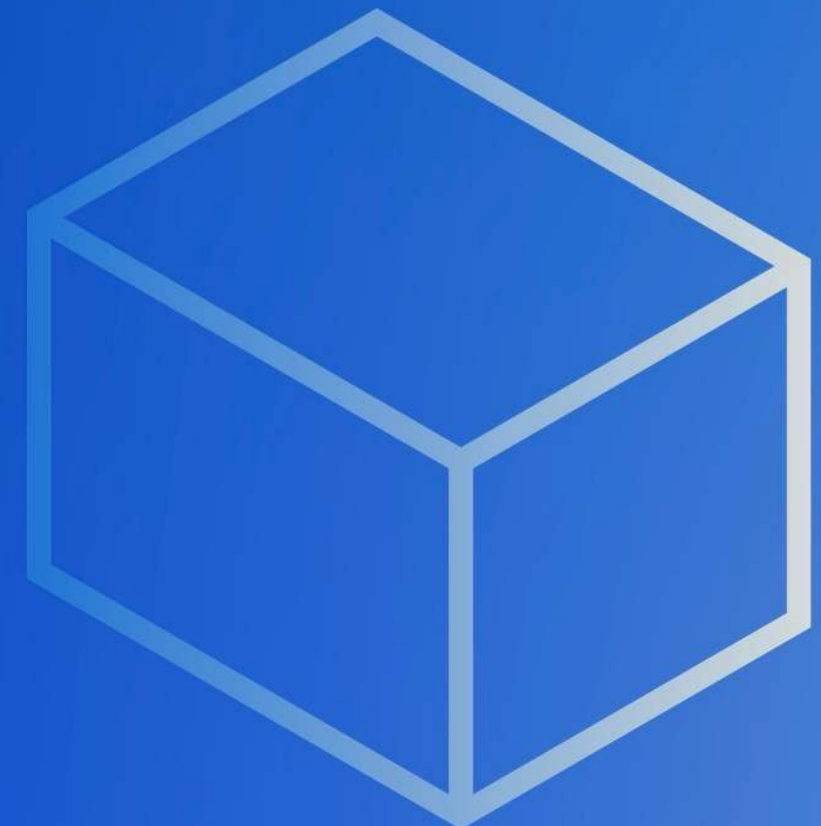




Improving Credit Scoring Models with Bias Correction Algorithms

Nikita Kozodoi, PhD

04/07/2024



About Me

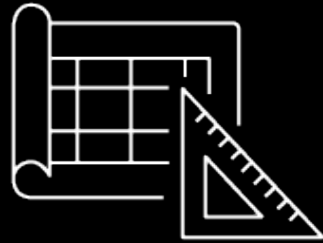


<https://kozodoi.me>

- Applied Scientist at Amazon Web Services
- Building **GenAI solutions** across industries
- Earned PhD in **ML for Credit Risk Analytics**
- Won 18 Kaggle **competition medals**



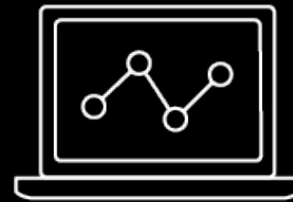
About My Team



Design

Design guidance:

- Select the GenAI use case with the highest business impact
- Design how to develop, train, and deploy it to production



Deploy

Deploy recommended solutions:

- Develop and fine-tune a GenAI solution to meet your business objectives and demonstrate what's possible



Drive

Drive adoption:

- Accelerate stickiness and adoption with a path to production for your GenAI solution integrated into your application.



Presentation Outline

1. Background

- What is credit scoring?
- What are the business goals?

2. Problem Description

- Sampling bias illustration
- Bias impact on ML models

3. Approach

- Improving model evaluation
- Improving model training

4. Results

- Offline evaluation
- Business impact

What is Credit Scoring?

Customer perspective:

Instant loan in 10 minutes

Amount

10 000 ₺

2 000 ₺ 15 000 ₺ 30 000 ₺

Duration

75 days

14 days 90 days

Get money

You pay back: 12 600 ₺

Due date: 7.06.2022

Name

First Name

Occupation

Years of experience

☐ 0-1 Year

☐ 1-2 Years

☐ 3-4 Years

☐ 5+ Years

Gross monthly income

ex: 1500

What is Credit Scoring?

Customer perspective:

Instant loan

Amount

<>

2 000 ₪

Duration

14 days

Get

You pay back:

Due date:

7.06.2022

Name

LOAN APPLICATION

REJECTED

LOAN APPLICATION

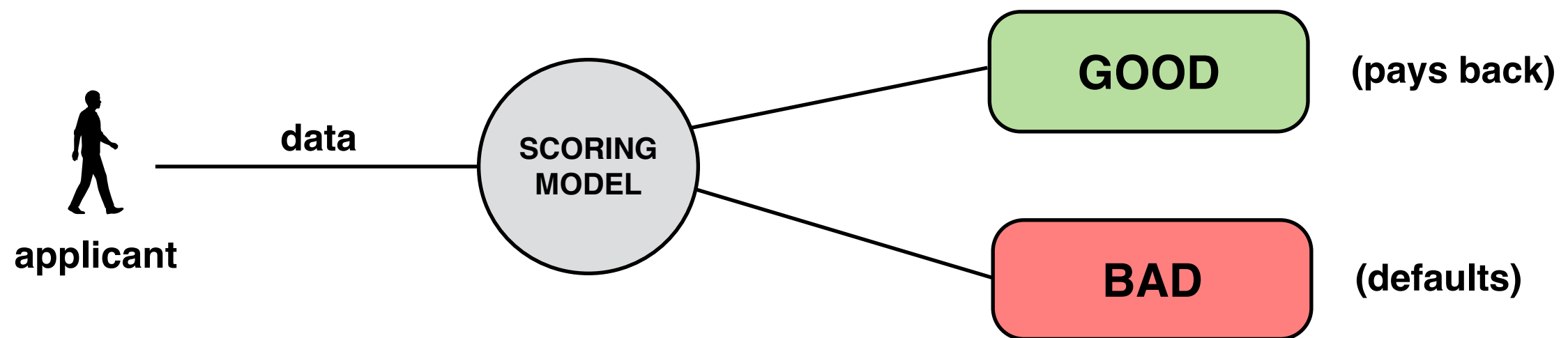
APPROVED

Image source: <https://www.indusind.com/>

What is Credit Scoring?

Business perspective:

- classification task of distinguishing **BAD** and **GOOD** loans
- scorecard — model that predicts probability of default
- increasing reliance on Machine Learning (*e.g., Wei et al. 2016*)
 - consumer credit in the US exceeds \$4,325 billion¹
 - FinTechs account for 49.4% of consumer loan market²



¹ The Federal Reserve: Statistical Release on Consumer Credit (2021)

² Experian: FinTech vs. Traditional FI Trends (2019)

Business Goals

Goal: improving accuracy of credit scoring models

Costs:

- **accepting BAD customer results in a high loss**
 - business: loss = amount that the client does not pay back
 - customer: long-term financial difficulties
- **rejecting GOOD customer results in a moderate loss**
 - business: loss = potential interest and fees earned from the client
 - customer: limited access to finance

Project goal:

- **maximize scorecard profitability**
 - minimize **BAD** rate among accepts

		<u>Decision</u>	
		Accept	Reject
<u>Outcome</u>	GOOD	+ interest	- interest
	BAD	- amount	0

Presentation Outline

1. Background

- What is credit scoring?
- What are the business goals?

2. Problem Description

- Sampling bias illustration
- Bias impact on ML models

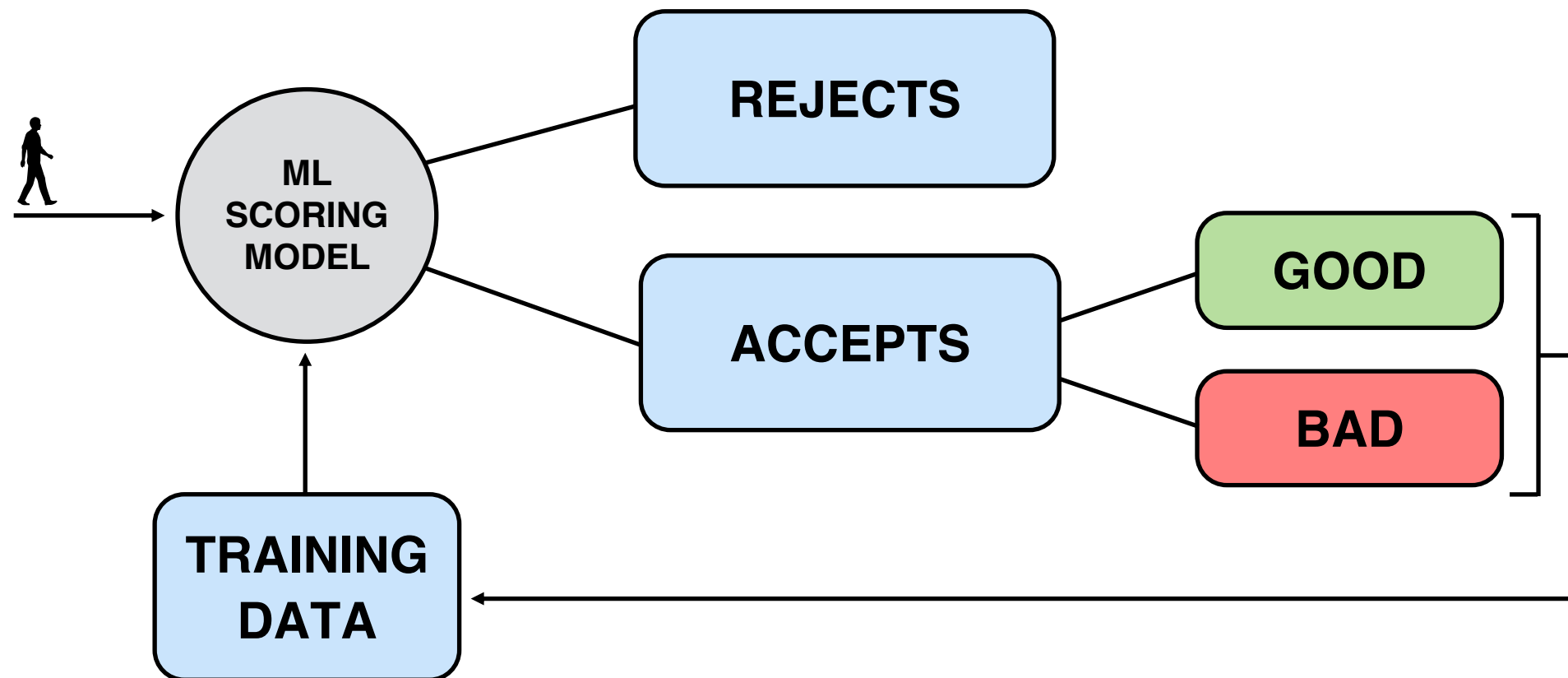
3. Approach

- Improving model evaluation
- Improving model training

4. Results

- Offline evaluation
- Business impact

Loan Approval Process at Monedo

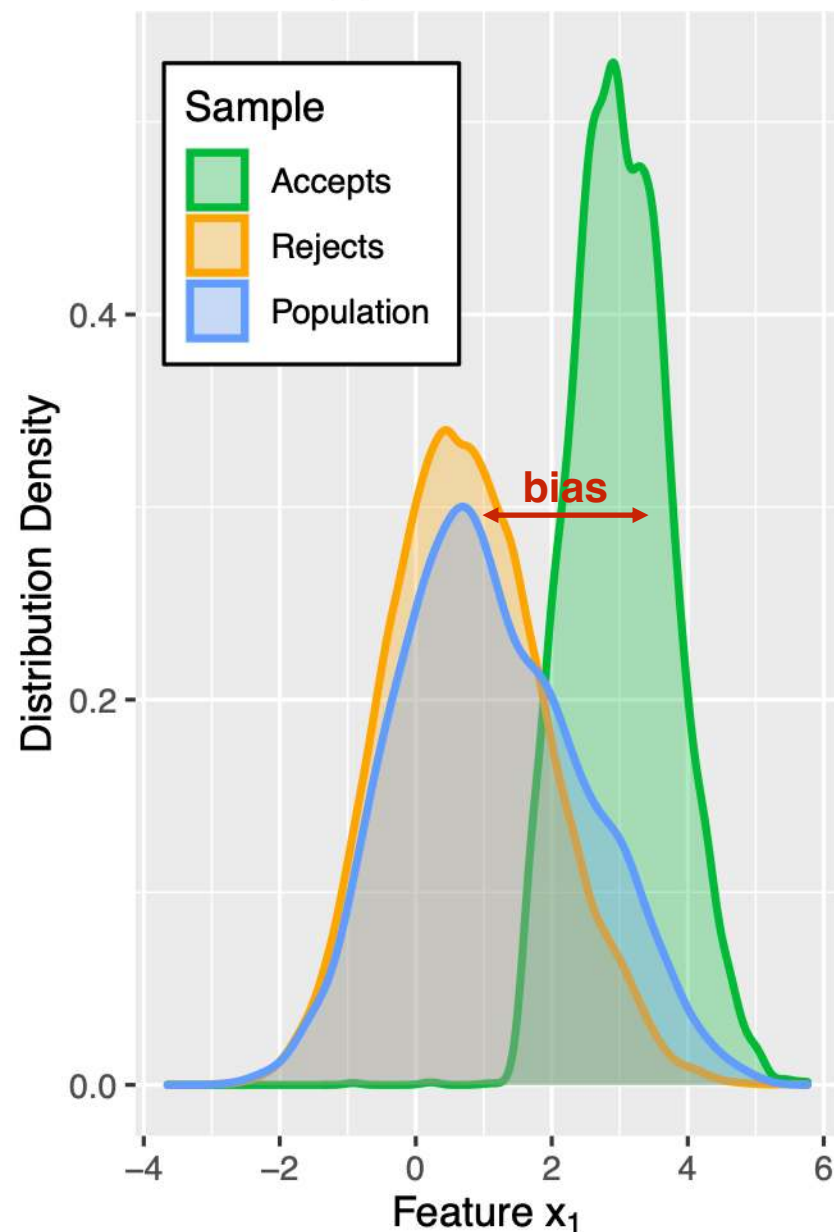


- **scoring model filters incoming loan applications**
 - ML model observes applicants' features and predicts **P(GOOD)**
 - top-ranked applicants are accepted and receive a loan
- **training a model requires data with known outcomes**
 - outcomes are only observed for previously **accepted clients**
 - labels of **rejects** are missing not at random (*Crook et al. 2004*)
 - historical data suffers from sampling bias

