The Data Minimization Principle in Machine Learning

Ferdinando Fioretto [1] **Prakhar Ganesh** [2 3] **Cuong Tran** [1] **Reza Shokri** [4]

Abstract

The principle of *data minimization* aims to reduce the amount of data collected, processed or retained to minimize the potential for misuse, unauthorized access, or data breaches. Rooted in privacy-by-design principles, data minimization has been endorsed by various global data protection regulations. However, its practical implementation remains a challenge due to the lack of a rigorous formulation. This paper addresses this gap and introduces an optimization framework for data minimization based on its legal definitions. It then adapts several optimization algorithms to perform data minimization and conducts a comprehensive evaluation in terms of their compliance with minimization objectives as well as their impact on user privacy. Our analysis underscores the mismatch between the privacy expectations of data minimization and the actual privacy benefits, emphasizing the need for approaches that account for multiple facets of real-world privacy risks.

1. Introduction

As data-driven systems and machine learning (ML) applications continue to proliferate, they introduce new privacy risks, necessitating robust data protection. Among these privacy risks stands the fundamental concern of unauthorized access to sensitive information, which involves disclosure, or acquisition of sensitive information, posing a threat to individuals' privacy (Nasr et al., 2019; Thomas et al., 2022; Wairimu & Fritsch, 2022). In response, several international data protection frameworks, including the European General Data Protection Regulation (GDPR), [1] the California Privacy Rights Act (CPRA), [2] and the Brazilian General Data Protection Law (LGPD), [3] have consequently adopted *data minimization* as a key principle to mitigate these risks (Biega & Finck, 2021).

At its core, the data minimization principle requires organizations to *collect, process, and retain only personal data that is adequate, relevant, and limited to what is necessary for specified objectives* (see Table 1 for further details). It's grounded in the expectation that not all collected data is essential for the objective, in this case training an ML model (Goldsteen et al., 2021; Paul et al., 2021; Sorscher et al., 2022; Shanmugam et al., 2022), and instead contributes to a heightened risk of information leakage. However, despite its legal significance and endorsement by data protection regulations, the data minimization principle lacks an appropriate mathematical representation, one that can be applied effectively to real-world ML applications. In particular, as reviewed in Section 6, the current discourse on data minimization practices often overlooks two crucial aspects **(1)** the individualized nature of minimization (e.g., information that is unimportant for an individual may be critical for another) and **(2)** its intrinsic link to data privacy.

To overcome these limitations, this paper introduces a formal framework that recasts the data minimization principle in ML as an optimization problem while being faithful to its legal and practical aspects. It further adapts and evaluates various optimization algorithms to solve the problem of data minimization. Next, we conduct an extensive evaluation of these algorithms and explores several key characteristics of minimized datasets, such as emergent individualized minimization during optimization and compatibility with real-world privacy metrics. In particular, we seek to answer a critical question:

"Do data minimization requirements in different regulations align with the privacy expectations set by legal frameworks?"

Our evaluations reveal that the answer is, unfortunately, negative. While being an implicit intention, the requirements of data minimization are not necessarily aligned with risk of reconstruction and re-identification and thus may not provide the expected privacy protection.

Finally, to address these shortcomings, we propose simple

[*]Equal contribution [1]University of Virginia [2]McGill University [3]Mila [4]National University of Singapore. Correspondence to: Ferdinando Fioretto <fioretto@virginia.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

[1] gdpr-info.eu/ [2] cppa.ca.gov/

[3] lgpd-brazil.info/

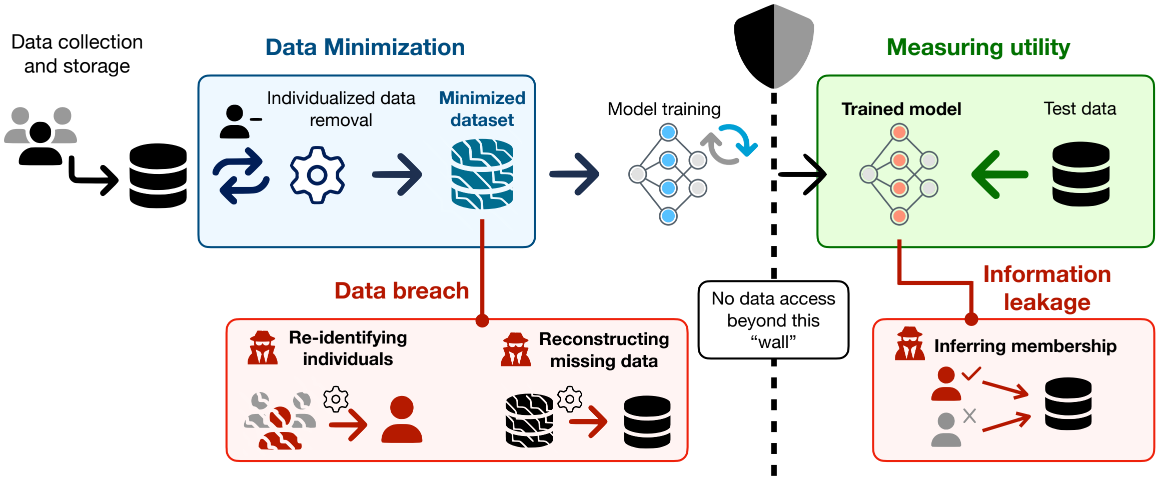

Data collection and storage
Data Minimization
Individualized data removal
Minimized dataset
Model training
Measuring utility
Trained model
Test data
Data breach
Re-identifying individuals
Reconstructing missing data
No data access beyond this "wall"
Information leakage
Inferring membership

Fig. 1: Data minimization in an ML pipeline. Behind the data access "wall", we highlight the formalization of data minimization and quantify the risks of a data breach. Outside the "wall" with no direct data access, we establish the data minimization objectives using utility measurement and study information leakage from the trained model.

yet effective modifications to data minimization algorithms that enhance their privacy-preserving capabilities, leading to far better privacy-utility trade-offs under minimization. Through these comprehensive analyses, the paper hopes to demonstrate that incorporating data minimization with appropriate modifications into ML systems not only satisfies the legal requirements but can also reduce the privacy risks of potential data breaches while retaining high accuracy.

In summary, this paper makes the following contributions:

1. It examines various data protection regulations from around the world and provides the first formalization of data minimization as an optimization problem that is in line with the legal frameworks faithfully incorporates the individualized nature of minimization.
2. It implements three classes of optimization algorithms for solving the data minimization problem in ML and introduces several threat models to empirically quantify privacy leakage within the context of data minimization.
3. Through extensive evaluations, it assesses the limitations of current regulatory requirements in meeting their inherent privacy expectations.
4. Finally, it proposes simple yet effective modifications to the data minimization algorithms to better trade-off user privacy and downstream utility.

These contributions aim to bridge the gap between the regulatory and technical domains, offering a robust solution to the data minimization challenge faced by ML practitioners. The proposed framework provides a solid groundwork for future research and a practical guide for developing privacy-preserving ML systems in compliance with the legal requirements of data minimization.

2. The data minimization principle

Prior developing the proposed framework (illustrated in Figure 1), we focus on a key question: "*How to translate regulatory laws into formal principles for data minimization in machine learning?*"

We start by inspecting the legal language of six global data protection regulations, whose language is summarized in Table 1, which enables us to discern three foundational pillars of the data minimization principle:

1. **Purpose Limitation**: Data should only be collected for a specific, legally justified purpose. In the context of machine learning, this aligns with the goal of achieving a specified task performance via model training, or retaining high performance at inference time.
2. **Data Relevance**: Regulations mandate that collected data be relevant and limited to what is necessary for the stated purpose. In ML tasks, this means striving to minimize data without affecting performance.
3. **Data Privacy**: Data protection laws define personal data as information that can identify an individual. This places an onus on data minimization to prevent any unnecessary usage of such identifiable data.

2.1. Implications in practice

We next examine three practical implications of data minimization as framed in the reviewed regulations: **(1) Individualized nature of data minimization:** Notice that different individuals may require different amounts and types of information to fulfill a given purpose. For instance, in a loan approval scenario, personal data such as age, income, and job history are crucial. However, the importance of additional data, like medical history, can vary significantly among applicants. This variability underscores that data

General Data Protection Regulation (GDPR), Europe gdpr-info.eu/

Article 4(1): “personal data” means any information relating to an identified or identifiable natural person (“data subject”) [. . .];
Article 5(1)(b): Personal data shall be collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes;
Article 5(1)(c): Personal data shall be adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed.

California Privacy Rights Act (CPRA), USA cppa.ca.gov/

Section 1798.100 (a)(1) & (a)(2): [. . .] A business shall not collect additional categories of (sensitive) personal information or use (sensitive) personal information collected for additional purposes that are incompatible with the disclosed purpose for which the (sensitive) personal information was collected without providing the consumer with notice consistent with this section.
Section 1798.140 (v)(1): “Personal information” means information that identifies, relates to, describes, is reasonably capable of being associated with, or could reasonably be linked, directly or indirectly, with a particular consumer or household.

General Personal Data Protection Law (LGPD), Brazil lgpd-brazil.info/

Article 5(I): personal data: information regarding an identified or identifiable natural person;

Article 6: Activities of processing of personal data shall be subject to the following principles,
I: processing done for legitimate, specific and explicit purposes of which the data subject is informed, with no possibility of subsequent processing that is incompatible with these purposes;
II: compatibility of the processing with the purposes communicated to the data subject, in accordance with the context of the processing;
III: limitation of the processing to the minimum necessary to achieve its purposes, covering data that are relevant, proportional and non-excessive in relation to purposes of the data processing [. . .];

Protection of Personal Information Act (POPIA), South Africa popia.co.za/

Section 1: “personal information” means information relating to an identifiable, living, natural person, [. . .]

Section 10: Personal information may only be processed if, given the purpose for which it is processed, it is adequate, relevant and not excessive.
Section 13(1): Personal information must be collected for a specific, explicitly defined and lawful purpose [. . .]

Consumer Data Rights (CDR), Australia www.legislation.gov.au/Details/F2023C00735

Rule 1.8(a): An accredited person complies with the data minimisation principle if when making a consumer data request on behalf of a CDR consumer, it does not seek to collect: (i) more CDR data than is reasonably needed; [. . .]
Rule 4.11(1)(a): When asking a CDR consumer to give a consent, an accredited person must allow the CDR consumer to choose the types of CDR data to which the consent will apply by enabling the CDR consumer to actively select or otherwise clearly indicate: (i)..; and (ii) in the case of a use consent— the specific uses of collected data to which they are consenting; [. . .]

Personal Information Protection Act (PIPA), South Korea www.pipc.go.kr/eng/index.do

Article 2(1): The term “personal information” means any of the following information relating to a living individual: (a) Information that identifies a particular individual [. . .]; (b) Information which, even if it by itself does not identify a particular individual, may be easily combined with other information to identify a particular individual [. . .];
Article 3(1): The personal information controller shall specify explicitly the purposes for which personal information is processed; and shall collect personal information lawfully and fairly to the minimum extent necessary for such purposes.

Table 1: Excerpts from various data protection regulations from across the globe on the principle of data minimization, highlighting language on purpose limitation, data relevance, and references to the expectations of data privacy.

relevance—-and therefore redundancy—-is not universal but contextually and individually defined. **(2) Data minimization is data dependent:** Regulations describe the term “data collection” as acquiring data from entities like government agencies, which typically involves selecting relevant data from an already gathered large dataset. For instance, census data is often used for various analysis tasks. This differs from *field* data collection, which lacks flexibility for individualized minimization adjustments. Our framework focuses on the former interpretation, wherein minimization is contingent on the data itself (as shown by the two distinct stages of “Data collection from individuals” and “Data minimization” in Figure 1). **(3) Privacy expectations through minimization:** There is an implicit expectation of privacy through minimization in data protection regulations (Leemann et al., 2022; Goldsteen et al., 2021; Staab et al., 2024). *However, this expectation overlooks a crucial aspect of real-world data–the inherent correlations among* *various features*. Information about individuals is rarely isolated, thus, merely removing or not collecting data, may allow for confident reconstruction inference (Garfinkel et al., 2019), as we will show in our work.

