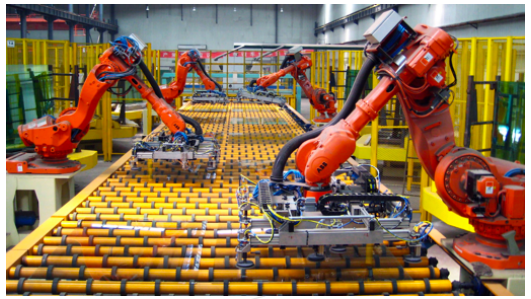
Theory & Systems for Weak Supervision

Chris Ré

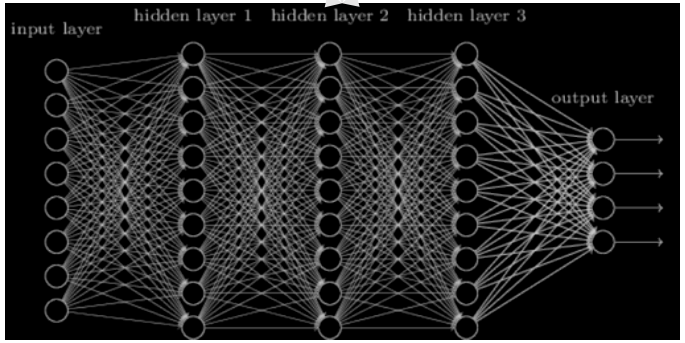
Stanford University

http://hazyresearch.stanford.edu/people/

Software 2.0 is eating Software 1.0



1000x Productivity: Google shrinks language translation code from 500k LoC to 500 lines of dataflow.



<u>Classical problems ML 1st</u>

- ETL & Cleaning (Holoclean.io)
- DB Tuning Peloton (CMU)
- Networks Pensieve (MIT) .
- NeuroCore (Stanford)

https://jack-clark.net/2017/10/09/import-ai-63-google-shrinks-language-translation-code-rom-500000-to-500-lines-with-ai-only-25-of-surveyed-people-believe-automationbetter-jobs

"Software 2.0", Andrej Karpathy, https://medium.com/@karpathy/software-2-0-a64152b37c35

Easier to build, deploy, and maintain

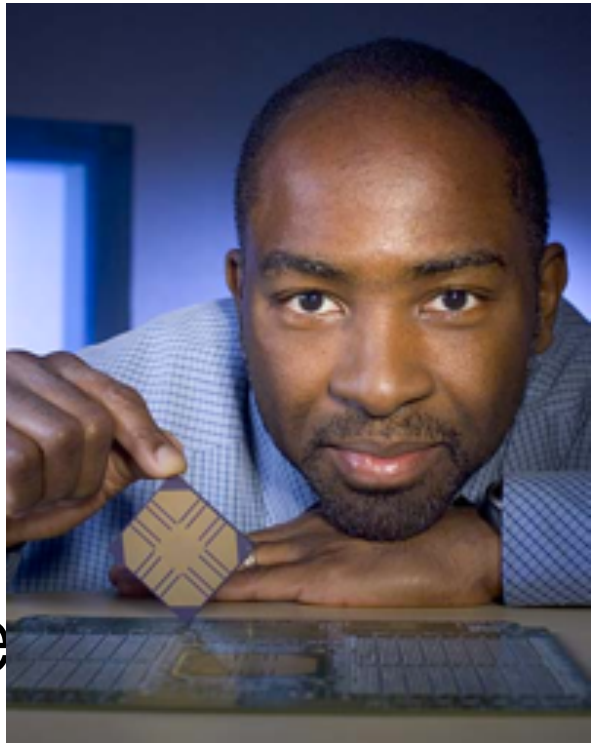
Build products faster. Speed is amazing.

Deploy is critical: NNs “new JVM”

- Dataflow has regular run-times.
  - Qualification easier means “ship faster.”
- See Kunle’s ISCA/NeurIPS keynote for more info.

Maintain: ”retrain”—no “ninja” dependence

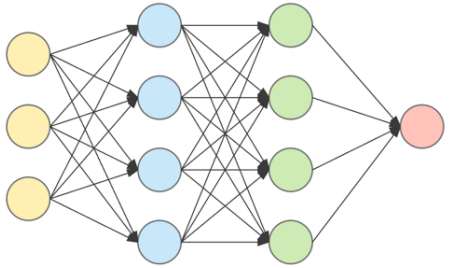
SW2.0 View: eng. changes are **significant**.



Kunle Olukotun

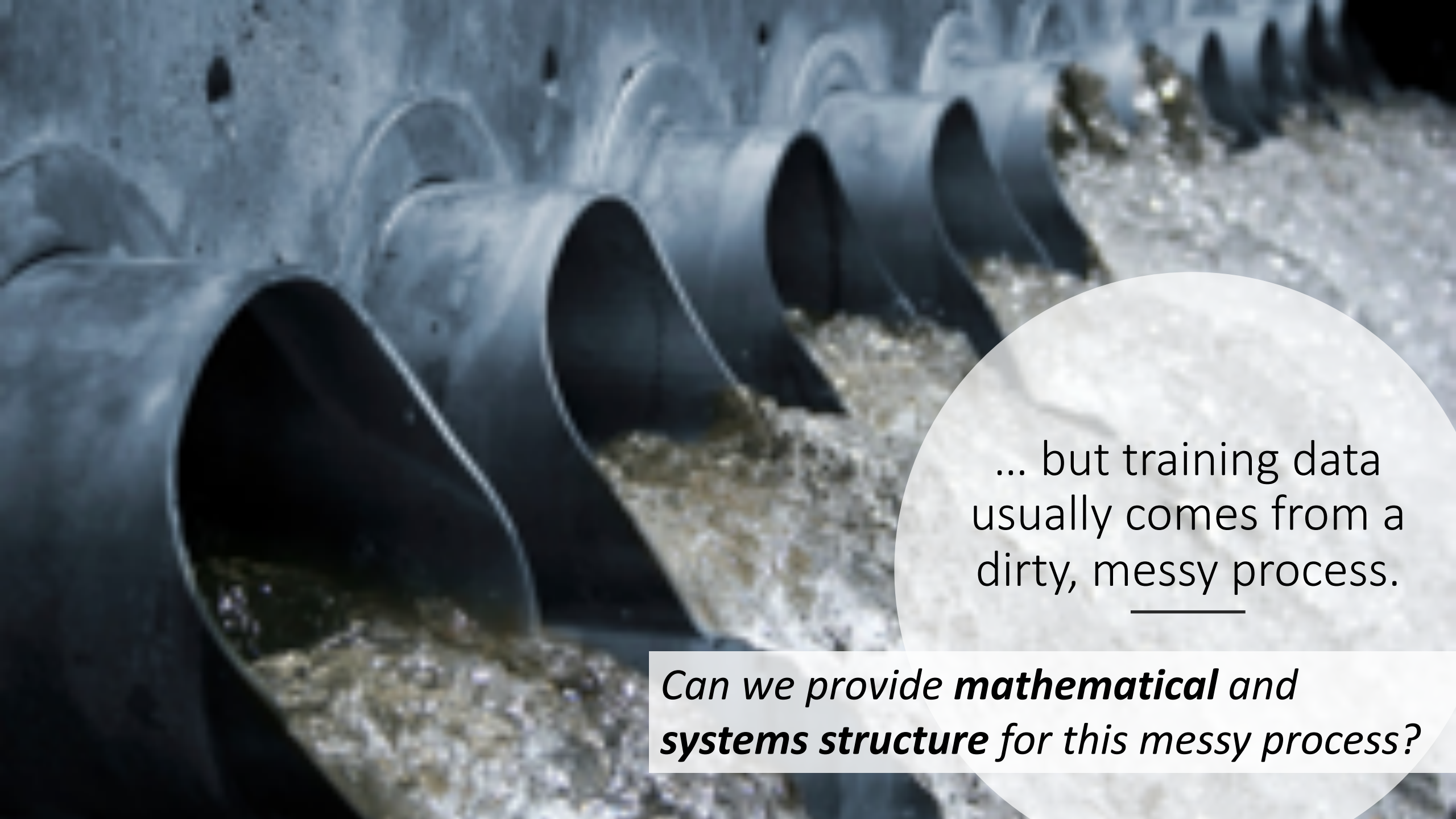
ML Application =

Model + Data + Hardware



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

But supervision comes from god herself….



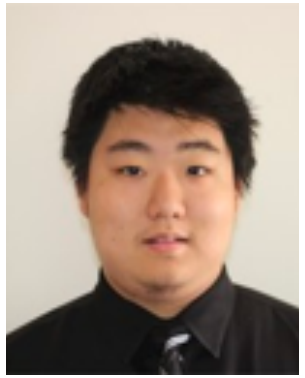
… but training data usually comes from a dirty, messy process.

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

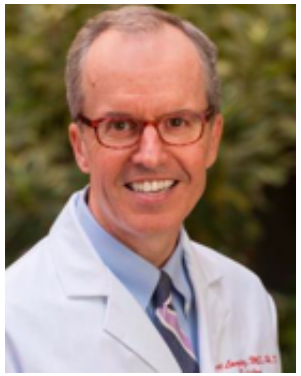
Supervision is where the action is…

*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 January 19

RSNA® Radiological Society of North America

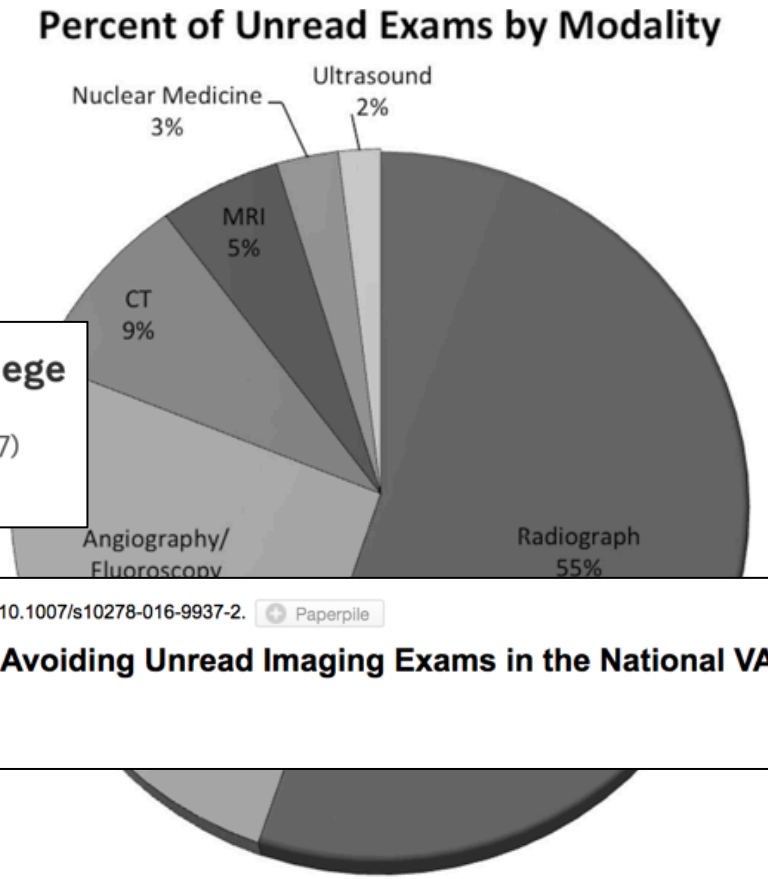
Radiology

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!

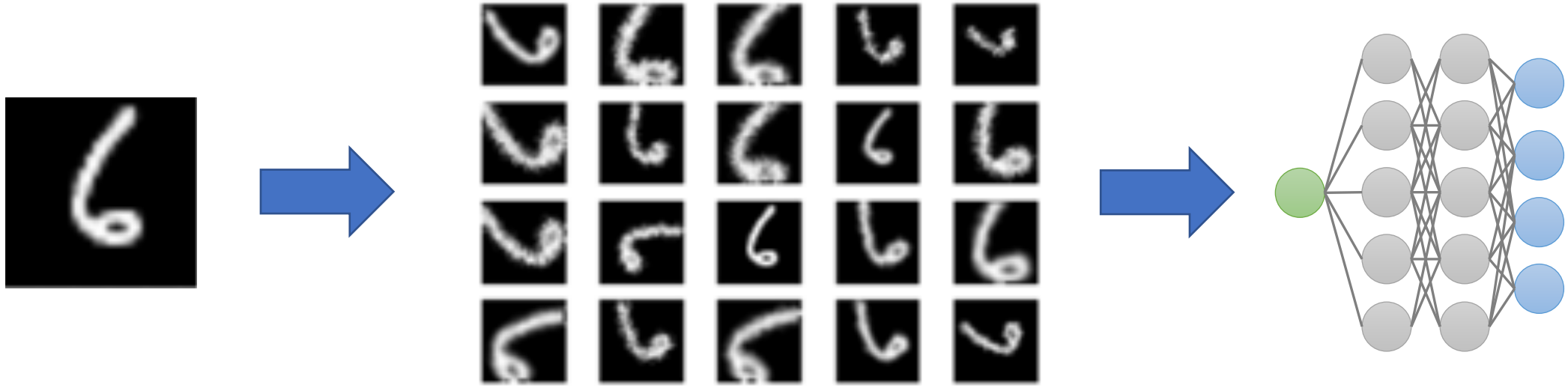
- Created large dataset of clinical labels
- Evaluated effect of label quality
- Work published in a *clinical journal*

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.

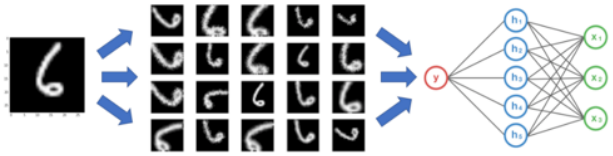
Even in Benchmarks: Data Augmentation is Critical



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**

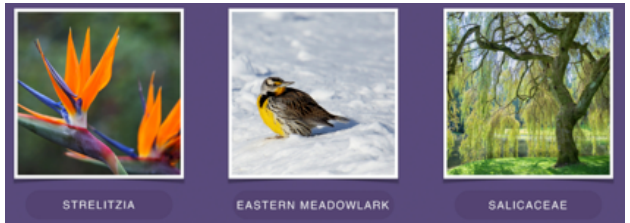
- **Augmentation Policies** first in Ratner et al. NIPS ’17

Henry Ehrenberg

Alex Ratner (to: Washington)

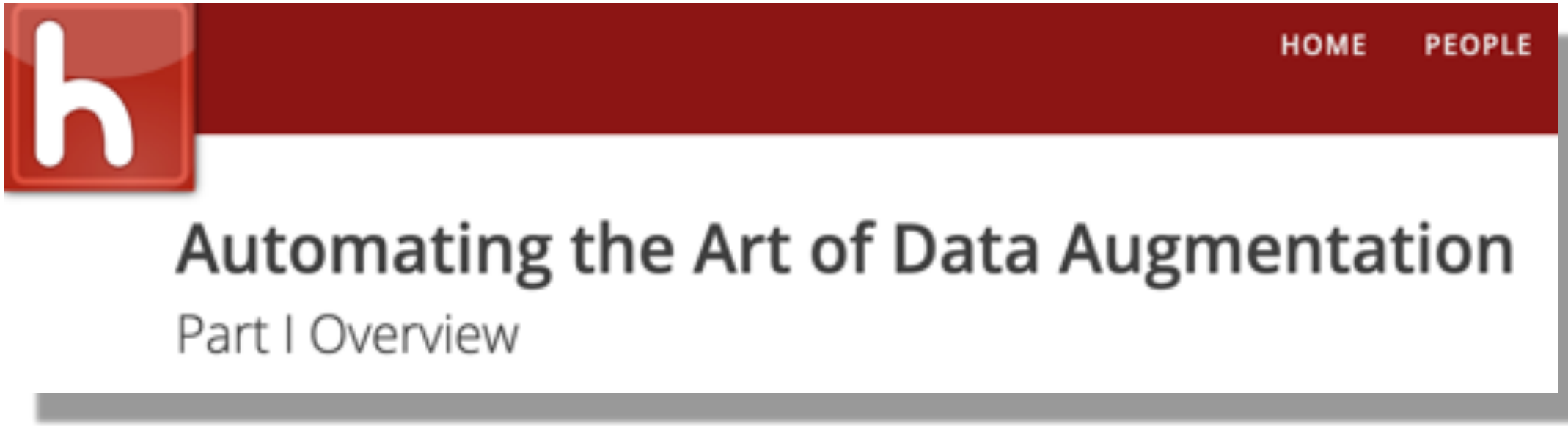
Facebook Hash tag weakly supervised pre-training

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

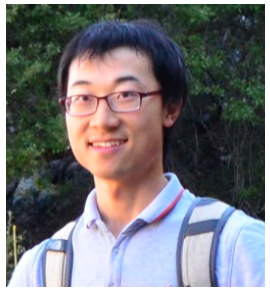


Sharon Y. Li (to: Wisconsin)

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



Sharon Y. Li
(to: Wisconsin)



Hongyang Zhang

Sen Wu

Our approach: Uncertainty-based sampling

- ***Key idea:*** Instead of randomly sampling, reduce the frequencies of transformations that the neural net has learned!
- ***Empirical result*:** **84.54%** on CIFAR-100 using Wide-ResNet-28-10, improves RandAugment (Cubuk et al.'19) by **1.24%**.
- ***Theory:*** Analyze the effect of different transformations in a high-dimensional setting, including revealing the *regularization effect* of a curious mixup augmentation!

