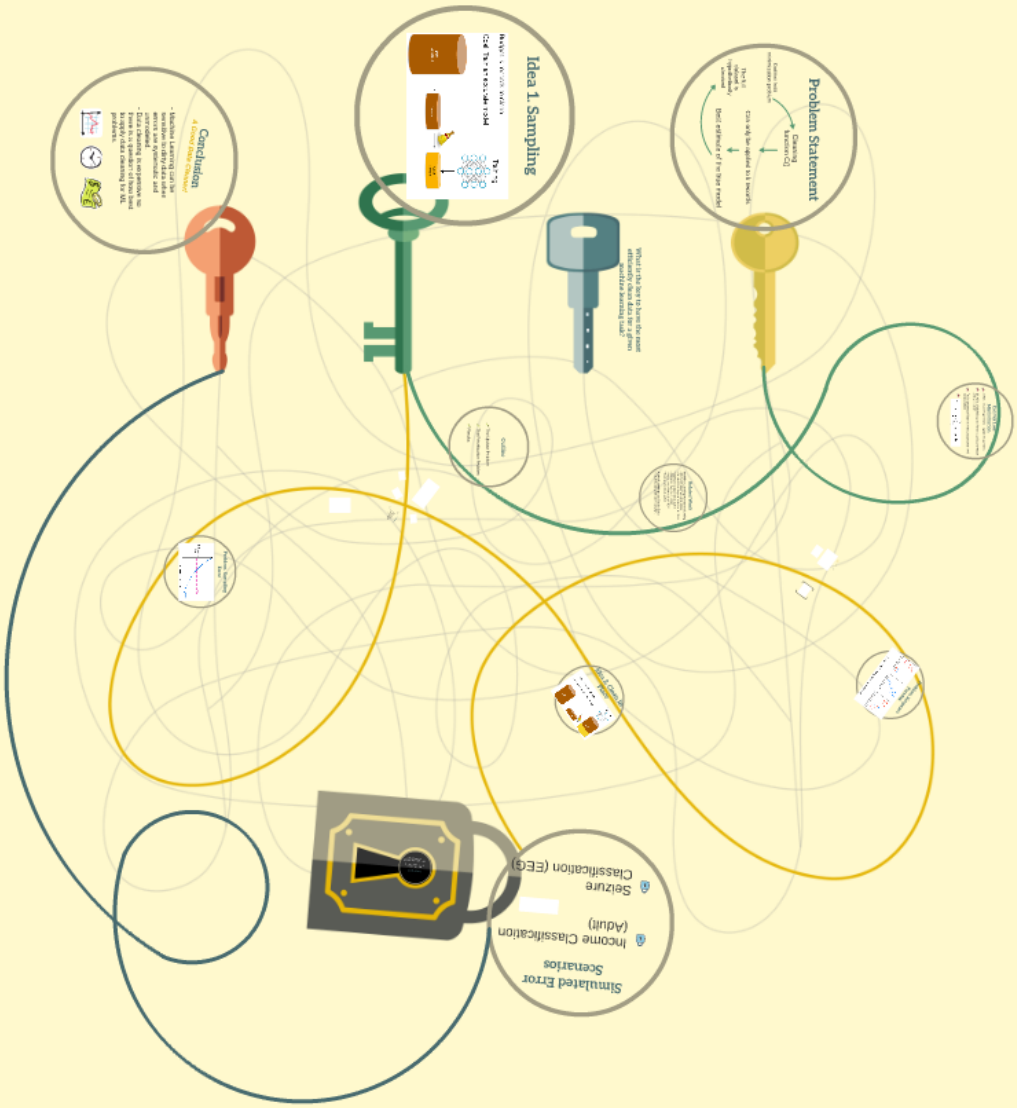
ActiveLearning

Interactive Data Cleaning For Statistical Modeling



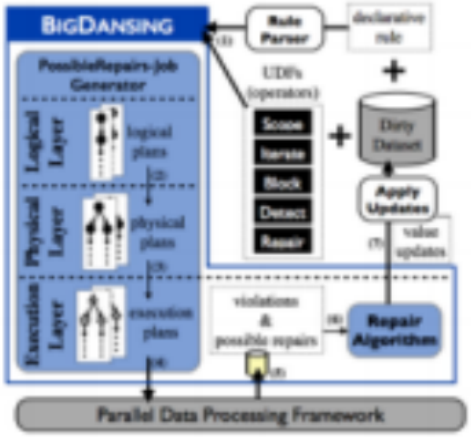

Prezi

Large Datasets, Sophisticated Models

Data Cleaning Is Expensive

