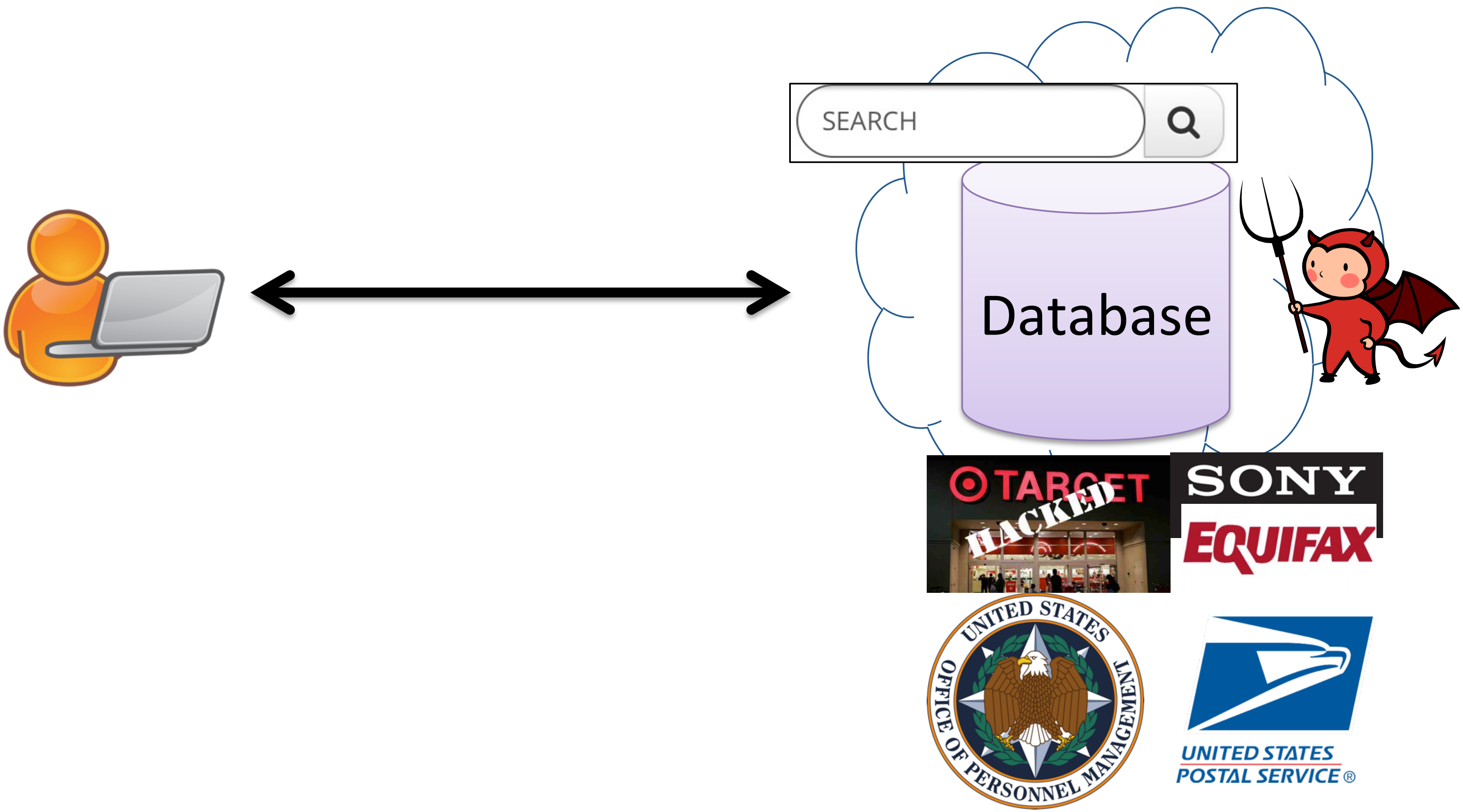
Learning to Reconstruct: Statistical Learning Theory and Encrypted Database Attacks

