Retrospective ICML99

Transductive Inference for Text Classification using Support Vector Machines

Thorsten Joachims

Then: Universität Dortmund, Germany
Now: Cornell University, USA

Outline

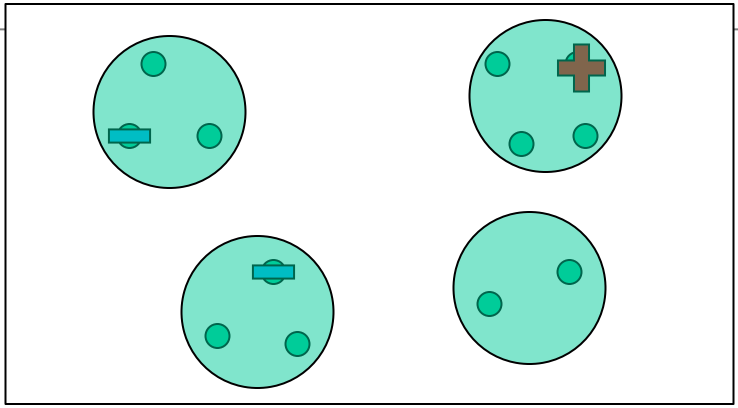
- **The paper in a nutshell**
- **Connections to other semi-supervised methods**
  - Co-training
  - Graph Mincuts
  - Normalized cuts
  - Harmonic functions
  - Manifold methods
  - Random walks
- **Post-mortem**
- **Valuable life lessons**

Input

Tom Mitchell

"What can we do with all the text data on the web?"

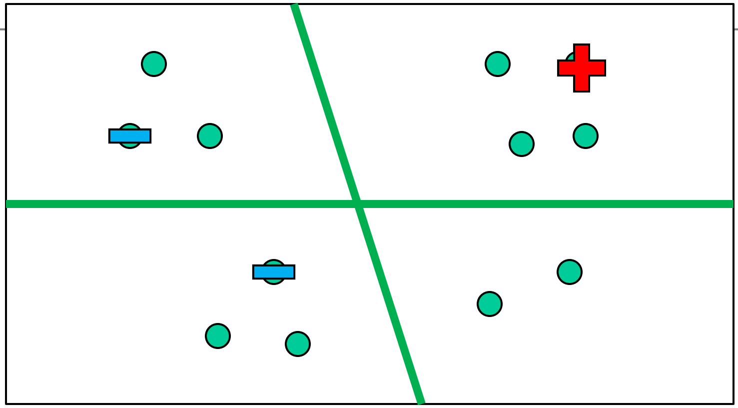
- [Blum/Mitchell] Co-training
  - Exploit redundant representations
- [Nigam/McCallum/Thrun/Mitchell] Semi-supervised Naïve Bayes
  - Generatively model clusters in P(X)
  - Mixture model



Vladimir Vapnik

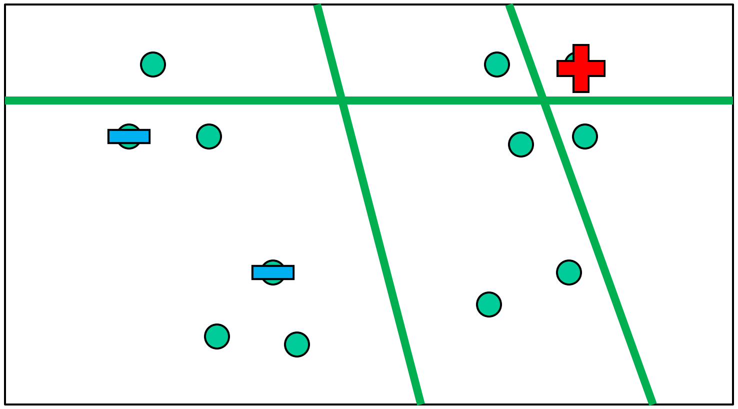
Transduction: Predicting only at known locations is easier

- Finite number of predictions vs. continuous function
- Define margin w.r.t. test points
- Generalization error bounds



Transductive SVMs

- **Objective [Vapnik]: Max margin on training and test set**
- **Input:**
  - Location of examples: $\{x_1 \ldots x_n\}$
  - Labels for subset L of examples



