Interpretable Attribute Discretization

EUGENIE LAI, MIT CSAIL, USA
INBAL CROITORU, Technion, Israel
BRIT YOUNGMANN, Technion, Israel
SAINYAM GALHOTRA, Cornell University, USA
EL KINDI REZIG, The University of Utah, USA
MICHAEL CAFARELLA, MIT CSAIL, USA

Data discretization, the conversion of numeric attributes into categorical bins, underpins many data-driven tasks, from visualization and causal analysis to symbolic learning. *Interpretable partitions* make results easier to communicate and act upon. Yet, a fundamental tension underlies this discretization process: partitions that are interpretable to humans often differ from those that maximize utility performance (e.g., causal analysis). Despite the abundance of discretization techniques, reconciling this utility-semantics trade-off remains an open challenge.

In this work, we formally define the interpretable attribute discretization problem and introduce a reinforcement-learning-based framework that jointly optimizes utility and semantic fidelity. Our approach *learns a partition selection policy* that can navigate a vast space of candidates and accurately estimate the optimal partitions by strategically sampling only 200 candidates out of millions.

We measure success by the average Hausdorff distance between the estimated and true *Pareto frontier* of the utility–semantics trade-off. On seven datasets across four tasks (visualization, modeling, imputation, causal analysis) and multiple semantic measures, our method achieves the lowest overall error, typically **2–4× smaller** than strong alternatives, yielding a more accurate Pareto frontier that reveals the complete utility–semantics trade-off.

CCS Concepts: • **Information systems** → **Information integration**; **Wrappers (data mining)**; *Data cleaning*.

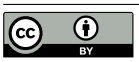
Additional Key Words and Phrases: Attribute Discretization, Semantic Interpretability, Reinforcement Learning, Multi-Objective Optimization

ACM Reference Format:
Eugenie Lai, Inbal Croitoru, Brit Youngmann, Sainyam Galhotra, El Kindi Rezig, and Michael Cafarella. 2026. Interpretable Attribute Discretization. *Proc. ACM Manag. Data* 4, 3 (SIGMOD), Article 198 (June 2026), 27 pages. https://doi.org/10.1145/3802075

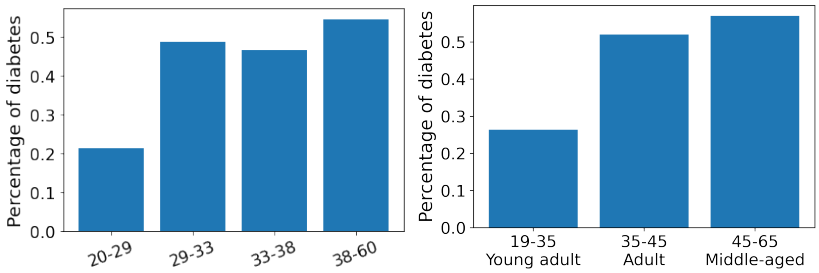
1 Introduction

Data discretization maps numeric attributes to categorical bins, turning raw values into intervals. For example, a continuous `Age` attribute can be grouped by decades (e.g., “0-10”, “10-20”) or life stages, such as “Young adult”, “Adult”, or “Middle-aged”. This transformation underpins many data-driven tasks, such as data visualization, causal inference, and symbolic learning [17], as it reduces noise and highlights structural patterns in the data.

Authors’ Contact Information: Eugenie Lai, eylai@mit.edu, MIT CSAIL, Cambridge, USA; Inbal Croitoru, inbal.cr@campus.technion.ac.il, Technion, Haifa, Israel; Brit Youngmann, brity@technion.ac.il, Technion, Haifa, Israel; Sainyam Galhotra, sg@cs.cornell.edu, Cornell University, Ithaca, USA; El Kindi Rezig, elkindi@cs.utah.edu, The University of Utah, Salt Lake City, USA; Michael Cafarella, michjc@csail.mit.edu, MIT CSAIL, Cambridge, USA.



ACM 2836-6573/2026/6-ART198
https://doi.org/10.1145/3802075



(a) Age by equal-frequency. (b) Age by life stages.

Fig. 1. (a) user's attempt; (b) our output in Ex.3.

Crucially, discretization serves two complementary yet often competing goals. (i) Utility: the discretized data support strong quality in downstream statistical or learning tasks. (ii) Interpretability: the resulting bins match human intuition and are semantically meaningful. For example, an automated splitter (e.g., ChiMerge) picks Age cuts at $(20, 24, 30, 33, 42, 43, 47, 48, 54, 61)$, yielding top validation accuracy for diabetes risk. Yet these jagged intervals resist interpretable justification, and small cut shifts change categories—hard to explain or defend to clinicians. Whereas, using guideline-aligned bins, e.g., $[19, 35)$, $[35, 45)$, $[45, 65)$, slightly trails the best accuracy, but produces stable, policy-relevant groups [34] that support clear triage rules, reporting, and interventions. *Interpretable discretization* thus acts as **a bridge between raw numerical values and human-readable patterns**.

Data-sharing efforts in science and governance further underscore the need for meaningful discretization. In a broad class of application domains, discretization is not merely preferred—it is *required* for policy design, communication, and auditing. Regulated settings, operational decision-making pipelines, and symbolic or explainable ML systems rely on threshold-based or category-based rules rather than raw numeric inputs. A concrete example is the U.S. National Cancer Institute's breast-cancer risk tables, which report risk using decade-based age bins (30, 40, 50, 60, 70), with statements such as "1 in 204 women at age 30-40" or "1 in 28 at age 60-70" [40]; raw numeric age is never presented to clinicians or patients, even though it is more precise. Similar discretization requirements arise in guidelines and protocols across many fields [9, 14, 33, 34, 53], all of which mandate interpretable categorical thresholds.

Meanwhile, standard databases often reflect *untargeted* design choices and must be adapted to specific analytical goals [17]. **Poor discretization can lead to low utility or limited interpretability**: bins that miss key patterns degrade utility, while semantically unclear bins hinder communication of insights. The following examples illustrate these challenges.

Example 1. Alex, an analyst in a research team, examines a diabetes dataset to understand the percentage of diabetes diagnoses w.r.t age for people between 21 and 60. She hopes to find an interpretable partition that reveals interesting trends in the data. Here, the utility of the visualization task is measured by the correlation between age and the probability of having diabetes. Alex starts with a standard equal-frequency method (quartiles) [17] to bin Age as $(21, 29]$, $(29, 33]$, $(33, 38]$, $(38, 60]$ (Figure 1a). The resulting bar chart exhibits an increase followed by a decrease and then another increase in diabetes prevalence across adjacent bins. She sees no clearly interpretable trend and finds it difficult to meaningfully label the bins. Alex knows there are countless ways to bin Age and wonders if at least one is satisfactory. Verifying her work without extensive trial and error is impossible. □

Example 2. In a separate task, Alex is now training a tree-based model for diabetes diagnosis, with the partitioned Age, BMI, and Glucose attributes, in the hope of finding interpretable and useful patterns in the data to guide diabetes diagnosis, as required by the clinical reporting and

decision rules used in this setting. She starts with an automated machine learning package and finds that the partition $(20, 24, 30, 33, 42, 43, 47, 48, 54, 61)$ for `Age` by ChiMerge [25] achieves the highest accuracy of 86%. After consulting with a diabetes expert, Alex realizes that this high-accuracy partition does not make much sense in medical contexts. She wonders if she can quickly find other partitions that maintain accuracy while improving the interpretability of the model and results. □

Together, these examples highlight a core dilemma: statistical discretization methods often fall short when both utility and interpretability matter. In Example 1, equal-width binning produces semantically uninformative bins that obscure trends. In Example 2, a high-accuracy ChiMerge partition yields clinically meaningless intervals. Despite the abundance of discretization techniques [17], existing methods offer limited help in situations like Alex's, where both goals are critical for downstream analysis and decision-making.

Broadly, existing approaches fall into two categories. (i) *Statistical methods*, such as equal-width or ChiMerge, optimize task metrics but often ignore domain knowledge, leading to bins that may be high-utility but uninterpretable. (ii) *Domain-specific methods* rely on handcrafted cutoffs, such as life-stage for age, which match human intuition but can degrade downstream utility. Reconciling this utility–semantics trade-off remains open.

Our Approach. We present an *interpretable attribute discretization approach* that bridges the gap by offering an automatic framework to discover *useful and interpretable* discretization strategies (referred to as partitions). In our framework, the user **inputs**: (1) a dataset with the numerical attributes to be discretized, (2) a downstream task with a corresponding utility measure (e.g., model accuracy for tree-based modeling, or interestingness measure for visualization). Our framework **outputs** a set of attribute-partition pairs (defining how to bin each attribute) that jointly optimize utility and interpretability. We revisit Example 1 and 2 to illustrate. Consider what happens when Alex uses our *interpretable-discretization* system:

Example 3. Alex configures the system to measure utility using the Spearman correlation [20] between the percentage of diabetes diagnoses and `Age`. By default, the interpretable-discretization system constructs a *candidate space* using a mix of statistical discretizers (e.g., equal-width, ChiMerge) and domain-informed proposals (e.g., life-stage bins surfaced via curated heuristics/LLM). From this pool, the system selects a life-stage partition $(19, 35, 45, 65)$ with labels ("Young adult," "Adult," "Middle-aged") (Figure 1b), commonly used in medical literature [34]. Alex quickly notices a sharp jump from young adult to adult, with a milder increase to middle-aged, yielding a visualization that is both interesting and interpretable. □

Example 4. Alex again trains a tree-based model with partitioned `Age`, `BMI`, `Glucose`. She configures the system to measure utility using model accuracy. The system then returns a set of attribute-partition pairs that balances both goals: `Age` $(19, 35, 45, 65)$ ("Young adult," "Adult," "Middle-aged"); `BMI` $(18.5, 25, 30, 68)$ ("Healthy," "Overweight," "Obese"); and `Glucose` $(0, 140, 200)$ ("Normal," "Impaired"). The resulting model attains 70% accuracy, which meets Alex's accuracy tolerance while substantially improving interpretability. Here, BMI has only one medical-aware partition, and the system finds it even though Alex is unaware of this domain-specific. Figure 2 illustrates the accuracy-semantics trade-offs for this scenario. □

The scenarios above suggest a naïve solution: exhaustively enumerate partitions per attribute and score every candidate by utility and interpretability. This brute-force approach is computationally infeasible even at moderate scales, as the number of candidates grows exponentially with the attribute domains and with the number of attributes. Moreover, jointly maximizing utility *and* interpretability is often ill-posed—improving one can degrade the other—so a single optimum may not exist (as illustrated in Figure 2). Instead, our goal is to efficiently approximate the *Pareto*

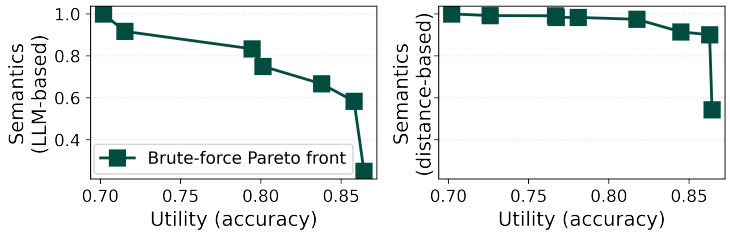


Fig. 2. Utility–semantics trade-off for Example 4. Each marker represents an optimal partition set from a brute-force search over 313,820 candidates for `Age`, `BMI`, and `Glucose`. Trade-offs vary slightly across interpretability measures (left: LLM-based; right: distance-based), highlighting the need for a framework robust to different semantic metrics.

frontier [36] of partition sets that best trade off utility and interpretability. To obtain such a system, we identify several challenges: **(C1)** being *task-agnostic* via a general, modular framework (applicable across data-driven tasks including modeling, causal inference, etc.); **(C2)** defining a robust, pluggable semantic interpretability measure for partitions; and **(C3)** navigating the exponentially large search space to discover partitions that balance both objectives.

To address **(C1)**, we provide a modular framework that treats utility as a pluggable component. Users instantiate it per task (e.g., accuracy or AUC for predictive modeling, correlation strength for exploration, or variance reduction for causal inference), without changing the overall architecture.

