DAVINZ: Data Valuation using Deep Neural Networks at Initialization

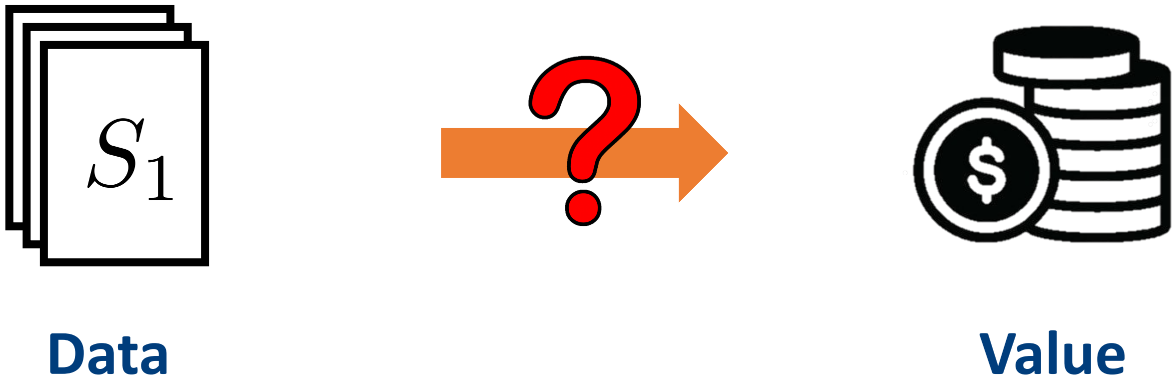
Zhaoxuan Wu[1], Yao Shu[2], Bryan Kian Hsiang Low[2]

Institute of Data Science, National University of Singapore[1]
Department of Computer Science, National University of Singapore[2]

