Introduction to Weak Supervision

Chris Ré

CS229

Messages for Today

Introduce **two key concepts**:

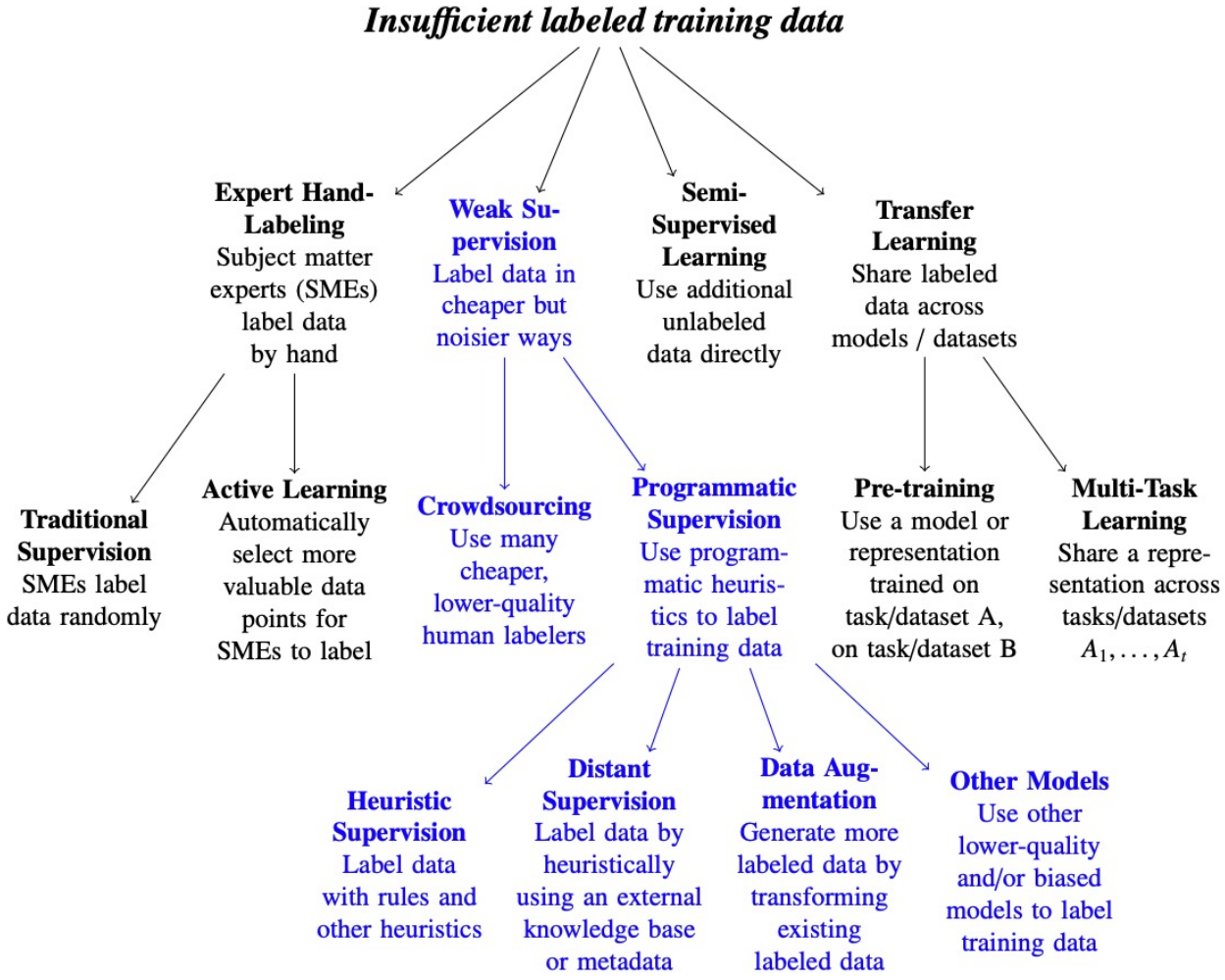
- **Method of moments for** latent probabilistic variable models
  - *They have provable global solution (*Compare with EM methods)
  - Widely used in "tensor methods"
- Probability **distributions on graphs** (graphical models)
  - Fun facts about Gaussians that are good for your soul (Inverse covariance matrix structure and graphs)

- High-level overview of new area called **weak supervision**.
  - Why supervision is so critical in this age and resources (nascent)
  - Very recent work & biased by our own group's work—but you have likely used it today!

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

Related Work in Weak Supervision

- ***Crowdsourcing:*** Dawid & Skene 1979, Karger et. al. 2011, Dalvi et. al. 2013, Ruvolo et. al. 2013, Zhang et. al. 2014, Berend & Kontorovich 2014, etc.
- ***Distant Supervision:*** Mintz et. al. 2009, Alfonesca et. al. 2012, Takamatsu et. al. 2012, Roth & Klakow 2013, Augenstein et. al. 2015, etc.
- ***Co-Training:*** Blum & Mitchell 1998
- ***Noisy Learning:*** Bootkrajang et. al. 2012, Mnih & Hinton 2012, Xiao et. al. 2015, etc.
- ***Indirect Supervision:*** Clarke et. al. 2010, Guu et. Al. et. al. 2017, etc.
- ***Feature and Class-distribution Supervision:*** Zaidan & Eisner 2008, Druck et. al. 2009, Liang et. al. 2009, Mann & McCallum 2010, etc.
- ***Boosting & Ensembling:*** Schapire & Freund, Platanios et. al. 2016, etc.
- ***Constraint-Based Supervision:*** Bilenko et. al. 2004, Koestinger et. al. 2012, Stewart & Ermon 2017, etc.

