Machine Unlearning: Challenges in Data Quality and Access

Miao Xu
School of Electrical Engineering and Computer Science,
University of Queensland, Brisbane, Australia
miao.xu@uq.edu.au

Abstract

Machine unlearning aims to remove specific knowledge from a well-trained machine learning model. This topic has gained significant attention recently due to the widespread adoption of machine learning models across various applications and the accompanying privacy, legal, and ethical considerations. During the unlearning process, models are typically presented with data that specifies which information should be erased and which should be retained. Nonetheless, practical challenges arise due to prevalent issues of data quality issues and access restrictions. This paper explores these challenges and introduces strategies to address problems related to unsupervised data, weakly supervised data, and scenarios characterized by zero-shot and federated data availability. Finally, we discuss related open questions, particularly concerning evaluation metrics, how the forgetting information is represented and delivered, and the unique challenges posed by large generative models.

1 Introduction

Machine unlearning [Bourtoule *et al.*, 2021] aims to remove specific information from a well-trained machine learning model. It arises from concerns about private information leakage, ensuring the "right-to-be-forgotten" for particular data [Kwak *et al.*, 2017]. This right has been formally incorporated into legal regulations, such as the GDPR [Voigt and Von dem Bussche, 2017] and CCPA [de la Torre, 2018]. With the development and widespread use of generative models, there are also concerns about generating unethical content, highlighting the need to remove learned knowledge from these models [Liu *et al.*, 2024]. Due to privacy, legal, and ethical requirements, machine unlearning has gained significant attention in recent years. Various methods have been proposed to address unlearning problems with different levels of accuracy and for various applications. For more comprehensive details on machine unlearning, refer to surveys such as [Xu *et al.*, 2024; Nguyen *et al.*, 2022].

A typical machine unlearning problem can be formalized as follows [Bourtoule *et al.*, 2021]. Initially, a well-trained model is obtained from the training data. When a forgetting request is submitted, it usually includes forgetting data, which represents the information to be removed, and remaining data, which represents the information to be retained. The forgetting data and remaining data are partitions of the original training data. An unlearning algorithm is then designed to remove the information associated with the forgetting data from the model. Retraining the model using only the remaining data is a potential solution to this unlearning request [Zhang *et al.*, 2022; Di *et al.*, 2023]. However, due to concerns about efficiency or other constraints, retraining may not always be feasible. Consequently, researchers are exploring better strategies to approximate such retraining [Thudi *et al.*, 2022; Chundawat *et al.*, 2023b; Kurmanji *et al.*, 2023; Shen *et al.*, 2024a].

In a typical machine learning process, one crucial factor is the data—specifically, the information to be forgotten and the information to be retained. However, data quality and access issues are prevalent in machine learning. A common data quality issue is the weak supervision problem [Zhou, 2018] in supervised learning, where annotations are low-quality due to being incomplete, inaccurate, or inexact. Research in this area has proposed methods for learning from noisy labels [Han *et al.*, 2018], positive-unlabeled data [Su *et al.*, 2021], partial labels [Lv *et al.*, 2024], and complementary labels [Gao *et al.*, 2023]. Additionally, data access can be problematic, necessitating algorithms for zero-shot learning [Romera-Paredes and Torr, 2015], which assumes no data is available for a particular domain, or federated learning [Konečný *et al.*,], which assumes data is distributed across different nodes and cannot be shared. These data challenges are fundamental in machine learning due to the costs of data collection and labeling, as well as restrictions on data sharing. In the context of machine unlearning, similar issues could also arise. It is worth exploring what strategies can be implemented to overcome data quality and access problems in unlearning.

