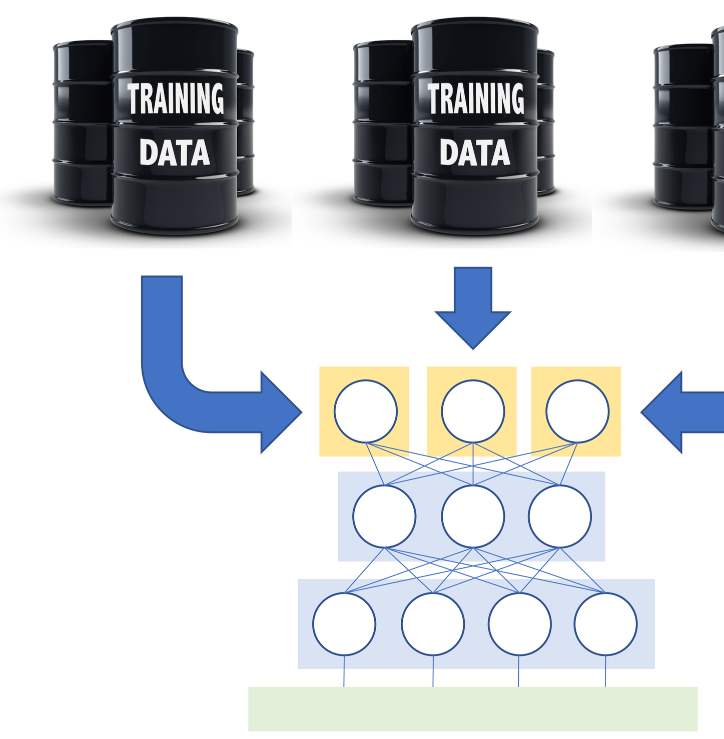
MeTaL: Training Complex Models with Multi-Task Weak Supervision

Alex Ratner, Braden Hancock, Jared Dunnmon, Fred Sala, Shreyash Pandey, and Chris Ré

STANFORD DAWN

Motivation: The Training Data Bottleneck



Multi-task learning models require large labeled training sets… for each task!

Multi-Task Programmatic Weak Supervision Example: Fine Grain NER

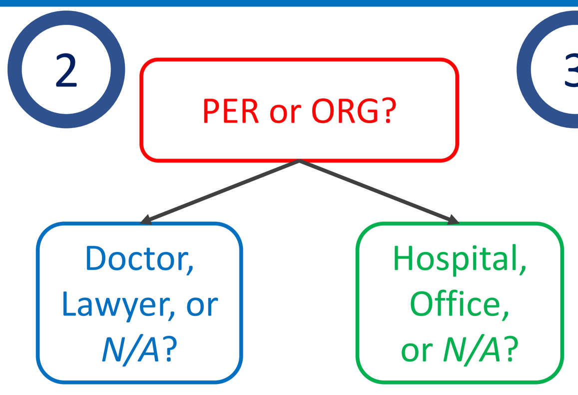
1

PER:DOCTOR

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

ORG:HOSPITAL

2



1. In a *fine grain named entity recognition (NER)* example, we want to tag mentions of {doctors, lawyers, hospitals, offices}.
2. We can decompose this into three hierarchically related sub-tasks
3. This then lets us easily use different programmatic sources of supervision, e.g. pattern matchers, existing classifiers, and external ontologies / databases to generate training labels

3

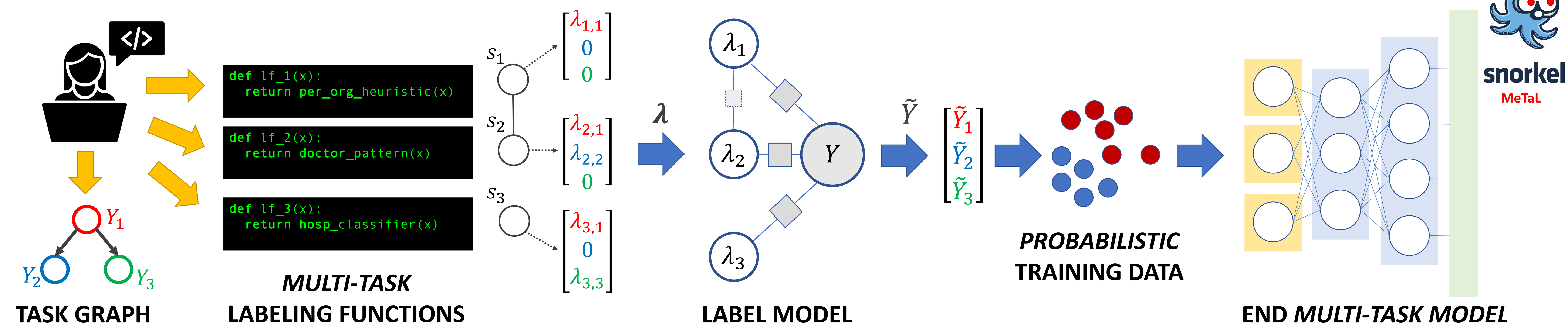
```
def is_all_upper(x):
  if all(map(is_upper, x.split())):
    return PERSON

def existing_classifier(x):
  if classifier(x) == 'org':
    return ORG

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

def is_in_corp_name_DB(x):
  if x in CORP_NAMES_DB:
    return OFFICE
```