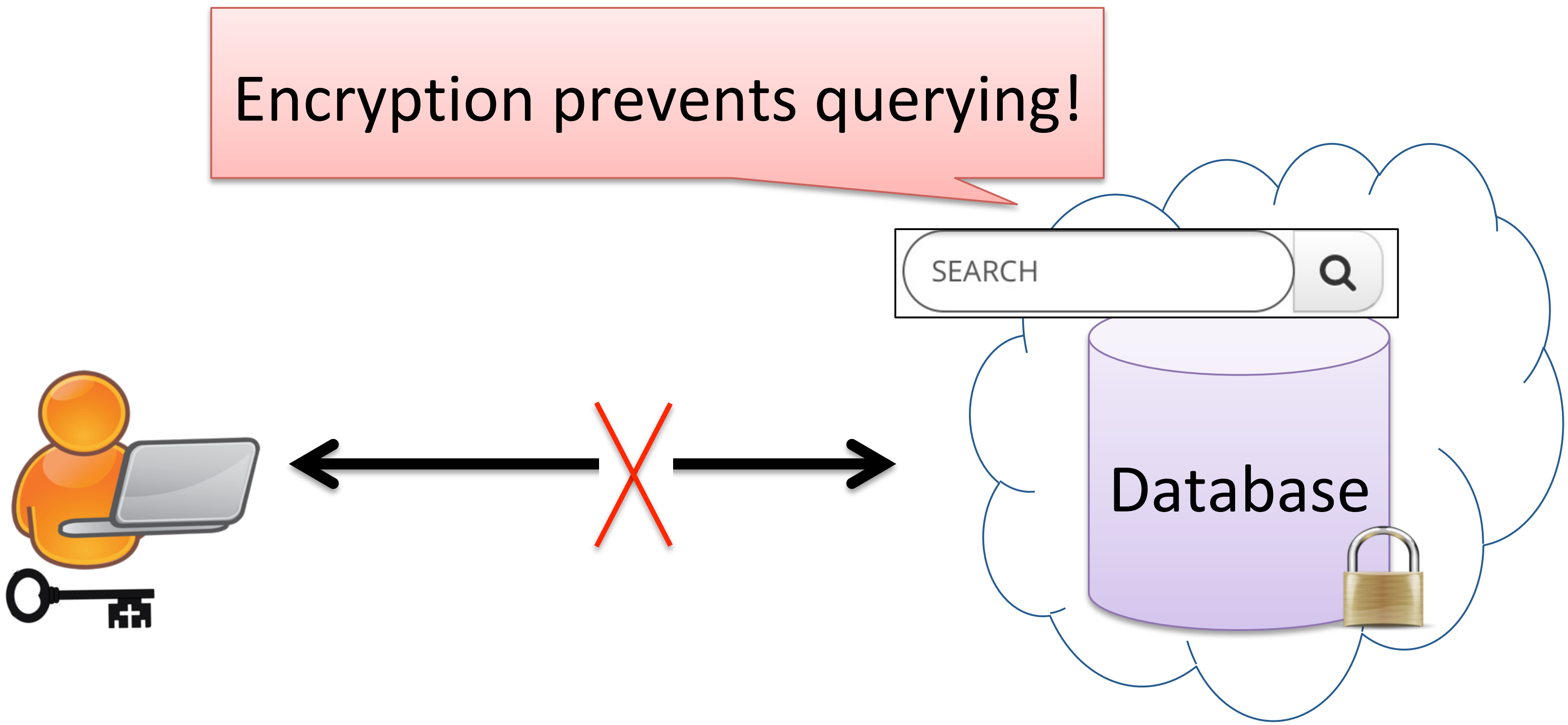
Paul Grubbs, Marie-Sarah Lacharité, Brice Minaud, Kenny Paterson

