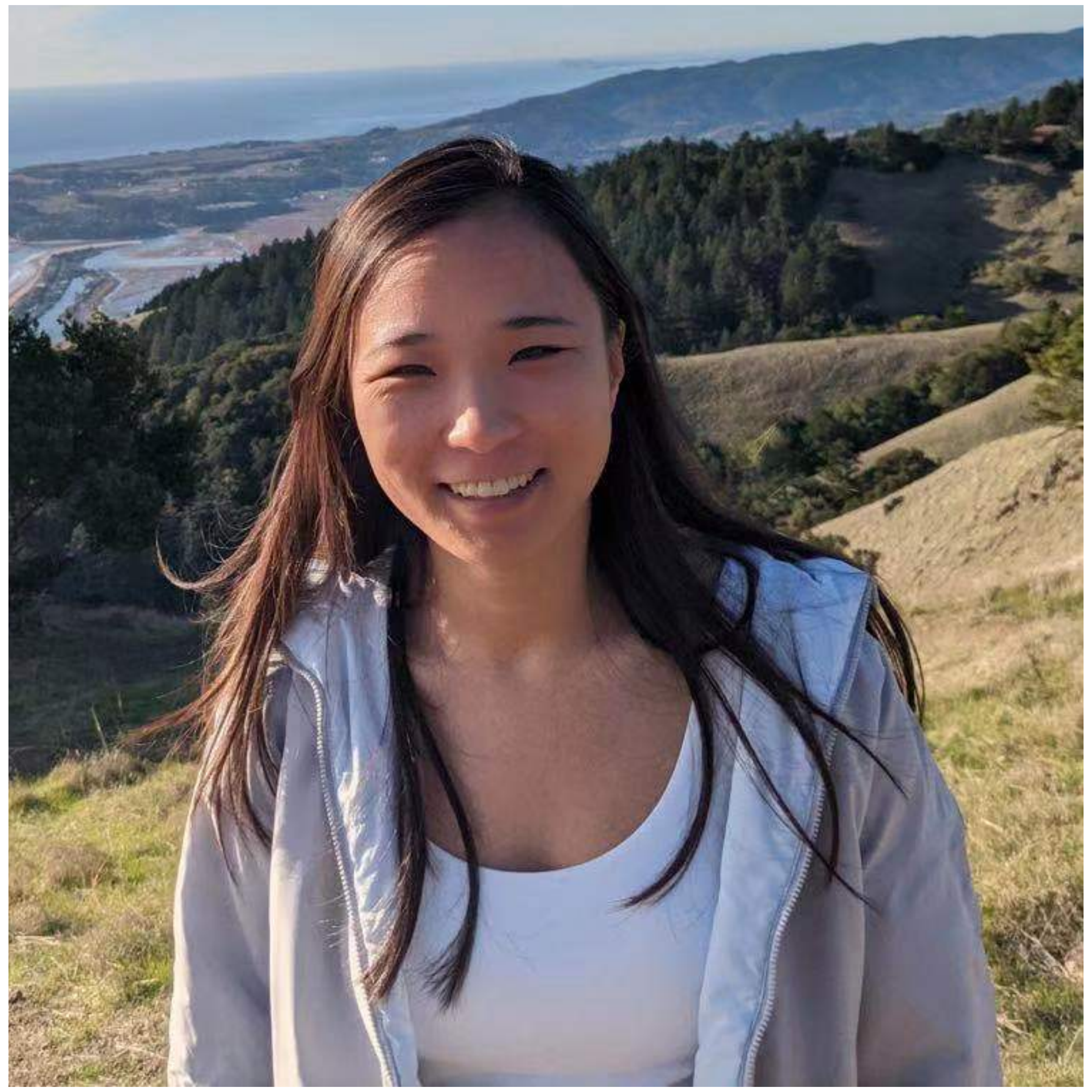
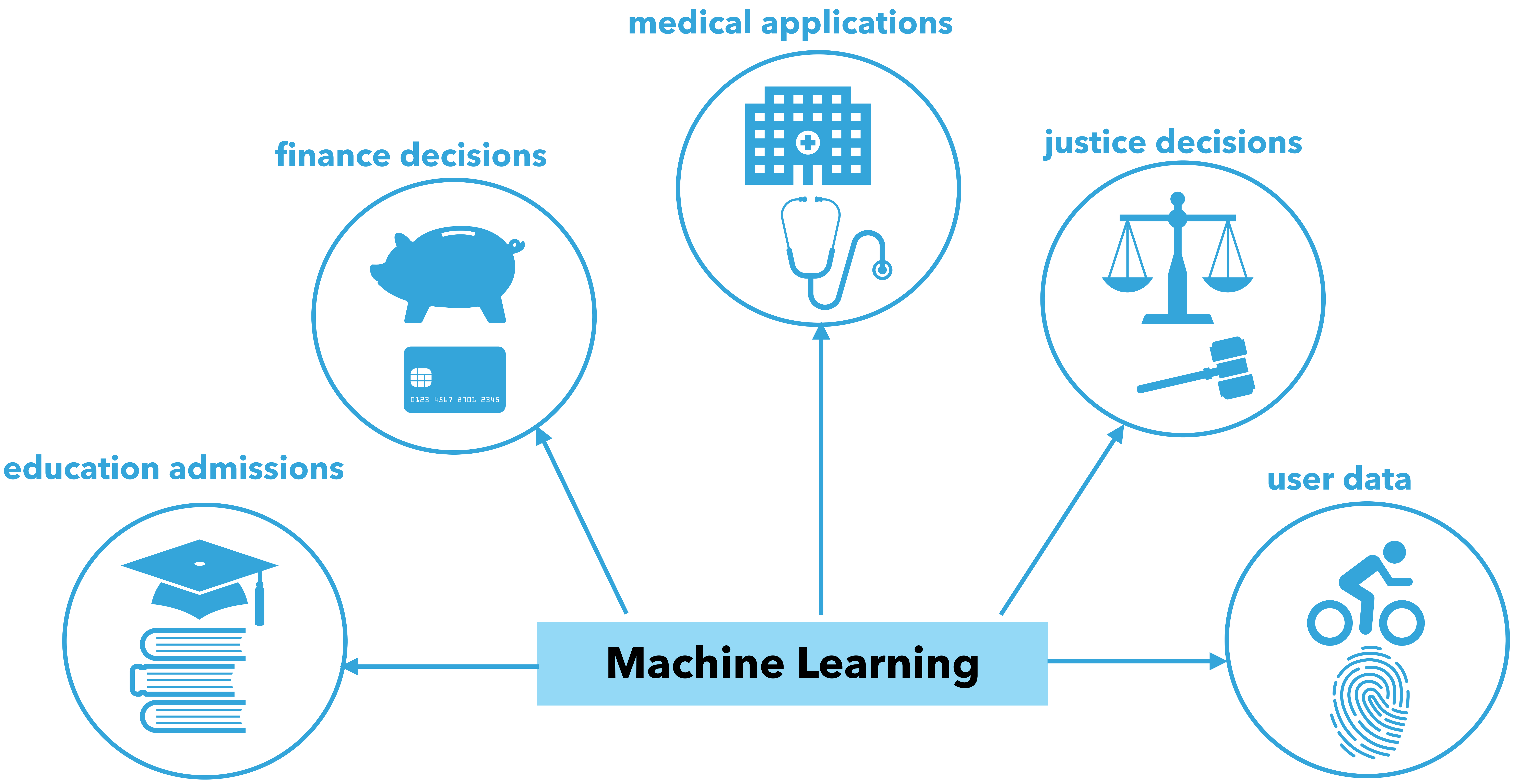
ACTIVE BAYESIAN ASSESSMENT OF BLACK-BOX CLASSIFIERS

Disi Ji (UC Irvine)



Joint work with **Robert L. Logan IV**, **Padhraic Smyth** and **Mark Steyvers**

BACKGROUND

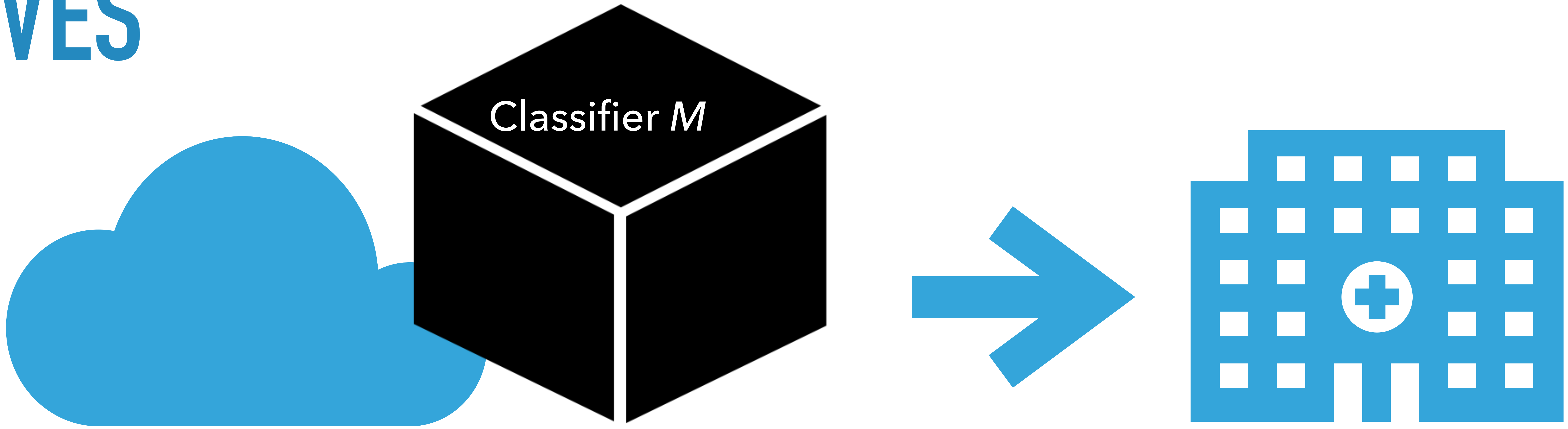


Disi Ji , Robert L. Logan IV , Padhraic Smyth , Mark Steyvers
UCIrvine

BACKGROUND

- **Assess** performance of machine learning models **independently** from the **training** procedures
  - **legal requirements**, e.g. General Data Protection Regulation (GDPR)
  - build consumers' **trust** in model predictions
  - **distribution change** at deployment time:
    - label shift [*Lipton et al. 2018*]
    - corruptions and perturbations [*Hendrycks et al. 2019, Ovadia et al. 2019b*]
  - models' **inability to generalize** [*Recht et al. 2019*]

OBJECTIVES

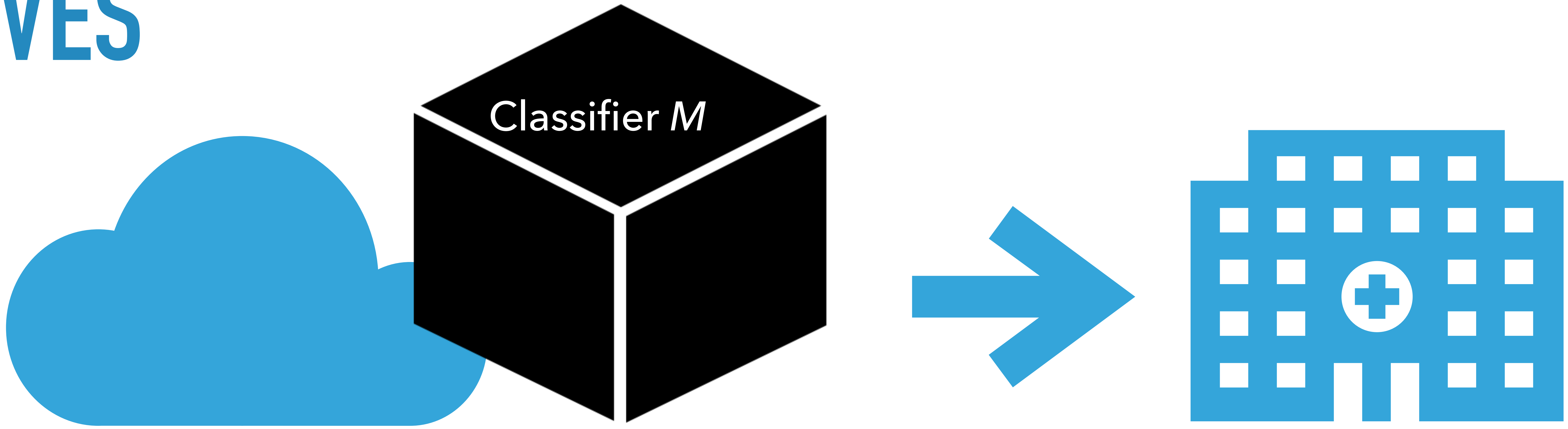


- **Estimation**: How accurate?
- **Identification**: Where is the model least accurate?
- **Comparison**: Is the model fair, e.g. equally accurate across different demographic groups?

