End-to-End Weak Supervision

Salva Rühling Cachay[1,2]* **Benedikt Boecking**[1] **Artur Dubrawski**[1]

[1] Carnegie Mellon University [2] Technical University of Darmstadt

Abstract

Aggregating multiple sources of weak supervision (WS) can ease the data-labeling bottleneck prevalent in many machine learning applications, by replacing the tedious manual collection of ground truth labels. Current state of the art approaches that do not use any labeled training data, however, require two separate modeling steps: Learning a probabilistic latent variable model based on the WS sources – making assumptions that rarely hold in practice – followed by downstream model training. Importantly, the first step of modeling does not consider the performance of the downstream model. To address these caveats we propose an end-to-end approach for directly learning the downstream model by maximizing its agreement with probabilistic labels generated by reparameterizing prior probabilistic posteriors with a neural network. Our results show improved performance over prior work in terms of end model performance on downstream test sets, as well as in terms of improved robustness to dependencies among weak supervision sources. [2]

1 Introduction

The success of supervised machine learning methods relies on the availability of large amounts of labeled data. The common process of manual data annotation by humans, especially when domain experts need to be involved, is expensive, both in terms of effort and cost, and as such presents a major bottleneck for deploying supervised learning methods to new domains and applications.

Recently, data programming, a paradigm that makes use of multiple sources of noisy labels, has emerged as a promising alternative to manual data annotation [30]. It encompasses previous paradigms such as distant supervision from external knowledge bases [29, 34], crowdsourcing [15, 26, 14, 41], and general heuristic and rule-based labeling of data [23, 20]. In the data programming framework, users encode domain knowledge into so called labeling functions (LFs), which are functions (e.g. domain heuristics or knowledge base derived rules) that noisily label subsets of data. The main task for learning from multiple sources of weak supervision is then to recover the sources' accuracies in order to estimate the latent true label, without access to ground truth data. In previous work [30, 32, 19], this is achieved by first learning a generative probabilistic graphical model (PGM) over the weak supervision sources and the latent true label to estimate *probabilistic labels*, which are then used in the second step to train a *downstream model* via a noise-aware loss function.

Data programming has led to a wide variety of success stories in domains such as healthcare [18, 17] and e-commerce [5], but the existing PGM based frameworks still come with a number of drawbacks. The separate PGM does not take the predictions of the downstream model into account, and indeed this model is trained independently of the PGM. In addition, current approaches for estimating the unknown class label via a PGM need to rely on computationally expensive approximate sampling methods [30], estimation of the full inverse of the LFs covariance matrix [32], or they need to make strong independence assumptions [19]. Furthermore, existing prior work and the associated

*salvaruehling@gmail.com.

[2]The most recent version can be retrieved at: `https://arxiv.org/abs/2107.02233`

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

Figure 1: For a task with unobserved ground truth labels Y, given m sources of weak supervision λ_i and training features X, WeaSEL trains a downstream model f_θ by maximizing the agreement of its predictions y_f with probabilistic labels $y_e = P(y = c_j \mid \lambda)$ generated by reparameterizing the posterior of prior work with sample-dependent accuracy scores produced by an encoder network e_ϕ.

theoretical analyses make assumptions that may not hold in practice [30, 32, 19], such as availability of a well-specified generative model structure (i.e. that the dependencies and correlations between the weak sources have been correctly specified by the user), that LF errors are randomly distributed across samples, and that the latent label is independent of the features given the weak labels (i.e. only the joint distribution between the sources and labels needs to be modeled).

We introduce WeaSEL, our Weakly Supervised End-to-end Learner model for training neural networks with, exclusively, multiple sources of weak supervision as noisy signals for the latent labels. WeaSEL is based on 1) reparameterizing previous PGM based posteriors with a neural encoder network that produces accuracy scores for each weak supervision source, and 2) training the encoder and downstream model on the same target loss, using the other model's predictions as constant targets to maximize the agreement between both models. The proposed method needs no labeled training data, and neither assumes sample-independent source accuracies nor redundant features for latent label modeling. We show empirically that it is not susceptible to highly correlated LFs. In addition, the proposed approach can learn from multiple probabilistic sources of weak supervision.

Our contributions include:

- We introduce a flexible, end-to-end method for learning models from multiple sources of weak supervision.
- We empirically demonstrate that the method is naturally robust to adversarial sources as well as highly correlated weak supervision sources.
- We release an open-source, end-to-end system for arbitrary PyTorch downstream models that will allow practitioners to take advantage of our approach[3].
- We show that our method outperforms, by as much as 6.1 F1 points, state-of-the-art latent label modeling approaches on 4 out of 5 relevant benchmark datasets, and achieves state-of-the-art performance on a crowdsourcing dataset against methods specifically designed for this setting.

2 Related Work

Multi-source Weak Supervision The data programming paradigm [30] allows users to programmatically label data through multiple noisy sources of labels, by treating the true label as a latent variable of a generative PGM. Several approaches for learning the parameters of the generative model have been introduced [32, 19, 11] to address computational complexity issues. Existing methods are susceptible to misspecification of the dependencies and correlations between the LFs, which can lead to substantial losses in performance [8]. Indeed, it is common practice to assume a conditionally independent model – without any dependencies between the sources – in popular libraries [5, 31] and related research [15, 2, 37, 7], even though methods to learn the intra-LF structure have been

[3] https://github.com/autonlab/weasel

proposed [4, 39, 38]. As in the approach proposed in this paper, the aforementioned methods do not assume any labeled training data, i.e. the downstream model is learned based solely on outputs of multiple LFs on unlabeled data. The traditional co-training paradigm [6] on the other hand is similar in spirit but requires some labeled data to be available. Recent methods that study the co-training setup where labeled training data supplements multiple WS sources, include [3, 25]. Note that the experiments in [3, 25] rely on large pre-trained language models, making the applicability of the approach without such models or to non-text domains unclear.

Crowdsourcing Aggregating multiple noisy labels is also a core problem studied in the crowdsourcing literature. Common approaches model worker performance and the unknown label jointly [15, 14, 41] using expectation maximization (EM) or similar approaches. Some core differences to learning from weak supervision sources are that errors by crowdworkers are usually assumed to be random, and that task assignment is not always fixed but can be optimized for. The benefits of jointly optimizing the downstream model and the aggregator of the weak sources have been recognized in multiple end-to-end methods that have been proposed for the crowdsourcing problem setting [13, 22, 40, 27, 35, 9]. They often focus on image labeling and EM-like algorithms for modeling and aggregating the workers. Importantly, our proposed approach can be used in general applications with weak supervision from multiple sources without any restrictive assumptions specific to crowdsourcing, and we show that our approach outperforms the aforementioned methods on a crowdsourcing benchmark task.

3 End-to-End Weak Supervision

In this section we present our flexible base algorithm that we call WeaSEL, which can be extended to probabilistic sources and other network architectures (Section 7). See Algorithm 1 for its pseudocode.

Algorithm 1 WeaSEL The proposed Weakly Supervised End-to-end Learning algorithm for learning from multiple weak supervision sources.

input: batch size n, networks e, f, inverse temperatures β_1, β_2, noise-aware loss function l, class balance $P(y)$.
for sampled minibatch $\{z^{(k)} = (x^{(k)}, \lambda^{(k)})\}_{k=1}^{n}$ do
  for all $k \in \{1, \dots, n\}$ do
    # Produce accuracy scores for all weak sources
    $z^{(k)} = \text{softmax}(e(z^{(k)}) \cdot \beta_1)$
    # Generate probabilistic labels
    define $s^{(k)}$ as $s^{(k)} = (z^{(k)})^T \bar{\lambda}^{(k)}$
    $y_e^{(k)} = P(y \mid \lambda^{(k)}) = \text{softmax}(s^{(k)} \cdot \beta_2 \cdot P(y))$
    # Downstream model forward pass
    $y_f^{(k)} = f(x^{(k)})$
  end for
  $L_f = \frac{1}{n} \sum_{k=1}^{n} l(y_f^{(k)}, \text{stop-grad}(y_e^{(k)}))$
  $L_e = \frac{1}{n} \sum_{k=1}^{n} l(y_e^{(k)}, \text{stop-grad}(y_f^{(k)}))$
  update e to minimize L_e, and f to minimize L_f
end for
return downstream network $f(\cdot)$

3.1 Problem Setup

Let $(x, y) \sim D$ be the data generating distribution, where the unknown labels belong to one of C classes: $y \in Y = \{1, \dots, C\}$. As in [30], users provide an unlabeled training set $X = \{x_i\}_{i=1}^{N}$, and m labeling functions $\lambda = (x) \in \{0, 1, \dots, C\}^m$, where 0 means that the LF abstained from labeling for any class. We write $\bar{\lambda} = (\mathbb{1}\{\lambda = 1\}, \dots, \mathbb{1}\{\lambda = C\}) \in \{0, 1\}^{m \cdot C}$ for the one-hot representation of the LF votes provided by the m LFs for C classes. Our goal is to train a downstream model $f : X \to Y$ on a noise-aware loss $l(y_f, y_e)$ that operates on the model's predictions $y_f = f(x)$ and probabilistic labels y_e generated by an encoder model e that has access to LF votes λ, and features x. Note that prior work restricts the probabilistic labels to only being estimated from the LFs.

