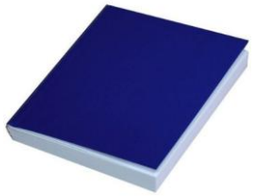
Unsupervised Domain Adaptation: From Practice to Theory

John Blitzer

Unsupervised Domain Adaptation



amazon.com

Running with Scissors

Title: Horrible book, horrible.

This book was horrible. I read half, suffering from a headache the entire time, and eventually i lit it on fire. 1 less copy in the world. Don't waste your money. I wish i had the time spent reading this book back. It wasted my life

Source

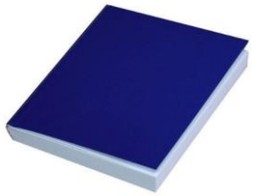
Avante Deep Fryer; Black

Title: lid does not work well...

I love the way the Tefal deep fryer cooks, however, I am returning my second one due to a defective lid closure. The lid may close initially, but after a few uses it no longer stays closed. I won't be buying this one again.

Target

Target-Specific Features



amazon.com

Running with Scissors

Title: Horrible book, horrible.

This book was horrible. I read half, suffering from a headache the entire time, and eventually i lit it on fire. 1 less copy in the world. Don't waste your money. I wish i had the time spent reading this book back. It wasted my life

Source

Avante Deep Fryer; Black

Title: lid **does not work** well...

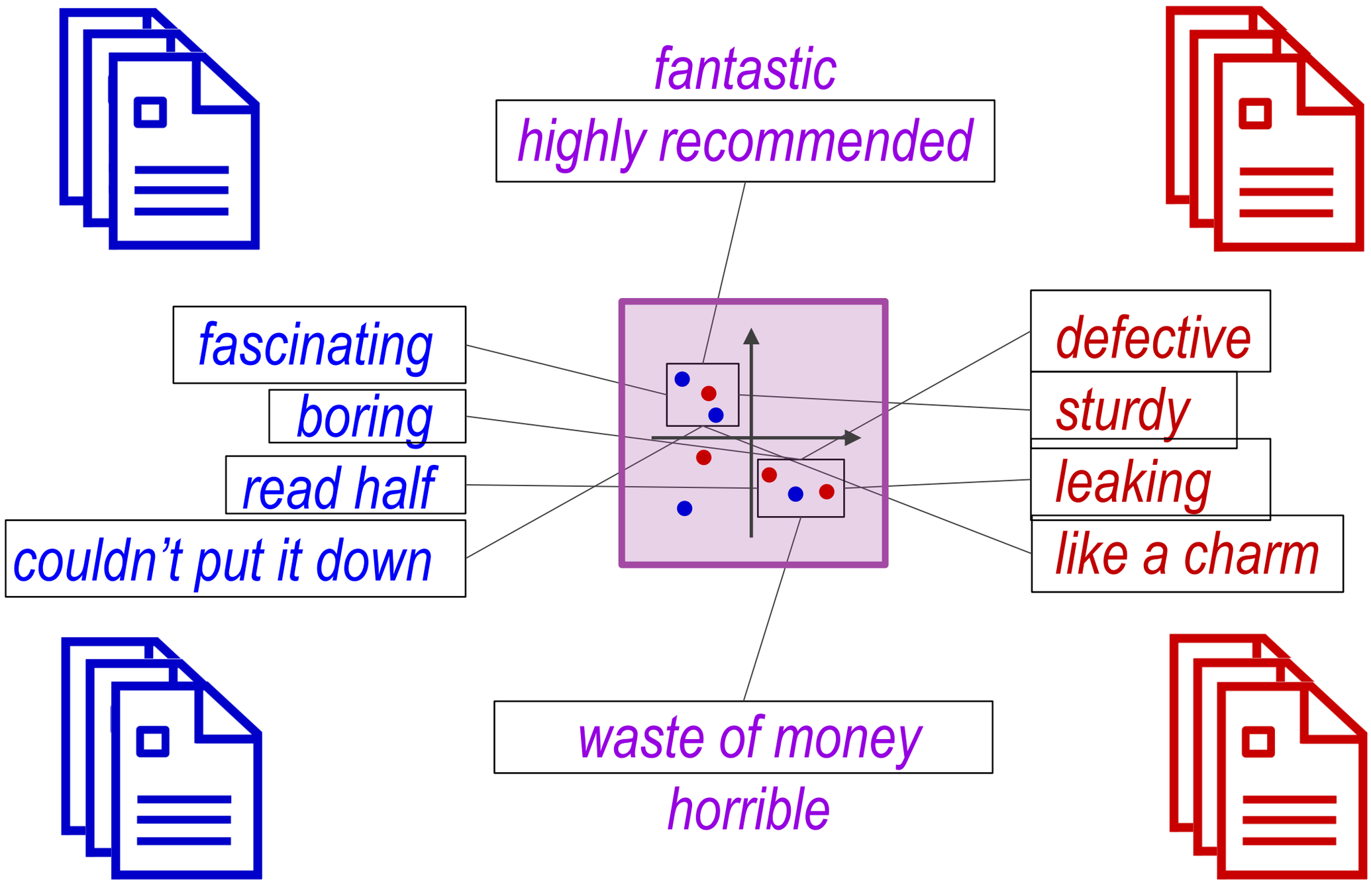
I love the way the Tefal deep fryer cooks, however, I am **returning** my second one due to a **defective** lid closure. The lid may close initially, but after a few uses it no longer stays closed. I won't be buying this one again.

Target

Learning Shared Representations

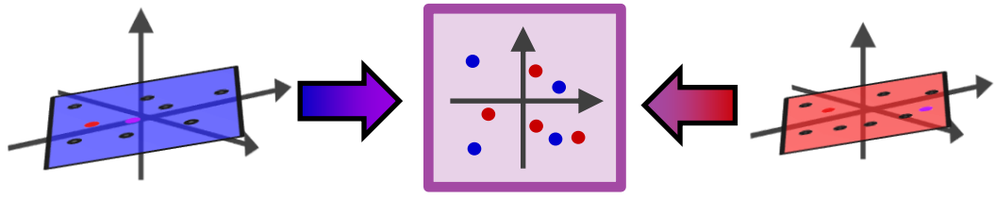
Source

Target



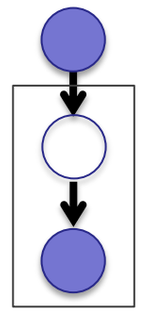
Shared Representations: A Quick Review

Blitzer et al. (2006, 2007). Shared CCA.



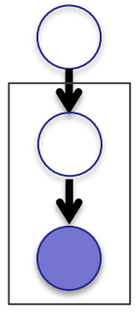
Tasks: Part of speech tagging, sentiment.

Xue et al. (2008). Probabilistic LSA



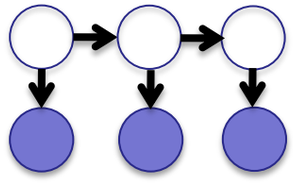
Task: Cross-lingual document classification.

Guo et al. (2009). Latent Dirichlet Allocation



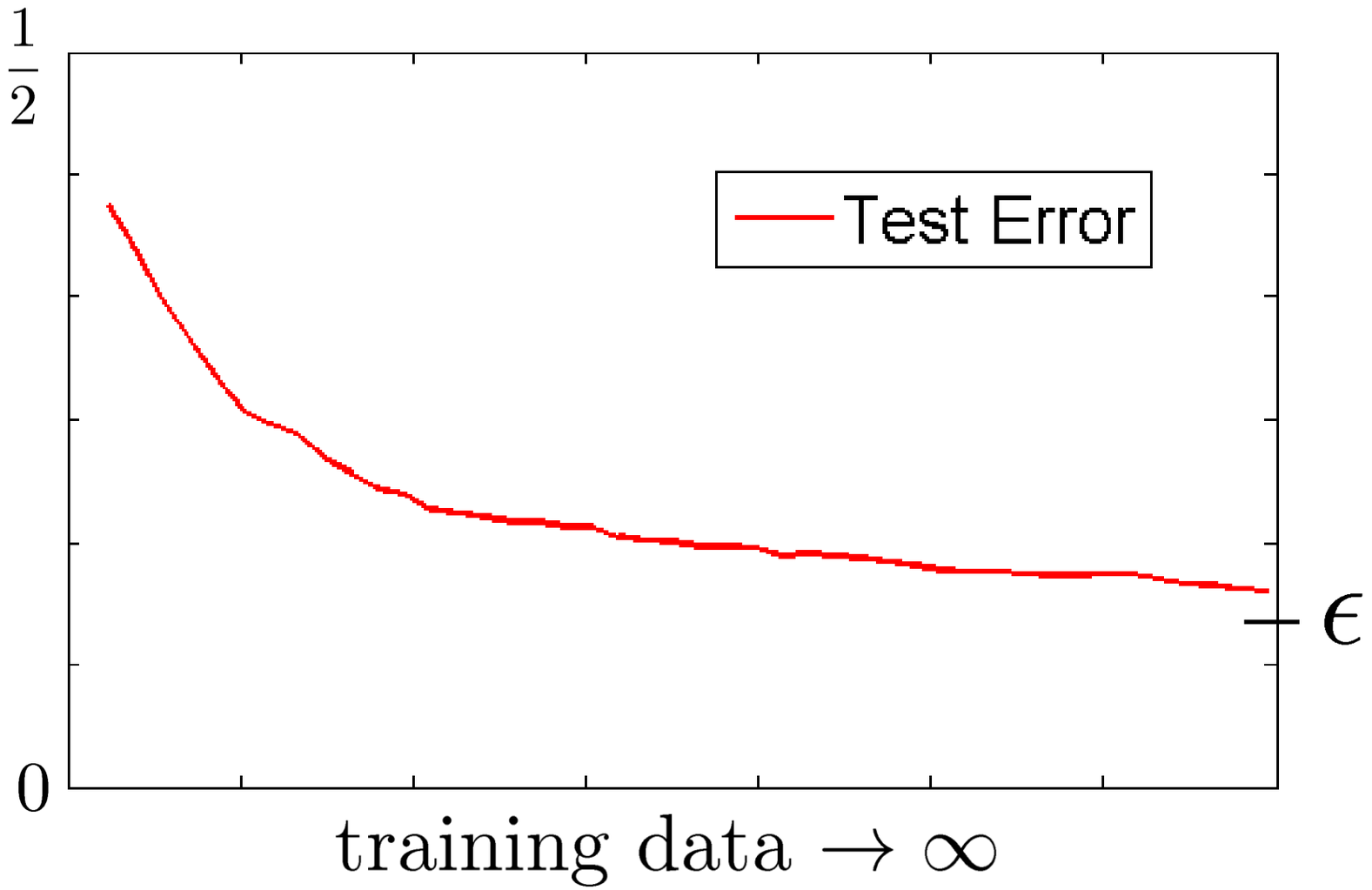
Task: Named entity recognition

Huang et al. (2009). Hidden Markov Models

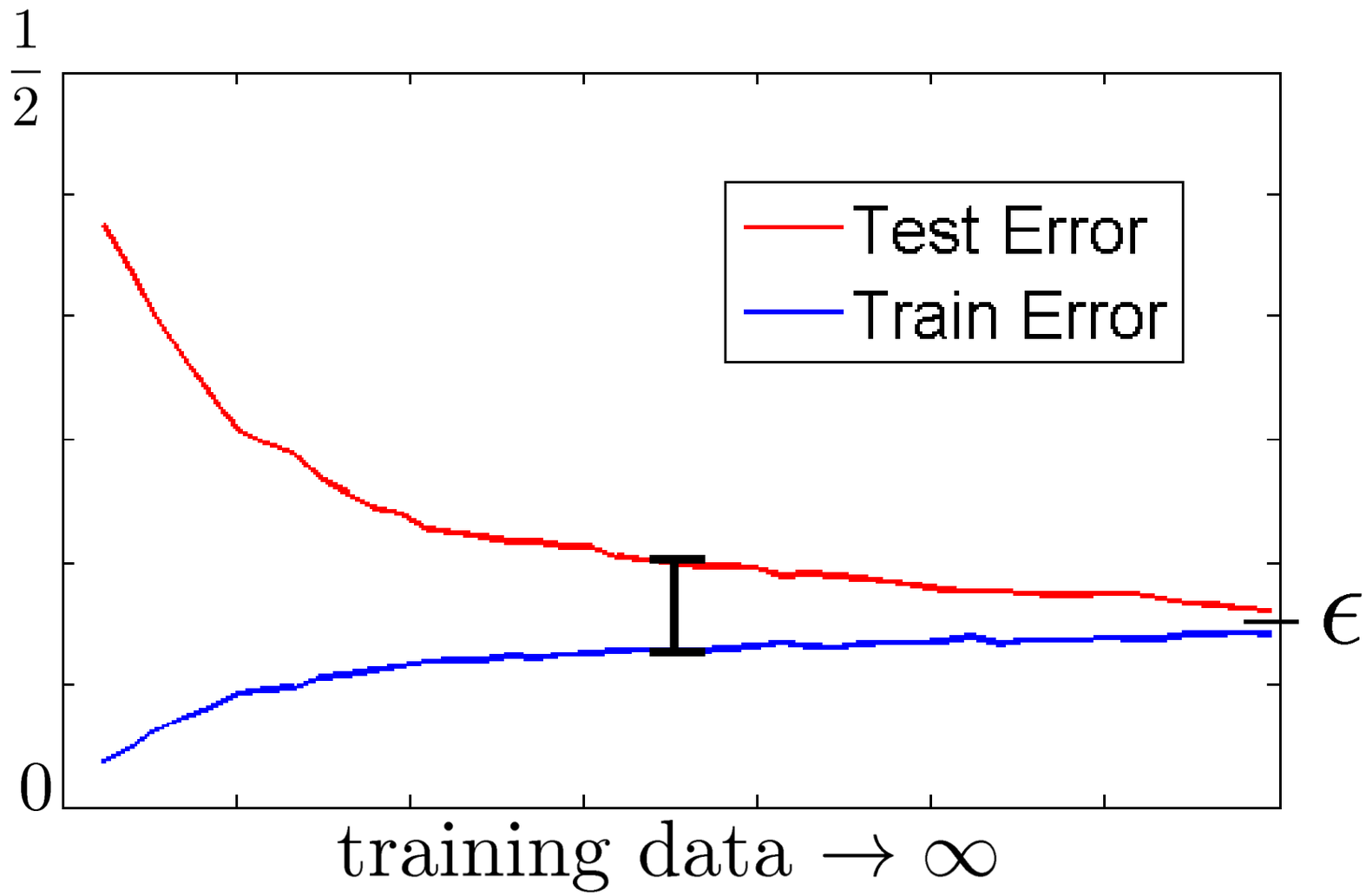


Task: Part of Speech Tagging

What do you mean, theory?

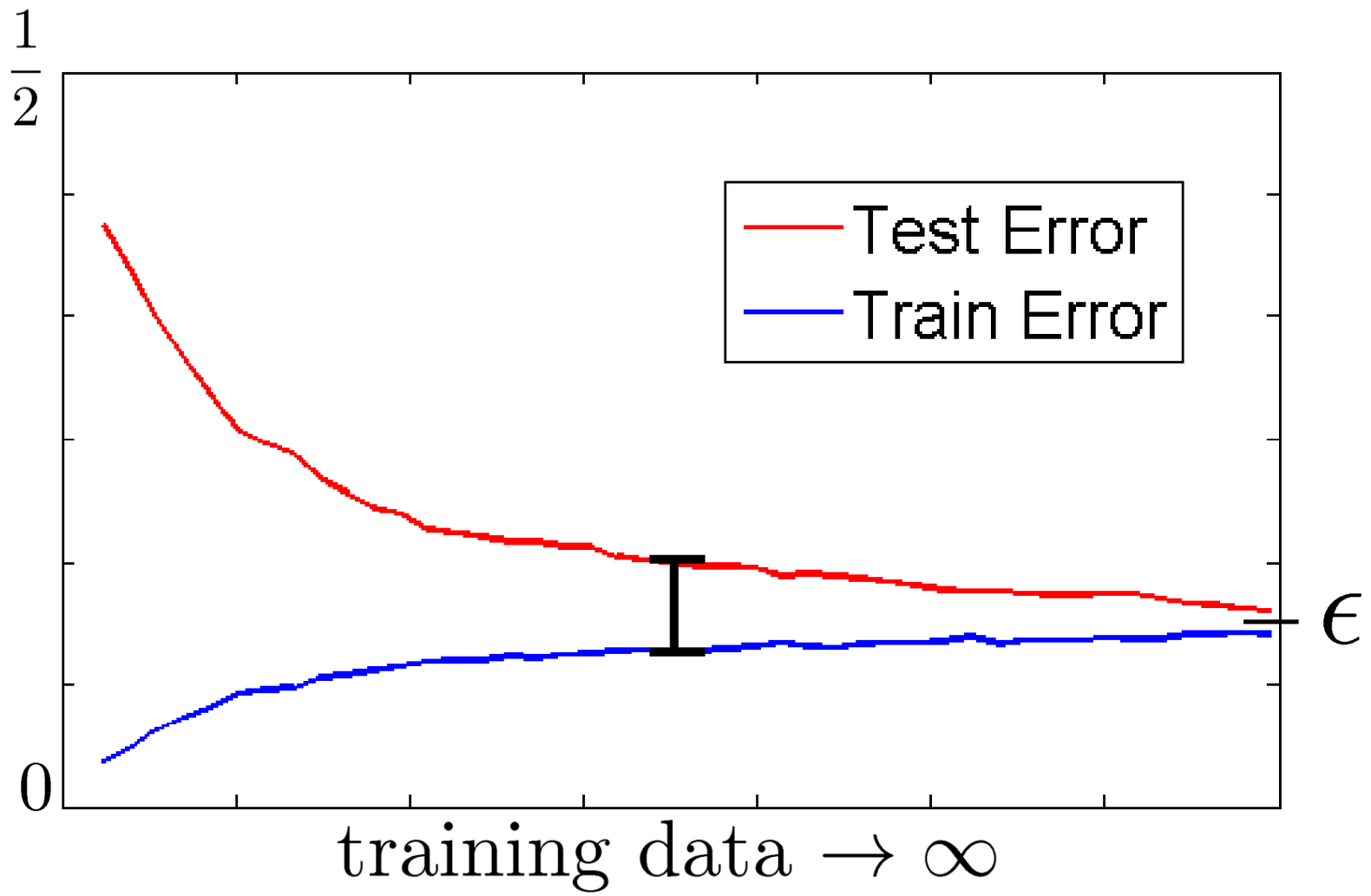


What do you mean, theory?



