Weak Supervision Overview

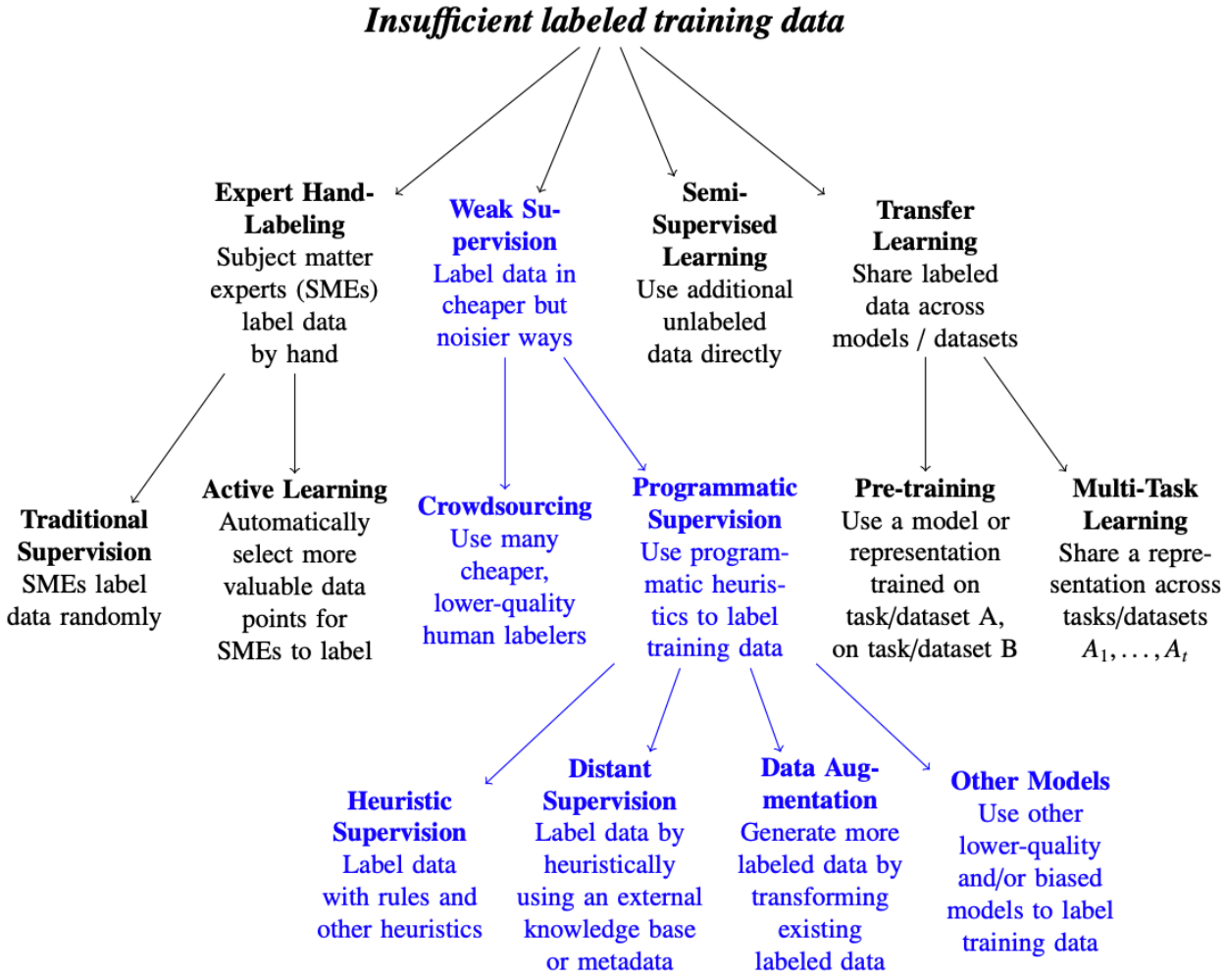
Chris Ré

Stanford University

Various techniques for limited labeled data

- **Active learning:** Select points to label more intelligently
- **Semi-supervised learning:** Use unlabeled data as well
- **Transfer learning:** Transfer from one training dataset to a new task
- **Weak supervision:** Label data in cheaper, higher-level ways

This lecture.

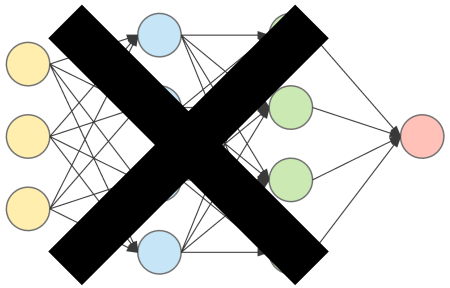


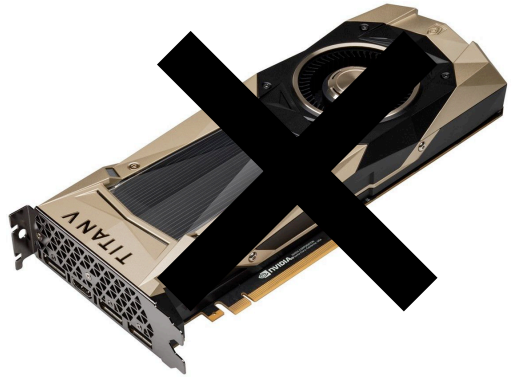
https://www.snorkel.org/blog/weak-supervision

Biased by active on-going work.

