Sparse Learning for Noisy Data/Labels:

A Simple yet Effective Framework for Vision Applications

Yikai Wang
yikai-wang.github.io

Yanwei Fu
http://yanweifu.github.io

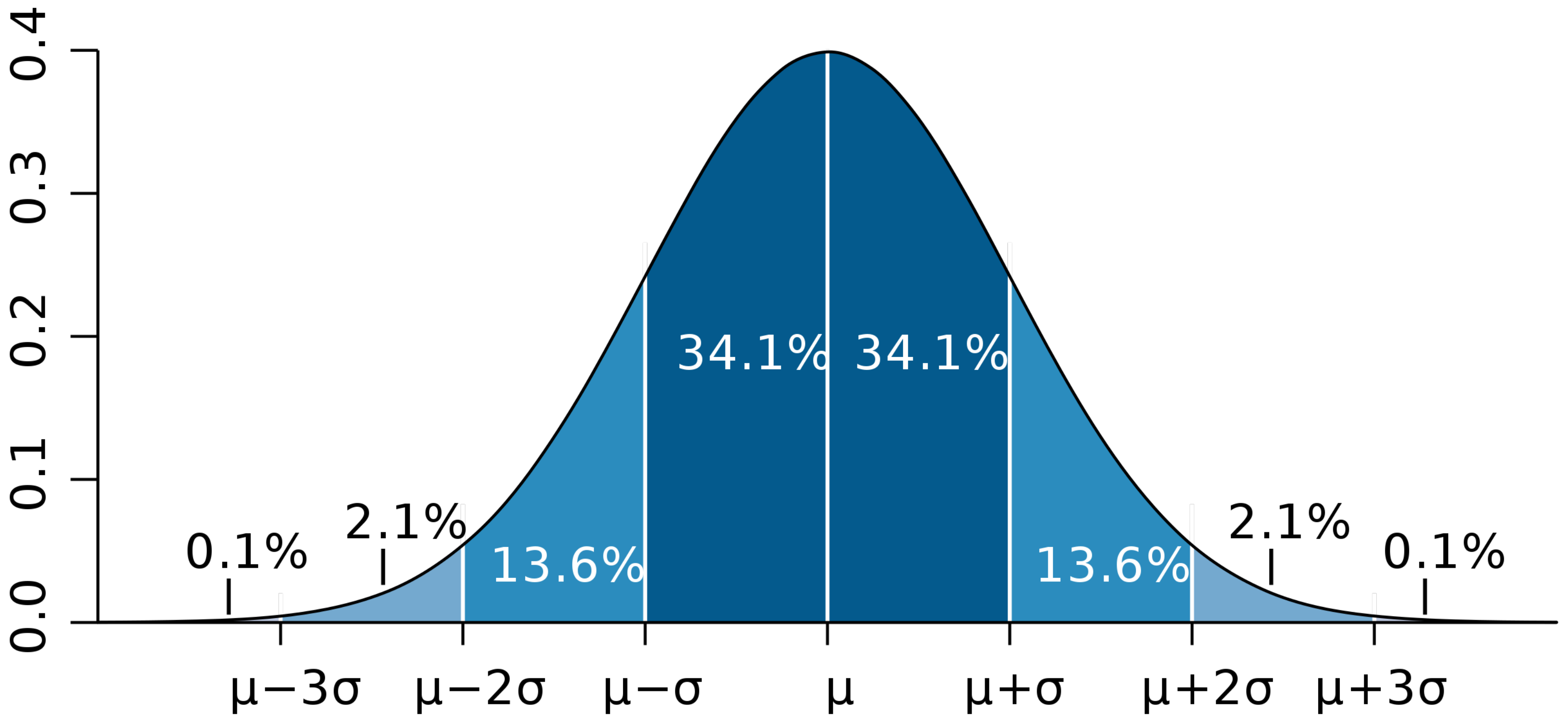
School of Data Science
Fudan University

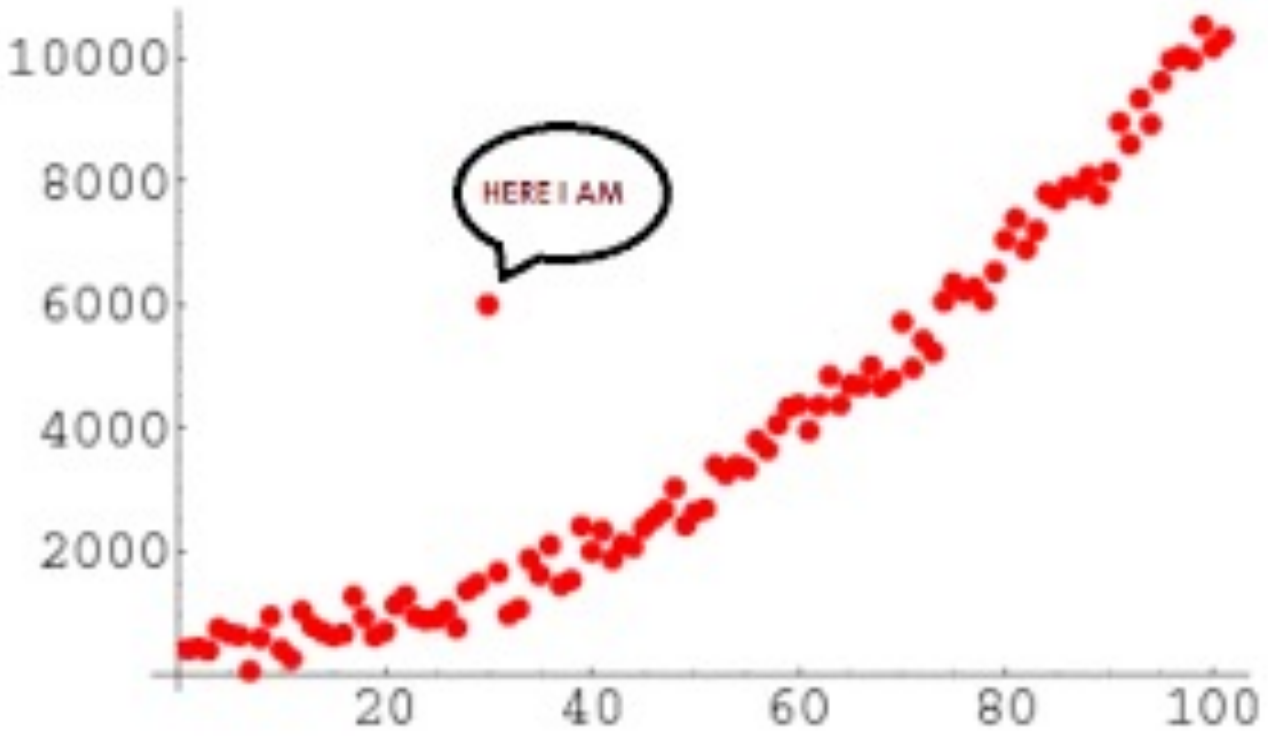
Sparse Learning

for Noisy Data Detection

大数据学院
School of Data Science

Examples of Noisy Data/Outliers





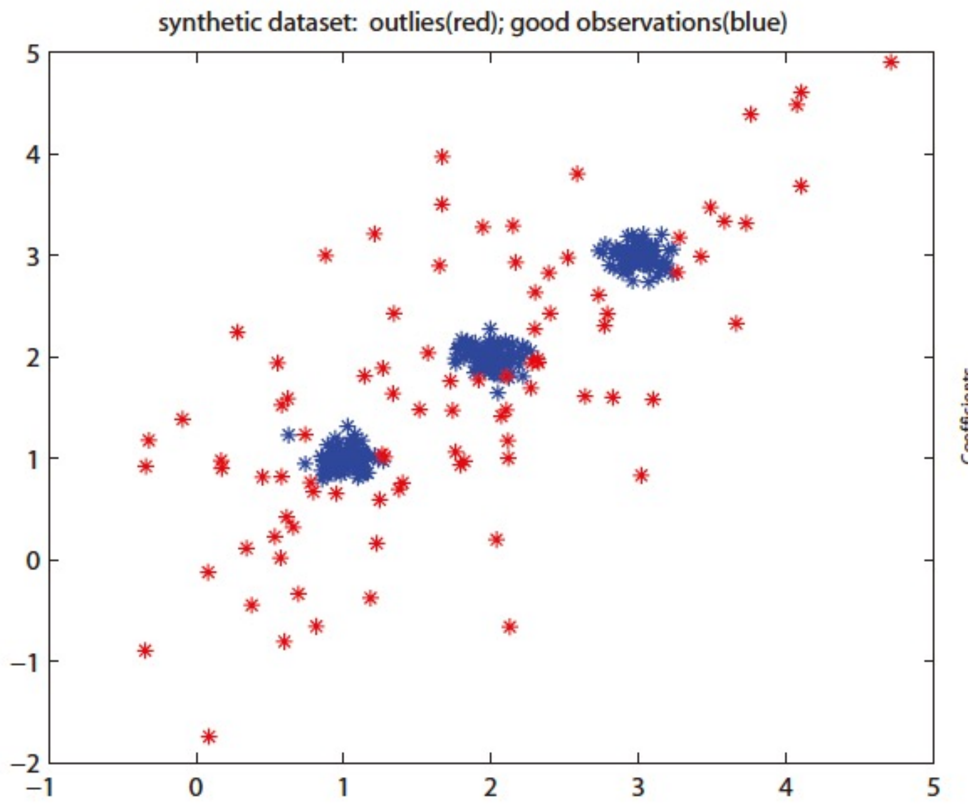
Outliers are the **irregular** data compared with the majority of the dataset.

Figures from
[1] towardsdatascience.com/this-article-is-about-identifying-outliers-through-funnel-plots-using-the-microsoft-power-bi-d7ad16ac9ccc
[2] en.wikipedia.org/wiki/Outlier#/media/File:Standard_deviation_diagram_micro.svg
[3] medium.com/analytics-vidhya/its-all-about-outliers-cbe172aa1309

