Performance Scaling via Optimal Transport: Enabling Data Selection from Partially Revealed Sources

Feiyang Kang [*1] Hoang Anh Just [*1] Anit Kumar Sahu [2] Ruoxi Jia [1]

Abstract

Data selection has been studied in settings where all samples from prospective sources are fully revealed, however, in practical data exchange scenarios, data providers often reveal only a limited subset of samples before an acquisition decision is made. We propose a novel framework `projektor` , which predicts model performance and supports data selection decisions based on partial samples of prospective data sources. We first leverage the Optimal Transport distance to predict the model's performance for any data mixture within the range of disclosed data sizes; then, we extrapolate the performance to larger undisclosed data sizes based on a novel parameter-free mapping technique inspired by neural scaling laws. We further derive an efficient gradient method to select data sources based on projected model performance. Evaluation over diverse applications demonstrates that `projektor` significantly outperforms existing performance scaling approaches in both prediction accuracy and computational costs.

1. Introduction

The choice of training data is one of the most crucial components when it comes to extracting the best performance out of a model. Since data is typically acquired from various sources (e.g., different organizations or vendors), machine learning practitioners often encounter a central question: *how to select and combine samples from these data sources?* Although data selection has been extensively studied in the literature related to active learning (Settles, 2012), coreset selection (Guo et al., 2022), and data valuation (Jia et al., 2019b; Ghorbani & Zou, 2019; Pruthi et al., 2020; Yan & Procaccia, 2021; Just et al., 2023), most techniques are designed for a *fully-observable* setting where all data sources are fully revealed to the model developer.

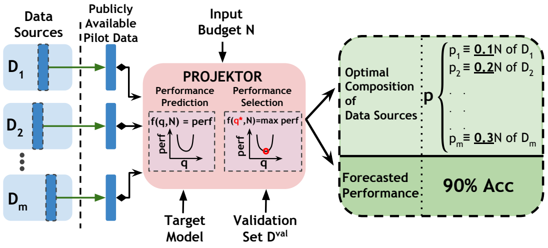

Figure 1. Overview of `projektor` , predicting performance on larger data scales and finding optimal data source composition from limited pilot data.

The core ideas behind these techniques are to compare the relative importance of different data points or enumerate possible combinations of data points, all of which require complete knowledge of the entire collection of data points. While these methods have shown promising results, their practical applications in real-world scenarios are limited due to a significant gap: the acquisition decision-making processes require knowledge of the entire data sets, while data owners may only reveal limited samples before an acquisition decision is made (e.g., (Dat; Inn) provide the examples in real-world data markets).

2. Methods

For m prospective **data providers** with datasets (*data sources*) $D_1^{\text{all}}, \ldots, D_m^{\text{all}}$, denote the public subset of each data source as D_i^{pi}. Each provider i, upon accepting the *purchasing order* for acquiring n_i samples ($n_i \leq \bar{N}_i$), will randomly sample a subset S_i of size n_i from D_i^{all} and return the subset to the requester.[1] Consider a **data collector** with a validation set D^{val}, who would like to acquire samples from the providers to train a model $\mathcal{A}$ with performance metric $\mathcal{L}$, where *acquisition decisions must be made based only on the pilot datasets*. Given a selection budget of N samples and a mixing ratio of data sources $\mathbf{p} = \{p_1, \ldots, p_m\}$, denote the selected dataset by $\mathcal{D}(N, \mathbf{p}) = S_1 \cup \cdots \cup S_m$ where $|S_i| = p_i N$. Consider a typical data acquisition goal where the collector seeks to maximize the resulting model performance by strategically choosing the mixing ratio $\mathbf{p}$ of m data sources at a *pre-specified* selection budget $N_s \leq \sum_{i=1}^{m} \bar{N}_i$–that is, $\max_{\mathbf{p}} \mathcal{L}(\mathcal{A}(\mathcal{D}(N_s, \mathbf{p})), D^{\text{val}})$.