Hard Margin:

$$\min_{\boldsymbol{y}} \min_{\boldsymbol{w}} \quad \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}$$

$$s.t. \quad \forall i : y_i[\boldsymbol{w}^T\boldsymbol{x} + b] \geq 1$$

$$\forall i \in L : y_i = 1/-1$$

$$\boldsymbol{y} \in \{+1, -1\}$$

$$\boldsymbol{y}^T 1 = c$$

Soft Margin:

$$\min_{\boldsymbol{y}} \min_{\boldsymbol{w}} \quad \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum \xi_i$$

$$s.t. \quad \forall i : y_i[\boldsymbol{w}^T\boldsymbol{x}+b] \geq 1-\xi_i$$

$$\forall i \in L : y_i = 1/-1$$

$$\boldsymbol{y} \in \{+1, -1\}$$

$$\boldsymbol{y}^T 1 = c$$

Class balance constraint

Text and Margins

	nuclear	physics	atom	pepper	basil	salt	and
D1	**1**						**1**
D2	1	1	1				1
D3			1				1
D4				1	1		1
D5				1		1	1
D6					**1**	**1**	**1**

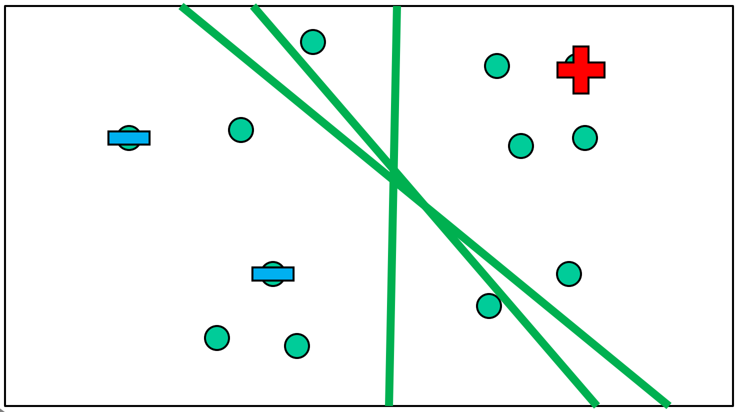
Altavista (1999)
- hits(pepper & salt) → 327K
- hits(pepper & physics) → 4.2K
- hits(physics) > hits(salt)

Google (2009)
- hits(pepper & salt) → 159M
- hits(pepper & physics) → 1.3M
- hits(physics) = 107M > hits(salt) = 56M

Prof. Michael Pepper → Prof. Sir Michael Pepper

Training Algorithm

- **Algorithm (http://svmlight.joachims.org)**
  - Assign labels to test examples (s.t. class balance constraint)
  - Train supervised SVM
  - DO
    - Find pair of test labels to flip
    - Retrain supervised SVM
  - WHILE objective decreased



Soft Margin:

Smoothed objective to avoid local optima
Smoothing reduced as optimization progresses

Criterion for selecting pair that guarantees descent
Criterion is efficiently computable

