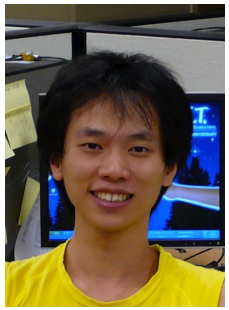
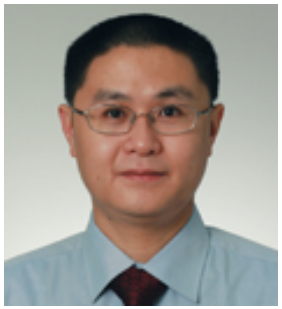
NANYANG TECHNOLOGICAL UNIVERSITY

# Co-Labeling: A New Multi-view Learning Approach for Ambiguous Problems

*presented by Wen Li, 12-Dec-2012*



**Wen Li, Lixin Duan, Ivor W.-H. Tsang, and Dong Xu**
*Centre for Multimedia and Networks (CeMNet)*
*School of Computer Engineering,*
*Nanyang Technological University, Singapore*

# Outline

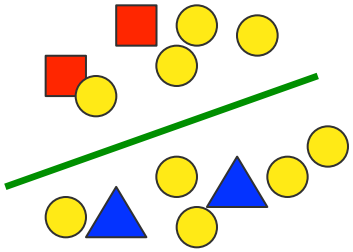
- Motivations
- Problem
  - Ambiguous problem
  - Multi-view ambiguous problem
- Solution
  - The co-labeling algorithm
- Experimental results
  - Documents/Webpage Classification
  - Web Image Retrieval

# Motivations

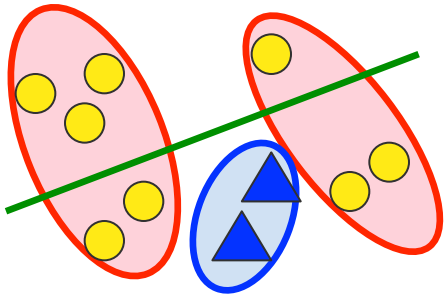
- Data are cheap but labeling them is expensive.
  - It is easy to collect a mass of images from the web, but is hard to label all of them.
  - A lot of learning models have been proposed to cope with less supervision, such as semi-supervised learning, multiple instance learning and clustering.
- Data are usually represented in multiple forms.
  - Different features can be easily extracted from an image, such as SIFT, HOG, LBP, etc.
  - Multi-view of features can enhance the performance and help us to reduce the supervision (for example, co-training).

