HEC MONTRÉAL
École affiliée à l'Université de Montréal

Investigating the Privacy/Utility tradeoff through the lens of data valuation

par
Patrick Mesana

Thèse présentée en vue de l'obtention du grade de Ph. D. en administration
(spécialisation Sciences de la décision)

Novembre 2025

HEC MONTRÉAL
École affiliée à l'Université de Montréal

Cette thèse intitulée :

Investigating the Privacy/Utility tradeoff through the lens of data valuation

Présentée par :

Patrick Mesana

a été évaluée par un jury composé des personnes suivantes :

Georges Zaccour
HEC Montréal
Président-rapporteur

Gilles Caporossi
HEC Montréal
Directeur de recherche

Sébastien Gambs
Université du Québec à Montréal
Codirecteur de recherche

Gregory Vial
HEC Montréal
Membre du jury

Yves-Alexandre de Montjoye
Imperial College London
Examinateur externe

Résumé

Cette thèse explore les méthodes de valuation des données comme un moyen de quantifier la valeur des points de données individuels tout en mesurant les compromis entre l'utilité des données et les risques pour la vie privée dans les contextes de l'analyse de données et de l'apprentissage automatique. Nous nous appuyons sur des perspectives de l'informatique, de la gestion des données et de la théorie des jeux pour explorer comment différentes hypothèses sur la vie privée façonnent la valeur des données. La thèse est structurée en trois essais indépendants, chacun abordant des facettes du compromis entre utilité et vie privée.

Le premier essai introduit un cadre basé sur la simulation pour évaluer les stratégies de publication de données préservant le format des données, en se concentrant sur un contexte d'organisation spécifique et en s'inspirant des notions juridiques du risque de ré-identification. Il utilise la théorie des jeux coopératifs comme un outil explicatif pour mettre en évidence les attributs qui contribuent le plus au risque pour la vie privée. Il permet également des comparaisons entre des stratégies de base telles que la suppression de données et des techniques d'anonymisation comme un algorithme de k-anonymat. Le deuxième présente WaKA, une méthode pour analyser les contributions individuelles des données dans les modèles k-plus proches voisins en calculant les distances de Wasserstein entre les distributions de perte du modèle, distinguant l'utilité des données du risque d'inférence d'appartenance. Le troisième essai modélise la valorisation des données comme un jeu de divulgation d'informations entre deux agents : un propriétaire de données qui contrôle stratégiquement les fuites d'informations par le biais de la confidentialité différentielle, et un consommateur de données qui sélectionne les données sous des contraintes d'utilité et de budget.

À travers ces contributions, la thèse montre que la valuation des données reflète des choix implicites sur les intérêts qui sont priorisés. Plutôt que de traiter la valuation des données comme une simple technique d'apprentissage automatique, la thèse montre comment elle peut servir de prisme technique pour explorer une perspective utilitariste de la vie privée.

Mots-clés

Valuation des Données, Compromis Utilité-Vie Privée, Apprentissage Automatique, Anonymisation des Données, Risque de Ré-identification, Inférence d'Appartenance, Théorie des Jeux, Confidentialité Différentielle.

Méthodes de recherche

Évaluation du Risque pour la Vie Privée, Évaluation par Simulation, Conception d'Algorithmes, Analyse Quantitative, Modélisation par la Théorie des Jeux.

Abstract

This thesis explores data valuation methods as a means to quantify the value of individual data points while measuring the tradeoffs between data utility and privacy risks in data analytics and machine learning settings. We draw on perspectives from computer science, data management, and game theory to explore how different assumptions about privacy shape what data is worth. The thesis is structured as three independent essays, each tackling facets of the privacy-utility tradeoff.

The first essay introduces a simulation-based framework for evaluating format-preserving data release strategies, focusing on a specific organisation problem and inspired by legal notions of the re-identification risk. It uses cooperative game theory as an explanatory tool to highlight which attributes contribute most to privacy risk. It also enables comparisons between basic strategies such as data removal and anonymization techniques like a k-anonymity algorithm. The second presents WaKA, a method for analyzing individual data contributions in k-nearest neighbor models by computing Wasserstein distances between prediction loss distributions—distinguishing data utility from membership inference risk. The third essay models data valuation as a an information disclosure game between two agents: a data owner who strategically controls information leakage through differential privacy, and a data consumer who selects data under utility and budget constraints.

Across these contributions, the thesis shows that data valuation reflects implicit choices about whose interests are prioritized — Rather than treating data valuation as merely a machine learning technique, the thesis shows how it can serve as a technical lens for exploring the incentive-driven view of privacy.

Keywords

Data Valuation, Privacy-Utility Tradeoff, Machine Learning, Data Anonymization, Re-identification Risk, Membership Inference, Game Theory, Differential Privacy.

Research Methods

Privacy Risk Assessment, Simulation-based Evaluation, Algorithm Design, Quantitative Analysis, Game-theoretic Modeling.

Contents

List of Tables

List of Figures

List of Acronyms

AI	Artificial Intelligence
AIC	Adult Income Census
ASR	Attack Success Rate
AUC	Area Under the Curve
BM	Bank Marketing
CCPA	California Consumer Privacy Act
CCD	Credit Card Dataset
CPRA	California Privacy Rights Act
CTGAN	Conditional Tabular Generative Adversarial Network
DC	Data Consumer
DO	Data Owner
DP	Differential Privacy
DP-SGD	Differentially Private Stochastic Gradient Descent
DSV	Data Shapley Value
FPR	False Positive Rate

GANs	Generative Adversarial Networks
GDPR	General Data Protection Regulation
HPC	High-Performance Computing
IDG	Information Disclosure Game
k-NN	k-Nearest Neighbor
LDP	Local Differential Privacy
LiRA	Likelihood Ratio Membership Inference Attack
LLM	Large Language Model
LOO	Leave-One-Out
MAB	Multi-Armed Bandit
MCC	Matthews Correlation Coefficient
MIA	Membership Inference Attack
NIST	National Institute of Standards and Technology
OECD	Organisation for Economic Co-operation and Development
OSFI	Office of the Superintendent of Financial Institutions
PCI DSS	Payment Card Industry Data Security Standard
PETs	Privacy Enhancing Technologies
PII	Personally Identifiable Information
PPDP	Privacy-Preserving Data Publishing
PrivInf	Privacy Influence

ROC	Receiver Operating Characteristic
RRS	Re-identification Risk Score
RSV	Re-identification Shapley Value
SGD	Stochastic Gradient Descent
TPR	True Positive Rate
TVAE	Tabular Variational Autoencoder
UCB	Upper Confidence Bound
VAE	Variational Autoencoder
ViT	Vision Transformer
WaKA	Wasserstein k-NN Attribution

To my beloved family.

Acknowledgements

I would first like to thank my two supervisors. To Gilles Caporossi, who was the first person I spoke to about my PhD project and who supported me from the very beginning. You gave me the freedom and trust I needed to explore the questions and ideas I cared about. Our many discussions were always rich and intellectually stimulating. To Sébastien Gambs, whose expertise in privacy and security I was lucky to benefit from. Thank you for your guidance not only on my research topic but also on methodology and scientific writing. Your fairness, rigor, and thoughtful reviews were an important compass for me throughout this journey.

I would also like to thank Gregory Vial for his openness and empathy. His curiosity and broad expertise—especially in data governance—gave me valuable perspective. To Hadrien Lautraite, thank you for your friendship and the many debates we had on privacy. Many of the ideas I developed were sparked by our conversations. To Julien Crowe, thank you for taking a chance on an outlier student like me and for trusting me to lead a research project at the National Bank of Canada. I also thank my co-authors, Pascal Jutras and Clément Benesse, for their valuable feedback during the writing of our papers. I gratefully acknowledge the PhD program at HEC Montréal, as well as Fin-ML & NSERC, and Mitacs for their financial support.

I want to express my deepest gratitude to those who are dearest to me in my personal life. To my parents, who have inspired me since childhood, who taught me never to settle, and who passed on strong values that have guided me throughout life. To my whole family, in Canada, the U.S., and France—thank you for your support during this long journey.

And finally, to the ones I live for—my brilliant wife Marie, who believed in me and supported me in so many ways, and my two unique daughters, Anna and Charlie. You give me the courage to be better.

Preface

For the sake of transparency, I note that I used generative artificial intelligence tools during the writing and experimentation phases of this thesis. Their use was limited to helping improve the clarity of certain English passages—since English is not my first language—and to generating preliminary code snippets to accelerate early experimentation. Except for the research in Chapter 1, most of the code used in this thesis has been open-sourced and is available on Github.

General Introduction

Over the past decade, data has become a foundational asset for organizations—driving innovation, shaping business models, and powering machine learning systems. This shift has sparked growing interest in how data creates value, and under what conditions. But as data becomes more tightly regulated and individuals assert greater control over their information, questions arise not just about privacy, but about its implications for the utility and value of data itself. To understand how privacy constrains or enables the value derived from data, it is essential to first examine what data privacy means.

Perspectives on Privacy

Privacy is a multifaceted concept with various definitions, depending on the lens through which it is analyzed. More precisely, four interrelated perspectives frequently appear in both academic and practical discussions of privacy.

First, *the legal perspective*, often embodied by the persona of a lawyer or legal expert, defines privacy primarily through compliance with regulations and laws. Historically, privacy-related legal frameworks have evolved significantly in response to technological and societal changes. One of the foundational legal conceptions of privacy—famously articulated by Warren and Brandeis in 1890—is the *right to be let alone*, a notion that continues to influence legislation even today. Early international standards—such as the Organisation for Economic Co-operation and Development (OECD) privacy guidelines (OECD, 2013) provided foundational principles like transparency, accountability, and purpose limitation. Modern privacy frameworks—including the European Union's General

Data Protection Regulation (Parliament and of the European Union, 2016), the California Privacy Rights Act (California State Legislature, 2020), and Law 25 in Quebec (National Assembly of Québec, 2021)—build on these earlier principles while introducing additional requirements. While these modern legal instruments share common objectives, they also reflect local legal, cultural, and institutional specificities.

Second, *the contextual perspective* is often seen through the eyes of a citizen or employee, who defines privacy in a normative way based on their personal and professional context. This perspective sees privacy as dynamic and situational, shaped by norms and practices around information flow—this is formalized in theories such as contextual integrity (Nissenbaum, 2004). Rather than framing privacy purely as control over information or secrecy, contextual integrity holds that privacy is maintained when information flows align with the norms of a specific context—including who is sharing, who is receiving, what is being shared and under what conditions. Thus, privacy is not violated simply when data is shared, but when it is shared outside of these contextual expectations. For example, one may expect a bank employee at a local branch to access his financial records when he is helping with respect to a transaction, but not a data analyst at headquarters to identify individual customers while data mining aggregate trends. This principle also surfaces in organizational data governance, in which privacy decisions are embedded in workflows, access controls and business processes. Regulations such as the GDPR reinforce this perspective by requiring that privacy constraints be integrated directly into how data is collected, stored and analyzed (Cavoukian, 2009; Hoofnagle et al., 2019; Mahanti and Mahanti, 2021).

Third, *the risk-based perspective* is typically associated with the security researcher or expert, who uses frameworks like the NIST Privacy Framework to assess and manage privacy risks in information systems (National Institute of Standards and Technology, 2020). It reflects a technical and scientific approach that formally evaluates risks such as re-identification, information leakage or unauthorized access. This approach often focuses on the likelihood and severity of harm, including those associated with data breaches caused by security vulnerabilities, misconfigurations or malicious attacks. In particular, these incidents can result in financial, reputational or psychological harm to individuals, as well as legal or regulatory consequences for organizations. However, not all privacy risks stem from large-scale events, as even small, routine or cumulative disclosures—so-called micro privacy breaches—can affect individuals in subtle but meaningful ways. A notable example includes a Quebec case where a nurse repeatedly accessed patient records without authorization, leading to disciplinary action despite the absence of a large-scale breach (Danakas, 2024).

Finally, *the utilitarian perspective* is often embraced by economists and many people in general, treating privacy as an economic decision. Individuals weigh the costs and benefits of disclosing their data and can be influenced by incentives or disincentives. A common interpretation under this view is the notion of a privacy calculus (Dinev and Hart, 2006), in which individuals subjectively assess potential benefits against perceived risks when deciding whether to share personal information. However, this calculus is often shaped by bounded rationality, as individuals frequently make these decisions without full understanding of how their data will be used, underestimate long-term risks, etc. (Acquisti and Grossklags, 2005; Acquisti et al., 2016). Moreover, on the organizational side, efforts to reduce privacy risks—through data minimization, obfuscation or anonymization—can diminish data utility and constrain the economic value that can be extracted from it (Acquisti, 2010). In this perspective, as individuals increasingly gain the ability to control, withhold or monetize their data, the interdependence between personal choices and institutional practices brings a central question to the foreground for both individuals and organizations: “How to define the value of personal data?”.

Data Valuation

The interplay between data privacy and utility is a critical consideration for organizations today. As they navigate the complexities of privacy regulations, contextual norms, security risks, and economic incentives, quantifying the value of data becomes essential. In its broadest sense, data valuation refers to the use of quantitative methods to assess the value of data—not limited to personal information (Bendechache et al., 2023; Fleckenstein et al., 2023; Ghorbani and Zou, 2019). Understanding how to systematically attribute value to individual data contributions is poised to become an essential tool for addressing the challenges of data management in the future. This thesis explores these techniques, focusing on their implications for balancing data utility with privacy considerations in analytics and machine learning. More precisely, we assume the presence of an organization, or data curator, that uses collected data within legal and regulatory boundaries. For instance, a financial institution may gather both descriptive (*e.g.*, age, sex or occupation) and behavioral (*e.g.*, credit history or spending habits) data to predict outcomes like loan repayment. Such predictions rely on machine learning models, in which utility is typically assessed using global performance metrics like classification accuracy or average loss under empirical risk minimization (Vapnik, 1991). This evaluation should ideally be performed on a separate test set to ensure an unbiased estimate of model performance on unseen data. While these metrics reflect the overall model utility, they reveal little about the marginal utility of specific data points. Organizations have long been interested in estimating the relative importance of individual data points—typically to refine models, detect noise or prioritize data sources—rather than to assess the value of user contributions. This concern relates to the broader concept of data quality (Strong et al., 1997), often addressed using heuristic or rule-based approaches. More recently, efforts have shifted towards formally estimating how individual points influence model learning, which is particularly challenging in large datasets, in which incremental contributions are small and interactions between data points create non-trivial synergies. In particular, numerous data attribution methods have been developed, such as influence functions (Koh and Liang, 2017b), which, despite their practical applications, possess certain limitations (Basu et al., 2021; Ghorbani et al., 2019b). For example, their estimates can become unstable near the decision boundary, where small perturbations to the input can cause large, erratic shifts in attribution—undermining their reliability for interpretation.

Coincidentally, methods based on the Shapley value—a solution concept from cooperative game theory—can serve both organizational and individual needs. Originally designed to allocate gains fairly among players in a coalition, the Shapley value satisfies a set of axioms: symmetry (equal contributors receive equal value), linearity (values can be added across games), efficiency (the total value is fully distributed), and null-player (non-contributors receive zero). This has motivated the development of methods such as Data Shapley (Ghorbani and Zou, 2019; Jia, Dao, Wang, Hubis, Hynes, et al., 2019b) as well as alternative solution concepts like the Banzhaf value, Beta-Shapley, and the Core, each offering different tradeoffs in fairness, robustness, or computational cost (Kwon and Zou, 2022; J. T. Wang and Jia, 2023; Yan and Procaccia, 2021). Though computationally demanding and model-dependent (e.g., k-NN Shapley by Jia, Dao, Wang, Hubis, Gurel, et al., 2019), these methods can accurately identify influential or problematic data points, whose removal significantly improves model performance. This effect is notably observed in fields such as medical imaging, where algorithmic valuation challenges conventional approaches based on manual inspection (Tang et al., 2021b). Therefore, compared to previously used leave-one-out methods, which assess influence via single-point removal, Shapley-based valuation better captures interactions between data points by modeling coalitions and synergies among players (data points) in a cooperative game and offers a more principled approach (grounded in Shapley fairness axioms) to quantifying individual data value.

Data valuation, beyond attributing predictive utility to data points, offers a technical lens through which to examine the privacy-utility tradeoff. If the marginal contribution of each data point to a machine learning model can be quantified, for example through Shapley-based methods, then the cost of privacy can in theory be defined by the utility loss incurred when that data is removed. This framing has motivated interest in using data valuation for pricing mechanisms in emerging data marketplaces. For instance, Pei's survey on data pricing (Pei, 2022) shows how this area bridges economic principles (*e.g.*, fairness, arbitrage-freeness and revenue maximization) with data-centric concerns such as aggregability, reusability and privacy. However, building markets around this framework is nontrivial. As noted by Tian et al., 2022; Xia et al., 2023, valuing data in a way that both protects privacy and ensures fair compensation introduces substantial challenges. This perspective also assumes rational and informed decision-making on the part of individuals, which is often not the case. In particular, behavioral evidence contradicts this assumption: users frequently make inconsistent privacy choices, exhibit limited understanding of data use and accept long-term risks for short-term gains—what is often termed as the privacy paradox (Kokolakis, 2017). This utilitarian framing connects closely to the emerging economy of privacy, in which questions of who benefits from personal data, how value is distributed and whether individuals recognize their data's worth remain largely unresolved (Acquisti, 2023). Regulatory efforts like the GDPR are partly a response to these imbalances- by helping to restore individual agency over data. At the same time, regulation reshapes the privacy economy by introducing constraints and incentives that influence how data is collected, valued and exchanged. While these developments do not resolve the underlying issue of fair valuation, they foreground the need for technical frameworks that can formalize privacy as a quantifiable and enforceable property—a goal that has been pursued most directly by the risk-based perspective in computer science and security research.

Defining Privacy Risks through Indistinguishability

We next review technical privacy definitions that aim to formalize the inherent tension between privacy and utility. A fundamental conjecture in the privacy science literature asserts that data cannot be fully anonymized while maintaining usefulness (Dwork and Roth, 2014a), prompting researchers to develop definitions where such tradeoffs are explicit. Many of these definitions rely on the concept of indistinguishability. For instance, definitions such as k-anonymity (Sweeney, 2002) ensure that datasets contain groups of at least k indistinguishable records, alongside its refinements like l-diversity (Machanavajjhala et al., 2007a) and t-closeness (N. Li et al., 2007). Differential Privacy (DP), another widely recognized definition, characterizes privacy as a property of the algorithm used, for example, by introducing controlled noise into aggregate calculations like averages (Dwork, 2006). DP ensures that the outputs of an algorithm do not reveal whether any specific individual's data was included or excluded, by bounding the likelihood ratio of observing any output given two neighboring datasets. The research literature reflects an ongoing adversarial dynamic, akin to a "blue team" building privacy mechanisms and a "red team" continuously attempting to challenge them. For instance, re-identification attacks have been demonstrated on released datasets, highlighting real-world vulnerabilities (*e.g.*, De Montjoye et al., 2015; Heffetz and Ligett, 2014; Machanavajjhala et al., 2007b). Other privacy attacks include membership inference—identifying if an individual's data is present in a dataset, the minimal leakage corresponding exactly to one bit of information—and attribute inference, which involves uncovering specific personal attributes. For example, inferring that someone appears in a dataset from an abortion clinic could reveal that they underwent the procedure. In jurisdictions where abortion is restricted or stigmatized, this could expose them to legal, social, or economic harm—even without revealing other personal details. Estimating membership inference risk through practical attacks such as (Carlini, Chien, et al., 2022b; Shokri et al., 2017b; Ye et al., 2022b) serves as a lower bound for assessing privacy risks, as shown in recent auditing frameworks (Steinke et al., 2023). DP offers a formal upper bound on the membership inference risk, but its practical effectiveness depends on implementation details such as the choice of mechanism and privacy budget (ε). Despite these challenges, it remains the gold standard in terms of privacy model within the privacy research community (Dwork et al., 2017).

Although DP has primarily been developed and utilized in the literature as a defensive privacy mechanism, it has also been introduced as a solution concept within social choice theory, modeling privacy as a cost associated with truthfully revealing personal information (McSherry and Talwar, 2007). Within this game-theoretic context, DP helps mitigate individuals' incentives to withhold participation or to misrepresent their data, suggesting its broader applicability to mechanisms aimed at achieving more balanced privacy-utility tradeoffs (Pai and Roth, 2013). To the best of my knowledge, this makes DP the only technical framework that directly and simultaneously models the privacy-utility tradeoff from both the risk-based and utilitarian perspectives. Indeed from the risk-based side, DP provides formal, worst-case guarantees against a broad class of inference attacks while from the utilitarian side, DP can be interpreted through a game-theoretic lens as introducing a measurable cost for truthful data contribution—thereby aligning individual incentives with collective outcomes. This dual role is grounded in the core principle of indistinguishability, which not only supports formal anonymization, but also reduces the incentive to withhold participation by ensuring that any one individual's data has a limited but quantifiable influence on the output. As explained in book on The Algorithmic Foundations of Differential Privacy Dwork and Roth, 2014a, this property enables DP to serve as both a protective mechanism and a tool for aligning privacy and utility within mechanism design.

Research Questions

Several fundamental research questions remain under-explored in the literature. As previously mentioned, data valuation can be used by organizations to guide data removal. Intuitively, removing specific data points—particularly those with low marginal utility—may seem like a straightforward way to reduce privacy risks. This strategy is appealing because it offers clear visibility into which data is retained and which is discarded. However, it is not clear if this intuition holds when examined through the lens of risk-based privacy frameworks. In addition, it should be compared to other commonly used risk mitigation techniques such as data anonymization (*e.g.*, *k*-anonymity Sweeney, 2002). These issues are especially relevant in organizational settings, in which privacy risks must be evaluated in relation to how data is accessed, shared and used—and in which the objective is not only to reduce exposure but also to preserve user trust and minimize the likelihood that individuals will withdraw consent.

This thesis also examines how the two technical perspectives—utilitarian and risk-based—interact, and whether they can be meaningfully aligned. A central question here is whether data valuation, which measures the contribution of individual data points to model performance, also reflects a form of dependency that could be exploited in privacy attacks—particularly membership inference. In other words, we aim at quantifying if higher utility at the level of individual data points implies greater privacy risk. This touches on the broader relationship between a model's generalization capacity and its potential for privacy leakage. Theoretical results (Kasiviswanathan et al., 2010) suggest that models can generalize well without depending heavily on any single data point but in practice, privacy-preserving training methods like differential privacy are still rarely adopted by organizations (Munilla Garrido et al., 2023). Moreover, the current literature does not offer a definitive answer as to whether data points deemed highly valuable from a utility perspective are also those most at risk of privacy exposure. This relationship may further depend on the nature of the data point—whether it is an inlier that reinforces dominant patterns in the data, or an outlier that stands out due to unique characteristics, potentially making it more vulnerable to privacy attacks despite limited utility.

Finally, the extent to which privacy considerations influence data valuation remains an open question. This issue connects to the stability of data value estimates—how consistent these attributions are across different training samples or modeling assumptions. Ghorbani, Kim, and Zou (Ghorbani et al., 2020) have proposed a distributional framework that stabilizes Shapley-based valuations by averaging over multiple resampled training sets. However, they also point out a critical limitation: Shapley values are inherently non-private, as the value assigned to any individual data point depends on its marginal contribution across all possible coalitions, effectively entangling it with the rest of the dataset. Similarly, Wang and collaborators (J. T. Wang and Jia, 2023) have designed semi-value alternatives—such as the Data Banzhaf method—that trade off some fairness properties for improved robustness and lower sensitivity to sampling variability. While DP has been shown to reduce sensitivity in model outputs and can, in principle, stabilize data valuation, most current approaches assume a global setting— which corresponds to full access to the entire dataset. However in contexts in which individuals and organizations make decisions based on data value, this assumption becomes problematic.

Roth and collaborators (Ghosh and Roth, 2011) have shown that in data pricing mechanisms, revealing too much about how prices are computed can itself leak sensitive information, since pricing functions often reflect statistical dependencies within the data. Although their focus is on pricing rather than valuation per se, the insight generalizes: information about how much a data point is "worth" can become a proxy for what it reveals. This introduces a paradox in which the ability to estimate individual data value depends on prior large-scale data collection under conditions of broad access and consent. Yet once such valuations are available, they may shift user behavior, undermine consent or create new privacy risks. Therefore, organizations must consider that the capacity to quantify data value is not just a technical achievement—it reflects a structural asymmetry in who holds the analytical power and whose data is being evaluated, potentially revealing sensitive information in the process.

Contributions

In the following, we briefly outline how each chapter addresses the research questions presented above.

First, in Chapter 1, we examine three prevalent format-preserving data release strategies: data removal, data anonymization and data synthesis, analyzing their effects on both data utility and the risk of re-identification. However, maintaining format consistency imposes inherent limitations on privacy, potentially leaving residual risks that are still considered personal information under privacy regulations. Organizations frequently require data to be shared internally without altering its structure—not only to support analytics but also to meet operational needs, such as using realistic test data. At the same time, obligations tied to traceability, auditability and explainability can come into tension with privacy-preserving strategies like data synthesis, which are often irreversible. For example, in Canada's financial sector, institutions regulated by the Office of the Superintendent of Financial Institutions (OSFI) must retain linkable records for fraud investigations and internal audits. Similar expectations under standards like the

Payment Card Industry Data Security Standard (PCI DSS) require data persistence for ongoing compliance and monitoring. These operational and regulatory demands constrain how much privacy can realistically be achieved within format-preserving data release strategies.

In this context, reducing re-identification risk remains a meaningful and actionable goal—even when formal guarantees such as differential privacy cannot be enforced. In principle, any strategy that reduces the identifiability or influence of individual data points is aligned with the foundational idea behind differential privacy: limiting the extent to which any one individual affects the output. From a utilitarian perspective, reducing risk also lowers the disincentive for individuals to contribute their data, especially in settings in which participation is voluntary or revocable. Concretely, mitigating re-identification risk helps prevent micro privacy breaches. In this spirit, we develop a new evaluation framework that quantifies the privacy and utility trade-off of format-preserving data release strategies.

To realize this, we have adopted key concepts from modern privacy regulations—namely, singling out, linkability and inference—to quantify the risk of re-identification through a composite re-identification score. This metric reflects the difficulty of finding a specific individual and inferring sensitive information about that individual. To assess the privacy-utility tradeoff, we aggregate re-identification scores and utility metrics at the dataset level. Moreover, we leverage cooperative game theory for two separate purposes: first, to compare strategic data removal guided by Data Shapley values (Ghorbani and Zou, 2019) against random removal strategies and second, to identify specific attributes (*e.g.*, age and income) that significantly enhance the strategic capability of an attacker attempting to re-identify an individual. Although primarily theoretical, we have validated our framework on three publicly available datasets.

This chapter also includes an addendum to the original paper that focuses on the individual-level tradeoff between privacy and utility. Specifically, we investigate whether reducing a data point's re-identification risk—thereby lowering disincentive—comes at the cost of removing points that are also highly valuable from a model performance standpoint, as measured by their Shapley value. This comparison helps clarify how valuation and anonymization strategies can be jointly analyzed to inform privacy-aware data governance.

In Chapter 2, we shift our focus from format-preserving data release to a simpler scenario, which is the *k*-nearest neighbor (k-NN) classifier. Our goal is to better understand data attribution in such context — how individual data points influence model outputs. Consider an organization aiming to publish an API for classification tasks, such as sentiment analysis on movie reviews. In this scenario, an organization may pursue a dual objective: data valuation, to assess each data point's individual contribution to utility, and privacy auditing, to evaluate membership inference risks — specifically, whether querying the classifier reveals an individual's participation in the dataset. We initially hypothesized that high-value data points are not necessarily those most at risk for privacy breaches, as this would imply a uniform risk-utility ratio across points, which seemed unlikely for *k*-NNs. For instance, a unique outlier point with distinctive characteristics might be misclassified yet still pose significant membership inference risk. To address this, we introduce WaKA (for Wasserstein k-NN Attribution), a method grounded in the same principles as the Likelihood Ratio Membership Inference Attack (LiRA) by Carlini, Chien, et al., quantifying the change in the loss distribution attributed to individual points. We validated our methodology across six datasets, covering tasks involving data addition, data removal, and adversarial security games. We also replicated experiments originally presented in the Onion paper (Carlini, Jagielski, et al., 2022), highlighting that data removal can reduce the membership inference risk but does not completely eliminate it.

In Chapter 3, we continue our exploration of the privacy-utility tradeoff from a utilitarian perspective, aiming to integrate privacy considerations directly into data valuation. We start by framing the problem as a Stackelberg game between a data owner agent and a data consumer agent: in which the former strategically controls information disclosure using DP, while the latter seeks to extract enough utility for a predictive task under budget constraints. Traditional methods like Data Shapley (Ghorbani and Zou, 2019) assume that individuals who contribute data are cooperative and accept fairness axioms inherent to the Shapley value. Such assumptions, however, can lead organizations to undervalue substantial portions of the dataset, implicitly suggesting these data points could have been not collected without significant utility loss—an insight ironically obtainable only through extensive data collection. While alternative game-theoretic approaches, such as the Core (Yan and Procaccia, 2021), have studied coalition robustness, we explicitly model individual privacy costs and their impact on valuation. We first analyze the limitations of full information disclosure—in which pricing and access are fixed upfront—and then propose a partial disclosure mechanism based on iterative, noisy releases. By leveraging DP, the data owner can modulate individual leakage and influence the consumer's data selection. Empirical validation is conducted on a review helpfulness prediction task using the Yelp dataset, demonstrating how valuation under privacy constraints induces an acquisition cost and reshapes incentives toward greater inclusiveness.

While the central theme of this thesis is the quantification of the privacy-utility tradeoff through data valuation, the work deliberately integrates perspectives from Computer and Data Science (*e.g.*, privacy auditing and mechanisms, machine learning), Information Systems (*e.g.*, data governance, compliance) as well as Decision Science (*e.g.*, game-theoretic modeling of incentives).

To summarize the thesis makes three main contributions, each corresponding to a different angle of the privacy-utility problem.

1. It introduces a methodological framework for evaluating format-preserving data release strategies, combining a composite re-identification risk metric with cooperative game theory to assess both global and individual-level impacts. This includes a comparative analysis of strategic data removal via Shapley-based attribution versus traditional anonymization approaches such as k-anonymity, highlighting their respective implications for minimizing identifiable information

while preserving utility.

2. It proposes *WaKA* (Wasserstein k-NN Attribution), a novel and computationally efficient method for quantifying the influence of individual data points in *k*-nearest neighbor models. WaKA explicitly captures the duality between data valuation and membership privacy risk, enabling the same attribution method to identify both high-utility and high-risk data points. This bridges two technical lenses—utilitarian (value-driven) and risk-based (vulnerability-driven)—within a unified empirical framework.

3. It presents a game-theoretic model of data valuation under privacy constraints, in which information is partially disclosed using a differential privacy mechanism. By framing data sharing as a Stackelberg game between a data owner and a data consumer, the model captures how noisy release affects data Shapley values. This work offers one of the first empirical attempts to measure the implicit privacy cost of data valuation.

Chapter 1

Measuring Privacy/Utility Tradeoffs of Format-Preserving Strategies for Data Release

Abstract [1]

In this paper, we introduce a novel approach to evaluate the risk of re-identification of individuals associated with format-preserving data release strategies, focusing on three strategies: data minimization (*i.e.*, through data removal using random sampling and data Shapley values), data anonymization (*i.e.*, through k-anonymity), and data synthesis (*i.e.*, through CTGAN and TVAE generative models). More precisely, our approach consists in simulating a security game in which (1) an attacker performs singling-out attacks as outlined in data protection regulations and (2) an evaluator scores attacks based on the linkability of records and the information gain obtained by the attacker. In addition, we further enhance our approach by simulating attacks as a cooperative game, in which the value of the attackers' information resources is determined using the Shapley value

[1] This essay was published in the Journal of Business Analytics (Taylor & Francis, Volume 8, Issue 3, 2025). First author: Patrick Mesana (HEC Montréal); co-authors: Gregory Vial (HEC Montréal), Pascal Jutras (Purdue University), Gilles Caporossi (HEC Montréal), Julien Crowe (National Bank of Canada), and Sebastien Gambs (UQAM). An addendum by Patrick Mesana has been included in this thesis.

borrowed from game theory. Re-identification Shapley value is proposed as a method to measure the level of re-identification potential of each feature in a dataset when combined with other features. We demonstrate the effectiveness of our approach using three datasets commonly used in the privacy literature. Overall, our work contributes to a better understanding of the inherent trade-offs that exist between data privacy and data utility in organizations.

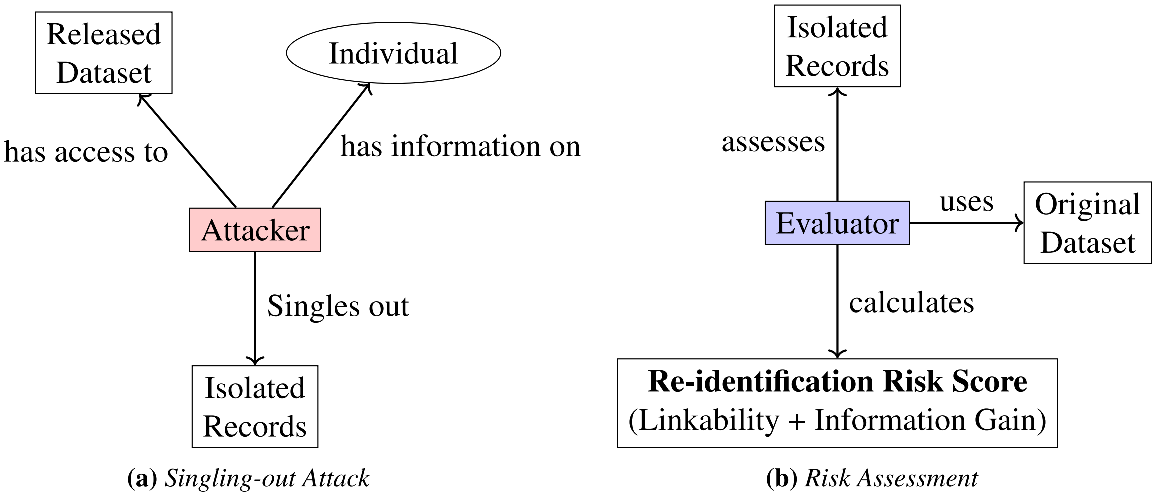
1.1 Introduction

According to the General Data Protection Regulation (GDPR) and other modern data protection laws, personal information has a broad definition that includes any data that can identify an individual, either directly or indirectly, through the linking of multiple pieces of information. More precisely, the term 'personal data' is defined in Art. 4(1) of the GDPR as "any information which are related to an identified or identifiable natural person" (Parliament and of the European Union, 2016). Entities serving as data custodians carry the obligation to protect their customers' privacy in the eventuality that personal data is released (*e.g.*, when it is shared with partners or with the public via an open data initiative, or through unlawful access to personal data - internal or external).

In the context of privacy-preserving data release, Privacy-Preserving Data Publishing (PPDP) is a term used in the literature (Clifton and Tassa, 2013; Fung et al., 2010; Rashid and Yasin, 2015) to refer to methods that enable the release of data while protecting the privacy of individuals. It is important to clarify that PPDP does not necessarily preserve the exact format of the original data. For instance, generalizing an attribute into an interval rather than a particular value, using dimensionality reduction, or building a different data structure from the original data are all techniques within PPDP that may alter the format to some extent. Within the broader scope of PPDP, format-preserving strategies represent a subset of strategies specifically focused on maintaining the original structure of the dataset. This means that while removing attributes with direct identifiers is acceptable, other kinds of attributes that are not considered direct identifiers are released. This ensures that, from an external observer's perspective, a sample of a released dataset appears similar to the original dataset, aside from potential changes in the data distribution. Such format-preserving strategies can be beneficial in scenarios in which maintaining the original format of a dataset is essential for usability purposes or to guarantee compatibility and compliance with existing systems and data workflows (Garfinkel, 2015). However, these strategies may also present challenges, including privacy vulnerabilities if the information contained is not adequately protected, and reduced value if the modifications applied to a dataset negatively affect data utility.

In this work, we focus on three distinct format-preserving strategies to ensure data privacy that we define hereafter: data minimization via data removal, data anonymization by generalization of attributes, and data synthesis using generative machine learning models. These strategies may also include the removal of direct identifiers, such as names or social security numbers, that do not bring any analytical value. Data anonymization transforms personal data to alter sensitive information while trying to preserve its utility. Techniques such as generalization-based transformations reduce the uniqueness of individual records. Uniqueness refers to the distinct characteristics or attributes of individual records that can uniquely identify an individual, making them stand out within a dataset. Among the most well-known models in this area of research is k-anonymity (Sweeney, 2002). Lastly, data synthesis involves creating artificial records using generative models trained on real data. This approach aims to generate data that share statistical similarities with the original dataset without revealing any personal information (Gootjes-Dreesbach et al., 2020; N. Park et al., 2018; Wan et al., 2017; Xu et al., 2019b).

Organizations are generally interested in anonymous data because most regulations no longer consider them as personal or sensitive information. However, in practice, it is widely acknowledged that all three strategies inevitably carry a risk of re-identification embodying the idea that there is an inherent trade-off between data privacy and data utility for organizations (Vial et al., 2024). Consistent with this idea, current regulations acknowledge the ability for organizations to use approaches that significantly reduce the risk of re-identification. In addition to legal requirements, an organization may have other motives for mitigating this risk. These include ethical standards, a commitment to preserving customer trust or a desire to uphold strong data governance practices while still being able to gain significant business value from personal data. These considerations are in line with the proactive principle of "privacy by design" (Cavoukian, 2009).



(a) *Singling-out Attack*

(b) *Risk Assessment*

Figure 1.1: *Overview of the security game between the attacker and the evaluator. (a)* ***Singling-out Attack****: The attacker has access to the released dataset and possesses some information about an individual. Using these resources, the attacker singles out specific records from the released dataset that they believe correspond to the individual, resulting in isolated records. (b)* ***Risk Assessment****: The evaluator receives the isolated records from the attacker and utilizes the original dataset to assess the risk of re-identification. By analyzing the linkability of the isolated records to the original data and calculating the potential information gain for the attacker, the evaluator computes a re-identification risk score.*

To address these issues, we propose a risk assessment approach aimed at evaluating the potential threat of an attacker attempting to single out an individual's records within a released dataset. For example, an attacker could be an external entity gaining unauthorized access to a compromised dataset, or an internal employee who might unknowingly or intentionally search someone, perhaps unaware of the associated privacy implications. The notion of singling-out (*Opinion 05/2014 on Anonymisation Techniques*, 2014) refers to the potential to isolate certain or all records that correspond to a single individual within a given dataset. It is presumed to form the foundation of the risk associated with re-identification and it has been discussed in the context of the GDPR, with its mathematical implications having been studied in prior work (A. Cohen and Nissim, 2020a). Our risk assessment approach consists of framing the singling-out problem as a security game between an attacker and an evaluator. This framing allows organizations to simulate adversarial scenarios and understand the risk of re-identification in various contexts. By using this approach, organizations can assess the likelihood and impact of singling-out attacks. For instance, employees may have access to a secured internal dataset, and the organization can monitor and record their queries on that dataset to evaluate the risk of singling-out attacks. By viewing these queries as potential attacks in a security game, the evaluator can analyze and score the queries to determine the risk of re-identification. (A. Cohen and Nissim, 2020b; Francis et al., 2019). In our approach, the attacker leverages information resources at their disposal along with the released dataset to isolate a specific group of records (Figure 1.1a). Subsequently, an evaluator scores the isolated records, using the original dataset, based on two factors: (1) the degree to which singled-out records can be linked, which we define as linkability, and (2) information gain, which refers to the amount of information an attacker could gain through the attack (Figure 1.1b). Unlike existing methodologies, we believe that these factors can be perceived as components within an attacker's valuation function, which concurrently serves as a re-identification risk score for an individual.

We estimate each score by simulating multiple attack scenarios, which can be modelled as independent attacks or as a collaborative game with multiple attackers. The latter configuration enables us to employ the Shapley value to assess the value of an information resource to an attacker. More precisely, the Shapley value is a concept from cooperative game theory developed by Shapley (1953) that has found applications in numerous fields, including economics, political science and computer science (Roth, 1988), among others. In a cooperative game setup, players form a coalition or group and work together to achieve a common goal. Here, the objective of the adversarial player(s) — referred to as attackers — is to re-identify and retrieve as much information as possible on individuals. For example, different attackers might have access to different pieces of information about the same individual, such as their job or their hobby. By evaluating the contribution of each piece of information to the overall success of an attack, Shapley values help in understanding the relative importance and impact of various factors in the risk of re-identification. This addresses the challenge that organizations often face in evaluating the re-identification potential of individual attributes that are not direct identifiers, making it difficult to assign appropriate security levels to them.

In our experiments, we primarily explore tabular data given its widespread use as a format for confidential information that underpins many of an organization's key decisions. We conduct our experiments on three well-known datasets: the Adult Income Census (AIC) dataset, the Bank Marketing (BM) dataset and the Credit Card Default (CCD) dataset (see Table 1.1 for more details). We assess three main strategies for data releases: the first is data minimization, which involves removing data through methods such as data valuation using Shapley values (Ghorbani and Zou, 2019; Jia, Dao, Wang, Hubis, Gurel, et al., 2019) and random sampling; the second involves achieving *k*-anonymity through the Mondrian algorithm (LeFevre, DeWitt, and Ramakrishnan, 2006), which generalizes all attributes in the dataset; and the third is synthesizing data using deep learning models, Tabular Variational Autoencoder (TVAE) and Conditional Tabular Generative Adversarial Network (CTGAN) introduced by Xu et al., 2019a. Our findings show that while data anonymization and data synthesis are not directly comparable, data synthesis can correspond to certain levels of *k*-anonymity in terms of the privacy-utility trade-off. However, when considering data utility, data synthesis tends to be more variable. For example, in the BM dataset, the predictive performance of synthetic data differs noticeably between CTGAN and TVAE. On the other hand, data minimization strategies are generally less effective at reducing the risk of re-identification compared to data anonymization or data synthesis. Nonetheless, data minimization offers stronger guarantees for removed data—specifically, individuals whose records are no longer present in the dataset are assigned zero risk, meaning there is absolutely no chance of re-identification for those individuals. In addition, our study provides insights into those features of our sampled datasets that are particularly valuable for attackers.

To summarize, our primary contribution is made to the literature on data privacy.

Specifically, our proposed measure for re-identification risk scores defines a privacy strength metric associated with data release strategies, producing results that are inherently dataset-specific. We tested this novel metric alongside data utility metrics (defined in the utility evaluation experiments section 1.5.1) to better understand the privacy-utility trade-offs. Our risk-based approach also has implications for practitioners as it does not focus solely on worst-case scenarios, such as differential privacy Dwork et al., 2016 or membership inference privacy attacks Dwork et al., 2017; Nasr et al., 2023. Instead, it provides insights on all individuals associated with a data release strategy. Moreover, we introduce the Re-identification Shapley Value as a robust mathematical framework that also offers an intuitive perspective on the re-identification risk synergies between dataset features, highlighting how they collectively contribute to the risk. Finally, our approach is compatible with economic analyses that can guide organizations in selecting the right data release strategy.

The outline of the paper is as follows. We begin with an overview of the literature on privacy threats and associated risks in Section 1.2 before in Section 1.3, outlining the elements of our risk-based approach for handling tabular data, discussing its key assumptions and measures. Following this, in Section 1.4, we explain how game theory and the concept of Shapley value can be harnessed to investigate re-identification. We then move to Section 1.5, in which we showcase the practical application of our approach using three illustrative datasets. Subsequently, in the discussion section (Section 1.6), we address limitations and explore potential applications of our findings. Finally, we summarize our key contributions and provide concluding remarks in Section 3.6.

1.2 Background and Related Work

In 2006, when a major streaming company published a so-called "anonymized dataset" for a public online contest with the goal of improving their recommendation engine, they believed they had sufficiently reduced the risk of re-identification, given the nature of the data being made public. Narayanan and Shmatikov (2008) managed to re-identify a small subset of the individuals from that dataset who were also included in the IMDB open database. It has been reported that the subsequent class action lawsuit caused the company to settle for 9 million dollars in 2011 (CNET, 2012). Other cases of re-identification attacks on published datasets have been documented since (Henriksen-Bulmer and Jeary, 2016), although details on their economic impacts for organizations often remain scarce. Nevertheless, there is a general consensus within the privacy community that the residual risk of re-identification associated with the use of data anonymization techniques, regardless of the type of technique being used, depends on the probability of success of re-identification attacks.

Measuring uniqueness served as a method to evaluate the risk of re-identification (Skinner and Holmes, 1998). For instance, de Montjoye et al. (2015) demonstrate that four spatiotemporal points are sufficient to uniquely re-identify 90% of individuals in a dataset containing three months of credit card records for 1.1 million people. Dankar et al. (2012) build on the work of Skinner and Elliot (2002) to evaluate the risk of re-identification in six datasets by estimating the uniqueness of individuals. Their method for estimating uniqueness is based on the assumption that the population originates from a larger, super-population, thus turning it into a question of parameter estimation. To minimize this risk, several approaches have been proposed, including generalization-based transformations such as *k*-anonymity, as well as synthetic data.

Data Anonymization. *k*-anonymity (Sweeney, 2002) is a property of a dataset that ensures each record is indistinguishable from at least $k-1$ other records with respect to certain attributes, known as quasi-identifiers. Quasi-identifiers are attributes such as age, gender, and ZIP code, which, when combined, can uniquely identify individuals. To achieve *k*-anonymity, methods such as the Mondrian algorithm (LeFevre, DeWitt, and Ramakrishnan, 2006) are often employed. This algorithm recursively partitions the dataset based on attribute values to form groups of at least *k* records. By generalizing the quasi-identifiers within each partition, the algorithm guarantees that no individual can be distinguished from a group of at least *k* others, though the group size may exceed *k*. This property reduces the probability of an attacker successfully associating a record with the correct individual to at most $1/k$.

Despite its advantages, it has been demonstrated that this method fails to encompass the entirety of privacy risks, in particular in the situation in which the risk associated with attribute disclosure is significant. More precisely, attribute disclosure can be viewed as a form of reconstruction attack in which the attacker's objective is to reconstruct sensitive information about individuals without necessarily re-identifying them. In addition, it is important to distinguish between the information that we aim at protecting versus the generic knowledge that can be drawn from the dataset (T. Li and Li, 2009). While our objective is to establish a definition of privacy safeguarding individuals, it is equally important to ensure that this definition allows for insightful data analysis (Dwork and Roth, 2014b). For example, consider a medical dataset in which each patient's health record is anonymized to protect their identity. The risk of re-identification in such cases primarily involves the potential harm to specific individuals when an attacker gains access to personal information about them. This distinction underscores that privacy risks extend beyond the mere identification of individual records and encompass broader implications for data utility and societal impact.

Data Synthesis. As a technique that is gaining attention in the industry, the use of synthetic data involves the generation of new data by a model that is trained on existing data. While it might seem counter-intuitive to attempt re-identification attacks on synthetic data, given the perception that its "fake" nature implies a reduced re-identification risk, the reality is more complex. It is unclear whether anonymization is worse than synthetization in this context of re-identification. However, like any other semantic privacy-preserving technique, it is crucial to acknowledge that there remains a possibility for leakage of personal information, for instance through reconstruction and membership inference attacks (Dwork et al., 2017). A membership inference attack can be understood as a process by which an adversary guesses that an individual was a member of a training dataset based on an observed output. There exists membership attacks specifically designed for data synthesis, such as distance-based attacks presented by Hilprecht et al. (2019). While synthetic data, therefore, holds promises in theory to preserve data privacy, recent findings suggest that it may not be a panacea.

Data protection laws that have been adopted over the past few years seek to regulate the use of personal data by organizations. Among those, the GDPR is considered the most comprehensive regulatory framework, and it has effectively influenced many aspects related to the governance, management and use of personal data by organizations, as well as data protection laws in other jurisdictions (*e.g.*, Canada). Within this context, the GDPR mentions anonymization as a valid technique for irreversibly transforming data, albeit without explicitly incorporating data synthesis as one of the possible ways to achieve this objective. One potential reason for this shortcoming is that data synthesis remained largely within the realm of academic research when the GDPR was adopted in 2016. Nevertheless, the GDPR highlights three main types of risk factors associated with data releases that should be mitigated, regardless of the nature of the data release strategy itself: "singling-out", as formally defined by A. Cohen and Nissim (2020b), involves isolating an individual or a specific record within a dataset; "linkability" refers to the ability to link records from the same individual across different datasets; and "inference" refers to the attribute disclosure risk discussed above. Giomi et al. (2022) have developed a unified framework for measuring these three factors separately to quantify the degree of risk associated with synthetic data. They argue that in synthetic data, linkability arises from the statistical similarities between the synthetic data and the original data. Stadler and colleagues also have portrayed membership attacks as a linkability risk factor to compare data synthesis against data anonymization (Stadler et al., 2022). Uniqueness could be seen as a way to measure the concept of singling-out. However, uniqueness measures alone often showcase bias, as the risk evaluator is limited by the available data distribution and the calibration of these measures. For instance, El Emam et al., 2013 highlights how variations in estimators for population uniqueness can impact the accuracy of re-identification risk assessments in the context of clinical data. Additionally, attackers generally have only partial access to background knowledge, indicating that uniqueness is not the sole determining factor in singling-out individuals in datasets.

Utility-Driven Data Minimization . Data removal is often an overlooked strategy to reduce re-identification risk, potentially because data scientists hypothesize that more data generally leads to better outcomes. However, this principle is critical in data protection laws. According to the GDPR's Article 5(1)(c), data minimization mandates that data controllers limit the processing of personal data to what is directly relevant and necessary for a specified purpose (*i.e.*, the finality of data). This ensures that only essential personal data is processed. In practice, data minimization can be achieved through techniques like data pruning and feature selection.

Effectively implementing data minimization requires identifying which pieces of information in a dataset are valuable enough to retain, while removing less valuable information to meet privacy goals without sacrificing utility. Cooperative game theory provides a powerful framework for such decisions, using the Shapley value to evaluate the contribution of players within a coalition. In this context, players correspond to distinct pieces of information, which could either represent individual data points or specific features. The Shapley value is grounded in four axioms that ensure fairness and consistency in the valuation of players:

1. **Efficiency**: The total value generated by a coalition is distributed among all players.
2. **Symmetry**: Players who contribute equally to all coalitions receive the same value.
3. **Null Element**: A player that does not contribute to any coalition receives no value.
4. **Additivity**: The value assigned to players remains consistent when coalitions from separate games are combined.

These axioms make the Shapley value a robust tool for measuring the contribution of information to a predictive model. When considering features as players, their contributions can be analyzed at different levels: globally, by evaluating the overall importance of the feature across the dataset, or locally, by examining how specific feature values influence individual predictions. For example, interpretability frameworks such as SHAP (*SHapley Additive exPlanations*) Lundberg and Lee, 2017 leverage Shapley values to explain a machine learning model's predictions by attributing the contribution of each feature. Similarly, Shapley values can be applied to data points in a dataset, a process known as data valuation. Data Shapley values, as introduced by Ghorbani and Zou (2019), provide a systematic method to quantify the utility of individual data points. By identifying and removing data points that do not significantly enhance task performance, this approach aligns with the principle of data minimization while preserving data utility. Organizations can use Shapley values to implement scientifically grounded strategies for achieving a balanced trade-off between privacy and performance, ensuring that only the most relevant information is used. Other techniques also exist for utility-driven data minimization. For example, influence functions (Koh and Liang, 2017a) estimate the importance of training points. However, these methods can be more fragile (Ghorbani et al., 2019a), often relying on simplifying assumptions or approximations that may not generalize well across diverse datasets or models. Notwithstanding the robustness of Shapley value-based methods, their computational complexity can be a drawback to their usefulness. Exact computation scales exponentially with the number of features or data points, making them challenging to apply to large datasets. Sampling-based approximations can alleviate this burden but may sacrifice precision, requiring organizations to carefully balance computational costs and accuracy.

