Machine Learning for Data Management Systems

Poisoning Attacks on LI

Amol Deshpande
February 23, 2023

Motivation

- Are learned systems vulnerable to adversarial attacks ?
  - An adversary can inject data that makes it perform badly
- What are potential failure modes?
  - Similar questions, but not adversarial
  - What range of datasets does the learned system work well for?

Background

- Recursive Model Indexes

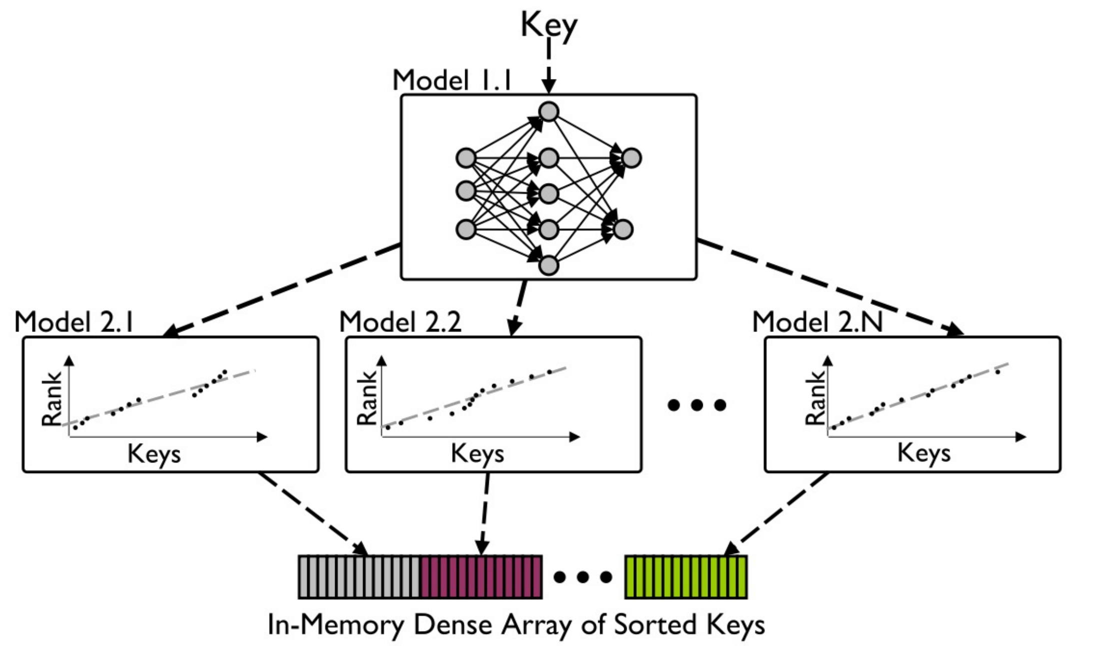


Figure 1: An illustration of the Recursive Model Index (RMI) with a two-stage architecture. The first stage is a single neural network model while the second stage is series of linear regression models on 1-out-of-*N* key partitions of equal size.

Threat Model

- An adversary who wants to worsen the performance
  - Perhaps better to think of failure scenarios in this scenario
- Can add a percent of “poisoning keys”
- White-box attacks
  - Adversary has access to all the data, or at least enough distributional information
  - Black-box attacks are harder to do
- Metrics
  - Ratio loss: MSE of poisoned vs non-poisoned dataset
  - Average memory offset: how off we are from the right location

