Keepin' It Real: Semi-Supervised Learning with Realistic Tuning

Andrew B. Goldberg
goldberg@cs.wisc.edu

Xiaojin Zhu
jerryzhu@cs.wisc.edu

Computer Sciences Department
University of Wisconsin-Madison

Gap between Semi-Supervised Learning (SSL) research and practical applications

Gap between Semi-Supervised Learning (SSL) research and practical applications

Semi-Supervised Learning:
Using unlabeled data to build better classifiers

Gap between Semi-Supervised Learning (SSL) research and practical applications

Semi-Supervised Learning: Using unlabeled data to build better classifiers

Real World

- natural language processing
- computer vision
- web search & IR
- bioinformatics
- etc

Gap between Semi-Supervised Learning (SSL) research and practical applications

Semi-Supervised Learning: Using unlabeled data to build better classifiers

Assumptions
- manifold? clusters?
- low-density gap?
- multiple views?

Parameters
- regularization?
- graph weights?
- kernel parameters?

Model Selection
- Little labeled data
- Many parameters
- Computational costs

Real World
- natural language processing
- computer vision
- web search & IR
- bioinformatics
- etc

Gap between Semi-Supervised Learning (SSL) research and practical applications

Assumptions

Wrong choices could hurt performance!

How can we ensure that SSL is never worse than supervised learning?

- Many parameters
- Computational costs

Our Focus

Our Focus

- Two critical issues
  - Parameter tuning
  - Choosing which (if any) SSL algorithm to use

Our Focus

- Two critical issues
  - Parameter tuning
  - Choosing which (if any) SSL algorithm to use

- Interested in realistic settings:
  - Practitioner is given some new labeled and unlabeled data
  - Must produce the best classifier possible

Our Contributions

Our Contributions

- Medium-scale empirical study

Our Contributions

- Medium-scale empirical study
  - Compares one supervised learning (SL) and two SSL methods

Our Contributions

- Medium-scale empirical study
  - Compares one supervised learning (SL) and two SSL methods
  - Eight less-familiar NLP tasks, three evaluation metrics

Our Contributions

- Medium-scale empirical study
  - Compares one supervised learning (SL) and two SSL methods
  - Eight less-familiar NLP tasks, three evaluation metrics
  - Experimental protocol explores several real-world settings

Our Contributions

- Medium-scale empirical study
  - Compares one supervised learning (SL) and two SSL methods
  - Eight less-familiar NLP tasks, three evaluation metrics
  - Experimental protocol explores several real-world settings
  - *All parameters are tuned realistically via cross validation*

Our Contributions

- Medium-scale empirical study
  - Compares one supervised learning (SL) and two SSL methods
  - Eight less-familiar NLP tasks, three evaluation metrics
  - Experimental protocol explores several real-world settings
  - *All parameters are tuned realistically via cross validation*
- Findings under these conditions:

OUR CONTRIBUTIONS

- Medium-scale empirical study
  - Compares one supervised learning (SL) and two SSL methods
  - Eight less-familiar NLP tasks, three evaluation metrics
  - Experimental protocol explores several real-world settings
  - *All parameters are tuned realistically via cross validation*
- Findings under these conditions:
  - Each SSL can be worse than SL on some data sets

Our Contributions

- Medium-scale empirical study
  - Compares one supervised learning (SL) and two SSL methods
  - Eight less-familiar NLP tasks, three evaluation metrics
  - Experimental protocol explores several real-world settings
  - *All parameters are tuned realistically via cross validation*
- Findings under these conditions:
  - Each SSL can be worse than SL on some data sets
  - Can achieve *agnostic SSL* by using cross validation accuracy to select among SL and SSL algorithms

OUTLINE

- Introduce "realistic tuning" for SSL
- Empirical study protocol
  - Data sets
  - Algorithms
  - Meta algorithm for SSL model selection
  - Performance metrics
- Results
- Conclusions

SSL With Realistic Tuning

SSL WITH REALISTIC TUNING

- Given labeled and unlabeled data, $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_l, y_l), \mathbf{x}_{l+1}, ..., \mathbf{x}_{l+u}\}$ how should you set parameters for some algorithm?

SSL WITH REALISTIC TUNING

- Given labeled and unlabeled data, $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_l, y_l), \mathbf{x}_{l+1}, ..., \mathbf{x}_{l+u}\}$ how should you set parameters for some algorithm?
  - Tune based on test set performance?

SSL WITH REALISTIC TUNING

- Given labeled and unlabeled data, $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_l, y_l), \mathbf{x}_{l+1}, ..., \mathbf{x}_{l+u}\}$ how should you set parameters for some algorithm?
  - Tune based on test set performance? No, this is cheating

SSL WITH REALISTIC TUNING

- Given labeled and unlabeled data, $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_l, y_l), \mathbf{x}_{l+1}, ..., \mathbf{x}_{l+u}\}$ how should you set parameters for some algorithm?
  - Tune based on test set performance? No, this is cheating
  - Use default values based on heuristics/experience?

SSL With Realistic Tuning

- Given labeled and unlabeled data, $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_l, y_l), \mathbf{x}_{l+1}, ..., \mathbf{x}_{l+u}\}$ how should you set parameters for some algorithm?
  - Tune based on test set performance? No, this is cheating
  - Use default values based on heuristics/experience? May fail on new data

SSL With Realistic Tuning

- Given labeled and unlabeled data, $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_l, y_l), \mathbf{x}_{l+1}, ..., \mathbf{x}_{l+u}\}$ how should you set parameters for some algorithm?
  - Tune based on test set performance? No, this is cheating
  - Use default values based on heuristics/experience? May fail on new data
  - k-fold cross validation?

SSL WITH REALISTIC TUNING

- Given labeled and unlabeled data, $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_l, y_l), \mathbf{x}_{l+1}, ..., \mathbf{x}_{l+u}\}$ how should you set parameters for some algorithm?
  - Tune based on test set performance? No, this is cheating
  - Use default values based on heuristics/experience? May fail on new data
  - k-fold cross validation? Little labeled data, but best available option

SSL With Realistic Tuning

- Given labeled and unlabeled data, $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_l, y_l), \mathbf{x}_{l+1}, ..., \mathbf{x}_{l+u}\}$ how should you set parameters for some algorithm?
  - Tune based on test set performance? No, this is cheating
  - Use default values based on heuristics/experience? May fail on new data
  - k-fold cross validation? Little labeled data, but best available option
- Cross validation choices:

SSL With Realistic Tuning

- Given labeled and unlabeled data, $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_l, y_l), \mathbf{x}_{l+1}, ..., \mathbf{x}_{l+u}\}$ how should you set parameters for some algorithm?
  - Tune based on test set performance? No, this is cheating
  - Use default values based on heuristics/experience? May fail on new data
  - k-fold cross validation? Little labeled data, but best available option
- Cross validation choices:
  - number of folds

SSL With Realistic Tuning

- Given labeled and unlabeled data, $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_l, y_l), \mathbf{x}_{l+1}, ..., \mathbf{x}_{l+u}\}$ how should you set parameters for some algorithm?
  - Tune based on test set performance? No, this is cheating
  - Use default values based on heuristics/experience? May fail on new data
  - k-fold cross validation? Little labeled data, but best available option
- Cross validation choices:
  - number of folds
  - how labeled and unlabeled data is divided into folds

SSL WITH REALISTIC TUNING

- Given labeled and unlabeled data, $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_l, y_l), \mathbf{x}_{l+1}, ..., \mathbf{x}_{l+u}\}$ how should you set parameters for some algorithm?
  - Tune based on test set performance? No, this is cheating
  - Use default values based on heuristics/experience? May fail on new data
  - k-fold cross validation? Little labeled data, but best available option
- Cross validation choices:
  - number of folds
  - how labeled and unlabeled data is divided into folds
  - parameter grid

RealSSL Procedure

RealSSL Procedure

Input:

- a single data set of labeled and unlabeled data *(one real-world scenario)*
- an algorithm (SSL or SL) and data-independent parameter grid
- performance metric *M*

REALSSL PROCEDURE

Input:

a single data set of labeled and unlabeled data *(one real-world scenario)*

an algorithm (SSL or SL) and data-independent parameter grid

performance metric *M*

Procedure:

1. Divide data into 5 folds *s.t. labeled/unlabeled ratio is preserved*

REALSSL PROCEDURE

Input:

a single data set of labeled and unlabeled data *(one real-world scenario)*

an algorithm (SSL or SL) and data-independent parameter grid

performance metric M

Procedure:

1. Divide data into 5 folds *s.t. labeled/unlabeled ratio is preserved*
2. For each parameter setting p in grid:

   Compute 5-fold average performance $M_{params=p}$

REALSSL PROCEDURE

Input:

a single data set of labeled and unlabeled data *(one real-world scenario)*

an algorithm (SSL or SL) and data-independent parameter grid

performance metric M

Procedure:

1. Divide data into 5 folds *s.t. labeled/unlabeled ratio is preserved*
2. For each parameter setting p in grid:

    Compute 5-fold average performance $M_{params=p}$

Output:

Model trained using the best parameters p = argmax M_{params}

Best average tuning performance (max M_{params})

Empirical Study Protocol

EMPIRICAL STUDY PROTOCOL

- Designed to simulate different settings a real-world practitioner might face for a new task and a set of algorithms to choose from

EMPIRICAL STUDY PROTOCOL

- Designed to simulate different settings a real-world practitioner might face for a new task and a set of algorithms to choose from
  - Labeled sizes = 10 or 100

EMPIRICAL STUDY PROTOCOL

- Designed to simulate different settings a real-world practitioner might face for a new task and a set of algorithms to choose from
  - Labeled sizes = 10 or 100
  - Unlabeled sizes = 100 or 1000

Empirical Study Protocol

- Designed to simulate different settings a real-world practitioner might face for a new task and a set of algorithms to choose from
  - Labeled sizes = 10 or 100
  - Unlabeled sizes = 100 or 1000
  - For each combination, run 10 trials with different random labeled and unlabeled data (same samples across algorithms)

Empirical Study Protocol

- Designed to simulate different settings a real-world practitioner might face for a new task and a set of algorithms to choose from
  - Labeled sizes = 10 or 100
  - Unlabeled sizes = 100 or 1000
  - For each combination, run 10 trials with different random labeled and unlabeled data (same samples across algorithms)
  - Same grid of algorithm-specific parameters used for all data sets

Empirical Study Protocol

Empirical Study Protocol

Input:

Fully labeled data set
Algorithm, Performance metric
Labeled sizes = {10, 100}, Unlabeled sizes = {100, 1000}

Empirical Study Protocol

Input:

Fully labeled data set
Algorithm, Performance metric
Labeled sizes = {10, 100}, Unlabeled sizes = {100, 1000}

Procedure:

Divide data into training data pool and a single test set
For each l and u value:

Repeat 10 times {
- Randomly select labeled & unlabeled data from training pool
- Use RealSSL for parameter tuning and model building
- Compute transductive and test performance

Empirical Study Protocol

Input:

Fully labeled data set
Algorithm, Performance metric
Labeled sizes = {10, 100}, Unlabeled sizes = {100, 1000}

Procedure:

Divide data into training data pool and a single test set
For each l and u value:

Repeat 10 times:
- Randomly select labeled & unlabeled data from training pool
- Use RealSSL for parameter tuning and model building
- Compute transductive and test performance

Output:

Tuning, transductive, and test performance for all l/u settings in 10 trials

DATA SETS

- Binary classification tasks

Name	d	$P(y=+)$	$\lvert D_{test} \rvert$	Description
MacWin	7511	0.51	846	Mac vs. Windows newsgroups
Interest	2687	0.53	1268	WSD: monetary sense vs. others
aut-avn	20707	0.65	70075	Auto vs. Aviation, SRAA corpus
real-sim	20958	0.31	71209	Real vs. Simulated, SRAA corpus
ccat	47236	0.47	22019	Corporate vs. rest, RCV1 corpus
gcat	47236	0.30	22019	Government vs. rest, RCV1 corpus
Wish-politics	13610	0.34	4999	Wish detection in political discussion
Wish-products	4823	0.12	129	Wish detection in product reviews