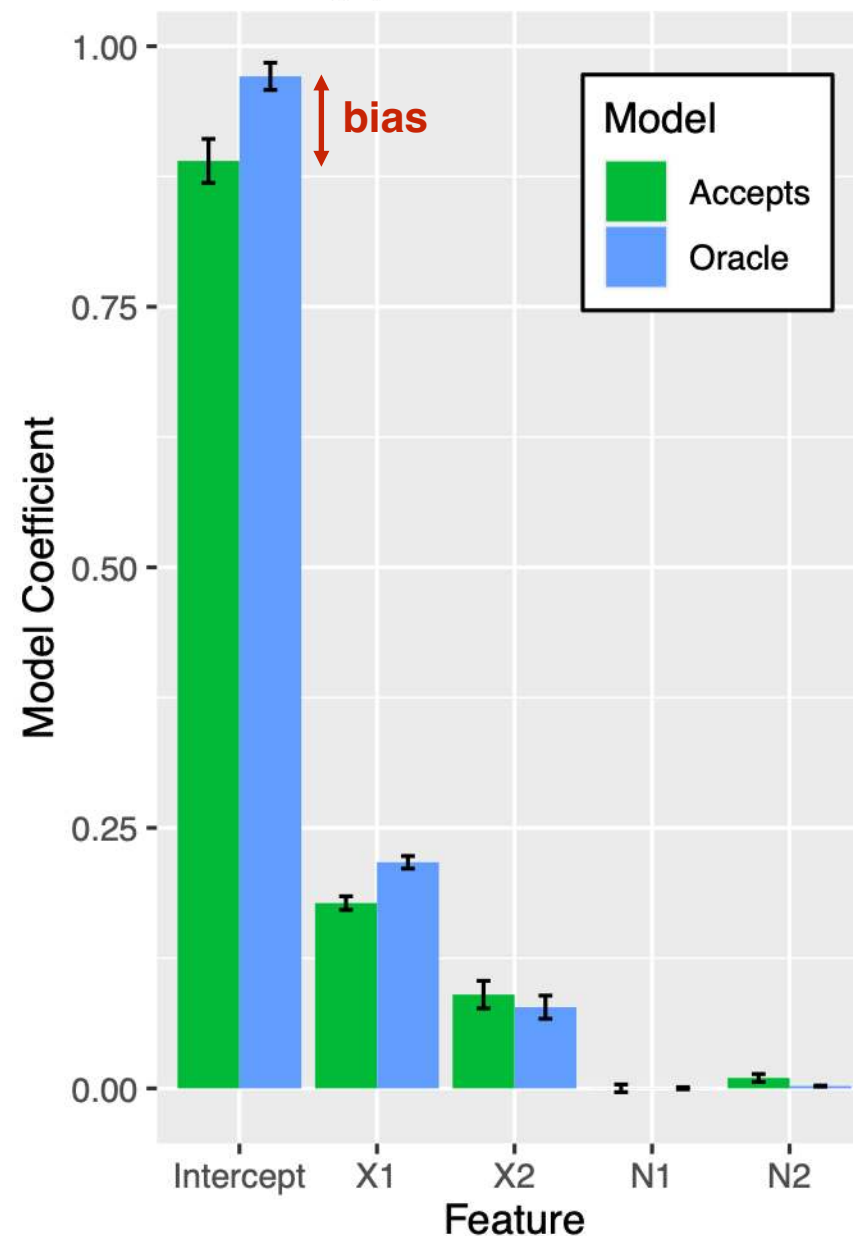
Sampling Bias Illustration

- **sampling bias** originates in the **training data**
- propagates to the **model parameters**
- and affects **model predictions**

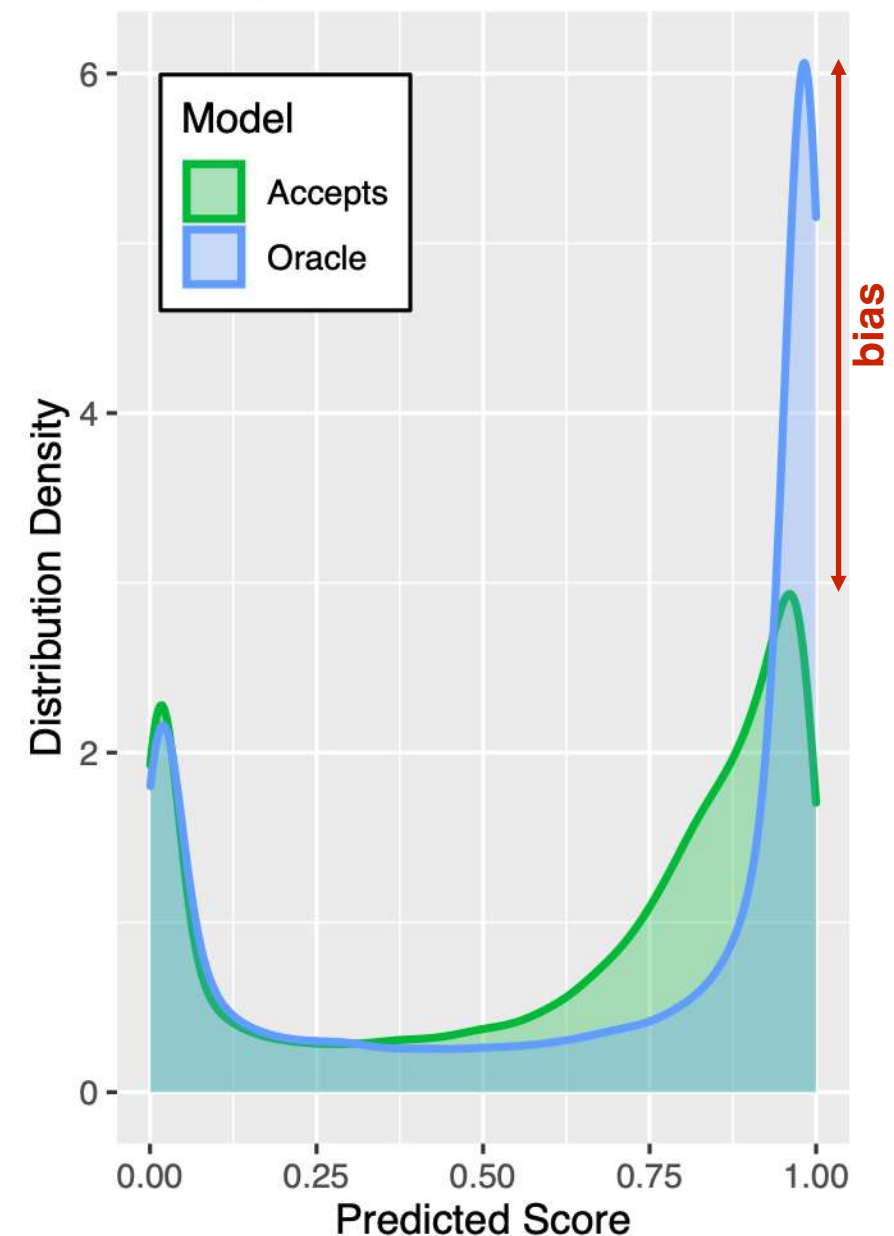
(a) Bias in Data



(b) Bias in Model



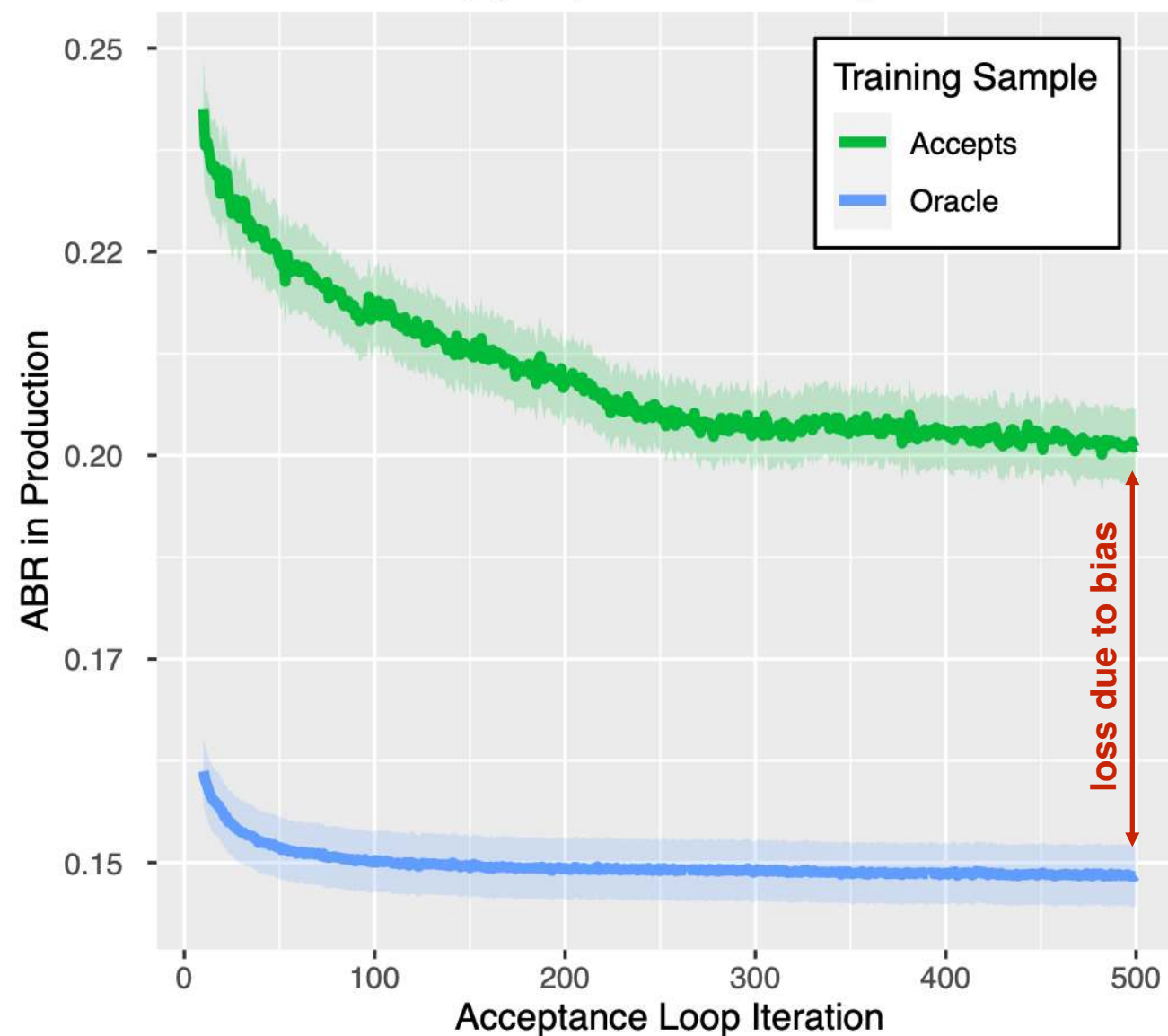
(c) Bias in Predictions



Sampling Bias Consequences

- training a model on a biased sample **decreases its production performance**

(d) Impact on Training



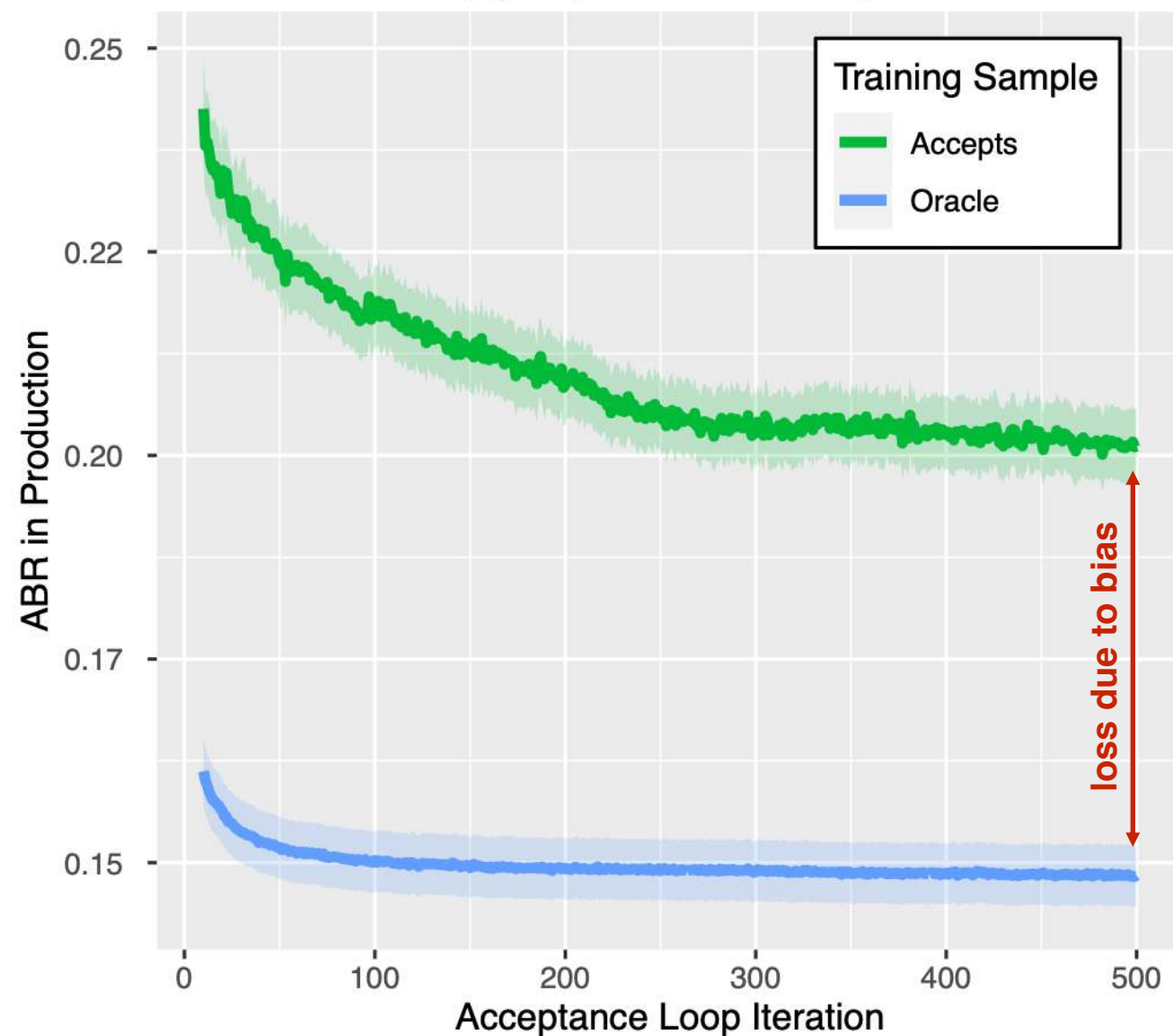
ABR = **BAD** rate when accepting top-30% applicants; lower is better