3.2 Posterior Reparameterization

Previous PGM based approaches assume that the joint distribution $p(\lambda, y)$ of the LFs and the latent true label can be modeled as a Markov Random Field (MRF) with pairwise dependencies between weak supervision sources [30, 31, 32, 19, 11]. These models are parameterized by a set of LF accuracy and intra-LF correlation parameters and in some cases by additional parameters to model LF and class label propensity. Note however, that the aforementioned models ignore features x

when modeling the latent labels and therefore disregard that LFs may differ in their accuracy across samples and data slices.

We relax these assumptions, and instead view the latent label as an aggregation of the LF votes that is a function of the entire set of LF votes and features, on a sample-by-sample basis. That is, we model the probability of a particular sample x having the class label $c \in Y$ as

$$P(y = c \mid \Lambda) = \text{softmax}(s)_c \, P(y = c); \tag{1}$$

$$s = \theta(\Lambda; x)^T \bar{\alpha} \in \mathbb{R}^C : \tag{2}$$

where $\theta(\Lambda; x) \in \mathbb{R}^m$ weighs the LF votes on a sample-by-sample basis and the softmax for class c on s is defined as

$$\text{softmax}(s)_c = \frac{\exp\left(\theta(\Lambda; x)^T \mathbb{1}\{y = c\}\right)}{\sum_{j=1}^{C} \exp\left(\theta(\Lambda; x)^T \mathbb{1}\{y = j\}\right)}.$$

While we do not use the class balance $P(y)$ in our experiments for our own model, WeaSEL, it is frequently assumed to be known [32, 19, 11], and can be estimated from a small validation set, or from unlabeled data as described in [32]. Our formulation can be seen as a reparameterization of the posterior of the pairwise MRFs in [31, 32, 19], where α corresponds to the LF accuracies that are fixed across the dataset and are solely learned via LF agreement and disagreement signals, ignoring the informative features. We further motivate this formulation and expand upon this connection in the appendix A.

3.3 Neural Encoder

Based on the setup introduced in the previous section and captured in Eq. (1), our goal is to estimate latent labels by means of learning sample-dependent accuracy scores $\theta(\Lambda, x)$, which we propose to parameterize by a neural encoder. This network takes as input the features x and the corresponding LF outputs $\Lambda(x)$ for a data point, and outputs unnormalized scores $\theta(\Lambda, x) \in \mathbb{R}^m$. Specifically, we define

$$\theta(\Lambda; x) = \beta_2 \, \text{softmax}(e(\Lambda; x)\, \beta_1); \tag{3}$$

where β_2 is a constant factor that scales the final softmax transformation in relation to the number of LFs m, and is equivalent to an inverse temperature for the output softmax in Eq. 1. It is motivated by the fact that most LFs are sparse in practice, and especially when the number of LFs is large this leads to small accuracy magnitudes without scaling (since, without scaling, the accuracies after the softmax sum up to one.)[4] β_1 is an inverse temperature hyperparameter that controls the smoothness of the predicted accuracy scores: The lower it is, the less emphasis is given to a small number of LFs – as $\beta_1 \to 0$, the model aggregates according to the equal weighted votes of the LFs. The softmax transformation naturally encodes our understanding of wanting to aggregate the weak sources to generate the latent label.

3.4 Training the Encoder

The key question now is how to train e, i.e. how can we learn an accurate mapping of the sample-by-sample accuracies, given that we do not observe any labels?

We hypothesize that in most practical cases, features, latent label, and labeling function aggregations are intrinsically correlated due to the design decisions made by the users defining the features and LFs. Thus, we can jointly optimize e and f by maximizing their agreement with respect to the target downstream loss ℓ in an end-to-end manner. See Algorithm 1 for pseudocode of the resulting WeaSEL algorithm. The natural classification loss is the cross-entropy, which we use in our experiments, but in order to encode our desire of maximizing the agreement of the two separate models that predict based on different views of the data, we adapt it[5] in the following form: The loss is symmetrized in order to compute the gradient of both models using the other model's predictions as targets. To that end, it is crucial to use the *stop-grad* operation on the targets (the second argument of ℓ), i.e. to treat them as though they were ground truth labels. This choice is supported by our synthetic experiment and ablations. This operation has also been shown to be crucial in siamese, non-contrastive, self-supervised learning, both empirically [21, 12] and theoretically [36]. By minimizing simultaneously,

[4] In our main experiments we set $\beta_2 = \sqrt{m}$.

[5] This holds for any asymmetric loss, while for symmetric losses this is not needed.

Table 1: The final test F1 performance of various multi-source weak supervision methods over seven runs, using different random seeds, are averaged and standard deviation. The top 2 performance scores are highlighted as First, Second. Triplet-median [11] is not listed as it only converged for IMDB with 12 LFs (F1 = 73.0 0.22), and Spouse (F1 = 48.7 1.0). The downstream model is the same for all methods. For Sup. (Val. set), and Majority vote it is trained on the hard labels induced by the labeled validation set and the majority vote of the LFs, respectively. For the rest it is trained on the probabilistic labels estimated by the respective state-of-the-art latent label model. For reference, we also report the Ground truthperformance of the same downstream model trained on the true training labels (which are unused by all other models, and not available for Spouse).

Model	Spouse (9 LFs)		ProfTeacher (99 LFs)		IMDB (136 LFs)		IMDB (12 LFs)		Amazon (175 LFs)	
Ground truth	–		90.65	0.29	86.72	0.40	86.72	0.40	92.93	0.68
Sup. (Val. set)	20.4	0.2	73.34	0.00	68.76	0.00	68.76	0.00	84.18	0.00
Snorkel	48.79	2.69	85.12	0.54	82.22	0.18	74.45	0.58	80.54	0.41
Triplet	45.88	3.64	74.43	10.59	75.36	1.92	73.15	0.95	75.44	3.21
Triplet-Mean	49.94	1.47	82.58	0.32	79.03	0.26	73.18	0.23	79.44	0.68
Majority vote	40.67	2.01	85.44	0.37	80.86	0.28	74.13	0.31	84.20	0.52
WeaSEL	51.98	1.60	86.98	0.45	82.10	0.45	77.22	1.02	86.60	0.71

both, $L(y_e; y_f)$ and $L(y_f; y_e)$ to jointly learn the network parameters for and the downstream model f respectively, we learn the accuracies of the noisy sources that best explain the patterns observed in the data, and vice versa the feature-based predictions that are best explained by aggregations of LF voting patterns.

3.5 WeaSEL Design Choices

Note that it is necessary to encode the inductive bias that the unobserved ground truth y is a (normalized) linear combination of LF votes – weighted by sample- and feature-dependent accuracy scores. Otherwise, if the encoder network directly predicts $P(y \mid x)$ instead of the accuracies (x), the pair of networks $e; f$ have no incentive to output the desired latent label, without observed labels. We do acknowledge that this two-player cooperation game with strong inductive biases could still allow for degenerate solutions. However, we empirically show that our simple WeaSEL model that goes beyond multiple earlier WS assumptions is 1) competitive and frequently outperforms state-of-the-art PGM-based and crowdsourcing models (see Tables 1 and 2); and 2) is robust against massive LF correlations and able to recover the performance of a fully supervised model on a synthetic example, while all other models break in this setting (see section 4.3 and appendix F).

4 Experiments

Datasets As in related work on label models for weak supervision [30, 32, 19, 11], we focus for simplicity on the binary classification case with unobserved ground truth labels $y \in \{-1, 1\}$. See Table 3 for details about dataset sizes and the number of LFs used. We also run an experiment on a multi-class, crowdsourcing dataset (see subsection 4.2). We evaluate the proposed end-to-end system for learning a downstream model from multiple weak supervision sources on previously used benchmark datasets in weak supervision work [31, 7, 11]. Specifically, we evaluate test set performance on the following classification datasets:

- The IMDB movie review dataset [28] contains movie reviews to be classified into positive and negative sentiment. We run two separate experiments, where in one we use the same 12 labeling functions as in [11], and for the other we choose 136 text-pattern based LFs. More details on the LFs can be found in the appendix C.
- A subset of the Amazon review dataset [24], where the task is to classify product reviews into positive and negative sentiment.
- We use the BiasBios biographies dataset [16] to distinguish between binary categories of frequently occurring occupations and use the same subset of professor vs teacher classification as in [7].

- Finally, we use the highly unbalanced Spouse dataset (90% negative class labels), where the task is to identify mentions of spouse relationships amongst a set of news articles from the Signal Media Dataset [13].

