# Active Learning for Reject Inference in Credit Scoring

**Nikita Kozodoi, Stefan Lessmann, Samantha Sizemore**

# Presentation Outline

## 1. Sampling Bias in Credit Scoring

- Problem setup & illustration
- Impact on scoring models

## 2. Correcting Sampling Bias

- Offline reject inference
- Active learning for online reject inference

## 3. Empirical Results

- Experimental setup
- Preliminary results

# Presentation Outline

## 1. Sampling Bias in Credit Scoring

- Problem setup & illustration
- Impact on scoring models

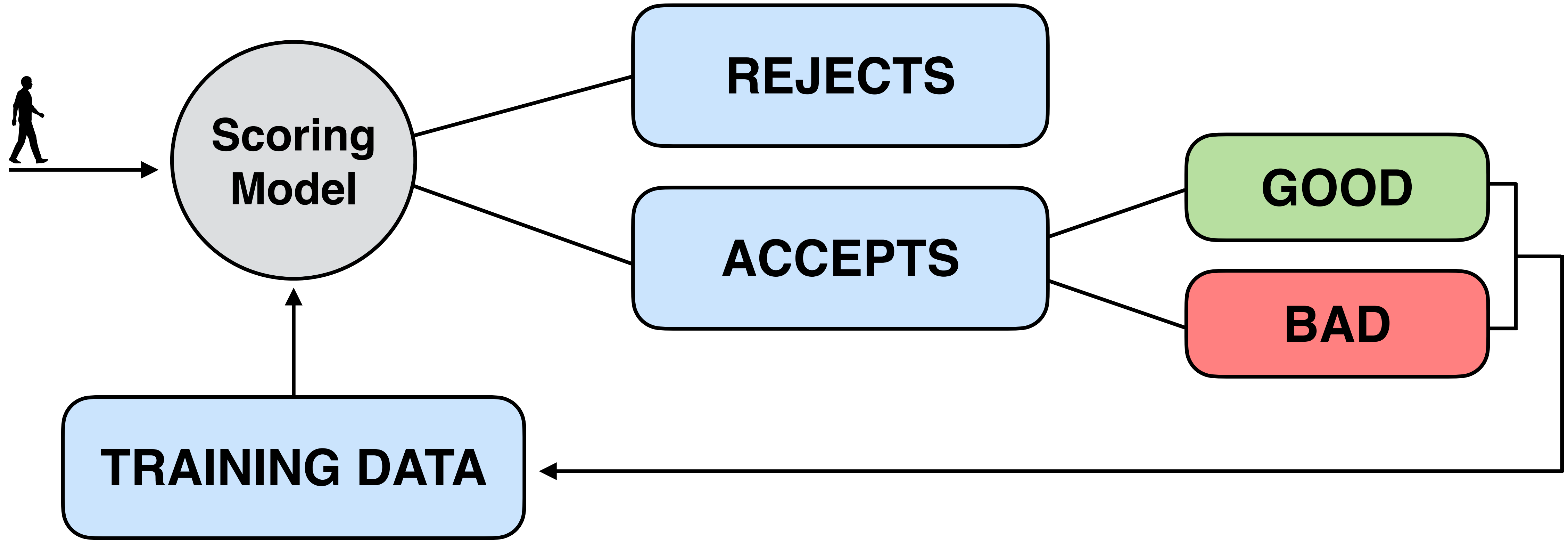
## 2. Correcting Sampling Bias

- Offline reject inference
- Active learning for online reject inference