大数据学院
School of Data Science

Noisy Data in Label Space

- Random Corruptions



- Noisy search engine results

Shogun: Total War - IGN
ign.com

Aug 21, 1192 CE: First S...
nationalgeographic.org

Baal Ascension Materials: What To Farm For G...
forbes.com

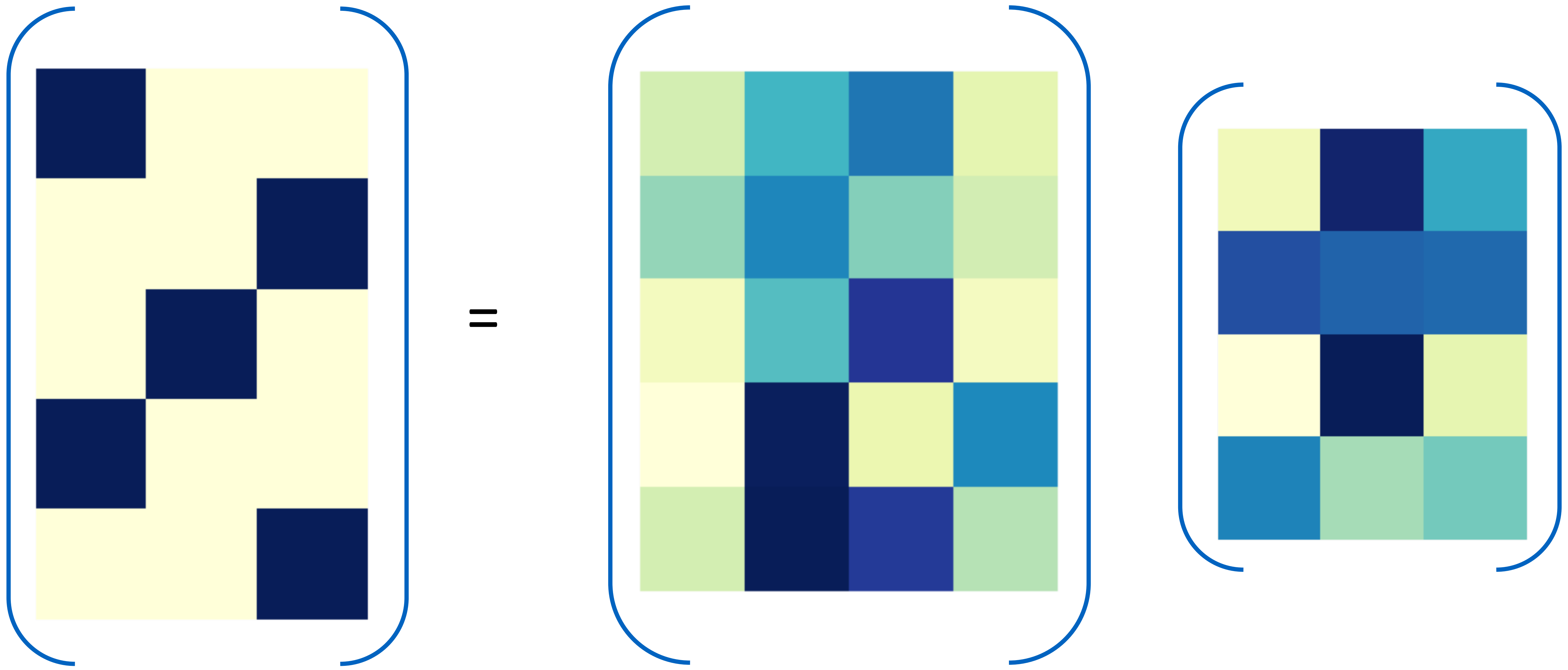
- Annotator mistakes

- Complex/Confusing items identified

大数据学院
School of Data Science

Identify Noisy Data in Label Space

Linear system $$Y = X\beta$$

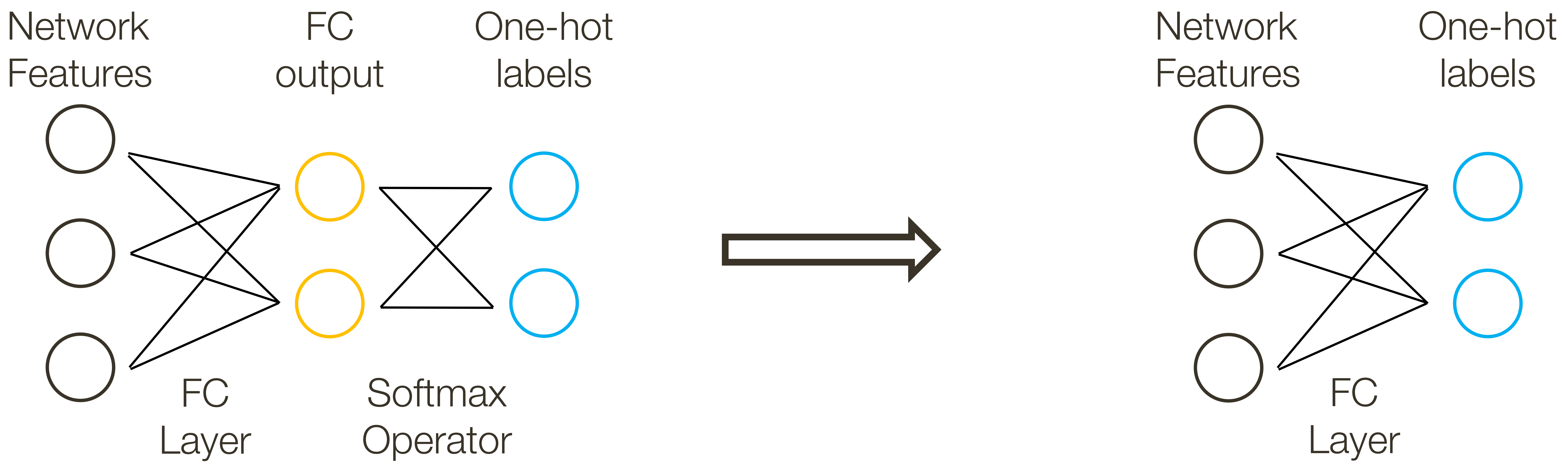


Noisy One-hot Labels **Deep Features** **Fitted Coef.**

$$Y \in \mathbb{R}^{n\times c} \quad X \in \mathbb{R}^{n\times d} \quad \beta \in \mathbb{R}^{d\times c}$$

β is sensitive to noisy data!

Approximated Linear Assumption in Networks



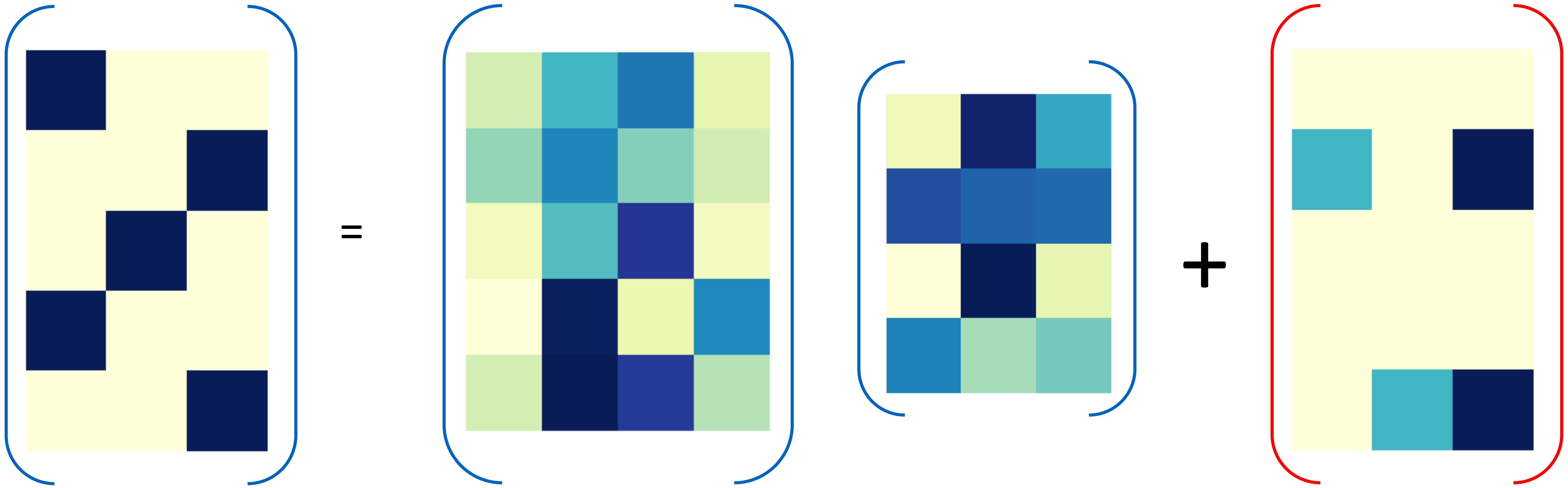
$$y_i = \mathrm{SoftMax}(\boldsymbol{x}_i^\top \beta)$$

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

Wang et al. Scalable Penalized Regression for Noise Detection in Learning with Noisy Labels. CVPR 2022.

Identify Noisy Data in Label Space: The Indicator

Linear system
with Noisy Data/Labels $Y = X\beta + \gamma$



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

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

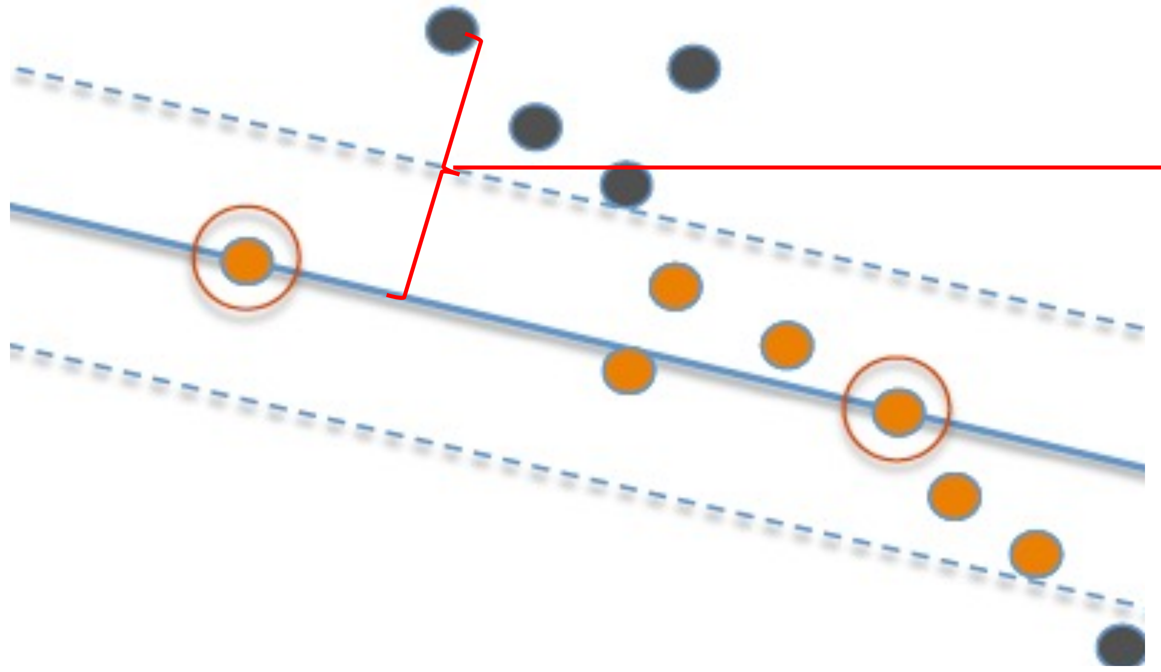
Fitted Coef. $\beta \in \mathbb{R}^{d\times c}$

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] [Wang et al. CVPR 20, TPAMI 21, CVPR 22]

