How many labels do you have?
Some perspectives on gold standard labels

John Duchi

Based on joint work with Chen Cheng and Hilal Asi

NSF
Office of Naval Research
ONR
Science & Technology
USN
STANFORD
DAWN

The "standard" story in statistics & ML

Data →

$$\{(X_i, Y_i)\}_{i=1}^{N}$$

$$y = X\beta + \varepsilon$$

$$\begin{aligned} &\text{minimize } f(x) \\ &\text{subject to } x \in X \end{aligned}$$

$$g(x) \in \int_{\mathcal{S}} \partial F(x; s) dP(s)$$

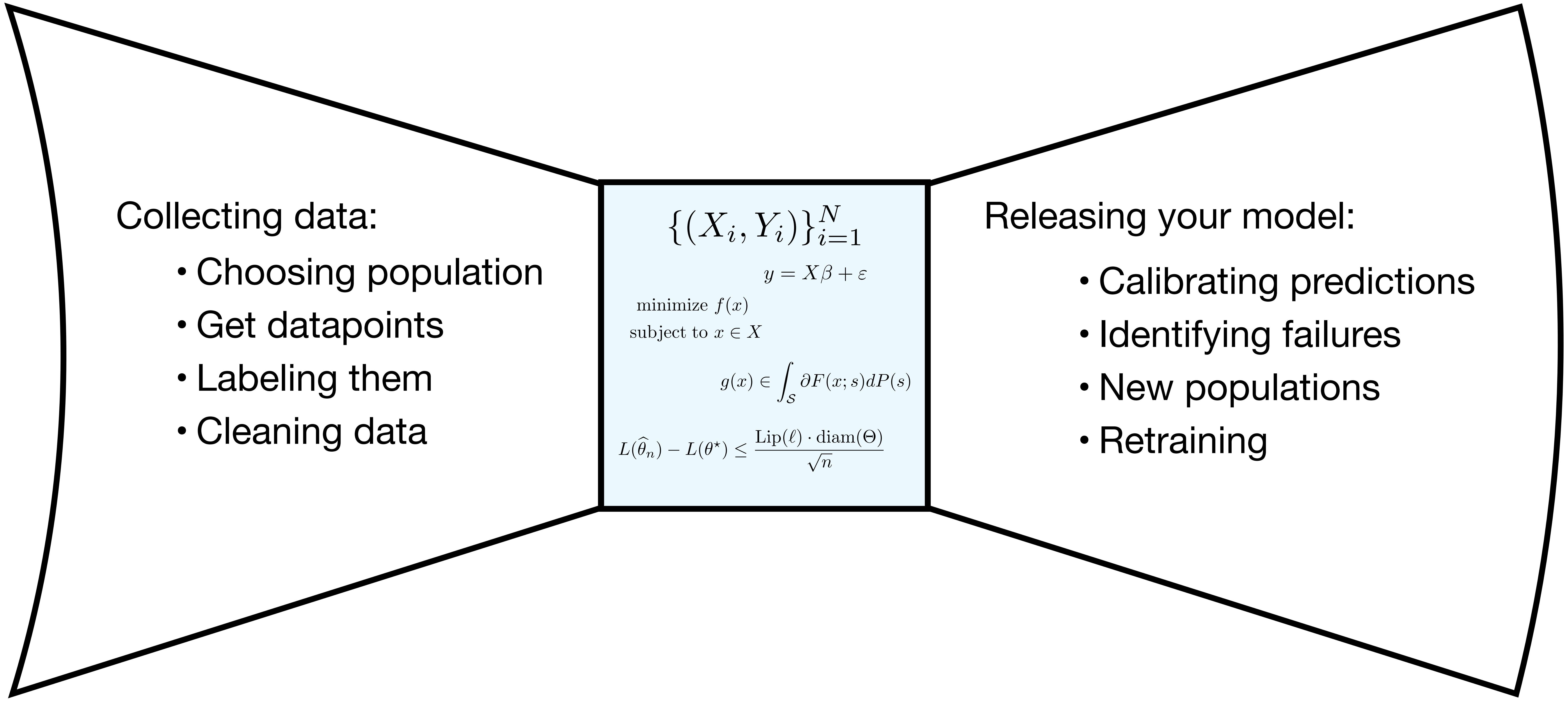
$$L(\widehat{\theta}_n) - L(\theta^\star) \leq \frac{\mathrm{Lip}(\ell) \cdot \mathrm{diam}(\Theta)}{\sqrt{n}}$$

Statistics and machine learning

→ Great model

The big picture

- Excited about the full pipeline of statistical machine learning



Motivation

Dave Donoho, “50 Years of Data Science”

> It is no exaggeration to say that the combination of a Predictive Modeling culture together with Common Task Framework is the ‘secret sauce’ of machine learning

Common Task Framework:

1. A publicly available training dataset
2. A set of enrolled competitors whose common task is to infer a class prediction rule from the training data
3. A scoring referee to which competitors submit their prediction rule(s)

