Clean Sample Selection Algorithms with Statistical Sparsity Analysis

Yikai Wang
Fudan University

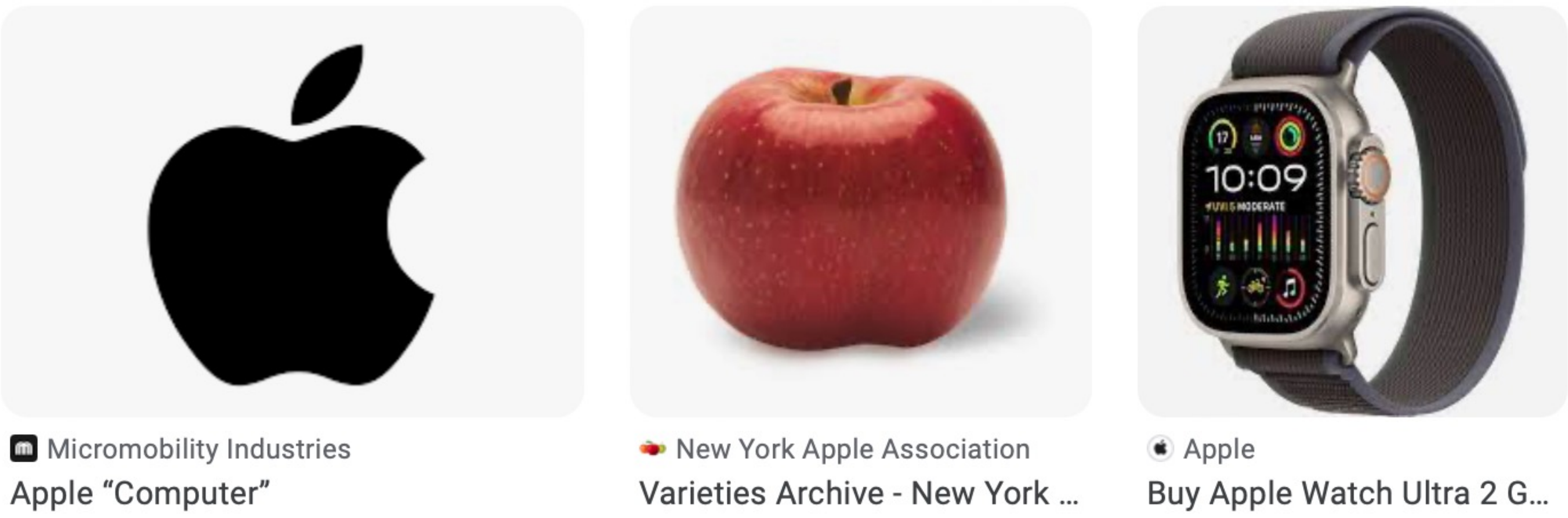
Background: Noisy Label in the Training Set

Noisy labels: mis-annotated labels. v.s. Clean labels: correctly-annotated labels.

- Annotator mistakes

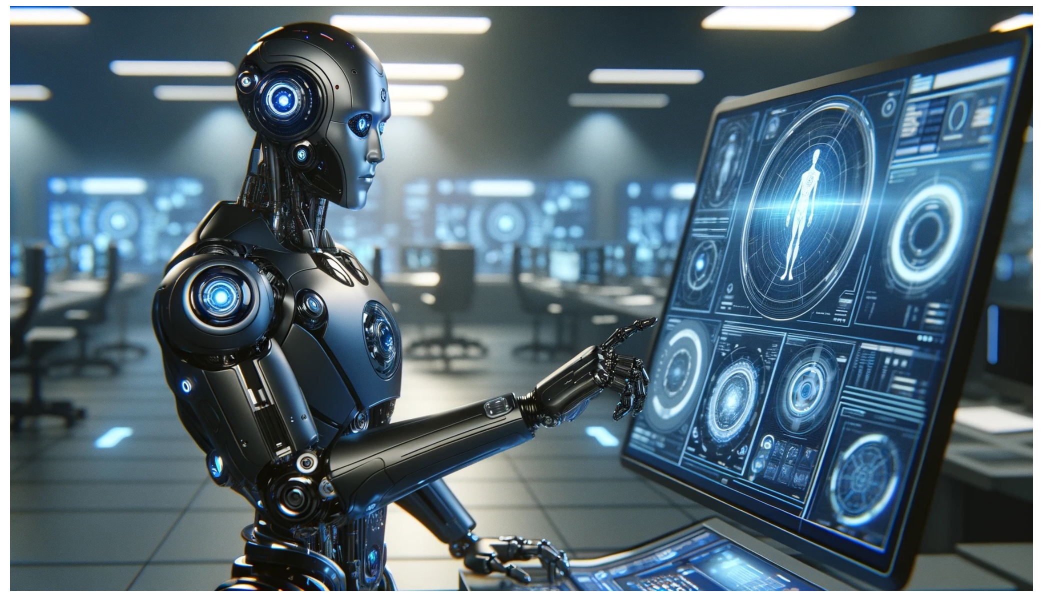
(supervised learning)

- Noisy search engine results



(webly/weak supervised learning)

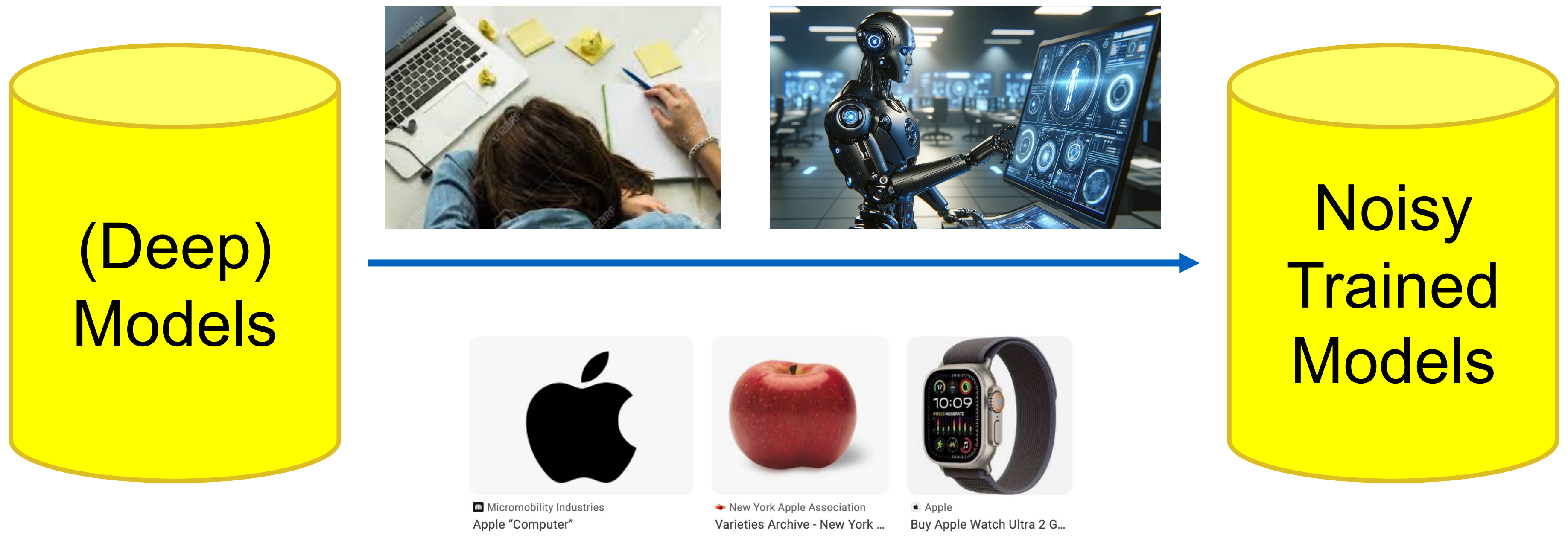
- Pseudo-labels

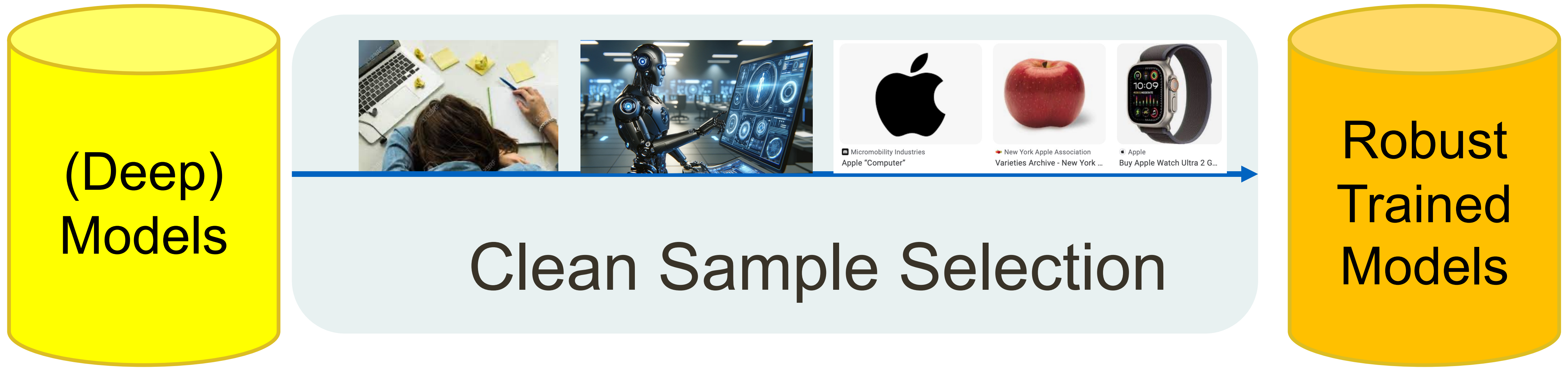


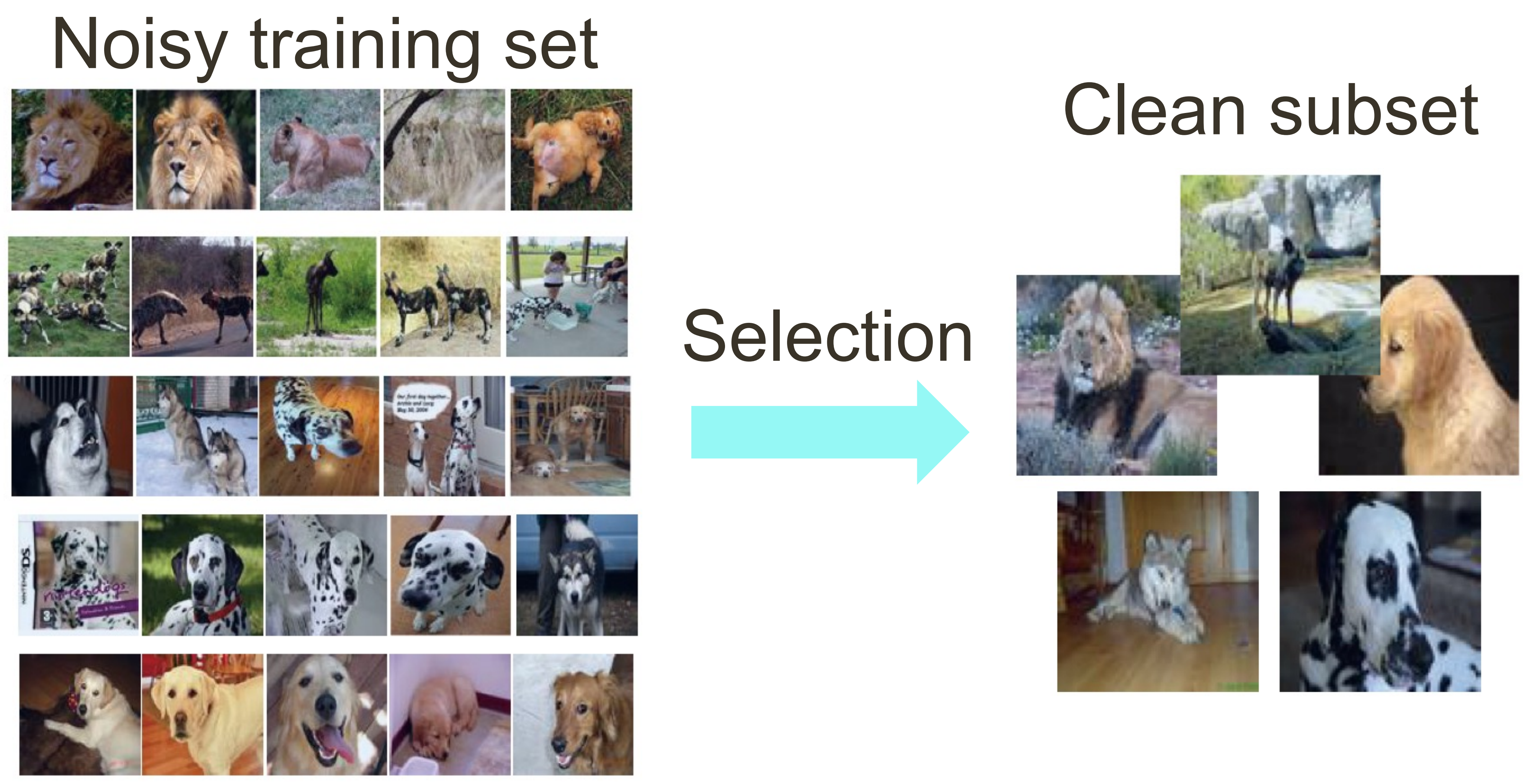
(semi-supervised learning)

The training data is corrupted in the label space with unknown corruption process.

Target: Identify Clean Subset to Improve Model Training







Motivation: Different behaviors between clean and noisy labels;

Method: Measure the different behaviors;

Theory: When will our method work?

1) The sufficient conditions to identify all the clean data;
2) Control the false-selection-rate in general scenarios;

Algorithm: How to incorporate sample selection with model training?

Application: semi-supervised few-shot learning; learning with noisy labels.

