**Motivation: Answering Macroscopic Questions**

*Ex: Given certain symptoms, which genetic mutations should we test for?*

<table>
<thead>
<tr>
<th>Unstructured Input Data (e.g., Pubmed articles)</th>
<th>Structured knowledgebase</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gene</strong></td>
<td><strong>Phenotype</strong></td>
</tr>
<tr>
<td>KRT16</td>
<td>Pachyonychia Congenita</td>
</tr>
<tr>
<td>LTBP2</td>
<td>Pachyonychia Congenita</td>
</tr>
<tr>
<td>DEAF1</td>
<td>Pachyonychia Congenita</td>
</tr>
</tbody>
</table>

We then train a “noise-aware” variant of our end model:

Results on benchmark extraction tasks:

- **TAC-KBP Relation Extraction Challenge**: Achieved what would have been winning score (2014 Slot Filling)
- **LSTM: 6 pts. F1 (20%) over state-of-the-art LSTM baseline**
- **CDR Chemical / Disease Tagging**: We match state-of-the-art results (using simple logistic regression model)

We can achieve optimal results with less input from users!

---

**Structure Learning: Identifying Correlated Sources**

Cannot assume all sources are conditionally independent

Developed a structure learning method to automatically identify correlations among labeling functions

Structure learning is 100x faster and has ⅓ the errors of MLE.

Leads to average 1.5 F1 boost on existing applications

<table>
<thead>
<tr>
<th>Application</th>
<th>Independent F1</th>
<th>Structure F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disease Tagging</td>
<td>66.3</td>
<td>68.9</td>
</tr>
<tr>
<td>Chemical-Disease</td>
<td>54.6</td>
<td>55.9</td>
</tr>
<tr>
<td>Tables</td>
<td>88.1</td>
<td>88.7</td>
</tr>
</tbody>
</table>

---

**Socratic Learning: Correcting Generative Models**

Sources can have different accuracies for latent classes in the data, which are difficult to identify manually.

Use features from the discriminative model to correct the generative model automatically.

---

**Data Programming with Snorkel**

Data programming is our framework for programmatically generating training data (*NIPS 2016, arxiv:1605.07723*)

Newly released as an open-source framework, Snorkel: [http://snorkel.stanford.edu](http://snorkel.stanford.edu)

To support these methods we need massive training datasets!

**Increasing shift towards automatic feature generation:**

- **Hand-Tuned Features**
- **Automatic Template-Based**
- **LSTMs (Deep Learning)**

**Users write labeling functions (LFs):**

- Encode domain heuristics
- Can be noisy / conflict
- Can subsume approaches like distant supervision, crowdsourcing, etc.
- Any scripting language works (e.g. Python)

We treat the LFs as a generative model, and learn their relative accuracies automatically

Can specify dependencies:

**Snorkel denoises training labels, features, and learns automatically**

---

**Snorkel: Lightweight Extraction with Simple Input**

Input: Unstructured information e.g., text, tables (plan to add images, diagrams)

Unlabeled data

Noisy Training Set Model

Learning & Inference

Snorkel users:

- User encodes domain heuristics as labeling functions

---

**Cannot assume all sources are conditionally independent**

Developed a structure learning method to automatically identify correlations among labeling functions

Structure learning is 100x faster and has ⅓ the errors of MLE.

Leads to average 1.5 F1 boost on existing applications

<table>
<thead>
<tr>
<th>Application</th>
<th>Independent F1</th>
<th>Structure F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disease Tagging</td>
<td>66.3</td>
<td>68.9</td>
</tr>
<tr>
<td>Chemical-Disease</td>
<td>54.6</td>
<td>55.9</td>
</tr>
<tr>
<td>Tables</td>
<td>88.1</td>
<td>88.7</td>
</tr>
</tbody>
</table>

---

**Socratic Learning: Correcting Generative Models**

Sources can have different accuracies for latent classes in the data, which are difficult to identify manually.

Use features from the discriminative model to correct the generative model automatically.

---

**Motivation: Answering Macroscopic Questions**

*Ex: Given certain symptoms, which genetic mutations should we test for?*

<table>
<thead>
<tr>
<th>Unstructured Input Data (e.g., Pubmed articles)</th>
<th>Structured knowledgebase</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Gene</strong></td>
<td><strong>Phenotype</strong></td>
</tr>
<tr>
<td>KRT16</td>
<td>Pachyonychia Congenita</td>
</tr>
<tr>
<td>LTBP2</td>
<td>Pachyonychia Congenita</td>
</tr>
<tr>
<td>DEAF1</td>
<td>Pachyonychia Congenita</td>
</tr>
</tbody>
</table>

We then train a “noise-aware” variant of our end model:

Results on benchmark extraction tasks:

- **TAC-KBP Relation Extraction Challenge**: Achieved what would have been winning score (2014 Slot Filling)
- **LSTM: 6 pts. F1 (20%) over state-of-the-art LSTM baseline**
- **CDR Chemical / Disease Tagging**: We match state-of-the-art results (using simple logistic regression model)

We can achieve optimal results with less input from users!

---

**Structure Learning: Identifying Correlated Sources**

Cannot assume all sources are conditionally independent

Developed a structure learning method to automatically identify correlations among labeling functions

Structure learning is 100x faster and has ⅓ the errors of MLE.

Leads to average 1.5 F1 boost on existing applications

<table>
<thead>
<tr>
<th>Application</th>
<th>Independent F1</th>
<th>Structure F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Disease Tagging</td>
<td>66.3</td>
<td>68.9</td>
</tr>
<tr>
<td>Chemical-Disease</td>
<td>54.6</td>
<td>55.9</td>
</tr>
<tr>
<td>Tables</td>
<td>88.1</td>
<td>88.7</td>
</tr>
</tbody>
</table>

---

**Socratic Learning: Correcting Generative Models**

Sources can have different accuracies for latent classes in the data, which are difficult to identify manually.

Use features from the discriminative model to correct the generative model automatically.

---

**Data Programming with Snorkel**

Data programming is our framework for programmatically generating training data (*NIPS 2016, arxiv:1605.07723*)

Newly released as an open-source framework, Snorkel: [http://snorkel.stanford.edu](http://snorkel.stanford.edu)

To support these methods we need massive training datasets!

**Increasing shift towards automatic feature generation:**

- **Hand-Tuned Features**
- **Automatic Template-Based**
- **LSTMs (Deep Learning)**

**Users write labeling functions (LFs):**

- Encode domain heuristics
- Can be noisy / conflict
- Can subsume approaches like distant supervision, crowdsourcing, etc.
- Any scripting language works (e.g. Python)

We treat the LFs as a generative model, and learn their relative accuracies automatically

Can specify dependencies:

**Snorkel denoises training labels, features, and learns automatically**

---

**Snorkel: Lightweight Extraction with Simple Input**

Input: Unstructured information e.g., text, tables (plan to add images, diagrams)

Unlabeled data

Noisy Training Set Model

Learning & Inference

Snorkel users:

- User encodes domain heuristics as labeling functions

---