DSC 261

Data Valuation

- Data valuation asks which examples help a chosen task.
- How can distributional similarity guide what we keep, clean, or acquire?

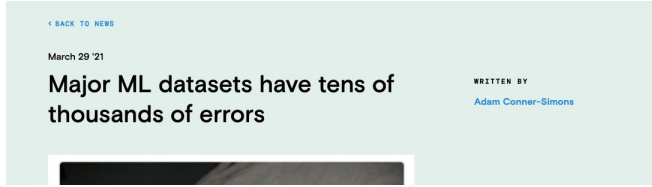
The case for trustworthy AI

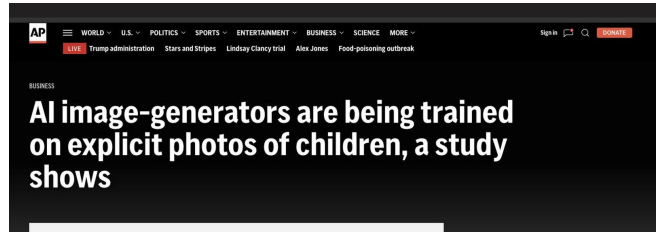
- AI increasingly shapes decisions that affect people's lives: who gets considered for a job, who receives credit, and who gets access to health care.
- Greater autonomy raises the stakes. As AI systems act with less human oversight, there are fewer opportunities to catch mistakes before they cause harm.
- We have already seen AI systems fail in consequential ways: they have favored one group over another in hiring, contributed to wrongful arrests through facial-recognition errors, fabricated legal citations, reproduced sensitive training text, and become less safe after fine-tuning.
- This motivates trustworthy AI: systems that are accurate, reliable, and robust in their intended use, while also being fair, privacy-preserving, safe, secure, transparent, and accountable.



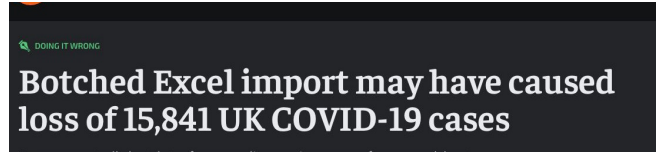
Data is a powerful point of intervention

- Data shapes what models learn: the capabilities they develop, the patterns they rely on, and the mistakes they make.
- Data problems can undermine learning and evaluation: biased samples, class imbalance, missing or corrupted values, incorrect or inconsistent labels, inconsistent formats or units, duplicates, outdated records, train–test leakage, and poisoned or harmful content.
- This lecture examines how data shapes learning and introduces principled methods to improve the data that learning systems rely on.



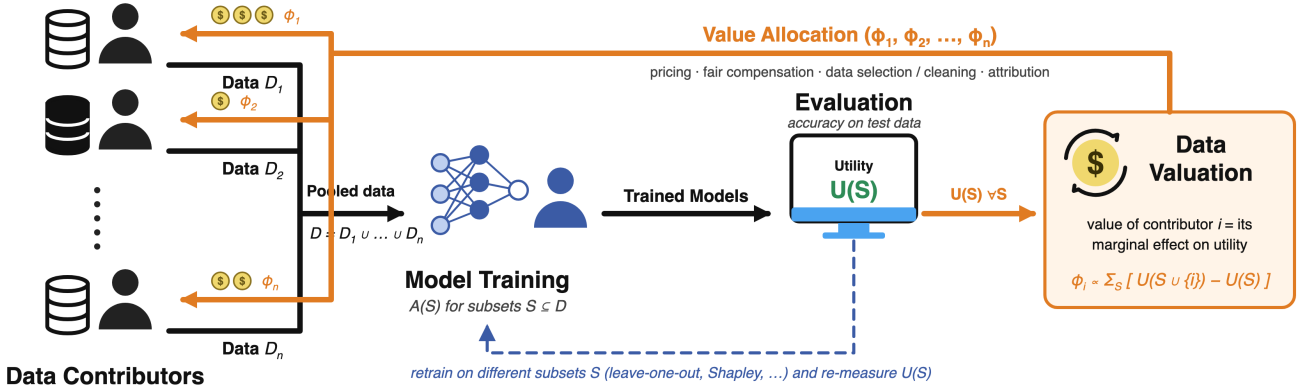






Data valuation

Data valuation assigns a score to a data contribution based on how much it helps or hurts a chosen objective.



- The unit can be one example, a batch, or a whole dataset.
- The objective sets the score: model performance or a data-level measure.
- A contribution can be negative. The same data can have different value for different tasks.

Application 1: data selection

Data selection chooses which existing data to keep for a task under a size or compute budget.

- Reduce training cost by retaining data with high estimated usefulness.
- Remove redundancy when other retained points provide similar information.
- Build task-specific subsets from a larger data pool.
- Preserve class and group coverage through explicit selection constraints.
- Update scores as the composition of the selected subset changes.

Valuation helps decide what to keep. The usefulness of a point depends on the rest of the subset.

Application 2: data cleaning and debugging

Data cleaning and debugging identify data problems and prioritize investigation and repair.

- Rank records for manual review when inspection time is limited.
- Flag possible label errors, inconsistencies, or harmful training contributions.
- Investigate links between performance problems and influential records or sources.
- Prioritize repairs by their expected benefit relative to review cost.
- Reassess the data and model after verified corrections.

Low value signals a reason to investigate. It does not establish that a record is wrong.

Application 3: acquisition and active learning

Acquisition adds new data. Active learning chooses which unlabeled data to label next.

- Identify gaps in the current dataset relative to the target task or population.
- Prioritize candidate additions by their expected contribution.
- Compare the value of more data with collection or labeling cost.
- Choose complementary batches that add coverage rather than repeat information.
- Update acquisition priorities as the available data changes.

Acquisition values a possible addition. A removal score does not automatically answer that question.

Application 4: adaptation and reweighting

Adaptation changes how available data is used to better support a target task or population.

- Use a representative target reference to assess source-data relevance.
- Select source records that better cover the target distribution.
- Assign training weights to emphasize useful data without discarding every other point.
- Reduce the influence of data that conflicts with the target objective.
- Check target performance and coverage as the reference population changes.

The reference defines what the adaptation favors. Weighting decisions still require evaluation.

Data valuation across the LLM pipeline

The data unit, objective, and available budget change across the LLM pipeline.

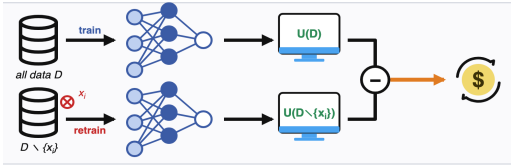
- Pretraining: choose corpus mixtures and allocate a token budget across sources.
- Instruction tuning: select instruction–response pairs for desired capabilities.
- Preference alignment: select comparisons that support the intended model behavior.
- Inference: select demonstrations and retrieved documents for a limited context window.