More Related work

- So much more! *Work was inspired by classics and new Cotraining , GANs, capsule networks, semi-supervised learning, crowd-sourcing and so much more!*

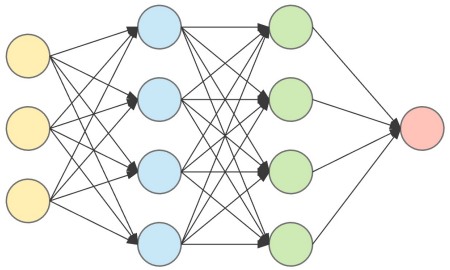
- Please see blog for summary. https://www.snorkel.org/blog/weak-supervision

snorkel

… Biased by on-going work…

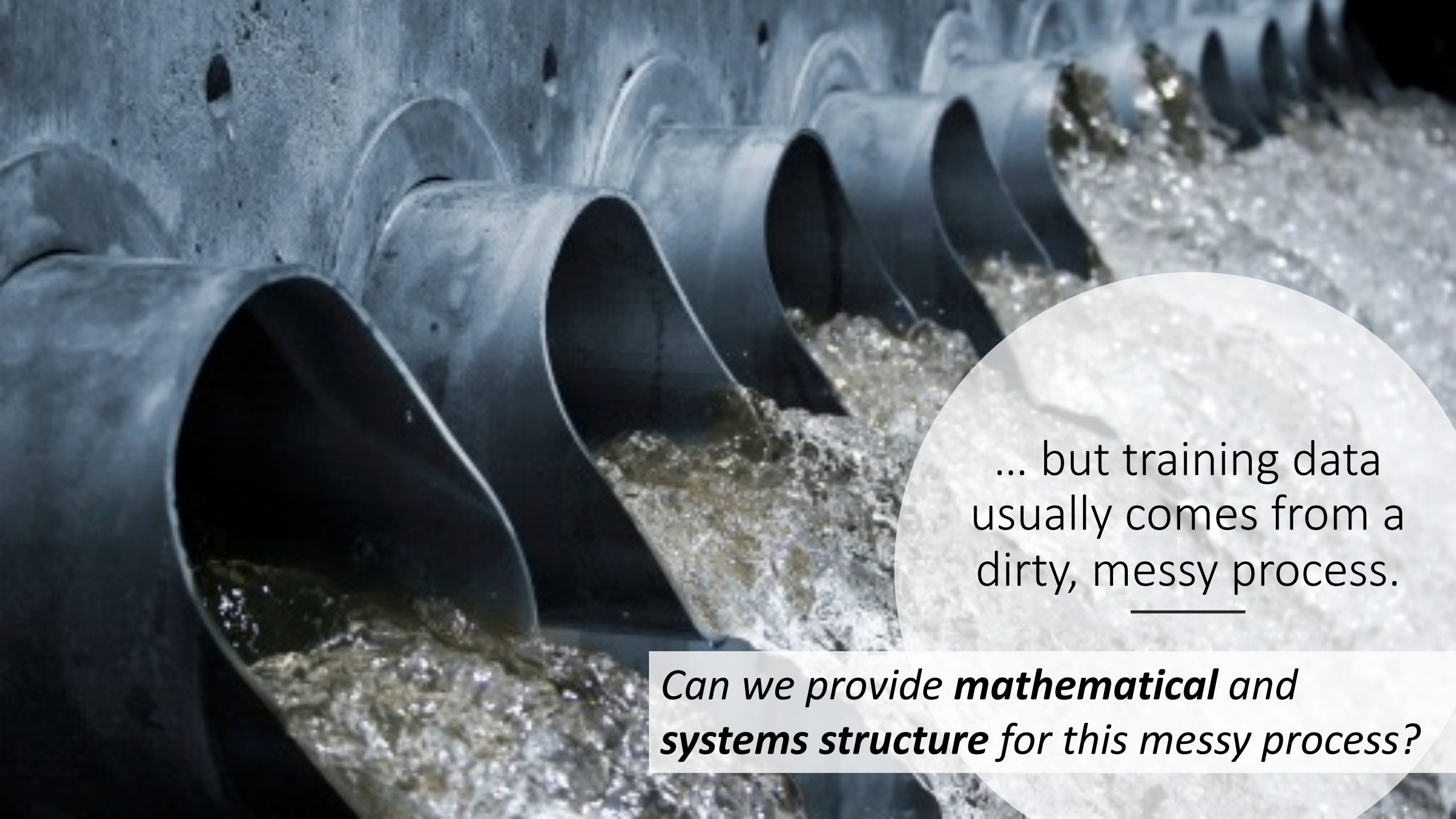
ML Application =

Model + Data + Hardware



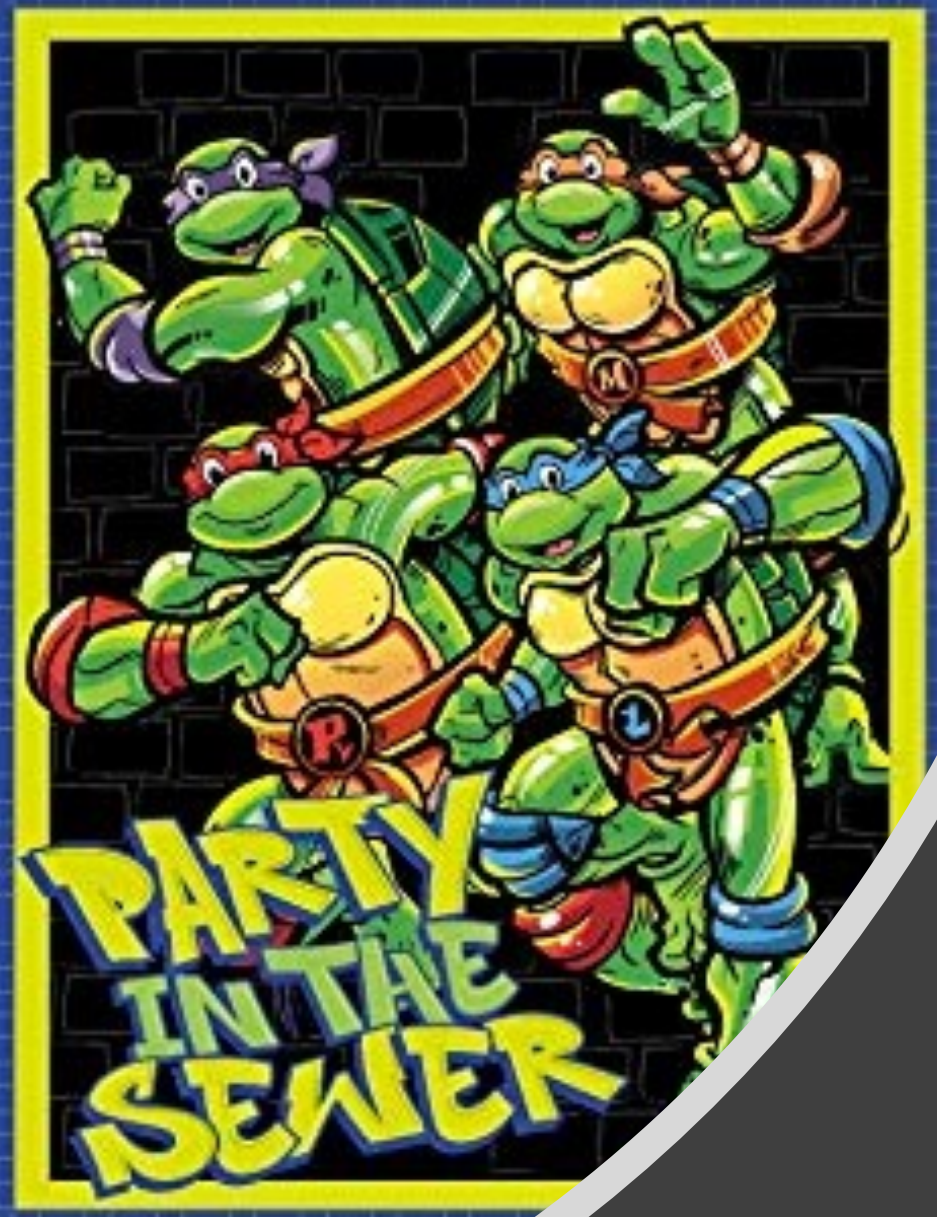
State-of-the-art models and hardware are available.
Training data is not