[*] Equal contribution [1] Bradley Department of Electrical and Computer Engineering, Virginia Tech, Blacksburg VA, USA [2] Amazon Alexa AI, Seattle WA, USA. Correspondence to: Feiyang Kang <fyk@vt.edu>.

Appears in *Data-centric Machine Learning Research (DMLR) Workshop at 40th International Conference on Machine Learning*, Honolulu HI, USA. July 29, 2023. Full paper at (Kang et al., 2023)

[1] Assume each provider *honestly* provides requested samples.

`projektor` : prediction, projection, and selection

Optimal Transport (OT) is a metric for measuring the discrepancy between probability distributions (Villani, 2009) with advantageous analytical properties (Genevay et al., 2018; Feydy et al., 2019). Given probability measures μ_t, μ_v over the space $\mathcal{Z}$, the OT distance is defined as $\mathrm{OT}(\mu_t, \mu_v) := \min_{\pi \in \Pi(\mu_t, \mu_v)} \int_{\mathcal{Z}^2} \mathcal{C}(z, z') d\pi(z, z')$. Inspired by the theoretical results that the upper bound on the difference between training loss and validation loss can be tightly bounded by an affine transformation of the OT distance (Edwards, 2011; Just et al., 2023), our first proposed approach is to directly estimate this transformation by empirically fitting data distances to model performance and then the estimated transformation can be used for predicting the model performance for different data mixtures, which is

$$\hat{\mathcal{L}}\left(\mathcal{A}(\mathcal{D}(N, \mathbf{p})), D^{\text{val}}\right) = a_1 \cdot \mathrm{OT}\left(\mathcal{D}(N, \mathbf{p}), D^{\text{val}}\right) + a_0, \quad (1)$$

where a_0, a_1 can be estimated through least-square fitting. We refer to it as *center-scaling* (`projektor/CS`). a_1 can be considered an *empirical estimate of the Lipschitz constant*. `projektor/CS` has only two parameters to be estimated, which brings an important benefit of efficiency.

For samples from different data sources (i.e., data lying in different manifolds of the input space), Lipschitz constant of the model along the combined manifold may vary with the mixing ratio. Hence, we supplement `projektor/CS` with simple nonlinear terms to characterize the dependence on each data source, leading to the *pseudo-quadratic* (`projektor/PQ`) method, which is

$$\begin{aligned}\hat{\mathcal{L}}(\mathcal{A}(\mathcal{D}(N, \mathbf{p})), D^{\text{val}}) = \sum_{i=1}^{m} (b_2^i \cdot p_i^2 + b_1^i \cdot p_i + b_0) \cdot \mathrm{OT}(\mathcal{D}(N, \mathbf{p}), D^{\text{val}}) \\ + \sum_{i=1}^{m} (c_2^i \cdot p_i^2 + c_1^i \cdot p_i + c_0),\end{aligned} \quad (2)$$

Then, we need to project these predictions onto the target data scales. Neural scaling laws showcase the predictability of empirical performance with respect to the size of the training dataset, where it typically follows a log-linear scaling relationship as $\mathbb{E}_V[\mathcal{L}(\mathcal{A}(\mathcal{D}(N, \mathbf{p})); D^{\text{val}})] \approx -\alpha \log(N) + C$ where α and C are some constants (Kaplan et al., 2020). Yet, model performance for different data mixtures $\mathbf{p}$ scales with different rates (Bahri et al., 2021). With the performance prediction tools proposed above, one can directly predict model performance of any data mixture at the scale the tool has been fitted. Thus, by performing the fitting process at different small scales for once, for any desired data mixture, we can directly fit the neural scaling laws for this particular distribution and project it onto larger data scales, which is