ImageNet

The (probably) currently preferred image classification benchmark

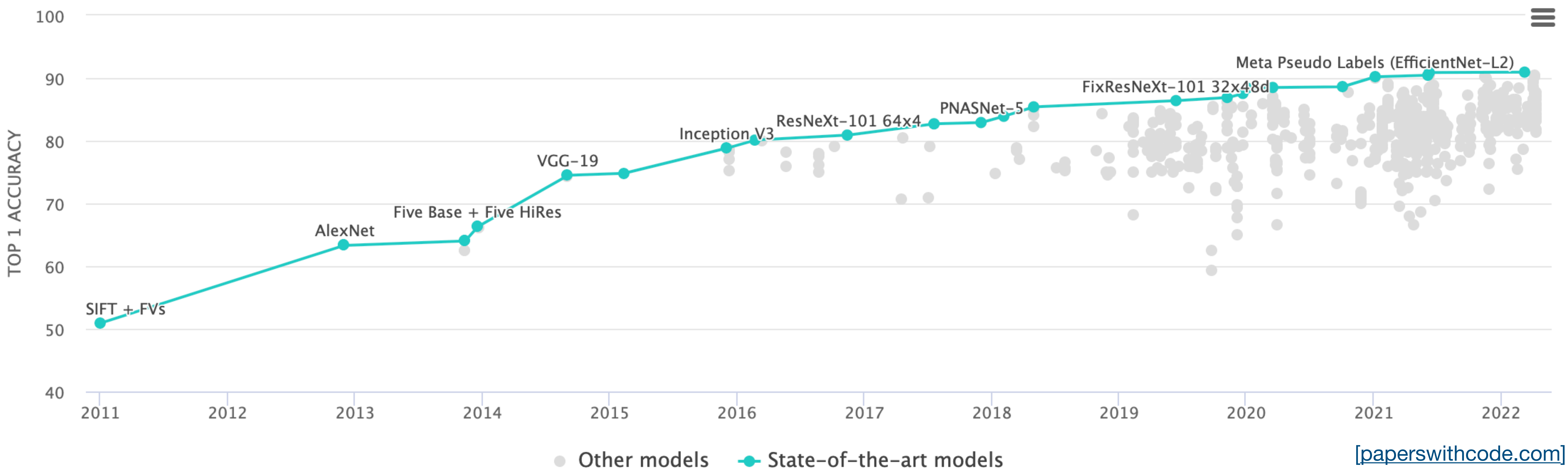
Input data X

Goal: assign label Y to this image (In this case, Y = Golden Retriever)

Dataset description: For each of 1000 image categories (e.g. cherry, bow and arrow, golden retriever, dachshund) there are 1000 representative images

[Deng et al., “ImageNet: A large-scale hierarchical image database” CVPR 2009]

ImageNet Progress



[paperswithcode.com]

Little exaggeration to say deep learning descends from ImageNet

Supervised Learning

The construction of ImageNet isn't really what we teach

- Usual machine learning story

Input data X

Noisy Label Y

"Dog"

"Golden Retriever Puppy"

"Cat"

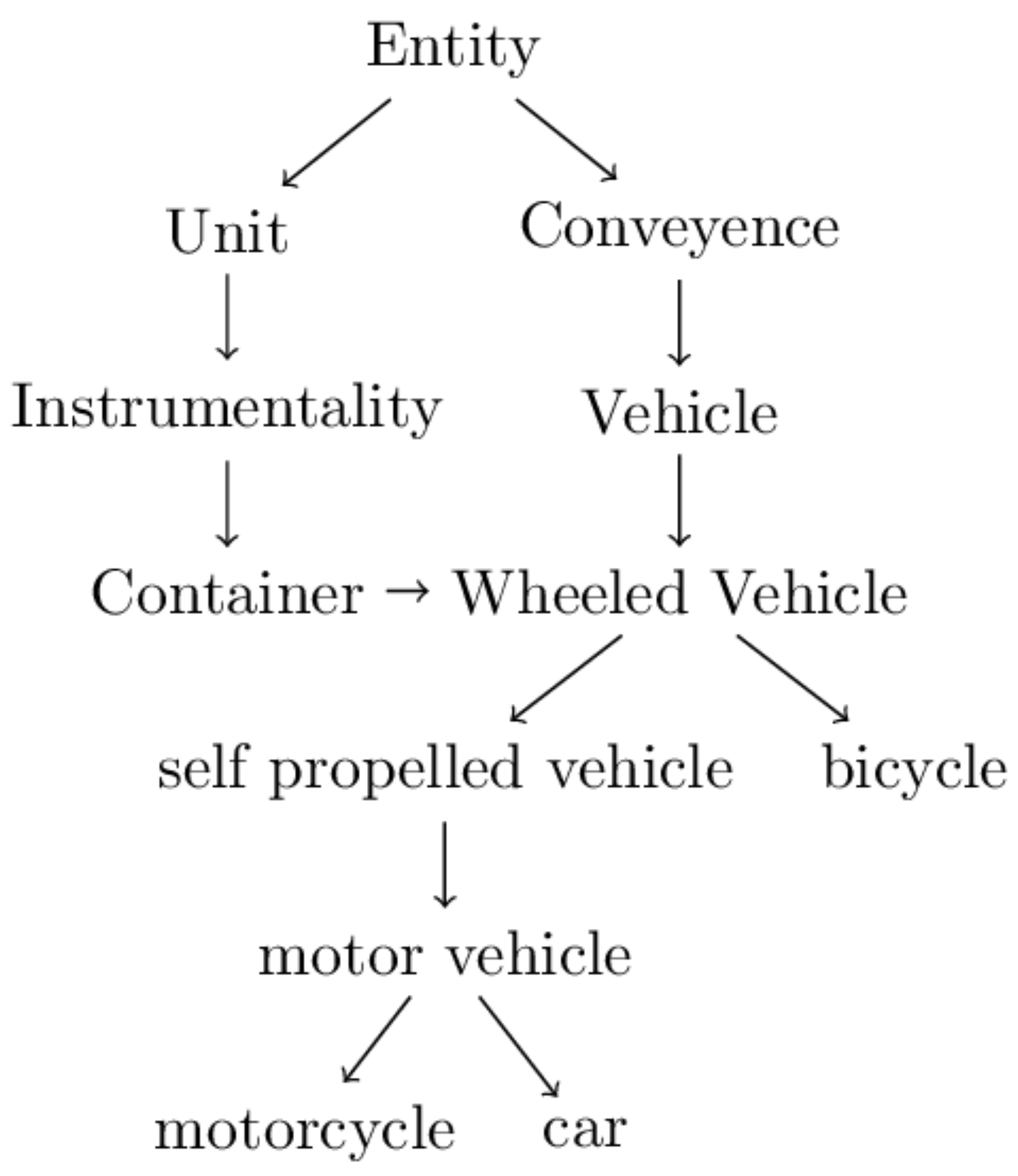
Machine learning pipeline:

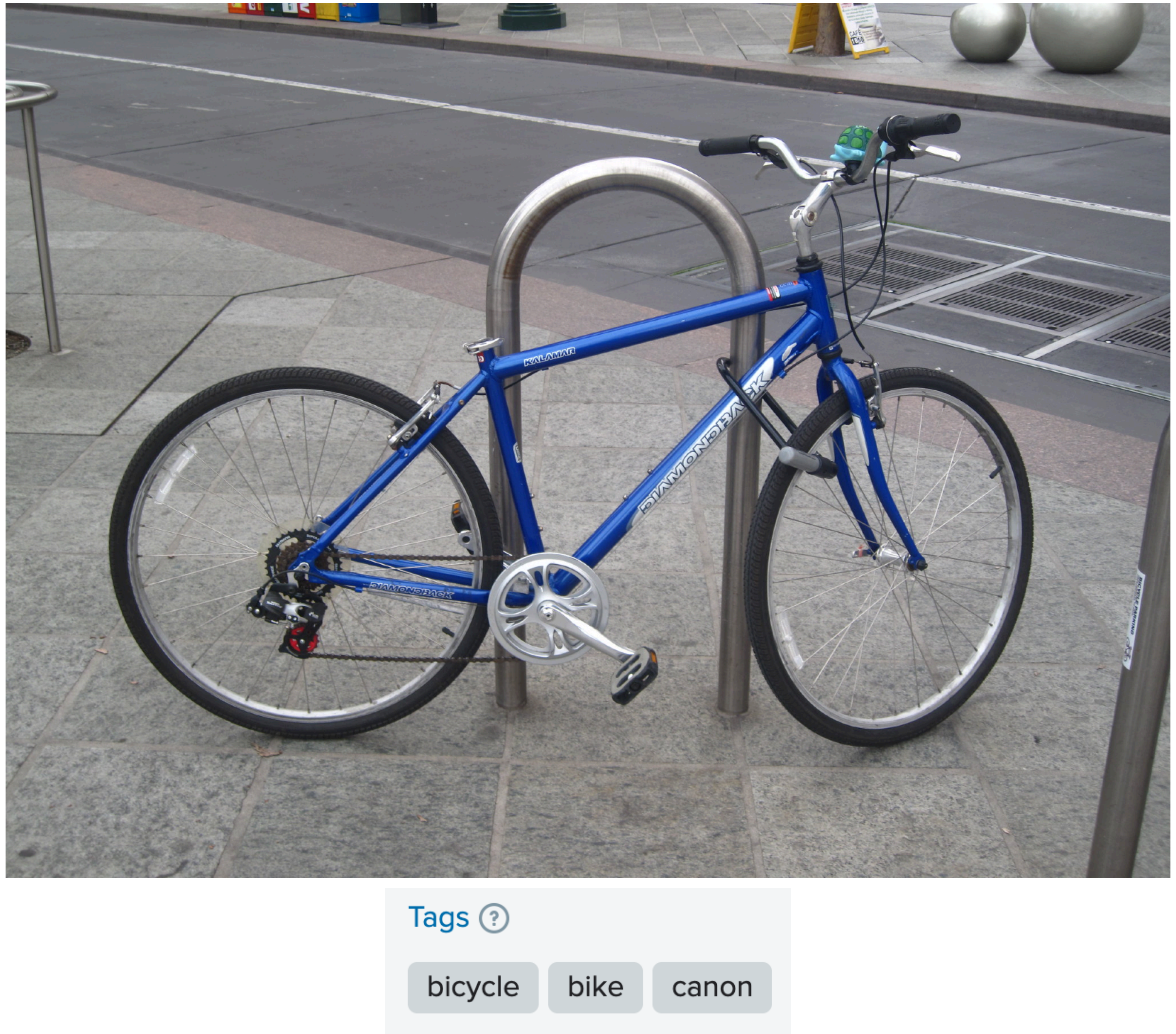
Feed in a bunch of pairs $\{(X_i, Y_i)\}_{i=1}^N$ ——(magical fitting…)——> Output a model $\widehat{Y} = f(X)$

ImageNet construction

flickr

WordNet hierarchy





[Deng et al., "ImageNet: A large-scale hierarchical image database" CVPR 2009]

ImageNet construction

Select all the images that contain a bicycle

[Deng et al., “ImageNet: A large-scale hierarchical image database” CVPR 2009]

ImageNet construction

Select all the images that contain a bicycle

Bicycle object class:

Include X as an example of Y = bicycle if selection frequency $\geq 70\%$

Repeat for each of 1000 classes

What this is:

Ingenious, clever, surprisingly effective

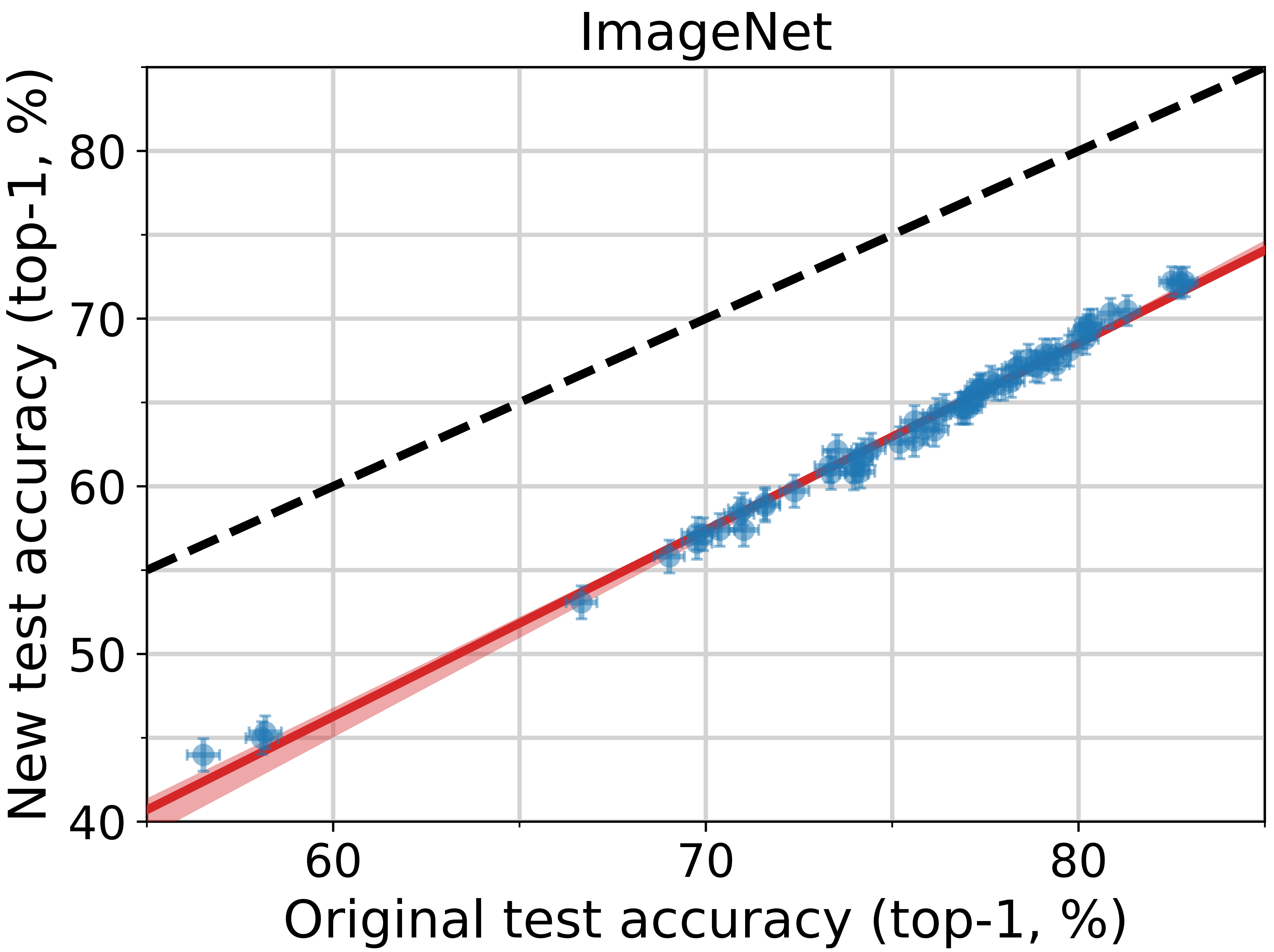
What this is not:

Noisy labels Y given X

[Deng et al., "ImageNet: A large-scale hierarchical image database" CVPR 2009]

How much does data construction matter?

Even when we're careful, things get weird



Also, methods are quite overconfident in predictions (e.g., predict classes with 90+% certainty)

[Recht, Roelofs, Schmidt, Shankar, "Does ImageNet generalize to ImageNet?", ICML 2019]

Remainder of this talk

Collecting data:

- Choosing population
- Get datapoints
- Labeling them
- Cleaning data

A very stylized model of part of this process

$$\{(X_i, Y_i)\}_{i=1}^N$$

$$y = X\beta + \varepsilon$$

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in X \end{array}$$

$$g(x) \in \int_{\mathcal{S}} \partial F(x; s) dP(s)$$

$$L(\widehat{\theta}_n) - L(\theta^\star) \le \frac{\text{Lip}(\ell) \cdot \text{diam}(\Theta)}{\sqrt{n}}$$

Releasing your model:

- Calibrating predictions
- Identifying failures
- New populations
- Retraining

Steps:

1. Propose a model
2. Analyze the model
3. It makes some predictions: test them!

The model

- Binary classification with *m* labelers

$$Y \in \{-1, 1\}, \quad X \in \mathbb{R}^d$$

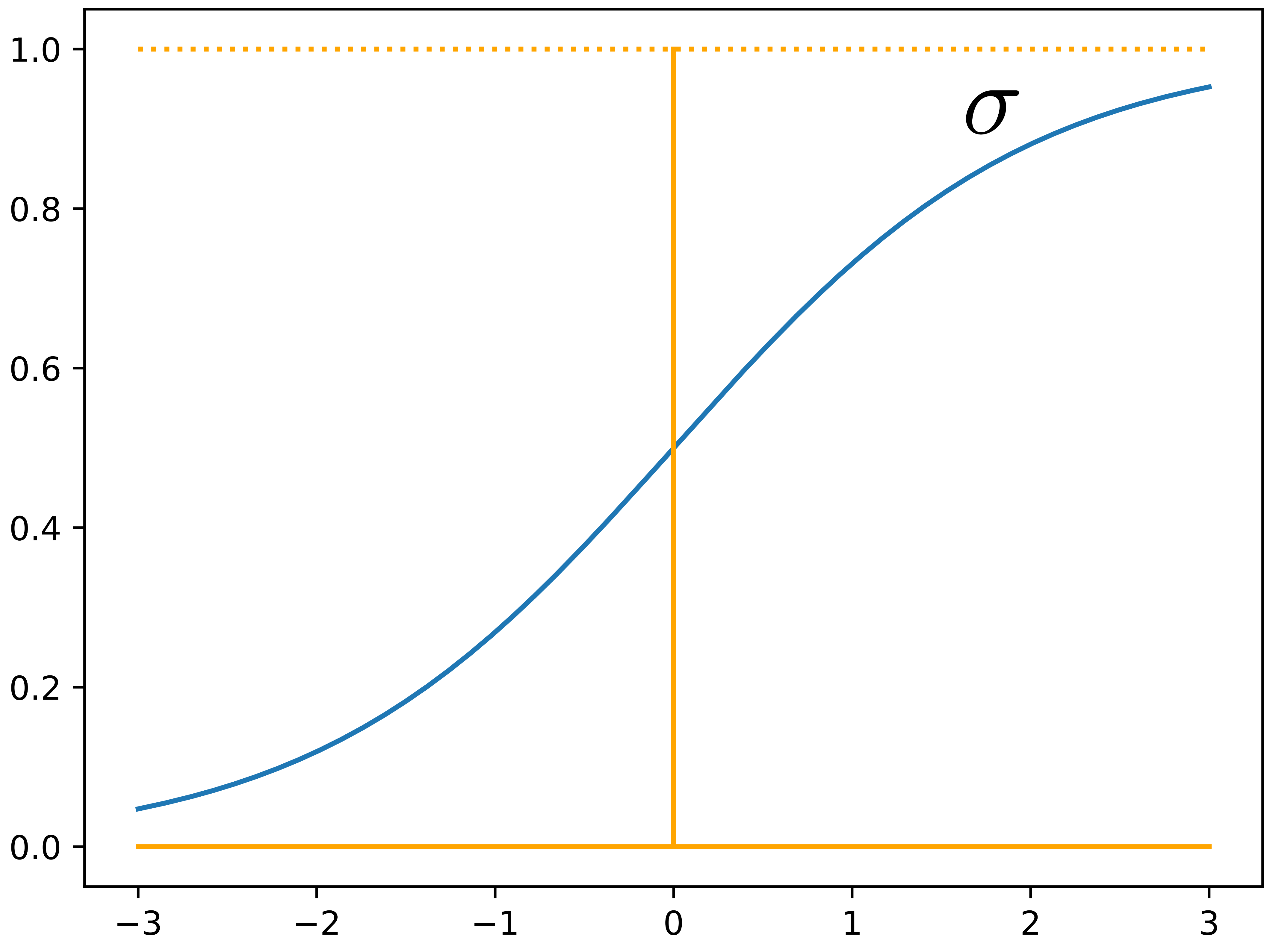
- Symmetric link function

$$\mathbb{P}(Y = y \mid X = x) = \sigma(y x^\top \theta^\star)$$

- Data in tuples (*n* total tuples)

$$(X, Y_1, \ldots, Y_m), \qquad Y_j \mid X \stackrel{\text{iid}}{\sim} \mathbb{P}(\cdot \mid X)$$

- Covariate vectors $X_i \stackrel{\text{iid}}{\sim} \mathsf{N}(0, I_d)$



The model (continued)

- Margin-based loss ℓ satisfying

$$\ell'(t) = -\sigma(-t)$$

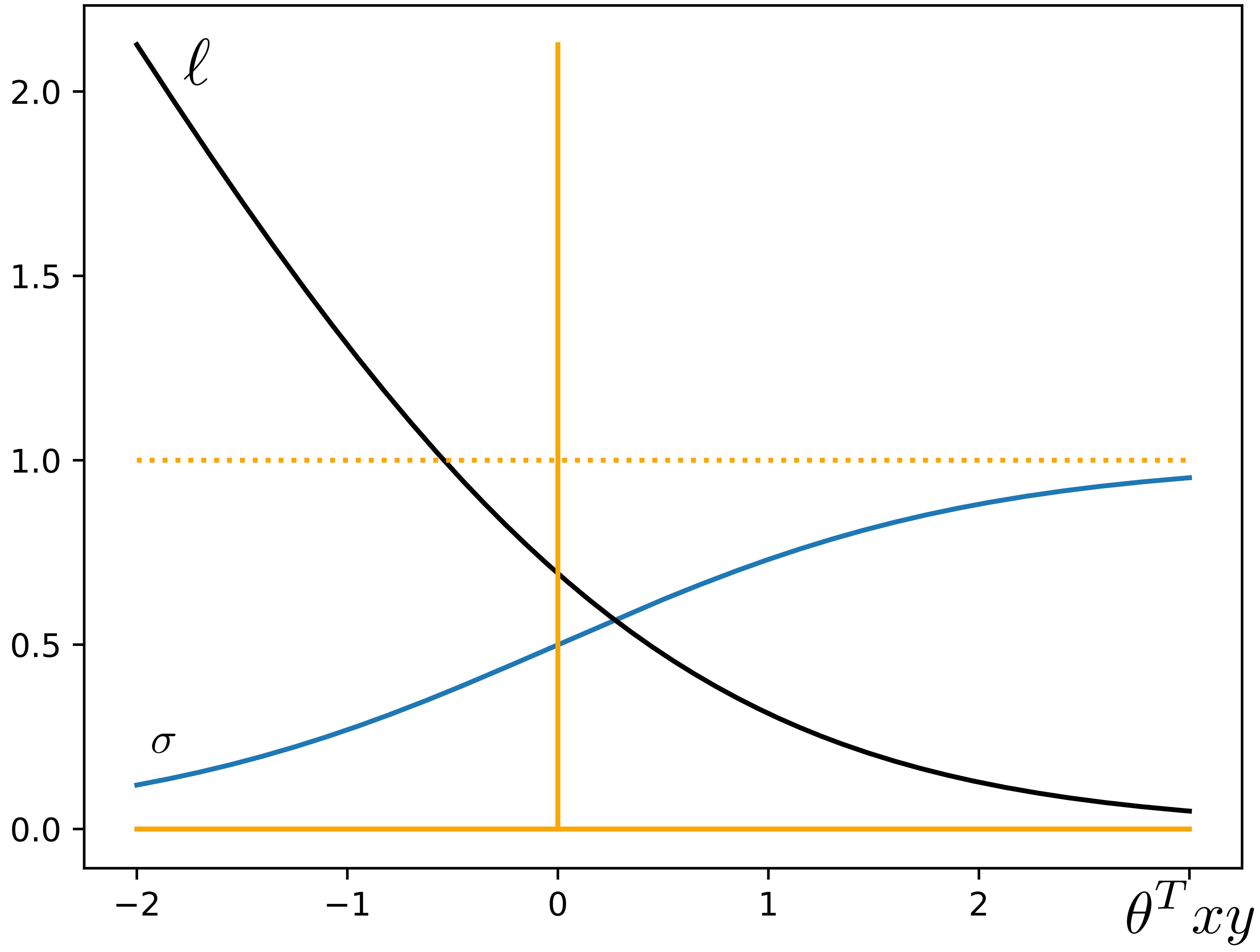
- Loss of parameter θ on (x, y)