## 3. Empirical Results

- Experimental setup
- Preliminary results

# Acceptance Loop in Credit Scoring

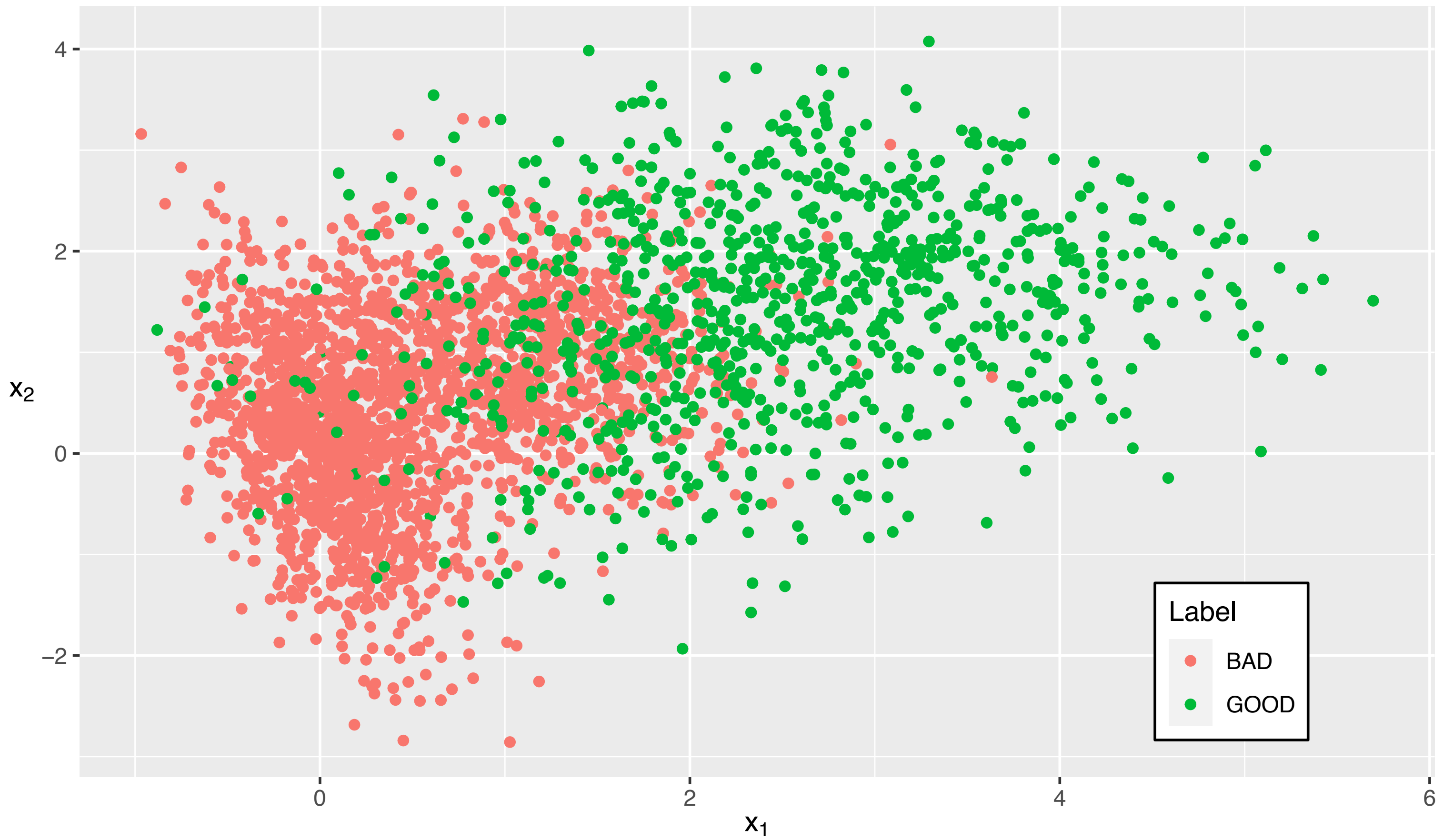



- **scoring model filters incoming loan applications**
  - ML model observes features of incoming applicants
  - predicts whether an applicant will repay the loan
- **training a model requires data with known outcomes**
  - repayment outcome is only observed for **accepted applicants**
  - application labels are missing **not completely at random**
- **acceptance loop creates sampling bias**
  - adverse impact of bias depends on the missingness type

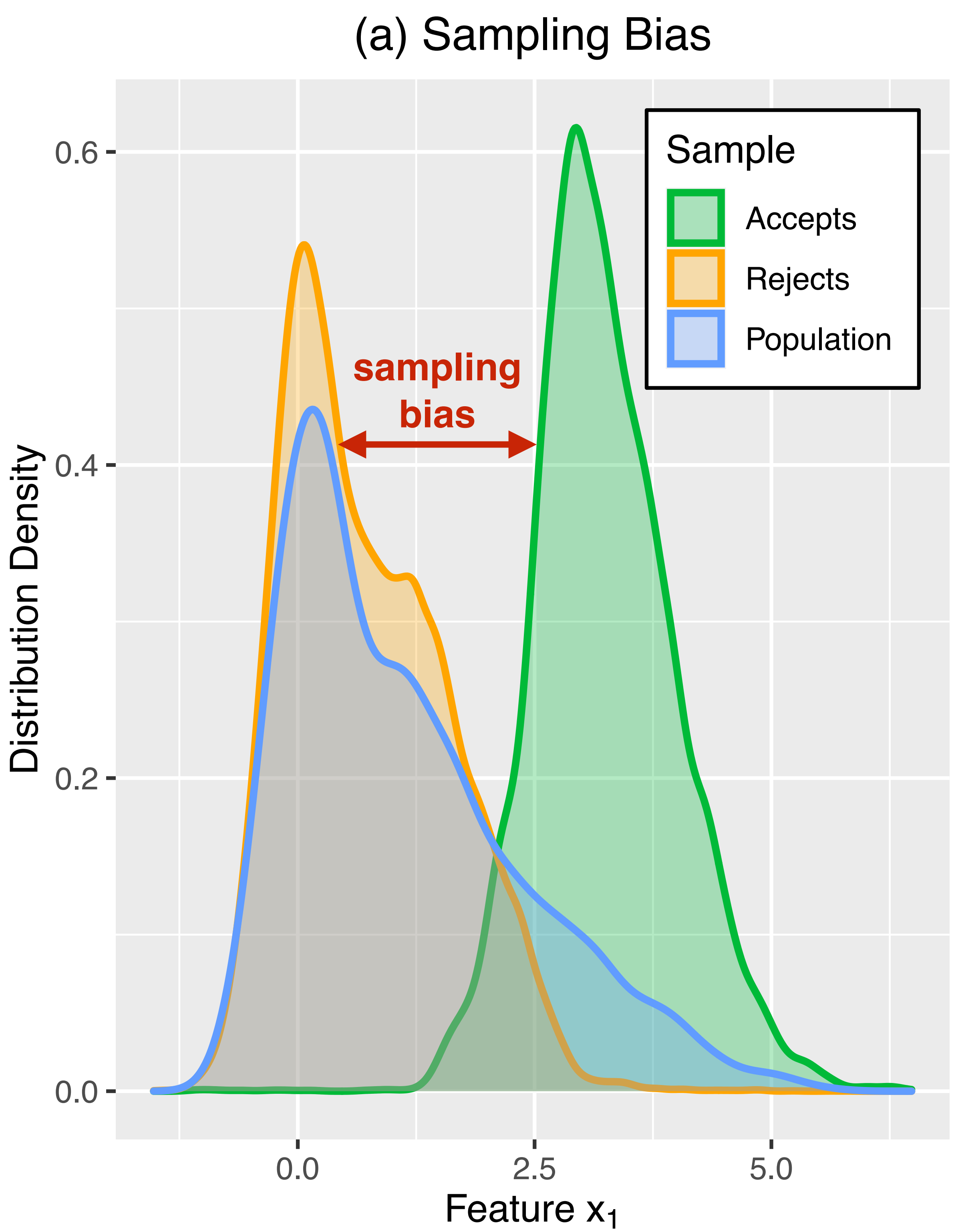
# Sampling Bias Illustration [1/3]