Single Linear Regression

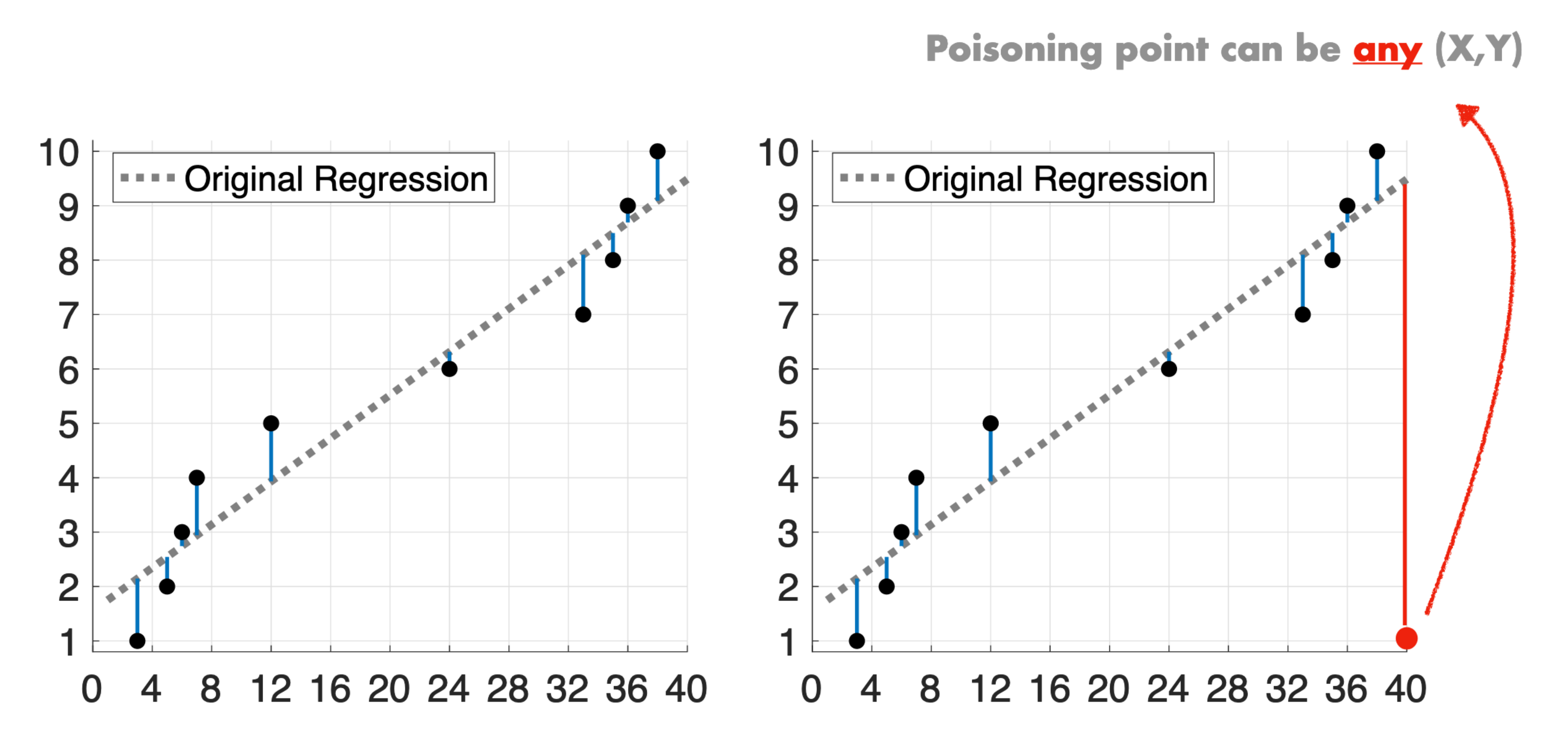
- Difference from prior work – when we add a poisoning key, it shifts the CDF slightly
- Linear regression objective

Definition 1 (Linear Regression on CDFs). *Let* $K = \{k_1, \cdots, k_n\} \subseteq \mathcal{K}$ *be the set of integers that correspond to the keys of the index. Every key* $k_i \in K$ *has its associated rank* $r_i \in [1, n]$. *The* linear regression model on a CDF *computes a pair of regression parameters* (w, b) *that minimizes the following mean squared error (MSE) function :*

$$\min_{w,b} \mathcal{L}\left(\{k_i, r_i\}_{i=1}^{n}, w, b\right) = \min_{w,b} \left(\sum_{i=1}^{n} (wk_i + b - r_i)^2\right).$$

Normal Linear Regression

- Poisoning points can be put anywhere



Single Linear Regression

- Poisoning problem definition

Definition 2 (**Poisoning Linear Regression on CDF**). *Let K be the set of n integers that correspond to the keys and let P be the set of p integers that comprise the poisoning points. The augmented set on which the linear regression model is trained is $\{(k'_1, r'_1), (k'_2, r'_2), \cdots, (k'_{n'}, r'_{n'})\}$, where $k'_i \in K \cup P$ and $r'_i \in [1, n+p]$. The goal of the adversary is to choose a set P of size at most λ so as to maximize the loss function of the augmented set $K \cup P$ which is equivalent to solving the bilevel optimization problem:*

$$\arg\max_{P \text{ s.t. } |P| \leq \lambda} \left(\min_{w,b} \mathcal{L}\left(\{k'_i, r'_i\}_{i=1}^{n+p}, w, b \right) \right)$$