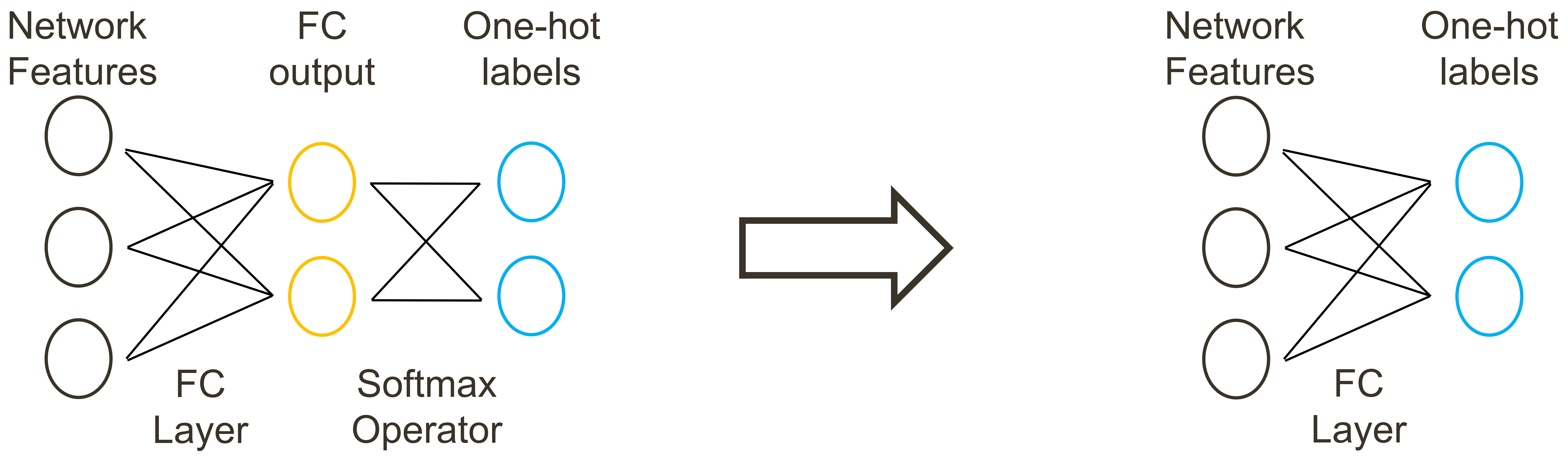
Outline

1. Method: Instance Credibility Inference
2. Theory: Noisy Set Recovery Theorem
3. Method: Knockoffs Comparison
4. Theory: False-Selection-Rate Control Theorem
5. Applications: Semi-supervised Few-Shot Learning & Learning with Noisy Labels

Clean Sample Selection
with Statistical Sparsity

1. Method: Instance Credibility Inference
2. Theory: Noisy Set Recovery
3. Method: Knockoffs Comparison
4. Theory: False-Selection-Rate Control
5. Applications

Identify Noisy Label: Linear Assumption in Networks



$$y_i = \mathrm{SoftMax}(\boldsymbol{x}_i^\top \beta)$$

$$y_i = \boldsymbol{x}_i^\top \beta + \varepsilon$$

"Sparse assumption": there are fewer single noisy *patterns* than clean *patterns*.

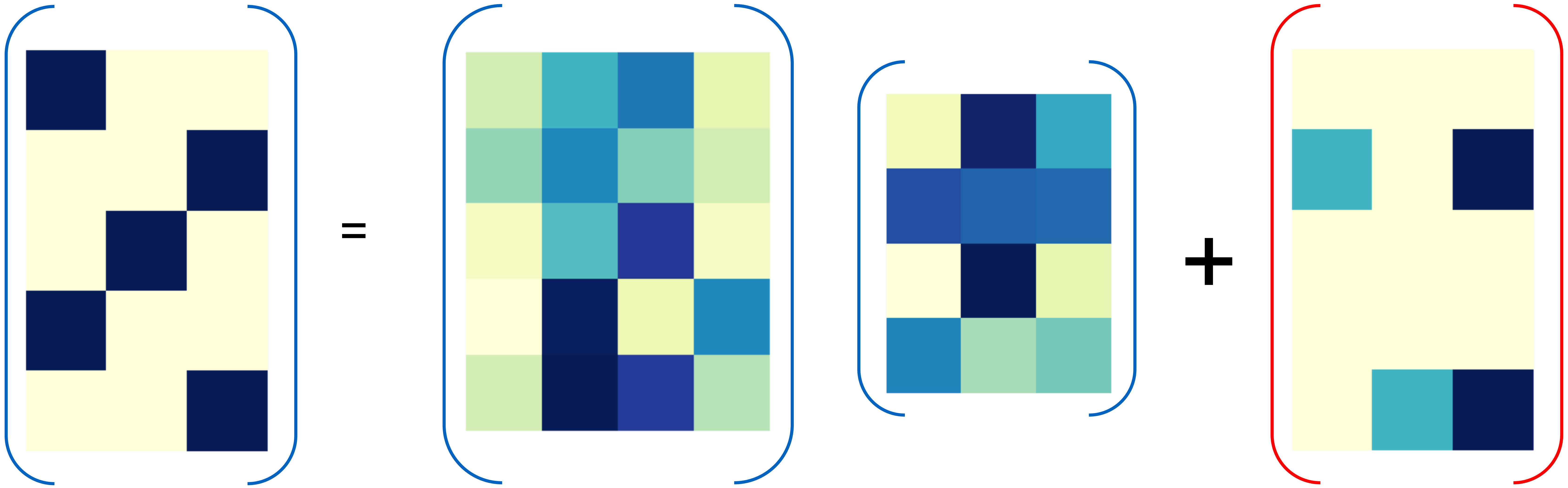
In a **2-class** classification task, there should be more clean samples in class A than **one-second** of all samples labeled as A.

Yikai Wang et al. Instance Credibility Inference for Few-Shot Learning. CVPR 2020.

Identify Noisy Data in Label Space: The Indicator

Linear model
with Noisy Labels

$$Y = X\beta + \gamma$$



Noisy One-hot Labels $Y \in \mathbb{R}^{n \times c}$

Deep Features $X \in \mathbb{R}^{n \times d}$

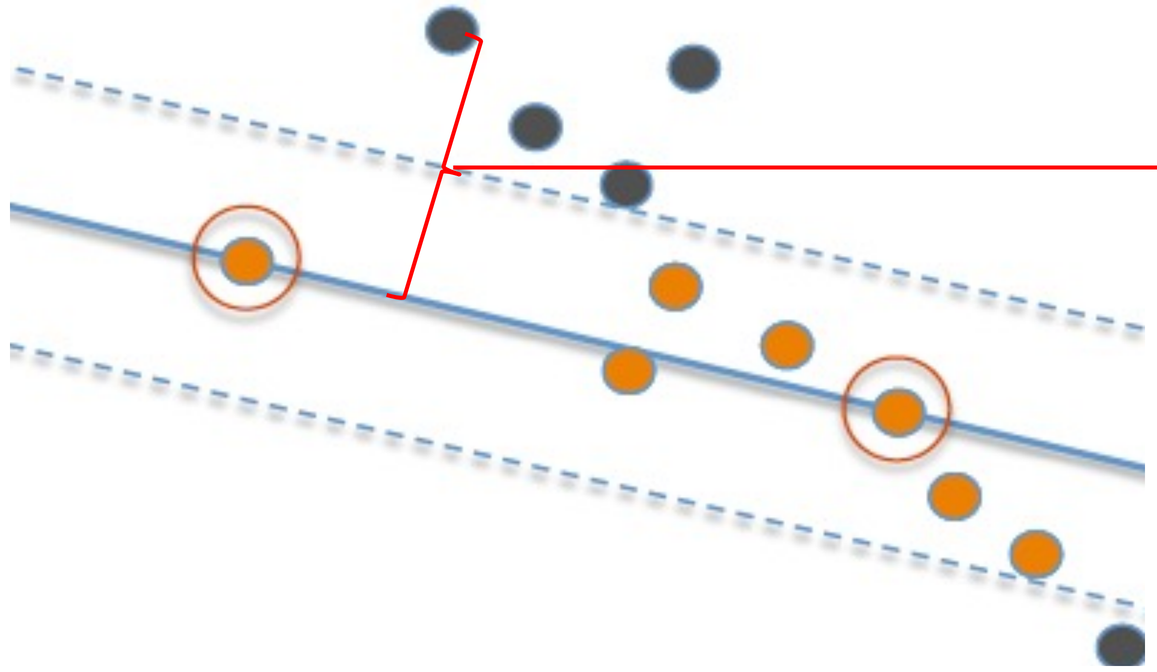
Fitted Coef. $\beta \in \mathbb{R}^{d \times c}$

Noisy Data Indicator $\gamma \in \mathbb{R}^{n \times c}$

[Wright et al. TPAMI 09] [She et al. JASA 11] [Fu et al. ECCV 14, TPAMI 16.] [Fan et al. Statistical Sinica 18] [Yikai Wang et al. CVPR 20, TPAMI 21, CVPR 22, TPAMI 23]

Motivation of γ

$$y = x^\top \beta + \varepsilon + \gamma$$



γ_i equals to the residual predict error $\gamma_i = y_i - x_i^\top \hat{\beta}$

↓

Leave-one-out externally studentized residual:

$$t_i = \frac{y_i - \boldsymbol{x}_i^\top \hat{\beta}_{(i)}}{\hat{\sigma}_{(i)}(1+\boldsymbol{x}_i(\boldsymbol{X}_{(i)}^\top \boldsymbol{X}_{(i)})^{-1}\boldsymbol{x}_i)^{1/2}}$$

$\Leftrightarrow$ test whether $\gamma = 0$ in $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \gamma 1_i + \boldsymbol{\varepsilon}$.

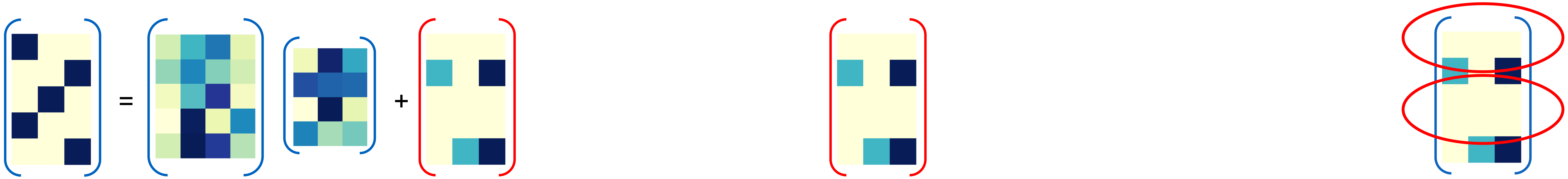
↓

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} + \boldsymbol{\gamma}$$

Yiyuan She and Art B Owen. Outlier detection using nonconvex penalized regression. Journal of the American Statistical Association, 2011.

Select Clean Sample in the Dataset

$$y_i = x_i^\top \beta + \varepsilon + \gamma_i \longrightarrow \hat{\gamma}_i \longrightarrow C = \{i : \hat{\gamma}_i = 0\}$$



clean data: zero $\|\gamma\|$;
noisy data: large $\|\gamma\|$.

$$\underset{\boldsymbol{\beta},\boldsymbol{\gamma}}{\text{argmin}} L(\boldsymbol{\beta}, \boldsymbol{\gamma}) := \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\gamma}\|_{\mathrm{F}}^2 + \lambda P(\boldsymbol{\gamma})$$

Yikai Wang et al. Instance Credibility Inference for Few-Shot Learning. CVPR 2020.

Simplification: Remove β

$$\underset{\boldsymbol{\beta},\boldsymbol{\gamma}}{\operatorname{argmin}} L(\boldsymbol{\beta},\boldsymbol{\gamma}) := \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\gamma}\|_{\mathrm{F}}^2 + \lambda P(\boldsymbol{\gamma})$$

$\frac{\partial L}{\partial \beta} = 0$ ↓ $\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^\top \boldsymbol{X}\right)^\dagger \boldsymbol{X}^\top (\boldsymbol{Y} - \boldsymbol{\gamma})$

$$\underset{\boldsymbol{\gamma}}{\operatorname{argmin}} \left\|\boldsymbol{Y} - \boldsymbol{X}\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^\dagger \boldsymbol{X}^\top (\boldsymbol{Y} - \boldsymbol{\gamma}) - \boldsymbol{\gamma}\right\|_{\mathrm{F}}^2 + \lambda P(\boldsymbol{\gamma})$$

$\boldsymbol{H} = \boldsymbol{X}\left(\boldsymbol{X}^\top \boldsymbol{X}\right)^\dagger \boldsymbol{X}^\top$ ↓ $\tilde{\boldsymbol{X}} = \boldsymbol{I} - \boldsymbol{H}, \tilde{\boldsymbol{Y}} = \tilde{\boldsymbol{X}}\boldsymbol{Y}$

$$\underset{\boldsymbol{\gamma}}{\operatorname{argmin}} \left\|\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}}\boldsymbol{\gamma}\right\|_{\mathrm{F}}^2 + \lambda P(\boldsymbol{\gamma})$$

A linear regression problem!

Yikai Wang et al. Instance Credibility Inference for Few-Shot Learning. CVPR 2020.

Simplification: How to decide λ?

