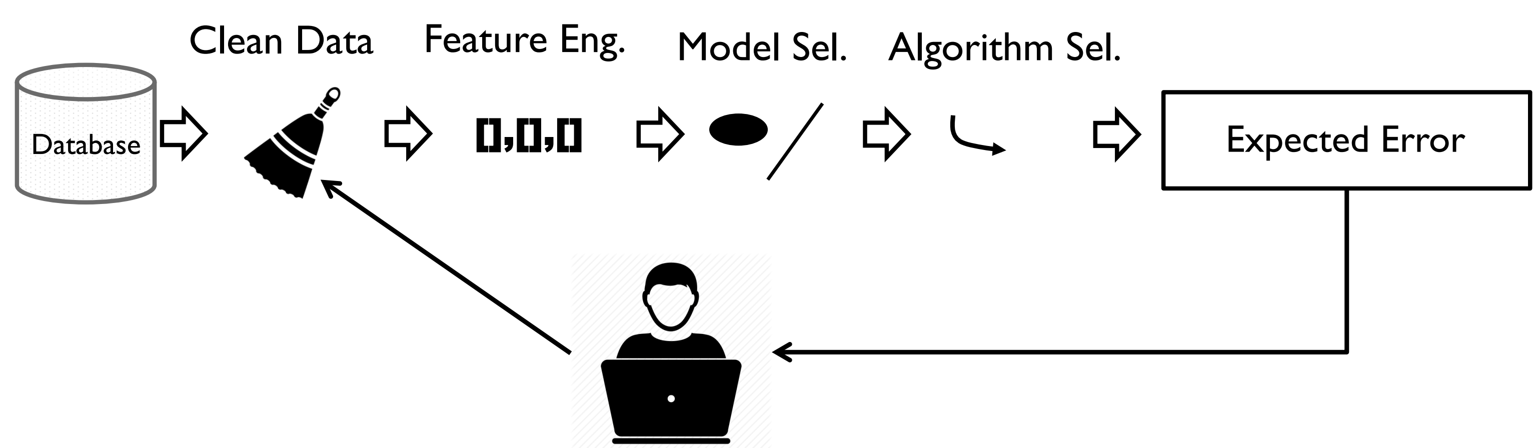
ActiveClean: Interactive Data Cleaning For Statistical Modeling



Sanjay Krishnan, Jiannan Wang, Eugene Wu,
Michael J. Franklin, Ken Goldberg

amplab

Berkeley
Artificial Intelligence Research Laboratory

Large Datasets, Sophisticated Models

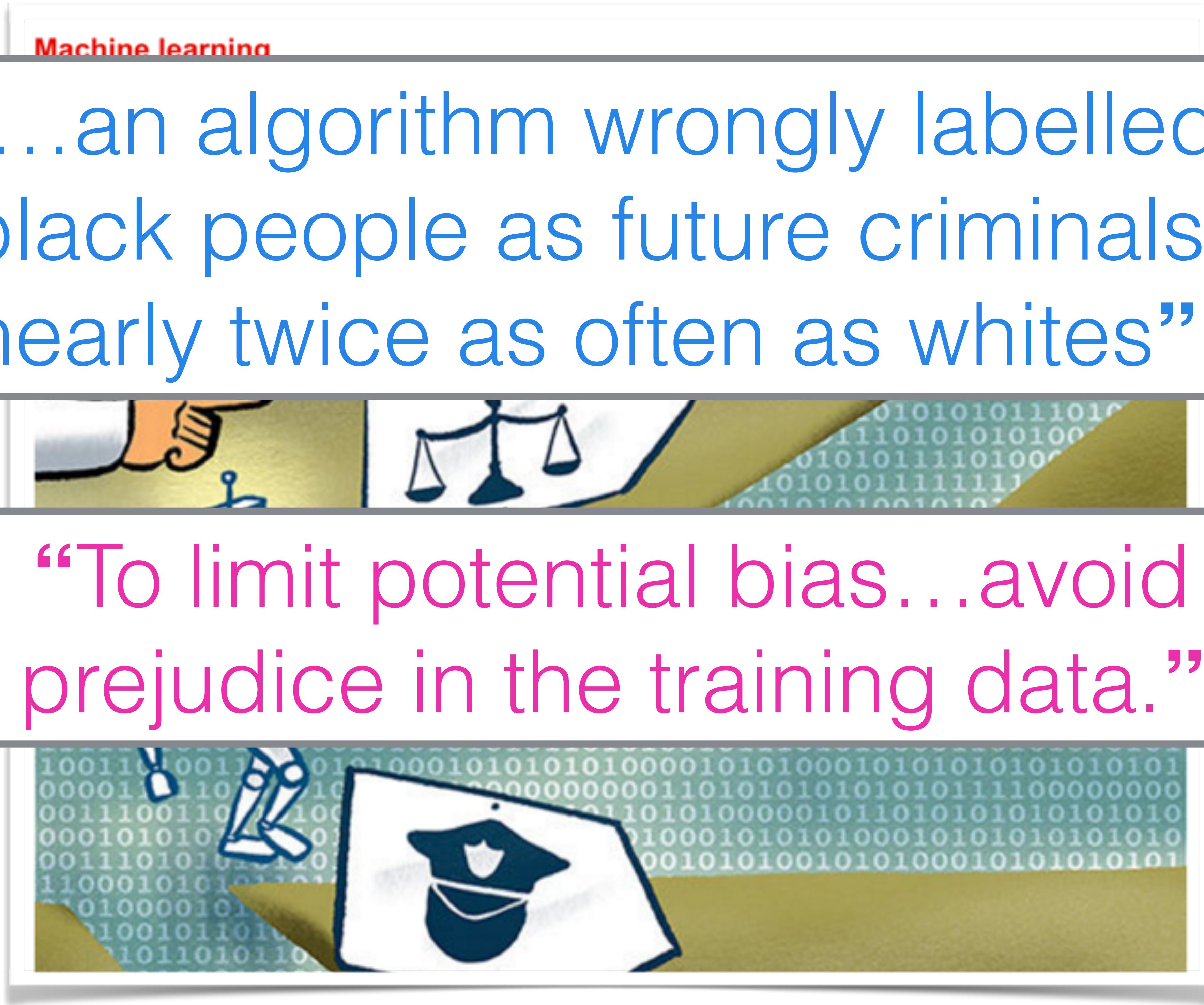
Spark MLlib

KeystoneML

TensorFlow™

ML

Biased Data = Biased Models



“…an algorithm wrongly labelled black people as future criminals nearly twice as often as whites”

“To limit potential bias…avoid prejudice in the training data.”

