projektor | 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]

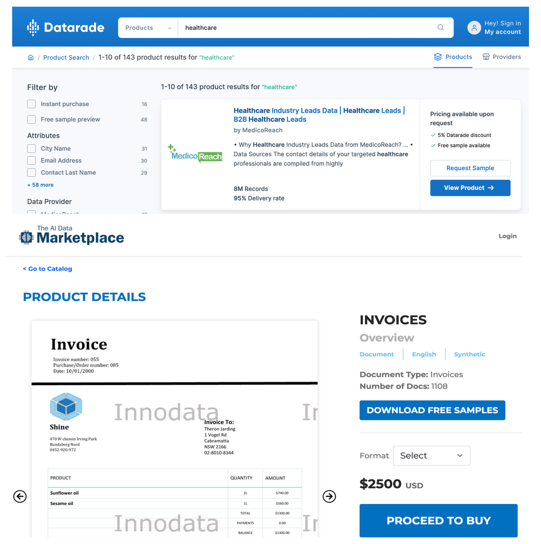
[1]Virginia Tech [2]Amazon Alexa AI

NEURAL INFORMATION PROCESSING SYSTEMS
NeurIPS | 2023

We approach practical data collection scenarios with multiple sources, where acquisition plans need to be made with only small samples. We propose a handy toolkit, `projektor`, that predicts model performance, projects it onto larger scales, and optimizes over predictions.

What's the problem?

"data is the new oil"–the choice of training data is crucial for extracting the best performance out of a model.



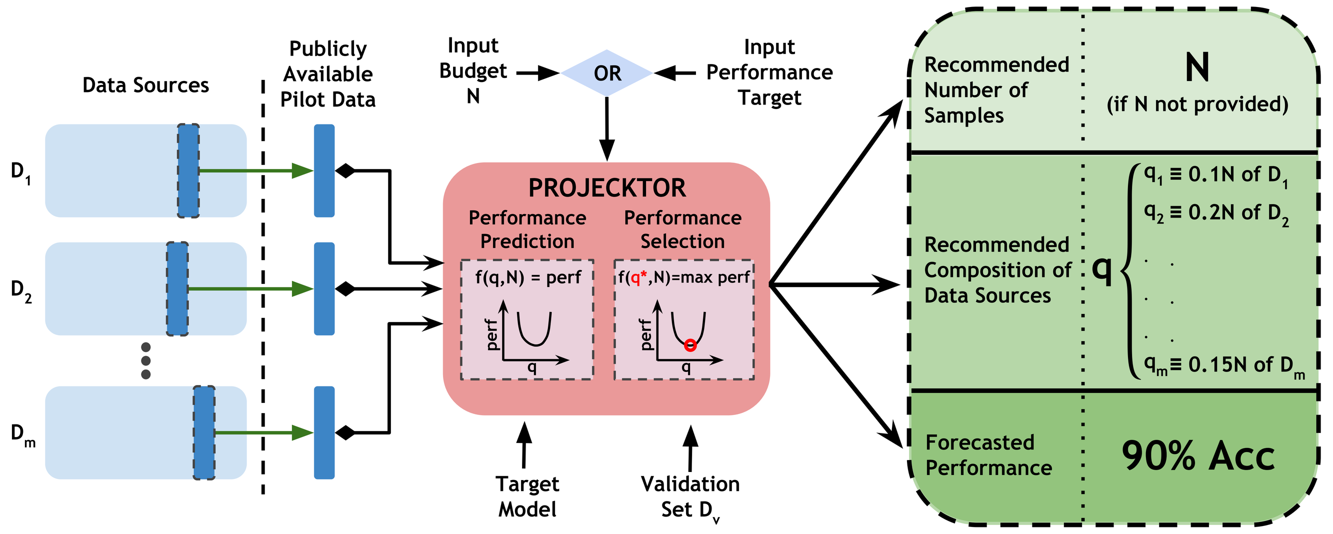
Data is typically acquired from **various sources**, such as different organizations or vendors (e.g., data marketplace)

in practical scenarios, data providers often reveal **only a limited subset** of samples before an acquisition decision is made.

How to select and combine samples from these data sources?