Define usefulness for the stage before choosing a valuation method.

Existing model-dependent valuation methods

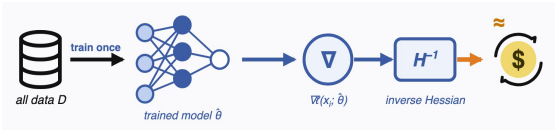
Both approaches tie data values to the chosen model.

Retraining-based valuation



- Compare utility after training with and without a data contribution.
- Repeated training can be costly across examples or data subsets.

Gradient-based methods



- Use per-example gradients to approximate a contribution's effect on reference loss.
- Less retraining, but accuracy depends on the approximation assumptions.

Can we value data efficiently without choosing a downstream model, while still capturing its usefulness for different learning tasks?

Training-data attribution for LLMs

Training-data attribution estimates how training records contribute to a chosen model behavior.

- Rank training sequences by their estimated influence.
- Identify data that supports or interferes with the desired behavior.
- Prioritize records for inspection and debugging.
- Guide decisions about correcting, removing, or reweighting data.
- Evaluate proposed changes before concluding that the behavior has improved.

Attribution estimates contribution. A proposed intervention needs a separate check.

Safety-oriented data curation for LLMs

Safety-oriented curation uses data assessments to decide which training content needs intervention.

- Prioritize documents or token spans for safety review.
- Investigate data associated with unwanted model behavior.
- Target interventions while retaining useful surrounding content.
- Choose whether to review, revise, filter, or reweight flagged content.
- Check safety improvements alongside general model performance.

An influence signal helps prioritize review. It does not by itself establish that content is harmful.

Instruction data selection for LLMs

Instruction data selection chooses which instruction–response pairs to use for fine-tuning.

- Prioritize pairs that support the capabilities the model should learn.
- Reduce redundant training data and unnecessary computation.
- Direct a limited training budget toward more useful data.
- Combine value scores with coverage constraints across skills and sources.
- Reassess the selection as the model and target tasks change.

Usefulness depends on the target capabilities, the current model, and the other selected pairs.

Preference data selection for LLMs

Preference data selection chooses which response comparisons to use for alignment training.

- Prioritize comparisons expected to help the current model.
- Identify feedback that may interfere with the desired preferences.
- Reduce the amount of feedback used during alignment.
- Allocate review effort to uncertain or conflicting comparisons.
- Evaluate the selection against independent preference judgments.

Preference labels and usefulness are different properties. The value of a comparison can change across models.

Selecting prompt demonstrations

Demonstration selection chooses which examples to include in an LLM's prompt.

- Prioritize demonstrations expected to improve task performance.
- Fit useful information within a limited context budget.
- Reduce duplication among demonstrations.
- Account for interactions and ordering within the prompt.
- Adapt the demonstration set to the model and task.

Here, valuation changes the prompt while the model's learned parameters stay fixed.

Selecting retrieved context for RAG

Retrieved-context selection chooses which documents to include when generating an answer.

- Rank documents by their estimated contribution to the answer.
- Filter content that adds little useful information.
- Allocate the context budget across the retrieved documents.
- Reduce distractions from irrelevant or misleading content.
- Check factual support and source reliability alongside model usefulness.

A document can be relevant yet add little value. Model usefulness alone does not establish factual reliability.

Model-agnostic data valuation

KAIROS: Scalable Model-Agnostic Data Valuation

Jiongli Zhu*, Parjanya Prajakta Prashant*, Alex Cloninger, Babak Salimi

University of California San Diego
jiz143@ucsd.edu, pprashant@ucsd.edu, acloninger@ucsd.edu, bsalimi@ucsd.edu

NeurIPS 2025



Jiongli Zhu*

UC San Diego



Parjanya Prajakta Prashant*

UC San Diego

* Equal contribution

THE QUESTION

How can we value data without committing to a particular learning algorithm?

An emerging research direction: data-centric, model-agnostic valuation.

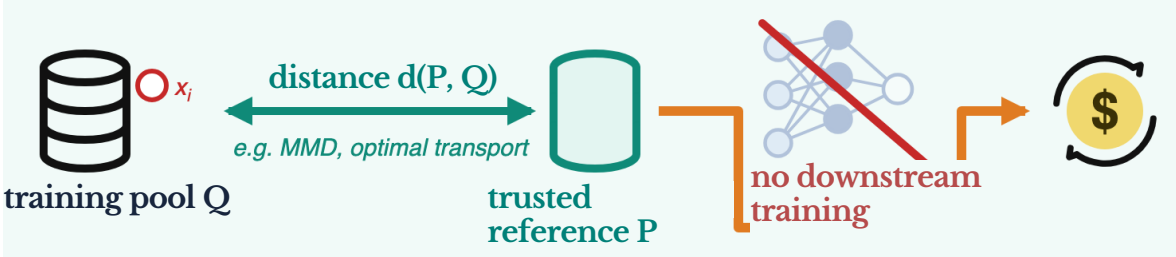
The reference population defines usefulness

P is the desired population. Q is the pool we can change.

Reference P	A clean, representative sample of the task and population we care about.
Training pool Q	The candidate examples, including possible redundancy, errors, and mismatch.
Held-out test	A separate evaluation sample that does not guide selection or tuning.

The same example can have different value for different references.

A data-level objective for valuation



Instead of training a model on Q and evaluating it on a clean, representative validation set P, we directly measure the distributional distance d(P, Q).

We value each training point by how it changes the distance between Q and P.

For the same model trained on Q, prior work establishes bounds for many distances, under suitable assumptions:

Target loss ≤ source loss + $C \cdot d(P, Q)$ + additional terms.

From a dataset discrepancy to a point's value



Before removal: Q and fixed P

remove z_i



After removal: Q without the point, same P

$$v_{LOO}(z_i) = d(P, Q_{-i}) - d(P, Q)$$

- Keep the reference P fixed and **remove one point** from the pool Q.
- If discrepancy increases, the point was **helping**. Its value is positive.
- If discrepancy decreases, the point was **hurting**. Its value is negative.
- Exact leave-one-out can require a new discrepancy computation **for every point**.

Which distance?

A value is a change in d(P, Q). Which distance?

- For **disjoint empirical supports**, KL can be infinite and total variation maximal, even when the samples are nearby.
- **Wasserstein and kernel MMD** use relationships between sample locations to capture more informative differences.

Wasserstein-1 and MMD share a test-function view: integral probability metrics.

