安全半监督学习：算法和理论分析初探

李宇峰

软件新技术国家重点实验室

南京大学

http://lamda.nju.edu.cn/liyf/

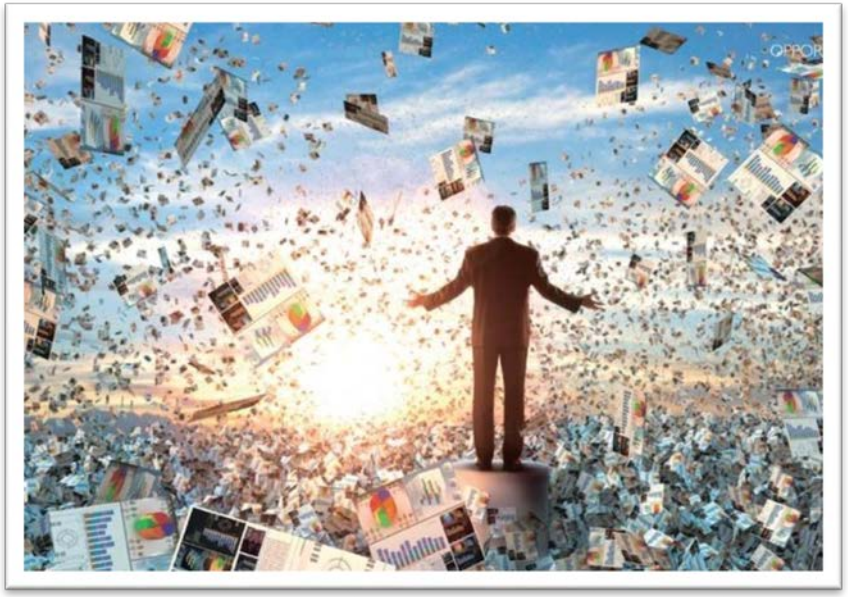
liyf@nju.edu.cn

LAMDA
Learning And Mining from DatA
http://lamda.nju.edu.cn

Big Data

Five “V”

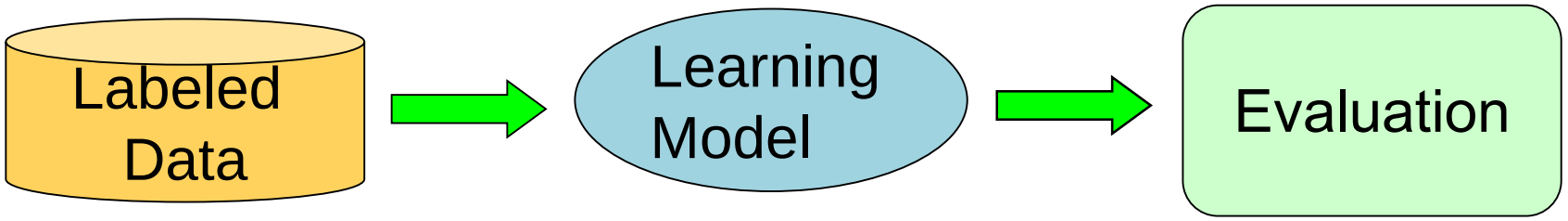
- Volume
- Velocity
- Variety
- Veracity
- Value



Big data analytics is the core of big data processing

ML is among the core techniques of big data analytics

Supervised Learning

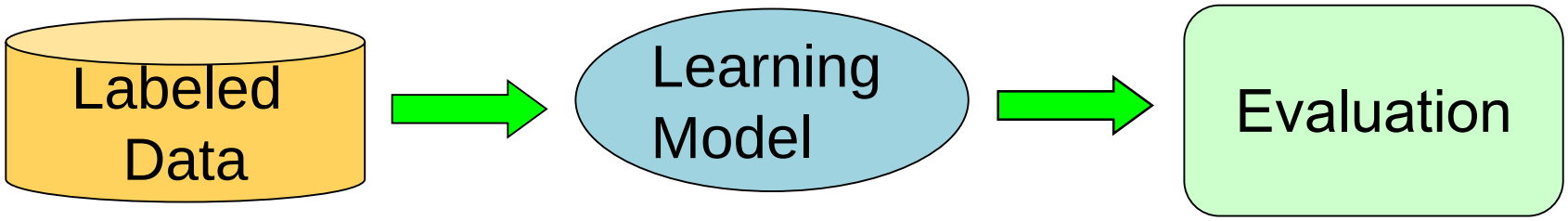


Big labeled data narrows the gap between generalization risk and empirical risk, bringing opportunity to achieve more accurate performance

e.g.,

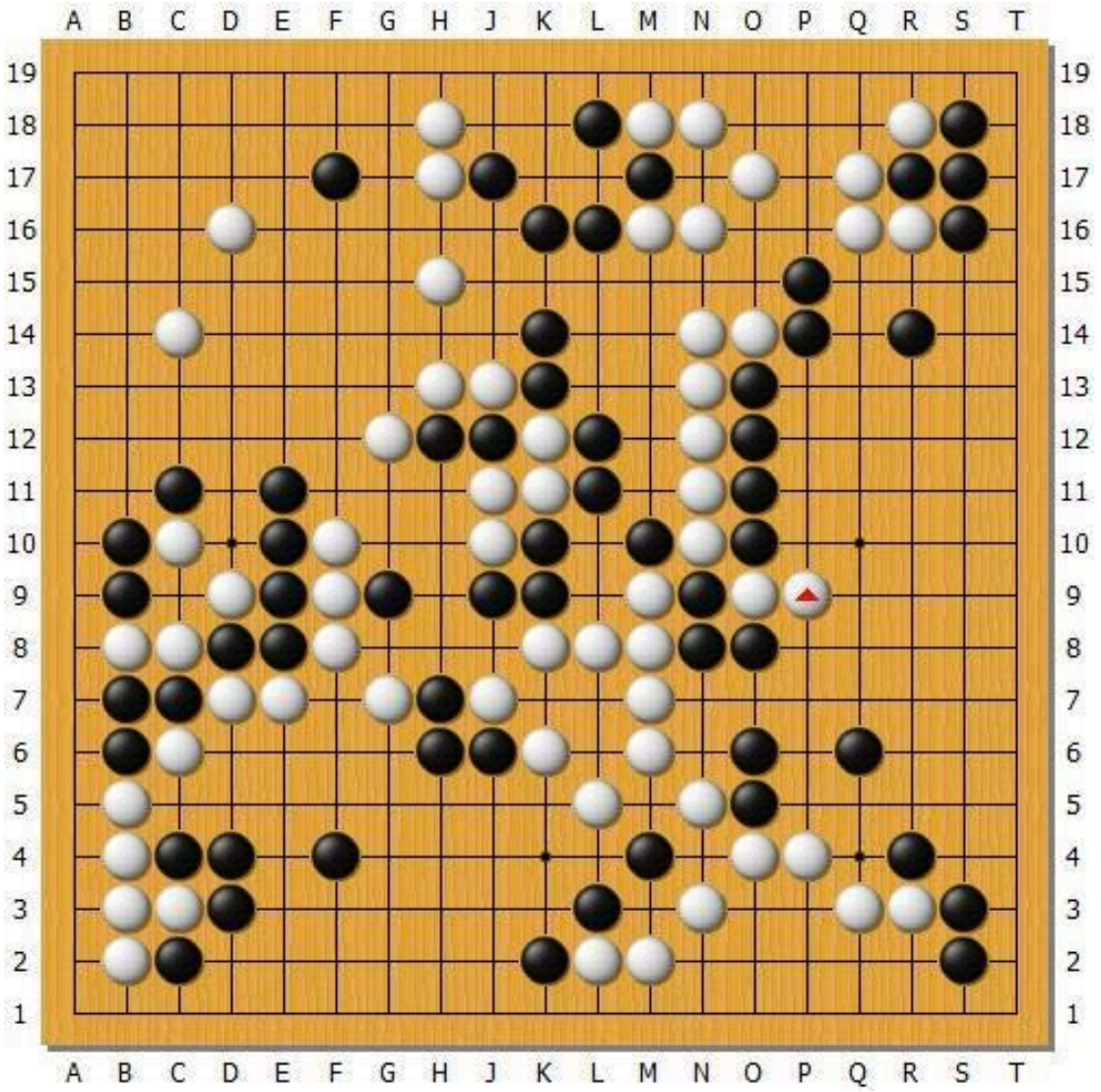
- Image classification
- Speech recognition
- Natural language processing

Supervised Learning



In order to have a good generalization performance, learning model often requires that a large amount of labeled data are available.

Labeled Data Remains Expensive

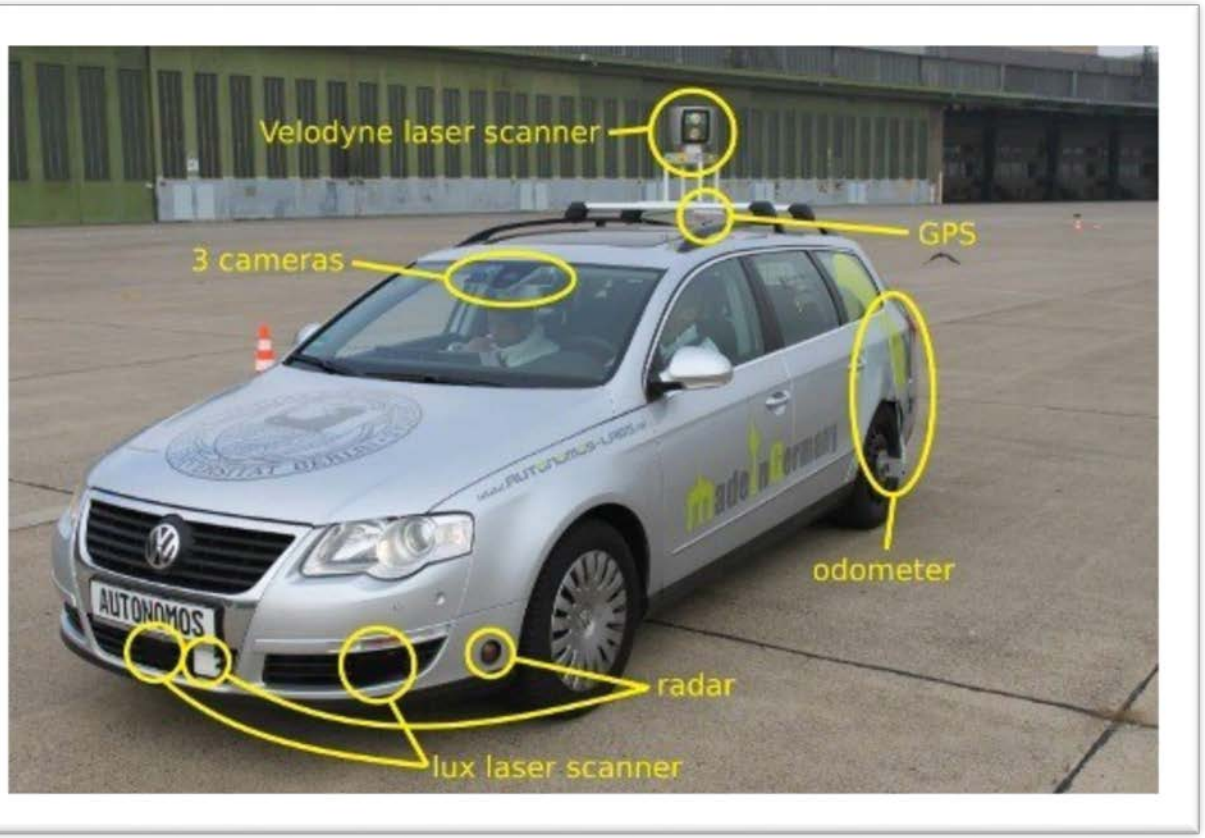


AlphaGo

Go/human expert requires to spend considerable energy to judge certain go game trend

Labeled Data Remains Expensive

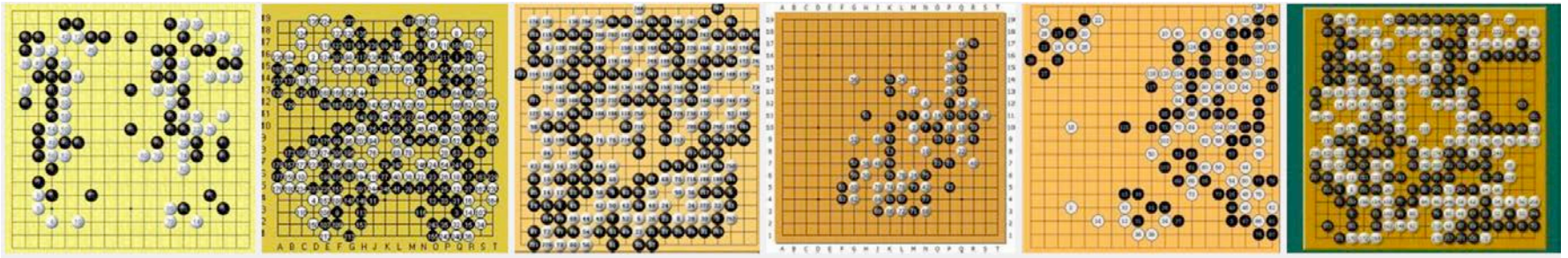
Google Autonomous Cars



Human needs to judge whether the collected sensor data is related to drive pattern or not

