LABEL-EFFICIENT BAYESIAN ASSESSMENT OF BLACK-BOX CLASSIFIERS

Disi Ji

November 18, 2020

Department of Computer Science

University of California, Irvine

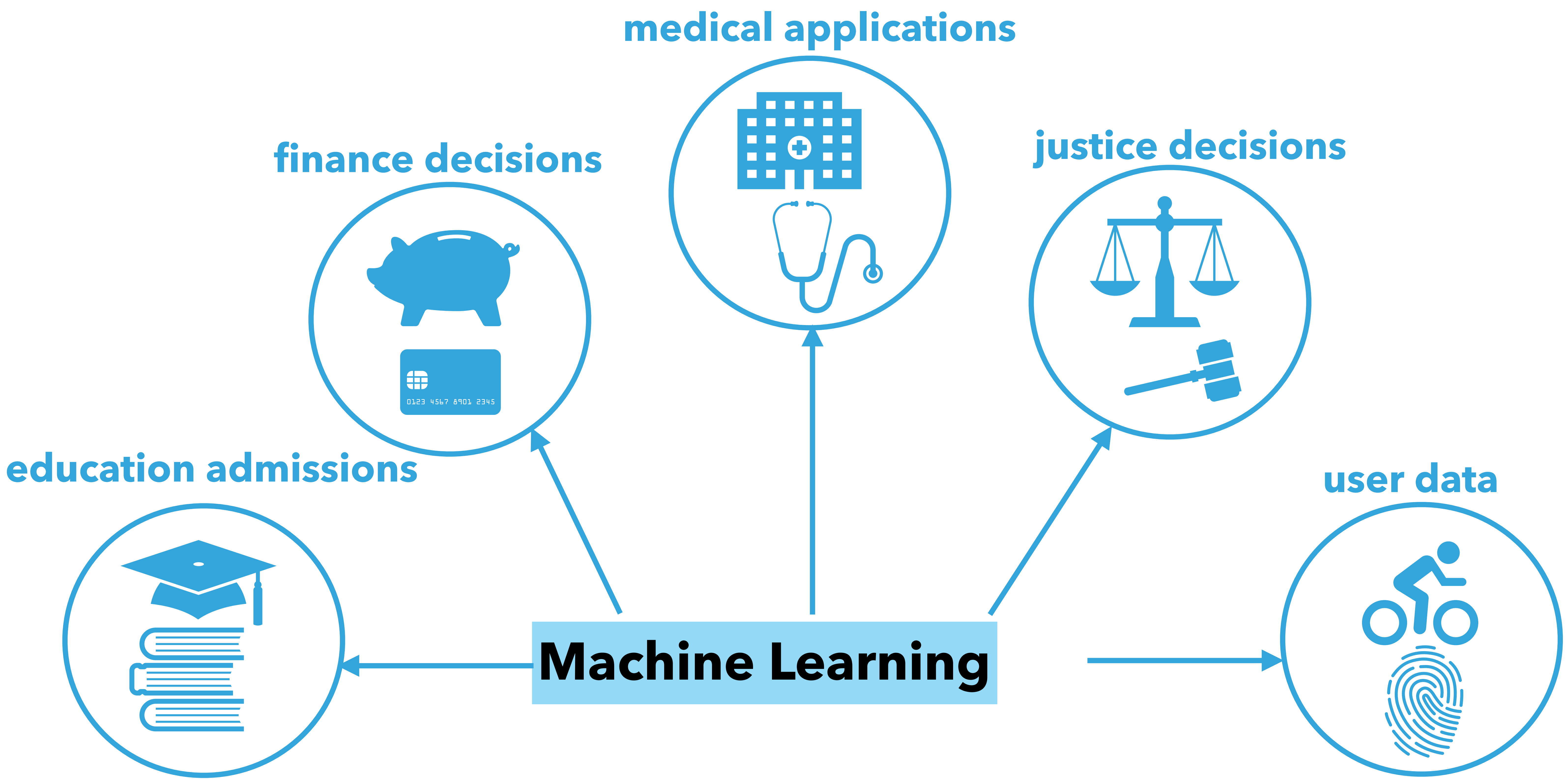
Committee members:

Chancellor's Professor Padhraic Smyth, chair

Assistant Professor Stephan Mandt

Professor Mark Steyvers

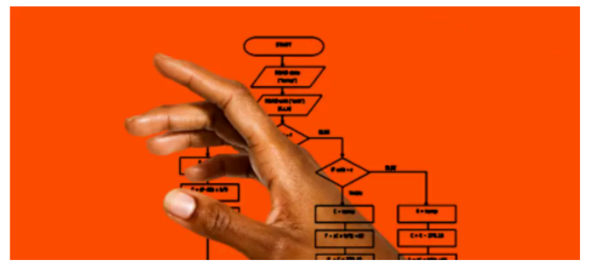
BACKGROUND


medical applications
finance decisions
justice decisions
education admissions
user data
Machine Learning

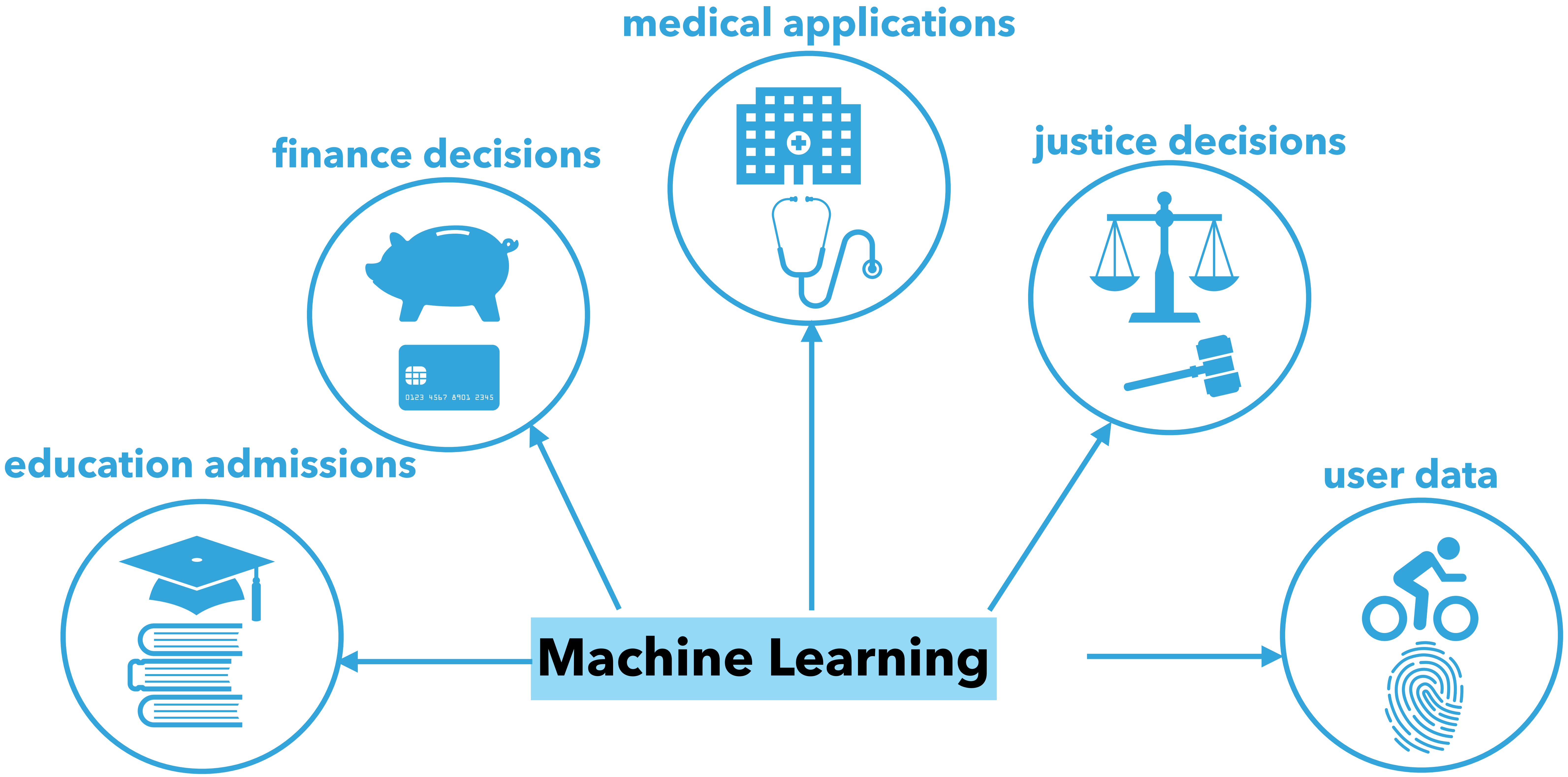
BACKGROUND

The lessons we all must learn from the A-levels algorithm debacle

Unless action is taken, similar systems will suffer from the same mistakes. And the consequences could be dire



By MATT BURGESS



BACKGROUND

ECONOMICS & SOCIETY

AI Can Make Bank Loans More Fair

by Sian Townson

November 06, 2020

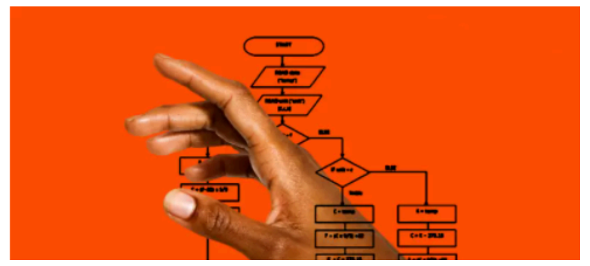
Summary Save Share Print $8.95 Buy Copies

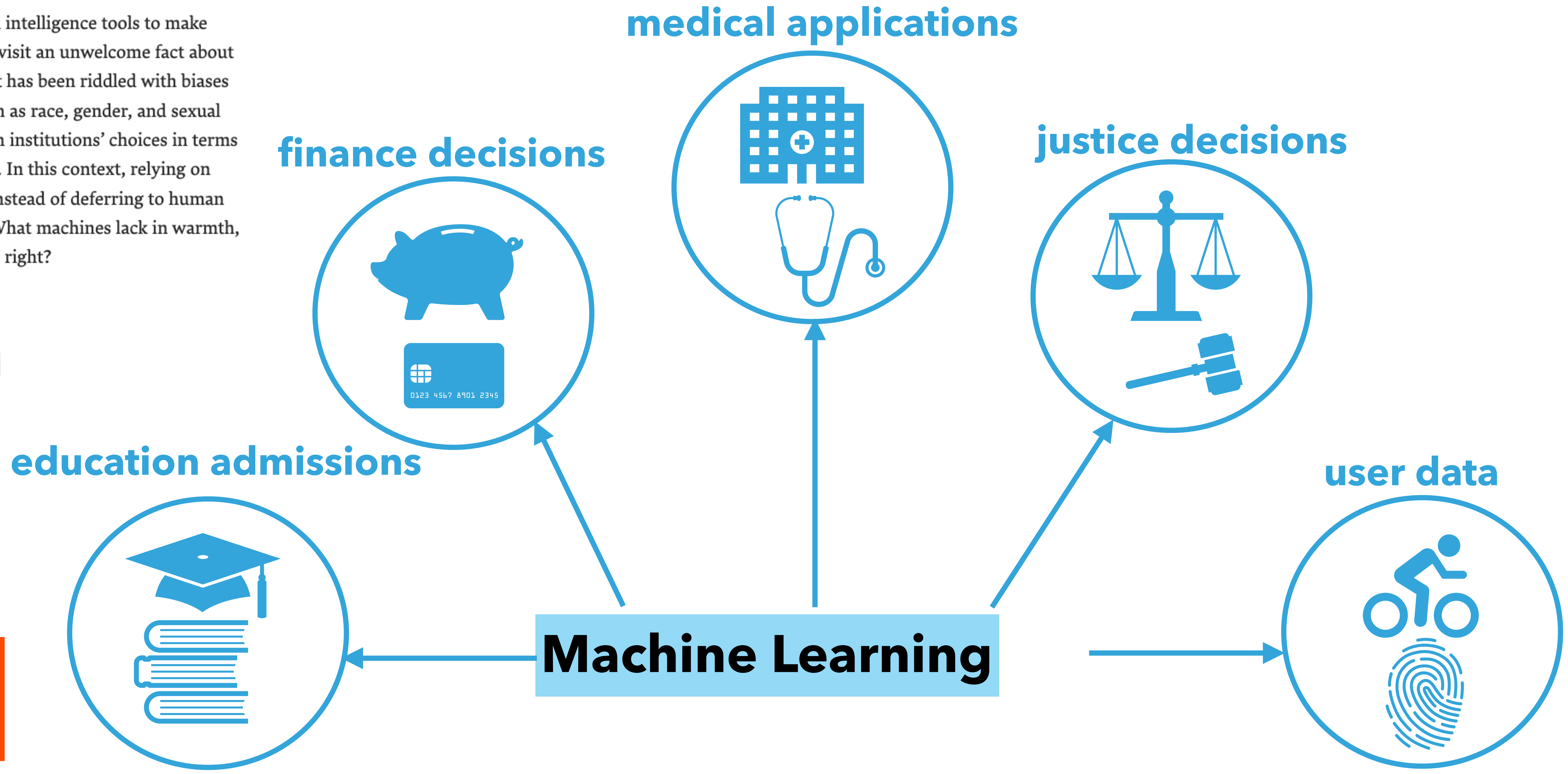
As banks increasingly deploy artificial intelligence tools to make credit decisions, they are having to revisit an unwelcome fact about the practice of lending: Historically, it has been riddled with biases against protected characteristics, such as race, gender, and sexual orientation. Such biases are evident in institutions' choices in terms of who gets credit and on what terms. In this context, relying on algorithms to make credit decisions instead of deferring to human judgment seems like an obvious fix. What machines lack in warmth, they surely make up for in objectivity, right?

The lessons we all must learn from the A-levels algorithm debacle

Unless action is taken, similar systems will suffer from the same mistakes. And the consequences could be dire

By MATT BURGESS





BACKGROUND

ECONOMICS & SOCIETY

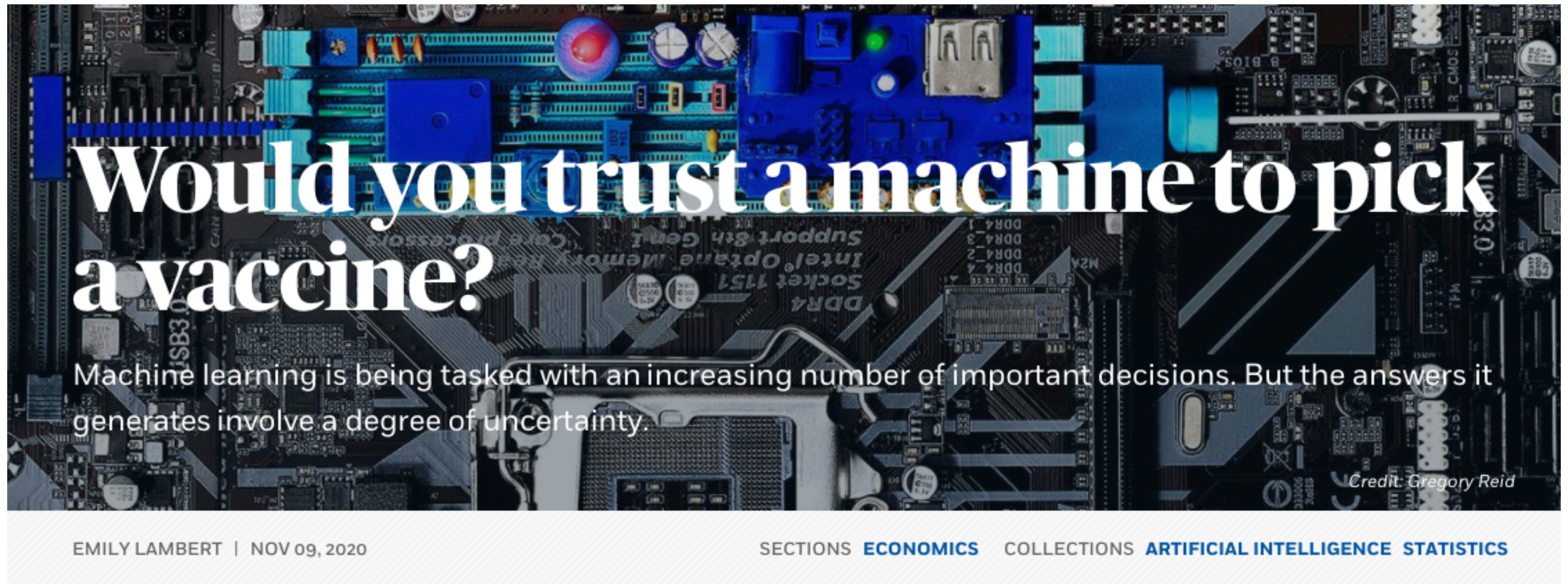
AI Can Make Bank Loans More Fair

by Sian Townson

November 06, 2020

Summary Save Share Print $8.95 Buy Copies

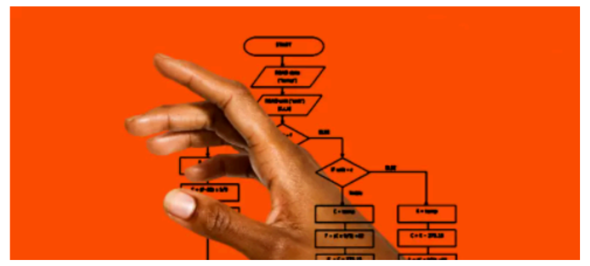
As banks increasingly deploy artificial intelligence tools to make credit decisions, they are having to revisit an unwelcome fact about the practice of lending: Historically, it has been riddled with biases against protected characteristics, such as race, gender, and sexual orientation. Such biases are evident in institutions' choices in terms of who gets credit and on what terms. In this context, relying on algorithms to make credit decisions instead of deferring to human judgment seems like an obvious fix. What machines lack in warmth, they surely make up for in objectivity, right?

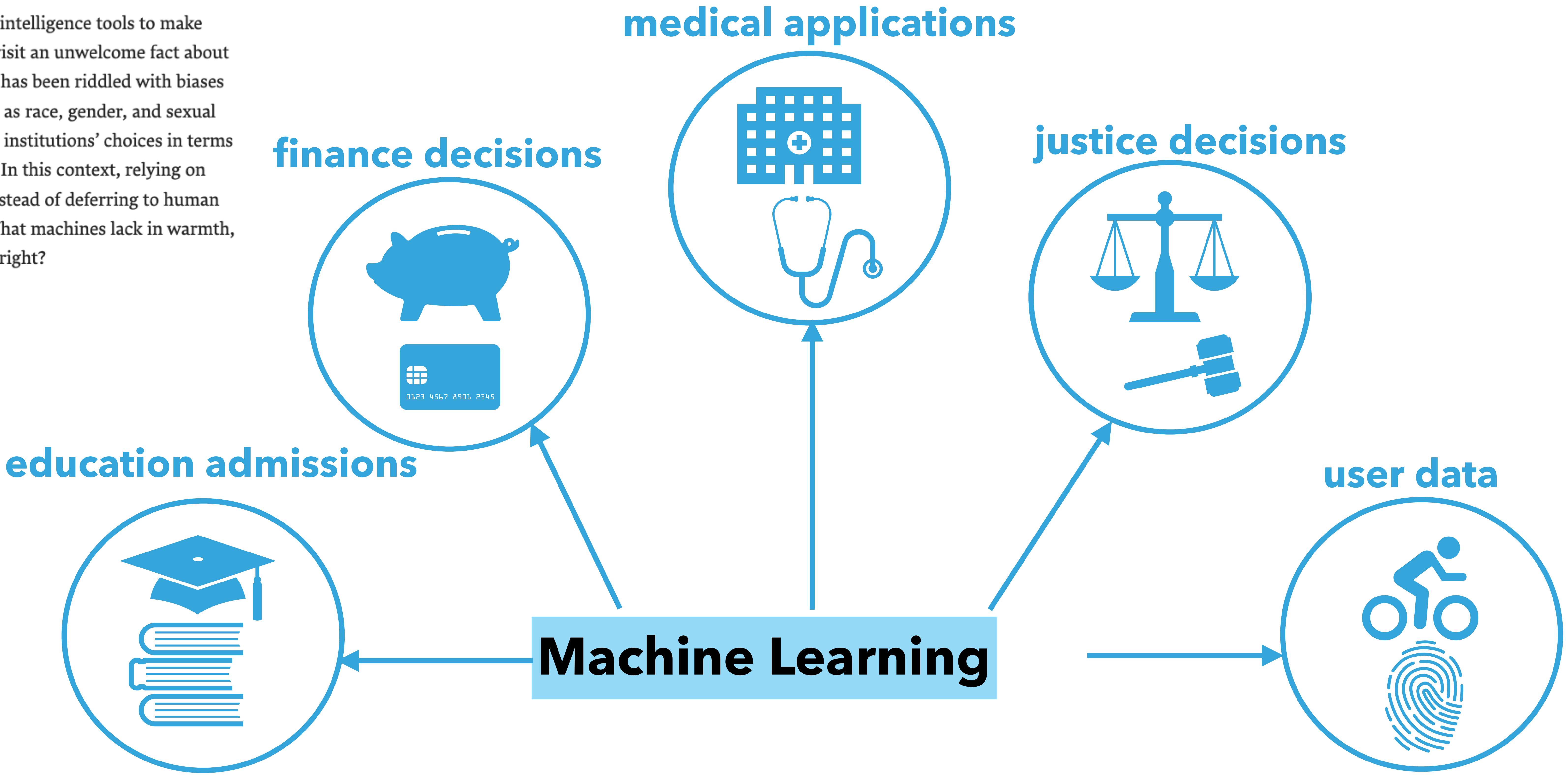


The lessons we all must learn from the A-levels algorithm debacle

Unless action is taken, similar systems will suffer from the same mistakes. And the consequences could be dire

By MATT BURGESS





BACKGROUND

ECONOMICS & SOCIETY

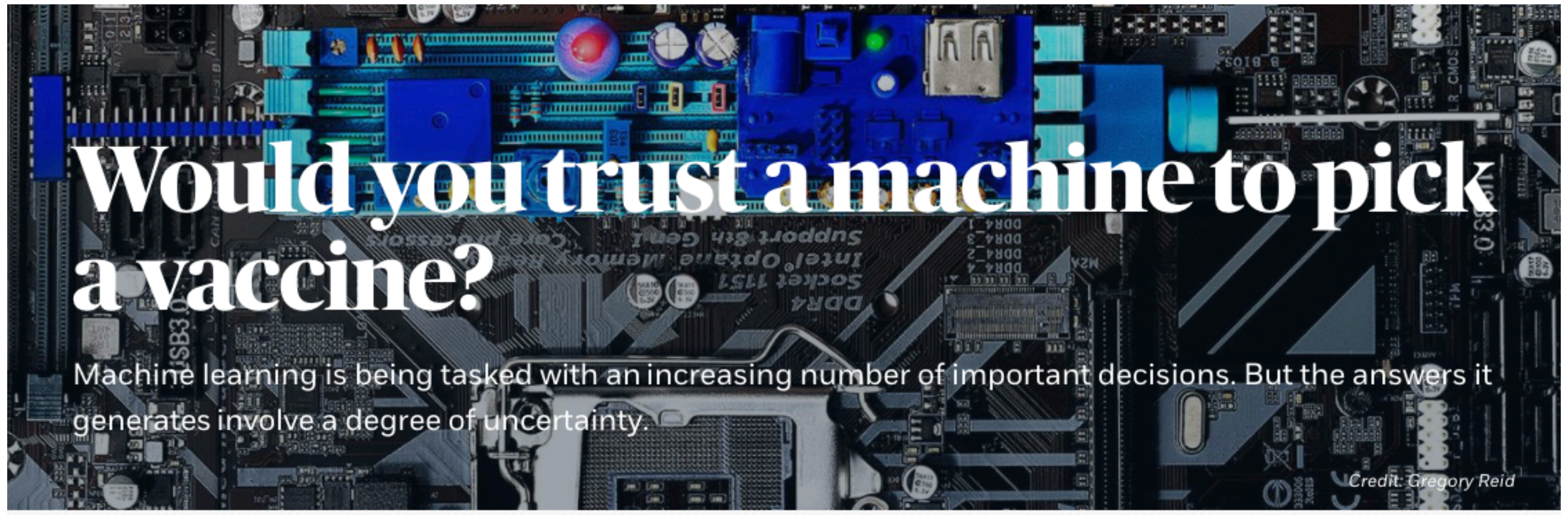
AI Can Make Bank Loans More Fair

by Sian Townson

November 06, 2020

Summary Save Share Print $8.95 Buy Copies

As banks increasingly deploy artificial intelligence tools to make credit decisions, they are having to revisit an unwelcome fact about the practice of lending: Historically, it has been riddled with biases against protected characteristics, such as race, gender, and sexual orientation. Such biases are evident in institutions' choices in terms of who gets credit and on what terms. In this context, relying on algorithms to make credit decisions instead of deferring to human judgment seems like an obvious fix. What machines lack in warmth, they surely make up for in objectivity, right?



EMILY LAMBERT | NOV 09, 2020   SECTIONS ECONOMICS   COLLECTIONS ARTIFICIAL INTELLIGENCE STATISTICS

Analysis

The United States of Risk Assessment: The Machines Influencing Criminal Justice Decisions

In every state, assessment tools help courts decide certain cases or correctional officers determine the supervision and programming an offender receives. But the tools each state uses varies widely, and how they're put into practice varies even more.

By Rhys Dipshan, Victoria Hudgins and Frank Ready | July 13, 2020 at 07:00 AM

medical applications

finance decisions

justice decisions

education admissions

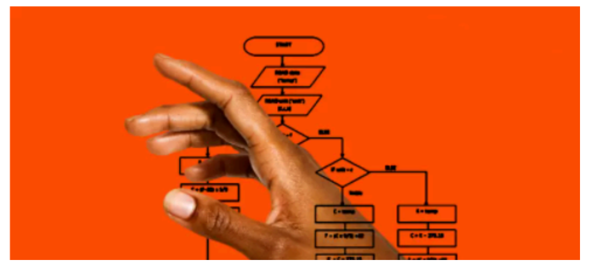
user data

Machine Learning

The lessons we all must learn from the A-levels algorithm debacle

Unless action is taken, similar systems will suffer from the same mistakes. And the consequences could be dire

By MATT BURGESS



BACKGROUND

ECONOMICS & SOCIETY

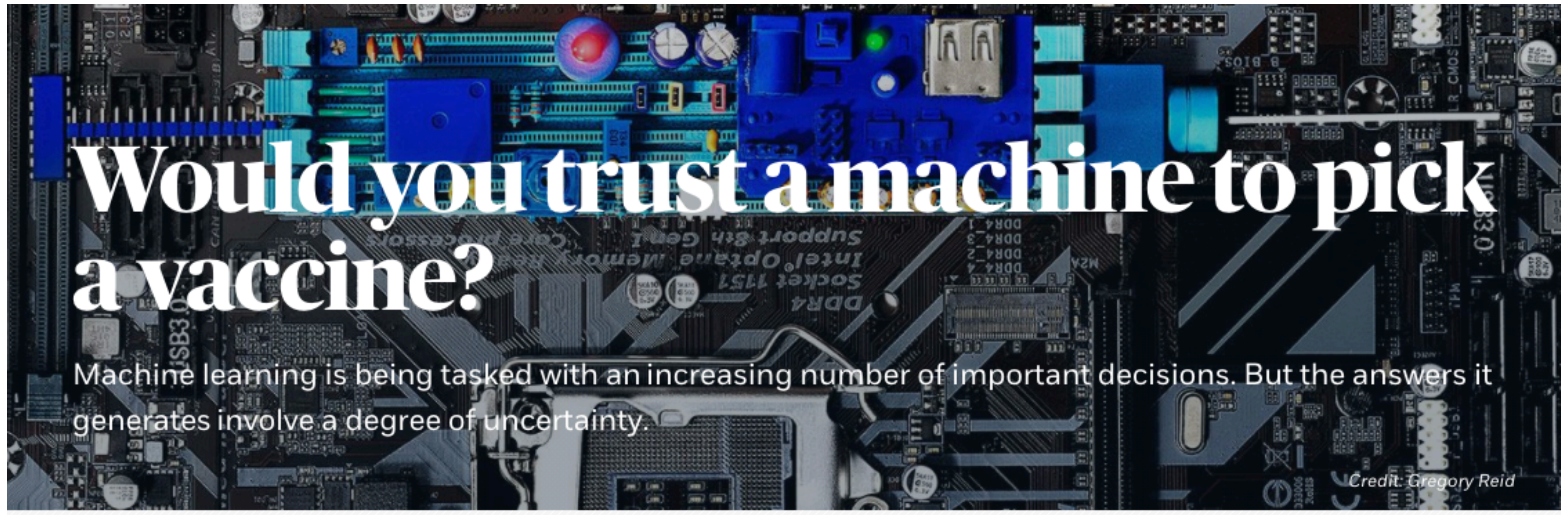
AI Can Make Bank Loans More Fair

by Sian Townson

November 06, 2020

Summary Save Share Print $8.95 Buy Copies

As banks increasingly deploy artificial intelligence tools to make credit decisions, they are having to revisit an unwelcome fact about the practice of lending: Historically, it has been riddled with biases against protected characteristics, such as race, gender, and sexual orientation. Such biases are evident in institutions' choices in terms of who gets credit and on what terms. In this context, relying on algorithms to make credit decisions instead of deferring to human judgment seems like an obvious fix. What machines lack in warmth, they surely make up for in objectivity, right?



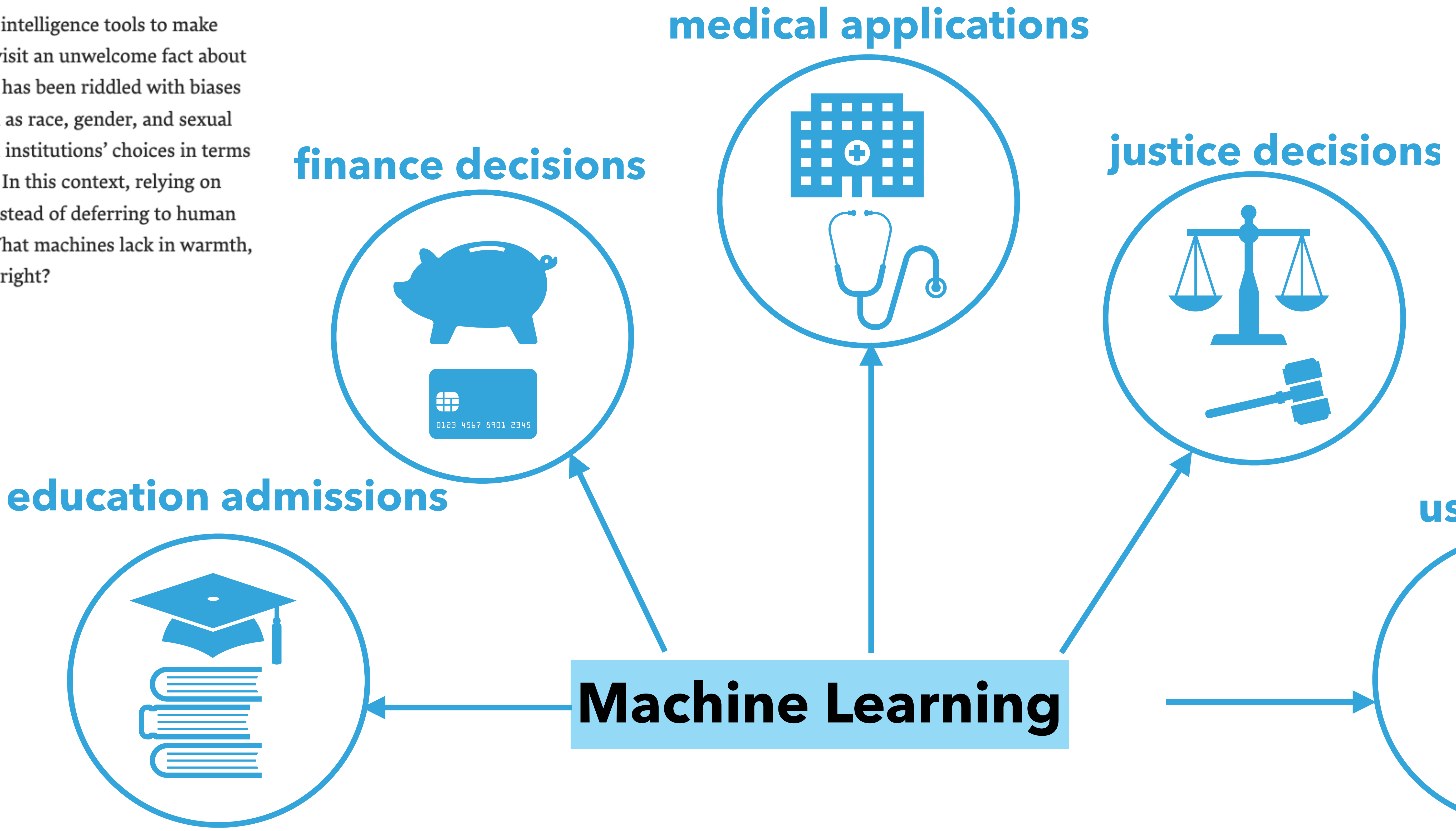
EMILY LAMBERT | NOV 09, 2020 SECTIONS ECONOMICS COLLECTIONS ARTIFICIAL INTELLIGENCE STATISTICS

Analysis

The United States of Risk Assessment: The Machines Influencing Criminal Justice Decisions

In every state, assessment tools help courts decide certain cases or correctional officers determine the supervision and programming an offender receives. But the tools each state uses varies widely, and how they're put into practice varies even more.

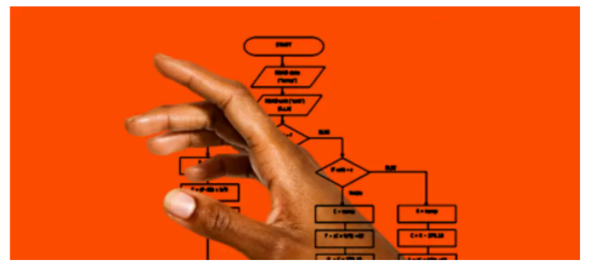
By Rhys Dipshan, Victoria Hudgins and Frank Ready | July 13, 2020 at 07:00 AM



The lessons we all must learn from the A-levels algorithm debacle

Unless action is taken, similar systems will suffer from the same mistakes. And the consequences could be dire

By MATT BURGESS



Research into Siri, Alexa, Google Assistant voice tech reveals bias in training data

By Mike Peterson
7 months ago

BACKGROUND

- Assessment machine learning models independently from the training procedures
  - legal requirement, 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*]

BACKGROUND

- Assessment machine learning models independently from the training procedures
  - legal requirement, 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*]

Do ImageNet Classifiers Generalize to ImageNet?

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, Vaishaal Shankar

We build new test sets for the CIFAR-10 and ImageNet datasets. Both benchmarks have been the focus of intense research for almost a decade, raising the danger of overfitting to excessively re-used test sets. By closely following the original dataset creation processes, we test to what extent current classification models generalize to new data. We evaluate a broad range of models and find accuracy drops of 3% – 15% on CIFAR-10 and 11% – 14% on ImageNet. However, accuracy gains on the original test sets translate to larger gains on the new test sets. Our results suggest that the accuracy drops are not caused by adaptivity, but by the models' inability to generalize to slightly "harder" images than those found in the original test sets.

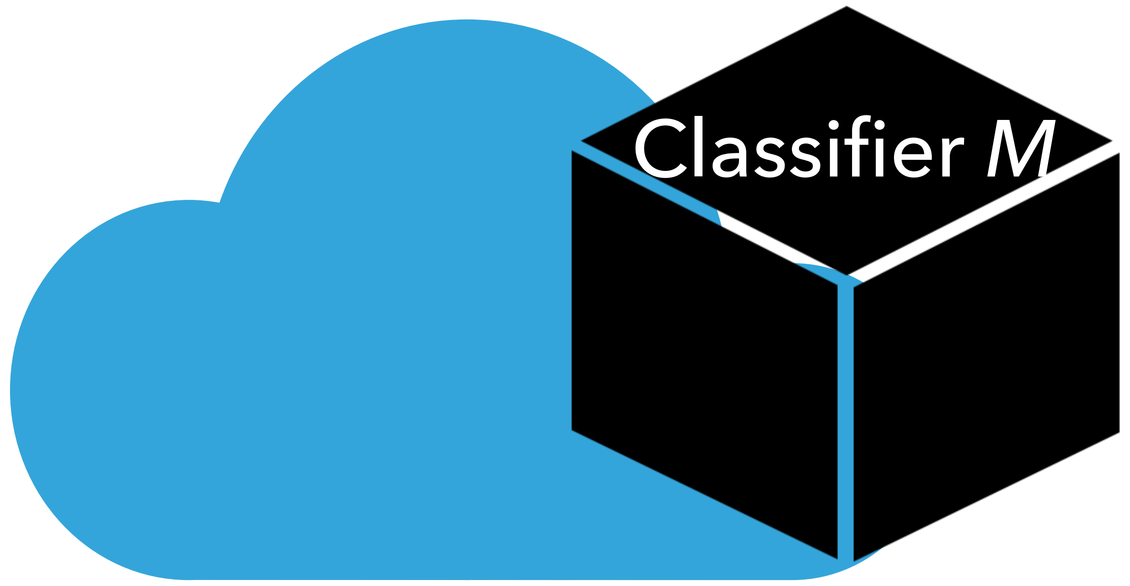
OBJECTIVES OF BLACKBOX CLASSIFIER ASSESSMENT

- How **accurate**?
- How **calibrated**?
- How **fair**?
- …

- And how much **confidence** should we have in this assessment?
- How to **increase our confidence** given the labeling budget?

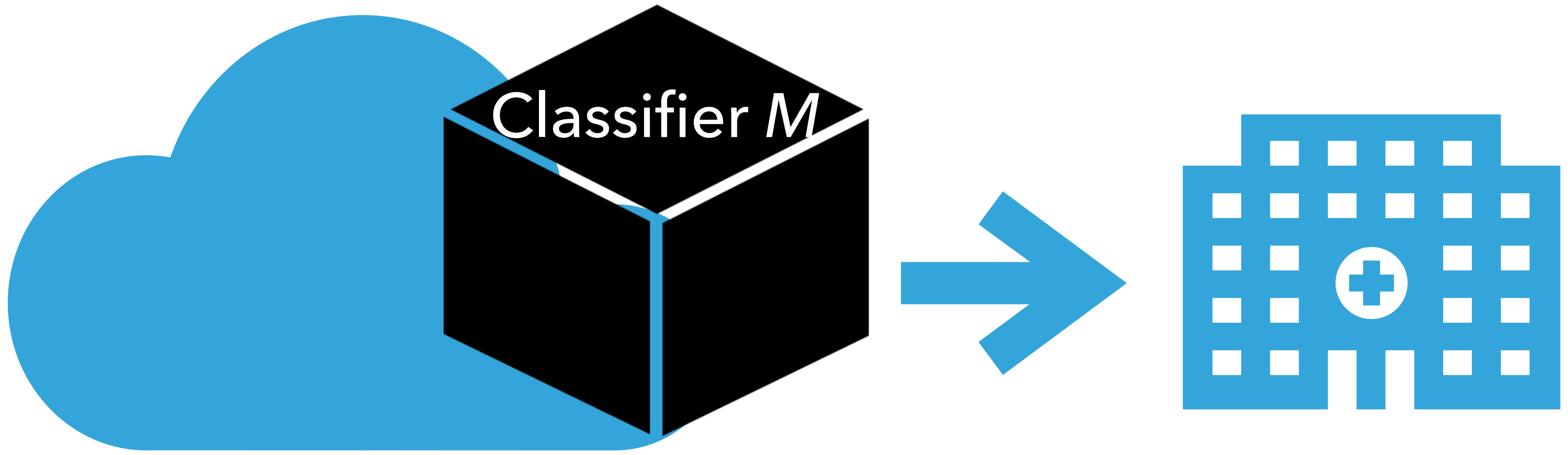
OBJECTIVES OF BLACKBOX CLASSIFIER ASSESSMENT

- How **accurate**?
- How **calibrated**?
- How **fair**?
- …
- And how much **confidence** should we have in this assessment?
- How to **increase our confidence** given the labeling budget?



OBJECTIVES OF BLACKBOX CLASSIFIER ASSESSMENT

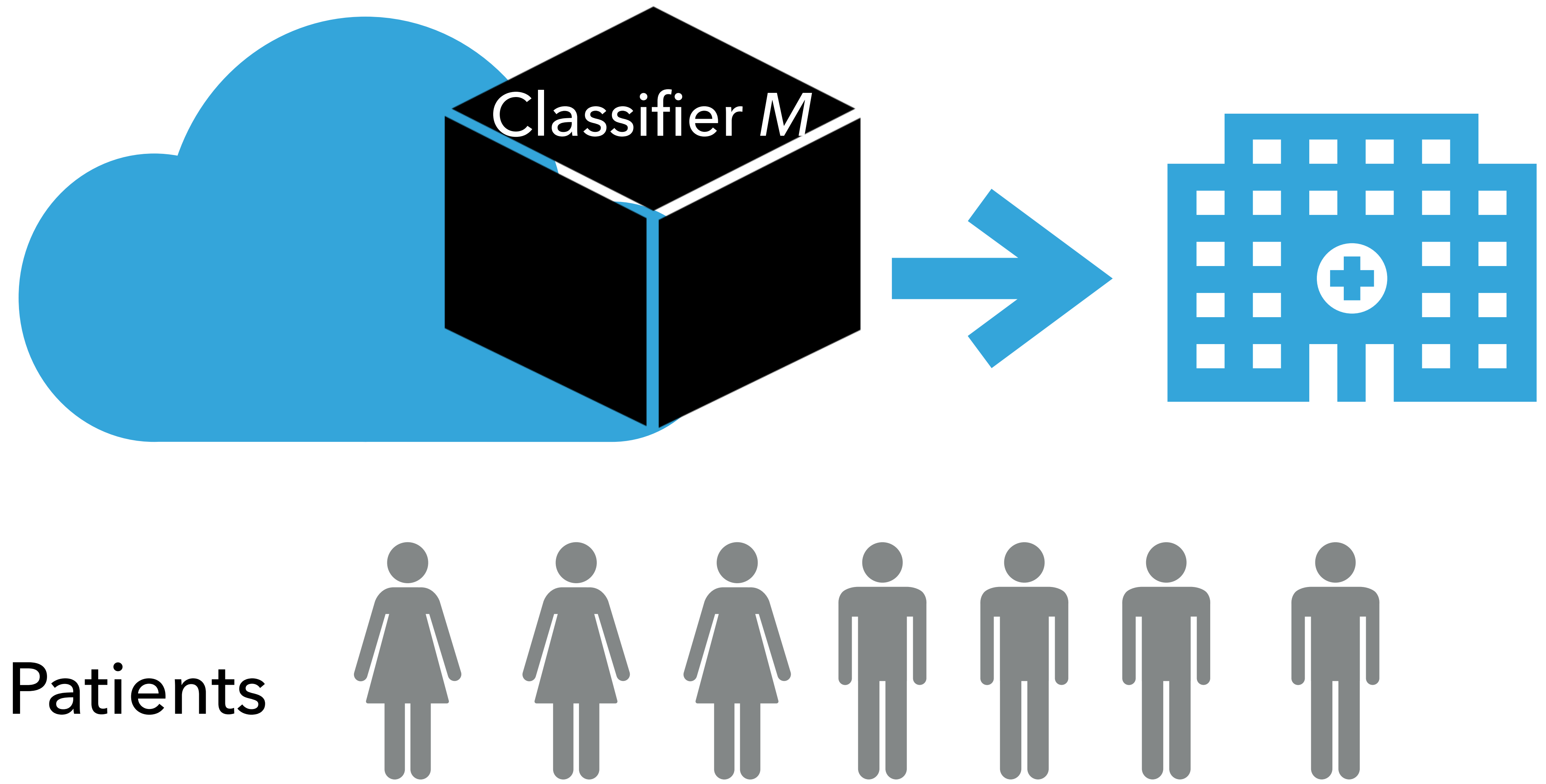
- How **accurate**?
- How **calibrated**?
- How **fair**?
- …
- And how much **confidence** should we have in this assessment?
- How to **increase our confidence** given the labeling budget?



OBJECTIVES OF BLACKBOX CLASSIFIER ASSESSMENT

- How **accurate**?
- How **calibrated**?
- How **fair**?
- …

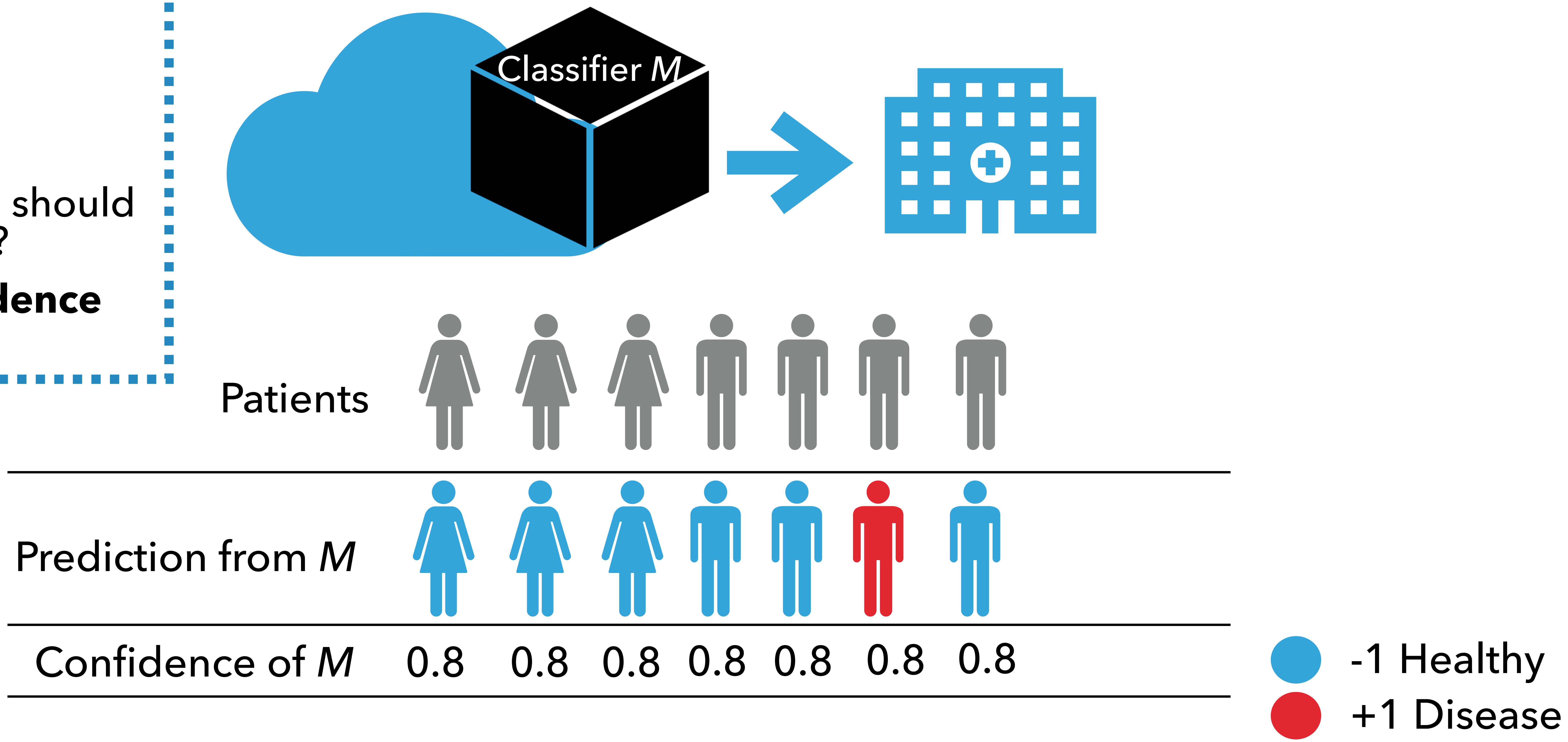
- And how much **confidence** should we have in this assessment?
- How to **increase our confidence** given the labeling budget?



