Weakly Supervised Natural Language Understanding Models for Clinical Text

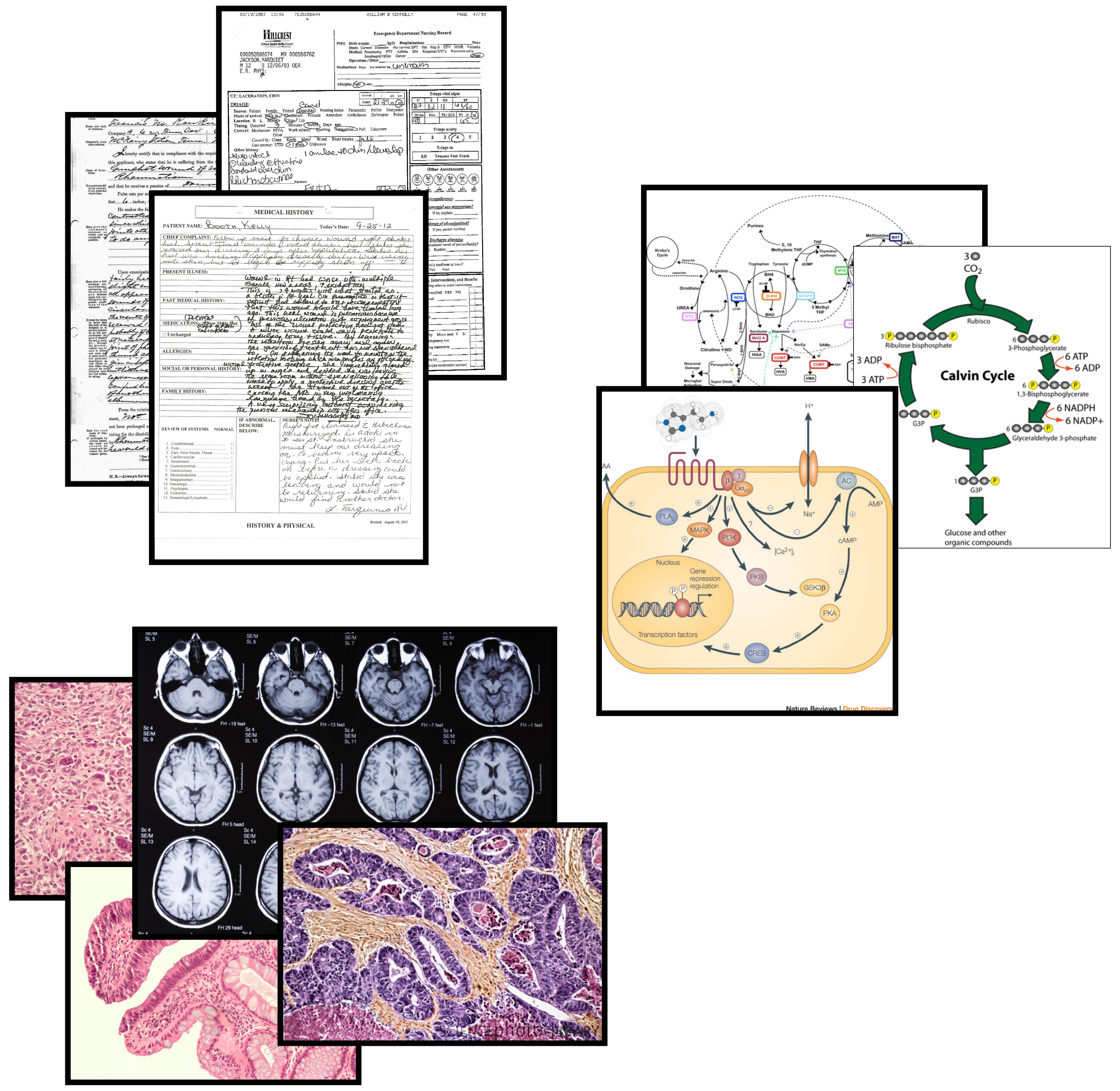
Jason Alan Fries, PhD
Research Scientist
Shah Lab, Stanford University

STANFORD DAWN Stanford MEDICINE Shah Lab BMIR Stanford Center for Biomedical Informatics Research

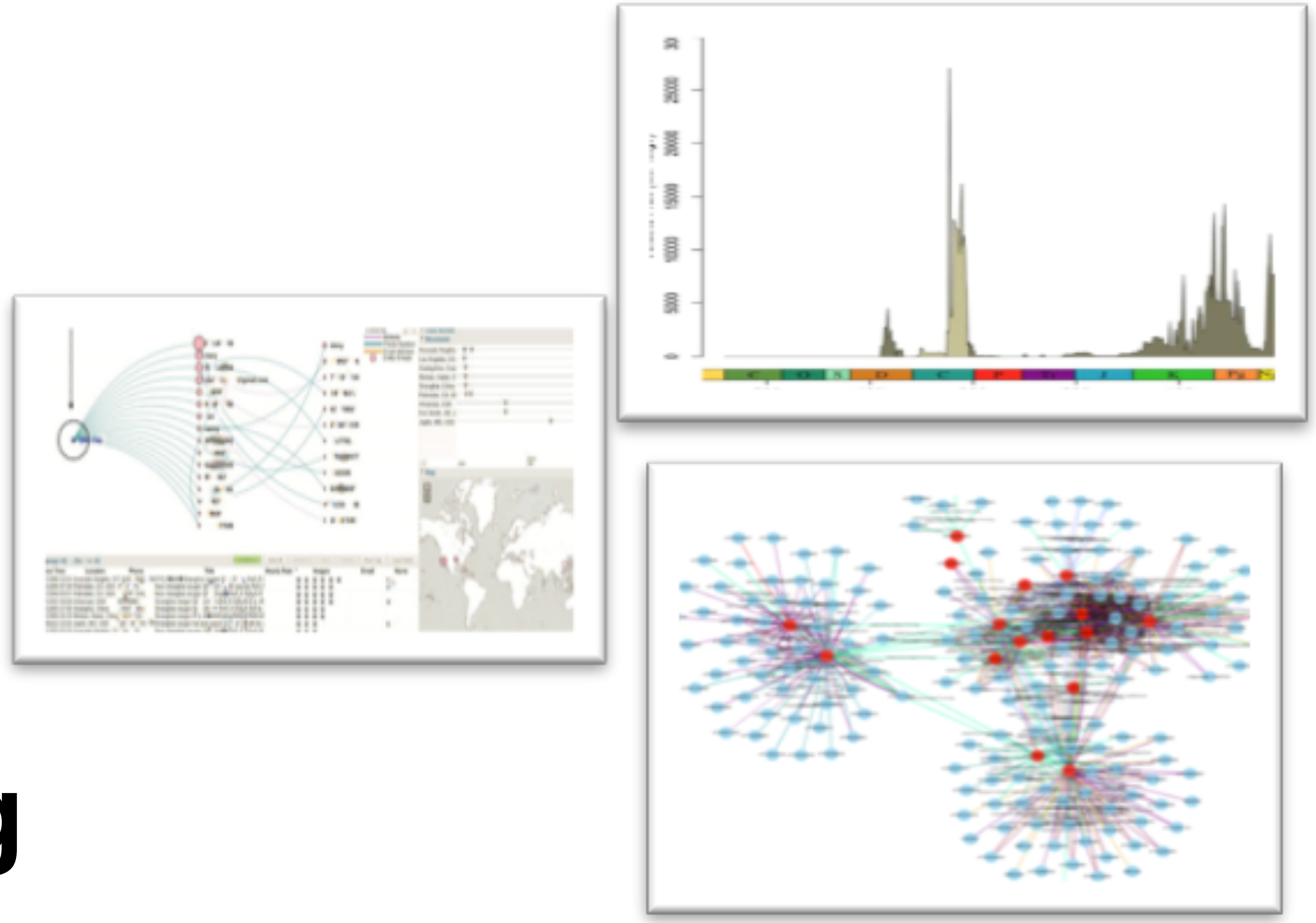
Outline

- Introduction: Snorkel & Programmatic Training Data
- Weakly Supervised Sequence Labeling for NLP
- Case Study: Medical Device Surveillance
- Closing Thoughts

Transforming Unstructured to Structured



Machine Learning

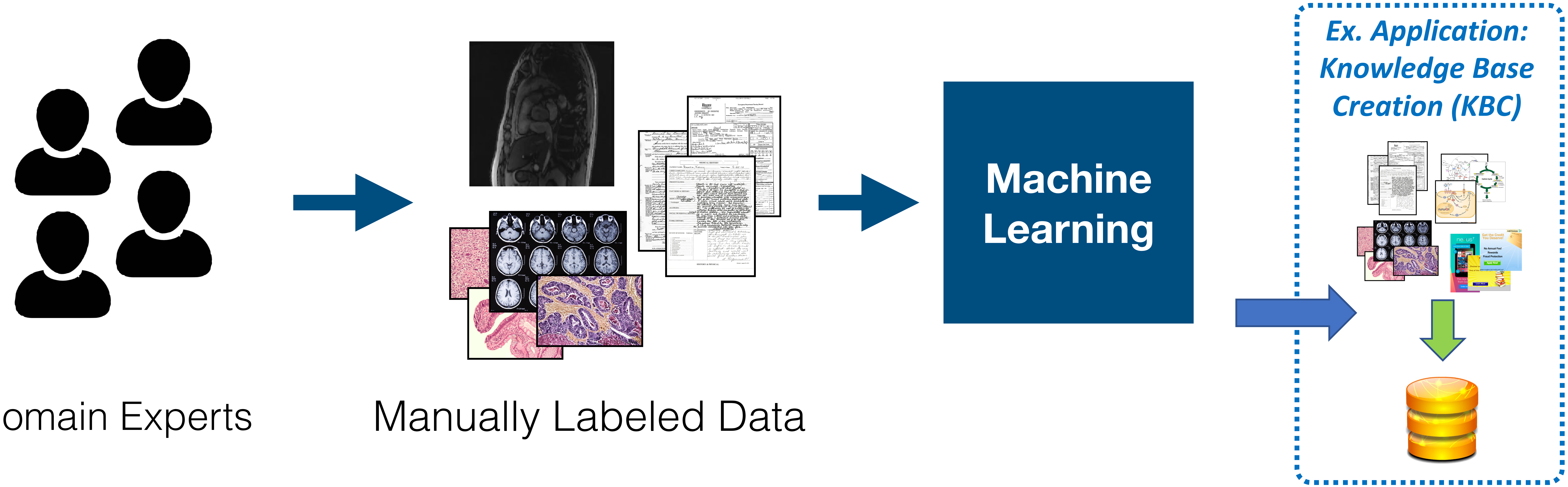


Dark Data: Text, Tables, Images, Diagrams, etc.

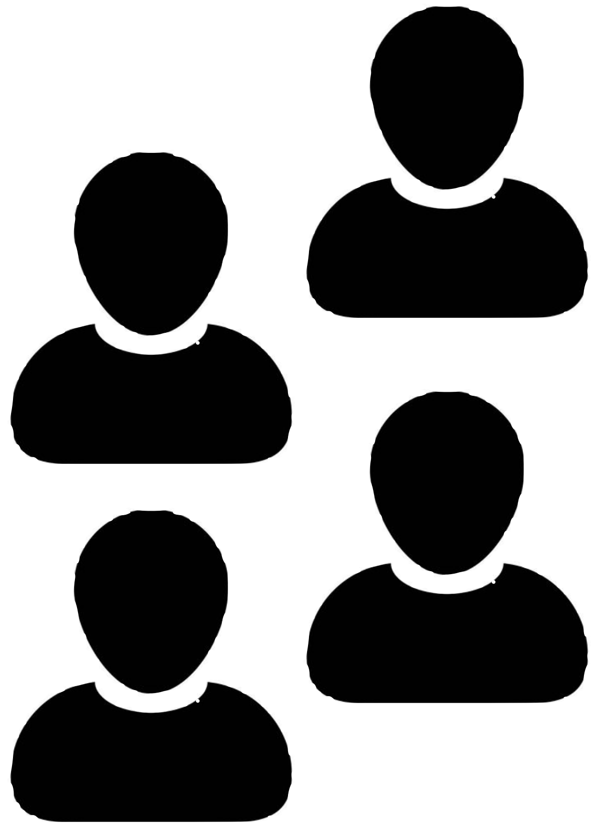
Structured Data: Enables analyses, interfaces, etc.