Overall, research on data privacy has greatly contributed to our understanding of the nature and the negative impacts of the risks associated with the re-identification of individuals as well as various techniques available to try and mitigate these risks. Notwithstanding, we observe that there is still a need to further develop approaches that can reconcile the demands of data protection regulations, the need for organizations to generate business value using data, and the rapid technological advances that allow attackers to perform re-identification attacks at low cost. In particular, we argue that decision-makers would benefit from an increased ability to assess the trade-offs between data privacy and data utility in the context of data releases. To help fill this important gap, the next sections detail the constituting elements of our approach in the context of tabular data. Our approach is based on the idea that as managers, decision-makers need to make decisions regarding data releases on a frequent basis. To help them make these decisions in an informed manner without remaining stuck in a state of “paralysis by analysis” that can hinder value creation, we draw from the scientific literature on data privacy and game theory to formally quantify the degree of risk associated with a data release strategy for a particular dataset.

1.3 Re-Identification Risk Evaluation for Tabular Data

The security game begins with attackers attempting to identify individuals within a dataset T that was released using a specific strategy, aiming to achieve the highest possible re-identification risk score (*RRS*) for each individual. We assume there is an original dataset D and that the type of attack is the same whether T was produced through data anonymization or data synthesis but that the background information an attacker has access to will change. In the literature, “background information” typically refers to the auxiliary data or knowledge that an attacker may have. In this paper, we extend this concept and call it “information resources” or “resources” in short, because when combined these can provide attackers with more re-identification power. For instance, an attacker who knows both the “age” and the “zip code” of an individual has more resources compared to an attacker with only the “age" information.

We now describe the type of singling-out attack implemented in our approach. The singling-out process involves isolating one record or a small group of records. The symbol α will be used interchangeably with the singling-out function that takes T and a resource r as inputs and returns a subset C of records in T, such that $C = \alpha(r, T)$. Obtaining a subset C instead of just a single record is consistent with scenarios in which T is a k-anonymized dataset in which each anonymized record has at least one duplicate. Due to the fact that the attacker looks at records in T that are similar to his resources r to re-identify one individual, we select the records minimizing the distance between r and any t contained within T.

$$\alpha(r,T) = \min_{t \in T} d(r,t) \tag{1.1}$$

More precisely, we use the unnormalized Gower distance (Gower, 1971) as our base method as this distance measure can handle a combination of continuous and categorical features that are characteristic of tabular data. We vectorize all resources to compute distances. A key advantage of the vectorization of resources is that it enables the use of efficient techniques, such as kd-trees or similar structures, to perform neighborhood searches. This allows us to significantly speed up the simulation of attack scenarios by quickly narrowing a set of records. In particular, instead of computing a complete similarity matrix between vectors r and t, we can use a subset of the matrix. We used a BallTree (Omohundro, 1989), mainly for its computational efficiency, to estimate the neighborhood of r in T.

Having introduced the singling-out function employed by the attacker, we now turn to the process of the evaluator. To simplify notations, we refer to an attack scenario s as an attacker α utilizing their resources r to single out a specific individual within T. The evaluator has the responsibility to score attacks, which serves also as a risk evaluation of any plausible attack scenario s. We assume that the evaluator is aware of the data transformation strategy employed to avoid any form of "privacy through obscurity". Records that have been singled out by the attacker constitute a risk only if they can be linked to a real individual. In our framework, the evaluator estimates the linkability score of an attack scenario, denoted L_s, of an attack scenario. Similar to other membership risk studies (Giomi et al., 2022; Hilprecht et al., 2019; Houssiau et al., 2022; Lu et al., 2019), our linkability measure uses the distance between a matched record and the original record of the individual. Our goal is to assess whether a given distance is likely to be observed in the original dataset. For example, a target point that is very close to the original point may appear suspicious, but this also depends on the region density, sometimes referred to as the uniqueness of the point.

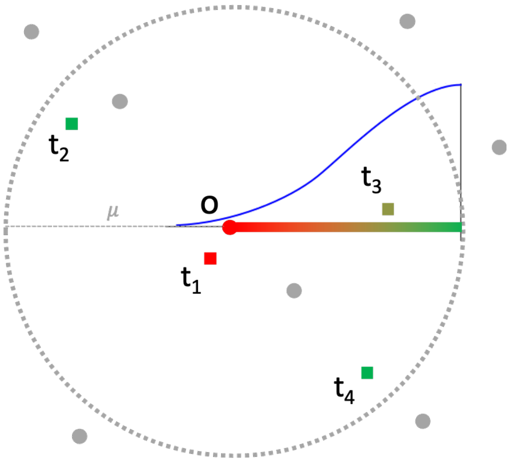


Figure 1.2: *The attacker singles out four specific data points, indicated by the squares t_1, t_2, t_3, and t_4. Nearby points from the original dataset, shown in grey, are used to establish the neighborhood of the targeted individual o and to calculate an average distance μ and a standard deviation σ. The color gradient illustrates the likelihood, based on a Gaussian distribution, that each point t_i can be "linked" to the original point o.*

We take inspiration from the offline version of the Likelihood Ratio Attack (LiRA) described by Carlini, Chien, et al., 2022a, in which the likelihood of the model is evaluated based on the assumption that the distribution of the loss function without the point of interest follows a Gaussian distribution. Similarly, in our context, we assume that an underlying model of the data exists, which generates points in the neighborhood of the original point. The loss function quantifies the error between the original point and a generated point, reflecting the deviation introduced by the generative process. The loss function can take the form of the distance between the two points. For instance, a conditional generative model for data synthesis creates points that resemble the original data, but with a likelihood influenced by the model's capacity and subject to a reconstruction error. The likelihood of the distance can serve as a proxy for estimating the likelihood of the model's error. Formally, let $d(o,n)$ denote the distance between the original record o and any neighboring record n in the original dataset D. Given a singled-out record t in the released dataset T, we aim to estimate the likelihood of $d(o,t)$. To do this, we model the distance distribution as a Gaussian distribution, parameterized by the average distance and the standard deviation of the distances within a neighborhood of the original record in the original dataset D. The neighborhood of o is defined by a parameter k, representing the number of nearest neighbors. For each original record o, we compute the mean

$$\mu = \frac{1}{k}\sum_{i=1}^{k} d(o,n_i)$$

and the standard deviation

$$\sigma = \sqrt{\frac{1}{k}\sum_{i=1}^{k}(d(o,n_i)-\mu)^2}$$

of the distances to its k nearest neighbors, which parameterize our Gaussian distribution $\mathcal{N}(\mu,\sigma^2)$, where n_i represents the i-th nearest neighbor of o. To compute the linkability score L_s for a given attack scenario, we calculate the likelihood of the distance $d(o,t)$ between the original record o and the singled-out record t. This likelihood is captured by the term $L(t)$, which measures how linkable t is to o. Specifically, $L(t)$ quantifies the extent to which the observed distance $d(o,t)$ lies within a "linkable" region, as determined by the Gaussian distribution $\mathcal{N}(\mu,\sigma^2)$. Since we only consider distances smaller than or equal to μ as indicative of linkability, $L(t)$ is set to 0 when $d(o,t) > \mu$. For distances within the linkable region $d(o,t) \leq \mu$, the likelihood is normalized by dividing by $(1-\mu)$ to account for the one-sided nature of the distribution. $L(t)$ is then computed as 1 minus this normalized likelihood.

The overall linkability score L_s is calculated as the average of $L(t)$ across all records $t \in C$:

$$L(t) = \begin{cases} 1 - \frac{\Pr[Z > d(o,t)]}{1-\mu}, & \text{if } d(o,t) \leq \mu \\ 0, & \text{if } d(o,t) > \mu \end{cases}$$

$$L_s = \frac{1}{|C|}\sum_{t \in C} L(t)$$

where $Z \sim \mathcal{N}(\mu,\sigma^2)$.

This formulation ensures that the linkability score reflects the probability of a specific record being a member of the original dataset based on its distance from the original record, while accounting for the expected distribution of distances. While we assume a Gaussian distribution to model the distances, we believe this assumption provides a reasonable approximation for the purposes of our analysis, as it effectively captures the spread of distances in most cases. Figure 1.2 illustrates the computation of the linkability score using the assumed Gaussian distribution model for the distances.

The evaluator also has to determine if an attack scenario leads to an information gain, denoted I_s. We claim that I_s depends on the reconstruction loss of each possible match t in C. Let o be the original record of an individual in D, $A' \subseteq A$ be the set of attributes to be reconstructed, and A the set of all attributes. We express the reconstruction loss, denoted $E(t)$, of a matched record t, as an unnormalized Gower distance between $o[A']$ and $t[A']$.

$$E(t) = \sum_{i \in A'} \zeta_i \cdot |o_i - t_i|. \tag{1.2}$$

We denote ζ_i as the information gain per distance unit on the i-th feature. By default, we consider all ζ_i to be equal. We can now express the average information gain I_s of scenario s as the following formula:

$$\begin{gathered} I_s = \frac{1}{|C|} \sum_{t \in C} \gamma_s \cdot \left(1 - \frac{E(t)}{\sum_{i \in A'} \zeta_i}\right), \\ \text{in which } \gamma_s = f\left(\frac{\sum_{i \in A'} \zeta_i}{\sum_{i \in A} \zeta_i}; \varepsilon_1\right). \end{gathered} \tag{1.3}$$

The function $f(x; \varepsilon_1)$ is defined as:

$$f(x; \varepsilon_1) = \frac{\log(1 + \varepsilon_1 \cdot x)}{\log(1 + \varepsilon_1)}.$$

When the reconstruction loss is 0, the information gain depends on the function f, which penalizes the ratio of the attributes to reconstruct based on the parameter ε_1. This allows for flexibility in adjusting the information gain: with a higher ε_1, the penalty is reduced, enabling more information gain even when fewer attributes need to be reconstructed. This ensures that the information gain can still be significant with a smaller set of attributes while maintaining proportionality to the sum of the normalized weights of the attributes. Following this rationale, when there are no attributes to reconstruct, the information gain has a value of 0. The logarithmic scaling function was chosen for ε_1 to transition from a linear relationship to a non-linear one, enabling more nuanced control over the impact of extreme scenarios (attackers reconstructing no features or all features) on the re-identification risk.

Finally, we can define the re-identification risk score *RRS* of an individual as follows.

Definition 1.3.1 (Re-identification risk score)**.** Let S be the set of plausible attack scenarios on an individual. Each scenario $s \in S$ has a linkability score, denoted by L_s, and an information gain, denoted by I_s, such that $0 \leq L_s \leq 1$ and $0 \leq I_s \leq 1$. The constant ε_2 allows the evaluator to give a minimum weight to linkability even when I_s is 0.

$$RRS = \frac{1}{|S|} \sum_{s \in S} L_s \cdot \max\{I_s, \varepsilon_2\} \tag{1.4}$$

The *RRS* of an individual measures how much we estimate the singled out records in T to be linked to the individual, moderated by how much information can be gained by the attacker. Equation 1.4 suggests that the information gain is taken into account only when a record is linkable. Furthermore, if the information gain for a particular scenario s is equal to 0 but the records are linkable, the associated risk is equal to ε_2. To score a release strategy, we can simply compute the expected re-identification risk score of all individuals present in dataset D.

1.4 Re-identification Shapley Value

Attack scenarios can be performed individually and independently to evaluate the *RRS* of each person. Nevertheless, we purport that modelling our problem as a re-identification cooperative game can provide additional insights. The entire simulation can be perceived as consisting of multiple attackers, each of them possessing their own resources and focusing on a specific group of individuals. Consider for instance a scenario in which multiple adversarial agents conduct attacks simultaneously, with the evaluator providing risk scores for each individual. The re-identification risk scores can be combined to calculate the score of a sample of individuals. More precisely, we repeat this process until the score converges. We create resources by using D to sample possible subsets of resources. As attackers' resources often overlap, this process is akin to a simulation that employs sampling with replacement.

We will now model the simulation as a cooperative game, in which attackers with varying resources collaborate to maximize their ability to re-identify individuals and gather as much information as possible. More precisely, α_1 and α_2 represent two attackers with resources r_i and r_j respectively, targeting the same individual. The combined attacker set $\alpha_{1,2}$, acts as a "super" attacker with combined resources (r_i, r_j). There are now three possible attack scenarios: s_1, s_2, and $s_{1,2}$. The scenarios s_1 and s_2 correspond to attacks using resources r_i and r_j, respectively, while $s_{1,2}$ involves attacking with the combined resources (r_i, r_j). We assume that the combined attacker set $\alpha_{1,2}$ is stronger than any individual attacker, such that $RRS_{s_{1,2}} \geq RRS_{s_1}$ and $RRS_{s_{1,2}} \geq RRS_{s_2}$. This assumption is consistent with the additive property of the value of information, suggesting that increased resources lead to higher re-identifying capabilities. However, this assumption may not universally hold true, as additional resources can sometimes negatively impact re-identification, especially when attackers encounter noisy or non-informative attributes. Nonetheless, in practice, this simplistic assumption appears to work well, primarily because released datasets tend to possess non-redundant features and preserve good utility.

To better understand the motives underpinning potential cooperation in the context of re-identification attacks, we now turn to the evaluation of the rewards associated with the success of an attack, as well as the question of how attackers should split the gains in such an instance. To answer this question, we borrow from the concept of Shapley value, which provides a way to allocate total gains or costs among players in cooperative games. In our context, we adapt this concept and refer to it as Re-identification Shapley Value (RSV):

Definition 1.4.1 (Re-identification Shapley Value)**.** Consider a re-identification game with a player set α_N of size n, let ϕ the valuation function of the re-identification of an individual and $S \subseteq \alpha_N \backslash \{i\}$ a subset of players that does not include player i. The re-identification Shapley value $\psi_i(phi)$ of player i is defined as follows:

$$\psi_i(\phi) = \frac{1}{n} \sum_S \left(\begin{array}{c} n-1 \\ |S| \end{array} \right)^{-1} \phi(S \cup \{i\}) - \phi(S). \tag{1.5}$$

If the valuation function is the *RRS* and if we assign specific resources to an attacker, the RSVs serve as an equitable measurement of how these resources are valuable for re-identification, from the point of view of the evaluator. One of the interesting properties of Shapley values is linearity. More precisely, the linearity axiom states that for any payoff function v that is a linear combination of two other payoff functions u and w, the Shapley values of v equal the corresponding linear combination of the Shapley values of u and w. Applying this axiom to RSVs, we can combine the Shapley values of attackers when they collaborate to re-identify more than one individual. If each attacker is responsible for one type of resource corresponding to a feature in the dataset T, the RSVs for all individuals in D can be interpreted as the most valuable resources or features to re-identify individuals in T.

RSVs represent an additional application of Shapley values, focusing on measuring the contribution of information in a dataset from an attacker's perspective. While data valuation assesses the importance of data points to model performance and explains attributes feature contributions to predictions, RSVs quantify how individual pieces of information contribute to the success of re-identification attacks. Similar to its use in explainability, the cooperative game framework ensures equitable attribution of contributions, emphasizing the synergies of attackers' resources. Without assuming collaboration, we cannot ensure a unique and fair method (following the Shapley axioms) of value distribution. This would require accounting for competing attacker dynamics, which is beyond the scope of this study. We primarily rely on RSVs to explain the *RRS* by quantifying the contribution of resources involved in re-identification attacks.

To approximate Shapley values using Monte Carlo methods (Maleki et al., 2013), we express the Shapley formula as an expectation. This approximation is necessary due to the exponential complexity associated with exact calculation of Shapley values when considering all subsets of features. The Shapley value for a feature i can thus be rewritten as:

$$\psi_i(\phi) = \mathbb{E}_{\pi \sim \Pi} [\phi(S_i^\pi \cup \{i\}) - \phi(S_i^\pi)], \tag{1.6}$$

in which Π represents the uniform distribution over all possible permutations of the features, and S_i^π denotes the set of features that precede feature i in the permutation π. If i is the first feature in the permutation, then $S_i^\pi = \emptyset$. To approximate this expectation, we randomly sample a set of M permutations $\pi_1, \pi_2, \ldots, \pi_M$ from Π. For each sampled permutation π_j, we compute the marginal contribution of feature i as $\phi(S_i^{\pi_j} \cup \{i\}) - \phi(S_i^{\pi_j})$. The Monte Carlo estimate of the Shapley value is then obtained by averaging these marginal contributions:

$$\hat{\psi}_i(\phi) = \frac{1}{M} \sum_{j=1}^{M} \left[\phi(S_i^{\pi_j} \cup \{i\}) - \phi(S_i^{\pi_j}) \right].$$

By using a sufficiently large number of permutations M, the Monte Carlo estimate $\hat{\psi}_i(\phi)$ converges to the true Shapley value $\psi_i(\phi)$.

1.5 Experiments

In this section, we introduce a series of experiments designed to evaluate the performance of our proposed re-identification risk assessment methodology. We applied our framework to three commonly used datasets in the privacy research community: the Adult Income Census (AIC) dataset, the Bank Marketing (BM) dataset and the Credit Card Default (CCD) dataset. All three datasets were split into train, validation and test sets with a 70%-15%-15% distribution. This split was performed for predictive utility evaluation and data Shapley valuation removal strategy, as detailed in the Appendix 1.10.

Each dataset is summarized in Table 1.1. They all include a variety of features, all of which contain sensitive information that could potentially be used for re-identification. We conducted experiments on these datasets to simulate different attack scenarios and evaluate the re-identification risk scores. These datasets, with their mix of categorical and continuous features, provide a robust foundation to test our methodology across various data types.

Table 1.1: *Description of the datasets used in the experiments*

Dataset	**Type**	**Categorical Attributes**	**Numerical Attributes**	**Training Size (70%)**	**Validation Size (15%)**	**Test Size (15%)**	**Reference**
AIC	Tabular	9 attributes	4 attributes	10,977	2,353	2,352	Kohavi, 1996
BM	Tabular	8 attributes	4 attributes	7,404	1,587	1,587	Moro et al., 2014
CCD	Tabular	14 attributes	7 attributes	700	150	150	Yeh and Lien, 2009

More precisely, the Adult Income Census dataset includes sensitive attributes such as age, race, sex, marital status, and native country, making it an ideal candidate for assessing re-identification risks. The target variable in this dataset is the income level, which is categorized as either greater than or less than 50K per year. The Bank Marketing dataset contains information from a Portuguese banking institution's marketing campaign, featuring attributes such as age, job, marital status and education. The target variable in this dataset is whether the client subscribed to a term deposit. Finally, the Credit Card Default dataset consists of records from a Taiwanese credit card company, including features such as age, gender, education and credit history. The target variable in this dataset is whether a client will default on their credit card payment.

The following three sections are structured as follows: in the **Utility Evaluation of Data Release Strategies** section, we introduce the data release strategies we used and discuss how we evaluated their data utility. Specifically, we assessed the effect of increasing k in k-anonymity, tested the impact of removing 25%, 50%, and 75% of the training data instances, and used the Synthetic Data Vault (Patki et al., 2016) synthetic data models, which includes CTGAN and TVAE generative models. Afterwards, we delve into the **Privacy Risk Scores** section, in which we focus on the individual privacy risk scores, analyzing how these scores vary across different datasets and data release strategies. We also examine the trade-offs between privacy and utility, providing insights into how some strategies can be removed using the Pareto frontier. Finally, the **Re-identification Shapley Value Analysis** section focuses on the features with the highest re-identification potential, using the concept of Re-identification Shapley Value to identify features contributing most significantly to re-identification risks.

1.5.1 Utility Evaluation of Data Release Strategies

In this section, we explain how we evaluated the utility of different data release strategies, focusing on data minimization, data anonymization, and data synthesis. Data utility has been explored in the literature through two complementary lenses: one focusing on the characteristics of the data itself and the other emphasizing its role in supporting decision-making, as explained by Even and Shankaranarayanan, 2007. In our experiments, we employed two metrics to capture these complementary views. The first metric is the Wasserstein distance, which measures the similarity between data distributions and is particularly effective at capturing distribution shifts due to its ability to account for the underlying geometry of the data (compared with distance metrics such as the KL divergence). To provide a more interpretable measure, we normalized this distance using a random baseline. Specifically, we computed the normalized Wasserstein distance as:

$$\widetilde{W} = \frac{W_{\text{rand}} - W}{W_{\text{rand}}}.$$

Here, W_{rand} is the Wasserstein distance between the original data and a randomly generated version, serving as a reference point, and W is the Wasserstein distance for the specific released dataset being evaluated. This normalization ensures that the metric reflects the improvement over random values. The second metric evaluates the performance of a k-Nearest Neighbors (kNN) classifier using the Matthews Correlation

Coefficient (MCC), focusing on the data's ability to support predictive analytical tasks. In the context of our results, we refer to these metrics as data fidelity and predictive utility, respectively.

The Rationale for Using kNNs in Predictive Utility Although XGBoost overall achieved better performance across all datasets (see Appendix 1.10), we selected kNN for the predictive utility measurement for two main reasons:

1. **Direct relationship with data:** kNN is recognized as an instance-based explainable method (Molnar, 2020), meaning its predictions are directly influenced by the training data without an explicit model abstraction. This property makes it sensitive to data transformations, and thus it is suitable for evaluating the impact of data release strategies on utility.

2. **Computational efficiency for data Shapley value:** The Data Shapley Value (DSV) for kNN (kNN-Shapley) can be computed exactly in $O(n)$ time (Jia, Dao, Wang, Hubis, Gurel, et al., 2019), in which n is the number of data points. In contrast, calculating DSV for models like XGBoost or Neural Networks requires approximation methods using Monte Carlo simulations, involving evaluating the utility on all possible subsets of the training set, which is computationally intensive (Ghorbani and Zou, 2019)

We explored the use of Data Shapley Value (DSV), particularly with kNN-Shapley, as a method to quantify each data point's contribution to model performance, given its relevance as a data release strategy. We specifically used data Shapley removal to identify and delete less valuable data points, aiming to reduce the dataset size while preserving predictive utility (see Appendix 1.10). However, we encountered challenges related to class imbalance in our datasets (AIC, BM and CCD), which affected the effectiveness of DSV. To mitigate this, we applied a weighting scheme during validation, adjusting the influence of the majority class in data valuation. Our experiments, summarized in Appendix 1.10, showed that an optimal balance ratio could enhance predictive metrics like F1 score and MCC, though there remained a tradeoff with data fidelity as represented by the Wasserstein distance. Additionally, we also explored an alternative approach using Monte Carlo Shapley Value approximations with XGBoost, but due to inconsistent results across datasets and computational intensity, we opted for kNN-based DSV as the primary method for our study. Ultimately, we decided on the strategies of removing 25% and 75% of the data using Data Shapley Value. Similarly, we implemented a random removal strategy, in which 25% and 75% of the data were removed using a uniform distribution. We conducted random removal experiments by repeating the train/validation/test split across multiple iterations. Shapley valuation is computed on the validation set, and utility results are reported on the test set.

As shown in Table 1.2, the Shapley-based removal strategy shows clear advantages in terms of predictive utility, measured by MCC, particularly at the 25% removal threshold. For instance, on the AIC dataset, the MCC for Shapley removal at 25% is 0.5039, compared to 0.4766 for random removal at the same level. This aligns with findings in the data valuation literature (Jia, Dao, Wang, Hubis, Gurel, et al., 2019), in which removing data points with low or negative Shapley values enhances predictive utility by eliminating noisy or detrimental data. However, when considering data fidelity, measured by the normalized Wasserstein distance, Shapley removal underperforms compared to random removal. For example, the Wasserstein distance on the AIC dataset at 25% removal is 0.8071 for Shapley removal, compared to 0.9901 for random removal. This reflects the trade-off between preserving predictive utility and maintaining data fidelity, highlighting the nuanced effects of Shapley removal strategies on different aspects of data utility. Moreover, the observed distributional shift resulting from Shapley-based removal may introduce a trade-off in terms of re-identification risk, potentially increasing privacy vulnerabilities. Our work therefore highlights this unexplored intersection between data valuation and privacy, which we further address in the next section.

For the choice of k parameters in k-anonymity, we selected k values as exponents of 2 (*i.e.*, $k = 2, 4, 8, 16, 32, 64$ and 128). This selection reduces the number of partitions by roughly half with each increase in k, ensuring a consistent trade-off between privacy and utility. We observed that for k values significantly higher than 128, data fidelity dropped significantly. However, the relationship between predictive utility and k-anonymity is not as straightforward. For the BM dataset, there is even a slight improvement when a small generalization (*e.g.*, $k = 2$) is applied to the data.

For the k-anonymity strategy, data fidelity consistently decreased as the value of k increased, a trend observed across all datasets. This behavior aligns with expectations, as higher k values introduce greater generalization. Predictive utility, measured by MCC, showed a similar pattern for the BM and CCD datasets, in which low k values ($k = 2$ and $k = 4$) even improved MCC compared to the original dataset, suggesting that slight generalization may help reduce noise in the data. However, for the AIC dataset, k-anonymity performed noticeably worse than other strategies in terms of predictive utility, even at lower k values. This discrepancy highlights potential differences in how k-anonymity's effect is data dependent.

For data synthesis, we used CTGAN and TVAE, two deep learning models that are widely adopted in the industry, especially for tabular data. CTGAN, based on Generative Adversarial Networks (GANs), addresses the challenges of tabular data, such as imbalanced categorical features and multimodal distributions, by conditioning the generator on specific feature values to produce realistic synthetic data. GANs optimize a loss function in which the generator creates data to fool the discriminator, which learns to distinguish real from synthetic samples. TVAE, built on the Variational Autoencoder (VAE) architecture, encodes input features into a latent space and decodes them to generate synthetic samples, with the synthetic data being directly sampled from the learned latent space. VAEs optimize an objective that balances accurate reconstruction of input data with learning a well-organized latent space. Both models are implemented in the Synthetic Data Vault (SDV) library, a widely adopted tool in the industry, which motivated the use of these models in our study. Referring to Table 1.2, the utility of the synthetic data varies significantly depending on the dataset.

For the AIC dataset, TVAE achieved an MCC score of 0.5014, outperforming CTGAN's MCC of 0.4789, while CTGAN demonstrated superior data fidelity with a

Wasserstein score of 0.9512 compared to TVAE's 0.9349. In contrast, for the BM dataset, CTGAN delivered better data fidelity with a Wasserstein score of 0.8227, outperforming TVAE's 0.7543, but TVAE excelled in predictive utility with an MCC of 0.5486, significantly higher than CTGAN's 0.4621. Similarly, for the CCD dataset, CTGAN achieved a lower Wasserstein distance of 0.8971 compared to TVAE's 0.8834, while TVAE recorded a higher MCC of 0.4893 versus CTGAN's 0.4502. Otherwise, both models performed similarly across the datasets, suggesting that either method can provide comparable utility in many scenarios. Notably, TVAE's MCC for the BM dataset approached the best predictive utility across all strategies. This suggests that while the models can generate synthetic data that approximate the original data, their utility scores indicate more variability in performance, demonstrating that the effectiveness of these generative models can depend heavily on the specific characteristics of the dataset being used.

Table 1.2: *Summary of Utility Metrics for AIC, BM, and CCD Datasets*

Dataset	**Strategy**	AIC	BM	CCD
		$\widetilde{W}$ | MCC	$\widetilde{W}$ | MCC	$\widetilde{W}$ | MCC
k_1	Original	1.0000 | 0.5070	1.0000 | 0.2992	1.0000 | 0.2234
k_2	Data Anonymization	0.9657 | 0.4476	0.9624 | 0.4164	0.7631 | 0.3211
k_4	Data Anonymization	0.9420 | 0.4097	0.9473 | 0.4100	0.6454 | 0.2812
k_8	Data Anonymization	0.8876 | 0.3960	0.9105 | 0.3906	0.4476 | 0.2302
k_16	Data Anonymization	0.8541 | 0.4381	0.8750 | 0.3150	0.3492 | 0.1400
k_32	Data Anonymization	0.7923 | 0.3981	0.8221 | 0.2540	-
k_64	Data Anonymization	0.7643 | 0.4340	0.7580 | 0.1794	-
k_128	Data Anonymization	0.6807 | 0.4396	0.6866 | 0.1368	-
tvae	Data Synthesis	0.9239 | 0.4461	0.8214 | 0.3859	0.6143 | 0.1717
ctgan	Data Synthesis	0.8433 | 0.4614	0.8305 | 0.1864	0.6420 | 0.1310
shapley_75	Data Minimization	0.9020 | 0.5057	0.9318 | 0.3238	0.9330 | 0.2463
shapley_50	Data Minimization	0.8909 | 0.5071	0.8956 | 0.3249	0.8644 | 0.2147
shapley_25	Data Minimization	0.8071 | 0.5039	0.8779 | 0.3053	0.8057 | 0.1645
random_75	Data Minimization	0.9968 | 0.4932	0.9940 | 0.2901	0.9699 | 0.2099
random_50	Data Minimization	0.9935 | 0.4895	0.9883 | 0.2925	0.9404 | 0.1997
random_25	Data Minimization	0.9901 | 0.4766	0.9737 | 0.2708	0.8908 | 0.1777

1.5.2 Privacy Risk Scores and Tradeoffs

For each dataset, we conducted 200 re-identification attack simulation iterations. In each iteration, a batch of 100 individuals was randomly sampled without replacement and targeted for the attack. The reported results represent the average performance across these 200 iterations for each batch of 100 individuals. One way to differentiate between simulations on a single dataset is by attack scenario type. For example, we ran a simulation to estimate the risk of re-identification in a scenario in which an attacker has access to all features as resources, with all features also considered quasi-identifiers (Figure 1.3). Although this may not be entirely realistic, as it is uncommon to treat all attributes as quasi-identifiers, this approach establishes a boundary for assessing how easily records in the dataset can be re-identified. Unsurprisingly, when an attacker has access to all resources and attacks the original AIC dataset, the average linkability score is 91.2% (Figure 1.3). This score is not 100% because a tabular dataset like AIC can have records that are nearly identical and possibly duplicates when examined through quasi-identifiers. However, it is important to note that the attacker has no attributes to infer in that case. The situation becomes more interesting when the attacker has fewer resources. For instance, we find that the average linkability score is 11.5% when the attacker has access to 3 features.

While it would ultimately be up to decision-makers within an organization to determine the plausibility of the different scenarios to consider, our approach remains versatile and adaptable to different conditions and contexts. One may compute the score of a given strategy by averaging all scenarios. Alternatively, there are other ways to compare "leave k out" atrributes scenarios. For instance, instead of weighting them equally, we can use a binomial weighting scheme. The rationale is to put less weight on scenarios that are unlikely to be run by attackers. For example, the scenario in which an attacker has all the information but still tries to re-identify someone is not very likely. Likewise, the scenario where an attacker can identify someone based on knowledge of just one feature is improbable, except when that feature functions as a direct identifier in the dataset for certain individuals (*e.g.*, a precise numerical value such as a salary or a specific monetary transaction amount). We believe it is reasonable to think these scenarios are less frequent, and thus, a binomial weighting scheme can better reflect the practical likelihood of various attack scenarios. This method allows us to assign more realistic weights to each scenario, enhancing the robustness and applicability of our risk assessment framework. We used this approach to aggregate the scenario scores and build a comprehensive score for each individual and each strategy (Table 1.3, Table 1.4). In a real-world setting, it would also be important to assign a minimum weight to the linkability scores. In our experiments, we chose $\varepsilon_2 = 0.5$ and $\varepsilon_1 = 100$. Smaller values of ε_1 (e.g., $\varepsilon_1 = 1$) treat all reconstructed features equally, while excessively high values ($\varepsilon_1 \gg 100$) make extreme scenarios negligible, such as reconstructing all or no features. Our choice of $\varepsilon_1 = 100$ strikes a balance, ensuring stability. For similar reasons, we picked $\varepsilon_2 = 0.5$ to provide a moderate weight to linkability even when information gain is zero, avoiding extreme parameters like $\varepsilon_2 = 0$ (which would put weight on the membership risk only when an attacker can reconstruct a feature) or $\varepsilon_2 = 1$ (which equates membership risk and re-identification risk). These values can be adjusted based on the specific context of the organization or the dataset being analyzed, and exploring a broader range of values in future work could provide further insights.

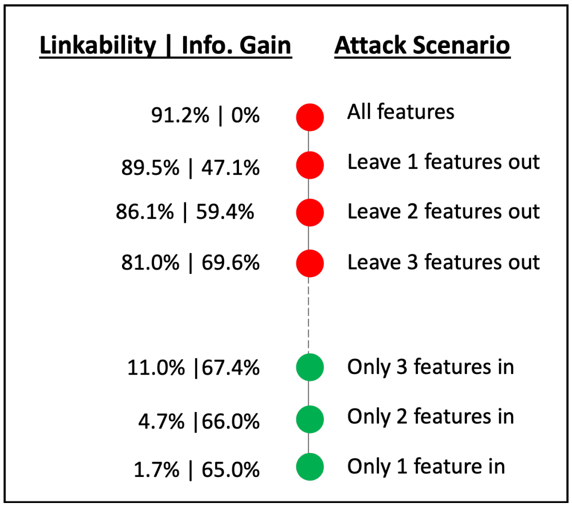
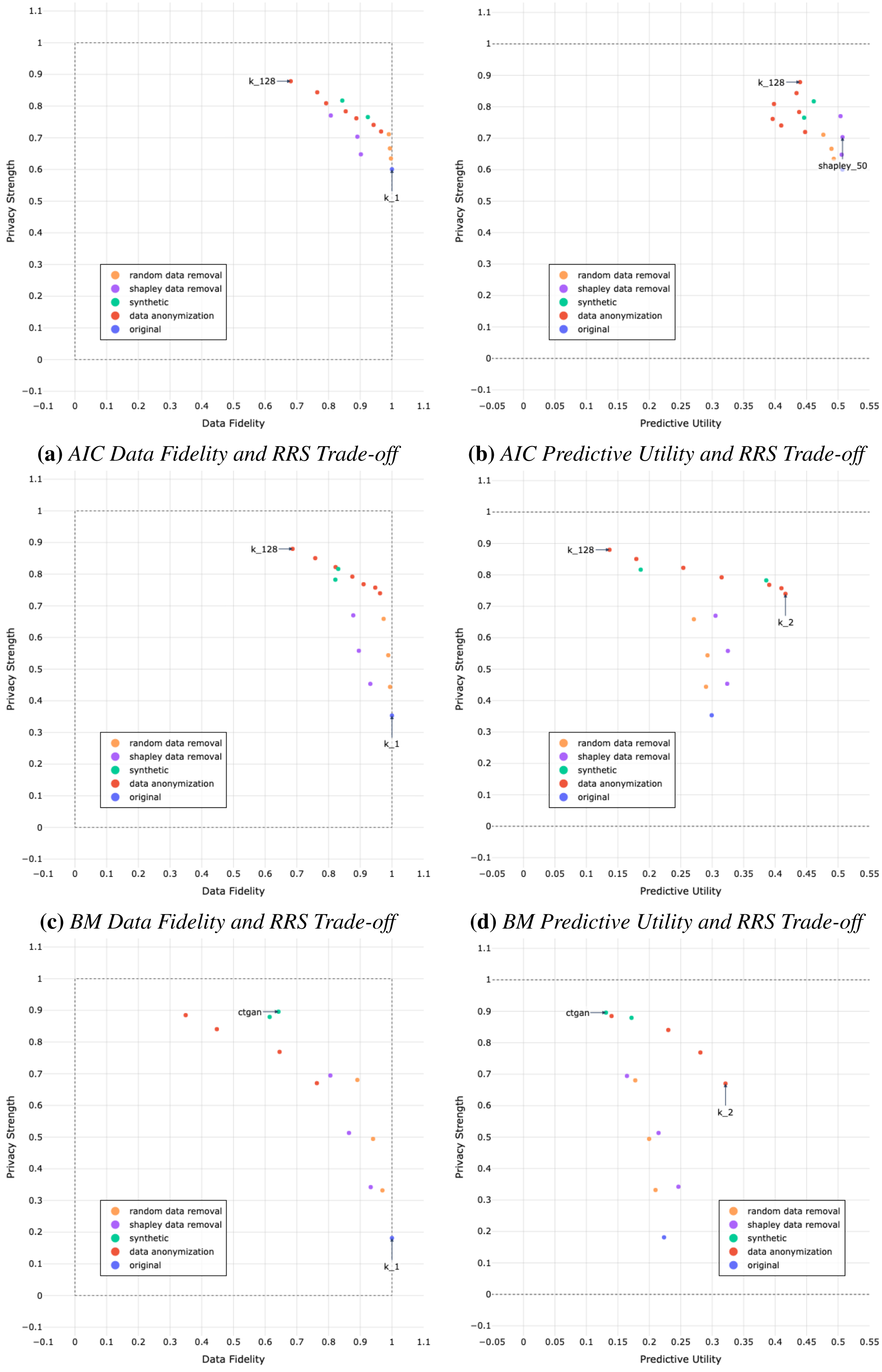


Figure 1.3: *Risk of re-identification by scenario on the AIC dataset*



(a) *AIC Data Fidelity and RRS Trade-off*

(b) *AIC Predictive Utility and RRS Trade-off*

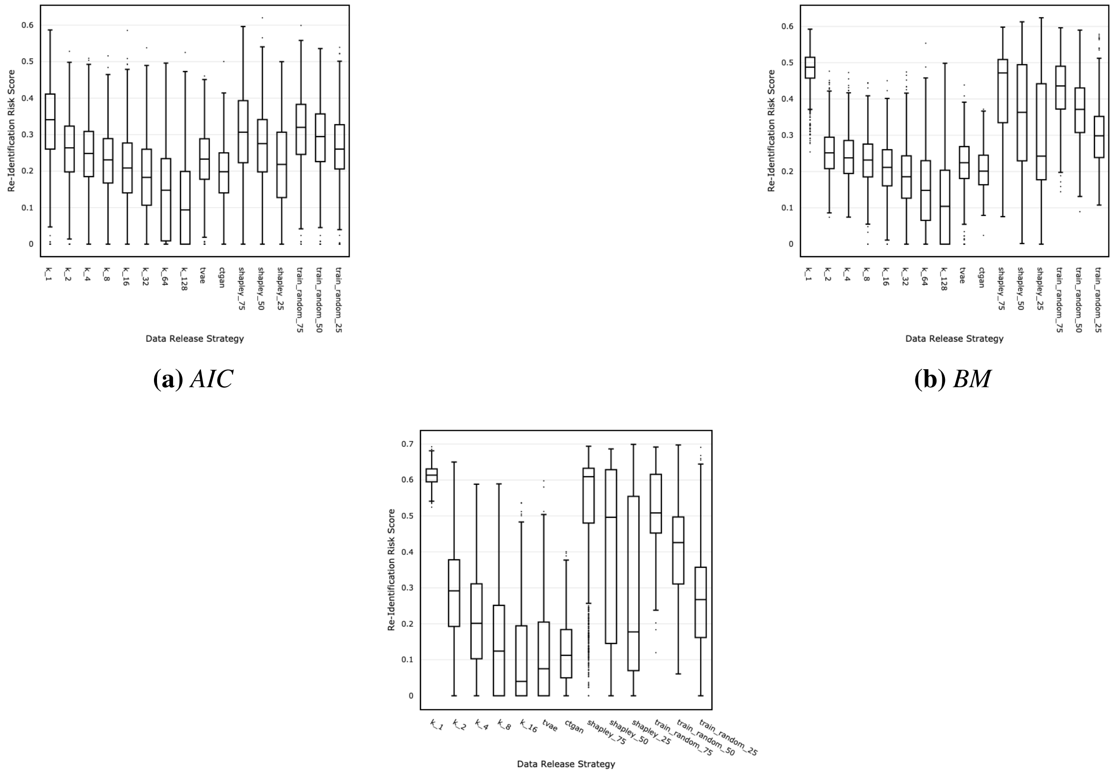
(c) *BM Data Fidelity and RRS Trade-off*

(d) *BM Predictive Utility and RRS Trade-off*

(e) *CCD Data Fidelity and RRS Trade-off*

(f) *CCD Predictive Utility and RRS Trade-off*

Figure 1.4: *This figure illustrates the trade-offs between privacy and utility, evaluated using two key metrics: Data Fidelity and Predictive Utility. Data Fidelity, measured using the Wasserstein distance, quantifies how well the statistical properties of the original dataset are preserved after privacy-preserving transformations. Predictive Utility, assessed using the Matthews Correlation Coefficient (MCC), evaluates the impact of these transformations on the performance of predictive models.*



(a) *AIC*

(b) *BM*

(c) *CCD*

Figure 1.5: *Box Plot Distributions of RRS for AIC, BM and CCD Datasets.*

Table 1.3: *AIC Re-Identification Simulations Results*

Dataset	**Strategy**	**Linkability (Avg ± Std)**	**Information Gain (Avg ± Std)**	**Re-identification Score (Avg ± Std)**
k_1	Original	$0.526 \pm 6.83e{-}03$	$0.700 \pm 2.77e{-}03$	$0.399 \pm 5.91e{-}03$
k_2	Data Anonymization	$0.398 \pm 6.77e{-}03$	$0.627 \pm 2.81e{-}03$	$0.280 \pm 5.18e{-}03$
k_4	Data Anonymization	$0.375 \pm 7.31e{-}03$	$0.610 \pm 2.91e{-}03$	$0.259 \pm 5.39e{-}03$
k_8	Data Anonymization	$0.348 \pm 7.09e{-}03$	$0.600 \pm 2.84e{-}03$	$0.239 \pm 5.11e{-}03$
k_16	Data Anonymization	$0.317 \pm 6.68e{-}03$	$0.590 \pm 2.87e{-}03$	$0.217 \pm 4.72e{-}03$
k_32	Data Anonymization	$0.284 \pm 6.18e{-}03$	$0.574 \pm 2.97e{-}03$	$0.191 \pm 4.25e{-}03$
k_64	Data Anonymization	$0.237 \pm 5.02e{-}03$	$0.548 \pm 3.13e{-}03$	$0.156 \pm 3.33e{-}03$
k_128	Data Anonymization	$0.189 \pm 3.57e{-}03$	$0.517 \pm 3.50e{-}03$	$0.122 \pm 2.30e{-}03$
tvae	Data Synthesis	$0.339 \pm 7.37e{-}03$	$0.619 \pm 3.16e{-}03$	$0.235 \pm 5.56e{-}03$
ctgan	Data Synthesis	$0.271 \pm 6.22e{-}03$	$0.580 \pm 3.11e{-}03$	$0.183 \pm 4.49e{-}03$
shapley_75	Data Minimization	$0.471 \pm 6.55e{-}03$	$0.678 \pm 2.62e{-}03$	$0.352 \pm 5.46e{-}03$
shapley_50	Data Minimization	$0.408 \pm 7.01e{-}03$	$0.654 \pm 2.83e{-}03$	$0.297 \pm 5.61e{-}03$
shapley_25	Data Minimization	$0.317 \pm 6.54e{-}03$	$0.636 \pm 3.01e{-}03$	$0.230 \pm 5.09e{-}03$
random_75	Data Minimization	$0.489 \pm 6.51e{-}03$	$0.681 \pm 2.69e{-}03$	$0.365 \pm 5.49e{-}03$
random_50	Data Minimization	$0.453 \pm 6.32e{-}03$	$0.663 \pm 2.66e{-}03$	$0.333 \pm 5.17e{-}03$
random_25	Data Minimization	$0.401 \pm 6.30e{-}03$	$0.641 \pm 2.74e{-}03$	$0.289 \pm 4.93e{-}03$

To provide an intuitive understanding of our metrics, we examine the AIC dataset using the original data ($k1$) and after applying k-anonymity with $k = 2$. In the original dataset, the **Linkability score** is 0.526, indicating a relatively high likelihood that a

Table 1.4: *BM Re-Identification Simulations Results*

Dataset	Strategy	Linkability (Avg ± Std)	Information Gain (Avg ± Std)	Re-identification Score (Avg ± Std)
k_1	Original	$0.803 \pm 8.35e-03$	$0.770 \pm 3.64e-03$	$0.647 \pm 8.22e-03$
k_2	Data Anonymization	$0.378 \pm 7.10e-03$	$0.572 \pm 3.39e-03$	$0.260 \pm 5.13e-03$
k_4	Data Anonymization	$0.359 \pm 7.14e-03$	$0.558 \pm 3.30e-03$	$0.243 \pm 5.00e-03$
k_8	Data Anonymization	$0.352 \pm 6.88e-03$	$0.540 \pm 3.11e-03$	$0.232 \pm 4.62e-03$
k_16	Data Anonymization	$0.325 \pm 6.46e-03$	$0.512 \pm 3.16e-03$	$0.208 \pm 4.11e-03$
k_32	Data Anonymization	$0.285 \pm 5.61e-03$	$0.483 \pm 3.34e-03$	$0.178 \pm 3.43e-03$
k_64	Data Anonymization	$0.244 \pm 4.36e-03$	$0.459 \pm 3.54e-03$	$0.149 \pm 2.55e-03$
k_128	Data Anonymization	$0.200 \pm 3.37e-03$	$0.433 \pm 3.67e-03$	$0.120 \pm 1.91e-03$
tvae	Data Synthesis	$0.324 \pm 7.43e-03$	$0.562 \pm 3.45e-03$	$0.217 \pm 5.26e-03$
ctgan	Data Synthesis	$0.284 \pm 7.29e-03$	$0.532 \pm 3.45e-03$	$0.183 \pm 5.01e-03$
shapley_75	Data Minimization	$0.695 \pm 6.60e-03$	$0.714 \pm 2.96e-03$	$0.547 \pm 6.32e-03$
shapley_50	Data Minimization	$0.582 \pm 5.80e-03$	$0.656 \pm 2.69e-03$	$0.442 \pm 5.03e-03$
shapley_25	Data Minimization	$0.454 \pm 5.46e-03$	$0.599 \pm 2.66e-03$	$0.330 \pm 4.08e-03$
random_75	Data Minimization	$0.704 \pm 6.81e-03$	$0.722 \pm 3.05e-03$	$0.556 \pm 6.54e-03$
random_50	Data Minimization	$0.594 \pm 5.81e-03$	$0.667 \pm 2.67e-03$	$0.456 \pm 5.08e-03$
random_25	Data Minimization	$0.469 \pm 6.09e-03$	$0.601 \pm 2.84e-03$	$0.341 \pm 4.56e-03$

Table 1.5: *CCD Re-Identification Simulations Results*

Dataset	Strategy	Linkability (Avg ± Std)	Information Gain (Avg ± Std)	Re-identification Score (Avg ± Std)
k_1	Original	$0.975 \pm 2.18e-03$	$0.834 \pm 1.14e-03$	$0.819 \pm 2.37e-03$
k_2	Data Anonymization	$0.495 \pm 4.06e-03$	$0.560 \pm 2.45e-03$	$0.330 \pm 2.86e-03$
k_4	Data Anonymization	$0.363 \pm 3.90e-03$	$0.525 \pm 2.59e-03$	$0.231 \pm 2.61e-03$
k_8	Data Anonymization	$0.254 \pm 3.16e-03$	$0.507 \pm 2.57e-03$	$0.159 \pm 2.08e-03$
k_16	Data Anonymization	$0.183 \pm 2.61e-03$	$0.500 \pm 2.68e-03$	$0.115 \pm 1.72e-03$
tvae	Data Synthesis	$0.188 \pm 2.63e-03$	$0.531 \pm 2.24e-03$	$0.121 \pm 1.74e-03$
ctgan	Data Synthesis	$0.179 \pm 3.07e-03$	$0.457 \pm 2.58e-03$	$0.104 \pm 1.87e-03$
shapley_75	Data Minimization	$0.804 \pm 2.72e-03$	$0.745 \pm 1.51e-03$	$0.658 \pm 2.55e-03$
shapley_50	Data Minimization	$0.617 \pm 2.84e-03$	$0.655 \pm 1.73e-03$	$0.487 \pm 2.31e-03$
shapley_25	Data Minimization	$0.411 \pm 2.90e-03$	$0.571 \pm 2.10e-03$	$0.306 \pm 2.14e-03$
random_75	Data Minimization	$0.818 \pm 2.78e-03$	$0.752 \pm 1.48e-03$	$0.668 \pm 2.57e-03$
random_50	Data Minimization	$0.643 \pm 2.93e-03$	$0.666 \pm 1.69e-03$	$0.506 \pm 2.34e-03$
random_25	Data Minimization	$0.429 \pm 3.03e-03$	$0.576 \pm 2.15e-03$	$0.320 \pm 2.28e-03$

singled-out record belongs to the original dataset. The **Information Gain** is 0.700, meaning the attacker achieves a low reconstruction error when attempting to infer the features of a singled-out record. The **Re-identification Score**, which combines both Linkability and Information Gain, is 0.399, reflecting a significant overall risk of re-identification. After applying k-anonymity with $k = 2$, the average linkability score decreases to 0.398, reducing the likelihood that a record belongs to the original dataset. The Information Gain is also reduced to 0.627, indicating a modest decrease in Information Gain. Notably, the reduction in the Linkability score is more substantial than the reduction in the Information Gain, suggesting that data anonymization primarily mitigates the likelihood of records being linked to individuals rather than significantly increasing reconstruction error. Consequently, the Re-identification Score drops to 0.280, demonstrating that *k*-anonymity effectively reduces the overall risk of re-identification, with a stronger impact on Linkability.

Figure 1.4 shows the aggregated risk of re-identification for each dataset alongside two utility metrics: data fidelity (Figures 1.4a, c, and e) and predictive utility (Figures 1.4b, d, and f). For data anonymization strategies, adjusting the k parameter in *k*-anonymity reveals clear trade-offs between risk and utility. As k increases, the size of generalized partitions grows, reducing the risk of re-identification but also leading to a loss of data fidelity due to diminished variance within each partition. For predictive utility, the trade-offs are also evident: smaller k values ($k = 2$ and $k = 4$) improve predictive utility for the BM and CCD datasets, as seen in higher MCC scores, while larger k values reduce utility. However, in the AIC dataset, *k*-anonymity performs worse for predictive utility than other strategies.

For data minimization strategies, both data Shapley removal and random removal significantly reduced re-identification risk. However, Shapley removal consistently achieved slightly better reductions, with the AIC dataset showing a notable improvement from a risk score of 0.289 (random removal) to 0.230 (Shapley removal) at the 75% removal level. This suggests that removing data points with lower Shapley values positively impacts re-identification risk in *k*-nearest neighbor evaluations. In terms of utility, Shapley removal sometimes improved predictive utility compared to using the full dataset, making the strategy better on both fronts.

Finally, for data synthesis methods (CTGAN and TVAE), both performed similarly in reducing re-identification risk, achieving results comparable to *k*-anonymity with k below 64. These results indicate that data synthesis methods can achieve similar levels of privacy protection as *k*-anonymity while maintaining comparable utility metrics. These results also demonstrate that our re-identification risk framework is robust to these specific data synthesis methods, ensuring they do not serve as a loophole within the framework.

Figures 1.5 display the distribution of *RRS* for each strategy across all datasets. We observe a positive effect from either increasing the k parameter in *k*-anonymity or increasing the percentage of data removed using Shapley values, although removing data points is not as effective. In all our benchmarks, synthetic data performed very similarly to k-anonymity strategies with a small degree of generalization (*e.g.*, $k = 4$ and $k = 5$).

