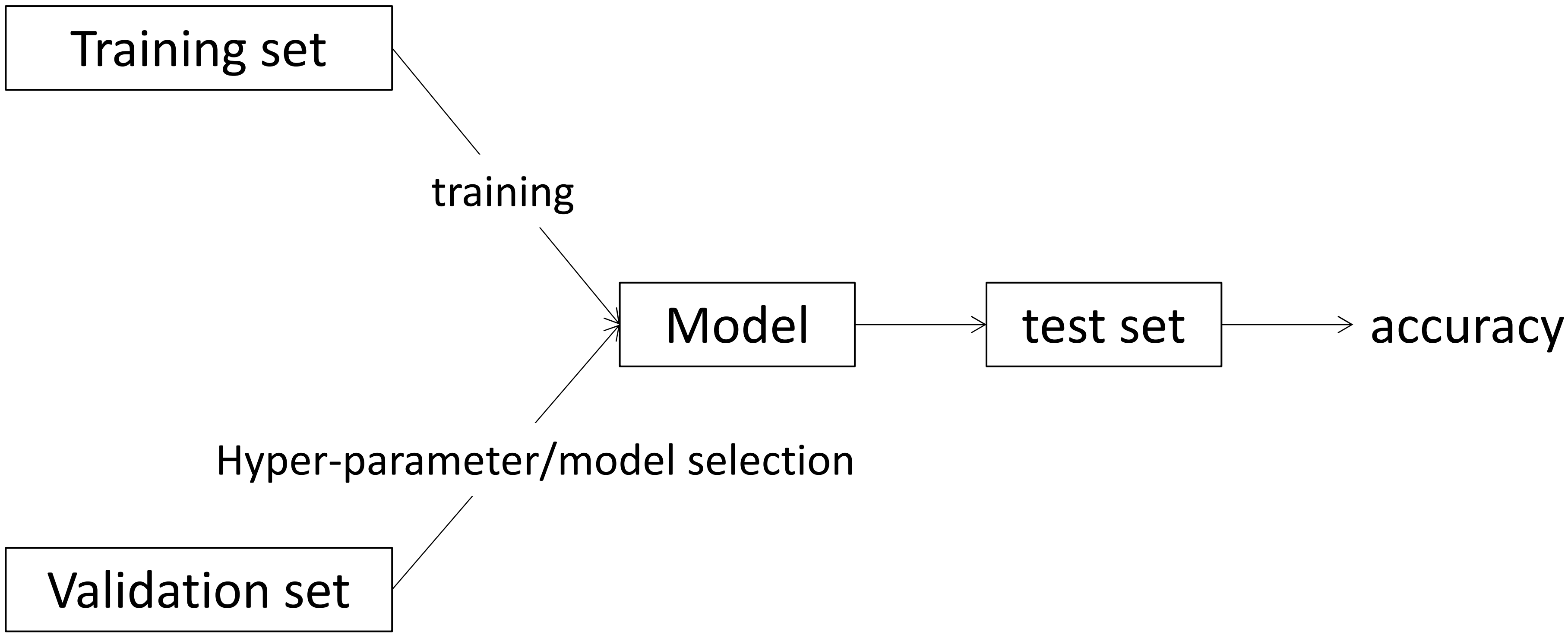
Data-centric Computer Vision

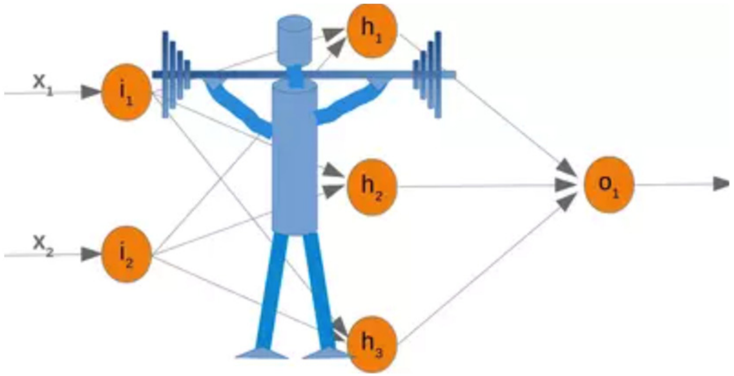
Liang Zheng
Australian National University
20 February 2023

Pillars in machine learning



Now suppose you are a researcher working at Google. You probably spend

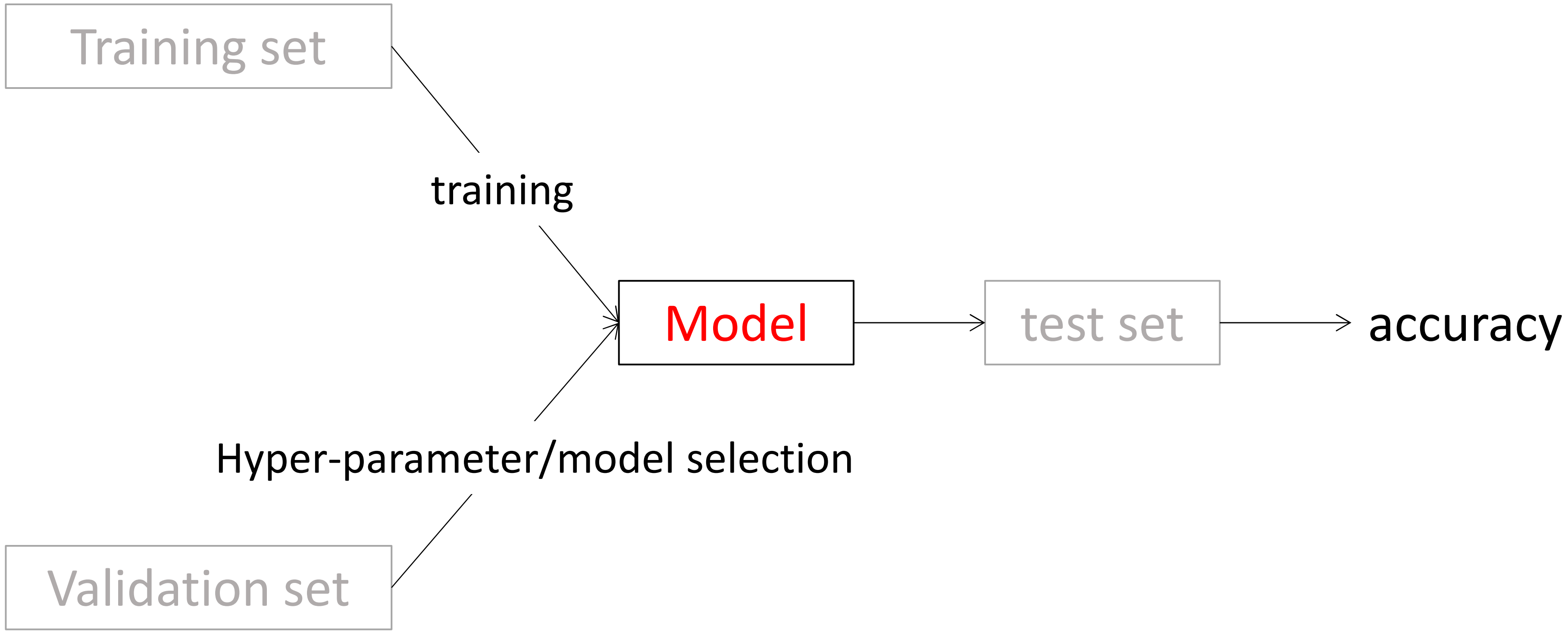
- half your time configuring your network



