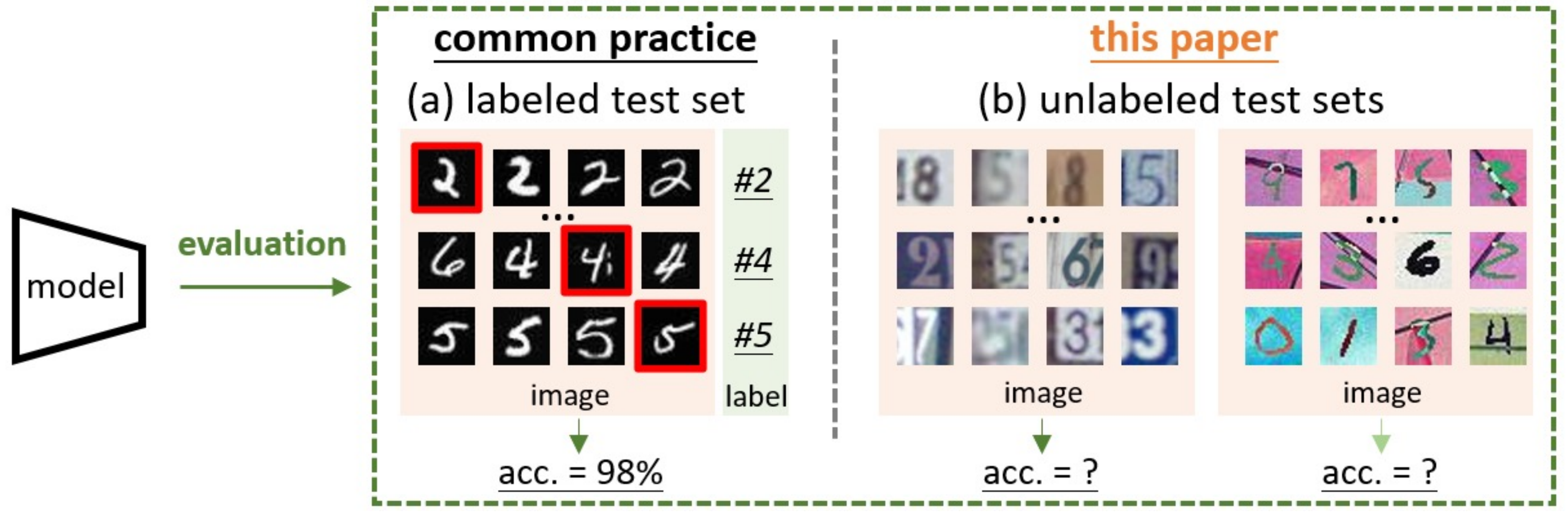
Are Labels Always Necessary for Classifier Accuracy Evaluation?

Weijian Deng and Liang Zheng
Australian National University

Automatic Model Evaluation

Given a trained classifier, the overall goal is to estimate its accuracy on various test datasets **without labels**



Motivation

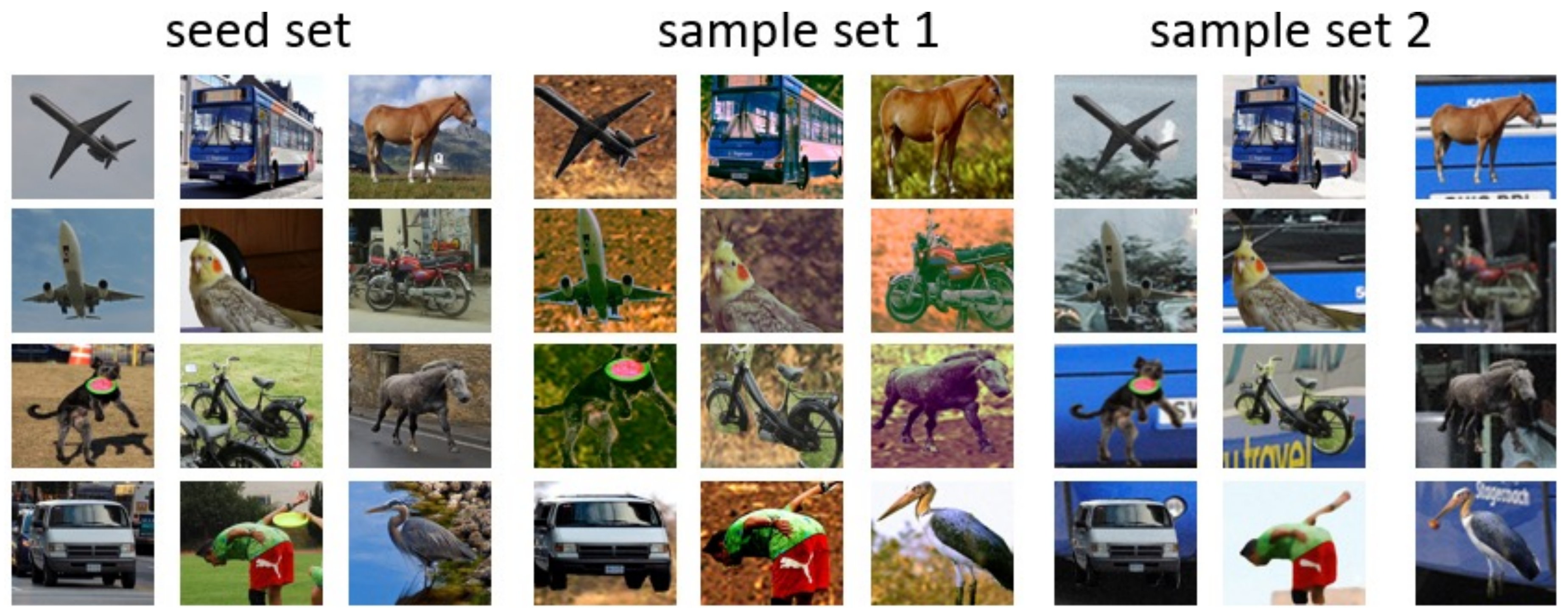
- Distribution shift degrades classifier accuracy