OBJECTIVES OF BLACKBOX CLASSIFIER ASSESSMENT

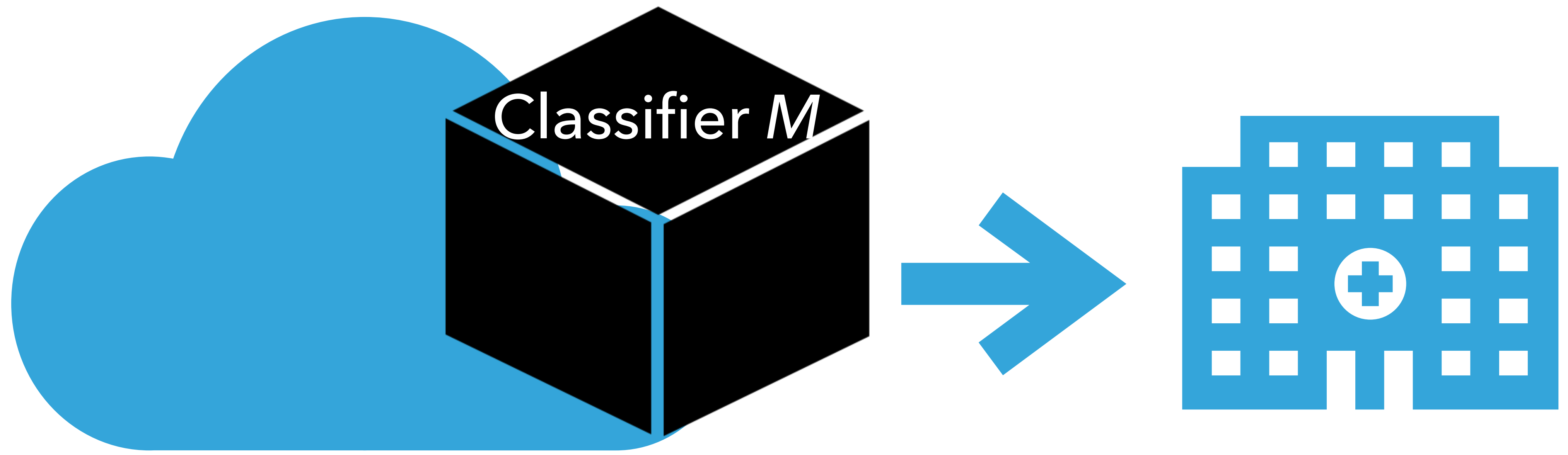
- How **accurate**?
- How **calibrated**?
- How **fair**?
- …

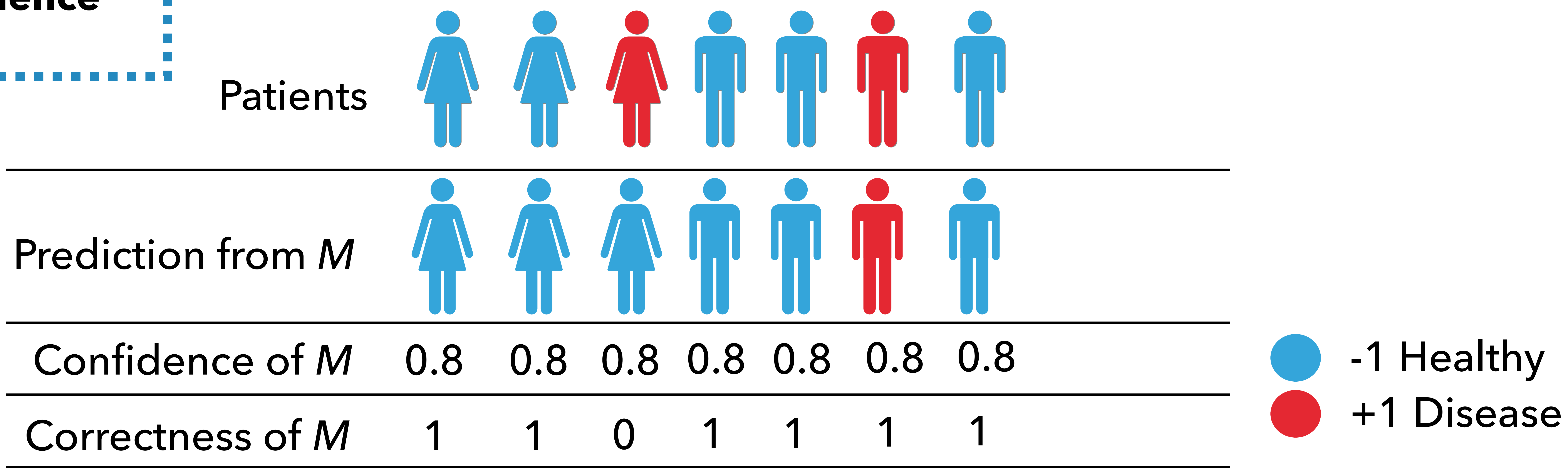
- And how much **confidence** should we have in this assessment?
- How to **increase our confidence** given the labeling budget?



OBJECTIVES OF BLACKBOX CLASSIFIER ASSESSMENT

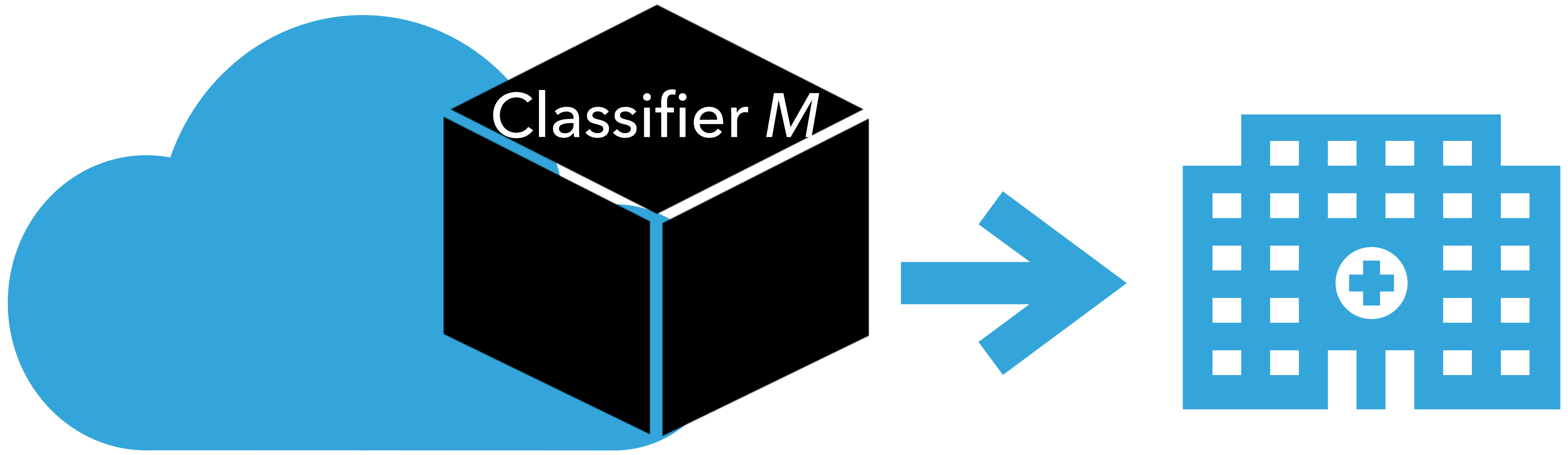
- How **accurate**?
- How **calibrated**?
- How **fair**?
- …
- And how much **confidence** should we have in this assessment?
- How to **increase our confidence** given the labeling budget?





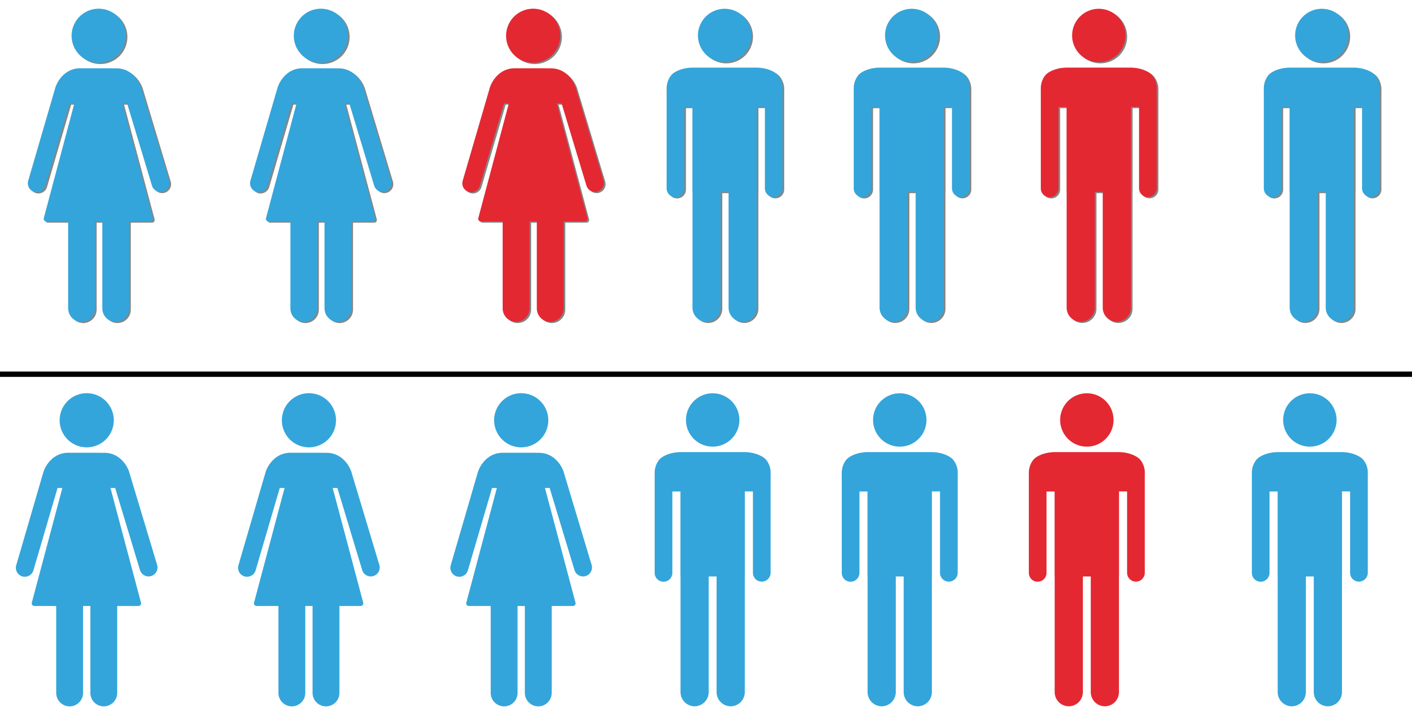
OBJECTIVES OF BLACKBOX CLASSIFIER ASSESSMENT

- How **accurate**?
- How **calibrated**?
- How **fair**?
- …
- And how much **confidence** should we have in this assessment?
- How to **increase our confidence** given the labeling budget?



Patients

Prediction from *M*



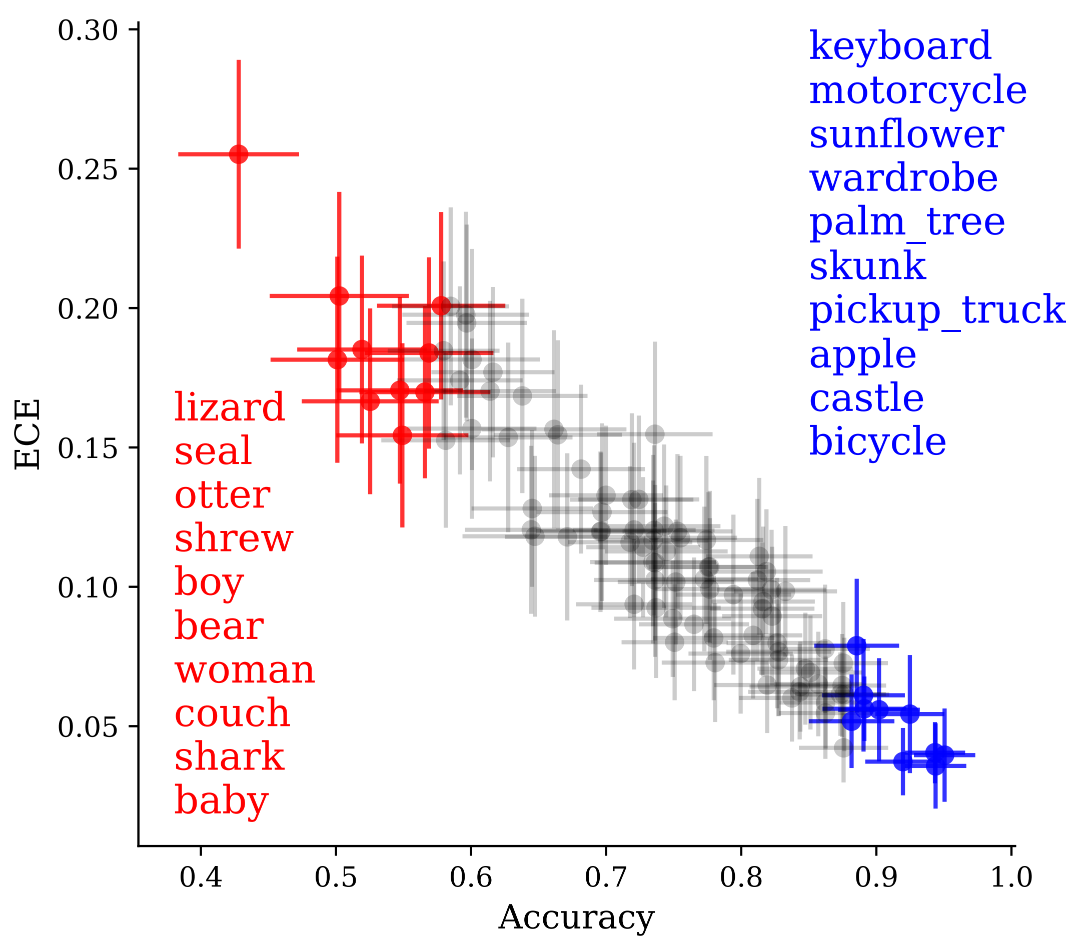
								Average
Confidence of *M*	0.8	0.8	0.8	0.8	0.8	0.8	0.8	0.8
Correctness of *M*	1	1	0	1	1	1	1	6/7

-1 Healthy
+1 Disease

ROAD MAP

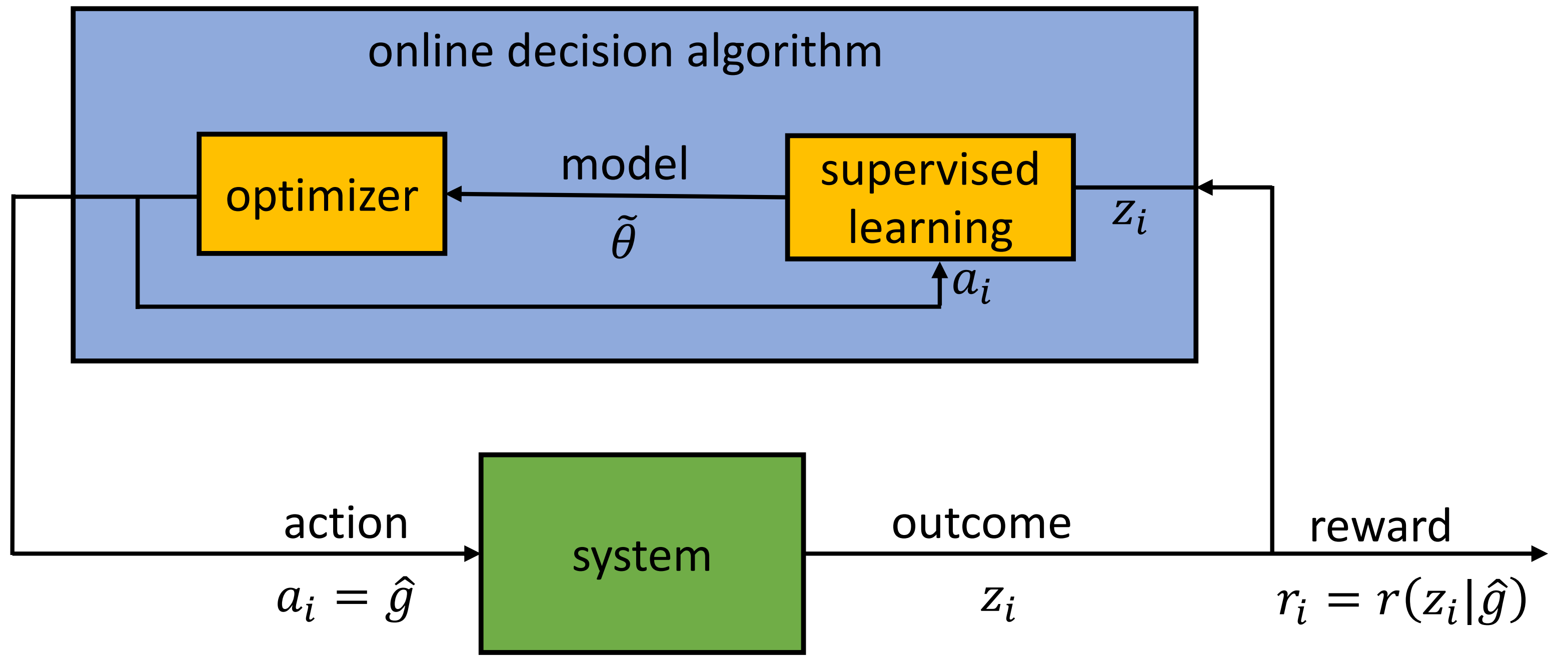
Bayesian assessment

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



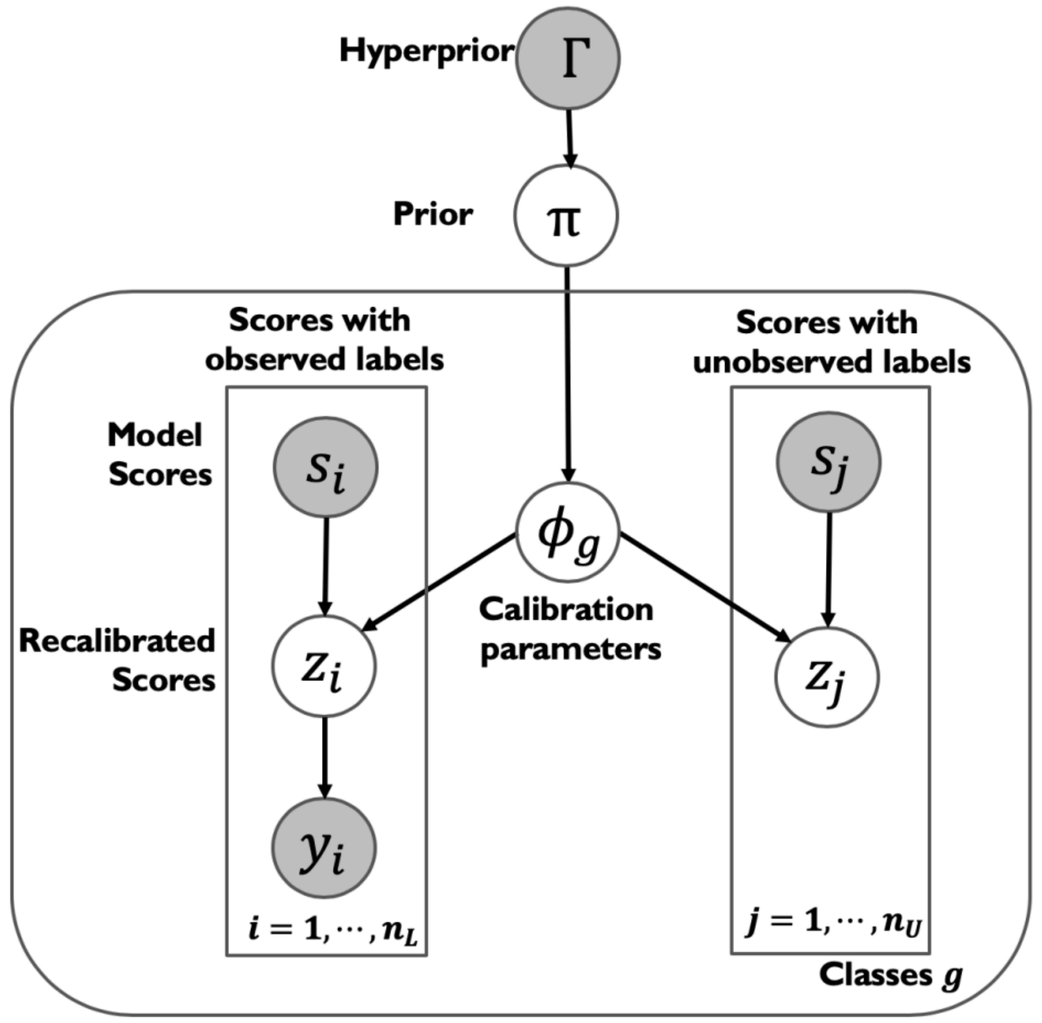
active Bayesian assessment

2. **Reduce uncertainty** of assessment, with **actively labeled data** selected from a pool of unlabeled data



assess with **unlabeled** data

3. **Reduce uncertainty** of assessment, by leveraging both **labeled and unlabeled data**

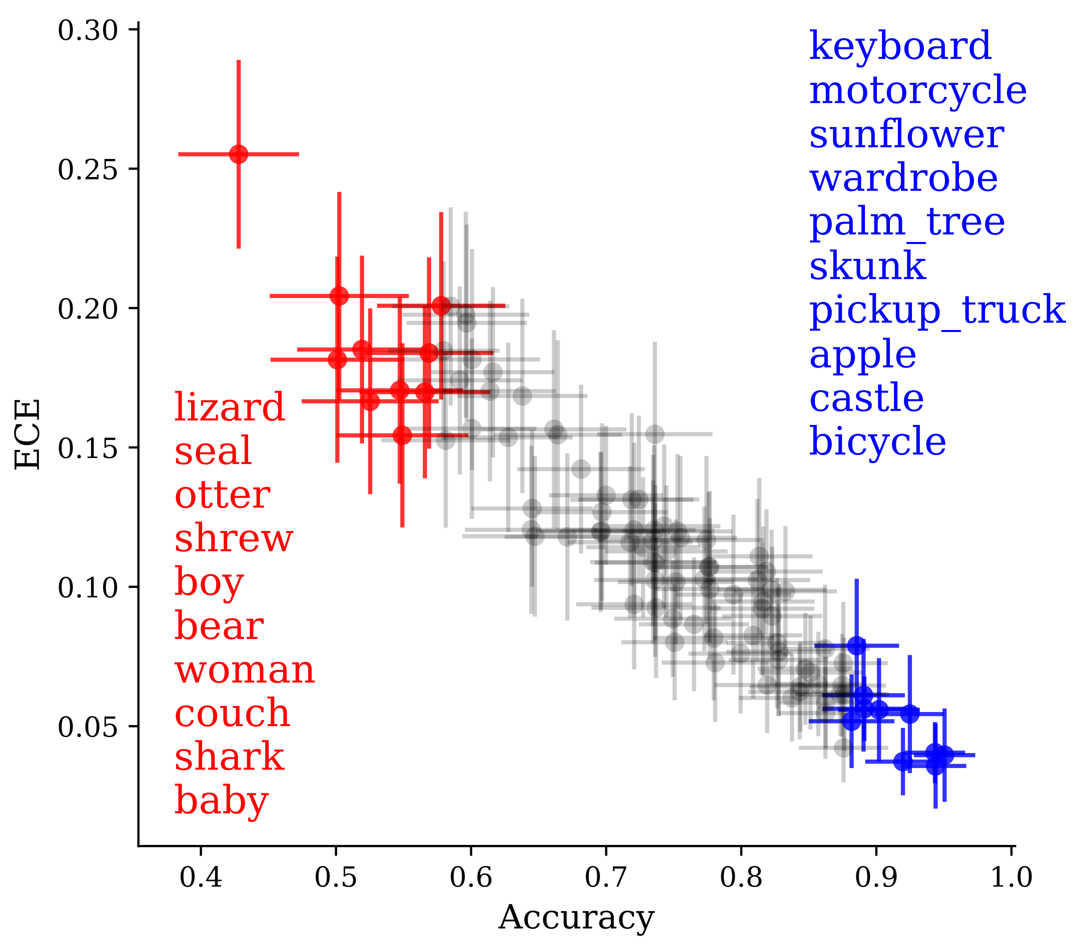


UCIRVINE

ROAD MAP

Bayesian assessment

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



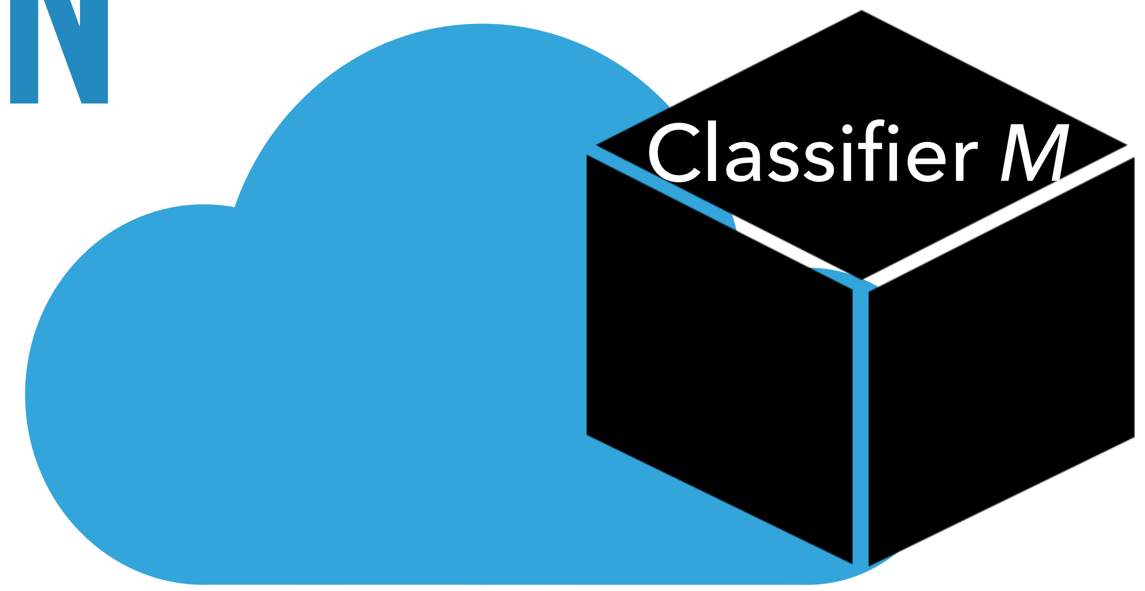
active Bayesian assessment

2. Reduce uncertainty of assessment, **with actively labeled data** selected from a pool of unlabeled data

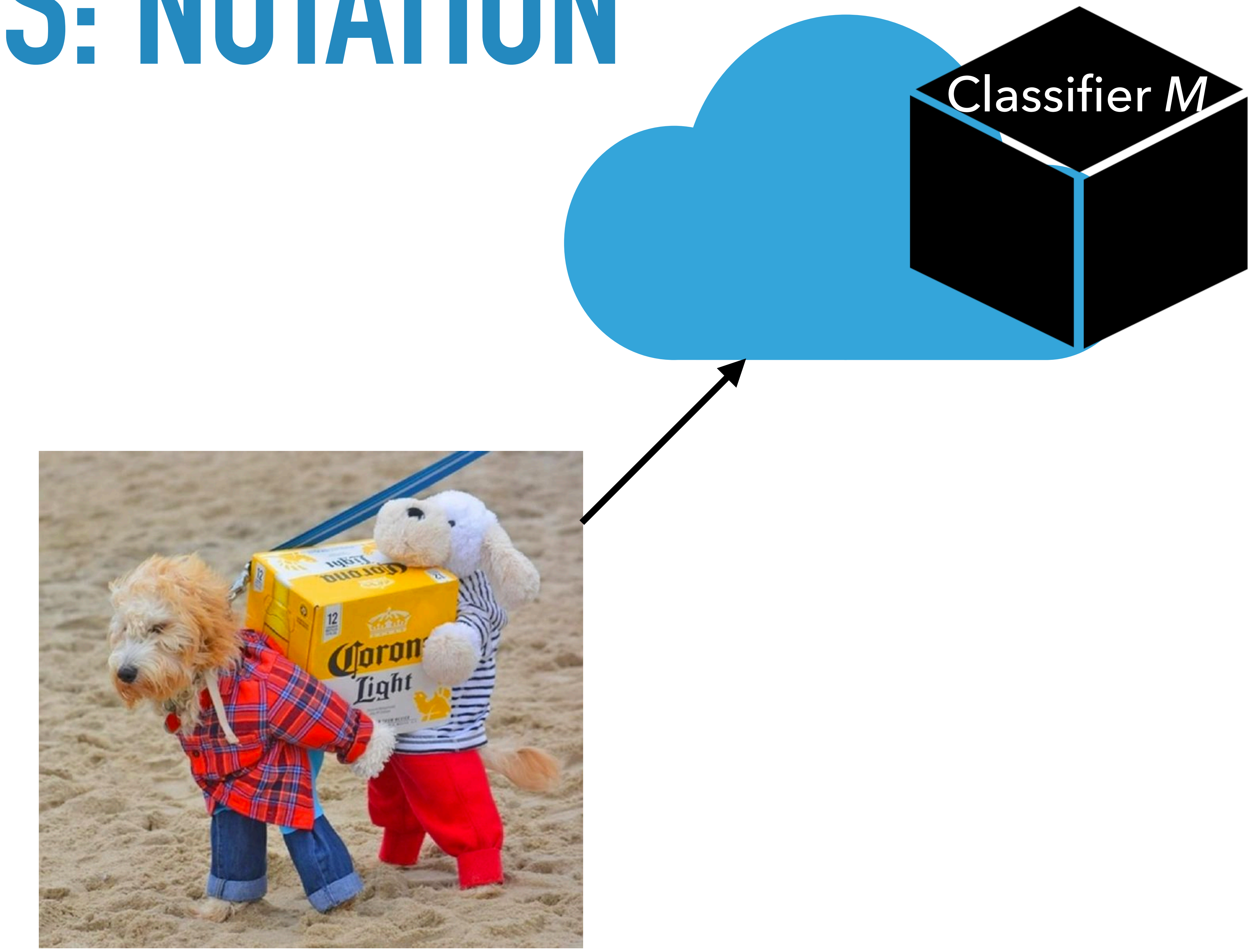
assess with **unlabeled** data

3. Reduce uncertainty of assessment, by leveraging both **labeled and unlabeled data**

PRELIMINARIES: NOTATION



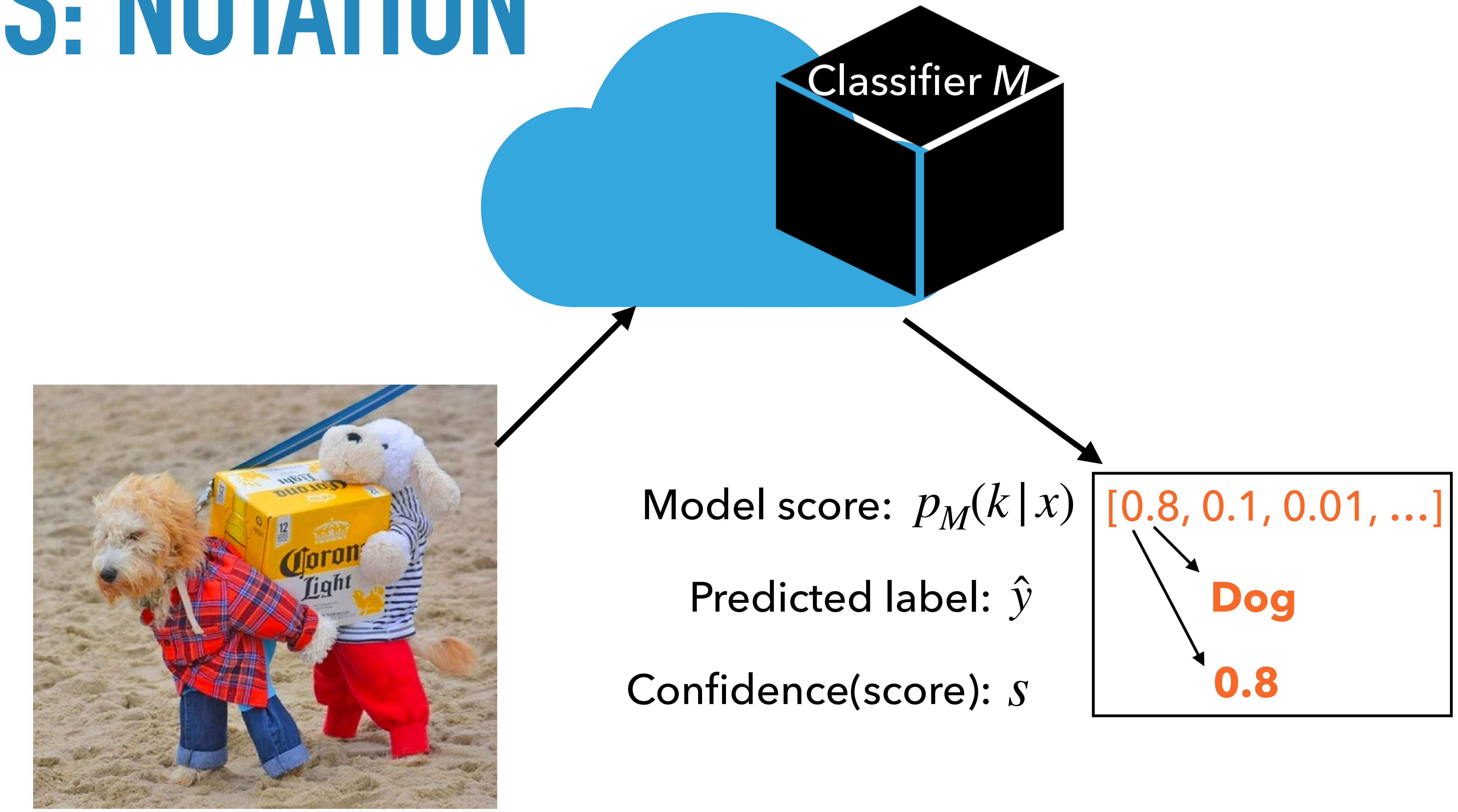
PRELIMINARIES: NOTATION



Input: x

PRELIMINARIES: NOTATION

Input: x

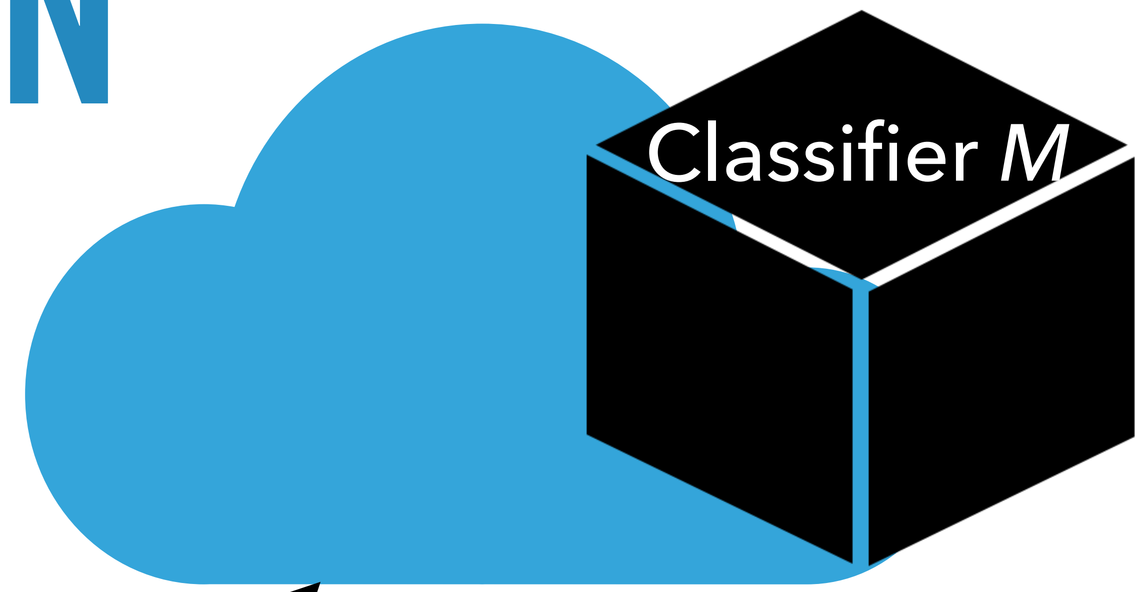


$\hat{y} = arg\,max_k p_M(k|x)$

$s = max_k p_M(k|x)$

UCIrvine

PRELIMINARIES: NOTATION



Input: x

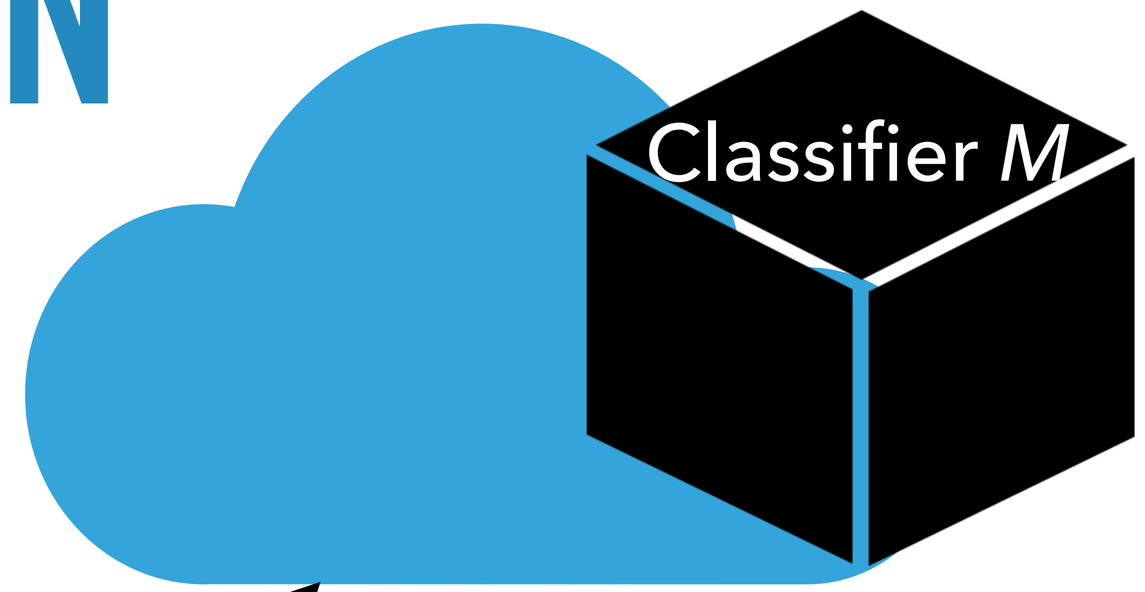
label: y **Dog**

Model score: $p_M(k|x)$ [0.8, 0.1, 0.01, ...]

Predicted label: $\hat{y}$ **Dog** $\hat{y} = arg\,max_k p_M(k|x)$

Confidence(score): s **0.8** $s = max_k p_M(k|x)$

PRELIMINARIES: NOTATION



$p(x, y)$

Input: x

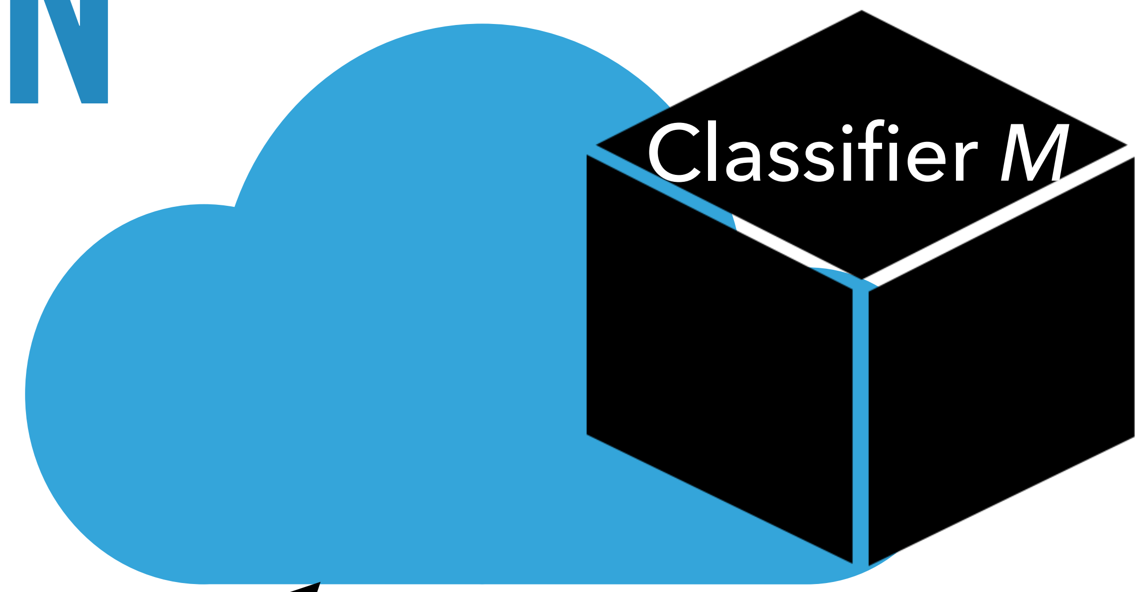
label: y **Dog**

Model score: $p_M(k \mid x)$ [0.8, 0.1, 0.01, ...]

Predicted label: $\hat{y}$ **Dog** $\hat{y} = arg\,max_k p_M(k \mid x)$

Confidence(score): s **0.8** $s = max_k p_M(k \mid x)$

PRELIMINARIES: NOTATION



$p(x, y)$

Input: x

label: y **Dog**

Model score: $p_M(k|x)$ [0.8, 0.1, 0.01, ...]