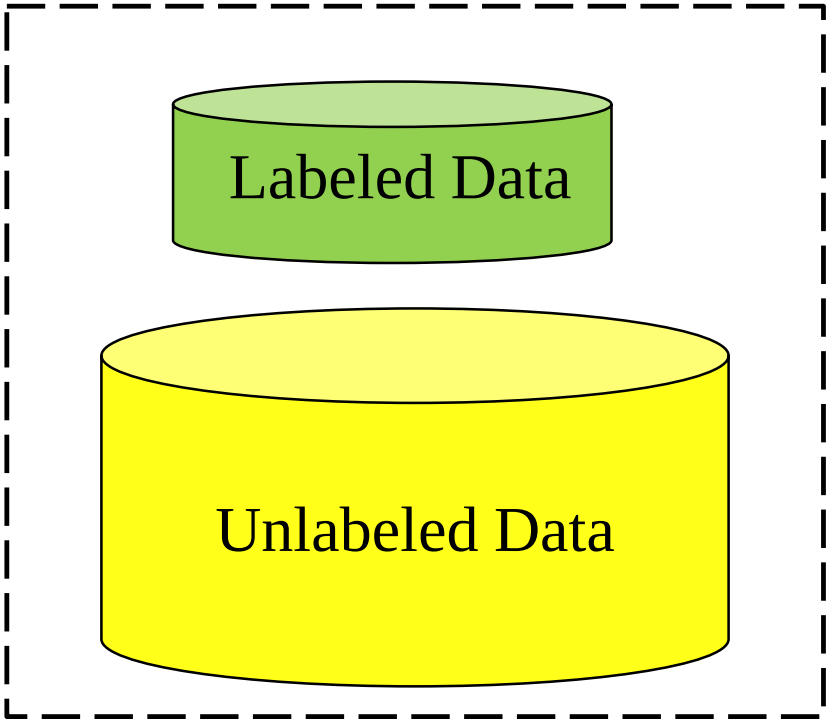
Unlabeled Data are Ubiquitous

- It is easy to collect/generate unlabeled games



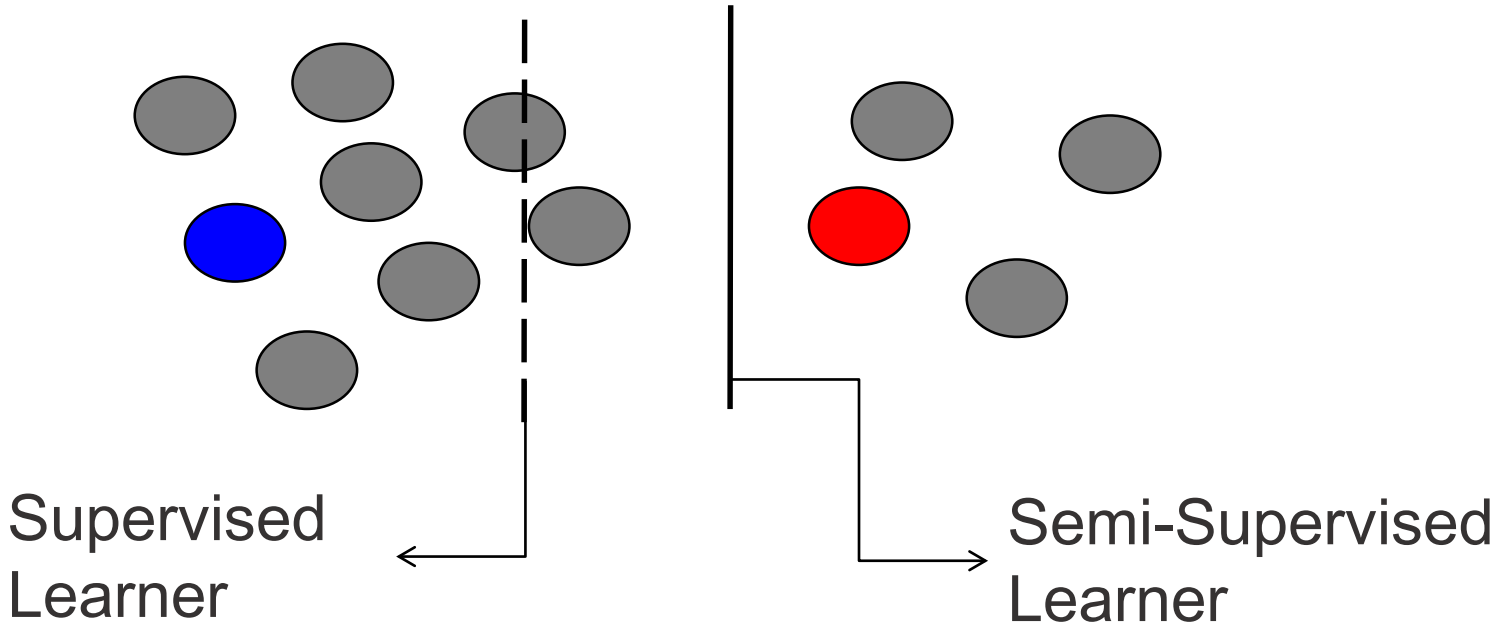
- Similarly, one can easily collect many unlabeled sensor data

Goal of SSL



When labeled data is scarce, how to exploit big unlabeled data to help machine learning models improve the performance

Why unlabeled data helps



unlabeled data helps when its distribution is beneficial to derive the distribution of determinative boundary

Four Popular Paradigms

- **Generative models** [B.M Shahshahani & D.A. Landgrebe, TGRS94; D.J. Miller & H.S. Uyar, NIPS96; etc.]
- **Disagreement-based methods** [Blum & Mitchell, ICML98; Balcan et al., NIPS05; Zhou & Li, TKDE10; etc.]
- **Graph-based methods** [Blum & Chawla, ICML01; Zhu et al.,ICML03; Zhou et al., NIPS05; Belkin et al., JMLR06; etc.]
- **Semi-Supervised SVMs** [Vapnik, STL98; Bennett & Demiriz, NIPS99; Joachims, ICML99; Chapelle & Zien, ICML05; etc.]

Recent Efforts

- Representation Learning + SSL
  - SSL Embedding [Waston et al., 2012]
  - Generative model [Kingma et al., NIPS2014]
  - Graph Embedding [Yang et al., ICML2016]
  - GCN [Kipf and Welling, ICLR2017]
  - …

- Basic idea：
  - SSL does not work well, because representation is not good => representation helps improve SSL
  - Provide another way to help improve performance

How SSL works in practice

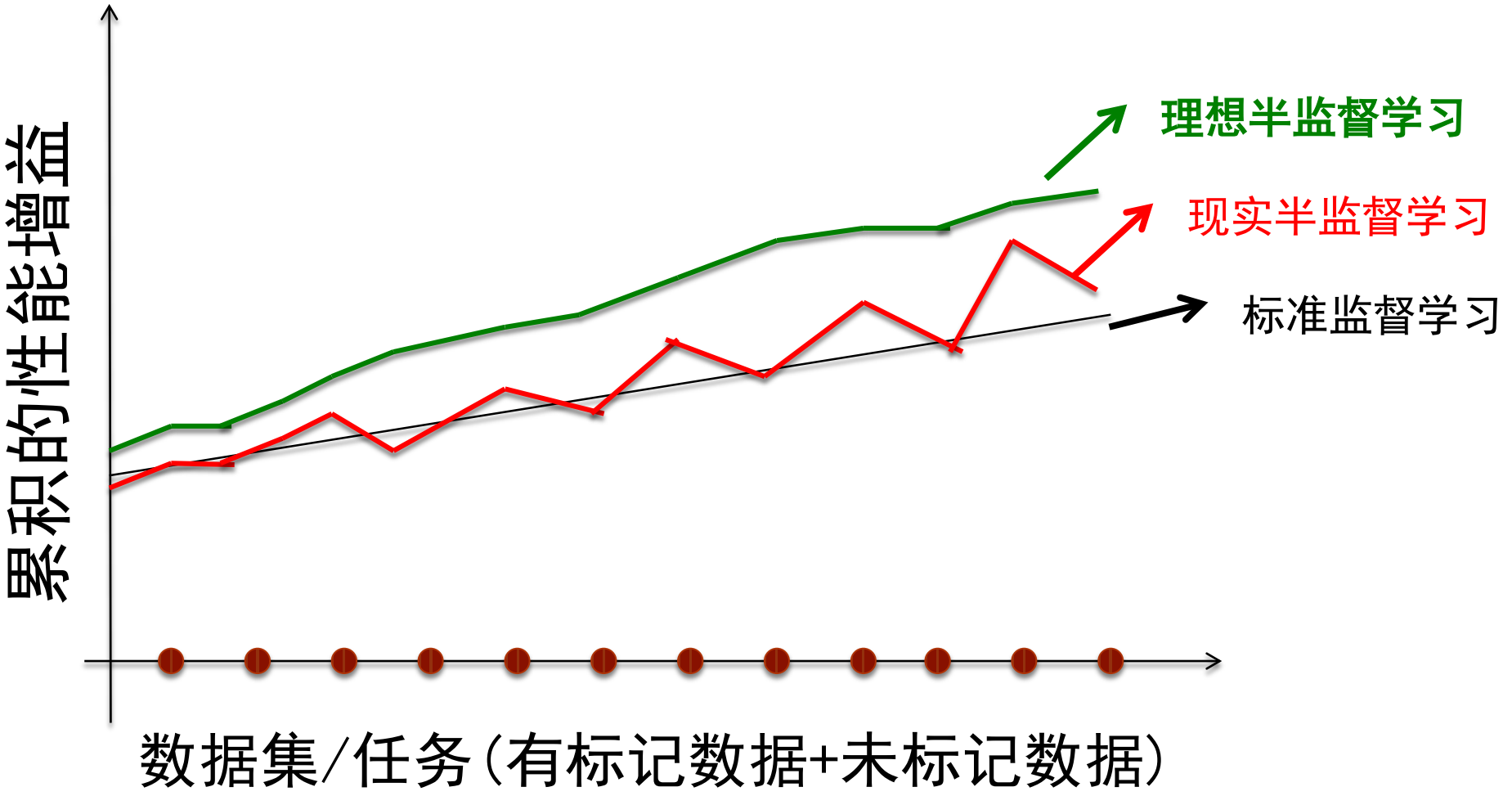
20 labeled examples

Data	SVM	TSVM(C=1.0)	CMN(5NN)	TSVM(LM)	CMN(LM)
blood	74.86±3.18	68.07±2.54	70.61±5.03	67.55±2.62	70.85±5.00
breast-cw	94.81±2.28	**96.52±0.21**	95.78±0.65	**96.52±0.21**	95.81±0.67
clean1	62.41±5.02	55.87±6.03	59.90±5.81	54.98±5.41	59.90±5.81
cylinder-bands	58.63±4.65	59.61±3.27	**64.19±1.92**	59.44±3.20	**63.81±2.03**
echocardiogram	80.49±3.10	75.58±4.46	78.75±3.51	75.58±4.46	78.97±3.57
ethn	77.45±2.58	**81.89±3.14**	80.72±11.9	**81.89±3.14**	79.91±13.4
heart-h	76.58±5.76	**80.29±1.48**	77.70±3.48	**80.58±1.44**	78.01±3.39
horse-colic	70.56±4.28	68.83±4.02	71.33±4.04	68.63±3.90	70.53±5.22
house	91.27±3.88	91.67±0.98	89.31±1.92	91.58±1.09	89.55±1.79
ilpd-il	66.16±3.65	64.17±3.64	**69.47±1.31**	64.17±3.64	**69.88±1.44**
ionosphere	80.89±4.34	**85.90±5.18**	78.11±6.54	**85.90±5.18**	77.99±6.77
isolet	96.85±1.48	**99.67±0.07**	96.91±2.01	**99.67±0.07**	96.91±2.01
liverDisorders	57.18±2.29	50.6±4.12	55.8±2.26	50.23±3.68	56.01±2.27
mammographic	77.82±3.25	77.93±0.73	72.17±4.19	78.11±0.42	72.17±4.19
monks-1	60.54±3.74	58.55±6.16	53.02±2.34	58.41±6.26	52.85±2.44
monks-2	56.85±2.91	53.13±1.95	**59.54±4.75**	52.91±2.31	59.26±5.06
monks-3	68.82±4.96	63.87±8.84	57.66±3.62	63.87±8.84	56.59±3.86
mushroom	84.65±3.66	86.30±3.45	86.78±4.59	85.66±3.57	86.78±4.55
oocytes-mn4	64.02±4.75	59.69±5.52	64.89±3.02	59.60±5.51	65.02±3.03
oocytes-tn2	61.58±5.89	52.75±3.95	56.75±2.53	52.74±3.94	56.39±2.69
optdigits	97.62±1.24	95.08±0.30	**99.68±0.04**	95.08±0.30	**99.70±0.06**
sat	99.51±0.19	**99.89±0.05**	**99.84±0.02**	**99.89±0.05**	**99.85±0.03**
spambase	80.00±4.03	75.00±7.95	65.31±4.09	77.09±6.30	65.10±4.10
spect	61.79±5.61	63.77±6.70	63.10±4.06	64.10±6.77	63.20±4.05
spectf	71.35±2.36	73.84±7.14	**72.73±1.31**	73.84±7.14	**72.85±0.42**
texture	99.50±0.76	95.96±0.31	**100.0±0.0**	95.95±0.29	**100.0±0.0**
tic-tac-toe	70.69±7.93	64.01±4.95	**99.52±0.79**	64.01±4.95	**99.52±0.79**
titanic	74.14±5.68	70.45±5.62	66.43±10.6	71.72±3.76	64.26±12.7
vertebral-c2	78.55±3.29	74.25±3.62	75.70±3.35	74.25±3.62	75.75±3.49
credit-a	77.41±4.07	80.17±5.12	77.39±4.54	**80.89±4.75**	77.39±4.54
heart	76.98±3.81	77.26±3.25	73.26±4.83	77.26±3.25	73.26±4.83
house-v	90.18±2.46	88.60±2.57	88.32±1.89	88.63±2.59	88.39±1.94
krvskp	73.46±4.47	59.19±7.14	60.64±3.90	57.09±6.65	60.02±3.99
pima	69.19±3.69	70.24±3.22	66.49±1.94	70.24±3.22	66.18±1.97
statlog-ac	54.85±3.05	**56.89±1.65**	**63.73±2.87**	**56.97±1.83**	**64.95±3.39**
statlog-gc	62.52±3.85	61.22±2.38	**68.29±2.02**	61.22±2.38	**68.67±2.05**
statlog-h	74.44±5.60	**79.98±3.82**	77.06±4.96	**81.48±3.12**	77.06±4.96
twonorm	92.88±1.86	**97.76±0.06**	71.52±20.3	**97.76±0.06**	69.34±21.3
vehicle	85.33±4.44	**93.04±3.68**	**92.30±4.05**	**93.04±3.68**	**92.30±4.05**
wdbc	85.93±3.43	**93.12±1.77**	80.52±5.33	**93.12±1.77**	79.79±6.58
Ave.Acc.	75.97±12.5	75.01±14.6	75.03±13.5	75.04±14.8	74.87±13.6
W/T/L against SVM		11/15/14	11/15/14	12/15/13	10/16/14