Theorem 1. *Consider log-linear performance scaling relationship depending on both data size N and data composition $\mathbf{p}$ given as $\mathbb{E}_V[\mathcal{L}(\mathcal{A}(\mathcal{D}(N, \mathbf{p})); D^{val})] = -\alpha(\mathbf{p}) \log(N) + C(\mathbf{p})$. Assume one has completed the fitting of the performance predictor on two different scales $N_0 < N_1$, which gives $\hat{\mathcal{L}}(\mathcal{A}(\mathcal{D}(N_0, \mathbf{p})); D^{val})$ and $\hat{\mathcal{L}}(\mathcal{A}(\mathcal{D}(N_1, \mathbf{p})); D^{val})$ for all data mixtures $\mathbf{p}$. Then, the model performance $\hat{\mathcal{L}}(\mathcal{A}(\mathcal{D}(N, \mathbf{p})); D^{val})$ for any data mixture $\mathbf{p}$ at any data scale N can be predicted as*

$$\begin{aligned}\hat{\mathcal{L}}(\mathcal{A}(\mathcal{D}(N, \mathbf{p}); D^{val}) = \left(\log \frac{N_1}{N_0}\right)^{-1} \Big[\log \frac{N}{N_0} \hat{\mathcal{L}}(\mathcal{A}(\mathcal{D}(N_1, \mathbf{p})); D^{val}) \\ - \log \frac{N}{N_1} \hat{\mathcal{L}}(\mathcal{A}(\mathcal{D}(N_0, \mathbf{p}); D^{val})\Big]\end{aligned} \quad (3)$$

Further, we expect the predictions to support determining the optimal data acquisition strategy $\mathbf{p}^* = \arg\max_{\mathbf{p}} \hat{\mathcal{L}}(\mathcal{A}(\mathcal{D}(N_s, \mathbf{p})), D^{\text{val}})$. These problems are convex and differentiable and thus can be solved effectively via gradient methods, where *calibrated gradient* of OT from (Just et al., 2023) allows almost free gradient c calculation.

3. Empirical Results

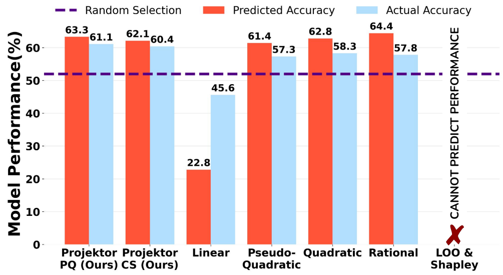

Figure 2. Performance projection vs. actual model performance of selected *optimal* mixture ratios onto 50K ImageNet-100 from 10K samples. Please refer to Appendix B.2 for baseline details.

Performance projection onto larger data scale N. In this experiment, we assume a more realistic setting, where data sources may contain some portion of mislabeled data due to labeling errors (Karimi et al., 2020). Given 3 data sources formed by sampling CIFAR-10, each of which releases a pilot dataset of size 1K, we project performance for various mixing ratios onto larger data sizes, i.e. 2K, 5K, 7K, and 10K. We then measure mean absolute errors across all data scales. We observe that `projektor` achieves the best projection performance compared to all baseline methods. `projektor/PQ` achieves the lowest MAE score below 2% and `projektor/CS` is slightly above 2%. The improved performance of our method can be attributed to the incorporation of actual data distance computation, which allows for a more accurate dataset representation (e.g. mislabeled information), whereas baseline methods completely neglect this crucial information. `projektor` not only excels at projecting performance but also at detecting irregularity in data sources.

Optimal mixing ratio with performance projection. We demonstrate improving model performance by choosing training data strategically. We consider a problem of finding the best mixing ratio of 50K samples from 3 data sources (with $10K$ pilot data each) to train ResNet-50 on ImageNet-100. Using our method, we can find a mixing ratio that achieves the best performance among all baseline methods with 2-3% accuracy improvement over the best baseline. Then, we predict performance for selected mixing ratios and observe that `projektor` achieves the lowest prediction error (Fig. 2). Unlike baseline methods that assume the same optimal composition for all data scales, our method finds the optimal composition specific to each data scale.