National University of Singapore

International Conference On Machine Learning

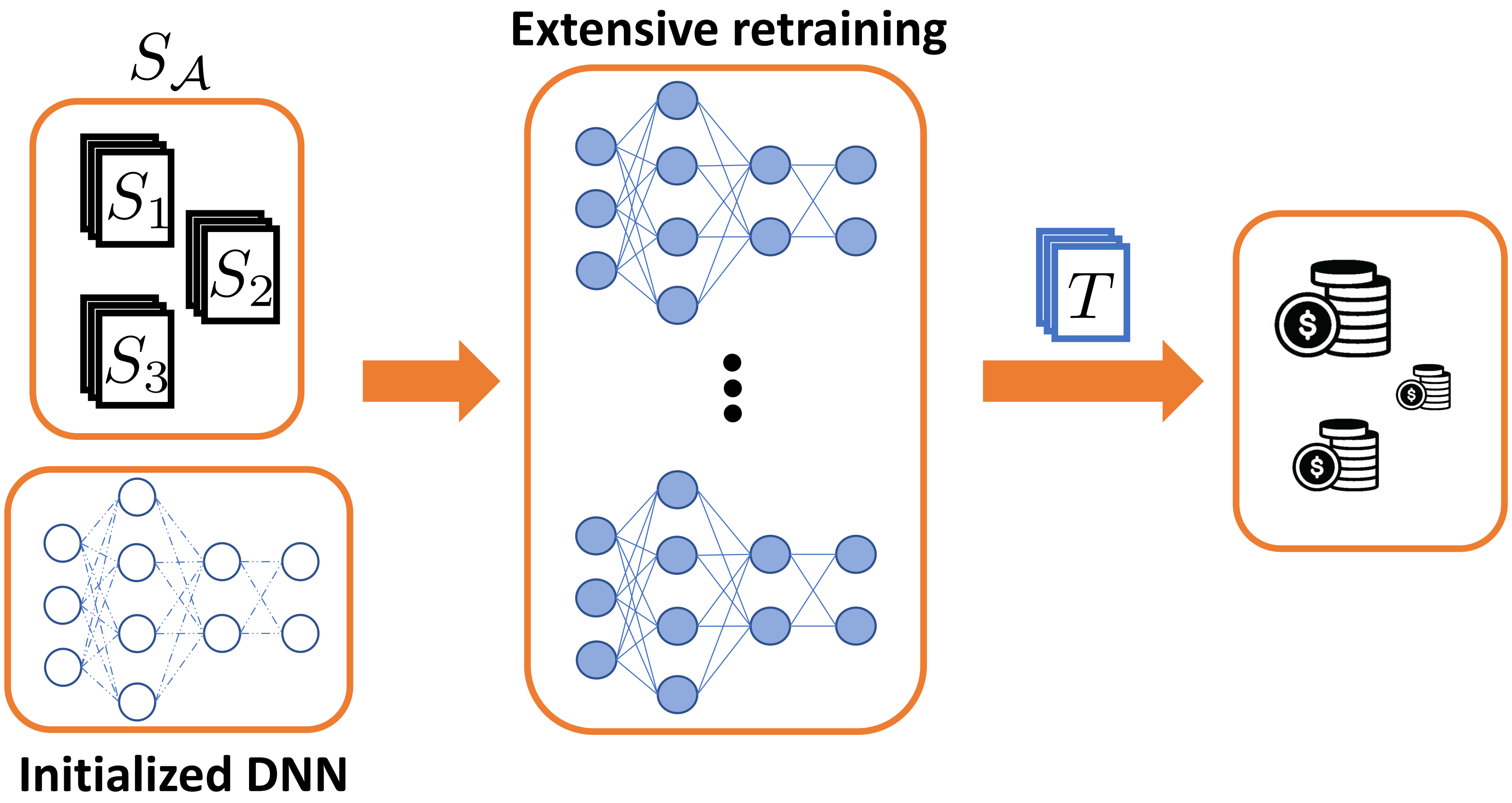
Background & Motivation



- Data valuation
  - Data with different qualities typically lead to diverse ML model performances