		Decision	
		Accept	Reject
Outcome	GOOD	+ interest	- interest
	BAD	- amount	0

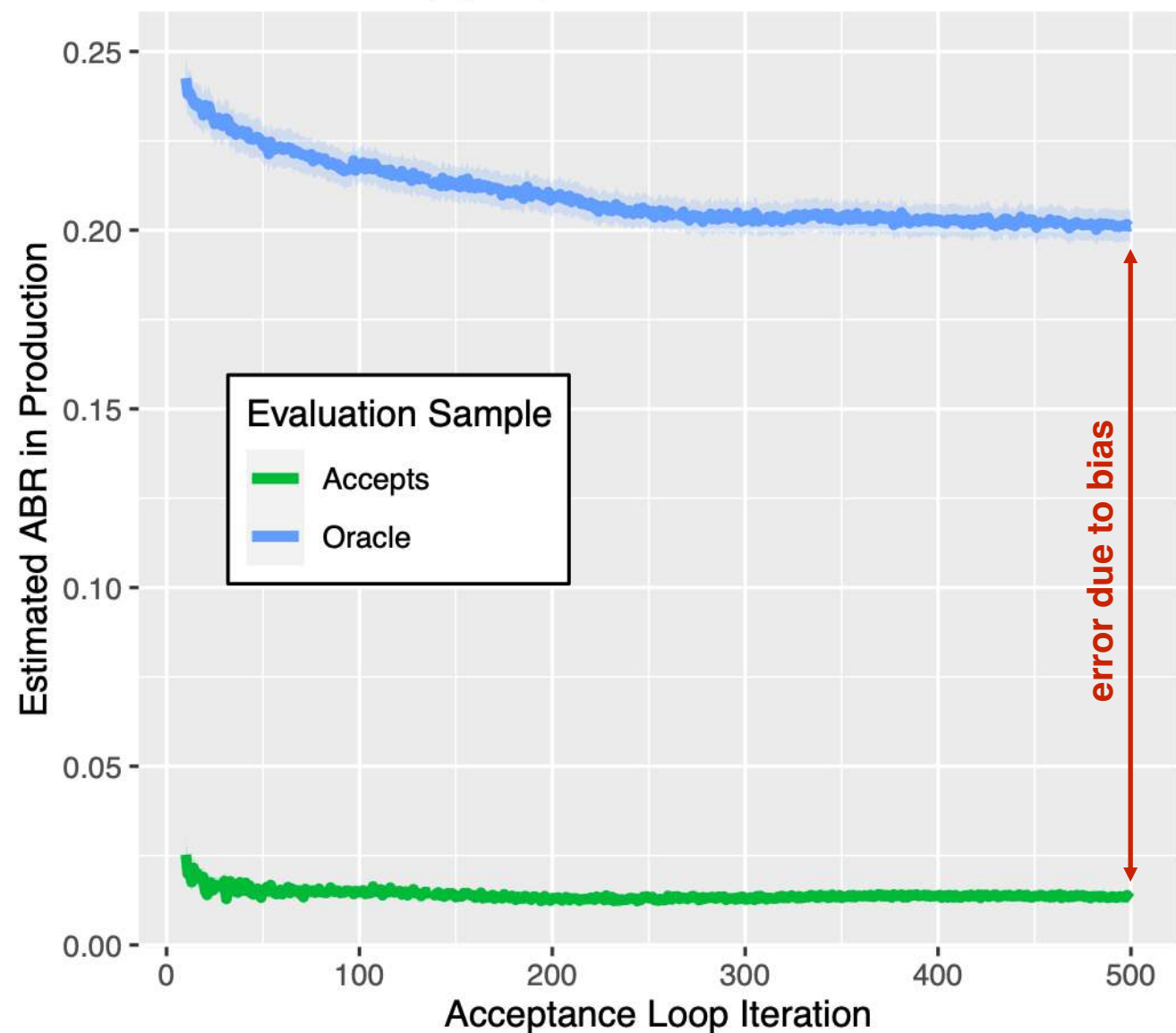
Sampling Bias Consequences

- training a model on a biased sample **decreases its production performance**
- evaluating a model on a biased sample provides a **misleading estimate**

(d) Impact on Training



(e) Impact on Evaluation

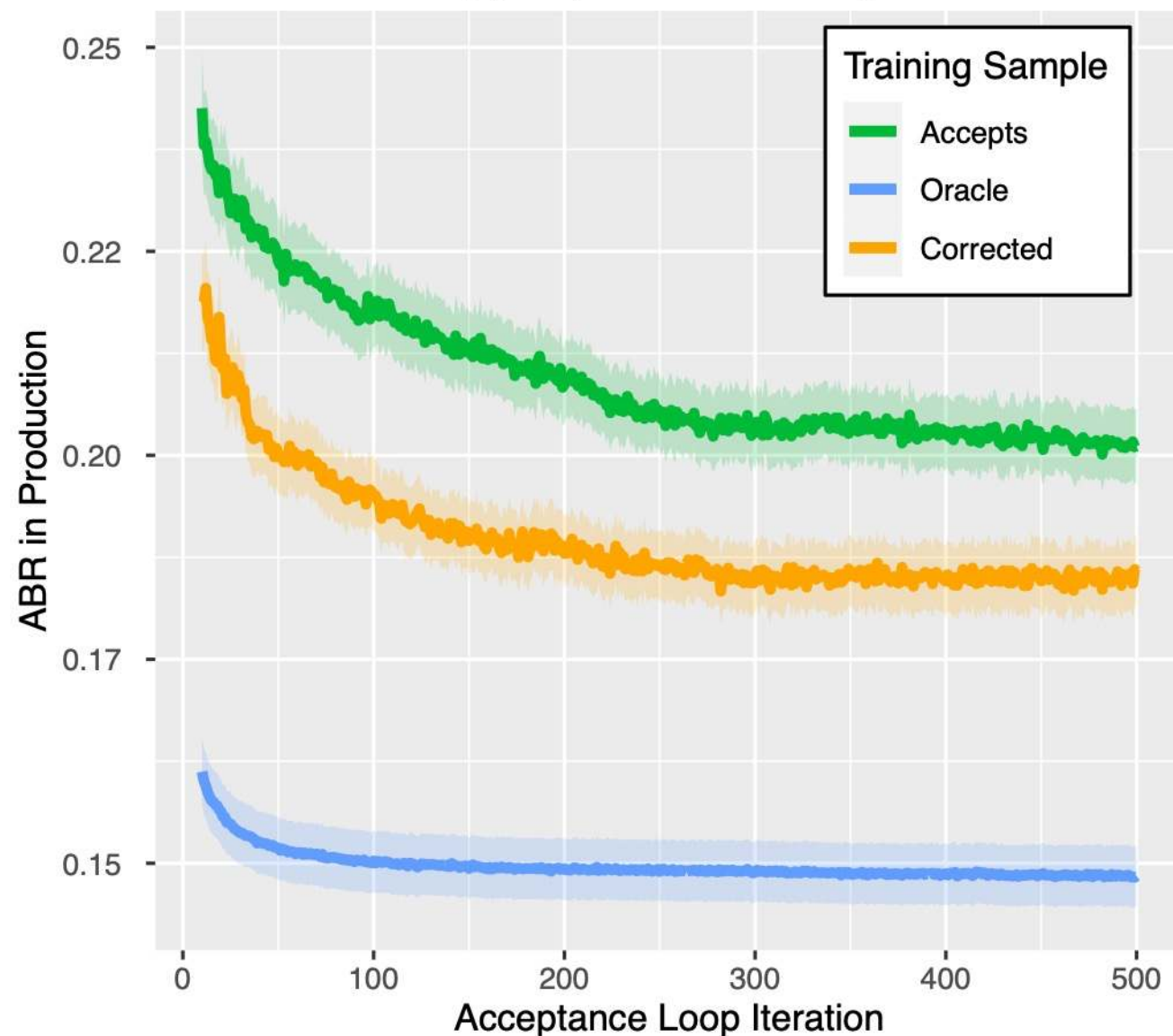


ABR = **BAD** rate when accepting top-30% applicants; lower is better

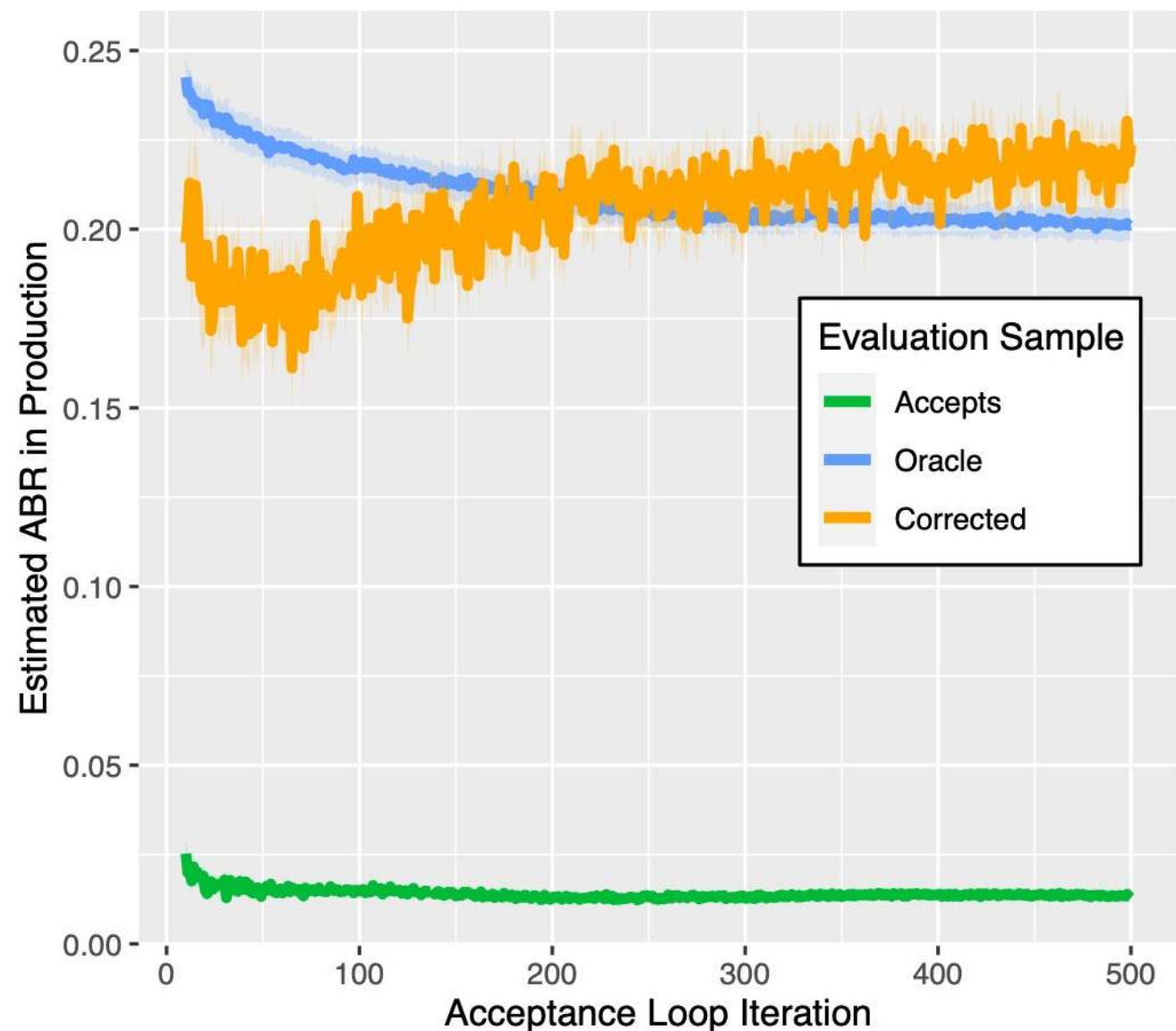
Potential Performance Gains

- **bias correction** can **improve the model performance in production**
- **bias correction** can provide a **better estimate of production performance**

(d) Impact on Training



(e) Impact on Evaluation



ABR = **BAD** rate when accepting top-30% applicants; lower is better

Presentation Outline

1. Background

- What is credit scoring?
- What are the business goals?