ML Application =

Model + Data + Hardware

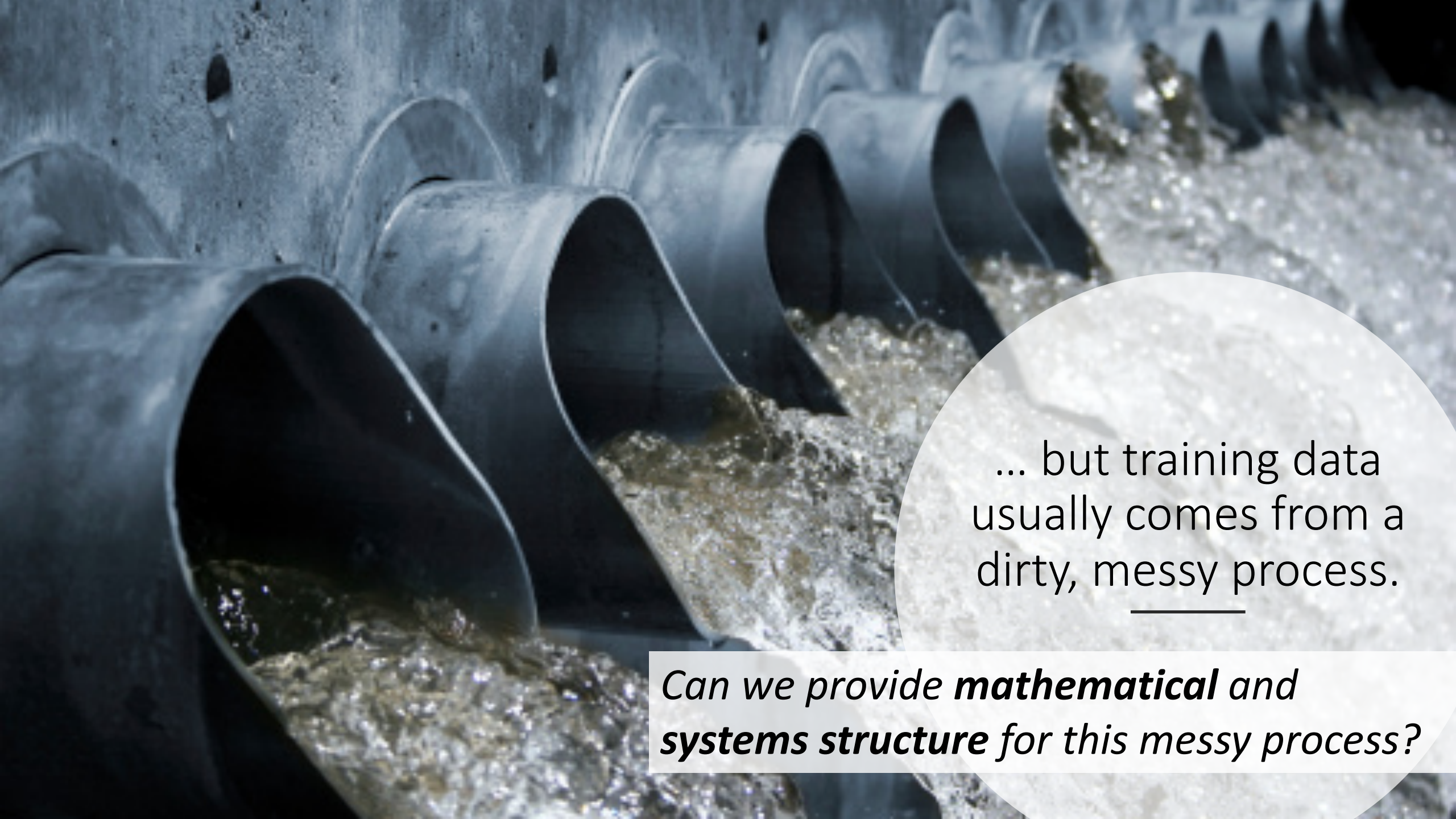




?

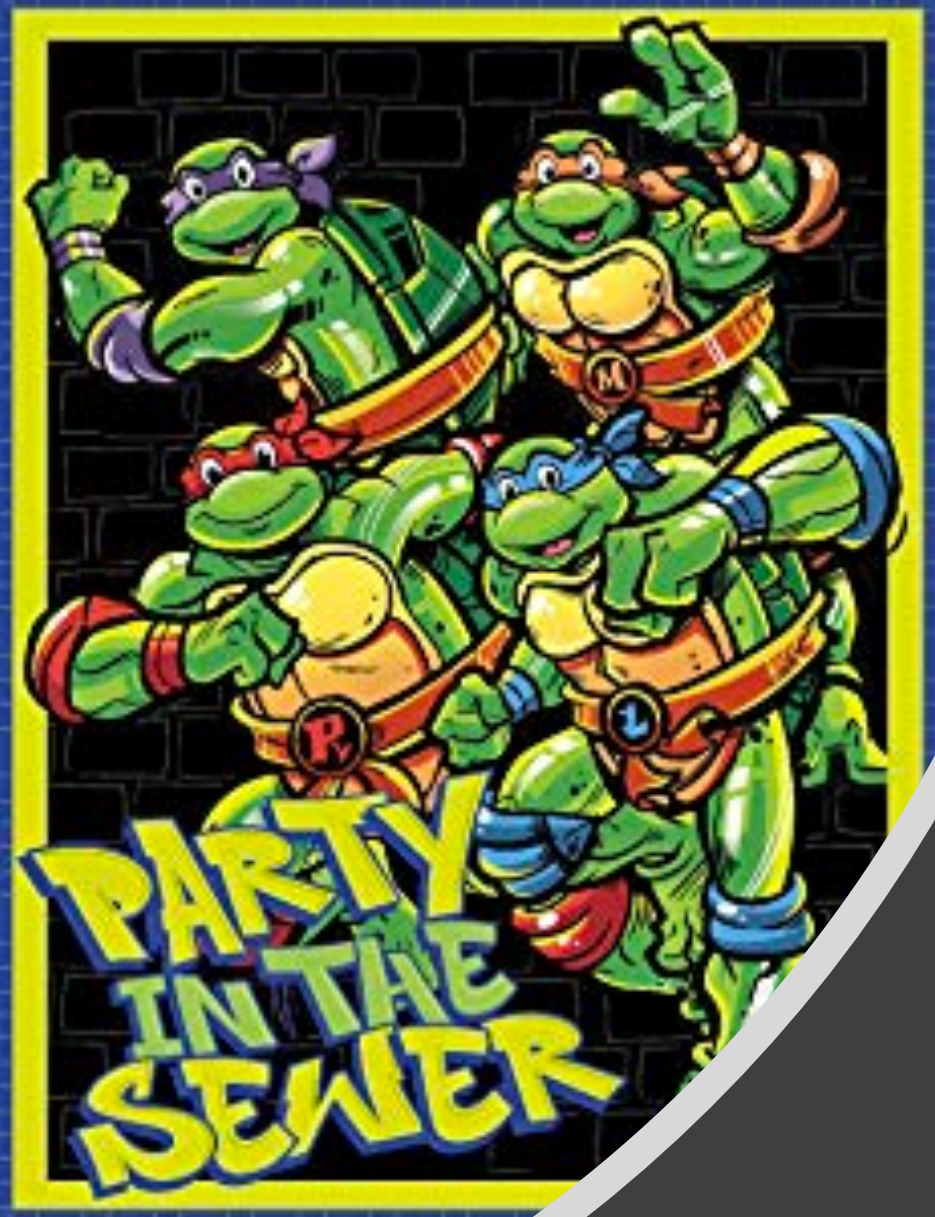
State-of-the-art models and hardware are available.
Training data is not

But supervision comes from god herself….



… but training data usually comes from a dirty, messy process.

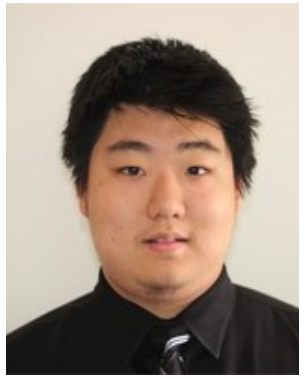
*Can we provide **mathematical** and **systems structure** for this messy process?*



