Risk-Aware Data Auditing for Resource-Efficient Learning under Distribution Shift

Data-CAPM: A Group-Level Factor Decomposition

Yingjie Bai
School of Computer Science
The University of Sydney, Australia
ybai0439@uni.sydney.edu.au

Weiming Zhi
School of Computer Science
The University of Sydney, Australia
Australian Centre for Robotics
The University of Sydney, Australia

Abstract

Resource-efficient learning is often constrained at the data level: labels, source acquisition, retraining, and manual audit budgets are finite. Yet scalar data valuation can allocate this budget poorly under distribution shift, because a high-return data group may carry shortcut-driven systematic exposure rather than robust contribution. We introduce Data-CAPM, a diagnostic framework for risk-aware group data auditing. For each evaluation scenario, a group return is the utility drop from leaving that group out; across scenarios, Data-CAPM decomposes returns into mean return, beta, alpha, residual risk, and Alpha/Risk. This profile helps decide whether a data group provides robust idiosyncratic value, redundant utility, volatile contribution, or shortcut-like exposure. In controlled synthetic and Adult-tabular shortcut benchmarks, the shortcut group ranks first by mean return but last by alpha and Alpha/Risk. Under the same top-k group budget, alpha-based acquisition improves robust-average balanced accuracy by 10.4 percentage points at $k = 2$, and risk-aware pruning removes the shortcut group first. Folktables/ACSIncome and RxRx1 official-embeddings pilots show smaller natural-metadata effects, illustrating CPU-friendly diagnostic separation rather than end-to-end performance superiority. The contribution is a data-level resource-efficiency audit tool, not a universal data valuation scalar.

Preprint.

1 Introduction

Resource-efficient learning for knowledge discovery includes data-level constraints: obtaining labels, integrating new data sources, retraining models, and inspecting data quality can dominate practical cost. In recommendation, healthcare, finance, social media, and scientific discovery pipelines, the question is often not whether more data exists, but which data source deserves scarce acquisition or audit budget.

Data valuation offers one answer by estimating which samples or groups are helpful, harmful, redundant, or worth acquiring. Data Shapley estimates average marginal contribution over coalitions [Ghorbani and Zou, 2019]; influence functions approximate effects on predictions or losses [Koh and Liang, 2017]; DVRL learns validation-guided sample values [Yoon et al., 2020]; scalable variants such as Beta Shapley and Data-OOB improve practicality [Kwon and Zou, 2022, 2023]. These methods are often consumed as scalar evidence: high score means acquire or retain; low score means inspect or remove.

Under distribution shift, this scalar interface can be misleading. A group may improve average validation utility because it encodes a shortcut that co-moves with easy scenarios, while hurting rare, OOD, or shifted slices. Conversely, a rare-domain group may have modest average return but high value in the scenarios where robustness matters. We therefore ask not only *how valuable is this data?*, but *what kind of value does this group provide under a fixed data budget and a chosen scenario design?*

Data-CAPM treats group valuation as a resource-aware audit problem. It decomposes each group return vector into mean return, systematic exposure beta, idiosyncratic alpha, residual risk, and Alpha/Risk. The analogy to finance is operational rather than literal: data groups are not assumed to satisfy financial-market assumptions. The purpose is to make acquisition and pruning decisions more transparent under shift.

Contributions.

1. We formulate group data valuation as a data-level resource-efficiency problem over scenario-wise return vectors.
2. We introduce Data-CAPM, a factor-style diagnostic profile for group acquisition, pruning, and inspection under distribution shift.
3. We show a scalar-valuation failure mode: a shortcut group can rank first by mean return but last by alpha and Alpha/Risk.
4. We provide operational probes showing that risk-aware diagnostics can improve robust performance under the same group budget, plus CPU-friendly natural-metadata sanity checks.

2 Data-CAPM

Let training data be partitioned into groups $D_1, \ldots, D_G$. A scenario $t \in \{1, \ldots, T\}$ may specify a random seed, model class, validation slice, domain, year, or metric. For validation utility U on V_t, the leave-one-group-out return is

$$R_{g,t} = U(f_{\text{all}}, V_t) - U(f_{-g}, V_t), \tag{1}$$

where f_{all} is trained on all groups and f_{-g} is trained without D_g. A positive return means removing g hurts utility.