Understanding γ in Statistics

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



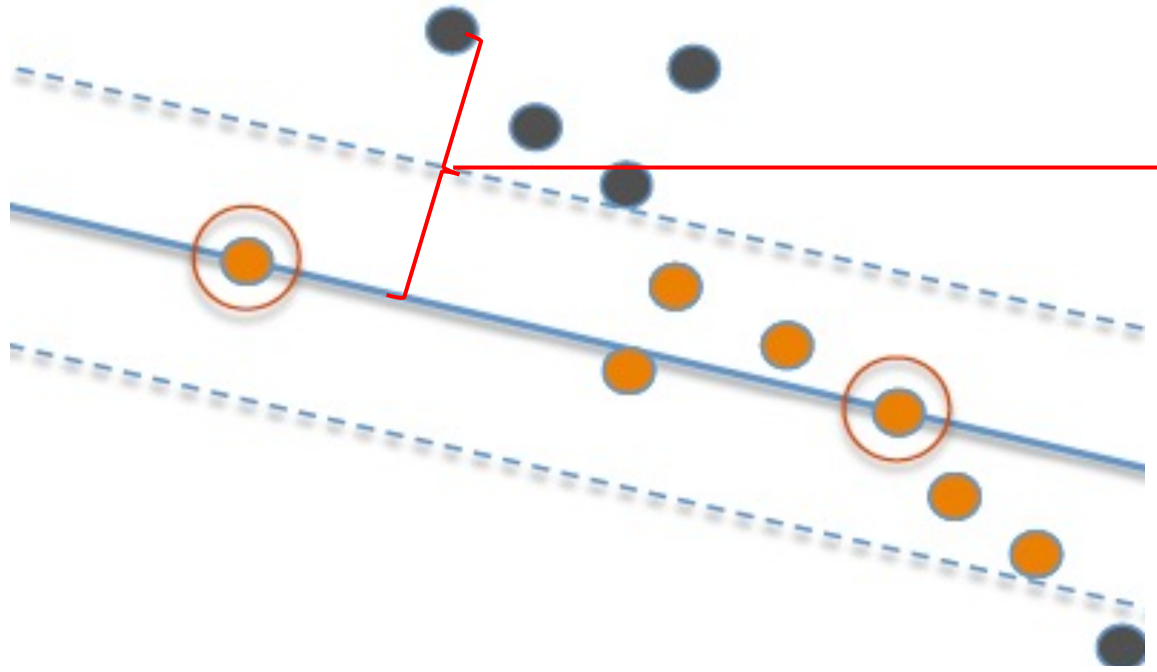
γ_i equals to the residual predict error $\gamma_i = y_i - x_i^\top \hat{\beta}$

Row residuals fail to detect outliers at *leverage points.*

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

Understanding γ in Statistics

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



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

Leave-one-out externally studentized residual:

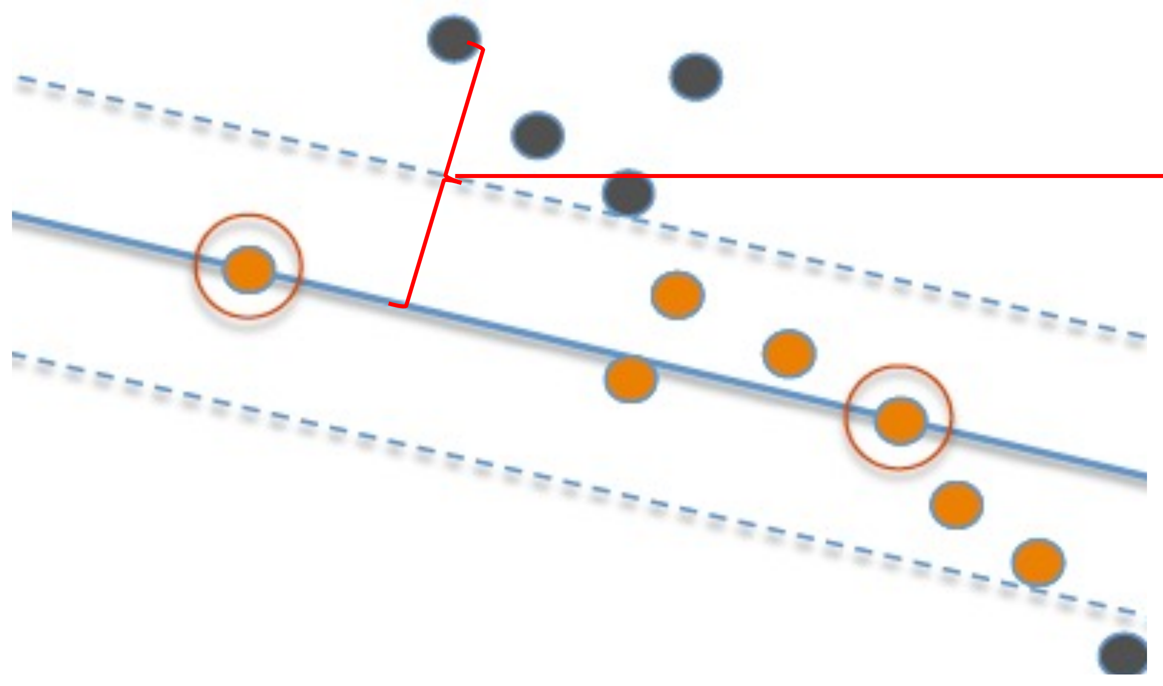
$$t_i = \frac{y_i - \boldsymbol{x}_i^\top \hat{\beta}_{(i)}}{\hat{\sigma}_{(i)}(1+\boldsymbol{x}_i(\boldsymbol{X}_{(i)}^\top \boldsymbol{X}_{(i)})^{-1}\boldsymbol{x}_i)^{1/2}}$$

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

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

Understanding γ in Statistics

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



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 $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \gamma 1_i + \boldsymbol{\varepsilon}$.

When there are multiple outliers:

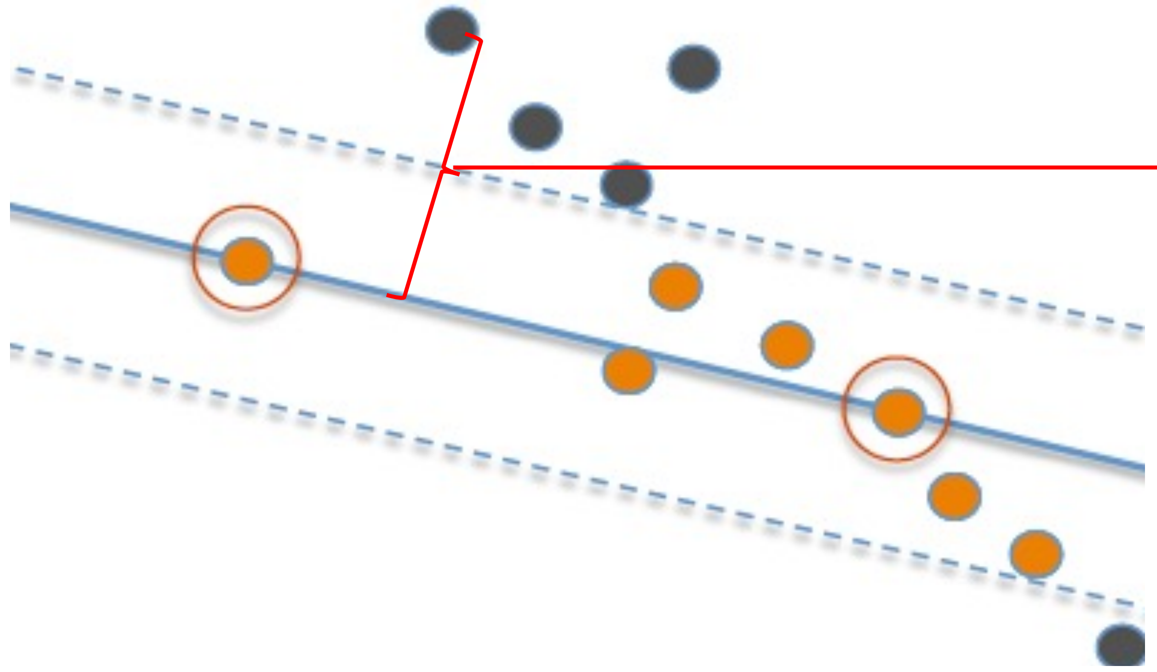
1. masking: multiple outliers may mask each other and being undetected;

2. swamping: multiple outliers may lead the large t_i for clean data.

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

Understanding γ in Statistics

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



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

Leave-one-out externally studentized residual:

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

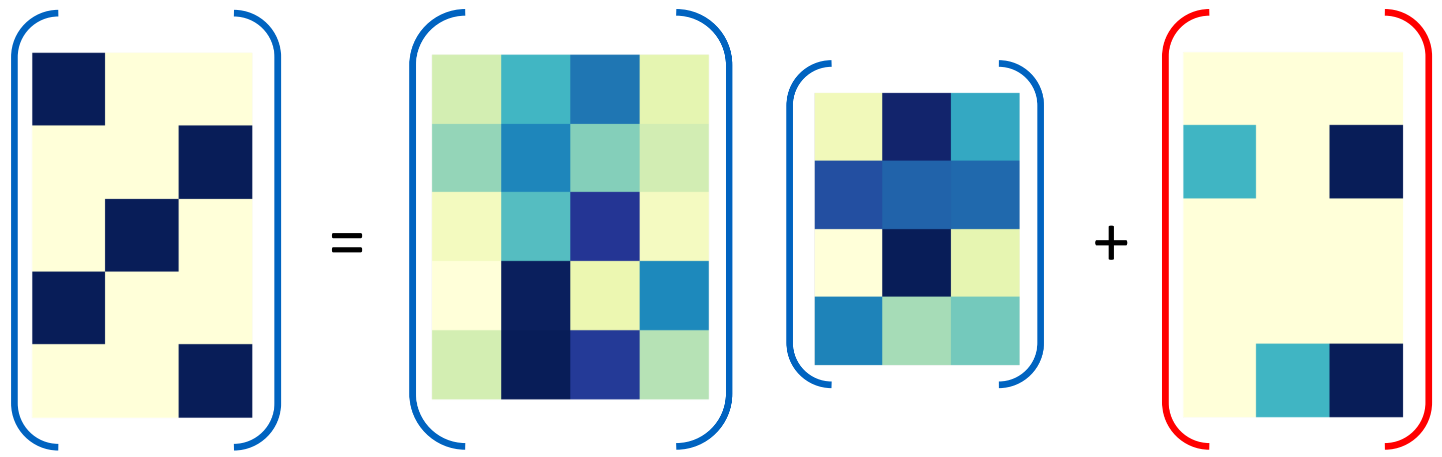
$\Leftrightarrow$ test whether $\gamma = 0$ in $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \gamma 1_i + \boldsymbol{\varepsilon}$.

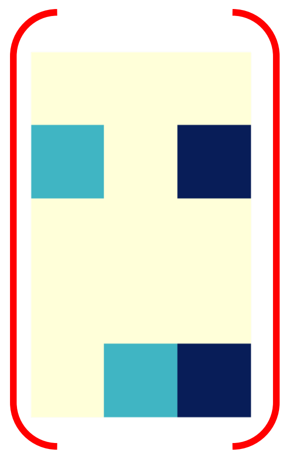
$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} + \boldsymbol{\gamma}$$

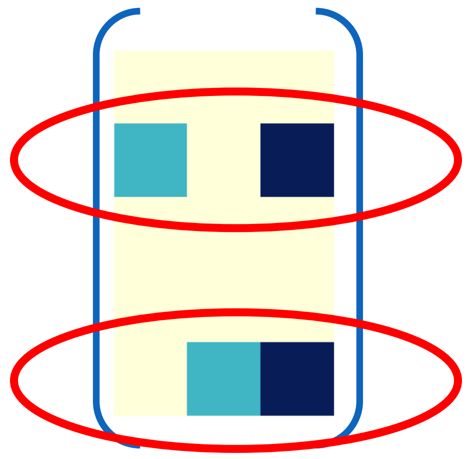
[1] Yiyuan She and Art B Owen. Outlier detection using nonconvex penalized regression. Journal of the American Statistical Association, 2011.

Identify Noisy Data in the Dataset

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







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

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

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

Wang et al. Scalable Penalized Regression for Noise Detection in Learning with Noisy Labels. CVPR 2022.

