Confident Learning for Machines and Humans

Curtis G. Northcutt
Massachusetts Institute of Technology

For Learning, the Data is as important as the Model

In machine learning, we tend to focus on the model

When algorithms mess up, the nearest human gets the blame

Source: MIT Technology Review (May 28, 2019)

For Learning, the Data is as important as the Model

In machine learning, we tend to focus on the model

When algorithms

Source: MIT Technology Review (May 28, 2019)

But learning depends on the quality of the data

Deep neural networks easily fit random labels.

- Zhang et al. (ICLR, 2017)

Despite their massive size, successful deep artificial neural networks can exhibit a remarkably small difference between training and test performance. Conventional wisdom attributes small generalization error either to properties of the model family, or to the regularization techniques used during training.

Through extensive systematic experiments, we show how these traditional approaches fail to explain why large neural networks generalize well in practice. Specifically, our experiments establish that state-of-the-art convolutional networks for image classification trained with stochastic gradient methods easily fit a random labeling of the training data. This phenomenon is qualitatively unaffected by explicit regularization, and occurs even if we replace the true images by completely unstructured random noise. We corroborate these experimental findings with a theoretical construction showing that simple depth two neural networks already have perfect finite sample expressivity as soon as the number of parameters exceeds the number of data points as it usually does in practice.

For Learning, the Data is as important as the Model

In machine learning, we tend to focus on the model

But learning depends on the quality of the data

When algorithms are trained with mislabeled data

RE-

Deep neural networks easily fit random labels.

Focus of this talk!

- *Zhang et al. (ICLR, 2017)*

Source: MIT Technology Review (May 28, 2019)

ily, or to the regularization techniques used during training.

Through extensive systematic experiments, we show how these traditional approaches fail to explain why large neural networks generalize well in practice. Specifically, our experiments establish that state-of-the-art convolutional networks for image classification trained with stochastic gradient methods easily fit a random labeling of the training data. This phenomenon is qualitatively unaffected by explicit regularization, and occurs even if we replace the true images by completely unstructured random noise. We corroborate these experimental findings with a theoretical construction showing that simple depth two neural networks already have perfect finite sample expressivity as soon as the number of parameters exceeds the number of data points as it usually does in practice.

Central Claim of My Thesis

Quantifying uncertainty in dataset labels

empowers machines and humans to learn and perform tasks **with confidence** in **noisy, real-world** environments

"When a system isn't performing well, teams instinctively try to improve the code. But for practical applications, ***it's more effective instead to focus on improving the data.***"

- Andrew Ng (April 6, 2021)

To support this claim, this talk addresses two questions

1. **In noisy, realistic settings, can we assemble a principled framework for quantifying, finding, and learning with label errors using a machine's confidence?**
    a. Traditionally, ML has focused on "Which model best learns with noisy labels?"
    b. In this talk I ask, "Which data is mislabeled?"

If Q1 works out, and there are label errors in datasets… does it matter? This leads us to Q2...

2. **Are we unknowingly benchmarking the progress of ML models, based on erroneous test sets? If so, can we quantify how much noise destabilizes benchmarks?**

Steps to Confident Learning for Machines and Humans

1

Precursors to CL

Machine learning for human learning requires dealing with real-world, noisy labels

Northcutt, Ho, & Chuang (C&E, 2016)
Northcutt, Wu, & Chuang (UAI, 2017)
Northcutt, Leon, & Chen (L@S, 2017)
Corrigan-Gibbs, Gupta, Northcutt, Cutrell, & Thies (TOCHI 2015, CHI 2016)

Contributions (in the context of what's already been done)

- Confident learning is the first framework to:
  - estimate the joint distribution of noisy labels and true labels directly
    - Prior work focuses on estimating conditionals/marginals of the joint (e.g. label flipping rates)
      - Sukhbaatar & Fergus (2015), Goldberger and BenReuven (2017), Northcutt et al. (2017), Clayton Scott (2015)
  - provide sufficient conditions for exactly finding label errors with per-example noisy model outputs
    - Prior theory with noisy labels (mostly) focuses on learnability/ estimators (not the data)
      - Angluin and Laird (1988), Clayton Scott (2015), Natarajan et al. (2013, 2017), Liu & Tao (2015), Ghosh et al. (2015)
- Label Errors + Implications for ML
  - First work to quantify noise and find label errors at scale across ten popular ML test sets.
    - Prior work on ImageNet, but it was not known that, e.g. MNIST also has many label errors
      - Shankar et al. (2020), Beyer et al. (2020), Recht et al. (2019), Tsipras et al., (2020), Taori et al. (2021)
  - First work to estimate the noise prevalence needed to destabilize benchmarks in popular datasets
    - Prior work has verified linear trends under distributional shift of test sets
      - Taori et al. (2021), Recht et al. (2019), Mania & Sra, (2021), Tsipras et al., (2020)

Steps to Confident Learning for Machines and Humans

1 — 2 — 3 — 4

Precursors to CL

Machine learning for human learning requires dealing with real-world, noisy labels

Northcutt, Ho, & Chuang (C&E, 2016)
Northcutt, Wu, & Chuang (UAI, 2017)
Northcutt, Leon, & Chen (L@S, 2017)
Corrigan-Gibbs, Gupta, Northcutt, Cutrell, & Thies (TOCHI 2015, CHI 2016)

Confident Learning

We develop a principled framework of theory and algorithms for quantifying, finding, and learning with label noise in datasets.

https://github.com/cgnorthcutt/cleanlab

Northcutt, Jiang, & Chuang (JAIR, 2021)

Label Errors in ML Datasets

We find tens of thousands (3.4%) of label errors in the most commonly benchmarked ML test sets.

labelerrors.com

Northcutt, Athalye, & Lin (NeurIPS Workshop on Dataset Curation and Security, 2020)

Implications for ML Practitioners

We study whether practitioners are unknowingly benchmarking the progress of ML based on erroneous test sets? How noisy is too noisy?

https://github.com/cgnorthcutt/label-errors

Northcutt, Athalye, & Mueller (ICLR RobustML Workshop, 2021) (ICLR WeaSuL Workshop, 2021)

Precursors to Confident Learning

Takeaway: Learning with confidence for human applications requires dealing with real-world noisy labels

Democratizing human learning via open online courses → Cheating detection → Key observation: no ground truth cheating labels. → Binary classification with noisy labels

Ho, Chuang, Reich, Coleman, Whitehill, Northcutt, Williams, Hansen, Lopez, Petersen, 2015)

Northcutt, Leon, & Chen (L@S, 2017)

Corrigan-Gibbs, Gupta, Northcutt, Cutrell, & Thies (TOCHI 2015, CHI 2016)

Northcutt, Ho, & Chuang (C&E, 2016)

*"Its exceedingly hard to prove that a person **didn't** do something"*
- Walter C. Northcutt

e.g. "innocence until proven guilty" (in lieu of ground truth "innocent" labels)

For human-inspired datasets, we often have noisy labels

Northcutt, Wu, & Chuang (UAI, 2017)

Consistent/exact estimation of false positives & false negative noisy labels

Steps to Confident Learning for Machines and Humans

1

2

3

4

Precursors to CL

Machine learning for human learning requires dealing with real-world, noisy labels

Northcutt, Ho, & Chuang (C&E, 2016)
Northcutt, Wu, & Chuang (UAI, 2017)
Northcutt, Leon, & Chen (L@S, 2017)
Corrigan-Gibbs, Gupta, Northcutt, Cutrell, & Thies (TOCHI 2015, CHI 2016)

Confident Learning

We develop a principled framework of theory and algorithms for quantifying, finding, and learning with label noise in datasets.

https://github.com/cgnorthcutt/cleanlab

Northcutt, Jiang, & Chuang (JAIR, 2021)

Label Errors in ML Datasets

We find tens of thousands (3.4%) of label errors in the most commonly benchmarked ML test sets.

labelerrors.com

Northcutt, Athalye, & Lin (NeurIPS Workshop on Dataset Curation and Security, 2020)

Implications for ML Practitioners

We study whether practitioners are unknowingly benchmarking the progress of ML based on erroneous test sets? How noisy is too noisy?

https://github.com/cgnorthcutt/label-errors

Northcutt, Athalye, & Mueller (ICLR RobustML Workshop, 2021) (ICLR WeaSuL Workshop, 2021)

Steps to Confident Learning for Machines and Humans

1

Precursors to CL

Machine learning for human learning requires dealing with real-world, noisy labels

Northcutt, Ho, & Chuang (C&E, 2016)
Northcutt, Wu, & Chuang (UAI, 2017)
Northcutt, Leon, & Chen (L@S, 2017)
Corrigan-Gibbs, Gupta, Northcutt, Cutrell, & Thies (TOCHI 2015, CHI 2016)

2

Confident Learning

We develop a principled framework of theory and algorithms for quantifying, finding, and learning with label noise in datasets.

https://github.com/cgnorthcutt/cleanlab

Northcutt, Jiang, & Chuang (JAIR, 2021)

Organization for this part of the talk:

1. What is confident learning?
2. Situating confident learning
   a. Types of noise
3. How does CL work? (methods)
4. Comparison with other methods
5. Why does CL work? (theory)
   a. Intuitions
   b. Principles
6. Examples + Dataset Curation

What is Confident learning (CL)?

Confident learning (CL) is a principled framework of theory and algorithms for classification with noisy labels.

CL provides affordances for:

- Complete characterization of label noise in a dataset
- Finding label errors in a dataset
- Learning with noisy labels
- Dataset curation

Situating Confident Learning within ML

Supervised Learning

> Classification with perfect observed labels

> Classification with **noisy observed labels**

> Classification with **noisy labels + noisy (real-world) model outputs**

(i.e, models that yield stochastic outputs/predicted class-probabilities)

Notation

$\tilde{y}$ - observed, noisy label

y^* - unobserved, latent, correct label

$\boldsymbol{X}_{\tilde{y}=i,y^*=j}$ - set of examples with noisy observed label *i*, but actually belong to class *j*

$\boldsymbol{C}_{\tilde{y}=i,y^*=j} = |\boldsymbol{X}_{\tilde{y}=i,y^*=j}|$ - counts in each set

$p(\tilde{y}=i, y^*=j)$ - joint distribution of noisy labels and true labels (estimated by normalizing $\boldsymbol{C}_{\tilde{y}=i,y^*=j}$)

$p(\tilde{y}=i | y^*=j)$ - transition probability that label *j* is flipped to label *i*

Organization for this part of the talk:

✓1. What is confident learning?
✓2. Situate confident learning
    a. Noise + related work
3. How does CL work? (methods)
4. Comparison with other methods
5. Why does CL work? (theory)
    a. Intuitions
    b. Principles
6. Examples + Dataset Curation

Types of label noise (how noisy labels are generated)

- Uniform/symmetric class-conditional label noise
  - $p(\tilde{y}{=}i \mid y^*{=}j) = \epsilon, \forall i \neq j$
  - Goldberger and BenReuven (2017); Arazo et al. (2019); Huang et al. (ICCV, 2019); Chen et al. (ICML, 2019)

0.6	0.1	0.1	0.1	0.1
0.1	0.6	0.1	0.1	0.1
0.1	0.1	0.6	0.1	0.1
0.1	0.1	0.1	0.6	0.1
0.1	0.1	0.1	0.1	0.6

Least assuming

- Systematic/Asymmetric Class-Conditional Label Noise
  - $p(\tilde{y}{=}i \mid y^*{=}j)$ can be <u>any</u> valid distribution $\leftarrow$ Confident Learning
  - Wang et al. (2019), Natarajan et al. (2017), Lipton et al. (2018), Goldberger & Ben-Reuven (2017), Sukhbaatar et al. (2015)
- Instance-Dependent Label noise
  - $p(\tilde{y}{=}i \mid y^*{=}j, \boldsymbol{x})$
  - Strong assumptions on the covariates of $\boldsymbol{x}$ to reduce to class-conditional case
  - Out of scope for this talk
  - Menon et al. (2016), Xia et al. (2020), Cheng et al. (2020), Berthon et al. (2020), Wang et al. (2021)

Is a label noise process assumption necessary? (yes)

Consider the predicted probabilities of a model

$$\hat{p}(\tilde{y}{=}i; \boldsymbol{x}, \boldsymbol{\theta})$$

$\hat{p}(\tilde{y}{=}i; \boldsymbol{x}, \boldsymbol{\theta})$ expresses both:

- noisy model outputs (**epistemic** uncertainty)
- label noise of every example (**aleatoric** uncertainty)

No noise process assumption → cannot **disambiguate** the two sources of noise

To disambiguate epistemic uncertainty from aleatoric uncertainty, we use a reasonable assumption to remove the dependency on $\boldsymbol{x}$

CL assumes **class-conditional** label noise

We **assume** labels are flipped based on an unknown transition matrix $p(\tilde{y}|y^*)$ that depends only on pairwise noise rates between classes, not the data $\boldsymbol{x}$

$$p(\tilde{y}|y^*;\boldsymbol{x}) = p(\tilde{y}|y^*)$$

This assumption is reasonable for real-world data. Let's look at some...

$\tilde{y}$ - observed, noisy label

y^* - unobserved, latent, correct label

Class-conditional noise process first introduced by Angluin and Laird (1988)

Label Errors in ML Test Sets About

In real-world images, lots of “boars” were mislabeled as “pigs”

But no “missiles” or “keyboards” were mislabeled as “pigs”

This “class-conditional” label noise depends on the class, not the image data $\boldsymbol{x}$ (what the pig looks like)

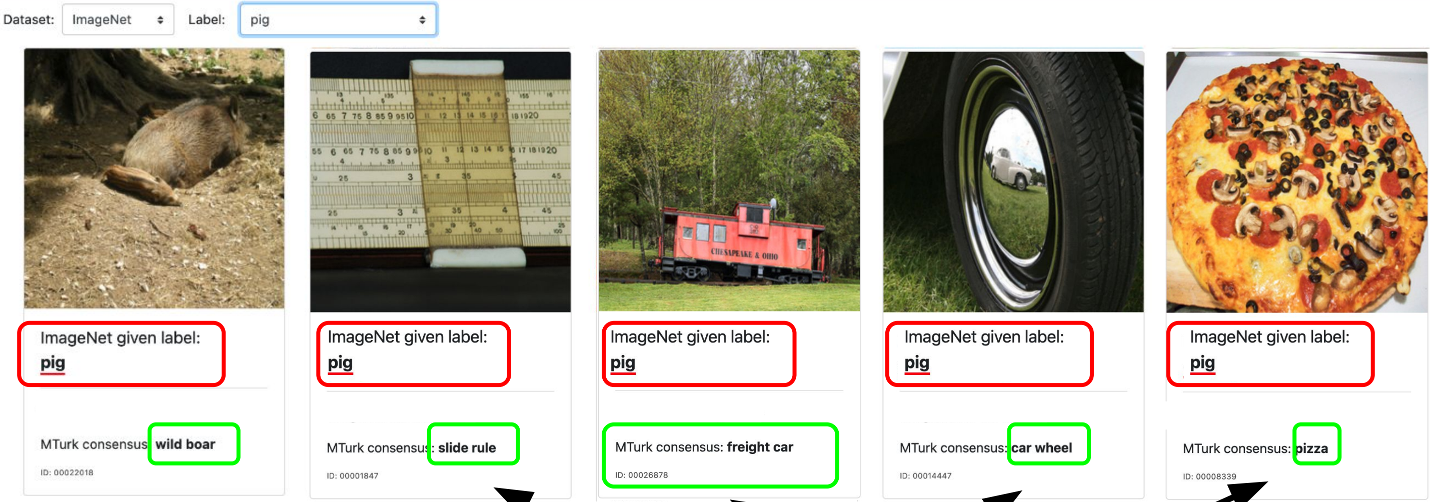
Given its realistic nature, we choose to solve for “class-conditional noise” in CL.

labelerrors.com

Label Errors in ML Test Sets About

What does uniform label noise look like?

Goldberger and BenReuven (2017)
Arazo et al. (2019)



Fictitious examples
(**not** naturally occurring)

Does label noise matter? Deep learning is robust to label noise… right?

(Jindal et al. ICDM 2016), (Krause et al. ECCV 2016) suggest that “with enough data, learning is possible with arbitrary amounts of uniformly random label noise”

Quotes across the literature:

- “label noise may be a limited issue if networks are trained on billions of images” *(Mahajan et al. ECCV 2018)*
- “it seems the scale of data can overpower noise in the label space” *(Sun et al. ICCV 2017)*
- “Successful learning is possible with an arbitrary amount of noise” *(Rolnick et al. arXiv 2017)*
- “[Neural networks] miraculously avoid bad minima [caused by label errors].” *(Huang et al. PMLR 2019)*

Does label noise matter? Deep learning is robust to label noise… right?

(Jindal et al. ICDM 2016), (Krause et al. ECCV 2016) suggest that “with enough data, learning is

Qu

These results assume $\boxed{\text{uniformly random label noise}}$ and usually don’t apply to **real-world settings**.

(Huang et al. PMLR 2019)

Types of Noise that CL does NOT cover

Noise in Data

Blurry images, adversarial examples, typos in text, background noise in audio

CL assumes **labels** are noisy, not data.

Annotator Label Noise

1 Annotation: Sports Car

2 Annotation: Toy Car

3 Annotation: Toy Car

Dawid and **Skene** (1979)

CL assumes **one** annotation per example

Types of methods for Learning with Noisy Labels

Model-Centric Methods

"Change the Loss"

- Use loss from another network
  - Co-Teaching (Han et al., 2018)
  - MentorNet (Jiang et al., 2017)
- Modify loss directly
  - SCE-loss (Wang et al., 2019)
- Importance reweighting
  - (Liu & Tao, 2015; Patrini et al., 2017; Reed et al., 2015; Shu et al., 2019; Goldberger & Ben-Reuven, 2017)

We'll see later why these approaches propagate error to the learned model

Data-Centric Methods

"Change the Data"

- Find label errors in datasets
- Then learn with(out) noisy labels by providing cleaned data for training
  - (Pleiss et al., 2020; Yu et al., ICML, 2019; Li et al., ICLR, 2020; Wei et al., CVPR, 2020, Northcutt et al., JAIR, 2021)

Our approach

Organization for this part of the talk:

✓1. What is confident learning?
✓2. Situate confident learning
    a. Noise + related work
3. How does CL work? (methods)
4. Comparison with other methods
5. Why does CL work? (theory)
    a. Intuitions
    b. Principles
6. Examples + Dataset Curation

How does confident learning work?

Directly estimate the joint distribution of observed noisy labels and latent true labels.

$p(\tilde{y}|y^*)$

$p(y^*)$

$p(y^*|\tilde{y})$

$p(\tilde{y}, y^*)$	$y^*=dog$	$y^*=fox$	$y^*=cow$
$\tilde{y}=dog$	0.25	0.1	0.05
$\tilde{y}=fox$	0.14	0.15	0
$\tilde{y}=cow$	0.08	$\boxed{0.03}$	0.2

Off-diagonals tell you what fraction of your dataset is mislabeled.
Example -- "3% of your cow images are actually foxes"

How does confident learning work?

To estimate $p(\tilde{y}, y^*)$ and find label errors, confident learning requires two inputs:

- Noisy labels, $\tilde{y}$
- Predicted probabilities, $\hat{p}(\tilde{y}{=}i; \boldsymbol{x}, \boldsymbol{\theta})$

Note: CL is scale-invariant w.r.t. outputs, i.e. raw logits work as well

How does confident learning work?

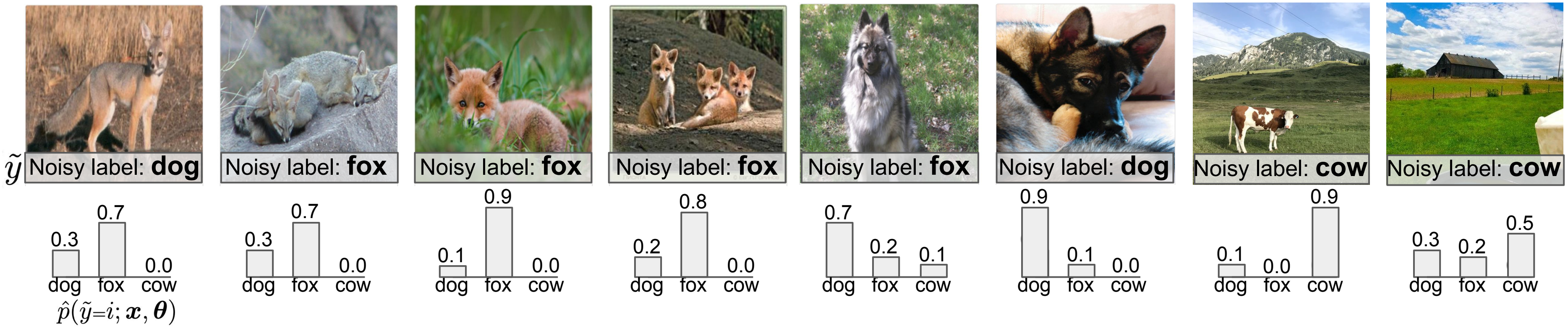
Key idea: First we find thresholds as a proxy for the machine's self-confidence, on average, for each task/class j

$$t_j = \frac{1}{|\boldsymbol{X}_{\tilde{y}=j}|} \sum_{\boxed{\boldsymbol{x} \in \boldsymbol{X}_{\tilde{y}=j}}} \hat{p}(\boxed{\tilde{y} = j}; \boldsymbol{x}, \boldsymbol{\theta})$$

How does confident learning work?