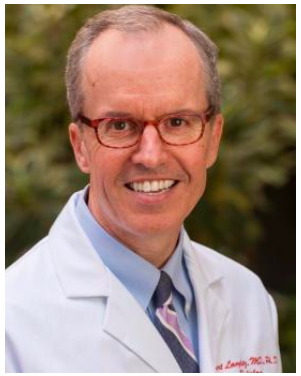
Supervision is where the action is…

*Model differences **overrated**, and supervision differences **underrated**.*

Alex Ratner



Darvin Yi



Curt Langlotz

Matt Lungren

Daniel Rubin

Jared Dunnmon

Automated Chest X-ray Triage

Optimizing Workflows with Automated Prioritization, Radiology 19

RSNA Radiological Society of North America

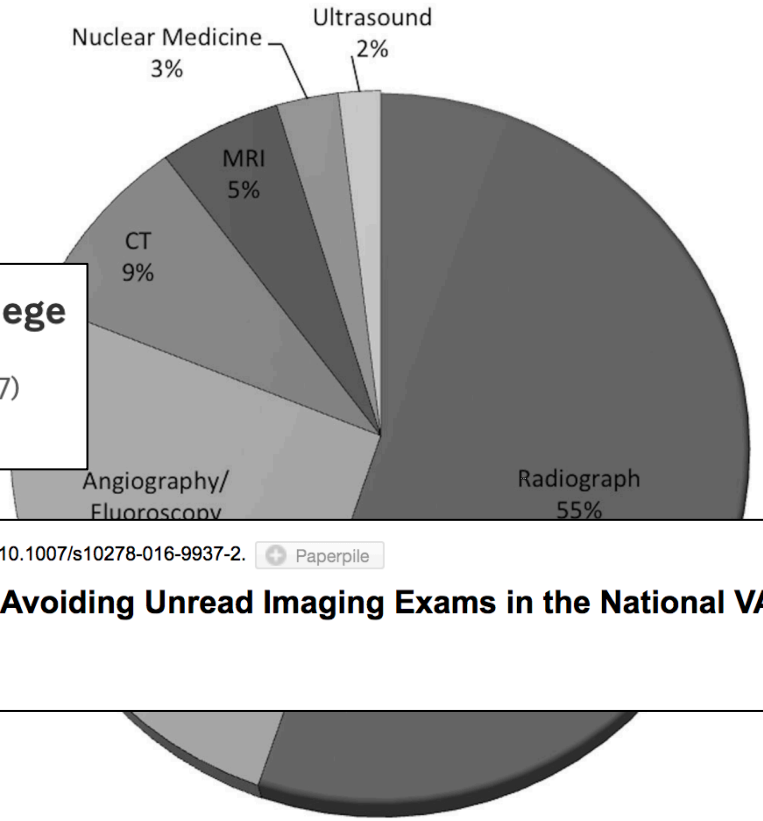
Radiology

J. Dunnnmon, D. Yi, C. Langlotz, C. Re, D. Rubin, M. Lungren. “Assessing Convolutional Neural Networks for Automated Radiograph Triage.” *Radiology*, 2019.

What's the Problem?

Radiologist shortage leaves patient care at risk, warns royal college

BMJ 2017 ; 359 doi: https://doi.org/10.1136/bmj.j4683 (Published 11 October 2017)
Cite this as: *BMJ* 2017;359:j4683



J Digit Imaging. 2017 Jun;30(3):309-313. doi: 10.1007/s10278-016-9937-2. Paperpile

Improving Patient Safety: Avoiding Unread Imaging Exams in the National VA Enterprise Electronic Health Record.

Bastawrous S[1,2], Carney B[3].

Too many of these!

Is Deep Learning the Answer?

This is not an easy question…

- No benchmark dataset
- Effects of data quality are unclear
- No assessment of existing algorithms
- No feedback from clinical community

…so we spent a year trying to answer it!

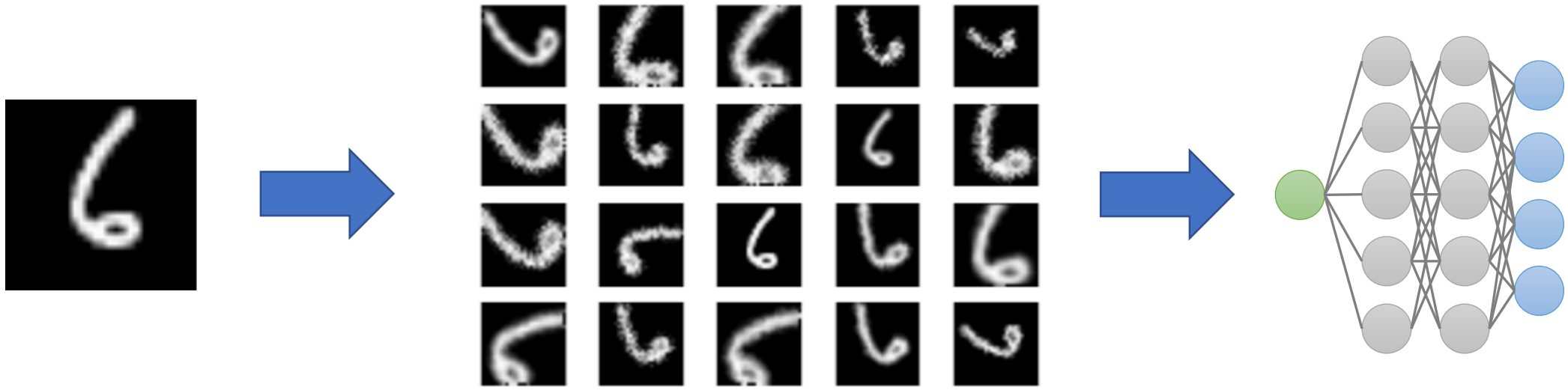
- Created large dataset of clinical labels
- Evaluated effect of label quality
- Work published in a *clinical journal*

Model	Test Accuracy
BOVW + KSVM	0.88
AlexNet	0.87
ResNet-18	0.89
DenseNet-121	**0.91**

Often: Differences in models ~ 2-3 points.

Later: Label quality & quantity > model choice.

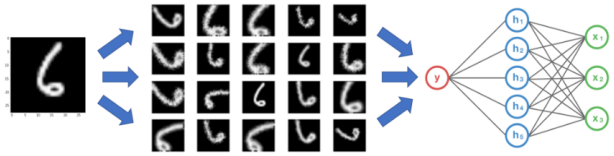
Even in Benchmarks: Data Augmentation is Critical



Ex: *13.4 pt.* avg. accuracy gain from data augmentation across top ten CIFAR-100 models—*difference in top-10 models is less!*

Training Signal is key to pushing SotA

New methods for gathering signal leading the state of the art

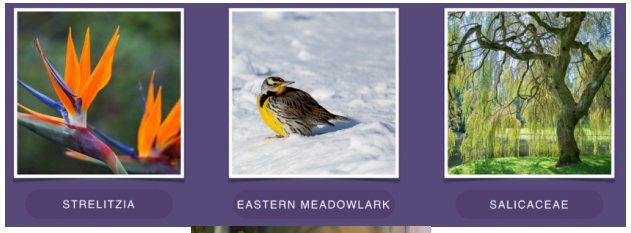


Google AI AutoAugment: Using learned **data augmentation policies (**RandAugment**)**

