Reviewer Integration and Performance Measurement for Malware Detection

Brad Miller, Alex Kantchelian, Michael Carl Tschantz, Sadia Afroz, Rekha Bachwani, Riyaz Faizullabhoy, Ling Huang, Vaishaal Shankar, Tony Wu, George Yiu, Anthony D. Joseph and J.D. Tygar

Google, UC Berkeley, ICSI, Intel Labs, DataVisor

Status Quo

- Malware detectors have room to improve
  - Only 66% of malware detected in first 24 hours*
  - 93% of malware detected in first month*

(*Damballa: State of Infections Report Q4 2014)

- ML research outperforms industry detectors
  - Multiple projects claiming >90% detection

Questions and Answers

- This talk explores two questions and answers

Questions and Answers

- This talk explores two questions and answers
- **Q:** Why does research outperform industry?

Questions and Answers

- This talk explores two questions and answers

- **Q:** Why does research outperform industry?
- **A:** Research is offline, accurate training labels

Questions and Answers

- This talk explores two questions and answers

- **Q:** Why does research outperform industry?
- **A:** Research is offline, accurate training labels

- **Q:** Can the performance gap be closed?

Questions and Answers

- This talk explores two questions and answers

- **Q:** Why does research outperform industry?
- **A:** Research is offline, accurate training labels

- **Q:** Can the performance gap be closed?
- **A:** Yes, by expert review of selected samples

Concrete Contributions

- Temporally consistent labels
  - Explains detection rate drop from 91% to 72%
- ML guided human reviewer integration
  - Increases detection from 72% to 89%
  - Detects 42% of previously undetected malware
- Open, scalable implementation & sample data

Overview

- Dataset analysis and design
  - Measure label shift; simulate reviewers at scale
- Experimental design
  - Accommodates time and integrated reviewers
- Experimental results
  - Demonstrated impact of labeling and reviewers

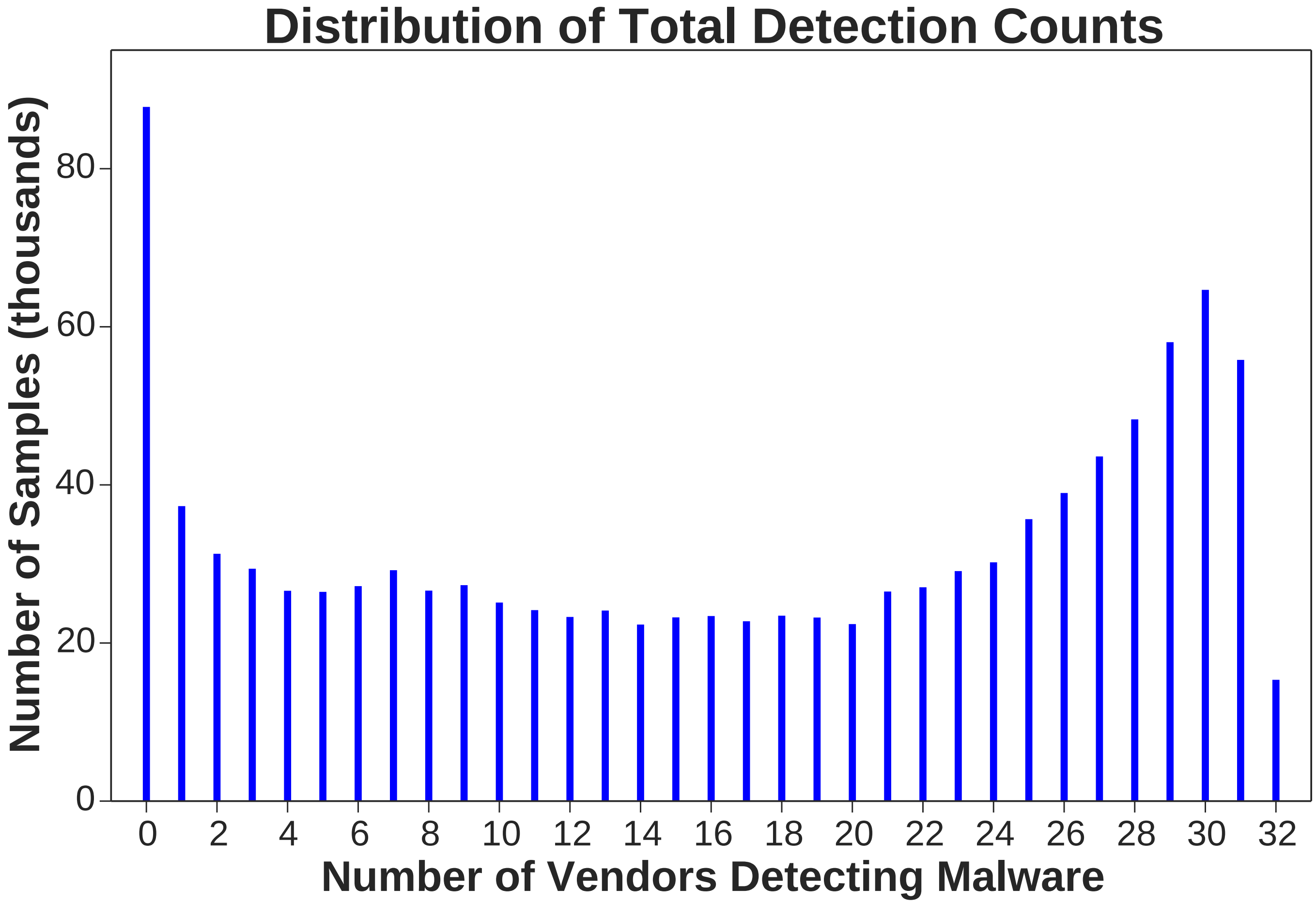
DATASET ANALYSIS AND DESIGN

Data Source

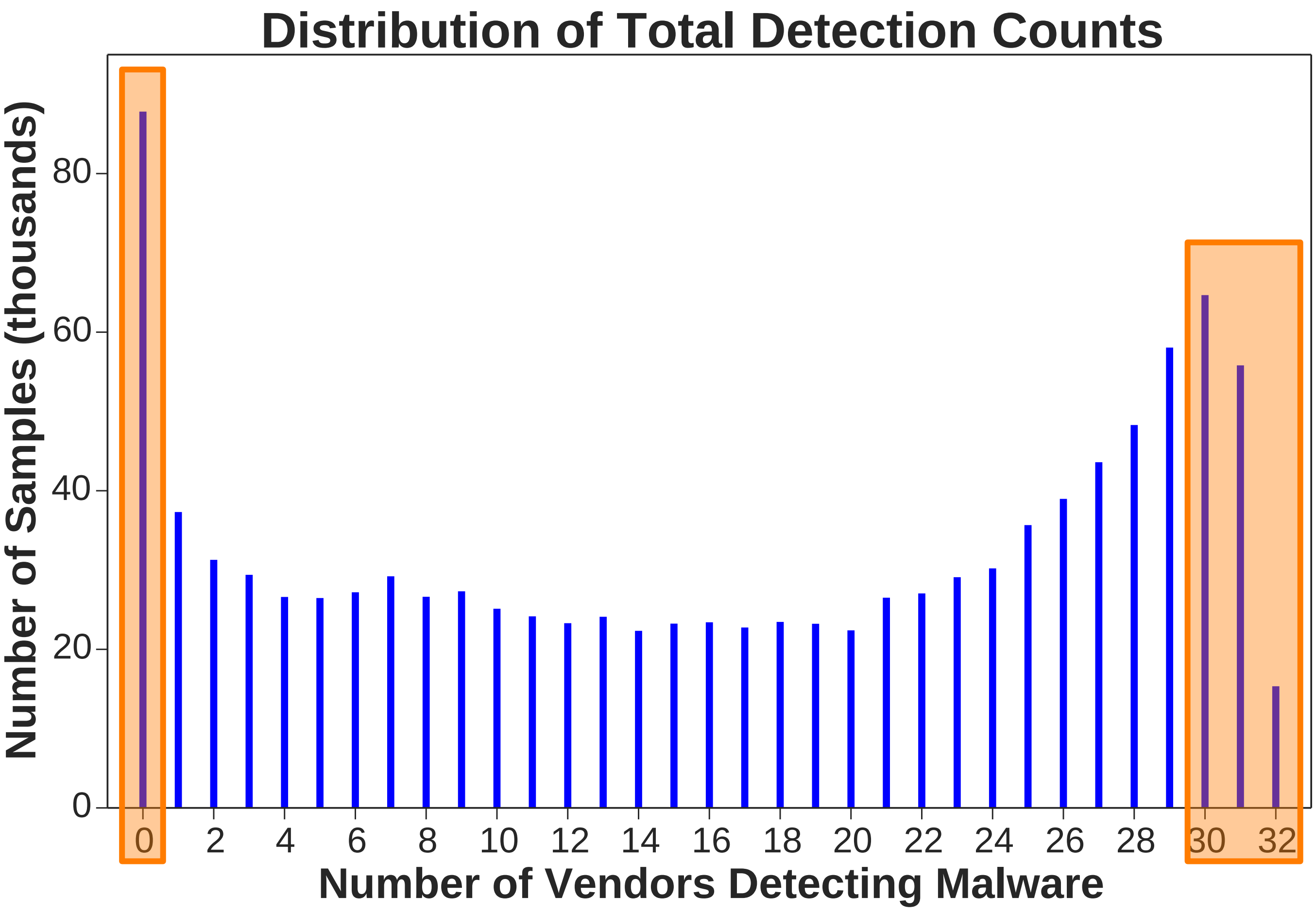
virustotal

- Scans submitted binaries with multiple AVs
- Each scan of a binary has a timestamp
- Re-scans occur upon request or re-submission
- 1M+ samples, 2M+ scans, spanning 2.5 years

