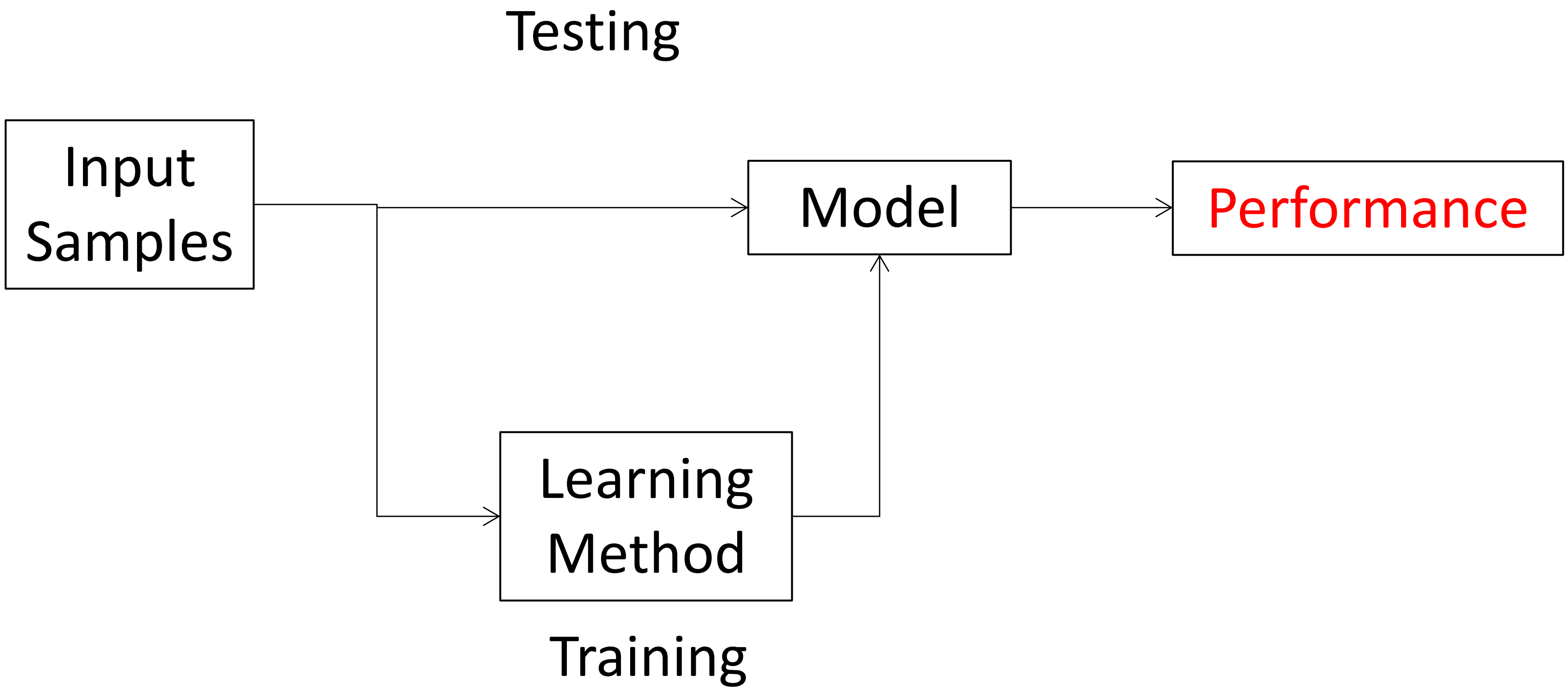
Do we really need ground truths to evaluate a model?

Liang Zheng

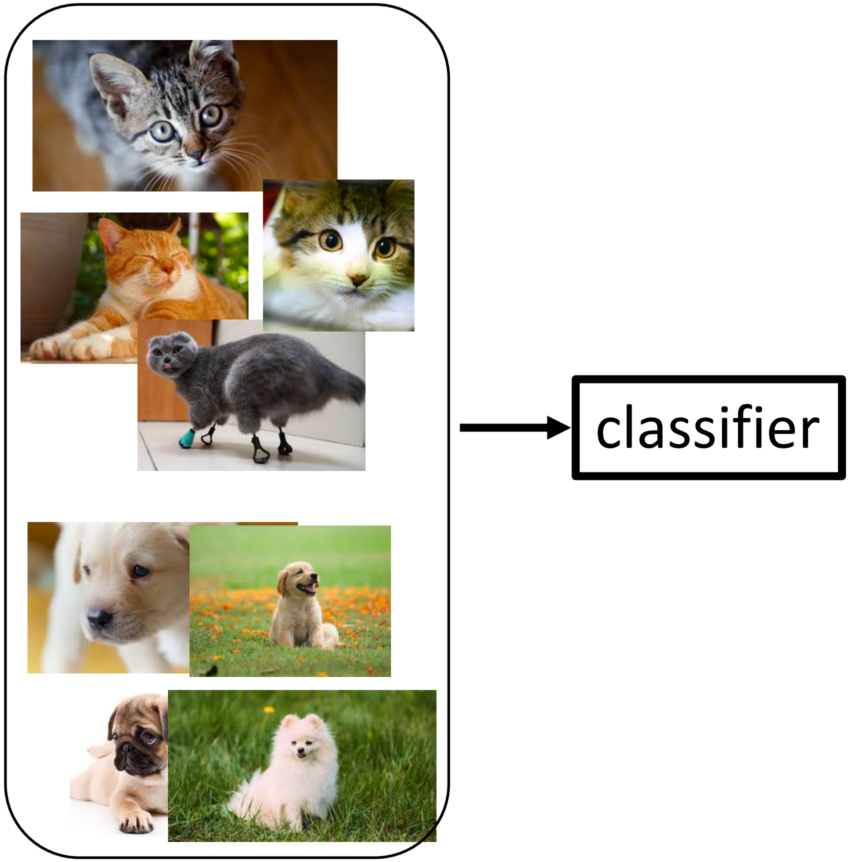
Australian National University

11-Nov-2020

Pillars in machine learning



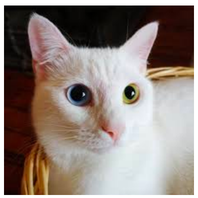
We start with training a classifier



Training data

We do a bit testing....