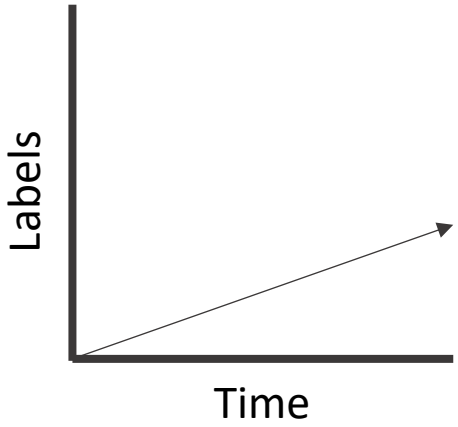
Blog post: hazyresearch.stanford.edu/data-aug-part-3, ***Code:*** https://bit.ly/32E2V7n

Training data: the new bottleneck

Slow, expensive, and static

Manual Labels

Slow



Expensive

$10 - $100/hr

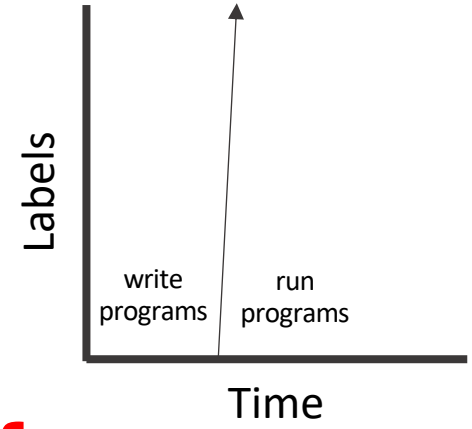
Static

{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

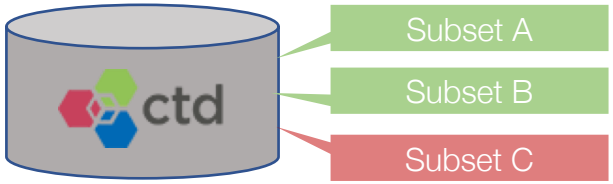
snorkel

Pattern Matching

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

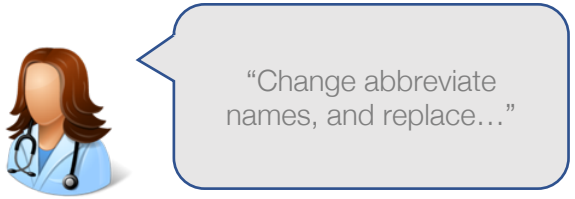
[e.g. Hearst 1992, Snow 2004]

Distant Supervision

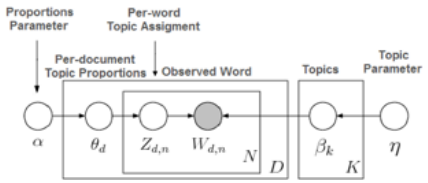


[e.g. Mintz 2009]

Augmentation

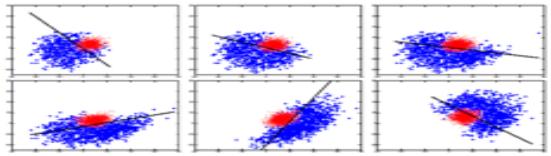


Topic Models



[e.g. Hingmire 2014]

Third-Party Models



[e.g. Schapire 1998]

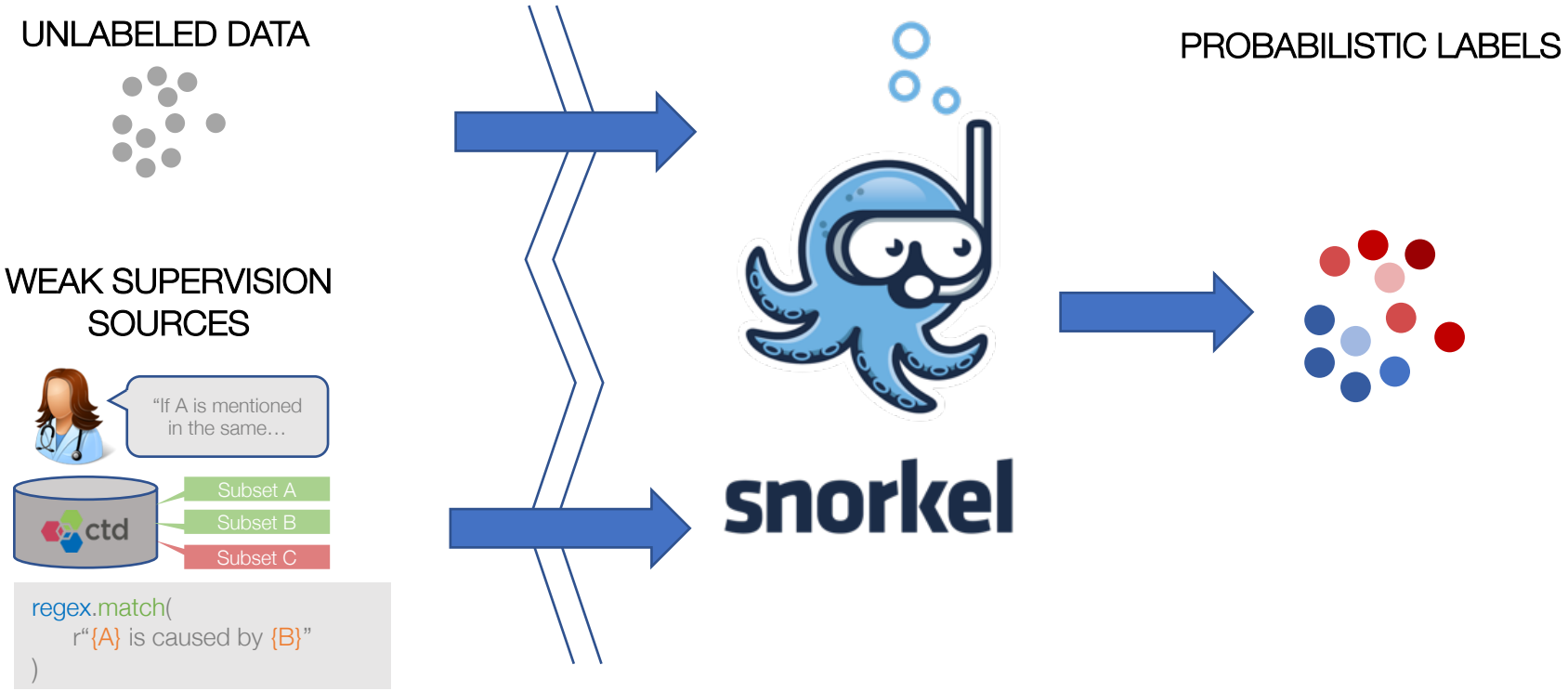
Crowdsourcing

amazon mechanical turk

[e.g. Dalvi 2013]

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

Snorkel: Formalizing Programmatic Labeling

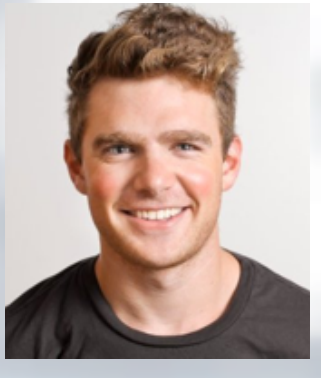


Goal: Replace *ad hoc* weak supervision with a formal, unified, theoretically grounded approach for programmatic labeling

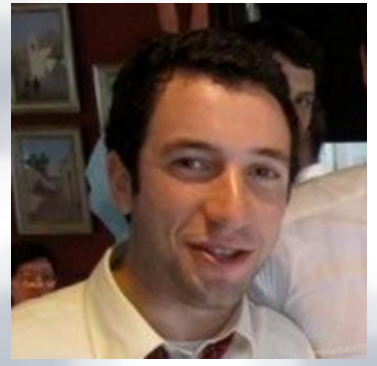
The Real Work

Stephen Bach

Braden Hancock



Henry Ehrenberg



Alex Ratner

Paroma Varma

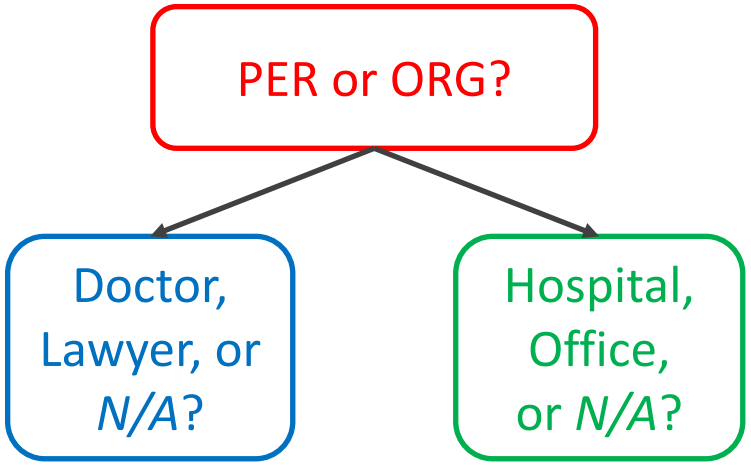
Snorkel.org

Running Example: NER

PER:DOCTOR

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

ORG:HOSPITAL



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
```

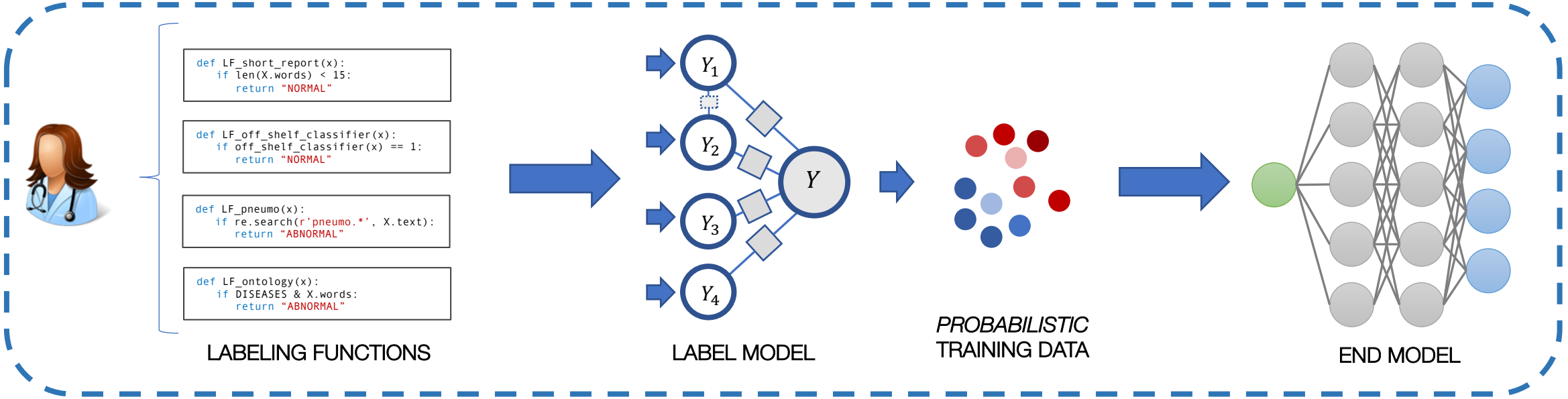
→ “PERSON”

```
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



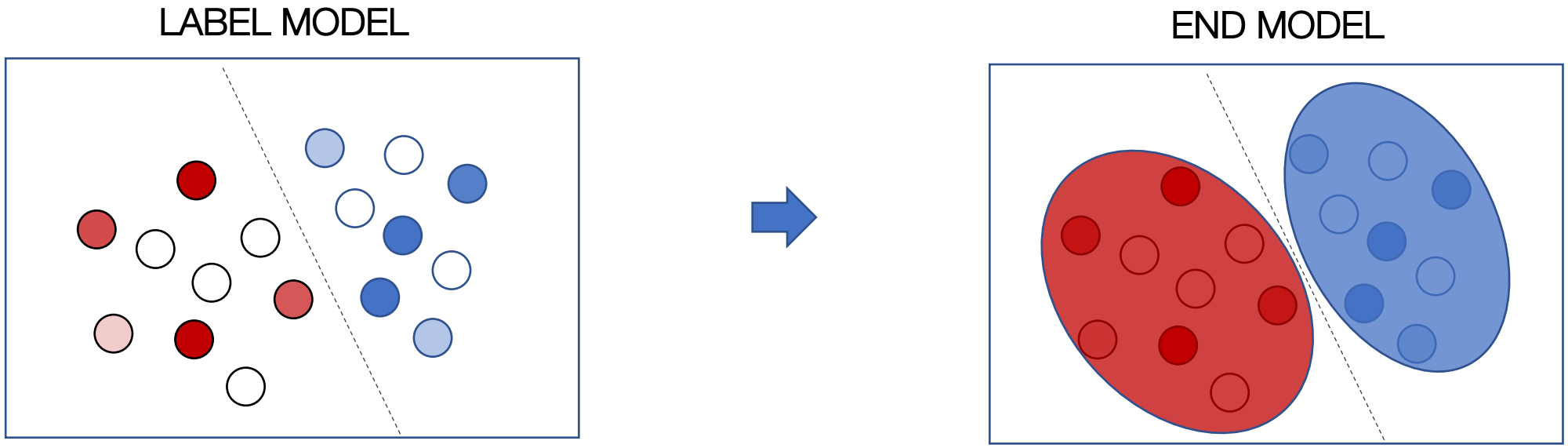
1. Users write *labeling functions* to generate noisy labels

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:

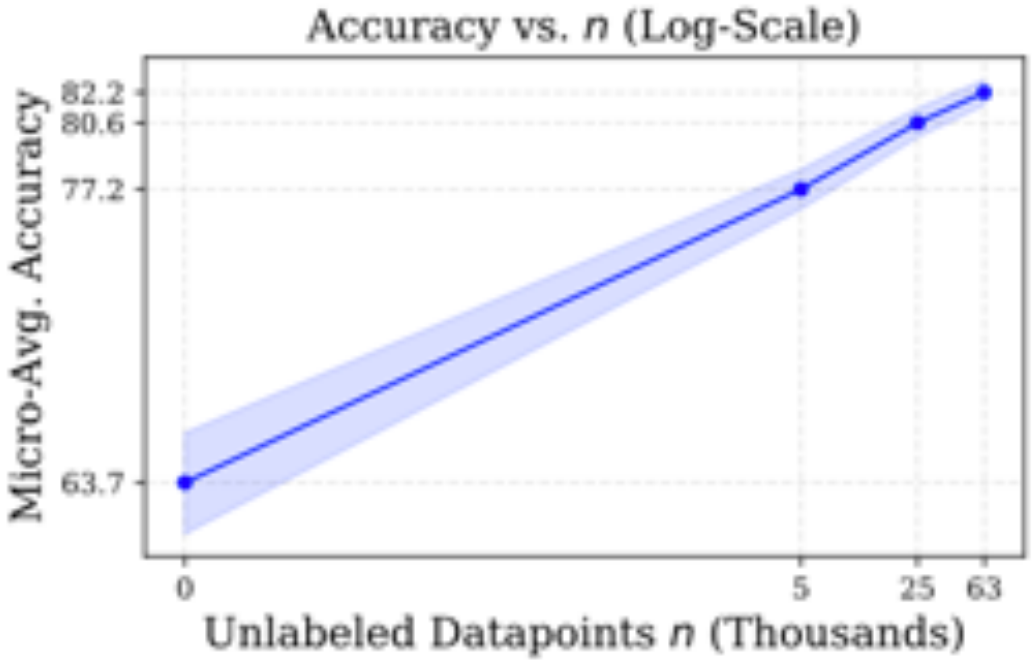
"treats", "causes", "induces", "prevents", …

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



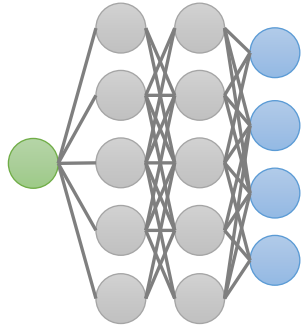
Add more unlabeled data—without changing the LFs—and performance improves!