# Learning with Less Supervision

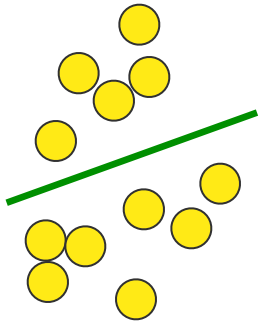
**Semi-supervised Learning (SSL)**



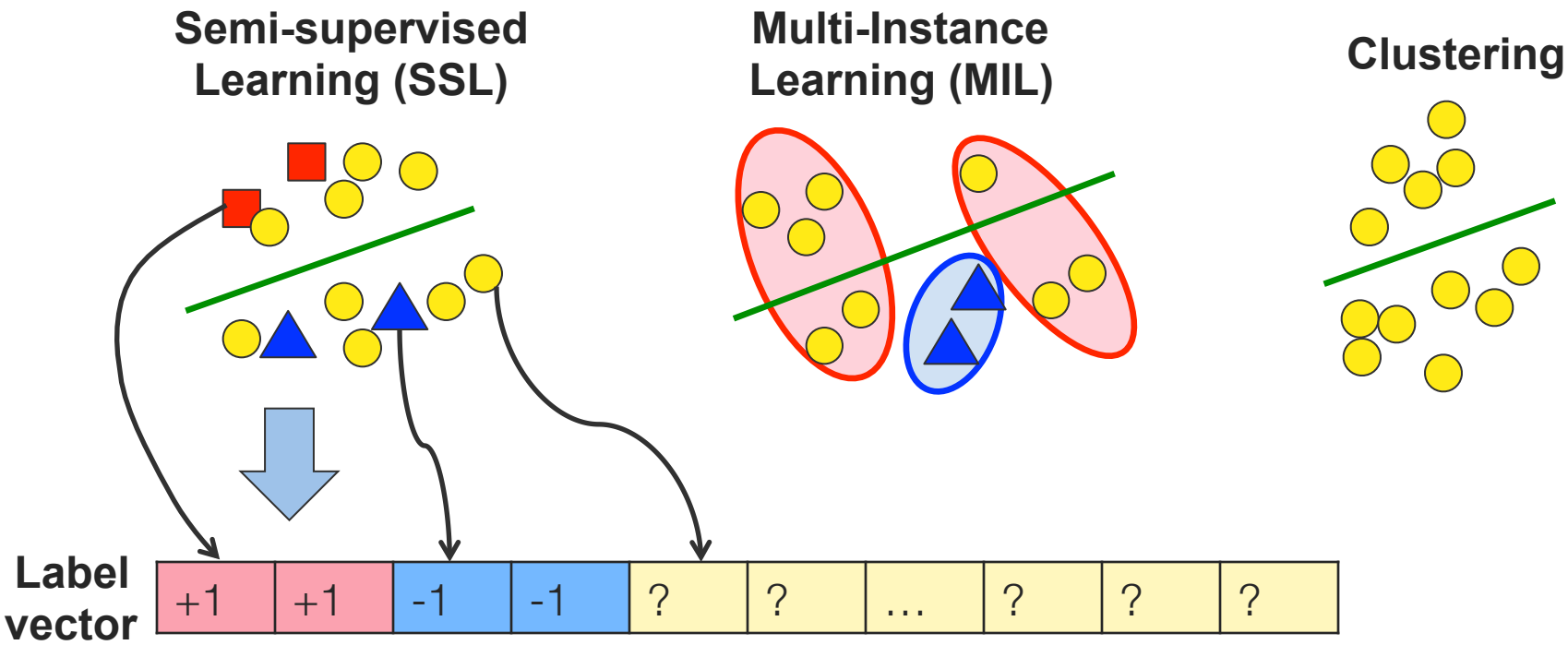
**Multi-Instance Learning (MIL)**



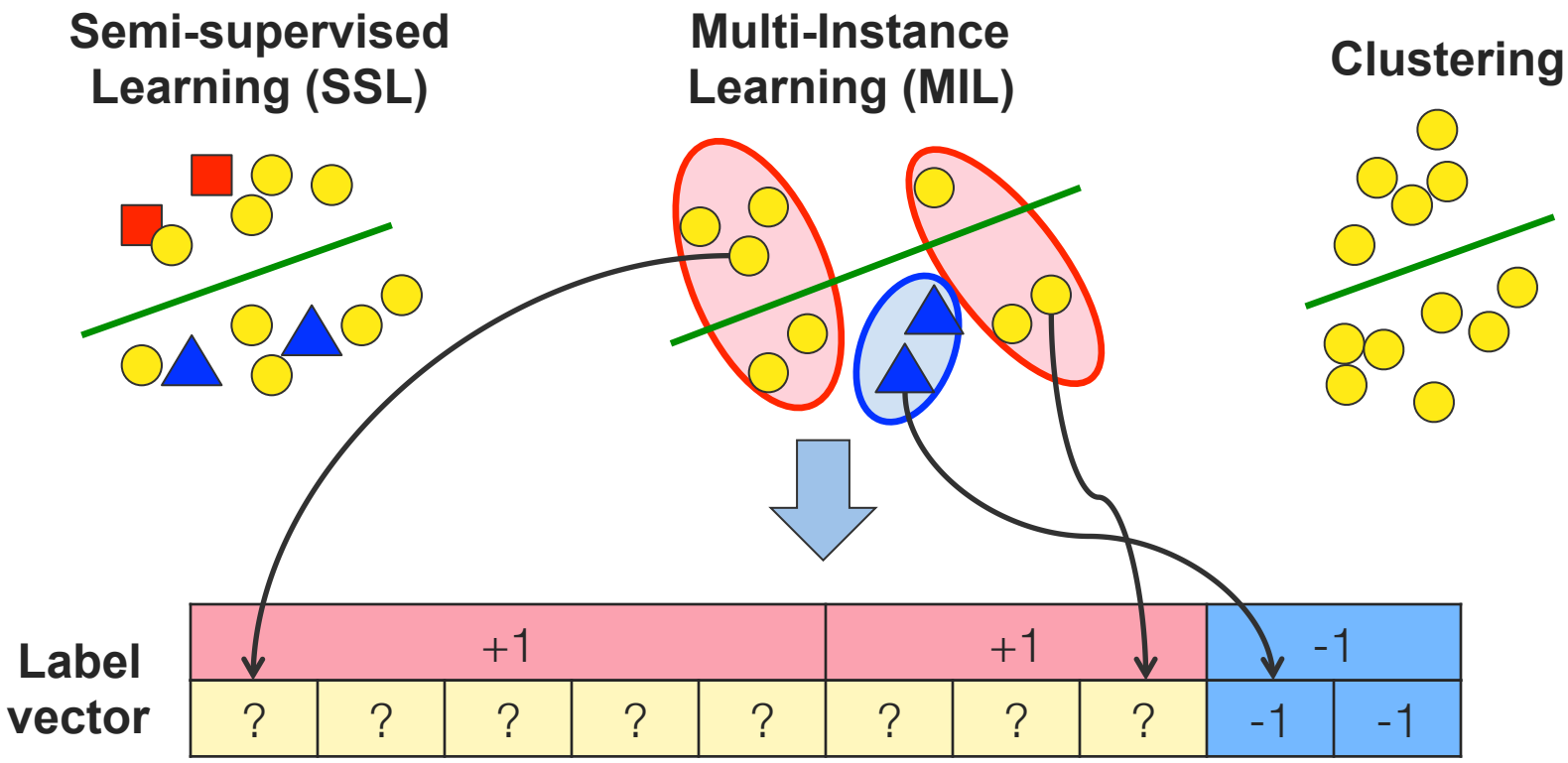
**Clustering**



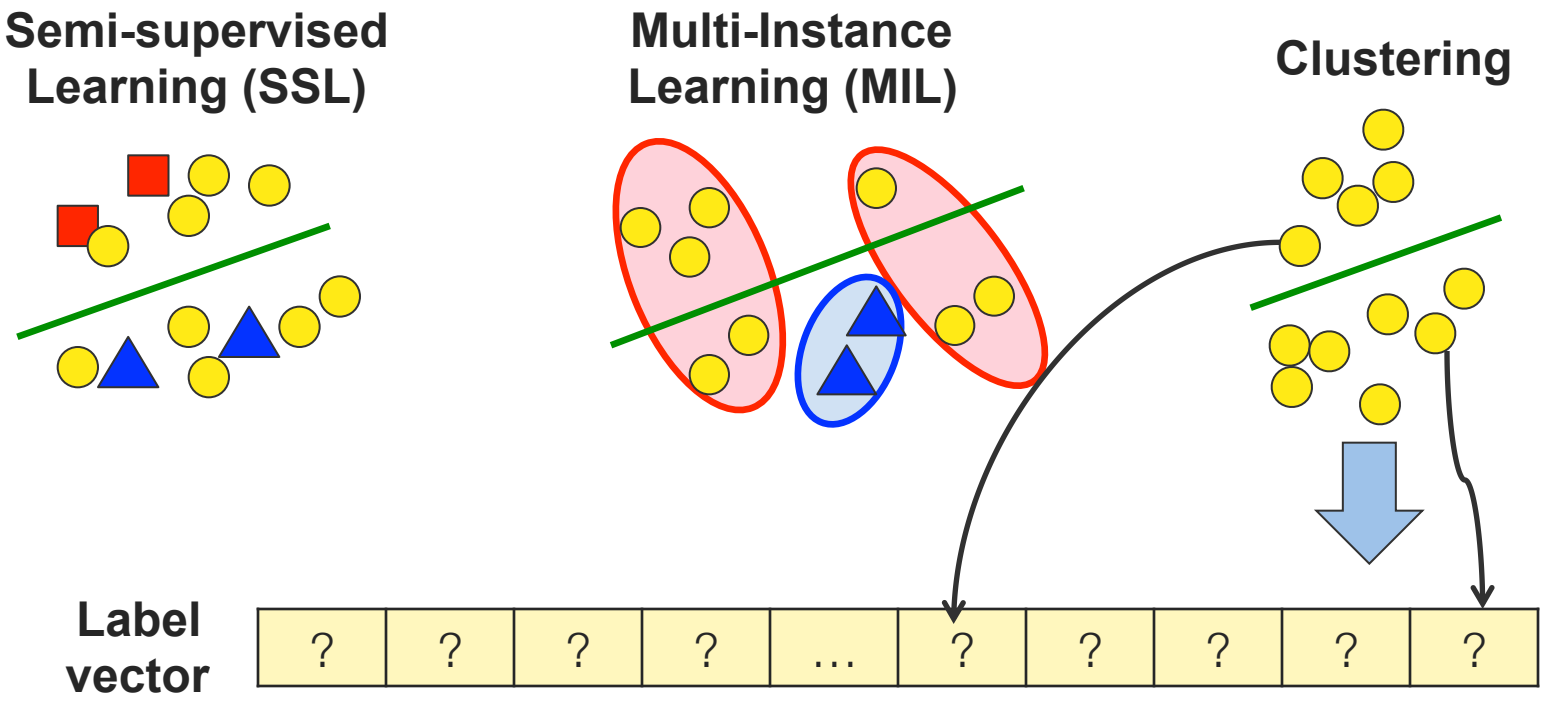
# Learning with Less Supervision



# Learning with Less Supervision

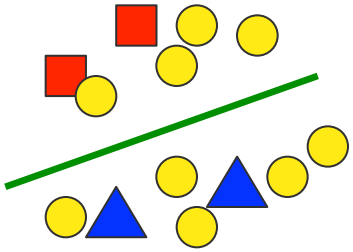