$$\forall i \in L : y_i = 1/-1$$

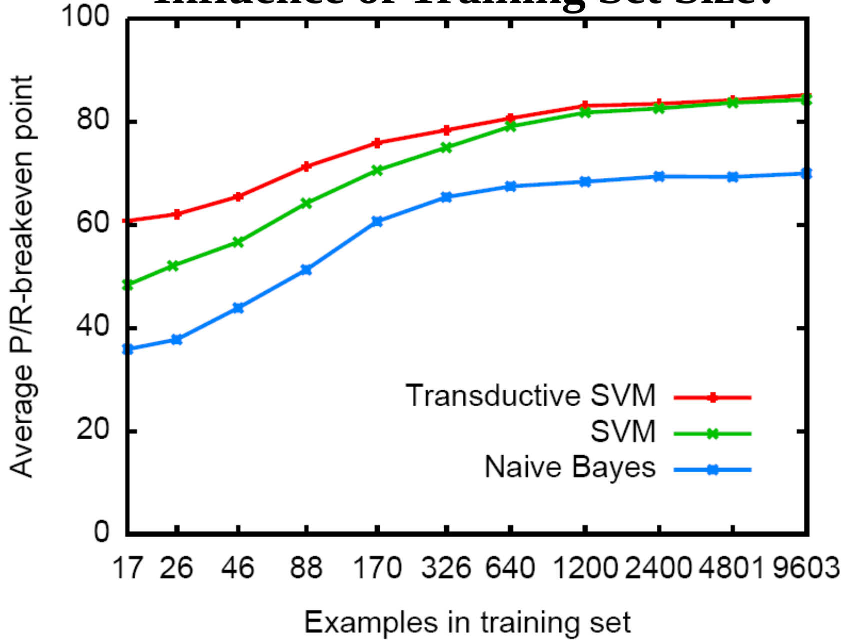
$$\boldsymbol{y} \in \{+1, -1\}$$

$$\boldsymbol{y}^T 1 = c$$

Experiment: Reuters-21587

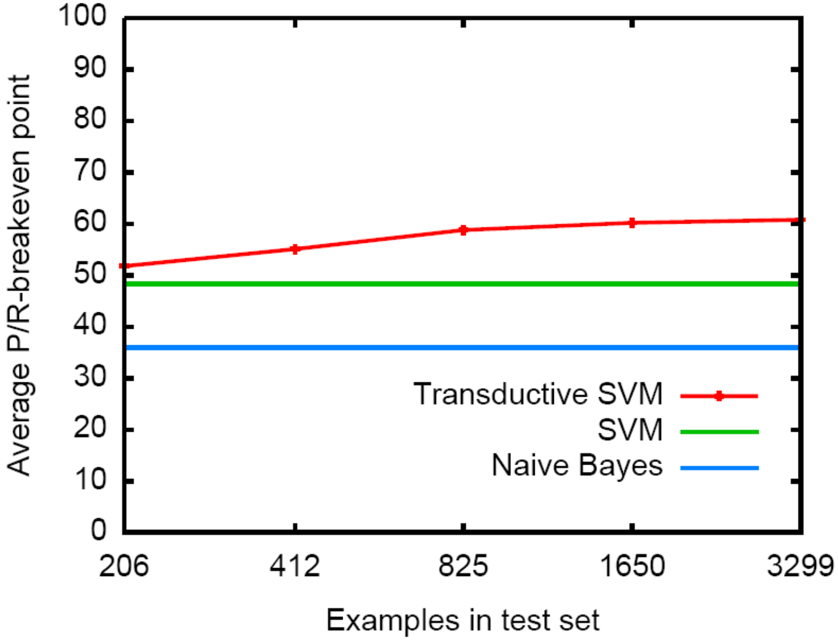
- **Setup**
  - Top 10 categories of Reuters-21587 dataset
  - ~12000 features after stemming and stopword removal
  - Macro-averaged precision/recall break-even point

Influence of Training Set Size?



(3299 test examples)

Influence of Test Set Size?



(17 training examples)

Experiment: WebKB

- **Setup**
  - 4 classes
  - 9 training examples, 3957 test examples
  - Precision/recall break-even point per class (and average)

	Bayes	SVM	TSVM
course	57.2	68.7	93.8
faculty	42.4	52.5	53.7
project	21.4	37.5	18.4
student	63.5	70.0	83.8
macro-average	46.1	57.2	62.4

Other Approaches

- **Optimization Methods for TSVM Objective**
  - Semi-definite Programming relaxation (convex) [Xu et al.]
  - Gradient Descent in Primal [Chapelle/Zien]
  - Concave Convex Procedure [Collobert et al.]
- **Other Objectives**
  - Manifolds and Graph Kernels [Belkin/Niyogi] [Chapelle et al.]
  - Harmonic Functions and Gaussian Processes [Zhu et al.]
  - Random Walks [Szummer/Jaakola]
  - Graph Cuts [Blum/Chawla]
  - Kernels from Generative Models [Jaakola/Haussler]
- **Special Structure of Problem**
  - Co-Training [Blum/Mitchell]
  - Structured Output Prediction [Brefeld/Scheffer]
- **Transductive Error Bounds**
- **Much more…**

Self-Consistency and Stability

- **Inductive Learner:** L_{ind}
- **Transductive Learner:** L_{trans} (based on L_{ind})
- **Assumption**
  - If whole sample was labeled, then L_{ind} would learn accurate classifier.
- **Reasoning**
  - If assumption holds, then L_{ind} will have low leave-one-out error.
  - If L_{trans} returns a labeling on which L_{ind} would have high leave-one-out error, it cannot be the correct labeling.

