ENCODING EXPERT KNOWLEDGE INTO FEDERATED LEARNING USING WEAK SUPERVISION

Sebastian Caldas, Mononito Goswami, Artur Dubrawski
Auton Lab
School of Computer Science
Carnegie Mellon University
scaldas@cmu.edu, {mgoswami,awd}@cs.cmu.edu

ABSTRACT

Learning from on-device data has enabled intelligent mobile applications ranging from smart keyboards to apps that predict abnormal heartbeats. However, due to the sensitive and distributed nature of this data, expert annotations are often unavailable. Consequently, existing federated learning techniques that learn from on-device data mostly rely on unsupervised approaches, and are unable to capture expert knowledge via data annotations. In this work, we explore an alternative way to codify this expert knowledge: using programmatic weak supervision, a principled framework that leverages *labeling functions* (i.e., heuristic rules) in order to label vast quantities of data without direct access to the data itself. We introduce Weak Supervision Heuristics for Federated Learning (WSHFL[1]), a method that interactively mines and leverages labeling functions to annotate on-device data in cross-device federated settings. Our experiments on two sentiment classification tasks demonstrate that WSHFL achieves competitive performance compared to a fully supervised baseline while reducing the need for direct data annotations.

1 INTRODUCTION

The potential for increasingly intelligent mobile applications through learning from on-device data has been widely acknowledged (McMahan et al., 2017). This includes applications such as smart keyboards that boost usability (Hard et al., 2018) and health apps that improve patient outcomes (Fitzpatrick et al., 2017; Bui & Liu, 2021). However, since on-device data is sensitive, it is often not available for external expert annotation (Wang et al., 2021). Thus, previous efforts to train distributed models using federated learning techniques on this type of data have mostly relied on unsupervised methods (Hard et al., 2018; Lu et al., 2021) and user contextual signals (Yang et al., 2018) for supervision.

One key challenge for federated learning in this cross-device setting is to encode expert knowledge into its models. For instance, suppose we wish to train an arrhythmia detection model using electrocardiogram (ECG) data generated through a sensor carried by the user. Federated techniques are in order, as this setting requires us to respect the sensitive nature of the data. To solve this particular task, we also wish to capture clinicians' expertise via annotations of the ECG data. However, this would imply giving clinicians access to the devices' data. How to capture expert knowledge in this federated setting is thus an active area of research (Jeong et al., 2020; Liu et al., 2021; Zhuang et al., 2021; Wu et al., 2021).

In this paper, we explore a particular strategy for codifying expert knowledge into a federated model using Labeling Functions (LFs): functions that assign imperfect labels to subsets of the data and that can be used to automatically label training data (Ratner et al., 2017; Rühling Cachay et al., 2021). Encoding supervision through LFs is referred to as Programmatic Weak Supervision (PWS) (Ratner et al., 2016; Zhang et al., 2022), and it has had success in centralized settings (Fries et al., 2019; Dunnmon et al., 2020; Goswami et al., 2021; Dey et al., 2022). To the best of our knowledge, PWS has not been explored in federated scenarios, where the focus has been on unsupervised and

[1]pronounced as in wishful.

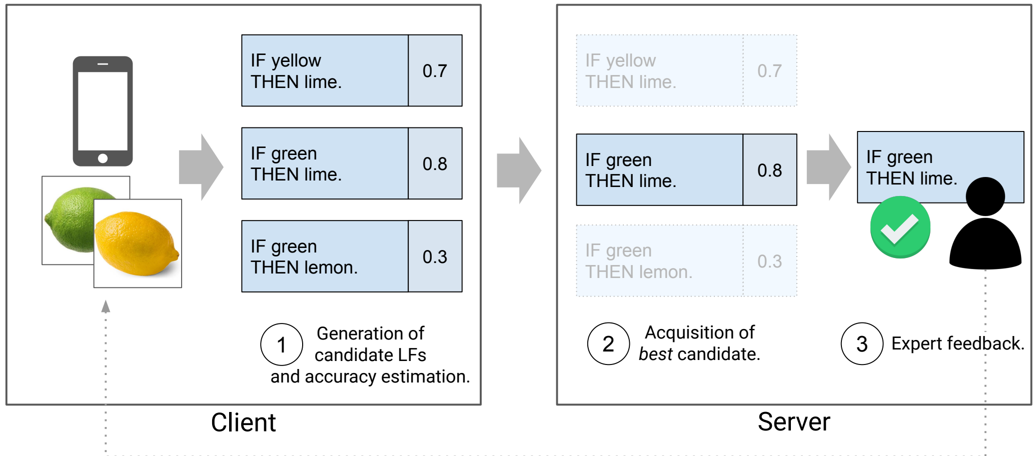

Figure 1: Visualization of WSHFL's strategy for generating LFs. Based on the on-device data, (1) candidate LFs are generated. These candidates, alongside an estimate for their accuracy, are then sent over to the server, where (2) one candidate is selected to be inspected by an expert. This (3) expert feedback is then used to estimate candidates' accuracies and iteratively pick accurate LFs for subsequent expert feedback.

