

Towards Inference-Aware Privacy Guidance for Data Preparation

Vishal Chakraborty
University of California, Irvine
USA
vchakrab@uci.edu

Felix Naumann
Hasso Plattner Institute, University of Potsdam
Germany
felix.naumann@hpi.de

ABSTRACT

Data preparation often begins with sensitive data and produces a releasable artifact for analysis, sharing, or model training. Existing workflows are primarily guided by utility: a curator drops attributes, coarsens values, filters populations, and suppresses tuples until the resulting dataset appears useful and safe. Privacy, when considered, is usually evaluated only on the final release.

We propose privacy-aware data preparation as an interactive guidance problem. We model a preparation plan as a sequence of deterministic curation operators and ask how each step changes the evidence available to an observer with prior knowledge about a target. Our semantics is based on compatibility sets, which capture the source tuples still plausible for the target after a released representation is observed. This view separates operators that remove evidence from those that remove ambiguity, explains why privacy effects can be non-monotone, and supports prefix-level feedback under a disclosure budget. The result is an inference-aware foundation for guiding curators throughout data preparation, rather than judging privacy only after the final artifact is produced. We conclude by identifying the key challenges in building interactive, inference-aware data preparation systems.

VLDB Workshop Reference Format:

Vishal Chakraborty and Felix Naumann. Towards Inference-Aware Privacy Guidance for Data Preparation. VLDB 2026 Workshop: 15th International Workshop on Quality in Databases (QDB'26).

1 INTRODUCTION

Data preparation/curation has become a fundamental database workflow: transforming a "raw" dataset, often containing private or sensitive information, into a releasable artifact for analysis, sharing, model training, or public use [4, 14, 19]. This transformation is rarely a single operation. A curator drops attributes, coarsens values, selects populations, suppresses tuples, and repeats these choices until the release appears useful and safe. This operator view is central to interactive data preparation systems such as Potter's Wheel and Wrangler [9, 18], to data-transformation-by-example systems [8], and to recent pipeline search and preparation frameworks [4, 14, 24]. In these settings, curation choices are primarily guided by utility: which attributes are needed, which values should be generalized, and which records should remain in scope. Privacy, when considered, is usually evaluated on the resulting release rather than on the sequence of operators that produced it.

This work is licensed under the Creative Commons BY-NC-ND 4.0 International License. Visit <https://creativecommons.org/licenses/by-nc-nd/4.0/> to view a copy of this license. For any use beyond those covered by this license, obtain permission by emailing info@vldb.org. Copyright is held by the owner/author(s). Publication rights licensed to the VLDB Endowment.
Proceedings of the VLDB Endowment. ISSN 2150-8097.

The same operators that shape utility also determine what evidence remains in the released artifact and how that evidence changes across curation steps. Dropping a sensitive column may not remove evidence for the sensitive fact, because retained attributes may still make the fact predictable. This concern is central to statistical disclosure control and anonymization, where releases may remain identifying or revealing even after explicit identifiers are removed [21, 23]. Refinements such as ℓ -diversity and t -closeness address the fact that indistinguishability alone may still leave sensitive values predictable within an equivalence class [13, 15]. More general privacy models and de-anonymization attacks show that auxiliary information and correlations can make a release revealing even when the sensitive attribute is absent [11, 17]. Conversely, deleting a tuple may make a release more revealing by removing the record that made the target ambiguous. Thus privacy is not only a constraint to check after a useful release is found. A utility-driven search may miss another feasible plan that is both more useful and less revealing, because disclosure depends on operator-induced comparison classes, correlations, and ambiguity.

Example 1.1 (Running example). Consider the curation steps in Fig. 1(a)–(d). A curator preparing the data for a treatment-pattern study begins with the source relation R in Fig. 1a, first drops *Diag*, bucketizes *Age*, and, finally, shortens *ZIP*, producing R_1 (Fig. 1b). The release appears privacy-conscious because the *diagnosis* column is absent. Yet an observer who knows that the target is a woman in her thirties in ZIP prefix 926 sees that the matching released records contain two tuples with *Med*=ART and one with *Med*=None. If ART is strongly associated with HIV, then R_1 still supports the sensitive inference.

Now consider two single-operator refinements. In R_2 (Fig. 1c), the curator also drops *Med*. This removes the visible indicator and restores the target to the larger prior group. In R_3 (Fig. 1d), the curator instead keeps *Med* but deletes t_3 , the only matching tuple with *Med*=None. This makes the release smaller, but not necessarily safer: it removes ambiguity and leaves only ART records in the target's visible profile. Thus two ordinary curation operators that look comparable from a data preparation perspective can have different privacy effects.