Correct prediction



classifier

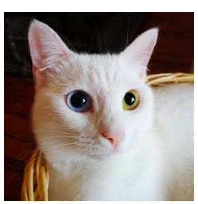
Dog 0.05
Cat 0.95

Cat

Testing image

Prediction result = Ground truth

Wrong prediction



classifier

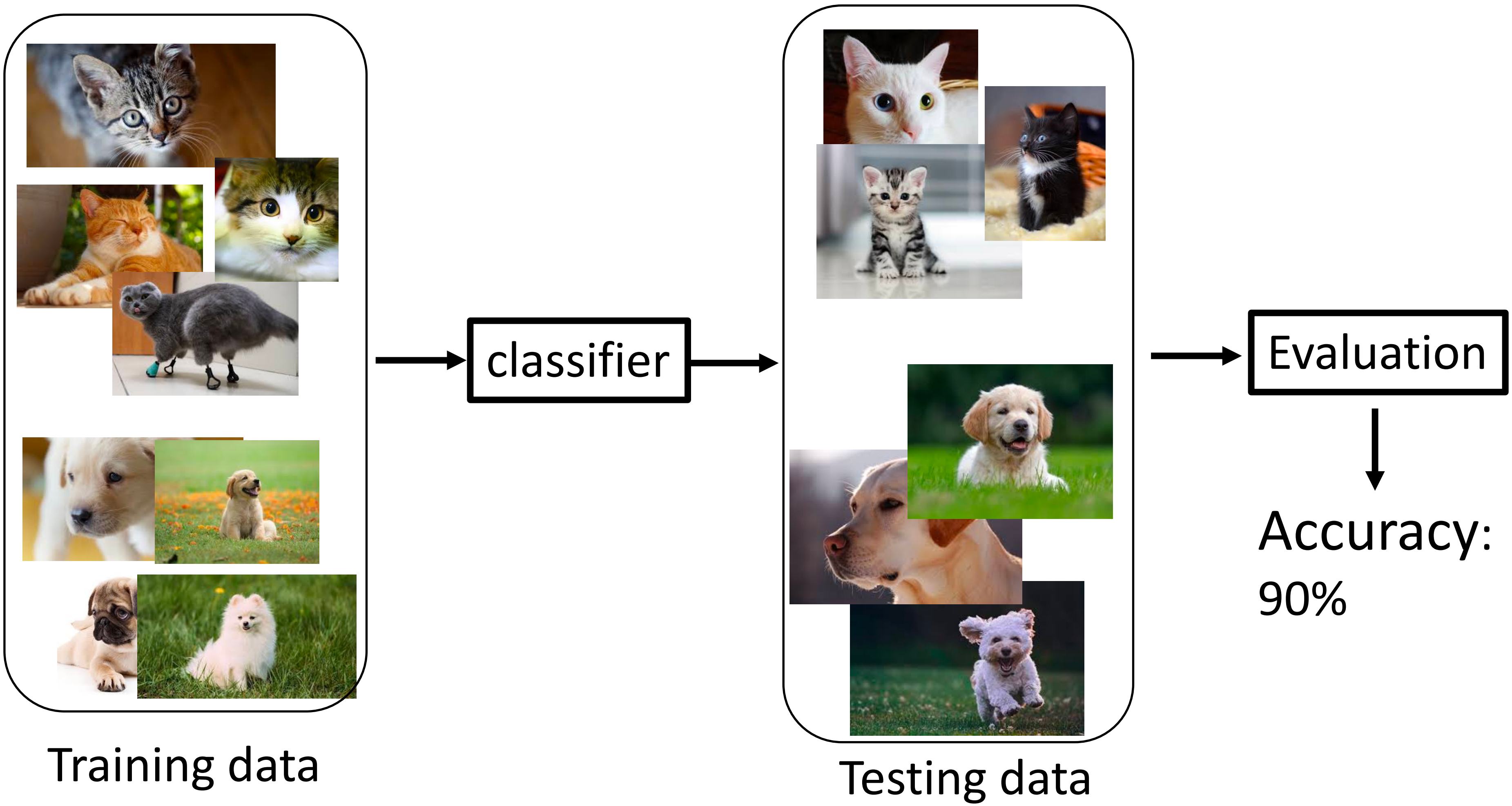
Dog 0.85
Cat 0.15

Cat

Testing image

Prediction result ≠ Ground truth

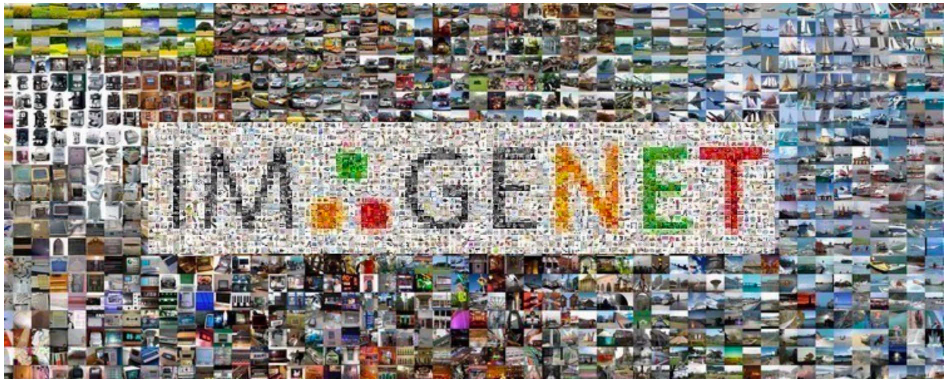
We now evaluate a model



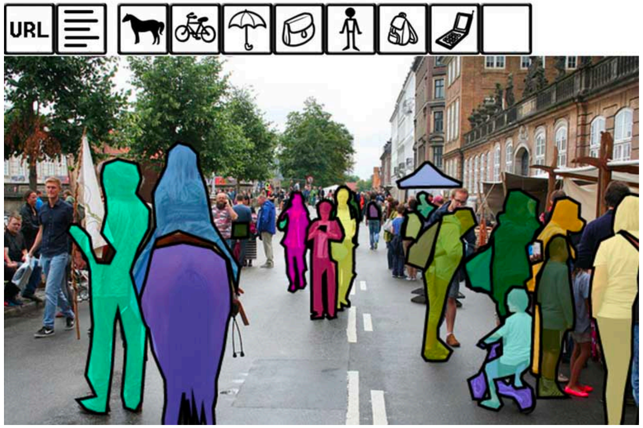
Ground truths provided

Is this way of evaluation feasible?

- Yes



ImageNet



MSCOCO

Ground truths provided

LFW

Is this way of evaluation feasible?

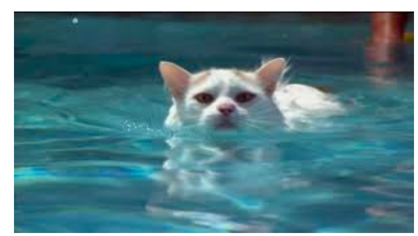
- No....

We can't calculate a classifier accuracy!!

Suppose we deploy our cat-dog classifier to a swimming pool





Ground truths not provided

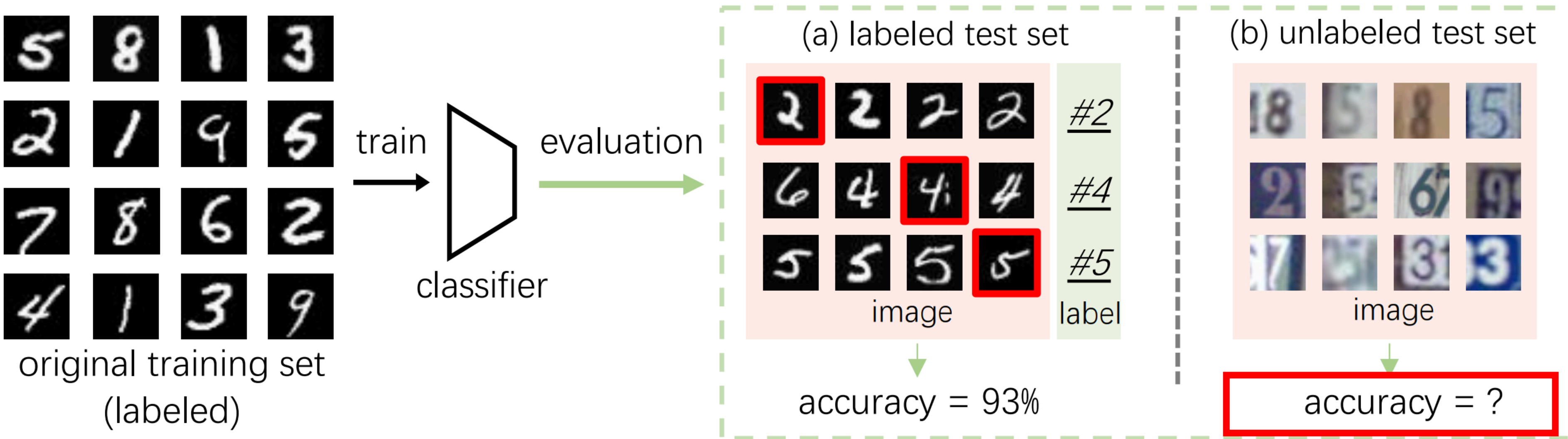
We encounter this problem too many times in CV applications….

