Training Complex Models with Multi-Task Weak Supervision

Alexander Ratner, Braden Hancock, Jared Dunnmon, Frederic Sala,
Shreyash Pandey, Christopher Ré
Department of Computer Science
Stanford University
{ajratner, bradenjh, jdunnmon, fredsala, shreyash, chrismre}@stanford.edu

Abstract

As machine learning models continue to increase in complexity, collecting large hand-labeled training sets has become one of the biggest roadblocks in practice. Instead, weaker forms of supervision that provide noisier but cheaper labels are often used. However, these weak supervision sources have diverse and unknown accuracies, may output correlated labels, and may label different tasks or apply at different levels of granularity. We propose a framework for integrating and modeling such weak supervision sources by viewing them as labeling different related sub-tasks of a problem, which we refer to as the *multi-task weak supervision* setting. We show that by solving a matrix completion-style problem, we can recover the accuracies of these *multi-task* sources given their dependency structure, but without any labeled data, leading to higher-quality supervision for training an end model. Theoretically, we show that the generalization error of models trained with this approach improves with the number of *unlabeled* data points, and characterize the scaling with respect to the task and dependency structures. On three fine-grained classification problems, we show that our approach leads to average gains of 20.2 points in accuracy over a traditional supervised approach, 6.8 points over a majority vote baseline, and 4.1 points over a previously proposed weak supervision method that models tasks separately.

1 Introduction

One of the greatest roadblocks to using modern machine learning models is collecting hand-labeled training data at the massive scale they require. In real-world settings where domain expertise is needed and modeling goals change frequently, hand-labeling training sets is prohibitively slow, expensive, and static. For these reasons, practitioners are increasingly turning to weak supervision techniques wherein noisier, often programmatically-generated labels are used instead. Common *weak supervision sources* include external knowledge bases (Mintz et al. 2009; Zhang et al. 2017a; Craven and Kumlien 1999; Takamatsu, Sato, and Nakagawa 2012), heuristic patterns (Gupta and Manning 2014; Ratner et al. 2018), feature annotations (Mann and McCallum 2010; Zaidan and Eisner 2008), and noisy crowd labels (Karger, Oh, and Shah 2011; Dawid and Skene 1979). The use of these sources has led to state-of-the-art results in a range of domains (Zhang et al. 2017a; Xiao et al. 2015). A theme of weak supervision is that using the full diversity of available sources is critical to training high-quality models (Ratner et al. 2018; Zhang et al. 2017a).

The key technical difficulty of weak supervision is determining how to combine the labels of multiple sources that have different, unknown accuracies, may be correlated, and may label at different levels of granularity. In our experience with users in academia and industry, the complexity of real world weak supervision sources makes this integration phase the key time sink and stumbling block. For example, if we are training a model to classify entities in text, we may have one available source of high-quality but coarse-grained labels (e.g. “Person” vs. “Organization”) and one source that provides lower-quality but finer-grained labels (e.g. “Doctor” vs. “Lawyer”); moreover, these sources might be correlated due to some shared component or data source (Bach et al. 2017; Varma et al. 2017). Handling such diversity requires addressing a core technical challenge: estimating the unknown accuracies of multi-granular and potentially correlated supervision sources without any labeled data.

To overcome this challenge, we propose `MeTaL`, a framework for modeling and integrating weak supervision sources with different unknown accuracies, correlations, and granularities. In `MeTaL`, we view each source as labeling one of several related sub-tasks of a problem—we refer to this as the *multi-task weak supervision* setting. We then show that given the dependency structure of the sources, we can use their observed agreement and disagreement rates to recover their unknown accuracies. Moreover, we exploit the relationship structure between tasks to observe additional cross-task agreements and disagreements, effectively providing extra signal from which to learn. In contrast to previous approaches based on sampling from the posterior of a graphical model directly (Ratner et al. 2016; Bach et al. 2017), we develop a simple and scalable matrix completion-style algorithm, which we are able to analyze by applying strong matrix concentration bounds (Tropp 2015). We use this algorithm to learn and model the accuracies of diverse weak supervision sources, and then combine their labels to produce training data that can be used to supervise arbitrary models, including increasingly popular multi-task learning models (Caruana 1993; Ruder 2017).

Compared to previous methods which only handled the single-task setting (Ratner et al. 2016; 2018), and generally considered conditionally-independent sources (Anandkumar et al. 2014; Dawid and Skene 1979), we demonstrate that our multi-task aware approach leads to average gains of 4.1 points in accuracy in our experiments, and has at least three additional benefits. First, many dependency structures between weak supervision sources may lead to non-identifiable models of their accuracies, where a unique solution cannot be recovered. We provide a compiler-like check to establish identifiability—i.e. the existence of a unique set of source accuracies—for arbitrary dependency structures, without resorting to the standard assumption of non-adversarial sources (Dawid and Skene 1979), alerting users to this potential stumbling block that we have observed in practice. Next, we provide sample complexity bounds that characterize the benefit of adding additional unlabeled data and the scaling with respect to the user-specified task and dependency structure. While previous approaches required thousands of sources to give non-vacuous bounds, we capture regimes with small numbers of sources, better reflecting the real-world uses of weak supervision we have observed. Finally, we are able to solve our proposed problem directly with SGD, leading to over $100\times$ faster runtimes compared to prior Gibbs-sampling based approaches (Ratner et al. 2016; Platanios et al. 2017), and enabling simple implementation using libraries like PyTorch.

We validate our framework on three fine-grained classification tasks in named entity recognition, relation extraction, and medical document classification, for which we have diverse weak supervision sources at multiple levels of granularity. We show that by modeling them as labeling hierarchically-related sub-tasks and utilizing unlabeled data, we can get an average improvement of 20.2 points in accuracy over a traditional supervised approach, 6.8 points over a basic majority voting weak supervision baseline, and 4.1 points over data programming (Ratner et al. 2016), an existing weak supervision approach in the literature that is not multi-task-aware. We also extend our framework to handle unipolar sources that only label one class, a critical aspect of weak supervision in practice that leads to an average 2.8 point contribution to our gains over majority vote. From a practical standpoint, we argue that our framework represents an efficient way for practitioners to supervise modern machine learning models, including new multi-task variants, for complex tasks by opportunistically using the diverse weak supervision sources available to them. To further validate this, we have released an open-source implementation of our framework.[1]

2 Related Work

Our work builds on and extends various settings studied in machine learning.

