Interactive Anonymization for Privacy aware Machine Learning

Bernd Malle[1,2], Peter Kieseberg[1,2], Andreas Holzinger[1]

[1] Holzinger Group HCI-KDD
Institute for Medical Informatics, Statistics & Documentation
Medical University Graz, Austria
b.malle@hci-kdd.org
[2] SBA Research gGmbH, Favoritenstrae 16, 1040 Wien
PKieseberg@sba-research.org

Abstract. Privacy aware Machine Learning is the discipline of applying Machine Learning techniques in such a way as to protect and retain personal identities during the process. This is most easily achieved by first anonymizing a dataset before releasing it for the purpose of data mining or knowledge extraction. Starting in June 2018, this will also remain the sole legally permitted way within the EU to release data without granting people involved the *right to be forgotten*, i.e. the right to have their data deleted on request. To governments, organizations and corporations, this represents a serious impediment to research operations, since any anonymization results in a certain degree of reduced data utility. In this paper we propose applying human background knowledge via interactive Machine Learning to the process of anonymization; this is done by eliciting human preferences for preserving some attribute values over others in the light of specific tasks. Our experiments show that human knowledge can yield measurably better classification results than a rigid automatic approach. However, the impact of interactive learning in the field of anonymization will largely depend on the experimental setup, such as an appropriate choice of application domain as well as suitable test subjects.

Keywords: Machine Learning, Privacy aware ML, interactive ML, Knowledge Bases, Anonymization, k-Anonymity, SaNGreeA, Information Loss, Weight Vectors

1 Introduction and Motivation

In many sectors of today's data-driven economies technical progress is dependent on data mining, knowledge extraction from diverse sources, as well as the analysis of personal information. Especially the latter constitutes a vital building-block for business intelligence and the provision of personalized services, which are practically demanded by modern society. Often, the insights necessary for enabling organizations to provide these goods require publication, linkage, and systematic analysis of personal data sets from heterogeneous sources, exposing

those data to the risk of leakage, with repercussions ranging from mild inconvenience (exposure of a social profile) to potentially catastrophic ramifications (leakage of health information to an employer).

Living up to those challenges, governments around the world are contemplating or already enacting new laws concerning the handling of personal data. For instance, under the new European General Data Protection Regulations (*GDPR*) taking effect on June 1st, 2018, customers are given a *right to be forgotten*, meaning that an organization is obligated to remove a customer's personal data upon request. An exception to this rule is only granted to organizations which anonymize data before analyzing them in any wholesale, automated fashion. This brings us to the field of Privacy aware machine learning (PaML), e.g. the application of ML algorithms only on previously anonymized data. Such anonymization can be provided by perturbing data (e.g. introduction noise into numerical values or *differential privacy* [4]) or *k-anonymity* [17] (clustering of data into equivalence groups), which has since become the industry standard.

The original requirement of *k-anonymity* has since been extended by the concepts of *l-diversity* [11] (where every cluster must contain at least l diverse sensitive values), *t-closeness* [9] (demanding that the local distribution over sensitive values must not diverge from its global distribution by more than a threshold of t) as well as *delta-presence* [15] (which incorporates the background knowledge of a potential attacker). Although all of those concepts are interesting in their own right, for the sake of comparing interactive ML algorithms to their fully automatic counterpart, we only took *k-anonymity* into consideration.

Based on our previous works on this topic [13] [12], in which we conducted a comparison study of binary classification performance on perturbed (selective deletion) vs. wholesale anonymized data, in this paper we introduce the notion of interactive Machine Learning for (k-)anonymization.

2 k-Anonymity

Given the original tabular concept of anonymization, we will usually encounter three different categories of attributes within a given dataset:

- **Personal identifiers** are data items which directly identify a person without having to cross-reference or further analyze them. Examples are email address or social security number (SSN). As personal identifiers are immediately dangerous, this category of data is usually removed.
- **Sensitive data,** also called 'payload', represents information that is crucial for further data mining or research purposes. Examples for this category would be disease classification, drug intake or personal income. This data shall be preserved in the anonymized dataset and can therefore not be deleted or generalized.
- **Quasi identifiers (QI's)**, are data which in themselves do not directly reveal the identity of a person, but might be used in aggregate to reconstruct it. For instance, [18] reported in 2002 that the identity of 87% of U.S. citizens

could be uncovered via just the 3 attributes *zip code*, *gender* and *date of birth*. Despite this danger, QI's may contain vital information to research applications (like ZIP code in a disease spread study); they are therefore generalized to an acceptable compromise between privacy (data loss) and information content (data utility).