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

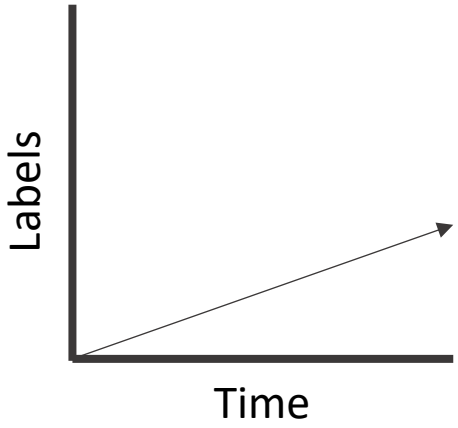
Stanford MEDICINE

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

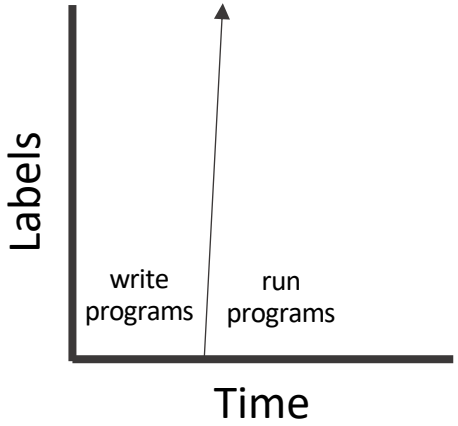
Static

{Positive, Negative}

{Positive, Neutral, Negative}

Programmatic Labels

Fast



Cheap

aws

$0.10/hr

Dynamic

Snorkel: In use at the world's largest companies

snorkel

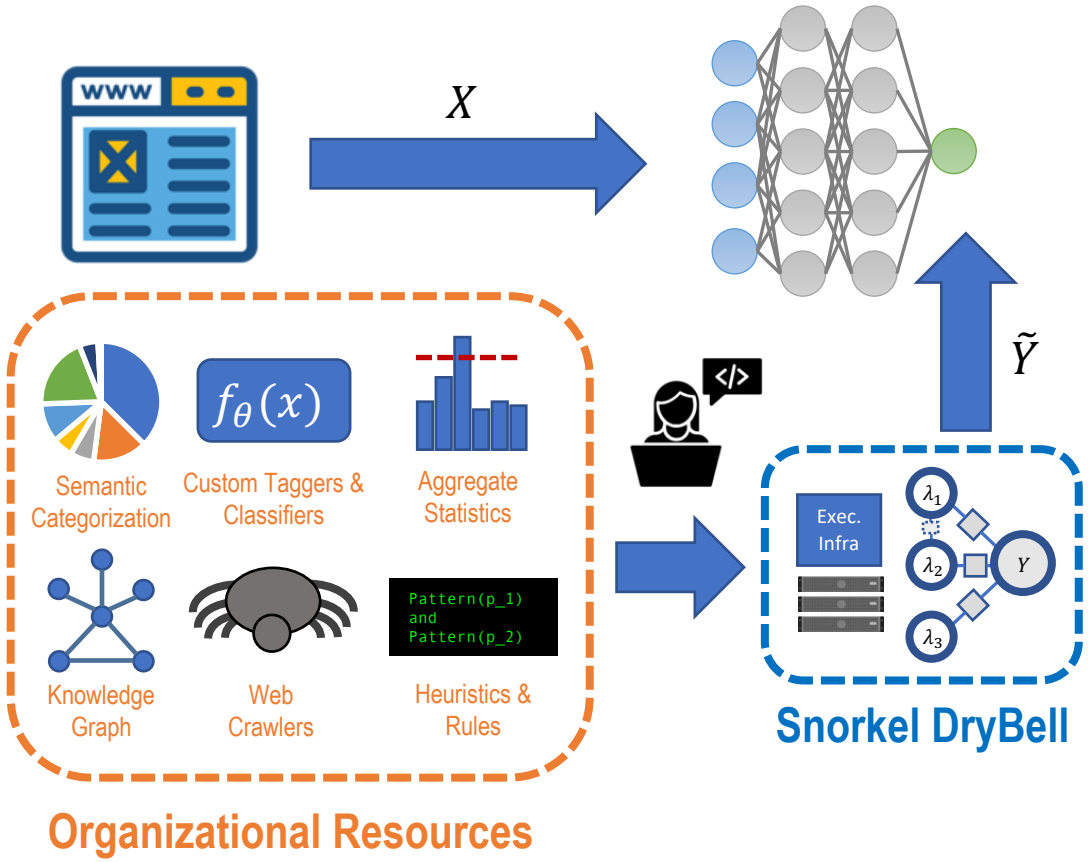
Http://snorkel.org

"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.



Thank you, Google!
(More soon)

[Bach et. al., SIGMOD 2019]

You probably 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

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.

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

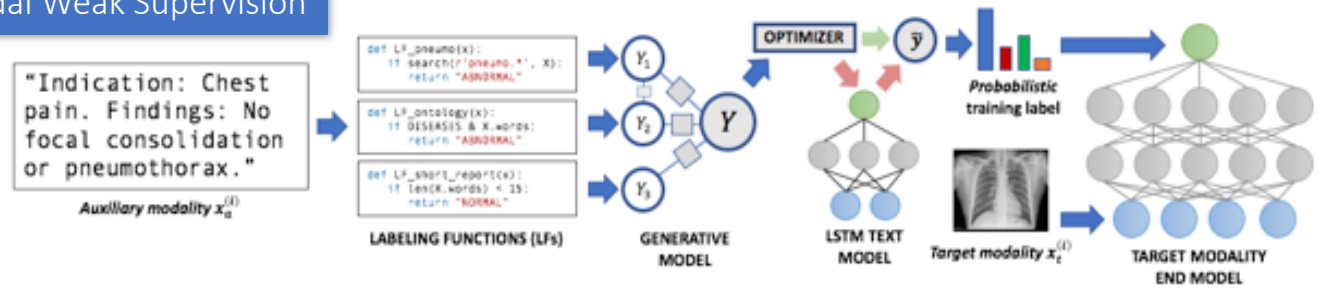
TensorFlow

PyTorch

Core ML

Weak Supervision in Science & Medicine

Cross-Modal Weak Supervision

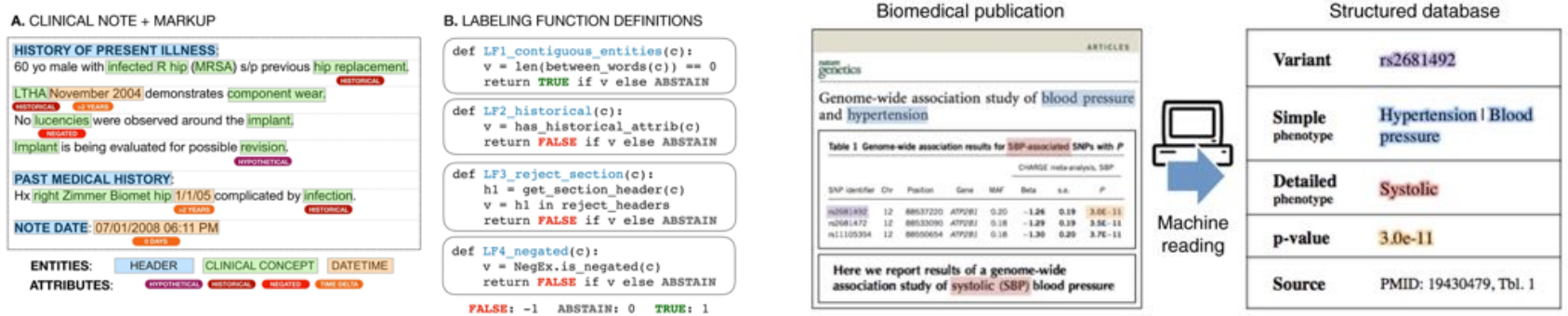


J. Dunnmon et al., "Cross-Modal Data Programming Enables Rapid Medical Machine Learning," *2020.*

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

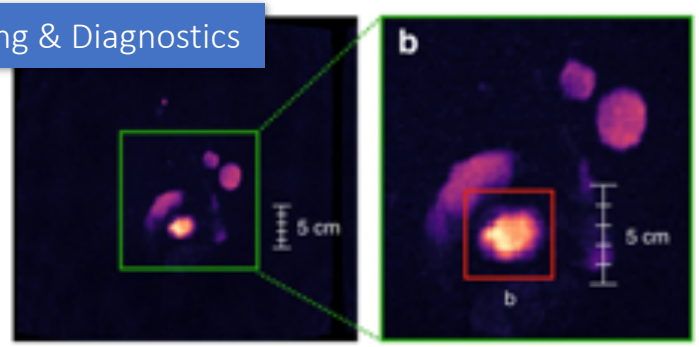
Text & Extraction

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

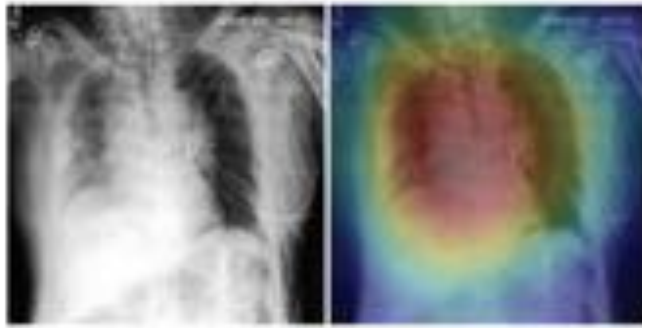


V. Kuleshov et al., Nat Comms, 2019

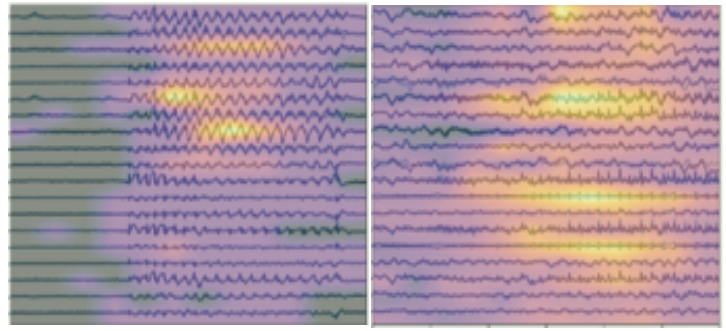
Imaging & Diagnostics



J. Fries et al., Nat Comms, 2019



J. Dunnmon et al., Radiology, 2019

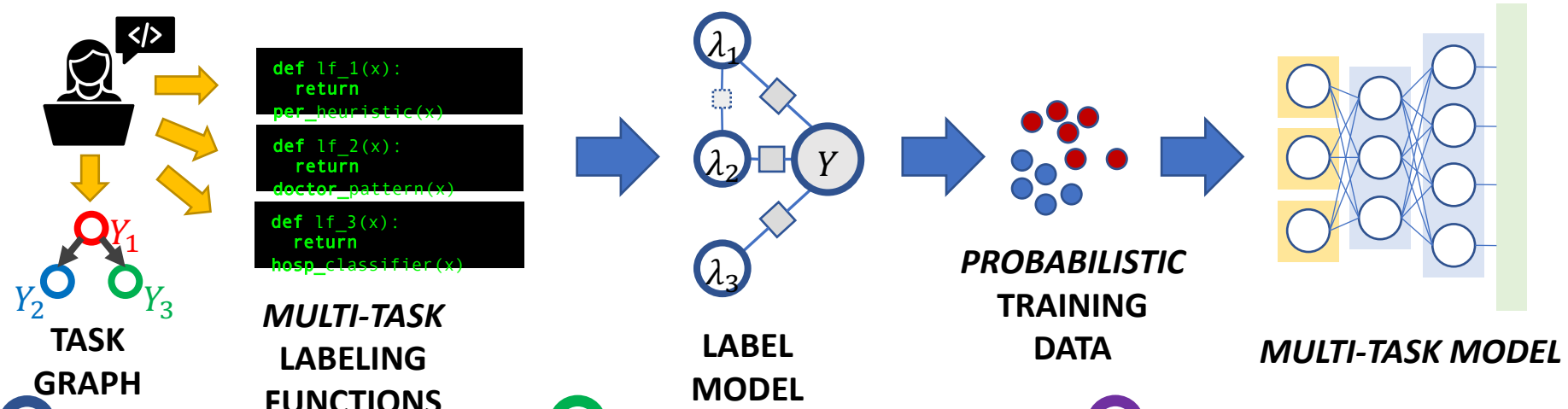


K. Saab et al., NPJ Dig Med, 2020

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

Fred Sala. *On the market **NOW!***

The Snorkel Pipeline



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

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

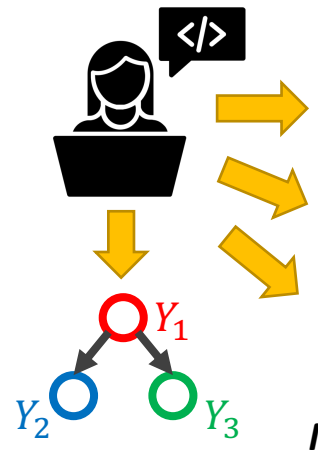
3. We use the *probabilistic* labels to train a ***multi-task*** model

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

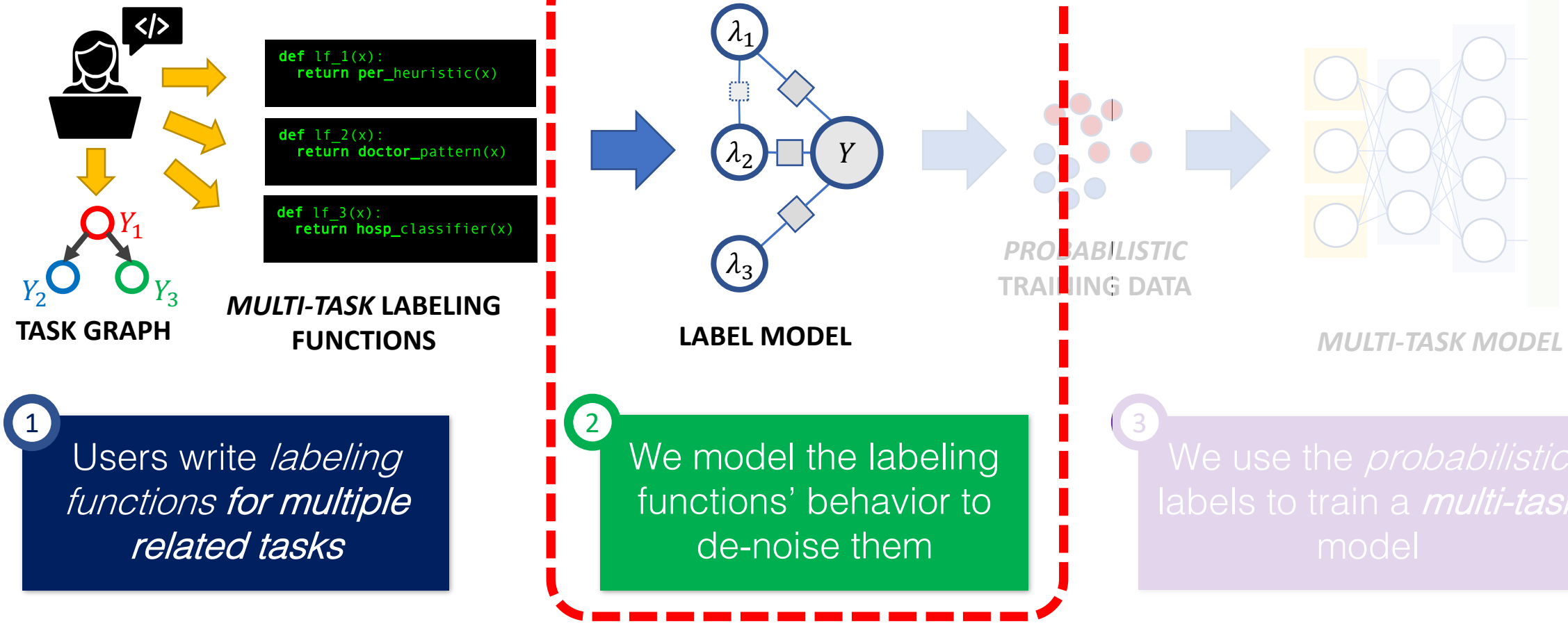
1 Users write *labeling functions for multiple related tasks*

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)
```

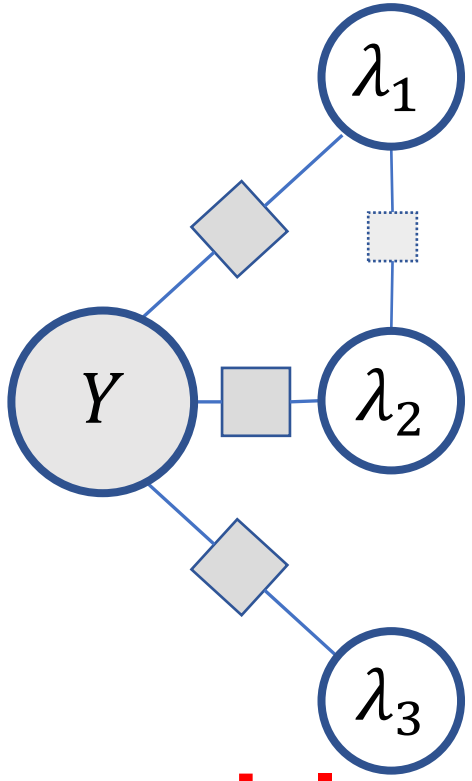
→ “PERSON”

```
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?