Weak Supervision: We draw motivation from recent work which models and integrates weak supervision using generative models (Ratner et al. 2016; 2018; Bach et al. 2017) and other methods (Guan et al. 2017; Khetan, Lipton, and Anandkumar 2017). These approaches, however, do not handle multi-granularity or multi-task weak supervision, require expensive sampling-based techniques that may lead to non-identifiable solutions, and leave room for sharper theoretical characterization of weak supervision scaling properties. More generally, our work is motivated by a wide range of specific weak supervision techniques, which include traditional distant supervision approaches (Mintz et al. 2009; Craven and Kumlien 1999; Zhang et al. 2017a; Hoffmann et al. 2011; Takamatsu, Sato, and Nakagawa 2012), co-training methods (Blum and Mitchell 1998), pattern-based supervision (Gupta and Manning 2014; Zhang et al. 2017a), and feature-annotation techniques (Mann and McCallum 2010; Zaidan and Eisner 2008; Liang, Jordan, and Klein 2009).

Crowdsourcing: Our approach also has connections to the crowdsourcing literature (Karger, Oh, and Shah 2011; Dawid and Skene 1979), and in particular to spectral and method of moments-based approaches (Zhang et al. 2014; Dalvi et al. 2013a; Ghosh, Kale, and McAfee 2011; Anandkumar et al. 2014). In contrast, the goal of our work is to support and explore settings not covered by crowdsourcing work, such as sources with correlated outputs, the proposed multi-task supervision setting, and regimes wherein a small number of labelers (weak supervision sources) each label a large number of items (data points). Moreover, we theoretically characterize the generalization performance of an end model trained with the weakly labeled data.

Multi-Task Learning: Our proposed approach is motivated by recent progress on multi-task learning models (Caruana 1993; Ruder 2017; Søgaard and Goldberg 2016), in particular their need for multiple large hand-labeled training datasets. We note that the focus of our paper is on generating supervision for these models, not on the particular multi-task learning model being trained, which we seek to control for by fixing a simple architecture in our experiments.

Our work is also related to recent techniques for estimating classifier accuracies without labeled data in the presence of structural constraints (Platanios et al. 2017). We use matrix structure estimation (Loh and Wainwright 2012) and concentration bounds (Tropp 2015) for our core results.

3 Programming Machine Learning with Weak Supervision

As modern machine learning models become both more complex and more performant on a range of tasks, developers increasingly interact with them by programmatically generating noisier or *weak* supervision. These approaches of effectively *programming* machine learning models have recently been formalized by the following pipeline (Ratner et al. 2016; 2018): First, users provide one or more *weak supervision sources*, which are applied to unlabeled data to generate a set of noisy labels. These labels may overlap and conflict; we model and combine them via a *label model* in order to produce a final set of training labels. These labels are then used to train some discriminative model, which we refer to as the *end model*. This programmatic weak supervision approach can utilize sources ranging from heuristic

[1] github.com/HazyResearch/metal

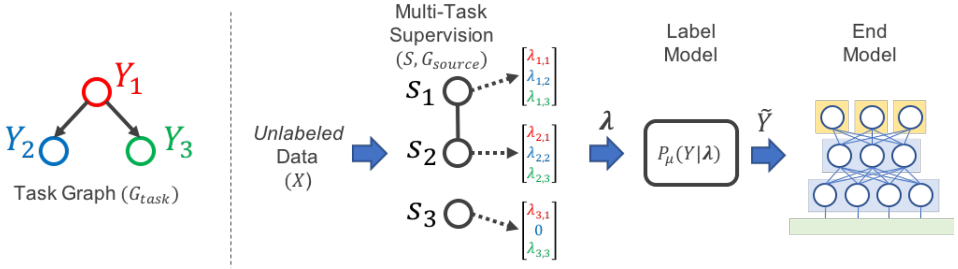
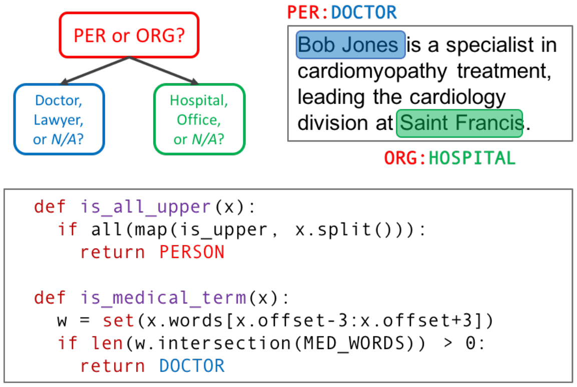

Y_1
Y_2
Y_3
Task Graph (G_{task})
Unlabeled Data (X)
Multi-Task Supervision (S, G_{source})
s_1
s_2
s_3
$\lambda_{1,1}$ $\lambda_{1,2}$ $\lambda_{1,3}$
$\lambda_{2,1}$ $\lambda_{2,2}$ $\lambda_{2,3}$
$\lambda_{3,1}$ 0 $\lambda_{3,3}$
$\boldsymbol{\lambda}$
Label Model
$P_\mu(Y|\boldsymbol{\lambda})$
$\tilde{Y}$
End Model

Figure 1: A schematic of the `MeTaL` pipeline. To generate training data for an *end model*, such as a multi-task model as in our experiments, the user inputs a *task graph* G_{task} defining the relationships between *task labels* $Y_1, ..., Y_t$; a set of *unlabeled* data points X; a set of *multi-task weak supervision sources* s_i which each output a vector $\boldsymbol{\lambda}_i$ of task labels for X; and the dependency structure between these sources, G_{source}. We train a *label model* to learn the accuracies of the sources, outputting a vector of probabilistic training labels $\tilde{\mathbf{Y}}$ for training the end model.



Bob Jones is a specialist in cardiomyopathy treatment, leading the cardiology division at Saint Francis.

```
def is_all_upper(x):
  if all(map(is_upper, x.split())):
    return PERSON

def is_medical_term(x):
  w = set(x.words[x.offset-3:x.offset+3])
  if len(w.intersection(MED_WORDS)) > 0:
    return DOCTOR
```

Figure 2: An example fine-grained entity classification problem, where weak supervision sources label three sub-tasks of different granularities: (i) `Person` vs. `Organization`, (ii) `Doctor` vs. `Lawyer` (or *N/A*), (iii) `Hospital` vs. `Office` (or *N/A*). The example weak supervision sources use a pattern heuristic and dictionary lookup respectively.

rules to other models, and in this way can also be viewed as a pragmatic and flexible form of multi-source *transfer learning*.