Compound Effect

- Poisoning problem definition

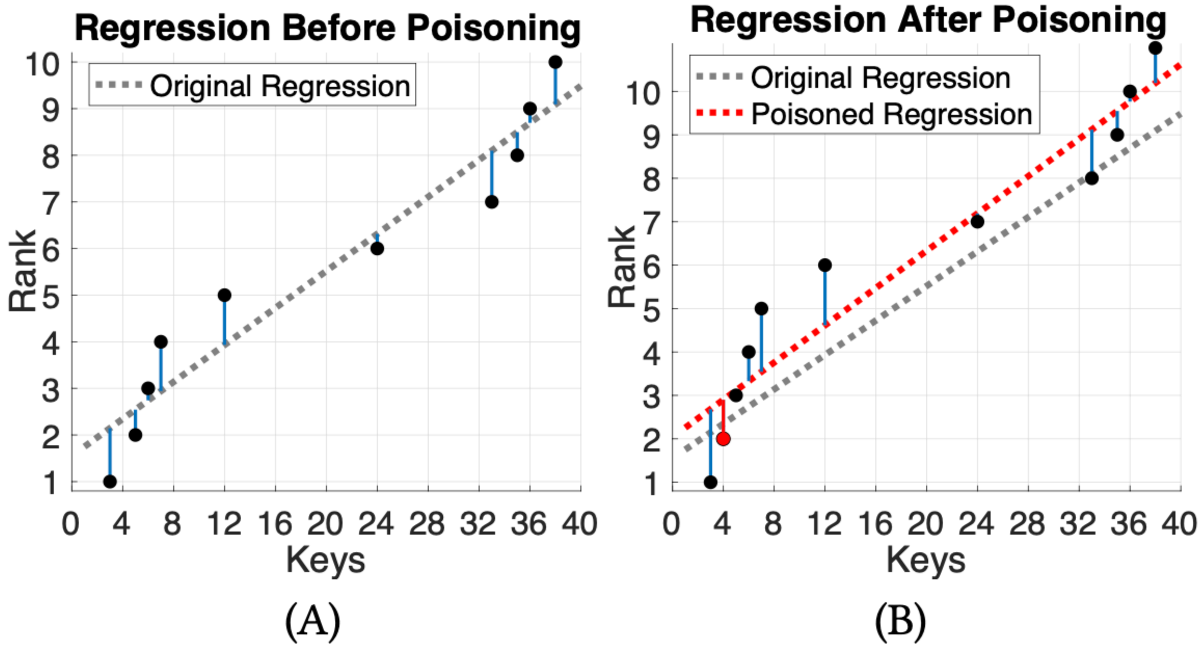