But supervision comes from god herself….



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

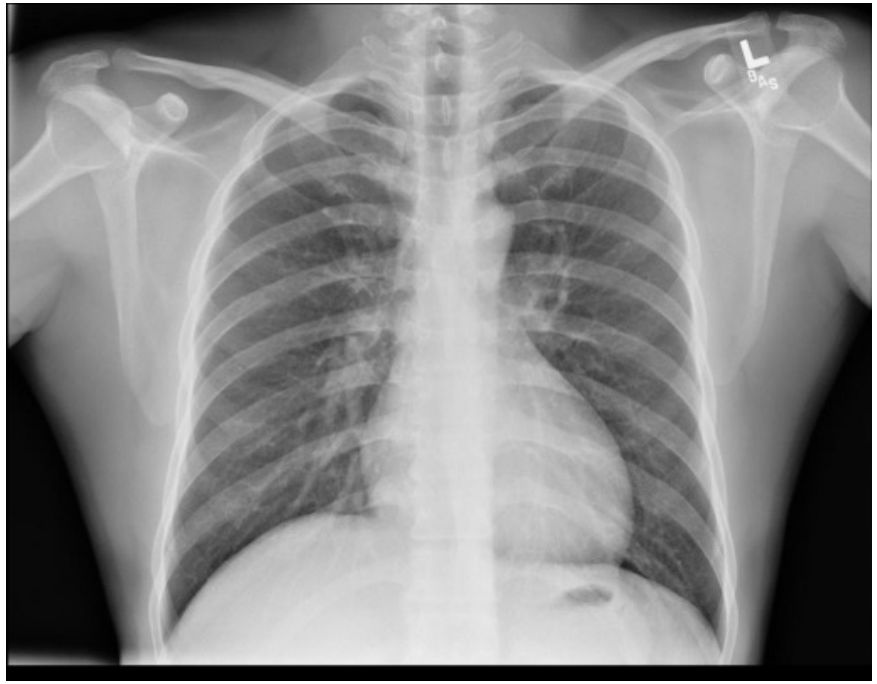
*Can we provide **mathematical** and **systems structure** for this messy process?*



Supervision is where the action is…

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

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



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

We spent a year on this challenge

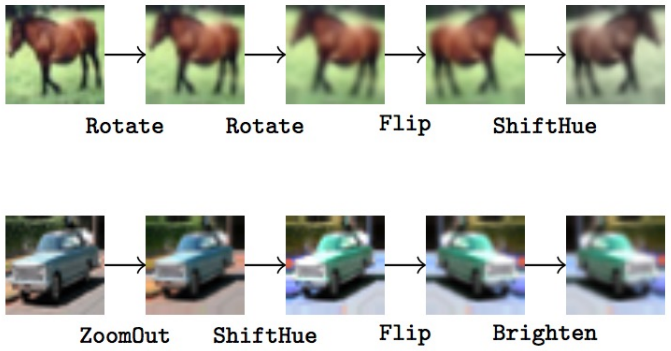
- Created large dataset of clinical labels
- Evaluated effect of label quality
- Work published in a *clinical journal*

Often: Differences in models ~ 2-3 points.

Label quality & quantity > model choice.

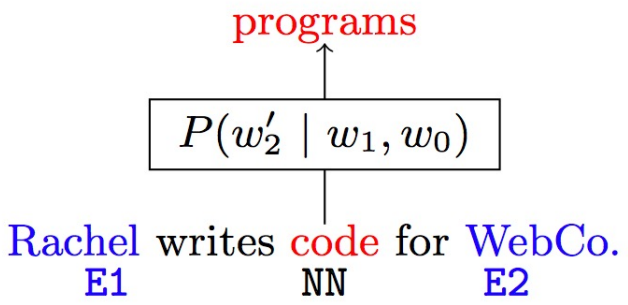
Data augmentation by specifying invariances

Images



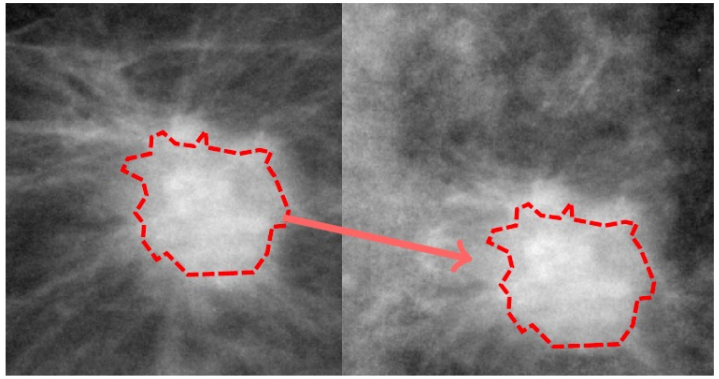
- Rotations
- Scaling / Zooms
- Brightness
- Color Shifts
- Etc…