Naively applying SSL to given data causes disappointed results

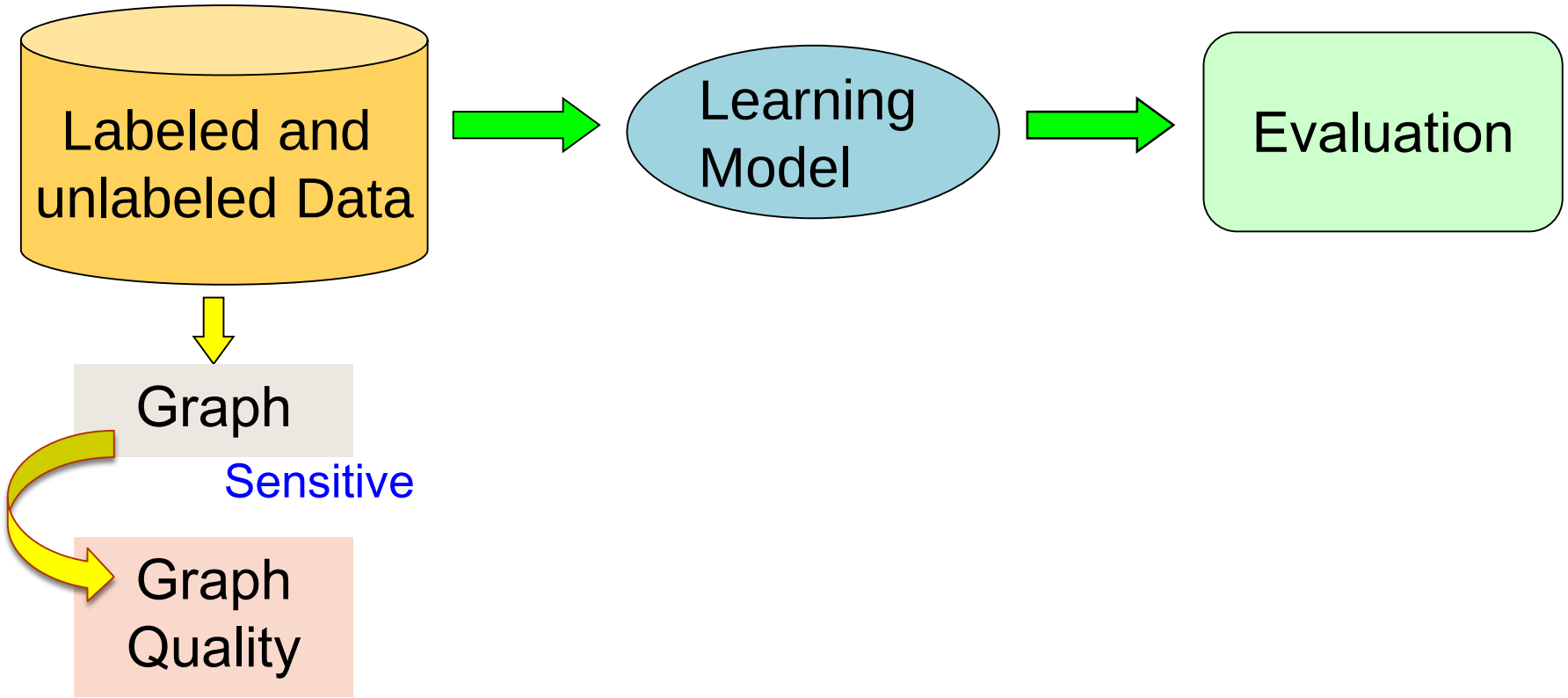
半监督学习的目标

理想的半监督学习应该是安全的，然而现实还做不到



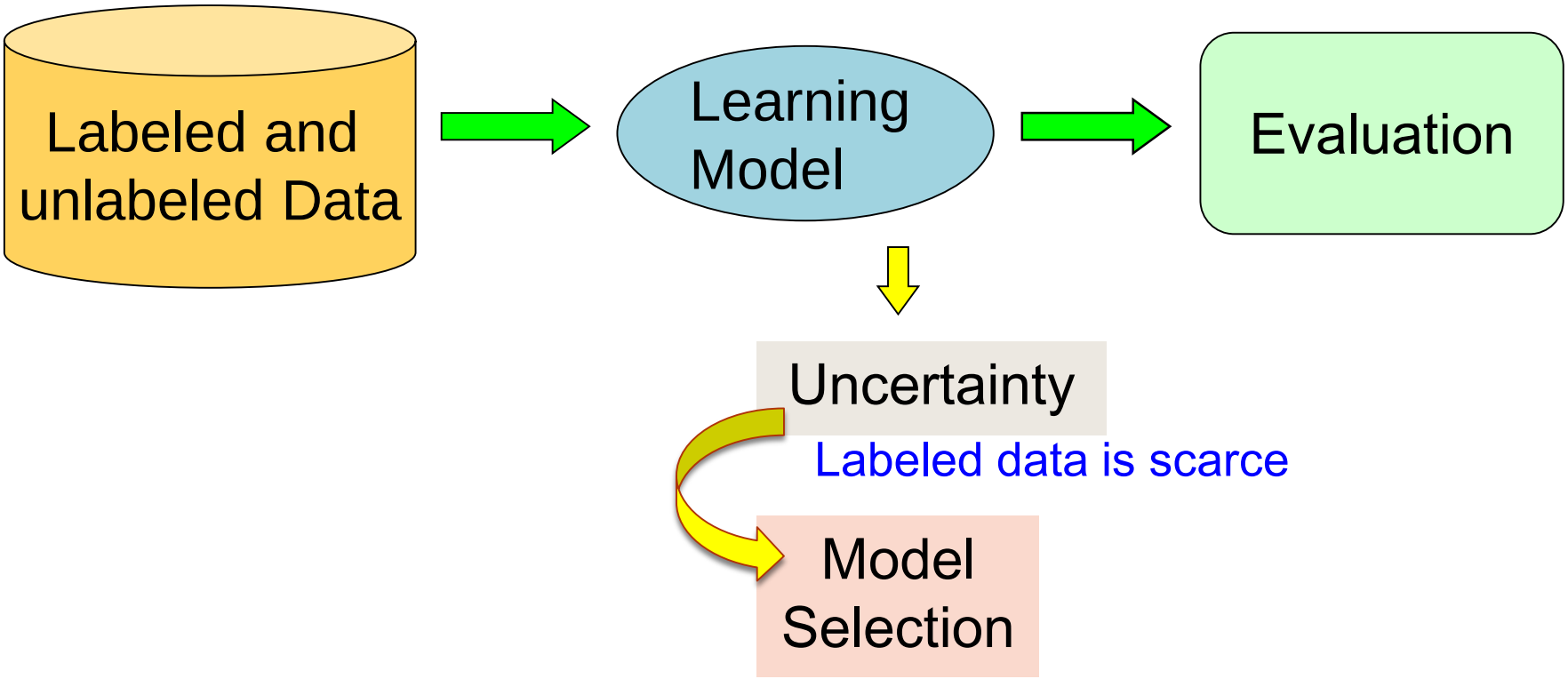
Why not safe?

A typical learning pipeline



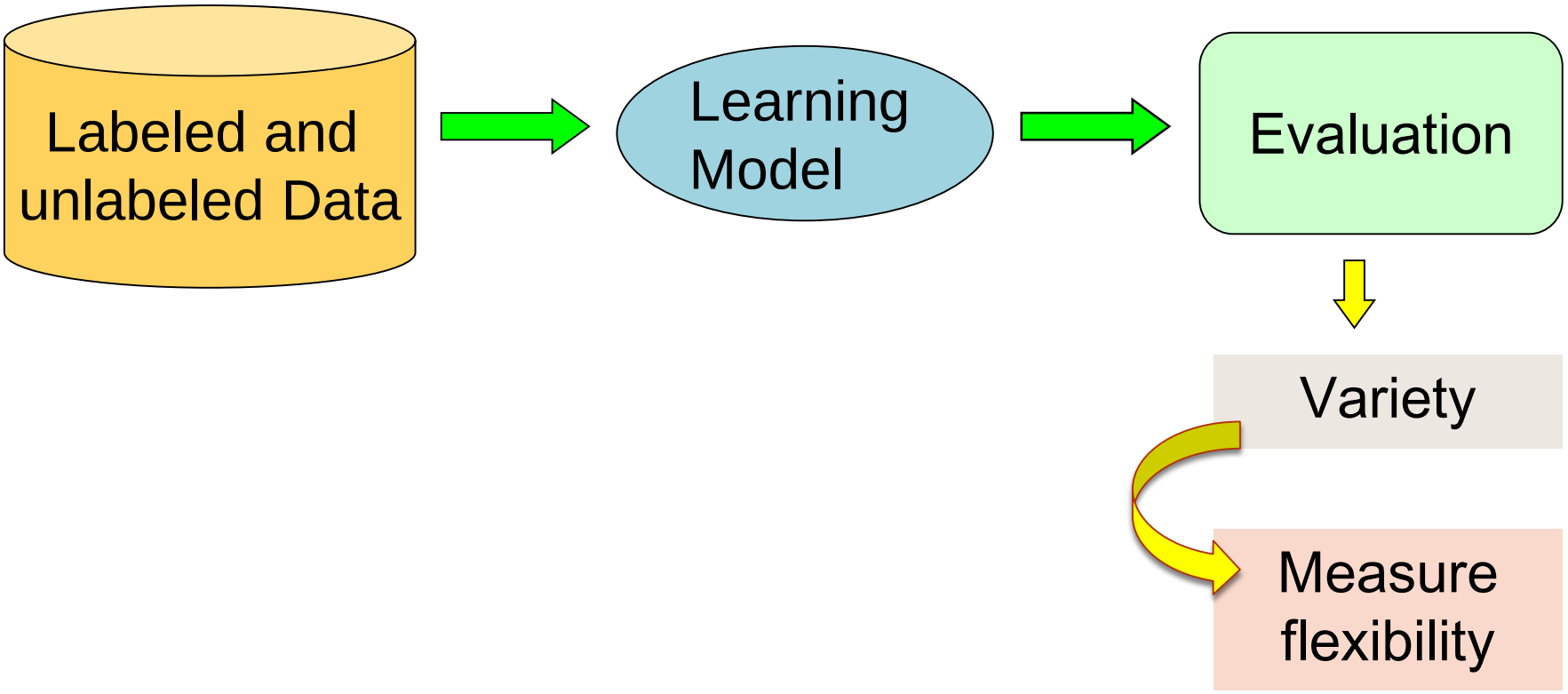
Why not safe?

A typical learning pipeline



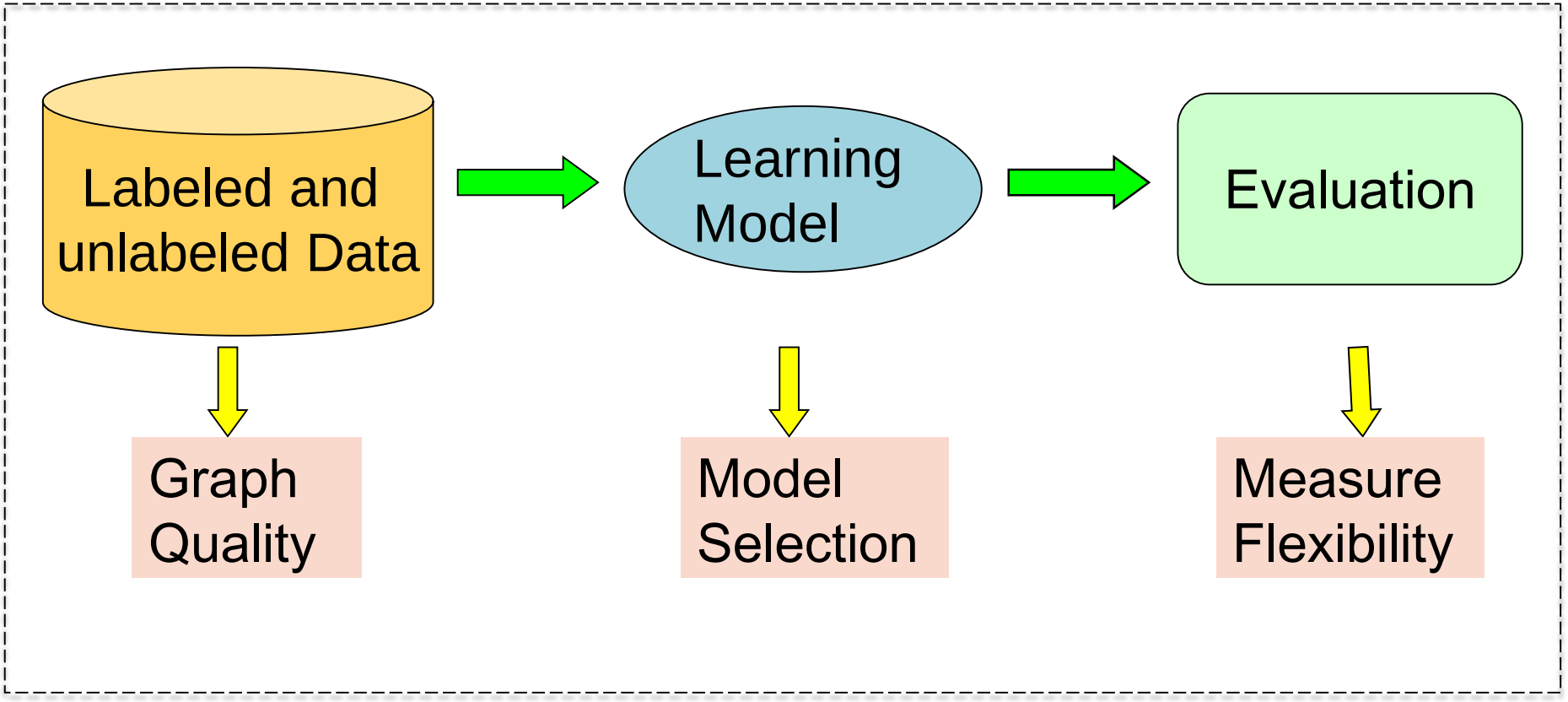
Why not safe?