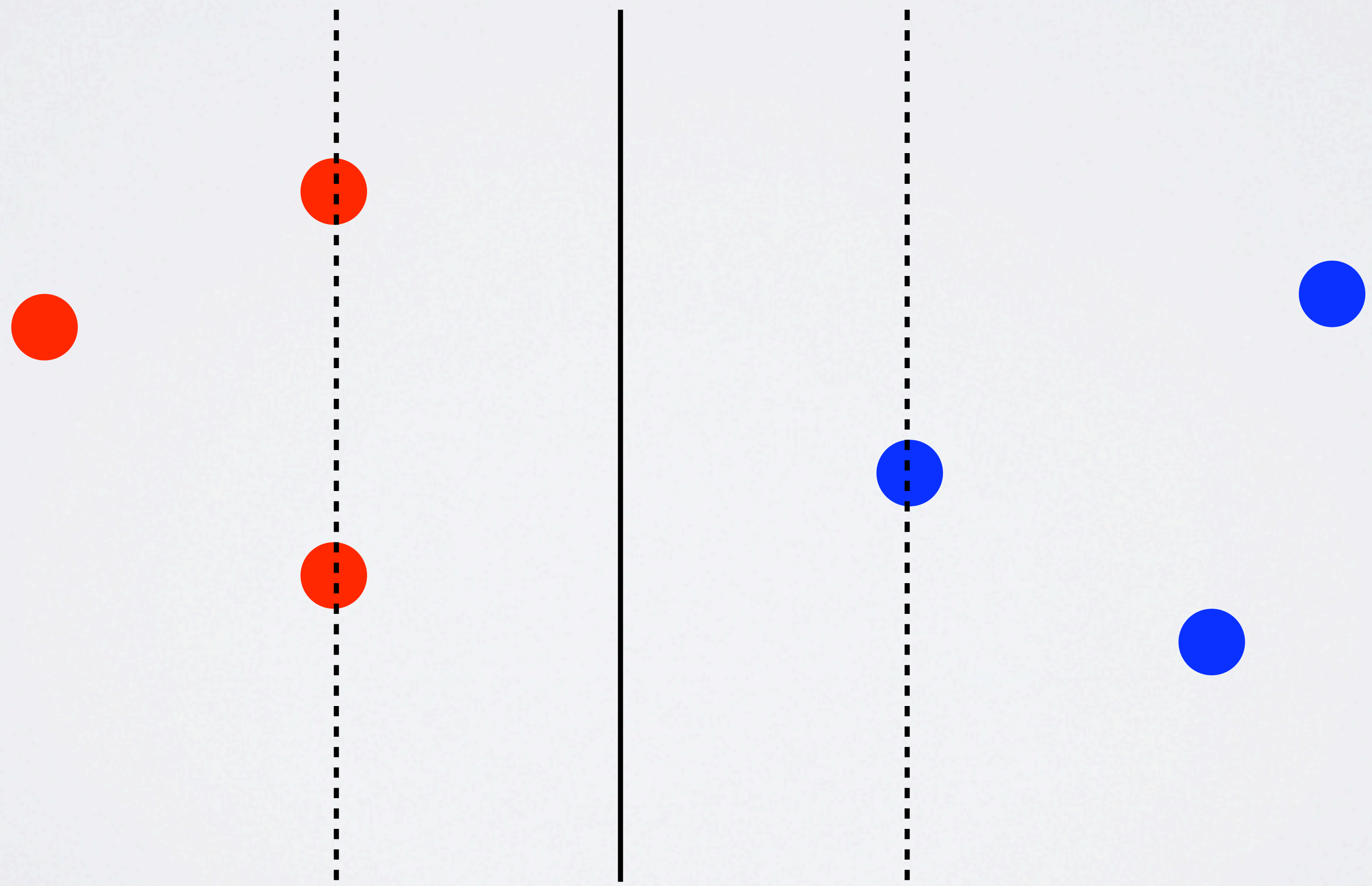
ALGORITHMS

- Linear classifiers only: $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b$
- **Supervised SVM**:
  - ignores the unlabeled data
- **Semi-Supervised SVM** (S^3VM):
  - assumes low density gap between classes
- **Manifold Regularization** (MR):
  - assumes smoothness w.r.t. graph

SUPERVISED SVM

Maximizes margin between decision boundary and labeled data

$$\min_f \frac{1}{2}\|f\|_2^2 + C\sum_{i=1}^{l}\max(0, 1 - y_i f(\mathbf{x}_i))$$

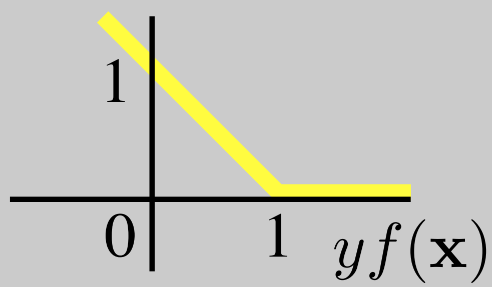


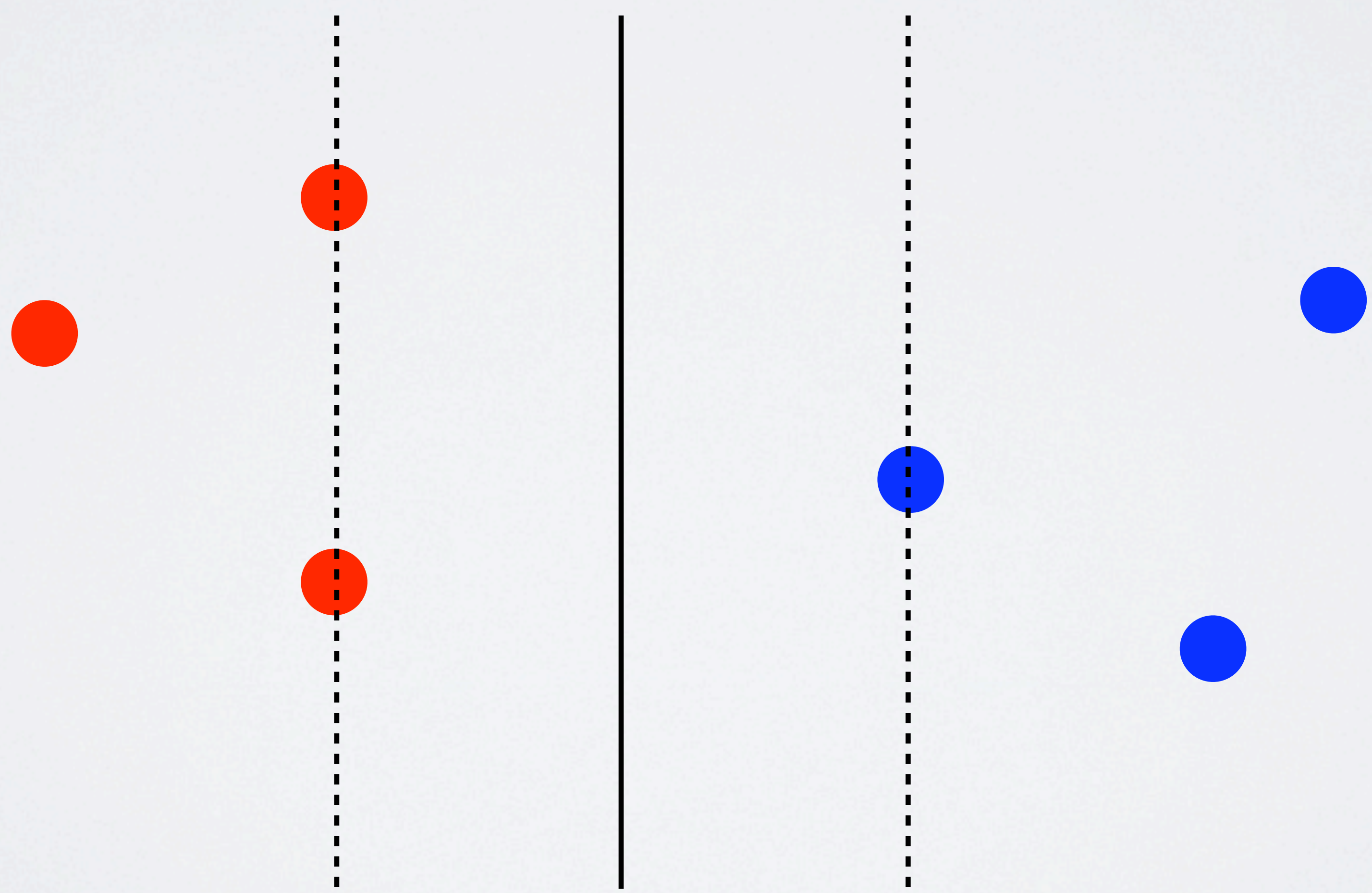
Supervised SVM

Maximizes margin between decision boundary and labeled data

$$\min_f \frac{1}{2}\|f\|_2^2 + C\sum_{i=1}^{l}\max(0, 1 - y_i f(\mathbf{x}_i))$$

Hinge loss



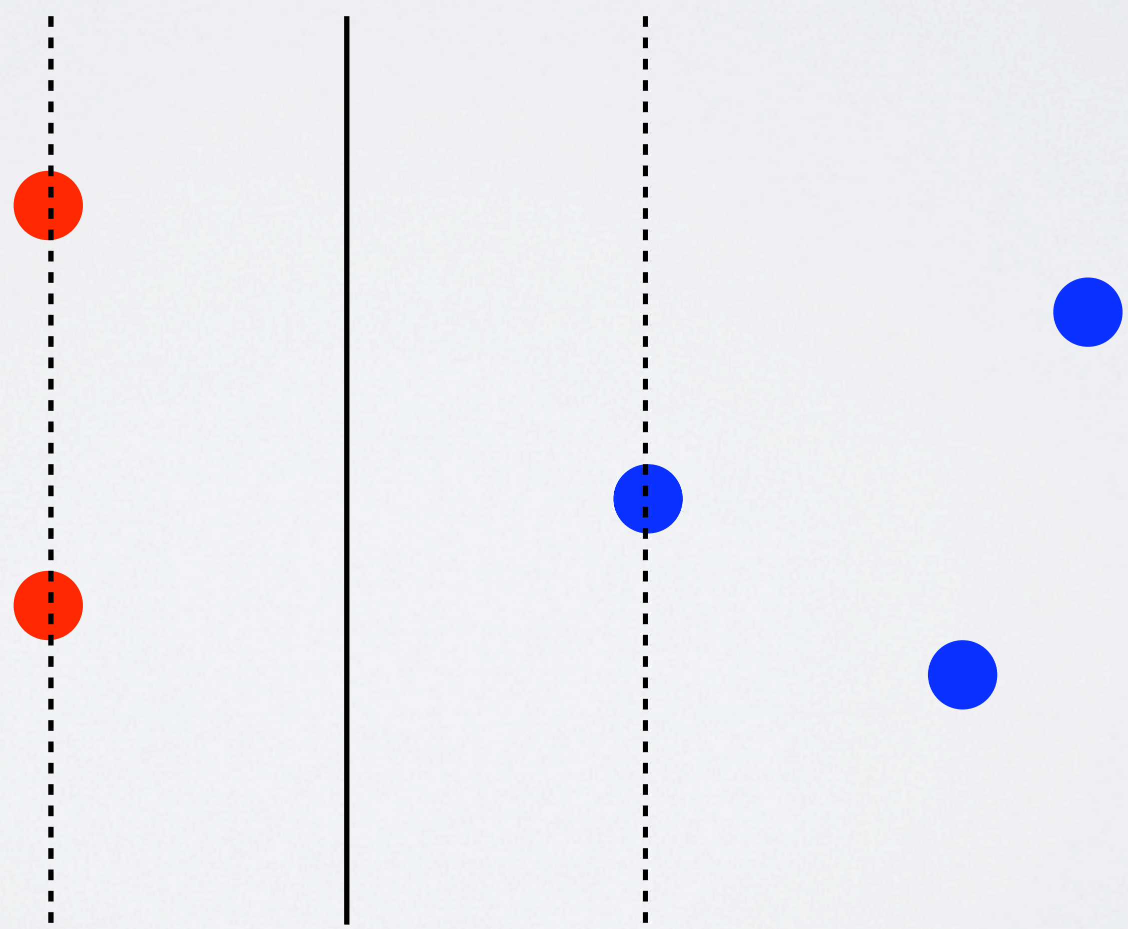


SUPERVISED SVM

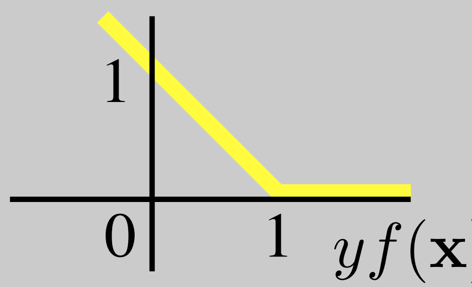
Maximizes margin between decision boundary and labeled data

$$\min_f \frac{1}{2}\|f\|_2^2 + C\sum_{i=1}^{l}\max(0, 1 - y_i f(\mathbf{x}_i))$$

