Practical Differentially Private Top-k Selection with Pay-what-you-get Composition

David Durfee and Ryan Rogers

Spotlight Presentation

NeurIPS 2019

Differential Privacy [Dwork, McSherry, Nissim, Smith ‘06]

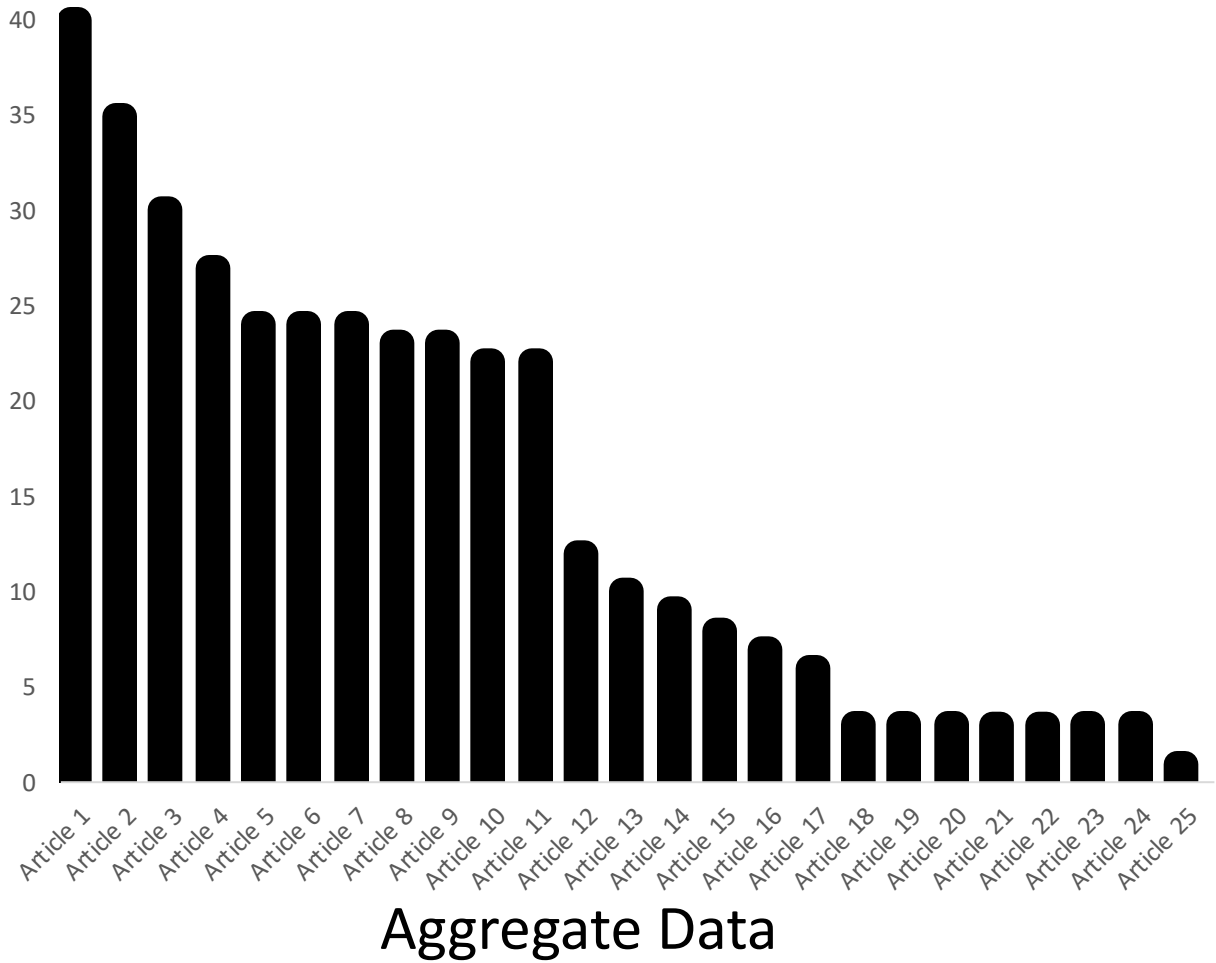
A randomized algorithm $A: \mathcal{D} \to \mathcal{Y}$ is (ε, δ) −DP if for any neighboring data sets $x, x' \in \mathcal{D}$ and any outcome $S \subseteq \mathcal{Y}$ we have:

$$P(A(x) \in S) \leq e^{\varepsilon} P(A(x') \in S) + \delta$$

Privacy loss

Privatizing Top-k Queries

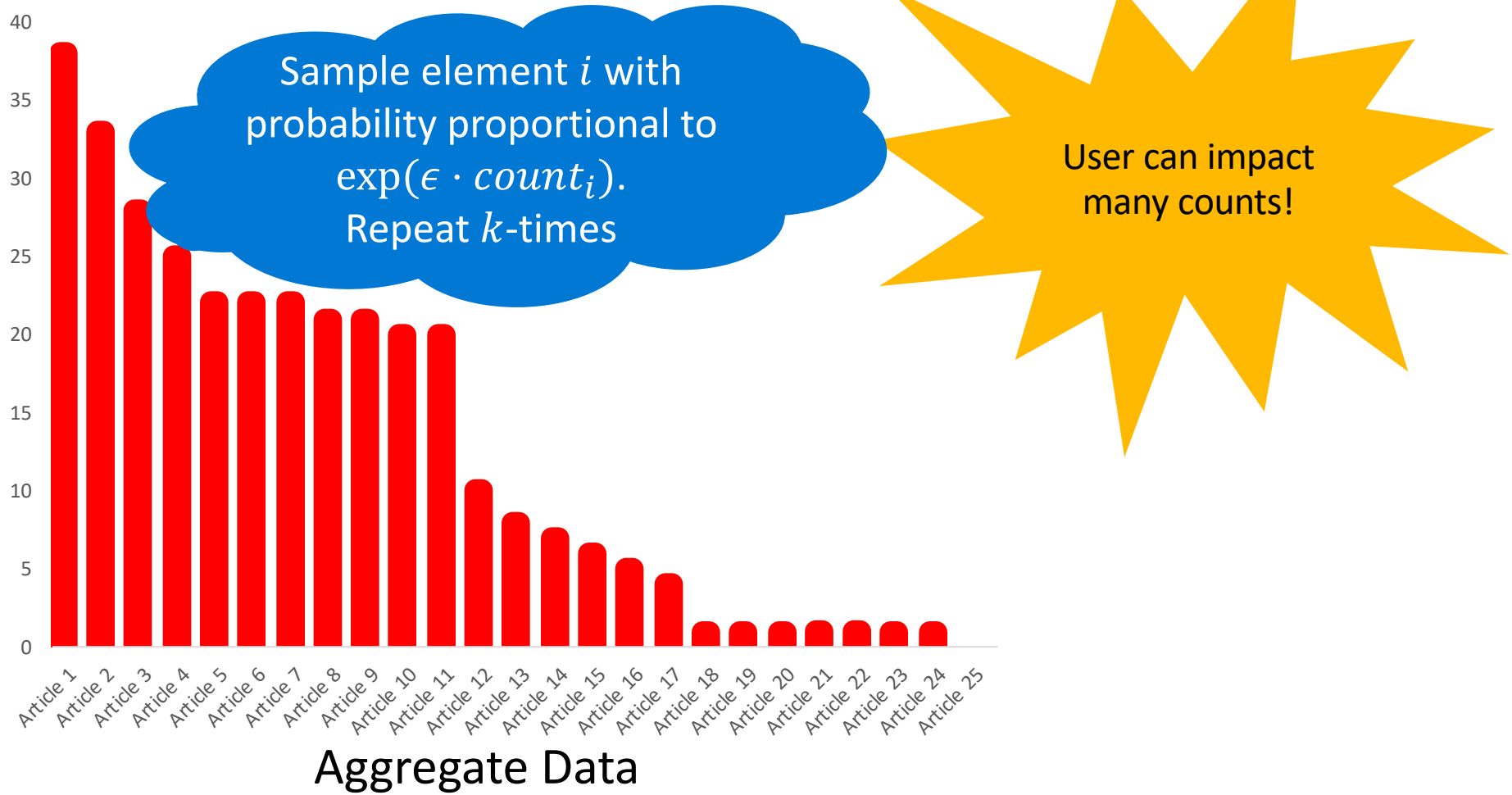
Query: Top-10 articles with distinct user engagement?



Aggregate Data

Privatizing Top-k Queries

Query: Top-10 articles with distinct user engagement?



Privatizing Top-k Queries

Query: Top-10 articles with distinct user engagement?