In our experiences with users from science and industry, we have found it critical to utilize all available sources of weak supervision for complex modeling problems, including ones which label at multiple levels of *granularity*. However, this diverse, multi-granular weak supervision does not easily fit into existing paradigms. We propose a formulation where each weak supervision source labels some sub-task of a problem, which we refer to as the *multi-task weak supervision* setting. We consider an example:

Example 1 *A developer wants to train a fine-grained Named Entity Recognition (NER) model to classify mentions of entities in the news (Figure 2). She has a multitude of available weak supervision sources which she believes have relevant signal for her problem—for example, pattern matchers, dictionaries, and pre-trained generic NER taggers. However, it is unclear how to properly use and combine them: some of them label phrases coarsely as* `PERSON` *versus* `ORGANIZATION`*, while others classify specific fine-grained types of people or organizations, with a range of unknown accuracies. In our framework, she can represent them as labeling tasks of different granularities—e.g.* $Y_1 = \{\texttt{Person}, \texttt{Org}\}$, $Y_2 = \{\texttt{Doctor}, \texttt{Lawyer}, \textit{N/A}\}$, $Y_3 = \{\texttt{Hospital}, \texttt{Office}, \textit{N/A}\}$*, where the label N/A applies, for example, when the type-of-person task is applied to an organization.*

In our proposed multi-task supervision setting, the user specifies a set of structurally-related *tasks*, and then provides a set of *weak supervision sources* which are user-defined functions that either label each data point or abstain for each task, and may have some user-specified dependency structure. These sources can be arbitrary black-box functions, and can thus subsume a range of weak supervision approaches relevant to both text and other data modalities, including use of pattern-based heuristics, distant supervision (Mintz et al. 2009), crowd labels, other weak or biased classifiers, declarative rules over unsupervised feature extractors (Varma et al. 2017), and more. Our goal is to estimate the unknown accuracies of these sources, combine their outputs, and use the resulting labels to train an end model.

4 Modeling Multi-Task Weak Supervision

The core technical challenge of the *multi-task weak supervision* setting is recovering the unknown *accuracies* of weak supervision sources given their dependency structure and a schema of the tasks they label, but without any ground-truth labeled data. We define a new algorithm for recovering the accuracies in this setting using a matrix completion-style optimization objective. We establish conditions under which the resulting estimator returns a unique solution. We then analyze the sample complexity of our estimator, characterizing its scaling with respect to the amount of *unlabeled data*, as well as the task schema and dependency structure, and show how the estimation error affects the generalization performance of the end model we aim to train. Finally, we highlight how our approach handles abstentions and *unipolar* sources, two critical scenarios in the weak supervision setting.

A Multi-Task Weak Supervision Estimator

Problem Setup Let $X \in \mathcal{X}$ be a data point and $\mathbf{Y} = [Y_1, Y_2, \ldots, Y_t]^T$ be a vector of categorical *task labels*, $Y_i \in \{1, \ldots, k_i\}$, corresponding to t tasks, where $(X, \mathbf{Y})$ is drawn i.i.d. from a distribution $\mathcal{D}$.[2]

The user provides a specification of how these tasks relate to each other; we denote this schema as the *task structure* G_{task}. The task structure expresses logical relationships between tasks, defining a *feasible set* of label vectors $\mathcal{Y}$, such that $\mathbf{Y} \in \mathcal{Y}$. For example, Figure 2 illustrates a hierarchical task structure over three tasks of different granularities pertaining to a fine-grained entity classification problem. Here, the tasks are related by logical subsumption relationships: for example, if $Y_2 = \texttt{DOCTOR}$, this implies that $Y_1 = \texttt{PERSON}$, and that $Y_3 = \textit{N/A}$, since the task label Y_3 concerns types of organizations, which is inapplicable to persons. Thus, in this task structure, $\mathbf{Y} = [\texttt{PERSON}, \texttt{DOCTOR}, \textit{N/A}]^T$ is in $\mathcal{Y}$ while $\mathbf{Y} = [\texttt{PERSON}, \textit{N/A}, \texttt{HOSPITAL}]^T$ is not. While task structures are often simple to define, as in the previous example, or are explicitly defined by existing resources—such as ontologies or graphs—we note that if no task structure is provided, our approach becomes equivalent to modeling the t tasks separately, a baseline we consider in the experiments.

In our setting, rather than observing the true label $\mathbf{Y}$, we have access to m *multi-task weak supervision* sources $s_i \in S$ which emit label vectors $\boldsymbol{\lambda}_i$ that contain labels for some subset of the t tasks. Let 0 denote a null or abstaining label, and let the *coverage set* $\tau_i \subseteq \{1, \ldots, t\}$ be the fixed set of tasks for which the ith source emits non-zero labels, such that $\boldsymbol{\lambda}_i \in \mathcal{Y}_{\tau_i}$. For convenience, we let $\tau_0 = \{1, \ldots, t\}$ so that $\mathcal{Y}_{\tau_0} = \mathcal{Y}$. For example, a source from our previous example might have a coverage set $\tau_i = \{1, 3\}$, emitting coarse-grained labels such as $\boldsymbol{\lambda}_i = [\texttt{PERSON}, 0, \textit{N/A}]^T$. Note that sources often label multiple tasks implicitly due to the constraints of the task structure; for example, a source that labels types of people (Y_2) also implicitly labels people vs. organizations ($Y_1 = \texttt{PERSON}$), and types of organizations (as $Y_3 = \textit{N/A}$). Thus sources tailored to different tasks still have agreements and disagreements; we use this additional *cross-task* signal in our approach.

The user also provides the conditional dependency structure of the sources as a graph $G_{\text{source}} = (V, E)$, where $V = \{\mathbf{Y}, \boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \ldots, \boldsymbol{\lambda}_m\}$ (Figure 3). Specifically, if $(\boldsymbol{\lambda}_i, \boldsymbol{\lambda}_j)$ is not an edge in G_{source}, this means that $\boldsymbol{\lambda}_i$ is independent of $\boldsymbol{\lambda}_j$ conditioned on $\mathbf{Y}$ and the other source labels. Note that if G_{source} is unknown, it can be estimated using statistical techniques such as (Bach et al. 2017). Importantly, we do not know anything about the strengths of the correlations in G_{source}, or the sources' accuracies.

Our overall goal is to apply the set of weak supervision sources $S = \{s_1, \ldots, s_m\}$ to an unlabeled dataset $\mathcal{X}_U$ consisting of n data points, then use the resulting weakly-labeled training set to supervise an *end model* $f_w : \mathcal{X} \mapsto \mathcal{Y}$ (Figure 1). This weakly-labeled training set will contain overlapping and conflicting labels, from sources with unknown accuracies and correlations. To handle this, we will learn a *label model* $P_\mu(\mathbf{Y}|\boldsymbol{\lambda})$, parameterized by a vector of source correlations and accuracies μ, which for each data point X takes as input the noisy labels $\boldsymbol{\lambda} = \{\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_m\}$ and outputs a single probabilistic label vector $\tilde{\mathbf{Y}}$. Succinctly, given a user-provided tuple $(\mathcal{X}_U, S, G_{\text{source}}, G_{\text{task}})$, our key technical challenge is recovering the parameters μ without access to ground truth labels $\mathbf{Y}$.

