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.

Classical problems ML 1st
• ETL & Cleaning (Holoclean.io)
• DB Tuning Peloton (CMU)
• Networks Pensieve (MIT).
• NeuroCore (Stanford)

“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.
Automated Chest X-ray Triage

Optimizing Workflows with Automated Prioritization, Radiology January 19

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

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

<table>
<thead>
<tr>
<th>Model</th>
<th>Test Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>BOVW + KSVM</td>
<td>0.88</td>
</tr>
<tr>
<td>AlexNet</td>
<td>0.87</td>
</tr>
<tr>
<td>ResNet-18</td>
<td>0.89</td>
</tr>
<tr>
<td>DenseNet-121</td>
<td>0.91</td>
</tr>
</tbody>
</table>

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

AutoAugment: Using learned data augmentation policies
  - Augmentation Policies first in Ratner et al. NIPS ’17

Facebook Hash tag weakly supervised pre-training
  - Pre-train using a massive dataset with hashtags
Automating the Art of Data Augmentation
Part I Overview

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

Static

Manual Labels

$10 - $100/hr

Programmatic Labels

Fast

Cheap

Dynamic

Programmatic Labels

write programs

run programs

Time

Labels

$0.10/hr

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

Observation: Weak supervision applied in *ad hoc* and isolated ways.
Goal: Replace *ad hoc* weak supervision with a formal, unified, theoretically grounded approach for programmatic labeling

Snorkel: Formalizing Programmatic Labeling
The Real Work

Stephen Bach  Braden Hancock  Henry Ehrenberg  Alex Ratner  Paroma Varma

Snorkel.org
Running Example: NER

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

Goal: Label training data using *weak supervision* strategies for these three tasks
Problem: These noisy sources conflict and are correlated.
The Snorkel Pipeline

Users write labeling functions to generate noisy labels

Snorkel models and combines the noisy labels into probabilities

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

Use training data as a medium for knowledge transfer

Available at test time
This is servable!

Not available at test time
Not servable

Report 47:
Indication: Chest pain.
Findings: Pneumothorax.
Operation recommended.

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

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

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

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

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

- **Slow**: $10 - $100/hr
- **Expensive**: {Positive, Negative}
- **Static**: {Positive, Neutral, Negative}

Programmatic Labels

- **Fast**: $0.10/hr
- **Cheap**: aws

Dynamic Labels
Snorkel: In use at the world’s largest companies

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)
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 Srisuwawanukorn
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
It has changed use real systems...

<table>
<thead>
<tr>
<th>Resourcing</th>
<th>Error Reduction</th>
<th>Amount of Weak Supervision</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>65% (2.9×)</td>
<td>80%</td>
</tr>
<tr>
<td>Medium</td>
<td>82% (5.6×)</td>
<td>96%</td>
</tr>
<tr>
<td>Medium</td>
<td>72% (3.6×)</td>
<td>98%</td>
</tr>
<tr>
<td>Low</td>
<td>40% (1.7×)</td>
<td>99%</td>
</tr>
</tbody>
</table>

A couple of highlights

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

CIDR2020
High-Level Related Work

LUDWIG

Snorkel

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

JAX

TensorFlow

PyTorch

Core ML
Cross-Modal Weak Supervision

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

A. Clinical Note + Markup

B. Labeling Function Definitions

Biomedical publication

Structured database

J. Dunnmon et al., Radiology, 2019


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

V. Kuleshov et al., Nat Comms, 2019

Imaging & Diagnostics

J. Fries et al., Nat Comms, 2019

J. Dunnmon et al., Nat Comms, 2019

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

Text & Extraction

Weak Supervision in Science & Medicine
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!**


Users write labeling functions for multiple related tasks.

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

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

How to represent diverse sources of weak supervision?
Users write labeling functions for multiple related tasks

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

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

How can we do anything without the ground truth labels?
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

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

Intuition: Learn from the Overlaps

Key idea: We can observe overlapping judgements on many points!
Solution Sketch: Using the covariance

Can only observe portion of the covariance ($\Sigma_0$)... if we observe rest, we’d be done! ($E[y\lambda] \sim$ accuracy).
Idea: Use graph-sparsity of the inverse

\[ \Sigma_0 \]

\[ (\Sigma^{-1})_0 \]

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

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

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

Incompletely Observed

matrix inversion lemma

Observed overlaps
Result: A matrix completion problem?