3. Operationalizing data minimization in ML

We next present a data minimization framework for ML that aligns with the regulatory objectives discussed above. Successively, we examine the tension between the minimization objectives and their privacy implication and define various threat models to quantify real-world privacy risks.

3.1. Data minimization as optimization

Consider a dataset D consisting of n datapoints $(\boldsymbol{x}_i, y_i)$, where $i \in [n]$, each drawn i.i.d. from an unknown distribution. Therein, $\boldsymbol{x}_i \in \mathcal{X}$ is a p-dimensional feature vector and $y_i \in \mathcal{Y}$ is the corresponding output label. As an illustrative example, consider a loan approval task (Ding et al., 2021). Here, $\boldsymbol{x}_i$ could describe an individual's age, income, race, and job, while y_i whether they will repay a loan. The objective is to train a predictor $f_\theta : \mathcal{X} \to \mathcal{Y}$, parametrized by $\theta \in \mathbb{R}^d$, to minimize the empirical risk:

$$\overset{\star}{\theta} = \operatorname*{argmin}_{\theta} \; J(\theta; \mathbf{X}, \mathbf{Y}) = {}^1\!/_n \sum_{i=1}^{n} \ell(f_\theta(\boldsymbol{x}_i), y_i),$$

where $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ is a non-negative loss function that evaluates model quality, and $\mathbf{X}$ and $\mathbf{Y}$ represent the matrix of all features and the vector of all labels in D, respectively.

The goal of *data minimization* is to reduce the size of D by selectively removing components from the feature vectors $\boldsymbol{x}_i$ (data relevance). This is achieved while also maintaining performance levels comparable to those achieved using the complete dataset (purpose). The privacy goal, in this interpretation, corresponds to retaining only the necessary data. This objective can be stated as a bilevel optimization:

$$\operatorname*{Minimize}_{\boldsymbol{B} \in \{\bot,1\}^{n \times p}} \quad \|\boldsymbol{B}\|_1 \tag{1a}$$

$$s.t.: \quad J(\hat{\theta}; \mathbf{X}, \mathbf{Y}) - J(\overset{\star}{\theta}; \mathbf{X}, \mathbf{Y}) \le \alpha \tag{1b}$$

$$\hat{\theta} = \operatorname*{argmin}_{\theta} \frac{1}{n} \sum_{i=1}^{n} \ell\left(f_\theta(\boldsymbol{x}_i \odot \boldsymbol{B}_i), y_i\right). \tag{1c}$$

Therein, $\boldsymbol{B}$ is an $n \times p$ binary matrix, which we call the *minimization matrix*, taking values in the set $\{\bot, 1\}$, and the ℓ_1-norm of $\boldsymbol{B}$, i.e., $\|\boldsymbol{B}\|_1$, is simply the sum of 1s in the minimization matrix. Here, the symbol $\bot$ represents the concealment or removal of redundant values in the dataset, i.e., $\forall a \in \mathbb{R} : a \times \bot = \bot$, and $\alpha \ge 0$ is an input parameter which thresholds the permitted drop in model quality due to data minimization.

The minimized input features $\mathbf{X}'$ are defined as the element-wise product of the original features $\mathbf{X}$ and the minimization matrix $\boldsymbol{B}$, i.e., $\mathbf{X}' = \mathbf{X} \odot \boldsymbol{B}$. Note that, data minimization produces features in the space $\mathcal{X} \cup \{\bot\}$. While some learning algorithms can handle missing values ($\bot$), data imputation is needed in other cases. A discussion of imputation methods is delegated to Section 4.1.

The optimization problem above defines an operational method to remove entries from the feature set $\mathbf{X}$ in a *personalized* manner (1a), while adhering to pre-specified accuracy requirements on the original dataset (1b), for the final model trained on the minimized dataset (1c). While this formulation captures the original goals expressed in the legal formulation of data minimization, it is however intractable to solve in practice. A discussion of tractable approximate algorithmic alternatives is delegated to Section 4 and we discuss next the privacy implications of this data minimization process.

3.2. Privacy leakage

The implicit objective of data minimization is to enhance privacy. Thus, we focus on assessing the minimized data using several threat models (Figure 1). Since the expectation of data privacy from minimization in various regulations is focused on the events of a data disclousre, we primarily focus on reconstruction and re-identification risks. This is different from conventional private ML settings (Dwork, 2006; Rigaki & Garcia, 2023), which focuses on information leakage due to ML inference (i.e., outside the "wall" in Figure 1). Nonetheless, we also examine how data minimization privacy promise may hold under membership inference attacks and delegate this analysis to Appendix C.

Reconstruction Risk (RCR). Real-world datasets often exhibit underlying associations between various features, making it possible to reconstruct minimized data (Garfinkel et al., 2019). Reconstruction attacks aim to recover missing information from a target dataset. Given the minimized dataset $\mathbf{X}'$, the attacker's goal is to generate their reconstruction $\mathbf{X}^R$ of the original set of features $\mathbf{X}$. The ReConstruction Risk (RCR) can be evaluated by measuring the similarity between the original features $\boldsymbol{x}_i \in \mathbf{X}$ and the reconstructed features $\boldsymbol{x}_i^R \in \mathbf{X}^R$, computed using a gaussian kernel with $\sigma = 1$ (Srebro, 2007) as:

$$\text{RCR} = \frac{1}{n} \sum_{i=1}^{n} e^{-\|\boldsymbol{x}_i - \boldsymbol{x}_i^R\|_2}. \tag{2}$$

The reconstruction risk metric can be adjusted prioritize the reconstruction of certain features over others by introducing appropriate weights to the similarity measurement.

Re-identification Risk (RIR). Data breaches often lead to re-identification of individuals using partial or anonymized

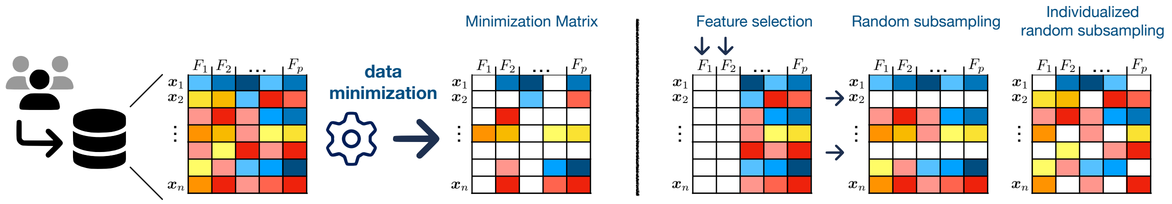


Fig. 2: **(Left)** Data minimization as a set of individualized binary decisions, visualizing the minimization matrix $\boldsymbol{B}$. **(Right)** Three baseline algorithms for data minimization.

data matched with an auxiliary dataset $\mathbf{X}^A$. The success of these re-identification attacks can be measured by the mean reciprocal rank (MRR) scores. For each data point $\boldsymbol{x}_i^A$ in $\mathbf{X}^A$, the adversary ranks data points in the minimized dataset $\mathbf{X}'$ based on the likelihood of matching identities. The Re-Identification Risk (RIR) is calculated as the average MRR score, assigning a score of 1 for a correct match and 0 otherwise, defined by the formula:

$$RIR = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{index(i, rank(\boldsymbol{x}_i^A, \mathbf{X}'))}. \tag{3}$$

Here, $rank(\boldsymbol{x}_i^A, \mathbf{X}')$ is the adversary's predicted ranking, and $index(e, A)$ the position of e in vector A. It assumes a one-to-one match between auxiliary and minimized datasets.

4. Data minimization and utility

Having defined the problem of data minimization for ML as a bilevel optimization, this section outlines how to operationalize it, providing three strong baselines and three additional algorithms from the bi-level optimization literature (1). Note that although the data minimization principle aims to optimize the dataset size while maintaining the model's quality (Constraint 1b), these algorithms adopt a dual approach, optimizing model quality when trained on minimized data under a given sparsity constraint $\|\boldsymbol{B}\|_1 \leq k$. This dual approach has the advantage of allowing these algorithms to find a sparsity parameter k that meets the desired α-drop in performance.

Baseline techniques. The experiments implement three baseline methods, depicted pictorially in Figure 2: **(1)** *Feature selection* (Blum & Langley, 1997) employs a breadth-based strategy that identifies and minimizes less important features within the dataset. **(2)** *Random subsampling* is a depth-based strategy that randomly selects a subset of data points, thereby reducing the dataset size by excluding specific rows. **(3)** *Individualized random subsampling* further tailors this approach by randomly selecting specific entries (feature, sample) for each individual, aiming to reduce dataset size in a more personalized manner. While these baselines help us to assess model quality degradation, they do not fully comply with legal standards of data minimization, which require adherence to principles of purpose and data relevance, as discussed earlier and further expanded in Appendix B.

Data minimization algorithms. We next briefly introduce three classes of algorithms adopted to solve the bilevel problem 1 and defer to extensive details, their theoretical underpinning, and comparative analysis to Appendix A.

1. *Approximating the target utility* (Hongli et al., 2011), which focuses on finding an approximate closed-form solution to the lower-level problem, thus simplifying the original optimization.
2. *Modelling the target utility* (Wang & Shan, 2006), which instead attempts to model the lower-level mapping and estimate it online again without solving the optimization.
3. *Evolutionary algorithms* (Sinha et al., 2014), which trade the advantage of carrying no assumption with slow convergence and high computational demand.

Throughout our experimental evaluation, baseline methods are depicted with blue colors while the optimization algorithms for data minimization above with red colors.

4.1. Experiment setup

Datasets. Our evaluations focus on classification for both tabular data (marked with symbol $^\diamond$) and image data (symbol $^\heartsuit$), to cover a diverse set of distributions and modalities. In this section, we use **(i) the bank marketing dataset** (Moro et al., 2014), taken from a telemarketing campaign with financial information of 11,162 customers, and **(ii) the handwritten digits dataset** (Xu et al., 1992), containing a total of 1,797 handwritten digit images across 10 different classes (i.e., digits), Results on several additional datasets and modalities, including text, are reported in Appendix D.

Data splits. Datasets are split into two equal groups, private and public data. The public data is used as the test data, to calculate statistics for data imputation, and train reference models for membership inference. The private data is further split into two equal groups, i.e. members and non-members. The members are our training data $\mathbf{X}$, i.e., also the data that will be minimized in our pipeline. The non-members are only used for evaluating membership inference attacks (details in Appendix C).