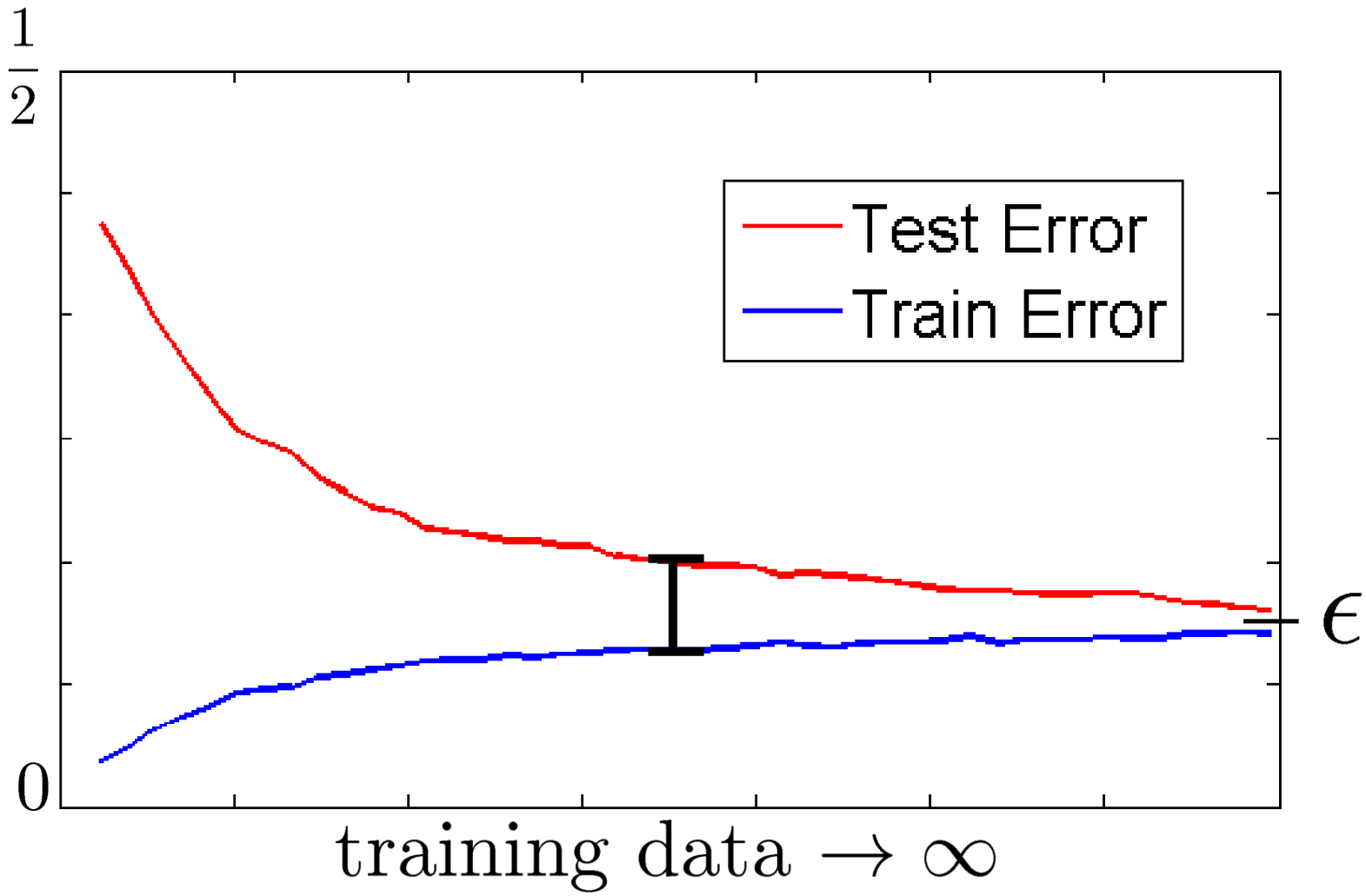
Statistical Learning Theory: $\epsilon_{\text{test}} \leq \hat{\epsilon}_{\text{train}} + \sqrt{\frac{\text{complexity}}{n}}$

What do you mean, theory?



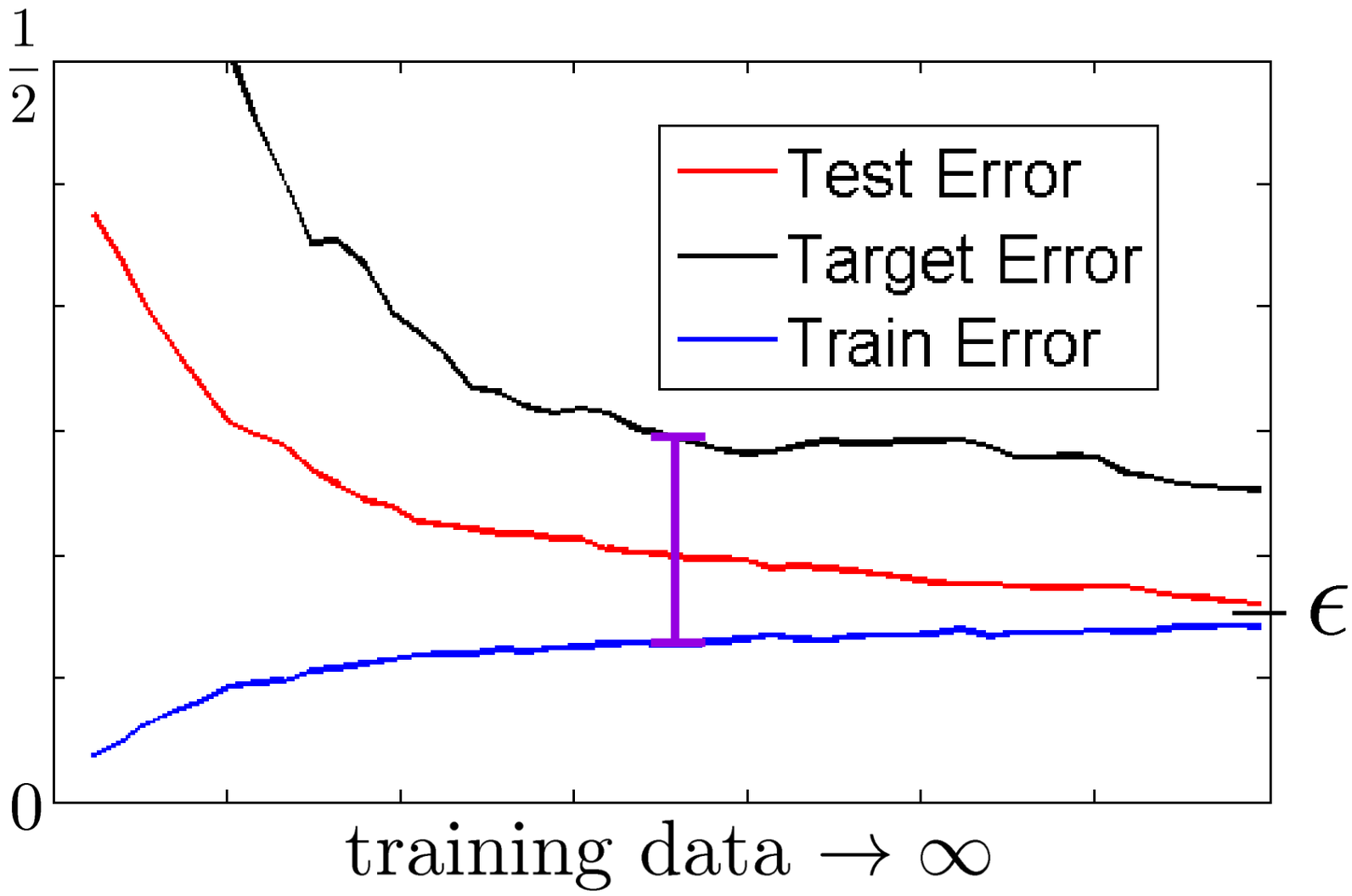
Statistical Learning Theory: $\epsilon_{\text{test}} \leq \hat{\epsilon}_{\text{train}} + \sqrt{\frac{\text{num features}}{n}}$

What do you mean, theory?

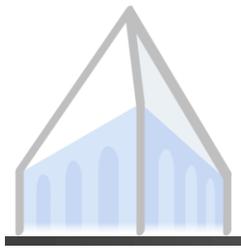


Classical Learning Theory: $\epsilon_{\text{test}} \leq \hat{\epsilon}_{\text{train}} + \sqrt{\frac{\text{num features}}{n}}$

What do you mean, theory?



Adaptation Learning Theory: $\epsilon_{\text{target}} \leq$??



Goals for Domain Adaptation Theory

1. A computable (source) sample bound on target error

2. A formal description of empirical phenomena
   - Why do shared representations algorithms work?

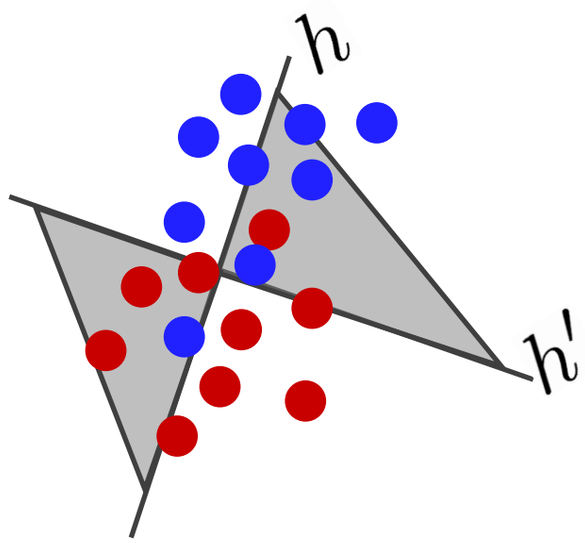
3. Suggestions for future research

Talk Outline

1. Target Generalization Bounds using Discrepancy Distance

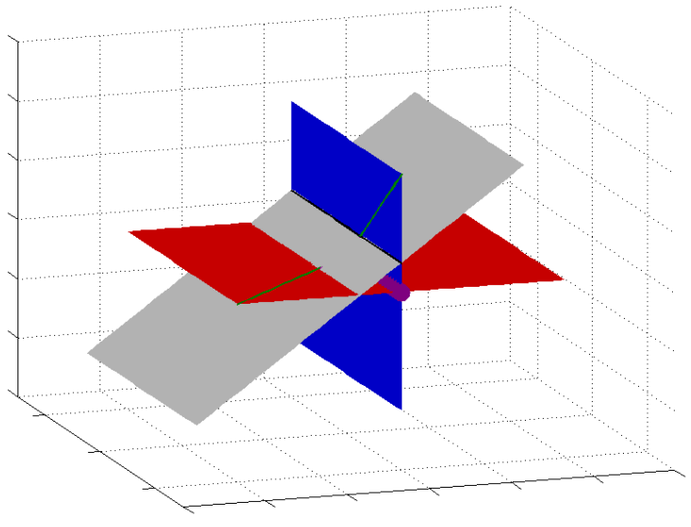
   [BBCKPW 2009]

   [Mansour et al. 2009]



2. Coupled Subspace Learning

   [BFK 2010]



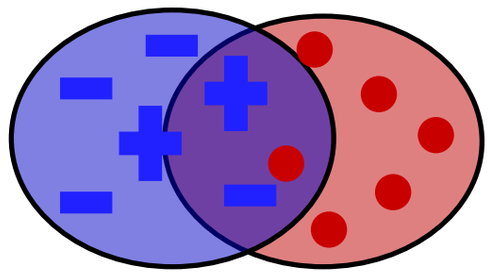
Formalizing Domain Adaptation

Source distribution

$$(x, y) \sim \Pr_S[x, y]$$

Target distribution

$$(x, y) \sim \Pr_T[x, y]$$



Source labeled data

$$x \sim \Pr_S[x]$$

$$y \sim \Pr_S[y|x]$$

Target **unlabeled** data

$$x \sim \Pr_T[x]$$

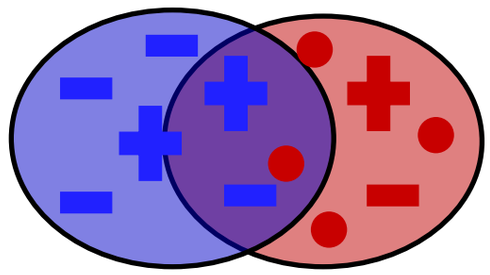
Formalizing Domain Adaptation

Source distribution

$$(x, y) \sim \Pr_S[x, y]$$

Target distribution

$$(x, y) \sim \Pr_T[x, y]$$



Source labeled data

$$x \sim \Pr_S[x]$$

$$y \sim \Pr_S[y|x]$$

Target **unlabeled** data

$$x \sim \Pr_T[x]$$

$$y \sim \Pr_T[y|x]$$

Semi-supervised adaptation

Some target labels

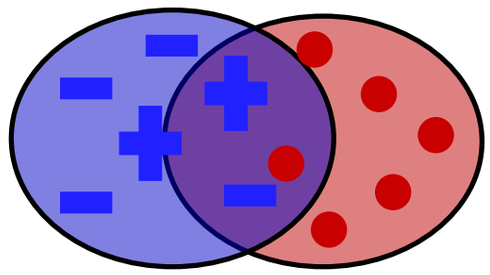
Formalizing Domain Adaptation

Source distribution

$$(x, y) \sim \Pr_S[x, y]$$

Target distribution

$$(x, y) \sim \Pr_T[x, y]$$



Source labeled data

$$x \sim \Pr_S[x]$$

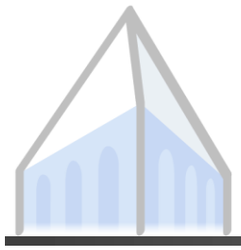
$$y \sim \Pr_S[y|x]$$

Target **unlabeled** data

$$x \sim \Pr_T[x]$$

Semi-supervised adaptation

Not in this talk



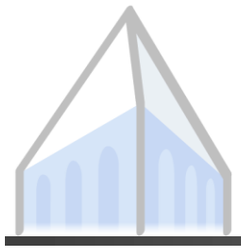
A Generalization Bound

S, T: Source and target $\quad \mathcal{H}$: Hypothesis class $\quad n$: Sample size

$\hat{S}$: Labeled S sample $\quad \hat{T}$: Unlabeled T sample $\quad h^*$: best $h \in \mathcal{H}$

With probability $1 - \delta$, for h the ERM of $\hat{S}$:

$$\epsilon_T(h) - \epsilon_T(h^*) \quad \leq$$



A new adaptation bound

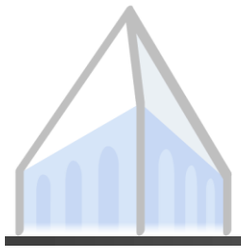
S, T: Source and target $\quad \mathcal{H}$: Hypothesis class $\quad n$: Sample size

$\hat{S}$: Labeled S sample $\quad \hat{T}$: Unlabeled T sample $\quad h^*$: best $h \in \mathcal{H}$

Bound from [MMR09]

With probability $1-\delta$, for h the ERM of $\hat{S}$:

$$\epsilon_T(h) - \epsilon_T(h^*) \leq \epsilon_{\hat{S}}(h, h^*) + \mathrm{disc}_{\mathcal{H}}(\hat{S}, \hat{T}) + \\ + O\left(\sqrt{\frac{\mathrm{complexity}(\mathcal{H}) + \log\frac{1}{\delta}}{n}}\right)$$