## Synthetic data:

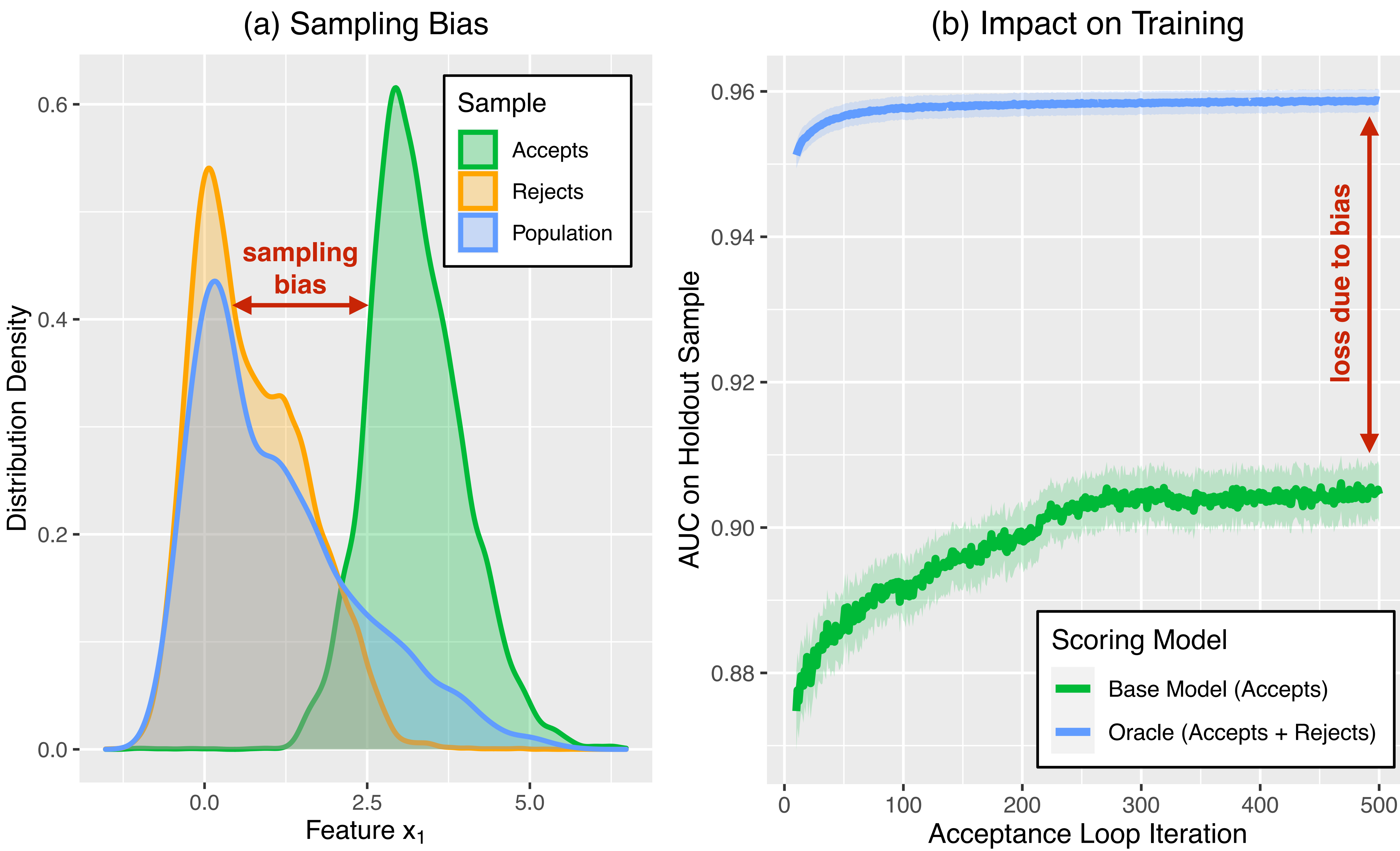
- sampling **GOOD** and **BAD** risks from multivariate Gaussian mixtures
- simulating real-world **acceptance loop:**
  - iteratively generating **batches** of new applications
  - using a scoring model to **accept** and **reject** new applications
  - **updating** the model after learning the labels of **accepts**
- evaluating performance on a **holdout sample** from population



# Sampling Bias Illustration [2/3]



# Sampling Bias Illustration [3/3]




**AUC** = area under the ROC curve; higher is better

# Presentation Outline

1. Sampling Bias in Credit Scoring
   - Problem setup & illustration
   - Impact on scoring models

2. **Correcting Sampling Bias**
   - Offline reject inference
   - Active learning for online reject inference

3. Empirical Results
   - Experimental setup
   - Preliminary results

# Background on Reject Inference [1/2]

**Reject inference mitigates sampling bias by using data on rejects**

- label **rejects** using one of the RI techniques
- train a scoring model on the augmented data
- **examples:** hard cutoff augmentation, parceling, Heckman model

# Background on Reject Inference [2/2]

**Reject inference mitigates sampling bias by using data on rejects**

- label **rejects** using one of the RI techniques
- train a scoring model on the augmented data
- **examples:** hard cutoff augmentation, parceling, Heckman model

## Hard cutoff augmentation (HCA):

- train a scoring model over **accepts**
- predict **P(BAD)** for **rejects** using this model
- assign labels based on a certain threshold

## Parceling:

- split **rejects** into groups based on the model score
- assign labels within groups proportionally to the expected **BAD** rate
- **BAD** rate for **rejects** is usually assumed to be higher than for **accepts**

# Offline vs Online Reject Inference [1/2]

- **traditional reject inference methods are offline**
  - sampling bias is mitigated by working with past **rejects**
- **offline reject inference has limitations**
  - actual labels of the **rejects** are never observed
  - rejects become less relevant with dataset shift (e.g., business cycle)
  - regulation may prohibit using data on rejected customers

