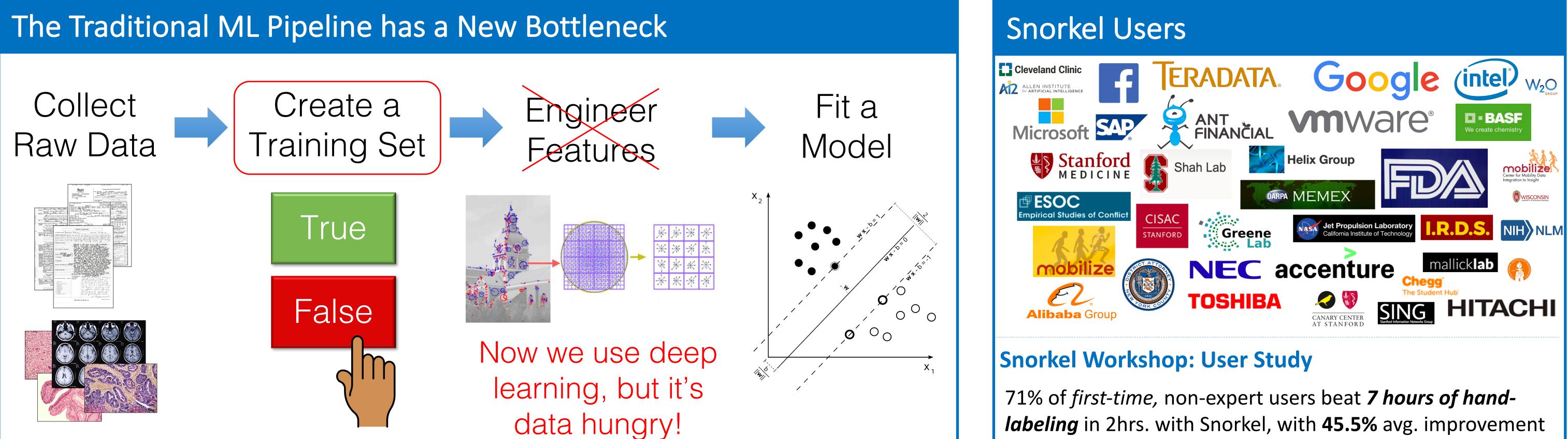
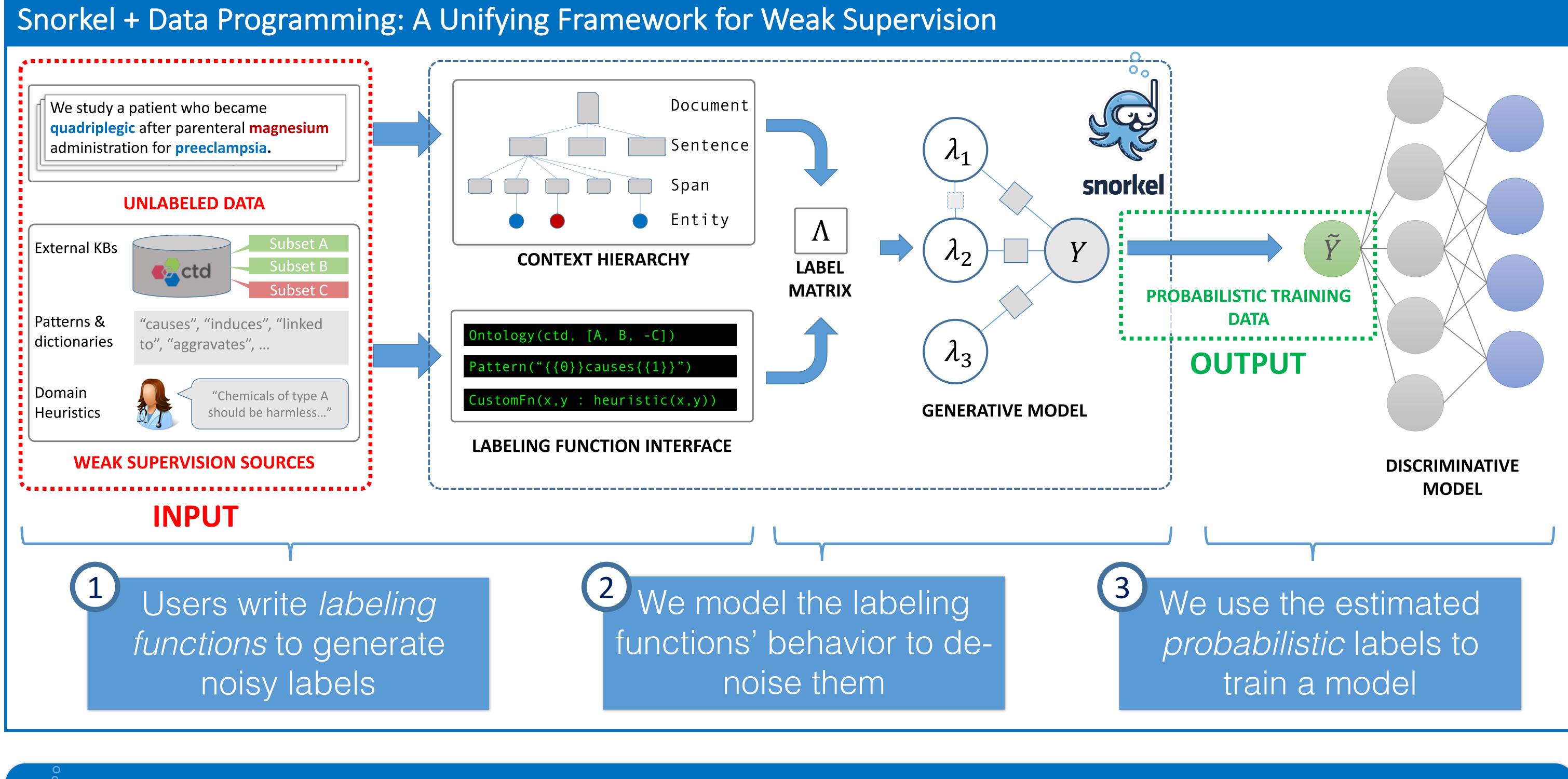