The primary estimator uses a shared arithmetic market factor

$$R_{M,t} = G^{-1} \sum_{g=1}^{G} R_{g,t}, \tag{2}$$

which keeps alpha values comparable across groups. For each group,

$$\beta_g = \frac{\text{Cov}(R_g, R_M)}{\text{Var}(R_M)}, \qquad \alpha_g = \mathbb{E}[R_g] - \beta_g \mathbb{E}[R_M]. \tag{3}$$

Residual risk and Alpha/Risk are

$$\sigma_g^{\text{res}} = \text{Std}(R_g - \beta_g R_M), \qquad S_g = \frac{\alpha_g}{\sigma_g^{\text{res}} + \epsilon}. \tag{4}$$

S_g is a Sharpe-like diagnostic, labeled Alpha/Risk. We also report a beta-penalized diagnostic score,

$$\text{CAPMScore}_g = \alpha_g - \lambda \max(\beta_g, 0), \tag{5}$$

for risk-aware selection policies. The recommended use is to inspect the full profile rather than replace valuation with a single new scalar.

Exact leave-one-group-out requires $O(T(G+1))$ model trainings. This is feasible for small source/domain audits and can be parallelized. The artifact also includes engineering hooks for influence-style approximation, median/trimmed market aggregators, beta shrinkage, and KMeans auto-grouping. These are treated as sensitivity diagnostics, not as primary evidence.

3 Experimental Protocol

Synthetic shortcut benchmark. We construct six groups: core clean data ($G0$), redundant clean core data ($G1$), rare clean data ($G2$), OOD clean data ($G3$), label-noise data ($G4$), and spurious shortcut data ($G5$). Validation scenarios include IID, rare, OOD, and spurious-shift slices. The default setup uses 5 seeds, 2 model classes, and 4 validation slices, yielding 40 return observations per group.

Adult controlled shortcut benchmark. We use UCI Adult Income covariates and labels [Becker and Kohavi, 1996], then inject controlled rare/OOD/noise/shortcut mechanisms. This is mechanism validation on real tabular covariates, not natural shortcut discovery.

Natural metadata pilots. Adult workclass, Folktables/ACSIncome state domains [Ding et al., 2021], and RxRx1 official embeddings [Sypetkowski et al., 2023] use naturally defined metadata groups without injected shortcuts. These pilots are sanity checks, not fairness audits, state-quality rankings, or biological batch-quality labels.

Resource-budget probes. Acquisition selects top-k groups under a fixed group budget and evaluates robust-average balanced accuracy over rare, OOD, and spurious-shift slices. Pruning removes bottom-k groups and retrains, simulating a limited inspection/removal budget. Baselines include mean return, loss-based ranking, worst-slice return, validation-guided proxies, approximate group Shapley, and random ranking.

4 Results

4.1 A High-Return Group Can Be a Poor Data-Budget Choice

Table 1 and Figure 1 show the core synthetic result. The shortcut group G5_spurious ranks first by mean return but last by alpha and Alpha/Risk. Its beta is 6.88, indicating strong systematic exposure to the shared scenario market. After accounting for this exposure, its alpha is negative. Rare and OOD clean groups have the strongest risk-adjusted profiles despite lower mean return.

Table 1: Synthetic Data-CAPM scores. Lower rank is better.

Group	Mean	Beta	Alpha	Alpha/Risk	R_m	R_α	R_S
G3_ood_clean	0.0189	-0.2680	0.0225	0.5820	2	1	2
G2_rare_clean	0.0142	-0.5463	0.0217	0.7438	3	2	1
G0_core_clean	-0.0016	-0.1735	0.0008	0.0671	5	3	3
G4_label_noise	0.0018	0.1757	-0.0006	-0.0913	4	4	4
G1_redundant_core	-0.0025	-0.0717	-0.0015	-0.1426	6	5	5
G5_spurious	0.0512	6.8839	-0.0429	-1.4173	1	6	6