- Deploy a ReID model to a new community
- Deploy face recognition in an airport
- Deploy a 3D object detection system to a new city
- ……

We can't quantitatively measure the performance of our model like we usually do!!

Unless we annotate the test data…, but environment will change over time…. We need to annotate test data again

Formally, we want to solve:

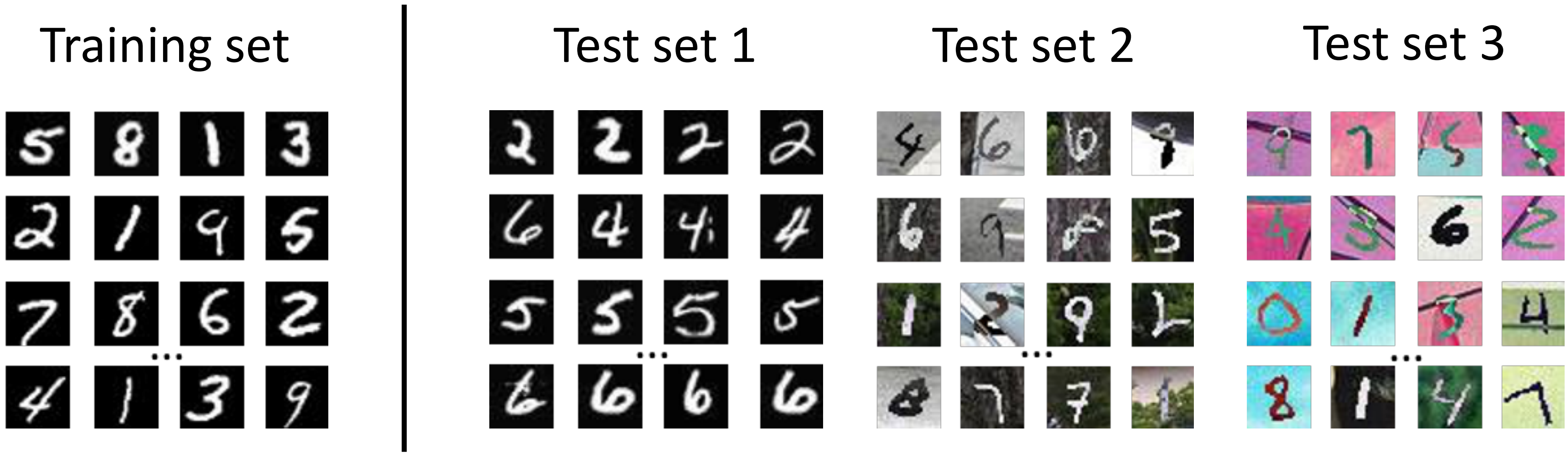


Given
- A training dataset
- A classifier trained on this dataset
- A test set without labels

We want to estimate:
Classification accuracy on the test set

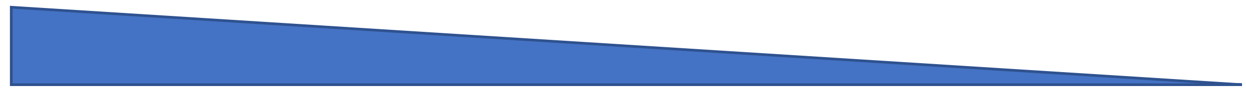
Our idea

Training set	Test set 1	Test set 2	Test set 3

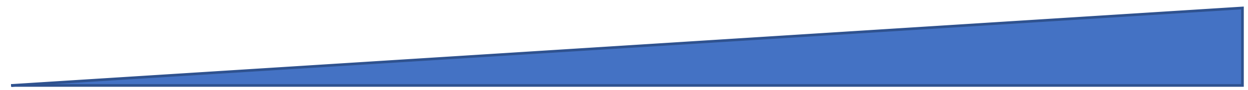


	97%	80%	60%

recognition accuracy:



Domain gap:



Negative correlation between recognition accuracy and domain gap

Our idea

Known (from existing literature)

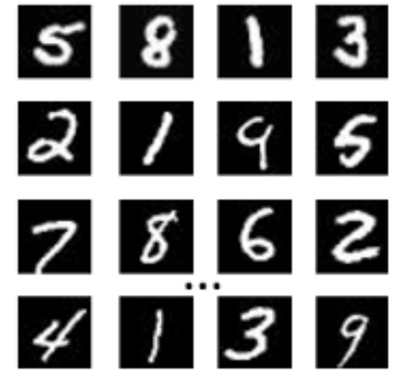
Larger domain gap -> lower recognition accuracy

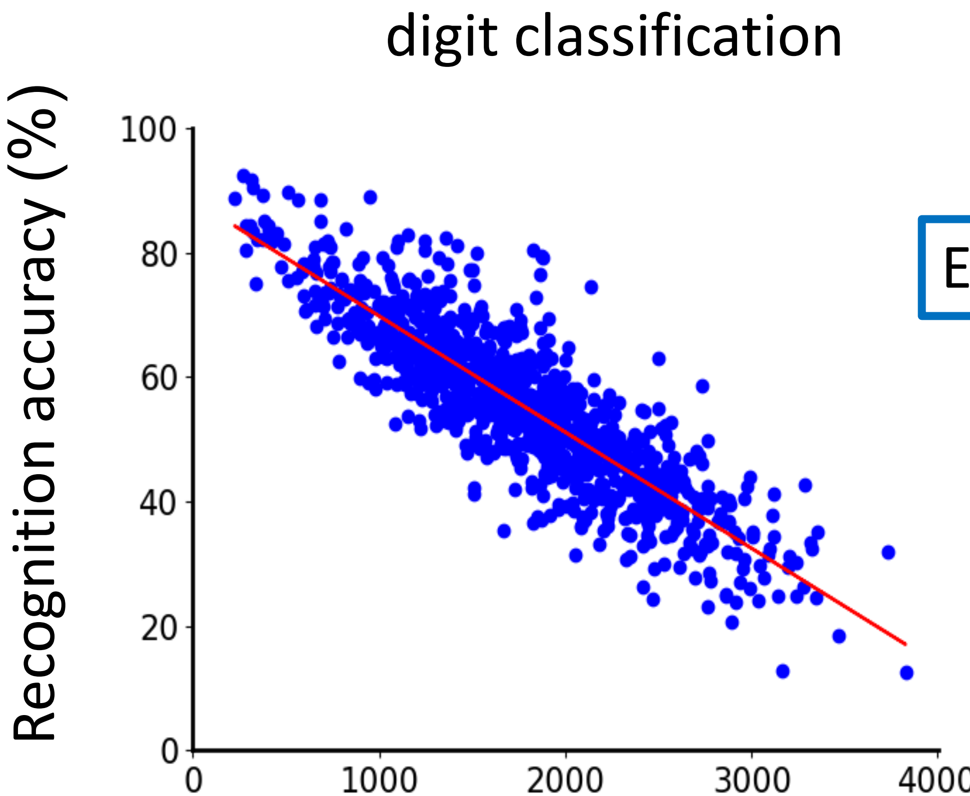
Unknown

Can we quantify this relationship?

A regression problem!