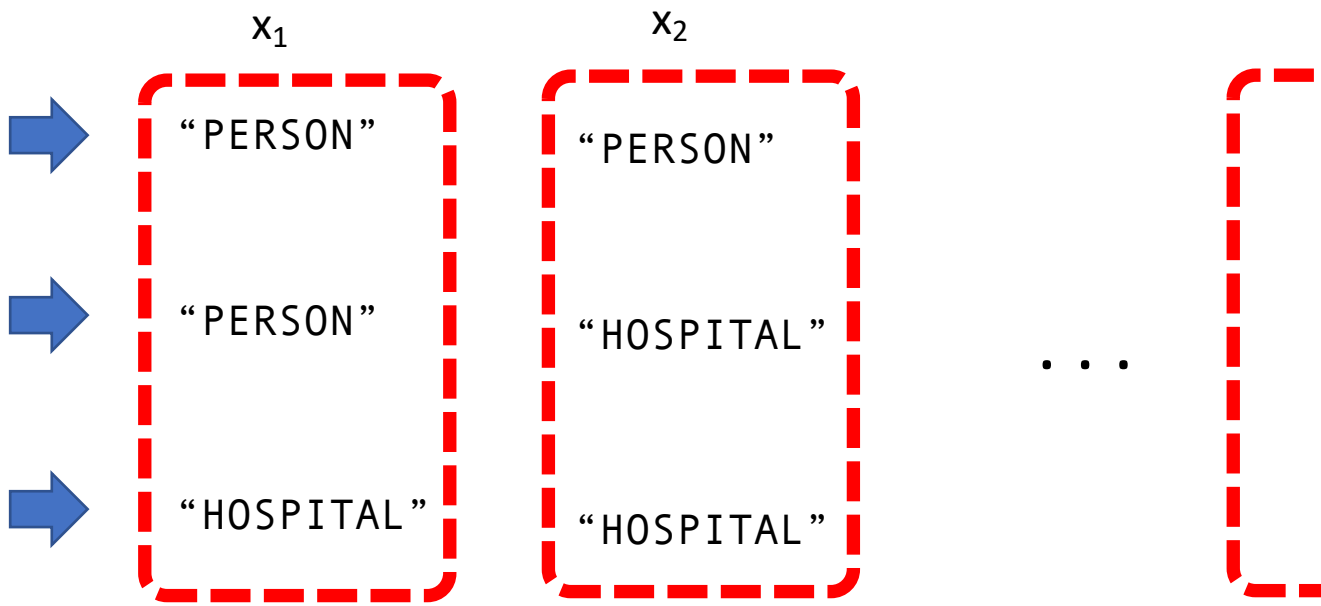
Intuition: Learn from the Overlaps

Sources.

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

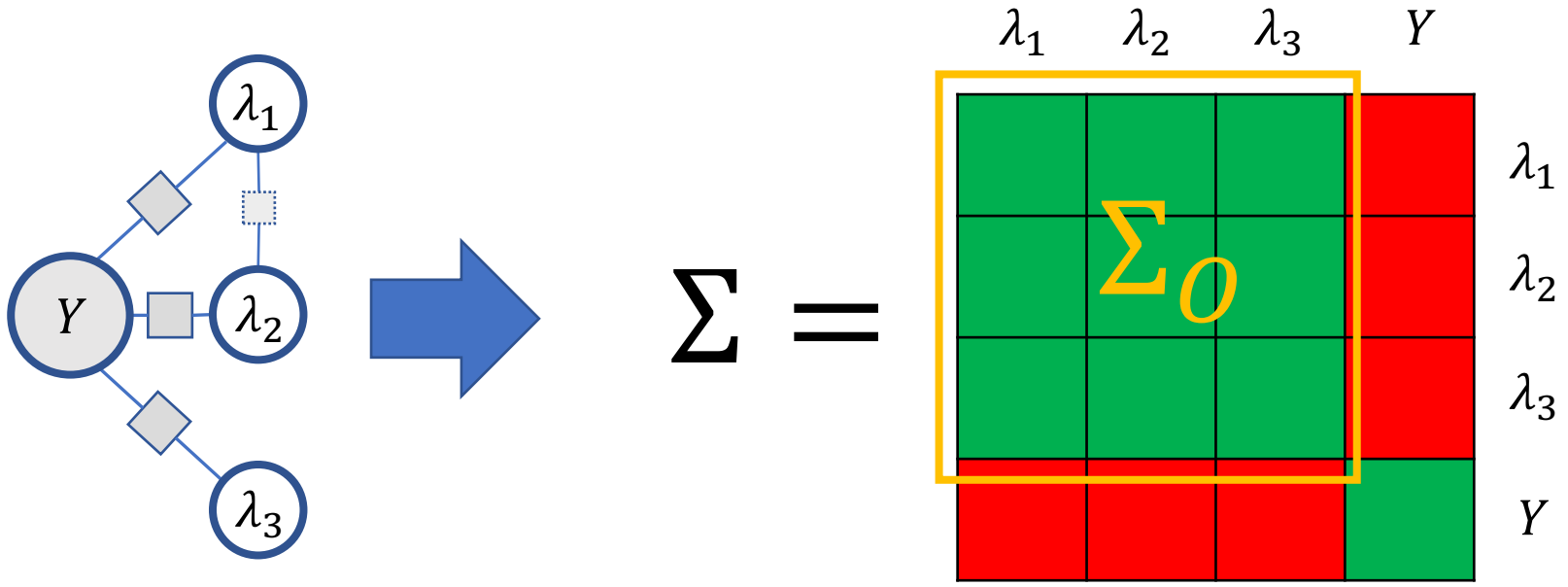
```
def upper_case_existing_classifier(x):
  if all(map(is_upper, x.split())) and \
off_shelf_classifier(x) == 'PERSON':
    return PERSON
```

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



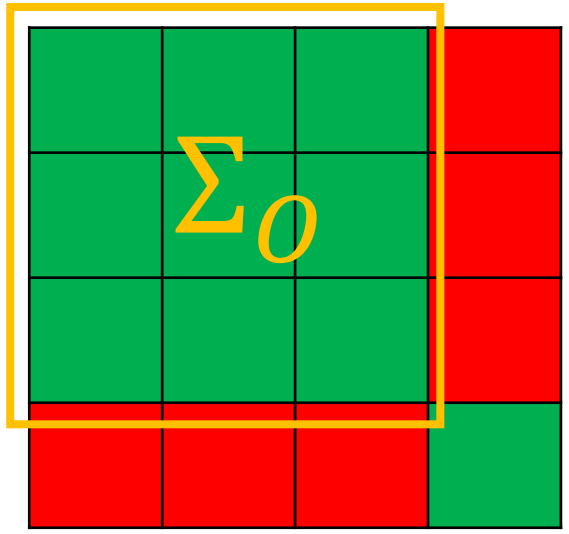
Key idea: We can observe overlapping judgements on many points!

Solution Sketch: Using the covariance



Can only observe ***portion*** of the covariance (Σ_O)…
if we observe rest, we'd be done! (E[yλ] ~ accuracy).

Idea: Use graph-sparsity of the inverse



λ_1
λ_2
λ_3
Y

Incompletely Observed

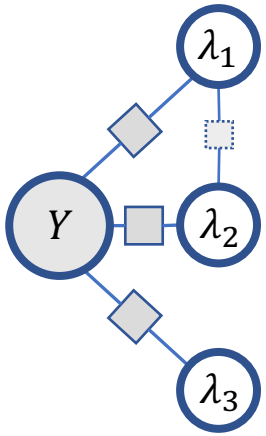
$$(\Sigma^{-1})_O$$

matrix inversion lemma

Observed overlaps

Rank-1 params to solve for (~ function of accuracies)

- $E[z_i] = 1$ if perfectly accurate
- $E[z_i] = 0$ if random noise



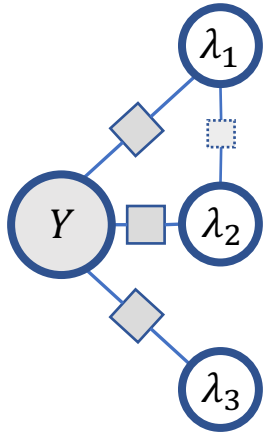
Fewer degrees of freedom: Roughly, zero where corresponding pair of variables has no edge

[Loh & Wainwright 2013, Ratner et al. 2019]

For now, assume we know the graph (dependency structure)…

Result: A matrix completion problem?

We get a set of equations. For any pair I != j with no edge in graph—the lhs is 0



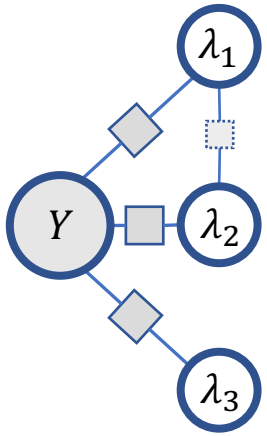
$$0 = \boxed{(\Sigma_O^{-1})_{i,j}} + \boxed{z_i z_j}$$

Observed overlaps

Low-rank parameters to solve for

Σ is full rank, so not really matrix completion…

Key: $\Sigma = I + uu^T$ for some u so intuitively close…



Couple of Technical Comments

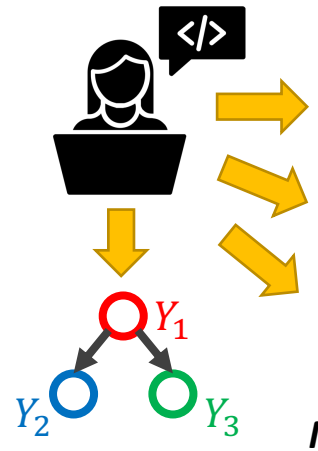
$$0 = \boxed{(\Sigma_O^{-1})_{i,j}} + \boxed{z_i z_j}$$

Observed overlaps

Low-rank parameters to solve for

- Symmetry: z and –z are solutions? What does this mean?
- $z_i = 0$ when accuracy 0.5, i.e., total noise! (more samples)
- Effective rank $er(\Sigma) = tr(\Sigma)/|\Sigma|_2$ (effectively, use this!)
  - small when single large: $|z|_2$ is large.
  - Scale inversely distance to noise (z_i = 0).

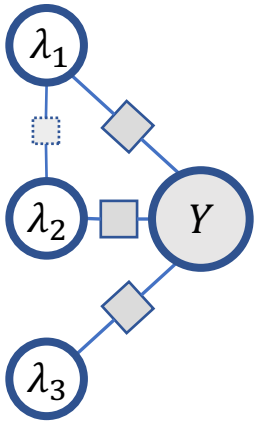
The Snorkel Pipeline



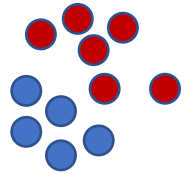
TASK GRAPH

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

def lf_2(x):
  return doctor_pattern(x)

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

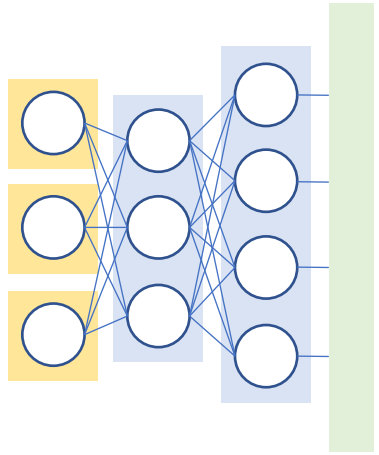
***MULTI-TASK* LABELING FUNCTIONS**



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

Recovery Results (Informal)

Result:

- Given n unlabeled data points--that overlap.
- And a sufficiently independent set of LFs (for recovery)
- The **end model test set error** should **decrease as $n^{-1/2}$**

$$E[\|l_{\hat{w}} - l_{w^*}\|] = O\left(\frac{1}{\sqrt{n}}\right)$$

Same asymptotic rate as with labeled data!

NB: Generalization straightforward—if you assume coverage.

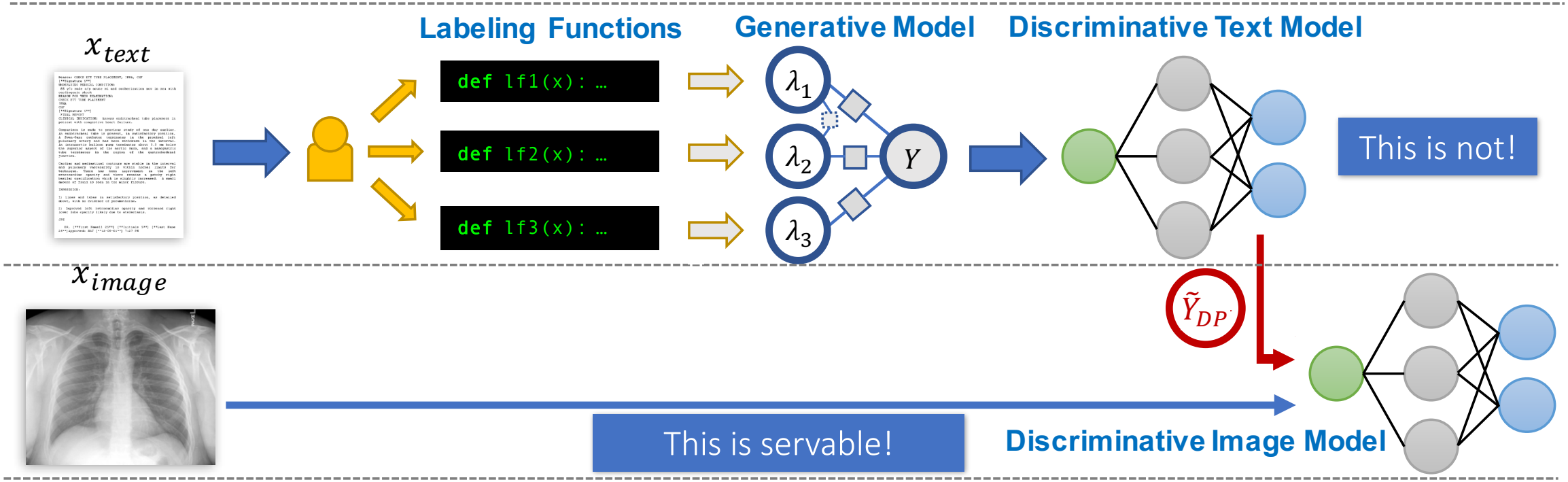
Empirical Results: NLP Experiments

	Ontonotes (Fine-grained NER)	TACRED (Relation Extraction)	OpenI (Document Classification)	Average
Gold Labels (n=300)	63.7 ± 2.1	28.4 ± 2.3	62.7 ± 4.5	51.6
Majority Vote	76.9 ± 2.6	43.9 ± 2.6	74.2 ± 1.2	65.0
Pipelined Snorkel	78.4 ± 1.2	49.0 ± 2.7	75.8 ± 0.9	67.7
Snorkel MeTaL	**82.2 ± 0.8**	**56.7 ± 2.1**	**76.6 ± 0.4**	**71.8**

- Avg. over Traditionally Supervised: **+ 20 points**
- Avg. over Majority Vote: **+ 7 points**
- Avg. over Single Task Modeling: **+ 4 points**