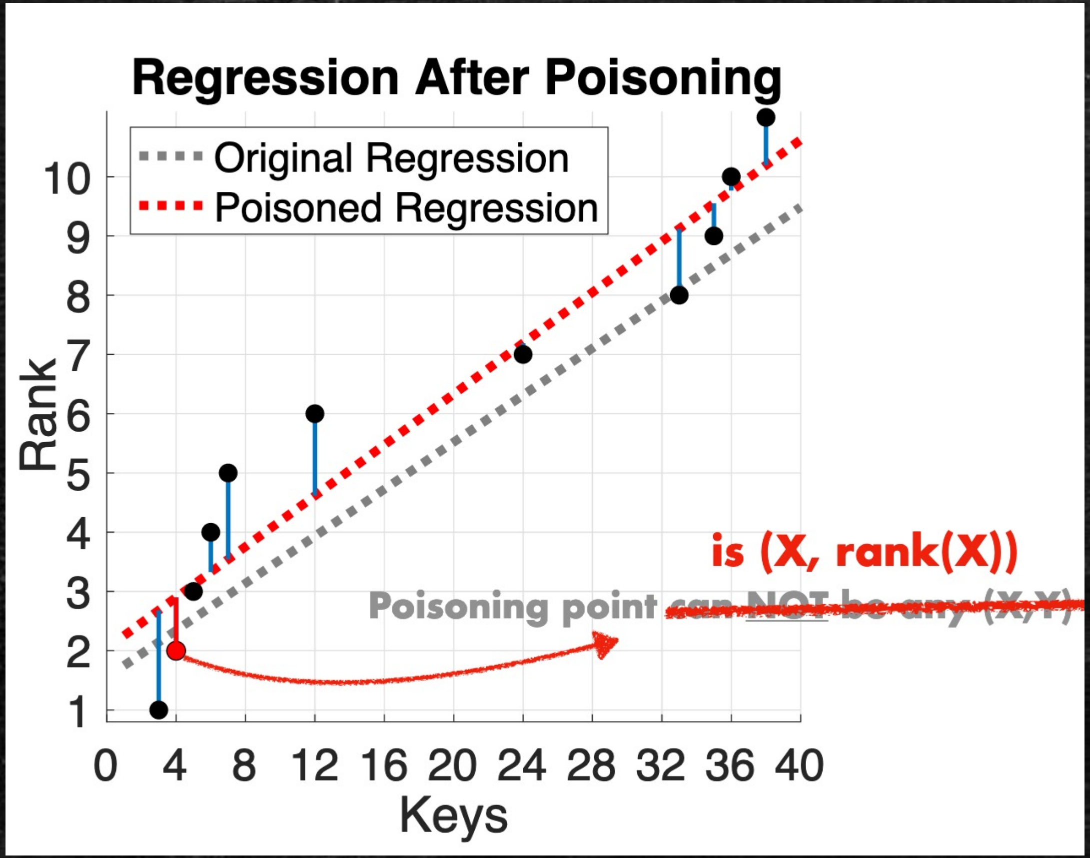
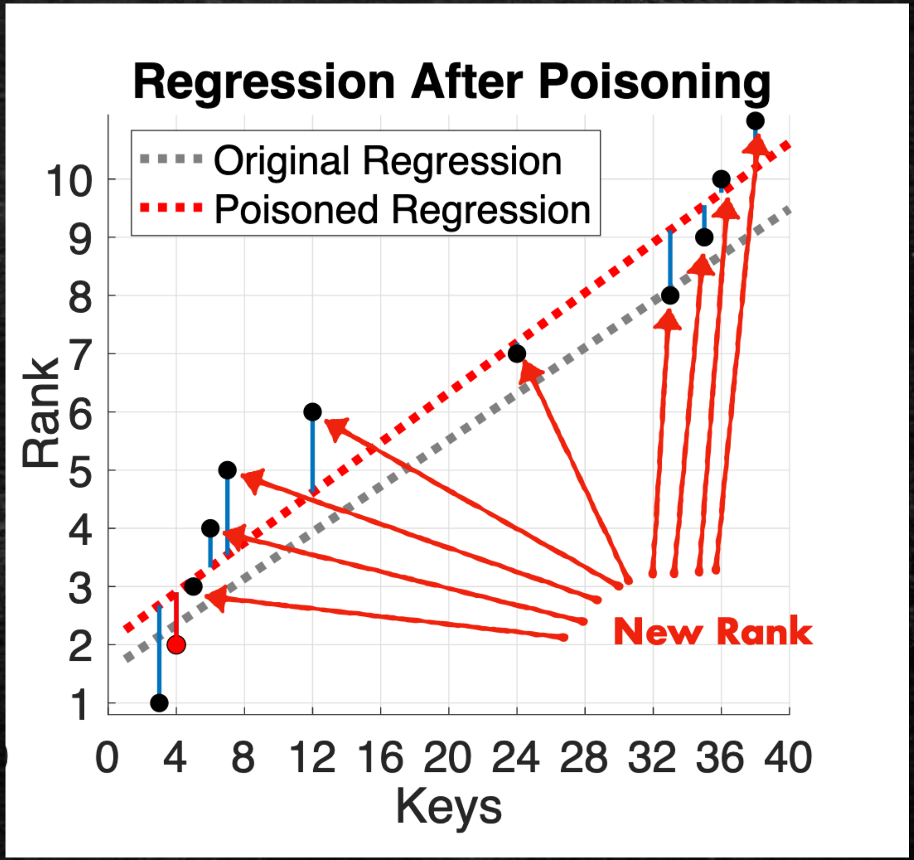


