Shallow Self-Learning for Reject Inference in Credit Scoring

Nikita Kozodoi[1,2] , Panagiotis Katsas[2],
Stefan Lessmann[1], Luis Moreira-Matias[2] and Konstantinos Papakonstantinou[2]

nikita.kozodoi@hu-berlin.de

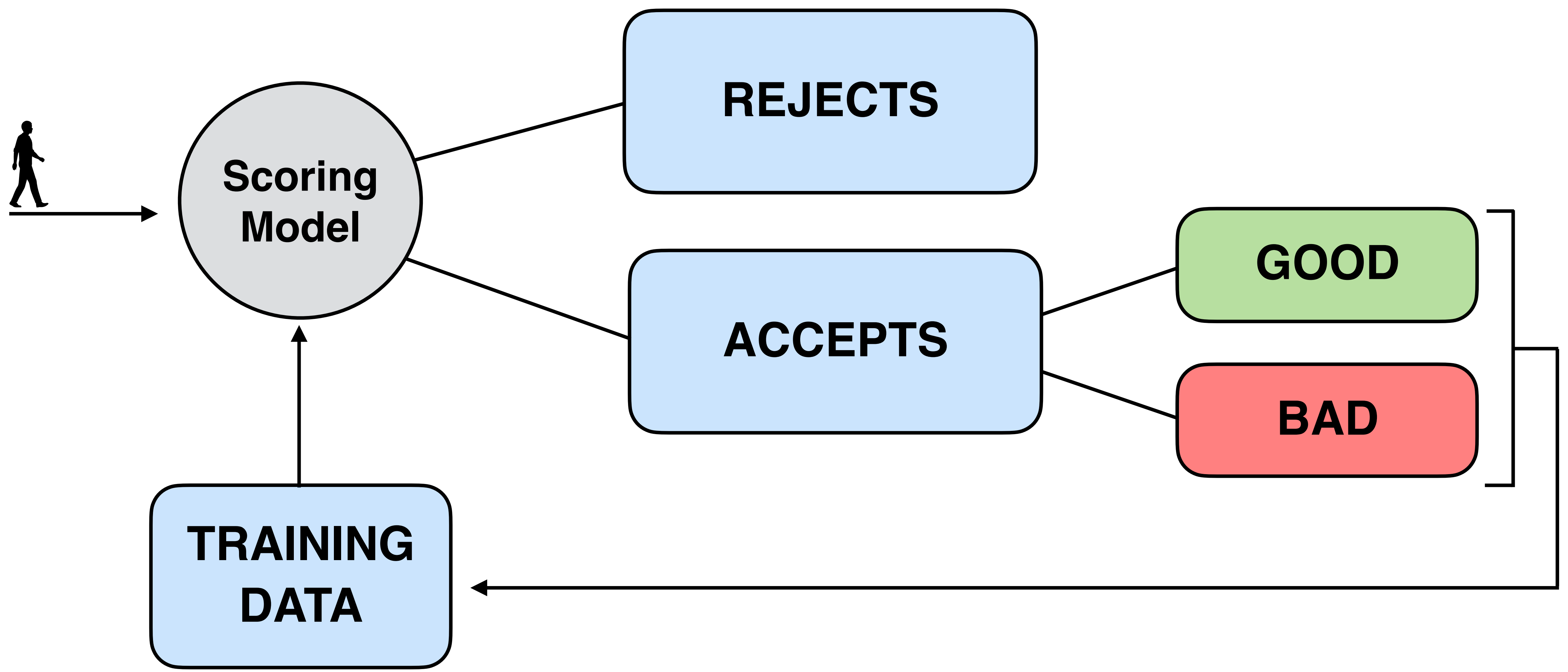
1 Humboldt University of Berlin

2 kreditech

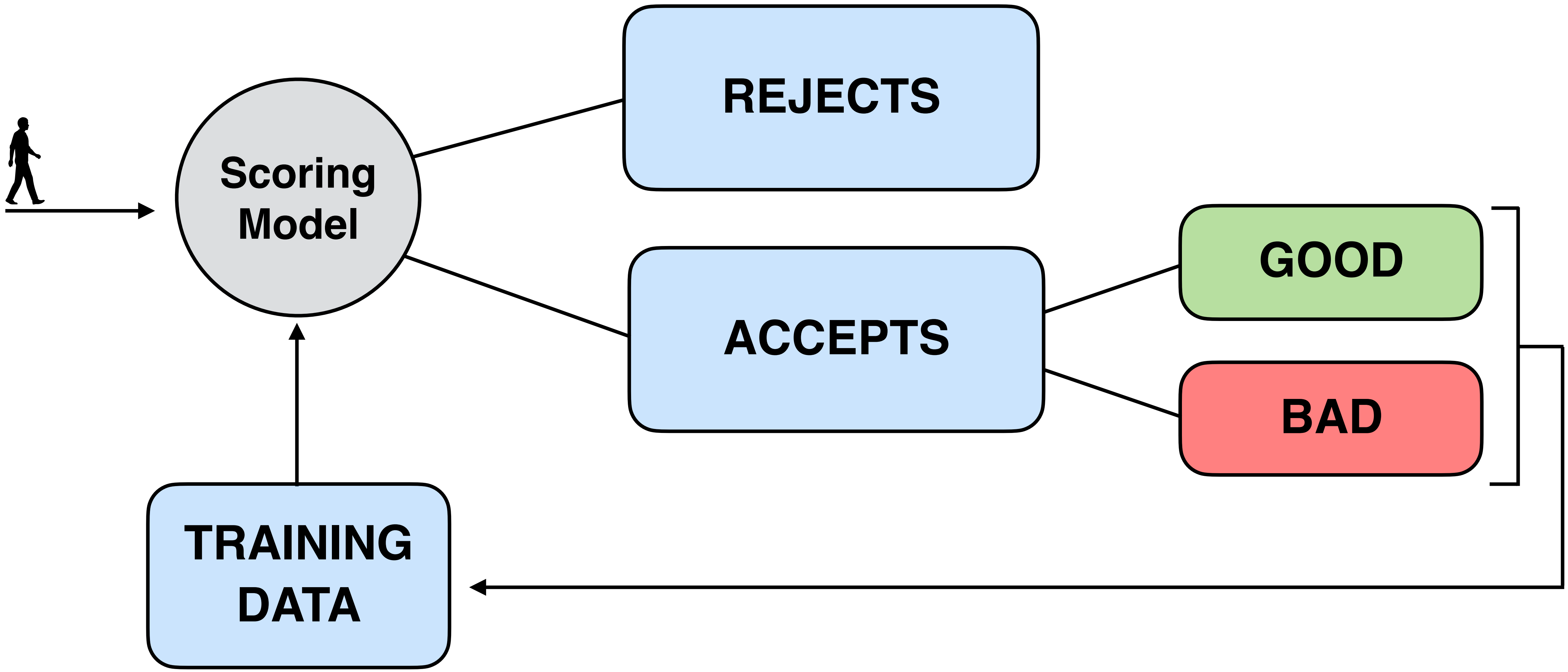
Presentation Outline

1. Sample Bias Problem
2. Shallow Self-Learning for Reject Inference
3. Evaluation Problem
4. Kickout Metric for Model Selection
5. Performance Evaluation