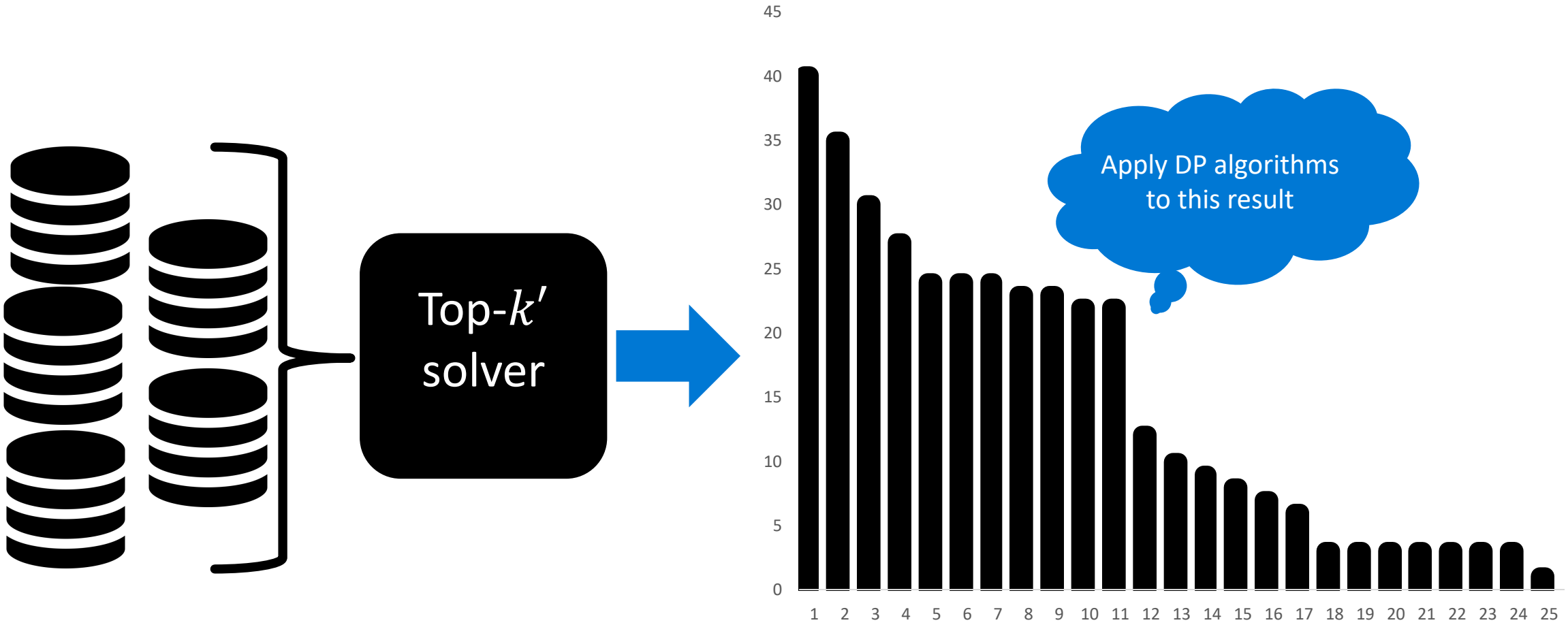
Sample element i with probability proportional to $\exp(\epsilon \cdot count_i)$.
Repeat k-times

[McSherry, Talwar '07]
Releasing only elements in top-k (not their counts) ensures $k\epsilon$-DP