Example 1.1 illustrates the central difficulty. Privacy under curation is not determined only by which sensitive attributes are removed from the final release. It depends on how each operator changes the observer's evidence. Projection may remove evidence. Generalization may weaken evidence. Selection and tuple deletion may either reduce or increase disclosure, depending on which tuples they remove and how those tuples affected ambiguity. A final-only privacy check cannot explain which operator caused the change, in what direction, or by how much.

pid	Age	Sex	ZIP	Med	Diagnosis
t_1	34	F	92612	ART	HIV
t_2	36	F	92617	ART	HIV
t_3	35	F	92612	None	Flu
t_4	61	M	90024	Ins.	Diab.

(a) Source R .

Bkt	Sex	ZIP3	Med
30–39	F	926	ART
30–39	F	926	ART
30–39	F	926	None
60–69	M	900	Ins.

(b) R_1 : drop Diag.

Source relation and three prepared releases.

Bkt	Sex	ZIP3
30–39	F	926
30–39	F	926
30–39	F	926
60–69	M	900

(c) R_2 : drop Med.

Bkt	Sex	ZIP3	Med
30–39	F	926	ART
30–39	F	926	ART
60–69	M	900	Ins.

(d) R_3 : delete t_3 .

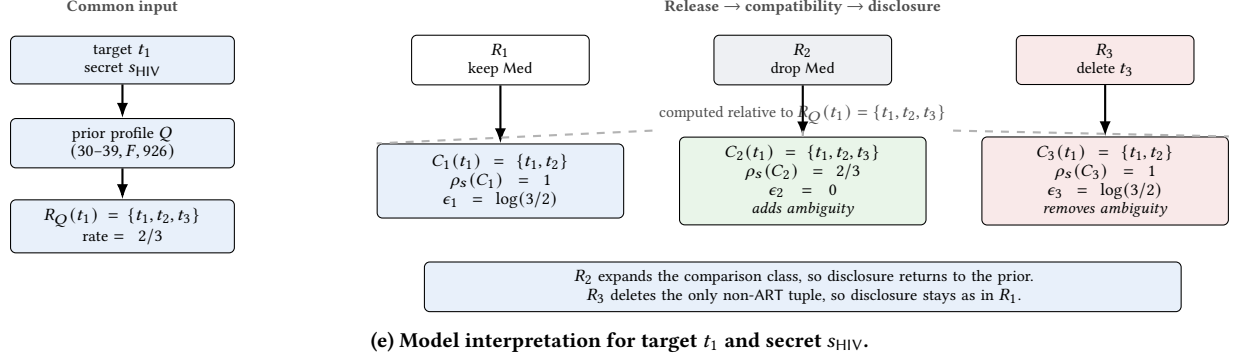


Figure 1: Running example. (a)–(d) Curation steps. (e) The prior profile identifies the comparison group $R_Q(t_1)$; each release induces a compatibility set $C_i(t_1)$; disclosure compares the secret rate in $C_i(t_1)$ against the prior rate in $R_Q(t_1)$. Gray denotes representation change, crimson tuple deletion, and green the privacy-improving outcome.

We adopt an *operator-level* view of privacy-aware data curation. Rather than asking only whether the final release satisfies a privacy test, we ask how each curation step changes the observer’s evidence about protected secrets. We model a curation plan as a sequence of deterministic operators over a source relation. The adversary observes the released data and has prior knowledge about a target, such as a visible demographic profile. This is close in spirit to privacy definitions that reason about belief change under side information [5, 7, 10, 11], but our mechanisms are ordinary deterministic curation operators rather than randomized query mechanisms.

Our semantics uses *compatibility sets*: for a target tuple, the compatibility set contains the source tuples that remain plausible after the adversary combines prior knowledge with the released representation. While alternative privacy semantics might be suitable, the compatibility sets expose the operator-level phenomenon. Column and value operators change how source tuples appear in the release. Row operators change which tuples remain present. These effects are not interchangeable. In Example 1.1, dropping Med expands the target’s compatible set and weakens the evidence for HIV; deleting t_3 removes contrary evidence and strengthens, or at least preserves, the inference. The usual vocabulary of “removing information” is therefore too coarse: what matters is whether an operator removes evidence, removes ambiguity, or changes the representation through which the target is compared to the source.

Existing work does not directly address this setting. Data preparation systems formalize, synthesize, and explain transformation sequences, but do not attach privacy semantics to individual preparation operators [4, 8, 9, 14, 18, 22, 24]. Statistical disclosure control and Bayesian privacy give semantics to released tables, but typically evaluate the release as a single object rather than as the result of a step-indexed construction process [11, 13, 15, 21, 23].