Learning setup and data imputation. The experiments

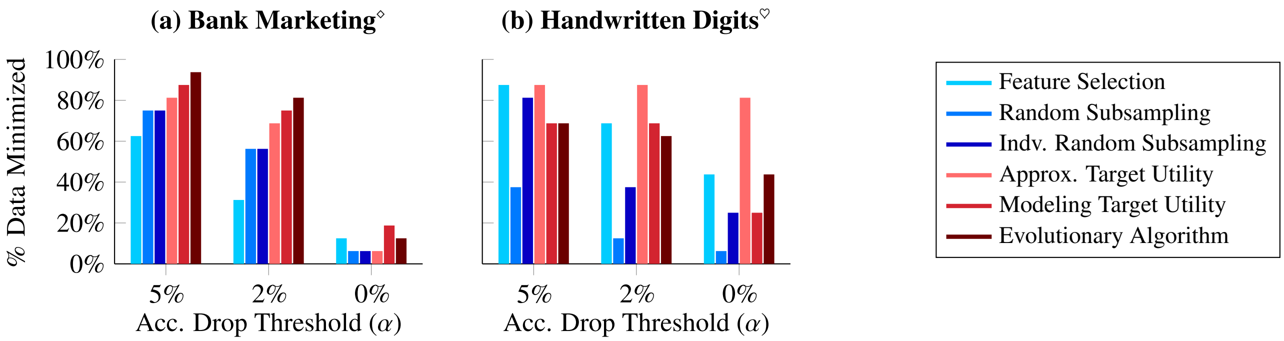


Fig. 3: Percentage of dataset minimized under different accuracy drop thresholds (α). Detailed results with changing minimization sparsities can be found in Appendix D.

use logistic regression for the tabular dataset, and a fully connected neural network for the handwritten digits dataset. Both datasets are normalized using MinMaxScaler. Data imputation is performed under the assumption of a multivariate Gaussian distribution, utilizing the mean and covariance matrix of the public data to fill missing values with the mean of the conditional distribution (Hoff, 2007).

4.2. Efficacy of data minimization

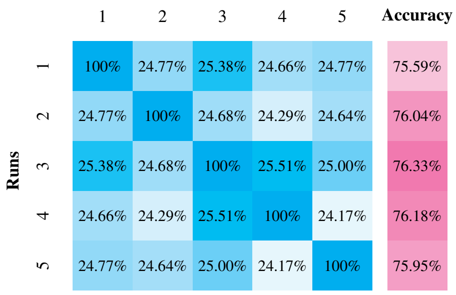
First, this section assesses the ability of data minimization algorithms to reduce the dataset size across several modalities. Figure 3 summarizes the results, comparing the amount of data that could be minimized (y-axis) given a maximum drop in accuracy (x-axis) when compared to the accuracy returned by the classifier trained over the original, not-minimized, dataset. The results indicate a strong trend: *A substantial amount of data can often be removed without sacrificing utility, suggesting that much of the collected data is superfluous in the datasets analyzed.*

Notably, the baseline methods for minimization are found effective in reducing substantial amounts of data, possibly because they leverage existing structural redundancies in the dataset. For instance, feature selection excels with the handwritten digit dataset, where outer boundary pixels are never covered by the digits, and thus can be easily removed with minimal impact on model performance. Nonetheless, the optimization algorithms consistently outperform the baselines across all datasets, with different algorithms outshining others in specific scenarios. Evolutionary algorithms, for instance, surpass others on the bank dataset, driving minimization to extreme sparsity. It highlights a strength of minimization at its core: reducing the dataset size to a mere 18.75% (and 6.25%) subset of the original bank dataset still allows the model to maintain utility with only a 2% (and 5%) drop in accuracy. Similarly, approximating the target utility excels on the handwritten digits dataset, presumably due to its compatibility with the values of redundant features in the dataset (empty pixels are represented with 0, compatible with the zero-imputation assumption of the algorithm). An in-depth discussion of the strengths and weaknesses of these algorithms is provided in Appendix A.

4.3. Multiplicity, emergent individualization, and privacy in data minimization

Building on the previous section's demonstration of data minimization's effectiveness in various settings, this section explores *the concept of multiplicity of data minimization*. We applied the evolutionary algorithm five times to the bank dataset under different randomness settings, each targeting 75% sparsity. Notably, the results, illustrated in Figure 4, show minimal overlap among the datasets retained from each run, with the highest overlap being only 25.51%. Clearly, there are many distinct ways to achieve data minimization while maintaining utility that meets data relevance constraints (i.e., no further data can be removed without compromising utility). Such variability underscores the flexibility of achieving data minimization through diverse algorithms. However, this also suggests that different minimized datasets may pose varying privacy risks related to the undisclosed features, as discussed in Section 2.1, pointing to a misalignment in privacy outcomes.

To better understand these effects and their relation to privacy, we visualize the minimized dataset and highlight some intriguing characteristics. Starting with the bank dataset,



Runs	1	2	3	4	5	**Accuracy**
1	100%	24.77%	25.38%	24.66%	24.77%	75.59%
2	24.77%	100%	24.68%	24.29%	24.64%	76.04%
3	25.38%	24.68%	100%	25.51%	25.00%	76.33%
4	24.66%	24.29%	25.51%	100%	24.17%	76.18%
5	24.77%	24.64%	25.00%	24.17%	100%	75.95%

Fig. 4: Despite achieving similar utility, there is minimal pairwise overlap in the datasets minimized over five iterations with varying randomness.

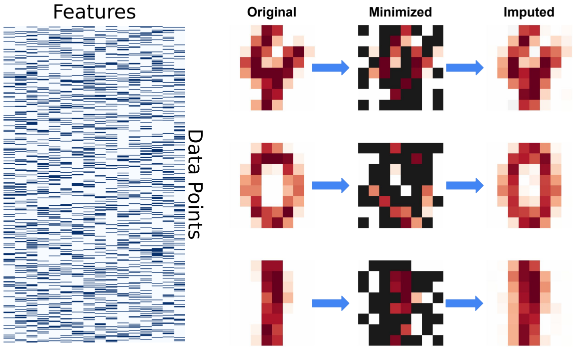


Fig. 5: Visualization of minimized datasets across different modalities illustrates: (Left) the emergence of individualized minimization strategies within a compressed view of a bank dataset; and (Right) a misalignment between minimization efforts and privacy objectives, as imputed images reveal reconstructed information despite minimization.

minimized to 75% sparsity using the evolutionary algorithm, Figure 5 (left) shows that individualization trends naturally emerge from the optimization process, where no single feature is consistently favoured over others. Similarly, the minimized images from the handwritten digits dataset, processed by the evolutionary algorithm at 50% sparsity, are shown in Figure 3 (right). The trends here are more interpretable; for example, the central vertical line is preserved in the image of the digit '1', while the outer curves are retained for '0'. Notice however that despite reducing the dataset to half its original size, a significant portion of the images can still be reconstructed using statistics learned from public data. *This provides a strong indication of privacy risks and suggests that, as we will show next, a minimized dataset does not equate to enhanced privacy.*

5. Data minimization and privacy

In the previous section, we showed that various algorithms can effectively minimize data while focusing on utility. However, as discussed in Section 2.1, there is an expectation of privacy associated with minimization, currently unexplored in literature (Rastegarpanah et al., 2021; Shanmugam et al., 2022; Biega et al., 2020). Next, we conduct an important empirical assessment of minimization algorithms on real-world privacy attacks and their alignment with the minimization objective.

Defining inference attacks. During a reconstruction attack, the attacker aims to recover missing information from the minimized dataset. To achieve this, we use the data imputation method described in Section 4.1. In contrast, a re-identification attack involves the attacker attempting to identify an individual using only partial information. Specifically, the attacker aims to find the best match in the minimized dataset $\mathbf{X}'$ for a data point from the auxiliary dataset (in our setup, $\mathbf{X}^A \equiv \mathbf{X}$). For this purpose, we use a modified Euclidean distance that disregards missing values and adjusts the distance scaling accordingly, to accurately rank the best matches.

5.1. Evaluating privacy leakage

Data minimization and re-identification risk. We first focus on the re-identification risks for the handwritten digits dataset, shown in Figure 6(b). A key observation is that re-identification risk remains remarkably high, even after extensive minimization. For instance, the re-identification risk is close to 1 for most algorithms even when the dataset is reduced to approximately 20% of its original size. This behaviour is due to the dataset's large feature space. Thus, while minimization algorithms were able to reduce dataset size significantly while maintaining utility (as shown in Figure 3), they do not achieve comparable reductions in re-identification risk. *This underscores a fundamental misalignment between the goals of data minimization and the actual privacy goals.*

While this misalignment is less pronounced in the bank dataset (Figure 6(a)), there are still regions in which reducing dataset size does not correspondingly decrease re-identification risk. Notably, the algorithm that approximates the target utility demonstrates a closer alignment with privacy goals compared to other algorithms. Considering the lower amount of minimization by approximating the target utility as observed in Figure 3, we can infer that the additional data minimized by the other two algorithms was not aligned with re-identification risk. Thus, while these algorithms would have been preferred if we only considered utility and dataset size (as we did in Figure 3), they do not provide proportionate privacy improvements, highlighting a critical concern of misalignment.

Finally, observe that our baseline algorithms display notable trends across both datasets. Random subsampling emerges as an effective strategy to reduce dataset size while being aligned with re-identification risks. Indeed, completely removing a data point also eliminates any risk of its re-identification. In contrast, feature selection, despite preserving reasonable utility post-minimization, fails to effectively mitigate re-identification risks in any dataset. This inadequacy can be attributed to the persistence of certain features in the minimized dataset that, even when isolated, can uniquely identify individuals. This phenomenon suggests a general misalignment between breadth-based minimization techniques (Rastegarpanah et al., 2021; Goldsteen et al., 2021; Staab et al., 2024) and the expected reduction in re-identification risks. Once again, these observations highlight the disparity between mere reductions in dataset size and actual improvements in privacy outcomes.

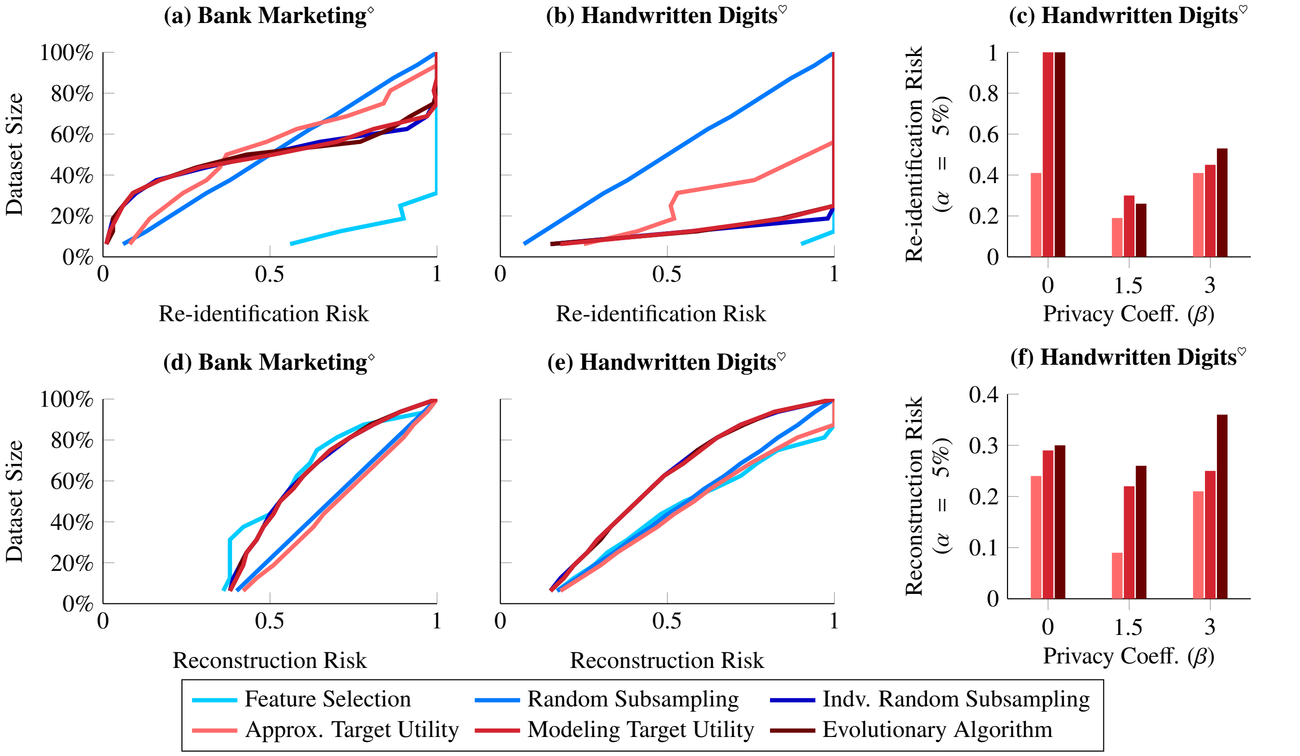


Fig. 6: **(a, b, d, e)** Re-identification and reconstruction risks under changing sparsity. While the overall trends show improvement in privacy with data minimization, they are not aligned with the trends of dataset size. **(c, f)** Privacy risks under feature-level privacy scores. The lowest risk values can be seen at $\beta = 1.5$, highlighting the importance of considering privacy during minimization.