For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights

By STEVE LOHR AUG. 17, 2014

Email

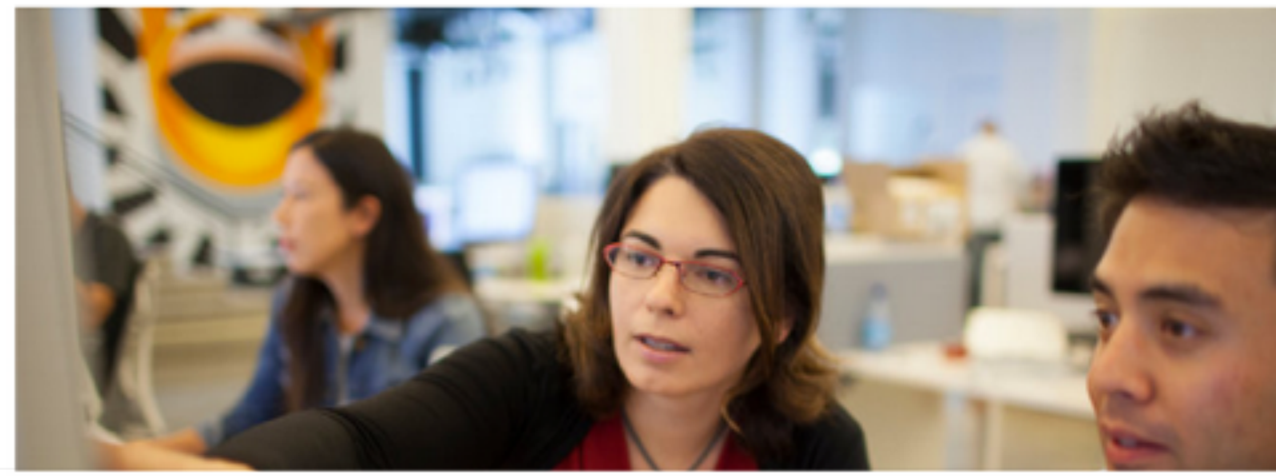
Share

Tweet

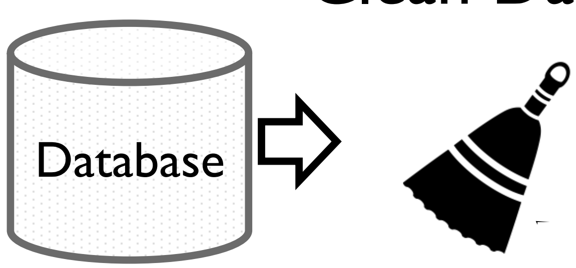
Save

Technology revolutions come in measured, sometimes foot-dragging steps. The lab science and marketing enthusiasm tend to underestimate the bottlenecks to progress that must be overcome with hard work and practical engineering.

The field known as "big data" offers a contemporary case study. The catchphrase



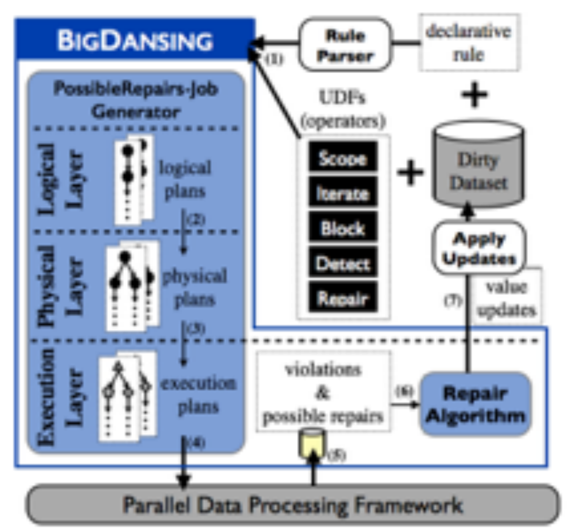
Clean Data



Data Cleaning Is Expensive

[1] Data Analyst Effort

[2] Crowdsourcing

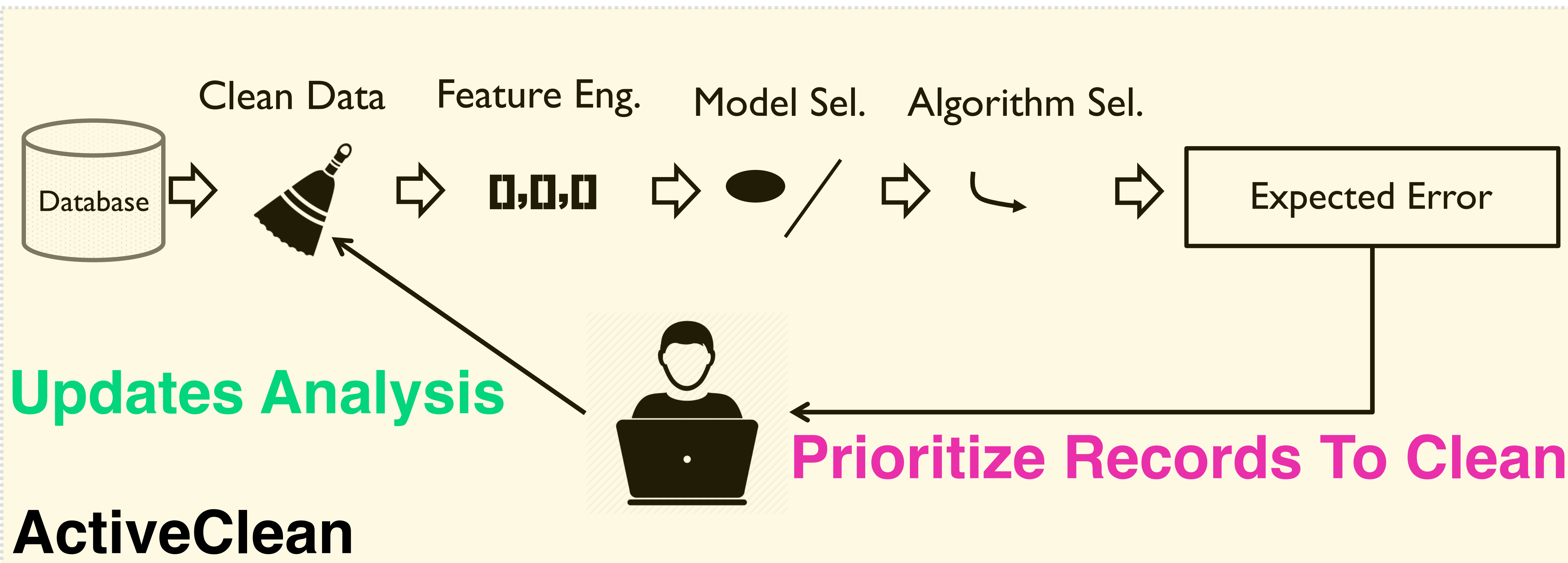


[3] Computational Cost

[1] Krishnan, Sanjay, et al. "Towards reliable interactive data cleaning: a user survey and recommendations." HILDA@SIGMOD. 2016.
[2] Marcus, Adam, and Aditya Parameswaran. "Crowdsourced data management industry and academic perspectives." Foundations and Trends in Databases 2015.
[3] Khayyat, Zuhair, et al. "Bigdansing: A system for big data cleansing." SIGMOD. 2015.

ActiveClean

- How do we most efficiently clean data for a given machine learning task?



Problem Statement

Given a convex loss minimization problem and a cleaning function C() which can only be applied to k records.

Find the best estimate of the true model (where the full dataset is hypothetically cleaned).