Differential privacy provides powerful composition and accounting principles, including systems for private data analysis [1, 5–7, 16, 20, 25]. These frameworks, however, account for randomized mechanisms, whereas many curation pipelines are deterministic. PrivateClean brings differential privacy to data cleaning, but still treats privacy through randomized DP mechanisms rather than through the operator-specific evidential effects of deterministic curation steps [12]. What is missing is a framework that treats deterministic curation operators themselves as privacy-relevant transformations. Our work is also related to inference-aware deletion, where the goal is to remove a value while limiting what remains inferable from dependencies in the retained database [2, 3]. In deletion, the request specifies a target value to remove. In curation, the analyst constructs a release through heterogeneous operators, and the privacy question is how those operators alter the evidence available to an observer during and after plan construction.

We study privacy-aware curation as an *interactive guidance* problem. Given a source relation, deterministic curation operators, protected secrets, and an adversary with prior knowledge about a target, how can a curator construct a useful release while keeping evidential disclosure below a budget? The challenge is that disclosure is a property of plan prefixes, not only endpoints. A release that appears safe at the end may be reached through revealing intermediate states. A plan that was safe when committed may become unsafe when the source data evolves. And a curator choosing one operator at a time needs feedback before committing to an operator, not only after the release is complete.

This paper makes three contributions.

- We define a curation and privacy framework in which formulate privacy-aware data curation as an operator-level inference problem (Sec. 2).

- We characterize the disclosure effects of common curation operators and show how they support local guidance under a disclosure budget (Sec. 3)
- We outline research questions for interactive privacy-aware curation, including per-operator semantics, prefix-level accounting, budgeted guidance, re-execution under drift, and richer disclosure models (Sec. 4).

2 CURATION PLANS AND DISCLOSURE SEMANTICS

A curation workflow transforms a source relation into a release by applying ordinary preparation operators. Let R be the source relation and let $p = o_t \circ \dots \circ o_1$ be a curation plan. Each o_i is a deterministic operator, such as projecting attributes, generalizing values, selecting a population, or deleting tuples. The prefix $p_i = o_i \circ \dots \circ o_1$ produces the intermediate release $R_i = p_i(R)$, with $R_0 = R$.

The prefix view is essential. A curator constructs the release one operator at a time, and each prefix may be inspected, committed, shared, cached, or reused as part of a later pipeline. The privacy question is therefore not only what the final release reveals. It is also how much inferential support each prefix leaves for protected facts.

Observation and reasoning. A privacy-aware curation framework has three components. The curation component describes how a prefix p_i transforms R into R_i . The observation component describes what the adversary sees. In the default regime, the adversary observes the released relation R_i and has prior knowledge about a target, but does not observe the internal operator sequence. In a plan-aware regime, the adversary also observes a description of the curation process, such as a transformation script, exclusion criteria, metadata, or intermediate artifacts. This regime matters when plans are published for reproducibility, audit, or regulatory transparency.

The reasoning component maps the adversary’s observation to disclosure about protected facts. In this paper, we instantiate this component using compatibility sets. This is a vehicle for making the operator-level problem concrete, not the only possible semantics. A dependency-explicit semantics could reason under a declared model Σ of dependencies or probabilistic rules; a Bayesian semantics could use richer priors; and a multi-release semantics could account for evidence accumulated across correlated outputs. The common requirement is that the semantics explain how each curation step changes the evidence available to the observer.

Secrets and prior visibility. A protected secret is a Boolean predicate over source tuples. For a tuple t , $s(t) = 1$ means that the protected fact holds for t . In Example 1.1, the secret

$$s_{\text{HIV}}(t) = 1 \iff t[\text{Diagnosis}] = \text{HIV}$$

captures whether the target has diagnosis HIV. Other secrets may describe membership in a cohort, the presence of a treatment class, or any application-specific property of a tuple.

The observer also has prior information about the target. We model this by a set Q of visible attributes. The observer knows $t[Q]$ before seeing the release. The corresponding prior group is

$$R_Q(t) = \{u \in R : u[Q] = t[Q]\}.$$

For E.x. 1.1, take $Q = \{\text{Bkt, Sex, ZIP3}\}$. For target t_1 , the prior group consists of t_1, t_2, t_3 : the women in their thirties in ZIP prefix 926.

Released representations. A curation prefix changes how a source tuple appears in the release. We write $\pi_i(u)$ for the representation of source tuple u after prefix p_i . Projection removes fields from this representation. Generalization replaces values by coarser values. Row operators do not change the representation itself; they change whether the tuple is present in the released relation.