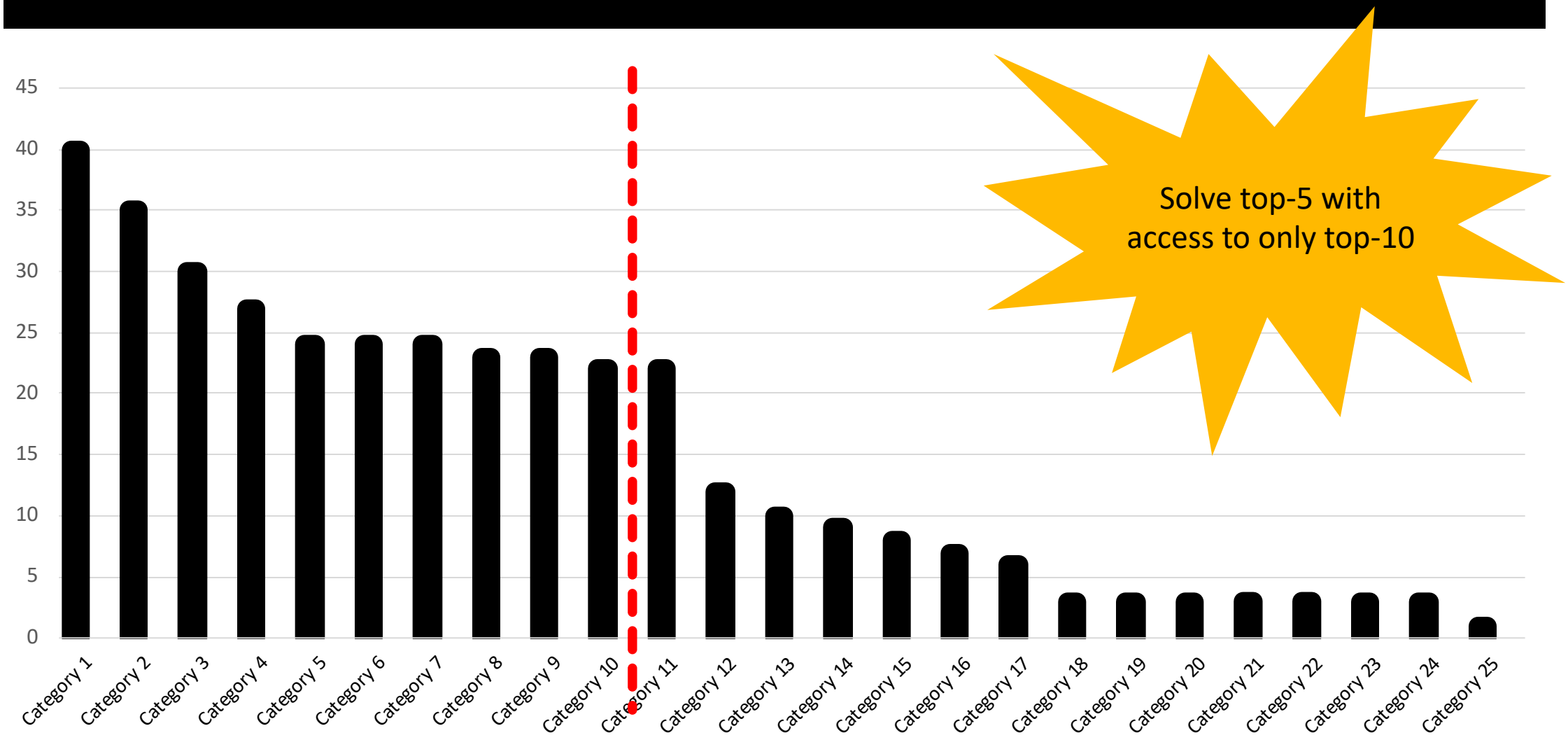
Unknown Domain Setting

- Previous algorithms require knowing the full data domain
- They require adding noise to counts even when the true count is zero
- Typically, the domain is unknown or very large (e.g. all possible articles)
- Lots of prior work for Frequent Itemsets, but requires knowing structure of the data domain universe.