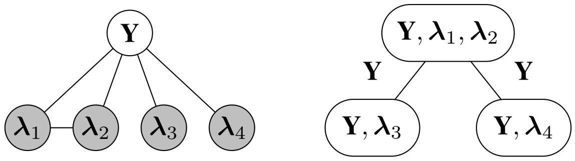


Figure 3: An example of a weak supervision source dependency graph G_{source} (left) and its junction tree representation (right), where $\mathbf{Y}$ is a vector-valued random variable with a feasible set of values, $\mathbf{Y} \in \mathcal{Y}$. Here, the output of sources 1 and 2 are modeled as dependent conditioned on $\mathbf{Y}$. This results in a junction tree with singleton separator sets, $\mathbf{Y}$. Here, the observable cliques are $O = \{\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \boldsymbol{\lambda}_3, \boldsymbol{\lambda}_4, \{\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2\}\} \subset \mathcal{C}$.

Modeling Multi-Task Sources To learn a label model over multi-task sources, we introduce sufficient statistics over the random variables in G_{source}. Let $\mathcal{C}$ be the set of cliques in G_{source}, and define an indicator random variable for the event of a clique $C \in \mathcal{C}$ taking on a set of values y_C:

$$\psi(C, y_C) = \mathbb{1}\left\{\cap_{i \in C} V_i = (y_C)_i\right\},$$

where $(y_C)_i \in \mathcal{Y}_{\tau_i}$. We define $\psi(C) \in \{0, 1\}^{\prod_{i \in C}(|\mathcal{Y}_{\tau_i}| - 1)}$ as the vector of indicator random variables for all combinations of all but one of the labels emitted by each variable in clique C—thereby defining a minimal set of statistics—and define $\psi(\mathbf{C})$ accordingly for any set of cliques $\mathbf{C} \subseteq \mathcal{C}$. Then $\mu = \mathbb{E}\left[\psi(\mathcal{C})\right]$ is the vector of sufficient statistics for the label model we want to learn.

We work with two simplifying conditions in this section. First, we consider the setting where G_{source} is *triangulated* and has a junction tree representation with singleton separator sets. If this is not the case, edges can always be added to G_{source} to make this setting hold; otherwise, we describe how our approach can directly handle non-singleton separator sets in the Appendix.

Second, we use a simplified *class-conditional* model of the noisy labeling process, where we learn one accuracy parameter for each label value $\boldsymbol{\lambda}_i$ that each source s_i emits. This is equivalent to assuming that a source may have a different accuracy on each different class, but that if it emits a certain label incorrectly, it does so uniformly over the different true labels $\mathbf{Y}$. This is a more expressive model than the commonly considered one, where each source is modeled by a single accuracy parameter, e.g. in (Dawid and Skene 1979; Ratner et al. 2016), and in particular allows us to capture the *unipolar* setting considered later on.

[2] The variables we introduce throughout this section are summarized in a glossary in the Appendix, which can be accessed at https://arxiv.org/abs/1810.02840.

Our Approach The chief technical difficulty in our problem is that we do not observe $\mathbf{Y}$. We overcome this by analyzing the covariance matrix of an observable subset of the cliques in G_{source}, leading to a matrix completion-style approach for recovering μ. We leverage two pieces of information: (i) the observability of *part of* $\textbf{Cov}\left[\psi(\mathcal{C})\right]$, and (ii) a result from (Loh and Wainwright 2012) which states that the inverse covariance matrix $\textbf{Cov}\left[\psi(\mathcal{C})\right]^{-1}$ is structured according to G_{source}, i.e., if there is no edge between $\boldsymbol{\lambda}_i$ and $\boldsymbol{\lambda}_j$ in G_{source}, then the corresponding entries are 0.

We start by considering two disjoint subsets of $\mathcal{C}$: the set of observable cliques, $O \subseteq \mathcal{C}$—i.e., those cliques not containing $\mathbf{Y}$—and the separator set cliques of the junction tree, $\mathcal{S} \subseteq \mathcal{C}$. In the setting we consider here, $\mathcal{S} = \{\mathbf{Y}\}$ (see Figure 3). We then write the covariance matrix of the indicator variables for $O \cup \mathcal{S}$, $\textbf{Cov}\left[\psi(O \cup \mathcal{S})\right]$, in block form, similar to (Chandrasekaran, Parrilo, and Willsky 2010), as:

$$\textbf{Cov}\left[\psi(O \cup \mathcal{S})\right] \equiv \Sigma = \begin{bmatrix} \Sigma_O & \Sigma_{O\mathcal{S}} \\ \Sigma_{O\mathcal{S}}^T & \Sigma_{\mathcal{S}} \end{bmatrix} \qquad (1)$$

and similarly define its inverse:

$$K = \Sigma^{-1} = \begin{bmatrix} K_O & K_{O\mathcal{S}} \\ K_{O\mathcal{S}}^T & K_{\mathcal{S}} \end{bmatrix} \qquad (2)$$

Here, $\Sigma_O = \textbf{Cov}\left[\psi(O)\right] \in \mathbb{R}^{d_O \times d_O}$ is the observable block of Σ, where $d_O = \sum_{C \in O} \prod_{i \in C}(|\mathcal{Y}_{\tau_i}| - 1)$. Next, $\Sigma_{O\mathcal{S}} = \textbf{Cov}\left[\psi(O), \psi(\mathcal{S})\right]$ is the unobserved block which is a function of μ, the label model parameters that we wish to recover. Finally, $\Sigma_{\mathcal{S}} = \textbf{Cov}\left[\psi(\mathcal{S})\right] = \textbf{Cov}\left[\psi(\mathbf{Y})\right]$ is a function of the class balance $P(\mathbf{Y})$.

We make two observations about $\Sigma_{\mathcal{S}}$. First, while the full form of $\Sigma_{\mathcal{S}}$ is the covariance of the $|\mathcal{Y}| - 1$ indicator variables for each individual value of $\mathbf{Y}$ but one, given our simplified class-conditional label model, we in fact only need a single indicator variable for $\mathbf{Y}$ (see Appendix); thus, $\Sigma_{\mathcal{S}}$ is a scalar. Second, $\Sigma_{\mathcal{S}}$ is a function of the class balance $P(\mathbf{Y})$, which we assume is either known, or has been estimated according to the unsupervised approach we detail in the Appendix. Thus, given Σ_O and $\Sigma_{\mathcal{S}}$, our goal is to recover the vector $\Sigma_{O\mathcal{S}}$ from which we can recover μ.