Motivation: Acceptance Cycle

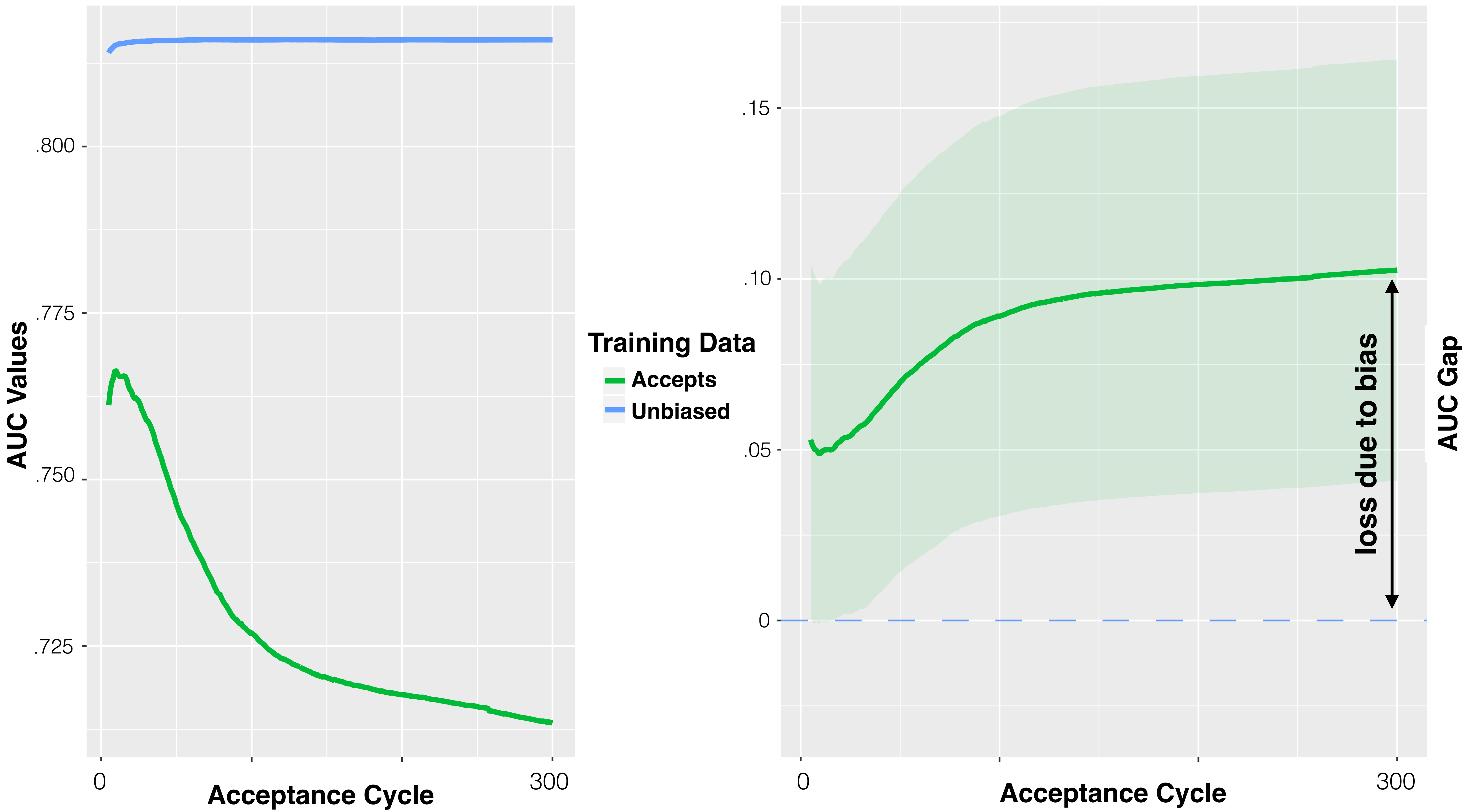


Motivation: Acceptance Cycle