Need to transform data into machine readable form

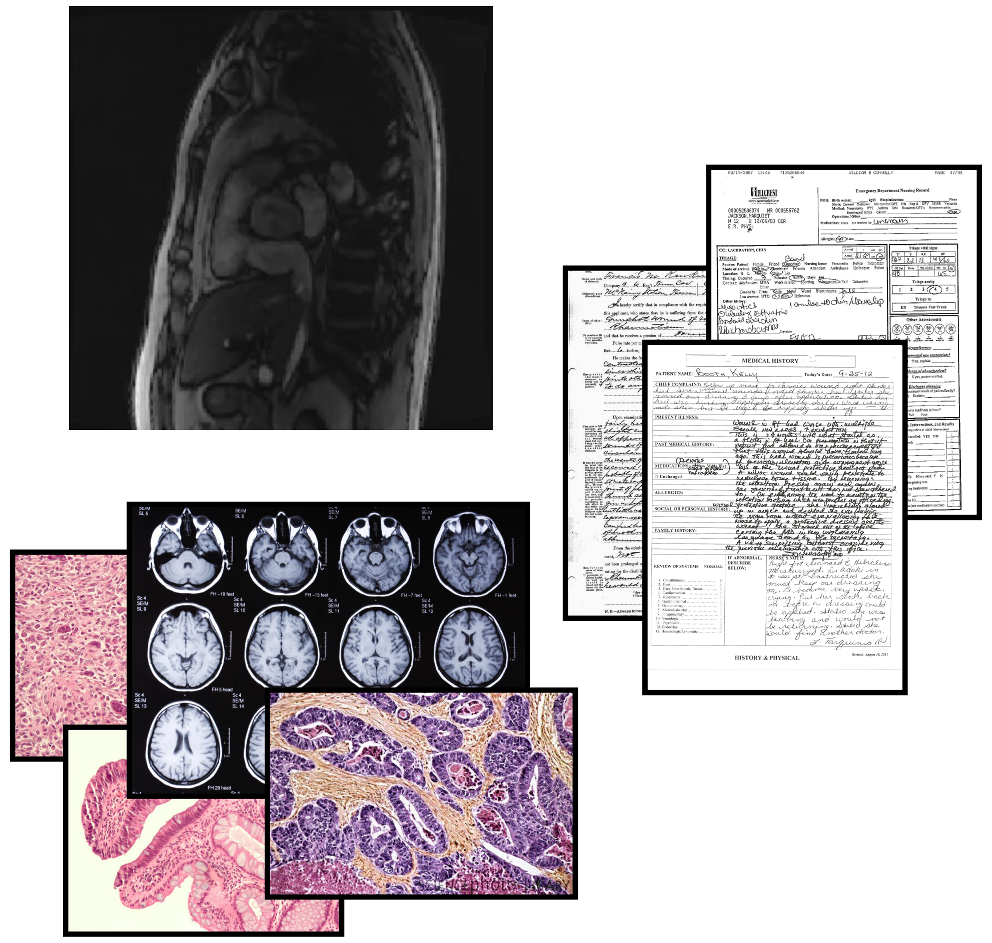
Standard Machine Learning Process



Standard Machine Learning Process



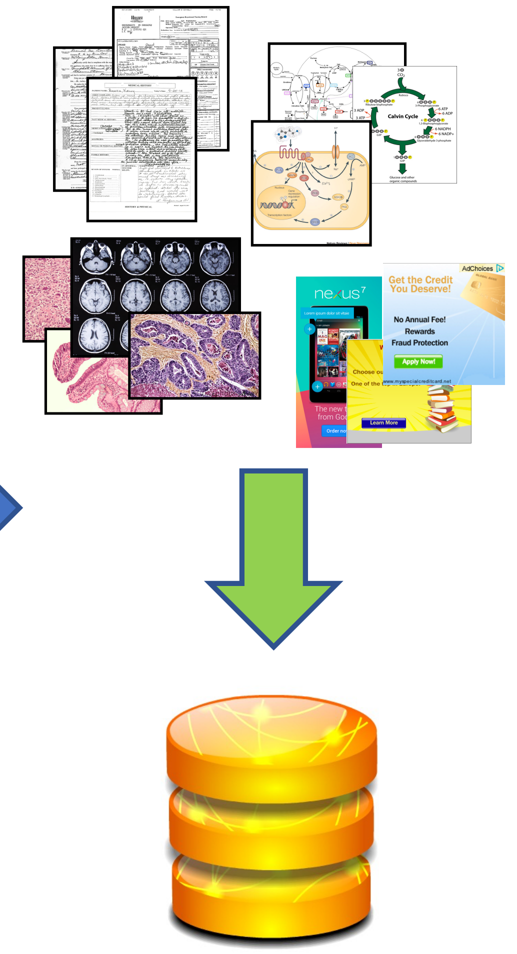
Domain Experts



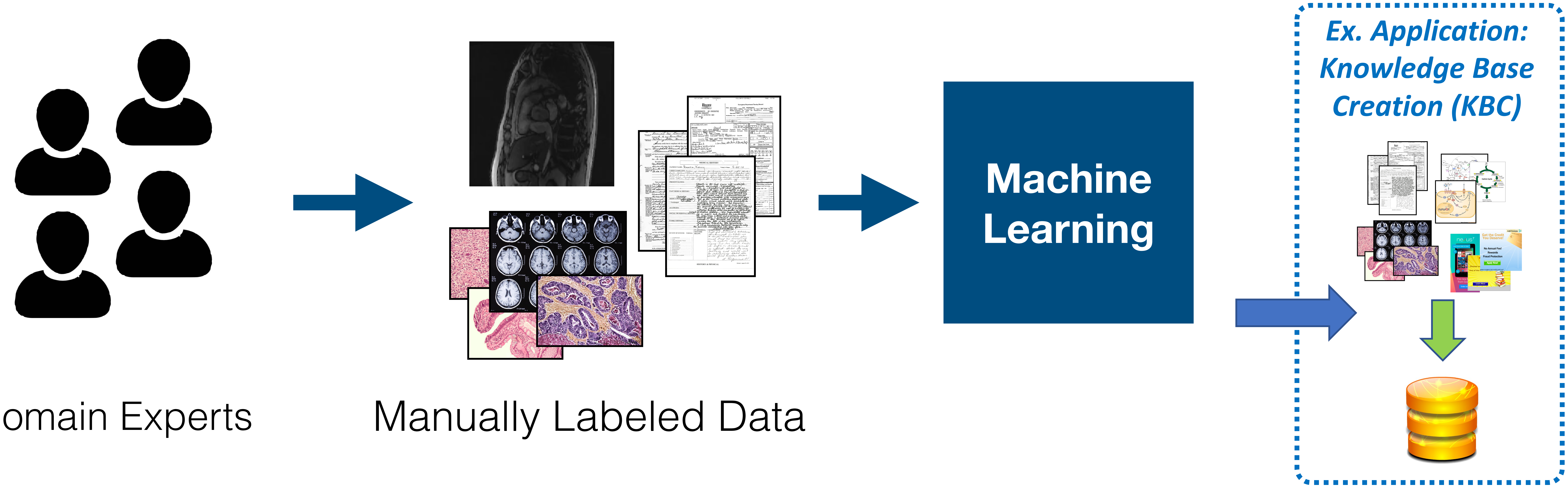
Manually Labeled Data

Machine Learning

Ex. Application: Knowledge Base Creation (KBC)

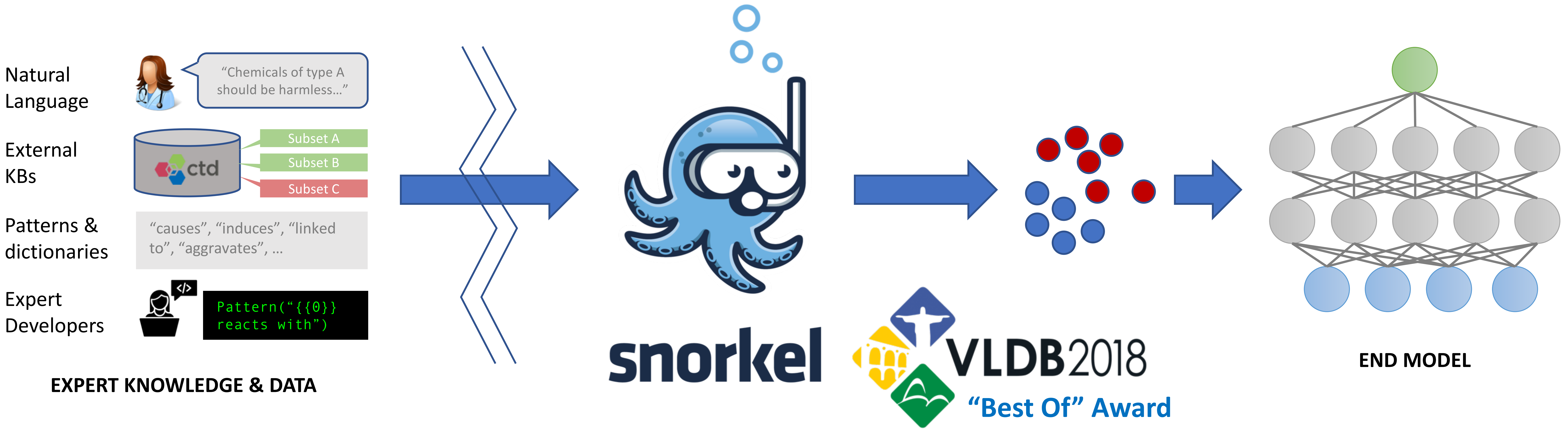


Standard Machine Learning Process



Building machine learning systems can take **months or years**!

Snorkel: (Ratner et al. 2017)
A System for Rapidly Creating Training Sets



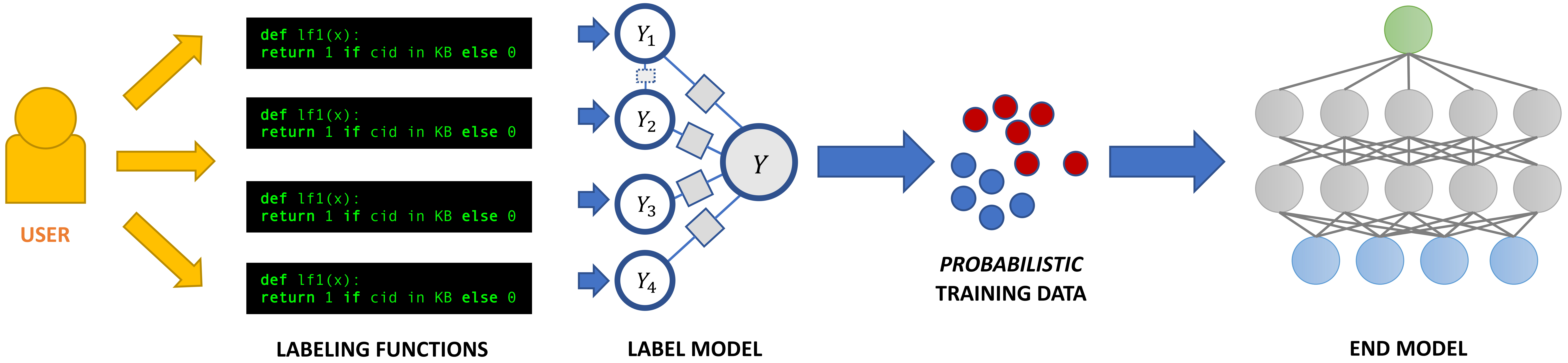
Program ML systems faster and easier

Slide credit: Alex Ratner

USERS & SPONSORS

Snorkel usage is growing in industry and research

The Snorkel Pipeline



1 Users write *labeling functions* to generate noisy labels

2 Snorkel models and combine these labels

3 We use the resulting *probabilistic* training labels to train a model

Key point: Input is *labeling functions*– ***No hand-labeled training sets***

Key Concepts: Labeling Functions

Labeling Functions (LFs)