Some experiments





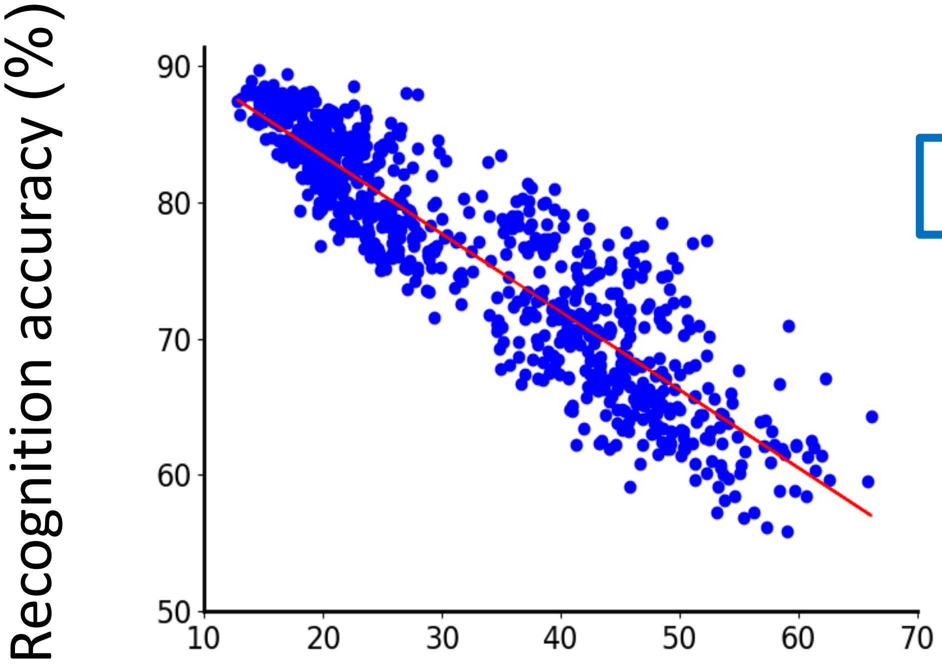
Every point is a dataset

Fréchet distance

Domain gap between a training set and test sets

Some experiments

natural image classification

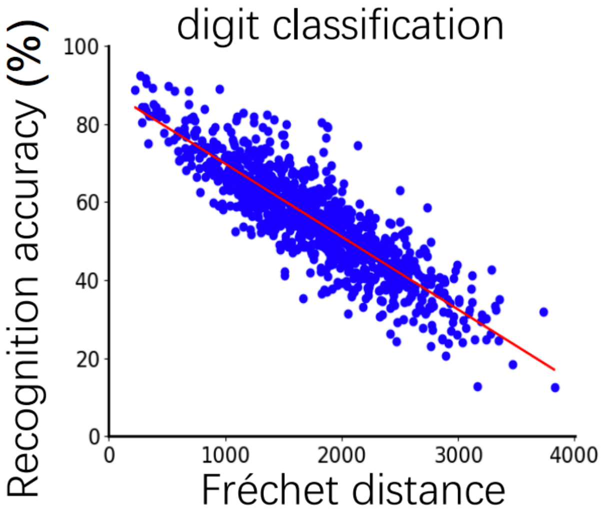


Every point is a dataset

Fréchet distance

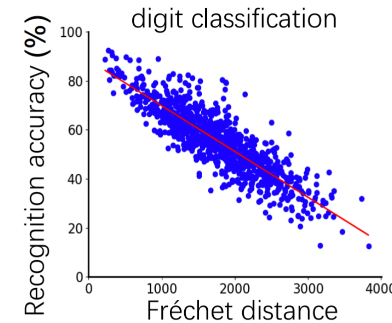
Domain gap between a training set and test sets

Method key points

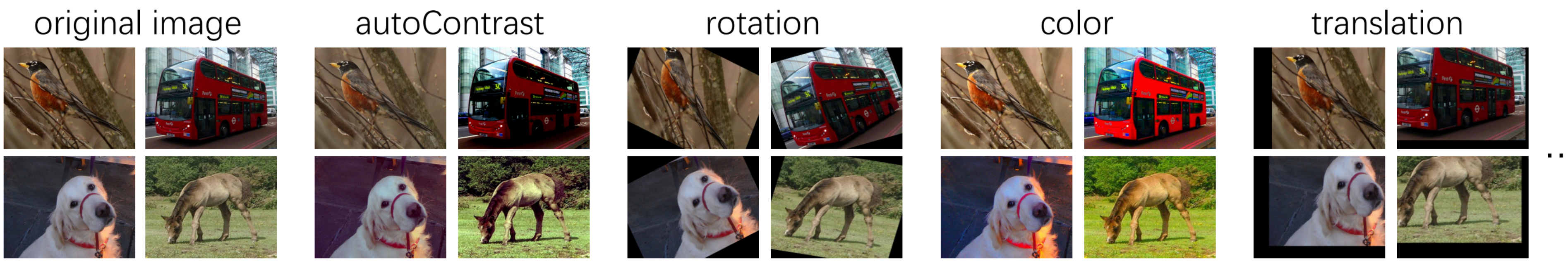


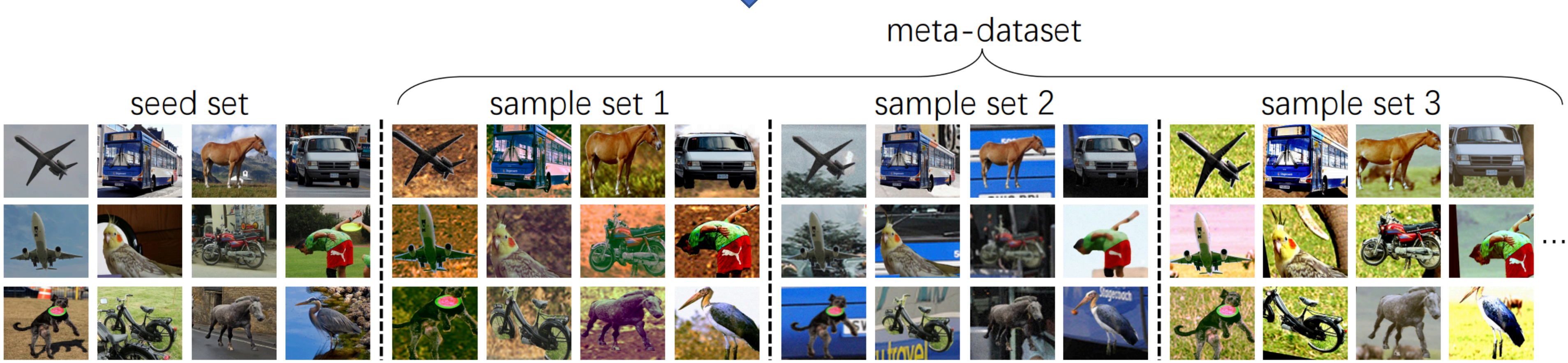
- How can we have MANY datasets?
- How to obtain the recognition accuracy for each dataset?
- Dataset representation
  - Fréchet distance?
  - Other representations?
- We use regression to relate dataset representation with recognition accuracy.

How can we have MANY datasets?