To address **(C2)**, we design a configurable semantic module that can use *any* semantic similarity function; in this paper, we demonstrate two instantiations: (i) distance-based metrics (e.g., KL divergence) and (ii) LLM-based judgments that capture domain knowledge (e.g., common age brackets in healthcare). While our implementation focuses on these two, our framework is extensible to other choices (e.g., ontology- or feedback-driven scores).

To address **(C3)**, we propose a reinforcement-learning-based (RL-based) search algorithm for navigating the vast search space. Evaluating the utility of a partition w.r.t a downstream task can be expensive. For example, predictive modeling requires training a model. Therefore, our goal is to assess the utility of as few partitions as possible. To this end, our framework first curates a diverse candidate pool by combining statistical discretization techniques [17] with semantically grounded suggestions from LLMs [41]. These candidates are then represented as distributions and grouped using density-based clustering with Earth Mover's Distance, to reduce the search space. Finally, we cast partition search as a Markov Decision Process (MDP) over clusters. The key observation is that evaluating a partition from one cluster provides evidence about others in that cluster. We adopt the Monte Carlo Control (MCC) algorithm and provide conditions under which the approach exhibits a provable quality guarantee. Our framework balances exploration of new clusters with exploitation of promising ones, allowing us to efficiently discover Pareto-optimal partition sets.

A demonstration of our system usability and its suitability for end-to-end employment was recently published, which provides only a high-level description of the system, whereas the present paper provides the theoretical foundations and algorithms underlying the demonstrated system, as well as the experimental study.

Contributions. Our main contributions are as follows:

- We formalize the *interpretable attribute discretization* task as a multi-objective optimization over utility and semantic fidelity, and evaluate success by the average Hausdorff distance [49] between the estimated and ground-truth Pareto frontiers.
- We present an RL-based framework that *learns a selection policy* over clustered candidate partitions. By clustering partitions and sharing evidence within clusters, Quill explores the space efficiently while preserving interpretability. We provide conditions under which the approach exhibits a provable quality guarantee.

- We instantiate scalable components, including distribution- and value-based clustering with auto-tuned hyperparameters, and budgeted exploration to control evaluation cost.
- On 7 real-world datasets across 4 tasks and multiple semantic measures, our system achieves the lowest overall error, typically **2–4× smaller** than strong baselines [41, 60], by strategically sampling only 200 candidates out of millions. The resulting Pareto frontiers expose the full utility–semantics trade-off, enabling users to select partitions that match task-specific preferences.

2 Related Work

Data discretization. There are numerous statistical data discretization methods. [17] evaluates the effect of 30 discretizers across 40 datasets on the accuracy of machine learning models. These methods range from the widely used (e.g., equal-width, ChiMerge [25]) to more sophisticated ones (e.g., distance-based discretizers [11], dependency-preserving approaches [37], correlation-preserving discretizers [35]). In this work, we consider existing discretization methods as a *search space*, from which we identify partitions that are interpretable and useful for downstream tasks. We experiment with the commonly used 11 discretizers [17] along with partitions suggested by LLMs (Section 6.1).

Interpretable AI. A large body of work studies interpretable models and when interpretability matters. Survey [13] articulates formal definitions of interpretability, evaluation taxonomies, and when interpretability matters in practice. Approaches such as TreeFARMS [60], which explores sets of sparse decision trees within the "Rashomon set" of near-optimal models, provide transparency by enumerating multiple high-performing models rather than relying on a single opaque one. Other works like [61] and Bayesian Rule Lists [29] offer sparse, human-readable logic; post hoc explanation methods like LIME [43] and SHAP [31] provide human-understandable rationales.

However, even when the model class is interpretable, the input features (e.g., discretized attributes) can lack semantic meaning. Many existing methods (including TreeFARMS [60]) require *post-processing* of continuous features or handcrafted discretization to improve interpretability. Our framework differs by producing semantically meaningful partitions *before* modeling, enabling end-to-end interpretability of both features and models (Section 6).

Automated data visualization (AutoViz) and automated machine learning (AutoML). AutoViz systems assist users in generating effective visualizations with minimal manual intervention. Systems like SeeDB [56] automatically recommend visualizations that maximize deviation between subsets, while others like Voyager [59] and Draco [38] guide chart generation using formal grammars and perceptual principles. A persistent challenge in AutoViz is the transformation of numerical attributes into human-understandable categories. Oscar [52] addresses this by learning semantic binning strategies from past visualizations, but its reliance on historical patterns limits applicability to unseen attributes.

AutoML frameworks automate the design of machine learning pipelines, including feature preprocessing and model selection [16]. While many pipelines include discretization as a preprocessing step, these methods—such as ChiMerge [25] or MDLP-based binning [15]—optimize purely for predictive accuracy. Recent systems like DeepLine [21] and GAMA [18] offer end-to-end automation but treat discretization as a low-level operation with limited interpretability. Some efforts in interpretable ML attempt to recover meaningful bins post hoc [27].

In this work, we demonstrate that our generic framework can efficiently identify interpretable and utility-preserving partitions for various downstream tasks, while letting users adjust the trade-off between utility and interpretability.

Discretization for causal inference. In causal inference, discretization plays a crucial role in defining treatment groups, confounder strata, or covariate balance [22, 24]. Many causal estimators—such as matching or stratification—rely on grouped versions of continuous covariates to

approximate conditional independence or reduce variance. However, inappropriate binning can introduce residual confounding, increase bias, or distort identification assumptions [23, 44]. Our framework identifies partitions that are both useful for causality tasks (e.g., estimating treatment effects) and meaningful to domain experts, enabling transparent analysis and communication.

3 Problem Formulation

The goal of the Interpretable Attribute Discretization Optimal Partition Sets problem is to identify meaningful discretizations of numerical attributes that balance two objectives: maximizing the utility of a given downstream task and ensuring semantic interpretability of the resulting partitions. We begin by introducing the notations and formal problem definition.

3.1 Partitions

Let $\mathbb{A}$ denote the schema of a dataset (i.e., table) D, where each attribute $A \in \mathbb{A}$ has an active domain $\text{adom}(A, D)$. A dataset instance D consists of a set of tuples of the form $t = (a_1, \ldots, a_n)$ where $a_i \in \text{adom}(A_i, D)$. Let $\mathbb{A}_{\text{num}} \subseteq \mathbb{A}$ denote the set of numerical attributes to be discretized. Namely, each attribute $A \in \mathbb{A}_{\text{num}}$ must be transformed into a categorical attribute via a valid *partition*.

The system allows flexibility in that a user may choose to exclude certain numerical attributes from partitioning if discretization is deemed unnecessary or undesirable for their analysis. However, for simplicity of presentation, we assume that all numerical attributes $\mathbb{A}_{\text{num}}$ are to be partitioned. In practice, the search space may include a "no-op" (non-discretized) discretizer, producing a single candidate partition for that attribute. Such a candidate naturally attains minimal semantic interpretability but may still achieve high task utility, and therefore can appear as an extreme point on the utility–semantics Pareto frontier.

We define a **partition** B for a numerical attribute A as an ordered list of cut points $B = (b_1, \ldots, b_{m-1})$, where each $b_i \in \text{adom}(A, D)$ and $b_1 < \cdots < b_{m-1}$. These cut points induce m contiguous bins[1]:

$$(-\infty, b_1),\ [b_1, b_2),\ \ldots,\ [b_{m-2}, b_{m-1}),\ [b_{m-1}, \infty).$$

Applying a partition B to the attribute A yields the discretized version $B(A)$. We denote by $\mathbf{B} = \{B_i \mid A_i \in \mathbb{A}_{\text{num}}\}$ a set of partitions applied across numerical attributes $\mathbb{A}_{\text{num}}$.

Example 5. Consider the Age attribute from Example 3, where

$$\text{adom}(A_{\text{Age}}, D) = [59, 59, 27, 27, 33, 50, \ldots, 35, 21, 46, 46].$$

Let us examine three candidate partitions:
$B_1 = (20, 40, 60)$ defines 4 bins $(-\infty, 20),\ [20, 40),\ [40, 60),\ [60, \infty)$;
$B_2 = (19,\ 35,\ 45,\ 60)$ defines 5 bins $(-\infty, 19),\ [19, 35),\ [35, 45),\ [45, 60),\ [60, \infty)$;
$B_3 = (15,\ 30,\ 45,\ 60)$ defines 5 bins $(-\infty, 15),\ [15, 30),\ [30, 45),\ [45, 60),\ [60, \infty)$. □

In this work, we adopt the commonly used 11 discretizers [17], ranging from the widely used (e.g., equal-width, ChiMerge [25]) to more sophisticated (e.g., distance-based [11]). We demonstrate that our framework is *agnostic* to discretization methods, allowing any (subsets) of them to be considered. The choice of discretization methods therefore defines only the candidate pool and does not restrict the applicability or correctness of the subsequent optimization procedure (details of candidates in Section 4.1).

[1]For simplicity, we adopt left-closed, right-open bins $[b_i, b_{i+1})$. An equivalent right-closed convention $(b_i, b_{i+1}]$ could be used without affecting our algorithms or metrics.

3.2 Problem Statement

We formalize the Interpretable Attribute Discretization (IAD) problem with two quality measures: a task-specific *utility* over discretized data (e.g., accuracy for modeling, correlation for data visualization) and a *semantic* score that captures interpretability. The goal is to balance these two objectives by finding a set of partitions that maximizes a weighted combination of both. We treat both functions abstractly here and instantiate them per task in Section 5.

The user provides (1) a dataset D with input attributes $\mathbb{A}_{\text{num}}$ and (2) a downstream task together with a utility function $M_{\text{util}}(\mathbf{B}, D)$ that evaluates performance on a discretized view of D. When an outcome or label attribute is present (e.g., for prediction), M_{util} incorporates it internally; in settings without an outcome, M_{util} is defined without the outcome attribute. We now state the problem.

Definition 1. [Interpretable Attribute Discretization] Given a dataset D over schema $\mathbb{A}$, a set of numerical attributes $\mathbb{A}_{\text{num}} \subseteq \mathbb{A}$, and a downstream task M_{util}, find valid partitions $\mathbf{B} = \{B_A \mid A \in \mathbb{A}_{\text{num}}\}$ that maximize:

$$\mathbf{B}^* = \arg\max_{\mathbf{B}} \ \mathcal{J}(\mathbf{B}, D) \tag{1}$$

with trade-off parameter $\alpha \in [0, 1]$ and objective

$$\mathcal{J}(\mathbf{B}, D) \ = \ \alpha \cdot M_{\text{sem}}(\mathbf{B}) \ + \ (1{-}\alpha) \cdot M_{\text{util}}\big(\mathbf{B}, D\big), \tag{2}$$

where M_{util} evaluates the utility of $\mathbf{B}$ after discretizing the input dataset D using partition set $\mathbf{B}$.

Note that the semantic score $M_{\text{sem}}(\mathbf{B})$ is a *pluggable* component that can be defined by the user, to quantify the interpretability of a candidate partition set $\mathbf{B}$. Higher values indicate more human-meaningful bins (e.g., alignment with domain conventions), independent of task utility. In our implementation (Section 4), we assume an LLM-based semantic measure (Section 5) and do not require the user to specify a semantic measure. □

The utility function M_{util} is treated as a black-box objective that maps a candidate discretization to a scalar score. Our framework imposes no assumptions on the functional form of M_{util}. As long as a utility score can be evaluated for a given discretized dataset, a user-defined M_{util} can be directly plugged into the framework. This abstraction is intentional and reflects practical settings, where utility is often defined by downstream pipelines (e.g., visualization interestingness or causal analysis) that are task-specific, expensive, but can be evaluated. While our method makes no assumptions about the form of the utility or semantic functions themselves, its empirical effectiveness relies on a smoothness assumption: partitions with similar empirical representations tend to yield similar objective values (details in Section 4.2).

The scalar α determines the relative importance of semantic interpretability and downstream task utility in the objective. Rather than fix a single α, we seek to *estimate the entire frontier* [36] of non-dominated solutions—partition sets where no other configuration achieves both a higher semantic score and a higher task utility (illustrated Figure 2). This provides a faithful picture of the utility–semantics trade-off.