Wasserstein distance: transport intuition

Wasserstein-1 is the least average distance needed to rearrange one distribution into another.

$$W_1(P,Q) = \inf_{\pi \in \Pi(P,Q)} \mathbb{E}_{(X,Y)\sim\pi}[\rho(X,Y)]$$

Moving probability mass

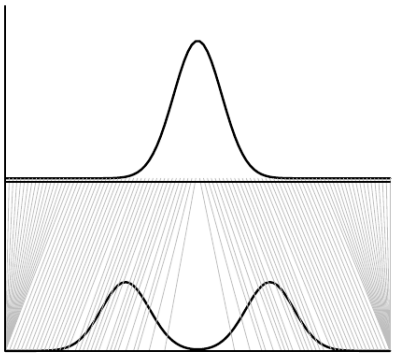


Figure: Wasserman, Fig. 3 (cropped).

Transport plans

P and Q are probability distributions. π is a joint distribution with marginals P and Q. Π(P,Q) is the set of all such plans.

Distance and minimum cost

ρ(X,Y) is the ground distance. Average it under each valid plan, then choose the least cost. Nearby, non-overlapping distributions can still be close.

For valuation: does removing an example make matching the target distribution harder?

Both are integral probability metrics

An IPM chooses a **critic** (or test function) whose average score differs most between P and Q.

$$IPM_F(P,Q) = \sup_{f \in F} \left| E_P[f] - E_Q[f] \right|$$

F: allowed critics, $f: X \to \mathbb{R}$

Wasserstein-1 the transport distance with an IPM form

$$W_1(P,Q) = \sup_{Lip(f) \le 1} \left| E_P[f] - E_Q[f] \right|$$

1-Lipschitz critics: scores change no faster than distances between inputs.

Classical exact OT: $O(n^3 \log n)$
+ $O(dn^2)$ to compute pairwise distances.

MMD

$$MMD(P,Q) = \sup_{\|f\|_H \le 1} \left| E_P[f] - E_Q[f] \right|$$

RKHS unit ball: kernel-defined critics with norm at most one.

Gaussian MMD: $O(dn^2)$
Closed-form pairwise kernel sums.

n = number of data points in each dataset; d = data dimension (number of features).

Why compare features instead of raw means?

Two distributions can have the same mean and still look very different.

Example	P puts equal mass at −1 and +1. Q puts all mass at 0. Both means equal 0.
A useful feature	The squared feature has average 1 under P and 0 under Q. It exposes the spread.
Kernel idea	Compare many features together through pairwise similarities, without listing every feature.

MMD means Maximum Mean Discrepancy: the means live in feature space.

RKHS and MMD: three simple ideas

1. **A kernel** k(x,y) measures similarity.

$$\text{Point feature: } \varphi(x) = k(x,\cdot)$$

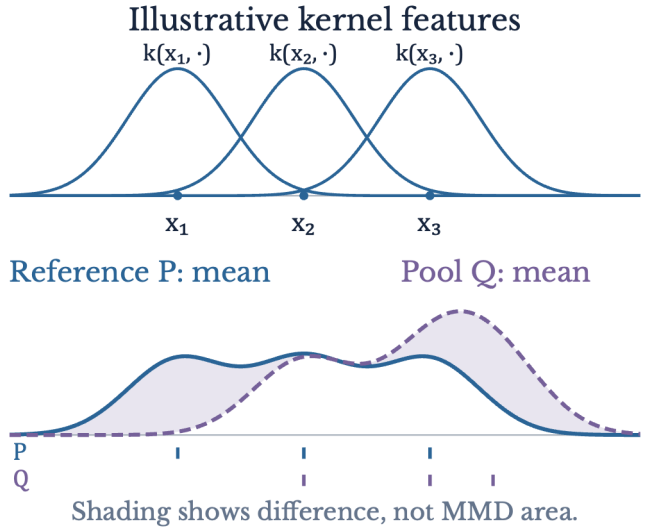
The RKHS contains weighted sums of these feature functions.

$$\langle k(x,\cdot), k(y,\cdot)\rangle_H = k(x,y)$$

2. **Dataset becomes an average.** Average its feature functions to get its mean embedding.

3. **MMD compares the averages.** Small MMD means the pool looks like the reference.

$$\text{MMD}_k(P,Q) = \|\mu_P - \mu_Q\|_H$$



RKHS norm

$$\text{MMD}_k^2(P,Q) = \underbrace{E_{X,X' \sim P}\, k(X,X')}_{\text{within P}} + \underbrace{E_{Y,Y' \sim Q}\, k(Y,Y')}_{\text{within Q}} - 2 \underbrace{E_{X \sim P, Y \sim Q}\, k(X,Y)}_{\text{across P and Q}}$$

Closed form from **pairwise similarities**, without explicit feature vectors.

The kernel determines what MMD detects

Linear kernel

$$k(x, y) = x^T y \quad MMD_k(P, Q) = \left\| E_P X - E_Q Y \right\|$$

Only mean differences, assuming the means exist.

Inhomogeneous quadratic kernel

$$k(x, y) = (1 + x^T y)^2$$

First and second raw moments, assuming finite second moments.

Gaussian kernel

$$k_\sigma(x, y) = \exp\left(-\frac{\|x - y\|_2^2}{2\sigma^2}\right), \qquad \sigma > 0$$

Characteristic on $\mathbb{R}^d$: MMD = 0 exactly when P = Q.

Sriperumbudur et al. (2010) · Muandet et al. (2017)

Bandwidth: how far does similarity reach?

Bandwidth σ sets how far apart points can be while still looking similar.

$$k_\sigma(x, y) = \exp\left(-\frac{\|x - y\|_2^2}{2\sigma^2}\right), \qquad \sigma > 0$$

Distances much smaller than σ get high similarity; distances much larger than σ get low similarity.

Small σ	Local view: only very close points look similar. Comparisons focus on small neighborhoods.
Large σ	Broader view: more distant points look similar too. Fine differences are smoothed over.
Choosing σ	Choose a scale that captures meaningful variation. Compare several bandwidths after scaling the features.

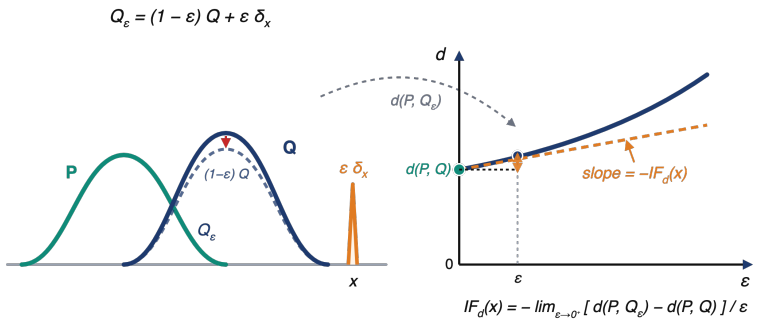
Similarity fades smoothly with distance. Bandwidth is a scale, not a hard cutoff.