For the Spouse dataset, the same data split and LFs as in [19] are used, while for the rest we take a small subset of the test set as validation set. This is common practice in the related works [31, 32, 19, 7] for tuning hyperparameters, and allows for a fair comparison of models.

4.1 Benchmarking Weak Supervision Label Models

To evaluate the proposed system, we benchmark it against state-of-the-art systems that aggregate multiple weak supervision sources for classification problems, without any labeled training data. We compare our proposed approach with the following systems: 1) Snorkel, a popular system proposed in [31, 32]; 2) Triplet, exploits a closed-form solution for binary classification under certain assumptions [19]; and 3) Triplet-mean and Triplet-median [11], which are follow-up methods based on Triplet with the aim of making the method more robust.

We report the held-out test set performance of WeaSEL's downstream model. Note that in many settings it is often not possible to apply the encoder model to make predictions at test time, since the LFs usually do not cover all data points (e.g. in Spouse only 25.8% of training samples get at least one LF vote), and can be difficult to apply to new samples (e.g. when the LFs are crowdsourced annotations). In contrast, the downstream model is expected to generalize to arbitrary unseen data points.

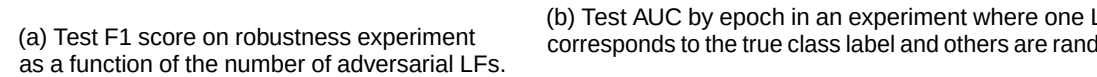
We observe strong results for our model, with 4 out of 5 top scores, and a lift of 6.1 F1 points over the next best label model-based method in the Amazon dataset. Our results are summarized in Table 1. Since our model is based on a neural network, we hypothesize that the large relative lift in performance on the Amazon review dataset is due to it being the largest dataset size on which we evaluate on – we expect this lift to hold or become larger as the training set size increases. To obtain the comparisons shown in Table 1, we run Snorkel over six different label model hyperparameter configurations, and train the downstream model on the labels estimated by the label model with the best AUC score on the validation set. We do not report Triplet-median in the main table, since it only converged for the two tasks with very small numbers of labeling functions. Interestingly, we observed that training the downstream model on the hard labels induced by majority vote leads to a competitive performance, better than triplet methods in four out of five datasets. This baseline is not reported in previous papers (only the raw majority vote is usually reported, without training a classifier). Our own model, WeaSEL, on the other hand consistently improves over the majority vote baseline (which in Table 4, in the appendix, can be seen to lead to similar performance as an untrained encoder network, that is left at its random initialization).

4.2 Crowdsourcing dataset

Data programming and crowdsourcing methods have been rarely compared against each other, even though the problem setup is quite similar. Indeed, end-to-end systems specifically for crowdsourcing have been proposed [33, 27, 35, 9]. These methods follow crowdsourcing-specific assumptions and modeling choices (e.g. independent crowdworkers, a confusion matrix model for each worker, and in general build upon [15]). Still, since crowdworkers can be seen as a specific type of labeling functions, the performance of general WS methods on crowdsourcing datasets is of interest, but has so far not been studied. We therefore choose to also evaluate our method on the multi-class LabelMe image classification dataset that was previously used in the core related crowdsourcing literature [35, 9]. The results are reported in Table 2, and more details on this experiment can be found in Appendix E. Note that the evaluation procedure in [9] reports the best test set performance for all models, while we

Table 2: Test accuracy scores on the crowdsourced, multi-class LabelMe image classification dataset.

Model	Accuracy	
Majority vote	79.23	0.5
MBEM [27]	76.84	0.4
DoctorNet [22]	81.31	0.4
CrowdLayer [35]	82.83	0.4
AggNet [1]	84.35	0.4
MaxMIG [9]	85.45	1.0
Snorkel+CE	82.89	0.7
WeaSEL+CE	82.46	0.8
Snorkel+MIG	85.15	0.8
WeaSEL+MIG	86.36	0.3



(a) Test F1 score on robustness experiment as a function of the number of adversarial LFs.

(b) Test AUC by epoch in an experiment where one LF corresponds to the true class label and others are random.

Figure 2: WeaSELs significantly more robust against correlated adversarial (left) or random (right) LFs than prior work whose assumptions make them equivalent to a Naive Bayes model. For subfigure (a), we duplicate a fake adversarial LF up to 10 times, and observe that our end-to-end system is robust against the adversarial LF, while other systems quickly degrade in performance (over ten random seeds). In (b), we let one LF be the true labels y and then duplicate a LF that votes according to a coin flip 2, 5, ..., 2000 times. We plot the test AUC performance curve as a function of the epochs, averaged out over the different number of duplications (and five random seeds). WeaSEL consistently recovers the test performance of the supervised end-model trained directly on the true labels y, whose end performance (AUC ≈ 0.967) is shown in red.

follow the more standard practice of reporting results obtained by tuning based on a small validation set – as in our main experiments. We find that our model, WeaSEL, is able to outperform Snorkel as well as multiple state-of-the-art methods that were specifically designed for crowdsourcing (including several end-to-end approaches). Interestingly, this is achieved by using the mutual information gain loss (MIG) function introduced in [9], which significantly boosts performance of both Snorkel (the end-model, f_θ, trained on the MIG loss with respect to soft labels generated by the first Snorkel label model step) and WeaSEL that use the cross-entropy (CE) loss. This suggests that the MIG loss is a great choice for the special case of crowdsourcing, due to its strong assumptions common to crowdsourcing which are much less likely to hold for general LFs. This is reflected in our ablations too, where using the MIG loss leads to a consistently worse performance on our main multi-source weak supervision datasets.

4.3 Robustness to Adversarial LFs and LF correlations

Users will sometimes generate sources they mistakenly think are accurate. This also encompasses the 'Spammer' crowdworker-type studied in the crowdsourcing literature. Therefore, it is desirable to build models that are robust against such sources. We argue that our system that is trained by maximizing the agreement between an aggregation of the sources and the downstream model's predictions should be able to distinguish the adversarial sources. In Fig. 2a we show that our system does not degrade in its initial performance, even after duplicating an adversarial LF ten times. Prior latent label models, on the other hand, rapidly degrade, given that they often assume the weak label sources to be conditionally independent given the latent label, equivalent to a Naive Bayes generative model. Note that the popular open-source implementation [31, 32] does not support user-provided LF dependencies modeling, while [10, 11] did not converge in our experiments when modeling dependencies, and as such we were not able to test their performance when the correlation dependencies between the duplicates are provided (which in practice, of course, are not known).

We also run a synthetic experiment inspired by [9], where one LF is set to the true labels of the ProfTeacher dataset, i.e. $\lambda_1 = y$, while the other LF simply votes according to a coin flip, i.e. $\lambda_2 \sim P(y)$, and we then duplicate this latter LF, i.e. $\lambda_3 = \dots = \lambda_m = \lambda_2$. Under this setting, our WeaSEL model is able to consistently recover the fully supervised performance of the same downstream model directly trained on the true labels y, even when we duplicate the random LF up to 2000 times ($m = 2001$). Snorkel and triplet methods, on the other hand, were unable to recover the true label (AUC ≈ 0.5). Importantly, we find that the design choices for WeaSEL are to a large extent key in order to recover the true labels in a stable manner as in Fig. 2b. Various other choices either collapse similarly to the baselines, are not able to fully recover the supervised performance, or lead to unstable test performance curves, see Fig. 5 in the appendix. More details on the experimental

Table 3: Dataset details, where training, validation and test set sizes are N_{train}, N_{val}, N_{test} respectively, and f denotes the downstream model type. We also report the total coverage Cov. of all LFs, which refers to the percentage of training samples which are labeled by at least one LF (the rest is not used). For IMDB we used two different sets of labeling functions of sizes 12 and 136.

Dataset	#LFs	N_{train}	Cov. (in %)	N_{val}	N_{test}	f
Spouse	9	22,254	2.58	2811	2701	LSTM
BiasBios	99	12,294	81.8	250	12,044	MLP
IMDB	12	25k	88.0	250	24,750	MLP
IMDB	136	25k	83.1	250	24,750	MLP
Amazon	175	160k	65.5	500	39,500	MLP

design and an extensive discussion, ablation, and gures based on the synthetic experiment can be found in the appendix F.

4.4 Implementation Details

Here we provide a high-level overview over the used encoder architecture, the LF sets, and the features. More details, especially hyperparameter and architecture details, are provided in Appendix C. All downstream models are trained with the (binary) cross-entropy loss, and our model with the symmetric version of it that uses stop-grad on the targets.

Encoder network The encoder network k does not need to follow a speci c neural network architecture and we therefore use a simple multi-layer perceptron (MLP) in our benchmark experiments.

Features for the encoder A big advantage of our model is that it is able to take into account the features x for generating the sample-by-sample source accuracies. For all datasets, we concatenate the LF outputs with the same features that are used by the downstream model as input of our encoder model (for Spouse we use smaller embeddings than the ones used by the downstream LSTM).