pag225@cornell.edu, @pag_crypto

Outsourced Databases

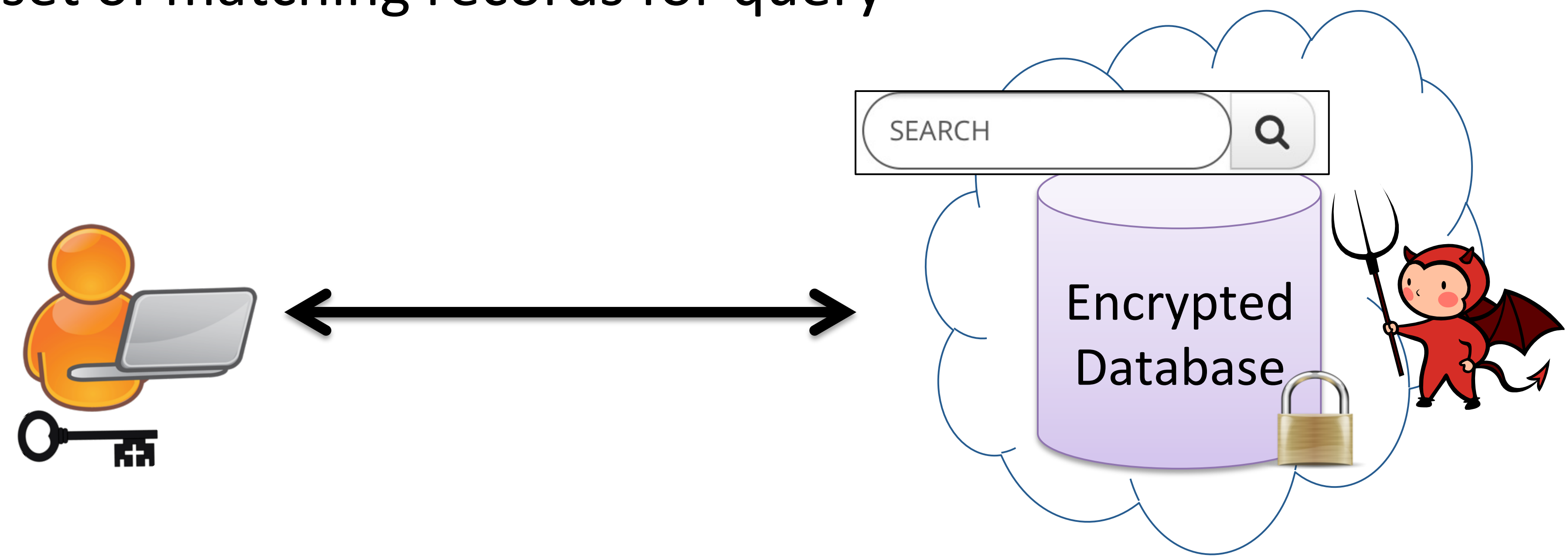


Encrypting Outsourced Databases?