# Learning with Less Supervision

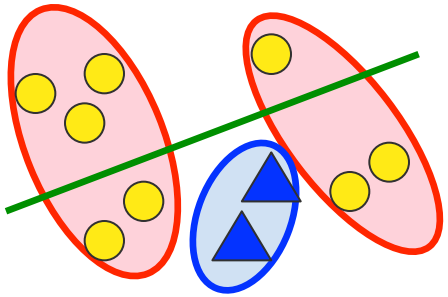


# Learning with Less Supervision

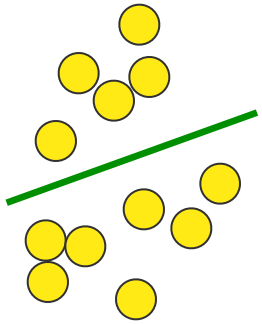
**Semi-supervised Learning (SSL)**



**Multi-Instance Learning (MIL)**



**Clustering**



**SSL**

| +1 | +1 | -1 | -1 | ? | ? | … | ? | ? | ? |
|---|---|---|---|---|---|---|---|---|---|

**MIL**

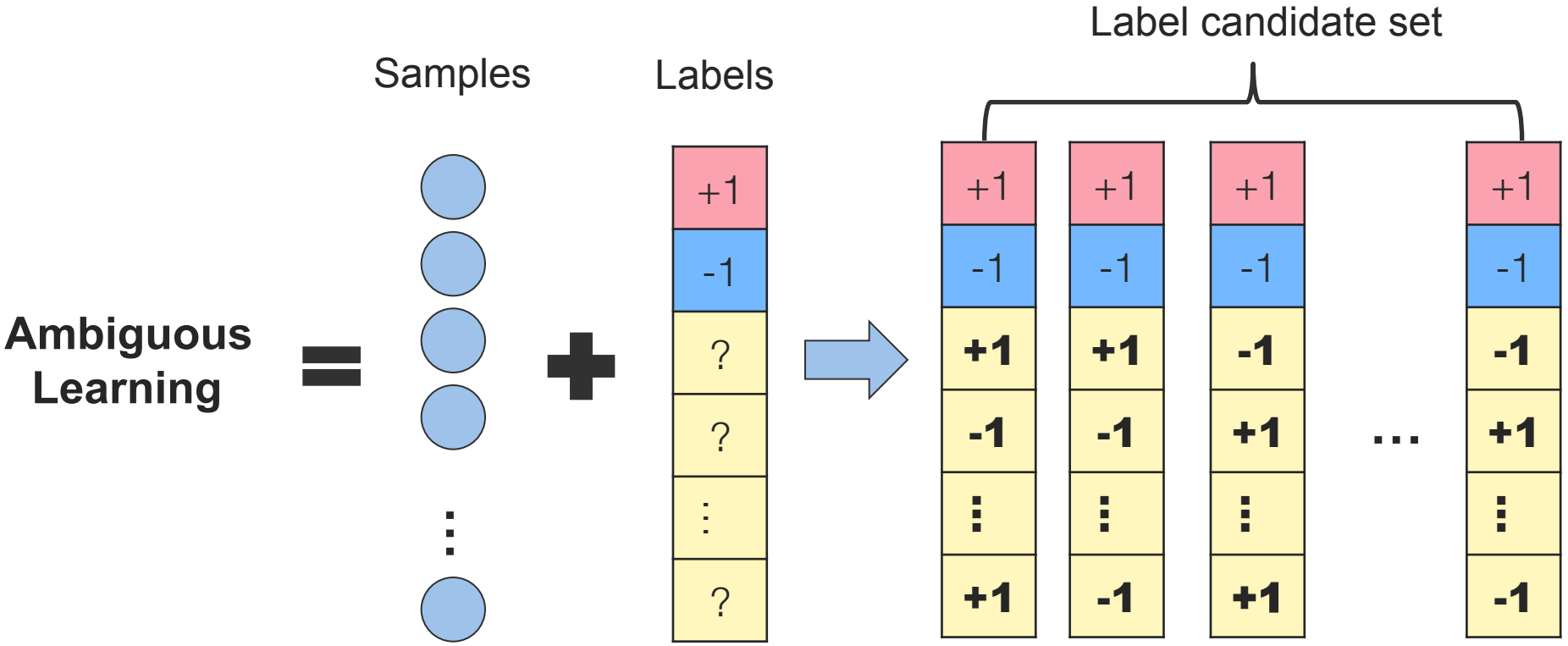
| ? | ? | ? | ? | ? | ? | ? | ? | -1 | -1 |
|---|---|---|---|---|---|---|---|---|---|