Black box functions that label subsets of data

`{-1, 0, 1}`

`{Negative, Abstain, Positive}`

Key Concepts: Labeling Functions

His father died secondary to **prostate cancer** and mother had Alzheimer's .

prostate cancer ∈ SNOMED CT **AND** STY == ‘Neoplastic Process’

prostate cancer —> **DISORDER**

Check membership in a knowledge base/ontology

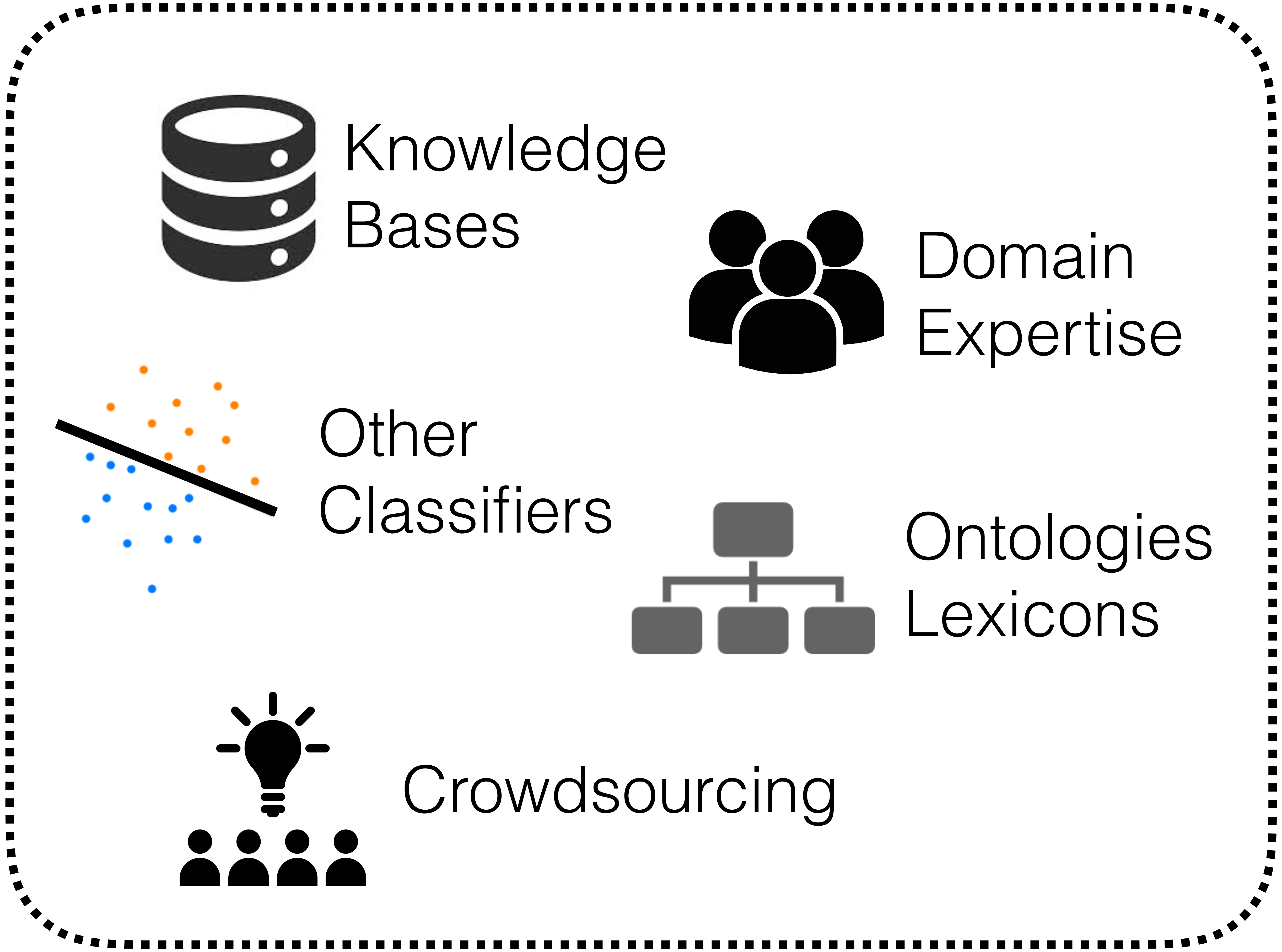
Key Concepts: Labeling Functions

His **father** died secondary to **prostate cancer** and mother had Alzheimer's .

```
def LF_is_a_relative(span):
    rgx = re.compile(r'''\b((grand)*(mother|father)|grand(m|p)a|
(parent|(daught|sist|broth)er|son|cousin)(s)*)\b''', re.I)
    text = get_left_span(span, window=6).text
    return FAMILY if rgx.search(text) else ABSTAIN
```

Match regular expression rules

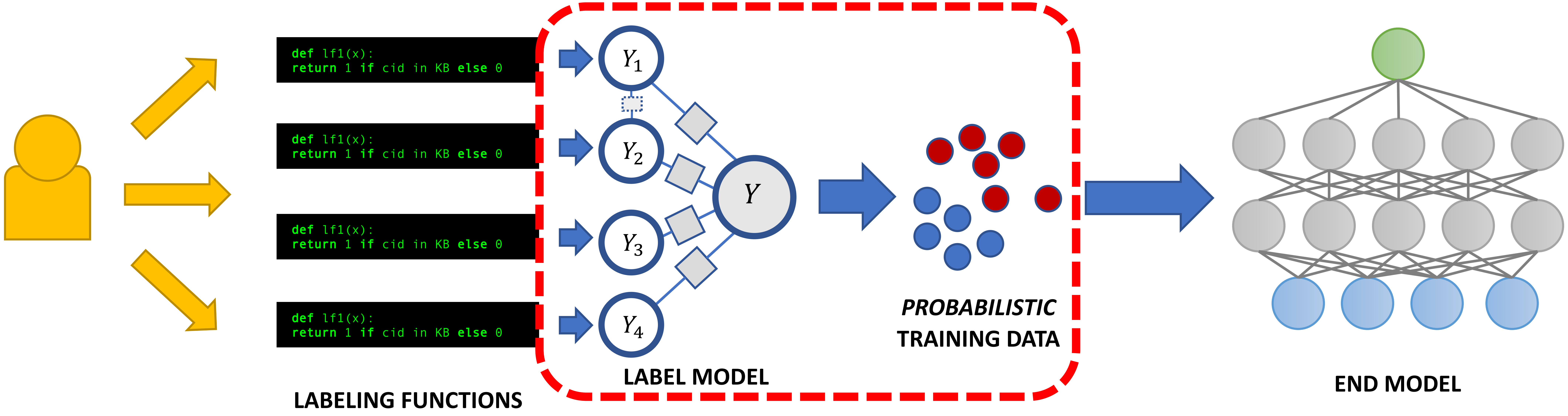
Key Concepts: Labeling Functions



Labeling functions provide a **unified interface** for label sources

Allows us to combine sources and model aspects like *accuracy* and *statistical dependencies* **without hand-labeled data**

How do we model and combine LFs?

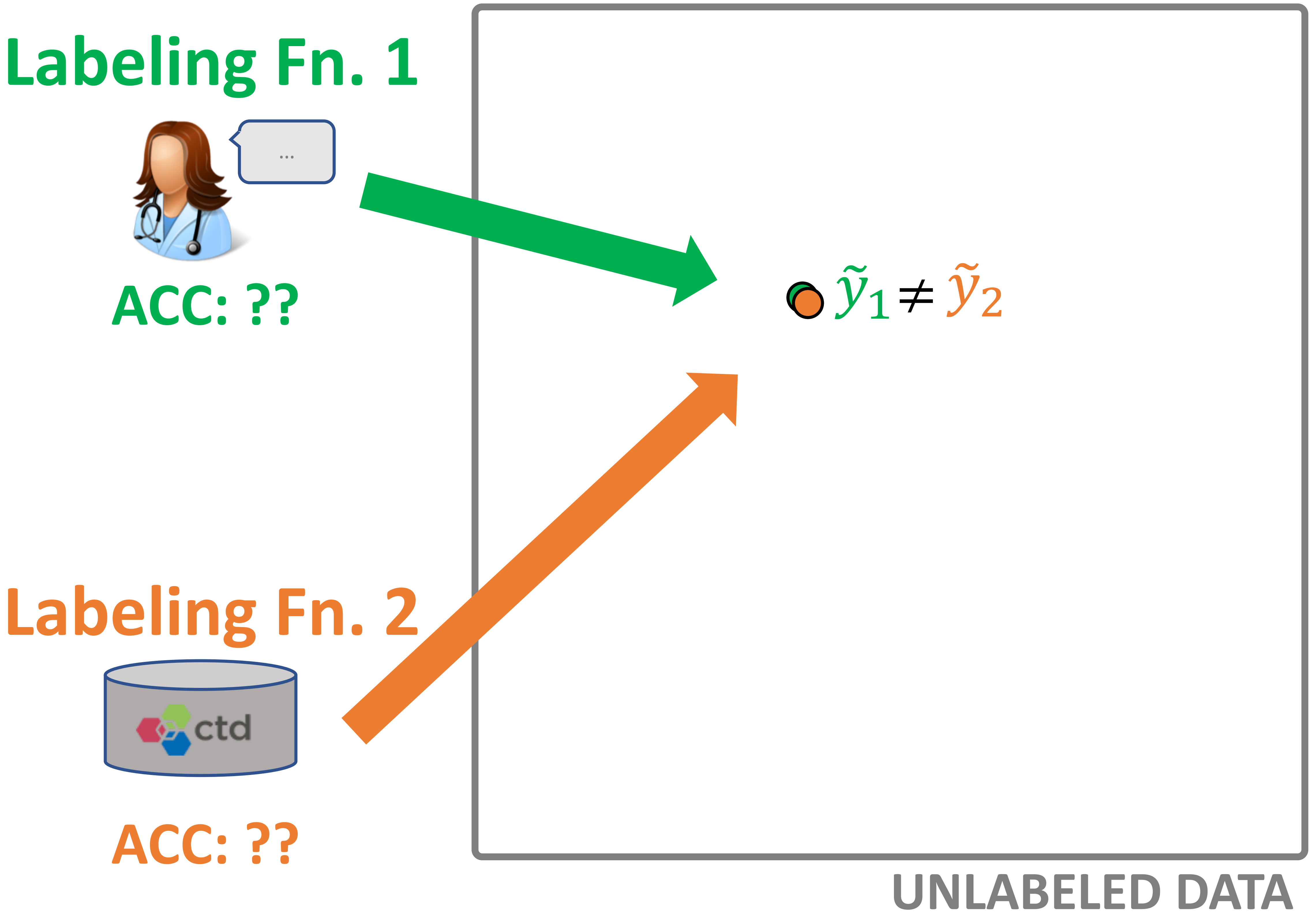


Key Technical Challenge: How to best reweight and combine the noisy supervision signal?

Challenges of Weak Supervision

- Problem 1: How do we resolve conflicts between weak label sources?
  - How can we estimate their accuracies without ground truth?