Convex Loss Minimization

- SVMs, Linear Regression, Logistic Regression
- (xi, yi) is a labeled tuple where x is a feature vector and y is a label.
- Find a parameter that minimize disagreement with the true label.

$$\theta^* = \arg\min_{\theta} \sum_{i=1}^{N} \phi(x_i, y_i, \theta)$$

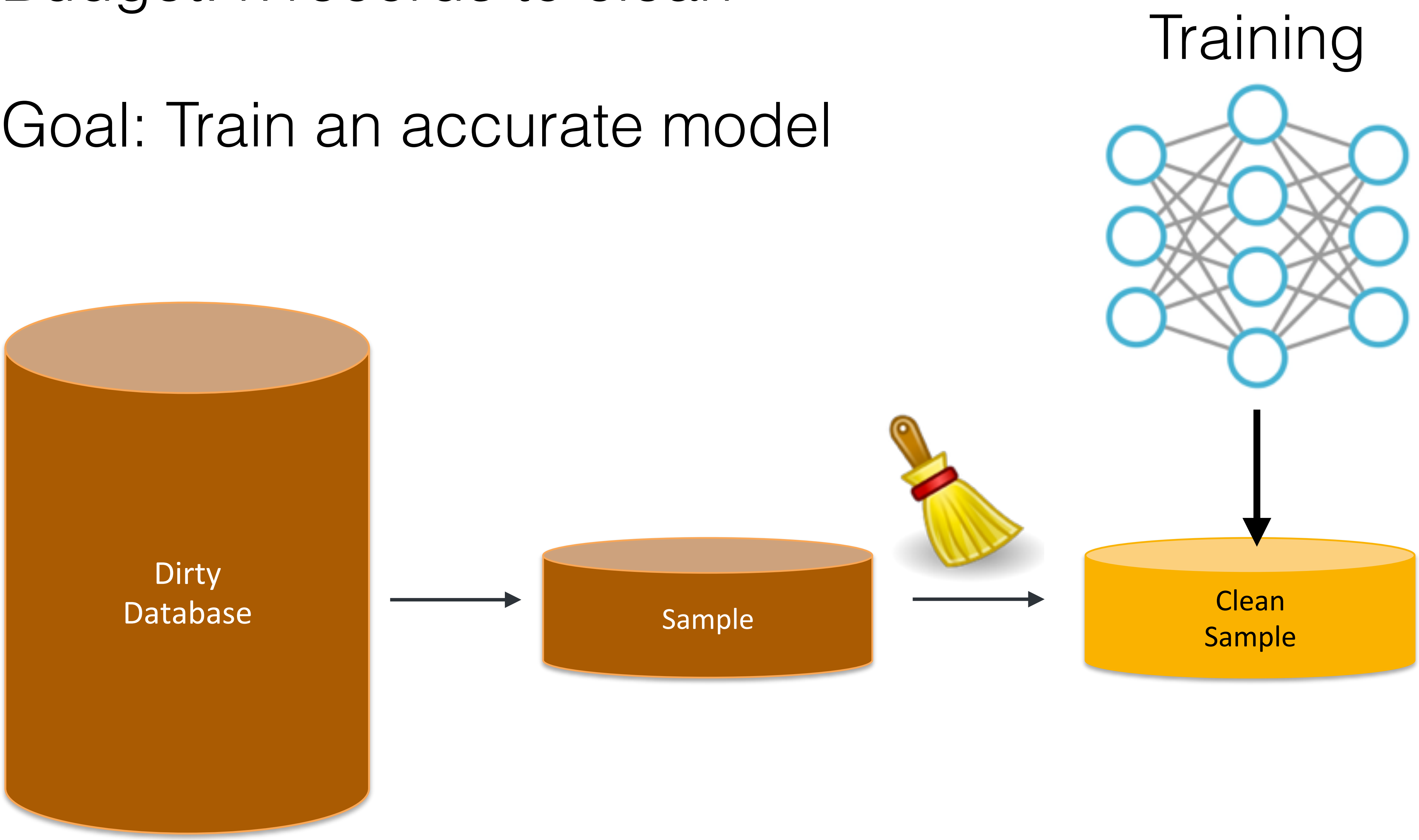
Outline

- Motivation
- **The Update Problem**
- The Prioritization Problem