Before confident learning,starts, a model is trained on this data using cross-validation, to produce $\hat{p}(\tilde{y}{=}i; \boldsymbol{x}, \boldsymbol{\theta})$, the out-of-sample predicted probabilities

How does confident learning work?



$$\frac{t_j}{}$$

$$t_{\text{dog}} = 0.7$$

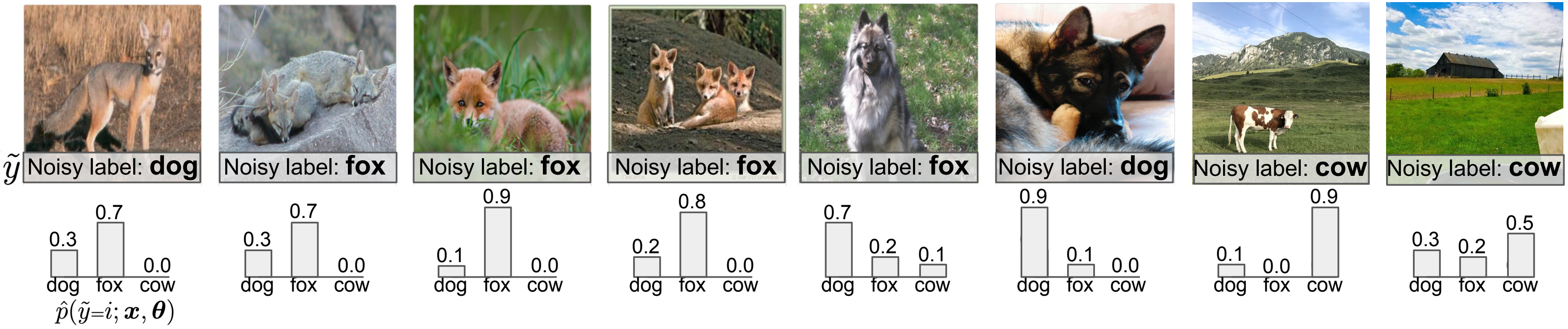
$$t_{\text{fox}} = 0.7$$

$$t_{\text{cow}} = 0.9$$

$$\hat{\mathbf{X}}_{\tilde{y}=i, y^*=j} = \{\boldsymbol{x} \in \mathbf{X}_{\tilde{y}=i} : \; \hat{p}(\tilde{y} = j; \boldsymbol{x}, \boldsymbol{\theta}) \geq t_j\}$$

CL estimates sets of label errors for each pair of (noisy label i, true label j)

How does confident learning work?



$$t_j$$

$$t_{\text{dog}} = 0.7$$
$$t_{\text{fox}} = 0.7$$
$$t_{\text{cow}} = 0.9$$

$$\hat{\boldsymbol{X}}_{\tilde{y}=i, y^*=j} = \{\boldsymbol{x} \in \boldsymbol{X}_{\tilde{y}=i} : \ \hat{p}(\tilde{y} = j; \boldsymbol{x}, \boldsymbol{\theta}) \geq t_j\}$$

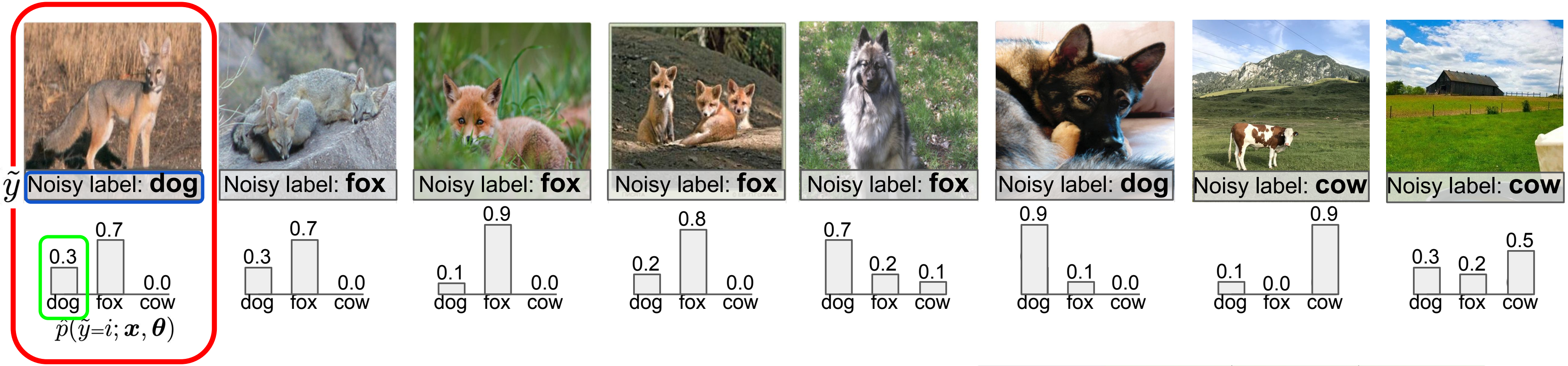
CL estimates sets of label errors for each pair of (noisy label i, true label j)

$\boldsymbol{C}_{\tilde{y}, y^*}$	$y^*{=}dog$	$y^*{=}fox$	$y^*{=}cow$
$\tilde{y}{=}dog$			
$\tilde{y}{=}fox$			
$\tilde{y}{=}cow$			

Creating a matrix of counts to estimate the unnormalized joint distribution

The confident joint $\boldsymbol{C}_{\tilde{y}, y^*}$ counts the size of each set $\rightarrow \boldsymbol{C}_{\tilde{y}, y^*}[i][j] = |\hat{\boldsymbol{X}}_{\tilde{y}=i, y^*=j}|$

How does confident learning work?



$$\underline{t_j}$$

$$\boxed{t_{\text{dog}} = 0.7}$$

$$t_{\text{fox}} = 0.7$$

$$t_{\text{cow}} = 0.9$$

$$\hat{\boldsymbol{X}}_{\tilde{y}=i, y^*=j} =$$

0.3 $\not\geq$ 0.7

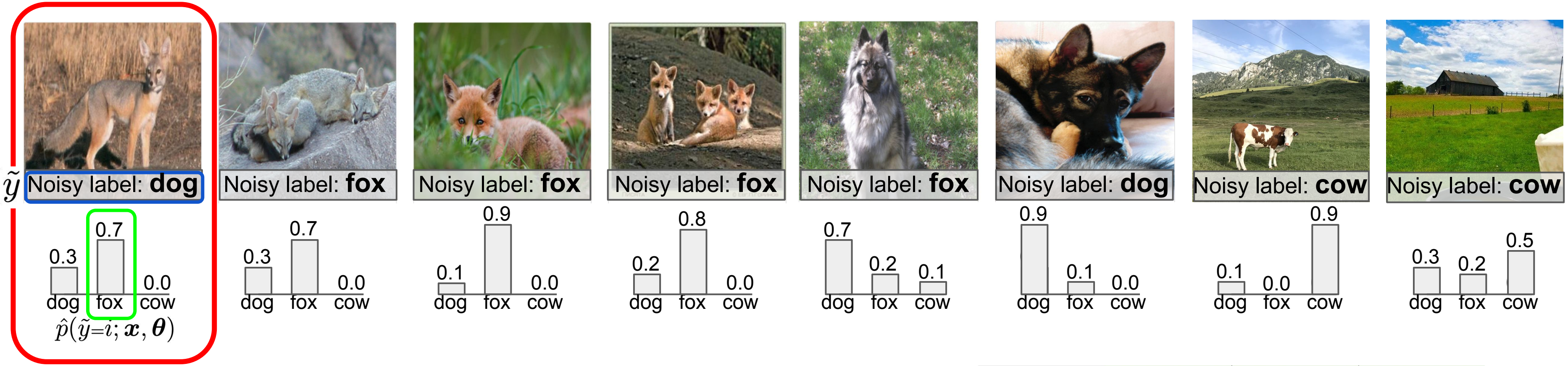
$$\{\boldsymbol{x} \in \boldsymbol{X}_{\boxed{\tilde{y}=i}} : \hat{p}(\tilde{y} = \boxed{j}; \boldsymbol{x}, \boldsymbol{\theta}) \geq \boxed{t_j}\}$$

$\boldsymbol{C}_{\tilde{y},y^*}$	$y^*=dog$	$y^*=fox$	$y^*=cow$
$\tilde{y}=dog$	0	0	0
$\tilde{y}=fox$	0	0	0
$\tilde{y}=cow$	0	0	0

$$\boldsymbol{C}_{\tilde{y},y^*}[i][j] = |\hat{\boldsymbol{X}}_{\boxed{\tilde{y}=i}, \boxed{y^*=j}}|$$

t_j - class self-confidence thresholds
$\hat{p}(\tilde{y}=i; \boldsymbol{x}, \boldsymbol{\theta})$ - out-of-sample predicted probabilities

How does confident learning work?



$$\underline{t_j}$$

$$t_{\text{dog}} = 0.7$$

$$\boxed{t_{\text{fox}} = 0.7}$$

$$t_{\text{cow}} = 0.9$$

$$\hat{\boldsymbol{X}}_{\tilde{y}=i, y^*=j} =$$

✓

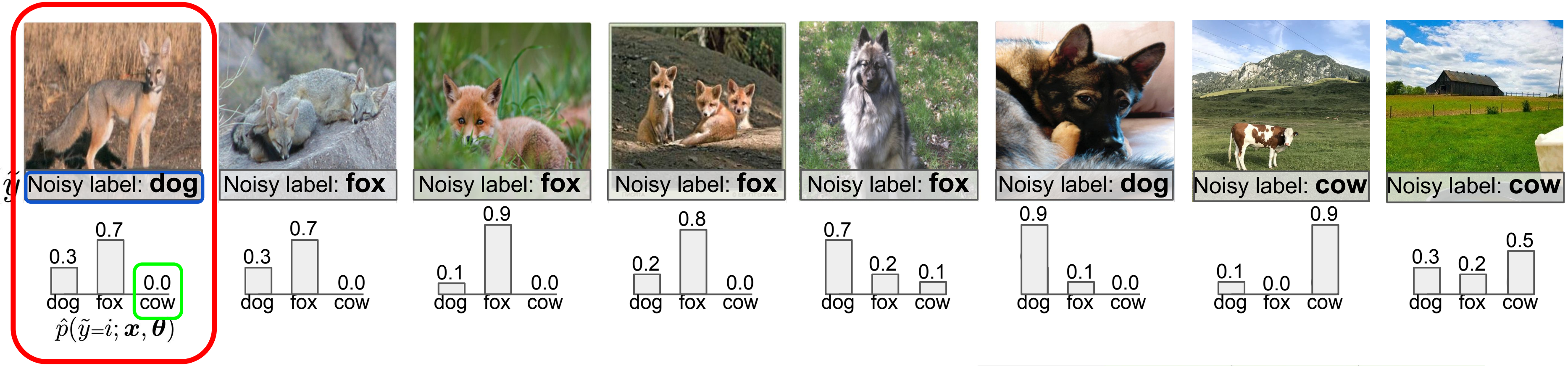
0.7 ≥ 0.7

$$\{\boldsymbol{x} \in \boldsymbol{X}_{\boxed{\tilde{y}=i}} : \ \hat{p}(\tilde{y} = \boxed{j}; \boldsymbol{x}, \boldsymbol{\theta}) \geq \boxed{t_j}\}$$

$\boldsymbol{C}_{\tilde{y},y^*}$	$y^*{=}dog$	$\boxed{y^*{=}fox}$	$y^*{=}cow$
$\boxed{\tilde{y}{=}dog}$	0	1	0
$\tilde{y}{=}fox$	0	0	0
$\tilde{y}{=}cow$	0	0	0

$$\boldsymbol{C}_{\tilde{y},y^*}[i][j] = |\hat{\boldsymbol{X}}_{\boxed{\tilde{y}=i}, \boxed{y^*=j}}|$$

How does confident learning work?



$$\frac{t_j}{}$$

$$t_{\text{dog}} = 0.7$$

$$t_{\text{fox}} = 0.7$$

$$\boxed{t_{\text{cow}} = 0.9}$$

$$\hat{\boldsymbol{X}}_{\tilde{y}=i, y^*=j} =$$

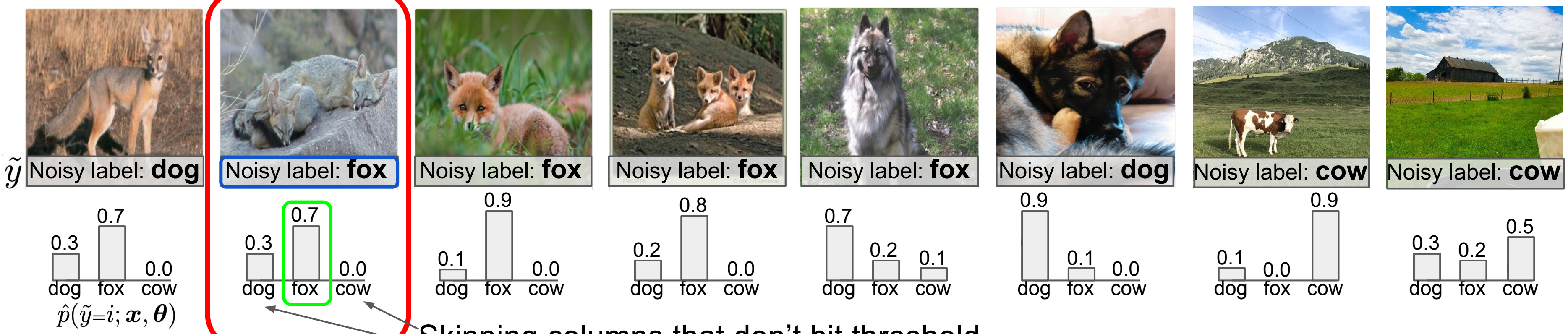
0.0 $\not\geq$ 0.9

$$\{\boldsymbol{x} \in \boldsymbol{X}_{\boxed{\tilde{y}=i}} : \ \hat{p}(\tilde{y} = \boxed{j}; \boldsymbol{x}, \boldsymbol{\theta}) \geq \boxed{t_j}\}$$

$\boldsymbol{C}_{\tilde{y}, y^*}$	$y^*{=}dog$	$y^*{=}fox$	$y^*{=}cow$
$\tilde{y}{=}dog$	0	1	0
$\tilde{y}{=}fox$	0	0	0
$\tilde{y}{=}cow$	0	0	0

$$\boldsymbol{C}_{\tilde{y}, y^*}[i][j] = |\hat{\boldsymbol{X}}_{\boxed{\tilde{y}=i}, \boxed{y^*=j}}|$$

How does confident learning work?



Skipping columns that don't hit threshold

$$t_j$$

$$t_{\text{dog}} = 0.7$$

$$\boxed{t_{\text{fox}} = 0.7}$$

$$t_{\text{cow}} = 0.9$$

$$\hat{\boldsymbol{X}}_{\tilde{y}=i, y^*=j} =$$

✓

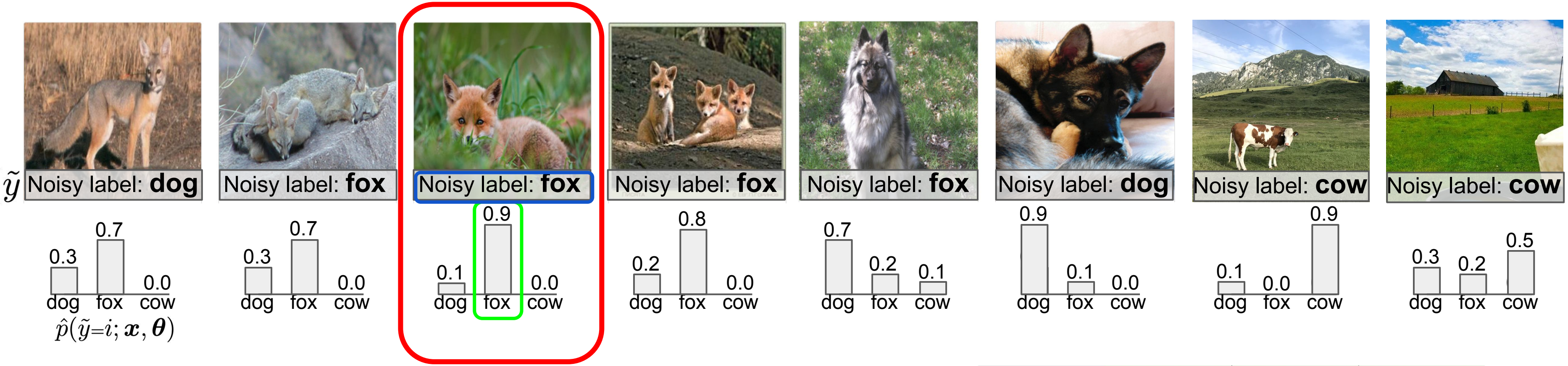
$0.7 \geq 0.7$

$$\{\boldsymbol{x} \in \boldsymbol{X}_{\boxed{\tilde{y}=i}} : \ \hat{p}(\tilde{y} = \boxed{j}; \boldsymbol{x}, \boldsymbol{\theta}) \geq \boxed{t_j}\}$$

$\boldsymbol{C}_{\tilde{y},y^*}$	$y^*=dog$	$\boxed{y^*=fox}$	$y^*=cow$
$\tilde{y}=dog$	0	1	0
$\boxed{\tilde{y}=fox}$	0	1	0
$\tilde{y}=cow$	0	0	0

$$\boldsymbol{C}_{\tilde{y},y^*}[i][j] = |\hat{\boldsymbol{X}}_{\boxed{\tilde{y}=i}, \boxed{y^*=j}}|$$

How does confident learning work?



t_j
$t_{\text{dog}} = 0.7$
$\boxed{t_{\text{fox}} = 0.7}$
$t_{\text{cow}} = 0.9$

$$\hat{\boldsymbol{X}}_{\tilde{y}=i, y^*=j} =$$

✓

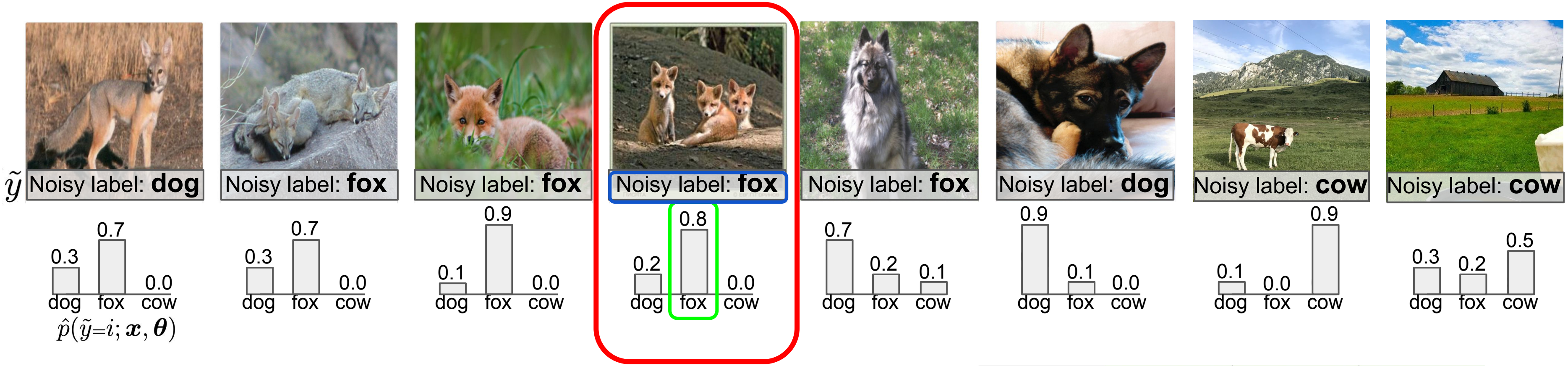
$0.9 \geq 0.7$

$$\{\boldsymbol{x} \in \boldsymbol{X}_{\boxed{\tilde{y}=i}} : \ \hat{p}(\tilde{y} = \boxed{j}; \boldsymbol{x}, \boldsymbol{\theta}) \geq \boxed{t_j}\}$$

$\boldsymbol{C}_{\tilde{y}, y^*}$	$y^*{=}dog$	$y^*{=}fox$	$y^*{=}cow$
$\tilde{y}{=}dog$	0	1	0
$\tilde{y}{=}fox$	0	2	0
$\tilde{y}{=}cow$	0	0	0

$$\boldsymbol{C}_{\tilde{y}, y^*}[i][j] = |\hat{\boldsymbol{X}}_{\boxed{\tilde{y}=i}, \boxed{y^*=j}}|$$

How does confident learning work?



$$\frac{t_j}{}$$

$t_{\text{dog}} = 0.7$

$\boxed{t_{\text{fox}} = 0.7}$

$t_{\text{cow}} = 0.9$

$$\hat{\boldsymbol{X}}_{\tilde{y}=i,y^*=j} =$$

✓

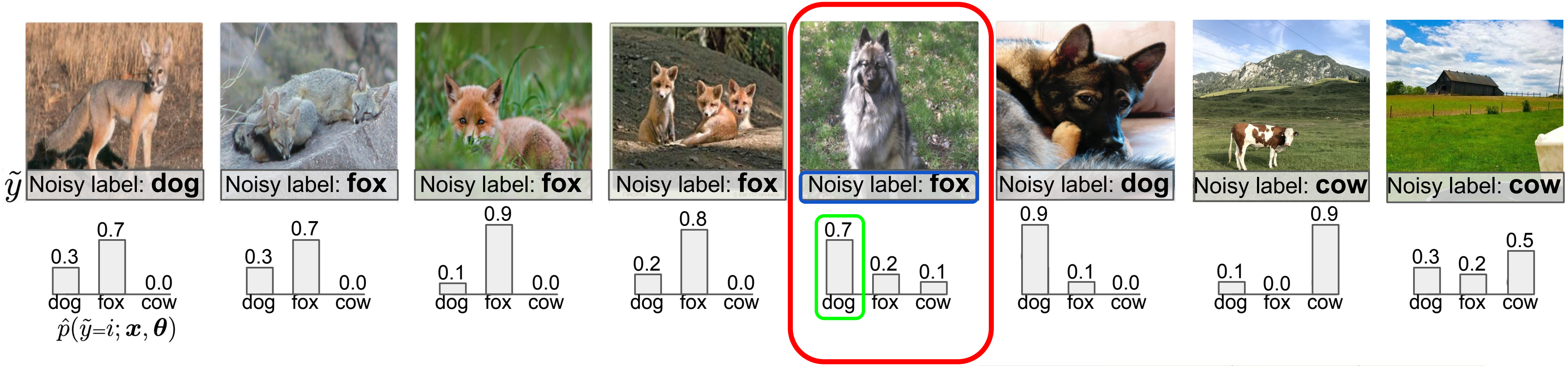
0.8 ≥ 0.7

$$\{\boldsymbol{x} \in \boldsymbol{X}_{\boxed{\tilde{y}=i}} : \ \hat{p}(\tilde{y} = \boxed{j}; \boldsymbol{x}, \boldsymbol{\theta}) \geq \boxed{t_j}\}$$