First-order influence of a training point

Approximating leave-one-out

Removing one of many points slightly changes the empirical distribution. We estimate the change in distance using its local sensitivity to that point's weight.

$$\mathrm{IF}_d(x; P, Q) = -\frac{d\big(P, (1-\varepsilon)Q + \varepsilon\delta_x\big) - d(P, Q)}{\varepsilon}\Big|_{\varepsilon \to 0^+}, \tag{1}$$



Reweight x and renormalize the other points. Features and labels stay fixed.

What does the sign mean?

Positive influence: more weight on x reduces the discrepancy.

For n > 1 equally weighted points. The removal approximation assumes differentiability across ε = 0.

Zhu, Prashant, Cloninger, Salimi. KAIROS: Scalable Model-Agnostic Data Valuation. NeurIPS 2025.

The reweighting path

Take n equally weighted observations, with n > 1. Keep the reference P fixed.

$$Q = \frac{1}{n}\sum_{j=1}^{n} \delta_{x_j}$$

Nudge Q toward the selected observation x_i

$$Q_\varepsilon = (1 - \varepsilon)Q + \varepsilon\delta_{x_i}$$

Selected observation x_i

$$w_i(\varepsilon) = \frac{1 - \varepsilon}{n} + \varepsilon$$

Each other observation

$$w_j(\varepsilon) = \frac{1 - \varepsilon}{n} \quad (j \neq i)$$

At ε = 0, Q is unchanged. The weights always sum to 1.
Positive ε upweights x_i and negative ε downweights it.

KAIROS, NeurIPS 2025, Definition 1. Leave-one-out algebra and Taylor expansion shown here.

IPM influence decomposition

Upweighting a point changes the average critic score and may change the best critic.

$$d(P,Q) = \sup_{f \in \mathcal{F}} \big(\mathbb{E}_P[f] - \mathbb{E}_Q[f] \big)$$

P is the reference; Q is the pool. The optimal critics before and after reweighting are f* and fε*, respectively.

$$-\frac{d(P,Q_\varepsilon) - d(P,Q)}{\varepsilon} = C_\varepsilon(x) + G_\varepsilon(x)$$

Critic-shift term

$$C_\varepsilon(x) = \frac{\mathbb{E}_Q[f_\varepsilon^\star - f^\star] - \mathbb{E}_P[f_\varepsilon^\star - f^\star]}{\varepsilon}$$

Point-gap term

$$G_\varepsilon(x) = f_\varepsilon^\star(x) - \mathbb{E}_Q[f_\varepsilon^\star]$$

Exact for ε > 0. Influence is the limit as ε approaches 0.

KAIROS, NeurIPS 2025, §3.1, Eqs. (2)–(3). Exact finite-step decomposition.

Reading the two terms

The decomposition separates a point's critic score from a change in the critic itself.

Point gap Does x score above the pool average? A positive gap means extra mass on x narrows the discrepancy for that fixed critic.

Critic shift Reweighting the pool can change which function best distinguishes it from the reference. This term accounts for that change.

$$\mathrm{IF}_d(x; P, Q) = \lim_{\varepsilon \downarrow 0} (C_\varepsilon(x) + G_\varepsilon(x))$$

Freezing the critic needs regularity; a changing or nonunique optimum requires care.

MMD lets us differentiate directly and obtain kernel averages.

KAIROS, NeurIPS 2025, §3.1 Eqs. (2)–(3), §3.2 and Appendix B.1.

Closed-form MMD influence and its cost

The MMD derivative is just kernel similarity.

Proposition 1. *The influence function for* MMD *as the distance metric is, up to additive and positive multiplicative constants, given by*

$$\mathrm{IF}_{\mathrm{MMD}}(x; P, Q) = \mathbb{E}_{x' \sim P}[k(x', x)] - \mathbb{E}_{x' \sim Q}[k(x', x)]. \tag{6}$$

- Higher similarity to the trusted reference raises the score.
- Higher similarity to the current pool lowers the score.

Closed-form scores: $O(n + m)$ per point; $O(n^2 + nm)$ for all n points.

Recomputing MMD for all n removals: $O(n^3 + n^2m)$ total.

n training points, m reference points. Kernel evaluations; a dense Gaussian kernel adds a factor d.

Zhu, Prashant, Cloninger, Salimi. KAIROS, NeurIPS 2025, Eq. (7) and §3.4. Recompute cost derived from pairwise sums.

A score is relevance minus redundancy

For a fixed pool and reference, compare two kernel averages.

Point A	Mean similarity to P: 0.80. Mean similarity to other pool points: 0.30. Score: +0.50.
Point B	Mean similarity to P: 0.40. Mean similarity to other pool points: 0.70. Score: −0.30.
Decision	A covers a reference-relevant gap. B is less relevant and more redundant in this example.

The ranking score is not the exact leave-one-out change or a probability.

Feature matching can miss incorrect labels

Identical features can have different label quality.

Counterexample	Keep every image unchanged but flip its label. The feature distribution and feature-only MMD stay unchanged.
KAIROS extension	Combine marginal feature discrepancy with an expected conditional mean discrepancy for labels.
Interpretation	Compare label behavior given features, using trusted labeled reference examples.

A feature-only score cannot detect a label error that leaves features unchanged.

Properties of KAIROS valuation

- **Symmetry**: equivalent effects on the discrepancy give equal scores.
- **Density separation**: under the proposition's assumptions and a suitable Gaussian bandwidth, scores separate specified density regions.
- **Labels**: a conditional discrepancy compares label behavior between the pool and reference.
- **Streaming**: reuse existing kernel sums and update them for new data.

Proposition 2 (Symmetry). *Let D^{train} and D^{val} be finite samples from distributions Q and P, respectively. If for all subsets $S \subseteq D^{train} \setminus \{x_i, x_j\}$,*

$$\widehat{\mathrm{MMD}}(D^{val}, S \cup \{x_i\}) - \widehat{\mathrm{MMD}}(D^{val}, S) = \widehat{\mathrm{MMD}}(D^{val}, S \cup \{x_j\}) - \widehat{\mathrm{MMD}}(D^{val}, S)$$

then $\widehat{\mathrm{IF}}(x_i) = \widehat{\mathrm{IF}}(x_j)$.