Data minimization and reconstruction risk. Next, we shift focus to the results of reconstruction risks, shown in Figure 6(d, e). Notably, even at the highest levels of data minimization (smallest dataset size), reconstruction risks remain significant for both domains. This aspect is linked to the ability to reconstruct features using overall dataset statistics, which, although decreasing in accuracy as minimization increases, still retain some reconstructive value. Interestingly, we find that the handwritten digits dataset is comparatively easier to protect against reconstruction, likely due to its higher feature variance relative to the bank dataset. While the same high variance in features led to worse re-identification risk, it instead helped the handwritten digits dataset achieve better alignment with reconstruction risk. Thus, not only is the dataset size not aligned with privacy, but even different privacy risks are not aligned with each other, motivating the explicit involvement of appropriate privacy constraints during minimization.

The trends observed in re-identification risk also manifest in reconstruction risks, albeit with some unique differences. Firstly, both random subsampling and individualized random subsampling have linear relationships between dataset size and privacy, highlighting a strong alignment with privacy. Secondly, the best-performing algorithm for dataset size might not always be the best choice overall. Specifically, while the algorithm that approximates the target utility shows strong performance on the bank marketing dataset and holds a significant advantage on the handwritten digits dataset, it consistently presents higher reconstruction risks for both datasets. These trends emphasize the importance of incorporating privacy considerations when selecting the most appropriate minimization algorithm for a specific scenario.

5.2. Adapting minimization for better privacy

We now present a simple yet effective adjustment to the minimization problem to align better with privacy and demonstrate the feasibility of a more comprehensive minimization objective, managing both utility and privacy. Given the output of a minimization algorithm $\boldsymbol{B}^a = [\overset{\star}{\boldsymbol{B}}]_{\perp\to 0}$, we define privacy scores $\boldsymbol{V}$, such that the score matrix $\boldsymbol{C}_{ij} = \boldsymbol{B}^a_{ij}+\beta\boldsymbol{V}_{ij}$ determines the indices that should be minimized. For simplicity, we will only define feature-level privacy scores, which are not personalized, i.e., $\boldsymbol{V}_{ij} = \boldsymbol{V}_{qj},\ \forall i, q \in [n], j \in [p]$. Here, β serves as a hyperparameter to tune the emphasis on privacy. Ultimately, we minimize values with the lowest $\boldsymbol{C}_{ij}$ scores to achieve the target sparsity. The privacy scores $\boldsymbol{V}$ are normalized to $[0, 1]$ before being combined with the minimization matrix.

- **Privacy scores to reduce re-identification.** In re-identification attacks, the risk arises from disclosing *unique* features (e.g., phone number, SSN) rather than non-unique features (e.g., gender, race), as they make it easier to identify individuals. Following this rationale, we propose using the negative of the number of unique values for a feature as its privacy score. Minimizing distinctive features can increase the difficulty of re-identification.
- **Privacy scores to reduce reconstruction.** In reconstruction attacks, the risk arises from the high correlation among features, which adversaries exploit to infer missing values. To mitigate reconstruction risks, we propose using the negative of the average correlation of each feature with every other feature as a measure of its independence and thus its privacy score.

Evaluating privacy score modifications. For a given level of sparsity, it can be expected that incorporating privacy scores will decrease data breach risks, but potentially reduce accuracy. The critical question is whether we can attain a more favourable trade-off between privacy and utility, regardless of the sparsity. We present the results in Figure 6(c, f) for the handwritten digits dataset, varying the hyperparameter β. Results for other datasets are present in Appendix D.

At $\beta = 1.5$, integrating privacy scores enhances the privacy-utility trade-off, cutting the re-identification risk by more than half while maintaining the same accuracy drop. The impact on reconstruction risk is less dramatic, but improvements are evident. Increasing the focus on privacy to $\beta = 3$ results in a less optimal trade-off, yet still better than at $\beta = 0$. This notable effect of a basic feature-level privacy score underscores the necessity of directly addressing privacy in the minimization process, instead of relying on its incidental emergence.

6. Related work

Existing research on minimization in the context of data protection regulations can be broadly divided into *breadth-based* and *depth-based* techniques. Breadth-based strategies aim to minimize data by limiting the number of features (Rastegarpanah et al., 2021) or introducing feature generalization (Goldsteen et al., 2021; Staab et al., 2024). On the other hand, depth-based approaches focus on reducing the number of unique data points by using methods like data pruning (Paul et al., 2021; Sorscher et al., 2022; Shanmugam et al., 2022). While there are some discussions on individualized minimization for recommender systems (Biega et al., 2020; Chen et al., 2023), they are limited in their ability to generalize to other settings in ML.

On a separate note, most studies in data minimization aim to simply reduce the raw size of the datasets (Rastegarpanah et al., 2021; Shanmugam et al., 2022; Biega et al., 2020; Chen et al., 2023) and don't give any attention to privacy concerns (Leemann et al., 2022). Although some works do go beyond dataset size and discuss other aspects of information leakage (Goldsteen et al., 2021), they still lack connections with real-world privacy risks. The work closest to ours is a concurrent work by Staab et al. (2024), which also introduces real-world privacy attacks to quantify privacy leakage after minimization. However, unlike our approach, Staab et al. (2024) concentrates on breadth-based methods, thus missing the individualized nature of minimization.

Some studies have also formalized data minimization during inference, emphasizing the personalized nature of the process and delving into its privacy implications (Tran & Fioretto, 2023; James et al., 2023). However, data minimization during inference is distinctly different from data minimization during training, which is the primary focus of our paper.

7. Discussion and conclusion

In proposing a formal framework for data minimization in ML, this paper reveals a disconnect between legal mandates and their practical implementation. While data protection regulations aim to limit data collection with an expectation of privacy, current objectives of minimization fall short of providing robust privacy safeguards. Notice, however, that this is not to say that minimization is incompatible with privacy; instead, we emphasize the need for approaches that incorporate privacy into their objectives (as done in §5.2), rather than treating them as an afterthought.

Addressing this concern also brings forth a variety of optimization challenges. The optimization problem is complicated further under alternative interpretations of data collection, such as gathering a range of data rather than exact values (Goldsteen et al., 2021; Staab et al., 2024). These challenges are particularly significant in large-scale applications where both time and accuracy are critical factors. Therefore, future work should focus on the development of efficient minimization algorithms able to make good trade-offs between utility and privacy. The ethical and fairness considerations of data minimization also add to its complexity. By design, data minimization is likely to remove data that resembles the majority (Biega et al., 2020), leaving the minority more vulnerable to privacy risks. Thus, developing fairness-aware mechanisms for minimization is an important avenue for future research.

This study marks a step in aligning the legal requirements with practical, technical solutions for data minimization in ML. We hope it could set the stage for future work aimed at developing comprehensive, efficient, and ethically sound methodologies for minimization.

Acknowledgements

This research is partially supported by NSF grants 2133169. Fioretto is also supported by a Google Research Scholar Award and an Amazon Research Award. Its views and conclusions are those of the authors only.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pp. 308–318, 2016.

Biega, A. J. and Finck, M. Reviving Purpose Limitation and Data Minimisation in Data-Driven Systems. *Technology and Regulation*, pp. 44–61 Pages, December 2021.

Biega, A. J., Potash, P., Daumé, H., Diaz, F., and Finck, M. Operationalizing the legal principle of data minimization for personalization. In *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval*, pp. 399–408, 2020.

Bishop, C. M. and Nasrabadi, N. M. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

Blum, A. L. and Langley, P. Selection of relevant features and examples in machine learning. *Artificial intelligence*, 97(1-2):245–271, 1997.

Chen, Z., Sun, F., Tang, Y., Chen, H., Gao, J., and Ding, B. Studying the impact of data disclosure mechanism in recommender systems via simulation. *ACM Transactions on Information Systems*, 41(3):1–26, 2023.

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., and Reis, J. Modeling wine preferences by data mining from physicochemical properties. *Decision support systems*, 47(4): 547–553, 2009.

Ding, F., Hardt, M., Miller, J., and Schmidt, L. Retiring adult: New datasets for fair machine learning. *Advances in neural information processing systems*, 34:6478–6490, 2021.

Dwork, C. Differential privacy. In *International colloquium on automata, languages, and programming*, pp. 1–12. Springer, 2006.

Garfinkel, S., Abowd, J. M., and Martindale, C. Understanding database reconstruction attacks on public data. *Communications of the ACM*, 62(3):46–53, 2019.

Goldsteen, A., Ezov, G., Shmelkin, R., Moffie, M., and Farkash, A. Data minimization for gdpr compliance in machine learning models. *AI and Ethics*, pp. 1–15, 2021.

Hoff, P. D. Extending the rank likelihood for semiparametric copula estimation. *Annals of Applied Statistics*, 1(1):265–283, 2007.

Hongli, G., Juntao, L., and Hong, G. A survey of bilevel programming model and algorithm. In *2011 Fourth International Symposium on Computational Intelligence and Design*, volume 2, pp. 199–203, 2011. doi: 10.1109/ISCID.2011.151.

James, H., Nagpal, C., Heller, K. A., and Ustun, B. Participatory personalization in classification. In *ICML 3rd Workshop on Interpretable Machine Learning in Healthcare (IMLH)*, 2023.

Leemann, T., Pawelczyk, M., Eberle, C. T., and Kasneci, G. I prefer not to say: Operationalizing fair and user-guided data minimization. *arXiv preprint arXiv:2210.13954*, 2022.

Moro, S., Cortez, P., and Rita, P. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62:22–31, 2014.

Nasr, M., Shokri, R., and Houmansadr, A. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *2019 IEEE symposium on security and privacy (SP)*, pp. 739–753. IEEE, 2019.

Paul, M., Ganguli, S., and Dziugaite, G. K. Deep learning on a data diet: Finding important examples early in training. *Advances in Neural Information Processing Systems*, 34: 20596–20607, 2021.

Rastegarpanah, B., Gummadi, K., and Crovella, M. Auditing black-box prediction models for data minimization compliance. *Advances in Neural Information Processing Systems*, 34:20621–20632, 2021.

Rigaki, M. and Garcia, S. A survey of privacy attacks in machine learning. *ACM Computing Surveys*, 56(4):1–34, 2023.

Shanmugam, D., Diaz, F., Shabanian, S., Finck, M., and Biega, A. Learning to limit data collection via scaling laws: A computational interpretation for the legal principle of data minimization. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 839–849, 2022.

Shokri, R., Stronati, M., Song, C., and Shmatikov, V. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pp. 3–18. IEEE, 2017.