$\boldsymbol{C}_{\tilde{y},y^*}$	$y^*=dog$	$y^*=fox$	$y^*=cow$
$\tilde{y}=dog$	0	1	0
$\tilde{y}=fox$	0	3	0
$\tilde{y}=cow$	0	0	0

$$\boldsymbol{C}_{\tilde{y},y^*}[i][j] = |\hat{\boldsymbol{X}}_{\boxed{\tilde{y}=i},\boxed{y^*=j}}|$$

How does confident learning work?



$$\underline{t_j}$$

$$\boxed{t_{\text{dog}} = 0.7}$$

$$t_{\text{fox}} = 0.7$$

$$t_{\text{cow}} = 0.9$$

$$\hat{\boldsymbol{X}}_{\tilde{y}=i, y^*=j} =$$

✓

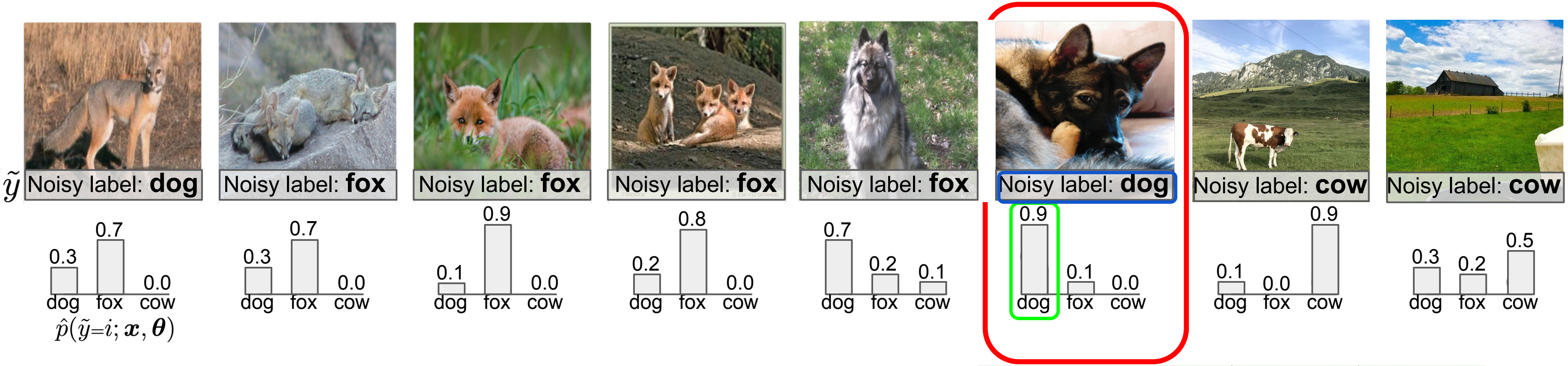
0.7 $\geq$ 0.7

$$\{\boldsymbol{x} \in \boldsymbol{X}_{\boxed{\tilde{y}=i}} : \ \hat{p}(\tilde{y} = \boxed{j}; \boldsymbol{x}, \boldsymbol{\theta}) \geq \boxed{t_j}\}$$

$\boldsymbol{C}_{\tilde{y},y^*}$	$y^*{=}dog$	$y^*{=}fox$	$y^*{=}cow$
$\tilde{y}{=}dog$	0	1	0
$\tilde{y}{=}fox$	1	3	0
$\tilde{y}{=}cow$	0	0	0

$$\boldsymbol{C}_{\tilde{y},y^*}[i][j] = |\hat{\boldsymbol{X}}_{\boxed{\tilde{y}=i}, \boxed{y^*=j}}|$$

How does confident learning work?



$$\underline{t_j}$$

$$\boxed{t_{\text{dog}} = 0.7}$$

$$t_{\text{fox}} = 0.7$$

$$t_{\text{cow}} = 0.9$$

$$\hat{\boldsymbol{X}}_{\tilde{y}=i,y^*=j} =$$

✓

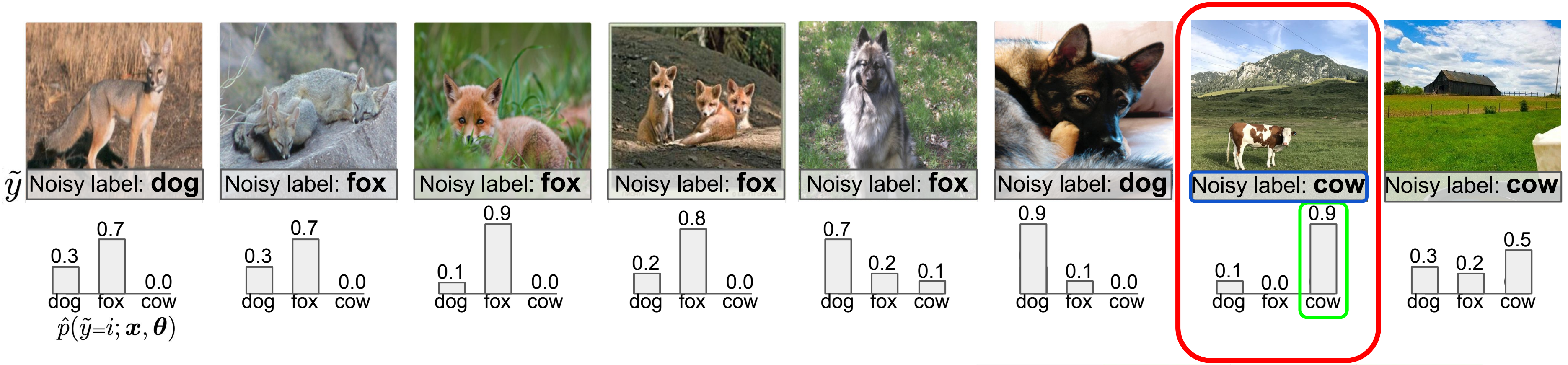
0.9 $\geq$ 0.7

$$\{\boldsymbol{x} \in \boldsymbol{X}_{\boxed{\tilde{y}=i}} : \hat{p}(\tilde{y} = \boxed{j}; \boldsymbol{x}, \boldsymbol{\theta}) \geq \boxed{t_j}\}$$

$C_{\tilde{y},y^*}$	$y^*=dog$	$y^*=fox$	$y^*=cow$
$\tilde{y}=dog$	1	1	0
$\tilde{y}=fox$	1	3	0
$\tilde{y}=cow$	0	0	0

$$C_{\tilde{y},y^*}[i][j] = |\hat{\boldsymbol{X}}_{\boxed{\tilde{y}=i},\boxed{y^*=j}}|$$

How does confident learning work?



$$
\begin{array}{c}
t_j \\ \hline
t_{\text{dog}} = 0.7 \\
t_{\text{fox}} = 0.7 \\
\boxed{t_{\text{cow}} = 0.9}
\end{array}
$$

$$\hat{\boldsymbol{X}}_{\tilde{y}=i, y^*=j} =$$

✓

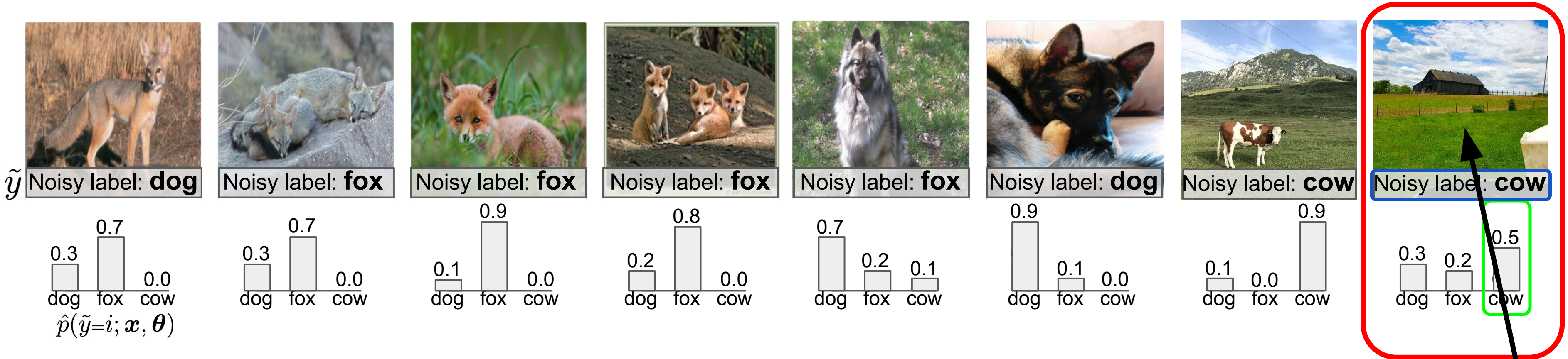
0.9 ≥ 0.9

$$\{\boldsymbol{x} \in \boldsymbol{X}_{\boxed{\tilde{y}=i}} : \; \hat{p}(\tilde{y} = \boxed{j}; \boldsymbol{x}, \boldsymbol{\theta}) \geq \boxed{t_j}\}$$

$\boldsymbol{C}_{\tilde{y},y^*}$	$y^*=dog$	$y^*=fox$	$y^*=cow$
$\tilde{y}=dog$	1	1	0
$\tilde{y}=fox$	1	3	0
$\tilde{y}=cow$	0	0	1

$$\boldsymbol{C}_{\tilde{y},y^*}[i][j] = |\hat{\boldsymbol{X}}_{\boxed{\tilde{y}=i}, \boxed{y^*=j}}|$$

How does confident learning work?



$$t_j$$

$$t_{\text{dog}} = 0.7$$

$$t_{\text{fox}} = 0.7$$

$$\boxed{t_{\text{cow}} = 0.9}$$

$$\hat{\boldsymbol{X}}_{\tilde{y}=i, y^*=j} =$$

$0.5 \not\geq 0.9$

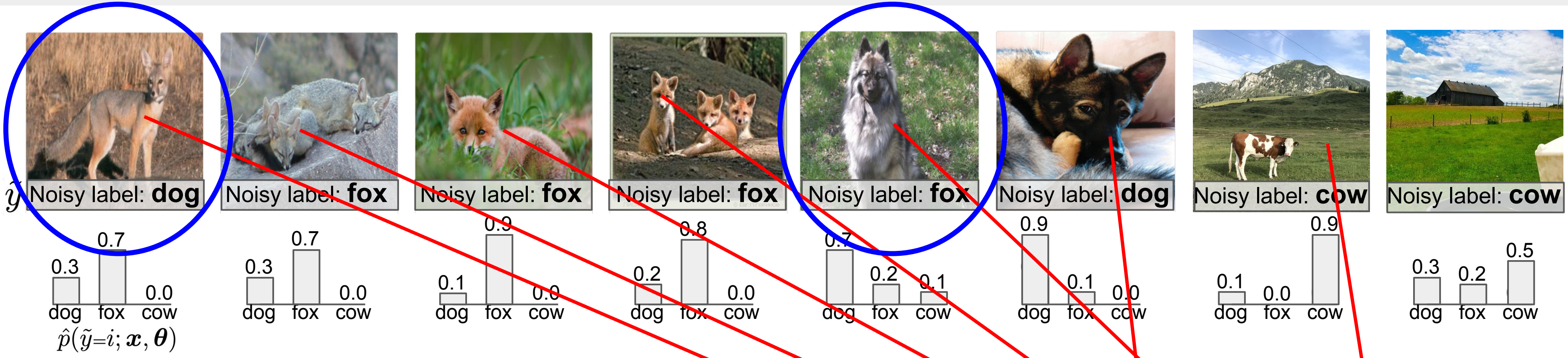
$$\{\boldsymbol{x} \in \boldsymbol{X}_{\boxed{\tilde{y}=i}} : \ \hat{p}(\tilde{y} = \boxed{j}; \boldsymbol{x}, \boldsymbol{\theta}) \geq \boxed{t_j}\}$$

$\boldsymbol{C}_{\tilde{y}, y^*}$	$y^*{=}dog$	$y^*{=}fox$	$\boxed{y^*{=}cow}$
$\tilde{y}{=}dog$	1	1	0
$\tilde{y}{=}fox$	1	3	0
$\boxed{\tilde{y}{=}cow}$	0	0	1

Out of distribution

$$\boldsymbol{C}_{\tilde{y}, y^*}[i][j] = |\hat{\boldsymbol{X}}_{\boxed{\tilde{y}=i}, \boxed{y^*=j}}|$$

How does confident learning work? (in 10 seconds)



$$t_j$$

$$t_{\text{dog}} = 0.7$$

$$t_{\text{fox}} = 0.7$$

$$t_{\text{cow}} = 0.9$$

$$\hat{\boldsymbol{X}}_{\tilde{y}=i, y^*=j} = \{\boldsymbol{x} \in \boldsymbol{X}_{\tilde{y}=i} : \ \hat{p}(\tilde{y} = j; \boldsymbol{x}, \boldsymbol{\theta}) \geq t_j\}$$

$\boldsymbol{C}_{\tilde{y}, y^*}$	$y^*{=}dog$	$y^*{=}fox$	$y^*{=}cow$
$\tilde{y}{=}dog$	1	1	0
$\tilde{y}{=}fox$	1	3	0
$\tilde{y}{=}cow$	0	0	1

$$\boldsymbol{C}_{\tilde{y}, y^*}[i][j] = |\hat{\boldsymbol{X}}_{\tilde{y}=i, y^*=j}|$$

Off diagonals are CL-guessed label errors

After looking through the entire dataset, we have:

$C_{\tilde{y},y^*}$	$y^*=dog$	$y^*=fox$	$y^*=cow$
$\tilde{y}=dog$	100	40	20
$\tilde{y}=fox$	56	60	0
$\tilde{y}=cow$	32	12	80

From $C_{\tilde{y},y^*}$ we obtain the joint distribution of label noise

Estimated

$\hat{p}(\tilde{y}, y^*)$	$y^*=dog$	$y^*=fox$	$y^*=cow$
$\tilde{y}=dog$	0.25	0.1	0.05
$\tilde{y}=fox$	0.14	0.15	0
$\tilde{y}=cow$	0.08	0.03	0.2

Organization for this part of the talk:

✓1. What is confident learning?
✓2. Situate confident learning
    a. Noise + related work
✓3. How does CL work? (methods)
4. Comparison with other methods
5. Why does CL work? (theory)
    a. Intuitions
    b. Principles
6. Examples + Dataset Curation

Compare Accuracy: Learning with 40% label noise in CIFAR-10

Fraction of zeros in the off-diagonals of $p(\tilde{y}|y^*)$

	0	0.6 ← More realistic (e.g. ImageNet)
Baseline (remove prediction != label)	83.9	84.2
Confident learning methods	84.8	86.2
	86.7	86.9
	87.1	**87.2**
	87.1	**87.2**
INCV (Chen et al., 2019)	84.4	73.6
Mixup (Zhang et al., 2018)	76.1	59.8
SCE-loss (Wang et al., 2019)	76.3	58.3
MentorNet (Jiang et al., 2018)	64.4	61.5
Co-Teaching (Han et al., 2018)	62.9	58.1
S-Model (Goldberger et al., 2017)	58.6	57.5
Reed (Reed et al., 2015)	60.5	58.6
Baseline	60.2	57.3

<u>Data-centric</u>
Train with errors removed

"Change the dataset"

Same perf

<u>Model-centric</u>
Train with errors

"adjust the loss"

Perf drop-off

Organization for this part of the talk:

✓1. What is confident learning?
✓2. Situate confident learning
    a. Noise + related work
✓3. How does CL work? (methods)
✓4. Comparison with other methods
5. Why does CL work? (theory)
    a. Intuitions
    b. Principles
6. Examples + Dataset Curation

Theory of Confident Learning

To understand CL performance, we studied conditions where CL exactly finds label errors, culminating in the following Theorem:

*As long as examples in class **i** are labeled **i** more than any other class, then...*

We prove realistic sufficient conditions (allowing significant error in all model outputs)

Such that CL still exactly finds label errors. $\hat{\boldsymbol{X}}_{\tilde{y}=i,y^*=j} \approxeq \boldsymbol{X}_{\tilde{y}=i,y^*=j}$

Intuition: CL theory builds on three principles

- The **Prune** Principle
  - remove errors, then train
  - Chen et al. (2019), Patrini et al. (2017), Van Rooyen et al. (2015)
- The **Count** Principle
  - use ratios of counts, not noisy model outputs
  - Page et al. (1997), Jiang et al. (2018)
- The **Rank** Principle
  - use rank of model outputs, not the noisy values
  - Natarajan et al. (2017), Forman (2005, 2008), Lipton et al. (2018)

CL Robustness Intuition 1: $\boxed{\text{Prune}}$

Key Idea:

Pruning enables robustness to stochastic/imperfect predicted probabilities $\hat{p}(\tilde{y}{=}i; \boldsymbol{x}, \boldsymbol{\theta})$

Pred probs are stochastic/erroneous for real-world models!!

Prior work
modifies the loss:
(e.g. importance reweighting)
(Liu & Tao, 2015; Patrini et al., 2017; Reed et al., 2015; Shu et al., 2019; Goldberger & Ben-Reuven, 2017)

$$\mathcal{L}'(\boldsymbol{\theta}) \sim g(\hat{p}(\tilde{y}; \boldsymbol{x}, \boldsymbol{\theta})) \cdot \mathcal{L}(\boldsymbol{\theta})$$

Error propagation

CL Robustness Intuition 1: Prune

Key Idea:

Pruning enables robustness to stochastic/imperfect predicted probabilities $\hat{p}(\tilde{y}{=}i; \boldsymbol{x}, \boldsymbol{\theta})$

Pred probs are stochastic/erroneous for real-world models!!

Prior work modifies the loss:

(e.g. importance reweighting)
(Liu & Tao, 2015; Patrini et al., 2017; Reed et al., 2015; Shu et al., 2019; Goldberger & Ben-Reuven, 2017)

$$\mathcal{L}'(\boldsymbol{\theta}) \sim g(\hat{p}(\tilde{y}; \boldsymbol{x}, \boldsymbol{\theta})) \cdot \mathcal{L}(\boldsymbol{\theta})$$

Error propagation

SGD weights update: $$\boldsymbol{\theta}^{(t+1)} := \boldsymbol{\theta}^{(t)} + \eta \nabla \mathcal{L}'(\boldsymbol{\theta})^{(t)}$$

Error propagation

CL Robustness Intuition 1: Prune

Key Idea:

Pruning enables robustness to stochastic/imperfect predicted probabilities $\hat{p}(\tilde{y}{=}i; \boldsymbol{x}, \boldsymbol{\theta})$

Pred probs are stochastic/erroneous for real-world models!!