- This is a real development burden that our users faced with prior “distant supervision” systems

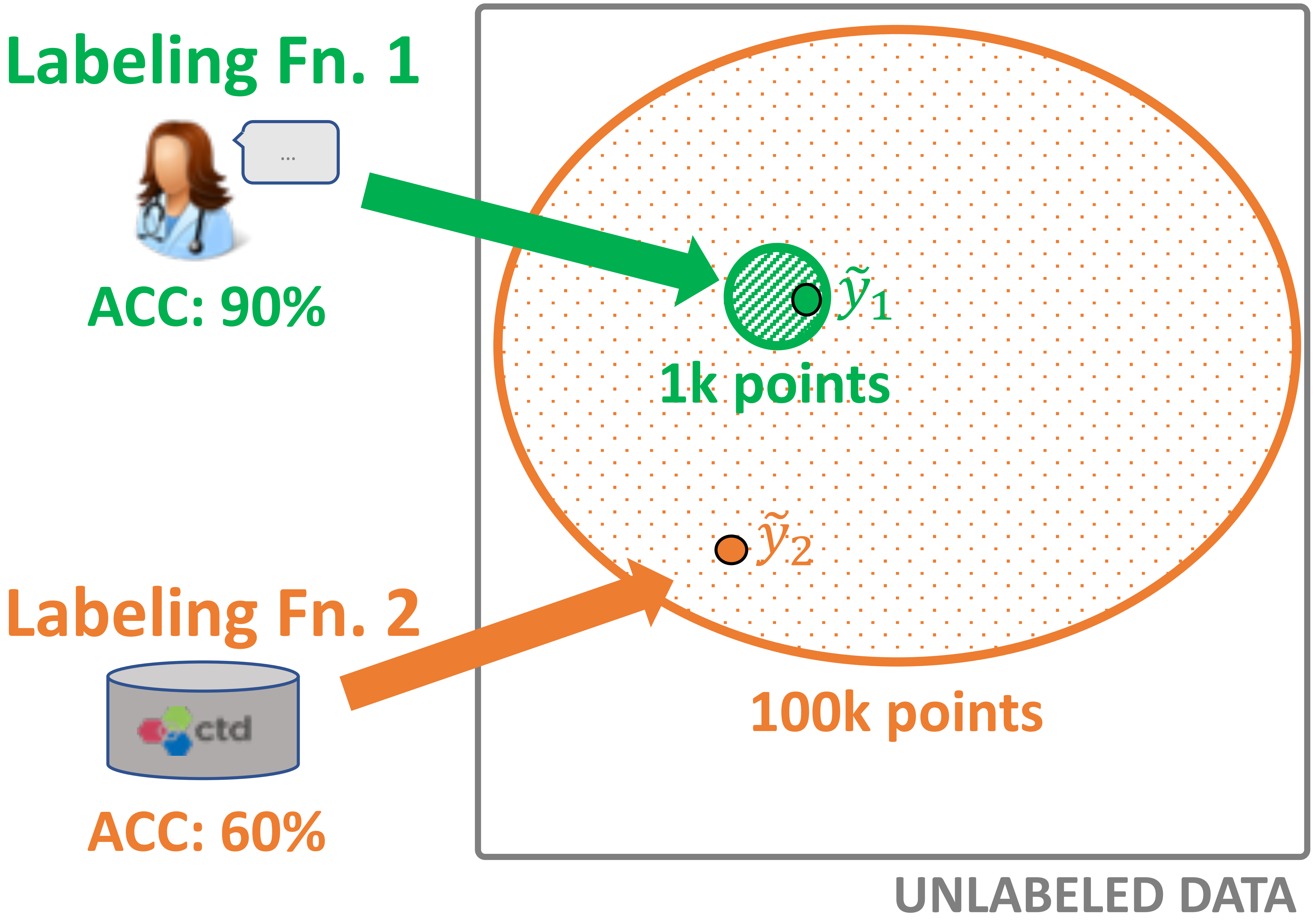


Need to be able to estimate source accuracies

Challenges of Weak Supervision

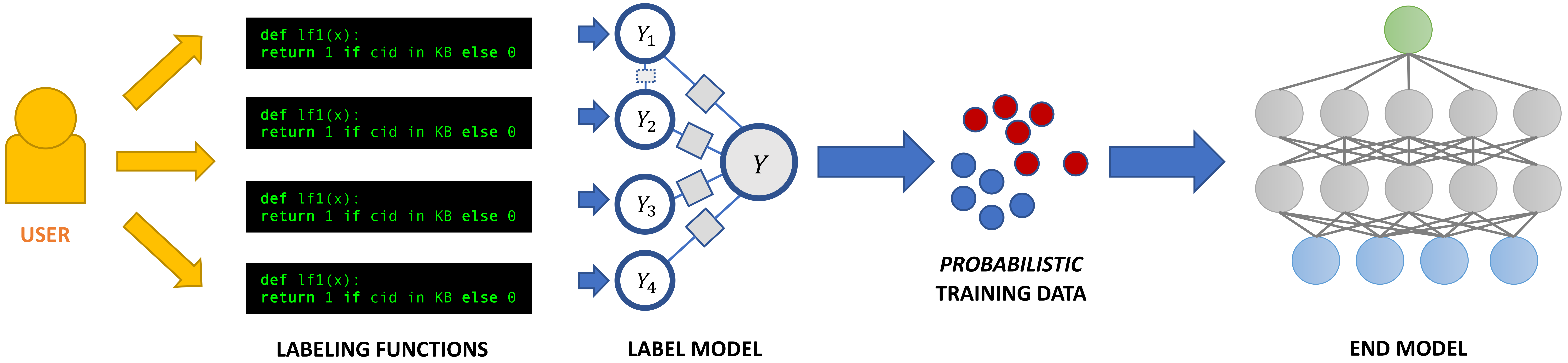
- Problem 2: Need to communicate training point lineage to model being trained

- Ex:
  - User writes one high-accuracy, low-coverage LF…
  - …and one low-accuracy, high-coverage LF
  - *If we just naively take the union of labels, expected acc. = 60.3%!*



Need to communicate training label *lineage*

The Snorkel Pipeline



1 Users write *labeling functions* to generate noisy labels

2 Snorkel models and combine these labels

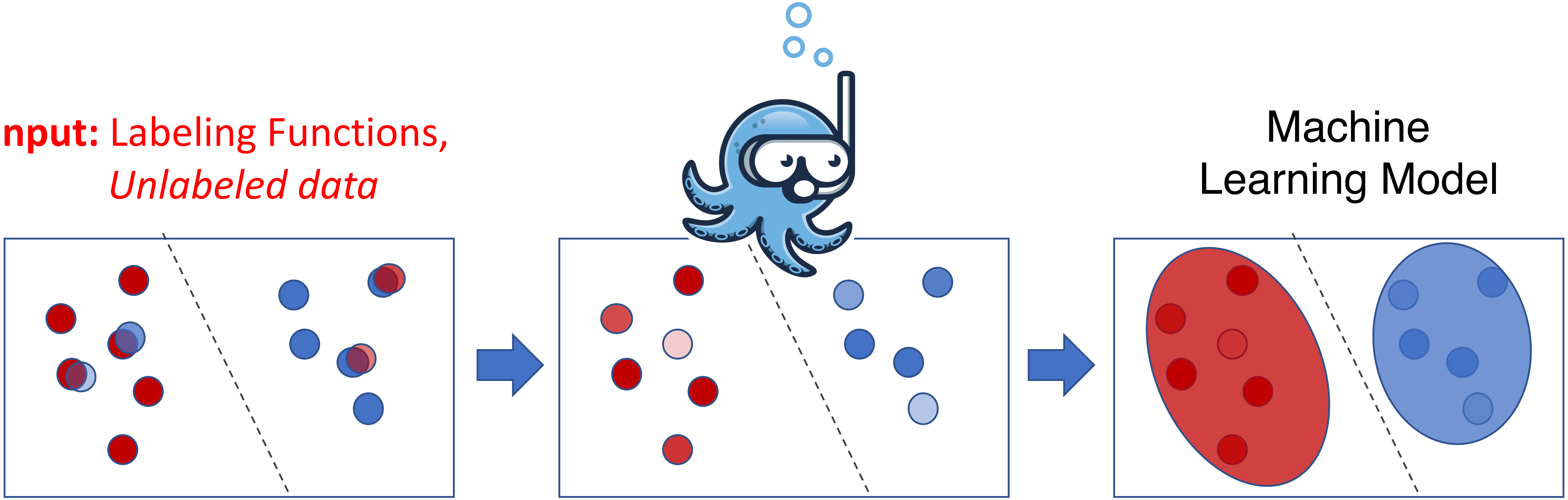
3 We use the resulting *probabilistic* training labels to train a model

Key point: Input is *labeling functions*– ***No hand-labeled training sets***

Generalize Beyond Labeling Functions

Input: Labeling Functions,
Unlabeled data

Machine
Learning Model



Noisy, conflicting labels

Resolve conflicts,
re-weight & combine

Generalize beyond the
labeling functions

Weakly Supervised Sequence Labeling for NLP

Many NLP Tasks Are Sequence Labeling Problems

His	father	died	secondary	to	prostate	cancer	and	mother	had	Alzheimer's	.
O	O	O	O	O	I	I	O	O	O	I	O

Named Entity Recognition

Many NLP Tasks Are Sequence Labeling Problems

His	father	died	secondary	to	prostate	cancer	and	mother	had	Alzheimer's	.
O	O	O	O	O	I	I	O	O	O	I	O

Named Entity Recognition

Building labeled training sets
for these style of tasks is very expensive

UMLS-based Labeling Functions

Let's look at named entity recognition for **disorders**

Map Semantic Types to Classes

Positive	Negative
disease_or_syndrome	manufactured_object
neoplastic_process	intellectual_product
injury_or_poisoning	body_location_or_region
sign_or_symptom	virus
pathologic_function	functional_concept
anatomical_abnormality	…
…	

Create LFs for k Source Vocabularies

- **Consumer Health Vocabulary (CHV)**
- **SNOMED CT**
- **Medical Subject Headings (MSH**

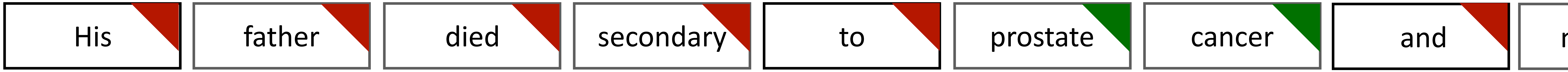
Sequence Labeling with Weak Supervision

His father died secondary to prostate cancer and mother had Alzheimer's .

I (Inside) O (Outside)

IO Disorder Tagging

Example: Apply 5 labeling functions (LFs) to a sentence

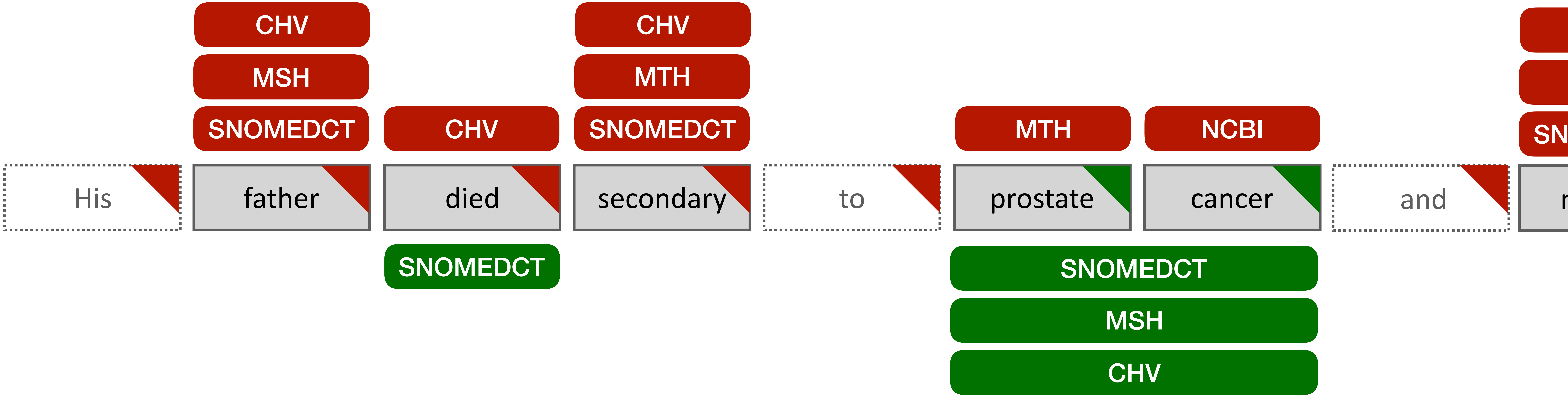


Sequence Labeling with Weak Supervision

His father died secondary to prostate cancer and mother had Alzheimer's .