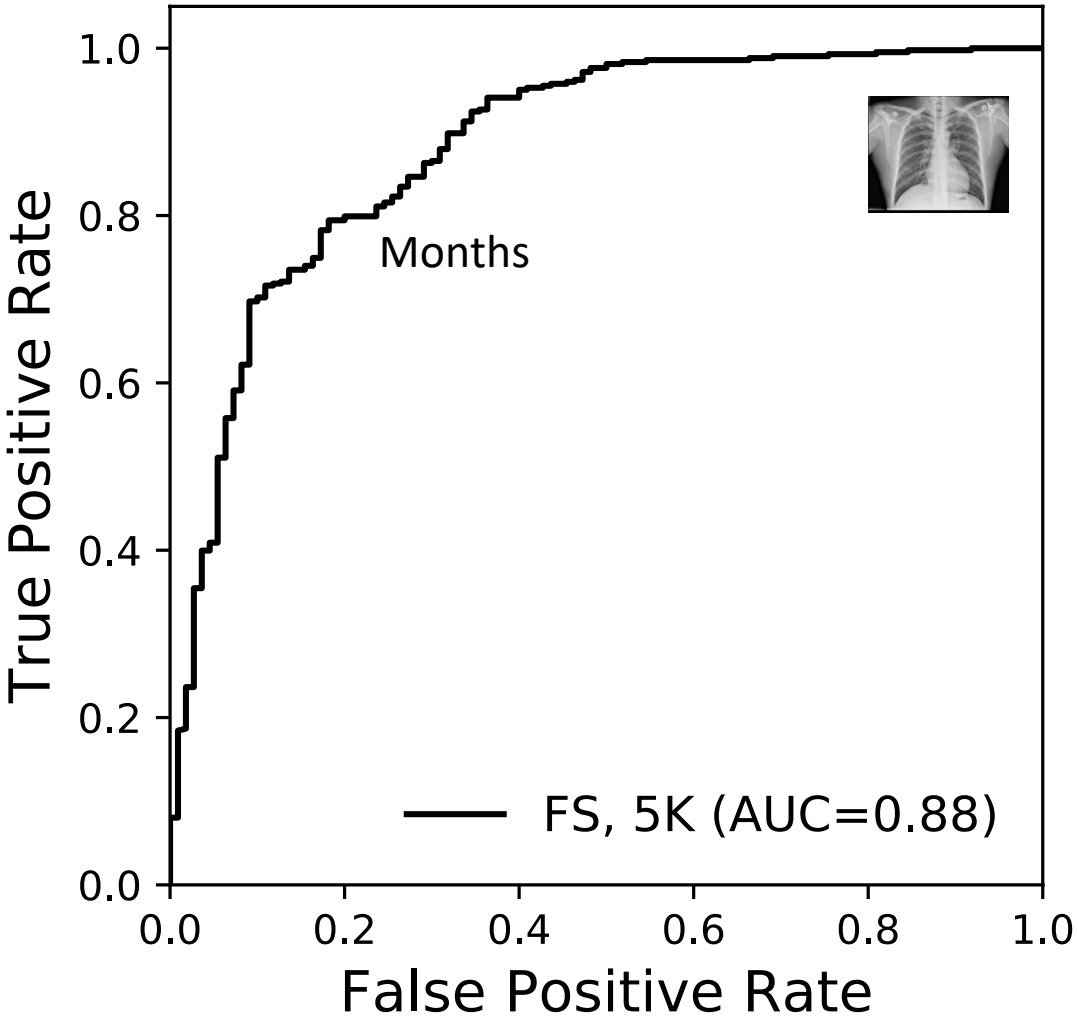
Let’s go back to radiology…

Applying Weak Supervision Across Modalities

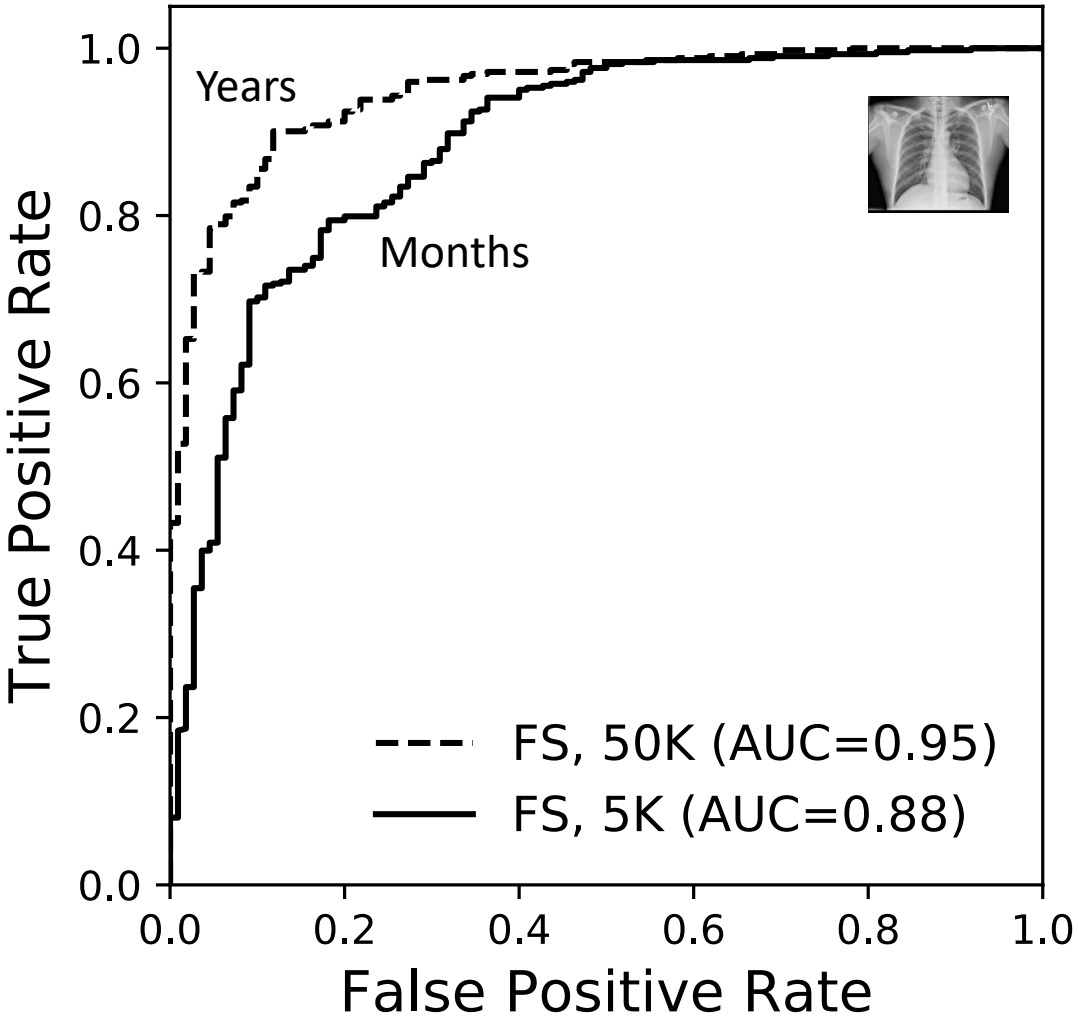


We can leverage data programming across modalities to make weak supervision of complex tasks easier!

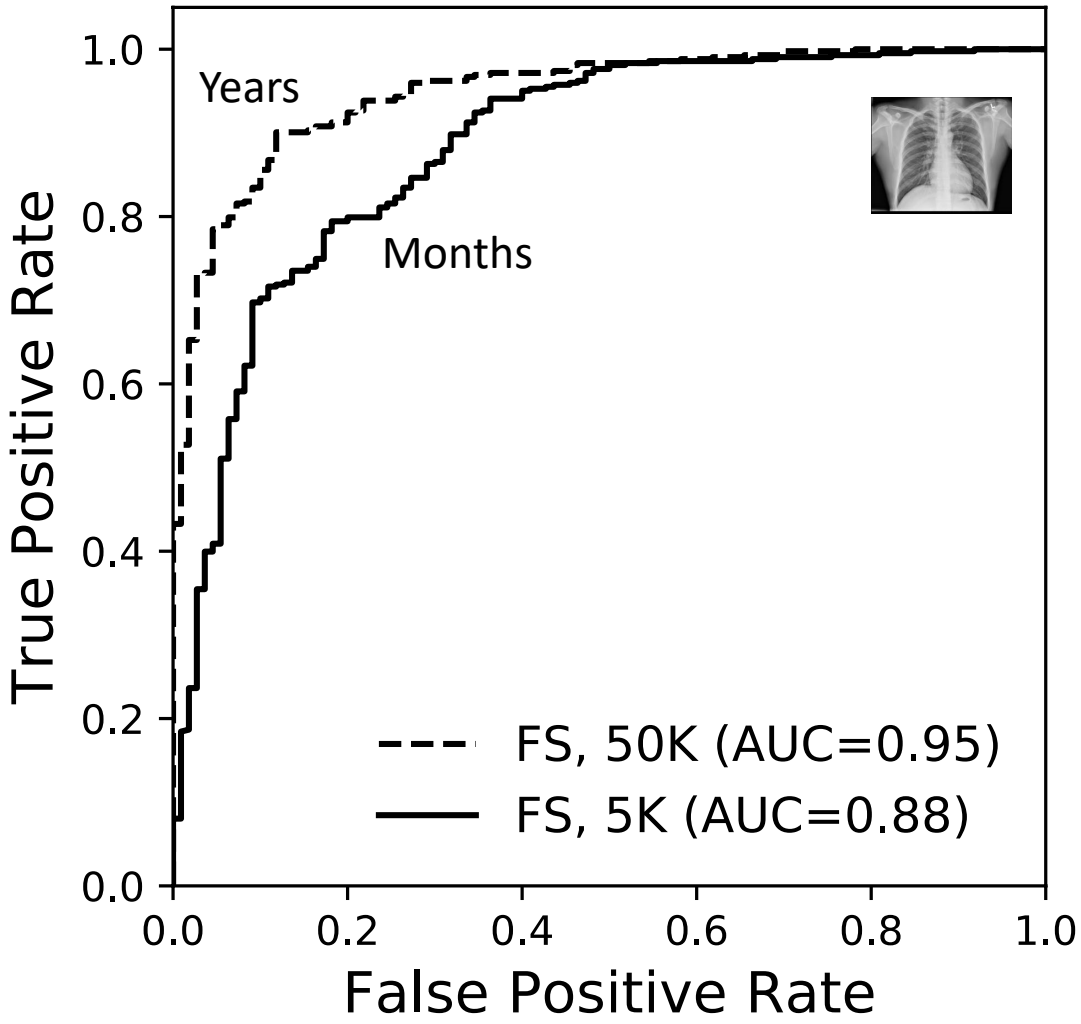
Cross-Modal Chest X-ray Classification



Cross-Modal Chest X-ray Classification



Cross-Modal Chest X-ray Classification

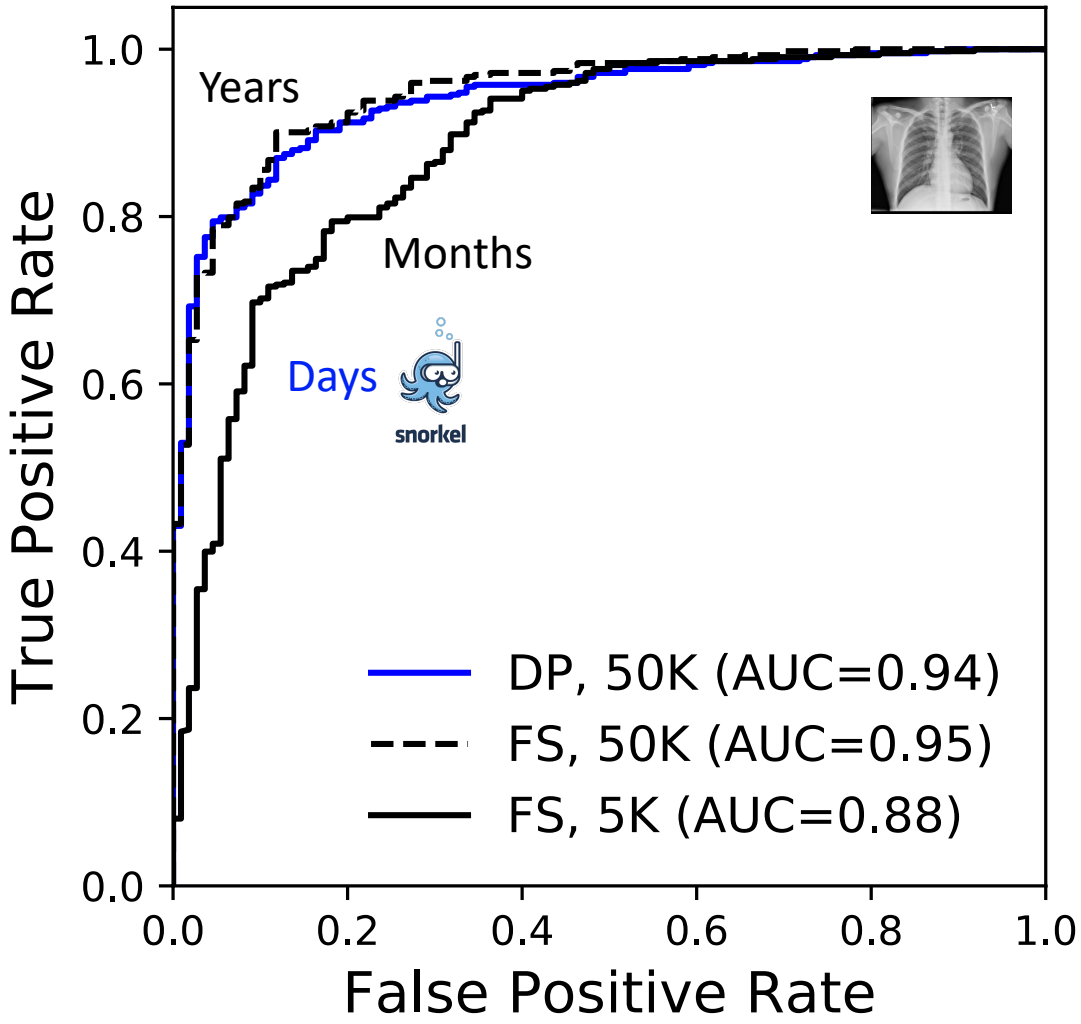



```
def LF_pneumothorax(c):
  if re.search(r'pneumo.*', c.report.text):
      return "ABNORMAL"

def LF_pleural_effusion(c):
  if "pleural effusion" in c.report.text:
      return "ABNORMAL"

def LF_normal_report(c, thresh=2):
    if len(NORMAL_TERMS.intersection(c.
     report.words)) > thresh:
      return "NORMAL"
```

Indication: Chest pain. Findings: Mediastinal contours are within normal limits. Heart size is within normal limits. No focal consolidation, pneumothorax or pleural effusion. Impression: No acute cardiopulmonary abnormality.

20 Labeling Functions
in 8 hours

Cross-Modal Chest X-ray Classification




```
def LF_pneumothorax(c):
  if re.search(r'pneumo.*', c.report.text):
      return "ABNORMAL"

def LF_pleural_effusion(c):
  if "pleural effusion" in c.report.text:
      return "ABNORMAL"

def LF_normal_report(c, thresh=2):
    if len(NORMAL_TERMS.intersection(c.
     report.words)) > thresh:
      return "NORMAL"
```

Indication: Chest pain. Findings: Mediastinal contours are within normal limits. Heart size is within normal limits. No focal consolidation, pneumothorax or pleural effusion. Impression: No acute cardiopulmonary abnormality.

20 Labeling Functions
in 8 hours

Related Work in Weak Supervision

- ***Distant Supervision:*** Mintz et. al. 2009, Alfonesca et. al. 2012, Takamatsu et. al. 2012, Roth & Klakow 2013, Augenstein et. al. 2015, etc.
- ***Crowdsourcing:*** Dawid & Skene 1979, Karger et. al. 2011, Dalvi et. al. 2013, Ruvolo et. al. 2013, Zhang et. al. 2014, Berend & Kontorovich 2014, 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.
- **Propensity SVMs:** Joachims 17

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

What if we don't have the dependency structure?

Paroma Varma

Fred Sala

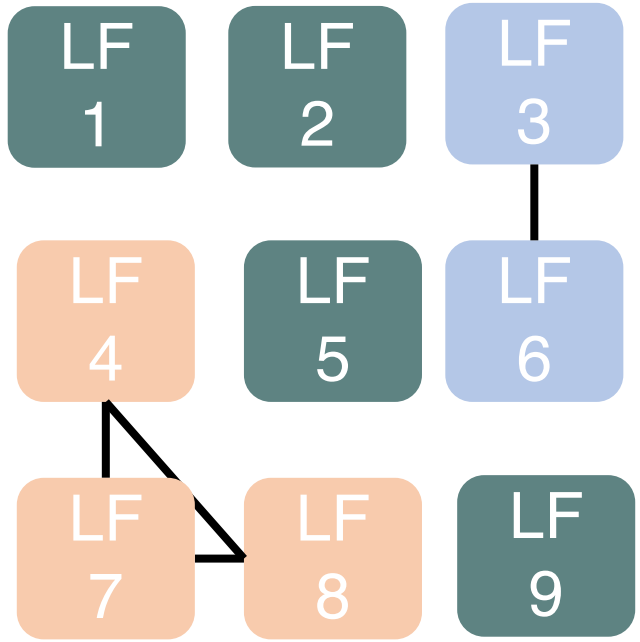
P. Varma, F. Sala, A. Ratner, C. Ré, "Learning Dependency Structures for Weak Supervision Models", ICML 2019.