**Clustering**

| ? | ? | ? | ? | … | ? | ? | ? | ? | ? |
|---|---|---|---|---|---|---|---|---|---|

NANYANG TECHNOLOGICAL UNIVERSITY

# Ambiguous Learning




- Ambiguous learning is to learn from some training samples and a set of label candidates.

# Ambiguous Learning: Formulation

- Based on the regularized empirical risk minimization principle:

$$\min_{f,\mathbf{y}\in\mathcal{Y}} \|f\|^2 + C\sum_{i=1}^{n} l(f,\mathbf{x}_i,y_i)$$

- $f$ is the target classifier, $l(.)$ is the loss function.
- $\mathbf{y}$ is a label candidate, and $\mathcal{Y}$ is the label candidate set:
  - Semi-supervised Learning (SSL):

$$\mathcal{Y} = \{\mathbf{y} | y_i = g_i, i = 1,\ldots,l; \sum_{i=l+1}^{n} y_i = \sigma\}$$

  - Multiple Instance Learning (MIL)

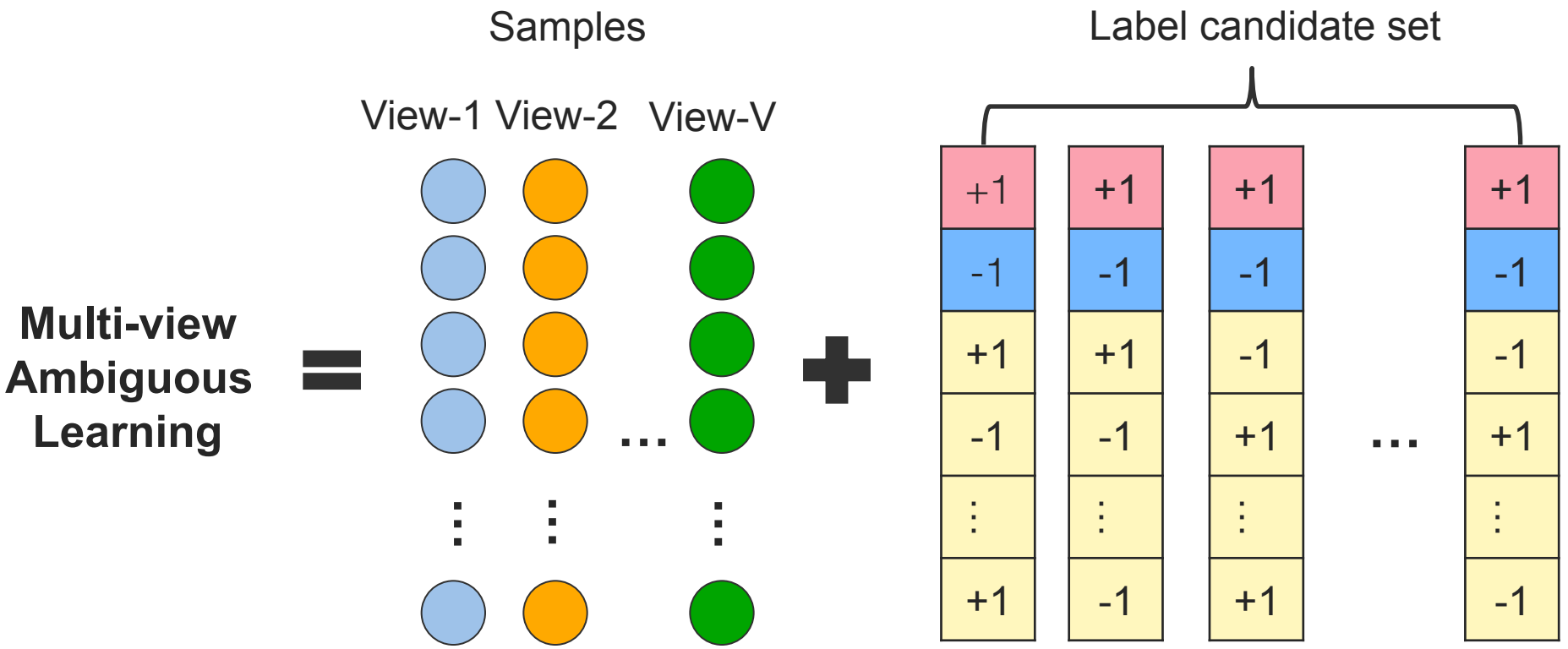
$$\mathcal{Y} = \{\mathbf{y} | \sum_{\mathbf{x}_i\in\mathcal{B}_I} \frac{y_i+1}{2} \geq 1, \text{if } Y_I = 1; y_i = -1, \text{otherwise}\}$$

  - Clustering:

$$\mathcal{Y} = \{\mathbf{y} | \sum_{i=1}^{n} y_i = \sigma\}$$

NANYANG TECHNOLOGICAL UNIVERSITY

# Multi-view Ambiguous Learning




- Multi-view ambiguous learning is to learn from multi-view training samples and a set of label candidates.
- Multi-view of features can enhance the performance and help to reduce the ambiguities.