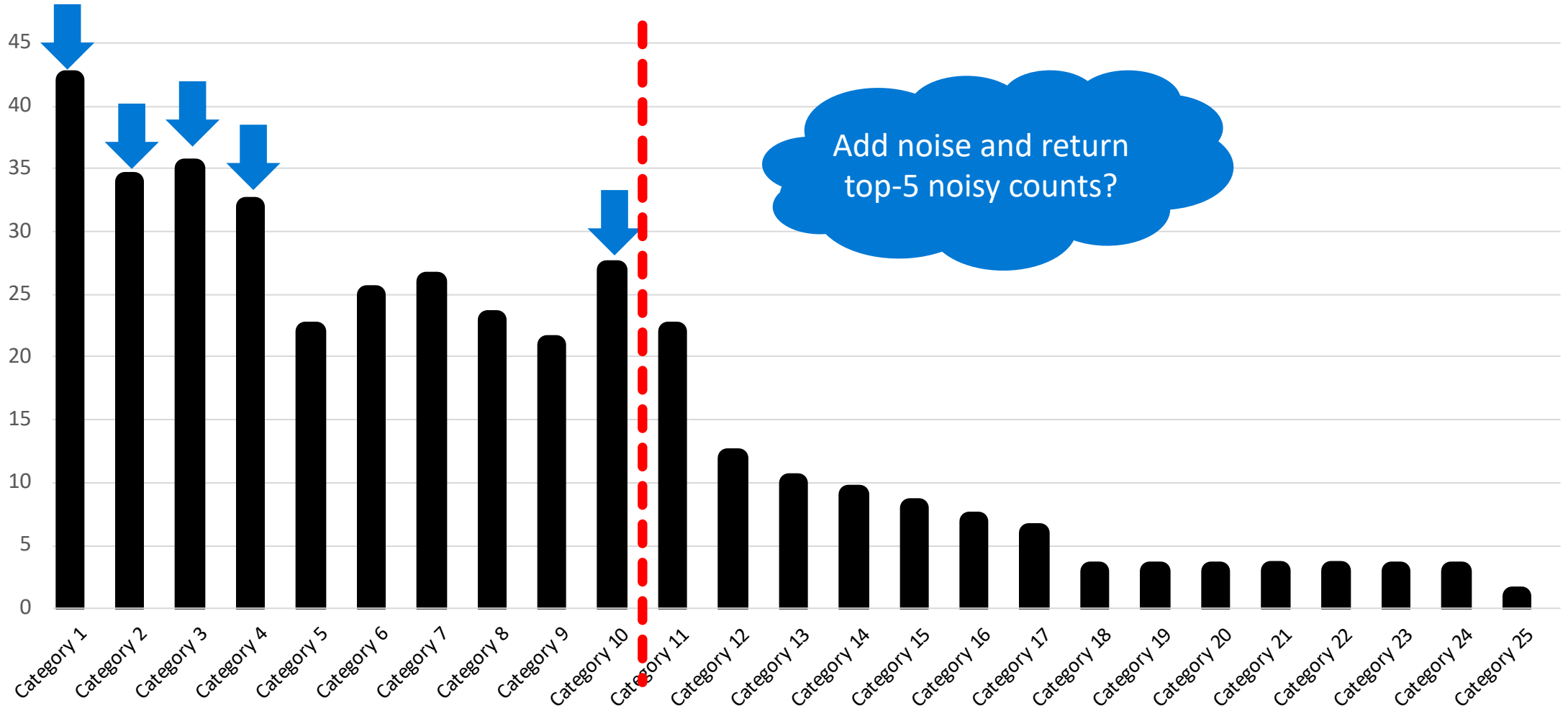
Existing Systems for Data Analytics



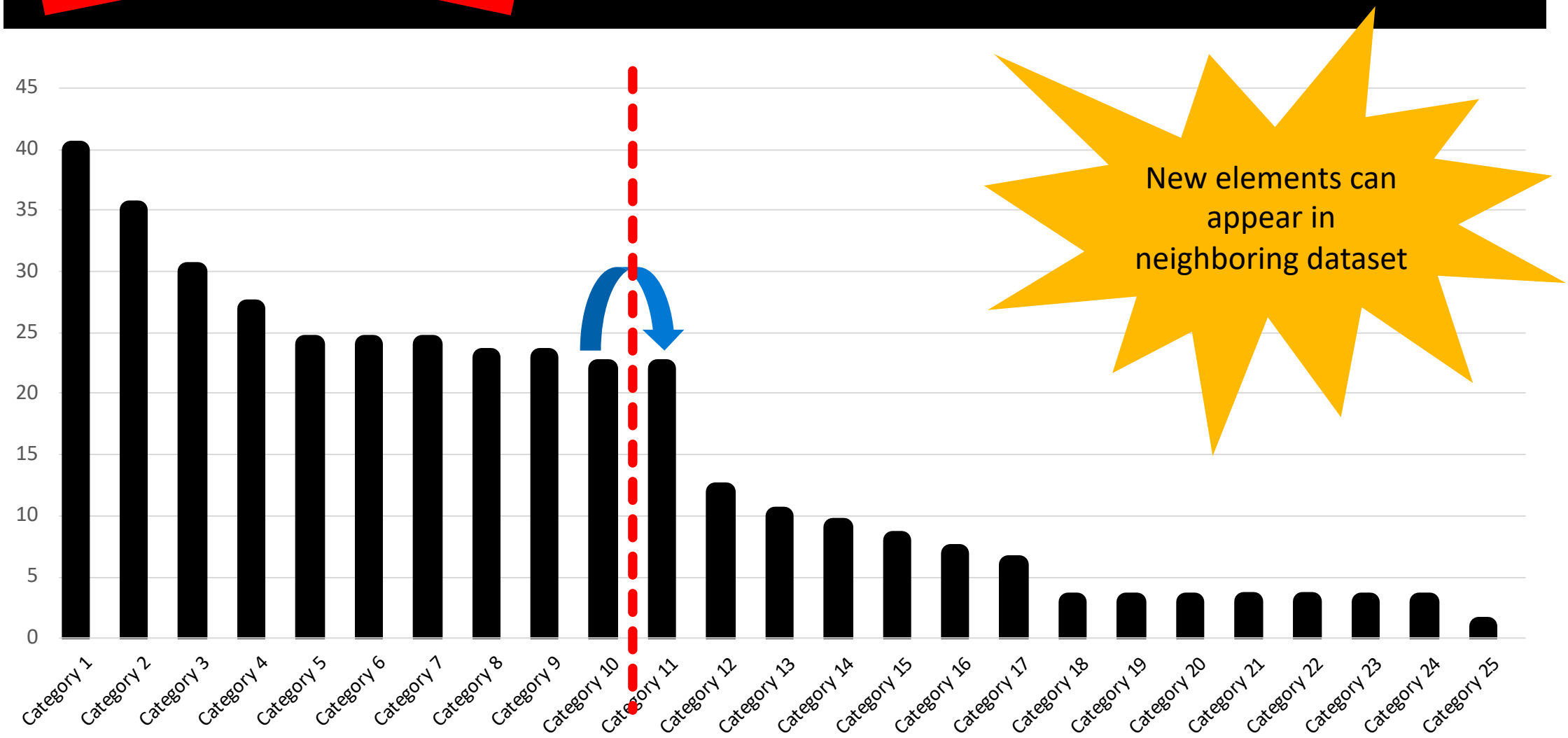