Ignore the dependencies?

Edge-based Primitives
Intensity-based Primitives
Morphology-based Primitives

...

Input Primitives for Labeling Functions



Rely on morphology-based features

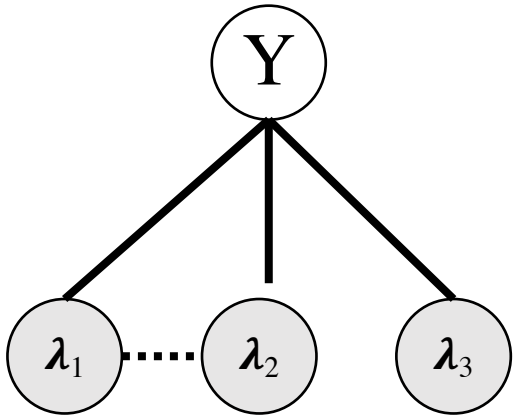
Rely on edge-based features

Labeling Functions Operate over Similar Primitives

Ignoring dependencies hurts end model by up to 4.61 F1 points!

Learn the dependencies?

Many structure learning techniques, but **key difference:** Label variable is **latent.**



New setting for dependency structure learning

Sample Complexity

m is # of LFs, d is largest degree for a dependency. A sample is an m x m matrix.

- **Prior work**: samples to recover WS dependency structure w. h. p.

S. Bach, B. He, A. Ratner, C. Ré, "Learning the structure of generative models without labeled data", ICML 2017.

$O(m \log m)$ Doesn't exploit ***d:*** sparsity of graph

- Recent application of RPCA for general latent-variable structure learning

C. Wu, H. Zhao, H. Fang, M. Deng, "Graphical model selection with latent variables", EJS 2017.

$O(d^2 m)$ is Linear in ***m***.

$\Omega(d^2 \log m)$ is the optimal possible sample complexity---*even in the supervised case* [Santhanam & Wainwright '10]

Robust PCA

Back to our covariance matrix—assume it's *somewhat sparse.*

$$\underbrace{\Sigma_O^{-1}}_{\text{Observed}} = \underbrace{(\Sigma^{-1})_O}_{\text{Sparse}} - \underbrace{zz^T}_{\text{Low-Rank}}$$

- Idea: decompose LHS into **sparse** and **low-rank** components; **sparse** part contains graph structure

- **Robust PCA** [Candès et al. 2010, Chandrasekaran et al. 2010].

Our Approach: Sample Complexity

m *is # of LFs,* ***d*** *is largest degree for a dependency*

Ours: for $\tau < 1$, an eigenvalue decay factor in blocks of LFs

$$O(d^2 m^{\tau})$$

Ours: When there is a **dominant block or independent** of correlated LFs

$$O(d^2 \log m)$$

Key Tool: exploit sharp concentration inequalities on sample covariance matrix Σ_O again via *effective rank* [Vershynin '12].

Comparison to Supervised Case.

m is # of LFs, d is largest degree for a dependency

- For some graphs (w/o singleton separators), improve the supervised case (which has cubic dependence on d) [Santhanam & Wainwright '10]

- We can also identify an **extra sample factor** for the weak supervision setting. Asymptotically,

$$\frac{n_{WS}}{n_S} \leq 2$$

- Need (at most) twice as many samples!

Does it help the end model?

Application	m	(s, d)	MV	Indep.	Bach et al.	Ours	Improvement Over Indep.	Improvement Over Bach et al.
Bone Tumor	17	(2,3)	65.72	67.32	67.83	71.96	+4.64	+4.13
CDR	33	(22,14)	47.74	54.60	55.90	56.81	+2.21	+0.91
IMDb	5	(1,4)	55.21	58.80	60.23	62.71	+3.91	+2.48
MS-COCO	3	(1,2)	57.95	59.47	59.47	63.88	+4.41	+4.41

 # LFs

(# cliques, max.degree)

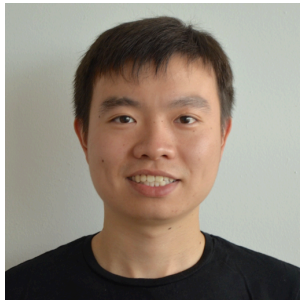
Yes! Improvement over

- Not modeling dependencies by up to 4.64 F1 points
- Previous approach by up to 4.41 F1 points

How many types of data?

HoloClean

Theo Retkatsinas



Xu Chu

Ihab Ilyas

SW 2.0 for **structured data.**

~90% precision & ~ 76% recall on real data sets—2x higher F1 score than SotA

Theodoros Rekatsinas, Xu Chu, Ihab F. Ilyas, C.Ré **VLDB17**

Tutorial - Data Integration and Machine Learning: A Natural Synergy

http://www.dataintegration.ml/

SIGMOD 2018

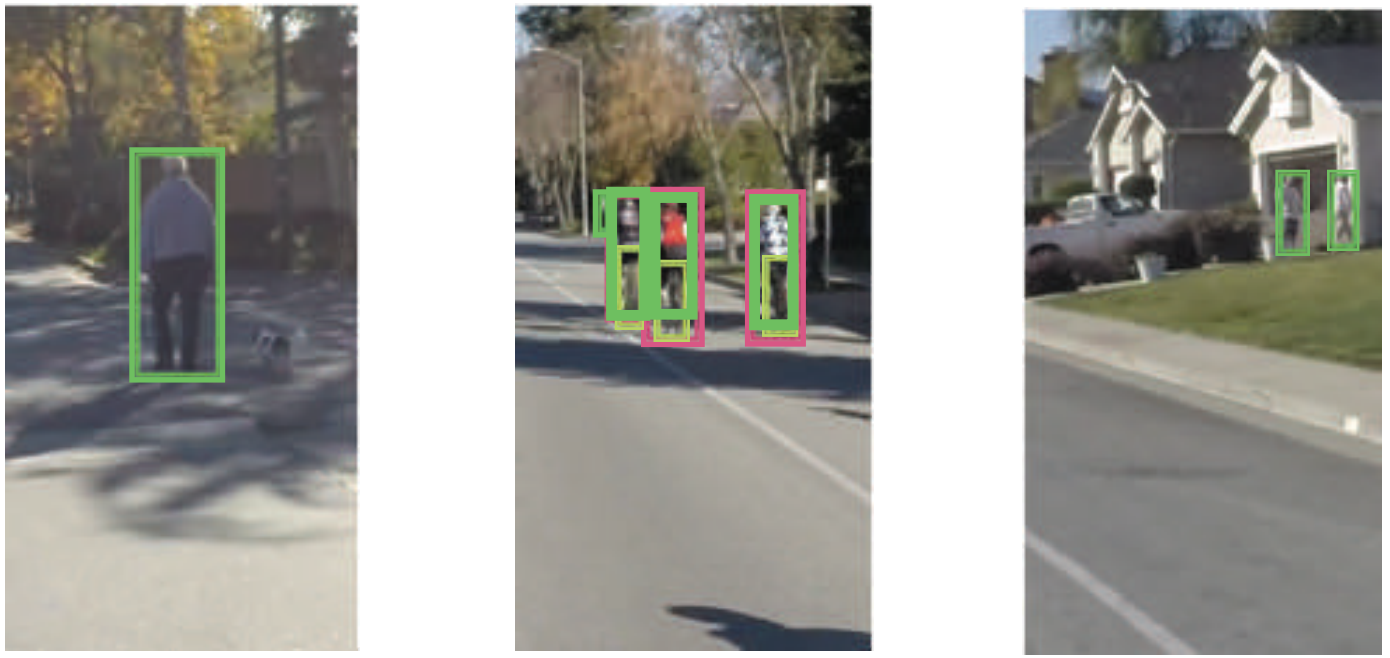
VLDB2018

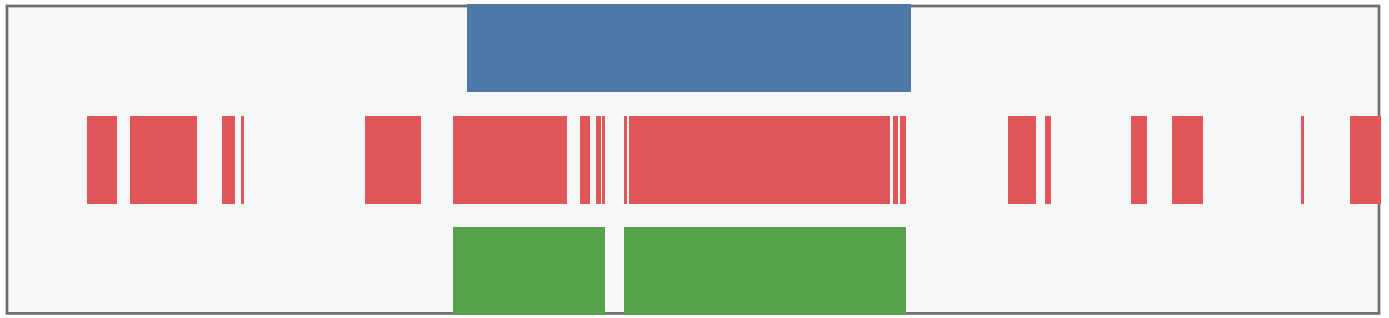
KDD 2019

Results: Detecting Cyclists in Videos

Supervision sources use:
Object detection output for `person` and `bike`

Distribution Prior states:
Cyclists likely to appear in consecutive frames





Avoid false positives using prior – improve by 37.5 F1 points

Utilizing Weak Supervision to Infer Complex Objects and Situations in Autonomous Driving Data. Z Weng, et al. 2019 IEEE Intelligent Vehicles

Tracking Medical Device Safety is an International Problem

1.7 Million Injuries
83,000 Deaths
Since 2010 in the U.S.

2018 global investigation on the medical device industry

We use Snorkel to automatically identify poor patient outcomes from medical record data

machine reading + patient notes

13 - 54% F1 improvement
over current NLP approaches

6x more complications found vs. billing codes

On the job market!

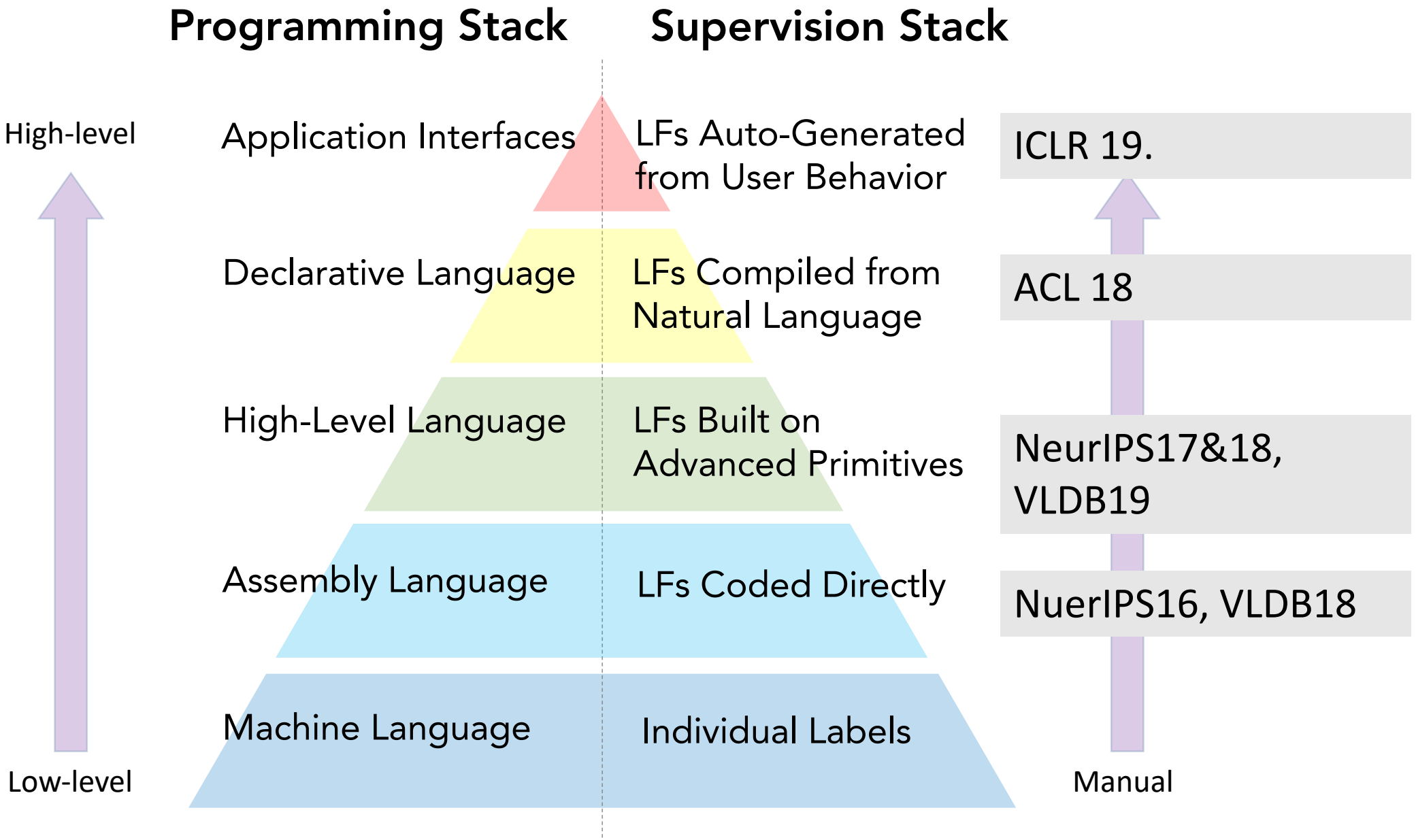
Jason Fries, PhD

npj nature partner journals

Medical Device Surveillance with Electronic Health Records
Callahan & Fries et al. *npj Digital Medicine*. 2019.

Some of our Future Directions?

More Problems, More Modalities, More Supervision


Programming Stack
Supervision Stack
High-level
Low-level
Application Interfaces
Declarative Language
High-Level Language
Assembly Language
Machine Language
LFs Auto-Generated from User Behavior
LFs Compiled from Natural Language
LFs Built on Advanced Primitives
LFs Coded Directly
Individual Labels
ICLR 19.
ACL 18
NeurIPS17&18, VLDB19
NuerIPS16, VLDB18
Manual

HOME PEOPLE

Software 2.0 and Data Programming: Lessons Learned, and What's Next

Dan Fu, Laurel Orr, and students of HazyResearch

Posted on February 28, 2020

Laurel Orr

Dan Fu

The only view that matters: student & postdoc view

http://Hazyresearch.Stanford.edu

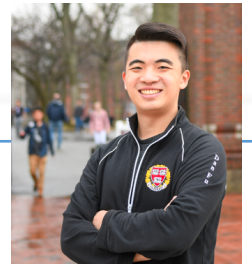
Towards Interactive Weak Supervision with FlyingSquid

Dan Fu, Mayee Chen, Fred Sala, Sarah Hooper, Kayvon Fatahalian, and Chris Ré

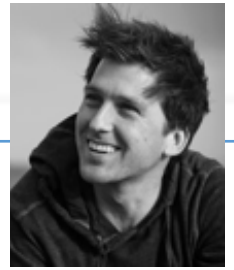
Posted on February 28, 2020

Weak supervision has become a popular technique for automatically generating labeled data for machine learning models from multiple noisy label sources and is in use in applications used by billions of people every day like Gmail AI products at Apple and search products at Google. But existing weak supervision frameworks...
[Read More]

Names in order above.







Hidden Stratification Causes Clinically Meaningful Failures in Machine Learning for Medical Imaging

New Abstractions, New Problems

Gustavo Carneiro (Adelaide)

Luke Oakden-Rayner (Adelaide)

Jared Dunnmon (Stanford)

These eyes haunt me…

Any model may pick out **unintended signal**.
Deep models may pick out ***more*** unintended signal.

Upshot:
Picked up on
mascara

Kuehlkamp et al. *Gender-from-Iris or Gender from-Mascara*

Do we know how well these models are really performing?

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

Are we sure those differences are causal? Anticausal?

t!

- 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.