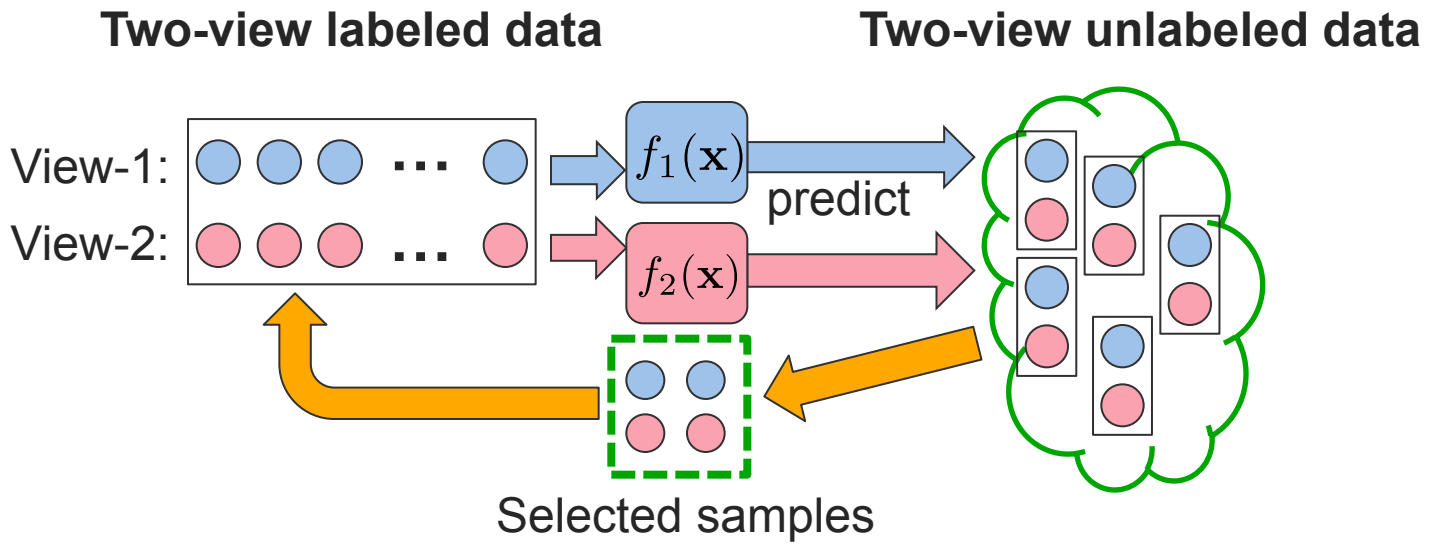
# Multi-view AL: Formulation

$$\min_{f^v, \mathbf{y}^v \in \mathcal{Y}^v} \sum_{v=1}^{V} \left( \|f^v\|^2 + C \sum_{i=1}^{n} l(f^v, \mathbf{x}_i^v, y_i^v) \right)$$

- Terms:
  - $f^v$ is the classifier on the $v$-th view
  - $\mathcal{Y}^v$ is a *small label candidate set* on the $v$-th view
- Key problem
  - How to construct a small label candidate set for each view.