➔ **Construct prior of L_{trans} via leave-one-out error of L_{ind}.**
  - Margin wrt. test set bounds leave-one-out error of inductive SVM.
  - Ridge Regression [Chapelle et al.]
  - Graph-cuts [Blum/Chawla]

see [Joachims/03a]

Redefining Margin

Primal:

$$\min_{\boldsymbol{y}} \min_{\boldsymbol{w}} \ \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}$$

$$s.t. \ \forall i : y_i \boldsymbol{w}^T \boldsymbol{x} \geq 1$$

$$\forall i \in L : y_i = 1/-1$$

$$\boldsymbol{y} \in \{+1, -1\}$$

Dual:

$$\min_{\boldsymbol{y}} \max_{\boldsymbol{\alpha} \geq 0} \ 1^T\boldsymbol{\alpha} - \frac{1}{2}\boldsymbol{\alpha}^T Y A Y \boldsymbol{\alpha}$$

$$s.t. \ \forall i : Y_{ii} = y_i$$

$$\forall i \in L : y_i = 1/-1$$

$$\boldsymbol{y} \in \{+1, -1\}$$

$$\alpha_1 = \ldots = \alpha_n$$

Classification Rule / Margin:

$$h(\boldsymbol{x}) = sign\left\{\sum_{i=1}^{n} y_i \cancel{\alpha_i} K(\boldsymbol{x}, \boldsymbol{x}_i)\right\}$$

$$m(\boldsymbol{x}, y) = 1 - y \sum_{i=1}^{n} y_i \cancel{\alpha_i} K(\boldsymbol{x}, \boldsymbol{x}_i)$$

Nearest Neighbor Rule

Simplified Dual:

Min bound on leave-one-out error of NN

$$\min_{\boldsymbol{y}} \ -\boldsymbol{y}^T A \boldsymbol{y}$$

$$s.t. \ \forall i \in L : y_i = 1/-1$$

$$\boldsymbol{y} \in \{+1, -1\}$$

Connection to Graph Cuts [Blum/Chawla]

Simplified Dual:

$$\min_{\boldsymbol{y}} \; -\boldsymbol{y}^T A \boldsymbol{y}$$

$$s.t. \; \forall i \in L : y_i = 1/-1$$

$$\boldsymbol{y} \in \{+1, -1\}$$

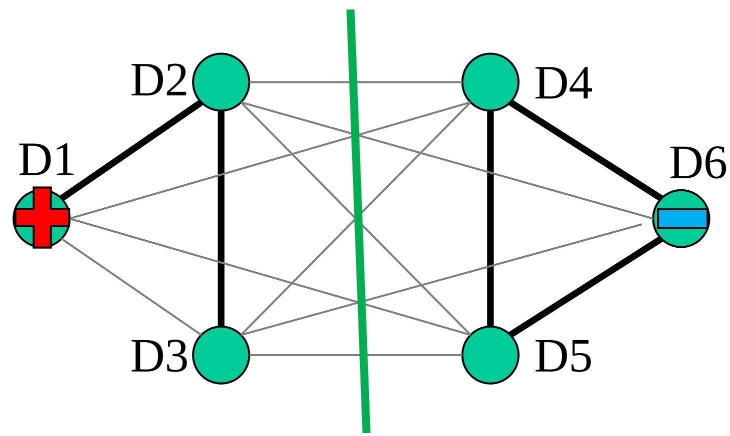
Graph Cut:

$$\min_{\boldsymbol{y}} \sum_{y_i \neq y_j} A_{ij} = \sum_{ij} A_{ij}(y_i - y_j)^2$$

$$s.t. \; \forall i \in L : y_i = 1/-1$$

$$\boldsymbol{y} \in \{+1, -1\}$$

	nuclear	physics	atom	pepper	basil	salt	and
D1	**1**						**1**
D2	1	1	1				1
D3			1				1
D4				1	1		1
D5				1		1	1
D6					**1**	**1**	**1**



➔ Fast algorithms for computing cuts for sparse graphs (e.g. k-NN)

Connection to Harmonic Functions [Zhu/Ghahramani/Lafferty]

Graph Cut:

$$\min_{\boldsymbol{y}} \sum_{ij} A_{ij}(y_i - y_j)^2$$
$$s.t.\ \forall i \in L : y_i = 1/-1$$
$$\boldsymbol{y} \in \{+1, -1\}$$