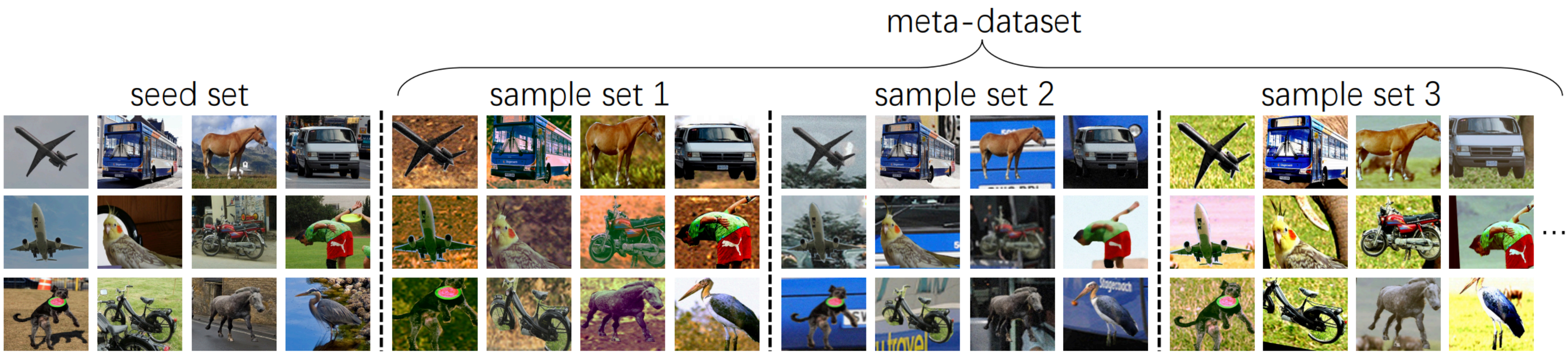


- Using image transformations





How to obtain the recognition accuracy for each dataset?



Labels of the sample sets are inherited from the seed set.

Given a classifier, the recognition accuracy on these sample sets can be easily calculated.

Dataset representation

- Method 1: Fréchet distance (FD) between a sample set and the original training set

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

- FD: distribution difference between two domains
- Including mean and covariance
- Dimension of f_{linear}: 1

- We thus can use linear regression to predict accuracy

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

Dataset representation

- Method 2: FD+mean+sum(covariance)

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

- We calculate $\boldsymbol{\sigma}$ by taking a weighted summation of each row of $\boldsymbol{\Sigma}$ to produce a single vector
- Dimension of f_{linear}: $2d + 1$
- d is the dimension of an image feature

- We use neural network regression

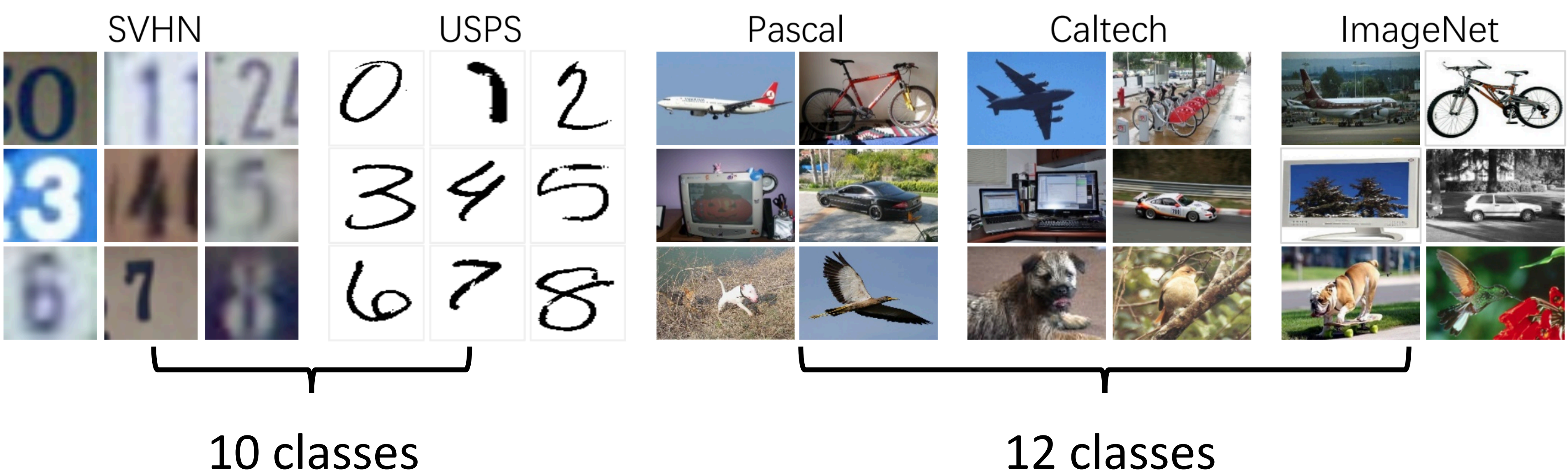
$$a_{neural} = A_{neural}(\boldsymbol{f}_{neural})$$

Experiment

Training set	**Seed set**
MNIST training set	MNIST test set
COCO training set	COCO validation set

Experiment

- We predict the classifier accuracy on five real-world datasets



- We use mean squared error (MSE) to evaluate the accuracy of recognition accuracy prediction.

Experiment

Method	Digits			Natural images			
	SVHN	USPS	MSE↓	Pascal	Caltech	ImageNet	MSE↓
Ground-truth accuracy	25.46	64.08	0	86.13	93.40	88.83	0

Experiment

	Digits			Natural images			
Method	SVHN	USPS	MSE↓	Pascal	Caltech	ImageNet	MSE↓
Ground-truth accuracy	25.46	64.08	0	86.13	93.40	88.83	0
Confidence ($\tau = 0.8$)	7.97	5.88	16.03	84.32	90.78	86.50	**1.32**
Confidence ($\tau = 0.9$)	37.22	27.95	20.55	78.61	87.71	87.71	4.02

"Confidence": a simple pseudo label method.

If the maximum value of the softmax vector is greater than τ, we view this sample as correctly classified.

Experiment

Method	Digits			Natural images			
	SVHN	USPS	MSE↓	Pascal	Caltech	ImageNet	MSE↓
Ground-truth accuracy	25.46	64.08	0	86.13	93.40	88.83	0
Confidence ($\tau = 0.8$)	7.97	5.88	16.03	84.32	90.78	86.50	**1.32**
Confidence ($\tau = 0.9$)	37.22	27.95	20.55	78.61	87.71	87.71	4.02
Linear reg.	26.28	50.14	6.98	83.87	79.11	83.19	4.98

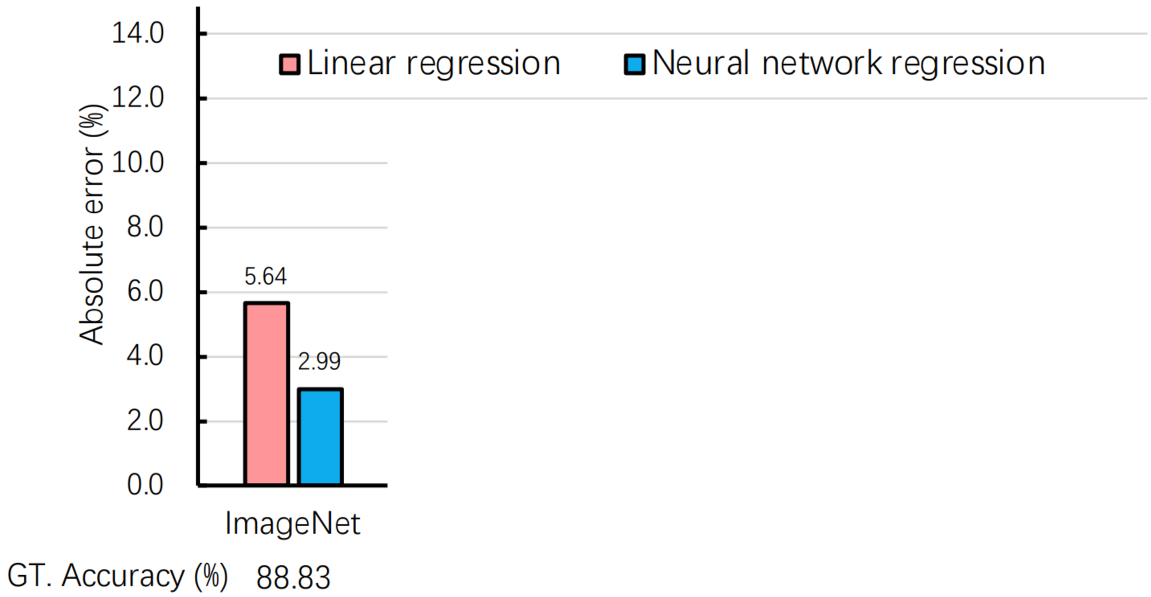
Experiment

Method	Digits			Natural images			
	SVHN	USPS	MSE↓	Pascal	Caltech	ImageNet	MSE↓
Ground-truth accuracy	25.46	64.08	0	86.13	93.40	88.83	0
Confidence ($\tau = 0.8$)	7.97	5.88	16.03	84.32	90.78	86.50	**1.32**
Confidence ($\tau = 0.9$)	37.22	27.95	20.55	78.61	87.71	87.71	4.02
Linear reg.	26.28	50.14	6.98	83.87	79.11	83.19	4.98
Neural network reg.	27.52	64.11	**1.03**	87.76	89.39	91.82	**1.75**

The two regression methods are stable and quite accurate.