(Can replace accuracy with other performance metrics, e.g., calibration metrics)

Requires labeled data!

OBJECTIVES



- **Estimation**: How accurate?
- **Identification**: Where is the model least accurate?
- **Comparison**: Is the model fair, e.g. equally accurate across different demographic groups?

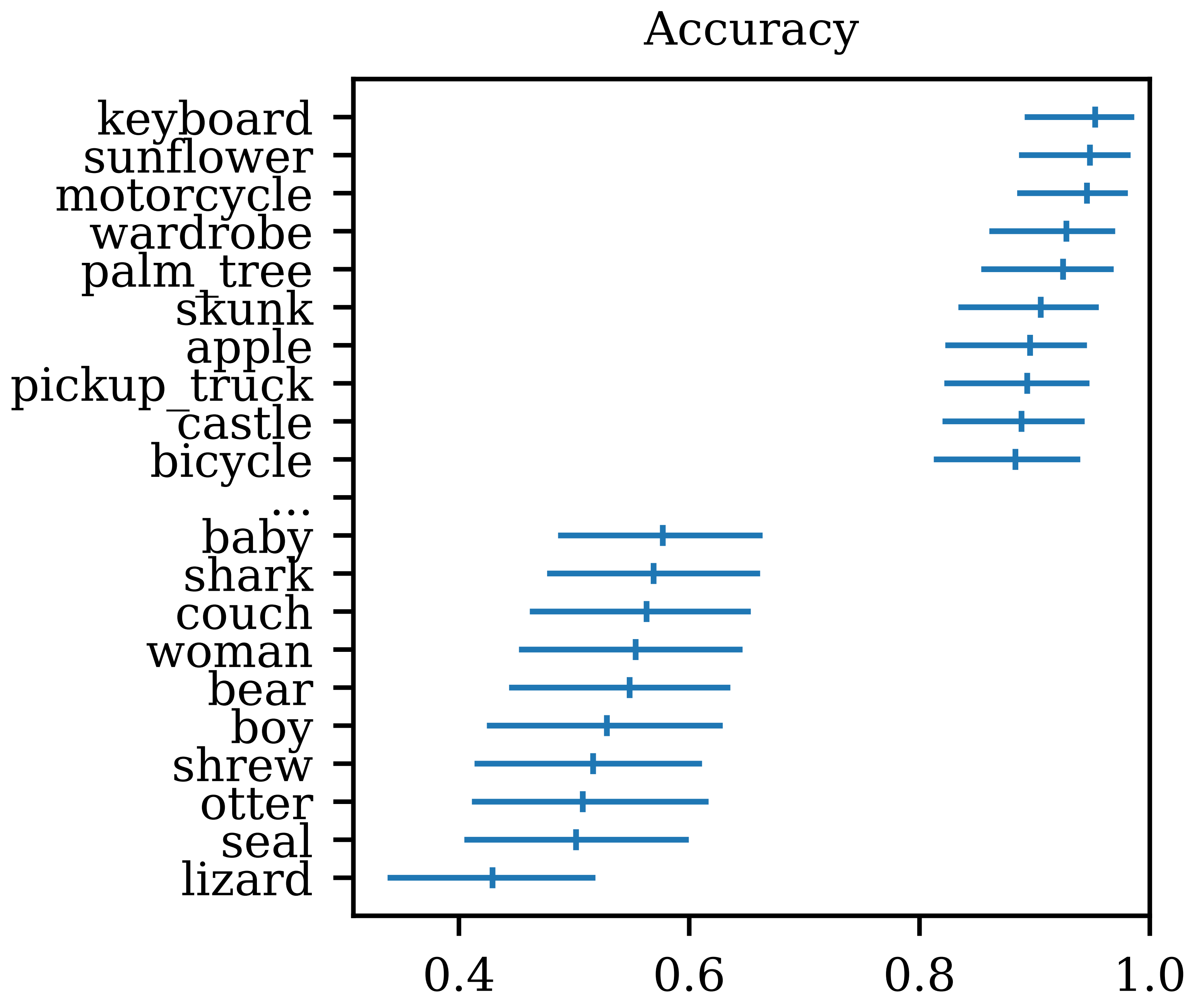
(Can replace accuracy with other performance metrics, e.g., calibration metrics)

Requires labeled data!

- How much **confidence** should we have in this assessment?
- How best to **increase our confidence** given a limited budget for labeled data?

Bayesian assessment

1. **Quantify uncertainty** of assessment with Bayesian methods, **with a set of labeled data**