Sinha, A., Malo, P., Frantsev, A., and Deb, K. Finding optimal strategies in a multi-period multi-leader–follower stackelberg game using an evolutionary algorithm. *Computers & Operations Research*, 41:374–385, 2014.

Sorscher, B., Geirhos, R., Shekhar, S., Ganguli, S., and Morcos, A. Beyond neural scaling laws: beating power law scaling via data pruning. *Advances in Neural Information Processing Systems*, 35:19523–19536, 2022.

Srebro, N. How good is a kernel when used as a similarity measure? In *Learning Theory: 20th Annual Conference on Learning Theory, COLT 2007, San Diego, CA, USA; June 13-15, 2007. Proceedings 20*, pp. 323–335. Springer, 2007.

Staab, R., Jovanović, N., Balunović, M., and Vechev, M. From principle to practice: Vertical data minimization for machine learning. In *2024 IEEE Symposium on Security and Privacy (SP)*, pp. 89–89. IEEE Computer Society, 2024.

Thomas, L., Gondal, I., Oseni, T., and Firmin, S. S. A framework for data privacy and security accountability in data breach communications. *Computers & Security*, 116:102657, 2022.

Tibshirani, R. Modeling basics: Assessment, selection, and complexity. 2015. URL `www.stat.cmu.edu/~ryantibs/statml`.

Tran, C. and Fioretto, F. Personalized privacy auditing and optimization at test time. *arXiv preprint arXiv:2302.00077*, 2023.

Wairimu, S. and Fritsch, L. Modelling privacy harms of compromised personal medical data-beyond data breach. In *Proceedings of the 17th International Conference on Availability, Reliability and Security*, pp. 1–9, 2022.

Wang, G. G. and Shan, S. Review of metamodeling techniques in support of engineering design optimization. In *International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*, volume 4255, pp. 415–426, 2006.

Xu, L., Krzyzak, A., and Suen, C. Y. Methods of combining multiple classifiers and their applications to handwriting recognition. *IEEE transactions on systems, man, and cybernetics*, 22(3):418–435, 1992.

Zarifzadeh, S., Liu, P., and Shokri, R. Low-cost high-power membership inference attacks. *arXiv preprint arXiv:2312.03262*, 2023.

A. Data minimization algorithms

We provide additional details on various algorithms used in the paper.

A.1. Baseline techniques

Feature Selection (Blum & Langley, 1997). Feature selection is a breadth-based minimization strategy that retains only the most important features in the data. The algorithm works by first sorting the features of the dataset in order of their importance, using a pre-established criterion, to identify a subset S of the least important features. The minimized dataset $\mathbf{X}'$ is formed from $\mathbf{X}$ by removing all features in S. In our paper, the importance criterion is the absolute correlation between each feature and the output label, and the algorithm sets:

$$\boldsymbol{B}_{ij} = \perp \quad \forall i \in [n], j \in S \tag{4}$$

where $|S|$ is a parameter controlling the minimization sparsity.

Random Subsampling. Random subsampling is a depth-based minimization strategy that randomly chooses a subset of data points from the original dataset. In the context of data minimization, it sets the minimization matrix $\boldsymbol{B}$ as:

$$\boldsymbol{B}_{ij} = \perp \quad \forall j \in [p], \text{ with probability } \ k_p, \tag{5}$$

where $0 \leq k_p \leq 1$ is a probability value chosen so that nk_p rows of the dataset D are minimized, in expectation.

Individualized Random Subsampling. Individualized random subsampling is an extension of random subsampling, but it removes individual entries (feature, sample) rather than complete rows, i.e., it performs individualized minimization. The minimization matrix $\boldsymbol{B}$ is defined as:

$$\boldsymbol{B}_{ij} = \perp \quad \text{with probability } \ k_p, \tag{6}$$

where $0 \leq k_p \leq 1$ is chosen so that npk_p elements of the dataset D are minimized, in expectation.

A.2. Data minimization algorithms

Approximating the Lower Level Program (Hongli et al., 2011). Solving the bi-level optimization discussed is challenging due to the nested structure of the problem which requires solving a non-convex lower-level optimization within a non-convex upper-level optimization. When the lower-level problem has a unique solution that can be explicitly expressed in a closed form, then the overall bilevel program can be rewritten as a single-level program which is much simpler to solve. The underlying idea behind the proposed framework lies in approximating the target utility via the original utility by the first Taylor approximation (see again Equation 20). Assuming that the second order component associated with $\|\mathbf{X}' - \mathbf{X}\|_2^2$ is negligible, the difference in model's utility is

$$J(\hat{\theta}; \mathbf{X}, \mathbf{Y}) - J(\overset{\star}{\theta}; \mathbf{X}, \mathbf{Y}) \approx -(\mathbf{X}' - \mathbf{X})^T LH^{-1}G.$$

where L, H, G are the gradients w.r.t. the model's parameters, the Hessian w.r.t. the model's parameters, and the second-order derivative w.r.t. the model's parameters and the dataset, respectively, of the original utility on the complete dataset $J(\overset{\star}{\theta}; \mathbf{X}, \mathbf{Y})$. This simplifies the original optimization as:

$$\underset{\boldsymbol{B} \in \{\perp, 1\}^{n \times p}}{\text{Minimize}} \ \ \|\boldsymbol{B}\|_1 \qquad s.t.: \ \ (\mathbf{X} - \mathbf{X}')^T LH^{-1}G \leq \alpha. \tag{7}$$

Suppose, we perform a simple zero imputation, i.e., setting $\perp = 0$, then $\mathbf{X}' = \mathbf{X} \odot \boldsymbol{B}$. The above problem now is a binary integer linear programming and the dual problem can be rewritten as below:

$$\underset{\boldsymbol{B} \in \{\perp, 1\}^{n \times p}}{\text{Minimize}} \ \ (\mathbf{X} \odot \boldsymbol{B} - \mathbf{X})^T LH^{-1}G \tag{8a}$$

$$s.t.: \ \ \|\boldsymbol{B}\|_1 \leq k. \tag{8b}$$

Recall again that it is often convenient to view the optimization expressed in (1) as a program that minimizes the loss $J(\hat{\theta}; \mathbf{X}, \mathbf{Y})$ under sparsity constraint over the minimization matrix $\boldsymbol{B}$. It turns out that for this binary linear programming, we can obtain a closed-form solution by setting $\boldsymbol{B}_{ij} = 1$ for k largest entries of $(\mathbf{X}_{ij} - \mathbf{X}'_{ij}) \odot (LH^{-1}G)_{ij}$, and $\boldsymbol{B}_{ij} = \perp$ otherwise. Note that this proposed method is applicable when the imputed data is given (e.g., zero imputation) before running the method. Complicated data imputation like mean/mode/median imputation by the non-missing entries of the same column in the minimized data will result in a non-linear objective in the dual problem. This poses a difficulty since there is no efficient method to solve the binary integer non-linear programming in general.

Modeling the Target Utility (Wang & Shan, 2006). Instead of relying on the assumptions of low sparsity to approximate the target utility as above, we can take a more general approach by directly modelling and learning the mapping between the target utility and the minimized dataset. In other words, we want to learn a parametrized function $m_\omega(\boldsymbol{B}) \approx J(\hat{\theta}; \mathbf{X}, \mathbf{Y})$, where ω is a vector of parameters, to estimate the target utility without solving the lower-level optimization for $\hat{\theta}$.

To learn a tractable mapping $m_\omega(\boldsymbol{B})$, we restrict ourselves to linear functions of $\boldsymbol{B}$ and assume that each index of the dataset has an independent influence on the target utility. Therefore, if we quantify the influence of each index on the target utility as $\mathcal{I}_{ij}$, the mapping $m_\omega(\boldsymbol{B})$ becomes:

$$J(\hat{\theta}; \mathbf{X}, \mathbf{Y}) \approx m_\omega(\boldsymbol{B}) = \frac{1}{np} \sum_{i=1}^{n} \sum_{j=1}^{p} \mathcal{I}_{ij}. \tag{9}$$

Here $\mathcal{I}_{ij}$ can only exist in one of two binary states, i.e., $\mathcal{I}_{ij} = \mathbb{1}_{[\boldsymbol{B}_{ij}=1]} \cdot \mathcal{I}^{1}_{ij} + \mathbb{1}_{[\boldsymbol{B}_{ij}=\perp]} \cdot \mathcal{I}^{\perp}_{ij}$. To learn the values of $\mathcal{I}^{1}_{ij}$ and $\mathcal{I}^{\perp}_{ij}$, we generate a large number of minimization matrix-target utility pairs by solving the lower-level optimization for each pair. We can then learn these parameters for each index ij by averaging the target utility when $\boldsymbol{B}_{ij} = 1$ and $\perp$, respectively. The final optimization thus becomes:

$$\underset{\boldsymbol{B} \in \{\perp, 1\}^{n \times p}}{\text{Minimize}} \ \ \|\boldsymbol{B}\|_1 \qquad s.t.: \ \frac{1}{np} \sum_{i=1}^{n} \sum_{j=1}^{p} \mathcal{I}_{ij} - J(\overset{\star}{\theta}; \mathbf{X}, \mathbf{Y}) \leq \alpha. \tag{10}$$

When considering a sparsity constraint $\|\boldsymbol{B}\|_1 \leq k$ on the minimization matrix, the solution to this formulation is retaining the k entries with the highest value of the term $\mathcal{I}^{\perp}_{ij} - \mathcal{I}^{1}_{ij}$. Although the assumption of independent influence may not hold in all cases, our results (illustrated in the next section) show that this approach can substantially improve the accuracy of existing baselines.

We need to generate minimization matrix-target utility pairs. We start by generating a random minimization matrix $\boldsymbol{B}$, with the target sparsity k. This is the same as performing personalized random subsampling. We then solve the lower-level program for this $\boldsymbol{B}$ and obtain the final target utility of the trained model. We repeat these steps multiple times to create a dataset to learn the parameters of the mapping $m_\omega(\boldsymbol{B})$.

To learn the mapping $m_\omega(\boldsymbol{B})$, we simply need to learn the parameters $\mathcal{I}^{1}_{ij}, \mathcal{I}^{\perp}_{ij}$. Given the assumption of the independent influence of each index on the minimization matrix, we can learn these parameters by calculating the average loss when $B_{ij} = 1$ and $B_{ij} = \perp$:

$$\mathcal{I}^{1}_{ij} = \frac{1}{|P^1|} \sum_{q=1}^{|P^1|} J(\hat{\theta}_{P^1_q}; \mathcal{X}, \mathcal{Y}) \quad \text{where} \quad P^1 := \{\boldsymbol{B} \,|\, \boldsymbol{B}_{ij} = 1\} \tag{11}$$

$$\mathcal{I}^{\perp}_{ij} = \frac{1}{|P^\perp|} \sum_{q=1}^{|P^\perp|} J(\hat{\theta}_{P^\perp_q}; \mathcal{X}, \mathcal{Y}) \quad \text{where} \quad P^\perp := \{\boldsymbol{B} \,|\, \boldsymbol{B}_{ij} = \perp\} \tag{12}$$

$$\text{where} \quad \hat{\theta}_P = \underset{\theta}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \ell\left(f_\theta(\boldsymbol{x}_i \odot P_i), y_i\right). \tag{13}$$

Evolutionary Algorithms (Sinha et al., 2014). So far, we have discussed methods that approximate the original optimization problem under various assumptions, enabling us to solve it more easily. An orthogonal class of algorithms, often applied to solve bi-level programs, are evolutionary algorithms. They trade the advantage of carrying no assumption with slow convergence, i.e., high computational demand, and higher risks of overfitting.