Prior work modifies the loss:

(e.g. importance reweighting)
(Liu & Tao, 2015; Patrini et al., 2017; Reed et al., 2015; Shu et al., 2019; Goldberger & Ben-Reuven, 2017)

$$\mathcal{L}'(\boldsymbol{\theta}) \sim g(\hat{p}(\tilde{y}; \boldsymbol{x}, \boldsymbol{\theta})) \cdot \mathcal{L}(\boldsymbol{\theta})$$

Error propagation

SGD weights update:

$$\boldsymbol{\theta}^{(t+1)} := \boldsymbol{\theta}^{(t)} + \eta \nabla \mathcal{L}'(\boldsymbol{\theta})^{(t)}$$

Error propagation

Takeaway

CL methods
↓
Prune Label Errors
↓
Avoid loss reweighting
↓
Avoid this form of error propagation

CL Robustness Intuition 2: $\boxed{\text{Count}}$ & $\boxed{\text{Rank}}$

Same idea: **Counting** and **Ranking** enable robustness to erroneous probabilities

$$\hat{p}(\tilde{y}{=}i; \boldsymbol{x}, \boldsymbol{\theta})$$

But this time: Let's look at noise transition estimation

Other methods:

(Elkan & Noto, 2008; Sukhbaatar et al., 2015)

$$p(y^* = j | \tilde{y} = i) \approx \mathbb{E}[p(\hat{y} = j | \boldsymbol{x} \in \boldsymbol{X}_i)]$$

CL Robustness Intuition 2: $\boxed{\text{Count}}$ & $\boxed{\text{Rank}}$

Same idea: **Counting** and **Ranking** enable robustness to erroneous probabilities

$\hat{p}(\tilde{y}{=}i; \boldsymbol{x}, \boldsymbol{\theta})$

But this time: Let's look at noise transition estimation

Other methods:

(Elkan & Noto, 2008; Sukhbaatar et al., 2015)

$$p(y^* = j | \tilde{y} = i) \approx \mathbb{E}[p(\hat{y} = j | \boldsymbol{x} \in \boldsymbol{X}_i)]$$

Confident Learning:

$$p(y^* = j | \tilde{y} = i) = \frac{p(y^* = j, \tilde{y} = i)}{p(\tilde{y} = i)} \approx \frac{\boxed{\text{count}}(y^* = j, \tilde{y} = i)}{\boxed{\text{count}}(\tilde{y} = i)}$$

Enables CL to disambiguate aleatoric (label noise) from epistemic (model noise)

Robust statistic w/ counts + rank (1 step removed erroneous probs)

$$p(y^* = j | \tilde{y} = i) \approx \boxed{\frac{|(\boldsymbol{x} \in \boldsymbol{X}_i \text{ with } \boxed{\text{large}}\ p(y = j; \boldsymbol{x})|}{|\boldsymbol{x} \in \boldsymbol{X}_i|}}$$

e.g. Median of Means

CL Robustness Intuition 2: Count & Rank

Same idea: **Counting** and **Ranking** enable robustness to erron

But this time: Let's look at noise transition estimation

Other methods:
(Elkan & Noto, 2008; Sukhbaatar et al., 2015)

$$p(y^* = j | \tilde{y} = i) \approx \mathbb{E}[p(\hat{y} = j | \boldsymbol{x} \in$$

Takeaway

CL methods

↓

Robust statistics to estimate with counts based on rank

↓

Robust to imperfect probabilities from model

Confident Learning: $p(y^* = j | \tilde{y} = i) = \dfrac{p(y^* = j, \tilde{y} = i)}{p(\tilde{y} = i)} \approx \dfrac{\boxed{\text{count}}(y^* = j, \tilde{y} = i)}{\boxed{\text{count}}(\tilde{y} = i)}$

Enables CL to disambiguate aleatoric (label noise) from epistemic (model noise) ← **Robust statistic w/ counts + rank (1 step removed erroneous probs)**

$$p(y^* = j | \tilde{y} = i) \approx \boxed{\frac{|(\boldsymbol{x} \in \boldsymbol{X}_i \ \text{ with } \boxed{\text{large}} \ p(y = j; \boldsymbol{x})|}{|\boldsymbol{x} \in \boldsymbol{X}_i|}}$$

e.g. Median of Means

What do “ideal” (non-erroneous) predicted probs look like?

$$\underbrace{\boldsymbol{x} \in \mathbf{X}_{\tilde{y}=i, y^*=j}}_{\text{error-free predicted probs}} = \underbrace{p(\tilde{y} = i | y^* = j)}_{\text{noise rate}}$$

Equipped with this understanding of ideal probabilities

And the prune, count, and rank principles of CL

We can see the intuition for our theorem (exact error finding with noisy probs)

Theorem Intuition

Let "ideal" $\hat{p}$ = 0.9.

$$\hat{\boldsymbol{X}}_{\tilde{y}=i,y^*=j} = \{\boldsymbol{x} \in \boldsymbol{X}_{\tilde{y}=i} : \; \hat{p}(\tilde{y} = j; \boldsymbol{x}, \boldsymbol{\theta}) \geq 0.6\}$$

The model can be up to (0.9 - 0.6) / 0.9 = 33% wrong in its estimate of $\hat{p}$

And $\boldsymbol{x}$ will be correctly counted.

Does this result still hold for systematic miscalibration (common in neural networks)?

Guo, Pleiss, Sun, & Weinberger (2017) "On Calibration of Modern Neural Networks." ICML

Final Intuition: Robustness to miscalibration

$$\boldsymbol{C}_{\tilde{y}=i, y^*=j} := \boxed{|\{\boldsymbol{x} : \boldsymbol{x} \in \boldsymbol{X}_{\tilde{y}=i},\ \hat{p}(\tilde{y}=j|\boldsymbol{x}) \geq t_j\}|}$$

Exactly finds label errors for "ideal" probabilities (Ch. 2, Thm 1, in thesis)

$$t_j = \frac{1}{|X_{\tilde{y}=j}|} \sum_{\boldsymbol{x} \in \boldsymbol{X}_{\tilde{y}=j}} \boxed{\hat{p}(\tilde{y}=j; \boldsymbol{x}, \boldsymbol{\theta})}$$

But neural networks have been shown (Guo et al., 2017) to be over-confident for some classes:

$$t_j^{\epsilon_j} = \frac{1}{|X_{\tilde{y}=j}|} \sum_{\boldsymbol{x} \in \boldsymbol{X}_{\tilde{y}=j}} \boxed{\hat{p}(\tilde{y}=j; \boldsymbol{x}, \boldsymbol{\theta}) + \epsilon_j}$$

$$= \boxed{t_j + \epsilon_j}$$

What happens to $\boldsymbol{C}_{\tilde{y}=i, y^*=j}$?

$$\boldsymbol{C}^{\epsilon_j}_{\tilde{y}=i, y^*=j} = |\{\boldsymbol{x} : \boldsymbol{x} \in \boldsymbol{X}_{\tilde{y}=i},\ \hat{p}(\tilde{y}=j|\boldsymbol{x}) + \cancel{\epsilon_j} \geq \boxed{t_j + \cancel{\epsilon_j}}\}|$$

exactly finds errors

Enough intuition, let's see some results

First we'll look at examples for dataset curation in ImageNet.

Then we'll look at CL with various distributions/models

Then we'll look at failure modes

Finally, we're ready for part 3: "label errors"

Organization for this part of the talk:

✓ 1. What is confident learning?
✓ 2. Situate confident learning
   a. Noise + related work
✓ 3. How does CL work? (methods)
✓ 4. Comparison with other methods
✓ 5. Why does CL work? (theory)
   a. Intuitions
   b. Principles
6. Examples + Dataset Curation

Dataset Curation: ImageNet Train Set

Rank	$\tilde{y}$ name	y^* name	$C(\tilde{y}, y^*)$
1	projectile	$\xleftarrow{\text{is a}}$ missile	645
2	tub	$\xleftarrow{\text{is a}}$ bathtub	539
3	breastplate	$\xleftarrow{\text{is a}}$ cuirass	476
4	green_lizard	chameleon	437
5	chameleon	green_lizard	435
6	missile	projectile	433
7	maillot	maillot	417
8	horned_viper	sidewinder	416
9	corn	ear	410
10	keyboard	space_bar	406

The largest off-diagonals of $C(\tilde{y}, y^*)$ reveal ontological issues.

Note the **(is a)** and **(has a)** relationships

Does this also work for val/test sets?

Dataset Curation: ImageNet Train Set

Rank	$\tilde{y}$ name		y^* name	$\tilde{y}$ nid	y^* nid	$C(\tilde{y}, y^*)$
1	projectile	← is a	missile	n04008		
2	tub	← is a	bathtub	n04493 $C(\tilde{y}, y^*)$	n02808440	539
3	breastplate	← is a	cuirass	n02895154	n03146219	476
4	green_lizard		chameleon	n01693334	n01682714	437
5	chameleon		green_lizard	n01682714	n01693334	435
6	missile		projectile	n03773504	n04008634	433
7	maillot		maillot	n03710637	n03710721	417
8	horned_viper		sidewinder	n01753488	n01756291	416
9	corn		ear	n12144580	n13133613	410
10	keyboard		space_bar	n04505470	n04264628	406

Same id for two different classes!

Does this also work for val/test sets?

Dataset Curation: ImageNet Val Set

```
26 | n02979186 cassette_player | n04392985 tape_player
23 | n03773504 missile | n04008634 projectile
23 | n03642806 laptop | n03832673 notebook
23 | n02808440 bathtub | n04493381 tub
23 | n13133613 ear | n12144580 corn
22 | n03710721 maillot | n03710637 maillot
22 | n01682714 American_chameleon | n01693334 green_lizard
21 | n02895154 breastplate | n03146219 cuirass
20 | n02412080 ram | n02415577 bighorn
19 | n04008634 projectile | n03773504 missile
18 | n01753488 horned_viper | n01756291 sidewinder
18 | n02107908 Appenzeller | n02107574 Greater_Swiss_Mountain_dog
18 | n12144580 corn | n13133613 ear
17 | n03146219 cuirass | n02895154 breastplate
17 | n02113624 toy_poodle | n02113712 miniature_poodle
16 | n03710637 maillot | n03710721 maillot
```

There are indistinguishable examples in these classes

Appenzeller Sennenhund

Dog breed

The Appenzeller Sennenhund is a medium-size breed of dog, one of the four regional breeds of Sennenhund-type dogs from the Swiss Alps. The name Sennenhund refers to people called Senn, herders in the Appenzell region of Switzerland. Wikipedia

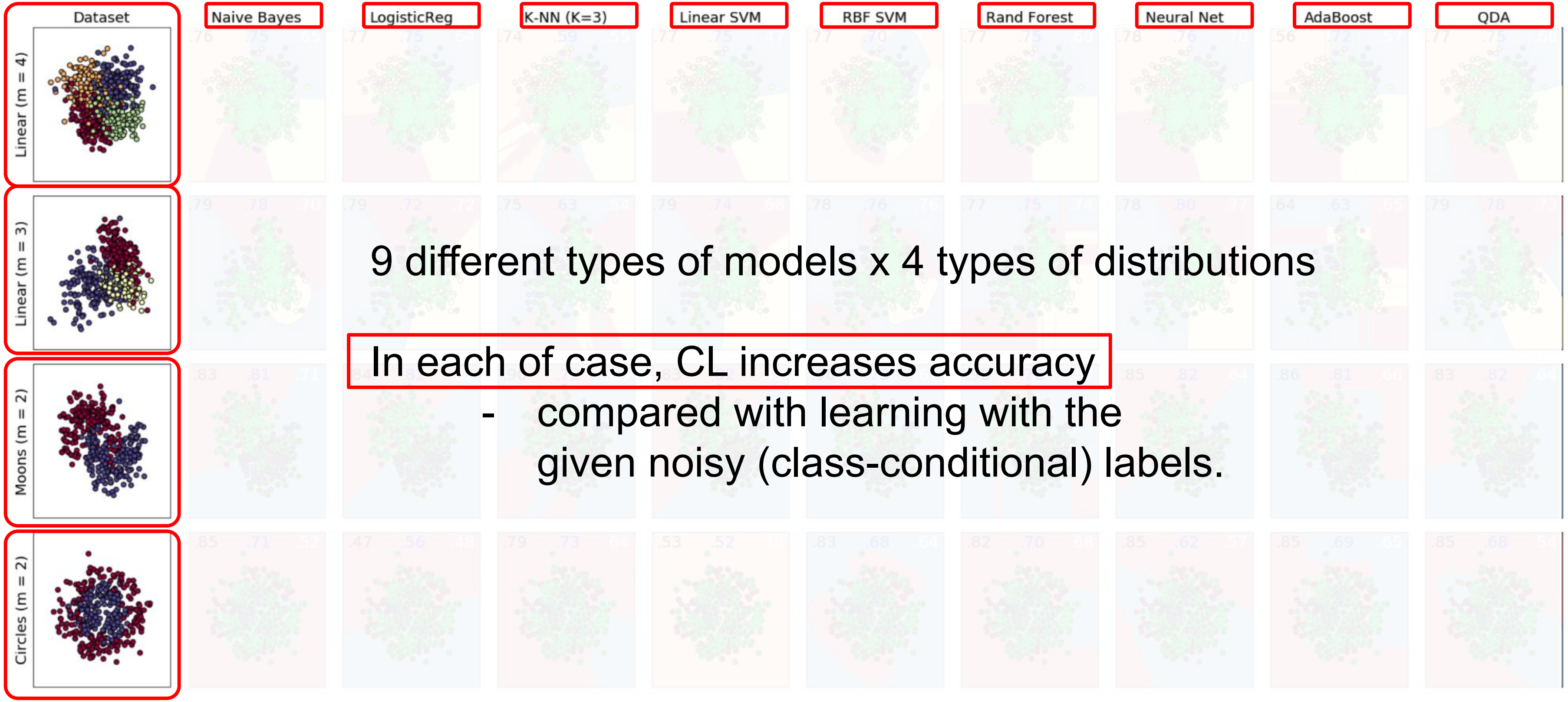
Greater Swiss Mountain Dog

Dog breed

The Greater Swiss Mountain Dog is a dog breed which was developed in the Swiss Alps. The name Sennenhund refers to people called Senn or Senner, dairymen and herders in the Swiss Alps. Wikipedia

Can you spot the difference?

CL is model-agnostic



9 different types of models x 4 types of distributions

In each of case, CL increases accuracy

- compared with learning with the given noisy (class-conditional) labels.

Failure Modes (when does CL fail?)

When the error in $\hat{p}(\tilde{y}{=}i; \boldsymbol{x}, \boldsymbol{\theta})$ exceeds the threshold margins.

When might this happen?

ImageNet given label:
sewing machine

We guessed: **manhole cover**

MTurk consensus: **Neither sewing machine nor manhole cover**

ID: 00001127

CIFAR-10 given label:
airplane

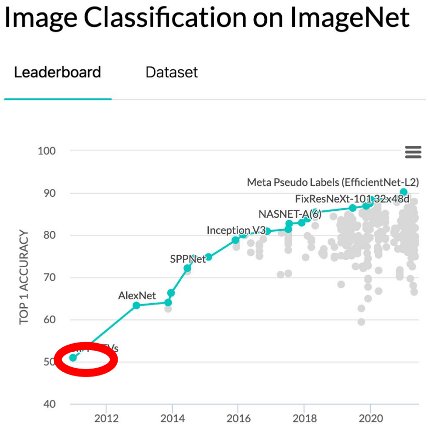
We guessed: **automobile**

MTurk consensus: **Neither airplane nor automobile**

ID: 2532

70%			
0	0.2	0.4	0.6
31.5	39.3	33.7	30.6
33.7	40.7	35.1	31.4
32.4	**41.8**	34.4	34.5
41.1	41.7	39.0	32.9
41.0	**41.8**	**39.1**	**36.4**

Acc. of CL-based methods for 70% noise for various settings.



(really) hard examples

too much (70+%) noise

inappropriate model

Steps to Confident Learning for Machines and Humans

1 — 2 — 3 — 4

Precursors to CL

Machine learning for human learning requires dealing with real-world, noisy labels

Northcutt, Ho, & Chuang (C&E, 2016)
Northcutt, Wu, & Chuang (UAI, 2017)
Northcutt, Leon, & Chen (L@S, 2017)
Corrigan-Gibbs, Gupta, Northcutt, Cutrell, & Thies (TOCHI 2015, CHI 2016)

Confident Learning

We develop a principled framework of theory and algorithms for quantifying, finding, and learning with label noise in datasets.

https://github.com/cgnorthcutt/cleanlab

Northcutt, Jiang, & Chuang (JAIR, 2021)

Label Errors in ML Datasets

We find tens of thousands (3.4%) of label errors in the most commonly benchmarked ML test sets.

labelerrors.com

Northcutt, Athalye, & Lin (NeurIPS Workshop on Dataset Curation and Security, 2020)

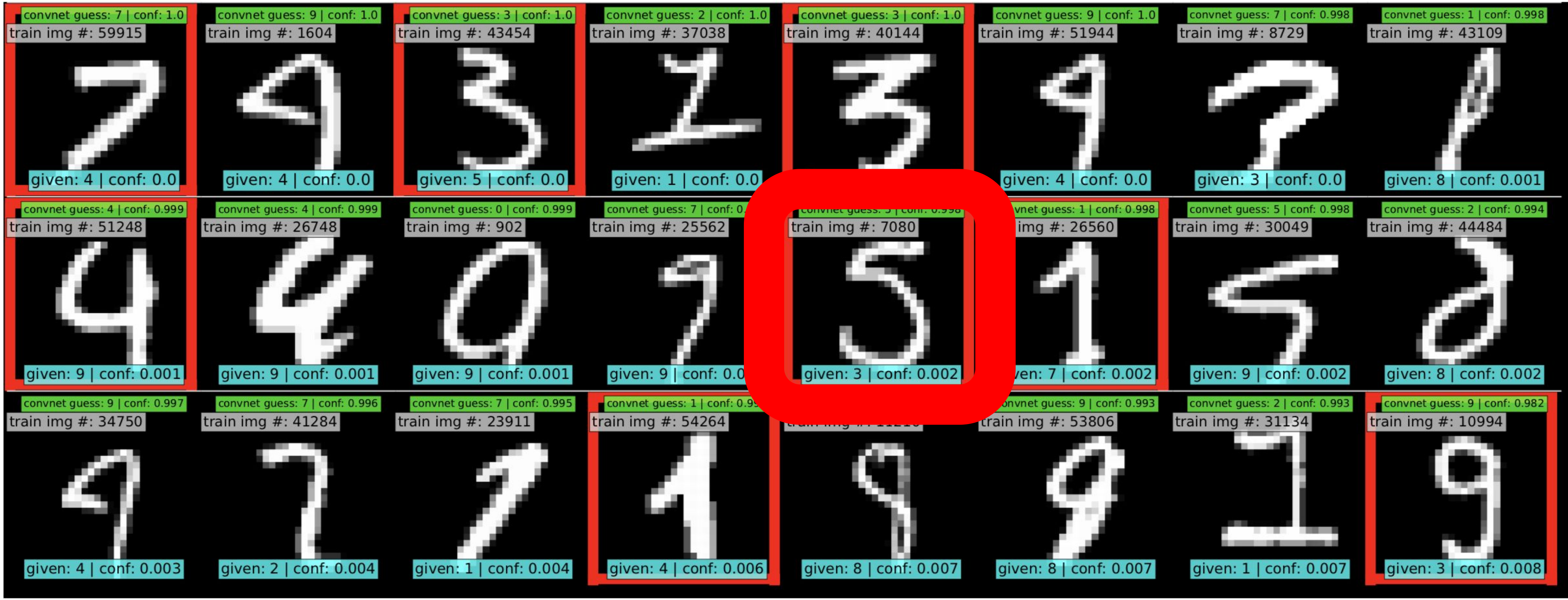
Implications for ML Practitioners

We study whether practitioners are unknowingly benchmarking the progress of ML based on erroneous test sets? How noisy is too noisy?

https://github.com/cgnorthcutt/label-errors

Northcutt, Athalye, & Mueller (ICLR RobustML Workshop, 2021) (ICLR WeaSuL Workshop, 2021)

A. MNIST is assumed error-free in tens of thousands of papers




“To conclude my talk, I will show that our method finds one label error in Yann’s MNIST dataset!”
Jun 17, 2016
Fri, 2:19 PM GMT-04:00
- Hinton (@Facebook AI Research, NYC)
MNIST Label: 3

Motivated by the surprising errors in MNIST, we found label errors in 10 of the most commonly used datasets in Machine Learning

labelerrors.com

Demo (click the link above)

3.4% of labels in popular ML test sets are erroneous

Dataset	Test Set Errors				
	CL guessed	MTurk checked	validated	estimated	% error
MNIST	100	100 (100%)	15	-	0.15
CIFAR-10	275	275 (100%)	54	-	0.54
CIFAR-100	2235	2235 (100%)	585	-	5.85
Caltech-256	4,643	400 (8.6%)	65	754	2.46
ImageNet*	5,440	5,440 (100%)	2,916	-	5.83
QuickDraw	6,825,383	2,500 (0.04%)	1870	5,105,386	10.12
20news	93	93 (100%)	82	-	1.11
IMDB	1,310	1,310 (100%)	725	-	2.9
Amazon	533,249	1,000 (0.2%)	732	390,338	3.9
AudioSet	307	307 (100%)	275	-	1.35

Images →

Text →

Audio →

There are pervasive label errors in test sets, but what are the implications for ML?

Steps to Confident Learning for Machines and Humans

1 — 2 — 3 — 4

1. Precursors to CL

Machine learning for human learning requires dealing with real-world, noisy labels

Northcutt, Ho, & Chuang (C&E, 2016)
Northcutt, Wu, & Chuang (UAI, 2017)
Northcutt, Leon, & Chen (L@S, 2017)
Corrigan-Gibbs, Gupta, Northcutt, Cutrell, & Thies (TOCHI 2015, CHI 2016)

2. Confident Learning

We develop a principled framework of theory and algorithms for quantifying, finding, and learning with label noise in datasets.

https://github.com/cgnorthcutt/cleanlab

Northcutt, Jiang, & Chuang (JAIR, 2021)

3. Label Errors in ML Datasets

We find tens of thousands (3.4%) of label errors in the most commonly benchmarked ML test sets.

labelerrors.com

Northcutt, Athalye, & Lin (NeurIPS Workshop on Dataset Curation and Security, 2020)

4. Implications for ML Practitioners

We study whether practitioners are unknowingly benchmarking the progress of ML based on erroneous test sets? How noisy is too noisy?

https://github.com/cgnorthcutt/label-errors

Northcutt, Athalye, & Mueller (ICLR RobustML Workshop, 2021) (ICLR WeaSuL Workshop, 2021)

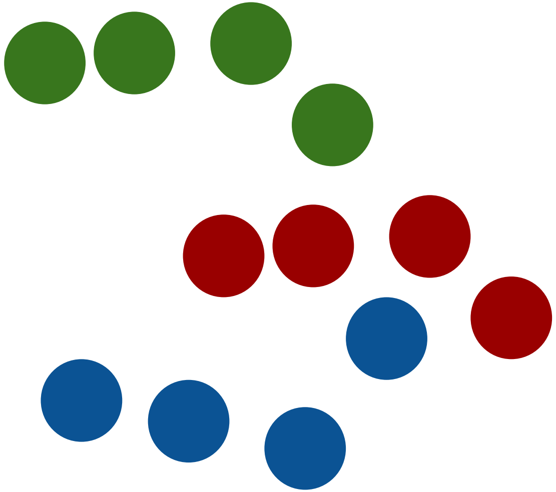
Are practitioners unknowingly benchmarking ML using erroneous test sets?