Encrypted Databases

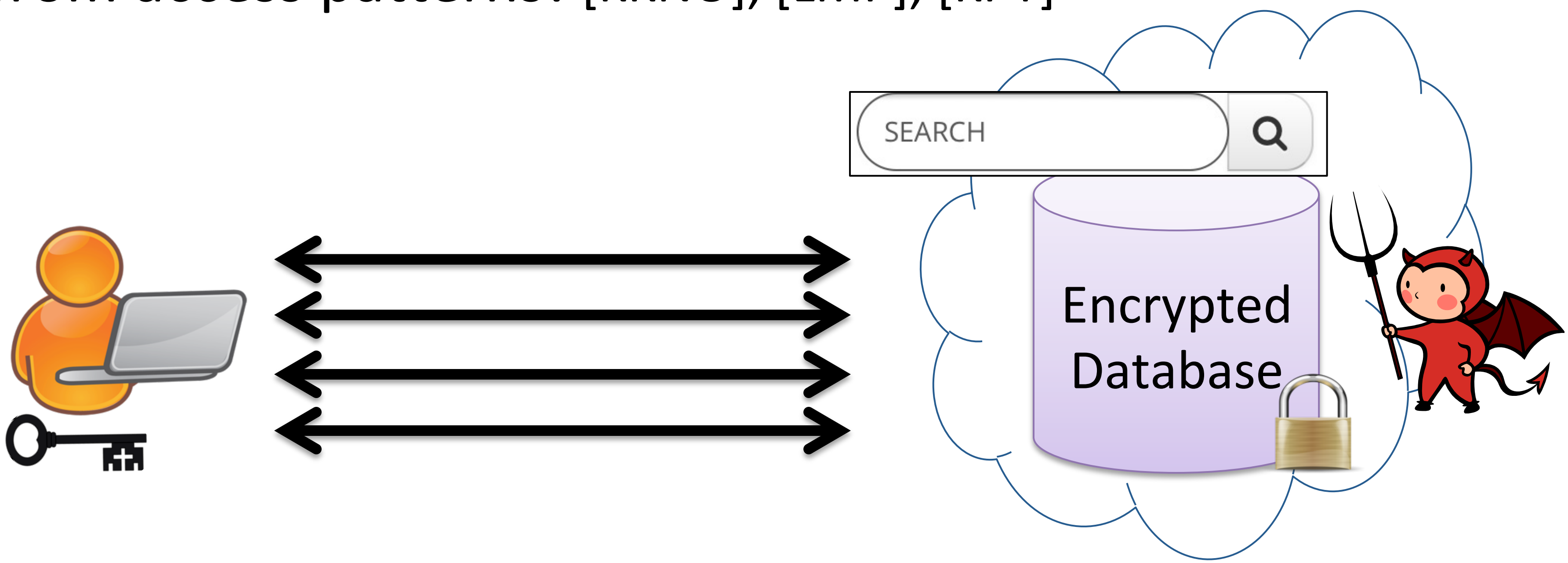
Efficient ones leak **access patterns:**
set of matching records for query



What can an attacker learn
from access pattern leakage?

Database Reconstruction (DR)

With enough queries, can learn data from access patterns! [KKNO], [LMP], [KPT]



Prior work:

huge numbers of queries, strong assumptions, specific query types.

[KKNO]: 10^{26} for salaries
[LMP]: dense database
[KPT]: kNN queries only

Our Contributions

- Enabling insight: access pattern leakage is a binary classification
  Use statistical learning theory (SLT) to build and analyze attacks
- New DR attacks on range queries
  Generalize and improve [KKNO], [LMP] with SLT + PQ trees
  On real data: with only 50 queries, predict salaries to 2% error
- Generic reduction from DR with known queries to PAC learning
- Give "minimal" attack for all query types via ε-nets
  Instantiate with first DR attack for prefix queries
- First general lower bound on #queries needed for DR

Full version: https://eprint.iacr.org/2019/011

Our Contributions

- Enabling insight: access pattern leakage is a binary classification
  - Use statistical learning theory (SLT) to build and analyze attacks
- New DR attacks on range queries
  - Generalize and improve [KKNO], [LMP] with SLT + PQ trees
  - On real data: with only 50 queries, predict salaries to 2% error
- Generic reduction from DR with known queries to PAC learning
- Give "minimal" attack for all query types via ε-nets
  - Instantiate with first DR attack for prefix queries
- First general lower bound on #queries needed for DR

