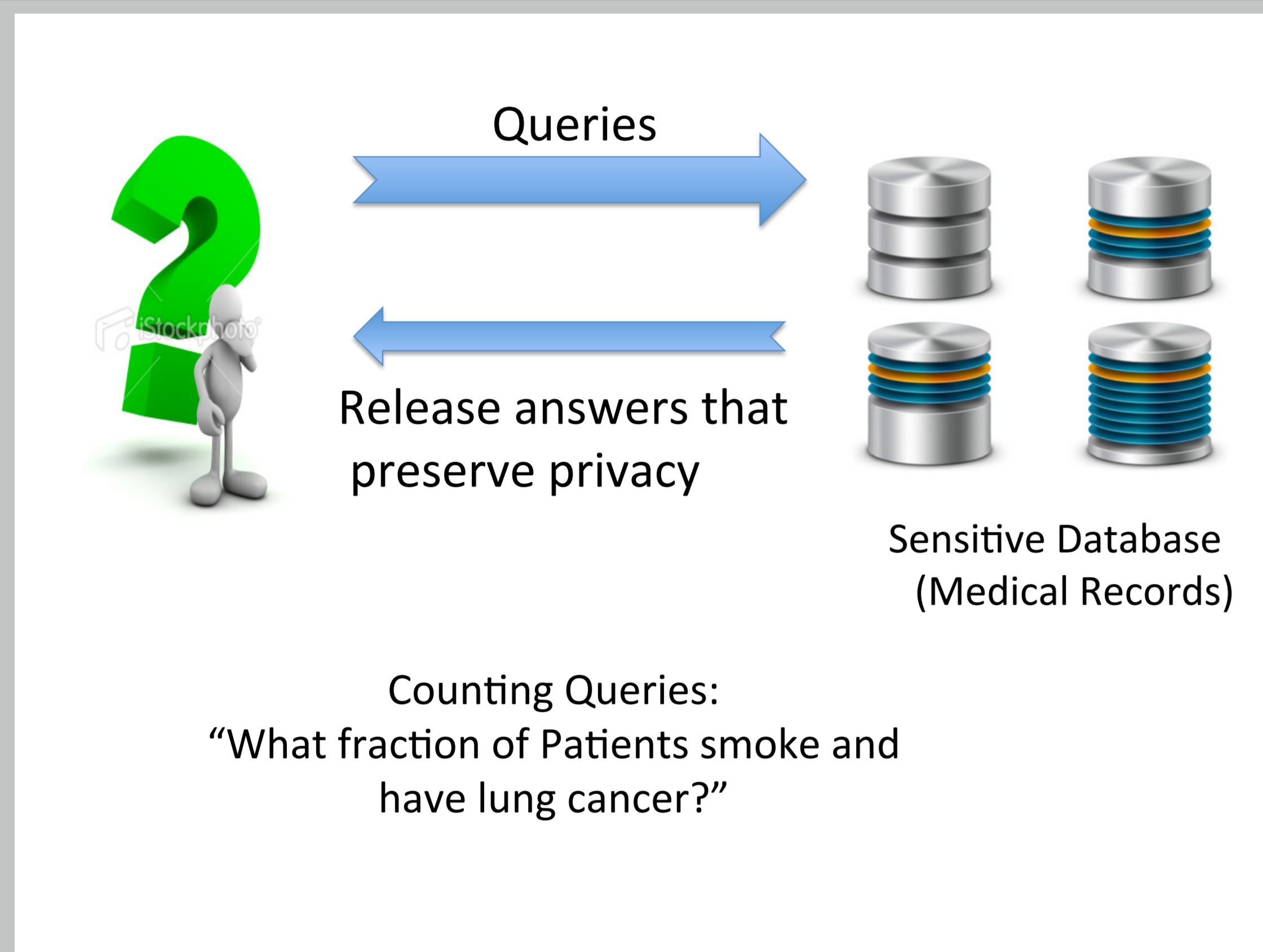


# Dual Query: Practical Private Query Release for High Dimensional Data

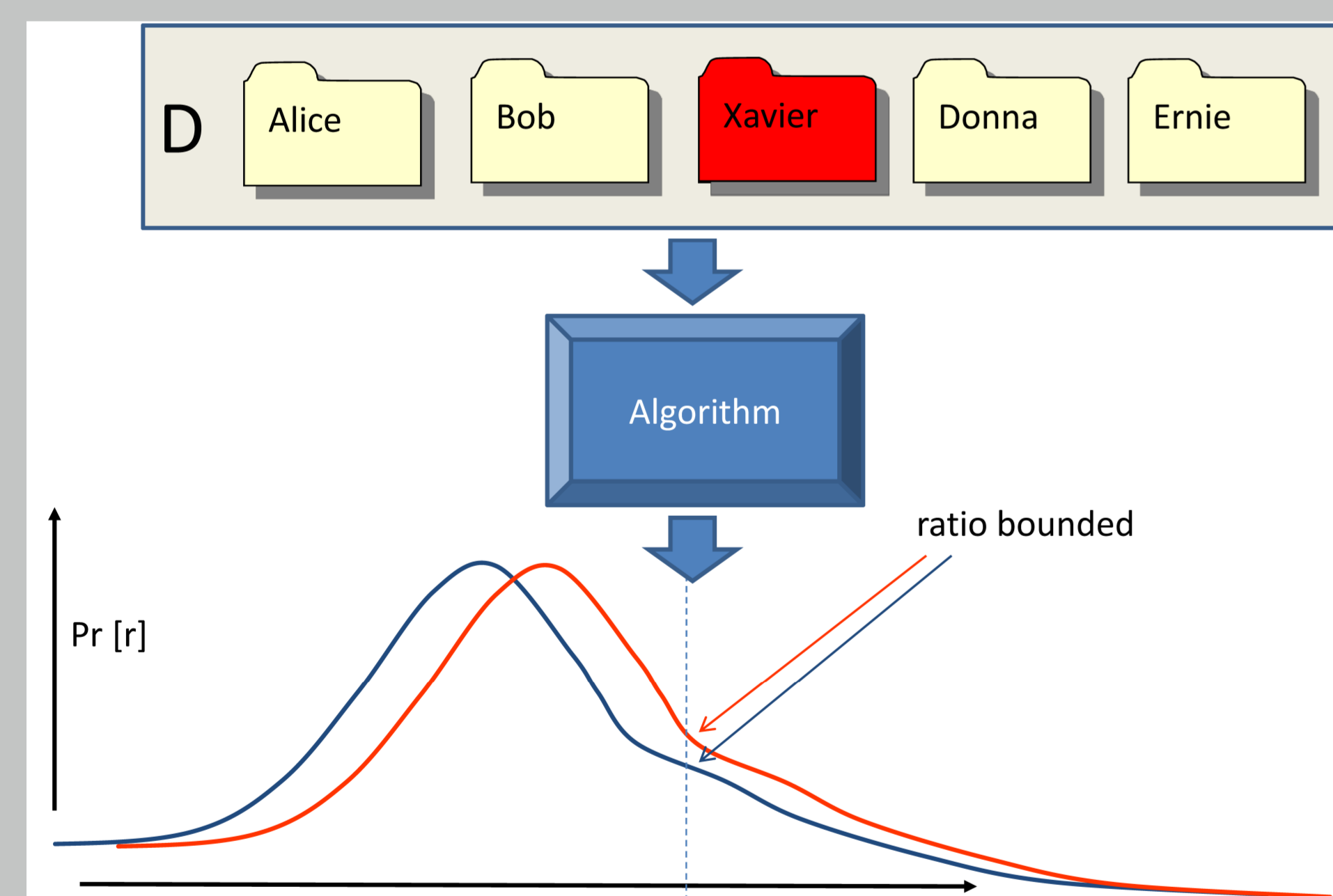
Marco Gaboardi, Emilio Jesús Gallego Arias, Justin Hsu, Aaron Roth, and Zhiwei Steven Wu  
ICML 2014



## Private Query Release



## Differential Privacy [DMNS06]



Neighboring Databases  $D$  and  $D'$ :  
 $\Pr[A(D) = r] \lesssim (1 + \epsilon)\Pr[A(D') = r]$

## Query Release as a Zero-Sum Game

Query Player Maximizes while Data Player Minimizes

- ▶ Actions for query player: query class  $\mathcal{Q}$
- ▶ Actions for data player: possible data records  $\mathcal{X} = \{0, 1\}^d$
- ▶ Payoff on  $(q, x)$  is  $q(D) - q(x)$
- ▶ Approximate Minimax Equilibrium  $\Rightarrow$  Accurate Answers

## Find the Equilibrium with No-Regret Learning

No-Regret Algorithm vs. Best Response  
→ converge to Equilibrium

- ▶ Previous idea: Data player runs no-regret learning
- ▶ Maintain approximate database  $\hat{D}$ , privately find queries with high error, update  $\hat{D}$  [HR10][HLM12]
- ▶  $\hat{D}$  is distribution over  $\mathcal{X}$  (HUGE!  $2^d$ )
- ▶ Problem: not scalable for high dimensional data. Existing work:  $\sim 100$  attributes [HLM12].