For additional visualization of our results, we refer the readers to Fig. 4 and Fig. 5 in App. B.

References

Datarade. https://datarade.ai/. 1

Innodata. https://innodata.com/ai-data-marketplace/. 1

Bahri, Y., Dyer, E., Kaplan, J., Lee, J., and Sharma, U. Explaining neural scaling laws. *arXiv preprint arXiv:2102.06701*, 2021. 2

Edwards, D. A. On the kantorovich–rubinstein theorem. *Expositiones Mathematicae*, 29(4):387–398, 2011. 2

Feydy, J., Séjourné, T., Vialard, F.-X., Amari, S.-i., Trouvé, A., and Peyré, G. Interpolating between optimal transport and mmd using sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 2681–2690. PMLR, 2019. 2

Genevay, A., Peyré, G., and Cuturi, M. Learning generative models with sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pp. 1608–1617. PMLR, 2018. 2

Ghorbani, A. and Zou, J. Data shapley: Equitable valuation of data for machine learning. In *International Conference on Machine Learning*, pp. 2242–2251. PMLR, 2019. 1

Guo, C., Zhao, B., and Bai, Y. Deepcore: A comprehensive library for coreset selection in deep learning. In *Database and Expert Systems Applications: 33rd International Conference, DEXA 2022, Vienna, Austria, August 22–24, 2022, Proceedings, Part I*, pp. 181–195. Springer, 2022. 1

Hashimoto, T. Model performance scaling with multiple data sources. In *International Conference on Machine Learning*, pp. 4107–4116. PMLR, 2021. 5

Jia, R., Dao, D., Wang, B., Hubis, F. A., Gurel, N. M., Li, B., Zhang, C., Spanos, C. J., and Song, D. Efficient task-specific data valuation for nearest neighbor algorithms. *arXiv preprint arXiv:1908.08619*, 2019a. 4

Jia, R., Dao, D., Wang, B., Hubis, F. A., Hynes, N., Gürel, N. M., Li, B., Zhang, C., Song, D., and Spanos, C. J. Towards efficient data valuation based on the shapley value. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1167–1176. PMLR, 2019b. 1

Just, H. A., Kang, F., Wang, T., Zeng, Y., Ko, M., Jin, M., and Jia, R. Lava: Data valuation without pre-specified learning algorithms. In *11th International Conference on Learning Representations, ICLR*, pp. to appear, 2023. 1, 2

Kang, F., Just, H. A., Sahu, A. K., and Jia, R. Performance scaling via optimal transport: Enabling data selection from partially revealed sources. *arXiv preprint arXiv:2307.02460*, 2023. 1

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., and Amodei, D. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020. 2

Karimi, D., Dou, H., Warfield, S. K., and Gholipour, A. Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. *Medical Image Analysis*, 65:101759, 2020. 2

Pruthi, G., Liu, F., Kale, S., and Sundararajan, M. Estimating training data influence by tracing gradient descent. *Advances in Neural Information Processing Systems*, 33: 19920–19930, 2020. 1

Settles, B. Active learning: Synthesis lectures on artificial intelligence and machine learning. *Long Island, NY: Morgan & Clay Pool*, 10:S00429ED1V01Y201207AIM018, 2012. 1

Villani, C. *Optimal transport: old and new*, volume 338. Springer, 2009. 2

Yan, T. and Procaccia, A. D. If you like shapley then you'll love the core. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 5751–5759, 2021. 1

A. `projektor` : Framework and Algorithms

A.1. `projektor` Pipeline

I. Training Data Preparation. For data distance-based performance prediction, the first step of the pipeline is to prepare "training data" (i.e., data mixture-model performance pairs) to fit parameters of Equations 1 and 2. The data consists of different data source compositions for some given data scales N_0, N_1, where for each composition $\mathbf{p}$, we compute the OT distance of the composed dataset $\mathcal{D}(N_1, \mathbf{p})$ to the validation data and then train a model on $\mathcal{D}(N_1, \mathbf{p})$ to get the actual model performance. For simplicity, we select compositions through grid search. This step is represented in Algorithm 1 from lines 4-9 and is the first (I) step in the pipeline Figure 3.