Full version: https://eprint.iacr.org/2019/011

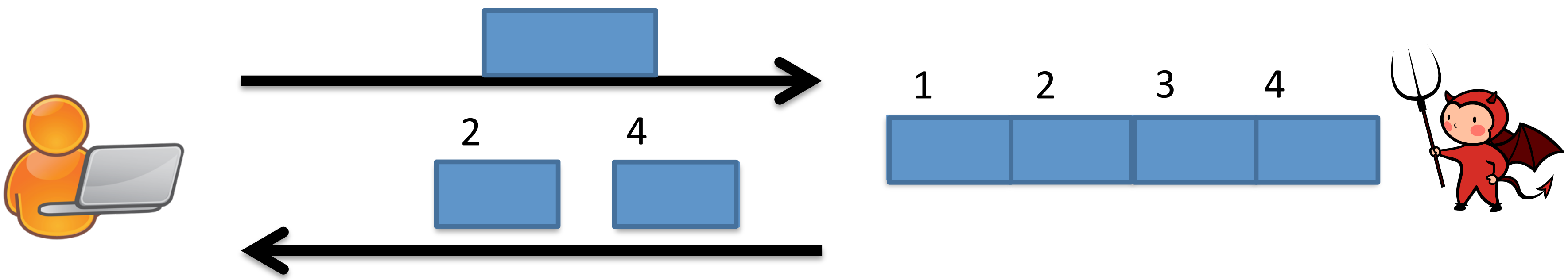
Notation and Terminology

N: number of possible values, wlog [1, …, N]

E.g., N=125 for age data

Range query: is a pair [a, b] where $1 \le a \le b \le N$.

Database: is composed of ***records***, each with values in [1, …, N]



Access Pattern: which records match

Full database reconstruction (DR): recovering exact record values

Approximate DR: recovering all record values within εN.

ε = 0.05 is recovery within 5%. ε = 1/N is full DR.

Scale-free: query complexity independent of #records or N.

DR For Range Queries: Our Work

[KKNO16]: full DR in O(N^4 log N) queries

Generalizes

Three attacks:

	Full DR	Lower Bound
GeneralizedKKNO: O($\varepsilon^{-4} \log \varepsilon^{-1}$) for approx. DR	$\boxed{O(N^4 \log N)}$	$\Omega(\varepsilon^{-4})$
ApproxValue: O($\varepsilon^{-2} \log \varepsilon^{-1}$) approx. DR*	O(N^2 log N)	$\Omega(\varepsilon^{-2})$
ApproxOrder: O($\varepsilon^{-1} \log \varepsilon^{-1}$) for approx. *order* rec*	$\boxed{O(N \log N)}$	$\Omega(\varepsilon^{-1} \log \varepsilon^{-1})$