2. Problem Description

- Sampling bias illustration
- Bias impact on ML models

3. Approach

- Improving model evaluation
- Improving model training

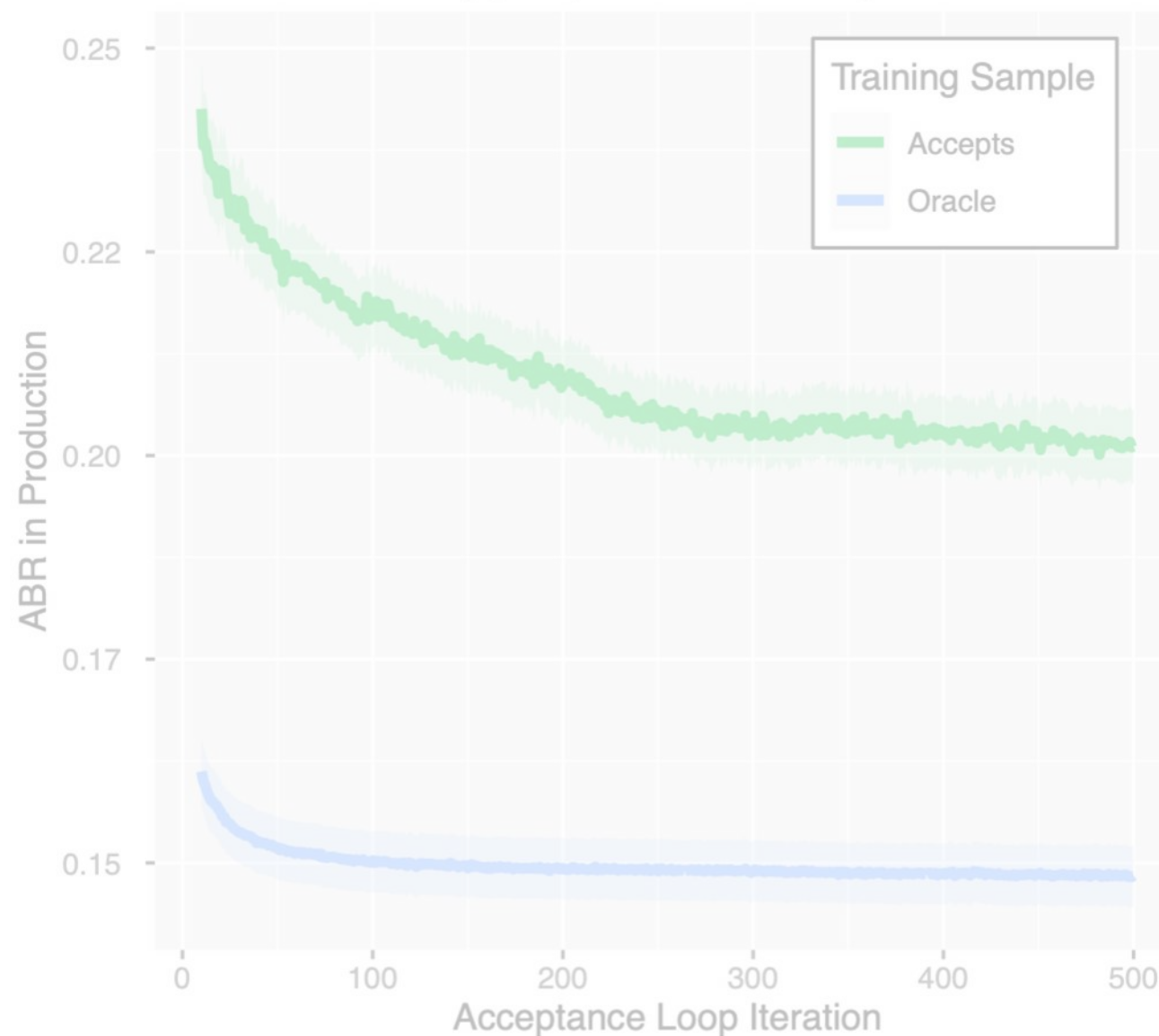
4. Results

- Offline evaluation
- Business impact

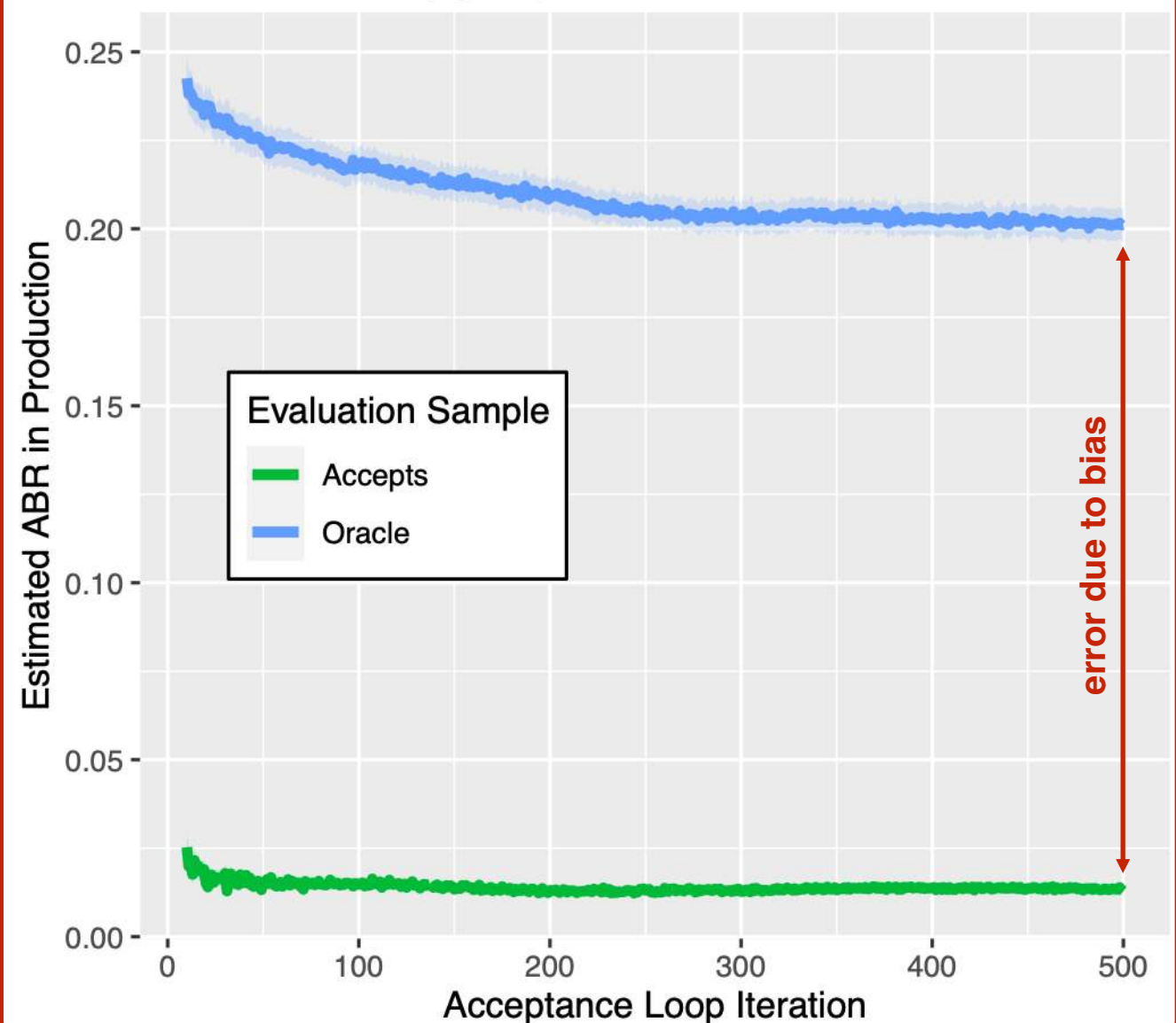
Bias Impact on Evaluation

- training a model on a biased sample **decreases its production performance**
- **evaluating a model** on a biased sample provides a **misleading estimate**

(d) Impact on Training



(e) Impact on Evaluation



ABR = **BAD** rate when accepting top-30% applicants; lower is better

Evaluation under Sampling Bias

How to improve evaluation?

Collect unbiased sample

- completely avoids sampling bias
- requires issuing loans to **random set of applicants** without scoring
- issue: very costly

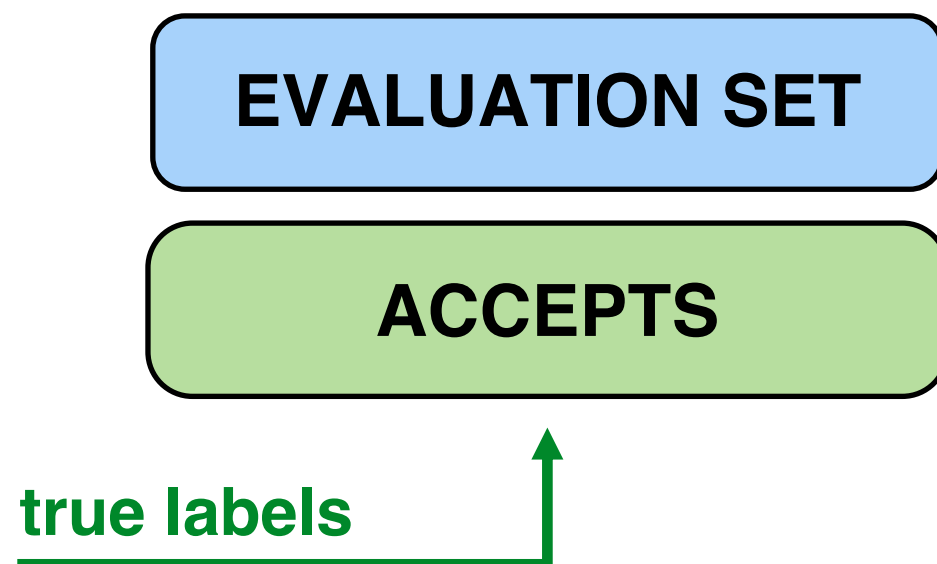
Adjust evaluation framework

- use bias correction methods to account for **distribution mismatch**
- issue: labels of **rejects** are unknown

Standard Practice: Evaluate on Accepts

Idea:

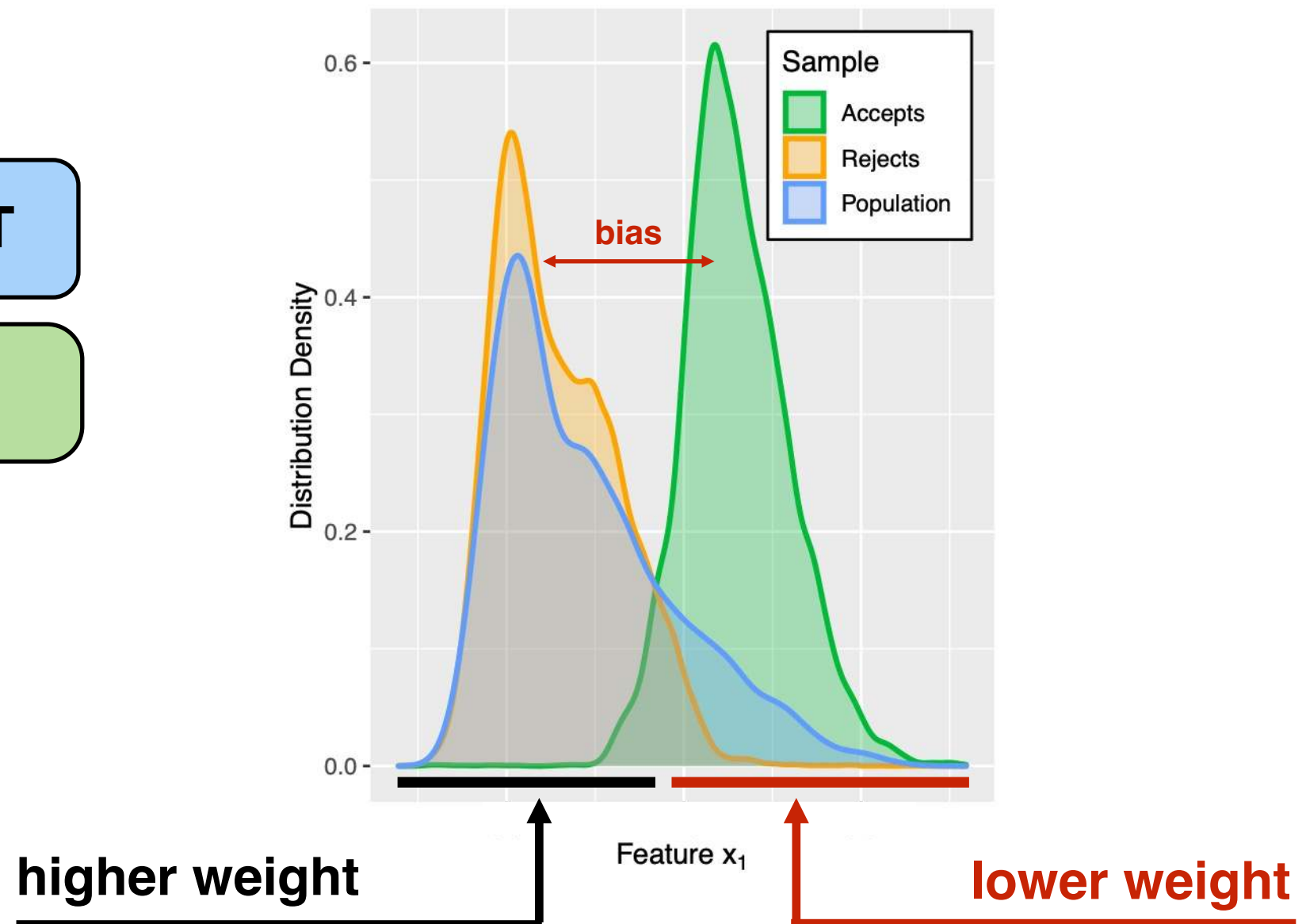
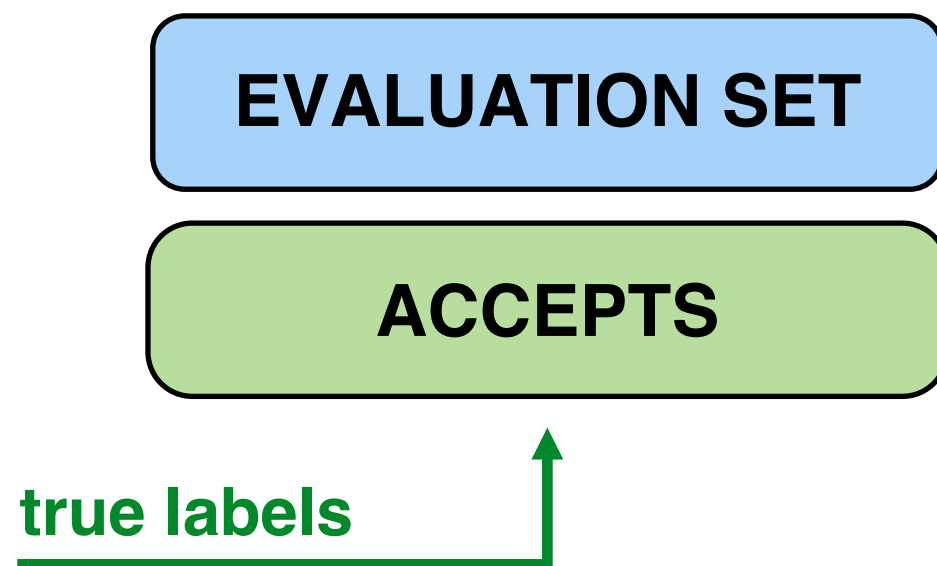
- evaluate metric M on evaluation set containing labeled **accepts**



State-of-the-Art: Reweighting

Idea:

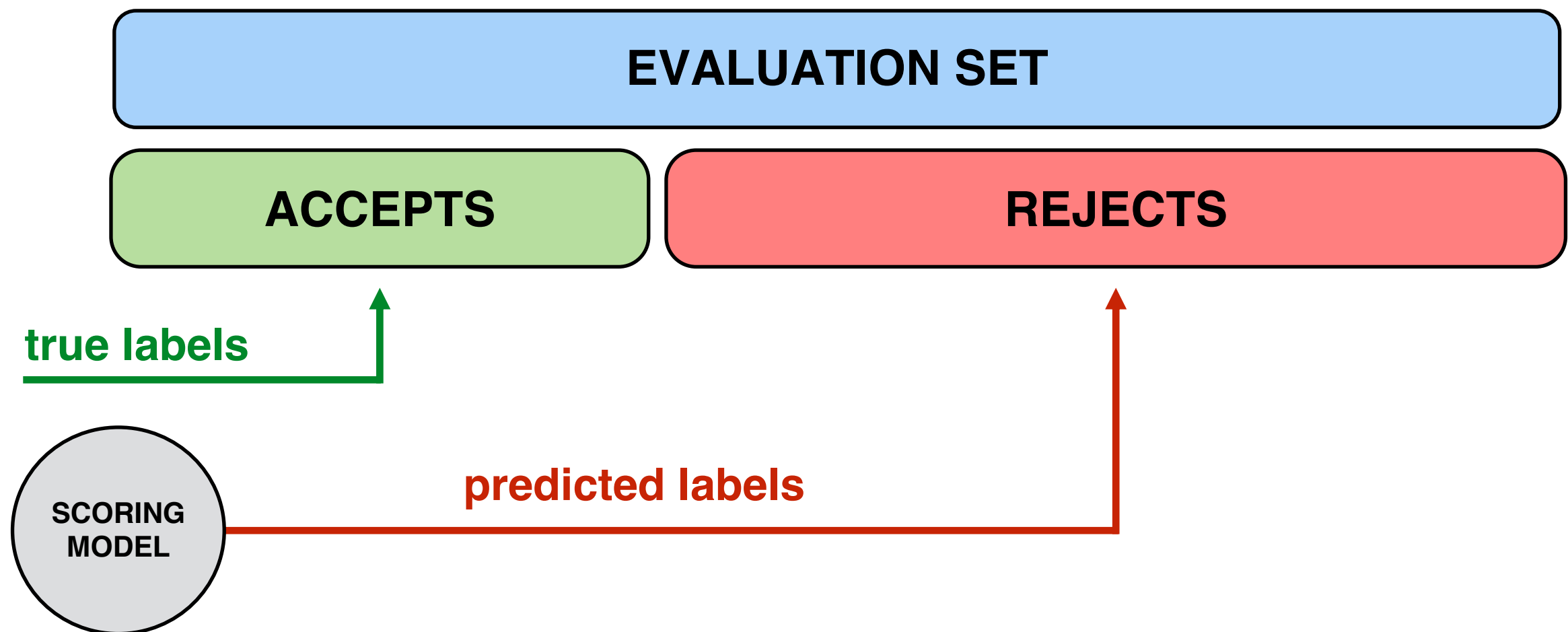
- evaluate metric M on evaluation set containing labeled **accepts**
- reweigh the metric to focus on representative cases