## Our Novelty: Switching the Roles

Query player runs no-regret learning

- ▶ Now: distribution over queries  $\mathcal{Q}$ , find record minimizing error
- ▶ Makes High Dimensional Data Possible!
- ▶ Space linear in  $|\mathcal{Q}|$  rather than  $|\mathcal{X}|$
- ▶ Best response problem for data player is NP-Hard but non-private and succinctly represented, can use existing solvers like CPLEX

## Theoretical Accuracy Guarantee

Max additive error over all queries (error 1 trivial):

$$O\left(\frac{\log |\mathcal{Q}|}{|\mathcal{D}|^{1/3}\epsilon^{1/3}}\right)$$

## Experimental Accuracy

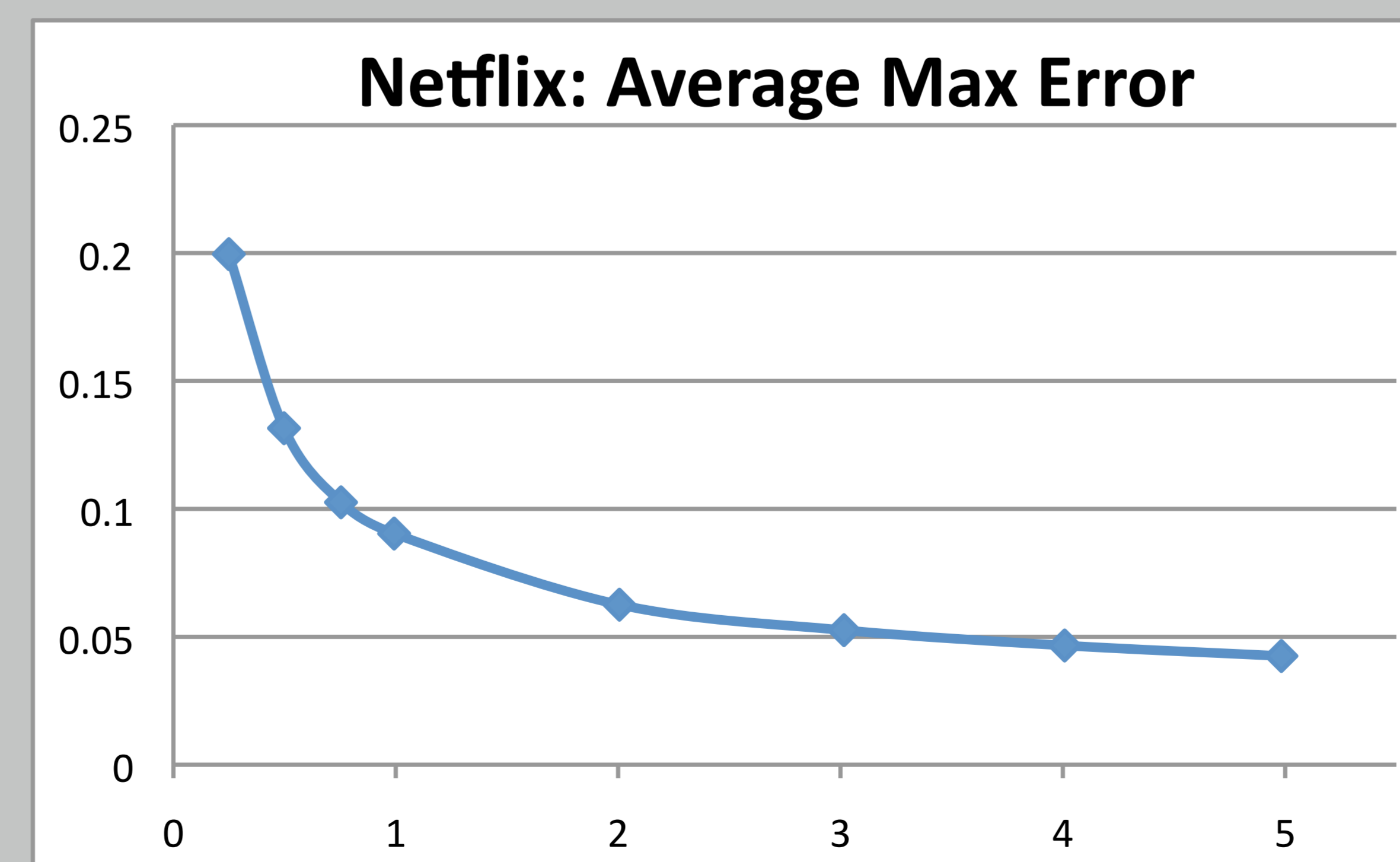


Figure 1: Accuracy versus  $\epsilon$  (privacy)