To answer this, let’s consider how ML traditionally creates test sets...

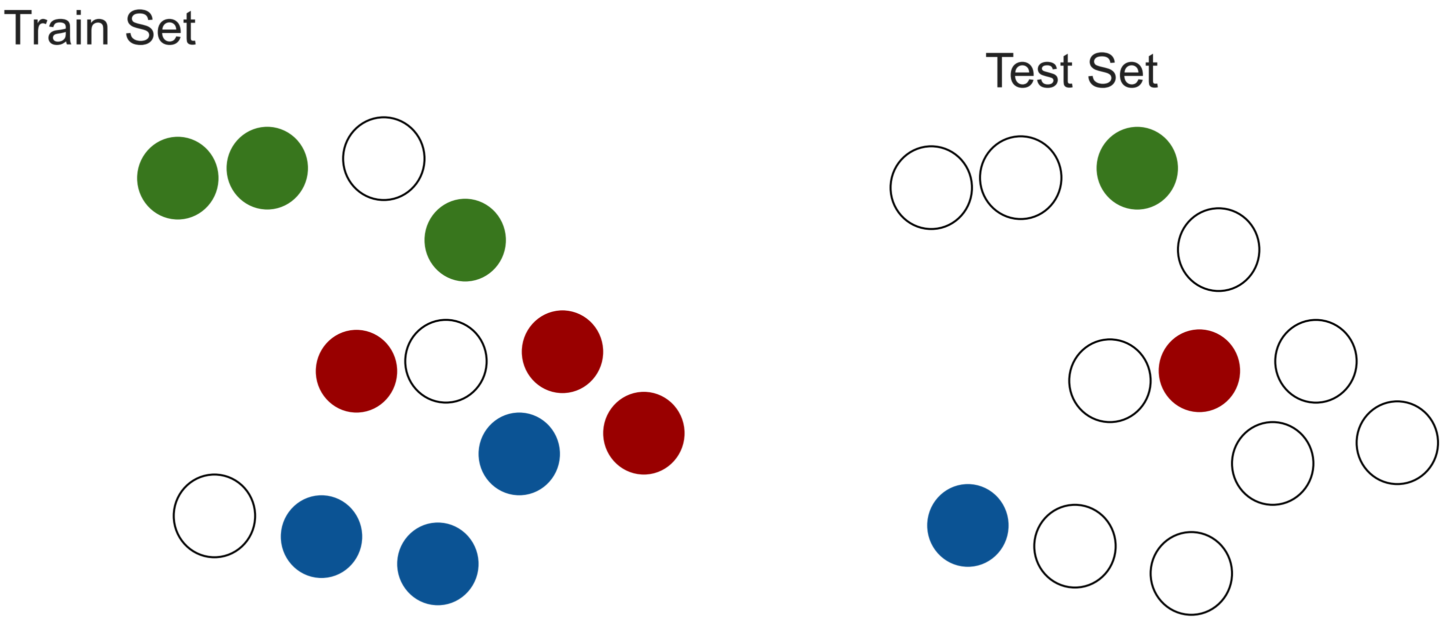
and why it can lead to problems for real-world deployed AI models.