Figure 2: Illustration of the compound effect of poisoning using a single key k_p colored in red. All original keys that are larger than k_p increase their rank by one. The new regression line, dotted red line, accumulates larger error from most of the original points due to the adjustment of ranks.

Compound Effect

- Poisoning problem definition





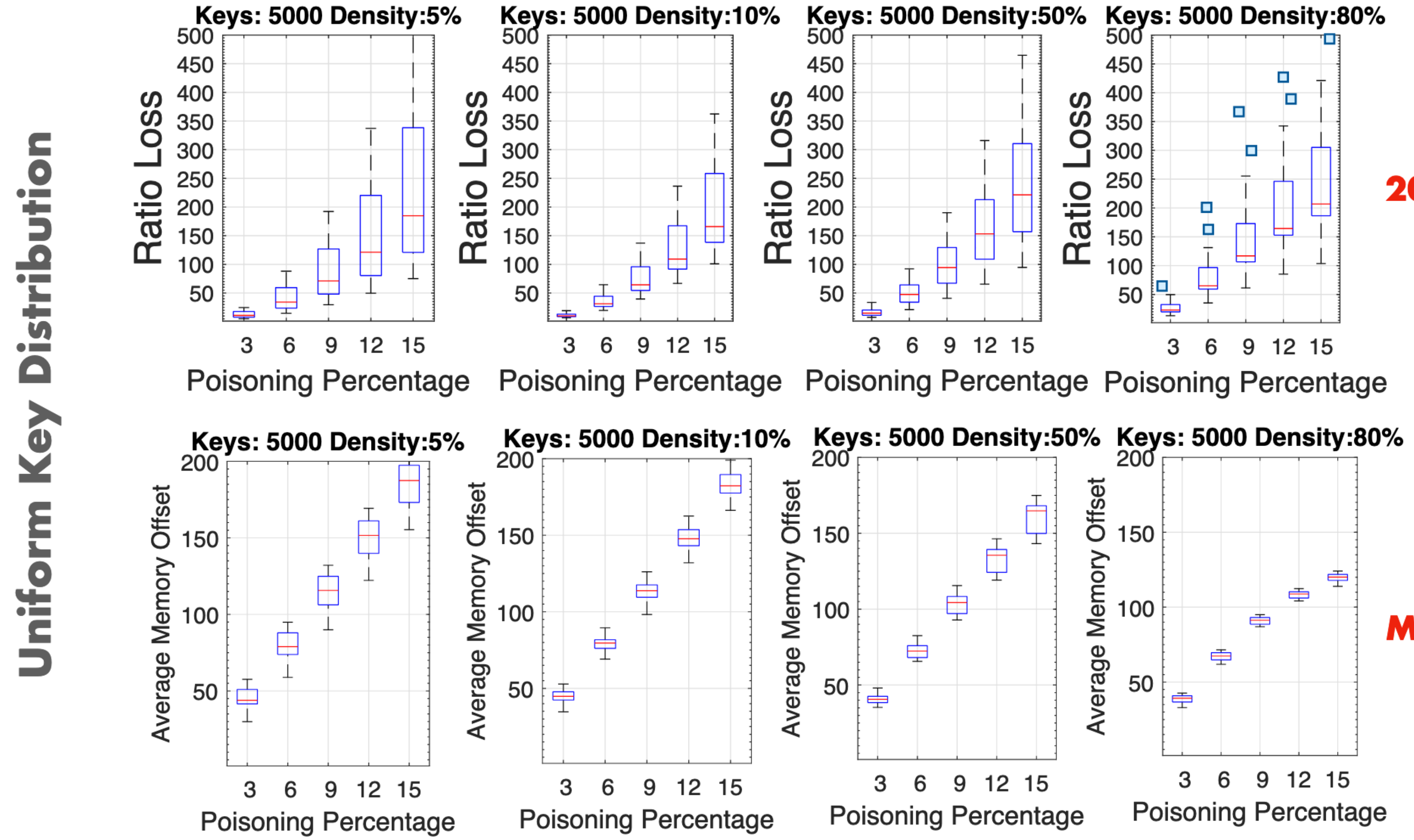
Algorithms

- Linear Poisoning Attack for a "single" point
  - Optimal
  - Expression for loss etc., can be computed incrementally

$$L(k_p) = \begin{cases} \min_{w,b} \left(\sum_{k' \in K \cup k_p} (wk' + b - r')^2 \right) & \text{, if } k_p \notin K \\ \perp & \text{, if } k_p \in K \end{cases} \quad (1)$$