- **Augmentation Policies** first in Ratner et al. NIPS ’17

Facebook Hash tag weakly supervised pre-training

- Pre-train using a massive dataset with *weak training signal*

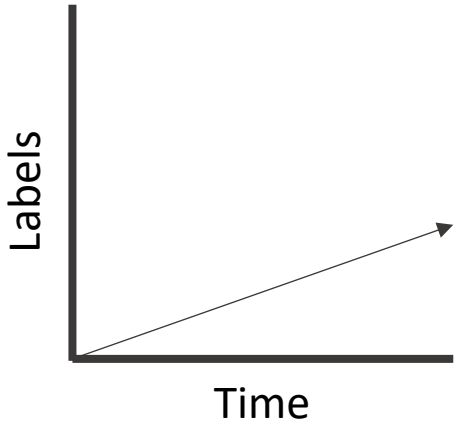


Training data: the new bottleneck

Slow, expensive, and static

Manual Labels

Slow



Expensive

$10 - $100/hr

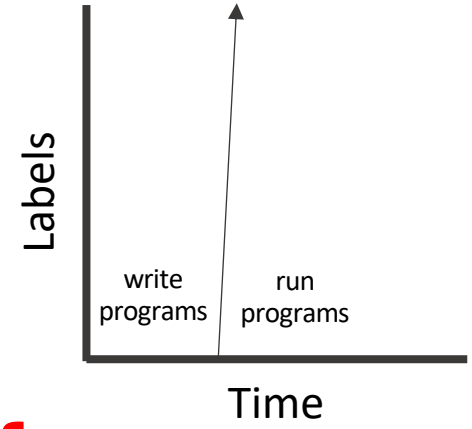
Static

{Positive, Negative}

{Positive, Neutral, Negative}

Programmatic Labels

Fast



Cheap

aws

$0.10/hr

Dynamic

Trade-off: programmatic labels are noisy…

Key Idea: Model Training Creation Process

This talk:

1. An interface for generating **training data via weak supervision**
2. An **approach** to learn quality and correlations of sources
3. Training an **end model---in various domains**

Snorkel: Formalizing Programmatic Labeling

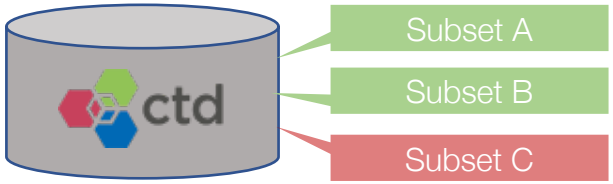
snorkel

Pattern Matching

```
regex.match(
    r"{A} is caused by {B}"
)
```

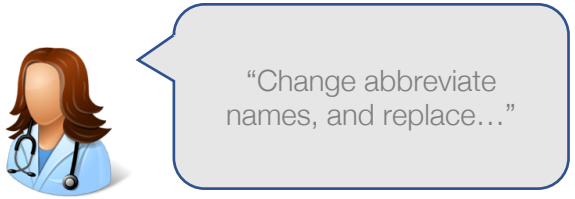
[e.g. Hearst 1992, Snow 2004]

Distant Supervision

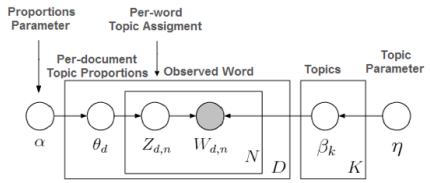


[e.g. Mintz 2009]

Augmentation

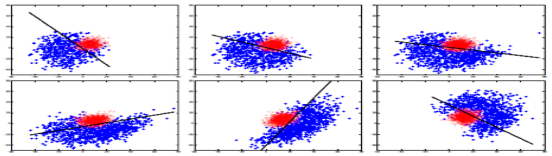


Topic Models



[e.g. Hingmire 2014]

Third-Party Models



[e.g. Schapire 1998]

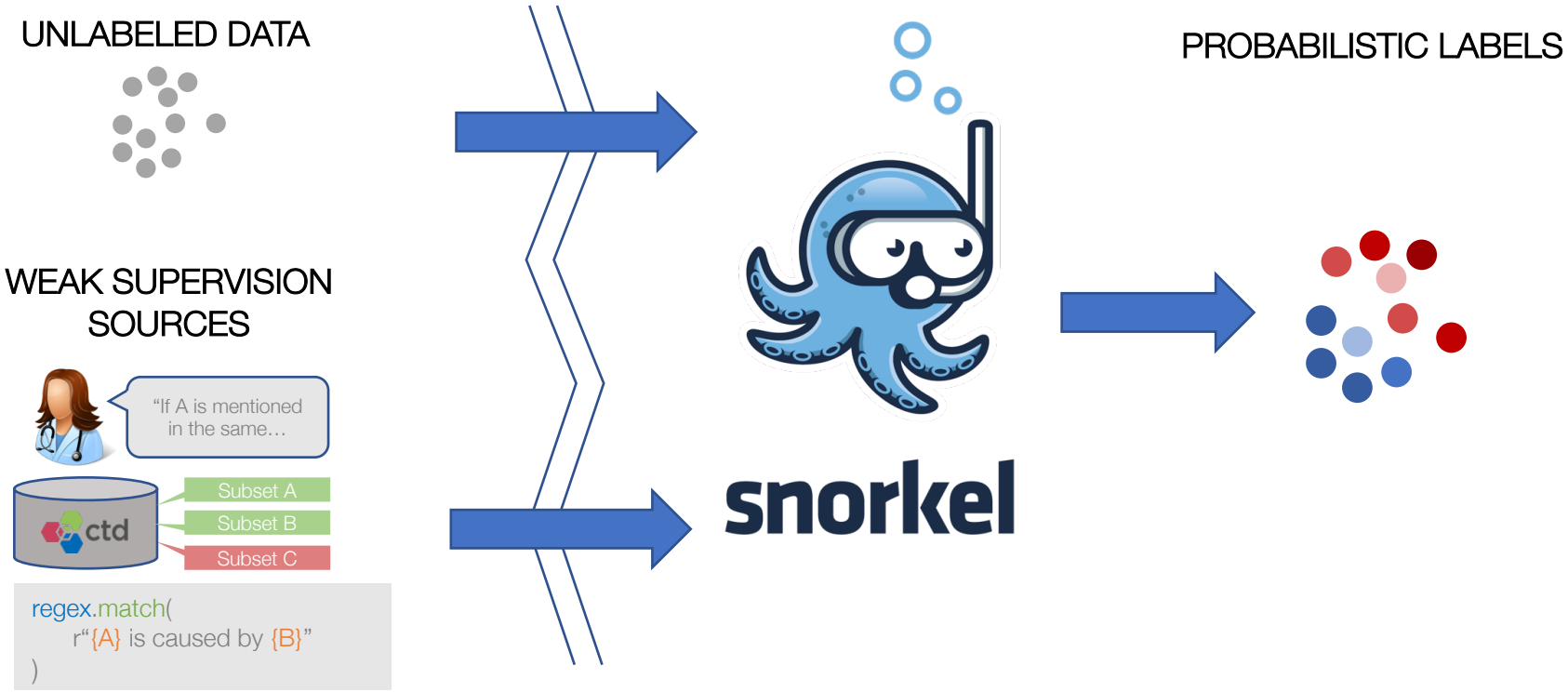
Crowdsourcing

amazon mechanical turk

[e.g. Dalvi 2013]

Observation: Weak supervision applied in *ad hoc* and isolated ways.