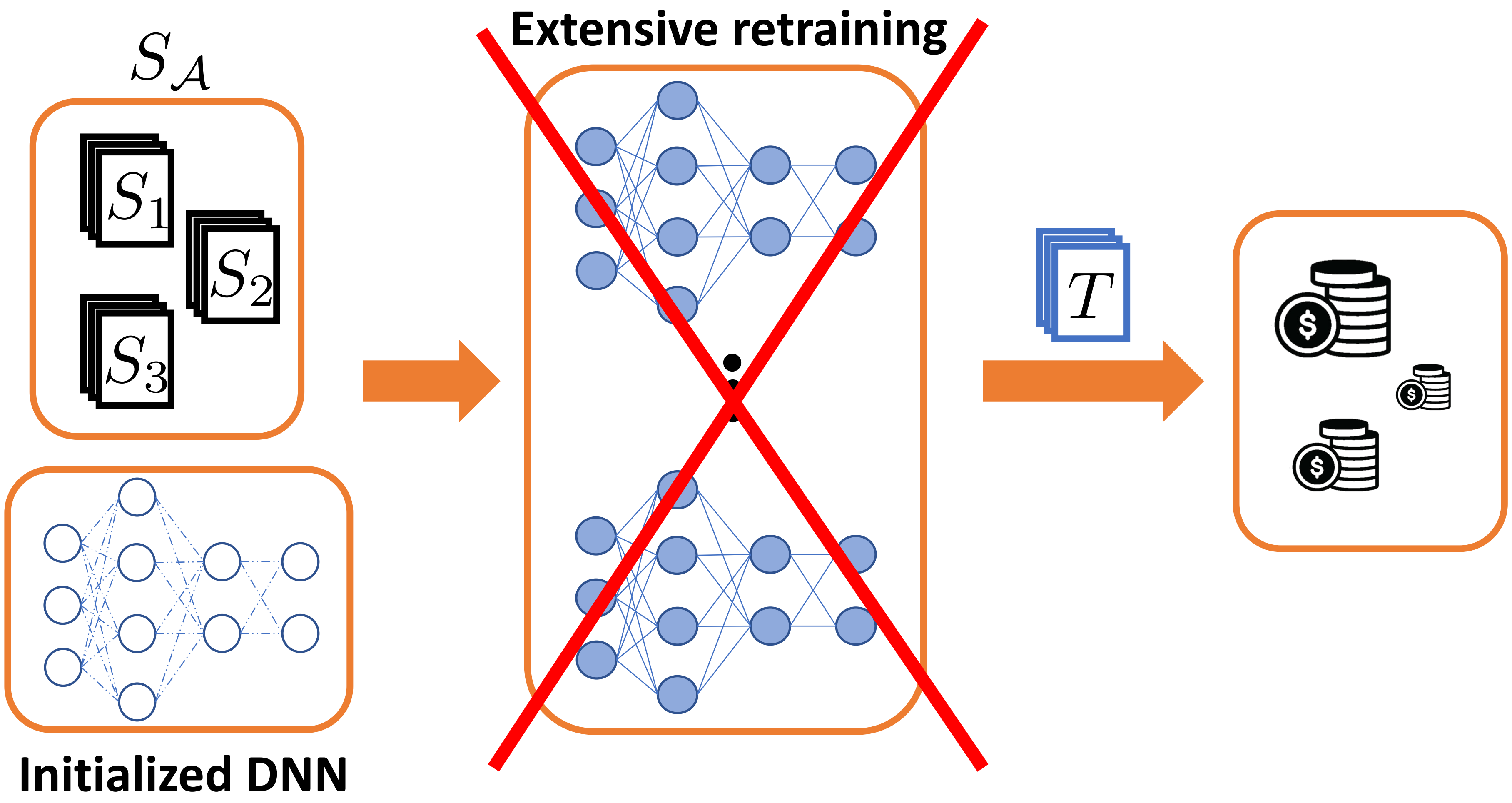
Background & Motivation

The conventional data valuation



Background & Motivation

The conventional data valuation



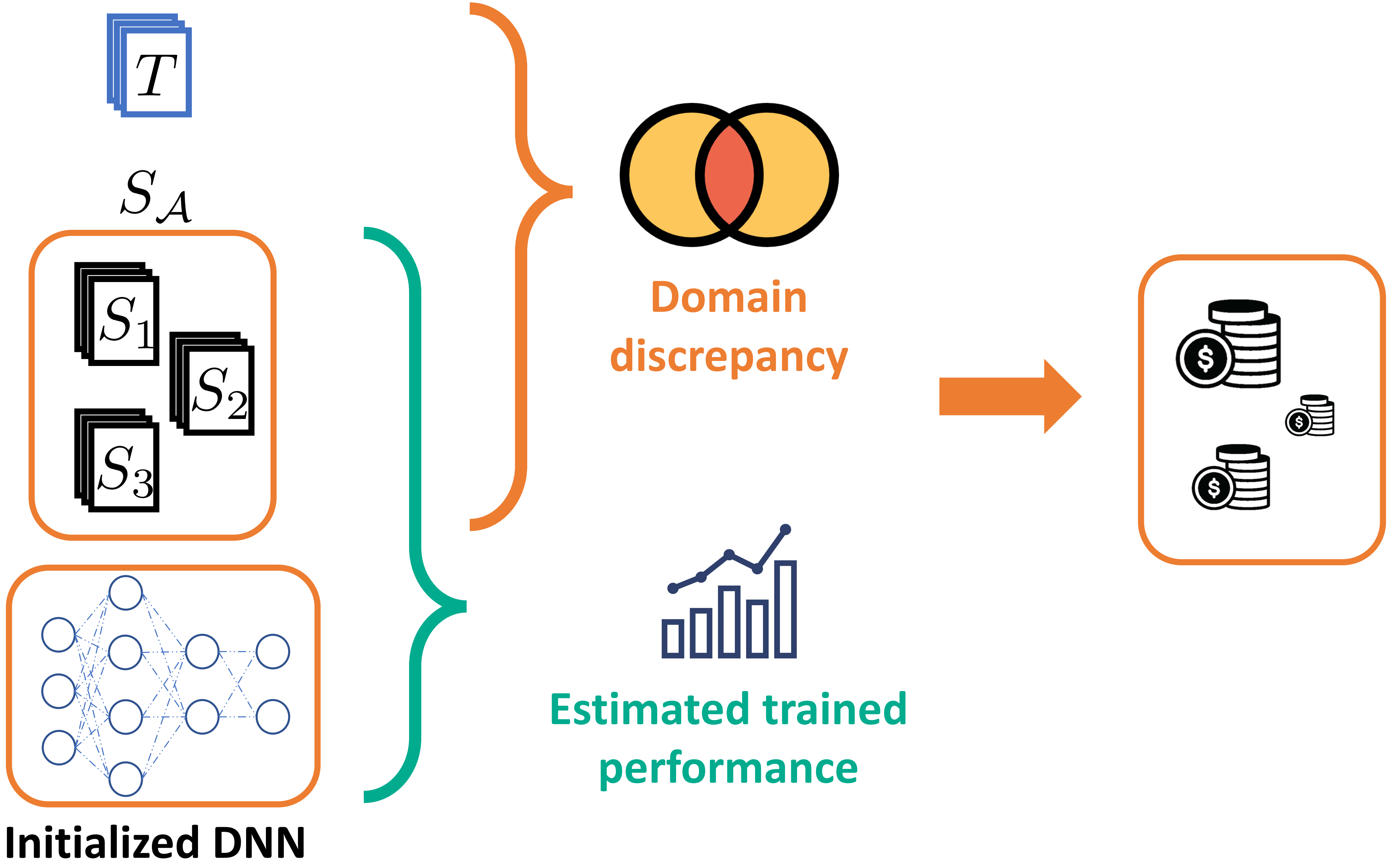
Motivation

- Estimate the *domain-aware generalization performance* of *DNNs without actual model training*

- Neural tangent kernel (NTK)
  - Characterize the training dynamics of DNNs with gradient descent
  - The generalization performance can be theoretically bounded using NTK

- Domain adaption
  - In data valuation, an agent's dataset typically has a different distribution from the test dataset
  - Characterizes the generalization error caused by train-test domain discrepancy

The Idea



T
$S_{\mathcal{A}}$
S_1
S_2
S_3
Initialized DNN
Domain discrepancy
Estimated trained performance

Definitions & Notations

- NTK matrix

$$\boldsymbol{\Theta}(\boldsymbol{x}, \boldsymbol{x}'; \boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} f(\boldsymbol{x}, \boldsymbol{\theta})^\top \nabla_{\boldsymbol{\theta}} f(\boldsymbol{x}', \boldsymbol{\theta}) \in \mathbb{R}^{m \times m}$$

- **Definition 1** (Empirical Domain Discrepancy [1])