# Offline vs Online Reject Inference [2/2]

- **traditional reject inference methods are <u>offline</u>**
  - sampling bias is mitigated by working with past **rejects**
- **<u>offline</u> reject inference has limitations**
  - actual labels of the **rejects** are never observed
  - rejects become less relevant with dataset shift (e.g., business cycle)
  - regulation may prohibit using data on rejected customers
- **we propose <u>online</u> reject inference with active learning (AL)**
  - working with applications about to be rejected by a scorecard
  - issuing a loan to selected **rejects** to learn the actual labels
- **<u>online</u> reject inference stands on the cost-benefit trade-off**
  - cost from issuing loans to risky customers
  - gain from obtaining a more representative training data

# What is Active Learning? [1/4]

**ML framework in which a learning algorithm interactively queries to label currently unlabeled data points**

# What is Active Learning? [2/4]

**ML framework in which a learning algorithm interactively queries to label currently unlabeled data points**

- consider a classification task with labeled and unlabeled data

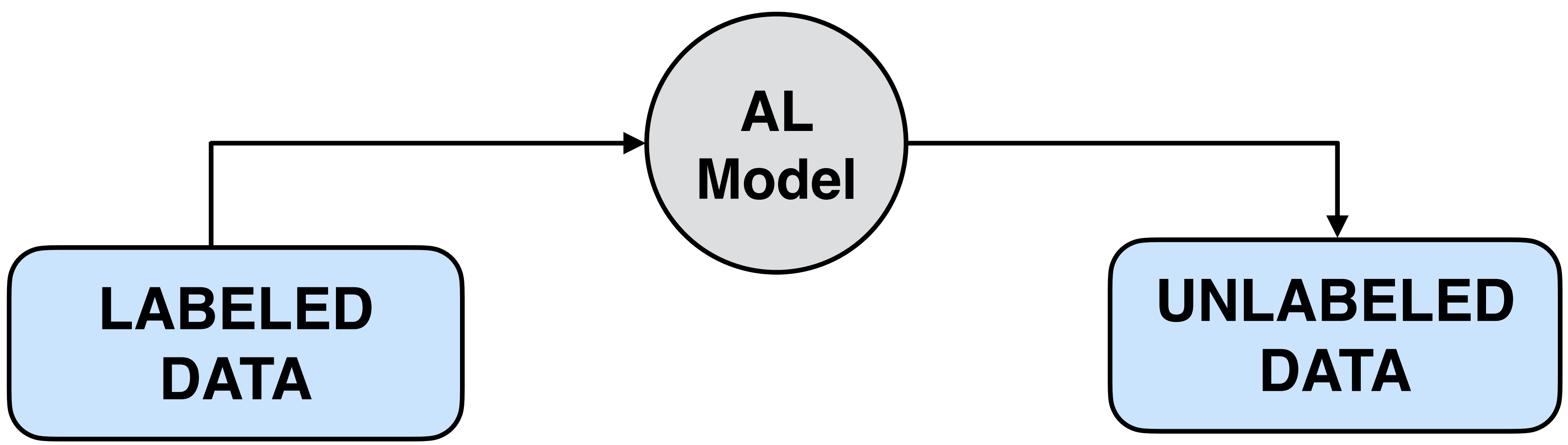
LABELED DATA

UNLABELED DATA

# What is Active Learning? [3/4]

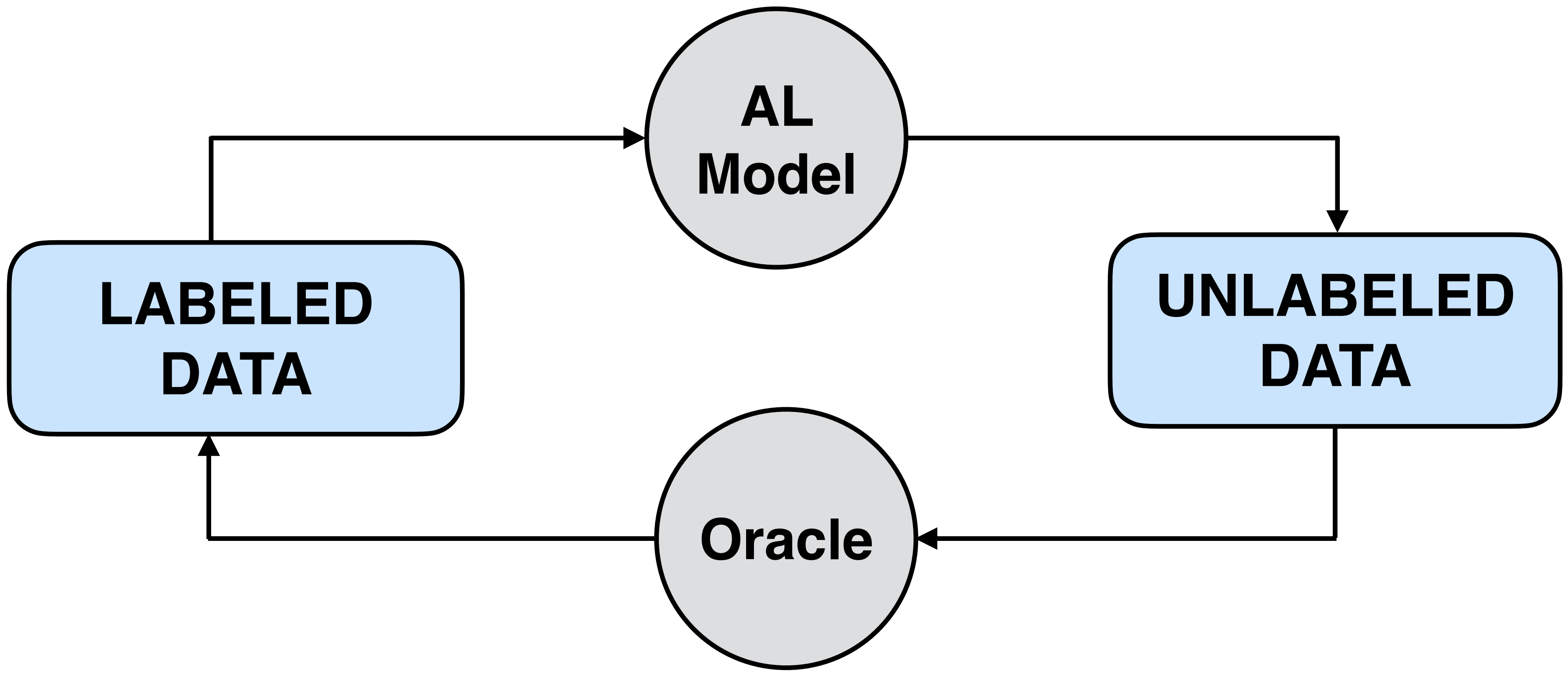
**ML framework in which a learning algorithm interactively queries to label currently unlabeled data points**