A traditional view

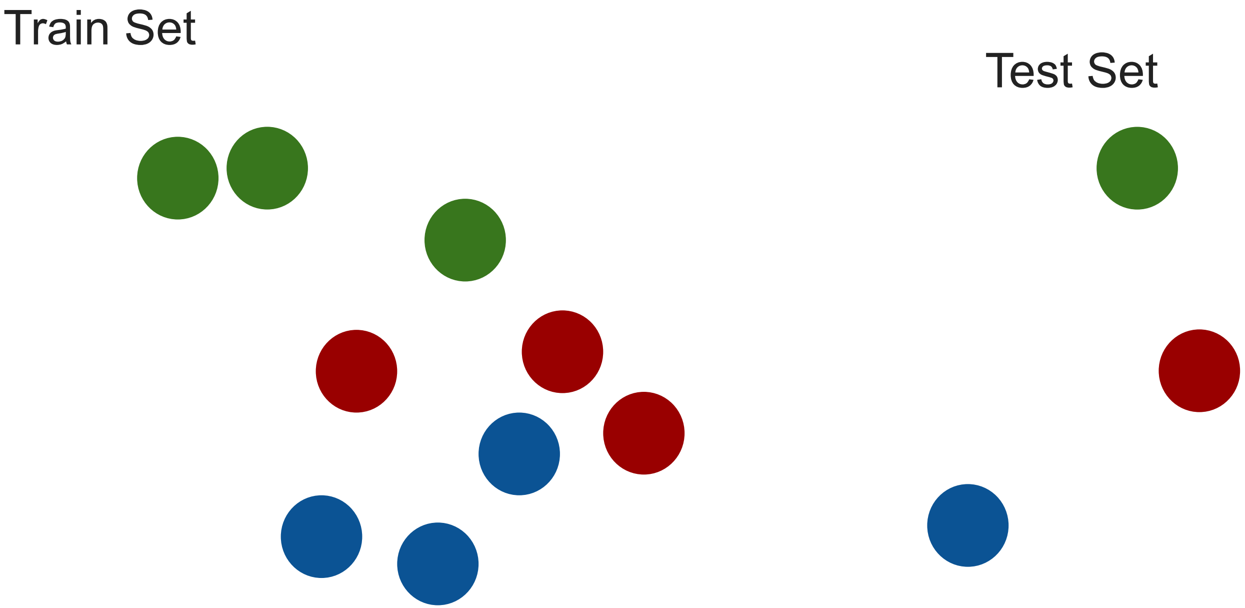
Data Set



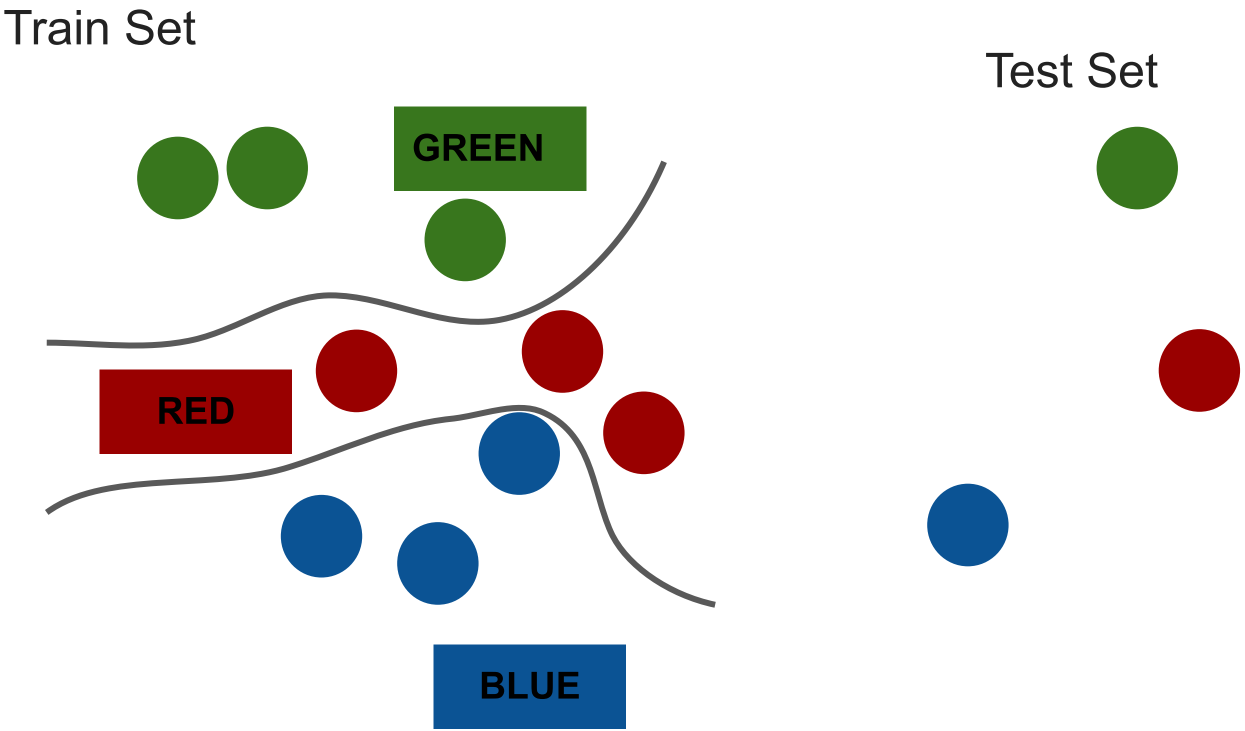
A traditional view



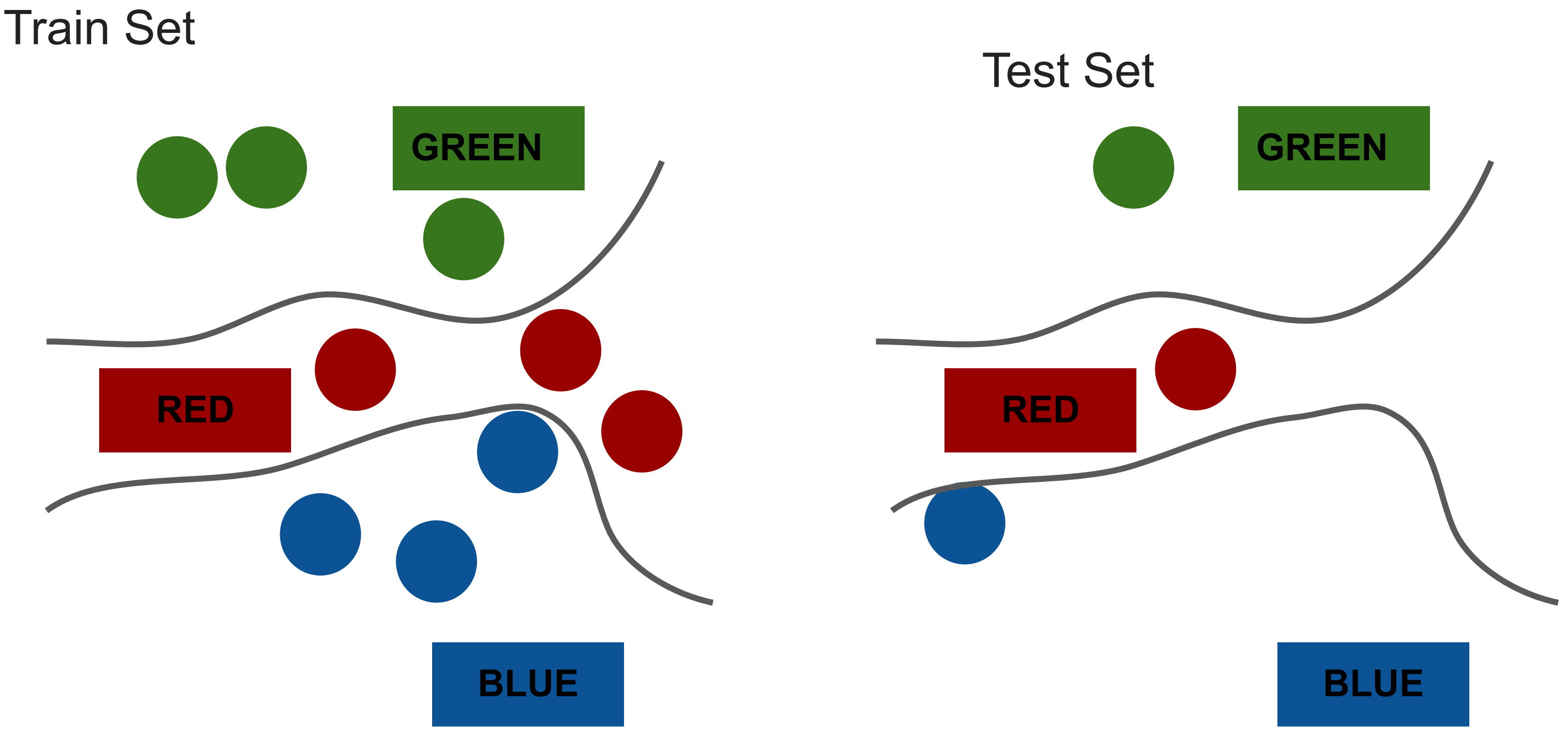
A traditional view



A traditional view

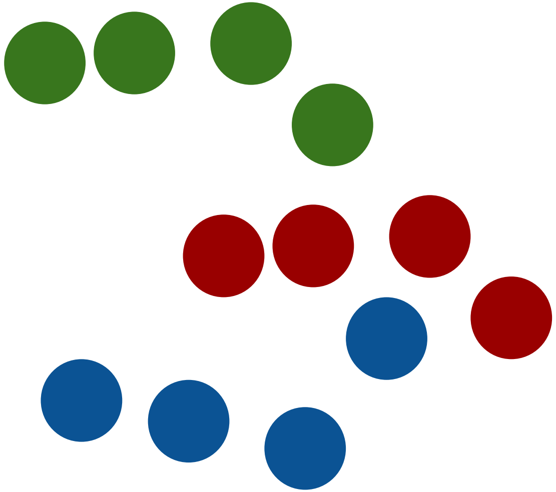


A traditional view



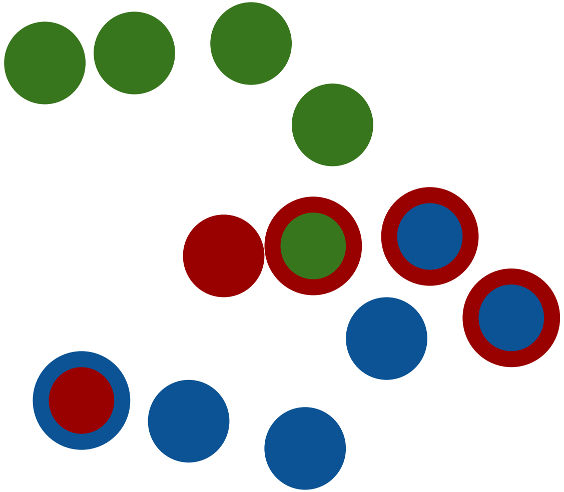
A real-world view

Data Set



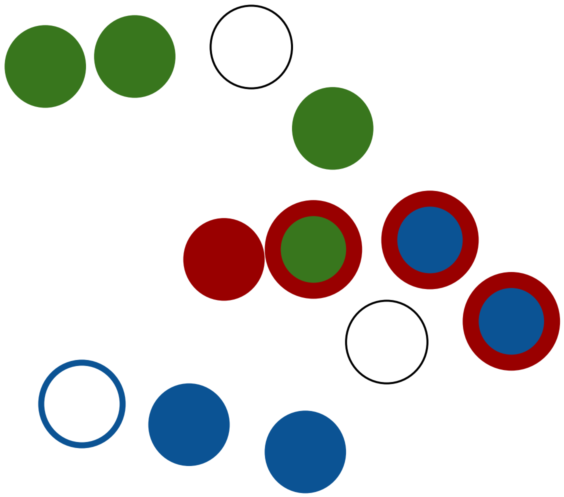
A real-world view

Data Set

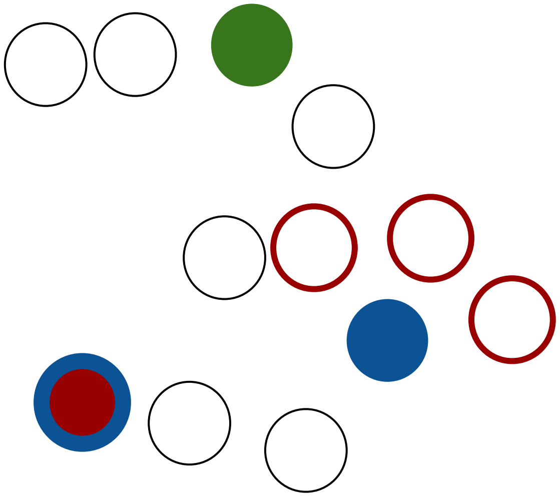


A real-world view

Train Set



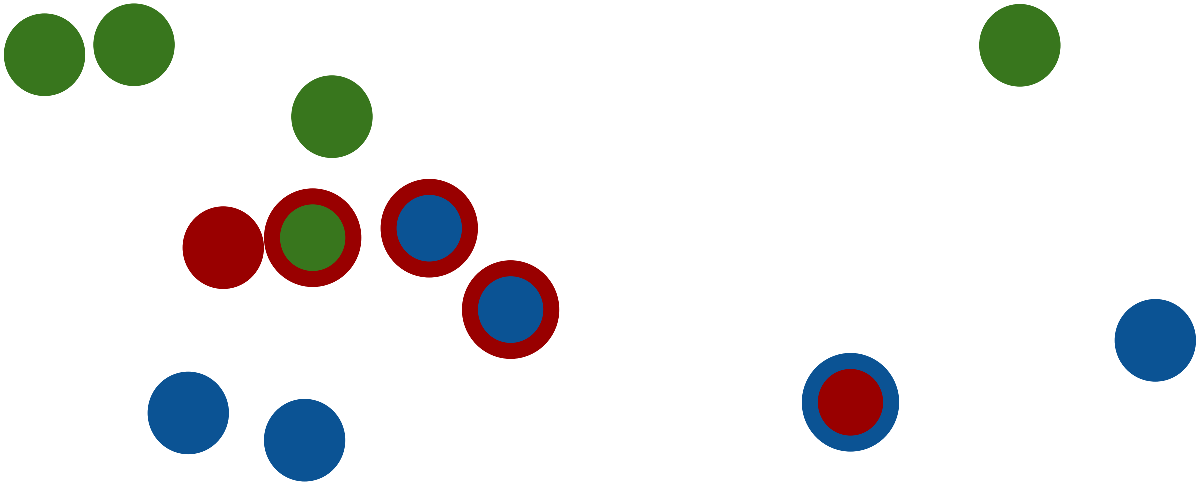
Test Set



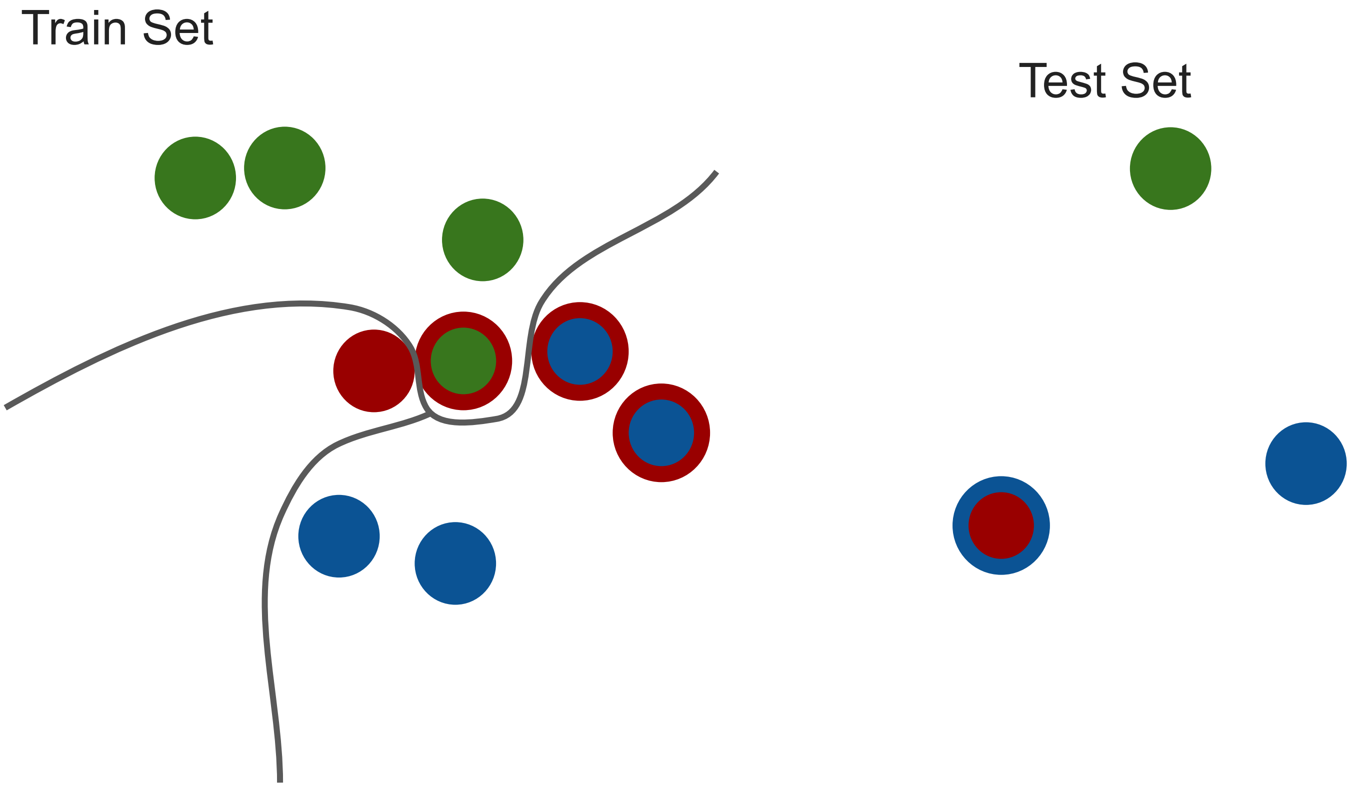
A real-world view

Train Set

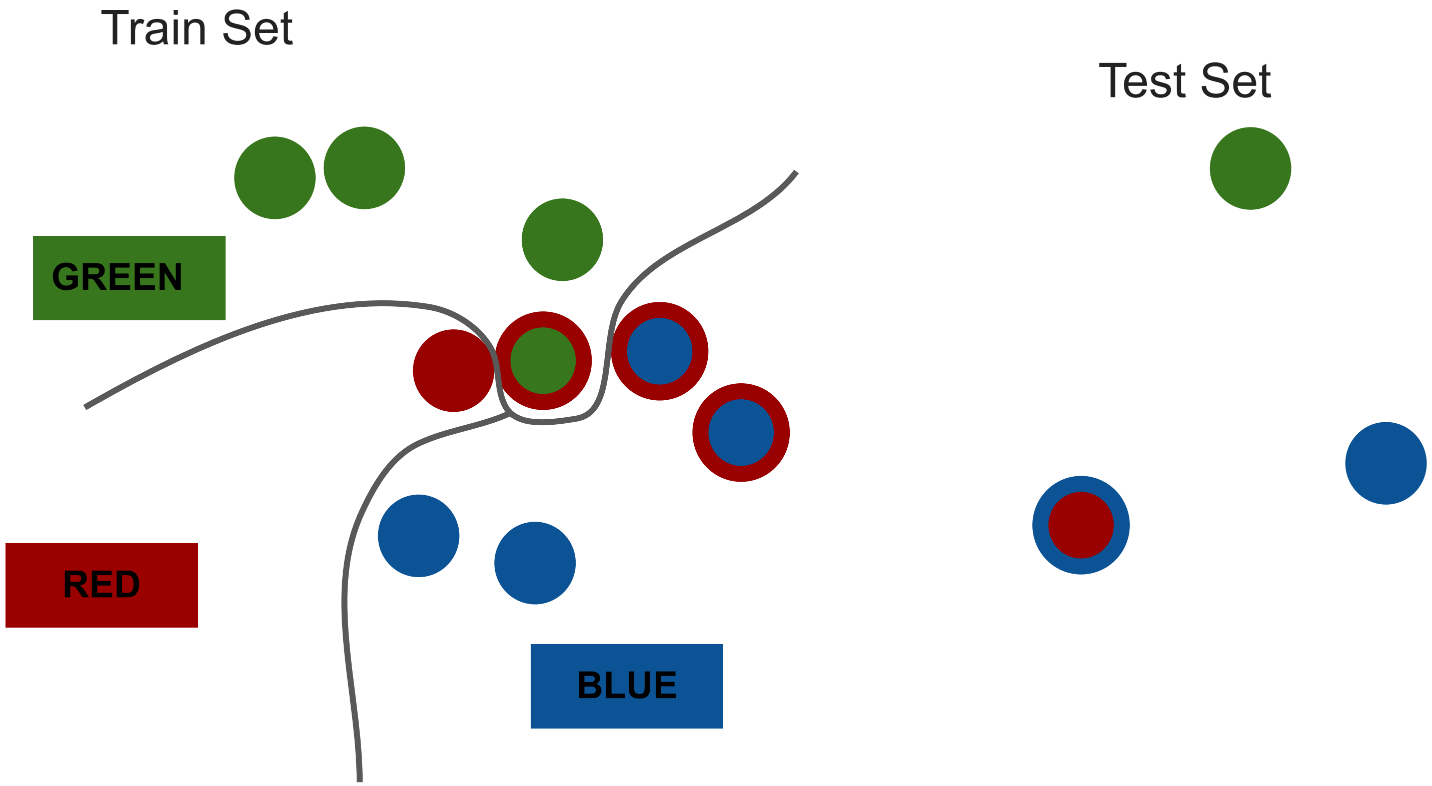
Test Set



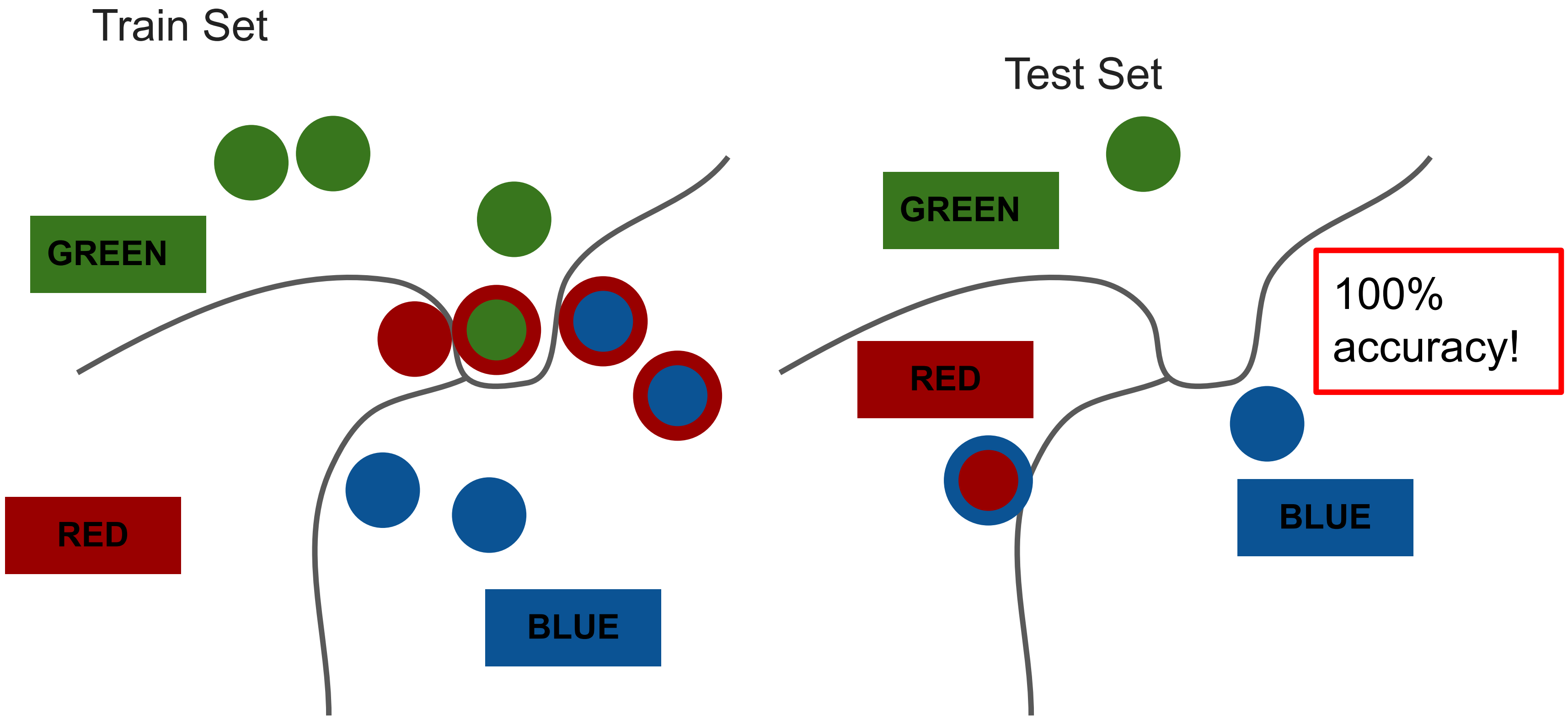
A real-world view



A real-world view

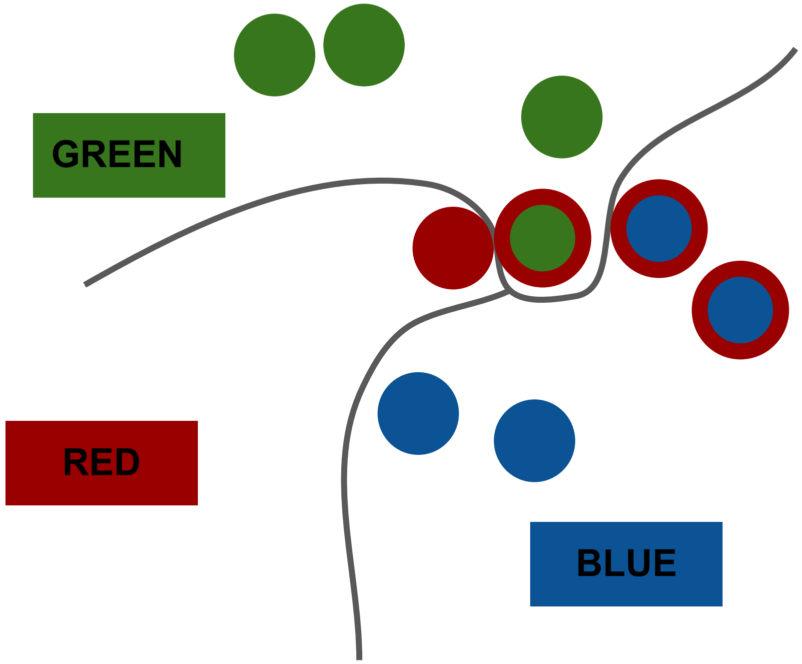


A real-world view



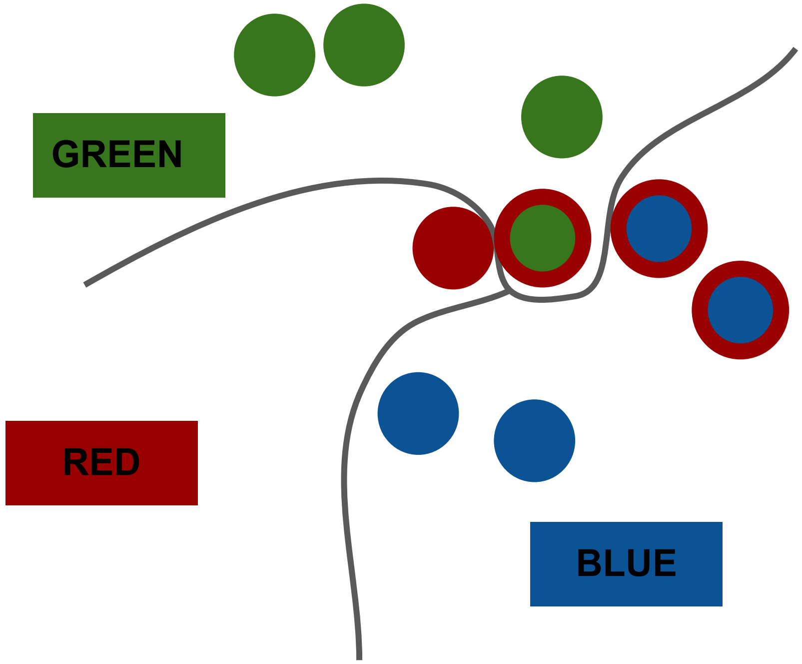
A real-world view

Trained Model with 100% test accuracy.

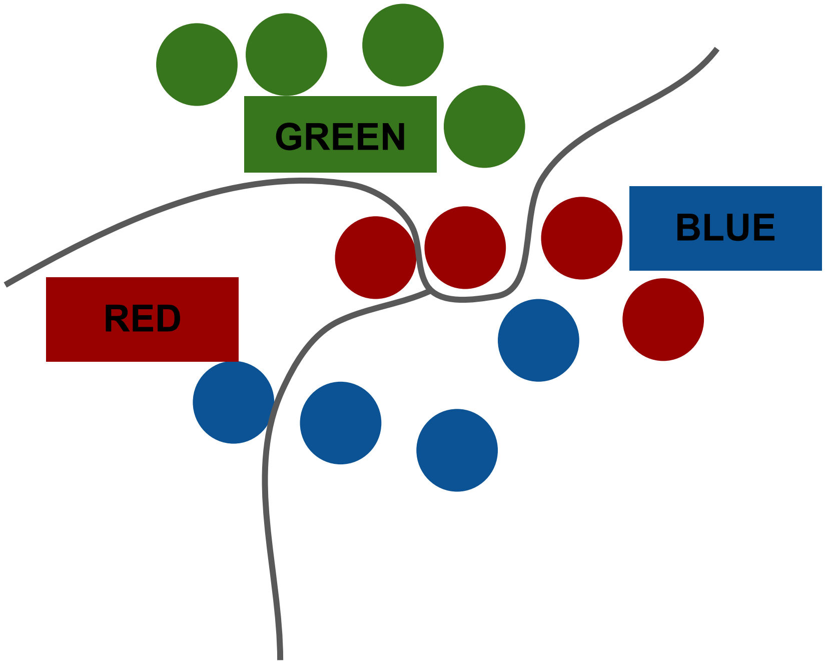


A real-world view

Trained Model with 100% test accuracy.



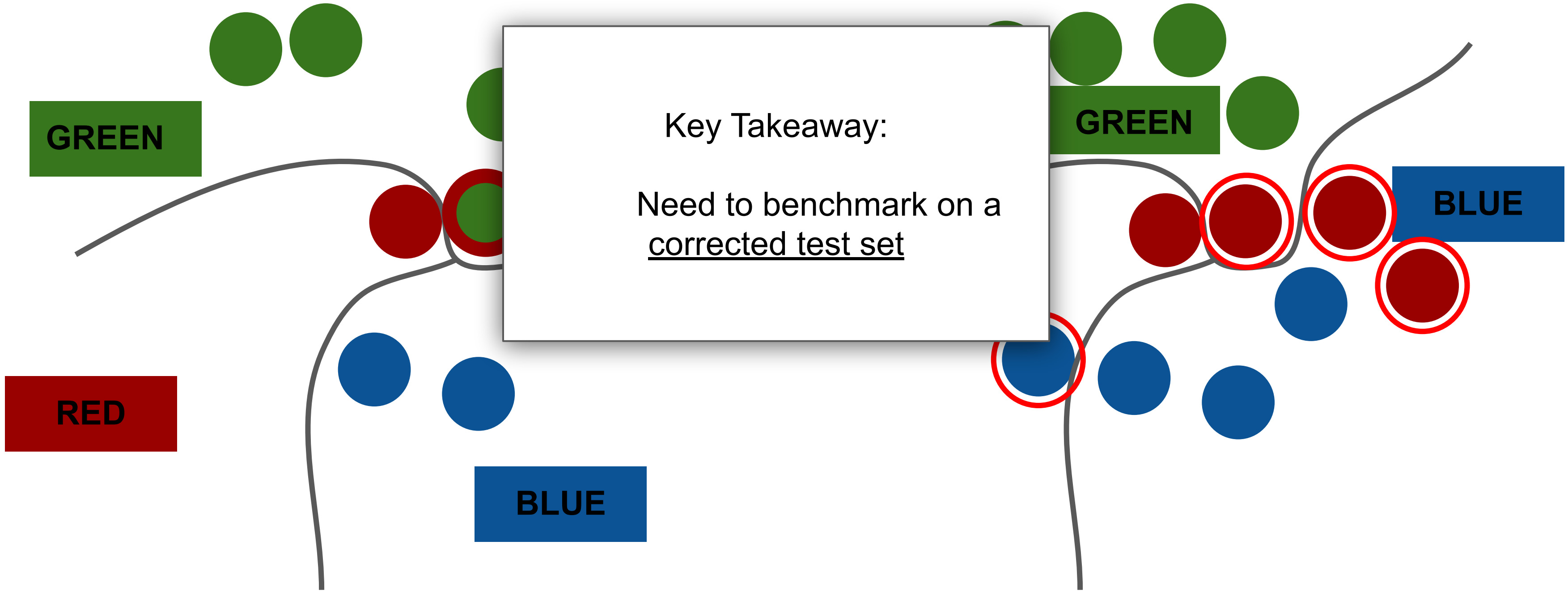
Real-world distribution
(the test set you actually care about)



A real-world view

Trained Model with 100% test accuracy.

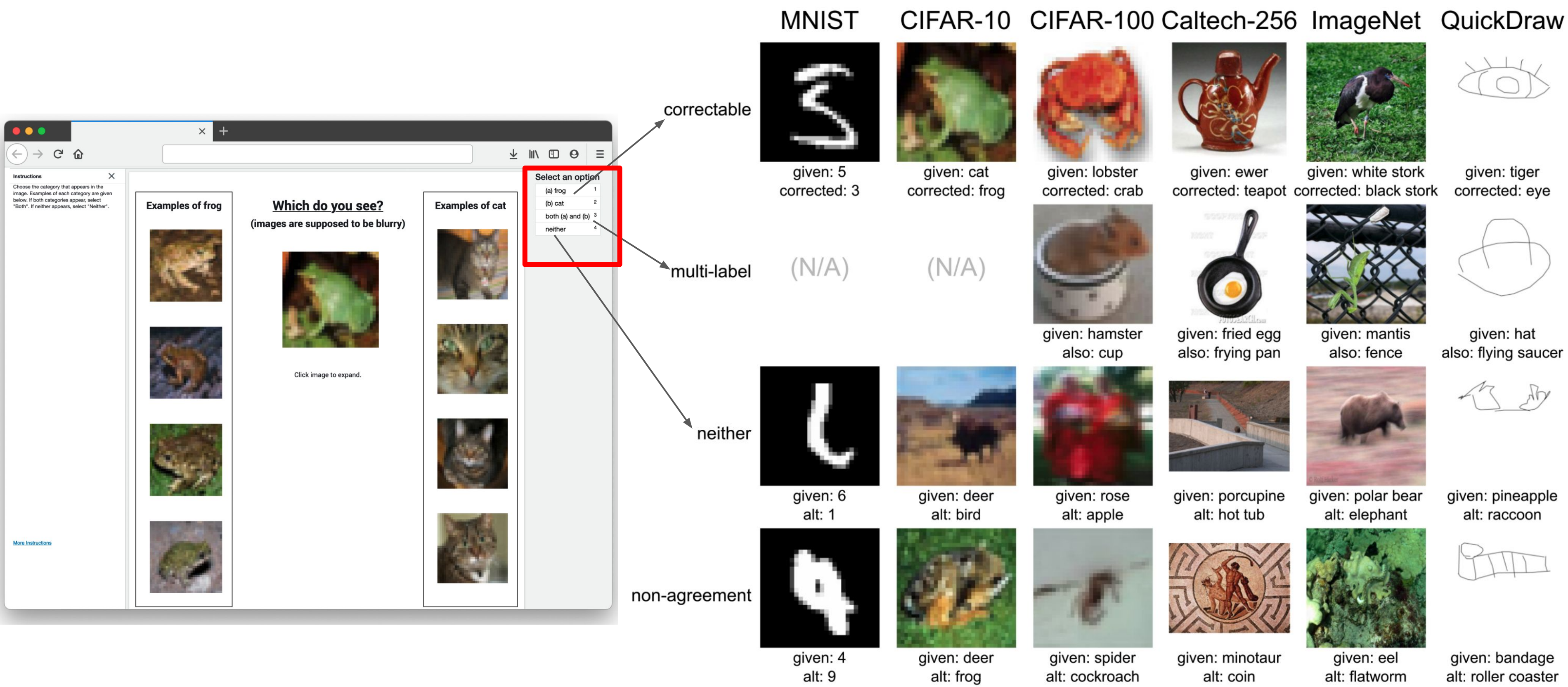
Real-world accuracy ~ 67%



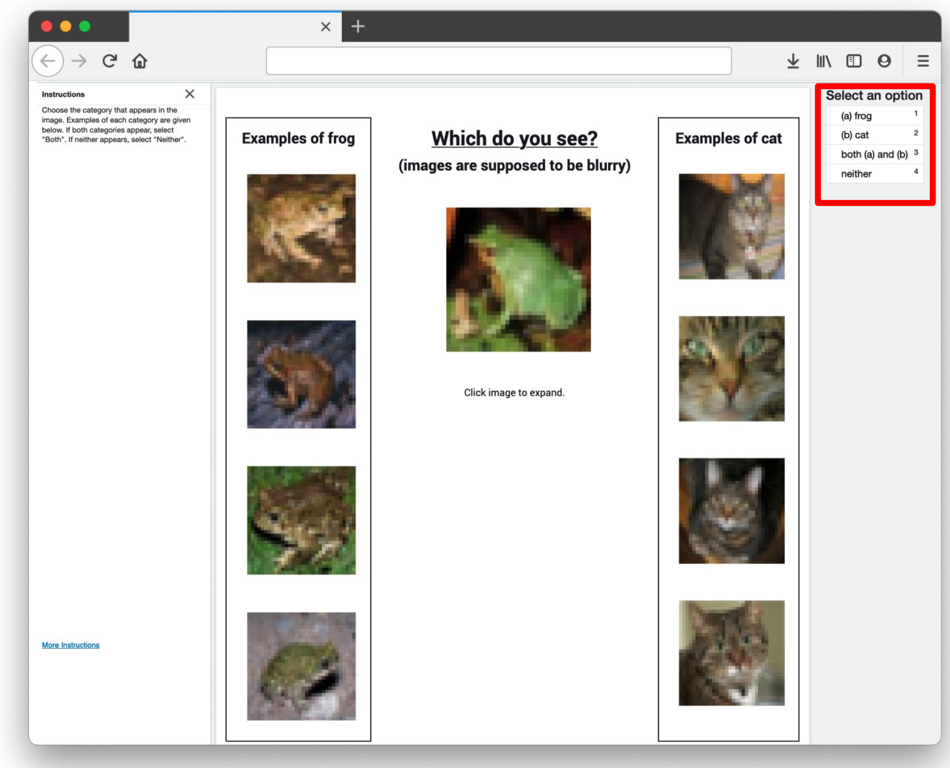
Key Takeaway:

Need to benchmark on a <u>corrected test set</u>

Correcting the test set



Correcting the test sets



Correct the label if a majority of reviewers:

- agree on our proposed label

Do nothing if a majority of reviewers:

- agree on the original label

Prune the example from the test set if the consensus is:

- Neither
- Both (multi-label)
- Reviewers cannot agree

Test Set Errors Categorization

Dataset	Test Set Errors Categorization correctable
MNIST	10
CIFAR-10	18
CIFAR-100	318
Caltech-256	22
ImageNet	1428
QuickDraw	1047
	22
	173
	302
AudioSet	-

Remember our two questions? Now we have the tools (corrected test sets) to answer Q2:

To support this claim, this talk addresses two questions

1. **In noisy, realistic settings, can we assemble a principled framework for quantifying, finding, and learning with label errors using a machine's confidence?**
   a. Traditionally, ML has focused on "Which model best learns with noisy labels?"
   b. In this talk I ask, "Which data is mislabeled?"