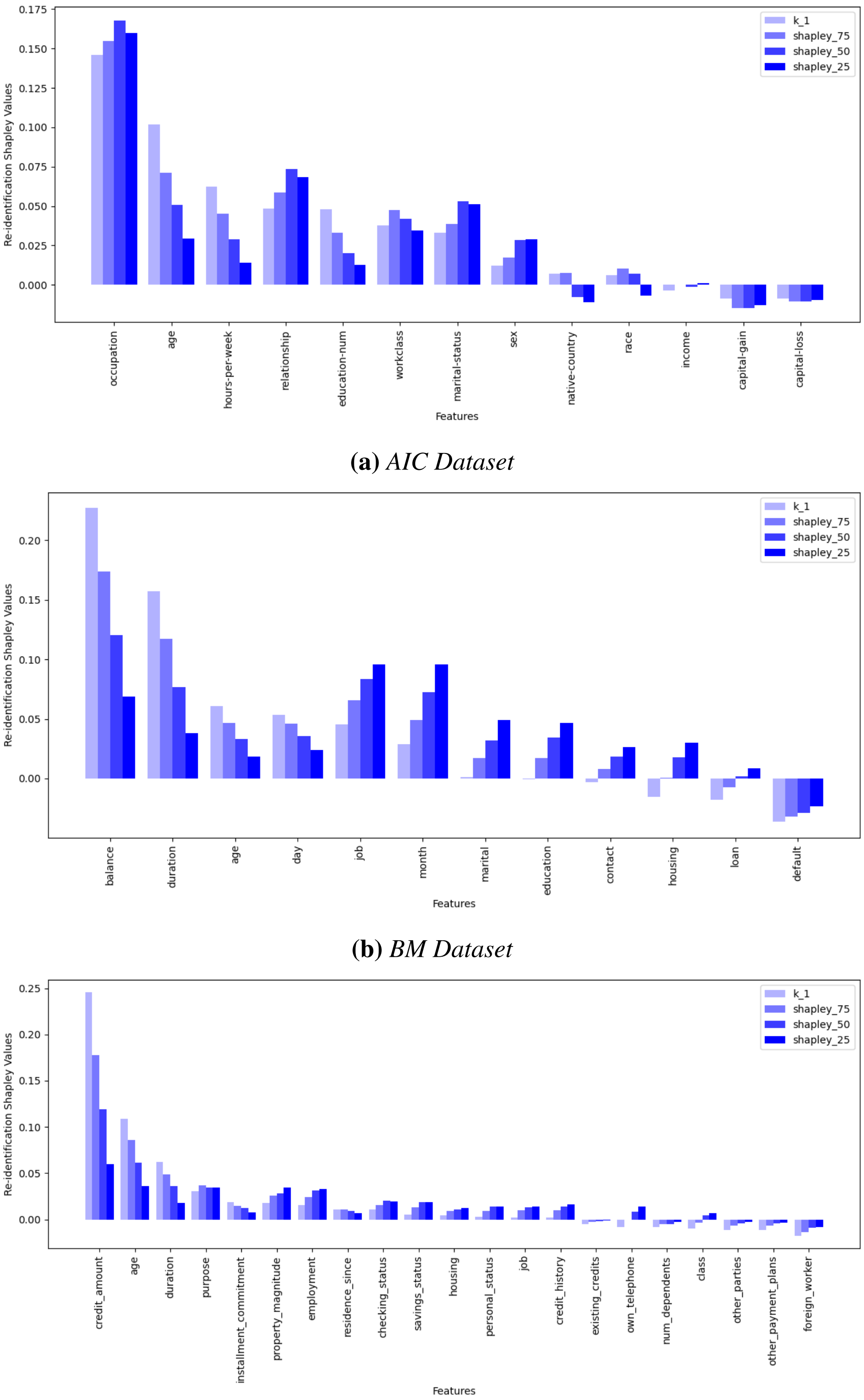
1.5.3 Re-identification Shapley Value Analysis

For each individual, we performed 50 permutations, in which each permutation represents a different order of features to compute the RSVs, providing a robust estimate of the contribution of each feature to the re-identification risk. We observed that this number was more than enough to achieve robust estimates when aggregating over all individuals. This observation is further supported by theoretical findings from Maleki et al., 2013, which demonstrate that sampling-based methods for Shapley value approximations can achieve accurate results with a limited number of samples.

In the AIC dataset, the two most important features for re-identification are “age” and “occupation,” while “balance” and “duration” dominate in the BM dataset, with age also being a significant feature. For the CCD dataset, the “credit_amount” attribute, which corresponds to the credit amount requested, exhibits the highest re-identification potential. This is notable because, in organizational contexts, “credit_amount” may not be perceived as sensitive as personally identifiable information (PII) like names or addresses, or even features such as account balance. These findings underscore the context-dependent nature of re-identification risks across datasets.

Figure 1.6 shows an analysis of the effect of data Shapley removal on the RSVs, while Figure 1.7 displays the RSVs for the effect of k-anonymity on the RSVs. For both types of strategies, we see a general positive effect in reducing the RSVs. We also observe that k-anonymity has a more significant impact on the RSVs than data Shapley value removal. This is consistent with the results of the privacy-utility trade-off analysis, in which k-anonymity strategies were more effective in reducing the risk of re-identification than data Shapley value removal strategies. We also observe that the top feature in AIC (namely “occupation”) requires $k = 128$ to have a significant impact on the RSVs, while other features require $k = 64$ or less to reduce the RSVs. Figure 1.8 illustrates the RSVs for data synthesis, revealing that “occupation” is once again a feature more accurately reproduced by both CTGAN and TVAE.

Interestingly, we note that some features, such as the capital-gain, capital-loss, and native-country in the AIC dataset, have little value in re-identification attacks. We decided to investigate the relationship between the Re-identification Shapley Values (RSVs) and the entropy of the AIC features, both calculated on the original dataset, suspecting that entropy could serve as a singling-out factor. Upon conducting a correlation analysis using Pearson’s correlation coefficient, we found a substantial positive correlation between the two. This suggests that higher entropy attributes are often associated with higher RSVs. For instance, the “age” attribute, which is present in all three datasets, exhibited high entropy and a high re-identification potential, highlighting the need for a higher security level for such features. For example, organizations could implement policies to monitor queries involving high-risk attributes like ’age’ and set stricter rules for data access or feature removal to mitigate re-identification risks.

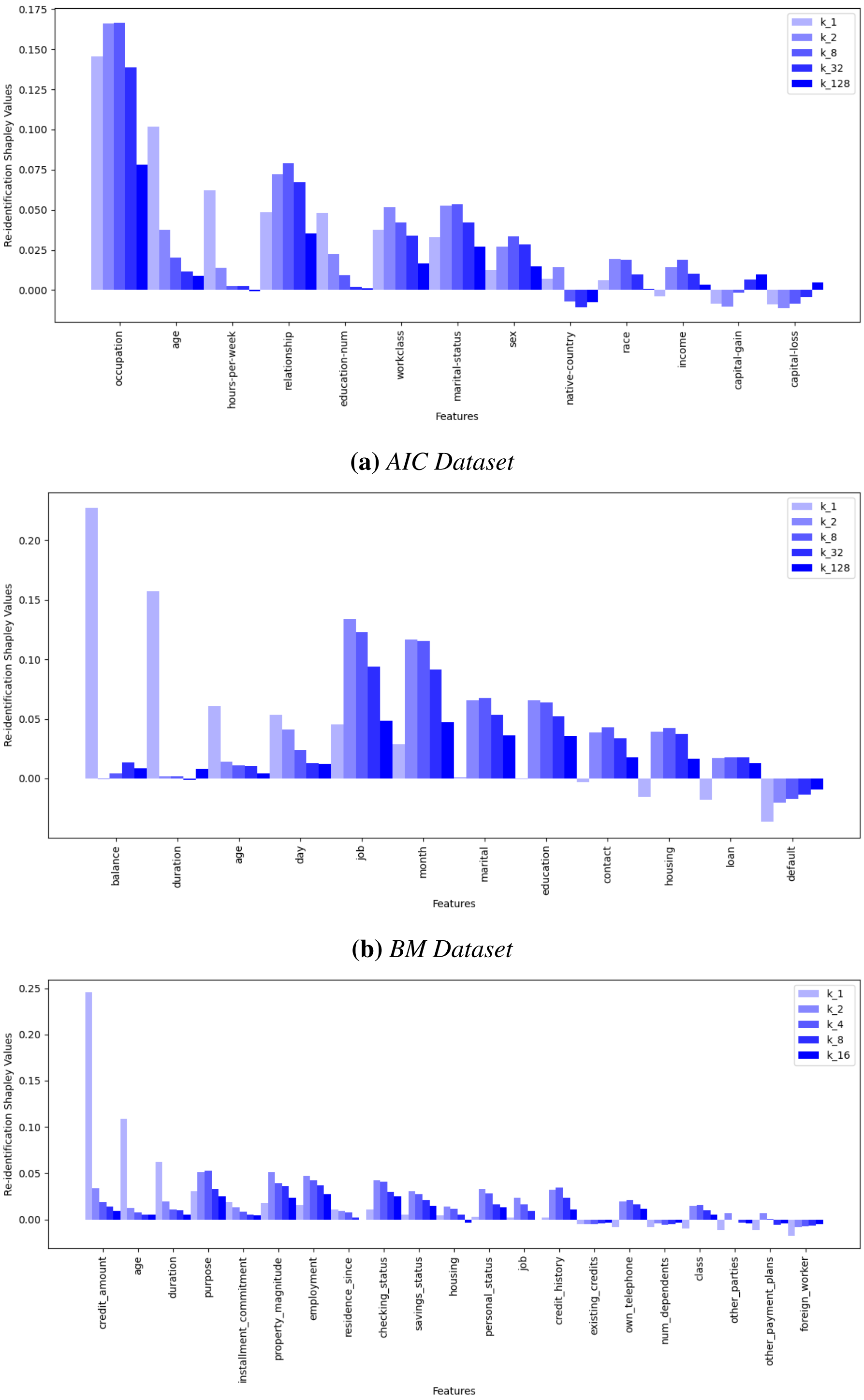


(a) *AIC Dataset*

(b) *BM Dataset*

(c) *CCD Dataset*

Figure 1.6: *RSVs on data minimization strategies using data Shapley values*

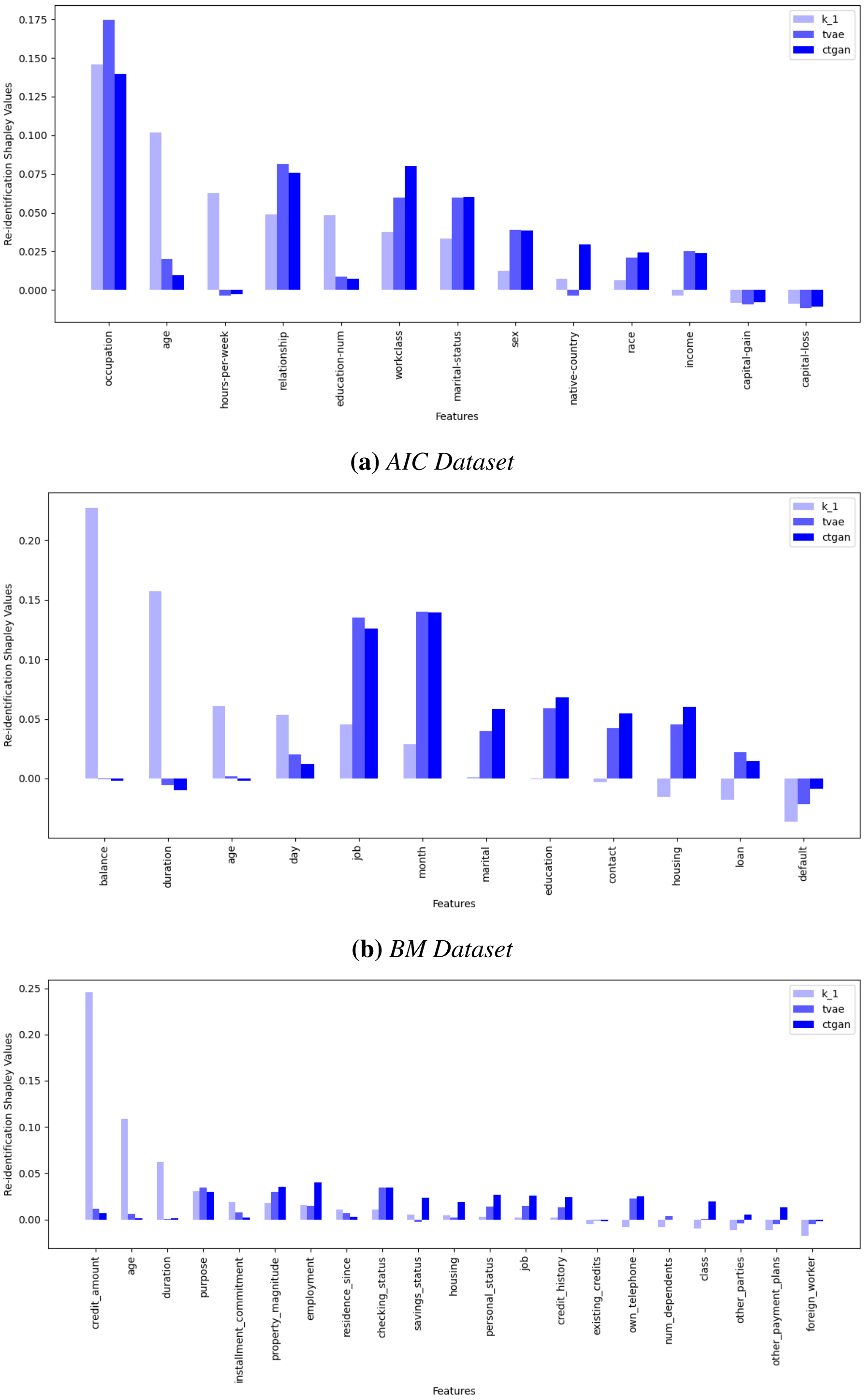


(a) *AIC Dataset*

(b) *BM Dataset*

(c) *CCD Dataset*

Figure 1.7: *RSVs on data anonymization strategies.*



(a) *AIC Dataset*

(b) *BM Dataset*

(c) *CCD Dataset*

Figure 1.8: *RSVs on data synthesis strategies.*

1.6 Discussion

Our experiments demonstrated the practicality of our approach in evaluating the risk of re-identification with respect to different data release strategies. More precisely, we focused on three main strategies: data minimization, data anonymization, and data synthesis. Data minimization and data anonymization still have a direct link to the original data, while data synthesis relies on a generative model and not the data itself. Altering the data changes its structure, while modeling (such as with synthetic data) has no direct link to the original data, only an indirect link via training. Each approach has unique advantages. In particular, synthetic data allows for the creation of additional data points and does not seem to require particular privacy mechanisms to offer some privacy-preserving properties. *k*-anonymity can perform well and provides an easier rationale for linkability, which is whyorganizations use *k*-anonymity variants (like *l*-diversity Machanavajjhala et al., 2007a). We contribute to the literature on data minimization, data anonymization and data synthesis by investigating the implications of each strategy individually. Future research could explore the effects of combining these strategies. Our framework is useful for data custodians such as data governance and cybersecurity teams. They could use our approach every time they are about to release data, either internally or to share outside of the organization. However, we caution against sharing outside the organization solely based on a post-hoc risk analysis tradeoff; our risk framework does not offer protections akin to techniques such as differential privacy (Dwork and Roth, 2014b). In particular, we did not use differential privacy with data synthesis because differential privacy is a property of the mechanism and not the data, making it difficult to compare directly to other parameters like *k*-anonymity, which is a property of the data. Nonetheless, our approach also remains versatile in that we do not explicitly define criteria to decide the right protection and utility thresholds, as these depend on the nature of the data, its classification (*e.g.*, PII), and the difficulty for an attacker to obtain access to information resources.

To gain further monetary insights into the costs an organization may incur from a partially anonymized or even synthetic dataset breach, a sensitivity analysis can be added to extend our privacy scores. For example, we could adapt our method to use different utility metrics depending on the chosen focus. A 1% decrease in utility may have varying costs across different departments within an organization. If both privacy strength and data utility are converted into monetary values, they can provide more pragmatic insights for decision-makers. Even if the complete estimation of the monetary costs associated with data breaches may be challenging, recent events attest to the importance of this issue. According to a recent IBM report, the average cost of a data breach in the U.S. is approximately $5 million (IBM, 2022).

Despite its usefulness, we acknowledge important assumptions that can limit the effectiveness of our framework. These limitations could lead to an underestimation or an overestimation of the risk we want to evaluate.

First, in our setup, we framed the risk evaluation as a type of security game in which attackers submit records and retrieve a score. Other setups could be explored to evaluate this risk. For instance, one could envision a security game in which an attacker attempts to infer an attribute about an individual. Second, our approach uses a specific implementation of singling-out attacks. In particular, our use of the unnormalized Gower distance relies on fixed weights associated with resources. A potential improvement to overcome this limitation would be to use a different type of predictor that is not subject to the same constraints, while remaining suitable for use with tabular data. Third, our linkability evaluation method depends heavily on fine-tuning of parameters, which may vary across datasets. As an alternative, an attack such as Shokri's membership attack (Shokri et al., 2017b) on synthetic data appears relevant. Similarly, in the evaluation of the information gain, a machine learning model could be employed to predict the information to gain, instead of relying solely on distance metrics without considering external knowledge. Our current implementation of information gain does not consider the potential inferences that could be drawn from the reconstructed information.

Additionally, we assume attackers have access to resources in the same format as the original dataset, which may lead to an overestimation of the risk. In real-life scenarios, attackers often have aggregated or noisy information, making their resources less precise than the clean datasets we simulate. While this means that our approach errs on the side of caution by being more conservative than may be necessary, future work could simulate different conditions for a more realistic assessment of the risk of re-identification. Additionally, not all attributes in a dataset are equally accessible to attackers. For example, quasi-identifiers such as age, gender, and ZIP code are often easier to obtain from external sources, whereas other attributes may be much harder to acquire. This distinction is not currently accounted for in our approach, as Shapley values treat all features as equally accessible resources. An interesting extension of our work would be to weight Shapley values with estimates of the difficulty of acquiring different attributes. Such an adjustment would provide a more nuanced understanding of the re-identification risks based on both the value and accessibility of resources. Because of these limitations, the risk results should not be considered as absolute levels of risk but rather as insights to understand the dynamics of a strategy (*e.g.*, *k*-anonymity) or how one strategy performs relative to another. Another limitation of our study is the lack of analysis of privacy risks for subgroups, such as majority and minority populations. Data Shapley removal has the potential to alter the group balance within datasets, as it favors data points based on their utility contribution. This could potentially expose these groups to higher re-identification risks. While our trade-off analysis showed that Shapley removal and random removal yield comparable re-identification risk scores, with Shapley removal showing slight improvements in some cases, the potential differential impact on subgroups was not explicitly explored. The observed reduction in re-identification risk primarily stems from assigning a risk of zero to removed individuals, but the implications for privacy equity between subgroups remain unclear. Analyzing subgroup-specific impacts across all strategies—data minimization, data anonymization, and data synthesis—could provide additional insights. Future work could extend this analysis to better understand how privacy risks vary across populations, ensuring more fair privacy management.

1.7 Conclusion

In this work, we have proposed an approach for evaluating the risk of re-identification in tabular data by considering the risk from the perspective of what would be valuable to an attacker in singling out individuals. Our approach allows us to account for the information resources at the disposal of attackers and their use to single out individuals. In the security game we introduced, an evaluator scores attacks to assess the risk of re-identification of individuals based on two factors: the linkability of singled out records and the information gain of an attacker. This evaluation yields a re-identification risk score for each individual, enabling comparisons across different data release strategies.

We also introduced the concept of Re-identification Shapley values (RSVs) to estimate the value of information in a privacy attack as a form of cooperative game involving multiple attackers based on our approach. We tested our approach on anonymized and synthetic versions of the Adult Income Census (AIC), Bank Marketing (BM) and Credit Card Default (CCD) datasets, demonstrating its usefulness in realistic scenarios. Our experiments showed that across data release strategies, some features like "age" (found in all three datasets) have a high synergy with other features for re-identification and should be assigned a higher security level. While data requirements and their associated impacts remain contextual to the organization in which they are considered, our approach provides valuable insights for data custodians, helping them balance the trade-off between data privacy and data utility.

1.8 Addendum: Individual Risk and Utility Analysis on the Folktables Dataset

In this addendum, we extend our published analysis by focusing on individual-level privacy and utility scores. While our main study evaluated re-identification risk at an aggregate level across data release strategies, we had not yet examined how this risk compares with individual utility as measured by data Shapley values. Here, we address this gap by jointly analyzing the re-identification risk scores and utility contributions of individual data points, focusing on two deterministic and traceable strategies: *k*-anonymity and data minimization through Shapley-based removal. The goal of this analysis is to better understand the tradeoffs at the individual level—how much value a data point contributes to predictive performance versus how much risk it incurs under a given release strategy. This is particularly relevant as our methodology emphasized individual risk exposure, but did not address in the experiments whether individuals whose data is more vulnerable are also contributing significantly to the utility of a model.

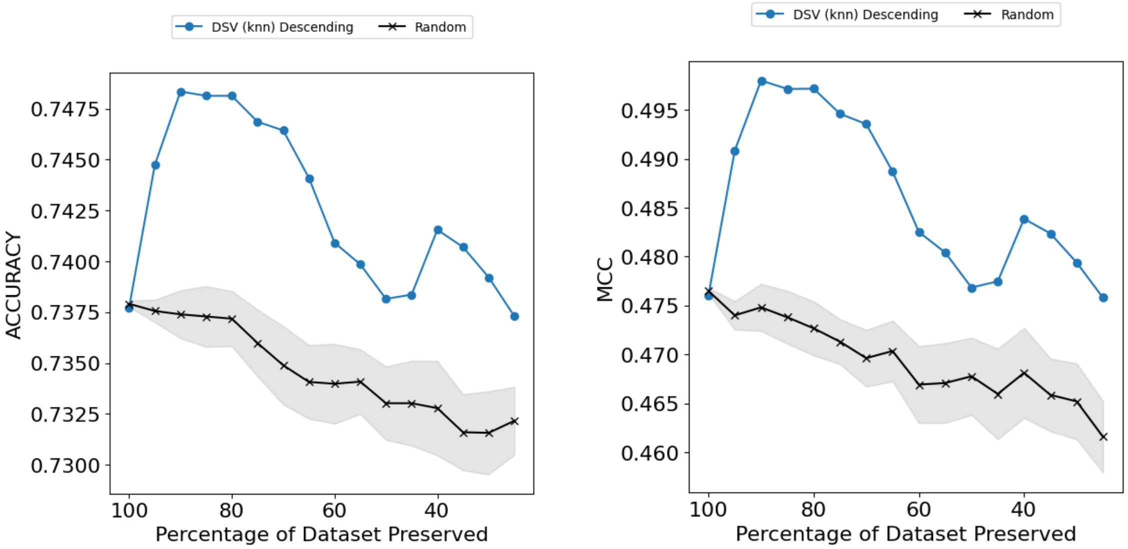
As previously discussed, we assume that individuals have already consent to the use of their data, with the release taking place internally within the organization—for instance, when data is shared between teams for modeling or evaluation. While individuals may not control how their data is processed post-consent, organizations should still be attentive to how release strategies affect individual-level tradeoffs. If a method decreases an individual's re-identification risk while preserving or increasing their utility (as reflected by Shapley value), it reduces the implicit disincentive to data contribution. This perspective aligns with the incentive-theoretic view of differential privacy introduced by McSherry and Talwar, which frames privacy defenses as a mechanism for encouraging truthful participation. Although our methods do not rely on differential privacy, we adopt a similar lens: evaluating strategies not just by their privacy guarantees, but by how they shape the distribution of incentives within the dataset.

To support this analysis, we rely on the Folktables dataset—a well-structured alternative to the widely used Adult dataset. As highlighted by Ding et al., the Adult dataset suffers from outdated features and social biases that can complicate fairness and generalizability assessments. In our case, we encountered specific technical limitations with the Adult dataset, particularly class imbalance, which distorted Shapley value estimation and introduced instability in utility assessments. By contrast, the Folktables dataset allows us to construct balanced prediction tasks—such as the ACSIncome task, in which income is derived from the PINCP attribute using a $25,000 threshold—which enabled more stable data Shapley values by mitigating the class imbalance issues we encountered with the Adult dataset. The rest of the features in the task are comparable to those in the Adult dataset.

This addendum focuses on predictive utility and Shapley value as a measure of individual value, maintaining the thesis's broader emphasis on model-based contribution. Our objective here is to understand how individual risk and utility across data release strategies, and how the distributions may inform organizational decisions that are both privacy-aware and incentive-aligned.

We reproduced the Shapley-based data minimization strategy on the Folktables dataset, using a *k*-NN classifier to evaluate the impact of removing low-value points. Figure 1.9 shows how accuracy and MCC evolve as the dataset is progressively reduced. In both cases, we observe that removing data points in descending order of their Shapley value yields better performance than random removal. Unlike the datasets used in the published paper, we did not observe strong effects of class imbalance, which previously distorted Shapley value estimates and degraded stability. This result confirms that Shapley-based data minimization is more effective at preserving utility by identifying and discarding low-contribution points, outperforming random baselines across a wide range of dataset sizes.

We similarly computed re-identification risk scores (RRS) for 4,400 individuals under the same data release strategies evaluated in the core paper, including data anonymization, synthetic data generation and both random and Shapley-based data minimization. These



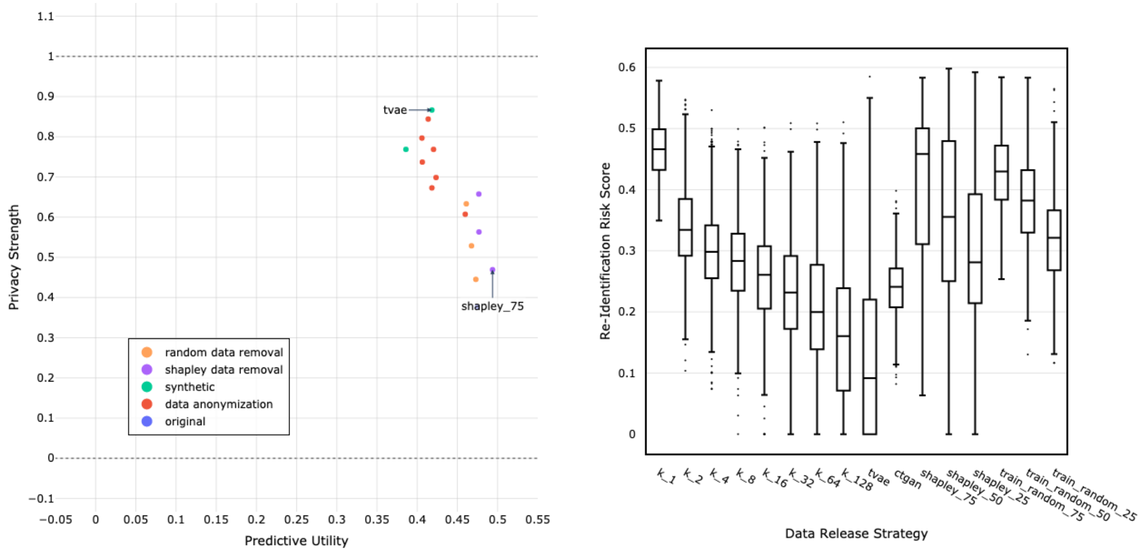
(a) *Accuracy as a function of the percentage of the dataset preserved.*

(b) *Matthews Correlation Coefficient (MCC) as a function of the percentage of the dataset preserved.*

Figure 1.9: *Reproduced results from a data minimization strategy applied to the Folktables dataset using a k-NN classifier. The plots compare Shapley-value-guided removal (descending DSV) with random removal. In both accuracy and MCC, the Shapley-based minimization strategy consistently outperforms the random baseline, especially in early stages, by preserving predictive utility while reducing dataset size.*

scores quantify how easy it is to isolate and infer information of each individual under different release strategies. Figure 1.10 summarizes the results by presenting predictive utility tradeoffs in relation to privacy strength, as well as box plots of individual RRS distributions across all strategies. This visualization allows us to compare the relative performance of each approach along both privacy and utility dimensions. As expected, we observe that both increasing the value of k in the k-anonymity strategy and removing data points (either randomly or based on Shapley value) have a noticeable effect on reducing re-identification scores.

To further understand how re-identification risk varies across individuals, we analyzed the full distribution of RRS values under different levels of k in the k-anonymity strategy. Figure 1.11 compares the distribution of risk scores for $k = 1$, $k = 2$, $k = 16$ and $k = 128$. The original dataset (*i.e.*, $k = 1$) not only exhibits a high average RRS but is also clearly bimodal, with one of the modes concentrated at high risk levels. This indicates



(a) *Privacy-utility tradeoffs for different data release strategies. Each point represents a strategy, with predictive utility on the x-axis and privacy strength (1 - average RRS) on the y-axis.*

(b) *Distribution of individual re-identification risk scores (RRS) across data release strategies. Each box plot summarizes the RRS spread for 4,400 individuals.*

Figure 1.10: *Evaluation of privacy and utility tradeoffs across data release strategies. Synthetic data, anonymization and minimization approaches are compared using both predictive utility and individual re-identification risk scores.*

that a substantial subset of individuals is highly exposed to re-identification attacks. As k increases, the bimodal structure gradually disappears and is replaced by a unimodal distribution that shifts toward lower risk values. When $k = 128$, the maximum setting used in our experiments, we observe a pronounced spike in the lowest bin (average RRS ≈ 0.01), with 908 individuals having risk scores close to zero. This demonstrates the strong protective effect of high k values, though it also reflects reduced informational specificity in the released data.

We also investigated how the distribution of re-identification risk scores (RRS) changes when applying data minimization based on Shapley values. Figure 1.12 shows the RRS distributions for the original dataset and after removing 25%, 50% and 75% of the least valuable data points. Unlike the k-anonymity strategy, the bimodal pattern present in the original data is largely preserved across at different Shapley-based removal percentages. While removing 75% of the data does reduce the concentration of high-risk individuals, earlier removal levels (25% and 50%) have a limited effect on altering

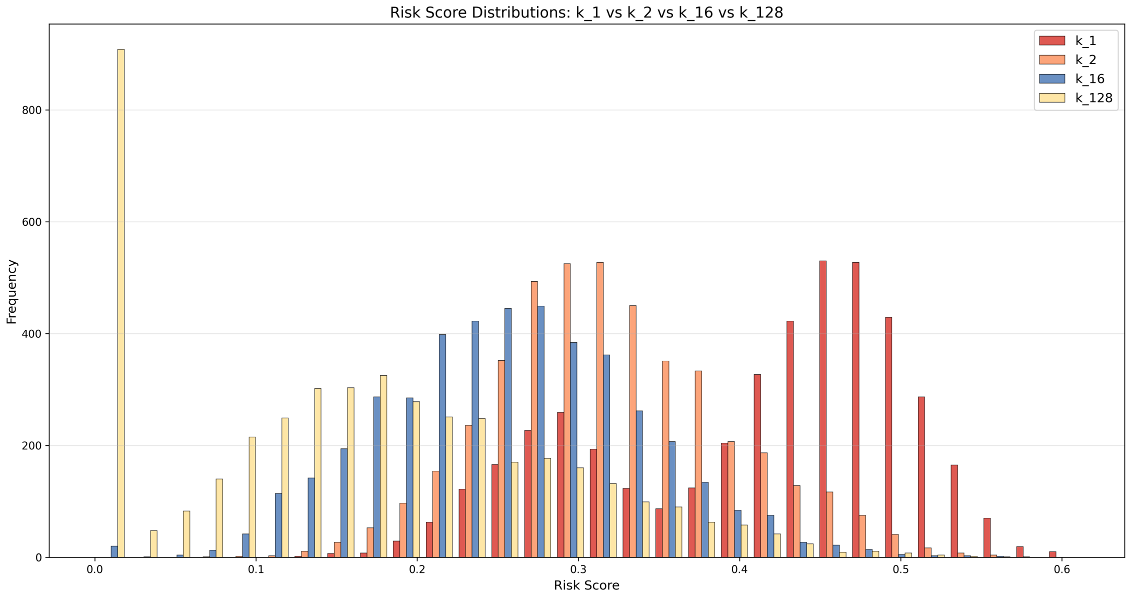


Figure 1.11: *Distribution of re-identification risk scores (RRS) for selected values of k in the k-anonymity strategy. The original dataset (*$k = 1$*) shows a bimodal distribution with many high-risk individuals. As k increases, the distribution becomes unimodal and shifts left. At* $k = 128$*, nearly 900 individuals fall into the first risk bin (average RRS* ≈ 0.01*), illustrating the strong protective effect of generalization.*

the overall distribution of RRS. This indicates that Shapley-based data removal, while beneficial for enhancing or preserving predictive utility, is less effective at eliminating high-risk individual data points unless a substantial portion of the dataset is removed. In fact, removing 25% of the data at random results in a slightly greater reduction in the average RRS compared to Shapley-based removal at the same level, as shown in Figure 1.10b.

We now explore whether some individuals experience better privacy-utility tradeoffs than others under anonymization. Specifically, we examine whether an individual's original Shapley value is predictive of how much their re-identification risk or utility changes when *k*-anonymity is applied. Figure 1.13 shows the average RRS in the original dataset ($k = 1$) grouped by Shapley value bins. We observe that individuals with negative and very low Shapley values tend to have higher risk scores. This suggests that some data points may be both uninformative for modeling and highly exposed to re-identification, making their inclusion particularly questionable.

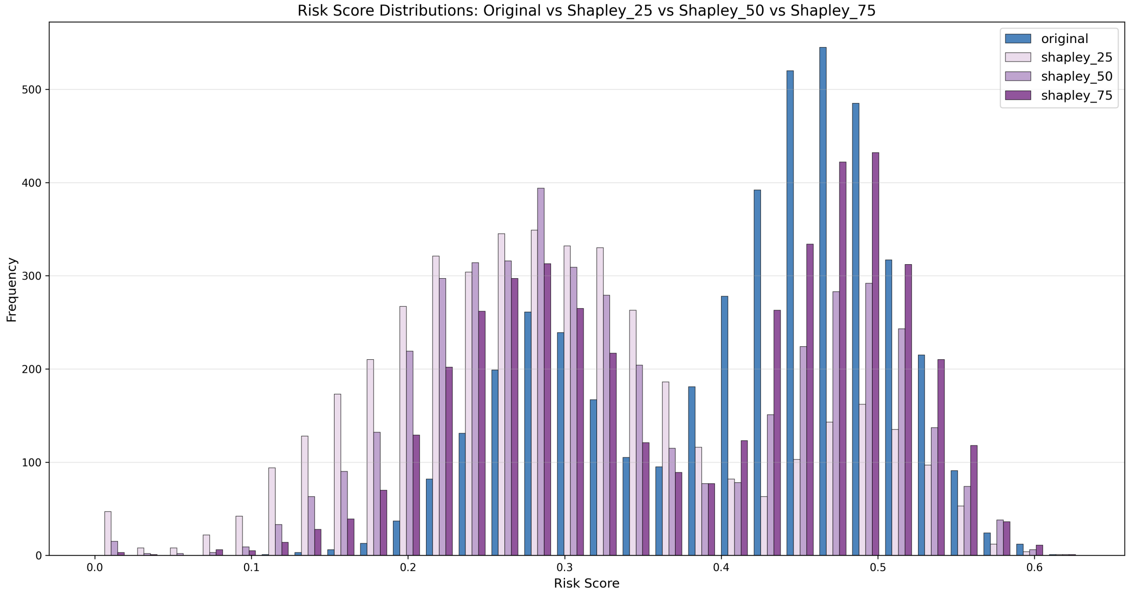


Figure 1.12: *Distribution of re-identification risk scores (RRS) under data Shapley value-based removal. Compared to the original dataset, the bimodal structure remains visible for 25% and 50% removal. Only the 75% removal setting shows a notable shift away from higher risk scores, suggesting limited privacy benefit unless a large portion of the dataset is pruned.*

To further analyze this, we assess how risk changes across Shapley value bins when increasing k in k-anonymity (Figure 1.14). From $k = 4$ and above, individuals with very low Shapley values experience larger reductions in RRS than higher-value individuals. This implies that, in terms of risk alone, individuals whose data is least useful stand to gain the most from anonymization.

Next, we evaluate how data Shapley value rankings change across the same k levels (Figure 1.15). We find that at higher k levels (especially $k = 64$ and $k = 128$), low-value individuals tend to gain value, likely because generalization merges them with more informative records. In contrast, high-value individuals tend to lose Shapley value without any notable reduction in their risk. These observations highlight a potential asymmetry in the privacy-utility tradeoff: while low-value individuals gain in both dimensions, high-value individuals primarily experience utility loss. This offers a more nuanced view of how anonymization affects individual incentives in a non-uniform way.

Finally, we investigate how k-anonymity affects the distribution of Shapley values

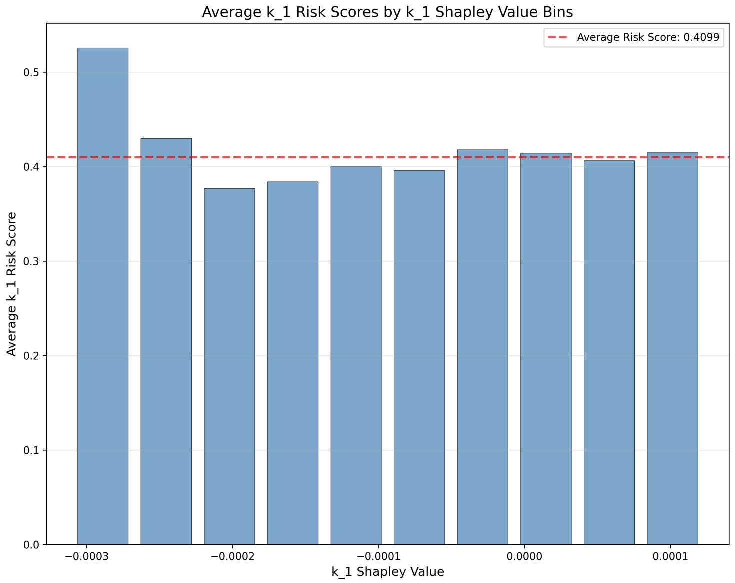


Figure 1.13: *Average re-identification risk score (RRS) in the original dataset ($k = 1$), grouped by Shapley value bins. Higher risk is concentrated among individuals with extremely low utility.*

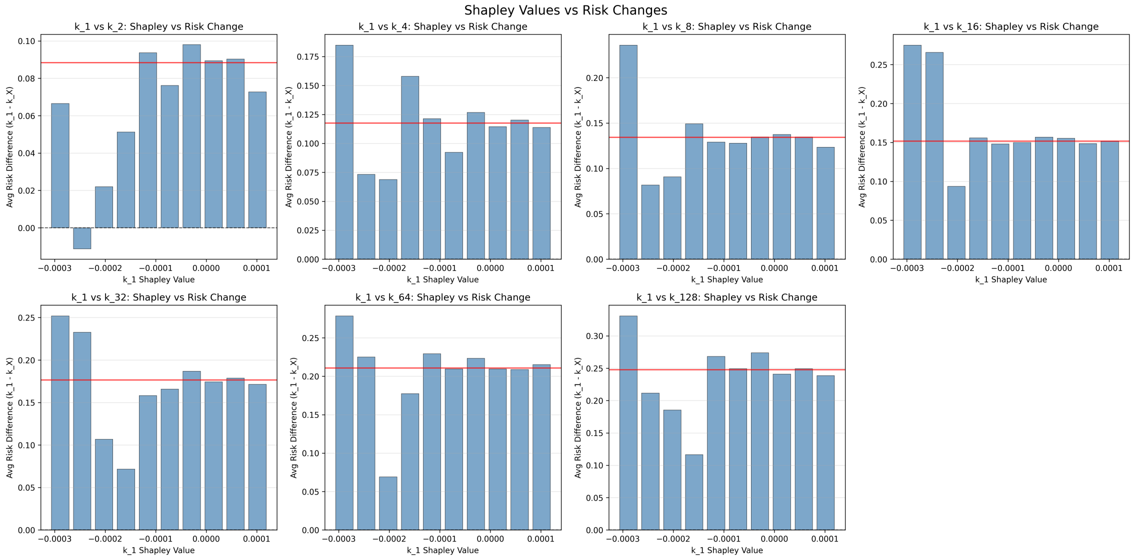


Figure 1.14: *Change in average RRS from $k = 1$ to higher k values across Shapley value bins. Risk decreases more substantially for low-value individuals, particularly at $k \geq 4$.*

across individuals. Specifically, we ask whether increasing k reduces the organization's ability to distinguish between data points in terms of their utility. To answer this, we analyze two metrics: the Spearman correlation between the original ($k = 1$) Shapley values and the ones calculated using the k-anonymity strategy and the range of Shapley

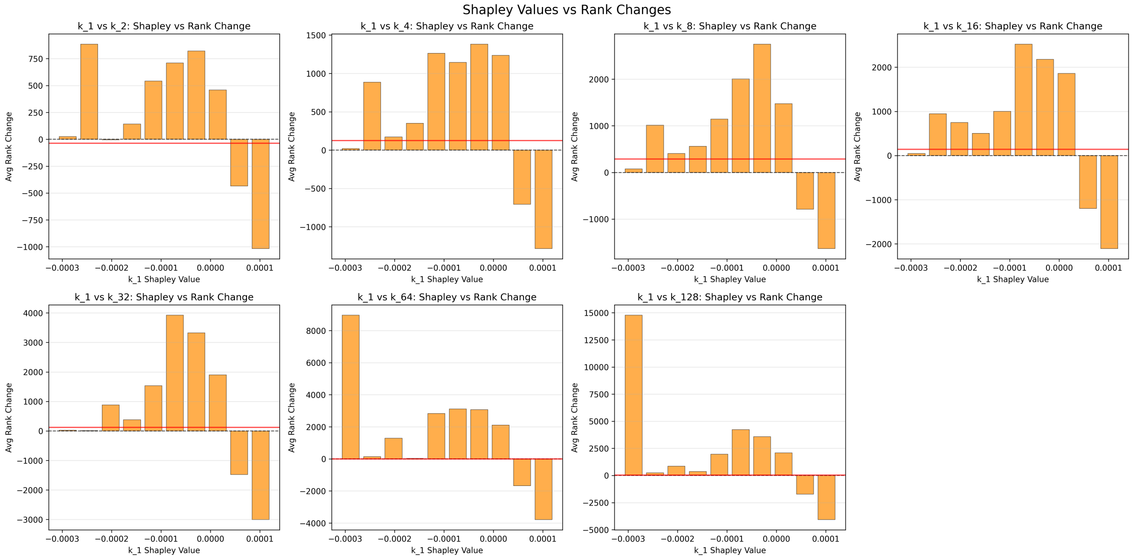


Figure 1.15: *Average change in Shapley rank from $k = 1$ to higher k values, grouped by Shapley value bins. Low-value individuals often gain utility, while high-value individuals typically lose it.*

values (*i.e.*, max minus min) under each k level.

As shown in Figure 1.16, both metrics decline as k increases. More precisely, the Spearman correlation drops steadily from 1.0 at $k = 1$ to just above 0.6 at $k = 128$, indicating a progressive loss of rank ordering in individual utility. Simultaneously, the Shapley value range contracts, suggesting that anonymization compresses the spread of contributions across individuals. From an organizational perspective, this homogenization reduces the granularity with which value can be attributed to individuals. Consequently, as k increases, the incentive to differentiate or reward specific individuals may weaken.

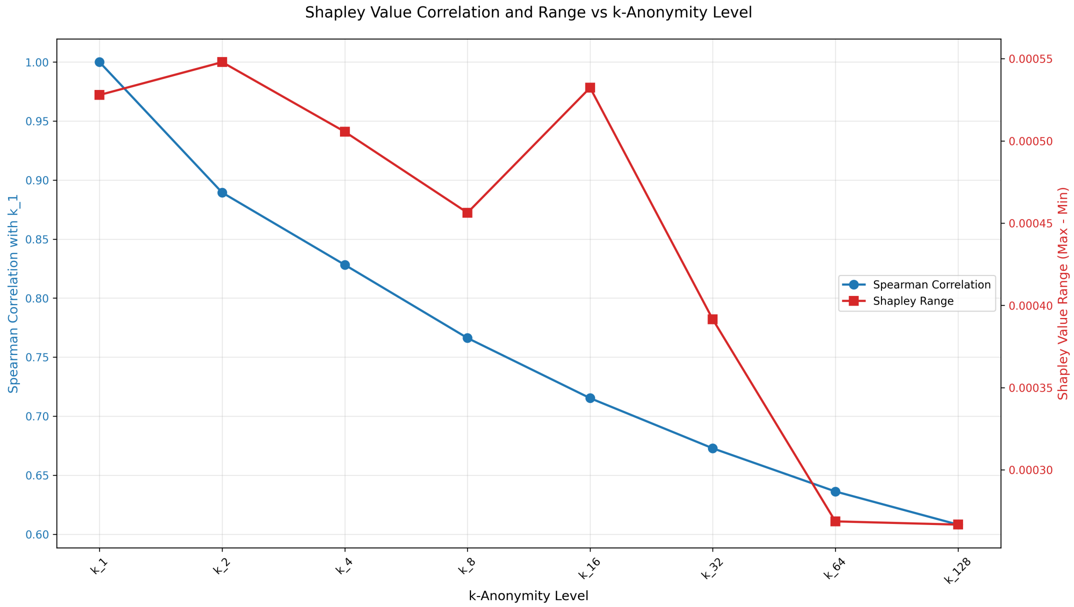


Figure 1.16: *Spearman correlation of Shapley values between the original data ($k = 1$) and anonymized versions (left axis, blue), and range of Shapley values (right axis, red), plotted across increasing k-anonymity levels. As k increases, both correlation and value range decline, indicating reduced differentiation of individual contributions.*

1.9 Appendix : Datasets Description

The following tables summarize the features in each dataset used in this study. Categorical attributes on a nominal scale were encoded using one-hot encoding to represent each category as a binary column, while continuous and ordinal attributes were normalized to a [0,1] range using a min-max scaler.

Table 1.6: *Summary of Attributes in the Adult Income Census (AIC) Dataset*

Attribute	**Description**	**Type**
Age	Age of the individual	Numeric
Workclass	Type of employment	Categorical
Education-num	Education level as numeric	Numeric
Marital-status	Marital status	Categorical
Occupation	Type of occupation	Categorical
Relationship	Relationship status	Categorical
Race	Race of the individual	Categorical
Sex	Gender of the individual	Categorical
Capital-gain	Capital gains	Numeric
Capital-loss	Capital losses	Numeric
Hours-per-week	Hours worked per week	Numeric
Native-country	Country of origin	Categorical
Income (Target)	Income category ($\leq 50K$, $\geq 50K$)	Categorical

Table 1.7: *Summary of Attributes in the Bank Marketing (BM) Dataset*

Attribute	**Description**	**Type**
Age	Age of the client	Numeric
Job	Job title	Categorical
Marital	Marital status	Categorical
Education	Education level	Categorical
Default	Has credit in default?	Categorical
Balance	Average yearly bank balance	Numeric
Housing	Has housing loan?	Categorical
Loan	Has personal loan?	Categorical
Contact	Contact communication type	Categorical
Day	Last contact day	Numeric
Month	Last contact month	Categorical
Duration	Last contact duration (seconds)	Numeric
Subscription (Target)	Subscribed to term deposit (yes/no)	Categorical

Table 1.8: *Summary of Attributes in the Credit Card Default (CCD) Dataset*

Attributes	**Description**	**Type**
Checking Status	Status of the checking account	Categorical
Duration	Duration of the credit in months	Continuous
Credit History	History of the client's credit	Categorical
Purpose	Purpose for which the credit is requested	Categorical
Credit Amount	Amount of credit requested	Continuous
Savings Status	Status of the savings account	Categorical
Employment	Employment status of the client	Categorical
Installment Commitment	Monthly installment commitment	Continuous
Personal Status	Personal status and gender	Categorical
Other Parties	Presence of other parties responsible for credit	Categorical
Residence Since	Number of years in current residence	Continuous
Property Magnitude	Value of property owned	Categorical
Age	Age of the client	Continuous
Other Payment Plans	Other payment plans held by the client	Categorical
Housing	Housing status of the client	Categorical
Existing Credits	Number of existing credits at this bank	Continuous
Job	Type of job held by the client	Categorical
Num Dependents	Number of dependents	Continuous
Own Telephone	Availability of a telephone	Categorical
Foreign Worker	Whether the client is a foreign worker	Categorical
Class (Target)	Classification of credit risk	Categorical

1.10 Appendix : Predictive Utility Evaluation Details

Classification Task Benchmark

We decomposed data utility into two metrics: data fidelity and predictive utility. For predictive utility, we used the performance on a classification task. Tables 1.9, 1.10 and 1.11 present the classification task benchmarks for respectively the AIC, BM and CCD datasets.

We began by setting a benchmark and evaluating a diverse set of models to assess their classification performance on these datasets. We evaluated the performance on the training, validation and test sets using several metrics: the F1 score on the minority class, the Matthews Correlation Coefficient (MCC) and the loss (calculated using the confidence score). The confidence score of the predictions is also reported.

For the k-Nearest Neighbors (kNN) classifier, we experimented with various values of k. We selected $k = 10$ for the AIC dataset and $k = 5$ for both the BM and CCD datasets for the rest of the experiments in the paper, as these values yielded the best validation performance (using MCC). Additionally, we fine-tuned parametric models such as XGBoost and neural networks to achieve optimal performance to find the best hyperparameters.

Evaluation Metrics We include the formulas for the F1 score and MCC here to ensure clarity, as some readers may be unfamiliar with these metrics.

The F1 score, specifically calculated on the minority class, is defined as:

$$\mathrm{F1} = 2 \times \frac{\mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}, \tag{1.7}$$

in which Precision and Recall are given by:

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \quad \mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}. \tag{1.8}$$

Here, TP, FP, and FN represent respectively the true positives (TP), false positives (FP), and false negatives (FN) for the minority class.

The Matthews Correlation Coefficient (MCC) is defined as:

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}}, \tag{1.9}$$

in which TN represents the number of true negatives. The MCC ranges from -1 to 1, in which 1 represents a perfect prediction, 0 indicates performance no better than random chance, and -1 corresponds to total disagreement between prediction and observation. MCC is preferred in our evaluation because it considers all four quadrants of the confusion matrix and provides a balanced measure, ensuring no preference toward either class, even when the classes differ significantly in size.

Mitigating Class Imbalance Effects on Data Shapley Values

A challenge in using DSV arises when dealing with imbalanced datasets, as is the case with our three datasets. This issue is a recognized challenge in data valuation, as noted by Tang et al., 2021a, and is especially pronounced for kNN, in which it arises from the reliance on a specific utility function for the efficiency property of DSV. Following Jia, Dao, Wang, Hubis, Gurel, et al., 2019, we define the utility function as:

$$v(S) = \frac{1}{K} \sum_{k=1}^{\min\{K,|S|\}} \mathbf{1}\left[y_{\alpha_k(S)} = y_{\text{test}}\right], \tag{1.10}$$

in which K represents the number of nearest neighbors, S is a subset of the data, $y_{\alpha_k(S)}$ is the predicted label of the k-th nearest neighbor in S, and y_{test} is the true label. This utility function ensures that the contribution of each subset S is based on the proportion of correct predictions within its K-nearest neighbors, providing a more direct alignment with the kNN classification accuracy.

A common approach to handle imbalanced datasets is to rebalance them, but this is less effective with kNN because the model relies on local neighborhoods for predictions. Downsampling the training set also impacts the data fidelity metric monitored using the Wasserstein distance. Instead, we employed a weighting scheme in the valuation of the validation set. Since Shapley Values are calculated using each point in the validation set and typically averaged to obtain the value of each training point, an imbalanced validation set causes the minority class to have less influence on the valuation. To mitigate this, we tested various weights on the majority class points in the validation set.

Figures 1.17, 1.18 and 1.19 display the results of our experiments for respectively the AIC, BM, and CCD datasets. In our weighting scheme, *kNN-DSV (1.0)* indicates a balance ratio of 1.0, meaning the minority and majority classes have the same frequency in the validation set. *kNN-DSV (2.0)* means the majority class has twice the frequency of the minority class, corresponding to a balance ratio of 2.0, and so on.

Let c_1 be the minority class and c_2 the majority class, with N_{c_1} and N_{c_2} as their respective counts. The weights are set to 1.0 for the minority class and $\frac{N_{c_1} \times \text{balance\_ratio}}{N_{c_2}}$

for the majority class. This weighting scheme is equivalent to averaging over all downsampled validation sets with c_2 downsampled to $N_{c_1} \times$ balance_ratio.

Our experiments indicate that without any weighting scheme (*i.e.*, maximum balance ratio corresponding to average weighting), the loss decreases as we remove points, which aligns with existing literature. However, imbalance-robust metrics such as the F1 score and MCC decrease rapidly. Notably, we identified optimal balance ratios for each dataset that yielded better results: a balance ratio of 2.0 for the AIC and BM datasets, and 1.5 for the CCD dataset. We did not extensively optimize these ratios, as it was beyond the scope of our study.