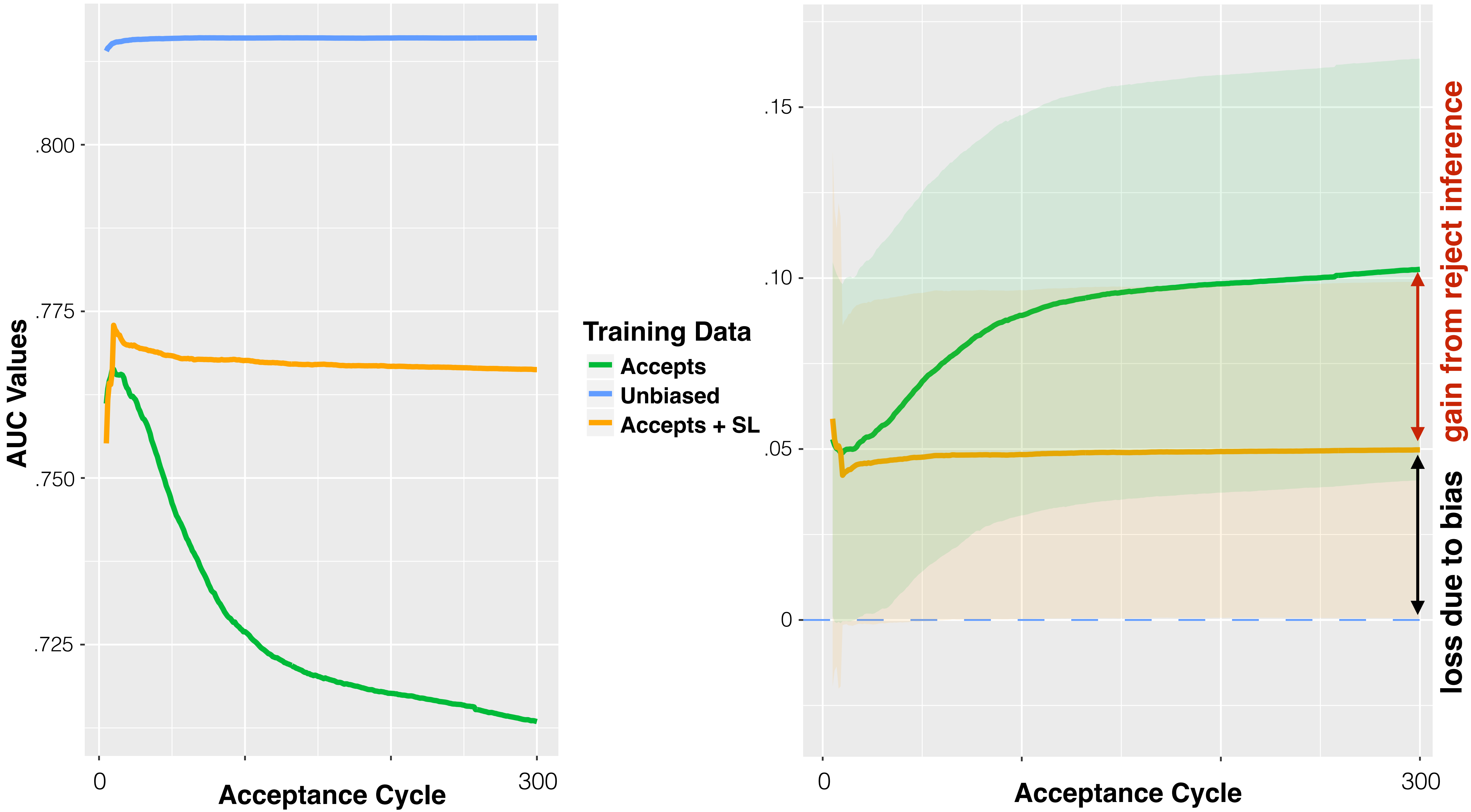
- acceptance cycle creates **sample bias**
- labels are **not missing at random**