$$\ell(\theta^T xy)$$

- E.g. logistic regression

$$\ell(t) = \log(1 + e^t)$$

$$\sigma(t) = \frac{1}{1 + e^{-t}}$$



The two estimators

- Use all the labels

$$L_n(\theta) := \frac{1}{n}\frac{1}{m}\sum_{i=1}^{n}\sum_{j=1}^{m}\ell(Y_{ij}X_i^\top\theta)$$

(Log likelihood for multiple labels)

$$\widehat{\theta}_n = \operatorname*{argmin}_\theta L_n(\theta)$$

- Use majority vote

$$L_n^{\mathrm{mv}}(\theta) := \frac{1}{n}\sum_{i=1}^{n}\ell(\overline{Y}_i X_i^\top\theta)$$

where $\overline{Y}_i = \text{Majority}(Y_{i1},\ldots,Y_{im})$

$$\widehat{\theta}_n^{\mathrm{mv}} = \operatorname*{argmin}_\theta L_n^{\mathrm{mv}}(\theta)$$

Main quantities of interest:

- Calibration error $\left\|\widehat{\theta} - \theta^\star\right\|_2$
- Classification error $\left\|\widehat{u} - u^\star\right\|_2$ where $u = \theta/\left\|\theta\right\|_2$ is unit

Convergence of the MLE

$$L_n(\theta) := \frac{1}{n}\frac{1}{m}\sum_{i=1}^{n}\sum_{j=1}^{m}\ell(Y_{ij}X_i^\top\theta) \qquad\qquad L(\theta) = \mathbb{E}[\ell(YX^\top\theta)]$$

$$\widehat{\theta}_n = \operatorname*{argmin}_{\theta} L_n(\theta)$$

Theorem:

Under the well-specified model, we have asymptotic normality

$$\sqrt{n}\left(\widehat{\theta}_n - \theta^\star\right) \overset{d}{\to} \mathsf{N}\Big(0, \frac{1}{m}\nabla^2 L(\theta^\star)^{-1}\operatorname{Cov}(\dot{\ell}_{\theta^\star})\nabla^2 L(\theta^\star)^{-1}\Big)$$

and

$$\sqrt{n}\left(\widehat{u}_n - u^\star\right) \overset{d}{\to} \mathsf{N}\Big(0, \frac{1}{m\|\theta^\star\|_2^2}(I - u^\star u^{\star\top})\Big)$$

Convergence of majority vote

$$L_n^{\mathrm{mv}}(\theta) := \frac{1}{n}\sum_{i=1}^{n} \ell(\overline{Y}_i X_i^\top \theta) \quad \text{where } \overline{Y}_i = \mathsf{Majority}(Y_{i1}, \ldots, Y_{im})$$

Decompose $X = u^\star Z + (I - u^\star u^{\star\top})X = u^\star Z + W$

Theorem: Under the model, we have "overconfident" convergence

$$\widehat{\theta}_n^{\mathrm{mv}} \xrightarrow{p} t_m u^\star \qquad \text{where} \qquad t_m \asymp \sqrt{m}$$

and asymptotic normality

$$\sqrt{n}\,(\widehat{u}_n^{\mathrm{mv}} - u^\star) \xrightarrow{d} \mathsf{N}\Big(0, \frac{1}{t_m^2} H(t_m)^\dagger C(t_m) H(t_m)^\dagger\Big)$$

$$\stackrel{\mathrm{dist}}{=} \mathsf{N}\Big(0, \frac{c(1+o_m(1))}{\sqrt{m}}(I - u^\star u^{\star\top})\Big)$$

for matrices $H(t) = \frac{1}{4t}\mathbb{E}[WW^\top](1+o(1))$ and $C(t) = \frac{c}{t}\mathbb{E}[WW^\top](1+o(1))$

Robustness of majority vote

Theorem: With misspecified link, we have "overconfident" convergence

$$\widehat{\theta}_n^{\mathrm{mv}} \stackrel{p}{\to} t_m u^\star \qquad \text{where} \qquad t_m \asymp \sqrt{m}$$

and asymptotic normality (for fixed Σ)

$$\sqrt{n}\,(\widehat{u}_n^{\mathrm{mv}} - u^\star) \stackrel{d}{\to} \mathsf{N}\Big(0, \frac{1 + o_m(1)}{\sqrt{m}}\Sigma\Big)$$

Take home messages:

- Majority vote is (unfixably) uncalibrated and overconfident
- More robust (doesn't matter if the link is correct)
- Less efficient when the link *is* correct

Extensions: semiparametric estimates

- Corrected estimator: fit the model, refit the link, refit the model

$$\widehat{\theta}^{\mathrm{mv}} = \operatorname{argmin} L_n^{\mathrm{mv}}(\theta)$$

$$\widehat{\sigma} = \operatorname{argmin} \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} (2\sigma(\widehat{u}^\top X_i Y_{ij}) - 1 - Y_{ij})^2$$