# Co-Labeling:
# A new multi-view ambiguous learning approach

# Review of Co-training: Feed samples

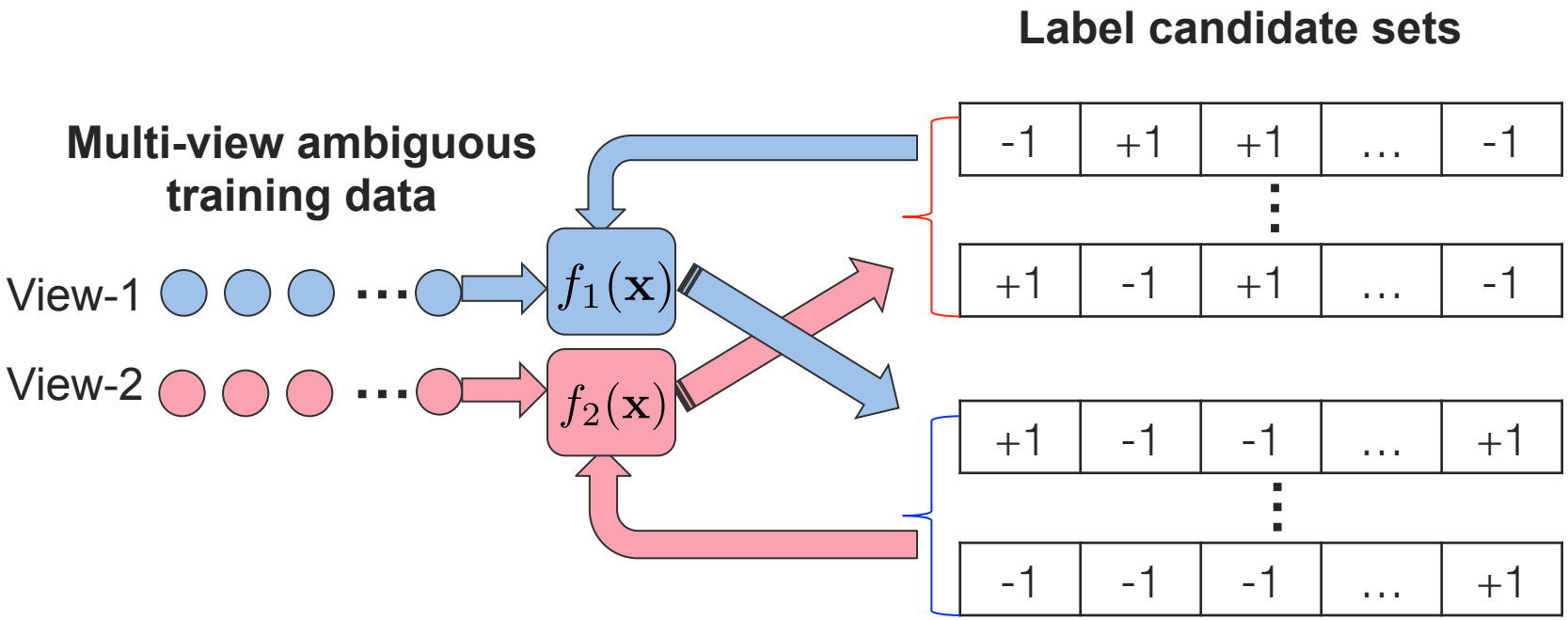



1. Training two classifier using labeled data on two views,
2. Predict the unlabeled data, and select a fixed number of samples which are *confident in one view* but *unconfident in the other view*.
3. Label the selected samples and merge them into the labeled set, and then retrain the classifiers.
4. Repeat the above 3 steps.

# Review of Co-training: Feed samples

- Highlight:
  - Using the classifier on one view to enhance the classifier on the other view by feeding samples.
- Limitations:
  - The selecting of samples cannot be applied to the training data associated with structures (MIL).
  - If the selected samples are incorrectly labeled, it may do harm to the classifiers trained in the following iterations.

# Co-Labeling: Feed the labeling




1. Training two classifiers on two views,
2. Predict the ambiguous training data.
3. Update the label candidate set by using the predictions (decision values on training data) from other views.
4. Repeat the above 3 steps.

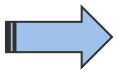
# Co-Labeling: Three Strategies to construct the label candidate set

**Strategy 1:** $\mathcal{Y}_{t+1}^{v} = \bigcup_{p \neq v}^{V} o_t^p$ *where* $o_t^p$ *is obtained by projecting the decision value from the p-th view* (i.e., $z^p$) *into the feasible set* $\mathcal{Y}$ *defined by the constraints on the ambiguous training samples.*

**Decision values from $f_2$**

| | | | | |
|---|---|---|---|---|
| -1.04 | 0.98 | 0.87 | … | -0.93 |

Projection



**Label candidate set on view-1**

| | | | | |
|---|---|---|---|---|
| -1 | +1 | +1 | … | -1 |

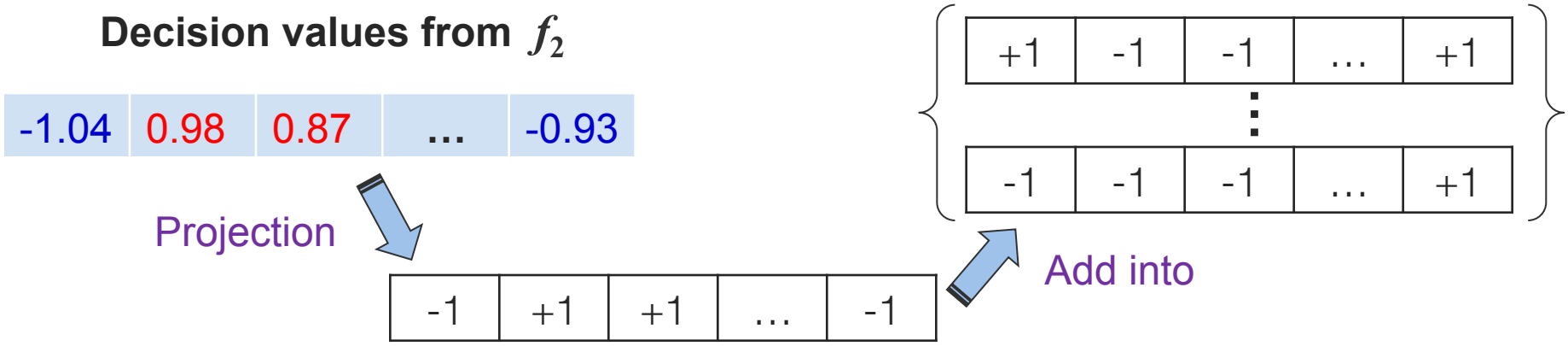
- The label candidate set is generated by using the prediction from classifier of another view, which is consistent with the philosophy that using one view to help another.
- The projection operation makes the label candidate to satisfy the constraints.
- The projection operation only needs to rank the decision values, which is very efficient.

# Co-Labeling: Three Strategies to construct the label candidate set

**Strategy 2:** $\mathcal{Y}_{t+1}^{v} = (\bigcup_{p \neq v}^{V} o_t^p) \bigcup \mathcal{Y}_t^v$ *where* $o_t^p$ *is obtained in the same manner as in Strategy 1.*

**Label candidate set on view-1**

**Decision values from** $f_2$




- If one sample is miss-labeled at one iteration, then it may be corrected by the label candidates obtained from other iterations.

NANYANG TECHNOLOGICAL UNIVERSITY

# Co-Labeling: Three Strategies to construct the label candidate set

**Strategy 3:** $\mathcal{Y}_{t+1}^{v} = (\bigcup_{p \neq v}^{V} \mathcal{O}_t^p) \bigcup \mathcal{Y}_t^v$ *where* $\mathcal{O}_t^p$ *is a set of label candidates obtained in the same manner as in Strategy 1 from the predictions with different biases.*

**Label candidate set on view-1**

**Decision values from** $f_2$

| | | | | |
|---|---|---|---|---|
| -1.04 | 0.98 | 0.87 | … | -0.93 |

Projection by varying biases

| | | | | |
|---|---|---|---|---|
| -1 | +1 | +1 | … | -1 |
| -1 | -1 | -1 | … | +1 |

Add into

| | | | | |
|---|---|---|---|---|
| +1 | -1 | -1 | … | +1 |
| -1 | -1 | -1 | … | +1 |

- Handle the possible bias problems caused by imbalanced training data, inaccurate initialization, and so on.

