Machine Unlearning and Model Editing

Naren Sivakumar
University of Baltimore County
1000 Hilltop Circle, Baltimore
narens1@umbc.edu

Abstract

The process of making a model "forget" specific data, machine unlearning ensures important data that has had its access revoked can be forgotten. This is especially important if the data the model is being trained on raises security and ethics concerns, like the GDPR mandate that gives individuals the "right to be forgotten". In order to comply with such ethical and legal responsibilities, there has to be some method in place to remove data that is considered to violate these conventions. Additionally, such techniques are also vital in making sure erroneous data that skews the model's accuracy and fairness can be removed to make the model better at the task at hand without retraining it completely. This is challenging because once the model learns data, it becomes difficult to "un-entangle" it from the parameters, and removing the entanglement is a challenge. This paper aims to study the field of machine unlearning in depth, starting from the origins and theory behind this idea and then summarizing all the state-of-the-art methods in this field. Finally, we discuss some of the problems that lay ahead in research in this field, and how researchers can hope to resolve them.

1. Introduction

With the ever-increasing influence of data and machine learning models in day-to-day life, data has become king in this information-driven era. With machine-learning models being developed to deal with tasks that increase in complexity cite models in real life here, a concern that grows with such models is data privacy. With privacy concerns growing over data being used unfairly to train large machine-learning models to be deployed in the real world and laws like the General Data Protection Regulation being passed, machine-learning scientists are faced with requests to delete data from machine-learning models that have been deployed. As a result, machine unlearning has grown in popularity, offering methods to "unlearn" certain data points while preserving the rest of the knowledge learned. Machine unlearning is the process of a model "forgetting data" that it has used to learn the task, something that is easier said than done. When a model trains on data, it encodes intricate dependencies and patterns from several data points, thereby showing us that machine unlearning is not simply a point-and-click delete operation that solves the problem. Researchers have been looking into methods that make machine unlearning a viable task, making sure the model's performance is unaffected by the removal of this data.

Some of the key objectives and challenges of machine unlearning are defined in figure 1. What we are more interested in this paper are the challenges faced by machine unlearning, with some of the most influential ones being:

- **Selective Data Removal:** Removing very specific pieces of data remains a challenging task to do. It is also hard to maintain the accuracy and overall performance of the model while doing so.
- **Provable Guarantees:** To comply with ethics and regulatory guidelines, there has to be some empirical proof that data has been removed. This mathematical proof is often difficult to get, and methods that do exist are extremely limited in scope and applicability.
- **Ethical Concerns:** On the flip side, machine unlearning may be used to skirt ethical regulations by unlearning data that promotes bias and unfair treatment.
- **Model Integrity and Stability:** Unlearning can cripple a model by inadvertently affecting other learned knowledge if carried out incorrectly, and also affect explainability and interpretability, something that is crucial in high-stakes applications that run on machine learning models.

In this paper, we will be learning about the theory behind machine unlearning, starting with the origins of this emerging field. The paper will summarize prevailing theories in the field, then shift to contemporary methods in this field that either prove or disprove these theories. Finally, this paper aims to understand and discuss some of the challenges and discuss potential solutions for them.