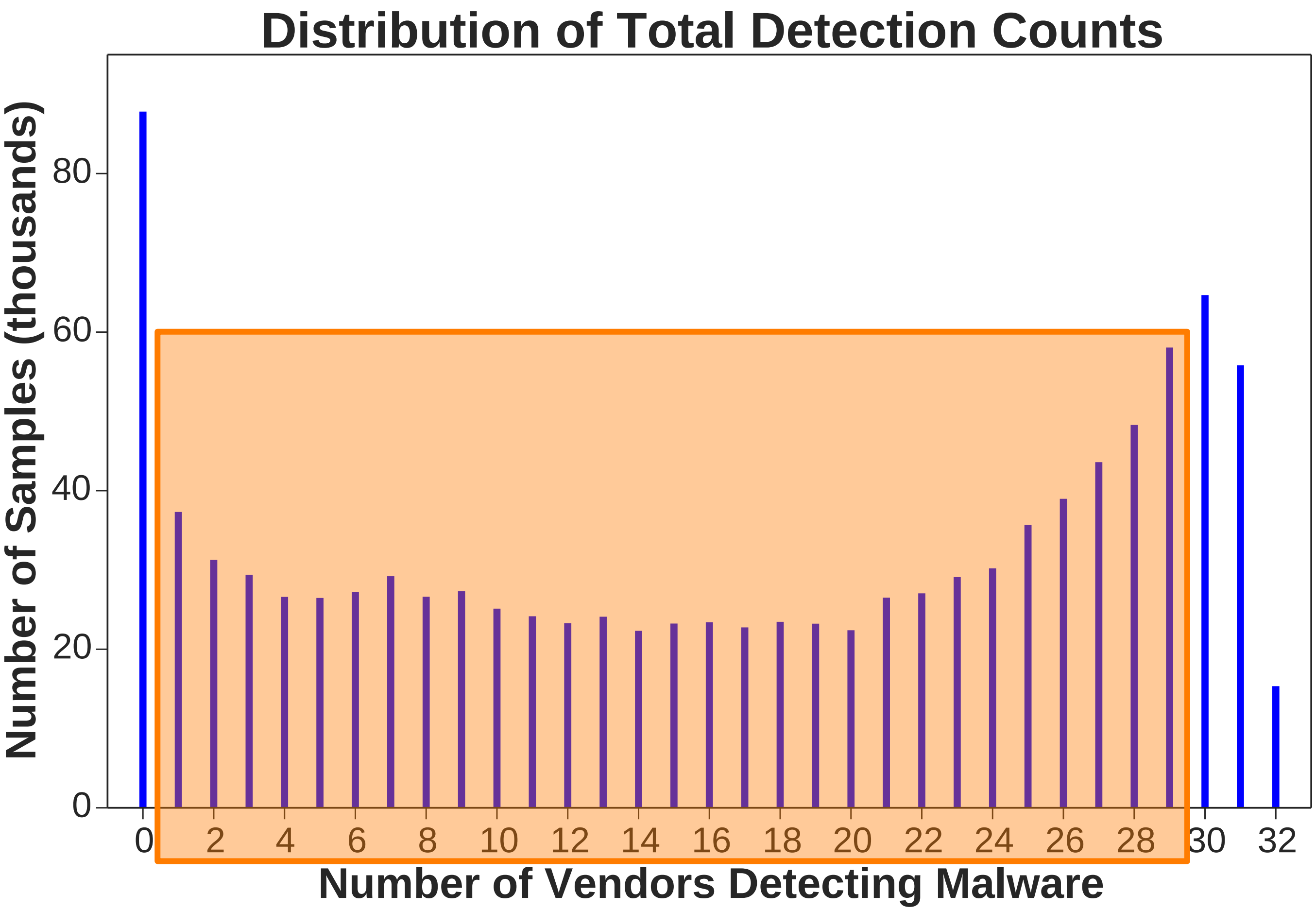
Initial Detection Results



Initial Detection Results

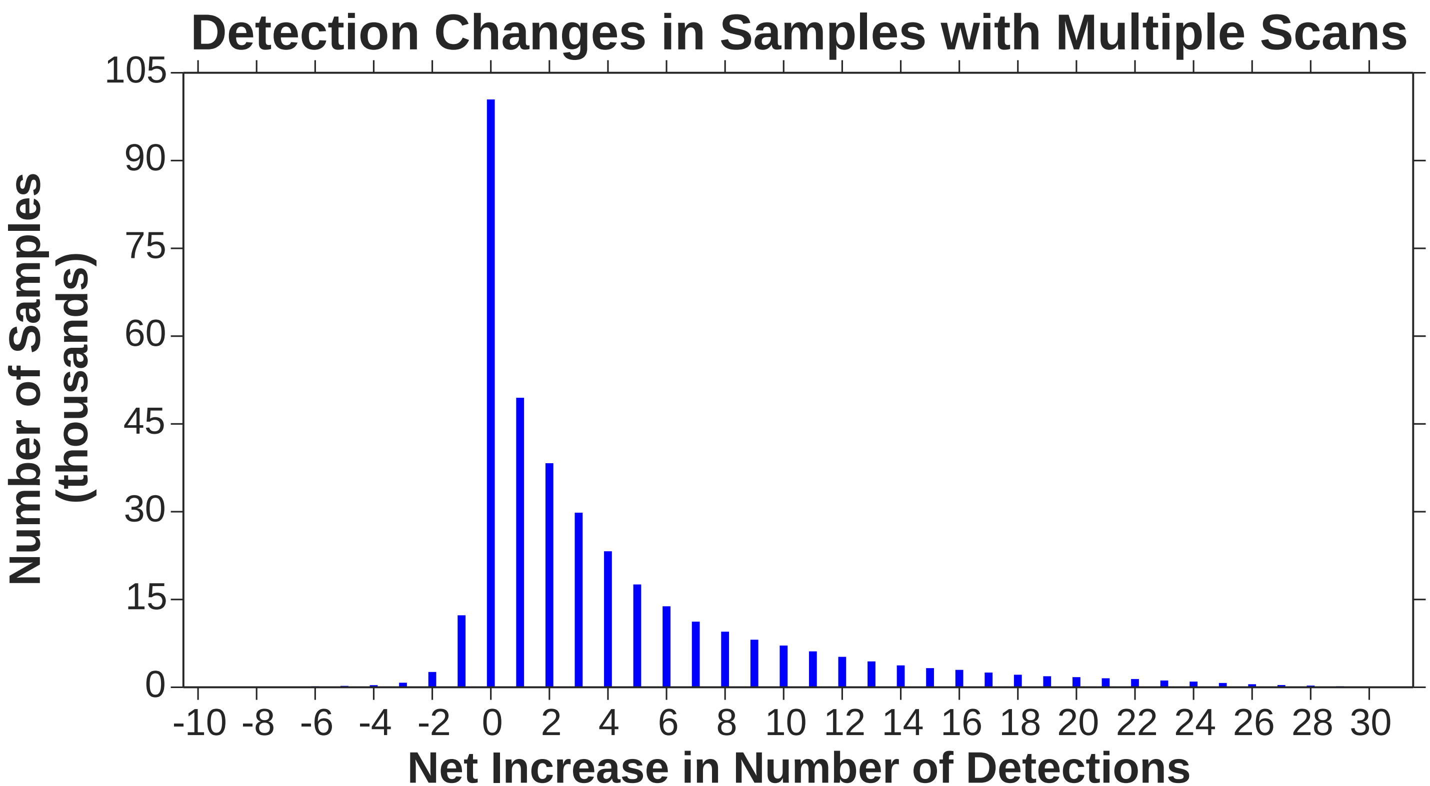


Initial Detection Results



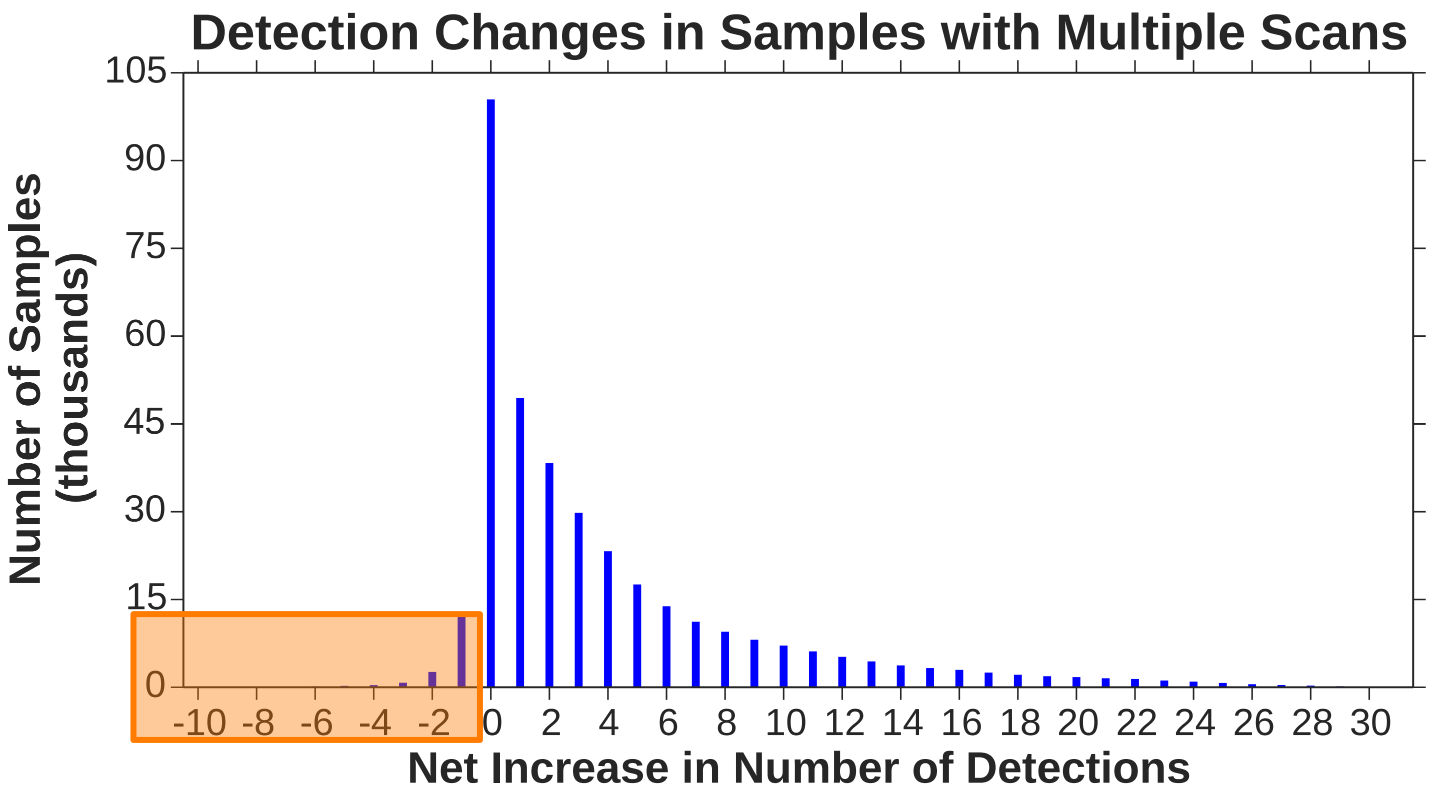
Detection Changes

- Detections generally increase with time



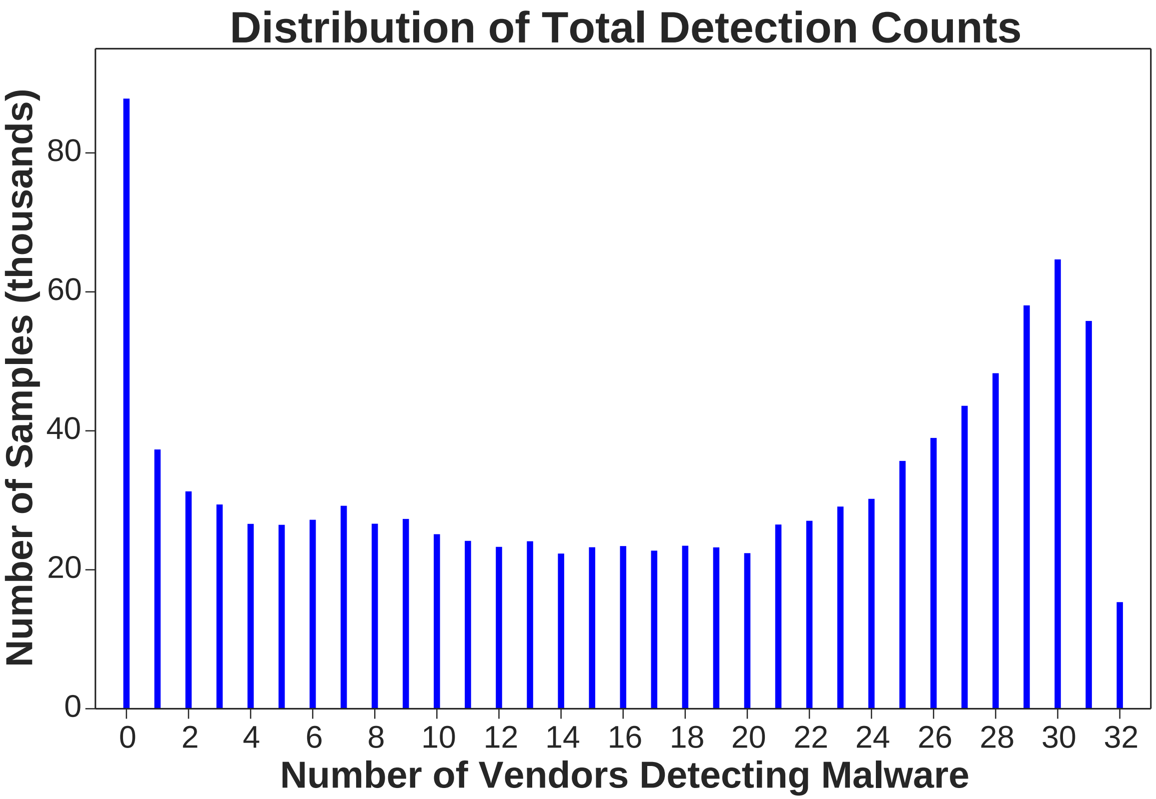
Detection Changes

- Detections generally increase with time



Final Detection Results

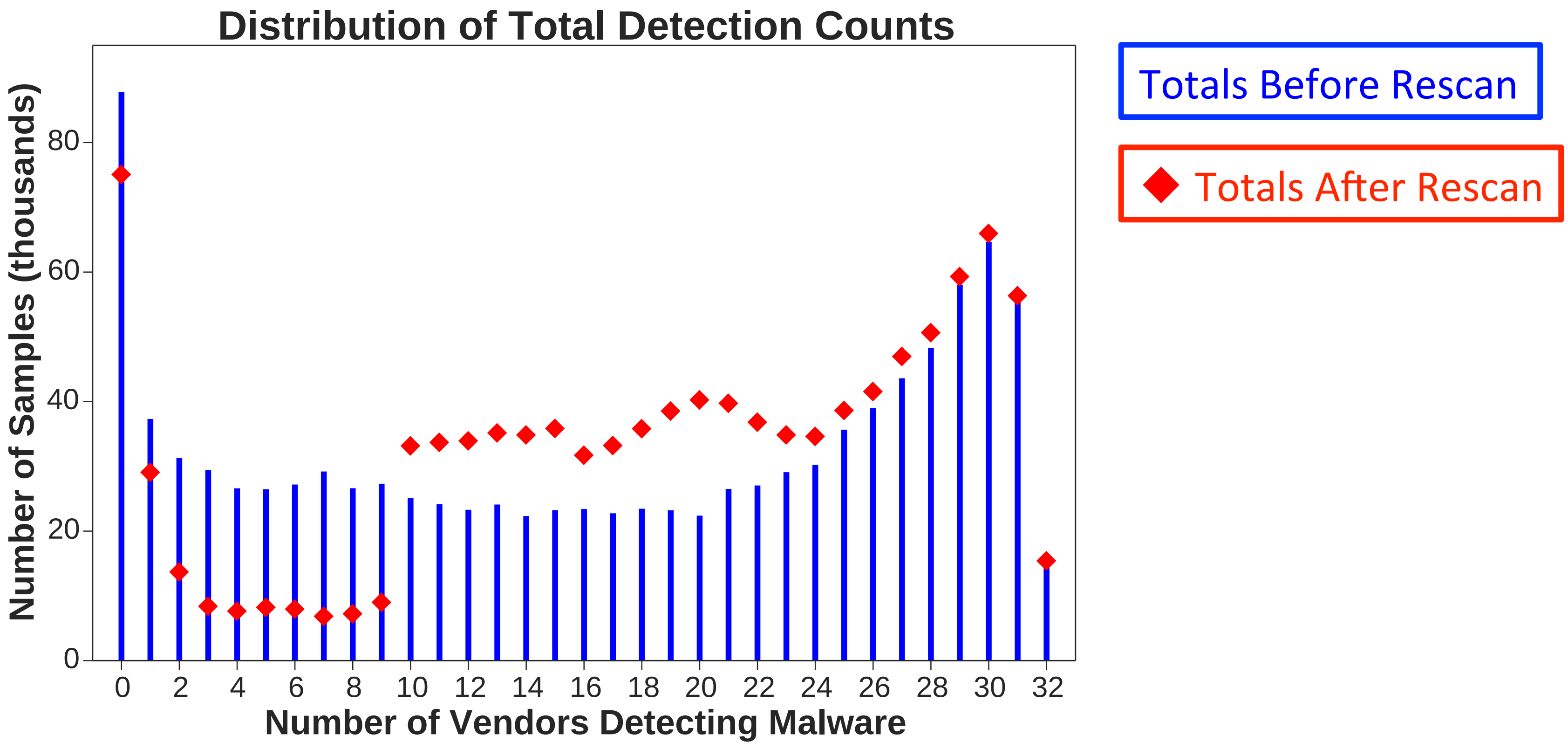
- Rescan ambiguous samples to clarify labels