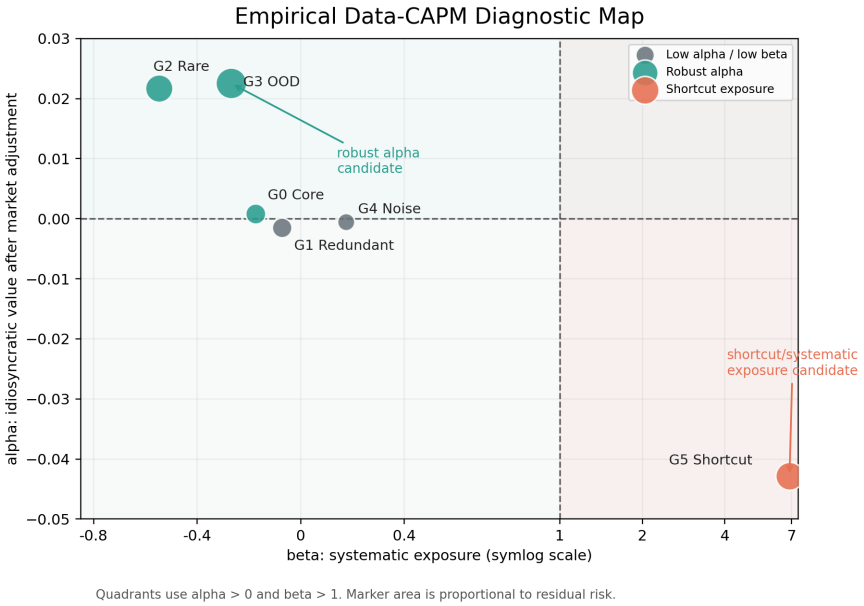

Figure 1: Alpha-beta diagnostic map on the synthetic benchmark. Marker area is proportional to residual risk.

Across 10 synthetic dataset seeds, $G2$ and $G3$ appear in the top two by alpha in every run; G5_spurious appears in the bottom two by alpha in every run and is ranked first by mean return in every run.

4.2 Operational Consequences under Fixed Group Budget

Table 2 summarizes the operational resource-budget probes. When acquisition is restricted to $k = 2$ groups, alpha-based selection chooses $G2$ and $G3$, avoids $G5$, and improves robust-average balanced accuracy by 0.1038 over mean-return top-k. Pruning by alpha removes the shortcut group first, whereas mean-return pruning removes a redundant clean group.

Table 2: Operational probes under fixed acquisition/pruning budget.

Setting	Method	k	Delta	Evidence
Synthetic acquisition	Alpha top-k	2	0.1038	selects $G2, G3$, avoids $G5$
Synthetic acquisition	Alpha/Risk top-k	2	0.1038	same top-2 as alpha
Synthetic acquisition	CAPM top-k	2	0.1038	same top-2 as alpha
Synthetic pruning	alpha	1	0.0131	removes $G5$ first
Synthetic pruning	mean return	1	0.0021	removes $G1$, not $G5$
Adult pruning	alpha	1	0.0305	removes $G5$ first
Adult pruning	mean return	1	0.0134	removes $G4$, not $G5$

On Adult with controlled shortcut intervention, the same mechanism appears on real covariates: the spurious group ranks first by mean return and last by alpha across three dataset seeds, with average beta 7.48 and average alpha -0.0388. This supports the mechanism claim while preserving the boundary that the shortcut is injected.

4.3 CPU-Friendly Natural Metadata Sanity Checks

Natural metadata pilots are smaller and more descriptive. In Folktables/ACSIncome state-domain and RxRx1 official-embeddings batch audits, absolute effects are on the order of 10^{-3}. We therefore make no end-to-end performance superiority claim. Their role is to show that Data-CAPM can

separate mean contribution, beta exposure, and idiosyncratic alpha on natural metadata groups using CPU scikit-learn models and tabular or official precomputed embedding features.

In the RxRx1 paper-minimum run, the audit uses 24 HUVEC experimental batches, 31 classes, 3 random seeds, and 24 held-out validation domains, yielding 72 return observations per group, 1,728 LOGO return entries, and 1,800 total classifier fits including full-model fits. This is a practical embeddings-first audit rather than raw-image training.

4.4 Boundary Checks

Sensitivity diagnostics show that the central exact-LOGO warning is stable under the default shared arithmetic market and mild beta shrinkage, but not under all variants. A median market can promote the shortcut group in the six-group setting, and an influence-style approximation does not reproduce the exact-LOGO shortcut ranking under non-smooth balanced accuracy. These failures are informative: alternative market constructions and approximations should be reported as sensitivity diagnostics, while exact LOGO with the shared arithmetic market remains the primary estimator in this study.

5 Limitations and Responsible Use

Diagnostic, not universal. Data-CAPM is a factor-model analogy and audit tool, not a claim that data markets satisfy financial CAPM assumptions or that one scalar score solves data valuation.