More importantly, the Wasserstein distance, representing data fidelity, deteriorates across all balance ratios, indicating a cost in data fidelity when removing data using the data shapley value strategy.

Alternative Approach Using DSV Monte Carlo Approximation

To offer an alternative approach for organizations that may prefer high-performance models like XGBoost over kNN, we experimented with computing the DSV using the MCC directly as the utility function. As mentioned earlier, this approach requires the use of Monte Carlo Shapley Value approximation.

The results, presented in Figure 1.20, show that although the MCC values are initially higher than those obtained using kNN, the data removal results were less consistent across datasets. Only the AIC dataset demonstrated a positive effect of data Shapley value removal compared to random removal. Since we aimed for a consistent method across all datasets, this was another reason why we finally chose kNN for calculating data Shapley values.

We also wish to reiterate that computing DSV using Monte Carlo approximation is computationally intensive. For the AIC dataset, it required approximately 30 hours on the Alliance Canada High-Performance Computing (HPC) cluster, utilizing 40-core machines (processing time for the BM dataset was similar). This makes the approach less practical for large-scale applications that would require DSV in near real-time.

Table 1.9: *AIC Classification Task Benchmark*

Model	**Val Loss**	**Val Conf**	**Val F1**	**Val MCC**	**Test Loss**	**Test Conf**	**Test F1**	**Test MCC**	**Train Time (s)**
kNN (k=1)	0.21	1.00	0.58	0.44	0.21	1.00	0.56	0.43	0.0019
kNN (k=5)	0.21	0.87	0.61	0.51	0.21	0.88	0.61	0.50	0.0015
kNN (k=10)	0.22	0.85	0.65	0.54	0.22	0.86	0.63	0.52	0.0014
kNN (k=15)	0.22	0.85	0.62	0.52	0.22	0.85	0.62	0.52	0.0017
kNN (k=20)	0.22	0.85	0.63	0.53	0.22	0.85	0.63	0.52	0.0012
kNN (k=25)	0.22	0.84	0.62	0.52	0.22	0.84	0.61	0.51	0.0014
Random Forest	0.20	0.89	0.65	0.54	0.19	0.89	0.65	0.54	1.3400
Gradient Boosting	0.20	0.86	0.67	0.59	0.20	0.86	0.69	0.61	2.3899
XGBoost	0.18	0.87	0.69	0.61	0.18	0.88	0.71	0.63	0.2245
Logistic Regression	0.22	0.85	0.64	0.55	0.21	0.85	0.65	0.55	0.1972
Neural Network	0.20	0.87	0.67	0.57	0.20	0.87	0.66	0.57	16.0851

Table 1.10: *BM Classification Task Benchmark*

Model	**Val Loss**	**Val Conf**	**Val F1**	**Val MCC**	**Test Loss**	**Test Conf**	**Test F1**	**Test MCC**	**Train Time (s)**
kNN (k=1)	0.13	1.00	0.41	0.34	0.13	1.00	0.39	0.32	0.0014
kNN (k=5)	0.14	0.93	0.36	0.34	0.14	0.93	0.34	0.32	0.0012
kNN (k=10)	0.14	0.92	0.29	0.31	0.14	0.93	0.28	0.31	0.0013
kNN (k=15)	0.14	0.92	0.28	0.30	0.14	0.92	0.27	0.31	0.0013
kNN (k=20)	0.14	0.92	0.26	0.30	0.14	0.92	0.24	0.29	0.0014
kNN (k=25)	0.14	0.92	0.26	0.29	0.14	0.92	0.23	0.27	0.0012
Random Forest	0.13	0.90	0.52	0.48	0.13	0.90	0.50	0.47	2.1951
Gradient Boosting	0.13	0.91	0.52	0.49	0.13	0.91	0.51	0.47	3.6068
XGBoost	0.12	0.92	0.57	0.52	0.12	0.92	0.55	0.51	0.3932
Logistic Regression	0.14	0.91	0.45	0.42	0.14	0.91	0.43	0.41	0.1512
Neural Network	0.13	0.91	0.59	0.53	0.13	0.91	0.56	0.50	112.1338

Table 1.11: *CCD Classification Task Benchmark*

Model	**Val Loss**	**Val Conf**	**Val F1**	**Val MCC**	**Test Loss**	**Test Conf**	**Test F1**	**Test MCC**	**Train Time (s)**
kNN (k=1)	0.31	1.00	0.43	0.25	0.28	1.00	0.54	0.34	0.0004
kNN (k=5)	0.34	0.81	0.42	0.33	0.32	0.80	0.39	0.22	0.0003
kNN (k=10)	0.36	0.77	0.38	0.22	0.35	0.75	0.43	0.24	0.0002
kNN (k=15)	0.36	0.76	0.23	0.20	0.36	0.75	0.33	0.18	0.0002
kNN (k=20)	0.36	0.75	0.25	0.22	0.37	0.73	0.32	0.15	0.0002
kNN (k=25)	0.36	0.75	0.14	0.18	0.37	0.73	0.23	0.12	0.0002
Random Forest	0.35	0.74	0.41	0.31	0.36	0.70	0.45	0.29	0.0797
Gradient Boosting	0.32	0.79	0.47	0.37	0.33	0.75	0.54	0.37	0.1072
XGBoost	0.31	0.80	0.51	0.37	0.31	0.77	0.59	0.40	0.0471
Logistic Regression	0.32	0.77	0.49	0.31	0.33	0.76	0.56	0.38	0.0192
Neural Network	0.32	0.78	0.54	0.35	0.34	0.78	0.53	0.32	1.0750

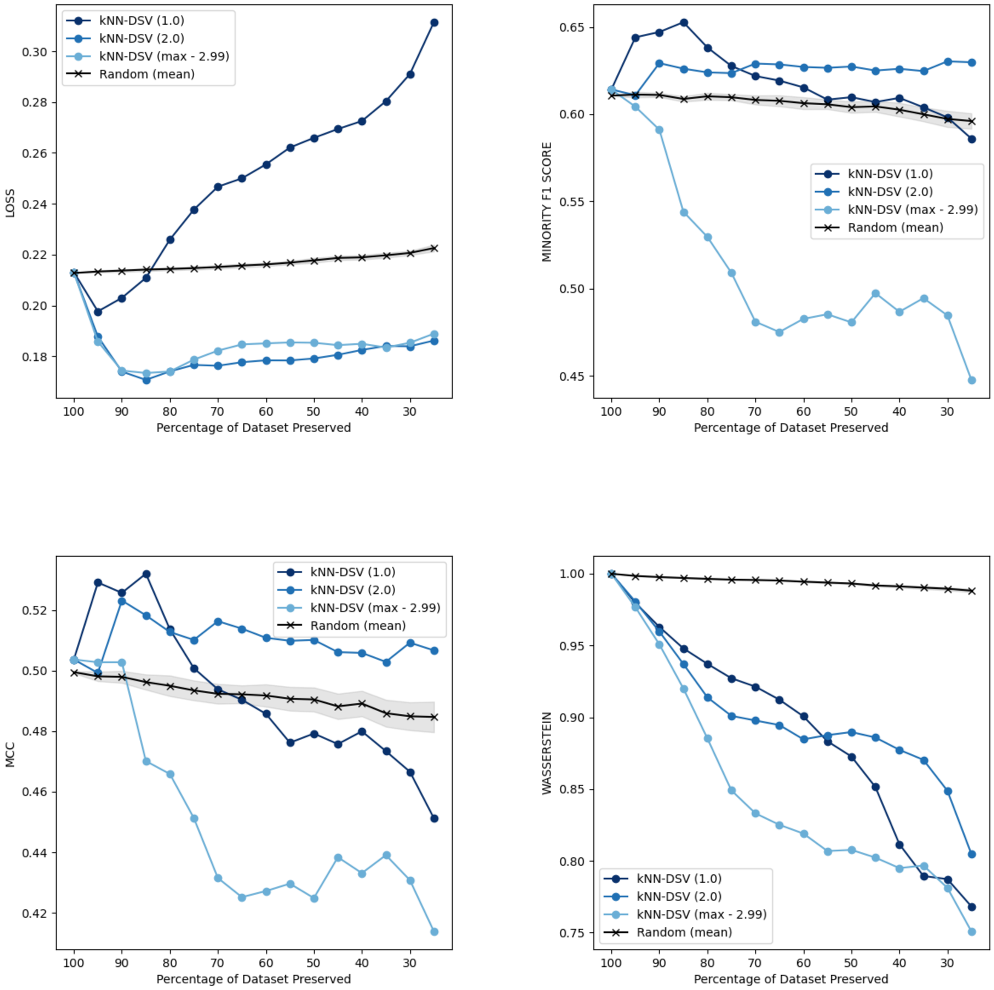


Figure 1.17: *Data downsampling effect on data sahpley removal for AIC (results are reported on test set but balance ratio was chosen on validation set).*

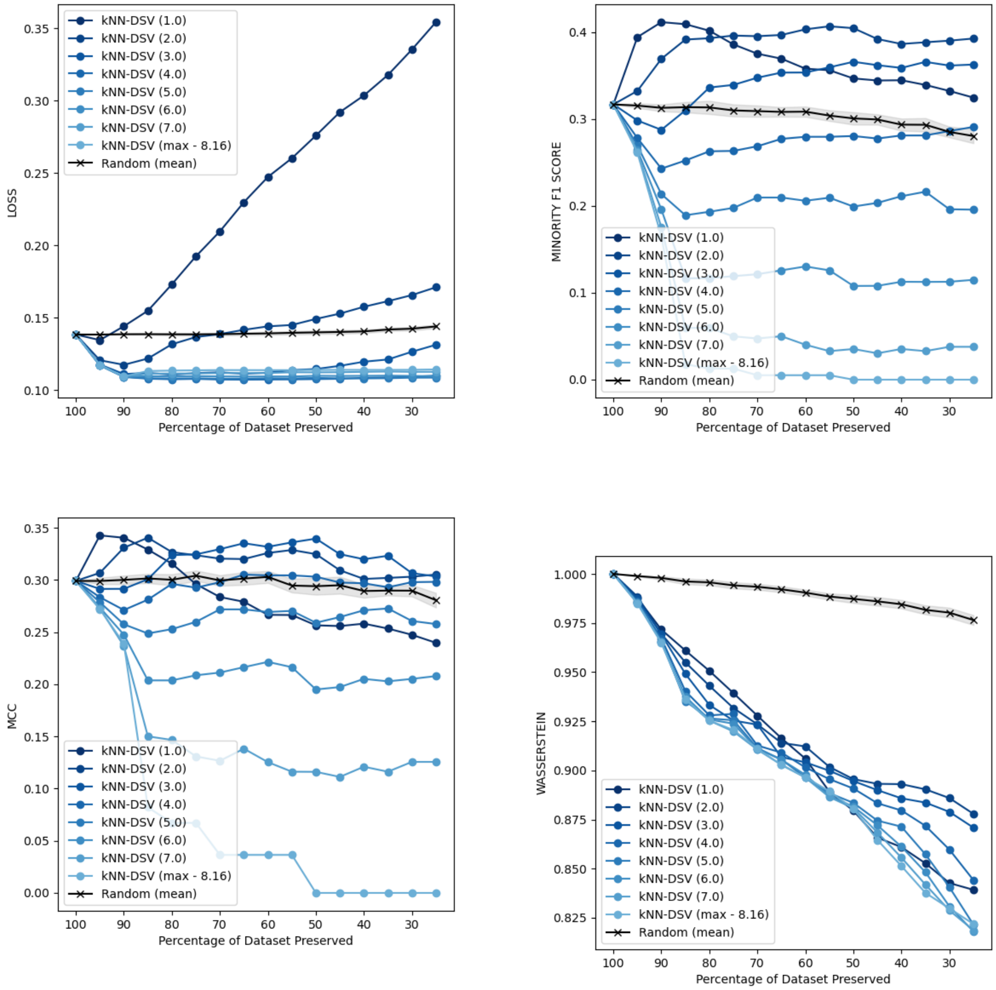


Figure 1.18: *Data downsampling effect on data sahpley removal for BM (results are reported on test set but balance ratio was chosen on validation set).*

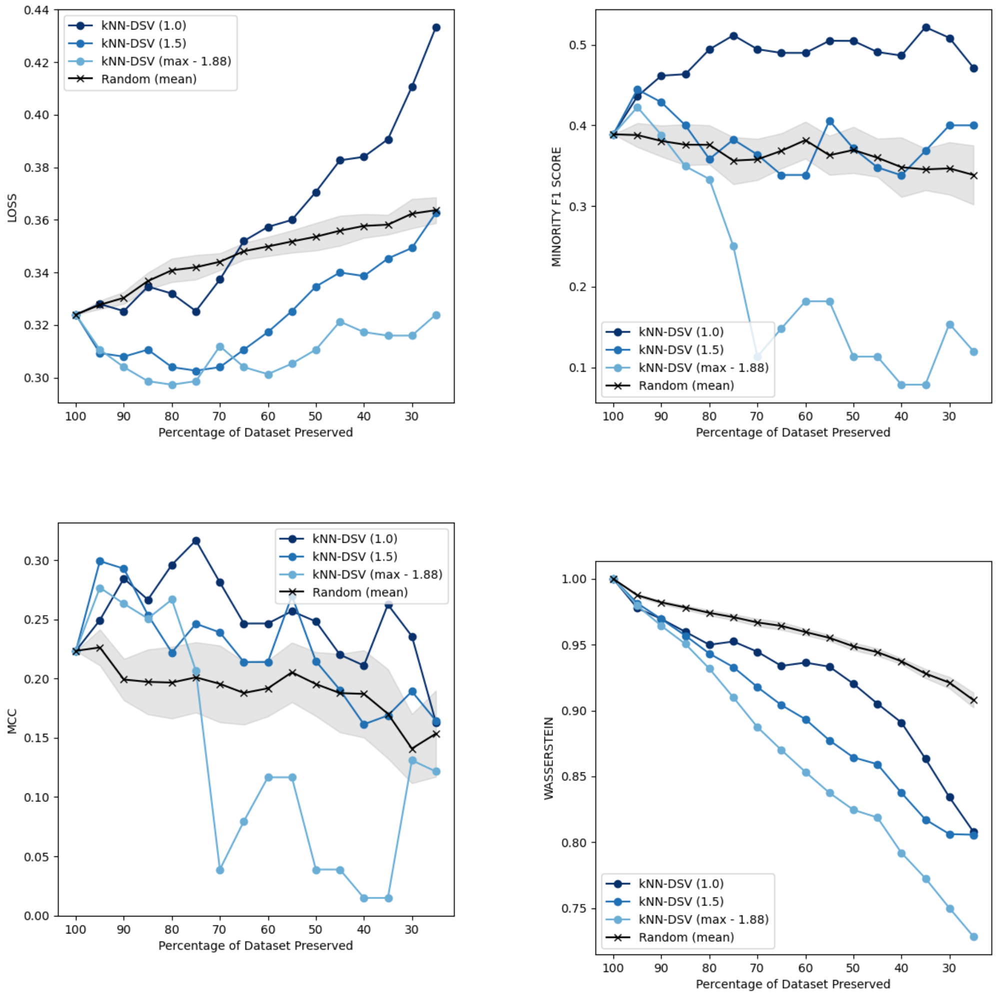
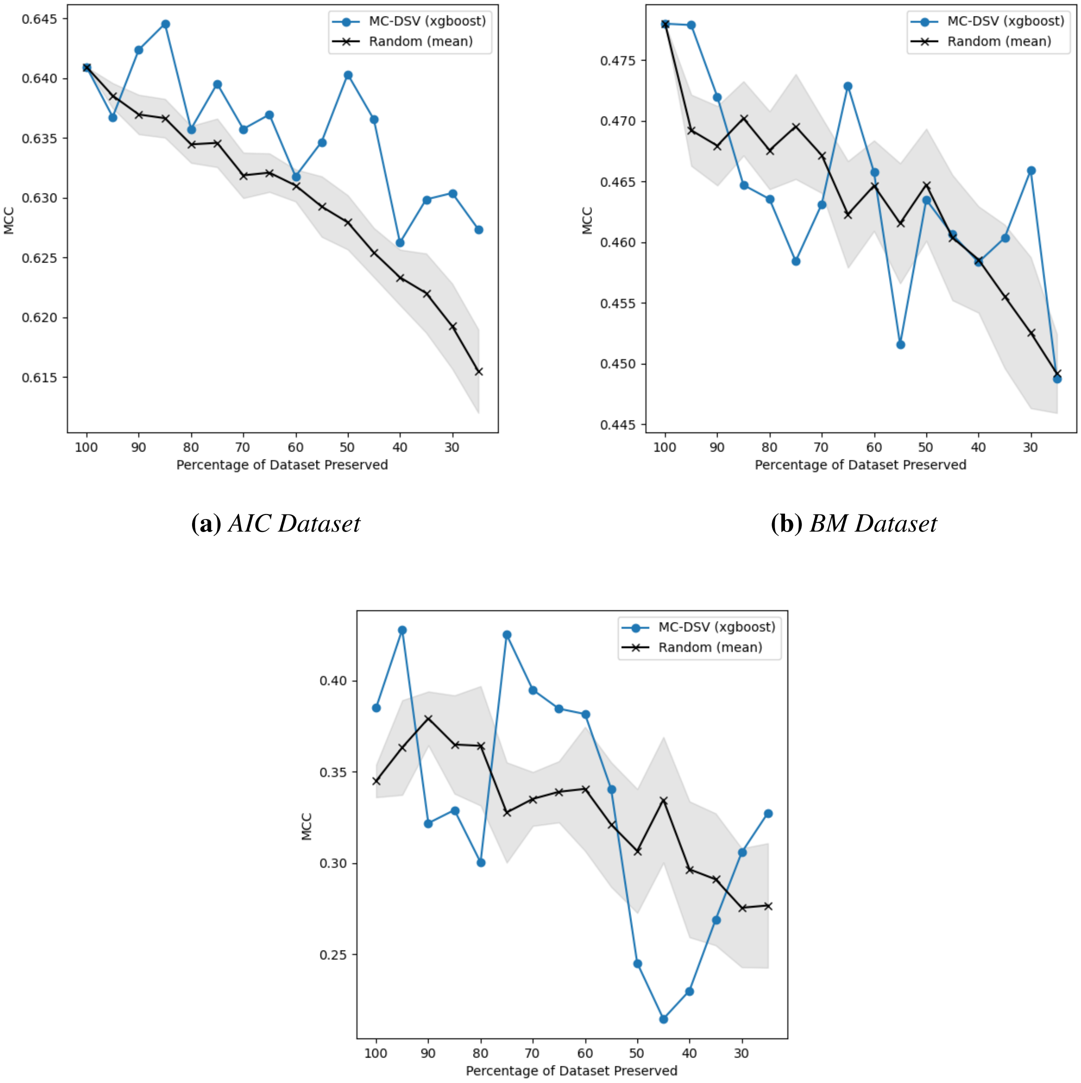


Figure 1.19: *Data downsampling effect on data sahpley removal for CCD (results are reported on test set but balance ratio was chosen on validation set).*

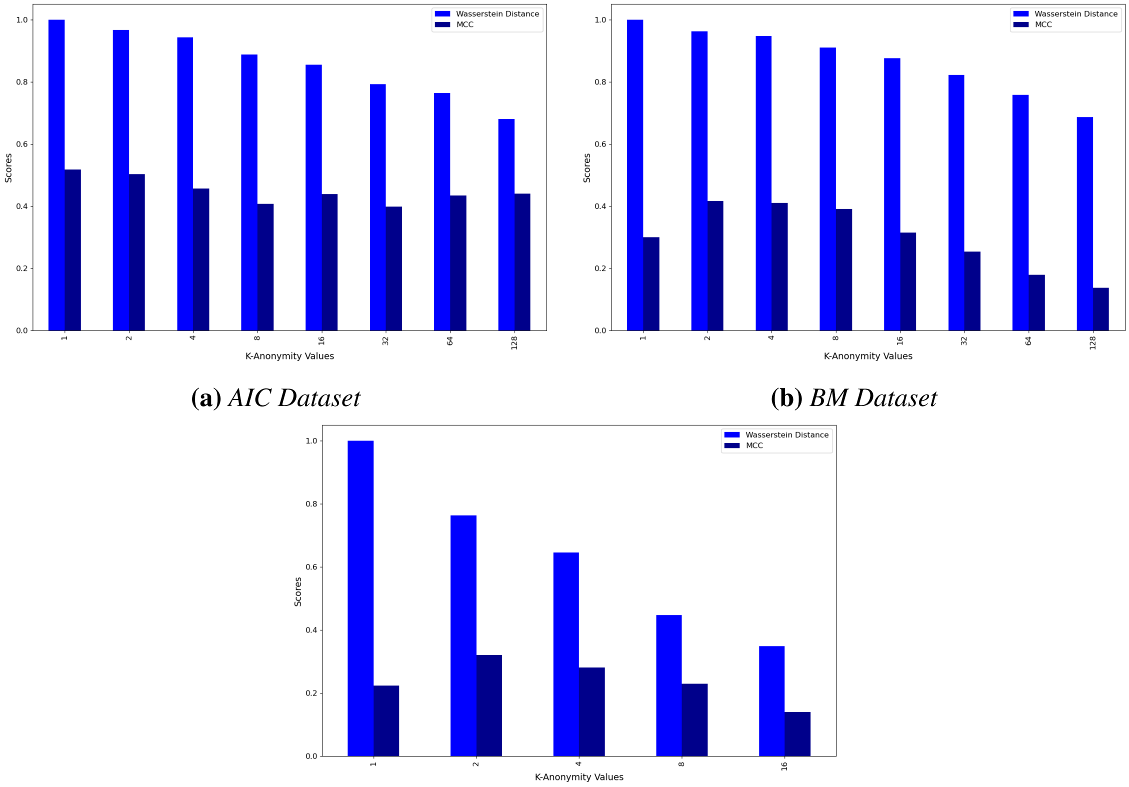


(a) *AIC Dataset*

(b) *BM Dataset*

(c) *CCD Dataset*

Figure 1.20: *Monte Carlo Data Shapley Value with XGBoost and MCC utility function.*



(a) *AIC Dataset*

(b) *BM Dataset*

(c) *CCD Dataset*

Figure 1.21: *kNN classifier average loss and Wasserstein distance after using k-anonymity strategies with multiple values of k.*

Chapter 2

WaKA: Data Attribution using K-Nearest Neighbors and Membership Privacy Principles

Abstract [1]

In this paper, we introduce WaKA (*Wasserstein K-nearest neighbors Attribution*), a novel attribution method that leverages principles from the LiRA (*Likelihood Ratio Attack*) framework and *k*-nearest neighbors classifiers (*k*-NN). WaKA efficiently measures the contribution of individual data points to the model's loss distribution, analyzing every possible *k*-NN that can be constructed using the training set, without requiring to sample subsets of the training set. WaKA is versatile and can be used *a posteriori* as a membership inference attack (MIA) to assess privacy risks or *a priori* for privacy influence measurement and data valuation. Thus, WaKA can be seen as bridging the gap between data attribution and membership inference attack (MIA) by providing a unified framework to distinguish between a data point's value and its privacy risk. For

[1] This essay was accepted to the 2025 Privacy Enhancing Technologies Symposium (PETS) and will appear in the Proceedings on Privacy Enhancing Technologies (PoPETs). First author: Patrick Mesana (HEC Montréal); co-authors: Clément Bénesse (UQAM), Hadrien Lautraite (UQAM), Gilles Caporossi (HEC Montréal), and Sébastien Gambs (UQAM).

instance, we have shown that self-attribution values are more strongly correlated with the attack success rate than the contribution of a point to the model generalization. WaKA's different usages were also evaluated across diverse real-world datasets, demonstrating performance very close to LiRA when used as an MIA on *k*-NN classifiers, but with greater computational efficiency. Additionally, WaKA shows greater robustness than Shapley Values for data minimization tasks (removal or addition) on imbalanced datasets.

2.1 Introduction

Data attribution methods have been developed originally to measure the contribution of individual data points in a training set to a model's output. These methods can serve different purposes depending on the context. One key application is data valuation, in which the objective is to quantify the "value" of each data point with respect to its impact on the model's ability to generalize. For example, Data Shapley Value (DSV), introduced in Ghorbani and Zou, 2019 and Jia, Dao, Wang, Hubis, Gurel, et al., 2019, is grounded in the game-theoretic Shapley Value framework and is often used for tasks such as data minimization via summarization, in which the objective is to remove a large fraction of data points while ensuring a high generalization performance of the model This approach aligns well with Article 5 of the General Data Protection Regulation (GDPR, European Parliament and Council, 2016), which emphasizes that personal data should be "adequate, relevant, and limited to what is necessary" in relation to the purposes for which they are processed.

As a motivating scenario, consider for instance an organization that collects a dataset and then trains a machine learning model over it, exposing its functionality via an API to monetize the access to the predictions of the model (*e.g.*, for classifying movie reviews or categorizing images). Data attribution can be used by this organization from two perspectives: data valuation and privacy with Figure 2.1 illustrating both viewpoints. On one hand, data valuation can help estimate the contribution of each data point to the model, guiding decisions on redistributing a share of the model's value to individuals who provided data points. This valuation, which aligns with Shapley value, also helps the organization determine which data points bring no value or even negative value, potentially avoiding unnecessary costs by not acquiring these data points.

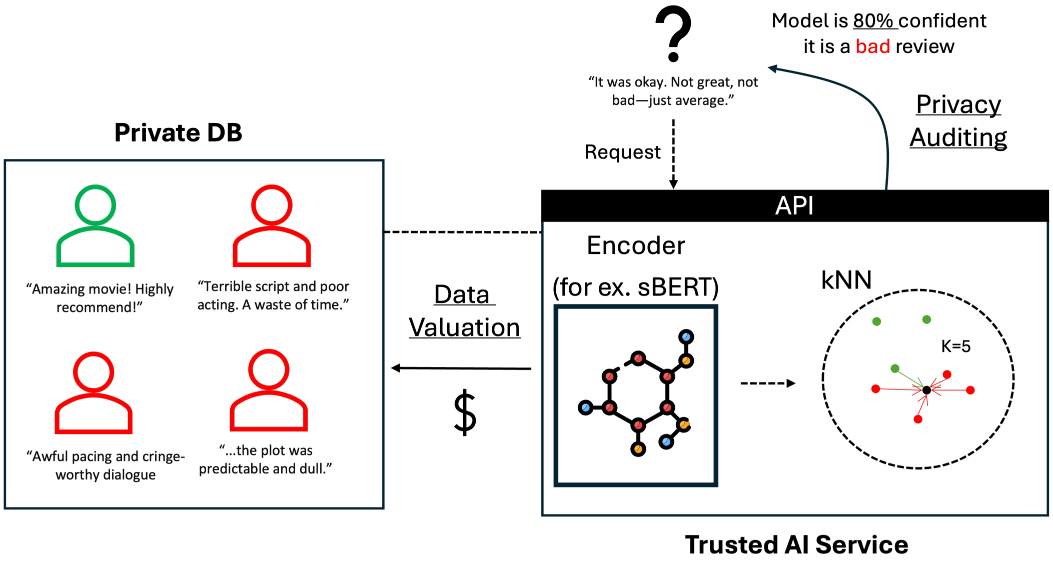

?
Model is 80% confident
it is a bad review
"It was okay. Not great, not bad—just average."
Request
Privacy Auditing
Private DB
"Amazing movie! Highly recommend!"
"Terrible script and poor acting. A waste of time."
"Awful pacing and cringe-worthy dialogue"
"...the plot was predictable and dull."
API
Encoder
(for ex. sBERT)
kNN
K=5
Data Valuation
$
Trusted AI Service

Figure 2.1: *Illustration of data attribution in a movie review classification scenario, highlighting the dual perspectives of data valuation (estimating data point value) and data privacy (measuring potential information leakage).*

On the other hand, from the privacy perspective, the organization might be concerned with how much information about data points could potentially be leaked through the API, which is akin to measuring the privacy risk of data leakage. This is closely related to membership inference attacks (MIAs), in which an attacker aims to determine whether a particular data point was part of the model's training set. While data valuation and privacy concerns are often related, they are usually addressed through very different methods. To solve this issue, we introduce WaKA (for 1-Wasserstein k-NN Attribution), which provides a unified framework for addressing both aspects using k-nearest neighbors (k-NN) models.

While being simple in their design, k-NN models offer a clear and intuitive way to perform data attribution. In particular, they are recognized as an instance-based explainable method (Molnar, 2020), meaning its predictions are directly influenced by the training data without an explicit model abstraction. However, a well-known limitation is that they do not perform well on high-dimensional data such as textual documents or images due to the curse of dimensionality. In such cases, the distance between points becomes nearly identical, making the identification of neighbors ineffective (Friedman, 1997). To overcome this challenge and apply them to datasets like images and textual data, we first employ a neural network to learn embeddings suitable for *k*-NN classification. These learned embeddings capture meaningful representations of the data, making *k*-NN effective even in high-dimensional settings. This approach is also commonly used in the data valuation literature in which case it is usually referred to as "surrogate models"(Jia et al., 2021). Additionally, while the majority of research on data attribution and membership inference attacks has been centered on neural networks, *k*-NN-based studies are highly relevant in practice. In particular, in industrial applications, such as retrieval-augmented generation (RAG) pipelines, *k*-NN with embeddings is widely adopted as the use of learned representations mitigates the high-dimensionality challenges typically associated with nearest neighbor search. Moreover, Yadav and Chaudhuri demonstrated that interpretations of *k*-NN models using embeddings are comparable to a softmax layer in neural networks, reinforcing their relevance to modern machine learning workflows.

The term "value" in data valuation is semantically charged, as it suggests an intrinsic worth of data points. In this context, the term is grounded in the Shapley Value axioms, which provide a unique way of attributing the contribution of each data point to the model's generalization performance. Data valuation methods such as DSV seek to uncover the intrinsic properties of data in relation to the category of model being used, whether it be a type of neural network, a *k*-NN classifier or another machine learning model.

In contrast, membership inference (Shokri et al., 2017a; Yeom et al., 2018) aims at determining whether a given data point was part of a specific model's training dataset. For instance, LiRA (Carlini, Chien, et al., 2022a) is a state-of-the-art approach for performing membership inference attacks (MIAs), which is based on the Likelihood Ratio Test

(LRT), a statistical test that compares the likelihood of a model's prediction for a given data point when trained with and without this point. To realize this, LiRA requires the training of shadow models via sampling and provides a membership score for each point in the training set, facilitating a more detailed attribution analysis. Similarly to LiRA, WaKA assumes that the adversary has access to the underlying data distribution, enabling probability mass function (PMF) computation. This assumption is fundamental to our framework but also to LiRA.

Related work. More recently, the intersection between data attribution and membership privacy has garnered significant attention as evidenced by a growing body of related literature (G. Cohen and Giryes, 2024; Duddu et al., 2021; Ye et al., 2023). One key concept is "self-influence", which has been investigated in differentiable models to measure the extent to which a data point influences its own prediction. This concept is particularly relevant in MIAs, as high self-influence scores often correlate with increased privacy risks (G. Cohen and Giryes, 2024). Self-influence is computed using influence functions and measures how much the loss changes for a data point when it is upweighted. In particular, this method has been used for capturing how a point's inclusion can lead to memorization (F. Liu et al., 2021), which is in turn relates to privacy vulnerability. Throughout the paper, we adopt a similar notion, which we refer to as "self-attribution", whose objective is to address the question "To what extent my data contribute to my own outcome?". More precisely, it can be quantified by the marginal contribution of a point to the model's prediction on that same point.

Beyond self-influence, several studies have analyzed how the model's performance relates to privacy risk. Prior work has shown that overfitting exacerbates MIAs (Shokri et al., 2017a; Yeom et al., 2018), but recent findings indicate that privacy leakage can occur even in well-generalized models (Carlini, Chien, et al., 2022a; Nasr et al., 2019). Additionally, per-record memorization suggests that not all training samples contribute equally to privacy risks—some points are more prone to memorization and thus more vulnerable to MIAs (Carlini et al., 2019; Feldman and Zhang, 2020). This suggests a complex relationship between the value of a data point and its privacy risk. Recent work on Leave-One-Out Distinguishability (LOOD, Ye et al., 2023) provides a unified perspective on data attribution and privacy auditing by quantifying the statistical distance between a model's outputs with and without a specific training point. This approach highlights the strong connection between influence, memorization and privacy leakage. While LOOD has been shown to predict MIA success and can serve as a privacy auditing tool, existing methods primarily focus on neural network-based models.

Summary of contributions. Our main contributions can be summarized as follows.

- We introduce the 1-Wasserstein *k*-NN Attribution (WaKA), a novel approach that leverages the Wasserstein distance from Optimal Transport, which has not been previously used in either membership inference attacks (MIAs) or data valuation. Unlike LiRA, which relies on hypothesis testing, WaKA accounts for the mass that a point moves positively or negatively. WaKA serves a dual purpose by providing a principled framework for both privacy risk assessment and data valuation, functioning as a general attribution method for data valuation while also offering privacy insights through self-attribution. More precisely, it can be adapted into t-WaKA to effectively perform membership inference attacks.

- In our experiments, we compare the performance of DSV and WaKA. These experiments have been conducted on six diverse datasets—two tabular datasets (Adult and Bank), two textual datasets (IMDB and Yelp) and two image datasets (CIFAR-10 and CelebA) —to demonstrate the versatility and robustness of our method. The evaluation of these scenarios focuses on two key aspects: utility, through two data minimization tasks as well as privacy, by measuring the attack success rate (ASR) on all training points.

- We explore the "onion effect" (Carlini, Jagielski, et al., 2022), a phenomenon observed previously in neural networks in which removing data points incrementally reveals deeper layers of vulnerable privacy points in the sense that these points suffer from a higher ASR after the removal. To investigate this effect, we have replicated some of experiments of the original paper by eliminating

10% of the training set using attribution methods, followed by a reassessment of privacy scores. More precisely, we have analyzed the relationship between privacy influences and WaKA influences, showing that they are correlated and can be used to predict, *a priori*, whether removing a data point will impact the ASR on other points.

- Finally, we have also conducted experiments using t-WaKA as an MIA on specific training points for k-NN models. t-WaKA displays a similar performance as LiRA but uses significantly less resources as it relies on a single reusable k-NN model trained on the entire dataset, thus avoiding the need for shadow models. More precisely, once this k-NN model is trained, t-WaKA has a computational complexity of $O(\log N)$ for attacking a specific point, which is much faster than LiRA for k-NN.

Outline. First in Section 2.2, we provide a brief overview of attribution methods (Leave one Out and Data Shapeley Value) for k-NN models as well as LiRA, before introducing the details of the Wasserstein k-NN Attribution (WaKA) method in Section 2.3. Afterwards, in Section 2.4, we conduct an extensive evaluation of WaKA as a new attribution method and t-WaKA for assessing the success of MIAs before finally concluding with a discussion in Section 3.5.

2.2 Background

In this section, we review the background notions necessary to the understanding of our work, namely the main existing attribution methods for k-NNs as well as the LiRA framework for conducting membership inference attacks.

2.2.1 Attribution Methods for k-NNs

Leave-One-Out (LOO) attribution methods are commonly used to assess the contribution of individual data points to a model's performance by examining the effect of removing a point from the training set. Although they can theoretically be applied with respect to any predictive target, they are typically computed on a test set D_{test} with the objective of measuring the contribution to the generalization performance. In this context, a positive contribution means reducing the generalization loss or increasing utility (Jia et al., 2021).

Consider a training set $D = \{z_i\}_{i=1}^N$, in which each z_i represents a feature-label pair (x_i, y_i). For k-NN models, the loss function ℓ for any data point z is generally defined as the fraction of neighbors that do not share the same label:

$$\ell(z; D, k) = 1 - \frac{1}{k} \sum_{j=1}^{k} \mathbf{1}(y_{\alpha_j} = y), \tag{2.1}$$

in which α_j is the index of the j-th closest neighbor to z in the training set D, and y is the true label of point z. This is referred to as a "loss" because it measures the error introduced by the model's prediction, which also quantifies the model's disagreement with the ground truth. The utility function $U(z_t; D)$ can be expressed as:

$$U(z_t; D) = 1 - \ell(z_t; D) = \frac{1}{k} \sum_{j=1}^{k} \mathbf{1}(y_{\alpha_j} = y_t).$$

Here, we evaluate utility using z_t, which we refer to as a test point, because utility typically reflects the performance of the k-NN model on data points outside the training set. This approach aligns with the goal of assessing the model's ability to generalize to unseen data, ensuring that its utility is not biased by the training data.

The LOO attribution method computes the difference in utility with respect to a test point, with and without the point z_i. Formally, the LOO attribution score for a point z_i in k-NN is given by:

$$v_{\text{loo}}(z_i) = U(z_t; D, k) - U(z_t; D_{-z_i}, k),$$

in which, D_{-z_i} represents the training set excluding z_i. If we want to compute the LOO value for each z_i in D, this approach requires evaluating N distinct neighborhood configurations.

Data Shapley Value. The Shapley Value, a concept from cooperative game theory, extends LOO by averaging the marginal contribution of each point across all possible subsets of the training set (Ghorbani and Zou, 2019; Jia, Dao, Wang, Hubis, Gurel, et al., 2019). The DSV is computed as the average contribution of each point z_i to the model's utility across subsets $S \subseteq D$:

$$v_{\text{shap}}(z_i) = \frac{1}{N!} \sum_{S \subseteq D \setminus \{z_i\}} \frac{1}{\binom{N-1}{|S|}} U(z_t; S \cup \{z_i\}) - U(z_t; S).$$

The Shapley value satisfies four key axioms: *Efficiency* (*i.e.*, the total value is distributed among all players), *Symmetry* (*i.e.*, identical contributions are rewarded equally), *Dummy* (*i.e.*, players that contribute nothing receive zero value), and *Linearity* (*i.e.*, the Shapley values from two games can be combined linearly).

While this formulation provides an axiomatic and robust way of quantifying the importance of each data point, the computational complexity of directly computing Shapley values can be as high as $O(2^N)$, making it infeasible for large datasets. Nonetheless, for k-NN classifiers, an exact formulation of the Shapley value exists, as shown by Jia, Dao, Wang, Hubis, Gurel, et al., which drastically reduces the complexity to $O(N \log N)$. More precisely, the exact Shapley value for k-NN can be computed as follows. For the farthest neighbor z_{α_N}, the Shapley value is:

$$v_{\text{shap}}(z_{\alpha_N}) = \frac{1[y_{\alpha_N} = y_t]}{N}.$$

Afterwards, for the remaining neighbors z_{α_i} with $i < N$, the Shapley value can be recursively computed as:

$$v_{\text{shap}}(z_{\alpha_i}) = v_{\text{shap}}(z_{\alpha_{i+1}}) + \frac{1[y_{\alpha_i} = y_t] - 1[y_{\alpha_{i+1}} = y_t]}{k} \cdot \frac{\min(k,i)}{i}.$$

By exploiting the structure of k-NN classifiers, this exact formulation avoids the need to evaluate individually all subsets, significantly reducing the computational cost.

Note that these attribution methods are agnostic to any specific model, unlike LiRA, which evaluates the impact of a point on a particular trained model. In contrast, DSV implicitly considers all possible k-NN models trained on subsets of the dataset, offering a more comprehensive measure of point importance across model variations.

2.2.2 Likelihood Ratio Attack (LiRA)

LiRA is a method for performing MIAs, which leverages a Likelihood Ratio Test (LRT) to compare the likelihood of a model when trained with and without a specific data point z_i (Carlini, Chien, et al., 2022a). More precisely for a given model f, the LiRA score for z_i is defined as:

$$\Lambda(f; z_i) = \frac{P(f \mid \mathbb{Q})}{P(f \mid \mathbb{Q}_{-z_i})}, \tag{2.2}$$

in which $\mathbb{Q}$ and $\mathbb{Q}_{-z_i}$ represent the distributions of models trained on datasets that respectively include or exclude a training point z_i. By "distributions of models" we mean the set of possible models that can be obtained through training on different subsets of data. For instance for a neural network, this distribution refers to the space of model weights while for k-NN classifiers, the distribution of models corresponds to the distribution of subsets of k data points from the training set D. While theoretically, the number of unique possible combinations is $\binom{N}{k}$, in practice, combinations including the closest neighbors to a test point are more likely. We assume that the attacker does not have access to the training set and can only observe the final prediction or the loss of the model, similar to the adversary model used by Carlini, Chien, et al.

To compute the LiRA score with respect to the loss, the LRT becomes a one-dimensional statistic:

$$\Lambda(\ell; z_i) = \frac{P(\ell(z_i) \mid \mathbb{Q})}{P(\ell(z_i) \mid \mathbb{Q}_{z-i})}. \tag{2.3}$$

Estimating this ratio requires sampling various subsets of the training data, which can be computationally expensive. This motivated the development of our attribution method specifically designed for k-NNs, named WaKA. More precisely, WaKA relies on the same principle as LiRA, focusing on comparing the loss distributions for members and non-members of the training dataset. However, instead of relying on a statistical test to distinguish these distributions, WaKA measures the 1-Wasserstein distance between them, providing a more flexible and computationally efficient approach for k-NNs.

2.3 Wasserstein k-NN Attribution (WaKA)

2.3.1 Attribution using 1-Wasserstein Distance

In our approach, we quantify the importance of each data point in a k-NN model by measuring how the inclusion or exclusion of a data point affects the distribution of the model's loss relative to a specific point z_t. Here, z_t represents the point at which we evaluate the model's loss. In the context of data valuation, z_t is typically a test point for which we aim to assess the influence of training data points on the model's performance. In the context of privacy, z_t could be the data point z_i itself, allowing us to analyze how the inclusion or exclusion of z_i affects the loss distribution relative to z_i. To achieve this, we leverage the 1-Wasserstein distance (Peyré and Cuturi, 2018) between the loss distributions with and without the data point. More formally, the 1-Wasserstein distance (also known as the Earth Mover's Distance) between two probability distributions μ and ν can be defined as:

$$W_1(\mu,\nu) = \int_{-\infty}^{\infty} \left|F_\mu(x) - F_\nu(x)\right| dx$$

in which F_μ and F_ν are respectively the cumulative distribution functions (CDFs) of μ and ν.

Let $\mathbb{L}$ denote the distribution of loss values relative to z_t using k-NN models trained on all possible subsets of D, and $\mathbb{L}_{-z_i}$ refer to the corresponding distribution when z_i is excluded from D. Our goal is to compute the 1-Wasserstein distance between these two distributions to assess the impact of data point z_i. Since the loss values in a k-NN classifier are discrete and belong to the finite set $\mathscr{L} = \left\{0, \frac{1}{k}, \frac{2}{k}, \ldots, 1\right\}$, the 1-Wasserstein distance can be computed directly by summing the absolute difference of cumulative distribution functions (CDFs) of the loss distributions.

We define $\mathscr{F} := \{k\text{-NN trained on subsets of } D\}$ and $\mathscr{F}_{-z_i} := \{k\text{-NN trained on subsets of } D \setminus \{z_i\}\}$, the two spaces of models that correspond to all k-NN models trained with or without the point z_i. In this setting, let $\mathbb{Q}$ denote the uniform distribution over models in $\mathscr{F}$, and $\mathbb{Q}_{-z_i}$ be the uniform distribution over models in $\mathscr{F}_{-z_i}$. Let $\ell : \mathscr{F} \to \mathbb{R}$ denote the loss function mapping a model $f \in \mathscr{F}$ to a real-valued loss $\ell(f)$ evaluated at the point z_t, *i.e.*, $\ell(f) = \ell(z_t; f)$. The distributions $\mathbb{L}$ and $\mathbb{L}_{-z_i}$ represent the pushforward distributions of $\mathbb{Q}$ and $\mathbb{Q}_{-z_i}$ through the loss function ℓ, denoted with the # symbol:

$$\mathbb{L} = \ell_{\#}\mathbb{Q}, \quad \mathbb{L}_{-z_i} = \ell_{\#}\mathbb{Q}_{-z_i}.$$

More precisely, $\mathbb{L}$ is the distribution of losses evaluated at z_t induced by models drawn from $\mathbb{Q}$, and $\mathbb{L}_{-z_i}$ is the distribution of losses evaluated at z_t induced by models drawn from $\mathbb{Q}_{-z_i}$.

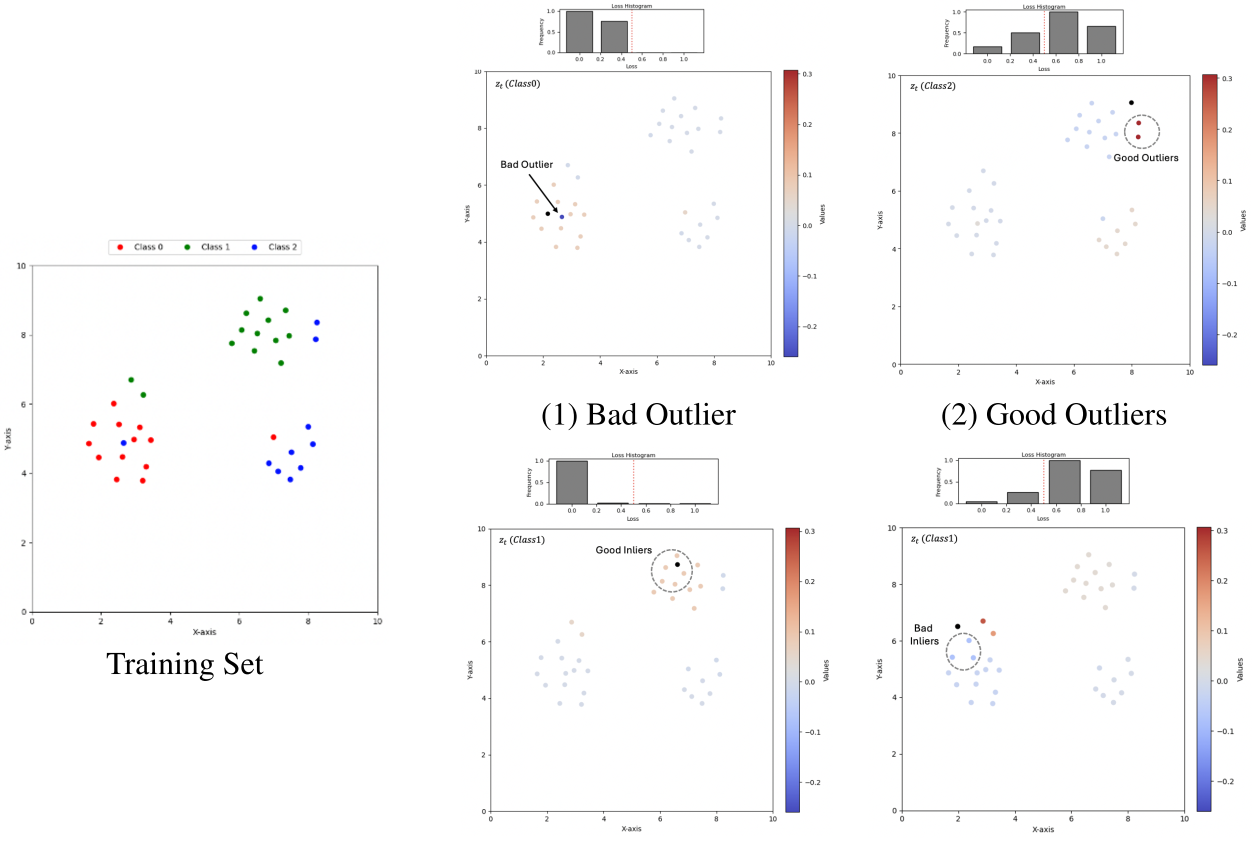
Definition 2.3.1 (WaKA)**.** We define the 1-Wasserstein k-NN attribution for data point z_i relative to any point z_t as follows:

$$W_1(\mathbb{L}, \mathbb{L}_{-z_i}) = \frac{1}{k} \sum_{l_{\min} \leq l \leq l_{\max}} \left| F_{\ell_{\#}\mathbb{Q}}(l) - F_{\ell_{\#}\mathbb{Q}_{-z_i}}(l) \right|, \tag{2.4}$$

in which $\mathbb{L}$ and $\mathbb{L}_{-z_i}$ are the loss distributions relative to z_t, restricted to the range $[l_{\min}, l_{\max}]$. The cumulative distribution functions (CDFs) $F_{\ell_{\#}\mathbb{Q}}(l)$ and $F_{\ell_{\#}\mathbb{Q}_{-z_i}}(l)$ are evaluated for these restricted distributions at the discrete loss values $l \in \mathscr{L} = \left\{0, \frac{1}{k}, \frac{2}{k}, \ldots, 1\right\}$. Here, $l_{\min} \geq 0$ and $l_{\max} \leq 1$ specify the bounds of the restricted range.

While DSV provides a unique set of values, WaKA introduces flexibility by focusing on how probability masses are moved positively or negatively, relative to particular loss values. For instance, when z_t and z_i share the same label ($y_t = y_i$), WaKA can measure how much loss mass is moved to improve the loss distribution. Conversely, when z_t and z_i have different labels ($y_t \neq y_i$), WaKA can measure how much z_i worsens the loss distribution. Additionally, we can refine this analysis by incorporating the decision threshold of the k-NN classifier. Typically, the decision threshold is set to $1/2$, but in imbalanced datasets, it can be adjusted to favor the minority class. WaKA can identify points skewed positively or negatively relative to the decision threshold, making it especially useful for tasks like data removal or data addition. Figure 2.2 illustrates four types of points that WaKA can identify well.

Figure 2.2: *Illustrating the importance of data points in data valuation tasks. The plots highlight: (1)* ***Bad Outlier*** *(Top-Left): Outlier point that negatively impact (low Shapley value) one class while offering little benefit on its own (Class 1). Important to identify for data removal; (2)* ***Good Outliers*** *(Top-Right): Outlier points that, despite being in less dense region, improve their class performance (high Shapley value). (3)* ***Good Inliers*** *(Bottom-Left): Inlier points contributing to loss distributions skewed toward zero, crucial for data addition; (4)* ***Bad Inlier*** *(Bottom-Right): Not outlier but harm Class 2 if added to dataset.*



Training Set

(1) Bad Outlier

(2) Good Outliers

(3) Good Inliers

(4) Bad Inlier

WaKA for data removal and addition. To address these data valuation tasks, we propose two formulations of WaKA: one for data removal (WaKA$_{\text{rem}}$) and one for data addition (WaKA$_{\text{add}}$). Both formulations rely on a fixed decision threshold $\tau \in [0,1]$, which divides the domain of loss values into two regions.

For data removal, WaKA$_{\text{rem}}$ identifies outliers that negatively affect one class while contributing little to other classes. It also accounts positively for points that, despite being outliers, improve their class. The formulation is:

$$\begin{aligned}\mathrm{WaKA}_{\mathrm{rem}}(z_i) = \mathbf{1}(y_{\alpha_i} = y_t) \sum_{l > 1-\tau} \left| F_{(1-\ell)\#Q}(l) - F_{(1-\ell)\#Q_{-z_i}}(l) \right| \\ - \mathbf{1}(y_{\alpha_i} \neq y_t) \sum_{l} \left| F_{\ell\#Q}(l) - F_{\ell\#Q_{-z_i}}(l) \right|.\end{aligned}$$

For data addition, $\mathrm{WaKA}_{\mathrm{add}}$ prioritizes inliers but it also penalizes points that are not outliers but negatively affect other classes. The formulation is:

$$\begin{aligned}\mathrm{WaKA}_{\mathrm{add}}(z_i) = \mathbf{1}(y_{\alpha_i} = y_t) \sum_{l} \left| F_{(1-\ell)\#Q}(l) - F_{(1-\ell)\#Q_{-z_i}}(l) \right| \\ - \mathbf{1}(y_{\alpha_i} \neq y_t) \sum_{l \leq \tau} \left| F_{\ell\#Q}(l) - F_{\ell\#Q_{-z_i}}(l) \right|.\end{aligned}$$

These formulations are straightforward applications of WaKA tailored for data removal and data addition tasks. While other formulations are possible, these provide a simple yet effective approach to prioritizing specific points based on their contributions to the loss distribution. In the context of data valuation, z_t typically refers to a test point; however, when $z_t = z_i$, we enter what we call self-attribution, in which the focus shifts to understanding the contribution of z_i to its own loss distribution. In the following section, we adapt WaKA for membership inference, not only by looking at the loss distribution of z_i but also by incorporating the loss of a specific model.