Totals Before Rescan

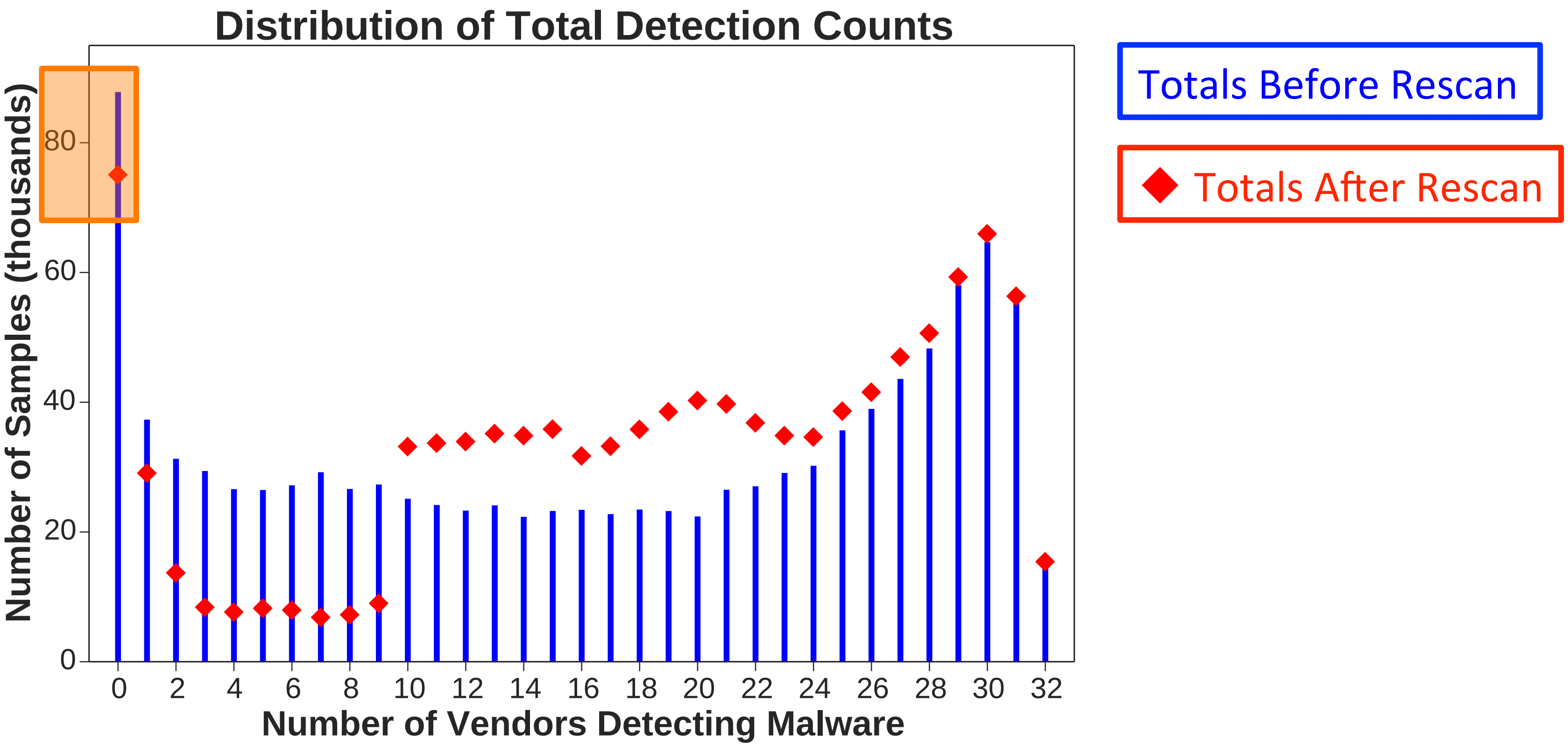
Final Detection Results

- Rescan ambiguous samples to clarify labels



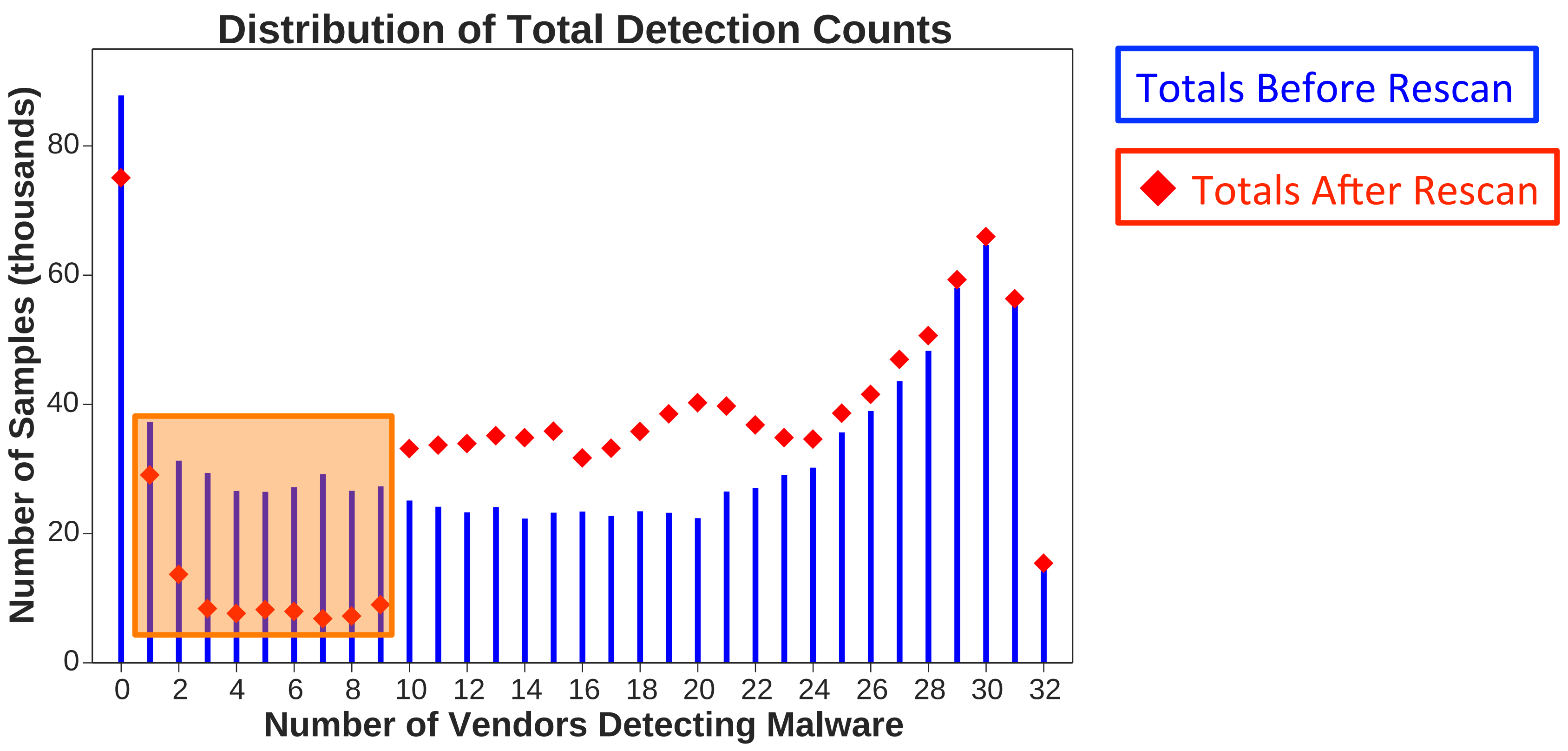
Final Detection Results

- Rescan ambiguous samples to clarify labels



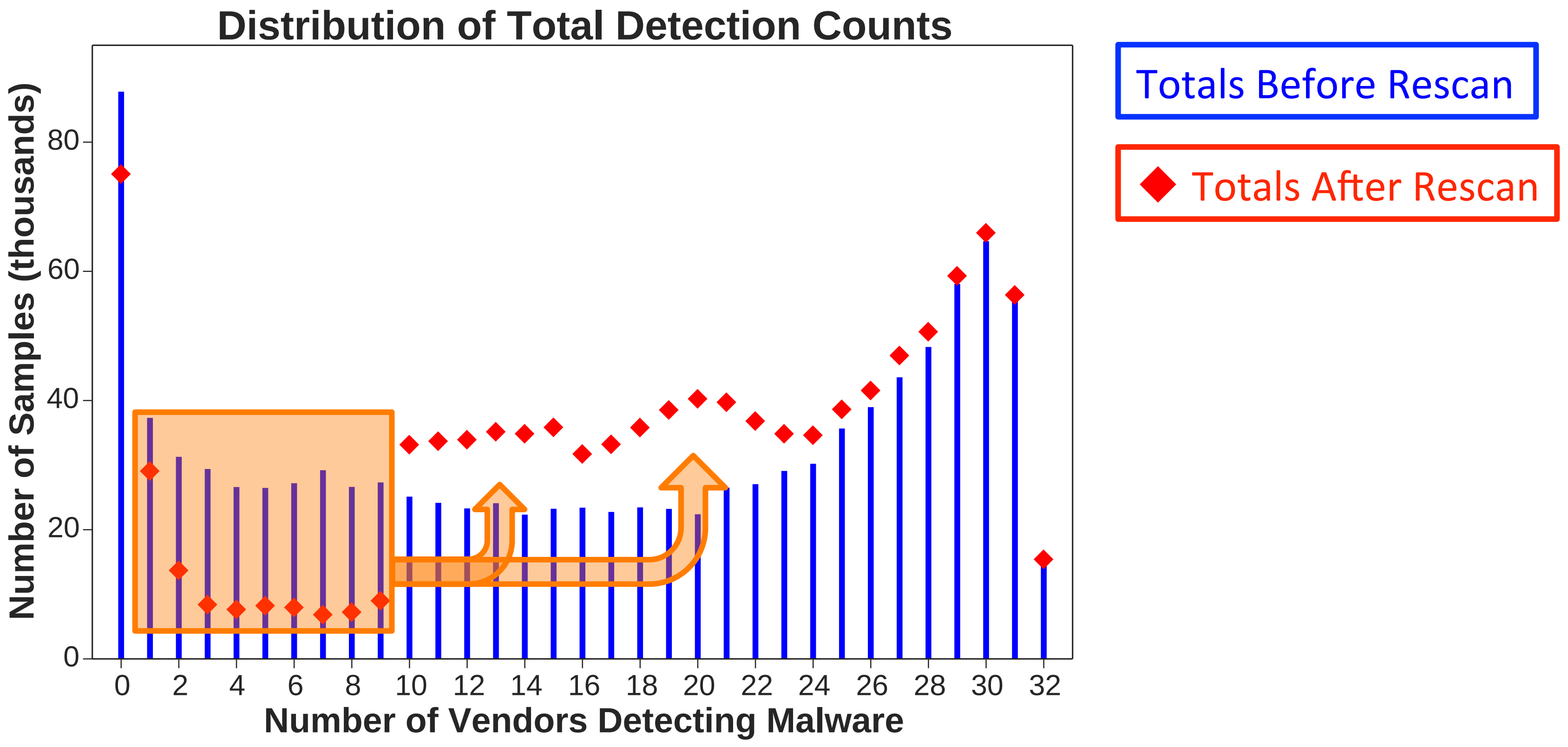
Final Detection Results

- Rescan ambiguous samples to clarify labels



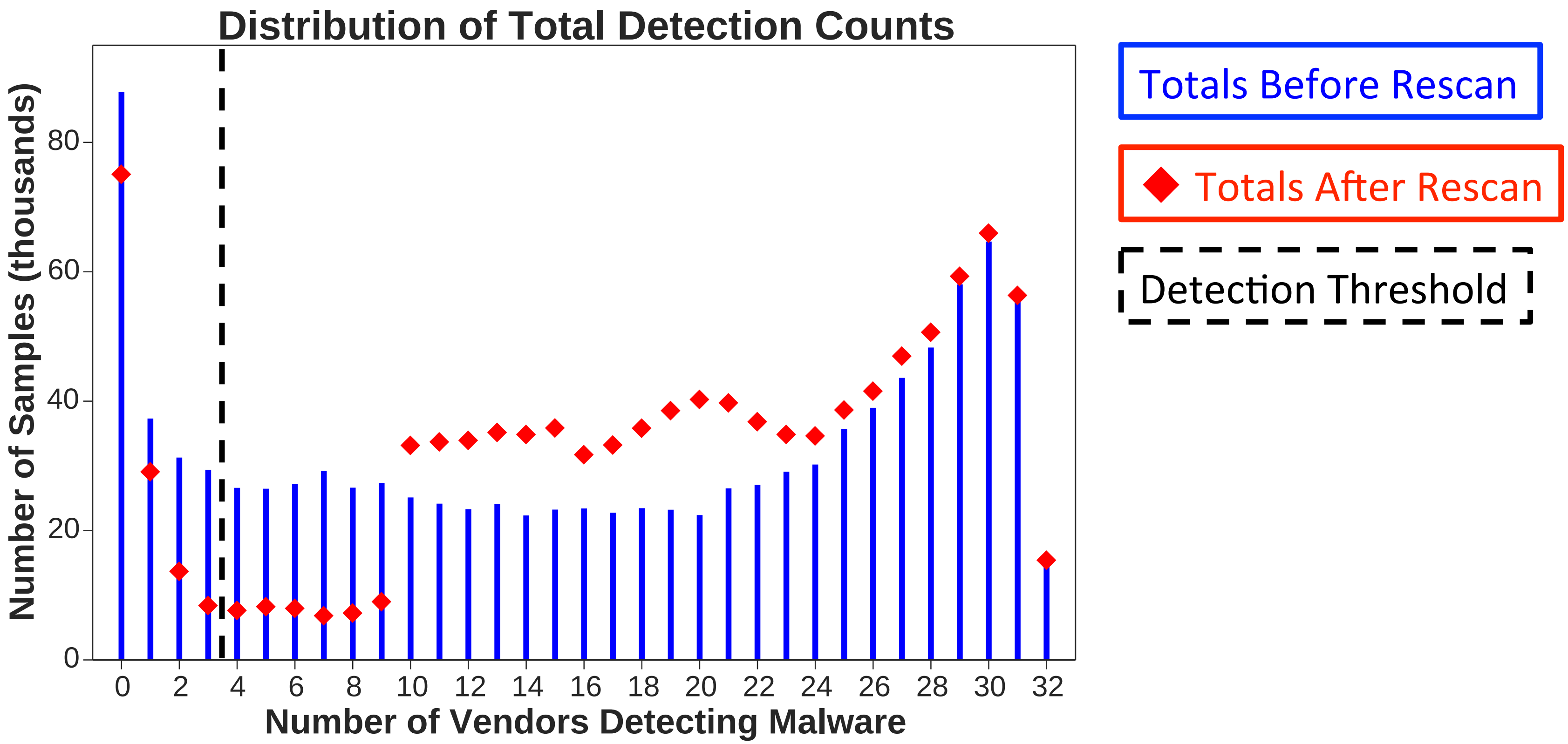
Final Detection Results

- Rescan ambiguous samples to clarify labels



Final Detection Results

- Rescan ambiguous samples to clarify labels



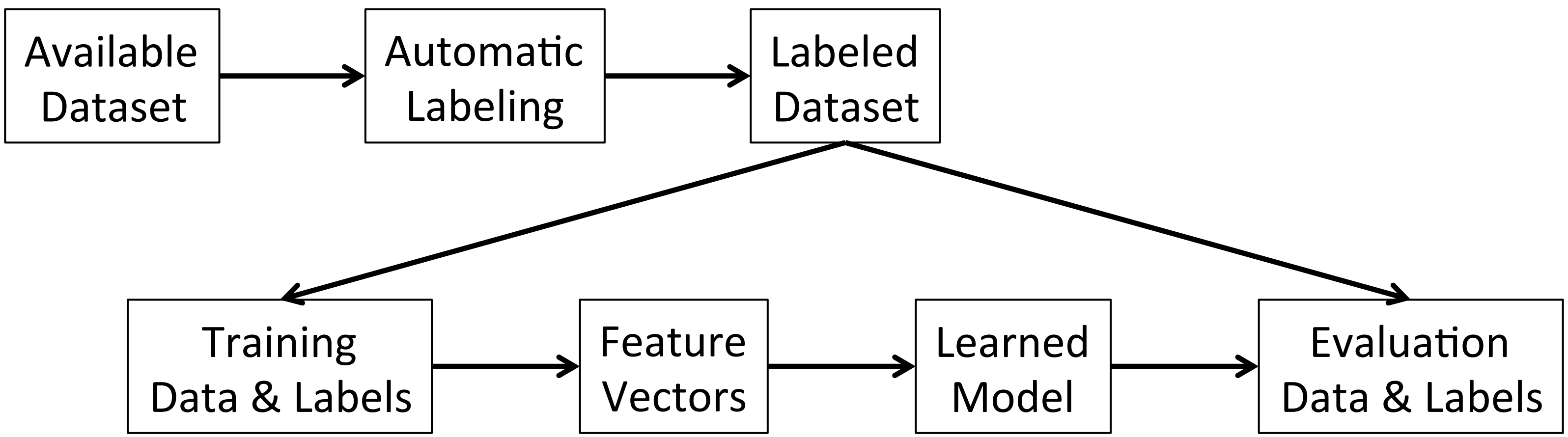
Reviewer Simulation at Scale

- Expert review not tractable for our scale
- We use simulation to study review at scale
- Final scan of sample simulates reviewer label
  - Added noise simulates imperfect reviewers

EXPERIMENTAL DESIGN

Classical ML Approach

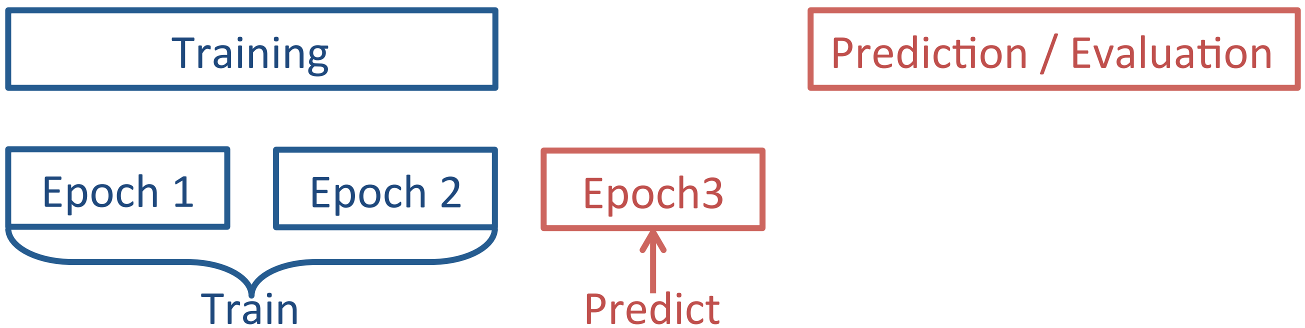