In this paper, we explore various approaches and methodologies to address data quality and access issues in the context of machine unlearning. For the quality challenge, we introduce works that enable machine unlearning when the data is weakly supervised—a supervision-free method Label-Agnostic Forgetting [Shen *et al.*, 2024b] and a universal weakly supervised method [Tang *et al.*, 2024]. These methods address the limitations of traditional unlearning methods that rely on well-annotated datasets. Label-agnostic forgetting leverages variational autoencoders to model data distributions, allowing the system to unlearn specific data without exact label information. Weakly supervised unlearning focuses on the uniformity of the model's prediction distribution, diluting distinctiveness in model predictions without direct label reliance. We then shift our attention to the data access problem, particularly when the intention is to forget a particular class of data but no forgetting or remaining data is provided, except the class name [Chundawat *et al.*, 2023a; Zhang *et al.*, 2024]. Additionally, we discuss briefly on a federated learning setting, where the objective is to forget knowledge learned from a particular node [Liu *et al.*, 2021; Cao *et al.*, 2023; Fraboni *et al.*, 2024]. These techniques collectively address both the quality and accessibility challenges in machine unlearning, paving the way for more robust and flexible unlearning methods. After introducing these current advancements, we will provide our reflections and discuss open problems related to these issues. Evaluation metrics need to be designed to address the unique challenges and diverse requirements of machine unlearning tasks, particularly those affected by data quality and access issues. Additionally, innovative methods are needed for online unlearning, interaction-based unlearning in large language models, and forgetting information expressed through natural language or specific data-based requests. These open problems require further research and exploration to advance the field.

2 Quality Challenge

Traditional unlearning methods predominantly rely on supervised learning processes, which necessitate well-annotated datasets. This requirement presents a substantial challenge, as acquiring extensive, high-quality labels is often costly, time-consuming, and sometimes infeasible, especially in scenarios where data privacy must be preserved during the unlearning process. Furthermore, in many real-world applications, significant portions of data remain unannotated or are weakly labeled, limiting the applicability of conventional unlearning methods that depend on precise label information to differentiate between data to be forgotten and data to be retained.

Addressing these challenges, two innovative works, the Label-Agnostic Forgetting [Shen *et al.*, 2024b] and the weakly supervised unlearning [Tang *et al.*, 2024], offer groundbreaking approaches to machine unlearning that minimize reliance on labeled data. As in the machine learning case, these two works also evolve from the unsupervised to weakly supervised paradigms reflects varying degrees of reliance on labeled data.

Label-Agnostic Forgetting. A deep model g is composed of a representation extractor g^e and a downstream classifier g^c. Since no label is available in label-agnostic forgetting, a potential approach would be to adjust the representation extractor but leave the downstream classifier untouched. Based on this basic idea, the Label Agnostic Forgetting (LAF) method [Shen *et al.*, 2024b] divides the unlearning into two steps– extractor unlearning and representation alignment.

In the extractor unlearning phase, LAF aims to learn an extractor that preserves the distribution of the remaining data while dissolving the distribution of the forgetting data. This is achieved by optimizing the following objective:

$$\theta^* = \arg\min_{\theta} \Delta(Q(D_r), \mathcal{P}_r) - \Delta(Q(D_f), \mathcal{P}_f),$$

where $\Delta(Q(D_r), \mathcal{P}_r)$ represents the discrepancy between the distributions of the objective extractor on reaming data and the true remaining data, and $\Delta(Q(D_f), \mathcal{P}_f)$ represents the discrepancy between the distributions of the objective extractor on forgetting data and the true forgetting data.

To mimic the distributions of the remaining data D_r and the forgetting data D_f, two VAEs (Variational Autoencoders) [Kingma and Welling, 2014] are trained to approximate them. For efficiency, the VAE for D_r is trained on the original model, leveraging its ability to capture the distribution of D to approximate D_r. This is more efficient given the typically smaller size of the forgetting data compared to the remaining data. In contrast, the VAE for D_f is specifically trained on the forgetting data.

After the extractor adjustment, LAF performs representation alignment to mitigate performance degradation caused by approximating D_r's representation using D's representation and to better align with the classifier layer. In this phase, a small subset of the remaining data can be employed in the following contrastive-learning style loss function to align the representation of the adjusted extractor g^e_U against the original extractor g^e_D:

$$\sum_{x \in X_r} \log \left(\frac{\exp(\text{simloss}(g^e_U(x), g^e_D(x)))}{\sum_{\hat{x} \in X_f} \exp(\text{simloss}(g^e_U(\hat{x}), g^e_D(\hat{x}))/\tau)} \right),$$

where X_r is a subset of the remaining data, and $\text{simloss}(\cdot, \cdot)$ is the similarity loss function, such as cosine similarity. This method enables the LAF approach to achieve unlearning and alignment without relying on labeled data. In cases where some training data with extra annotations is available, an additional supervised repairing stage can be employed to finetune the model using labeled data.