Definition 2. [Interpretable Attribute Discretization Optimal Partition Sets (OPS)] Let $\mathcal{B}$ denote the set of all partition sets (one partition per numerical attribute). Each $\mathbf{B} \in \mathcal{B}$ induces a point

$$p(\mathbf{B}) \ = \ \big(M_{\text{sem}}(\mathbf{B}),\ M_{\text{util}}(\mathbf{B}, D)\big) \in [0, 1]^2, \tag{3}$$

where the first coordinate is the semantic interpretability score and the second is the task utility. For points $p, q \in [0, 1]^2$, we say p *dominates* q (write $p \succ q$) if p is at least as large in both coordinates and strictly larger in one.

The *Pareto frontier* [36] is the set of non-dominated points:

$$\mathcal{P} \ = \ \big\{\, p(\mathbf{B}) \ \big|\ \mathbf{B} \in \mathcal{B},\ \nexists\, \mathbf{B}' \in \mathcal{B} \text{ s.t. } p(\mathbf{B}') \succ p(\mathbf{B}) \,\big\}. \tag{4}$$

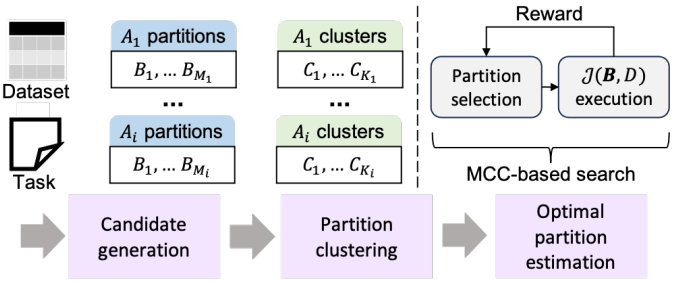


Fig. 3. Quill overview.

Optimal partition sets. We call $\mathbf{B} \in \mathcal{B}$ *optimal for IAD* iff its image lies on the frontier, i.e., $p(\mathbf{B}) \in \mathcal{P}$. Equivalently, the set of optimal partition sets is

$$\mathcal{B}^* = \left\{ \mathbf{B} \in \mathcal{B} \;\middle|\; \nexists\, \mathbf{B}' \in \mathcal{B} \text{ s.t. } p(\mathbf{B}') \succ p(\mathbf{B}) \right\}.$$

Moreover, for any scalarized objective $\mathcal{J}(\mathbf{B}, D)$ with $\alpha \in [0, 1]$, every maximizer $\mathbf{B}^*$ lies in $\mathcal{B}^*$. □

Selecting a single partition set from the Pareto frontier is inherently application-dependent: in general, no single solution simultaneously maximizes both semantic interpretability and task utility. Accordingly, our framework does not prescribe a unique optimum, but instead exposes the full set of Pareto-optimal candidates. End users may resolve this trade-off in two complementary ways: (1) *Trade-off parameter α.* Users may specify a trade-off parameter $\alpha \in [0, 1]$, which induces a scalarized objective (Equation 2) and allows the system to automatically select the corresponding solution from the estimated Pareto frontier. (2) *Interactive exploration.* We have developed an interactive UI (described in a separate Demo paper; details omitted for anonymity) that allows users to explore each point on the Pareto frontier by visualizing each discretized attribute. This helps practitioners understand the semantic implications of each candidate and then choose a suitable α if desired.

Example 6. We revisit Example 5. Assume $\mathbb{A}_{\text{num}} = \{\texttt{Age}\}$ so we have $\mathbf{B_1} = \{(20, 40, 60)\}$, $\mathbf{B_2} = \{(19, 35, 45, 60)\}$, $\mathbf{B_3} = \{(15, 30, 45, 60)\}$. Suppose the downstream task M_{util} is to train an interpretable model, say a decision tree classifier, predicting diabetes outcome. Let the measured utility scores (model accuracy) for these partitions be: $M_{\text{util}}(D_{\mathbf{B}_1}) = 0.72$, $M_{\text{util}}(D_{\mathbf{B}_2}) = 0.70$, $M_{\text{util}}(D_{\mathbf{B}_3}) = 0.74$. Meanwhile, let the semantic interpretability scores be: $M_{\text{sem}}(\mathbf{B_1}) = 0.75$, $M_{\text{sem}}(\mathbf{B_2}) = 1.0$, $M_{\text{sem}}(\mathbf{B_3}) = 0.75$.

The resulting Pareto frontier consists of $\mathbf{B_2}$ and $\mathbf{B_3}$. Partition $\mathbf{B_1}$ is Pareto-dominated by $\mathbf{B_3}$, which achieves both higher utility ($0.74 > 0.72$) and equal semantic score (0.75). In contrast, $\mathbf{B_2}$ and $\mathbf{B_3}$ are non-dominated: $\mathbf{B_2}$ offers higher semantic interpretability, while $\mathbf{B_3}$ achieves higher task utility.

Given a trade-off parameter $\alpha = 0.5$, scalarization selects a single solution from the Pareto frontier. The combined objective scores are $\mathcal{J}(\mathbf{B_2}, D) = 0.5 \cdot 1.0 + 0.5 \cdot 0.70 = 0.85$ and $\mathcal{J}(\mathbf{B_3}, D) = 0.5 \cdot 0.75 + 0.5 \cdot 0.74 = 0.745$. Although $\mathbf{B_3}$ yields higher predictive performance, $\mathbf{B_2}$ attains a higher overall objective due to its stronger semantic score, and is therefore selected for this value of α. □

The main challenge in solving the optimal partition sets for IAD problem lies in navigating the vast combinatorial search space of partition sets. If each $A \in \mathbb{A}_{\text{num}}$ admits up to k cuts from v candidates, the per-attribute count is $\sum_{i=1}^{k} \binom{v}{i}$ and the joint space over d attributes is $\left(\sum_{i=1}^{k} \binom{v}{i}\right)^d$, exponential in d. We address this by combining structure-aware clustering of partitions with MDP-based search (Section 4).

4 Quill

4.1 Overview

Our framework architecture is illustrated in Figure 3, which consists of three main components: (1) a candidate generation step, (2) a partition clustering step to group similar partitions, and (3) a search step to estimate optimal partition sets.

Candidate generation. We curate a diverse pool of candidate discretization strategies for each numerical attribute, combining 11 widely used statistical methods (e.g., equal-width, ChiMerge [25]) with interpretable partitions generated by GPT-4o [41] that reflect domain practices (e.g., age brackets in healthcare). This hybrid design ensures coverage of both utility-optimized and semantically meaningful partitions (details in [19].). Importantly, candidate generation is independent of the rest of the framework: additional discretization methods can be incorporated by adding new partition generators without modifying the clustering or search components. We include widely used methods with publicly available implementations [17] for comparability, but the framework readily supports larger or domain-specific partitions. In this work, we instantiate each statistical discretizer with varying bin counts ranging from 2 to 10, yielding a typically dense set of ∼70 unique candidate partitions per attribute (details in Table 3).

Partition clustering. For each attribute we enumerate candidate partitions, yielding a combinatorially large space of partition sets. To shrink this space, we (i) represent each partition as a distribution vector over the attribute's domain, and (ii) cluster partitions by similarity in that representation. This groups near-equivalent candidates and lets the search operate over clusters rather than raw partitions. The intuition and details explained in Section 4.2.

Optimal partition estimation. We holistically consider partition candidates across all attributes, defined in the Optimal Partition Sets (OPS) problem (Definition 2). Specifically, after clustering similar partitions for each attribute, we cast our OPS problem as a Markov Decision Process over clusters. The high-level intuition is that evaluating a partition from one cluster provides information about the others in the same cluster (details in Section 4.3). We adopt the Monte Carlo Control (MCC) algorithm to balance exploration across clusters and exploitation of promising ones (Section 4.4).

Our framework is general and modular. Candidate generation, partition clustering, semantic scoring, and utility evaluation are independent components that can be swapped or extended, while still utilizing the same clustering-and-search backbone, which is a salient feature of Quill. Specifically for utility evaluation, because the search procedure interacts with M_{util} only through observed scalar rewards, it remains agnostic to, allowing utility definitions to be swapped without affecting the algorithmic structure. We note that these components introduce hyperparameters; to improve robustness, Quill includes an automatic hyperparameter tuner (Section 4.6). The resulting Pareto frontier can be consumed either programmatically via a user-specified α or interactively by inspecting candidate discretizations.

4.2 Partition Representations

In this section, we will introduce how we represent the entire search space, including candidate partition sets across all attributes.

Partition distribution vectors. We will start by each candidate partition B as a distribution vector that encodes the distribution of data values of an attribute A across its bins, denoted by $\mathbf{v}(B)$.

Definition 3. [Partition distribution vector] Let A be a numerical attribute with observed values $\text{adom}(A, D) = \{x_1, \ldots, x_N\}$, and let $B = \{b_1, \ldots, b_{m-1}\}$ be a partition of A that induces m contiguous bins: $(-\infty, b_1)$, $[b_1, b_2)$, $\ldots$, $[b_{m-1}, \infty)$. We define $n_j(B)$ as the number of values in

adom(A, D) that fall into the jth bin:

$$n_j(B) = |\{x \in \text{adom}(A, D) \mid x \in \text{bin } j\}|, \quad j = 1, \ldots, m.$$

The *distribution vector* of B is then:

$$\mathbf{v}(B) = \left(\frac{n_1(B)}{N}, \ldots, \frac{n_m(B)}{N}\right), \quad \text{where } N = |\text{adom}(A, D)|.$$

This vector $\mathbf{v}(B)$ represents the empirical distribution of values across bins, enabling comparisons between partitions B using distance metrics over vectors. □

Example 7. Let $A = \texttt{Age}$ with $|\text{adom}(A, D)| = 114$ in this example for illustration. Consider three candidate partitions:

$$\begin{aligned} B_1 &: \{[19, 35),\ [35, 45),\ [45, 60]\}, && (n_1, n_2, n_3) = (33, 57, 24), \\ B_2 &: \{[19, 30),\ [30, 45),\ [45, 60]\}, && (n_1, n_2, n_3) = (25, 65, 24), \\ B_3 &: \{[19, 60]\}, && (n_1) = (114). \end{aligned}$$

By Definition 3, the distribution vectors $\mathbf{v}(B_1), \mathbf{v}(B_2), \mathbf{v}(B_3)$ are

$$\mathbf{v}(B_1) = \left(\tfrac{33}{114},\ \tfrac{57}{114},\ \tfrac{24}{114}\right) \approx (0.289,\ 0.500,\ 0.211),$$
$$\mathbf{v}(B_2) = \left(\tfrac{25}{114},\ \tfrac{65}{114},\ \tfrac{24}{114}\right) \approx (0.219,\ 0.570,\ 0.211),\ \mathbf{v}(B_3) = \big(1.0\big).$$

□

Partition clusters. We group partitions with similar distribution vectors, using DBSCAN [48] with Earth Mover's Distance (EMD) [46], to reduce the search space.

Definition 4. [Partition clusters] For attribute A_i, let $\mathbf{B}_i$ be its candidate partitions, and $\mathbf{V}_i = \{\mathbf{v}(B) \mid B \in \mathbf{B}_i\}$ the corresponding distribution vectors of the partitions $\mathbf{B}_i$. Let $\text{EMD}(\mathbf{v}, \mathbf{v}')$ denote the Earth Mover's Distance [46] between two vectors. We cluster $\mathbf{V}_i$ using DBSCAN with parameters $(radius, \text{minPts})$, where: (i) a vector $\mathbf{v}(\mathbf{B})$ is a *core point* if it has at least minPts neighbors within $radius$; (ii) a *cluster* C is any maximal group of core points and their density-connected neighbors; (iii) vectors not assigned to any cluster are treated as singleton clusters. We denote the resulting clusters by $\mathcal{C}_i = \{C_1, \ldots, C_{K_i}\} \subseteq 2^{\mathbf{V}_i}$. □

The clustering is task-agnostic: it depends only on $\mathbf{v}(B)$ and the distance measure, not on downstream utility. We use DBSCAN [48] with EMD [46] because it: (i) makes no parametric assumptions on cluster shape, (ii) works directly with a precomputed, non-Euclidean distance (EMD), (iii) automatically detects outliers as *noise* so they can be identified as their individual clusters (singletons), instead of forcing them into a cluster with other partitions, (iv) does not require the number of clusters K to be specified, and (v) has a small, interpretable parameter set $(radius, \text{minPts})$ that we auto-tune (Section 4.6). Ablations in Section 6 validate this choice of clustering method and distance metric.