Predicted label: $\hat{y}$ **Dog** $\hat{y} = arg\,max_k p_M(k|x)$

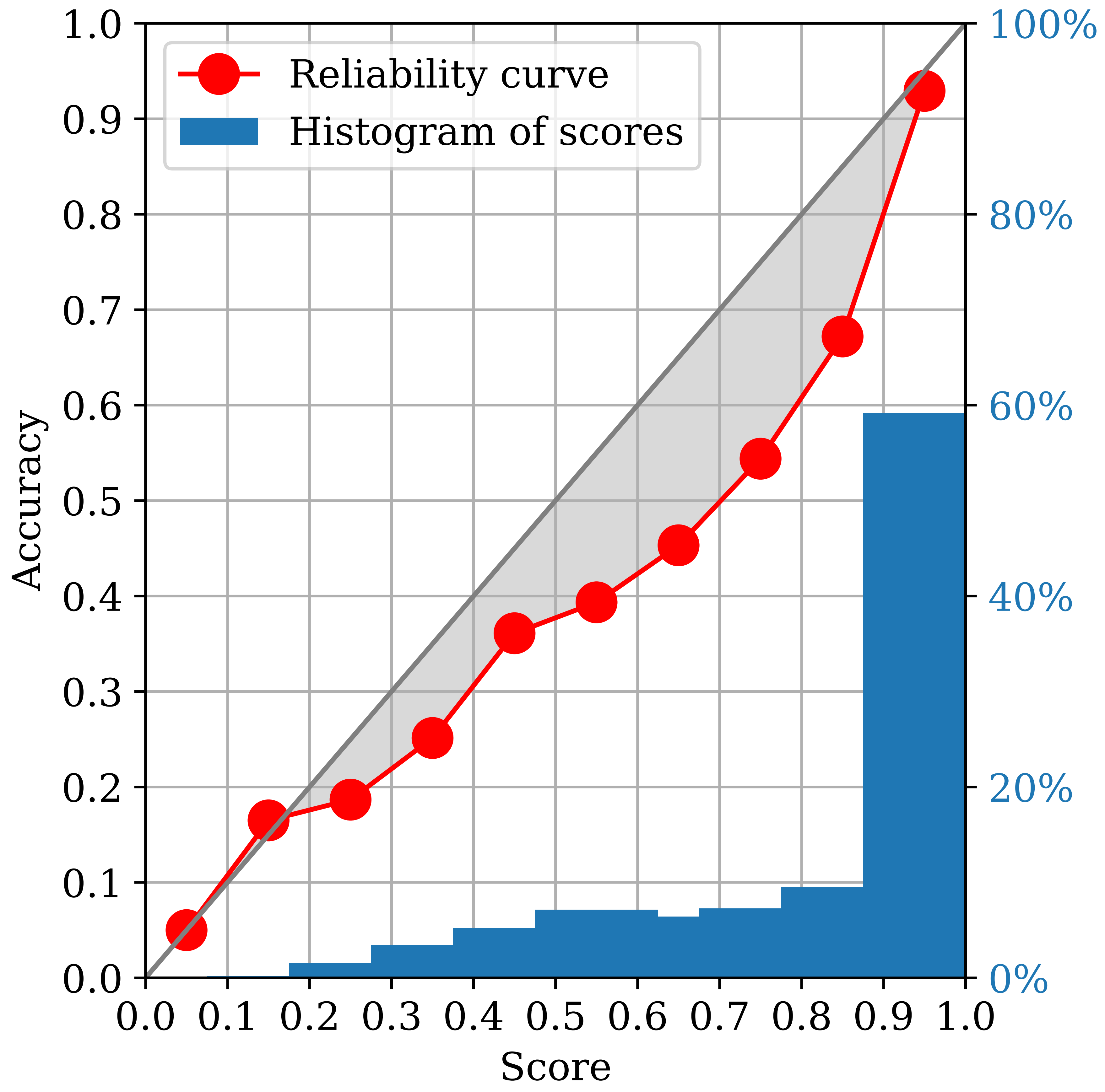
Confidence(score): s **0.8** $s = max_k p_M(k|x)$

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

Empirical accuracy $\hat{\theta} = \frac{1}{N} \sum_{i=1}^{N} 1(y_i = \hat{y}_i)$

PRELIMINARIES: CALIBRATION

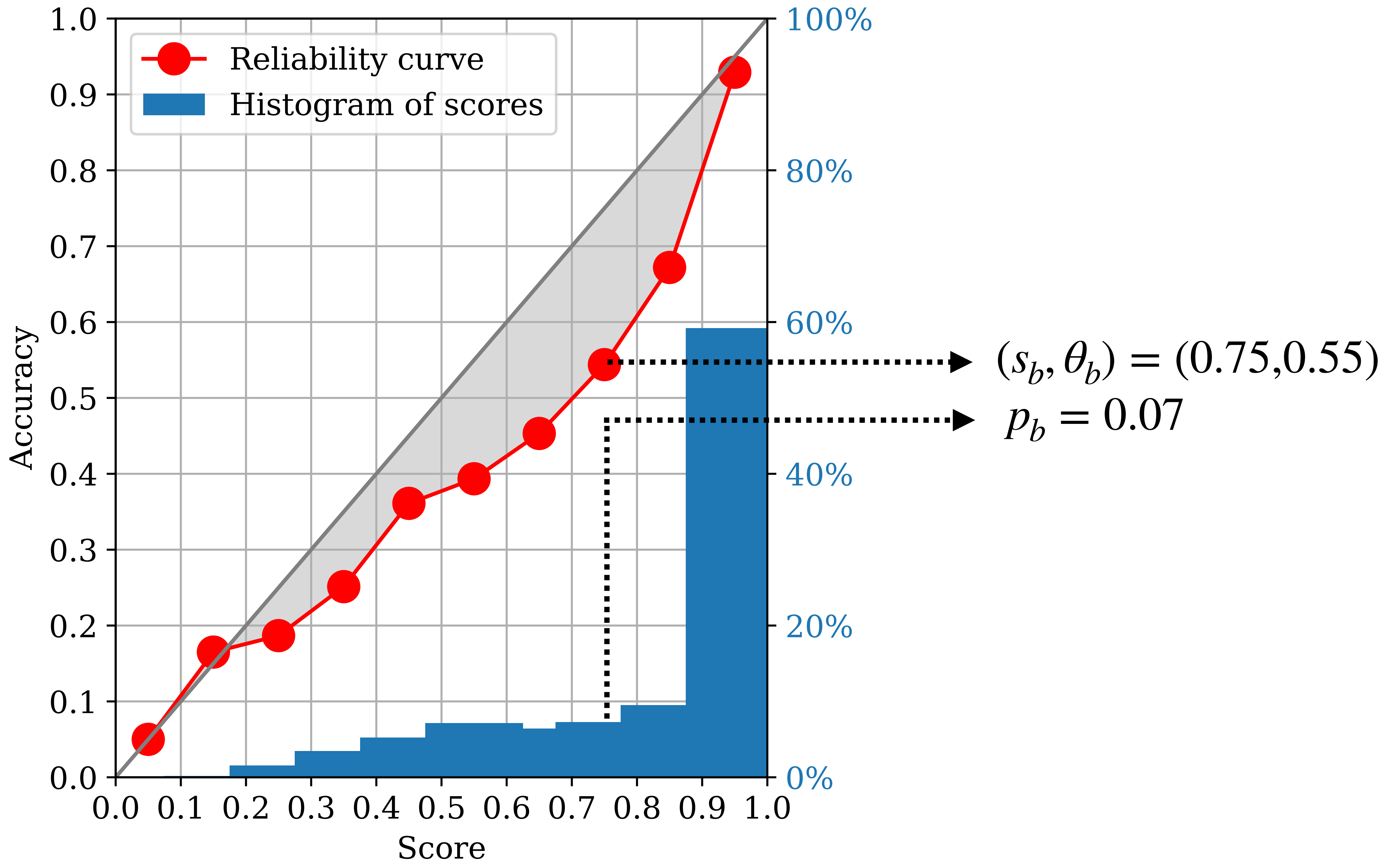
- Deep neural networks are miscalibrated [*Guo et al. 2017*]
  - e.g. ResNet-110 on CIFAR-100
- Reliability diagram
- Expected calibration error (ECE)



Reliability diagram for ResNet-110 on CIFAR-100

PRELIMINARIES: CALIBRATION

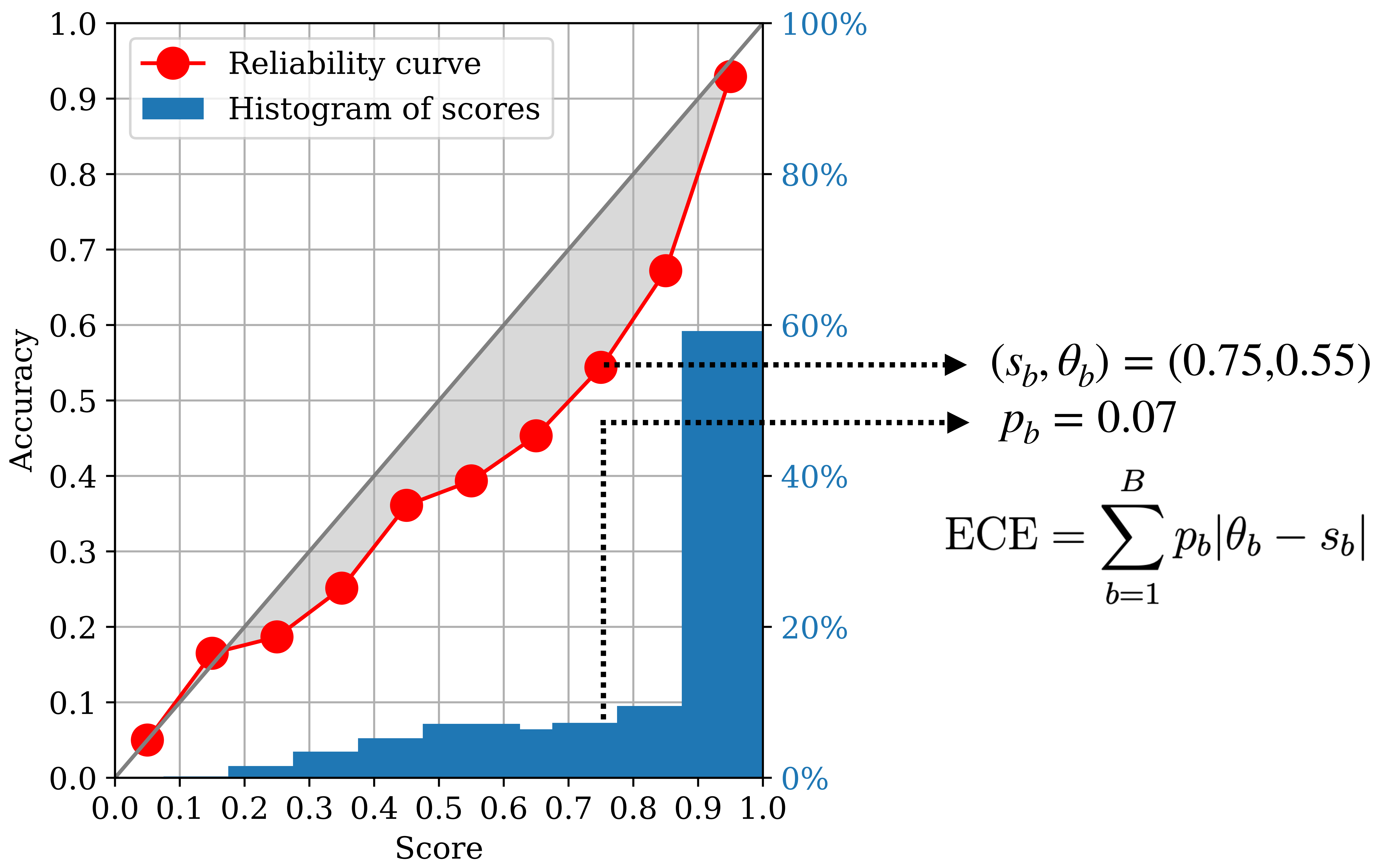
- Deep neural networks are miscalibrated [*Guo et al. 2017*]
  - e.g. ResNet-110 on CIFAR-100
- Reliability diagram
- Expected calibration error (ECE)



Reliability diagram for ResNet-110 on CIFAR-100

PRELIMINARIES: CALIBRATION

- Deep neural networks are miscalibrated [*Guo et al. 2017*]
  - e.g. ResNet-110 on CIFAR-100
- Reliability diagram
- Expected calibration error (ECE)

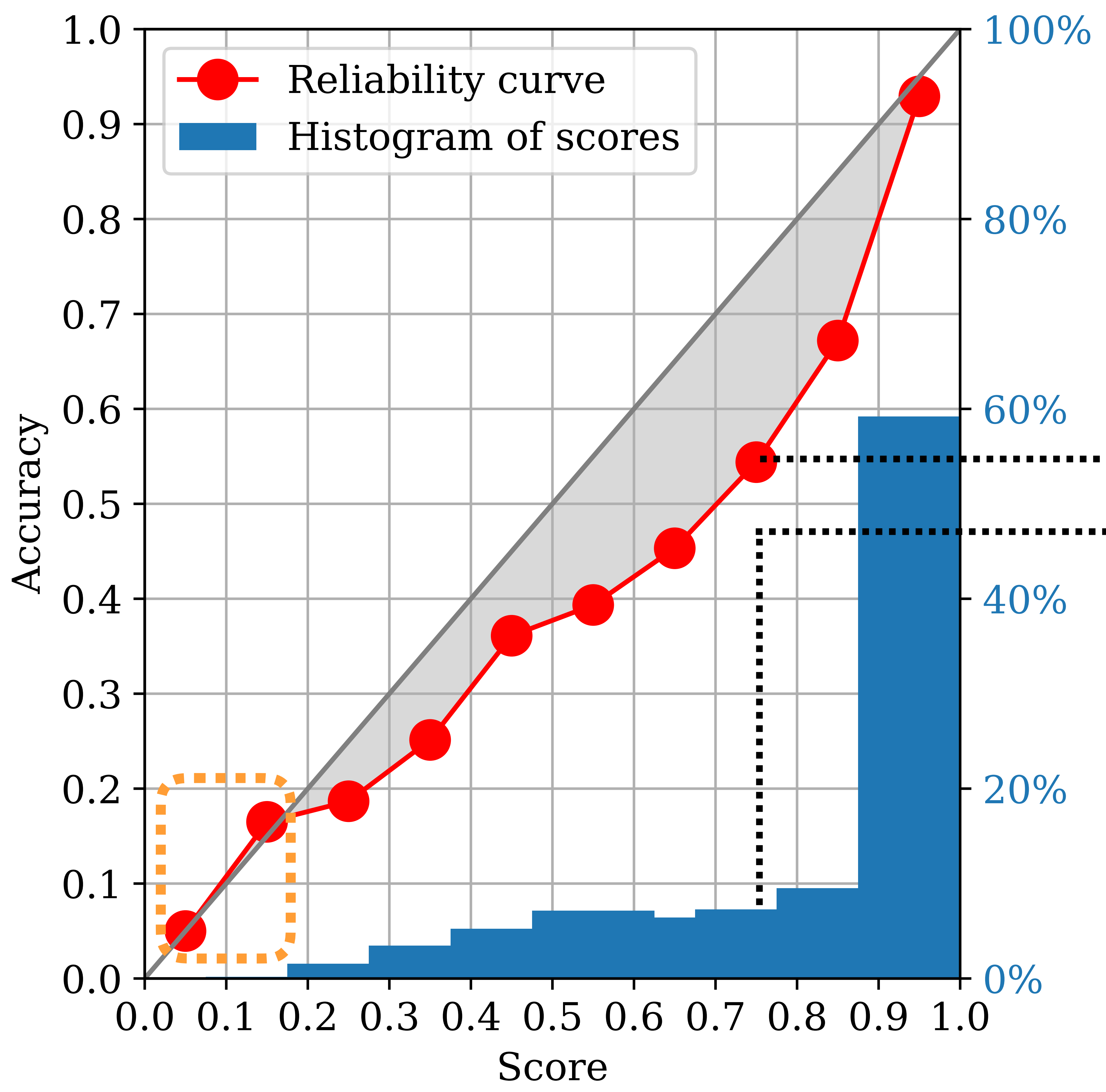


$$\mathrm{ECE} = \sum_{b=1}^{B} p_b |\theta_b - s_b|$$

Reliability diagram for ResNet-110 on CIFAR-100

PRELIMINARIES: CALIBRATION

- Deep neural networks are miscalibrated [*Guo et al. 2017*]
  - e.g. ResNet-110 on CIFAR-100
- Reliability diagram
- Expected calibration error (ECE)



$(s_b, \theta_b) = (0.75, 0.55)$

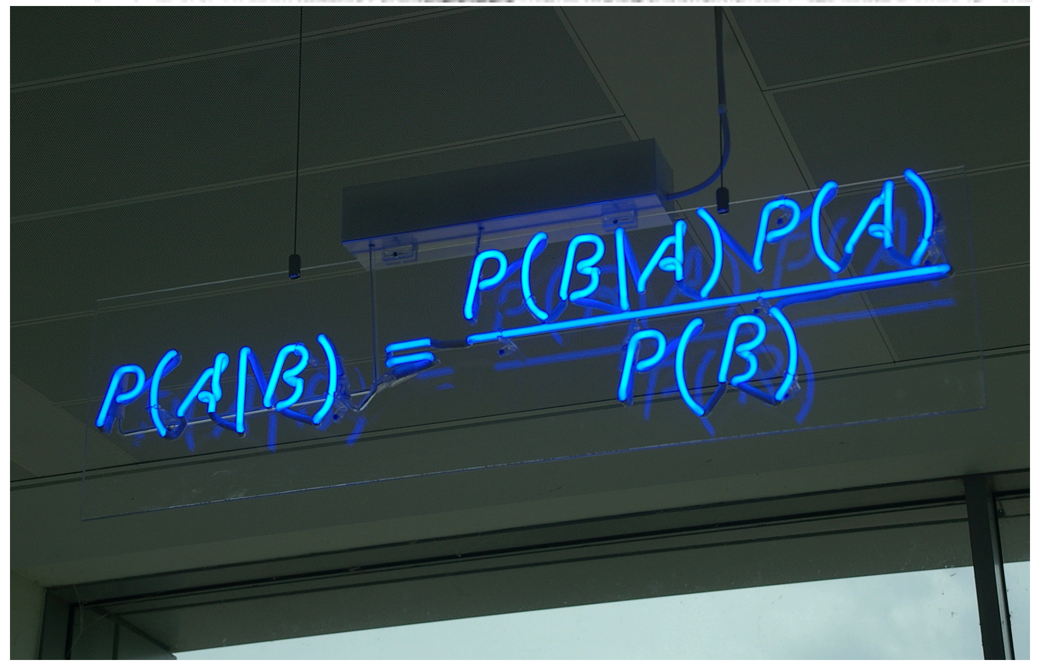
$p_b = 0.07$

$$\mathrm{ECE} = \sum_{b=1}^{B} p_b |\theta_b - s_b|$$

Reliability diagram for ResNet-110 on CIFAR-100 computed with 10,000 data points

BAYESIAN ASSESSMENT: HOW ACCURATE

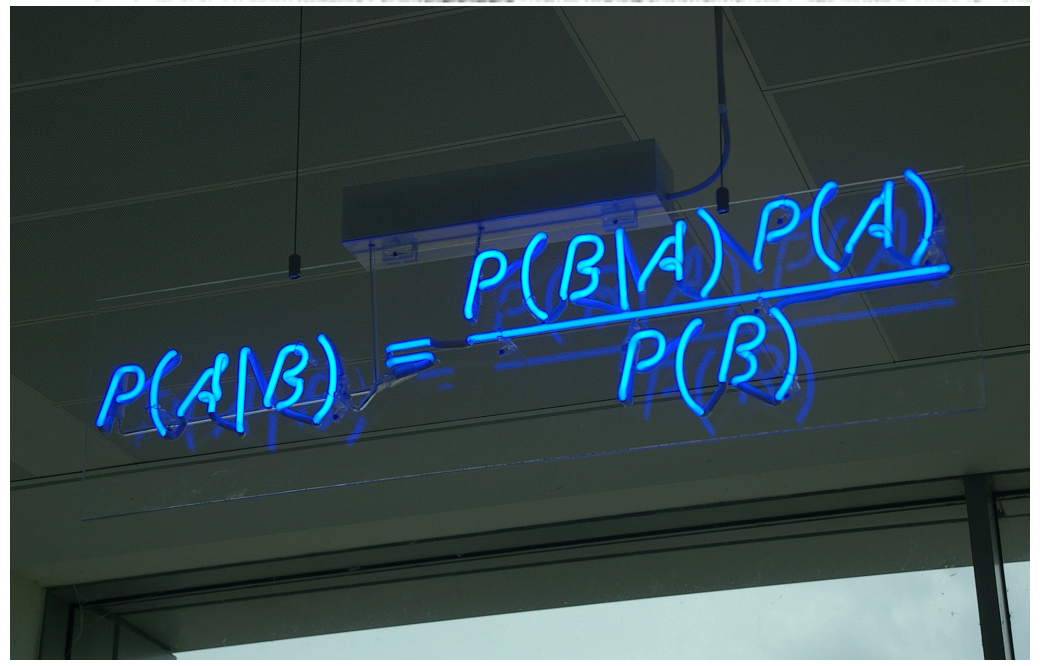
Accuracy	$\theta = \mathbb{E}_{p(x,y)} 1(y = \hat{y})$
Empirical accuracy	$\hat{\theta} = \frac{1}{N} \sum_{i=1}^{N} 1(y_i = \hat{y}_i)$



BAYESIAN ASSESSMENT: HOW ACCURATE

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

Empirical accuracy $\hat{\theta} = \frac{1}{N} \sum_{i=1}^{N} 1(y_i = \hat{y}_i)$



Beta posterior

Beta prior

Bernoulli likelihood

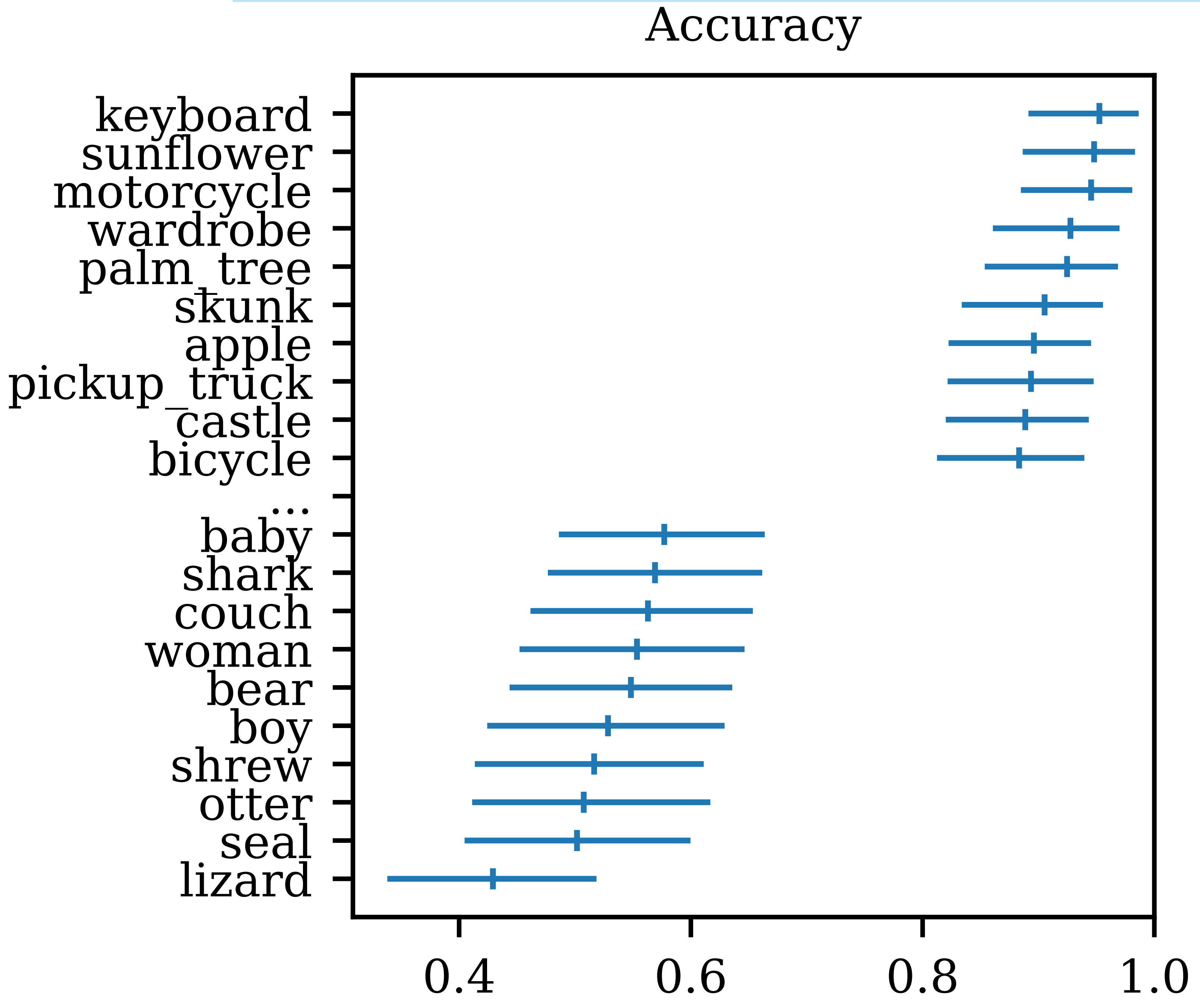
$$p(\theta|\mathcal{D}) = \frac{p(\theta) \cdot \prod_{i=1}^{N} q_\theta(z_i)}{\int_\theta p(\theta) \cdot \prod_{i=1}^{N} q_\theta(z_i)\, d\theta}$$

$$\mathcal{D} = \{(x_i, y_i), i = 1, 2, \cdots, N\}$$

BAYESIAN ASSESSMENT: HOW ACCURATE

Accuracy of the k-th predicted class:

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



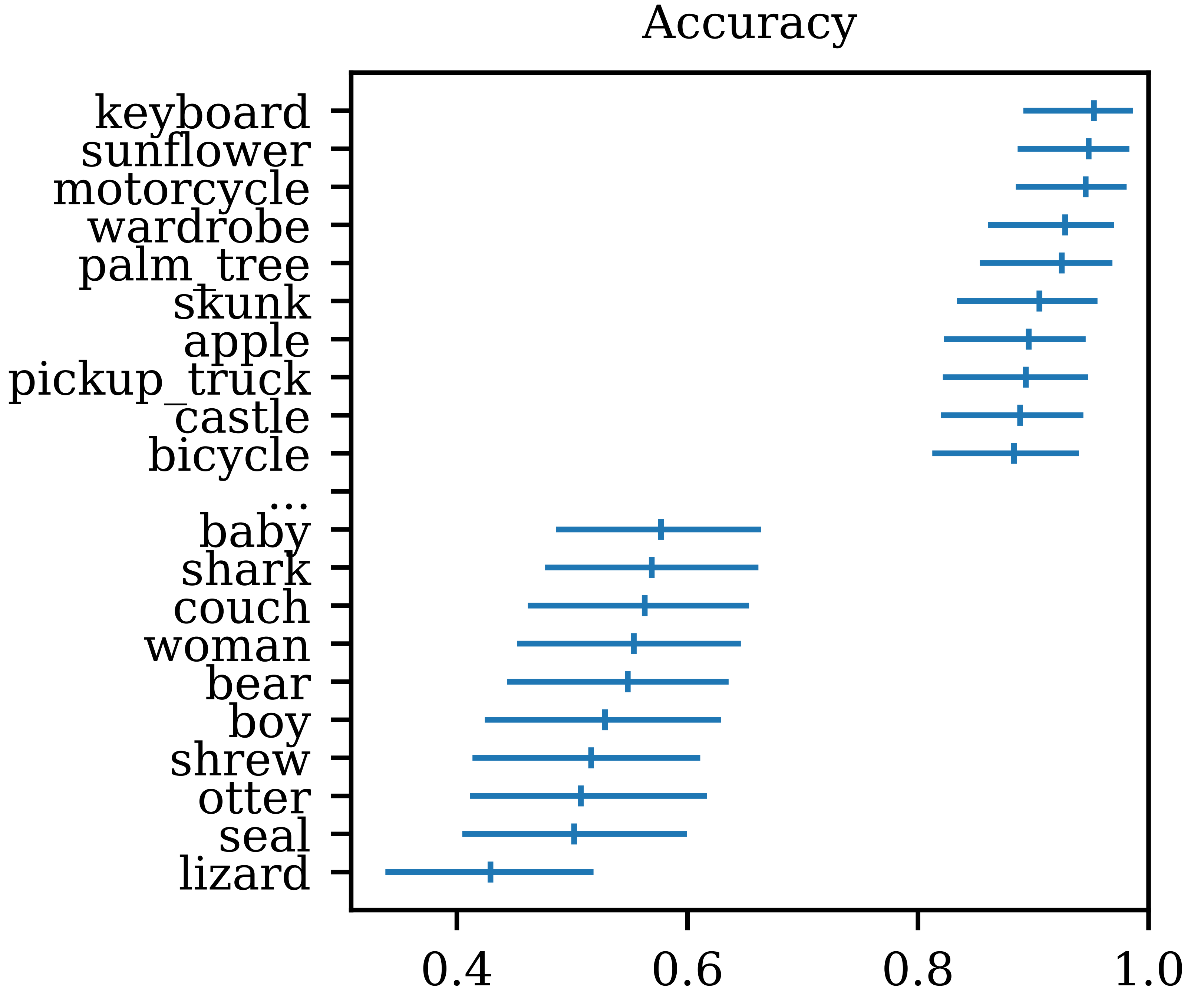
classwise accuracy

for ResNet-110 on CIFAR-100

BAYESIAN ASSESSMENT: HOW ACCURATE

Accuracy of the k-th predicted class:

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

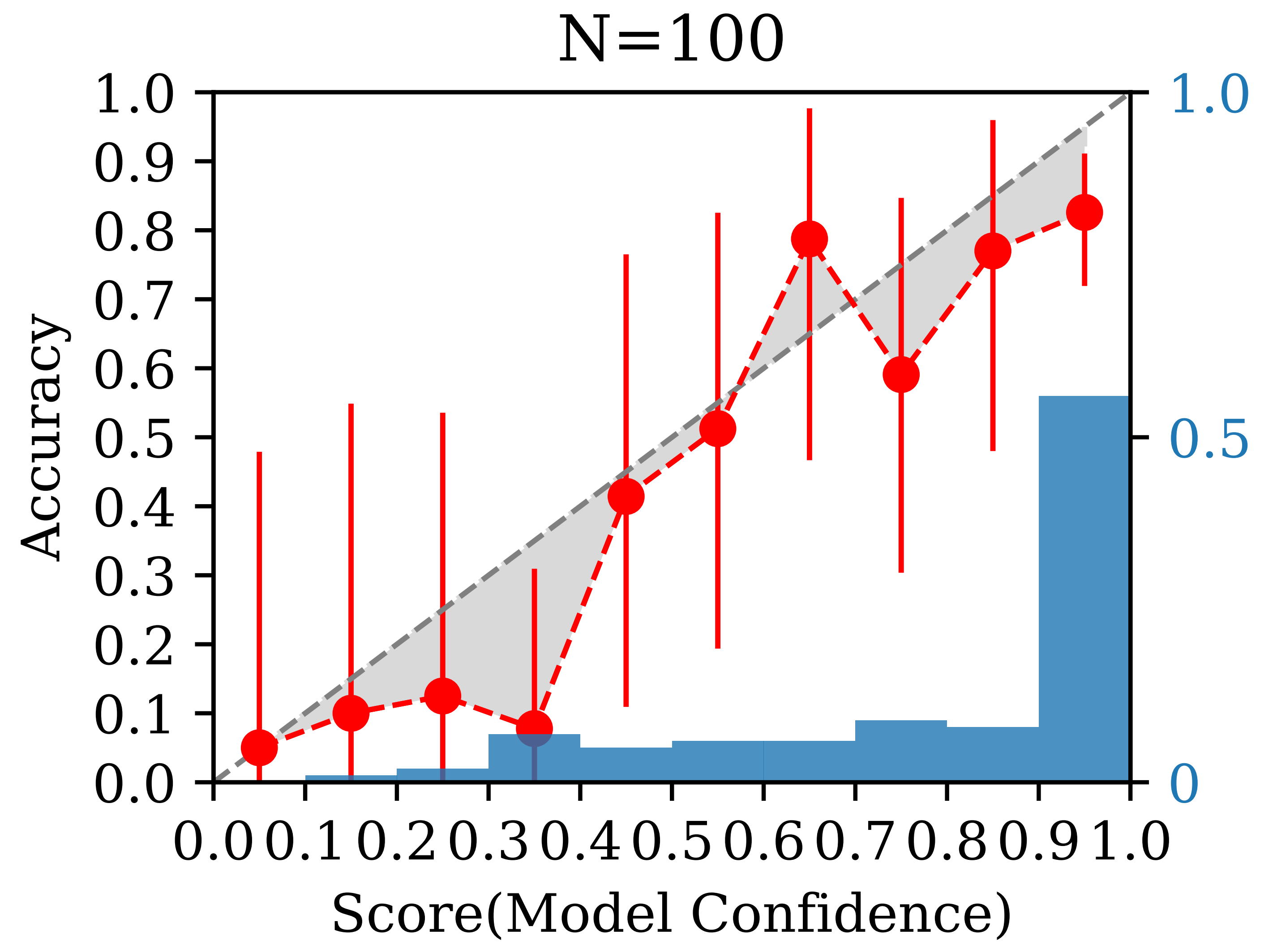


classwise accuracy

for ResNet-110 on CIFAR-100

Accuracy of the b-th bin:

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



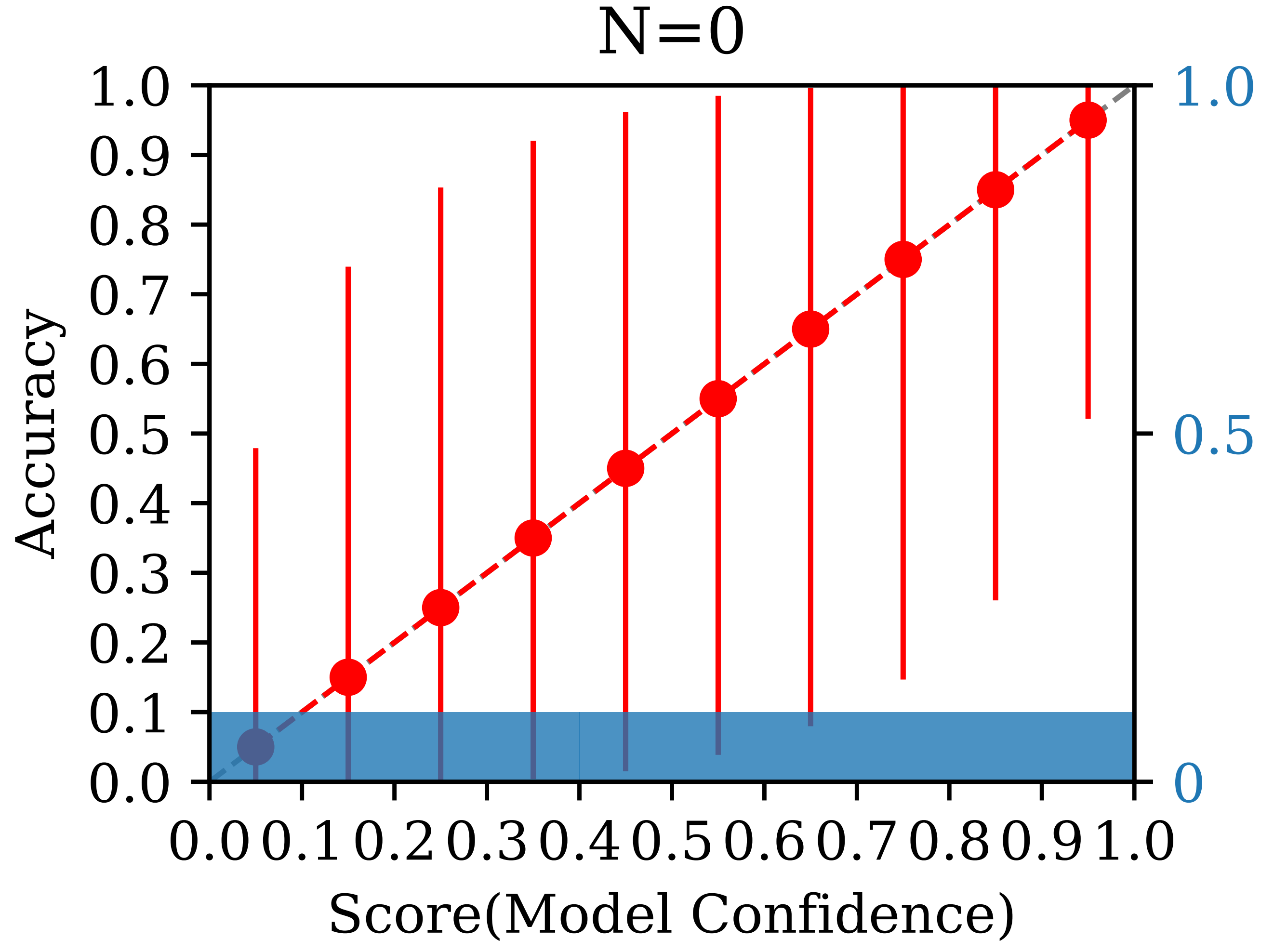
binwise accuracy

for ResNet-110 on CIFAR-100

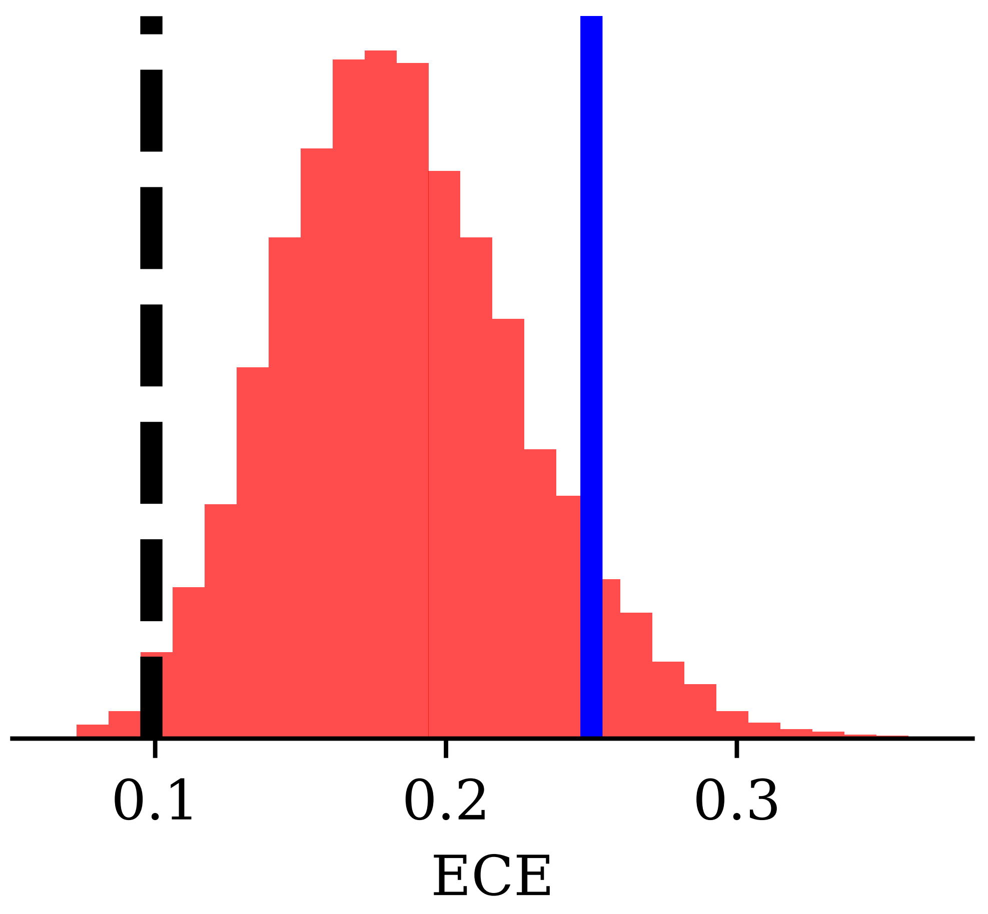
BAYESIAN ASSESSMENT: HOW CALIBRATED

Use self-assessment as informative prior:
assume the classifier is calibrated *a priori*

$\theta_b = \text{Beta}(\alpha_b, \beta_b)$



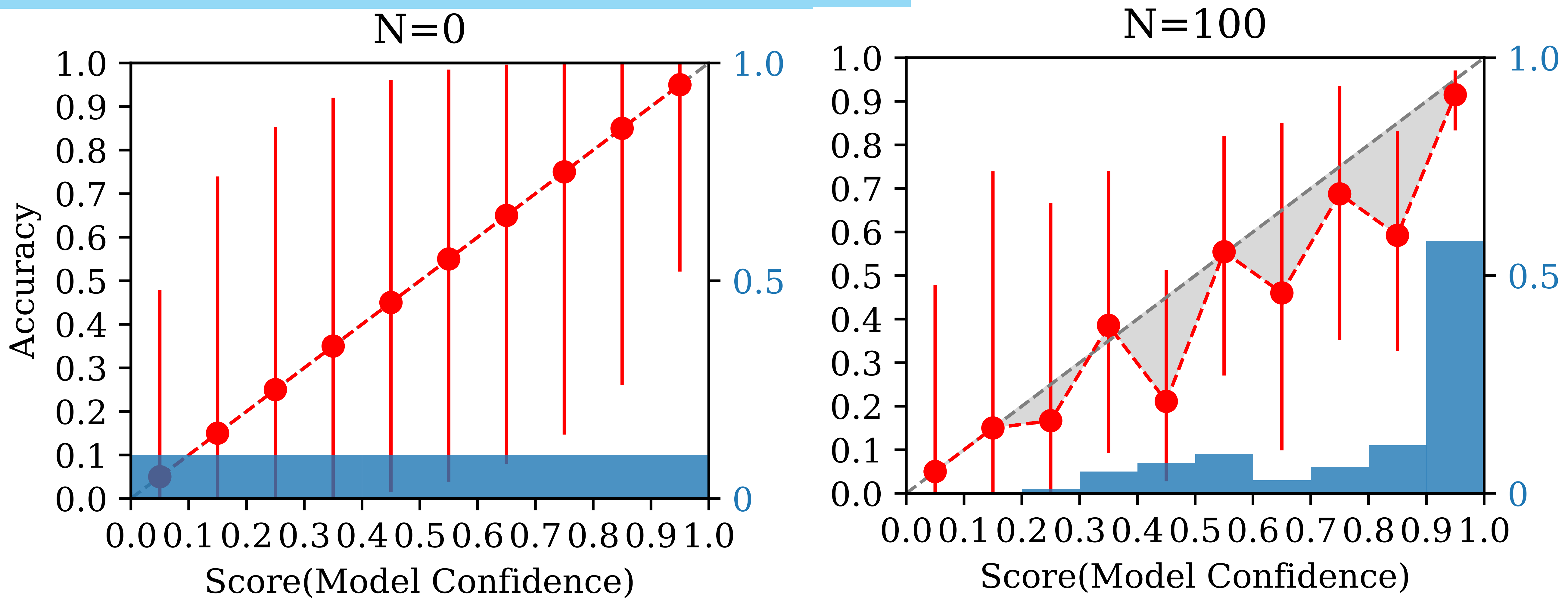
$$\text{ECE} = \sum_{b=1}^{B} p_b |\theta_b - s_b|$$



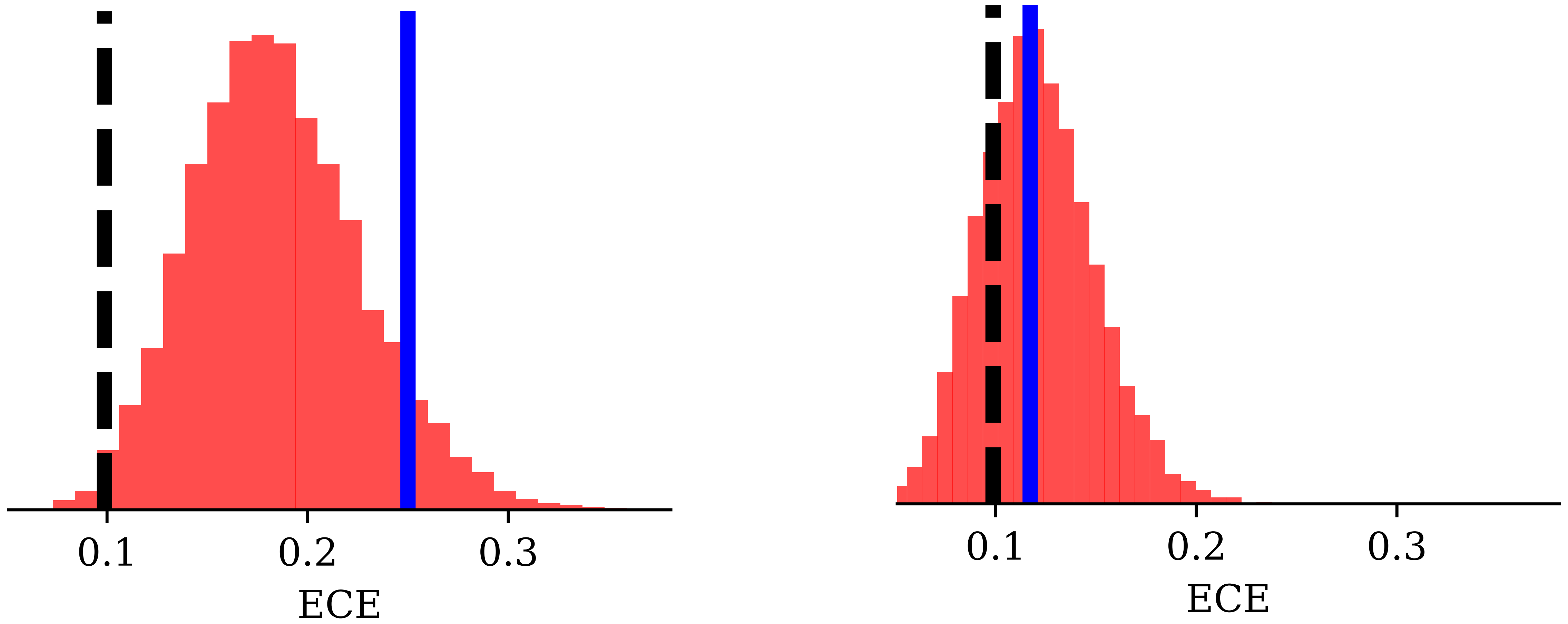
BAYESIAN ASSESSMENT: HOW CALIBRATED

Use self-assessment as informative prior:
assume the classifier is calibrated *a priori*

$$\theta_b = \text{Beta}(\alpha_b, \beta_b)$$



$$\text{ECE} = \sum_{b=1}^{B} p_b |\theta_b - s_b|$$



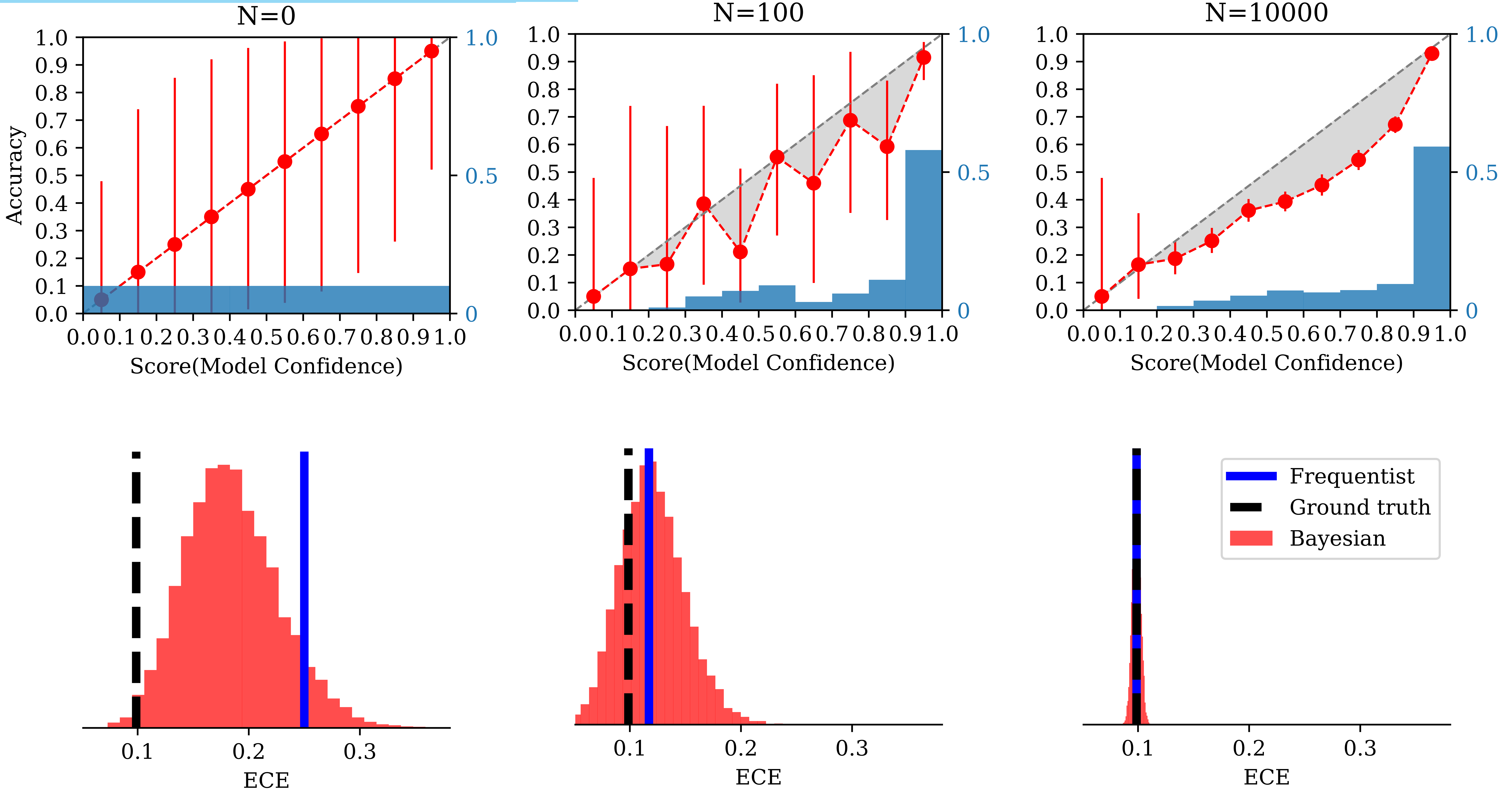
UCIrvine

BAYESIAN ASSESSMENT: HOW CALIBRATED

Use self-assessment as informative prior:
assume the classifier is calibrated *a priori*