- the other half of your time collecting/cleaning data

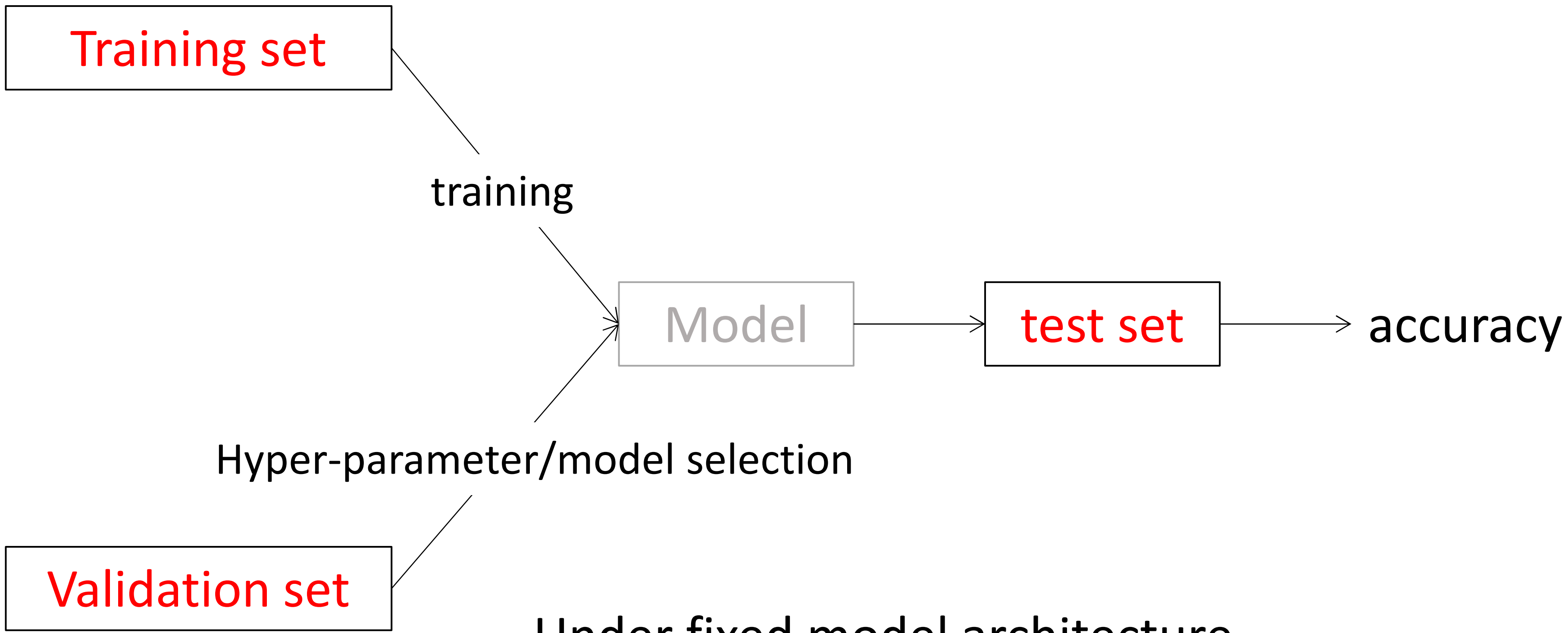
What most works are studying

algorithm-centric research



Features: SIFT, GIST, color histogram, etc
Hand-crafted models: SVM, boosting, sparse coding etc
Deep models: ResNet, DenseNet, Transformers…

What I'm going to talk about
data-centric research



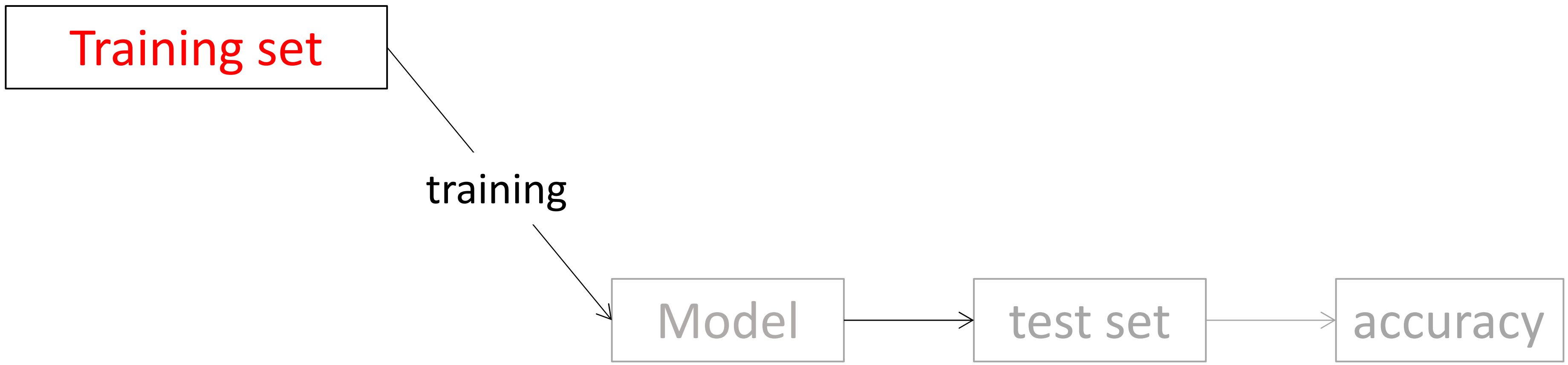
Under fixed model architecture,

- Can we improve the training data?
- Can we find good validation data?
- Can we estimate test set difficulty?

Outline

- Training data optimization
- Validation data search
- Label-free model evaluation (estimate test set difficulty)

Training data optimization

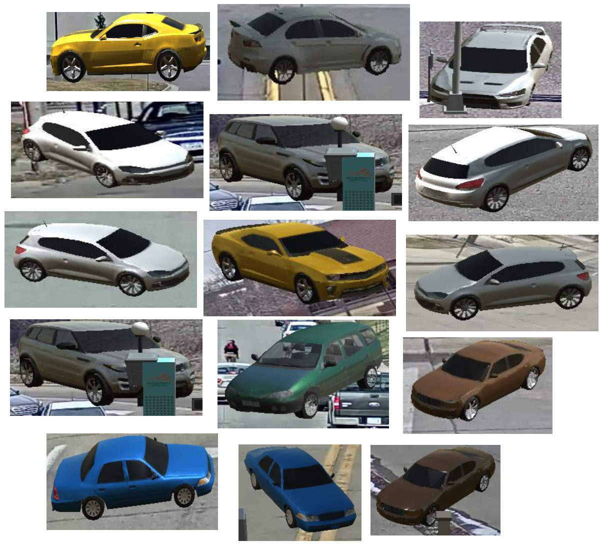


Objective: Given a model and a test set, we want to create a training set that gives us possibly high accuracy.

Yao et al., Simulating content consistent vehicle datasets with attribute descent, ECCV 2020

Training (source) data optimization

source



target

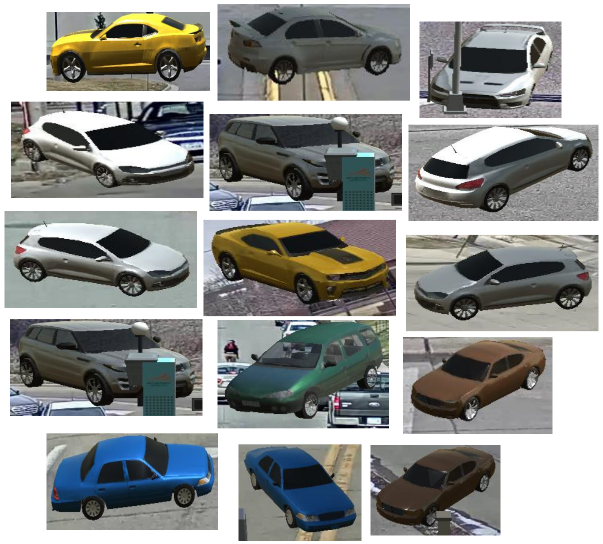
domain gap?

Style/feature alignment

Content alignment

Training (source) data optimization

source



idea

target

Objective: create a training set that has similar content with target data

We propose to use synthetic data

+ large-scale, quickly, accurately, cheaply Sun and Zheng, CVPR 2019

+ controllability and editability

+ challenging situation (danger forecast)

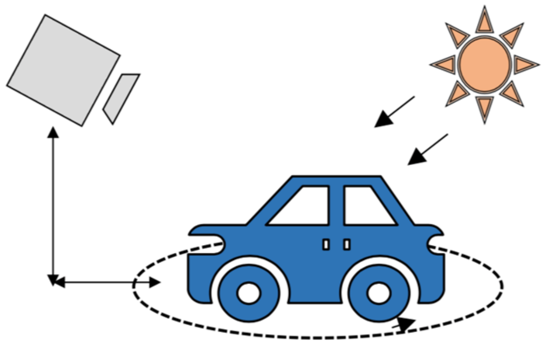
+ security and privacy issues

+ corner cases (heavy occlusion)