$$d_{\mathcal{H}}(T, S) \triangleq \sup_{h \in \mathcal{H}} \left| \frac{1}{m_T} \sum_{i=1}^{m_T} h(\boldsymbol{x}_i') - \frac{1}{m_S} \sum_{i=1}^{m_S} h(\boldsymbol{x}_i) \right|$$

[1] Gretton, A., Borgwardt, K. M., Rasch, M. J., Sch¨olkopf, B., and Smola, A. A kernel two-sample test. JMLR, 13(25): 723–773, 2012a.

Domain-aware Generalization Bound

- **Theorem 1**

$$\mathcal{L}_{\mathcal{D}_T}(f_t) \leq \mathcal{L}_S(f_t) + 2\rho\sqrt{\widehat{\boldsymbol{y}}^\top \boldsymbol{\Theta}_0^{-1}\widehat{\boldsymbol{y}}/m_S} + d_{\mathcal{H}}(T, S) + \varepsilon$$

- Two sources of error: (a) *in-domain error* and (b) *out-of-domain* error
- In-domain error
  - More complex the data, higher the generalization error $\mathcal{L}_{\mathcal{D}_T}$
- Out-of-domain error
  - More different S is from T, higher the generalization error $\mathcal{L}_{\mathcal{D}_T}$

Training-free Data Valuation

- Based on Theorem 1, we propose the scoring function

$$\nu(S) = -\kappa\sqrt{\widehat{\boldsymbol{y}}^\top \boldsymbol{\Theta}_0^{-1} \widehat{\boldsymbol{y}} / m_S} - d_{\mathcal{H}}(T, S) \quad (3)$$