- Greedy algorithm for multiple points
  - One at a time

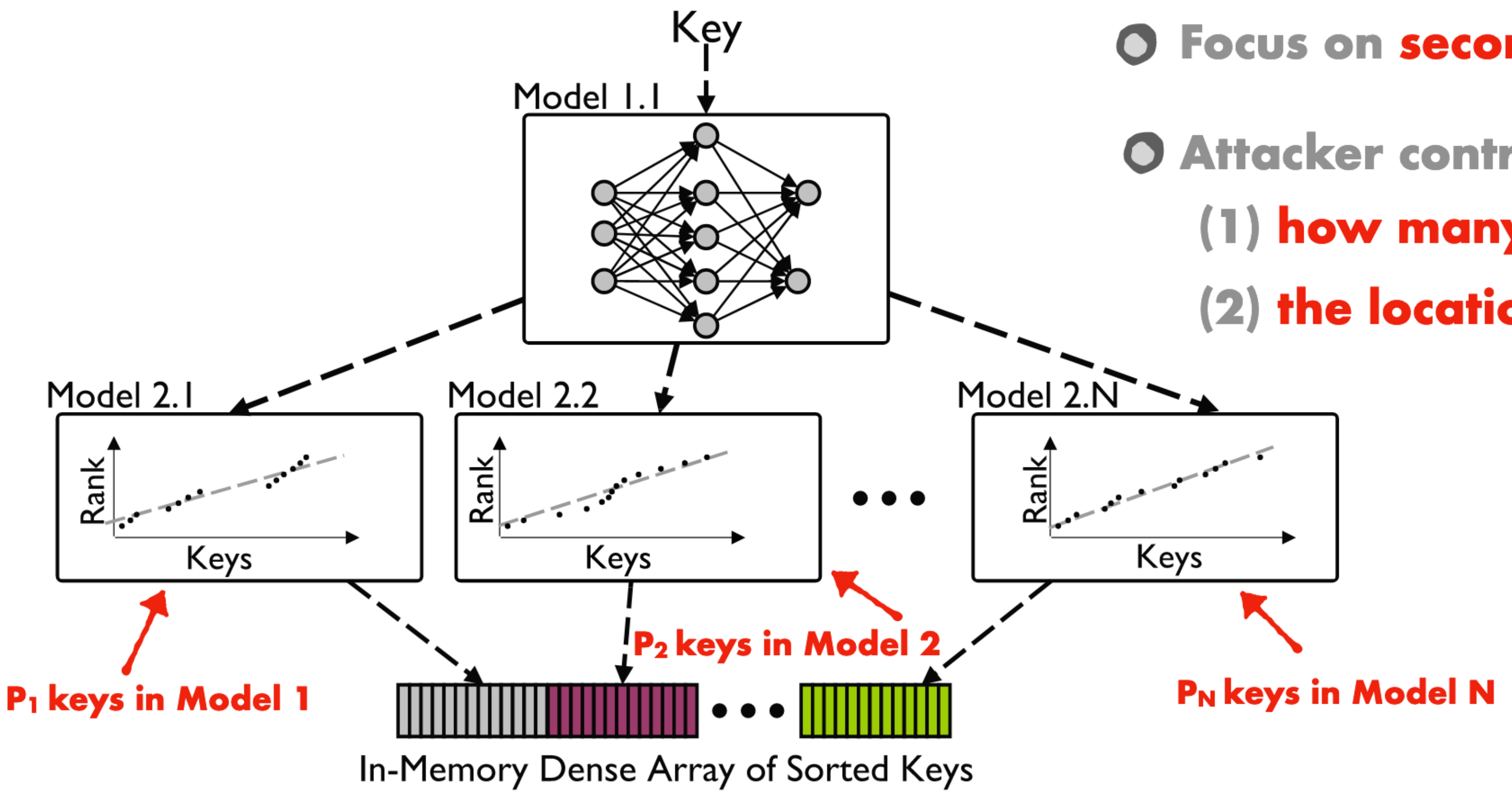
Results for a Single LR



200x larger MSE

Memory Offset 125-180

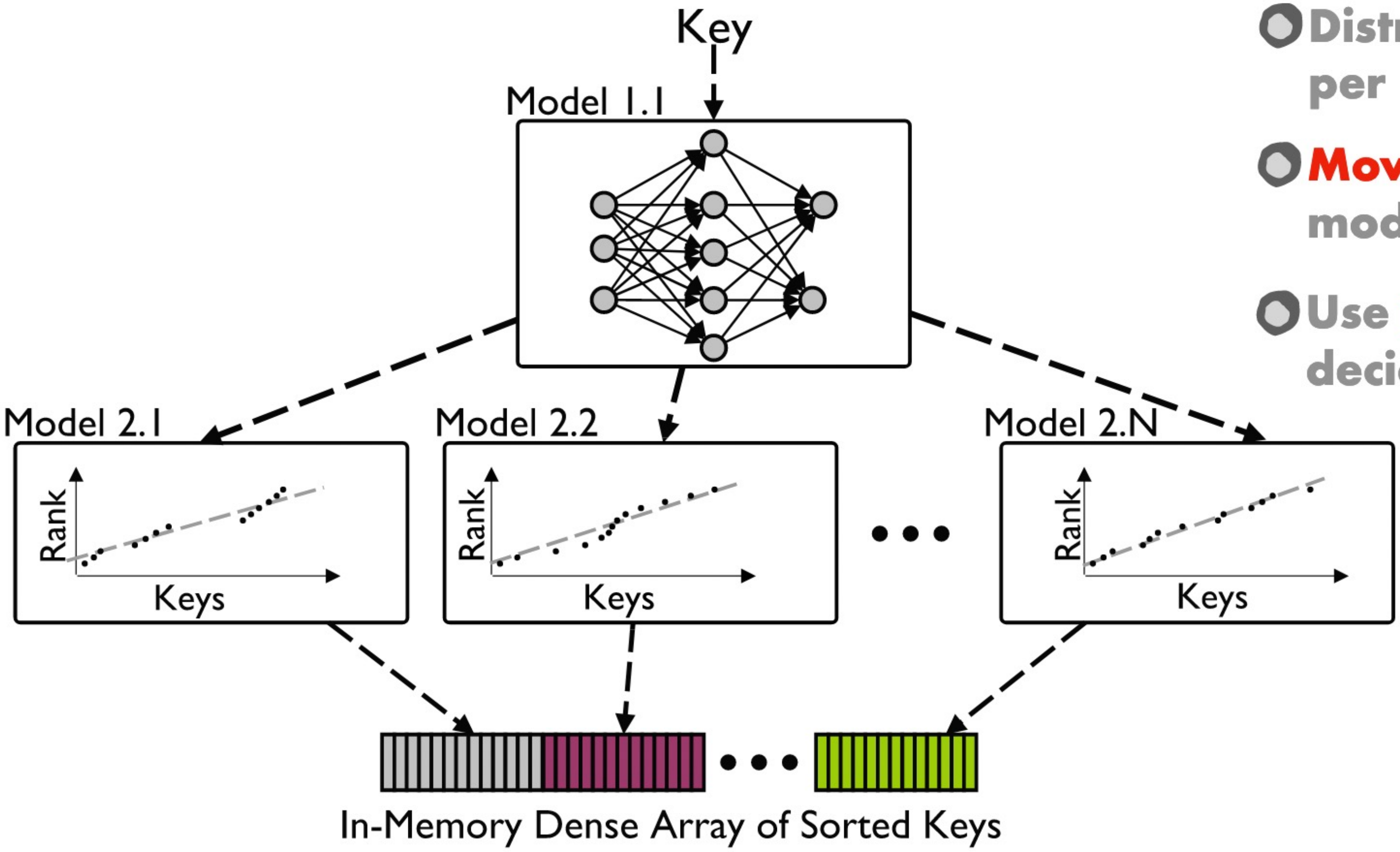
Attacks on Hierarchical Model



ADVERSARIAL APPROACH

- Focus on **second-level** poisoning (regression)
- Attacker controls
  (1) **how many** keys per model (Volume)