Weak supervision sources For the Spouse dataset, and the IMDB variant with 12 LFs, we use the same LFs as in [9, 11] respectively. The remaining three LF sets were selected by us prior to running experiments. These LFs are all pattern- and regex-based heuristics, while the Spouse experiments also contain LFs that are distant supervision sources based on DBPedia.

5 Ablations

In this section we demonstrate the strength of WeaSEL model design decisions. We perform extensive ablations on all four main datasets but Spouse for twenty con gurations of WeaSEL with different encoder architectures, hyperparameters, and loss functions. The tabular results and a more detailed discussion than in the following can be found in Appendix D.

We observe that ignoring the features when modeling the sample-dependent accuracies ($a(\lambda, x) = a(\lambda)$), usually underperforms by up to 1.2 F1 points. A more drastic drop in performance, up to 4.9 points, occurs when the encoder network is linear, i.e. without hidden layers, as in [9]. It also proves helpful to scale the softmax in Eq. 3 by $\sqrt{m}$ via the inverse temperature parameter $\tau = \frac{1}{2}$. Further, while the MIG loss proved important for WeaSEL to achieve state-of-the-art performance on the crowdsourcing dataset (with a similar lift in performance observable for Snorkel using MIG for downstream model training), this does not hold for the main datasets. This indicates that the MIG loss is a good choice for crowdsourcing, but not for more general WS settings.

Our ablations also show that it is important to restrict the accuracies to a positive interval (e.g. (0, 1), with the sigmoid function being a good alternative to the softmax we use). On the one hand, this encodes the inductive bias that LFs are not adversarial, i.e. can not have negative accuracies, (using tanh to output accuracy scores does not perform well), and on the other hand does not give the encoder network too much freedom in the scale of the scores (using ReLU underperforms signi cantly as well).

Additionally, we nd that our choice of using the symmetric cross-entropy loss with stop-grad applied to the targets is crucial for the obtained strong performance of WeaSEL. Removing the

stop-grad operation, or using the standard cross-entropy (without stop-grad on the target) leads to significantly worse scores and a very brittle model. Losses that already are symmetric (e.g. L1 or Squared Hellinger loss) neither need to be symmetrized nor use stop-grad. While the L1 loss consistently underperforms, we find that the Squared Hellinger loss can lead to better performance on two of the four datasets.

However, only the symmetric cross-entropy loss with stop-grad on the targets is shown to be robust and able to recover the true labels in our synthetic experiment in Section 4.3. Thus, to complement the above ablation on real datasets, we additionally run extensive ablations on this synthetic setup in Appendix F. This synthetic ablation gives interesting insights, and strongly supports the proposed design of WeaSEL. Indeed, many choices for WeaSEL that perform well enough on the real datasets, such as no features for the encoder ($h_2 = 1$), sigmoid parameterized accuracies, and all other losses that we evaluated, lead to significantly worse performance and less robust learning on the synthetic adversarial setups.

6 Practical Aspects and Limitations

On why it works & degenerate solutions Overall, WeaSEL avoids trivial overfitting and degenerate solutions by hard-coding the encoder generated labels as a (normalized) linear combination of the LF outputs, weighted by sample-dependent accuracy scores. This design choice also ensures that the randomly initialized encoder will lead the downstream model f that is trained on soft labels generated by the random encoder, to obtain performance similar to when f is trained on majority vote labels. In fact, the random-encoder WeaSEL variant itself often outperforms other baselines, and triplet methods in particular (see appendix B).

Empirically, we only observed degenerate solutions when training for too many epochs. Early-stopping on a small validation set ensures that a strong final solution is returned, and should be done whenever such a set exists or is easy to create. When no validation set is available, we find that choosing the temperature hyperparameter in Eq. 3 such that $\tau_1 = 3$ avoids collapsed solutions on all our datasets. This can be explained by the fact that a lower inverse temperature forces the encoder-predicted label to always depend on multiple LF votes when available, rather than a single one (which happens when the softmax in Eq. 3 becomes a max as $\tau_1 \to 1$). This makes it harder for the encoder to overfit to individual LFs. Our ablations indicate that this temperature parameter setting comes at a small cost in terms of loss in downstream performance, compared to when using a validation set for early stopping. Thus, when no validation set is available, we advise to lower τ_1.

Complex downstream models We have shown that WeaSEL achieves competitive or state-of-the-art performance on all datasets we tried it on, for a given set of LFs. In practice, however, this LF set needs to first be defined by users. This can be done via an iterative process, where the feedback is sourced from the quality of the probabilistic labels generated by the label model. A limitation of our model, is that each such iteration would require training the downstream model. When f is slow to train, this may slow down the LF development cycle and lead to unnecessary energy consumption. A practical solution to this can be to a) do the iteration cycle with a less complex downstream model; or b) use the fast to train PGM-based label models to choose a good LF set, and then move to WeaSEL in order to achieve better downstream performance.

7 Extensions

Probabilistic labeling functions Our learning method can easily support labeling functions that output continuous scores instead of discrete labels as in Eq. 1. In particular, this includes probabilistic sources that output a distribution over the potential class labels. This can be encoded in our model by changing the one-hot representation of our base model to a continuous representation $\bar{\lambda} \in [0,1]^{m \times C}$.

Modeling more structure While we use a simple multi-layer perceptron (MLP) as our encoder in our benchmark experiments, our formulation is flexible to support arbitrarily complex networks. In particular, we can naturally model dependencies amongst weak sources via edges in a Graph Neural Network (GNN), where each LF is represented by a node that is given the LF outputs as features. Furthermore, while we only explicitly reparameterized the accuracy parameters of the sources in our

base model, it is straightforward to augment $\bar{\pi}$ with additional sufficient statistics, e.g. the fixing or priority dependencies from $\mathcal{D}$, [8] that encode that one source fixes (i.e. should be given priority over) the other whenever both vote.

8 Conclusion

We proposed WeaSEL, a new approach for end-to-end learning of neural network models for classification from, exclusively, multiple sources of weak supervision that streamlines prior latent variable models. We evaluated the proposed approach on benchmark datasets and observe that the downstream models outperform state-of-the-art data programming approaches in 4 out of 5 cases while remaining highly competitive on the remaining task, and outperforming several state-of-the-art crowdsourcing methods on a crowdsourcing task. We also demonstrated that our integrated approach can be more robust to dependencies between the labeling functions as well as to adversarial labeling scenarios. The proposed method works with discrete and probabilistic labeling functions and can utilize various neural network designs for probabilistic label generation. This end-to-end approach can simplify the process of developing effective machine learning models using weak supervision as the primary source of training signal, and help adoption of this form of learning in a wide range of practical applications.

References

[1] S. Albarqouni, C. Baur, F. Achilles, V. Belagiannis, S. Demirci, and N. Navab. Aggnet: Deep learning from crowds for mitosis detection in breast cancer histology images. IEEE Transactions on Medical Imaging, 35(5):1313–1321, 2016.

[2] Animashree Anandkumar, Rong Ge, Daniel Hsu, Sham M Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. Journal of Machine Learning Research, 15: 2773–2832, 2014.

[3] Abhijeet Awasthi, Sabyasachi Ghosh, Rasna Goyal, and Sunita Sarawagi. Learning from rules generalizing labeled exemplars. ICLR, 2020.

[4] Stephen H. Bach, Bryan He, Alexander Ratner, and Christopher Ré. Learning the structure of generative models without labeled data. In Proceedings of the 34th International Conference on Machine Learning - Volume 70, ICML'17, page 273–282, 2017.

[5] Stephen H. Bach, Daniel Rodriguez, Yintao Liu, Chong Luo, Haidong Shao, Cassandra Xia, Souvik Sen, Alex Ratner, Braden Hancock, Houman Alborzi, Rahul Kuchhal, Chris Ré, and Rob Malkin. Snorkel drybell: A case study in deploying weak supervision at industrial scale. In Proceedings of the 2019 International Conference on Management of Data, SIGMOD '19, page 362–375, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450356435. doi: 10.1145/3299869.3314036.

[6] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In Proceedings of the Eleventh Annual Conference on Computational Learning Theory, COLT' 98, page 92–100, New York, NY, USA, 1998. Association for Computing Machinery. ISBN 1581130570. doi: 10.1145/279943.279962.

[7] Benedikt Boecking, Willie Neiswanger, Eric Xing, and Artur Dubrawski. Interactive weak supervision: Learning useful heuristics for data labeling. ICLR, 2021.

[8] Salva Rühling Cachay, Benedikt Boecking, and Artur Dubrawski. Dependency structure misspeci cation in multi-source weak supervision models. ICLR Workshop on Weakly Supervised Learning 2021.

[9] Peng Cao, Yilun Xu, Yuqing Kong, and Yizhou Wang. Max-mig: an information theoretic approach for joint learning from crowds. ICLR, 2019.