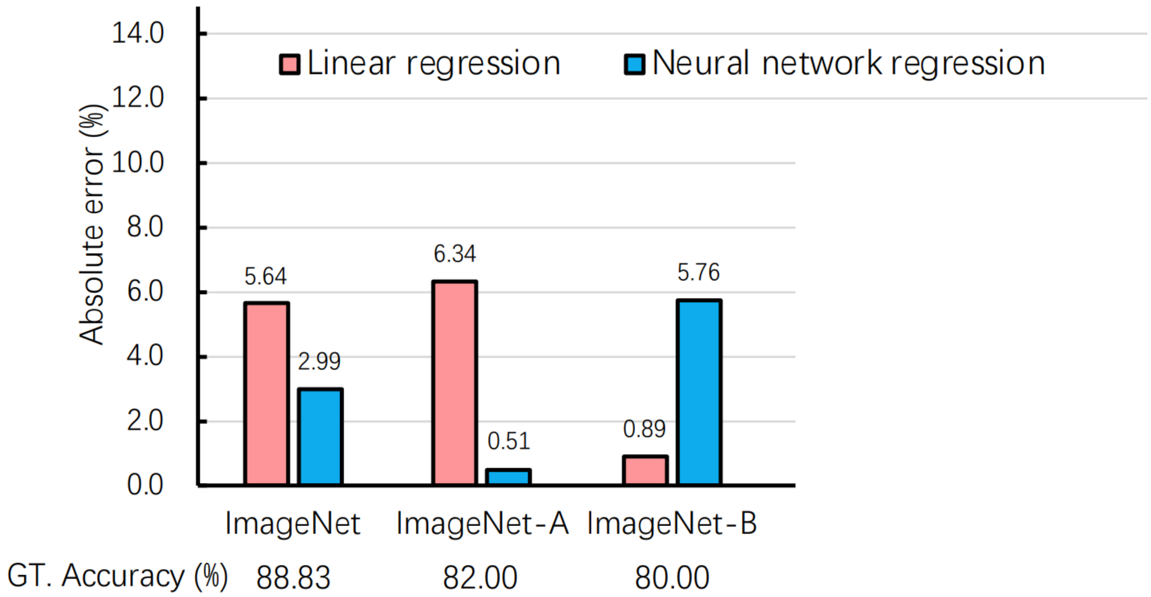
Test sets undergo new transformations

- We add new image transformations to the test sets.
- Random erasing / cutout, Shear, Equalize and ColorTemperature



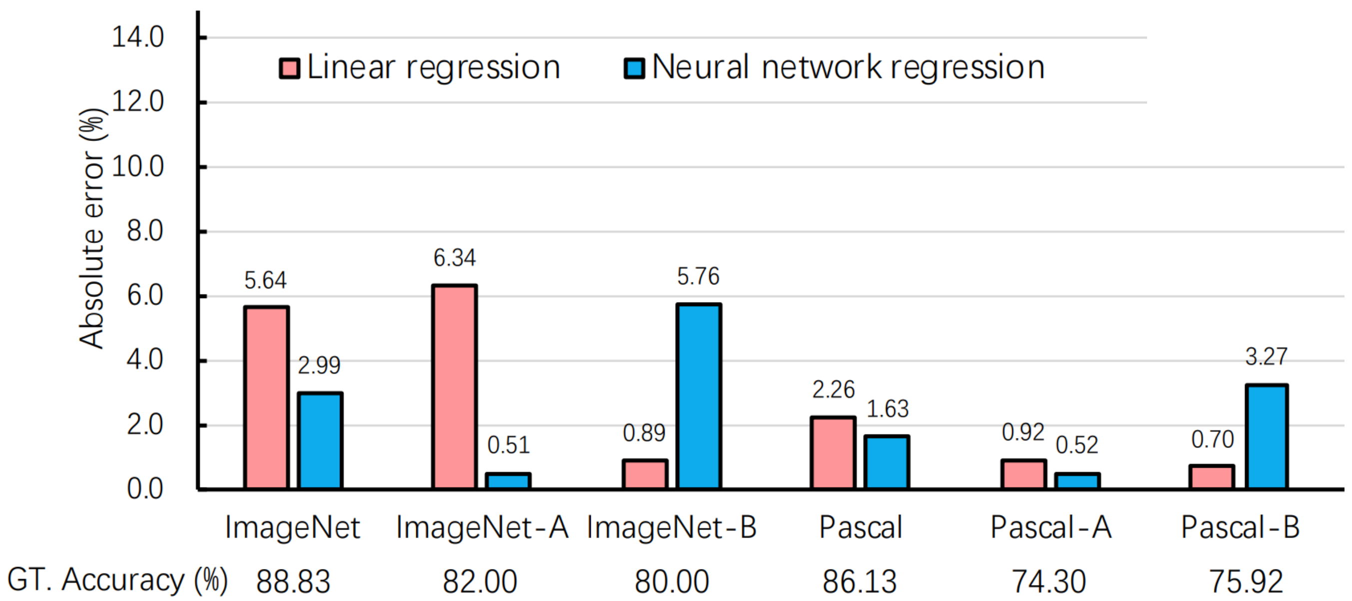
Test sets undergo new transformations

- We add new image transformations to the test sets.
- Random erasing / cutout, Shear, Equalize and ColorTemperature