semi-supervised approaches (Hard et al., 2018; Lu et al., 2021; Jeong et al., 2020; Liu et al., 2021; Zhuang et al., 2021; Wu et al., 2021). Thus, we introduce Weak Supervision Heuristics for Federated Learning (WSHFL), a method for mining and leveraging LFs in a cross-device federated setting. WSHFL proceeds in two stages:

1. **Mining of LFs (or heuristics)**: WSHFL automates the crafting of LFs (Varma & Ré, 2018; Boecking et al., 2020), incorporating expert feedback on which ones they consider useful. Only parameterized LFs are exchanged while the data is kept isolated on-device. Figure 1 presents a brief overview of this strategy.
2. **Training of the PWS model**: Once we have a set of LFs, we can use them to train a PWS model. WSHFL leverages the architecture proposed by Rühling Cachay et al. (2021), training it in a federated manner (McMahan et al., 2017).

The primary contribution of this work is the introduction of PWS to the federated setting. We argue that the main challenge of adopting PWS into cross-device federated methods is the crafting of LFs. In practice, crafting these functions is a data-dependent process, as experts rely on available data in order to extract dataset-specific heuristics (Zhang et al., 2022). In federated learning, however, experts cannot freely explore the on-device data. To tackle this obstacle, we extend the work of Boecking et al. (2020) in automatic generation of LFs to the particulars of cross-device federated learning. Preliminary results show that our approach is successful on two sentiment analysis tasks when we use unigram-based LFs.

2 Related Work

Programmatic Weak Supervision. Programmatic weak supervision (PWS) has been proposed as an alternative framework to the expensive and time-consuming process of point-by-point labeling used for supervised machine learning. PWS leverages multiple sources of noisy supervision, expressed as LFs, to label large quantities of data (Zhang et al., 2022). LFs (see an example in Fig. 2) are imperfect and may generate conflicting labels on certain data points. Thus, a *label model* (Ratner et al., 2016; Rühling Cachay et al., 2021) is used to aggregate the noisy votes of labeling functions into training labels. These labels are then used to train an *end model*, which learns to generalize the relationship between features and the learned training labels. Recent studies have also explored end-to-end approaches that couple the label and end models, leading to state-of-the-art performance (Rühling Cachay et al., 2021). To the best of our knowledge, the PWS literature has only focused on centralized settings.

Automatic Mining of LFs. Hand-crafting LFs requires expertise and data exploration, which can be resource-intensive (Boecking et al., 2020; Varma & Ré, 2018). To address this, previous methods have aimed to automate the creation of LFs given some extra supervision such as seed LFs (Li et al., 2021), labeled data (Varma & Ré, 2018; Awasthi et al., 2020), class descriptors (Gao et al., 2022), or instance-wise expert feedback (Nashaat et al., 2020). Interactive Weak Supervision (IWS) by Boecking et al. (2020) is a related algorithm that learns useful heuristics from user feedback at the LF level. WSHFL leverages this particular type of expert supervision while tackling challenges inherent to the federated scenario.

```python
def nice_lf(review):
    if "nice" in review:
        return POSITIVE
    else:
        return ABSTAIN
```

Figure 2: An example labeling function. If the unigram "nice" appears in a review, then it votes for the positive class, otherwise it abstains from voting.

Semi-supervised methods in Federated Learning. We can also codify expert knowledge into federated models by using semi-supervised learning. Recent works that have studied this alternative rely on a centralized dataset available for annotation by the experts, and augment the federated learning procedure with techniques such as consistency regularization (Jeong et al., 2020; Liu et al., 2021) and contrastive learning (Zhuang et al., 2021; Wu et al., 2021). These techniques come with their own set of assumptions, such as the existence of a centralized annotated dataset that is from the same (or a similar) distribution as the on-device data.