- different data distribution

We collected the VehicleX Dataset

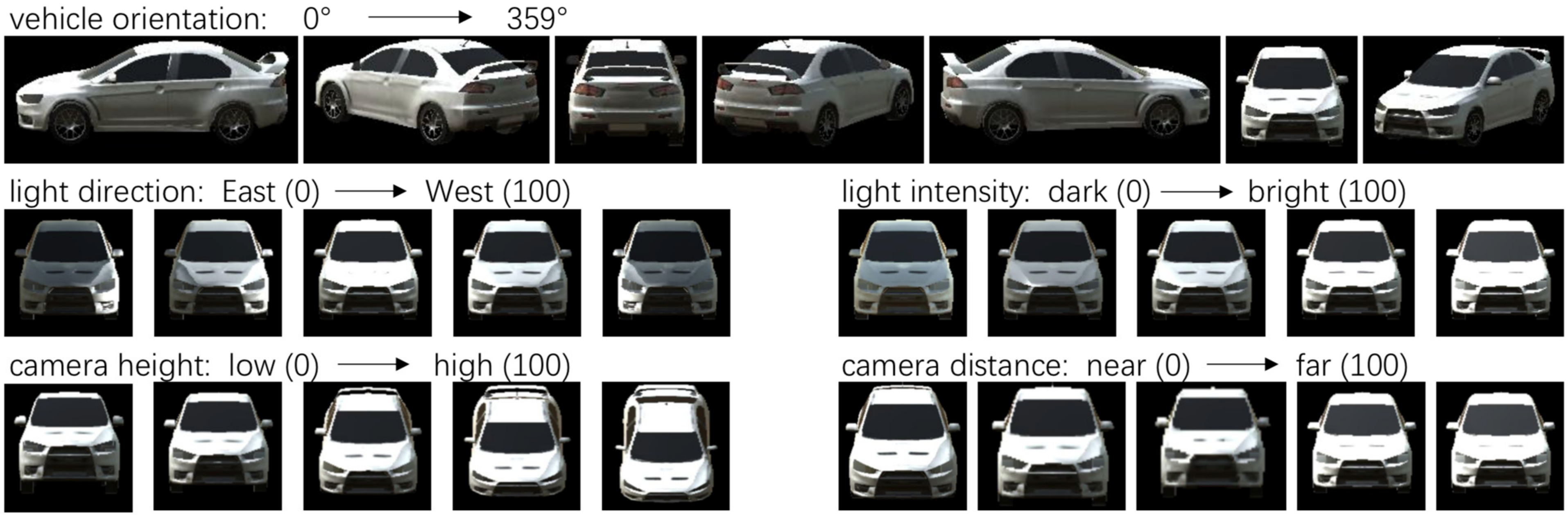
- 1,209 vehicles
- ~350 types of vehicles
- Platform: Unity
- Editable attributes: lighting direction, lighting intensity, vehicle orientation, camera height, camera distance