Proposition 3 (Density Separation). *Let P and Q be two probability distributions on $\mathcal{X} \subseteq \mathbb{R}^n$. For any $\epsilon > 0$ and $r > 0$, there exists a Gaussian isotropic kernel k such that for* $\mathrm{IF}(x) = \mathbb{E}_{x' \sim P}[k(x, x')] - \mathbb{E}_{x' \sim Q}[k(x, x')]$, *the following holds*

$$\begin{cases} \mathrm{IF}(x) > 0, & \text{if } P(x') - Q(x') \geq \epsilon \, \forall \, x' \in B(x, r), \\ \mathrm{IF}(x) < 0, & \text{if } P(x') - Q(x') \leq -\epsilon \, \forall \, x' \in B(x, r), \end{cases}$$

where $B(x, r) = \{x' : \|x' - x\| < r\}$.

Proposition 4. *The influence function for* E-MCMD *as the distance metric is, up to additive and positive multiplicative constants, given by*

$$\mathrm{IF}_{\text{E-MCMD}}(x, y; P, Q) = -\|\mu^P_{Y|X}(x) - \phi(y)\|_{\mathcal{H}_\mathcal{Y}} \tag{8}$$

For existing points (x_i, y_i) where $i \leq n_t$, the terms $\mathcal{B}_i$ and $\mathcal{R}_i$ remain unchanged. The training kernel average $\mathcal{A}_i$ can be efficiently updated as:

$$\mathcal{A}_i^{(t+1)} = \frac{1}{n_{t+1} - 1} \left((n_t - 1) \cdot \mathcal{A}_i^{(t)} + \sum_{j=1}^{m} k(x^{\text{train}}_{n_t + j}, x_i) \right)$$

The kernel form supports interpretable scores and efficient updates.

A bound linking discrepancy to transfer loss

Feature and label discrepancies control a loss-gap bound under stated assumptions.

Theorem 1 (Bounding transfer loss (simplified)). *Let $(\mathcal{X}, d_{\mathcal{X}})$ and $(\mathcal{Y}, d_{\mathcal{Y}})$ be compact metric spaces and $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. Let $L : \mathcal{Z} \to \mathbb{R}$ be a continuous loss function. Then, for some constant c,*

$$\mathbb{E}_{(x,y)\sim Q}[L(x,y)] \leq \mathbb{E}_{(x,y)\sim P}[L(x,y)] + c\Big(\mathrm{MMD}(P_X, Q_X) + \mathbb{E}_{x\sim Q_X}\left[\mathrm{MCMD}_{P,Q}(x)\right]\Big).$$

The theorem bounds loss using **feature mismatch and label mismatch**. Reducing those discrepancies tightens the bound's discrepancy term.

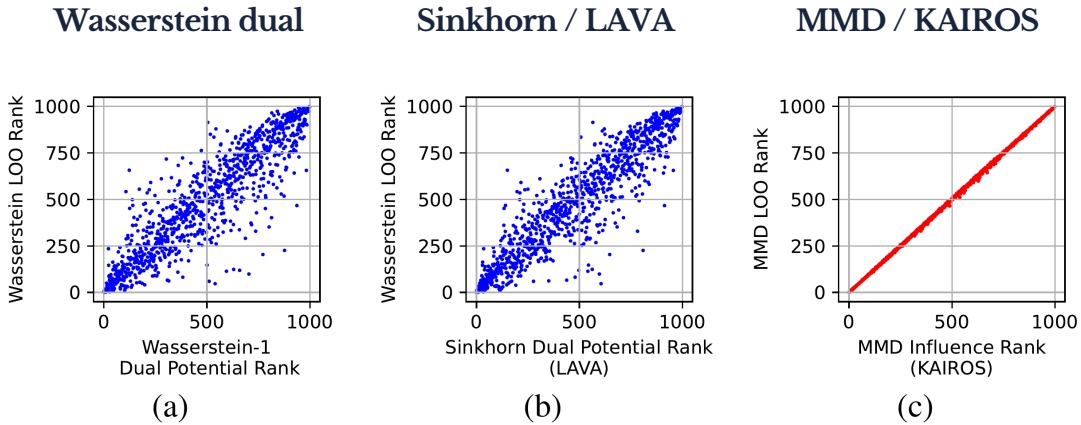
Tightens a bound; not a guarantee of better test accuracy.

Zhu, Prashant, Cloninger, Salimi. KAIROS: Scalable Model-Agnostic Data Valuation. NeurIPS 2025.

Do the scores predict the effect of removing data?

Comparison with prior work: LAVA (Just et al., ICLR 2023).

LAVA values data using sensitivity of a class-wise Wasserstein distance. Here we compare its transport-based scores with KAIROS and actual point removal.



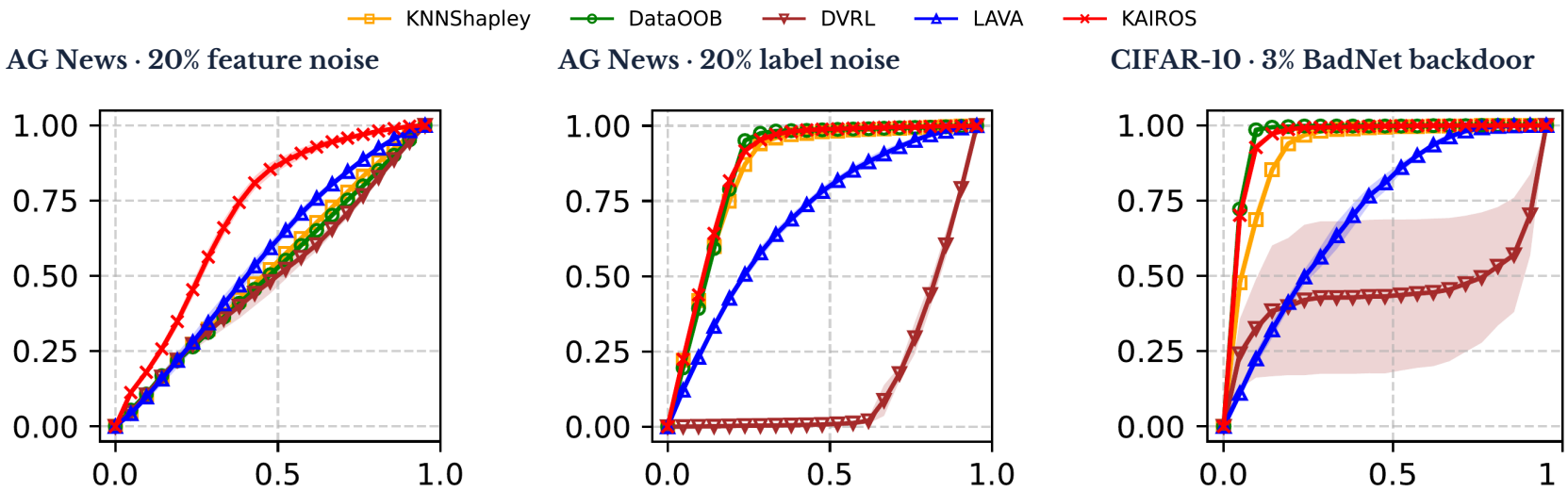
One dot per training point. x: rank by score. y: rank by leave-one-out removal.