II. Fitting Predictor Function. Once the training data is prepared, we proceed to fit our function in Eqs. 1 and 2. This step is shown in lines 10-11 in Algorithm 1 and is the second (II) step of the pipeline Figure 3. Then, with the performance predictors fitted for data scales N_0 and N_1, we move on to the inference stage, where we can perform 2 tasks: performance projection and data source selection.

III. Two-Stage Performance Projection. For performance projection, we project the performance to any data size N for any mixing ratio $\mathbf{p}$ in two stages. **(1)** We predict the performance given the mixing ratio $\mathbf{p}$ at data scales N_0 and N_1. **(2)** We use Eq. 3 to project performance prediction to any data size N. This process is represented in line 12 of Algorithm 1 and is the third (III) step of the pipeline Figure 3.

IV. Optimal Data Source Selection. For optimal data source selection, we solve an optimization problem through gradient descent. The gradient computation uses parameters of the fitted functions from step II and the process terminates when the mixing ratio converges. This process is represented in line 12 of Algorithm 2 and is the fourth (IV) step of pipeline Figure 3.

Algorithm 1: `projektor` performance predictor.

```
In  : Pilot Datasets $D_1^{pi}, D_2^{pi}, \ldots, D_m^{pi}$; Query Data Budget $N$; Query Mixing Ratio $\mathbf{p}$;
      0-Data Scale Size $N_0$; 1-Data Scale Size $N_1$; Learning Algorithm $\mathcal{A}$; Performance Metric
      Function $\mathcal{L}(\cdot, D^{val})$; OT Distance Function $OT(\cdot, D^{val})$.
Out : Projected Model Performance $\rightarrow [0, 1]$.
1  P $\leftarrow$ Generate mixing ratios
2  DT$_0$, DT$_1$ $\leftarrow$ Initialize empty lists to store OT distances
3  L$_0$, L$_1$ $\leftarrow$ Initialize empty lists to store performance values
4  for Mixing Ratio $\mathbf{p_i}$ in P do
5      $S_0, S_1 = \mathcal{D}(N_0, \mathbf{p_i}), \mathcal{D}(N_1, \mathbf{p_i})$ newly composed datasets of size $N_0, N_1$
6      DT$_0$ $\leftarrow$ append $OT(S_0, D^{val})$ Optimal Transport distance between $S_0$ and $D^{val}$
7      DT$_1$ $\leftarrow$ append $OT(S_1, D^{val})$ Optimal Transport distance between $S_1$ and $D^{val}$
8      L$_0$ $\leftarrow$ append $\mathcal{L}(\mathcal{A}(S_0), D^{val})$ Performance of a model trained on $S_0$
9      L$_1$ $\leftarrow$ append $\mathcal{L}(\mathcal{A}(S_1), D^{val})$ Performance of a model trained on $S_1$
10 $\hat{\mathcal{L}}(\mathcal{A}(\mathcal{D}(N_0, \cdot)), D^{\text{val}})$ $\leftarrow$ Fit the function from Eq. 2 with ((P, DT$_0$), L$_0$)
11 $\hat{\mathcal{L}}(\mathcal{A}(\mathcal{D}(N_1, \cdot)), D^{\text{val}})$ $\leftarrow$ Fit the function from Eq. 2 with ((P, DT$_1$), L$_1$)
12 $\hat{\mathcal{L}}(\mathcal{A}(\mathcal{D}(N, \mathbf{p})); D^{\text{val}})$ $\leftarrow$ Project performance by substituting $\hat{\mathcal{L}}(\mathcal{A}(\mathcal{D}(N_0, \mathbf{p})), D^{\text{val}})$ and $\hat{\mathcal{L}}(\mathcal{A}(\mathcal{D}(N_1, \mathbf{p})), D^{\text{val}})$
     into Eq. 3
13 return $\hat{\mathcal{L}}(\mathcal{A}(\mathcal{D}(N, \mathbf{p}); D^{\text{val}})$
```