Simplification

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

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

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

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

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

A linear regression problem!

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

Solving Gamma in Linear Regression

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

How to select λ?

- heuristics rules $\lambda = 2.5\hat{\sigma}$?
- Cross-validation?
- Data adaptive techniques?
- AIC, BIC?

It is hard to select a proper λ.

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

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

With $R(\gamma) = \sum_{i=1}^{n} \|\gamma_i\|_2$,
γ vanishes instance by instance.

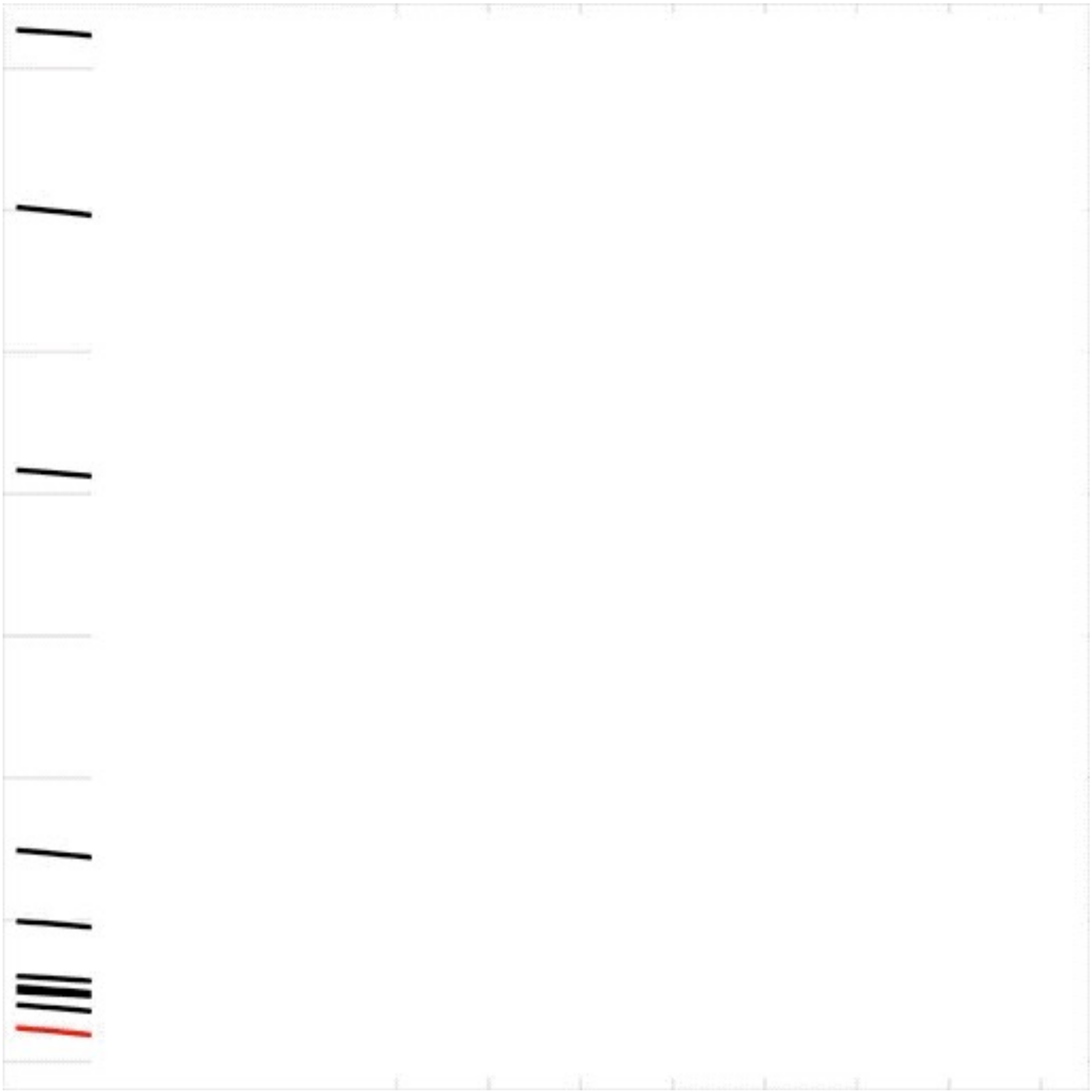
$C_i = \sup\{\lambda : \|\hat{\gamma}_i(\lambda)\| \neq 0\}$

This can be sovled by GLMnet[1].

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

Solving Gamma in Linear Regression

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



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

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

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

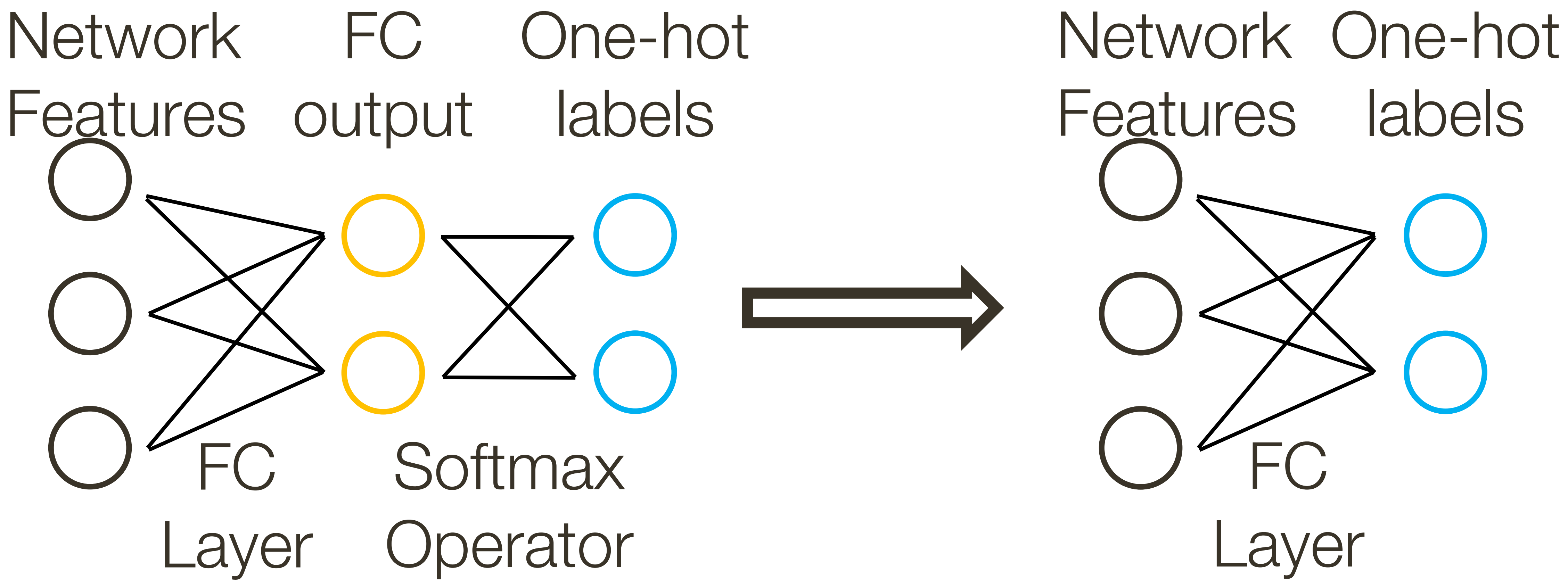
$C_i = \sup\{\lambda : \|\hat{\gamma}_i(\lambda)\| \neq 0\}$

This can be sovled by GLMnet[1].

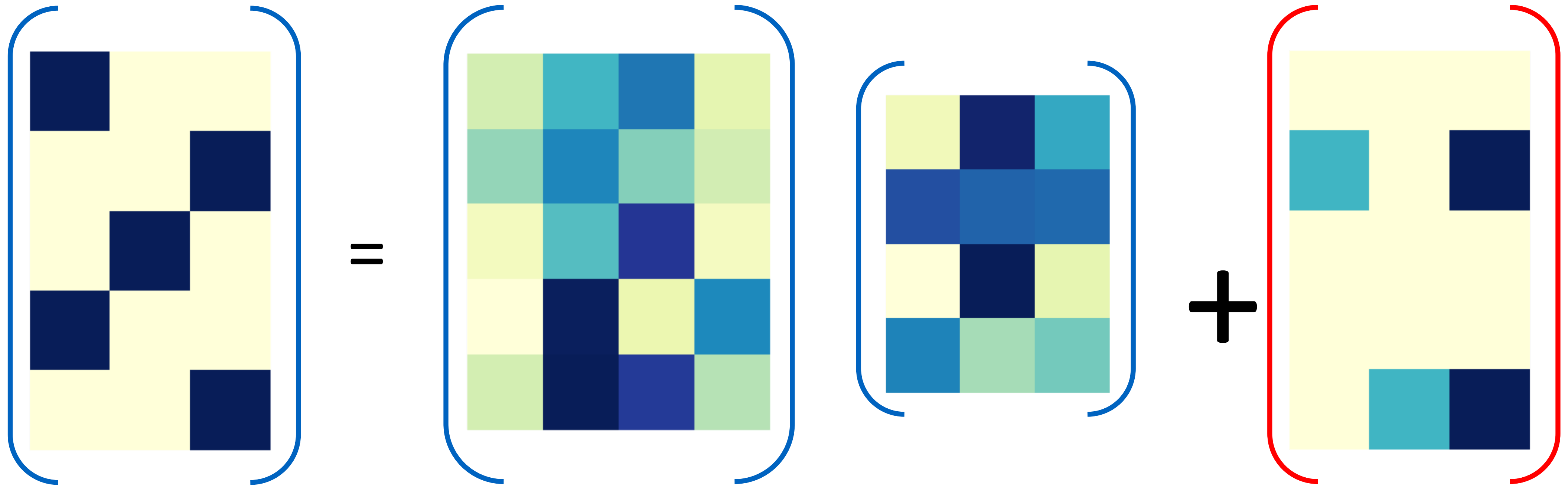
[1] Friedman, et al. 2010. "Regularization Paths for Generalized Linear Models via Coordinate Descent." Journal of Statistical Software.