Sample Bias: Impact on Performance



Data: multivariate Gaussians with class-specific means and covariance

Sample Bias: Gain from Reject Inference



Data: multivariate Gaussians with class-specific means and covariance

Background on Reject Inference

Reject Inference Methods

Credit Scoring Literature

- label all as BAD
- hard cutoff augmentation
- parcelling

Semi-Supervised Learning

- self-learning
- semi-supervised SVMs
- graph-based methods

Label Noise Correction

- CV-based voting
- neighbor-based labeling
- evolutionary algorithms

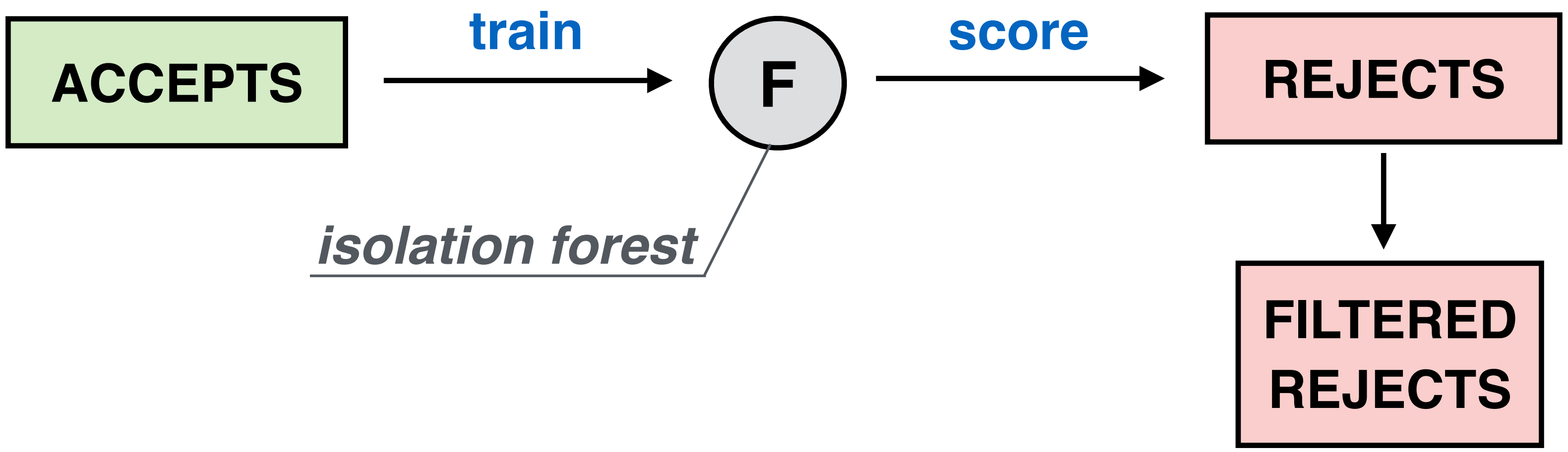
Empirical results:

- studies provide little evidence of gains from reject inference
  (Banasik et al 2005, Chen et al 2001, Cook et al 2004, Verstraeten et al 2005)
- data is often incomplete, low-dimensional or synthetic
  (e.g., Bücker et al 2013, Maldonado et al 2010)