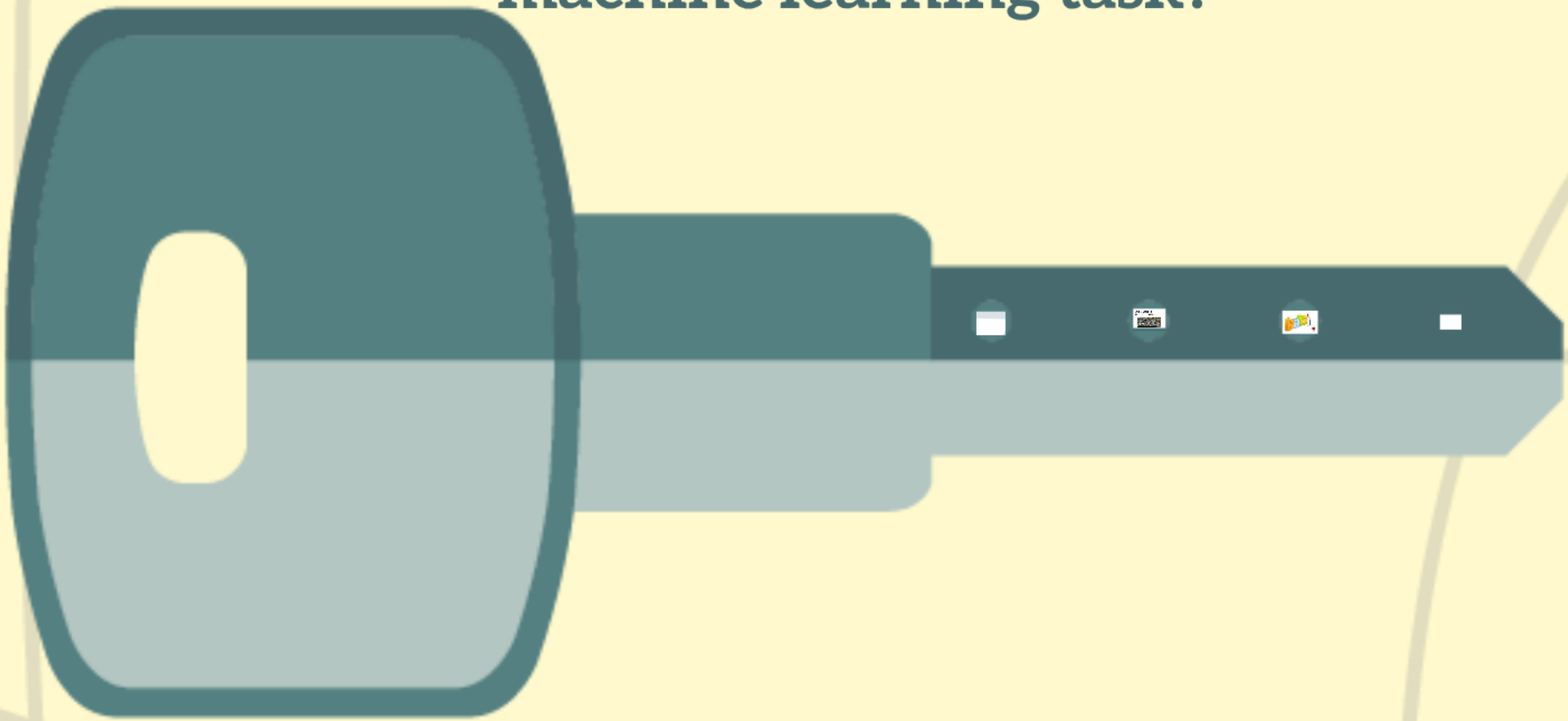
[1] Data Analyst Effort

[2] Crowdsourcing



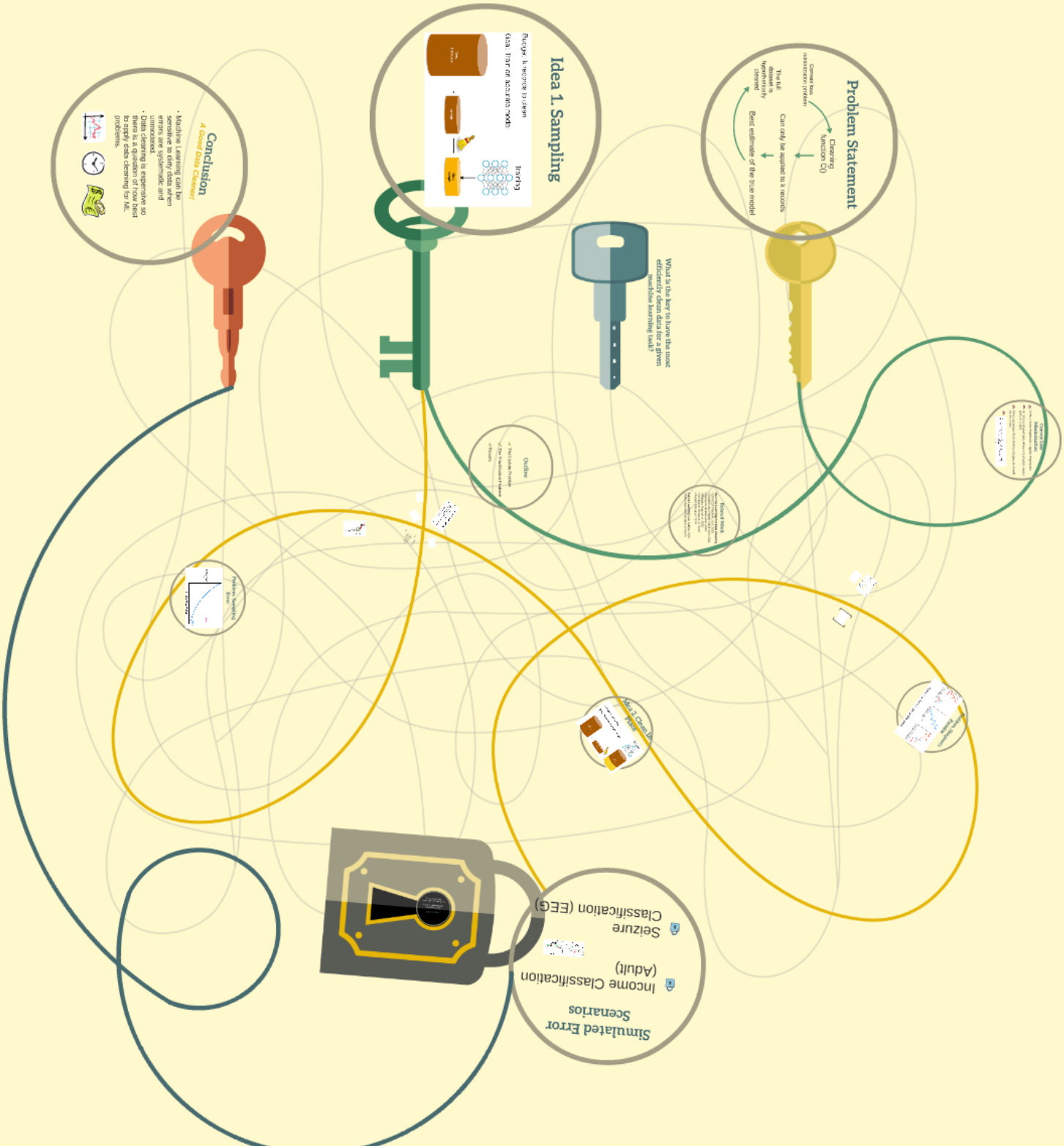
[3] Computational Cost

What is the key to have the most efficiently clean data for a given machine learning task?