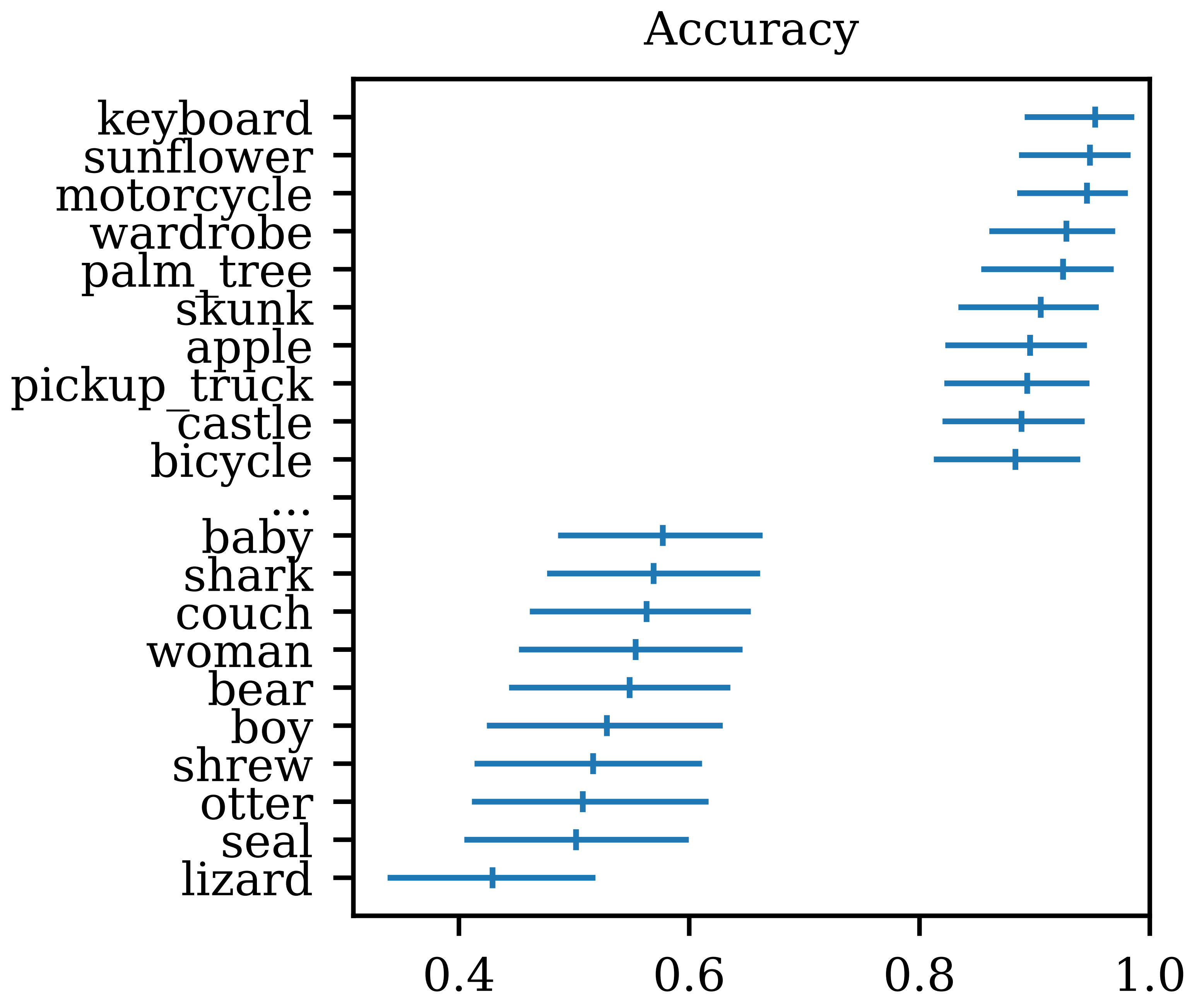


Disi Ji , Robert L. Logan IV , Padhraic Smyth , Mark Steyvers

ROAD MAP

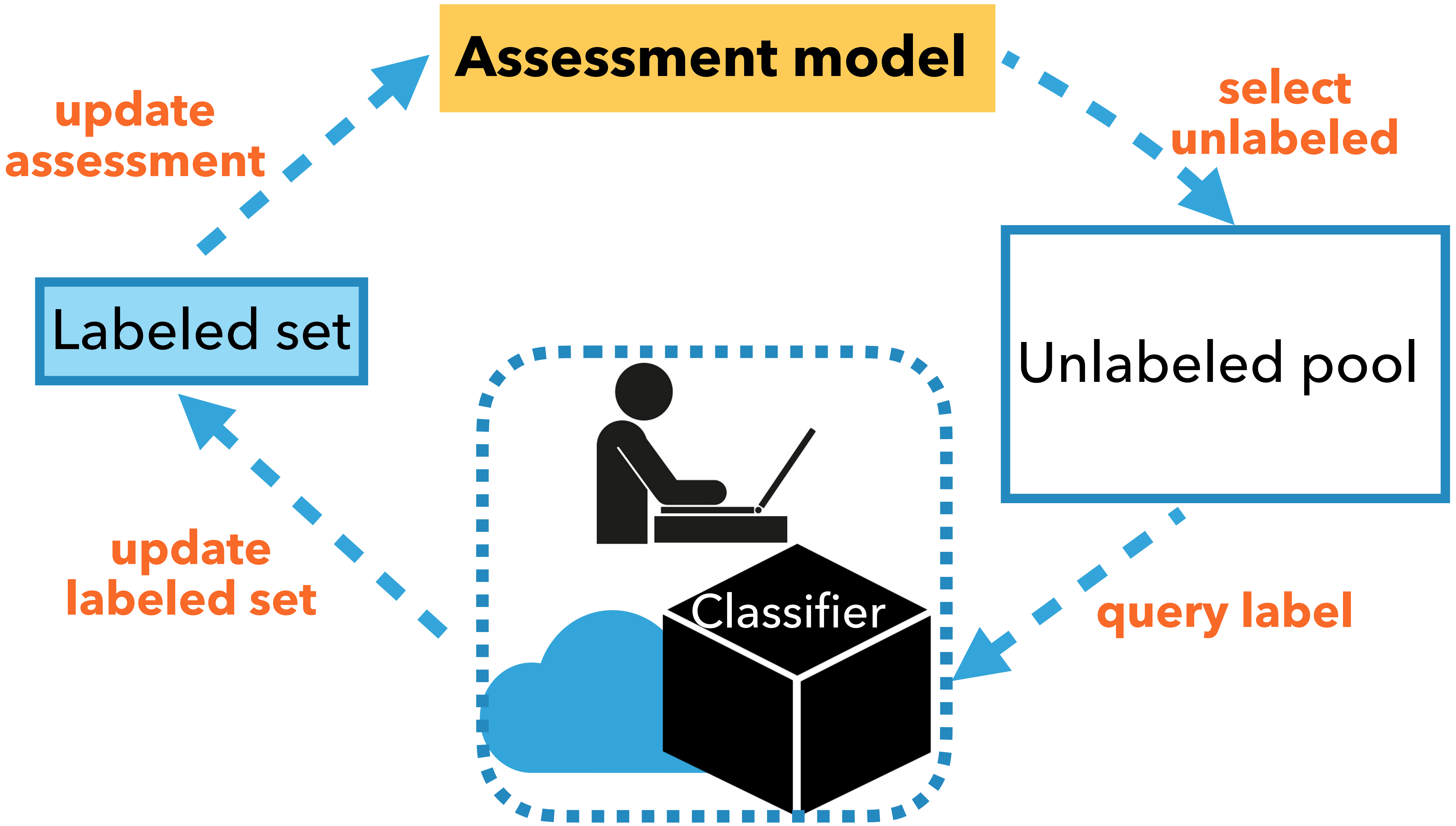
Bayesian assessment

1. **Quantify uncertainty** of assessment with Bayesian methods, **with a set of labeled data**



Active Bayesian assessment

2. **Reduce uncertainty** of assessment, by **actively labeling data** from **a pool of unlabeled data**



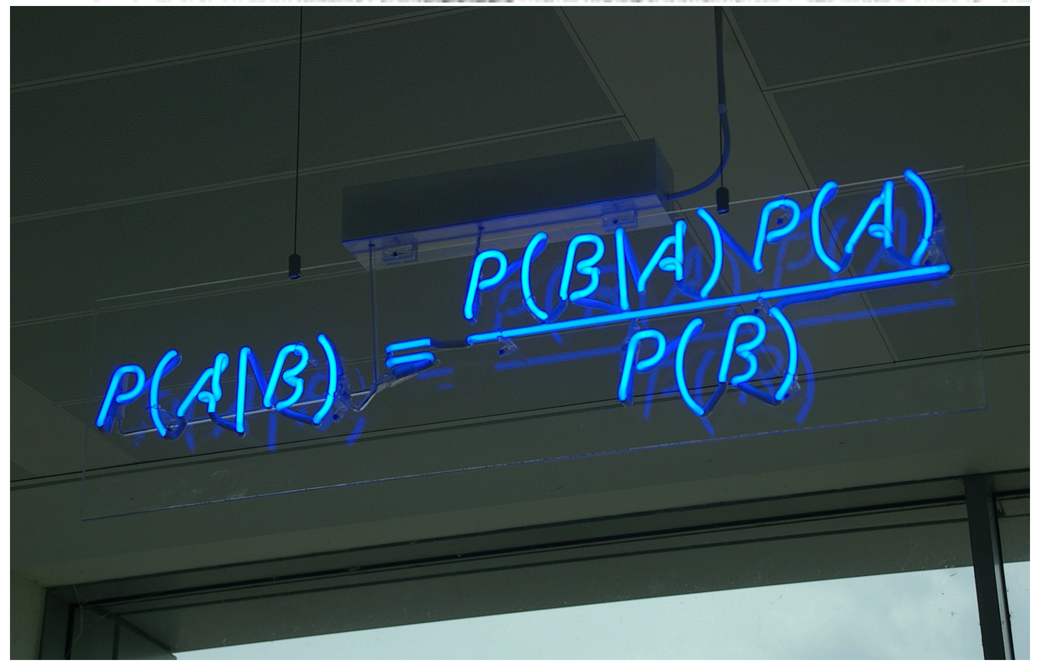
BAYESIAN ASSESSMENT: HOW ACCURATE

Performance metric of interest: θ

Labeled data: $D = \{(x_i, y_i) \,|\, i = 1,2,\cdots,N\}$

Label outcome: $z_i = 1(y_i = \hat{y}_i)$

BAYESIAN ASSESSMENT: HOW ACCURATE



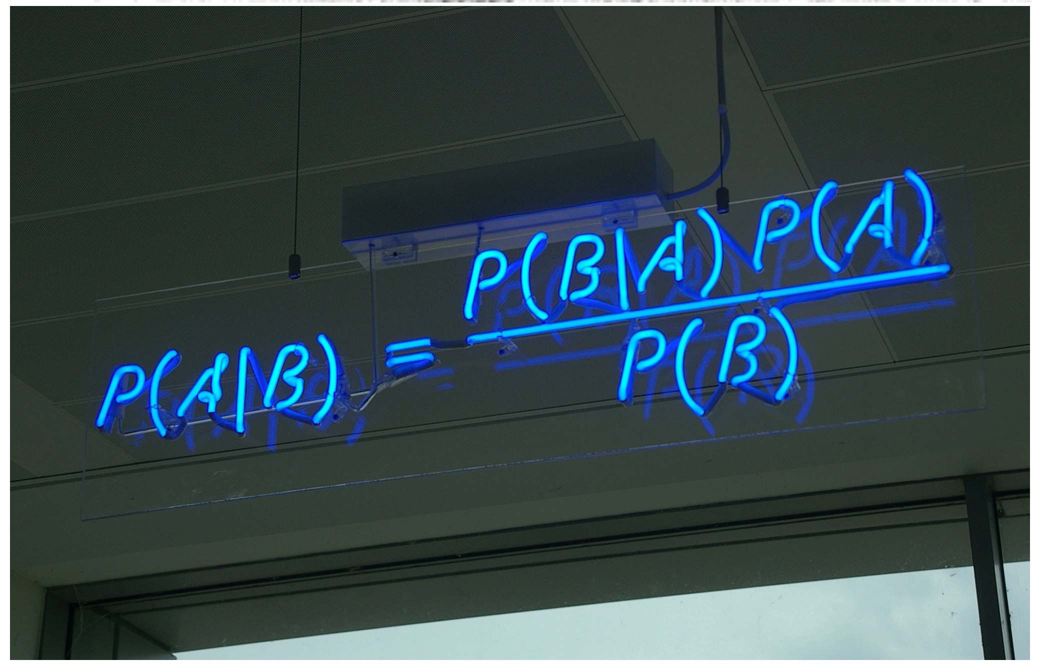
Performance metric of interest: θ

Labeled data: $D = \{(x_i, y_i) \,|\, i = 1,2,\cdots,N\}$

Label outcome: $z_i = 1(y_i = \hat{y}_i)$

$$p(\theta|\mathcal{D}) = \frac{p(\theta) \cdot \prod_{i=1}^{N} q_\theta(z_i)}{\int_\theta p(\theta) \cdot \prod_{i=1}^{N} q_\theta(z_i)\, d\theta}$$

BAYESIAN ASSESSMENT: HOW ACCURATE



Performance metric of interest: θ

Labeled data: $D = \{(x_i, y_i) \,|\, i = 1, 2, \cdots, N\}$

Label outcome: $z_i = 1(y_i = \hat{y}_i)$

Beta posterior

Beta prior

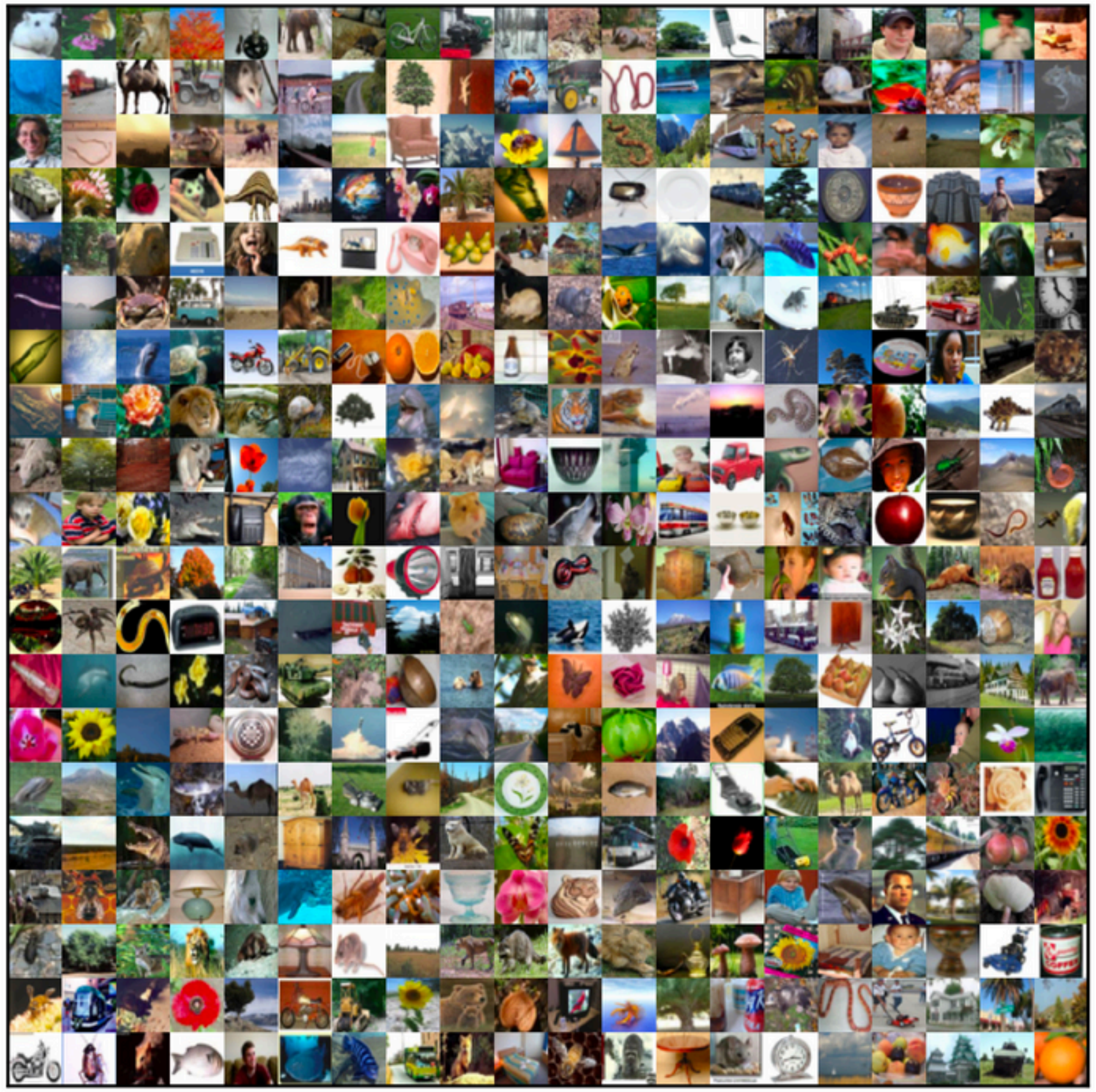
Bernoulli likelihood

$$p(\theta|\mathcal{D}) = \frac{p(\theta) \cdot \prod_{i=1}^{N} q_\theta(z_i)}{\int_\theta p(\theta) \cdot \prod_{i=1}^{N} q_\theta(z_i)\, d\theta}$$