Based on this categorization *k-anonymity* [16] was introduced as a formal concept of privacy, in which a record is released only if its quasi-identifiers are indistinguishable from at least $k-1$ other entities in the dataset. This can be imagined like a clustering of data into so-called *equivalence groups* of at least size k, with all internal QI's being generalized to the exact same level.

Generalization in this setting means an abstraction of attribute value: e.g. given two ZIP codes of '8010' and '8045', we could first generalize to '80**', then incorporate another data point showing ZIP '8500' by generalizing the cluster to '8***', and finally merging with any other ZIP code to the highest level of 'all', also denoted as '*'.

3 interactive Machine Learning

Interactive ML algorithms adjust their inner workings by continuously interacting with an outside *oracle*, drawing positive / negative reinforcement from this interaction [7]. Such systems are especially useful for highly-personalized predictions or decision support [8]; moreover many real-world problems exhibit (super)exponential algorithmic runtime; in such cases human brains dwarf machines at approximating solutions and learning from very small samples, thus enabling us to 'intuit' solutions efficiently [6].

By incorporating humans as oracles into this process, we can elicit background knowledge regarding specific use cases unknown to automatic algorithms [19]. This however is highly dependent on the users' experience in a certain field as well as data / classification complexity; domain experts can of course be expected to contribute more valuable decision points than laymen; likewise, a low-dimensional dataset and simple classification tasks will result in higher quality human responses than convoluted problem sets.

While the authors of [14] propose a system that interacts with a user in order to set a certain k-factor and subsequently provides a report on information loss and Kurtosis of QI distributions, the algorithm is not *interactive* by our definition in that it does not influence the inner workings of the algorithm during the learning phase. This is also true in case of the Cornell Anonymization Toolkit (Cat) [20], which conducts a complete anonymization run and only afterwards lets the user decide if they are satisfied with the results. In contrast, our approach alters algorithmic parameters upon every (batch of) human decisions, letting the algorithm adapt in real-time.

[10] describe an approach incorporating humans into the anonymization process by allowing them to set constraints on attribute generalization; moreover they construct generalization hierarchies involving domain-specific ontologies.

Although this technique marks a departure from wholesale automatic anonymization, it still lacks the dynamic human-computer interaction of our approach.

Apart from the field of privacy, interactive ML is present in a wide spectrum of applications, from bordering medical fields like protein interactions / clusterings [1] via on-demand group-creation in social networks [2] to even teaching algorithms suitable mappings from gestures to music-generating parameters [5].

4 Experiments

The following sections will describe our experiment in detail, encompassing the general iML setting, chosen data set, anonymization algorithm used as well as a description of the overall processing pipeline employed to obtain the final results as presented.

4.1 General setting

The basic idea of our experiment was to compare different weight vectors representing attribute (quasi-identifier) importance during anonymization: Let's say that a doctor needs ro release a dataset for the purpose of studying disease-spread; in this case 'ZIP code' information is probably (but not necessarily) of much greater importance then 'occupation' or 'race'. However, if a skin cancer study is to be performed, 'race' information might be of utmost importance, whereas 'ZIP code' might be negligible.

In our experiment, the task was to classify a people dataset on the target attributes *income*, *education level* and *marital status*. Therefore, we tested an *equal* weight vector setting against two others obtained from human experiments: 1) *bias* in which the user just specified which attributes they thought would be important for a specific classification by moving sliders, and 2) *iML* in which the user was tasked to decide a series of clustering possibilities by moving a data row to one of two partly anonymized clusters presented, thereby conveying which attributes were more important to preserve than others (Figure 1). Only the last method constitutes an interactive learning approach by introducing an oracle into the process.