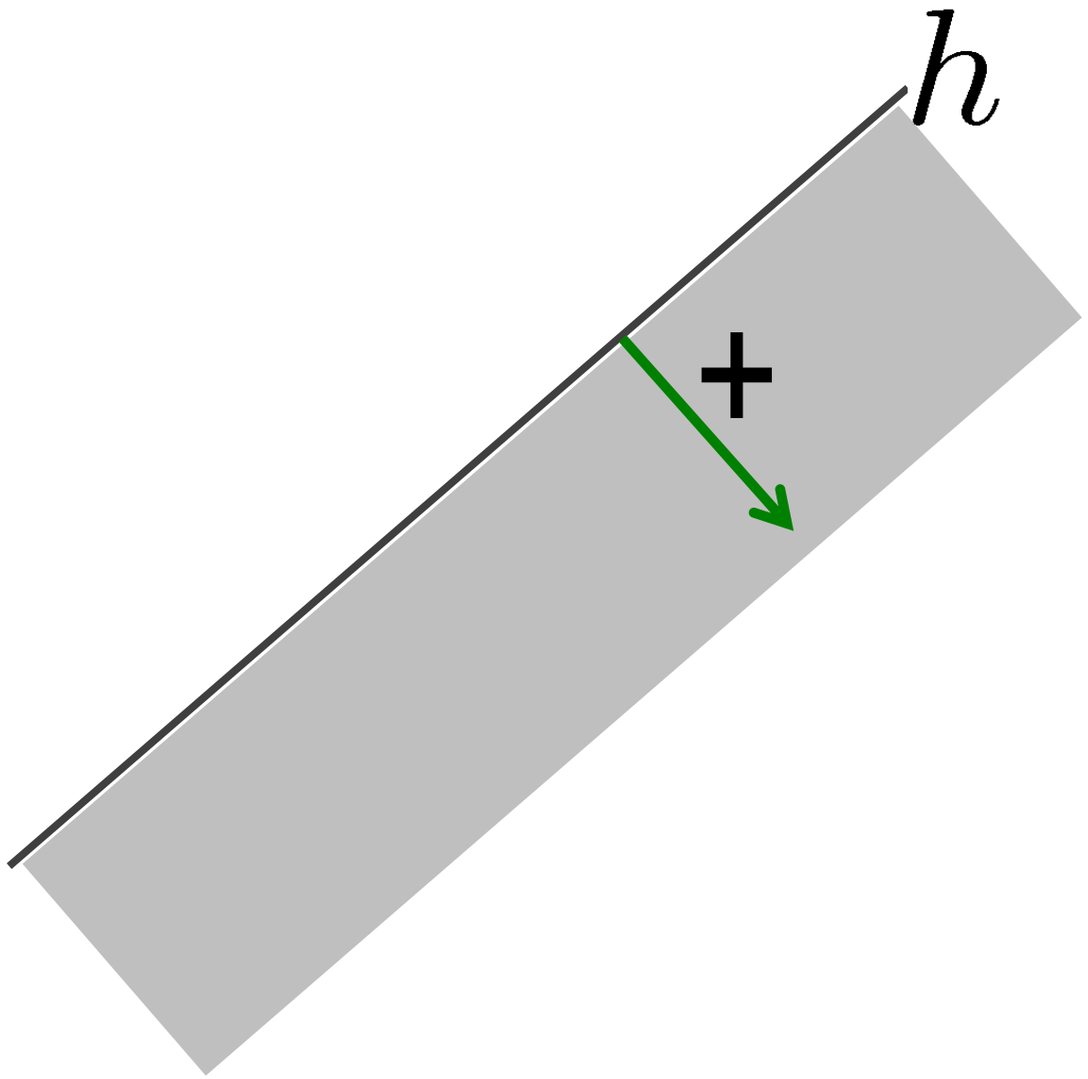
Discrepancy Distance

When good source models go bad

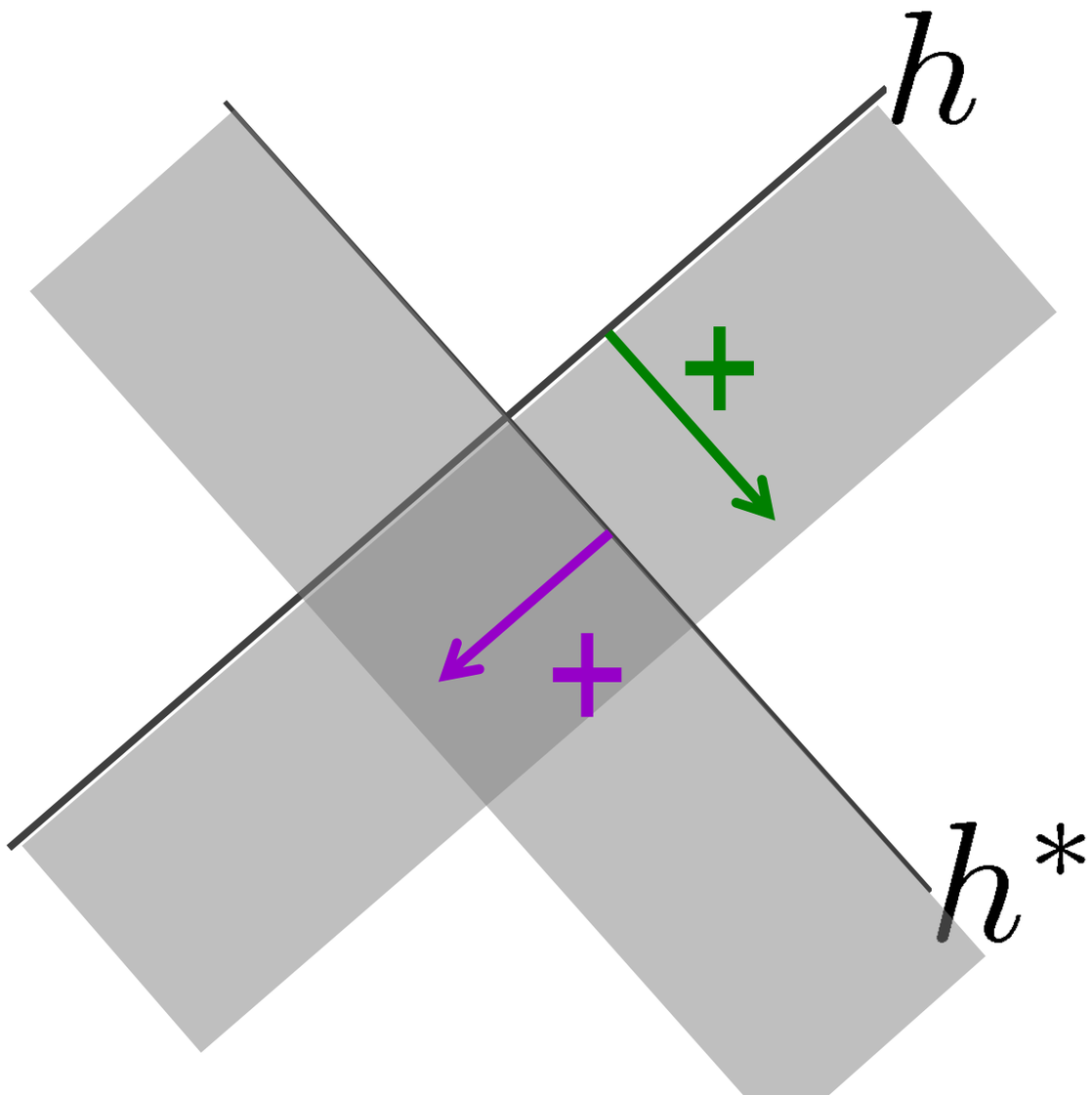
$$\mathrm{disc}_{\mathcal{H}}(S,T) =$$

$$\max_{h,h^* \in \mathcal{H}} |E_S[h(x) \neq h^*(x)] - E_T[h(x) \neq h^*(x)]|$$

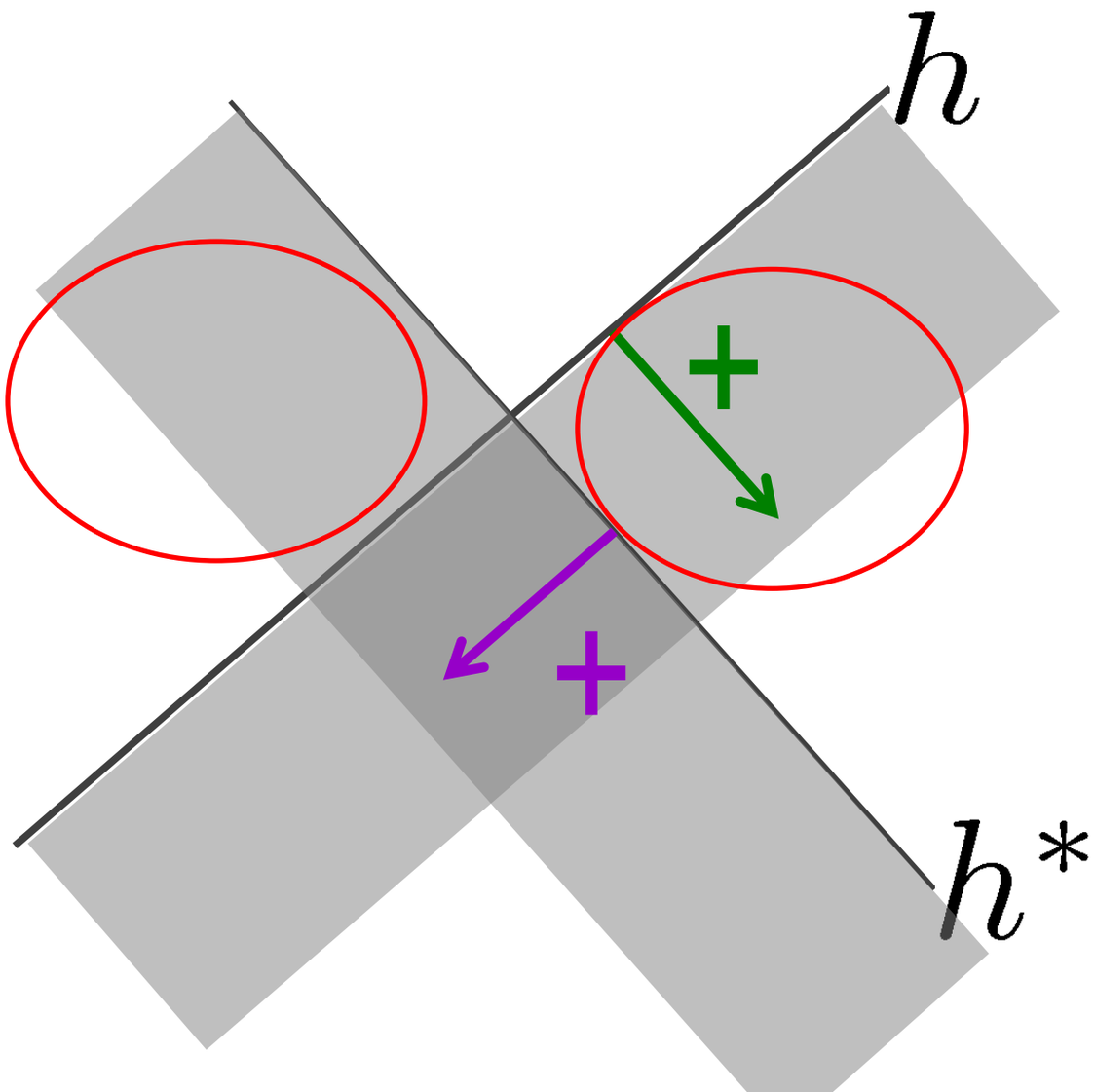
Binary Hypothesis Error Regions



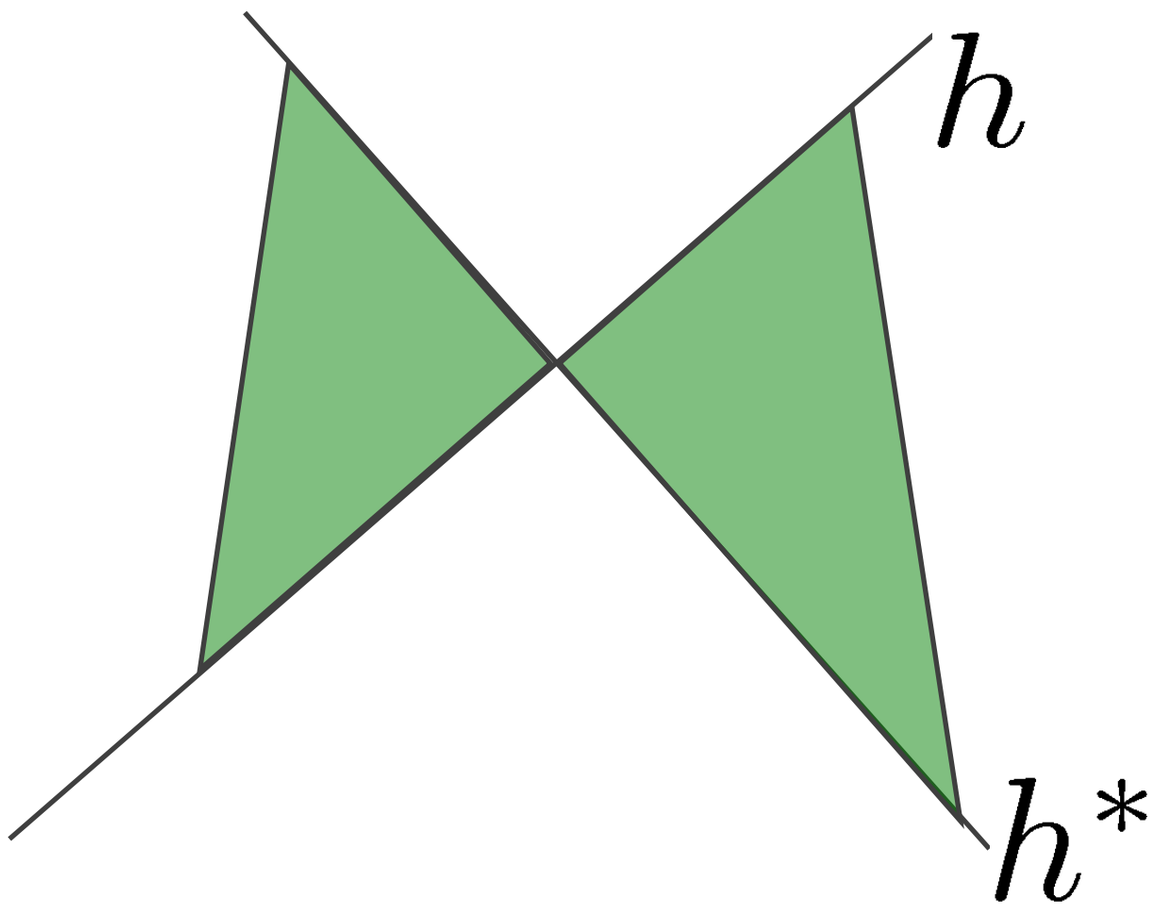
Binary Hypothesis Error Regions



Binary Hypothesis Error Regions



Binary Hypothesis Error Regions

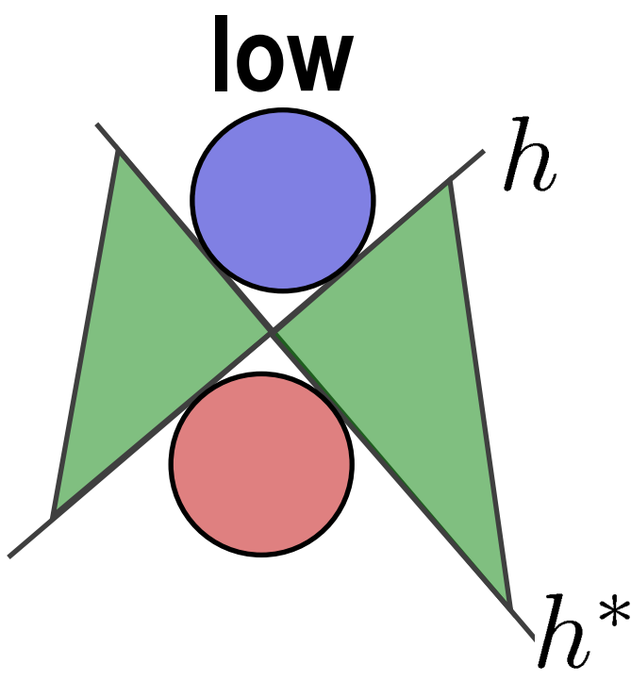


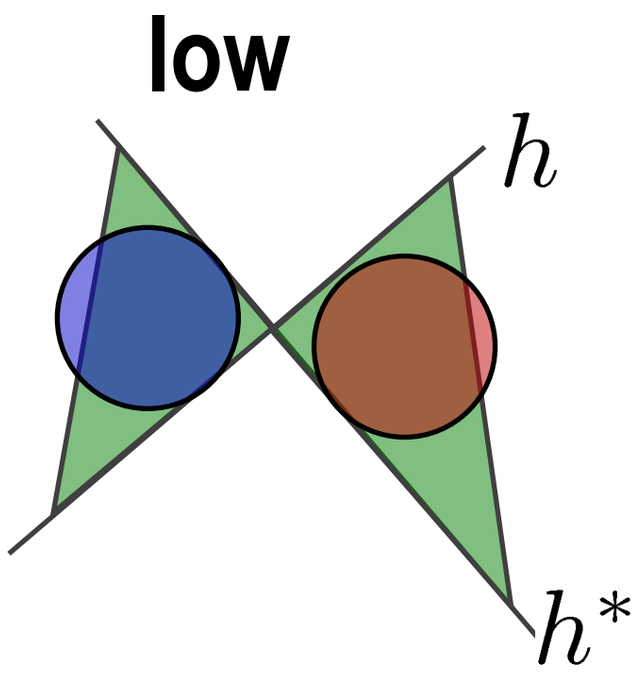
Discrepancy Distance

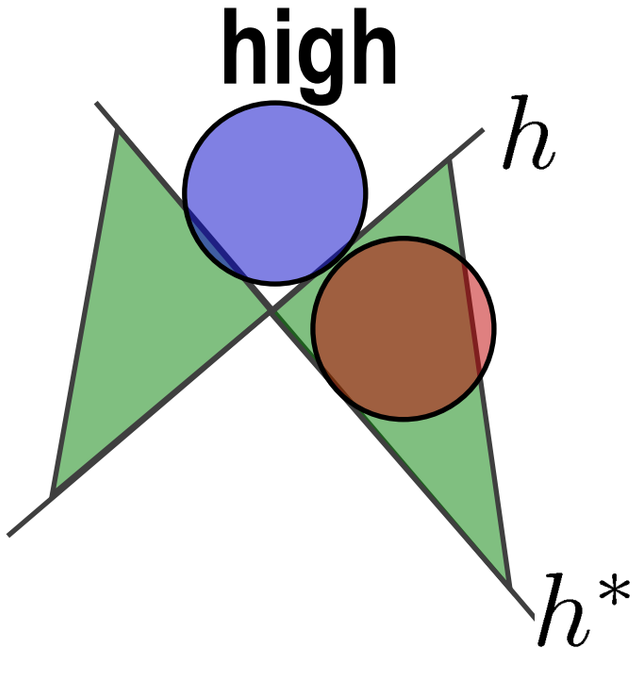
When good source models go bad

$$\mathrm{disc}_{\mathcal{H}}(S,T) =$$

$$\max_{h,h^* \in \mathcal{H}} |E_S[h(x) \neq h^*(x)] - E_T[h(x) \neq h^*(x)]|$$

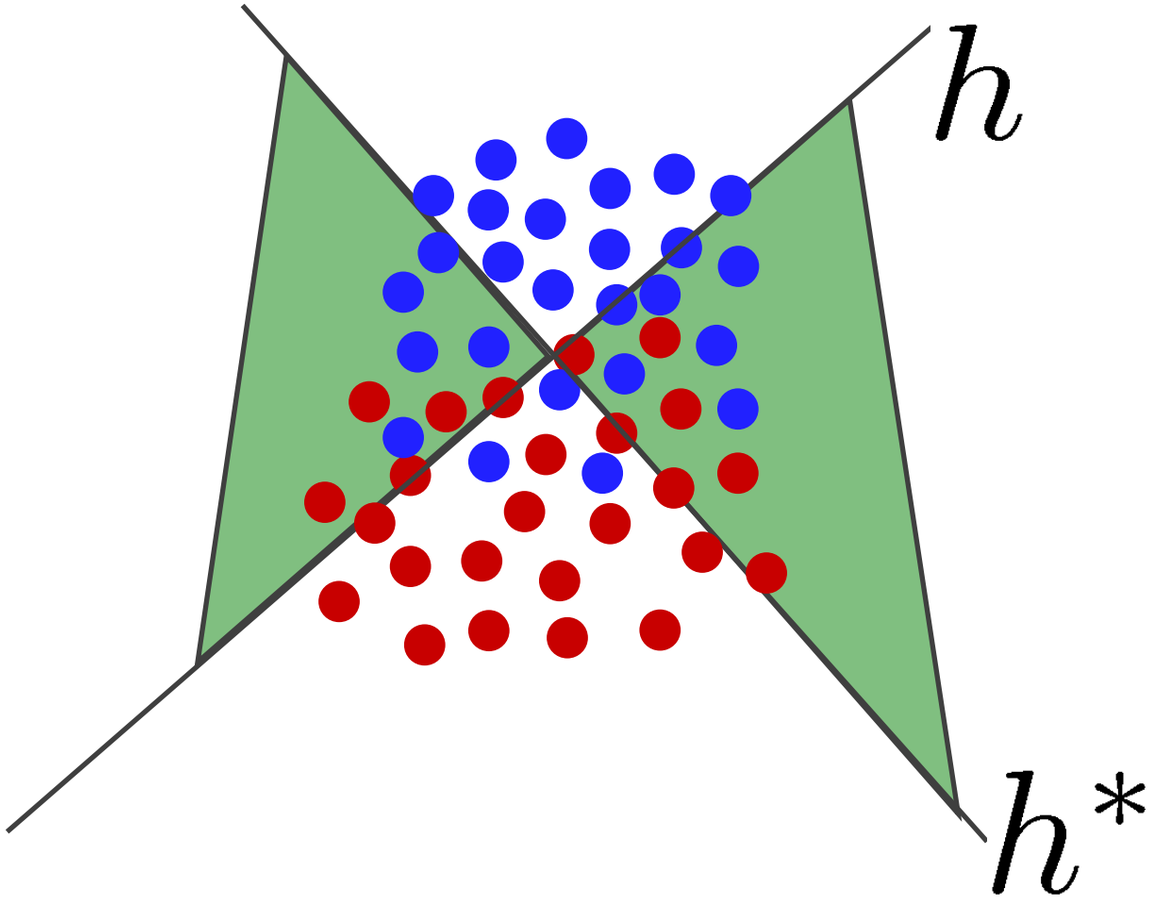






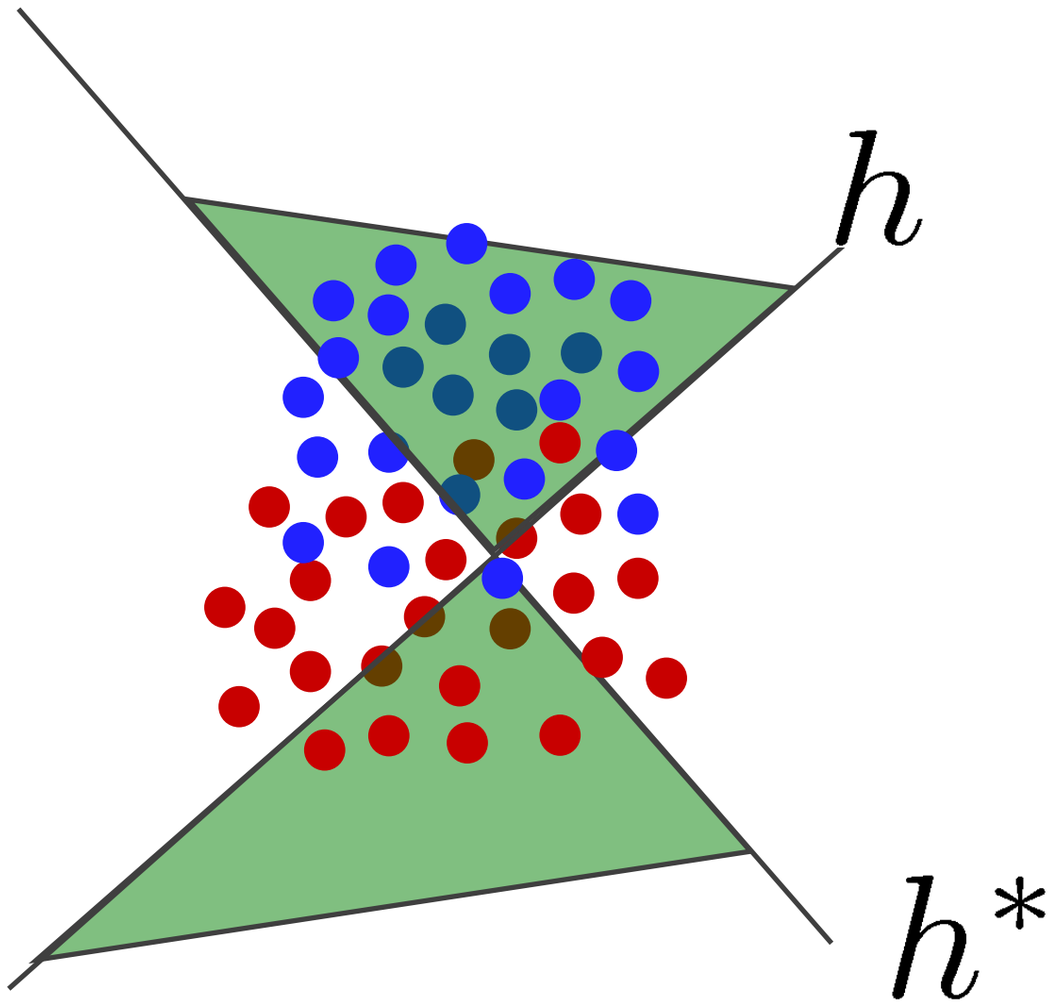
Computing Discrepancy Distance

Learn pairs of hypotheses to discriminate source from target



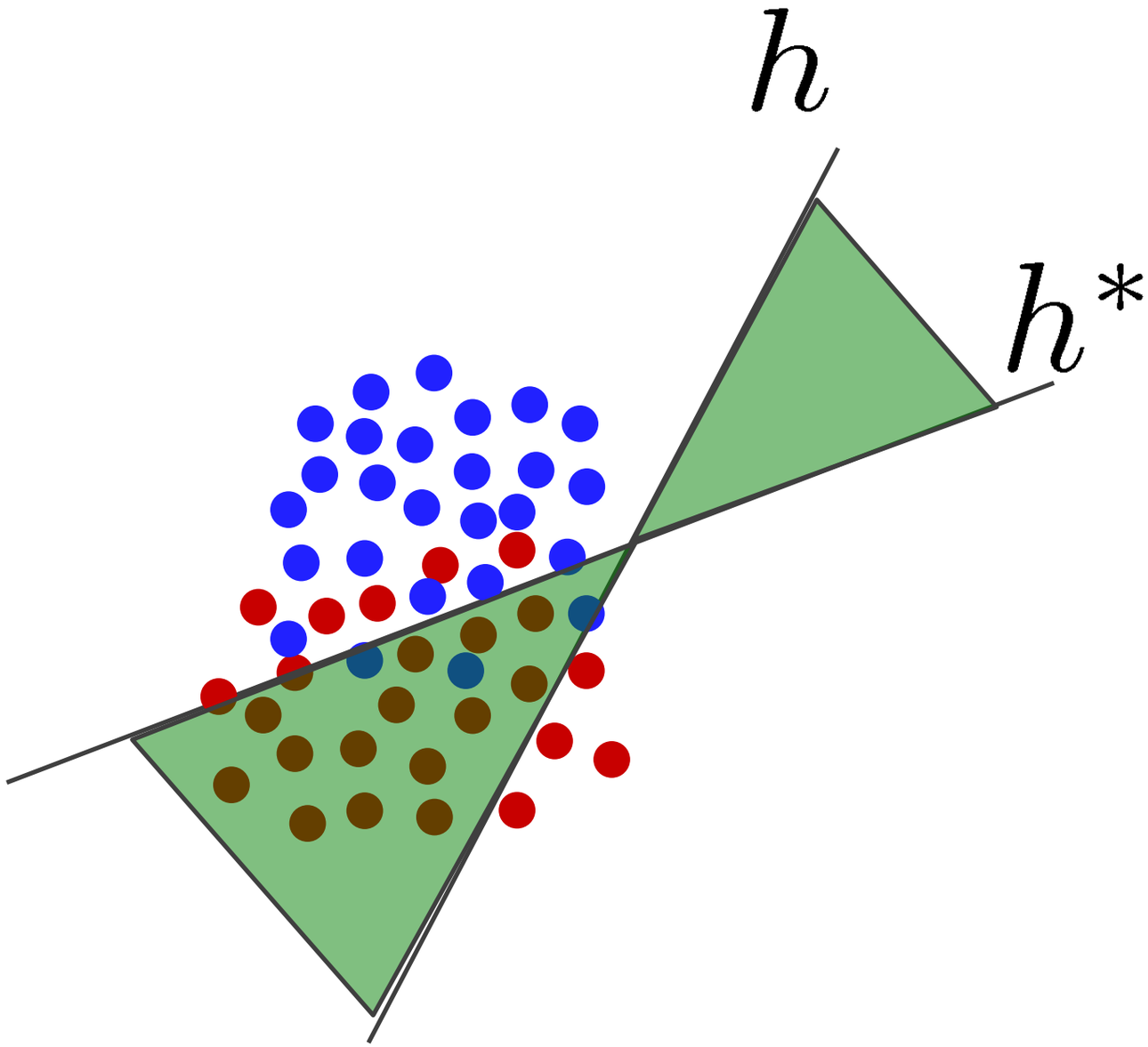
Computing Discrepancy Distance

Learn pairs of hypotheses to discriminate source from target



Computing Discrepancy Distance

Learn pairs of hypotheses to discriminate source from target



Hypothesis Classes & Representations

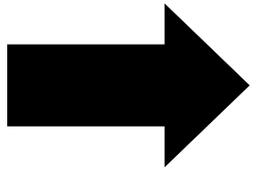
Linear Hypothesis Class: $h(x) = \text{sgn}\,(\beta \cdot x)$

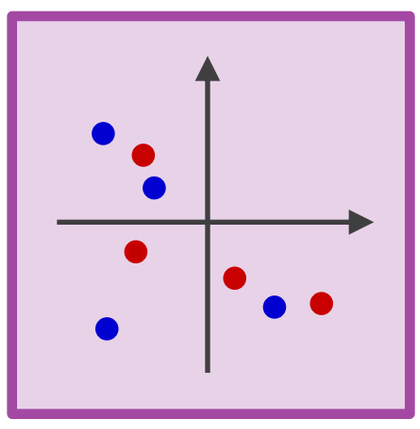