- An empirical hyper-parameter κ
  - Balances the averaged scales of the in-domain and out-of-domain error

$$\kappa = \frac{\sum_{i=1}^{K} d_{\mathcal{H}}(T, S_i)}{\sum_{i=1}^{K} \left(\widehat{\boldsymbol{y}}_{S_i}^\top \boldsymbol{\Theta}_{0,S_i}^{-1} \widehat{\boldsymbol{y}}_{S_i} / m_{S_i}\right)^{1/2}}$$

Training-free Data Valuation

- Algorithm

Algorithm 1 **Da**ta **V**aluation at **In**itiali**z**ation (DAVINZ)

1: **Input:** Datasets $\{S_i\}_{i=1}^K$ from K data contributors, validation dataset T, DNN model f with initialized parameters $\boldsymbol{\theta}_0$, kernel k for $d_{\mathcal{H}}$, weighting factors $\alpha_{\mathcal{C}}$
2: **for** contributor $i = 1, \ldots, K$ **do**
3: **for** coalition $\mathcal{C} \subseteq \mathcal{A} \setminus \{i\}$ **do**
4: Evaluate the scores $\nu(S_{\mathcal{C}\cup\{i\}})$ and $\nu(S_{\mathcal{C}})$ by (3)
5: Evaluate the marginal $\Delta_{i,\mathcal{C}} = \nu(S_{\mathcal{C}\cup\{i\}}) - \nu(S_{\mathcal{C}})$
6: **end for**
7: $\phi_i = \sum_{\mathcal{C}\subseteq\mathcal{A}\setminus\{i\}} \alpha_{\mathcal{C}} \times \Delta_{i,\mathcal{C}}$
8: **end for**

Properties of DAVINZ

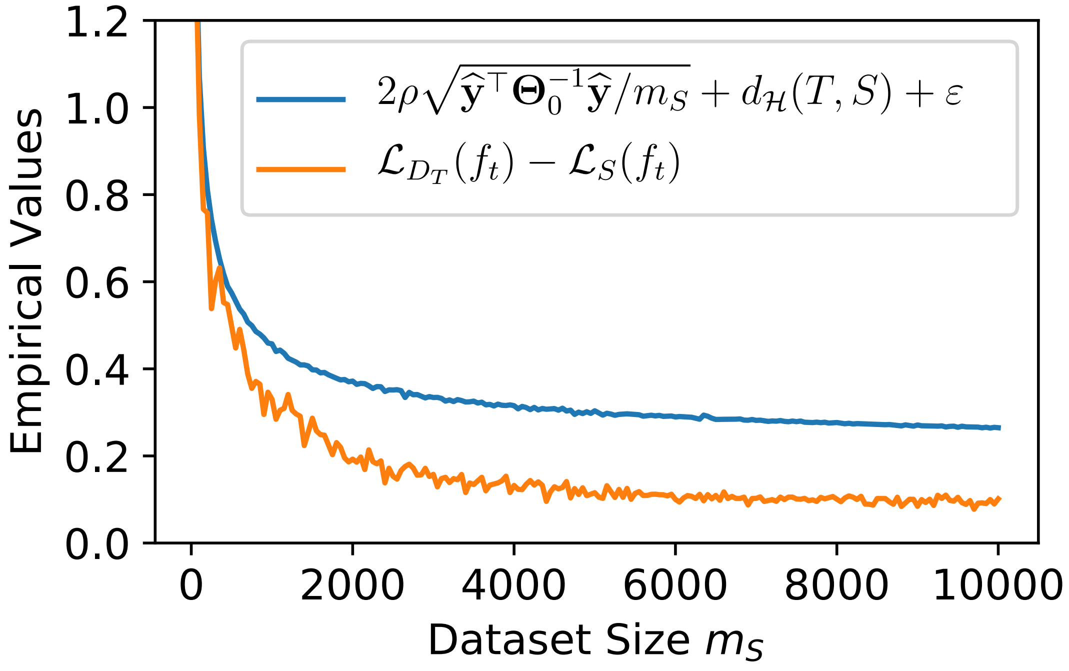
We theoretically prove the following properties:

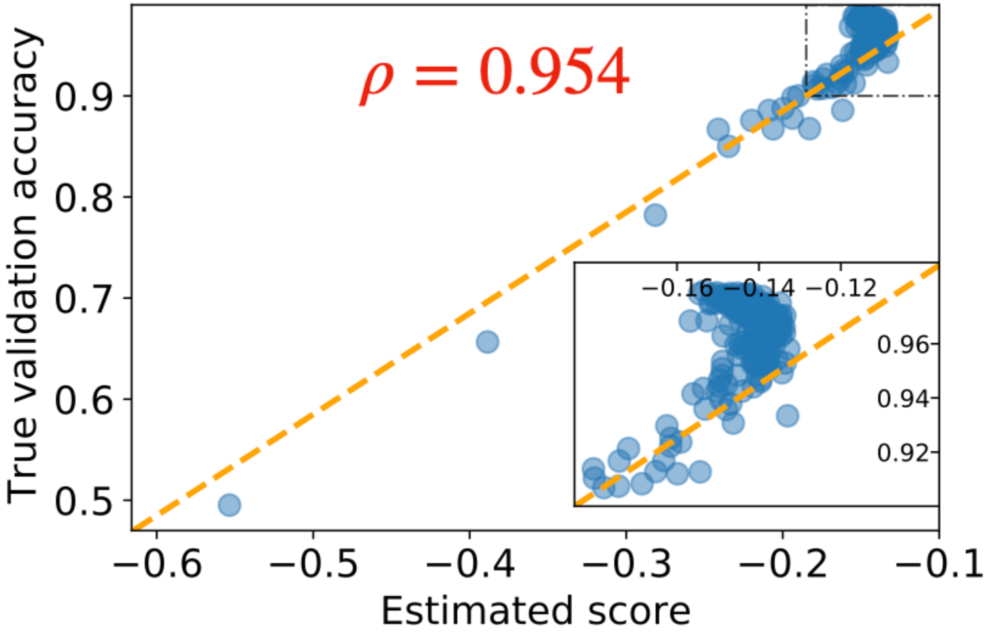
1. [**Theorem 1**] Awareness of Data Preference
2. [**Proposition 1**] Awareness of Data Quantity
3. [**Proposition 2**] Stability to Noise
4. [**Proposition 3**] Robustness to Model

Experiments

Valid scoring function in practice

- Construct 200 datasets each consisting of up to 10K MNIST images
- DNN with 2 Conv layers





Experiments

Effective and efficient DAVINZ

Method	Model	MNIST			CIFAR-10		
		Pearson	Spearman	Cost (Min.)	Pearson	Spearman	Cost (Min.)
VP	VGG13	1.00±0.00	0.98±0.01	88.6	0.53±0.28	0.77±0.09	88.4
	ResNet18	0.99±0.00	0.97±0.01	185.9	0.63±0.17	0.70±0.09	211.8
IF	VGG13	0.17±0.04	0.30±0.07	11.0	0.55±0.04	0.57±0.03	11.0
	ResNet18	0.42±0.05	0.55±0.07	22.6	0.08±0.07	0.07±0.10	26.3
RV	VGG13	−0.01±0.05	−0.14±0.08	9.7	0.17±0.03	0.32±0.06	9.6
	ResNet18	−0.36±0.11	−0.30±0.05	18.8	0.18±0.05	0.22±0.07	21.6
DAVINZ	VGG13	0.84±0.01	0.52±0.02	**2.5**	0.46±0.10	0.44±0.12	**2.0**
	ResNet18	0.85±0.00	0.62±0.00	**3.3**	0.55±0.03	0.67±0.03	**3.2**