Applying the block matrix inversion lemma, we have:

$$K_O = \Sigma_O^{-1} + c\Sigma_O^{-1}\Sigma_{O\mathcal{S}}\Sigma_{O\mathcal{S}}^T\Sigma_O^{-1}, \qquad (3)$$

where $c = \left(\Sigma_{\mathcal{S}} - \Sigma_{O\mathcal{S}}^T\Sigma_O^{-1}\Sigma_{O\mathcal{S}}\right)^{-1} \in \mathbb{R}^+$. Let $z = \sqrt{c}\Sigma_O^{-1}\Sigma_{O\mathcal{S}}$; we can then express (3) as:

$$K_O = \Sigma_O^{-1} + zz^T \qquad (4)$$

The right hand side of (4) consists of an empirically observable term, Σ_O^{-1}, and a rank-one term, zz^T, which we can solve for to directly recover μ. For the left hand side, we apply an extension of Corollary 1 from (Loh and Wainwright 2012) (see Appendix) to conclude that K_O has graph-structured sparsity, i.e., it has zeros determined by the structure of dependencies between the sources in G_{source}. This suggests an algorithmic approach of estimating z as a matrix completion problem in order to recover an estimate of μ (Algorithm 1). In more detail: let Ω be the set of indices (i, j) where $(K_O)_{i,j} = 0$, determined by G_{source}, yielding a system of equations,

$$0 = (\Sigma_O^{-1})_{i,j} + \left(zz^T\right)_{i,j} \text{ for } (i,j) \in \Omega, \qquad (5)$$

which is now a matrix completion problem. Define $||A||_\Omega$ as the Frobenius norm of A with entries not in Ω set to zero; then we can rewrite (5) as $\left|\left|\Sigma_O^{-1} + zz^T\right|\right|_\Omega = 0$. We solve this equation to estimate z, and thereby recover $\Sigma_{O\mathcal{S}}$, from which we can directly recover the label model parameters μ algebraically.

Checking for Identifiability A first question is: which dependency structures G_{source} lead to unique solutions for μ? This question presents a stumbling block for users, who might attempt to use non-identifiable sets of correlated weak supervision sources.

We provide a simple, testable condition for identifiability. Let G_{inv} be the inverse graph of G_{source}; note that Ω is the edge set of G_{inv} expanded to include all indicator random variables $\psi(\mathcal{C})$. Then, let M_Ω be a matrix with dimensions $|\Omega| \times d_O$ such that each row in M_Ω corresponds to a pair $(i, j) \in \Omega$ with 1's in positions i and j and 0's elsewhere.

Taking the log of the squared entries of (5), we get a system of linear equations $M_\Omega l = q_\Omega$, where $l_i = \log(z_i^2)$ and $q_{(i,j)} = \log(((\Sigma_O^{-1})_{i,j})^2)$. Assuming we can solve this system (which we can always ensure by adding sources; see Appendix), we can uniquely recover the z_i^2, meaning our model is identifiable *up to sign*.

Given estimates of the z_i^2, we can see from (5) that the sign of a single z_i determines the sign of all other z_j reachable from z_i in G_{inv}. Thus to ensure a unique solution, we only need to pick a sign for each connected component in G_{inv}. In the case where the sources are assumed to be independent, e.g., (Dalvi et al. 2013b; Zhang et al. 2014; Dawid and Skene 1979), it suffices to make the assumption that the sources are *on average* non-adversarial; i.e., select the sign of the z_i that leads to higher average accuracies of the sources. Even a single source that is conditionally independent from all the other sources will cause G_{inv} to be fully connected, meaning we can use this symmetry breaking assumption in the majority of cases even with correlated sources. Otherwise, a sufficient condition is the standard one of assuming non-adversarial sources, i.e. that all sources have greater than random accuracy.

Source Accuracy Estimation Algorithm Now that we know when a set of sources with correlation structure G_{source} is identifiable, yielding a unique z, we can estimate the accuracies μ using Algorithm 1. We also use the function ExpandTied, which is a simple algebraic expansion of tied parameters according to the simplified class-conditional model used in this section; see Appendix for details. In Figure 4, we plot the performance of our algorithm on synthetic data, showing its scaling with the number of unlabeled data points n, the density of pairwise dependencies in G_{source}, and the runtime performance as compared to a prior Gibbs sampling-based approach. Next, we theoretically analyze the scaling of the error $||\hat{\mu} - \mu^*||$.

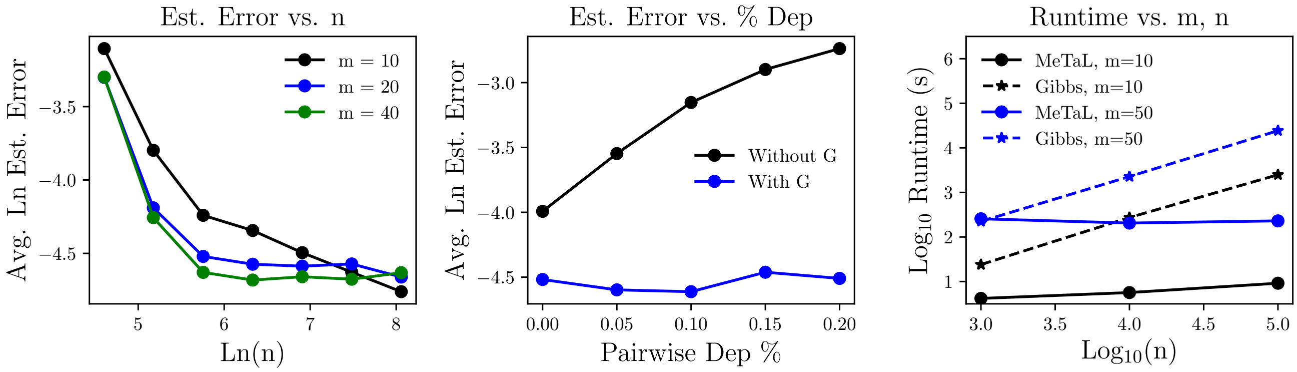


Figure 4: (Left) Estimation error $||\hat{\mu} - \mu^*||$ decreases with increasing n. (Middle) Given G_{source}, our model successfully recovers the source accuracies even with many pairwise dependencies among sources, where a naive conditionally-independent model fails. (Right) The runtime of MeTaL is independent of n after an initial matrix multiply, and can thus be multiple orders of magnitude faster than Gibbs sampling-based approaches (Ratner et al. 2016).