- Standard machine learning workflow



- Randomly divides training and evaluation data
- Training and evaluation labels are high quality

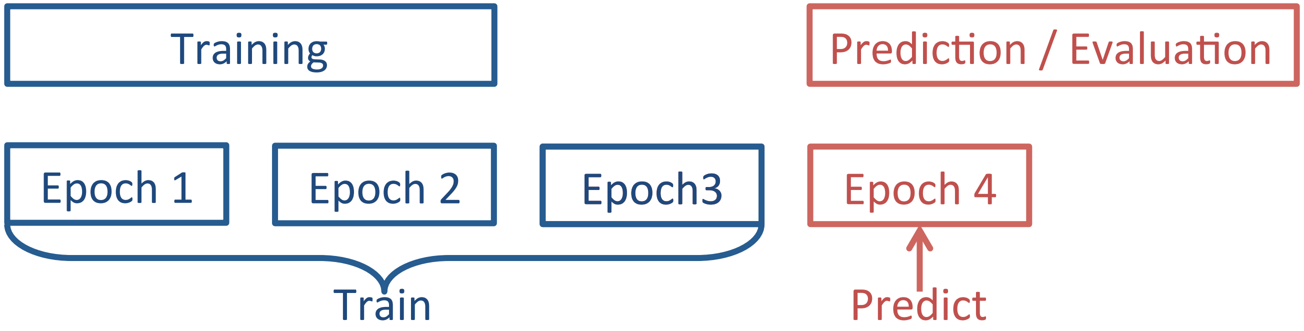
Chronological Sample Epochs

- Epochs provide sample temporal consistency



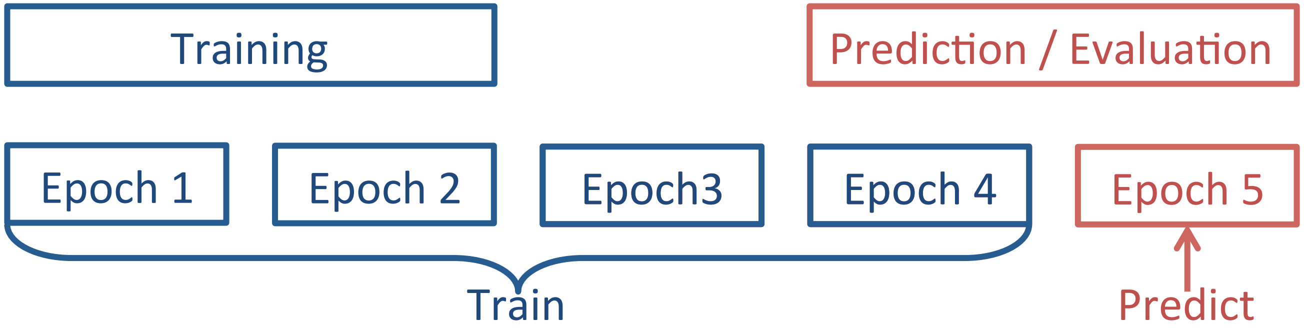
Chronological Sample Epochs

- Epochs provide sample temporal consistency



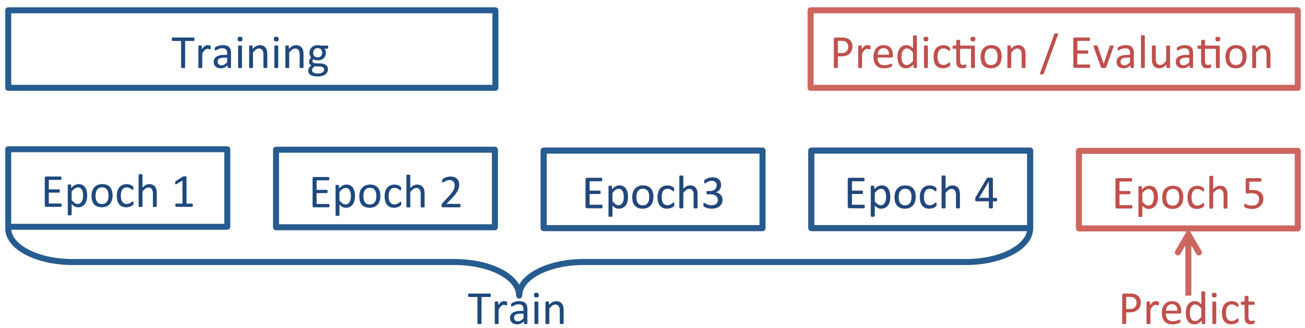
Chronological Sample Epochs

- Epochs provide sample temporal consistency

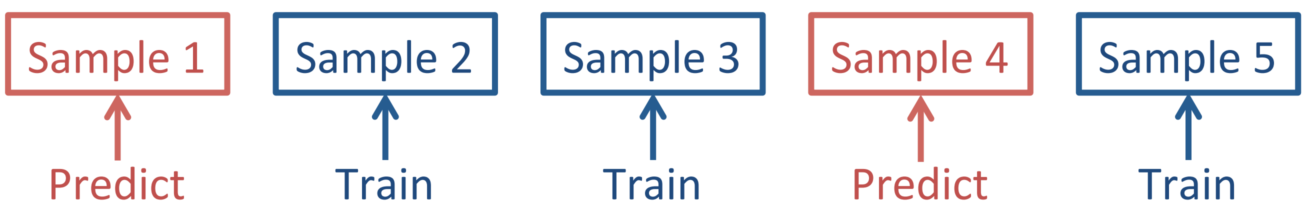


Chronological Sample Epochs

- Epochs provide sample temporal consistency

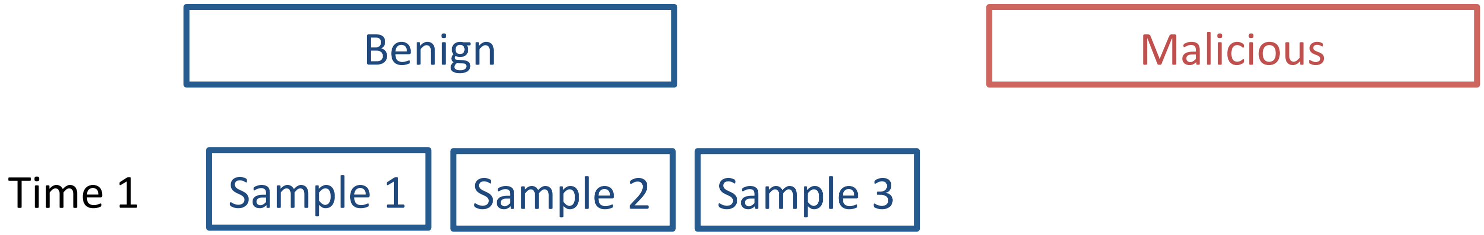


- Random division breaks temporal consistency



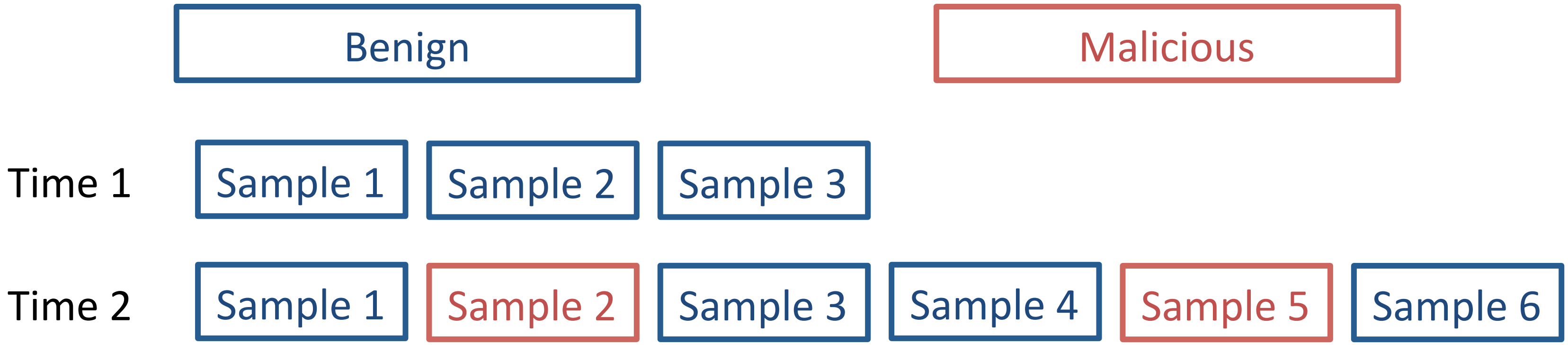
Temporally Consistent Labels

- Training labels must be known at training time
- Best possible labels used for evaluation