Weakly-Supervised Unlearning. Compared to the label-agnostic unlearning introduced earlier, weakly-supervised unlearning directly utilizes prepared weak labels, particularly partial labels [Lv *et al.*, 2024], where an instance is associated with a candidate label set instead of a single label, and noisy labels [Han *et al.*, 2018], where the provided labels may be incorrect. The designed weakly-supervised method is based on the belief that a trained model, after removing specific data, should exhibit similar predictions for these specific data as an untrained model that makes random guesses. With this belief, weakly-supervised unlearning [Tang *et al.*, 2024] is designed to access only the forgetting data, without remaining data.

In [Tang *et al.*, 2024], a label transformation method, denoted as $\boldsymbol{T}_i(\boldsymbol{z}, y)$, is designed. This method transforms a classifier's softmax output $\boldsymbol{z}$ for label i depending on the ground truth y. This transformation ensures an output as the soft labels for unlearning, which attain minimal divergence from the uniform distribution while maintaining performance on all outputs except the ground-truth one. Since the ground truth y may not be accurate in partial label learning or noisy label learning, Tang *et al.* adjust $\boldsymbol{T}_i(\cdot,\cdot)$ to ensure that all potential labels are considered and all relevant information is forgotten. Learning is then conducted using these transformed labels as the soft labels for the forgetting data. This weakly-supervised learning approach not only directly uses the weakly supervised labels through transformation but also eliminates the reliance on the remaining data by using only the forgetting data.

3 Access Challenge

When introducing weakly-supervised unlearning, we emphasized that it does not rely on remaining data. Generally, due to various regulations, data access may not always be possible, but at the very least, forgetting data may be essential. However, in a specific unlearning scenario—class-unlearning, which aims to unlearn an entire class of data—it is possible to achieve pure zero-shot unlearning, meaning only a class name is given without any data. Below, we first introduce this zero-shot unlearning problem, or more accurately, zero-shot class-unlearning [Chundawat *et al.*, 2023a; Zhang *et al.*, 2024]. Next, we briefly discuss federated unlearning [Liu *et al.*, 2021; Cao *et al.*, 2023; Fraboni *et al.*, 2024], another form of data access restriction, where data needs to be distributed across different nodes and cannot be transferred between them. Note that for general zero-shot unlearning, the scenario where unlearning occurs with forgetting data but without remaining data is still under-explored. This may be a promising future direction to investigate.

Zero-shot Unlearning. The Gated Knowledge Transfer (GKT) method [Chundawat *et al.*, 2023a] first addressed the zero-shot unlearning problem, which involves removing specific class information without accessing the original training data. It uses pseudo data optimization through error minimization-maximization, generating anti-samples to maximize the target loss for the forgetting class and minimize it for the retaining class. This allows GKT to adjust the model's memory to selectively erase data traces while preserving its utility for other tasks. GKT also includes a gated mechanism that filters generated samples, ensuring those less likely to belong to the forgetting class are used for retraining. It then employs a "teacher-student" dynamic, where a newly initialized student network is trained under the guidance of the original (teacher) model, enabling the transfer of essential knowledge while omitting unwanted information. By using generated pseudo samples, GKT performs unlearning tasks without needing the original data, addressing privacy concerns and reducing storage and computational demands.

Inspired by GKT, a new method called GENIU [Zhang *et al.*, 2024] has been proposed. GENIU follows the generative-based unlearning pipeline in GKT and uses a set of generated proxy data to facilitate unlearning. Unlike GKT, GENIU includes both a training phase and an unlearning phase. During the training phase, a generator is trained alongside the classifier using noise samples to preserve information about class features. In the unlearning phase, these noise samples and the generator produce reliable proxy samples, which are then used to update the classifier using an in-batch tuning method. Similar to LAF, the proxy generator in GENIU employs a VAE structure to generate proxies that accurately represent class characteristics, even under imbalanced data conditions. The noise samples are optimized to be correctly classified by the classifier, serving as prior knowledge for generating proxy samples. In-batch tuning ensures that proxies to be unlearned increase the model's error while proxies to be retained reduce the error. Supervision samples, selected based on maximum logit entropy, guide the training of the generator to ensure proxies are near the decision boundary. In the unlearning phase, the generator creates proxies used to adjust the classifier, resulting in an unlearned model that effectively forgets specified classes without degrading performance on retained knowledge. GENIU emphasizes the need to prepare for potential unlearning tasks early in the training phase. It incorporates additional information within the trained model to enable efficient and reliable unlearning without needing the original training data.