Reject Inference with Shallow SL

Stage I

filtering

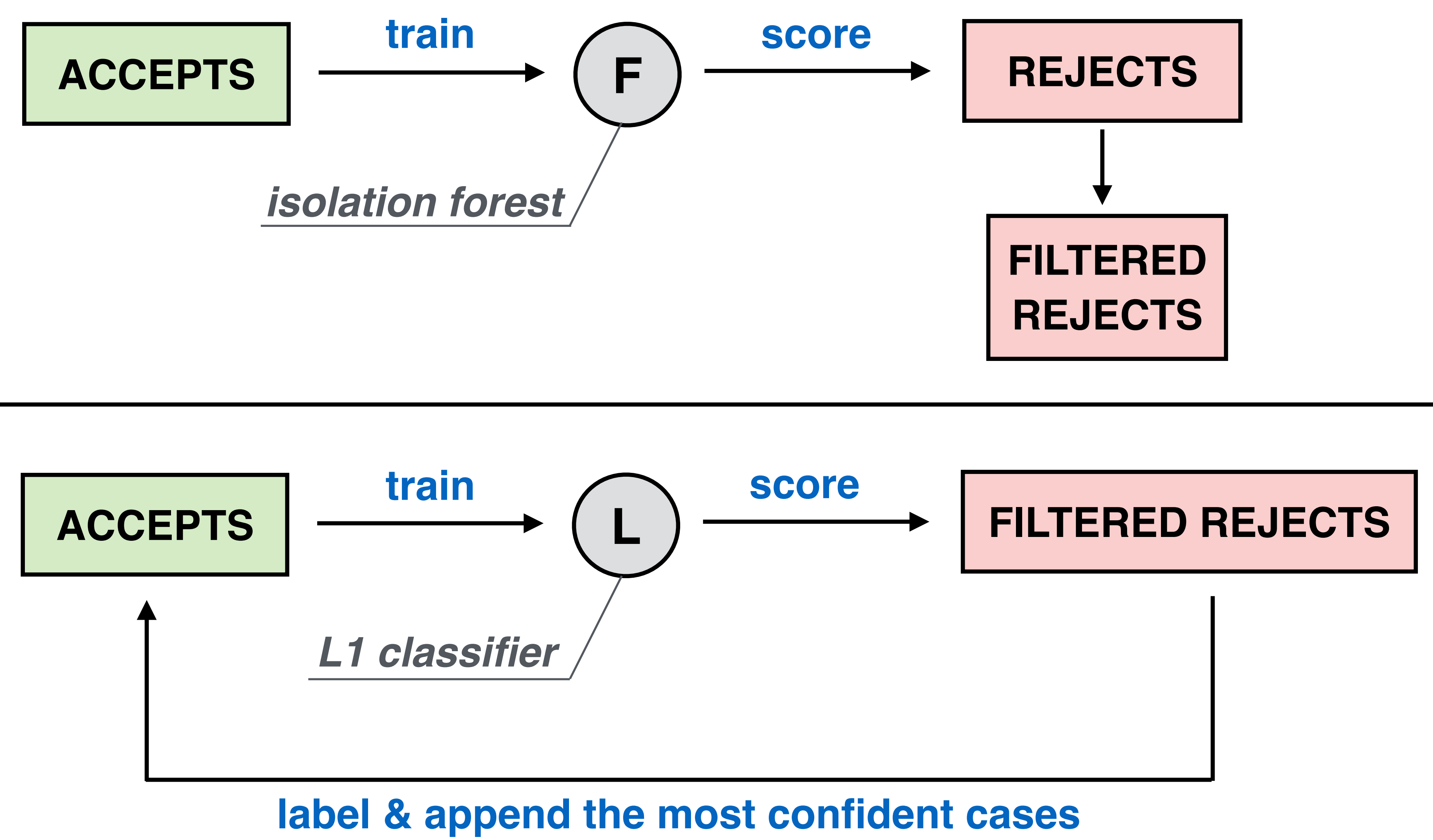


- removing rejects whose distribution is **most different** from the accepts
- reduces the risk of **error propagation** due to noise in predictions

Reject Inference with Shallow SL

Stage I
filtering

Stage II
labeling

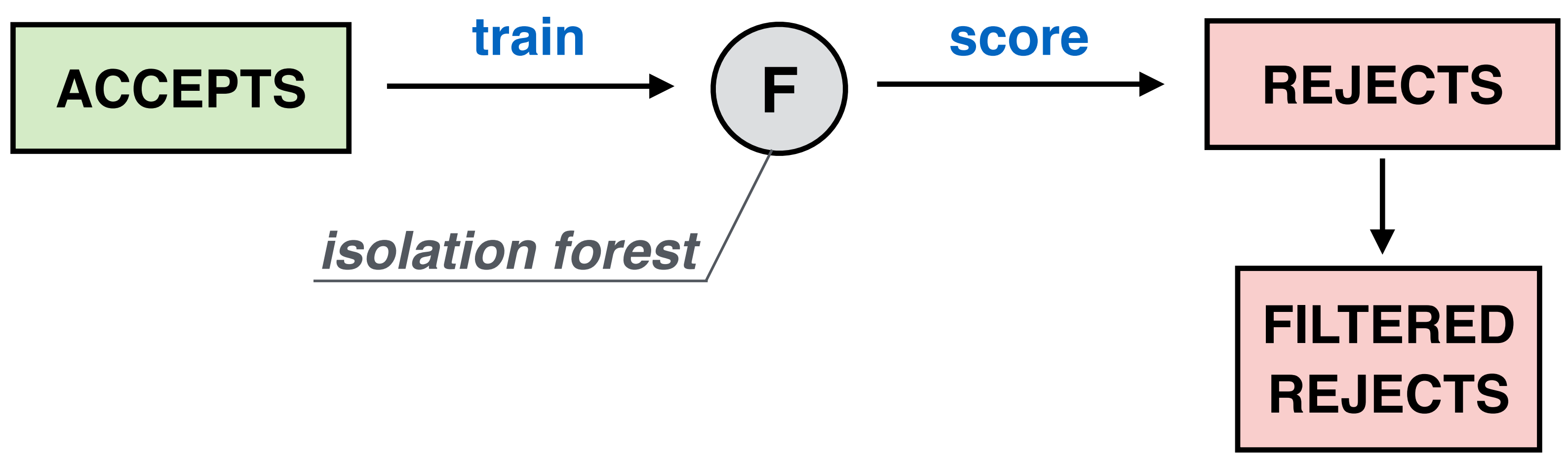


- only label rejects if the model's **confidence is high**
- using **weak learner** (L1) to get well-calibrated probabilities
- **imbalance parameter** θ to account for higher BAD rate among rejects
- **stopping criteria:** confidence threshold & scoring model performance