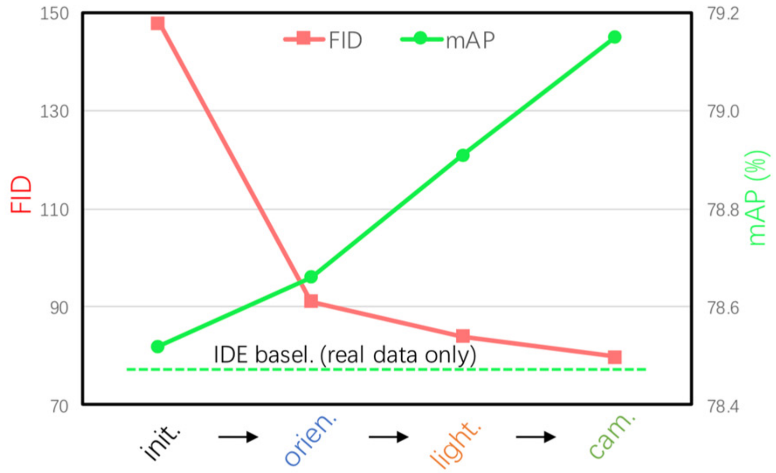
A Platform

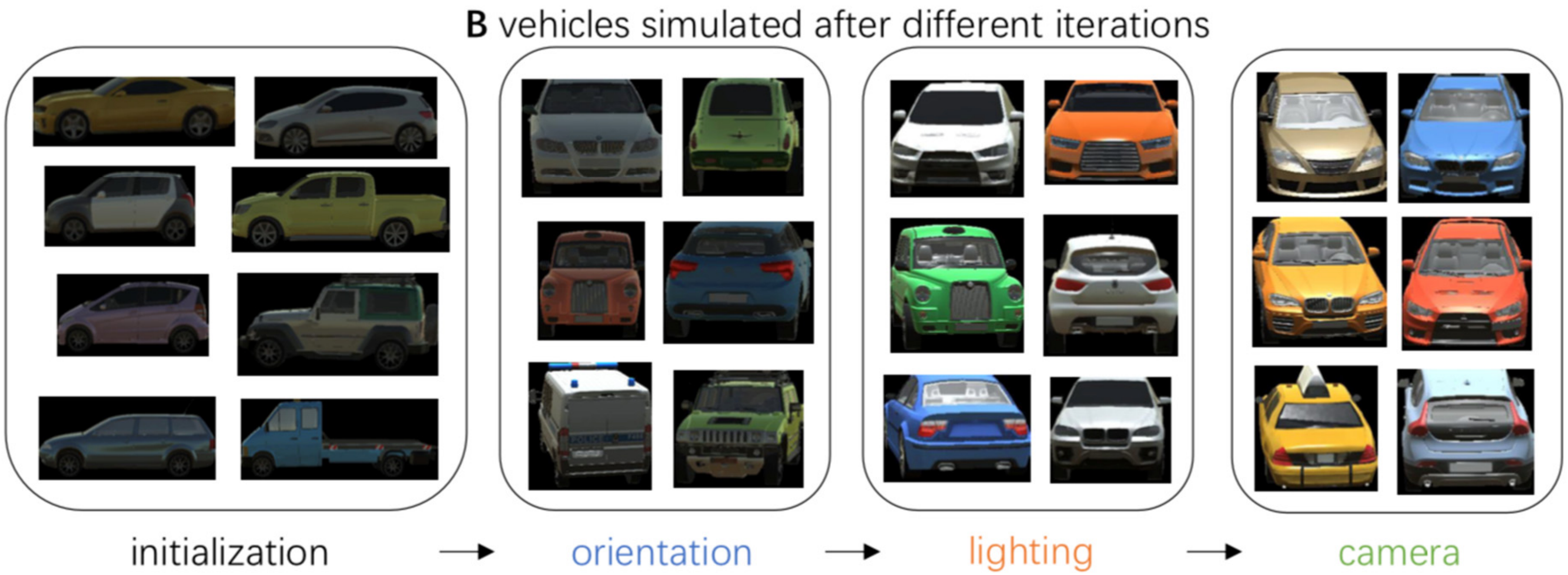
B Vehicle identities

Editable Attributes



Attribute descent



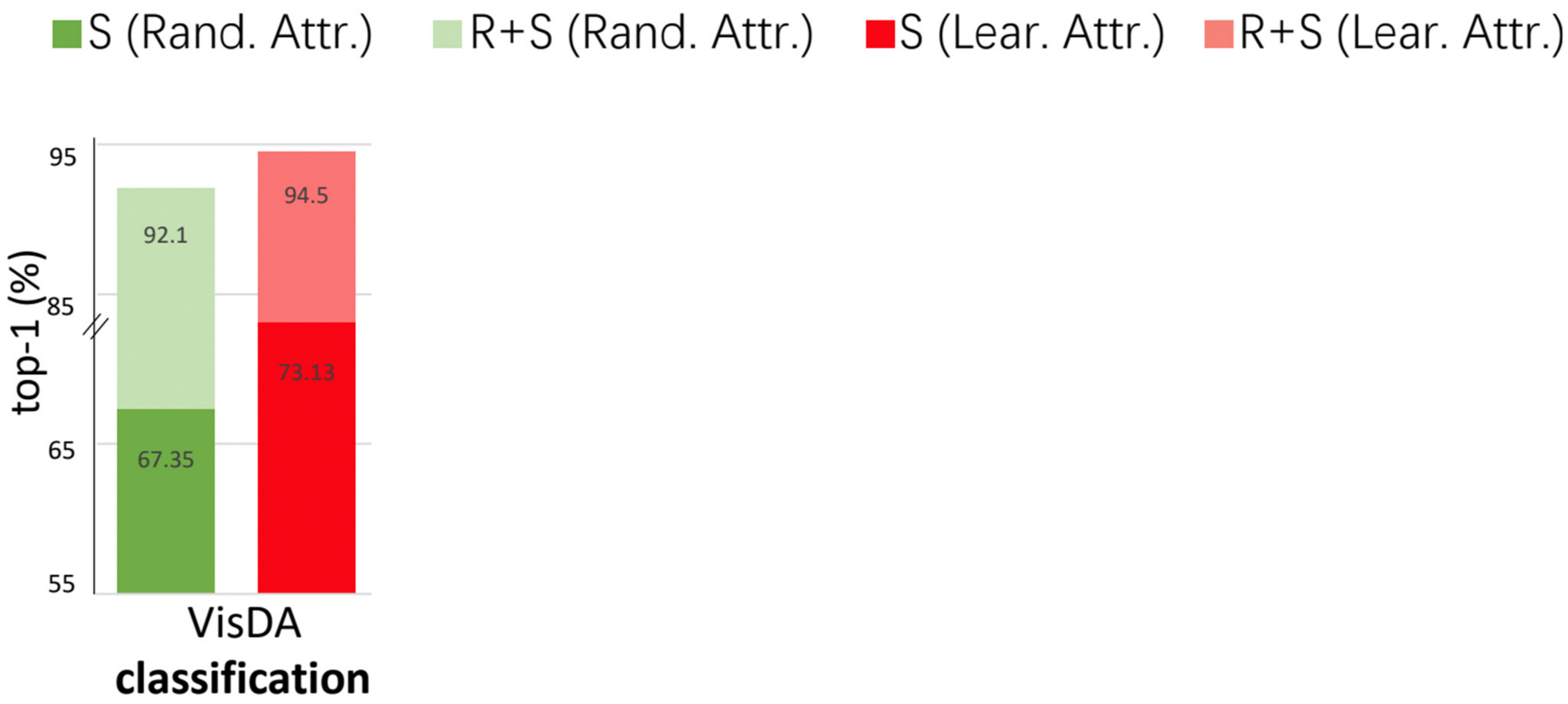


We optimize the value of each attributes successively

For a given attribute, we search (brute-force) for its optimum value such that FID is minimized