A typical learning pipeline



A line of research

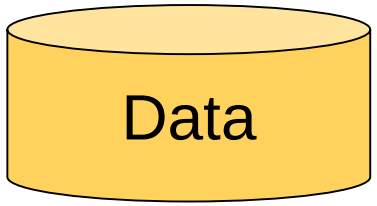
A typical learning pipeline

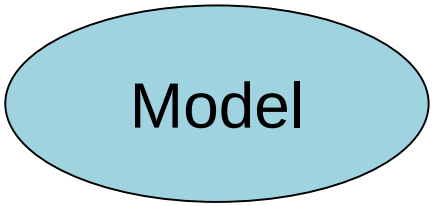


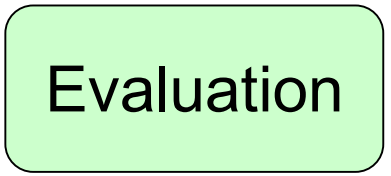
Towards safe semi-supervised learning

Outline

- Graph structure
- Model selection
- Performance measures
- Conclusion

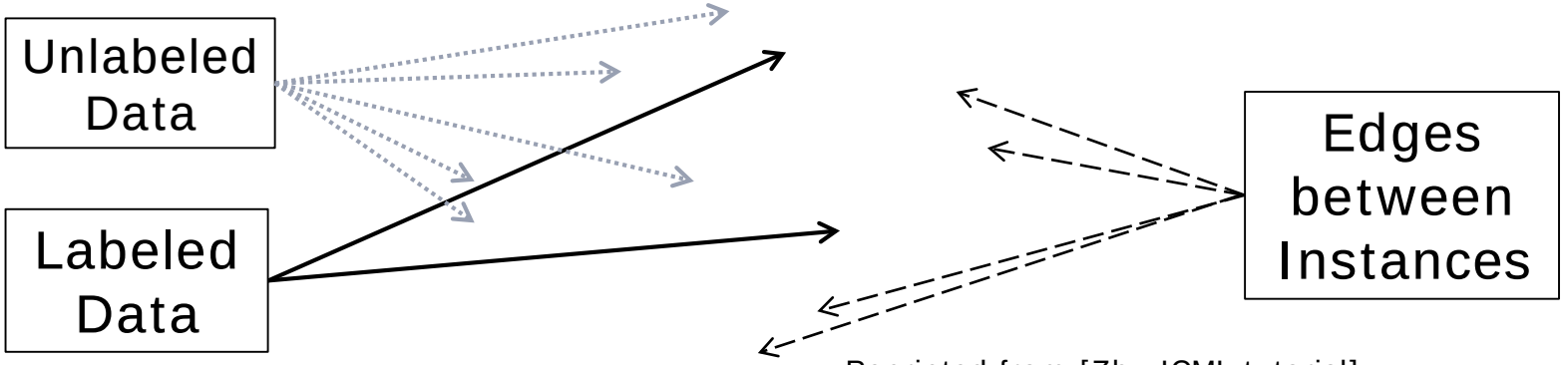






Graph-based SSL (GSSL)

A graph is given on the labeled and unlabeled data



Reprinted from [Zhu ICML tutorial]

GSSL assumption: instances connected by heavy edge tend to have the same label

$$\min_{\mathbf{z}} \quad \sum_{e_{ij} \in \mathcal{E}} w_{ij} \|z_i - z_j\|^2$$

$$\text{s.t.} \quad z_i = y_i, \; i = 1, \ldots, l;$$

$$z_j \in [-1, 1], \; j = l+1, \ldots, l+u;$$

Similar instances share similar labels

Graph Quality

However, the quality of the graph seriously affects the performance of GSSL methods

- widely accepted in many literatures [Zhu, survey07; Belkin and Niyogi, ICML08; Wang and Zhang, TKDE08; Jebara et al., ICML09]
- many kinds of graphs have been developed, e.g., k-NN graph, b-matching graph [Jebara et al., ICML09], LLE graph [Wang and Zhang, TKDE08], minimal spanning tree [Carreira-Perpinan and Zemel, NIPS05], etc.
- many distance metrics have been proposed, e.g., Euclidean distance, cosine distance, Manhattan distance, etc. [Zhu, survey07]

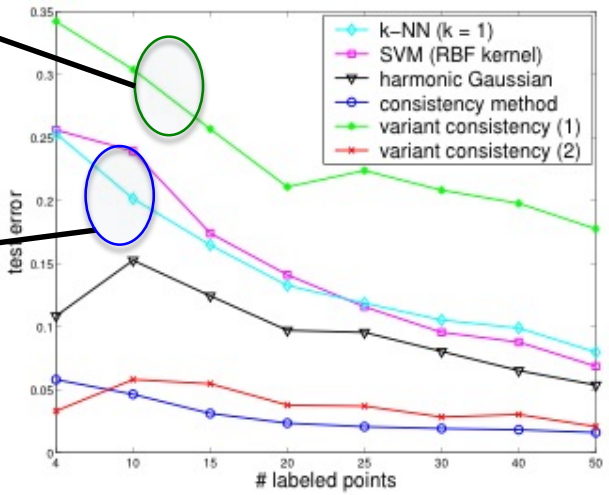
How to judge the graph quality remains challenging

Graph Quality (cont.)

An impropriate graph may even deteriorate the performance

GSSL method

Supervised 1NN



Reprinted from [Zhou NIPS04]

the higher the worse

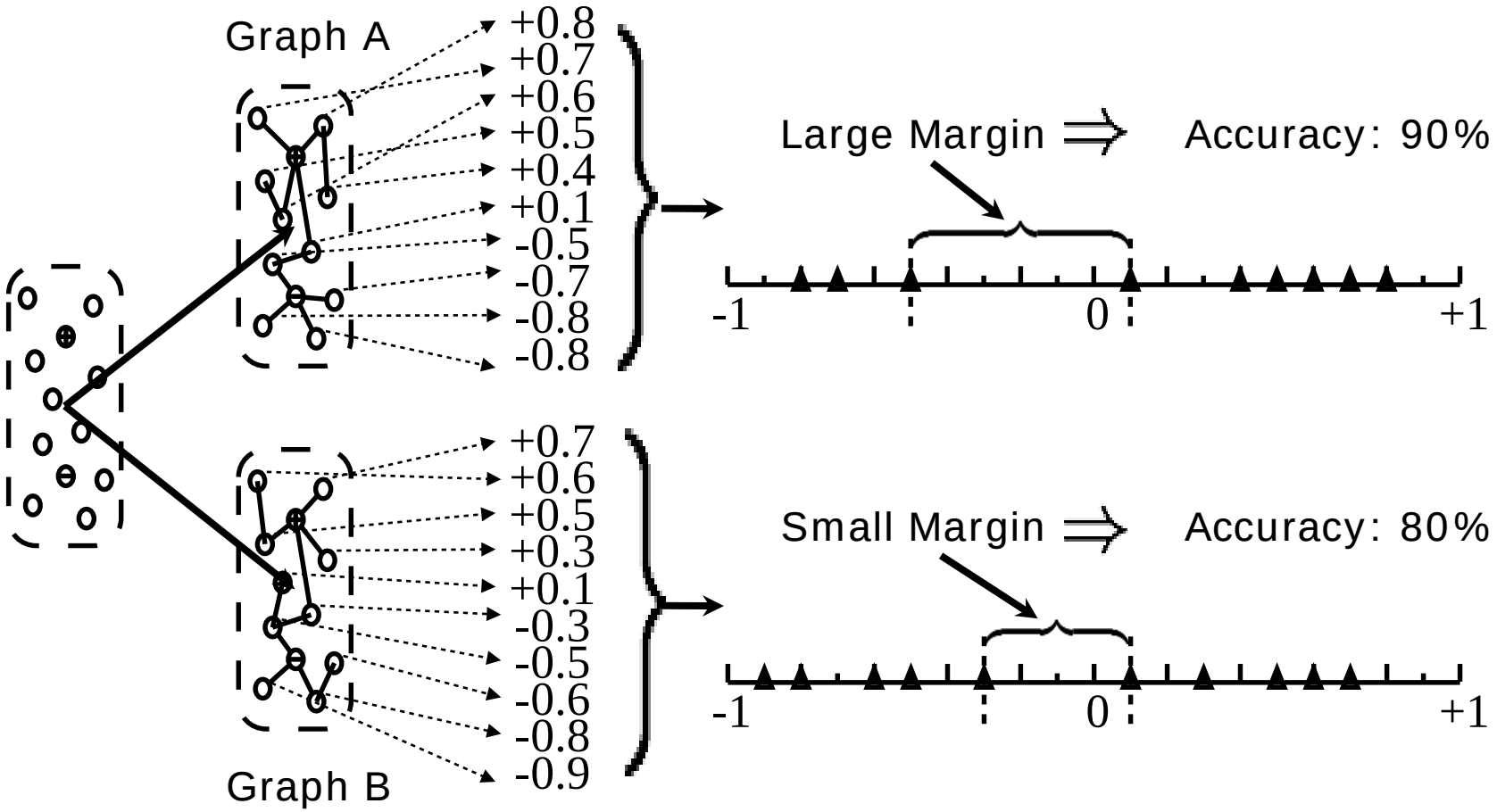
Impropriate graph degenerates the performance

Similar observations can be found in many literatures, e.g., [Zhou et al., 2004; Belkin and Niyogi, 2004; Wang and Zhang, 2008; Karlen et al., 2008]

How to judge the quality of graph, so as to construct a robust GSSL method?

Basic Idea

Given a set of candidate graphs, when one graph has a high quality, its predictive results may have a large margin separation



Illustrated Experiments

We perform large margin separation on the predictive results of GSSL methods on different graphs

Higher accuracy
Lower accuracy
Lower hinge loss
Higher hinge loss

GSSL with two different graphs

Dataset	Domain	5NN Graph with Euclidean Distance Accuracy	5NN Graph with Euclidean Distance Hinge Loss	5NN Graph with Manhattan Distance Accuracy	5NN Graph with Manhattan Distance Hinge Loss
breast-cancer	life	91.6±3.1	0.529±0.110	92.6±2.7	0.370±0.106
coil	image	60.9±6.2	0.341±0.109	62.4±6.9	0.276±0.109
musk-1	physical	57.7±4.9	0.632±0.139	56.4±4.5	0.671±0.145
text	text	52.3±3.3	0.964±0.006	50.3±0.0	0.994±0.009

Higher accuracy, lower hinge loss

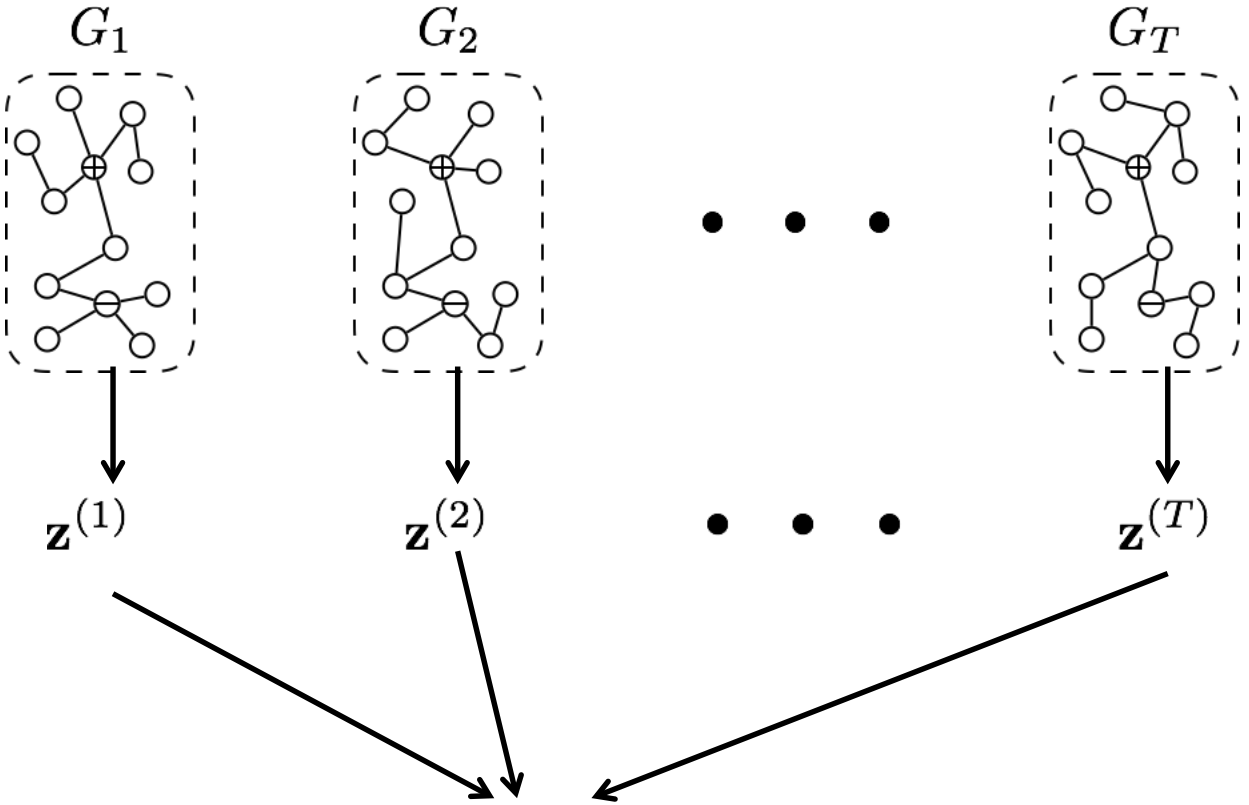
Large margin principle is helpful to judge the quality of graph

The Proposed Method

Given a set of candidate graphs

Perform GSSL and obtain the predictive results

Regenerate a new SSL data set by treating the predictive results as features



$$\mathbf{u}_i = [z_i^{(1)}, z_i^{(2)}, \ldots, z_i^{(T)}], \text{ for } i = 1, \ldots, l+u$$

$$\{\mathbf{u}_i, y_i\}_{i=1}^{l} \text{ and } \{\mathbf{u}_j\}_{j=l+1}^{l+u}$$

The Proposed Method

Given a new SSL data set

$$\{\mathbf{u}_i, y_i\}_{i=1}^{l} \text{ and } \{\mathbf{u}_j\}_{j=l+1}^{l+u}$$

Construct a large margin classifier

$$f(\mathbf{u}) = \mathbf{w}'\mathbf{u} + b$$

$$\min_{\mathbf{w},\hat{\mathbf{y}}} \quad \frac{1}{2}\|\mathbf{w}\|_2^2 + C_1\sum_{i=1}^{l}\ell(y_i f(\mathbf{u}_i)) + C_2\sum_{j=l+1}^{l+u}\ell(\hat{y}_j f(\mathbf{u}_j))$$

$$\text{s.t.} \quad \hat{y}_{l+j} \in \{+1, -1\}, \quad j = 1, \ldots, u;$$

$$\left|\frac{\sum_{j=l+1}^{l+u}\hat{y}_j}{u} - \frac{\sum_{i=1}^{l} y_i}{l}\right| \leq \beta \quad \Rightarrow$$

Semi-Supervised SVM optimization

Experimental setting

- 20 data sets that cover a broad range of properties
- Compared methods
  - Baseline method: Supervised 1NN
  - Two classical GSSL methods
    - Harmonic [Zhu et al., ICML03]
    - LLGC [Zhou et al., NIPS04]
  - Two ensemble GSSL methods
    - Majority Voting
    - CGL [Argyriou et al., NIPS05]
  - Model selection method
    - GSSL-CV

Experimental results

10 labeled data; 20 splits;