Federated Unlearning. Federated unlearning [Gao *et al.*, 2024; Liu *et al.*, 2021; Cao *et al.*, 2023; Liu *et al.*, 2022; Jin *et al.*, 2023; Wang *et al.*, 2022; Fraboni *et al.*, 2024; Che *et al.*, 2023; Xiong *et al.*, 2023] is a scenario where data access is limited due to data location and regulatory restrictions on data exchange. In federated unlearning tasks, the goal is often to unlearn data associated with a particular data location. A key challenge in federated unlearning is achieving efficiency, whether in terms of computational costs from retraining or storage costs from maintaining historical information. Current works on federated unlearning typically focus on either computational efficiency or storage efficiency. For example, FedEraser [Liu *et al.*, 2021] and SIFU [Fraboni *et al.*, 2024] use historical data storage for rapid client unlearning, while methods like Quantized FL [Xiong *et al.*, 2023] and Projected Gradient Ascent (PGA) [Halimi *et al.*, 2022] also require historical data storage. Other approaches [Liu *et al.*, 2022; Jin *et al.*, 2023] need a Hessian computation for storage efficiency. Thus, it is essential to develop an efficient federated unlearning method that balances both computational and storage requirements.

4 Open Problems

Despite the progress made in improving data quality and access for machine unlearning, many challenges still persist. Addressing these issues is essential for further advancements in the field. In this section, we highlight some of the key open problems that require additional research and exploration.

Evaluation. One of the key challenges in machine unlearning is the evaluation problem. As summarized in [Xu *et al.*, 2024], various metrics have been proposed to evaluate the performance of machine unlearning, including retraining, member inference attack [Shokri *et al.*, 2017], and relearning [Tarun *et al.*, 2021]. However, similar to other machine learning paradigms that employ different evaluation metrics, such as multi-label learning [Xu *et al.*, 2020], it remains unclear which evaluation method is most suitable for a given machine unlearning task. This is particularly challenging because different requirements, such as utility and privacy, may not always align towards the same goal. Recent work [Kurmanji *et al.*, 2023] has highlighted that unlearning is application-dependent and may have varying needs in different scenarios. This application-dependent need is especially relevant for unlearning tasks involving data quality and access issues. For instance, when dealing with various poor data quality or limited data access, retraining may not produce a sufficiently accurate model to serve as the gold standard in evaluation. Designing evaluation metrics that address the unique challenges and diverse requirements of machine unlearning tasks, particularly those affected by data quality and access issues, remains an open problem.

Forgetting Information Access and Expression. We have previously explored the zero-shot unlearning problem, where neither the data to be forgotten nor the remaining data is accessible for class unlearning. We proposed potential future research focused on instance unlearning without the availability of remaining data. In these models, the information targeted for forgetting is typically a subset of the training data. However, ethical regulation scenarios often necessitate unlearning information presented in different forms, such as natural language instructions (e.g., “forget all personal identifiers in a model trained on healthcare data”) or specific data-based requests (e.g., submitting a dataset containing outdated medical treatment information and requesting the model to forget all related outdated practices). These practical requests arise when users need to directly remove specific, potentially harmful or unethical information without access or time to examine the original training data. Relevant preliminary work in this area uses a one-word concept to accurately describe the unlearning request in complex data scenarios, moving beyond traditional data points to define what needs to be forgotten [Chang *et al.*, 2024]. These requests introduce several challenges to existing unlearning frameworks, which typically rely on data-based forgetting methods. One initial challenge is identifying the specific data within the training set that needs to be forgotten, followed by the challenge of verifying the success of the unlearning process, especially since only part of the information in these training subsets requires removal. This issue is particularly relevant for large generative models, which may not easily trace how generated content relates to the training data, yet still need to responsibly exclude sensitive topics from their outputs.