# Co-Labeling: Detailed Formulation

- Multi-view ambiguous learning:

$$\min_{f^v, \mathbf{y}^v \in \mathcal{Y}^v} \sum_{v=1}^{V} \left( \|f^v\|^2 + C \sum_{i=1}^{n} l(f^v, \mathbf{x}_i^v, y_i^v) \right)$$

- Based on the rho-SVM and squared hinge loss:

$$\min_{\mathbf{y}^v \in \mathcal{Y}^v} \min_{\mathbf{w}^v, b^v, \rho^v, \xi_i} \frac{1}{2} \left( \|\mathbf{w}^v\|^2 + b^{v\,2} + |C \sum_{i=1}^{n} \xi_i^2 \right) - \rho^v,$$

$$\text{s.t.} \quad y_i^v (\mathbf{w}^{v'} \phi(\mathbf{x}_i^v) + b^v) \geq \rho^v - \xi_i, \; i = 1, \ldots, n,$$

- Terms:
  - The classifier: $f^v(\mathbf{x}^v) = w^{v\prime} \mathbf{x}^v + b^v$

NANYANG TECHNOLOGICAL UNIVERSITY

# Co-Labeling: An MKL Solution

- We write the dual form as:

$$\min_{\mathbf{y}\in\mathcal{Y}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} -\frac{1}{2}\boldsymbol{\alpha}'\left(\mathbf{K}\circ\mathbf{y}\mathbf{y}'+\frac{1}{C}\mathbf{I}\right)\boldsymbol{\alpha},$$

- Convex relaxation by using the linear combination of label candidates, which results in an MKL problem.

$$\min_{\mathbf{d}\in\mathcal{D}} \max_{\boldsymbol{\alpha}\in\mathcal{A}} -\frac{1}{2}\boldsymbol{\alpha}'\left(\sum_{t=1}^{|\mathcal{Y}|} d_t\mathbf{K}\circ\mathbf{y}_t\mathbf{y}_t'+\frac{1}{C}\mathbf{I}\right)\boldsymbol{\alpha}$$

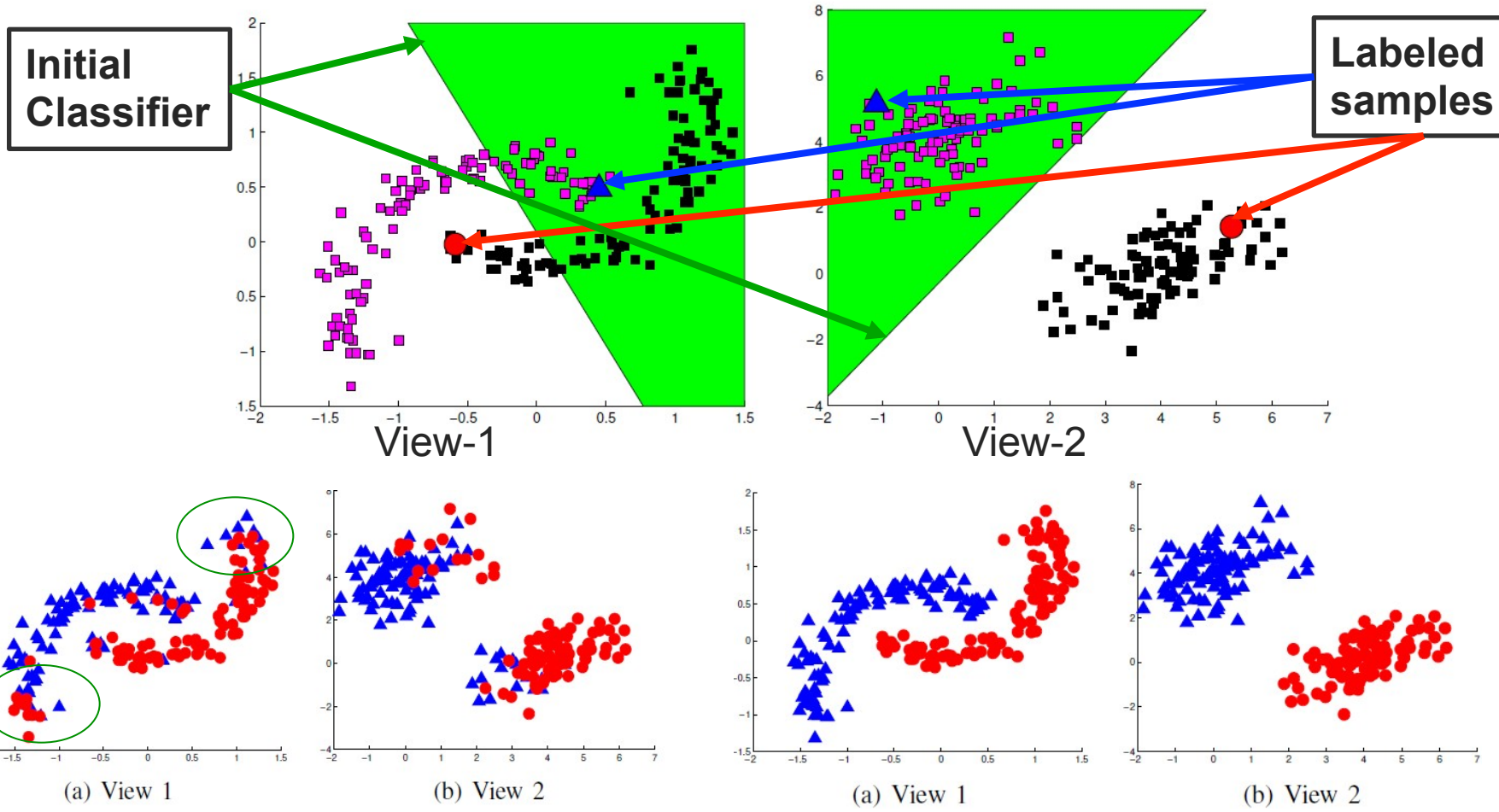
- Final classifier:

$$f(\mathbf{x})=\sum_{v=1}^{V} f^v(\mathbf{x}^v)=\sum_{v=1}^{V}\frac{1}{\rho_v}\left(\sum_{i=1}^{n}\alpha_i^v\sum_{t=1}^{|\mathcal{Y}^v|} d_t^v y_{t,i}^v\left(k(\mathbf{x}_i^v,\mathbf{x}^v)+1\right)\right)$$