Induced classes from projections $\beta \cdot \Pi x$ $\Pi = \Pi\Pi$

x

Π

3
0
⋮
1
0
0
1





Goals for Π

1) $\text{Minimize } \text{disc}_{\Pi}(S, T)$

2) $\epsilon_{\Pi,T}(\beta^*) - \epsilon_T(\beta^*)$ small

A Proxy for the Best Model

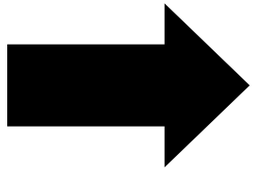
Linear Hypothesis Class: $h(x) = \text{sgn}\,(\beta \cdot x)$

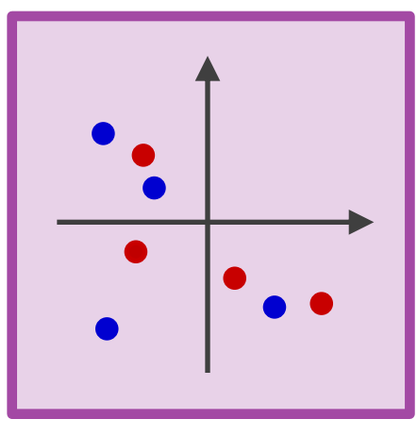
Induced classes from projections $\beta \cdot \Pi x$ $\Pi = \Pi\Pi$

x

Π

3
0
⋮
1
0
0
1





Goals for Π

1) Minimize $\text{disc}_\Pi(S, T)$

2) $\hat{\epsilon}_{S,\Pi}(\hat{\beta})$ small

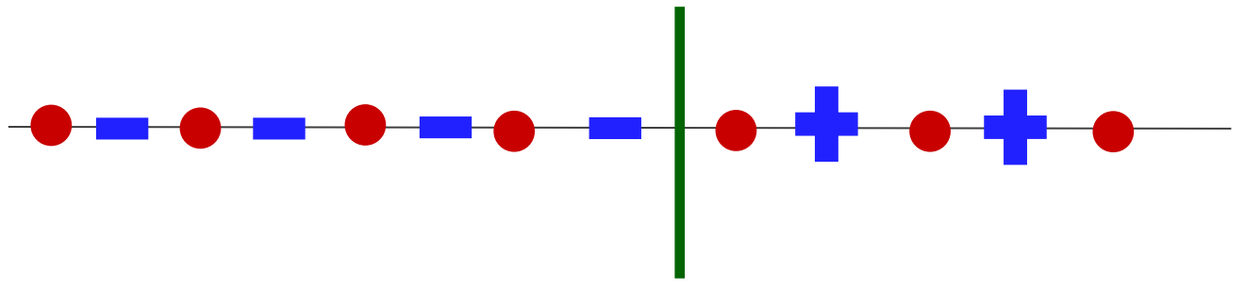
Problems with the Proxy

Π

x

3
0
⋮
1
0
0

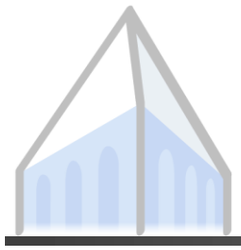
$$\Pi = \hat{\beta}\hat{\beta}^\top \quad \hat{\beta} = \text{source ERM}$$



1) Minimize $\text{disc}_\Pi(S, T)$
2) $\hat{\epsilon}_{S,\Pi}(\hat{\beta})$ small

Π ignores target-unique features!

~~$\epsilon_{\Pi,T}(\beta^*) - \epsilon_T(\beta^*)$ small~~



Goals

1. **A computable bound**

2. **Description of shared representations**

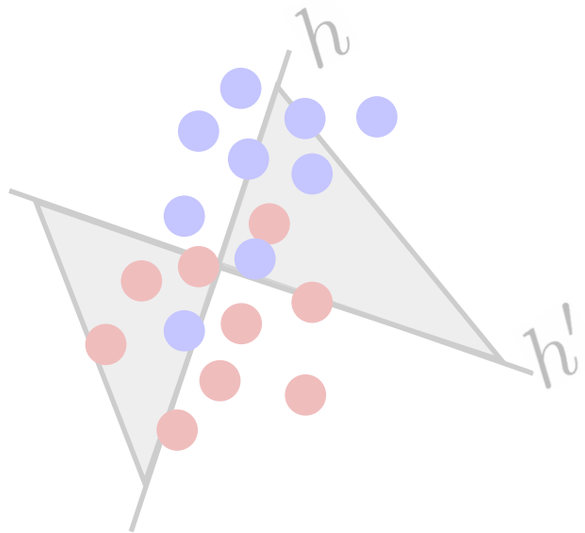
3. **Suggestions for future research**

Talk Outline

1. Target Generalization Bounds using Discrepancy Distance

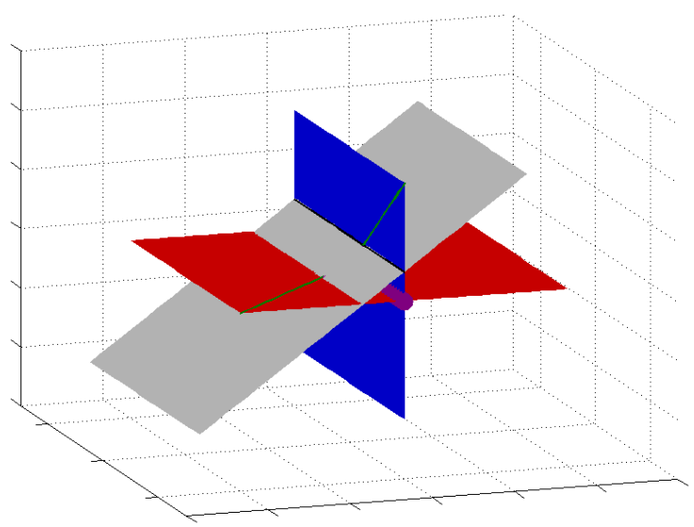
   [BBCKPW 2009]

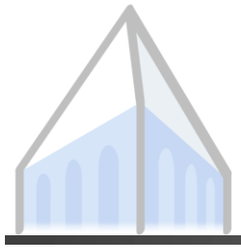
   [Mansour et al. 2009]



2. Coupled Subspace Learning

   [BFK 2010]