Online Unlearning. One pressing open problem in machine unlearning is handling continuous, incremental data deletion requests, which we term *online unlearning* to draw a parallel with online learning [Cesa-Bianchi and Lugosi, 2006], where models are updated continuously with incoming data rather than being trained in batches. One example of such online unlearning is on a social media platform, where users frequently request the deletion of their personal data, such as browsing history or interaction records, to protect their privacy. These requests are made independently and at different times, reflecting real-world scenarios where data removal requests arrive incrementally rather than in batches. Traditional batch unlearning methods struggle with this dynamic, facing significant challenges: accumulated performance degradation, where each unlearning operation causes the model’s performance to drop, resulting in a substantial decline over time; and inefficiency with high computational costs, as managing frequent requests individually requires repeated reprocessing of the remaining data, which is time-consuming and expensive. Additionally, these methods often require access to the original training data, which can be restricted by privacy policies and data access regulations. Addressing online unlearning necessitates developing innovative approaches that can efficiently handle a stream of data removal requests, maintain model performance, and operate without continuous access to the original training data, adapting to the dynamic nature of data deletion while minimizing performance degradation and computational costs.

Interaction-based Unlearning in LLM. The widespread use of LLM, such as ChatGPT, among non-expert users presents significant challenges to the process of unlearning. Non-expert users primarily interact with these models through chat interfaces, without access to the underlying training processes. This raises critical questions about how to ensure our sensitive information is forgotten by these models when we cannot directly control their training. One potential solution involves designing methods that allow users to provide specific data through interactions that effectively “fool” the model into forgetting or erasing previously learned sensitive information. Additionally, well-designed prompts are needed to test whether the model has successfully forgotten particular information. Current prompt approaches [Liu *et al.*, 2023] usually focus on improving interaction quality but do not intentionally serve the purpose of unlearning. However, if prompting approaches were adapted to ensure the erasure of sensitive information and to verify the success of the unlearning process, it could potentially be an effective solution. This approach would promote fairness in information privacy, allowing non-experts to manage their privacy without needing access to the model’s training processes.

5 Conclusion

In this paper, we focus particularly on the data quality and access issues for machine unlearning with classification problems. We introduce the supervision-free and weakly supervised machine unlearning problems and solutions, and the machine unlearning with limited data access, particularly class-based zero-shot unlearning and federated learning, challenges, and solutions. We give some open problems regarding evaluation, how the forgetting request is represented (instruction-based unlearning), how the forgetting request comes (online unlearning), and give our reflection on what new challenges generative model unlearning can face. We hope this paper can provide valuable insights and inspire future research in developing more efficient and effective machine unlearning techniques.

Acknowledgments

I would like to thank all my close collaborators on the topic of machine unlearning, especially PhD students, Mr Shaofei Shen and Mr Chenhao Zhang. This work was supported by the Australian Research Council (DE230101116, DP240103070).

References

[Bourtoule *et al.*, 2021] Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *Proceedings of the IEEE Symposium on Security and Privacy (SP)*, pages 141–159, 2021.

[Cao *et al.*, 2023] Xiaoyu Cao, Jinyuan Jia, Zaixi Zhang, and Neil Zhenqiang Gong. Fedrecover: Recovering from poisoning attacks in federated learning using historical information. In *Proceedings of the IEEE Symposium on Security and Privacy (SP)*, pages 1366–1383, 2023.

[Cesa-Bianchi and Lugosi, 2006] Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.

[Chang *et al.*, 2024] Wenhan Chang, Tianqing Zhu, Heng Xu, Wenjian Liu, and Wanlei Zhou. Class machine unlearning for complex data via concepts inference and data poisoning. *arXiv preprint arXiv:2405.15662*, 2024.

[Che *et al.*, 2023] Tianshi Che, Yang Zhou, Zijie Zhang, Lingjuan Lyu, Ji Liu, Da Yan, Dejing Dou, and Jun Huan. Fast federated machine unlearning with nonlinear functional theory. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 4241–4268, 2023.