Text



- Synonymy
- Positional Swaps
- Etc…

Medical

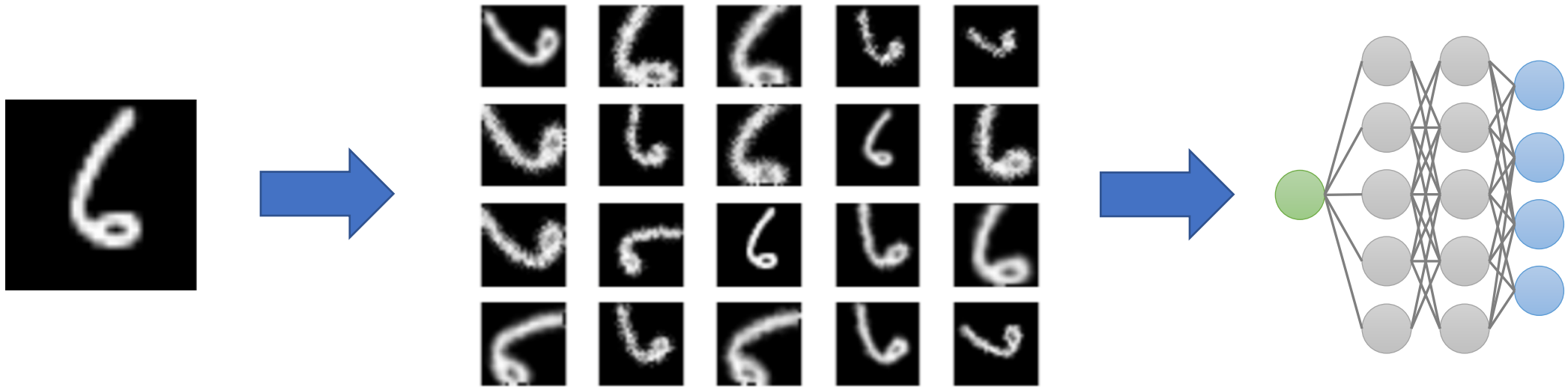


Domain-specific transformations. Ex:

1. *Segment tumor mass*
2. *Move*
3. *Resample background tissue*
4. *Blend*

How do we choose which to apply? In what order?

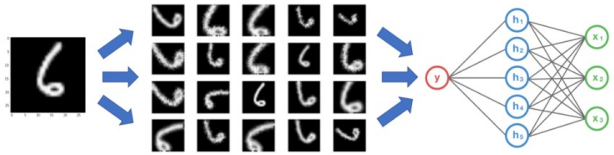
Simple Benchmarks:
Data Augmentation is Critical



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

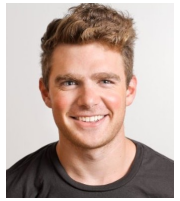
Training Signal is key to pushing SotA

New methods for gathering signal leading the state of the art, lots of exciting ML progress here (SotA due to noisy teacher!)



Google AI AutoAugment: Using learned **data augmentation policies**

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



Henry Ehrenberg

Alex Ratner (to: Washington)

Facebook Hash tag weakly supervised pre-training

- Pre-train using a massive dataset with *hashtags*

Sharon Y. Li (to: Wisconsin)

Check out Sharon's series on hazyresearch.Stanford.edu

HOME PEOPLE

Automating the Art of Data Augmentation

Part I Overview

The Stanford AI Lab Blog

Sharon Y. Li
(to: Wisconsin)

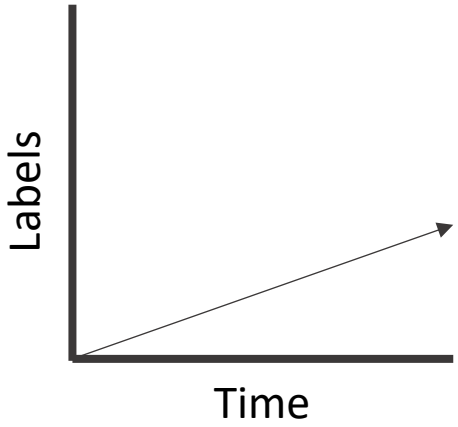
http://ai.stanford.edu/blog/data-augmentation/

Training data: the new bottleneck

Slow, expensive, and static

Manual Labels

Slow



Expensive

$10 - $100/hr

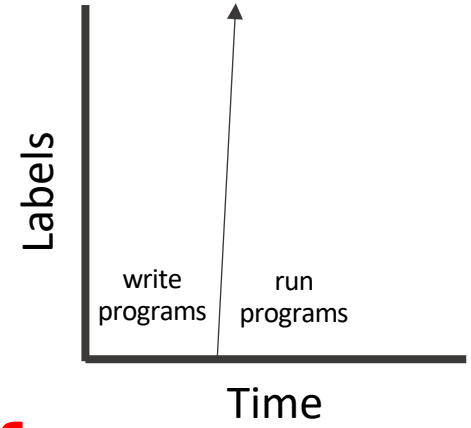
Static

{Positive, Negative}

{Positive, Neutral, Negative}

Programmatic Labels

Fast



Cheap

aws

$0.10/hr

Dynamic

Trade-off: programmatic labels are noisy…

Snorkel: Formalizing Programmatic Labeling

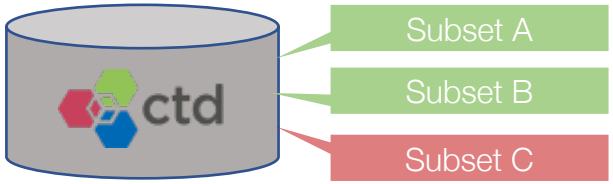
snorkel

Pattern Matching

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

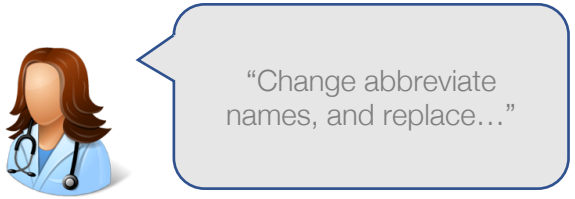
[e.g. Hearst 1992, Snow 2004]

Distant Supervision

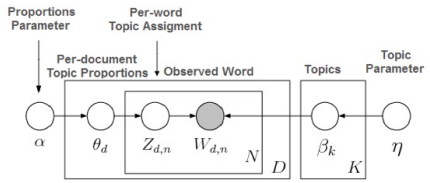


[e.g. Mintz 2009]

Augmentation

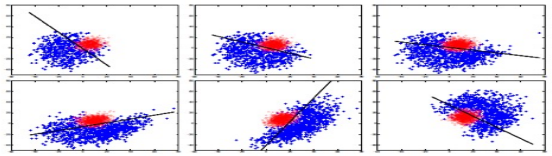


Topic Models



[e.g. Hingmire 2014]

Third-Party Models



[e.g. Schapire 1998]