Example 8. We continue with Example 7. Using EMD on the distribution vectors $\mathbf{v}(B)$, we obtain:

$$\text{EMD}(\mathbf{v}(B_1),\ \mathbf{v}(B_2)) \approx \mathbf{0.026} \quad \text{and} \quad \text{EMD}(\mathbf{v}(B_1),\ \mathbf{v}(B_3)) \approx \mathbf{0.121},$$

with the pairs involving B_3 substantially larger than (B_1, B_2). With DBSCAN (radius between 0.03 and 0.10, minPts = 2), B_1 and B_2 form a cluster, while B_3 is identified as noise (or a singleton), illustrating how clustering groups semantically/statistically similar partitions and isolates outliers. □

We posit that if two partitions B, B' of an attribute A have similar empirical distributions, formalized as a small EMD between their distribution vectors $\mathbf{v}(B)$ and $\mathbf{v}(B')$, then their objective values are close, i.e., $|\mathcal{J}(B, D) - \mathcal{J}(B', D)|$ is small. This is well motivated for the *utility* term: many task scores (e.g., impurity- or likelihood-based criteria, correlations, simple-model accuracies on discretized inputs) depend smoothly on bin frequencies, so a small distance induces small utility changes. For the *semantic* term, the relationship can be less smooth (e.g., domains with conventionally "hard" boundaries like BMI), and small boundary shifts may cause larger semantic deltas; we acknowledge this limitation.

In practice, we adopt this assumption to make search tractable and interactive, and we find it holds well enough to guide exploration: clustering by EMD lets us evaluate a few representatives and generalize to nearby partitions, dramatically reducing evaluations while still locating high-utility, high-semantic solutions. Empirically (Section 6), this design yields strong quality and consistent gains across datasets and metrics.

4.3 OPS as a Markov Decision Process Problem

Using the partition vectors and partition clusters, we now describe how the OPS problem in Definition 2 can be solved using a reinforcement learning algorithm we design in this work.

To solve OPS, we must select one partition per input attribute from a combinatorially large space to maximize the overall objective $\mathcal{J}(\mathbf{B}, D)$. This is inherently a multi-step process: the objective can only be evaluated once all attribute–partition choices are made, and the goal is not a single optimum but the entire Pareto frontier capturing the utility–semantics trade-off.

To make this search tractable, we first cluster partitions using their distribution vectors (Section 4.2), allowing decisions to operate at the cluster level rather than over individual partitions. Evaluating one representative within a cluster provides informative feedback about others, enabling efficient generalization and exploration. These properties make OPS naturally suited to a reinforcement learning (RL) formulation, where a policy iteratively selects one cluster per attribute and is rewarded based on the joint quality of the resulting partition set.

We are now ready to cast OPS as a finite-horizon Markov Decision Process (MDP), which cleanly exposes the choices, constraints, and objectives of our search.

Definition 5. [Optimal Partition Sets as a Markov Decision Process (OPS-MDP)] Let the numerical attributes be $\mathbb{A}_{\text{num}} = \{A_1, \ldots, A_d\}$, and let each attribute A_i have a set of candidate partitions $\mathbf{B}_i = \{B_1, \ldots, B_{j_i}, \ldots, B_{M_i}\}$, where B_{j_i} is a candidate partition for A_i. From these candidates, we form clusters of similar partitions $C_i = \{C_1, \ldots, C_{k_i}, \ldots C_{K_i}\}$, where each cluster $C_{k_i} \subseteq \mathbf{B}_i$ groups partitions of A_i that are semantically and statistically similar (as defined in Section 4.2). The union of all such clusters across attributes $\mathbb{A}_{\text{num}}$ constitutes the reduced search space for OPS.

OPS-MDP is defined as follows:

- **States:** A state s_t represents a partial assignment of partitions to attributes, i.e., a set $\mathbf{B}_t = \{B_{j_i} \mid i \in \{1, \ldots, d\}\}$ where $d = |\mathbb{A}_{\text{num}}|$.
- **Actions:** An action $a = (A_i, C_{k_i})$ selects one of the clusters $C_{k_i} \in C_i$ for an *undiscretized* attribute A_i. A concrete partition (representative) $B_{j_i} \sim C_{k_i}$ is then sampled from this cluster and applied to discretize A_i.
- **Transitions:** Deterministic, updating the state by adding A_i and its chosen partition B_{j_i} to the current partial assignment.
- **Episodes:** Run for $d = |\mathbb{A}_{\text{num}}|$ steps, terminating when all d attributes have been assigned partitions, resulting in a complete partition set $\mathbf{B}_d = \{(A_1, B_{j_1}), \ldots, (A_d, B_{j_d})\}$, where each $B_{j_i} \in \mathbf{B}_i$ is the chosen partition for attribute A_i.

- **Reward:** The agent receives reward $r = 0$ for $t < d$, and $r = \mathcal{J}(\mathbf{B}_d, D)$ once all attributes have been discretized.
- **Policy:** A mapping $\pi(a \mid s)$ from partial states s to distributions over feasible actions a, where each action $a = (A_i, C_{k_i})$ selects a cluster C_{k_i} for an undiscretized attribute A_i.

This formulation allows us to learn a policy that constructs a high-quality partition set by leveraging generalization within clusters and delayed reward feedback from the final objective. □

Valid solutions to OPS-MDP. We first show how a valid solution to OPS would look like in the Markov Decision Process setup, before describing our RL solution.

We discretize a set of input attributes $\mathbb{A}_{\text{num}}$. Since we are applying exactly one partition to each attribute A_i, then out of all clusters $\mathcal{C}_i$ for A_i, only exactly one partition should be chosen (as the representative) $B_{j_i} \sim C_{k_i}$ from exactly one cluster $C_{k_i} \in \mathcal{C}_i$. The union of such partitions across all attributes $A_i \in \mathbb{A}_{\text{num}}$ is then the overall selected partition set $\mathbf{B}_d = \{(A_1, B_{j_1}), \ldots, (A_d, B_{j_d})\}$.

Definition 6. [Valid solutions to OPS-MDP] Given the set of attributes $\mathbb{A}_{\text{num}} = \{A_1, \ldots, A_d\}$, each attribute A_i has an associated set of partition clusters $\mathcal{C}_i = \{C_1, \ldots, C_{K_i}\}$, where each cluster C_{k_i} contains candidate partitions $B_{j_i} \in \mathbf{B}_i$. A valid OPS solution must select exactly one partition for each attribute. Each valid $\mathbf{B}_d$ thus represents a complete discretization—one partition per attribute—corresponding to an episode in OPS-MDP. □

Given that there exist many valid solutions (exponential in the number of attributes), we resort to the objective function of OPS to evaluate the relative merits of these solutions as $\mathcal{J}(\mathbf{B_d}, D)$.

4.4 Solving OPS-MDP via Monte Carlo Control

We now describe our RL algorithm for OPS-MDP, that can solve OPS-MDP efficiently, with conditional provable quality guarantees.

Monte Carlo Control (MCC) is a classic reinforcement learning algorithm for learning optimal policies in episodic, model-free settings [55]. It is particularly well-suited to our OPS-MDP for three reasons: (i) OPS-MDP requires a delayed reward available only after all attribute–partition selections are complete, which matches MCC's episodic, return-based updates. (ii) The problem has a finite horizon and discrete action space—one decision per attribute—making sample-based value estimation both stable and efficient. (iii) OPS-MDP has deterministic transitions and requires no transition model, aligning naturally with MCC's model-free formulation. These properties make MCC an ideal and principled choice for optimizing policies in OPS-MDP while maintaining computational tractability.

Q-table setup for OPS-MDP. We model the OPS environment as a finite-horizon Markov Decision Process (MDP) in which each state represents a *partially filled partition set* $\mathbf{B}_t$, and each action corresponds to selecting a *partition cluster* C_{k_i} for attribute A_i.

Let K_i denote the number of clusters available for attribute A_i, and define the total number of clusters across all attributes as $\sum_{i=1}^{d} K_i$. We construct a Q-table with $n = \sum_{i=1}^{d} K_i + 2$ total states and actions, where: one dummy *start state* s_0 marks the beginning of each episode; $\sum_{i=1}^{d} K_i$ *intermediate states* correspond to selecting each of the $\sum_{i=1}^{d} K_i$ attribute–cluster pairs; one *terminal state* s_n signifies completion of the partitioning process. For example, with 3 attributes—`Age`, `BMI`, and `Glucose`—each having 3 clusters, we have $\sum_{i=1}^{d} K_i = 3 + 3 + 3 = 9$ and $n = 11$, yielding an 11×11 Q-table.

Thus, the Q-table is an $n \times n$ matrix, where each entry $Q(s, a)$ estimates the expected return of taking action a in state s. In OPS-MDP, where the reward is only observed at the end of each episode, these values are updated retrospectively using the final reward, which is shared by all state–action pairs visited within that trajectory.

Algorithm 1: Solve OPS-MDP via Monte Carlo Control

Input: Attributes $\{A_1, \ldots, A_d\}$, clusters $\{C_1, \ldots, C_d\}$, number of episodes T
Output: Estimated Pareto-optimal set of partitions $\mathcal{B}^*$

1 Initialize Q-table Q and returns dictionary $\texttt{returns}[(s,a)] \leftarrow []$;
2 Initialize policy π_θ **for** $n = 1$ *to* T **do**
3 Initialize state s_0, empty partition set $\mathbf{B_t}$, and episode history $H \leftarrow []$;
4 **while** *not all attributes covered* **do**
5 Select valid cluster C_{k_i} from state s using π_θ;
6 Sample $B_{j_i} \sim C_{k_i}$, update $\mathbf{B_t}$ and transition to next state s';
7 Append (s, a) to H;
8 Set $s \leftarrow s'$;
9 Compute final reward $r = \mathcal{J}(\mathbf{B_d})$;
10 **foreach** $(s_t, a_t) \in H$ **do**
11 Append r to $\texttt{returns}[(s_t, a_t)]$;
12 Update $Q(s_t, a_t) \leftarrow \frac{1}{|\texttt{returns}[(s_t,a_t)]|} \sum r' \in \texttt{returns}[(s_t, a_t)] r'$;
13 Update policy π_θ using updated Q;
14 Compute the Pareto front $\hat{\mathcal{B}} = \texttt{compute\_pareto}(\{\mathbf{B}_1, \ldots, \mathbf{B}_T\})$;
15 **return** $\hat{\mathcal{B}}$;

Valid actions and Q-table masking. To enforce the one-partition-per-attribute constraint (Definition 6), we apply a masking strategy that prunes invalid transitions from the Q-table. Specifically, we initialize invalid actions with a value of $-\infty$, preventing the agent from: (i) selecting more than one cluster for the same attribute (ii) transitioning back to the start state s_0, (iii) or taking actions from the terminal state s_n, after a complete partition set has been formed.

This *Q-table masking*, detailed in [19], is a critical design choice in our MCC formulation. It ensures that all episodes represent valid solutions, reduces the effective action space, and accelerates convergence by guiding exploration. Notably, the masking procedure is *generalizes* to any number of attributes and clusters, preserving the modularity and correctness of Quill framework.

Policy for OPS-MDP. At each step in an episode, the agent selects a valid cluster C_{k_i}, samples a partition $B_{j_i} \sim C_{k_i}$, adds B_{j_i} to the current partition set $\mathbf{B}_t$, and computes the reward as the change in $\mathcal{J}$. Let $Q(s, a)$ denote the state-action value for state s and action a, and let $\mathcal{A}_s$ denote the set of valid actions in state s, after masking out. Then the ϵ-greedy policy is defined as:

$$\pi_\epsilon(s) = \begin{cases} \arg\max_{a \in \mathcal{A}_s} Q(s, a), & \text{with probability } 1 - \epsilon \\ \text{Uniform}(\mathcal{A}_s), & \text{with probability } \epsilon \end{cases} \tag{5}$$

This formulation ensures that the agent selects among only valid actions at each state, while balancing exploration and exploitation.

Solving OPS-MDP. With all components described above, we are now ready to introduce our Algorithm 1, which builds on top of Monte Carlo Control (MCC) [55] to solve OPS-MDP.