Assumption: Single Linear Predictor

$$\text{Assumption 1: } \mathbb{E}_S[Y|x] = \mathbb{E}_T[Y|x] = \beta \cdot x$$

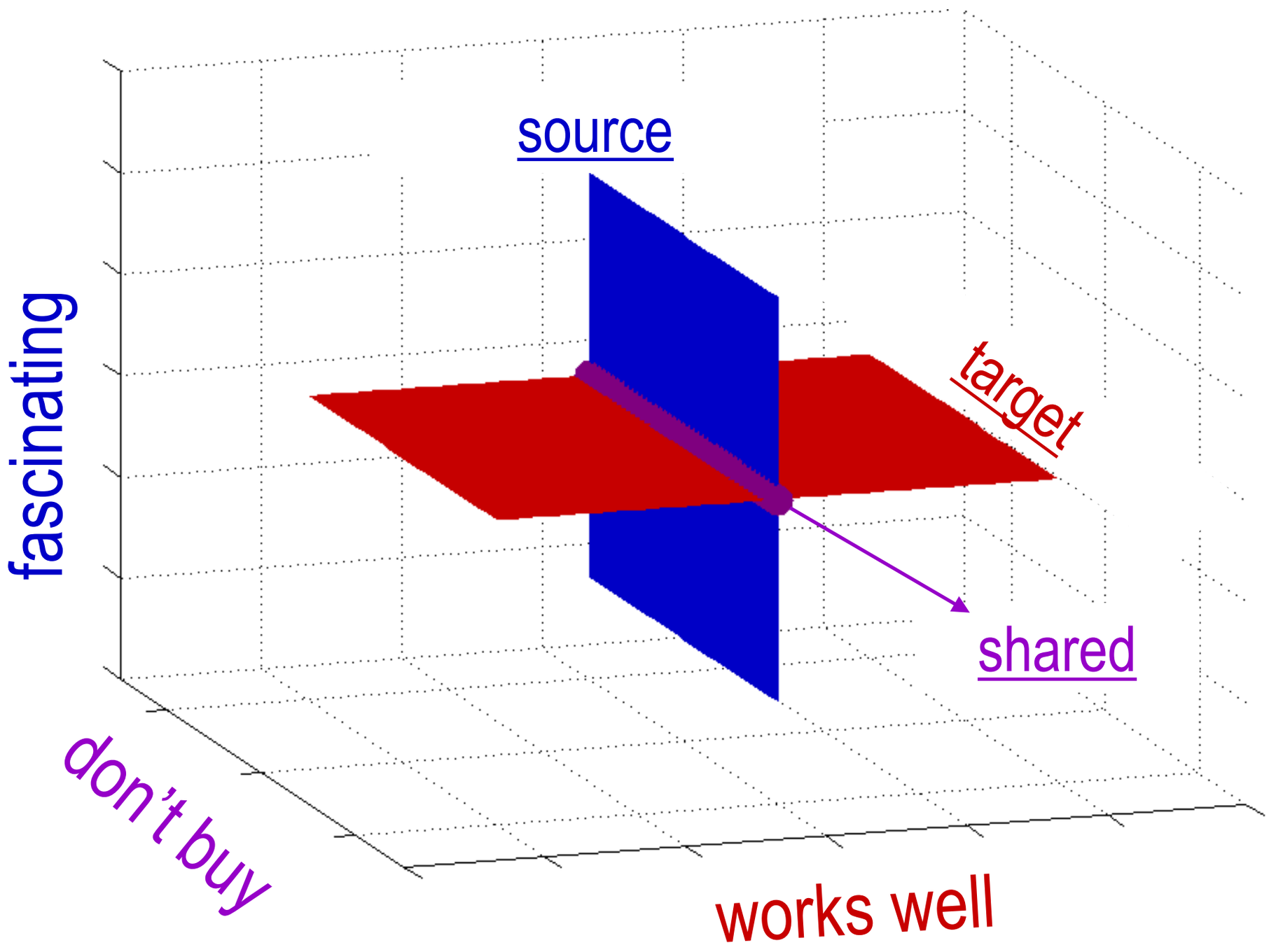
$\beta \cdot x$ can be decomposed as

source-specific | shared | target-specific

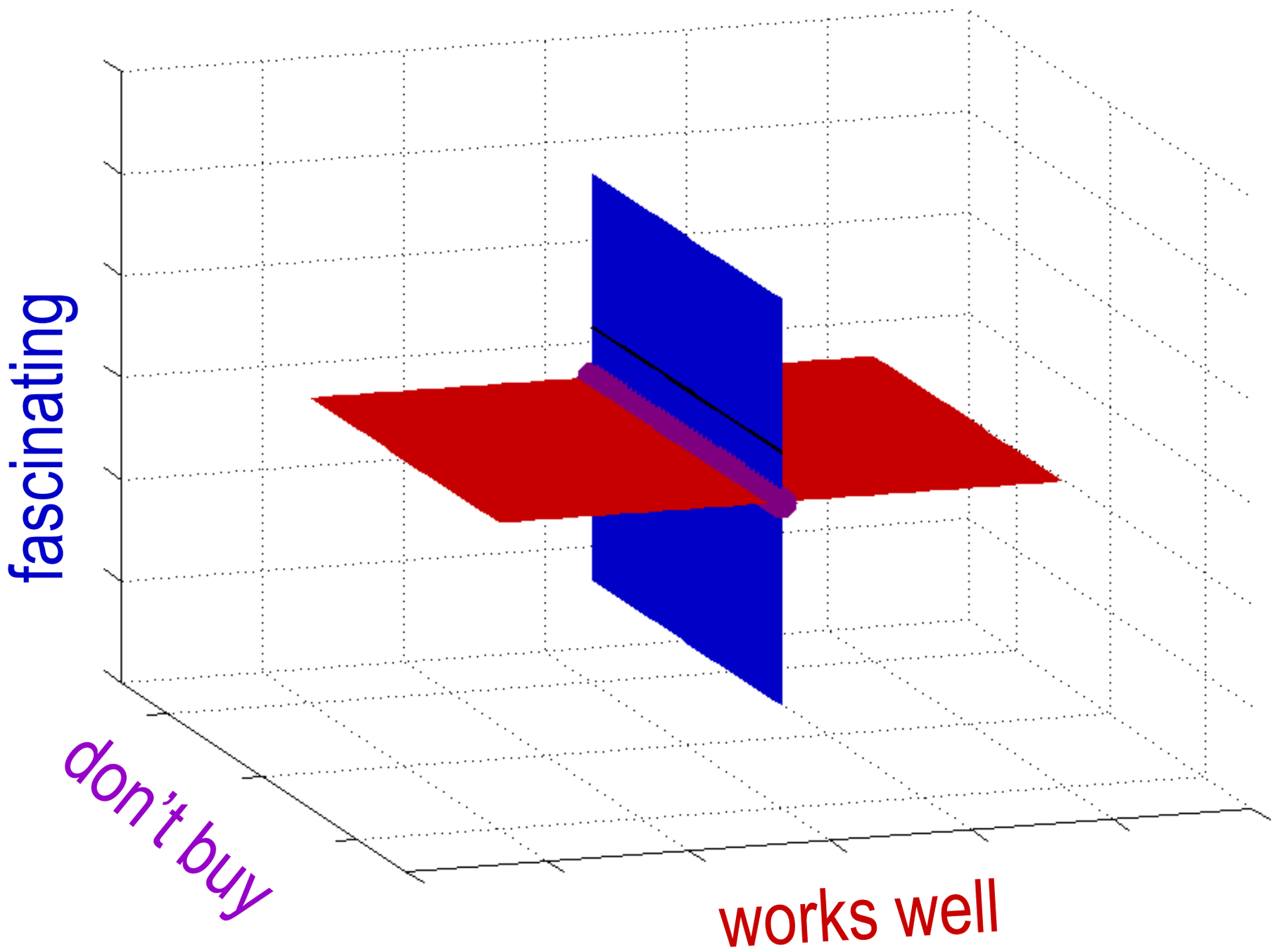
$$[\beta]_{S,\perp} \cdot [x]_{S,\perp} + [\beta]_{S,T} \cdot [x]_{S,T} + [\beta]_{T,\perp} \cdot [x]_{T,\perp}$$

target-specific can't be estimated from source alone . . . yet

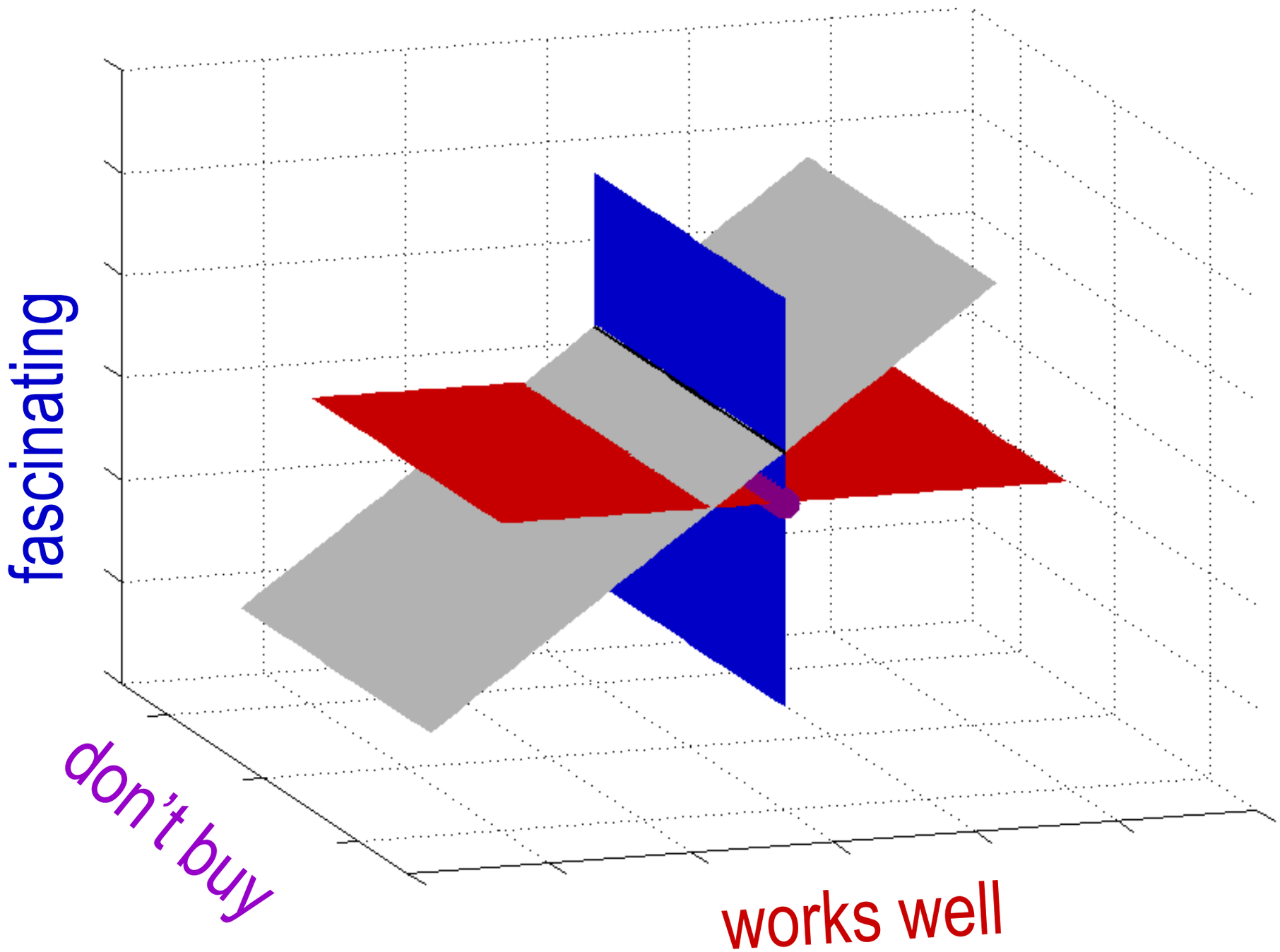
Visualizing Single Linear Predictor



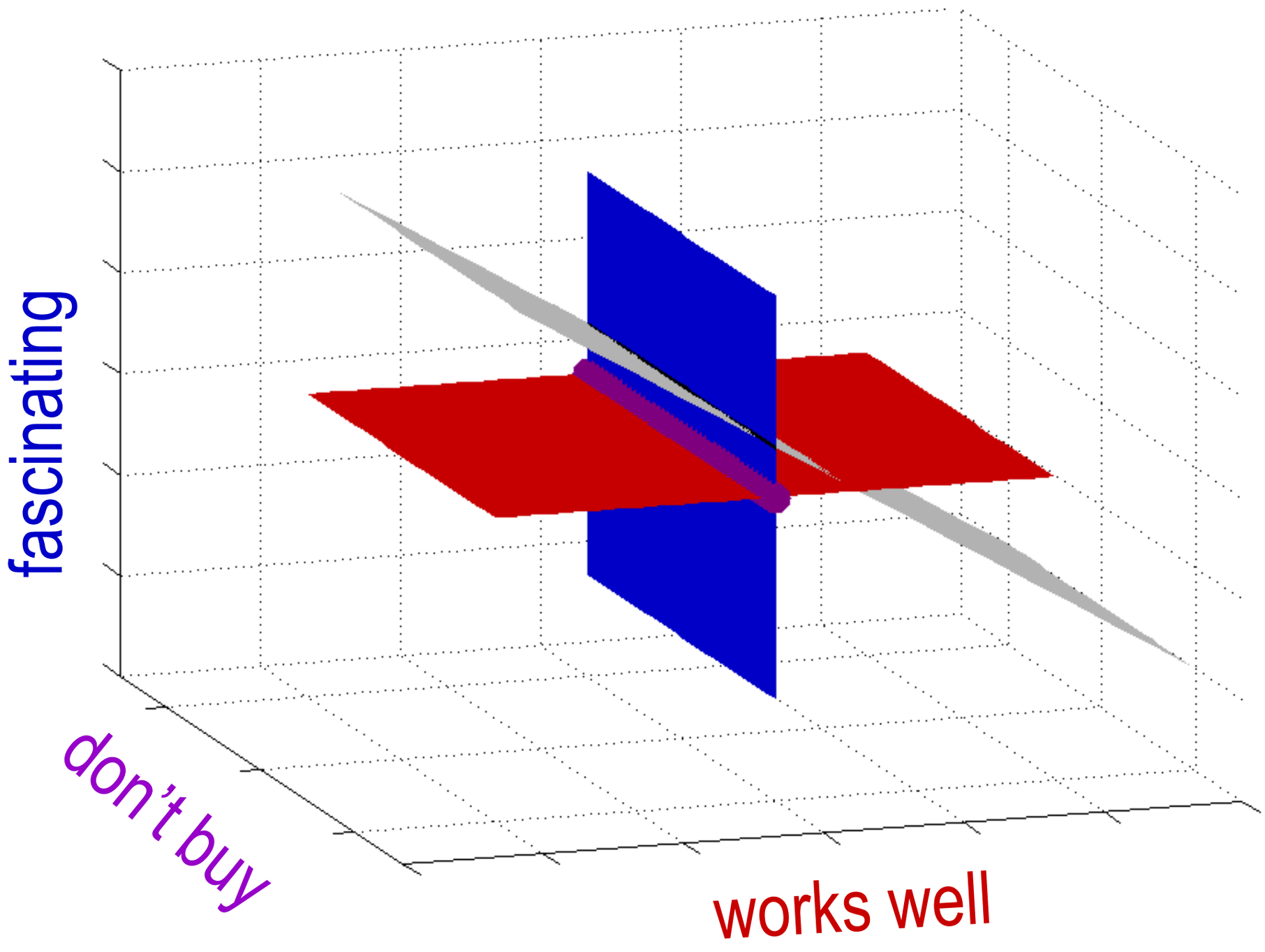
Visualizing Single Linear Predictor



Visualizing Single Linear Predictor



Visualizing Single Linear Predictor



Dimensionality Reduction Assumption

Assumption 2: $\mathbb{E}_S[Y|x] = \beta_S \cdot (\Pi_S x)$

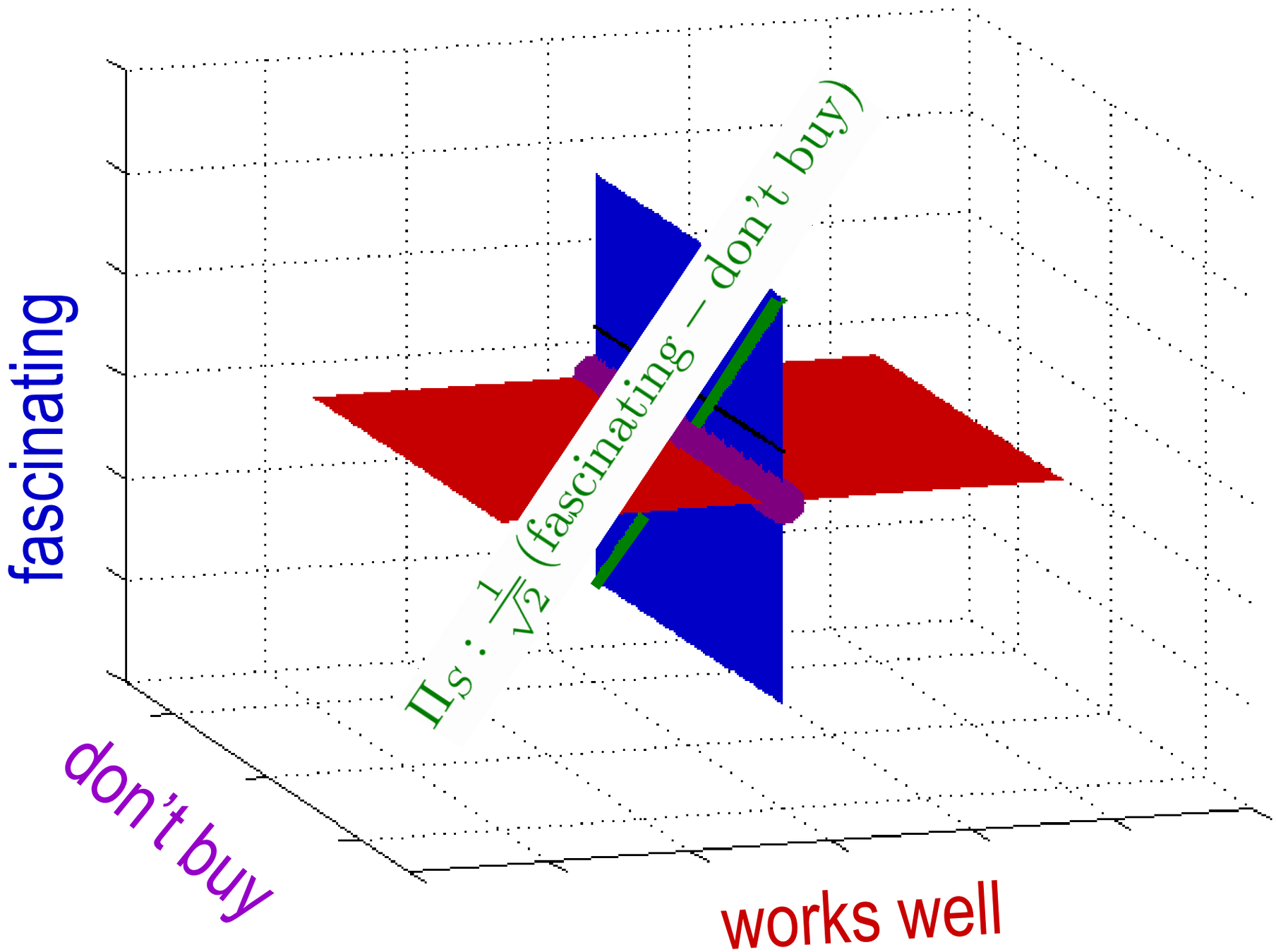
$\mathbb{E}_T[Y|x] = \beta_T \cdot (\Pi_T x)$