Let $S_i \subseteq R$ be the source tuples that survive in R_i . The release can then be viewed as the multiset

$$\{\pi_i(u) : u \in S_i\}.$$

This separates two effects that are often conflated: representation-changing operators modify π_i , while presence-changing operators modify S_i .

Compatibility-set instantiation. Under the compatibility-set semantics, the adversary compares the target with source tuples that agree with the prior profile and have the same released representation. The compatibility set after prefix p_i is

$$C_i(t) = \{u \in S_i : u[Q] = t[Q] \text{ and } \pi_i(u) = \pi_i(t)\}.$$

This is the set of source tuples still plausible as the target under the observer’s prior information and the visible release.

Fig. 1 gives the intuition. The top panel shows the source and the three releases from Example 1.1; the bottom panel shows how the compatibility-set semantics reads the same releases for target t_1 . Keeping Med visible in R_1 compares the target only with the two ART tuples. Dropping Med in R_2 restores the larger comparison group and adds ambiguity. Deleting t_3 in R_3 does the opposite: it removes the only non-ART tuple in the target profile, so the release is smaller but not less revealing for this target.

Evidential disclosure. The compatibility set induces an empirical belief update. For $X \subseteq R$, define

$$\rho_s(X) = \frac{|X \cap s^{-1}(1)|}{|X|}.$$

The prior support for the secret is $\rho_s(R_Q(t))$. The support after observing prefix p_i is $\rho_s(C_i(t))$. We define evidential disclosure as

$$\epsilon_i(s, t) = \log \frac{\rho_s(C_i(t))}{\rho_s(R_Q(t))}.$$

A positive value means that the prefix increases support for the secret relative to the prior profile. A value of zero means that the release adds no support beyond the prior. A negative value means that the release weakens support for the secret.

In the running example, the prior group for t_1 has HIV rate $2/3$. Release R_1 raises the compatible secret rate to 1, so $\epsilon_1(s_{\text{HIV}}, t_1) = \log(3/2)$. Release R_2 returns the comparison class to the prior group, so disclosure returns to 0. Release R_3 again leaves only the ART tuples compatible with t_1 , so disclosure remains the same as under R_1 . The key point is not the arithmetic itself, but the direction of evidence: R_2 adds ambiguity, while R_3 removes it.

Dependencies as evidence. The compatibility-set instantiation does not require the curator to declare a dependency set. Dependencies enter through the secret rate inside the target’s comparison class. If Med=ART is strongly associated with Diagnosis=HIV,

then a release that leaves the target compatible only with ART records increases support for the HIV secret. Thus, the basic model treats dependencies as evidential regularities in the source relation. A dependency-explicit version can replace the empirical rate $\rho_s(C_i(t))$ by inference under a declared semantics model Σ .

Plan-level disclosure. The marginal effect of operator o_i is

$$\delta_i(s, t) = \epsilon_i(s, t) - \epsilon_{i-1}(s, t).$$

This quantity tells us whether the next curation step removed evidence, added evidence, or left the inferential support unchanged.

Problem definition. A release is evaluated over a set $\mathcal{S}$ of protected secrets and a set T of target tuples. We write

$$\alpha_i = \max_{s \in \mathcal{S}, t \in T} \epsilon_i(s, t)$$

for the disclosure of prefix p_i . A curation plan satisfies budget τ if every prefix remains below the budget:

$$\alpha_i \leq \tau \quad \text{for all } i \in \{1, \dots, \ell\}.$$

The curation problem is to find a useful release without exceeding this disclosure budget:

$$\max_{p \in \mathcal{P}} U(R_p) \quad \text{s.t.} \quad \alpha_i \leq \tau \text{ for every prefix } p_i \preceq p.$$

Here U is a task-specific utility measure, such as retained attribute value, workload accuracy, or downstream model quality. The prefix constraint captures the interactive setting: the curator needs to know what the next operator does before committing to it. The next section analyzes this marginal effect under the compatibility-set instantiation.

3 OPERATOR EFFECTS AND CURATION GUIDANCE

Under the compatibility-set instantiation, the privacy effect of a curation operator becomes local. At prefix p_i , disclosure for a target t is determined by two objects: the representation map π_i and the surviving tuple set S_i . A candidate operator can change one or both. Projection, generalization, and refinement act primarily on π_i : they change how source tuples appear in the release. Selection and deletion act primarily on S_i : they change which tuples remain present. This distinction gives one concrete explanation for why privacy under curation is non-monotone.