Snorkel: Rapid Training Set Creation with Weak Supervision Alex Ratner, Stephen Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Chris Ré







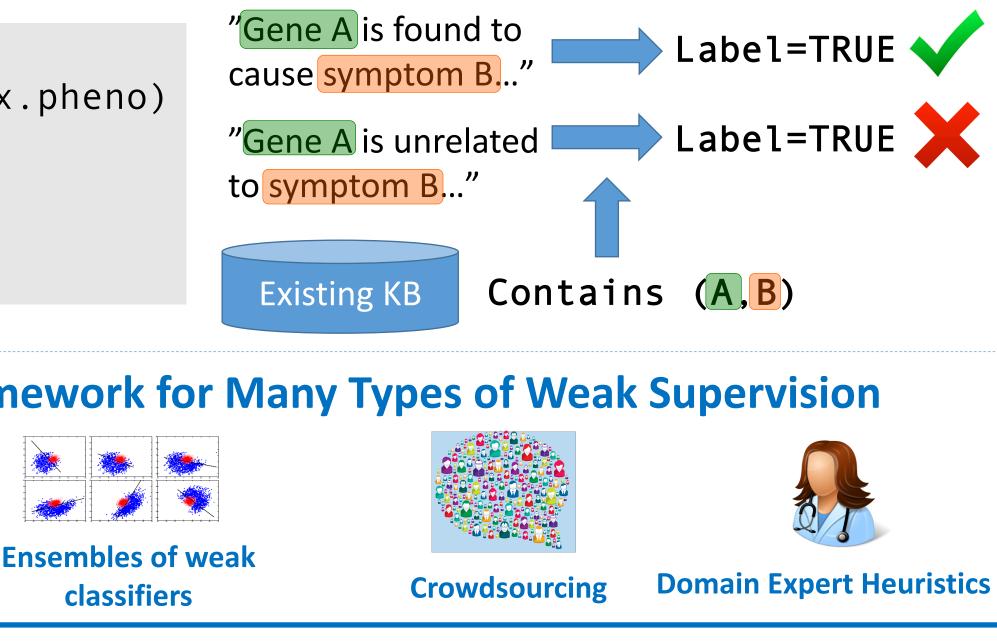
Open-source code and tutorials: snorkel.stanford.edu