I (Inside) O (Outside)

IO **Disorder Tagging**

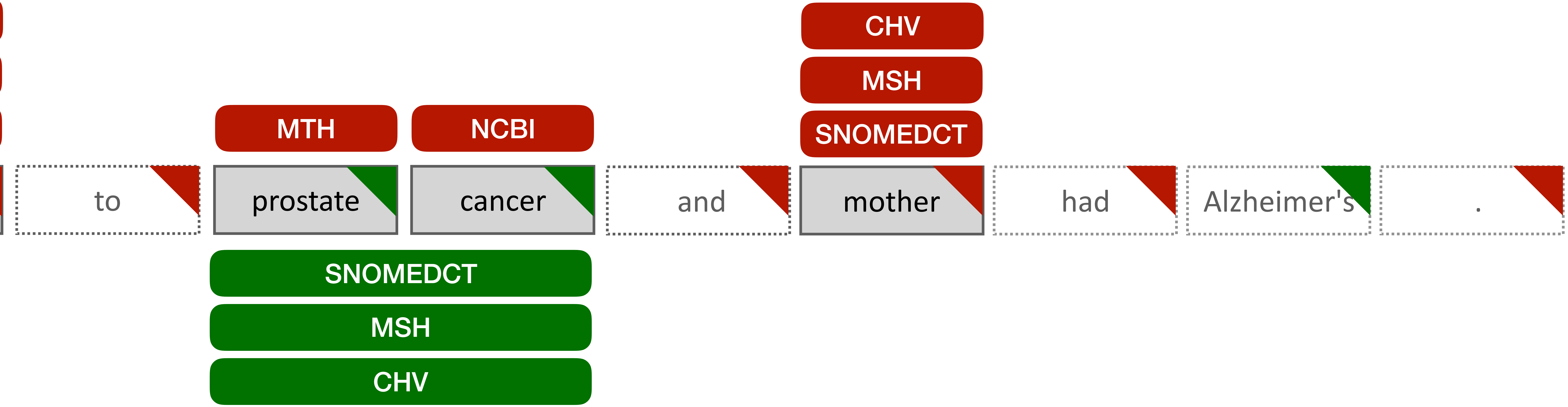


Sequence Labeling with Weak Supervision

His father died secondary to prostate cancer and mother had Alzheimer's .

I (Inside) O (Outside)

IO **Disorder Tagging**

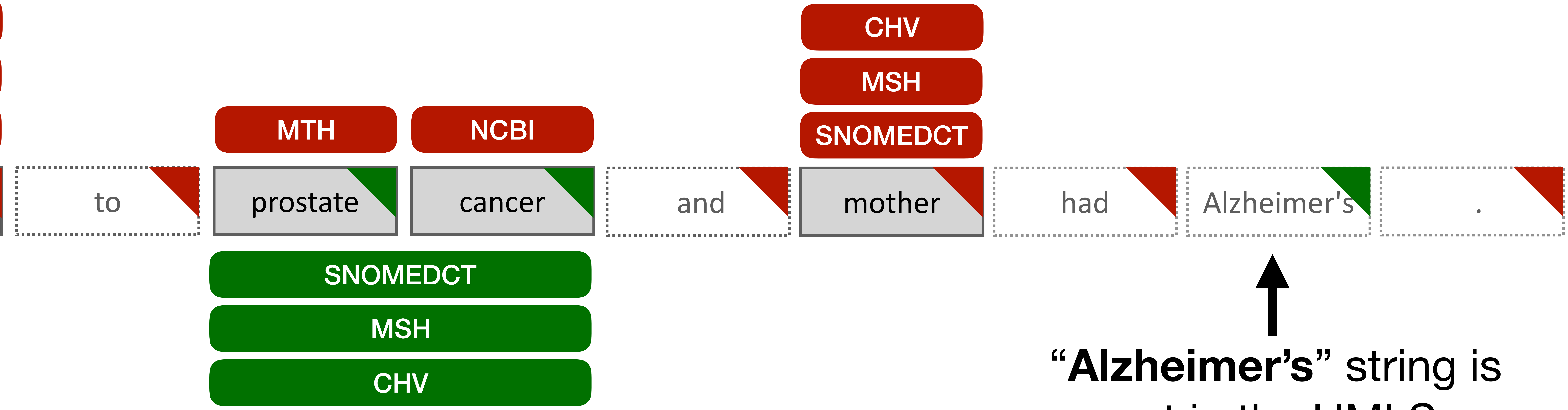


Sequence Labeling with Weak Supervision

His father died secondary to prostate cancer and mother had Alzheimer's .

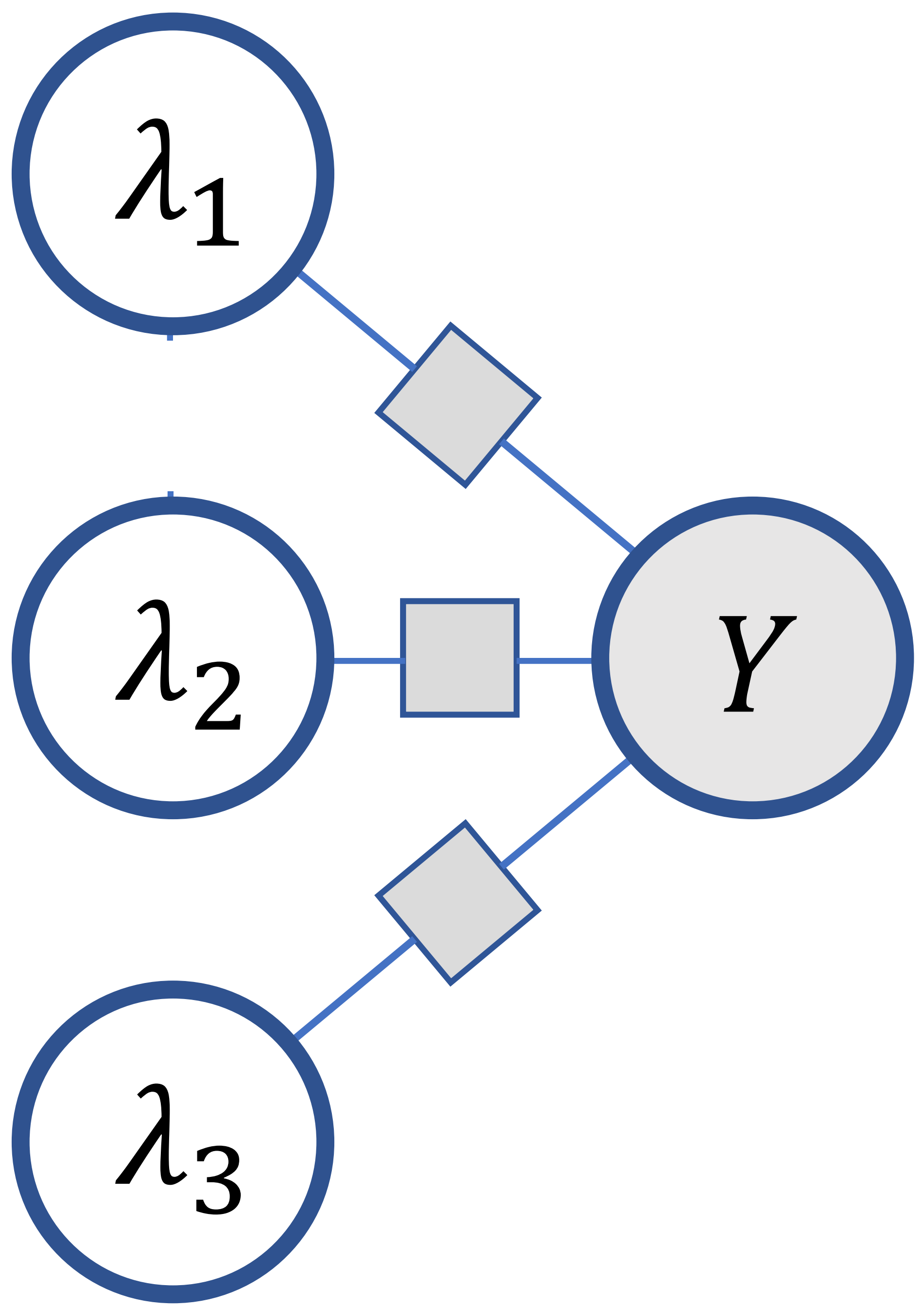
I (Inside) O (Outside)

IO **Disorder Tagging**



“**Alzheimer’s**” string is not in the UMLS

Sequence Labeling with Weak Supervision



Factor Graph-based Label Model

$\lambda_1, \ldots, \lambda_n$	Labeling functions
m	Words
$\mathbf{\Lambda} \in \{-1, 0, 1\}^{m \times n}$	Label matrix
$\mathbf{Y} := y_1, \ldots, y_m$	True label (unobserved)

$$\phi_j^{Acc}(\Lambda_i, y_i) := y_i \Lambda_{ij}$$

$$p_\theta(\mathbf{\Lambda}, \mathbf{Y}) \propto \exp\left(\sum_{i=1}^{m} \sum_{j=1}^{n} \theta_j^{Acc} \phi_j^{Acc}(\Lambda_i, y_i)\right)$$

Sequence Labeling with Weak Supervision

His father died secondary to prostate cancer and mother had Alzheimer's .

I (Inside) O (Outside)

IO **Disorder Tagging**

Label model output

(1) Probabilistic label per-word
(2) A mask of uncovered words

X	His	father	died	secondary	to	prostate	cancer	and	mother	had	Alzheimer's	.
Y	0	0	0	0	0	1	1	0	0	0	1	0
$\hat{Y}$	-	0.01	0.35	0.01	-	0.85	0.95	-	0.01	-	-	-
Mask	0	1	1	1	0	1	1	0	1	0	0	0

Weakly-labeled Training Set

X	His	father	died	secondary	to	prostate	cancer	and	mother	had	Alzheimer's	.
Y	0	0	0	0	0	1	1	0	0	0	1	0
$\hat{Y}$	-	0.01	0.35	0.01	-	0.85	0.95	-	0.01	-	-	-
Mask	0	1	1	1	0	1	1	0	1	0	0	0

This ***weakly-labeled training set*** can be used with many end models by using a **simple noise-aware loss function**

$$\hat{w} = argmin_w \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{y \sim \hat{Y}} \left[L(w, x_i, y) \right]$$

LSTM

GPT2

BERT

RoBERTa

Logistic Regression

End Model Generalization

Powerful ***representation learning*** algorithms allow us to generalize beyond our labeling function output

X	His	father	died	secondary	to	prostate	cancer	and	mother	had	Alzheimer's	.
Y	0	0	0	0	0	1	1	0	0	0	1	0
$\hat{Y}$	-	0.01	0.35	0.01	-	0.85	0.95	-	0.01	-	-	-
Pred	0.00	0.00	0.25	0.01	0.00	0.90	0.95	0.00	0.00	0.00	**0.85**	0.00

End model provides predictions for uncovered words

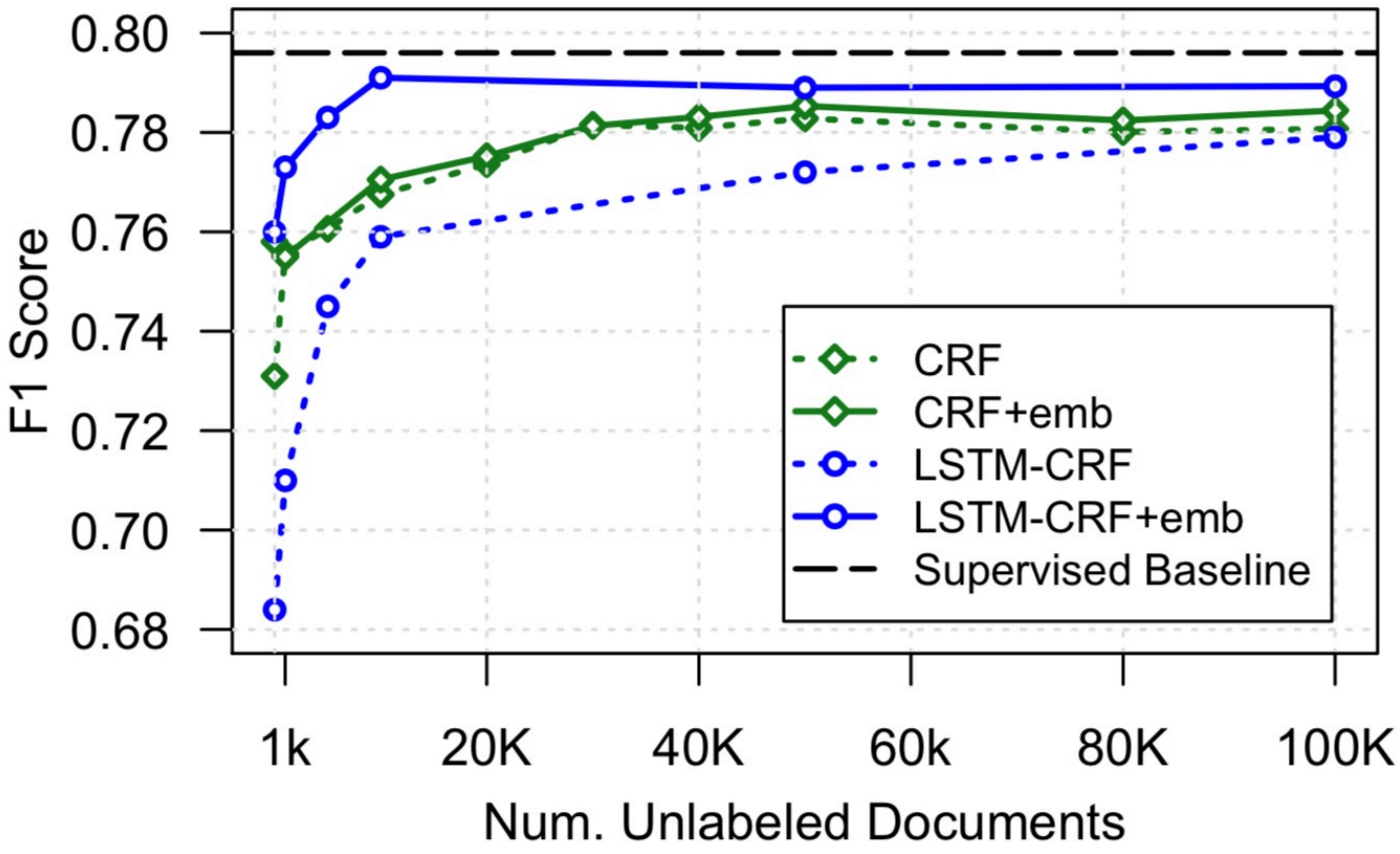
i2b2 Medication Challenge (2009)