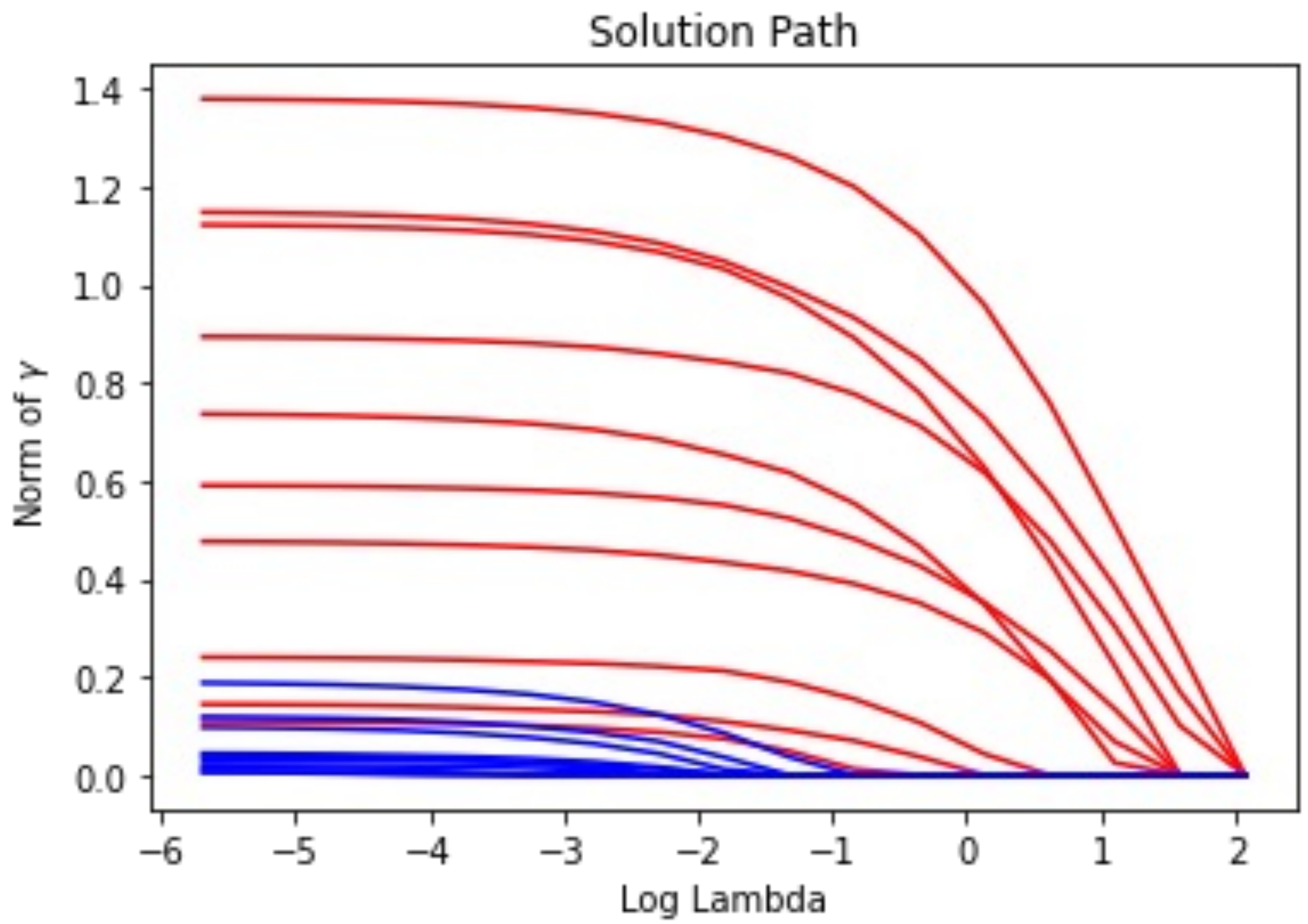
$$\underset{\boldsymbol{\gamma}}{\operatorname{argmin}} \left\| \tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}}\boldsymbol{\gamma} \right\|_{\mathrm{F}}^2 + \lambda P(\boldsymbol{\gamma})$$

We regard $\hat{\boldsymbol{\gamma}} = f(\lambda)$.

When $\lambda \to \infty$, $\hat{\boldsymbol{\gamma}} \to 0$.

With $P(\boldsymbol{\gamma}) = \sum_{i=1}^{n} \|\boldsymbol{\gamma}_i\|_2$,
$\boldsymbol{\gamma}$ vanishes instance by instance.

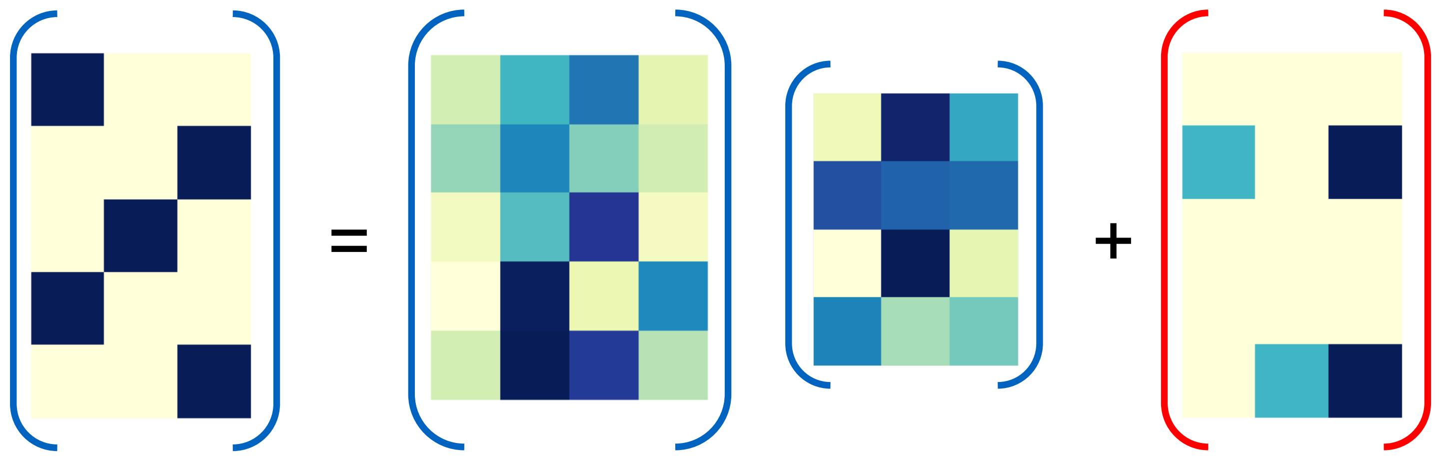
$Z_i = \sup\{\lambda : \|\hat{\boldsymbol{\gamma}}_i(\lambda)\| \neq 0\}$



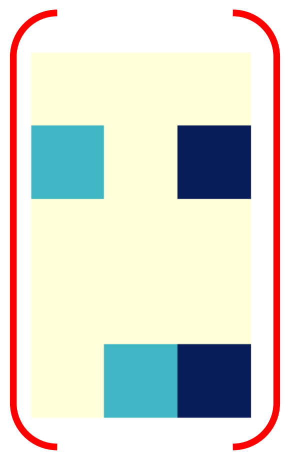
[1] Friedman, et al. 2010. "Regularization Paths for Generalized Linear Models via Coordinate Descent." Journal of Statistical Software.

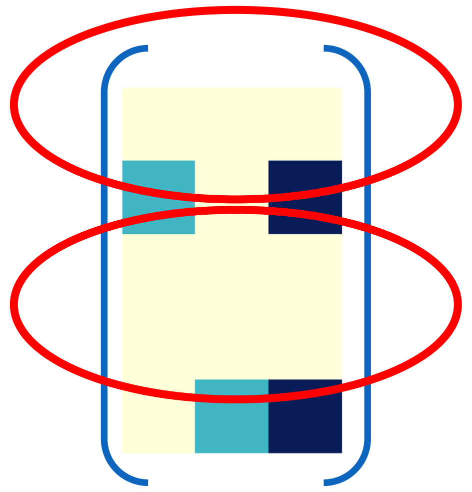
Select Clean Sample in the Dataset (Callback)

$$y_i = x_i^\top \beta + \varepsilon + \gamma_i \longrightarrow \hat{\gamma}_i \longrightarrow C = \{i : \hat{\gamma}_i = 0\}$$



clean data: zero $\|\gamma\|$;
noisy data: large $\|\gamma\|$.





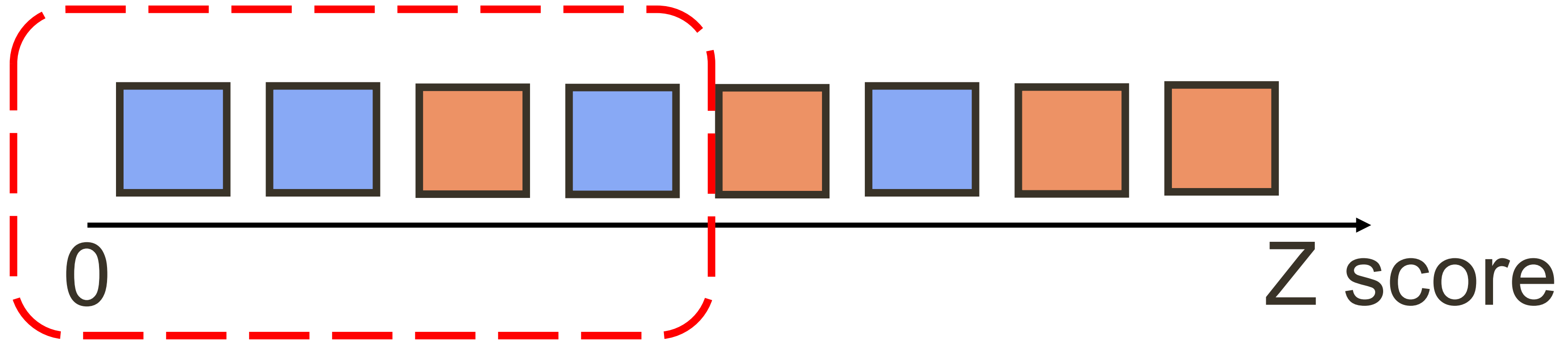
$$\underset{\boldsymbol{\beta},\boldsymbol{\gamma}}{\operatorname{argmin}} L(\boldsymbol{\beta},\boldsymbol{\gamma}) := \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\gamma}\|_{\mathrm{F}}^2 + \lambda P(\boldsymbol{\gamma})$$

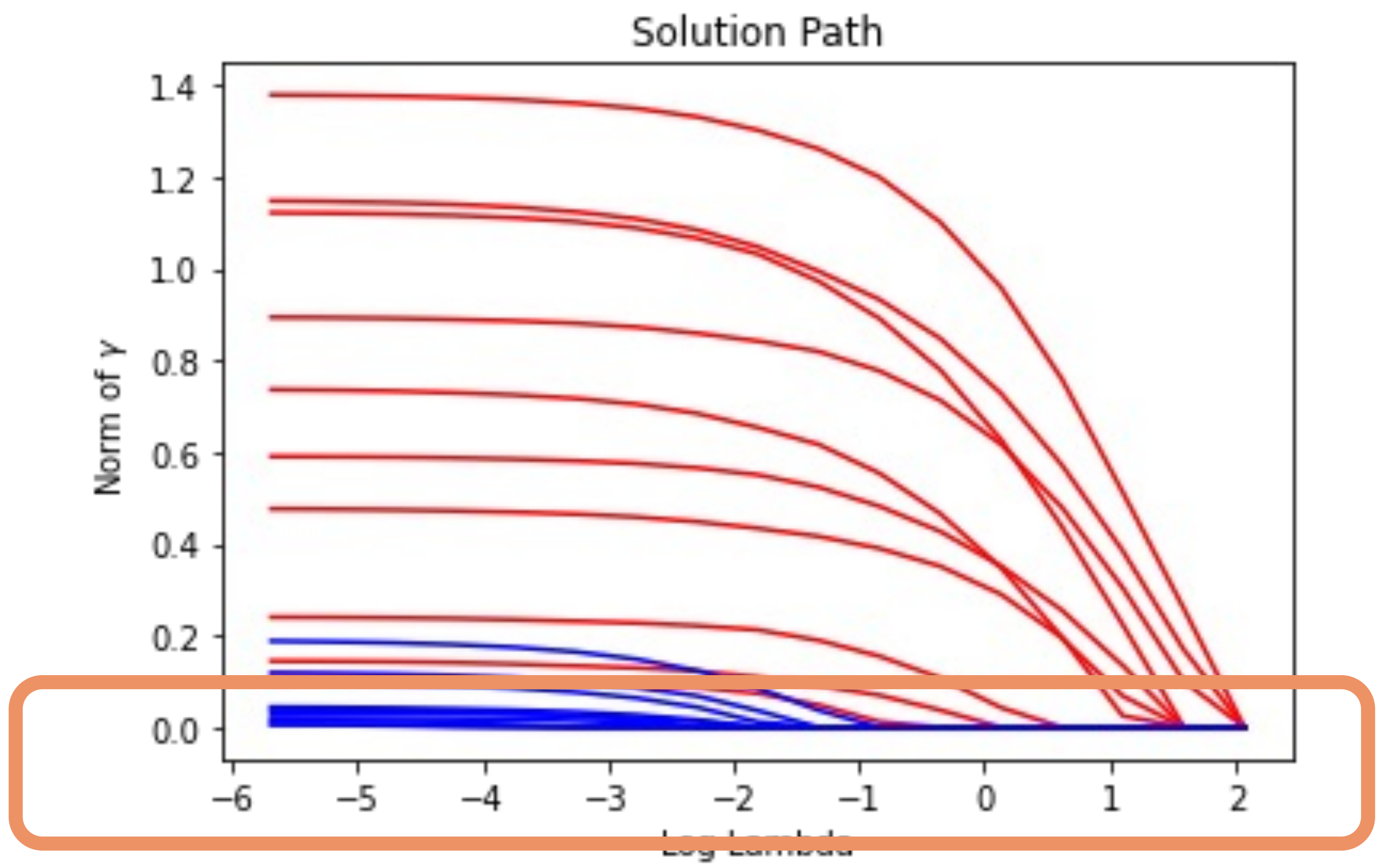
$$\downarrow$$

$$\underset{\boldsymbol{\gamma}}{\operatorname{argmin}} \left\|\tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}}\boldsymbol{\gamma}\right\|_{\mathrm{F}}^2 + \lambda P(\boldsymbol{\gamma})$$

$$\downarrow$$

$$Z_i = \sup\{\lambda : \|\hat{\gamma}_i(\lambda)\| \neq 0\}$$





Yikai Wang et al. Instance Credibility Inference for Few-Shot Learning. CVPR 2020.

Noisy Set Recovery

Advantages and Disadvantages

1. Method: Instance Credibility Inference
2. Theory: Noisy Set Recovery
3. Method: Knockoffs Comparison
4. Theory: False-Selection-Rate Control
5. Applications

Noisy Set Recovery

$$y_i = x_i^\top \beta + \varepsilon + \gamma_i \qquad \underset{\gamma}{\operatorname{argmin}} \left\| \tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}} \gamma \right\|_{\mathrm{F}}^2 + \lambda P(\gamma)$$

When can our method identify all the clean/noisy data?

Theorem 1 (Noisy set recovery). *Assume that:*
C1, Restricted eigenvalue: $\lambda_{\min}(\mathring{\boldsymbol{X}}_{\mathcal{S}}^\top \mathring{\boldsymbol{X}}_{\mathcal{S}}) = C_{\min} > 0$;
C2, Irrepresentability: *there exists a* $\eta \in (0,1]$, *such that* $\|\mathring{\boldsymbol{X}}_{\mathcal{S}^c}^\top \mathring{\boldsymbol{X}}_{\mathcal{S}} (\mathring{\boldsymbol{X}}_{\mathcal{S}}^\top \mathring{\boldsymbol{X}}_{\mathcal{S}})^{-1}\|_\infty \leq 1 - \eta$;
C3, Large error: $\vec{\gamma}^*_{\min} := \min_{i \in \mathcal{S}} |\vec{\gamma}^*_i| > h(\lambda, \eta, \mathring{\boldsymbol{X}}, \vec{\gamma}^*)$;
where $\|\boldsymbol{A}\|_\infty := \max_i \sum_j |A_{i,j}|$, *and* $h(\lambda, \eta, \mathring{\boldsymbol{X}}, \vec{\gamma}^*) = \lambda\eta / \sqrt{C_{\min}\mu_{\mathring{\boldsymbol{X}}}} + \lambda \|(\mathring{\boldsymbol{X}}_{\mathcal{S}}^\top \mathring{\boldsymbol{X}}_{\mathcal{S}})^{-1} \operatorname{sign}(\vec{\gamma}^*_{\mathcal{S}})\|_\infty$.
Let $\lambda \geq \frac{2\sigma\sqrt{\mu_{\mathring{\boldsymbol{X}}}}}{\eta}\sqrt{\log cn}$. *Then with probability greater than* $1 - 2(cn)^{-1}$, *model Eq. (8) has a unique solution* $\hat{\vec{\gamma}}$ *such that: 1) If C1 and C2 hold,* $\hat{\mathcal{C}}^c \subseteq \mathcal{C}^c$;*2) If C1, C2 and C3 hold,* $\hat{\mathcal{C}}^c = \mathcal{C}^c$.