Fig. 1 shows this distinction. Dropping Med in R_2 changes the representation of the target profile: tuples that were separated by medication now become comparable again, so the compatibility set for t_1 expands. Deleting t_3 in R_3 changes presence instead: it removes the only non-ART tuple in the target profile. The two operations are both ordinary curation steps, and both remove information in an informal sense, but they have different effects on evidence. One adds ambiguity; the other removes it.

Fig. 2 abstracts this example into the interactive setting. At a current prefix, the curator has a set Ω_i of candidate actions. The missing primitive is not merely a final privacy test. It is a way to score each candidate before it is committed: which compatibility sets can it change, in which direction does it move disclosure, and what utility does it retain?

Representation-changing operators. If an operator changes π_i , it changes the comparison used by the observer to relate the target to source tuples. A projection or generalization merges representation classes. For a target t , the new compatible set has the form

$$C_{i+1}(t) = C_i(t) \cup A,$$

where A is the set of newly indistinguishable tuples inside the same visible profile. The new secret rate is

$$\rho_s(C_{i+1}(t)) = \frac{|C_i(t)|\rho_s(C_i(t)) + |A|\rho_s(A)}{|C_i(t)| + |A|}.$$

The direction of the privacy effect is therefore determined by the tuples that join the target's comparison class. If the added tuples have lower secret rate than $C_i(t)$, the operator weakens evidence for the secret. If they have higher secret rate, it strengthens it. Refinement has the reverse form: it splits a class and may remove the tuples that were providing ambiguity.

This is the case for R_2 in Fig. 1. Dropping Med merges the ART and None representations inside the target's visible profile. The compatible set for t_1 expands from $\{t_1, t_2\}$ to $\{t_1, t_2, t_3\}$. Since the added tuple t_3 does not satisfy s_{HIV} , the secret rate falls from 1 to $2/3$, and the disclosure returns to the prior level. The operator helps because it adds ambiguity.

Presence-changing operators. Selection and deletion behave differently. They change S_i , not the released representation itself. If a removed tuple d is not in $C_i(t)$, then the target's evidence does not change:

$$C_{i+1}(t) = C_i(t).$$

If $d \in C_i(t)$, then

$$C_{i+1}(t) = C_i(t) \setminus \{d\}.$$

The sign of the privacy effect depends on what d contributed. If $s(d) = 1$, the operator removes supporting evidence and may reduce disclosure. If $s(d) = 0$, it removes contrary evidence and may increase disclosure. Equivalently, for $C = C_i(t)$,

$$\rho_s(C \setminus \{d\}) = \frac{|C|\rho_s(C) - s(d)}{|C| - 1}.$$

Tuple deletion is therefore not inherently privacy-improving. It is protective when it removes evidence for the secret, and harmful when it removes the ambiguity that shielded the target.

This is the case for R_3 in Fig. 1. Tuple t_3 is the only matching non-ART tuple in the target's visible profile. Deleting it does not remove the visible ART evidence; it removes the tuple that made the profile mixed. The release becomes smaller, but the evidence for HIV remains concentrated on the two ART records. Hence R_3 has the same disclosure for t_1 as R_1 .

Local guidance. The same local view gives the primitive needed for interactive curation. For each candidate next operator $o \in \Omega_i$, the framework should identify the targets whose compatibility sets can change. Representation-changing operators affect targets whose classes are merged or split. Presence-changing operators affect only targets whose compatibility sets contain removed tuples. All other targets have zero marginal change under the compatibility-set semantics.

