Few-shot learning of weak supervision sources in Snorkel

(or, learning weakly supervised weak supervisors)

Jesse Mu

Project Outline

- Replicate Snorkel causal relation extraction system
- Learn weak supervision sources from tiny sets of annotated examples, and compare performance to (1)

The *New* New Oil

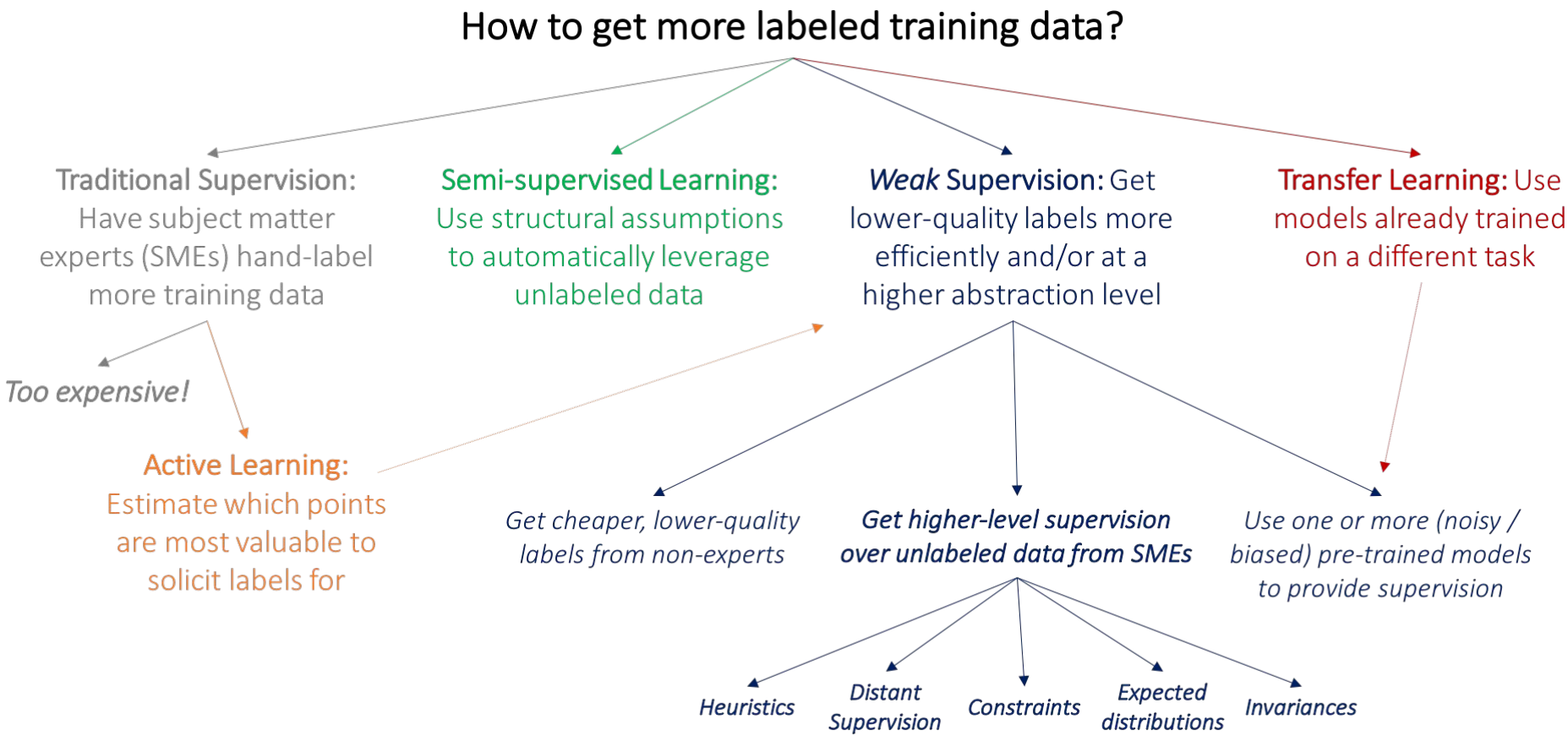
snorkel

A training data creation and management system focused on information extraction

Stanford **DAWN**

*"We don't have **better algorithms** than anyone else; we just have **more data**"*