Noisy Set Recovery (in natural language):
1. With C1-C3, we can identify all the noisy data.
2. With C1-C2, the identified noisy data is the subset of ground-truth noisy data.

Yikai Wang et al. How to Trust Unlabeled Data? Instance Credibility Inference for Few-Shot Learning. TPAMI 2021.

Verification: Will satisfying conditions lead to improved accuracy?

Satisfied Assumptions	None	C1	C1 and C2	All
Improved Episodes	0	424	1035	40
Total Episodes	0	793	1164	43
I/T	–	53.5%	88.9%	93.0%

1) In more than half of the experiments the assumptions C1-C2 are satisfied. Most of them (89.0%) will achieve better performance after self-taught with ICI.

2) When all the assumptions are satisfied, we will get better performance in a high ratio (93.0%).

3) Even if C2-C3 are not satisfied, we still have the chance of improving the performance (53.5%).

Yikai Wang et al. How to Trust Unlabeled Data? Instance Credibility Inference for Few-Shot Learning. TPAMI 2021.

Challenges of Noisy Set Recovery

$$y_i = x_i^\top \beta + \varepsilon + \gamma_i$$

$$\underset{\gamma}{\operatorname{argmin}} \left\| \tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}}\boldsymbol{\gamma} \right\|_{\mathrm{F}}^2 + \lambda P(\boldsymbol{\gamma})$$

Theorem 1 (Noisy set recovery). *Assume that:*
C1, Restricted eigenvalue: $\lambda_{\min}(\mathring{\boldsymbol{X}}_{\mathcal{S}}^\top \mathring{\boldsymbol{X}}_{\mathcal{S}}) = C_{\min} > 0$;
C2, Irrepresentability: *there exists a* $\eta \in (0,1]$, *such that* $\|\mathring{\boldsymbol{X}}_{\mathcal{S}^c}^\top \mathring{\boldsymbol{X}}_{\mathcal{S}} (\mathring{\boldsymbol{X}}_{\mathcal{S}}^\top \mathring{\boldsymbol{X}}_{\mathcal{S}})^{-1}\|_\infty \leq 1 - \eta$;
C3, Large error: $\vec{\gamma}^*_{\min} := \min_{i \in \mathcal{S}} |\vec{\gamma}^*_i| > h(\lambda, \eta, \mathring{\boldsymbol{X}}, \vec{\gamma}^*)$;
where $\|\boldsymbol{A}\|_\infty := \max_i \sum_j |A_{i,j}|$, *and* $h(\lambda, \eta, \mathring{\boldsymbol{X}}, \vec{\gamma}^*) = \lambda\eta/\sqrt{C_{\min}\mu_{\mathring{\boldsymbol{X}}}} + \lambda\|(\mathring{\boldsymbol{X}}_{\mathcal{S}}^\top \mathring{\boldsymbol{X}}_{\mathcal{S}})^{-1}\mathrm{sign}(\vec{\gamma}^*_{\mathcal{S}})\|_\infty$.
Let $\lambda \geq \frac{2\sigma\sqrt{\mu_{\mathring{\boldsymbol{X}}}}}{\eta}\sqrt{\log cn}$. *Then with probability greater than* $1 - 2(cn)^{-1}$, *model Eq. (8) has a unique solution* $\hat{\vec{\gamma}}$ *such that: 1) If C1 and C2 hold,* $\hat{\mathcal{C}}^c \subseteq \mathcal{C}^c$;*2) If C1, C2 and C3 hold,* $\hat{\mathcal{C}}^c = \mathcal{C}^c$.

Uncontrollable Challenges:

- The C2 requires knowledge about the ground-truth noisy set, which is unknown in practice.
- Our target is to select clean data, but in most cases (C1-C2 satisfied), we will still falsely-select noisy data, and we do not know the false-selection-rate.

Can we control the false-selection-rate in general scenarios?

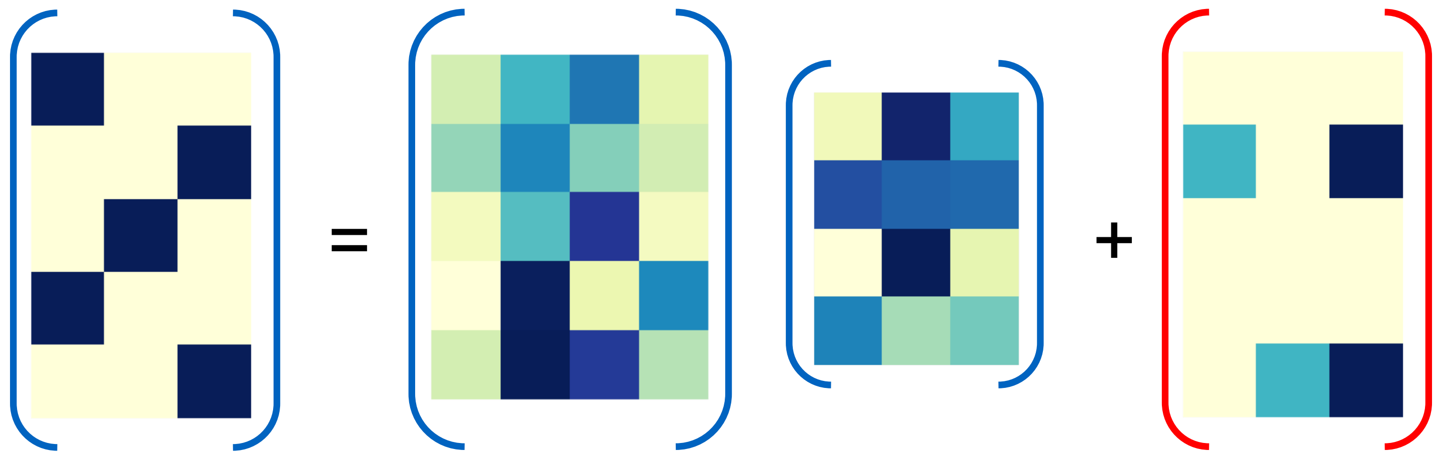
Yikai Wang et al. How to Trust Unlabeled Data? Instance Credibility Inference for Few-Shot Learning. TPAMI 2021.

Clean Sample Selection
with Controlled False-Selection-Rate

1. Method: Instance Credibility Inference
2. Theory: Noisy Set Recovery
3. Method: Knockoffs Comparison
4. Theory: False-Selection-Rate Control
5. Applications

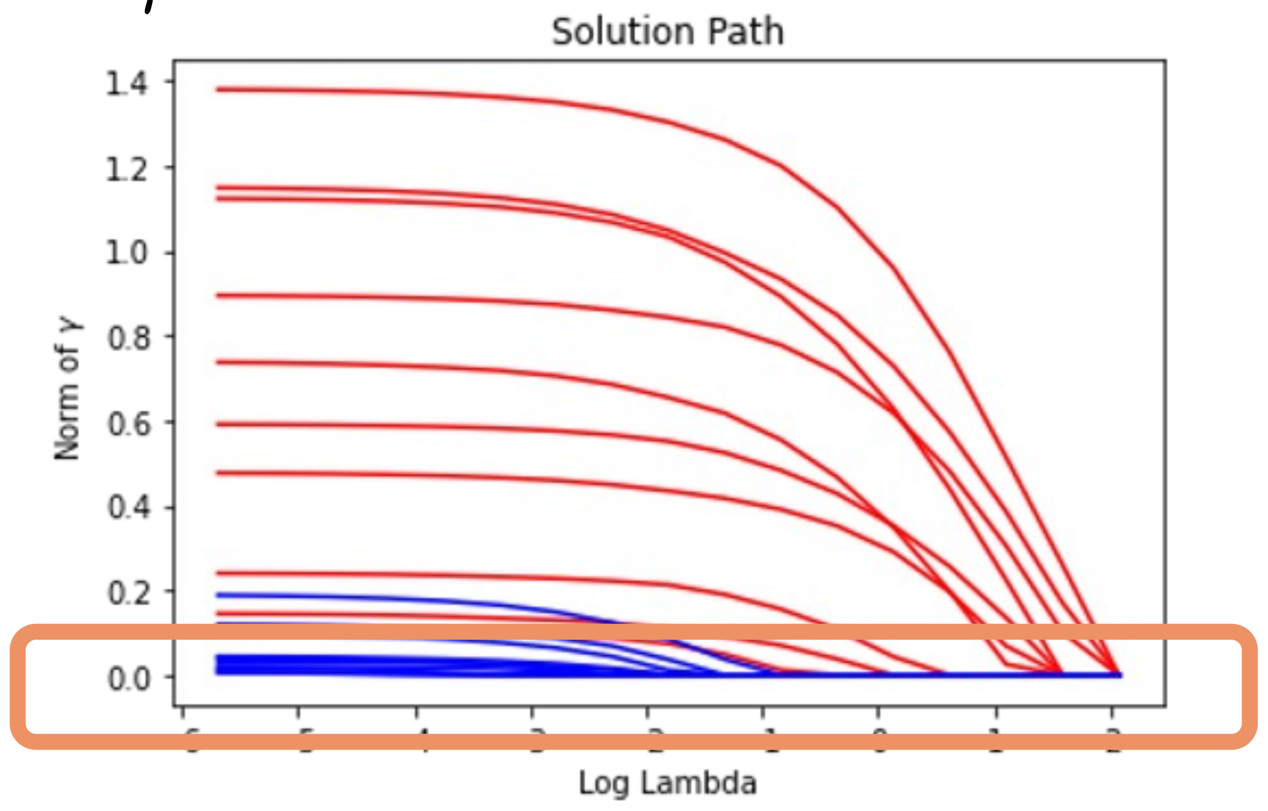
Motivation: Bi-Level Comparison

$$y_i = x_i^\top \beta + \varepsilon + \gamma_i$$



clean data: zero $\|\gamma\|$;
noisy data: large $\|\gamma\|$.

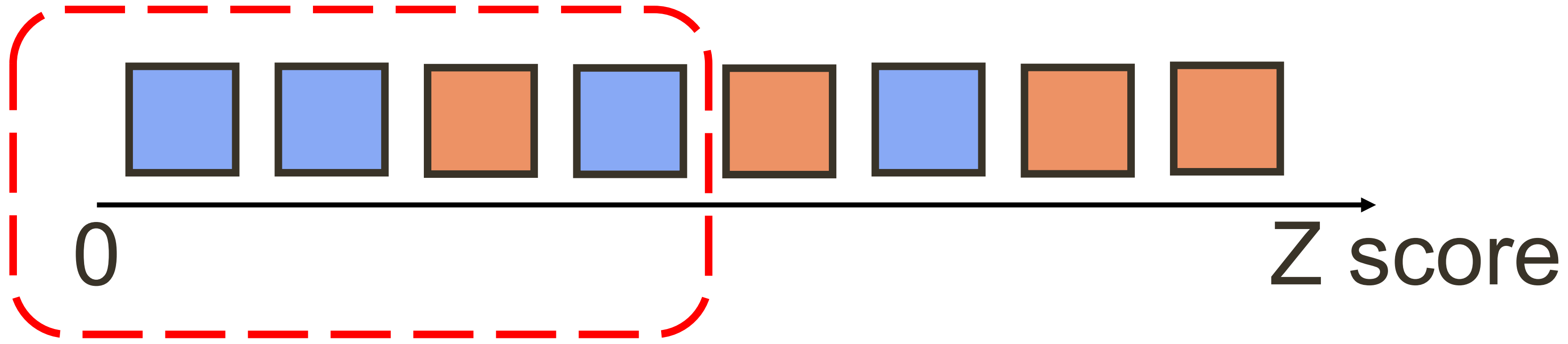
$$\underset{\boldsymbol{\gamma}}{\operatorname{argmin}} \left\| \tilde{\boldsymbol{Y}} - \tilde{\boldsymbol{X}} \boldsymbol{\gamma} \right\|_{\mathrm{F}}^2 + \lambda P(\boldsymbol{\gamma})$$



$$Z_i = \sup\{\lambda : \|\hat{\gamma}_i(\lambda)\| \neq 0\}$$

We transform the sample selection problem into a ranking problem:

$Z_i < Z_j \Leftrightarrow$ sample i is more reliably than sample j



Is the label of sample i more reliably than another label?