2.3.2 Adapting WaKA for Membership Inference

In line with the “security game” framework employed in LiRA for evaluating MIAs, the adversary’s objective is to ascertain whether a specific point i is part of the training dataset. This framework involves an interaction between a challenger and an adversary. First, the challenger samples a training dataset and trains a model over it. Then, depending on the outcome of a private random bit, the challenger sends either a fresh challenge point to the adversary or a point from the training set. The adversary, having query access to both the distribution (*i.e.*, in the sense that it can have access to samples for this distribution) and the trained model, must then decide if the given point was part of the training set. This structured game can be used to assess the adversary's ability to correctly infer membership in terms of his advantage compared to a random guess, thereby evaluating the robustness of the model against such attacks.

Since we are dealing with k-NN models, adding the point z_i to the training set can only improve the model's performance, meaning that the loss will decrease. As such, the distribution $\mathbb{L}_{-z_i}$ is shifted towards lower values when transitioning to $\mathbb{L}$. However, in the context of the security game, we are given a specific loss value $\ell(z_i)^*$ for the true model. As a result, models with loss greater than $\ell(z_i)^*$ are incentivized to indicate that z_i is part of the training set, as the distribution is shifting towards $\ell(z_i)^*$, while models with loss less than $\ell(z_i)^*$ suggest the opposite, since adding z_i to the training set would further decrease the loss. Therefore, we can refine the model spaces $\mathscr{F}$ and $\mathscr{F}_{-z_i}$ into two partitions:

$$\begin{aligned}
\mathscr{F}^{+} &= \{f \in \mathscr{F} : \ell(z_i;f) \geq \ell(z_i)^*\},\\
\mathscr{F}^{-} &= \{f \in \mathscr{F} : \ell(z_i;f) < \ell(z_i)^*\},\\
\mathscr{F}^{+}_{-z_i} &= \{f \in \mathscr{F}_{-z_i} : \ell(z_i;f) \geq \ell(z_i)^*\},\\
\mathscr{F}^{-}_{-z_i} &= \{f \in \mathscr{F}_{-z_i} : \ell(z_i;f) < \ell(z_i)^*\}.
\end{aligned} \tag{2.5}$$

Definition 2.3.2 (t-WaKA)**.** The target-WaKA attribution score for point z_i, denoted as t-WaKA(z_i), is defined as:

$$\begin{aligned}
\text{t-WaKA}(z_i) = W_1 &\left(\mathbb{L} \mid \mathscr{F}^{+}, \mathbb{L}_{-z_i} \mid \mathscr{F}^{+}_{-z_i}\right)\\
&- W_1 \left(\mathbb{L} \mid \mathscr{F}^{-}, \mathbb{L}_{-z_i} \mid \mathscr{F}^{-}_{-z_i}\right).
\end{aligned} \tag{2.6}$$

We can use the previous formula to obtain the following simplification:

$$\begin{aligned}
\text{t-WaKA}(z_i) = \frac{1}{k} \sum_{l \geq \ell(z_i)^*} &\left| F_{\ell_{\#}\mathbb{Q}}(l) - F_{\ell_{\#}\mathbb{Q}_{-z_i}}(l) \right|\\
&- \frac{1}{k} \sum_{l < \ell(z_i)^*} \left| F_{\ell_{\#}\mathbb{Q}}(l) - F_{\ell_{\#}\mathbb{Q}_{-z_i}}(l) \right|.
\end{aligned} \tag{2.7}$$

Algorithm 1 Counting the marginal contributions of a point of interest z_i with respect to a test point z_t

```
**Input:** Sorted training set $D$, point of interest $i$, test point $t$, number of neighbors $k$.
**Output:** Contributions for all losses of point $i$
Initialize $\text{CPV}[j] = \sum_{m \neq i}^{j} \mathbf{1}(y_{\alpha_m(D)} = y_t)$ for $j = 1, \ldots, N$  ▷ Cumulative positive votes up to $z_j$ excluding $z_i$
Initialize $\text{CNV}[j] = \sum_{m \neq i}^{j} \mathbf{1}(y_{\alpha_m(D)} \neq y_t)$ for $j = 1, \ldots, N$  ▷ Cumulative negative votes up to $z_j$ excluding $z_i$
Contributions $\leftarrow \mathbf{0}$
**for** $j = i + 1$ to $N$ **do**
    **if** $y_{\alpha_j(D)} \neq y_{\alpha_i(D)}$ and $j > k$ **then**
        **for all** $l$ in $\{0, 1/k, \ldots, 1\}$ **do**
            $\text{NV} \leftarrow \text{round}(l \cdot k)$  ▷ Number of negative votes for loss $l$
            $\text{PV} \leftarrow k - \text{NV}$  ▷ Number of positive votes for loss $l$
            $\delta_{\text{PV},i} \leftarrow \mathbf{1}(y_{\alpha_i(D)} = y_t)$
            $\delta_{\text{NV},i} \leftarrow \mathbf{1}(y_{\alpha_i(D)} \neq y_t)$
            $\delta_{\text{PV},j} \leftarrow \mathbf{1}(y_{\alpha_j(D)} = y_t)$
            $\delta_{\text{NV},j} \leftarrow \mathbf{1}(y_{\alpha_j(D)} \neq y_t)$
            **if** $\text{CPV}[j] \geq \text{PV} - \delta_{\text{PV},j}$ and $\text{CNV}[j] \geq \text{NV} - \delta_{\text{NV},j}$ and $\text{CPV}[j] \geq \text{PV} - \delta_{\text{PV},i}$ and $\text{CNV}[j] \geq \text{NV} - \delta_{\text{NV},i}$ **then**  ▷ k-nearest neighbors combinations should be valid
                $\text{Count}_{-z_i} \leftarrow \binom{\text{CPV}[j]}{\text{PV} - \delta_{\text{PV},j}} \cdot \binom{\text{CNV}[j]}{\text{NV} - \delta_{\text{NV},j}}$
                $\text{Count} \leftarrow \binom{\text{CPV}[j]}{\text{PV} - \delta_{\text{PV},i}} \cdot \binom{\text{CNV}[j]}{\text{NV} - \delta_{\text{NV},i}}$
                $\text{Contributions}[l] \leftarrow \text{Contributions}[l] + \frac{\text{Count} - \text{Count}_{-z_i}}{2^j}$
            **end if**
        **end for**
    **end if**
**end for**
**return** Contributions
```

2.3.3 Attribution Algorithm

To compute the 1-Wasserstein distance, we assume access to approximations of Probability Mass Functions (PMFs) of $\mathbb{L}$ and $\mathbb{L}_{-i}$, which take the form of histograms. More precisely, for each possible loss value l in the set $\{0, 1/k, \ldots, 1\}$, we want to count the associated number of existing k-NN models, including and excluding z_i. In the context of k-NNs, there is no need to sample many subsets of the training set to compute these values as we can count exactly the contribution of a point z_i. A key insight is the observation that computing the difference in PMFs due to z_i is proportional to calculating the marginal contribution of z_i when z_i is added to the training set D. A contribution occurs for any y_i if $y_i \neq y_j$ for all $j > i$, with the points sorted relative to the test point z_t. To realize this, we have designed Algorithm 1 to provide an efficient method for computing the marginal contributions.

In a nutshell, the algorithm starts by ordering the training set with respect to the test point z_t before iterating through the sorted labels to store the cumulative positive and negative votes, CPV[j] and CNV[j], at each index j. Note that if $z_t = z_i$, we are in the context of self-attribution, or in the context of an MIA with t-WaKA. Afterwards, the algorithm loops over the training set to find labels that differ from y_i (this step could be parallelized), the label of the point of interest, which could be precomputed in the previous step. For each differing label with an index greater than k, a pass is done over the possible loss values to compute the marginal contribution of the point of interest z_i. This process consists of two steps: first, we count the combinations of k nearest neighbors (of z_t) a particular loss, in which z_i is included are counted. Similarly, we count the combinations in which z_j is excluded for the same loss. Then, this count is normalized using the term 2^j, which corresponds to all supersets of the k nearest neighbors. Finally, the marginal contribution is either stored or aggregated. Once the marginal contributions are calculated, the computation of the 1-Wasserstein distance is straightforward using histograms of the losses (a proof and more details are provided in Appendix 2.8.

2.3.4 Computational analysis

The worst-case time complexity of the algorithm is $O(N \log N + kN)$. More precisely, the $O(N \log N)$ term comes from sorting the N training points with respect to their distance to the test point z_t. Initializing the cumulative vote functions for any j, CPV[j] and CNV[j], requires $O(N)$. The main loop, which iterates over the training set and processes each point's contributions to each loss, runs in $O(k \cdot N)$. This complexity is reduced by using an approximation leveraging the fact that contributions decrease exponentially due to the 2^j factor and that k becomes much smaller than j. This enables to focus on a fixed-size neighborhood around the target point rather than considering all training points.

In practice in our experiments, we used a neighborhood of 100 points to compute these values. By restricting the computation to this fixed neighborhood, CPV[j] and CNV[j] can also be measured with a constant complexity. To efficiently identify the neighborhood, we can use an optimized data structure such as a kd-tree, which allows for identifying the nearest neighbors in $O(\log N)$ time. With this approximation, only a single reusable k-NN model needs to be trained, which reduces significantly the memory usage compared to using multiple k-NN shadow models as done in LiRA. More precisely, the complexity of computing marginal contributions for any point of interest in this approximation becomes $O(\log N + K)$, in which K is the fixed size of the neighborhood. This approach is particularly efficient for large datasets. In the following experiments, we have employed this approximation and observed a minimal impact on the estimated value.

2.4 Experimental Evaluation

In this section, we present our experimental evaluation of WaKA as a general attribution method and t-WaKA as a membership inference attack (MIA) on six widely-used public datasets (see Table 1.1 in Appendix 2.6). Our main objective with the following experiments is to explore the dual use of WaKA: first, as a tool for understanding the contribution of individual data points to the utility and privacy of k-NN models and second, as an efficient approach for conducting MIAs on specific data points.

2.4.1 Experimental Setting

To apply WaKA to standard image and text datasets in machine learning, we prioritized using pre-trained neural network embeddings to minimize dependencies to specific datasets. For CIFAR-10, we used a custom pre-trained embedding based on ImageNet, extracting feature representations from its last layer to obtain a transferable encoding. We evaluated both ImageNet embeddings and custom embeddings trained on a reserved portion of CIFAR-10. While the latter showed a slight performance gain, the overall improvement was negligible. Therefore, we chose to use ImageNet embeddings, as this avoids partitioning the training set and reduces the risk of data leakage. For the CelebA dataset, we employed a pre-trained Vision Transformer (ViT) embedding Dosovitskiy, 2020. For the IMDB and Yelp reviews textual datasets, we used a pre-trained version of the Sentence-BERT (SBERT) model Reimers, 2019, ensuring that it did not include the two datasets during training, thereby aligning with our goal of reduced data dependencies. In a nutshell, SBERT is a modification of the BERT architecture that produces sentence embeddings optimized for semantic similarity tasks and downstream classification. Finally, for tabular datasets, we employed a straightforward encoding approach, which includes one-hot encoding for categorical features and normalization for numerical features. Note that the CelebA and Yelp datasets were added and only used for data valuation experiments.

Our experiments have been designed around two key scenarios for evaluating attribution methods. The first scenario, which we call "test-attribution", represents the classical data valuation setting, in which the validation set is used to compute aggregated attribution scores and measure their impact on utility through performance on a separate test set. The second scenario, which we name "self-attribution", focuses on computing the attribution value of a point z_i based on how well the model predicts the corresponding label y_i. This situation quantifies the direct contribution of each point to its own label prediction and we conjecture that this self-attribution is more correlated to privacy scores than utility. In a nutshell, to compute self-attribution, we only need to use the values attributed to each point to predict itself while for test-attribution, we must decide on how to aggregate the values calculated for each test point. More precisely, for test-based attribution, the Shapley value, denoted as $\mathrm{DSV}_{\mathrm{test}}(z_i)$, computes the average contribution of each training point across the test set, formulated as:

$$\mathrm{DSV}_{\mathrm{test}}(z_i) = \frac{1}{|D_{\mathrm{test}}|} \sum_{z_t \in D_{\mathrm{test}}} v_{\mathrm{shap}}(z_i; z_t),$$

in which $v_{\mathrm{shap}}(z_i; z_t)$ is the Shapley value of training point z_i with respect to test point

z_t. This method benefits from Shapley's linearity axiom, which ensures that contributions from individual data points can sum to the total utility. Note that a test point can assign both positive and negative valuations to a training point, depending on its effect on the model's performance. Similarly, for WaKA_{rem} and WaKA_{add}, we compute the average across the test set by evaluating each test point $z_t \in D_{\text{test}}$.

In our experiments, we selected $k = 1$ as the least privacy-preserving parameter value, since the k-NN model's predictions are directly influenced by individual points. We also chose $k = 5$ because it is a commonly used parameter in the literature, offering more label anonymization through generalization, as predictions are influenced by a larger set of neighboring points.

2.4.2 WaKA for Utility-driven Data Minimization

The state-of-the-art approach for utility-driven data minimization is Data Shapley Value (DSV, Jia, Dao, Wang, Hubis, Gurel, et al., 2019). In addition to WaKA and DSV, the basic Leave-One-Out (LOO) method was also included, serving as a benchmark due to its simplicity in utility-driven data removal and addition tasks. In data addition, models are trained incrementally by adding a certain percent of the dataset in descending order of importance based on an attribution method. Conversely, in data removal, models start with the full dataset and are retrained after progressively removing the least valuable points. This setup allows us to assess the impact of each method on model performance as data is added or removed. Looking at the results (Figure 2.6), it is clear that the LOO method generally underperforms as a data valuation technique, consistent with the findings of Ghorbani and Zou, 2019. This reinforces the limitations of LOO in effectively identifying the most influential points for utility-driven data minimization. In contrast, WaKA_{rem} and WaKA_{add} consistently match or outperform DSV, particularly in imbalanced datasets. For instance, Yelp, with a minority class of 0.4, and Adult, with a minority class of 0.24, are noticeably more imbalanced compared to CIFAR and IMDB, which are perfectly balanced through preprocessing. As shown in Figure 2.6, WaKA demonstrates greater robustness in these cases. On the Yelp dataset, we observe that while DSV's macro F1 score is comparable to LOO for data removal, WaKA$_{\text{rem}}$ maintains significantly better performance (Figure 2.3). Similarly, on the Adult dataset, Figure 2.4 illustrates that during data addition, DSV predominantly favors adding majority class samples, leading to performance worse than random addition, whereas WaKA$_{\text{add}}$ effectively balances the dataset, yielding superior results.

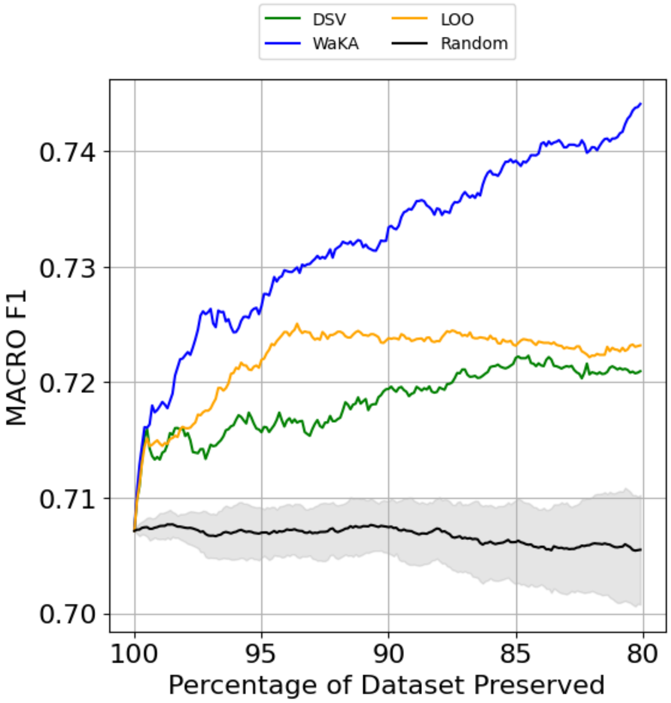


Figure 2.3: *F1 score on the Yelp dataset for data removal. WaKA$_{rem}$ maintains significantly better performance compared to DSV and LOO, highlighting its robustness on imbalanced datasets.*

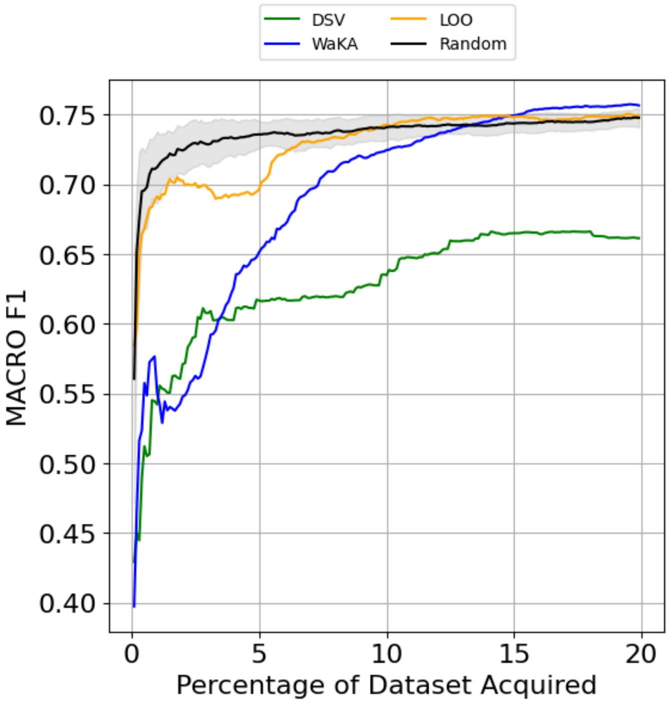


Figure 2.4: *F1 score on the Adult dataset for data addition. DSV predominantly favors adding majority class samples, leading to performance worse than random addition. WaKA$_{add}$, by contrast, effectively balances the dataset and achieves better results.*

An important aspect of WaKA's formulation is its sensitivity to the parameter τ, which controls the weighting of outliers and inliers during data removal and data addition. The role of τ is to introduce flexibility in how probability mass is moved within the loss distribution, influencing whether points are prioritized based on their effect on model utility. Specifically, when $\tau = 1.0$ for addition and $\tau = 0.0$ for removal, the formulations become identical. In data addition, the goal is to emphasize inliers—points that contribute significantly to reducing loss—while penalizing those that negatively impact other points depending on their loss distribution. In data removal, we focus on outliers, particularly those that increase loss, while still accounting for outliers that provide positive contributions. Empirical evaluations (see Appendix 2.9) indicate that $\tau = 0.5$ leads to stable results across datasets, though higher variance is observed in imbalanced datasets such as Bank and CelebA. This suggests that further exploration of τ could be valuable in contexts where dataset imbalance affects data attribution.

Finally, to further explore the impact of data valuation methods, we examined their influence on class balance during data removal. Figure 2.5 highlights that DSV tends to disproportionately remove points from the minority class, exacerbating class imbalance. In contrast, WaKA_{rem} demonstrates a more balanced removal strategy, effectively preserving class proportions. We include all results, such as CelebA and Bank datasets, in Appendix 2.9. These additional experiments confirm the bias of Shapley values towards the majority class and further underscore the robustness of WaKA in handling imbalanced data scenarios.

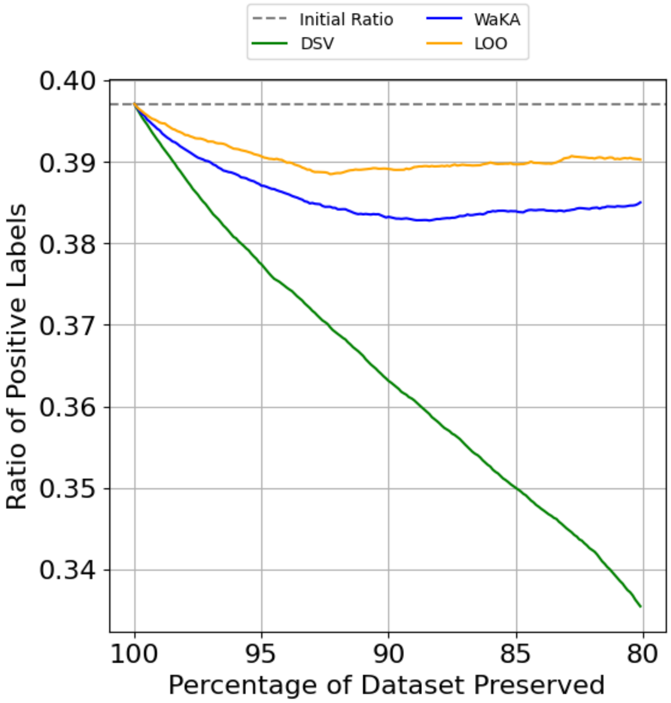
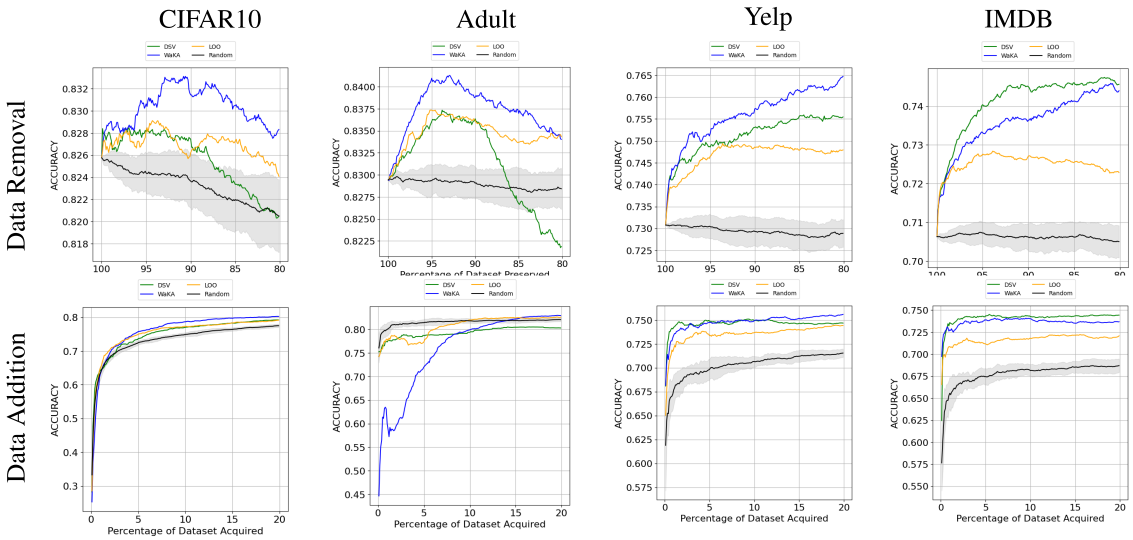


Figure 2.5: *Effect of data removal on class balance for the Yelp dataset. WaKA_{rem} preserves class proportions, while DSV disproportionately removes points from the minority class, exacerbating imbalance.*

Figure 2.6: *Attribution using WaKA$_{rem}$ for data removal and WaKA$_{add}$ for data addition, compared with two test-attribution methods: Data Shapley and Leave-One-Out (LOO).* ***Data addition*** *starts with an empty dataset (0%) and progressively adds points, either randomly (black line) or in descending order of importance as ranked by an attribution method (e.g., Shapley, WaKA$_{add}$, LOO).* ***Data removal*** *begins with the full dataset (100%) and iteratively removes the least valuable points according to each method. The x-axis represents the percentage of data added or removed, while the y-axis tracks the model's accuracy on the test set*



2.4.3 WaKA for Privacy Evaluation and Auditing

Privacy scores. In our privacy evaluation, we followed a similar approach to Carlini, Jagielski, et al., 2022 by computing the average ASR on all training points. In practice, this was done by simulating security games as described previously, in which multiple random partitions of the training set were created, and the LiRA method was applied to evaluate the ASR. Once the ASR values were obtained, the attribution scores were computed from each method and sorted to analyze their correlation with the ASR. This allows to identify how well each attribution method aligns with the likelihood of a successful MIA.

Both self-Shapley and self-WaKA have shown a monotonic increase in self-attribution values as the ASR increases (see Figure 2.7), for both values of the k-NN parameter (*i.e.*, $k = 1$ and $k = 5$). However, it is important to note that the correlation between ASR and self-attribution values does not behave uniformly across all datasets. For instance, in the Bank dataset, the attack accuracy remains close to 0.5 (*i.e.*, random guessing) until about the 70-th percentile of self-WaKA values, whereas other datasets, such as IMDB, exhibit a more gradual increase in ASR across percentiles. This indicates that different datasets exhibit varying levels of correlation between ASR and self-attribution. Furthermore, test-attribution methods showed different correlations, with notably a higher average ASR towards extremes. This means that very low or very high attribution scores tend to correlate more strongly with higher ASR values. More precisely, data points that are highly detrimental to utility were found to have consistently high ASR values across all datasets. This observation challenges the common belief that only high-value points pose significant privacy risks as lower-value points can also be privacy vulnerable.

Additionally, increasing the parameter value k reduces the average ASR (Table 2.8 in Appendix 2.10) while shifting higher ASR values towards the upper percentiles of the attribution scores. We further tested the Spearman rank correlation between ASR and attribution scores, finding that they are highly correlated across datasets (see Table 2.1).

Table 2.1: *Spearman Correlation between self-WaKA and self-Shapley for K=1 and K=5*

Dataset	**K=1**		**K=5**	
	Spearman	**P-Value**	**Spearman**	**P-Value**
IMDB	0.99	0.00	0.93	0.00
CIFAR10	1.00	0.00	0.99	0.00
adult	0.99	0.00	0.98	0.00
bank	0.98	0.00	0.98	0.00

This suggests that self-attribution methods, such as in particular self-WaKA, can more reliably indicate privacy risks in k-NN models than test attribution. This can be explained by the fact that self-Shapley can be understood through a game-theoretic lens, in which each data point z_i is a "player" contributing to the prediction. In this setting, a high self-Shapley value means that z_i plays a dominant role in predicting itself, indicating that whatever the coalition, its neighbors contribute little. Similarly, a high self-WaKA value means that removing z_i significantly changes the loss distribution of predicting itself across all possible subsets of the training set, which aligns more closely with LiRA.

Finally, we have conducted additional experiments to determine whether self-WaKA values could provide privacy insights for models beyond k-NN classifiers. To investigate, LiRA and trained logistic regression shadow models we first looked at the logits of the models confidence, following the approach outlined by Carlini and co-authors. More precisely, the LiRA scores were compared with the self-WaKA values for $k = 1$ to see if there was a correlation with the ASR. Our observations revealed that, across all datasets, the average ASR for the logistic regression model was lower than that of k-NN with $k = 5$ (see Table 2.8). Although the correlation between value percentiles and ASR was less pronounced than in the k-NN case, a significant relationship can still be observed (see Figure 2.22 in Appendix 2.10). This suggests that self-attribution on k-NN models may offer insights into the data that can be extrapolated to other types of models, although we leave as future works the detailed investigation of such research avenes.

Privacy influences. The "onion effect" as explored in Carlini, Jagielski, et al., 2022 throught the introduction the concept of privacy influence (PrivInf) as a method to quantify the influence that removing a specific training example z has on the privacy risk of a target example z'. Specifically, it is defined as the change in the membership inference accuracy on z' after removing z, averaged over models trained without z:

$$\text{PrivInf}(\text{Remove}(z) \Rightarrow z') := \mathbb{E}_{f\in\mathscr{F},S\subseteq D}\left[\mathbf{1}(z' \in S) \mid z \notin S\right]$$

This definition can be used to analyze how the removal of certain training samples (*e.g.*, inliers or outliers) affects the ASR of other data points.

Following the same experimental setting as in Carlini, Jagielski, et al., 2022, we started by removing 10% of the training set with the highest self-WaKA values before re-evaluating the ASR on the remaining points. The success of MIAs across all data points is compared by plotting the AUC curve on a log scale, following the authors' recommendation (see Figure 2.23 in Appendix 2.10). As observed by Carlini and collaborators, while the overall risk is reduced after removing certain training points, it remains higher than the expected decrease. Similar patterns were observed when removing points with the highest ASR or self-Shapley values, whereas random removal or other attribution methods resulted in no significant change in ASR reduction (see Tables 2.6 and 2.7 in Appendix 2.10). This confirms that k-NN models exhibit similar privacy layers, for which removing vulnerable data does not proportionally reduce the vulnerability to privacy attacks, thus confirming the "onion effect". The distribution of ASR was also analyzed across all points. For CIFAR10, many points still exhibit high ASR values close to 1.0 after removal. For the Bank, Adult and IMDB datasets, a general shift occurs towards lower ASR values after data removal, though some points with high ASR remain, particularly near the upper percentiles. This highlights the varying impact of data removal on privacy risks across datasets.

Computing privacy influence (PrivInf) is computationally expensive, as it requires removing a point z or a subset of points Z and re-evaluating privacy scores for all other points z' in the training set to determine which points have the most influence on their privacy score. Instead of directly computing PrivInf, we investigated whether changes in the ASR of the remaining points after removing 10% of the training set could be explained by their self-attribution values. To achieve this, an approach to compute point-wise influence on self-attribution values is needed. In particular for Shapley, this task is non-trivial, as exact Shapley values are computed recursively, starting from the last sorted data point (*i.e.*, complexity of $O(n)$. Efficiently re-computing Shapley values after removing a point would require adapting the Shapley k-NN algorithm to support faster recalculations. For WaKA, the situation is different as the contributions of each point to the WaKA value for a point z_i are independent and can be stored, allowing for re-computation of the self-WaKA value in $O(1)$ time. This provides a strong advantage for self-WaKA over self-Shapley in terms of computational efficiency. To compute the influence of removing a subset $Z \subseteq D$ on the self-WaKA values—denoted as $\text{WaKAInf}(\text{Remove}(Z) \Rightarrow z)$ —the contributions of each removed point $z_j \in Z$ on the self-WaKA values are summed on the remaining points. The total WaKA influence is defined as:

$$\text{WaKAInf}(\text{Remove}(Z) \Rightarrow z) = \sum_{z_j \in \mathcal{N}(z) \cap Z} \text{WaKA}_{self}(z; D_{-z_j}) - \text{WaKA}_{self}(z; D), \tag{2.8}$$

in which $\mathcal{N}(z)$ represents the fixed neighborhood around z, and z_j corresponds to a point that was part of the subset Z removed. If no influencing points are included, the total influence is zero. Figure 2.9 shows the correlation between ASR change and WaKAInf, using k=1, the value of the k parameter with the highest privacy risks. Three distinct regimes across all datasets can be observed. In the leftmost region, points with negative WaKAInf values correspond to negative ASR changes, indicating reduced vulnerability after removal. In the middle, in which WaKAInf values are near zero, the ASR changes are close to 0, suggesting little impact on privacy. Finally, in the rightmost region, higher WaKAInf values align with positive ASR changes, indicating increased vulnerability.

Figure 2.7: *Correlation between ASR and self-attribution values across different datasets for k-NN values of $k = 1$ and $k = 5$. **Self-attribution** measures the extent to which a data point contributes to its own prediction, while ASR (Attack Success Rate) represents the likelihood of a successful membership inference attack, serving as a measure of privacy risk per point. The ASR increases monotonically with self-attribution values in most datasets, but the behavior varies across datasets. For example, the Bank dataset exhibits a steep increase in ASR around the 70th percentile of self-WaKA values, while other datasets, such as IMDB, show a more gradual augmentation. Additionally, **test-attribution** is less correlated with ASR compared to self-attribution, indicating that self-attribution is a stronger predictor of privacy risk. The darker curves represent self-attribution methods, including self-Shapley (red) and self-WaKA (blue).*

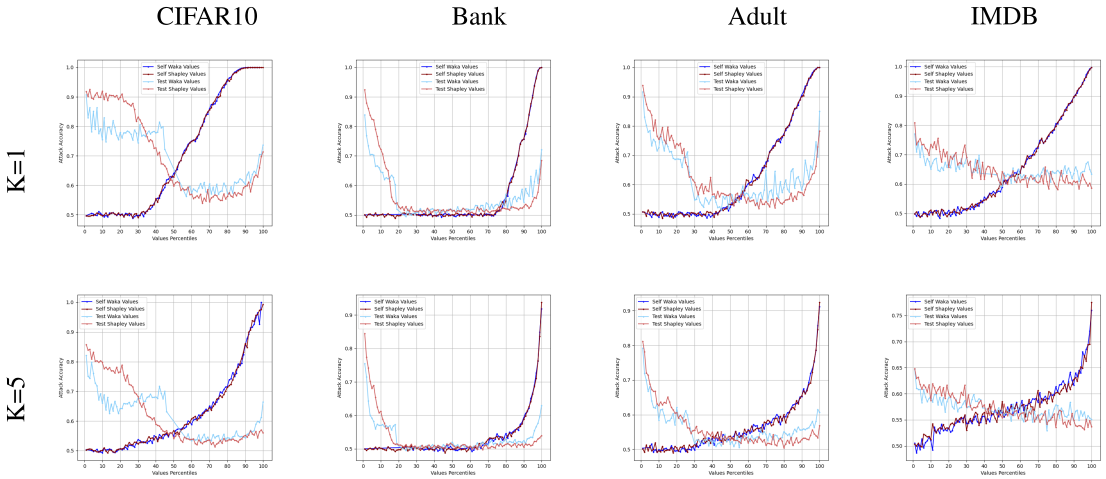


Figure 2.8: *Comparison of data points ASR histograms at $k = 1$. The blue bars represent the 100% dataset scenario, while the red bars show the 90% dataset scenario after removing the 10% of points with the highest self-WaKA values. While the overall ASR distributions appear similar, the removal of these high-risk points significantly reduces the probability of having data points with ASR close to 100%, indicating a mitigation of extreme privacy vulnerabilities.*

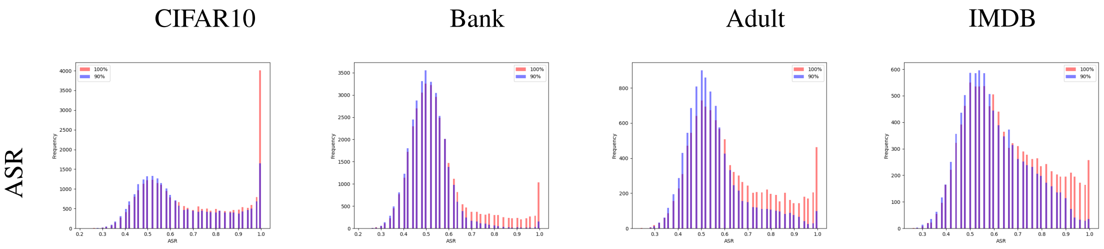
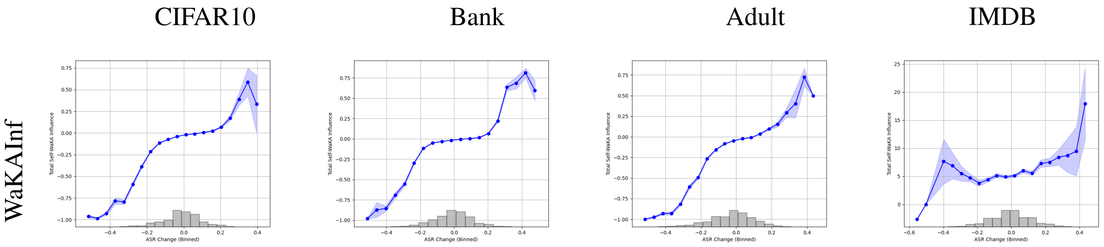


Figure 2.9: *The relationship between ASR change and total self-WaKA influence (WaKAInf) for $k = 1$. Three regimes can be seen: negative WaKAInf values lead to negative ASR changes (reduced vulnerability), medium WaKAInf values show minimal ASR change while and higher WaKAInf values correlate with positive ASR changes (increased privacy risks). ASR histograms before and after removal for $k = 1$.*



Membership inference attack through target-WaKA (t-WaKA). We have conducted extensive experiments using target-WaKA with multiple parameters ($k = 1$ to 5) across all six datasets. The results of the attacks, specifically TPR at a low FPR, are shown in Table 2.2. Additional results, including AUC curves (in log scale) and their corresponding values, are provided in Appendix 2.10. Inspired by the confidence-based attacks described in Ye et al., 2022a, we introduce two new attacks on k-NN: a "confidence" attack (Conf) and calibrated confidence attack (Conf-calib). Both leverage the target model's confidence on each target point and compare it to that of a k-NN model built on a small neighborhood (i.e. 100 points). For Conf, this is done with a single k-NN, whereas Conf-calib samples multiple k-NNs (akin to using shadow models) to calibrate the confidence score against variations in the local neighborhood. Since t-WaKA and LiRA share the same principles of leveraging a model's response for membership inference, LiRA will serve as our main point of comparison. To compute the mean AUC and TPRs (True Positive Rates) for all FPRs (False Positive Rates) used in ROC curves, a bootstrap approach was employed in which 48 security games are ran in parallel, using a predefined list of seeds. In each game, the training set was split in half, with a k-NN model trained on one half. From these two halves, 100 points were drawn at random for evaluation. The LiRA method was implemented with 16 shadow models and although we have tested with an increase in the number of shadow models, an early convergence was observed across all datasets. Training a k-NN model essentially involves storing the entire training set and optionally building an optimized structure, such as a kd-tree, to facilitate quick neighborhood searches during inference. In our experiment, we did not use any such structures but rather, we ordered all points with respect to the target or test point.

These experiments differ from the previous privacy evaluations, as 100 randomly sampled points from the dataset are targeted for a specific model, repeating the process 48 times. While one might expect that using attribution values would lead to effective attacks on that repeated security game, we found the opposite. More precisely, all attribution methods produced an average AUC close to 0.5, indicating random performance. We believe that this can be explained by the fact that attribution methods evaluate data points with respect to all possible models, rather than focusing on a specific model's loss. In contrast, both LiRA and t-WaKA can incorporate the loss of a particular model, making them more effective for this type of targeted attack.

Table 2.2: *Comparison of TPR at FPR=0.05 among LiRA, t-WaKA, Conf, and Conf-calib for* $K = 5$

Dataset	LiRA	t-WaKA	Conf	Conf-calib
Adult	0.13 ± 0.17	0.13 ± 0.15	0.06 ± 0.11	0.14 ± 0.19
Bank	0.10 ± 0.16	0.10 ± 0.19	0.05 ± 0.07	0.12 ± 0.17
CelebA	0.07 ± 0.09	0.08 ± 0.12	0.06 ± 0.07	0.08 ± 0.11
CIFAR10	0.16 ± 0.24	0.15 ± 0.16	0.05 ± 0.10	0.14 ± 0.17
IMDB	0.10 ± 0.16	0.14 ± 0.18	0.06 ± 0.11	0.12 ± 0.19
Yelp	0.11 ± 0.17	0.13 ± 0.17	0.07 ± 0.09	0.10 ± 0.16

For almost all values of *k*, the AUCs of LiRA, t-WaKA, and Conf-calib are remarkably comparable. Following the recommendation of Carlini, Chien, et al. to assess the success of membership inference attacks, we focus on TPR at a low FPR (5As the value of the parameter *k* increases, the success rate of the attacks decreases similarly for all three methods. We observed that t-WaKA performs slightly worse and sometimes a little better than LiRA, but overall, the results are very close (see Table 2.2). The Conf attack performs worse than the other methods, but Conf-calib performs just as well as LiRA and t-WaKA, sometimes even surpassing them. Interestingly, t-WaKA remains the strongest attack on textual datasets (IMDB and Yelp). In some scenarios Additionally, the results confirmed that t-WaKA shows significant correlation in rankings using the Spearman test, indicating consistent performance across different settings (Table **??**).

Note that as *k* increase there is less and less rank correlation with LiRA. This outcome is likely due to the same number of shadow models being used for all *k*, but further investigation is needed to confirm this.

In terms of execution time, t-WaKA demonstrates significant efficiency compared to LiRA. We report the execution times for attacking the IMDB dataset across all values of *k* from 1 to 5. IMDB is the largest dataset we experimented with, featuring s-BERT embeddings of size 383, in contrast to the custom CIFAR10 embeddings, which are of size 191. For this dataset, t-WaKA completed the experiment in approximately 140.07 seconds, significantly faster than LiRA, which took 1708.83 seconds. Both methods were run on an Apple M1 Pro (16 GB, 8 cores), utilizing all available cores through a single Python process managed by scikit-learn. LiRA employed 16 shadow *k*-NN models using kd-trees, while t-WaKA reused the same *k*-NN model across all security games. For fairness that, re LiRA's implementation is model-agnostic and not optimized specifically for *k*-NN classifiers.

To better understand the distinction between high self-attribution points and high test-attribution points, we conducted an experiment involving data minimization on synthetic data (see Figure 2.24 in Appendix 2.10) using a dataset generated with the `scikit-learn` library. Both test-Shapley and self-Shapley values were computed for a k-NN model with $k = 5$. Each iteration consists in removing the top 20% of highest value points, ranked by their Shapley values. The results highlights a clear distinction between the two types of Shapley values. More precisely, test-Shapley values tend to concentrate between the decision boundary and the outer edges of the data distribution while self-Shapley values consistently highlight points that lie directly on the decision boundary. This underscores the effectiveness of self-Shapley values in identifying critical points near the decision boundary, which seem to be more vulnerable to membership inference attacks in k-NNs.

2.5 Discussion and Conclusion

In this work, we have introduced WaKA, a novel attribution method specifically designed for k-NN classifiers that also functions as a MIA. WaKA leverages the 1-Wasserstein distance to efficiently assess the contribution of individual data points to the model's loss distribution. Our motivation for this paper stemmed from the realization that while membership inference and data attribution both focus on point-wise contributions, they are related to very different concepts, namely privacy and value. Indeed, membership inference aims to measure information leakage, whereas data valuation seeks to extract intrinsic insights about data utility. Recognizing this distinction, we have design an attribution method inspired by membership inference principles, one that could serve multiple purpose. The insights from this study emphasize the need for further exploration of the relationship between model parameters, data attribution and privacy.

Data Valuation. WaKA proves to be highly effective for utility-driven data minimization by identifying both low- and high-value points. This targeted approach allows for optimized dataset refinement. Our experiments further demonstrate WaKA's robustness over Data Shapley Value (DSV) in the context of imbalanced datasets. This makes WaKA a versatile and reliable tool for data valuation.

Self-attribution methods. Due of its relationship with MIAs, WaKA is particularly interesting as a self-attribution method. We also introduced the self-Shapley term as the

DSV of a point with respect to itself, a concept that has not been discussed in the literature to date. From an utilitarian point of view, self-Shapley value could be particularly appealing to individuals who prioritize their own utility over the collective utility captured by average DSV calculated using a test set, especially when their main interest is whether contributing their data directly improves their own prediction rather than the model's overall performance. For example, in the context of a financial institution, an individual may only be willing to provide their data if it directly improves the accuracy of predictions that affect them personally, such as determining their creditworthiness. Self-Shapley offers individuals a way to know if their data would be beneficial for them only, making it a valuable tool in cases where personal utility is prioritized over collective fairness (*i.e.*, favoring a greedy strategy). Self-WaKA correlates with self-Shapley, and thus it can be used for similar purposes. However, additionally Self-WaKA also brings a more direct rationale concerning membership privacy, due to its relation with LiRA. This demonstrates that privacy is not always directly correlated with value, as it largely depends on how value is defined. In the case of k-NN, we showed it is more closely tied to "self-utility" rather than overall contribution to the model's generalization performance.

Privacy insights and the onion effect. In our evaluation, we confirmed the "onion effect" previously described by Carlini and co-authors, in which removing certain data points only partially mitigates the risk of MIAs, leaving the remaining data still vulnerable. WaKA, not only corroborates this phenomenon but also provides an efficient approach to measuring privacy risk in k-NN models. This aligns with recent work by Ye et al., 2022a, which underscores that membership inference risk is influenced not only by individual data points but also by their neighborhood. If k-NN is used as part of a Machine Learning pipeline—such as combining embeddings with k-NN for downstream tasks—WaKA can offer a way to understand membership privacy risks. We leave as future work the exploration to whether self-WaKA values and WaKA influences can generalize to other types of models, such as interpreting privacy risks in the last layer of a large language model (LLM) or other deep learning architectures.

Ethical Considerations. The development of WaKA raises ethical concerns.

Adversaries could potentially exploit WaKA to selectively identify and target particularly vulnerable data points, either to enhance inference attacks or to manipulate datasets to maximize privacy leakage. However, WaKA also provides defensive capabilities, enabling practitioners to identify and remove high-risk data points or design privacy-aware data-selection strategies. It is essential that attribution insights derived from methods like WaKA be leveraged responsibly, aiming to strengthen privacy and avoid introducing new vulnerabilities.

Protection against membership inference attack. Our experiments reveal that increasing the parameter *k* in *k*-NN models has a significant effect on reducing the success rate of MIAs. We believe that this issue is under-explored in the current literature and warrants further investigation. In particular, the hybrid models employed in our study, such as the combination of s-BERT and *k*-NN used on the IMDB and Yelp datasets, suggest that the parameter *k* could be strategically use as a defense mechanism against membership inference. In addition, we believe that WaKA values could provide valuable insights for data removal task for other types of models, such as neural networks. While our current research focuses on *k*-NN classifiers, we will also investigate the extension of this approach to neural networks as future work. In particular, testing WaKA's applicability to these contexts could significantly enhance the understanding of data attribution and its impact on model robustness and privacy.

2.6 Appendix: Datasets

2.7 Appendix: Notations

Symbol	Description
$W_1(\mu, \nu)$	1-Wasserstein distance between two probability distributions μ and ν.
$\mathbb{L}$	Distribution of loss values for k-NN models trained on all possible subsets of the dataset D.
$\mathbb{L}_{-z_i}$	Distribution of loss values for k-NN models trained on subsets of D excluding point z_i.
$\mathscr{L}$	Finite set of possible discrete loss values for a k-NN classifier.
$\mathscr{F}$	Space of all k-NN models trained on subsets of D.
$\mathscr{F}_{-z_i}$	Space of all k-NN models trained on subsets of D excluding point z_i.
$\mathbb{Q}$	Uniform distribution over models in $\mathscr{F}$ trained on subsets S with $z_i \in S$.
$\mathbb{Q}_{-z_i}$	Uniform distribution over models in $\mathscr{F}$ trained on subsets S with $z_i \notin S$.
$\ell(f)$	Loss function mapping a k-NN model f to a real-valued loss.
$\ell_{\#\mathbb{Q}}$	Push-forward distribution of $\mathbb{Q}$ through the loss function ℓ.
$\ell_{\#\mathbb{Q}_{-z_i}}$	Push-forward distribution of $\mathbb{Q}_{-z_i}$ through ℓ.
$F_{\ell_{\#\mathbb{Q}}}(l)$	CDF of $\mathbb{L}$ evaluated at loss value l.
$F_{\ell_{\#\mathbb{Q}_{-z_i}}}(l)$	CDF of $\mathbb{L}_{-z_i}$ evaluated at loss value l.
$\mathscr{F}^+, \mathscr{F}^-$	Partition of $\mathscr{F}$ into subsets with losses $\geq$ or $<$ the target loss $\ell(z_i)^*$.
$\mathscr{F}^+_{-z_i}, \mathscr{F}^-_{-z_i}$	Same partition defined on $\mathscr{F}_{-z_i}$.
WaKA(z_i)	Wasserstein k-NN Attribution (WaKA) score for point z_i, i.e. $W_1(\mathbb{L}, \mathbb{L}_{-z_i})$.
t-WaKA(z_i)	Target-WaKA score for z_i, the difference in W_1 distances across the partitions $\mathscr{F}^+$ and $\mathscr{F}^-$.

Table 2.5: *Summary of notation used throughout the chapter.*

2.8 Appendix: Algorithm 1 Details and Proof

To compute the differences between the loss distributions $\mathbb{L}$ and $\mathbb{L}_{-z_i}$, we first need to calculate the number of possible subsets S of the training set D that lead to a particular loss value l. The loss is computed using the k-nearest neighbors of a test point z_t.

For each subset S, we are interested in how the inclusion of the point z_i affects the loss, which happens when z_i pushes out a point z_j from the k-th position to the $(k+1)$-th position, i.e. $\alpha_j(D) > \alpha_i(D)$.

To formalize this, we define the count $C_{-z_i}(l, z_j)$ for subsets that exclude z_i and yield a loss l as:

$$C_{-z_i}(l, z_j) = \#\left\{S \subseteq D \mid z_i \notin S, \alpha_k(S) = \alpha_j(D), \ell(z_t; S, k) = l\right\} \cdot 2^{N-1-j}.$$

where:

- $\alpha_k(S) = \alpha_j(D)$ means that the point z_j is both the j-th sorted point w.r.t z_t and the k-th nearest neighbor in the subset S,
- $\ell(z_t; S, k)$ is the loss function parameterized by the subset S and the number of neighbors k,
- The factor 2^{N-1-j} accounts for the number of possible supersets of S, considering the positions of other points in the training set.

Similarly, $C(l, z_j)$ counts the subsets where z_i is included, with z_j now at $(k+1)$-th position, and the loss is l.

$$C(l, z_j) = \#\left\{S \subseteq D \mid z_i \in S, \alpha_{k+1}(S) = \alpha_j(D), \ell(z_t; S, k) = l\right\} \cdot 2^{N-j},$$

Note that we are considering the same subsets as before but with the inclusion of z_i. The normalized contribution to the frequency of the loss value l, denoted as $\delta(l, z_j)$ and $\delta_{-z_i}(l, z_j)$, is obtained by dividing $C(l, z_j)$ and $C_{-z_i}(l, z_j)$ by the total number of possible subsets. Specifically:

$$\delta(l, z_j) = \frac{C(l, z_j)}{2^N}, \quad \delta_{-z_i}(l, z_j) = \frac{C_{-z_i}(l, z_j)}{2^{N-1}}.$$

To compare the differences between the loss distributions $\mathbb{L}$ and $\mathbb{L}_{-z_i}$, we sum the differences in the normalized contributions of the loss values in both cases:

$$\sum_{l\in\mathscr{L}}\sum_{z_j}\delta(l,z_j)-\delta_{-z_i}(l,z_j)$$

Here, $\mathscr{L}$ is the set of possible loss values $\left\{0,\frac{1}{k},\frac{2}{k},\dots,1\right\}$, and the summation over z_j accounts for all points that can occupy the k-th nearest neighbor position in the subsets.

We simplify the term inside the summation as follows:

$$\frac{C\cdot 2^{N-j}}{2^N}-\frac{C_{-z_i}\cdot 2^{N-1-j}}{2^{N-1}}=\frac{1}{2^{N-1}}\cdot\left(\frac{C}{2}\cdot 2^{N-j}-C_{-z_i}\cdot 2^{N-1-j}\right) \tag{2.9}$$

$$=\frac{1}{2^{N-1}}\left(C\cdot 2^{N-1-j}-C_{-z_i}\cdot 2^{N-1-j}\right) \tag{2.10}$$

$$=\frac{2^{N-1-j}}{2^{N-1}}\left(C-C_{-z_i}\right) \tag{2.11}$$

$$=\frac{C-C_{-z_i}}{2^j}. \tag{2.12}$$

Thus, the difference between the normalized contributions simplifies to $\frac{C-C_{-z_i}}{2^j}$, where C and C_{-z_i} are shorthand notations for the counts of subsets for a given loss value l with and without z_i, and a given point z_j. Notice that when z_j and z_i have the same label, i.e. $y_j=y_i$, the k-nearest neighbors are identical, thus $C-C_{-z_i}=0$.