Bayesian Evaluation (BE)

Idea:

- evaluate metric M on evaluation set containing:
 - labeled **accepts**
 - **pseudo-labeled rejects**
- estimate prior **P(BAD)** for **rejects** using the current scorecard $f(X)$



Bayesian Evaluation (BE)

Idea:

- evaluate metric M on evaluation set containing:
 - labeled accepts
 - pseudo-labeled rejects
- estimate prior $P(\text{BAD})$ for rejects using the current scorecard $f(X)$

input : model $f(X)$, evaluation sample S consisting of labeled accepts $S^a = \{(\mathbf{X}^a, \mathbf{y}^a)\}$ and unlabeled rejects $\mathbf{X}^r$, prior $\mathbf{P}(\mathbf{y}^r | \mathbf{X}^r)$, evaluation metric $M(f, S, \tau)$, meta-parameters

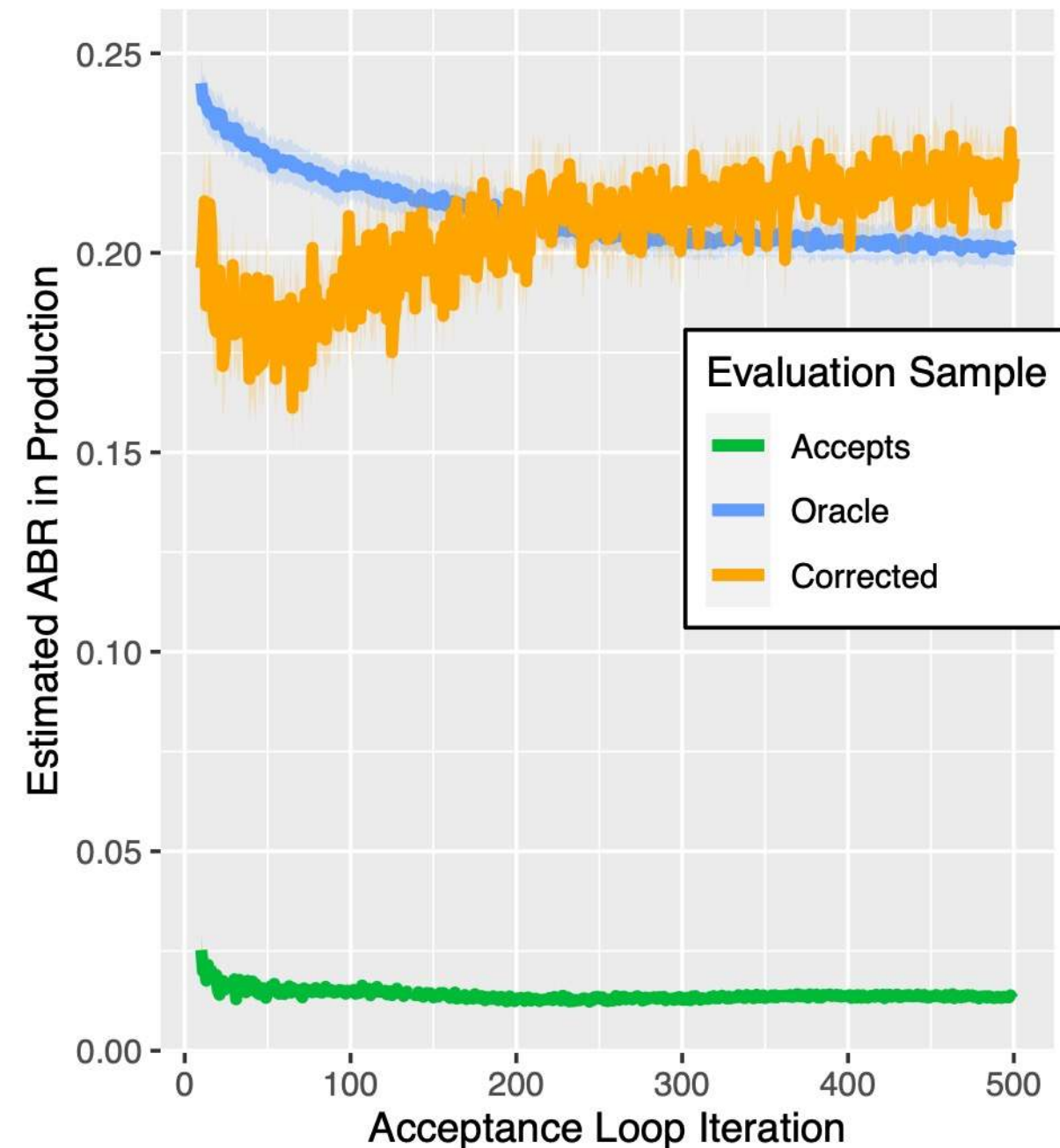
j_{max}, ϵ

output: Bayesian evaluation metric $BM(f, S, \tau)$

```
1  $j = 0; \Delta = \epsilon; E^c = \{\}$  ; // initialization
2 while ( $j \leq j_{max}$ ) and ( $\Delta \geq \epsilon$ ) do
3    $j = j + 1$ 
4    $\mathbf{y}^r = \text{binomial}(1, \mathbf{P}(\mathbf{y}^r | \mathbf{X}^r))$  ; // generate labels of rejects
5    $S_j = \{(\mathbf{X}^a, \mathbf{y}^a)\} \cup \{(\mathbf{X}^r, \mathbf{y}^r)\}$  ; // construct evaluation sample
6    $E_j^c = \sum_{i=1}^j M(f(X), S_i, \tau) / j$  ; // evaluate
7    $\Delta = E_j^c - E_{j-1}^c$  ; // check convergence
8 end
9 return  $BM(f, S, \tau) = E_j^c$ 
```


BE: Simulation Results

Performance Dynamics



Aggregated Results

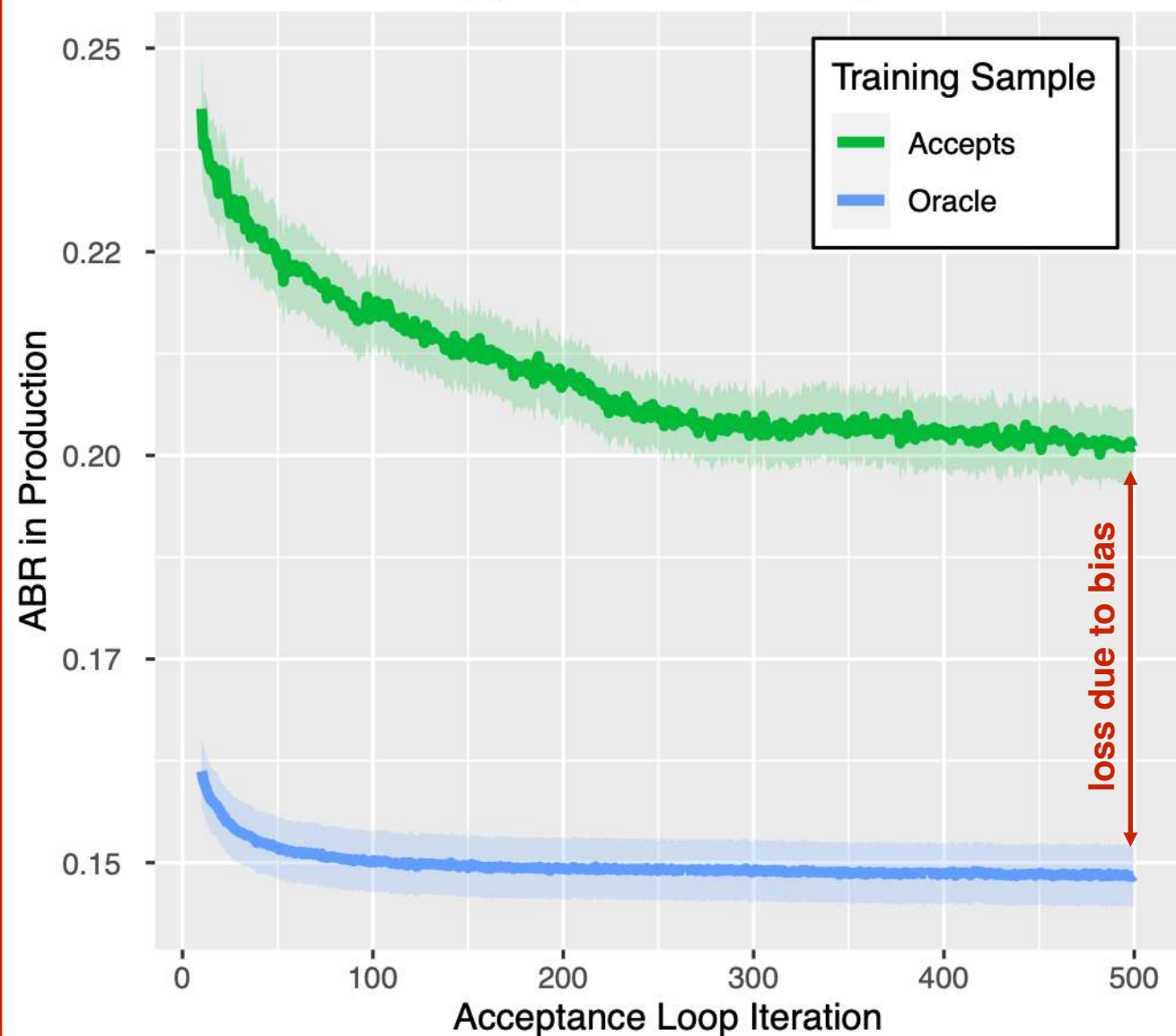
Metric	RMSE due to bias	Gains from BE
ABR	.2058	55.83%
BS	.0829	36.55%
AUC	.2072	67.57%
PAUC	.2699	70.80%

- BE improves **performance estimates**
- gains are **statistically significant** at 5%

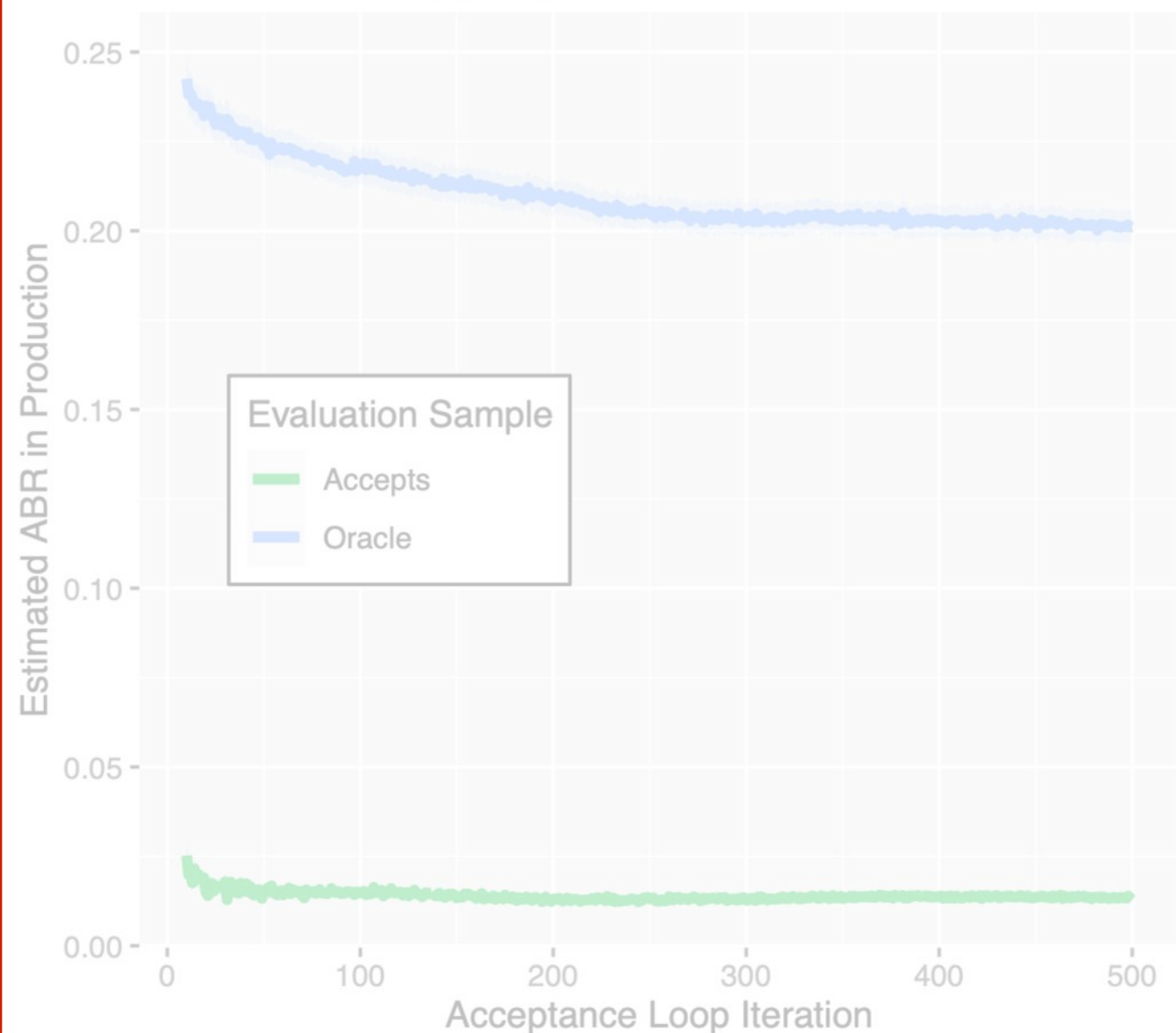
Bias Impact on Training

- training a model on a biased sample **decreases its production performance**
- evaluating a model on a biased sample provides a **misleading estimate**

(d) Impact on Training



(e) Impact on Evaluation



ABR = **BAD** rate when accepting top-30% applicants; lower is better

Training under Sampling Bias

How to improve training?