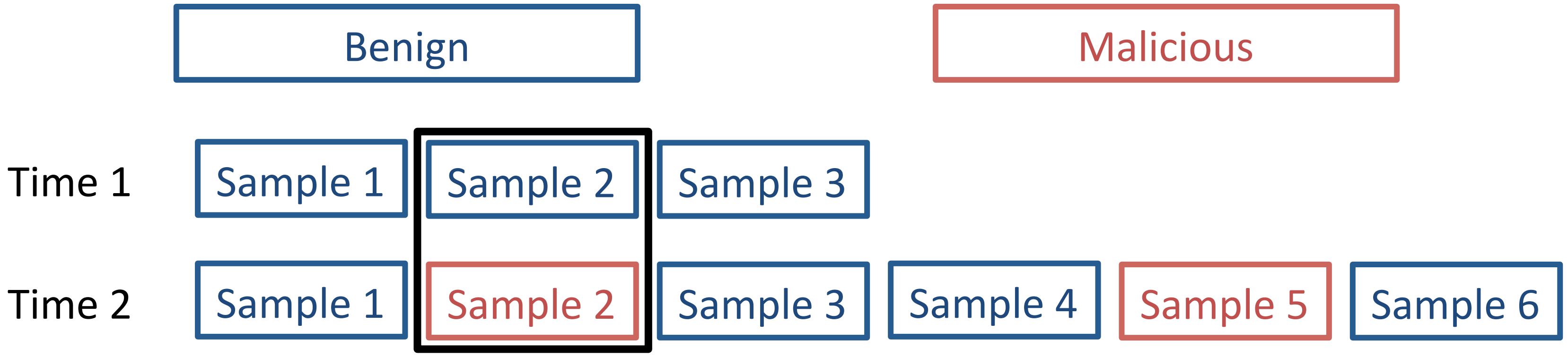
Temporally Consistent Labels

- Training labels must be known at training time
- Best possible labels used for evaluation



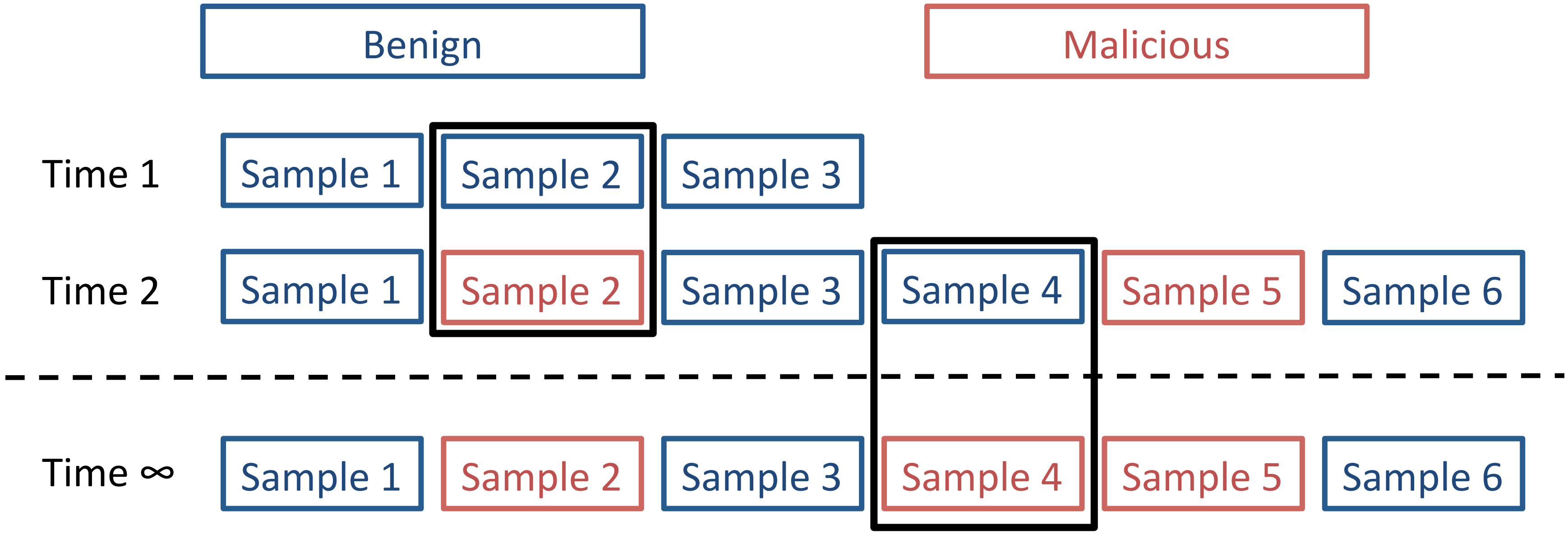
Temporally Consistent Labels

- Training labels must be known at training time
- Best possible labels used for evaluation

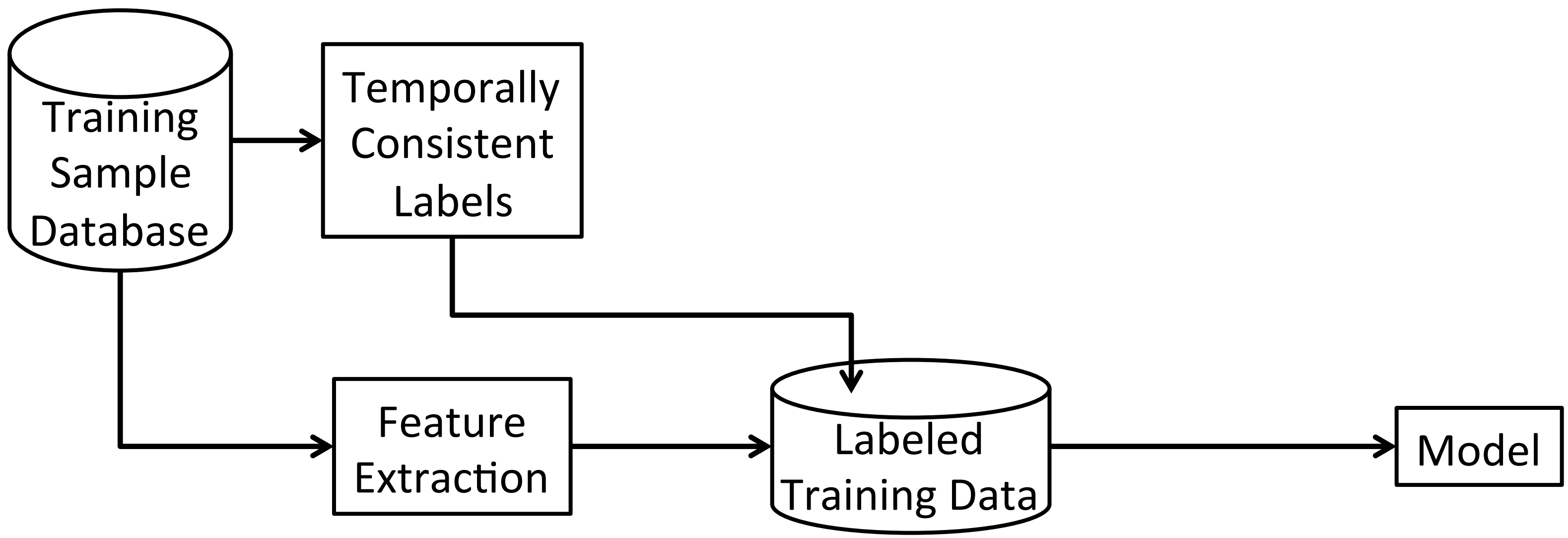


Temporally Consistent Labels

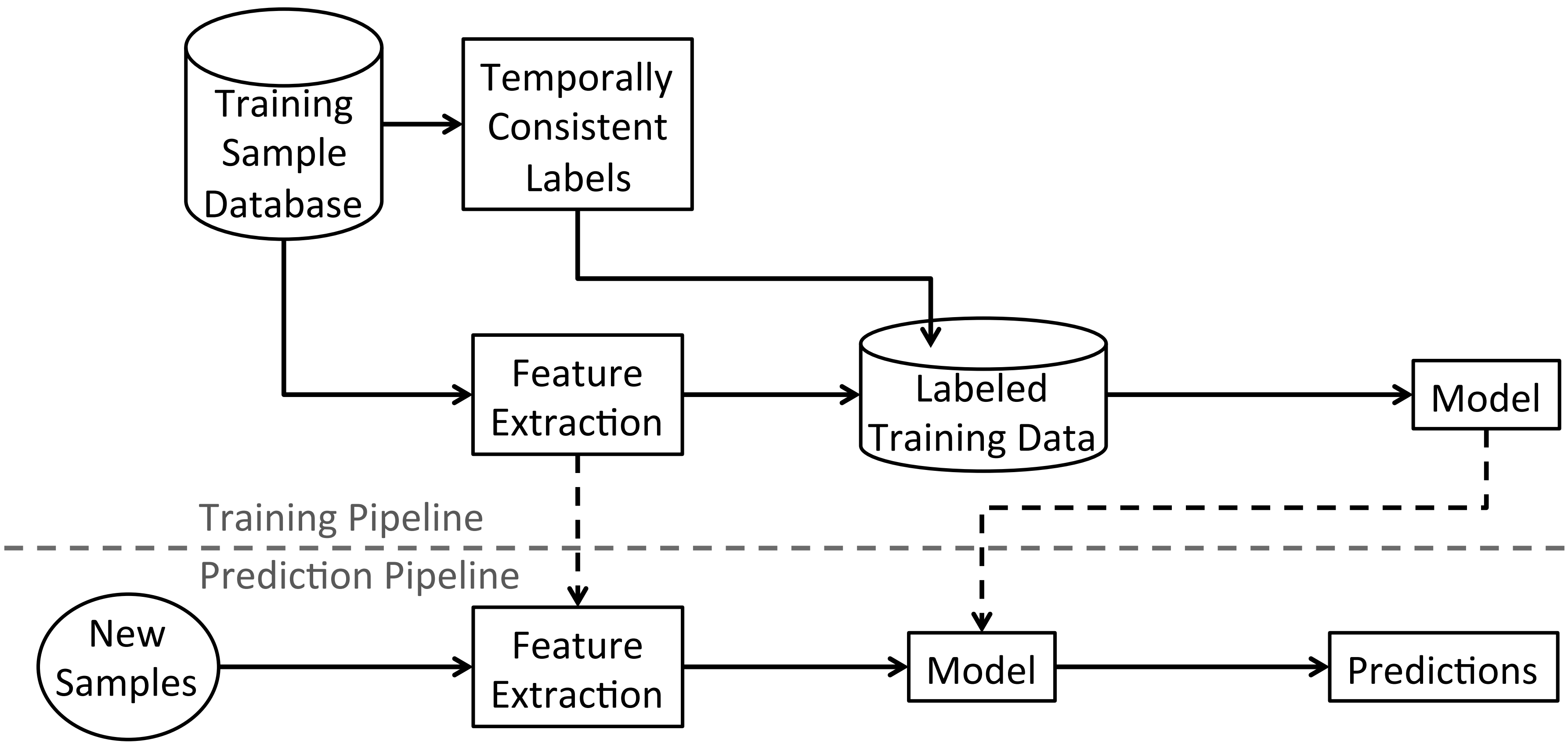
- Training labels must be known at training time
- Best possible labels used for evaluation