Snorkel MeTaL: An Open-Source System for *Multi-Task* Weak Supervision



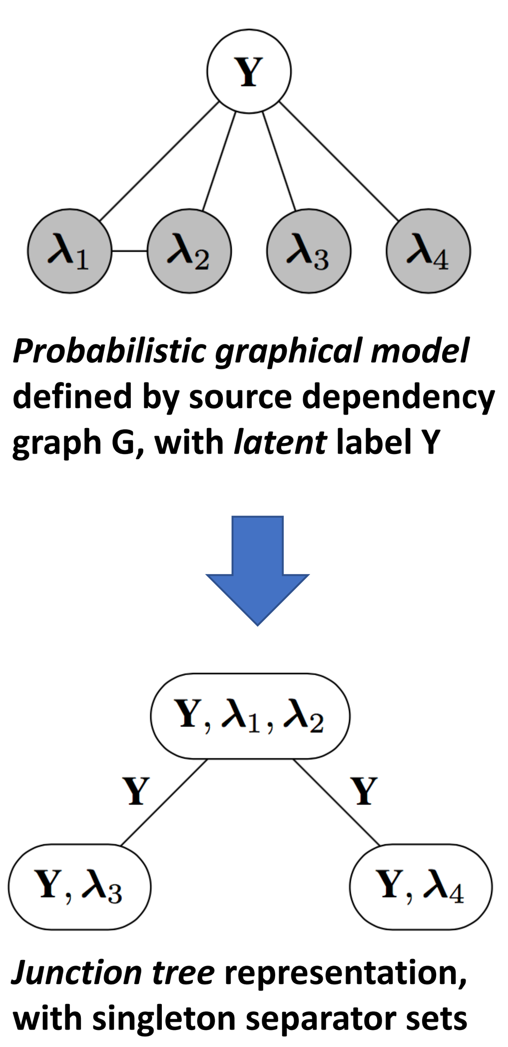
1 Users write *labeling functions* ***for multiple related tasks***

2 We model the labeling functions' behavior to de-noise them

3 We use the *probabilistic* labels to train a ***multi-task*** model

Modeling Multi-Task Supervision as Matrix Completion

Representation & Model



Algorithm: Modeling Weak Supervision as Matrix Completion

How to learn the above model given Y is not observed? We show how to formulate as matrix completion & provably recover!

1 Split into **observed (O)** and **separator set (S)** cliques; consider the covariance & inverse covariance matrices

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

$$\boxed{K} = \Sigma^{-1} = \begin{bmatrix} K_O & K_{OS} \\ K_{OS}^T & K_S \end{bmatrix}$$

We use the recent result that K has graph-structured sparsity [Loh & Wainwright 2013]; i.e. has zeros for indices Ω where no edge in our graphical model

2 Use block matrix inversion lemma to rearrange into an equation w/ **observed** matrix **low-rank** parameters, and a **graph-structured sparse** matrix

$$K_O = \Sigma_O^{-1} + c\Sigma_O^{-1}\Sigma_{OS}\Sigma_{OS}^T\Sigma_O^{-1}$$

$$\underbrace{\boxed{K_O}}_{\textit{Sparse}} = \underbrace{\boxed{\Sigma_O^{-1}}}_{\textit{Empirically observed}} + \underbrace{\boxed{zz^T}}_{\textit{Low-rank}}$$

3 **Result:** A **matrix completion** problem w.r.t. the inverse observed covariance!

$$\hat{z} \leftarrow \operatorname{argmin}_z \left\| \hat{\Sigma}_O^{-1} + zz^T \right\|_{\boxed{\Omega}}$$

This is a standard problem to solve!

We then show that a simple, deterministic check of the dependency graph G can determine identifiability of this solution

Theory

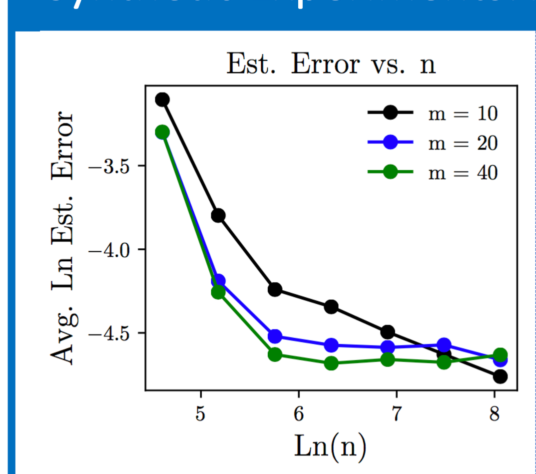
I **Result: Given n unlabeled data points,** the source accuracy & correlation estimation error **decreases as $n^{-1/2}$**

$$\underbrace{\boxed{E\left[\|\hat{\mu} - \mu^*\|\right]}}_{\text{Estimation error of the source accuracies and correlations}} = \underbrace{\boxed{O\left(\frac{1}{\sqrt{n}}\right)}}_{\text{Scales with amount of } \textit{unlabeled} \text{ data n!}}$$