Algorithm 1 Source Accuracy Estimation for Multi-Task Weak Supervision

```
    Input: Observed labeling rates $\hat{\mathbb{E}}[\psi(O)]$ and covariance $\hat{\Sigma}_O$; class balance $\hat{\mathbb{E}}[\psi(\mathbf{Y})]$ and variance $\Sigma_{\mathcal{S}}$; correlation sparsity structure $\Omega$
    $\hat{z} \leftarrow \text{argmin}_z \left|\left|\hat{\Sigma}_O^{-1} + zz^T\right|\right|_\Omega$
    $\hat{c} \leftarrow \Sigma_{\mathcal{S}}^{-1}(1 + \hat{z}^T \hat{\Sigma}_O \hat{z}),\ \hat{\Sigma}_{O\mathcal{S}} \leftarrow \hat{\Sigma}_O \hat{z}/\sqrt{\hat{c}}$
    $\hat{\mu}' \leftarrow \hat{\Sigma}_{O\mathcal{S}} + \hat{\mathbb{E}}[\psi(\mathbf{Y})]\,\hat{\mathbb{E}}[\psi(O)]$
    return ExpandTied$(\hat{\mu}')$
```

Theoretical Analysis: Scaling with Diverse Multi-Task Supervision

Our ultimate goal is to train an *end model* using the source labels, denoised and combined by the label model $\hat{\mu}$ we have estimated. We connect the generalization error of this end model to the estimation error of Algorithm 1, ultimately showing that the generalization error scales as $n^{-\frac{1}{2}}$, where n is the number of unlabeled data points. This key result establishes the same asymptotic scaling as traditionally supervised learning methods, but with respect to *unlabeled* data points.

Let $P_{\hat{\mu}}(\tilde{\mathbf{Y}} \mid \boldsymbol{\lambda})$ be the probabilistic label (i.e. distribution) predicted by our label model, given the source labels $\boldsymbol{\lambda}$ as input, which we compute using the estimated $\hat{\mu}$. We then train an *end* multi-task discriminative model $f_w : \mathcal{X} \mapsto \mathcal{Y}$ parameterized by w, by minimizing the expected loss with respect to the label model over n unlabeled data points. Let $l(w, X, \mathbf{Y}) = \frac{1}{t}\sum_{s=1}^{t} l_t(w, X, \mathbf{Y}_s)$ be a bounded multi-task loss function such that without loss of generality $l(w, X, \mathbf{Y}) \leq 1$; then we minimize the empirical *noise aware loss*:

$$\hat{w} = \text{argmin}_w \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\tilde{\mathbf{Y}} \sim P_{\hat{\mu}}(\cdot|\boldsymbol{\lambda})}\left[l(w, X_i, \tilde{\mathbf{Y}})\right], \qquad (6)$$

and let $\tilde{w}$ be the w that minimizes the true noise-aware loss. This minimization can be performed by standard methods and is not the focus of our paper; let the solution $\hat{w}$ satisfy $\mathbb{E}\left[\|\hat{w} - \tilde{w}\|^2\right] \leq \gamma$. We make several assumptions, following (Ratner et al. 2016): (1) that for some label model parameters μ^*, sampling $(\boldsymbol{\lambda}, \mathbf{Y}) \sim P_{\mu^*}(\cdot)$ is the same as sampling from the true distribution, $(\boldsymbol{\lambda}, \mathbf{Y}) \sim \mathcal{D}$; and (2) that the task labels Y_s are independent of the features of the end model given $\boldsymbol{\lambda}$ sampled from $P_{\mu^*}(\cdot)$, that is, the output of the optimal label model provides sufficient information to discern the true label. Then we have the following result:

Theorem 1 *Let $\tilde{w}$ minimize the expected noise aware loss, using weak supervision source parameters $\hat{\mu}$ estimated with Algorithm 1. Let $\hat{w}$ minimize the empirical noise aware loss with $\mathbb{E}\left[\|\hat{w} - \tilde{w}\|^2\right] \leq \gamma$, $w^* = \min_w l(w, X, \mathbf{Y})$, and let the assumptions above hold. Then the generalization error is bounded by:*

$$\mathbb{E}\left[l(\hat{w}, X, \mathbf{Y}) - l(w^*, X, \mathbf{Y})\right] \leq \gamma + 4|\mathcal{Y}|\,||\hat{\mu} - \mu^*||\,.$$

Thus, to control the generalization error, we must control $||\hat{\mu} - \mu^*||$, which we do in Theorem 2:

Theorem 2 *Let $\hat{\mu}$ be an estimate of μ^* produced by Algorithm 1 run over n unlabeled data points. Let $a := \left(\frac{d_O}{\Sigma_{\mathcal{S}}} + \left(\frac{d_O}{\Sigma_{\mathcal{S}}}\right)^2 \lambda_{max}(K_O)\right)^{\frac{1}{2}}$ and $b := \frac{\|\Sigma_O^{-1}\|^2}{(\Sigma_O^{-1})_{\min}}$. Then, we have:*

$$\mathbb{E}\left[||\hat{\mu} - \mu^*||\right] \leq 16(|\mathcal{Y}| - 1)d_O^2 \sqrt{\frac{32\pi}{n}}\, ab\sigma_{\max}(M_\Omega^+)$$
$$\times \left(3\sqrt{d_O}\, a\lambda_{\min}^{-1}(\Sigma_O) + 1\right)\left(\kappa(\Sigma_O) + \lambda_{\min}^{-1}(\Sigma_O)\right).$$

Interpreting the Bound We briefly explain the key terms controlling the bound in Theorem 2; more detail is found in the Appendix. Our primary result is that the estimation error scales as $n^{-\frac{1}{2}}$. Next, $\sigma_{\max}(M_\Omega^+)$, the largest singular value of the pseudoinverse M_Ω^+, has a deep connection to the density of the graph G_{inv}. The smaller this quantity, the more information we have about G_{inv}, and the easier it is to estimate the accuracies. Next, $\lambda_{\min}(\Sigma_O)$, the smallest eigenvalue of the observed covariance matrix, reflects the conditioning of Σ_O; better conditioning yields easier estimation, and is roughly determined by how far away from random guessing the worst weak supervision source is, as well as how conditionally independent the sources are. $\lambda_{\max}(K_O)$, the largest eigenvalue of the upper-left block of the inverse covariance matrix, similarly reflects the overall conditioning of Σ. Finally, $(\Sigma_O^{-1})_{\min}$, the smallest entry of the inverse observed matrix, reflects the smallest non-zero correlation between source accuracies; distinguishing between small correlations and independent sources requires more samples.

Extensions: Abstentions & Unipolar Sources

We briefly highlight two extensions handled by our approach which we have found empirically critical: handling *abstentions*, and modeling *unipolar* sources.

Handling Abstentions. One fundamental aspect of the weak supervision setting is that sources may abstain from labeling a data point entirely—that is, they may have incomplete and differing coverage (Ratner et al. 2018; Dalvi et al. 2013b). We can easily deal with this case by extending the coverage ranges $\mathcal{Y}_{\tau_i}$ of the sources to include the vector of all zeros, $\vec{0}$, and we do so in the experiments.

Handling Unipolar Sources. Finally, we highlight the fact that our approach models *class conditional* source accuracies, in particular motivated by the case we have frequently observed in practice of *unipolar* weak supervision sources, i.e., sources that each only label a single class or abstain. In practice, we find that users most commonly use such unipolar sources; for example, a common template for a heuristic-based weak supervision source over text is one that looks for a specific pattern, and if the pattern is present emits a specific label, else abstains. As compared to prior approaches that did not model class-conditional accuracies, e.g. (Ratner et al. 2016), we show in our experiments that we can use our class-conditional modeling approach to yield an improvement of 2.8 points in accuracy.