[10] Oishik Chatterjee, Ganesh Ramakrishnan, and Sunita Sarawagi. Data programming using continuous and quality-guided labeling functions. AAAI, 2020.

[11] Mayee F. Chen, Benjamin Cohen-Wang, Steve Mussmann, Frederic Sala, and Christopher Ré. Comparing the value of labeled and unlabeled data in method-of-moments latent variable estimation. AISTATS, 2021.

[12] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 15750–15758, 2021.

[13] David Corney, Dyaa Albakour, Miguel Martinez-Alvarez, and Samir Moussa. What do a million news articles look like? Workshop on Recent Trends in News Information Retrieval, pages 42–47, 2016.

[14] Nilesh Dalvi, Anirban Dasgupta, Ravi Kumar, and Vibhor Rastogi. Aggregating crowdsourced binary ratings. In Proceedings of the 22nd international conference on World Wide Web, pages 285–294, 2013.

[15] A. P. Dawid and A. M. Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics), 28(1): 20–28, 1979. ISSN 00359254, 14679876.

[16] Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. Bias in bios: A case study of semantic representation bias in a high-stakes setting. Proceedings of the Conference on Fairness, Accountability, and Transparency, pages 120–128, 2019.

[17] 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.

[18] 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 classi cation of aortic valve malformations using unlabeled cardiac mri sequences. Nature Communications, 10(1):1–10, 2019.

[19] Daniel Y. Fu, Mayee F. Chen, Frederic Sala, Sarah Hooper, Kayvon Fatahalian, and Christopher Ré. Fast and three-rious: Speeding up weak supervision with triplet methods. ICML, 2020.

[20] Garrett B Goh, Charles Siegel, Abhinav Vishnu, and Nathan Hodas. Using rule-based labels for weak supervised learning: a chemnet for transferable chemical property prediction. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pages 302–310, 2018.

[21] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. Advances in Neural Information Processing Systems, volume 33, pages 21271–21284, 2020.

[22] Melody Guan, Varun Gulshan, Andrew Dai, and Geoffrey Hinton. Who said what: Modeling individual labelers improves classi cation. Proceedings of the AAAI Conference on Arti cial Intelligence, 2018.

[23] Sonal Gupta and Christopher D Manning. Improved pattern learning for bootstrapped entity extraction. In Proceedings of the Eighteenth Conference on Computational Natural Language Learning, pages 98–108, 2014.

[24] Ruining He and Julian McAuley. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative ltering. In Proceedings of the 25th International Conference on World Wide Web, pages 507–517, 2016.

[25] Giannis Karamanolakis, Subhabrata (Subho) Mukherjee, Guoqing Zheng, and Ahmed H. Awadallah. Self-training with weak supervision. In NAACL, 2021.

[26] David Karger, Sewoong Oh, and Devavrat Shah. Iterative learning for reliable crowdsourcing systems. In Advances in Neural Information Processing Systems, volume 24, pages 1953–1961. Curran Associates, Inc., 2011.

[27] Ashish Khetan, Zachary C Lipton, and Anima Anandkumar. Learning from noisy singly-labeled data. ICLR, 2018.

[28] Andrew L. 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, pages 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

[29] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, pages 1003–1011, 2009.

[30] Alexander Ratner, Christopher De Sa, Sen Wu, Daniel Selsam, and Christopher Ré. Data programming: Creating large training sets, quickly. Advances in neural information processing systems, 29, 05 2016.

[31] Alexander Ratner, Stephen Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. Snorkel: rapid training data creation with weak supervision. The VLDB Journal, 29, 07 2019. doi: 10.1007/s00778-019-00552-1.

[32] Alexander Ratner, Braden Hancock, Jared Dunnmon, Frederic Sala, Shreyash Pandey, and Christopher Ré. Training complex models with multi-task weak supervision. Proceedings of the AAAI Conference on Arti cial Intelligence, 33:4763–4771, 07 2019. doi: 10.1609/aaai.v33i01.33014763.

[33] Vikas C. Raykar, Shipeng Yu, Linda H. Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. Learning from crowds. Journal of Machine Learning Research, 11(43):1297–1322, 2010.

[34] Sebastian Riedel, Limin Yao, and Andrew McCallum. Modeling relations and their mentions without labeled text. In Machine Learning and Knowledge Discovery in Databases, pages 148–163, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg. ISBN 978-3-642-15939-8.

[35] Filipe Rodrigues and Francisco Pereira. Deep learning from crowds. In Proceedings of the AAAI Conference on Arti cial Intelligence, 2018.

[36] Yuandong Tian, Xinlei Chen, and Surya Ganguli. Understanding self-supervised learning dynamics without contrastive pairs. ICML, 2021.

[37] 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, page 223. NIH Public Access, 2018.

[38] Paroma Varma, Bryan D He, Payal Bajaj, Nishith Khandwala, Imon Banerjee, Daniel Rubin, and Christopher Ré. Inferring generative model structure with static analysis. In Advances in neural information processing systems, pages 240–250, 2017.

[39] Paroma Varma, Frederic Sala, Ann He, Alexander Ratner, and Christopher Ré. Learning dependency structures for weak supervision models. In International Conference on Machine Learning, pages 6418–6427. PMLR, 2019.

[40] Peter Welinder, Steve Branson, Pietro Perona, and Serge Belongie. The multidimensional wisdom of crowds. In Advances in Neural Information Processing Systems, volume 23, pages 2424–2432. Curran Associates, Inc., 2010.

[41] Yuchen Zhang, Xi Chen, Dengyong Zhou, and Michael I Jordan. Spectral methods meet em: A provably optimal algorithm for crowdsourcing. The Journal of Machine Learning Research, 17 (1):3537–3580, 2016.

Appendix

A Posterior Reparameterization

In this section we motivate the design choices and inductive biases that we encode into our neural encoder network θ, which is the network that is used to model the relative accuracies of the weak supervision sources. Recall that we model the probability of a particular sample $x \in X$ having the class label $y \in Y = \{1, \ldots, C\}$ as

$$P_\theta(y|x) = \text{softmax}(s)_y \, P(y); \tag{4}$$

$$s = \theta(\lambda, x)^T \bar{\lambda} \in \mathbb{R}^C: \tag{5}$$

where $\theta(\lambda, x) \in \mathbb{R}^m$ weighs the LF votes on a sample-by-sample basis and the softmax for class y on s is defined as

$$\text{softmax}(s)_y = \frac{\exp\left(\theta(\lambda, x)^T \mathbb{1}\{\lambda = y\}\right)}{\sum_{y' \in Y} \exp\left(\theta(\lambda, x)^T \mathbb{1}\{\lambda = y'\}\right)}.$$

Connection to prior PGM models We now motivate this choice by deriving a less expressive variant of it from the standard Markov Random Field (MRF) used in the related work. If we view the attention scores $\theta(\lambda, x) \in \mathbb{R}^m$, that assign sample-dependent accuracies to each labeling function, as sample-independent parameters θ_1 and, by that, drop the features from the equation – as is done in the related work [30, 32, 19, 11] – we can rewrite Eq. 4 as

$$\frac{\exp\left(\theta_1^T \mathbb{1}\{\lambda = y\}\right)}{\sum_{y' \in Y} \exp\left(\theta_1^T \mathbb{1}\{\lambda = y'\}\right)} P(y)$$

Let $\phi_1(\lambda, y) = \mathbb{1}\{\lambda = y\}$, and, for clarity of writing, we drop the class balance, then this becomes

$$\begin{aligned}
&= \frac{\exp\left(\theta_1^T \phi_1(\lambda, y)\right)}{\sum_{y' \in Y} \exp\left(\theta_1^T \phi_1(\lambda, y')\right)} \\
&= \frac{Z^{-1} \exp\left(\theta_1^T \phi_1(\lambda, y) + \theta_2^T \phi_2(\lambda)\right)}{\sum_{y' \in Y} Z^{-1} \exp\left(\theta_1^T \phi_1(\lambda, y') + \theta_2^T \phi_2(\lambda)\right)} \\
&= \frac{P_\theta(\lambda, y)}{\sum_{y' \in Y} P_\theta(\lambda, y')} \\
&= \frac{P_\theta(\lambda, y)}{P_\theta(\lambda)} \\
&= P_\theta(y|\lambda);
\end{aligned}$$

where in the second step we multiplied the denominator and numerator with the same quantity $\frac{1}{Z} \exp\left(\theta_2^T \phi_2(\lambda)\right)$, and now θ parameterizes the joint distribution of the latent label and weak sources as

$$P_\theta(\lambda, y) = \frac{1}{Z} \exp\left(\theta_1^T \phi_1(\lambda, y) + \theta_2^T \phi_2(\lambda)\right) = \frac{1}{Z} \exp\left(\theta^T \phi(\lambda, y)\right).$$

We can recognize P_θ as a distribution from the exponential family, and more specifically as a pairwise MRF, or factor graph, with canonical parameters $\theta = (\theta_1; \theta_2)$ and corresponding sufficient statistics, or factors, $\phi(\lambda, y) = (\phi_1(\lambda, y); \phi_2(\lambda))$, as well as the log partition function Z. The accuracy factors and parameters ϕ_1, θ_1 are the core component of this model and sometimes take the form $\phi_1(\lambda, y) = \lambda y$ in binary models as in [30, 19, 11]. The label-independent factors $\phi_2(\lambda)$ have, as can be seen from the derivation above, no direct influence on the latent label posterior, but are often used to model labeling propensities $\mathbb{1}\{\lambda \neq 0\}$ and correlation dependencies $\mathbb{1}\{\lambda_i = \lambda_j\}$, which can be important for PGM parameter learning, but are susceptible to misspecification [30, 11, 8]. Our own parameterization therefore is a more expressive variant of these latent-variable PGM models, where we are able to assign LF accuracies on a sample-by-sample basis. Furthermore, our neural encoder network outputs them as a function of the LF outputs and features, and is expected to learn the easy to misspecify dependencies and label-independent statistics implicitly. Indeed, our empirical findings and subsection 4.3 support this.

