Snorkel: Rapid Training Data Creation with Weak Supervision

RATNER, A., BACH, S. H., EHRENBERG, H., FRIES, J., WU, S., & RÉ, C. (2017)

PRESENTED BY: MAREK STRELEC

Motivation

- Problem: Users struggle to write good features
- DNNs to rescue:
  - perform well without any hand-engineered features
- State-of-the-art machine learning models require massive labeled training sets
  - Often do not exist for real-world applications
- Hand-labeled training data is expensive and slow to collect
- A common scenario
  - access to tons of *unlabeled* training data, and have some idea of how to label it programmatically
- Key idea: model the process of training set creation

Weak Supervision

- Generate training data using heuristics, rules-of-thumb, existing databases, ontologies, …
- It isn't perfectly accurate, possibly consists overlapping and conflicting signals
- Sources of weak supervision
  - Domain heuristics (e.g. common patterns, rules of thumb, etc.)
  - Distant supervision - Existing ground-truth data that is not an exact fit for the specific task
  - Weak classifiers (boosting)
  - Unreliable non-expert annotators (e.g. crowdsourcing)
- Data programming (Ratner, Alexander J., et al., 2016)
  - Domain experts encode various weak supervision signals as *labeling functions*
  - These labeling functions can be noisy but can be reconciled and denoised automatically
  - Used to train a discriminative model

Extracting Spouse Relations - Preprocessing

- Process documents into sentences and tokens
- Define Candidate Schema

```
Spouse = candidate_subclass('Spouse', ['person1', 'person2'])
```

- Define Candidate Extractor
  - Named Entity Recognition - PersonMatcher
  - Extract Candidate objects for all pairs of n-grams that were tagged as people

```
CandidateExtractor(Spouse, [ngrams, ngrams], [person_matcher, person_matcher])
```

- Apply Candidate Extractor to all preprocessed documents

Extracting spouse relations - Generating and modeling noisy training labels

- Create Labeling Functions
  - Marks each Candidate as ‘true, ‘false’, or ‘abstain’
  - Pattern-based
    - E.g. Checking whether the last names match
  - Distant Supervision
    - E.g. DB of known spouse pairs
- Apply over all training candidates
- Fit the Generative Model
  - Train a model of the LFs to estimate their accuracies
  - Once the model is trained, outputs of the LFs are combined into a single, noise-aware training label set

	j	Coverage	Overlaps	Conflicts
LF_distant_supervision	0	0.001481	0.001481	0.000628
LF_distant_supervision_last_names	1	0.008080	0.007856	0.004758
LF_husband_wife	2	0.104642	0.066798	0.017867
LF_husband_wife_left_window	3	0.078021	0.057910	0.010774
LF_same_last_name	4	0.016700	0.014994	0.010011
LF_no_spouse_in_sentence	5	0.603026	0.081657	0.009472
LF_and_married	6	0.000673	0.000539	0.000404
LF_familial_relationship	7	0.104283	0.091489	0.021413
LF_family_left_window	8	0.073352	0.067651	0.012076
LF_other_relationship	9	0.009337	0.006868	0.001122

[0.07592901, 0.07395425, 0.11954169, 0.11397737, 0.07065144,
0.6901572 , 0.07358515, 0.15698341, 0.13658573, 0.08221857]

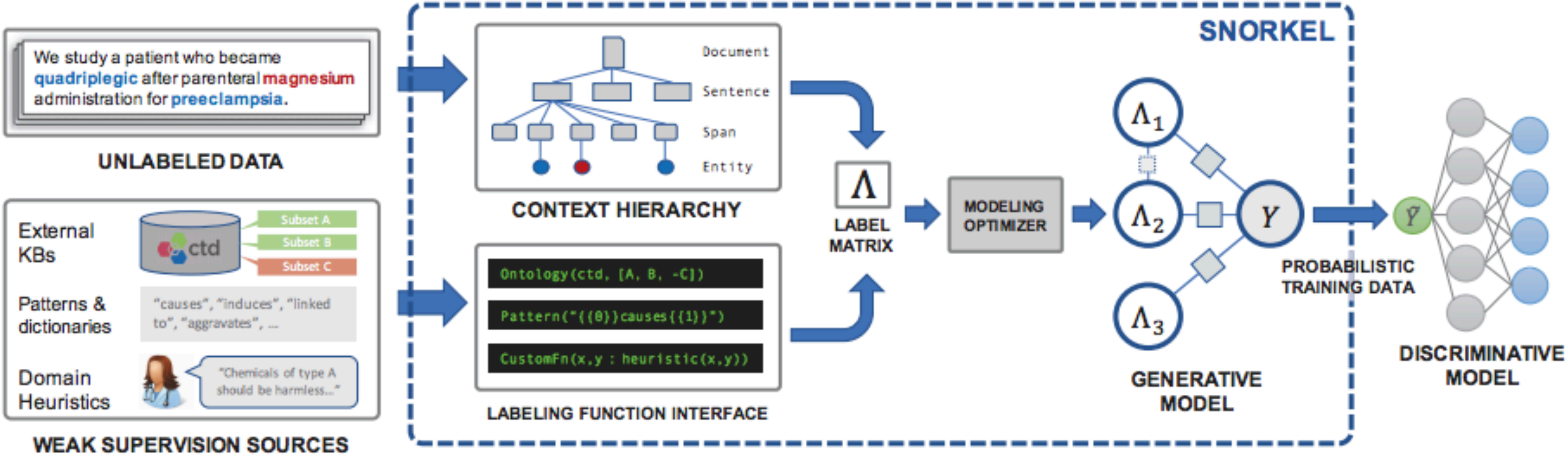
Extracting spouse relations - Generating and modeling noisy training labels