Training-based

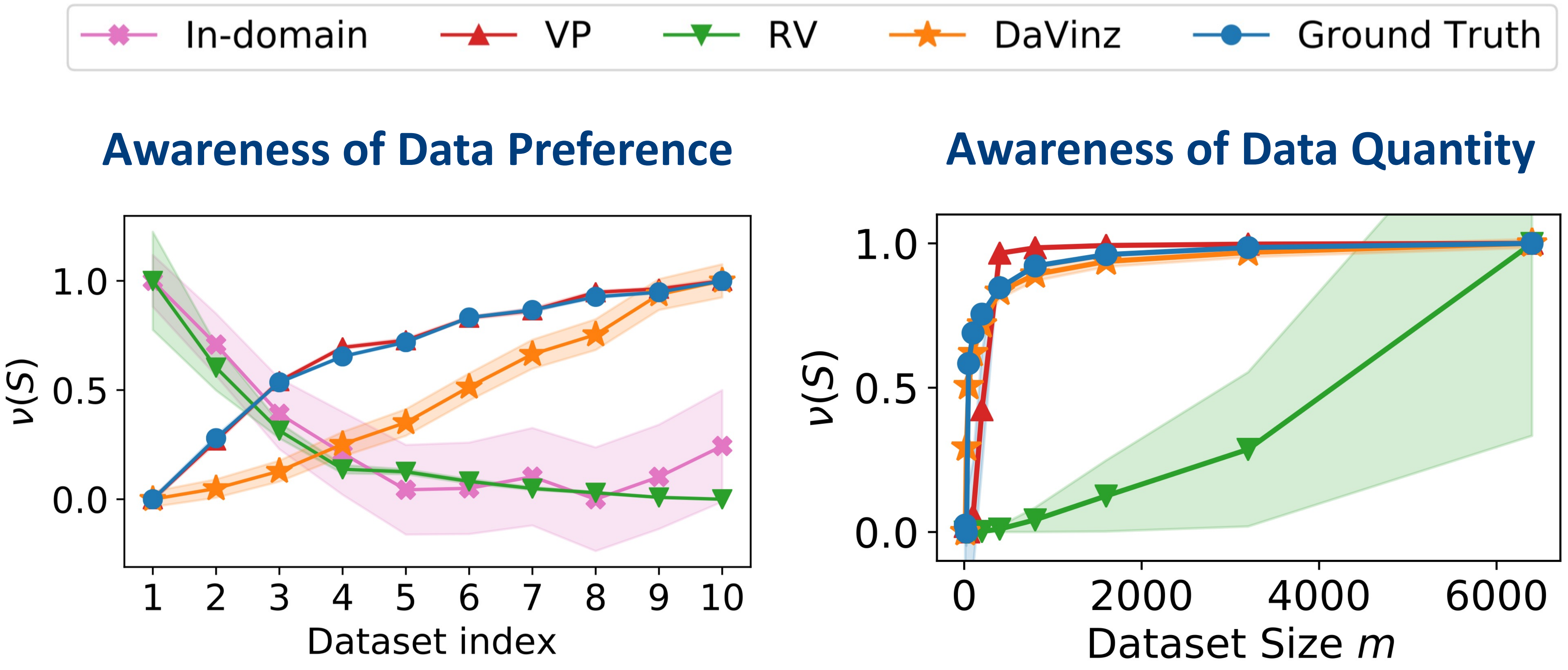
Ours training-free

Experiments

Effective and efficient DAVINZ (regression)