Group-level and scenario-dependent. The method audits predefined groups under chosen scenarios, metrics, and model classes. A low-alpha result under one design is an inspection signal, not an intrinsic property of a population, state, hospital, experimental batch, or user group.

Controlled evidence. The strongest empirical evidence uses constructed shortcut mechanisms. Adult covariates are real, but the shortcut is injected. Folktables/ACSIncome and RxRx1 are natural-metadata sanity checks, not proof of natural shortcut discovery.

Redundancy blindness. Exact leave-one-group-out can underestimate redundant but useful clean groups, because another similar source may substitute for the removed group. Combining Data-CAPM with Group Shapley or interaction influence estimators is a natural extension.

Human-centered interpretation. When groups correspond to people, institutions, clinical sites, or regions, low alpha or high beta must not be read as human value. It may reflect a poor evaluation environment, an inadequate model class, a utility metric that omits relevant value, or unequal opportunity to contribute. Data-CAPM is intended to improve documentation and inspection, not to narrow access to opportunity.

6 Reproducibility

The workshop artifact is CPU-oriented and uses scikit-learn models. It does not require GPU training, raw-image processing, Docker, sudo, or system package installation. The main controlled suite can be run with:

```
python run_experiment.py --paper_suite true
```

or through:

```
bash scripts/run_paper_suite.sh
```

The RxRx1 official-embeddings audit can be regenerated without downloading raw images:

```
bash experiments/rxrx1_batch_audit/scripts/\
run_rxrx1_auto_download_mac_paper_min.sh
```

The artifact writes resolved `run_config.json` files, CSV outputs, table sources, and figures. Third-party raw datasets are not redistributed; users should follow upstream terms for UCI Adult, Folktables/ACS, and RxRx1.

7 Conclusion

Data-CAPM reframes group data valuation as resource-aware data auditing under distribution shift. Its value is not a new universal scalar, but a diagnostic profile that separates average utility from systematic exposure, idiosyncratic alpha, and residual risk. This profile can prevent scarce acquisition or pruning budget from being spent on high-mean shortcut groups and can make data-level resource decisions more transparent in knowledge-discovery pipelines.

References

Barry Becker and Ronny Kohavi. Adult [dataset]. UCI Machine Learning Repository, 1996. URL https://archive.ics.uci.edu/dataset/2/adult.

Frances Ding, Moritz Hardt, John P. Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. In *Advances in Neural Information Processing Systems*, volume 34, pages 6478–6490, 2021. URL https://papers.nips.cc/paper_files/paper/2021/hash/32e54441e6382a7fbacbbbaf3c450059-Abstract.html.

Amirata Ghorbani and James Zou. Data shapley: Equitable valuation of data for machine learning. In *Proceedings of the 36th International Conference on Machine Learning*, pages 2242–2251, 2019. URL https://proceedings.mlr.press/v97/ghorbani19c.html.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1885–1894, 2017. URL http://proceedings.mlr.press/v70/koh17a.html.

Yongchan Kwon and James Zou. Beta shapley: A unified and noise-reduced data valuation framework for machine learning. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, pages 8780–8802, 2022. URL https://proceedings.mlr.press/v151/kwon22a.html.

Yongchan Kwon and James Zou. Data-OOB: Out-of-bag estimate as a simple and efficient data value. In *Proceedings of the 40th International Conference on Machine Learning*, pages 18135–18152, 2023. URL https://proceedings.mlr.press/v202/kwon23e.html.

Maciej Sypetkowski, Morteza Rezanejad, Saber Saberian, Oren Kraus, John Urbanik, James Taylor, Ben Mabey, Mason Victors, Jason Yosinski, Alborz Rezazadeh Sereshkeh, Imran Haque, and Berton Earnshaw. RxRx1: A dataset for evaluating experimental batch correction methods. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 4285–4294, 2023. URL https://openaccess.thecvf.com/content/CVPR2023W/CVMI/html/Sypetkowski_RxRx1_A_Dataset_for_Evaluating_Experimental_Batch_Correction_Methods_CVPRW_2023_paper.html.

Jinsung Yoon, Sercan O. Arik, and Tomas Pfister. Data valuation using reinforcement learning. In *Proceedings of the 37th International Conference on Machine Learning*, pages 10842–10851, 2020. URL https://proceedings.mlr.press/v119/yoon20a.html.