# Co-Labeling: The Algorithm

- We summarize the algorithm as follows:
  - Initialize the label candidate set for each view.
  - Repeat: (for each view)
    - Solve the MKL problem.
    - Use the learnt classifier to predict the training samples.
    - Obtain a set of label candidates for each view by projecting decision values from other views into feasible labeling set with different biases.
    - Add the new label candidates into the label candidate set and retrain the classifiers.
  - Until the stop criterion is reached.

# Co-Labeling vs Co-Training




NANYANG TECHNOLOGICAL UNIVERSITY

# Experimental Results

- Document/Webpage Classification (SSL)
  - Dataset: WebKB
  - BBC, BBCSports
- Image Retrieval (MIL)
  - NUS-WIDE dataset.

# Experiments: Document Classification

| | BBC | | | BBCSport | | |
|---|---|---|---|---|---|---|
| | View1 | View2 | View1+2 | View1 | View2 | View1+2 |
| SVM | 66.53(4.08) | 63.11(3.67) | 74.26(3.37) | 70.69(3.42) | 66.43(3.98) | 76.99(3.76) |
| TSVM | 71.99(5.48) | 66.83(3.54) | 75.72(3.16) | 74.62(5.73) | 65.51(3.36) | 79.21(6.43) |
| Co-LapSVM | 70.30(3.39) | 68.04(4.56) | 76.97(3.41) | 70.70(3.43) | 66.43(3.99) | 77.14(3.29) |
| 2V-TSVM | 52.70(3.96) | 52.61(5.34) | 58.39(5.25) | 64.00(3.08) | 63.50(3.86) | 69.82(3.78) |
| PMC | — | — | 71.57(6.37) | — | — | 79.48(5.41) |
| Co-Labeling | **78.41(3.79)**↑ | **77.61(3.01)** ↑ | **81.37(3.14)** ↑ | **82.10(5.41)** ↑ | **79.60(4.44)** ↑ | **84.22(5.11)** ↑ |

- On BBC and BBCSport (MAP):
  - Co-Labeling is significantly better than other methods on the combined results as well as on each view.
- On WebKB (PRBEP):
  - Co-Labelling also gets the best combined result.
  - The improvement is not significant possibly because it is already a very high performance (it is 99.11% in the measurement of MAP)

| | WebKB | | |
|---|---|---|---|
| | page | link | page+link |
| SVM | 74.4 | 77.8 | 84.4 |
| TSVM | 85.5 | 91.4 | 92.2 |
| Co-LapSVM | **94.3** | **93.3** | 94.2 |
| 2V-TSVM | 85.7 | 86.7 | 87.3 |
| PMC | | | 88.6 |
| Co-Labeling | 92.5 | 93.1 | **95.1** |

NANYANG TECHNOLOGICAL UNIVERSITY

# Experiment: Web Image Retrieval

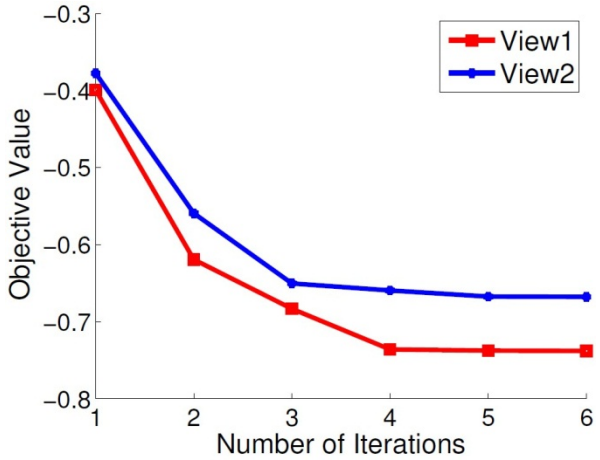
- On NUS-WIDE (MAP)
  - View-1: text + global visual features (color, etc.)
  - View-2: text + SIFT feature with LLC coding.

| | TG | TL | TG+TL |
|---|---|---|---|
| MIL-CPB | 61.43 | 57.84 | 77.07 |
| mi-SVM | 59.25 | 59.26 | 77.18 |
| sMIL | 60.01 | 62.09 | 75.48 |
| Co-Labeling | **62.56** | 61.71 | **79.09** |

  - Our method also gets the best result.

# Experiments: Convergence & Time

- Convergence




Converge fast.
Usually no more than 10 rounds.

- Comparison of training time

| | BBC | BBCSport | WebKB |
|---|---|---|---|
| Co-LapSVM | 52.45 | 2.136 | 16.69 |
| 2V-TSVM | 1108 | 497.3 | 446.4 |
| PMC | 30.64 | 7.215 | 55.41 |
| Co-Labeling | 36.50 | 5.111 | 21.27 |

Comparable with other methods in terms of training time

# Summary

- People always say that

  *A general algorithm can hardly beat a specific designed one,*

- But I would argue

  Except you have found the key to the problem.

- Conclusion
  - An *general* multi-view learning method which unifies and outperforms the traditional semi-supervised learning and multi-instance learning.
  - Where the key is the perspective from *label candidates*.

# Thank you!

NANYANG TECHNOLOGICAL UNIVERSITY