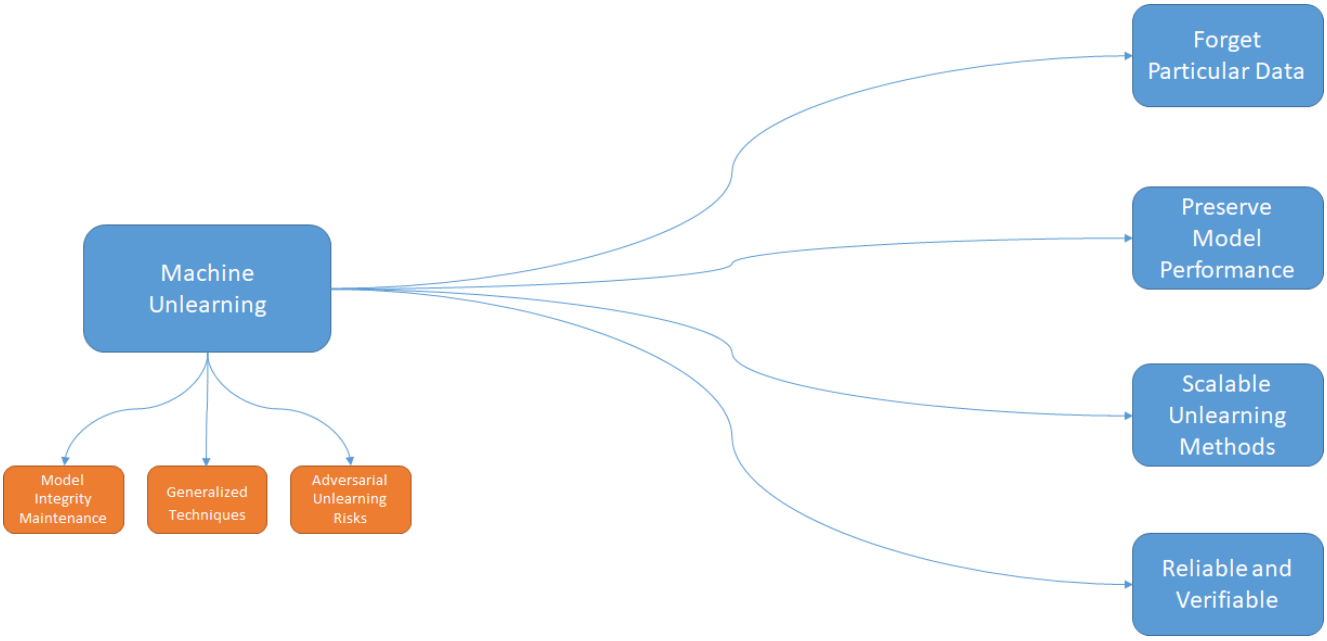

Figure 1. Some objectives and challenges in machine unlearning. Objectives are highlighted in blue, and challenges are in orange.

2. Machine Unlearning

Machine unlearning has many real-world applications that claim to be where the concept originated. In reality, however, it was formulated as a privacy-preserving problem [35]. This is further supported by [4, 8, 13, 14, 38] who show us the true meaning of data deletion, and how to make AI "forget" us. To formally define machine unlearning, we can assume that we have a sample space of data Z. If all the datasets D that train a model are some combination of Z, then the data that has to be removed Df that belongs to D. Now, the machine unlearning algorithm A has to take Df as input and return an unlearned model U that performs as well as a retrained model A* on the remainder of the data after removing Df. A problem that arises with machine learning models is that a majority of the models trained are stochastic, and therefore we cannot predict what impact each data point has on the model itself.

Now that we have a formal definition of machine unlearning, let's look at some of the removal tasks it could be faced with and some of the design requirements to build a good machine unlearning algorithm.

2.1. Unlearning Tasks

We cannot make the assumption that all data removal requests will be singular by nature. While data is certainly skewed towards such singular requests, such as in [39, 43, 44] we must also take into consideration some of the other types of requests to remove data from a machine learning model. Some of these requests are given below, from [35] in increasing order of size.

- **Item Removal:** Quite possibly the most popular understanding of the subject, a singular item removal is the most common type of removal task that one can come across [4].
- **Feature and Class Removal:** Data points that collectively influence the model to make decisions in the incorrect manner are seen in some cases where machine unlearning is required. There are several attempts to handle such cases, with methods ranging from using influence functions to unlearn features [46], to disentanglement functions [20]. Similarly, methods that focus on class unlearning also exist, as seen in [3, 41].
- **Task Removal:** With the rise of continual learning, task removal is also becoming a common field of study in machine unlearning. Task removal is challenging because of how the tasks are learned sequentially, which might cause problems when unlearning. Some attempts at task unlearning have been made in [27].
- **Stream Removal:** Finally, machine learning models can be faced with streams of removal requests, such as in the case where multiple users delete their accounts (following a boycott, say). Gupta et al [21] attempt to deal with this type of removal from machine learning.