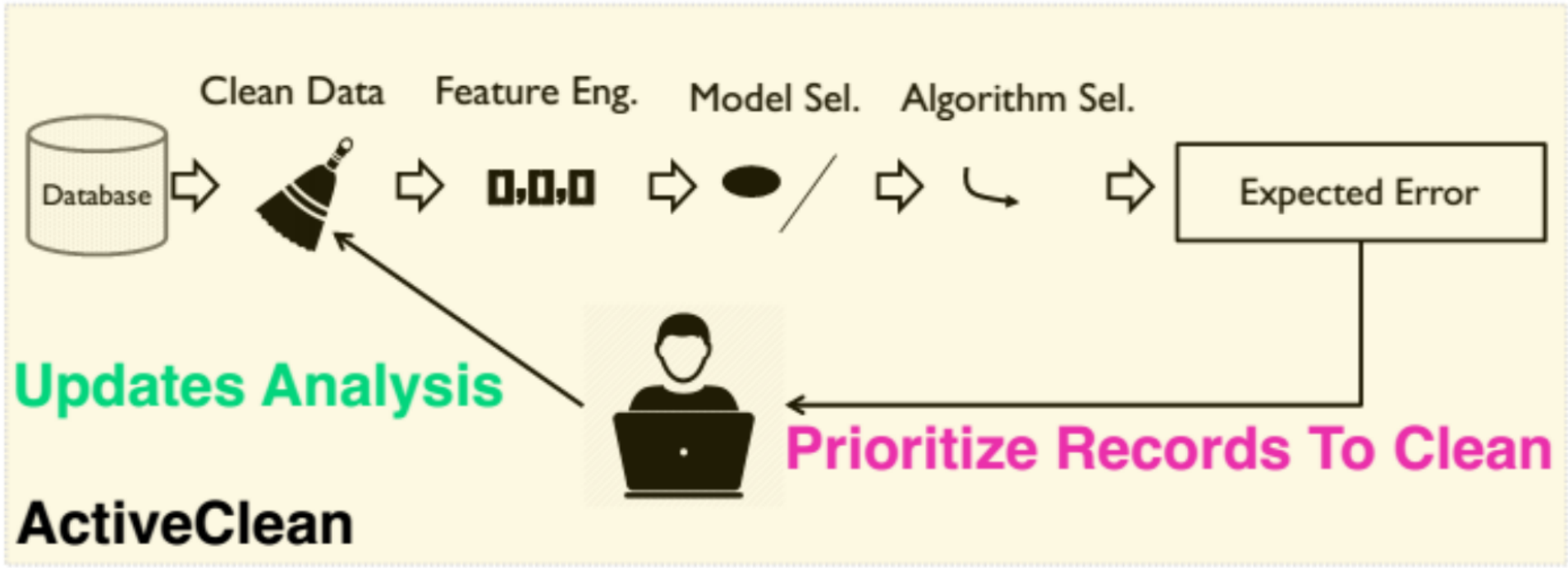


ActiveLearning

Interactive Data Cleaning For Statistical Modeling

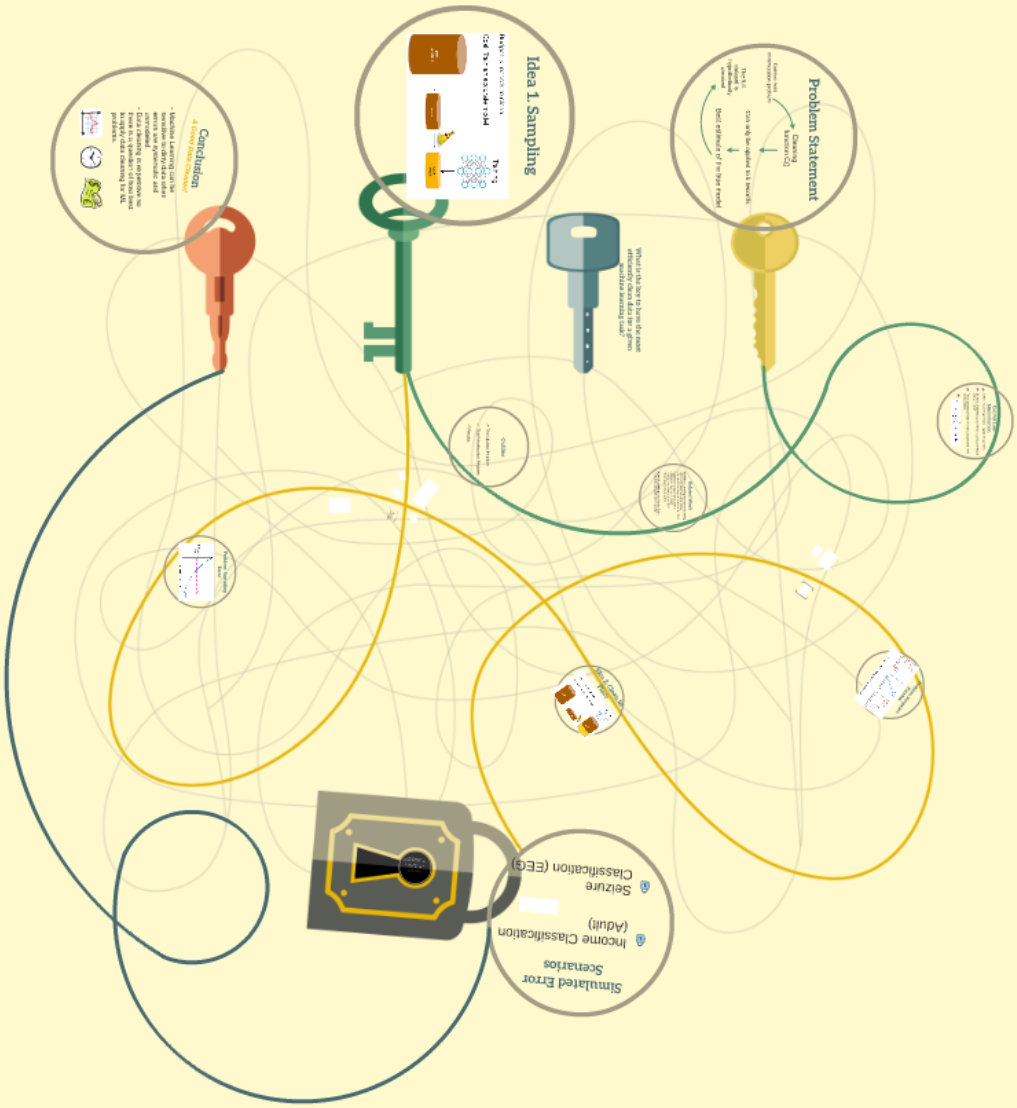


ActiveClean

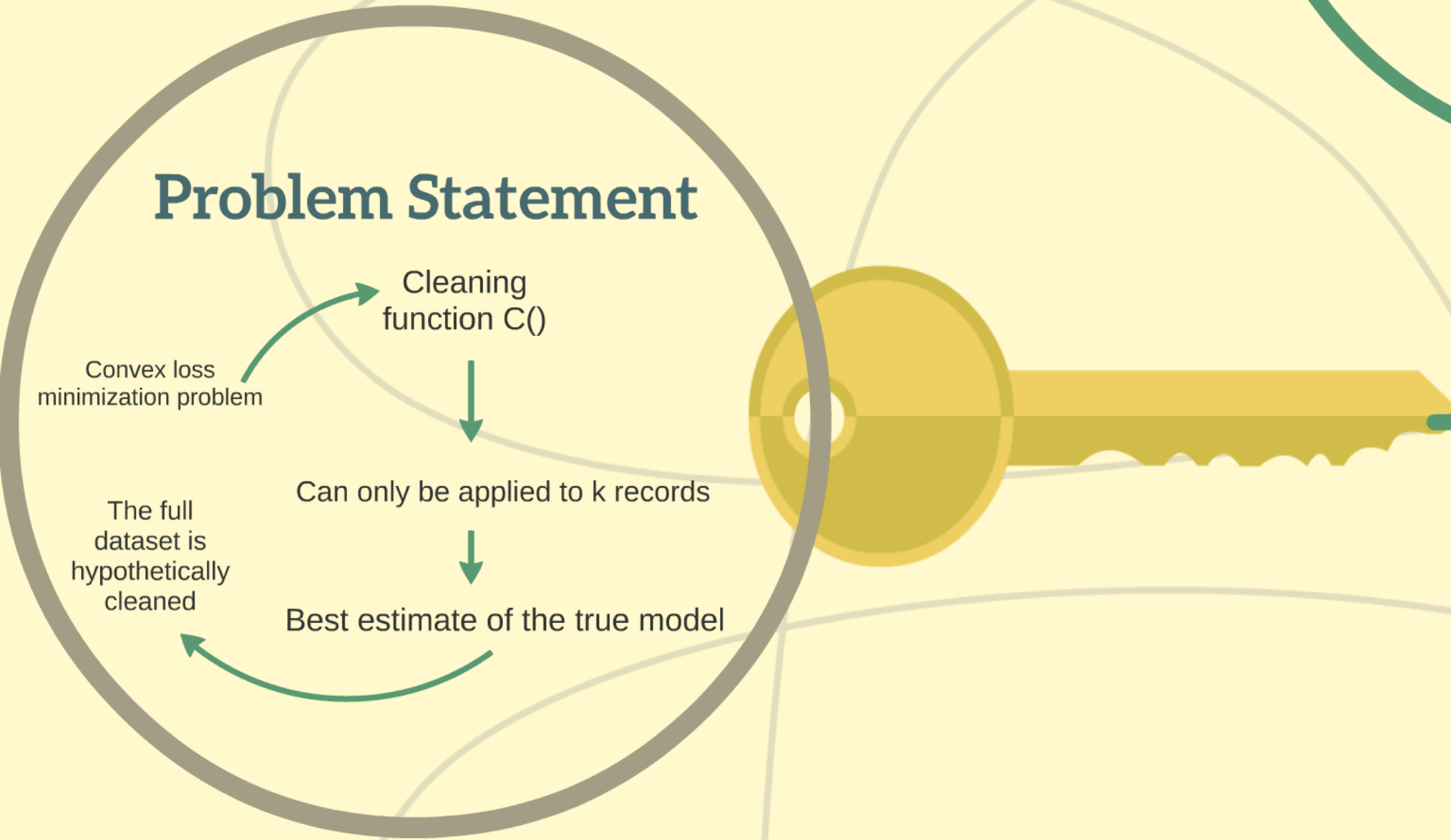


ActiveLearning

Interactive Data Cleaning For Statistical Modeling



Problem Statement



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.

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

Related Work

Machine Learning For Data Cleaning

- ERACER Mayfield et al. 2010
- Guided Data Repair Yakut et al. 2011
- Corleone Gokale et al. 2014
- Wisteria Haas et al. 2015
- Deep Dive Zhang et al. 2014
- Katara Chu et al. 2014

Active Learning (see Settles 2010 “Active learning literature survey”)