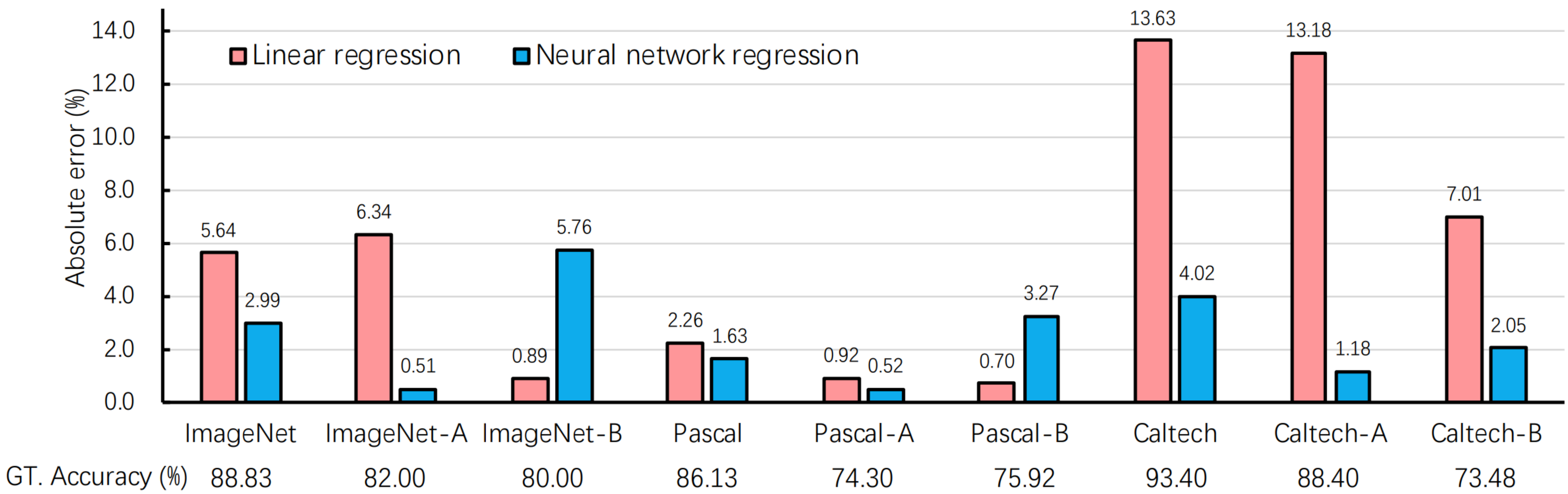
Test sets undergo new transformations

- We add new image transformations to the test sets.
- Random erasing / cutout, Shear, Equalize and ColorTemperature

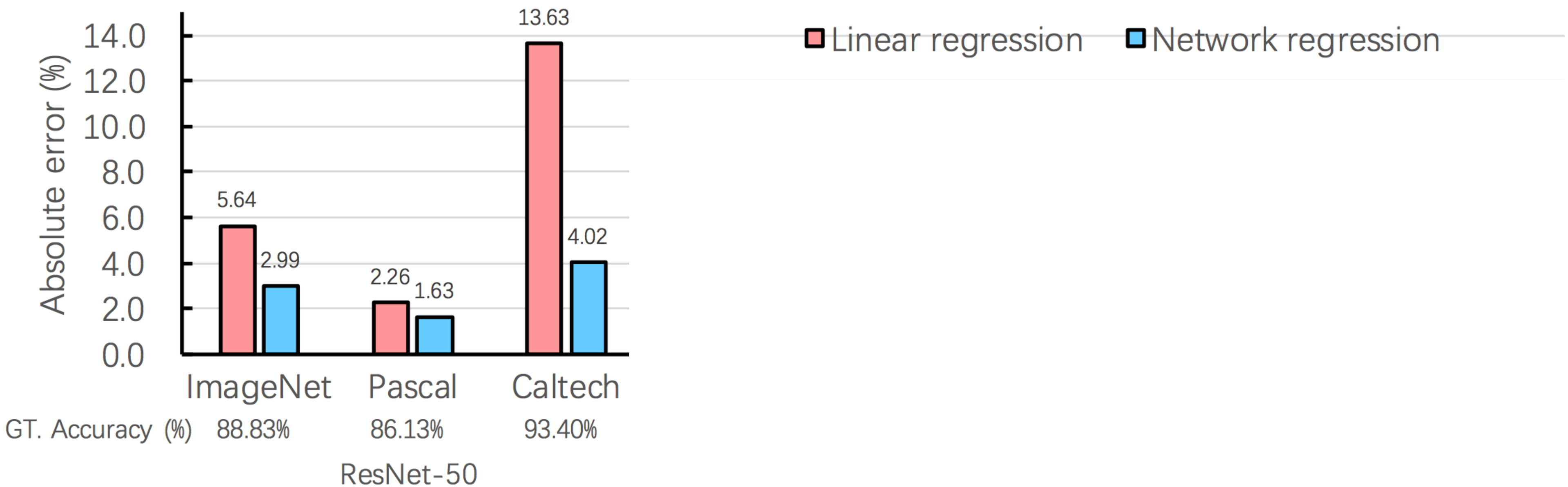


Test sets undergo new transformations

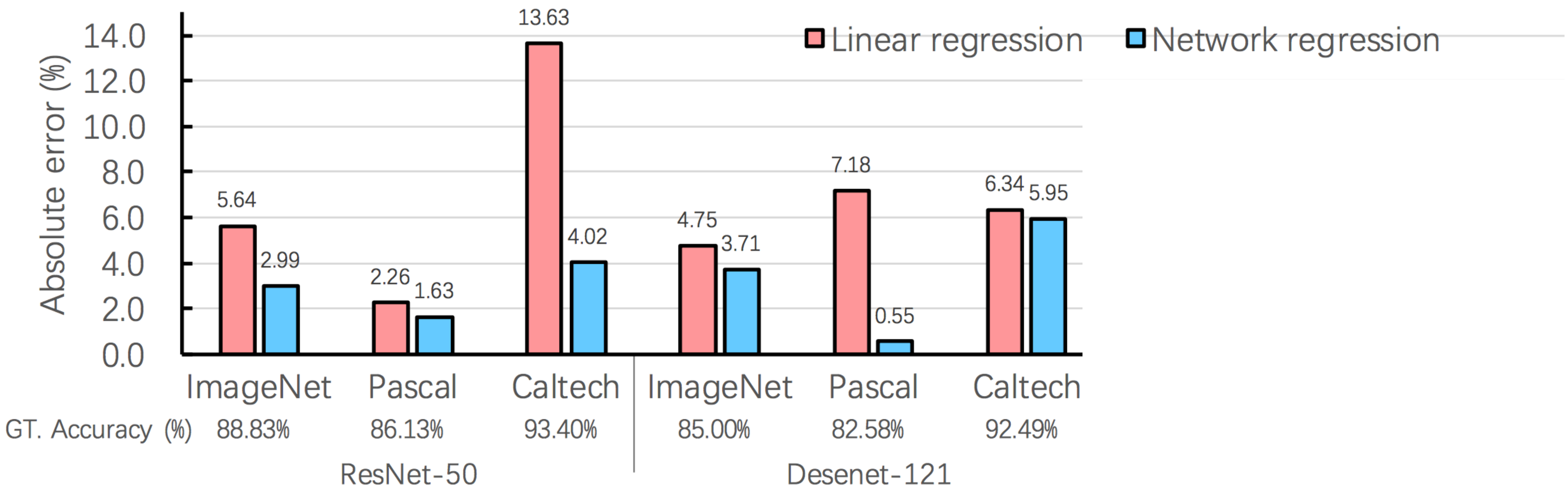
- We add new image transformations to the test sets.
- Random erasing / cutout, Shear, Equalize and ColorTemperature