First Attempt



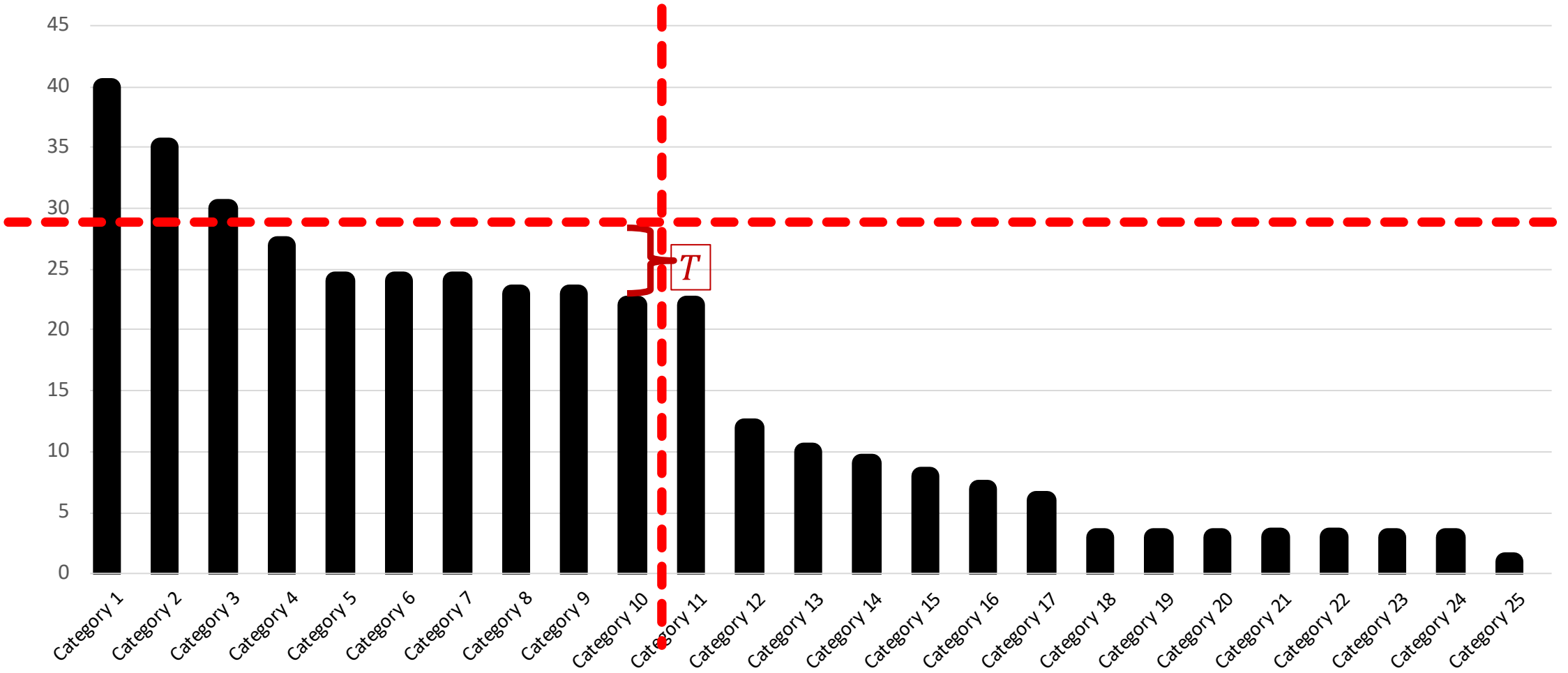
First Attempt



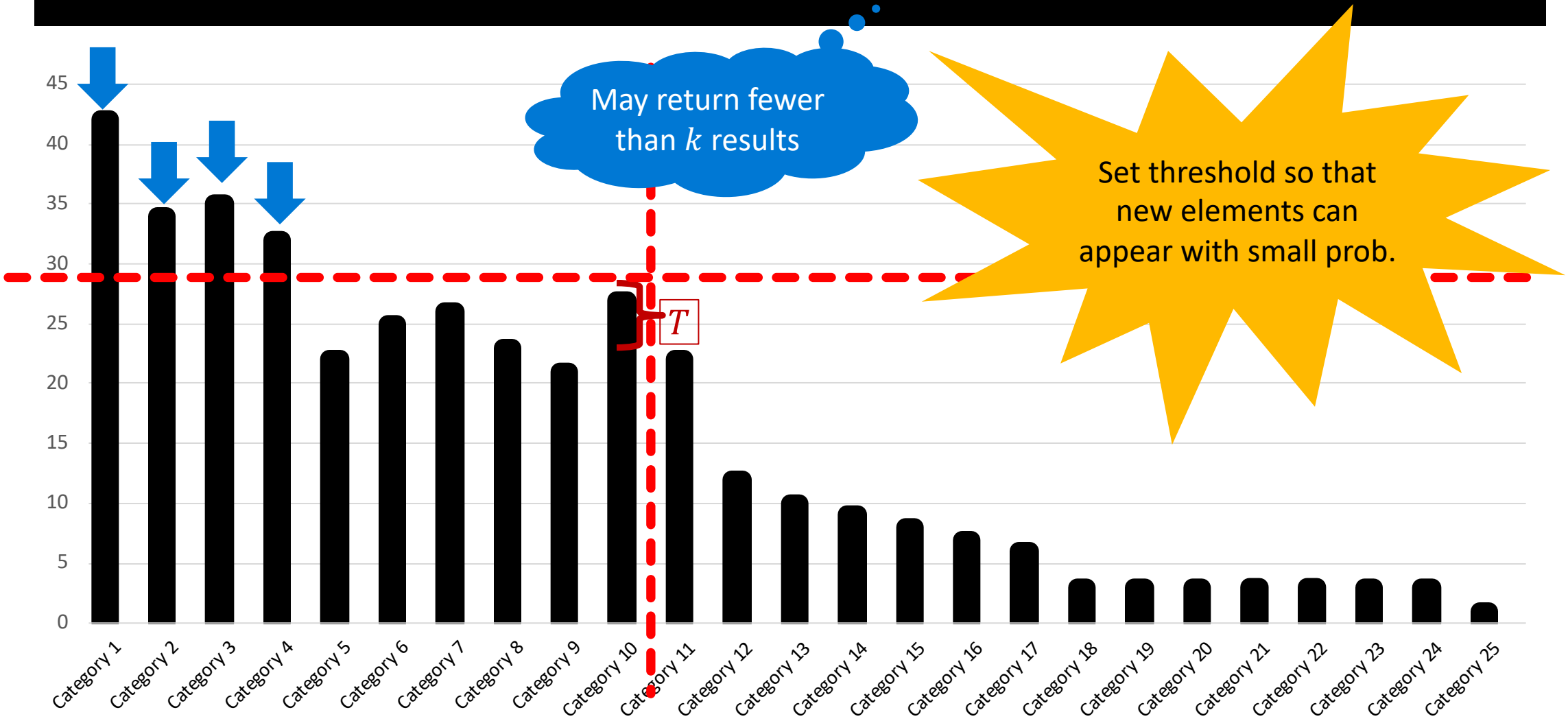