Peter Norvig
Chief Scientist, Google


How to get more labeled training data?
Traditional Supervision: Have subject matter experts (SMEs) hand-label more training data
Too expensive!
Active Learning: Estimate which points are most valuable to solicit labels for
Semi-supervised Learning: Use structural assumptions to automatically leverage unlabeled data
Weak Supervision: Get lower-quality labels more efficiently and/or at a higher abstraction level
Get cheaper, lower-quality labels from non-experts
Get higher-level supervision over unlabeled data from SMEs
Heuristics
Distant Supervision
Constraints
Expected distributions
Invariances
Transfer Learning: Use models already trained on a different task
Use one or more (noisy / biased) pre-trained models to provide supervision

Data Programming: Creating Large Training Sets, Quickly

Alexander Ratner, Christopher De Sa, Sen Wu, Daniel Selsam, Christopher Ré
Stanford University
{ajratner,cdesa,senwu,dselsam,chrismre}@stanford.edu

Large labeled training sets are the critical building blocks of supervised learning methods and are key enablers of deep learning techniques. For some applications, creating labeled training sets is the most time-consuming and expensive part of applying machine learning. We therefore propose a paradigm for the programmatic creation of training sets called *data programming* in which users express weak supervision strategies or domain heuristics as *labeling functions*, which are programs that label subsets of the data, but that are noisy and may conflict. We show that by explicitly representing this training set labeling process as a generative model, we can "denoise" the generated training set, and establish theoretically that we can recover the parameters of these generative models in a handful of settings. We then show how to modify a discriminative loss function to make it noise-aware, and demonstrate

NIPS 2016

BioMed Central

The Open Access Publisher

BioMed Central
The Open Access Publisher

```
def lf1(x):
  cid = (x.chemical_id, x.disease_id)
  return 1 if cid in KB else 0
```

"Chemical A is found to cause disease B under certain conditions..." → Label=TRUE