Theorem: Under appropriate conditions,

$$\widehat{\theta} = \operatorname{argmin} \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \ell_{\widehat{\sigma}}(\theta^\top X_i Y_{ij})$$

is efficient: $$\sqrt{n}\left(\widehat{\theta} - \theta^\star\right) \stackrel{d}{\to} \mathsf{N}\left(0, \frac{1}{m} I(\theta^\star)^{-1}\right)$$

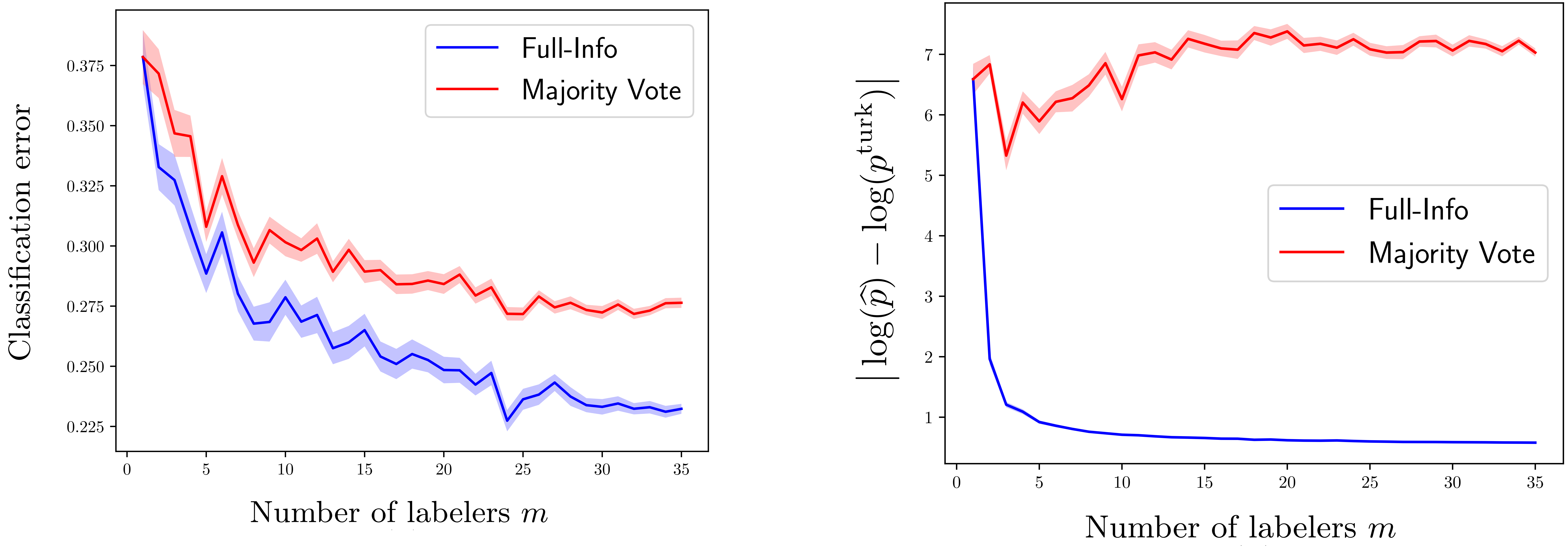
Experimental results

If our model is reasonable, it should make real predictions

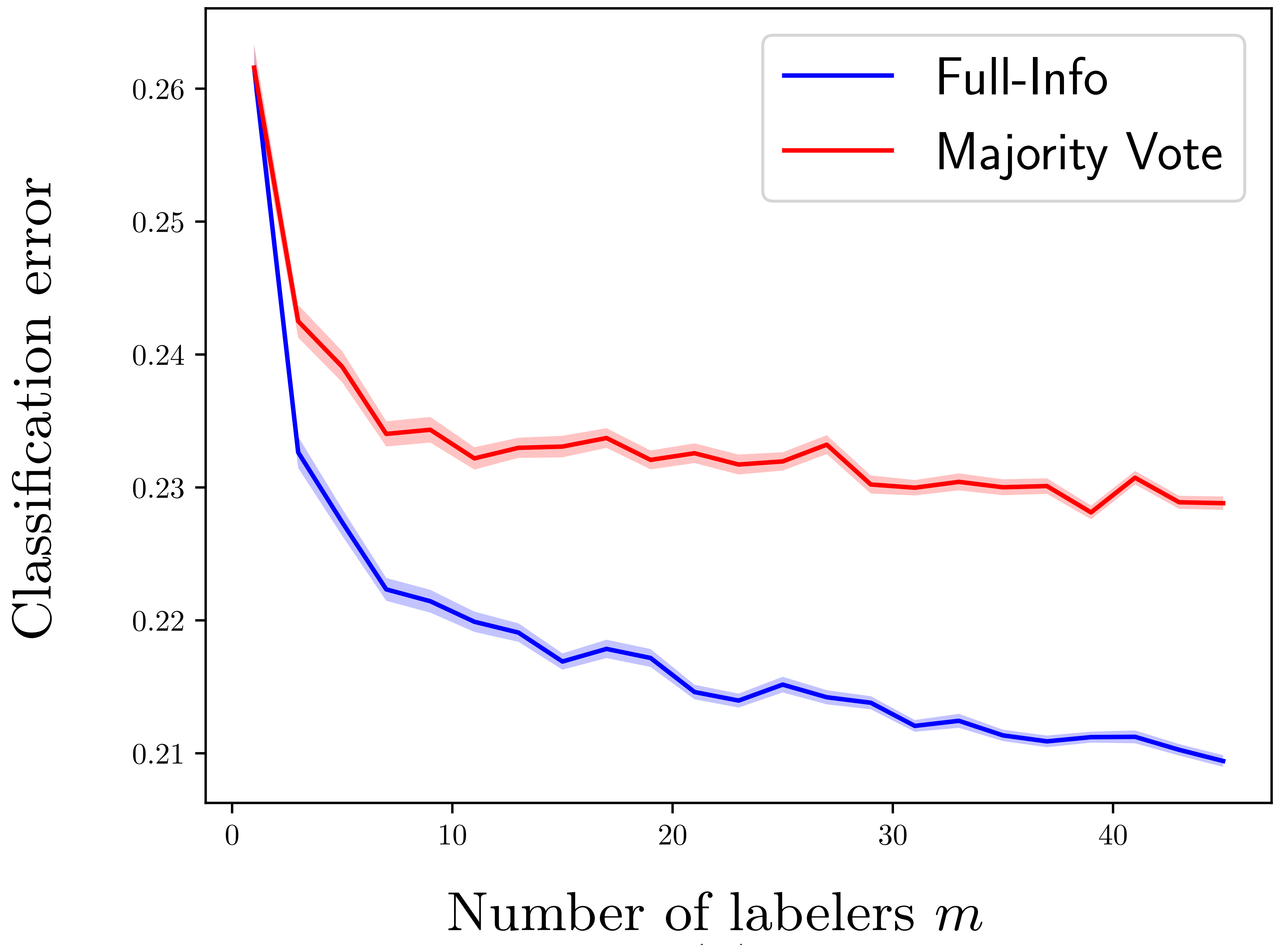
- BlueBirds: *Indigo Bunting* versus *Blue Grosbeak* [Welinder, Branson, Perona, Belongie 10]