Parameter:
C



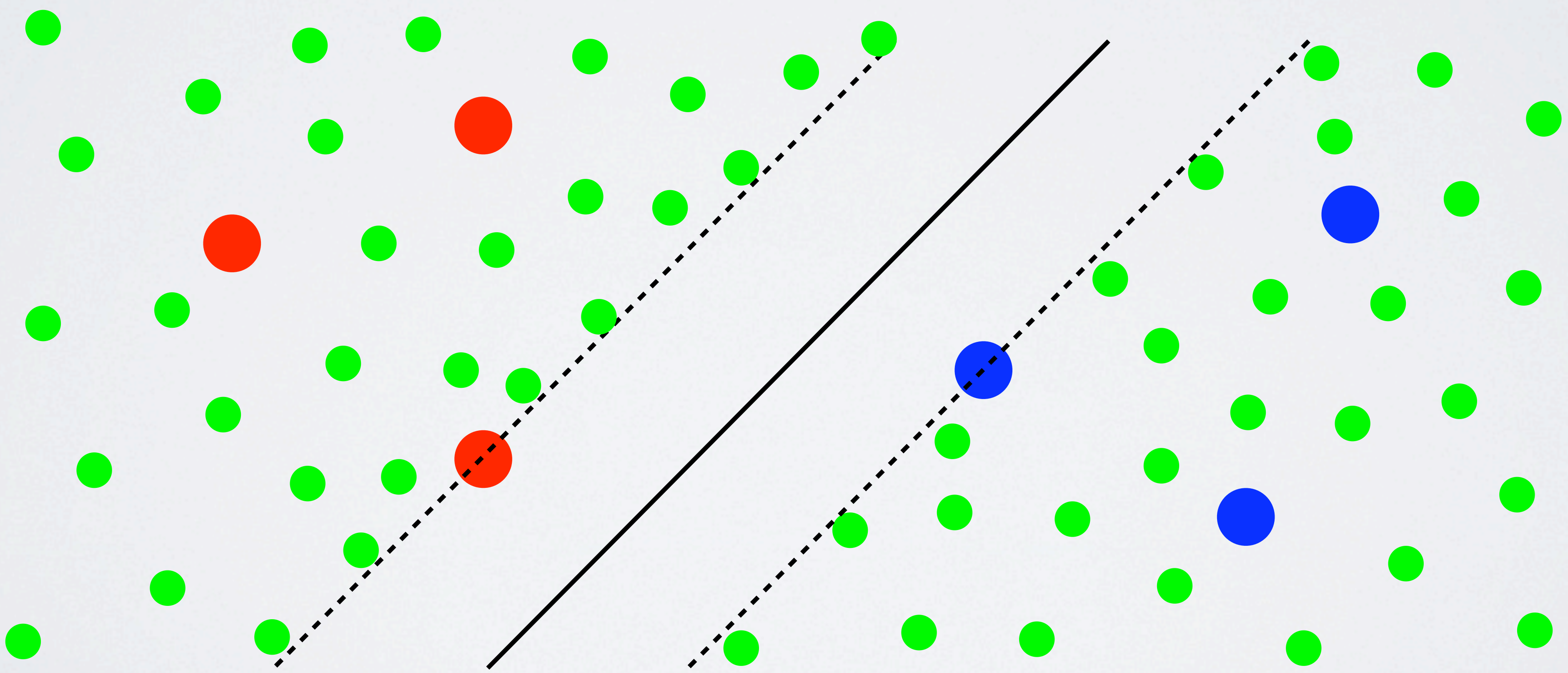
Hinge loss



Semi-Supervised SVM (S3VM)

Places decision boundary in low density region

$$\min_f \frac{\lambda}{2}\|f\|_2^2 + \frac{1}{l}\sum_{i=1}^{l}\max(0, 1 - y_i f(\mathbf{x}_i)) + \frac{\lambda'}{u}\sum_{j=l+1}^{l+u}\max(0, 1 - |f(\mathbf{x}_j)|)$$

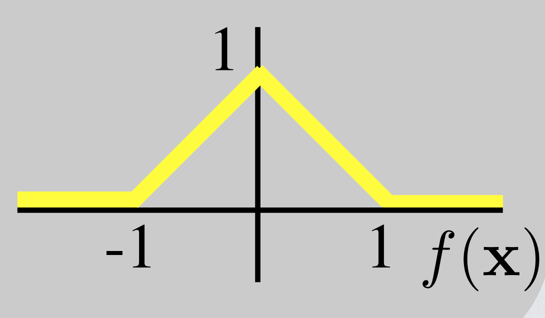


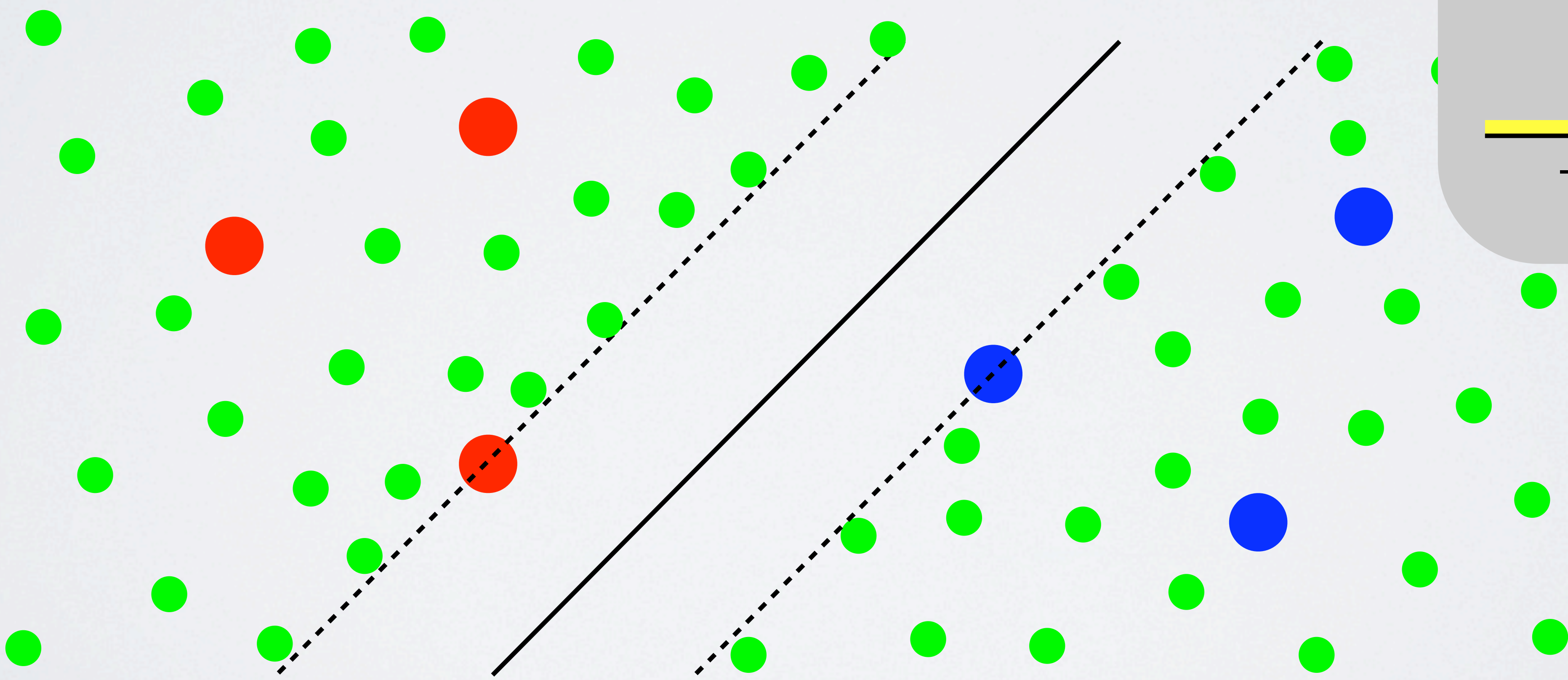
Semi-Supervised SVM (S3VM)

Places decision boundary in low density region

$$\min_f \frac{\lambda}{2}\|f\|_2^2 + \frac{1}{l}\sum_{i=1}^{l}\max(0, 1 - y_i f(\mathbf{x}_i)) + \frac{\lambda'}{u}\sum_{j=l+1}^{l+u}\max(0, 1 - |f(\mathbf{x}_j)|)$$

Hat loss



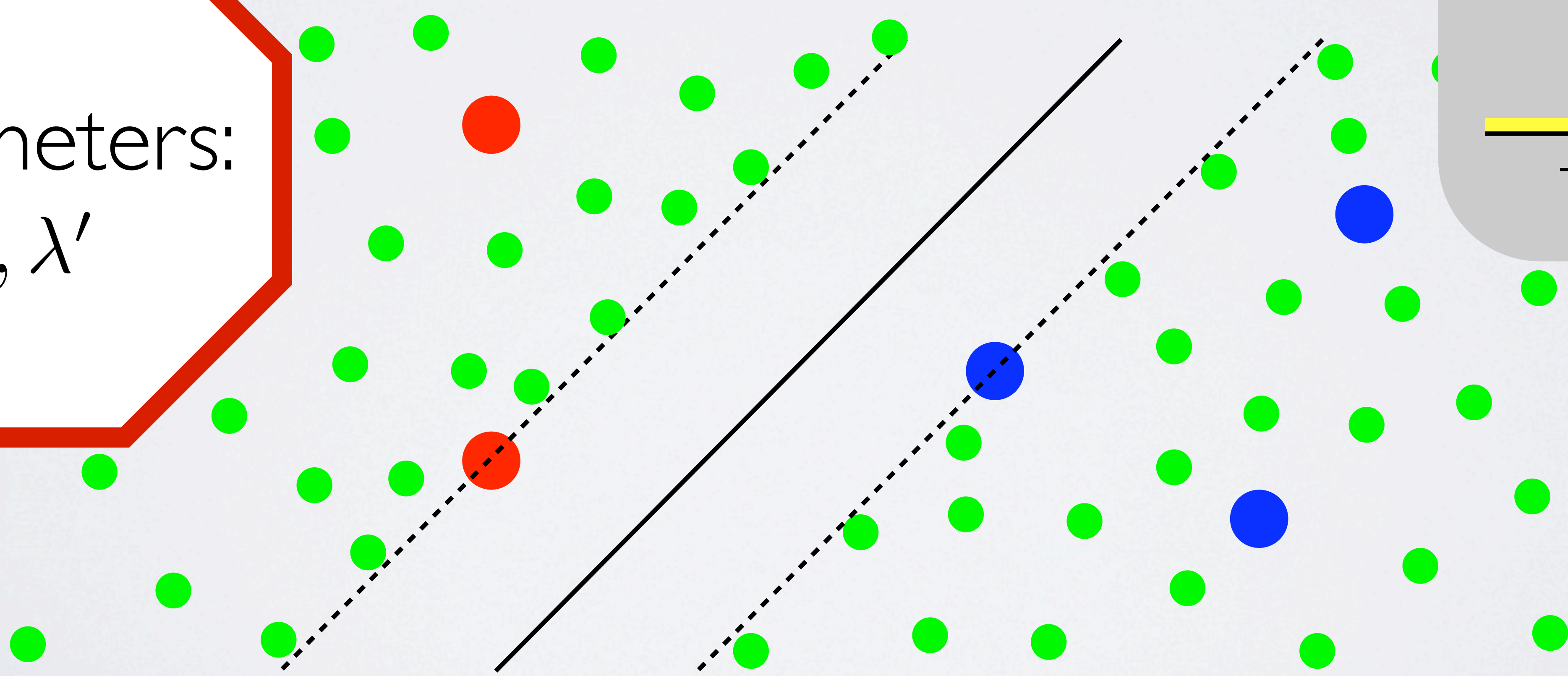


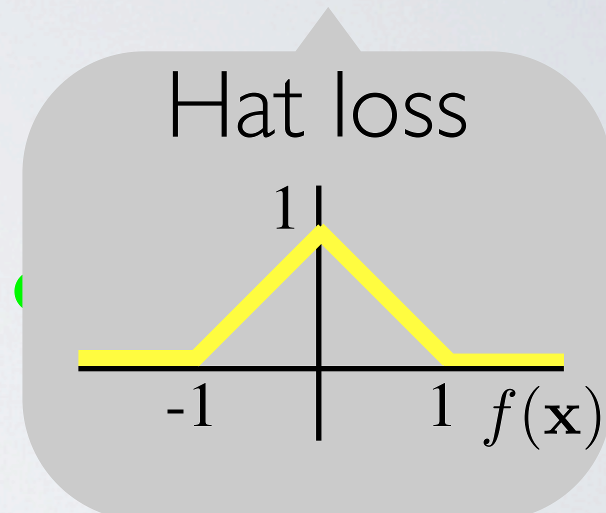
Semi-Supervised SVM (S3VM)