大数据学院 School of Data Science

Instance Credibility Inference



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



$$Y \in \mathbb{R}^{n\times c} \quad X \in \mathbb{R}^{n\times d} \quad \beta \in \mathbb{R}^{d\times c} \quad \gamma \in \mathbb{R}^{n\times c}$$

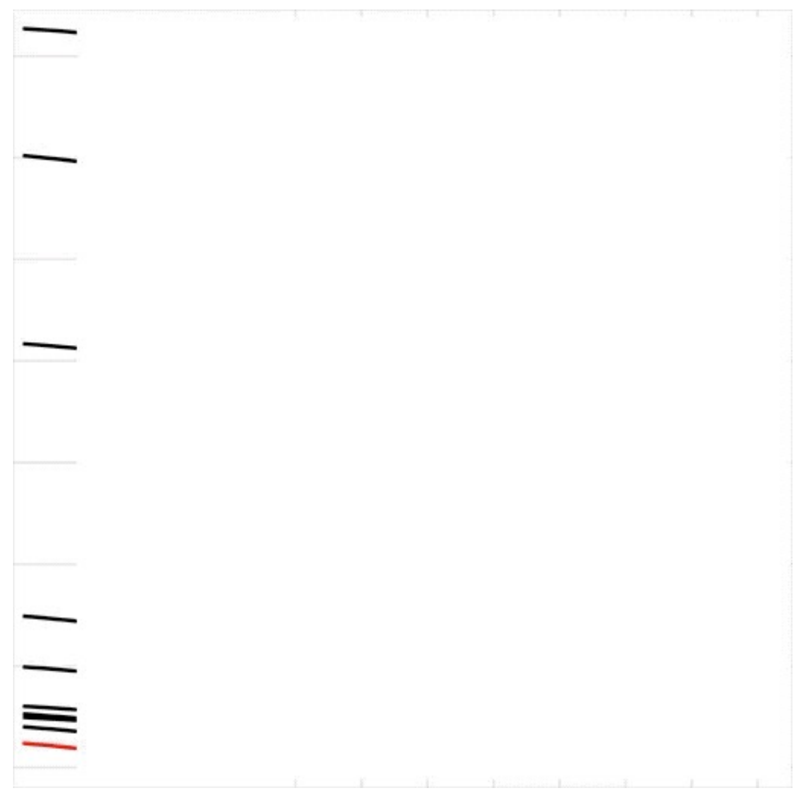
$$\underset{\boldsymbol{\beta},\boldsymbol{\gamma}}{\operatorname{argmin}} L(\boldsymbol{\beta},\boldsymbol{\gamma}) := \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\gamma}\|_{\mathrm{F}}^2 + \lambda R(\boldsymbol{\gamma})$$

$$\downarrow$$

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

$$\downarrow$$

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



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

大数据学院 School of Data Science

Noise Set Recovery

When will the model identify all the outliers?

Assume $\boldsymbol{\varepsilon}$ is i.i.d zero-mean sub-Gaussian noise. We give three conditions:

- (C1: Restricted eigenvalue)

$$\lambda_{\min}\left(\tilde{\boldsymbol{U}}_S^\top \tilde{\boldsymbol{U}}_S\right) = C_{\min} > 0.$$

- (C2: Irrepresentability) $\exists \eta \in (0,1]$,

$$\left\| \tilde{\boldsymbol{U}}_{S^c}^\top \tilde{\boldsymbol{U}}_S \left(\tilde{\boldsymbol{U}}_S^\top \tilde{\boldsymbol{U}}_S\right)^{-1} \right\|_\infty \le 1 - \eta.$$

- (C3: Large error)

$$\vec{\gamma}_{\min} := \min_{i \in S} |\vec{\gamma}^*| > h\left(\lambda, \eta, \tilde{\boldsymbol{U}}, \vec{\gamma}^*\right).$$

[1]: M. J. Wainwright, Sharp thresholds for high-dimensional and noisy sparsity recovery using l1-constrained quadratic programming. TIT 2009.

A non-asymptotic probabilistic result

Based on these conditions, we could provide the following theorem:

Theorem 1 (Identifiability of ICI). *Let* $\lambda \geq \frac{2\sigma\sqrt{\mu_{\tilde{U}}}}{\eta}\sqrt{\log cn}$. *Then with probability greater than* $1 - 2\,(cn)^{-1}$, *the problem has a unique solution* $\hat{\gamma}$ *satisfies the following properties:*

1) If C1 and C2 hold, the wrong-predicted instances indicated by ICI has no false positive error, i.e., $\hat{S} \subseteq S$ *and hence* $\hat{O} \subseteq O$, *and*

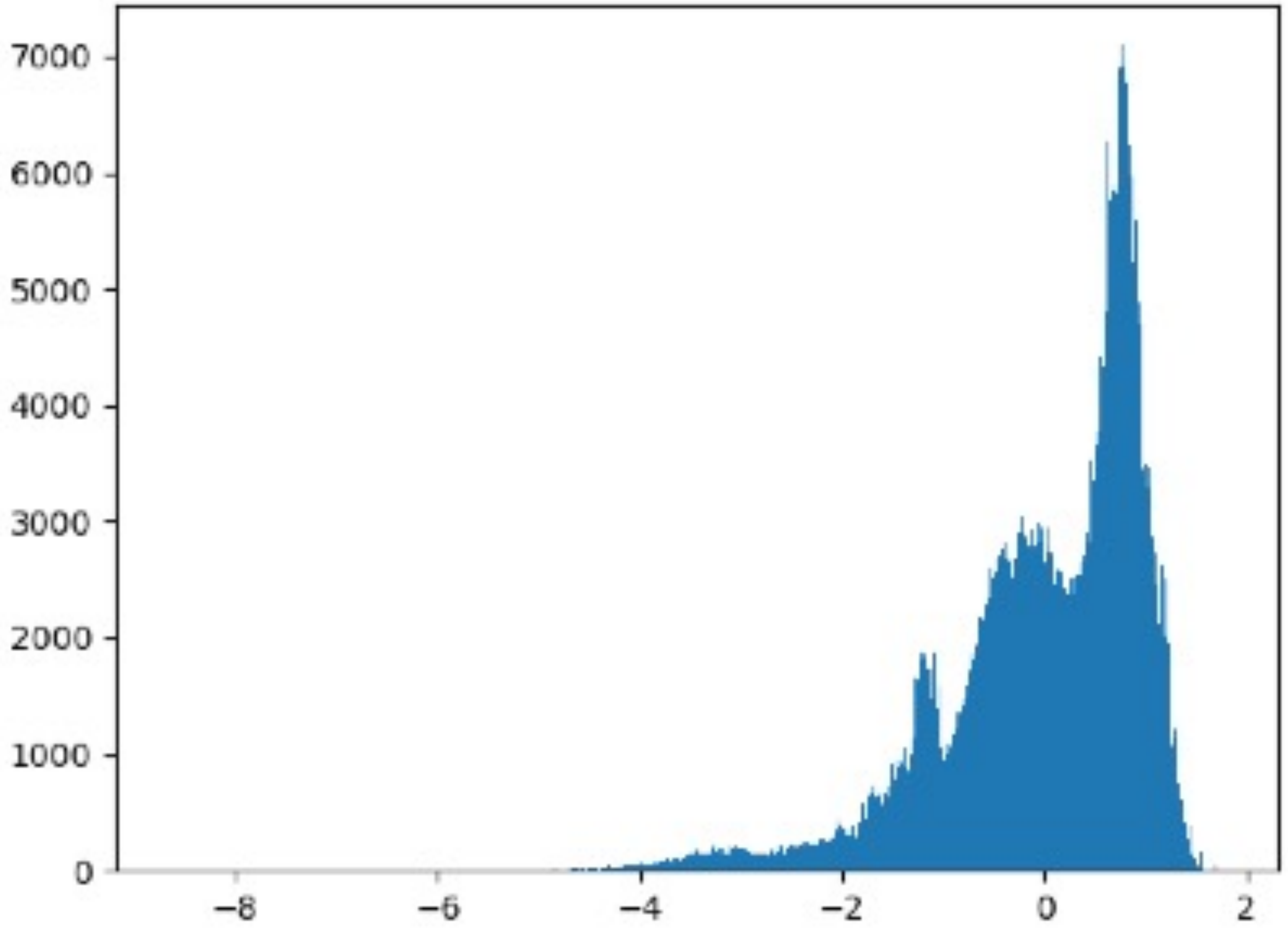
$$\left\|\hat{\vec{\gamma}}_S - \vec{\gamma}^*_S\right\|_\infty \leq h\left(\lambda, \eta, \tilde{\boldsymbol{U}}, \vec{\gamma}^*\right);$$

2) If C1, C2, and C3 hold, ICI will identify all the correctly-predicted instance, i.e., $\hat{S} = S$ *and hence* $\hat{O} = O$ *(in fact* $\mathrm{sign}\left(\hat{\vec{\gamma}}\right) = \mathrm{sign}\left(\vec{\gamma}^*\right)$*).*

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

大数据学院 School of Data Science

Identifiability in reality: sub-Gaussian noise



$$\mathbb{E}\left[\hat{\varepsilon}\right] \approx 10^{-19}$$

$$\mathrm{Var}\left[\hat{\varepsilon}\right] \approx 0.99$$

Sparse Learning

in Few-Shot Learning

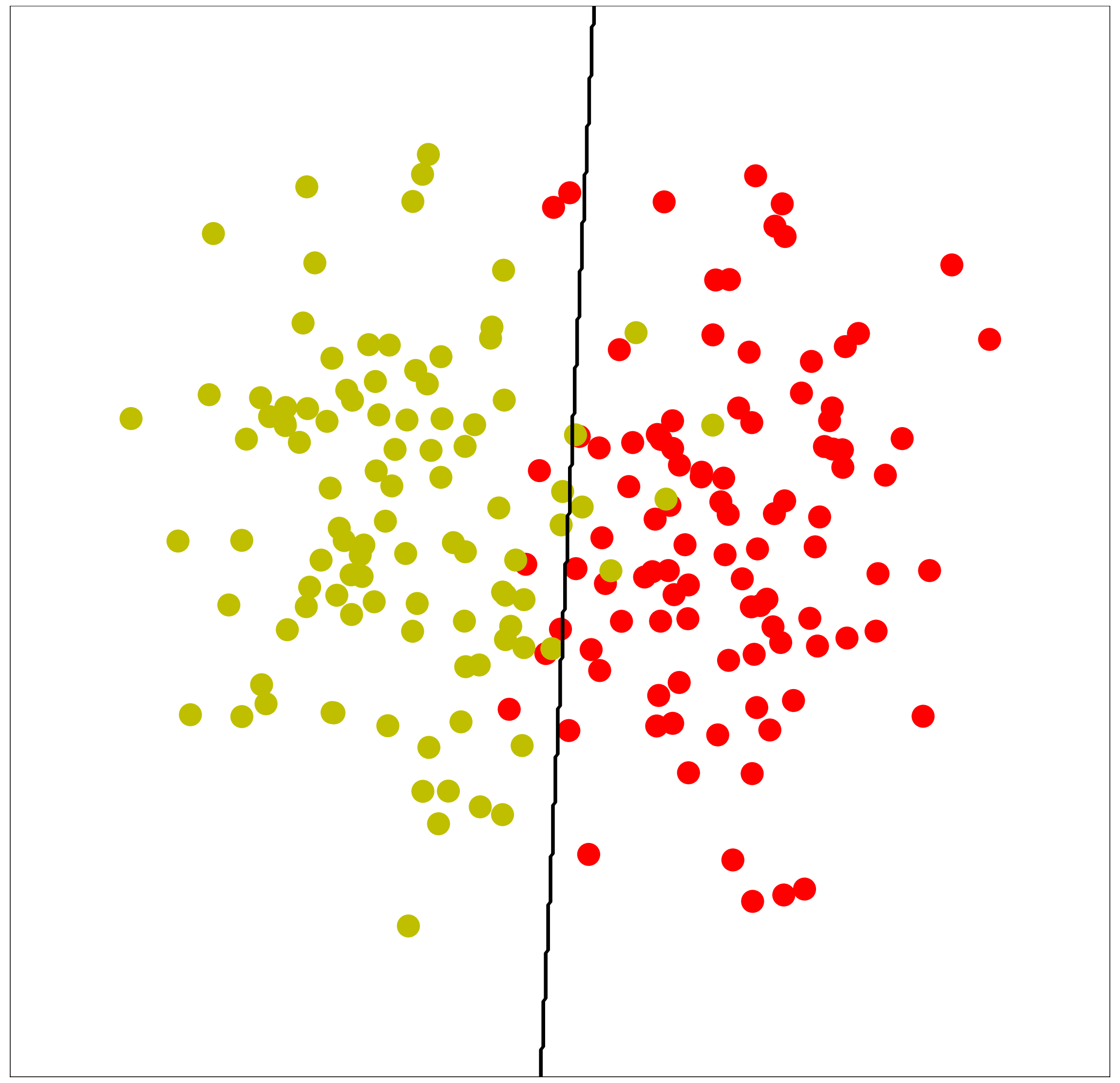
Definition of Few-Shot Learning

Tackle machine learning problem with only limited training data provided.