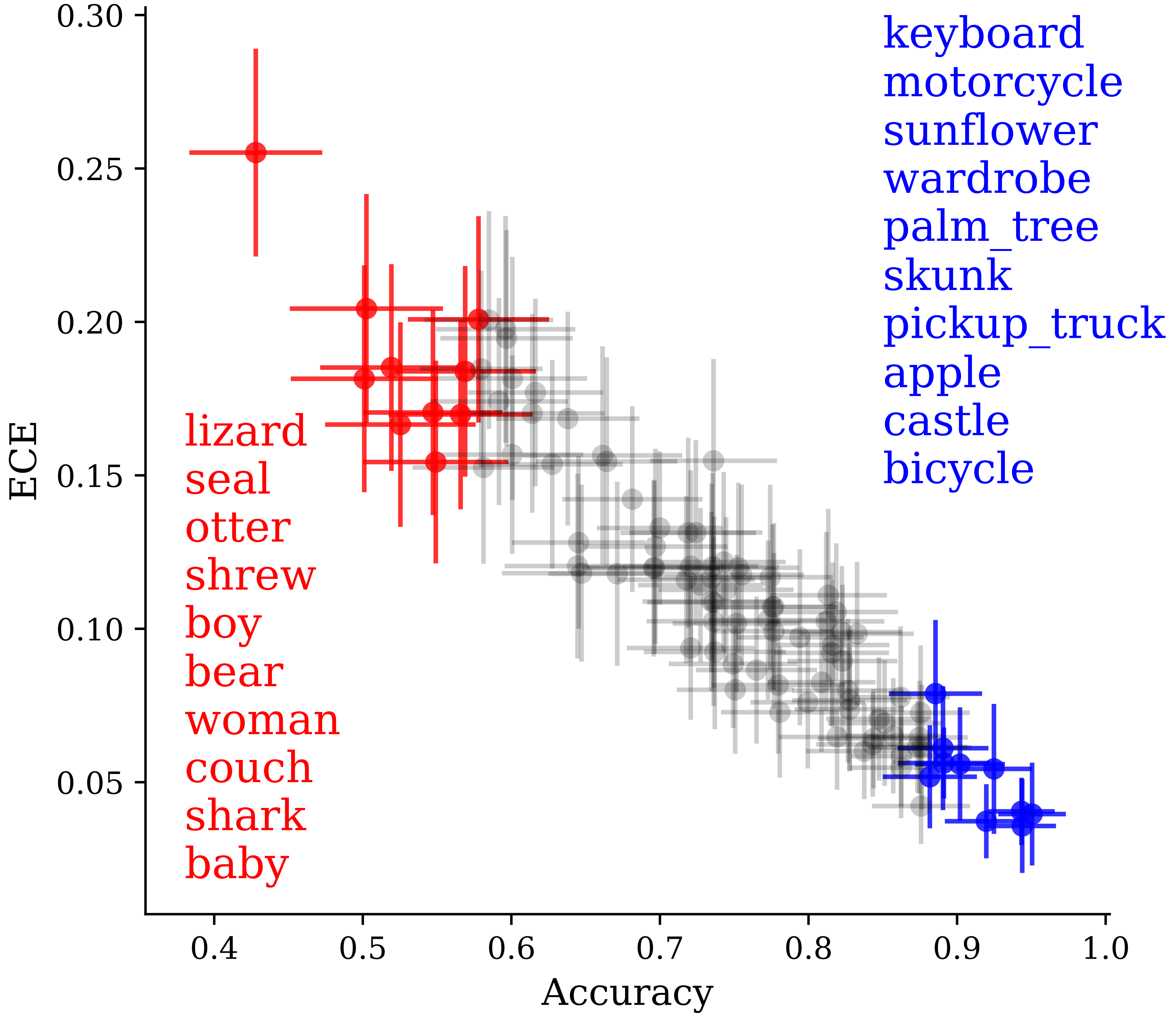
$$\theta_b = \mathrm{Beta}(\alpha_b, \beta_b)$$

$$\mathrm{ECE} = \sum_{b=1}^{B} p_b |\theta_b - s_b|$$



UCIRVINE

BAYESIAN ASSESSMENT: HOW CALIBRATED



classwise accuracy vs classwise ECE

ResNet-110 on CIFAR-100

Accuracy of the b-th bin of the k-th predicted class:

$$\theta_{kb} = \text{Beta}(\alpha_{kb}, \beta_{kb})$$

$$k = 1,2,\cdots,K; \quad b = 1,2,\cdots,B$$

SUMMARY

- ✓ How **accurate**?
- ✓ How **calibrated**?
- ◉ How **fair**?
- ◉ …other metrics…

- ✓ And how much **confidence** should we have in this assessment?
- ◉ How to **increase our confidence** given the labeling budget?

classwise accuracy

Accuracy of the k-th predicted class:

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

binwise accuracy(ECE)

Accuracy of the *b*-th bin:

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

Classwise ECE

Accuracy of the *b*-th bin of the *k*-th predicted class:

$$\theta_{kb} = \text{Beta}(\alpha_{kb}, \beta_{kb})$$

$$k = 1,2,\cdots,K; \quad b = 1,2,\cdots,B$$

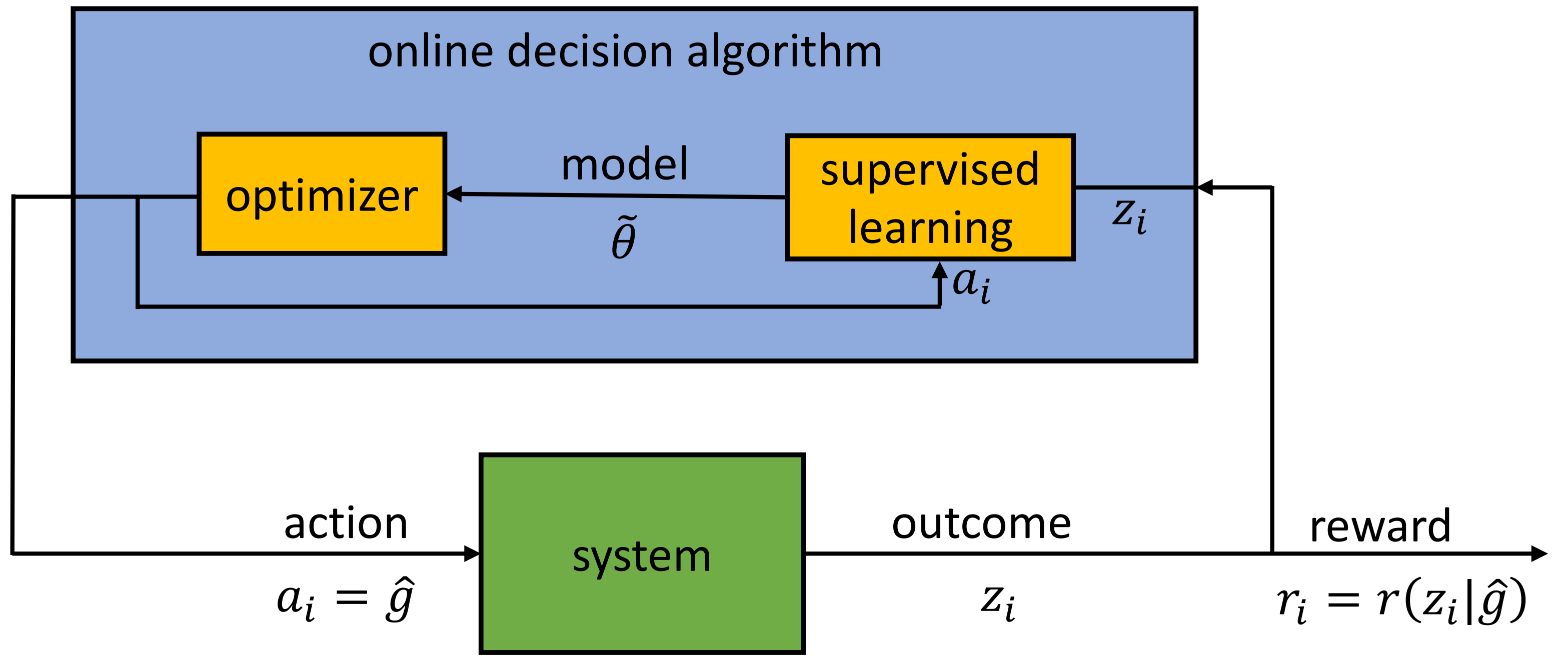
ROAD MAP

Bayesian assessment

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

active Bayesian assessment

2. **Reduce uncertainty** of assessment, with **actively labeled data** selected from a pool of unlabeled data



assess with **unlabeled** data

3. **Reduce uncertainty** of assessment, by leveraging both **labeled and unlabeled data**

ASSESSMENT TASKS

	Partition of Input space	Index of groups
classwise accuracy Accuracy of the k-th predicted class: $\theta_k = \text{Beta}(\alpha_k, \beta_k), k = 1,2,\cdots,K$	Predicted class	$g = k$
binwise accuracy(ECE) Accuracy of the *b*-th bin: $\theta_b = \text{Beta}(\alpha_b, \beta_b), b = 1,2,\cdots,B$	Model score	$g = b$
Classwise ECE Accuracy of the *b*-th bin of the *k*-th predicted class: $\theta_{kb} = \text{Beta}(\alpha_{kb}, \beta_{kb})$ $k = 1,2,\cdots,K; \quad b = 1,2,\cdots,B$	Predicted class X model score	$g = kb$

ASSESSMENT TASKS

	Partition of Input space	Index of groups
classwise accuracy Accuracy of the k-th predicted class: $\theta_k = \text{Beta}(\alpha_k, \beta_k), k = 1,2,\cdots,K$	Predicted class	$g = k$
binwise accuracy(ECE) Accuracy of the *b*-th bin: $\theta_b = \text{Beta}(\alpha_b, \beta_b), b = 1,2,\cdots,B$	Model score	$g = b$
Classwise ECE Accuracy of the *b*-th bin of the *k*-th predicted class: $\theta_{kb} = \text{Beta}(\alpha_{kb}, \beta_{kb})$ $k = 1,2,\cdots,K; \quad b = 1,2,\cdots,B$	Predicted class X model score	$g = kb$

performance metrics to estimate $\theta = (\theta_0, \theta_1, \cdots, \theta_G)$

labeled data: $\{(x_i, y_i) \mid i = 1,2,\cdots,N\}$

label outcome (e.g. prediction correctness): $z_i = f_M(x_i, y_i)$

prior distribution of of metrics: $\theta \sim p(\theta)$

likelihood of label outcome: $z_i \sim q_\theta(z_i)$

posterior of metircs: $p(\theta \mid \mathcal{D}) = \dfrac{p(\theta) \cdot \prod_{i=1}^{N} q_\theta(z_i)}{\int_\theta p(\theta) \cdot \prod_{i=1}^{N} q_\theta(z_i)\, d\theta}$.

ASSESSMENT TASKS

Performance metrics to estimate $\theta = (\theta_0, \theta_1, \cdots, \theta_G)$

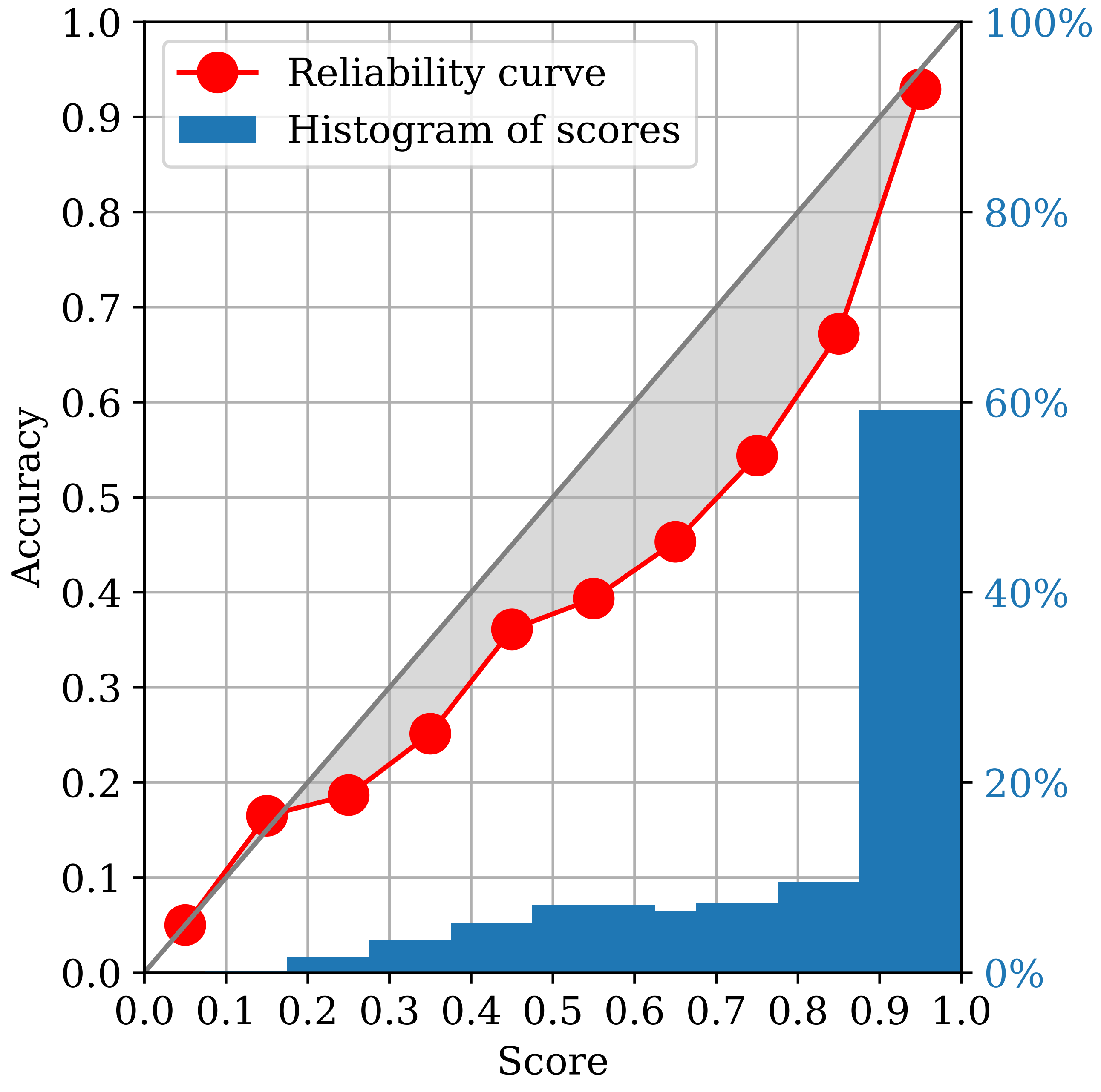
- **Estimation**: estimate model performance across all groups[1]
  - e.g. minimize RMSE $= (\sum_g p_g(\hat{\theta}_g - \theta_g^*)^2)^{\frac{1}{2}}$
- **Identification**: identify extreme groups, e.g. least accurate, least calibrated
  - e.g. identify $\hat{g} = \arg\max_g \theta_g$
- **Comparison**: compare performance between two groups
  - e.g. $\theta_0 > \theta_1$?

[1] Sawade et al. [2010] and Kumar and Raj [2018] use importance sampling and stratified sampling respectively to allocate labeling resources among different groups.

ASSESSMENT TASKS

Performance metrics to estimate $\theta = (\theta_0, \theta_1, \cdots, \theta_G)$

- **Estimation**: estimate model performance across all groups[1]
  - e.g. minimize RMSE $= (\sum_g p_g(\hat{\theta}_g - \theta_g^*)^2)^{\frac{1}{2}}$
- **Identification**: identify extreme groups, e.g. least accurate, least calibrated
  - e.g. identify $\hat{g} = \arg\max_g \theta_g$
- **Comparison**: compare performance between two groups
  - e.g. $\theta_0 > \theta_1$?



$$\mathrm{ECE} = \sum_{b=1}^{B} p_b |\theta_b - s_b|$$

[1] Sawade et al. [2010] and Kumar and Raj [2018] use importance sampling and stratified sampling respectively to allocate labeling resources among different groups.

ASSESSMENT TASKS

Performance metrics to estimate $\theta = (\theta_0, \theta_1, \cdots, \theta_G)$

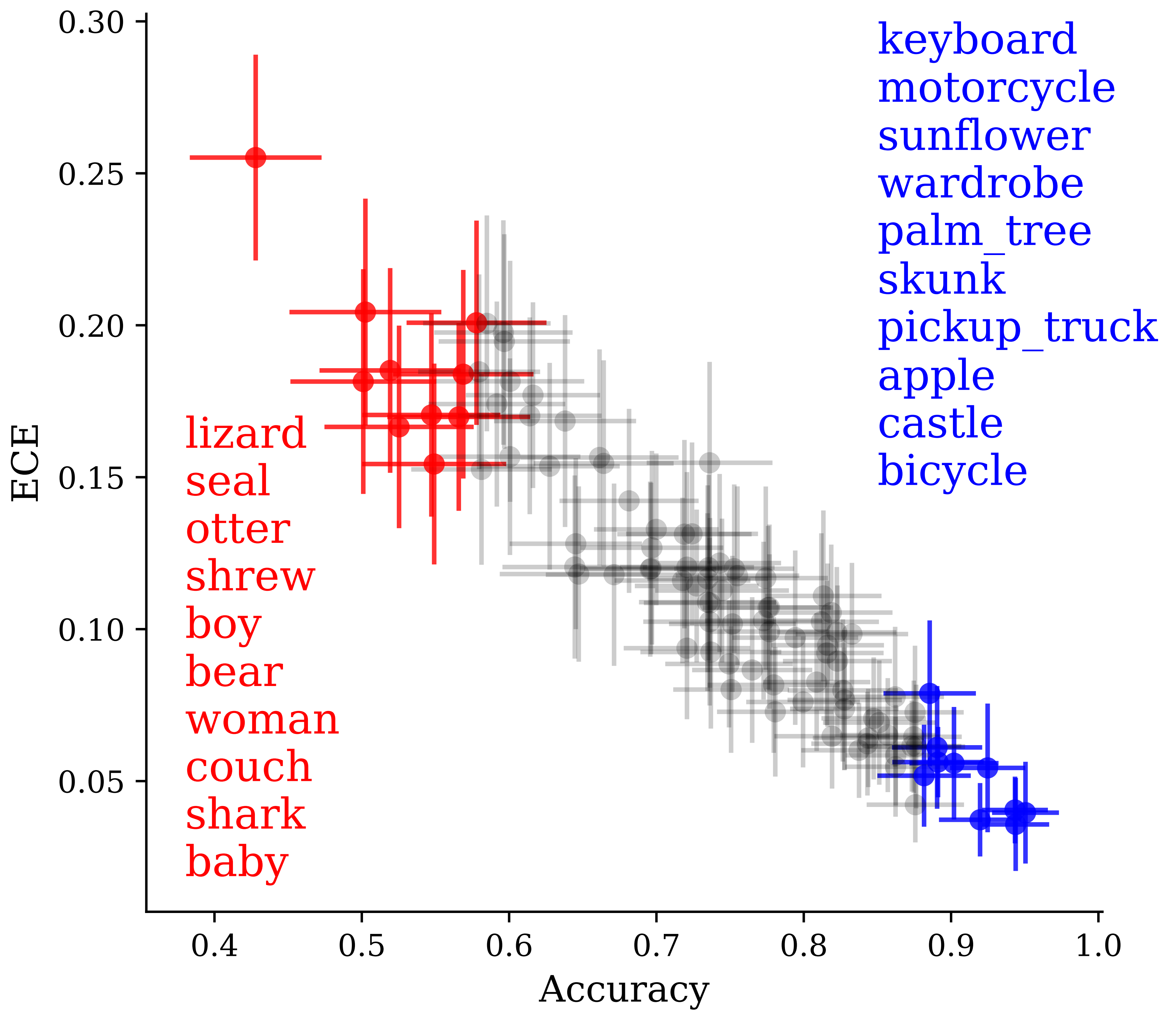
- **Estimation**: estimate model performance across all groups[1]
  - e.g. minimize RMSE $= (\sum_g p_g(\hat{\theta}_g - \theta_g^*)^2)^{\frac{1}{2}}$
- **Identification**: identify extreme groups, e.g. least accurate, least calibrated
  - e.g. identify $\hat{g} = \arg\max_g \theta_g$
- **Comparison**: compare performance between two groups
  - e.g. $\theta_0 > \theta_1$?

[1] Sawade et al. [2010] and Kumar and Raj [2018] use importance sampling and stratified sampling respectively to allocate labeling resources among different groups.

ASSESSMENT TASKS

Performance metrics to estimate $\theta = (\theta_0, \theta_1, \cdots, \theta_G)$

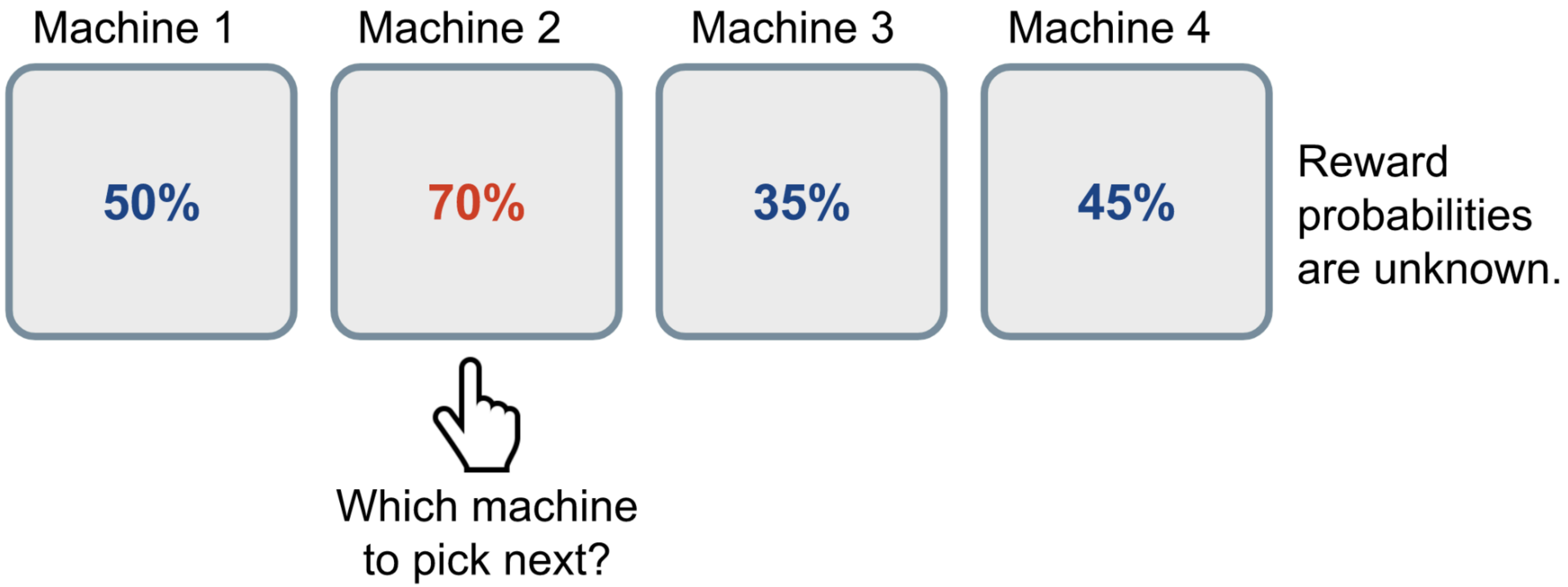
- **Estimation**: estimate model performance across all groups[1]
  - e.g. minimize RMSE $= (\sum_g p_g(\hat{\theta}_g - \theta_g^*)^2)^{\frac{1}{2}}$
- **Identification**: identify extreme groups, e.g. least accurate, least calibrated
  - e.g. identify $\hat{g} = \arg\max_g \theta_g$
- **Comparison**: compare performance between two groups
  - e.g. $\theta_0 > \theta_1$?



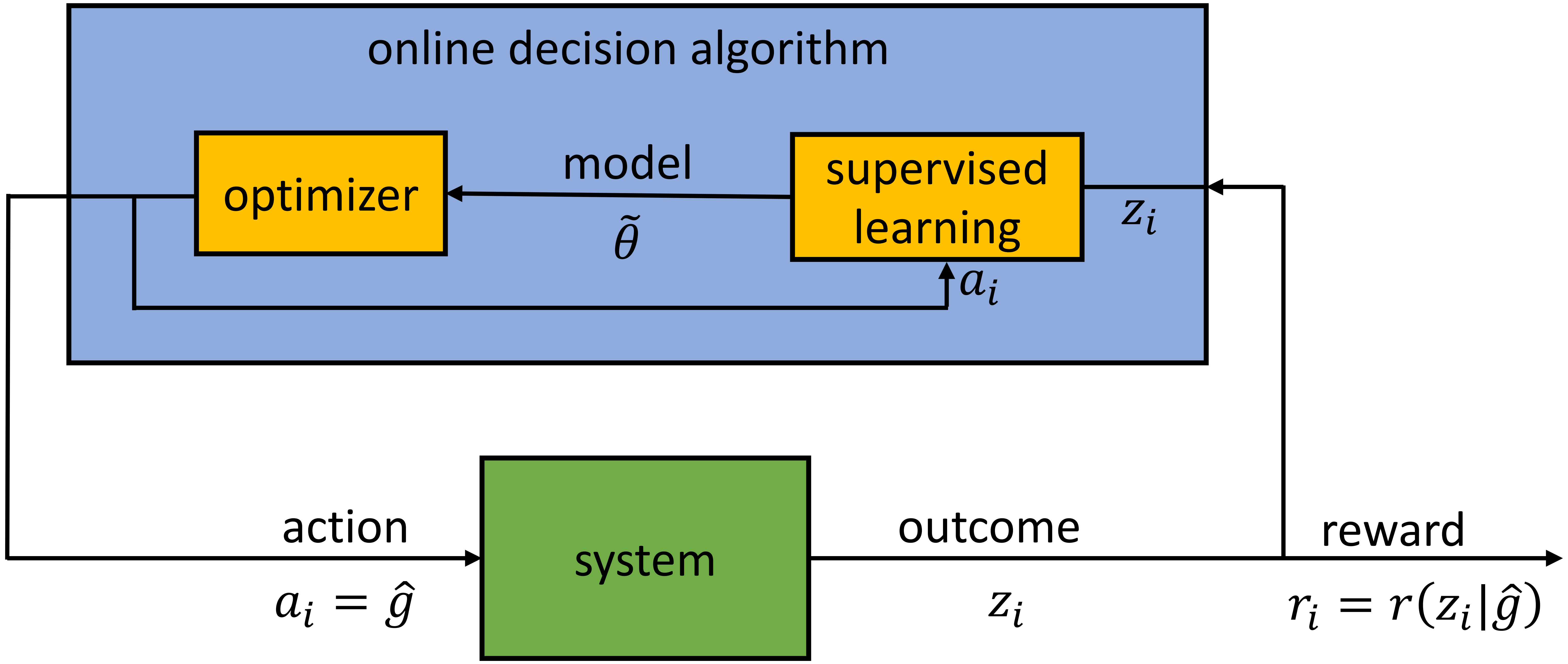
[1] Sawade et al. [2010] and Kumar and Raj [2018] use importance sampling and stratified sampling respectively to allocate labeling resources among different groups.

MULTI-ARMED BANDIT PROBLEMS

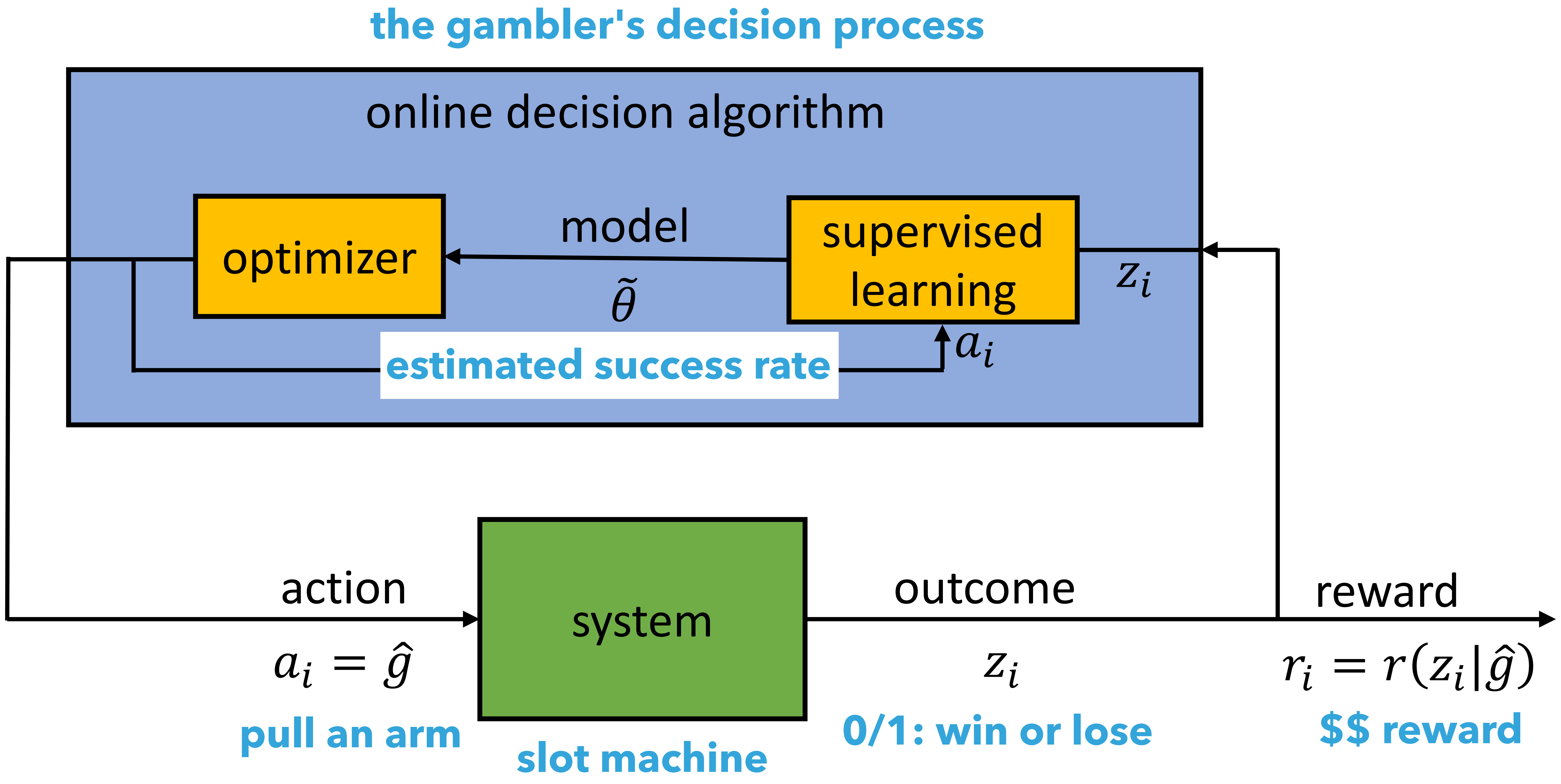
- reward probabilities of each arm are not told in advance
- **objective**: maximize cumulative reward
- **exploration-exploitation** trade-off
- **budget**: decide when to switch from more exploration to more exploitation
- **Sequential** decision making



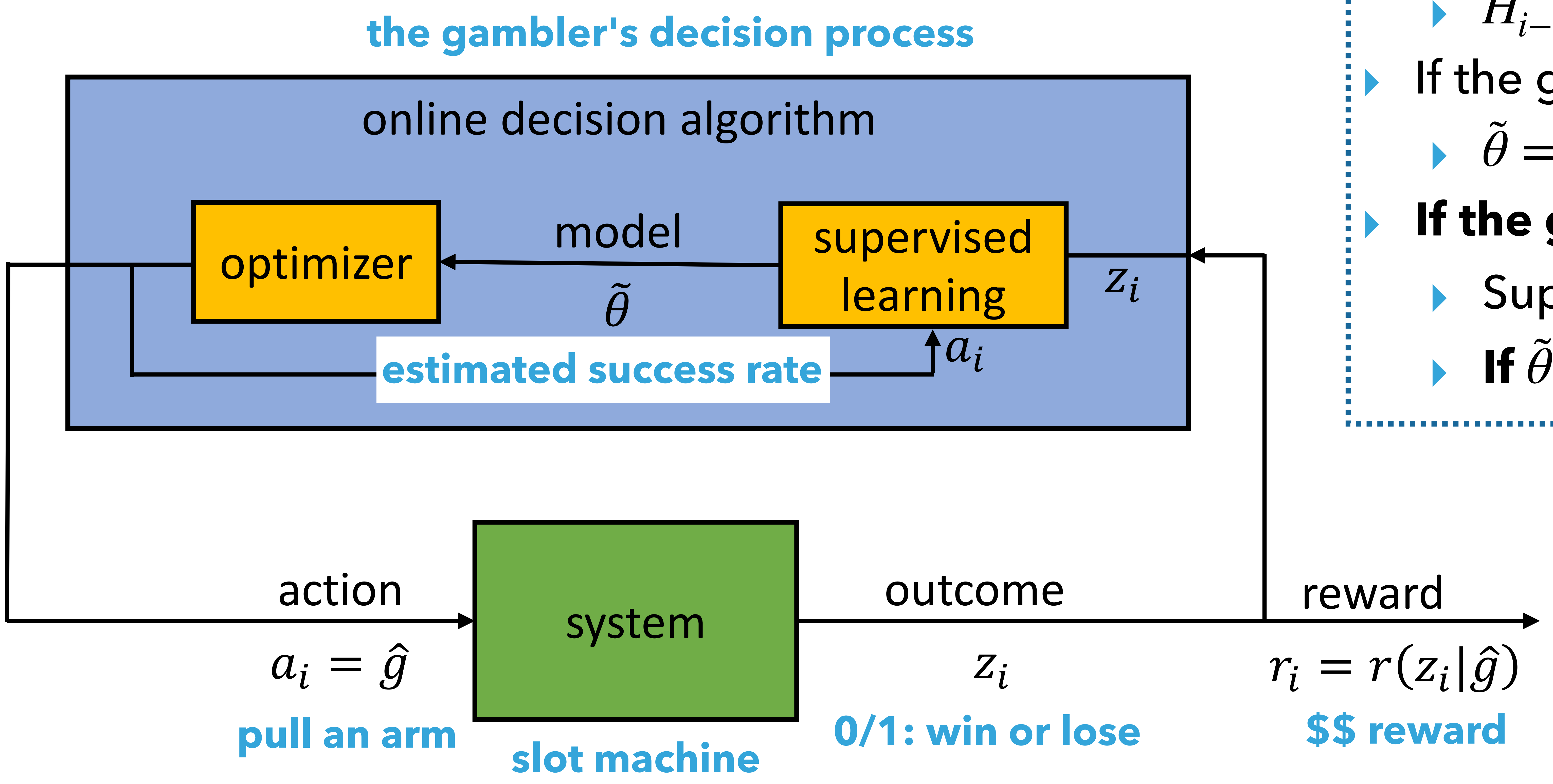
ONLINE DECISION SYSTEM



ONLINE DECISION SYSTEM

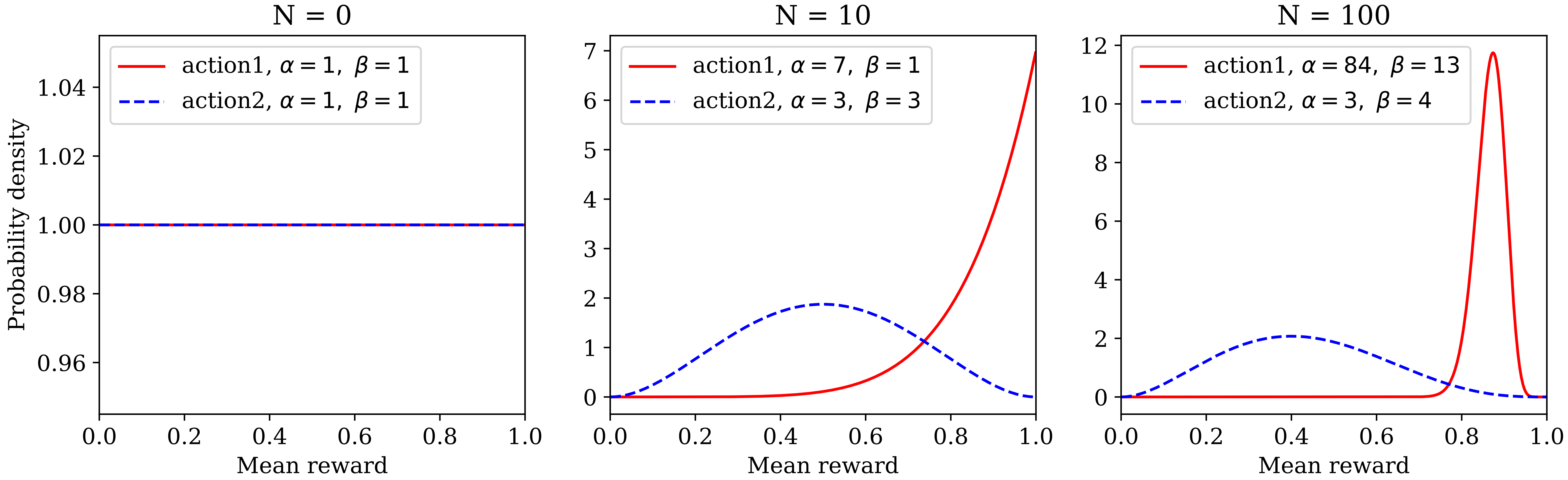


ONLINE DECISION SYSTEM



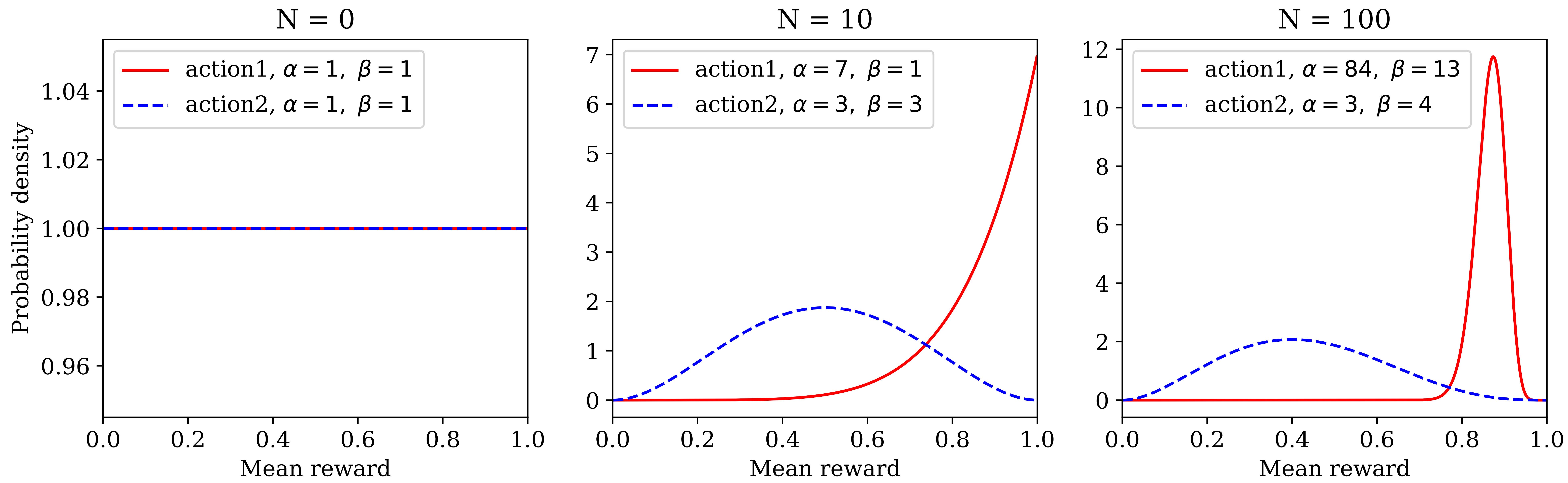
- At i-th step, fit decision model to H_{i-1}
  - $H_{i-1} = \{(a_j, z_j) \,|\, j = 1,2,\cdots, i-1\}$
- If the gambler is frequentist…
  - $\tilde{\theta} = \theta$
- **If the gambler is Bayesian…**
  - Supervised learning: $p_{i-1}(\theta)$
  - **If $\tilde{\theta} \sim p_{i-1}(\theta)$: Thompson sampling**

THOMPSON SAMPLING: EXAMPLE

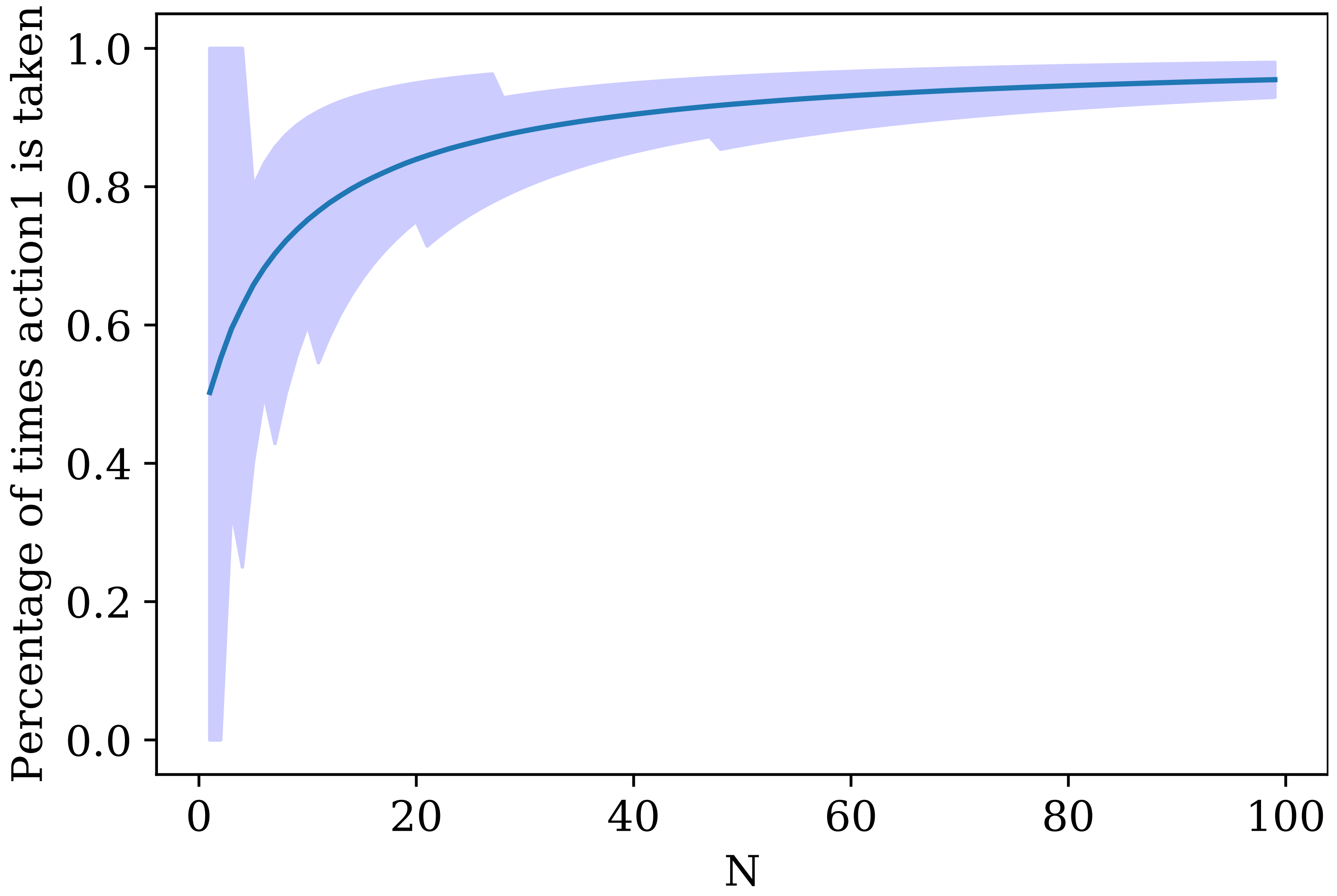


- True reward distributions
  - action1: $r \sim$ Bern(0.8)
  - action2: $r \sim$ Bern(0.2)