2.2. Algorithm Requirements

There are some prerequisites that have to be satisfied for any machine unlearning algorithm. This is explained in more detail in the paper by Salvatore Mercuri's paper [31]. Ideally, a machine unlearning algorithm must be:

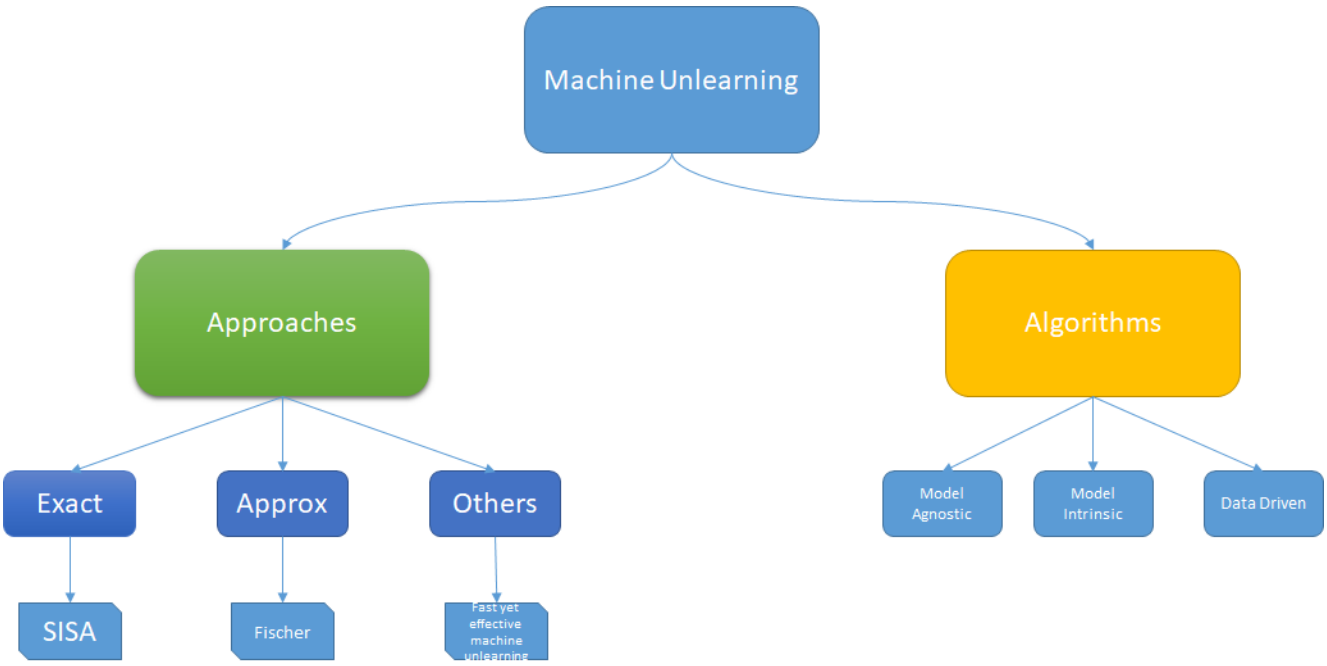

Figure 2. Approaches and Algorithms in Machine Unlearning

- Effective: Effectiveness is the comparison of the test predictions from the unlearned model to the naive retrained model. An effective unlearning model has a comparable performance to the naive model.
- Efficient: The ratio of time taken to obtain the unlearned model to the retrained model, a small turnaround time is crucial in developing a good machine unlearning algorithm.
- Consistency: This is a measure of how close the untrained model is to the naive retrained model in terms of prediction of output. There are many ways of determining consistency, with one of them being the distance between the unlearned parameters and naive retrained parameters. This method was used by Wu in their paper [47]. Consistency can also be measured by the number of predictions that are common between the naive retrained model and the untrained model output by the algorithm, as in [23]. Another measure of consistency was using KL Divergence in Golatkar's paper [15].
- Certifiability: There are various definitions for certifiability available in the literature. There are many theoretical measures of certifiability such as in Guo's paper [19]. Other empirical measures of certifiability are used in [23, 28, 40] and so on.
- Timeliness: Another critical design requirement, timeliness refers to the difference in speed between development of an unlearning model and just naive relearning of a new model. There has to be a tradeoff between completeness and timeliness of an unlearning model.
- Model Agnostic: Any unlearning alorithm developed must be agnostic to different learning techniques and models, something that is challenging to do with the wide array of available techniques in the machine learning field right now.

3. Approaches to Machine Unlearning

With the formal definition and requirements of machine unlearning listed, we can now shift our focus to classifying contemporary approaches to the field. Figure 2 gives a broad overview, followed by a breakdown by approach and algorithm.

The different approaches are listed in the subsections that follow.

3.1. Exact Unlearning

Exact unlearning is defined as the unlearned model being the same as the naive retrained model. Since it is hard to define exactness in machine learning and unlearning, where most models operate on a black box assumption, we can look at exact unlearning in two different ways [5, 14]:

- Exact Unlearning by weight distribution: The weight distribution of the unlearned model is the same as the weight distribution from the naive retrained model.
- Exact Unlearning by distribution of output: The output distribution from both models is the same.

Some exact unlearning approaches that exist are SISA and DARE algorithms[4, 5]. However, to accurately judge

how good an unlearned model is, you would need to train a model from scratch, which might be computationally expensive. As a result, there are other methods that are being explored that are more viable.