Results on the dev set:

	Accuracy	Coverage	Precision	Recall
0	0.534134	0.6665	0.541630	0.360980
1	0.532118	0.6694	0.539711	0.358431
2	0.565538	0.6645	0.574173	0.380980
3	0.565055	0.6702	0.574157	0.377255
4	0.543329	0.6716	0.553972	0.358235
5	0.796530	0.7146	0.804610	0.574902
6	0.526747	0.6711	0.535660	0.343137
7	0.577701	0.6673	0.588448	0.383529
8	0.568233	0.6661	0.572989	0.370196
9	0.540609	0.6698	0.551551	0.366078

Extracting spouse relations - Training an End Extraction Mode

- Train a predictive model
  - A state-of-the-art deep neural network
- Snorkel provides API for frameworks such as TensorFlow, PyTorch
- Uses probabilistic training labels from the generative model
- Binary output - spouse/non-spouse candidate



We study a patient who became quadriplegic after parenteral magnesium administration for preeclampsia.
UNLABELED DATA
External KBs
ctd
Subset A
Subset B
Subset C
Patterns & dictionaries
"causes", "induces", "linked to", "aggravates", ...
Domain Heuristics
"Chemicals of type A should be harmless..."
WEAK SUPERVISION SOURCES
SNORKEL
Document
Sentence
Span
Entity
CONTEXT HIERARCHY
Ontology(ctd, [A, B, -C])
Pattern("{{0}}causes{{1}}")
CustomFn(x,y : heuristic(x,y))
LABELING FUNCTION INTERFACE
Λ
LABEL MATRIX
MODELING OPTIMIZER
Λ1
Λ2
Λ3
Y
GENERATIVE MODEL
PROBABILISTIC TRAINING DATA
Ỹ
DISCRIMINATIVE MODEL

USER:

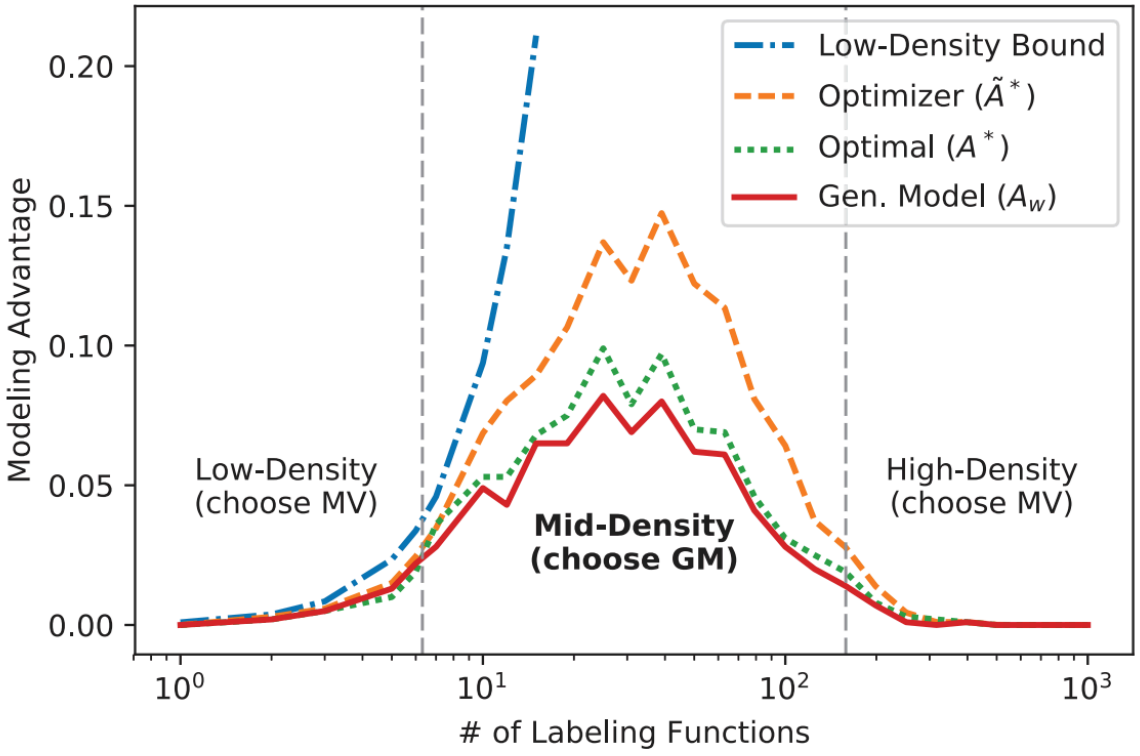
- Provide unlabeled data
- Writes labeling functions
- Chooses a discriminative model (e.g. Bi-LSTM)

SNORKEL:

- Creates a noisy training data
- Learns a model of this noise
- Trains a noise-aware discriminative model

Generative Model or Majority Voting?