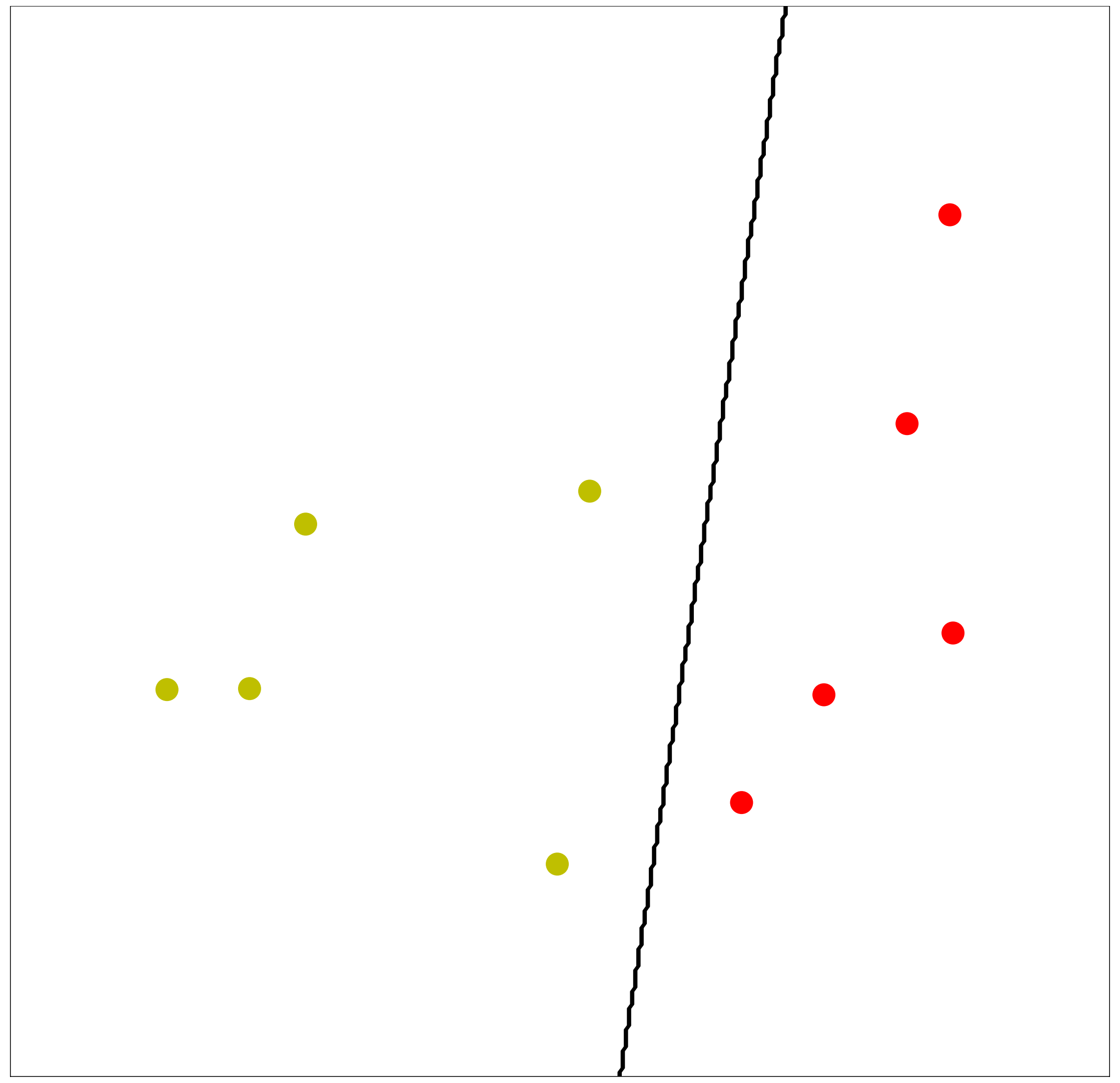
Few-Shot Learning

Low cost → **High cost**

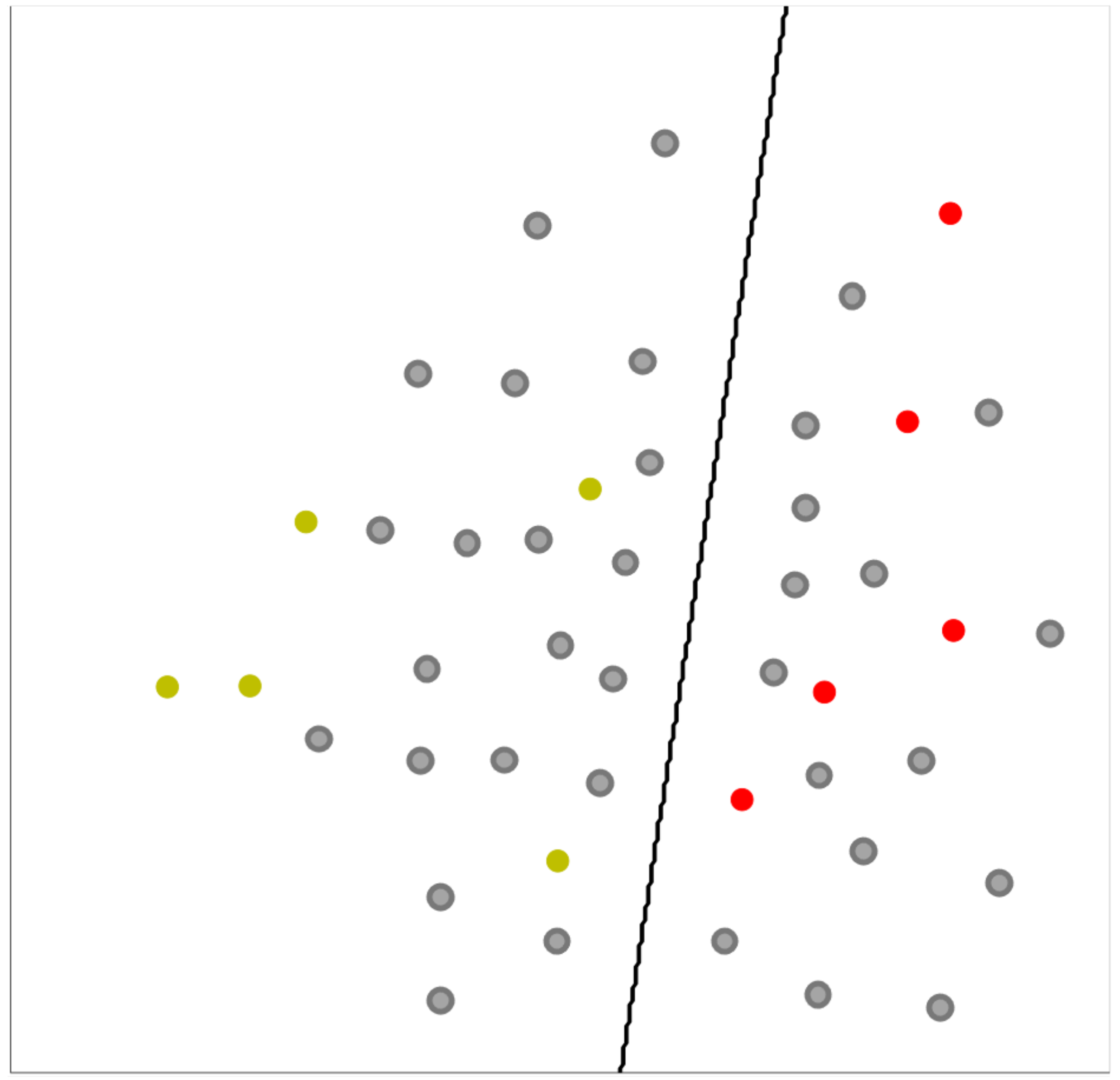
Unsupervised Learning | **Learning From Noisy Labels** | **Semi-Supervised Learning** | **Supervised Learning**