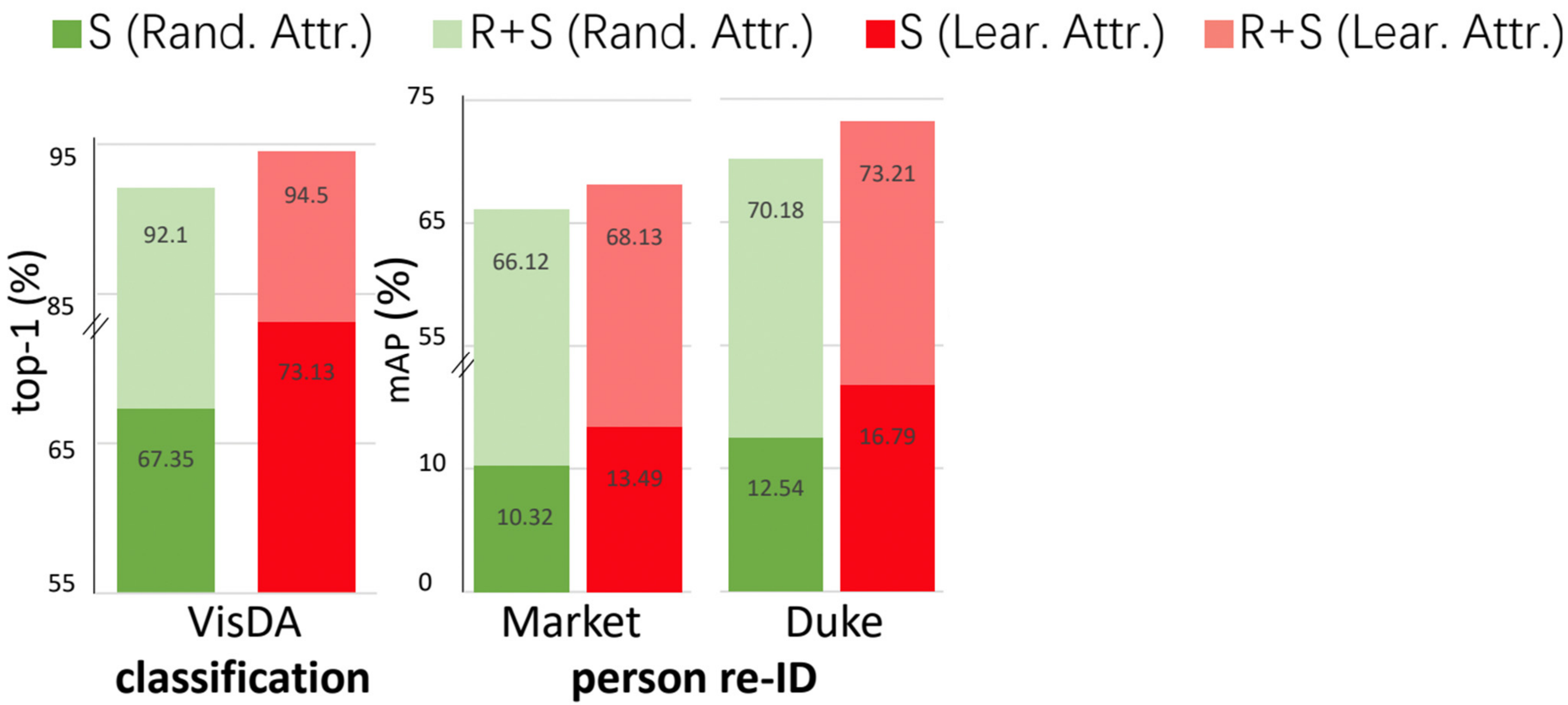
Experiment – statistical significance

- Learned attribute vs. random attribute



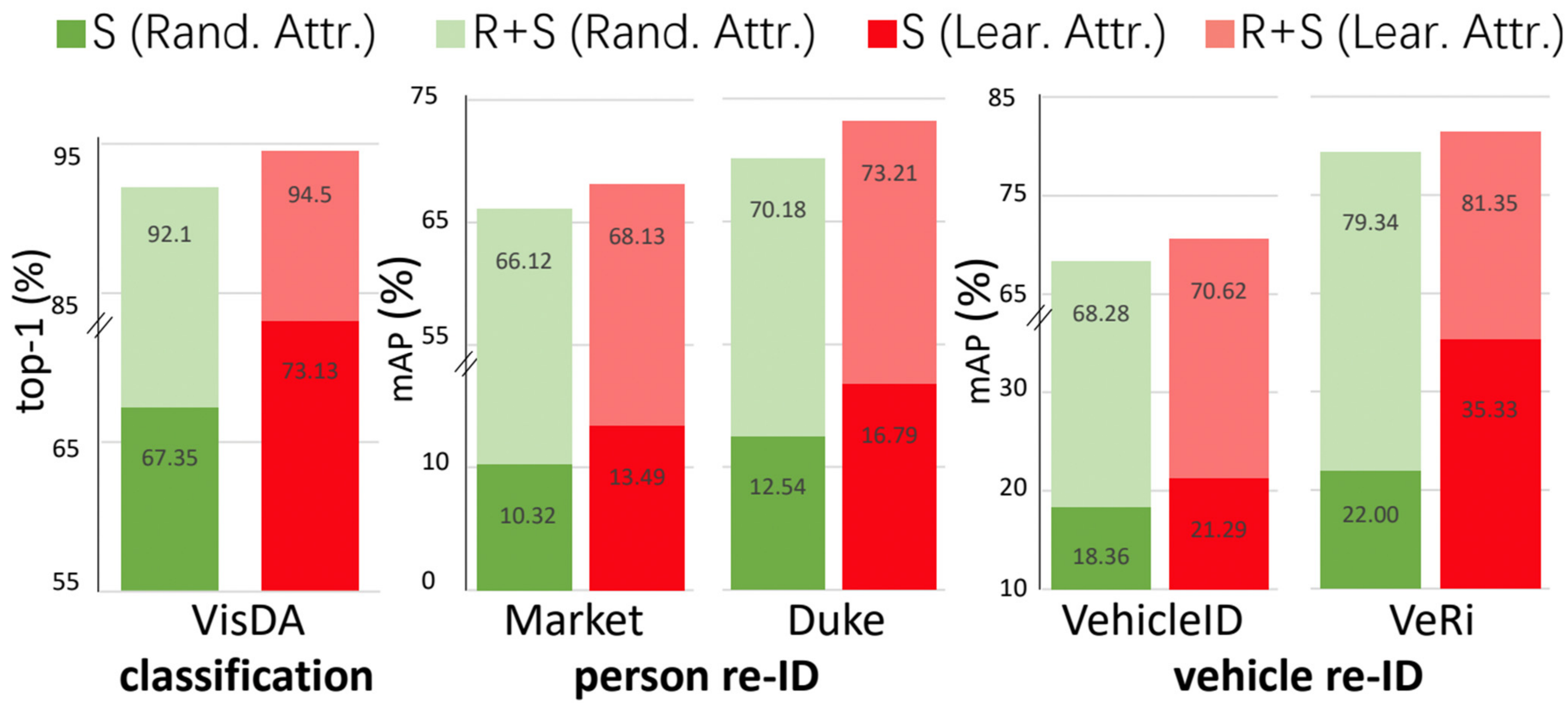
Experiment – statistical significance

- Learned attribute vs. random attribute



Experiment – statistical significance

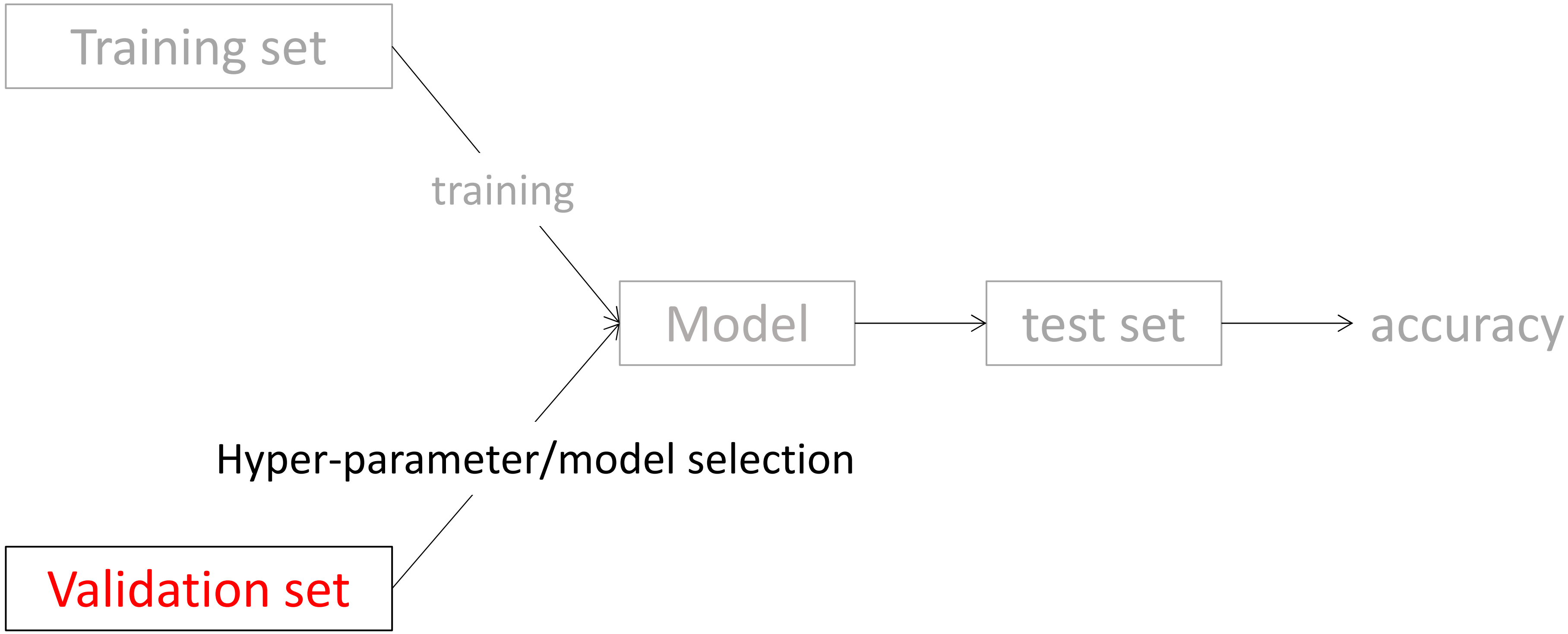
- Learned attribute vs. random attribute



Outline

- Training data optimization
- Validation data search
- Label-free model evaluation (estimate test set difficulty)