3 Weak Supervision Heuristics for Federated Learning

We want to train an end model f on data distributed across clients, where each client k has access to an unlabeled dataset $X_k = \{x_i^k\}_{i=1}^{n_k}$, where n_k is the number of data points in client k. WSHFL uses X_k to generate a set of p_k candidate LFs $\mathcal{L}_k = \{\lambda_j^k\}_{j=1}^{p_k}$, which are then used to build $\mathcal{L} = \bigcup_k \mathcal{L}_k$. Like IWS (Boecking et al., 2020), WSHFL then sequentially queries an expert to inspect candidate LFs from $\mathcal{L}$, building an optimal $\mathcal{L}^* \subset \mathcal{L}$, which is then used to train our PWS model on the clients' data.

Algorithm 1: WSHFL mining of labeling functions

Input: Number of iterations T, seeds S, number of rounds R, number of clients per round K, confidence $1-\delta$.
$Q_0 \leftarrow S$, $\mathcal{L}_0 \leftarrow \emptyset$.
for $t = 1, \ldots, T$ **do**
  $\lambda_t \leftarrow$ `FederatedAcquisition`(Q, R, K, δ)
  $u_t \leftarrow$ `ExpertQuery`(λ_t)
  $Q_t \leftarrow Q_{t-1} \cup (\lambda_t, u_t)$
Output: $\{\lambda_j \in Q_T : u_j = 1\}$
Function `FederatedAcquisition`(Q, R, K, δ)
  $\mathcal{L}_0 \leftarrow \emptyset$
  for $r = 1, \ldots, R$ **do**
    Select K clients at random.
    retrieve from each client
      $\mathcal{L}_k, \mathcal{U}_k \leftarrow$ `TrainClient`(Q)
    $\mathcal{L}_r = \mathcal{L}_{r-1} \cup \bigcup_{k=1}^{K} \mathcal{L}_k$
  Count how many clients m_j have proposed each λ_j.
  Compute $\epsilon_j = \sqrt{\log(2/\delta)/2m_j}$ for all j.
  Return λ_j with the highest $\mu(\hat{u}_j) - \epsilon_j$
Function `TrainClient`(Q)
  Generate candidate LFs $\mathcal{L}_k$.
  Initialize neural network h_k.
  Train h_k by predicting u_j given $\tau_k(\lambda_j)$ for pairs in Q.
  Use h_k to predict $\hat{u}_j^k$ for $\lambda_j^k \in \mathcal{L}_k$.
  Return $\mathcal{L}_k$, $\{\hat{u}_j^k\}_{j=1}^{p_k}$

Mining of Labeling Functions

The main objectives of this step are (1) to construct $\mathcal{L}$ in a federated scenario, and (2) to select the next LFs to be shown to the expert. Algorithm 1 describes our general procedure. We proceed to highlight its main points.

To construct $\mathcal{L}$ in a federated setting, we leverage a domain specific process that can produce $\mathcal{L}_k$ in each individual client. We assume that a parameterization of these $\mathcal{L}_k$ can be shared with the server and (after inspection by the expert) other clients. As an example, in our experiments, our candidate LFs are parameterized by a particular unigram and label, assigning the label if the unigram is present in data point x. Thus, a client can automatically generate

$\mathcal{L}_k$ from the cross product of unigrams in its vocabulary within a document frequency range[2], and the set of possible labels. These parameters would then be sent to the server to obtain $\mathcal{L} = \bigcup_k \mathcal{L}_k$.

To select the next $\lambda_j \in \mathcal{L}$ to show to the expert, we rely on estimates $\hat{\alpha}_j$ of the LFs accuracy $\alpha_j = P(\lambda_j(x) = y | \lambda_j(x) \neq 0)$[3]. To obtain these estimates, when the expert inspects a given candidate λ_j, they assign it a label $u_j \in \{0, 1\}$, corresponding to whether they believe α_j is better than random. WSHFL can then train a model h_k in each client that predicts u_j given the client-specific representation $\tau_k(\lambda_j) = (\lambda_j(x_1^k), \ldots, \lambda_j(x_{n_k}^k))$. This model is then used to obtain $\hat{u}_j^k$ for the λ_j^k that the expert has not inspected. We show the expert the λ_j with the highest $\hat{\alpha}_j = \frac{1}{m_j} \sum_k \hat{u}_j^k - \epsilon_j$, where m_j is the number of clients that observed λ_j and $\epsilon_j = \sqrt{\log(2/\delta)/2m_j}$ is obtained from a Hoeffding confidence interval (Maron & Moore, 1993). We use this lower confidence bound to account for the high variance of the simple mean. Finally, we build an estimate of $\mathcal{L}^*$ with those LFs that have been marked as useful.