$Y_i\ (1, 0, 0)$ **Permute** $\longrightarrow$ $(0, 1, 0)\ \tilde{Y}_i$

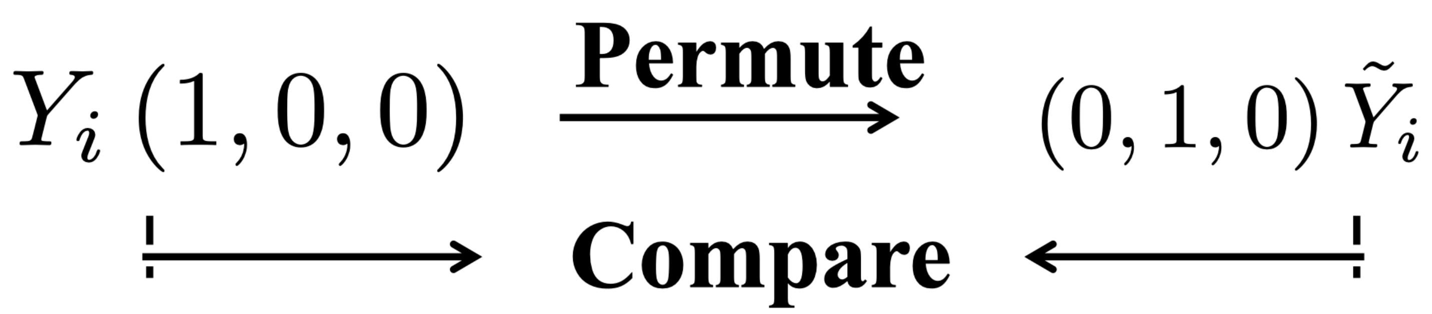
$\longrightarrow$ **Compare** $\longleftarrow$

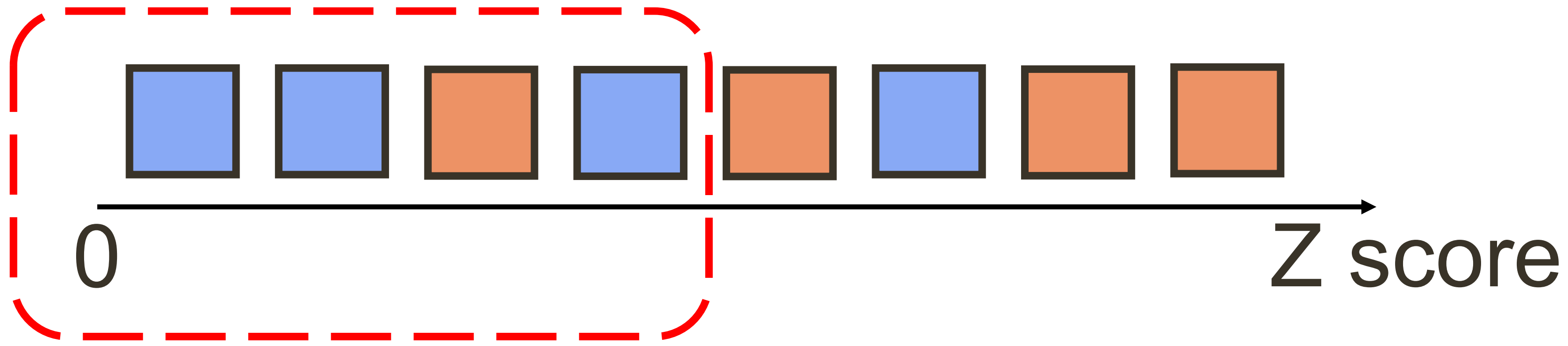
Yikai Wang et al. Knockoffs-SPR: Clean Sample Selection in Learning with Noisy Labels. TPAMI 2023.

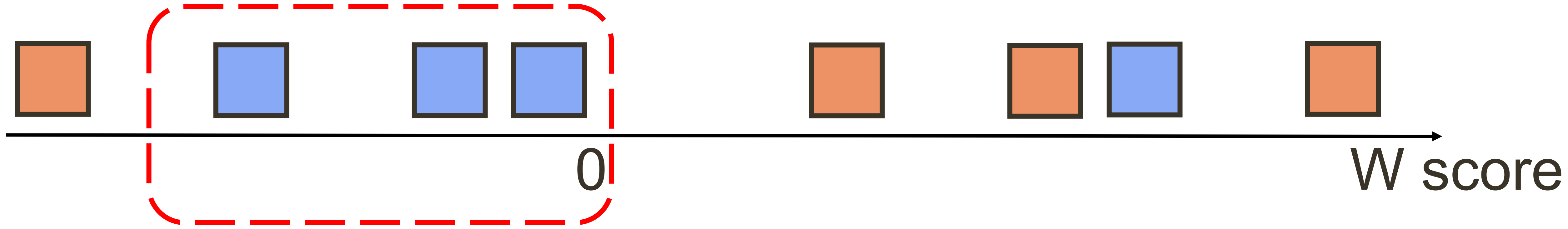
Motivation: An extra sign comparison

$$y_i = x_i^\top \beta + \varepsilon + \gamma_i$$

+







Yikai Wang et al. Knockoffs-SPR: Clean Sample Selection in Learning with Noisy Labels. TPAMI 2023.

Label-Knockoff Comparison

$$\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\gamma}\|_{\mathrm{F}}^2 + \lambda P(\boldsymbol{\gamma})$$

+

$$Y_i\,(1,0,0) \xrightarrow{\textbf{Permute}} (0,1,0)\,\tilde{Y}_i$$

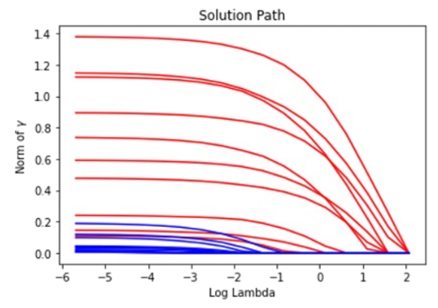
$\longrightarrow$ **Compare** $\longleftarrow$

$\longrightarrow$

$$\begin{cases} \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\gamma}\|_{\mathrm{F}}^2 + \lambda P(\boldsymbol{\gamma}), \\ \left\|\tilde{\mathbf{Y}} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\gamma}\right\|_{\mathrm{F}}^2 + \lambda P(\boldsymbol{\gamma}). \end{cases}$$

$\downarrow$

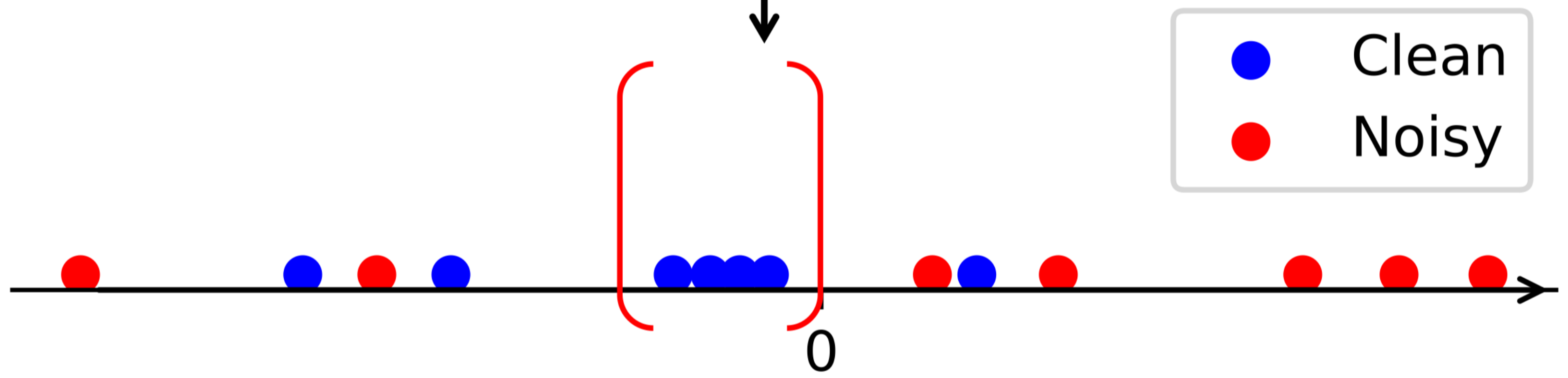
$$\begin{cases} Z_i = \sup\{\lambda : \|\boldsymbol{\gamma}_i(\lambda)\| \neq 0\} \\ \tilde{Z}_i = \sup\{\lambda : \|\tilde{\boldsymbol{\gamma}}_i(\lambda)\| \neq 0\} \end{cases}$$



clean data: zero $\|\gamma\|$, small Z;
noisy data: large $\|\gamma\|$, large Z.

$\longleftarrow$

$$W_i := Z_i \cdot \mathrm{sign}(Z_i - \tilde{Z}_i).$$



Select data with small negative statistics:

$$C_2 := \{j : -T \leq W_j < 0\}, \quad T = \max\left\{t > 0 : \frac{1 + \#\{j : 0 < W_j \leq t\}}{\#\{j : -t \leq W_j < 0\} \vee 1} \leq q\right\}$$

Yikai Wang et al. Knockoffs-SPR: Clean Sample Selection in Learning with Noisy Labels. TPAMI 2023.

Knockoff Comparison：Why Permutation Label?

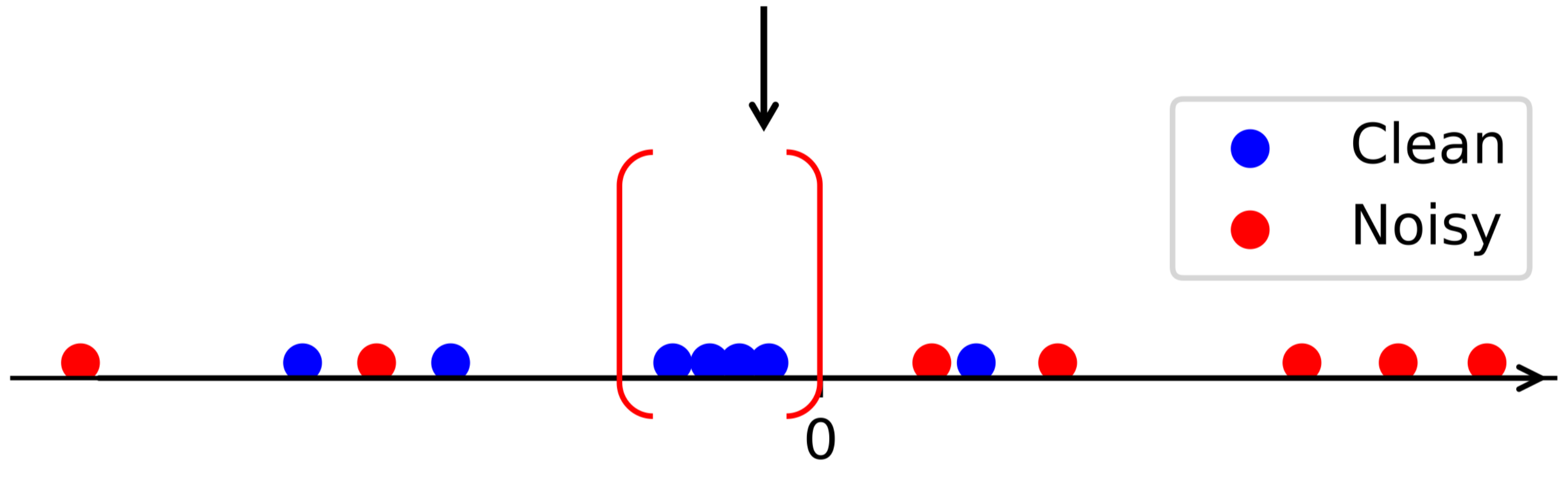
$$\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{\gamma}\|_{\mathrm{F}}^{2}+\lambda P(\boldsymbol{\gamma})$$

+

$Y_i\,(1,0,0)$ **Permute** → $(0,1,0)\,\tilde{Y}_i$

→ **Compare** ←

$$W_i := Z_i \cdot \operatorname{sign}(Z_i - \tilde{Z}_i).$$



clean data: zero $\|\gamma\|$, small Z;
noisy data: large $\|\gamma\|$, large Z.

- Clean label → noisy label:
  Ideally small negative W.

- Noisy label → clean label ($\frac{1}{c-1}$), noisy label ($\frac{c-2}{c-1}$),
  where c denotes the number of classes.
  i) Noisy → clean:
  large positive W.
  ii) Noisy → noisy:
  large W.
  approximately equal probability to be positive or negative.

Select data with small negative statistics:

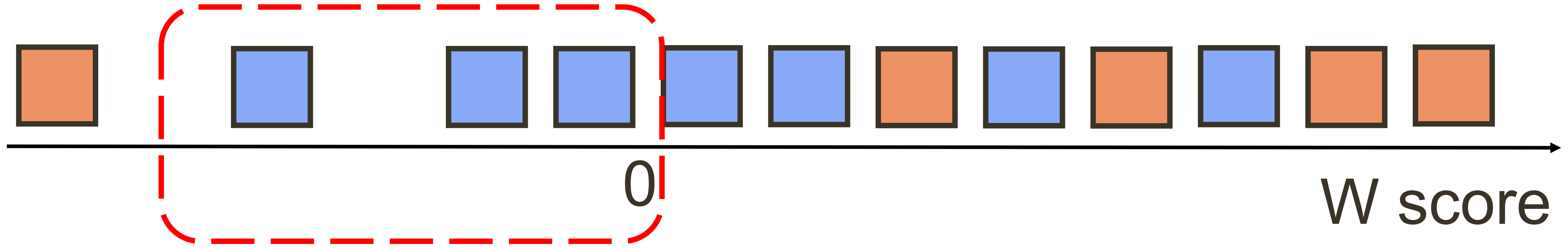
Yikai Wang et al. Knockoffs-SPR: Clean Sample Selection in Learning with Noisy Labels. TPAMI 2023.

Knockoff Comparison: How to decide T (intuitively)?

Select data with small negative statistics:

$$C_2 := \{j : -T \leq W_j < 0\}, \quad T = \max\left\{t > 0 : \frac{1 + \#\{j : 0 < W_j \leq t\}}{\#\{j : -t \leq W_j < 0\} \vee 1} \leq q\right\}$$

- Clean label → noisy label:
  Ideally small negative W.



- Noisy label → clean label ($\frac{1}{c-1}$), noisy label ($\frac{c-2}{c-1}$),
  where c denotes the number of classes.
  i) Noisy → clean:
    large positive W.
  ii) Noisy → noisy:
    large W.
    approximately equal probability to be positive or negative.

The samples fall in the negative interval:

1. clean labels, great!
2. noisy labels, bad..
   Can we know the number?
   Yes, approximately the number of samples in the positive interval!

Yikai Wang et al. Knockoffs-SPR: Clean Sample Selection in Learning with Noisy Labels. TPAMI 2023.

Knockoff Comparison: How to decide T (formally)?

Select data with small negative statistics:

$$C_2 := \{j : -T \leq W_j < 0\}, \quad T = \max\left\{t > 0 : \frac{1 + \#\{j : 0 < W_j \leq t\}}{\#\{j : -t \leq W_j < 0\} \vee 1} \leq q\right\}$$

We aim to control the false selection rate:

$$\mathrm{FSR} = \mathbb{E}\left[\frac{\#\left\{j : j \notin \mathcal{H}_0 \cap \hat{\mathcal{C}}\right\}}{\#\left\{j : j \in \hat{\mathcal{C}}\right\} \vee 1}\right]$$

And in our problem, FSR becomes:

$$\mathrm{FSR}(t) = \mathbb{E}\left[\frac{\#\{j : \boldsymbol{\gamma}_j \neq 0 \text{ and } -t \leq W_j < 0\}}{\#\{j : -t \leq W_j < 0\} \vee 1}\right]$$

We can decompose it into:

$$\mathbb{E}\left[\frac{\#\{\boldsymbol{\gamma}_j \neq 0,\ -t \leq W_j < 0\}}{1 + \#\{\boldsymbol{\gamma}_j \neq 0,\ 0 < W_j \leq t\}} \cdot \frac{1 + \#\{\boldsymbol{\gamma}_j \neq 0,\ 0 < W_j \leq t\}}{\#\{-t \leq W_j < 0\} \vee 1}\right]$$

$$\leq \mathbb{E}\left[\frac{\#\{\boldsymbol{\gamma}_j \neq 0,\ -t \leq W_j < 0\}}{1 + \#\{\boldsymbol{\gamma}_j \neq 0,\ 0 < W_j \leq t\}} \frac{1 + \#\{0 < W_j \leq t\}}{\#\{-t \leq W_j < 0\} \vee 1}\right]$$

$$\leq \mathbb{E}\left[\frac{\#\{\boldsymbol{\gamma}_j \neq 0,\ -t \leq W_j < 0\}}{1 + \#\{\boldsymbol{\gamma}_j \neq 0,\ 0 < W_j \leq t\}} q\right]$$

Yikai Wang et al. Knockoffs-SPR: Clean Sample Selection in Learning with Noisy Labels. TPAMI 2023.