4.2 Data

We chose the adults dataset from the UCI Machine Learning repository which was generated from US census data from 1994 and contains approximately 50k entries in it's original; this data-set is used by many anonymization researchers and therefore constitutes a quasi-standard. After initial preprocessing we chose the first 500 complete data rows as our iML experimental data to be presented to users. After obtaining bias / iml weights from the experiment, we chose the first 3k entries of the original data as the basis for producing 775 new, anonymized data sets. Although 3k rows might seem overly frugal on our part, we have asserted via random deletion of original data points that classifier performance remains stable for as little as 1.5k rows. Of the original attributes (data columns)

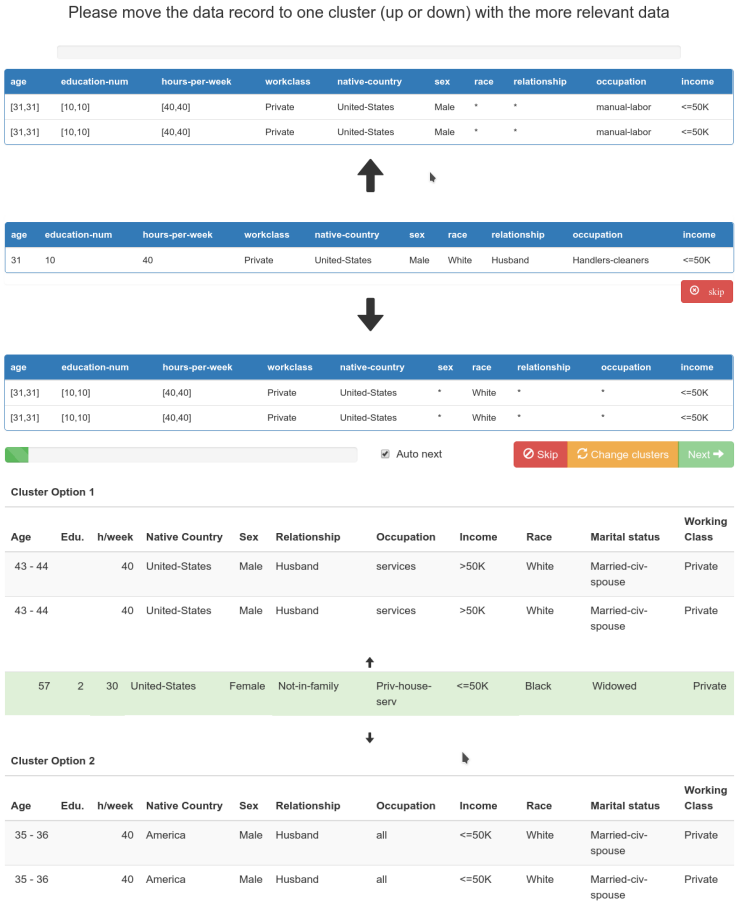

Fig. 1. Two different implementations of the iML interface design.

provided 4 were deleted: 'capital-gain' & 'capital-loss' (both were too skewed to be useful for humans), 'fnlwgt' (a mere weighting factor) as well as 'education' which is also represented by 'education_num'.

4.3 Anonymization Algorithm

In order to conduct our experiments, it was necessary to choose an algorithm which would enable us to easily hook into its internal logic - we therefore chose a greedy clustering algorithm called *SaNGreeA* (Social network greedy clustering) which was introduced by [3] and implemented it in JavaScript. This enabled us to execute it within a browser environment during our iML experiments as well as server-side for batch-execution of all derived datasets afterwards. As a greedy clustering algorithm SaNGreeA's runtime lies in $O(n^2)$ - which we were willing to accept in exchange for it's white-box internals.

Besides its capacity to anonymize graph structures (which we did not utilize during this work), it is a relatively simple algorithm considering *General information loss* - or GIL - during anonymization. This GIL can be interpreted by the sum of information loss occurring during generalization of continuous (range) as well as hierarchical attributes:

$$\mathrm{GIL}(cl) = |cl| \cdot \Big(\sum_{j=1}^{s} \frac{size(gen(cl)[N_j])}{size(min_{x \epsilon N}(X[N_j]), max_{x \epsilon N}(X[N_j]))} + \sum_{j=1}^{t} \frac{height(\Lambda(gen(cl)[C_j]))}{height(H_{C_j})}\Big)$$

where:
- $|cl|$ denotes the cluster cl's cardinality;
- $size([i1, i2])$ is the size of the interval $[i1, i2]$, i.e., $(i2 - i1)$;
- $\Lambda(w), w \epsilon H_{C_j}$ is the sub-hierarchy of H_{C_j} rooted in w;
- $height(H_{C_j})$ denotes the height of the tree hierarchy H_{C_j};