Validation data search



We usually select models using a validation set

Training set

→ Models A, B, C, D, E

validation set

Model comparison

B ≻D ≻C ≻A ≻E

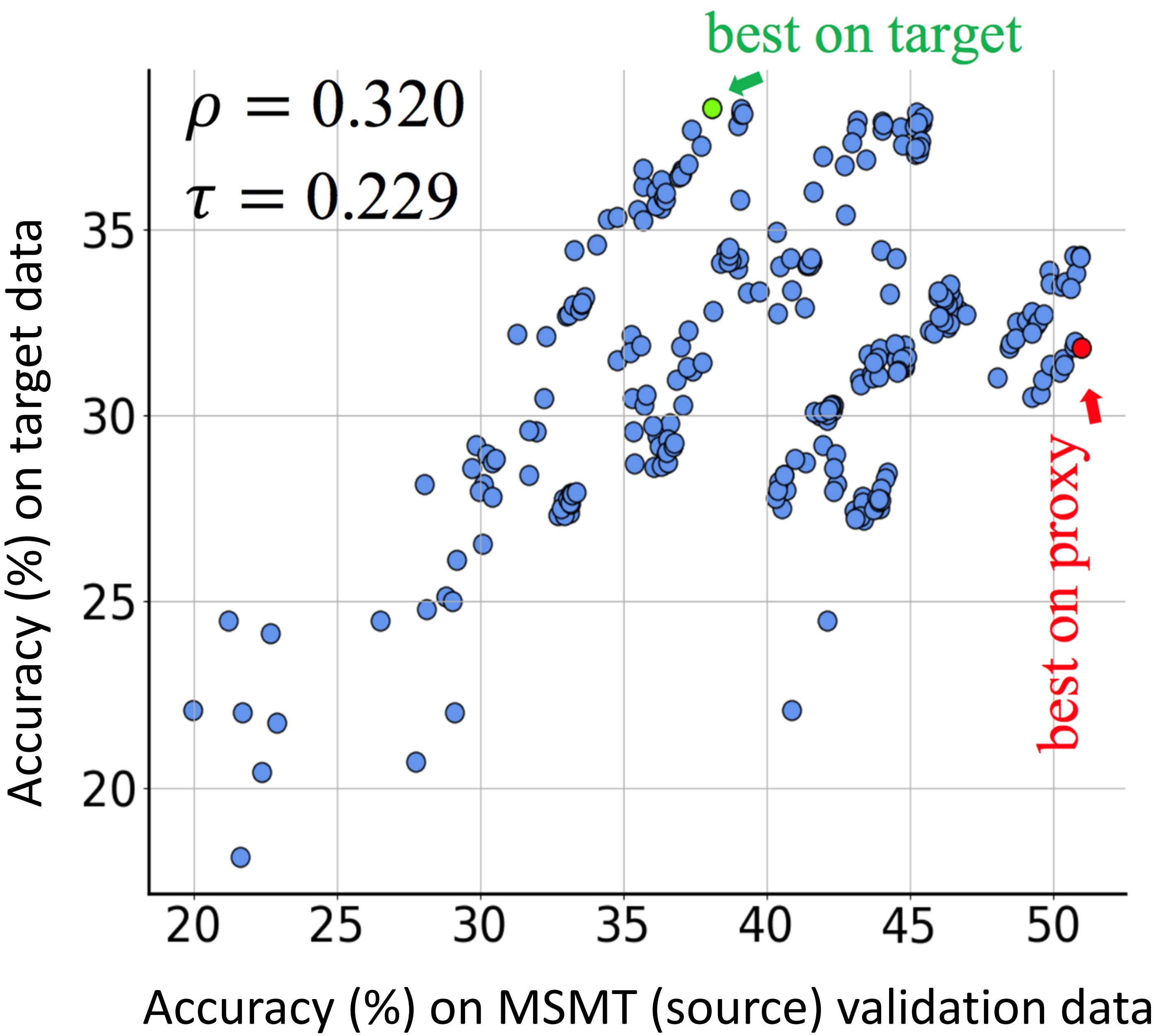
We will deploy B in testing

However, if we deploy the models to another domain…

Training (source) data

Target data

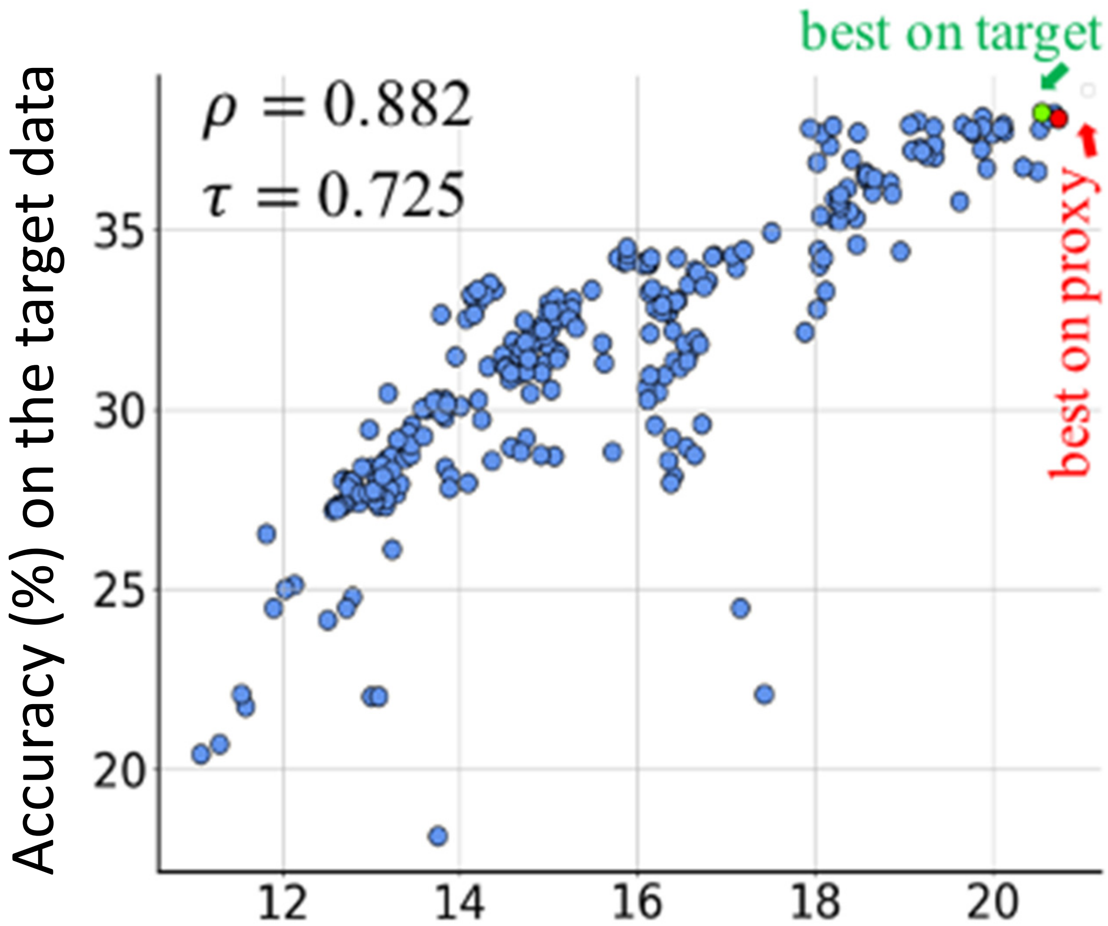
Will we still have B ≻D ≻C ≻A ≻E on this target domain?



280 models trained on the source

We want to search a validation set that

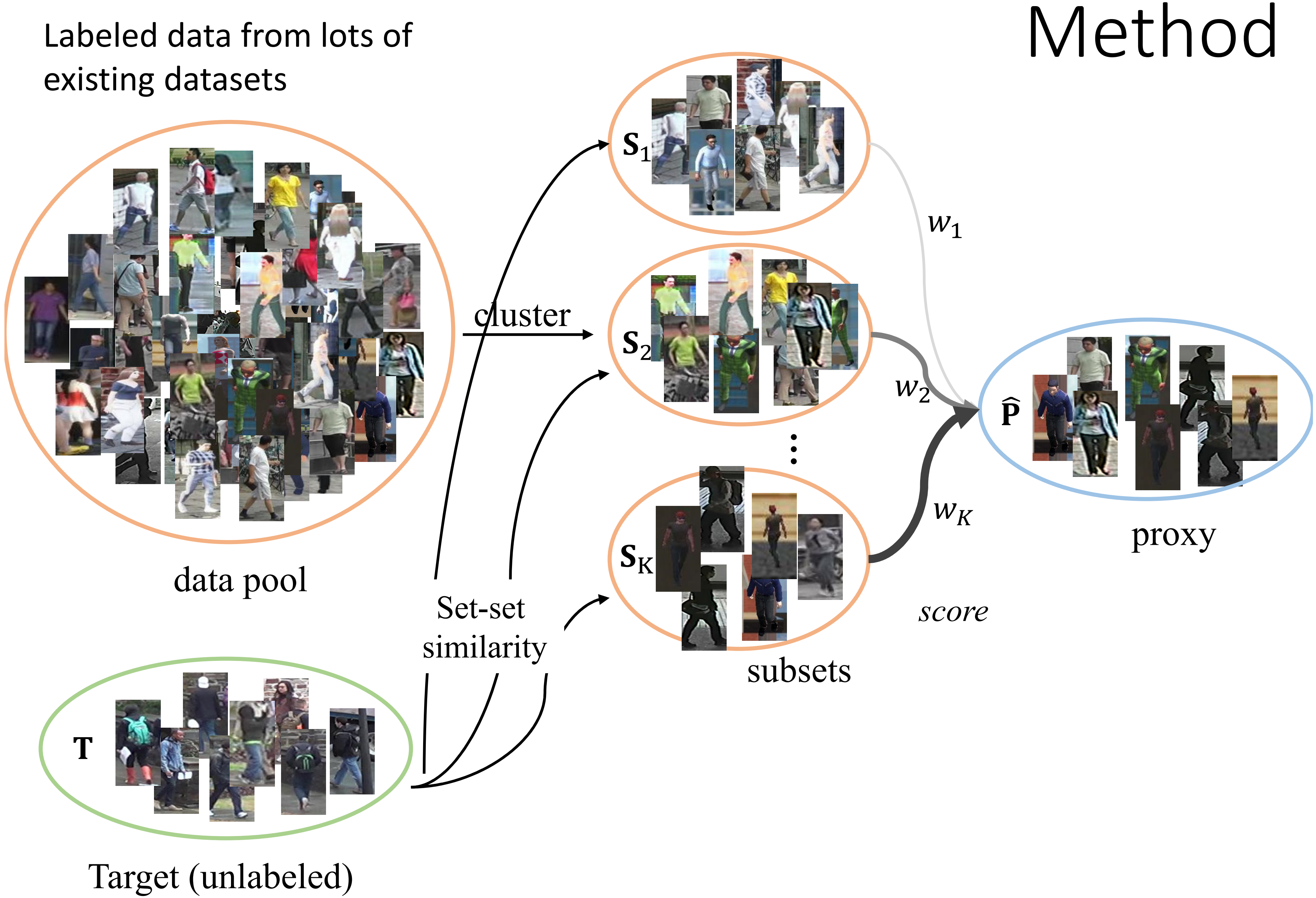
- is fully labeled
- has similar distributions with the target data