- Results

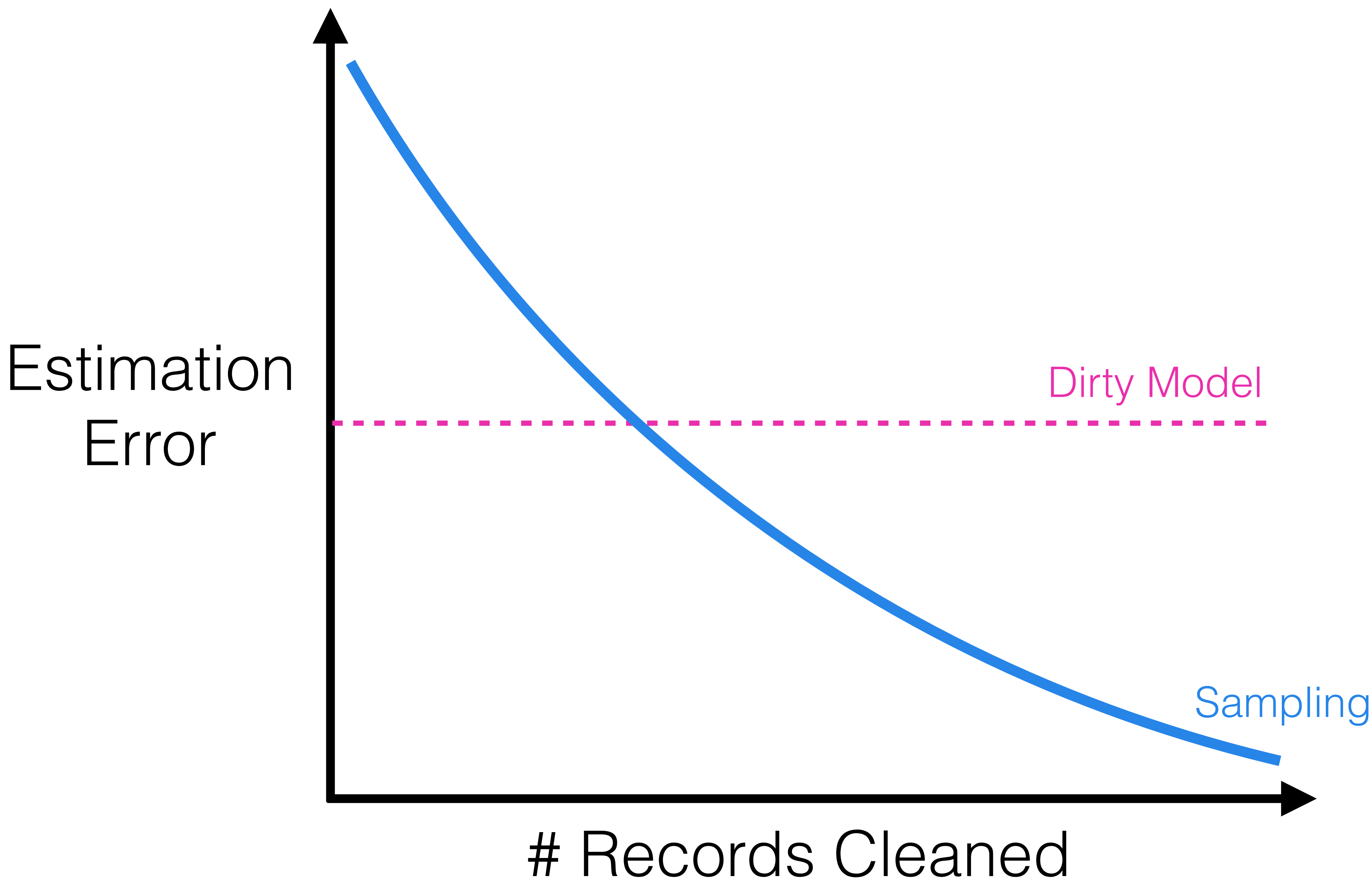
Idea 1. Sampling

Budget: k records to clean

Goal: Train an accurate model



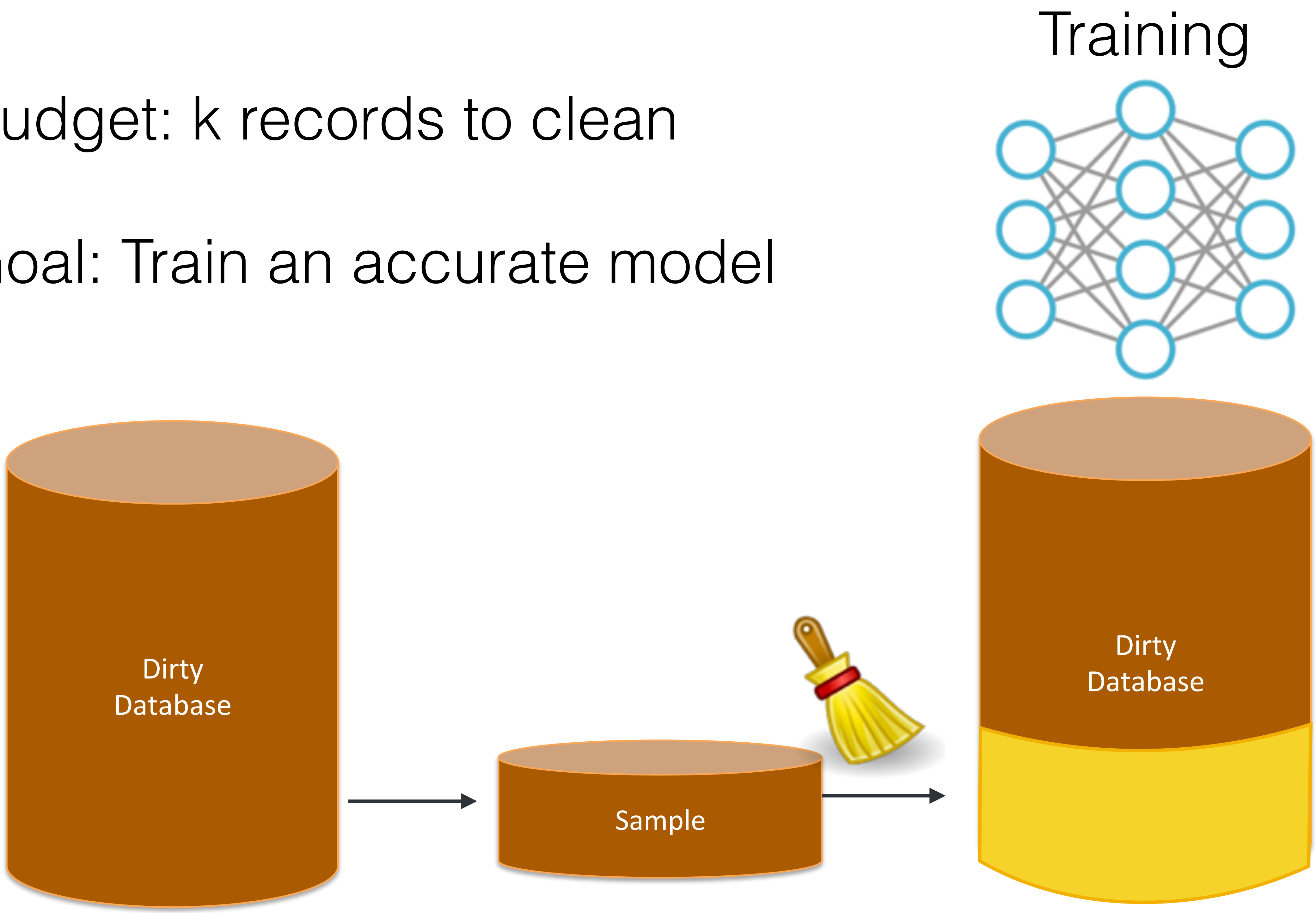
Problem. Sampling Error



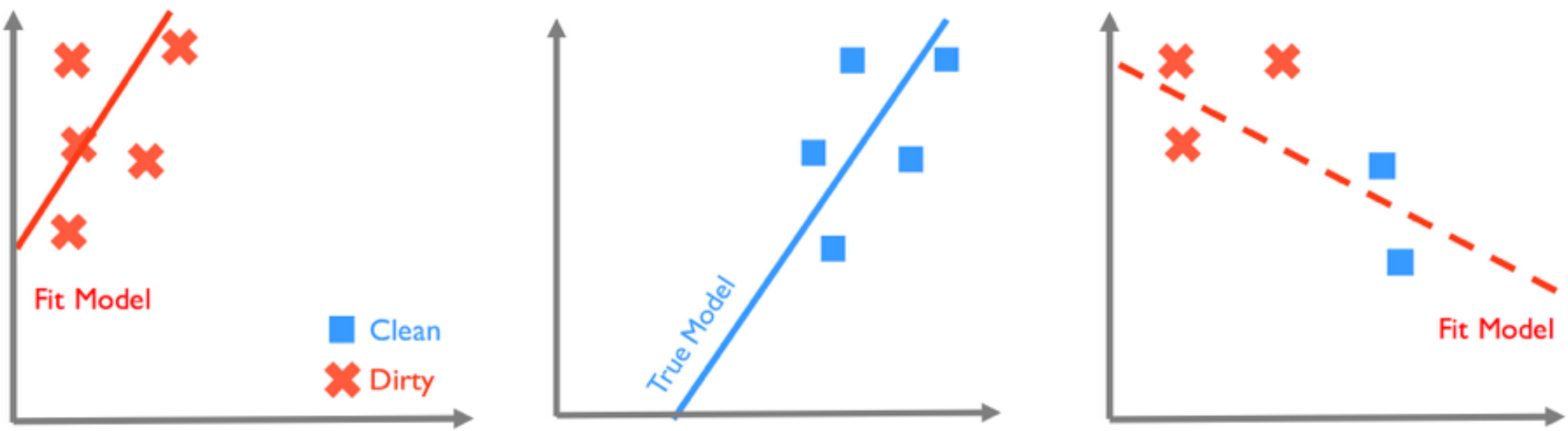
Idea 2. Clean In Place

Budget: k records to clean

Goal: Train an accurate model



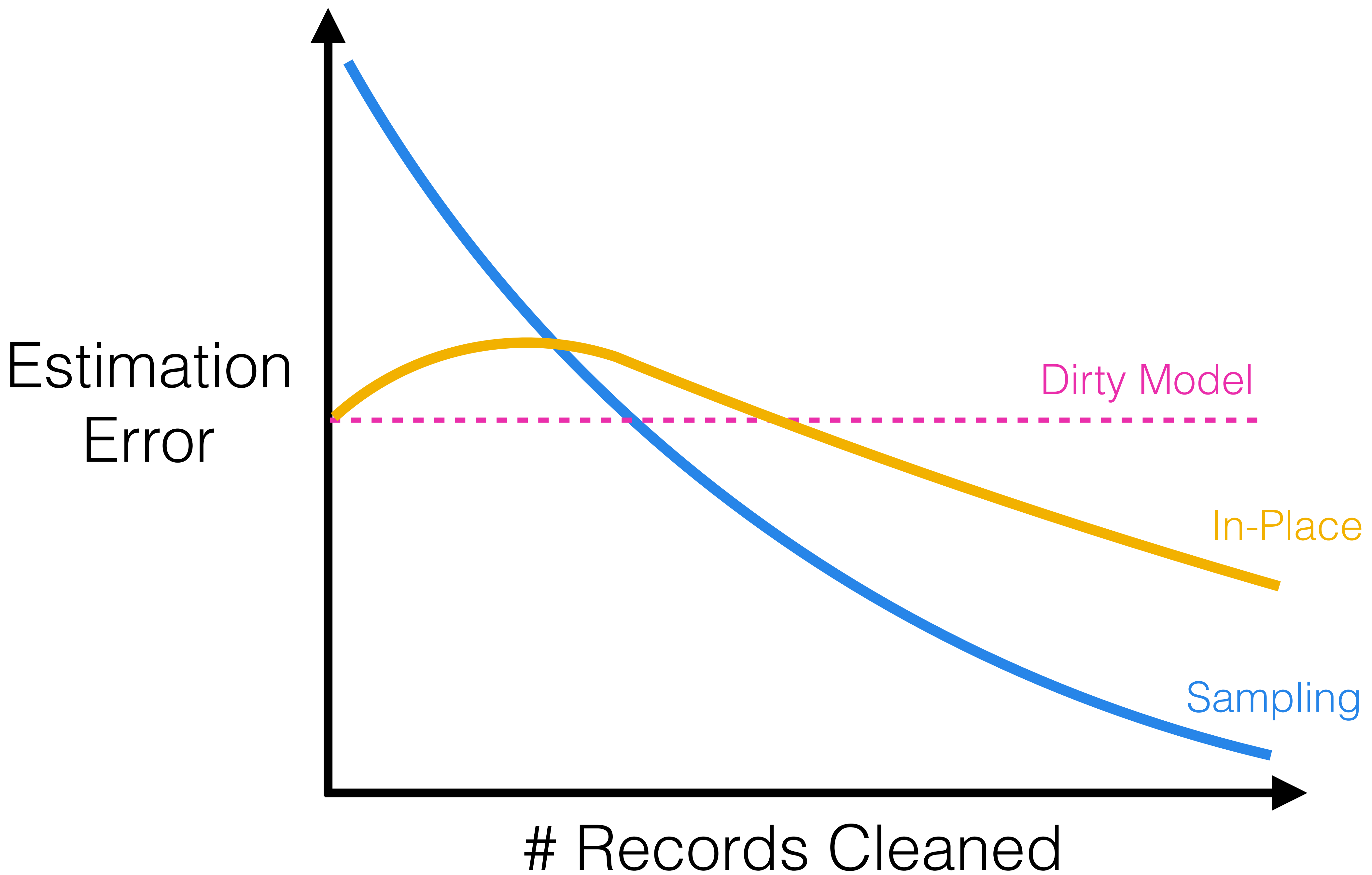
Problem. Simpson's Paradox