Collect unbiased sample

- completely avoids sampling bias
- issue: very costly

Data augmentation (label rejects)

- **predict labels** of **rejects**
- use combined data of **accepts** and **rejects** for model training
- issue: high risk of error propagation

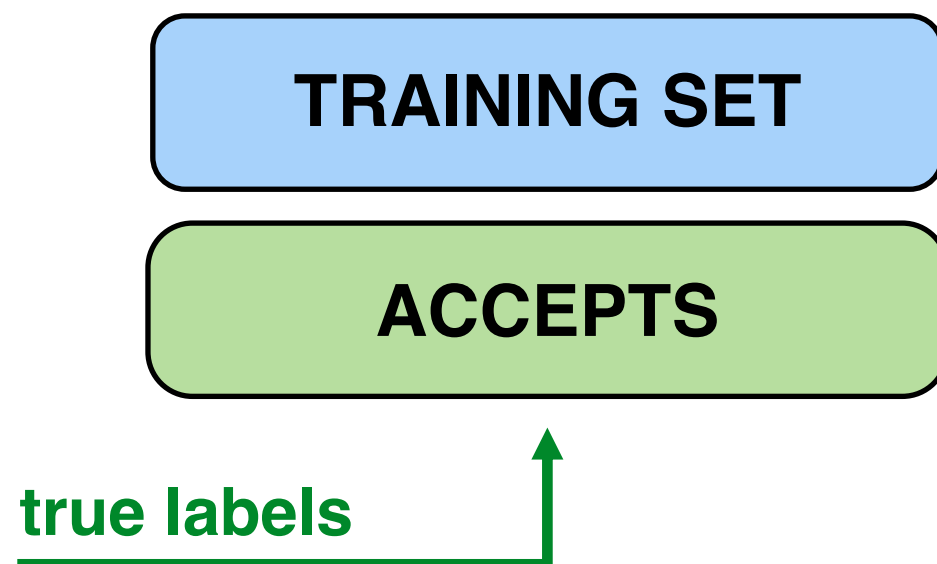
Extract information from rejects

- estimate **distribution mismatch** between **accepts** and **rejects**
- modify training procedure
- issue: hard in high-dimensional data

Standard Practice: Train on Accepts

Idea:

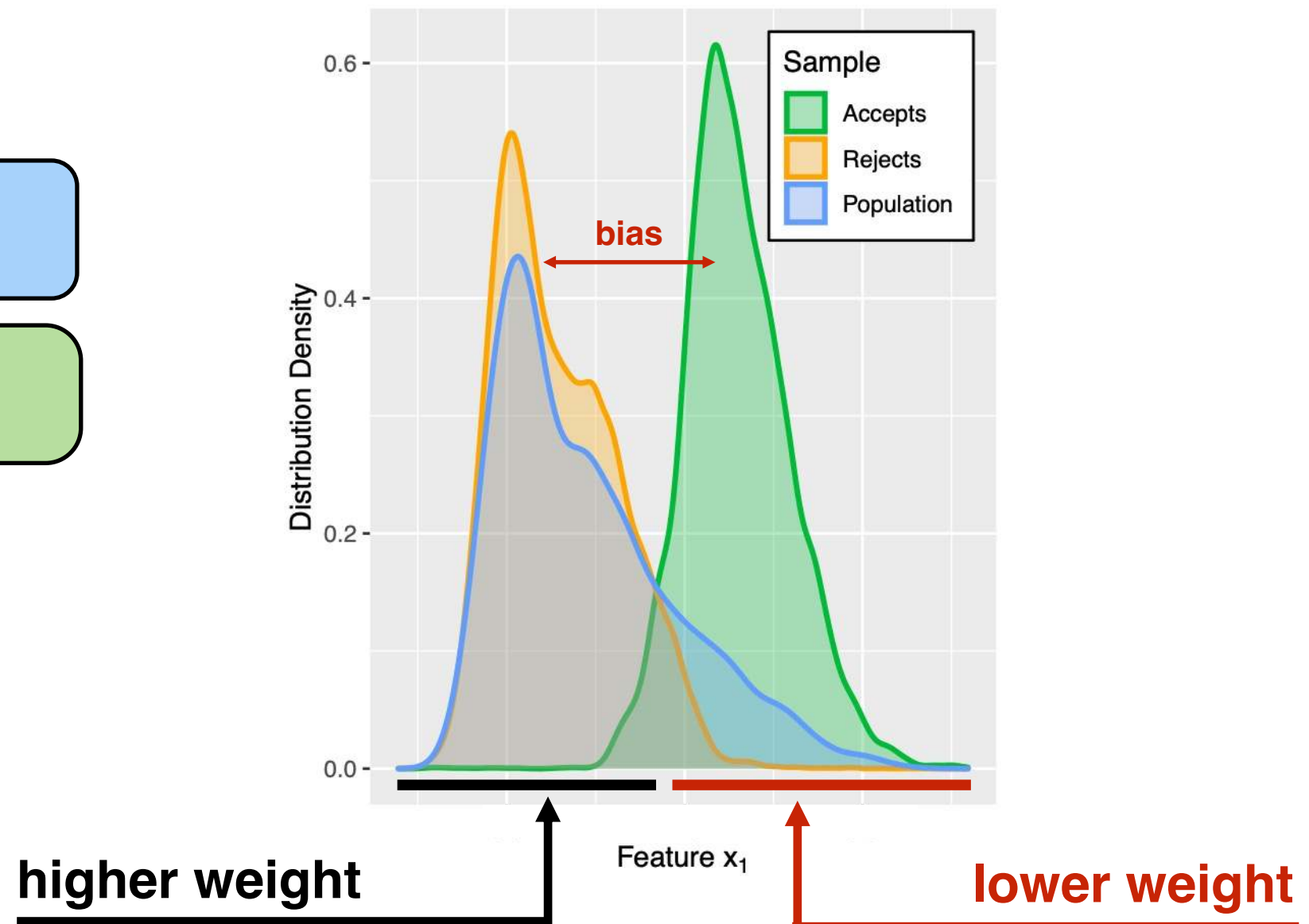
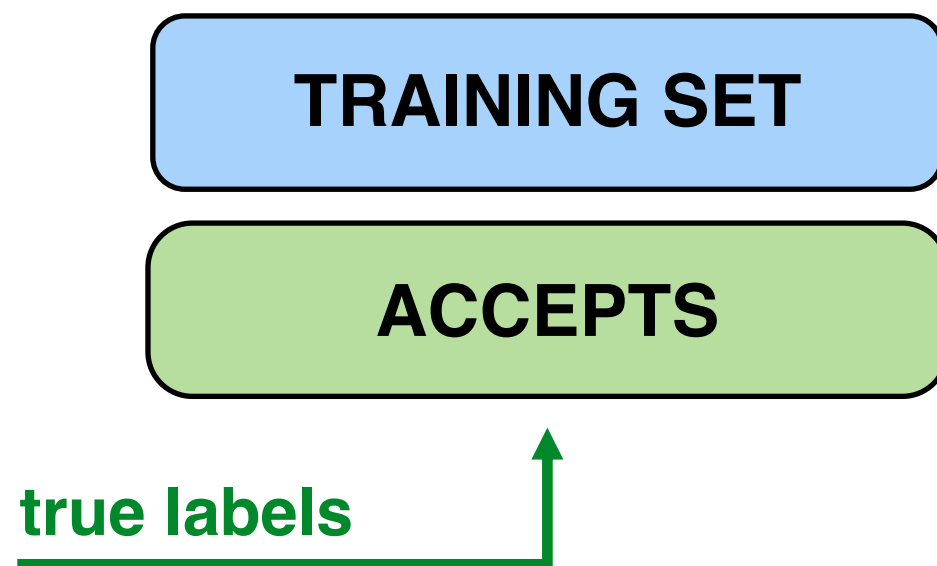
- train model $f(x)$ on training set containing labeled **accepts**



State-of-the-Art: Reweighting

Idea:

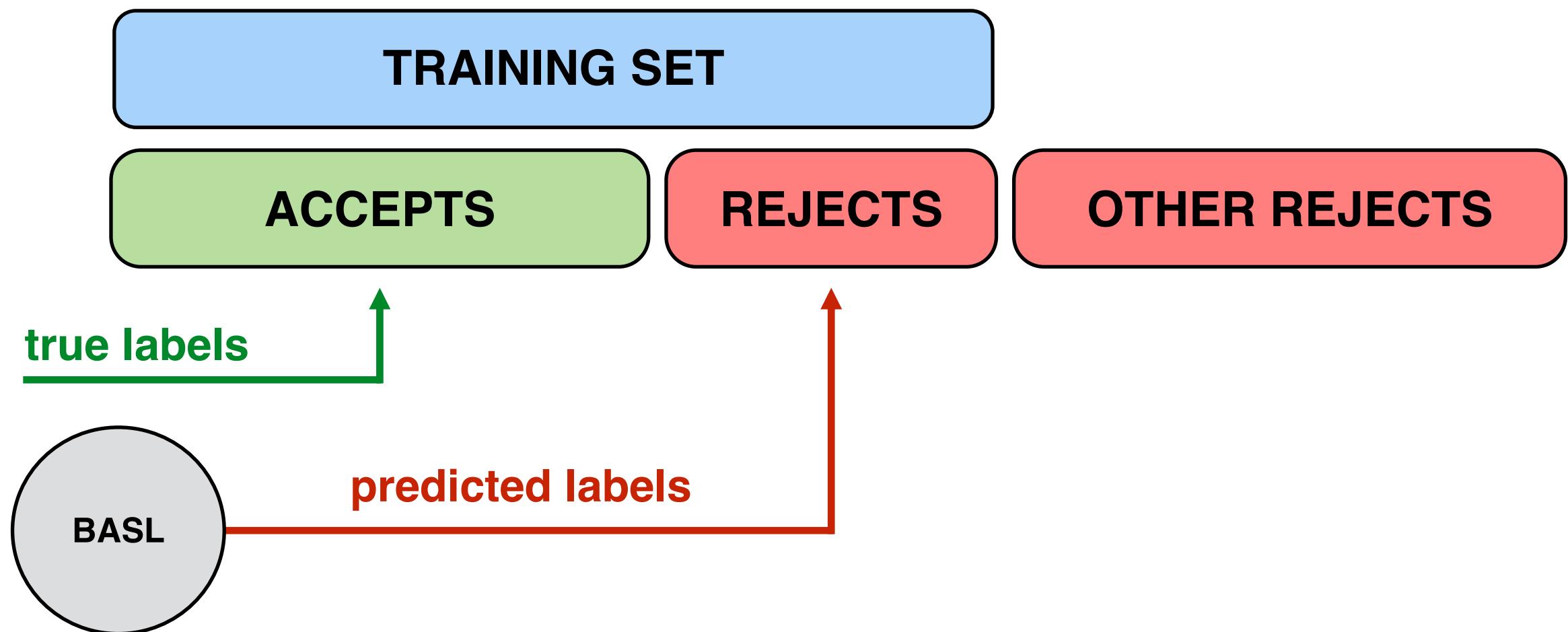
- train model $f(x)$ on training set containing labeled **accepts**
- reweigh model loss to focus on representative cases



Bias-Aware Self-Learning (BASL)

Idea:

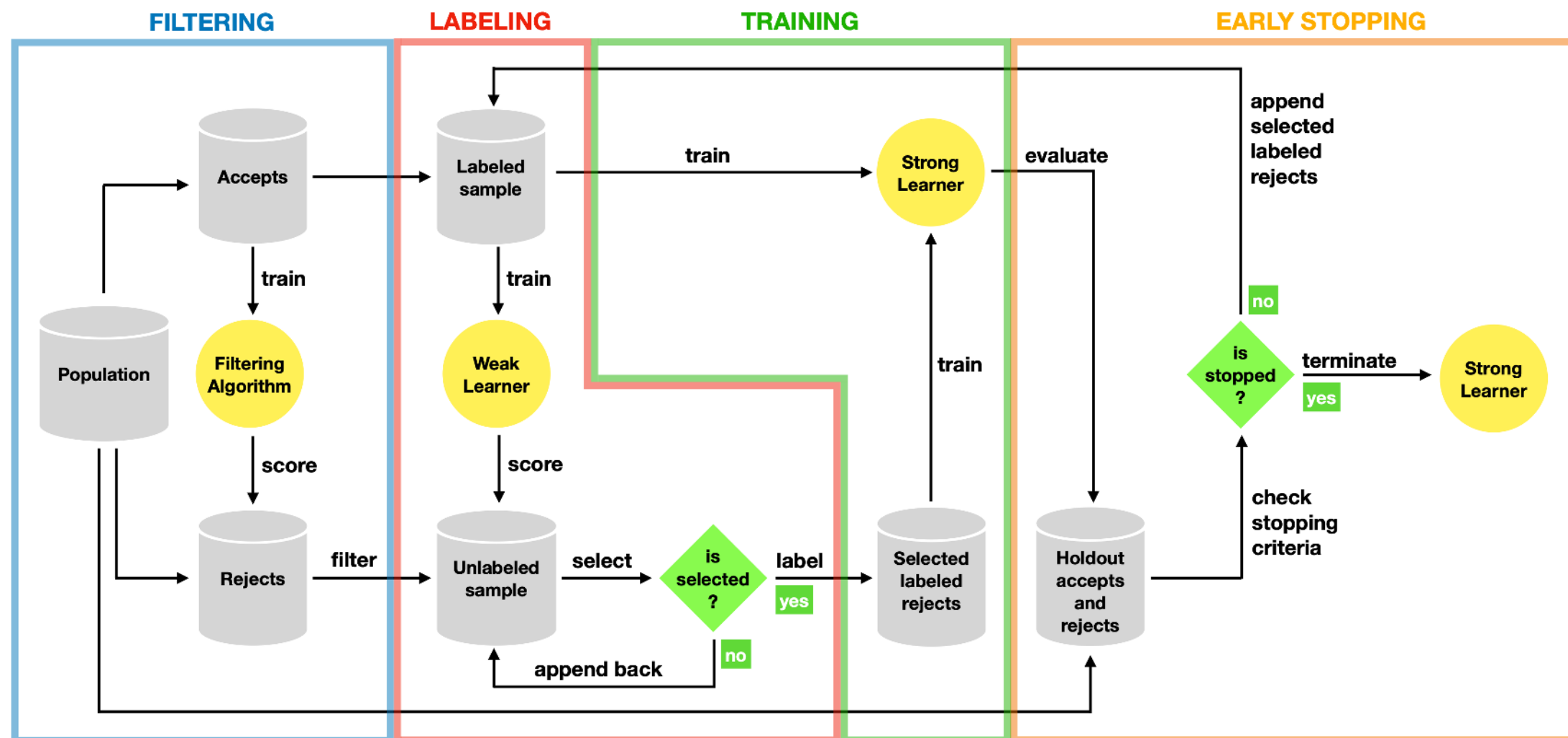
- train model $f(x)$ on augmented training set containing:
 - labeled **accepts**
 - selected pseudo-labeled **rejects**
- use modified self-learning framework (e.g., *Triguero et al. 2013*)
 - implement techniques to reduce the risk of error propagation