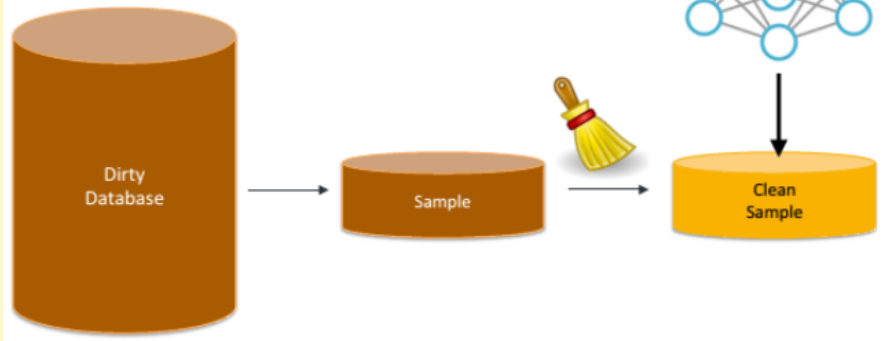
Outline

- The Update Problem
- The Prioritization Problem
- Results

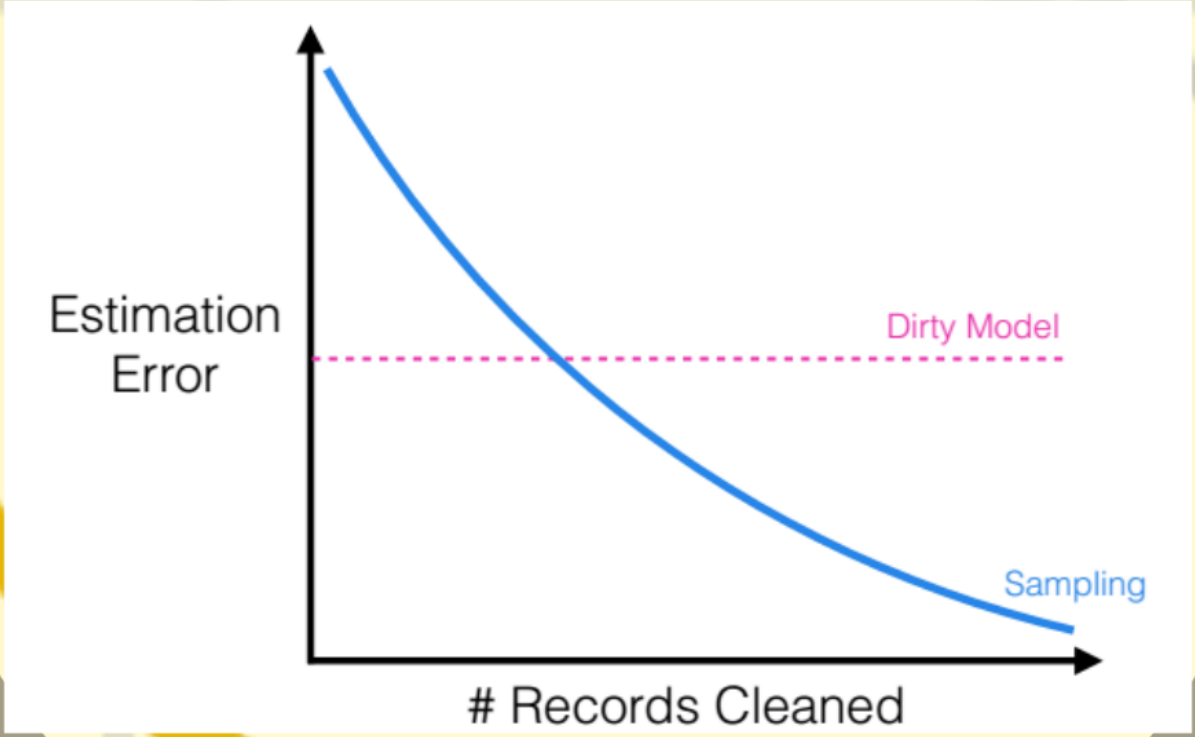
Idea 1. Sampling

Budget: k records to clean

Goal: Train an accurate model



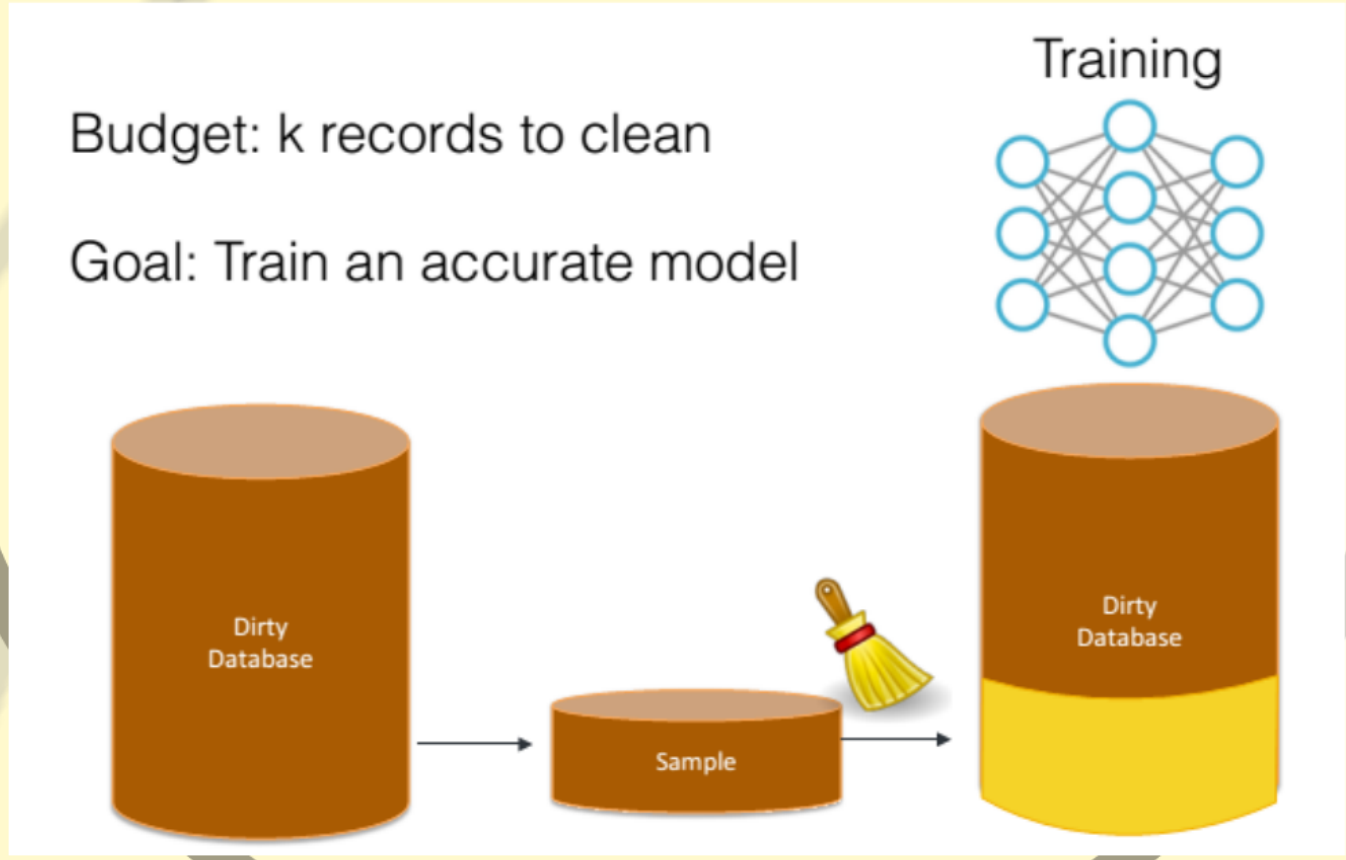
Problem. Sampling Error



Idea 2. Clean In Place

Budget: k records to clean

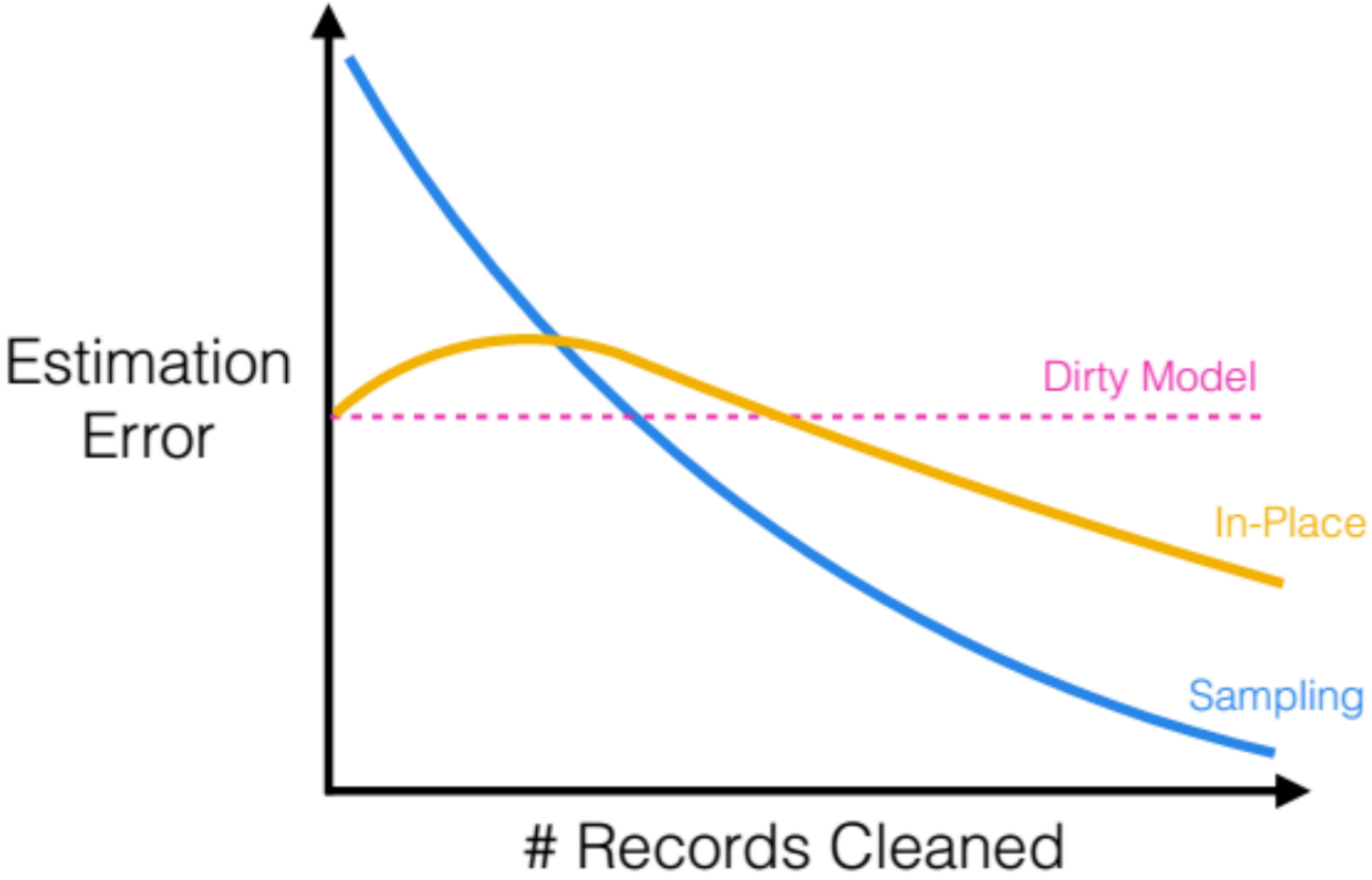
Goal: Train an accurate model



Problem. Simpson's Paradox

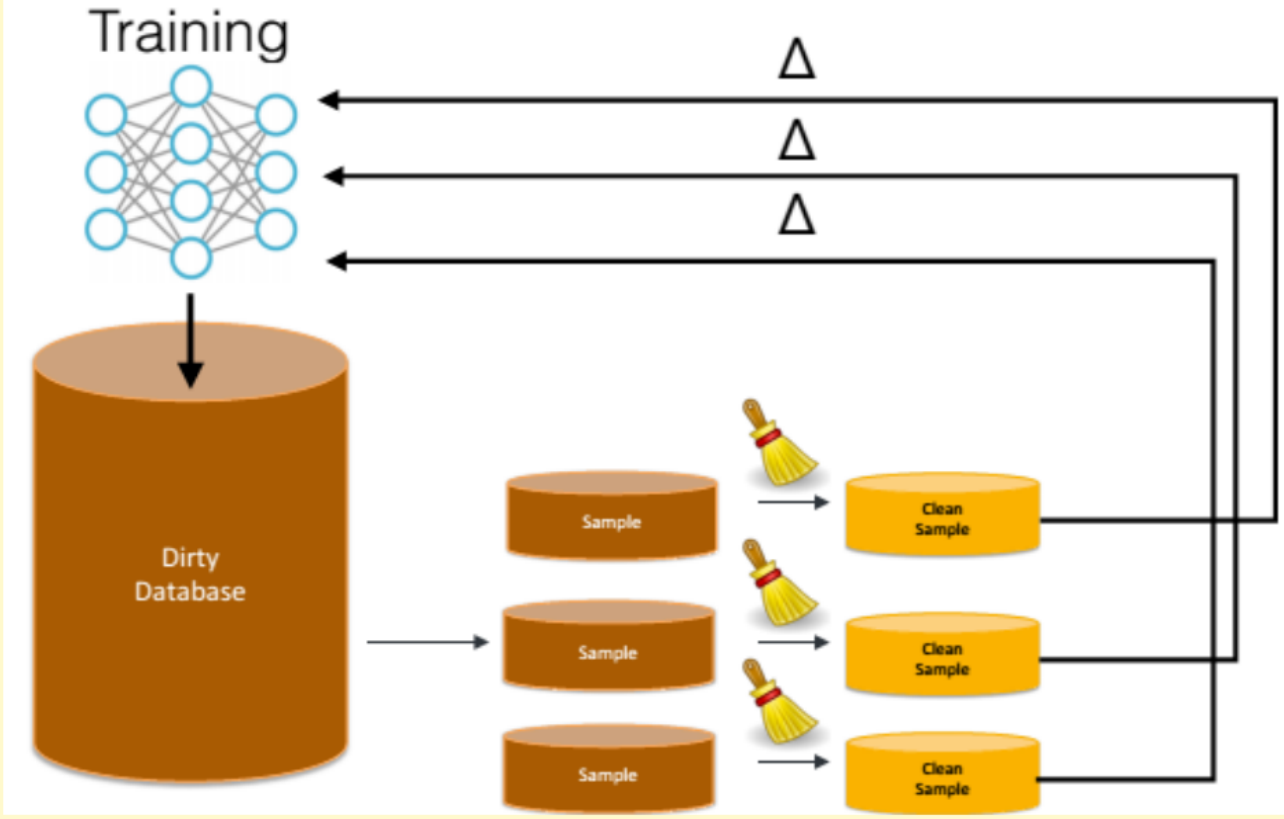
Partial Data Cleaning Can Be Misleading

Problem. Simpson's Paradox

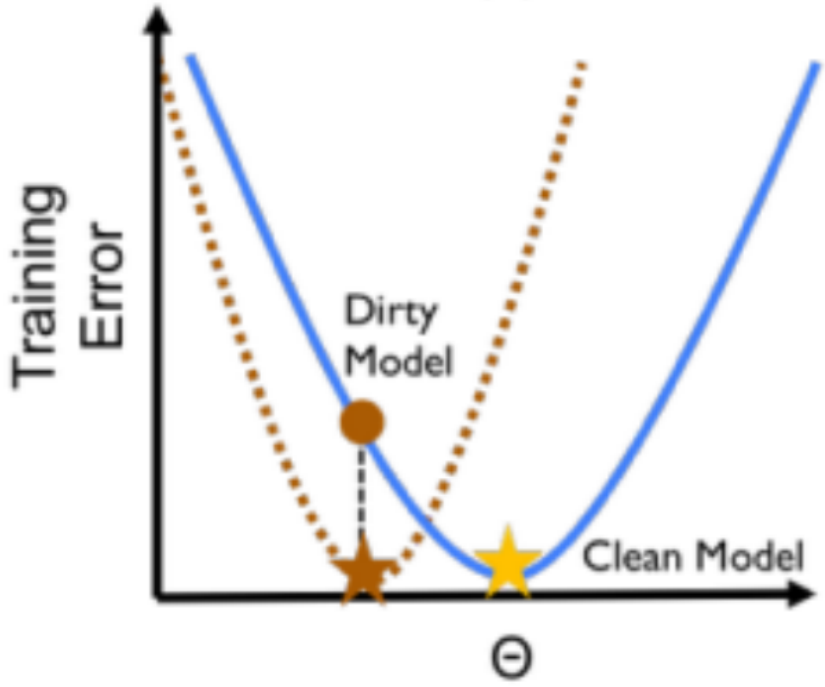


Active Clean

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.

Math time!

$$g(\theta) = \frac{|\, R_{clean} \,|}{|\, R \,|} \cdot g_C(\theta) + \frac{|\, R_{dirty} \,|}{|\, R \,|} \cdot g_S(\theta)$$

$$g_C(\theta) = \frac{1}{|\, R_{clean} \,|} \sum_{i \in R_{clean}} \nabla \phi(x_i^{(c)}, y_i^{(c)}, \theta)$$

$$g_S(\theta) = \frac{1}{|\, S \,|} \sum_{i \in S} \frac{1}{p(i)} \nabla \phi(x_i^{(c)}, y_i^{(c)}, \theta)$$

Model Update Algorithm

1. Take a sample of data S from R_{dirty}
2. Calculate the gradient over the sample of newly clean data and call the result $g_S(\theta^{(t)})$
3. Calculate the average gradient over all of the already clean records in $R_{clean} = R - R_{dirty}$, and call the result $g_C(\theta^{(t)})$
4. Apply the following update rule, which is a weighted average of the gradient on the already clean records and newly cleaned records:

$$\theta^{(t+1)} \leftarrow \theta^{(t)} - \gamma \cdot \left(\frac{|R_{dirty}|}{|R|} \cdot g_S(\theta^{(t)}) + \frac{|R_{clean}|}{|R|} \cdot g_C(\theta^{(t)}) \right)$$