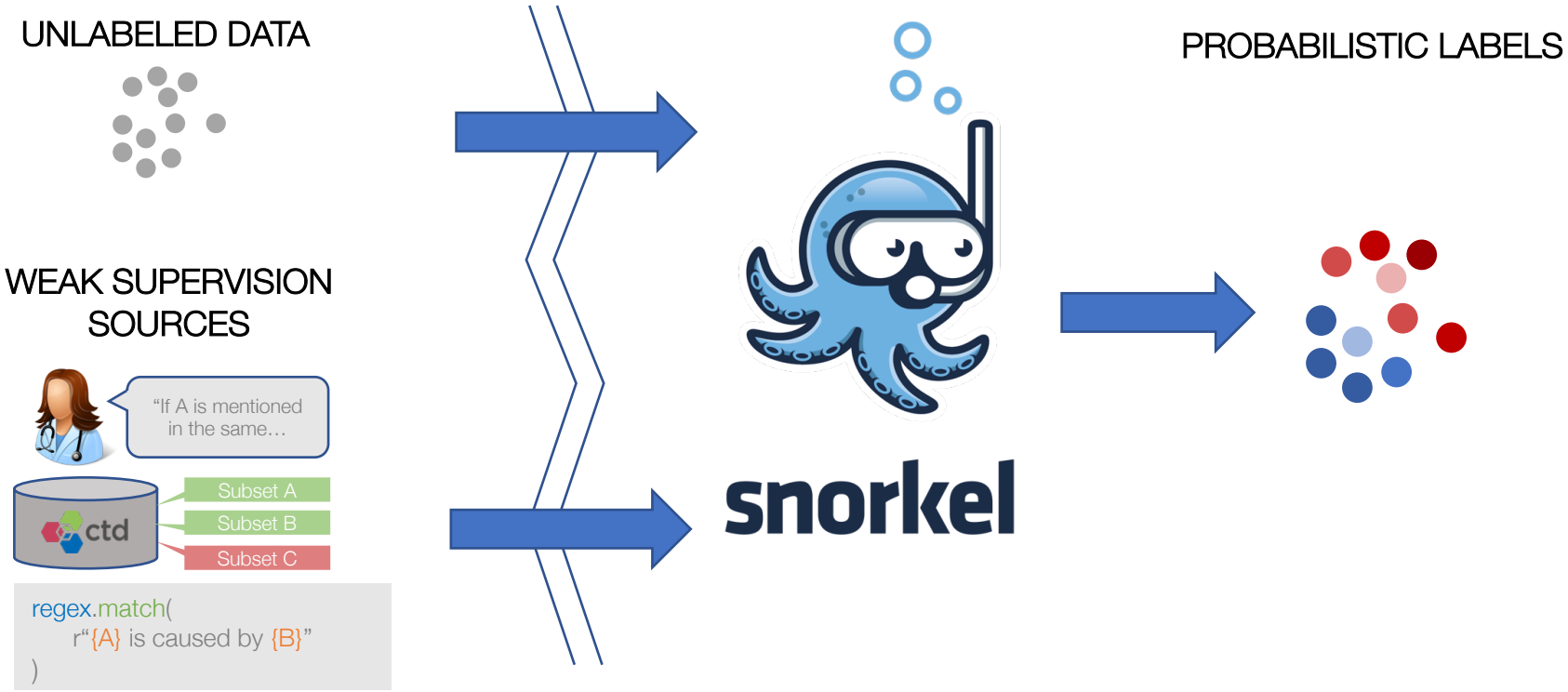
Crowdsourcing

amazon mechanical turk

[e.g. Dalvi 2013]

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

Snorkel: Formalizing Programmatic Labeling

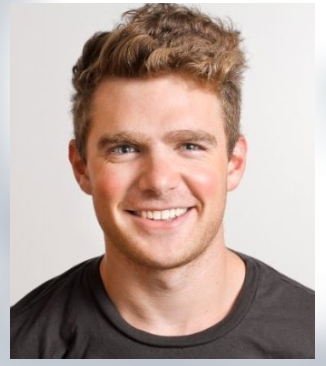


Goal: Replace *ad hoc* weak supervision with a formal, unified, theoretically grounded approach for programmatic labeling

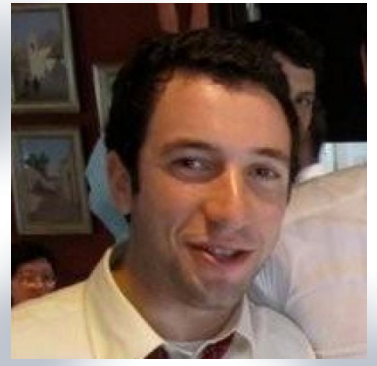
The Real Work

Stephen Bach

Braden Hancock



Henry Ehrenberg



Alex Ratner

Paroma Varma

Snorkel.org

Running Example: NER

PER:DOCTOR

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

ORG:HOSPITAL

Let's look at labeling "Person" versus "Hospital"