## Scaling with Number of Attributes

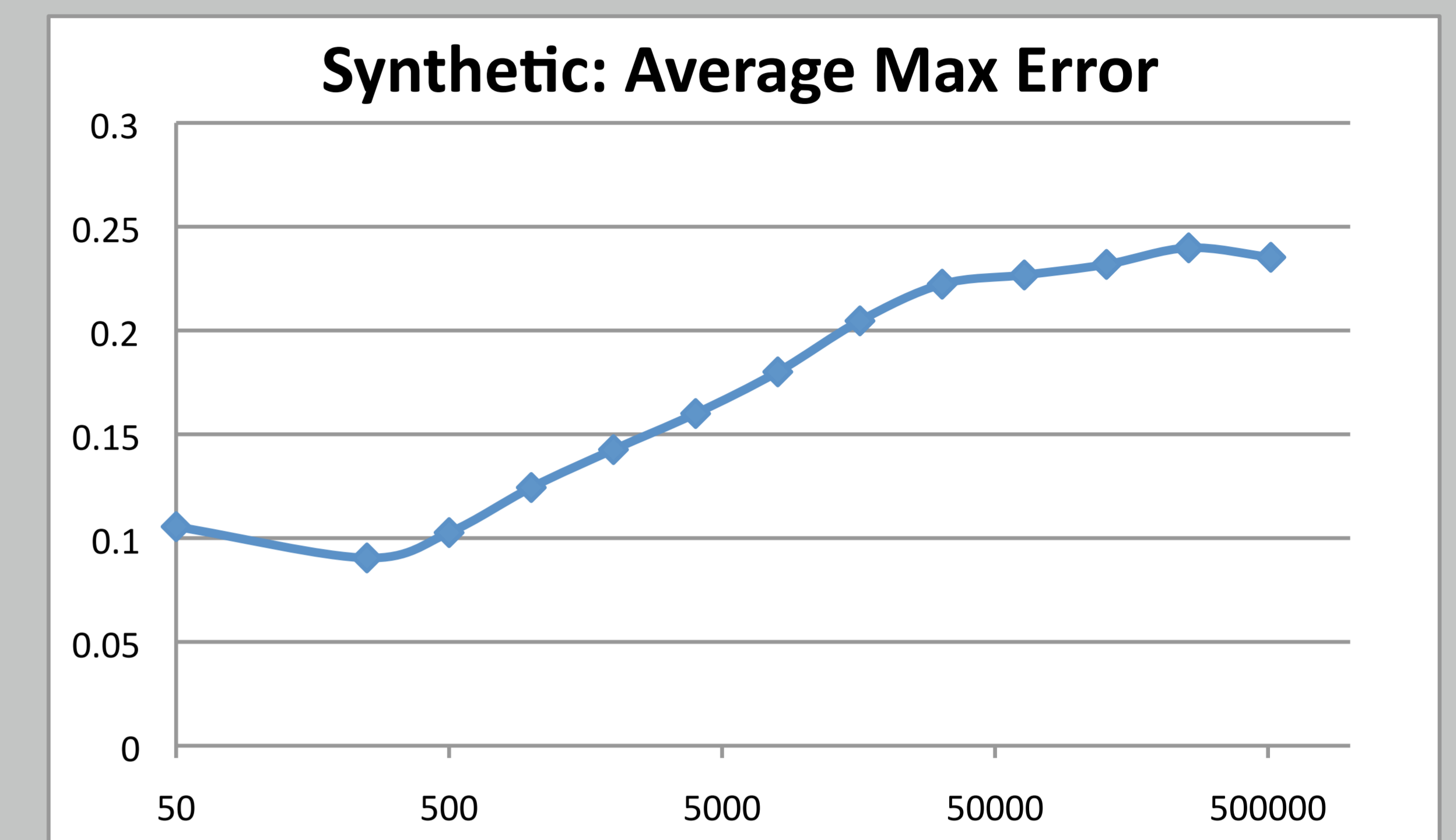


Figure 2: Accuracy versus number of attributes

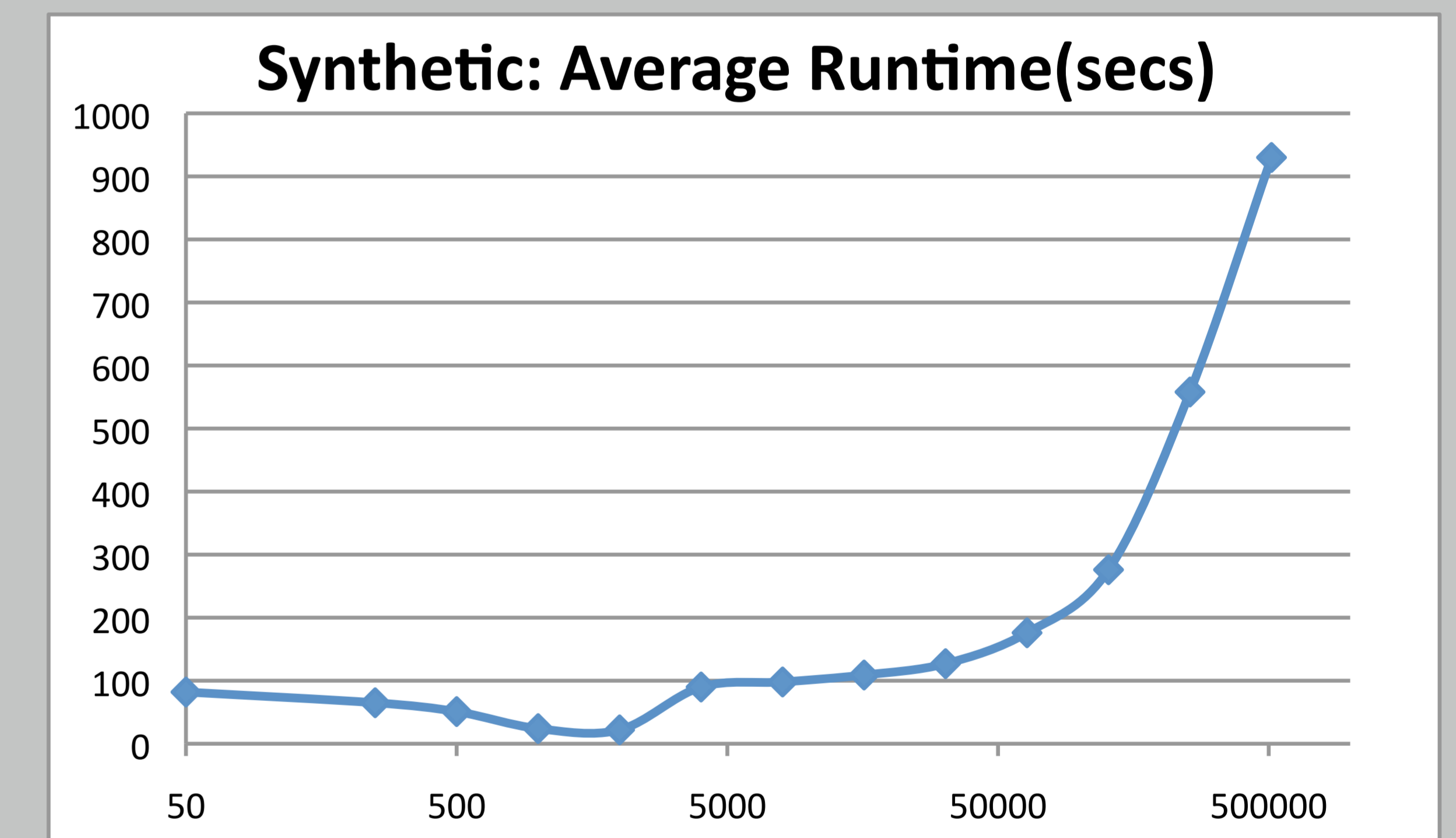


Figure 3: Runtime versus number of attributes

## Conclusion and Open Problems

- ▶ Dual Query: A new private query release mechanism that can handle datasets with dimensionality multiple orders of magnitude larger than what was previously possible.
- ▶ Open problems:
  - ▷ Parameter setting under differential privacy
  - ▷ Incorporate sparsity of the dataset
  - ▷ Subclass of queries with "easy" best response problem
  - ▷ Allow queries to arrive online