THOMPSON SAMPLING: EXAMPLE

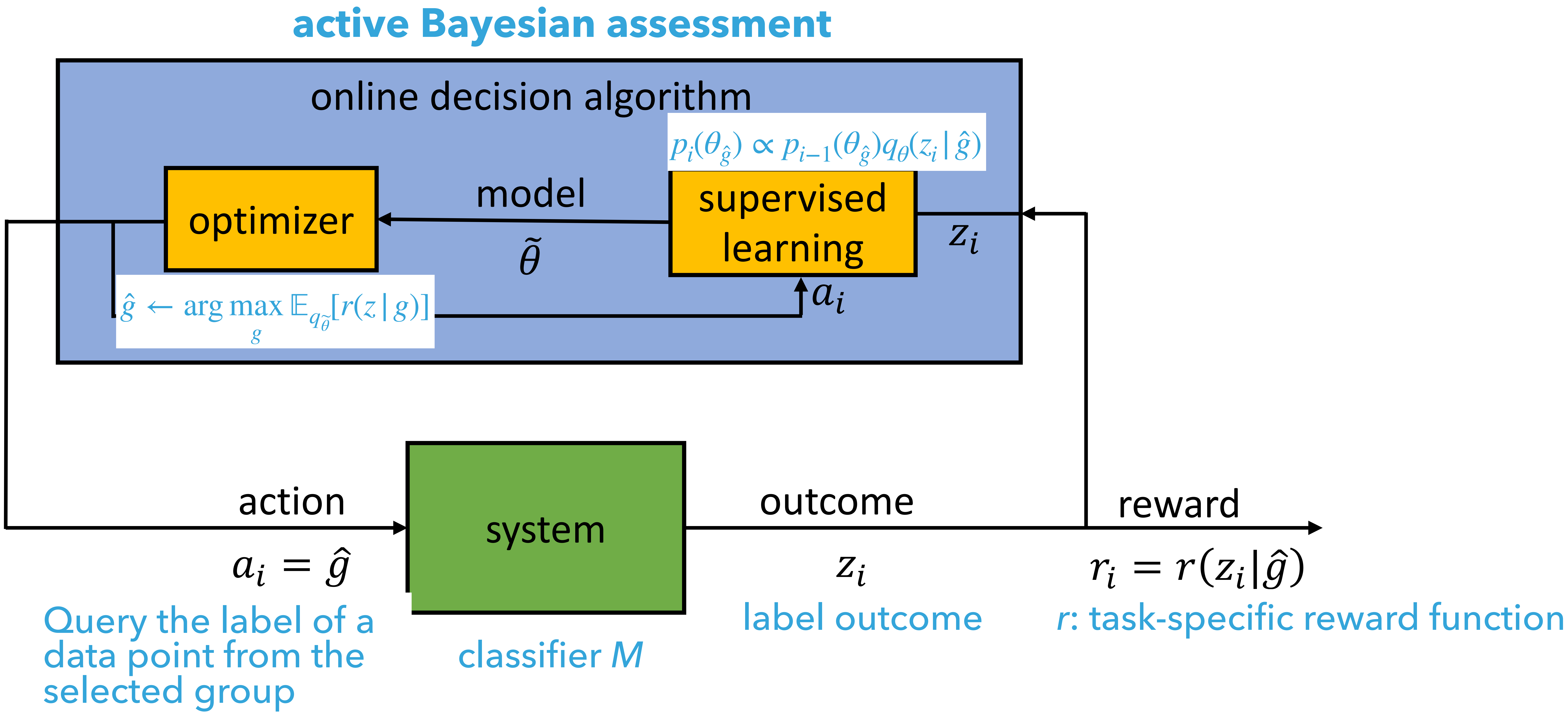


- True reward distributions
  - action1: $r \sim \text{Bern}(0.8)$
  - action2: $r \sim \text{Bern}(0.2)$

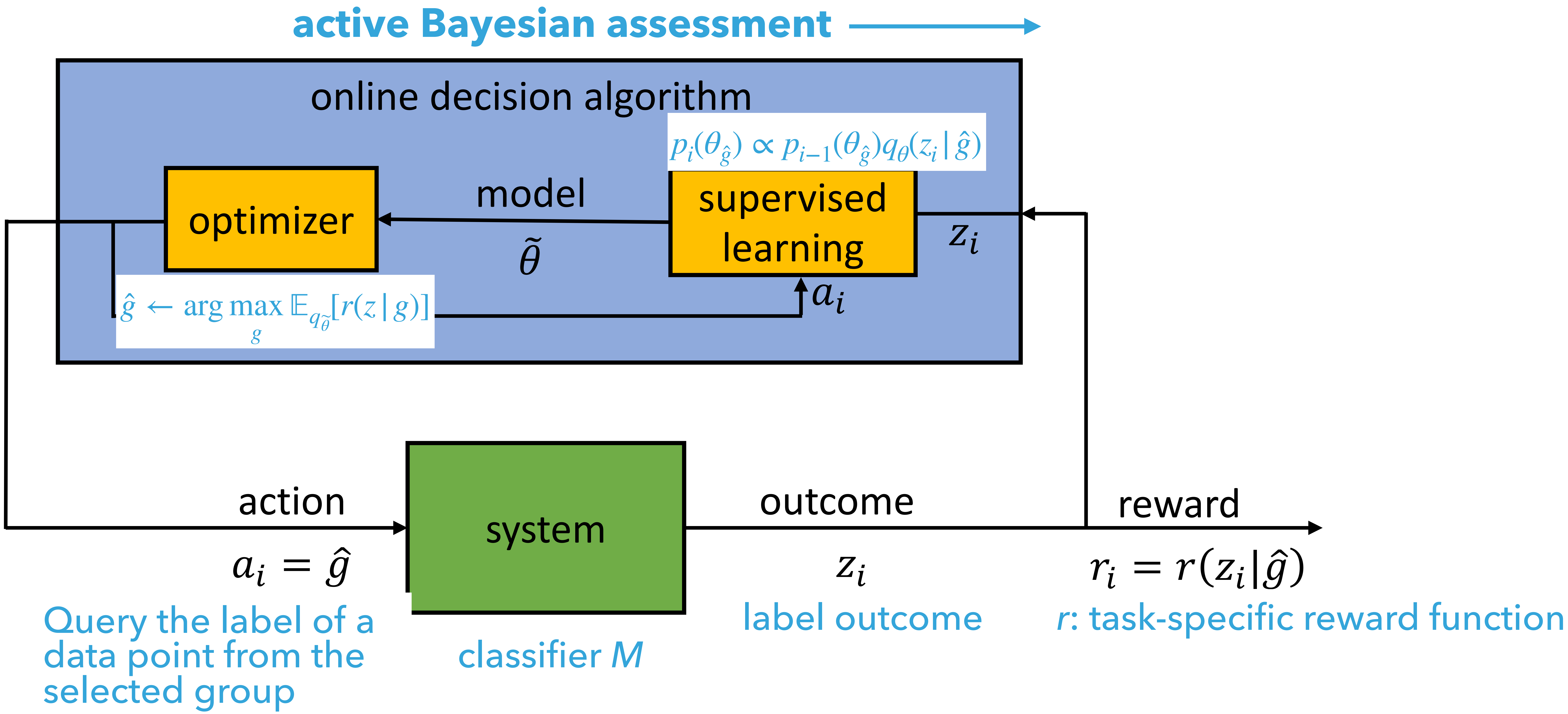


exploration -> exploitation

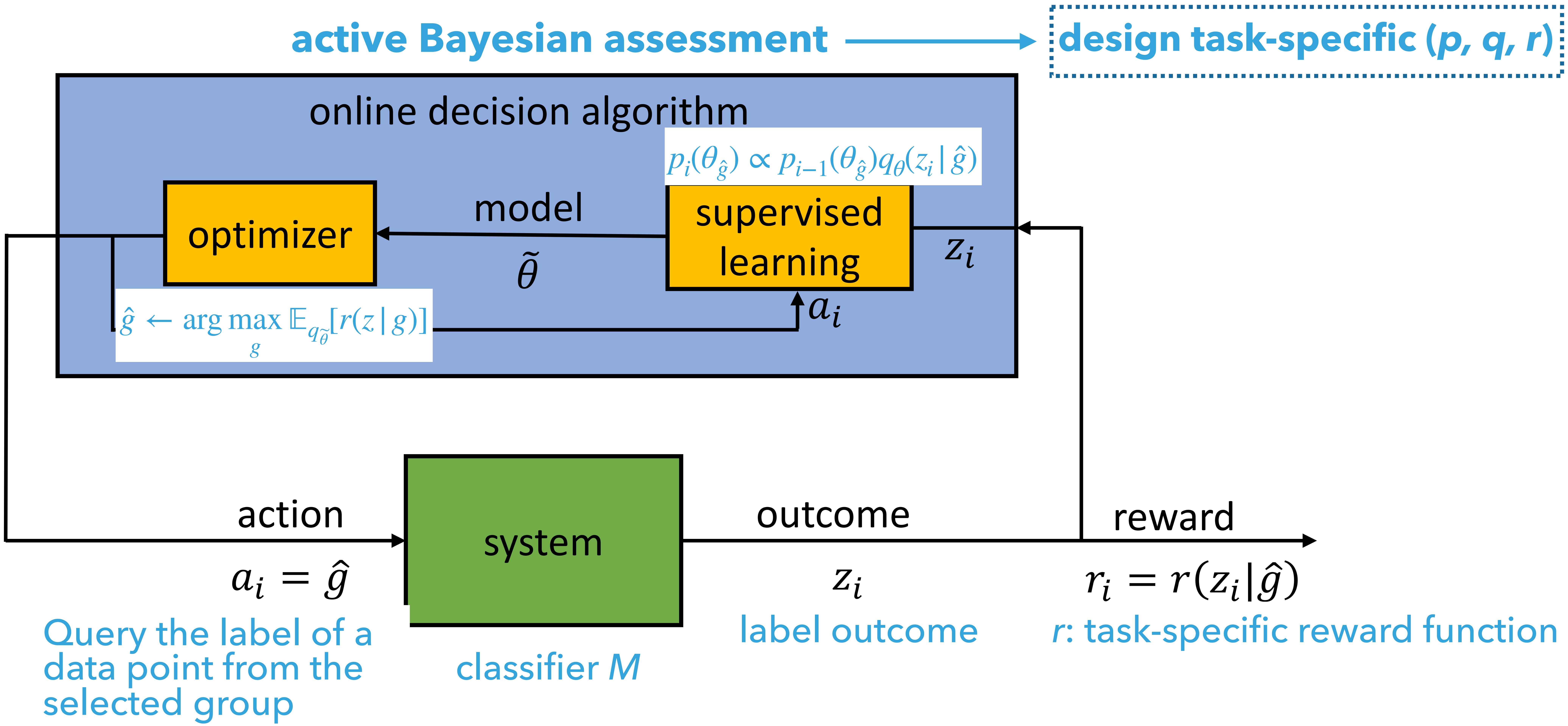
ACTIVE BAYESIAN ASSESSMENT



ACTIVE BAYESIAN ASSESSMENT



ACTIVE BAYESIAN ASSESSMENT



ACTIVE BAYESIAN ASSESSMENT

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

ACTIVE BAYESIAN ASSESSMENT

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

ACTIVE BAYESIAN ASSESSMENT

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

ACTIVE BAYESIAN ASSESSMENT

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

previously labeled data

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

EXPERIMENTS: MATERIAL

- Difference mode, varying size and number of classes
- Kudos to Robby for training the classification models

	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	$\text{BERT}_{\text{BASE}}$

EXAMPLE: 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 90%

EXAMPLE: 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 90%

We obtained similar performance gain across multiple datasets, prediction models, and assessment tasks

DISCUSSION

- **Other Bayesian active learning method to TS?**
  - Comparisons with alternative active learning algorithms
  - e.g. Epsilon-greedy, Bayesian upper-confidence bound
  - Thompson sampling is broadly more reliable and more consistent

- TS is not designed for **exploration-only problems** (best arm identification)
  - Comparisons between TS and top-two TS
  - TS and TTTS gave very similar performance

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

ROAD MAP

Bayesian assessment

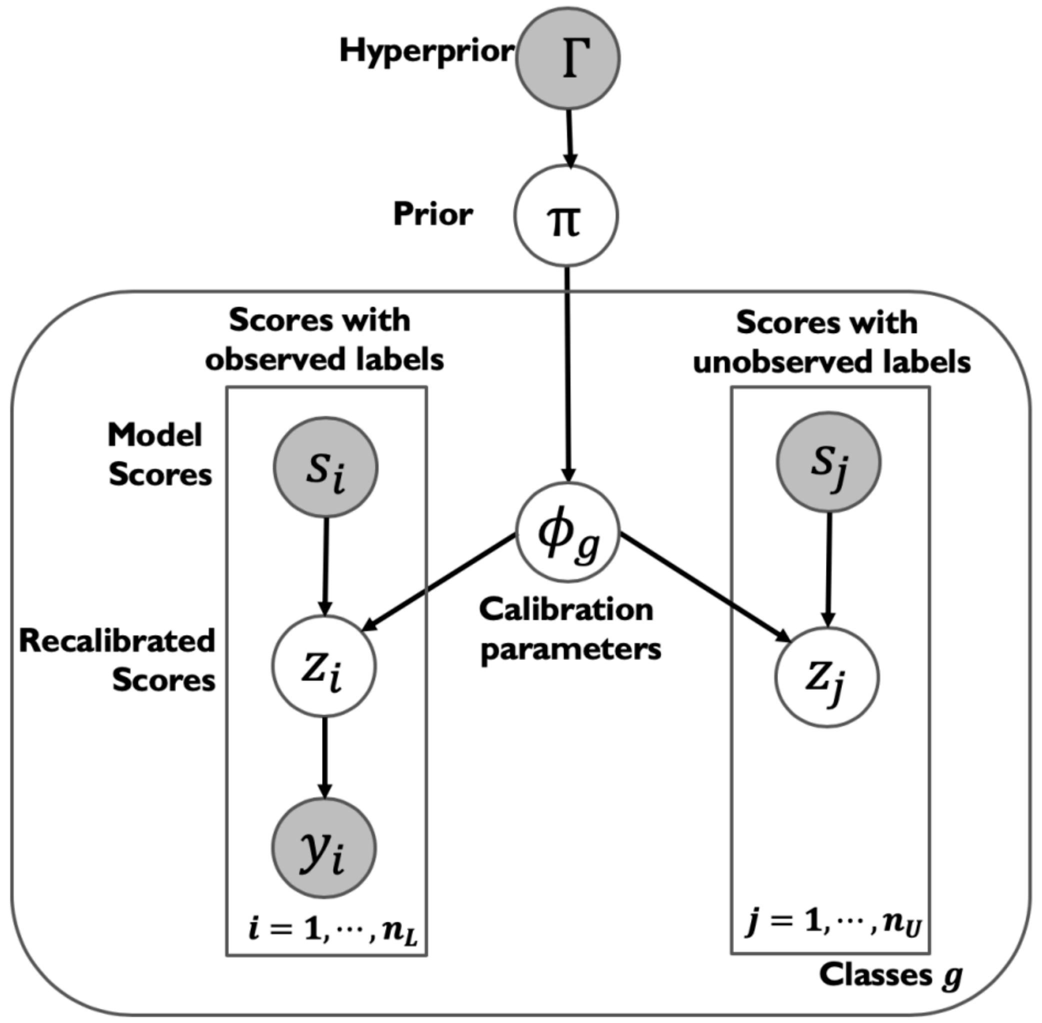
1. **Quantify uncertainty** of assessment with Bayesian models, with a set of **labeled data**

active Bayesian assessment

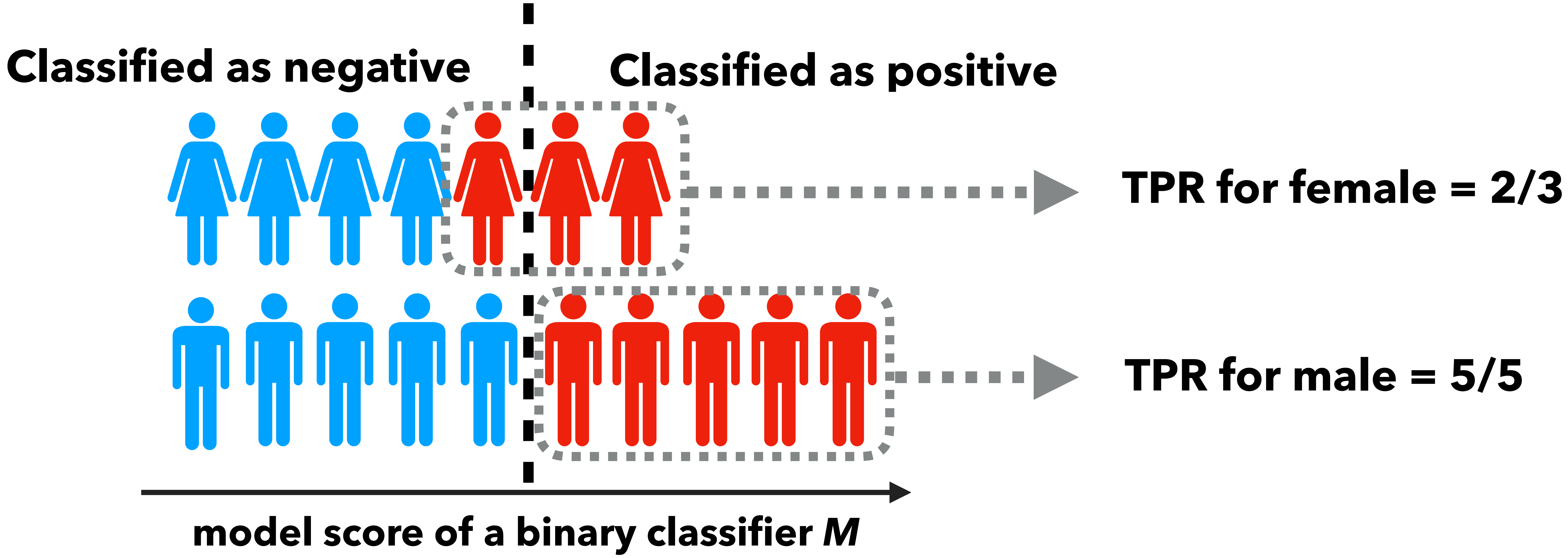
2. **Reduce uncertainty** of assessment, with **actively labeled data** selected from a pool of unlabeled data

assess with **unlabeled** data

3. **Reduce uncertainty** of assessment, by leveraging both **labeled and unlabeled data**



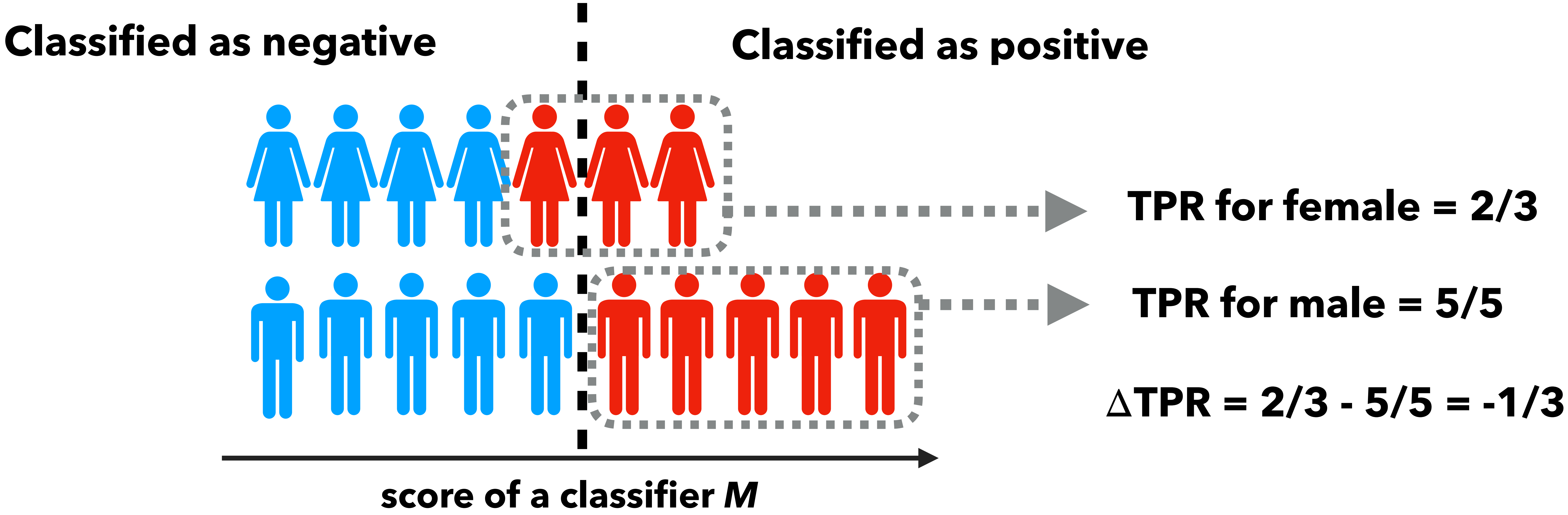
IS THE CLASSIFIER REALLY UNFAIR?

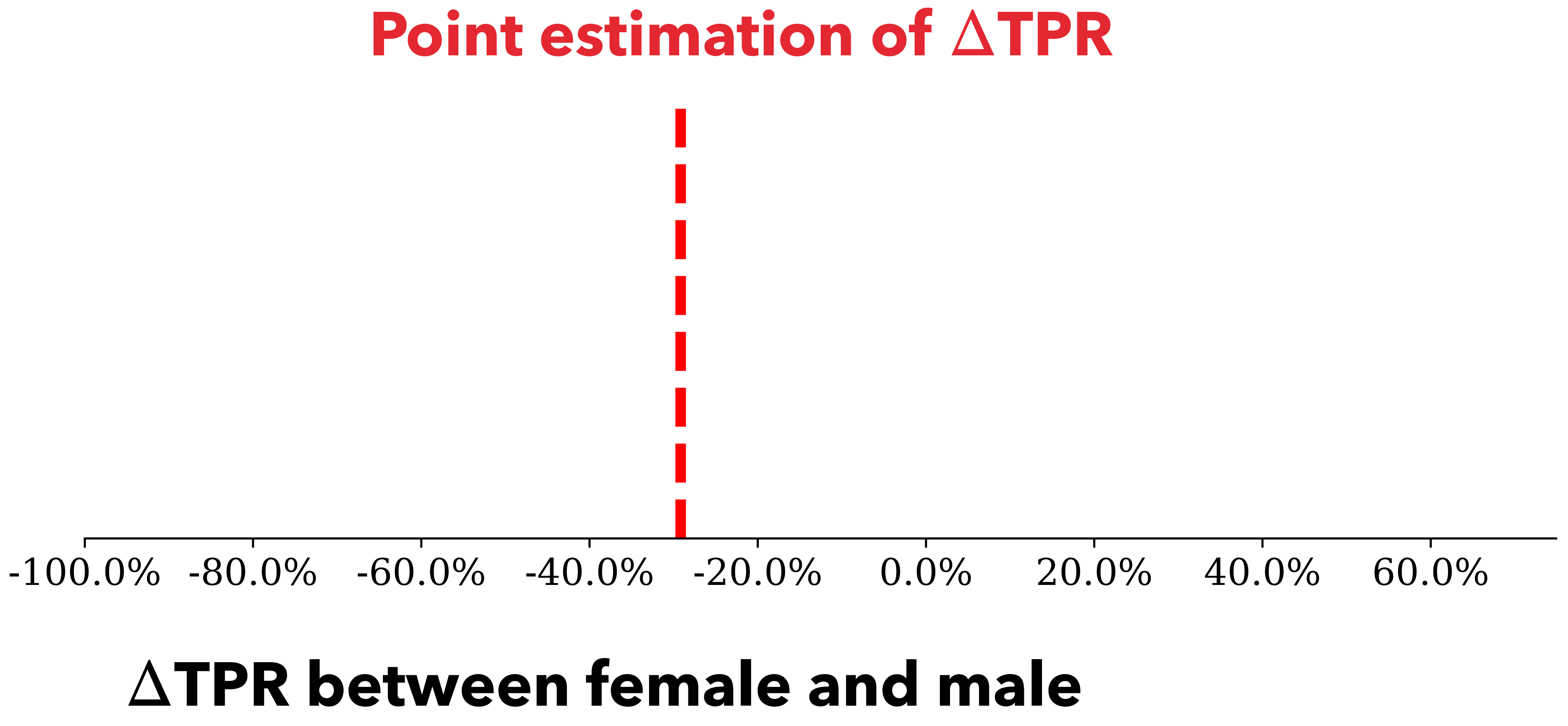


- **Equality of opportunity:** equal TPR across different groups[1]
  - *"people who pay back their loan, have an equal opportunity of getting the loan in the first place"*
- Due to small sample size, the estimated TPR is noisy!

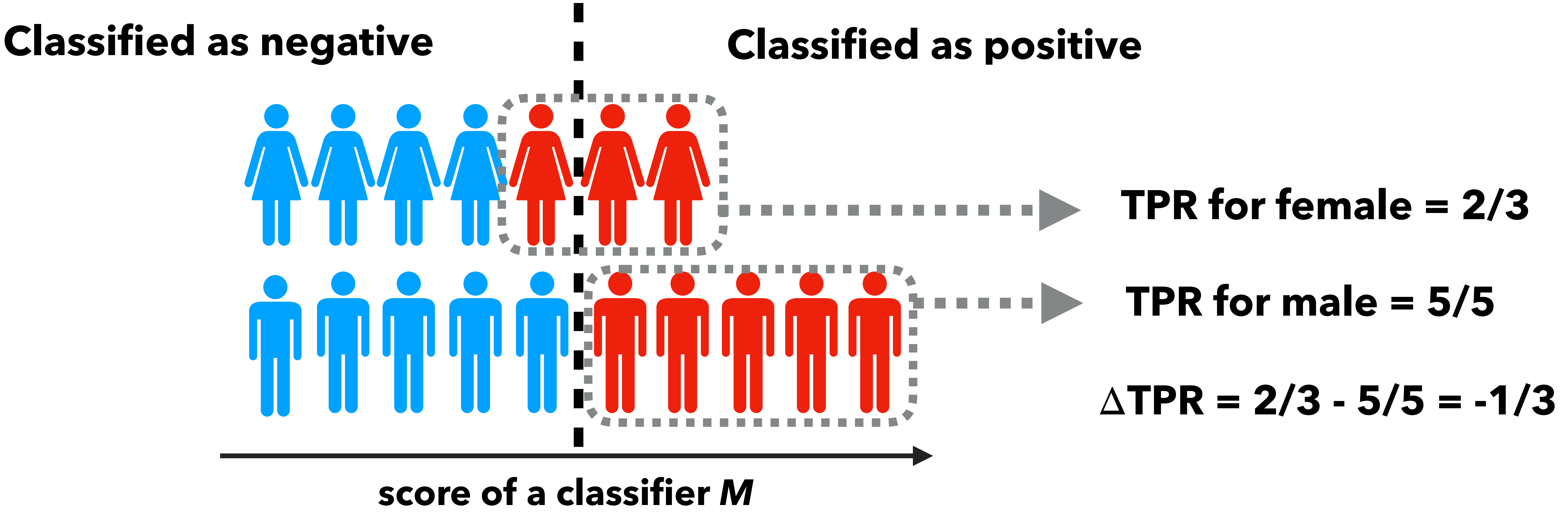
[1] "Equality of Opportunity in Supervised Learning". Hardt, Price & Srebro. NeurIPS 2016.

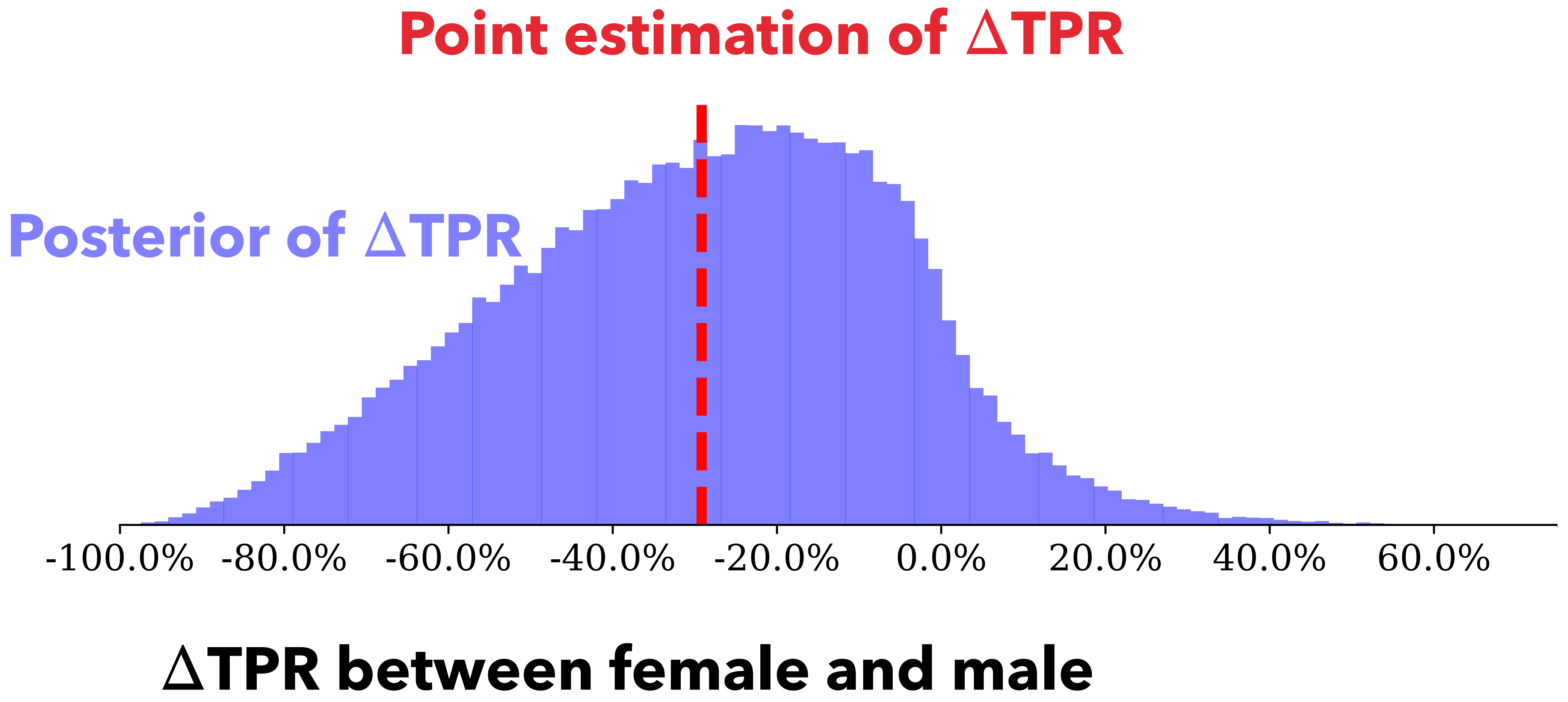
MODEL FAIRNESS METRICS WITH UNCERTAINTY



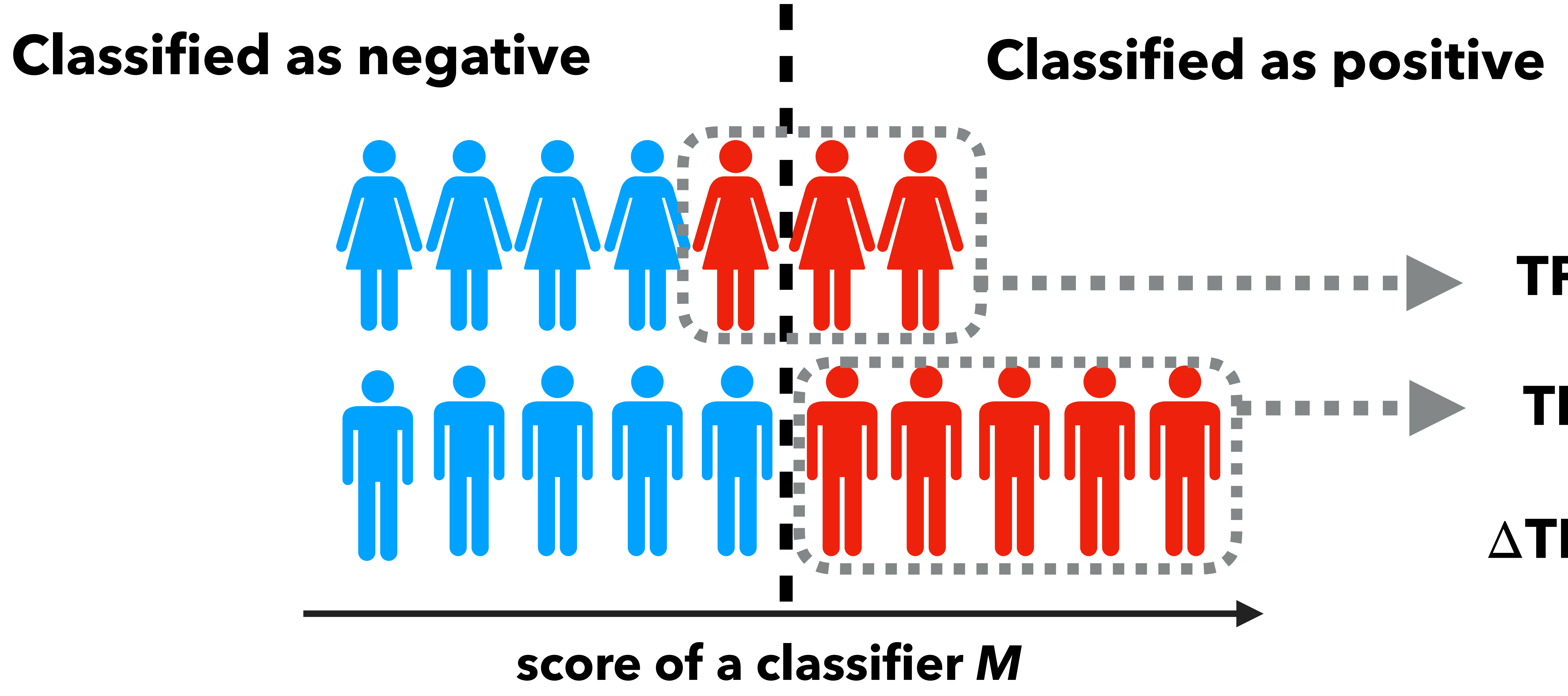


MODEL FAIRNESS METRICS WITH UNCERTAINTY





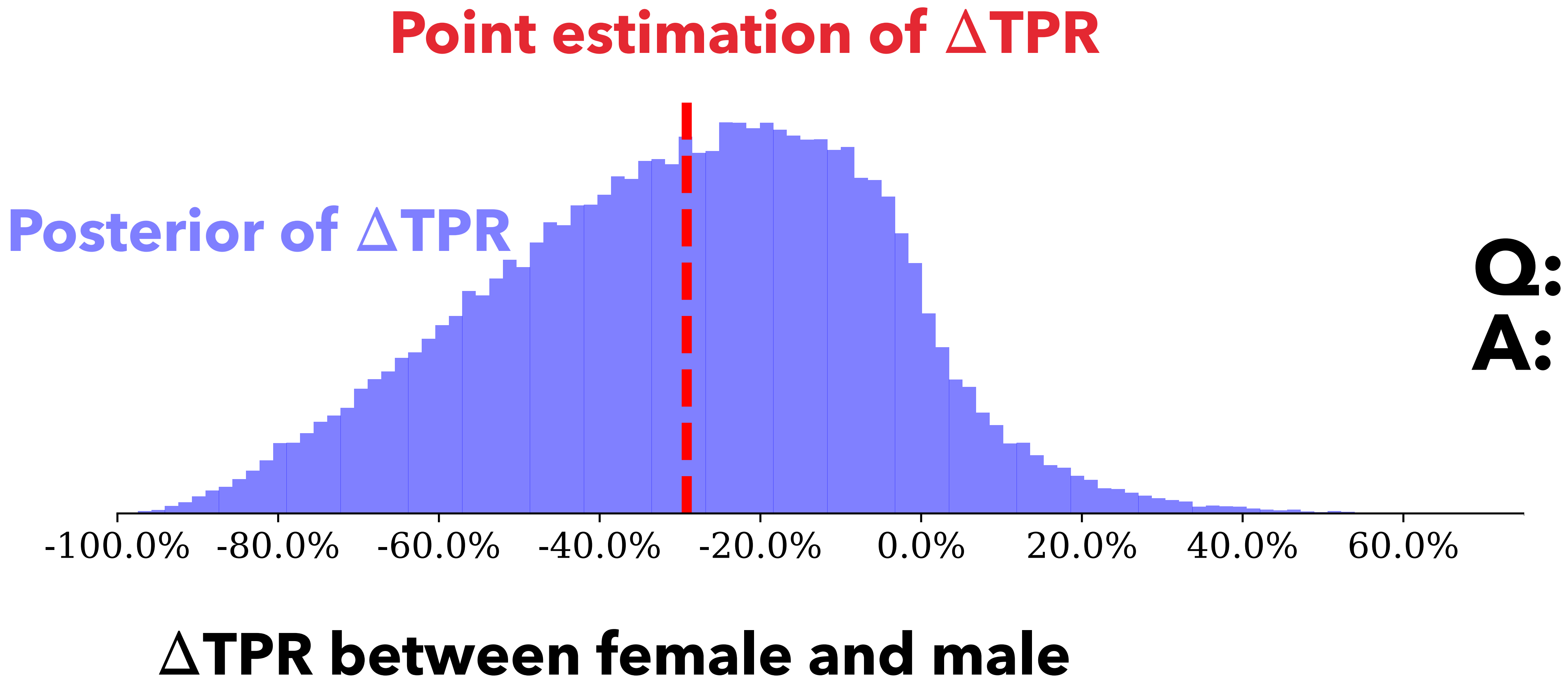
MODEL FAIRNESS METRICS WITH UNCERTAINTY



TPR for female = 2/3

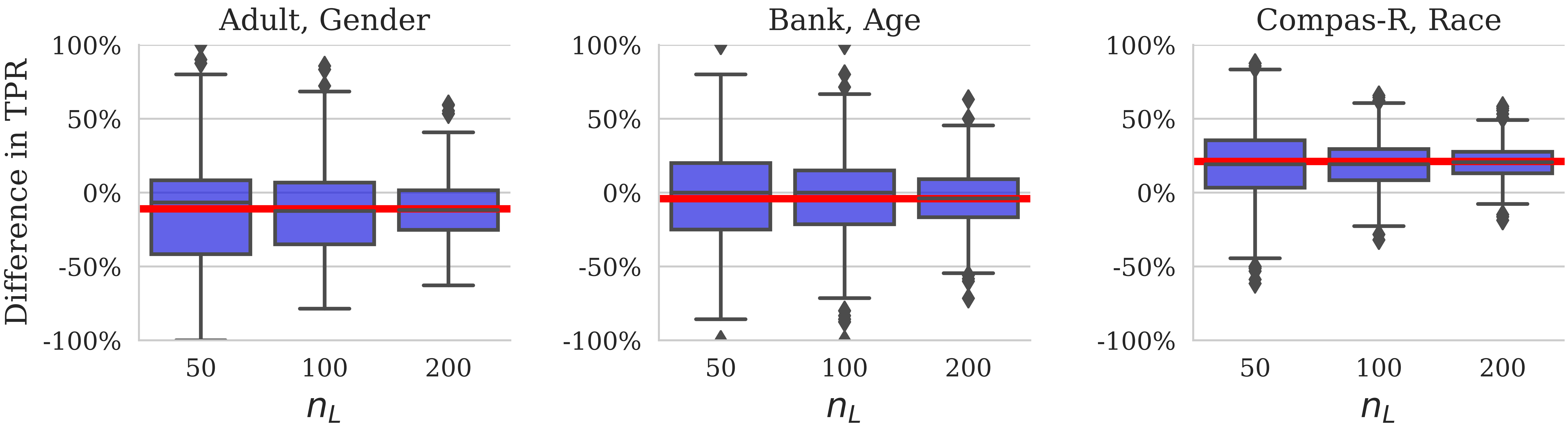
TPR for male = 5/5

ΔTPR = 2/3 - 5/5 = -1/3



Q: The uncertainty is high! How to reduce it?
A: Collect more data! Labeled or **unlabeled!**

HIGH UNCERTAINTY FOR REAL-WORLD DATA



whether income exceeds $50,000 per year

whether the individual has subscribed to a term deposit account or not

COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) risk assessment tool for recidivism

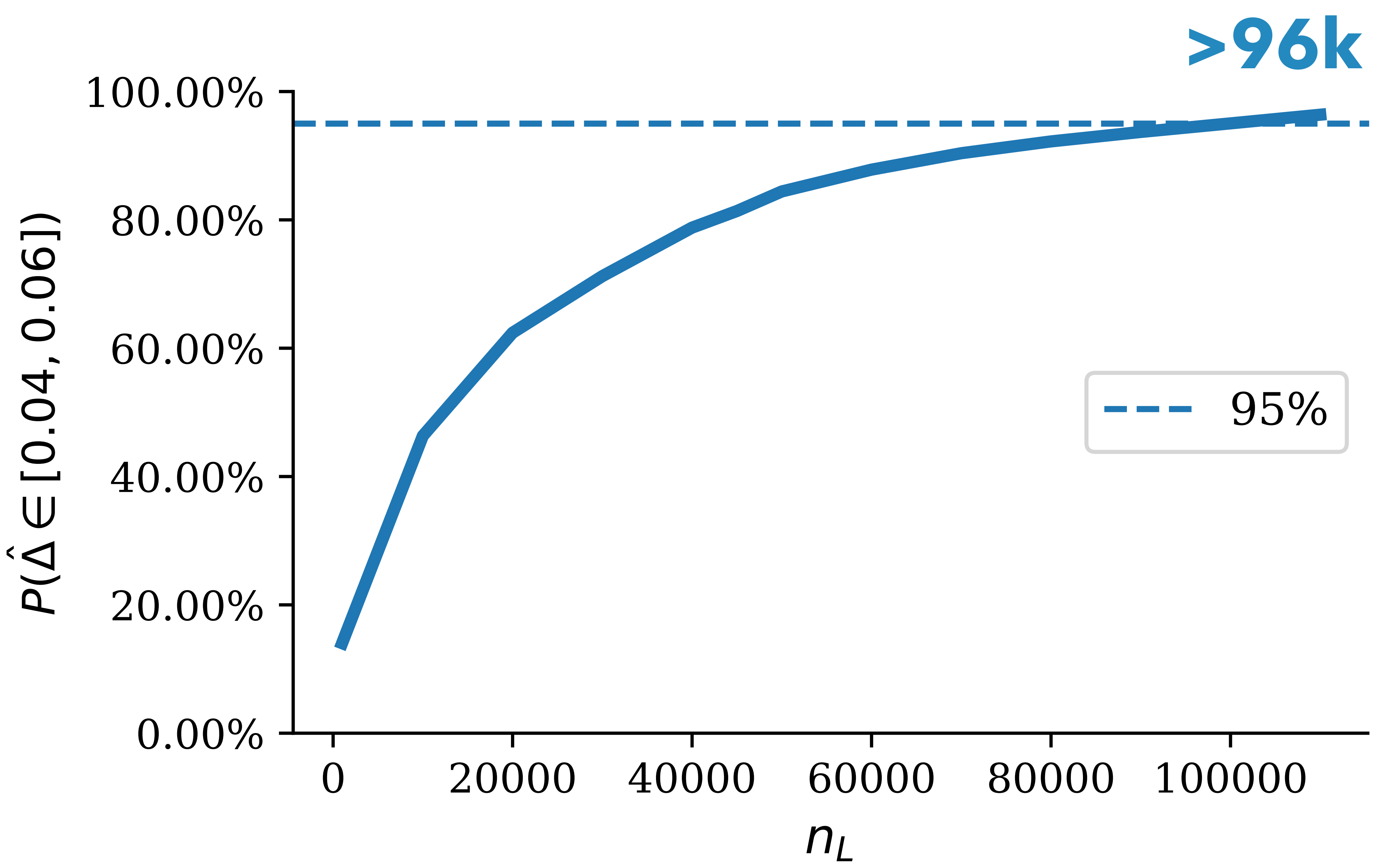
frequency-based estimates of the difference in true positive rate (TPR)

HOW MANY LABELED DATA DO I NEED TO COLLECT?

- Simulation:
  - p(g=0) = 20%
  - groupwise positive rates p(y = 1) are both 20%
  - the true groupwise TPRs are 95% and 90%.
- Compute frequentist estimation of ΔTPR for 10000 times

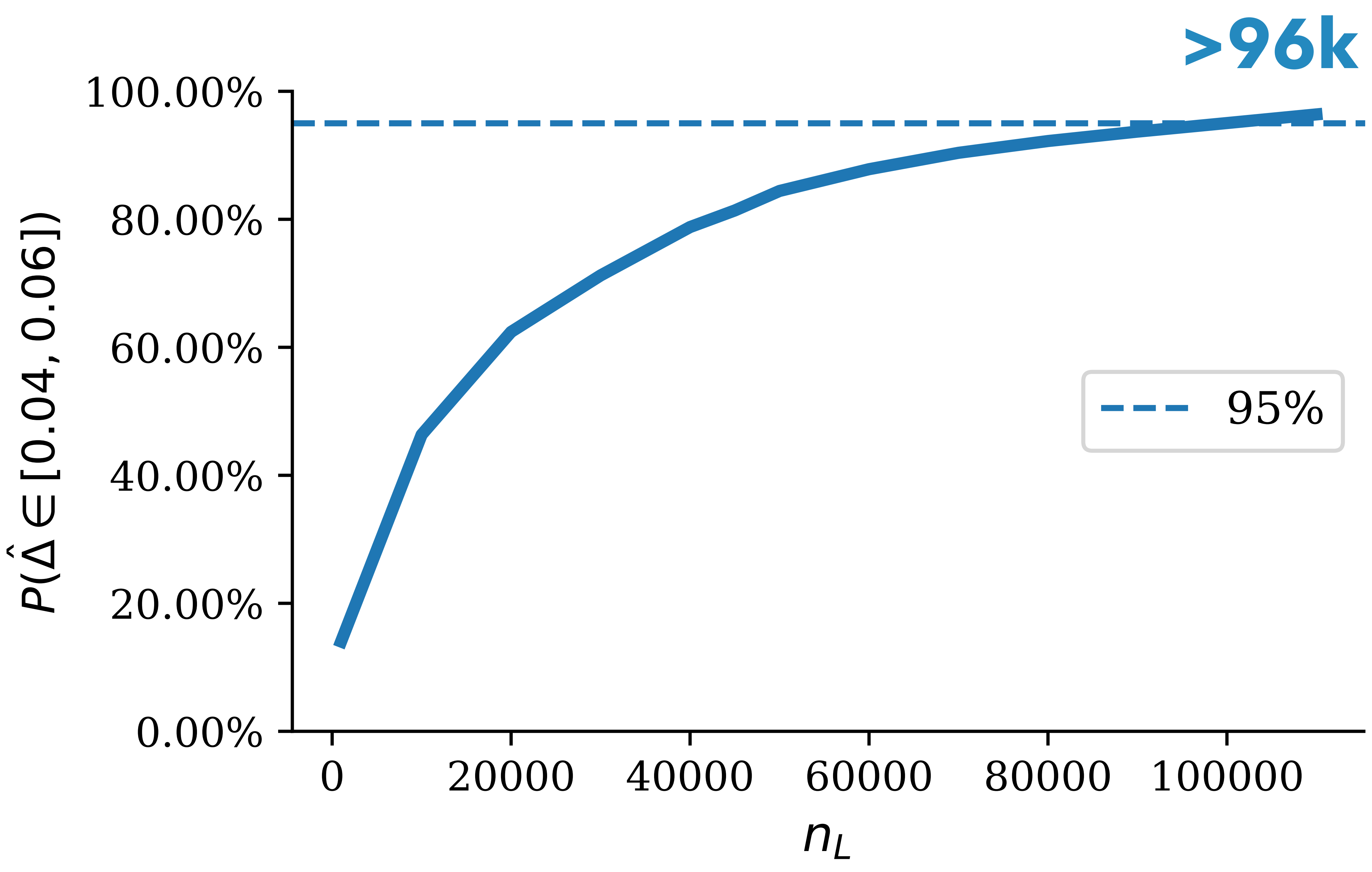
HOW MANY LABELED DATA DO I NEED TO COLLECT?