- consider a classification task with labeled and unlabeled data
- AL identifies **“most interesting”** unlabeled data points
  - which observations would improve classifier performance if they had labels?
  - can be measured as **uncertainty**, **correlation**, **expected error decrease**, etc.

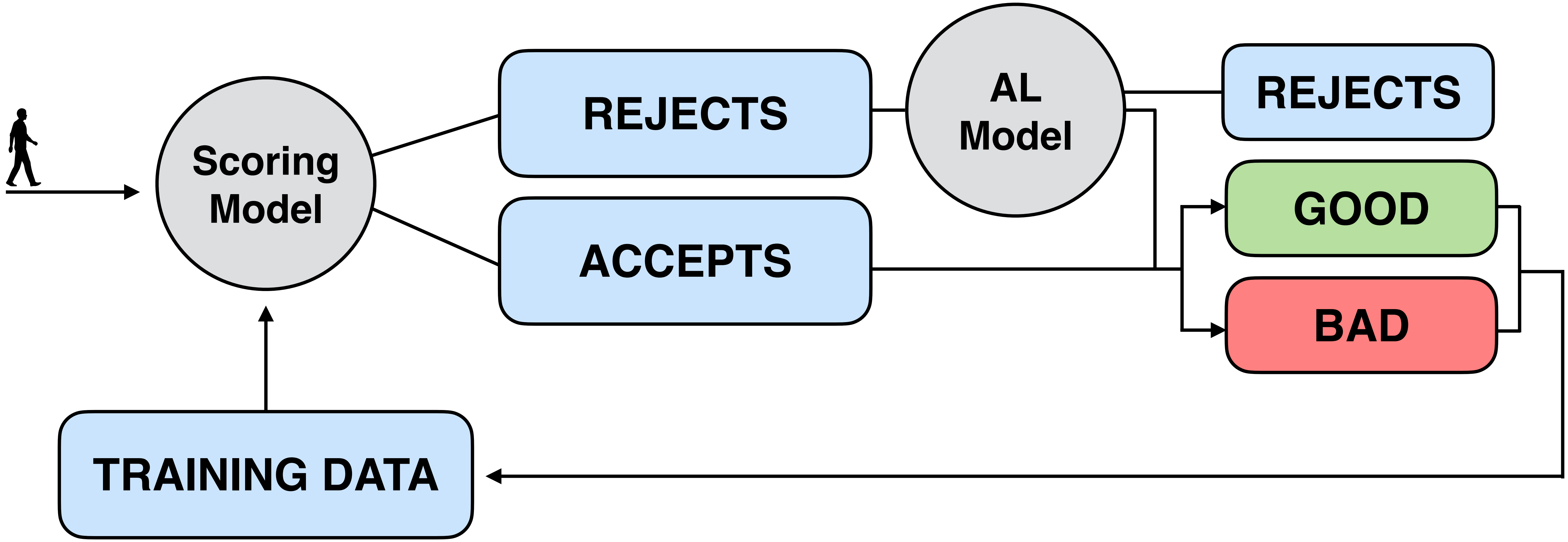


**ML framework in which a learning algorithm interactively queries to label currently unlabeled data points**

- consider a classification task with labeled and unlabeled data
- AL identifies **"most interesting"** unlabeled data points
  - which observations would improve classifier performance if they had labels?
  - can be measured as **uncertainty**, **correlation**, **expected error decrease**, etc.
- identified data points are labeled by oracle
- the classifier is trained on augmented data



# Acceptance Loop with AL




- **scoring model filters incoming loan applications**
  - **ML model** observes features of incoming applicants
  - predicts whether an applicant will repay the loan
- **active learning selects additional cases rejected by a scorecard**
  - **AL model** observes features of rejects and scorecard predictions
  - predicts whether an applicant will be «useful»

# Selected AL Techniques

## Uncertainty sampling:

- selects observations that the ML model is **least confident about**
- e.g., cases with predicted **P(BAD)** close to 0.5

## Query-by-committee (QBC):

- trains a set (committee) of ML models (e.g., on different training folds)
- selects observations where the committee **disagrees the most**
- e.g., cases with the highest Kullback-Leibler divergence over predictions

## Optimized probabilistic active learning (OPAL):

- measures **«spatial usefulness»** of an unlabeled observation
- selects observations that maximize the expected reduction in (asymmetric) misclassification cost
- e.g., cases from high-density areas with potentially higher error costs

# Presentation Outline

## 1. Sampling Bias Problem

- Problem setup & illustration
- Impact on scoring models

## 2. Correcting Sampling Bias

- Traditional «offline» reject inference
- Active learning for «online» reject inference

## 3. Empirical Results

- Experimental setup
- Preliminary results

# Data Summary

## Real data:

- consumer credit scoring data provided by LendingClub
- repayment behavior of actual **rejects** is not available
- treating most risky **accepts** as «**rejects**»

## Synthetic data:

- full control over the data generation process
- repayment behavior of both **accepts** and **rejects** is available

| Data set | Observations | Features | BAD rate |
|---|---|---|---|
| **LendingClub** | 100,000 | 17 | 8 % |
| **Synthetic Data** | 50,000 | 19 | 40 % |

