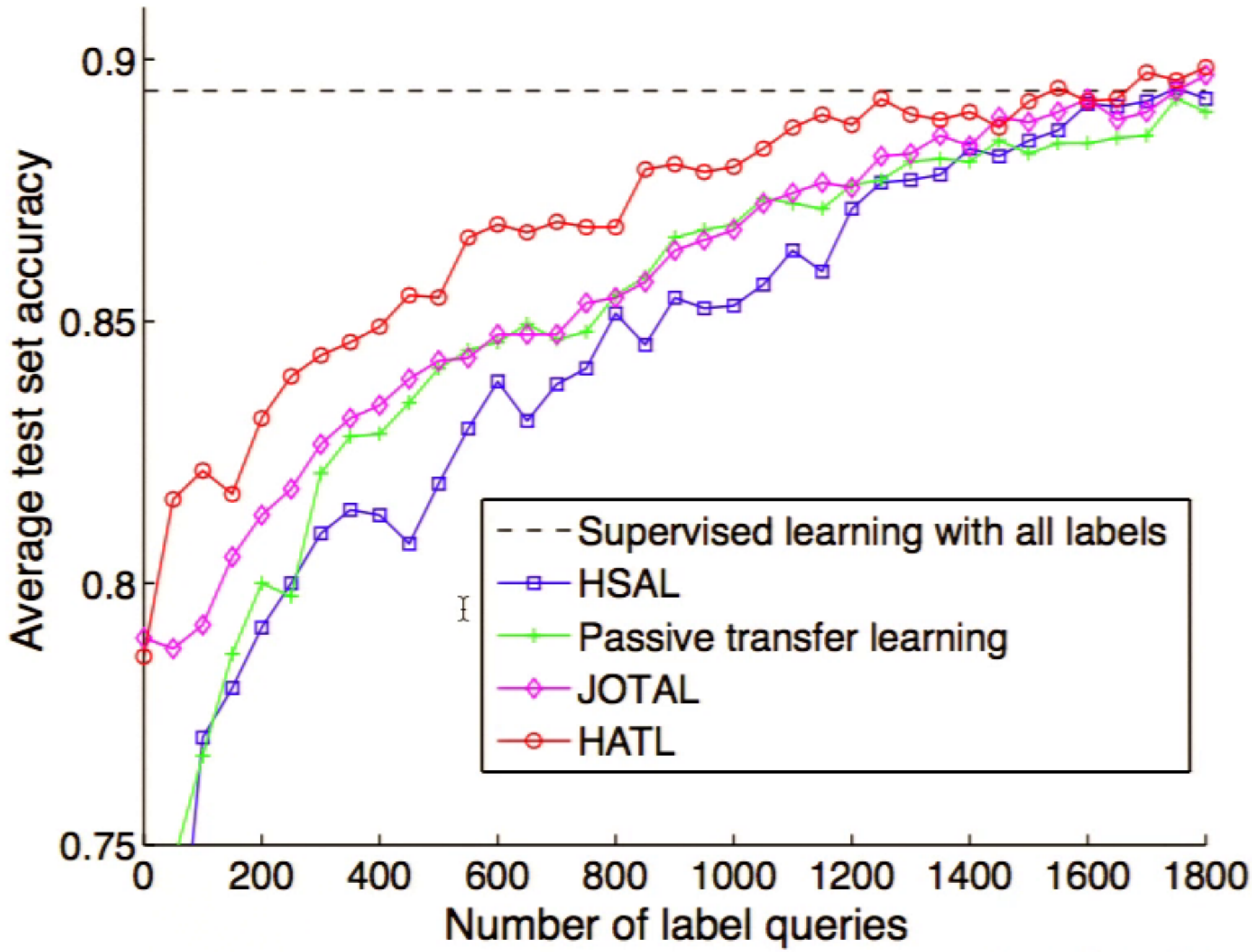
Compare against passive transfer learning, HSAL [1], JOTAL [**?**]