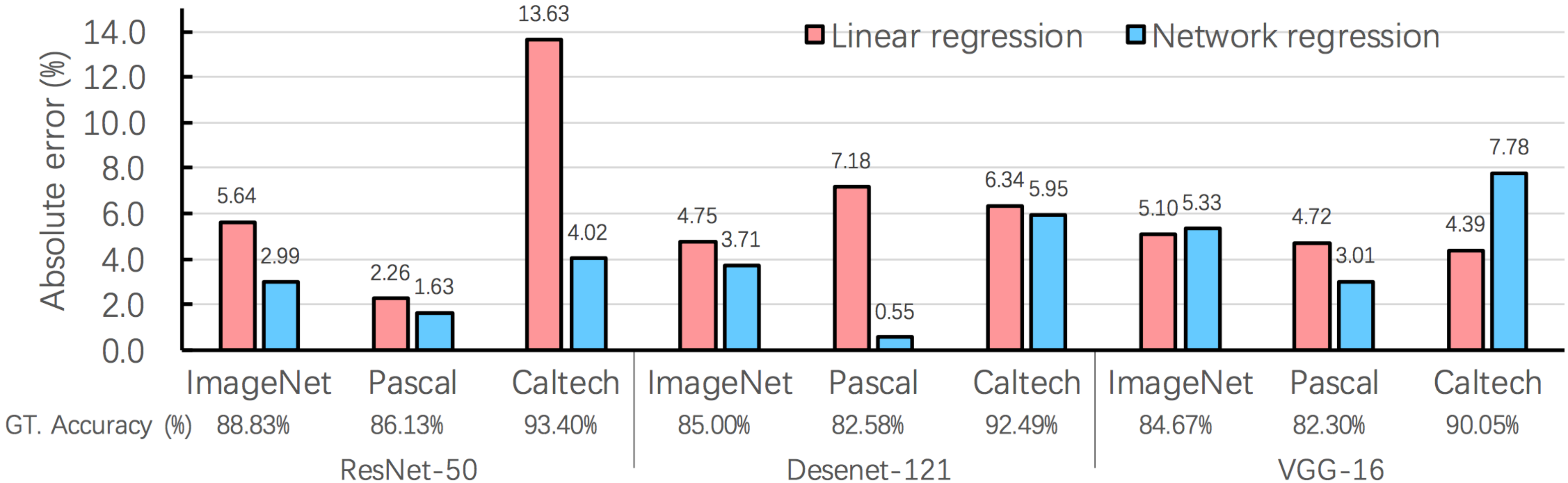
Predicting the accuracy of various classifiers



Predicting the accuracy of various classifiers

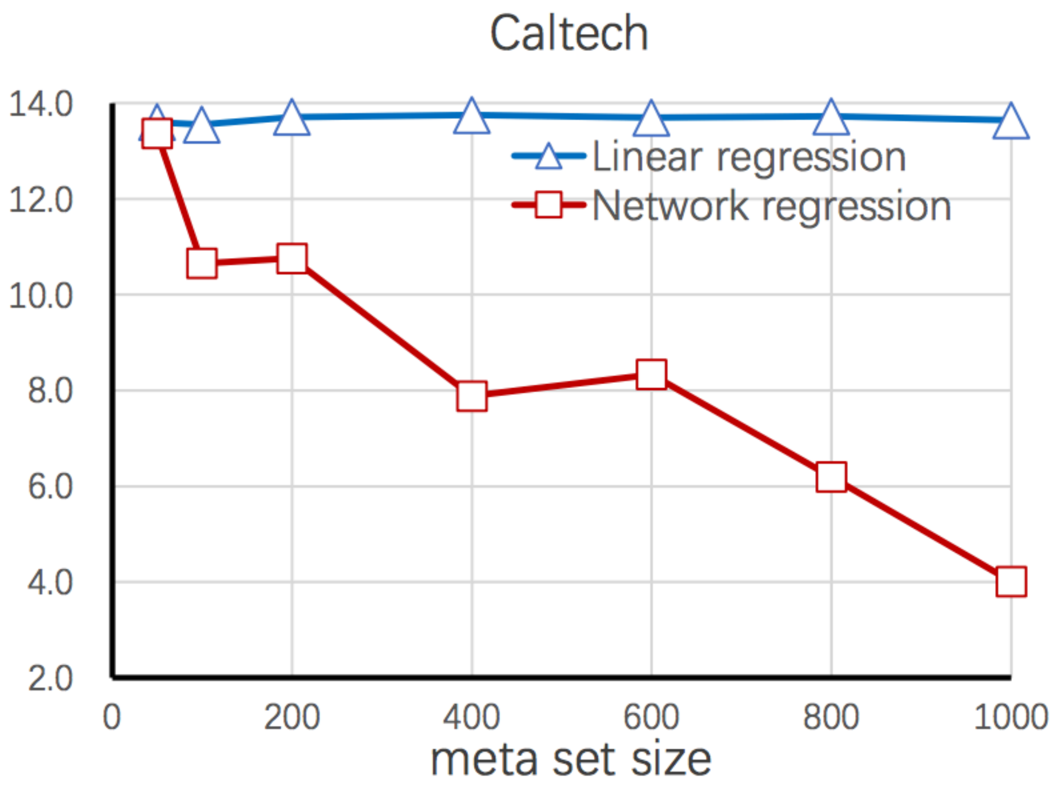


Predicting the accuracy of various classifiers



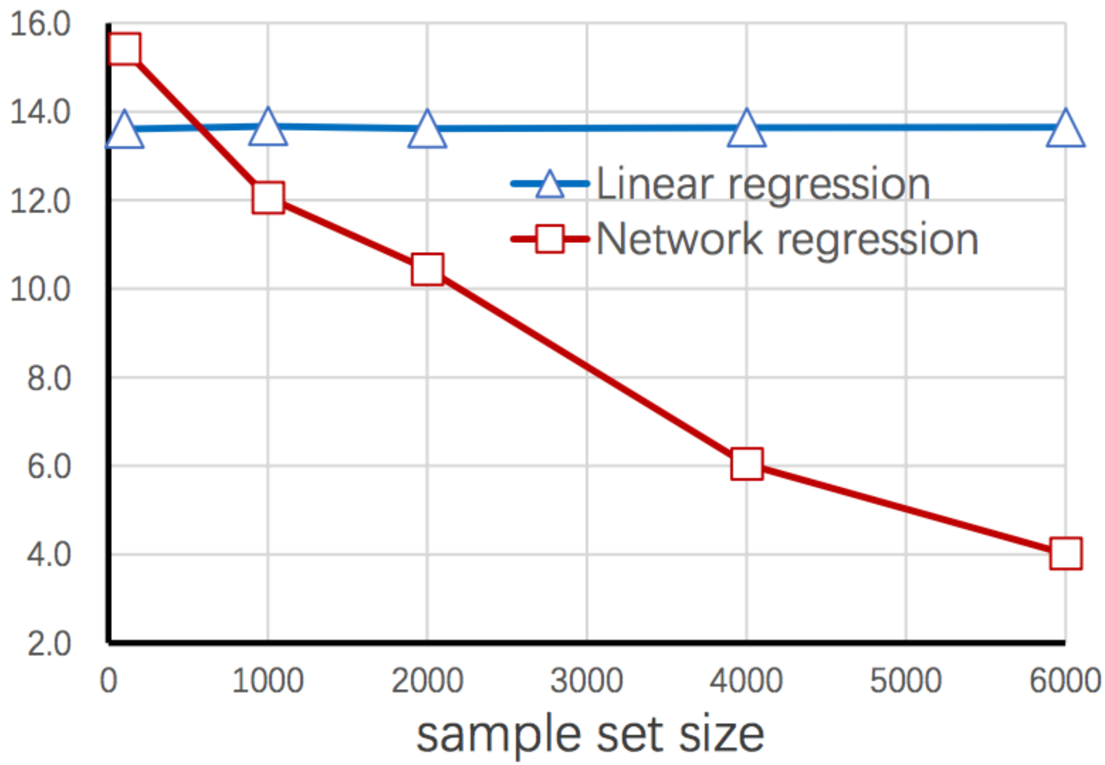
Some important parameters

- The number of synthetic datasets (sample sets)



Some important parameters

- The size of each synthetic dataset (sample set)



Conclusions and insights

- We study a very interesting problem:
- Evaluating model performance without ground truths

- We use a very simple method:
- Regression

- Potential Applications:
- Object recognition, detection, segmentation, re-ID, etc.

Conclusions and insights

- Application scope
  - The space spanned by the sample sets should cover the test sets.
  - If not, there will be failure cases
- Dataset representation
  - A less studied problem
  - We use first- and second-order feature statistics and FD
  - Better representations?
- Dataset similarity
  - We use FD score
  - Better similarity estimation?