Significantly hurt the performance

Data sets	1NN	Harmonic 3NN Graph	Harmonic 5NN Graph	Harmonic 7NN Graph	CGL	Majority Voting	GSSL-CV	LEAD
breast-cancer	94.0 ± 2.6	**95.7 ± 2.1**	**95.6 ± 1.1**	**95.3 ± 1.4**	94.7 ± 2.7	92.2 ± 0.8	**95.8 ± 2.1**	**94.1 ± 2.5**
coil	60.4 ± 5.2	**67.7 ± 8.0**	**64.2 ± 5.9**	62.8 ± 6.9	**65.6 ± 6.3**	**65.3 ± 6.9**	**67.6 ± 7.1**	**63.5 ± 6.9**
credit	72.0 ± 6.1	70.2 ± 6.0	70.3 ± 8.4	67.9 ± 9.6	68.0 ± 6.7	71.5 ± 6.9	70.1 ± 6.8	72.2 ± 5.8
digit1	73.3 ± 3.9	**87.4 ± 4.9**	**87.4 ± 5.6**	**86.1 ± 6.7**	**93.6 ± 2.3**	**92.4 ± 2.7**	**90.8 ± 4.5**	**79.1 ± 3.8**
heart-hungarian	76.2 ± 7.3	73.9 ± 5.8	74.9 ± 4.5	76.2 ± 5.5	75.7 ± 10.0	76.2 ± 6.4	75.4 ± 5.7	76.3 ± 7.2
horse	62.9 ± 5.0	64.5 ± 7.6	63.6 ± 8.4	64.3 ± 6.7	60.6 ± 8.1	64.6 ± 5.0	65.3 ± 7.0	62.9 ± 5.1
ionosphere	73.4 ± 6.8	72.0 ± 6.3	72.7 ± 8.6	72.8 ± 8.9	69.9 ± 2.9	75.1 ± 5.7	75.1 ± 7.0	74.0 ± 6.9
mammographic	73.5 ± 5.9	66.6 ± 5.8	67.8 ± 5.1	69.3 ± 5.2	71.7 ± 6.2	70.5 ± 6.1	66.0 ± 5.6	74.2 ± 4.9
mnist1vs7	92.3 ± 3.3	**97.7 ± 4.7**	**97.6 ± 4.3**	95.4 ± 9.5	-	**97.4 ± 0.3**	**97.8 ± 4.7**	**96.8 ± 1.1**
mnist3vs8	79.1 ± 3.8	63.5 ± 18.2	67.7 ± 21.0	62.1 ± 18.2	-	84.3 ± 19.5	77.8 ± 21.0	**80.5 ± 3.8**
mnist4vs9	67.0 ± 6.2	67.8 ± 16.5	62.7 ± 14.5	58.6 ± 11.8	-	**76.6 ± 13.7**	72.4 ± 15.8	**67.4 ± 6.3**
mnist7vs9	76.0 ± 3.7	70.2 ± 20.0	63.7 ± 18.5	59.4 ± 14.1	-	79.3 ± 15.3	80.9 ± 15.2	**77.9 ± 4.3**
mushroom	79.3 ± 7.9	79.3 ± 7.9	79.3 ± 7.9	79.3 ± 7.9	-	79.1 ± 7.9	**80.4 ± 7.0**	80.1 ± 7.3
musk-1	60.5 ± 4.0	61.1 ± 5.8	60.3 ± 6.0	60.6 ± 5.4	60.3 ± 5.3	61.8 ± 5.0	61.0 ± 6.6	60.5 ± 4.1
musk-2	73.9 ± 6.7	**76.2 ± 5.9**	**78.5 ± 5.9**	**78.7 ± 6.4**	57.3 ± 5.7	**77.2 ± 4.6**	**80.0 ± 5.4**	75.2 ± 5.6
spambase	71.9 ± 5.5	62.4 ± 6.3	62.5 ± 5.6	61.8 ± 7.6	-	62.7 ± 10.3	62.8 ± 6.5	**72.5 ± 5.0**
text	59.2 ± 5.6	54.9 ± 4.0	53.2 ± 3.0	53.9 ± 5.4	**63.1 ± 6.3**	56.0 ± 4.6	54.3 ± 4.0	59.2 ± 5.6
twonorm	89.2 ± 3.6	60.3 ± 18.2	57.5 ± 15.3	62.6 ± 19.0	**96.8 ± 0.2**	90.5 ± 13.8	88.3 ± 14.8	**89.2 ± 3.6**
usps	81.3 ± 2.8	80.8 ± 0.9	80.5 ± 0.8	80.3 ± 0.7	62.6 ± 6.8	70.7 ± 3.3	80.8 ± 0.9	81.3 ± 2.8
vertebral	70.3 ± 7.5	71.0 ± 3.4	72.1 ± 2.6	71.9 ± 3.1	68.0 ± 9.0	71.0 ± 4.9	72.4 ± 5.0	70.4 ± 7.5
Ave. Accuracy	74.3 ± 9.9	72.2 ± 11.4	71.6 ± 12.0	71.0 ± 11.9	N/A	75.7 ± 11.3	75.7 ± 11.6	75.4 ± 10.2
Win/Tie/Loss		5/10/5	5/9/6	3/9/8	4/6/4	5/10/5	6/11/3	**9/11/0**

GSSL often hurt the performance
LEAD never hurt the performance

LEAD achieves highly competitive performance

Experimental results

10 labeled data; 20 splits;

Significantly hurt the performance

Data sets	1NN	LLGC 3NN Graph	LLGC 5NN Graph	LLGC 7NN Graph	CGL	Majority Voting	GSSL-CV	LEAD
breast-cancer	94.0 ± 2.6	**95.9 ± 0.7**	**95.4 ± 1.1**	94.7 ± 1.5	94.7 ± 2.7	92.7 ± 0.6	**96.1 ± 0.7**	**94.1 ± 2.5**
coil	60.4 ± 5.2	**66.8 ± 6.5**	**63.6 ± 6.2**	61.7 ± 6.4	**65.6 ± 6.3**	**64.0 ± 6.1**	**67.4 ± 6.7**	**64.3 ± 6.6**
credit	72.0 ± 6.1	72.3 ± 5.3	69.7 ± 7.6	66.1 ± 8.2	68.0 ± 6.7	73.3 ± 6.2	72.6 ± 7.1	72.5 ± 5.5
digit1	73.3 ± 3.9	**90.4 ± 3.5**	**90.4 ± 3.3**	**90.2 ± 3.7**	**93.6 ± 2.3**	**91.5 ± 3.1**	**91.2 ± 3.7**	**84.2 ± 3.7**
heart-hungarian	76.2 ± 7.3	75.7 ± 4.9	73.8 ± 4.2	72.7 ± 6.0	75.7 ± 10.0	74.9 ± 7.2	75.7 ± 5.4	**76.4 ± 7.2**
horse	62.9 ± 5.0	65.2 ± 6.1	64.4 ± 5.7	62.9 ± 5.3	60.6 ± 8.1	63.3 ± 5.1	65.3 ± 6.2	62.9 ± 5.1
ionosphere	73.4 ± 6.8	70.6 ± 7.7	68.2 ± 7.0	67.2 ± 6.4	69.9 ± 2.9	71.5 ± 7.6	71.1 ± 7.2	73.3 ± 6.9
mammographic	73.5 ± 5.9	67.7 ± 5.9	69.8 ± 4.3	71.5 ± 4.8	71.7 ± 6.2	71.7 ± 4.8	67.3 ± 5.7	74.1 ± 4.7
mnist1vs7	92.3 ± 3.3	**98.6 ± 1.5**	**98.8 ± 0.8**	**98.8 ± 0.7**	-	**97.4 ± 0.4**	**98.4 ± 1.7**	**98.7 ± 1.1**
mnist3vs8	79.1 ± 3.8	**95.5 ± 1.7**	**95.5 ± 2.0**	**95.4 ± 2.1**	-	**95.8 ± 1.8**	**95.8 ± 2.0**	**93.6 ± 2.1**
mnist4vs9	67.0 ± 6.2	**88.8 ± 7.7**	**87.8 ± 7.4**	**87.1 ± 7.7**	-	**88.3 ± 7.7**	**88.8 ± 7.6**	**84.6 ± 7.4**
mnist7vs9	76.0 ± 3.7	**94.7 ± 3.9**	**93.7 ± 4.5**	**93.5 ± 4.4**	-	**94.1 ± 4.1**	**93.8 ± 4.2**	**91.5 ± 4.3**
mushroom	79.3 ± 7.9	79.3 ± 7.9	79.3 ± 7.9	79.3 ± 7.9	-	79.1 ± 7.9	**80.4 ± 7.0**	80.2 ± 7.1
musk-1	60.5 ± 4.0	61.4 ± 5.2	59.4 ± 5.2	60.0 ± 4.6	60.3 ± 5.3	60.7 ± 5.1	61.4 ± 5.1	60.2 ± 4.1
musk-2	73.9 ± 6.7	74.7 ± 6.3	**77.3 ± 4.6**	**78.6 ± 4.9**	57.3 ± 5.7	**77.7 ± 3.8**	**77.9 ± 5.6**	75.9 ± 5.1
spambase	71.9 ± 5.5	72.1 ± 5.3	73.2 ± 4.8	72.7 ± 6.3	-	**79.4 ± 5.1**	71.3 ± 5.7	**78.1 ± 2.9**
text	59.2 ± 5.6	58.7 ± 4.0	58.0 ± 4.6	57.8 ± 5.8	**63.1 ± 6.3**	60.1 ± 5.3	57.9 ± 5.2	59.2 ± 5.6
twonorm	89.2 ± 3.6	**96.0 ± 0.5**	**96.3 ± 0.6**	**96.5 ± 0.7**	**96.8 ± 0.2**	**97.1 ± 0.3**	**96.6 ± 0.5**	**94.6 ± 1.0**
usps	81.3 ± 2.8	**82.9 ± 2.2**	81.8 ± 1.9	81.0 ± 1.4	62.6 ± 6.8	76.1 ± 4.6	**83.0 ± 2.2**	**81.5 ± 2.8**
vertebral	70.3 ± 7.5	69.6 ± 2.2	68.4 ± 2.2	68.0 ± 1.3	68.0 ± 9.0	64.9 ± 7.1	71.3 ± 4.0	**70.7 ± 7.2**
Ave. Accuracy	74.3 ± 9.9	78.8 ± 13.0	78.2 ± 13.4	77.8 ± 13.6	N/A	78.7 ± 12.8	79.2 ± 13.0	78.5 ± 11.9
Win/Tie/Loss		9/9/2	9/9/2	7/10/3	4/6/4	9/9/2	11/8/1	**12/8/0**

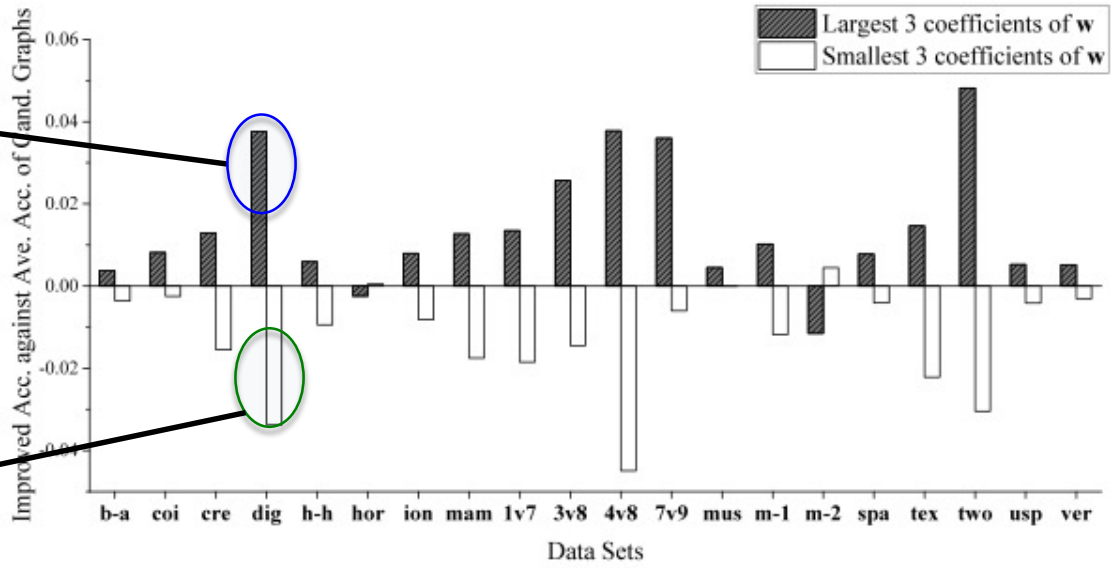
GSSL often hurt the performance
LEAD never hurt the performance

LEAD obtains highly competitive performance

Why the proposal is effective

Average Accuracy for the graphs with the **largest 3 coefficients**

Average Accuracy for the graphs with the **smallest 3 coefficients**



Large margin principle can avoid less accurate graphs.

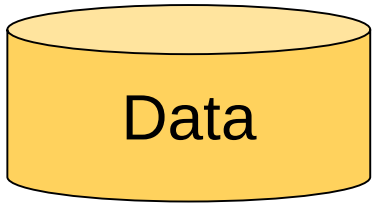
Hints

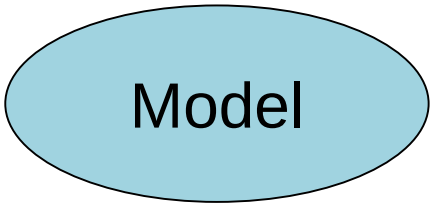
To judge the graph quality

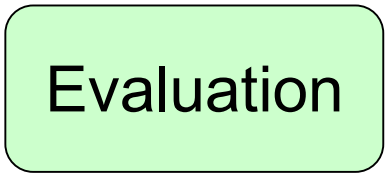
- ✓ Large margin principle helps
- ✓ Ensemble learning + large margin principle could robustly exploit unlabeled data

Outline

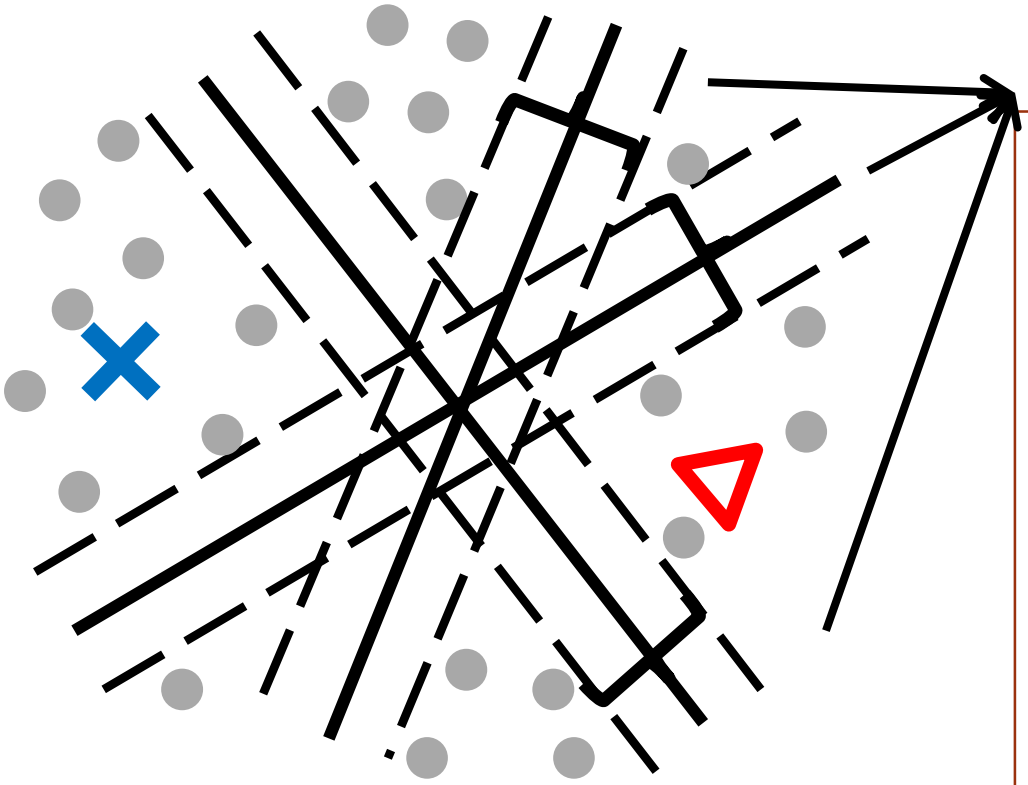
- Graph structure
- Model selection
- Performance measures
- Conclusion







Observation



S3VMs

Large Margin Separator

i) **More than one** Large Margin Separators!!

ii) Current S3VMs **randomly** select one of them as the output.