5 Experiments

We validate our approach on three fine-grained classification problems—entity classification, relation classification, and document classification—where weak supervision sources are available at both coarser and finer-grained levels (e.g. as in Figure 2). We evaluate the predictive accuracy of end models supervised with training data produced by several approaches, finding that our approach outperforms traditional hand-labeled supervision by 20.2 points, a baseline majority vote weak supervision approach by 6.8 points, and a prior weak supervision denoising approach (Ratner et al. 2016) that is not multi-task-aware by 4.1 points.

Datasets Each dataset consists of a large (3k-63k) amount of unlabeled training data and a small (200-350) amount of labeled data which we refer to as the *development set*, which we use for (a) a traditional supervision baseline, and (b) for hyperparameter tuning of the end model (see Appendix). The average number of weak supervision sources per task was 13, with sources expressed as Python functions, averaging 4 lines of code and comprising a mix of pattern matching heuristics, external knowledge base or dictionary lookups, and pre-trained models. In all three cases, we choose the decomposition into sub-tasks so as to align with weak supervision sources that are either available or natural to express.

Named Entity Recognition (NER): We represent a fine-grained named entity recognition problem—tagging entity mentions in text documents—as a hierarchy of three sub-tasks over the OntoNotes dataset (Weischedel et al. 2011): $Y_1 \in \{\text{Person}, \text{Organization}\}$, $Y_2 \in \{\text{Businessperson}, \text{Other Person}, \textit{N/A}\}$, $Y_3 \in \{\text{Company}, \text{Other Org}, \textit{N/A}\}$, where again we use *N/A* to represent "not applicable".

Relation Extraction (RE): We represent a relation extraction problem—classifying entity-entity relation mentions in text documents—as a hierarchy of six sub-tasks which either concern labeling the subject, object, or subject-object pair of a possible or *candidate* relation in the TACRED dataset (Zhang et al. 2017b). For example, we might label a relation as having a Person subject, Location object, and Place-of-Residence relation type.

Medical Document Classification (Doc): We represent a radiology report triaging (i.e. document classification) problem from the OpenI dataset (National Institutes of Health 2017) as a hierarchy of three sub-tasks: $Y_1 \in \{\text{Acute}, \text{Non-Acute}\}$, $Y_2 \in \{\text{Urgent}, \text{Emergent}, \textit{N/A}\}$, $Y_3 \in \{\text{Normal}, \text{Non-Urgent}, \textit{N/A}\}$.

End Model Protocol Our goal was to test the performance of a basic multi-task end model using training labels produced by various different approaches. We use an architecture consisting of a shared bidirectional LSTM input layer with pre-trained embeddings, shared linear intermediate layers, and a separate final linear layer ("task head") for each task. Hyperparameters were selected with an initial search for each application (see Appendix), then fixed.

Core Validation We compare the accuracy of the end multi-task model trained with labels from our approach versus those from three baseline approaches (Table 1):

- *Traditional Supervision* **[Gold (Dev)]**: We train the end model using the small hand-labeled development set.
- *Hierarchical Majority Vote* **[MV]**: We use a hierarchical majority vote of the weak supervision source labels: i.e. for each data point, for each task we take the majority vote and proceed down the task tree accordingly. This procedure can be thought of as a hard decision tree, or a cascade of if-then statements as in a rule-based approach.
- *Data Programming* **[DP]**: We model each task separately using the data programming approach for denoising weak supervision (Ratner et al. 2018).

In all settings, we used the same end model architecture as described above. Note that while we choose to model these problems as consisting of multiple sub-tasks, we evaluate with respect to the broad primary task of fine-grained

	NER	RE	Doc	Average
Gold (Dev)	63.7 ± 2.1	28.4 ± 2.3	62.7 ± 4.5	51.6
MV	76.9 ± 2.6	43.9 ± 2.6	74.2 ± 1.2	65.0
DP (Ratner et al. 2016)	78.4 ± 1.2	49.0 ± 2.7	75.8 ± 0.9	67.7
`MeTaL`	**82.2** ± 0.8	**56.7** ± 2.1	**76.6** ± 0.4	**71.8**

Table 1: **Performance Comparison of Different Supervision Approaches.** We compare the micro accuracy (avg. over 10 trials) with 95% confidence intervals of an end multi-task model trained using the training labels from the hand-labeled development set (Gold Dev), hierarchical majority vote (MV), data programming (DP), and our approach (`MeTaL`).

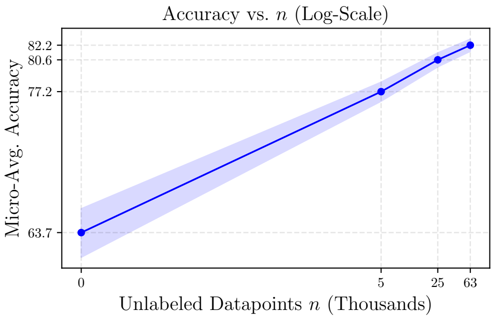


Figure 5: In the OntoNotes dataset, end model accuracy scales with the amount of available *unlabeled* data.

classification (for subtask-specific scores, see Appendix). We observe in Table 1 that our approach of leveraging multi-granularity weak supervision leads to large gains—20.2 points over traditional supervision with the development set, 6.8 points over hierarchical majority vote, and 4.1 points over data programming.

Ablations We examine individual factors:

Unipolar Correction: Modeling unipolar sources (Sec 4), which we find to be especially common when fine-grained tasks are involved, leads to an average gain of 2.8 points of accuracy in `MeTaL` performance.

Joint Task Modeling: Next, we use our algorithm to estimate the accuracies of sources for each task separately, to observe the empirical impact of modeling the multi-task setting jointly as proposed. We see average gains of 1.3 points in accuracy (see Appendix).

End Model Generalization: Though not possible in many settings, in our experiments we can directly apply the label model to make predictions. In Table 6, we show that the end model improves performance by an average 3.4 points in accuracy, validating that the models trained do indeed learn to generalize beyond the provided weak supervision. Moreover, the largest generalization gain of 7 points in accuracy came from the dataset with the most available unlabeled data (n=63k), demonstrating scaling consistent with the predictions of our theory (Fig. 5). This ability to leverage additional unlabeled data and more sophisticated end models are key advantages of the weak supervision approach in practice.

	# Train	LM	EM	*Gain*
NER	62,547	75.2	82.2	*7.0*
RE	9,090	55.3	57.4	*2.1*
Doc	2,630	75.6	76.6	*1.0*

Figure 6: Using the label model (LM) predictions directly versus using an end model trained on them (EM).