By applying this simplification to the previous expression, we have:

$$\sum_{l\in\mathscr{L}}\sum_{z_j}\delta(l,z_j)-\delta_{-z_i}(l,z_j)=\sum_{l\in\mathscr{L}}\sum_{\substack{z_j\\ y_i\neq y_j}}\frac{C(l,z_j)-C_{-z_i}(l,z_j)}{2^j}.$$

The 1-Wasserstein distance between the loss distributions $\mathbb{L}$ and $\mathbb{L}_{-z_i}$ is defined as:

$$W_1(\mathbb{L},\mathbb{L}_{-z_i})=\sum_{l\in\mathscr{L}}\left|F_{\mathbb{L}}(l)-F_{\mathbb{L}_{-z_i}}(l)\right|\cdot\Delta l,$$

where $F_{\mathbb{L}}(l)$ and $F_{\mathbb{L}_{-z_i}}(l)$ represent the cumulative distribution functions for $\mathbb{L}$ and $\mathbb{L}_{-z_i}$ evaluated at the loss value l, and $\Delta l=\frac{1}{k}$ is the difference between successive loss values.

We know that the cumulative distribution functions $F_{\mathbb{L}}(l)$ and $F_{\mathbb{L}_{-z_i}}(l)$ can be described as the sum of their respective normalized contributions up to the loss value l. However, since we are ultimately interested in the differences between the normalized contributions $\delta(l,z_j)$ and $\delta_{-z_i}(l,z_j)$ (rather than computing the cumulative distributions explicitly), we can substitute the expression for the difference of these terms directly into the Wasserstein distance formula.

Therefore, instead of expressing the 1-Wasserstein distance in terms of cumulative distributions, we can express it directly as the sum of the absolute differences in the normalized contributions:

$$W_1(\mathbb{L},\mathbb{L}_{-z_i}) = \frac{1}{k}\sum_{l\in\mathscr{L}}\left|\sum_{z_j}\left(\delta(l,z_j)-\delta_{-z_i}(l,z_j)\right)\right|.$$

Substituting the earlier expression for $\delta(l,z_j)-\delta_{-z_i}(l,z_j)$, we get:

$$W_1(\mathbb{L},\mathbb{L}_{-z_i}) = \frac{1}{k}\sum_{l\in\mathscr{L}}\left|\sum_{\substack{z_j\\ y_i\neq y_j}}\frac{C(l,z_j)-C_{-z_i}(l,z_j)}{2^j}\right|.$$

2.9 Appendix: Utility-driven Data Minimization Results and Analysis

Figure 2.10: *Results for the CIFAR10 dataset across various metrics and tasks. Rows show data addition and removal tasks, while columns display Accuracy, Macro-F1, and Label Ratio Evolution (for data removal only).*

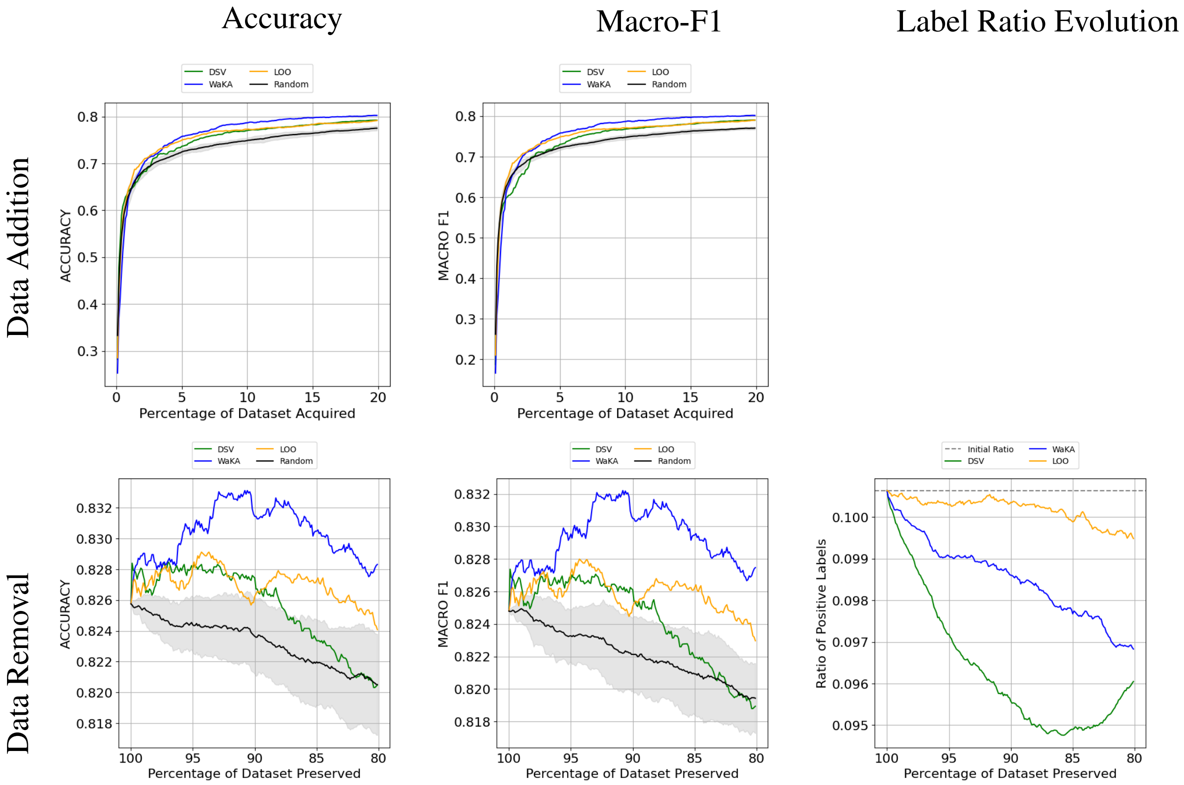


Figure 2.11: *Analysis of τ values for the CIFAR10 dataset computed using the* WaKA_{add} *and* WaKA_{rem} *formulas. τ is varied from 0.0 to 1.0 in increments of 0.2.*

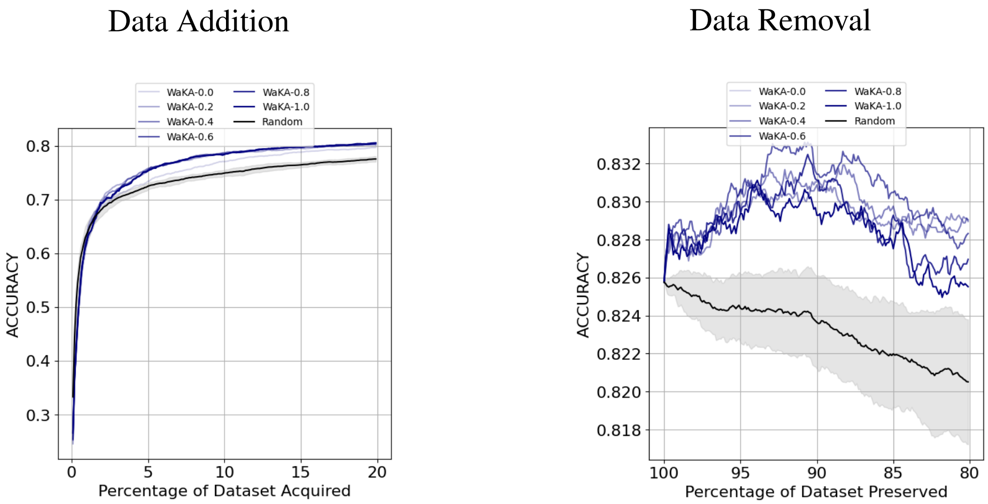


Figure 2.12: *Results for the Adult dataset across various metrics and tasks. Rows show data addition and removal tasks, while columns display Accuracy, Macro-F1, and Label Ratio Evolution (for data removal only).*

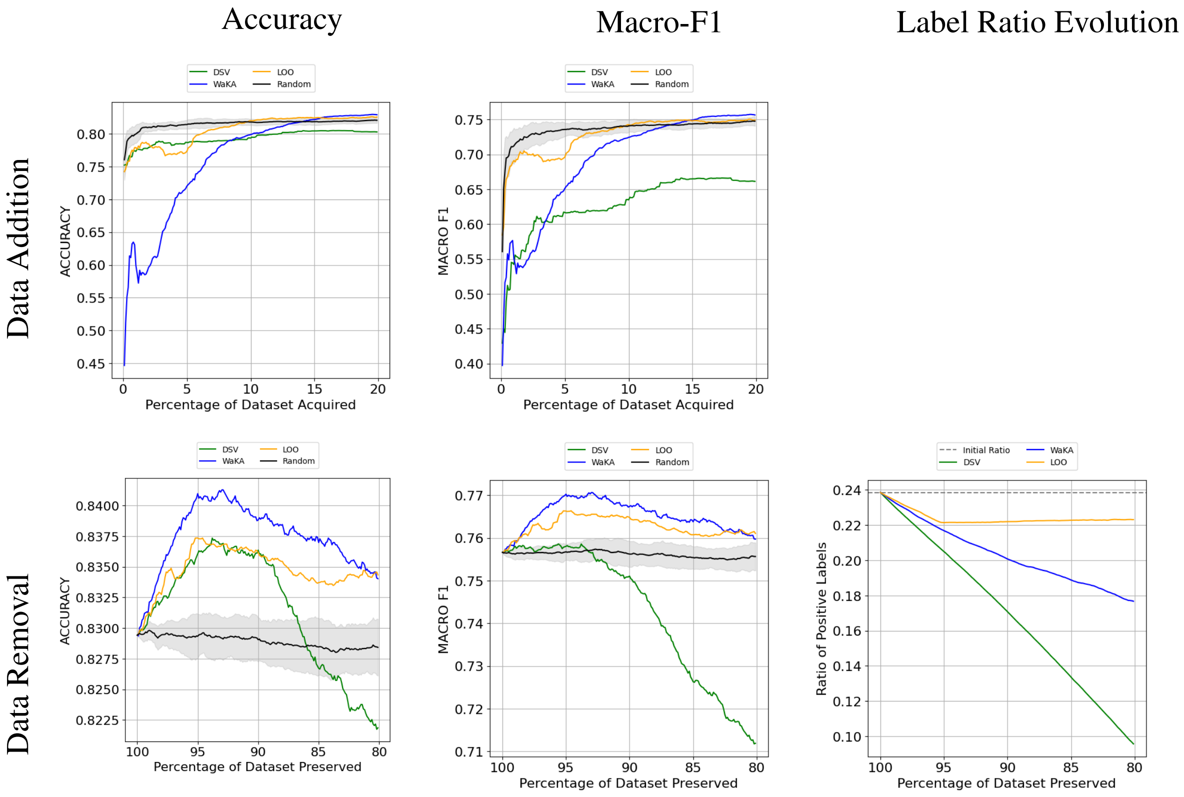


Figure 2.13: *Analysis of τ values for the Adult dataset computed using the* $\mathrm{WaKA}_{\mathrm{add}}$ *and* $\mathrm{WaKA}_{\mathrm{rem}}$ *formulas. τ is varied from 0.0 to 1.0 in increments of 0.2.*

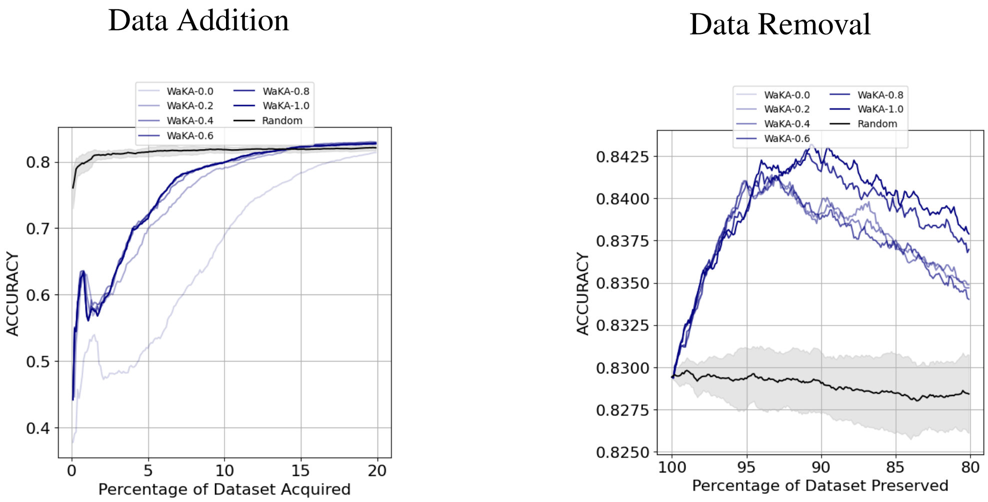


Figure 2.14: *Results for the Yelp dataset across various metrics and tasks. Rows show data addition and removal tasks, while columns display Accuracy, Macro-F1, and Label Ratio Evolution (for data removal only).*

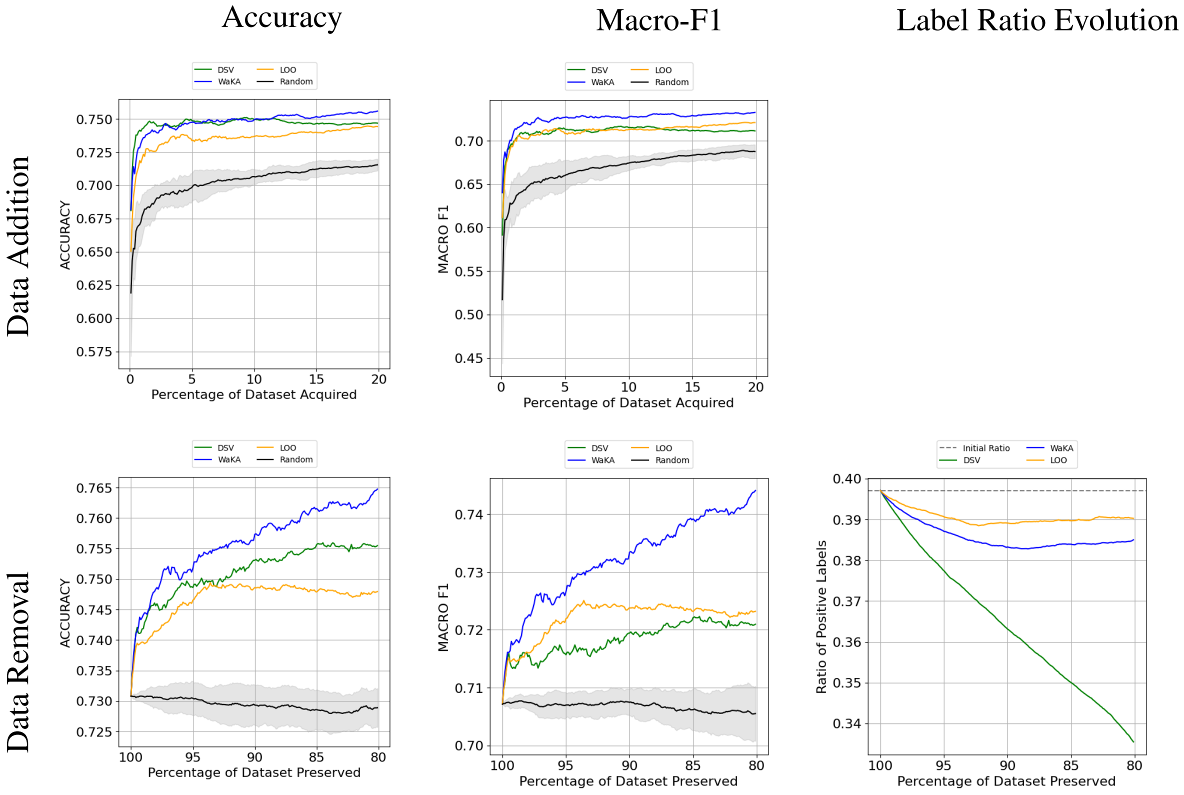


Figure 2.15: *Analysis of τ values for the Yelp dataset computed using the* $\mathrm{WaKA}_{\mathrm{add}}$ *and* $\mathrm{WaKA}_{\mathrm{rem}}$ *formulas. τ is varied from 0.0 to 1.0 in increments of 0.2.*

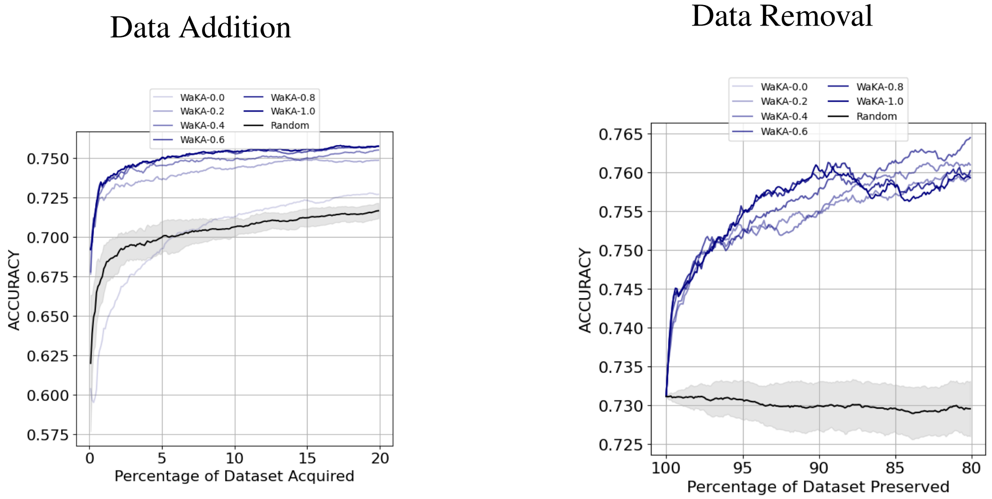


Figure 2.16: *Results for the IMDB dataset across various metrics and tasks. Rows show data addition and removal tasks, while columns display Accuracy, Macro-F1, and Label Ratio Evolution (for data removal only).*

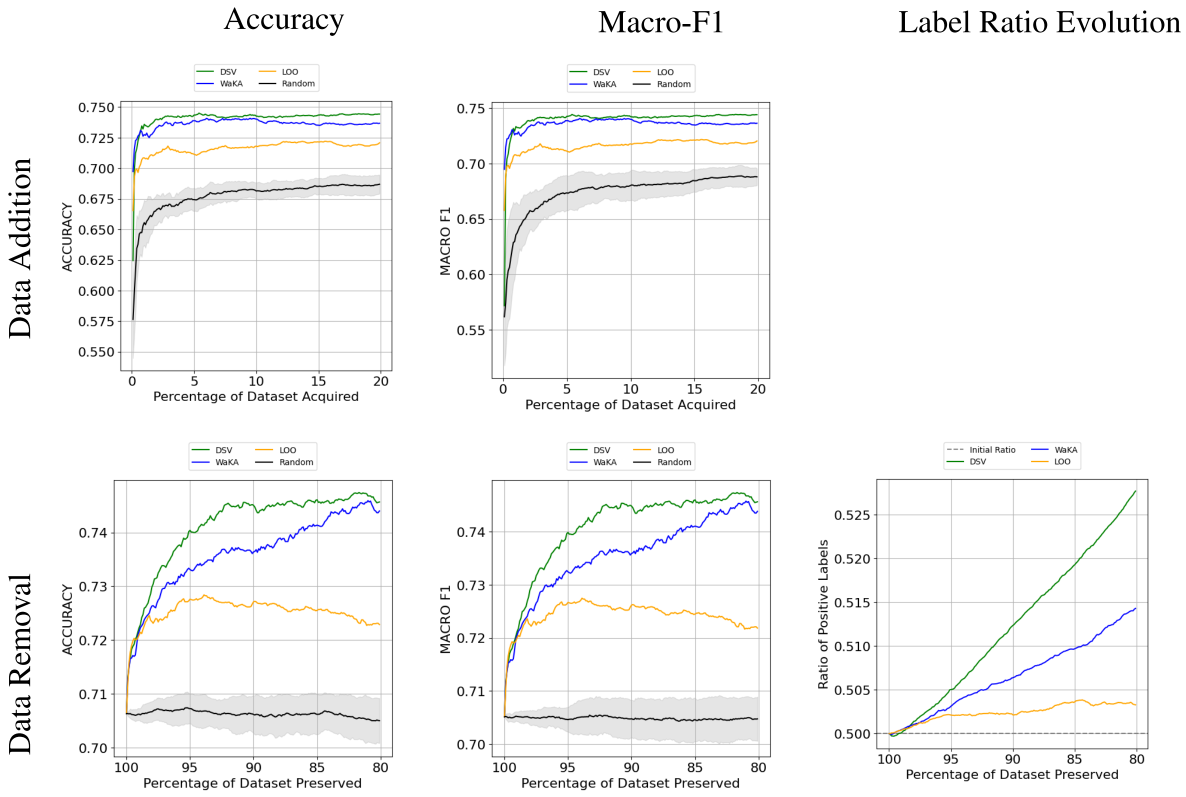
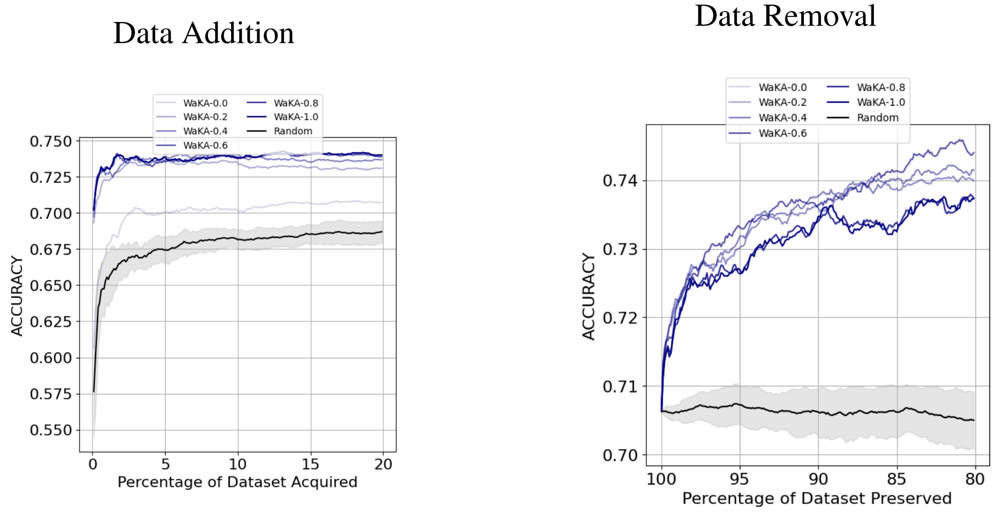


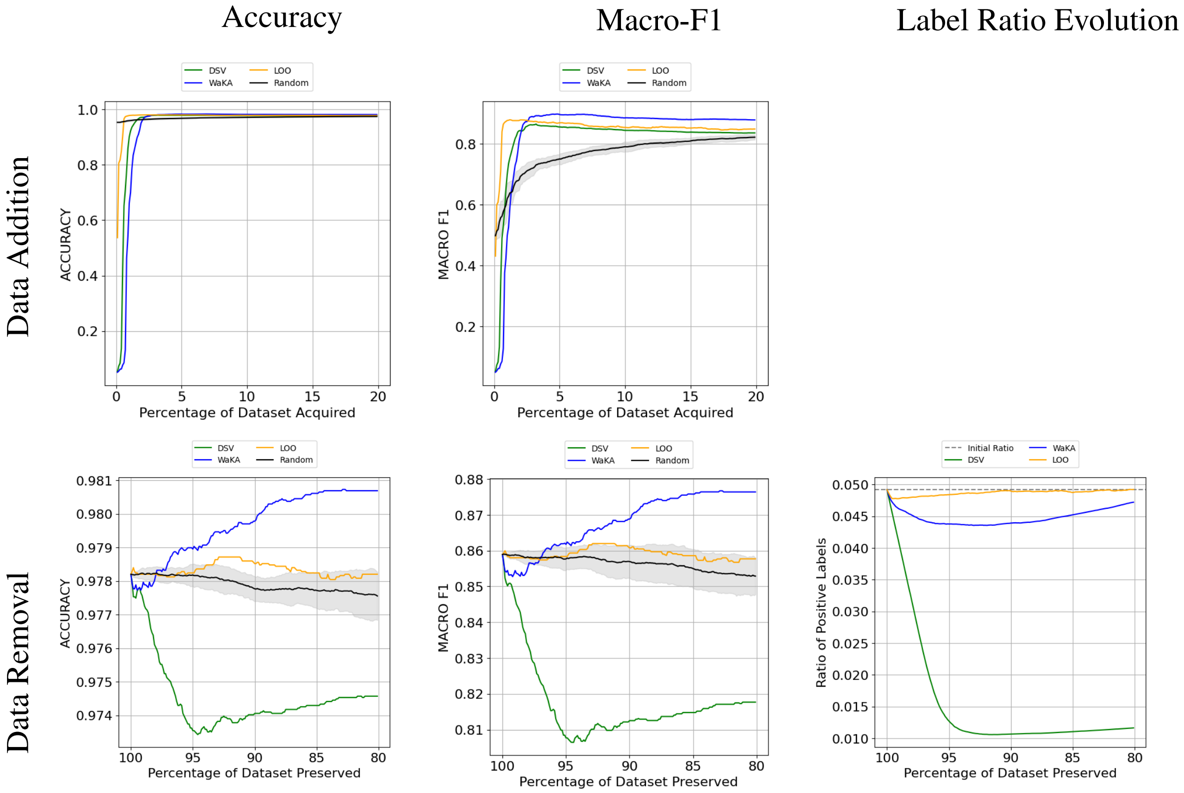
Figure 2.17: *Analysis of τ values for the IMDB dataset computed using the* WaKA_{add} *and* WaKA_{rem} *formulas. τ is varied from 0.0 to 1.0 in increments of 0.2.*



Data Addition

Data Removal

Figure 2.18: *Results for the CelebA dataset across various metrics and tasks. Rows show data addition and removal tasks, while columns display Accuracy, Macro-F1, and Label Ratio Evolution (for data removal only).*



Accuracy

Macro-F1

Label Ratio Evolution

Data Addition

Data Removal

Figure 2.19: *Analysis of τ values for the CelebA dataset computed using the* $\mathrm{WaKA_{add}}$ *and* $\mathrm{WaKA_{rem}}$ *formulas. τ is varied from 0.0 to 1.0 in increments of 0.2.*

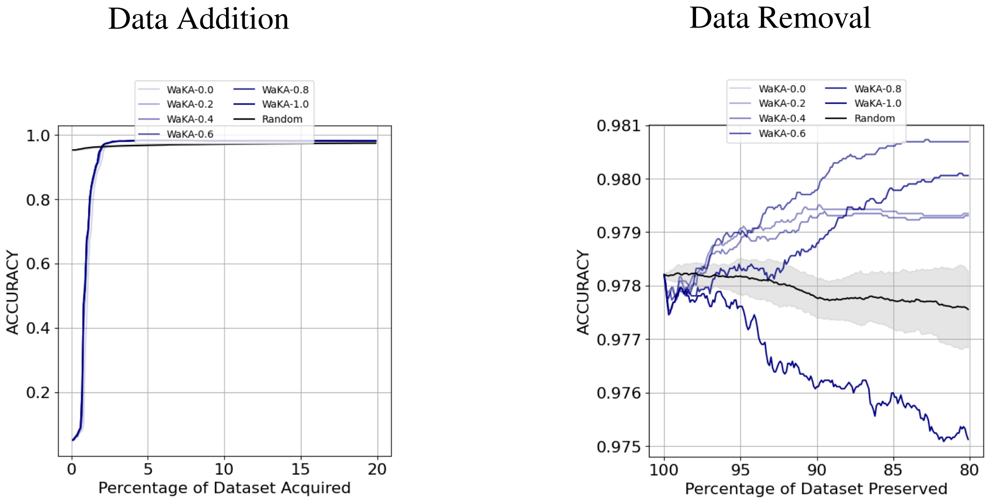


Figure 2.20: *Results for the Bank dataset across various metrics and tasks. Rows show data addition and removal tasks, while columns display Accuracy, Macro-F1, and Label Ratio Evolution (for data removal only).*

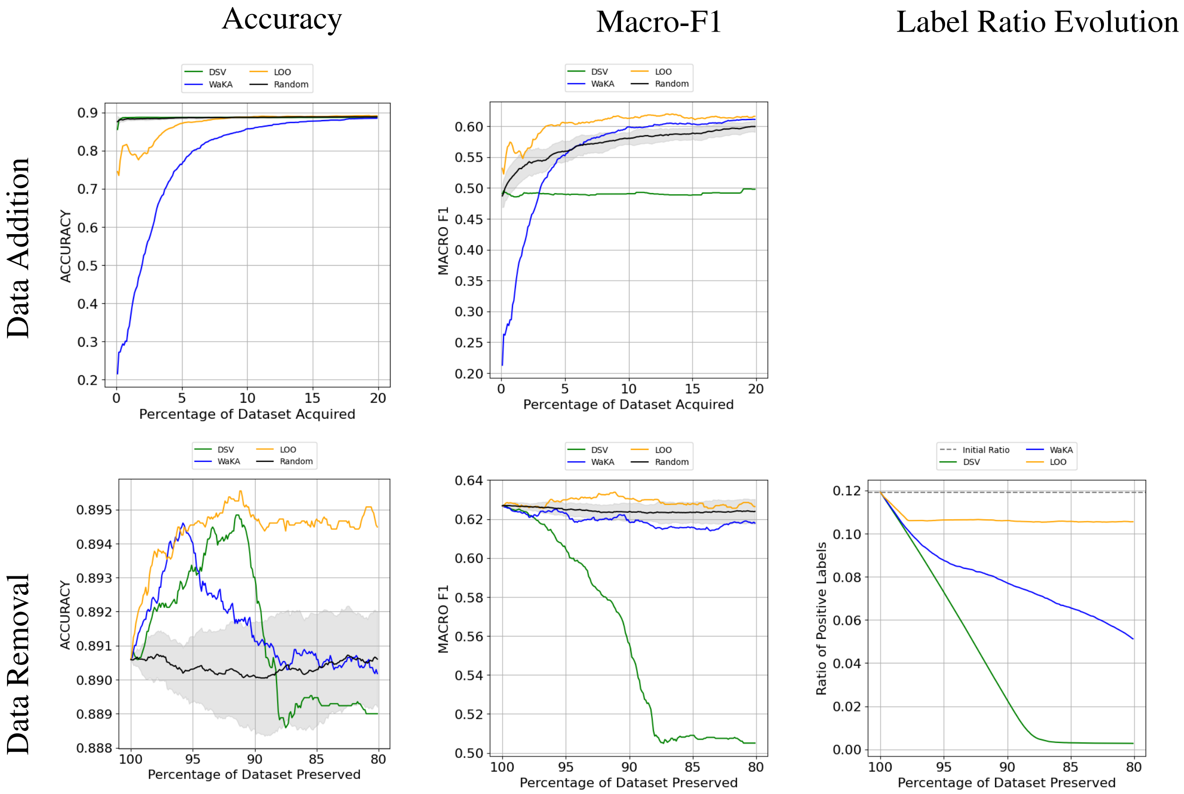
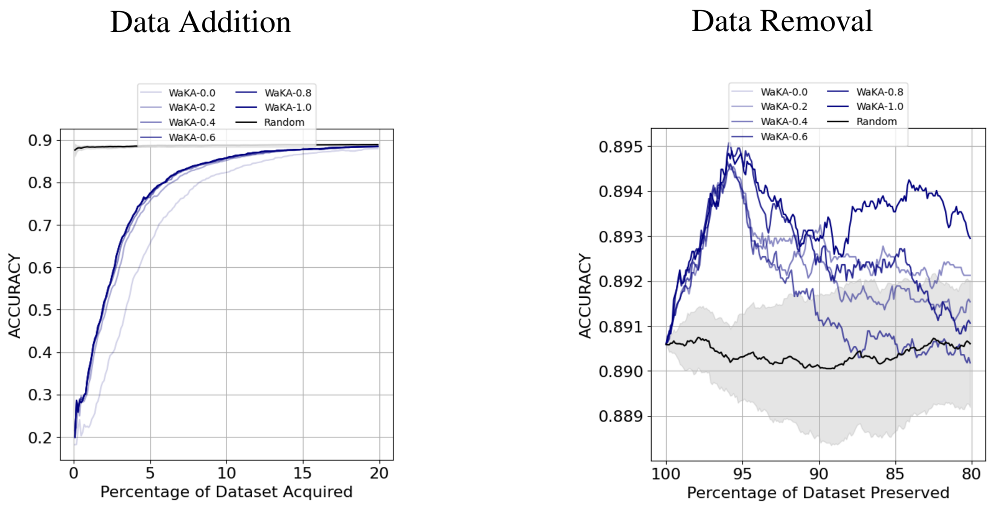


Figure 2.21: *Analysis of τ values for the Bank dataset computed using the* $\mathrm{WaKA}_{\mathrm{add}}$ *and* $\mathrm{WaKA}_{\mathrm{rem}}$ *formulas. τ is varied from 0.0 to 1.0 in increments of 0.2.*


Data Addition

Data Removal

2.10 Appendix: Privacy Evaluation and Auditing Results

Figure 2.22: *Correlation between self-WaKA values ($k = 1$) and ASR for logistic regression models. The results show a weaker correlation compared to k-NN models but a significant relationship still exists.*

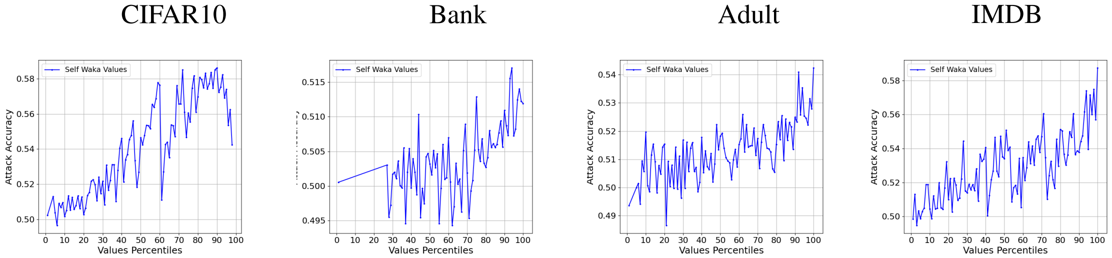


Figure 2.23: *Comparison of AUC curves in log-log scale for privacy scores at $k = 1$. The blue line represents the 100% dataset scenario while the red line shows the 90% dataset scenario.*

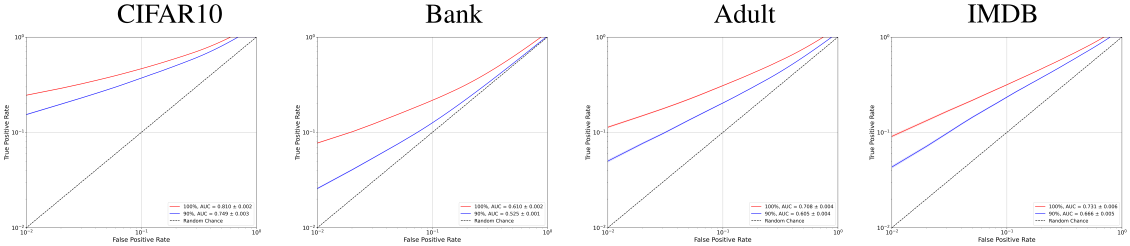


Table 2.6: *AUC of LiRA before and after removing 10% of the dataset, using various attribution methods.*

Dataset	K	AUC 100%	AUC accuracy (90%)	AUC self-waka (90%)	AUC self-shapley (90%)	AUC test-waka (90%)	AUC test-shapley (90%)
adult	1	0.731	0.610	0.605	0.604	0.696	0.704
adult	5	0.604	0.566	0.555	0.555	0.594	0.598
bank	1	0.810	0.527	0.525	0.525	0.597	0.608
bank	5	0.691	0.524	0.512	0.510	0.552	0.558
CIFAR10	1	0.708	0.752	0.749	0.747	0.819	0.826
CIFAR10	5	0.596	0.626	0.631	0.628	0.696	0.699
IMDB	1	0.610	0.671	0.666	0.666	0.740	0.754
IMDB	5	0.555	0.592	0.580	0.582	0.611	0.617

Table 2.7: *TPR at 5% FPR of LiRA before and after removing 10% of the dataset, using various attribution methods.*

Dataset	K	TPR 100%	TPR accuracy (90%)	TPR self-waka (90%)	TPR self-shapley (90%)	TPR test-waka (90%)	TPR test-shapley (90%)
adult	1	0.217	0.137	0.133	0.134	0.213	0.216
adult	5	0.080	0.076	0.069	0.072	0.074	0.082
bank	1	0.373	0.077	0.075	0.075	0.144	0.152
bank	5	0.189	0.070	0.062	0.060	0.079	0.078
CIFAR10	1	0.221	0.284	0.279	0.276	0.383	0.390
CIFAR10	5	0.081	0.123	0.125	0.123	0.197	0.199
IMDB	1	0.154	0.150	0.144	0.142	0.219	0.228
IMDB	5	0.080	0.066	0.060	0.059	0.090	0.096

Figure 2.24: *Experiment of data minimization on the synthetic data Moons generated using the scikit-learn library for a k-NN model with $k = 5$. The darker the points, the higher the corresponding Shapley values. The results show that higher test-Shapley values tend to concentrate between the decision boundary and the external boundaries of the data, while higher self-Shapley values consistently identify points lying directly on the decision boundary.*

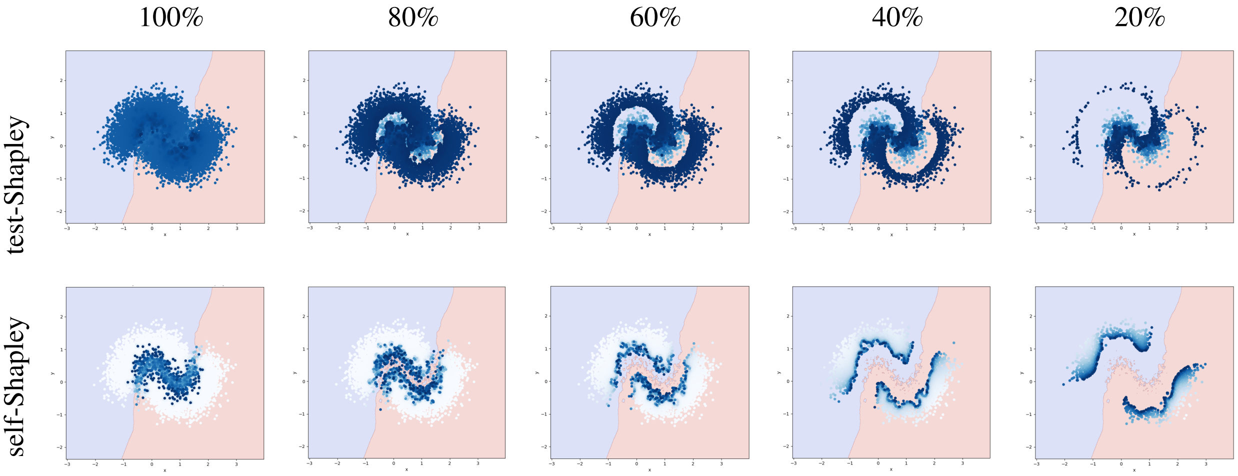


Table 2.8: *Average ASR with Standard Error for $k = 1$, $k = 5$ and Logistic Regression (LogReg)*

Dataset	**K=1**	**K=5**	**LogReg**
Adult	0.625 ± 0.002	0.561 ± 0.001	0.512 ± 0.001
Bank	0.56 ± 0.001	0.531 ± 0.001	0.503 ± 0.0004
CIFAR	0.702 ± 0.001	0.621 ± 0.001	0.541 ± 0.001
IMDB	0.65 ± 0.002	0.573 ± 0.001	0.529 ± 0.001

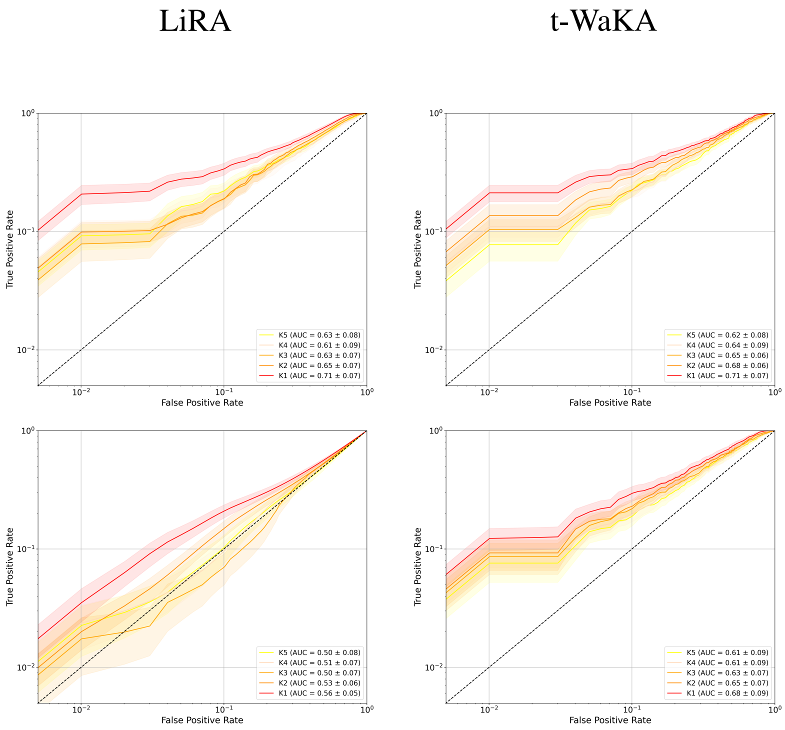


Figure 2.25: *CIFAR10: Membership Inference Attacks (MIA) AUC and ROC curves (in log-scale) using LiRA, t-WaKA, confidence, and calibrated confidence attacks for k values between 1 and 5.*

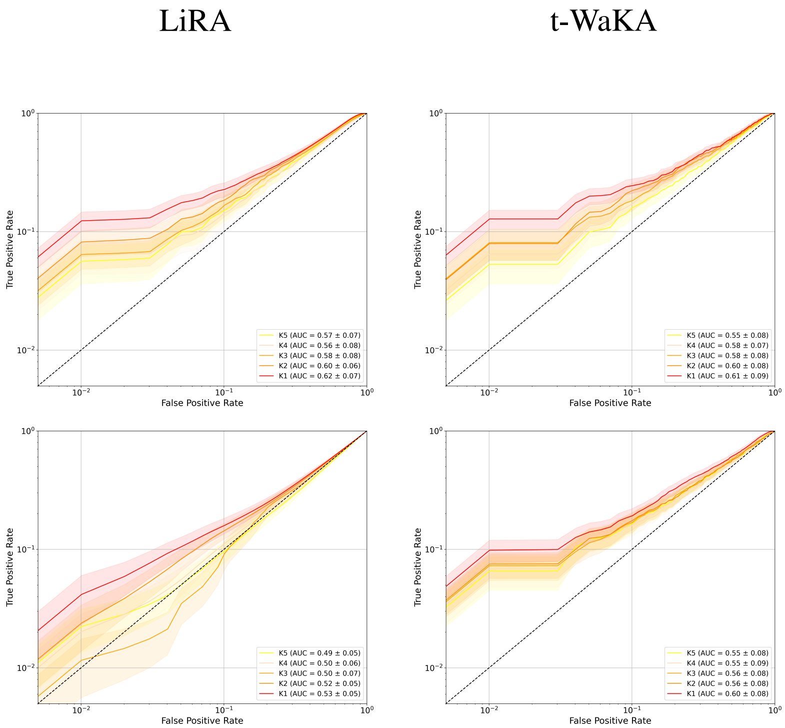


Figure 2.26: *Bank: Membership Inference Attacks (MIA) AUC and ROC curves (in log-scale) using LiRA, t-WaKA, confidence, and calibrated confidence attacks for k values between 1 and 5.*

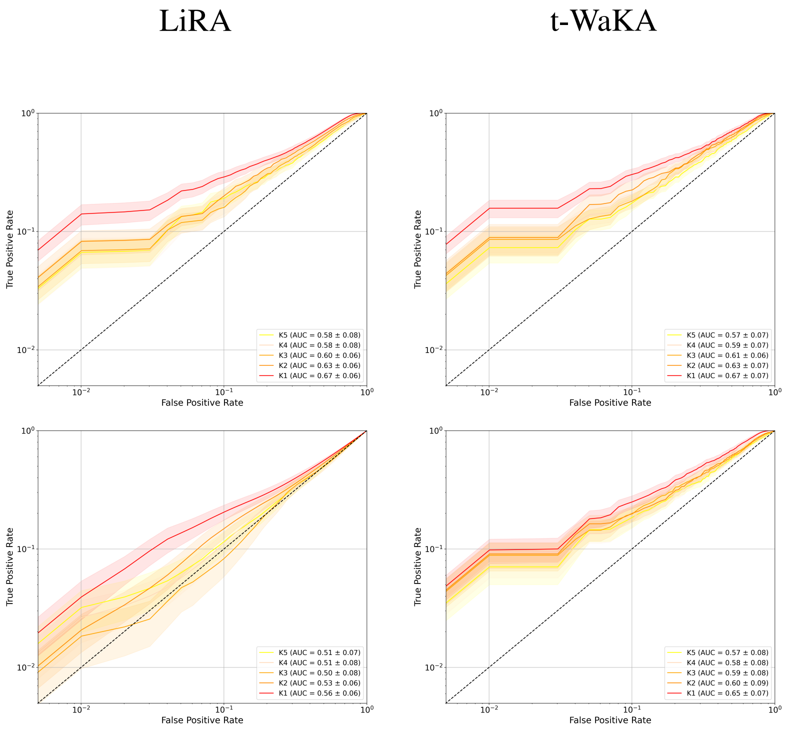


Figure 2.27: *Adult: Membership Inference Attacks (MIA) AUC and ROC curves (in log-scale) using LiRA, t-WaKA, confidence, and calibrated confidence attacks for k values between 1 and 5.*

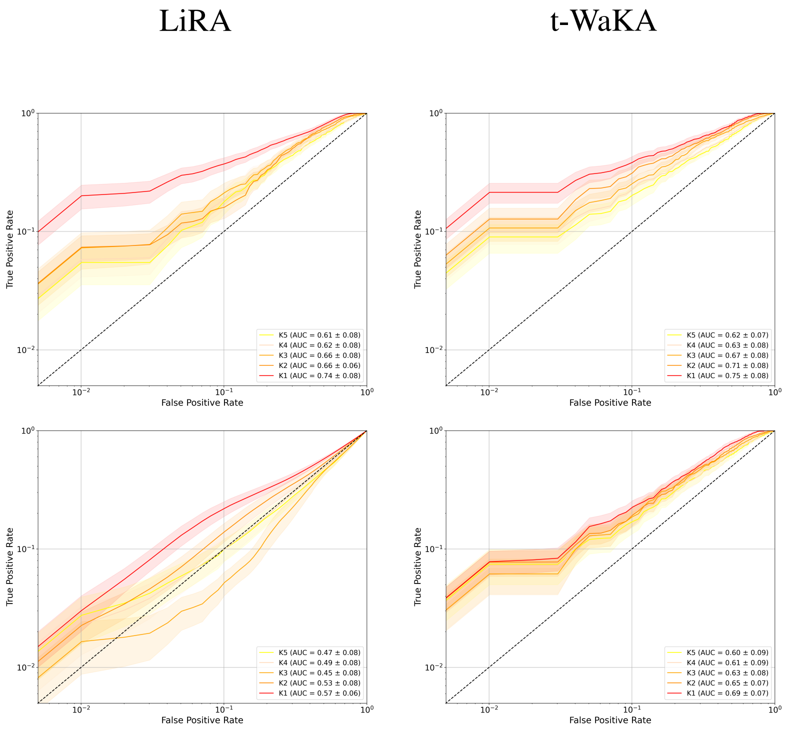


Figure 2.28: *IMDB: Membership Inference Attacks (MIA) AUC and ROC curves (in log-scale) using LiRA, t-WaKA, confidence, and calibrated confidence attacks for k values between 1 and 5.*

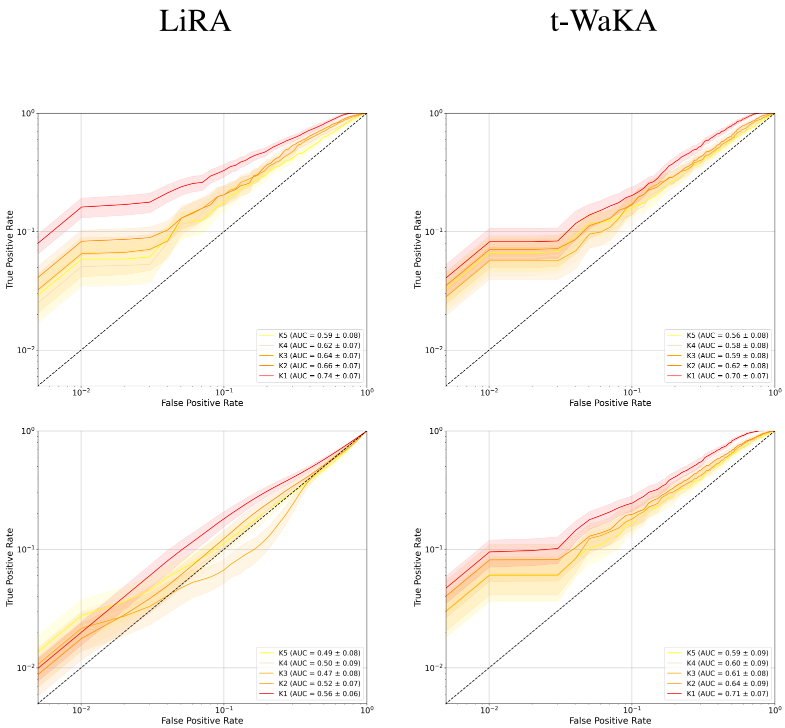


Figure 2.29: *Yelp: Membership Inference Attacks (MIA) AUC and ROC curves (in log-scale) using LiRA, t-WaKA, confidence, and calibrated confidence attacks for k values between 1 and 5.*

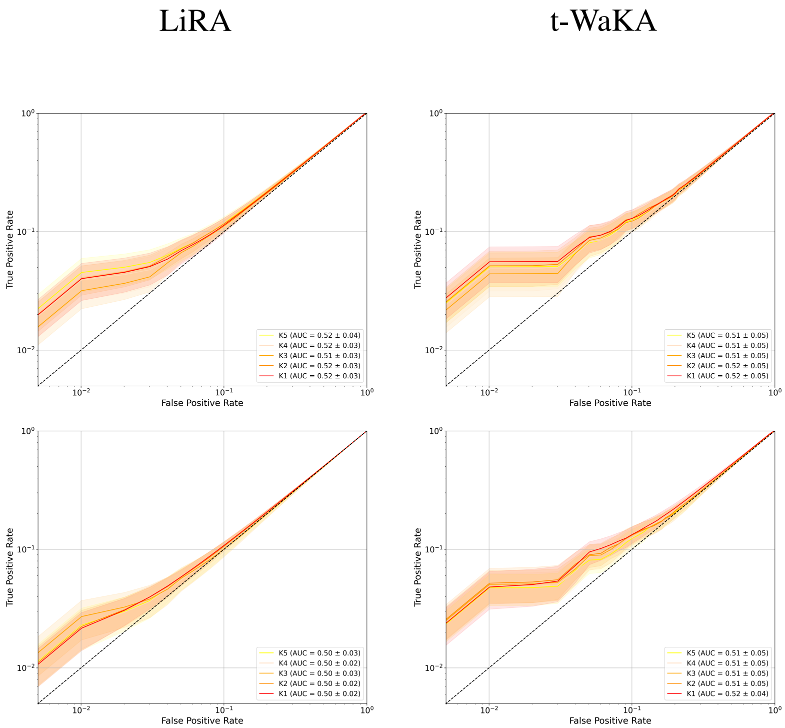


Figure 2.30: *Celeba: Membership Inference Attacks (MIA) AUC and ROC curves (in log-scale) using LiRA, t-WaKA, confidence, and calibrated confidence attacks for k values between 1 and 5.*

Chapter 3

Data Valuation via an Information Disclosure Game

Abstract [1]

Data valuation frameworks, such as data Shapley values, quantify the contribution of individual data points within a dataset to a particular predictive performance metric. As such, they can be used by organizations to redistribute value in a fair manner to individuals. However, computing value directly from contributions often results in some individuals receiving nothing, creating a catch-22 in which people may withdraw consent if their data is deemed worthless. To address this, we propose a game-theoretic framework— the *Information Disclosure Game*—between a *Data Owner* (DO) and a *Data Consumer* (DC). Instead of directly releasing the full raw data, the DO discloses information progressively using Laplacian noise under a differential privacy model. In this work, we focus on *k*-nearest neighbors (*k*-NN) due to its direct relation between individual points and predictive performance, although our framework is more generic. For this setting, we simulate plausible DC behaviors, including strategies guided by data Shapley values or multi-armed bandit exploration. In particular, we empirically

[1] This essay has been submitted to the Conference on Information and Knowledge Management (CIKM 2025). First author: Patrick Mesana (HEC Montréal); co-authors: Sébastien Gambs (UQAM) and Gilles Caporossi (HEC Montréal).

validate our approach on a Yelp review helpfulness prediction task using text embeddings, demonstrating that data valuation inherently incurs a cost, which is made explicit in the DC's budget consumption strategy.