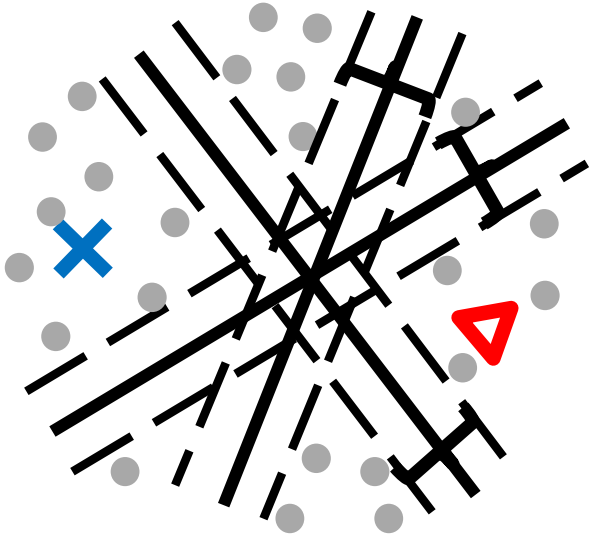
iii) Large Margin Separators are usually **diverse**.

iv) **Incorrect selection** degenerates the performance!

S4VM (Safe S3VM) is presented

S4VM: A simple algorithm

- Step 1: Generate a pool of large-margin separators (LMS).



- Step 2: Construct S4VM by optimizing the performance improvement under the worst-case

S4VM Formulation

- Maximize accuracy

$$\max_{\mathbf{y}\in\{\pm 1\}^u} \; J(\mathbf{y},\mathbf{y}^*,\mathbf{y}^{svm}) = gain(\mathbf{y},\mathbf{y}^*,\mathbf{y}^{svm}) - \lambda\, loss(\mathbf{y},\mathbf{y}^*,\mathbf{y}^{svm})$$

  - gain(): gained accuracy against SVM (without using unlabeled data)
  - loss(): lost accuracy against SVM (without using unlabeled data)
  - λ : measure the risk that user would like to undertake
  - $\mathbf{y}^*$: ground-truth label assignment
  - Difficulty: The ground-truth is unknown.

- Note that ground-truth is a LMS, we assume that $\mathbf{y}^* \in \{\hat{\mathbf{y}}_t\}_{t=1}^T$
- S4VM maximizes the worst-case accuracy

$$\bar{\mathbf{y}} = \arg\max_{\mathbf{y}\in\{\pm 1\}^u} \; \min_{\hat{\mathbf{y}}\in\{\hat{\mathbf{y}}_t\}_{t=1}^T} \; gain(\mathbf{y},\hat{\mathbf{y}},\mathbf{y}^{svm}) - \lambda\, loss(\mathbf{y},\hat{\mathbf{y}},\mathbf{y}^{svm})$$

Theoretical Analysis

Theorem 1: If $\mathbf{y}^* \in \{\hat{\mathbf{y}}_t\}_{t=1}^T$ and $\lambda \geq 1$, the accuracy of $\bar{\mathbf{y}}$ is never worse than that of $\mathbf{y}^{svm}$.

Under the assumption employed in S3VMs, that is the ground-truth is realized by a large-margin separator, S4VM is provable safe

Proposition 2: If $\mathbf{y}^* \in \{\hat{\mathbf{y}}_t\}_{t=1}^T$ and $\lambda = 1$, the accuracy of $\bar{\mathbf{y}}$ achieves the maximal performance improvement over that of $\mathbf{y}^{svm}$ in the worst case.

Under the assumption employed in S3VMs, S4VM already achieves the largest performance improvement.

Experiment

10 Lab	SVM (linear / rbf)	TSVM (linear / rbf)	$S4VM_s$ (linear / rbf)
aust	67.2±8.0 / 66.6±7.5	67.8±12.1 / 67.8±13.2	68.7±9.8 / **68.1±9.8**
ausl	77.3±7.1 / 72.9±6.1	**81.1±6.1 / 77.0±8.2**	77.8±7.5 / 73.3±6.4
brea	94.4±3.5 / 95.7±2.8	94.5±0.4 / 94.7±0.1	**96.5±0.6 / 96.8±0.4**
clea	56.6±5.0 / 57.1±8.1	56.7±5.7 / 55.8±7.8	56.7±5.1 / 56.9±8.2
diab	65.7±5.4 / 65.6±5.5	65.7±5.4 / 65.3±6.2	66.4±5.2 / 65.5±6.2
germ	63.1±8.0 / 64.8±12.1	62.2±5.5 / 61.6±5.1	63.3±7.9 / 64.8±11.8
habe	64.6±7.0 / 65.9±8.1	62.1±7.9 / 62.4±6.2	65.0±6.5 / 65.9±7.9
hear	72.2±6.7 / 71.8±6.8	71.9±6.9 / 72.5±7.6	72.2±6.7 / 72.7±7.0
hous	89.0±3.0 / 88.0±2.8	**90.6±3.3 / 89.8±1.8**	**89.6±3.1** / 88.5±2.3
houv	88.0±2.1 / 87.0±3.1	85.4±6.4 / 85.0±5.8	87.9±4.4 / 86.8±3.9
iono	72.6±6.8 / 74.7±9.1	72.1±9.4 / 77.4±8.6	73.7±6.7 / 75.2±10.0
isol	89.6±6.8 / 81.0±14.7	86.6±10.0 / **86.7±9.9**	**91.4±9.0 / 85.2±17.2**
live	55.3±5.4 / 55.5±5.9	53.3±5.0 / 54.2±4.8	54.8±5.7 / 55.2±6.0
optd	92.2±4.4 / 85.8±9.8	87.1±8.7 / 87.2±8.7	**93.5±5.6 / 87.9±11.3**
vehi	74.0±7.9 / 74.3±8.1	76.2±9.2 / **78.7±8.7**	74.7±8.5 / 76.2±9.8
wdbc	81.3±7.4 / 80.7±7.5	**84.3±6.6 / 84.4±6.3**	82.1±7.2 / **82.5±9.1**
digi	76.2±7.1 / 57.2±12.1	**80.7±3.9 / 84.2±4.4**	76.0±7.1 / **63.6±12.6**
USPS	78.5±2.9 / 80.0±0.1	72.2±3.7 / 71.6±4.9	78.7±2.7 / **80.1±0.2**
COIL	56.5±5.4 / 56.9±4.2	54.9±6.2 / 57.2±3.9	56.8±5.4 / 57.0±4.2
BCI	52.6±2.4 / 51.3±2.0	50.9±2.8 / 51.2±2.1	51.8±2.3 / 51.3±2.2
g241c	54.5±4.2 / 52.3±4.4	**78.3±4.5 / 59.5±3.4**	54.6±4.5 / **52.8±4.4**
g241n	56.4±5.2 / 52.5±5.3	53.3±7.3 / 52.8±5.2	56.3±5.1 / 52.7±5.1
Text	52.2±3.0 / 52.6±4.0	**64.8±7.9** / 52.4±6.1	52.1±2.9 / 52.6±4.0
Avg.Acc	71.3 / 69.7	72.2 / 71.3	71.8 / 70.6
SVM vs. Semi-Supervised: W/T/L		12/19/13	**0/32/12**

100 Lab	SVM (linear / rbf)	TSVM (linear / rbf)	$S4VM_s$ (linear / rbf)
aust	83.7±1.8 / 78.7±2.9	83.1±2.1 / 78.6±2.8	83.7±1.6 / 78.8±3.0
ausl	79.5±2.8 / 80.6±2.3	79.0±2.9 / 81.1±2.5	**81.1±2.1 / 81.3±2.0**
brea	95.0±1.3 / 95.4±1.0	**95.7±0.7 / 95.8±0.7**	95.2±1.3 / 95.4±1.0
clea	73.3±3.1 / 83.1±2.0	73.6±2.9 / 83.2±2.2	73.5±2.9 / 83.3±2.1
diab	74.6±1.6 / 70.3±2.1	73.7±1.9 / 70.0±2.1	74.3±1.9 / 70.6±2.4
germ	65.3±3.0 / 70.9±1.0	**66.1±2.2** / 68.8±2.3	**65.8±2.9** / 71.0±1.1
habe	71.7±3.1 / 68.3±2.8	69.1±3.0 / 66.3±2.6	71.9±3.2 / 68.1±2.6
hear	82.2±2.5 / 76.3±3.4	82.2±2.9 / 76.0±3.4	82.4±2.5 / 76.5±3.7
hous	94.0±1.8 / 94.9±1.7	92.1±3.2 / 92.4±3.3	94.2±1.6 / 94.9±1.7
houv	91.1±1.4 / 92.5±1.7	89.9±2.2 / 90.9±2.4	91.2±1.3 / 92.6±1.6
iono	83.9±2.7 / 91.5±2.1	81.8±3.0 / 90.6±2.8	83.4±2.6 / 91.6±2.1
isol	99.0±0.4 / 99.2±0.5	96.3±3.3 / 96.4±3.4	98.9±1.2 / **99.3±0.5**
live	64.3±3.6 / 66.5±2.6	64.8±3.1 / 66.1±2.3	63.1±4.2 / 66.8±3.3
optd	99.2±0.3 / 99.5±0.2	96.2±3.3 / 96.2±3.3	99.1±1.2 / 99.6±0.4
vehi	93.6±1.9 / 97.7±1.0	93.2±2.0 / 96.0±2.1	93.7±2.0 / **97.9±0.8**
wdbc	95.2±1.4 / 93.6±1.7	93.4±2.6 / 92.4±2.7	94.7±1.7 / 93.6±1.7
digi	90.8±0.7 / 94.2±1.5	**92.0±1.6** / 94.5±1.9	**91.5±0.8 / 94.9±1.4**
USPS	87.2±1.0 / 83.1±1.9	86.7±1.5 / **91.7±2.5**	**87.7±1.0 / 91.0±2.4**
COIL	80.2±2.2 / 87.1±2.0	80.8±2.7 / 87.0±1.5	**80.8±2.3** / 87.2±2.1
BCI	70.4±3.4 / 66.2±2.9	70.5±4.1 / 65.4±2.8	70.5±3.4 / 66.1±2.9
g241c	74.5±2.0 / 70.1±8.5	**80.0±1.4 / 77.8±1.6**	**75.3±1.8 / 74.8±4.1**
g241n	71.8±2.7 / 59.4±9.4	**75.4±4.5** / 65.0±13.8	**72.2±2.8 / 60.9±8.4**
Text	69.8±2.1 / 54.4±4.9	**74.2±1.4 / 58.8±4.9**	69.9±2.0 / 54.1±4.1
Avg.Acc	82.4 / 81.8	82.3 / 82.0	82.5 / 82.5
SVM vs. Semi-Supervised: W/T/L		17/17/10	**0/30/14**

In terms of average performance,
S4VM is highly competitive with TSVM

Experiment

Significantly degenerated performance

10 Lab	SVM (linear / rbf)	TSVM (linear / rbf)	$S4VM_s$ (linear / rbf)
aust	67.2±8.0 / 66.6±7.5	67.8±12.1 / 67.8±13.2	68.7±9.8 / **68.1±9.8**
ausl	77.3±7.1 / 72.9±6.1	**81.1±6.1 / 77.0±8.2**	77.8±7.5 / 73.3±6.4
brea	94.4±3.5 / 95.7±2.8	94.5±0.4 / 94.7±0.1	**96.5±0.6 / 96.8±0.4**
clea	56.6±5.0 / 57.1±8.1	56.7±5.7 / 55.8±7.8	56.7±5.1 / 56.9±8.2
diab	65.7±5.4 / 65.6±5.5	65.7±5.4 / 65.3±6.2	66.4±5.2 / 65.5±6.2
germ	63.1±8.0 / 64.8±12.1	62.2±5.5 / 61.6±5.1	63.3±7.9 / 64.8±11.8
habe	64.6±7.0 / 65.9±8.1	62.1±7.9 / 62.4±6.2	65.0±6.5 / 65.9±7.9
hear	72.2±6.7 / 71.8±6.8	71.9±6.9 / 72.5±7.6	72.2±6.7 / 72.7±7.0
hous	89.0±3.0 / 88.0±2.8	**90.6±3.3 / 89.8±1.8**	**89.6±3.1** / 88.5±2.3
houv	88.0±2.1 / 87.0±3.1	85.4±6.4 / 85.0±5.8	87.9±4.4 / 86.8±3.9
iono	72.6±6.8 / 74.7±9.1	72.1±9.4 / 77.4±8.6	73.7±6.7 / 75.2±10.0
isol	89.6±6.8 / 81.0±14.7	86.6±10.0 / **86.7±9.9**	**91.4±9.0 / 85.2±17.2**
live	55.3±5.4 / 55.5±5.9	53.3±5.0 / 54.2±4.8	54.8±5.7 / 55.2±6.0
optd	92.2±4.4 / 85.8±9.8	87.1±8.7 / 87.2±8.7	**93.5±5.6 / 87.9±11.3**
vehi	74.0±7.9 / 74.3±8.1	76.2±9.2 / **78.7±8.7**	74.7±8.5 / 76.2±9.8
wdbc	81.3±7.4 / 80.7±7.5	**84.3±6.6 / 84.4±6.3**	82.1±7.2 / **82.5±9.1**
digi	76.2±7.1 / 57.2±12.1	**80.7±3.9 / 84.2±4.4**	76.0±7.1 / **63.6±12.6**
USPS	78.5±2.9 / 80.0±0.1	72.2±3.7 / 71.6±4.9	78.7±2.7 / **80.1±0.2**
COIL	56.5±5.4 / 56.9±4.2	54.9±6.2 / 57.2±3.9	56.8±5.4 / 57.0±4.2
BCI	52.6±2.4 / 51.3±2.0	50.9±2.8 / 51.2±2.1	51.8±2.3 / 51.3±2.2
g241c	54.5±4.2 / 52.3±4.4	**78.3±4.5 / 59.5±3.4**	54.6±4.5 / **52.8±4.4**
g241n	56.4±5.2 / 52.5±5.3	53.3±7.3 / 52.8±5.2	56.3±5.1 / 52.7±5.1
Text	52.2±3.0 / 52.6±4.0	**64.8±7.9** / 52.4±6.1	52.1±2.9 / 52.6±4.0
Avg.Acc	71.3 / 69.7	72.2 / 71.3	71.8 / 70.6