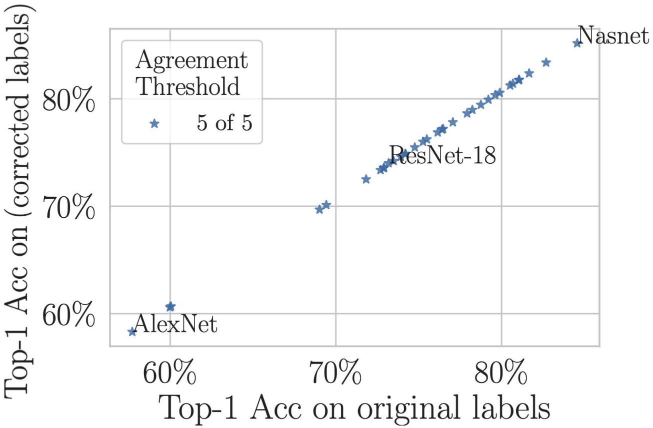
If Q1 works out, and there are label errors in datasets… does it matter? This leads us to Q2...

2. **Are we unknowingly benchmarking the progress of ML models, based on erroneous test sets? If so, can we quantify how much noise destabilizes benchmarks?**

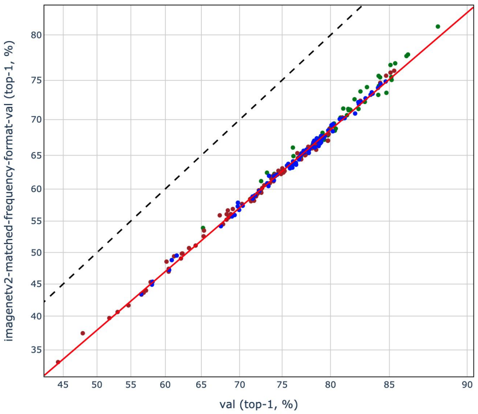
	Categorization
	correctable
	10
	18
	318
Caltech-256	22
ImageNet	1428
QuickDraw	1047
	22
	173
	302
AudioSet	-

Remember our two questions? Now we have the tools (corrected test sets) to answer Q2:

34 pre-trained black-box models on ImageNet

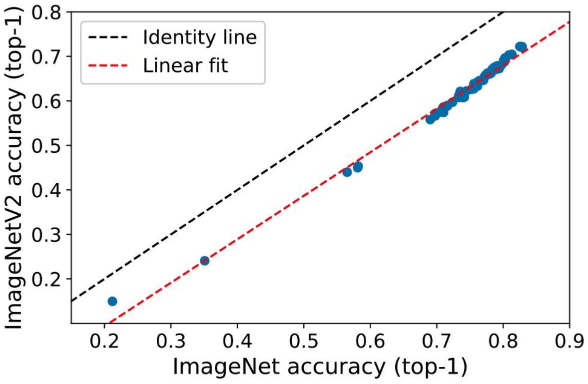


Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks
(Northcutt, Athalye, & Mueller 2021)



Measuring Robustness to Natural Distribution Shifts in Image Classification
(Taori, Dave, Shankar, Carlini, Recht, & Schmidt, 2021)

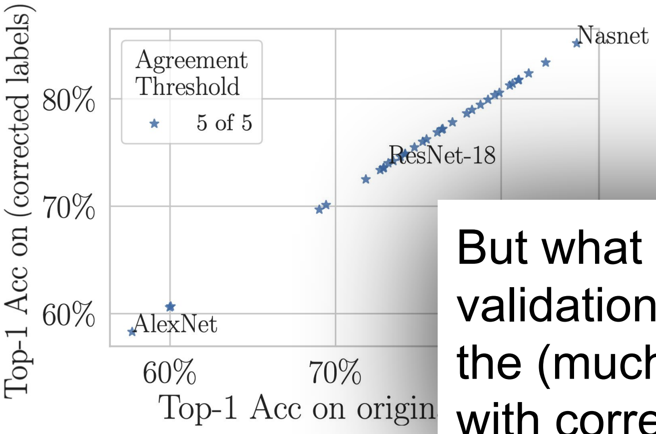
From ImageNet to Image Classification: Contextualizing Progress on Benchmarks
(Tsipras, Santurkar, Engstrom, Ilyas, Madry, 2020)

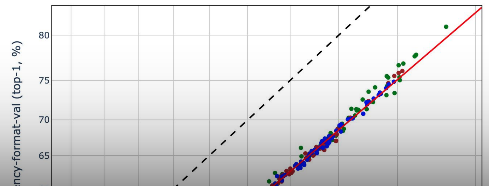


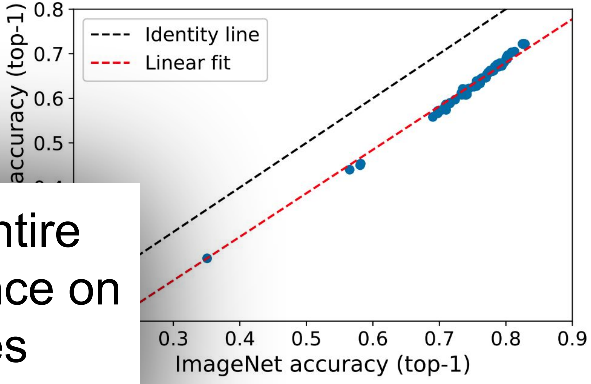
Do ImageNet classifiers generalize to ImageNet?
(Recht, Roelofs, Schmidt, & Shankar, 2019)

Why do classifier accuracies show linear trends under distribution shift?
(Mania & Sra, 2021)

34 pre-trained black-box models on ImageNet







But what if instead of looking at the entire validation set, we compare performance on the (much smaller) subset of examples with corrected labels?

Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks
(Northcutt, Athalye, & Mueller 2021)

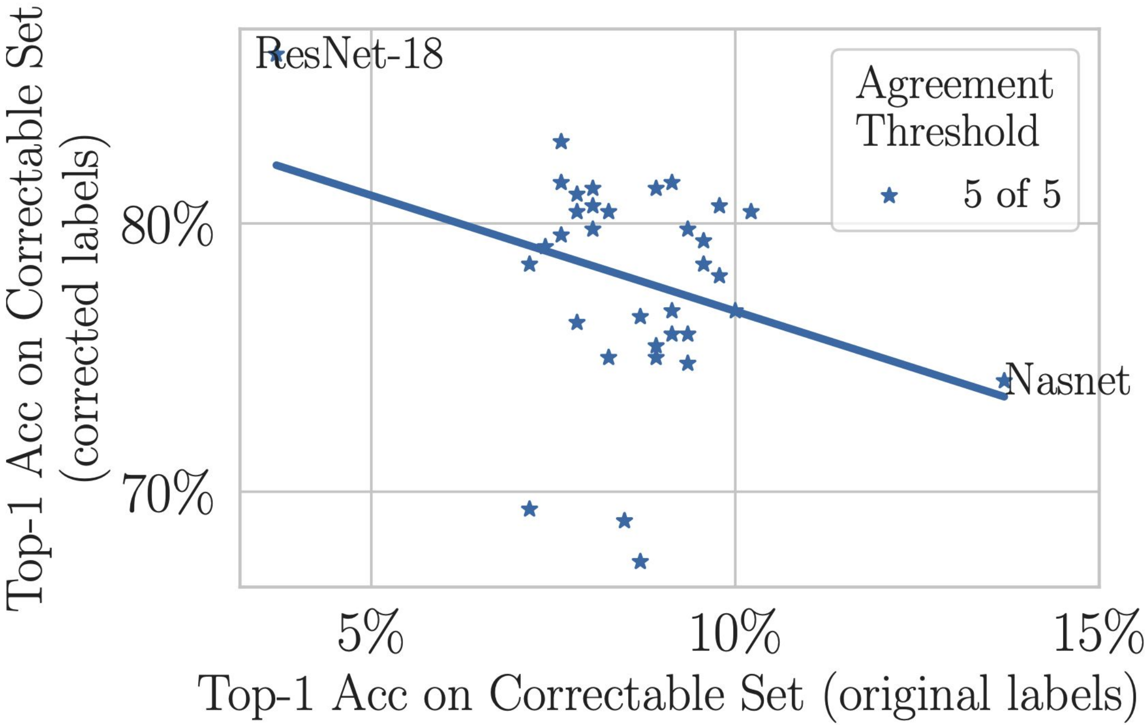
Measuring Robustness to Natural Distribution Shifts in Image Classification
(Taori, Dave, Shankar, Carlini, Recht, & Schmidt, 2021)

From ImageNet to Image Classification: Contextualizing Progress on Benchmarks
(Tsipras, Santurkar, Engstrom, Ilyas, Madry, 2020)

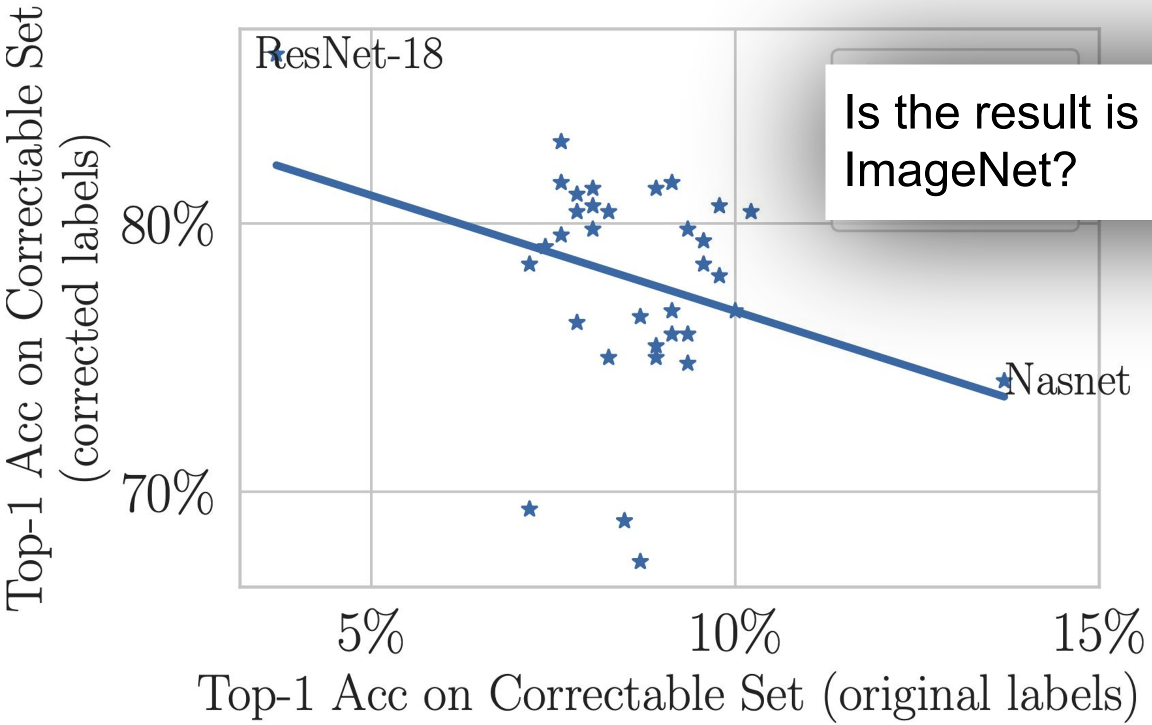
Do ImageNet classifiers generalize to ImageNet?
(Recht, Roelofs, Schmidt, & Shankar, 2019)

Why do classifier accuracies show linear trends under distribution shift?
(Mania & Sra, 2021)

34 pre-trained black-box models on ImageNet

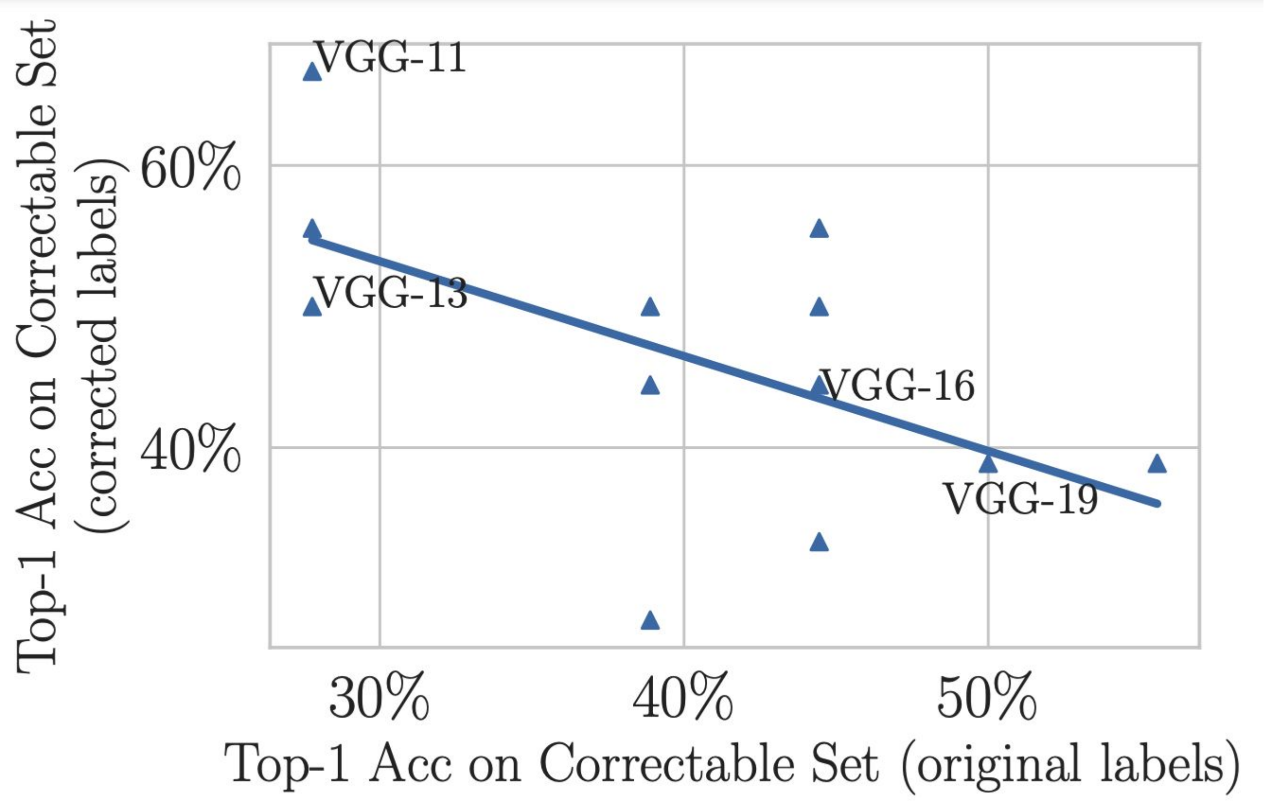


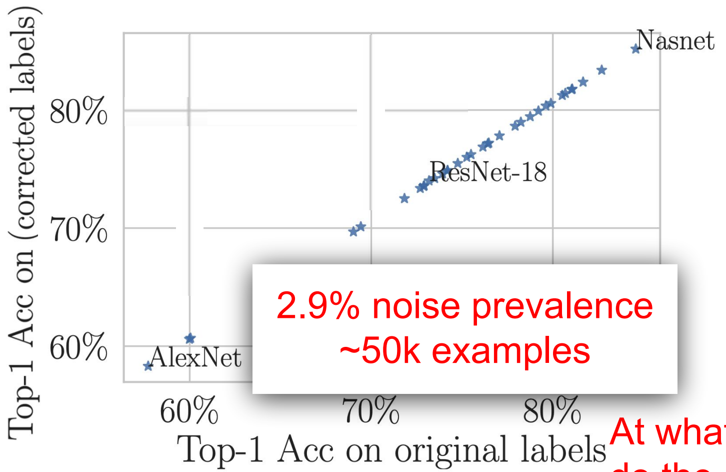
34 pre-trained black-box models on ImageNet



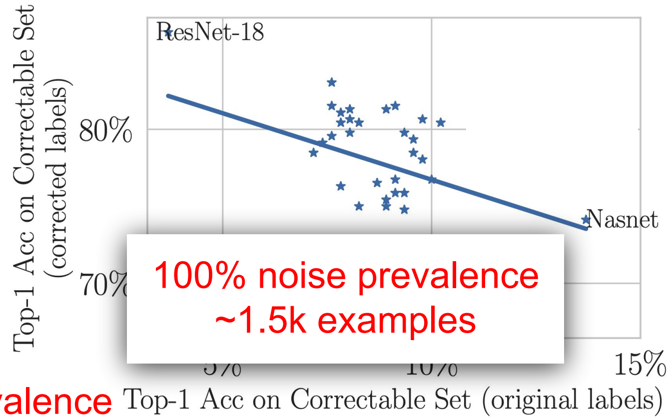
Is the result is specific to ImageNet?

The same finding, this time on CIFAR-10





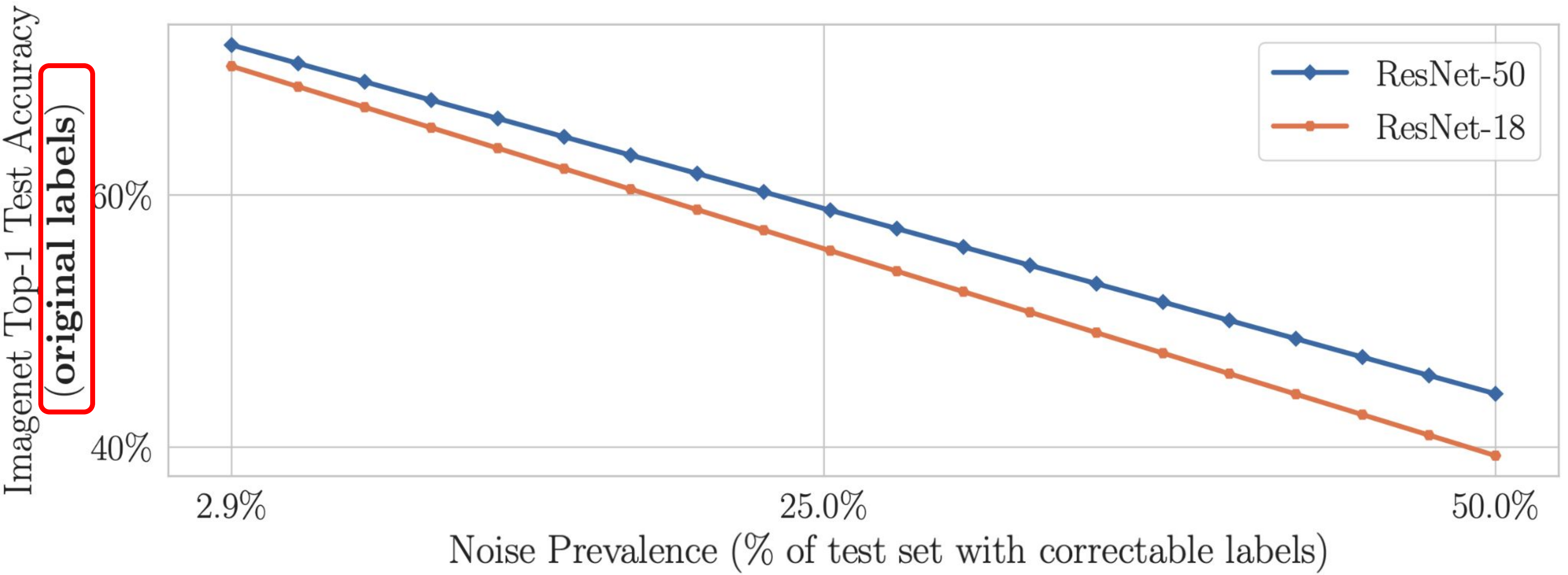
2.9% noise prevalence
~50k examples



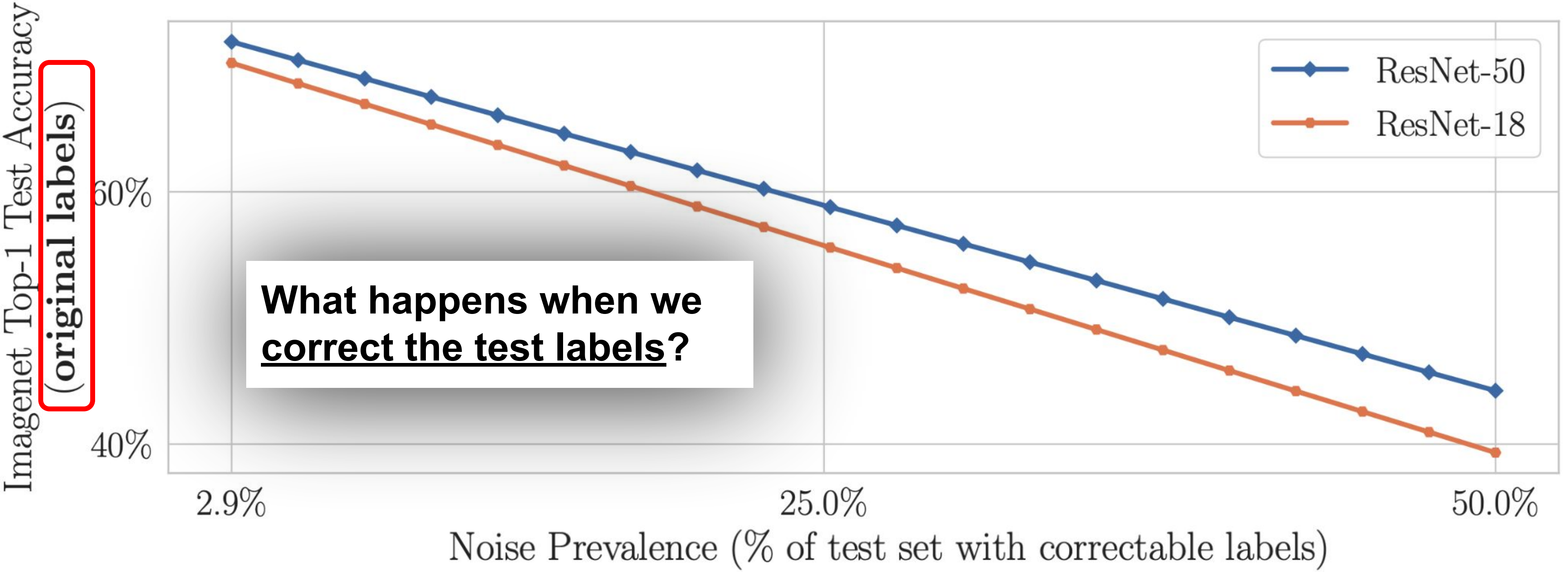
100% noise prevalence
~1.5k examples

At what noise prevalence do the rankings start to change?