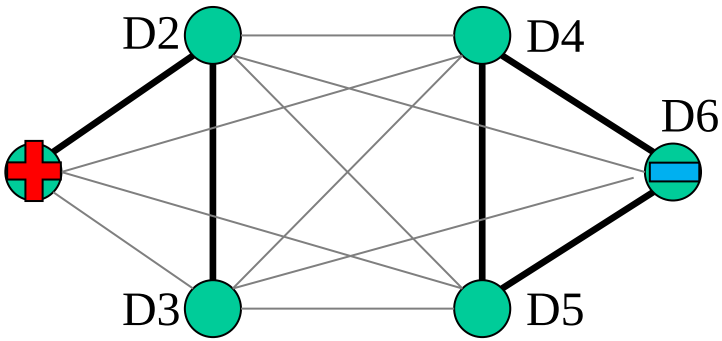
relax

Harmonic:

$$\min_{\boldsymbol{y}} \sum_{ij} A_{ij}(y_i - y_j)^2$$
$$s.t.\ \forall i \in L : y_i = 1/-1$$
$$\boldsymbol{y} \in [+1, -1]$$

Interpretations:

- Gaussian process
- Electric network
- Probability that random walk hits positively labeled node first
  → Connection to [Szummer/Jaakkola]



→ Closed form solution and/or very efficient iterative methods

Connection to Normalized Cuts [Joachims]

Graph Cut:

$$\min_{\boldsymbol{y}} \sum_{ij} A_{ij}(y_i - y_j)^2$$
$$s.t.\ \forall i \in L : y_i = 1/-1$$
$$\boldsymbol{y} \in \{+1, -1\}$$

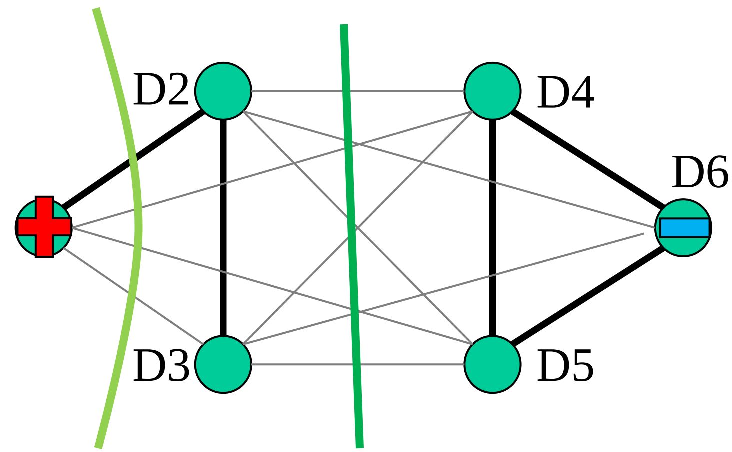
norm

Normalized (Ratio) Cut:

$$\min_{\boldsymbol{y}} \sum_{ij} A_{ij}(y_i - y_j)^2 / \sum_{ij}(y_i - y_j)^2$$
$$s.t.\ \forall i \in L : y_i = 1/-1$$
$$\boldsymbol{y} \in \{+1, -1\}$$

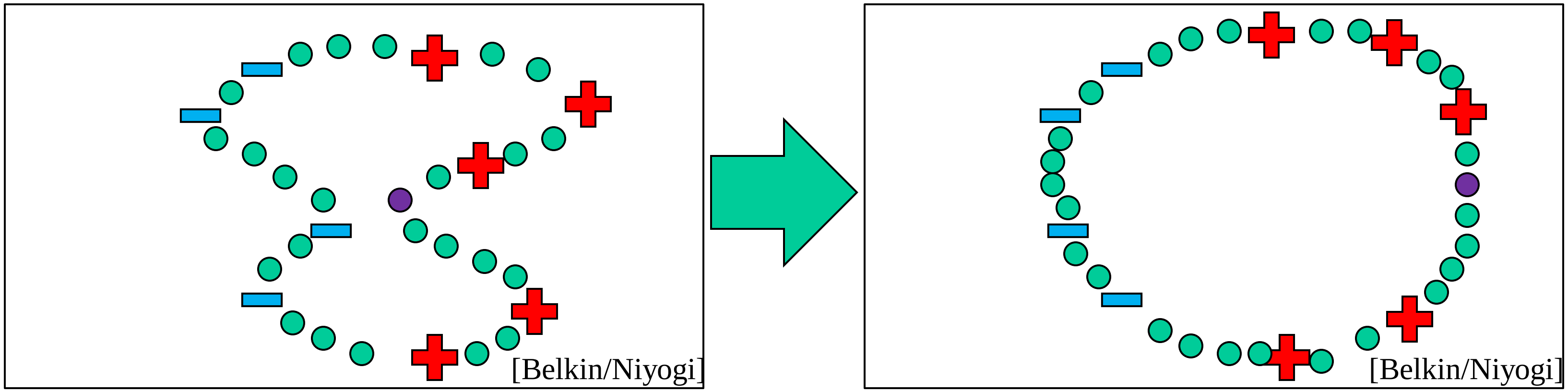
Interpretations:

- Minimize average weight of cut edge
- Spectral relaxation has efficient solution
  - →Normalized cuts [Shi/Malik]
- "Supervised" normalized cut
  - → Supervised clustering [Yu/Gross/Shi]



➔ Efficient solution of spectral relaxation

Connection to Manifolds and Graph Kernels [Belkin/Niyogi] [Chapelle et al.]



Exploit Manifold Structure

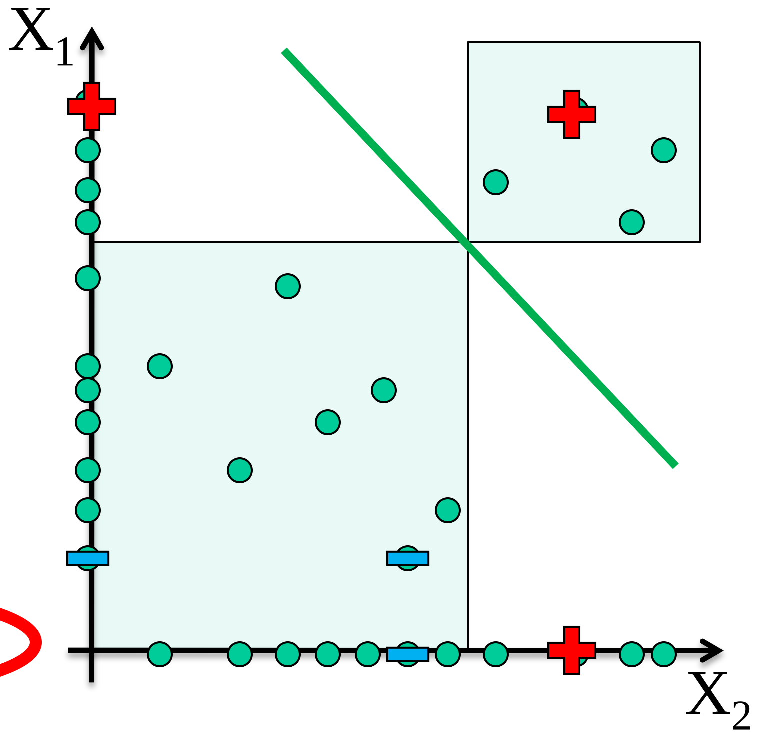
- Smoothness criterion $\sum_{ij} A_{ij}(y_i - y_j)^2 = \boldsymbol{y}^T L \boldsymbol{y}$ related to graph Laplacian L=D-A
- Not Euclidian distance, but geodesic distance in local neighborhood graph
- Use eigenvectors $U\Lambda U^T = L$ of graph Laplacian L to
  - explicitly re-represent data [Roweis/Saul] [Tennenbaum et al.]
  - define a kernel (e.g. to use in inductive SVM) [Kondor/Lafferty]

Connection to Co-Training [Blum/Mitchell]

- **Idea:**
  - Exploit two sufficiently redundant representations
- **Example:**
  - Learn threshold on X_1 / X_2

  ➔ Co-training implies margin
- **Experiment:**
  - Error rate on WebKB "course"

	SVM	TSVM	B&M
page	21.6	4.6	12.9
link	18.5	8.9	12.4
co-train	20.3	4.3	5.0



Post Mortem

- **Why does Transductive Learning Work?**
  - Smoothness: labels change smoothly with structure of unlabeled data (clusters, manifold).
  - Self-Consistency: if all examples were labeled, supervised learner has low leave-one-out error.
- **Transduction vs. Semi-supervised?**
  - Transduction = semi-supervised
- **Discriminative vs. Generative?**
  - No need for density estimate of P(X)
- **Use in Practice?**
  - Largest benefits for small training sets
  - Better mean, but (still) large variance
- **How can we use ALL the (text) data on the web?**