Model	# Train Docs	P	R	F1	Diff.
Expert-labeled + LSTM	124	**90.4**	**88.5**	**89.4**	**-**
Lexicon (UMLS)	-	31.9	67.6	43.3	**-52%**
Amazon Comprehend Medical (Aug. 2019)	?	69.4	79.9	74.3	**-17%**
Snorkel (UMLS) + LSTM	1000	82.2	74.7	78.3	**-12%**
Snorkel (UMLS + Manual LFs) + LSTM	1000	83.9	82.9	83.4	**-7%**

Weakly supervised models score within **7-12%** of supervised baseline

Test Set: 125 expert-labeled docs

Theory Benefit: Scaling with Unlabeled Data



PubMed Disease Tagging (Fries et al. 2017)

Log-linear performance improvements with unlabeled data

In (Bach et al. 2019), matched performance of models trained on **12 - 80k hand-labeled instances** at Google.

Google AI

Clinical Text Sequence Labeling Tasks

Named Entities

Disorders (CLEF)
Drugs (i2b2)

We have labeling functions for all these benchmark tasks
(3 clinical NLP datasets)

Attributes

Temporality (THYME) $\in$ `{before, before_overlaps, overlaps, after}`
Negation (THYME, CLEF) $\in$ `{positive, negative}`
BodyLocation (CLEF) $\in$ `{CUIs}`
Experiencer (CLEF) $\in$ `{patient, other}`

Case Study: Medical Device Surveillance

Learning from unlabeled electronic health records for medical device surveillance

Alison Callahan, BMIR
Jason A. Fries, Stanford CS/BMIR
Chris Ré, Stanford CS
Scott Delp, Stanford Bioengineering
Nicholas J Giori, Stanford Medicine, Palo Alto VA
James I Huddleston, Stanford Medicine
Nigam Shah, BMIR

Early Failure of Implants is Very Expensive

Metal-on-metal hip implants

13% failure rate within 5 years
expected rate is **0.5%**!

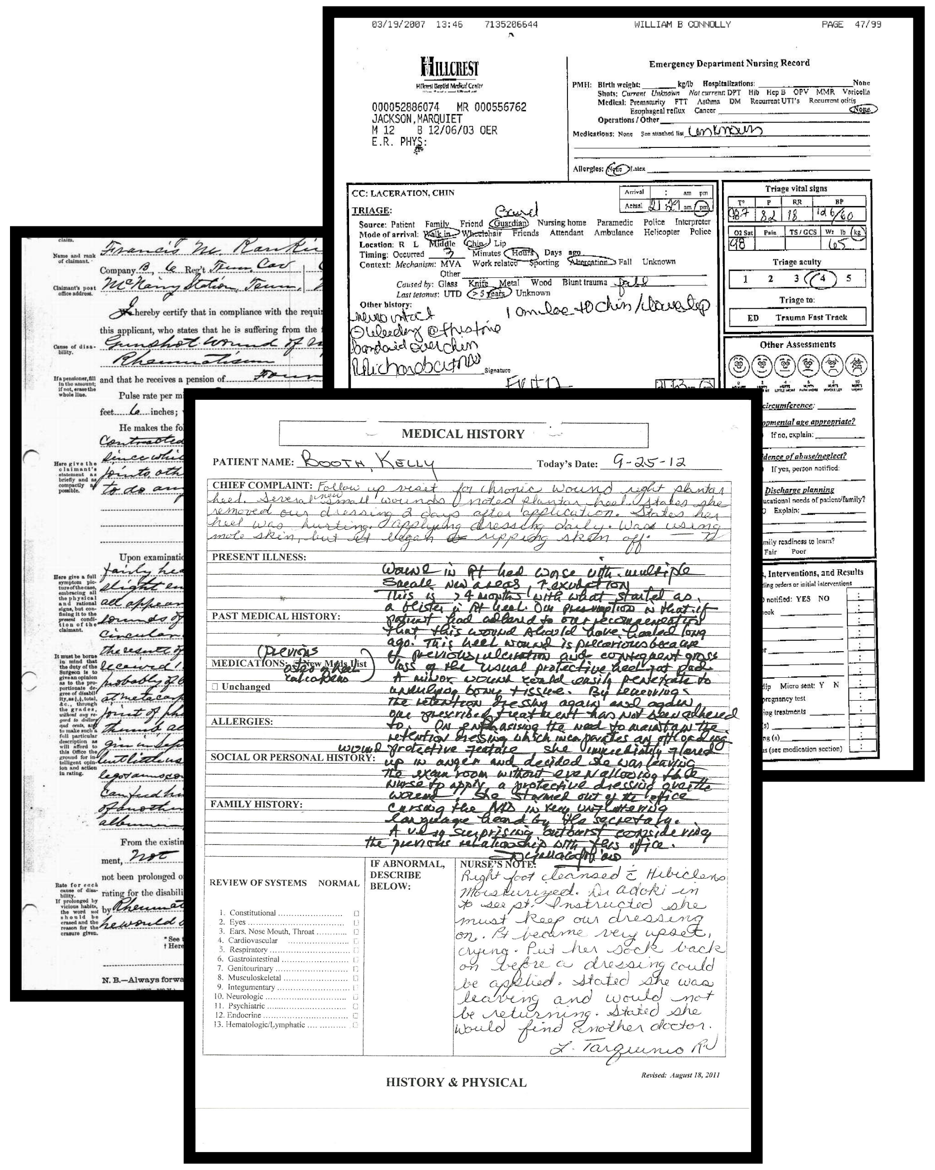
$4 Billion Dollars
in legal settlements

On the market for ~5 years before issuing a recall

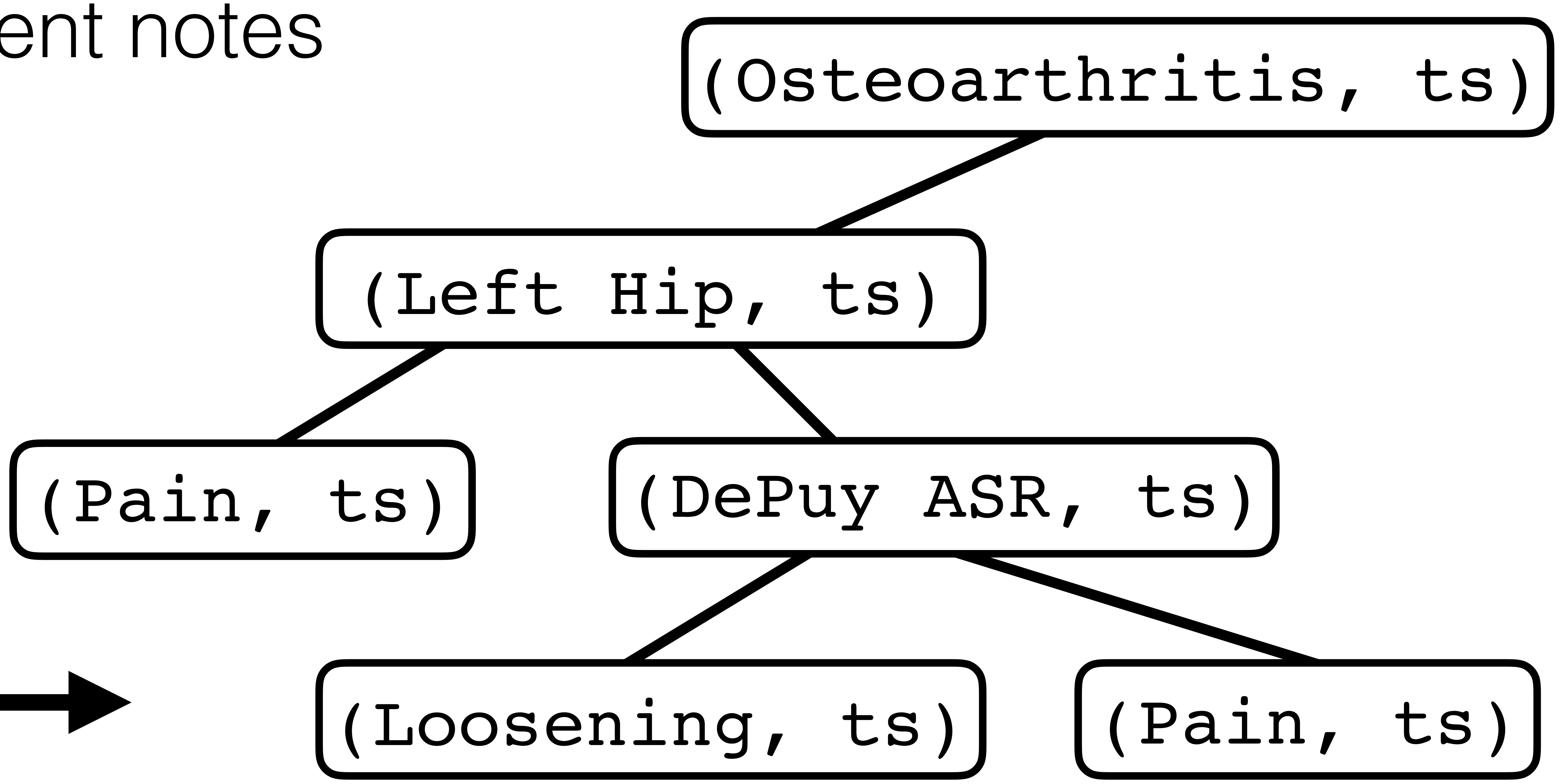
We need faster strategies for evaluating devices

Automating Medical Device Surveillance with EHRs

Treat this as a ***knowledge base construction*** task using patient notes



Transform Patient Notes
into Structured Data



Orthopedic Devices
(hip replacements)

Extracting Implant-related Complications

Given removal of all **hardware**, chances of clearance of **infection** with IV antibiotics improves significantly.

THE **ACETABULAR COMPONENT** HAS BEEN **REPLACED**, AND THE FEMORAL HEAD APPEARS WELL CENTERED.

There is also a **lucency** surrounding the **right acetabular cup** which has increased since prior film suggestive of osteolysis.

The **hip** was **dislocated** with a hook, and the trial femoral component head and neck were removed.

(this is from a surgical procedure — not a complication!)