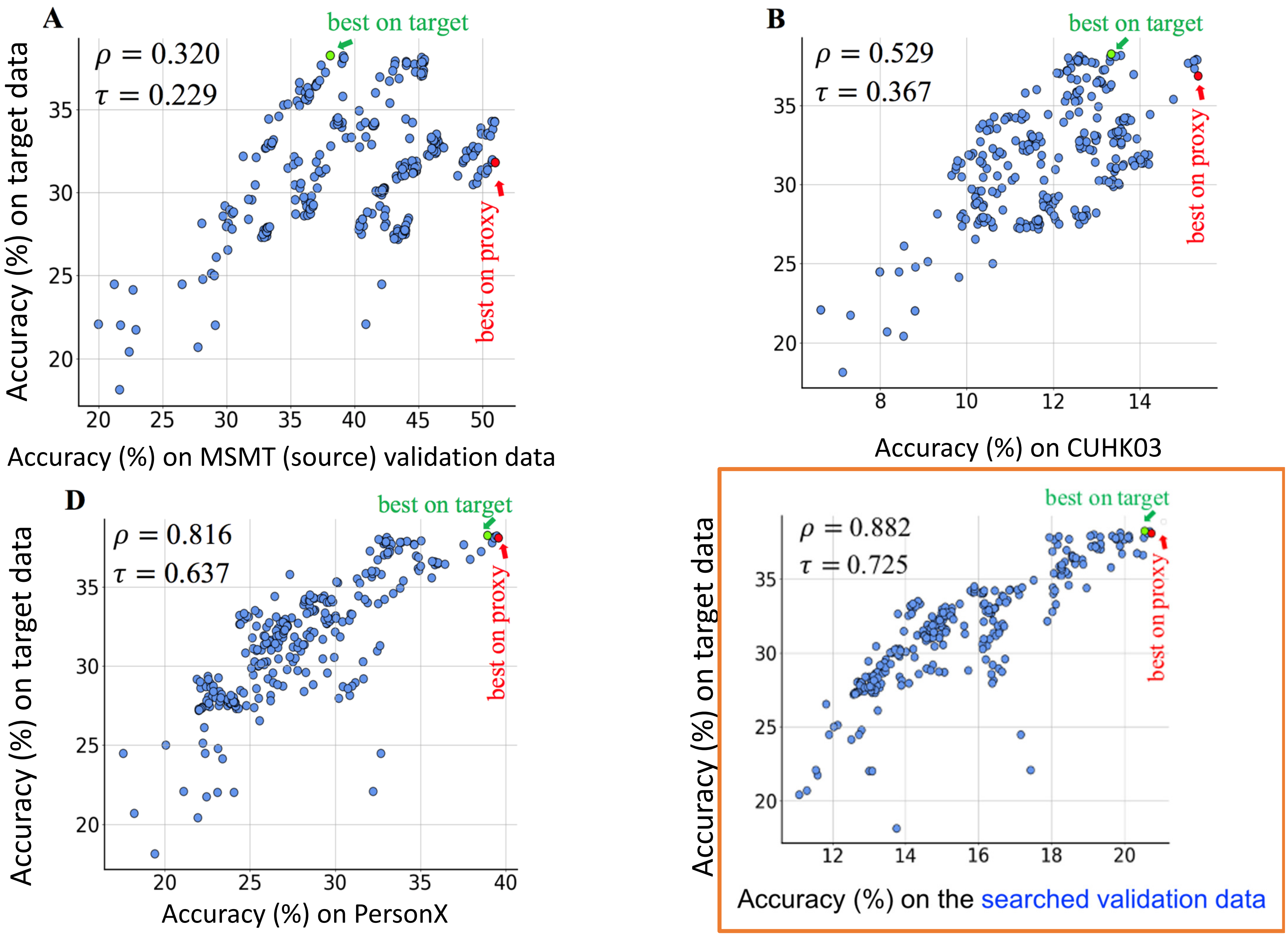


280 models trained on the source

Accuracy (%) on the searched validation data

Method

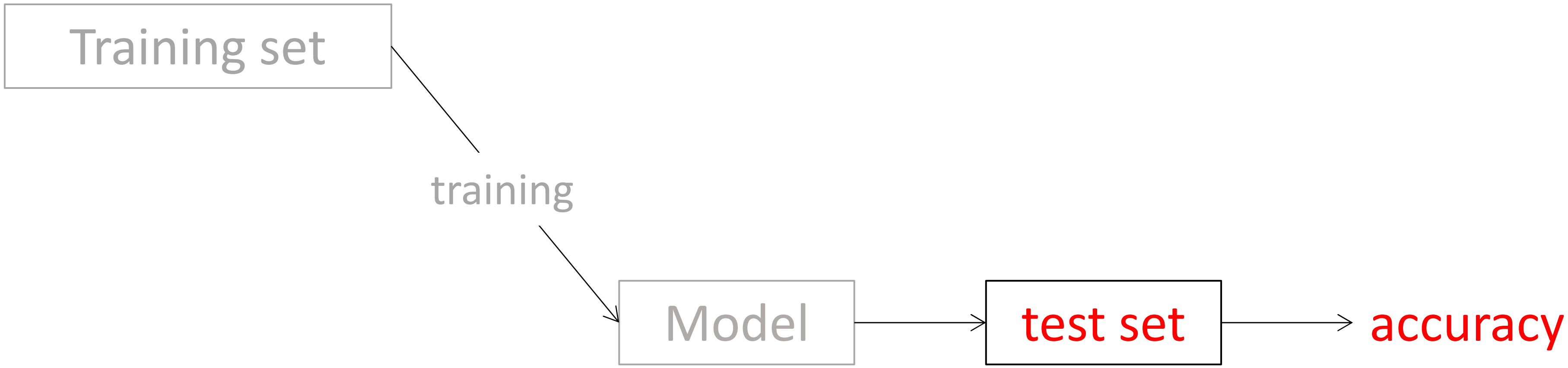

Labeled data from lots of existing datasets
data pool
cluster
$\mathbf{S}_1$
$\mathbf{S}_2$
$\mathbf{S}_K$
subsets
w_1
w_2
w_K
score
$\hat{\mathbf{P}}$
proxy
Set-set similarity
$\mathbf{T}$
Target (unlabeled)


A
best on target
$\rho = 0.320$
$\tau = 0.229$
best on proxy
Accuracy (%) on target data
Accuracy (%) on MSMT (source) validation data
B
best on target
$\rho = 0.529$
$\tau = 0.367$
best on proxy
Accuracy (%) on CUHK03
D
best on target
$\rho = 0.816$
$\tau = 0.637$
best on proxy
Accuracy (%) on target data
Accuracy (%) on PersonX
best on target
$\rho = 0.882$
$\tau = 0.725$
best on proxy
Accuracy (%) on target data
Accuracy (%) on the searched validation data

Outline

- Training data optimization
- Validation data search
- Label-free model evaluation (estimate test set difficulty)

Estimate test set difficulty (label-free model evaluation)

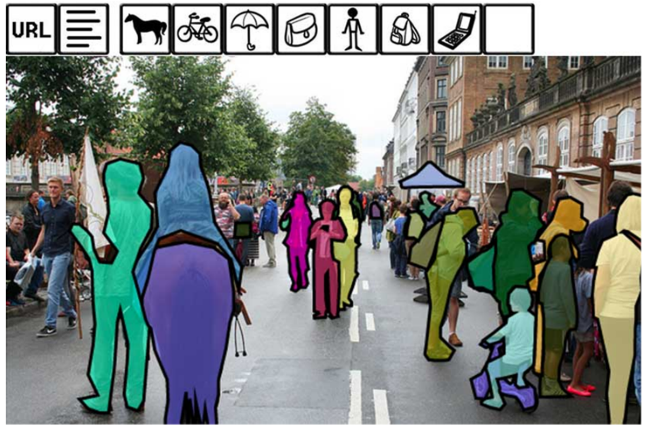


W. Deng and L. Zheng, Are Labels Necessary for Classifier Accuracy Evaluation? *CVPR, TPAMI, 2021*

Our usual way of evaluating models

- Yes

ImageNet



MSCOCO

Ground truths provided

LFW

However,...

We can’t calculate a classifier accuracy!!

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






Ground truths not provided

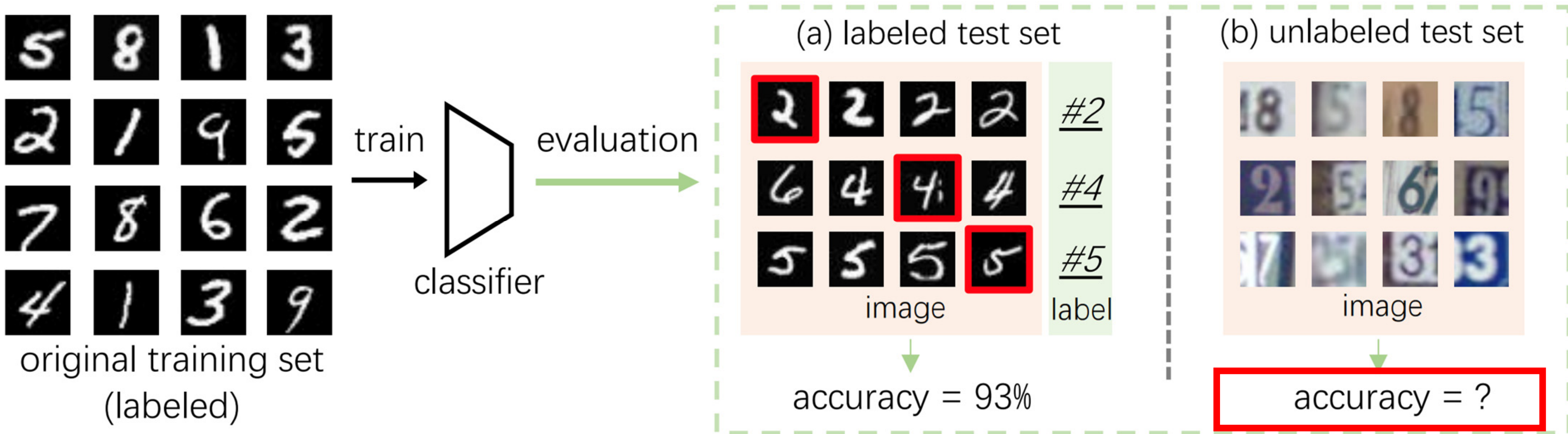
We encounter this problem too many times in CV applications….