Snorkel: Formalizing Programmatic Labeling



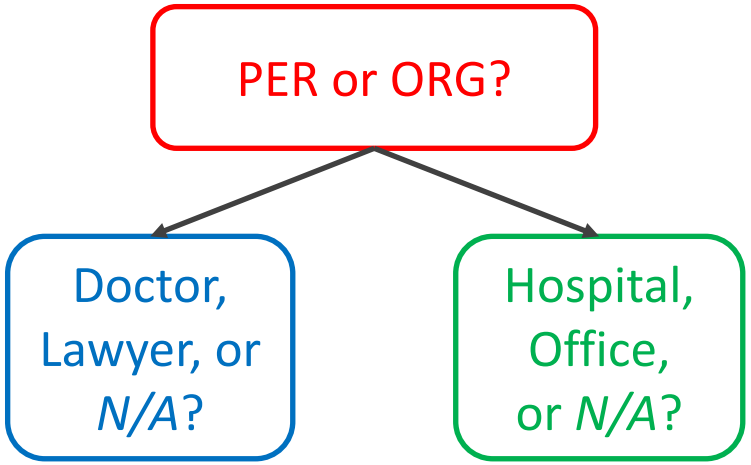
Goal: Replace *ad hoc* weak supervision with a formal, unified, theoretically grounded approach for programmatic labeling

Running Example: NER

PER:DOCTOR

Dr. Bob Jones is a specialist in cardiomyopathy treatment, leading the cardiology division at Saint Francis.

ORG:HOSPITAL



Goal: Label training data using *weak supervision* strategies for these three tasks

Weak Supervision as Labeling Functions

Dr. Bob Jones is a specialist in cardiomyopathy treatment, leading the cardiology division at Saint Francis.

```
def existing_classifier(x):
  return off_shelf_classifier(x)
```

→ “PERSON”

```
def upper_case_existing_classifier(x):
  if all(map(is_upper, x.split())) and \
off_shelf_classifier(x) == ‘PERSON’:
    return PERSON
```

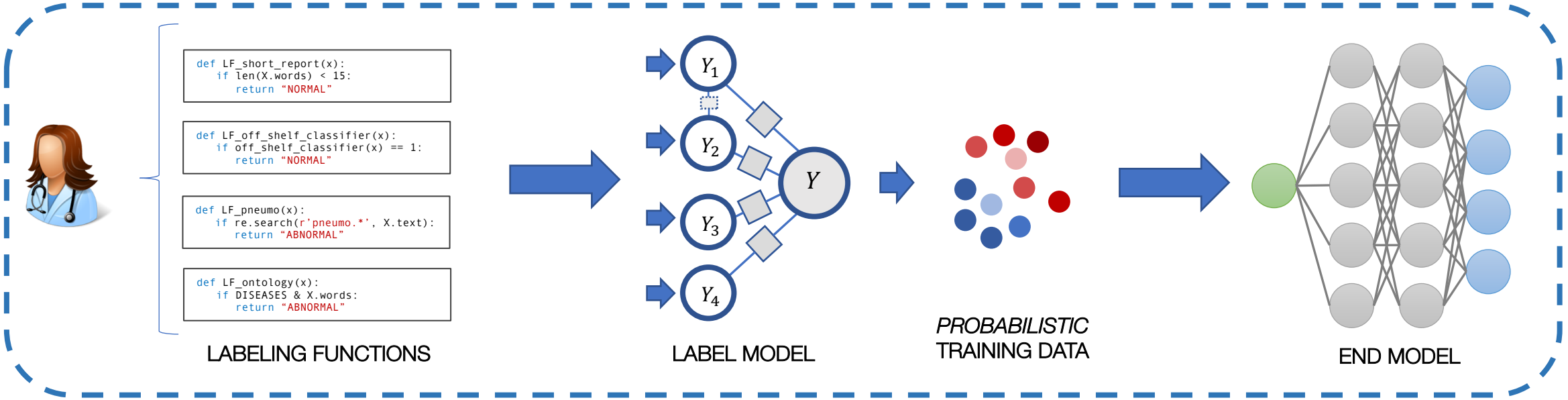
→ “PERSON”

```
def is_in_hospital_name_DB(x):
  if x in HOSPITAL_NAMES_DB:
    return HOSPITAL
```

→ “HOSPITAL”

Problem: These noisy sources *conflict and are correlated*

The Snorkel Pipeline



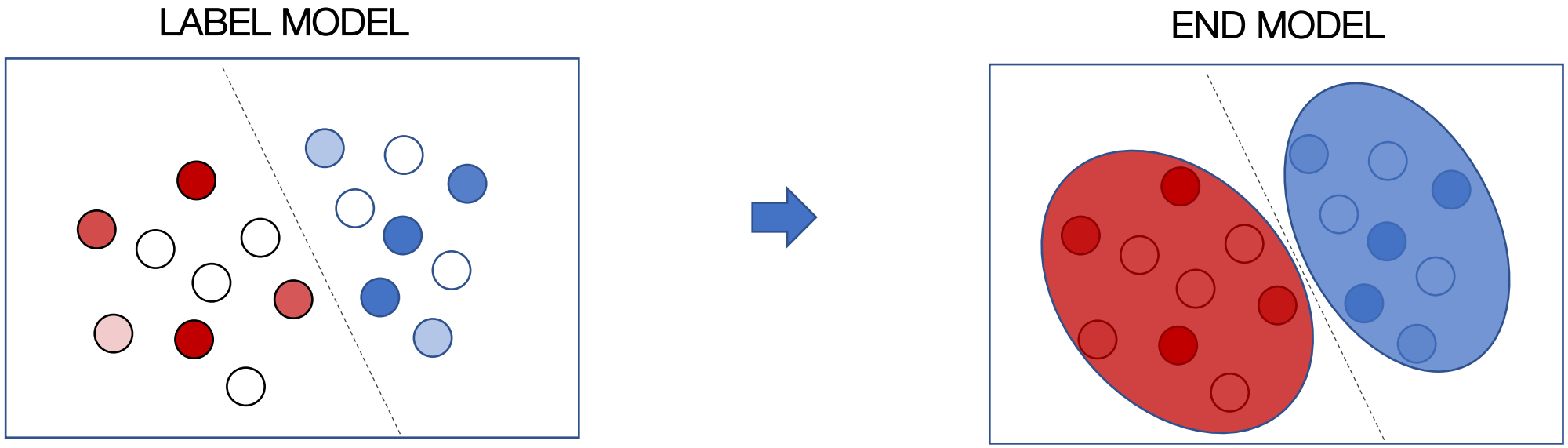
1. Users write *labeling functions* to generate noisy labels

2. Snorkel *models and combines* the noisy labels into probabilities

3. The resulting *probabilistic* labels train a model

KEY IDEA: Probabilistic training point carries accuracy. No hand labeled data needed.

Reason #1: Improved Generalization



Empirically, the end model boosts recall by **43%** on average!