Accuracy $\theta = \mathbb{E}_{p(x,y)} 1(y = \hat{y})$

BAYESIAN ASSESSMENT: HOW ACCURATE

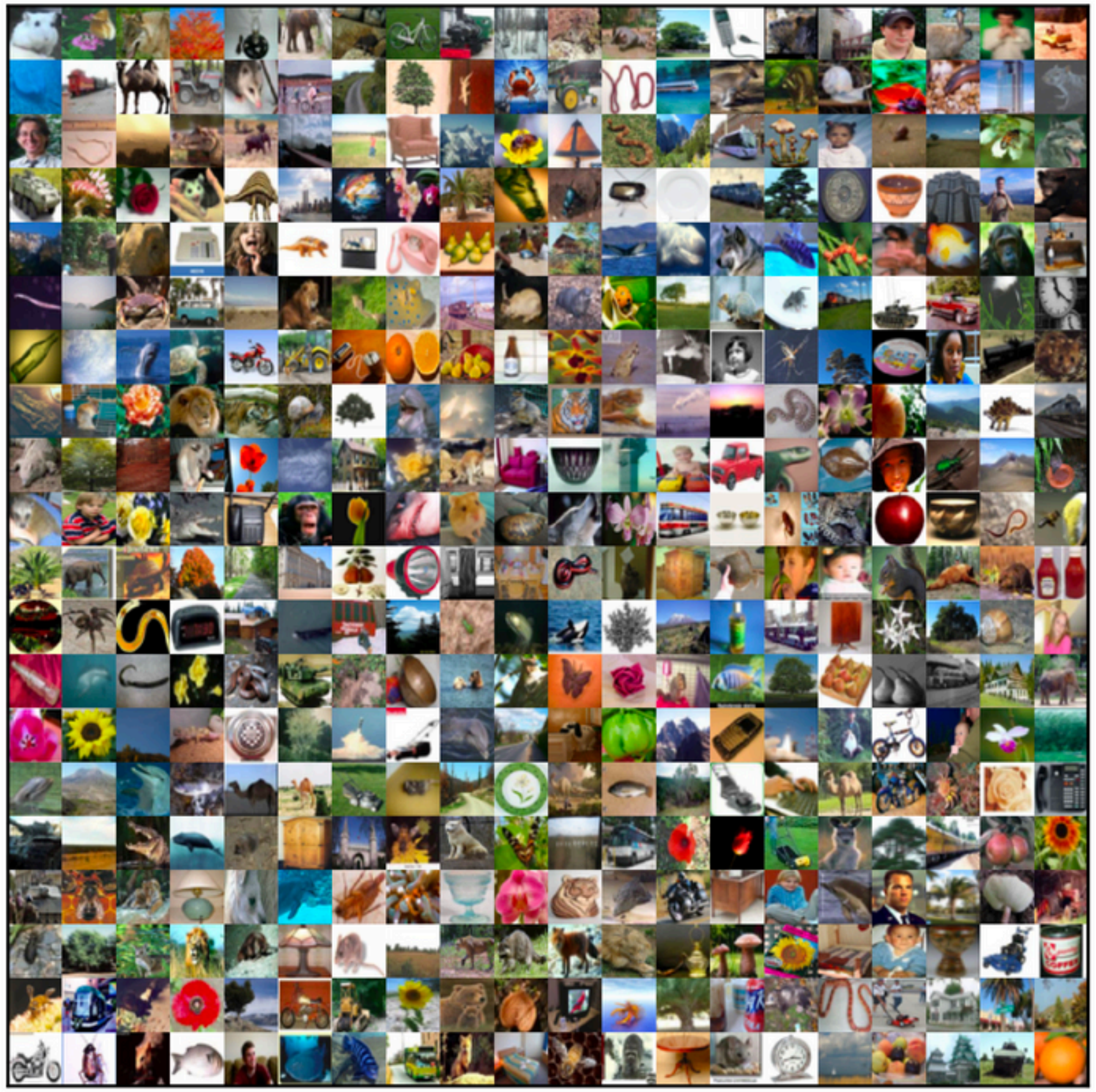
- **CIFAR100**
  - 100 balanced classes
  - 50,000 images for training
  - 10,000 images for testing
  - prediction model: the ResNet model with110 layers
  - overall accuracy on all test data: ~80%



BAYESIAN ASSESSMENT: HOW ACCURATE

- **CIFAR100**
  - 100 balanced classes
  - 50,000 images for training
  - 10,000 images for testing
  - prediction model: the ResNet model with110 layers
  - overall accuracy on all test data: ~80%

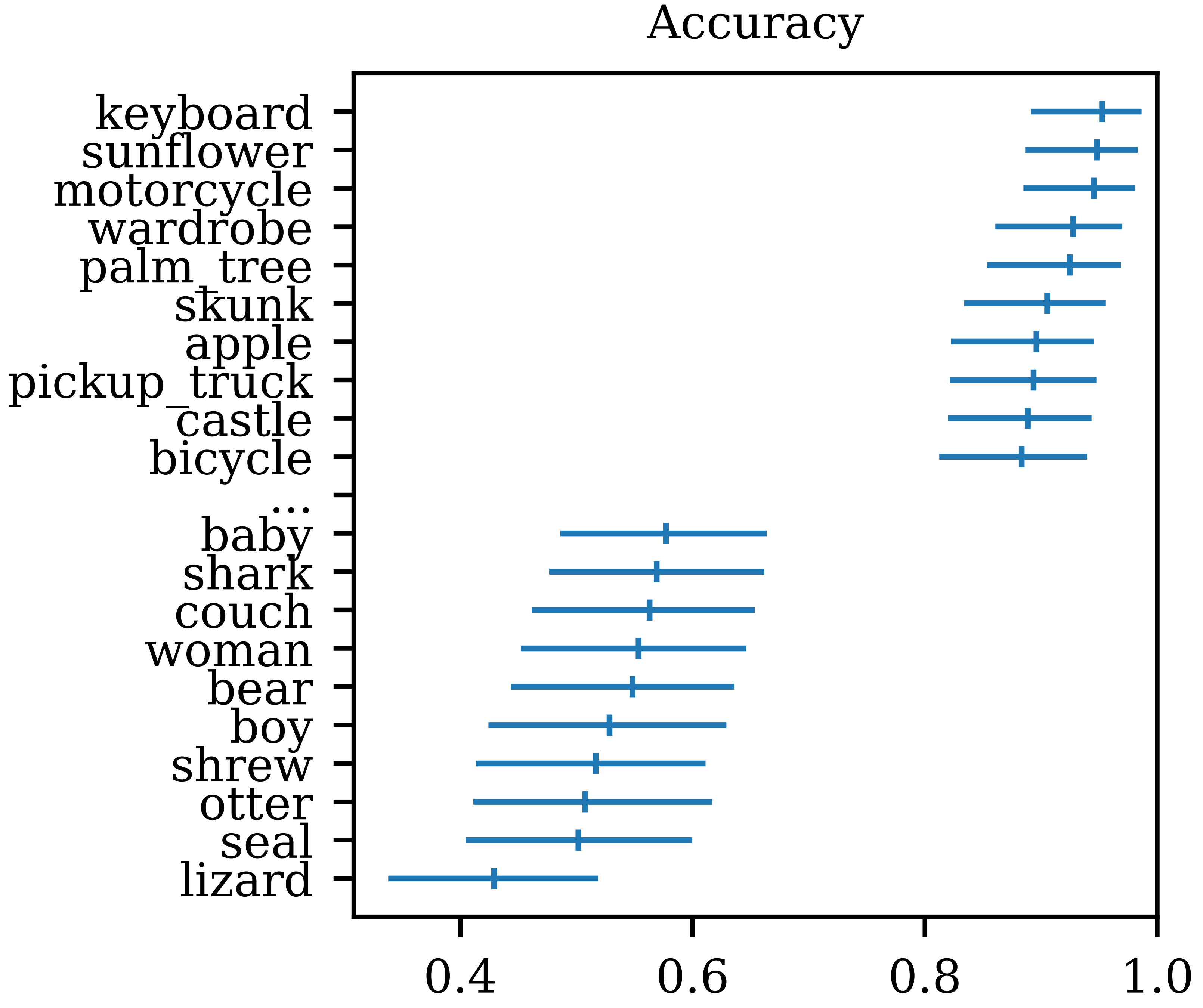
Predicted as Tiger with score $p(\hat{y} \mid x) = 0.99$



BAYESIAN ASSESSMENT: HOW ACCURATE

Accuracy of the k-th predicted class:

$\theta_k = \text{Beta}(\alpha_k, \beta_k), k = 1,2,\cdots,K$



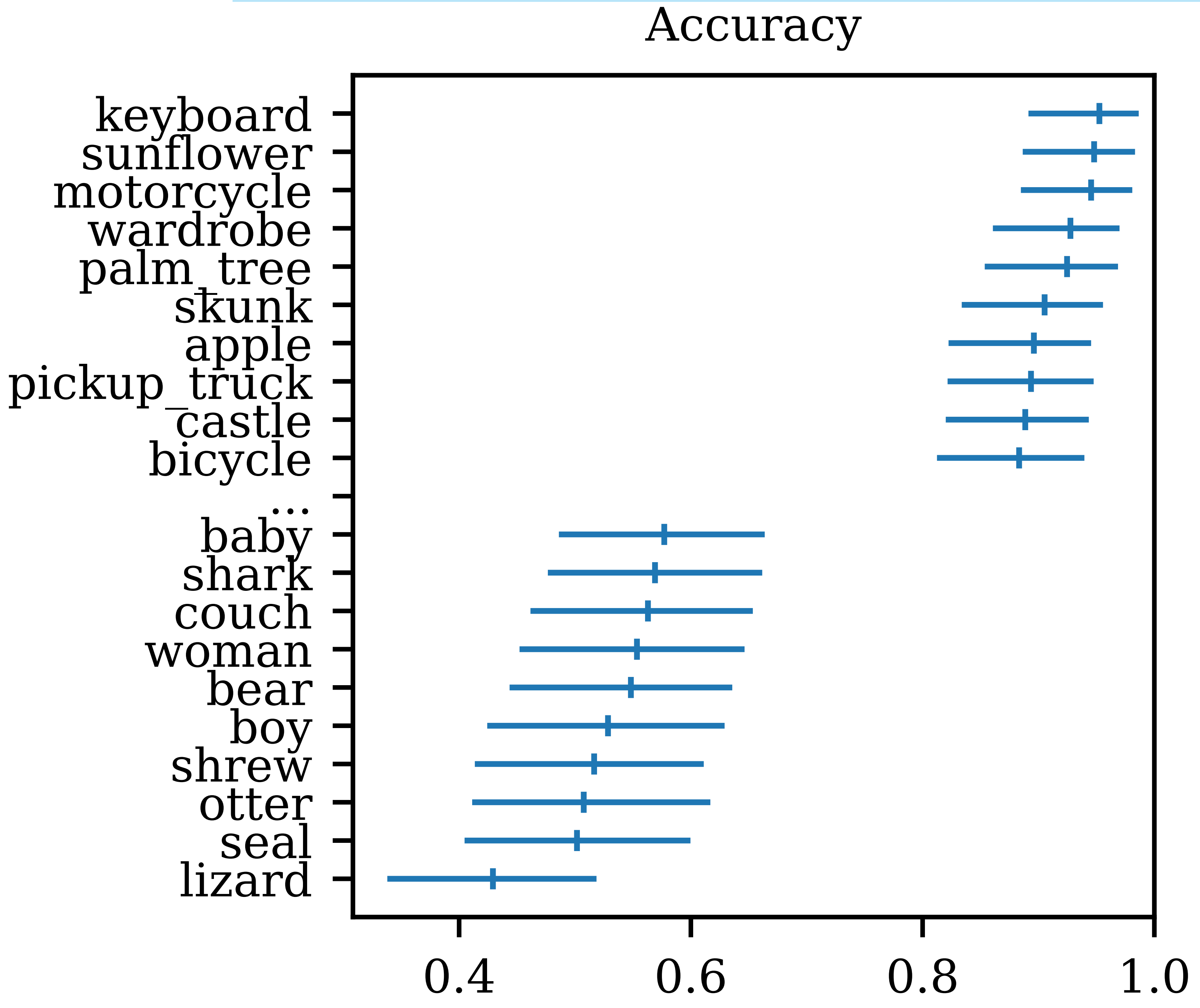
Classwise accuracy

for ResNet-110 on CIFAR-100

BAYESIAN ASSESSMENT: HOW ACCURATE

Accuracy of the k-th predicted class:

$\theta_k = \text{Beta}(\alpha_k, \beta_k), k = 1,2,\cdots,K$

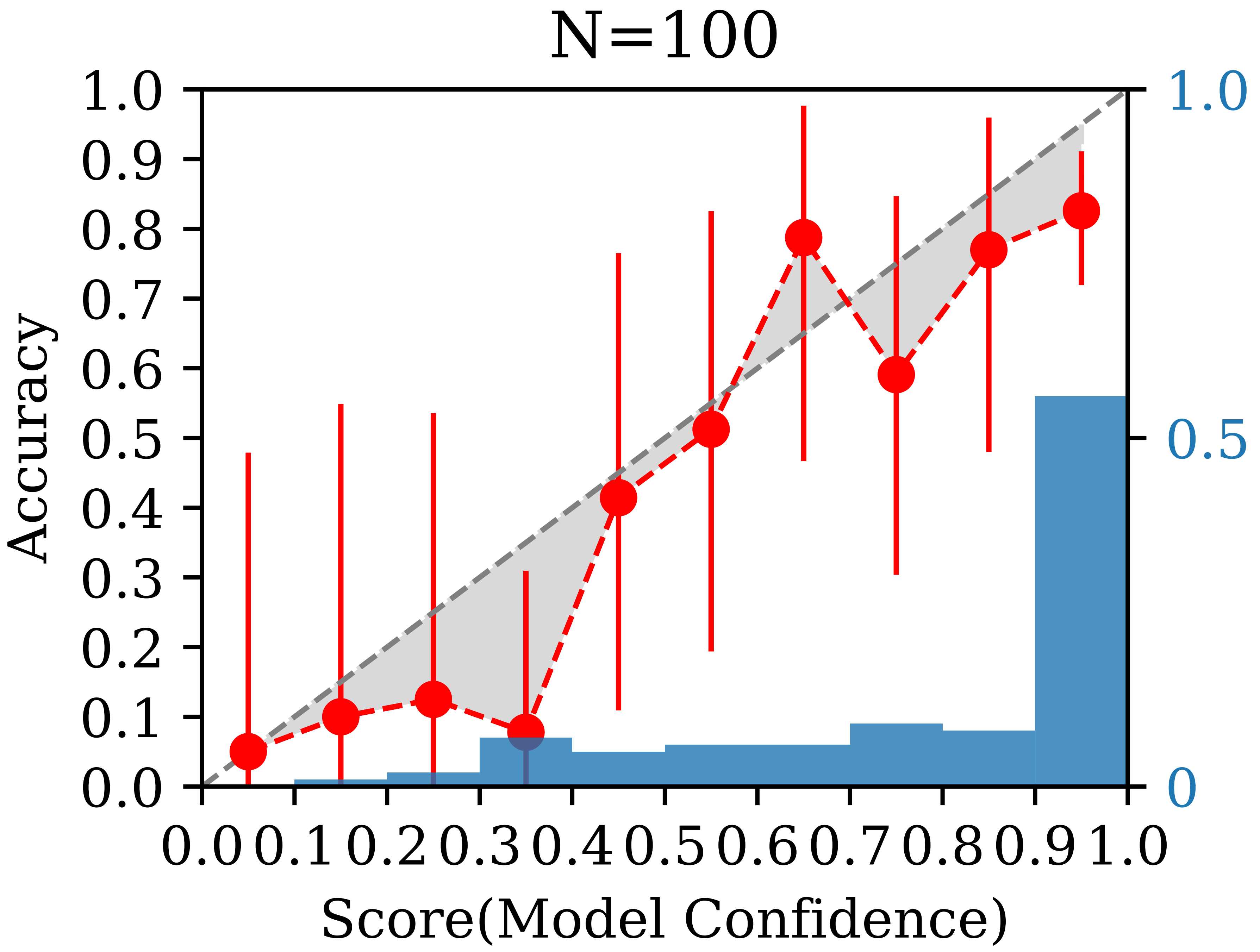


Classwise accuracy

for ResNet-110 on CIFAR-100

Accuracy of the b-th bin:

$\theta_b = \text{Beta}(\alpha_b, \beta_b), b = 1,2,\cdots,B$



Reliability diagram

for ResNet-110 on CIFAR-100

Disi Ji , Robert L. Logan IV , Padhraic Smyth , Mark Steyvers

BAYESIAN ASSESSMENT: ASSESSMENT TASKS

We can obtain $p(\theta_g \mid D)$ for different groupings and performance metrics $\theta = (\theta_1, \theta_2, \cdots, \theta_G)$

Grouped by predicted class, model score etc.
θ can be accuracy, precision, ECE, etc.

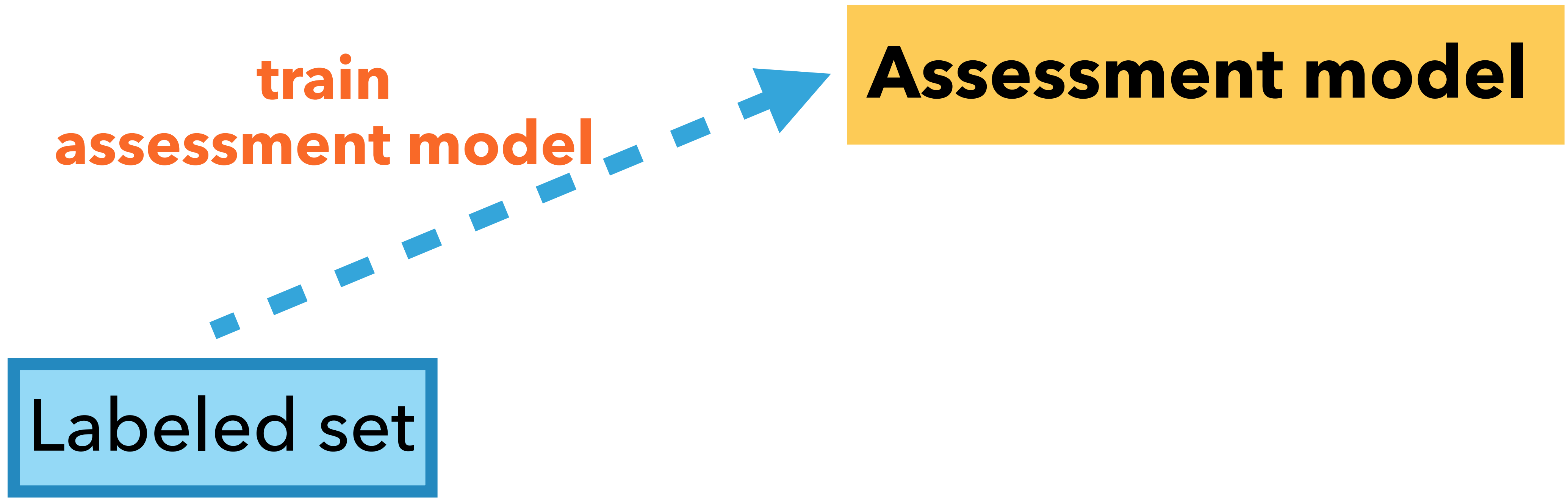
BAYESIAN ASSESSMENT: ASSESSMENT TASKS

We can obtain $p(\theta_g \mid D)$ for different groupings and performance metrics $\theta = (\theta_1, \theta_2, \cdots, \theta_G)$

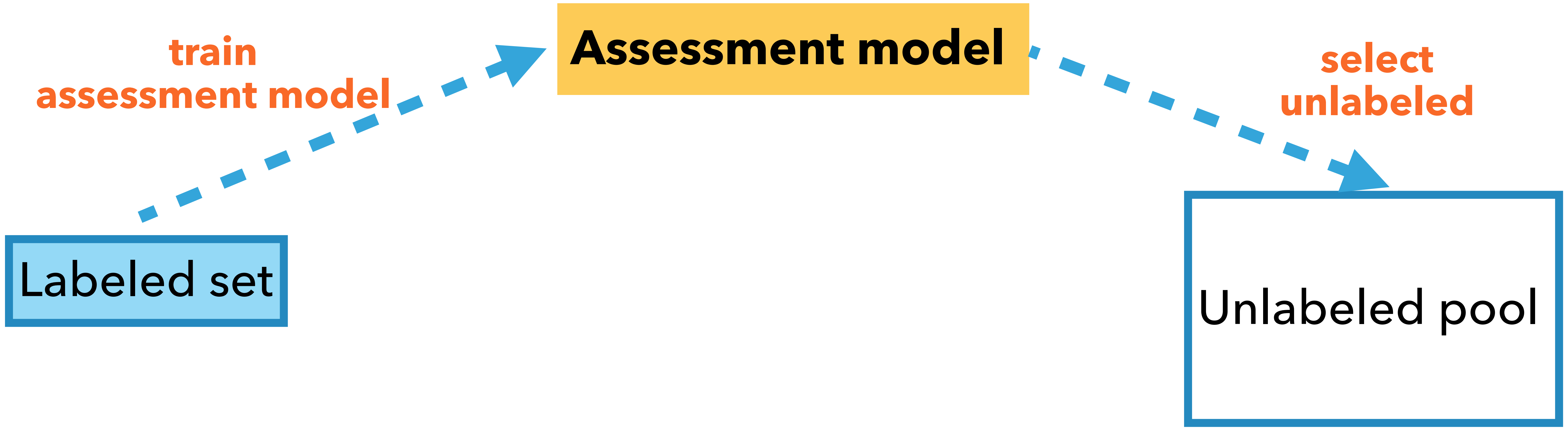
Grouped by predicted class, model score etc.
θ can be accuracy, precision, ECE, etc.

- **Estimate** model performance across all groups
  - e.g. minimize RMSE = $(\sum_g p_g(\hat{\theta}_g - \theta_g^*)^2)^{\frac{1}{2}}$
- **Identify** extreme groups
  - e.g. identify the least accurate group $\hat{g} = \arg\max_g \theta_g$
- **Compare** performance between two groups
  - e.g. $\theta_i > \theta_j$?