In our implementation of the evolutionary algorithm, we begin with a population of randomly generated minimization matrices $\boldsymbol{B}$ and then evolve them across iterations to reach our objective. We mutate and breed the current population at each stage of the process to create a larger pool of choices and carry out the lower-level optimization for every member of this pool, only retaining the best performers for the next generation. The entire process is repeated until convergence, or for a fixed number of iterations.

Mutation: In our paper, we mutate a minimization matrix $\boldsymbol{B}$ by flipping the value of exactly 10 randomly chosen indices from $1 \to \perp$, and exactly 10 randomly chosen indices from $\perp \to 1$.

Breeding: When breeding between two parent minimization matrices $\boldsymbol{B}^1$ and $\boldsymbol{B}^2$, we keep the same value in the child $\boldsymbol{B}^c$ at indices where both parents agree to be the same, while we randomly choose values for indices where they don't agree to maintain target sparsity. In simpler terms,

$$\boldsymbol{B}^c_{ij} \Leftarrow \boldsymbol{B}^1_{ij} \text{ if } \boldsymbol{B}^1_{ij} = \boldsymbol{B}^2_{ij}, \tag{14}$$

$$\boldsymbol{B}^c_{ij} \Leftarrow \perp \text{ with probability } k', \text{ otherwise} \tag{15}$$

where $0 \leq k' \leq 1$ is a value chosen so that the sparsity k is maintained, in expectation.

A.3. Summary

We provide a summary of various strengths and weaknesses of all algorithms used in our paper in Table 2.

	Approximating Lower-Level Program	Modeling Target Utility	Evolutionary Algorithms
Assumptions	All errors between the original and minimized data beyond the first order are considered insignificant and ignored. This might not hold when sparsity is high.	The influence of the presence or absence of every value in the data is modelled independently. This can break for datasets with highly correlated features.	No assumptions are made about the structure of the problem setting.
Hyperparameter Sensitivity	No hyperparameters.	Large number of iterations required to create better data when modelling the lower optimization.	Large number of iterations as well as a large active population size will facilitate better solutions.
Consistency	No randomness in the process, but it inherits the randomness of the lower-level learning model.	Highly inconsistent across changing randomness. But consistent across sparsity, i.e. data minimized at lower sparsity will also be minimized at higher sparsity.	Highly inconsistent across both randomness and sparsity. Data minimized may not remain minimized under changing randomness or increasing sparsity.
Convergence Behavior	Provides an exact solution under the given assumptions.	No convergence guarantees.	Convergence guarantees under a sufficiently large number of iterations.
Runtime Considerations	Closed form solution, but requires second-order derivatives. Fast for simpler settings, but does not scale well with either dataset size or model complexity.	Requires training the learning model multiple times. Relatively slower for simpler settings, but scales better with increasing complexity.	Requires training the learning model a significantly larger number of times. Furthermore, needs to be repeated for every unique output sparsity required.
Selection Criteria	Choose in simple settings with a small amount of minimization required for fast and accurate results.	Choose in more complex settings and when algorithm runtime is as important as the accuracy of the method.	Choose in a setting where the accuracy of the method is of the utmost importance, even at the cost of compute.

Table 2: A summary of strengths and weaknesses of various algorithms.

B. Theoretical analysis of baseline techniques

We provide the theoretical properties of the model learned on minimized data for feature selection and individualized random subsampling in this section.

B.1. Feature selection

As introduced in the main text, the feature selection framework sorts the importance of features based on their importance in learning task and then remove the least important features. We denote $S \in [p]$ to be the subset of the weakest features that will be not used. The following Theorem 1 provides the Bayes Mean Squared Error (MSE) when using all features $[p]$ and using a subset of features $[p] \setminus S$.

Theorem 1. *Suppose all input features and labels are jointly Gaussian, i.e.,* $[\boldsymbol{x}, y] \sim \mathcal{N}(\mu, \Sigma)$*, where* $\Sigma = \begin{bmatrix} \Sigma_{\boldsymbol{x},\boldsymbol{x}}, \Sigma_{\boldsymbol{x},y} \\ \Sigma_{y,\boldsymbol{x}}, \Sigma_{y,y} \end{bmatrix}$. *Furthermore, we assume that all input features are mutually independent, i.e., the covariance matrix* $\Sigma_{\boldsymbol{x},\boldsymbol{x}} = diag([\sigma_i^2]_{i=1}^p)$ *is a diagonal matrix and* $\sigma_i^2 = Var[\boldsymbol{x}^i]$ *the variance of* i^{th} *feature. Then the Bayes MSE when using all input features* $[p]$ *and using a subset of input features in* $[p] \setminus S$ *in turn are:* $Var[y] - \sum_{i=1}^p \frac{(Cov(y,\boldsymbol{x}^i))^2}{\sigma_i^2}]$ *and* $Var[y] - \sum_{i \in [p] \setminus S} \frac{Cov(y,\boldsymbol{x}^i))^2}{\sigma_i^2}$

Theorem 1 suggests data minimization procedure based on feature selection introduces an additional MSE of $\sum_{i=1}^p \frac{(Cov(y,\boldsymbol{x}^i))^2}{\sigma_i^2} - \sum_{i \in [p] \setminus S} \frac{Cov(y,\boldsymbol{x}^i))^2}{\sigma^2)_i} = \sum_{i \in S} \frac{(Cov(y,\boldsymbol{x}^i))^2}{\sigma_i^2}$. As long as all features in the removed feature set S have small correlation with the label, i.e. $Cov(y, \boldsymbol{x}^i) \approx 0$, such additional MSE is neligible.

Proof. Our proof relies on the properties of multivariate Gaussian variables. In particular, when feature and label are jointly Gaussian, $[\boldsymbol{x}, y] \sim \mathcal{N}(\mu, \Sigma)$ then for any subset of features $A \in [p]$ we can derive the following conditional density of label y given partial input features $\boldsymbol{x}_A$(Bishop & Nasrabadi, 2006):

$$P(y|x_A) = \mathcal{N}(\mu_y + \Sigma_{\boldsymbol{x}_A,y}\Sigma_{y,y}^{-1}\Sigma_{y,\boldsymbol{x}}, \Sigma_{y,y} - \Sigma_{\boldsymbol{x}_A,y}\Sigma_{\boldsymbol{x},\boldsymbol{x}}^{-1}\Sigma_{y,\boldsymbol{x}_A}).$$

In the above equation, $\Sigma_{y,y}$ is the variance of the label y, i.e., $\Sigma_{y,y} = \text{Var}(y)$, while $\Sigma_{\boldsymbol{x}_A,y}$ or $\Sigma_{y,\boldsymbol{x}_A}$ is the covariance between y and a subset of features $\boldsymbol{x}_A$. The Bayes MSE for Gaussian distribution is just the conditional variance (Tibshirani, 2015). Hence the Bayes MSE when using all features, i,e $A = [p]$ is: $\Sigma_{y,y} - \Sigma_{y,\boldsymbol{x}}\Sigma_{\boldsymbol{x},\boldsymbol{x}}^{-1}\Sigma_{\boldsymbol{x},y}$. By the assumption that the input features are mutually independent $\Sigma_{\boldsymbol{x},\boldsymbol{x}} = \text{diag}([\sigma_i^2]_{i=1}^p)$, the Bayes MSE using all features can be further reduced as: $\text{Var}[y] - \Sigma_{y,\boldsymbol{x}}\text{diag}([\frac{1}{\sigma_i^2}]_{i=1}^p)\Sigma_{\boldsymbol{x},y} = \text{Var}[y] - \sum_{i \in [p]} \frac{(Cov(y,\boldsymbol{x}_i))^2}{\sigma_i^2}$.

Similarly, we can derive the Bayes MSE when using a subset of features $d \setminus S$ as: $\text{Var}[y] - \sum_{i \in [d] \setminus S} \frac{(Cov(y,\boldsymbol{x}_i))^2}{\sigma_i^2}$. □

B.2. Personalized random subsampling

As introduced in the above, the personalized random subsampling method works by randomly setting the entries of the minimization $\boldsymbol{B}_{ij} = \perp$ with a probability k_p, which controls the sparsity of the minimized dataset (Here, $k_p = \frac{k}{np}$). This can also be rewritten as randomly selecting a subset $S \in \{(i, j)|i \in [n], j \in [p]]\}$ of a given size $|S| = np - k$, and remove the entries in S. To understand the theoretical behaviour of this method, we first consider the optimal model parameter $\boldsymbol{\theta}^*(\mathbf{X}) = \text{argmin}_{\boldsymbol{\theta}} J(\boldsymbol{\theta}; \mathbf{X}, \mathbf{Y})$ as a function of the data $\mathbf{X}$. Using this notation, the model learnt on minimized data can be represented as $\hat{\theta} = \theta^*(\mathbf{X}')$ while the model parameter learned on original data $\overset{\star}{\boldsymbol{\theta}} = \theta^*(\mathbf{X})$. We then have the following Lemma 1 that derives the gradient of the optimal model parameter $\theta^*(\mathbf{X})$ w.r.t data $\mathbf{X}$

Lemma 1 (Sensitivity of model parameter w.r.t input $\mathbf{X}$)**.** *Assume the loss function* $J(\boldsymbol{\theta}; \mathbf{X}, \mathbf{Y})$ *is differentiable w.r.t* $\boldsymbol{\theta}$ *and* $\mathbf{X}$*, suppose* $\boldsymbol{\theta}^*(\mathbf{X}) = \text{argmin}_{\boldsymbol{\theta}} J(\boldsymbol{\theta}; \mathbf{X}, \mathbf{Y})$ *then the following holds:*

$$\frac{\partial \boldsymbol{\theta}^*(\mathbf{X})}{\partial \mathbf{X}} = -H^{-1}G, \tag{16}$$

where $H = \frac{\partial^2 J(\boldsymbol{\theta}^*(\mathbf{X});\mathbf{X},\mathbf{Y})}{\partial \boldsymbol{\theta}^2}$ *is the Hessian matrix of the loss w.r.t model parameter, while* $G = \frac{\partial^2 J(\boldsymbol{\theta}^*(\mathbf{X});\mathbf{X},\mathbf{Y})}{\partial \boldsymbol{\theta}, \partial \mathbf{X}}$ *is the second derivative of the loss w.r.t model parameter and input data.*

Proof. Since the model parameter $\boldsymbol{\theta}^*(\mathbf{X}) = \text{argmin}_{\boldsymbol{\theta}} J(\boldsymbol{\theta}; \mathbf{X}, \mathbf{Y})$ hence the gradient of the loss $J(.)$ w.r.t model parameter vanishes at $\boldsymbol{\theta}^*(\mathbf{X})$. In other words

$$\frac{\partial J(\boldsymbol{\theta}^*(\mathbf{X}); \mathbf{X}, \mathbf{Y})}{\partial \boldsymbol{\theta}} = \mathbf{0}^T.$$

Take the derivative w.r.t $\mathbf{X}$ on both sides of the above equation, it follows:

$$\frac{\partial}{\partial \mathbf{X}} \frac{\partial J(\boldsymbol{\theta}^*(\mathbf{X}); \mathbf{X}, \mathbf{Y})}{\partial \boldsymbol{\theta}} = \mathbf{0}^T. \tag{17}$$

The L.H.S can be rewritten as

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\theta}} \frac{\partial J(\boldsymbol{\theta}^*(\mathbf{X}); \mathbf{X}, \mathbf{Y})}{\partial \mathbf{X}} &= \frac{\partial}{\partial \boldsymbol{\theta}}\Big[\frac{\partial J(\boldsymbol{\theta}^*(X); \mathbf{X}, \mathbf{Y})}{\boldsymbol{\theta}} \frac{\partial \boldsymbol{\theta}^*(\mathbf{X})}{\partial \mathbf{X}} + \frac{\partial J(\boldsymbol{\theta}^*(X); \mathbf{X}, \mathbf{Y})}{\partial \mathbf{X}} \frac{\partial \mathbf{X}}{\partial \mathbf{X}}\Big] \\ &= \frac{\partial^2 J(\boldsymbol{\theta}^*(\mathbf{X}); \mathbf{X}, \mathbf{Y})}{\partial \boldsymbol{\theta}^2} \frac{\partial \boldsymbol{\theta}^*(\mathbf{X})}{\partial \mathbf{X}} + \frac{\partial^2 J(\boldsymbol{\theta}^*(\mathbf{X}); \mathbf{X}, \mathbf{Y})}{\partial \boldsymbol{\theta} \partial \mathbf{X}},\end{aligned}$$

where the first equation is due to the chain rule. If we put $H = \frac{\partial^2 J(\boldsymbol{\theta}^*(\mathbf{X}); \mathbf{X}, \mathbf{Y})}{\partial \boldsymbol{\theta}^2}$ (the Hessian matrix) and $G = \frac{\partial^2 J(\boldsymbol{\theta}^*(\mathbf{X}); \mathbf{X}, \mathbf{Y})}{\partial \boldsymbol{\theta} \partial \mathbf{X}}$, then the L.H.S of Equation 17 can be rewritten as:

$$H \frac{\partial \boldsymbol{\theta}^*(\mathbf{X})}{\partial \mathbf{X}} + G = \mathbf{0}^T,$$

which implies $\frac{\partial \boldsymbol{\theta}^*(\mathbf{X})}{\partial \mathbf{X}} = -H^{-1}G$.