# Experimental Setup

## Acceptance loop:

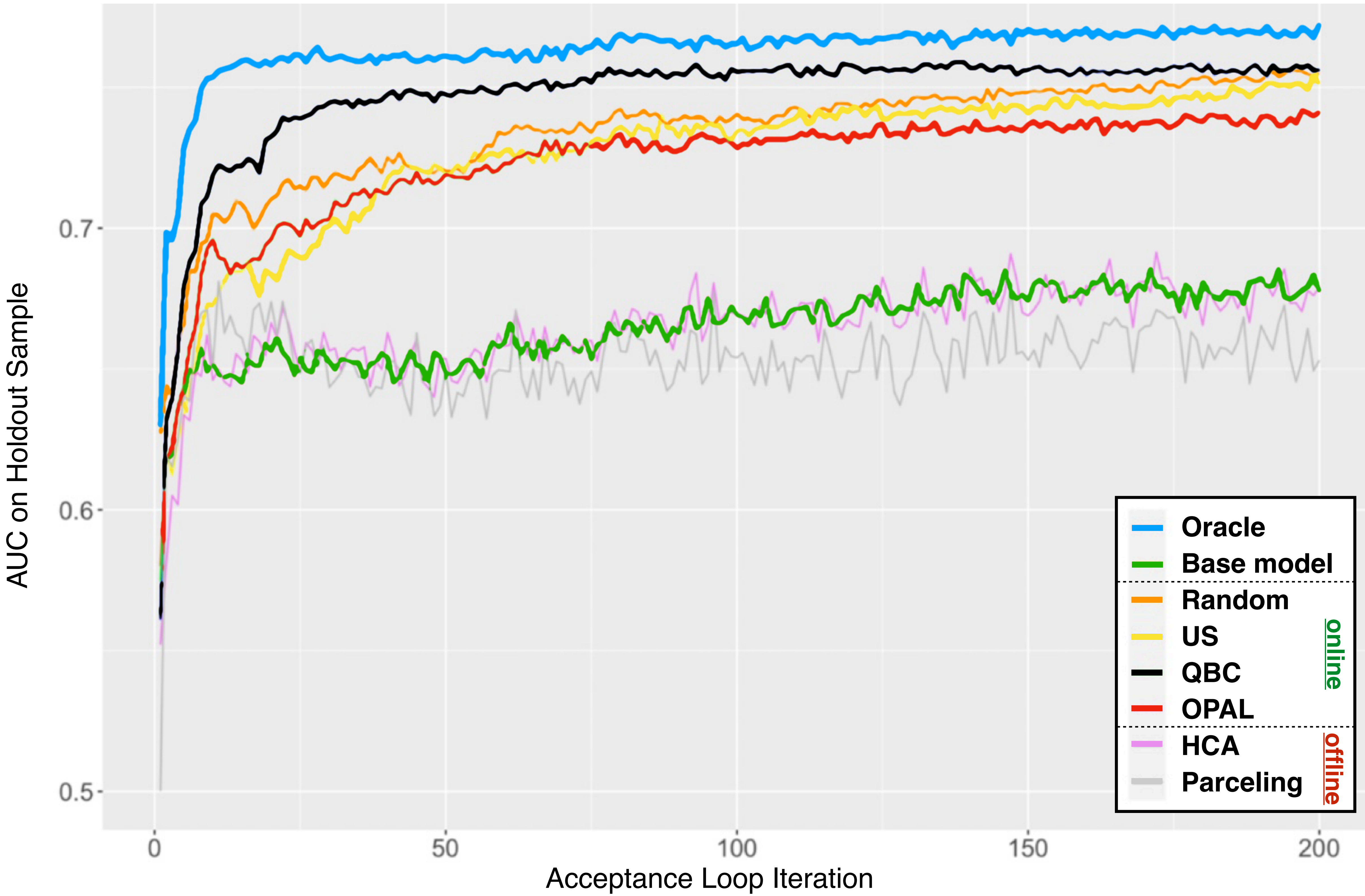
- draw / generate a batch of **new applications**
- **accept a subset** of loan applications
  - select 20% low-risk cases with **ML model**
  - select 10% «useful» cases with **AL model**
- **augment training data** with labeled accepts
- **retrain the scoring model** on new data
- **evaluate** performance on a holdout sample

