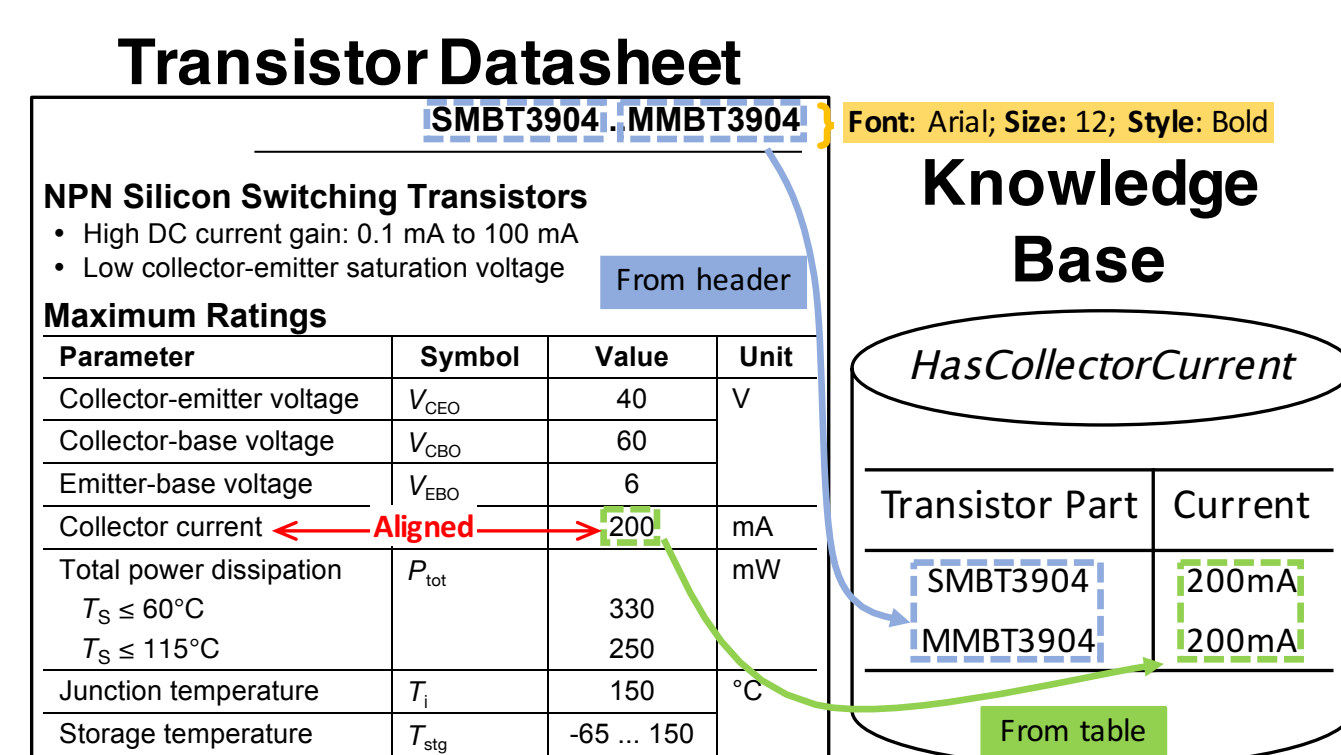


# Fonduer: Knowledge Base Construction from Richly Formatted Data

Sen Wu, Luke Hsiao, Xiao Cheng, Braden Hancock, Theodoros Rekatsinas\*, Philip Levis, Christopher Ré  
 {senwu, lwhsiao, xiao, bradenjh, pal, chrismre}@cs.stanford.edu \*thodrek@wisc.edu  
 Stanford University \*University of Wisconsin-Madison

## Introduction and Background

Fonduer is a machine-learning based knowledge base construction (KBC) framework for richly formatted data, where entity relations and attributes are conveyed via structural, tabular, visual, and textual expressions.