The following formulas then give the total / normalized GIL, respectively:

$$\mathrm{GIL}(G, S) = \sum_{j=1}^{v} \mathrm{GIL}(cl_j) \quad \text{and} \quad \mathrm{NGIL}(G, S) = \frac{\mathrm{GIL}(G, S)}{n \cdot (s + t)}$$

The algorithm starts by picking a (random or pre-defined) data row as its first cluster, then iteratively picking best candidates for merging by minimizing GIL until the cluster reaches size k, at which point a new data point is chosen as the initiator for the next cluster; this process continues until all data points are merged into clusters, satisfying the k-anonymity criterion for the given dataset.

4.4 Processing pipeline for obtaining results

Once our iML experiments had yielded enough weight vectors, we had to generate a whole new set of anonymized datasets on which we subsequently applied 4 classifiers on each of the 3 target attributes (columns) described; therefore we designed the following processing pipeline:

1. Taking the first 5k rows of the original, preprocessed dataset as input and applying k-anonymization with a k-factor range of [5, 10, 20, 50, 100, 200] and 129 different weight vectors (equal, bias, iml) from our experiments on it, we produced 774 anonymized datasets (775 including the original).
2. We executed 4 classifiers on all of the datasets and compared their F1 score; the reason for selecting multiple algorithms was to explore if anonymization would yield different behaviors on different mathematical approaches for classification. The four algorithms used were *linear SVC* (as a representative of Support Vector Machines), *logistic regression* (gradient descent),

gradient boosting (ensemble, boosting) as well as *random forest* (ensemble, bagging). While reading the datasets pertaining to the classification target of *education*, the 14 different education levels present within the adult dataset were grouped into 4 categories 'pre-high-school', 'high school', '<=bachelors' and 'advanced studies'.

3. For each combination of classification target (*income*, *marital status*, *education*) and weight category (*equal*, *bias*, *iml*) we averaged the respective results. Results were plotted per target, as this allows better comparison between different classifiers. The leftmost point in all plots designates the original, un-anonymized dataset.

5 Results & Discussion

As per the results in our previous work on PaML [13] [12] we generally expected $1/x$ shaped curves for classifier performance as factors of k are increasing. These expectations held to only a small degree; moreover for targets *education* as well as *income* there was no clear winner amongst the weight categories, with some achieving better or worse depending on a specific factor of k.

We got the smoothest results for the *marital status* target, with human bias winning consistently over equal weights as well as human interaction (Figure 2). We interpret this as stemming from the fact that there is a significant correlation between the attributes 'marital-status' and 'relationship' in the dataset, which led users to consciously overvalue the latter when prompted for their bias. It is not completely clear why the iML results were not able to keep up in this case, but since this seems to be a general phenomenon throughout our results, we will discuss this in a later paragraph.

On classification target *education*, bias still mostly outperforms iML-obtained attribute weights, with equal weights slightly winning out at very high factors of k (Figure 3). Although we assume that apparently important clues towards education might be misleading (like income or working hours), this cannot explain the difference between bias- and iML-based results. It has to be noted however, that results on this target are distinctly inferior to those of the other scenarios which might diminish the gap's significance.

Only on target *income* did we observe a partly reversed order between human bias and iML - however at the cost of both being usually inferior to a simple setting with equal attribute weights (Figure 4). This is especially surprising because *income* was the only binary classification task in our experiments, which should have given humans a slight advantage over the algorithm. On the other hand, human bias seems most susceptible to falling prey to certain stereotypes in the area of money (w.r.t. gender, race, marital status...), which would explain the reversal of results.