  (2) **the location** of poisoning keys per model

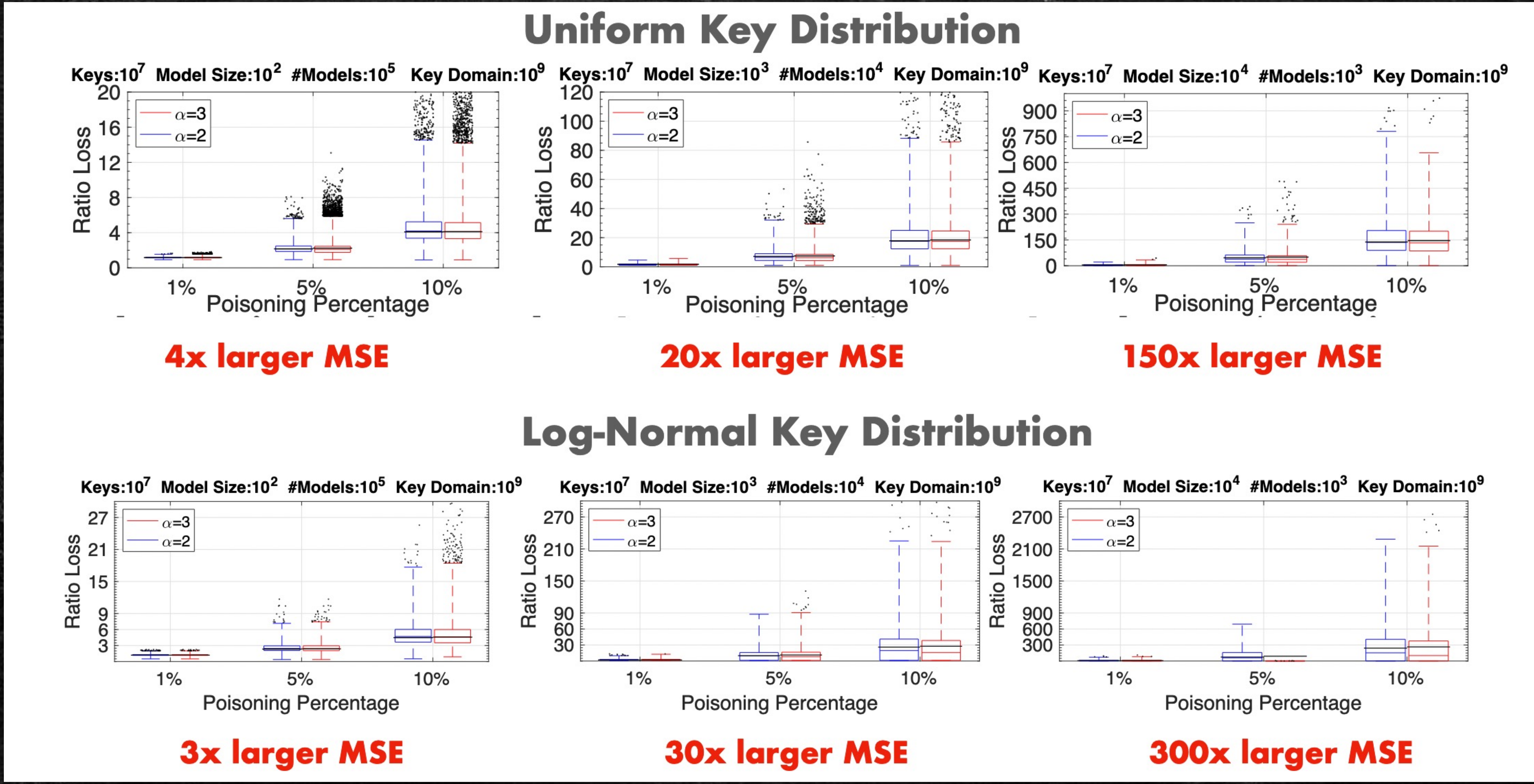
Attacks on Hierarchical Model



(HIGH-LEVEL) ALGORITHM

- Distribute the **same number** of poisoning keys per model
- **Move** a poisoning key to the next/previous model if it increases the total error
- Use **previous multipoint regression attack** to decide which poisoning points to insert