5. Append the newly cleaned records to set of previously clean records $R_{clean} = R_{clean} \cup S$

Optimal Sampling Problem

Given a set of candidate dirty data R(dirty),find sampling probabilities p(r) such that over all samples S of size k it minimizes the variance:

$$\arg\min_{p} \mathbb{E}(\|g_S - g^*\|^2)$$

It can be shown that the optimal distribution over records in R(dirty) is proportional to:

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

EXPERIMENTS

Setup

- Naive-Mix (NM):
- Naive-Sampling (NS)
- Active Learning (AL)
- Active Clean (AC)
- Oracle (O)

Metrics

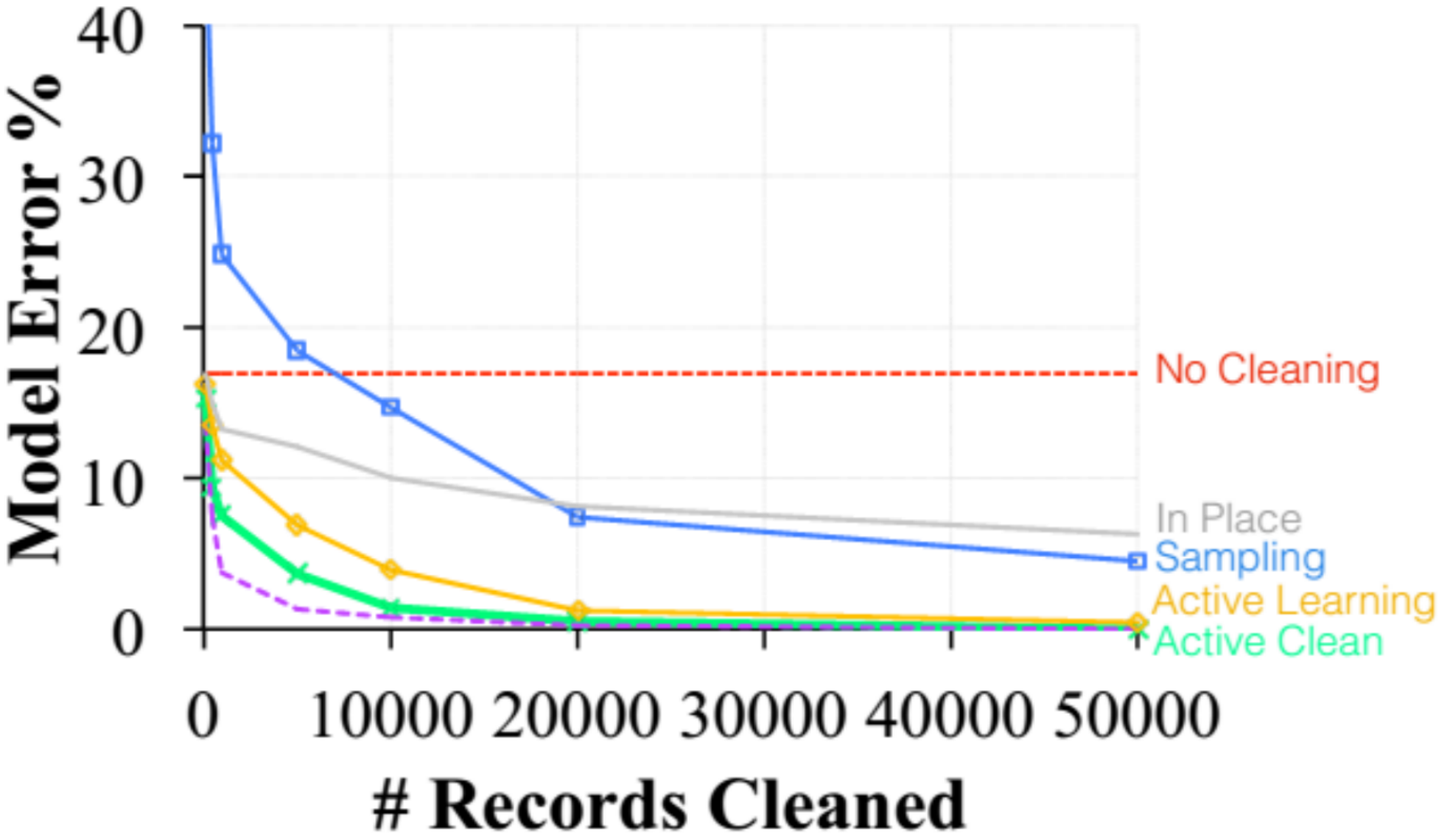
Experimental Setup

- 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

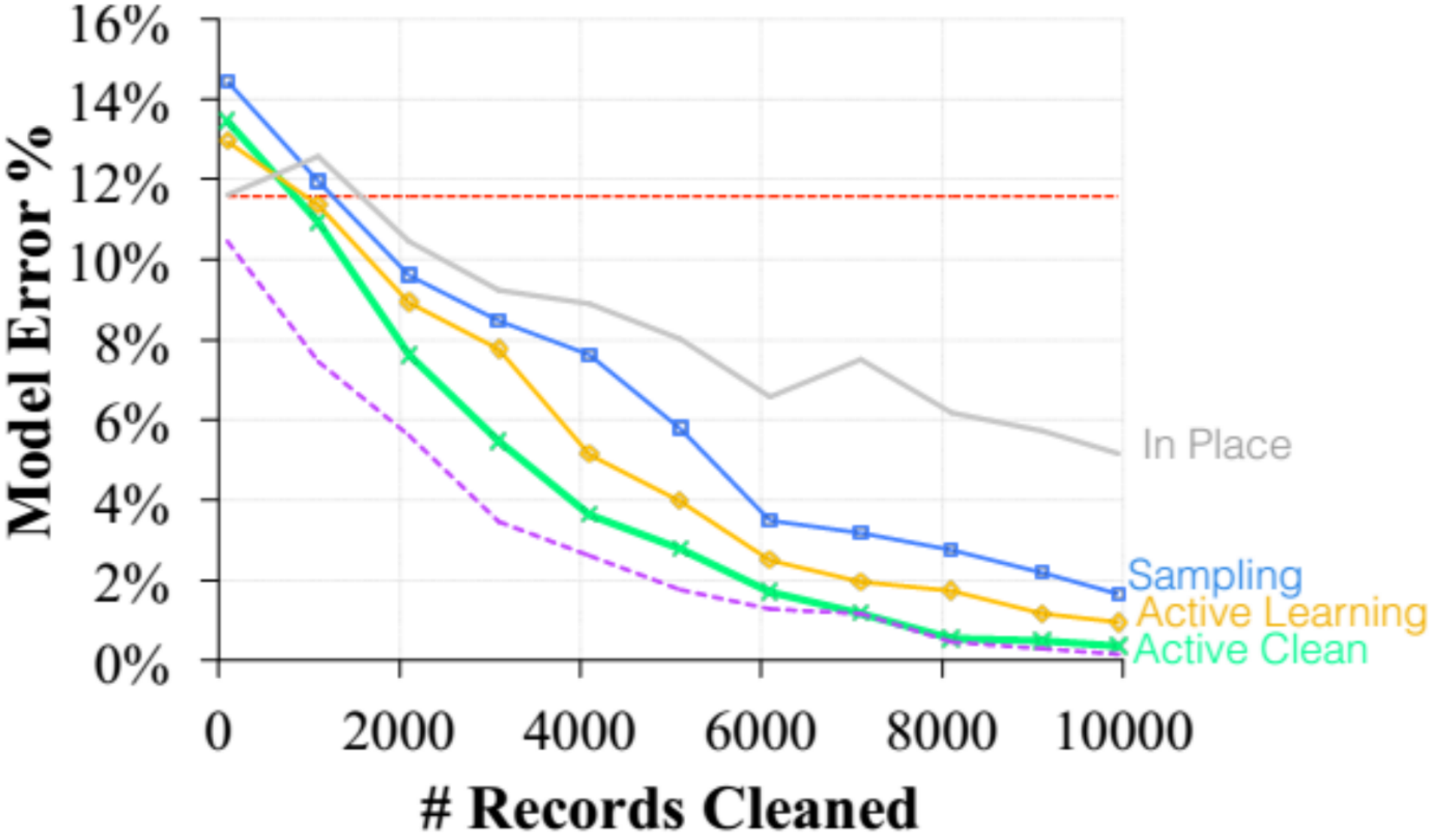


Yahoo Movies

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

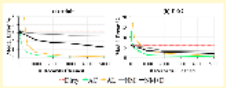
Bloodrage (1979) A psychotic killer stalks the streets of New York City. ***Comedy***