Query target labels, then train linear SVM with L2 regularization

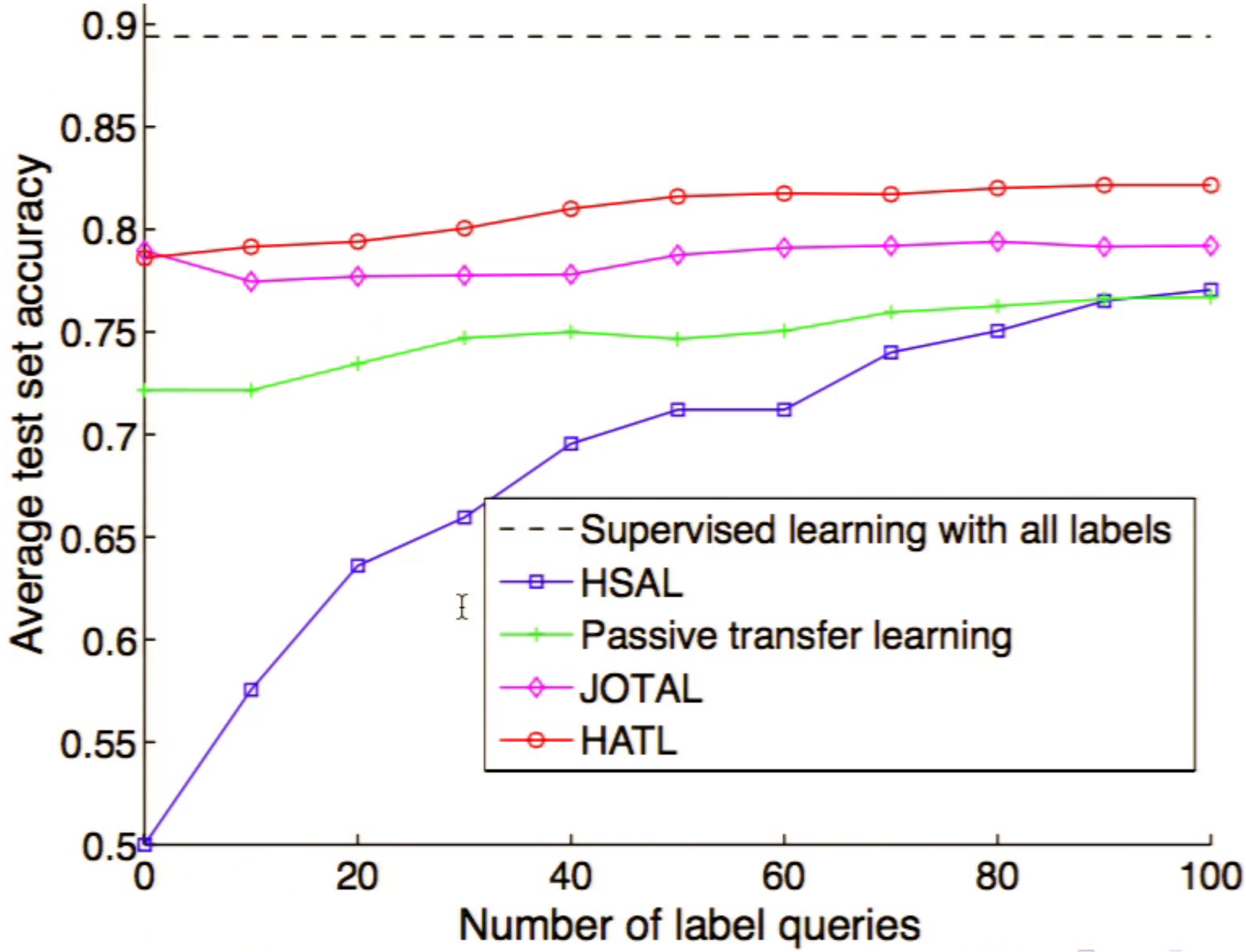
Measure SVM's classification performance on held out target data