Training of the PWS Model

Once we have an estimate of $\mathcal{L}^*$, we can use these LFs to train label model g and the resulting end model f on the clients' data. In this work, we leverage the architecture proposed by Rühling Cachay et al. (2021), who cast g as a neural encoder and jointly train it with f. We train both models in a federated manner using Federated Averaging (McMahan et al., 2017).

4 Experimental Setup

Datasets. We demonstrate the potential of our method on two binary sentiment analysis tasks, using two datasets: the Amazon product reviews dataset (Ni et al., 2019) and IMDB movie reviews dataset (Maas et al., 2011). On the Amazon dataset, we treat each unique reviewer as a different client, whereas on the IMDb dataset, we split reviews uniformly at random between clients. We pre-processed both public datasets to make them comparable to the datasets used by Boecking et al. (2020), and adapted them to a federated context. In Table 1, we provide further details about our datasets.

	Amazon dataset			**IMDb dataset**		
	Train	**Validation**	**Test**	**Train**	**Validation**	**Test**
Number of reviews	119,725	20,090	60,366	20,000	5,000	25,000
Number of clients	738	123	369	1000	-	-
Mean reviews per client (std)	162.22 (73.36)	-	-	20.0 (0.0)	-	-
Fraction of positive reviews	0.54	0.54	0.55	0.50	0.49	0.50

Table 1: *Details for Amazon and IMDb datasets used in our WSHFL experiments.*

Experimental Settings. We featurize our the data using a pre-trained sentence transformer (Reimers & Gurevych, 2019). As described above, we use LFs which output a given label if a unigram appears in a document. Previous studies applying programmatic weak supervision to text data have found n-grams to be excellent sources of supervision (Gao et al., 2022; Boecking et al., 2020). This is particularly true for unigrams due to their larger support in data. Hence, in our preliminary experiments we chose to parameterize our labeling functions using unigrams.

Client models h, label model g and end model f are all two-layer perceptrons. As an expert, we use an oracle that labels a LF as useful if it has an accuracy in the training data of at least 0.7. We optimize our parameters on our validation dataset, and report the area under the receiver operating characteristics curve (AUC ROC) on the test dataset.

[2]The document frequency of a unigram is defined as the fraction of documents which contain at least one occurrence of the unigram. Unigrams with high document frequency are likely to have low discriminative power, whereas unigrams with very low document frequency have limited utility due to low support in data.

[3]We define $\lambda_j(x) = 0$ as an abstention. α_j is defined for examples on which λ_j does not abstain.

Baselines. We compare the predictive performance of WSHFL against three baselines:

1. Random: Showing random LFs to the expert. A method that's effectively using the expert feedback should beat this baseline.
2. Naive Greedy: Showing the expert the LF with the highest $\frac{1}{m_j}\sum_k \hat{u}_j^k$. Simple mean estimates are natural in federated settings, but they may have high variance in this particular scenario.
3. Fully supervised: We also present a result of using all ground truth labels, trained with FedAvg (McMahan et al., 2017).

5 Results and Discussion

We show our results in Figure 3. We observe how WSHFL outperforms the other strategies as we show the oracle more LFs. We also observe how the performance of Naive Greedy deteriorates to that of Random. This happens because, with this strategy, the best LF usually has only one reported $\hat{u}_j^k$ from one optimistic client. WSHFL corrects for this by targeting LFs with high means and low variance, which usually implies having several $\hat{u}_j^k$ reported.