Reason #1: Improved Generalization

Task: identify disease-causing chemicals

Phrases mentioned in LFs:

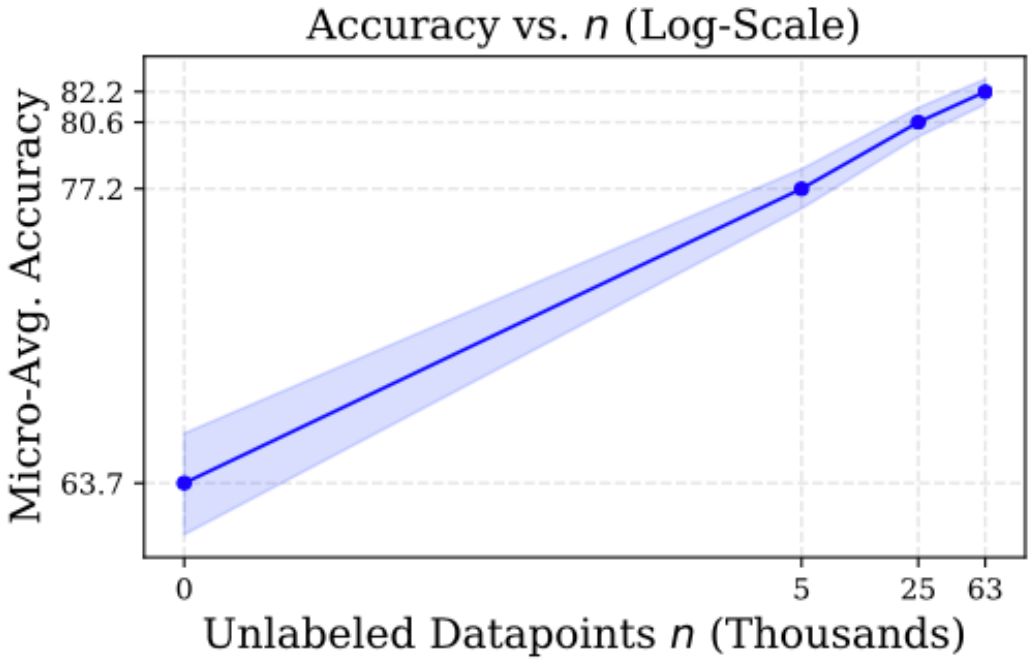
"treats", "causes", "induces", "prevents", …

Phrases given large weights by end model:

"could produce a", "support diagnosis of", …

The end model learned to take advantage of features that were helpful for prediction, but never explicitly mentioned in the LFs

Reason #2: Scaling with Unlabeled Data

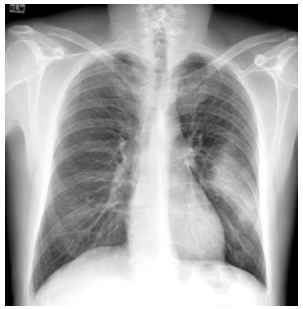


Add more unlabeled data—without changing the LFs—and performance improves!

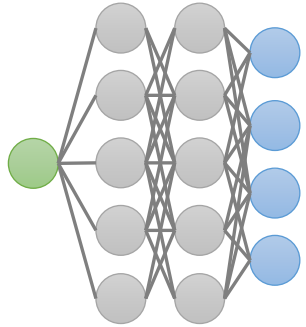
Reason #3: Cross-Model Supervision

Available at test time

This is servable!



ABNORMAL



```
def LF_short_report(x):
    if len(X.words) < 15:
        return "NORMAL"

def LF_off_shelf_classifier(x):
    if off_shelf_classifier(x) == 1:
        return "NORMAL"

def LF_pneumo(x):
    if re.search(r'pneumo.*', X.text):
        return "ABNORMAL"

def LF_ontology(x):
    if DISEASES & X.words:
        return "ABNORMAL"
```

snorkel

Report 47:

Indication:
Chest pain.
Findings:
Pneumothorax.

Operation
recommended.

ABNORMAL

Not available at test time

Not servable

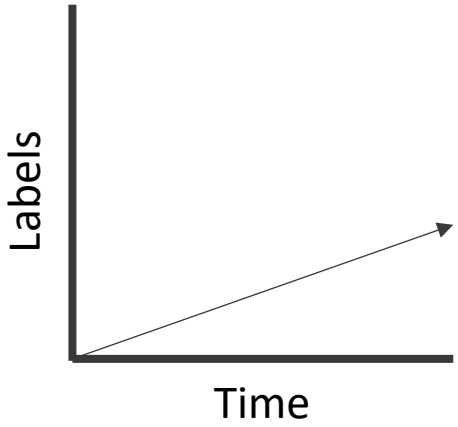
Stanford MEDICINE

Hours of weak supervision
matches manual labels
collected over ***person years!***

Use training data as a medium for knowledge transfer

Manual Labels

Slow



Expensive

$10 - $100/hr

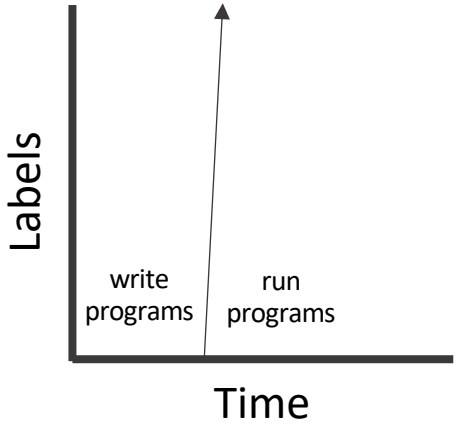
Static

{Positive, Negative}

{Positive, Neutral, Negative}

Programmatic Labels

Fast



Cheap

aws

$0.10/hr

Dynamic

Snorkel: In use at the world's largest companies

snorkel

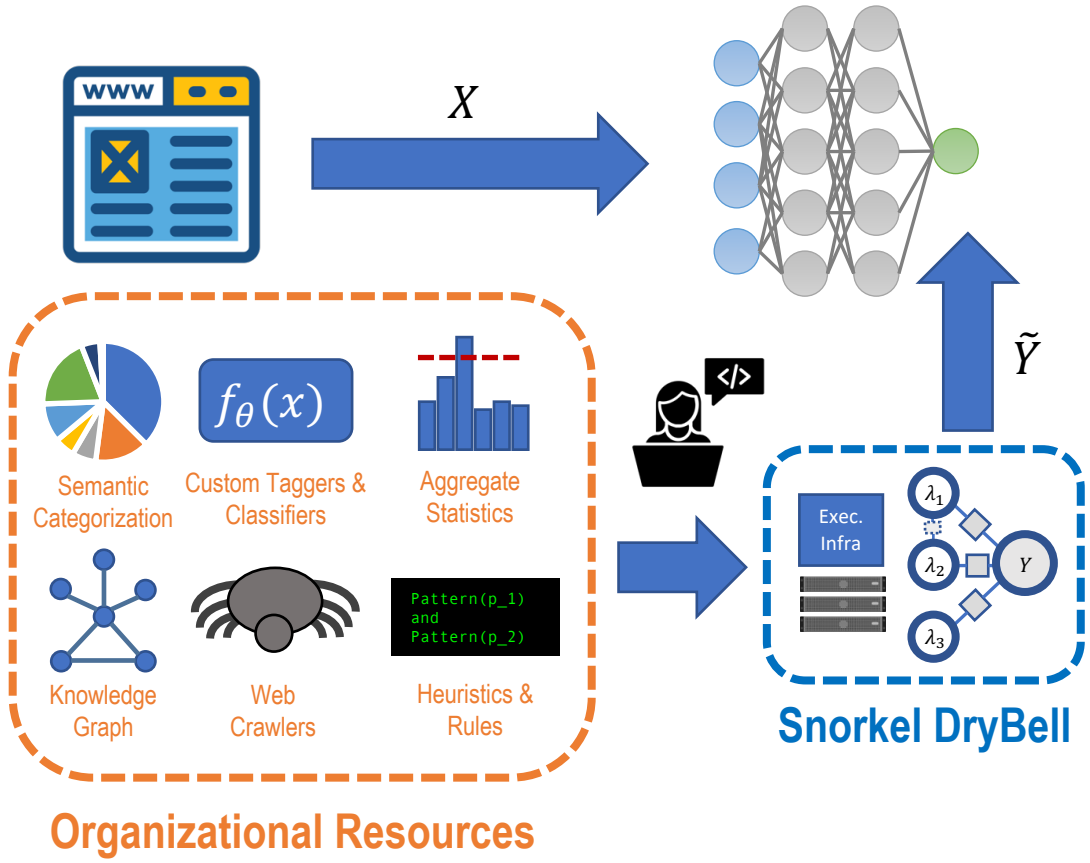
Http://snorkel.org

"Snorkel DryBell" collaboration with Google Ads. Bach et al. SIGMOD19.