□

Lemma 1 tells us how much the optimal model parameter changes when the data input changes. Based on this Lemma 1 we can prove the following Lemma 2 regarding the target utility.

Lemma 2 (Sensitivity of target utility w.r.t input $\mathbf{X}$)**.** *Given the same settings and conditions as in Lemma 1, then the following holds:*

$$\frac{\partial J(\boldsymbol{\theta}^*(\mathbf{X}), \mathbf{X}, \mathbf{Y})}{\partial \mathbf{X}} = -LH^{-1}G, \tag{18}$$

where H and G were provided in Lemma 1, and $L = \frac{\partial J(\boldsymbol{\theta}^*(\mathbf{X}), \mathbf{X}, \mathbf{Y})}{\partial \boldsymbol{\theta}}$ is the gradient of loss w.r.t model parameter.

Proof. The proof is based directly on the chain rule:

$$\frac{\partial J(\boldsymbol{\theta}^*(\mathbf{X}), \mathbf{X}, \mathbf{Y})}{\partial \mathbf{X}} = \frac{\partial J(\boldsymbol{\theta}^*(\mathbf{X}), \mathbf{X}, \mathbf{Y})}{\partial \boldsymbol{\theta}^*(bX)} \frac{\boldsymbol{\theta}^*(bX)}{\partial \mathbf{X}} = -LH^{-1}G,$$

where the last equation is by Lemma 1. □

Based on Lemma 2 we have the following Theorem 2 that derives the target utility $J(\hat{\boldsymbol{\theta}}; \mathbf{X}, \mathbf{Y})$ bound based on the original utiltiy $J(\overset{\star}{\boldsymbol{\theta}}; \mathbf{X}, \mathbf{Y})$ as follows:

Theorem 2. *Consider the personalized random subsampling framework, in which all features in a random set $S \in \{(i, j) | i \in [n], j \in [p]\}$ is removed to form the minimized dataset $\mathbf{X}'$, then the following holds:*

$$J(\hat{\boldsymbol{\theta}}; \mathbf{X}, \mathbf{Y}) \leq J(\overset{\star}{\boldsymbol{\theta}}; \mathbf{X}, \mathbf{Y}) + \sqrt{2|S|} \|\mathbf{X}\|_\infty \|LH^{-1}G\|_2 + O(|S| \|X\|_\infty^2). \tag{19}$$

Proof. The proof relies on the first-order Taylor approximation. First, we consider both optimal model parameters $\hat{\boldsymbol{\theta}}, \overset{\star}{\boldsymbol{\theta}}$ as a function of the input data, i.e., $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}^*(\mathbf{X}')$ and $\overset{\star}{\boldsymbol{\theta}} = \boldsymbol{\theta}^*(\mathbf{X})$. Based on the first-order Taylor approximation around $\mathbf{X}$ it follows that:

$$J(\boldsymbol{\theta}^*(\mathbf{X}'); \mathbf{X}, \mathbf{Y}) \approx J(\boldsymbol{\theta}^*(\mathbf{X}); \mathbf{X}, \mathbf{Y}) + (\mathbf{X}' - \mathbf{X})^T \frac{\partial J(\boldsymbol{\theta}^*(\mathcal{X}), \mathbf{X}, \mathbf{Y})}{\partial \mathbf{X}} + O(\|\mathbf{X}' - \mathbf{X}\|_2^2) \tag{20}$$

By Lemma 2, $\frac{\partial J(\boldsymbol{\theta}^*(\mathcal{X}),\mathbf{X},\mathbf{Y})}{\partial \mathbf{X}} = -LH^{-1}G$. Furthermore

$$\begin{aligned} \|\mathbf{X}' - \mathbf{X}\|_2^2 &= \sum_{i=1}^{n}\sum_{j=1}^{p}(\mathbf{X}'_{i,j} - \mathbf{X}_{i,j})^2 = \sum_{(i,j)\in S}(\mathbf{X}'_{i,j} - \mathbf{X}_{i,j})^2 \\ &\leq \sum_{(i,j)\in S} 2\|\mathbf{X}\|_\infty^2 = 2|S|\,\|\mathbf{X}\|_\infty^2, \end{aligned} \tag{21}$$

where the second equation is due to the fact we only remove features in S while the other entries are kept the same. The inequality is due to the fact that $|\mathbf{X}'_{i,j} - \mathbf{X}_{i,j}| \leq 2\max_{i,j}|\mathbf{X}_{i,j}| = 2\|\mathbf{X}\|_\infty$, since the the imputed data is in the range $\mathbf{X}'_{i,j} \in [\min_{i,j}\mathbf{X}_{i,j}, \max_{i,j}\mathbf{X}_{i,j}]$.

By Equation 20 and Cauchy-Schwarz inequality for vectors it follows that:

$$\begin{aligned} (\mathbf{X}' - \mathbf{X})^T \frac{\partial J(\boldsymbol{\theta}^*(\mathcal{X}), \mathbf{X}, \mathbf{Y})}{\partial \mathbf{X}} &\leq \|\mathbf{X}' - \mathbf{X}\|_2\| - LH^{-1}G\|_2 \\ &\leq \sqrt{2|S|}\|\mathbf{X}\|_\infty\|LH^{-1}G\|_2. \end{aligned} \tag{22}$$

Applying the results from Equation 21 and Equation 22 to Equation 20 we verify the correctness of the statement.

□

C. Membership inference attacks and data minimization

In this section, we focus on threat models outside the "wall", specifically addressing inference attacks on the trained model without direct access to the minimized dataset.

C.1. Membership inference risk and inference attack

Membership Inference Risk (MIR). Membership inference attacks (Shokri et al., 2017) aim to discern if an individual's data was in the dataset before minimization, focusing on an attacker accessing the model $\hat{\theta}$ trained on this minimized dataset. In these attacks, the adversary calculates a likelihood score $L(\boldsymbol{x}_q)$ for each query $\boldsymbol{x}_q$, representing its probability of being in the original dataset $\mathbf{X}$ using the model $\hat{\theta}$. The score is given by $L(\boldsymbol{x}_q) = \Pr\left[\boldsymbol{x}_q \in \mathbf{X} \,|\, \hat{\theta}\right]$. Using these scores for both original dataset $\mathbf{X}$ and non-members $\mathbf{X}_{nm}$, binary membership predictions can be made at any threshold t, denoted as $\mathbb{1}_{L(\boldsymbol{x}_q)\geq t}$. The overall Membership Inference Risk (MIR) is assessed as the area under the curve (AUC) of true positive rates ($\overrightarrow{\text{tpr}}$) and false positive rates ($\overrightarrow{\text{fpr}}$) across various thresholds:

$$\text{MIR} = \text{AUC}(\overrightarrow{\text{tpr}}, \overrightarrow{\text{fpr}}) \tag{23}$$

Attack Details. The attacker aims to determine the presence or absence of an individual in the original training dataset. We use the SOTA membership inference attack RMIA (Zarifzadeh et al., 2023), by training 8 reference models on random 50% subsets of the public data split. We operate under the practical assumption that the adversary is unaware of the data minimization applied before training the target model and evaluate the membership inference on the original dataset.

C.2. Privacy leakage through membership inference

Finally, we assess the membership inference risk under various minimization algorithms in Figure 7(a). As previously mentioned, information leakage through a trained model is not an expected benefit of data minimization, which mainly aims to address data breach scenarios. Nevertheless, we still observe that certain minimization algorithms are effective at reducing membership inference risks with decreasing dataset size, i.e., models trained on minimized datasets leak less information. Yet, these improvements are not perfectly aligned, carrying forward the same trends we saw in the two data breach scenarios above.

C.3. DP-SGD to improve privacy-utility trade-off

We analyze modifications in data minimization to counter membership inference. An effective mitigation can be obtained by introducing a differentially private learning algorithm, DP-SGD (Abadi et al., 2016), into the lower-level program of the

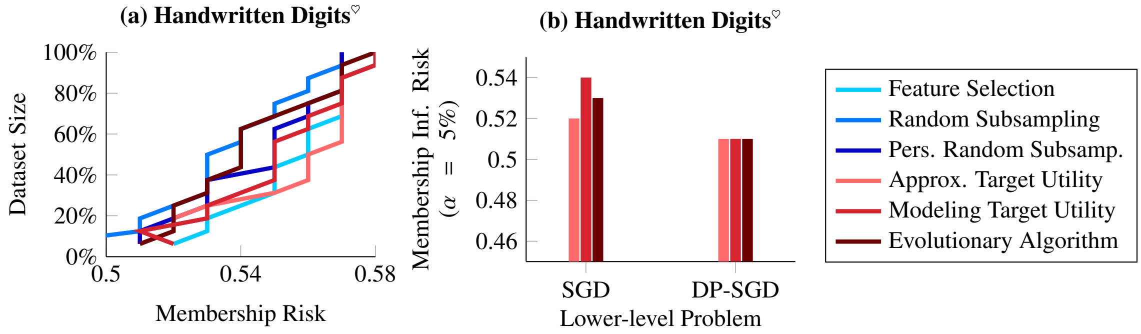


Fig. 7: Membership inference risks under changing sparsity on the handwritten digits dataset. The minimization algorithms can reduce inference risks and are pushed to even better trade-offs by introducing DP-SGD in the lower-level objective.

bi-level optimization in (1). We test its compatibility with data minimization by re-evaluating membership inference in this new setting. Note, we do not introduce DP-SGD into utility calculation, i.e., once the data is minimized, the rest of the pipeline outside the "wall" remains unchanged.

Evaluating DP-SGD Modifications. By incorporating DP-SGD into the minimization optimization without altering other components, we want to assess the compatibility between these two methods. The results, collected in Figure 7(b), clearly demonstrate a reduction in membership inference risks at the same accuracy threshold, with prominent improvements for methods that were more susceptible to information leakage, such as modelling target utility algorithms. DP-SGD is indeed compatible with minimization, reinforcing the benefits of considering privacy during minimization.