100 Lab	SVM (linear / rbf)	TSVM (linear / rbf)	$S4VM_s$ (linear / rbf)
aust	83.7±1.8 / 78.7±2.9	83.1±2.1 / 78.6±2.8	83.7±1.6 / 78.8±3.0
ausl	79.5±2.8 / 80.6±2.3	79.0±2.9 / 81.1±2.5	**81.1±2.1 / 81.3±2.0**
brea	95.0±1.3 / 95.4±1.0	**95.7±0.7 / 95.8±0.7**	95.2±1.3 / 95.4±1.0
clea	73.3±3.1 / 83.1±2.0	73.6±2.9 / 83.2±2.2	73.5±2.9 / 83.3±2.1
diab	74.6±1.6 / 70.3±2.1	73.7±1.9 / 70.0±2.1	74.3±1.9 / 70.6±2.4
germ	65.3±3.0 / 70.9±1.0	**66.1±2.2** / 68.8±2.3	**65.8±2.9** / 71.0±1.1
habe	71.7±3.1 / 68.3±2.8	69.1±3.0 / 66.3±2.6	71.9±3.2 / 68.1±2.6
hear	82.2±2.5 / 76.3±3.4	82.2±2.9 / 76.0±3.4	82.4±2.5 / 76.5±3.7
hous	94.0±1.8 / 94.9±1.7	92.1±3.2 / 92.4±3.3	94.2±1.6 / 94.9±1.7
houv	91.1±1.4 / 92.5±1.7	89.9±2.2 / 90.9±2.4	91.2±1.3 / 92.6±1.6
iono	83.9±2.7 / 91.5±2.1	81.8±3.0 / 90.6±2.8	83.4±2.6 / 91.6±2.1
isol	99.0±0.4 / 99.2±0.5	96.3±3.3 / 96.4±3.4	98.9±1.2 / **99.3±0.5**
live	64.3±3.6 / 66.5±2.6	64.8±3.1 / 66.1±2.3	63.1±4.2 / 66.8±3.3
optd	99.2±0.3 / 99.5±0.2	96.2±3.3 / 96.2±3.3	99.1±1.2 / 99.6±0.4
vehi	93.6±1.9 / 97.7±1.0	93.2±2.0 / 96.0±2.1	93.7±2.0 / **97.9±0.8**
wdbc	95.2±1.4 / 93.6±1.7	93.4±2.6 / 92.4±2.7	94.7±1.7 / 93.6±1.7
digi	90.8±0.7 / 94.2±1.5	**92.0±1.6** / 94.5±1.9	**91.5±0.8 / 94.9±1.4**
USPS	87.2±1.0 / 83.1±1.9	86.7±1.5 / **91.7±2.5**	**87.7±1.0 / 91.0±2.4**
COIL	80.2±2.2 / 87.1±2.0	80.8±2.7 / 87.0±1.5	**80.8±2.3** / 87.2±2.1
BCI	70.4±3.4 / 66.2±2.9	70.5±4.1 / 65.4±2.8	70.5±3.4 / 66.1±2.9
g241c	74.5±2.0 / 70.1±8.5	**80.0±1.4 / 77.8±1.6**	**75.3±1.8 / 74.8±4.1**
g241n	71.8±2.7 / 59.4±9.4	**75.4±4.5** / 65.0±13.8	**72.2±2.8 / 60.9±8.4**
Text	69.8±2.1 / 54.4±4.9	**74.2±1.4 / 58.8±4.9**	69.9±2.0 / 54.1±4.1
Avg.Acc	82.4 / 81.8	82.3 / 82.0	82.5 / 82.5

TSVM often degenerate the performance
while S4VM does not significantly degenerate the performance.

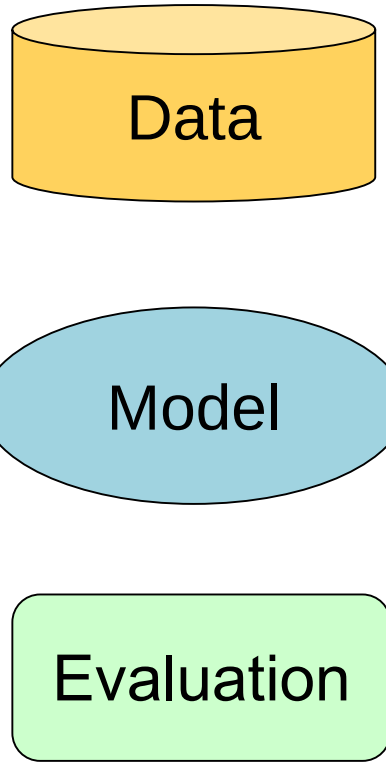
Hints

To alleviate the uncertainty of model selection

- ✓ Worst-case analysis helps
- ✓ Ensemble learning helps

Outline

- Graph structure
- Model selection
- Performance measure
- Conclusion



Variety of Performance Measures

- Previous efforts all focus on robustness of accuracy.
- However, real situations often require various performance measures.

For example

- In ranking applications
  - AUC
  - Top-k precision
- In text application
  - F1-Score
  - Precision-recall breakeven point
- In Information retrieval
  - Precision and recall
- …

Variety of Performance Measures

The robustness in accuracy is not equal to the robustness in other performance measures.

For example,

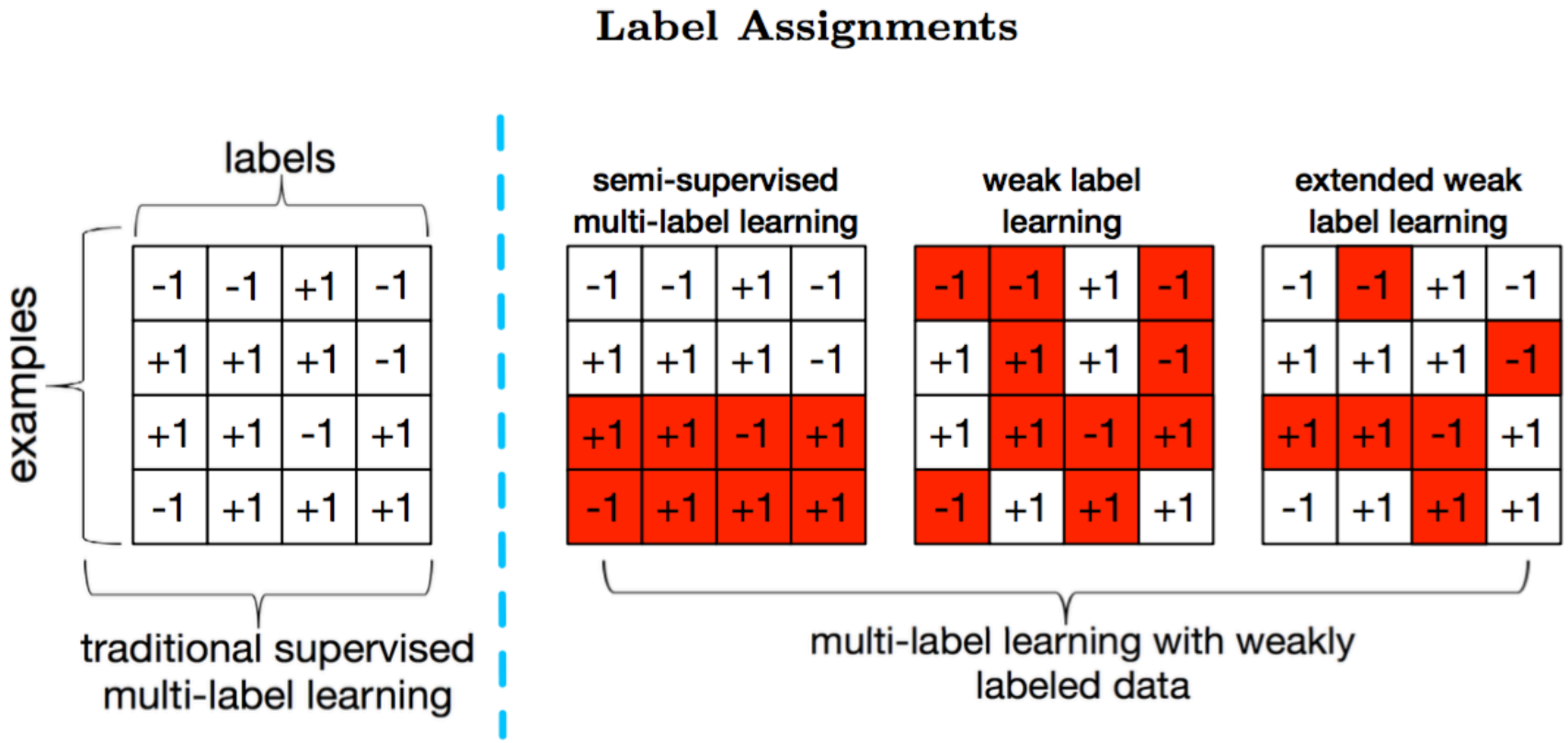
Doc ID	1	2	3	4	5	6	7	8	9	10	11
p	1	0	0	0	0	1	1	1	1	0	0
$rank(h_1(\mathrm{x}))$	11	10	9	8	7	6	5	4	3	2	1
$rank(h_2(\mathrm{x}))$	1	2	3	4	5	6	7	8	9	10	11

Hypothesis	MAP	Best Acc.
$h_1(q)$	0.56	0.64
$h_2(q)$	0.51	0.73

Reprinted from [Yue KDD07]

How to develop robust SSL methods for the adaption of various performance measures

Weakly Labeled Data

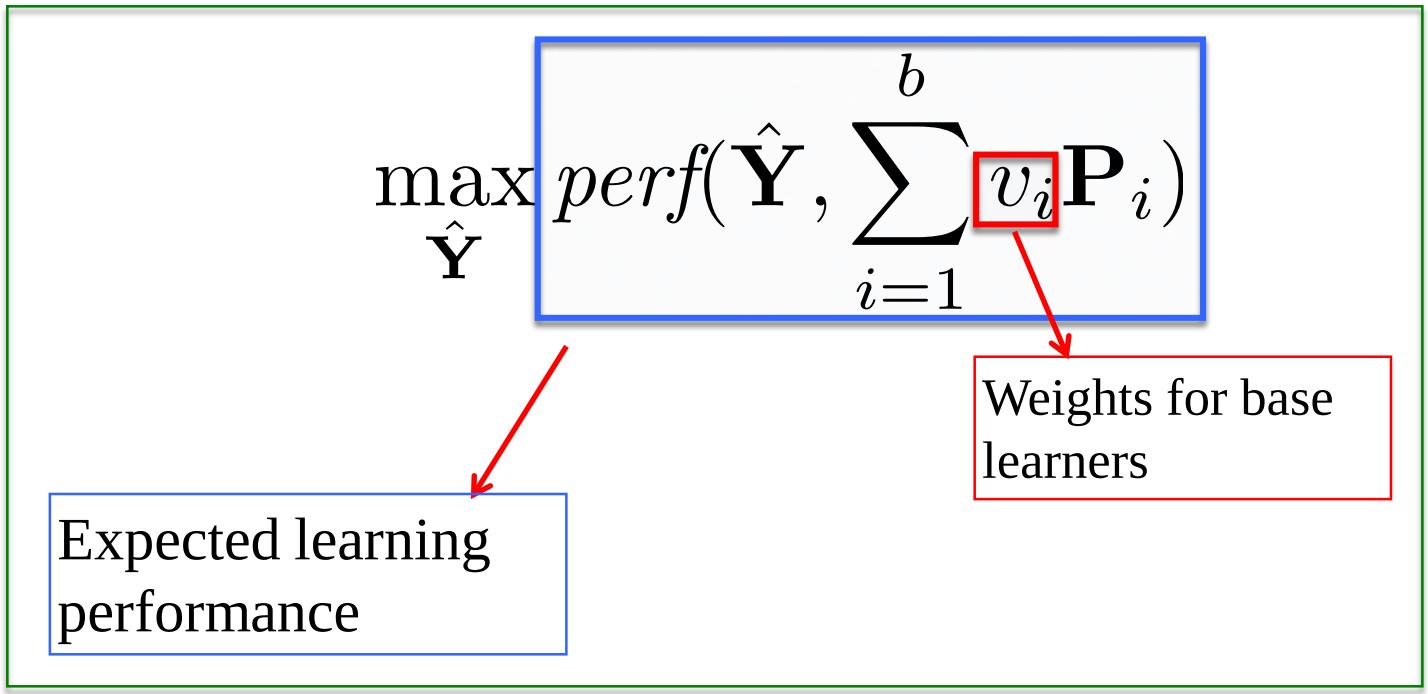


many kinds of weakly labeled data

- ✓ **semi-supervised multi-label learning**
- ✓ **weak label learning**
- ✓ **extended weak label learning**

A Direct Approach

Suppose the reliability or weights for these base learners, then one can have a very direct approach



However

The reliability for these base learners are hard to obtain

No additional domain knowledge for judging the quality of these base learners. We consider the worst case performance

SafeML

$$\max_{\hat{\mathbf{Y}}} \min_{\mathbf{Y}\in\Omega} \; perf(\hat{\mathbf{Y}}, \mathbf{Y})$$

$$\text{s.t. } \Omega = \left\{ \mathbf{Y} \middle| \mathbf{Y} = \sum_{i=1}^{b} v_i \mathbf{P}_i \right\}$$

Worst case consideration [Li and Zhou, ICML2011/TPAMI2015; Balsubramani and Freund, COLT2015]