(a) Full Dirty Data (b) Full Cleaned Data (c) Mixed Dirty and Clean

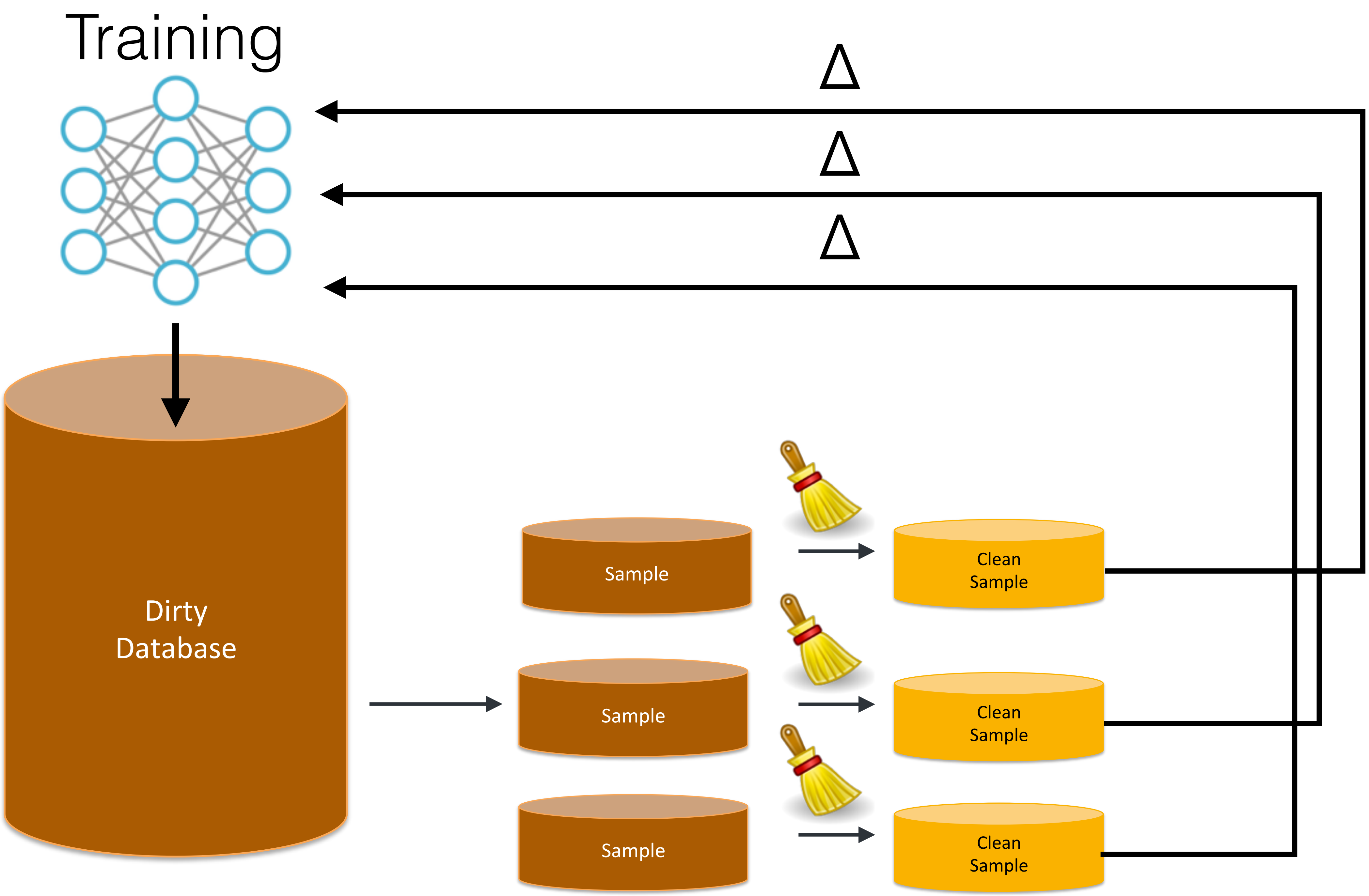
Partial Data Cleaning Can Be Misleading

Problem. Simpson's Paradox

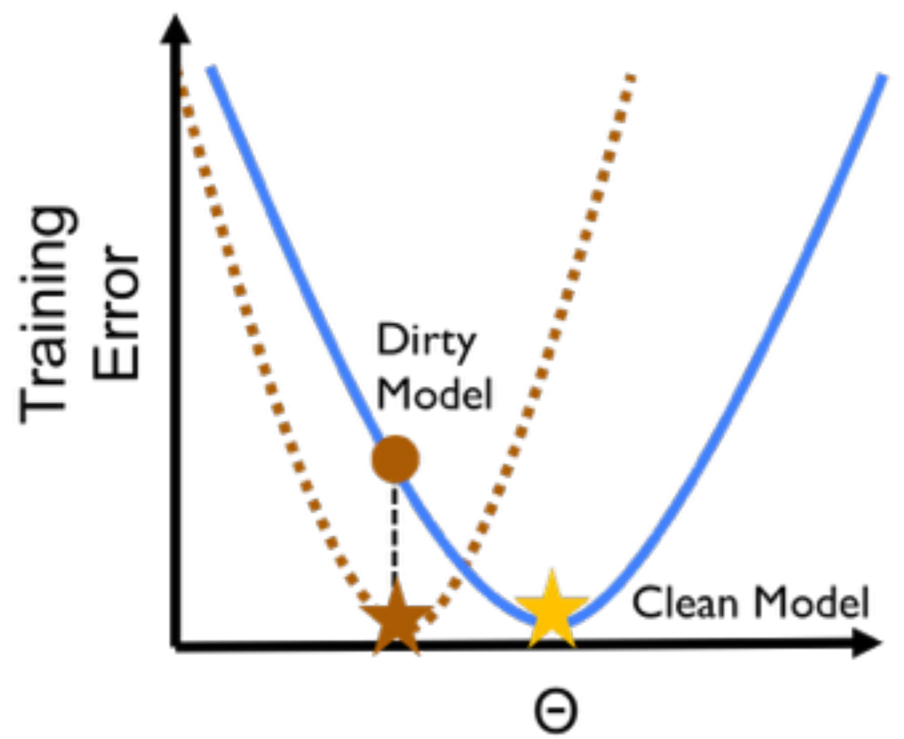


Active Clean

Model as incremental optimization



Intuition

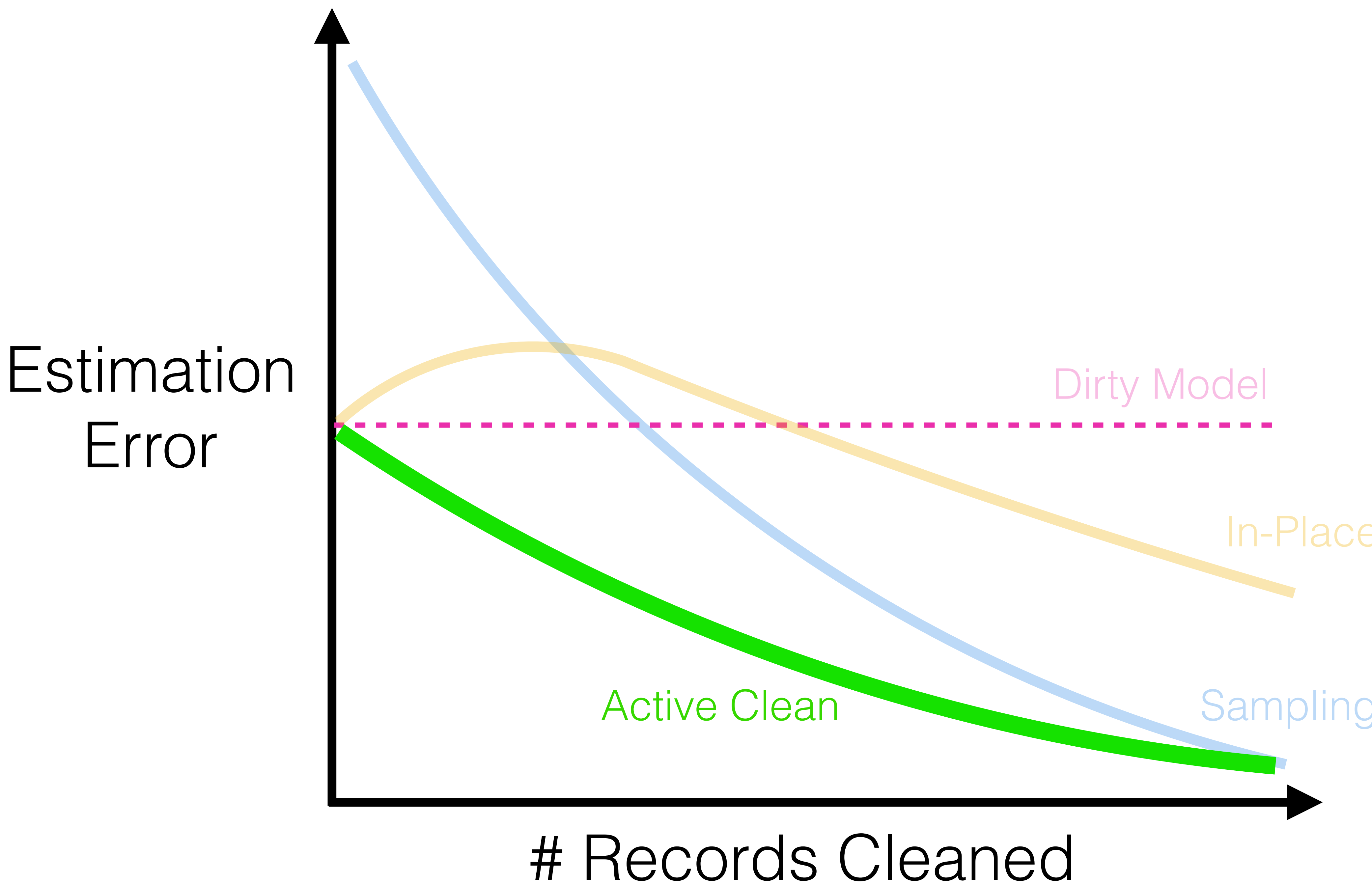


- Stochastic Gradient Descent.

$$\theta^{(t+1)} \leftarrow \theta^{(t)} - \gamma \cdot E[\nabla\phi(\theta^{(t)})]$$

- Make each step unbiased.

Intuition


Estimation Error
Dirty Model
In-Place
Active Clean
Sampling
Records Cleaned

Analysis

For a batch size b and iterations T, the ActiveClean stochastic gradient descent updates converge with rate:

$$O(\frac{1}{\sqrt{bT}})$$

For strongly-convex models:

$$O(\frac{1}{T\sqrt{b}})$$

For L-Lipschitz loss (e.g., SVM):