D. Additional experiments

D.1. Results on additional datasets

We also provide results for utility and privacy across changing sparsity on additional datasets. This includes the wine quality dataset (Cortez et al., 2009), a dataset containing various attributes of 6,463 unique wines and a binary label to classify them as red or white wine, and the ACSIncome and ACSEmployment tasks of the folktables dataset (Ding et al., 2021), with census information about individuals and a label marking whether their income is above $50, 000 for ACSIncome or whether they are employed or not for ACSEmployment. Similar to 20 newsgroup dataset, we only choose a random subset of 5000 data points from both ACSIncome and ACSEmployment datasets. We also provide detailed results on the bank dataset, the handwritten digits dataset, and the 20 newsgroup dataset, in this section.

The results for utility and privacy are collected in Figure 8 and Figure 9, respectively. The trends of utility match the behaviour seen in the main text, i.e., these datasets have redundant information that can be removed without suffering any performance loss. Similarly, for privacy results, we see the trends of main text like feature selection highly misaligned with re-identification, and all methods containing high reconstruction risks even after extreme minimization, replicated in these datasets. Thus, there is a clear misalignment between data minimization and privacy expectations.

D.2. Additional results for privacy-based modifications

We provide additional results for privacy-based modifications to the data minimization algorithm on other datasets, as well as the raw trends on the handwritten dataset. The results for the handwritten dataset are collected in Figure 10, for the bank dataset are collected in Figure 11, and for the employment dataset are collected in Figure 12. The results show similar improvement as in the main text for the bank dataset, however, we don't see similar benefits on the employment dataset. We believe that's because these results are quite sensitive to the choice of hyperparameter β, and a more thorough search for β can show improvements for other datasets as well. Despite this discrepancy, the aim of these modifications was not to propose a novel method of incorporating privacy in minimization, but instead to highlight that minimization and privacy are compatible, and thus one can perform data minimization in line with the regulations while making sure they respect individual privacy in the dataset.

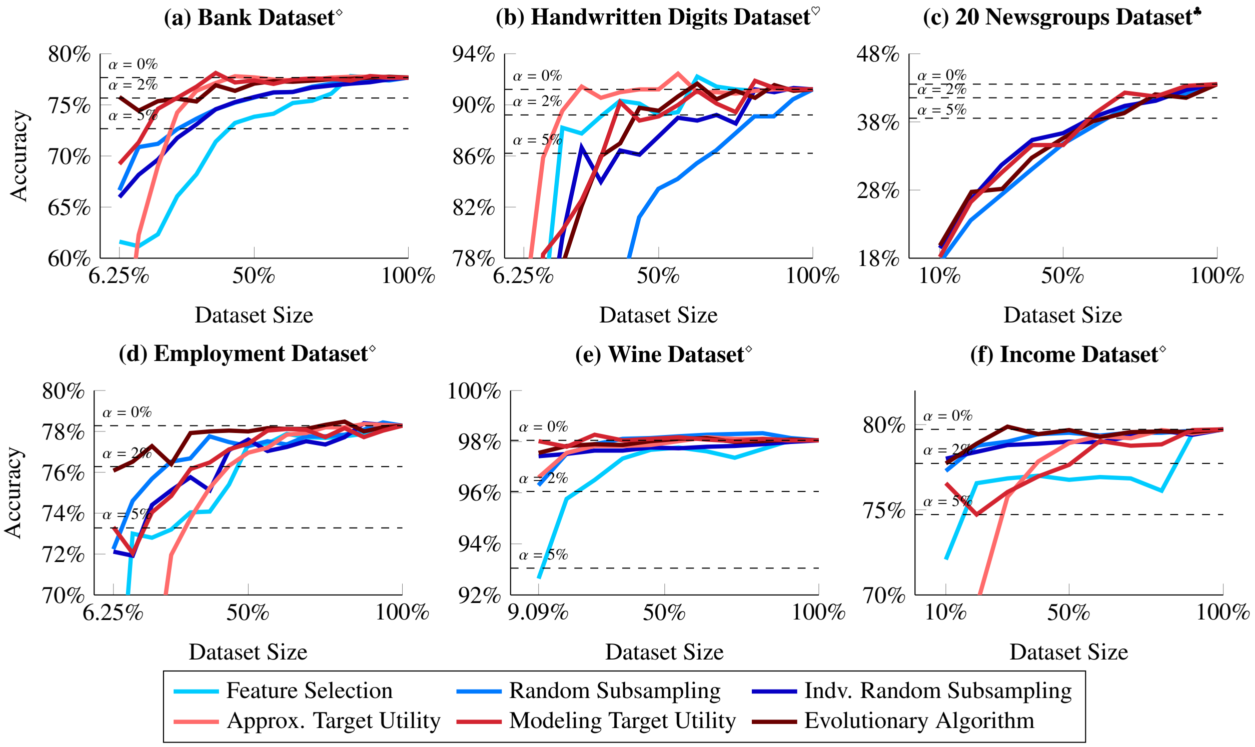


Fig. 8: Utility of the minimized data across various sparsities on all datasets.

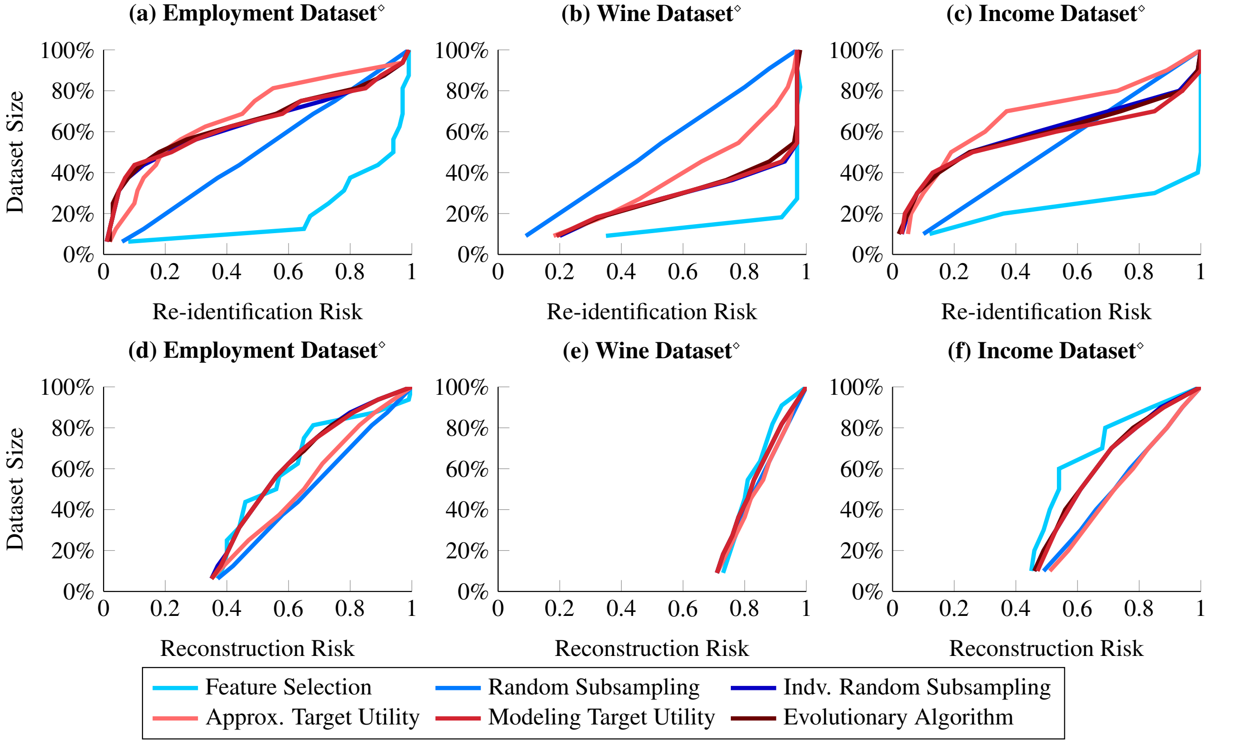


Fig. 9: Re-identification and reconstruction risks under changing sparsity on additional datasets.

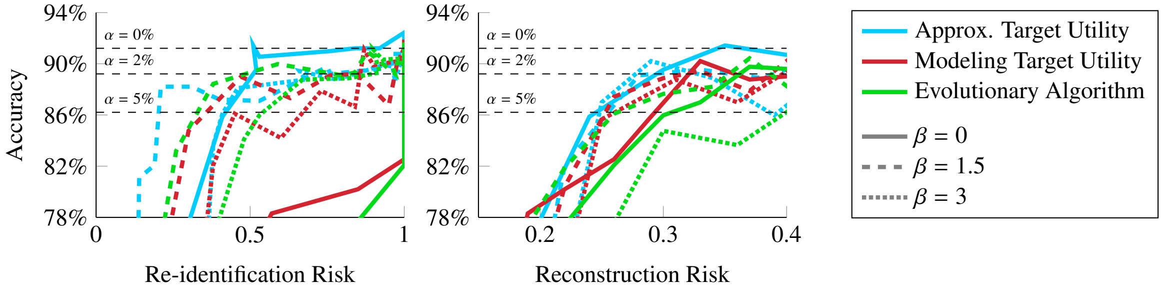


Fig. 10: Re-identification and reconstruction risks (zoomed in) on the handwritten digits dataset, using feature-level privacy scores.

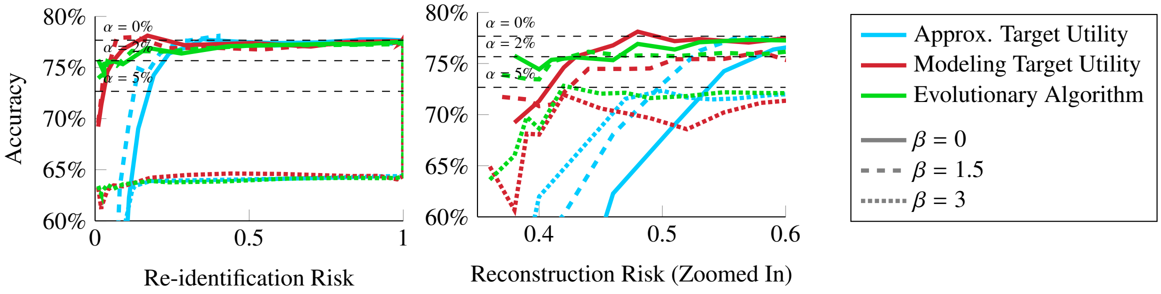


Fig. 11: Re-identification and reconstruction risks (zoomed in) on the bank dataset, using feature-level privacy scores.

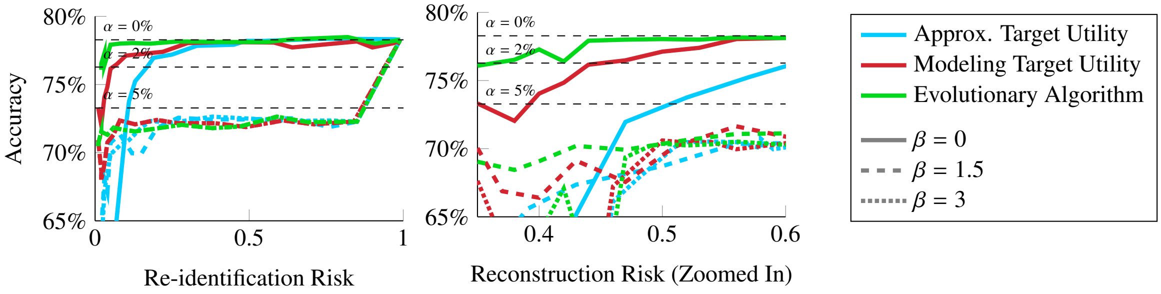


Fig. 12: Re-identification and reconstruction risks (zoomed in) on ACSEmployment, using feature-level privacy scores.