3.2. Approximate Unlearning

In approximate learning, the focus is on reducing cost-related constraints. This can either be done by:

- Perform computationally less costly actions on the final weights [18, 19, 38]
- modify the architecture and filter the outputs[3]

While these two methods are the prevailing techniques to develop unlearning techniques, a few other methods are being studied for their use cases in situations where there is no access to training data.

3.3. Zero Glance Unlearning

While traditional machine unlearning algorithms assume that access to training data is always available, this might not be the case. Some researchers like Tarun [41] prefer to work in stricter settings where the trainers of the model are not allowed to use data to even tweak or modify the model's weights. They further proposed a noise matrix for the classes that maximizes loss on the classes that have to be dropped. Then, the model is trained for 1 epoch to damage the model parameters on the forgotten classes, thus inducing unlearning.

3.4. Zero-shot Unlearning

Similarly, if the training data is not available, then the scenario becomes zero-shot unlearning. In this case, there is a strong focus on using error minimization and gated knowledge transfer to achieve this objective, as seen in [10].

3.5. Few-shot Unlearning

Finally, in few-shot learning, the scenario is more similar to zero-glance learning in that only a small portion of the data to be deleted is available. In this case, ideas like model inversion and influence approximation is used [36, 49]. These methods often have limitations, however, like the first case, which only works on models that use cross-entropy loss.

4. Algorithms in Machine Unlearning

We now move on to different algorithms that can be developed based on the requirements and approaches listed above. While there are several algorithms, the main types of algorithms that can be developed are:

- Model Agnostic: The definition of model agnostic algorithms are frameworks of unlearning algorithms that work on all models, regardless of the type or implementation of the model. Some sub-implementations in this category include differential privacy[12, 22], certified removal[15, 19, 32, 42], statistical query learning, and parameter sampling[6, 7, 9, 25, 34, 45, 51].
- Model Intrinsic: On the opposite end of the spectrum, we have model intrinsic approaches that are limited by the type of model. Although they are limited by the model type, they are not necessarily narrow, since many machine learning models can share the same type. Some subfields in model intrinsic approaches include softmax classifiers, linear and tree based models, DNN Based models, and Bayesian models[2, 16, 17, 24, 26, 30, 33, 37, 48, 50].
- Data Driven: Finally, in data-driven approaches, the focus is on speeding up retraining or untraining by using data manipulation techniques. Bourtoule's paper[4], in which he introduces the SISA algorithm talks about splitting the data into shards and slices. There is another group of researchers who work on effective retraining by augmenting the data[1, 11].