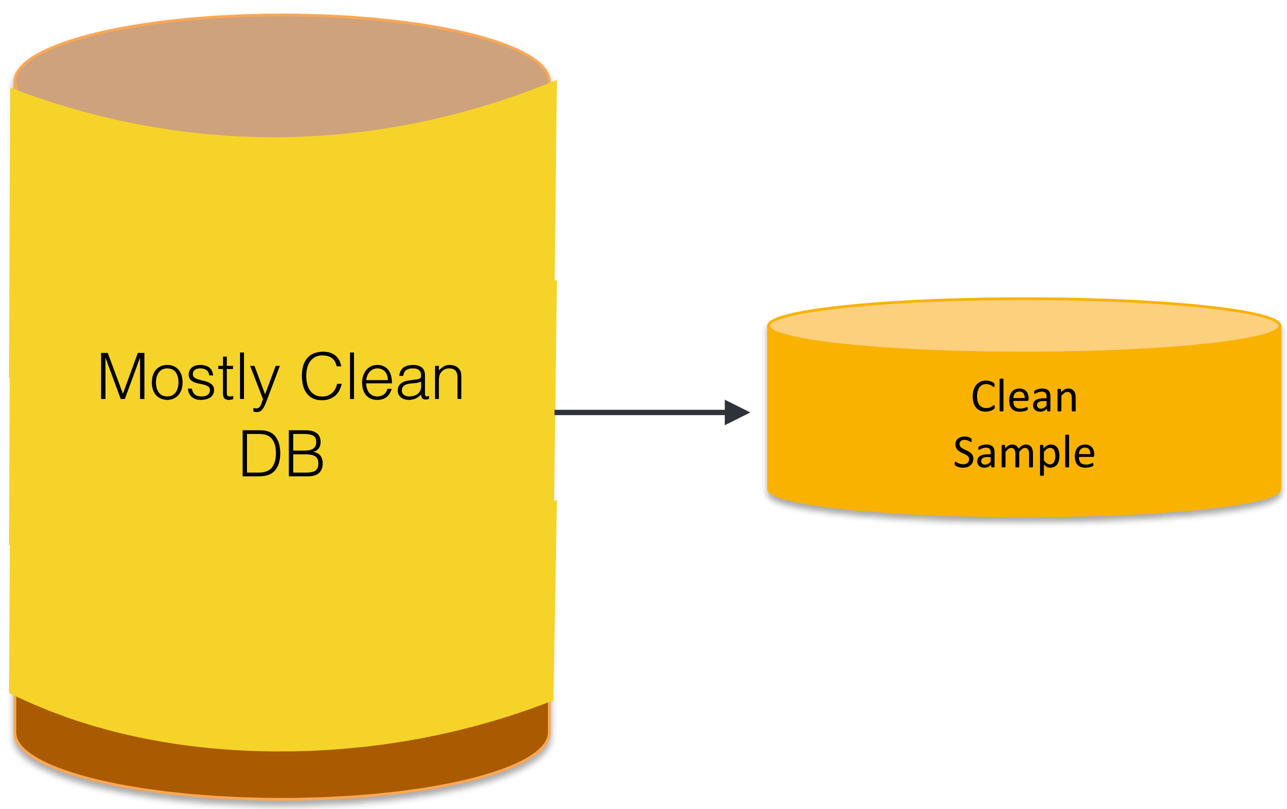
$$O(\frac{L}{\sqrt{bT}})$$

Outline

- Motivation
- The Update Problem
- **The Prioritization Problem**
- Results

Sparsity of Errors

- Uniform random sampling is not efficient for sparse errors.
- Rare errors can amplify



Data Likely To Be Dirty

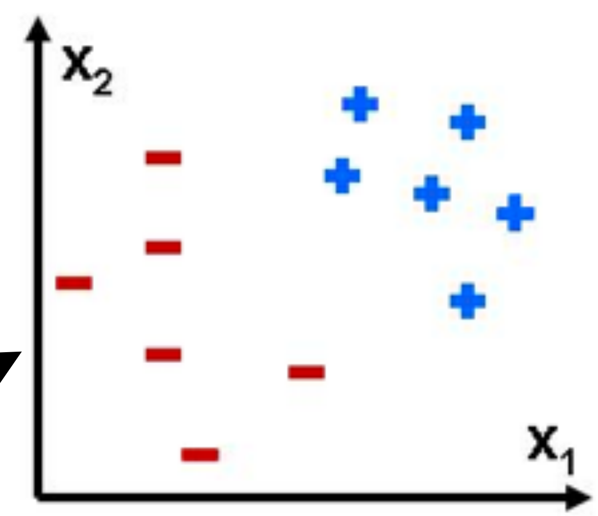
- If most of the dataset is clean, random sampling will result in wasted effort.

- Active Clean integrates with detection techniques

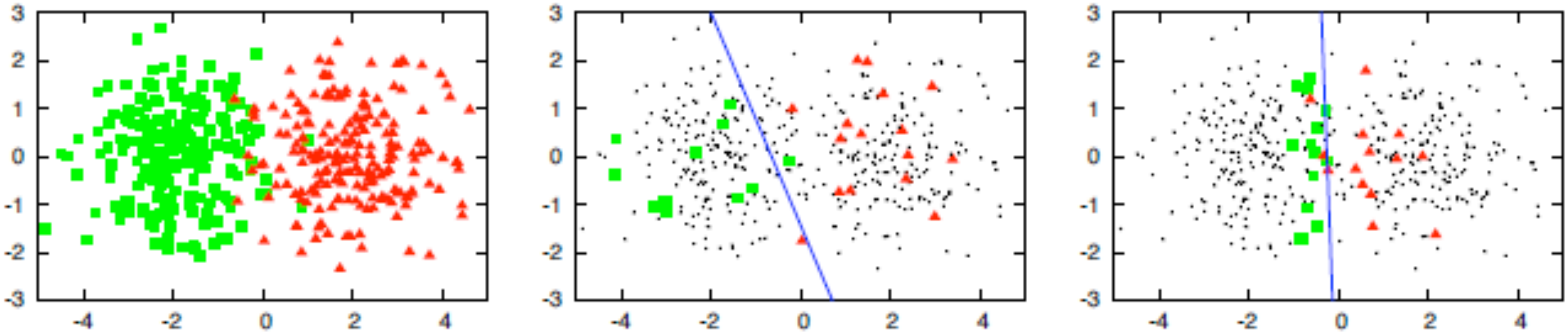
Rules

$$\exists\ \boldsymbol{\sigma} \in \boldsymbol{\Sigma} : \boldsymbol{\sigma}(\mathbf{r}) = \text{True}$$

Learning



Data Valuable To The Model



(a) (b) (c)

- Some data points are more valuable to the model

Non-Uniform Sampling

- Stochastic Gradient Descent.

$$\theta^{(t+1)} \leftarrow \theta^{(t)} - \gamma \cdot \underline{E[\nabla\phi(\theta^{(t)})]}$$

- Importance Sample: Expectations can be calculated over different distributions with the same support.