Projections $\Pi_S = \Pi_S \Pi_S$ $\Pi_T = \Pi_T \Pi_T$

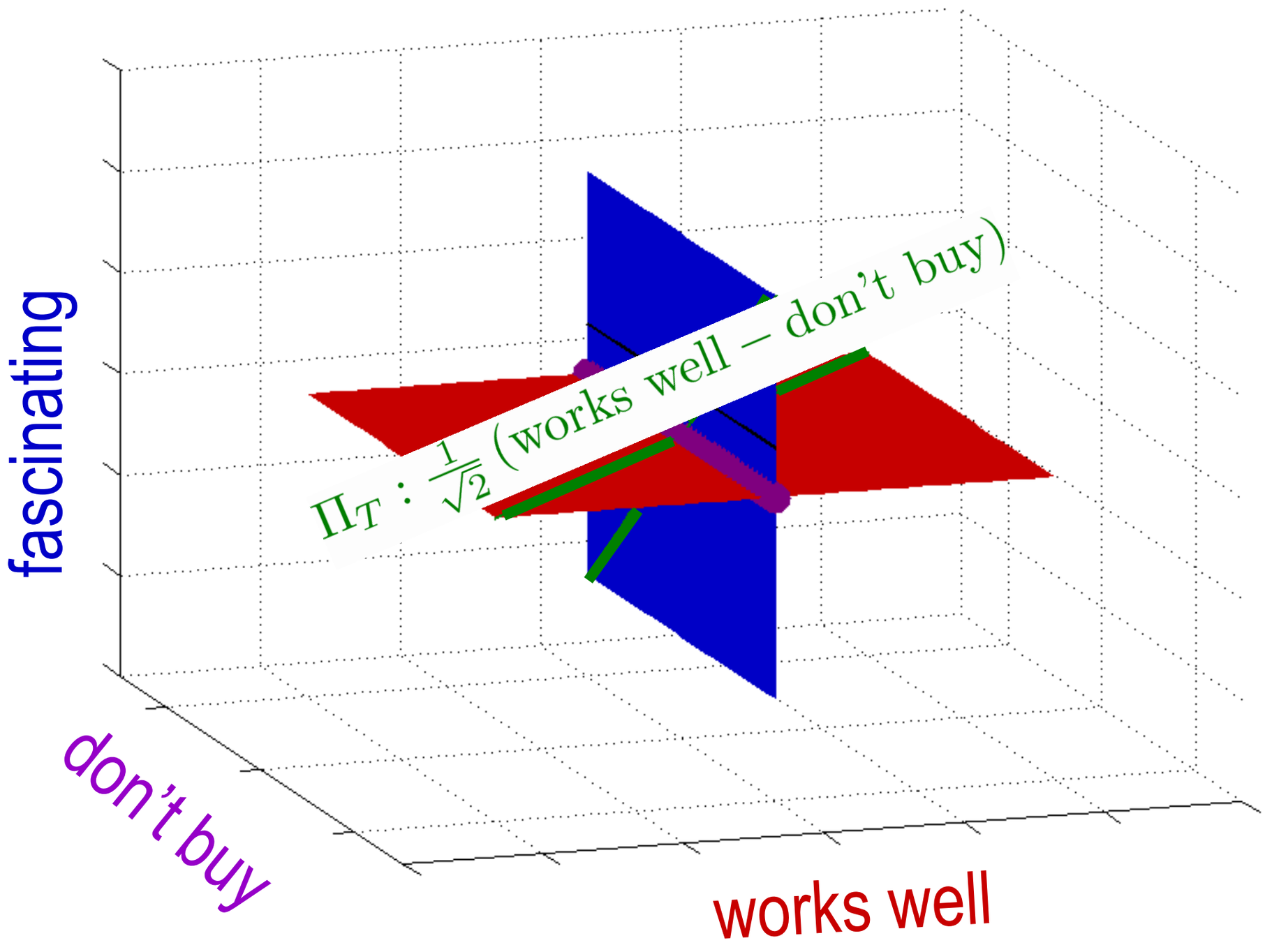
Π_T couples (works well) and -(don't buy)

Π_S & Π_T learned from unlabeled data

Visualizing Dimensionality Reduction

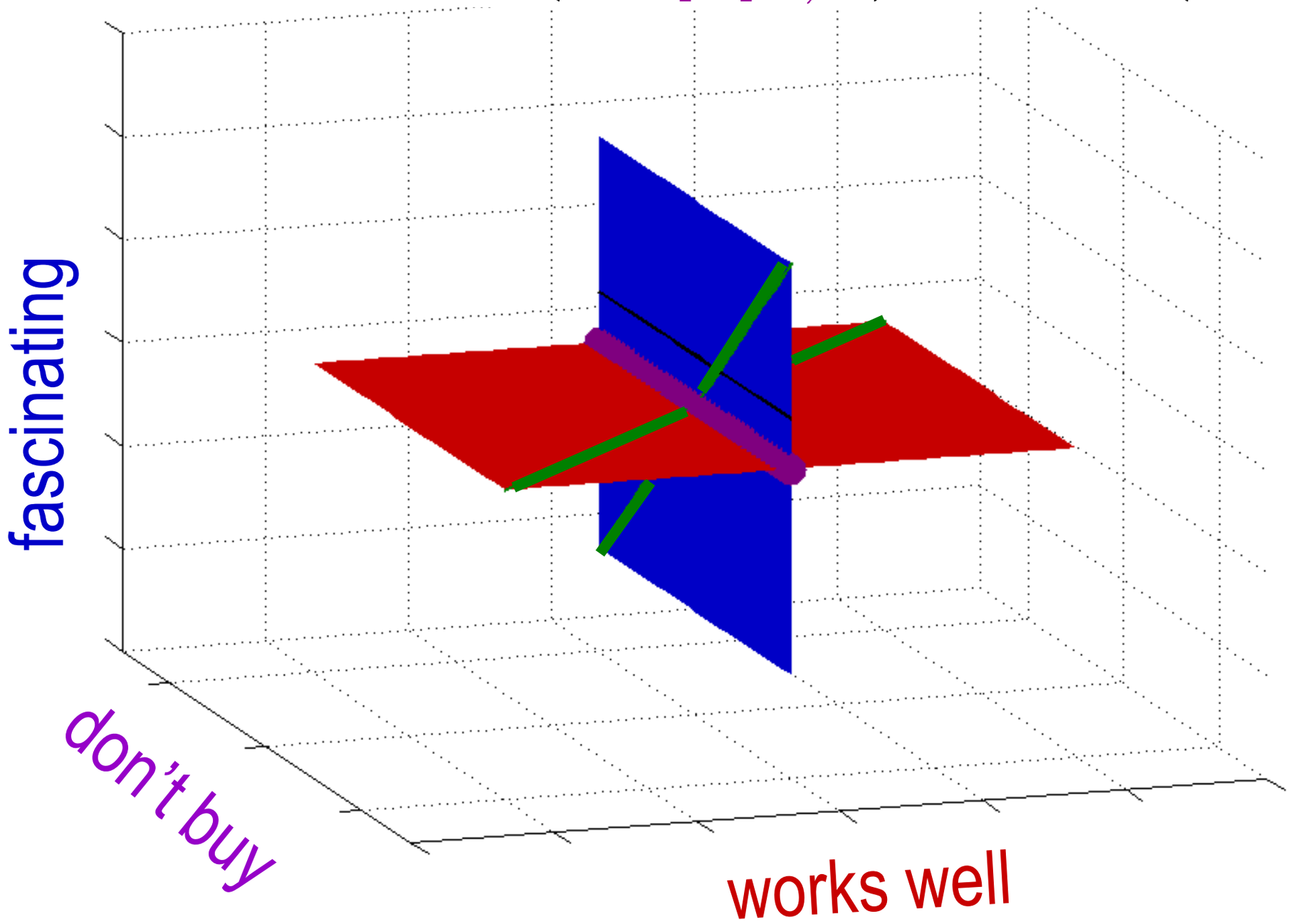


Visualizing Dimensionality Reduction



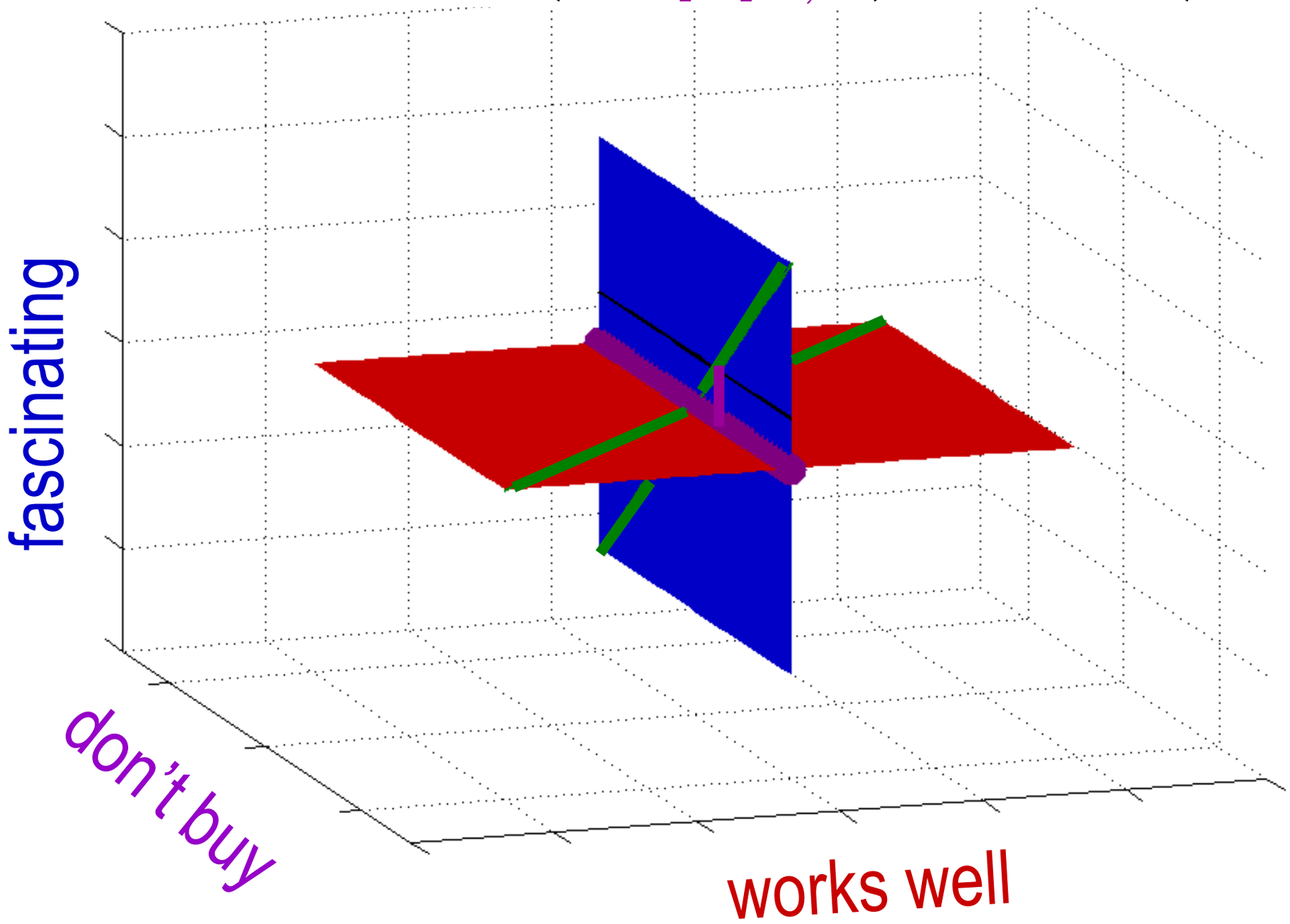
Representation Soundness

Lemma: $\forall x \quad \beta_S \cdot (\Pi_S[x]_{S,T}) = \beta_T \cdot (\Pi_T[x]_{S,T})$



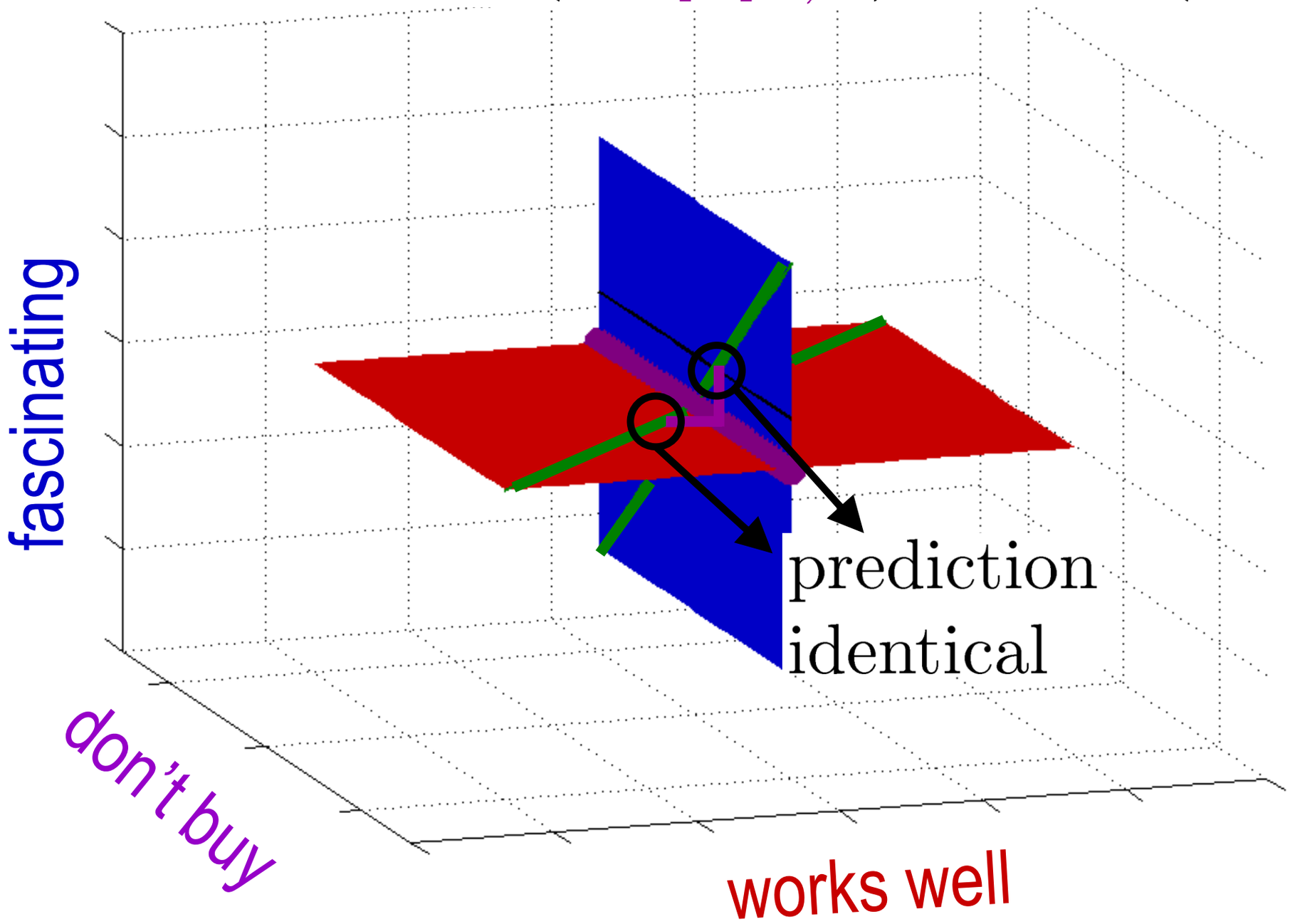
Representation Soundness

Lemma: $\forall x \quad \beta_S \cdot (\Pi_S[x]_{S,T}) = \beta_T \cdot (\Pi_T[x]_{S,T})$