[Chundawat *et al.*, 2023a] V. S. Chundawat, A. K. Tarun, M. Mandal, and M. S. Kankanhalli. Zero-shot machine unlearning. *IEEE Transactions on Information Forensics and Security*, pages 2345–2354, 2023.

[Chundawat *et al.*, 2023b] Vikram S. Chundawat, Ayush K. Tarun, Murari Mandal, and Mohan S. Kankanhalli. Can bad teaching induce forgetting? unlearning in deep networks using an incompetent teacher. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 7210–7217, 2023.

[de la Torre, 2018] Lydia de la Torre. A guide to the california consumer privacy act of 2018. *SSRN 3275571*, 2018.

[Di *et al.*, 2023] Jimmy Z. Di, Jack Douglas, Jayadev Acharya, Gautam Kamath, and Ayush Sekhari. Hidden poison: Machine unlearning enables camouflaged poisoning attacks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[Fraboni *et al.*, 2024] Yann Fraboni, Martin Van Waerebeke, Kevin Scaman, Richard Vidal, Laetitia Kameni, and Marco Lorenzi. Sifu: Sequential informed federated unlearning for efficient and provable client unlearning in federated optimization. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 3457–3465, 2024.

[Gao *et al.*, 2023] Yi Gao, Miao Xu, and Min-Ling Zhang. Unbiased risk estimator to multi-labeled complementary label learning. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3732–3740, 2023.

[Gao *et al.*, 2024] Xiangshan Gao, Xingjun Ma, Jingyi Wang, Youcheng Sun, Bo Li, Shouling Ji, Peng Cheng, and Jiming Chen. Verifi: Towards verifiable federated unlearning. *IEEE Transactions on Dependable and Secure Computing*, page PrePrints, 2024.

[Halimi *et al.*, 2022] Anisa Halimi, Swanand Kadhe, Ambrish Rawat, and Nathalie Baracaldo. Federated unlearning: How to efficiently erase a client in fl? *arXiv preprint arXiv:2207.05521*, 2022.

[Han *et al.*, 2018] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in Neural Information Processing Systems (NeurIPS)*, 31:8536–8546, 2018.

[Jin *et al.*, 2023] Ruinan Jin, Minghui Chen, Qiong Zhang, and Xiaoxiao Li. Forgettable federated linear learning with certified data removal. *arXiv preprint arXiv:2306.02216*, 2023.

[Kingma and Welling, 2014] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR)*, 2014.

[Konečný *et al.*,] Jakub Konečný, H. Brendan McMahan, Felix X. Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492*.

[Kurmanji *et al.*, 2023] Meghdad Kurmanji, Peter Triantafillou, Jamie Hayes, and Eleni Triantafillou. Towards unbounded machine unlearning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[Kwak *et al.*, 2017] Chanhee Kwak, Junyeong Lee, Kyuhong Park, and Heeseok Lee. Let machines unlearn - machine unlearning and the right to be forgotten. In *Americas Conference on Information Systems (AMCIS)*, 2017.

[Liu *et al.*, 2021] Gaoyang Liu, Xiaoqiang Ma, Yang Yang, Chen Wang, and Jiangchuan Liu. Federaser: Enabling efficient client-level data removal from federated learning models. In *IEEE/ACM International Symposium on Quality of Service (IWQOS)*, pages 1–10, 2021.

[Liu *et al.*, 2022] Yi Liu, Lei Xu, Xingliang Yuan, Cong Wang, and Bo Li. The right to be forgotten in federated learning: An efficient realization with rapid retraining. In *IEEE Conference on Computer Communications (INFOCOM)*, pages 1749–1758, 2022.

[Liu *et al.*, 2023] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):195:1–195:35, 2023.

[Liu *et al.*, 2024] Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Xiaojun

Xu, Yuguang Yao, Hang Li, Kush R Varshney, et al. Rethinking machine unlearning for large language models. *arXiv preprint arXiv:2402.08787*, 2024.

[Lv *et al.*, 2024] Jiaqi Lv, Biao Liu, Lei Feng, Ning Xu, Miao Xu, Bo An, Gang Niu, Xin Geng, and Masashi Sugiyama. On the robustness of average losses for partial-label learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):2569–2583, 2024.