While this list is not conclusive, it lists the broad algorithmic approaches that are being developed in the unlearning space.

5. Conclusion

In conclusion, machine unlearning is becoming a must-study field for many, especially after landmark data protection mandates like GDPR[29]. This paper aimed to gather information about the latest developments in this field, from theory to methodologies that are being applied practically worldwide to satisfy various guidelines on customer data protection. While there is certainly a lot of development in this field, there are still a lot of challenges that researchers must overcome to find something that can be objectively proven to remove data while being better than naive retraining.

References

[1] Nasser Aldaghri, Hessam Mahdavifar, and Ahmad Beirami. Coded machine unlearning. *IEEE Access*, 9:88137–88150, 2021. 4

[2] Samyadeep Basu, Philip Pope, and Soheil Feizi. Influence functions in deep learning are fragile. *arXiv preprint arXiv:2006.14651*, 2020. 4

[3] Thomas Baumhauer, Pascal Schöttle, and Matthias Zeppelzauer. Machine unlearning: Linear filtration for logit-based classifiers. *Machine Learning*, 111(9):3203–3226, 2022. 2, 4

[4] Lucas Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 141–159, 2021. 2, 3, 4

[5] Jonathan Brophy and Daniel Lowd. Machine unlearning for random forests. In *International Conference on Machine Learning*, pages 1092–1104. PMLR, 2021. 3

[6] Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *2015 IEEE symposium on security and privacy*, pages 463–480. IEEE, 2015. 4

[7] Yuantao Chen, Jie Xiong, Weihong Xu, and Jingwen Zuo. A novel online incremental and decremental learning algorithm based on variable support vector machine. *Cluster Computing*, 22:7435–7445, 2019. 4

[8] Rishav Chourasia and Neil Shah. Forget unlearning: Towards true data-deletion in machine learning. In *International Conference on Machine Learning*, pages 6028–6073. PMLR, 2023. 2

[9] Vikram S Chundawat, Ayush K Tarun, Murari Mandal, and Mohan Kankanhalli. Can bad teaching induce forgetting? unlearning in deep networks using an incompetent teacher. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7210–7217, 2023. 4

[10] Vikram S Chundawat, Ayush K Tarun, Murari Mandal, and Mohan Kankanhalli. Zero-shot machine unlearning. *IEEE Transactions on Information Forensics and Security*, 18: 2345–2354, 2023. 4

[11] Yonatan Dukler, Benjamin Bowman, Alessandro Achille, Aditya Golatkar, Ashwin Swaminathan, and Stefano Soatto. Safe: Machine unlearning with shard graphs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17108–17118, 2023. 4

[12] 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 *International Conference on Artificial Intelligence and Statistics*, pages 3457–3465. PMLR, 2024. 4

[13] Sanjam Garg, Shafi Goldwasser, and Prashant Nalini Vasudevan. Formalizing data deletion in the context of the right to be forgotten. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pages 373–402. Springer, 2020. 2

[14] Antonio Ginart, Melody Guan, Gregory Valiant, and James Y Zou. Making ai forget you: Data deletion in machine learning. *Advances in neural information processing systems*, 32, 2019. 2, 3

[15] Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9304–9312, 2020. 3, 4

[16] Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Forgetting outside the box: Scrubbing deep networks of information accessible from input-output observations. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX 16*, pages 383–398. Springer, 2020. 4

[17] Adit Goyal, Vikas Hassija, and Victor Hugo C de Albuquerque. Revisiting machine learning training process for enhanced data privacy. In *Proceedings of the 2021 Thirteenth International Conference on Contemporary Computing*, pages 247–251, 2021. 4