$$p_i \propto ||\nabla\phi(x_i, y_i, \theta^{(t)})||$$

- 2.5x improvement in experiments

Outline

- Motivation
- The Update Problem
- The Prioritization Problem
- **Results**

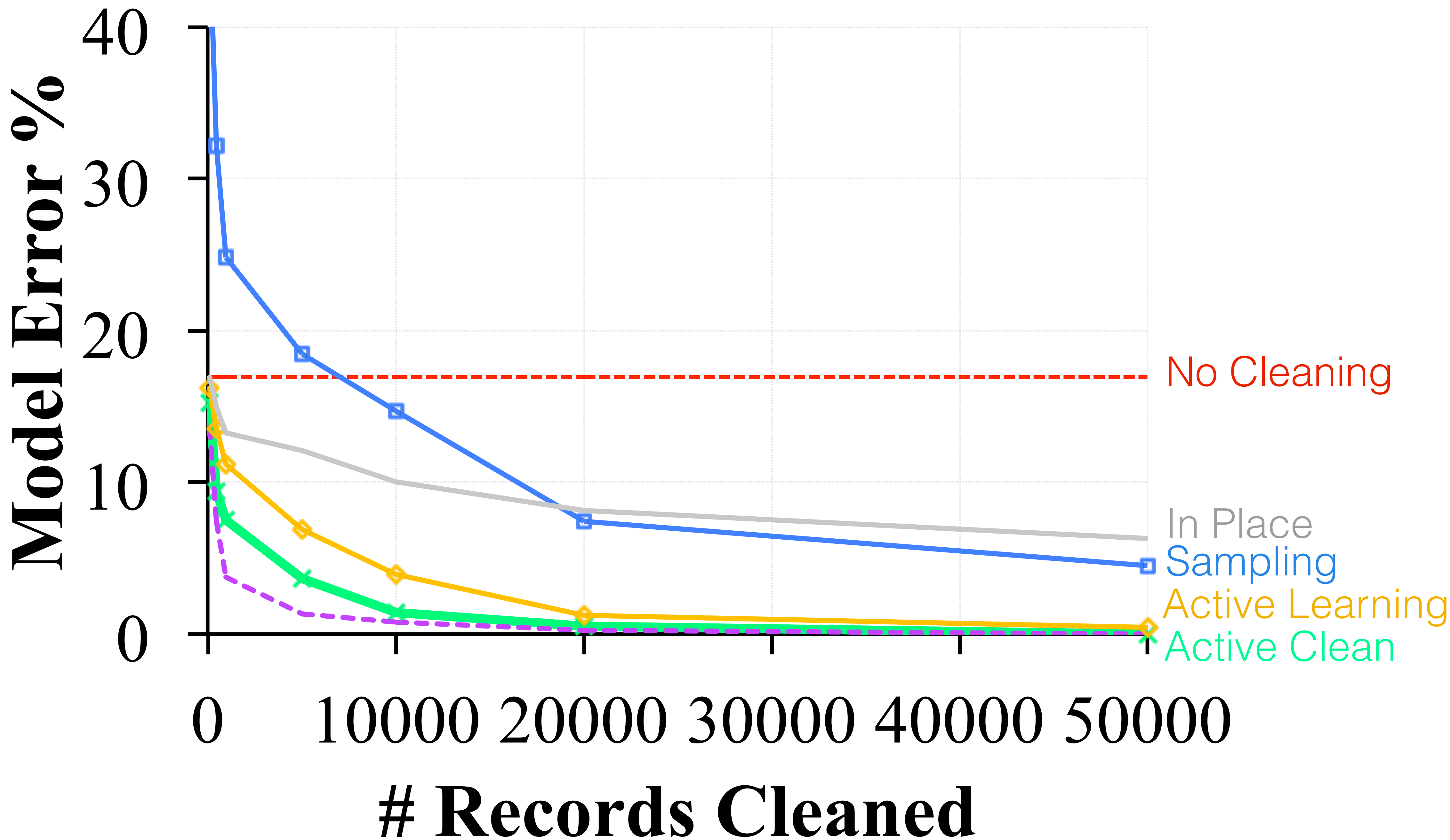
Experimental Setup

- Real datasets and real errors.
- Cleaned all of the errors up front, then simulated an analyst cleaning incrementally.
- Measured test and training error w.r.t true model

Dollars For Docs

- 250,000 medical contribution records
- Manually labeled as suspicious or not
- Entity resolution errors in company and drug names

Dollars For Docs

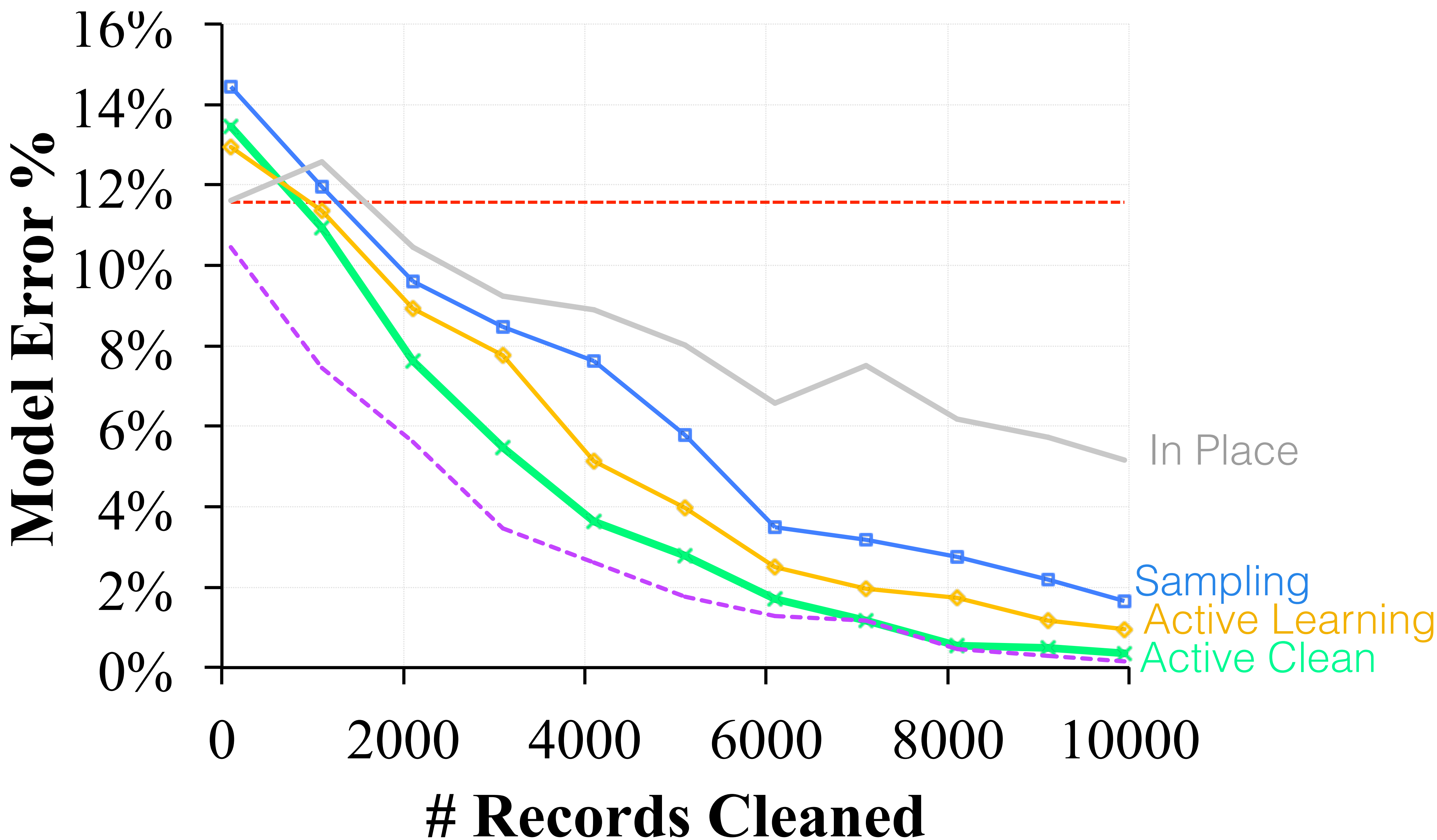


Yahoo Movies

- 900,000 Records of Plot Descriptions with Genres
- Classify Comedy vs. Horror

Bloodrage (1979) A psychotic killer stalks the streets of New York City. ***Comedy***

Yahoo Movies



Conclusion

- Machine Learning can be sensitive to dirty data when errors are systematic and unmodeled.
- Data cleaning is expensive so there is a question of how best to apply data cleaning for ML problems.
- Many open questions in future work.

sampleclean.org
sanjay@eecs.berkeley.edu