As for the failure of iML to significantly outperform both the equal weight setting and especially human bias, we conjecture that our experimental setup has produced those effects: Since we wanted our users to conduct their experiment in real-time but needed a simple implementation of an anonymization algorithm to

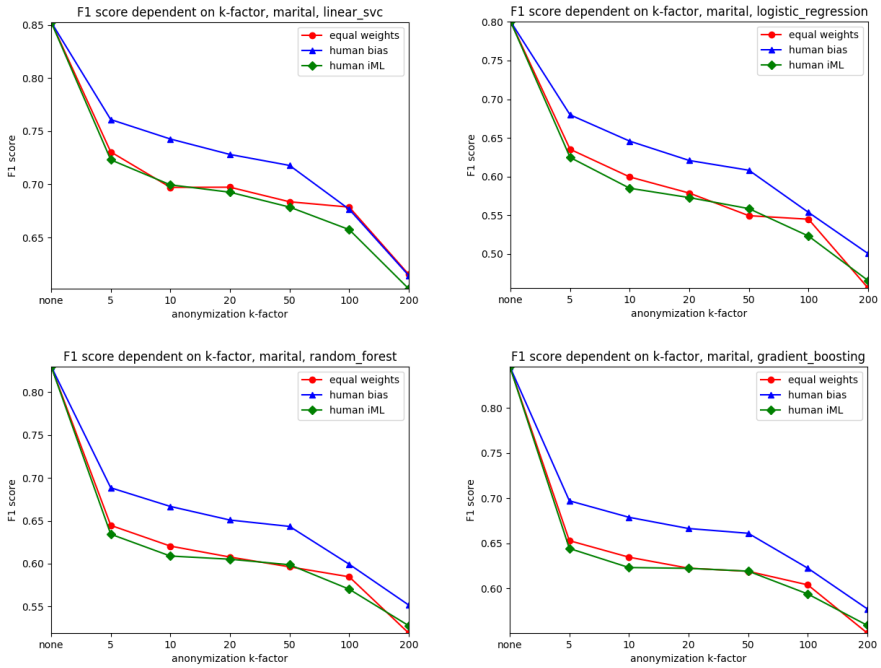

Fig. 2. Results on target *marital status* - human bias wins consistently over both equal weights and human interaction with the algorithm.

enable this interaction (which resulted in an $O(n^2)$ algorithmic runtime), we had to limit ourselves to just a tiny subset of data (500 rows, merely 1% of the original dataset). This choice apparently resulted in generalizations proceeding far too quickly, reaching suppression ('all') levels prematurely, thereby denying our users sensible clustering choices. On the other hand, the effect could also stem from users not really trying to contribute to the experiments in a meaningful way; this effect could only be mitigated by selecting more serious users or choosing some less serious (more social?) application domain.

Overall, we were also surprised that a seemingly absurd k-factor of 200 would still yield comparably good results (and in some cases even improve performance..).

6 Open problems & future challenges

As iML for anonymization is still a fledgling sub-area in the larger fields of privacy as well as Machine Learning, there are certainly innumerable possibilities for even basic progress & development. The following list is only a tiny subset of possible research venues we deem suitable for our own future work:

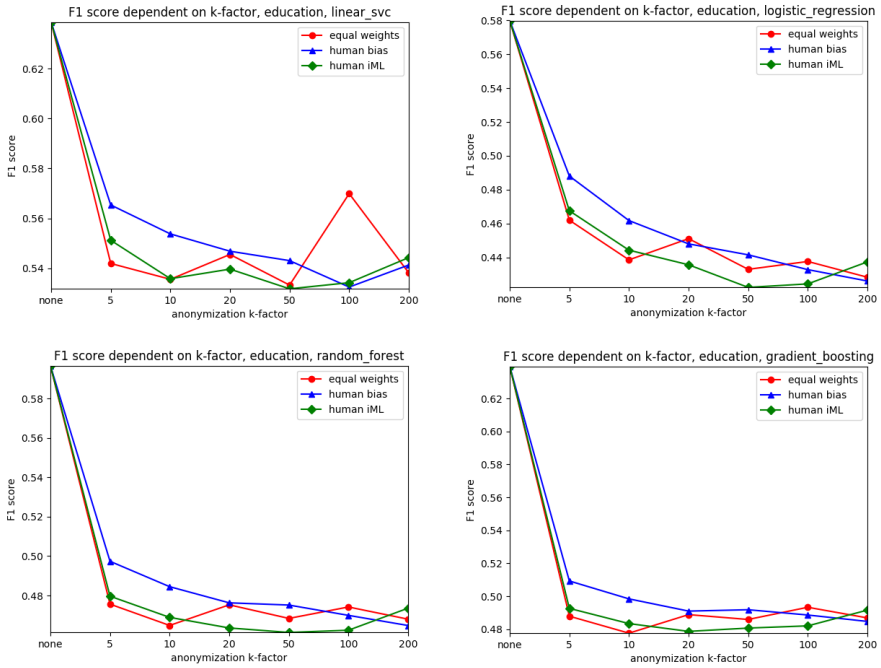

Fig. 3. Results on target *education* - we still see human bias performing slightly better than equal weights / iML in most cases of k, but not as consequently as above.

- **Explain the unexpected behavior** of linear SVC on the *income* target at high levels of k; probably by performing comparison studies on synthetically generated datasets.
- **Faster algorithm.** Repeat the experiments with a faster algorithmic implementation so that we can use thousands of data points even in real time within a Browser: this would lead to more relaxed generalizations, allowing the user to make better interactive choices, thus presumably improving results by quite some margin.
- **Expert domain, domain experts.** Choosing an expert domain like cancer studies in combination with proper experts like medical professionals, we would expect both human bias as well as iML results to significantly outperform a pre-defined weight vector.
- **Different setting.** On the other hand, a more 'gamified' setting such as recommendations within a social network could motivate amateur users to get more immersed into the experiment, yielding better results even for mundane application tasks.
- **Different data formats.** As Artificial Intelligence is slowly reaching maturity, it is now also applied to non- and semi-structured data like audio/video

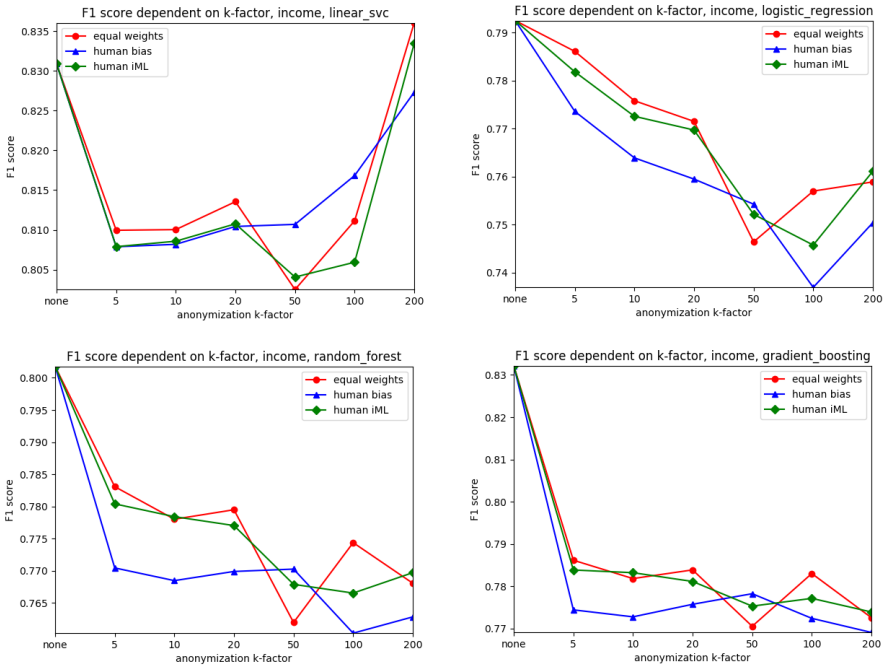

Fig. 4. Results on target *income* - only in this scenario do we see iML-based results generally outperforming bias (except linear SVC), nevertheless incapable of outperforming the rigidly equal setting.

or even *omics data. Since images are clearly relevant for medical research, and humans extremely efficient at processing them, studying interactive ML on visual data promises great scientific revenue.