Generating Noisy Labels with Labeling Functions

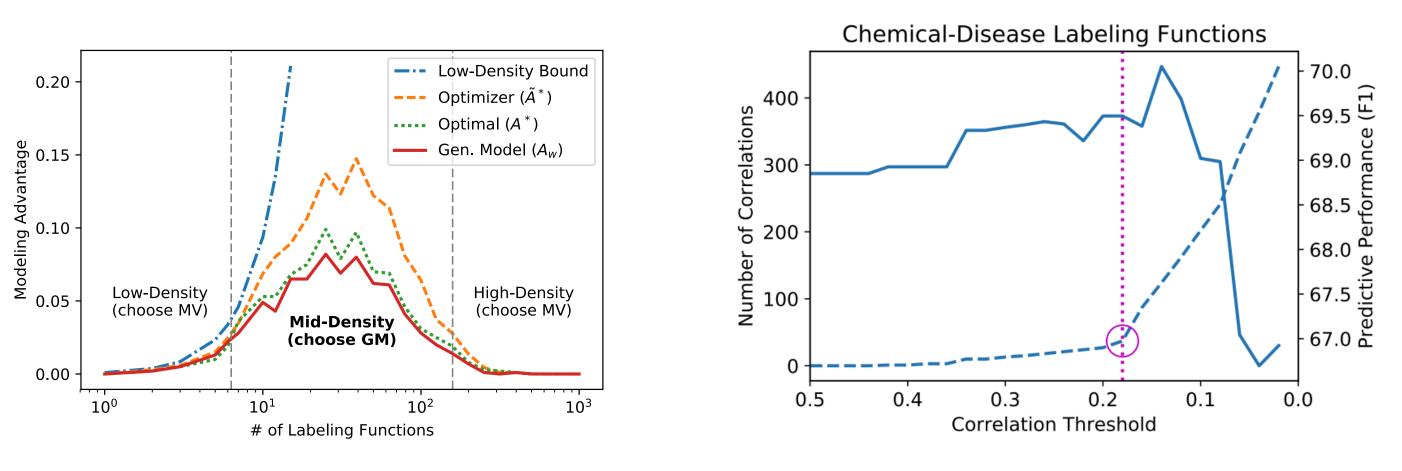
In Snorkel, users write *labeling functions (LFs)*, which are just scripts that noisily label subsets of the data. Ex: Labeling relations in text based on an existing knowledge base:

<pre>def lf1(x): cid = (x.gene, x. if cid in KB: return 1 else: return 0</pre>	p
A Unifying Frame	51

Distant supervision



Computational Tradeoffs Modeling Weak Supervision

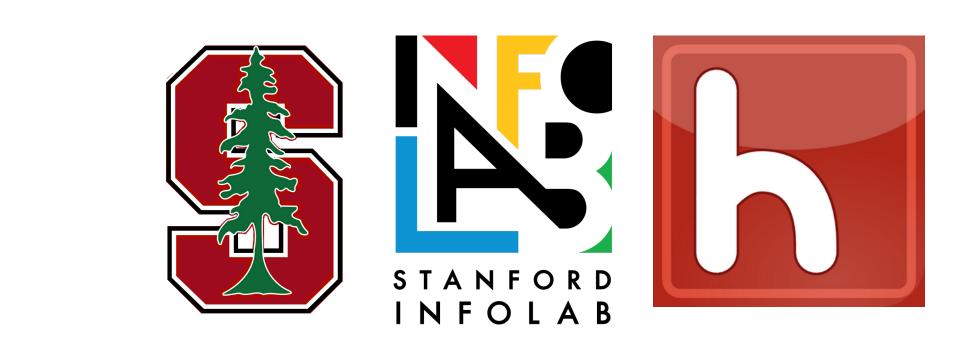


When to model the source accuracies: We provide an optimizer with theoretical guarantees that can speed up initial dev. cycles *How much structure to model:* Can speed up modeling by up to 10x!

New time-accuracy tradeoff space for modeling weak supervision:

Results on Real-World Text & Image Tasks

Application (Metric)	Distant Supervision (Baseline)	Snorkel- Gen. Model	Snorkel- Disc. Model [Unweighted]	Snorkel- Disc. Model	Hand-labeled training data
Chem (F1)	17.6	+ 16.2	+ 14.8	+ 5.5	N/A
EHR (F1)	72.2	+ 2.7	+ 6.0	+ 0.5	N/A
CDR (F1)	29.4	+ 9.1	+ 3.5	+ 3.3	+ 2.0
Radiology (AUC)	N/A	N/A	70.2	+ 1.3	+ 4.7
Crowd. (Acc)	N/A	N/A	62.5	+ 3.1	+ 3.2



With *tens* of LFs, we improve on DS / heuristic baselines, and come within points of expensive, hand-labeled datasets