Binary classification
with many labeled data

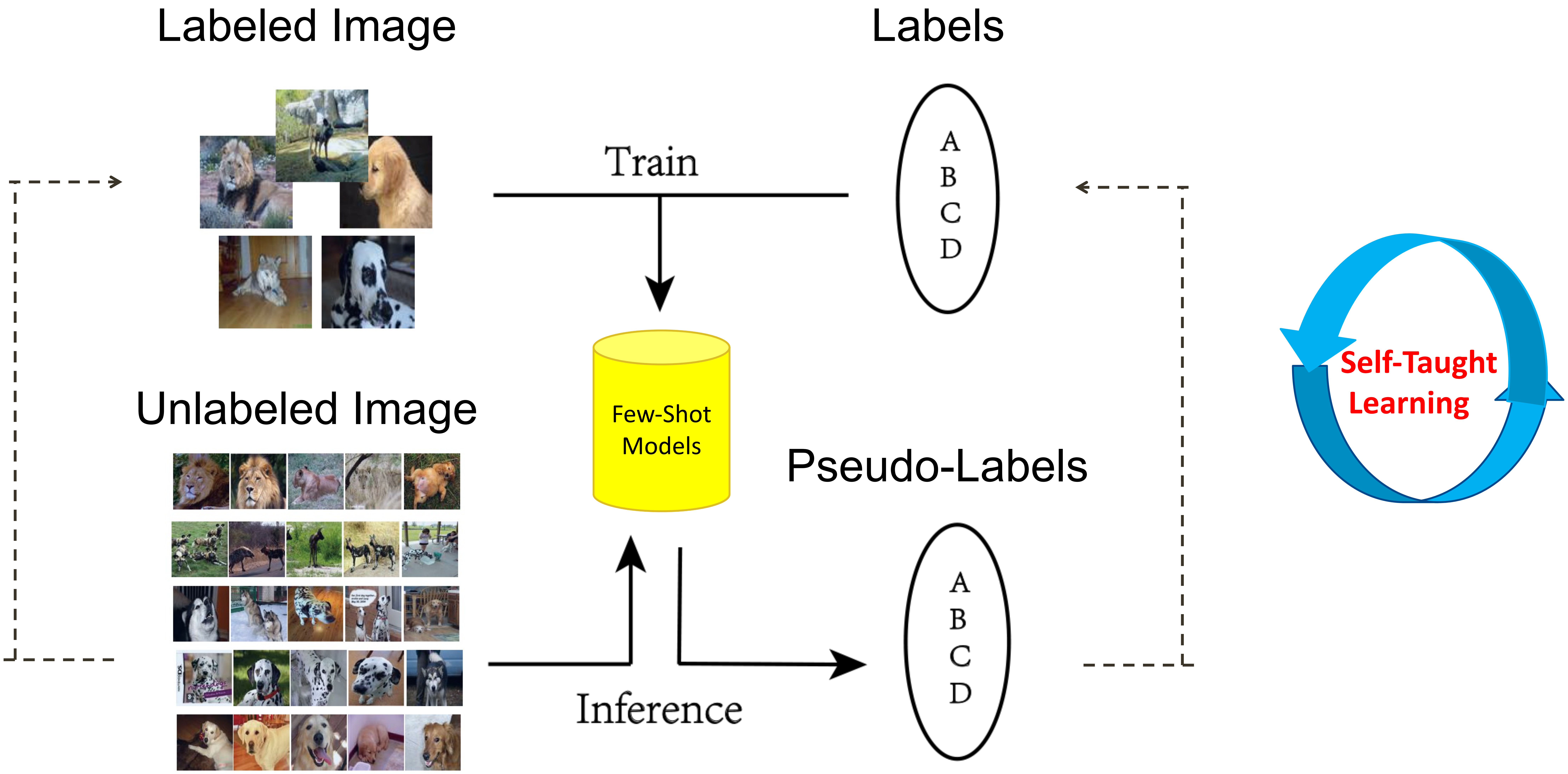


Few-shot binary classification



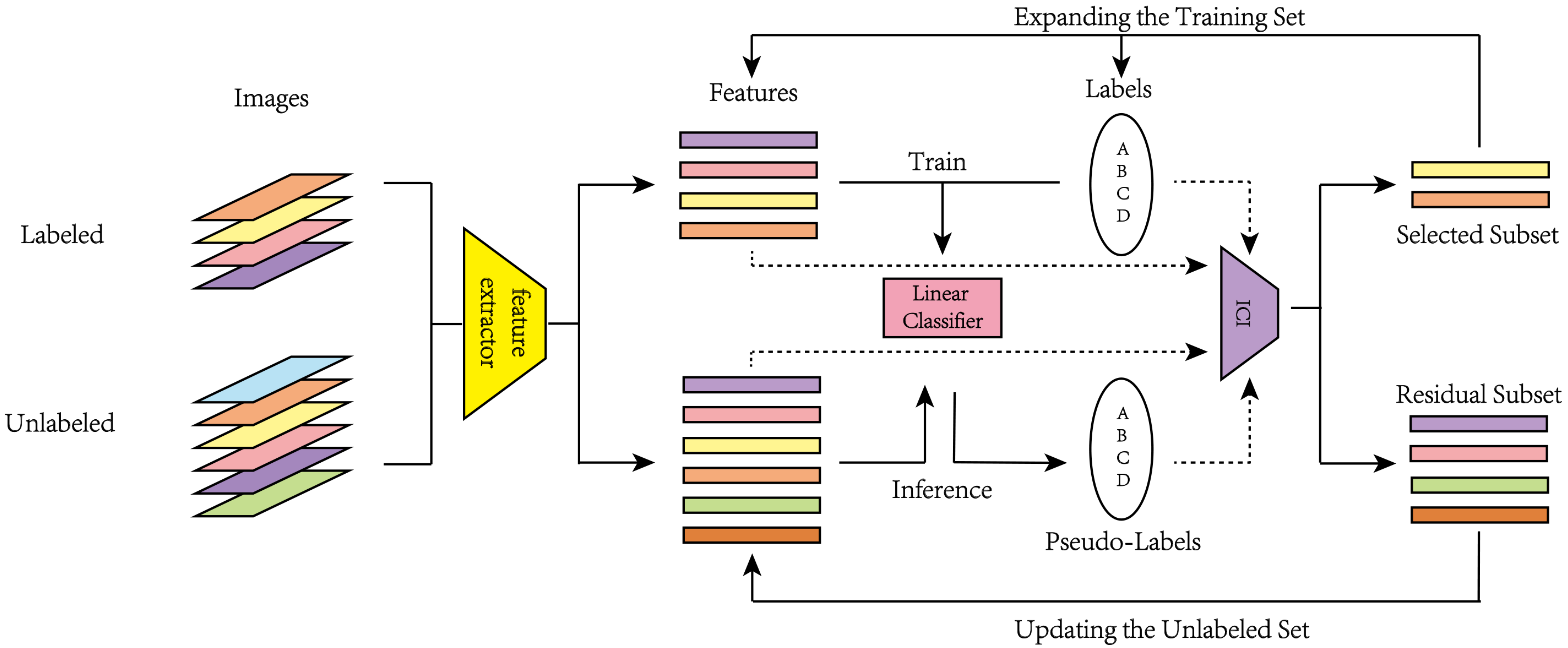
Few-shot binary classification
with unlabeled data

Motivation



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

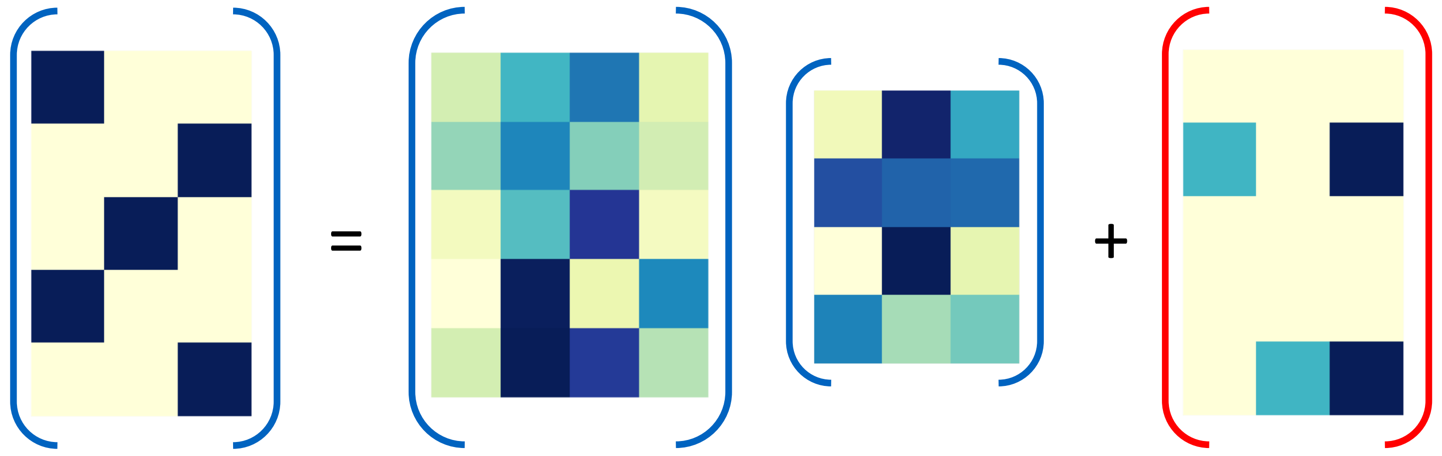
Framework



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

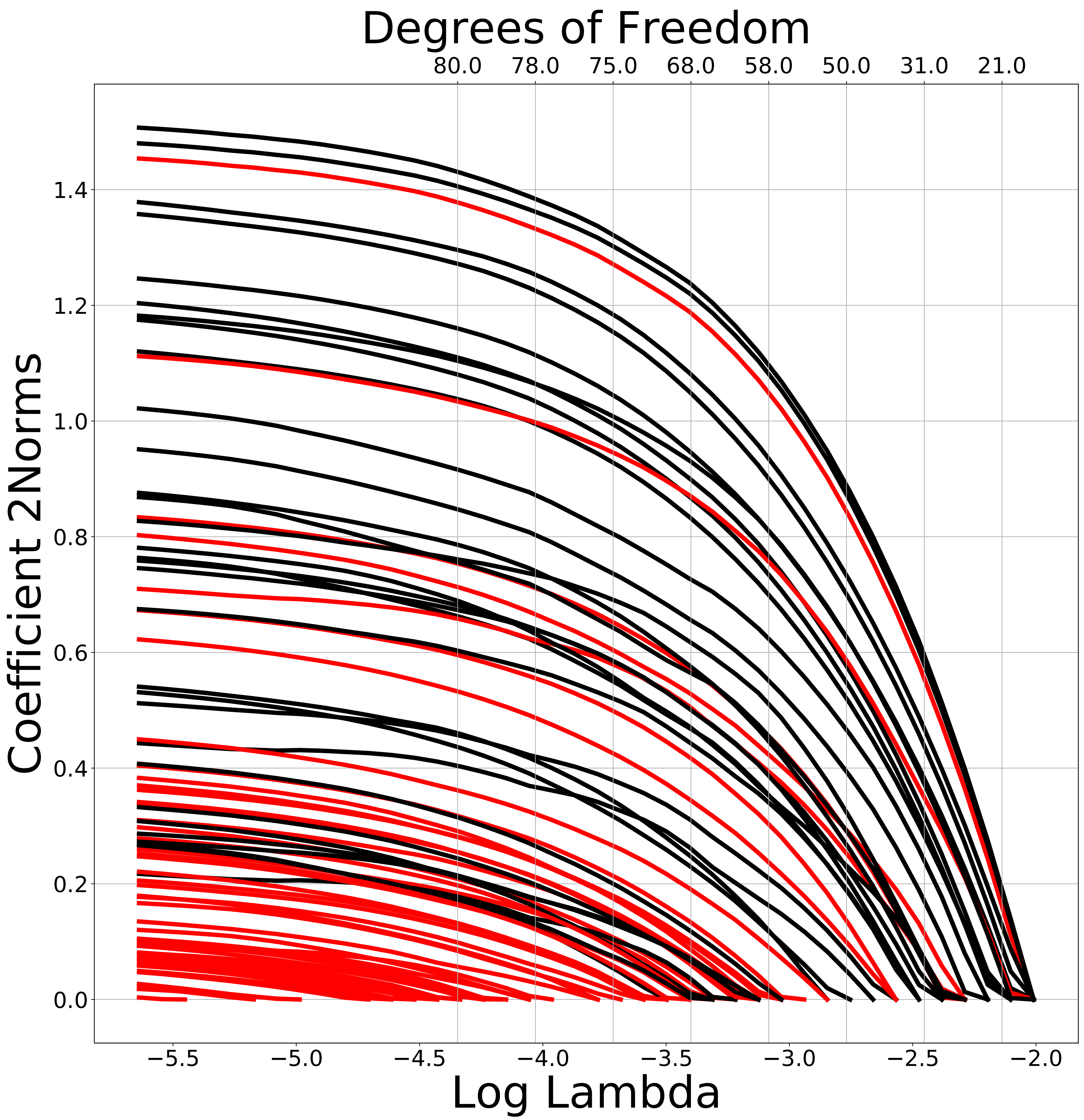
Sparse Learning in ICI

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



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

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



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

Sparse Learning: Extend to Logistic Regression

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

$$\downarrow$$

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

$$\downarrow$$

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

$$\boldsymbol{Y}_{i,c} = \frac{\exp\left(\boldsymbol{X}_{i,\cdot}\boldsymbol{\beta}_{\cdot,c} + \color{red}{\boldsymbol{\gamma}_{i,c}}\right)}{\sum_{l=1}^{C} \exp\left(\boldsymbol{X}_{i,\cdot}\boldsymbol{\beta}_{\cdot,l} + \color{red}{\boldsymbol{\gamma}_{i,l}}\right)} + \boldsymbol{\varepsilon}_{i,c}$$

$$\bar{\boldsymbol{X}} = (\boldsymbol{X}, \boldsymbol{I}) \quad \downarrow \quad \bar{\boldsymbol{\beta}} = (\boldsymbol{\beta}, \boldsymbol{\gamma})^\top$$

$$\boldsymbol{Y}_{i,c} = \frac{\exp\left(\bar{\boldsymbol{X}}_{i,\cdot}\bar{\boldsymbol{\beta}}_{\cdot,c}\right)}{\sum_{l=1}^{C} \exp\left(\bar{\boldsymbol{X}}_{i,\cdot}\bar{\boldsymbol{\beta}}_{\cdot,l}\right)} + \boldsymbol{\varepsilon}_{i,c}$$

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

Identifiability in Reality: Conditions and 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.