Reject Inference with Shallow SL

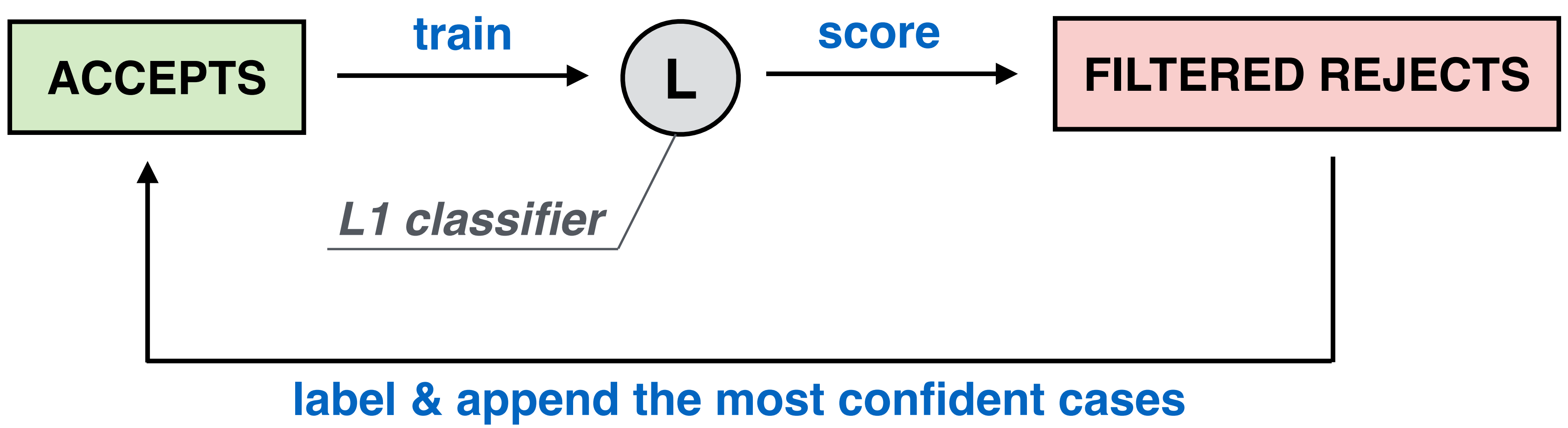
Stage I

filtering



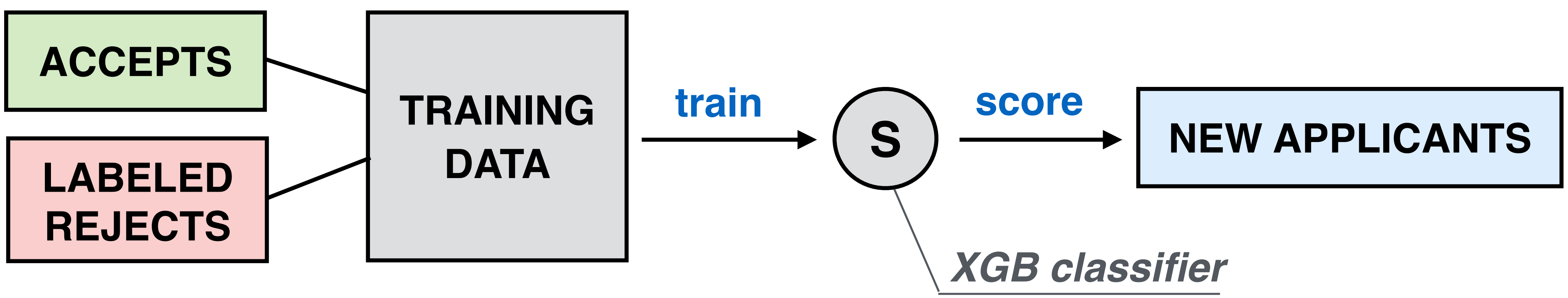
Stage II

labeling

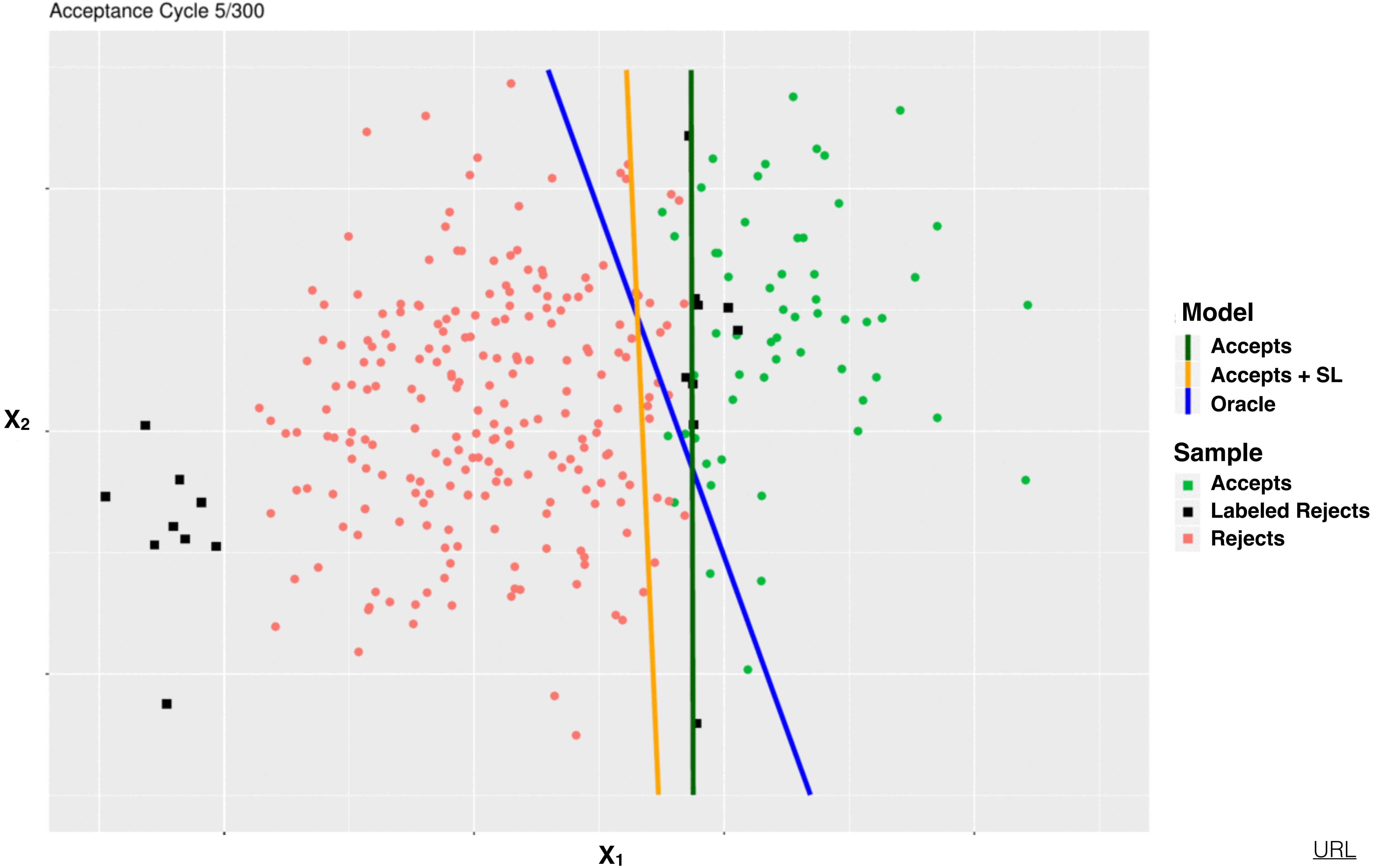


Stage III

scoring

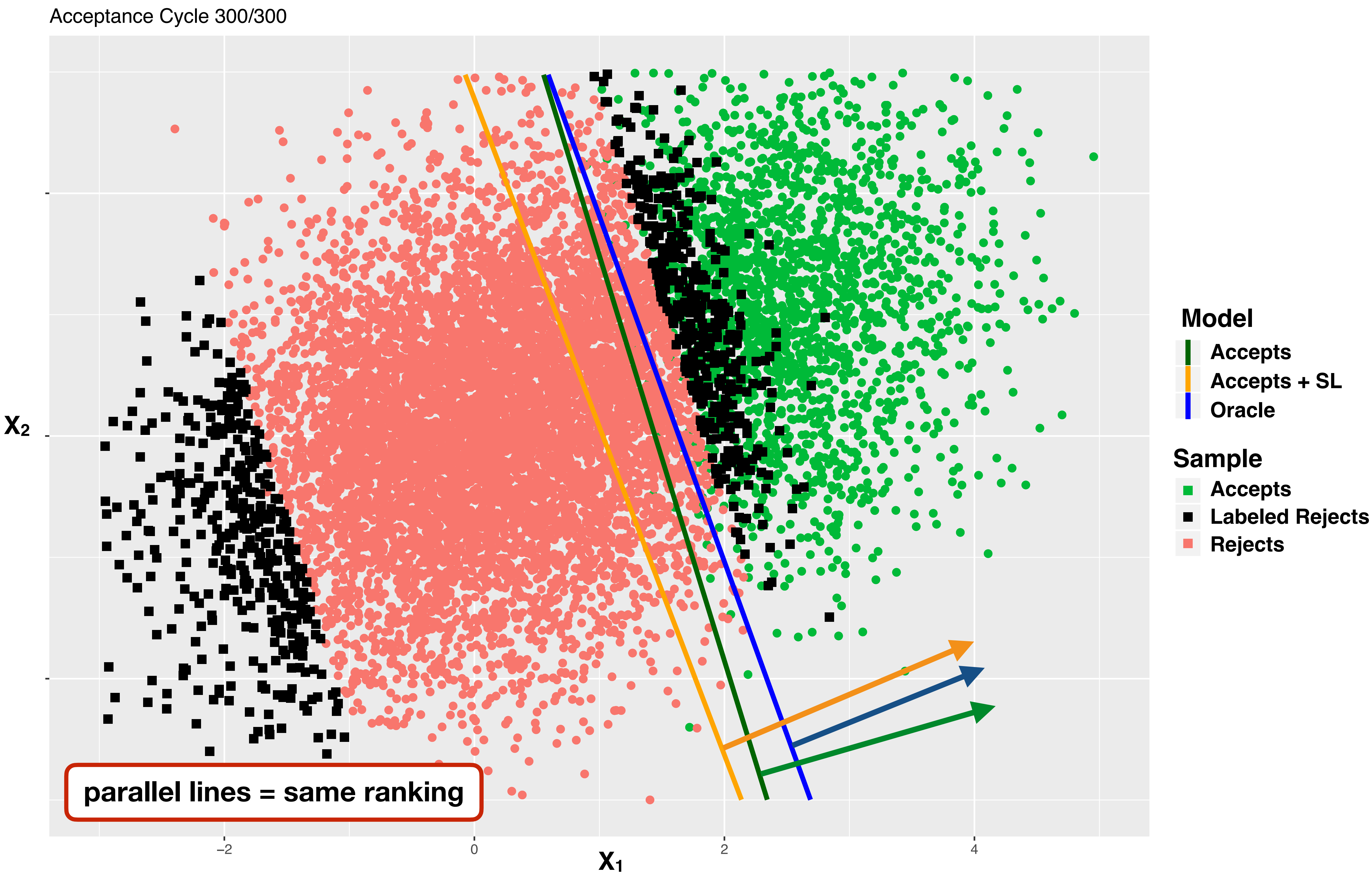


Illustrative Example on Synthetic Data



URL

Illustrative Example on Synthetic Data



Evaluation Problem: Correlation Analysis

	AUC (accepts)	AUC (unbiased)
AUC (accepts)	1	
AUC (unbiased)	0.12	1

- **AUC (accepts)** = experimental AUC on a biased holdout sample of accepts
- **AUC (unbiased)** = production AUC on a representative holdout sample of clients

Data: real-world credit scoring data with synthetic labels (bureau scores)

Evaluation Problem: Correlation Analysis

	AUC (accepts)	AUC (unbiased)	Kickout
AUC (accepts)	1		
AUC (unbiased)	0.12	1	
Kickout	0.01	0.30	1

Kickout metric better correlates with performance on unbiased sample

Data: real-world credit scoring data with synthetic labels (bureau scores)

Introducing the Kickout Metric

Intuition:

- compare two scoring models: **before** and **after** reject inference [RI]
- count **GOOD** and **BAD** cases that are "kicked out" - rejected after RI