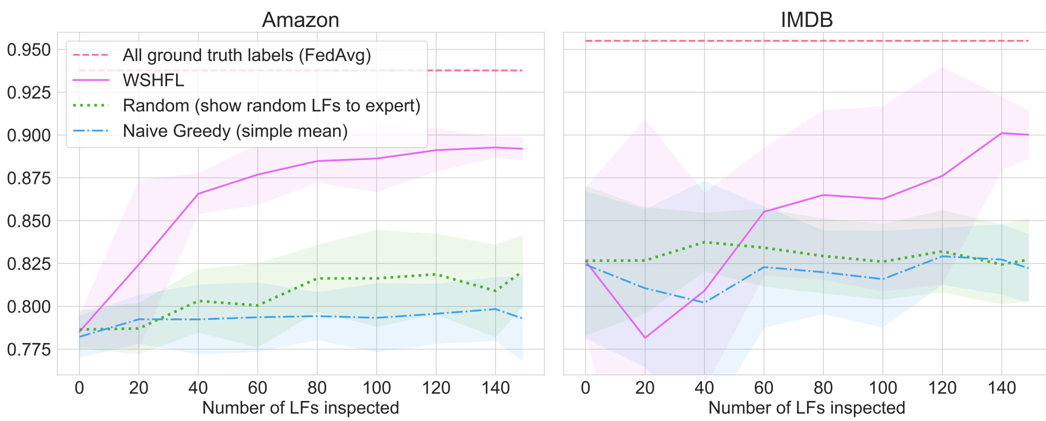

Figure 3: Results for WSHFL on the Amazon Reviews and IMDB datasets. We present the test ROC AUC of the end model vs. the number of LFs inspected by the expert. We repeat each experiment 10 times with different random seeds and show the mean (line) and standard deviation (shaded).

5.1 Future Work and Limitations

Extension to further data modalities. To further establish the utility of the proposed approach, we intend to test its performance on modalities beyond text, such as time-series and tabular data. The main challenge with these modalities is the definition of human-interpretable LFs that can be mined from the clients. Furthermore, if the parameterization of these LFs is not discrete, the algorithm in its current form has to be adapted. Alternatives include defining a way of merging similar elements in $\mathcal{L}_k$, or warm-starting and pre-training a global $\mathcal{L}$.

Exploration of non-myopic strategies. WSHFL uses a greedy strategy when selecting the next candidate LF to show to the expert. We do not consider the potential contribution that a LF could have on the functions that could be selected in the future. As such, future work may want to explore non-myopic algorithms (Jiang et al., 2017) to to better balance exploitation and exploration of the candidate LFs.

Exchange of labeling functions. WSHFL exchanges LFs at different stages of the algorithm, which means that some information may leak from the clients through these LFs. For example, in our experiments, unigrams extracted from client data are exchanged, first with experts and then with other clients. Future work may wish to study both the privacy risks of sharing a particular type of LF, and the mechanisms to mitigate these risks.

REFERENCES

Abhijeet Awasthi, Sabyasachi Ghosh, Rasna Goyal, and Sunita Sarawagi. Learning from rules generalizing labeled exemplars. *arXiv preprint arXiv:2004.06025*, 2020.

Benedikt Boecking, Willie Neiswanger, Eric Xing, and Artur Dubrawski. Interactive weak supervision: Learning useful heuristics for data labeling. In *International Conference on Learning Representations*, 2020.

Peggy Bui and Yuan Liu. Using ai to help find answers to common skin conditions. *online*, 2021. URL https://blog.google/technology/health/ai-dermatology-preview-io-2021/.

Arnab Dey, Mononito Goswami, Joo Heung Yoon, Gilles Clermont, Michael Pinsky, Marilyn Hravnak, and Artur Dubrawski. Weakly supervised classification of vital sign alerts as real or artifact. In *AMIA Annual Symposium Proceedings*, volume 2022. American Medical Informatics Association, 2022.

Jared A Dunnmon, Alexander J Ratner, Khaled Saab, Nishith Khandwala, Matthew Markert, Hersh Sagreiya, Roger Goldman, Christopher Lee-Messer, Matthew P Lungren, Daniel L Rubin, et al. Cross-modal data programming enables rapid medical machine learning. *Patterns*, 1(2):100019, 2020.

Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): A randomized controlled trial. *JMIR Ment Health*, 4(2):e19, Jun 2017. ISSN 2368-7959. doi: 10.2196/mental.7785. URL http://www.ncbi.nlm.nih.gov/pubmed/28588005.

Jason A Fries, Paroma Varma, Vincent S Chen, Ke Xiao, Heliodoro Tejeda, Priyanka Saha, Jared Dunnmon, Henry Chubb, Shiraz Maskatia, Madalina Fiterau, et al. Weakly supervised classification of aortic valve malformations using unlabeled cardiac mri sequences. *Nature communications*, 10(1):1–10, 2019.

Chufan Gao, Mononito Goswami, Jieshi Chen, and Artur Dubrawski. Classifying unstructured clinical notes via automatic weak supervision. In *Proceedings of the Machine Learning for Healthcare Conference*, Proceedings of Machine Learning Research. PMLR, 2022.