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

Identifiability in Reality: Conditions and 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%

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

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

Identifiability in Reality: Conditions and 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%

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

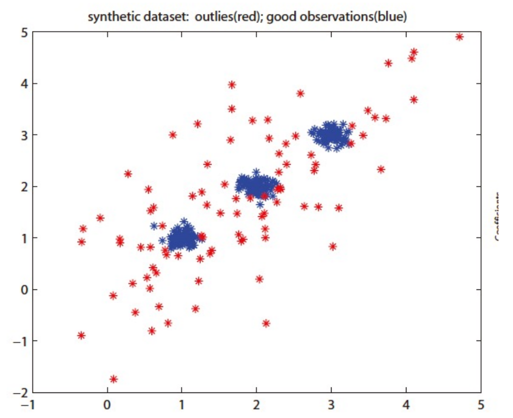
Wang et al. Instance Credibility Inference for Few-Shot Learning. CVPR 2020
Wang et al. How to Trust Unlabeled Data? Instance Credibility Inference for Few-Shot Learning. IEEE TPAMI 2021

Sparse Learning

in Learning with Noisy Labels

Definition of learning with noisy labels

Deep Models

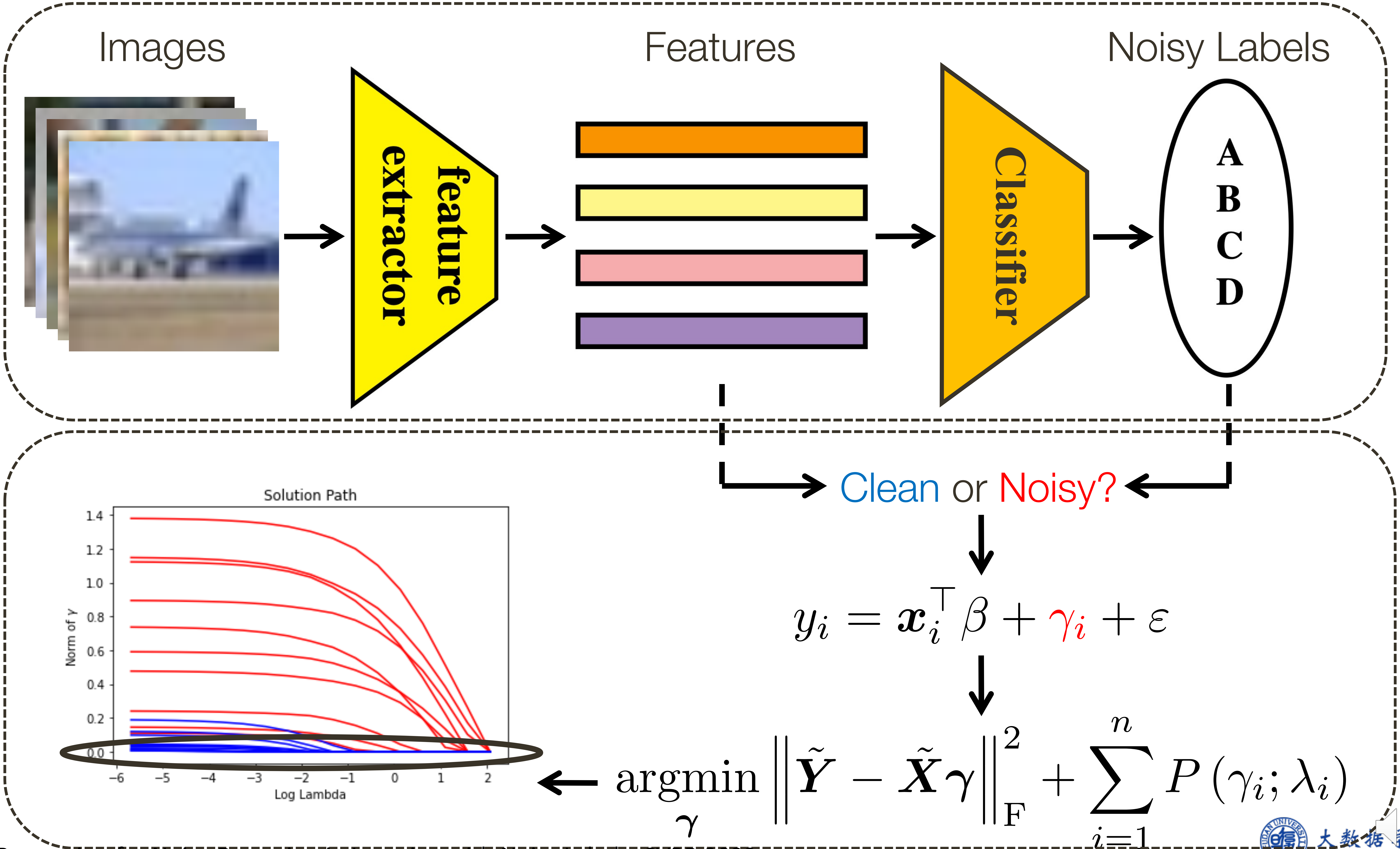


Robustly Trained Models

Framework

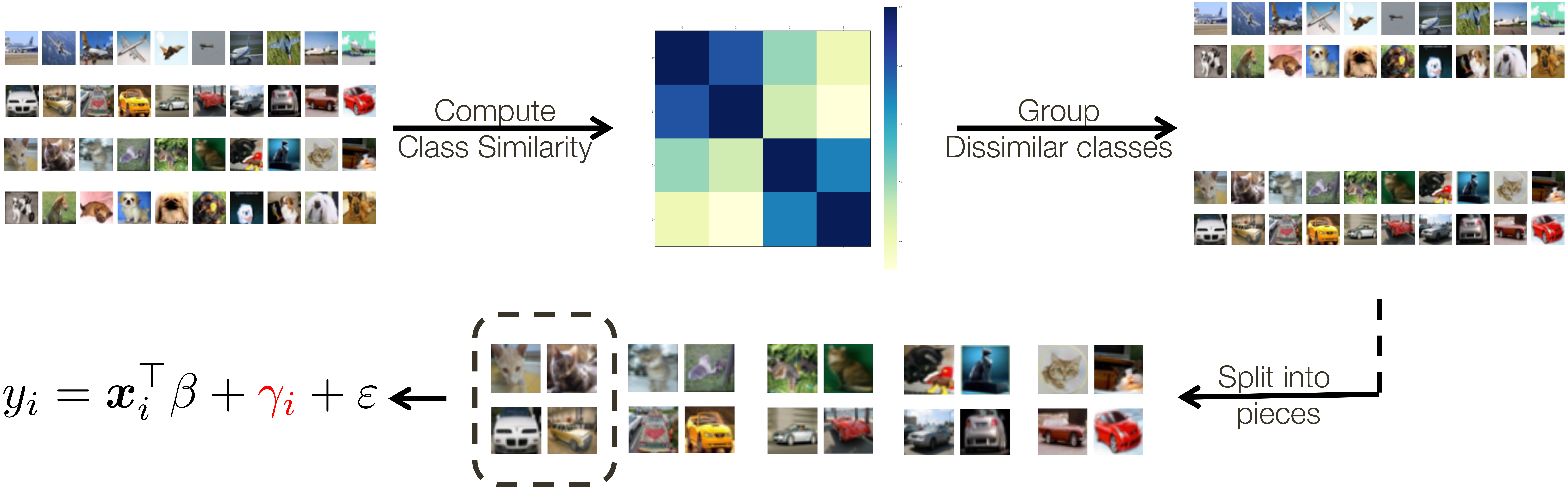
Stage 1:
Feature Learning

Stage 2:
Sample Selection



Wang et al. Scalable Penalized Regression for Noise Detection in Learning with Noisy Labels. CVPR 2022.

Make it scalable to large datasets



Wang et al. Scalable Penalized Regression for Noise Detection in Learning with Noisy Labels. CVPR 2022.

Strategies to help train the network

- Append a $\ell_q (q < 1)$ penalty to encourage the linear relation between feature and one-hot encoded vector:

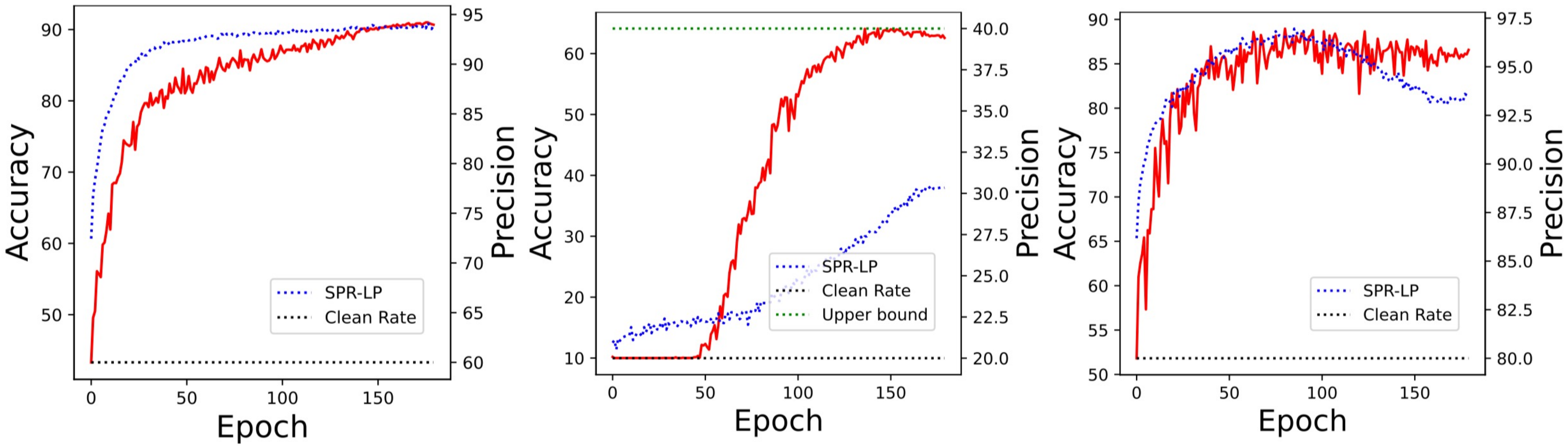
$$\mathcal{L}(\boldsymbol{x}_i, \boldsymbol{y}_i) = 1_{i \notin O} \left(\mathcal{L}_{\mathrm{CE}}(\boldsymbol{x}_i, \boldsymbol{y}_i) + \lambda \left\| \boldsymbol{x}_i^\top W_{\mathrm{fc}} \right\|_q \right)$$

- Use CutMix to further exploit the support of noisy data

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

Wang et al. Scalable Penalized Regression for Noise Detection in Learning with Noisy Labels. CVPR 2022.

Label precision performance



(a) Symmetric-40% (b) Symmetric-80% (c) Asymmetric-40%

Wang et al. Scalable Penalized Regression for Noise Detection in Learning with Noisy Labels. CVPR 2022.

THANKS