Our Framework

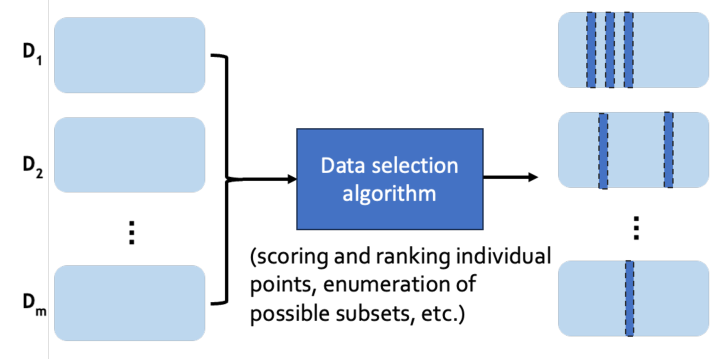
strategic data selection in **partially observable** settings, where only limited samples of data sources (pilot datasets) are accessible.



→ The goal is to determine an **optimal allocation** of the selection budget to each source, based only on pilot datasets, such that the model trained on the mixture of collected data achieves the best result on given objectives.

Limitations of Past Work

Most assumes **complete** access to potential data sources (coreset selection, active learning, data valuation, etc.)



+: Fine-grained selection

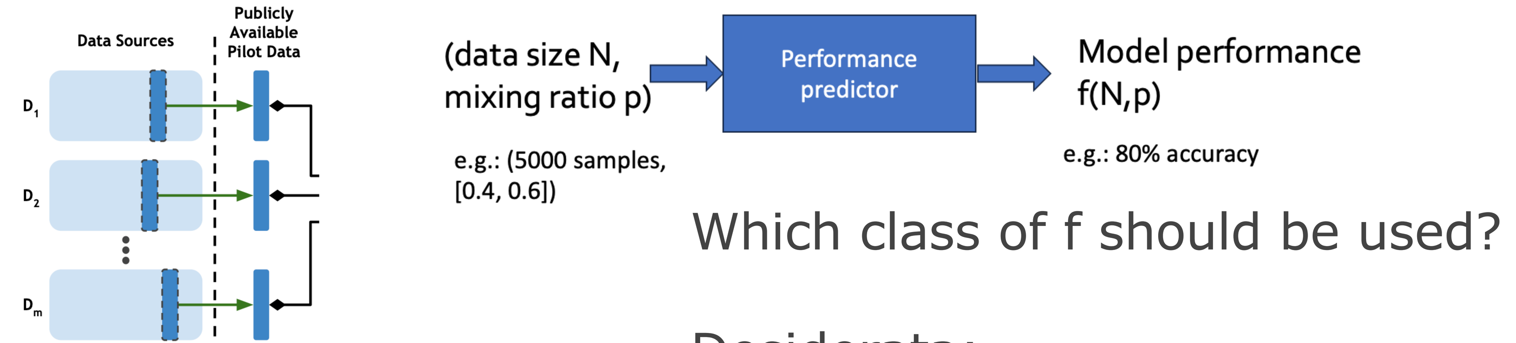
-: mismatch with many practical scenarios

-: scalability issues

Without complete access of data, we cannot directly evaluate a plan

Our Approach: projektor

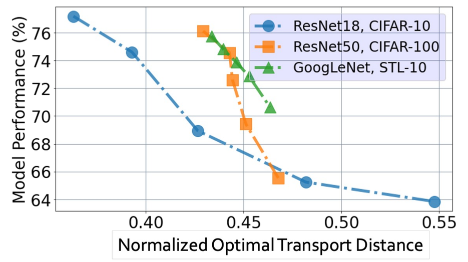
Key idea: learn a performance predictor



Which class of f should be used?