Extracting Implant-related Complications

- INFECTION
- COMPONENT WEAR
- PAIN
- REVISION
- RADIOGRAPHIC ABNORMALITY
- MECHANICAL FAILURE

Complication Types

→

Let's train a **relational inference model** to to link these to specific implants

IMPLANT TYPE

Binary classification over sentences w/ two arguments

> There is also a **lucency** surrounding the **right acetabular cup** wh
> suggestive of osteolysis.

Dataset

1.7 million patients

Stanford HEALTH CARE

6,583 patients

Primary THA and/or revision surgery

500k Notes

Expert Labeled Data

60 patient notes

233 mentions

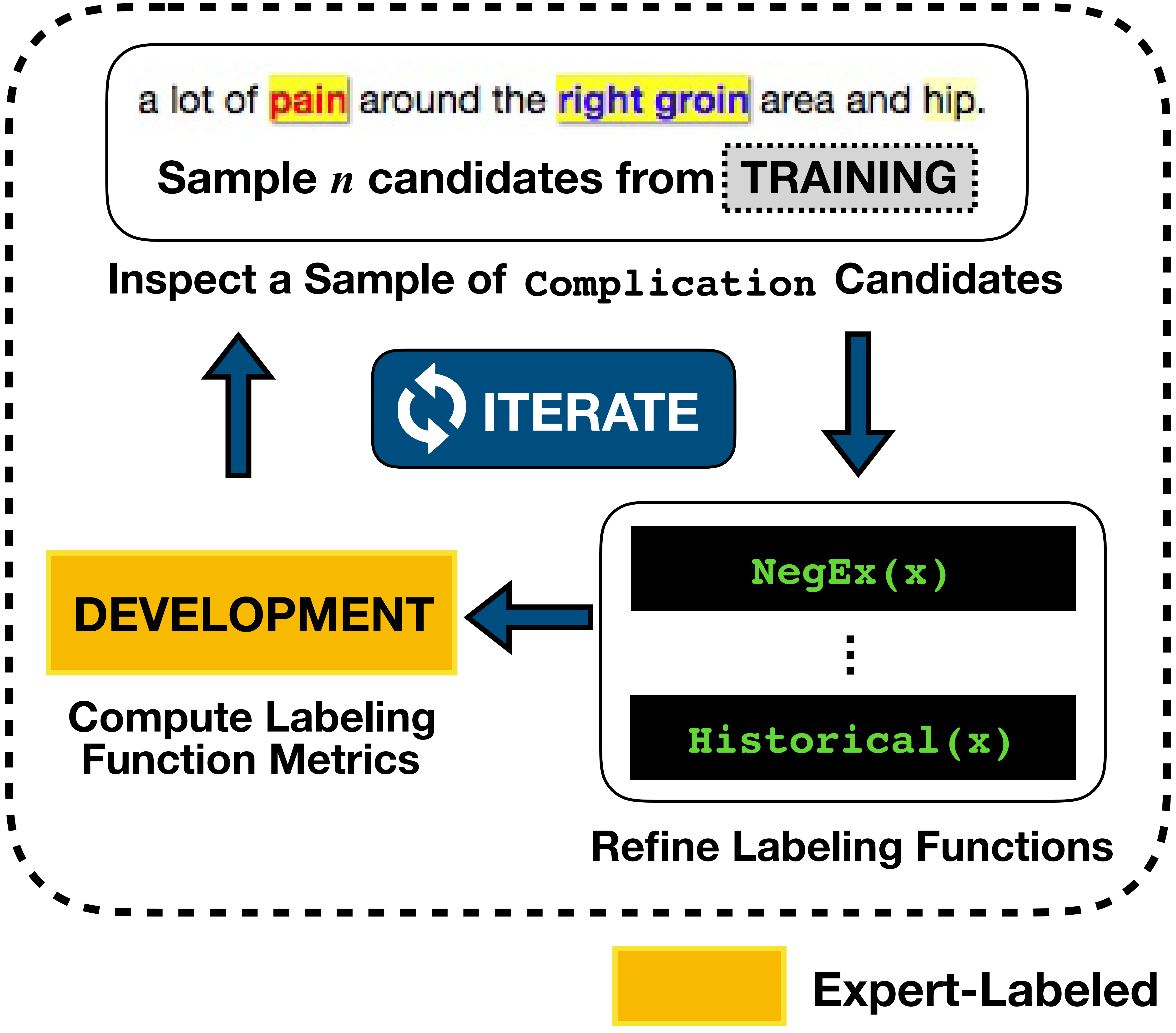
5 clinical annotators

DEVELOPMENT

TEST

30 notes each

Developing Labeling Functions



Iteratively tune labeling functions by examining unlabeled data

Clinical Note Markup

HISTORY OF PRESENT ILLNESS:
60 yo male with infected R hip (MRSA) s/p previous hip replacement.

LTHA November 2004 demonstrates component wear.

No lucencies were observed around the implant.

Implant is being evaluated for possible revision.

PAST MEDICAL HISTORY:
Hx right Zimmer Biomet hip 1/1/05 complicated by infection.

NOTE DATE: 07/01/2008 06:11 PM

Clinical Note Markup

HISTORY OF PRESENT ILLNESS:

60 yo male with infected R hip (MRSA) s/p previous hip replacement.
HISTORICAL

LTHA November 2004 demonstrates component wear.
HISTORICAL >2 YEARS

No lucencies were observed around the implant.
NEGATED

Implant is being evaluated for possible revision.
HYPOTHETICAL

PAST MEDICAL HISTORY:

Hx right Zimmer Biomet hip 1/1/05 complicated by infection.
>2 YEARS HISTORICAL

NOTE DATE: 07/01/2008 06:11 PM
0 DAYS

ENTITIES: HEADER | CLINICAL CONCEPT | DATETIME

ATTRIBUTES: HYPOTHETICAL | HISTORICAL | NEGATED | TIME DELTA

Labeling Function Examples

```
def LF2_historical(c):
    v = has_historical_attrib(c)
    return FALSE if v else ABSTAIN
```

```
def LF3_reject_section(c):
    h1 = get_section_header(c)
    v = h1 in reject_headers
    return FALSE if v else ABSTAIN
```

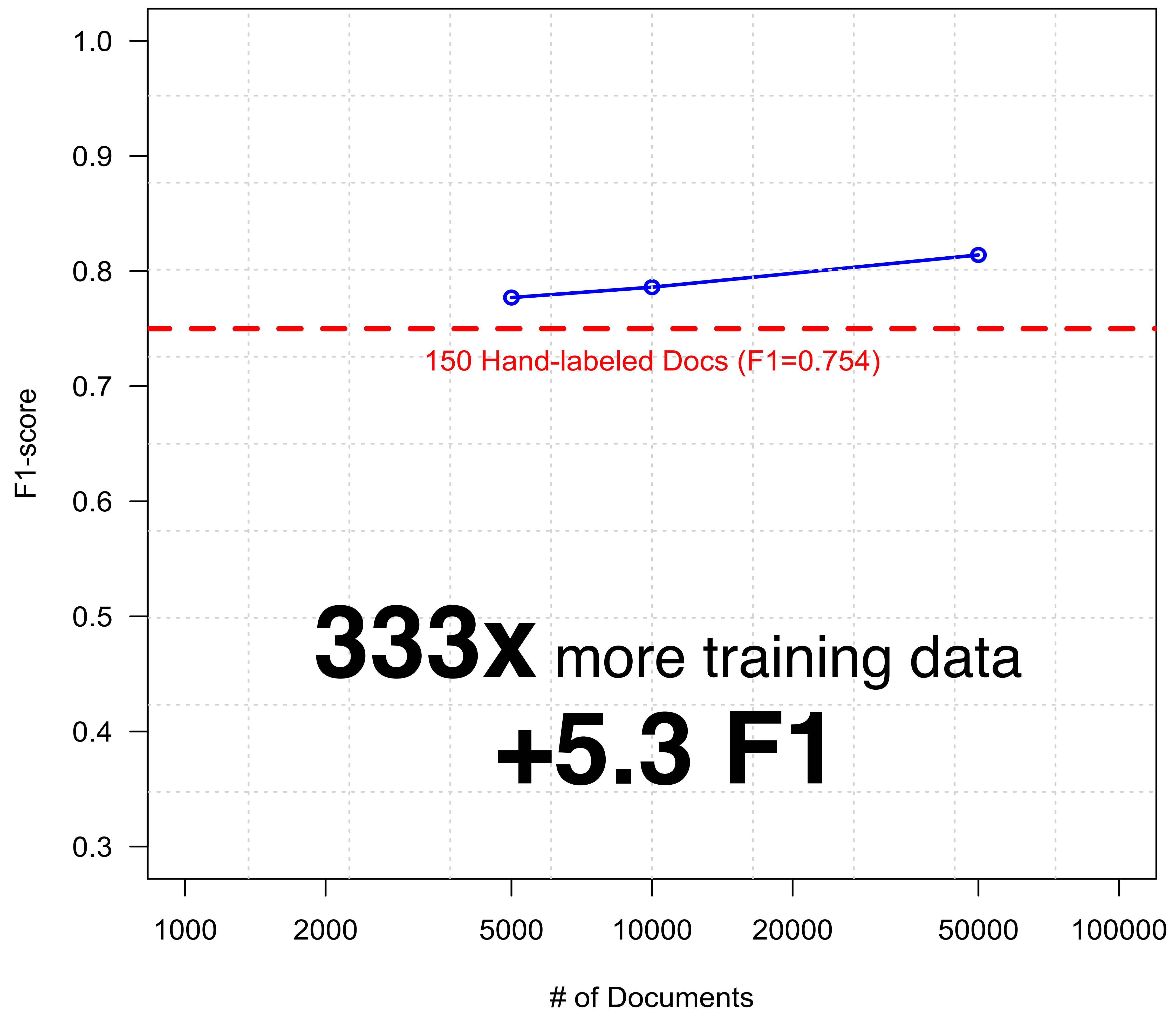
```
def LF4_negated(c):
    v = NegEx.is_negated(c)
    return FALSE if v else ABSTAIN
```

```
FALSE: -1   ABSTAIN: 0   TRUE: 1
```