Table 4: The final test F1 performance of various multi-source weak supervision methods over seven runs, using different random seeds, are averaged and standard deviation. The top 2 performance scores are highlighted as First, Second. Triplet-median [11] is not listed as it only converged for IMDB with 12 LFs (F1 = 73.0 0.22), and Spouse (F1 = 48.7 1.0). Sup. (Val. set) is the performance of the downstream model trained in a supervised manner on the labeled validation set. The rest are state-of-the-art latent label models. For reference, we also report the Ground truth performance of a fully supervised model trained on true training labels (which are unused by all other models, and not available for Spouse). We also report the performance of WeaSEL-Random, where only the downstream model of WeaSEL is trained (and the encoder network is left at its randomly initialized state). All models are run twice, where only the learning rate differs (either 10^{-4} or $4 \cdot 10^{-5}$), and the model with best ROC-AUC on the validation set is reported. The probabilistic labels from Snorkel used for downstream model training are chosen over six different configurations of the learning rate and number of epochs (again with respect to validation set ROC-AUC).

Model	Spouse (9 LFs)		ProfTeacher (99 LFs)		IMDB (136 LFs)		IMDB (12 LFs)		Amazon (175 LFs)	
Ground truth	–		90.65	0.29	86.72	0.40	86.72	0.40	92.93	0.68
Sup. (Val. set)	20.4	0.2	73.34	0.00	68.76	0.00	68.76	0.00	84.18	0.00
Snorkel	48.79	2.69	85.12	0.54	82.22	0.18	74.45	0.58	80.54	0.41
Triplet	45.88	3.64	74.43	10.59	75.36	1.92	73.15	0.95	75.44	3.21
Triplet-Mean	49.94	1.47	82.58	0.32	79.03	0.26	73.18	0.23	79.44	0.68
WeaSEL-Random	46.43	3.29	83.47	0.64	79.80	0.48	74.22	0.45	82.22	0.57
Majority vote	40.67	2.01	85.44	0.37	80.86	0.28	74.13	0.31	84.20	0.52
WeaSEL	51.98	1.60	86.98	0.45	82.10	0.45	77.22	1.02	86.60	0.71

B Extended Results

We provide more detailed results in Table 4. Here, we include WeaSEL-Random, which corresponds to WeaSEL with a randomly initialized encoder network that is not trained/updated. As expected, this setting produces performance often similar compared to training an end model on the hard majority vote labels. This is due to the strong inductive bias in our encoder model that constrains the encoder labels to be a normalized linear combination of the LF votes, weighted by positive accuracy scores. In fact, WeaSEL-Random itself is often able to outperform the PGM-based baselines, in particular the triplet methods. Our results show that WeaSEL consistently improves significantly upon these baselines via training the encoder network to maximize its agreement with the downstream model.

C Extended Implementation Details

Weak supervision sources For the Spouses dataset, and the IMDB variant with 12 LFs, we use the same LFs as in [19] and [11], respectively[6]. The set of 12 IMDB LFs was specifically chosen to have a large coverage, see Table 3. These LFs and the larger set of LFs that we introduce for the second IMDB experiment are all pattern- and regex-based heuristics, i.e. LFs that label whenever a certain word or bi-gram appears in a text document. For instance, 'excellent' would label for the positive movie review sentiment (and would do so with 80% accuracy on the samples where it does not abstain). This holds for the other text datasets as well, while the Spouse experiments also contain LFs that are distant supervision sources based on DBPedia.
For the remaining datasets (IMDB with 136 LFs, Bias Bios, and Amazon), we created the respective LF sets ourselves, prior to running experiments.

Encoder network architectures In all experiments, we use a simple multi-layer perceptron (MLP) as the encoder, with two hidden layers, batch normalization, and ReLU activation functions. For the Spouse dataset, we use a bottleneck-structured network of sizes 50, 5. This is motivated by the small size of the set of samples labeled by at least one LF. For all other datasets we use hidden dimensions of 70, 70. We show in the ablations (Table 5), that our end-to-end model also succeeds for different encoder architecture choices.

[6] All necessary label matrices are available in our research source code. The Spouse LFs and data are also available at the following URL: https://github.com/snorkel-team/snorkel-tutorials/blob/master/spouse/spouse_demo.ipynb

Downstream models For all datasets besides Spouse, we use a three-layer MLP with hidden dimensions 50, 50, 25. For Spouse, we use a single-layer bidirectional LSTM with a hidden dimension of 150, followed by two fully-connected readout layers with dimensions 64, 32. All fully-connected, layers use ReLU activation functions. We choose simple downstream architectures as we are interested in the relative improvements over other label models. More sophisticated architectures are expected to further improve the performances, however.

Hyperparameters Unless explicitly mentioned, all reported experiments are averaged out over seven random seeds. We use an L2 weight decay of $7e{-}7$ and dropout of 0.3 for both encoder and downstream model for all datasets but Spouse (where the LSTM does not use dropout). All models are optimized with Adam, with early-stopping based on AUC performance on the small validation set, and a maximum number of 150 epochs (75 for Spouse). The batch size is set to 64. The loss function is set to the (binary) cross-entropy. For each dataset and each model/baseline, we run the same experiment for learning rates of $1e{-}4$ and $3e{-}5$, and then report the model chosen according to the best ROC-AUC performance on the small validation set. For Spouse we additionally run experiments with a L2 weight decay of $1e{-}4$ which due to the risk of over tting to the small size of LF-covered data points boosts performance for all models. For our own model, WeaSEL, we also run additional experiments for Spouses with different con gurations of the temperature hyperparameter, $_1 \in \{1, 1/3\}$ and again report the test performance as measured by the best validation ROC-AUC. The probabilistic labels from Snorkel used for downstream model training are chosen over six different con gurations of the learning rate and number of epochs for Snorkel's label model (again with respect to validation set ROC-AUC). For all binary classi cation datasets (i.e. all except for LabelMe), we tune the downstream model's decision threshold based on the resulting F1 validation score for all models. We believe that this, alternatively to reporting test ROC-AUC scores, makes the comparison fairer, since F1 is a threshold dependent metric. All label model baselines are provided with the class balance, which WeaSEL does not use (but which is expected to be helpful for unbalanced classes, where no validation set is available).

D Extended Ablations

The full ablations are reported in Table 5, where in each row we change or remove exactly one component of our proposed model, WeaSEL. We nd that the design choices of WeaSEL, which were inspired by sensible inductive biases for an encoder label model are hard to beat by various changes to the architecture, loss function, or hyperparameters. Indeed, most changes consistently underperform WeaSEL, and the occasional positive changes – 1e-4 weight decay, and the Squared Hellinger loss instead of the symmetric cross-entropy – only beat the WeaSEL performance in at most two datasets, and never signi cantly. In practice, we advise to explore these strongest con gurations if a small validation set is available.

We nd that letting the accuracy scores depend on the input features (rst row), usually boosts performance, but not by much (1.2 F1 points at most). On the other hand, it proves very important to allow these accuracy scores to depend non-linearly on the LF votes and the features: A linear encoder network, as in [9], signi cantly underperforms WeaSEL with at least one hidden layer by up to 4.9 F1 score points. Conversely, a deeper encoder network (of hidden dimensions $[75, 50, 25]$, $[50, 50, 50]$, $[75, 75, 75]$, see fourth row) does not improve results. This may be due to the sample-dependent accuracies not being a too complex function to learn.

While the effect of the inverse temperature parameter $_1$, which controls the softness of the encoder-predicted accuracy scores–on downstream performance is not large, it can have signi cant effects on the learning dynamics and robustness, see Fig 3 for such learning curves as a function of epoch number. In particular, a lower $_1$ makes the dynamics more robust, since the accuracy score weights are more evenly distributed across LFs, which appears to help avoid over tting. When over tting is not easily detectable due to a lack of a validation set, it is therefore advisable to use a lower $_1$. It also proves helpful to scale the softmax in Eq. 3 by $\sqrt{m}$, rather than not scaling it ($_1 = 1$ row) or scaling by m.