Used in production in many industries, startups, and other tech companies!

Collaboration Highlight: Google + Snorkel

- *Snorkel DryBell* is a production version of Snorkel focused on:
  - Using *organizational knowledge resources* to train ML models
  - Handling *web-scale* data
  - Non-servable to servable feature transfer.

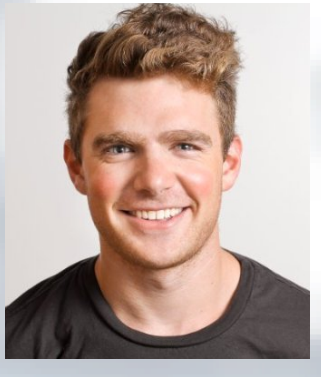


[Bach et. al., SIGMOD 2019]

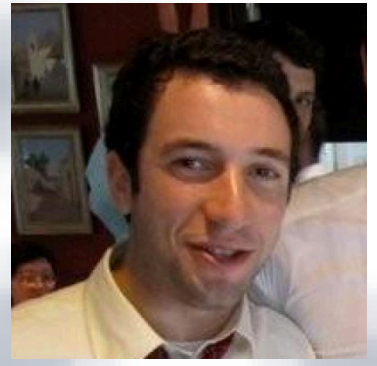
The Real Work

Stephen Bach

Braden Hancock



Henry Ehrenberg



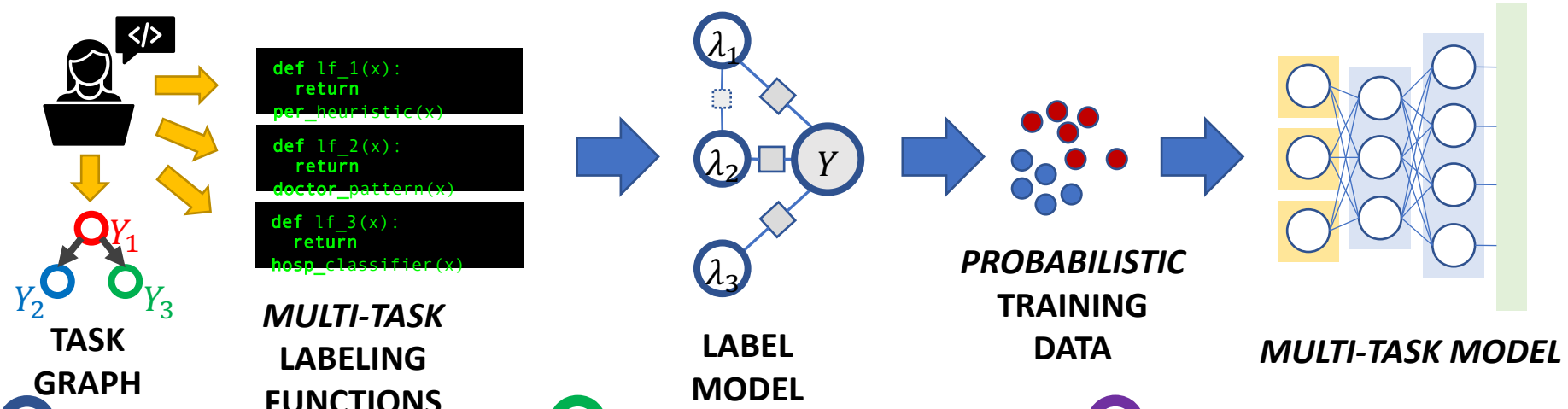
Alex Ratner

Paroma Varma

Snorkel.org

Let’s look under the hood and take a peak at some math

The Snorkel Pipeline



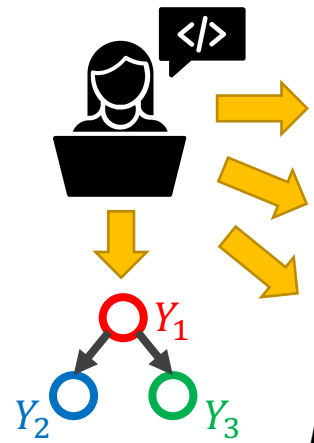
1. Users write *labeling functions* ***for multiple related tasks***
2. We model the labeling functions' behavior to de-noise them
3. We use the *probabilistic* labels to train a ***multi-task*** model

No hand-labeled training data!

A. Ratner, C. De Sa, S. Wu, D. Selsam, C. Ré, "Data programming: Creating large training sets, quickly", NIPS 2016.

A. Ratner, B. Hancock, J. Dunnmon, F. Sala, C. Ré, "Training complex models with multi-task weak supervision", AAAI 2019.

The Snorkel Pipeline



TASK GRAPH

```
def lf_1(x):
  return per_heuristic(x)

def lf_2(x):
  return doctor_pattern(x)

def lf_3(x):
  return hosp_classifier(x)
```