Places decision boundary in low density region

$$\min_f \frac{\lambda}{2}\|f\|_2^2 + \frac{1}{l}\sum_{i=1}^{l}\max(0, 1 - y_i f(\mathbf{x}_i)) + \frac{\lambda'}{u}\sum_{j=l+1}^{l+u}\max(0, 1 - |f(\mathbf{x}_j)|)$$

Parameters:
λ, λ'

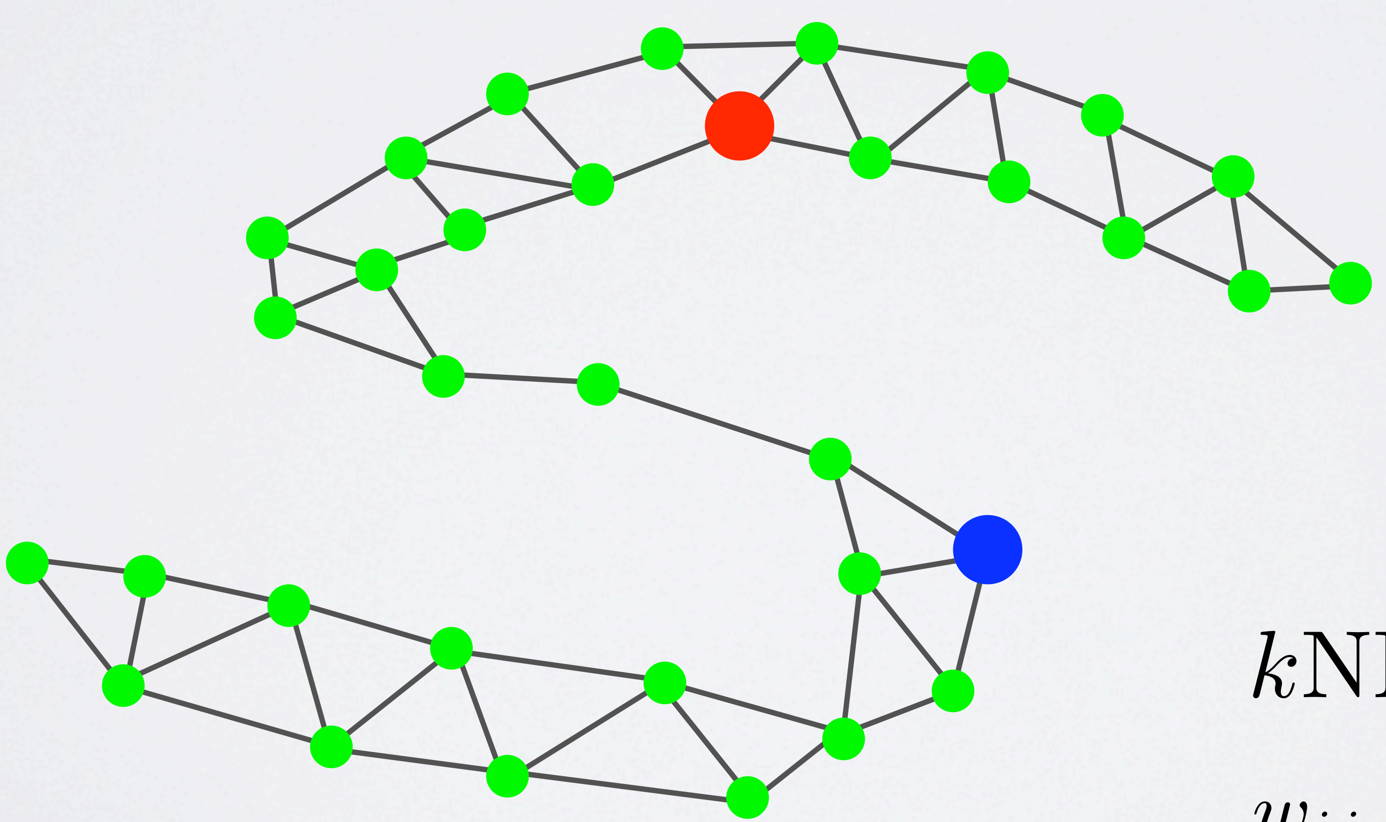




Manifold Regularization (MR)

Assumes smoothness w.r.t. graph over labeled/unlabeled data (similar examples should get similar labels)

$$\min_f \gamma_A \|f\|_2^2 + \frac{1}{l}\sum_{i=1}^{l} V(y_i f(\mathbf{x}_i)) + \gamma_I \sum_{i=1}^{l+u}\sum_{j=1}^{l+u} w_{ij}(f(\mathbf{x}_i) - f(\mathbf{x}_j))^2$$



kNN graph, where

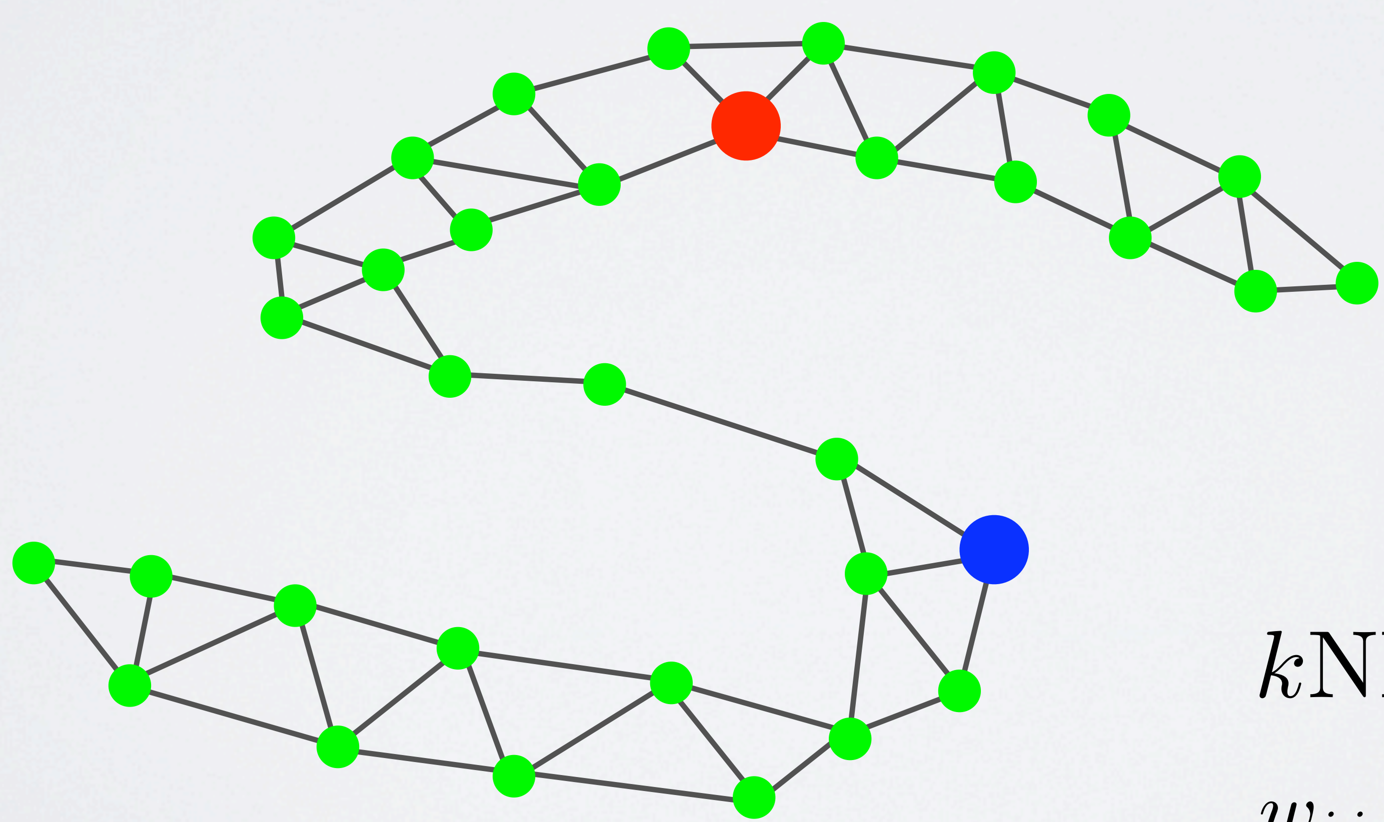
$$w_{ij} = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right)$$

MANIFOLD REGULARIZATION (MR)

Assumes smoothness w.r.t. graph over labeled/unlabeled data
(similar examples should get similar labels)

$$\min_f \gamma_A \|f\|_2^2 + \frac{1}{l}\sum_{i=1}^{l} V(y_i f(\mathbf{x}_i)) + \gamma_I \sum_{i=1}^{l+u}\sum_{j=1}^{l+u} w_{ij}(f(\mathbf{x}_i) - f(\mathbf{x}_j))^2$$

"Unsmoothness" penalty: if w_{ij} is large, $(f(\mathbf{x}_i) - f(\mathbf{x}_j))^2$ should be small.



kNN graph, where

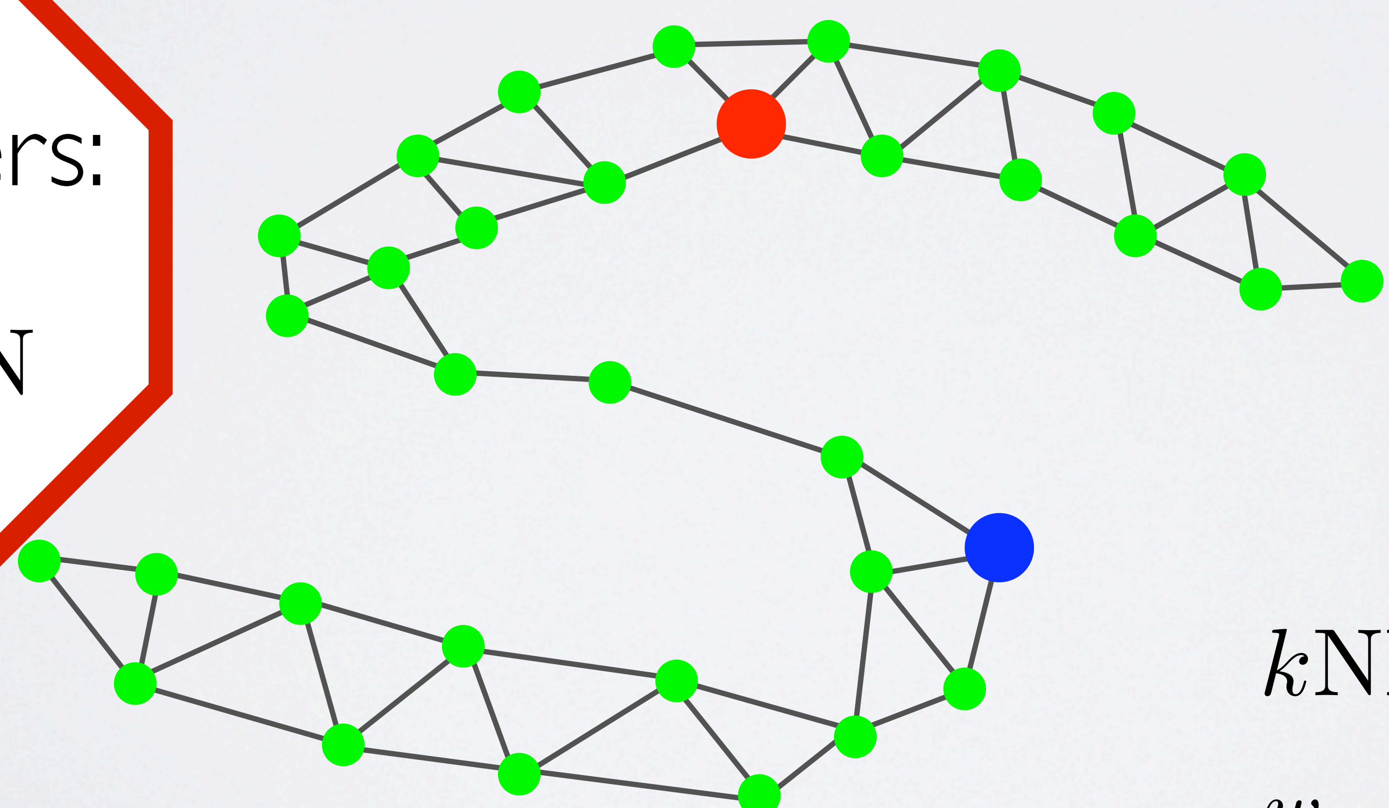
$$w_{ij} = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right)$$

Manifold Regularization (MR)

Assumes smoothness w.r.t. graph over labeled/unlabeled data
(similar examples should get similar labels)

$$\min_f \gamma_A \|f\|_2^2 + \frac{1}{l}\sum_{i=1}^{l} V(y_i f(\mathbf{x}_i)) + \gamma_I \sum_{i=1}^{l+u}\sum_{j=1}^{l+u} w_{ij}(f(\mathbf{x}_i) - f(\mathbf{x}_j))^2$$

Parameters:
γ_A, γ_I
k in kNN
σ



"Unsmoothness" penalty: if w_{ij} is large, $(f(\mathbf{x}_i) - f(\mathbf{x}_j))^2$ should be small.

kNN graph, where

$$w_{ij} = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right)$$

Toward Agnostic SSL

Important question: How can we automatically choose between SL={SVM}, SSL={S3VM, MR}?