For affected targets, the candidate induces a marginal disclosure

$$\delta_{i+1}^o(s, t) = \epsilon_{i+1}^o(s, t) - \epsilon_i(s, t),$$

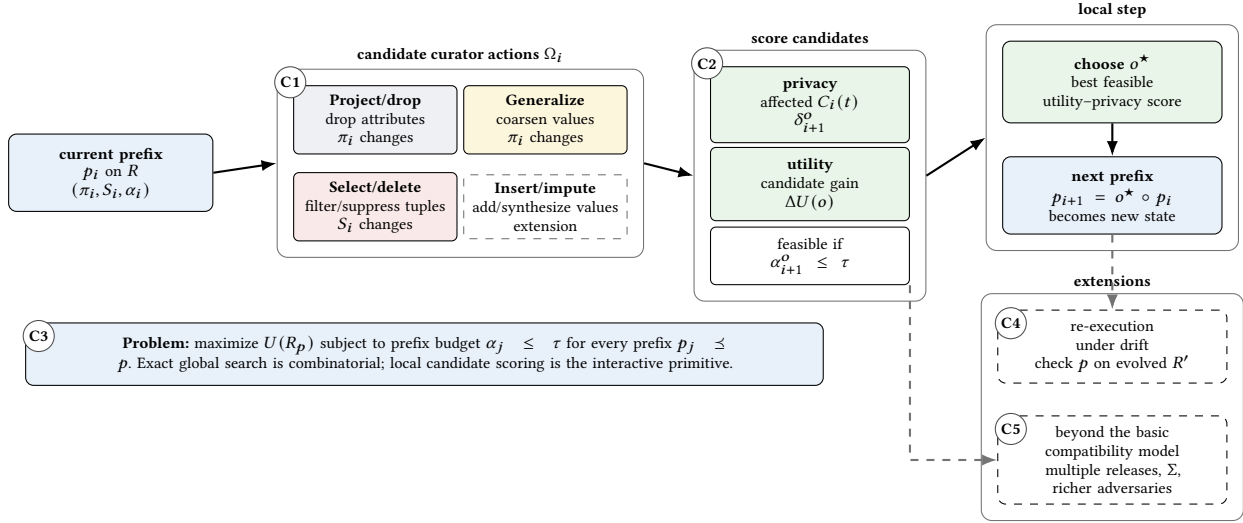


Figure 2: Interactive privacy-aware curation and its associated research challenges.

and a candidate prefix disclosure

$$\alpha_{i+1}^o = \max_{s \in S, t \in T} \epsilon_{i+1}^o(s, t).$$

The candidate is feasible only if $\alpha_{i+1}^o \leq \tau$. Among feasible candidates, the curator can compare utility gains. This gives the local primitive shown in Fig. 2: score candidate actions by privacy and utility, keep only those below the disclosure budget, and commit one action to obtain the next prefix.

Beyond rejecting unsafe final releases, the curator should see what each proposed operator does before committing to it. In the running example, the guidance loop distinguishes the two one-step refinements of R_1 : dropping Med changes the representation and adds ambiguity, while deleting t_3 changes presence and removes ambiguity. The former moves the prefix toward the disclosure budget; the latter does no privacy work for the target.

This analysis is specific to the compatibility-set instantiation, but the requirement is broader. Any disclosure semantics for privacy-aware curation should support the same operator question: given the current prefix and a candidate next step, what evidence does the step add, remove, or leave unchanged?

4 RESEARCH CHALLENGES

The compatibility-set view gives privacy-aware curation a concrete semantic object, but it also exposes why the problem is not a direct application of existing release-level privacy tests. A curation plan is not only a final table. It is a sequence of choices, and each choice changes the evidence available to an observer. Fig. 2 marks the four challenges that follow.

C1. Per-operator privacy semantics. The first challenge is to give each curation operator a privacy semantics. Existing preparation systems already treat curation as a sequence of typed transformations, but the types are usually understood operationally: project removes attributes, generalize coarsens values, select restricts the population, and delete removes tuples. For privacy, these descriptions are too coarse. In Example 1.1, both R_2 and R_3 are

one-step refinements of R_1 , and both can be described informally as “removing information.” Yet they have different privacy effects. Dropping Med in R_2 changes the representation of the target profile and merges the ART and None records, thereby adding ambiguity. Deleting t_3 in R_3 changes tuple presence and removes the only non-ART record, thereby removing ambiguity. Thus, the operator name alone does not determine the privacy effect.

What is needed is a semantics that predicts, before an operator is committed, whether it expands, shrinks, merges, or splits the evidence classes that determine disclosure. In the compatibility-set instantiation, this is the role of the representation/presence distinction from Sec. 3: column and value operators act on how tuples are represented, while row operators act on which tuples remain available as evidence.

C2. Prefix-level accounting. The second challenge is that disclosure must be accounted for over plan prefixes, not only at the endpoint. Example 1.1 already exhibits the phenomenon. Suppose the curator eventually publishes R_2 , where Med has been dropped. For target t_1 , the final disclosure is 0: the compatible set is again $\{t_1, t_2, t_3\}$, and the secret rate returns to the prior. But one natural construction path reaches R_2 through R_1 : first drop Diag, bucketize Age, and shorten ZIP; only later drop Med. Along this path, the intermediate prefix R_1 has disclosure $\log(3/2)$, because the target is compatible only with the two ART tuples. Thus, the final release may be acceptable even though the curation process passed through a more revealing state.