Changing the loss function from the symmetric cross-entropy to the MIG function [9] or the L1 loss consistently leads to worse performance. The former is interesting, since using the MIG loss for the crowdsourcing dataset LabelMe, see subsection 4.2, was important in order to achieve state-of-the-art crowdsourcing performance (with a similar lift in performance observable for Snorkel using MIG for downstream model training). The result provides some evidence that the MIG loss

Table 5: Ablative study on the subcomponents of our algorithm as in Alg. 1 (over 5 random seeds). In each row below we change exactly one component of WeaSEL and report the resulting F1 score. Note that the scores for WeaSEL are slightly different to the ones in the main results table, since they were run separately, with fewer seeds, and for only one learning rate (1e-4). Configurations that outperform base WeaSEL are highlighted in bold font, while the four worst performing configurations are highlighted in red for each dataset. Note that bold font does not indicate significant differences.

Change	Prof	Teacher	IMDB-136 LFs		IMDB-12 LFs		Amazon	
WeaSEL	86.8	0.4	82.1	0.7	77.3	0.5	86.6	0.5
$(\lambda; x) = \lambda(\lambda)$	85.6	1.6	82.1	0.5	75.9	0.8	86.6	0.4
Linear e	81.9	0.7	80.0	0.6	73.2	0.6	82.6	0.5
1 hidden layer e	87.1	0.7	81.8	0.6	76.8	0.9	85.3	0.8
75x50x25x50x75 e	84.3	2.1	81.9	0.6	75.8	1.1	86.1	0.6
$_1 = 2$	86.7	1.0	81.9	0.3	77.3	0.5	85.5	1.0
$_1 = 1=2$	86.5	0.8	81.8	0.5	76.0	1.4	86.4	0.3
$_1 = 1=4$	84.5	1.2	81.8	0.2	73.9	0.9	85.6	1.0
$_2 = 1$	85.2	1.6	82.2	0.4	76.6	1.0	84.3	1.2
$_2 = m$	86.1	0.7	81.2	0.6	76.4	0.4	85.7	0.2
No BatchNorm	82.6	1.4	81.9	0.5	74.7	0.7	85.3	0.8
1e-4 weight decay	87.4	0.4	80.9	1.3	77.9	0.6	85.2	0.5
MIG loss	86.7	0.4	78.7	0.4	74.1	0.4	84.7	1.8
L1 loss	86.2	0.6	81.1	0.5	75.6	0.9	84.1	0.9
Squared Hellinger loss	87.4	0.3	82.2	0.6	75.7	1.1	86.3	0.4
$CE(P_f; P_e)$ asymm. loss	77.3	3.7	77.7	1.1	71.7	0.3	78.7	1.2
$CE(P_e; P_f)$ asymm. loss	73.1	6.8	71.9	1.9	69.7	0.7	70.1	1.1
No stop-grad	80.4	2.1	76.2	0.5	71.0	0.6	79.3	0.6
$(\lambda; x) = \sqrt{m}\ \text{sigmoid}(\lambda(\lambda; x))$	85.5	0.6	81.8	0.5	78.0	0.7	86.9	0.3
$(\lambda; x) = \text{ReLU}(\lambda(\lambda; x)) + 1e\text{-}5$	83.0	2.3	78.3	1.1	69.1	2.1	74.2	2.7
$(\lambda; x) = \text{Tanh}(\lambda(\lambda; x))$	71.9	4.0	67.0	0.8	67.0	1.1	67.3	1.1

may be inappropiate for weak supervision settings other than crowdsourcing, while its use may be recommended for that specific setting.

We find that it is important to constrain the accuracy score space to a positive interval, either by viewing them as an aggregation of the LFs via the scaled softmax in Eq. 3, or by replacing the softmax with a sigmoid function. Indeed, using a less constrained activation function for the estimated accuracies (last two rows, where the 1e-5 in the ReLU row avoids accuracy scores equal to zero) significantly underperforms: Allowing the accuracies to be negative (last row) leads to collapse and bad downstream performance. This is likely due to the removal of the inductive bias that LFs are better-than-random, which makes the joint optimization more likely to find trivial solutions. Additionally, we find that our choice of using the symmetric cross-entropy loss with stop-grad applied to the targets is crucial for the strong performance of WeaSEL. Removing the stop-grad operation, or using the standard cross-entropy (with stop-grad on the target) leads to significantly worse scores and a very brittle model. This is somewhat expected, since conceptually our goal is to have an objective that maximizes the agreement between a pair of models that predict based on two different views of the latent label, the features and the LF votes. The cross-entropy with stop-grad on the target[7] naturally encodes this understanding, since each model uses the other model's predictions as a reference distribution. Losses that already are symmetric (e.g. L1 or Squared Hellinger loss) neither need to be symmetrized nor use stop-grad. While the L1 loss consistently underperforms, we find that the Squared Hellinger loss can lead to better performance on two out of four datasets.

However, only the symmetric cross-entropy loss with stop-grad on the targets is shown to be robust and able to recover the true labels in our synthetic experiments in appendix F, see Fig. 5 in particular. The synthetic ablation in appendix F gives interesting insights, and strongly supports the proposed design of WeaSEL. Indeed, many choices for WeaSEL that perform well enough on the real datasets,

[7] or, due to the stop-grad operation, equivalently the KL divergence

such as no features for the encoder, $_2 = 1$, sigmoid parameterized accuracies, and all other objectives that we evaluated, lead to significantly worse performance and less robust learning on the synthetic adversarial setups.

E Crowdsourcing dataset

As the crowdsourcing dataset, we choose the multi-class LabelMe image classification dataset that was previously used in the most related crowdsourcing literature [35, 9]. Note that this dataset consists of $10k$ samples, of which only $1k$ are unique, in the sense that the rest are augmented versions of the $1k$. They were annotated by 59 crowdworkers, with a mean overlap of 2.55 annotations per image. The downstream model is identical to the previously reported one [35, 9]. That is, a VGG-16 neural network is used as feature extractor, and a single fully-connected layer (with 128 units and ReLU activation) and one output layer is put on top, using 50 % dropout.

Experiments were conducted over seven random seeds with a learning rate of 1e-4 and 50 epochs. The reported scores are the ones with best validation set accuracy for a L2 weight decay $2 f$ 7e-7, 1e-4 g. The validation set is of size 200, and was split at random from the training set prior to running the experiments.

As is usual in the related work for multi-class settings [31], we employ class-conditional accuracies $(; x) 2 R^{m \times C}$ instead of only m class-independent accuracies. Recall the LF outputs indicator matrix, $\bar{} 2 R^{m \times C}$. To compute the resulting output softmax logits $s 2 R^{C}$, we set $A = (; x) \bar{} 2 R^{m \times C}$ and $s_j = \sum_i A_{ij} 2 R$, where $$ is the element-wise matrix product and we sum up the resulting matrix A across the LF votes dimension.

Snorkel+MIG indicates that the downstream model f was trained on the MIG loss with respect to soft labels generated by the first Snorkel step, label modeling. Snorkel+CE refers analogously to the same training setup, but using the cross-entropy (CE) loss. All crowdsourcing baseline models are based on the open-source code from [9].

F Robustness experiments

In this section we give more details on the experiments that validate the robustness of our approach against (strongly) correlated LFs that are not better than a random coin flip. In addition, we present one further experiment where the random LFs are independent of each other – a more difficult setup for learning (but which does not violate any assumptions of the PGM-based methods) – and our model, `WeaSEL`, again is shown to be robust to a large extent.

In contrast to `WeaSEL`, prior PGM-based work [31, 19, 11] attain significantly worse performance under these settings, due to assuming a Naive Bayes generative model where the weak label sources are conditionally independent given the latent label.

F.1 Adversarial LF duplication

For this experiment we use our set of 12 LFs for the IMDB dataset and generate a fake adversarial source by flipping the abstain votes, of the 80%-accurate LF that labels for the positive sentiment on 'excellent', to negative ones.

F.2 Recovery of true labels under massive LF noise

In this set of synthetic experiments we again validate the robustness of our approach. We focus on the Bias in Bios dataset, and use the features and true labels, y^*, therein. We let our initial LF set consist of 1) a 100% accurate LF, that is we set $_1 = y^*$, and 2) a LF that votes according to the class balance (i.e. a coin flip with probabilities for tail/head set according to the class balance), i.e. $_2 \quad P(y)$. In the first experiment we then add the same random LF $_2$ multiple times into the LF set (i.e. we duplicate it), see F.2.1, while in the second one, we incrementally add random LFs independently of $_2$ (and independently of any other LF already in the LF set), see F.2.2. For both setups, our model, `WeaSEL`, is able to recover the performance of the same downstream model, f, that is directly trained on the true labels, y^* (F1 = 90.65, ROC-AUC = 0.967, see Table 4). In contrast, the PGM-based baselines quickly collapse.