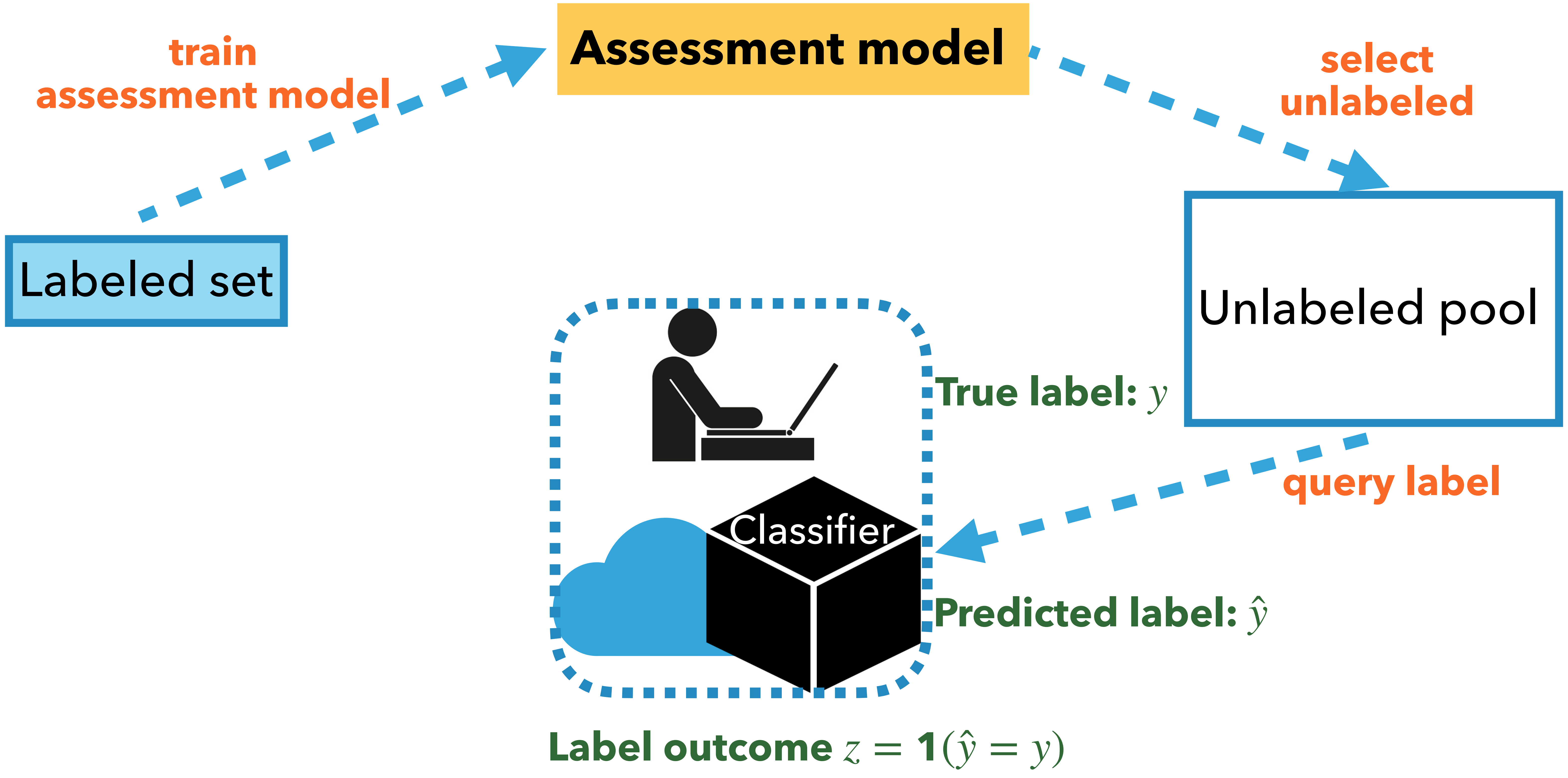
ACTIVE BAYESIAN ASSESSMENT



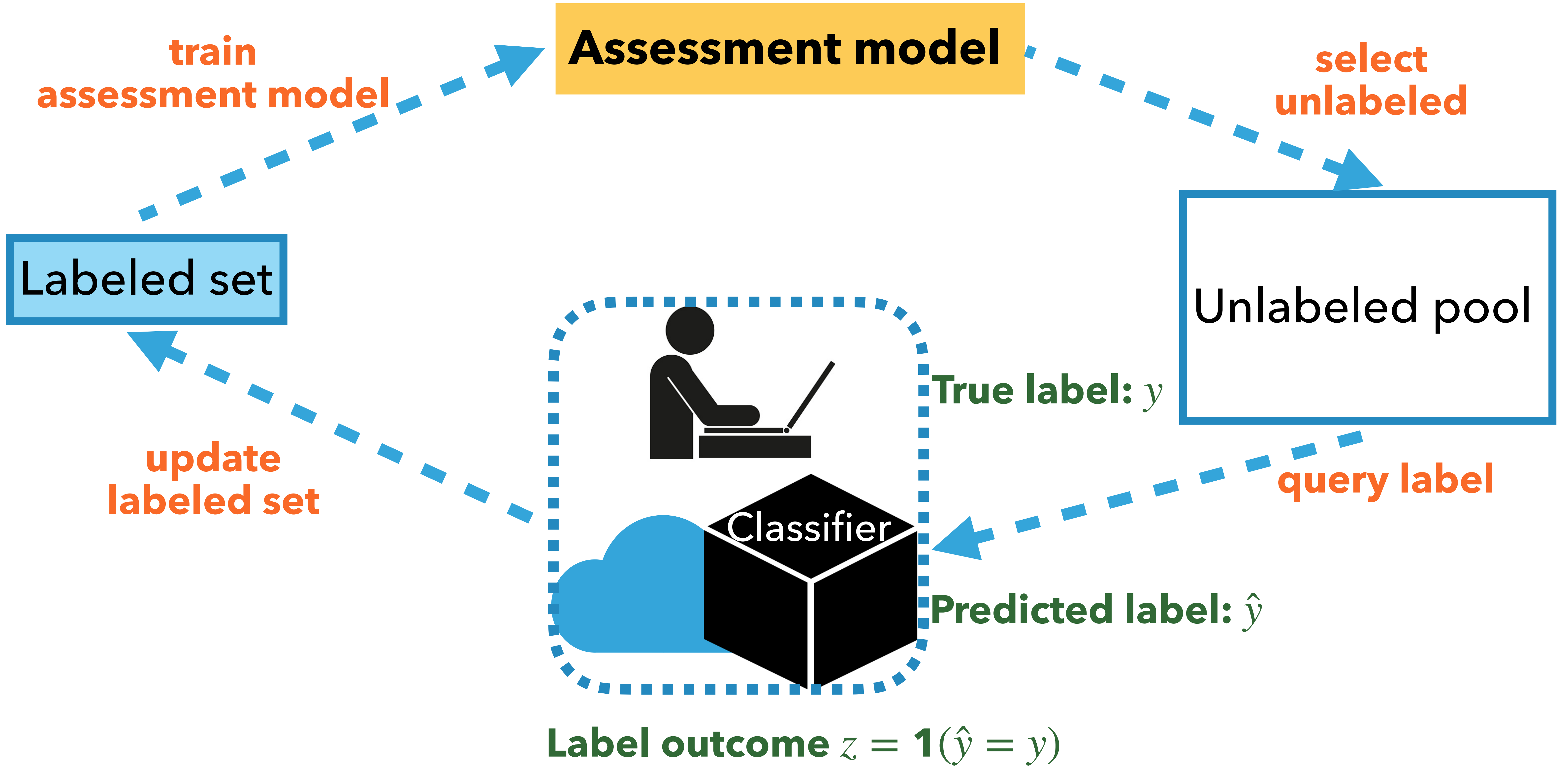
ACTIVE BAYESIAN ASSESSMENT



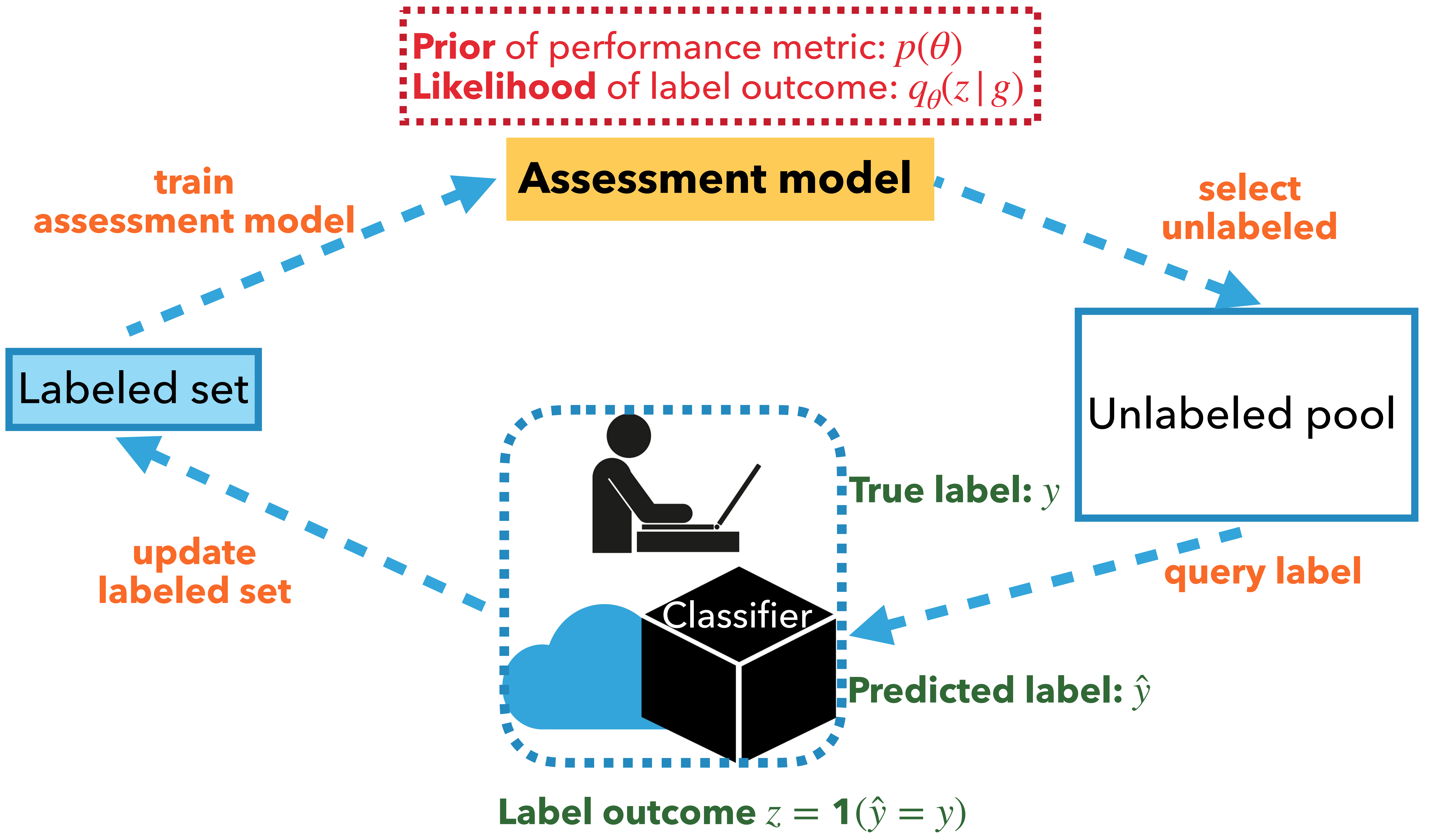
ACTIVE BAYESIAN ASSESSMENT



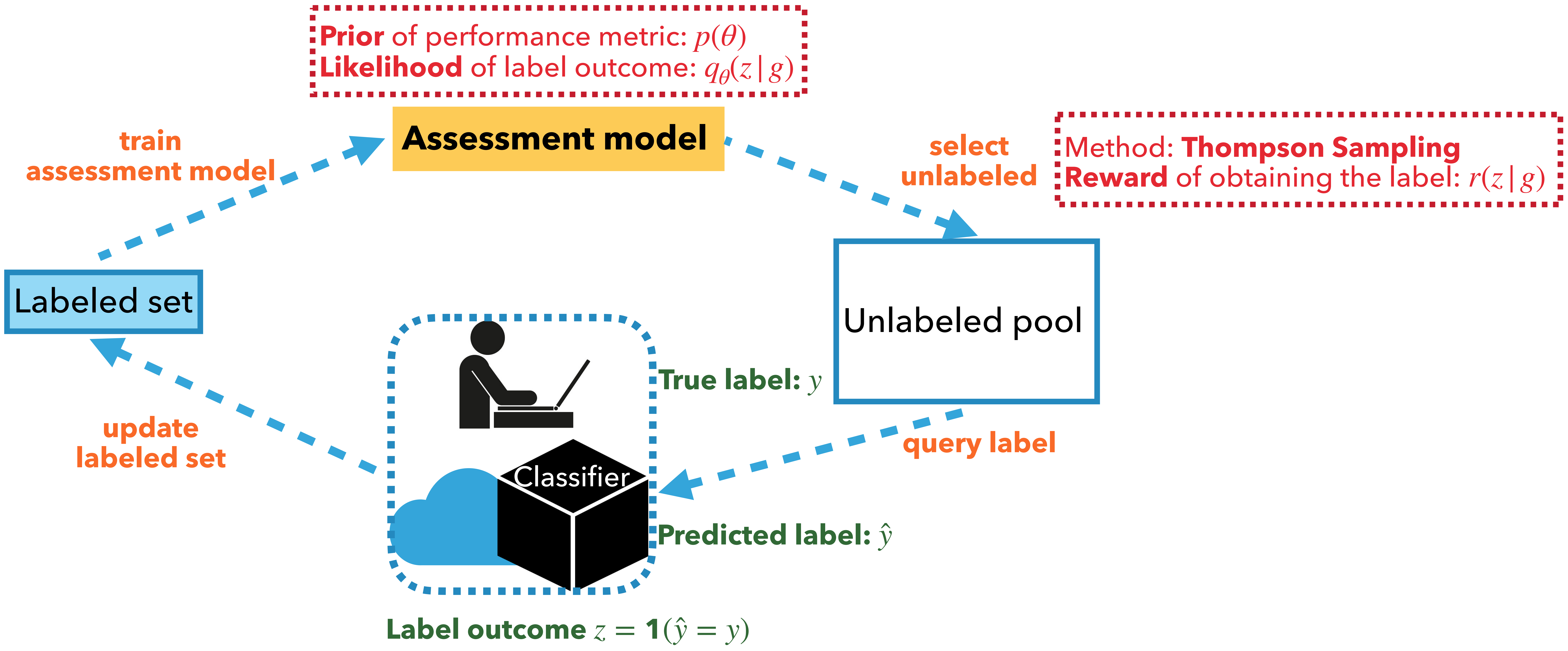
ACTIVE BAYESIAN ASSESSMENT



ACTIVE BAYESIAN ASSESSMENT

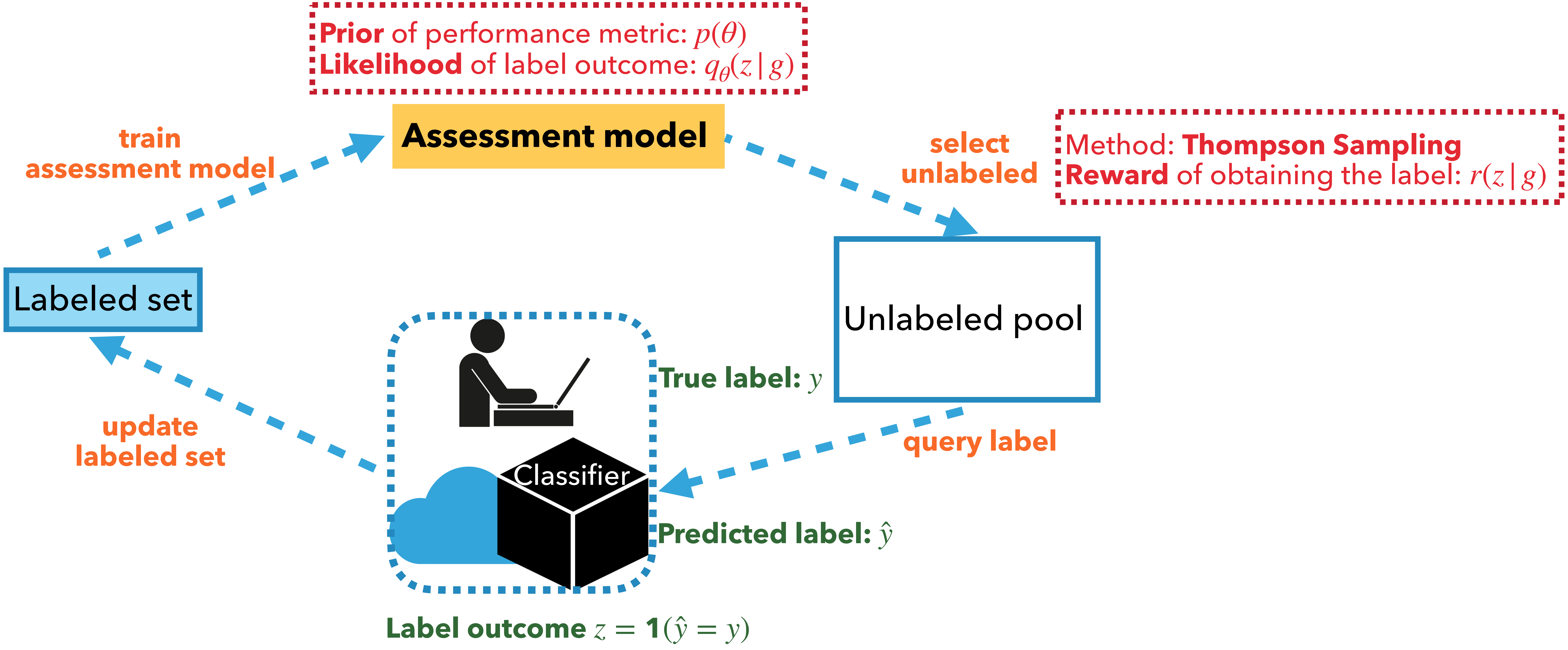


ACTIVE BAYESIAN ASSESSMENT



ACTIVE BAYESIAN ASSESSMENT

Active Bayesian assessment ⟶ Design task-specific (p, q, r)



ACTIVE BAYESIAN ASSESSMENT

	Assessment Task	$p(\theta)$	$q_\theta(z\|g)$	$r(z\|g)$
Estimation	Groupwise Accuracy	$\theta_g \sim \mathrm{Beta}(\alpha_g, \beta_g)$	$z \sim \mathrm{Bern}(\theta_g)$	$p_g \cdot (\mathrm{Var}(\hat{\theta}_g\|\mathcal{L}) - \mathrm{Var}(\hat{\theta}_g\|\{\mathcal{L}, z\}))$
	Confusion Matrix$(g = k)$	$\theta_{\cdot k} \sim \mathrm{Dirichlet}(\alpha_{\cdot k})$	$z \sim \mathrm{Multi}(\theta_k)$	$p_k \cdot (\mathrm{Var}(\hat{\theta}_k\|\mathcal{L}) - \mathrm{Var}(\hat{\theta}_k\|\{\mathcal{L}, z\}))$
Identification	Least Accurate Group	$\theta_g \sim \mathrm{Beta}(\alpha_g, \beta_g)$	$z \sim \mathrm{Bern}(\theta_g)$	$-\widetilde{\theta}_g$
	Least Calibrated Group	$\theta_{gb} \sim \mathrm{Beta}(\alpha_{gb}, \beta_{gb})$	$z \sim \mathrm{Bern}(\theta_{gb})$	$\sum_{b=1}^{B} p_{gb} \left\|\widetilde{\theta}_{gb} - s_{gb}\right\|$