This matters in interactive settings, where partial results may be inspected, cached, shared with collaborators, or used to decide the next transformation. It also matters conceptually: if the privacy effect of an operator is its marginal change $\delta_i(s, t) = \epsilon_i(s, t) - \epsilon_{i-1}(s, t)$, then the process cannot be reconstructed from the endpoint alone. A privacy-aware curation system should therefore maintain a trajectory $\alpha_0, \alpha_1, \dots, \alpha_\ell$, not only a final value α_ℓ .

C3. Guidance under a disclosure budget. The third challenge is to make the semantics useful to the curator. A data scientist does not only ask whether a release is safe; she asks which operator to apply next. This turns curation into a budgeted search problem:

$$\max_{p \in \mathcal{P}} U(R_p) \quad \text{s.t.} \quad \alpha_i \leq \tau \text{ for every prefix } p_i \preceq p.$$

Exact search is unlikely to be the right abstraction for interactive curation. The plan space is combinatorial, and utility is usually task-specific. The useful primitive is local: given the current prefix and a small set of candidate operators, estimate the marginal disclosure of each candidate, rule out those that exceed the budget, and explain the remaining privacy-utility tradeoff. In Example 1.1, this is the difference between telling the curator that dropping Med removes a visible indicator and telling her that deleting t_3 does no privacy work for the target.

C4. Re-execution under drift. The fourth challenge is that curated releases are often not one-shot artifacts. The same preparation plan may be committed once and re-executed whenever the source changes. A plan that satisfied a disclosure budget at commit time may fail later because new tuples enter a visible profile, existing tuples change their representation, or correlations between visible attributes and secrets shift. This makes privacy a maintained property of the curation pipeline, not a one-time certificate. The relevant question is not only whether $p(R)$ is safe today, but whether the same plan remains safe on R' before the next release is published.

C5. Beyond compatibility sets. The compatibility-set semantics measures how the release changes the target's plausible comparison class. Several extensions are natural. Insert and impute operators introduce synthetic or externally sourced values that are not captured by pure suppression and coarsening. Multiple releases require modeling how evidence accumulates across correlated outputs. A dependency-explicit version would replace the empirical compatibility score by inference under a declared semantics model Σ , connecting privacy-aware curation to inference-aware deletion under dependencies [2, 3]. These extensions should refine, not replace, the operator-level view: the central question remains how each curation step changes the evidence that a released artifact provides about protected facts.

Together, these challenges outline the problem space of privacy-aware curation. The aim is not merely to certify a final table, but to make privacy visible during the construction of that table: which operator changed disclosure, in which direction, for which targets, and under what budget.

Acknowledgements. Chakraborty acknowledges the support of the Hasso-Plattner-Institut.