Reminder: HATL, HSAL relabel ALL *training data.*

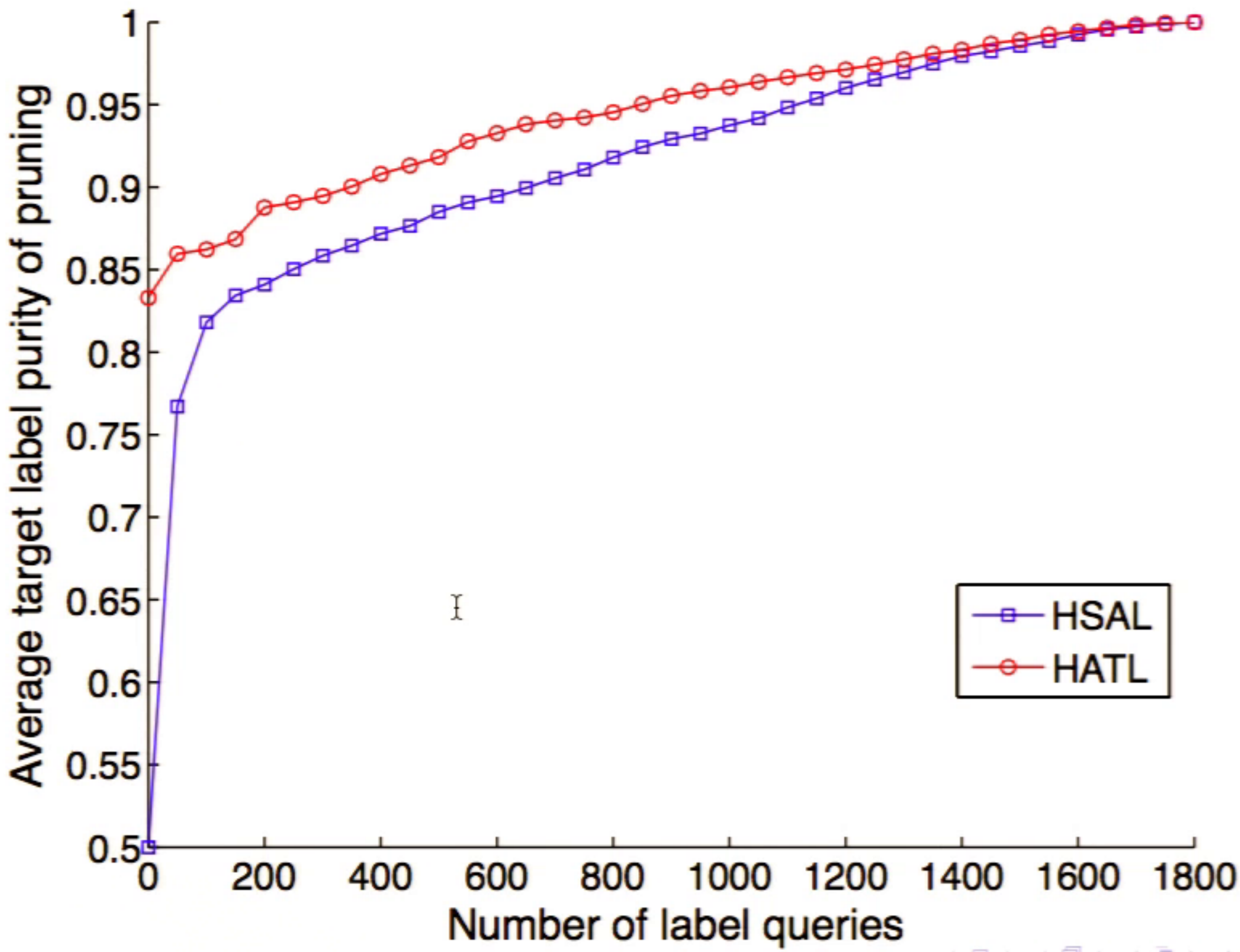
dvd→kitchen: SVM test set accuracy



dvd→kitchen: test set accuracy (over first 100 queries)



Effective SSL: dvd→kitchen imputation error



HATL: conceptually simple, agnostic, effective

+ Intuitive strategy: exploits cluster structure in data
  (handles differences in $\mathcal{D}_{\mathrm{S}}$, $\mathcal{D}_{\mathrm{T}}$)
+ Leverages a theoretically sound AL algorithm (HSAL)
  (inherits good properties, e.g., no bias)
+ Simple, agnostic TL: no commitment to single TL framework
  (we use one in analysis but could use others)
+ No hyperparameters to tune, etc.
+ Efficient, scalable (slowest part is initial clustering!)

Generic framework for TL, AL, SSL

Thank you!

Code: `http://www-bcf.usc.edu/~liu32/code.html`.
Extended: `http://www-scf.usc.edu/~dkale/publications.html`

Thank you to my collaborators:
Marjan Ghazvininejad, Anil Ramakrishna, Jingrui He, Yan Liu

Support: NSF, DARPA, IBM Faculty Award, Alfred E. Mann Institute.

Thank you and fight on!