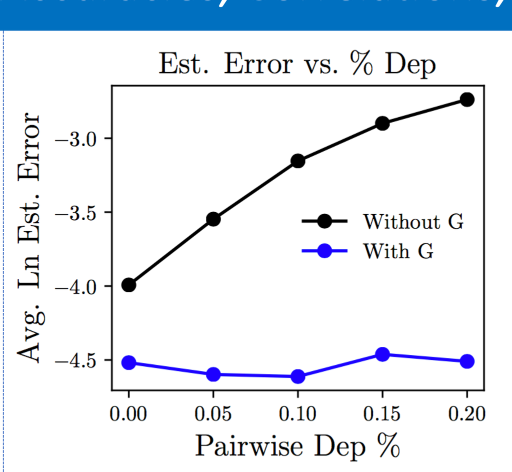
II **Result:** Given standard assumptions on relationship between end model features and our labels (see paper), **end model generalization controlled** by the above as well:

$$\underbrace{\boxed{E\left[\|l_{\hat{w}} - l_{w^*}\|\right]}}_{\text{Generalization error of end model with learned weights } \hat{w}} \leq \underbrace{\boxed{\gamma}}_{\text{Error in standard ERM procedure}} + 4|\mathcal{Y}|\boxed{E\left[\|\hat{\mu} - \mu^*\|\right]}$$

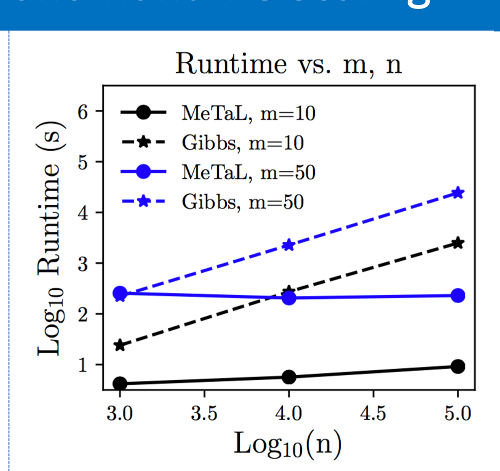
Synthetic Experiments: Accuracies, Correlations, and Runtime Scaling



Our estimates of the label sources' accuracies ***improve with more unlabeled data***



Our approach ***successfully handles the effect of a given source dependency structure (G)***



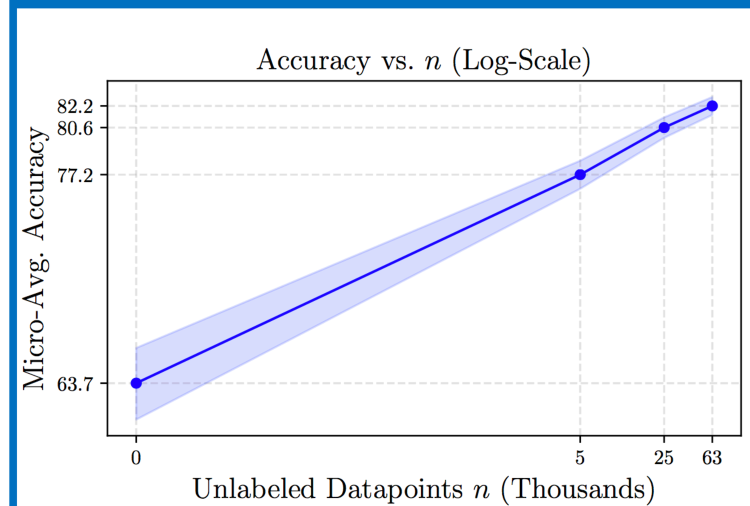
Our approach *is* ***orders-of-magnitude faster than previous Gibbs sampling-based techniques***

Real-World Experiments: Fine Grain Text Tagging & Classification

	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 [28]	78.4 ± 1.2	49.0 ± 2.7	75.8 ± 0.9	67.7
`MeTaL`	$\mathbf{82.2} \pm 0.8$	$\mathbf{56.7} \pm 2.1$	$\mathbf{76.6} \pm 0.4$	**71.8**

Significant performance gains in three text applications over models trained with: (i) a small hand-labeled dataset, corresponding to similar time as developing the weak supervision (Gold Dev); (ii) majority vote of the sources (MV); (iii) a prior single-task weak supervision approach (DP)

Scaling with Unlabeled Data



As predicted by theory, ***our end model accuracy scales with more unlabeled data***

Open-source code and tutorials: `github.com/HazyResearch/metal`