Reduce the shift to make model generalizable to the target domain

- Meta-dataset (dataset of datasets)

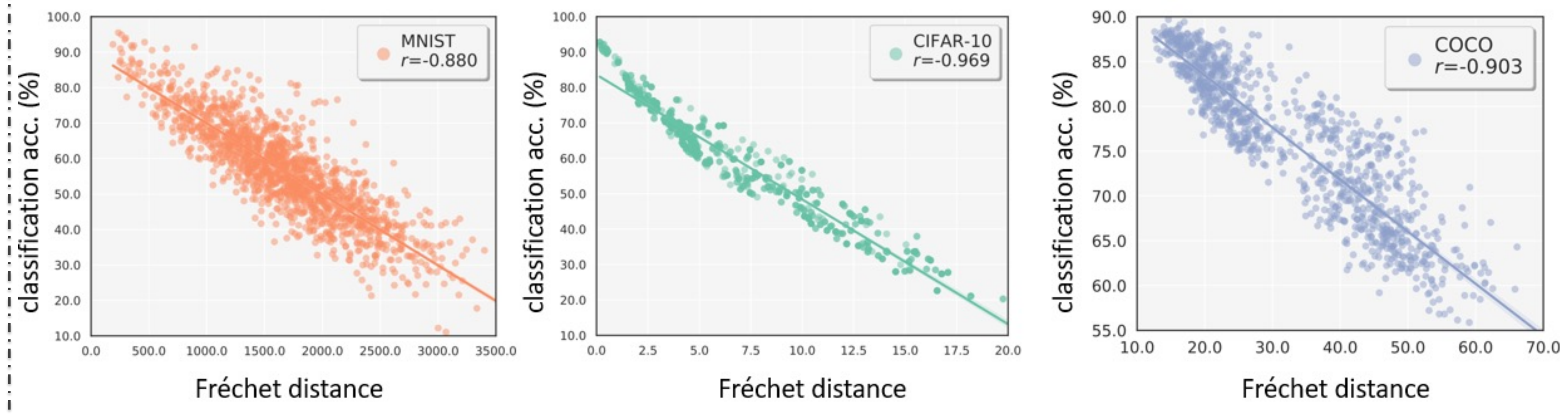
It contains many datasets from different distributions

Data synthesis: 1) image transform and 2) background change



Correlation Study

- Negative Linear Correlation between Test Accuracy and Distribution Shift



Classifier tends to have a ***low accuracy*** on the sample set which has a ***high distribution shift*** from the training set

Dataset-level Regression

Predict classifier accuracy from ***distribution-related statistics*** on an unlabeled test set

- Linear regression

Model: $a_{linear} = A_{linear}(\boldsymbol{f}) = w_1 f_{linear} + w_0$

Feature: $f_{linear} = \mathrm{FD}(\mathcal{D}_{ori}, \mathcal{D}) = \|\boldsymbol{\mu}_{ori} - \boldsymbol{\mu}\|_2^2 + Tr(\boldsymbol{\Sigma}_{ori} + \boldsymbol{\Sigma} - 2(\boldsymbol{\Sigma}_{ori}\boldsymbol{\Sigma})^{\frac{1}{2}}$

Fréchet distance (FD) *measures the difference between training and test distributions*