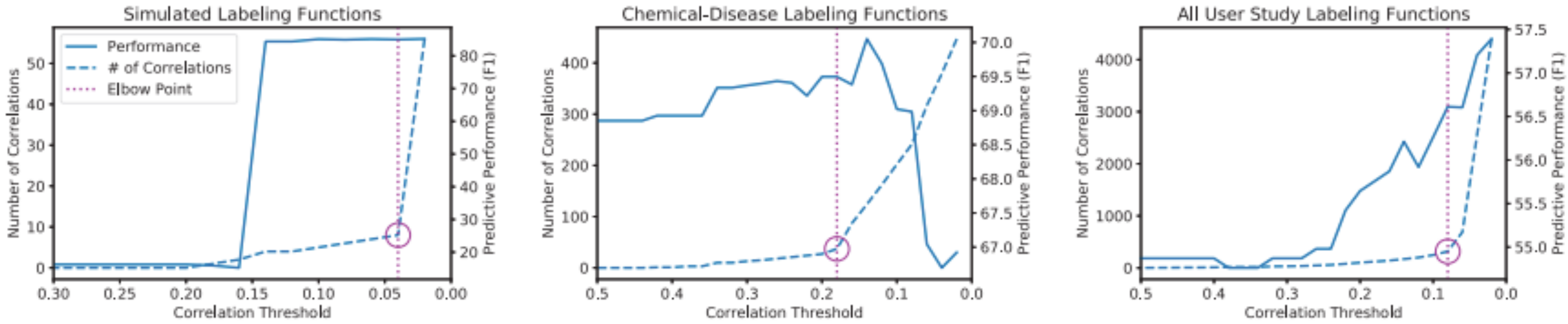
- “When does modeling the accuracies of sources improve end-to-end predictive performance?”
- Heuristic - ratio of positive to negative labels



Correlated labels

- Snorkel users writing labeling functions that are statistically dependent.
  - LF are variations of each other
  - LF operate on correlated inputs
  - LF use correlated sources of knowledge
- This affects estimates of the true labels
- Getting users to somehow indicate dependencies by hand is difficult and error-prone
- Pseudo-likelihood estimator
  - Selecting which dependencies to model
  - Hyper-parameter *e:* trades-off between predictive performance and computational cost
  - Large e = no correlations included
  - Choice of e determines the model's complexity

Correlated labels



Evaluation – User Study

- How quickly subject-matter experts could learn to write labelling functions
- 4.5 hours of instruction on how to use and evaluate models developed using Snorkel
- 2.5 hours to write labelling functions
- Snorkel users: 30.4 F1 average score
- The average hand-supervision: 20.9 F1 average score

Evaluation - Applications

Task	# LFs	% Pos.	# Docs	# Candidates
Chem	16	4.1	1,753	65,398
EHR	24	36.8	47,827	225,607
CDR	33	24.6	900	8,272
Spouses	11	8.3	2,073	22,195
Radiology	18	36.0	3,851	3,851
Crowd	102	-	505	505

Evaluation

	Distant Supervision			Snorkel (Gen.)				Snorkel (Disc.)				Hand Supervision		
Task	P	R	**F1**	P	R	**F1**	**Lift**	P	R	**F1**	**Lift**	P	R	**F1**
Chem	11.2	41.2	17.6	78.6	21.6	33.8	+16.2	87.0	39.2	54.1	+36.5	-	-	-
EHR	81.4	64.8	72.2	77.1	72.9	74.9	+2.7	80.2	82.6	81.4	+9.2	-	-	-
CDR	25.5	34.8	29.4	52.3	30.4	38.5	+9.1	38.8	54.3	45.3	+15.9	39.9	58.1	47.3
Spouses	9.9	34.8	15.4	53.5	62.1	57.4	+42.0	48.4	61.6	54.2	+38.8	47.8	62.5	54.2

Effect of Generative Modeling

Task	Disc. Model on Unweighted LFs	Disc. Model	Lift
Chem	48.6	54.1	+5.5
EHR	80.9	81.4	+0.5
CDR	42.0	45.3	+3.3
Spouses	52.8	54.2	+1.4
Crowd (Acc)	62.5	65.6	+3.1
Rad. (AUC)	67.0	72.0	+5.0

Conclusion

- Snorkel provides a new paradigm for managing weak supervision to create training data sets
- Users provide Labeling Functions that capture domain knowledge and resources
- Discriminative models trained on Snorkel’s probabilistic labels produce consistently better labeling
- Labeling functions written in Snorkel, even by SME users, can match or exceed a traditional hand-labeling approach

Thank you!