- Simulation:
  - p(g=0) = 20%
  - groupwise positive rates p(y = 1) are both 20%
  - the true groupwise TPRs are 95% and 90%.
- Compute frequentist estimation of ΔTPR for 10000 times



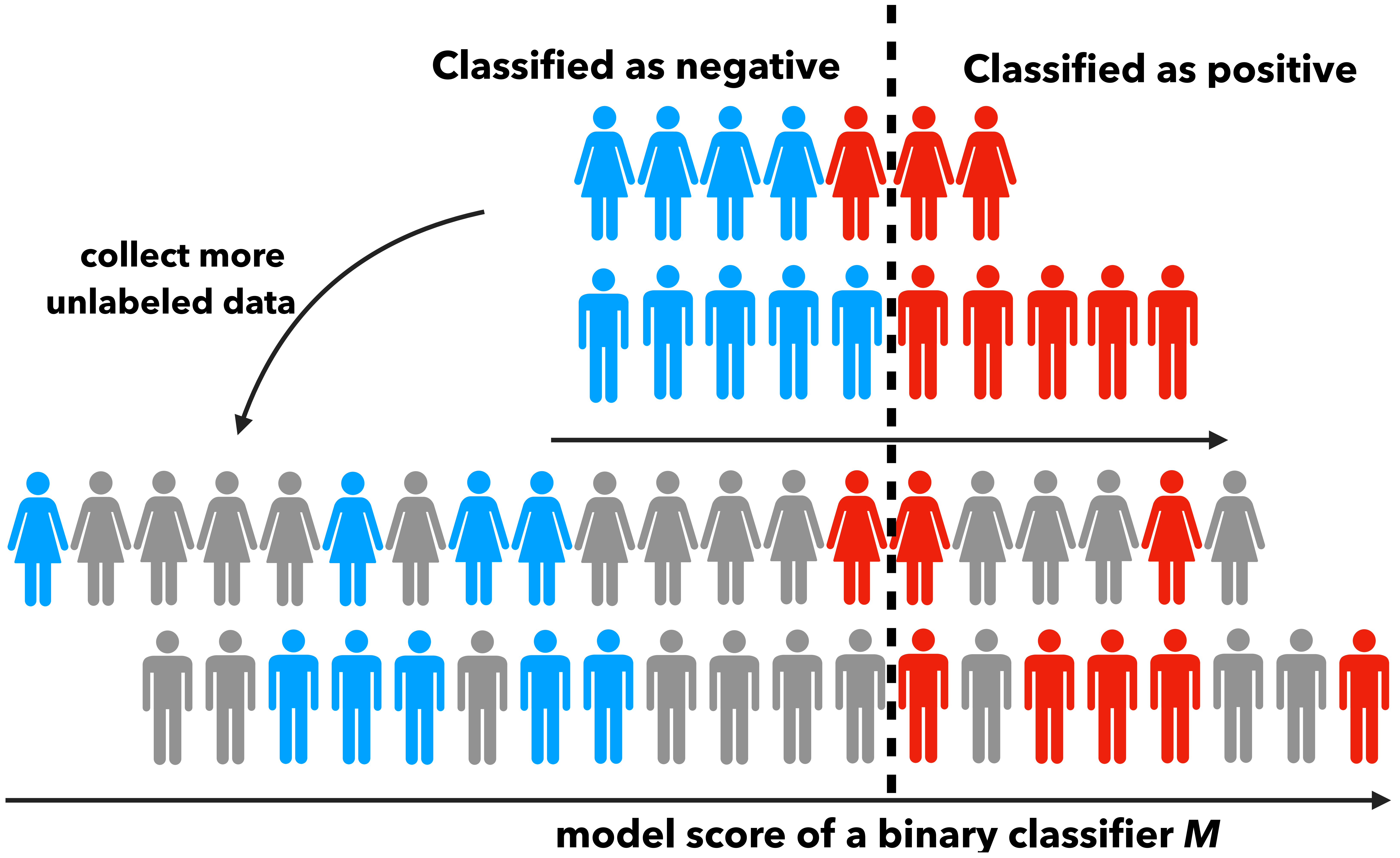
HOW MANY LABELED DATA DO I NEED TO COLLECT?

- Simulation:
  - p(g=0) = 20%
  - groupwise positive rates p(y = 1) are both 20%
  - the true groupwise TPRs are 95% and 90%.
- Compute frequentist estimation of ΔTPR for 10000 times

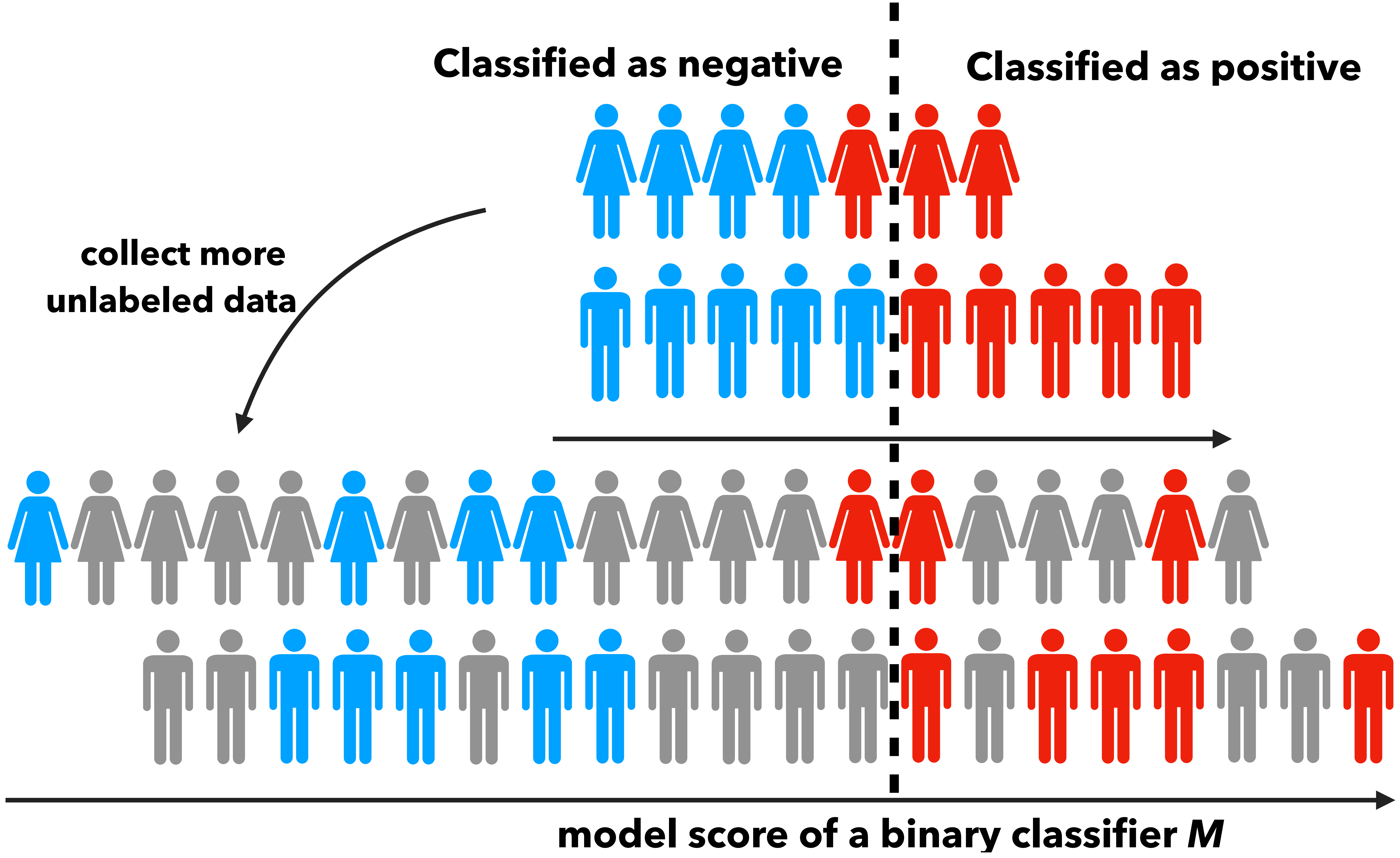


Dataset	Test Size	G	$p(g=0)$	$p(y=1)$
Adult	10054	gender, race	0.68, 0.86	0.25
Bank	13730	age	0.45	0.11
German	334	age, gender	0.79, 0.37	0.17
Compas-R	2056	gender, race	0.7, 0.85	0.69
Compas-VR	1337	gender, race	0.8, 0.34	0.47
Ricci	40	race	0.65	0.50

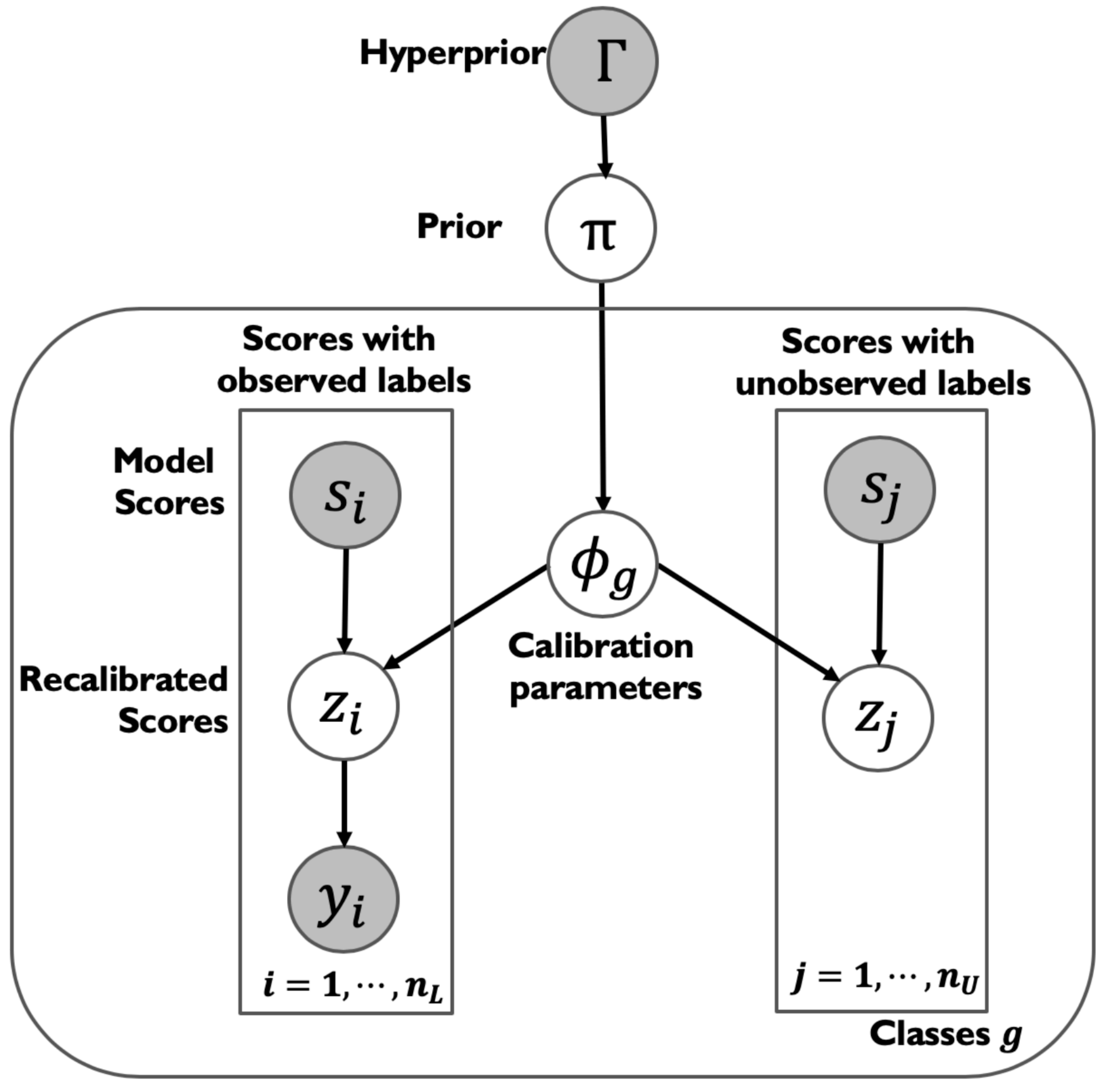
REDUCE UNCERTAINTY OF FAIRNESS WITH MORE UNLABELED DATA



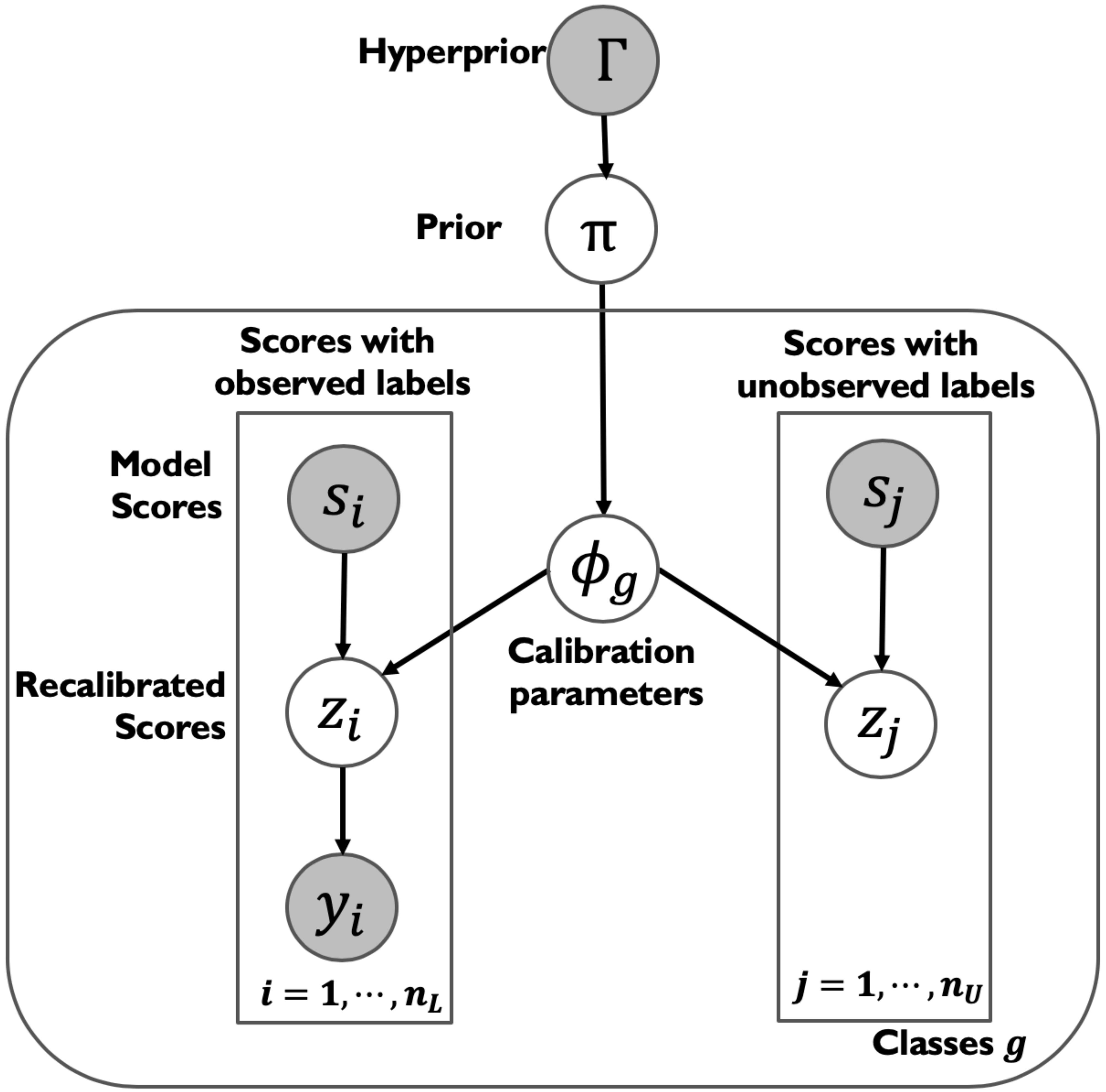
REDUCE UNCERTAINTY OF FAIRNESS WITH MORE UNLABELED DATA



Method: train a hierarchical Bayesian calibration model to predict the model performance on unlabeled data

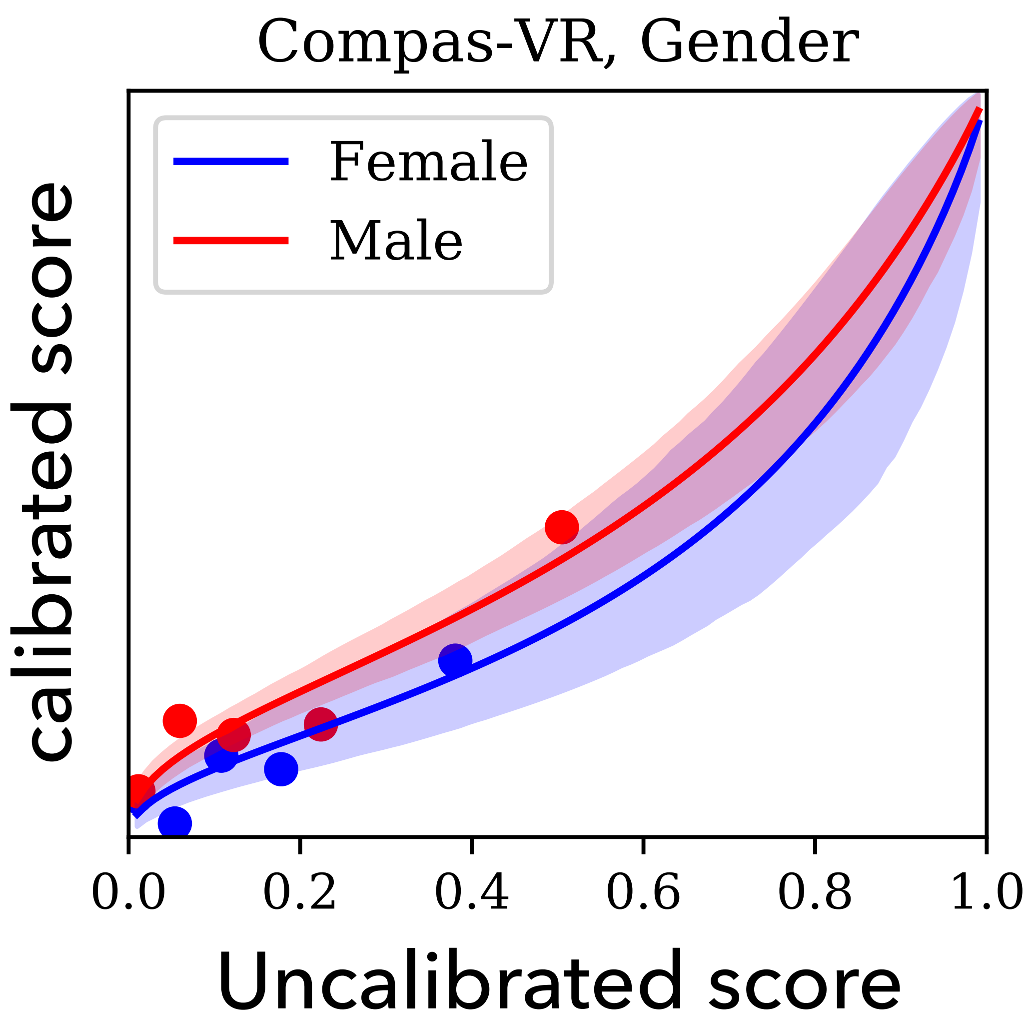


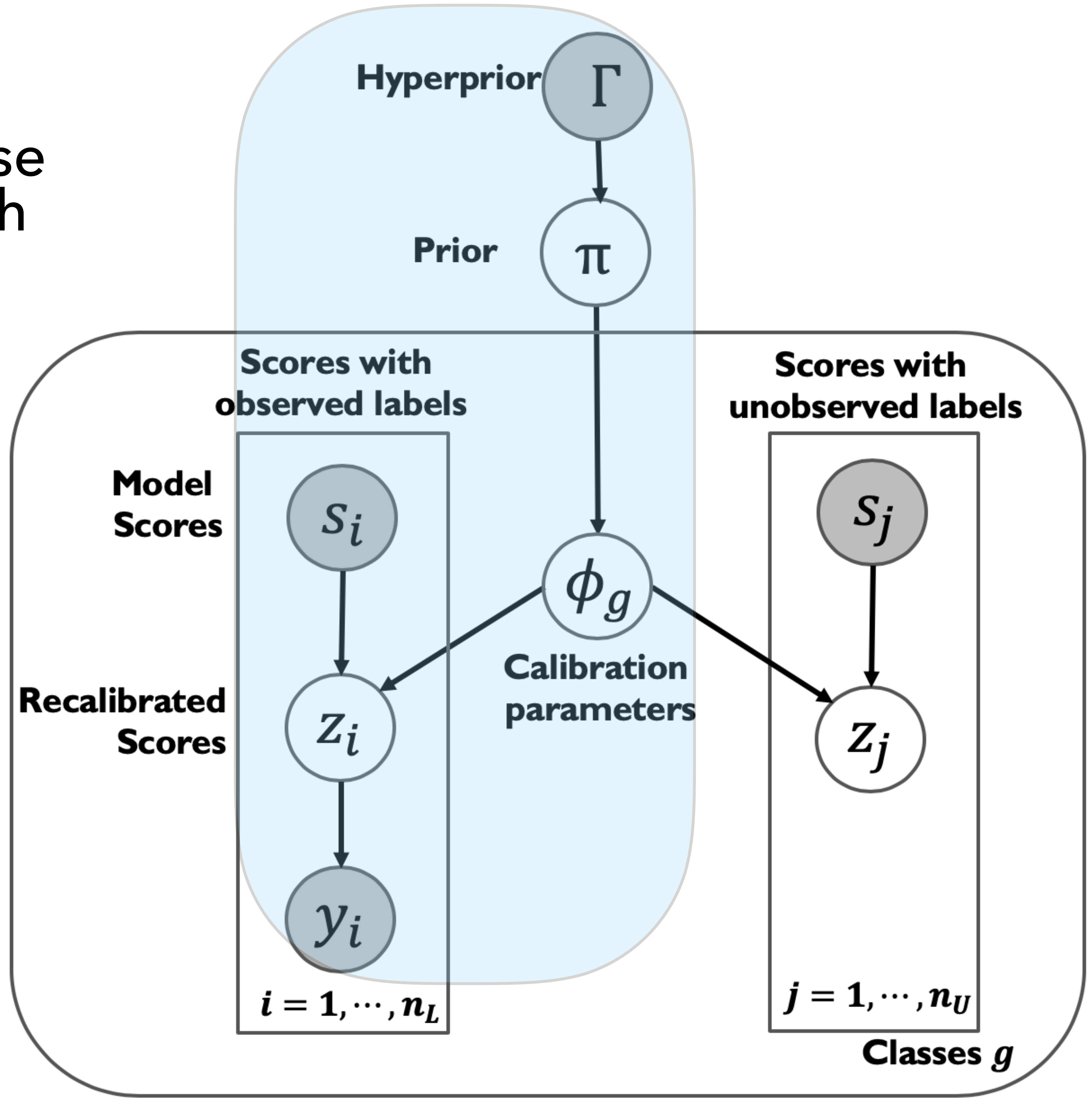
ASSESS FAIRNESS WITH BAYESIAN CALIBRATION



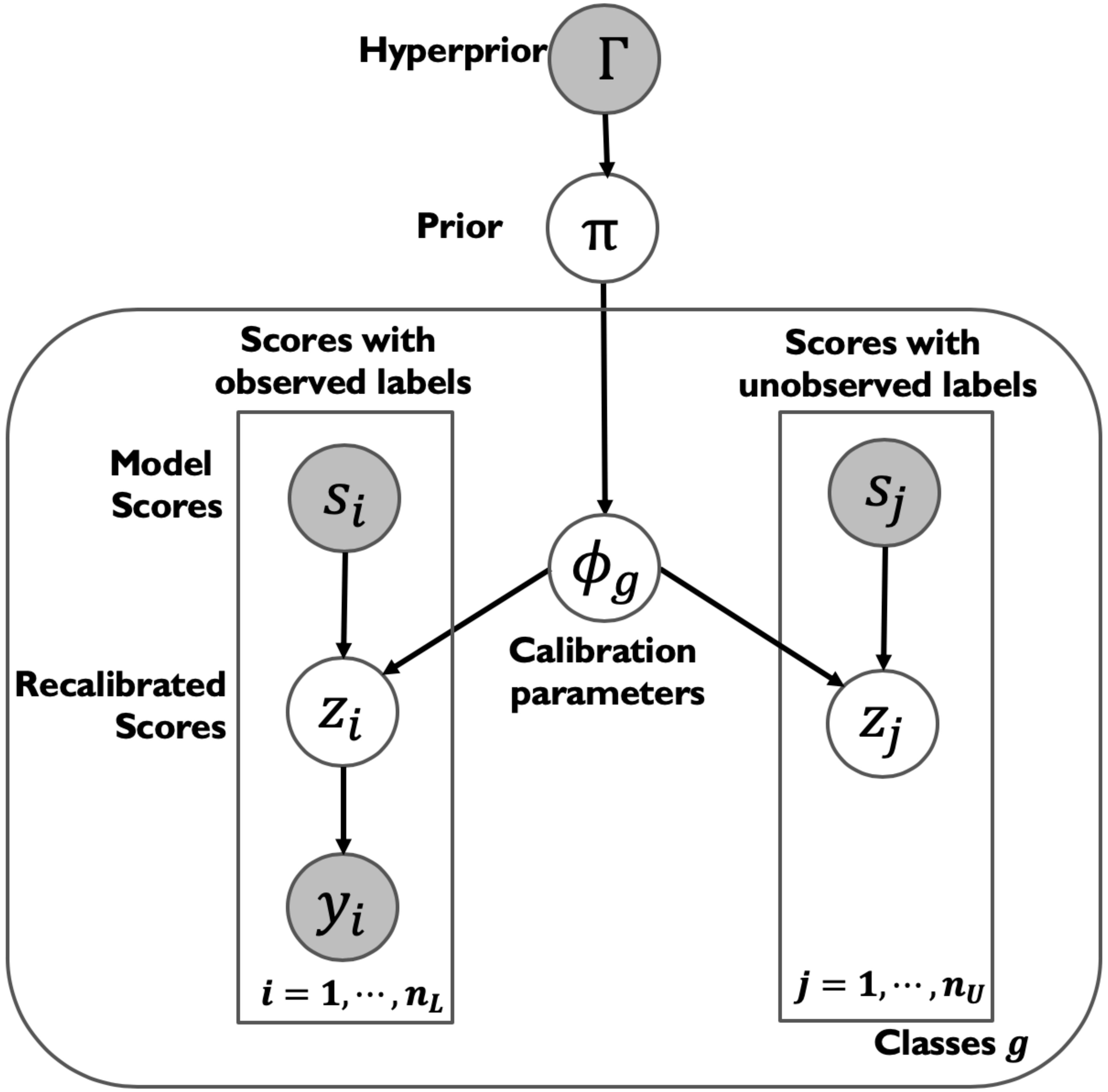
ASSESS FAIRNESS WITH BAYESIAN CALIBRATION

train: estimate groupwise calibration functions with parameters ϕ_g

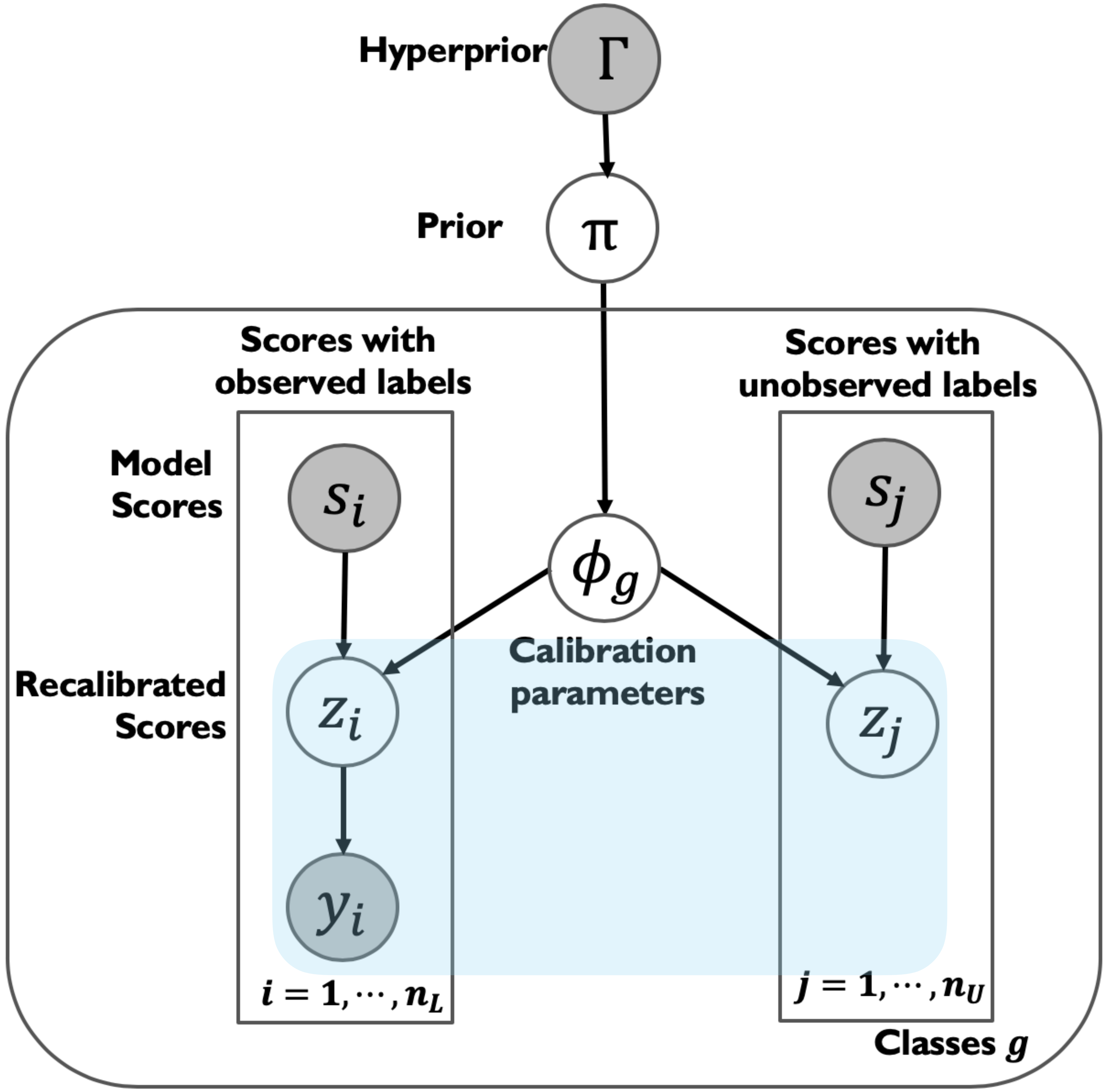




ASSESS FAIRNESS WITH BAYESIAN CALIBRATION



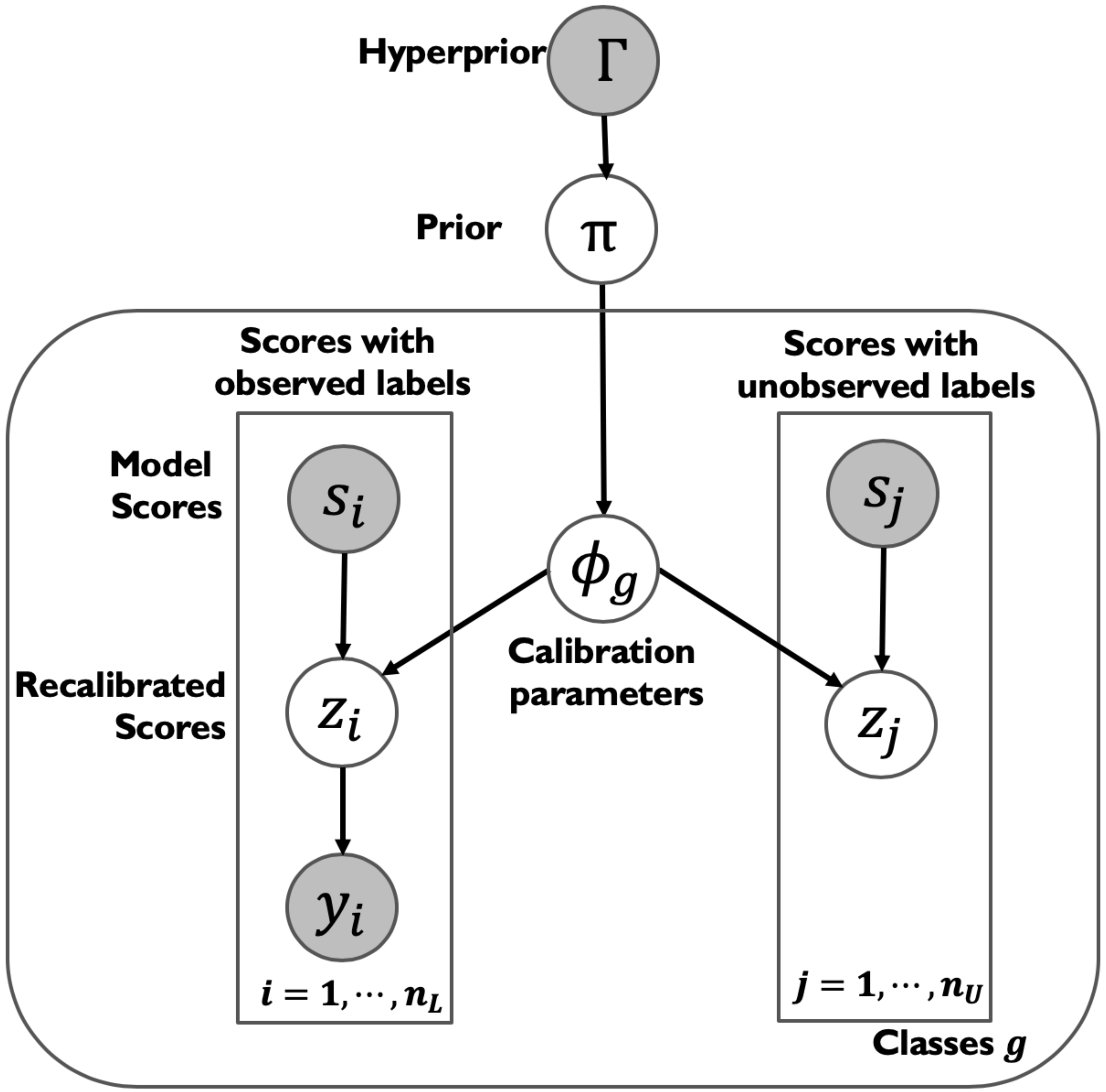
ASSESS FAIRNESS WITH BAYESIAN CALIBRATION



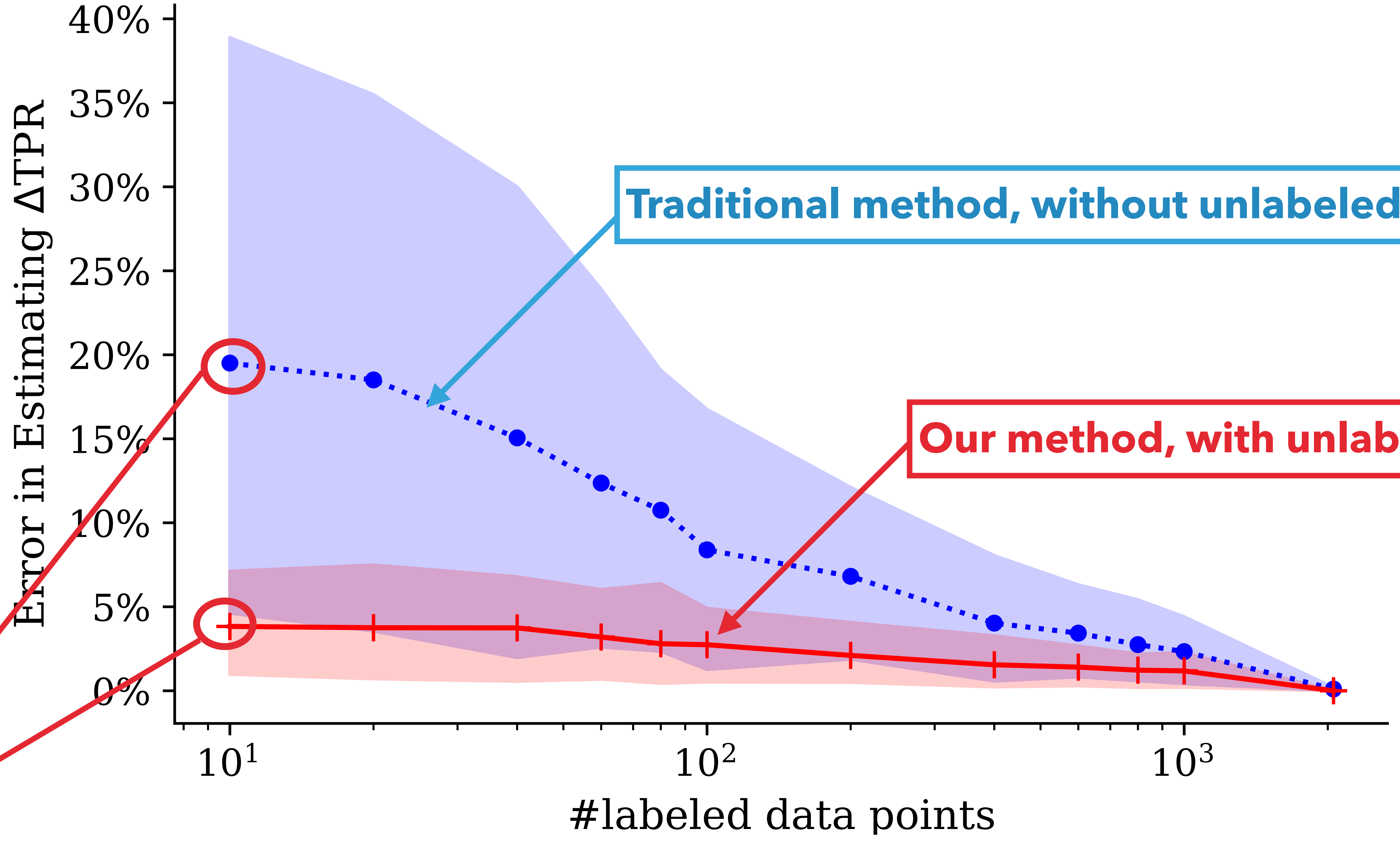
predict: generate estimates of the groupwise metrics θ_g and the difference in metrics Δ

ASSESS FAIRNESS WITH BAYESIAN CALIBRATION(BC)

- **#labeled data in some groups is small:** use Hierarchical Bayesian calibration to share statistical strength among groups
- **Variance of the estimates is high:** augment with unlabeled data by predicting labeling outcomes with BC
- **Calibration model**: any parametric calibration model, e.g. Beta calibration

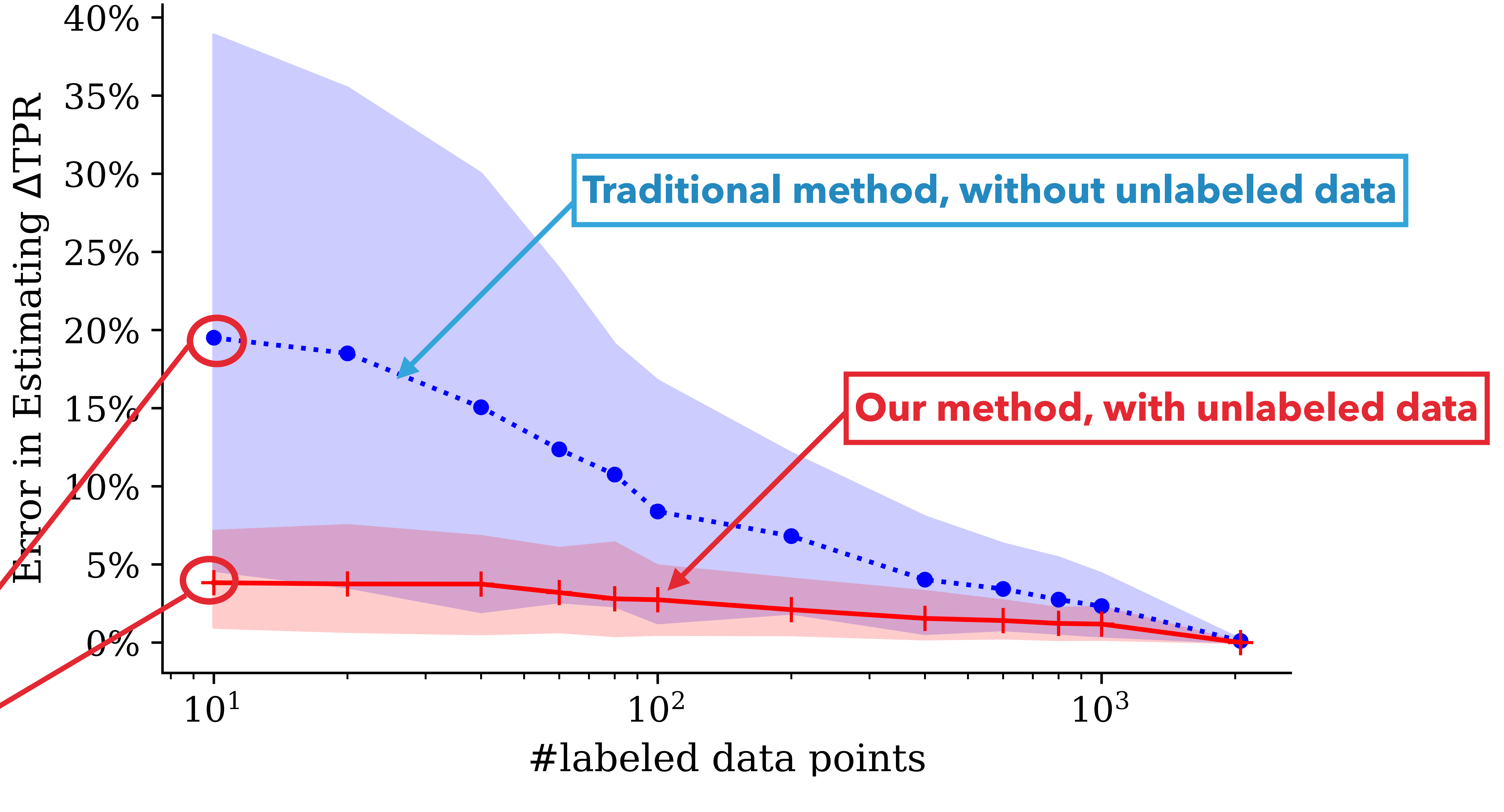


EXAMPLE: ASSESS DELTA TPR OF COMPAS RECIDIVISM



With **10** labeled data and ~**2000** unlabeled data, error in estimating TPR is **5%** for our method versus **20%** with only labeled data

EXAMPLE: ASSESS DELTA TPR OF COMPAS RECIDIVISM



With **10** labeled data and ~**2000** unlabeled data, error in estimating TPR is **5%** for our method versus **20%** with only labeled data

We obtained similar performance gain across multiple dataset-attribute combinations, prediction models, and fairness metrics

DISCUSSION

- bias-variance tradeoff
- potential **error in the calibration mapping** (e.g., due to misspecification of the parametric form of the calibration function) to **error in the estimate of Δ itself**

Lemma 4.5.1. *Given a prediction model* M *and score distribution* $P(s)$*, let* $f_g(s;\phi_g) : [0,1] \to [0,1]$ *denote the calibration model for group* g*; let* $f_g^*(s) : [0,1] \to [0,1]$ *be the optimal calibration function which maps* $s = P_M(\hat{y}=1|g)$ *to* $P(y=1|g)$*; and* Δ^* *is the true value of the metric. Then the absolute error of the expected estimate w.r.t.* ϕ *can be bounded as:* $|\mathbb{E}_\phi \Delta - \Delta^*| \leq \|\bar{f}_0 - f_0^*\|_1 + \|\bar{f}_1 - f_1^*\|_1$*, where* $\bar{f}_g(s) = \mathbb{E}_{\phi_g} f_g(s;\phi_g), \forall s \in [0,1]$*, and* $\|\cdot\|_1$ *is the expected* $L1$ *distance w.r.t.* $P(s|g)$*.*

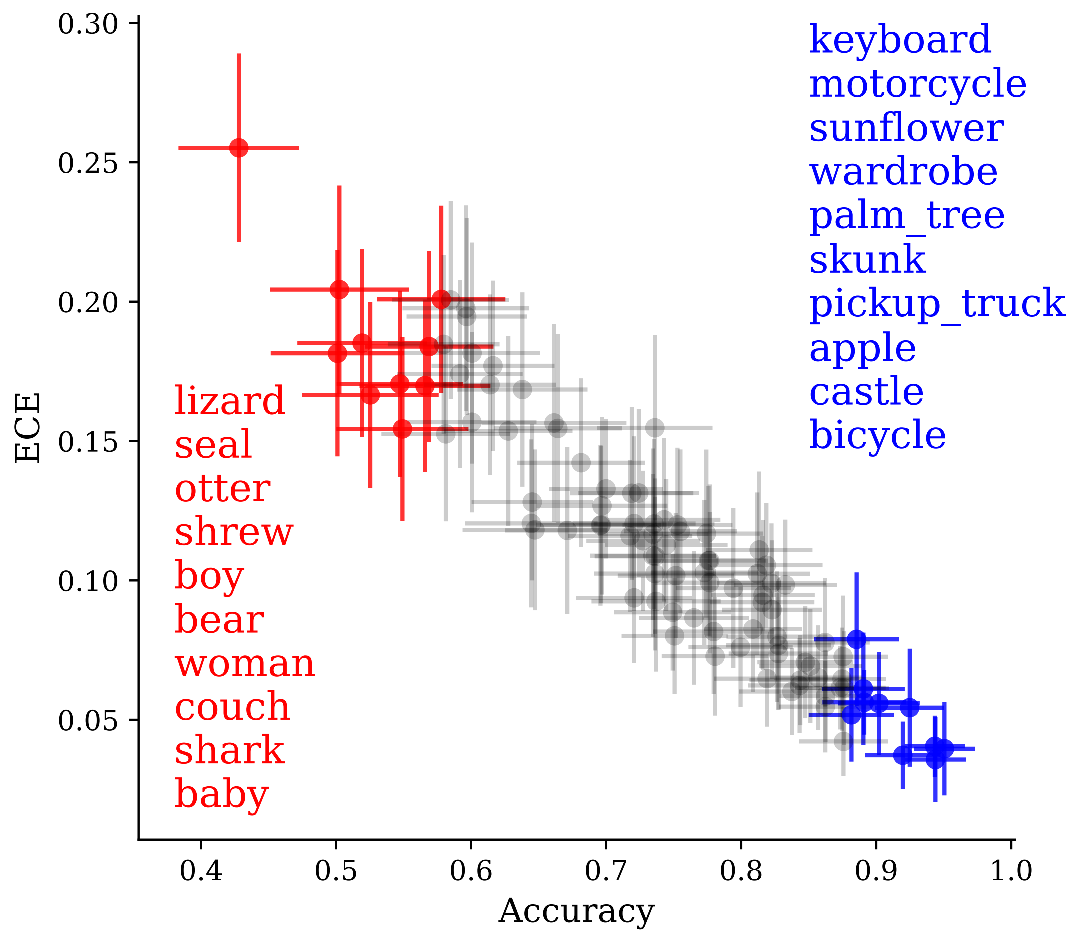
DISCUSSION

- **Calibration of the posterior probability**
  - a perfectly calibrated 95% credible interval would have 95% coverage.
  - generally not far from 95% there is room for improvement (model misspecification)
- **How about other calibration models?**
  - comparisons with an alternative calibration model, i.e. LLO calibration
  - two calibration methods tends to be very similar
- **Is the hierarchical structure necessary?**
  - ablation study by comparing with non-hierarchical Bayesian calibration
  - Hierarchical structure helps with avoiding occasional catastrophic errors
- **Sensitivity analysis for the calibration priors**
  - robust to the settings of prior variances