Shared structure makes writing labeling functions easier

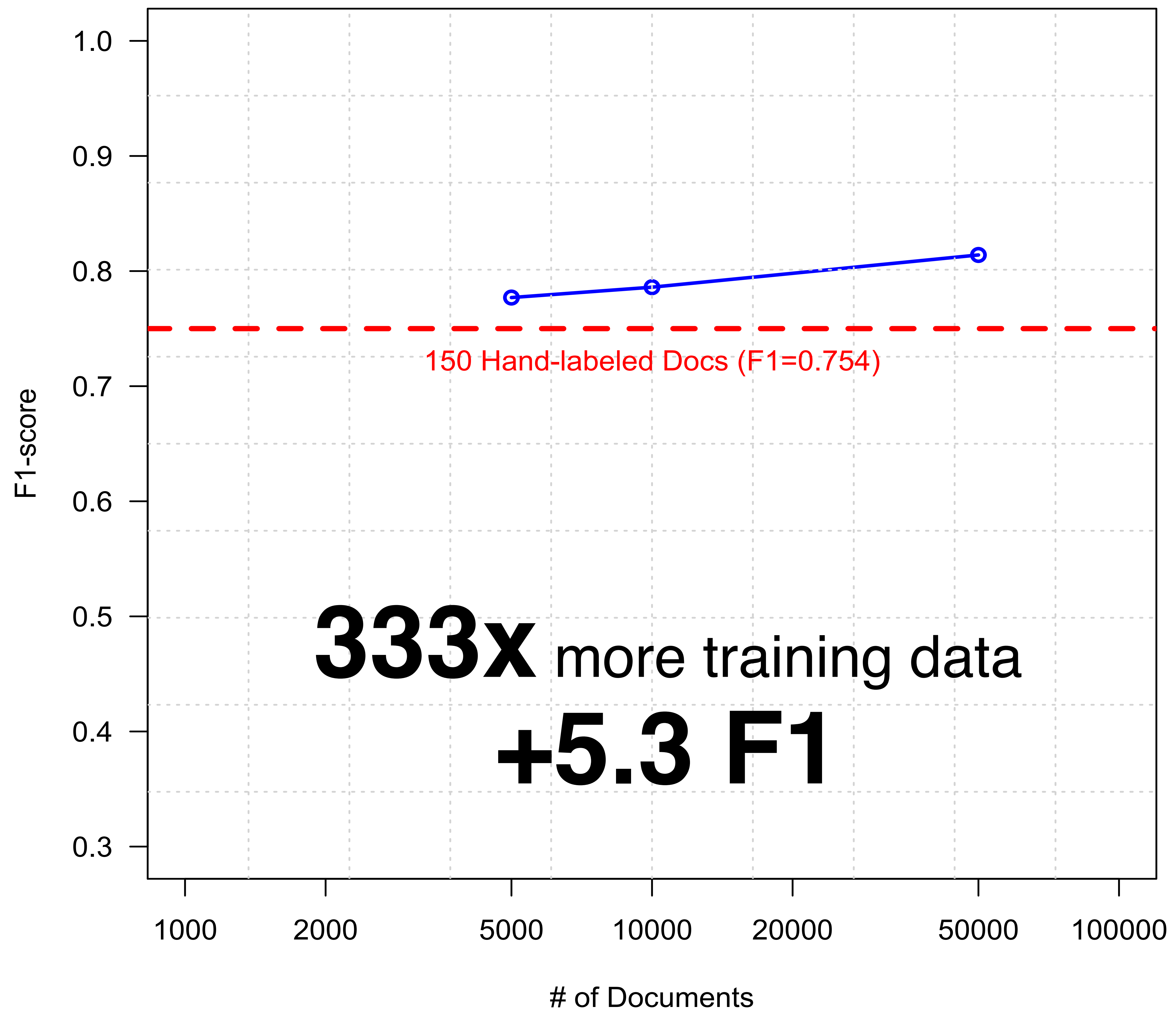
~ 20 - 40
Labeling Functions

Scaling with Unlabeled Data

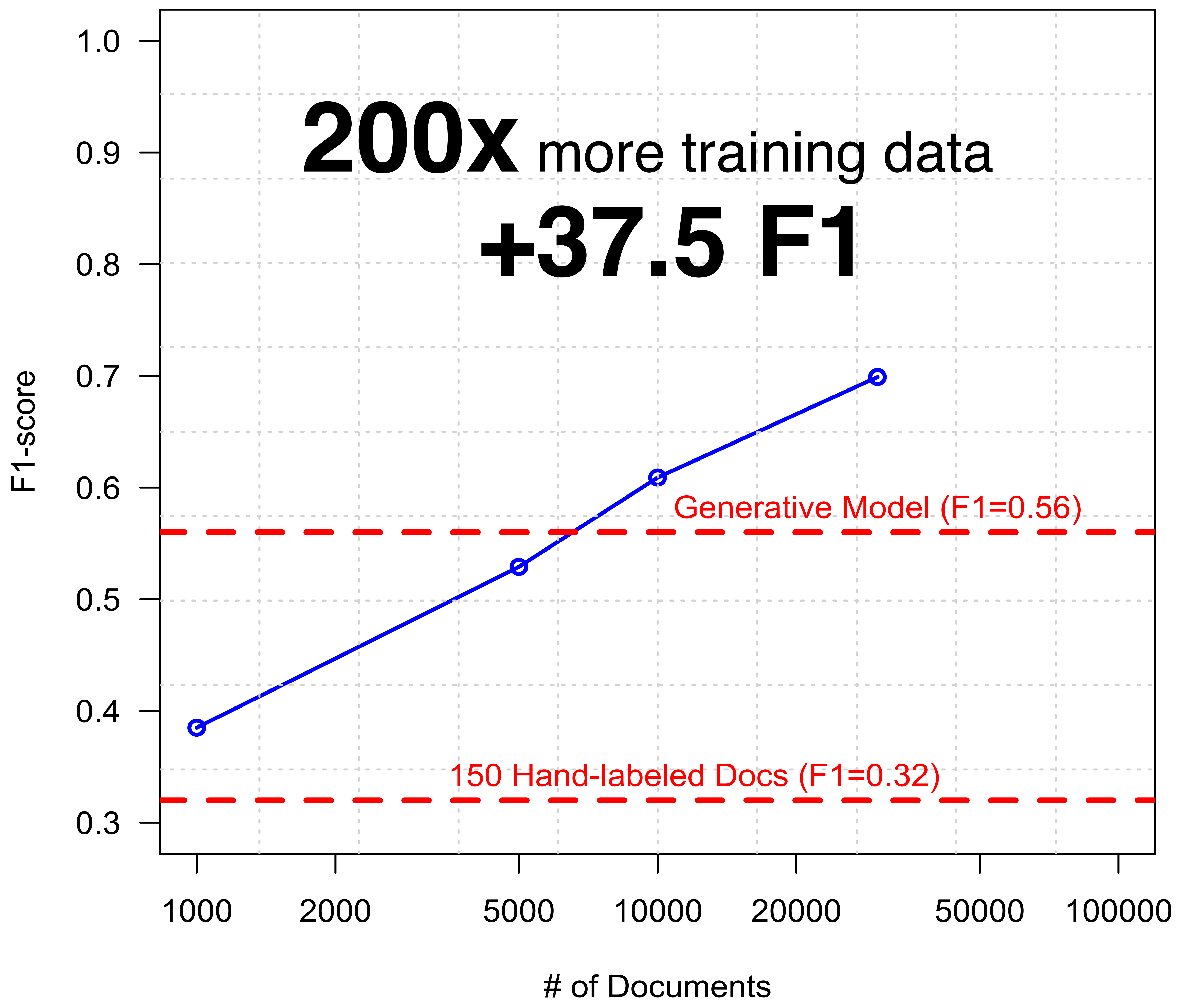


Pain

Scaling with Unlabeled Data



Pain



Complications

CATEGORY	NUM.	PRECISION	RECALL	F1	+/- F1
Revision	63	74.4	46.0	56.9	
Component Wear	48	71.4	41.7	52.6	
Mechanical Failure	25	87.5	28.0	42.4	
Particle Disease	65	80.0	6.2	11.4	
Radiographic Abnormality	17	**100.0**	**37.5**	**54.5**	
Infection	58	**100.0**	39.7	56.8	
Implant-Complications	276	81.7	32.4	46.4	
Pain-Anatomy	236	**81.4**	64.8	72.2	

Soft Majority Vote of Labeling Functions

CATEGORY	NUM.	PRECISION	RECALL	F1	+/- F1
Revision	63	**75.5**	**58.7**	**66.1**	**+16.2%**
Component Wear	48	72.9	**72.9**	**72.9**	**+38.6%**
Mechanical Failure	25	**91.7**	**44.0**	**59.5**	**+40.3%**
Particle Disease	65	**97.1**	**52.3**	**68.0**	**+496.5%**
Radiographic Abnormality	17	60.0	25.3	44.4	**-18.5%**
Infection	58	90.7	**84.5**	**87.5**	**+54.0%**
Implant-Complications	276	**82.7**	**62.3**	**71.1**	**+53.2%**
Pain-Anatomy	236	**80.2**	**82.6**	**81.4**	**+12.7%**

20k Imperfectly Labeled Documents

Improvements over a Rule-based Approach

MODEL	PRECISION	RECALL	F1
Majority Vote of LFs	81.7	32.4	46.4
Machine Learning	**82.7**	**62.3**	**71.1**

We trade little-to-no precision for
a big boost in recall

Closing Thoughts

The Benefits of Programmatic Supervision

Manually labeled datasets are
static artifacts with sunk costs

- Real machine learning tasks **change over time**
- Labeling functions are easily **shared and modified**
- Labeling functions can be **applied to unseen data**

~~Model~~ Labeling Function Zoos

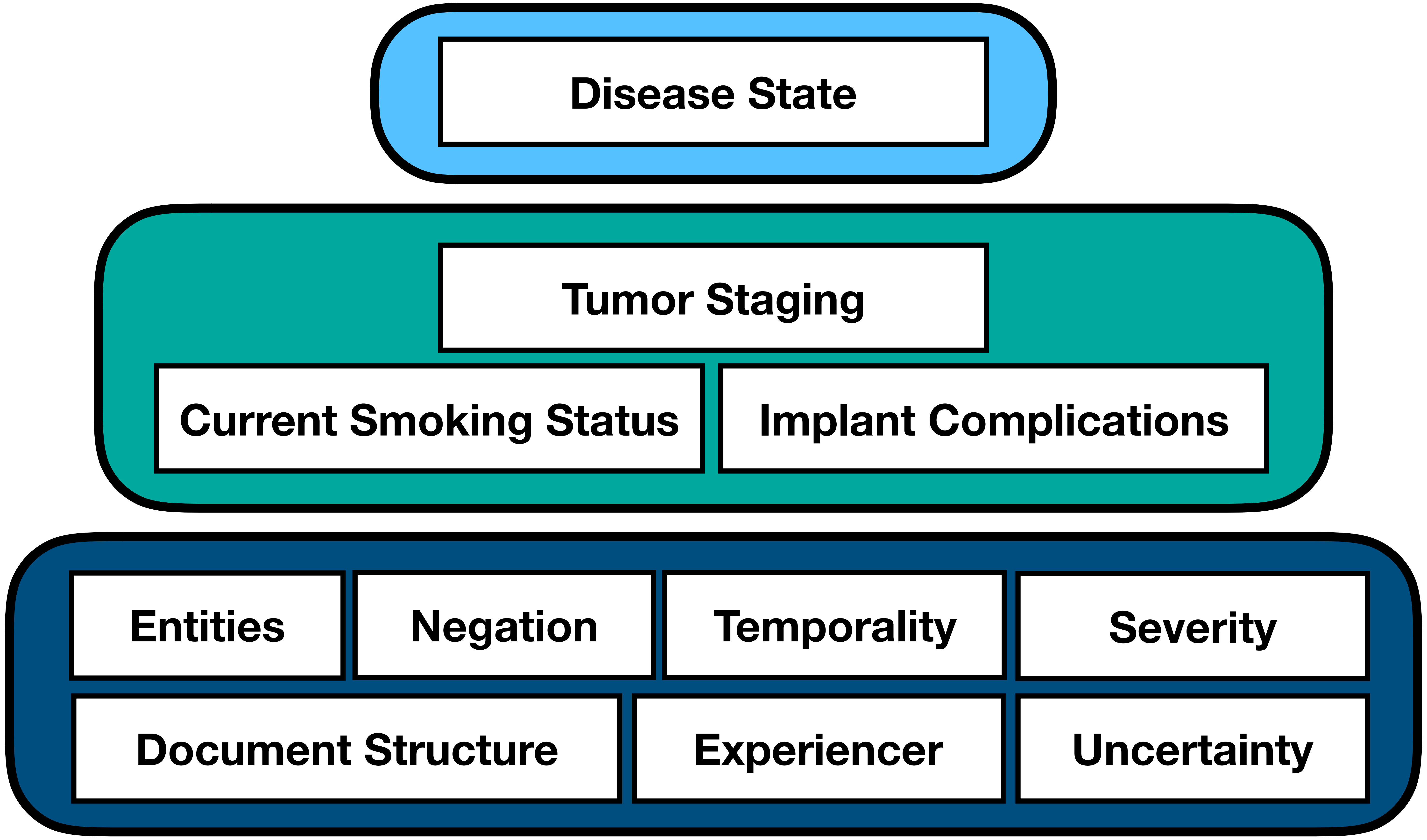
Downloadable **pre-trained, state-of-the-art models are common now** for text & images (model zoos)

…but clinical text models (especially large, language models like BERT) pose **considerable privacy issues**.

Share labeling functions instead!

Enables training high-performance NLP models with **orders of magnitude less hand-labeled data**

API / Programming Stack



Cohort Building
Phenotyping

Relations

Entities / Attributes

Reusable Supervision

Resources / Reading

Blogs, papers & more at: https://www.snorkel.org/

Academic Papers

Snorkel: Rapid Training Data Creation with Weak Supervision.
Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, Christopher Ré
Proceedings VLDB Endowment. 2017

SwellShark: A Generative Model for Biomedical Named Entity Recognition without Labeled Data
Jason Fries, Sen Wu, Alexander Ratner, Christopher Ré. 2017.

Medical device surveillance with electronic health records.
Alison Callahan, Jason A Fries, Christopher Ré, James I Huddleston III, Nicholas J Giori, Scott Delp, Nigam H Shah. 2019

Thank you!

jason-fries@stanford.edu