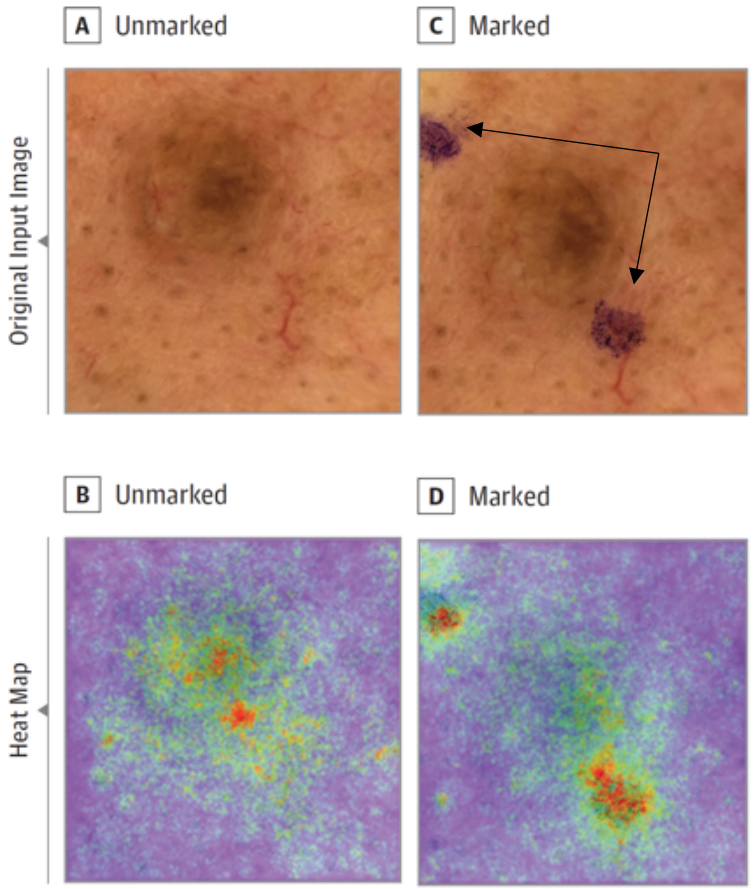
Image Credits

Valchanov, Kamen, Nicola Jones, and Charles W. Hogue, eds. *Core Topics in Cardiothoracic Critical Care*. Cambridge University Press, 2018.

Winkler, Julia K., et al. "Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition." *JAMA dermatology* (2019).

It's not just those eyes…

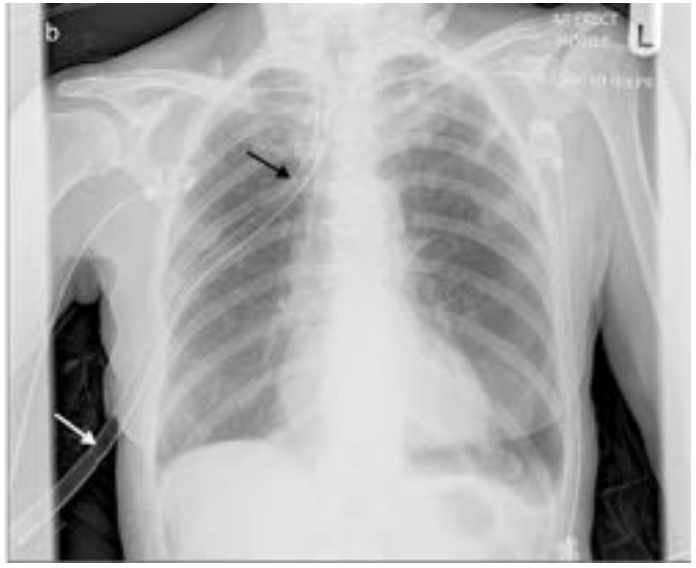
Melanoma Recognition (Surgical Marks)



Pneumonia Detection

No Drain

With Drain



Pneuomthorax detection 0.87 AUC, which is superhuman

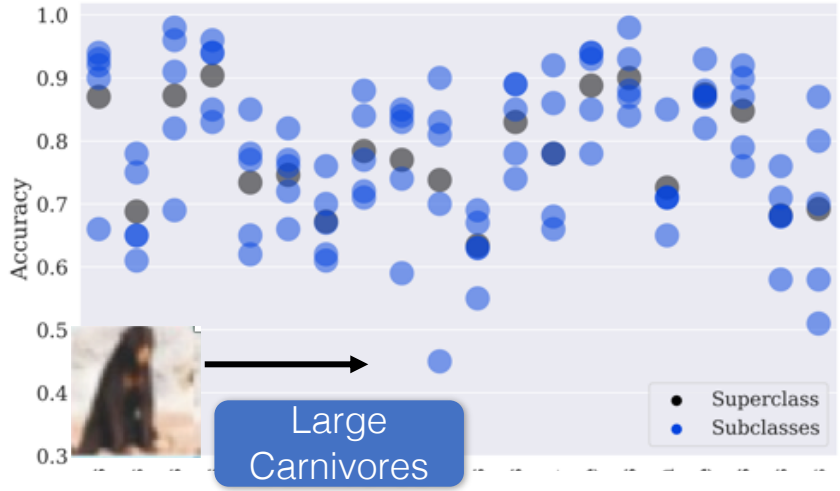
… with chest drains—***Chest drain means already treated!*** *Down to* 0.77 when removed…

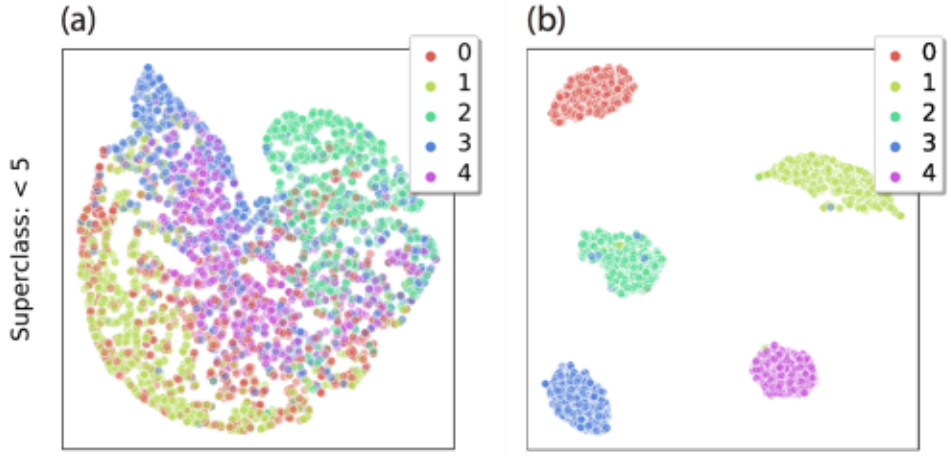
One issue: Hidden Stratification.

- Classical: Never write features that say
  - If drain then pneumonia
  - if purple dot then cancer
  - **But new SW abstraction, new bugs**

- Accidental—not adversarial—attacks. A subset of a class (stratum) performs worse.
  - E.g., Abnormal consists of **many** unlabeled subclasses or strata.

Develop a theory & techniques to handle hidden stratification?





Sometimes separated in representation!

Conclusion

- **Snorkel**: A framework for rapidly creating training sets for multi-task models used in a wide array of places.
- Nugget: Latent variable formulation w/ connection to **statistical estimation**, **structure learning.**
- The change to programming by supervision changes what systems you build and how you build them.

snorkel

Snorkel.org

User Study

Snorkel User Study

How easily can **non-machine learning experts** use Snorkel?

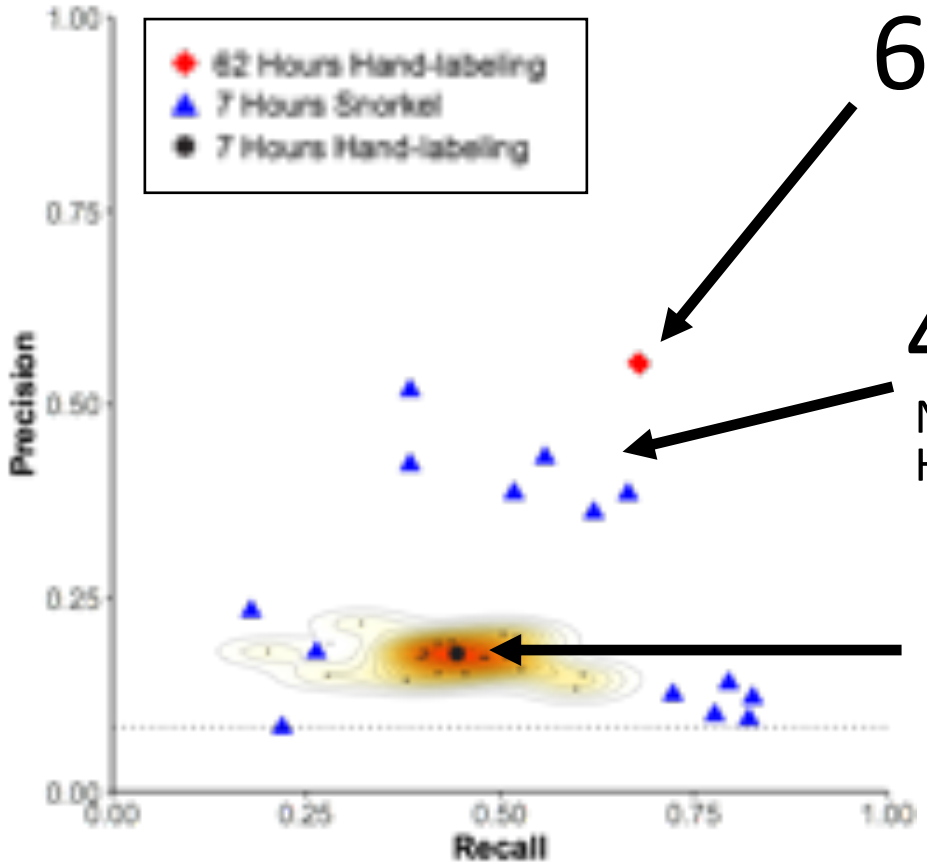
7 hours of human labeling
Amazon Mechanical Turk

VS.

14 New Snorkel Users
during a **7 hour** workshop

mobilize
Center for Mobility Data Integration to Insight
NIH NLM

Jason Fries, PhD



62 Hours
Hand-labeling
22,195 labels

43%
New Snorkel Users Surpassed
Hand-labeling

7 Hours
Hand-labeling
2500 randomly
sampled labels

3rd Place Score
No machine learning experience
Beginner-level Python

Median Crowdsourced Model
F1= 24.6

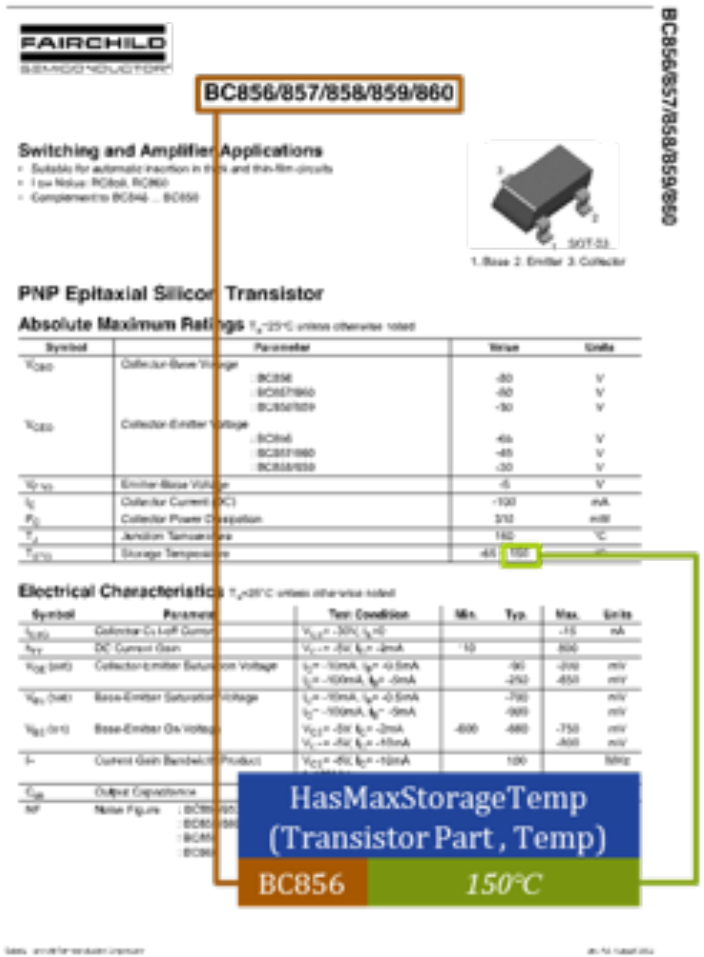
VS.

Best **Snorkel** Mod
F1= 48.7

Structured data

Fonduer: Handling Richly-Formatted Data

SIGMOD 2018



Challenges:
(1) Document-level relations
(2) Multimodal information
(3) Data variety

Doc. level Candidates		Multimodal Supervision: Horizontal Align with '°C'	Row Ngrams Gontain 'Junction'	Temp Value in Table
BC856	*160*	✓	✗	✓
BC856	*-65*	✓	Ø	✗
BC856	*150*	✓	Ø	✓

Data programming with labeling functions written over richly formatted data in unified data model

	Prec.	Rec.
MEMEX	87%	89%
IoT	73%	81%
GWAS	89%	81%
Paleo	72%	38%

DARPA MEMEX

SING Stanford Information Networks Group

Macrostrat Lab

Stanford MEDICINE

Alibaba Group

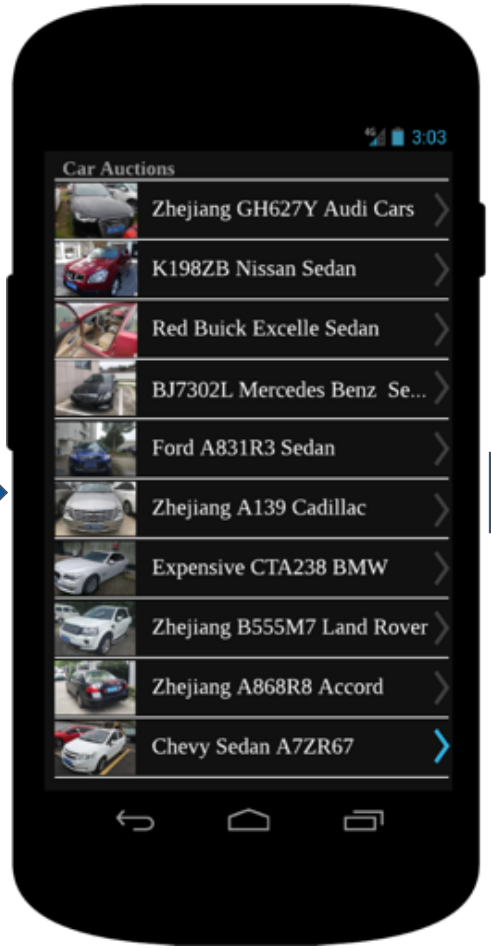
Input: User-customized HTML auction pages → **Output:** Structured knowledge base

Extract key facts (make, model, license etc.)