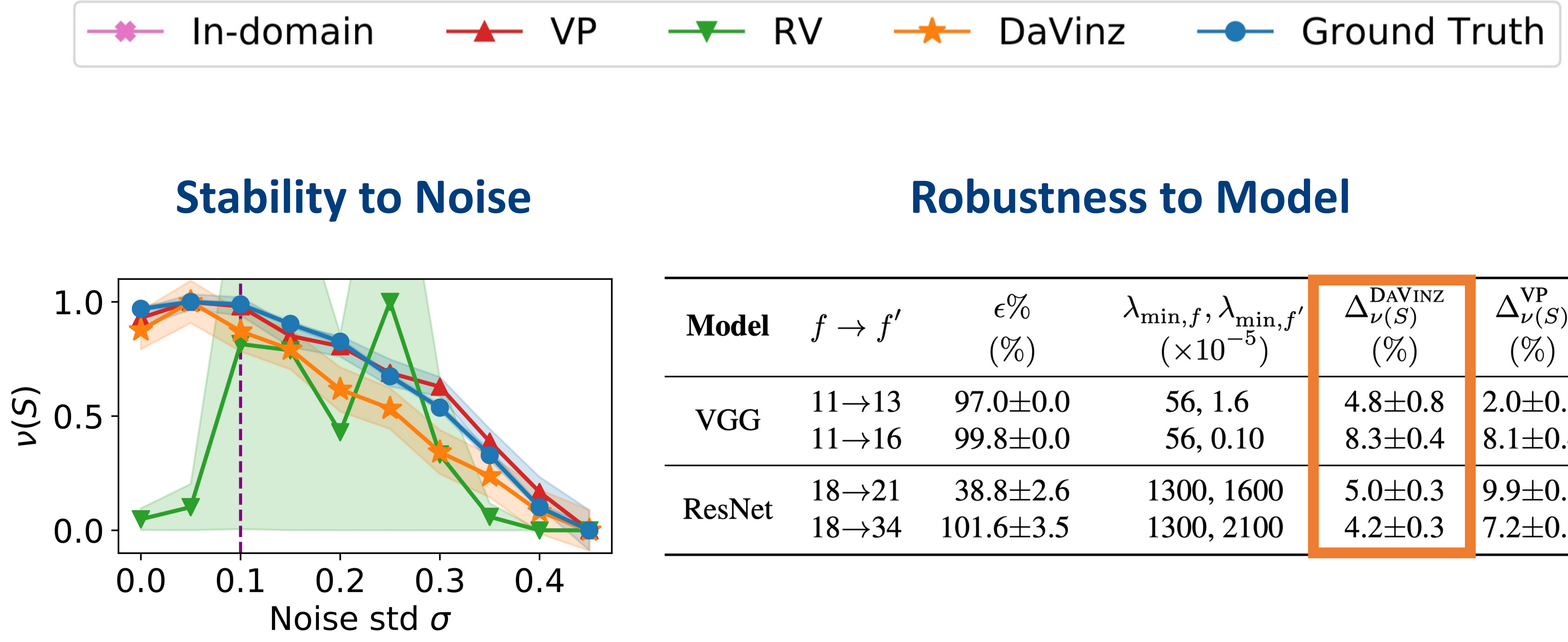
Method	Model	Ising Physical Model Dataset		
		Pearson	Spearman	Cost (Min.)
VP	MLP10	0.998±0.001	0.978±0.007	17.1
	CNN8	0.317±0.169	0.273±0.137	34.4
IF	MLP10	0.095±0.250	−0.006±0.072	1.9
	CNN8	0.189±0.142	0.001±0.124	4.1
RV	MLP10	0.727±0.231	0.699±0.182	2.0
	CNN8	0.805±0.009	0.818±0.041	4.1
DAVINZ	MLP10	0.994±0.001	0.905±0.018	**1.7**
	CNN8	0.823±0.003	0.702±0.063	**2.0**

Theoretical properties of DAVINZ



Theoretical properties of DAVINZ

Stability to Noise



Robustness to Model

Model	$f \to f'$	$\epsilon\%$ (%)	$\lambda_{\min,f}, \lambda_{\min,f'}$ ($\times 10^{-5}$)	$\Delta^{\text{DAVINZ}}_{\nu(S)}$ (%)	$\Delta^{\text{VP}}_{\nu(S)}$ (%)
VGG	11→13	97.0±0.0	56, 1.6	4.8±0.8	2.0±0.2
	11→16	99.8±0.0	56, 0.10	8.3±0.4	8.1±0.4
ResNet	18→21	38.8±2.6	1300, 1600	5.0±0.3	9.9±0.3
	18→34	101.6±3.5	1300, 2100	4.2±0.3	7.2±0.9

Conclusion

- A training-free method for efficient and trustworthy data valuation in complex DNNs
  - Derived a *domain-aware generalization bound* for DNNs using the NTK theory
  - Used the bound as the utility function to design a *training-free data valuation method*
  - Proved four desirable *theoretical properties* enjoyed by this method
- Applications: Enables large-scale SV calculation, data summarization, etc.