Toward Agnostic SSL

Important question: How can we automatically choose between SL={SVM}, SSL={S3VM, MR}?

- Recall our goal of ensuring that unlabeled data doesn't hurt us

Toward Agnostic SSL

Important question: How can we automatically choose between SL={SVM}, SSL={S3VM, MR}?

- Recall our goal of ensuring that unlabeled data doesn't hurt us
- Common view is that model selection with CV is unreliable with little labeled data

Toward Agnostic SSL

Important question: How can we automatically choose between SL={SVM}, SSL={S3VM, MR}?

- Recall our goal of ensuring that unlabeled data doesn't hurt us
- Common view is that model selection with CV is unreliable with little labeled data
- We explicitly tested this hypothesis

Toward Agnostic SSL

Important question: How can we automatically choose between SL={SVM}, SSL={S3VM, MR}?

- Recall our goal of ensuring that unlabeled data doesn't hurt us
- Common view is that model selection with CV is unreliable with little labeled data
- We explicitly tested this hypothesis
- Also use meta-level model selection procedure
  - Select model family as well as member within the family

MODEL SELECTION

Model Selection

Given several algorithms (e.g., SL={SVM}, SSL={S3VM, MR})

Model Selection

Given several algorithms (e.g., SL={SVM}, SSL={S3VM, MR})

1. Tune parameters of each algorithm using 5-fold CV

Model Selection

Given several algorithms (e.g., SL={SVM}, SSL={S3VM, MR})

1. Tune parameters of each algorithm using 5-fold CV

2. Compare best 5-fold average performance across algorithms

Model Selection

Given several algorithms (e.g., SL={SVM}, SSL={S3VM, MR})

1. Tune parameters of each algorithm using 5-fold CV

2. Compare best 5-fold average performance across algorithms

3. Select the algorithm with the best tuning performance (favoring SL if it is tied with any SSL algorithm)

Model Selection

Given several algorithms (e.g., SL={SVM}, SSL={S3VM, MR})

1. Tune parameters of each algorithm using 5-fold CV

2. Compare best 5-fold average performance across algorithms

3. Select the algorithm with the best tuning performance (favoring SL if it is tied with any SSL algorithm)

Note: On a per-trial basis to simulate single real-world training set

Performance Metrics

Three commonly used metrics in NLP

- Accuracy: $\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{[f(\mathbf{x}_i)=y_i]}$

- Maximum F1 value achieved over entire precision-recall curve

- AUROC: area under the ROC curve

Each is used for both parameter tuning and evaluation

Overall Results

		accuracy						maxF1						AUROC						
		$u = 100$			$u = 1000$			$u = 100$			$u = 1000$			$u = 100$			$u = 1000$			
Dataset	l	SVM	S3VM	MR	SVM	S3VM	MR	SVM	S3VM	MR	SVM	S3VM	MR	SVM	S3VM	MR	SVM	S3VM	MR	
[MacWin]	10	0.60	**0.72**	**0.83**	0.60	0.72	**0.86**	**0.66**	**0.67**	**0.67**	**0.66**	**0.67**	**0.67**	**0.63**	**0.69**	**0.67**	**0.63**	**0.69**	**0.69**	Tune
		0.51	0.51	**0.70**	0.51	0.50	**0.69**	0.74	**0.77**	**0.80**	**0.74**	**0.74**	**0.75**	0.72	**0.75**	**0.82**	0.72	**0.71**	**0.80**	Trans
		0.53	0.50	**0.71**	0.53	0.50	**0.68**	0.74	0.75	**0.79**	**0.74**	**0.75**	**0.74**	0.73	**0.72**	**0.83**	**0.73**	**0.71**	**0.76**	Test
	100	0.87	0.87	**0.91**	0.87	0.87	**0.90**	**0.94**	**0.95**	**0.95**	**0.94**	**0.95**	**0.95**	0.96	**0.97**	**0.97**	**0.96**	**0.96**	**0.96**	Tune
		0.89	**0.89**	**0.89**	**0.89**	**0.89**	**0.89**	0.91	**0.93**	**0.92**	**0.91**	**0.90**	**0.90**	**0.97**	**0.97**	**0.96**	**0.97**	**0.97**	0.96	Trans
		0.89	0.89	**0.91**	**0.89**	**0.89**	**0.90**	**0.92**	**0.92**	**0.92**	**0.92**	**0.91**	**0.91**	**0.97**	**0.97**	**0.97**	**0.97**	**0.97**	0.97	Test
[Interest]	10	**0.68**	**0.75**	**0.78**	0.68	**0.75**	**0.79**	0.73	**0.77**	**0.77**	0.73	**0.78**	**0.77**	0.52	**0.66**	**0.66**	0.52	**0.68**	**0.64**	Tune
		0.52	**0.56**	**0.56**	**0.52**	**0.56**	**0.56**	**0.72**	**0.72**	**0.72**	**0.72**	**0.71**	**0.71**	**0.55**	**0.54**	**0.54**	0.55	**0.56**	**0.61**	Trans
		0.52	**0.57**	**0.57**	0.52	**0.57**	**0.58**	**0.68**	**0.69**	**0.69**	**0.68**	**0.69**	**0.69**	**0.58**	**0.56**	**0.61**	0.58	**0.58**	**0.62**	Test
	100	0.77	**0.78**	0.76	0.77	**0.78**	**0.77**	0.84	**0.85**	**0.85**	0.84	**0.85**	0.84	0.89	**0.90**	0.89	**0.89**	**0.85**	**0.84**	Tune
		0.79	**0.79**	0.71	**0.79**	**0.79**	**0.77**	**0.84**	**0.83**	0.82	**0.84**	0.81	0.81	**0.91**	**0.91**	**0.89**	**0.91**	0.79	0.87	Trans
		0.81	**0.80**	0.78	**0.81**	**0.80**	0.79	**0.82**	**0.81**	**0.81**	**0.82**	**0.81**	**0.81**	**0.90**	**0.91**	**0.89**	**0.90**	**0.81**	0.88	Test
[aut-avn]	10	**0.72**	**0.76**	**0.82**	**0.72**	**0.76**	**0.79**	**0.89**	**0.92**	**0.91**	**0.89**	**0.92**	**0.91**	**0.58**	**0.67**	**0.65**	**0.58**	**0.67**	**0.65**	Tune
		0.65	**0.63**	**0.67**	**0.65**	**0.61**	**0.69**	**0.83**	**0.83**	**0.84**	**0.83**	**0.81**	**0.82**	**0.71**	**0.67**	**0.73**	**0.71**	**0.65**	**0.72**	Trans
		0.62	**0.61**	**0.67**	**0.62**	**0.61**	**0.67**	0.80	**0.81**	**0.82**	**0.80**	**0.81**	**0.81**	**0.71**	**0.70**	**0.73**	**0.71**	**0.65**	**0.69**	Test
	100	0.75	0.82	**0.87**	0.75	0.82	**0.86**	0.94	**0.94**	**0.95**	**0.94**	**0.94**	**0.94**	0.93	**0.94**	**0.94**	0.93	**0.94**	0.93	Tune
		0.77	0.79	**0.88**	0.77	0.83	**0.87**	**0.92**	**0.92**	**0.91**	**0.92**	**0.91**	**0.90**	**0.93**	**0.93**	0.91	**0.93**	**0.94**	**0.93**	Trans
		0.77	0.82	**0.89**	0.77	0.83	**0.87**	**0.91**	**0.91**	**0.91**	**0.91**	**0.91**	**0.91**	**0.95**	**0.94**	**0.95**	**0.95**	**0.95**	**0.95**	Test
[real-sim]	10	0.53	0.63	**0.82**	0.53	0.63	**0.78**	**0.65**	**0.66**	**0.66**	**0.65**	**0.66**	**0.65**	**0.77**	**0.81**	**0.81**	**0.77**	**0.81**	**0.77**	Tune
		0.64	0.63	**0.72**	**0.64**	**0.64**	**0.70**	0.57	**0.66**	**0.70**	0.57	**0.62**	**0.56**	0.65	**0.75**	**0.79**	**0.65**	**0.74**	**0.67**	Trans
		0.65	**0.66**	**0.74**	**0.65**	**0.66**	**0.68**	0.53	0.58	**0.63**	0.53	**0.59**	**0.53**	0.64	**0.73**	**0.80**	**0.64**	**0.74**	**0.66**	Test
	100	0.74	0.73	**0.86**	0.74	0.73	**0.84**	0.88	**0.90**	**0.90**	0.88	**0.91**	0.89	0.93	**0.94**	**0.94**	0.93	**0.94**	0.93	Tune
		0.78	0.76	**0.84**	0.78	0.78	**0.85**	**0.81**	**0.83**	**0.79**	**0.81**	**0.81**	**0.81**	**0.94**	**0.93**	**0.91**	**0.94**	**0.94**	0.94	Trans
		0.79	0.78	**0.85**	0.79	0.78	**0.85**	0.78	**0.79**	**0.78**	0.78	**0.79**	0.79	**0.93**	**0.93**	**0.93**	0.93	**0.94**	0.93	Test
[ccat]	10	0.54	0.60	**0.82**	0.54	0.60	**0.81**	**0.84**	**0.85**	**0.85**	**0.84**	**0.85**	**0.84**	**0.74**	**0.78**	**0.78**	**0.74**	**0.78**	**0.74**	Tune
		0.50	0.49	**0.65**	0.50	0.51	**0.67**	**0.69**	**0.69**	**0.73**	**0.69**	0.67	**0.69**	**0.60**	**0.61**	**0.71**	0.60	**0.59**	**0.72**	Trans
		0.49	0.52	**0.64**	0.49	0.52	**0.66**	0.66	0.66	**0.69**	**0.66**	**0.67**	**0.67**	**0.61**	**0.63**	**0.72**	**0.61**	**0.59**	**0.71**	Test
	100	0.80	0.80	**0.84**	0.80	0.80	**0.84**	**0.89**	**0.89**	**0.90**	**0.89**	**0.89**	**0.89**	**0.91**	**0.92**	**0.92**	**0.91**	**0.92**	**0.91**	Tune
		0.80	**0.79**	**0.80**	**0.80**	0.81	**0.83**	**0.83**	**0.85**	**0.84**	**0.83**	**0.82**	**0.82**	**0.91**	0.91	0.89	**0.91**	**0.90**	**0.91**	Trans
		0.81	**0.80**	**0.81**	**0.81**	**0.80**	**0.82**	**0.80**	**0.81**	**0.81**	**0.80**	**0.81**	**0.81**	**0.90**	0.90	**0.90**	**0.90**	**0.90**	**0.90**	Test
[gcat]	10	0.74	**0.83**	**0.82**	**0.74**	**0.79**	**0.81**	**0.44**	**0.47**	**0.46**	**0.44**	**0.47**	**0.46**	**0.69**	**0.79**	**0.75**	**0.69**	**0.79**	**0.75**	Tune
		0.69	**0.68**	**0.75**	**0.69**	**0.72**	**0.76**	0.60	**0.62**	**0.69**	**0.60**	**0.59**	**0.62**	0.71	**0.73**	**0.82**	**0.71**	**0.69**	**0.76**	Trans
		0.66	**0.67**	**0.73**	**0.66**	**0.71**	**0.74**	0.58	**0.61**	**0.66**	**0.58**	**0.60**	**0.59**	0.69	**0.69**	**0.81**	**0.69**	**0.69**	**0.75**	Test
	100	0.77	0.77	**0.90**	0.77	0.77	**0.91**	**0.92**	**0.92**	**0.93**	**0.92**	**0.92**	**0.92**	**0.97**	0.96	**0.97**	**0.97**	**0.96**	**0.96**	Tune
		0.81	0.80	**0.89**	0.81	0.81	**0.90**	**0.88**	**0.88**	0.84	**0.88**	**0.86**	**0.85**	**0.96**	**0.97**	**0.95**	**0.96**	**0.96**	**0.96**	Trans
		0.80	0.80	**0.89**	0.80	0.80	**0.90**	**0.86**	**0.86**	**0.85**	**0.86**	**0.86**	**0.86**	**0.96**	**0.96**	**0.96**	**0.96**	**0.96**	**0.96**	Test
[WISH-politics]	10	**0.70**	**0.77**	**0.79**	**0.70**	**0.77**	**0.82**	**0.61**	**0.62**	**0.61**	**0.61**	**0.62**	**0.61**	**0.74**	**0.78**	**0.74**	**0.74**	**0.78**	**0.76**	Tune
		0.50	**0.56**	**0.63**	**0.50**	**0.62**	**0.56**	**0.58**	**0.58**	**0.61**	**0.58**	**0.55**	**0.53**	0.62	**0.62**	**0.69**	**0.62**	**0.62**	**0.61**	Trans
		0.52	**0.56**	**0.60**	**0.52**	**0.62**	**0.53**	0.52	**0.53**	**0.53**	0.52	**0.54**	**0.52**	**0.57**	**0.58**	**0.61**	**0.57**	**0.62**	**0.60**	Test
	100	**0.75**	**0.75**	**0.75**	**0.75**	**0.75**	**0.74**	0.74	**0.75**	**0.76**	**0.74**	**0.75**	**0.75**	**0.79**	**0.80**	**0.80**	**0.79**	**0.80**	**0.80**	Tune
		0.73	**0.73**	**0.71**	**0.73**	**0.73**	**0.70**	**0.65**	**0.66**	**0.67**	**0.65**	**0.64**	**0.64**	**0.76**	**0.74**	**0.75**	**0.76**	**0.75**	**0.76**	Trans
		0.75	**0.75**	0.72	**0.75**	**0.75**	0.71	**0.64**	**0.63**	**0.63**	**0.64**	**0.63**	**0.64**	**0.78**	**0.76**	**0.77**	**0.78**	**0.76**	**0.77**	Test
[WISH-products]	10	**0.89**	**0.89**	0.67	**0.89**	**0.89**	0.67	0.19	**0.22**	0.16	0.19	**0.22**	0.16	**0.76**	**0.80**	**0.74**	**0.76**	**0.80**	**0.74**	Tune
		0.87	**0.87**	0.66	**0.87**	**0.87**	0.61	**0.31**	**0.29**	**0.32**	**0.31**	0.24	0.25	**0.56**	0.52	**0.58**	**0.56**	**0.54**	**0.56**	Trans
		0.90	**0.90**	0.67	**0.90**	**0.90**	0.61	0.22	0.23	**0.30**	0.22	**0.24**	**0.27**	0.50	0.53	**0.62**	0.50	**0.54**	**0.59**	Test
	100	**0.90**	**0.90**	0.82	**0.90**	**0.90**	0.81	0.49	**0.50**	**0.54**	0.49	**0.52**	**0.52**	0.73	0.73	**0.77**	**0.73**	**0.78**	**0.75**	Tune
		0.88	**0.88**	0.81	**0.88**	**0.88**	0.80	**0.34**	0.28	**0.37**	**0.34**	**0.27**	**0.30**	**0.60**	**0.55**	**0.57**	**0.60**	**0.57**	**0.61**	Trans
		0.90	**0.90**	0.79	**0.90**	**0.91**	0.76	**0.33**	**0.28**	**0.33**	**0.33**	**0.32**	**0.38**	**0.59**	**0.56**	**0.60**	**0.59**	**0.56**	**0.60**	Test