- Neural network regression

Model: $a_{neural} = A_{neural}(\boldsymbol{f}_{neural})$

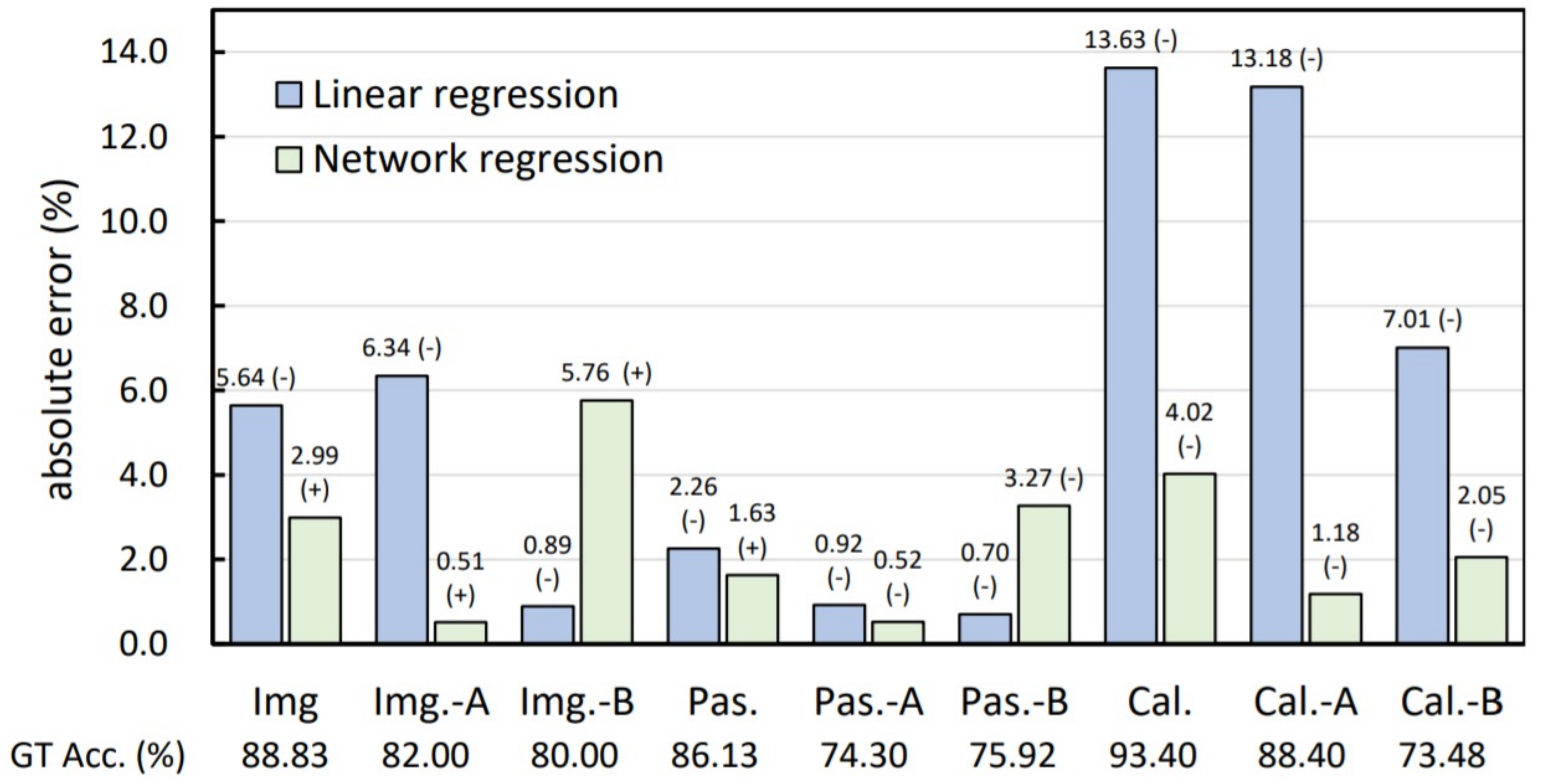
Feature: $\boldsymbol{f}_{neural} = [f_{linear}; \boldsymbol{\mu}; \boldsymbol{\sigma}]$

*Network regression uses **mean**, **co-variance**, and **FD** to represent each test set*

Experiment

Method	Digits: SVHN	Digits: USPS	Digits: RMSE↓	Natural images: Pascal	Natural images: Caltech	Natural images: ImageNet	Natural images: RMSE↓
Ground-truth accuracy	25.46	64.08	-	86.13	93.40	88.83	-
Predicted score ($\tau = 0.8$)	7.97	37.22	22.66	84.32	90.78	86.50	**2.28**
Predicted score ($\tau = 0.9$)	7.03	32.94	25.59	78.61	87.71	81.33	6.96
Linear reg.	26.28	50.14	9.87	83.87	79.77	83.19	8.62
Neural network reg.	27.52	64.11	**1.46**	87.76	89.39	91.82	**3.04**

Predicted score-based baseline is sensitive to the threshold;
Our regression methods gain promising estimations



Comparing linear regression and neural network regression when test data undergo new image transformations: A) Cutout and Shear; B) Equalize and ColorTemperature

Network regression is more robust than linear regression

Reference

A. Torralba and A. A. Efros. Unbiased look at dataset bias. In CVPR, 2011
Ben-David, Shai, et al. "A theory of learning from different domains." Machine learning, 2010

The code is available at
https://weijiandeng.xyz/AutoEval