MULTI-TASK LABELING FUNCTIONS

LABEL MODEL

PROBABILISTIC TRAINING DATA

MULTI-TASK MODEL

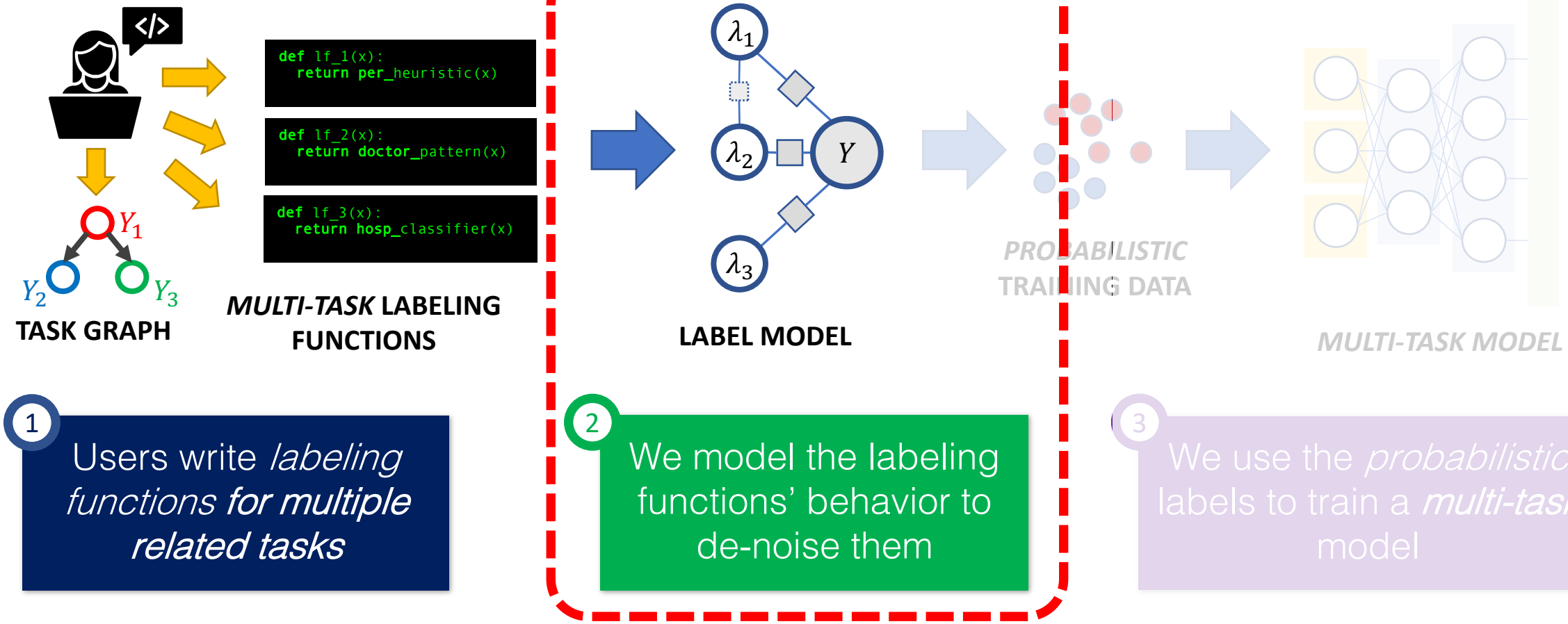
1 Users write *labeling functions for multiple related tasks*

2 We model the labeling functions' behavior to de-noise them

3 We use the *probabilistic* labels to train a *multi-task* model

How to represent diverse sources of weak supervision?

The Snorkel Pipeline



1. Users write *labeling functions* ***for multiple related tasks***

2. We model the labeling functions' behavior to de-noise them

3. We use the *probabilistic* labels to train a *multi-task* model

How can we do anything without the ground truth labels?

Model as Generative Process

```
def existing_classifier(x):
  return off_shelf_classifier(x)
```

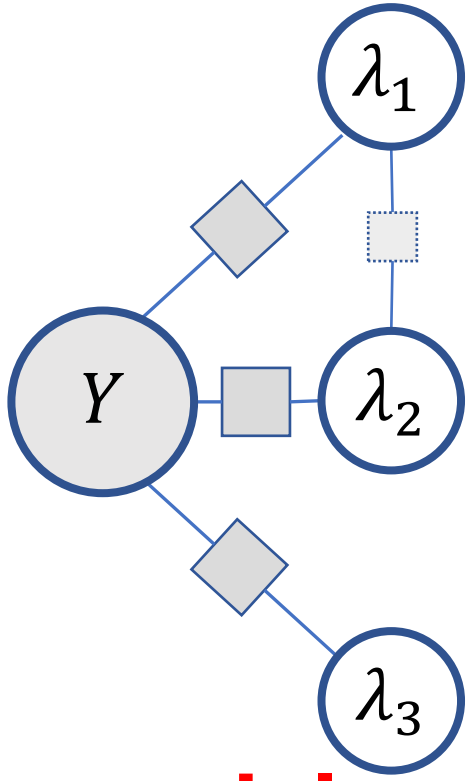
→ “PERSON”

```
def upper_case_existing_classifier(x):
  if all(map(is_upper, x.split())) and \
off_shelf_classifier(x) == ‘PERSON’:
    return PERSON
```

→ “PERSON”

```
def is_in_hospital_name_DB(x):
  if x in HOSPITAL_NAMES_DB:
    return HOSPITAL
```

→ “HOSPITAL”



How to learn the parameters of this model (accuracies & correlations) without Y?

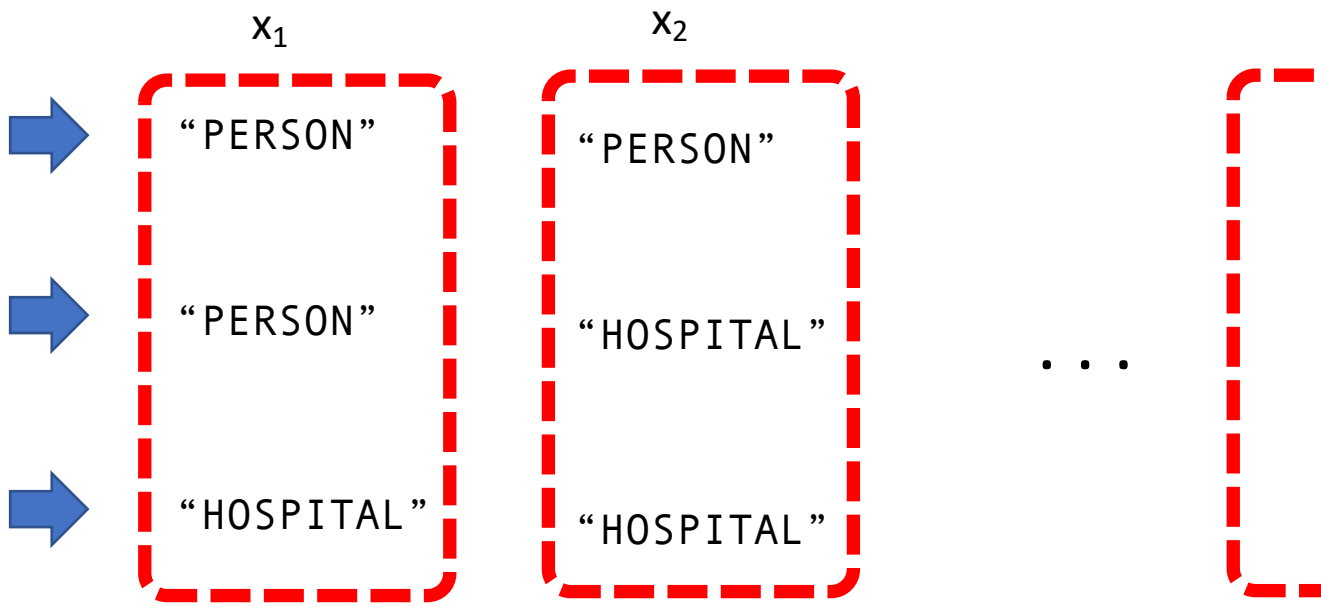
Intuition: Learn from the Overlaps

Sources.

```
def existing_classifier(x):
  return off_shelf_classifier(x)
```

```
def upper_case_existing_classifier(x):
  if all(map(is_upper, x.split())) and \
off_shelf_classifier(x) == 'PERSON':
    return PERSON
```

```
def is_in_hospital_name_DB(x):
  if x in HOSPITAL_NAMES_DB:
    return HOSPITAL
```



Key idea: We can observe overlapping judgements on many points!