- 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

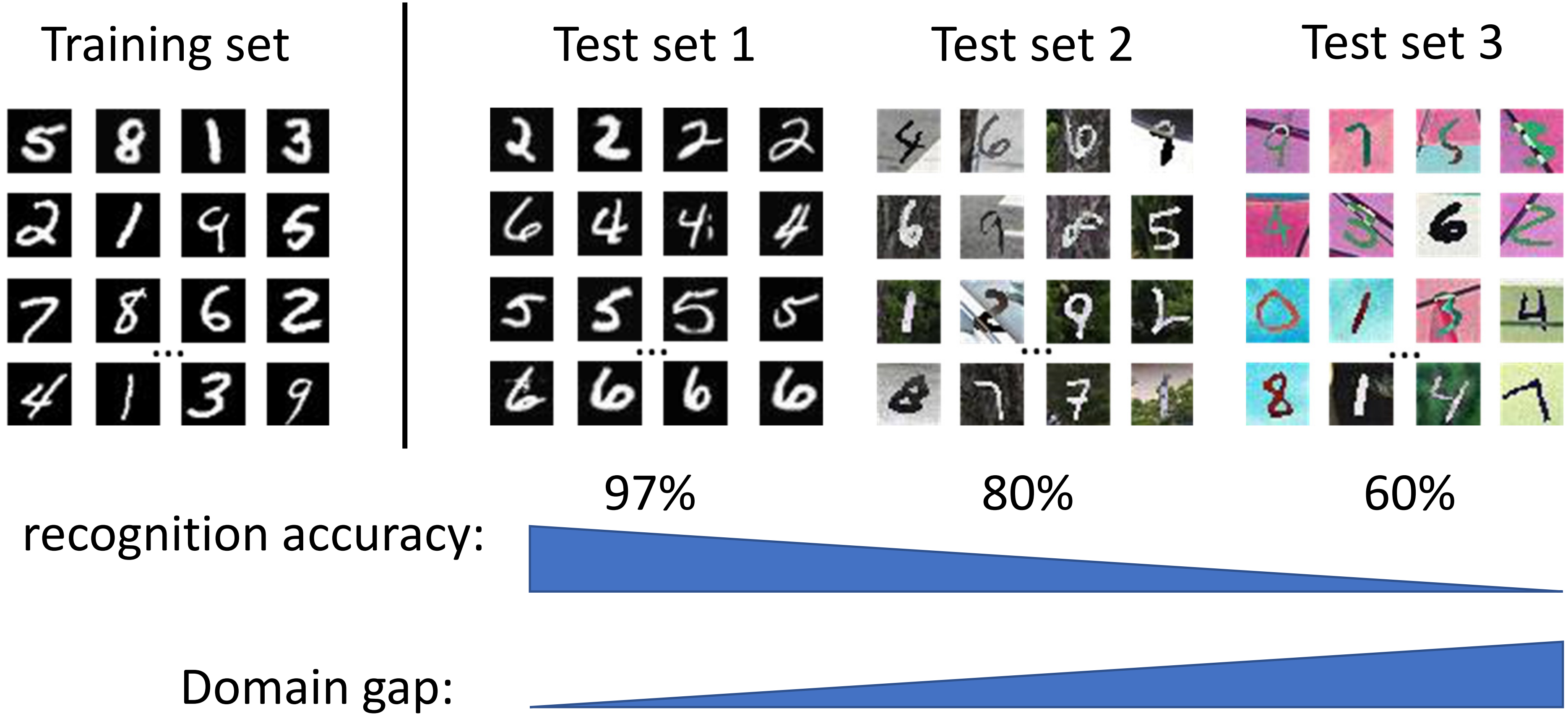
Formally, we want to solve:



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



Negative correlation between recognition accuracy and domain gap

Our idea

Known (from existing literature)

Larger domain gap -> lower recognition accuracy

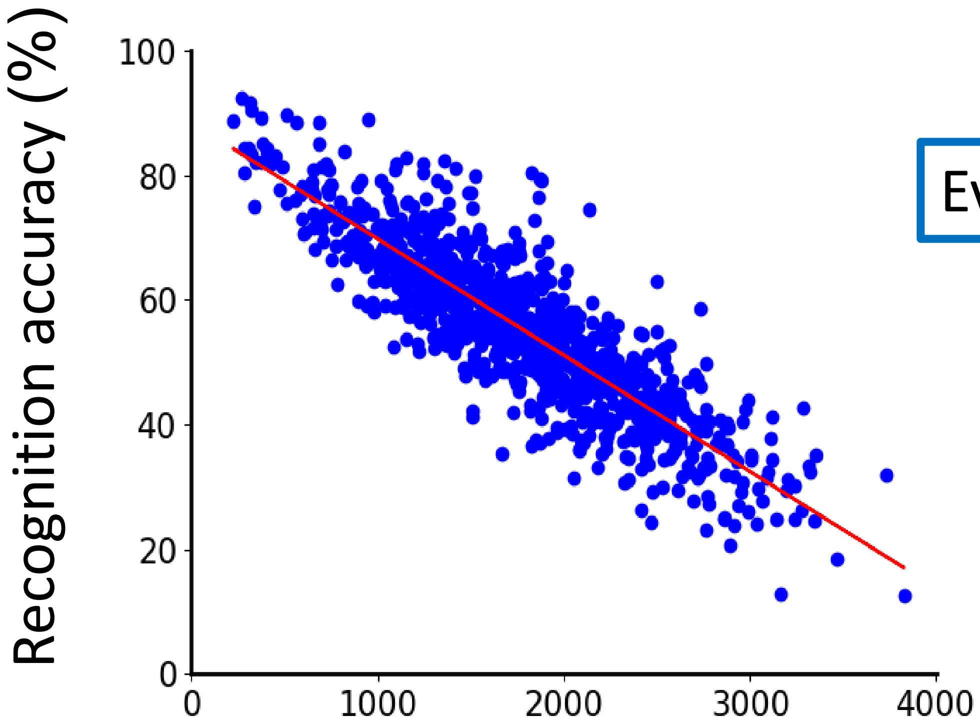
Unknown

Can we quantify this relationship?

A regression problem!

Some experiments

digit classification



Every point is a dataset

Fréchet distance

Domain gap between a training set and test sets

Qualitative examples

Train

GT: 71.07%
ours: 75.39%

GT: 40.19%
ours: 38.43%

GT: 90.16%
ours: 89.68%

We are organising the DataCV challenge @ CVPR 2023, on this label-free model evaluation problem.
https://sites.google.com/view/vdu-cvpr23/competition

Conclusions and insights

- We study data-centric computer vision problems
- Optimize the training set
  - given the test set and model architect
- Search and compose a validation set
  - Given the training set, a test set and models
- Estimate test set difficulty
  - Given the training set, test set and model

Conclusions and insights

- What else problems are data-centric?
  - Given a fine-tuning dataset, find a good pre-training dataset
  - Or the opposite
  - Estimate the noise level of a dataset
  - ….

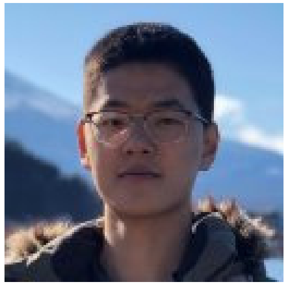
- Key techniques
  - Dataset representation
    - attribute values, feature mean, covariance etc..
  - Dataset-dataset similarity estimation
    - Frechet distance etc.

Thank you! Any question?

Collaborators

Xiaoxiao Sun
ANU

Yue Yao
ANU



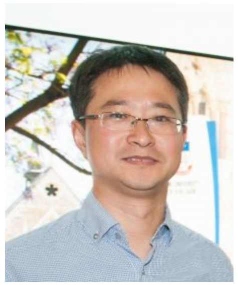
Yunzhong Hou
ANU

Weijian Deng
ANU

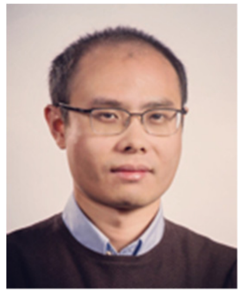
Stephen Gould
ANU

Milind Naphade
NVIDIA

Tom Gedeon
ANU



Hongdong Li
ANU



Xiaodong Yang
NVIDIA