REFERENCES

- [1] Mark Bun and Thomas Steinke. 2016. Concentrated Differential Privacy: Simplifications, Extensions, and Lower Bounds. In *Proceedings of the 14th International Conference on Theory of Cryptography (TCC) (Lecture Notes in Computer Science)*, Vol. 9985. Springer, Beijing, China, 635–658.
- [2] Vishal Chakraborty, Youri Kaminsky, Arnav Abhijit Dhariya, Sharad Mehrotra, Felix Naumann, and Sarvesh Pandey. 2026. Inference-Aware and Privacy-Preserving Deletion in Databases. arXiv preprint arXiv:2604.00326. <https://arxiv.org/abs/2604.00326>
- [3] Vishal Chakraborty, Youri Kaminsky, Sharad Mehrotra, Felix Naumann, Faisal Nawab, Primal Pappachan, Mohammad Sadoghi, and Nalini Venkatasubramanian. 2025. Meaningful Data Erasure in the Presence of Dependencies. *Proc. VLDB Endow.* 18, 10 (June 2025), 3435–3448.
- [4] Sibe Chen, Nan Tang, Ju Fan, Xuemi Yan, Chengliang Chai, Guoliang Li, and Xiaoyong Du. 2023. HAIPIpe: Combining Human-generated and Machine-generated Pipelines for Data Preparation. *Proceedings of the ACM on Management of Data (PACMOD)* 1, 1 (2023), 91:1–91:26.
- [5] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating Noise to Sensitivity in Private Data Analysis. In *Proceedings of the 3rd Theory of Cryptography Conference (TCC) (Lecture Notes in Computer Science)*, Vol. 3876. Springer, New York, NY, USA, 265–284.
- [6] Cynthia Dwork and Guy N. Rothblum. 2016. Concentrated Differential Privacy. arXiv preprint arXiv:1603.01887.
- [7] Cynthia Dwork, Guy N. Rothblum, and Salil Vadhan. 2010. Boosting and Differential Privacy. In *Proceedings of the 51st Annual IEEE Symposium on Foundations of Computer Science (FOCS)*. IEEE, Las Vegas, NV, USA, 51–60.
- [8] Zhongjun Jin, Michael R. Anderson, Michael J. Cafarella, and H. V. Jagadish. 2017. Foofah: Transforming Data By Example. In *Proceedings of the 2017 ACM International Conference on Management of Data (SIGMOD)*. ACM, Chicago, IL, USA, 683–698.
- [9] Sean Kandel, Andreas Paepcke, Joseph M. Hellerstein, and Jeffrey Heer. 2011. Wrangler: Interactive Visual Specification of Data Transformation Scripts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*. ACM, Vancouver, BC, Canada, 3363–3372.
- [10] Daniel Kifer and Ashwin Machanavajjhala. 2011. No Free Lunch in Data Privacy. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data*. ACM, 193–204.
- [11] Daniel Kifer and Ashwin Machanavajjhala. 2014. Pufferfish: A Framework for Mathematical Privacy Definitions. *ACM Transactions on Database Systems (TODS)* 39, 1 (2014), 3:1–3:36.
- [12] Sanjay Krishnan, Jiannan Wang, Michael J. Franklin, Ken Goldberg, and Tim Kraska. 2016. PrivateClean: Data Cleaning and Differential Privacy. In *Proceedings of the 2016 ACM SIGMOD International Conference on Management of Data (SIGMOD)*. ACM, San Francisco, CA, USA, 937–951.
- [13] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. 2007. t -Closeness: Privacy Beyond k -Anonymity and t -Diversity. In *Proceedings of the 23rd International Conference on Data Engineering (ICDE)*. IEEE, Istanbul, Turkey, 106–115.
- [14] Peng Li, Zhiyi Chen, Xu Chu, and Kexin Rong. 2023. DiffPrep: Differentiable Data Preprocessing Pipeline Search for Learning over Tabular Data. *Proceedings of the ACM on Management of Data (PACMOD)* 1, 2 (2023), 183:1–183:26.
- [15] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkatasubramanian. 2007. t -Diversity: Privacy Beyond k -Anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 1, 1 (2007), 3:1–3:52.
- [16] Frank D. McSherry. 2009. Privacy Integrated Queries: An Extensible Platform for Privacy-Preserving Data Analysis. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data (SIGMOD)*. ACM, Providence, RI, USA, 19–30.
- [17] Arvind Narayanan and Vitaly Shmatikov. 2008. Robust De-anonymization of Large Sparse Datasets. In *Proceedings of the 2008 IEEE Symposium on Security and Privacy (S&P)*. IEEE Computer Society, 111–125.
- [18] Vijayshankar Raman and Joseph M. Hellerstein. 2001. Potter's Wheel: An Interactive Data Cleaning System. In *Proceedings of the 27th International Conference on Very Large Data Bases (VLDB)*. Morgan Kaufmann, Roma, Italy, 381–390.
- [19] Jen Rogers and Anamaria Crisan. 2023. Tracing and Visualizing Human-ML/AI Collaborative Processes through Artifacts of Data Work. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, 837:1–837:22.
- [20] Ryan M. Rogers, Aaron Roth, Jonathan Ullman, and Salil Vadhan. 2016. Privacy Odometers and Filters: Pay-as-You-Go Composition. In *Advances in Neural Information Processing Systems 29 (NeurIPS)*. Curran Associates, Barcelona, Spain.
- [21] Pierangela Samarati. 2001. Protecting Respondents' Identities in Microdata Release. *IEEE Transactions on Knowledge and Data Engineering (TKDE)* 13, 6 (2001), 1010–1027.
- [22] Roece Shraga and Renée J. Miller. 2023. Explaining Dataset Changes for Semantic Data Versioning with Explain-Da-V. *Proceedings of the VLDB Endowment (PVLDB)* 16, 6 (2023), 1587–1600.
- [23] Latanya Sweeney. 2002. k -Anonymity: A Model for Protecting Privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10, 5 (2002), 557–570.
- [24] Junwen Yang, Yeye He, and Surajit Chaudhuri. 2021. Auto-Pipeline: Synthesizing Complex Data Pipelines By-Target Using Reinforcement Learning and Search. *Proceedings of the VLDB Endowment (PVLDB)* 14, 11 (2021), 2563–2575.
- [25] Dan Zhang, Ryan McKenna, Ios Kotsogiannis, Michael Hay, Ashwin Machanavajjhala, and Gerome Miklau. 2018. EKTELO: A Framework for Defining Differentially-Private Computations. In *Proceedings of the 2018 ACM SIGMOD International Conference on Management of Data (SIGMOD)*. ACM, Houston, TX, USA, 115–130.