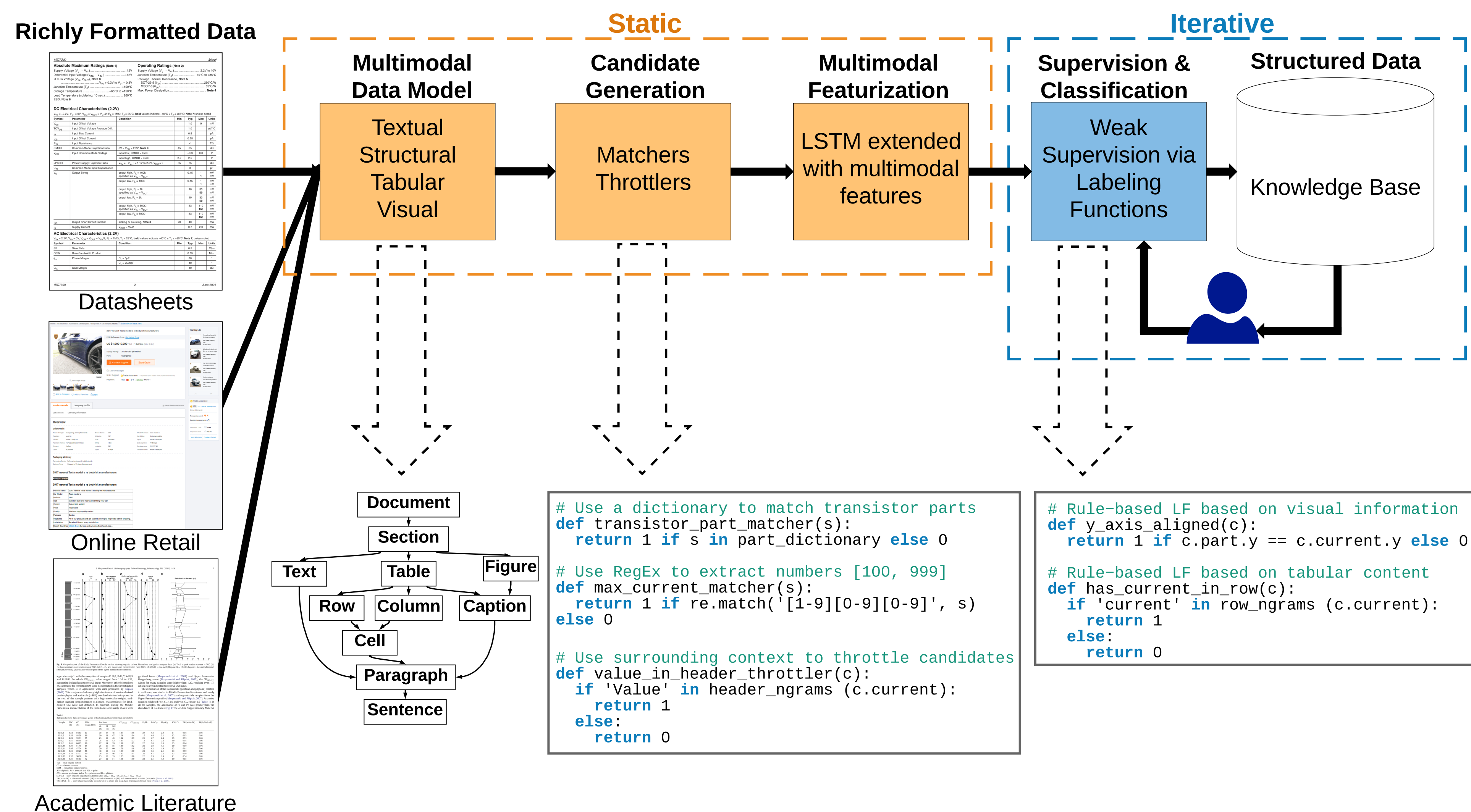


Challenges of KBC from Richly Formatted Data:

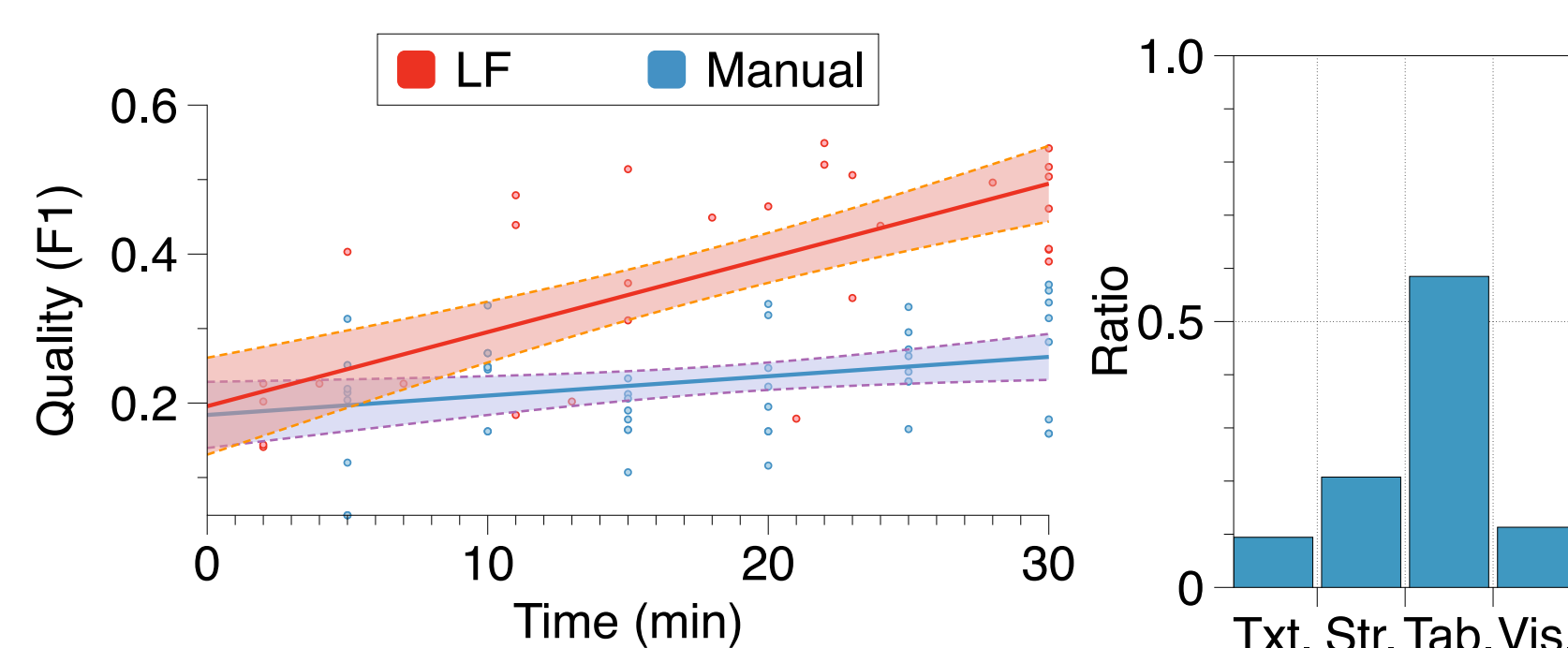
- **Prevalent Document-level Relations:** For richly formatted data, many relations rely on information from throughout the entire document to be extracted.
- **Multimodality:** Semantics are expressed as multiple modalities—textual, structural, tabular, and visual.
- **Data Variety:** The same information can be presented in many different formats and styles, in addition to linguistic variations.