- updated model should kick out more **BAD** and less **GOOD** cases
  - *kicked out cases are replaced by rejects with **unknown labels***
  - *kicking out a **BAD** case has a positive expected value*
  - *kicking out a **GOOD** case has a negative expected value*

$$\boxed{kickout = \frac{\frac{K_B}{p(B)} - \frac{K_G}{1-p(B)}}{\frac{S_B}{p(B)}}}$$

- $\mathbf{K_B}$, $\mathbf{K_G}$ - kicked-out **BADs** and **GOODs**
- **p(B)** - probability of selecting **BAD** example
- $\mathbf{S_B}$ - number of selected **BAD** examples

Experiment on Real-World Data

Data description:

- consumer loans provided by kreditech
- contains data on **accepted** and **rejected** applicants
- also contains **unbiased sample:** loans that were randomly accepted

Characteristic	Accepts	Rejects	Unbiased
Number of cases	39,579	18,047	1,967
Number of features	2,410	2,410	2,410
BAD rate	39 %	-	66 %

Experimental Results: Performance

Method	Mean AUC* (unbiased)
Ignore rejects	**0.8007**
Label all rejects as BAD	0.6797
Bureau score based inference	0.7911
Hard cutoff augmentation	0.7994
Parceling	0.8041
Shallow Self-Learning + Kickout	**0.8072**

*average across **50 bootstrap samples**

Experimental Results: Business Value

Assumptions:

- acceptance rate = **30%** (applicants with the lowest predicted score)
- average loan amount = **$17,100**[1]
- average interest rate = **10.36%**[1]
- average loss given default = **25%**[2]

Business value:

- difference between ignoring rejects and proposed method translates to **60** less defaulted loans for every **10,000** accepted clients
- potential gains = **$1.13 million** * **0.25** = **$283,073**

[1] Source: https://www.supermoney.com/studies/personal-loans-industry-study/

[2] Source: https://www.globalcreditdata.org/system/files/documents/gcd_lgd_report_large_corporates_2018.pdf

Summary & Questions

1. **Demonstrated the sample bias problem**

2. **Introduced a new reject inference method**
   - labeling rejects with **shallow self-learning** to mitigate bias

3. **Introduced a new evaluation metric**
   - performance on accepts poorly correlates with performance on the unbiased sample
   - **kickout metric** is a more suitable measure for model selection

4. **Evaluated performance gains**
   - proposed method increases **AUC** compared to benchmarks
   - potential monetary gains are ~ **$300k** per **10,000** loans