False-Selection-Rate Control

in general scenarios

1. Method: Instance Credibility Inference
2. Theory: Noisy Set Recovery
3. Method: Knockoffs Comparison
4. Theory: False-Selection-Rate Control
5. Applications

False-Selection-Rate Control

$$y_i = x_i^\top \beta + \varepsilon + \gamma_i$$

Theorem 1 (FSR control). *For c-class classification task, and for all* $0 < q \leq 1$*, the solution of our method holds*

$$\mathrm{FSR}(T) \leq q \tag{1}$$

with the threshold T for two subsets defined respectively as

$$T_i = \max\left\{t \in \mathcal{W} : \frac{1 + \#\left\{j : 0 < W_j \leq t\right\}}{\#\left\{j : -t \leq W_j < 0\right\} \vee 1} \leq \frac{c-2}{2c} q\right\}.$$

Advantages:

1. No complicate conditions;
2. Able to guide practical applications;

Limitations:

1. Too small q leads to empty selected clean subset;
2. Extra requirement: independence between β and γ.

Yikai Wang et al. Knockoffs-SPR: Clean Sample Selection in Learning with Noisy Labels. TPAMI 2023.

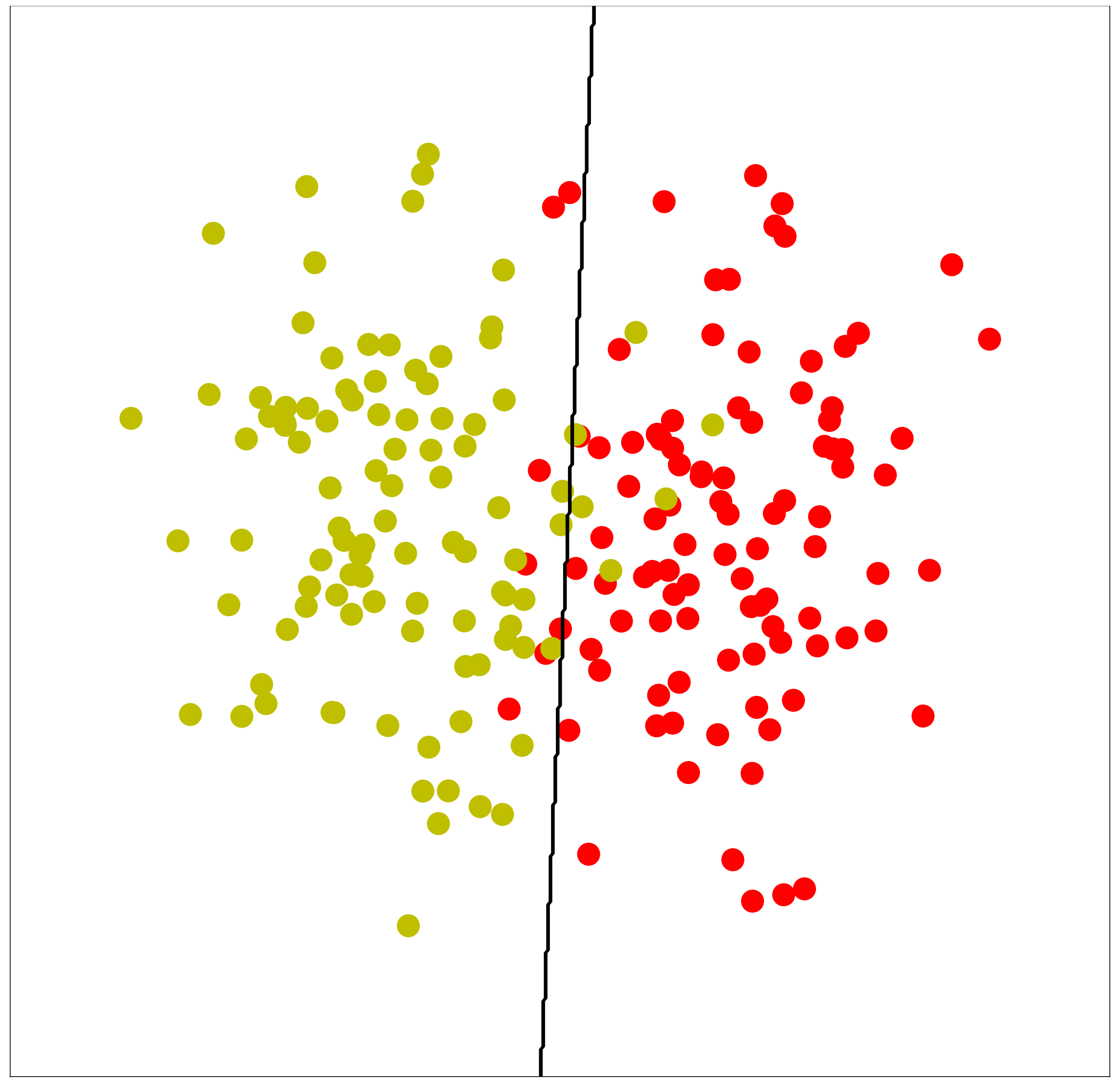
Clean Sample Selection

in Real Problems

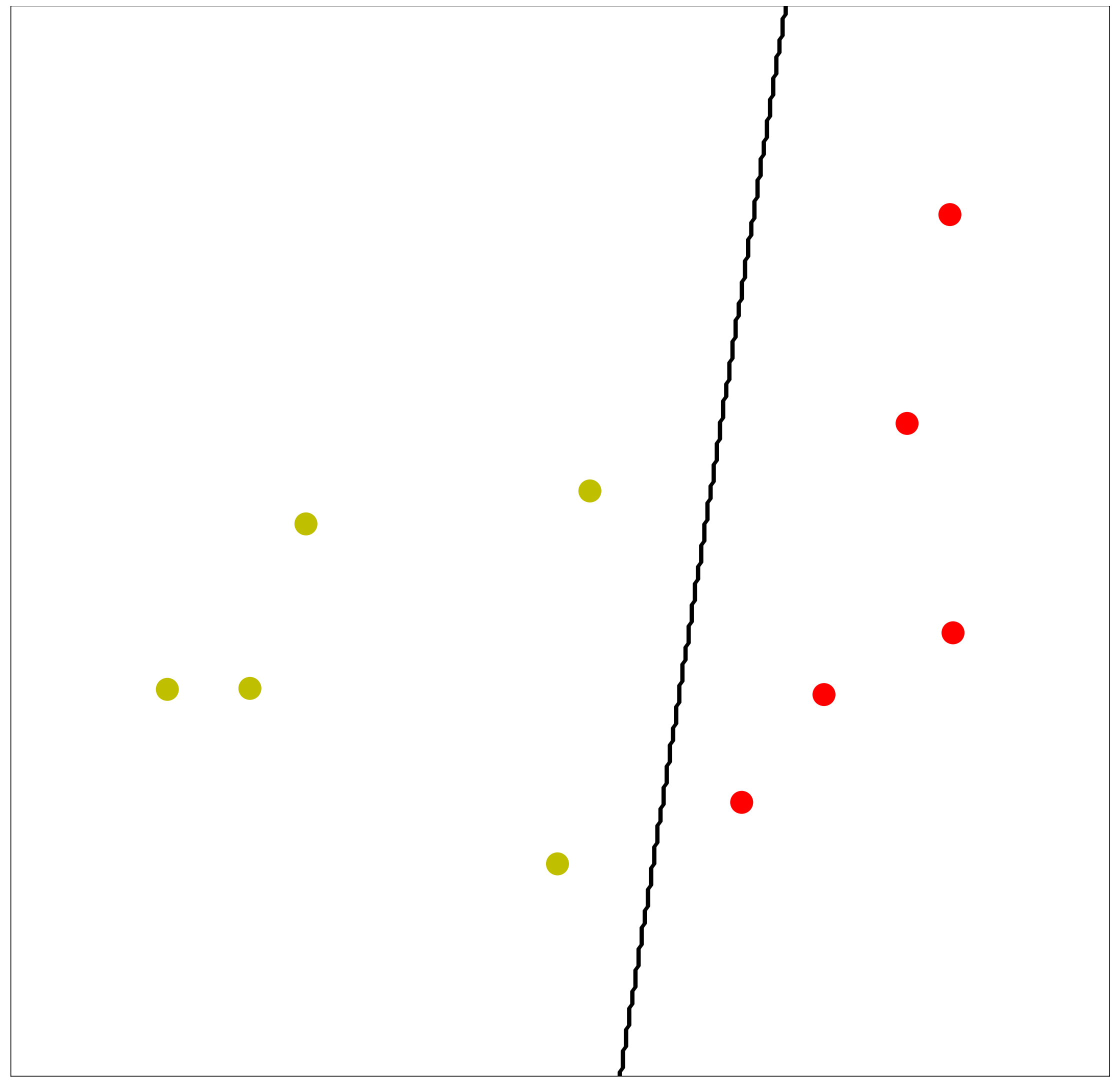
1. Method: Instance Credibility Inference
2. Theory: Noisy Set Recovery
3. Method: Knockoffs Comparison
4. Theory: False-Selection-Rate Control
5. Applications

Application 1: Semi-Supervised Few-Shot Learning

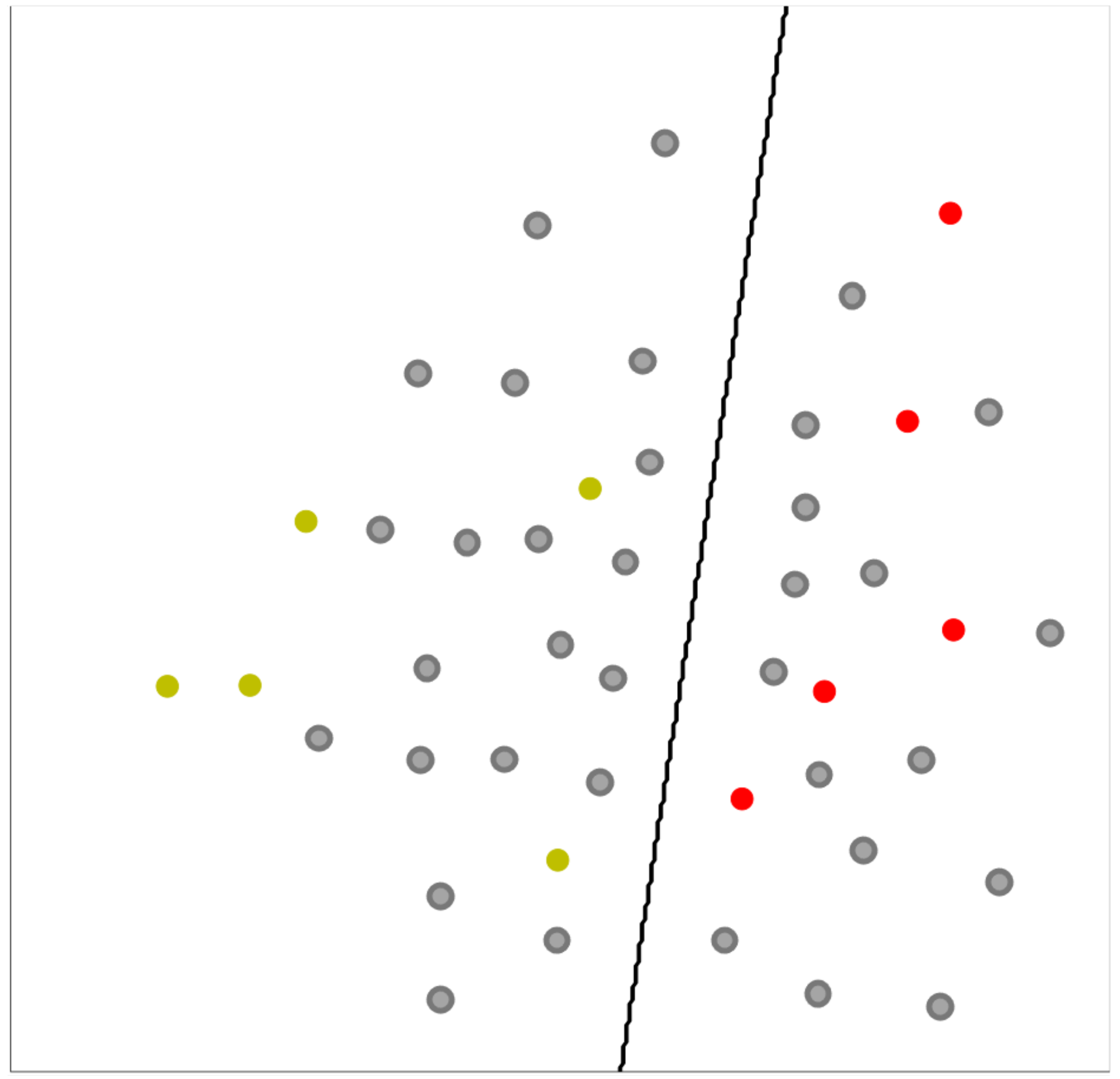
Tackle machine learning problem with only limited training data provided.