Yahoo Movies

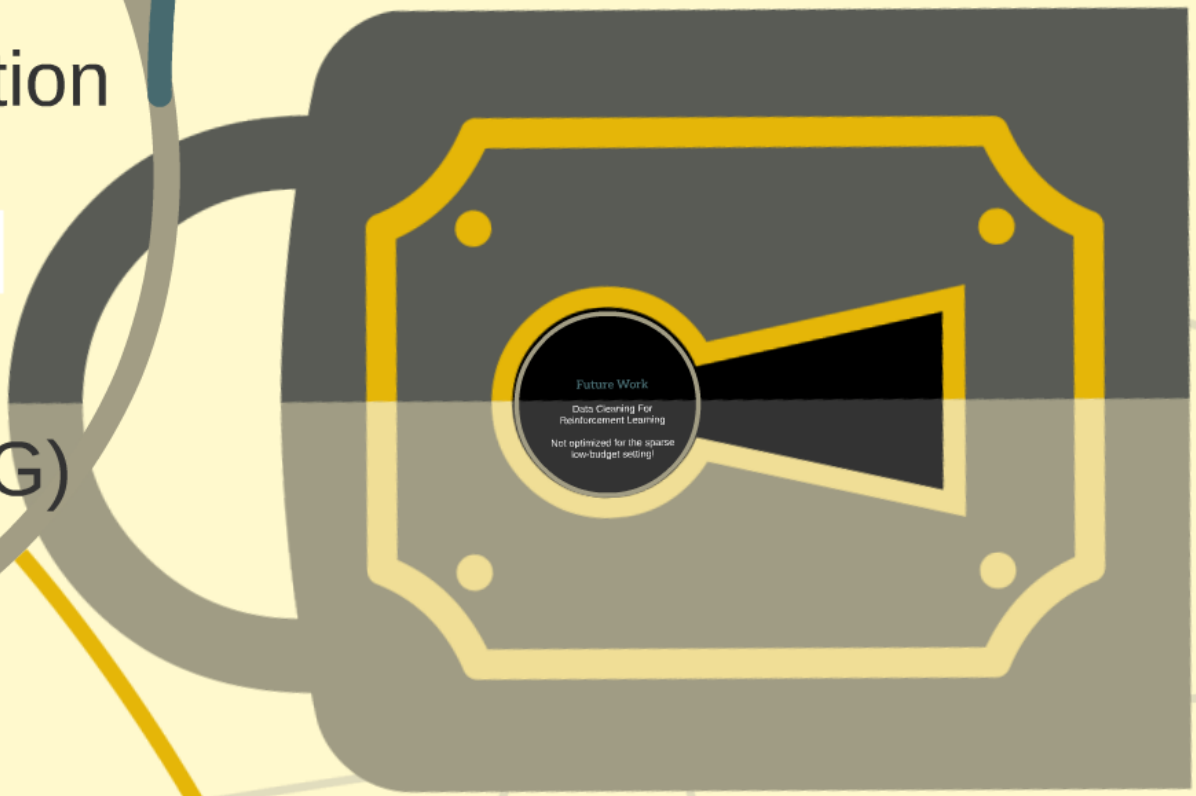


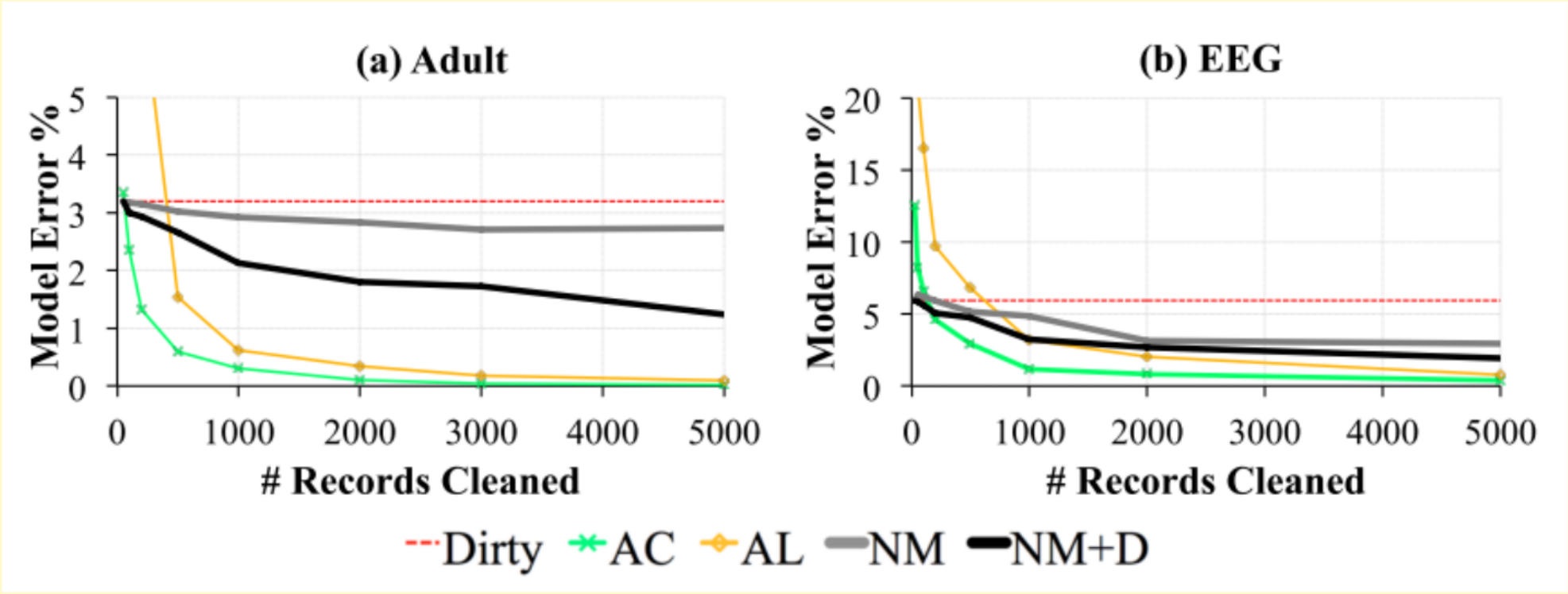
Simulated Error Scenarios

- Income Classification (Adult)



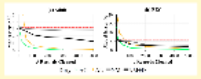
- Seizure Classification (EEG)



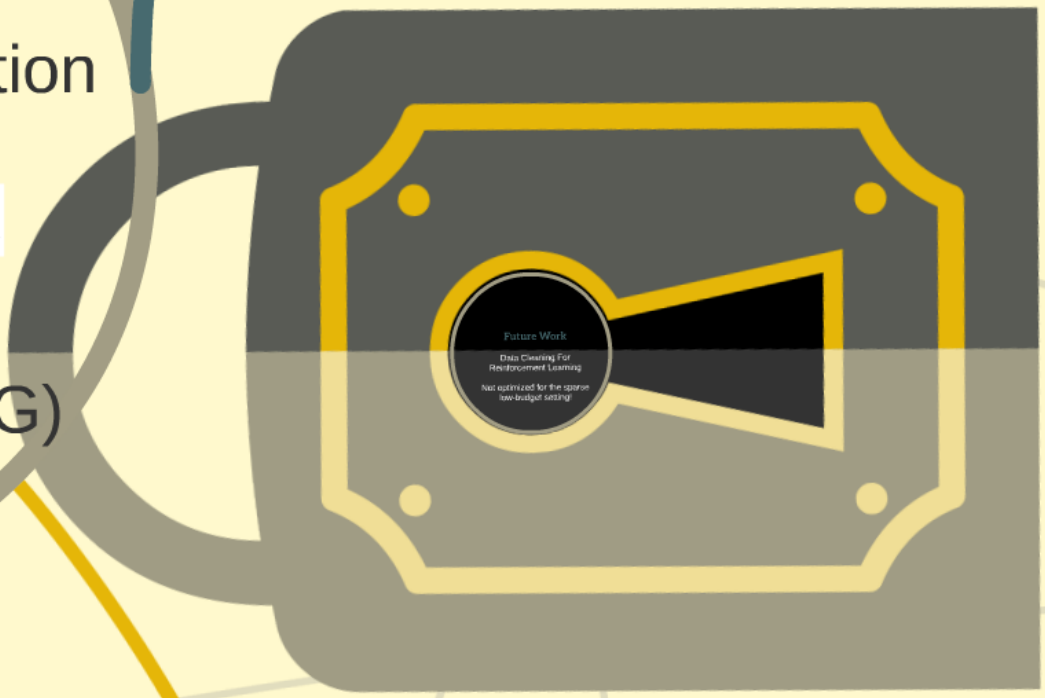

(a) Adult
Model Error %
0
1
2
3
4
5
0
1000
2000
3000
4000
5000
Records Cleaned
(b) EEG
Model Error %
0
5
10
15
20
0
1000
2000
3000
4000
5000
Records Cleaned
Dirty
AC
AL
NM
NM+D