7 Conclusion

Based on the emerging necessity of Privacy aware data processing, in this work we presented a fundamental approach of bringing human knowledge to bear on the task of anonymization via interactive Machine Learning. We devised an experiment involving clustering of data points with respect to human preference for attribute preservation and tested the resulting parameters on classification of anonymized people data into classes of *marital status*, *education* and *income*. Our preliminary results show that human bias can definitely contribute to even mundane application areas, whereas more complex or convoluted tasks may require trained professionals or better data preparation (dimensionality reduction etc.). We also described our insights regarding technical details for iML experiments and closed by outlining promising future research venues.

References

1. Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. Power to the People: The Role of Humans in Interactive Machine Learning. *AI Magazine*, 35(4):105–120, 2014.
2. Saleema Amershi, James Fogarty, and Daniel Weld. ReGroup: interactive machine learning for on-demand group creation in social networks. *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems - CHI '12*, page 21, 2012.
3. Alina Campan and Traian Marius Truta. Data and structural k-anonymity in social networks. In *Privacy, Security, and Trust in KDD*, pages 33–54. Springer, 2009.
4. Cynthia Dwork. Differential privacy: A survey of results. In *International Conference on Theory and Applications of Models of Computation*, pages 1–19. Springer, 2008.
5. R. Fiebrink, D. Trueman, and P.R. Cook. A metainstrument for interactive, on-the-fly machine learning. *Proc. NIME*, 2:3, 2009.
6. A Holzinger, M Plass, K Holzinger, GC Crisan, CM Pintea, and V Palade. Towards interactive machine learning (iml): Applying ant colony algorithms to solve the traveling salesman problem with the human-in-the-loop approach. In *IFIP International Cross Domain Conference and Workshop (CD-ARES)*, pages 81–95. Springer, Heidelberg, Berlin, New York, 2016.
7. Andreas Holzinger. Interactive machine learning for health informatics: When do we need the human-in-the-loop? *Springer Brain Informatics (BRIN)*, 3(2):119–131, 2016.
8. Peter Kieseberg, Bernd Malle, Peter Frhwirt, Edgar Weippl, and Andreas Holzinger. A tamper-proof audit and control system for the doctor in the loop. *Brain Informatics*, pages 1–11, 2016.
9. Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *IEEE 23rd International Conference on Data Engineering, ICDE 2007*, pages 106–115. IEEE, 2007.
10. Brian C.S. Loh and Patrick H.H. Then. Ontology-enhanced interactive anonymization in domain-driven data mining outsourcing. *Proceedings - 2nd International Symposium on Data, Privacy, and E-Commerce, ISDPE 2010*, (June):9–14, 2010.
11. Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):1–52, 2007.
12. Bernd Malle, Peter Kieseberg, and Andreas Holzinger. Do not disturb? classifier behavior on perturbed datasets. In *Machine Learning and Knowledge Extraction, IFIP CD-MAKE, Lecture Notes in Computer Science LNCS Volume 10410*, pages 155–173. Springer, Cham, 2017.
13. Bernd Malle, Peter Kieseberg, Edgar Weippl, and Andreas Holzinger. The right to be forgotten: towards machine learning on perturbed knowledge bases. In *International Conference on Availability, Reliability, and Security*, pages 251–266. Springer, 2016.
14. Carlos Moque, Alexandra Pomares, and Rafael Gonzalez. AnonymousData.co: A Proposal for Interactive Anonymization of Electronic Medical Records. *Procedia Technology*, 5:743–752, 2012.
15. M. E. Nergiz and C. Clifton. delta-presence without complete world knowledge. *IEEE Transactions on Knowledge and Data Engineering*, 22(6):868–883, 2010.

16. Pierangela Samarati. Protecting respondents identities in microdata release. *IEEE Transactions on Knowledge and Data Engineering*, 13(6):1010–1027, 2001.
17. Latanya Sweeney. Achieving k-anonymity privacy protection using generalization and suppression. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):571–588, 2002.
18. Latanya Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.
19. MALCOLM WARE, EIBE FRANK, GEOFFREY HOLMES, MARK HALL, and IAN H WITTEN. Interactive machine learning: letting users build classifiers. *International Journal of Human-Computer Studies*, 55(3):281–292, 2001.
20. Xiaokui Xiao, Guozhang Wang, and Johannes Gehrke. Interactive anonymization of sensitive data. *Proceedings of the 35th SIGMOD international conference on Management of data - SIGMOD '09*, page 1051, 2009.