Algorithm 1 begins by initializing the Q-table Q and the returns dictionary for tracking observed returns, along with the policy π_θ using an ϵ-greedy exploration schedule (lines 1–2). For each of the T episodes, the agent starts from an initial state s_0 with an empty partition set $\mathbf{B}_t$ and initializes an episode history H to store encountered (state, action) pairs (line 3). The agent then iteratively selects a valid cluster C_{k_i} from the current state using the policy π_θ, samples a partition B_{j_i} from C_{k_i}, adds it to $\mathbf{B}_t$, transitions to the next state, and records the (s, a) pair in the episode history (lines 4–8). Once all attributes are covered and the episode ends, the final reward r is computed as the overall objective score $\mathcal{J}(\mathbf{B_d}, D)$ (line 9). This final reward is then propagated to all (state, action) pairs

in the episode history: for each pair, the reward is appended to the list of observed returns, and the corresponding Q-value is updated as the mean of these returns (lines 10–12). The policy π_θ is then updated using the newly estimated Q-values (line 13). After all episodes are completed, the algorithm computes the final Pareto-optimal set $\hat{\mathcal{B}}$ over all evaluated partition sets and returns the final estimation (lines 14–15).

Our framework offers several compelling advantages. First, it enables coordinated decision-making across attributes, learning joint discretization strategies rather than treating attributes independently. This coordination allows the framework to naturally capture cross-attribute interactions, which are critical in many real-world datasets where utility quality depend on attribute combinations (e.g., causal inference, modeling). Second, it is metric-agnostic: it requires no assumptions about the form of the objective $\mathcal{J}$, and can flexibly accommodate any combination of utility and semantic measures. These properties make it a flexible, general-purpose approach for OPS. We additionally include an automatic hyperparameter tuner (Section 4.6).

4.5 Special Case: Single-Attribute OPS-MDP

Attribute discretization spans many use cases (Section 5), including both joint discretization across multiple attributes (e.g., causal analysis) and single-attribute settings (e.g., visualization). While our OPS-MDP formulation applies to both, formal regret guarantees are only possible in the single-attribute case. Accordingly, we analyze the $d = 1$ setting here, where the problem admits a reduction to a classical *multi-armed bandit*.

With d=1, the MCC formulation collapses to a single decision: choose one partition for the attribute. Each episode has exactly one action, after which we observe the final reward via the joint objective $\mathcal{J}$. This is precisely a multi-armed bandit: each cluster is an arm, and sampling a partition corresponds to pulling that arm.

Since *multi-armed bandits are degenerate MDPs* with a single state [55], classical algorithms such as Upper Confidence Bound (UCB) solve this setting and provide regret guarantees. When d=1, MCC reduces to UCB and *inherits UCB's regret bounds*, giving theoretical grounding for our framework, only in this case.

Theorem 1. If (i) d=1 (single attribute, one action per episode) and (ii) MCC uses the UCB index for action selection and updates, then our Algorithm 1 becomes a UCB-guided MCC variant and reduces exactly to a stochastic multi-armed bandit with K arms (one per cluster) and reward in $[0, 1]$ given by $\mathcal{J}$, and this equivalence does not extend to the multi-attribute case ($d > 1$). Running Algorithm 1 for T episodes yields the regret guarantees, inherited from UCB [10]—gap-free regret $O(\sqrt{KT \log T})$ and gap-dependent regret $O(\sum_i \frac{\log T}{\Delta_i})$, where Δ_i is how far cluster C_i's expected reward is below the optimal expected reward $\mathcal{J} = 1$.

Proof by induction on the number of episodes t. We show that for all episodes $t \geq 1$, the behavior of Algorithm 1 under $d = 1$ is equivalent to that of UCB. Their selected arms, maintained statistics, and update rules are the same across both algorithms.

Base case ($t = 1$): At the beginning, all cluster arms are untried. UCB initializes by pulling each arm once to get an initial reward estimate. In MCC, when $d = 1$, there is only one decision per episode: selecting one cluster, sampling a partition from it, and observing the reward $\mathcal{J}(B, D)$. We simplify the partition set notation from $\mathbf{B}$ to B here, since there is only one partition B in the set when $d = 1$.

Suppose there are K clusters (arms) $\{C_1, \ldots, C_K\}$. Then in the first k episodes, MCC samples each cluster once. These samples populate the initial reward statistics, just like UCB. So after $t = k$,

both UCB and MCC have exactly one reward value for each arm, and their empirical means match:

$$\hat{r}_i = \mathcal{J}(B_i, D) \quad \text{for each arm } i.$$

Inductive step: Assume that after $(t-1)$ episodes, the following invariants hold: (1) Each cluster C_i has been sampled n_i times in both MCC and UCB. (2) The empirical mean reward $\hat{r}_i$ for each cluster is the same in both algorithms. (3) The cluster selected in episode $t-1$ by MCC is the same as that selected by UCB.

Now consider episode t.

- **Action selection:** Since both use the UCB index for action selection and updates.

$$a_t = \arg\max_i \left(\hat{r}_i + \sqrt{\frac{2 \log t}{n_i}} \right)$$

  Thus, both algorithms choose the same cluster arm in episode t.
- **Reward observation and update:** MCC samples a partition $B \sim C_{a_t}$, observes its reward $\mathcal{J}(B, D)$, and updates the empirical mean via the incremental update rule:

$$\hat{r}_{a_t} \leftarrow \hat{r}_{a_t} + \frac{\mathcal{J}(B, D) - \hat{r}_{a_t}}{n_{a_t} + 1}.$$

  This is equivalent to recomputing the average over all samples for a_t, matching UCB's update step.
- **Statistics consistency:** After the update, both algorithms have: $n_{a_t} \leftarrow n_{a_t} + 1$, and updated $\hat{r}_{a_t}$ with the same formula. Hence, all maintained statistics remain identical between the two.

Conclusion: By induction, we have shown that for all t, the arm selected and the statistics maintained by MCC match exactly those of UCB. Thus, MCC reduces to UCB when $d = 1$. □

This UCB-guided MCC variant is effective in the single-attribute ($d = 1$) setting because choices are independent and the reward is one-step. For $d > 1$, the problem no longer satisfies the core assumptions of stochastic bandits—independent arms and immediate rewards—making UCB regret guarantees inapplicable.

Moreover, naïve UCB extensions are suboptimal. *Independent UCB* runs UCB per attribute, implicitly assuming partition choices can be made separately and ignoring cross-attribute interactions. *Sequential UCB* imposes an attribute order and applies UCB step by step, but early choices cannot be revised, inviting cascading errors. We evaluate both in Section 6.

4.6 Tuning Hyperparameters

Our end-to-end framework includes three key hyperparameters: ($radius$, minPts) required to form a dense region in DBSCAN [48], and the exploration rate ϵ in the ϵ-greedy policy in MCC. Exhaustive grid search over these parameters would be computationally expensive. Instead, we employ lightweight, data-driven tuning strategies that adapt hyperparameter values to the specific input dataset D.

The silhouette score [45] measures how well a data point fits into its assigned cluster compared to other clusters. Higher silhouette scores indicate better cluster separation and cohesion. Our auto-tuner selects DBSCAN's $radius$ and minPts from data, using Wasserstein distances between candidate distributions. For every ($radius$, minPts) pair, it runs DBSCAN with the precomputed distance matrix and scores the result by silhouette. Our framework uses the pair with the highest score. This way, clustering quality is ensured without using the objective score $\mathcal{J}$, which requires running the framework end-to-end.

To tune the exploration rate in MCC, we use an exponential decay schedule rather than fixed ϵ values, which shrinks the tuning space and adapts exploration during training. The exploration rate at episode n is defined as:

$$\epsilon_n = \max\left(\epsilon_{\min}, \epsilon_0 \cdot e^{-kn}\right), \quad \text{where} \quad k = \frac{\log(\epsilon_0/\epsilon_{\min})}{T},$$

and T is the total number of training episodes. This schedule enables a smooth exploration-to-exploitation transition and requires only two parameters (ϵ_0, $\epsilon_{\min}$), while still covering a broad spectrum of ϵ behaviors in a single run. In our experiments we set $\epsilon_0 = 0.7$, $\epsilon_{\min} = 0.5$ as default. This choice performs strongly across the majority of our 7 datasets and 14 (utility × semantic) scenarios. We include an ablation study against the set of hand-tuned hyperparameters (fit on 2 datasets), to show the effectiveness of the auto-tuner (Section 6.2.3).

5 Instantiations of Utility and Interpretability

To demonstrate our framework's general, modular design, we instantiate it with representative semantic measures and utilities. These cases serve as concrete testbeds to showcase effectiveness and efficiency, and to empirically verify that the framework is agnostic to both task choice and semantic metric (Section 6).

5.1 Measuring Interpretability

Our first objective is to quantify the interpretability of a given partition with respect to a numerical attribute. As no off-the-shelf metric directly addresses this, we propose a few intuitive measures tailored to this objective.

LLM-based semantic measure. We use corpus attestations as a proxy for semantic meaningfulness: partitions that appear frequently in literature or code likely reflect domain expertise and practical utility. Direct counting references is not trivial (e.g., attributes with different synonym names, partitions only in source code). We therefore query LLMs to aggregate this knowledge, due to their comprehensive knowledge base and ability to capture nuances. We investigated two models: GPT-4o [41] for broad domain knowledge, and Perplexity AI [8] for source-grounded responses.

We use a 1-4 scale for the semantic value of a given partition for an attribute, where 1 represents poor semantic value (not used at all) and 4 represents excellent semantic value (very commonly used). The specific prompts used are given in [19].

To validate our approach, we ran a user study with 54 Prolific participants [7]. Each rated 26 items—an attribute description paired with a candidate partition—on a 1–4 semantic scale (matching our LLM prompts; form in repo). Inter-rater agreement (Krippendorff's α [26]) was 0.20, reflecting task difficulty yet indicating modest consensus. Human ratings correlated strongly with LLM scores: Spearman $\rho = 0.90$ for GPT-4o and $\rho = 0.83$ for Perplexity (both $p < .05$). We therefore use GPT-4o for subsequent experiments.

Distance-based semantic measures. When interpretable partitions are known (e.g., binning `Age` into $(19, 35, 45, 65)$ with labels ("Young adult", "Adult", "Middle-aged") in life sciences [34]), users can supply a *gold* partition B_{gold}. Our semantic module then scores any candidate B by its distance to the gold.

Note that when distance-based semantic measures are used, the reference ("gold") partition is not assumed to be uniquely correct. Rather, such measures are intended for settings where one or more preferred binning schemes are available, e.g., from established guidelines or expert conventions, and serve as anchors for measuring deviation. In cases where multiple alternatives are considered reasonable, users may supply multiple gold partitions, and the semantic score of a candidate partition can be defined via an aggregation (e.g., maximum distance) over these references.

Importantly, this choice affects only the semantic scoring module; the rest of the framework operates unchanged. In this work, we consider the following distance-based semantic measures.

L2 norm. L2 norm is a symmetric and intuitive metric, ideal for situations where partitions differ structurally (e.g., different numbers of cut points), and where bin-wise deviation should be penalized uniformly. Given a candidate B (defined with cut points in Section 3.1) and the gold, we pad the shorter partition with zeros to equalize their lengths and compute the L2 norm.

Kullback–Leibler (KL) divergence [12]. Viewing $\mathbf{v}(B)$ and $\mathbf{v}(B_{\text{gold}})$ as discrete probability distributions, $\mathrm{KL}\big(\mathbf{v}(B) \,\|\, \mathbf{v}(B_{\text{gold}})\big)$ measures how well B_{gold} is approximated by B; it is 0 iff the distributions match. KL divergence measures the information loss (coding inefficiency) when the candidate partition is used in place of the gold. It penalizes moving probability mass in ways that contradict the gold, thereby aligning with domain priorities.

When to use which measure. If domain guidelines exist (e.g., BMI or age cutoffs), distance-based scores (L2 norm or KL) give transparent alignment to a reference. When such priors are unavailable or hard to formalize, LLM-based measures offer flexible, knowledge-grounded semantics. Our framework is agnostic to the choice and performs consistently across these measures (Section 6).

5.2 Measuring Usefulness

Table 1. Downstream tasks and utility definitions used in our benchmark. $\mathbf{B}(D)$ denotes the discretized (binned) data.