Binary classification
with many labeled data

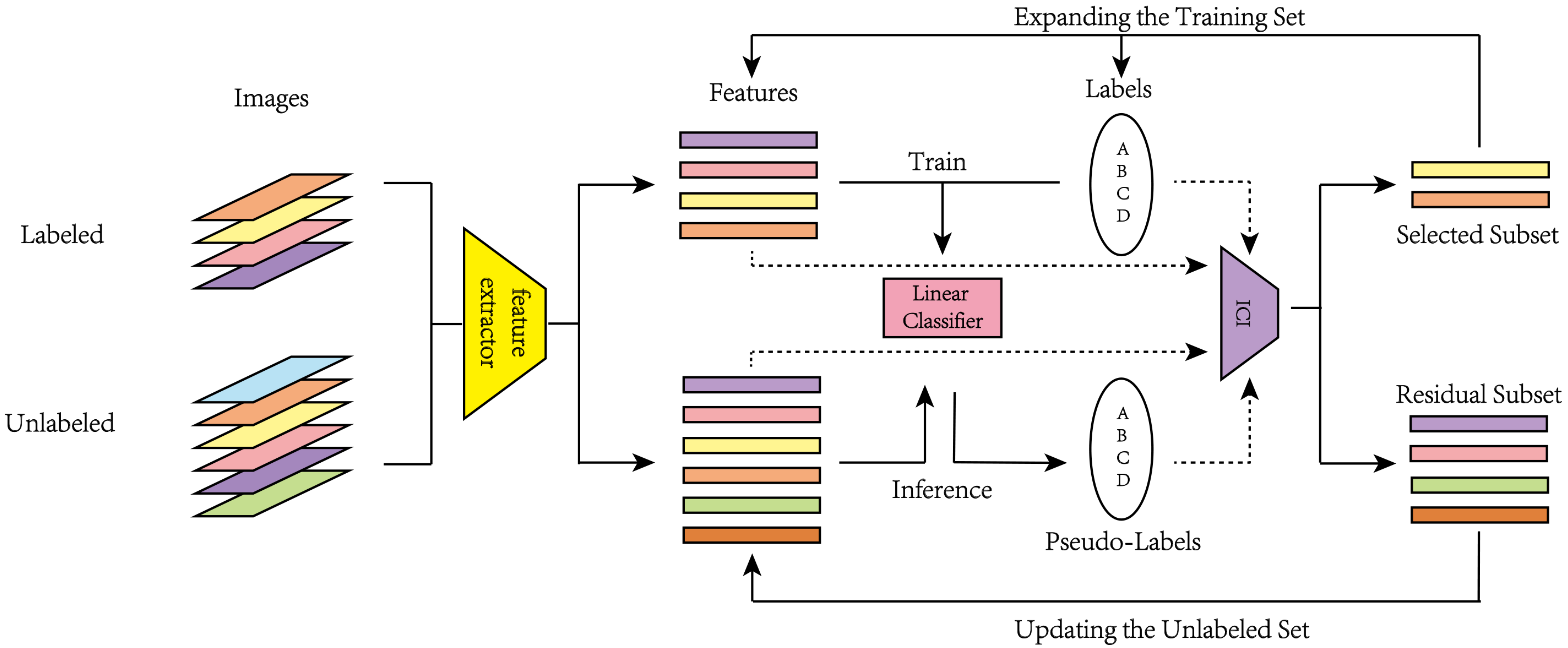


Few-shot binary classification



Few-shot binary classification
with unlabeled data

Framework for Semi-Supervised Few-Shot Learning

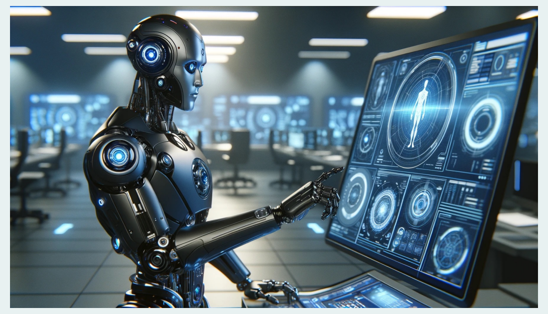


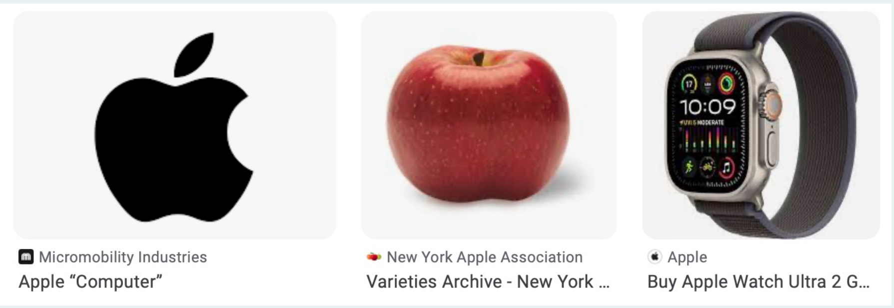
Yikai Wang et al. Instance Credibility Inference for Few-Shot Learning. CVPR 2020.
Yikai Wang et al. How to Trust Unlabeled Data? Instance Credibility Inference for Few-Shot Learning. IEEE TPAMI 2021.

Application 2: Learning with Noisy Labels

Directly trains a neural network from large scale noisy training dataset.

(Deep) Models





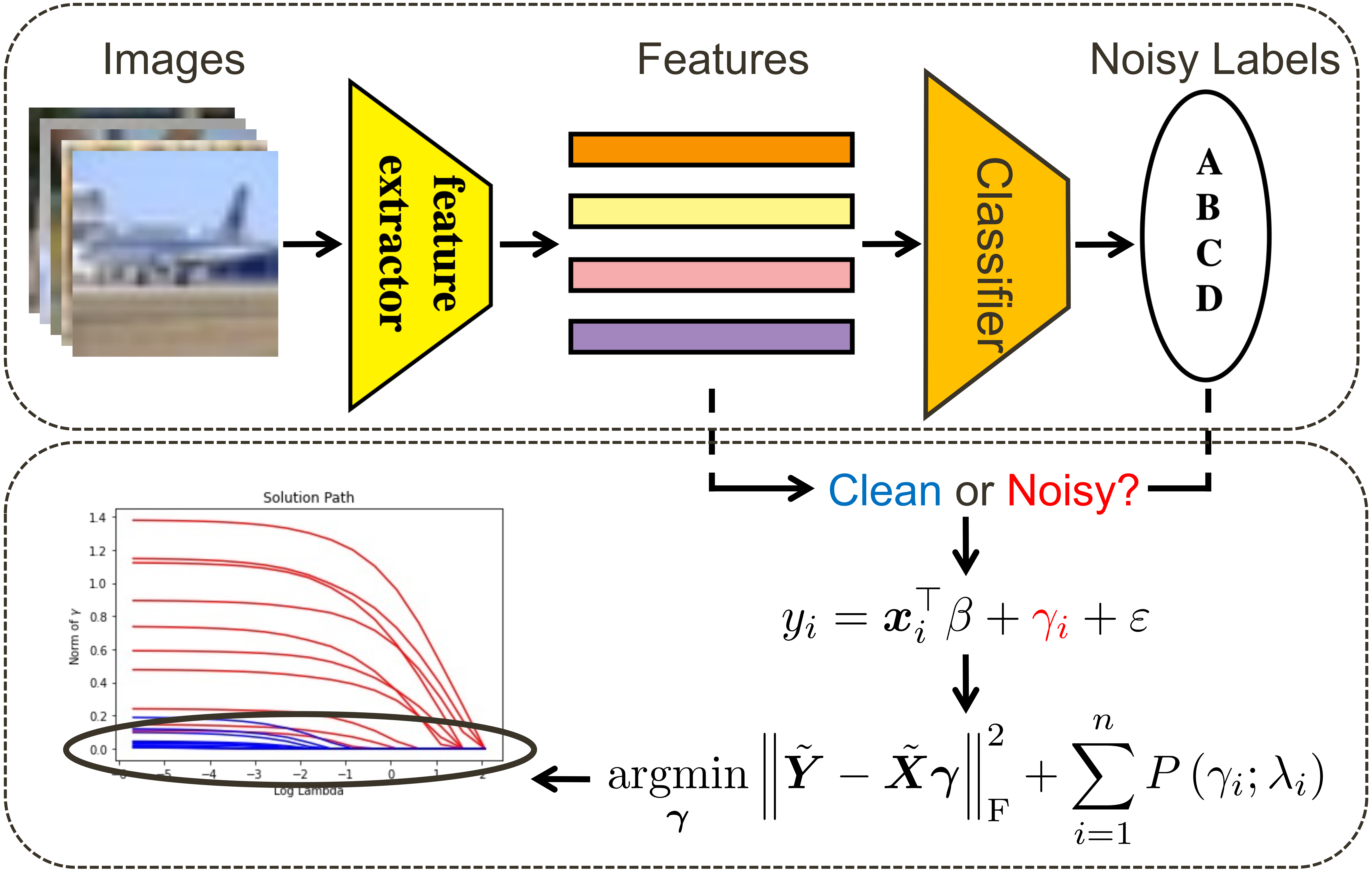
Clean Sample Selection

Robust Trained Models

Framework for Learning with Noisy Labels

Stage 1:
Feature Learning

Stage 2:
Sample Selection



Yikai Wang et al. Scalable Penalized Regression for Noise Detection in Learning with Noisy Labels. CVPR 2022.
Yikai Wang et al. Knockoffs-SPR: Clean Sample Selection in Learning with Noisy Labels. TPAMI 2023.

Bag of Tricks to Better Utilize Clean Sample Selection Algorithm

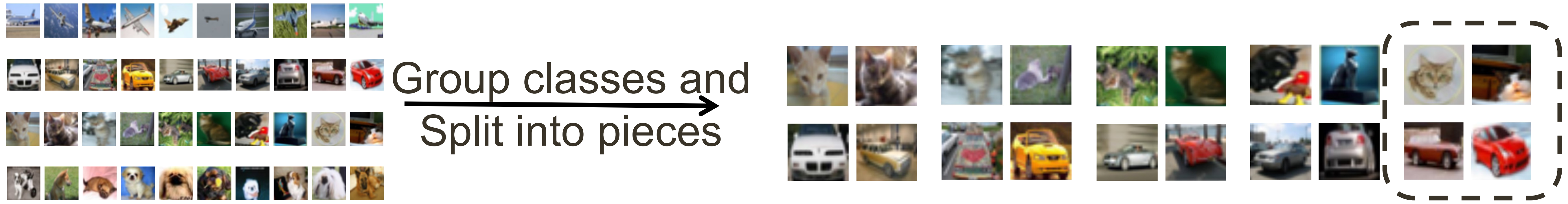
Encourage the linear relationship:

- In semi-supervised few-shot learning:
  We have pre-trained feature extractor, and we have ground-truth clean training set.
- In learning with noisy labels:
  1. Our first attempt is to append a sparse penalty on the network prediction:

$$\ell(x_i, y_i) = 1_{i \in \mathcal{C}}(\ell_{\text{CE}}(x_i, y_i) + \lambda \|x_i^\top W_{\text{fc}}\|_q)$$

  2. We can use self-supervised training to pre-train the backbone.

Scale up to large datasets:



$$\longrightarrow \quad y_i = \boldsymbol{x}_i^\top \beta + \gamma_i + \varepsilon$$

Fully utilize the noisy data:

$$\tilde{\text{img}} = \boldsymbol{M} \odot \text{img}_{\text{clean}} + (1 - \boldsymbol{M}) \odot \text{img}_{\text{noisy}}$$
$$\tilde{\boldsymbol{y}} = \lambda \boldsymbol{y}_{\text{clean}} + (1 - \lambda) \boldsymbol{y}_{\text{noisy}}$$

$$\mathcal{L}\left(\tilde{\text{img}}, \tilde{\boldsymbol{y}}\right) = \mathcal{L}_{\text{CE}}\left(\tilde{\text{img}}, \tilde{\boldsymbol{y}}\right)$$

Yikai Wang et al. [CVPR20][TPAMI21][CVPR22][TPAMI23]

Classification Performance on Few-Shot Learning

The Averaged Accuracies With 95 percent Confidence Intervals Over 2000 Episodes on Several Datasets

Setting	Model	*mini*ImageNet		*tiered*ImageNet		CIFAR-FS		CUB	
		1shot	5shot	1shot	5shot	1shot	5shot	1shot	5shot
In.	Baseline* [20]	51.75±0.80	74.27±0.63	-	-	-	-	65.51±0.87	82.85±0.55
	Baseline++* [20]	51.87±0.77	75.68±0.63	-	-	-	-	67.02±0.90	83.58±0.54
	MatchingNet* [10]	52.91[1]±0.88	68.88[1]±0.69	-	-	-	-	72.36[1]±0.90	83.64[1]±0.60
	ProtoNet* [8]	54.16[1]±0.82	73.68[1]±0.65	-	-	72.20[3]	83.50[3]	71.88[1]±0.91	87.42[1]±0.48
	MAML* [7]	49.61[1]±0.92	65.72[1]±0.77	-	-	-	-	69.96[1]±1.01	82.70[1]±0.65
	RelationNet* [9]	52.48[1]±0.86	69.83[1]±0.68	-	-	-	-	67.59[1]±1.02	82.75[1]±0.58
	adaResNet [86]	56.88	71.94	-	-	-	-	-	-
	TapNet [87]	61.65	76.36	63.08	80.26	-	-	-	-
	CTM† [88]	64.12	80.51	68.41	84.28	-	-	-	-
	MetaOptNet [82]	64.09	80.00	65.81	81.75	72.60	84.30	-	-
Tran.	TPN [22]	59.46	75.65	58.68[4]	74.26[4]	65.89[4]	79.38[4]	-	-
	TEAM* [26]	60.07	75.90	-	-	70.43	81.25	80.16	87.17
	CAN+T [53]	67.19±0.55	80.64±0.35	73.21±0.58	84.93±0.38	-	-	-	-
	DPGN [56]	67.77±0.32	**84.60**±0.43	72.45±0.51	**87.24**±0.39	77.90±0.50	**90.20**±0.40	75.71±0.47	91.48±0.33
Semi.	MSkM + MTL	62.10[2]	73.60[2]	68.6[2]	81.00[2]	-	-	-	-
	TPN + MTL	62.70[2]	74.20[2]	72.10[2]	83.30[2]	-	-	-	-
	MSkM [23]	50.40	64.40	52.40	69.90	-	-	-	-
	TPN [22]	52.78	66.42	55.70	71.00	-	-	-	-
	LST [24]	70.10	78.70	77.70	85.20	-	-	-	-
Tran.	ICIC	71.29±0.59	83.12±0.33	76.13±0.62	86.73±0.36	78.47±0.60	86.41±0.36	90.38±0.42	94.30±0.20
	ICIR	72.39±0.62	83.27±0.33	77.48±0.62	86.84±0.36	79.19±0.63	86.66±0.36	90.89±0.43	94.36±0.20
Semi. 15/15	ICIC	70.97±0.56	82.69±0.33	76.00±0.60	86.19±0.36	78.44±0.58	86.10±0.36	89.89±0.42	94.00±0.20
	ICIR	72.32±0.58	82.78±0.33	76.98±0.61	86.24±0.36	79.20±0.58	86.14±0.36	90.45±0.42	94.00±0.20
Semi. 30/50	ICIC	71.43±0.62	83.41±0.35	78.01±0.63	86.86±0.37	80.25±0.58	86.99±0.36	91.75±0.39	94.42±0.20
	ICIR	**73.12**±0.65	83.28±0.37	**78.99**±0.66	86.76±0.39	**80.74**±0.61	87.16±0.36	**92.12**±0.40	**94.52**±0.20

Yikai Wang et al. Instance Credibility Inference for Few-Shot Learning. CVPR 2020.
Yikai Wang et al. How to Trust Unlabeled Data? Instance Credibility Inference for Few-Shot Learning. IEEE TPAMI 2021.