In this experiment, top-100 overlap with leave-one-out is **99% for KAIROS** and 60% for LAVA.

Just et al. LAVA: Data Valuation without Pre-Specified Learning Algorithms. ICLR 2023.
Zhu, Prashant, Cloninger, Salimi. KAIROS: Scalable Model-Agnostic Data Valuation. NeurIPS 2025.

Finding corrupted examples early

Inspect from the lowest score upward. How much of the corruption is found?



x: fraction inspected. y: fraction of corrupted points found.

BadNets: trigger-marked images with wrong labels (3% poisoned).

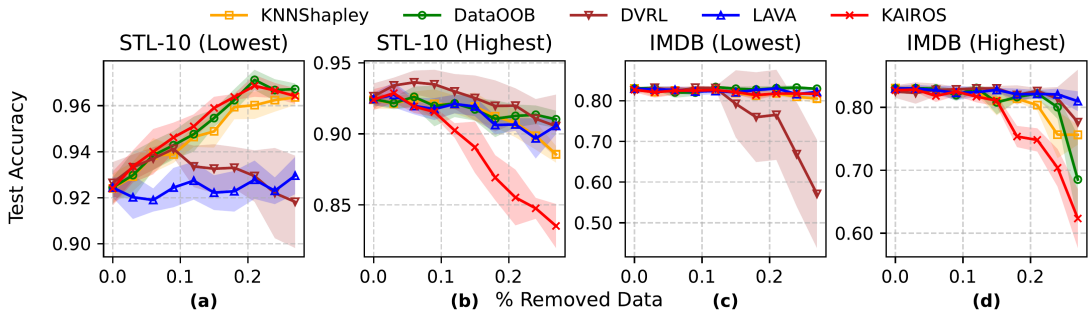
Across eight feature/label-noise settings, KAIROS ranks **first in six and second in two**. The backdoor panel is a separate test.

Zhu, Prashant, Cloninger, Salimi. KAIROS: Scalable Model-Agnostic Data Valuation. NeurIPS 2025.

Testing the ranking and scaling to ImageNet

Does pruning by score behave, and does it scale?

PRUNING BY SCORE · 20% MISLABELED



Test accuracy after removing the lowest-valued (a, c) or highest-valued (b, d) points.

Compare accuracy after removing **low-valued** versus **high-valued** points.

IMAGENET · 1.28M IMAGES

Feature noise	Runtime	AUC
KAIROS	7m56s	0.869
SAVA	1h58m	0.817
Label noise	**Runtime**	**AUC**
KAIROS	7m52s	0.861
SAVA	1h58m	0.484

ResNet-50 features, one A100. AUC of the curve on the previous slide.

SAVA: a scalable, batched version of LAVA.
Kessler, Le & Nguyen, ICLR 2025.

> On ImageNet features, KAIROS scores 1.28M images in **about 8 minutes**, versus about 2 hours for SAVA, with higher reported detection AUC.

Zhu, Prashant, Cloninger, Salimi. KAIROS: Scalable Model-Agnostic Data Valuation. NeurIPS 2025.

Trustworthiness depends on the objective and checks

Valuation helps prioritize interventions. Evaluation establishes their effects.

Reference coverage	Missing groups or outdated reference data can make useful examples look unhelpful.
Action quality	Review errors before deletion. Preserve coverage and compare with random or stratified baselines.
Separate outcomes	Measure accuracy, robustness, and group behavior directly. Privacy and safety need their own checks.

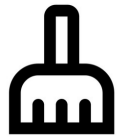
Reducing a discrepancy is a surrogate objective, not a general certificate of trust.

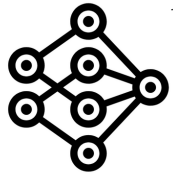
Data selection and prediction uncertainty

- **KAIROS** asks which examples to keep, using a trusted reference to define their value.
- Missing or ambiguous values admit **several plausible repairs**. Different repairs can produce different models.
- **ZORRO** asks how much a prediction can vary across those repairs.

INPUT DATA → DATA CLEANING → MODEL → INFERENCE

ID	Age	Income	...	Loan
1	25	50K	...	5K
2	NULL	60K	...	8K
3	35	NULL	...	[10K, 12K]





Loan Denied: **High** Risk

Loan Approved: **Low** Risk



How do data-cleaning choices change predictions?

Learning from uncertain training data

Learning from Uncertain Data: From Possible Worlds to Possible Models

Jiongli Zhu[1] Su Feng[2] Boris Glavic[3] Babak Salimi[1]

[1]University of California, San Diego [2]Nanjing Tech University [3]University of Illinois, Chicago

NeurIPS 2024



Jiongli Zhu

UC San Diego

THE QUESTION

How much can a prediction change across plausible repairs of the training data?

UNCERTAINTY-AWARE · SYMBOLIC · CERTIFIED

Possible Worlds Semantics

ID	Age	Income	...	Loan
1	25	50K	...	5K
2	NULL	60K	...	8K
3	35	NULL	...	[10K, 12K]
...	...	...	...	...

Possible Worlds Semantics

ID	Age	Income	...	Loan
1	25	50K	...	5K
2	NULL	60K	...	8K
3	35	NULL	...	[10K, 12K]
...	...	...	...	...

Q: Total income of the three shown rows

ID	Age	Income	...	Loan
1	25	50K	...	5K
2	30	60K	...	8K
3	35	55K	...	11K
...	...	...	...	...

→ Q(D1) = 165K

ID	Age	Income	...	Loan
1	25	50K	...	5K
2	35	60K	...	8K
3	35	60K	...	12K
...	...	...	...	...

→ Q(D2) = 170K

...

ID	Age	Income	...	Loan
1	25	50K	...	5K
2	35	60K	...	8K
3	35	60K	...	12K
...	...	...	...	...

→ Q(D3) = 170K

...

Shown range [165K, 170K]

Range of displayed completions

Answer variability

- Agreement across all represented worlds gives a certain answer. Otherwise, report the range and assess whether it matters for the decision.

ML from Possible Repairs

Dataset with Quality Issues

ID	Age	Income	...	Loan
1	25	50K	...	5K
2	NULL	60K	...	8K
3	35	NULL	...	[10K, 12K]
...	...	...	...	...