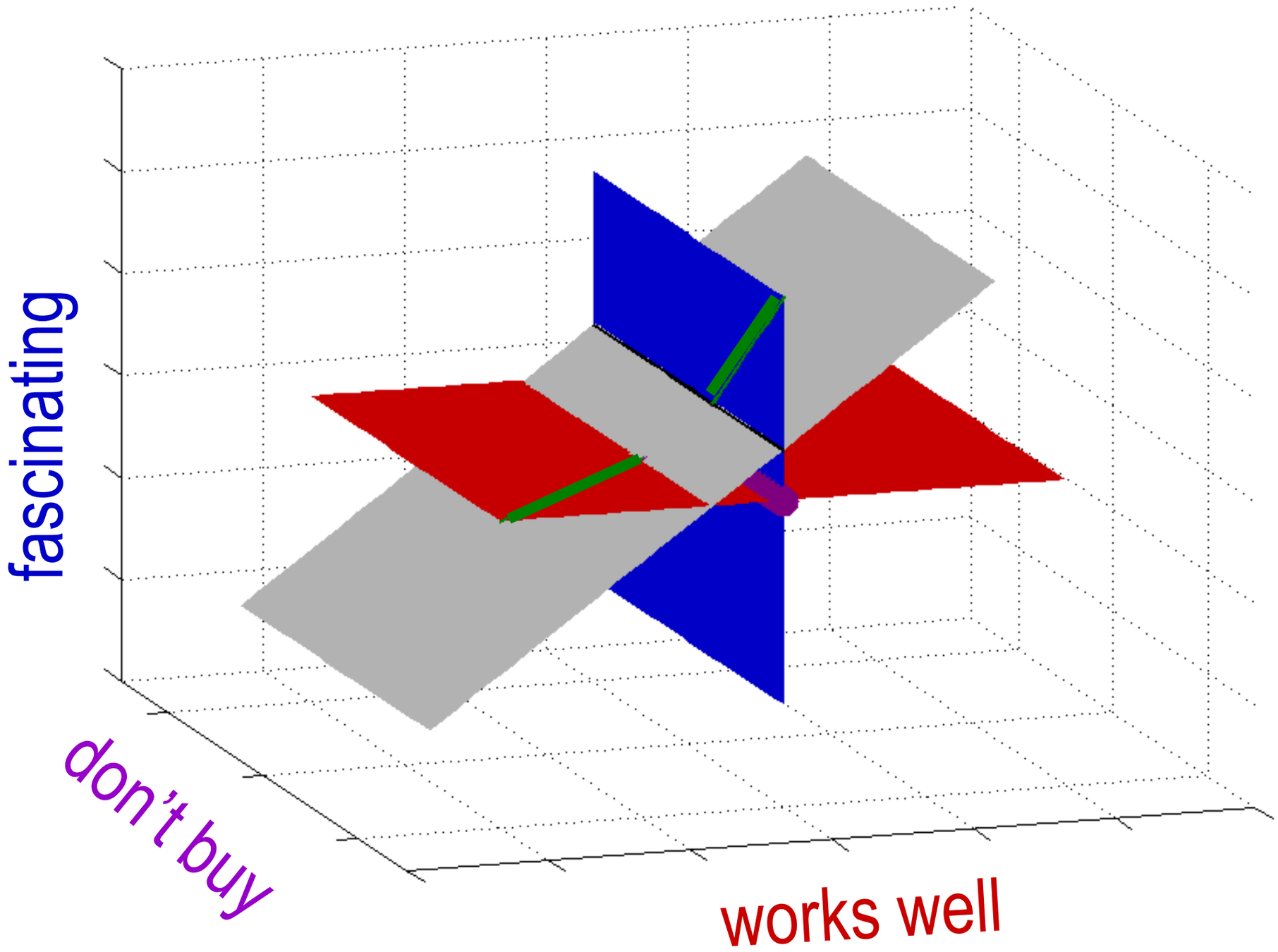


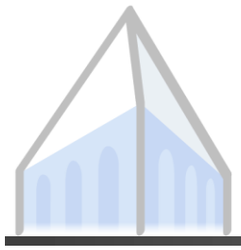
Representation Soundness

Lemma: $\forall x \quad \beta_S \cdot (\Pi_S[x]_{S,T}) = \beta_T \cdot (\Pi_T[x]_{S,T})$



Perfect Adaptation





Algorithm

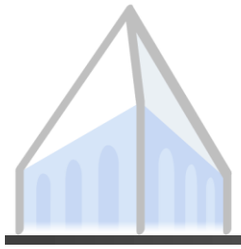
Input: Labeled source instances $(x_i, y_i)_{i=1}^n$

Unlabeled target instances x_T

1) Compute Π_S and Π_T (LDA, HMM, CCA)

2) $\left([\hat{\beta}]_{S,T}\right) = \underset{[\beta]_{S,T}}{\operatorname{argmin}} \sum_i \left([\beta]_{S,T}\Pi_T[x_i]_{S,T} - y_i\right)^2$

3) For target instance x, predict $[\beta]_{S,T}\Pi_T x$



Generalization

$$\text{Let } \Sigma_T = I \quad n = \text{num source instances}$$

$$\Sigma_{S \to T} = \sum_i \left(\Pi_T [x_i]_{S,T}\right) \left(\Pi_T [x_i]_{S,T}\right)^\top$$

$$\lambda_j = \text{eigenvalues of } \Sigma_{S \to T}$$

Under perfect adaptation, we have

$$\ell_T([\hat{\beta}]_{S,T}) - \ell_T(\beta_T^*) \leq$$

Generalization

$$\frac{d}{n} \text{ when } S = T$$

Under perfect adaptation, we have

$$\ell_T([\hat{\beta}]_{S,T}) - \ell_T(\beta_T^*) \leq \frac{\sum_j \frac{1}{\lambda_j}}{n}$$

Computing Π_S and Π_T

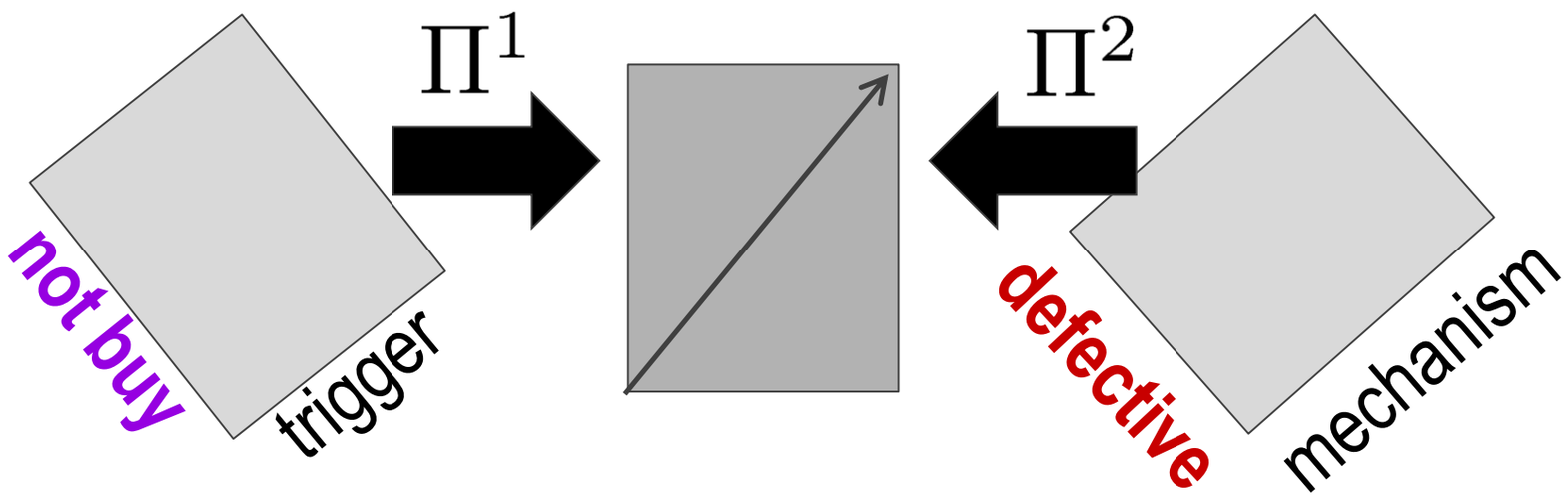
Canonical Correlation Analysis (CCA) [Hotelling 1935]

1) Divide feature space into disjoint views

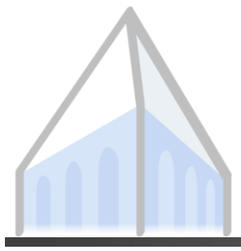
Do **not buy** the Shark portable steamer. The trigger mechanism is **defective**.

$$x^1 = \begin{array}{|c|}\hline \textbf{not buy} \\ \hline \textbf{trigger} \\ \hline \end{array} \qquad x^2 = \begin{array}{|c|}\hline \textbf{defective} \\ \hline \textbf{mechanism} \\ \hline \end{array}$$

2) Find maximally correlating projections



$$\Pi_T = \begin{bmatrix} \Pi_T^1 & 0 \\ 0 & \Pi_T^2 \end{bmatrix}$$



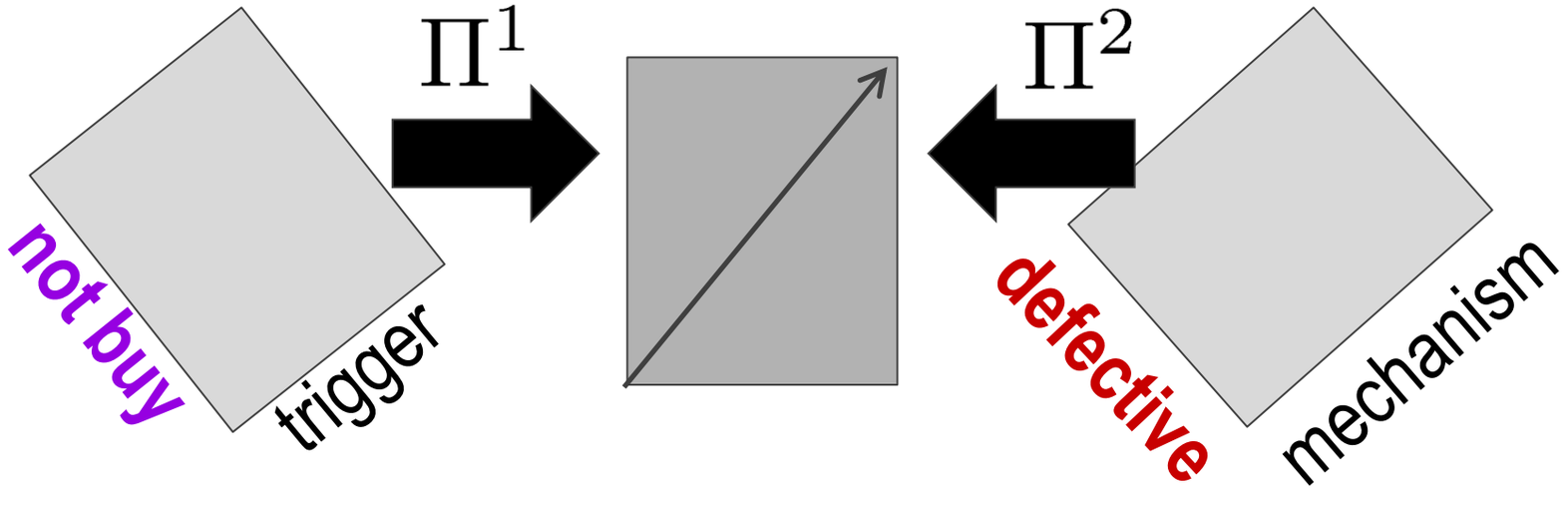
Computing Π_S and Π_T

Canonical Correlation Analysis (CCA) [Hotelling 1935]

Ando and Zhang (ACL 2005)

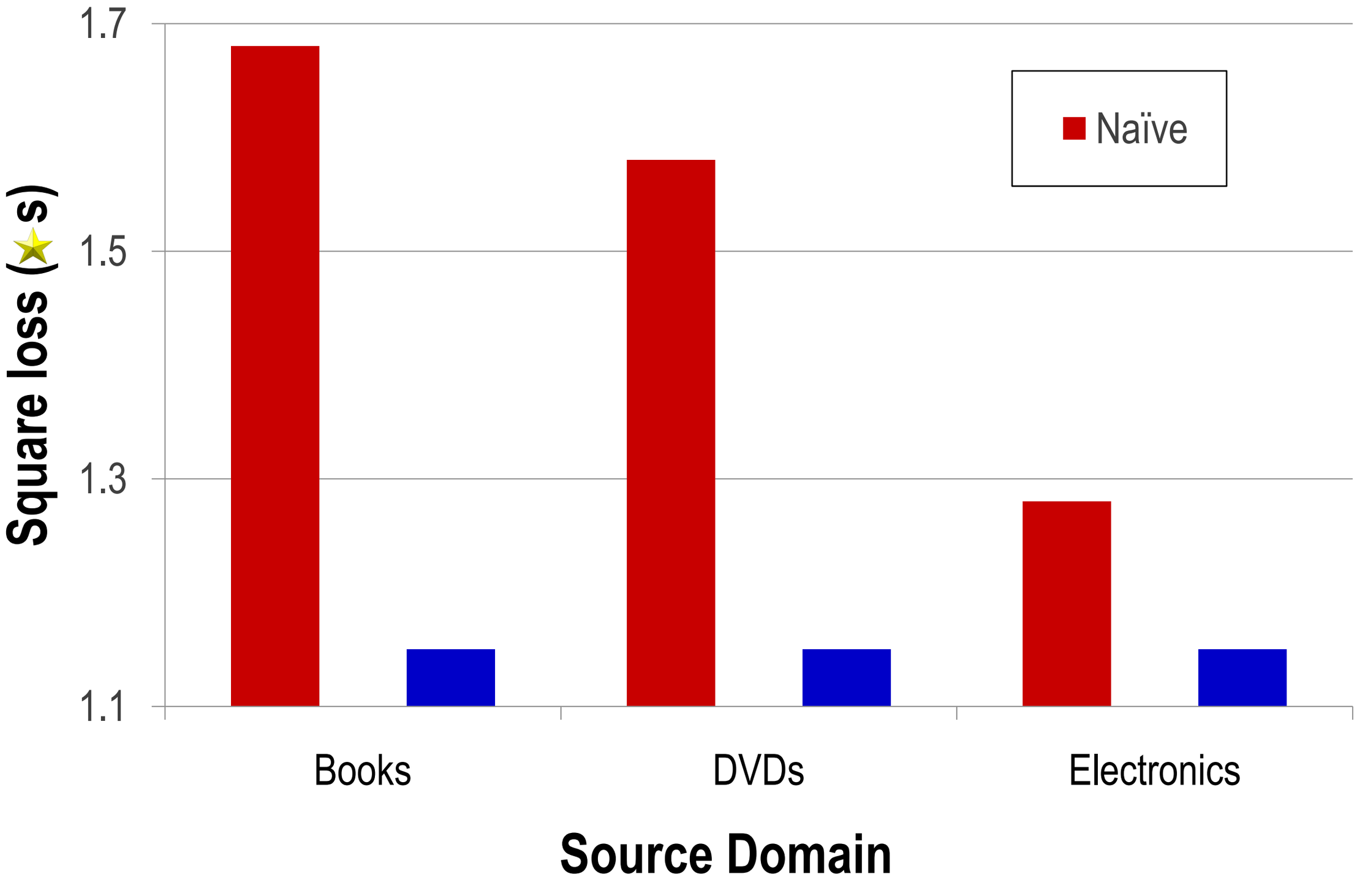
Kakade and Foster (COLT 2006)