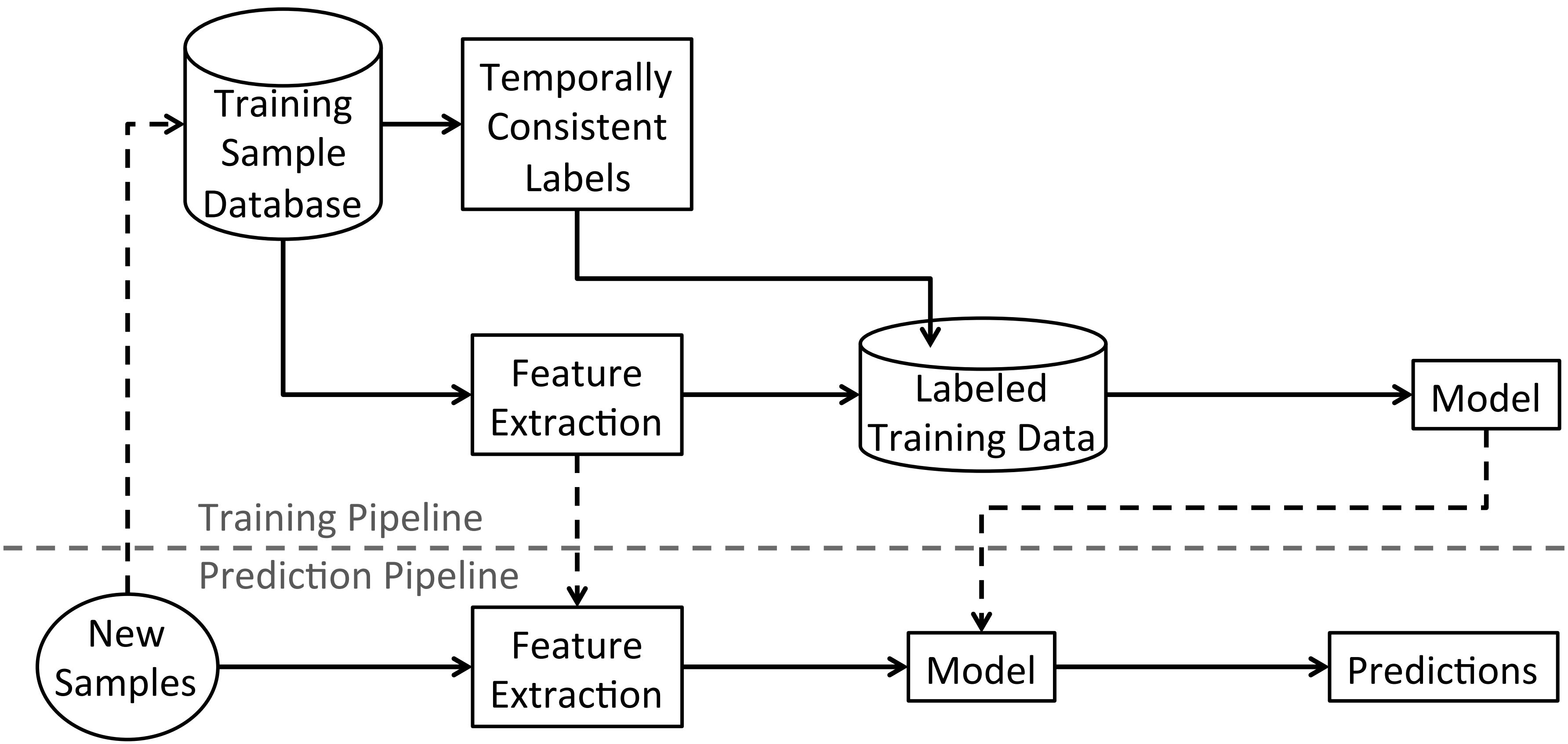
Design Overview



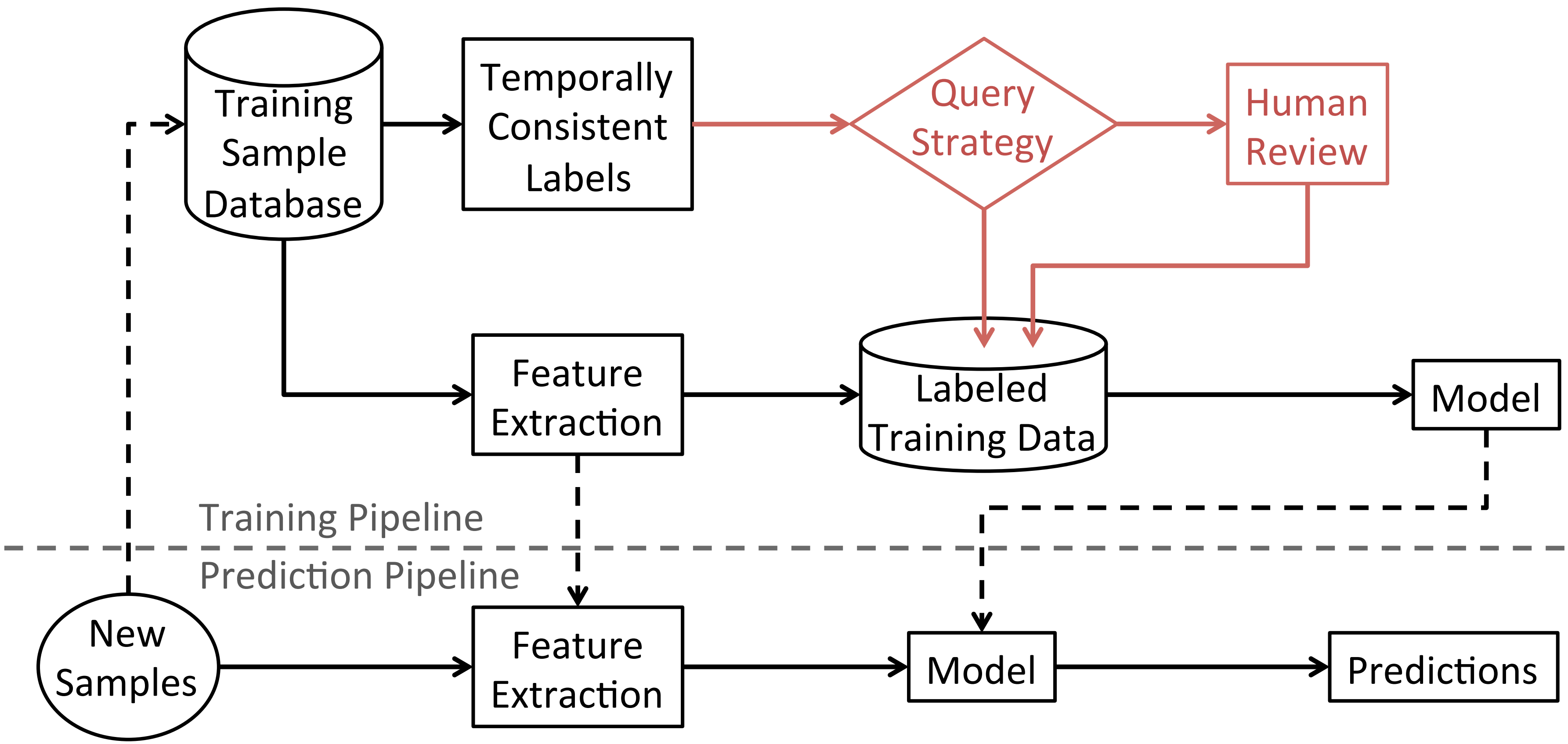
Design Overview



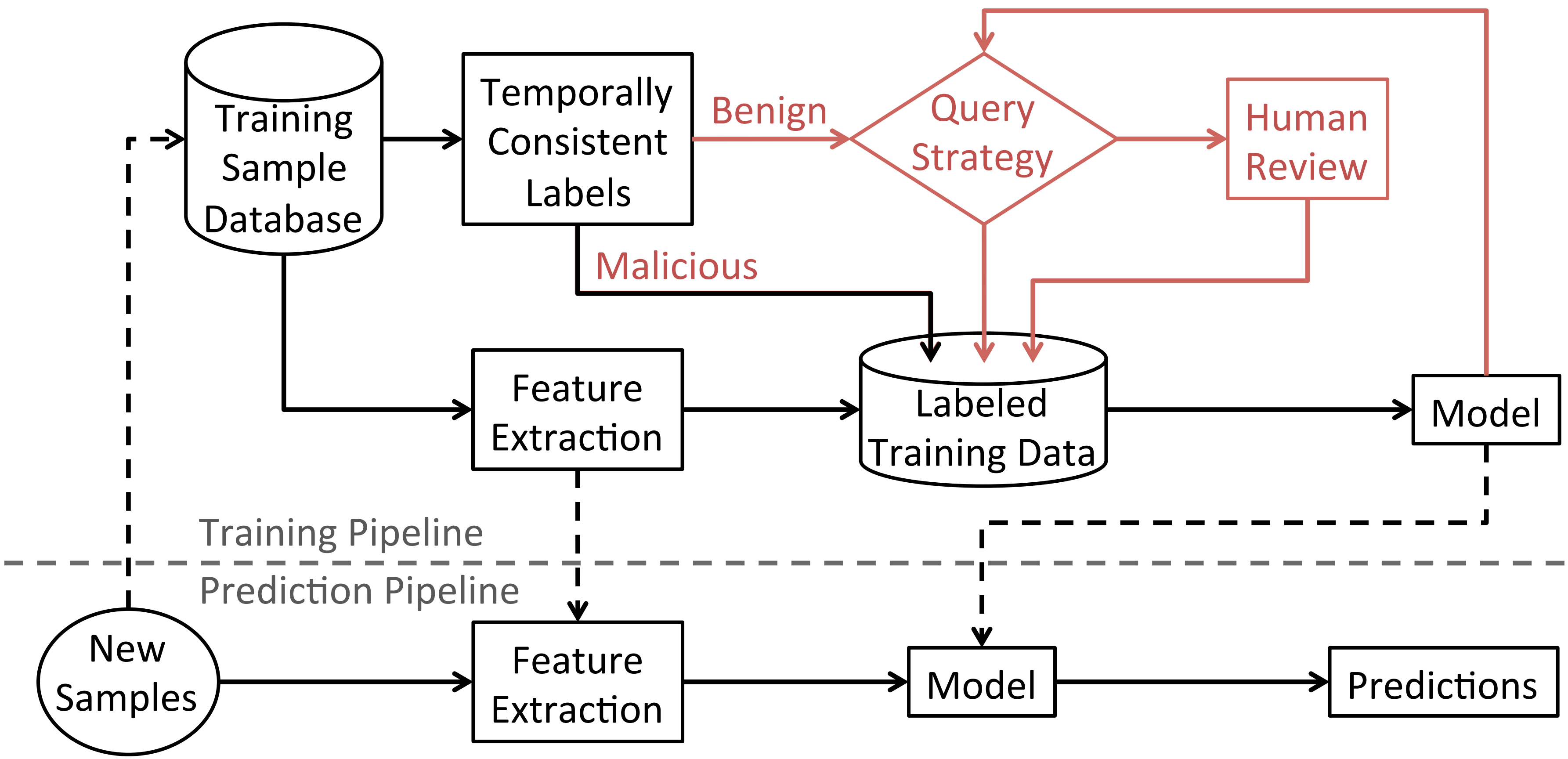
Design Overview



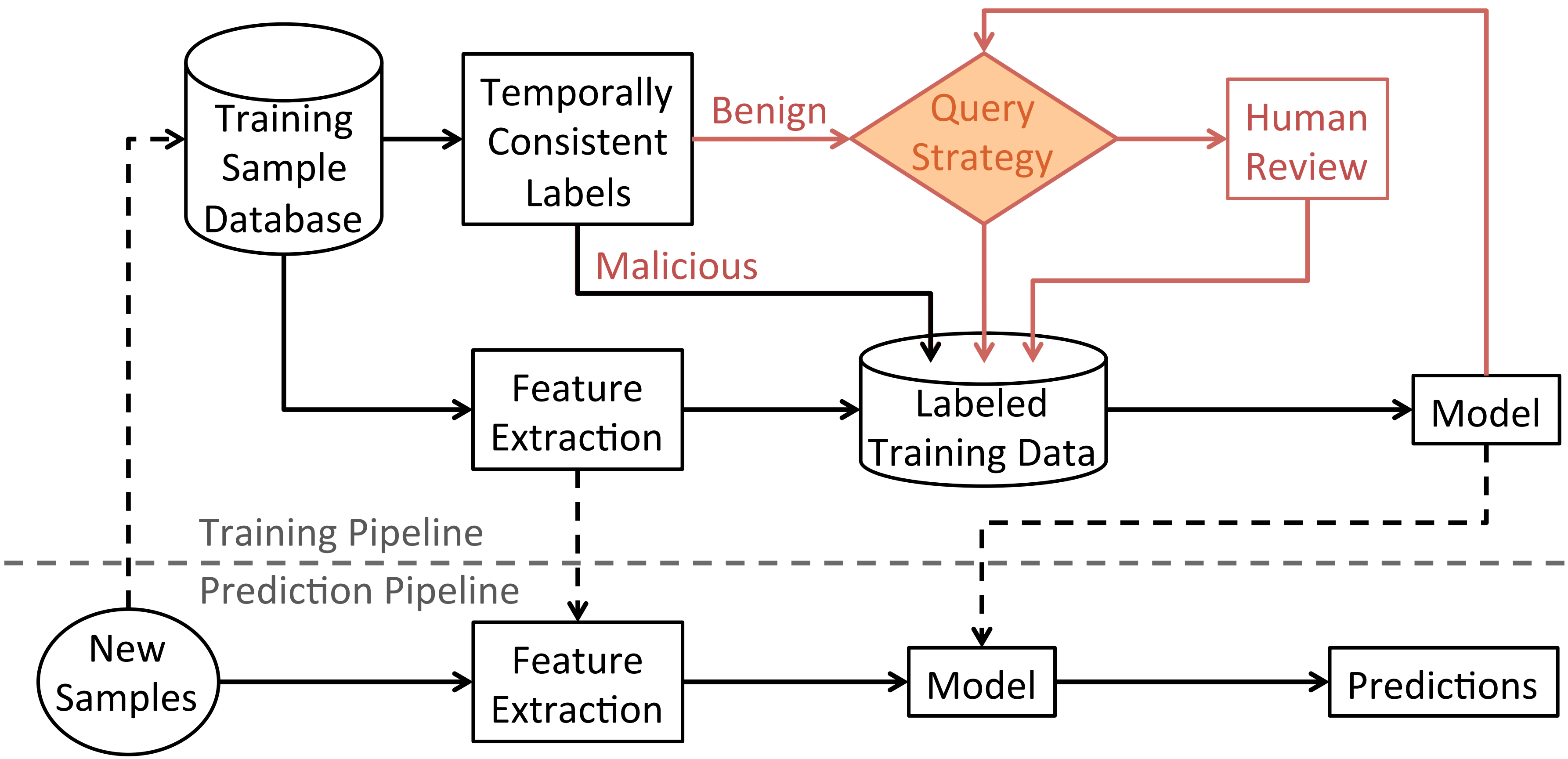
Design Overview



Design Overview

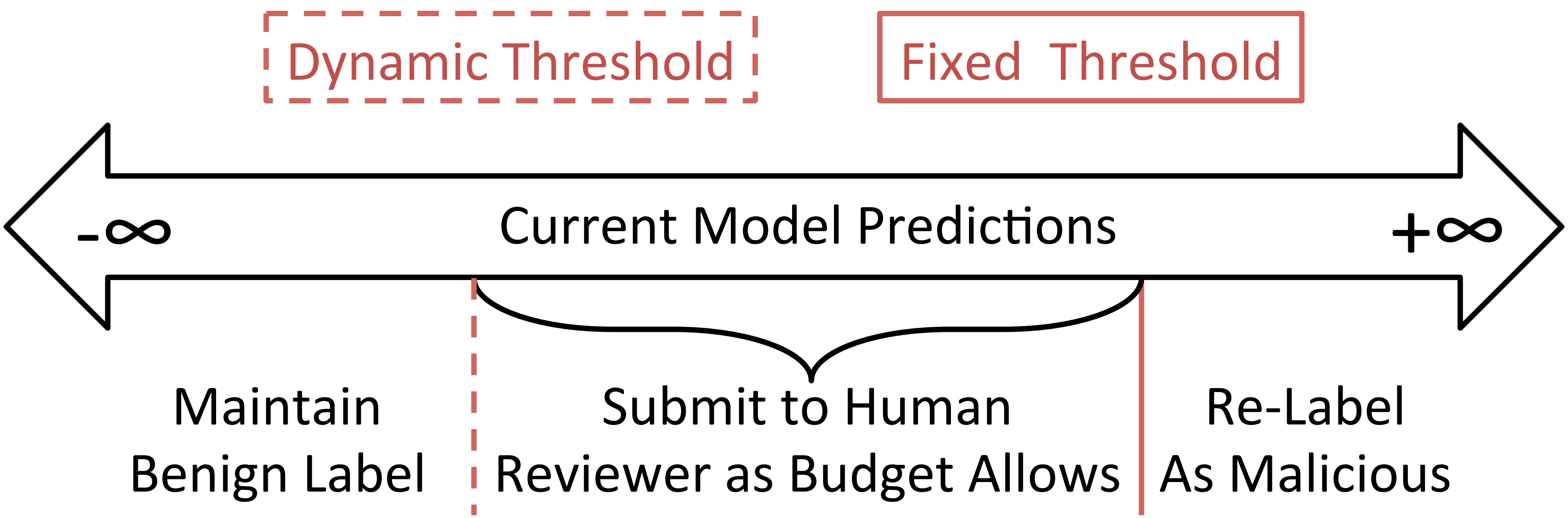


Design Overview



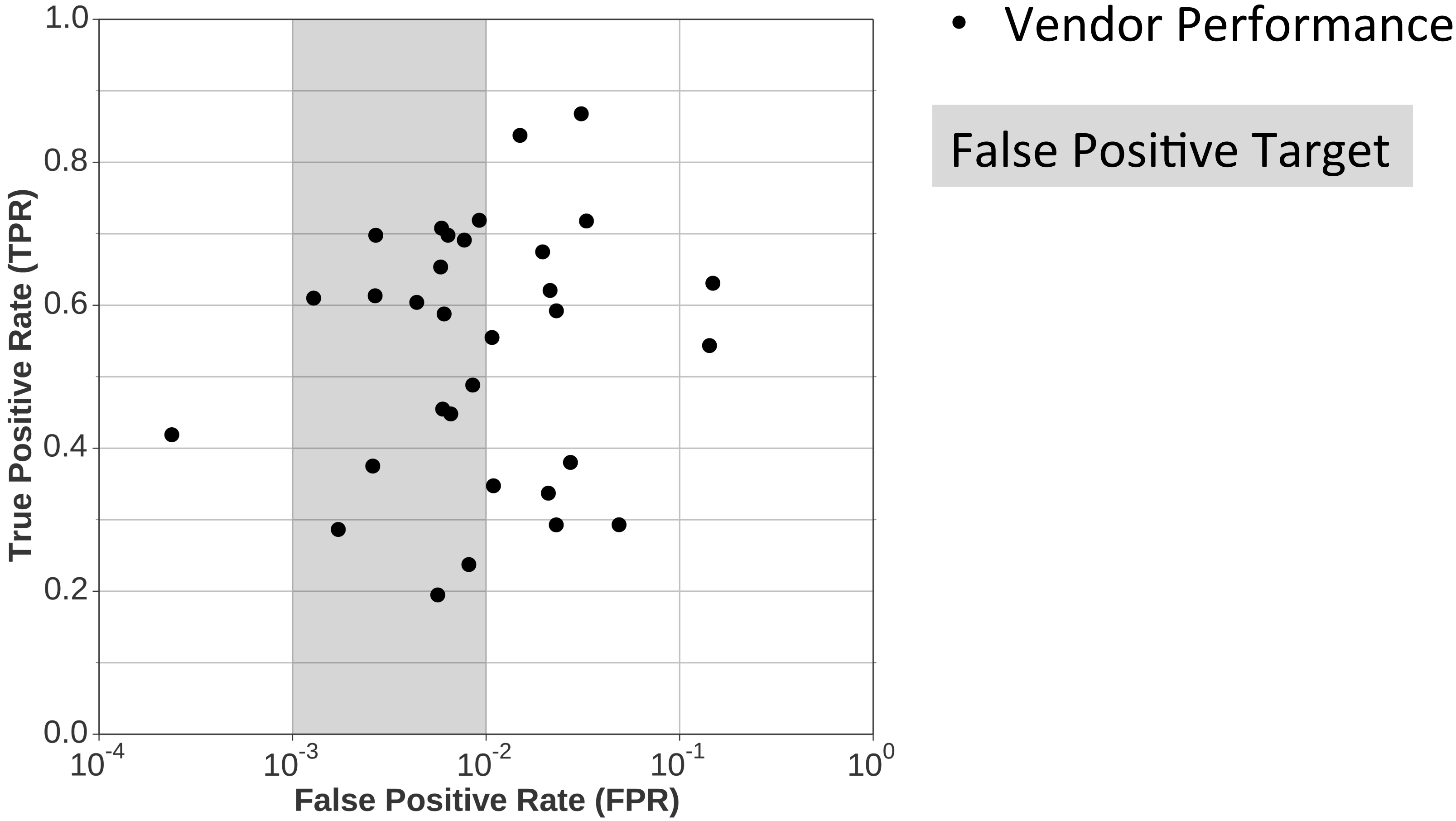