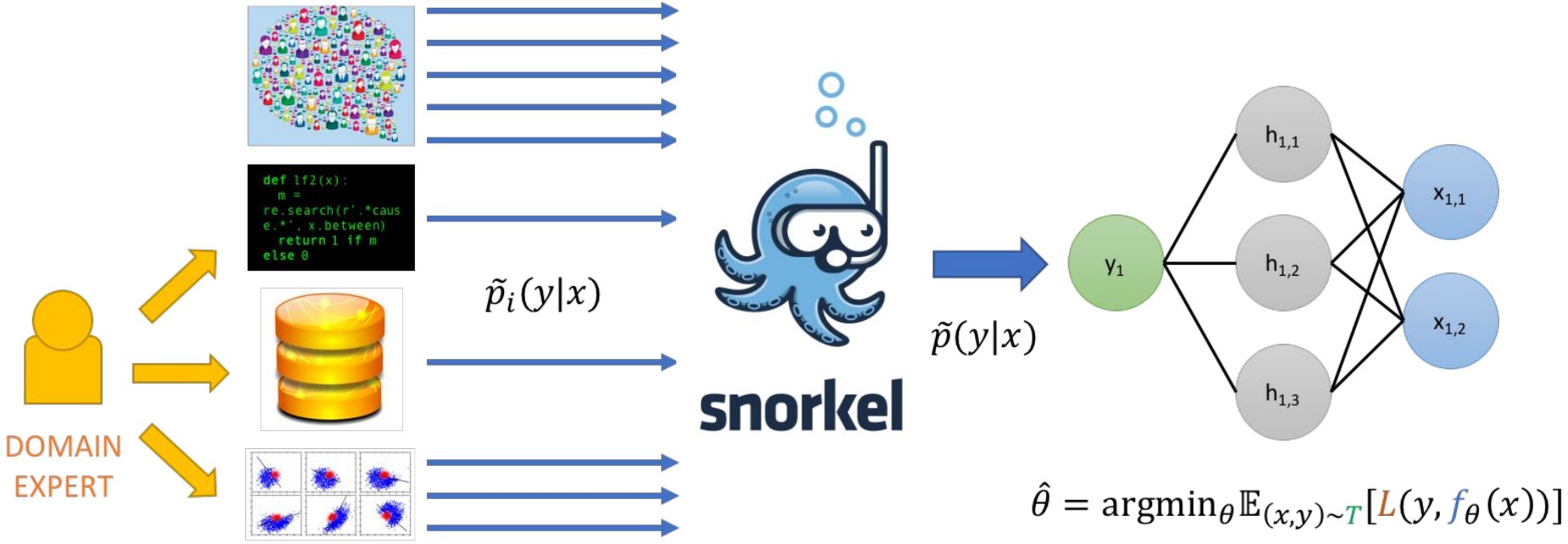
Existing KB

Contains(A, B)

```
def lf2(x):
  m = re.search(r'.*cause.*', x.between)
  return 1 if m else 0
```

"Chemical A is found to cause disease B under certain conditions..." → Label=TRUE

r'.*cause.*'



Example Weak Supervision Sources

Technical Challenge: Integrating & Modeling Diverse Sources

Use Weak Supervision to Train End Model

Extension: from *examples* to *labeling functions*

- Labeling functions (LFs) require programming experience and abstraction.
- Can we learn noisy labelers from few examples, without a single line of code?
- Given sentences and relations, generate many candidate LFs that distinguish LF from surrounding sentences

Several diseases that appear to be heritable, but not genetically defined, have been observed at low frequency in the breed.[11] , [12] , [13] Many of these disorders have evolved with the domestic dog over time and inherited by descent as breeds have been created [3]. Except for hip dysplasia, which is considered one of the more serious disorders of Samoyed, most heritable and potentially heritable disease traits of the breed have been of minor importance.[11] ***There are only three simple deleterious genetic disorders in Samoyed with defined causes, X-linked glomerulopathy [4], X-linked progressive retinal atrophy [5], and an incomplete dominant short-limbed defect with ocular abnormalities [6, 7].***

causes(e1=genetics, e2=retinal atrophy)

causes(e1=genetics, e2=glomerulopathy)

```
def candidate_lf1(s, e1, e2):
    return 'causes' in s

def candidate_lf2(s, e1, e2):
    return 'deleterious' in s
```

Extension: from *examples* to *labeling functions*

- Labeling functions (LFs) require programming experience and abstraction.
- Can we learn noisy labelers from few examples, without a single line of code?
- Given sentences and relations, generate many candidate LFs that distinguish LF from surrounding sentences

2 questions:

- How dumb are LFs generated in this way?
- How dumb can LFs be before Snorkel begins to break down?

Few-shot learning of weak supervision sources in Snorkel

(or, learning weakly supervised weak supervisors)

Jesse Mu

Application	# of LFs	Coverage	$\lvert S_{\lambda \neq 0} \rvert$	Overlap	Conflict	F1 Score Improvement	
						HT	LSTM
KBP (News)	40	29.39	2.03M	1.38	0.15	1.92	3.12
Genomics	146	53.61	256K	26.71	2.05	1.59	0.47
Pharmacogenomics	7	7.70	129K	0.35	0.32	3.60	4.94
Diseases	12	53.32	418K	31.81	0.98	N/A	N/A

Table 2: Labeling function (LF) summary statistics, sizes of generated training sets $S_{\lambda \neq 0}$ (only counting non-zero labels), and relative F1 score improvement over baseline IRT methods for hand-tuned (HT) and LSTM-generated (LSTM) feature sets.