THESIS CONTRIBUTIONS

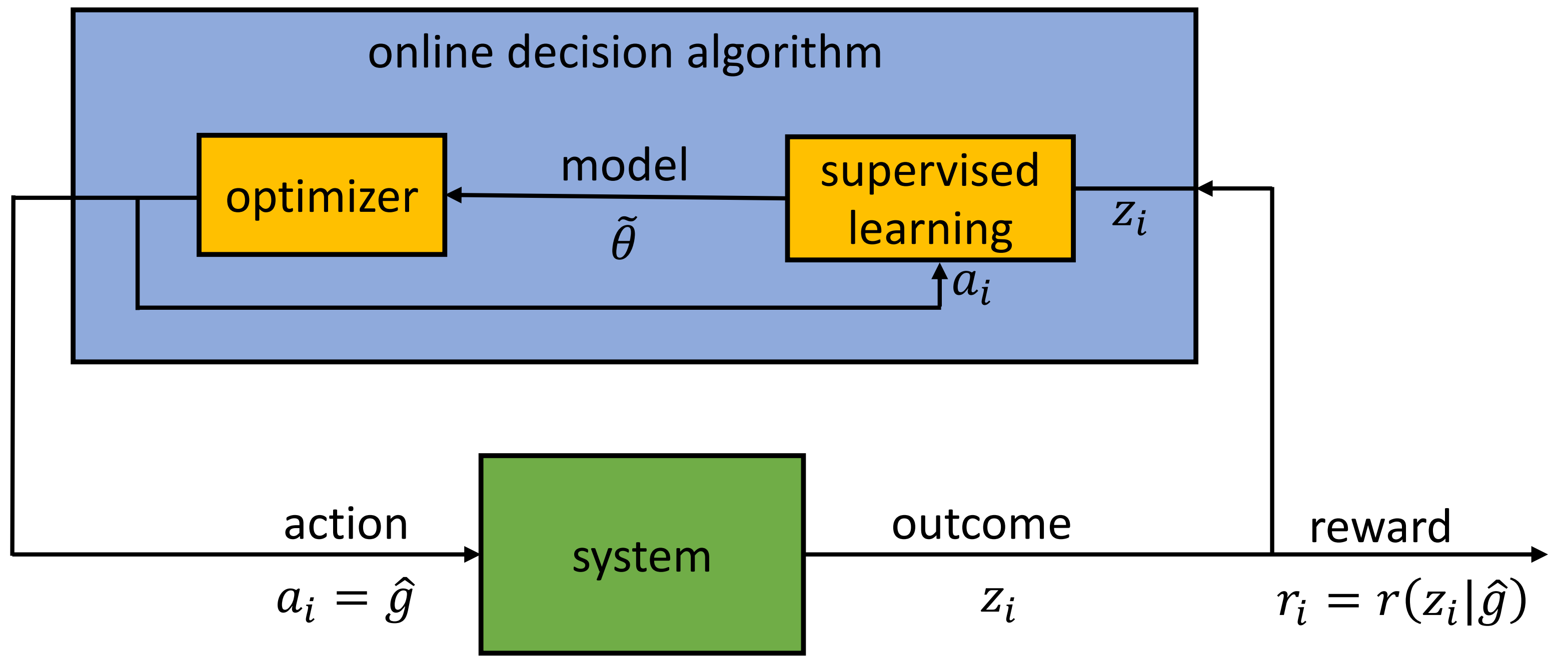
Bayesian assessment

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



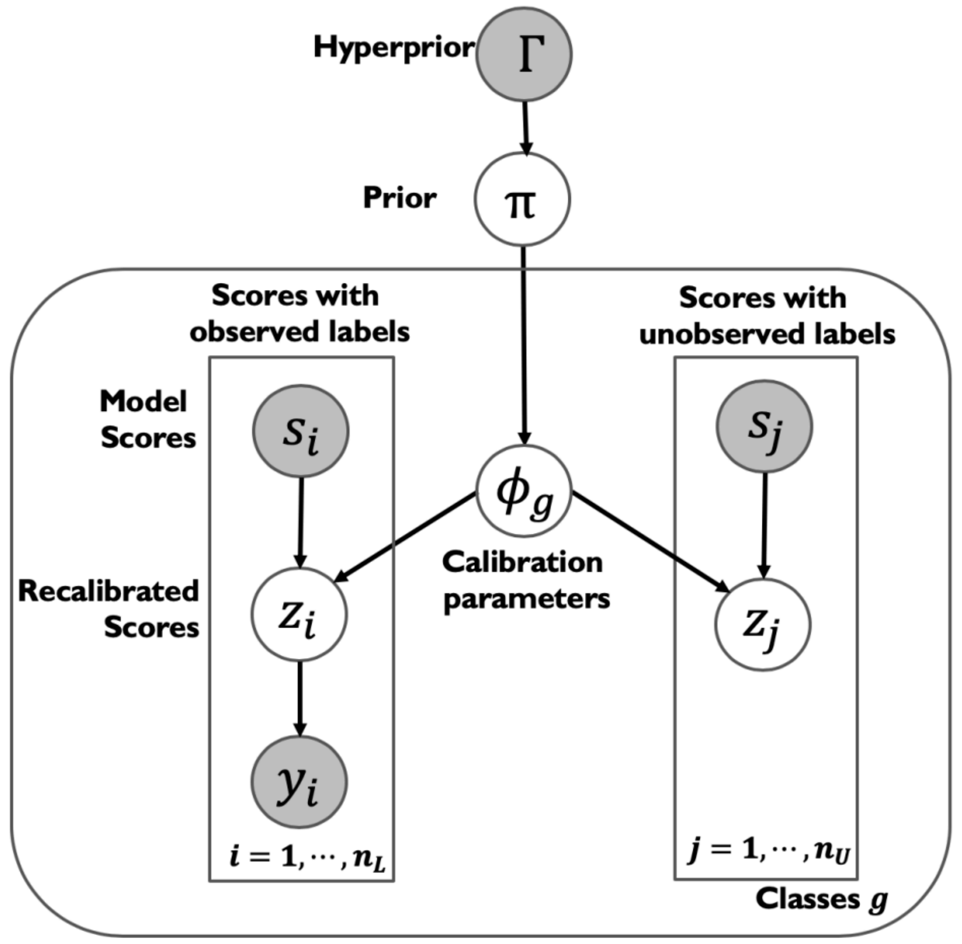
active Bayesian assessment

2. **Reduce uncertainty** of assessment, with **actively labeled data** selected from a pool of unlabeled data



assess with **unlabeled** data

3. **Reduce uncertainty** of assessment, by leveraging both **labeled and unlabeled data**

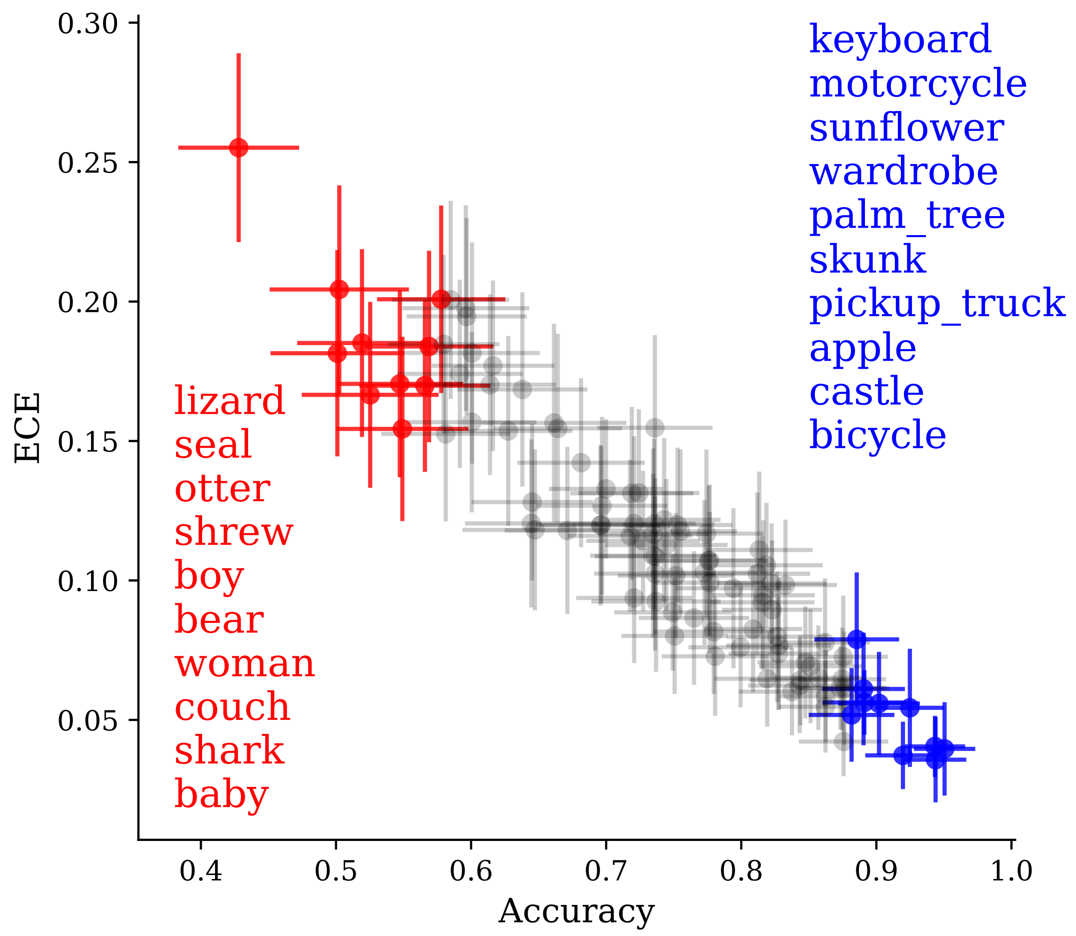


THESIS CONTRIBUTIONS

Bayesian estimation of performance metrics
(1) accuracy, reliability diagram, ECE
(2) Performance difference
(3) Confusion matrix, misclassification cost
Use self-assessment as informative priors

Bayesian assessment

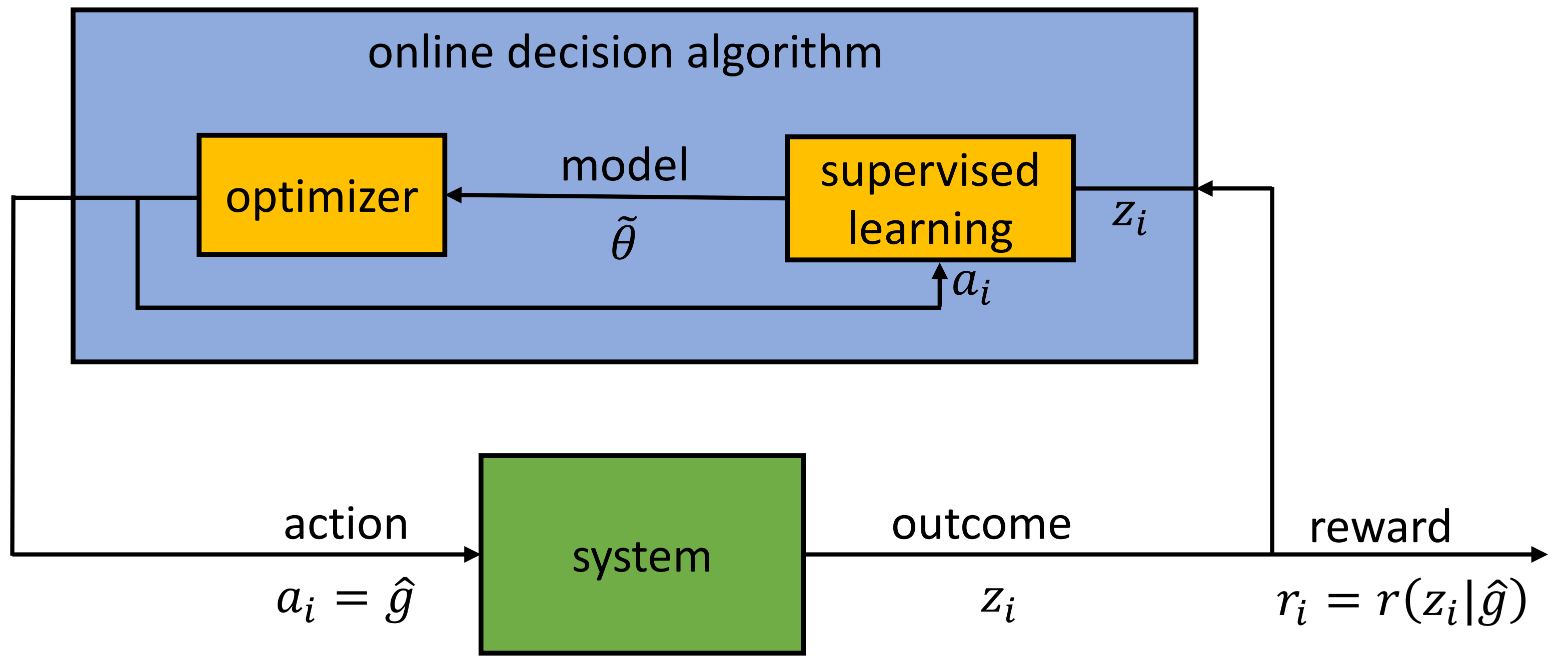
1. **Quantify uncertainty** of assessment with Bayesian models, with a set of **labeled data**



[Ji, Logan, Smyth, Steyvers 2019 ICML UDL]

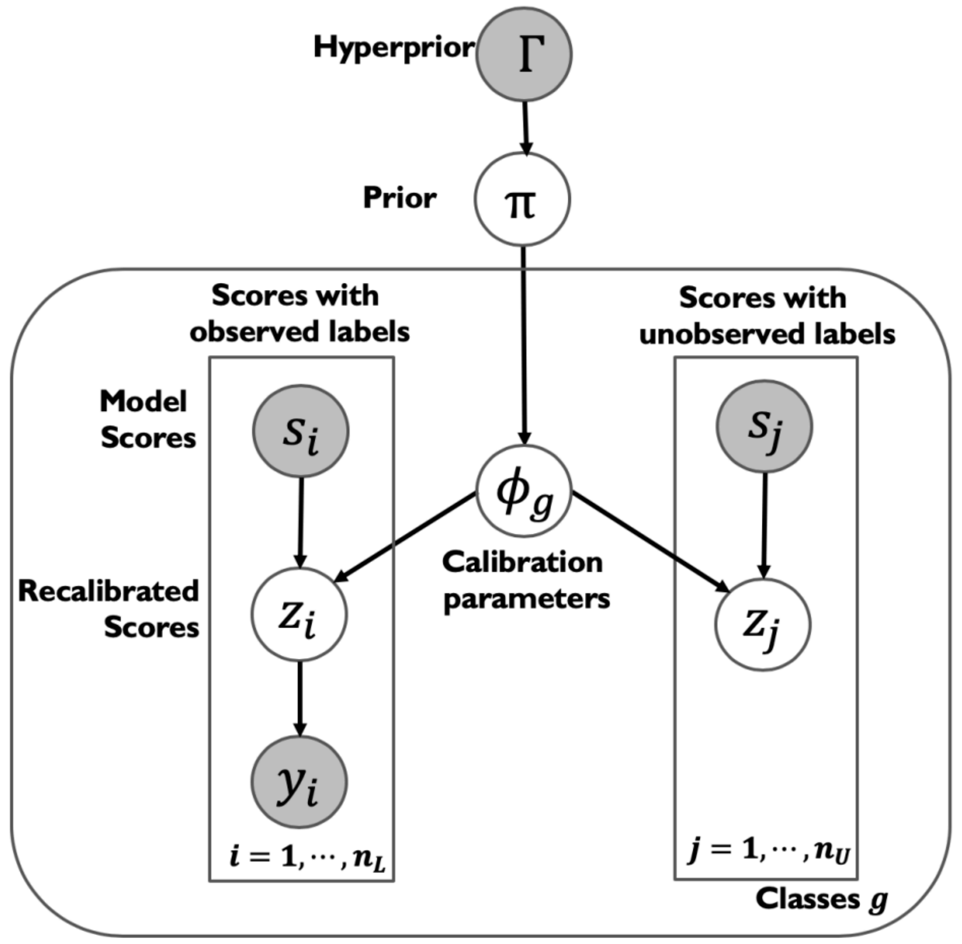
active Bayesian assessment

2. **Reduce uncertainty** of assessment, with **actively labeled data** selected from a pool of unlabeled data



assess with **unlabeled** data

3. **Reduce uncertainty** of assessment, by leveraging both **labeled and unlabeled data**



UCIRVINE

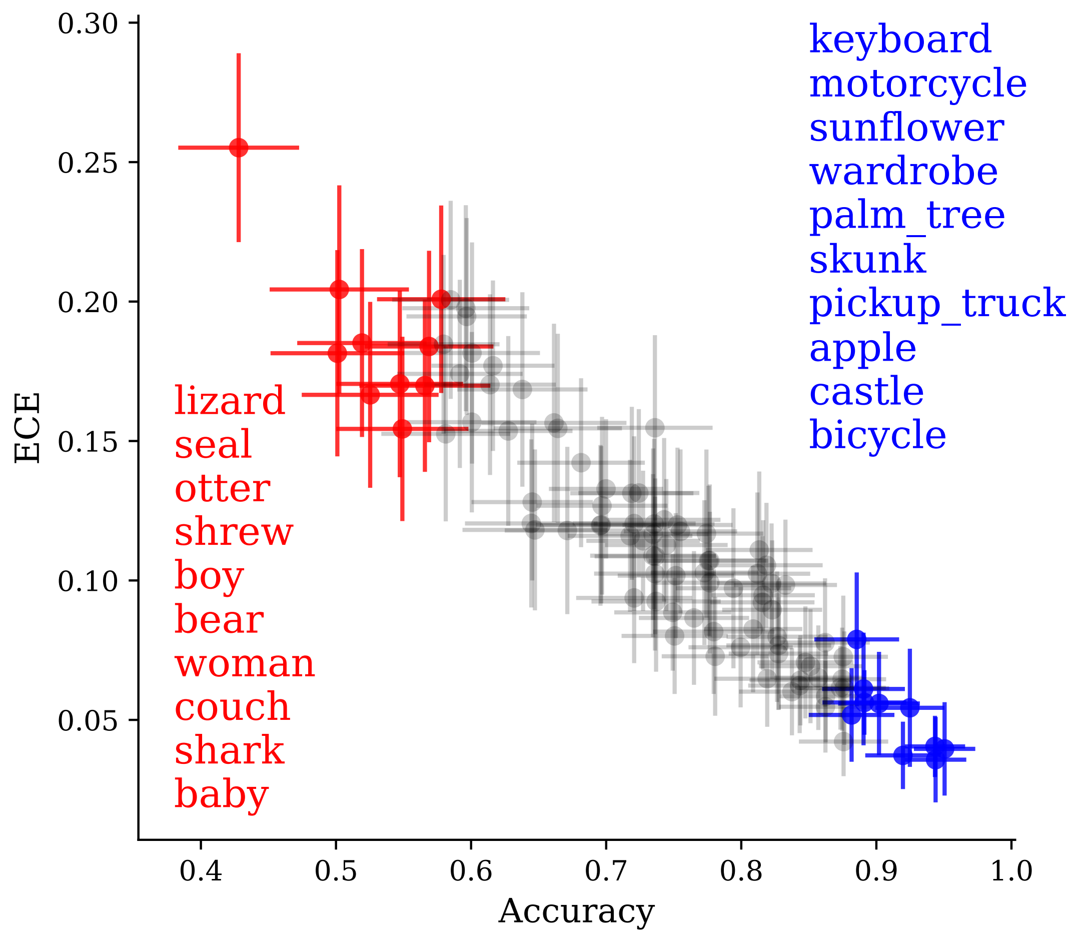
THESIS CONTRIBUTIONS

Bayesian estimation of performance metrics
(1) accuracy, reliability diagram, ECE
(2) Performance difference
(3) Confusion matrix, misclassification cost
Use self-assessment as informative priors

Developed active assessment framework for
(1) estimation of model performance;
(2) identification of model deficiencies;
(3) performance comparison between groups
Developed a set of Thompson sampling algorithms

Bayesian assessment

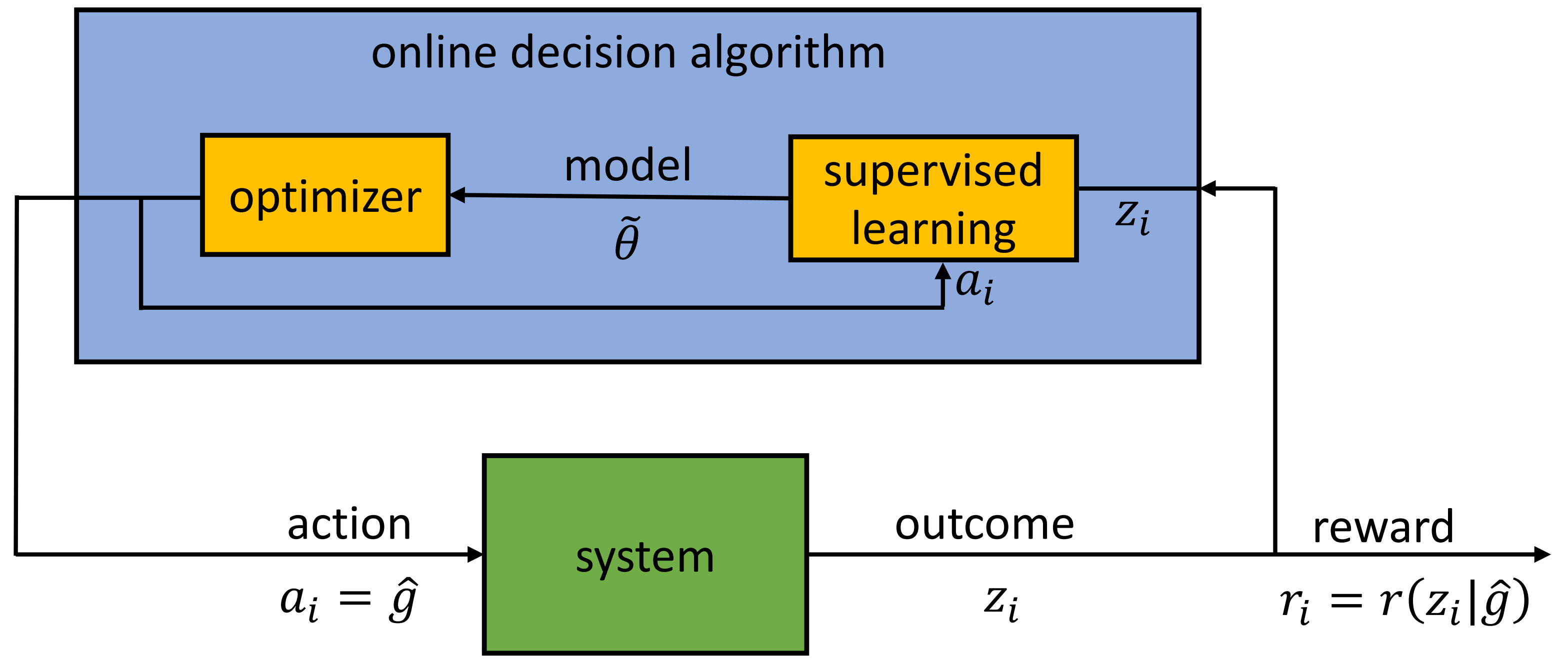
1. **Quantify uncertainty** of assessment with Bayesian models, with a set of **labeled data**



[Ji, Logan, Smyth, Steyvers 2019 ICML UDL]

active Bayesian assessment

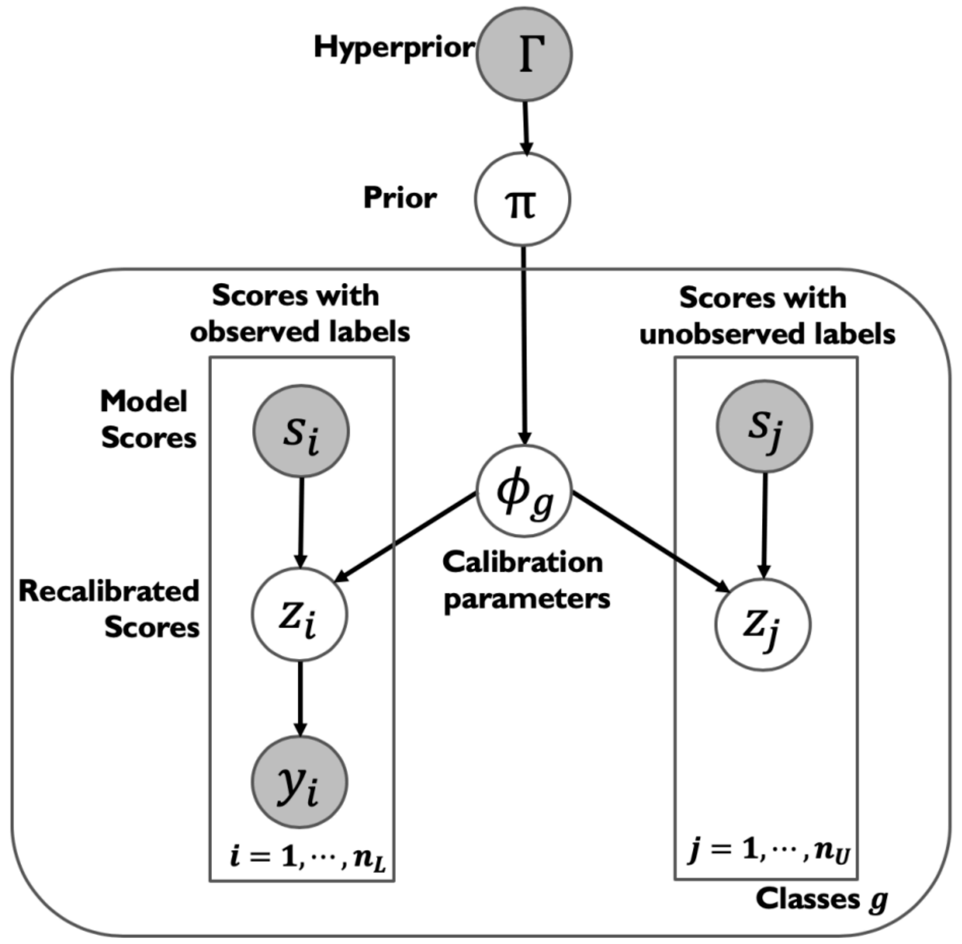
2. **Reduce uncertainty** of assessment, with **actively labeled data** selected from a pool of unlabeled data



[Ji, Logan, Smyth, Steyvers 2021 AAAI?]

assess with **unlabeled** data

3. **Reduce uncertainty** of assessment, by leveraging both **labeled and unlabeled data**



THESIS CONTRIBUTIONS

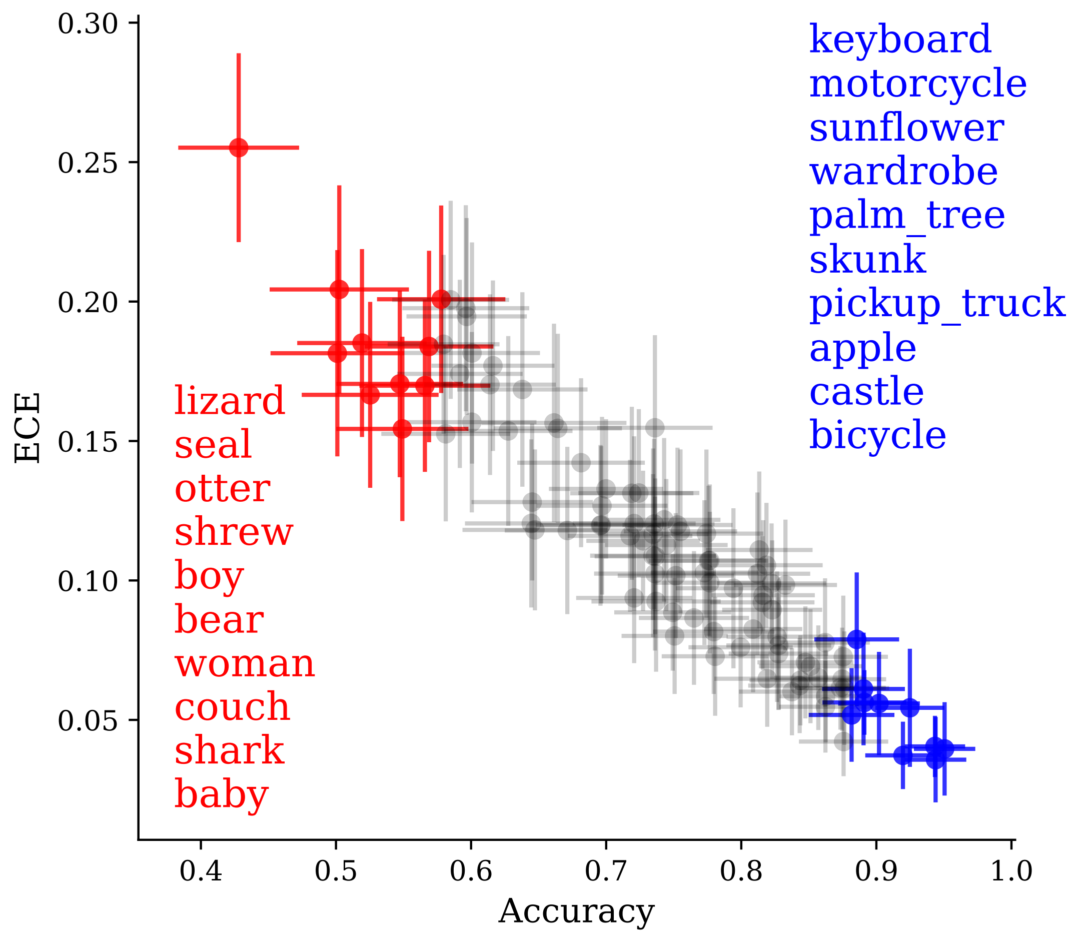
Bayesian estimation of performance metrics
(1) accuracy, reliability diagram, ECE
(2) Performance difference
(3) Confusion matrix, misclassification cost
Use self-assessment as informative priors

Developed active assessment framework for
(1) estimation of model performance;
(2) identification of model deficiencies;
(3) performance comparison between groups
Developed a set of Thompson sampling algorithms

(1) Proposed a comprehensive Bayesian treatment of fairness assessment
(2) Developed a new hierarchical Bayesian model to leverage information from both unlabeled and labeled examples

Bayesian assessment

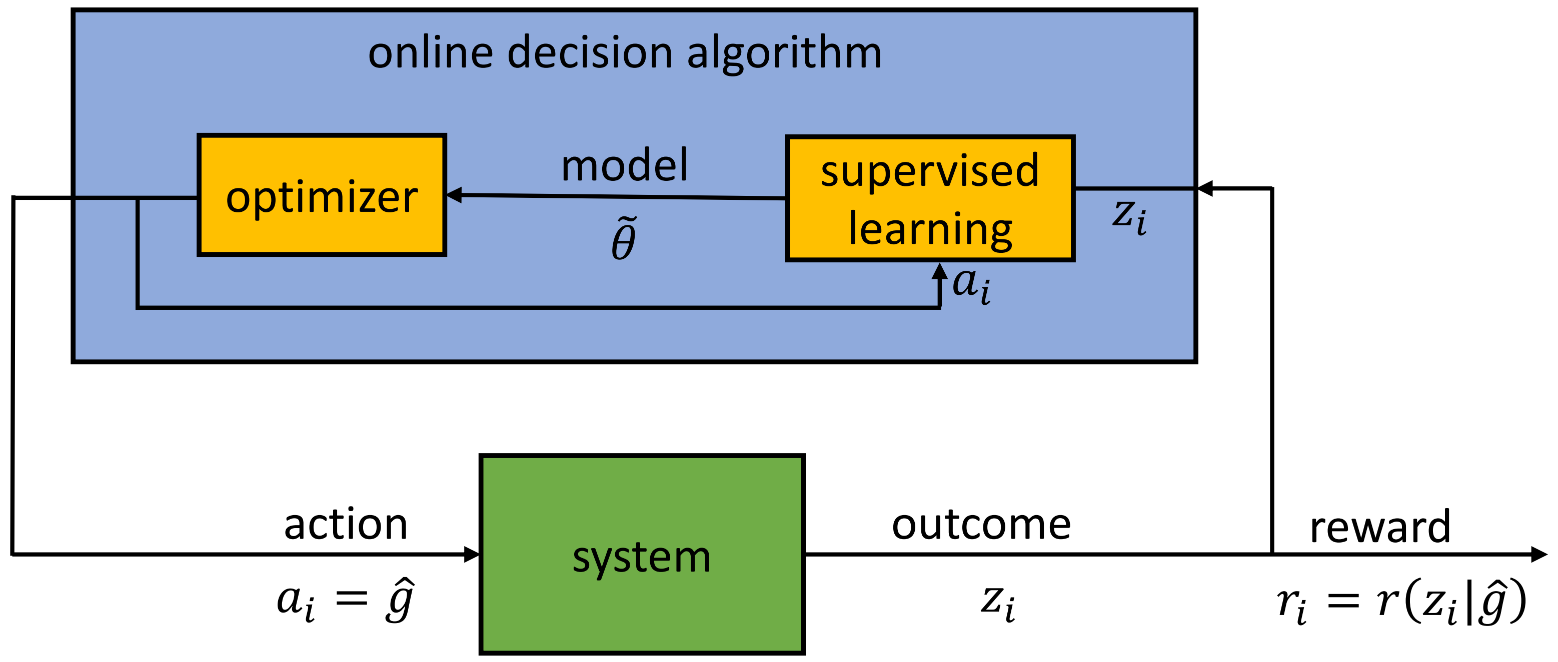
1. **Quantify uncertainty** of assessment with Bayesian models, with a set of **labeled data**



[Ji, Logan, Smyth, Steyvers 2019 ICML UDL]

active Bayesian assessment

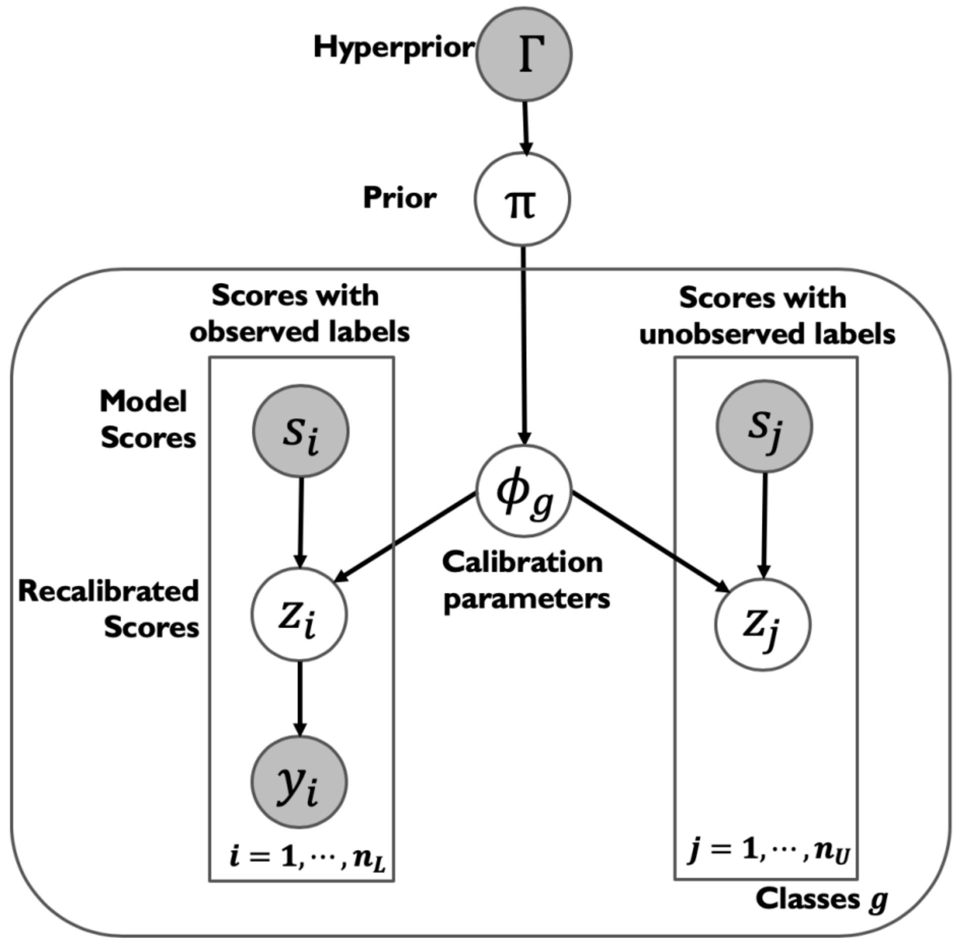
2. **Reduce uncertainty** of assessment, with **actively labeled data** selected from a pool of unlabeled data



[Ji, Logan, Smyth, Steyvers 2021 AAAI?]

assess with **unlabeled** data

3. **Reduce uncertainty** of assessment, by leveraging both **labeled and unlabeled data**



[Ji, Smyth, Steyvers 2020 NeurIPS]

LIST OF PUBLICATIONS

- Bayesian Evaluation of Black-Box Classifiers. [Ji, Logan, Smyth, Steyvers ICML UDL 2019]
- Can I Trust My Fairness Metric? Assessing Fairness with Unlabeled Data and Bayesian Inference. [Ji, Smyth, Steyvers NeurIPS 2020]
- Active Bayesian Assessment for Black-Box Classifiers. [Ji, Logan, Smyth, Steyvers AAAI 2021?]

LIST OF PUBLICATIONS

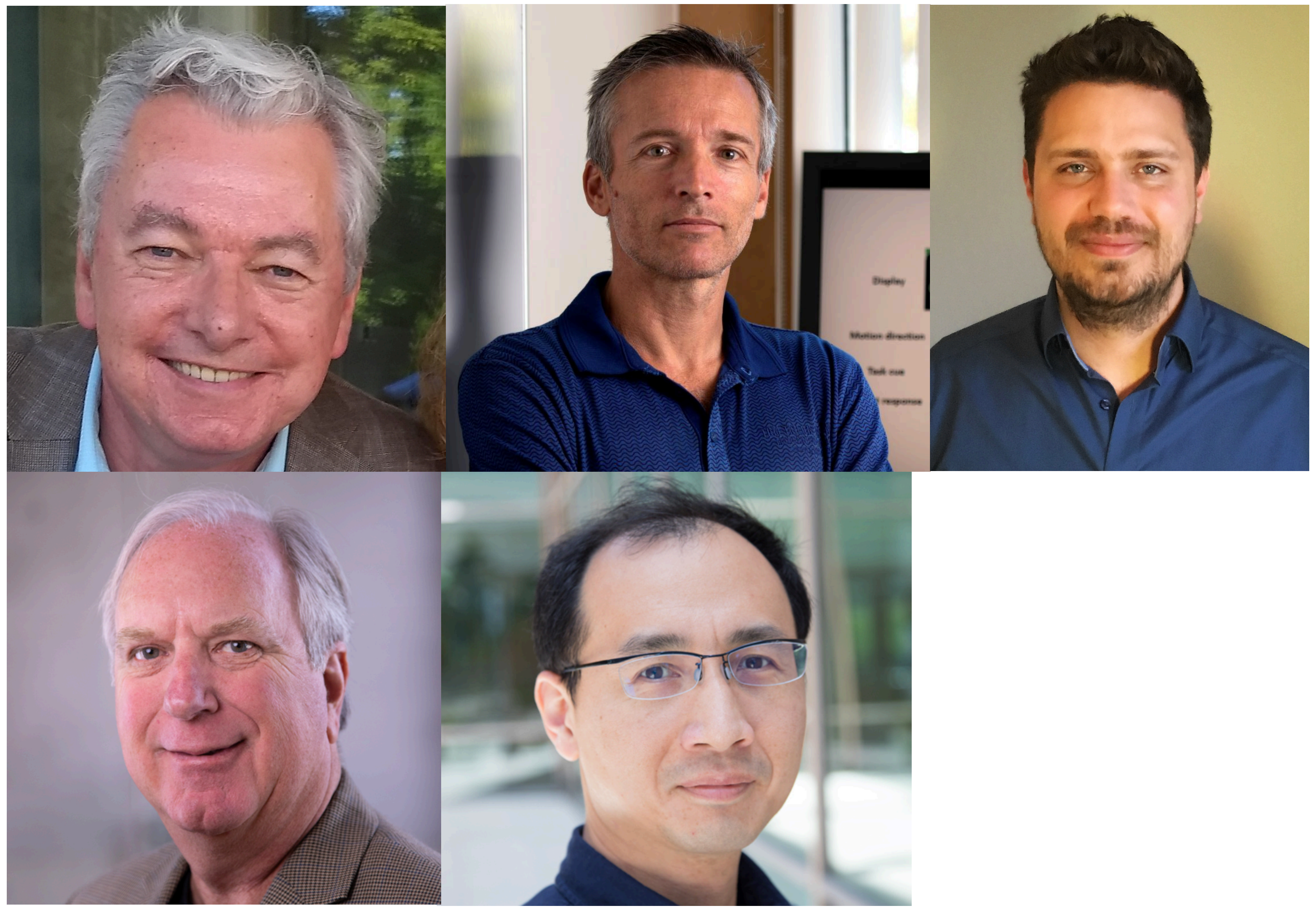
- Bayesian Evaluation of Black-Box Classifiers. [Ji, Logan, Smyth, Steyvers ICML UDL 2019]
- Can I Trust My Fairness Metric? Assessing Fairness with Unlabeled Data and Bayesian Inference. [Ji, Smyth, Steyvers NeurIPS 2020]
- Active Bayesian Assessment for Black-Box Classifiers. [Ji, Logan, Smyth, Steyvers AAAI 2021?]