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

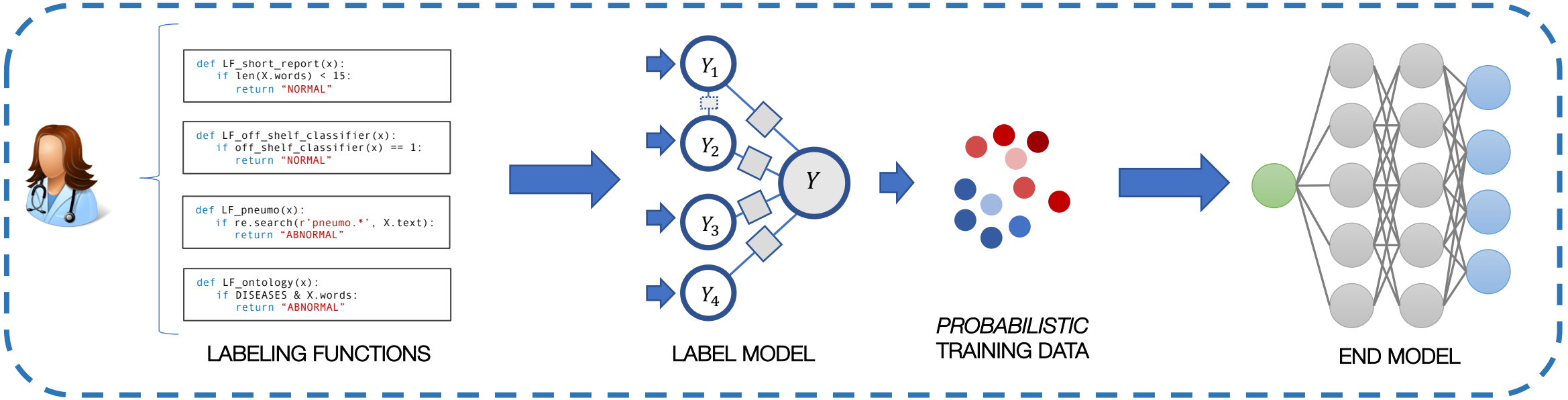
→ “PERSON”

```
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.

People use it…

snorkel

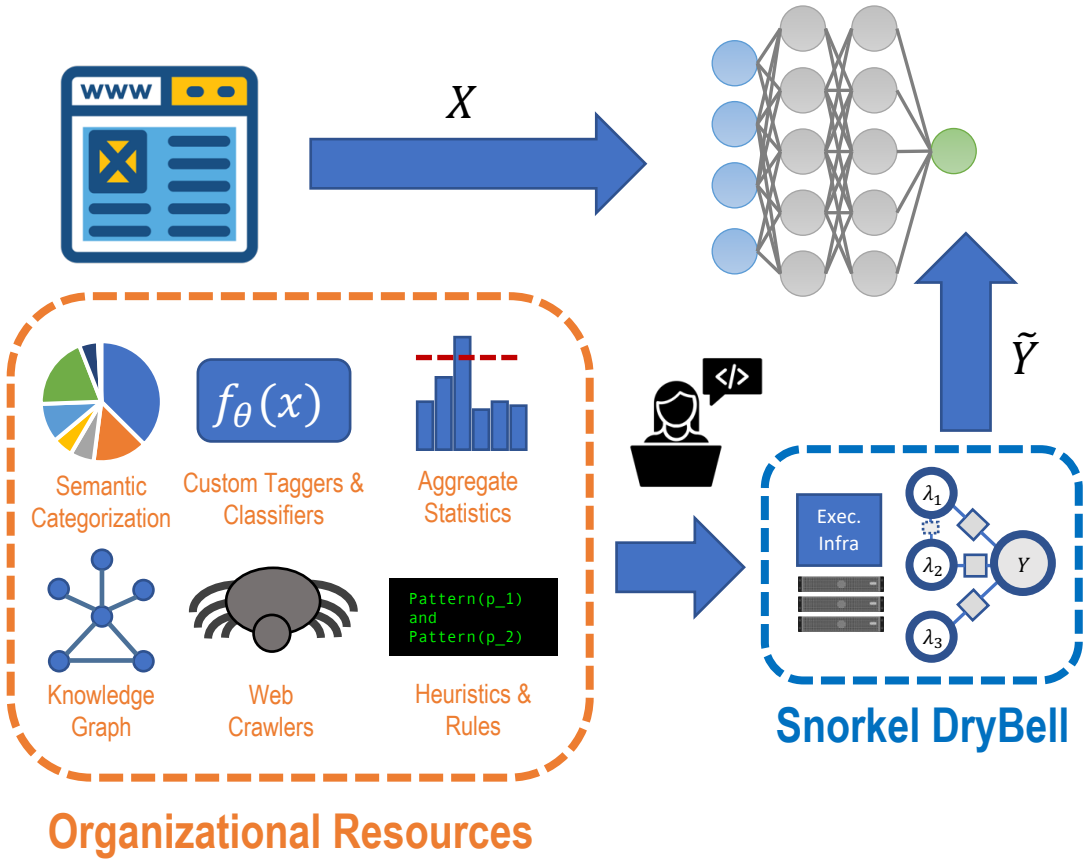
Http://snorkel.org

"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.



Thank you, Google!
Even best funded teams…

[Bach et. al., SIGMOD 2019]

Maybe you have used it?

Overton: A Data System for Monitoring and Improving Machine-Learned Products

Christopher Ré, Apple
Feng Niu, Apple
Pallavi Gudipati, Apple
Charles Srisuwananukorn, Apple

Migrating a Privacy-Safe Information Extraction System to a Software 2.0 Design

Ying Sheng, Google, Mountain View, CA, USA, yingsheng@google.com
Nguyen Vo, Google, Mountain View, CA, USA, nguyenvo@google.com
James B. Wendt, Google, Mountain View, CA, USA, jwendt@google.com
Sandeep Tata, Google, Mountain View, CA, USA, tata@google.com
Marc Najork, Google, Mountain View, CA, USA, najork@google.com

Leveraging Organizational Resources to Adapt Models to New Data Modalities

Sahaana Suri[†]*; Raghuveer Chanda, Neslihan Bulut, Pradyumna Narayana, Yemao Zeng
Peter Bailis[†], Sugato Basu, Girija Narlikar, Christopher Ré[†], Abishek Sethi
Google, Stanford[†]

https://arxiv.org/abs/1909.05372

CIDR2020

It has changed use real systems…

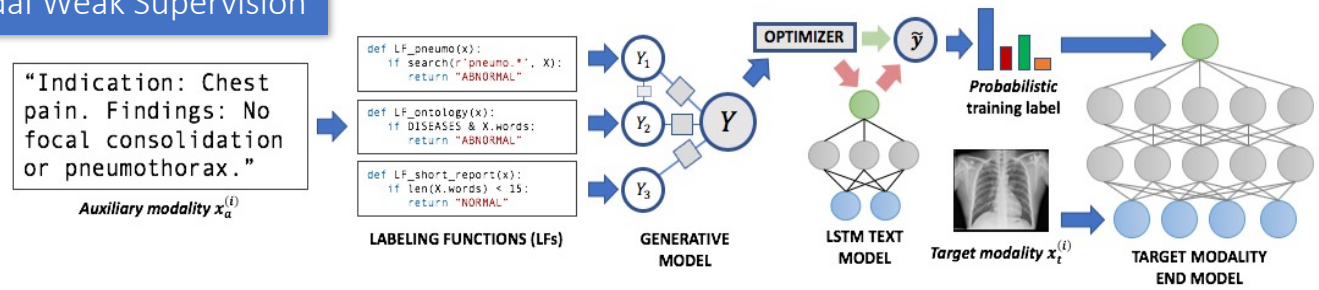
Resourcing	Error Reduction	Amount of Weak Supervision
High	65% (2.9×)	80%
Medium	82% (5.6×)	96%