Mononito Goswami, Benedikt Boecking, and Artur Dubrawski. Weak supervision for affordable modeling of electrocardiogram data. In *AMIA Annual Symposium Proceedings*, volume 2021, pp. 536. American Medical Informatics Association, 2021.

Andrew Hard, Kanishka Rao, Rajiv Mathews, Swaroop Ramaswamy, Françoise Beaufays, Sean Augenstein, Hubert Eichner, Chloé Kiddon, and Daniel Ramage. Federated learning for mobile keyboard prediction. *arXiv preprint arXiv:1811.03604*, 2018.

Wonyong Jeong, Jaehong Yoon, Eunho Yang, and Sung Ju Hwang. Federated semi-supervised learning with inter-client consistency & disjoint learning. In *International Conference on Learning Representations*, 2020.

Shali Jiang, Gustavo Malkomes, Geoff Converse, Alyssa Shofner, Benjamin Moseley, and Roman Garnett. Efficient nonmyopic active search. In *International Conference on Machine Learning*, pp. 1714–1723. PMLR, 2017.

Jiacheng Li, Haibo Ding, Jingbo Shang, Julian McAuley, and Zhe Feng. Weakly supervised named entity tagging with learnable logical rules. *arXiv preprint arXiv:2107.02282*, 2021.

Quande Liu, Hongzheng Yang, Qi Dou, and Pheng-Ann Heng. Federated semi-supervised medical image classification via inter-client relation matching. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 325–335. Springer, 2021.

Nan Lu, Zhao Wang, Xiaoxiao Li, Gang Niu, Qi Dou, and Masashi Sugiyama. Federated learning from only unlabeled data with class-conditional-sharing clients. In *International Conference on Learning Representations*, 2021.

Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pp. 142–150, 2011.

Oded Maron and Andrew Moore. Hoeffding races: Accelerating model selection search for classification and function approximation. *Advances in neural information processing systems*, 6, 1993.

H Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, 2017.

Mona Nashaat, Aindrila Ghosh, James Miller, and Shaikh Quader. Asterisk: Generating large training datasets with automatic active supervision. *ACM Transactions on Data Science*, 1(2):1–25, 2020.

Jianmo Ni, Jiacheng Li, and Julian McAuley. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 188–197, 2019.

Alexander Ratner, Stephen H Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. Snorkel: Rapid training data creation with weak supervision. In *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases*, volume 11, pp. 269. NIH Public Access, 2017.

Alexander J Ratner, Christopher M De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. Data programming: Creating large training sets, quickly. *Advances in neural information processing systems*, 29, 2016.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL `https://arxiv.org/abs/1908.10084`.

Salva Rühling Cachay, Benedikt Boecking, and Artur Dubrawski. End-to-end weak supervision. *Advances in Neural Information Processing Systems*, 34, 2021.

Paroma Varma and Christopher Ré. Snuba: Automating weak supervision to label training data. In *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases*, volume 12, pp. 223. NIH Public Access, 2018.

Jianyu Wang, Zachary Charles, Zheng Xu, Gauri Joshi, H Brendan McMahan, Maruan Al-Shedivat, Galen Andrew, Salman Avestimehr, Katharine Daly, Deepesh Data, et al. A field guide to federated optimization. *arXiv preprint arXiv:2107.06917*, 2021.

Yawen Wu, Dewen Zeng, Zhepeng Wang, Yi Sheng, Lei Yang, Alaina J James, Yiyu Shi, and Jingtong Hu. Federated contrastive learning for dermatological disease diagnosis via on-device learning. In *2021 IEEE/ACM International Conference On Computer Aided Design (ICCAD)*, pp. 1–7. IEEE, 2021.

Timothy Yang, Galen Andrew, Hubert Eichner, Haicheng Sun, Wei Li, Nicholas Kong, Daniel Ramage, and Françoise Beaufays. Applied federated learning: Improving google keyboard query suggestions. *arXiv preprint arXiv:1812.02903*, 2018.

Jieyu Zhang, Cheng-Yu Hsieh, Yue Yu, Chao Zhang, and Alexander Ratner. A survey on programmatic weak supervision. *arXiv preprint arXiv:2202.05433*, 2022.

Weiming Zhuang, Xin Gan, Yonggang Wen, Shuai Zhang, and Shuai Yi. Collaborative unsupervised visual representation learning from decentralized data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4912–4921, 2021.