~~First Attempt~~



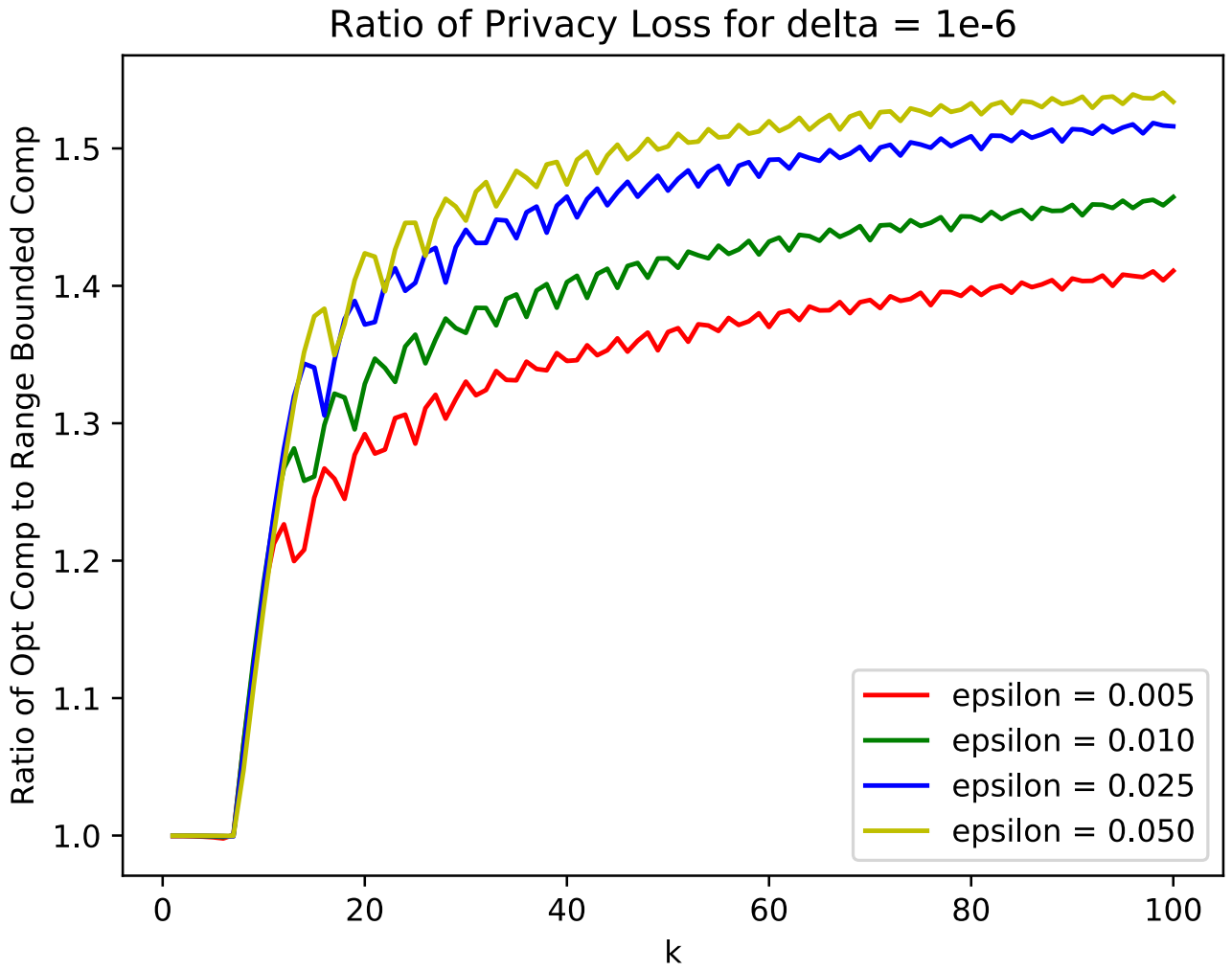
Second Attempt – Include Threshold



Second Attempt – Include Threshold



BR Composition versus Optimal DP Composition



Come to Poster #161 for more details