Each plausible repair produces a model. The same input can receive different predictions. Training on every repair separately is infeasible.

Possible Worlds

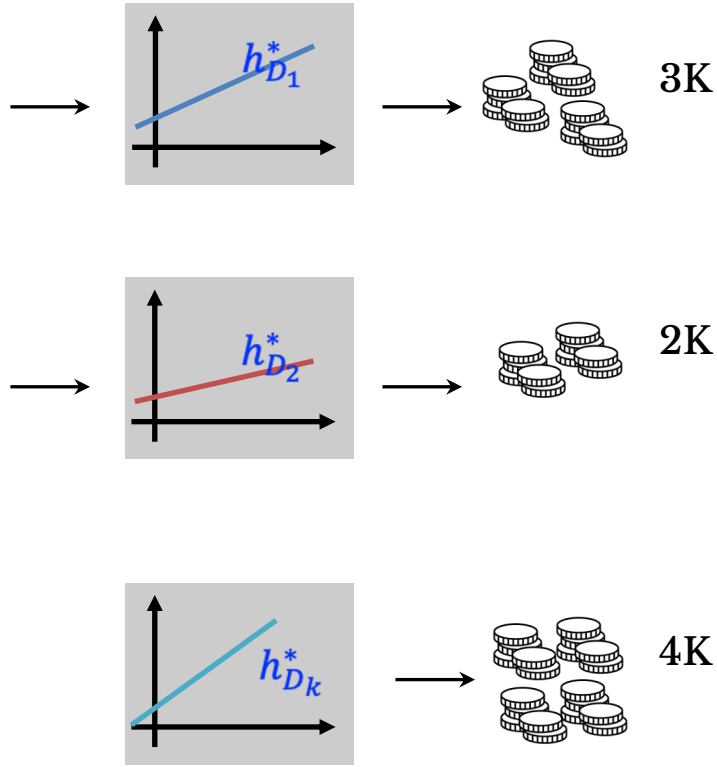
ID	Age	Income	...	Loan
1	25	50K	...	5K
2	30	60K	...	8K
3	35	55K	...	11K
...	...	...	...	...

ID	Age	Income	...	Loan
1	25	50K	...	5K
2	35	60K	...	8K
3	35	60K	...	12K
...	...	...	...	...

ID	Age	Income	...	Loan
1	25	50K	...	5K
2	35	60K	...	8K
3	35	60K	...	12K
...	...	...	...	...

Inference

- All represented models agree: certain prediction.
- Otherwise: a range, here 2K–4K.



Compact reasoning over possible models

- Databases represent every possible world compactly: c-tables, probabilistic databases — and relational queries run on the representation.

  [Imieliński & Lipski, JACM 1984] [Sundarmurthy et al., ICDT 2017] [Suciu, Olteanu, Ré & Koch, 2011]
- Reasoning over them is still NP-hard: certain answers are coNP-complete, prediction ranges NP-hard.

  [Abiteboul, Kanellakis & Grahe, TCS 1991] [Zhu, Feng, Glavic & Salimi, ZORRO, NeurIPS 2024]
- No support for numerical optimization: closed under relational algebra, not iterated arithmetic; symbolic gradient descent grows exponentially.
- Abstract interpretation: a representation closed under arithmetic — every training step maps the shape to a shape of fixed size, in polynomial time; sound, precise up to a little slack.

  [Cousot & Cousot, POPL 1977]

ZORRO uses zonotopes to represent uncertainty and its dependencies.

Zonotopes encode possible worlds

Dataset with Quality Issues

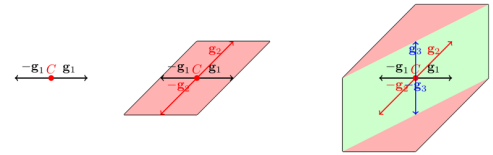
ID	Age	Income	...	Loan
1	25	50K	...	5K
2	NULL	60K	...	8K
3	35	NULL	...	[10K, 12K]
...	...	...	...	...

Possible Worlds

ID	Age	Income	...	Loan
1	25	50K	...	5K
2	30	60K	...	8K
3	35	55K	...	11K
...	...	...	...	...

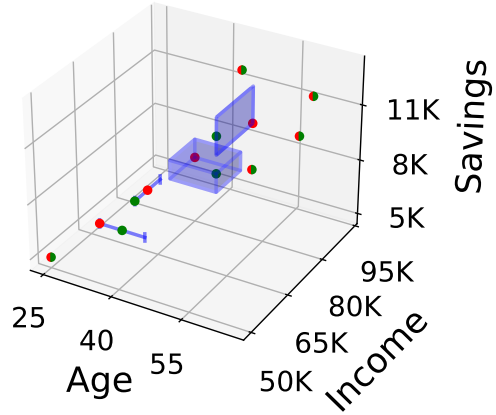
ID	Age	Income	...	Loan
1	25	50K	...	5K
2	35	60K	...	8K
3	35	60K	...	12K
...	...	...	...	...

ID	Age	Income	...	Loan
1	25	50K	...	5K
2	35	60K	...	8K
3	35	60K	...	12K
...	...	...	...	...



$c + \varepsilon_1 g_1$ $\quad c + \varepsilon_1 g_1 + \varepsilon_2 g_2$ $\quad c + \varepsilon_1 g_1 + \varepsilon_2 g_2 + \varepsilon_3 g_3$

Each ε_i varies freely in $[-1, 1]$.