Classification Performance on Learning with Noisy Labels (synthetic label noise)

Dataset	Method	Sym. Noise Rate 0.2	Sym. 0.4	Sym. 0.6	Sym. 0.8	Asy. Noise Rate 0.2	Asy. 0.3	Asy. 0.4
CIFAR-10	Standard	85.7 ± 0.5	81.8 ± 0.6	73.7 ± 1.1	42.0 ± 2.8	88.0 ± 0.3	86.4 ± 0.4	84.9 ± 0.7
	Forgetting	86.0 ± 0.8	82.1 ± 0.7	75.5 ± 0.7	41.3 ± 3.3	89.5 ± 0.2	88.2 ± 0.1	85.0 ± 1.0
	Bootstrap	86.4 ± 0.6	82.5 ± 0.1	75.2 ± 0.8	42.1 ± 3.3	88.8 ± 0.5	87.5 ± 0.5	85.1 ± 0.3
	Forward	85.7 ± 0.4	81.0 ± 0.4	73.3 ± 1.1	31.6 ± 4.0	88.5 ± 0.4	87.3 ± 0.2	85.3 ± 0.6
	Decoupling	87.4 ± 0.3	83.3 ± 0.4	73.8 ± 1.0	36.0 ± 3.2	89.3 ± 0.3	88.1 ± 0.4	85.1 ± 1.0
	MentorNet	88.1 ± 0.3	81.4 ± 0.5	70.4 ± 1.1	31.3 ± 2.9	86.3 ± 0.4	84.8 ± 0.3	78.7 ± 0.4
	Co-teaching	89.2 ± 0.3	86.4 ± 0.4	79.0 ± 0.2	22.9 ± 3.5	90.0 ± 0.2	88.2 ± 0.1	78.4 ± 0.7
	Co-teaching+	89.8 ± 0.2	86.1 ± 0.2	74.0 ± 0.2	17.9 ± 1.1	89.4 ± 0.2	87.1 ± 0.5	71.3 ± 0.8
	IterNLD	87.9 ± 0.4	83.7 ± 0.4	74.1 ± 0.5	38.0 ± 1.9	89.3 ± 0.3	88.8 ± 0.5	85.0 ± 0.4
	RoG	89.2 ± 0.3	83.5 ± 0.4	77.9 ± 0.6	29.1 ± 1.8	89.6 ± 0.4	88.4 ± 0.5	86.2 ± 0.6
	PENCIL	88.2 ± 0.2	86.6 ± 0.3	74.3 ± 0.6	45.3 ± 1.4	90.2 ± 0.2	88.3 ± 0.2	84.5 ± 0.5
	GCE	88.7 ± 0.3	84.7 ± 0.4	76.1 ± 0.3	41.7 ± 1.0	88.1 ± 0.3	86.0 ± 0.4	81.4 ± 0.6
	SL	89.2 ± 0.5	85.3 ± 0.7	78.0 ± 0.3	44.4 ± 1.1	88.7 ± 0.3	86.3 ± 0.1	81.4 ± 0.7
	TopoFilter	90.2 ± 0.2	87.2 ± 0.4	80.5 ± 0.4	45.7 ± 1.0	90.5 ± 0.2	89.7 ± 0.3	87.9 ± 0.2
	SPR	92.0 ± 0.1	$\mathbf{94.6 \pm 0.2}$	91.6 ± 0.2	80.5 ± 0.6	89.0 ± 0.8	90.3 ± 0.8	91.0 ± 0.6
	Knockoffs-SPR	$\mathbf{95.4 \pm 0.1}$	94.5 ± 0.1	$\mathbf{93.3 \pm 0.1}$	$\mathbf{84.6 \pm 0.8}$	$\mathbf{95.1 \pm 0.1}$	$\mathbf{94.5 \pm 0.2}$	$\mathbf{93.6 \pm 0.2}$
CIFAR-100	Standard	56.5 ± 0.7	50.4 ± 0.8	38.7 ± 1.0	18.4 ± 0.5	57.3 ± 0.7	52.2 ± 0.4	42.3 ± 0.7
	Forgetting	56.5 ± 0.7	50.6 ± 0.9	38.7 ± 1.0	18.4 ± 0.4	57.5 ± 1.1	52.4 ± 0.8	42.4 ± 0.8
	Bootstrap	56.2 ± 0.5	50.8 ± 0.6	37.7 ± 0.8	19.0 ± 0.6	57.1 ± 0.9	53.0 ± 0.9	43.0 ± 1.0
	Forward	56.4 ± 0.4	49.7 ± 1.3	38.0 ± 1.5	12.8 ± 1.3	56.8 ± 1.0	52.7 ± 0.5	42.0 ± 1.0
	Decoupling	57.8 ± 0.4	49.9 ± 1.0	37.8 ± 0.7	17.0 ± 0.7	60.2 ± 0.9	54.9 ± 0.1	47.2 ± 0.9
	MentorNet	62.9 ± 1.2	52.8 ± 0.7	36.0 ± 1.5	15.1 ± 0.9	62.3 ± 1.3	55.3 ± 0.5	44.4 ± 1.6
	Co-teaching	64.8 ± 0.2	60.3 ± 0.4	46.8 ± 0.7	13.3 ± 2.8	63.6 ± 0.4	58.3 ± 1.1	48.9 ± 0.8
	Co-teaching+	64.2 ± 0.4	53.1 ± 0.2	25.3 ± 0.5	10.1 ± 1.2	60.9 ± 0.3	56.8 ± 0.5	48.6 ± 0.4
	IterNLD	57.9 ± 0.4	51.2 ± 0.4	38.1 ± 0.9	15.5 ± 0.8	58.1 ± 0.4	53.0 ± 0.3	43.5 ± 0.8
	RoG	63.1 ± 0.3	58.2 ± 0.5	47.4 ± 0.8	20.0 ± 0.9	67.1 ± 0.6	65.6 ± 0.4	58.8 ± 0.1
	PENCIL	64.9 ± 0.3	61.3 ± 0.4	46.6 ± 0.7	17.3 ± 0.8	67.5 ± 0.5	66.0 ± 0.4	61.9 ± 0.4
	GCE	63.6 ± 0.6	59.8 ± 0.5	46.5 ± 1.3	17.0 ± 1.1	64.8 ± 0.9	61.4 ± 1.1	50.4 ± 0.9
	SL	62.1 ± 0.4	55.6 ± 0.6	42.7 ± 0.8	19.5 ± 0.7	59.2 ± 0.6	55.1 ± 0.7	44.8 ± 0.1
	TopoFilter	65.6 ± 0.3	62.0 ± 0.6	47.7 ± 0.5	20.7 ± 1.2	68.0 ± 0.3	66.7 ± 0.6	62.4 ± 0.2
	SPR	72.5 ± 0.2	$\mathbf{75.0 \pm 0.1}$	$\mathbf{70.9 \pm 0.3}$	$\mathbf{38.1 \pm 0.8}$	71.9 ± 0.2	72.4 ± 0.3	70.9 ± 0.5
	Knockoffs-SPR	$\mathbf{77.5 \pm 0.2}$	74.3 ± 0.2	67.8 ± 0.4	30.5 ± 1.0	$\mathbf{77.3 \pm 0.4}$	$\mathbf{76.3 \pm 0.3}$	$\mathbf{73.9 \pm 0.6}$

Yikai Wang et al. Scalable Penalized Regression for Noise Detection in Learning with Noisy Labels. CVPR 2022.
Yikai Wang et al. Knockoffs-SPR: Clean Sample Selection in Learning with Noisy Labels. TPAMI 2023.

Classification Performance on Learning with Noisy Labels (real-world label noise)

TABLE 2
Test accuracies(%) on WebVision and ILSVRC12.

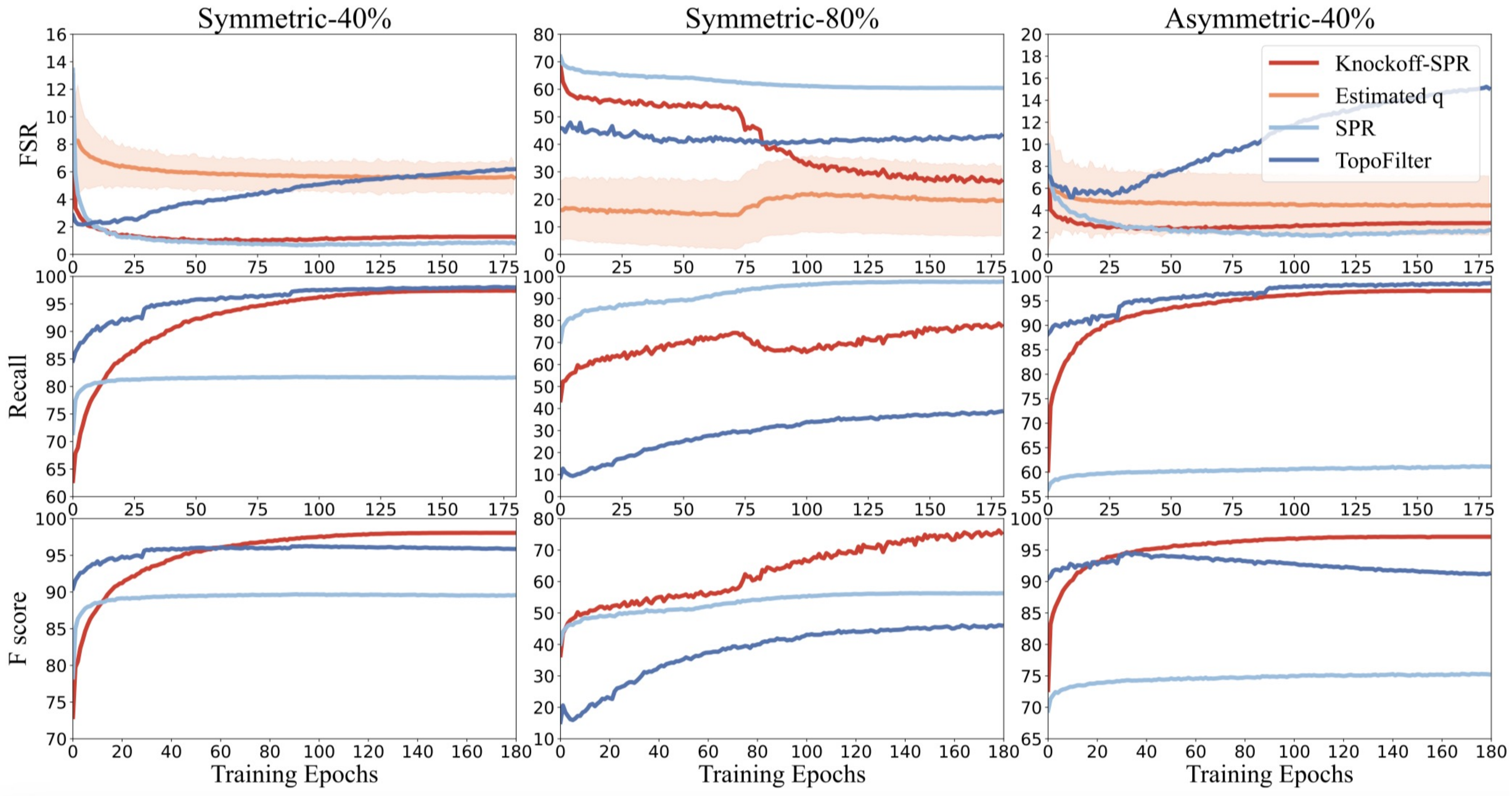
Method	WebVision top1	WebVision top5	WebVision → ILSVRC12 top1	WebVision → ILSVRC12 top5
F-correction	61.12	82.68	57.36	82.36
Decoupling	62.54	84.74	58.26	82.26
D2L	62.68	84.00	57.80	81.36
MentorNet	63.00	81.40	57.80	79.92
Co-teaching	63.58	85.20	61.48	84.70
Iterative-CV	65.24	85.34	61.60	84.98
DivideMix	77.32	91.64	**75.20**	90.84
SPR	77.08	91.40	72.32	90.92
Knockoffs-SPR	**77.96**	**92.28**	74.72	**92.88**

TABLE 3
Test accuracies(%) on Clothing1M.

Method	Accuracy
Cross-Entropy	69.21
F-correction	69.84
M-correction	71.00
Joint-Optim	72.16
Meta-Cleaner	72.50
Meta-Learning	73.47
P-correction	73.49
TopoFiler	74.10
DivideMix	74.76
SPR	71.16
Knockoffs-SPR	**75.20**

Yikai Wang et al. Scalable Penalized Regression for Noise Detection in Learning with Noisy Labels. CVPR 2022.
Yikai Wang et al. Knockoffs-SPR: Clean Sample Selection in Learning with Noisy Labels. TPAMI 2023.

Sample Selection Performance



Yikai Wang et al. Scalable Penalized Regression for Noise Detection in Learning with Noisy Labels. CVPR 2022.
Yikai Wang et al. Knockoffs-SPR: Clean Sample Selection in Learning with Noisy Labels. TPAMI 2023.

Summary

- **Ideologically**, we focus on the clean sample selection where the training data is not accurately-labeled.
- **Methodologically**, we propose a series of methods to identify clean samples in the training dataset, with a focus on sufficient noisy set recovery and false-selection-rate control, respectively.
- **Theoretically**, we prove the noisy set recovery theorem and false-selection-rate control theorem, to provide theoretical guarantees of our proposed methods.
- **Algorithmically**, we design algorithms to better train the learning model with our proposed clean sample selection algorithms, enabling balanced identifiability and complexity to scale up to large datasets.
- **Experimentally**, we demonstrate the effectiveness and efficiency of our method on semi-supervised few-shot learning and learning with noisy labels.

Yikai Wang et al. Instance Credibility Inference for Few-Shot Learning. CVPR 2020.
Yikai Wang et al. How to Trust Unlabeled Data? Instance Credibility Inference for Few-Shot Learning. IEEE TPAMI 2021.
Yikai Wang et al. Scalable Penalized Regression for Noise Detection in Learning with Noisy Labels. CVPR 2022.
Yikai Wang et al. Knockoffs-SPR: Clean Sample Selection in Learning with Noisy Labels. TPAMI 2023.

THANKS