Bias-Aware Self-Learning (BASL)

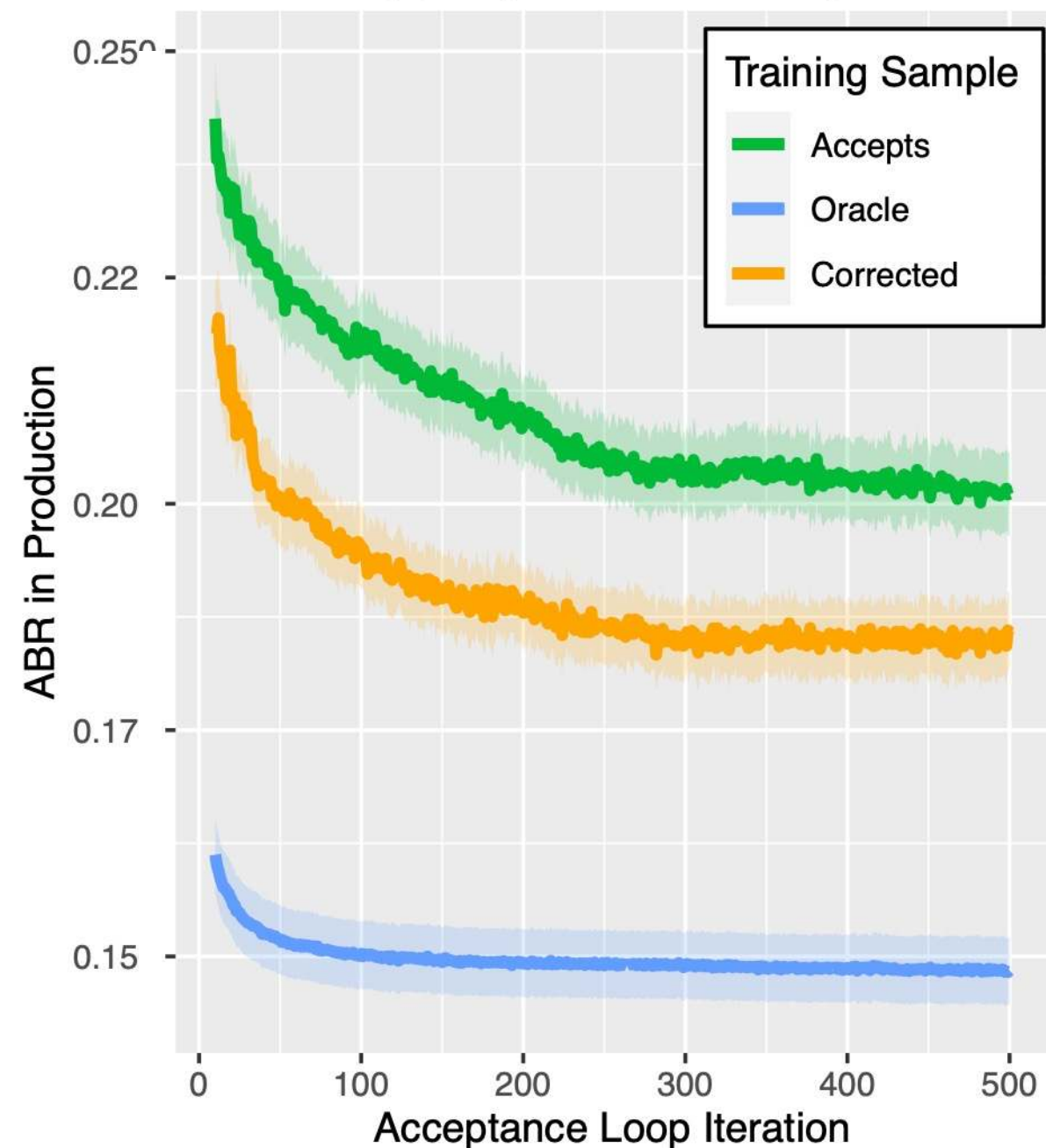
Idea:

- train model $f(x)$ on augmented training set containing:
 - labeled **accepts**
 - selected pseudo-labeled **rejects**
- use modified self-learning framework (e.g., Triguero et al. 2013)
 - implement techniques to reduce the risk of error propagation



BASL: Simulation Results

Performance Dynamics



Aggregated Results

Metric	Loss due to bias	Gains from BASL
ABR	.0547	36.86%
BS	.0404	45.28%
AUC	.0589	48.84%
PAUC	.0488	33.93%

- BASL improves **model performance**
- gains are **statistically significant** at 5%

Presentation Outline

1. Background

- What is credit scoring?
- What are the business goals?

2. Problem Description

- Sampling bias illustration
- Bias impact on ML models

3. Approach

- Improving model evaluation
- Improving model training

4. Results

- Offline evaluation
- Business impact

Offline Evaluation: Experimental Setup

Data description:

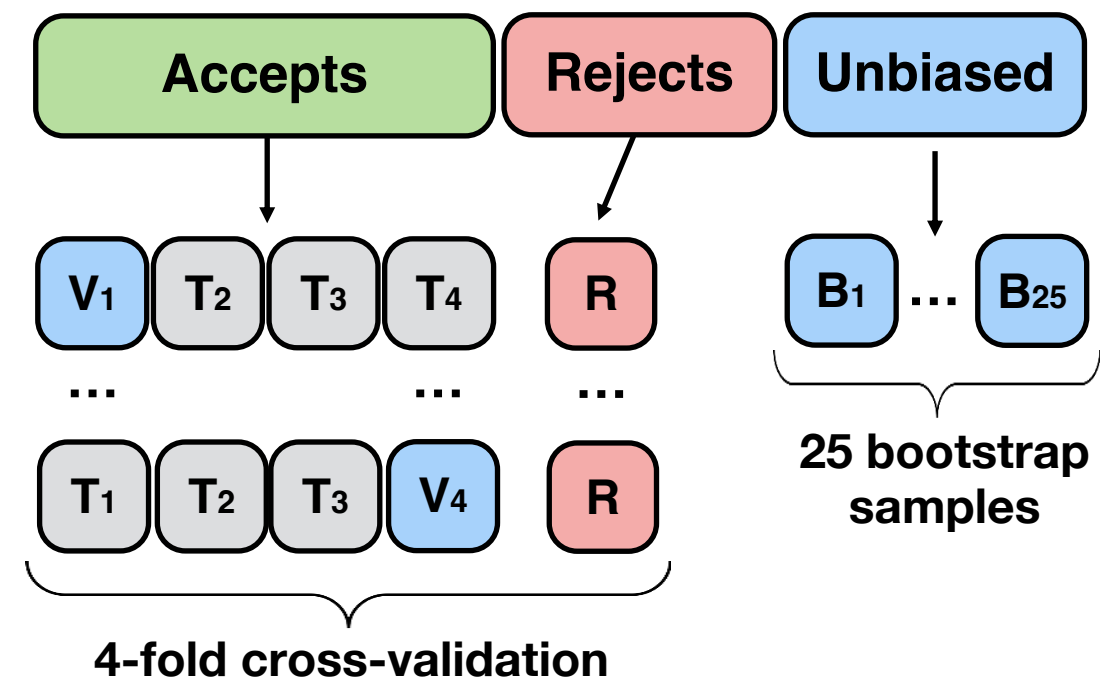
- consumer loans issued by  Monedo in Spain in 2017 - 2019
- contains **labeled accepts** and **unlabeled rejects**
- includes **unbiased sample**: loans from randomized trial

Data summary:

	Accepts	Rejects	Unbiased
No. clients	39,579	18,047	1,967
No. features	2,410	2,410	2,410
BAD* rate	39 %	-	66 %

* missed payments for **3** consecutive months

Data organization:



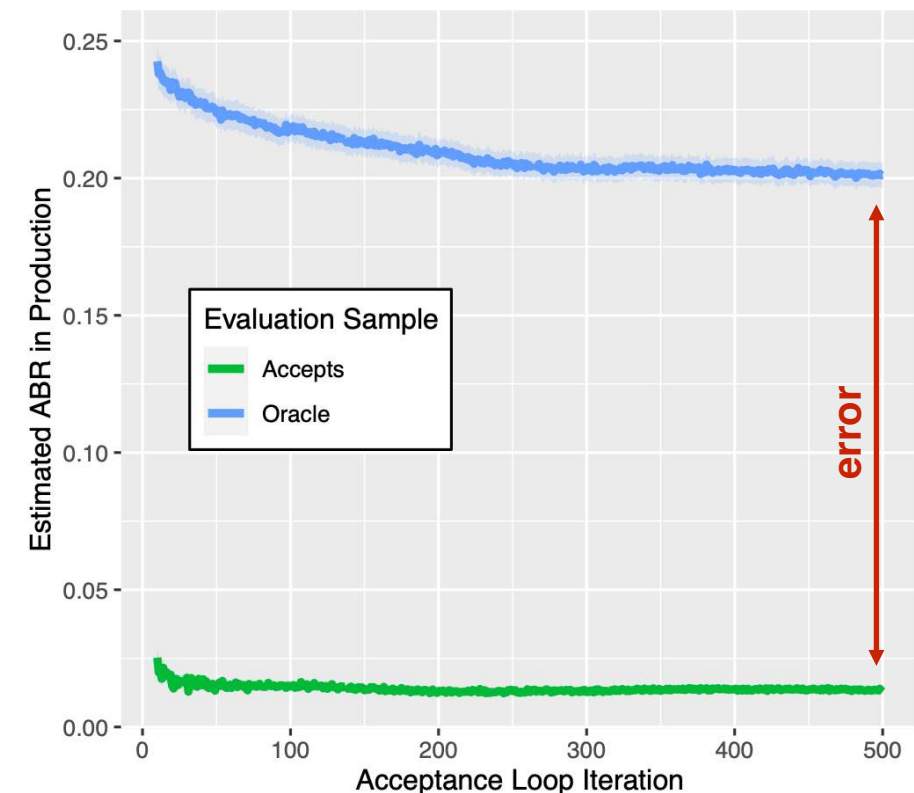
Experiment I: Improving Evaluation

Goal:

- compare accuracy of evaluation methods

Methodology:

- build a scoring model and assess it on **unbiased sample**
 - four evaluation metrics: ABR, BS, AUC, PAUC
- evaluate the same model on historical data
 - Bayesian evaluation
 - benchmarks
- compute RMSE between the two estimates



Experiment II: Results

Evaluation Method	ABR	BS	AUC	PAUC
Standard practice	.0356	.0983	.1234	.0306
Doubly robust evaluation	.1167	.0506	-	-
Reweighting	.0315	.0826	.1277	.0348
Bayesian evaluation	.0130	.0351	.0111	.0073

- **ABR** = BAD rate at 30% acceptance
- **BS** = Brier Score
- **AUC** = area under the ROC curve
- **PAUC** = partial AUC at FNR in [0, 0.2]

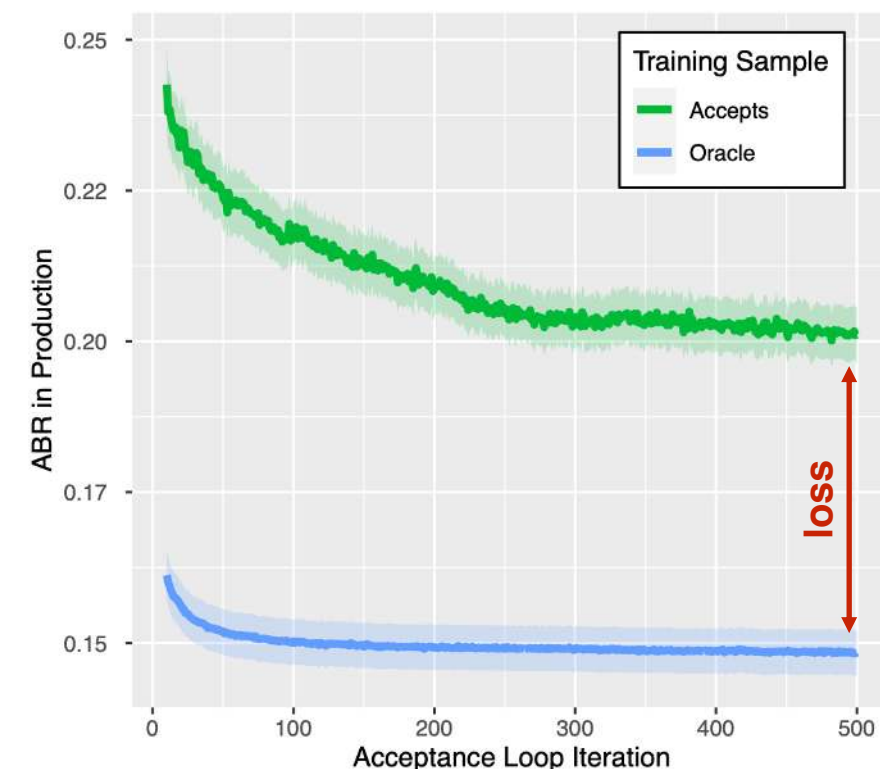
Experiment II: Improving Training

Goal:

- compare performance of bias correction methods

Methodology:

- build a scoring model on **accepts**
- assess performance on **unbiased sample**
 - four evaluation metrics: ABR, BS, AUC, PAUC
- improve the model with bias correction methods
 - BASL
 - benchmarks



Experiment II: Results

Training Method	ABR	BS	AUC	PAUC
Standard practice	.2388	.1819	.7984	.6919
Label all rejects as BAD	.3141	.2347	.6676	.6384
Bias-removing autoencoder	.3061	.2161	.7304	.6373
Heckman model	.3018	.2124	.7444	.6397
Bureau score based labels	.2514	.1860	.7978	.6783
Hard cutoff augmentation	.2458	.1830	.8033	.6790
Parceling	.2396	.1804	.8038	.6885
Reweighting	.2346	.1840	.8040	.6961
Bias-Aware Self-Learning	.2211	.1761	.8166	.7075

- **ABR** = BAD rate at 30% acceptance
- **BS** = Brier Score
- **AUC** = area under the ROC curve
- **PAUC** = partial AUC at FNR in [0, 0.2]

Business Impact: Setup

Parameters:

- acceptance rate
- loan principal
- interest rate

	Micro loans	Installment loans
Acceptance rate α	[20%, 40%]	[10%, 20%]
Loan principal A	\$375 (SD = \$100)	\$17,100 (SD = \$1,000)
Total interest i	17.33% (SD = 1%)	10.36% (SD = 1%)

Two markets:

- micro-loans
- installment loans

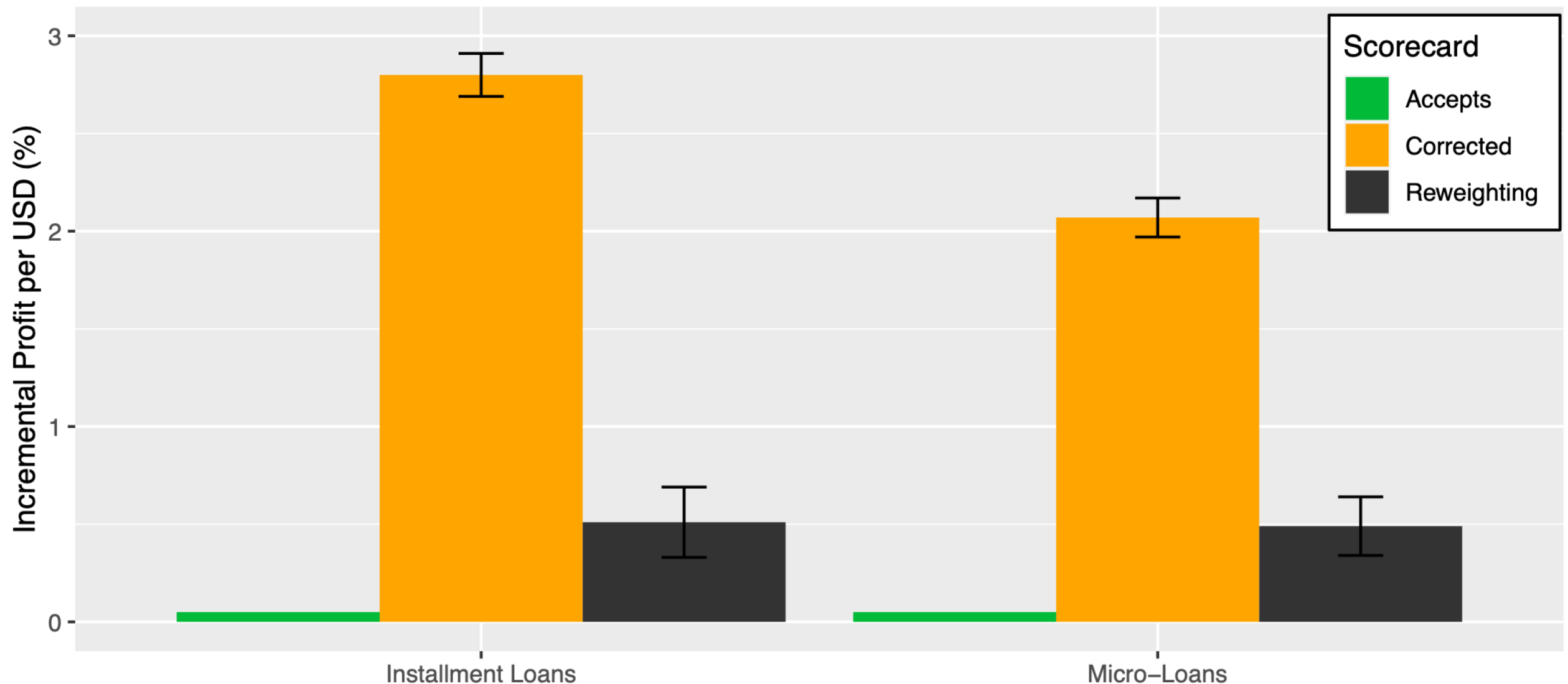
Calculations:

- average profit per loan for each algorithm:

$$\pi = \frac{1}{100} \sum_{j=1}^{100} \left[\underbrace{(1 - ABR_j) \times A \times (1 + i)}_{\text{GOOD clients}} - \underbrace{ABR_j \times A \times (1 + i)}_{\text{BAD clients}} - A \right]$$

- averaging over 100 values (4-fold CV x 25 bootstrap samples)

Business Impact: Results

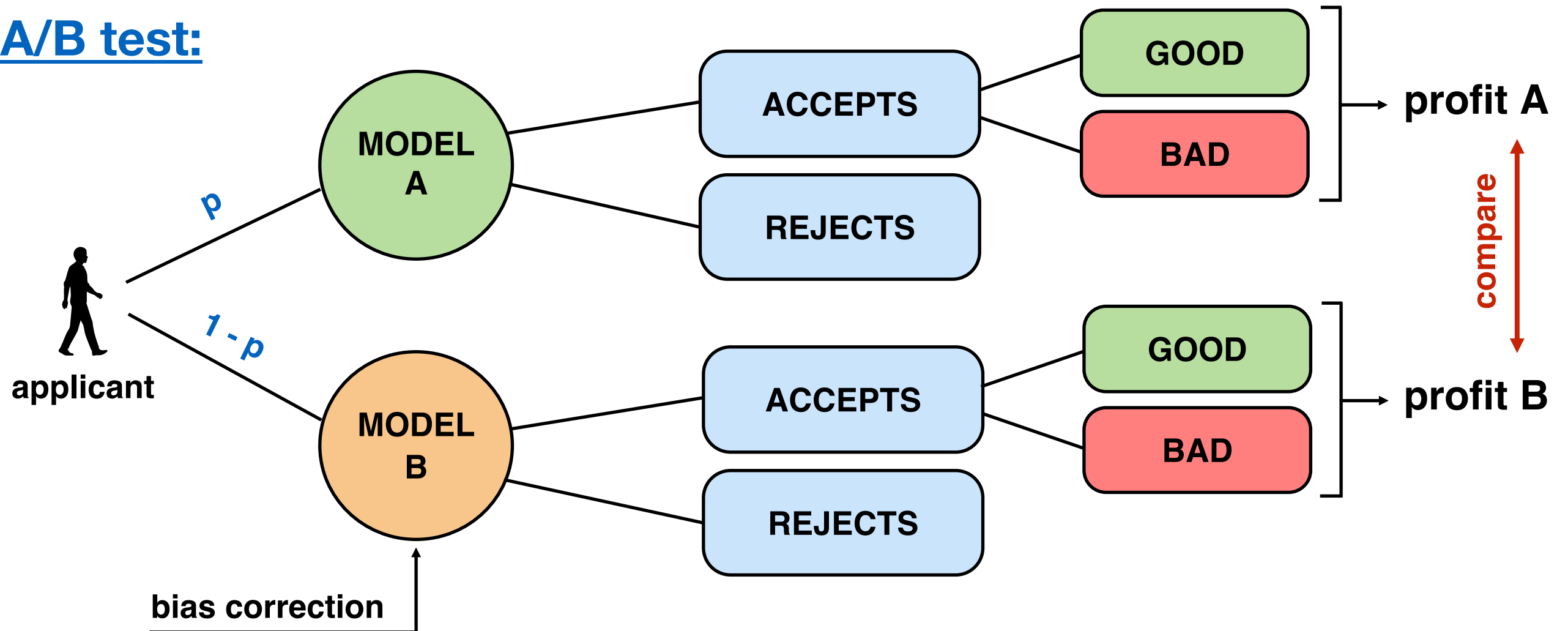


Incremental gains:

- installment loans: up to **\$461.70** per loan
- micro-loans: up to **\$7.78** per loan

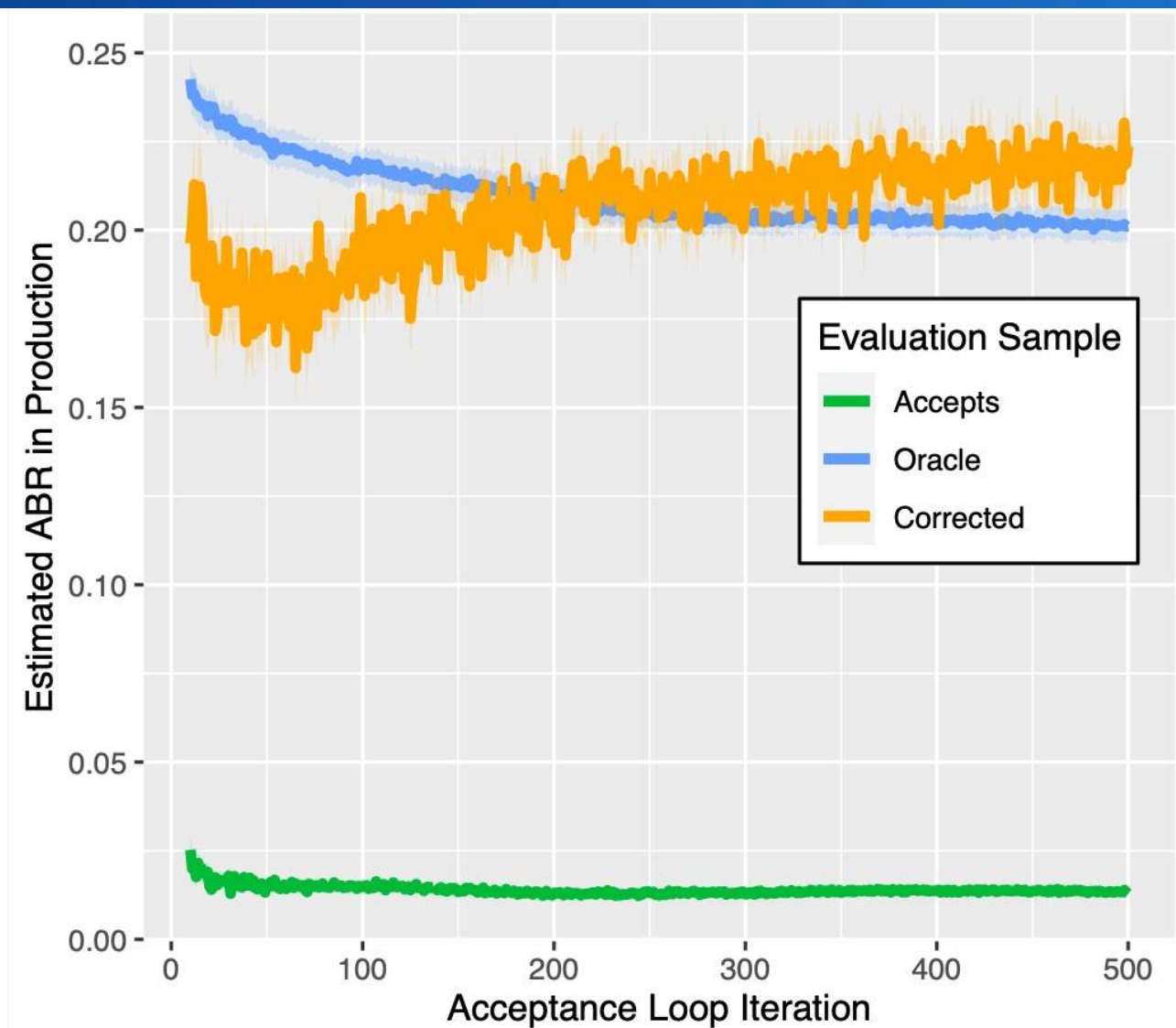
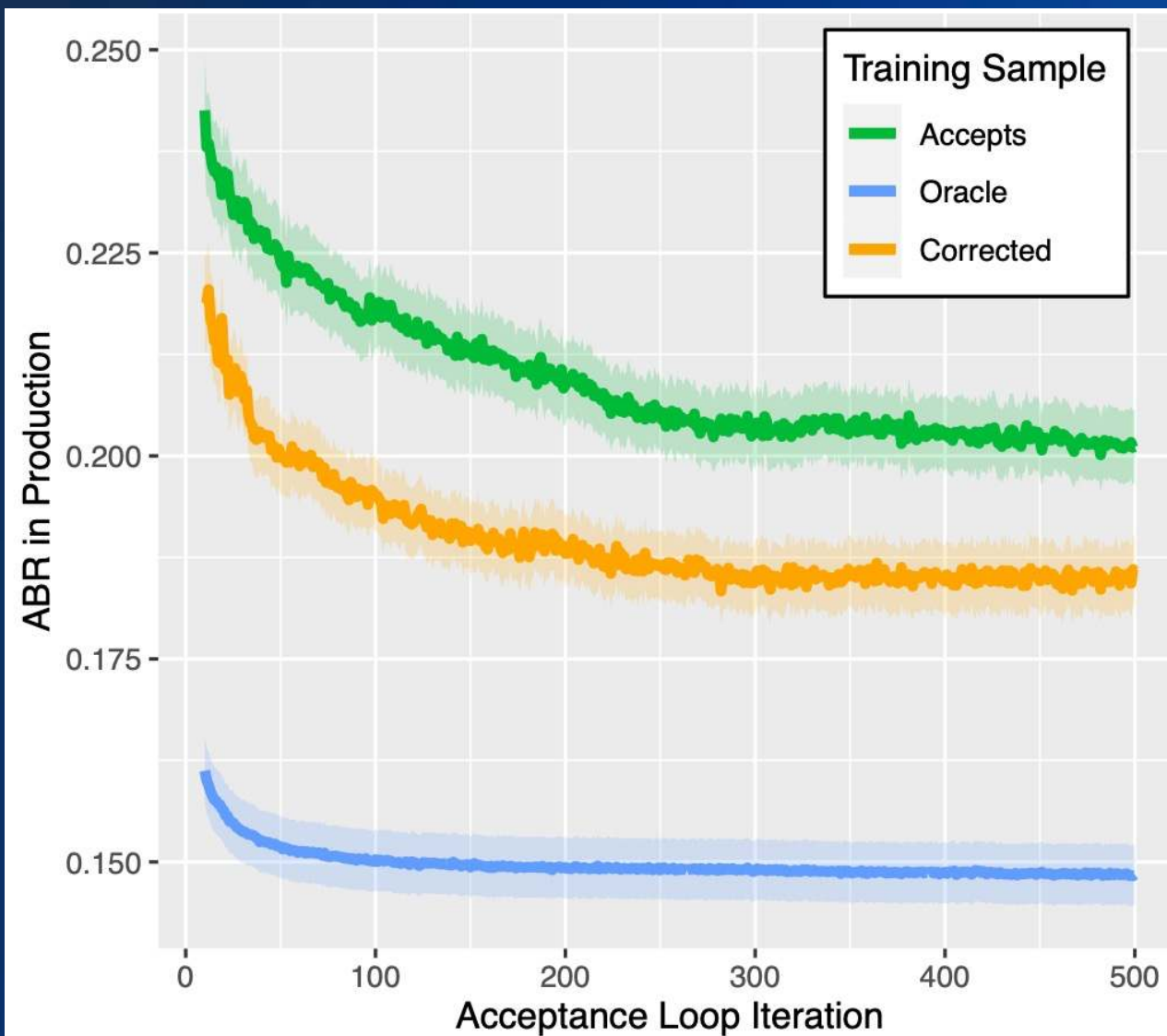
From Offline to Online

A/B test:



Challenges:

- long delay before observing the metrics
- regulations regarding data on rejected clients



Incremental gains:

- installment loans: up to **\$461.70** per loan
- micro-loans: up to **\$7.78** per loan