Task	Attribute setting	Description (utility measure)
Visualization (visual)	Single-attribute	Spearman correlation between the partitioned attribute and the outcome on $\mathbf{B}(D)$ (higher is better).
Interpretable modeling	Multi-attribute	Predictive accuracy of a tree-based model trained on $\mathbf{B}(D)$.
Data imputation	Multi-attribute	Discretize, apply KNN imputation on $\mathbf{B}(D)$, then train a tree-based model; utility is predictive accuracy.
Causal analysis (known ATE)	Multi-attribute	Preservation of causal effects measured by exponential ATE deviation decay (closer to 1 is better).
Causal analysis (unknown ATE)	Multi-attribute	Macro-averaged F1 of preserving high/low Spearman correlations before vs. after discretization.

Next, we instantiate utilities for common tasks—visualization, interpretable modeling, and causal analysis. Table 1 summarizes the measures used in our experiments.

Data visualization. Data visualizations (e.g., bar plots) help users spot patterns that raw tables can obscure. Visualization recommendation systems [51, 56, 58] often score *interestingness* via simple statistics (e.g., correlations). In this work, we quantify interestingness (utility) using the Spearman correlation coefficient [20]. We choose a partition of a *single* numerical attribute to maximize the correlation between the partitioned A_i and the outcome attribute.

Interpretable modeling. Many applied domains (e.g., economics, social science, agriculture) emphasize transparency and favor inherently interpretable models over post-hoc explanations [47]. Discretizing numeric attributes is especially useful for inherently interpretable models—decision trees, rule-based classifiers, and scorecards—yielding simpler splits, reduced noise sensitivity, and closer alignment with domain knowledge. In this use case, we discretize all numerical attributes in $\mathbb{A}_{\text{num}}$ prior to training. We assume a tree-based model, and we measure utility by predictive accuracy of the outcome model.

Data imputation. Data imputation replaces missing values with estimates (e.g., mean/median, most-frequent value, or nearest-neighbor) so that analysis and modeling need not discard incomplete rows. Discretizing numeric attributes *before* imputation simplifies this step: a smaller, structured set of bin labels makes plausible replacements easier and more consistent (e.g., imputing the most frequent bin within a subgroup). For this task, we select a partition for each $A_i \in \mathbb{A}_{\text{num}}$, discretize

the data, and apply KNN-imputation [50] on the discretized data. We then train a tree-based model; utility is the model's accuracy on the outcome attribute.

Causal analysis. Causal analysis estimates the effect of a treatment on an outcome while adjusting for confounders. We use the Average Treatment Effect (ATE), the expected treated–vs–untreated outcome difference under proper adjustment [42]. Discretizing numeric confounders simplifies conditioning on continuous variables, can improve robustness, and reveals heterogeneous effects across subgroups. Because counterfactuals are unobserved, evaluation requires specialized metrics. We therefore propose two utility measures for ATE estimation, reflecting different levels of available knowledge; details appear in [19].

Known ATE. When the true ATE is known, utility is the exponential of the negative absolute error between the true ATE (on D) and the ATE computed after discretization (on $\mathbf{B}(D)$): $M_{\text{util}}(\mathbf{B}, D) = \exp^{-|\text{ATE}(D)-\text{ATE}(\mathbf{B}(D))|}$, where $\text{ATE}(D)$ denotes the causal effect evaluated on dataset D. Thus, smaller ATE deviations yield values closer to 1 (better), and larger deviations shrink toward 0. Confounder adjustment is implicit in $\text{ATE}(\cdot)$.

Unknown ATE. When the true ATE is unavailable (typical in observational data), we evaluate how well discretization preserves dependency structure. We compute pairwise Spearman correlations [54] before and after discretization, label each as *high* or *low* using quartile thresholds ($\tau_{\text{low}} = \frac{1}{3}$, $\tau_{\text{high}} = \frac{2}{3}$), and score agreement via macro-averaged F1 [32]. This correlation-order preservation serves as a proxy for causal stability; on synthetic data it aligns with the known-ATE utility ($R^2 = 0.7989$, $p < 0.05$).

6 Experimental Evaluation

We embody our framework in a prototype system, Quill, which was written in 3,000 lines of Python source code. Our default implementation supports the 5 utility (from 4 downstream tasks) and 3 semantic measures in Section 5. Quill allows users to define any measures of their interest. Our data and code are in [19]. We empirically demonstrate the following claims about our method:

(1) Quill discovers binning partitions that are *semantically meaningful* while achieving *high downstream utility*, tracing the Pareto trade-off explicitly.
(2) Quill treats discretizers, utilities, and semantic metrics as plug-ins (11+ off-the-shelf discretizers and LLM suggestions; multiple task utilities). Across **14 dataset × utility cases** and **three semantic measures**, Quill consistently returns partitions on or near the Pareto frontier.
(3) Each component of Quill contributes meaningfully.
(4) Quill is efficient and can scale to large datasets.

6.1 Setup

6.1.1 Evaluation Metrics. We assess (i) accuracy of the estimated Pareto frontier and (ii) end-to-end efficiency.

Accuracy. Given the true frontier $\mathcal{P}^*$ and the estimated frontier $\hat{\mathcal{P}}$, we use a normalized **averaged Hausdorff distance** [49]:

$$\Delta(\hat{\mathcal{P}}, \mathcal{P}^*) = \frac{1}{Z} \max\{\text{GD}(\hat{\mathcal{P}}, \mathcal{P}^*),\ \text{IGD}(\hat{\mathcal{P}}, \mathcal{P}^*)\}, \tag{6}$$

where $Z = \sqrt{2}$ since utility and semantics are in $[0, 1]$. GD (Generational Distance) is the mean distance from each $\hat{p} \in \hat{\mathcal{P}}$ to its nearest neighbor in $\mathcal{P}^*$; IGD (Inverted Generational Distance) is the mean distance from each $p \in \mathcal{P}^*$ to its nearest neighbor in $\hat{\mathcal{P}}$. Taking max captures both *missing regions* and *spurious predictions*. Smaller Δ is better; $\Delta = 0$ is perfect.

Table 2. Dataset statistics.

Dataset	Description	#Tuples	$\lvert\mathbb{A}\rvert$	$\lvert\mathbb{A}_{\text{num}}\rvert$	Size (MB)
TITANIC [6]	Passenger data	0.9K	10	4	< 1
BANK [2]	Customer loan status	100K	19	10	18.8
CROP [3]	Agriculture variables	2.2K	8	7	< 1
LALONDE [28]	Causal effect of training	0.4k	12	3	< 1
PIMA [4]	Diabetes predictors	0.8K	9	8	< 1
SONGS [1]	Properties of songs	19K	15	13	2.2
TAXI [5]	Details of taxi trips	8.5M	5	4	265

Efficiency metrics. We report (i) wall-clock runtime and (ii) budget, which is the number of the candidates evaluated. Runtime captures end-to-end cost but depends on external components (e.g., model training), whereas percentage explored is a model-agnostic proxy for intrinsic search efficiency. We present both for fair comparison.

Table 3. Benchmark task–dataset mapping. "time-out" means that the brute-force computation of the true Pareto frontier exceeded a cap of 1500 hours. "n/a" means that the dataset–task pair excluded due to *<3 Pareto points*. "–" means that the dataset is not applicable to the task.

Dataset	Avg. #B	Total #**B**	Modeling	Imputation	Causal	Visual
TITANIC [6]	71	339,300	✓	✓	✓	✓
BANK [2]	74	28,392,912	time-out	time-out	-	✓
CROP [3]	73	27,625,536	✓	✓	-	n/a
LALONDE [28]	58	870	-	-	✓	-
PIMA [4]	73	313,820	✓	✓	-	✓
SONGS [1]	67	20,170,080	✓	-	-	✓
TAXI [5]	62	217,728	✓	✓	-	✓

6.1.2 Benchmark dataset. Table 2 summarizes dataset statistics, while Table 3 maps each dataset to the downstream tasks on which it is evaluated. Avg. #B denotes the average number of candidate partitions per numerical attribute, with bin counts ranging from 2 to 10. Total #**B** denotes the total number of candidate partition sets in the search space, computed as the Cartesian product of per-attribute candidates.

We assess accuracy by comparing each method's estimated Pareto frontier $\hat{\mathcal{P}}$ against a brute-force ground truth $\mathcal{P}^*$. Frontiers with fewer than three Pareto-optimal points provide insufficient resolution for trade-off evaluation and lead to unstable distance metrics; we therefore exclude any scenario with fewer than three Pareto points (*<3 Pareto points*). For the causal analysis task, we use the canonical LALONDE and TITANIC datasets due to their clear treatment–outcome structure. Overall, our benchmark comprises 4 visualizations, 5 modeling, 3 imputation, and 2 causal cases, ensuring that evaluations reflect each method's ability to recover meaningful and diverse trade-offs between utility and semantic quality. For each case, we evaluate both LLM-based and distance-based (L2 and KL) semantic measures, as described in Section 5.1.

6.1.3 Methods Compared. As the OPS problem lacks established baselines, we compare QUILL against naive heuristics, ablations of our framework, related methods, and LLM-based approaches.

- **QUILL**: Our framework with a hyperparameter auto-tuner; for single-attribute cases, we use the UCB-guided MCC variant.
- **UCB-I (Independent UCB)** [10]: Applies UCB independently to each attribute to obtain high-quality per-attribute candidate sets; evaluates joint objective $\mathcal{J}$ over the Cartesian product.
- **UCB-S (Sequential UCB)** [10]: Orders attributes and applies UCB sequentially. After fixing B_i for A_i, evaluation for A_{i+1} conditions on the current partial selection; $\mathcal{J}$ is recomputed as partitions are added, capturing inter-attribute interactions. We enumerate all attribute orderings and report the average.

Table 4. Average error $\Delta(\hat{\mathcal{P}}, \mathcal{P}^*)$ across datasets (lower is better). "–" means the method is not applicable.

Method	Single-attr.	Multi-attr.			
	Visual	Modeling	Imputation	Causal	**Overall**
LLM-based semantic measure					
Quill	**0.04**	**0.10**	**0.08**	**0.16**	**0.11**
UCB-I	**0.04**	0.28	0.17	0.42	0.28
UCB-S	**0.04**	0.46	0.23	0.39	0.38
Random	0.17	0.44	0.47	0.45	0.45
TF	–	0.29	–	–	n/a
LLM	0.23	0.30	0.32	0.48	0.34
Distance-based semantic measure (L2 norm)					
Quill	**0.04**	**0.04**	0.17	**0.03**	**0.08**
UCB-I	**0.04**	0.07	0.15	0.09	0.10
UCB-S	**0.04**	0.09	0.15	0.18	0.13
Random	0.40	0.09	**0.07**	0.38	0.15
TF	–	0.18	–	–	n/a
LLM	0.40	0.38	0.24	0.45	0.35
Distance-based semantic measure (KL-divergence)					
Quill	**0.03**	**0.03**	**0.02**	**0.05**	**0.03**
UCB-I	**0.03**	0.06	0.07	0.08	0.07
UCB-S	**0.03**	0.07	0.07	0.19	0.09
Random	0.50	0.05	0.06	0.39	0.12
TF	–	0.23	–	–	n/a
LLM	0.27	0.26	0.38	0.35	0.31

- **Random**: Estimates the Pareto frontier via uniform random sampling from the space of candidate partition sets.
- **TreeFARMS (TF)** [60]: Enumerates near-optimal decision trees; applied to modeling tasks and post-process the resulting tree sets to estimate the Pareto frontier. We cap max runtime to 2 hours.
- **LLM**. We use GPT-4o [41] with in-context prompting as a transformation baseline, testing prompt variants (few-shot, chain-of-thought [57], and RAG [30]). For each downstream task, the prompt specifies the utility and semantic measures; the LLM then proposes a set of partition sets $\hat{\mathcal{B}}$ it estimates to lie on the Pareto frontier. This is done *without* enforcing an evaluation budget.

Budget-time trade-off. UCB-I inherently explores a limited subset of candidates, while Quill, UCB-S, and Random are budget-controlled. To ensure comparability, we evaluate under a *matched budget* equal to the average number of candidates explored by UCB-I (222 candidates) for multi-attribute and 30 for single-attribute, unless stated otherwise. We note a trade-off between budget and runtime: higher budgets can improve accuracy at the cost of longer runtime. We study this trade-off in Section 6.2.2.

6.2 Experimental Results

6.2.1 Quality Comparison.

We evaluate quality by the average Hausdorff distance to the ground-truth Pareto frontier, $\Delta(\hat{\mathcal{P}}, \mathcal{P}^*)$ (lower is better).