Simulated Error Scenarios

- Income Classification (Adult)



- Seizure Classification (EEG)



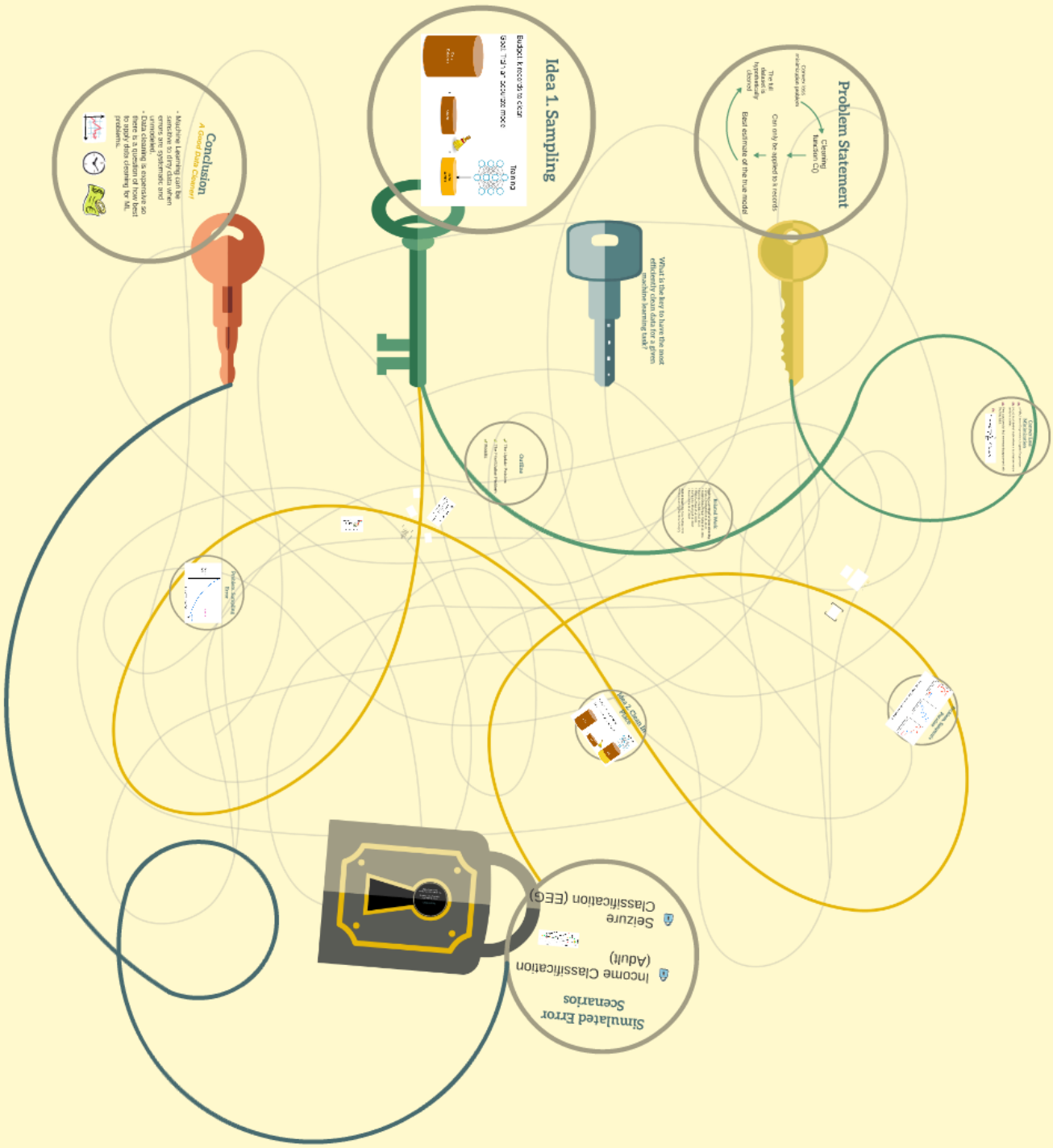
Future Work

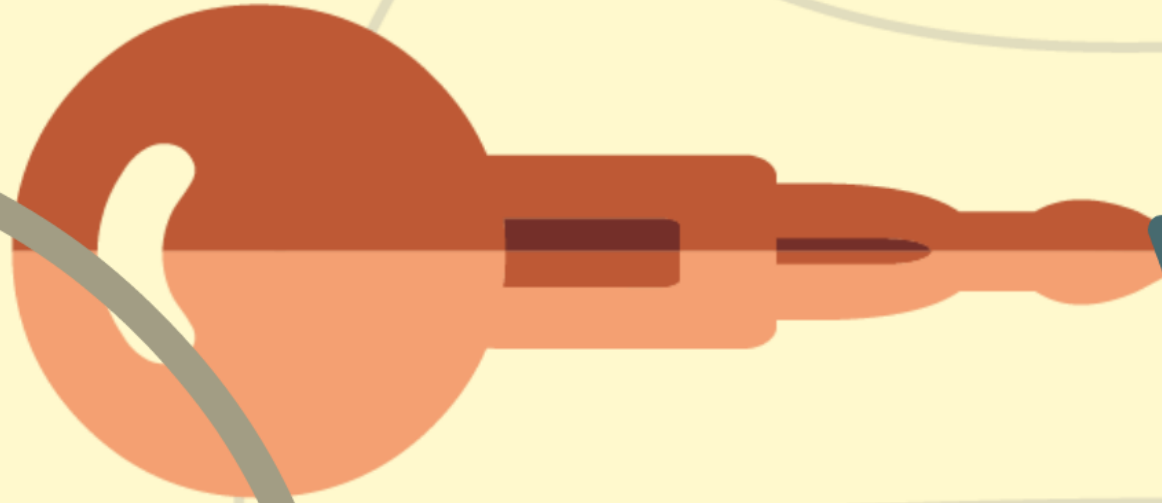
Data Cleaning For Reinforcement Learning

Not optimized for the sparse low-budget setting!

ActiveLearning

Interactive Data Cleaning For Statistical Modeling

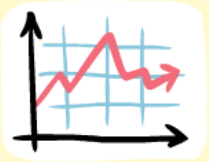




Conclusion

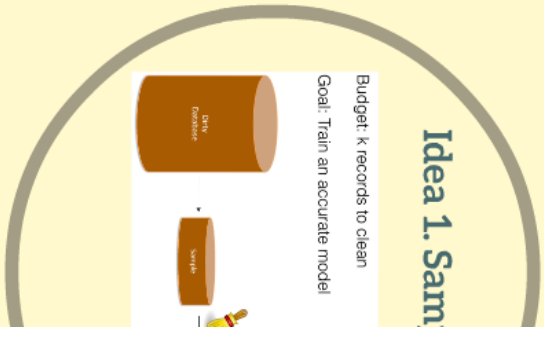
A Good Data Cleaner!

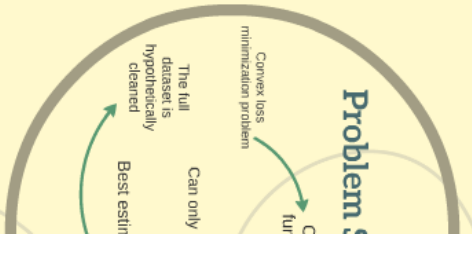
- Machine Learning can be sensitive to dirty data when errors are systematic and unmodeled.
- Data cleaning is expensive so there is a question of how best to apply data cleaning for ML problems.



ActiveLearning

Interactive Data Cleaning For Statistical Modeling






Thank you!
Prezi