## Knowledge Base Construction Using Fonduer

**Input:** Richly formatted documents (e.g. PDF/HTML/XML/etc.) → **Output:** Structured knowledge base



## User Study



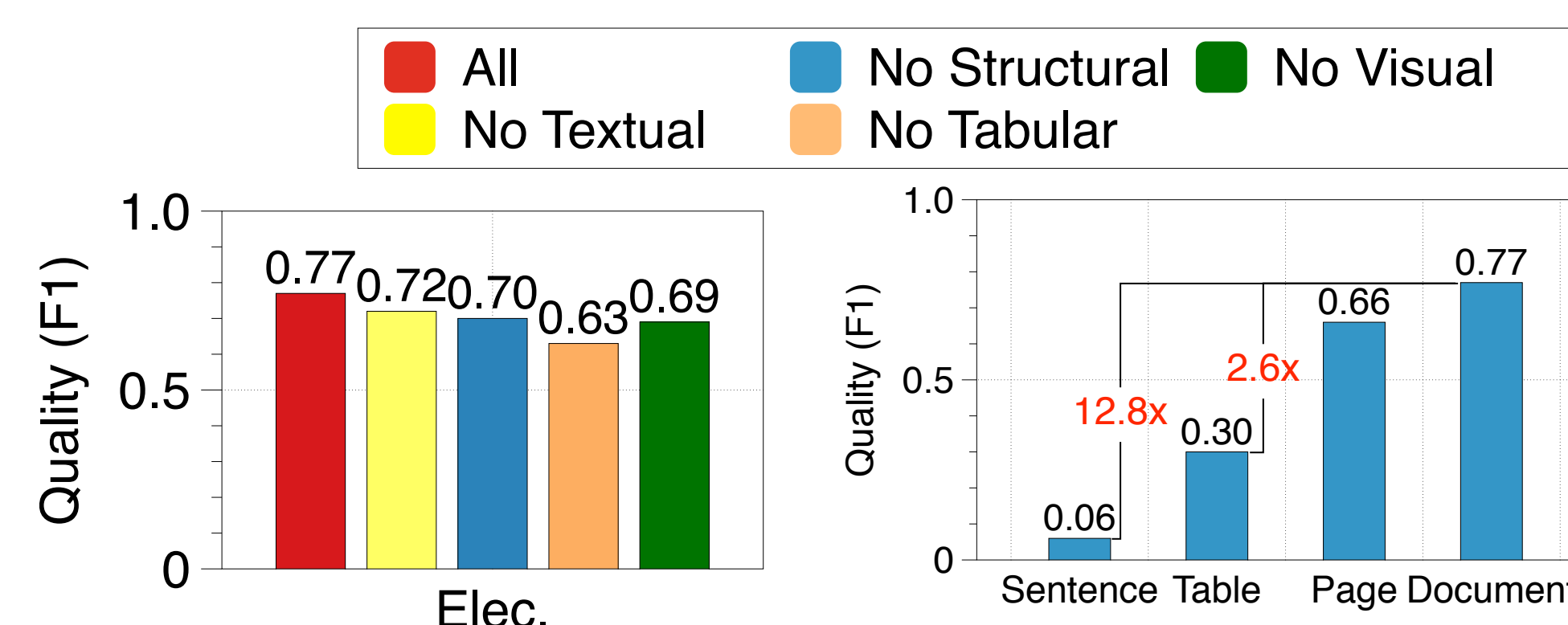
- Users relied 9× more on non-textual signals than textual information alone to identify candidates and provide weak supervision.
- Leveraging multimodal supervision allowed users to create knowledge bases more effectively than traditional manual annotations alone.

## Experimental Results

### End-to-end Quality vs. Public Data

System	Elec.	Gen.	
Knowledge Base	Digi-Key	GWAS Central	GWAS Catalog
# Entries in KB	376	3,008	4,023
# Entries in Fonduer	447	6,420	6,420
Coverage	0.99	0.82	0.80
Accuracy	0.87	0.87	0.89
# New Correct Entries	17	3,154	2,486
Increase in Correct Entries	1.05×	1.87×	1.42×

### Ablation Studies



## Our Users



## References

**Blog:** [hazyresearch.github.io/snorkel/blog/fonduer](https://hazyresearch.github.io/snorkel/blog/fonduer)  
**Paper:** [arxiv.org/abs/1703.05028](https://arxiv.org/abs/1703.05028)  
**Code:** [github.com/HazyResearch/fonduer](https://github.com/HazyResearch/fonduer)