2) Find maximally correlating projections

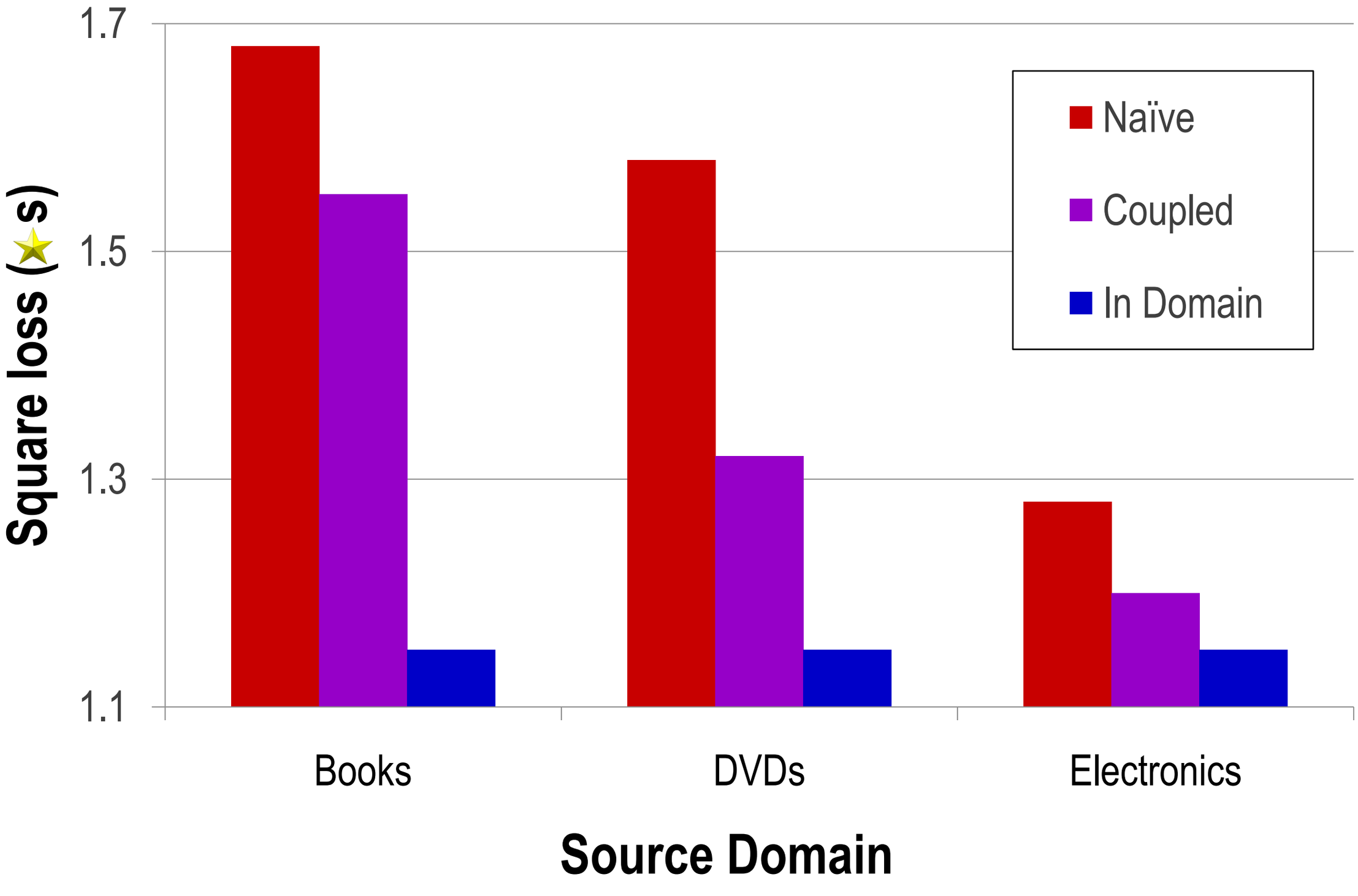


$$\Pi_T = \begin{bmatrix} \Pi_T^1 & 0 \\ 0 & \Pi_T^2 \end{bmatrix}$$

Square Loss: Kitchen Appliances

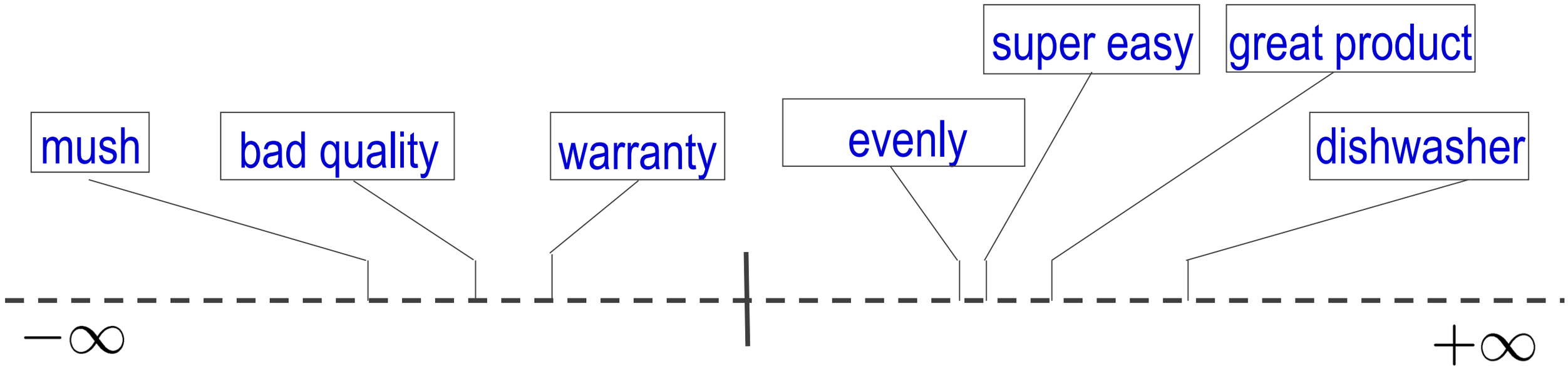


Square Loss: Kitchen Appliances

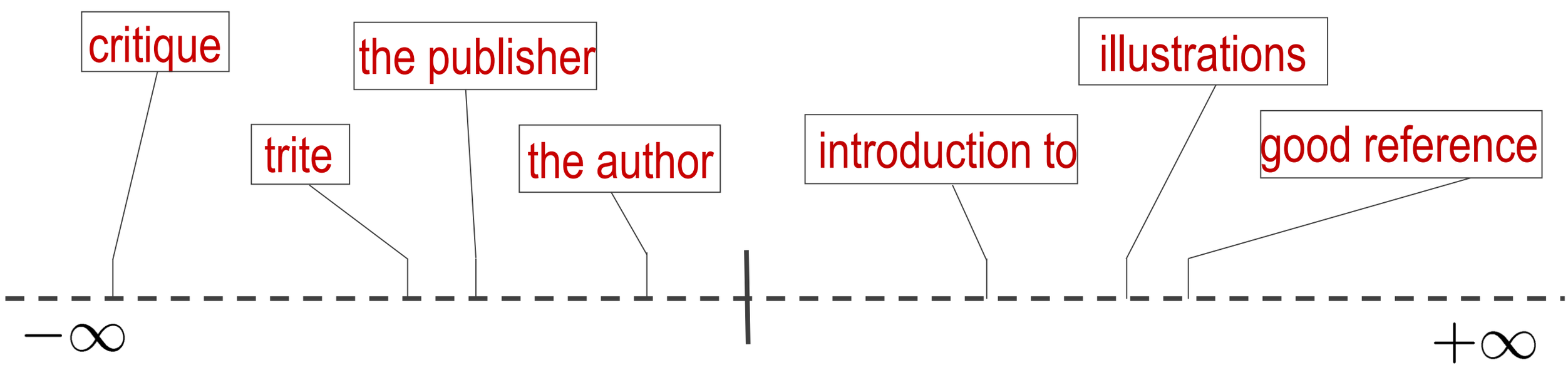


Using Target-Specific Features

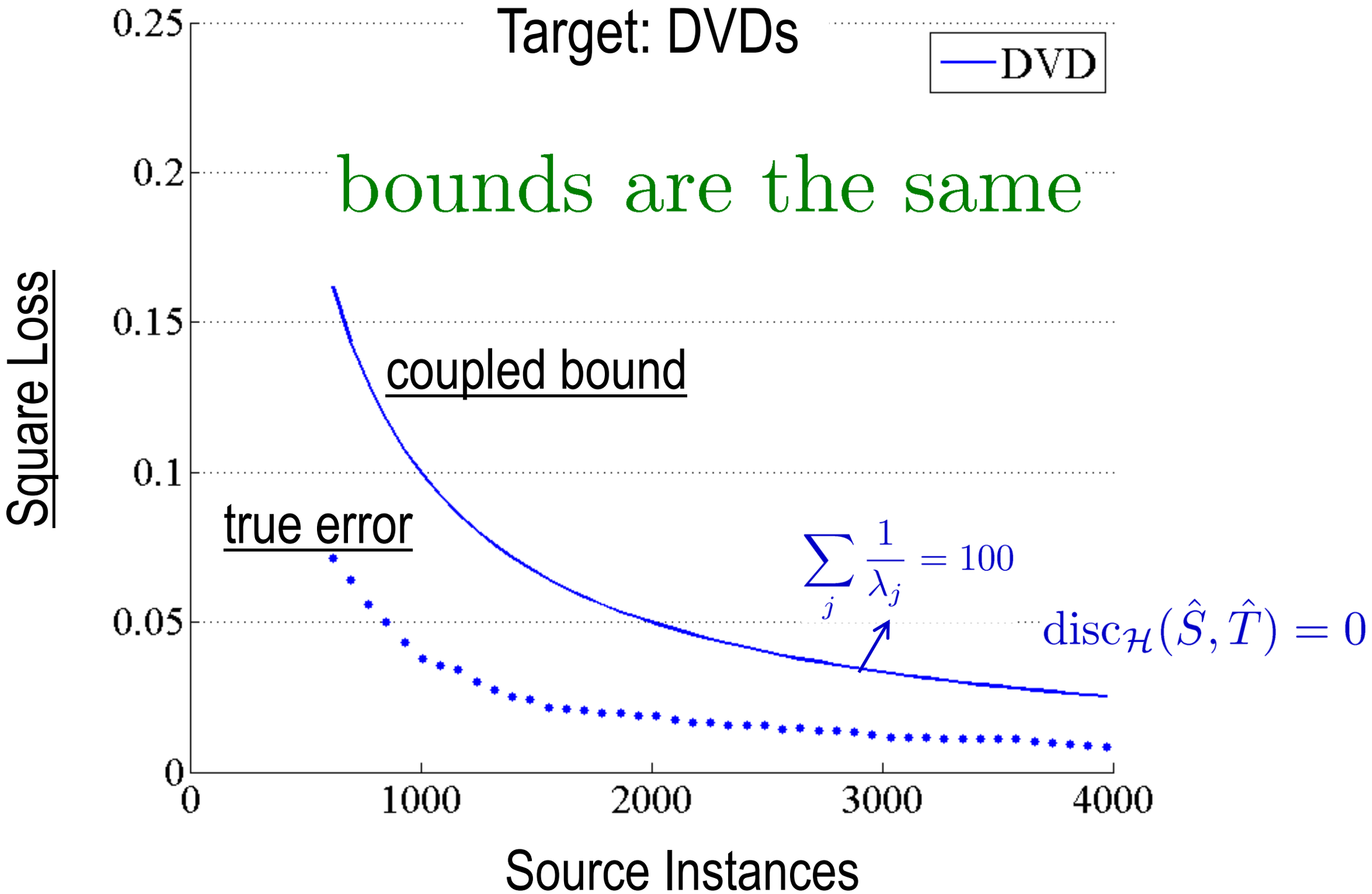
books → kitchen



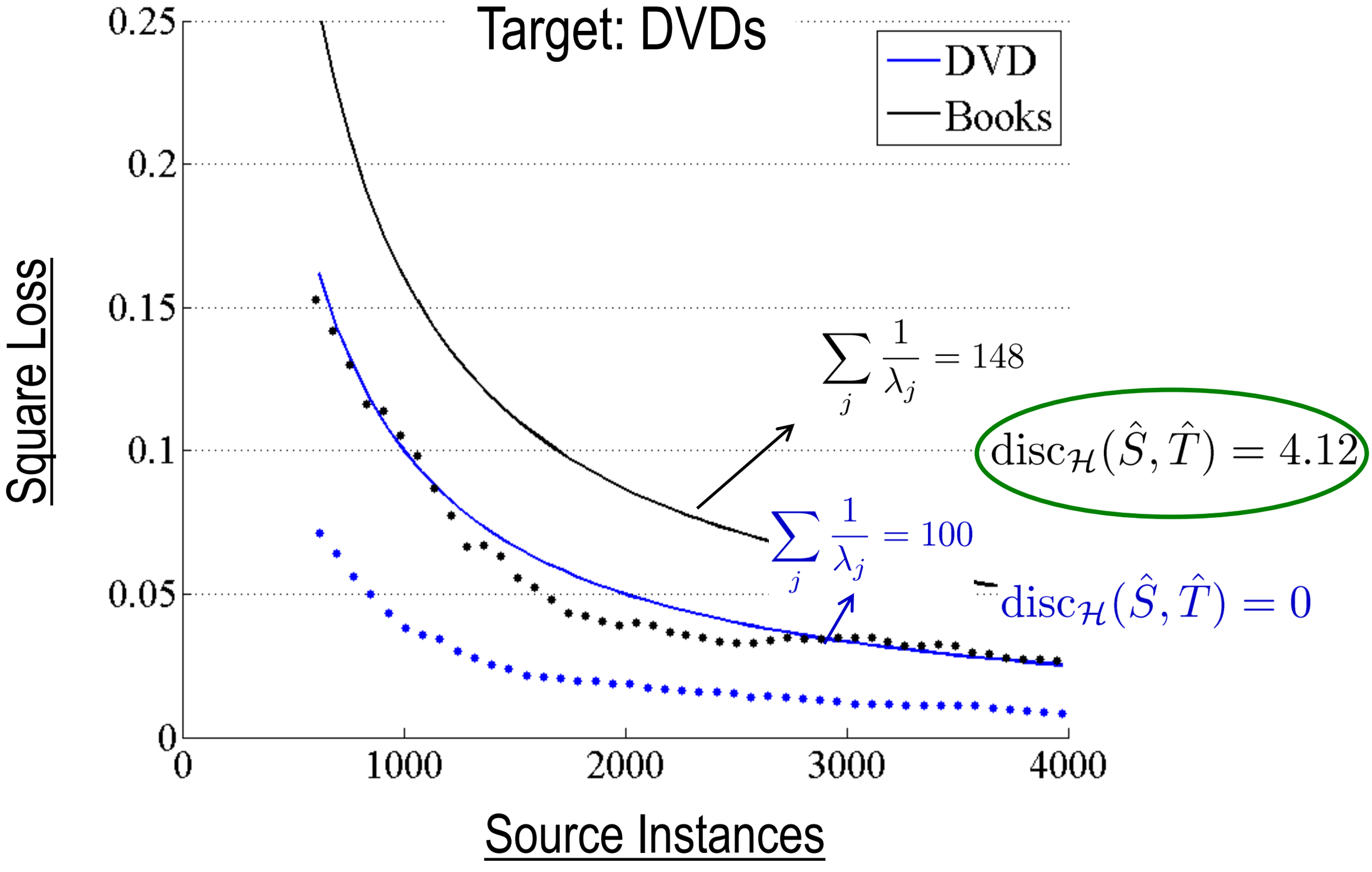
kitchen → books



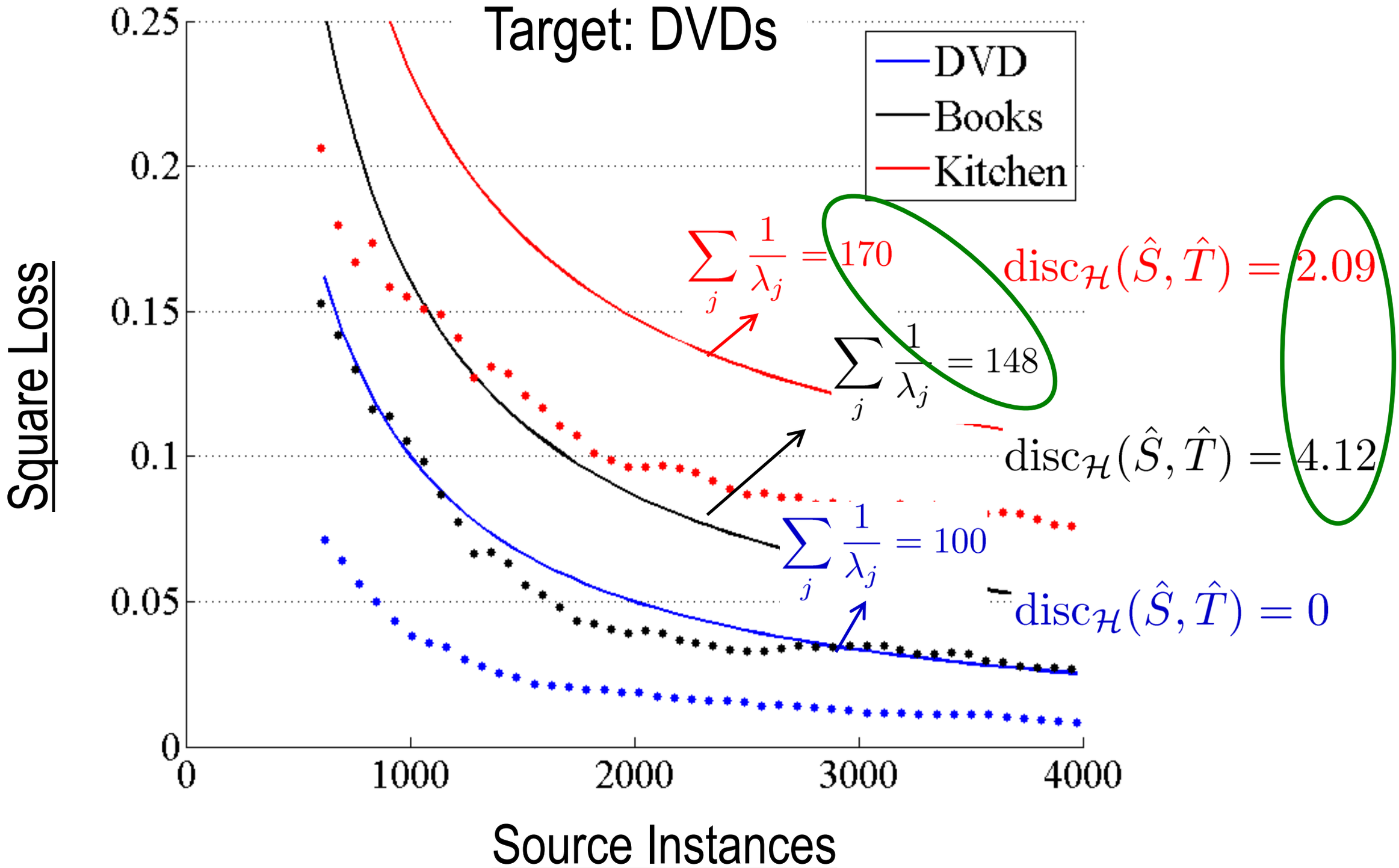
Comparing Discrepancy & Coupled Bounds

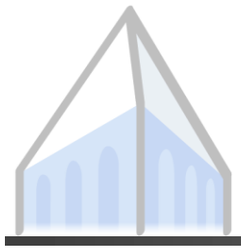


Comparing Discrepancy & Coupled Bounds



Comparing Discrepancy & Coupled Bounds





Idea: Active Learning

Only label really new target instances

$$\Sigma_{T \to S} = \sum_{x_i \in T} \left(\Pi_S [x_i]_{S,T}\right) \left(\Pi_S [x_i]_{S,T}\right)^\top$$

Order $x \in \mathcal{X}_T$ by

$$\frac{||\Pi_T x||^2_{\Sigma_T^{-1}}}{||\Pi_T x||^2_{\Sigma_{T \to S}^{-1}}}$$

Piyush Rai et al. (2010)

Ratio is

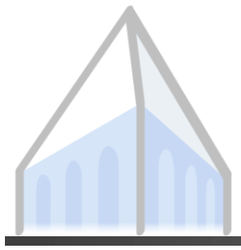
- 1 when $S = T$
- ∞ when $\Pi_T x$ has no shared part

Goals

1. **A computable bound** ✓

2. **Description of shared representations** ✓

3. **Suggestions for future research** ✓



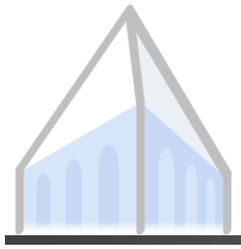
Conclusion

1. Theory can help us understand domain adaptation better

2. Good theory suggests new directions for future research

3. There's still a lot left to do

- Connecting supervised and unsupervised adaptation
- Unsupervised adaptation for problems with structure



Thanks

Collaborators

Shai Ben-David

Koby Crammer

Dean Foster

Sham Kakade

Alex Kulesza

Fernando Pereira

Jenn Wortman

References

Ben-David et al. A Theory of Learning from Different Domains. Machine Learning 2009.

Mansour et al. Domain Adaptation: Learning Bounds and Algorithms. COLT 2009.