We get a set of equations. For any pair \( i \neq j \) with no edge in graph—the lhs is 0.

\[
0 = (\Sigma^{-1}_O)_{i,j} + Z_iZ_j
\]

\( \Sigma \) is full rank, so not really matrix completion…

Key: \( \Sigma = I + uu^\top \) for some \( u \) so intuitively close…
Couple of Technical Comments

\[ 0 = (\Sigma_0^{-1})_{i,j} + Z_i Z_j \]

- 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 \( \text{er}(\Sigma) = \text{tr}(\Sigma)/|\Sigma|_2 \) (effectively, use this!)
  - small when single large: \( |z|_2 \) is large.
  - Scale inversely distance to noise (\( z_i = 0 \)).
Users write labeling functions for multiple related tasks.

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

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

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

- 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

![Graph showing ROC curves for different datasets.](image)

- **FS, 50K (AUC=0.95)**
- **FS, 5K (AUC=0.88)**

*Note: ROC curves illustrate the performance of the classification models.*
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.
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.
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
What if we don’t have the dependency structure?

Paroma Varma

Fred Sala

Ignore the dependencies?

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

Labeling Functions Operate over Similar Primitives

Input Primitives for Labeling Functions

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

...
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 LF{s}, $d$ is largest degree for a dependency. A sample is an $m \times m$ matrix.

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

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

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


  $O(d^2m)$ 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*.

\[
\Sigma^{-1}_0 = (\Sigma^{-1}_0)_{\text{Sparse}} - ZZ^T_{\text{Low-Rank}}
\]

- 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 \( \Sigma_o \) 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?

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

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/
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

Ground Truth Labels

Traditional Weak Supervision

Multi-Resolution Weak Supervision

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.

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

13 - 54% F1 improvement over current NLP approaches

6x more complications found vs. billing codes

2018 global investigation on the medical device industry

www.icij.org

IMPLANT FILES

2018 global investigation on the medical device industry

machine reading + patient notes

Medical Device Surveillance with Electronic Health Records

On the job market!

Jason Fries, PhD
Some of our Future Directions?

More Problems, More Modalities, More Supervision
Programming Stack

- Application Interfaces
- Declarative Language
- High-Level Language
- Assembly Language
- Machine Language

Supervision Stack

- 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
- NeurIPS16, VLDB18
- Manual
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

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.
New Abstractions, New Problems

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

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

Are we sure those differences are causal? Anticausal?

<table>
<thead>
<tr>
<th>Model</th>
<th>Test Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>BOVW + KSVM</td>
<td>0.88</td>
</tr>
<tr>
<td>AlexNet</td>
<td>0.87</td>
</tr>
<tr>
<td>ResNet-18</td>
<td>0.89</td>
</tr>
<tr>
<td>DenseNet-121</td>
<td><strong>0.91</strong></td>
</tr>
</tbody>
</table>

*Often:* Differences in models ~ 2-3 points.

Later: Label quality & quantity > model choice.
It’s not just those eyes...

Melanoma Recognition (Surgical Marks)

Pneumonia Detection

No Drain

With Drain

Pneumothorax 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?
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.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

Jason Fries, PhD
No machine learning experience

Beginner-level Python

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

Median Crowdsourced Model
F1 = 24.6

Best Snorkel Model
F1 = 48.7
Structured data
Fonduer: Handling Richly-Formatted Data

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

<table>
<thead>
<tr>
<th>Doc. level Candidates</th>
<th>Multimodal Supervision</th>
</tr>
</thead>
<tbody>
<tr>
<td>BC856 160</td>
<td>✓</td>
</tr>
<tr>
<td>BC856 -65</td>
<td>✓</td>
</tr>
<tr>
<td>BC856 150</td>
<td>✓</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>Prec.</th>
<th>Rec.</th>
</tr>
</thead>
<tbody>
<tr>
<td>MEMEX</td>
<td>87%</td>
<td>89%</td>
</tr>
<tr>
<td>IoT</td>
<td>73%</td>
<td>81%</td>
</tr>
<tr>
<td>GWAS</td>
<td>89%</td>
<td>81%</td>
</tr>
<tr>
<td>Paleo</td>
<td>72%</td>
<td>38%</td>
</tr>
</tbody>
</table>

SIGMOD 2018
**Input:** User-customized HTML auction pages → **Output:** Structured knowledge base

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

Improve auction searching quality and UX

Fonduer
Variant rs2681492

Simple phenotype Hypertension | Blood pressure

Detailed phenotype Systolic

p-value 3.0e-11

Source PMID: 19430479, Tbl. 1

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

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*
Time Series & Video


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

There is a lack of labeled datasets targeting BAV subjects

Source: www.umcvc.org
## 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

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

< 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

Ground Truth Labels
Traditional Weak Supervision
Multi-Resolution Weak Supervision

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
**Result:** 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**