**repeat for 200 iterations**

## Performance evaluation:

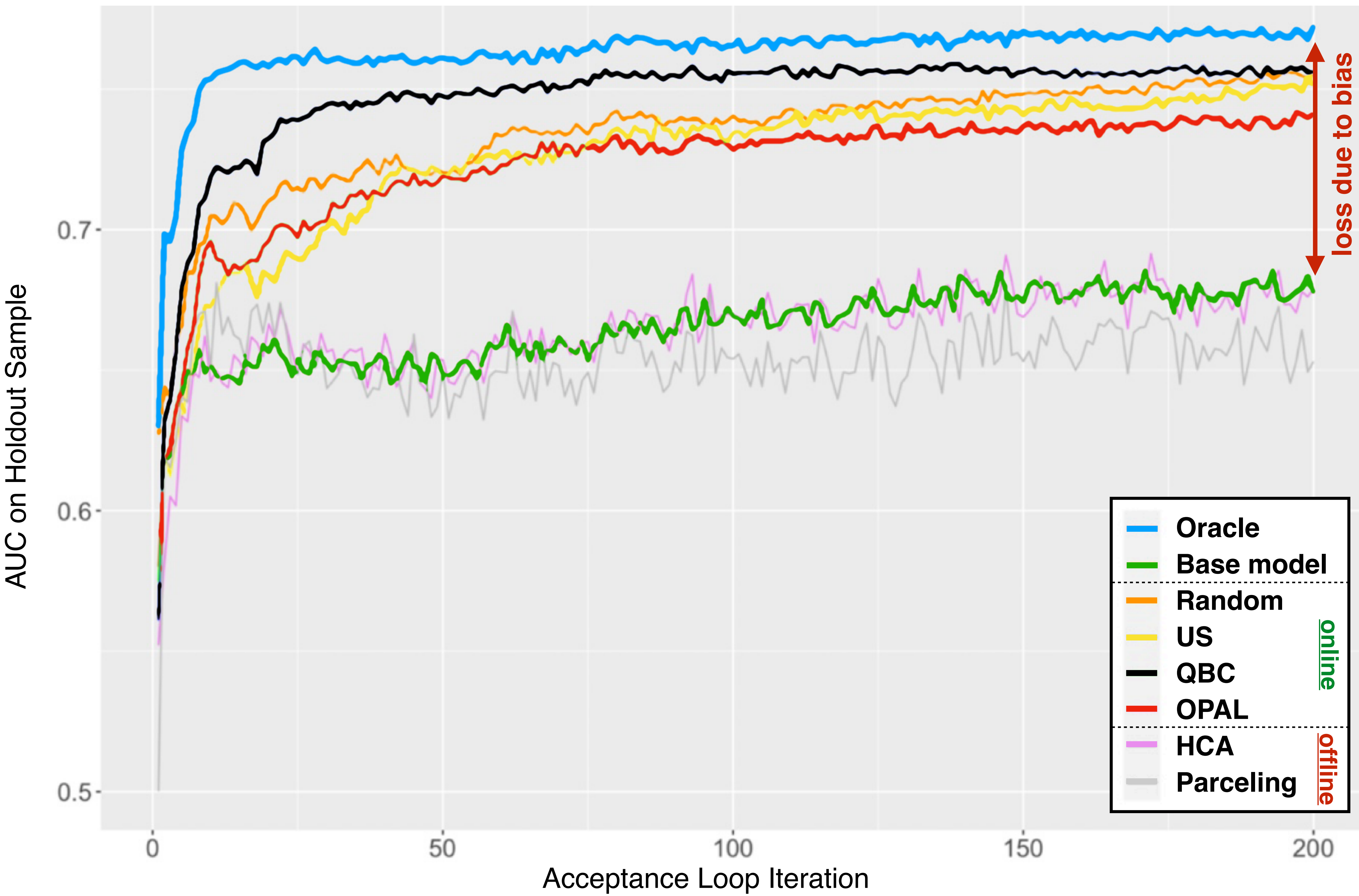
- **two cost / benefit components compared to base model:**
  - **model performance:** improved accuracy of the retrained **ML model**
  - **data augmentation:** accepting extra applicants with the **AL model**