Reviewer Query Strategy

- Score candidate samples with current model
- Submit samples as query budget allows

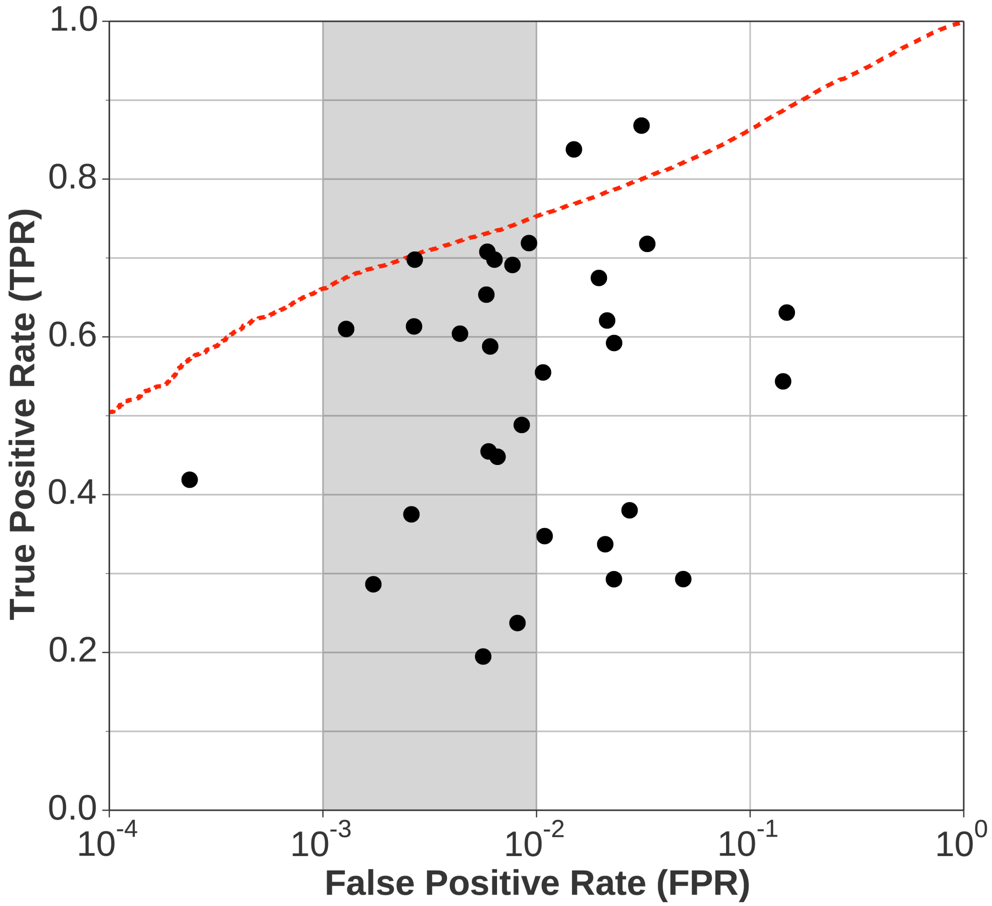


EXPERIMENTAL RESULTS

Performance Overview



Performance Overview

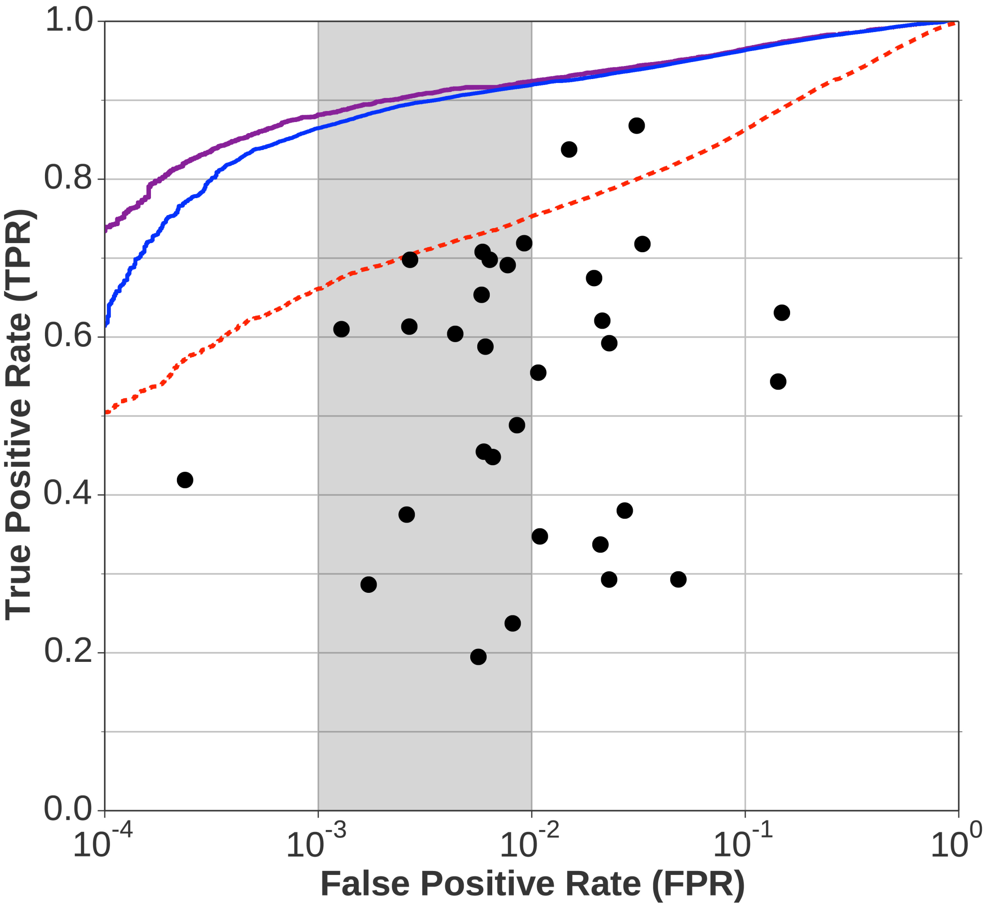


- Vendor Performance

False Positive Target

Online: Temporally Consistent (0 reviews)