3.1 Introduction

Data valuation is an emerging field in machine learning that seeks to address the fundamental question about the worth of data. It has recently gained significant attention with the development of Data Shapley Value (DSV, Jia, Dao, Wang, Hubis, Hynes, et al., 2019a) as well as due to its fundamental applications, such as data summarization and efficient data acquisition, as demonstrated by Ghorbani and Zou. More precisely, these applications suggest that a significant proportion of the data points in a training set are not needed to achieve good performance in terms of predictive utility. This insight also opens the door to considerations around privacy – since minimizing the amount of data used might imply a potential privacy gain — although this perspective does not account for all privacy risks, such as the "onion effect" described by Carlini, Jagielski, et al.

Nevertheless, data valuation aligns naturally with a utilitarian perspective on privacy. In data valuation frameworks, each data point is treated analogously to a rational agent seeking to maximize its own utility. This resonates closely with the approach taken by Differential Privacy (DP, Dwork, 2006), in which privacy mechanisms are designed to limit the privacy risks incurred by individual participants, thereby reducing their incentive to withhold information (McSherry and Talwar, 2007). While DP is often associated with protecting personal information, its applications extend beyond privacy. For instance, DP learning algorithms hide the influence of one training point on the model performance and bound data value estimation (Jia et al., 2021).

However, using DSV directly as the foundation for a pricing mechanism raises the following concern: if only a small subset of data points is necessary to achieve high utility, certain data points (*e.g.*, specific individuals for scenarios involving personal information), will receive no fraction of the generated value. This leads to a data

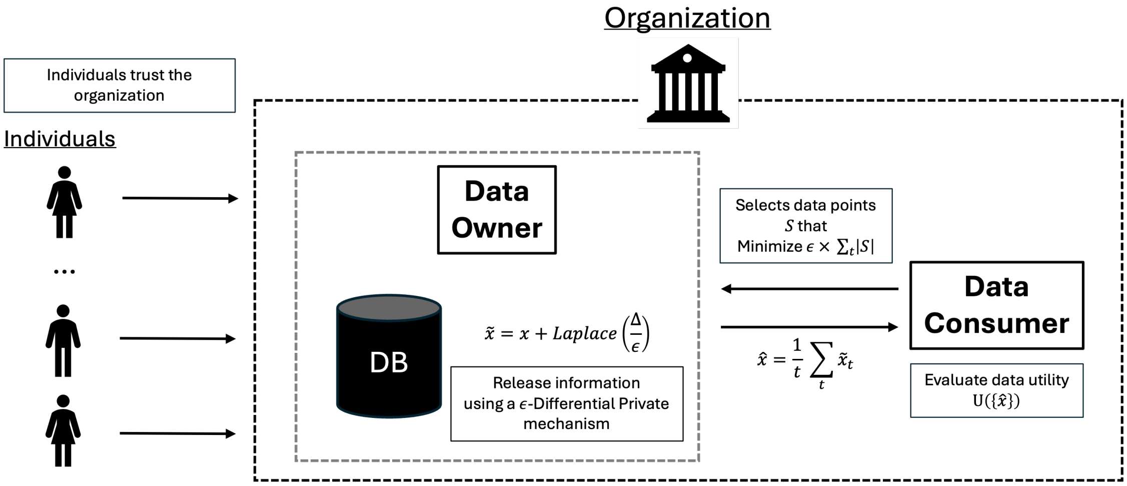

Organization
Individuals trust the organization
Individuals
...
Data Owner
DB
$\tilde{x} = x + Laplace\left(\frac{\Delta}{\epsilon}\right)$
Release information using a ϵ-Differential Private mechanism
Selects data points S that Minimize $\epsilon \times \sum_t |S|$
Data Consumer
$\hat{x} = \frac{1}{t} \sum_t \tilde{x}_t$
Evaluate data utility $U(\{\hat{x}\})$

Figure 3.1: *Illustration of the Information Disclosure Game. A Data Owner (DO) holds a private dataset and releases information to a Data Consumer (DC) by adding Laplacian noise to data points under ε-differential privacy. The DC incrementally acquires these noisy version of the data points, denoise them using an average, and train a model to reach a utility target. In this work, we focus on non-parametric k-Nearest Neighbors (k-NN).*

inclusiveness concern, as individuals that consent to share their data with an organization usually expect a benefit in return. However, if their data is not deemed valuable enough to be used, they might receive no benefits. We call this the *hidden cost of data valuation*. Specifically, identifying which data points have little or no value usually requires first collecting a sufficiently large set of data points to compute the data value estimation.

To address this paradox, we introduce a game theoretical framework with two agents: a *Data Owner* (DO) – responsible for managing data from individuals – and a *Data Consumer* (DC) – who seeks to use data to generate value, such as training machine learning models. Our inspiration comes from the tensions that can arise in organizations between data governance teams (*i.e.*, the DOs) and teams focused on creating value from data (*i.e.*, the DCs). While the DO's implicit goal is to value data in a responsible and fair manner, the DC is primarily concerned with maximizing utility, which might involve using only a subset of the data.

More explicitly, we characterize it as a Stackelberg game (Fudenberg and Tirole, 1991; Simaan and Cruz Jr, 1973), wherein the DO acts as a leader deciding what information to disclose by leveraging DP as a release mechanism, and the DC acts as a follower making optimal decisions based on the disclosed information. Figure 3.1 provides a high-level overview of the *Information Disclosure Game* (IDG).

Outline. The remainder of this paper is structured as follows. First in Section 3.2, we review the literature on cooperative game theory applied to data valuation and we review the utilitarian view of DP. Afterwards in Section 3.3, we provide an overview of our IDG framework by defining the interactions between the DO and DC as well as their respective objectives and constraints.

More precisely, the partial information disclosure process is modeled as an iterative decision-making problem from the DC point of view, in which he must balance exploration (*i.e.*, identifying high-value data points) and exploitation (*i.e.*, minimizing cost while achieving target utility). Although the DC's problem is iterative, the overall structure of the game remains a two-phase Stackelberg game. In Section 3.4, we present a concrete case study focused on review helpfulness using the Yelp dataset. To ensure computational feasibility, we use pretrained embeddings combined with a *k*-Nearest Neighbors (k-NN) classifier.

Under four different data acquisition strategies, our analysis aims to quantify the hidden cost of valuation, assess whether incentives could lead the DC to acquire data points different from those favored by Shapley-based approaches, and examine how the addition of noise affects privacy leakage from the resulting models. In Section 3.5, we discuss extensions to differentiable models and private learning mechanisms. Finally, in Section 3.6, we summarize our contributions and suggest directions for future research.

3.2 Background

3.2.1 Data Valuation

In cooperative game theory, an *imputation* is a way to distribute the total value of a game among its players (Osborne and Rubinstein, 1994).

In the context of data valuation, each data point in a dataset D is treated as a *player*, and a *value function* $v : 2^D \to \mathbb{R}$ measures the utility (*e.g.*, accuracy or loss reduction) of any subset $D' \subseteq D$. The objective is to find an imputation that fairly allocates the total value $v(D)$ among individual data points based on their contributions. Several approaches have been explored for data valuation—each with different fairness, computational and interpretability trade-offs—including *Data Shapley* (Ghorbani and Zou, 2019; Jia, Dao, Wang, Hubis, Hynes, et al., 2019a), *Data Banzhaf* (J. T. Wang and Jia, 2023), *Beta Shapley* (Kwon and Zou, 2022) and the *Core* (Yan and Procaccia, 2021).

Data Shapley Value. Formally, the Shapley value of a data point $z_i \in D$ is given by:

$$\phi_{z_i}(v) = \mathbb{E}_{D' \sim \mathscr{P}(D \setminus \{z_i\})} \left[v(D' \cup \{z_i\}) - v(D') \right],$$

in which D' is a subset of $D \setminus \{z_i\}$ sampled uniformly at random and $v(D')$ is the value function evaluated on D'. The Shapley value thus quantifies the *expected* incremental contribution of each data point across all possible subsets. The uniqueness of the Shapley value can be derived from satisfying four foundational axioms: *Efficiency* (*i.e.*, the total value is distributed among all players), *Symmetry* (*i.e.*, identical contributions are rewarded equally), *Dummy* (*i.e.*, players that contribute nothing receive zero value) and *Linearity* (*i.e.*, the Shapley values from two games can be combined linearly).

Data valuation methods are primarily evaluated through tasks such as data removal and data acquisition, in which data points are ranked based on their values to guide the removal of less valuable ones or the acquisition of the most valuable ones. A key assumption in these settings is the presence of a central entity, such as a curator or custodian, responsible for data collection and aggregation. This assumption is similar to the global setting of DP. One of the significant applications of data valuation is pricing in data marketplaces (Jia, Dao, Wang, Hubis, Hynes, et al., 2019b; Pei, 2022; Tian et al., 2022; Xia et al., 2023), in which the Shapley value framework provides a theoretically grounded approach to determine data prices, with the underlying hypothesis being that data points contributing to multiple subsets should lead to a higher value, reflecting the demand of the task at hand.

Data selection. Methods like data Shapley can serve as a foundation for data selection strategies. Indeed, by efficiently estimating which examples are most beneficial, practitioners can curate a smaller high-quality training set. However, its effectiveness relies on the structure of the utility function and is particularly well-suited when utility follows a monotonically transformed modular form (J. T. Wang, Yang, et al., 2024). A notable challenge in making data valuation practical for selection is also the high computational cost of estimating data values, especially for models with high capacity. Methods like G-Shapley (Ghorbani and Zou, 2019) tackle this by leveraging gradient-based approximations instead of costly combinatorial subset evaluations. Similarly, reinforcement learning approaches (Yoon et al., 2020) offer another way to optimize data valuation dynamically during training. For k-NN classifiers, an exact formulation of the Shapley value exists, as proposed in Jia, Dao, Wang, Hubis, Gurel, et al., 2019, which drastically reduces the complexity to $O(N\log N)$, for N the number of data points. This exact formulation avoids the need to evaluate individually all subsets by exploiting the k-NN structure.

3.2.2 Differential Privacy

Differential Privacy (DP, Dwork, 2006) provides a mathematically rigorous privacy framework by ensuring that the inclusion or exclusion of any individual in a dataset does not significantly alter the probability distribution of outputs. For instance, the trade-off between privacy and utility can be controlled through noise addition, thus reducing the re-identification risk and providing provable privacy guarantees.

Let $D = \{z_1, z_2, \ldots, z_n\}$ be a dataset consisting of n individual data points. Formally, a randomized mechanism M satisfies (ε, δ)-differential privacy if, for any dataset D, any data point $z_i \in D$, and any subset of possible outputs O, the following holds:

$$\Pr[M(D) \in O] \leq e^{\varepsilon} \Pr[M(D \setminus \{z_i\}) \in O] + \delta,$$

in which ε represents the privacy loss and δ allows for a small probability that strict ε-differential privacy does not hold. A key feature of DP is its **composability** property, which states that if all individual mechanisms in a system satisfy differential privacy, then their combination also satisfies differential privacy (Dwork and Roth, 2014a).

This modular property enables the design of complex privacy-preserving mechanisms and allows to set up trade-offs between privacy and accuracy, helping organizations choose mechanisms optimizing utility while maintaining strong privacy guarantees (McSherry and Talwar, 2007).

Utilitarian View. Privacy concerns are not solely about protection; they also involve the incentives individuals face when deciding whether to share their data. A fundamental concept in mechanism design is the *Revelation Principle* (Myerson, 1983), which asserts that if an outcome can be achieved through any mechanism (truthful or not), there exists a direct revelation mechanism in which agents truthfully reveal their private information to achieve the same outcome. The principles of truthfulness in mechanism design also have direct applications in privacy-preserving mechanisms. For instance, as put forward by McSherry and Talwar, privacy-preserving mechanisms can be viewed as social choice games in which privacy is modeled as an explicit cost, similar to the cost of lying. DP (Dwork, 2006) plays a key role in these settings, offering guarantees of approximate truthfulness. In this perspective, agents are incentivized to participate truthfully because their data is protected with formal guarantees that limit how much information can be inferred about them from the output.

Although cooperative game-theoretic approaches to data valuation offer valuable insights and practical tools, they do not take into account the cost of valuation nor the incentives from individuals. To address this concern, we introduce a framework that explicitly models these antagonistic objectives, offering a different perspective on data valuation.

3.3 Information Disclosure Game

System Model. We focus on modeling the interactions between two agents: a *Data Owner (DO)* and a *Data Consumer (DC)*. The DO has the broad mission to valorize data (*e.g.*, via analytics or monetization), ensuring that data receives appropriate valuation while accounting for concerns such as individual privacy. Meanwhile, the DC seeks to optimize utility—often by acquiring data points at minimal cost. In our framework, pricing mechanisms are designed to influence the DC's strategy, enabling the DO to shape acquisition behaviors in a way that balances utility and inclusiveness considerations. These interactions are formalized as a two-phase Stackelberg game, in which the DO acts as a leader by setting how to disclose information and at what price, and the DC, as a follower, responds through strategic data acquisition.

To begin, consider the scenario of complete information disclosure, which leads to implementation challenges if the objective is to achieve higher data inclusiveness compared to the Shapley valuation method. In this setting, the DO provides raw data points (*i.e.*, without the addition of noise), offering full utility per data point to the DC. All acquired data points are released simultaneously in a single batch, limiting the DO's ability to influence acquisition dynamics. The interactions between the DO and the DC can be modeled using microeconomic principles, focusing on a pricing strategy that influences the DC's purchasing decisions. The DO aims to maximize the DC's minimized total cost by setting prices, considering the DC's optimization behavior. To remain relatively close to the Shapley valuation, we assume a simple pricing structure that uses the Shapley value as a base price while incorporating a privacy cost as a function of this base price. We further assume that the DO communicates the Shapley values of all data points to the DC at no cost, providing guidance for the DC's acquisition decisions.

The DO sets a price p_i for each data point z_i, which is composed of two components:

$$p_i = b_i + c_i,$$

in which b_i is the *base price* of data point i, representing its intrinsic value while c_i is the *privacy cost* associated with data point i, reflecting the individual's additional demand for compensation when aware of the data's value to the DC. The base price b_i is set to be equal to the ϕ_{z_i}:

$$b_i = \phi_{z_i}.$$

Recall that DSV quantifies the contribution of each data point to the overall utility of the model, serving as a fair baseline price. In its general form, we can model the privacy cost c_i as a function of the base price and a parameter vector θ that controls how the privacy cost scales with respect to the base price:

$$c_i = f(\phi_{z_i}, \theta),$$

DC's Objective. Given the prices p_i, the DC aims to minimize the total cost of purchasing data points while achieving at least the target utility U_{target}:

$$\begin{aligned}
\min_{\{x_i\}} \quad & C_{\text{DC}} = \sum_i p_i x_i \\
\text{s.t.} \quad & \sum_i \phi_{z_i} x_i \geq U_{\text{target}}, \\
& x_i \in \{0,1\}, \quad \forall i.
\end{aligned}$$

DO's Objetive. The DO seeks to maximize the DC's minimized total cost by choosing the optimal γ and θ, anticipating the DC's response:

$$\max_{\theta} \left\{ C^*_{\text{DC}} = \min_{\{x_i\}} \sum_i p_i x_i \right\},$$

The DO adjusts θ to influence the DC's data selection, ensuring feasibility. Assume for instance that the cost function has a simple form such as $f(\phi_{z_i}) = \gamma \phi_{z_i}^{\theta}$, the acquisition price of data point i becomes $\phi_{z_i} + \gamma\phi_{z_i}^{\theta}$. When the privacy cost is zero or constant, the cost-effectiveness ratio per point does not change because the price for a point is either equal or proportional to ϕ_{z_i}. The optimal selection process involves ranking data points in descending order of their value and selecting the smallest subset such that:

$$\sum_{j=1}^{k} v_j \geq U_{\text{target}}.$$

Only when $\theta > 1$, that DO can influence DC by introducing a non-uniform cost-effectiveness across data points, in which higher-value points incur disproportionately larger costs. Although this approach provides a straightforward pricing mechanism for releasing data, it has several significant limitations. First, releasing the complete data set compromises privacy, as individuals have their exposure to privacy risks increased as well as a diminishing control over their personal information. Second, employing a static valuation method locks in the data's value at a single point in time, failing to account for future shifts in utility or opportunities for renegotiation. Lastly, despite the ability of pricing parameters to guide the acquisition of certain data points, the binary nature of these acquisition decisions inevitably results in some data points being excluded, raising concerns about data inclusiveness. To address these challenges, we propose to depart from classical pricing models by incorporating the privacy costs into the points themselves through the use of DP in an iterative release process.

Partial Information Disclosure Game. In this setting, the DO aims to incentivize the DC by adding noise to the data or learning process at each iteration, using a fixed differential privacy budget per data point per iteration, denoted as ε. Crucially, the DC does not know the Shapley values of the data points *a priori* and must infer their value over time through the noisy feedback received during the iterative acquisition process. The DO faces two strategic decisions: (1) determining the privacy budget ε allocated for querying a single data point and (2) establishing the total number of queries T that can be made on each data point. The DC attempts to minimize the total budget spent to achieve the target utility while the DO aims to maximize this budget.

At each iteration t, the DC selects a subset of data points S_t to query or use in the learning process. For each selected data point $z_i \in S_t$, the DC incurs a cost associated with the differential privacy budget ε used to obtain the noisy version of the data point or to perform a differentially private update. If noise is added directly to the data, the mechanism is similar to Local Differential Privacy (LDP) Cormode et al., 2018; N. Wang et al., 2019, with the exception that it is applied in a centralized setting since the DO has already collected the data. In this scenario, noise is added directly to individual data points before they are used in the learning process. A benefit of this setting is that it does not require any specific model type, making it particularly suitable for non-parametric models such as k-NN, which rely on access to raw data points for making predictions. Additionally, since the data is centralized, noise can be calibrated based on the local sensitivity of the dataset, which measures the maximum impact that a single data point can have on a given function, allowing for more precise noise addition compared to traditional LDP.

By configuring the maximum budget B_{MAX}, the DO directly shapes the trade-offs between privacy and utility within the learning process. A higher value of ε allows for more accurate results by reducing noise but comes at the cost of increased privacy in the form of information leakage. Conversely, a lower ε enhances privacy but might degrade model performance. Similarly, the total number of iterations T directly affects how frequently data points are accessed. The budget spent by the DC over all iterations is:

$$B = \varepsilon \times \sum_{t=1}^{T} |S_t|,$$

in which ε is the fixed differential privacy budget per data point per iteration, $|S_t|$ is the number of data points selected by the DC at iteration t and finally T is the total number of iterations until the DC achieves the target utility U_{target} or reaches T_{MAX}. By setting ε and

T_{MAX}, the DO establishes the maximum budget B_{MAX} per point, which is determined as:

$$B_{\text{MAX}} = T_{\text{MAX}} \times \varepsilon$$

The DC is confronted with a data selection optimization problem, his objective being to select the most valuable data points to achieve a predefined utility target while spending the least budget possible.

DC's Objective. Formally, the DC aims to minimize the budget spent while achieving the target utility, which can be formalized by the following optimization problem:

$$\min_{\{S_t\}} \quad \varepsilon \times \sum_{t=1}^{T} |S_t|,$$

subject to:

$$U^{(T)} \geq U_{\text{target}} \quad \text{and} \quad T \leq T_{\text{MAX}},$$

in which ε is the privacy budget per data point per iteration, $|S_t|$ is the number of data points selected at iteration t and T is the number of iterations required to meet the target utility U_{target}.

DO's Objective. The DO's objective can be framed as a max-min problem again, in which he seeks to maximize the minimum budget spent by the DC under the constraints of the DC's behavior while also ensuring a balanced spread of spending across data points. We assume that $T_{\max}$ plays a key role in this distribution as increasing $T_{\max}$ allows for more iterations, potentially leading to a more even allocation of the budget across data points and increasing overall spending. By controlling ε, the DO can further adjust the trade-off between privacy and utility as a higher ε allows for more accurate data contributions (less noise) but increases information leakage, whereas lowering ε raises the DC's costs. The joint optimization of ε and $T_{\max}$ enables the DO to shape the DC's acquisition strategy and effectively balance privacy and utility.

3.4 Review Helpfulness Prediction Case Study

As review platforms, such as Yelp or Amazon, increasingly prioritize high-quality content and services, the ability to assess and reward helpful reviews becomes strategically important. We rely on the dataset released by Yelp Asghar, 2016 for this case study, which includes text reviews along with helpfulness votes. This task is well-suited for our data valuation framework, as it involves clear individual contributions to predictive performance. More precisely, we will use as reference the work of Bilal and Almazroi, who used this dataset and reported a k-NN classifier accuracy of 59.6%. We use this result as a benchmark for evaluating the effectiveness of our information disclosure game using a similar k-NN model. Additionally, to improve upon this baseline, we evaluated two recent sentence embedding models: `sentence-transformers/all-mpnet-base-v2` and `intfloat/multilingual-e5-large-instruct`. Both models were selected for their strong performance on semantic similarity tasks while also ensuring that they have not been trained on the Yelp dataset to avoid indirect data leakage. Reviews were encoded using each model, and we trained a k-Nearest Neighbors (k-NN) classifier, tuning the number of neighbors k based on validation performance. The best results were obtained with the E5 model for which at $k = 67$, we achieved a validation accuracy of 69.60%, a test accuracy of 66.00% and a test F1-score of 65.21%. This configuration brings k-NN performance closer to that of fine-tuned transformer models while preserving the point-level interpretability necessary for data valuation analysis. We also used the same dataset split—8000 training, 1000 validation and 1000 test examples.

We replicated the experiment from Jia, Dao, Wang, Hubis, Hynes, et al. to verify that Shapley-based data acquisition outperforms random sampling on test accuracy. Figure 3.2 compares the test performance as data points are incrementally acquired using k-NN Shapley values versus a random strategy. In this experiment, accuracy improves more quickly under Shapley selection, confirming its effectiveness. In addition, the performance plateaus between 40% and 60% of the dataset, indicating diminishing returns beyond that point.

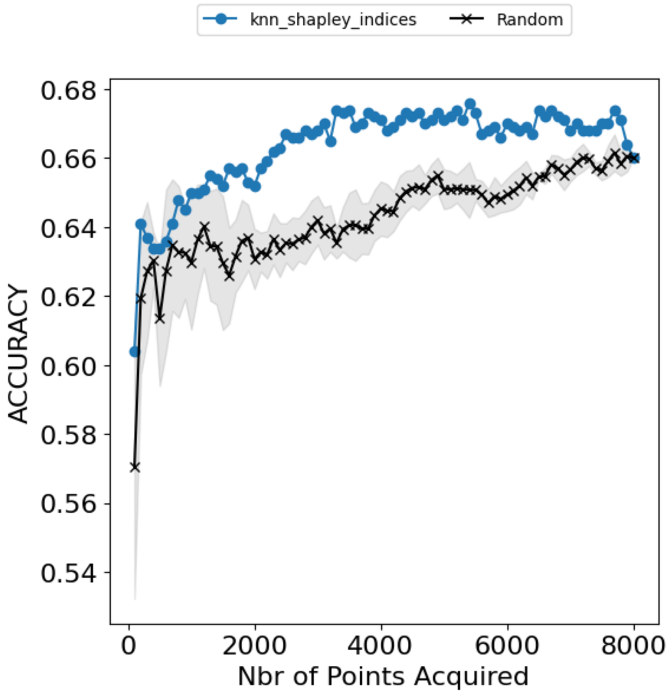


Figure 3.2: *Test accuracy as a function of data points acquired, comparing k-NN Shapley-based selection (blue) with random selection (black).*

To implement a differentially private iterative release mechanism, we have added Laplacian noise independently to each feature of every data point, each point being represented as a 1024-dimensional feature vector. We have allocated a privacy budget of $\varepsilon = 1$ per feature, resulting in a total budget of 1024 per point. The Laplace mechanism applied to a feature x_j follows:

$$\tilde{x}_j = x_j + \text{Laplace}\left(\frac{\Delta_j}{\varepsilon_j}\right),$$

in which Δ_j is feature j's sensitivity computed as the empirical range $\max(x_j) - \min(x_j)$ and ε_j is the allocated budget for that feature. Noise scales were determined by the ratio Δ_j / ε_j.

On the DC side, each noisy version of a point is averaged to form a denoised estimate. After observing t noisy versions of point i, its center is computed incrementally as:

$$\hat{x}_i^{(t)} = \frac{1}{t} \sum_{k=1}^{t} \tilde{x}_i^{(k)}.$$

This simple averaging strategy improves fidelity over time as more noisy samples are gathered. Figure 3.3 shows the average Euclidean distance between the true point and its denoised center. As expected, the error decreases non-linearly with more iterations. Figure 3.4 shows the Spearman correlation between original k-NN Shapley values and those computed on the noisy centers. Correlation improves steadily but also follows a non-linear trajectory with respect to the number of complete iterations (*i.e.*, one pass over all points).

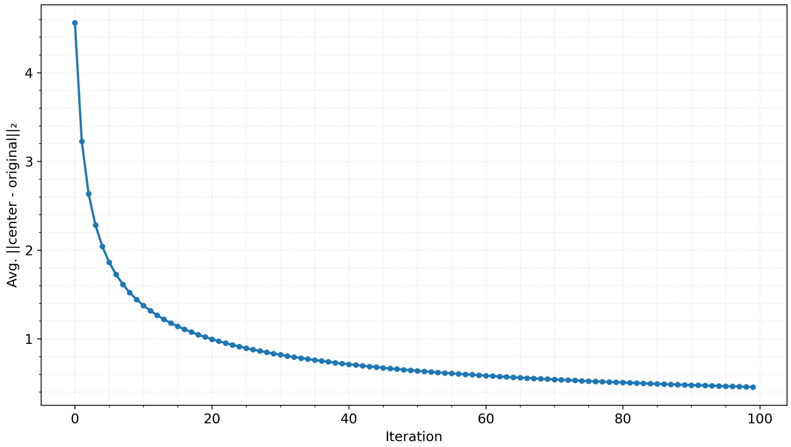


Figure 3.3: *Average ℓ_2 distance between original points and their denoised centers as a function of iterations. Error drops sharply early on with diminishing returns over time.*

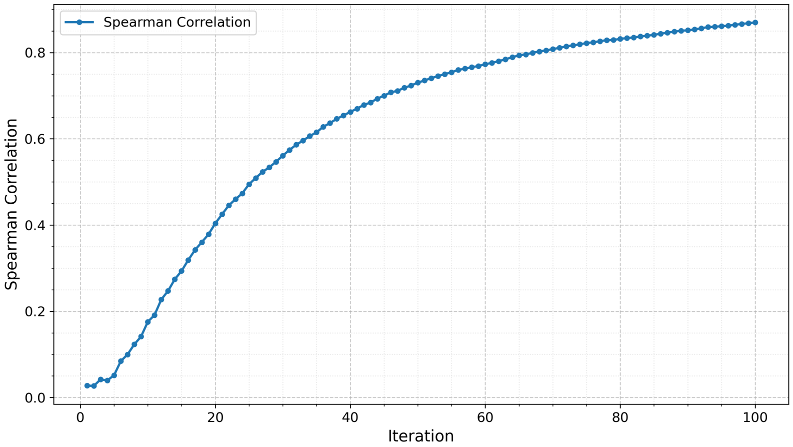


Figure 3.4: *Spearman correlation between original Shapley values and those computed on denoised centers over iterations.*

In the remainder of our analysis, we do not attempt to solve the full game between the

DO and DC. Instead, we simulate the DC's behavior by implementing and comparing different data selection strategies under the noisy iterative release mechanism. The next two subsections explore these strategies: one based on rankings (random and Shapley-based) and another using adaptive n-armed bandit algorithm. In our experiments, we define the utility target as the validation accuracy (0.696) achieved when training the k-NN model on the full dataset. For all data selection strategies, we use a fixed privacy budget ε per point per query and assume a constant maximum budget per data point.

3.4.1 Data Selection Using Random and Shapley-based Strategies

Random Data Selection. In this baseline, the DC commits to a random subset of data points for all iterations with Figure 3.5 showing the results averaged over 10 random permutations. As shown in Figure 3.5, this approach fails to reach the target utility target within 100 iterations unless nearly 100% of the dataset is acquired, which confirms that a purely random acquisition strategy from the DO is not viable under budget constraints.

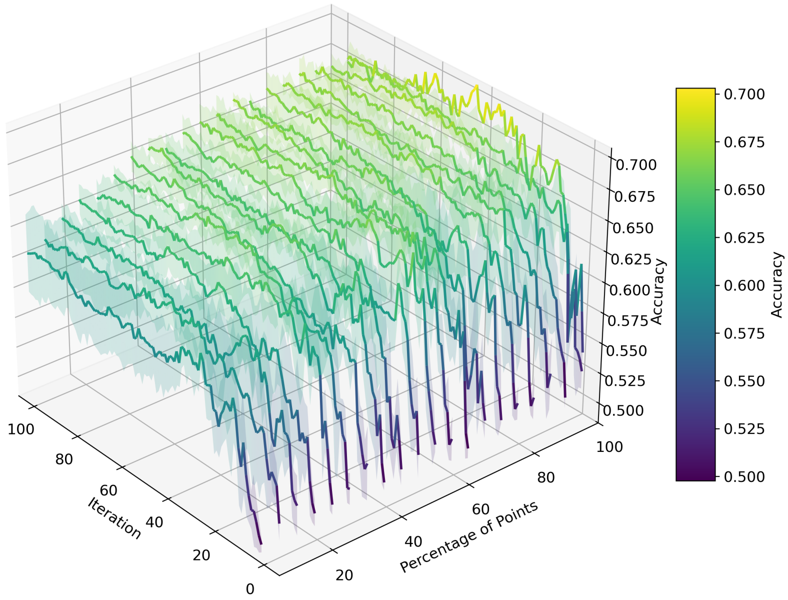


Figure 3.5: *Validation accuracy for random data selection across varying percentages and iterations. Unlike Shapley-based methods, random selection fails to consistently reach the utility target (0.696) within the budgeted iteration range.*

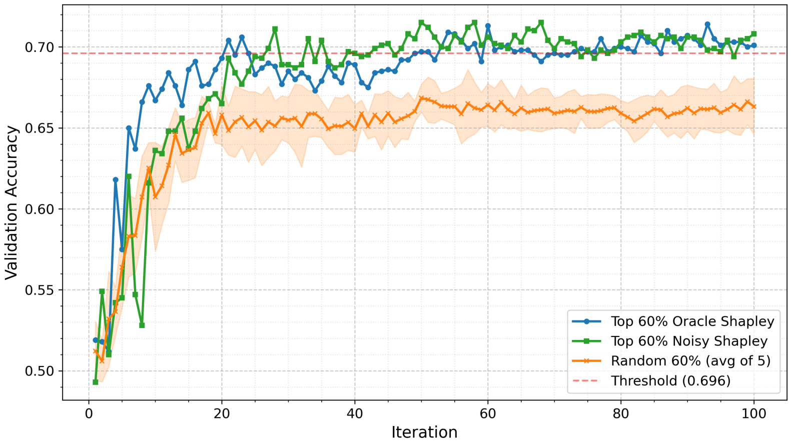


Figure 3.6: *Validation accuracy over iterations for 60% data selection using random, esimated (noisy) data Shapley strategies compared to exact Shapley values given by an oracle. Estimated data Shapley selection reaches the utility target in 25 iterations, significantly outperforming random selection.*

Data Shapley Selection. Afterwards, we assess whether Shapley valuation works over noisy data. More precisely, the experiment consists of the DC selecting all data points for a number of iterations (*i.e.*, bootstrap iterations) and estimating DSVs based on the averaged center points. As illustrated in Figure 3.7, selecting the top-valued centers based on noisy Shapley estimates allows the DC to reach the target utility. Successful acquisition starts at just 10% of the dataset but performance varies depending on both the percentage of selected data and the number of iterations. For example, in the 60% selection case (Figure 3.6), data Shapley selection achieves the utility target after roughly 25 bootstrap iterations. This figure also compares estimated data Shapley to the exact values given by an oracle, indicating that estimating Shapley values has a cost but it diminishes as more budget is consumed. Finally, we tested whether the DC can first estimate Shapley values for a number of bootstrap iterations, then commit to the top-selected points for continued acquisition potentially saving budget and increasing performance. Figure 3.8 explores this strategy across commitment ratios and bootstrap iterations. We find no consistent advantage in committing early, as committing after a fixed iteration does not reduce the number of total iterations needed to reach target utility as shown in Figure 3.9. Although committing may stabilize performance for a fixed estimation, our findings suggest that for Shapley-based strategies, the DC is better off continuing complete iterations to refine noisy point estimates.

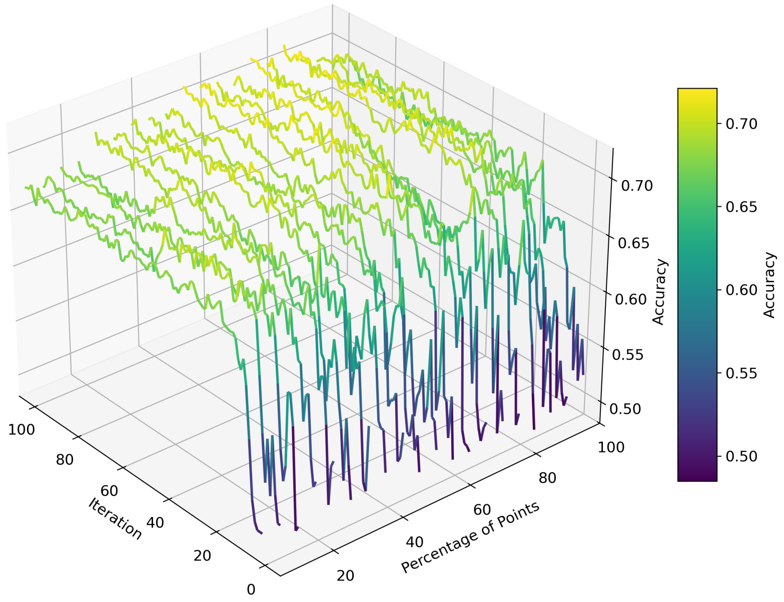


Figure 3.7: *Validation accuracy using estimated data Shapley selection across dataset percentages and iterations. The target accuracy is achieved with as little as 10% of data.*

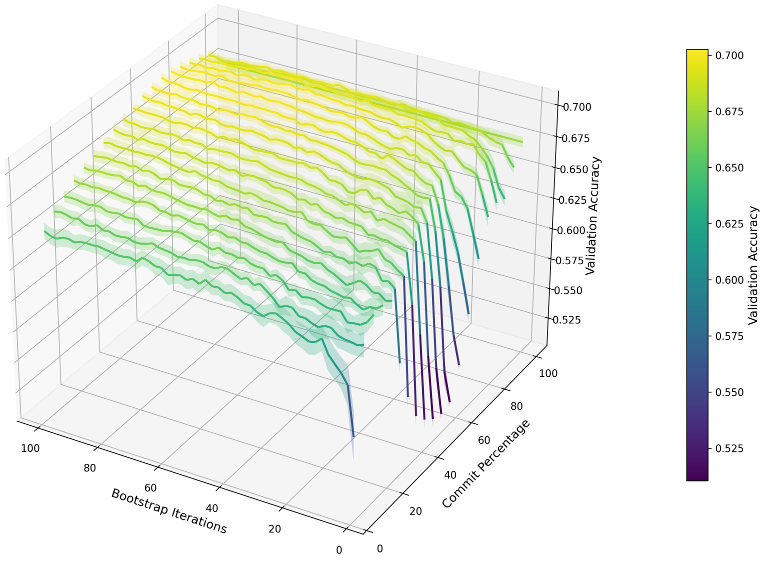


Figure 3.8: *Performance of data Shapley+commitment selection strategy for different combinations of bootstrap and commit parameters. No clear benefit is observed compared to no commitment.*

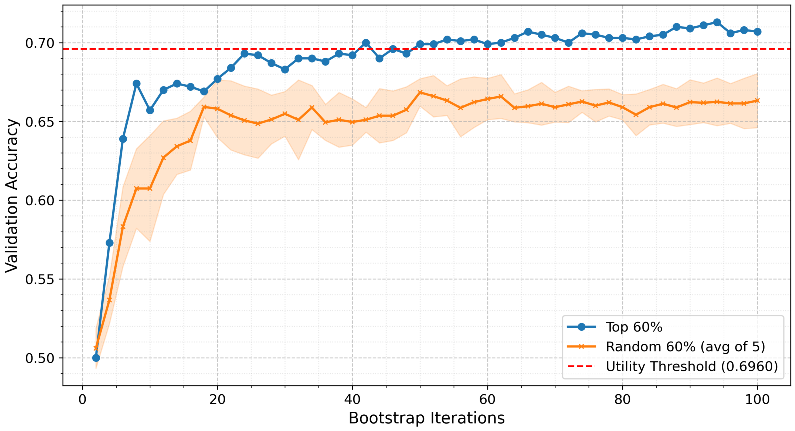


Figure 3.9: *Shapley+commitment strategy on 60% of the data. While utility is eventually reached, performance does not improve over dynamic Shapley.*

3.4.2 Data Selection Using n-Armed Bandits

As a final strategy, we evaluate a multi-armed bandit (MAB) approach for data selection, in which each data point z_i can be seen as an individual arm to pull. While many other selection mechanisms could be explored, we chose to test MABs to assess whether they naturally align with our prior findings on the budget required to reach the utility targets. Unlike the Shapley-based strategies, the MAB formulation does not rely on bootstrap iterations as the DC can choose to exploit a single data point repeatedly up to the constraint of a fixed privacy budget B_{MAX} per point.

For each iteration t (here we talk about point iterations, not bootstrap iterations), the DC selects a data point (*i.e.*, an arm) to query, thereby incurring a privacy cost and obtaining a reward in the form of an incremental improvement in model utility. This formulation captures the inherent exploration-exploitation trade-off as the DC must balance acquiring new data to enhance model performance with the limitations imposed by privacy constraints. As more queries are made, the accumulated privacy budget allows for distinctions between points, gradually increasing the estimated rewards for certain data points.

Upper Confidence Bound (UCB) with a Budget Constraint. We model the DC's behavior as an n-armed bandit problem, in which each action $a \in \{1, \ldots, n\}$ corresponds to selecting a data point z_a. The objective is to identify and refine the most valuable data points for a k-NN classifier, using the fraction of positive votes as a utility metric under a fixed per-point differential privacy budget.

At each iteration t, the DC selects one data point to query, receives a new noisy sample and updates its averaged center. Rewards are defined based on changes on overall utility of the current k-NN model using the available centers. More precisely, each point maintains $Q_t(a)$, the estimated value of point a, $N_t(a)$ the number of times point a was queried, center_a the current average of noisy observations as well as B_a^{re}, the remaining privacy budget. UCB scores are computed as:

$$\text{UCB}_t(a) = \begin{cases} Q_t(a) + c \cdot \sqrt{\frac{1}{N_t(a)+\varepsilon}} \cdot \frac{B_a^{\text{re}}}{B_a^{\text{MAX}}}, & \text{if } B_a^{\text{re}} > 0 \\ -\infty, & \text{otherwise} \end{cases}$$

Finally, the data utility is computed as the average vote agreement among the k neighbors (*i.e.*, the proportion of neighbors with the same label as the query point). See Algorithm 2 for more details.

Hyperparameter Search. Identifying an equilibrium in this setting can be viewed as a hyperparameter search problem as the dynamics between the DO's release policy and the DC's acquisition strategy create a complex stochastic interaction that does not admit an analytic equilibrium. Therefore, we conduct a grid search over two key hyperparameters: the *maximum budget per point* (set by the DO) and the *exploration coefficient c* (set by the DC in the UCB algorithm). These parameters influence both the total budget usage and the diversity of points selected over time.

Figure 3.10 shows a 3D sweep of these hyperparameters, in which each point represents a run color-coded by whether the utility threshold was achieved. We find that the bandit consistently achieves the threshold across a wide range of configurations, except when the maximum budget per point is too low or exploration is entirely disabled. Notably, reaching the target is highly unlikely below a budget-per-arm of 20. Some solutions were found under that budget but closer inspection shows they result from

Algorithm 2 Budget-Aware UCB for Data Selection

```
Require:
    $\mathscr{D} = \{x_1, \ldots, x_n\}$: dataset of $n$ data points (arms)
    $\varepsilon$: Differential privacy budget per query
    $T_{\max}$: Maximum number of iterations
    $U_{\text{target}}$: Target utility (average positive vote ratio)
    $c$: Exploration coefficient
    $\alpha$: Learning rate
    $\varepsilon$: Small constant for numerical stability
    $B^{\text{MAX}}(a)$: Maximum privacy budget per point
    NoisyRelease($x_a$, $\varepsilon$): Returns a noisy version of $x_a$
    knnUtility($\mathscr{X}$; $\{\text{center}_a\}, k$): Average positive votes ratio over validation set $\mathscr{X}$
    with current centers
Initialize:
for $a = 1$ to $n$ do
    $Q(a) \leftarrow 0, N(a) \leftarrow 0, B^{\text{re}}(a) \leftarrow B^{\text{MAX}}(a)$
    $\text{center}_a \leftarrow 0$
end for
$U \leftarrow \text{knnUtility}(\mathscr{X}; \{\text{center}_a\}, k)$
$t \leftarrow 1$
while $t \leq T_{\max}$ and $U < U_{\text{target}}$ do
    for $a = 1$ to $n$ do
        if $B^{\text{re}}(a) > 0$ then
            $\text{UCB}(a) \leftarrow Q(a) + c \cdot \sqrt{\frac{1}{N(a)+\varepsilon} \cdot \frac{B^{\text{re}}(a)}{B^{\text{MAX}}(a)}}$
        else
            $\text{UCB}(a) \leftarrow -\infty$
        end if
    end for
    $A_t \leftarrow \arg\max_a \text{UCB}(a)$
    $\tilde{x}_{A_t} \leftarrow \text{NoisyRelease}(x_{A_t}, \varepsilon)$
    $\text{center}_{A_t} \leftarrow \frac{\text{center}_{A_t} \cdot N(A_t) + \tilde{x}_{A_t}}{N(A_t)+1}$
    $U_{\text{new}} \leftarrow \text{knnUtility}(\mathscr{X}; \{\text{center}_a\}, k)$
    $R_t \leftarrow U_{\text{new}} - U$
    $U \leftarrow U_{\text{new}}$
    $Q(A_t) \leftarrow Q(A_t) + \alpha \cdot (R_t - Q(A_t))$
    $N(A_t) \leftarrow N(A_t) + 1$
    $B^{\text{re}}(A_t) \leftarrow B^{\text{re}}(A_t) - \varepsilon$
    $t \leftarrow t + 1$
end while
return $\{\text{center}_a\}, U$
```

what we term *lucky point selection* in which the bandit happened to identify valuable arms early and focused its budget there. This finding is supported by Figure 3.11, which visualizes the Gini coefficient of budget consumption. High Gini values indicate heavy concentration of queries on a few points, which is characteristic of narrow exploitation.

Finally, we examine the relationship between learned Q-values and Shapley values. Figure 3.12 shows that the Spearman correlation between these metrics increases with budget but remains moderate overall. This is expected and desired as the objective is for Q-values to be influenced by data utility but not perfectly mimic Shapley values, thereby offering a different valuation with more inclusiveness.

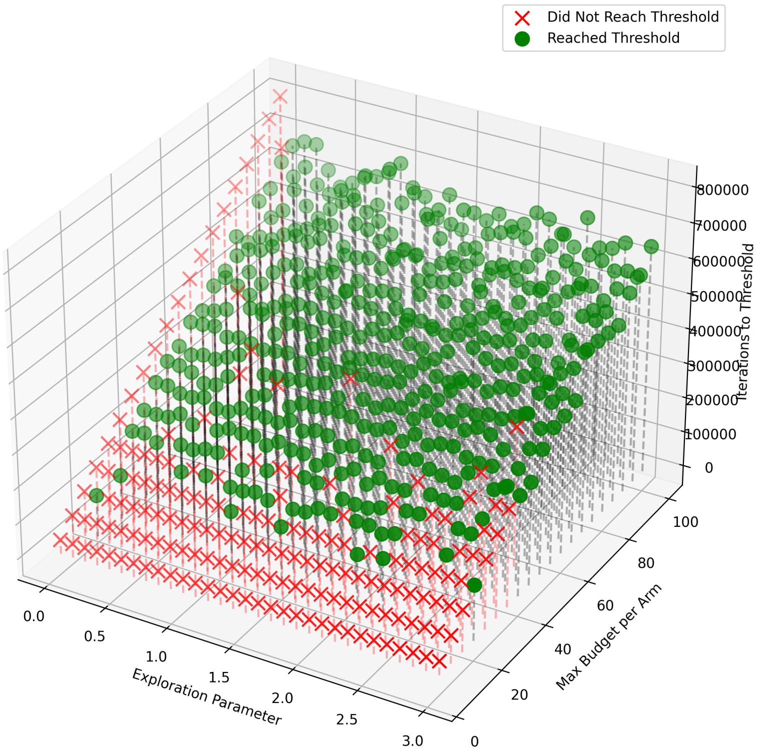


Figure 3.10: *3D visualization of hyperparameter combinations. Green dots represent successful runs in which the DC reached the utility threshold while red crosses represent failures. Success becomes unlikely with budget-per-arm below 20 or when exploration is zero.*

Across all strategies, we find that a similar minimum budget is needed for the DC to reach the utility target. The Gini analysis further confirms that exploration increases the likelihood of success by promoting a more balanced budget allocation

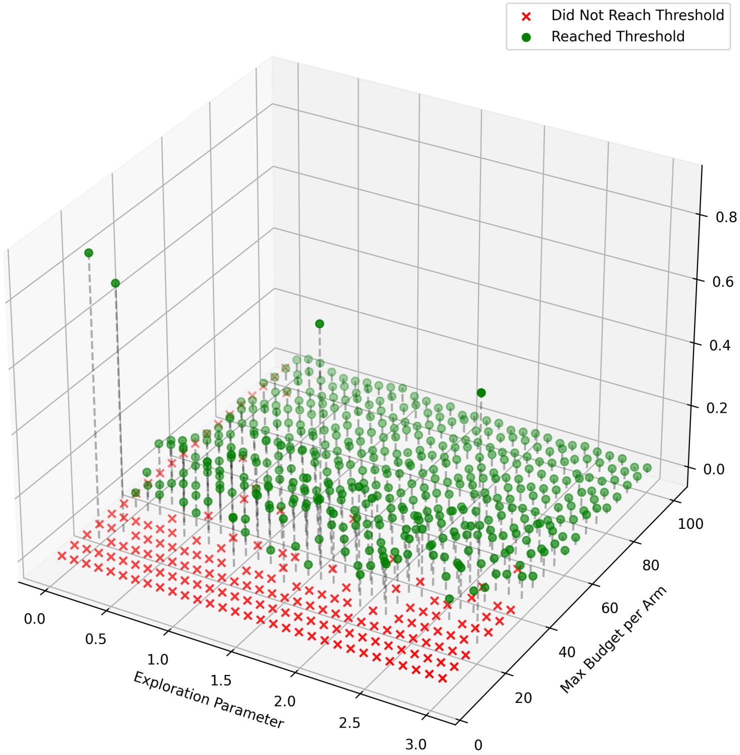


Figure 3.11: *Gini coefficient of budget usage by hyperparameter setting. High values indicate budget concentrated on few data points, typical in "lucky" early selections.*

across points. It is also worth noting that we observed that models trained on the final denoised centers—despite being derived from noisy data—achieved test performance comparable to or exceeding that of models trained on the original dataset. For instance, one configuration with a maximum per-point budget of 50 reached a test accuracy of 0.699, surpassing the original benchmark of 0.660.

3.4.3 Privacy Leakage Analysis

As a final experiment, we conducted a privacy leakage analysis to assess the risks of exposing a k-NN model trained on data released through our iterative ε-differentially private mechanism. This is relevant for instance for scenarios in which the organization makes the model accessible externally such as via an API. Among k-NN configurations, $k = 1$ is particularly vulnerable to membership inference attacks due to its reliance on individual training points (Mesana et al., 2024), making it a conservative setting for

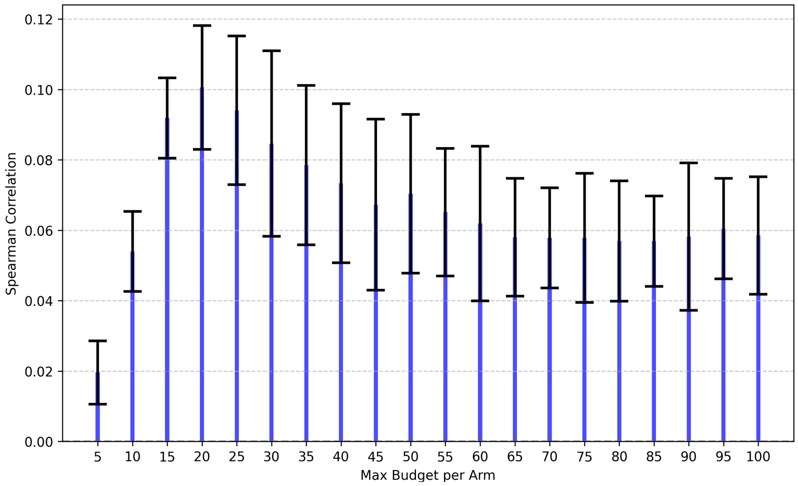


Figure 3.12: *Mean and standard deviation of Spearman correlation between learned Q-values and Shapley values, grouped by maximum budget. Correlation rises with budget but remains modest indicating partially aligned but distinct prioritization.*

privacy evaluation.

We used LiRA (Carlini, Chien, et al., 2022a), a standard membership inference attack to compare the leakage between a model trained on the original dataset and one trained on noisy centers after 25 iterations of Laplacian releases. As shown in Figure 3.13, the privacy leakage is notably reduced in the noisy case. More precisely, the AUC drops from 0.81 (original) to 0.74 (with noise), and the true positive rate at 5% false positive rate decreases from 0.282 to 0.089. These results suggest that even without end-to-end formal DP guarantees, progressively injecting noise offers meaningful privacy benefits by limiting the success of membership inference attacks.

3.5 Discussion

While our experiments focused on the *k*-NN classifier, the core ideas of our framework extend naturally to other types of models, which opens several avenues for further work.