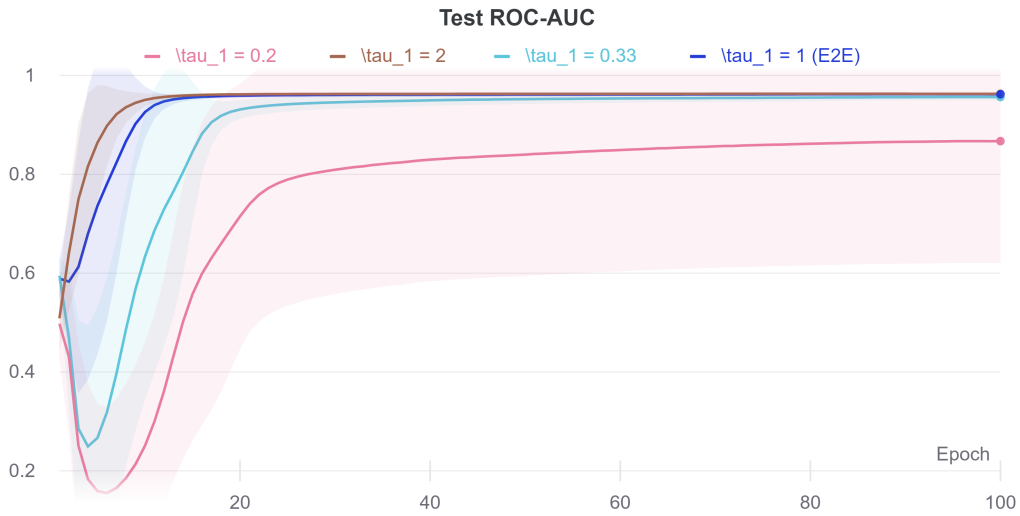


Figure 3: Test AUC performance at each training epoch for different choices of $_1$ 2 $f1{=}5; 1{=}3; 1; 2g$ on our synthetic experiment, see appendix F.2.1, averaged out over the number of duplicates and five random seeds. A lower $_1$ leads to slower or worse convergence in this specific case. A lower $_1$ corresponds to smoother accuracies, which makes their induced label depend on more LFs. Since in this specific case only one LF is 100% accurate and the rest are not better than a coin flip, the shown behavior is expected.

F.2.1 Random LF duplication

This experiment is inspired by the theoretical comparison in Appendix E of [9] between the authors' end-to-end system and maximum likelihood estimation (MLE) approaches that assume mutually independent LFs. The authors show that such MLE methods are not robust against the following simple example with correlated LFs. Based on the setup described above in F.2, we duplicate the random LF $_2$ multiple times, i.e. $_3 = \quad = \ _m = \ _2$. We run experiments for varying number of duplicates 2 $f2; 25; 100; 500; 2000g$. With this synthetic set of m LFs, where one LF is 100% accurate while the other m 1 LFs are just as good as a random guess, we train WeaSEL in the usual way on the features from the Bias in Bios dataset as well as the corresponding, just created, LF votes. WeaSEL is able to consistently and almost completely *recover this fully supervised performance, even when the number of duplicates is very high* ($m = 2001$). Snorkel and triplets methods, on the other hand, fare far worse (AUC 0:5) for all numbers of duplicates. This behavior is similar to the one observed in F.1 (see Fig. 2 for the performance of the baselines and WeaSEL averaged out over the varying number of duplicates, and Fig. 5a-c for the separate performance of WeaSEL for each number of duplicates).

We also run an additional ablation study on this synthetic experiment that shows that the observed robustness does not hold for all configurations of WeaSEL. In Fig. 5 we plot the test performance curves over the training epochs for each number of LF duplications.
Our proposed model, WeaSEL enjoys a stable and robust test curve (Fig. 5c) and quickly recovers the fully supervised performance, even with 2000 LF duplicates (although convergence becomes slower as the LF set contains more duplicates). On the other hand, we find that many other configurations and designs of WeaSEL lead to less robust and worse converging curves, collapses or bad performances. Indeed, for this experiment it is key to use as the loss function the proposed symmetric cross-entropy with stop-grad applied to the targets (see Fig. 5e, 5f), accuracies parameterized by a scaled (Fig. 5h) softmax (Fig. 5g), and, to a lesser extent, using the features an input to the encoder (Fig. 5d).
While the impact of not using stop-grad, or using an asymmetric cross-entropy loss is similarly bad in the main ablations on our real datasets, other configurations, and in particular sigmoid-parameterized accuracies (the choice in [25]), an unscaled softmax, and no features for the encoder, often perform well there. This additional ablation, however, provides support for why the good performances on the real datasets notwithstanding, our proposed design choices are most appropriate in order to attain strong test performances as well as stable and robust learning.

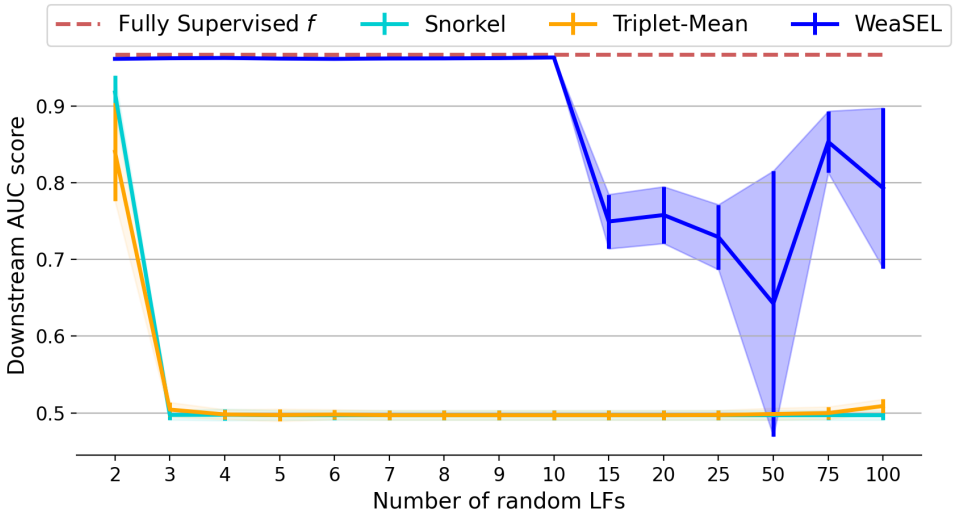


Figure 4: We start with a 100% accurate LF (i.e. ground truth labels) and incrementally add new, independent LFs that are no better than a random guess. WeaSEL recovers the performance of training directly on the ground truth labels (Fully Supervised f), for up to 10 such randomly voting LFs that are independent of each other. The PGM-based prior work, rapidly degrades in performance (AUC 0.5) and is not able to recover any of the 100% accurate signal of the true-labels-LF, as soon as the LF set is corrupted by three or more random LFs. Performances are averaged out over five random seeds, and the standard deviation is shaded. For more details, see F.2.2

F.2.2 Random, independent LFs

We start with the same setup as above in F.2, but instead of duplicating the same LF multiple times as in F.2.1, we now draw a new, independent random LF at each iteration. That is, we start with $_1 = y^*$; $_2$ $P(y)$ as our initial LFs, and the incrementally add new LFs $_i$ $P(y)$ that have no better skill than a coin flip. Note that this is arguably a harder setup than the one in the previous experiments, since there the LF set was corrupted by a single LF voting pattern. In this experiment, multiple equally bad, but independent, LFs corrupt the 100% accurate signal of $_1$. Notably, since these $_2; \ldots; _m$ are independent, we are not violating the independence assumptions of PGM-based methods. Nonetheless, we find that these PGM-based baselines break with only three ($m = 4$) of such random, but independent LFs, while WeaSEL is shown to be fully robust and able to recover the ground truth LF $_1$ for up to 10 random LFs ($m = 11$). For more LFs, WeaSEL starts deteriorating in performance, but is still able to consistently outperform the trivial solution of voting randomly according to the class balance (i.e. based on $_2; \ldots; _m$) and the baselines, see Fig. 4.

G Broader Impact

Large labeled datasets are important to many machine learning applications. Reducing the expensive human effort required to annotate such datasets is an important step towards making machine learning more accessible, more manageable, more beneficial, and therefore used more broadly. Our proposed end-to-end learning for weak supervision approach provides another step towards the practical utility of learning from multiple sources of weak labels on large datasets. Methods such as the one presented in our paper must be applied with care. One of the risks to consider and mitigate in a particular application is the possibility of incorporating biases from subjective humans who chose weak labeling sources. This is particularly the case when heuristics might apply differently to different subgroups in data, such as may be the case in scenarios highlighted in recent research towards fairness in machine learning.