Performance Overview



- Vendor Performance

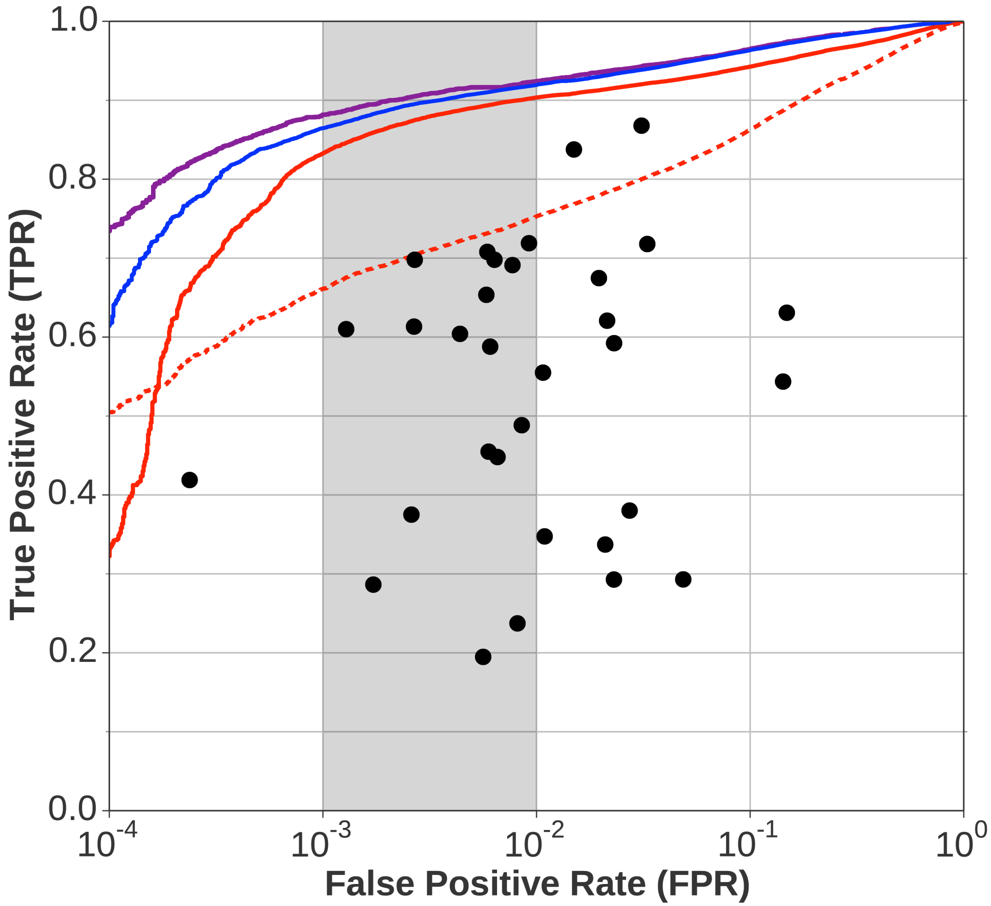
False Positive Target

Online: Temporally Consistent (0 reviews)

Offline: Random Division

Offline: Temporal Samples

Performance Overview



- Vendor Performance

False Positive Target

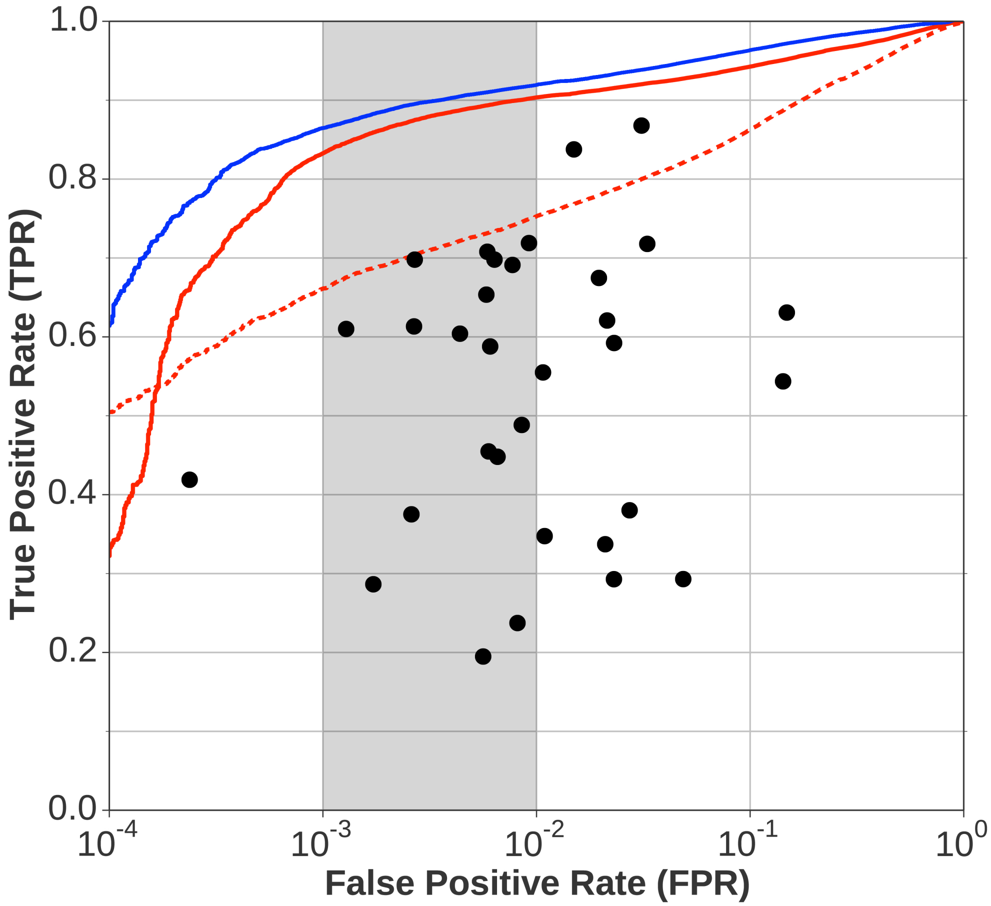
Online: Temporally Consistent (0 reviews)

Offline: Random Division

Offline: Temporal Samples

Online: Temporally Consistent (80 reviews/day)

Performance Overview



- Vendor Performance

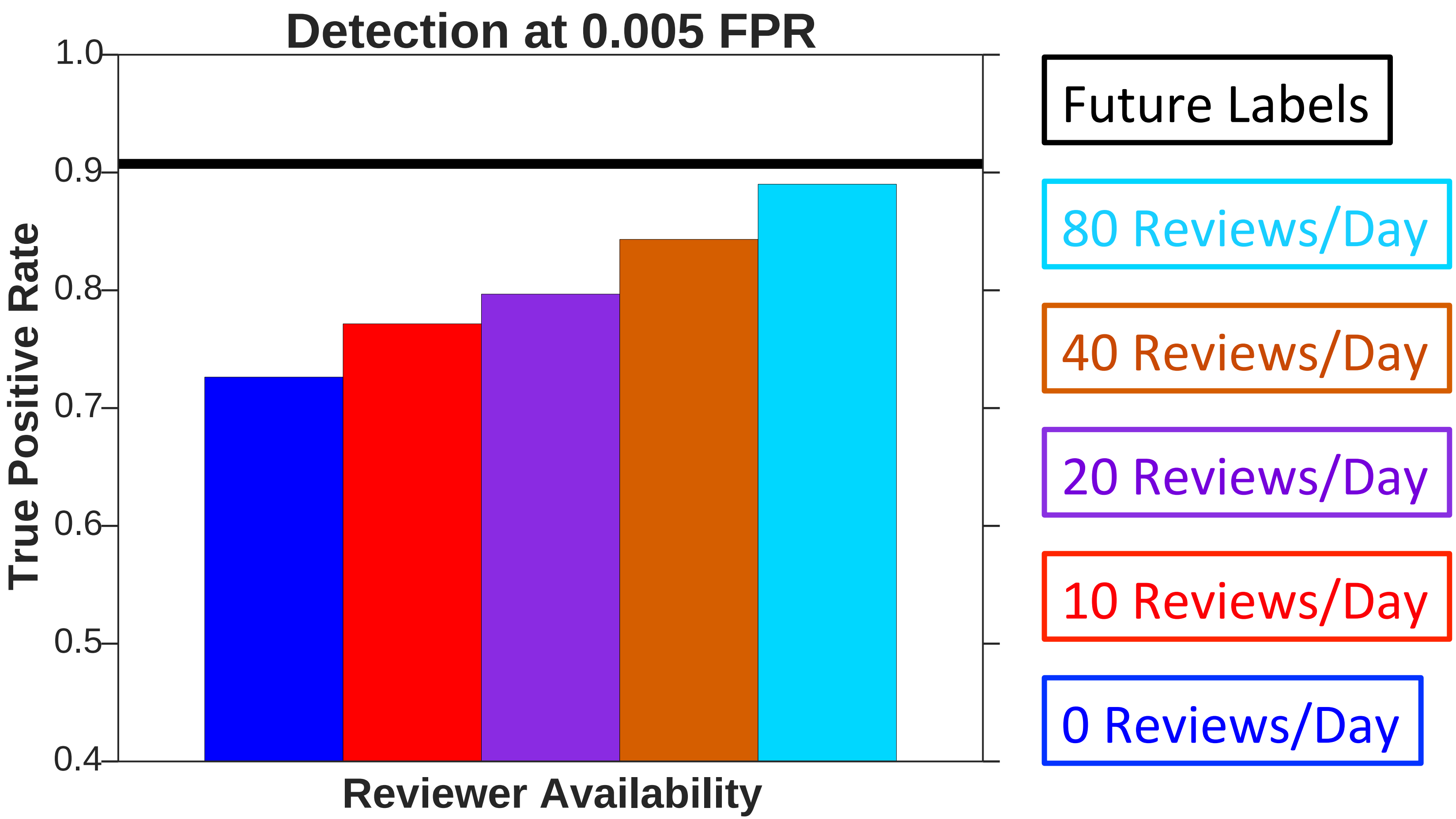
False Positive Target

Online: Temporally Consistent (0 reviews)

Offline: Temporal Samples

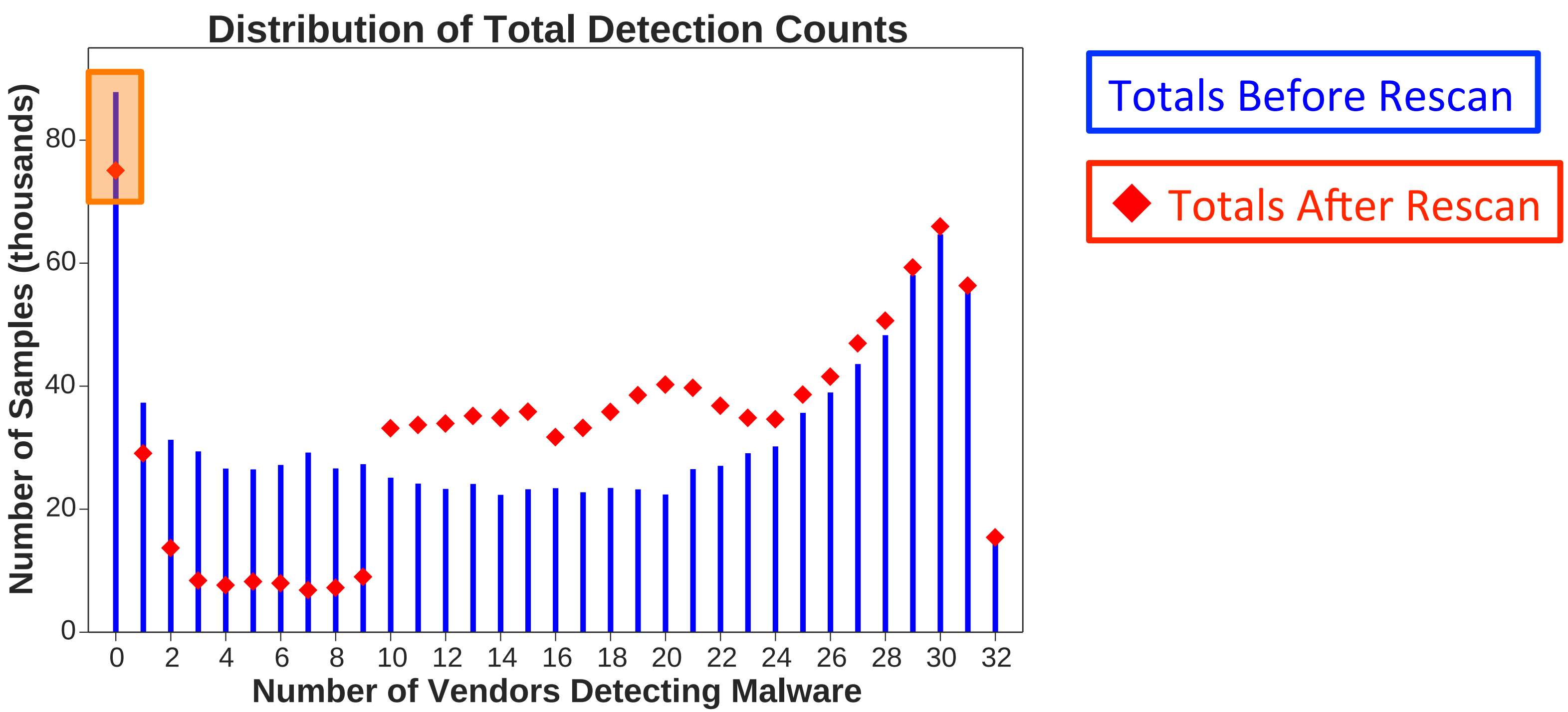
Online: Temporally Consistent (80 reviews/day)

Impact of Reviewer Queries



Catching Undetected Malware

- ML + reviewers increases detector robustness
- Detects 42% of previously undetected malware



Open Source & Data Release

- Modular design facilitates future work
  - Portable across application domains
  - Agnostic to learning algorithm and label source
- Scales well to large amounts of data
  - 778GB of raw data in ~12 hours with 40 cores
  - Apache Spark manages computation
- Data release enables reproducible results
  - 3% of our entire data set
  - List of all hashes

CONCLUSION

Key Results

- Account for industry performance gap
  - Offer improved technique for academic evaluation
- Offer solution to improve performance gap
  - Increases detection from 72% to 89%
  - Detects 42% of previously undetected malware
- Publicly release implementation and data