$\mathcal{M} = \{\mathbf{v} | \sum_{i=1}^{b} v_i = 1, v_i \geq 0\}$ is a simplex

Performance Measure

We direct optimize the multi-label performance measure

✓ F1 score

$$F_1(TP, FP, TN, FN) = \frac{2TP}{2TP + FN + FP}.$$

✓ Top-k precision

$$Pre@k(\mathbf{y}_i, \hat{\mathbf{y}}_i) = \frac{1}{k} \sum_{l \in \text{rank}_k(\hat{\mathbf{y}}_i)} (\mathbf{y}_{i,l} + 1)/2$$

Optimize F1 Performance

Substitute the definition of F1 measure, the objective function becomes

$$\max_{\hat{\mathbf{Y}}} \min_{\mathbf{Y}\in\Omega} \frac{\mathrm{tr}\left(\left(\frac{\hat{\mathbf{Y}}+1}{2}\right)^{\top}\left(\frac{\mathbf{Y}+1}{2}\right)\right)}{\sum_{i,j}\mathbb{I}(\mathbf{Y}_{i,j}=1)+\sum_{i,j}\mathbb{I}(\hat{\mathbf{Y}}_{i,j}=1)}$$

Assume # of labels each instance closes to a prior ratio

$$\max_{\hat{\mathbf{Y}}} \min_{\mathbf{Y}\in\Omega} \mathrm{tr}\left(\left(\frac{\hat{\mathbf{Y}}+1}{2}\right)^{\top}\left(\frac{\mathbf{Y}+1}{2}\right)\right)$$

$$\text{s.t.} \left|\sum_{i,j}\mathbb{I}(\mathbf{Y}_{i,j}=1)-\gamma_0\right| \le \epsilon, i=1\cdots N,\ j=1\cdots L$$

$$\left|\sum_{i,j}\mathbb{I}(\hat{\mathbf{Y}}_{i,j}=1)|-\gamma_0\right| \le \epsilon, i=1\cdots N,\ j=1\cdots L$$

Optimize F1 Performance

$$\max_{\hat{\mathbf{Y}},\theta} \ \theta$$

$$\text{s.t.} \ \ \theta \leq \text{tr}\left((\frac{\hat{\mathbf{Y}}+1}{2})^{\top} (\frac{\mathbf{Y}+1}{2}) \right), \forall \, \mathbf{Y} \in \Omega$$

$$\left| \sum_{i,j} \mathbb{I}(\mathbf{Y}_{i,j} = 1) - \gamma_0 \right| \leq \epsilon, \ i = 1 \cdots N, \ j = 1 \cdots L$$

$$\left| \sum_{i,j} \mathbb{I}(\hat{\mathbf{Y}}_{i,j} = 1) - \gamma_0 \right| \leq \epsilon, \ i = 1 \cdots N, \ j = 1 \cdots L$$

An exponential number of constraints; we employ the **cutting-plane algorithm** [Li et al., JMLR13] to solve this problem

Top-k precision optimization can be solved in a similar way

Analysis

Theorem 1 *Let* $\boldsymbol{Y}^{GT}$ *be the ground-truth label matrix and* $\hat{\mathbf{Y}}^{*}$ *be the prediction of* SafeML, *i.e., the optimal solution to Eq. (5). The performance of our proposal* $\text{perf}(\hat{\mathbf{Y}}^{*}, \boldsymbol{Y}^{GT})$ *w.r.t.* F_1 *score and Top-k precision is lower bounded by* $\max_{i=1,\ldots,b} \min_{j=1,\ldots,b} \text{perf}(\boldsymbol{P}_i, \boldsymbol{P}_j)$ *as long as* $\boldsymbol{Y}^{GT} \in \Omega$.

Data set	the lower bound in Theorem 1	direct binary relevance SVM
emotions	0.774	0.539
enron	0.194	0.076
image	0.378	0.105
scene	0.739	0.422
yeast	0.501	0.318

Theorem 1 shows the performance of SafeML is related to the maximin correlation of base learners.

In practice, as shown in above Table, it is often much larger than direct supervised multi-label learning with only labeled data.

Experimental Setting

- 9 data sets from a broad range of properties
- 8 competing methods
  - Baseline method: Binary Relevance (BR) [Tsoumakas et al, 2009]
  - S4VM [Li and Zhou, 2015]
  - ML-kNN [Zhang and Zhou, 2007]
  - ECC [Read et al, 2011]
  - CNMF [Liu et al, 2006]
  - LEML [Yu et al., 2014]
  - TRAM [Kong et al., 2013]
  - WELL [WELL et al., 2010]
- 3 evaluation metrics
  - Macro F1, Micro F1, Top-k precision

Macro F1 Setting on SSML setting

Macro-F1 score								
Data set	BR	S4VM	ECC	ML-kNN	CNMF	LEML	TRAM	SafeML
emotions	0.539	**0.608**	**0.589**	0.489	0.330	0.417	**0.586**	**0.624**
enron	0.076	0.082	0.083	0.067	**0.092**	**0.098**	**0.123**	**0.113**
image	0.105	**0.509**	**0.280**	**0.401**	**0.271**	**0.511**	**0.532**	**0.516**
scene	0.422	**0.702**	**0.596**	**0.617**	0.315	**0.567**	**0.684**	**0.657**
yeast	0.318	**0.405**	**0.346**	0.307	0.257	0.183	**0.355**	**0.408**
arts	0.075	**0.093**	**0.107**	0.068	**0.129**	**0.131**	**0.168**	**0.136**
bibtex	0.185	**0.204**	**0.247**	0.031	0.179	0.112	**0.229**	**0.272**
tmc2007	0.443	0.452	**0.474**	0.220	0.138	0.274	0.384	**0.475**
Ave. Perf.	0.279	0.381	0.340	0.275	0.214	0.286	0.383	0.408
win/tie/loss against BR		6/2/0	**7/1/0**	2/2/4	3/1/4	4/0/4	7/0/1	**8/0/0**

SafeML achieves highly competitive performance

Directly using current MLL methods often degenerate performance while SafeML does not degenerate performance

Top-k Precision on Weak Label setting

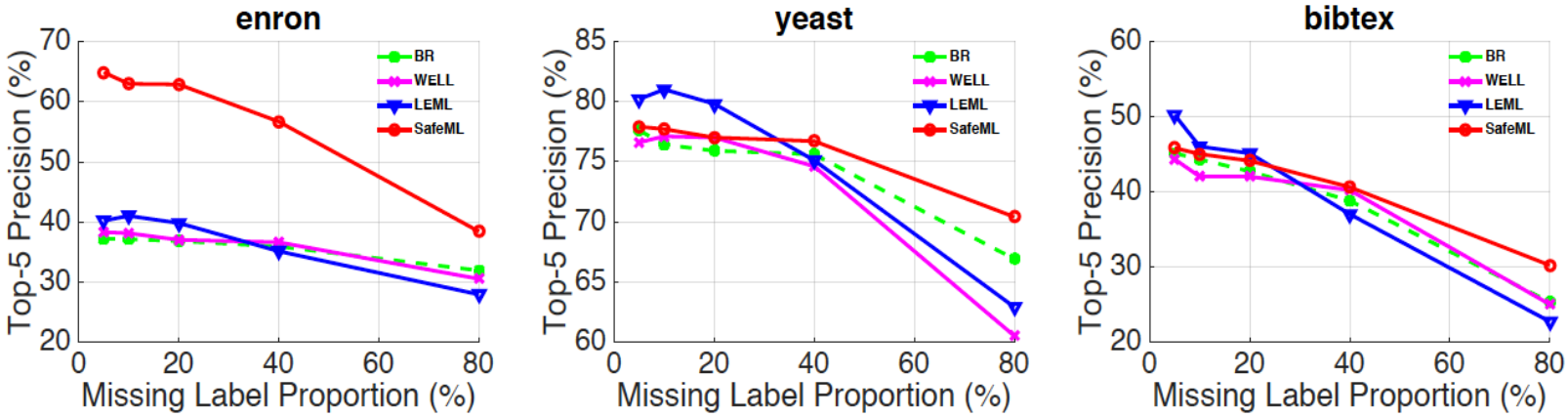


Fig. 4: Top-5 precision on weak label learning

SafeML obtains highly competitive performance and does not significantly hurt performance with the use of weak label

Top-k Precision on Extended Weak Label setting

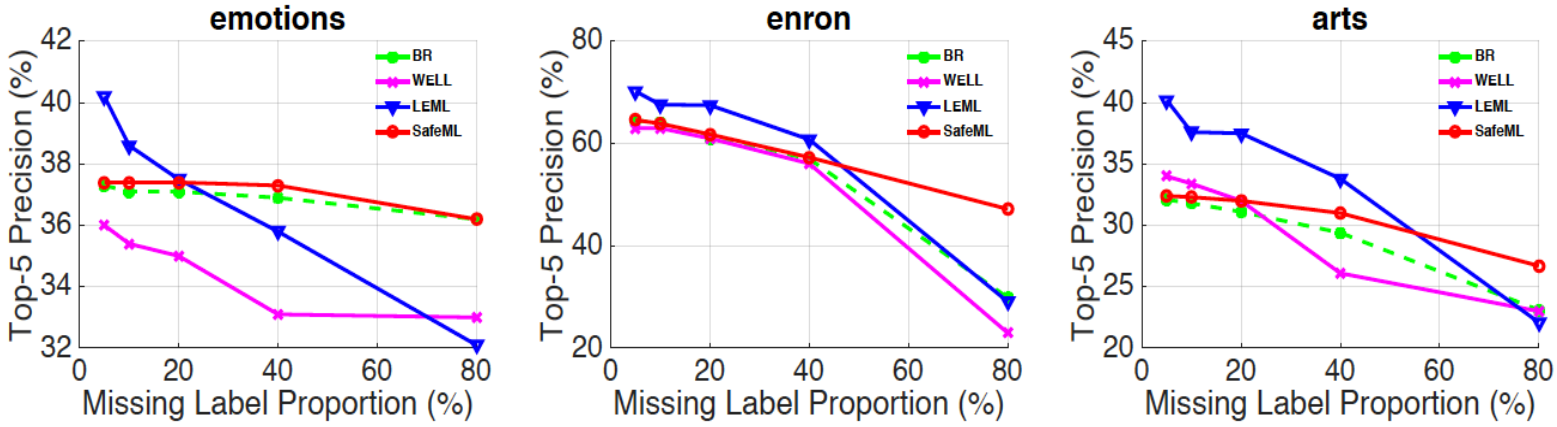


Fig. 5: Top-5 precision on extended weak label learning

We have similar observations on extended weak label learning setting

Hints

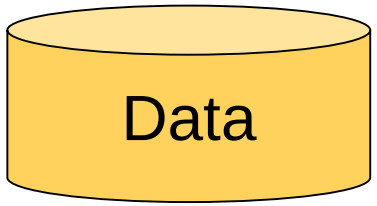
To safe semi-supervised learning under complicated

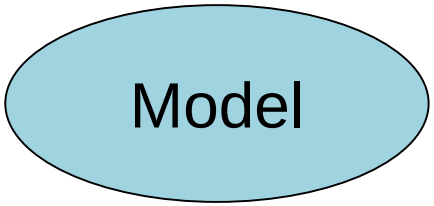
performance measure

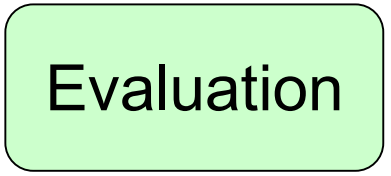
- ✓ Worst-case analysis helps
- ✓ Ensemble learning + accurate optimization helps

Outline

- Graph structure
  - LEAD [Li et al., IJCAI2016]
- Model selection
  - S4VM [Li and Zhou, TPAMI2015]
  - Safer [Li, Zha and Zhou, AAAI2017]
  - SafeW [Guo and Li, AAAI2018]
- Performance measure
  - SafeML [Tong, Guo and Li, MLJ2018]
  - UMVP [Li et al., AAAI2016]
- Conclusion



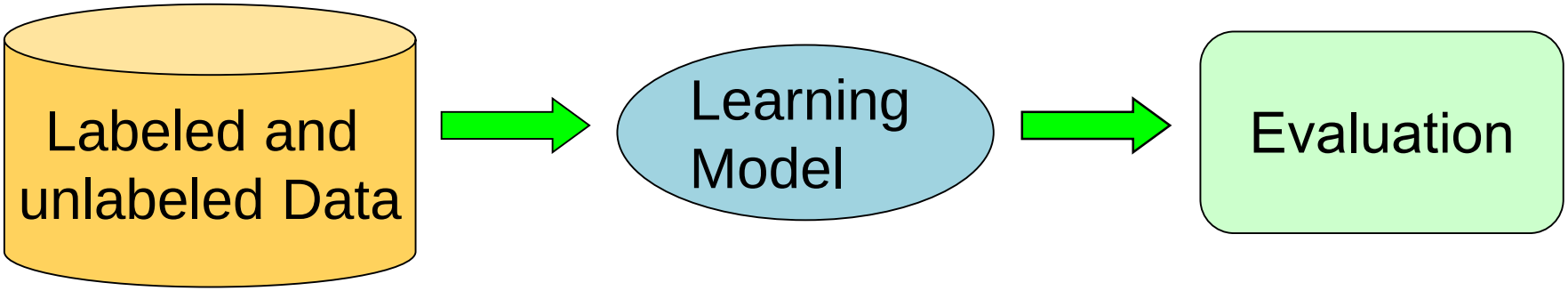


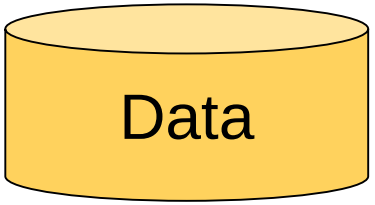


Summary

Unlabeled data is of important value in data analytics

How to safely exploit unlabeled data has become a crucial issue





Graph quality

- ✓ Large margin principle helps
- ✓ Ensemble learning + large margin principle helps

Summary

Model selection

- ✓ Worst-case analysis helps
- ✓ Ensemble learning helps

Model

Performance measure

- ✓ Worst-case analysis helps
- ✓ Ensemble learning + accurate optimization helps

Evaluation

Summary

Codes

- S4VM http://lamda.nju.edu.cn/code_S4VM.ashx
- LEAD http://lamda.nju.edu.cn/code_LEAD.ashx
- SAFER http://lamda.nju.edu.cn/code_SAFER.ashx
- SAFEW http://lamda.nju.edu.cn/code_SAFEW.ashx
- SAFEML http://lamda.nju.edu.cn/code_SAFEML.ashx

Joint work with

Zhi-Hua Zhou, James Kwok, Ivor Tsang, Shao-Bo Wang, Tong Wei, Lan-Zhe Guo

Discussions

The study on safely exploiting unlabeled data is young

Many things unknown.

- Safe representation ?
- Systematical approach? Uniform framework?
- Theoretical support is still not solid enough

Thanks!