[18] Laura Graves, Vineel Nagisetty, and Vijay Ganesh. Amnesiac machine learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11516–11524, 2021. 4

[19] Chuan Guo, Tom Goldstein, Awni Hannun, and Laurens Van Der Maaten. Certified data removal from machine learning models. *arXiv preprint arXiv:1911.03030*, 2019. 3, 4

[20] Tao Guo, Song Guo, Jiewei Zhang, Wenchao Xu, and Junxiao Wang. Efficient attribute unlearning: Towards selective removal of input attributes from feature representations. *arXiv preprint arXiv:2202.13295*, 2022. 2

[21] Varun Gupta, Christopher Jung, Seth Neel, Aaron Roth, Saeed Sharifi-Malvajerdi, and Chris Waites. Adaptive machine unlearning. In *Advances in Neural Information Processing Systems*, pages 16319–16330. Curran Associates, Inc., 2021. 2

[22] Varun Gupta, Christopher Jung, Seth Neel, Aaron Roth, Saeed Sharifi-Malvajerdi, and Chris Waites. Adaptive machine unlearning. *Advances in Neural Information Processing Systems*, 34:16319–16330, 2021. 4

[23] Yingzhe He, Guozhu Meng, Kai Chen, Jinwen He, and Xingbo Hu. Deepobliviate: a powerful charm for erasing data residual memory in deep neural networks. *arXiv preprint arXiv:2105.06209*, 2021. 3

[24] Zachary Izzo, Mary Anne Smart, Kamalika Chaudhuri, and James Zou. Approximate data deletion from machine learning models. In *International Conference on Artificial Intelligence and Statistics*, pages 2008–2016. PMLR, 2021. 4

[25] Hyunjune Kim, Sangyong Lee, and Simon S Woo. Layer attack unlearning: Fast and accurate machine unlearning via layer level attack and knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 21241–21248, 2024. 4

[26] Yuantong Li, Chi-Hua Wang, and Guang Cheng. Online forgetting process for linear regression models. *arXiv preprint arXiv:2012.01668*, 2020. 4

[27] Bo Liu, Qiang Liu, and Peter Stone. Continual learning and private unlearning. In *Conference on Lifelong Learning Agents*, pages 243–254. PMLR, 2022. 2

[28] Ananth Mahadevan and Michael Mathioudakis. Certifiable machine unlearning for linear models. *arXiv preprint arXiv:2106.15093*, 2021. 3

[29] Alessandro Mantelero. The eu proposal for a general data protection regulation and the roots of the 'right to be forgotten'. *Computer Law & Security Review*, 29(3):229–235, 2013. 4

[30] Ronak Mehta, Sourav Pal, Vikas Singh, and Sathya N Ravi. Deep unlearning via randomized conditionally independent hessians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10422–10431, 2022. 4

[31] Salvatore Mercuri, Raad Khraishi, Ramin Okhrati, Devesh Batra, Conor Hamill, Taha Ghasempour, and Andrew Nowlan. An introduction to machine unlearning, 2022. 2

[32] Seth Neel, Aaron Roth, and Saeed Sharifi-Malvajerdi. Descent-to-delete: Gradient-based methods for machine unlearning. In *Algorithmic Learning Theory*, pages 931–962. PMLR, 2021. 4

[33] Quoc Phong Nguyen, Bryan Kian Hsiang Low, and Patrick Jaillet. Variational bayesian unlearning. *Advances in Neural Information Processing Systems*, 33:16025–16036, 2020. 4

[34] Quoc Phong Nguyen, Ryutaro Oikawa, Dinil Mon Divakaran, Mun Choon Chan, and Bryan Kian Hsiang Low. Markov chain monte carlo-based machine unlearning: Unlearning what needs to be forgotten. In *Proceedings of the 2022 ACM on Asia Conference on Computer and Communications Security*, pages 351–363, 2022. 4

[35] Thanh Tam Nguyen, Thanh Trung Huynh, Zhao Ren, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, and Quoc Viet Hung Nguyen. A survey of machine unlearning, 2024. 2