Attacks on Hierarchical Model



Attacks on Hierarchical Model

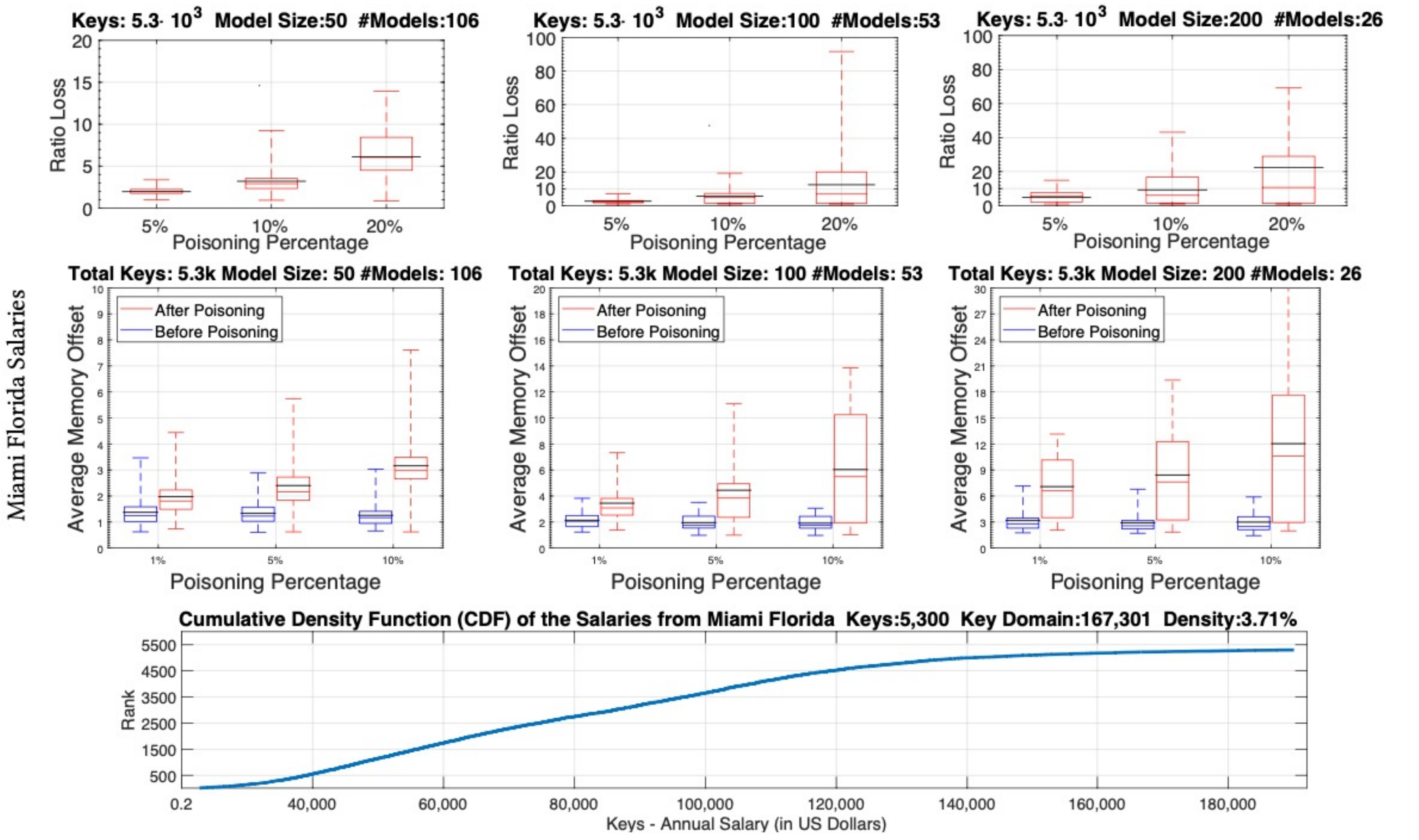


Figure 7: Evaluation of the multi-point poisoning for RMI applied on the CDF of the unique salaries of employees from Dada County in Miami. The X-axis represents different overall poisoning percentage where the second-stage poisoning threshold α takes value $\alpha = 3$. The third row presents the CDF.

Some Discussion Points

- What’s the main take-away from this paper?
- Major concerns with the paper?
- Possible improvements?