[Nguyen *et al.*, 2022] Thanh Tam Nguyen, Thanh Trung Huynh, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, and Quoc Viet Hung Nguyen. A survey of machine unlearning. *arXiv preprint arXiv:2209.02299*, 2022.

[Romera-Paredes and Torr, 2015] Bernardino Romera-Paredes and Philip Torr. An embarrassingly simple approach to zero-shot learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 2152–2161, 2015.

[Shen *et al.*, 2024a] Shaofei Shen, Chenhao Zhang, Alina Bialkowski, Weitong Chen, and Miao Xu. Camu: Disentangling causal effects in deep model unlearning. In *Proceedings of the SIAM Conference on Data Mining (SDM)*, 2024.

[Shen *et al.*, 2024b] Shaofei Shen, Chenhao Zhang, Yawen Zhao, Alina Bialkowski, Tony Weitong Chen, and Miao Xu. Label-agnostic forgetting: A supervision-free unlearning in deep models. In *International Conference on Learning Representations (ICLR)*, 2024.

[Shokri *et al.*, 2017] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *Proceedings of the IEEE Symposium on Security and Privacy (SP)*, pages 3–18, 2017.

[Su *et al.*, 2021] Guangxin Su, Weitong Chen, and Miao Xu. Positive-unlabeled learning from imbalanced data. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2995–3001, 2021.

[Tang *et al.*, 2024] Yi Tang, Yi Gao, Yong-Gang Luo, Ju-Cheng Yang, Miao Xu, and Min-Ling Zhang. Unlearning from weakly supervised learning. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2024.

[Tarun *et al.*, 2021] Ayush K. Tarun, Vikram S. Chundawat, Murari Mandal, and Mohan S. Kankanhalli. Fast yet effective machine unlearning. *arXiv preprint arXiv:2111.08947*, 2021.

[Thudi *et al.*, 2022] Anvith Thudi, Gabriel Deza, Varun Chandrasekaran, and Nicolas Papernot. Unrolling SGD: understanding factors influencing machine unlearning. In *IEEE European Symposium on Security and Privacy (EuroSP)*, pages 303–319, 2022.

[Voigt and Von dem Bussche, 2017] Paul Voigt and Axel Von dem Bussche. The eu general data protection regulation (GDPR). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 2017.

[Wang *et al.*, 2022] Junxiao Wang, Song Guo, Xin Xie, and Heng Qi. Federated unlearning via class-discriminative pruning. In *Proceedings of the ACM Web Conference (WWW)*, pages 622–632, 2022.

[Xiong *et al.*, 2023] Zuobin Xiong, Wei Li, Yingshu Li, and Zhipeng Cai. Exact-fun: An exact and efficient federated unlearning approach. In *IEEE International Conference on Data Mining (ICDM)*, pages 1439–1444, 2023.

[Xu *et al.*, 2020] Miao Xu, Yufeng Li, and Zhi-Hua Zhou. Robust multi-label learning with PRO loss. *IEEE Transactions on Knowledge and Data Engineering*, 32(8):1610–1624, 2020.

[Xu *et al.*, 2024] Heng Xu, Tianqing Zhu, Lefeng Zhang, Wanlei Zhou, and Philip S. Yu. Machine unlearning: A survey. *ACM Computing Surveys*, 56(1):9:1–9:36, 2024.

[Zhang *et al.*, 2022] Peng-Fei Zhang, Guangdong Bai, Zi Huang, and Xin-Shun Xu. Machine unlearning for image retrieval: A generative scrubbing approach. In *ACM International Conference on Multimedia (MM)*, pages 237–245, 2022.

[Zhang *et al.*, 2024] Chenhao Zhang, Shaofei Shen, Yawen Zhao, Weitong Tony Chen, and Miao Xu. Geniu: A restricted data access unlearning for imbalanced data. *arXiv preprint arXiv:2406.07885*, 2024.

[Zhou, 2018] Zhi-Hua Zhou. A brief introduction to weakly supervised learning. *National science review*, 5(1):44–53, 2018.