B. Experiment Details and Additional Results

B.1. Details on Baseline Methods

For N samples from m data sources with a mixing ratio $\mathbf{p} = \{p_1, \ldots, p_m\}$, we consider the following baselines:

Linear: $\hat{\mathcal{L}}(\mathcal{A}(\mathcal{D}(N, \mathbf{p}); D^{\text{val}}) := \mathbf{a}'\mathbf{p} + b\log(N) + c$, where $\mathbf{a} = \{a_0, a_1, ..., a_m\}$, b, and c are coefficients to be fitted.

Leave-one-out (LOO) *and* ***Shapley*** *can be considered special cases for* ***Linear****, where the coefficients are calculated as the marginal contribution of the data source (****LOO****) or its averaged contribution to different combinations of other data sources (****Shapley****) (Jia et al., 2019a), as opposed to the least-square fitting as in* ***Linear****.*

Pseudo-quadratic: $\hat{\mathcal{L}}(\mathcal{A}(\mathcal{D}(N, \mathbf{p}); D^{\text{val}}) := \sum_{i=1}^{m}(c_2^i \cdot p_i^2 + c_1^i \cdot p_i + c_0) + b\log(N)$

Algorithm 2: Optimal data source composition $\mathbf{p}^*$ Update.

```
In  : Pilot Datasets $D_1^{pi}, D_2^{pi}, \dots, D_m^{pi}$; Query Data Budget $N$; Query Mixing Ratio $\mathbf{p}$;
      0-Data Scale Size $N_0$; 1-Data Scale Size $N_1$; Trained projektor Models with $N_0$ and $N_1$ Data Budgets:
      $f_0, f_1$; Enquired Data Budget: $N$; OT Distance Function $OT(\cdot, D^{val})$ Validation Set: $D^{val}$.
Out : Optimal data source composition $\mathbf{p}^*$.
p ← Initialize Random Data Source Composition
while p not converged do
    $S_0, S_1 = \mathcal{D}(N_0, \mathbf{p_i}), \mathcal{D}(N_1, \mathbf{p_i})$ newly composed datasets of size $N_0, N_1$
    gradient ← Compute gradient w.r.t. p
    p ← Update composition p with the gradient update
return p
```

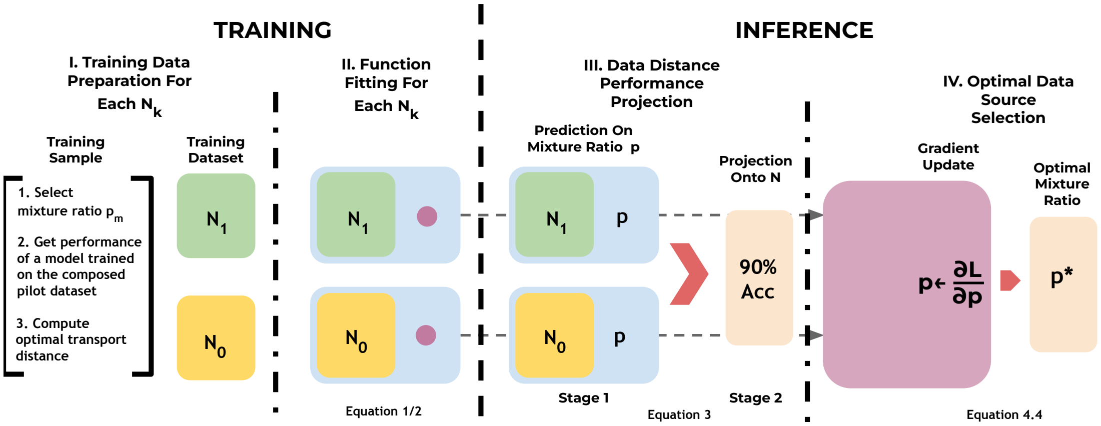
TRAINING
INFERENCE
I. Training Data Preparation For Each N_k
II. Function Fitting For Each N_k
III. Data Distance Performance Projection
IV. Optimal Data Source Selection
Training Sample
Training Dataset
1. Select mixture ratio p_m
2. Get performance of a model trained on the composed pilot dataset
3. Compute optimal transport distance
N_1
N_0
Prediction On Mixture Ratio p
p
Projection Onto N
90% Acc
Gradient Update
$p \leftarrow \frac{\partial L}{\partial p}$
Optimal Mixture Ratio
p*
Equation 1/2
Stage 1
Equation 3
Stage 2
Equation 4.4