	Most Costly Class$(g = k)$	$\theta_{\cdot k} \sim \mathrm{Dirichlet}(\alpha_{\cdot k})$	$z \sim \mathrm{Multi}(\theta_k)$	$\sum_{j=1}^{K} c_{jk} \widetilde{\theta}_{jk}$
Comparison	Accuracy Comparison	$\theta_g \sim \mathrm{Beta}(\alpha_g, \beta_g)$	$z \sim \mathrm{Bern}(\theta_g)$	$\lambda\|\{\mathcal{L}, (g, z)\}$

ACTIVE BAYESIAN ASSESSMENT

	Assessment Task	$p(\theta)$	$q_\theta(z\|g)$	$r(z\|g)$
Estimation	Groupwise Accuracy	$\theta_g \sim \text{Beta}(\alpha_g, \beta_g)$	$z \sim \text{Bern}(\theta_g)$	$p_g \cdot (\text{Var}(\hat{\theta}_g \| \mathcal{L}) - \text{Var}(\hat{\theta}_g \| \{\mathcal{L}, z\}))$

ACTIVE BAYESIAN ASSESSMENT

	Assessment Task	$p(\theta)$	$q_\theta(z\|g)$	$r(z\|g)$
Estimation	Groupwise Accuracy	$\theta_g \sim \mathrm{Beta}(\alpha_g, \beta_g)$	$z \sim \mathrm{Bern}(\theta_g)$	$p_g \cdot (\mathrm{Var}(\hat{\theta}_g\|\mathcal{L}) - \mathrm{Var}(\hat{\theta}_g\|\{\mathcal{L}, z\}))$

Prior of groupwise accuracy **Binary label outcome**

ACTIVE BAYESIAN ASSESSMENT

	Assessment Task	$p(\theta)$	$q_\theta(z\|g)$	$r(z\|g)$
Estimation	Groupwise Accuracy	$\theta_g \sim \mathrm{Beta}(\alpha_g, \beta_g)$	$z \sim \mathrm{Bern}(\theta_g)$	$p_g \cdot (\mathrm{Var}(\hat{\theta}_g\|\mathcal{L}) - \mathrm{Var}(\hat{\theta}_g\|\{\mathcal{L}, z\}))$