Medium	72% (3.6×)	98%
Low	40% (1.7×)	99%

A couple of highlights

- Used by multiple teams with good error reduction over production.
- Take away: many systems are almost entirely weak supervision based.

Weak Supervision in Science & Medicine

Cross-Modal Weak Supervision

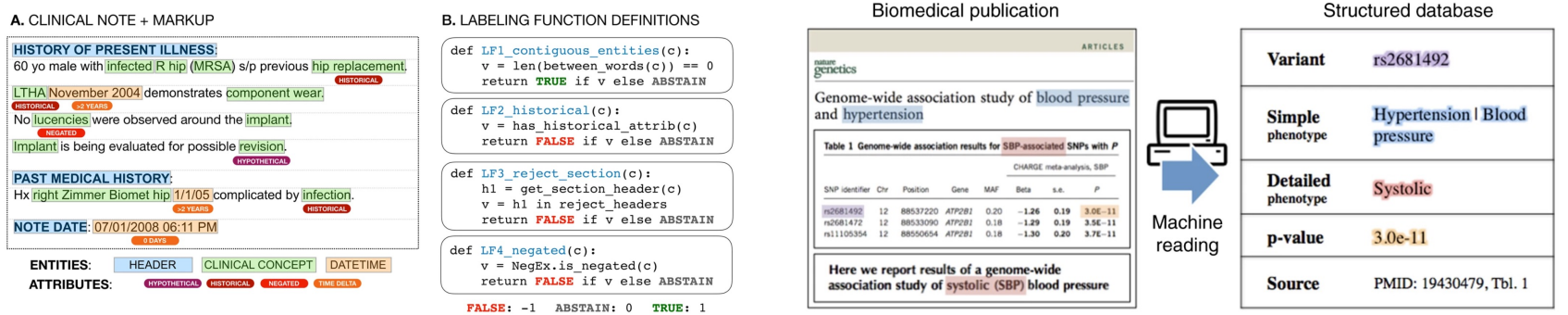


J. Dunnmon et al., "Cross-Modal Data Programming Enables Rapid Medical Machine Learning," *2020.*

Blog: http://hazyresearch.stanford.edu/ws4science

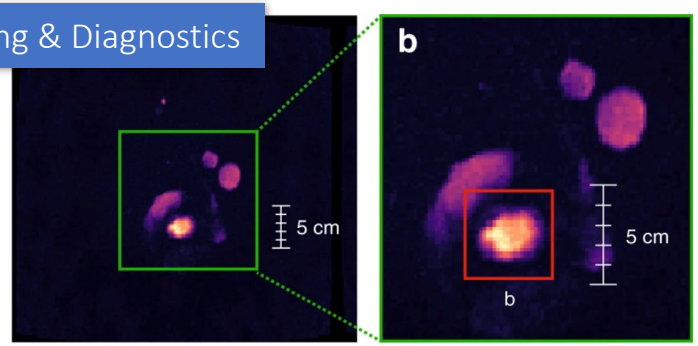
Text & Extraction

A. Callahan et al., NPJ Dig Med, 2020

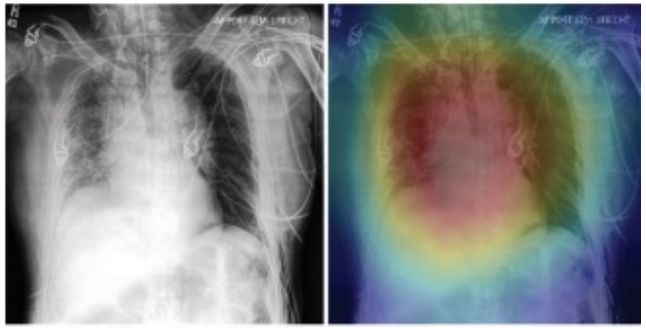


V. Kuleshov et al., Nat Comms, 2019

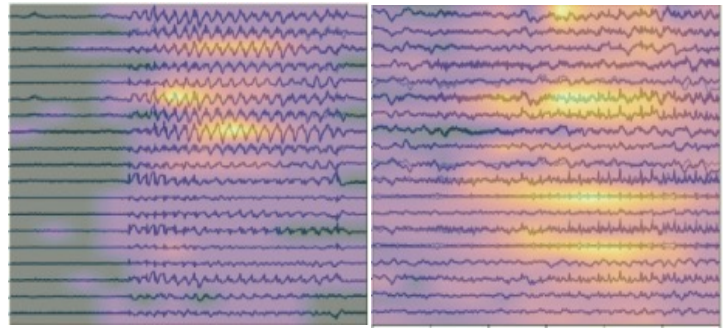
Imaging & Diagnostics



J. Fries et al., Nat Comms, 2019



J. Dunnmon et al., Radiology, 2019



K. Saab et al., NPJ Dig Med, 2020

High-Level Related Work

LUDWIG

snorkel

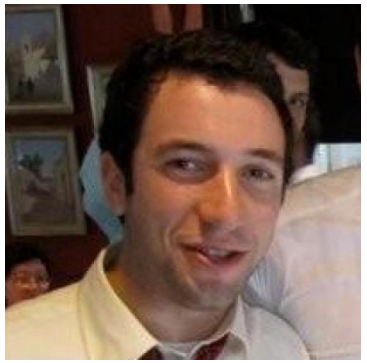
Software 2.0

Andrej Karpathy Follow

Nov 11, 2017 · 8 min read

Snorkel DryBell: A Case Study in Deploying Weak Supervision at Industrial Scale

Core ML

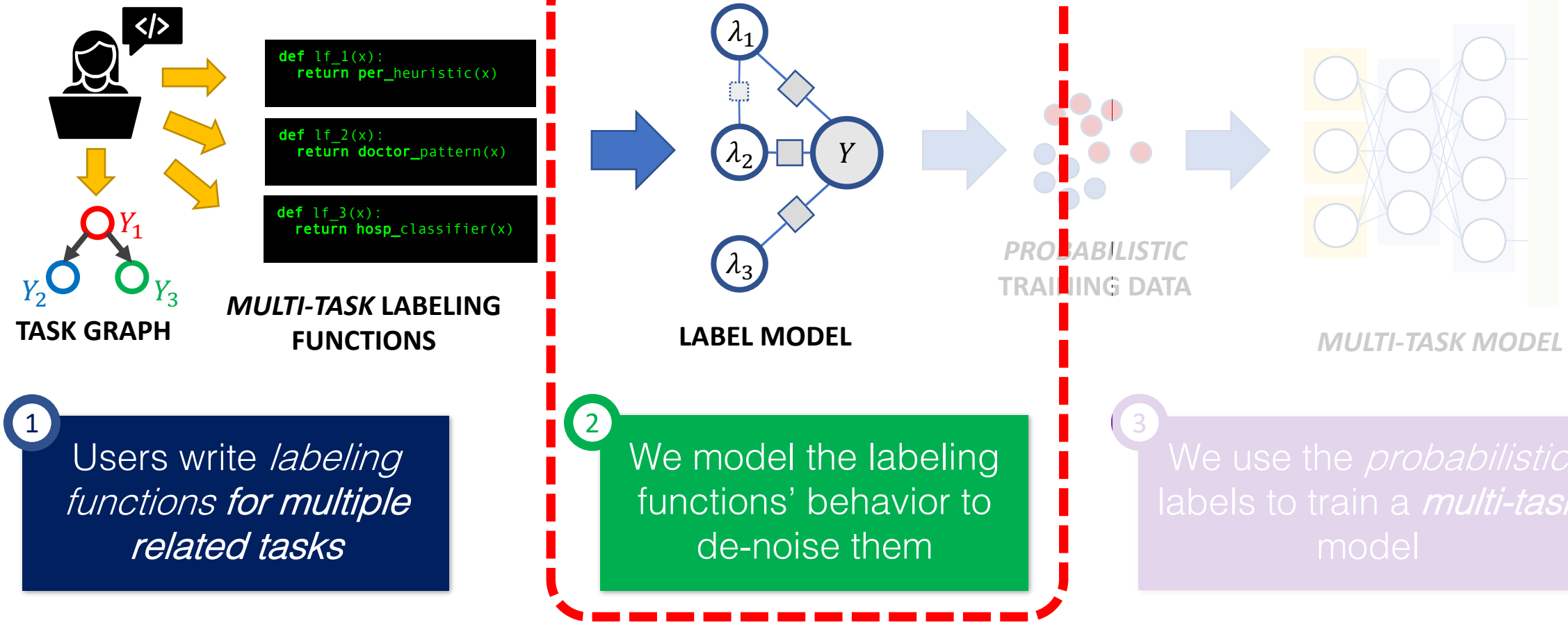


Alex Ratner
(to Washington)

Fred Sala
(to Wisconsin)

Let’s look under the hood and take a peak at some math (to the whiteboard soon..)

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

Later: We will define what this picture means precisely.

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

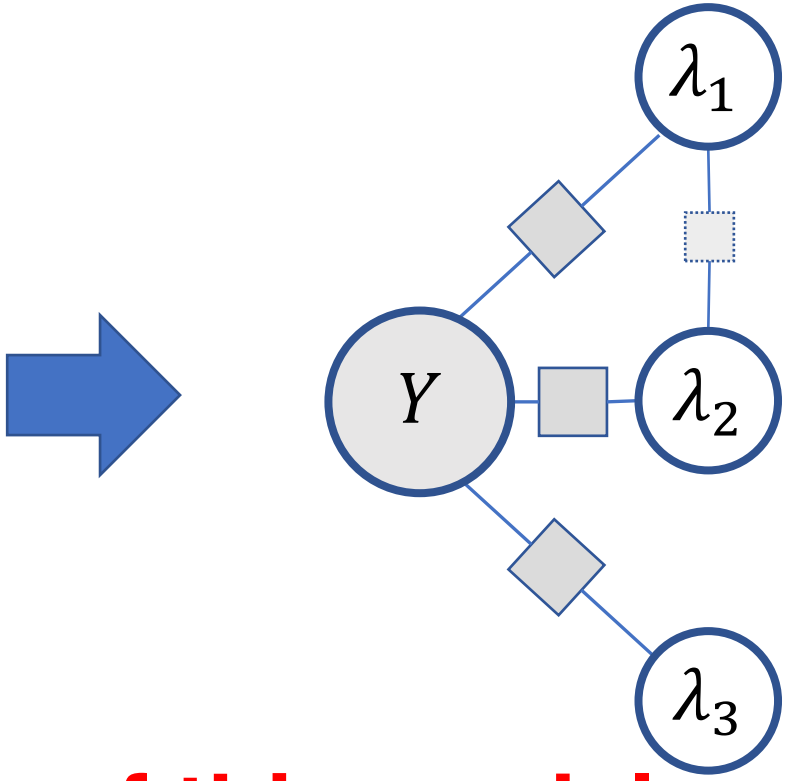
→ "PERSON"