# Results: LendingClub [1/3]



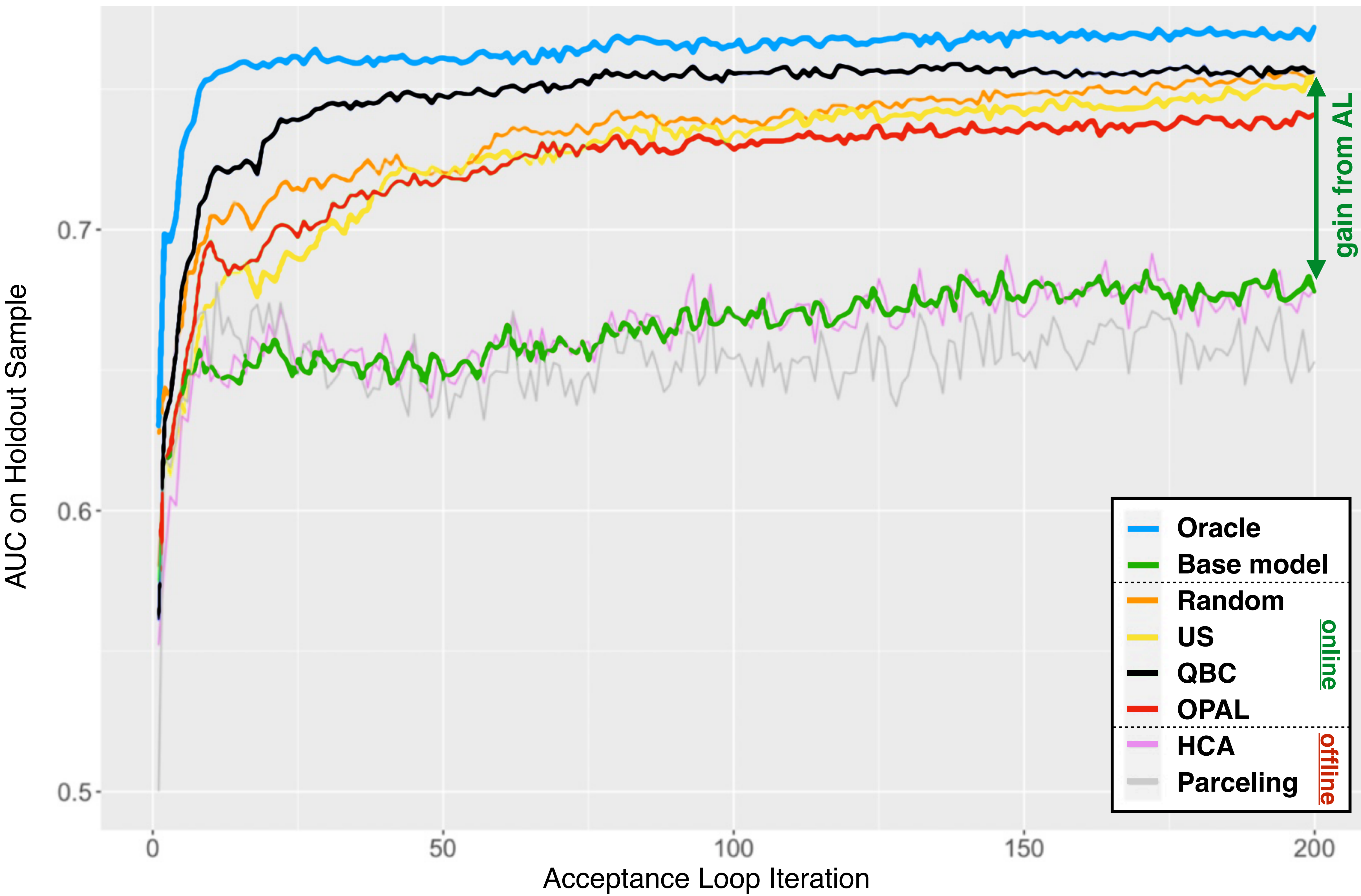

**AUC** = area under the ROC curve; higher is better

# Results: LendingClub [2/3]




**AUC** = area under the ROC curve; higher is better

# Results: LendingClub [3/3]




**AUC** = area under the ROC curve; higher is better

# Results: Model Performance [1/2]

**LendingClub Dataset**

| Method | AUC gain | BS gain | ABR gain |
|---|---|---|---|
| **Random** | .069 | .101 | .057 |
| **US** | | | |
| **QBC** | | | |
| **OPAL** | | | |
| **Oracle** | .098 | .107 | .405 |

**Synthetic Dataset**

| Method | AUC gain | BS gain | ABR gain |
|---|---|---|---|
| **Random** | .025 | .009 | .897 |
| **US** | | | |
| **QBC** | | | |
| **OPAL** | | | |
| **Oracle** | .037 | .013 | 1.380 |

- average gains per iteration in area under the learning curve relative to base model
- positive numbers indicate improvement over the base model

**AUC** = area under the ROC curve; **BS** = Brier score; **ABR** = BAD rate then accepting top-20% applicants

# Results: Model Performance [2/2]

**LendingClub Dataset**

| Method | AUC gain | BS gain | ABR gain |
|---|---|---|---|
| **Random** | .069 | .101 | .057 |
| **US** | .061 | **.103** | .245 |
| **QBC** | **.082** | .097 | **.264** |
| **OPAL** | .058 | .094 | .172 |
| **Oracle** | .098 | .107 | .405 |

**Synthetic Dataset**

| Method | AUC gain | BS gain | ABR gain |
|---|---|---|---|
| **Random** | .025 | **.009** | **.897** |
| **US** | .026 | **.009** | .798 |
| **QBC** | **.027** | .008 | .830 |
| **OPAL** | .025 | .008 | .857 |
| **Oracle** | .037 | .013 | 1.380 |

- average gains per iteration in area under the learning curve relative to base model
- positive numbers indicate improvement over the base model

**AUC** = area under the ROC curve; **BS** = Brier score; **ABR** = BAD rate then accepting top-20% applicants

# Results: Overall Profit [1/2]

**LendingClub Dataset**

| Method | Data profit | Model profit | Total profit |
|---|---|---|---|
| Random | | | |
| US | | | |
| QBC | | | |
| OPAL | | | |
| Oracle | 1.271 | .002 | 1.272 |

**Synthetic Dataset**

| Method | Data profit | Model profit | Total profit |
|---|---|---|---|
| Random | | | |
| US | | | |
| QBC | | | |
| OPAL | | | |
| Oracle | -1.068 | .005 | -1.062 |

- **data profit** = profit from assigning loans to applicants selected with AL
- **model profit** = profit from model improvement after data augmentation
- values represent average profit per EUR issued

# Results: Overall Profit [2/2]

**LendingClub Dataset**

| Method | Data profit | Model profit | Total profit |
|---|---|---|---|
| Random | .124 | .000 | .125 |
| US | .132 | **.001** | .133 |
| QBC | **.154** | **.001** | **.155** |
| OPAL | .095 | **.000** | .096 |
| Oracle | 1.271 | .002 | 1.272 |

**Synthetic Dataset**

| Method | Data profit | Model profit | Total profit |
|---|---|---|---|
| Random | -.098 | .002 | -.095 |
| US | -.115 | **.003** | -.112 |
| QBC | **-.040** | **.003** | **-.036** |
| OPAL | -.167 | **.003** | -.163 |
| Oracle | -1.068 | .005 | -1.062 |

- **data profit** = profit from assigning loans to applicants selected with AL
- **model profit** = profit from model improvement after data augmentation
- values represent average profit per EUR issued

# Summary

- **AL improves performance and profitability of credit scorecards**
  - positive gains in different performance metrics
  - query-by-committee demonstrates most potential
- **trade-off between labeling cost and model improvement**
  - labeling cost can outweigh the model improvement
  - percentage of labeled cases is an important meta-parameter
  - when to stop labeling?
- **further experiments needed to clarify the potential of AL**
  - strong impact of the data characteristics on costs & benefits
  - in which environments AL is useful?

# References


Banasik, J., Crook, J., & Thomas, L. (2003). **Sample selection bias in credit scoring models.** Journal of the Operational Research Society, 54(8), 822-832.

Culver, M., Kun, D., & Scott, S. (2006). **Active learning to maximize area under the ROC curve.** In Sixth International Conference on Data Mining (ICDM'06) (pp. 149-158). IEEE.

Krempl, G., Kottke, D. (2017). **On Optimising Sample Selection in Credit Scoring with Active Learning.** In Credit Scoring and Credit Control XV. (pp. 2). Credit Research Centre.

Krempl, G., Kottke, D., & Lemaire, V. (2015). **Optimised probabilistic active learning (OPAL): For fast, non-myopic, cost-sensitive active classification**, Machine Learning, 100(2–3), 449–476.

Settles, B. (2012). **Active Learning.** Synthesis Lectures on Artificial Intelligence and Machine Learning #18. Morgan & Claypool Publishers.

Seung, H.S., Opper, M., & Sompolinsky, H. (1992). **Query by committee.** In Proceedings of the ACM Workshop on Computational Learning Theory, 287-294.

# Active Learning for Reject Inference in Credit Scoring

**Nikita Kozodoi, Stefan Lessmann, Samantha Sizemore**

## Contact

*nikita.kozodoi@hu-berlin.de*

*linkedin.com/in/kozodoi*

*bit.ly/kozodoi_hu*

## Slides