Prior of groupwise accuracy

Binary label outcome

previously labeled data

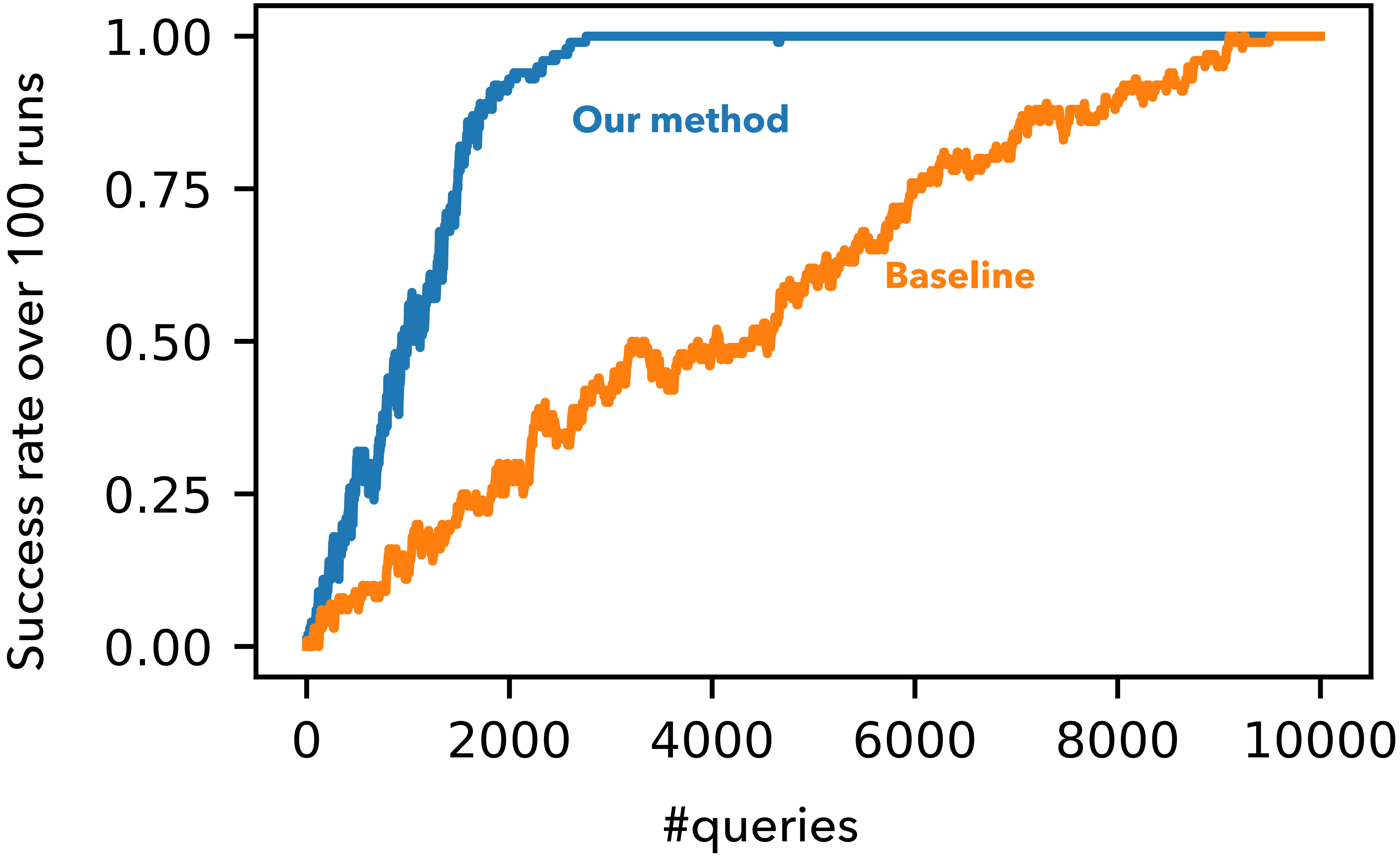
$$r(z\,|\,g) = p_g \cdot (\underbrace{\mathrm{Var}(\hat{\theta}_g\,|\,\mathscr{L}) - \mathrm{Var}(\hat{\theta}_g\,|\,\{\mathscr{L}, z\})}_{\text{reduction of posterior variance}})$$

group probability

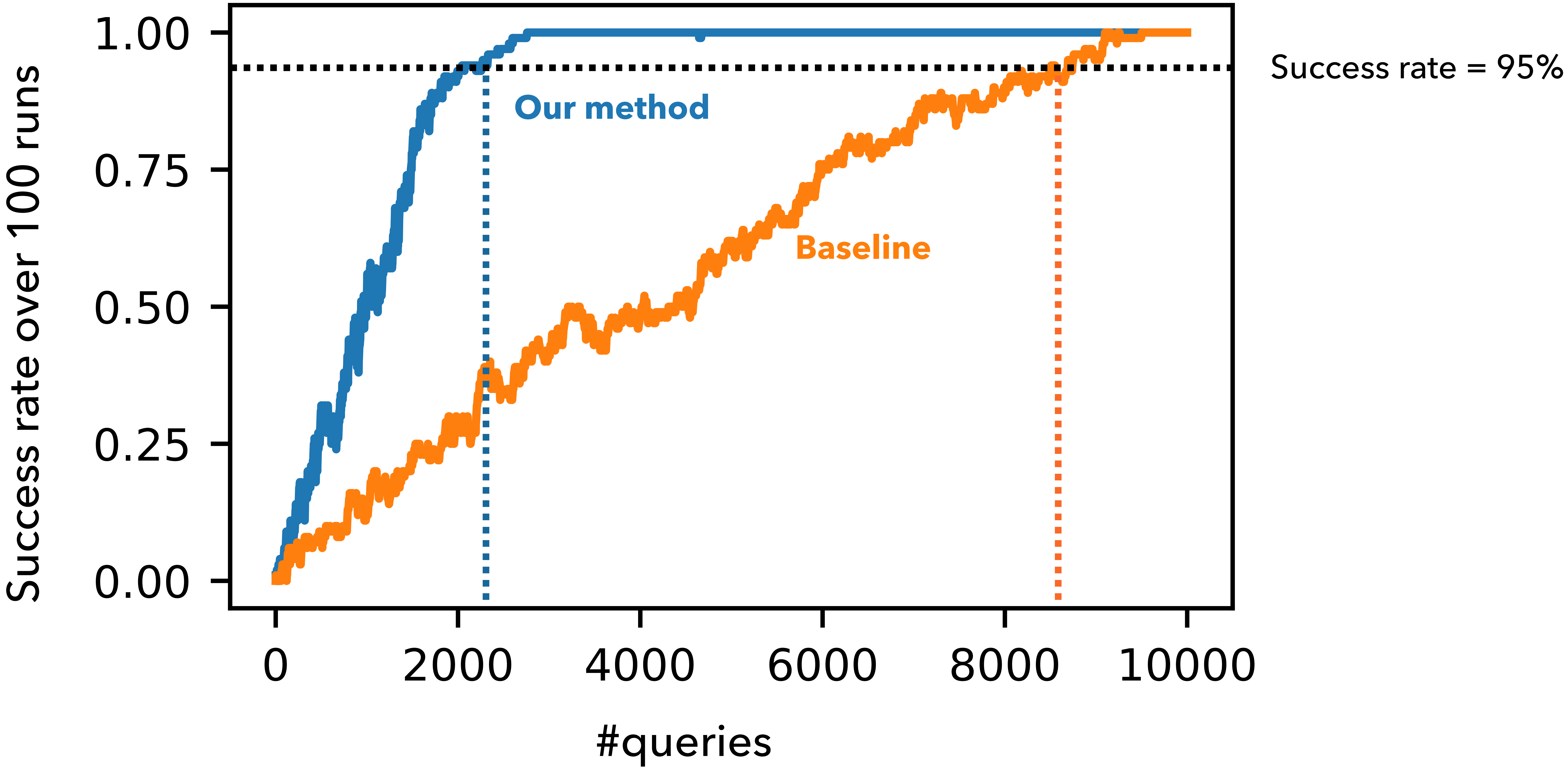
reduction of posterior variance

Maximal expected model change strategy [*Freytag et al., 2014, Vezhnevets et al., 2012*]

ACTIVELY IDENTIFY THE LEAST ACCURATE CLASS OF CIFAR100



ACTIVELY IDENTIFY THE LEAST ACCURATE CLASS OF CIFAR100



Disi Ji , Robert L. Logan IV , Padhraic Smyth , Mark Steyvers

UCIRVINE

ACTIVELY IDENTIFY THE LEAST ACCURATE CLASS

- Datasets with varying size and number of classes

	Mode	Size	Classes	Model
CIFAR-100	Image	10K	100	ResNet-110
ImageNet	Image	50K	1000	ResNet-152
SVHN	Image	26K	10	ResNet-152
20 Newsgroups	Text	7.5K	20	$\text{BERT}_{\text{BASE}}$
DBpedia	Text	70K	14	$\text{BERT}_{\text{BASE}}$

ACTIVELY IDENTIFY THE LEAST ACCURATE CLASS

Percentage of labeled samples needed to identify the least accurate classes

Dataset	Top m	UPrior (baseline)	IPrior (our work)	IPrior+TS (our work)
CIFAR-100	1	81.1	83.4	**24.9**
	10	99.8	99.8	**55.1**
ImageNet	1	96.9	94.7	**9.3**
	10	99.6	98.5	**17.1**
SVHN	1	90.5	89.8	**82.8**
	3	100.0	100.0	**96.0**
20 Newsgroups	1	53.9	55.4	**16.9**
	3	92.0	92.5	**42.5**
DBpedia	1	8.0	**7.6**	11.6
	3	91.9	90.2	**57.1**

Dropped by 71%

Dropped by 90%

ACTIVELY IDENTIFY THE LEAST ACCURATE CLASS

Percentage of labeled samples needed to identify the least accurate classes

Dataset	Top m	UPrior (baseline)	IPrior (our work)	IPrior+TS (our work)	
CIFAR-100	1	81.1	83.4	**24.9**	Dropped by 71%
	10	99.8	99.8	**55.1**	
ImageNet	1	96.9	94.7	**9.3**	Dropped by 90%
	10	99.6	98.5	**17.1**	
SVHN	1	90.5	89.8	**82.8**	
	3	100.0	100.0	**96.0**	
20 Newsgroups	1	53.9	55.4	**16.9**	
	3	92.0	92.5	**42.5**	
DBpedia	1	8.0	**7.6**	11.6	
	3	91.9	90.2	**57.1**	

We obtained similar performance gain for other assessment tasks! (full results in paper)

DISCUSSION

- **Other Bayesian active learning method to TS?**
  - e.g. Epsilon-greedy, Bayesian upper-confidence bound
  - e.g. top-two TS (TTTS) [*Russo, 2016*], multi-play TS (MPTS) [*Komiyama et al. 2015*]
  - Thompson sampling is broadly more reliable and more consistent

- **Sensitivity analysis** for hyperparameters
  - appears to be relatively robust to the prior strength

OUR CONTRIBUTIONS

- Developed a general **Bayesian framework** to assess classification performance metrics, including
  - (1) accuracy, reliability diagram, ECE;
  - (2) performance difference;
  - (3) confusion matrix, misclassification cost, etc
- Developed an **active assessment framework** for
  - (1) estimation of model performance;
  - (2) identification of model deficiencies;
  - (3) performance comparison between groups
- Demonstrated that our proposed approaches need significantly fewer labels than baselines

RELATED WORK

- **Prior related work on Bayesian assessment has focused on much more specific metrics and tasks:**
  - Goutte et al. [2005]: Bayesian estimation of precision, recall, and F-score in information retrieval
  - Benavoli et al. [2017]: Bayesian framework for comparing multiple classifiers
  - Johnson et al. [2019]: Bayesian mixture models of diagnostic metrics for medical tests
  - etc…

- **Prior related work of label-efficient assessment are mostly non-Bayesian or use a narrower set of metrics:**
  - Kumar and Raj [2008]: stratified sampling for risk estimation
  - Sawade et al. [2010]: importance sampling for risk estimation
  - Nguyen et al. [2018]: assess with large scale noisy labels for applications in computer vision
  - Ji et al. [2020]: used Bayesian estimation with scores from unlabeled data to assess group fairness

THANK YOU FOR LISTENING!