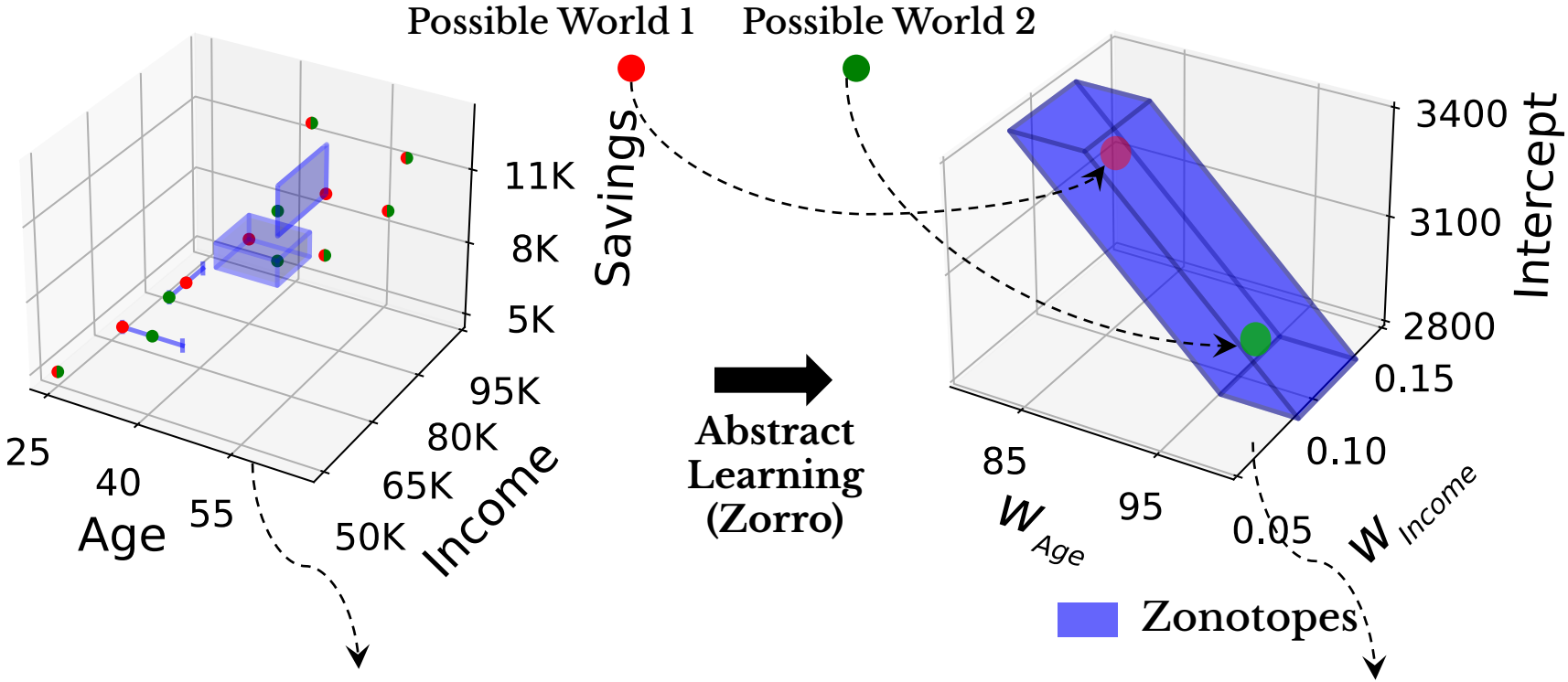
Age: 50

$Income$: $80K + 10K\ \varepsilon_1$

$Savings$: $11K + 1K\ \varepsilon_2$

$\varepsilon_1, \varepsilon_2 \in [-1, 1]$

Data uncertainty becomes model uncertainty



Age:
Income:
Savings:

$$\begin{bmatrix} 50 \\ 80K + 10K \cdot \varepsilon_1 \\ 11K + \varepsilon_2 \end{bmatrix}$$

$$, \varepsilon_1, \varepsilon_2 \in [-1, 1]$$

W_{Age}:
W_{Income}:
Intercept:

$$\begin{bmatrix} 91 + 7.7 \cdot \varepsilon_1 + 2.78 \cdot \varepsilon_3 \\ 0.1 + 0.02 \cdot \varepsilon_2 \\ 3100 - 278 \cdot \varepsilon_1 + 0.08 \cdot \varepsilon_3 \end{bmatrix}$$

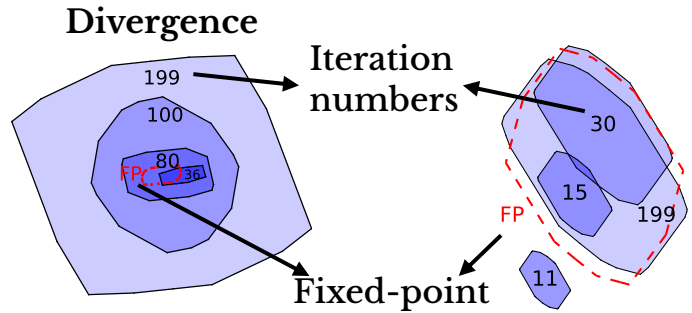
ZORRO: challenges and contributions

The challenges

- Training with uncertain values makes **symbolic expressions grow**.
- Accumulating error makes **enclosures too loose**.

Our contributions

- Control approximation and representation size to keep **updates compact**.
- For the analyzed ridge-regression setting, compute a **stable enclosure** of models across represented repairs.

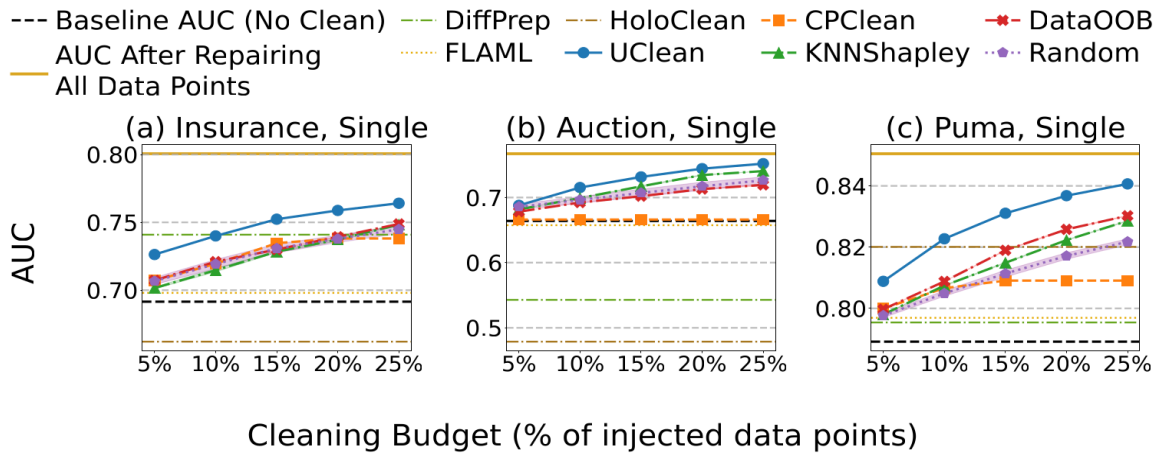


WHAT THE ENCLOSURE ENABLES

1. Establish a range of possible predictions.
2. Check whether a decision survives every represented repair.
3. **Follow-up: prioritize cells for cleaning using loss sensitivity.**

Targeted cleaning under a limited budget

UClean uses ZORRO loss sensitivity to prioritize uncertain cells for repair.



x: cleaning budget (% of injected data points). y: test ROC-AUC (0–1); higher means better class discrimination.

At 25%, UClean leads the plotted budgeted methods. Full oracle repair and non-oracle full-dataset cleaners are separate references.

DiffPrep, HoloClean and FLAML clean the full dataset without the ground-truth oracle.