Desiderata:

- Accurate extrapolation to large N
- Easy to optimize over p



Intuition: The **more relevant** training data is to the validation, the **higher model performance**

Paper & Repository

Paper. openreview.net/pdf?id=quMBEd27x9

Code. github.com/ruoxi-jia-group/projektor

Our Approach: projektor

Our approach: performance scaling via optimal transport

Stage 1: Performance prediction **at any p but small N**

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

Validation set

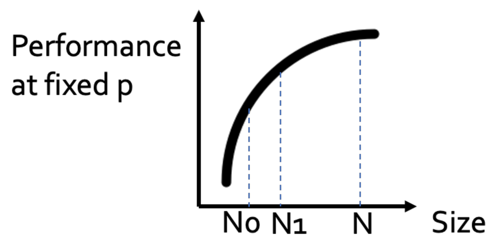
Performance of the model trained on D(N,p) and evaluated on D^{val}

Training set of size N and mixing ratio p

Stage 2: Parameter-free projection to **larger N**

Consider $\mathbb{E}_V[\mathcal{L}(\mathcal{A}(\mathcal{D}(N,\mathbf{p})); D^{val})] = -\alpha(\mathbf{p})\log(N) + C(\mathbf{p})$

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



Stage 3: Selection

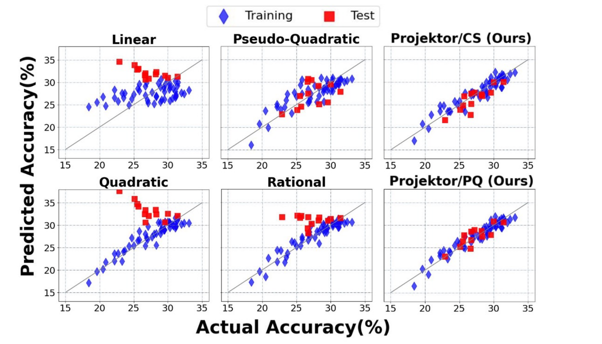
$$\mathbf{p}^* = \arg\max_{\mathbf{p}} \hat{\mathcal{L}}(\mathcal{A}(\mathcal{D}(N_s,\mathbf{p})), D^{\text{val}})$$

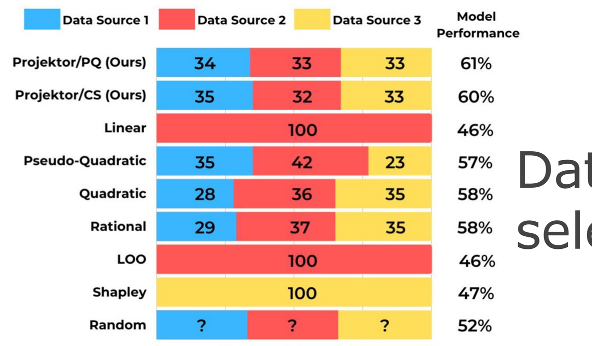
Optimization is solved via gradient descent

$$\mathbf{p}^{t+1} \leftarrow \mathbf{p}^t + d^t \cdot \left.\frac{\partial\hat{\mathcal{L}}(\mathcal{A}(\mathcal{D}(N_s,\mathbf{p})), D^{\text{val}})}{\partial\mathbf{p}}\right|_{\mathbf{p}=\mathbf{p}^t}$$

Applications

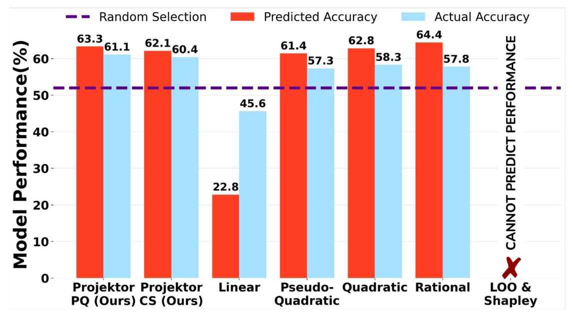
Evaluation: Performance prediction at unseen mixing ratio (extrapolation on p)

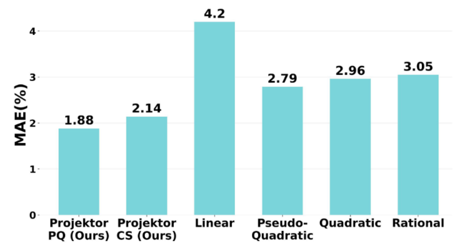




Data mixture ratio selection for ImageNet-100 selection for 50K budget from 10K samples.

Projected performance for the selected mixture ratio (from above) and the actual performance.





Performance projection from 1K samples to larger data scales (2-10K).