```
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?

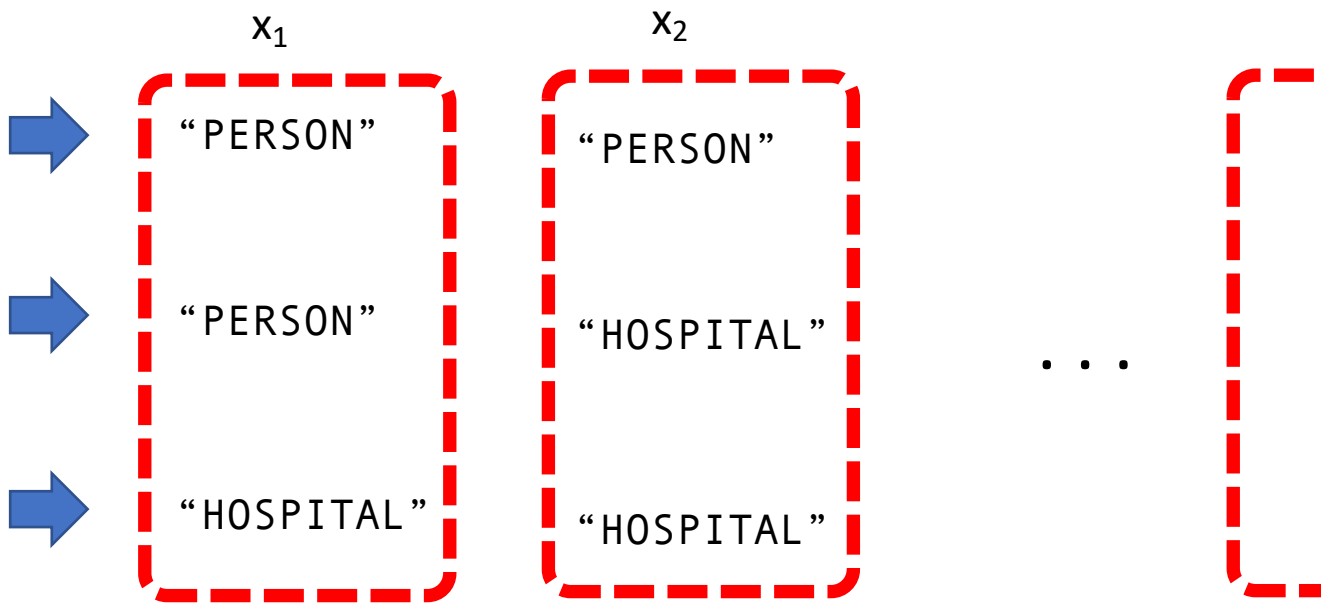
Intuition: Learn from the Overlaps

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 observe agreements (+1) and disagreements (-1) on many points! (More later!)