OVERALL RESULTS

		accuracy						maxF1						AUROC						
		$u = 100$			$u = 1000$			$u = 100$			$u = 1000$			$u = 100$			$u = 1000$			
Dataset	l	SVM	S3VM	MR	SVM	S3VM	MR	SVM	S3VM	MR	SVM	S3VM	MR	SVM	S3VM	MR	SVM	S3VM	MR	
[MacWin]	10	0.60	0.72	0.83	0.60	0.72	0.86	0.66	0.67	0.67	0.66	0.67	0.67	0.63	0.69	0.67	0.63	0.69	0.69	Tune
		0.51	0.51	0.70	0.51	0.50	0.69	0.74	0.77	0.80	0.74	0.74	0.75	0.72	0.75	0.82	0.72	0.71	0.80	Trans
		0.53	0.50	0.71	0.53	0.50	0.68	0.74	0.75	0.79	0.74	0.75	0.74	0.73	0.72	0.83	0.73	0.71	0.76	Test
	100	0.87	0.87	0.91	0.87	0.87	0.90	0.94	0.95	0.95	0.94	0.95	0.95	0.96	0.97	0.97	0.96	0.96	0.96	Tune
		0.89	0.89	0.89	0.89	0.89	0.89	0.91	0.93	0.92	0.91	0.90	0.90	0.97	0.97	0.96	0.97	0.97	0.96	Trans
		0.89	0.89	0.91	0.89	0.89	0.90	0.92	0.92	0.92	0.92	0.91	0.91	0.97	0.97	0.97	0.97	0.97	0.97	Test
[Interest]	10	0.68	0.75	0.78	0.68	0.75	0.79	0.73	0.77	0.77	0.73	0.78	0.77	0.52	0.66	0.66	0.52	0.68	0.64	Tune
		0.52	0.56	0.56	0.52	0.56	0.56	0.72	0.72	0.72	0.72	0.71	0.71	0.55	0.54	0.54	0.55	0.56	0.61	Trans
		0.52	0.57	0.57	0.52	0.57	0.58	0.68	0.69	0.69	0.68	0.69	0.69	0.58	0.56	0.61	0.58	0.58	0.62	Test
	100	0.77	0.78	0.76	0.77	0.78	0.77	0.84	0.85	0.85	0.84	0.85	0.84	0.89	0.90	0.89	0.89	0.85	0.84	Tune
		0.79	0.79	0.71	0.79	0.79	0.77	0.84	0.83	0.82	0.84	0.81	0.81	0.91	0.91	0.89	0.91	0.79	0.87	Trans
		0.81	0.80	0.78	0.81	0.80	0.79	0.82	0.81	0.81	0.82	0.81	0.81	0.90	0.91	0.89	0.90	0.81	0.88	Test
[aut-avn]	10	0.72	0.76	0.82	0.72	0.76	0.79	0.89	0.92	0.91	0.89	0.92	0.91	0.58	0.67	0.65	0.58	0.67	0.65	Tune
		0.65	0.63	0.67	0.65	0.61	0.69	0.83	0.83	0.84	0.83	0.81	0.82	0.71	0.67	0.73	0.71	0.65	0.72	Trans
		0.62	0.61	0.67	0.62	0.61	0.67	0.80	0.81	0.82	0.80	0.81	0.81	0.71	0.70	0.73	0.71	0.65	0.69	Test
	100	0.75	0.82	0.87	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.94	0.93	0.94	0.93	Tune
		0.77	0.79	0.88	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.91	0.93	0.94	0.93	Trans
		0.77	0.82	0.89	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.95	0.95	0.95	0.95	Test
[real-sim]	10	0.53	0.63	0.82	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.81	0.77	0.81	0.77	Tune
		0.64	0.63	0.72	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.79	0.65	0.74	0.67	Trans
		0.65	0.66	0.74	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.80	0.64	0.74	0.66	Test
	100	0.74	0.73	0.86	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.94	0.93	0.94	0.93	Tune
		0.78	0.76	0.84	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.91	0.94	0.94	0.94	Trans
		0.79	0.78	0.85	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.93	0.93	0.94	0.93	Test
[ccat]	10	0.54	0.60	0.82	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.78	0.74	0.78	0.74	Tune
		0.50	0.49	0.65	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.71	0.60	0.59	0.72	Trans
		0.49	0.52	0.64	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.72	0.61	0.59	0.71	Test
	100	0.80	0.80	0.84	0.80	[illegible]	[illegible]	0.89	0.89	0.90	0.89	0.89	0.89	0.91	0.92	0.92	0.91	0.92	0.91	Tune
		0.80	0.79	0.80	0.80	[illegible]	0.83	0.83	0.85	0.84	0.83	0.82	0.82	0.91	0.91	0.89	0.91	0.90	0.91	Trans
		0.81	0.80	0.81	[illegible]	0.80	0.82	0.80	0.81	0.81	0.80	0.81	0.81	0.90	0.90	0.90	0.90	0.90	0.90	Test
[gcat]	10	0.74	0.83	0.82	0.74	0.79	0.81	0.44	0.47	0.46	0.44	0.47	0.46	0.69	0.79	0.75	0.69	0.79	0.75	Tune
		0.69	0.68	0.75	0.69	0.72	0.76	0.60	0.62	0.69	0.60	0.59	0.62	0.71	0.73	0.82	0.71	0.69	0.76	Trans
		0.66	0.67	0.73	0.66	0.71	0.74	0.58	0.61	0.66	0.58	0.60	0.59	0.69	0.69	0.81	0.69	0.69	0.75	Test
	100	0.77	0.77	0.90	0.77	0.77	0.91	0.92	0.92	0.93	0.92	0.92	0.92	0.97	0.96	0.97	0.97	0.96	0.96	Tune
		0.81	0.80	0.89	0.81	0.81	0.90	0.88	0.88	0.84	0.88	0.86	0.85	0.96	0.97	0.95	0.96	0.96	0.96	Trans
		0.80	0.80	0.89	0.80	0.80	0.90	0.86	0.86	0.85	0.86	0.86	0.86	0.96	0.96	0.96	0.96	0.96	0.96	Test
[WISH-politics]	10	0.70	0.77	0.79	0.70	0.77	0.82	0.61	0.62	0.61	0.61	0.62	0.61	0.74	0.78	0.74	0.74	0.78	0.76	Tune
		0.50	0.56	0.63	0.50	0.62	0.56	0.58	0.58	0.61	0.58	0.55	0.53	0.62	0.62	0.69	0.62	0.62	0.61	Trans
		0.52	0.56	0.60	0.52	0.62	0.53	0.52	0.53	0.53	0.52	0.54	0.52	0.57	0.58	0.61	0.57	0.62	0.60	Test
	100	0.75	0.75	0.75	0.75	0.75	0.74	0.74	0.75	0.76	0.74	0.75	0.75	0.79	0.80	0.80	0.79	0.80	0.80	Tune
		0.73	0.73	0.71	0.73	0.73	0.70	0.65	0.66	0.67	0.65	0.64	0.64	0.76	0.74	0.75	0.76	0.75	0.76	Trans
		0.75	0.75	0.72	0.75	0.75	0.71	0.64	0.63	0.63	0.64	0.63	0.64	0.78	0.76	0.77	0.78	0.76	0.77	Test
[WISH-products]	10	0.89	0.89	0.67	0.89	0.89	0.67	0.19	0.22	0.16	0.19	0.22	0.16	0.76	0.80	0.74	0.76	0.80	0.74	Tune
		0.87	0.87	0.66	0.87	0.87	0.61	0.31	0.29	0.32	0.31	0.24	0.25	0.56	0.52	0.58	0.56	0.54	0.56	Trans
		0.90	0.90	0.67	0.90	0.90	0.61	0.22	0.23	0.30	0.22	0.24	0.27	0.50	0.53	0.62	0.50	0.54	0.59	Test
	100	0.90	0.90	0.82	0.90	0.90	0.81	0.49	0.50	0.54	0.49	0.52	0.52	0.73	0.73	0.77	0.73	0.78	0.75	Tune
		0.88	0.88	0.81	0.88	0.88	0.80	0.34	0.28	0.37	0.34	0.27	0.30	0.60	0.55	0.57	0.60	0.57	0.61	Trans
		0.90	0.90	0.79	0.90	0.91	0.76	0.33	0.28	0.33	0.33	0.32	0.38	0.59	0.56	0.60	0.59	0.56	0.60	Test

Just kidding...