[36] Alexandra Peste, Dan Alistarh, and Christoph H Lampert. Ssse: Efficiently erasing samples from trained machine learning models. *arXiv preprint arXiv:2107.03860*, 2021. 4

[37] Sebastian Schelter, Stefan Grafberger, and Ted Dunning. Hedgecut: Maintaining randomised trees for low-latency machine unlearning. In *Proceedings of the 2021 International Conference on Management of Data*, pages 1545–1557, 2021. 4

[38] Ayush Sekhari, Jayadev Acharya, Gautam Kamath, and Ananda Theertha Suresh. Remember what you want to forget: Algorithms for machine unlearning. *Advances in Neural Information Processing Systems*, 34:18075–18086, 2021. 2, 4

[39] Saurabh Shintre, Kevin A Roundy, and Jasjeet Dhaliwal. Making machine learning forget. In *Privacy Technologies and Policy: 7th Annual Privacy Forum, APF 2019, Rome, Italy, June 13–14, 2019, Proceedings 7*, pages 72–83. Springer, 2019. 2

[40] David Marco Sommer, Liwei Song, Sameer Wagh, and Prateek Mittal. Towards probabilistic verification of machine unlearning. *arXiv preprint arXiv:2003.04247*, 2020. 3

[41] Ayush K Tarun, Vikram S Chundawat, Murari Mandal, and Mohan Kankanhalli. Fast yet effective machine unlearning. *IEEE Transactions on Neural Networks and Learning Systems*, 2023. 2, 4

[42] Enayat Ullah, Tung Mai, Anup Rao, Ryan A Rossi, and Raman Arora. Machine unlearning via algorithmic stability. In *Conference on Learning Theory*, pages 4126–4142. PMLR, 2021. 4

[43] Michael Veale, Reuben Binns, and Lilian Edwards. Algorithms that remember: model inversion attacks and data protection law. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 376 (2133):20180083, 2018. 2

[44] Eduard Fosch Villaronga, Peter Kieseberg, and Tiffany Li. Humans forget, machines remember: Artificial intelligence and the right to be forgotten. *Computer Law Security Review*, 34(2):304–313, 2018. 2

[45] Lingzhi Wang, Tong Chen, Wei Yuan, Xingshan Zeng, Kam-Fai Wong, and Hongzhi Yin. Kga: A general machine unlearning framework based on knowledge gap alignment. *arXiv preprint arXiv:2305.06535*, 2023. 4

[46] Alexander Warnecke, Lukas Pirch, Christian Wressnegger, and Konrad Rieck. Machine unlearning of features and labels. *arXiv preprint arXiv:2108.11577*, 2021. 2

[47] Yinjun Wu, Edgar Dobriban, and Susan Davidson. Deltagrad: Rapid retraining of machine learning models. In *International Conference on Machine Learning*, pages 10355–10366. PMLR, 2020. 3

[48] Zhaomin Wu, Junhui Zhu, Qinbin Li, and Bingsheng He. Deltaboost: Gradient boosting decision trees with efficient machine unlearning. *Proceedings of the ACM on Management of Data*, 1(2):1–26, 2023. 4

[49] Youngsik Yoon, Jinhwan Nam, Hyojeong Yun, Jaeho Lee, Dongwoo Kim, and Jungseul Ok. Few-shot unlearning by model inversion. *arXiv preprint arXiv:2205.15567*, 2022. 4

[50] Peng-Fei Zhang, Guangdong Bai, Zi Huang, and Xin-Shun Xu. Machine unlearning for image retrieval: A generative scrubbing approach. In *Proceedings of the 30th ACM international conference on multimedia*, pages 237–245, 2022. 4

[51] Yongjing Zhang, Zhaobo Lu, Feng Zhang, Hao Wang, and Shaojing Li. Machine unlearning by reversing the continual learning. *Applied Sciences*, 13(16):9341, 2023. 4