Varying semantic scoring. Table 4 shows that Quill is consistently best across semantic measures. For Visual (single-attribute), the UCB methods match Quill exactly, which empirically aligns with our single-attribute proof in Section 4.5—when there is only one attribute, UCB and Quill learn the same policy to select clusters. In multi-attribute settings, however, the methods diverge: Quill

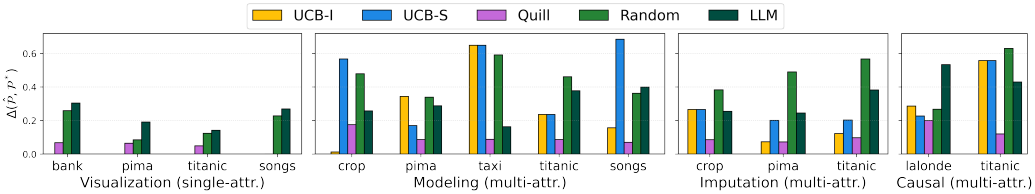


Fig. 4. Average error by downstream tasks and datasets (LLM-based semantics); UCB variants omitted for single-attribute.

jointly optimizes partitions for all attributes, whereas UCB variants explore only a limited subset of combinations.

With the *LLM-based* semantic measure, Quill achieves the lowest overall error (0.11), reducing error by 61% vs. UCB-I (0.28), 71% vs. UCB-S (0.38), 76% vs. Random (0.45), and 68% vs. LLM (0.34). With *distance-based* measures, since we need to supply a gold partition for each attribute, we take the top-ranked partitions under the LLM-based semantic measure as the reference ("gold") and compute distances to these partitions, to quantify the interpretability of the candidates. We use this proxy because those high-ranked partitions are most frequently attested in literature or code and most align with human intuition (Section 5.1), and thus plausibly reflect domain expertise and practical utility. The pattern still holds: under L2, Quill attains 0.08 overall (20% vs. UCB-I at 0.10; 39% vs. UCB-S at 0.13; 47% vs. Random at 0.15; 77% vs. LLM at 0.35). Under KL, Quill reaches 0.03 overall (57% vs. UCB-I at 0.07; 67% vs. UCB-S at 0.09; 75% vs. Random at 0.12; 90% vs. LLM at 0.31).

We note one exception in Table 4, where the random baseline slightly outperforms Quill on imputation under L2. Quill relies on a smoothness assumption: partitions with similar empirical distributions tend to have similar objective values, enabling clustering-based generalization. In this setting, the objective surface is noisier with respect to distribution vectors, making cluster representatives less predictive of other partitions and reducing the benefit of cluster-guided exploration under a limited budget. As a result, random exploration can occasionally identify a high-scoring partition earlier. This behavior is empirically rare: across all other datasets, tasks, and semantic measures, Quill consistently outperforms both random and UCB-based baselines.

Overall, while UCB matches Quill on single-attribute visualization (as expected), Quill dominates in the multi-attribute regime across metrics at a matched budget, reflecting its ability to more accurately estimate the Pareto frontier when attribute interactions matter. Note that for LLM, we extensively experimented with 8 prompting strategies to optimize the prompt for GPT-4o, including using few-shot vs. zero-shot, using COT [57] vs. no-COT, and the use of RAG [30] vs. no-RAG. We report the best prompt for LLM, which uses few-shot with COT, where few-shot examples provide demonstrations, with COT further enhancing structured reasoning. Despite extensive optimizations, LLM still lags behind. Details of our results with different prompt strategies can be found in [19].

Varying datasets. Figure 4 breaks results down by task–dataset pairs (LLM-based semantics). In nearly every panel, Quill is the lowest bar, showing small, stable errors across datasets. In Visualization, Quill is uniformly best across BANK, PIMA, TITANIC, and SONGS; in Causal, it leads on both LALONDE and TITANIC. Overall, the per-dataset breakdown indicates that Quill's gains are systematic (not driven by a few datasets) and that its performance is less sensitive to dataset idiosyncrasies than the baselines, yielding more reliable Pareto-frontier estimates across tasks.

6.2.2 *Scalability Analysis.*

We perform a sensitivity analysis to understand how the performance of Quill changes with different parameters (Figure 5).

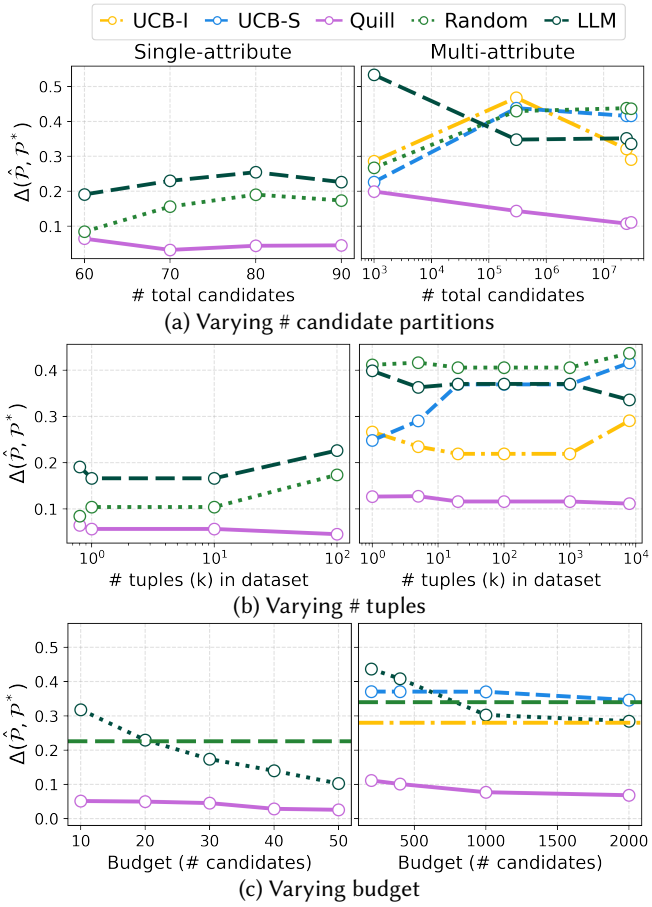


Fig. 5. Sensitivity analysis (LLM-based semantics).

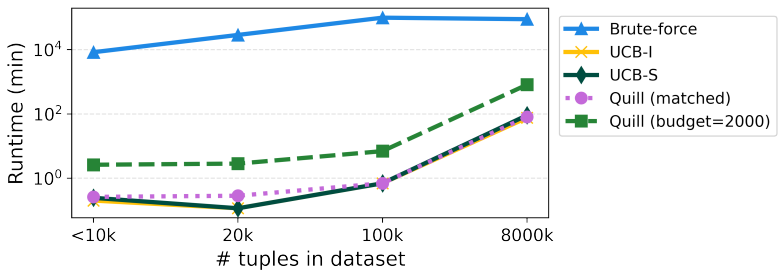


Fig. 6. Comparison of average latency. The Random baseline is omitted from the plot for clarity, as its runtime closely overlaps with the UCB variants and Quill under the matched evaluation budget.

Varying numbers of total candidates. Figure 5a shows that, in the *single-attribute* setting, Quill remains low and nearly flat as the candidate pool grows, whereas LLM and Random are consistently higher. In the *multi-attribute* setting, scaling the total candidates from 10^3 to 10^7 affects baselines much more than Quill. This indicates that Quill's cluster-based policy generalizes across larger search spaces without loss in estimation accuracy.

Varying data sizes. With more tuples (Figure 5b), both single- and multi-attribute panels show Quill achieving the lowest error across the entire range. UCB-I/S are especially sensitive in the

Table 5. Ablation studies of Quill (budget = 2000).

	Full	No auto-tuning	No DBSCAN
Visual	**0.04**	0.10	0.08
Modeling	**0.06**	0.08	0.21
Imputation	**0.05**	0.07	0.13
Causal	**0.12**	0.08	0.09
Overall	**0.06**	0.08	0.14

multi-attribute case as the data scale increases. Overall, Quill is the least sensitive to dataset size, suggesting robust Pareto-frontier estimation as dataset sizes grow.

Varying budgets. Increasing the evaluation budget monotonically improves quality (Figure 5c), with Quill clearly maintaining the best performance across the budget range. Baselines either lack a budget knob (LLM, UCB-I/S) or improve far less with budget (Random), leaving a consistent gap to Quill. Returns for Quill are diminishing but steady—higher budgets further reduce error, reflecting effective use of additional exploration.

6.2.3 *Ablation Studies.*

We ablate two design choices of Quill (Table 5).

No hyperparameter auto-tuner. Replacing the auto-tuned configuration with a fixed, hand-tuned setting (fit on 2 datasets, detail in Section 4.6) degrades the *overall* error from **0.06** to **0.08** (+33%). The drop is visible across *Visual*, *Modeling*, and *Imputation*. The only exception is *Causal* (0.12→0.08), suggesting that a single global setting can occasionally align with a specific task, but the auto-tuner is more reliable across tasks and datasets.

No DBSCAN (hierarchical instead). Substituting DBSCAN with hierarchical clustering [39] substantially hurts performance (overall **0.14**, more than ×2 worse than full Quill). The largest regression appears in *Modeling* (0.06→0.21). This indicates that density-based clustering better captures the structure of the candidate space than purely linkage-based grouping, which is crucial for forming semantically coherent, utility-relevant clusters, enabling the lowest overall error.

6.2.4 *Efficiency Comparison.*

We benchmark search latency by bucketizing datasets by size and averaging runtimes per bin (Figure 6). As input size grows, the wall-clock time *per* candidate rises (e.g., fitting models on larger data), amplifying brute-force costs. In contrast, Quill (matched) strategically evaluates only ∼200 candidates, versus millions in the full space (Table 2)—a 10^3–$10^5\times$ reduction (median $\approx 1.7 \times 10^3$). Consequently, for datasets under 100k tuples, Quill (matched) completes in under a minute on average, whereas brute force exceeds 16 hours even for < 10k tuples, underscoring the efficiency gains of our RL-based search.

Figure 6 also reports runtimes for all baseline methods under a matched evaluation budget (∼200 candidates). Under this setting, Random, UCB-I, UCB-S, and Quill exhibit very similar wall-clock runtimes, since all methods perform the same number of expensive utility evaluations. Relative to Random, Quill incurs a small additional overhead due to the initial clustering step (under one second on average), while its runtime is effectively identical to the UCB variants, which share the same clustering stage and differ only in their exploration policies.

6.3 Deeper Dive: Case Analysis

We analyze a representative run on Pima where the goal is to bin `Age`, `BMI`, and `Glucose` for a tree-based diabetes predictor (recall Example 2). Our framework returns multiple partition sets on the Pareto frontier, each trading semantic interpretability for predictive utility. For brevity, we highlight four illustrative partition sets from this run (the full output contains six).

Table 6. Representative partition sets illustrating the utility–semantics trade-off.

Set	Util.	Sem.	Partitions (BMI | Glucose | Age)	Notes
A (high sem.)	0.70	1.00	**BMI:** [0, 18.5, 25, 30, 68] **Glucose:** [0, 140, 200] **Age:** [0, 18, 35, 45, 65, 100]	WHO cutoffs Clinical thresholds Life stages
B (balanced)	0.71	0.83	**BMI:** [0, 17, 34, 51, 68] **Glucose:** [0, 140, 200] **Age:** [0, 25, 50, 75, 100]	Coarse equal steps Clinical thresholds Broad quartiles/decades
C (balanced, coarse)	0.73	0.58	**BMI:** [0, 35, 68] **Glucose:** [0, 140, 200] **Age:** [0, 20, 40, 60, 80, 100]	Coarse, WHO-aligned Clinical thresholds Broad decades
D (max utility)	0.81	0.06	**BMI:** [0, 7.46, 14.91, 22.37, 29.82, 37.28, 44.73, 52.19, 59.64, 67.10] **Glucose:** [0, 99, 157, 199] **Age:** [0, 29, 60]	Near-uniform slices Nonstandard Very coarse

Observations. Table 6 traces a clear Pareto frontier: (A) *high semantics*—clinically meaningful and readable with strong utility (0.70/1.00); (B) *balanced*—still intelligible with slightly improved utility (0.71/0.83); (C) *balanced, coarse*—further utility gain with reduced semantic alignment (0.73/0.58); and (D) *max utility*—highest accuracy (0.81) at minimal interpretability (0.06). Our framework (i) recovers domain-relevant partitions when they exist and (ii) surfaces nearby, higher-utility options so practitioners can select the point that best matches their interpretability–utility preference.