Fonduer

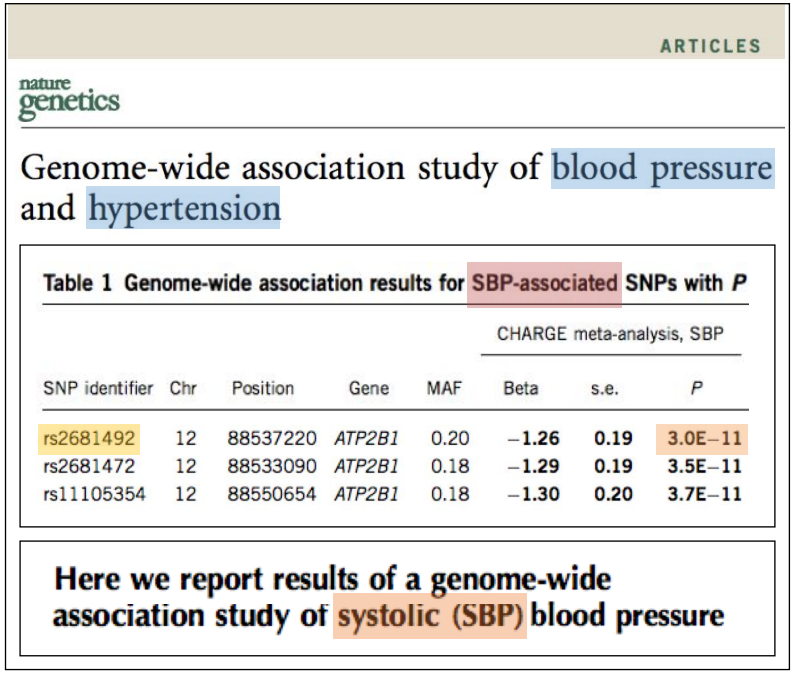
Alibaba.com

Improve auction searching quality and UX



A Machine Compiled Database of Genome Wide Association Studies

Biomedical Publication



ARTICLES

nature genetics

Genome-wide association study of blood pressure and hypertension

Table 1 Genome-wide association results for SBP-associated SNPs with *P*

					CHARGE meta-analysis, SBP		
SNP identifier	Chr	Position	Gene	MAF	Beta	s.e.	*P*
rs2681492	12	88537220	*ATP2B1*	0.20	−1.26	0.19	3.0E−11
rs2681472	12	88533090	*ATP2B1*	0.18	−1.29	0.19	3.5E−11
rs11105354	12	88550654	*ATP2B1*	0.18	−1.30	0.20	3.7E−11

Here we report results of a genome-wide association study of systolic (SBP) blood pressure

Machine Reading

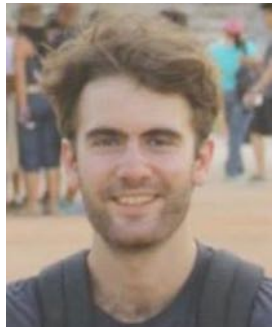
Structured Database

Variant	rs2681492
Simple phenotype	Hypertension | Blood pressure
Detailed phenotype	Systolic
p-value	3.0e-11
Source	PMID: 19430479, Tbl. 1

Database	Statistics over open-access papers	
	Associations	Unique Associations
GWAS Catalog	8,384	**2,026**
GWAS Central	5,914	**364**
GwasKB (ours)	6,231	**2,777**

Existing databases are incomplete

GwasKB finds 2,700 new associations



Volodymyr Kuleshov

HoloClean: Weakly-supervised Data Cleaning

Goal: Detect and repair errors in structured data

Diverse errors:

(i) Typos and formatting

(ii) Conflicting values

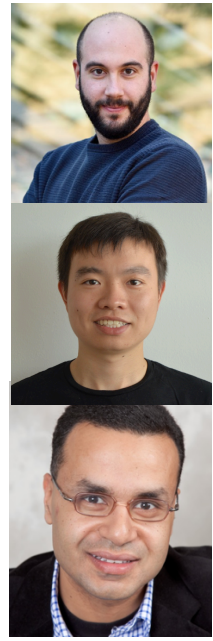
(iii) Outlier values

	DBAName	AKAName	Address	City	State	Zip
t1	John Veliotis Sr.	Johnnyo's	3465 S Morgan ST	***Chicago***	IL	***60608***
t2	John Veliotis Sr.	Johnnyo's	3465 S Morgan ST	Chicago	IL	***60609***
t3	John Veliotis Sr.	Johnnyo's	3465 S Morgan ST	Chicago	IL	***60609***
t4	***Johnnyo's***	Johnnyo's	3465 S Morgan ST	***Cicago***	IL	60608

Conflicts

Does not obey data distribution

Conflict



Users provide *high-level qualitative constraints* and external data. **No other supervision required!**

HoloClean has ~ 90% precision & ~ 76% recall on real data sets—2x higher F1 score than SotA

Theodoros Rekatsinas, Xu Chu, Ihab F. Ilyas C.Ré **VLDB17**

Paroma Varma

Jason Fries

James Priest

Fred Sala

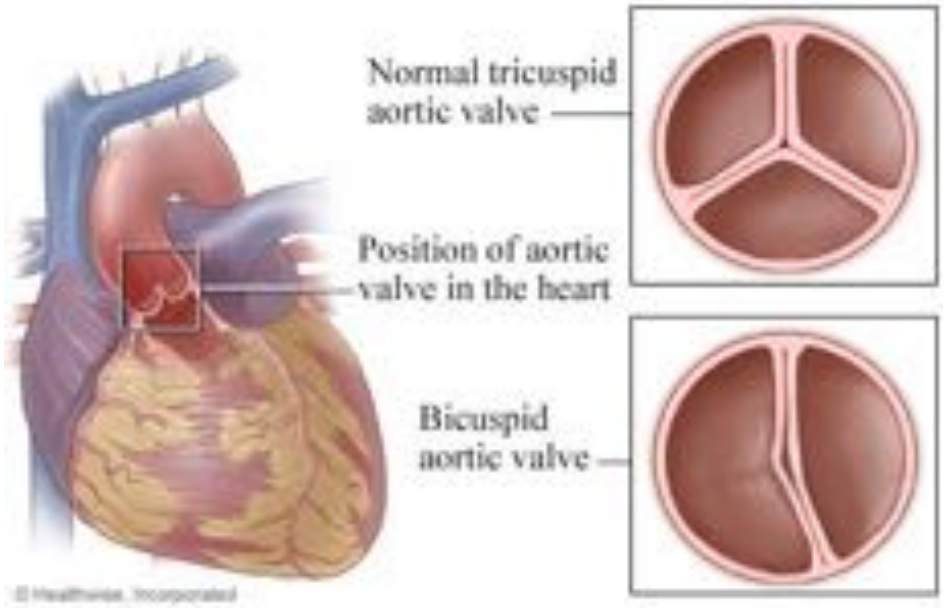
Time Series & Video

J. Fries, “Weakly supervised classification of rare aortic valve malformations using unlabeled cardiac MRI sequences”, Nature Communications, 2019.

Paroma Varma, Fred Sala et al. “Multi-Resolution Weak Supervision for Sequential Data.”, NeurIPS 2019.

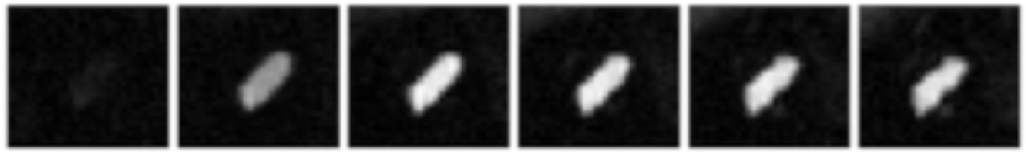
Classifying Heart Valve in MRI Video

- Bicuspid aortic valve (BAV) is a congenital malformation, incidence 0.5-2% -- look at lots of data!
- Can lead to cardiovascular issues and may require surgical valve replacement

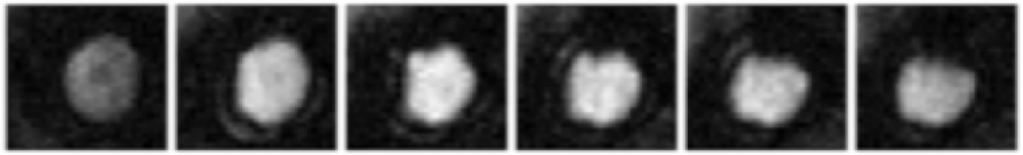


Source: www.umcvc.org

BAV Frames

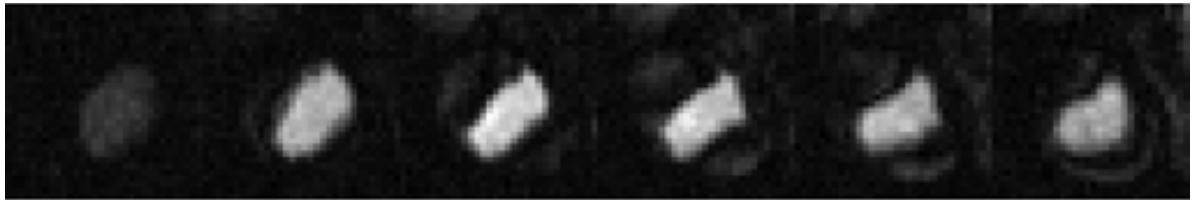


TAV Frames



There is a lack of labeled datasets targeting BAV subjects

Results: Bicuspid Aortic Valve Detection



Distribution Prior states:
If heart valve is abnormal, it should look abnormal across all frames

Supervision sources use:
Valve shape and *shape change* information

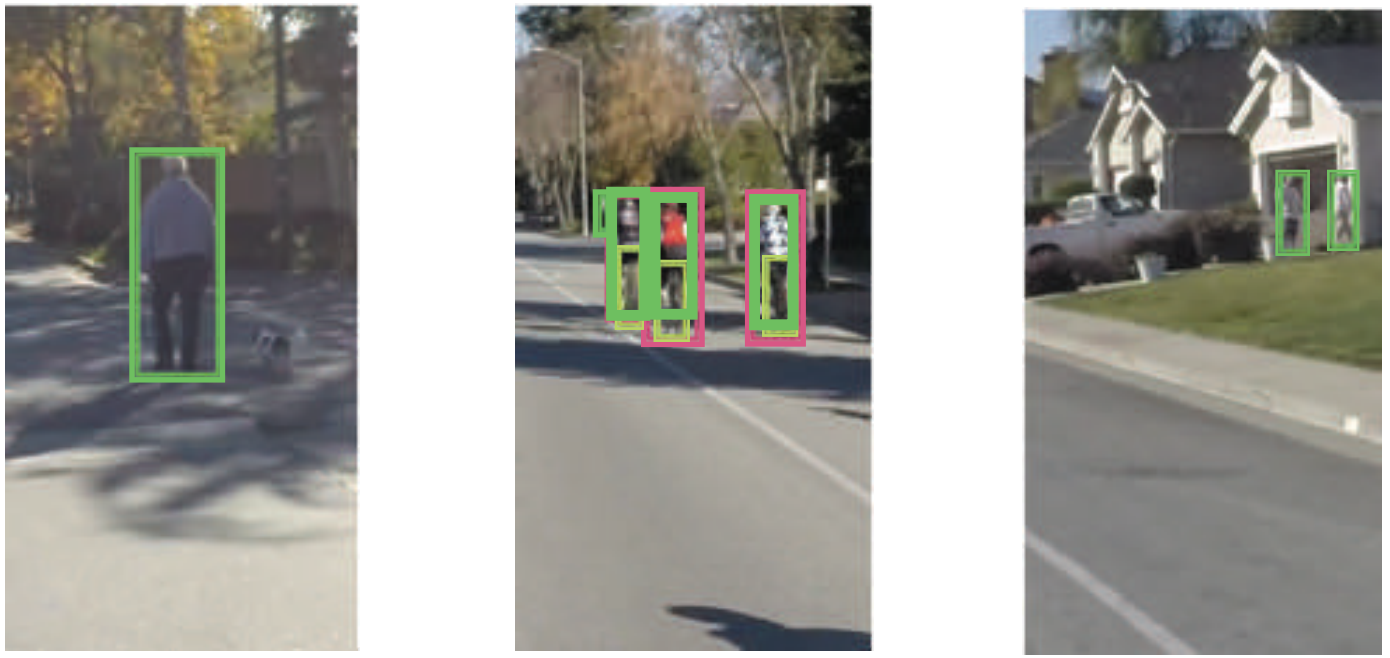
Model	Train Size	Precision	Recall	F1
Baseline (Hand-label)	106	26.1 ± 3.8	20.0 ± 7.0	22.1 ± 5.1
Traditional Weak Supervision (Nature Comms)	4239	70.0 ± 19.8	**45.7 ± 5.7**	53.2 ± 4.4
Multi-Resolution Weak Supervision	4239	**95.0 ± 10.0**	42.9 ± 0.0	**58.9 ± 2.2**

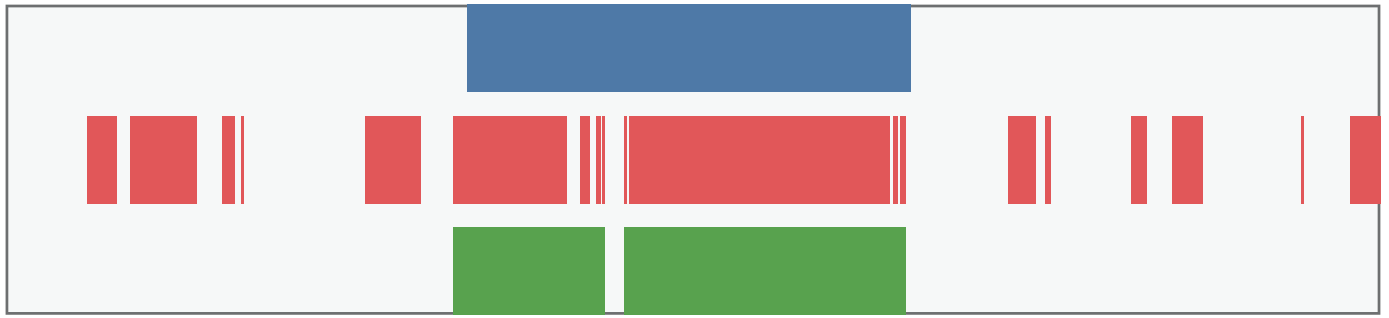
< 2% prevalence of positive cases – combine information across frames to improve by 25 points precision

Results: Detecting Cyclists in Videos

Supervision sources use:
Object detection output for `person` and `bike`

Distribution Prior states:
Cyclists likely to appear in consecutive frames

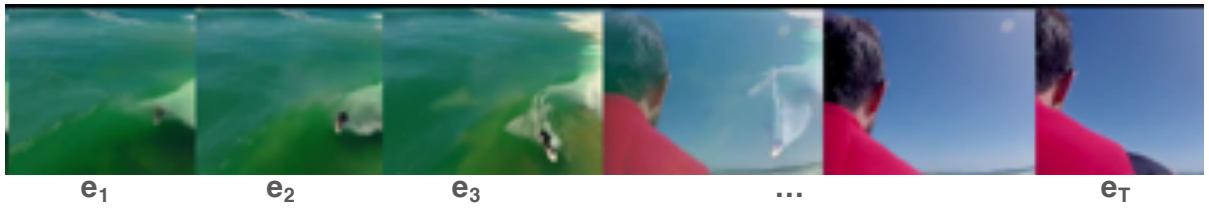




Avoid false positives using prior – improve by 37.5 F1 points

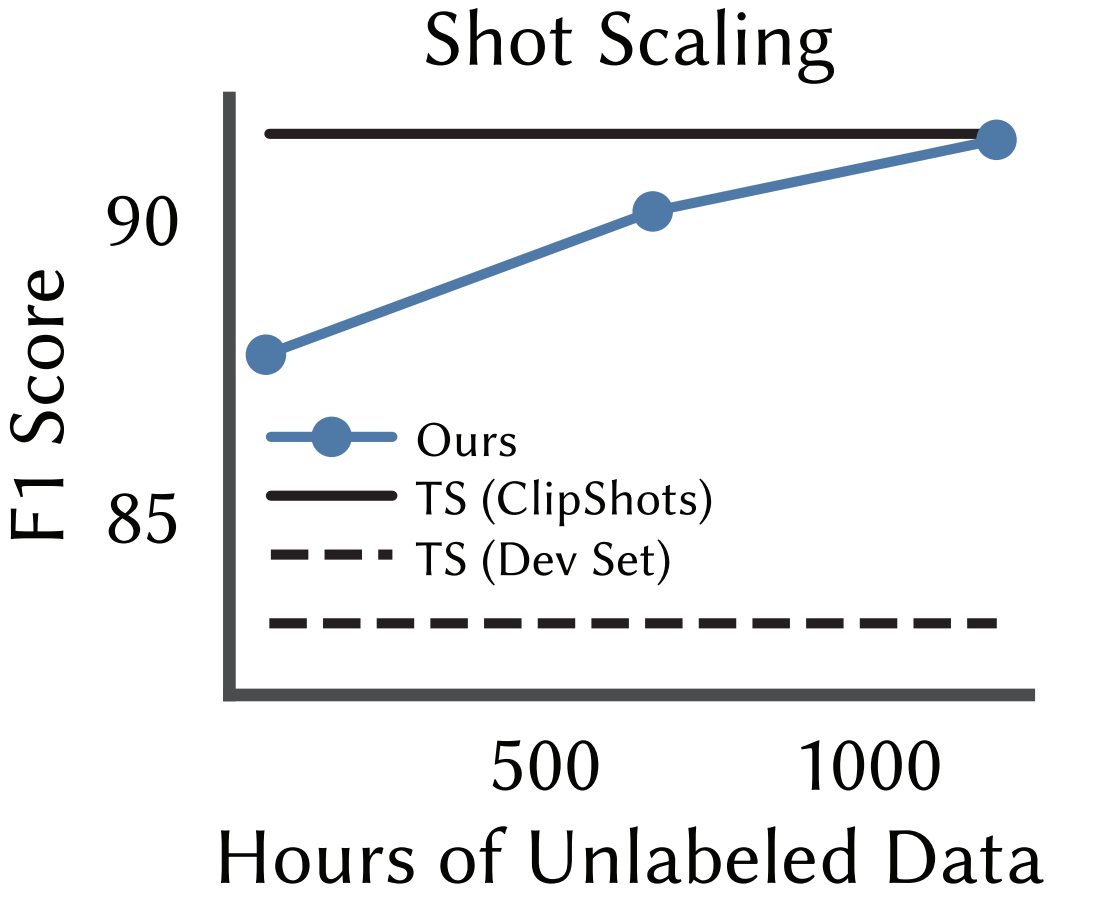
Utilizing Weak Supervision to Infer Complex Objects and Situations in Autonomous Driving Data. Z Weng, et al. 2019 IEEE Intelligent Vehicles

Results: Scene Change (Shot) Detection in Videos



Distribution Prior states:
Scene changes occur infrequently

Supervision sources use:
Frame and scene level color information



Match oracle model performance using 686x fewer ground truth labels