6 Conclusion

We presented `MeTaL`, a framework for training models with weak supervision from diverse, *multi-task* sources having different granularities, accuracies, and correlations. We tackle the core challenge of recovering the unknown source accuracies via a scalable matrix completion-style algorithm, introduce theoretical bounds characterizing the key scaling with respect to unlabeled data, and demonstrate empirical gains on real-world datasets. In future work, we hope to learn the task relationship structure and cover a broader range of settings where labeled training data is a bottleneck.

Acknowledgements We gratefully acknowledge the support of DARPA under Nos. FA87501720095 (D3M) and FA86501827865 (SDH), NIH under No. N000141712266 (Mobilize), NSF under Nos. CCF1763315 (Beyond Sparsity), CCF1563078 (Volume to Velocity), and DGE-114747 (NSF GRF), ONR under No. N000141712266 (Unifying Weak Supervision), the Moore Foundation, NXP, Xilinx, LETI-CEA, Intel, Google, NEC, Toshiba, TSMC, ARM, Hitachi, BASF, Accenture, Ericsson, Qualcomm, Analog Devices, the Okawa Foundation, and American Family Insurance, the Stanford Interdisciplinary Graduate and Bio-X fellowships, the Intelligence Community Postdoctoral Fellowship, and members of the Stanford DAWN project: Intel, Microsoft, Teradata, Facebook, Google, Ant Financial, NEC, SAP, and VMWare. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views, policies, or endorsements, either expressed or implied, of DARPA, NIH, ONR, or the U.S. Government.

References

Anandkumar, A.; Ge, R.; Hsu, D.; Kakade, S. M.; and Telgarsky, M. 2014. Tensor decompositions for learning latent variable models. *The Journal of Machine Learning Research* 15(1):2773–2832.

Bach, S. H.; He, B.; Ratner, A. J.; and Ré, C. 2017. Learning the structure of generative models without labeled data.

Blum, A., and Mitchell, T. 1998. Combining labeled and unlabeled data with co-training.

Caruana, R. 1993. Multitask learning: A knowledge-based source of inductive bias.

Chandrasekaran, V.; Parrilo, P. A.; and Willsky, A. S. 2010. Latent variable graphical model selection via convex optimization. In *Communication, Control, and Computing (Allerton), 2010 48th Annual Allerton Conference on*, 1610–1613. IEEE.

Craven, M., and Kumlien, J. 1999. Constructing biological knowledge bases by extracting information from text sources.

Dalvi, N.; Dasgupta, A.; Kumar, R.; and Rastogi, V. 2013a. Aggregating crowdsourced binary ratings.

Dalvi, N.; Dasgupta, A.; Kumar, R.; and Rastogi, V. 2013b. Aggregating crowdsourced binary ratings.

Dawid, A. P., and Skene, A. M. 1979. Maximum likelihood estimation of observer error-rates using the em algorithm. *Applied statistics* 20–28.

Ghosh, A.; Kale, S.; and McAfee, P. 2011. Who moderates the moderators?: Crowdsourcing abuse detection in user-generated content.

Guan, M. Y.; Gulshan, V.; Dai, A. M.; and Hinton, G. E. 2017. Who said what: Modeling individual labelers improves classification. *arXiv preprint arXiv:1703.08774*.

Gupta, S., and Manning, C. D. 2014. Improved pattern learning for bootstrapped entity extraction.

Hoffmann, R.; Zhang, C.; Ling, X.; Zettlemoyer, L.; and Weld, D. S. 2011. Knowledge-based weak supervision for information extraction of overlapping relations.

Karger, D. R.; Oh, S.; and Shah, D. 2011. Iterative learning for reliable crowdsourcing systems.

Khetan, A.; Lipton, Z. C.; and Anandkumar, A. 2017. Learning from noisy singly-labeled data. *arXiv preprint arXiv:1712.04577*.

Liang, P.; Jordan, M. I.; and Klein, D. 2009. Learning from measurements in exponential families.

Loh, P.-L., and Wainwright, M. J. 2012. Structure estimation for discrete graphical models: Generalized covariance matrices and their inverses.

Mann, G. S., and McCallum, A. 2010. Generalized expectation criteria for semi-supervised learning with weakly labeled data. *JMLR* 11(Feb):955–984.

Mintz, M.; Bills, S.; Snow, R.; and Jurafsky, D. 2009. Distant supervision for relation extraction without labeled data.

National Institutes of Health. 2017. Open-i.

Platanios, E.; Poon, H.; Mitchell, T. M.; and Horvitz, E. J. 2017. Estimating accuracy from unlabeled data: A probabilistic logic approach.

Ratner, A. J.; De Sa, C. M.; Wu, S.; Selsam, D.; and Ré, C. 2016. Data programming: Creating large training sets, quickly.

Ratner, A.; Bach, S.; Ehrenberg, H.; Fries, J.; Wu, S.; and Ré, C. 2018. Snorkel: Rapid training data creation with weak supervision.

Ruder, S. 2017. An overview of multi-task learning in deep neural networks. *CoRR* abs/1706.05098.

Søgaard, A., and Goldberg, Y. 2016. Deep multi-task learning with low level tasks supervised at lower layers.

Takamatsu, S.; Sato, I.; and Nakagawa, H. 2012. Reducing wrong labels in distant supervision for relation extraction.

Tropp, J. A. 2015. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning* 8(1-2):1–230.

Varma, P.; He, B. D.; Bajaj, P.; Khandwala, N.; Banerjee, I.; Rubin, D.; and Ré, C. 2017. Inferring generative model structure with static analysis.

Weischedel, R.; Hovy, E.; Marcus, M.; Palmer, M.; Belvin, R.; Pradhan, S.; Ramshaw, L.; and Xue, N. 2011. Ontonotes: A large training corpus for enhanced processing. *Handbook of Natural Language Processing and Machine Translation. Springer*.

Xiao, T.; Xia, T.; Yang, Y.; Huang, C.; and Wang, X. 2015. Learning from massive noisy labeled data for image classification.

Zaidan, O. F., and Eisner, J. 2008. Modeling annotators: A generative approach to learning from annotator rationales.

Zhang, Y.; Chen, X.; Zhou, D.; and Jordan, M. I. 2014. Spectral methods meet em: A provably optimal algorithm for crowdsourcing.

Zhang, C.; Ré, C.; Cafarella, M.; De Sa, C.; Ratner, A.; Shin, J.; Wang, F.; and Wu, S. 2017a. DeepDive: Declarative knowledge base construction. *Commun. ACM* 60(5):93–102.

Zhang, Y.; Zhong, V.; Chen, D.; Angeli, G.; and Manning, C. D. 2017b. Position-aware attention and supervised data improve slot filling.