7 Conclusions and Future Work

Observing the need to identify partitions that are both useful and interpretable, we propose and formalize a new Optimal Partition Sets problem that has been overlooked thus far, design a robust, modular framework, with RL-based algorithms that jointly optimize both objectives. Our framework has several limitations that suggest future work. (1) *Data cleanliness.* We assume clean inputs; outliers and missingness can skew counts and splits. Robust statistics and integrated imputation (with uncertainty) would improve reliability. (2) *Utility specification.* Utility is user-defined and optimized one metric per run. Many settings are multi-objective (accuracy, fairness, calibration, cost); extending to vector utilities and constraints is needed. (3) *Semantic subjectivity.* Interpretability depends on human and domain norms. We expose a Pareto frontier, but learned, domain-adapted semantic models (with uncertainty) could raise fidelity. (4) *Human-in-the-loop final decision.* Users still pick among Pareto candidates. Added decision support (provenance, counterfactuals, stability analyses) would further ease selection.

Our future work includes (i) *LLM-as-agent exploration*, where we replace exploration policies with an LLM-driven agent that consumes utility/semantic feedback, reasons over candidates, and adaptively decides what to sample next (active, budget-aware exploration); and (ii) *Parallelization and result sharing to cut latency*, especially on large datasets and expensive downstream tasks.

Acknowledgments

This work was supported by the ARPA-H Biomedical Data Fabric project, the DARPA ASKEM Award HR00112220042, and grants from Liberty Mutual and Jane Street. Additional support was provided by Amazon, Google, and Intel through the MIT Data Systems and AI Lab (DSAIL) at MIT. Brit Youngmann was supported by the United States–Israel Binational Science Foundation (Grant No. 2024101) and the Israel Science Foundation (Grant No. 934/25). Sainyam Galhotra was supported by the U.S. Army Research Office (ARO) under grant W911NF-25-1-0254, the United States–Israel Binational Science Foundation (Grant No. 2024101), and a grant from InfoSys.

References

[1] [n. d.]. 19000 Spotify Songs Dataset. https://www.kaggle.com/datasets/edalrami/19000-spotify-songs. Accessed: 2025-06-29.

[2] [n. d.]. Credit Train Dataset. https://www.kaggle.com/datasets/zaurbegiev/my-dataset/data?select=credit_train.csv. Accessed: 2025-06-29.

[3] [n. d.]. Crop Recommendation Dataset. https://www.kaggle.com/datasets/atharvaingle/crop-recommendation-dataset/data. Accessed: 2025-06-29.

[4] [n. d.]. Pima Indians Diabetes Database. https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database. Accessed: 2025-06-29.

[5] [n. d.]. Taxi Trip Fare Data 2023. https://www.kaggle.com/datasets/hrish4/taxi-trip-fare-data-2023. Accessed: 2025-06-29.

[6] [n. d.]. Titanic Dataset. https://www.kaggle.com/datasets/yasserh/titanic-dataset. Accessed: 2025-06-29.

[7] 2025. Prolific — Online research participant recruitment platform. https://www.prolific.com/.

[8] Perplexity AI. 2021. https://www.perplexity.ai/. Accessed: January 12, 2025.

[9] American Diabetes Association. 2025. Diabetes Diagnosis & Tests. https://diabetes.org/about-diabetes/diagnosis. Accessed: 2026-01-15.

[10] P Auer. 2002. Finite-time Analysis of the Multiarmed Bandit Problem.

[11] Jesús Cerquides and Ramon López De Màntaras. 1997. Proposal and Empirical Comparison of a Parallelizable Distance-Based Discretization Method.. In *KDD*. 139–142.

[12] I. Csiszar. 1975. *I*-Divergence Geometry of Probability Distributions and Minimization Problems. *The Annals of Probability* 3, 1 (1975), 146 – 158. doi:10.1214/aop/1176996454

[13] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[14] Eurostat and UNESCO Institute for Statistics. 2015. *ISCED 2011 Operational Manual Guidelines for Classifying National Education Programmes and Related Qualifications: Guidelines for Classifying National Education Programmes and Related Qualifications*. OECD publishing.

[15] Usama M Fayyad and Keki B Irani. 1993. Multi-interval discretization of continuous-valued attributes for classification learning. In *Ijcai*, Vol. 93. 1022–1029.

[16] Matthias Feurer, Aaron Klein, Katharina Eggensperger, Jost Springenberg, Manuel Blum, and Frank Hutter. 2015. Efficient and robust automated machine learning. *Advances in neural information processing systems* 28 (2015).

[17] Salvador Garcia, Julian Luengo, José Antonio Sáez, Victoria Lopez, and Francisco Herrera. 2012. A survey of discretization techniques: Taxonomy and empirical analysis in supervised learning. *TKDE* (2012), 734–750.

[18] Pieter Gijsbers and Joaquin Vanschoren. 2021. *GAMA: A General Automated Machine Learning Assistant*. Springer International Publishing, 560–564. doi:10.1007/978-3-030-67670-4_39

[19] GitHub repository. 2025. Quill. https://github.com/ey-l/quill/blob/main/full_technical_report.pdf

[20] Jan Hauke and Tomasz Kossowski. 2011. Comparison of values of Pearson's and Spearman's correlation coefficients on the same sets of data. *Quaestiones geographicae* (2011), 87–93.

[21] Yuval Heffetz, Roman Vainstein, Gilad Katz, and Lior Rokach. 2019. DeepLine: AutoML Tool for Pipelines Generation using Deep Reinforcement Learning and Hierarchical Actions Filtering. arXiv:1911.00061 [cs.LG] https://arxiv.org/abs/1911.00061

[22] Miguel A Hernán and James M Robins. 2020. *Causal Inference: What If*. Chapman & Hall/CRC. Available at https://www.hsph.harvard.edu/miguel-hernan/causal-inference-book/.

[23] Kosuke Imai, Gary King, and Elizabeth A Stuart. 2008. Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society Series A: Statistics in Society* 171, 2 (2008), 481–502.

[24] Guido W Imbens and Donald B Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press.

[25] Randy Kerber. 1992. Chimerge: Discretization of numeric attributes. In *AAAI*. 123–128.

[26] Klaus Krippendorff. 2011. Computing Krippendorff's alpha-reliability. (2011).

[27] Himabindu Lakkaraju, Stephen H Bach, and Jure Leskovec. 2016. Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1675–1684.

[28] Robert J. LaLonde. 1986. Evaluating the econometric evaluations of training programs with experimental data. *The American Economic Review* 76, 4 (1986), 604–620.

[29] Benjamin Letham, Cynthia Rudin, Tyler H McCormick, and David Madigan. 2015. Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model. *The Annals of Applied Statistics* 9, 3 (2015), 1350–1371.

[30] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems* 33 (2020), 9459–9474.
[31] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems* 30 (2017).
[32] Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.
[33] Cian P McCarthy, Kazem Rahimi, and John W McEvoy. 2024. Age-stratified risk categories for cardiovascular disease prevention therapies. *JAMA cardiology* 9, 12 (2024), 1069–1070.
[34] Morris L Medley. 1980. Life satisfaction across four stages of adult life. *The International Journal of Aging and Human Development* (1980), 193–209.
[35] S. Mehta, S. Parthasarathy, and Hui Yang. 2005. Toward unsupervised correlation preserving discretization. *IEEE Transactions on Knowledge and Data Engineering* 17, 9 (2005), 1174–1185. doi:10.1109/TKDE.2005.153
[36] Kaisa Miettinen. 1999. *Nonlinear multiobjective optimization*. Vol. 12. Springer Science & Business Media.
[37] Stefano Monti and Gregory F Cooper. 1999. A latent variable model for multivariate discretization.. In *AISTATS*.
[38] Dominik Moritz, Chenglong Wang, Greg L Nelson, Halden Lin, Adam M Smith, Bill Howe, and Jeffrey Heer. 2018. Formalizing visualization design knowledge as constraints: Actionable and extensible models in draco. *IEEE transactions on visualization and computer graphics* 25, 1 (2018), 438–448.
[39] Daniel Müllner. 2011. Modern hierarchical, agglomerative clustering algorithms. arXiv:1109.2378 [stat.ML] https://arxiv.org/abs/1109.2378
[40] National Cancer Institute. 2020. Breast Cancer Risk in American Women. https://www.cancer.gov/types/breast/risk-fact-sheet SEER-based, data from 2015–2017, posted April 2020.
[41] OpenAI. 2024. GPT-4o. arXiv:2410.21276 [cs.CL] https://arxiv.org/abs/2410.21276
[42] Judea Pearl. 2009. Causal inference in statistics: An overview. (2009).
[43] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1135–1144.
[44] Paul R Rosenbaum and Donald B Rubin. 1984. Reducing bias in observational studies using subclassification on the propensity score. *J. Amer. Statist. Assoc.* 79, 387 (1984), 516–524.
[45] Peter J Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* 20 (1987), 53–65.
[46] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. 2000. The earth mover's distance as a metric for image retrieval. *International journal of computer vision* 40 (2000), 99–121.
[47] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1, 5 (2019), 206–215. doi:10.1038/s42256-019-0048-x
[48] Erich Schubert, Jörg Sander, Martin Ester, Hans Peter Kriegel, and Xiaowei Xu. 2017. DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. *ACM Transactions on Database Systems (TODS)* 42, 3 (2017), 1–21.
[49] Oliver Schütze, Xavier Esquivel, Adriana Lara, and Carlos A. Coello Coello. 2012. Using the Averaged Hausdorff Distance as a Performance Measure in Evolutionary Multiobjective Optimization. *IEEE Trans. Evol. Comput.* 16, 4 (2012), 504–522. https://doi.org/10.1109/TEVC.2011.2161872
[50] scikit-learn developers. 2025. *sklearn.impute.KNNImputer*. scikit-learn. https://scikit-learn.org/stable/modules/generated/sklearn.impute.KNNImputer.html Version 1.7.1.
[51] Jinwook Seo and Ben Shneiderman. 2005. A rank-by-feature framework for interactive exploration of multidimensional data. *Information visualization* 4, 2 (2005), 96–113.
[52] Vidya Setlur, Michael Correll, and Sarah Battersby. 2022. Oscar: A semantic-based data binning approach. In *IEEE VIS*. IEEE, 100–104.
[53] Fact sheet No. 2011. 3-Water in the atmosphere. https://web.archive.org/web/20120114162401/http://www.metoffice.gov.uk/media/pdf/4/1/No._03_-_Water_in_the_Atmosphere.pdf. *National Meteorological Library and Archive, Met Office, UK*. (2011).
[54] Charles Spearman. 1961. The proof and measurement of association between two things. (1961).
[55] Richard S Sutton, Andrew G Barto, et al. 1998. *Reinforcement learning: An introduction*. Vol. 1. MIT press Cambridge.
[56] Manasi Vartak, Sajjadur Rahman, Samuel Madden, Aditya Parameswaran, and Neoklis Polyzotis. 2015. Seedb: Efficient data-driven visualization recommendations to support visual analytics. In *PVLDB*. NIH Public Access.
[57] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems* 35 (2022), 24824–24837.

[58] Graham Wills and Leland Wilkinson. 2010. Autovis: automatic visualization. *Information Visualization* 9, 1 (2010), 47–69.
[59] Kanit Wongsuphasawat, Dominik Moritz, Anushka Anand, Jock Mackinlay, Bill Howe, and Jeffrey Heer. 2015. Voyager: Exploratory analysis via faceted browsing of visualization recommendations. *IEEE transactions on visualization and computer graphics* 22, 1 (2015), 649–658.
[60] Rui Xin, Chudi Zhong, Zhi Chen, Takuya Takagi, Margo Seltzer, and Cynthia Rudin. 2022. Exploring the whole rashomon set of sparse decision trees. *Advances in neural information processing systems* 35 (2022), 14071–14084.
[61] Hongyu Yang, Cynthia Rudin, and Margo Seltzer. 2017. Scalable Bayesian rule lists. In *International conference on machine learning*. PMLR, 3921–3930.

Received October 2025; revised January 2026; accepted February 2026