Figure 3. `projektor`

Quadratic: $\hat{\mathcal{L}}(\mathcal{A}(\mathcal{D}(N, \mathbf{p}); D^{\text{val}}) := \sum_{i=1}^{m}(c_2^i \cdot p_i^2 + c_1^i \cdot p_i + c_0) + \sum_{i=1}^{m}\sum_{j=1}^{i}(c_3^{ij} \cdot p_i p_j) + b\log(N)$

Rational: $\hat{\mathcal{L}}(\mathcal{A}(\mathcal{D}(N, \mathbf{p}); D^{\text{val}}) := \sum_{i=1}^{m}\left(\sum_{j=1}^{m} c^{ij} \cdot p_j\right)^{-1} + b\log(N)$

*We fit the **Rational** baseline according to the setup detailed in (Hashimoto, 2021) and to our best effort. Originally, the method is intended for predicting log loss, whereas in our case, we aim to predict model accuracy. Thus, we replaced the log loss with* $\log(1 - accuracy)$ *for the prediction target.*

B.2. Evaluation Metric

We use mean absolute error (**MAE**) to evaluate the performance of our method by calculating the absolute difference between the predicted and the actual accuracy.

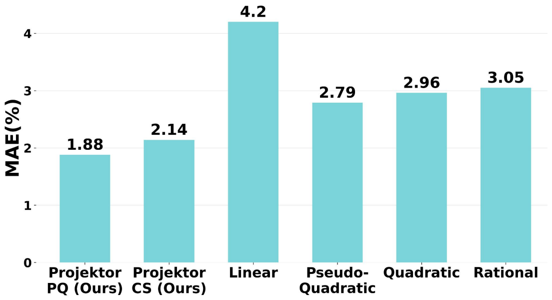

Figure 4. Performance projections from 1K CIFAR-10 samples across various mixing ratios and larger data scales: 2K, 5K, 7K, 10K. Comparison between `projektor` and baselines.

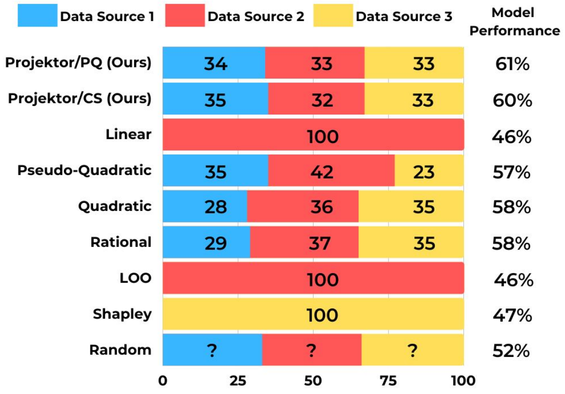

Figure 5. Optimal data source composition selection for 50K ImageNet-100 from 10K samples and actual model performance.