In the case of differentiable models (*e.g.*, neural networks), our UCB-based data selection shares a fundamental assumption with *G-Shapley* (Ghorbani and Zou, 2019), which is that a data point's contribution can be estimated by measuring contribution in model performance after incorporating that point into training. G-Shapley approximates

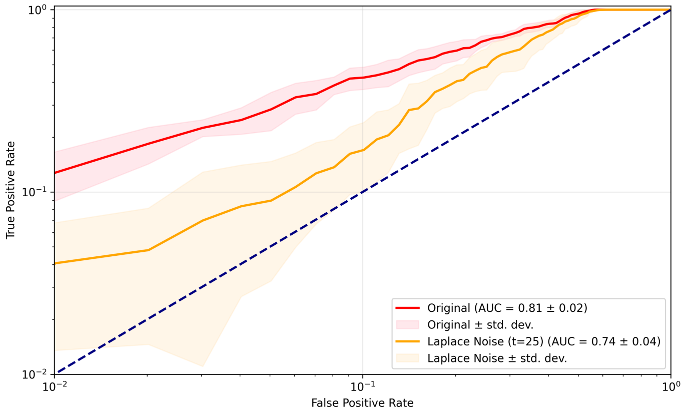


Figure 3.13: *Membership inference risk evaluated using LiRA. The model trained on noisy centers after 25 iterations (orange) shows reduced leakage: AUC drops from 0.81 to 0.74, and TPR@FPR = 0.05 drops from 0.282 to 0.089. Shaded areas show standard deviation.*

the Shapley value of a data point z_i by considering multiple random permutations π of the dataset. Given a performance function $V(\theta)$, which evaluates model utility (*e.g.*, accuracy or loss on a validation set), G-Shapley sequentially updates model parameters via stochastic gradient descent (SGD) following the order in π:

$$\theta_j = \theta_{j-1} - \alpha \nabla_\theta L(z_{\pi[j]}; \theta_{j-1}),$$

in which $L(z;\theta)$ is the loss function for an individual data point z. The Shapley value estimate for $z_{\pi[j]}$ is then updated based on the marginal change in performance:

$$\phi_{z_{\pi[j]}} \leftarrow \frac{t-1}{t}\phi_{z_{\pi[j]}} + \frac{1}{t}\big(V(\theta_j) - V(\theta_{j-1})\big).$$

While G-Shapley retrospectively estimates contributions based on random partitions, our approach *actively* selects data points based on an exploration-exploitation trade-off, using an Upper Confidence Bound (UCB) strategy.

Another related approach is *DP-SGD* (Abadi et al., 2016), which ensures DP during model training by modifying standard stochastic gradient descent (SGD). DP-SGD applies noise at every step of training and privacy loss accumulates over multiple gradient updates. A *privacy accountant* method can be used to track cumulative privacy loss. In contrast, our framework follows the standard composition rule. Unlike DP-SGD, we do not perform batch updates—each data point is queried individually and each query incurs a fixed privacy cost ε. Thus, privacy costs accumulate additively without interactions between data points. However, this approach comes with a trade-off: without larger batch updates and a privacy accountant, convergence could be slower as fewer data points are used per iteration. Additionally, the lack of tighter composition bounds means that the total privacy budget may be consumed more rapidly compared to DP-SGD, potentially leading to a less favorable privacy-utility trade-off.

3.6 Conclusion

We proposed a game-theoretic framework for data valuation that integrates privacy and inclusiveness concerns through an information disclosure game between a data owner (DO) and a data consumer (DC). Rather than assuming direct access to data, our approach models acquisition as a iterative process using a differential-private mechanism to enforce a privacy cost on data valuation. More precisely, in our case study using the Yelp dataset and the reviews helpfulness classification task, we showed that Shapley-based strategies remain effective under noise but require a minimal budget consumption—typically at least 10 full bootstrap iterations—to identify high-value points. This means that the data Shapley value comes with an explicit acquisition cost in the information disclosure game. Additionally, a multi-armed bandit approach achieved similar utility with comparable budget consumption, also suggesting an inherent acquisition cost imposed by noisy data. Combined with the Gini analysis of budget consumption, we observe that the DC's strategy shifts toward more inclusiveness as a reaction, spreading its acquisition effort across more data points. Despite this cost, DCs consistently reached target utility, and in some cases, test performance exceeded that of models trained on original data. Furthermore, privacy leakage was significantly reduced. While our experiments focused on *k*NN classifiers, the framework is general and can be extended to other models. For instance, adapting our information disclosure game to differentiable models would enable more expressive utility functions and allow for privacy-aware data valuation using DP-SGD-like mechanisms.

General Discussion

This thesis has explored privacy from multiple perspectives—legal, contextual, risk-based, and utilitarian—but consistently focused on the latter two through a technical lens. A key idea explored is that, although privacy auditing and data valuation rely on fundamentally different tools, they are often applied in ways that reveal deep connections—particularly when evaluating individual-level tradeoffs. Privacy is typically audited through techniques like membership inference attacks (MIAs) while data valuation uses methods such as Shapley value estimation. Despite their methodological differences, these approaches can be jointly analyzed to study the privacy-utility tradeoff. For instance, at the aggregate level of course, it is possible to aim for Pareto optimality, but our main interest lies in understanding how risk and value interact at the level of individual data points. In particular, our motivation was to examine how this interaction shapes incentives. While framing tradeoffs in terms of incentives is not new—differential privacy (DP) for example, treats privacy loss as a disincentive to participate—the integration of these views outside the DP framework, particularly in *a posteriori* analyses, remains underexplored.

In the first chapter, the individual's disincentive does not come from any privacy attack or from an assumed membership inference attack, which DP is designed to protect against. Instead, we consider adversaries attempting to single out a specific individual in a database by leveraging background knowledge they already possess. We argue that this threat model plays a significant role in diminishing individual trust in organizations, thereby creating a disincentive to share personal data. This assumption aligns with the view of privacy as contextual integrity (Nissenbaum, 2004), which holds that privacy violations occur when established norms of information flow are disrupted, even if formal access rights remain unchanged. A well-known example is the introduction of Facebook's News Feed, in which users experienced a loss of privacy—not because new information was revealed, but because scattered data became easily discoverable. The discomfort arose from a reduction in the cost of search, which violated contextual expectations despite unchanged consent and access control policies (Grimmelmann, 2008). In our case, the key distinction is that we focus on reduced search costs within the confines of an organization's internal systems. This also resonates with the legal perspective, as GDPR Recital 26 (European Union, 2016) states that data enabling singling out is not considered anonymous.

Most organizations focus on reducing large-scale breaches through improved governance, stronger cybersecurity, stricter data-sharing policies or a combination of these strategies. However, they often fail to address the types of risks emphasized in the computer science literature. In particular, DP appears to see limited practical adoption in real-world settings (Amin et al., 2024; Garrido et al., 2022). Indeed, many organizations consider privacy risks to be minimal as long as they have obtained user consent and secured their data assets. For example, in a recent survey conducted by the International Association of Privacy Professionals (IAPP)—a leading industry organization focused on privacy and data protection—nine out of ten respondents reported being at least somewhat confident in their organizations' privacy governance programs (International Association of Privacy Professionals, 2024). Moreover, employees themselves often struggle to see how Privacy Enhancing Technologies (PETs) contribute to meeting compliance goals (Klymenko et al., 2023). One of our objectives in the first chapter was therefore to raise organizational awareness of micro-level privacy breaches—particularly by analyzing data flows that are already familiar to them, such as how internal analysts query databases.

While risk management priorities may be subject to debate, it is clear that individual disincentives to share data can stem from specific, measurable risks—and these need to be assessed. To make our approach more interpretable to organizations, we drew inspiration from the widely used triplet representation of risk: scenario, probability and consequence (Kaplan and Garrick, 1981). In this view, a *scenario* represents a particular chain of events or system behavior; *probability* quantifies the likelihood of that scenario occurring and *consequence* refers to the impact if it does. In our framework, we argue that risk should not be reduced to the probability of identifying an individual record alone. While that is a critical factor, it must also account for information gain, as a rational individual may not be concerned if the only inference is something trivial—such as mere membership—or information that is already publicly available. In fact, in some situations membership inference does not lead directly to a privacy breach. For instance, in Québec, the two largest financial institutions—Desjardins and the National Bank—serve an estimated 70% to 80% of the population (based on official 2024 reports). As a result, inferring that someone is a client of at least one of them is often trivial from an attacker's perspective.

The framework we propose in the first chapter, allows organizations to adjust the scenarios they consider relevant. For example, most organizations already log database queries—whether in a data warehouse or data lake—and these logs are often monitored for signs of malicious activity. This kind of analysis parallels recent work showing that software logs can reveal sensitive composite identifiers, underscoring the value of systematically examining operational traces for privacy risk (Aghili et al., 2024). A use case we have not yet tested, but which should be straightforward to implement, involves re-executing logged queries, analyzing and scoring the resulting subset of records as if it were the output of a re-identification attempt. Organizations could then compare real queries to the uniformly distributed attack simulations we conducted in Chapter 1—in which no prior assumptions are made about which queries are more likely—in order to estimate the likelihood that a given query behaves like a re-identification attempts. This could offer a valuable tool for cybersecurity teams to detect and investigate queries that disproportionately narrow down to specific individuals. Another of our contributions, Re-identification Shapley Values (RSV), is particularly actionable and can support organizations in managing attribute-level security. For instance, consider the scenario in which analysts from an organization can access attributes such as the primary bank branch ZIP code, job title and type of recurring monthly expense (*e.g.*, private school tuition). While none of these attributes may seem highly sensitive in isolation, they function as quasi-identifiers—attributes, which when combined, can uniquely identify individuals. Their re-identification power arises not from their individual informativeness, but from their synergy. RSV is designed to capture these joint effects more accurately than traditional sensitivity assessments and can help quantify the incentive for an attacker to acquire specific pieces of background knowledge.

In the second chapter, WaKA further explores how mathematical tools from both the risk-based and utilitarian lenses of privacy can be unified. Specifically, it uses the same method—Wasserstein distance computed on the k-NN loss distribution—to assess both membership privacy risk and data valuation. This dual use allowed us to address a central question in the thesis: how does a data point's contribution to model performance, or what we call data value, relate to its vulnerability to privacy attacks? A common assumption is that data value and privacy risk are positively correlated. Using LiRA, a state-of-the-art MIA, we observed that on k-NN models across multiple datasets, the correlation between data Shapley values and privacy risk scores is nonlinear. In some cases, data points with negative Shapley values—those that degrade model performance—are actually more susceptible to privacy attacks than high-value points. We further confirmed this insight in the Chapter 1 addendum using our contextually defined re-identification risk scores. This has important implications for understanding individual-level tradeoffs and consent dynamics. In particular, an organization might naively assume that low-value data implies low individual concern—"if the data is not useful, why would anyone care if it is used?". But our findings suggest the opposite: low-value data can carry higher-than-average privacy risk and thus deserves careful consideration. Conversely, decisions by organizations and individuals could be better aligned when a data point is both low in value and high in risk—making its removal beneficial to both parties. However, we also replicate the onion effect described by Carlini, Jagielski, et al., who show that removing such points can inadvertently increase the risk for remaining individuals. This highlights the complexity of incentive dynamics when processing personal information.

It is important to note that data removal is not, *a priori*, a differentially private operation. While it may reduce empirically observed re-identification risk and thus be perceived as a privacy enhancer from some perspectives, it does not constitute a formal privacy guarantee in absolute terms or under any definition unless combined with a mechanism that enforces such guarantees. Moreover, when removal is guided by data Shapley values, the selection process itself can be exploited: observing which records are removed or retained can reveal partial orderings of Shapley values and, by implication, properties of the underlying data points. This leakage risk is closely related to the "privacy cost of valuation" we discuss in Chapter 3.

In the second chapter, we also show that attribution with respect to the individual—what we refer to as self-attribution—is more strongly correlated with the success of membership inference attacks. This implies that data valuation methods, which typically estimate a point's value with respect to the test set, may not serve as reliable indicators of privacy risk. In contrast, the self-Shapley value—a notion we did not find explicitly formulated in the existing literature—shows strong correlation with both self-WaKA scores in k-NN classifiers and membership attack success rates. From a utilitarian perspective, the self-Shapley value can be interpreted as a measure of how much including your data improves predictions about yourself. For example, if a financial institution is building a credit scoring model, a high self-Shapley value would suggest that the model performs significantly worse on your profile without your data. If individuals were aware of this, it could influence their willingness to share data, depending on the potential benefits. Conversely, a low self-Shapley value suggests that many substitute data points exist, meaning the model performs similarly whether or not your data is included. This aligns with the ideas behind privacy-preserving techniques, where the goal is indistinguishability: for instance, k-anonymity enforces indistinguishability at the data level, while DP enforces indistinguishability at the process (mechanism) level. This connection illustrates how Shapley value and DP can reflect similar reasoning about individual participation and data substitutability, despite differing methodologies. However, DP goes further by incorporating truthfulness as a core principle, which means that participants have no incentive to misrepresent their preferences or valuations—truth-telling becomes their dominant strategy. As McSherry and Talwar explain, privacy-preserving mechanisms can be modeled as social choice games, in which privacy loss is treated as an explicit cost analogous to the cost of lying. Under DP, the expected outcome of the mechanism remains nearly unchanged whether or not any single individual's data is included, giving agents strong incentives to participate truthfully without fearing privacy loss.

In the third chapter, we extend this connection further by using a DP mechanism to control and partially release information, which we then evaluate using k-NN Shapley valuation. This experimental setup allows us to demonstrate that obtaining stable and meaningful estimates of data value requires a large number of queries. We coin this phenomenon the privacy cost of data valuation. A limitation of our approach is that it operates in a global setting, meaning the information disclosure game we model is largely conceptual. A promising extension would be to explore a decentralized setting using local differential privacy (LDP, Cormode et al., 2018; N. Wang et al., 2019). However, it would introduce additional challenges, notably the possibility of collusion among individuals. Collusion could bias data valuation outcomes and complicate the design of fair reward mechanisms or data selection strategies. In other words, this would require simulating a much more complex game, where both privacy guarantees and incentive compatibility must be carefully balanced.

Finally, in the last part of this discussion, we want to position the work and contributions of this thesis within current trends in machine learning and artificial intelligence. In particular, three rapidly evolving areas intersect with our work: data attribution at scale, privacy auditing for large language models (LLMs) and LLM alignment. Each of these fields engages with questions central to this thesis—how data contributes to outcomes, how privacy risks can be measured and mitigated, and how systems can be designed to reflect the values and expectations of individuals.

Data attribution has recently gained significant attention and is now understood as a broad framework encompassing a variety of approaches—not just game-theoretic ones—that aim to trace the influence of training data on a machine learning model's behavior and outputs (Ilyas et al., 2022; S. M. Park et al., 2023; Worledge et al., 2024). In particular, its applications now extend beyond data valuation. For instance, data attribution in the context of LLMs can be used for citation generation (Gao et al., 2023). More precisely, this is an instance of what Worledge et al. refers to as corroborative attribution, in which the goal is to identify specific training data points that directly support a generated output—enabling the model to "cite" its sources. Datamodels (Ilyas et al., 2022) are also a representative example of this kind of corroborative attribution, as they aim to approximate the training process in order to predict a model's output had it been trained on a particular subset of data. In contrast, contributive attribution focuses on identifying more causal relationships between individual data points and model behavior. Data valuation also falls under this latter category of contributive attribution. Rather than simply identifying which data points are linked to specific outputs, it seeks to characterize the functional role of each point—particularly the effect of its inclusion or removal on overall model behavior. While this can support applications like pricing mechanisms, the foundational goal is to understand how individual data points shape model performance in a stable and interpretable way. Crucially, this also accounts for the supply or substitutability of a point within the dataset. For instance, a point that can be easily replaced by others may be assigned lower value. Although some empirical applications of data valuation have proven robust, prior work has shown that the results of several influential economics papers can be overturned by removing less than 1% of the data (Broderick et al., 2020).

The primary obstacle to applying data valuation to LLMs—particularly at scale—is computational in nature. Several papers have emerged that address this issue in different ways, each with its own strengths and limitations (Just et al., 2023; S. M. Park et al., 2023; J. T. Wang, Mittal, et al., 2024). One promising alternative is to use a proxy or surrogate model that approximates the behavior of the generative neural network and then compute data attributions on that proxy. For instance, as we did with WaKA, one can apply a k-NN directly to the final-layer embeddings (Jia et al., 2021). This idea is supported by growing evidence that pretrained LLMs behave similarly to non-parametric models due to their few-shot learning capabilities—a phenomenon often referred to as in-context learning (Kim et al., 2024; Y. Zhang et al., 2023). Furthermore, Nguyen, 2024 shows that n-gram statistics can explain a large portion of LLM behavior: many next-token predictions follow patterns that could be captured by simple n-gram rules derived from training data. In other words, always selecting the top-matching rule from a fixed datastore can be surprisingly effective. A related idea is presented in k-NN-LLMs (Khandelwal et al., 2019), which separates the generative task into two steps: embedding the sentence prefix and using those representations to query a k-NN datastore for next-token prediction. They show that this approach can even outperform end-to-end LLMs by explicitly retrieving relevant data points, especially for rare patterns that may be memorized in model parameters. Taken together, this growing body of evidence suggests that k-NN Shapley—or more specifically, WaKA—could be applied to generative LLMs as an efficient method for data valuation and privacy auditing. In addition, WaKA could also be extended to compute privacy influence scores (PrivInf), helping identify data points that would be most vulnerable after data removal. We believe this opens a promising direction for designing interpretable AI systems—ones in which understanding data influence and privacy risk is prioritized, even if k-NN imposes some performance cost.

LLMs raise significant privacy concerns due to their tendency to memorize and reproduce training data, which may include sensitive personal information (Staab et al., 2023). MIAs remain a useful tool—for instance, they are often used as a first step toward extracting specific training data, such as inserted canaries, from LLMs (Carlini et al., 2021). However, recent work has questioned the reliability of MIAs in proving that a model was trained on specific data (J. Zhang et al., 2024). That study argues that MIAs alone cannot provide statistically robust evidence of training data inclusion—a claim often central to lawsuits involving large-scale AI models. The main challenge is that estimating the false positive rate, which is crucial for statistical confidence, is practically impossible without access to the full training set or the ability to retrain the model. The researchers illustrate this with the example of Harry Potter: even if the books were not part of the training data, a model might still respond accurately to related prompts, simply because the concept is so prevalent in public web content. This undermines claims that a model's outputs necessarily reflect direct exposure to specific data. Instead, the authors recommend alternatives such as inserting trackable canaries, watermarking data or demonstrating data extraction—methods that provide stronger, statistically grounded evidence of training data usage. Hayes et al. further distinguishes between weak and strong MIAs. While strong MIAs—such as LiRA—can still be effective, they require substantial computational resources, particularly when applied to large models. More research is needed to clarify how MIAs relate to other privacy metrics, such as data extraction and to assess whether they can serve as reliable indicators of risk in ways that are actionable for decision makers. This resonates with points raised in Chapter 1 of this thesis: Privacy attacks must be contextualized to be meaningful to organizations.

So far, our discussion on LLMs has focused on their pre-training phase, which mirrors traditional machine learning in that it trains models to perform next-token prediction. However, this is only the initial step. Modern LLMs also undergo post-training processes such as supervised fine-tuning and alignment, which are intended to shape their behavior according to specific stakeholder preferences. These models are increasingly deployed with capabilities and autonomy that resemble agent-like systems. This raises a natural question: does this agentic paradigm shift change how we should think about privacy risk? Alignment refers to training a model so that its behavior reflects the preferences, values or goals of specific stakeholders (Ngo et al., 2022). Current alignment techniques carry inherent risks—especially when applied to agentic models. For instance, recent work shows that deceptive behaviors can persist through safety training (Hubinger et al., 2024), and that large reasoning models can unintentionally leak private user data through intermediate reasoning steps (Green et al., 2025). As noted in Chapter 1, our threat models originally assumed human adversaries, but the same concerns clearly extend to LLM-based agents, which could automate privacy attacks and scale fraudulent behavior

(Abdullaev et al., 2025). Recent ML literature has also revived privacy definitions beyond DP, including contextual integrity and the notion of "revealing secrets" (Mireshghallah et al., 2023). These perspectives raise a broader question: can we align models to avoid unintended information disclosure?

This is closely tied to the privacy-utility tradeoff and the incentives individuals face. How much people value their data is intimately linked to their perception of privacy. For instance, only 43% of women with wearables report sharing fitness data with doctors, compared to 57% of men; overall, 58% of users fear data breaches (Hupfer et al., 2023). Perceptions are also dynamic: Hsu et al. notes that even after events like the Cambridge Analytica scandal, many users returned to platforms due to lack of alternatives. This illustrates that the privacy tradeoff is not only technical—it is social, psychological and contextual. Still, formal tools can promote clearer thinking by helping actors reason from first principles. As shown in Chapter 2, the choice of valuation function (*e.g.*, self-Shapley vs. test-based Shapley) leads to different interpretations. Even with shared axioms, stakeholders may diverge on objectives. Optimizing for both utility and privacy is often infeasible; Pareto optimality may be the best alternative, in which no stakeholder's gain comes without someone else's loss. Keeping this in mind, alignment—specifically privacy alignment—cannot be captured by a single objective, but rather emerges from external mechanisms, such as DP noise or constraints tied to user risk profiles.

In Chapter 3, we proposed a setting with two agents: protecting each individual's interests, the other pursuing a utility goal. This framing shows that data valuation with privacy is not necessarily a zero-sum game. Utility goals can be achieved while minimizing queries and mechanisms like DP can help balance this tradeoff. Recent work frames alignment as a two-player game between models and human oversight (Cheng et al., 2023) or uses mechanism design to incentivize truthful reporting of preferences during fine-tuning (Sun et al., 2024). We believe game-theoretic approaches are promising for privacy-aware alignment. In such a setting, one agent might aim to enhance the value of private data, another may simulate an attacker, and a third could act as an analyst creating value. Alignment, then, becomes an equilibrium across competing incentives—not reducible to a single goal.

General Conclusion

This thesis positions data valuation as a lens through which to study the privacy-utility tradeoff—an angle that has received little attention in prior work. While much of the privacy and security literature focuses on guarantees and the search for provable or formal defense mechanisms—a pursuit we view as both important and necessary—we argue that disincentive-based reasoning deserves more emphasis. This perspective opens the door to alternative formulations of privacy that, while distinct from DP, remain consistent with its core principle: "we want to minimize any reason individuals have to opt out". Crucially, it also invites another key question that any individual might ask: What is my data worth?

This thesis makes three contributions at the intersection of data valuation and privacy in machine learning. First, it develops a simulation-based framework for evaluating format-preserving data release strategies and assigns both attribute-level and individual-level re-identification risk scores. Second, it introduces *WaKA*, a method for k-nearest neighbor models that supports both data valuation and membership privacy auditing by measuring the impact of specific records on the loss function of the model. Third, it models data sharing as a strategic interaction in which a data owner controls information leakage through a differential privacy mechanism, showing how noisy data release affects data Shapley values and the incentives of a data consumer.

Beyond academic contributions, we believe the methods and insights developed here can support practical decision-making across sectors. Information security teams, data governance professionals, and regulatory bodies can apply these tools to better assess individualized privacy risk and balance it against utility. One especially promising application lies in cyber-insurance, in which contextual and fine-grained risk assessments could help inform premium pricing and claims evaluation. Ultimately, integrating data valuation into privacy risk frameworks provides not only deeper technical understanding, but also a foundation for aligning economic incentives with more ethical uses of data.

Future research related to this thesis could explore the privacy-utility tradeoff in large language models (LLMs) using a similar data attribution lens. Instance-based interpretability shares methodological similarities with privacy auditing, as both involve tracing model outputs back to influential training data. A concrete example where the goals diverge is citation generation: whereas privacy auditing aims to prevent the model from revealing sensitive training data, citation generation encourages the model to surface specific sources that support its outputs. Nonetheless, both directions ultimately aim to make models more transparent and accountable — even if pursuing both at once may yet again surface fundamental tensions between privacy and utility.

Bibliography

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

Abdullaev, A., Loshchev, A., & Baldakov, M. (2025, February). The dark side of automation and rise of ai agents: emerging risks of card testing attacks [Blog post].

Acquisti, A. (2010). *The economics of personal data and the economics of privacy* (tech. rep.). Citeseer.

Acquisti, A. (2023). The economics of privacy at a crossroads. *Economics of Privacy. University of Chicago Press*.

Acquisti, A., & Grossklags, J. (2005). Privacy and rationality in individual decision making. *IEEE security & privacy*, *3*(1), 26–33.

Acquisti, A., Taylor, C., & Wagman, L. (2016). The economics of privacy. *Journal of economic Literature*, *54*(2), 442–492.

Aghili, R., Li, H., & Khomh, F. (2024). An empirical study of sensitive information in logs. *arXiv preprint arXiv:2409.11313*.

Amin, K., Kulesza, A., & Vassilvitskii, S. (2024). Practical considerations for differential privacy. *arXiv preprint arXiv:2408.07614*.

Asghar, N. (2016). Yelp dataset challenge: review rating prediction. *arXiv preprint arXiv:1605.05362*.

Basu, S., Pope, P., & Feizi, S. (2021, February). Influence Functions in Deep Learning Are Fragile [arXiv:2006.14651 [cs, stat]].

Bendechache, M., Attard, J., Ebiele, M., & Brennan, R. (2023). A systematic survey of data value: models, metrics, applications and research challenges. *IEEE Access*, *11*, 104966–104983.

Bilal, M., & Almazroi, A. A. (2023). Effectiveness of fine-tuned bert model in classification of helpful and unhelpful online customer reviews. *Electronic Commerce Research*, *23*(4), 2737–2757.

Broderick, T., Giordano, R., & Meager, R. (2020). An automatic finite-sample robustness metric: when can dropping a little data make a big difference? *arXiv preprint arXiv:2011.14999*.

California State Legislature. (2020). California privacy rights act of 2020 (proposition 24) [Major provisions became operative on January 1, 2023].

Carlini, N., Chien, S., Nasr, M., Song, S., Terzis, A., & Tramer, F. (2022a). Membership inference attacks from first principles. *Proceedings of the 2022 IEEE Symposium on Security and Privacy (SP)*, 1897–1914.

Carlini, N., Chien, S., Nasr, M., Song, S., Terzis, A., & Tramer, F. (2022b, April). Membership Inference Attacks From First Principles [arXiv:2112.03570 [cs]].

Carlini, N., Jagielski, M., Zhang, C., Papernot, N., Terzis, A., & Tramer, F. (2022). The privacy onion effect: memorization is relative. *Advances in Neural Information Processing Systems*, *35*, 13263–13276.

Carlini, N., Liu, C., Erlingsson, Ú., Kos, J., & Song, D. (2019). The secret sharer: evaluating and testing unintended memorization in neural networks. *Proceedings of the 28th USENIX Security Symposium (USENIX Security 19)*, 267–284.

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., & Erlingsson, U. (2021). Extracting training data from large language models. *Proceedings of the 30th USENIX Security Symposium (USENIX Security 21)*, 2633–2650.

Cavoukian, A. (2009). Privacy by Design - the 7 Foundational Principles. *Office of the Information and Privacy Commissioner*.

Cheng, P., Yang, Y., Li, J., Dai, Y., & Du, N. (2023). Adversarial preference optimization. *arXiv preprint arXiv:2311.08045*.

Clifton, C., & Tassa, T. (2013). On syntactic anonymity and differential privacy. *2013 IEEE 29th International Conference on Data Engineering Workshops (ICDEW)*, 88–93.

CNET, D. K. (2012, February). *Netflix pays $9 million to settle privacy violation lawsuit*. https://www.cnet.com/tech/services-and-software/netflix-pays-9-million-to-settle-privacy-violation-lawsuit/

Cohen, A., & Nissim, K. (2020a). Towards formalizing the gdpr's notion of singling out. *Proceedings of the National Academy of Sciences*, *117*(15), 8344–8352.

Cohen, A., & Nissim, K. (2020b). Towards formalizing the GDPR's notion of singling out. *Proceedings of the National Academy of Sciences*, *117*(15), 8344–8352.

Cohen, G., & Giryes, R. (2024). Membership inference attack using self influence functions. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 4892–4901.

Cormode, G., Jha, S., Kulkarni, T., Li, N., Srivastava, D., & Wang, T. (2018). Privacy at Scale: Local Differential Privacy in Practice. *Proceedings of the 2018 International Conference on Management of Data*, 1655–1658.

Danakas, D. (2024, January). *Consultation illicite de dossiers médicaux : quand la curiosité a des conséquences fâcheuses* [Billet de blogue]. Blogue SOQUIJ. Retrieved July 8, 2025, from https://blogue.soquij.qc.ca/2024/01/25/consultation-illicite-de-dossiers-medicaux-quand-la-curiosite-a-des-consequences-facheuses/

Dankar, F. K., El Emam, K., Neisa, A., & Roffey, T. (2012). Estimating the re-identification risk of clinical data sets. *BMC Medical Informatics and Decision Making*, *12*(1), 66.

De Montjoye, Y.-A., Radaelli, L., Singh, V. K., & Pentland, A. " (2015). Unique in the shopping mall: on the reidentifiability of credit card metadata. *Science*, *347*(6221), 536–539.

de Montjoye, Y.-A., Radaelli, L., Singh, V. K., & Pentland, A. " (2015). Unique in the shopping mall: On the reidentifiability of credit card metadata. *Science*, *347*(6221), 536–539.

Dinev, T., & Hart, P. (2006). An extended privacy calculus model for e-commerce transactions. *Information systems research*, *17*(1), 61–80.

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

Dosovitskiy, A. (2020). An image is worth 16x16 words: transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Duddu, V., Szyller, S., & Asokan, N. (2021). Shapr: an efficient and versatile membership privacy risk metric for machine learning. *arXiv preprint arXiv:2112.02230*.

Dwork, C. (2006). Differential privacy. *International colloquium on automata, languages, and programming*, 1–12.

Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2016). Calibrating noise to sensitivity in private data analysis. *Journal of Privacy and Confidentiality*, *7*(3), 17–51.

Dwork, C., & Roth, A. (2014a). The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, *9*(3–4), 211–407.

Dwork, C., & Roth, A. (2014b). The Algorithmic Foundations of Differential Privacy. *Foundations and Trends® in Theoretical Computer Science*, *9*(3–4), 211–407.

Dwork, C., Smith, A., Steinke, T., & Ullman, J. (2017). Exposed! A Survey of Attacks on Private Data. *Annual Review of Statistics and Its Application (2017)*.

El Emam, K., Dankar, F. K., Neisa, A., & Jonker, E. (2013). Evaluating the risk of patient re-identification from adverse drug event reports. *BMC medical informatics and decision making*, *13*, 1–14.

European Parliament and Council. (2016). General data protection regulation (gdpr) - article 5 principles relating to processing of personal data [Accessed: 2024-09-27].

European Union. (2016). Recital 26: not applicable to anonymous data: Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 [Accessed: 2025-07-11].

Even, A., & Shankaranarayanan, G. (2007). Utility-driven assessment of data quality. *ACM SIGMIS Database: the DATABASE for Advances in Information Systems*, *38*(2), 75–93.

Feldman, V., & Zhang, C. (2020). What neural networks memorize and why: discovering the long tail via influence estimation. *Advances in Neural Information Processing Systems*, *33*, 2881–2891.

Fleckenstein, M., Obaidi, A., & Tryfona, N. (2023). A review of data valuation approaches and building and scoring a data valuation model. *Harvard Data Science Review*, *5*(1).

Francis, P., Probst-Eide, S., Obrok, P., Berneanu, C., Juric, S., & Munz, R. (2019, August 21). *Diffix-Birch: Extending Diffix-Aspen*. arXiv: 1806.02075 [cs].

Friedman, J. H. (1997). On bias, variance, 0/1—loss, and the curse-of-dimensionality. *Data mining and knowledge discovery*, *1*, 55–77.

Fudenberg, D., & Tirole, J. (1991). *Game theory*. MIT press.

Fung, B. C., Wang, K., Chen, R., & Yu, P. S. (2010). Privacy-preserving data publishing: a survey of recent developments. *ACM Computing Surveys (Csur)*, *42*(4), 1–53.

Gao, T., Yen, H., Yu, J., & Chen, D. (2023). Enabling large language models to generate text with citations. *arXiv preprint arXiv:2305.14627*.

Garfinkel, S. (2015). *De-identification of personal information:*. US Department of Commerce, National Institute of Standards; Technology.

Garrido, G. M., Liu, X., Matthes, F., & Song, D. (2022). Lessons learned: surveying the practicality of differential privacy in the industry. *arXiv preprint arXiv:2211.03898*.

Ghorbani, A., Abid, A., & Zou, J. (2019a). Interpretation of neural networks is fragile. *Proceedings of the AAAI conference on artificial intelligence*, *33*(01), 3681–3688.

Ghorbani, A., Abid, A., & Zou, J. (2019b). Interpretation of Neural Networks Is Fragile [Number: 01]. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(01), 3681–3688.

Ghorbani, A., Kim, M., & Zou, J. (2020). A distributional framework for data valuation. *International Conference on Machine Learning*, 3535–3544.

Ghorbani, A., & Zou, J. (2019). Data shapley: equitable valuation of data for machine learning. *International conference on machine learning*, 2242–2251.

Ghosh, A., & Roth, A. (2011). Selling privacy at auction. *Proceedings of the 12th ACM conference on Electronic commerce*, 199–208.

Giomi, M., Boenisch, F., Wehmeyer, C., & Tasnádi, B. (2022, November 18). *A Unified Framework for Quantifying Privacy Risk in Synthetic Data*. arXiv: 2211.10459 [cs].

Gootjes-Dreesbach, L., Sood, M., Sahay, A., Hofmann-Apitius, M., & Fröhlich, H. (2020). Variational Autoencoder Modular Bayesian Networks for Simulation of Heterogeneous Clinical Study Data. *Frontiers in Big Data*, *3*.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, 857–871.

Green, T., Gubri, M., Puerto, H., Yun, S., & Oh, S. J. (2025). Leaky thoughts: large reasoning models are not private thinkers. *arXiv preprint arXiv:2506.15674*.

Grimmelmann, J. (2008). Saving facebook. *Iowa L. Rev.*, *94*, 1137.

Hayes, J., Shumailov, I., Choquette-Choo, C. A., Jagielski, M., Kaissis, G., Lee, K., Nasr, M., Ghalebikesabi, S., Mireshghallah, N., & Annamalai, M. S. M. S. (2025). Strong membership inference attacks on massive datasets and (moderately) large language models. *arXiv preprint arXiv:2505.18773*.

Heffetz, O., & Ligett, K. (2014). Privacy and data-based research. *Journal of Economic Perspectives*, *28*(2), 75–98.

Henriksen-Bulmer, J., & Jeary, S. (2016). Re-identification attacks—A systematic literature review. *International Journal of Information Management*, *36*, 1184–1192.

Hilprecht, B., Härterich, M., & Bernau, D. (2019). Monte Carlo and Reconstruction Membership Inference Attacks against Generative Models. *Proceedings on Privacy Enhancing Technologies*, *2019*(4), 232–249.

Hoofnagle, C. J., Van Der Sloot, B., & Borgesius, F. Z. (2019). The european union general data protection regulation: what it is and what it means. *Information & Communications Technology Law*, *28*(1), 65–98.

Houssiau, F., Jordon, J., Cohen, S. N., Daniel, O., Elliott, A., Geddes, J., Mole, C., Rangel-Smith, C., & Szpruch, L. (2022, November 11). *TAPAS: a Toolbox for Adversarial Privacy Auditing of Synthetic Data*. arXiv: 2211.06550 [cs].

Hsu, C.-L., Liao, Y.-C., Lee, C.-W., & Chan, L. K. (2022). Privacy concerns and information sharing: the perspective of the u-shaped curve. *Frontiers in Psychology*, *13*, 771278.

Hubinger, E., Denison, C., Mu, J., Lambert, M., Tong, M., MacDiarmid, M., Lanham, T., Ziegler, D. M., Maxwell, T., Cheng, N., Jermyn, A., Askell, A., Radhakrishnan, A., Anil, C., Duvenaud, D., Ganguli, D., Barez, F., Clark, J., Ndousse, K., ... Perez, E. (2024). Sleeper agents: training deceptive llms that persist through safety training. *arXiv preprint arXiv:2401.05566*.

Hupfer, S., Radin, J., Silverglate, P., & Steinhart, M. (2023, November). *Tech companies have a trust gap to overcome—especially with women*. Deloitte Insights. Retrieved July 4, 2025, from https://www.deloitte.com/us/en/insights/industry/technology/bridge-data-privacy-concerns-in-women-with-technology.html

IBM. (2022). Cost of a data breach 2022.

Ilyas, A., Park, S. M., Engstrom, L., Leclerc, G., & Madry, A. (2022). Datamodels: predicting predictions from training data. *arXiv preprint arXiv:2202.00622*.

International Association of Privacy Professionals. (2024). *Privacy governance report 2024: executive summary* (tech. rep.) (Accessed: 2025-07-02). International Association of Privacy Professionals (IAPP).

Jia, R., Dao, D., Wang, B., Hubis, F. A., Gurel, N. M., Li, B., Zhang, C., Spanos, C. J., & Song, D. (2019). Efficient task-specific data valuation for nearest neighbor algorithms. *arXiv preprint arXiv:1908.08619*.

Jia, R., Dao, D., Wang, B., Hubis, F. A., Hynes, N., Gürel, N. M., Li, B., Zhang, C., Song, D., & Spanos, C. J. (2019a). Towards efficient data valuation based on the shapley value. *The 22nd International Conference on Artificial Intelligence and Statistics*, 1167–1176.

Jia, R., Dao, D., Wang, B., Hubis, F. A., Hynes, N., Gürel, N. M., Li, B., Zhang, C., Song, D., & Spanos, C. J. (2019b). Towards Efficient Data Valuation Based on the Shapley Value. *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, 1167–1176.

Jia, R., Wu, F., Sun, X., Xu, J., Dao, D., Kailkhura, B., Zhang, C., Li, B., & Song, D. (2021). Scalability vs. utility: do we have to sacrifice one for the other in data importance quantification? *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8239–8247.

Just, H. A., Kang, F., Wang, J. T., Zeng, Y., Ko, M., Jin, M., & Jia, R. (2023). Lava: data valuation without pre-specified learning algorithms. *arXiv preprint arXiv:2305.00054*.

Kaplan, S., & Garrick, B. J. (1981). On the quantitative definition of risk. *Risk analysis*, *1*(1), 11–27.

Kasiviswanathan, S. P., Lee, H. K., Nissim, K., Raskhodnikova, S., & Smith, A. (2010, February). What Can We Learn Privately? [arXiv:0803.0924 [cs]].

Khandelwal, U., Levy, O., Jurafsky, D., Zettlemoyer, L., & Lewis, M. (2019). Generalization through memorization: nearest neighbor language models. *arXiv preprint arXiv:1911.00172*.

Kim, J., Nakamaki, T., & Suzuki, T. (2024). Transformers are minimax optimal nonparametric in-context learners. *Advances in Neural Information Processing Systems*, *37*, 106667–106713.

Klymenko, O., Meisenbacher, S., & Matthes, F. (2023). Identifying practical challenges in the implementation of technical measures for data privacy compliance. *arXiv preprint arXiv:2306.15497*.

Koh, P. W., & Liang, P. (2017a). Understanding black-box predictions via influence functions. *International conference on machine learning*, 1885–1894.

Koh, P. W., & Liang, P. (2017b). Understanding Black-box Predictions via Influence Functions [ISSN: 2640-3498]. *Proceedings of the 34th International Conference on Machine Learning*, 1885–1894.

Kohavi, R. (1996). Scaling up the accuracy of naive-bayes classifiers: a decision-tree hybrid. *Kdd*, *96*, 202–207.

Kokolakis, S. (2017). Privacy attitudes and privacy behaviour: a review of current research on the privacy paradox phenomenon. *Computers & security*, *64*, 122–134.

Krizhevsky, A., & Hinton, G. (2009). Learning multiple layers of features from tiny images.

Kwon, Y., & Zou, J. (2022, January 18). *Beta Shapley: a Unified and Noise-reduced Data Valuation Framework for Machine Learning*. arXiv: `2110.14049 [cs, stat]`. Retrieved March 13, 2023, from http://arxiv.org/abs/2110.14049

LeFevre, K., DeWitt, D., & Ramakrishnan, R. (2006). Mondrian Multidimensional K-Anonymity. *22nd International Conference on Data Engineering (ICDE'06)*, 25–25.

LeFevre, K., DeWitt, D. J., & Ramakrishnan, R. (2006). Mondrian multidimensional k-anonymity. *22nd International conference on data engineering (ICDE'06)*, 25–25.

Li, N., Li, T., Venkatasubramanian, S., & Labs, T. (2007). T-Closeness: Privacy Beyond k-Anonymity and -Diversity.

Li, T., & Li, N. (2009). On the tradeoff between privacy and utility in data publishing. *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 517–526.

Liu, F., Lin, T., & Jaggi, M. (2021). Understanding memorization from the perspective of optimization via efficient influence estimation. *arXiv preprint arXiv:2112.08798*.

Liu, Z., Luo, P., Wang, X., & Tang, X. (2018). Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, *15*(2018), 11.

Lu, P.-H., Wang, P.-C., & Yu, C.-M. (2019). Empirical Evaluation on Synthetic Data Generation with Generative Adversarial Network. *Proceedings of the 9th International Conference on Web Intelligence, Mining and Semantics*, 1–6.

Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, *30*.

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). Learning word vectors for sentiment analysis. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 142–150.

Machanavajjhala, A., Kifer, D., Gehrke, J., & Venkitasubramaniam, M. (2007a). L-diversity: privacy beyond k-anonymity. *Acm transactions on knowledge discovery from data (tkdd)*, *1*(1), 3–es.

Machanavajjhala, A., Kifer, D., Gehrke, J., & Venkitasubramaniam, M. (2007b). L-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data*, *1*(1), 3–es.

Mahanti, R., & Mahanti, R. (2021). *Data governance and compliance*. Springer.

Maleki, S., Tran-Thanh, L., Hines, G., Rahwan, T., & Rogers, A. (2013). Bounding the estimation error of sampling-based shapley value approximation. *arXiv preprint arXiv:1306.4265*.

McSherry, F., & Talwar, K. (2007). Mechanism design via differential privacy. *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, 94–103.

Mesana, P., Bénesse, C., Lautraite, H., Caporossi, G., & Gambs, S. (2024). Waka: data attribution using k-nearest neighbors and membership privacy principles. *arXiv preprint arXiv:2411.01357*.

Mireshghallah, N., Kim, H., Zhou, X., Tsvetkov, Y., Sap, M., Shokri, R., & Choi, Y. (2023). Can llms keep a secret? testing privacy implications of language models via contextual integrity theory. *arXiv preprint arXiv:2310.17884*.

Molnar, C. (2020). *Interpretable machine learning*. Lulu. com.

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

Munilla Garrido, G., Liu, X., Matthes, F., & Song, D. (2023). Lessons Learned: Surveying the Practicality of Differential Privacy in the Industry. *Proceedings on Privacy Enhancing Technologies*, *2023*(2), 151–170.

Myerson, R. B. (1983). Mechanism design by an informed principal. *Econometrica: Journal of the Econometric Society*, 1767–1797.

Narayanan, A., & Shmatikov, V. (2008). Robust De-anonymization of Large Sparse Datasets. *2008 IEEE Symposium on Security and Privacy (Sp 2008)*, 111–125.

Nasr, M., Hayes, J., Steinke, T., Balle, B., Tramèr, F., Jagielski, M., Carlini, N., & Terzis, A. (2023). Tight auditing of differentially private machine learning. *Proceedings of the 32nd USENIX Security Symposium (USENIX Security 23)*, 1631–1648.

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

National Assembly of Québec. (2021). An act to modernize legislative provisions as regards the protection of personal information (law 25) [Phased implementation from 2022 to 2024].

National Institute of Standards and Technology. (2020). *Nist privacy framework: a tool for improving privacy through enterprise risk management* (tech. rep. No. NISTIR 8062) (Version 1.0). National Institute of Standards and Technology.

Ngo, R., Chan, L., & Mindermann, S. (2022). The alignment problem from a deep learning perspective. *arXiv preprint arXiv:2209.00626*.

Nguyen, T. (2024). Understanding transformers via n-gram statistics. *Advances in neural information processing systems*, *37*, 98049–98082.

Nissenbaum, H. (2004). Privacy as contextual integrity. *Wash. L. Rev.*, *79*, 119.

OECD. (2013). The oecd privacy framework [Updated version of the 1980 Guidelines on the Protection of Privacy and Transborder Flows of Personal Data].

Omohundro, S. M. (1989). *Five Balltree Construction Algorithms | PDF | Algorithms And Data Structures | Areas Of Computer Science*.

Opinion 05/2014 on anonymisation techniques (Technical report No. 0829/14/EN, WP216). (2014, April). ARTICLE 29 DATA PROTECTION WORKING PARTY.

Osborne, M. J., & Rubinstein, A. (1994). *A course in game theory*. MIT press.

Pai, M. M., & Roth, A. (2013). Privacy and mechanism design. *ACM SIGecom Exchanges*, *12*(1), 8–29.

Park, N., Mohammadi, M., Gorde, K., Jajodia, S., Park, H., & Kim, Y. (2018). Data Synthesis based on Generative Adversarial Networks. *Proceedings of the VLDB Endowment*, *11*(10), 1071–1083.

Park, S. M., Georgiev, K., Ilyas, A., Leclerc, G., & Madry, A. (2023). Trak: attributing model behavior at scale. *arXiv preprint arXiv:2303.14186*.

Parliament, E., & of the European Union, C. (2016). Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation), art. 4(1) [Official Journal of the European Union, L 119, 1-88; cor. OJ L 127, 23.5.2018].

Patki, N., Wedge, R., & Veeramachaneni, K. (2016). The synthetic data vault. *IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 399–410.

Pei, J. (2022). A Survey on Data Pricing: From Economics to Data Science. *IEEE Transactions on Knowledge and Data Engineering*, *34*(10), 4586–4608.

Peyré, G., & Cuturi, M. (2018). Computational optimal transport [arXiv: 1803.00567]. *arXiv:1803.00567 [stat]*.

Rashid, A. H., & Yasin, N. B. M. (2015). Privacy preserving data publishing. *International Journal of Physical Sciences*, *10*(7), 239–247.

Reimers, N. (2019). Sentence-bert: sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Roth, A. E. (1988, October 28). *The Shapley Value: Essays in Honor of Lloyd S. Shapley*. Cambridge University Press.

Shapley, L. S. (1953). Stochastic Games*. *Proceedings of the National Academy of Sciences*, *39*(10), 1095–1100.

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

Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017b). Membership Inference Attacks Against Machine Learning Models. *IEEE Symposium on Security and Privacy*, 3–18.

Simaan, M., & Cruz Jr, J. B. (1973). On the stackelberg strategy in nonzero-sum games. *Journal of Optimization Theory and Applications*, *11*(5), 533–555.

Skinner, C. J., & Elliot, M. J. (2002). A measure of disclosure risk for microdata. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*(4), 855–867.

Skinner, C. J., & Holmes, D. J. (1998). Estimating the Re-identi®cation Risk Per Record in Microdata.

Staab, R., Vero, M., Balunović, M., & Vechev, M. (2023). Beyond memorization: violating privacy via inference with large language models. *arXiv preprint arXiv:2310.07298*.

Stadler, T., Oprisanu, B., & Troncoso, C. (2022, January 24). *Synthetic Data – Anonymisation Groundhog Day*. arXiv: 2011.07018 [cs].

Steinke, T., Nasr, M., & Jagielski, M. (2023). Privacy Auditing with One (1) Training Run. *Advances in Neural Information Processing Systems*, *36*, 49268–49280.

Strong, D. M., Lee, Y. W., & Wang, R. Y. (1997). Data quality in context. *Communications of the ACM*, *40*(5), 103–110.

Sun, H., Chen, Y., Wang, S., Chen, W., & Deng, X. (2024). Mechanism design for llm fine-tuning with multiple reward models. *arXiv preprint arXiv:2405.16276*.

Sweeney, L. (2002). K-ANONYMITY: A MODEL FOR PROTECTING PRIVACY. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *10*(05), 557–570.

Tang, S., Ghorbani, A., Yamashita, R., Rehman, S., Dunnmon, J. A., Zou, J., & Rubin, D. L. (2021a). Data valuation for medical imaging using shapley value and application to a large-scale chest x-ray dataset. *Scientific reports*, *11*(1), 8366.

Tang, S., Ghorbani, A., Yamashita, R., Rehman, S., Dunnmon, J. A., Zou, J., & Rubin, D. L. (2021b). Data valuation for medical imaging using Shapley value and application to a large-scale chest X-ray dataset. *Scientific Reports*, *11*(1), 8366.

Tian, Z., Liu, J., Li, J., Cao, X., Jia, R., & Ren, K. (2022, December 21). *Private Data Valuation and Fair Payment in Data Marketplaces*. arXiv: `2210.08723 [cs]`. Retrieved February 7, 2023, from http://arxiv.org/abs/2210.08723

Vapnik, V. (1991). Principles of risk minimization for learning theory. *Advances in neural information processing systems*, *4*.

Vial, G., Crowe, J., & Mesana, P. (2024). Managing data privacy risk in advanced analytics [Vial et al.]. *MIT Sloan Management Review*.

Wan, Z., Zhang, Y., & He, H. (2017). Variational autoencoder based synthetic data generation for imbalanced learning. *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, 1–7.

Wang, J. T., & Jia, R. (2023, March 1). *Data Banzhaf: A Robust Data Valuation Framework for Machine Learning*. arXiv: `2205.15466 [cs, stat]`. Retrieved July 6, 2023, from http://arxiv.org/abs/2205.15466

Wang, J. T., Mittal, P., Song, D., & Jia, R. (2024). Data shapley in one training run. *arXiv preprint arXiv:2406.11011*.

Wang, J. T., Yang, T., Zou, J., Kwon, Y., & Jia, R. (2024, May 6). *Rethinking Data Shapley for Data Selection Tasks: Misleads and Merits*. arXiv: 2405.03875 `[cs]`.

Wang, N., Xiao, X., Yang, Y., Zhao, J., Hui, S. C., Shin, H., Shin, J., & Yu, G. (2019). Collecting and analyzing multidimensional data with local differential privacy. *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, 638–649.

Warren, S. D., & Brandeis, L. D. (1890). The right to privacy [Seminal articulation of privacy as the "right to be let alone"]. *Harvard Law Review*, *4*(5), 193–220.

Worledge, T., Shen, J. H., Meister, N., Winston, C., & Guestrin, C. (2024). Unifying corroborative and contributive attributions in large language models. *2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, 665–683.

Xia, H., Liu, J., Lou, J., Qin, Z., Ren, K., Cao, Y., & Xiong, L. (2023). Equitable Data Valuation Meets the Right to Be Forgotten in Model Markets. *Proceedings of the VLDB Endowment*, *16*(11), 3349–3362.

Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019a). Modeling tabular data using conditional gan. *Advances in neural information processing systems*, *32*.

Xu, L., Skoularidou, M., Cuesta-Infante, A., & Veeramachaneni, K. (2019b, October 27). *Modeling Tabular data using Conditional GAN*. arXiv: 1907.00503 `[cs, stat]`.

Yadav, C., & Chaudhuri, K. (2021). Behavior of k-nn as an instance-based explanation method. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 90–96.

Yan, T., & Procaccia, A. D. (2021). If You Like Shapley Then You'll Love the Core. *Proceedings of the AAAI Conference on Artificial Intelligence*, *35*(6), 5751–5759.

Ye, J., Borovykh, A., Hayou, S., & Shokri, R. (2023). Leave-one-out distinguishability in machine learning. *arXiv preprint arXiv:2309.17310*.

Ye, J., Maddi, A., Murakonda, S. K., Bindschaedler, V., & Shokri, R. (2022a). Enhanced membership inference attacks against machine learning models. *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, 3093–3106.

Ye, J., Maddi, A., Murakonda, S. K., Bindschaedler, V., & Shokri, R. (2022b, September). Enhanced Membership Inference Attacks against Machine Learning Models [arXiv:2111.09679 [cs]].

Yeh, I.-C., & Lien, C.-h. (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert systems with applications*, *36*(2), 2473–2480.

Yeom, S., Giacomelli, I., Fredrikson, M., & Jha, S. (2018). Privacy risk in machine learning: analyzing the connection to overfitting. *2018 IEEE 31st computer security foundations symposium (CSF)*, 268–282.

Yoon, J., Arik, S., & Pfister, T. (2020). Data valuation using reinforcement learning. *International Conference on Machine Learning*, 10842–10851.

Zhang, J., Das, D., Kamath, G., & Tramèr, F. (2024). Membership inference attacks cannot prove that a model was trained on your data. *arXiv preprint arXiv:2409.19798*.

Zhang, X., Zhao, J., & LeCun, Y. (2015). Character-level convolutional networks for text classification. *Advances in neural information processing systems*, *28*.

Zhang, Y., Zhang, F., Yang, Z., & Wang, Z. (2023). What and how does in-context learning learn? bayesian model averaging, parameterization, and generalization. *arXiv preprint arXiv:2305.19420*.