Two pre-trained ImageNet models tested on original (noisy) labels

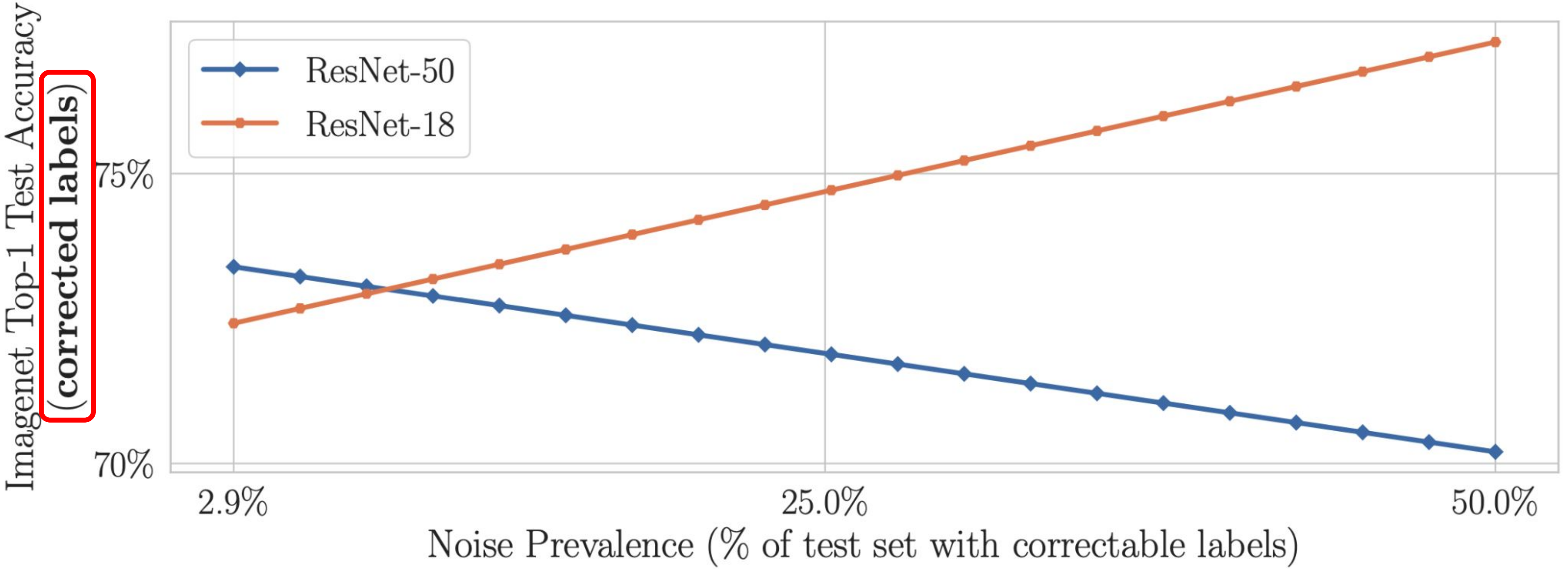


Two pre-trained ImageNet models tested on original (noisy) labels

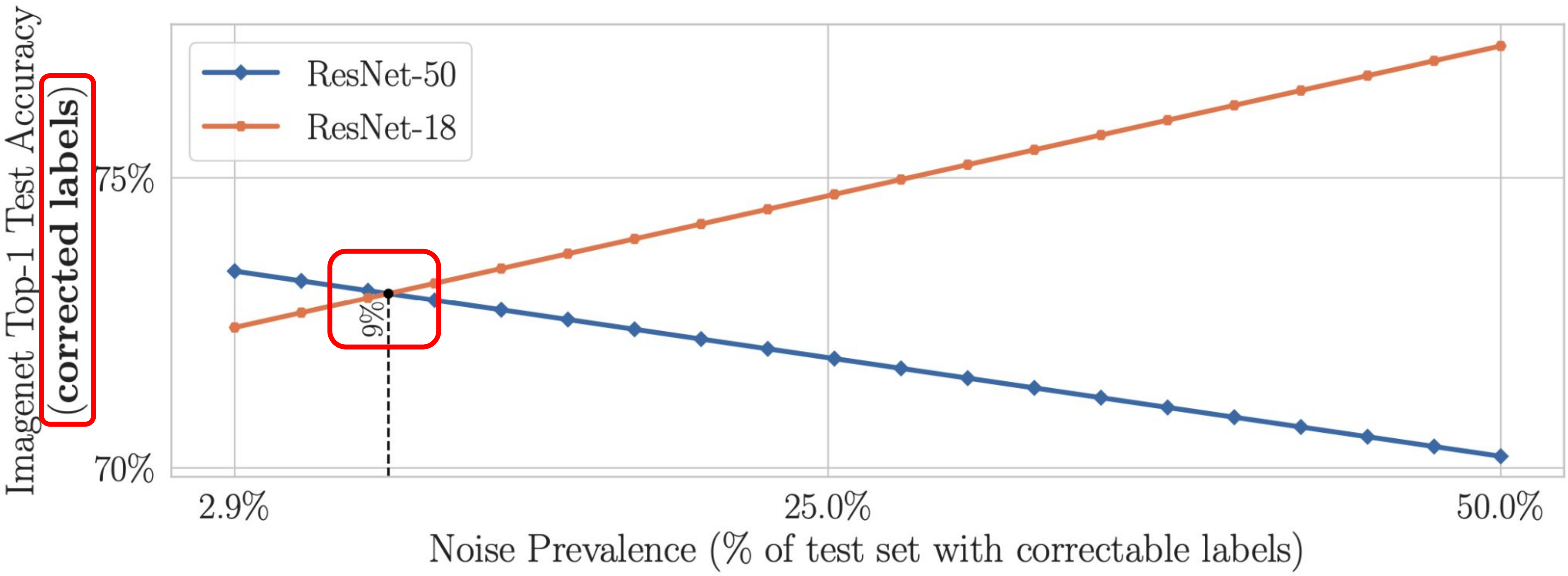


What happens when we <u>correct the test labels</u>?

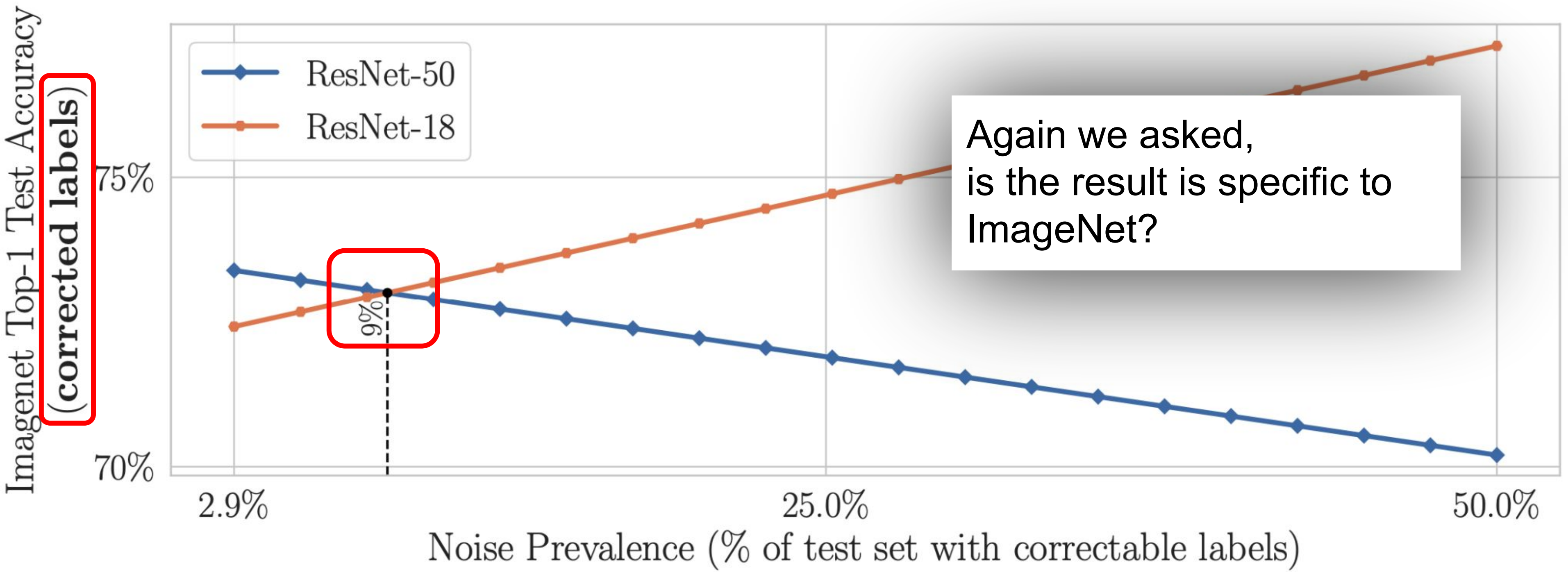
But when we correct the test set, benchmark rankings destabilize



But when we correct the test set, benchmark rankings destabilize

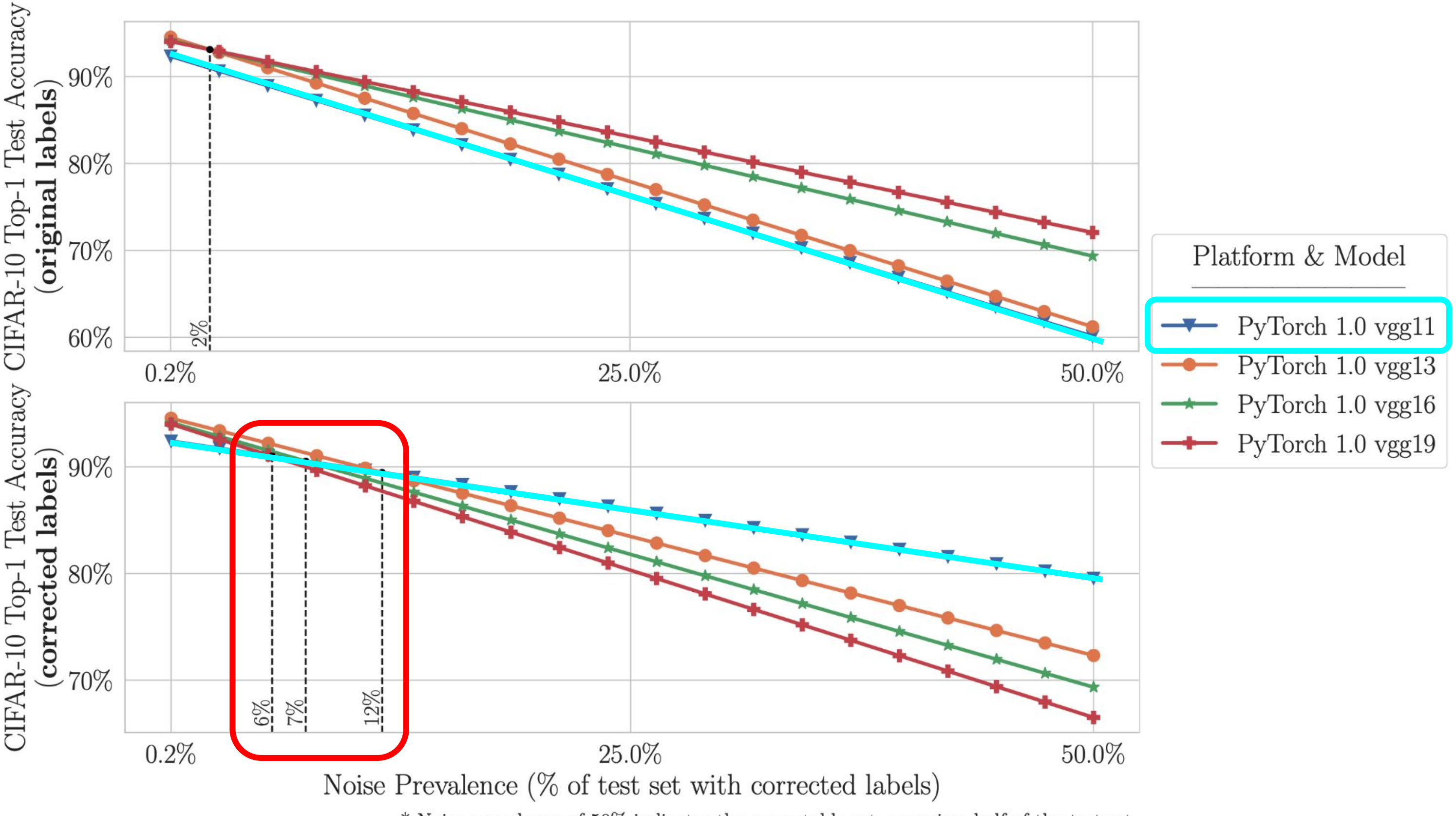


But when we correct the test set, benchmark rankings destabilize



Again we asked,
is the result is specific to ImageNet?

Same story on CIFAR-10 benchmark rankings



* Noise prevalance of 50% indicates the correctable set comprises half of the test set.

Conclusions

- Model rankings can change with just 6% increase in noise prevalence (even in these highly-curated test sets)
  - ML practitioners cannot know this unless they benchmark with corrected test set labels.
- The fact that simple models regularize (reduce overfitting to label noise) is not surprising. (Li, Socher, & Hoi, 2020)
  - The surprise -- test sets are far noisier than the ML community thought (labelerrors.com)
  - An ML practitioner's "best model" may underperform other models in real-world deployment.
- For humans to deploy ML models with confidence -- noise in the test set must be quantified
  - confident learning addresses this problem with realistic sufficient conditions for finding label errors -- and we have shown its efficacy for ten of the most popular ML benchmark test sets.

Steps to Confident Learning for Machines and Humans

1 — 2 — 3 — 4

1. Precursors to CL

Machine learning for human learning requires dealing with real-world, noisy labels

Northcutt, Ho, & Chuang (C&E, 2016)
Northcutt, Wu, & Chuang (UAI, 2017)
Northcutt, Leon, & Chen (L@S, 2017)
Corrigan-Gibbs, Gupta, Northcutt, Cutrell, & Thies (TOCHI 2015, CHI 2016)

2. Confident Learning

We develop a principled framework of theory and algorithms for quantifying, finding, and learning with label noise in datasets.

https://github.com/cgnorthcutt/cleanlab

Northcutt, Jiang, & Chuang (JAIR, 2021)

3. Label Errors in ML Datasets

We find tens of thousands (3.4%) of label errors in the most commonly benchmarked ML test sets.

labelerrors.com

Northcutt, Athalye, & Lin (NeurIPS Workshop on Dataset Curation and Security, 2020)

4. Implications for ML Practitioners

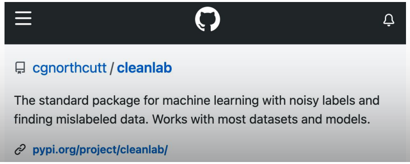
We study whether practitioners are unknowingly benchmarking the progress of ML based on erroneous test sets? How noisy is too noisy?

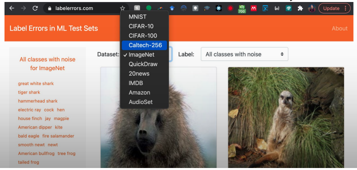
https://github.com/cgnorthcutt/label-errors

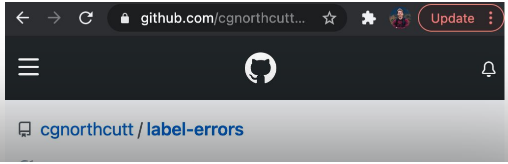
Northcutt, Athalye, & Mueller (ICLR RobustML Workshop, 2021) (ICLR WeaSuL Workshop, 2021)

Contributions of my Thesis (covered in this talk)

- Confident learning is the first framework to
  - estimate the joint distribution of noisy labels and true labels directly
    - Prior work focuses on estimating conditionals/marginals of the joint (e.g. label flipping rates)
  - provide sufficient conditions for exactly finding label errors with per-example noisy models outputs
    - Prior theory with noisy labels (mostly) focuses on learnability/ estimators (not the data)
- Label Errors + Implications for ML
  - First work to quantify noise and find label errors at scale across ten popular ML test sets.
    - Prior work on ImageNet, but it was not known that e.g., MNIST also has many label errors
  - First work to estimate the noise prevalence needed to destabilize benchmarks in popular datasets
    - Prior work has verified linear trends under distributional shift of test sets
- Public release of cleanlab, labelerrors.com, and corrected test sets







Using cleanlab, most of the results presented in this talk are reproducible in a few lines of code.

This talk focused on two (boxed in red) of five papers covered in my thesis. The other three papers/chapters focus on dealing with noisy real world data to augment human capabilities

- Curtis G. Northcutt, Lu Jiang, and Isaac L. Chuang (2021). Confident Learning: Estimating Uncertainty in Dataset Labels. In *Journal of Artificial Intelligence Research (JAIR)*.
- Curtis G. Northcutt, Anish Athalye, and Jonas Mueller (2021). Pervasive Label Errors in Test Sets Destabilize ML Benchmarks. In two *ICLR 2021 Workshops on Robust ML and Weakly Supervised Learning.*
- Curtis G. Northcutt, Cindy Zha, Steven Lovegrove, and Richard Newcombe (2020). EgoCom: A Multi-person Multi-modal Egocentric Communications Dataset. In *Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*. ← Augmented conversational cues (turn taking prediction)
- Nikola I. Nikolov, Eric Malmi, Curtis G. Northcutt, and Loreto Parisi (2020). Conditional Rap Lyrics Generation with Denoising Autoencoders. In *International Conference on Natural Language Generation (INLG).* ← Augmented writing of rap lyrics
- Curtis G. Northcutt, Kim Leon, and Naichun Chen (2017). Comment Ranking Diversification in Forum Discussions. In *Proceedings of the ACM Conference on Learning @ Scale (L@S).* ← Augmented learning in discussion forums

Thank you to my incredible co-authors!

For me, the greatest gift of grad school at MIT is the friends and colleagues I made along the way - thank you!

Thank you to my committee

Isaac Chuang, Suvrit Sra, Roz Picard

That concludes the talk portion of my defense.

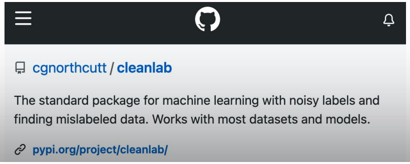
And to friends/colleagues:

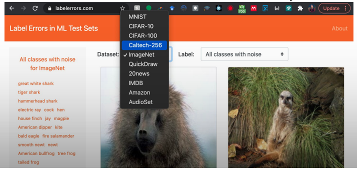
Anish Athalye (MIT), **Jonas** Mueller (Amazon), **Lisa** Vo (ChipBrain), and **my family**

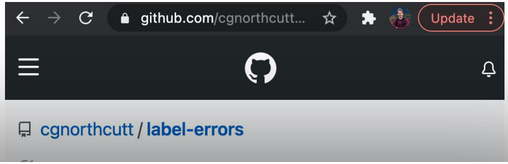
also... Lu Jiang (Google), Tailin Wu (Stanford), Gabriel Mintzer (MIT), Robin Cooper (Univ. of Kentucky), Gautam Biswas (Vanderbilt), Martin Segado (MIT), Arkopal Dutt (MIT), Natarajan Subramanyam (Google), Marek Hempel (MIT), Ludwig Schmidt (Berkeley), Nikola Nikolov (ETH Zurich), Eric Malmi (Google), Loreto Parisi (MusixMatch), Cindy Zha (Facebook Research), Steve Lovegrove (Oculus Research), Richard Newcombe (Facebook Reality Labs), and many others...

Contributions of my Thesis (covered in this talk)

- Confident learning is the first framework to
  - estimate the joint distribution of noisy labels and true labels directly
    - Prior work focuses on estimating conditionals/marginals of the joint (e.g. label flipping rates)
  - provide sufficient conditions for exactly finding label errors with per-example noisy models outputs
    - Prior theory with noisy labels (mostly) focuses on learnability/ estimators (not the data)
- Label Errors + Implications for ML
  - First work to quantify noise and find label errors at scale across ten popular ML test sets.
    - Prior work on ImageNet, but it was not known that e.g., MNIST also has many label errors
  - First work to estimate the noise prevalence needed to destabilize benchmarks in popular datasets
    - Prior work has verified linear trends under distributional shift of test sets
- Public release of cleanlab, labelerrors.com, and corrected test sets







Using cleanlab, most of the results presented in this talk are reproducible in a few lines of code.