Automated diagnosis of Leukemia with cytometry data analysis

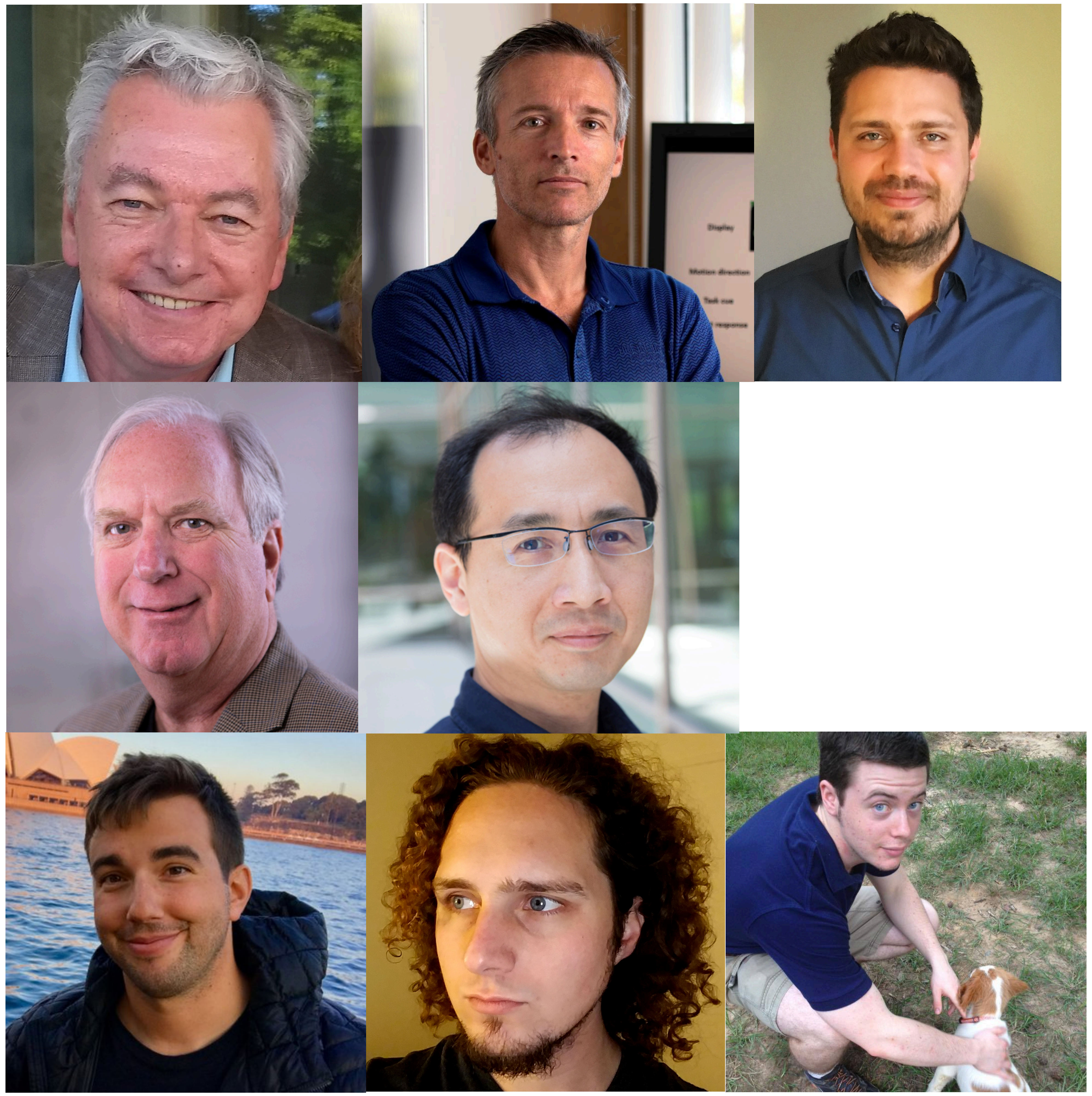
- Mondrian Processes for Flow Cytometry Analysis. [Ji, Nalisnick, Smyth NeurIPS ML4H 2017]
- Bayesian Trees for Automated Cytometry Data Analysis. [Ji, Nalisnick, Qian, Scheuermann, Smyth MLHC 2018]
- Learning Discriminative Gating Representations for Cytometry Data. [Ji, Putzel, Qian, Scheuermann, Bui, Wang, Smyth ICML Workshop on Computational Biology 2019]
- Optimization of Automated Gating for Clinical Diagnosis using Discriminative Gates. [Ji, Putzel, Qian, Scheuermann, Bui, Wang, Smyth Cytometry: Part A 2019]

ACKNOWLEDGEMENTS

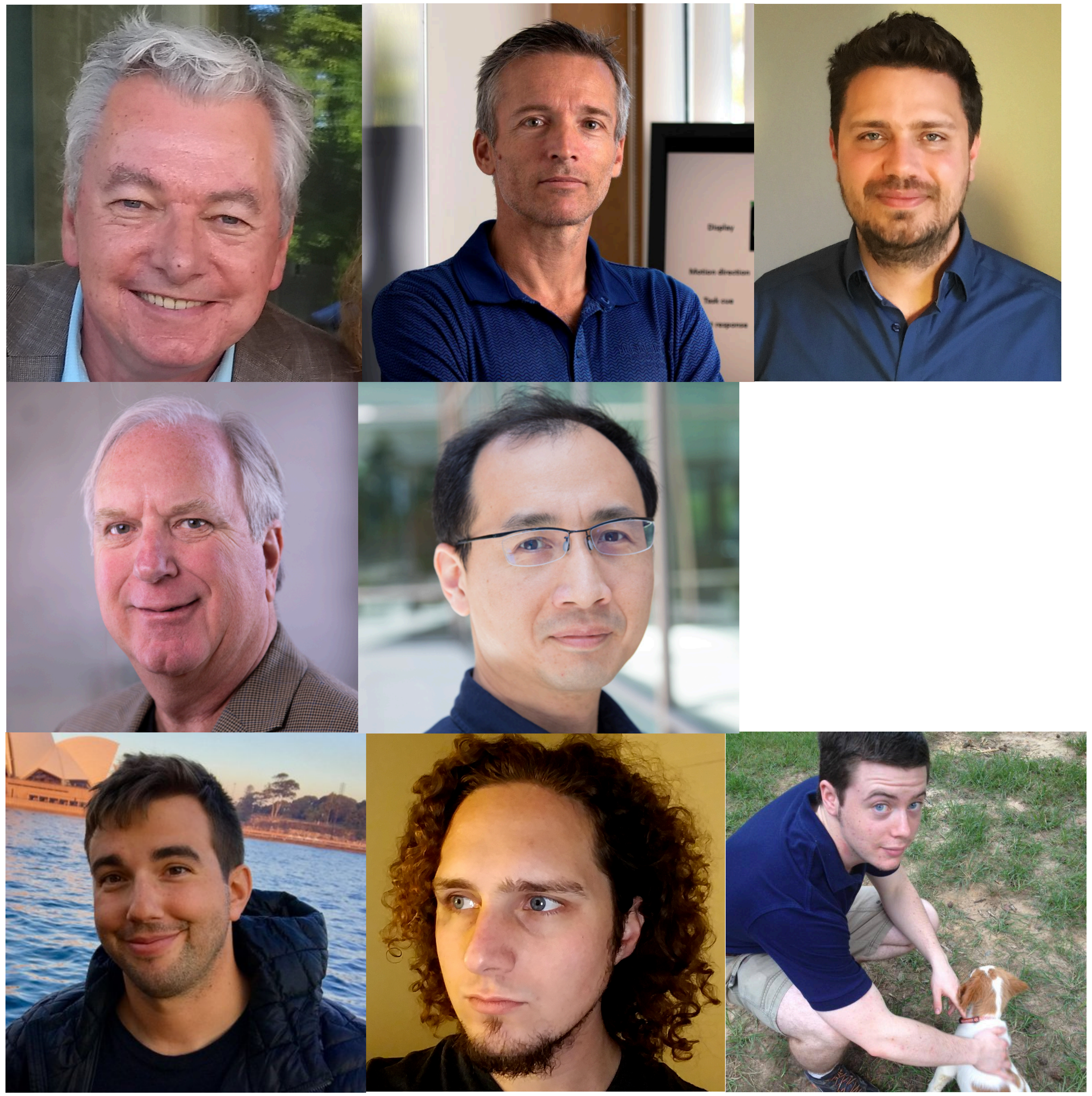
ACKNOWLEDGEMENTS



ACKNOWLEDGEMENTS



ACKNOWLEDGEMENTS



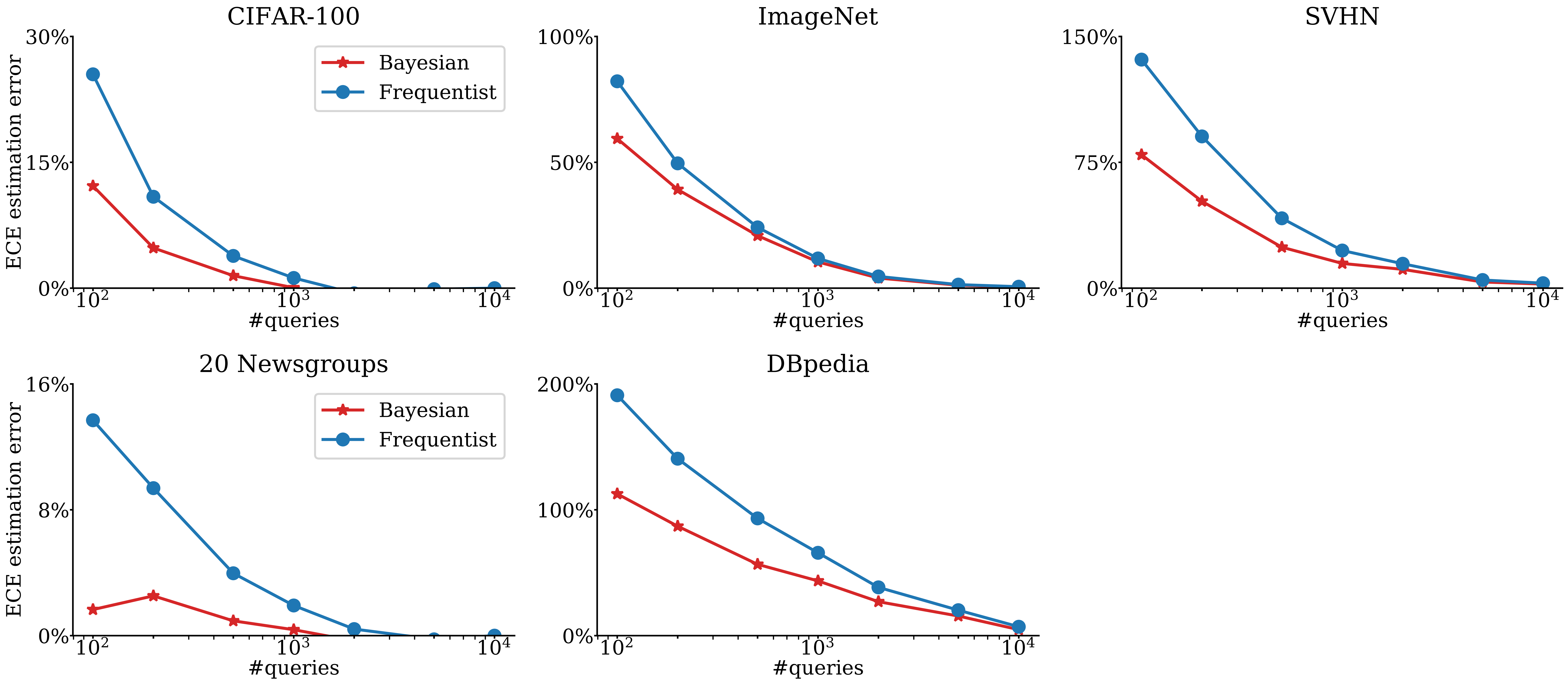
ACKNOWLEDGEMENTS

ACKNOWLEDGEMENTS

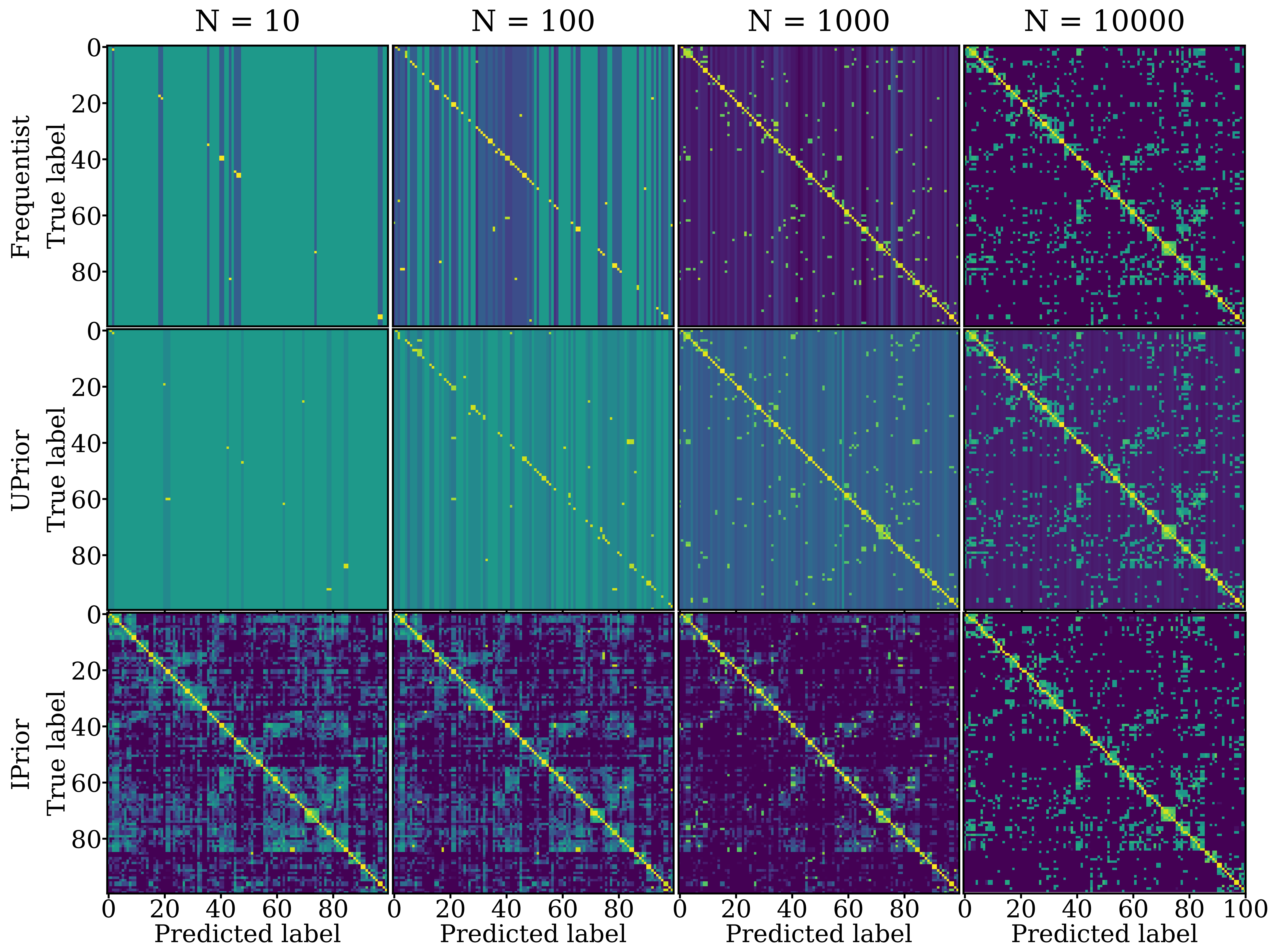
UCIRVINE

ACKNOWLEDGEMENTS

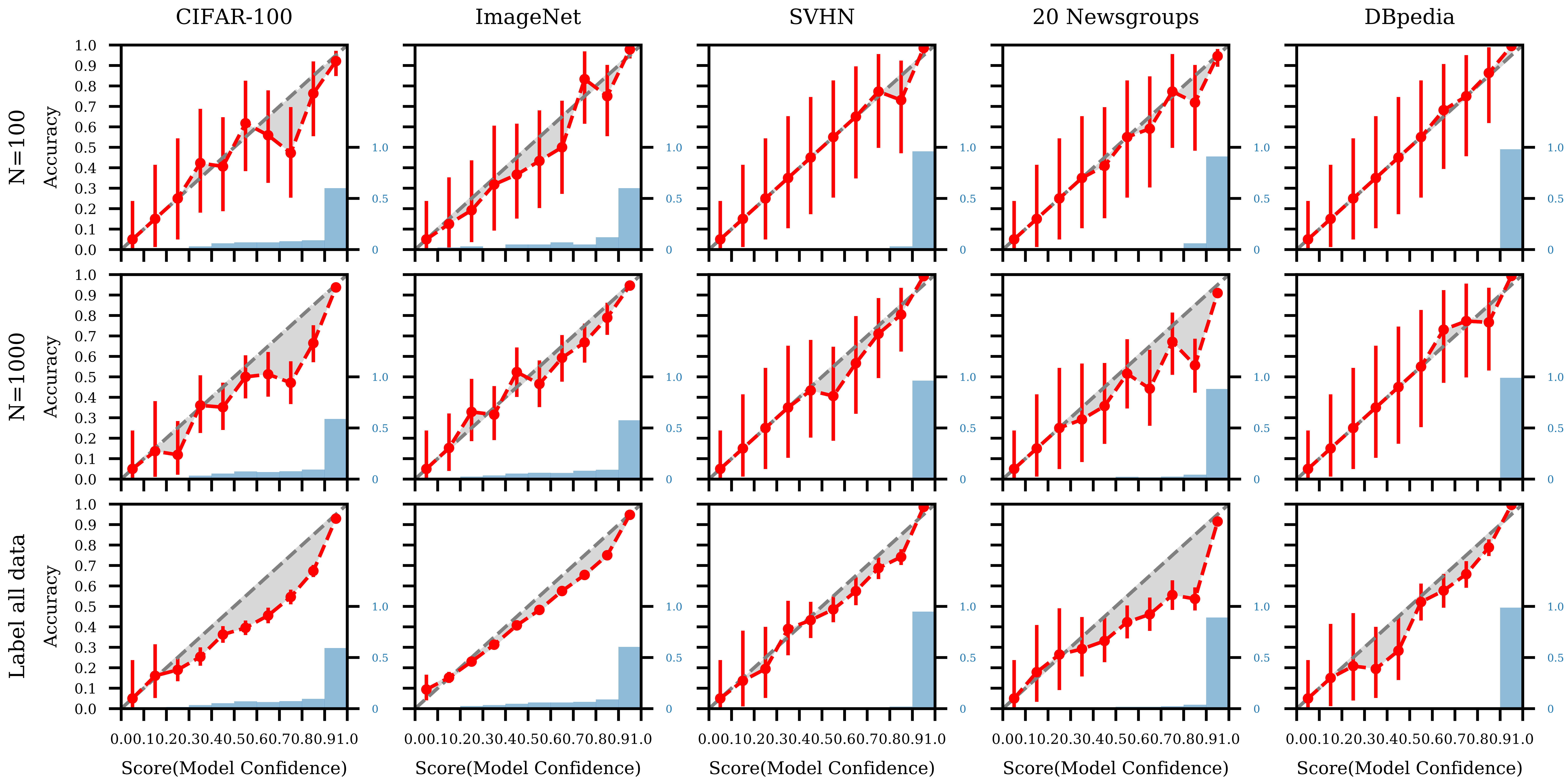
EXPERIMENTS: BAYESIAN ESTIMATION OF ECE



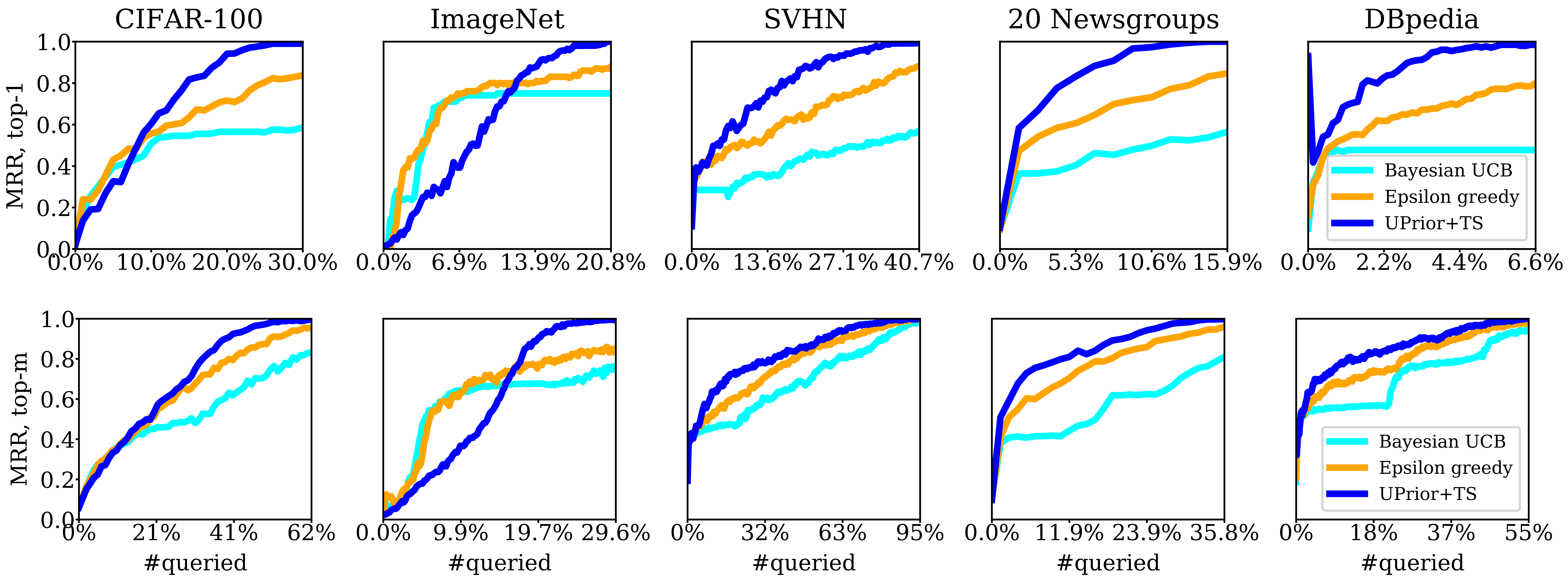
USE SELF-ASSESSMENT AS INFORMATIVE PRIOR



BAYESIAN RELIABILITY DIAGRAMS

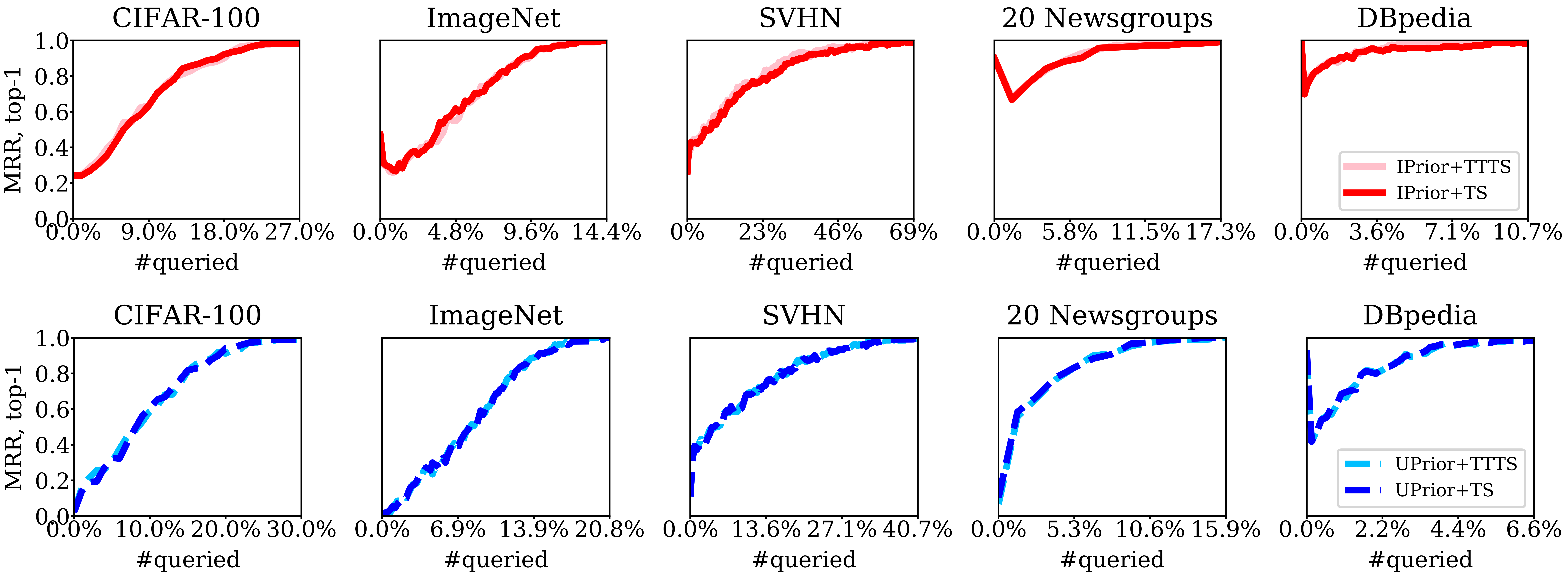


COMPARISONS WITH ALTERNATIVE ACTIVE LEARNING ALGORITHMS



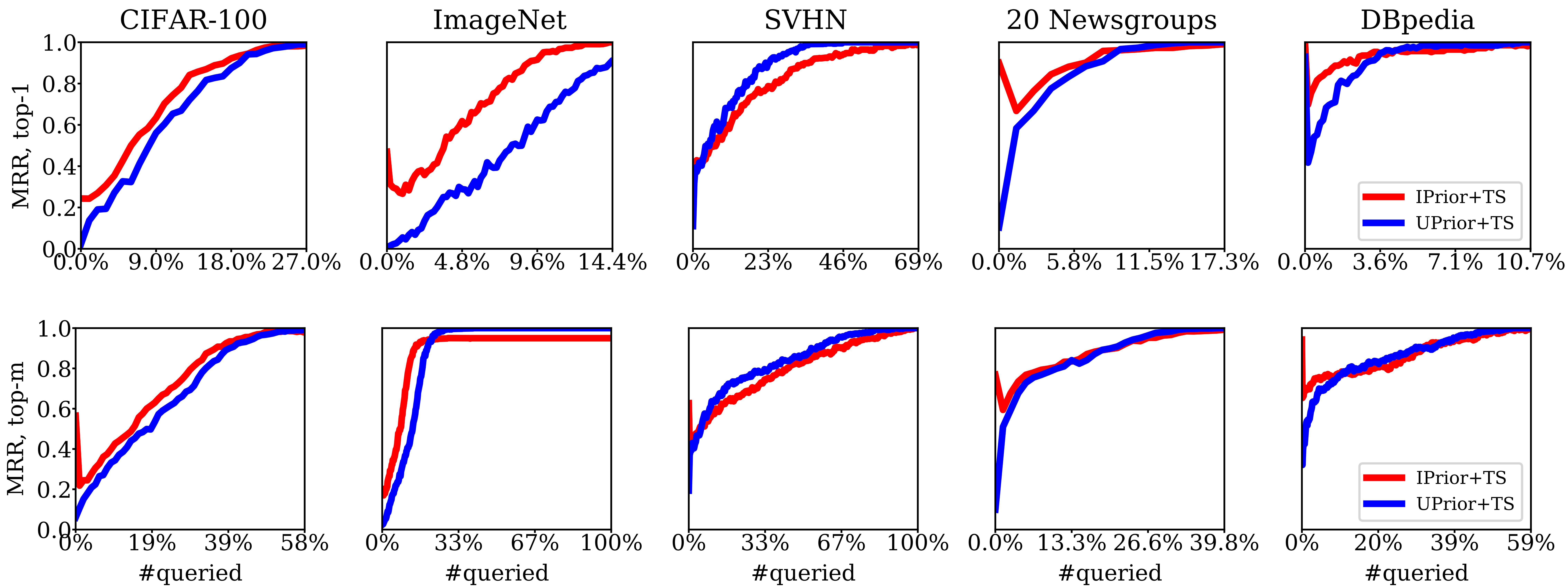
Least accuracy classes

COMPARISONS BETWEEN TS AND TTTS



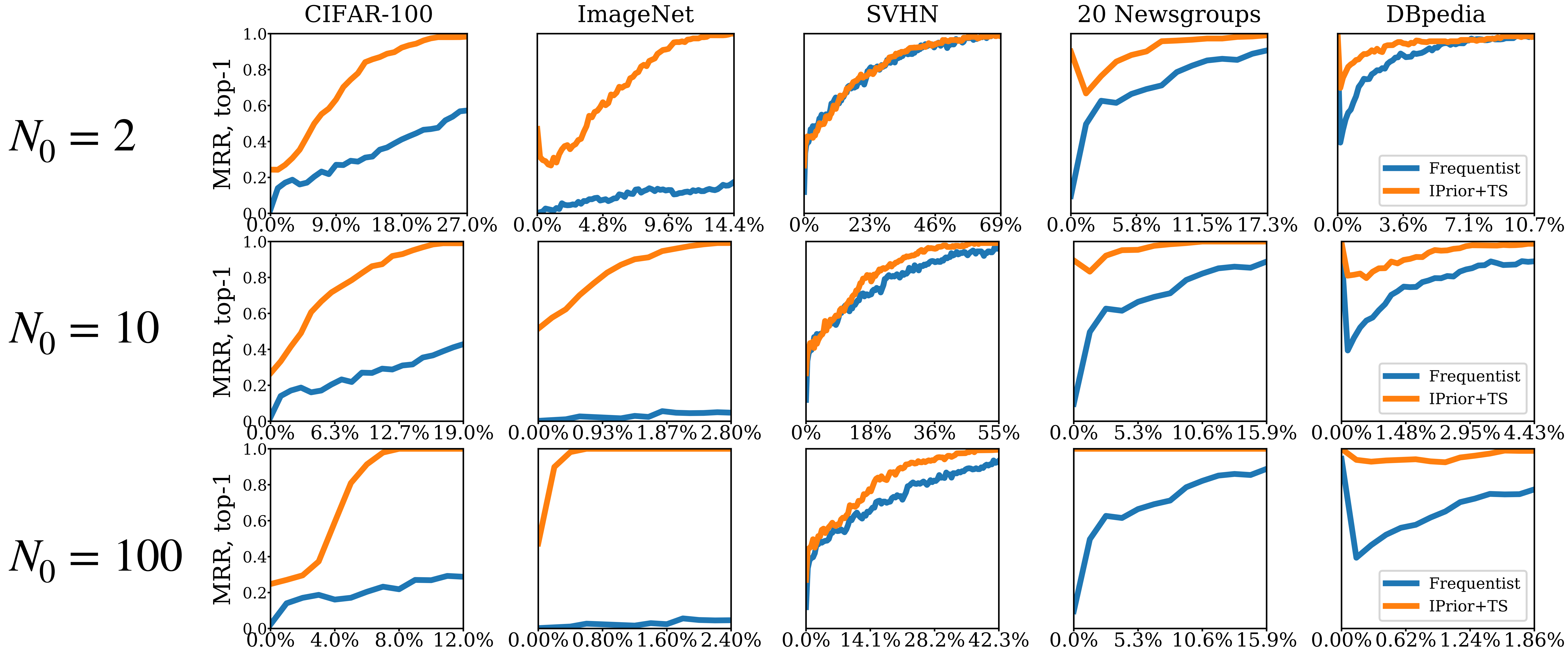
Least accuracy classes

COMPARISONS BETWEEN IPRIOR+TS AND UPRIOR+TS



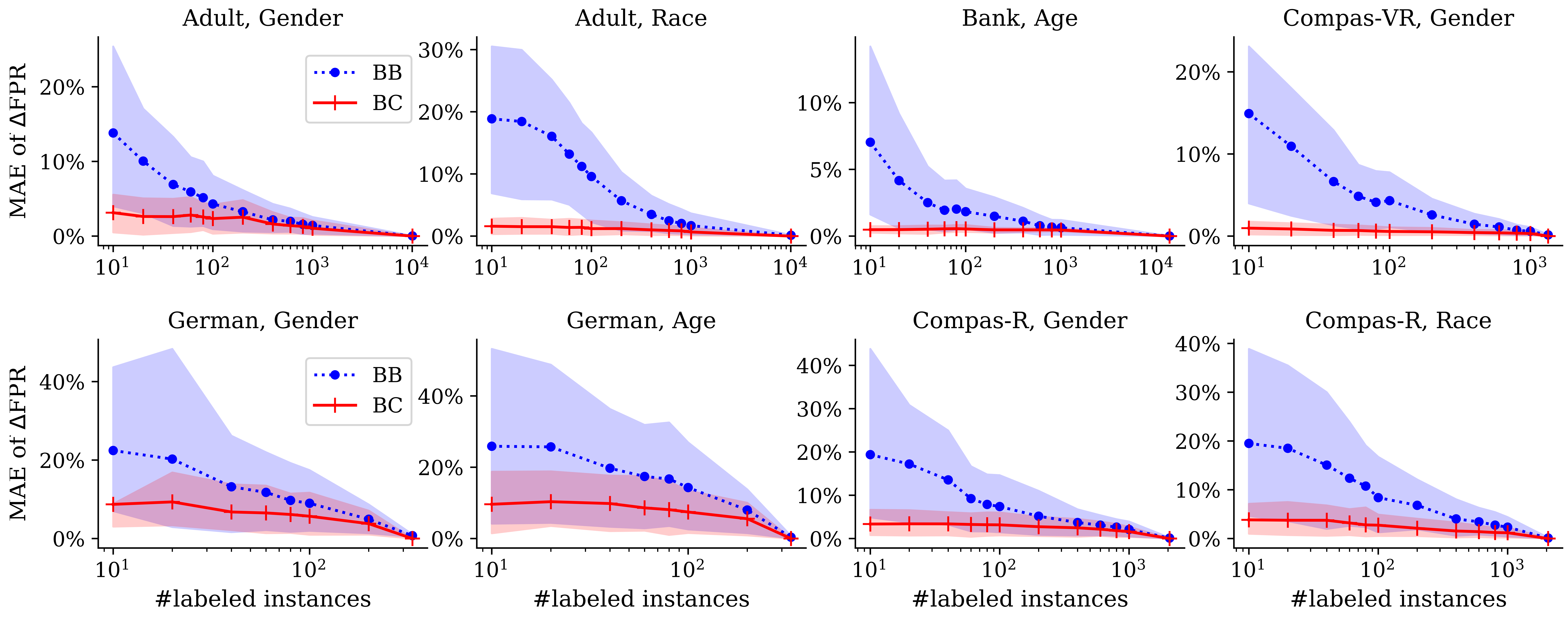
Least accuracy classes

SENSITIVITY ANALYSIS FOR HYPERPARAMETERS



Least accuracy classes

ΔFPR ESTIMATION



(FAIRNESS) CALIBRATION OF THE POSTERIOR PROBABILITY

Table 4.4: **Calibration Coverage of Posterior Credible Intervals Comparison**, across 1000 runs of labeled samples of different sizes n_L for 10 different dataset-group combinations (rows). Estimation methods are BC (Bayesian-Calibration) and BB (beta-bernoulli). Trained model is multi-layer perceptron.

	$n_L = 10$		$n_L = 20$		$n_L = 40$		$n_L = 100$	
Group	BC	BB	BC	BB	BC	BB	BC	BB
Adult, Race	99.9	97.7	98.6	93.5	96.2	93.2	92.3	95.3
Adult, Gender	100.0	96.4	99.7	95.5	99.2	94.9	96.8	95.5
Bank, Age	99.4	98.7	98.8	98.5	98.0	96.4	93.7	95.3
German, age	99.9	98.8	99.6	98.1	99.0	98.3	96.9	98.3
German, Gender	99.1	97.4	99.1	97.4	97.7	96.4	94.6	97.8
Compas-R, Race	99.3	98.8	99.4	97.2	99.1	96.7	99.3	96.6
Compas-R, Gender	99.3	97.7	99.3	97.0	98.6	95.9	97.6	96.5
Compas-VR, Race	99.6	100.0	98.6	97.8	97.9	95.2	97.5	93.1
Compas-VR, Gender	96.3	97.2	94.3	96.5	95.4	96.1	95.8	97.1
Ricci, Race	93.2	99.7	91.4	99.7	—	—	—	—

COMPARISONS WITH LLO CALIBRATION

		Multi-layer Perceptron		Logistic Regression		Random Forest		Gaussian Naive Bayes	
Group	n	BC	LLO	BC	LLO	BC	LLO	BC	LLO
Adult Race	10	3.9	3.8	2.9	2.8	3.2	3.2	3.6	3.5
	100	3.5	3.4	3.2	3.1	3.1	2.9	2.8	2.4
	1000	1.6	2.3	1.7	2.0	1.4	1.5	1.4	1.6
Adult Gender	10	5.1	5.1	2.2	2.3	4.8	4.7	5.4	5.0
	100	4.4	4.3	1.9	2.0	4.1	3.7	2.7	2.7
	1000	1.6	2.2	1.1	1.0	2.0	1.5	1.1	1.1
Bank Age	10	2.5	2.3	1.4	1.2	1.0	0.9	1.7	1.7
	100	2.0	2.0	1.2	1.2	0.9	0.9	1.1	1.2
	1000	1.1	1.2	0.7	0.7	0.5	0.5	0.8	0.9
German age	10	5.0	4.6	8.7	8.0	8.2	7.5	11.5	10.7
	100	3.9	4.1	3.8	4.7	4.3	4.0	4.2	6.0
	200	3.1	3.9	3.3	4.2	3.3	3.1	3.5	6.0
German Gender	10	8.2	6.4	6.3	5.0	8.6	6.9	6.5	5.3
	100	5.4	5.1	3.7	3.6	4.8	4.5	2.8	3.1
	200	3.0	3.4	2.9	2.8	2.9	3.1	2.2	2.9
Compas-R Race	10	4.2	4.6	4.8	5.2	2.4	2.5	8.4	8.2
	100	2.8	4.4	3.4	4.8	1.8	1.4	6.0	5.6
	1000	1.6	5.0	1.6	4.4	1.2	1.1	1.8	2.9
Compas-R Gender	10	5.0	4.3	3.8	3.9	4.4	4.1	13.7	13.0
	100	3.3	2.7	2.6	2.3	2.7	2.8	8.0	7.4
	1000	1.4	2.1	1.3	1.3	1.4	3.0	1.8	2.4
Compas-VR Race	10	4.0	3.9	4.4	4.7	2.4	2.9	6.5	6.4
	100	3.1	2.8	3.4	3.3	2.0	2.1	3.7	3.6
	1000	0.8	1.5	0.8	0.8	0.8	2.5	0.9	1.8
Compas-VR Gender	10	5.4	4.8	5.3	5.2	6.3	8.2	9.8	9.0
	100	3.4	3.0	3.1	3.3	4.4	5.4	4.5	4.2
	1000	0.9	1.2	0.9	1.5	1.0	1.7	0.9	0.9
Ricci Race	10	14.6	14.2	7.9	8.1	2.1	2.0	1.6	2.1
	20	9.8	13.6	7.1	6.6	1.5	1.6	2.1	2.5
	30	6.5	12.1	4.6	4.2	1.1	1.4	2.0	2.3

IS THE HIERARCHICAL STRUCTURE NECESSARY?

Group	n	Multi-layer Perceptron BB	NHBC	BC	Logistic Regression BB	NHBC	BC	Random Forest BB	NHBC	BC	Gaussian Naive Bayes BB	NHBC	BC
Adult Race	10	18.4	**3.2**	3.9	18.8	**2.7**	2.9	18.1	**2.8**	3.2	18.9	4.5	**3.6**
	20	16.1	**3.3**	4.4	16.7	**2.9**	3.4	16.3	**3.0**	3.7	16.8	4.1	**3.7**
	40	13.1	**2.8**	4.5	14.0	**2.9**	3.7	14.4	**2.9**	3.8	14.4	3.7	**3.3**
	100	8.6	**2.7**	3.5	9.2	**3.0**	3.2	9.0	**2.6**	3.1	9.6	**2.4**	2.8
	1000	2.5	**1.4**	1.6	2.3	2.1	**1.7**	2.1	**0.7**	1.4	2.3	1.8	**1.4**
Adult Gender	10	17.4	**4.1**	5.1	16.3	2.6	**2.2**	17.3	5.3	**4.8**	16.3	7.2	**5.4**
	20	12.9	**4.4**	5.1	12.2	2.6	**2.2**	12.4	5.3	**4.9**	11.6	6.7	**4.5**
	40	9.0	**4.1**	4.9	9.2	2.5	**2.1**	9.6	5.1	**4.5**	9.7	6.3	**3.9**
	100	5.4	**3.1**	4.4	5.5	2.0	**2.0**	5.9	4.7	**4.1**	6.0	4.8	**2.7**
	1000	1.9	**1.4**	1.6	1.7	**1.0**	1.1	**1.5**	1.8	2.0	1.5	**0.9**	1.0
Bank Age	10	14.0	**1.7**	2.5	12.8	1.5	**1.4**	11.2	1.1	**1.0**	13.7	**1.4**	1.7
	20	11.6	**2.3**	2.9	10.9	1.9	**1.7**	8.8	1.4	**1.2**	10.3	**1.6**	1.7
	40	8.0	**2.3**	2.6	7.3	1.7	**1.4**	6.5	1.5	**1.1**	7.5	1.7	**1.5**
	100	4.3	2.2	**2.0**	4.3	1.4	**1.2**	4.2	1.2	**0.9**	4.9	1.3	**1.1**
	1000	1.5	1.2	**1.1**	1.6	0.8	**0.7**	1.4	0.6	**0.5**	1.7	**0.7**	0.8
German age	10	19.7	5.6	**5.0**	21.3	10.3	**8.7**	19.1	**8.2**	8.2	20.4	14.2	**11.5**
	20	18.1	6.0	**4.4**	18.6	6.7	**6.4**	16.7	**7.0**	7.0	18.8	9.9	**9.0**
	40	15.9	6.7	**4.8**	15.0	5.6	**4.9**	11.7	6.6	**5.8**	14.9	**6.4**	6.9
	100	7.9	5.8	**3.9**	7.5	5.5	**3.8**	8.2	6.5	**4.3**	9.1	4.4	**4.2**
	200	4.2	3.7	**3.1**	4.4	4.1	**3.3**	4.7	4.1	**3.3**	4.7	3.8	**3.5**
German Gender	10	21.5	10.5	**8.2**	17.6	7.0	**6.3**	19.4	**8.5**	8.6	20.0	**5.9**	6.5
	20	16.2	10.0	**7.8**	13.2	7.1	**5.1**	14.1	8.4	**7.8**	15.4	5.9	**4.9**
	40	11.6	9.2	**6.6**	11.4	8.4	**4.5**	11.1	7.7	**5.9**	11.1	6.1	**3.8**
	100	7.1	6.5	**5.4**	6.9	6.6	**3.7**	7.0	6.1	**4.8**	5.9	6.4	**2.8**
	200	3.2	3.3	**3.0**	4.0	4.0	**2.9**	3.6	3.4	**2.9**	4.0	4.0	**2.2**
Compas-R Race	10	21.1	**2.9**	4.2	20.7	**4.0**	4.8	20.3	**1.4**	2.4	23.1	**6.6**	8.4
	20	14.8	**2.8**	3.3	15.2	3.9	**3.8**	15.8	**2.0**	2.5	16.6	**7.8**	8.0
	40	11.7	**3.0**	3.0	12.1	3.9	**3.6**	11.6	**2.0**	2.0	10.9	9.9	**8.1**
	100	6.8	2.9	**2.8**	7.4	3.7	**3.4**	8.5	2.1	**1.8**	7.9	7.7	**6.0**
	1000	2.0	**1.5**	1.6	1.9	**1.6**	1.7	1.9	1.3	**1.2**	1.9	1.9	**1.8**
Compas-R Gender	10	21.3	**3.8**	5.0	22.0	**3.4**	3.8	23.4	**3.5**	4.4	25.4	19.1	**13.7**
	20	18.5	**3.8**	5.1	18.4	**3.3**	4.0	17.4	**3.3**	4.6	21.4	23.8	**12.3**
	40	12.2	**3.4**	4.0	13.0	**3.0**	3.3	13.7	**2.8**	3.6	15.0	23.8	**9.5**
	100	8.8	**3.2**	3.3	9.1	2.7	**2.6**	8.5	**2.1**	2.7	9.8	15.5	**8.0**
	1000	2.0	1.7	**1.4**	2.2	1.4	**1.3**	2.4	1.6	**1.4**	1.9	1.9	**1.8**
Compas-VR Race	10	17.4	4.0	**4.0**	15.6	**4.4**	4.4	15.7	2.6	**2.4**	19.7	**6.1**	6.5
	20	13.5	4.7	**4.3**	13.7	5.0	**4.8**	13.6	3.3	**2.9**	15.9	10.7	**6.5**
	40	9.6	4.5	**3.8**	9.6	4.5	**3.9**	9.9	3.1	**2.4**	11.1	8.8	**5.5**
	100	5.6	3.6	**3.1**	5.2	3.8	**3.4**	6.2	2.6	**2.0**	6.6	6.8	**3.7**
	1000	0.9	0.8	**0.8**	0.9	**0.8**	0.8	0.9	0.8	**0.8**	1.1	1.2	**0.9**
Compas-VR Gender	10	17.2	5.6	**5.4**	16.8	5.7	**5.3**	19.0	**5.8**	6.3	21.3	18.9	**9.8**
	20	13.3	5.4	**5.1**	14.1	5.4	**4.9**	14.0	**5.7**	6.2	16.0	28.2	**8.7**
	40	9.3	5.1	**4.7**	9.7	4.9	**4.5**	10.5	**5.3**	5.7	12.4	30.9	**6.9**
	100	6.4	3.7	**3.4**	5.9	3.5	**3.1**	6.3	**4.2**	4.4	7.1	18.5	**4.5**
	1000	1.0	**0.8**	0.9	1.0	**0.9**	0.9	0.9	**0.9**	1.0	1.4	0.9	**0.9**
Ricci Race	10	17.7	16.1	**14.6**	14.4	**7.5**	7.9	12.2	**1.9**	2.1	13.1	1.7	**1.6**
	20	11.2	11.8	**9.8**	9.3	7.2	**7.1**	8.5	**1.5**	1.5	9.5	**2.0**	2.1
	30	7.4	7.7	**6.5**	5.8	5.1	**4.6**	6.0	1.1	**1.1**	6.4	**1.9**	2.0

SENSITIVITY ANALYSIS FOR THE CALIBRATION PRIORS

Method	Multi-layer Perceptron			Logistic Regression			Random Forest			Gaussian Naive Bayes		
	10	100	1000	10	100	1000	10	100	1000	10	100	1000
BB	18.52	8.48	2.46	18.74	9.14	2.30	18.24	9.00	2.12	18.88	9.54	2.32
BC, α=0.1	2.63	2.60	2.27	2.46	2.49	2.13	2.87	2.84	2.43	4.67	4.51	0.78
BC, α=0.2	2.63	2.56	2.08	2.46	2.51	2.06	2.85	2.83	2.09	4.63	3.95	0.82
BC, α=0.3	2.60	2.52	1.88	2.42	2.51	1.95	2.85	2.79	1.86	4.44	3.36	0.97
BC, α=0.4	2.49	2.46	1.74	2.41	2.57	1.90	2.74	2.82	1.70	4.25	3.06	1.11
BC, α=0.5	2.49	2.38	1.71	2.44	2.60	1.82	2.82	2.77	1.65	4.01	2.86	1.43
BC, α=0.6	2.47	2.37	1.62	2.55	2.62	1.75	2.82	2.88	1.60	3.81	2.79	1.46
BC, α=0.7	2.61	2.48	1.51	2.36	2.63	1.70	2.90	2.86	1.54	3.54	2.80	1.50
BC, α=0.8	2.86	2.30	1.47	2.52	2.73	1.63	2.87	2.86	1.46	3.51	2.77	1.60
BC, α=0.9	2.93	2.27	1.43	2.44	2.82	1.64	2.87	2.90	1.46	3.14	2.91	1.58
BC, α=1.0	3.05	2.31	1.50	2.71	2.74	1.57	2.99	2.96	1.42	3.31	2.85	1.68
BC, α=1.1	3.14	2.37	1.45	2.65	2.86	1.55	2.90	3.10	1.40	3.25	3.03	1.65
BC, α=1.2	3.11	2.19	1.49	2.73	2.80	1.52	3.27	3.01	1.39	3.20	3.03	1.68
BC, α=1.3	3.48	2.30	1.51	2.91	2.94	1.54	3.11	3.21	1.39	3.15	2.96	1.71
BC, α=1.4	3.76	2.28	1.47	3.17	3.01	1.51	3.26	3.21	1.30	3.48	3.21	1.75
BC, α=1.5	3.67	2.20	1.49	3.12	2.94	1.51	3.46	3.05	1.34	3.23	3.19	1.66
BC, α=1.6	4.06	2.24	1.45	3.26	2.93	1.47	3.56	3.13	1.33	3.48	3.17	1.69
BC, α=1.7	4.02	2.27	1.46	3.46	3.15	1.46	3.75	3.10	1.27	3.43	3.19	1.74
BC, α=1.8	4.35	2.14	1.42	3.36	3.09	1.50	3.76	3.26	1.29	3.67	3.22	1.81
BC, α=1.9	4.35	2.30	1.48	3.48	2.94	1.42	3.54	3.30	1.28	3.82	3.35	1.84
BC, α=2.0	4.69	2.16	1.44	3.87	2.99	1.54	3.91	3.46	1.21	3.83	3.18	1.81
BC, α=5.0	8.11	2.54	1.63	6.31	3.32	1.53	5.32	4.13	1.31	5.25	3.82	2.13
BC, α=10.0	10.39	2.63	1.63	7.18	3.83	1.70	7.19	4.41	1.42	6.32	4.08	2.33

$\mu_a \sim \mathrm{N}(0, .4\alpha), \sigma_a \sim \mathrm{TN}(0, .15\alpha)$

$\mu_b \sim \mathrm{N}(0, .4\alpha), \sigma_b \sim \mathrm{TN}(0, .15\alpha)$

$\mu_c \sim \mathrm{N}(0, 2\alpha), \sigma_c \sim \mathrm{TN}(0, .75\alpha)$