Observations

Observations

- No algorithm is universally superior

Observations

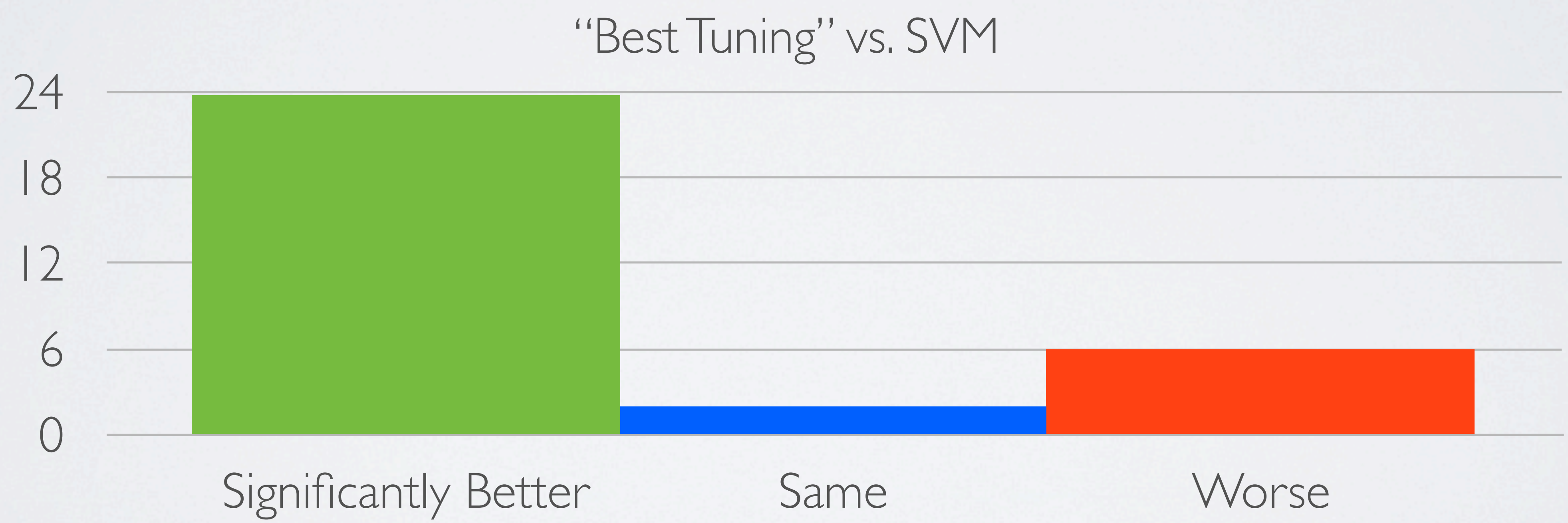
- No algorithm is universally superior
- Each of the SSL algorithms can be *significantly worse* than SL

Observations

- No algorithm is universally superior
- Each of the SSL algorithms can be *significantly worse* than SL
- **Tuning with accuracy as the metric is valid for SSL model selection**

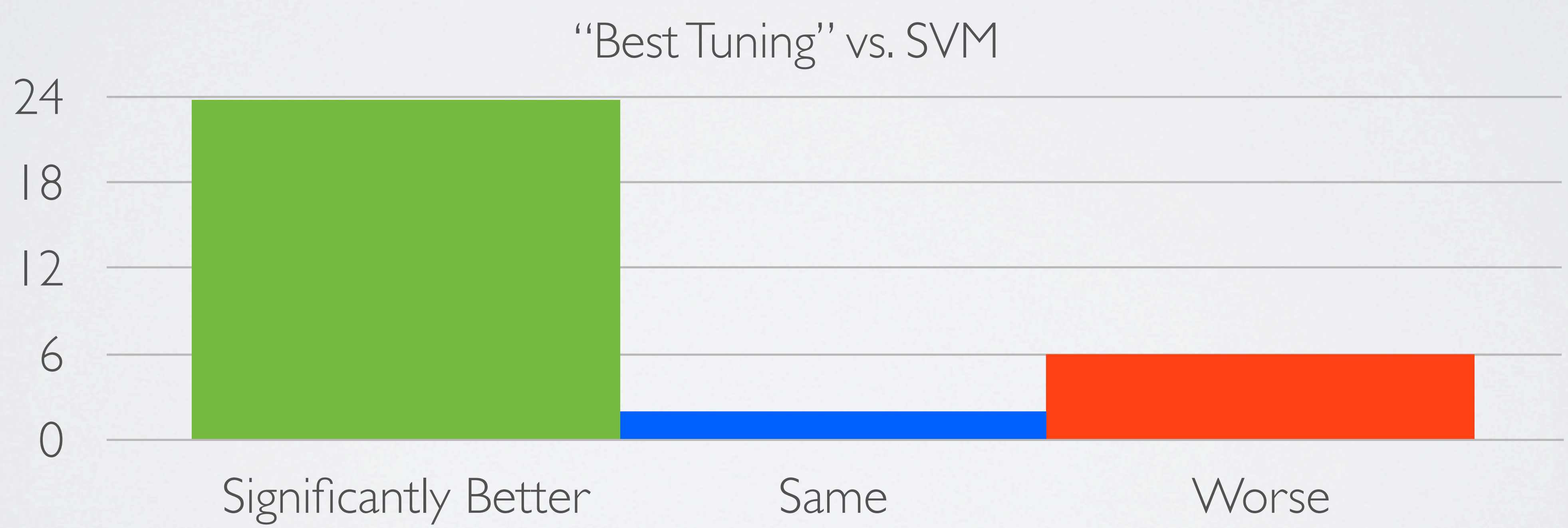
Observations

- No algorithm is universally superior
- Each of the SSL algorithms can be *significantly worse* than SL
- **Tuning with accuracy as the metric is valid for SSL model selection**
  - Out of 32 settings (8 data sets x 4 labeled/unlabeled sizes):



Observations

- No algorithm is universally superior
- Each of the SSL algorithms can be *significantly worse* than SL
- **Tuning with accuracy as the metric is valid for SSL model selection**
  - Out of 32 settings (8 data sets x 4 labeled/unlabeled sizes):



- Tuning with maxF1 or AUROC as the metric is less reliable

Aggregate Results

Compared relative performance *across all data sets* in terms of:

1. #trials where each method is worse/same/better than SVM
2. overall average test performance

Aggregate Results

(#trials worse than SVM, #trials equal to SVM, #trials better than SVM) out of 80 trials (10 trials x 8 data sets) per *l*/*u* setting

		$u = 100$			$u = 1000$		
Metric	l	S3VM	MR	Best Tuning	S3VM	MR	Best Tuning
accuracy	10	$(14, 27, 39)$	$(27, 0, 53)$	$(8, 31, 41)$	$(14, 25, 41)$	$(27, 0, 53)$	$(8, 29, 43)$
	100	$(27, 7, 46)$	$(38, 0, 42)$	$(20, 16, 44)$	$(27, 6, 47)$	$(37, 0, 43)$	$(16, 19, 45)$

Metric	l	S3VM	MR	Best Tuning	S3VM	MR	Best Tuning
maxF1	10	$(29, 2, 49)$	$(16, 1, 63)$	$(14, 55, 11)$	$(27, 0, 53)$	$(24, 0, 56)$	$(13, 53, 14)$
	100	$(39, 0, 41)$	$(34, 4, 42)$	$(31, 15, 34)$	$(39, 1, 40)$	$(44, 4, 32)$	$(26, 21, 33)$

Metric	l	S3VM	MR	Best Tuning	S3VM	MR	Best Tuning
AUROC	10	$(26, 0, 54)$	$(11, 0, 69)$	$(12, 57, 11)$	$(25, 0, 55)$	$(25, 0, 55)$	$(11, 56, 13)$
	100	$(43, 0, 37)$	$(37, 0, 43)$	$(38, 8, 34)$	$(38, 0, 42)$	$(46, 0, 34)$	$(28, 24, 28)$

Aggregate Results

(#trials worse than SVM, #trials equal to SVM, #trials better than SVM) out of 80 trials (10 trials x 8 data sets) per *l*/*u* setting

		$u = 100$			$u = 1000$		
Metric	l	S3VM	MR	Best Tuning	S3VM	MR	Best Tuning
accuracy	10	$(14, 27, 39)$	$(27, 0, 53)$	$(8, 31, 41)$	$(14, 25, 41)$	$(27, 0, 53)$	$(8, 29, 43)$
	100	$(27, 7, 46)$	$(38, 0, 42)$	$(20, 16, 44)$	$(27, 6, 47)$	$(37, 0, 43)$	$(16, 19, 45)$

Metric	l	S3VM	MR	Best Tuning	S3VM	MR	Best Tuning
maxF1	10	$(29, 2, 49)$	$(16, 1, 63)$	$(14, 55, 11)$	$(27, 0, 53)$	$(24, 0, 56)$	$(13, 53, 14)$
	100	$(39, 0, 41)$	$(34, 4, 42)$	$(31, 15, 34)$	$(39, 1, 40)$	$(44, 4, 32)$	$(26, 21, 33)$

Metric	l	S3VM	MR	Best Tuning	S3VM	MR	Best Tuning
AUROC	10	$(26, 0, 54)$	$(11, 0, 69)$	$(12, 57, 11)$	$(25, 0, 55)$	$(25, 0, 55)$	$(11, 56, 13)$
	100	$(43, 0, 37)$	$(37, 0, 43)$	$(38, 8, 34)$	$(38, 0, 42)$	$(46, 0, 34)$	$(28, 24, 28)$

Aggregate Results

(#trials worse than SVM, #trials equal to SVM, #trials better than SVM) out of 80 trials (10 trials x 8 data sets) per *l*/*u* setting

		$u = 100$			$u = 1000$		
Metric	l	S3VM	MR	Best Tuning	S3VM	MR	Best Tuning
accuracy	10	(14, 27, 39)	(27, 0, 53)	(8, 31, 41)	(14, 25, 41)	(27, 0, 53)	(8, 29, 43)
	100	(27, 7, 46)	(38, 0, 42)	(20, 16, 44)	(27, 6, 47)	(37, 0, 43)	(16, 19, 45)

Metric	l	S3VM	MR	Best Tuning	S3VM	MR	Best Tuning
maxF1	10	(29, 2, 49)	(16, 1, 63)	(14, 55, 11)	(27, 0, 53)	(24, 0, 56)	(13, 53, 14)
	100	(39, 0, 41)	(34, 4, 42)	(31, 15, 34)	(39, 1, 40)	(44, 4, 32)	(26, 21, 33)

Best Tuning
(11, 56, 13)
(28, 24, 28)

CV using accuracy and maxF1 mitigates some risk in applying SSL: worse than SVM in fewer trials

AGGREGATE RESULTS

(#trials worse than SVM, #trials e
out of 80 trials (10 trials x 8 data

Even with only 10 labeled points!

		$u = 100$			$u = 1000$		
Metric	l	S3VM	MR	Best Tuning	S3VM	MR	Best Tuning
accuracy	10	(14, 27, 39)	(27, 0, 53)	(8, 31, 41)	(14, 25, 41)	(27, 0, 53)	(8, 29, 43)
	100	(27, 7, 46)	(38, 0, 42)	(20, 16, 44)	(27, 6, 47)	(37, 0, 43)	(16, 19, 45)

Metric	l	S3VM	MR	Best Tuning	S3VM	MR	Best Tuning
maxF1	10	(29, 2, 49)	(16, 1, 63)	(14, 55, 11)	(27, 0, 53)	(24, 0, 56)	(13, 53, 14)
	100	(39, 0, 41)	(34, 4, 42)	(31, 15, 34)	(39, 1, 40)	(44, 4, 32)	(26, 21, 33)

Best Tuning
(11, 56, 13)
(28, 24, 28)

CV using accuracy and maxF1 mitigates some risk in applying SSL: worse than SVM in fewer trials

Aggregate Results

(#trials worse than SVM, #trials e
out of 80 trials (10 trials × 8 data

Even with only 10 labeled points!