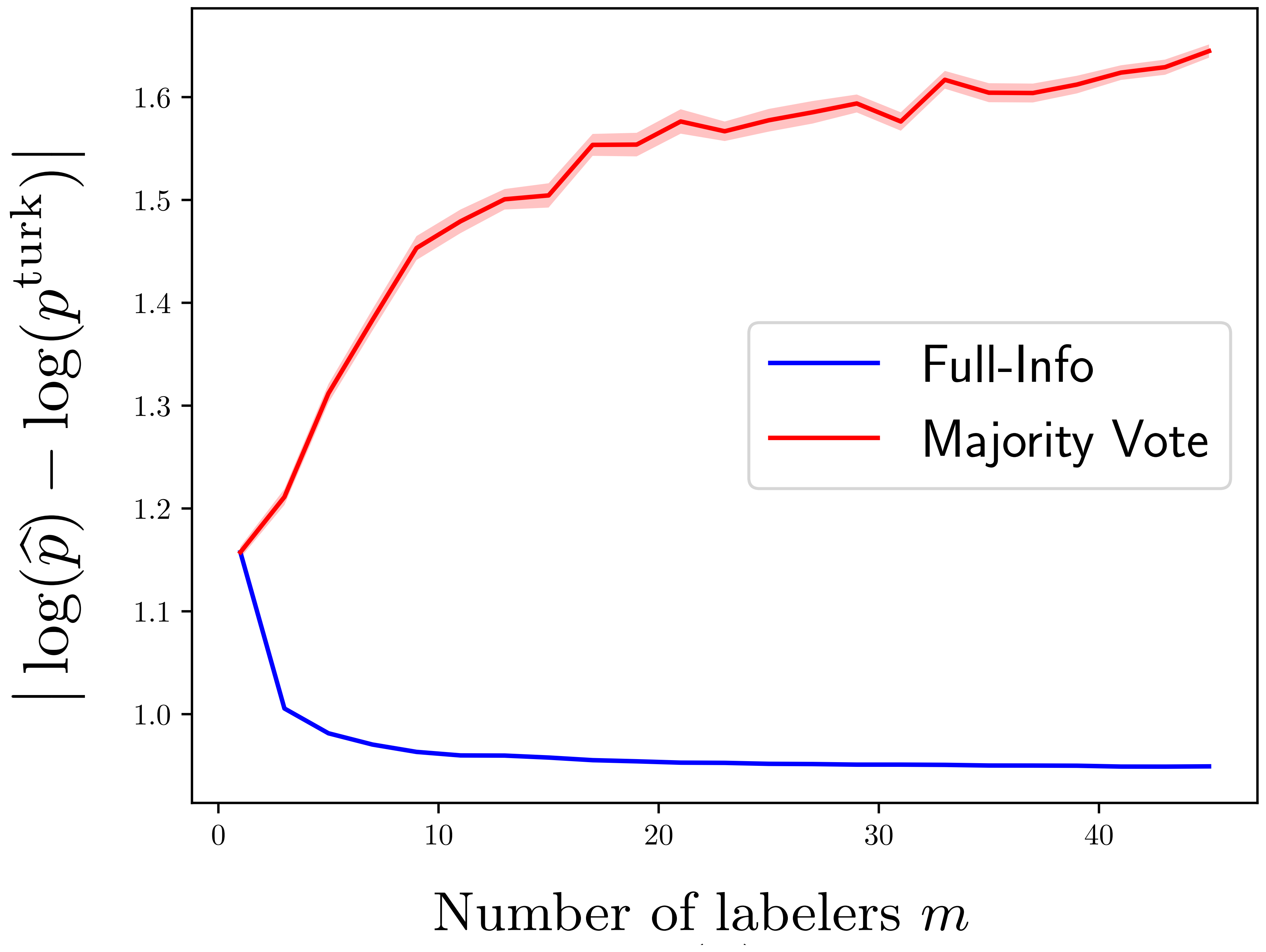
- CIFAR-10H: soft labels of CIFAR-10 test set [Peterson, Battleday, Griffiths, Russakovsky 19]

Experimental results: bluebirds



Experimental results: CIFAR-10H





Conclusions and next steps

- Interesting to think about dataset construction: a place for statistics to lay down some intellectual foundations
- Would obtaining data with (human) perceptual uncertainty help build better prediction methods?
- Currently limited datasets like those above: develop datasets to drive progress we want to see
- Fun to make (theoretical) predictions that can be wrong