Metric	l	$u = 100$ S3VM	MR	Best Tuning	$u = 1000$ S3VM	MR	Best Tuning
accuracy	10	(14, 27, 39)	(27, 0, 53)	(8, 31, 41)	(14, 25, 41)	(27, 0, 53)	(8, 29, 43)
	100	(27, 7, 46)	(38, 0, 42)	(20, 16, 44)	(27, 6, 47)	(37, 0, 43)	(16, 19, 45)

Metric	l	S3VM	MR	Best Tuning	S3VM	MR	Best Tuning
maxF1	10	(29, 2, 49)	(16, 1, 63)	(14, 55, 11)	(27, 0, 53)	(24, 0, 56)	(13, 53, 14)
	100	(39, 0, 41)	(34, 4, 42)	(31, 15, 34)	(39, 1, 40)	(44, 4, 32)	(26, 21, 33)

Best Tuning
(11, 56, 13)
(28, 24, 28)

CV using accuracy and maxF1 mitigates some risk in applying SSL: worse than SVM in fewer trials

But...due to conservative tie-breaking strategy, outperforms SVM in fewer trials as well

AGGREGATE RESULTS

(#trials worse than SVM, #trials equal to SVM, #trials better than SVM) out of 80 trials (10 trials x 8 data sets) per *l*/*u* setting

		$u = 100$			$u = 1000$		
Metric	l	S3VM	MR	Best Tuning	S3VM	MR	Best Tuning
accuracy	10	(14, 27, 39)	(27, 0, 53)	(8, 31, 41)	(14, 25, 41)	(27, 0, 53)	(8, 29, 43)
	100	(27, 7, 46)	(38, 0, 42)	(20, 16, 44)	(27, 6, 47)	(37, 0, 43)	(16, 19, 45)

Metric	l	S3VM	MR	Best Tuning	S3VM	MR	Best Tuning
maxF1	10	(29, 2, 49)	(16, 1, 63)	(14, 55, 11)	(27, 0, 53)	(24, 0, 56)	(13, 53, 14)
	100	(39, 0, 41)	(34, 4, 42)	(31, 15, 34)	(39, 1, 40)	(44, 4, 32)	(26, 21, 33)

Metric	l	S3VM	MR	Best Tuning	S3VM	MR	Best Tuning
AUROC	10	(26, 0, 54)	(11, 0, 69)	(12, 57, 11)	(25, 0, 55)	(25, 0, 55)	(11, 56, 13)
	100	(43, 0, 37)	(37, 0, 43)	(38, 8, 34)	(38, 0, 42)	(46, 0, 34)	(28, 24, 28)

AGGREGATE RESULTS

(#trials worse than SVM, #trials equal to SVM, #trials better than SVM) out of 80 trials (10 trials x 8 data sets) per *l*/*u* setting

		$u = 100$			$u = 1000$		
Metric	l	S3VM	MR	Best Tuning	S3VM	MR	Best Tuning
accuracy	10	(14, 27, 39)	(27, 0, 53)	(8, 31, 41)	(14, 25, 41)	(27, 0, 53)	(8, 29, 43)
	100	(27, 7, 46)	(38, 0, 42)	(20, 16, 44)	(27, 6, 47)	(37, 0, 43)	(16, 19, 45)

Metric	l	S3VM	MR	Best Tuning	S3VM	MR	Best Tuning
maxF1	10	(29, 2, 49)	(16, 1, 63)	(14, 55, 11)	(27, 0, 53)	(24, 0, 56)	(13, 53, 14)
	100	(39, 0, 41)	(34, 4, 42)	(31, 15, 34)	(39, 1, 40)	(44, 4, 32)	(26, 21, 33)

Metric	l	S3VM	MR	Best Tuning	S3VM	MR	Best Tuning
AUROC	10	(26, 0, 54)	(11, 0, 69)	(12, 57, 11)	(25, 0, 55)	(25, 0, 55)	(11, 56, 13)
	100	(43, 0, 37)	(37, 0, 43)	(38, 8, 34)	(38, 0, 42)	(46, 0, 34)	(28, 24, 28)

AUROC as the performance metric is less reliable

AGGREGATE RESULTS

Average test performance over the 80 runs in each setting:

Metric	l	$u = 100$ SVM	S3VM	MR	Best Tuning	$u = 1000$ SVM	S3VM	MR	Best Tuning
accuracy	10	0.61	0.62	0.67	0.68	0.61	0.63	0.64	0.67
	100	0.81	0.82	0.83	0.85	0.81	0.82	0.83	0.85

Metric	l	SVM	S3VM	MR	Best Tuning	SVM	S3VM	MR	Best Tuning
maxF1	10	0.59	0.61	0.64	0.59	0.59	0.61	0.61	0.59
	100	0.76	0.75	0.76	0.75	0.76	0.76	0.76	0.76

Metric	l	SVM	S3VM	MR	Best Tuning	SVM	S3VM	MR	Best Tuning
AUROC	10	0.63	0.64	0.72	0.61	0.63	0.64	0.67	0.61
	100	0.87	0.87	0.87	0.87	0.87	0.86	0.87	0.86

Aggregate Results

CV with accuracy metric: better than any single model due to per-trial selection strategy

Average test perfor

		$u = 100$				$u = 1000$			
Metric	l	SVM	S3VM	MR	Best Tuning	SVM	S3VM	MR	Best Tuning
accuracy	10	0.61	0.62	0.67	0.68	0.61	0.63	0.64	0.67
	100	0.81	0.82	0.83	0.85	0.81	0.82	0.83	0.85

Metric	l	SVM	S3VM	MR	Best Tuning	SVM	S3VM	MR	Best Tuning
maxF1	10	0.59	0.61	0.64	0.59	0.59	0.61	0.61	0.59
	100	0.76	0.75	0.76	0.75	0.76	0.76	0.76	0.76

Metric	l	SVM	S3VM	MR	Best Tuning	SVM	S3VM	MR	Best Tuning
AUROC	10	0.63	0.64	0.72	0.61	0.63	0.64	0.67	0.61
	100	0.87	0.87	0.87	0.87	0.87	0.86	0.87	0.86

Aggregate Results

Average test perfor

CV with accuracy metric: better than any single model due to per-trial selection strategy

		$u = 100$				$u = 1000$			
Metric	l	SVM	S3VM	MR	Best Tuning	SVM	S3VM	MR	Best Tuning
accuracy	10	0.61	0.62	0.67	0.68	0.61	0.63	0.64	0.67
	100	0.81	0.82	0.83	0.85	0.81	0.82	0.83	0.85

Metric	l	SVM	S3VM	MR	Best Tuning	SVM	S3VM	MR	Best Tuning
maxF1	10	0.59	0.61	0.64	0.59	0.59	0.61	0.61	0.59
	100	0.76	0.75	0.76	0.75	0.76	0.76	0.76	0.76

Metric	l	SVM	S3VM	MR	Best Tuning	SVM	S3VM	MR	Best Tuning
AUROC	10	0.63	0.64	0.72	0.61	0.63	0.64	0.67	0.61
	100	0.87	0.87	0.87	0.87	0.87	0.86	0.87	0.86

Mixed results based on maxF1
Poor results based on AUROC

Take-Home Message

Model selection + cross validation + accuracy metric = agnostic SSL with as few as 10 labeled points!

Take-Home Message

Model selection + cross validation + accuracy metric = agnostic SSL with as few as 10 labeled points!

Future Work:

- Expand empirical study to more data sets and algorithms
- Extend beyond binary classification tasks
- More sophisticated model selection techniques

TAKE-HOME MESSAGE

Model selection + cross validation + accuracy metric = agnostic SSL with as few as 10 labeled points!

Future Work:

- Expand empirical study to more data sets and algorithms
- Extend beyond binary classification tasks
- More sophisticated model selection techniques

Thank you! Questions?

EXTRA SLIDES

REALSSL PROCEDURE

Input: dataset $D_{labeled} = \{x_i, y_i\}_{i=1}^{l}$, $D_{unlabeled} = \{x_j\}_{j=1}^{u}$, $algorithm$, performance $metric$

Randomly partition $D_{labeled}$ into 5 equally-sized disjoint subsets $\{D_{l1}, D_{l2}, D_{l3}, D_{l4}, D_{l5}\}$.
Randomly partition $D_{unlabeled}$ into 5 equally-sized disjoint subsets $\{D_{u1}, D_{u2}, D_{u3}, D_{u4}, D_{u5}\}$.
Combine partitions: Let $D_{fold\ k} = D_{lk} \cup D_{uk}$ for all $k = 1, \dots, 5$.
foreach *parameter configuration in grid* **do**
    foreach *fold k* **do**
        Train model using $algorithm$ on $\cup_{i \neq k} D_{fold\ i}$.
        Evaluate $metric$ on $D_{fold\ k}$.
    end
    Compute the average $metric$ value across the 5 folds.
end
Choose parameter configuration that optimizes average $metric$.
Train model using $algorithm$ and the chosen parameters on $D_{labeled}$ and $D_{unlabeled}$.

Output: Optimal model; Average $metric$ value achieved by optimal parameters during tuning.

REALSSL PROCEDURE

Input: dataset $D_{labeled} = \{x_i, y_i\}_{i=1}^{l}$, $D_{unlabeled} = \{x_j\}_{j=1}^{u}$, $algorithm$, performance $metric$

Randomly partition $D_{labeled}$ into 5 equally-sized disjoint subsets $\{D_{l1}, D_{l2}, D_{l3}, D_{l4}, D_{l5}\}$.
Randomly partition $D_{unlabeled}$ into 5 equally-sized disjoint subsets $\{D_{u1}, D_{u2}, D_{u3}, D_{u4}, D_{u5}\}$.
Combine partitions: Let $D_{fold\ k} = D_{lk} \cup D_{uk}$ for all $k = 1, \ldots, 5$.
foreach *parameter configuration in grid* **do**
  foreach *fold k* **do**
    Train model using $algorithm$ on $\cup_{i \neq k} D_{fold\ i}$.
    Evaluate $metric$ on $D_{fold\ k}$.
  end
  Compute the average $metric$ value across the 5 folds.
end
Choose parameter configuration that optimizes average $metric$.
Train model using $algorithm$ and the chosen parameters on $D_{labeled}$ and $D_{unlabeled}$.

Output: Optimal model; Average $metric$ value achieved by optimal parameters during tuning.

5-fold cross validation over parameter grid; Folds maintain labeled/ unlabeled proportion

Empirical Study Protocol

Input: dataset $D = \{x_i, y_i\}_{i=1}^{n}$, $algorithm$, performance $metric$, set L, set U, trials T
Randomly divide D into D_{pool} (of size $\max(L) + \max(U)$) and D_{test} (the rest).
foreach l *in* L **do**
  foreach u *in* U **do**
    foreach *trial 1 up to* T **do**
      Randomly select $D_{labeled} = \{x_j, y_j\}_{j=l}^{l}$ and $D_{unlabeled} = \{x_k\}_{k=1}^{u}$ from D_{pool}.
      Run RealSSL($D_{labeled}, D_{unlabeled}, algorithm, metric$) to obtain model and tuning performance value (see Algorithm 1).
      Use model to classify $D_{unlabeled}$ and record transductive $metric$ value.
      Use model to classify D_{test} and record test $metric$ value.
    end
  end
end
Output: Tuning, transductive, and test performance for T runs of $algorithm$ using all l and u combinations.

EMPIRICAL STUDY PROTOCOL

Repeat each labeled and unlabeled size for 10 trials;
Tune parameters and build model using RealSSL

Input: dataset I
Randomly divid
foreach l *in* L **d**
  foreach u *in* U **do**
    foreach *trial 1 up to* T **do**
      Randomly select $D_{labeled} = \{x_j, y_j\}_{j=l}^{l}$ and $D_{unlabeled} = \{x_k\}_{k=1}^{u}$ from D_{pool}.
      Run RealSSL($D_{labeled}, D_{unlabeled}, algorithm, metric$) to obtain model and tuning performance value (see Algorithm 1).
      Use model to classify $D_{unlabeled}$ and record transductive $metric$ value.
      Use model to classify D_{test} and record test $metric$ value.
    end
  end
end

Output: Tuning, transductive, and test performance for T runs of $algorithm$ using all l and u combinations.