With DB distribution info, get approx. DR

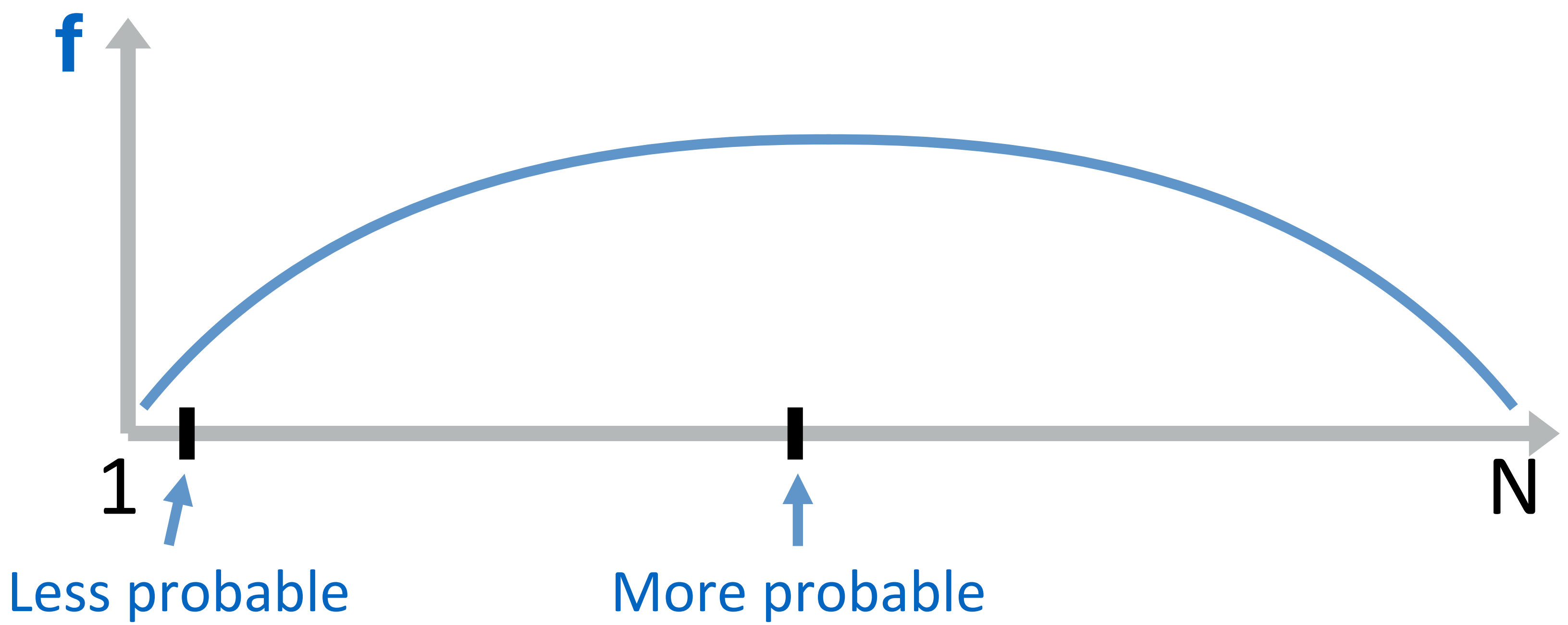
Implies

[LMP18]: Full DR for dense DB in O(N log N).

*Requires a mild hypothesis about data

Bypass [LMP] lower bound via relaxing to "sacrificial" recon.

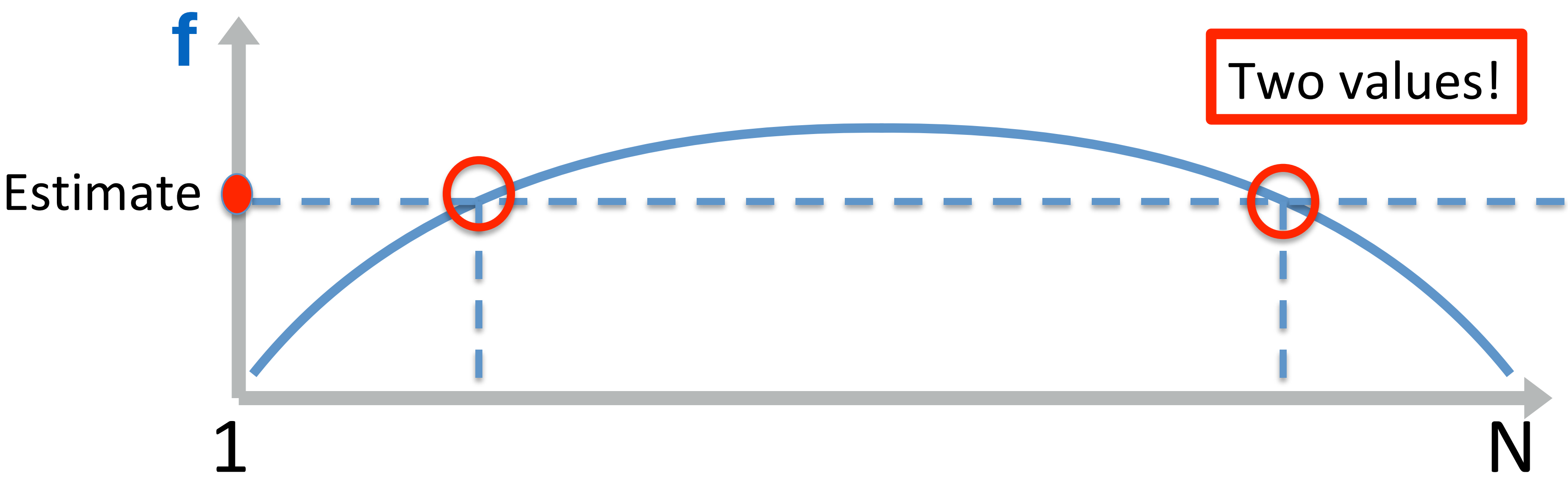
GeneralizedKKNO



Assume **uniform distribution** on range queries + static database.

Induces a distribution **f** on the probability that a value is accessed.

GeneralizedKKNO



Idea: for each record…

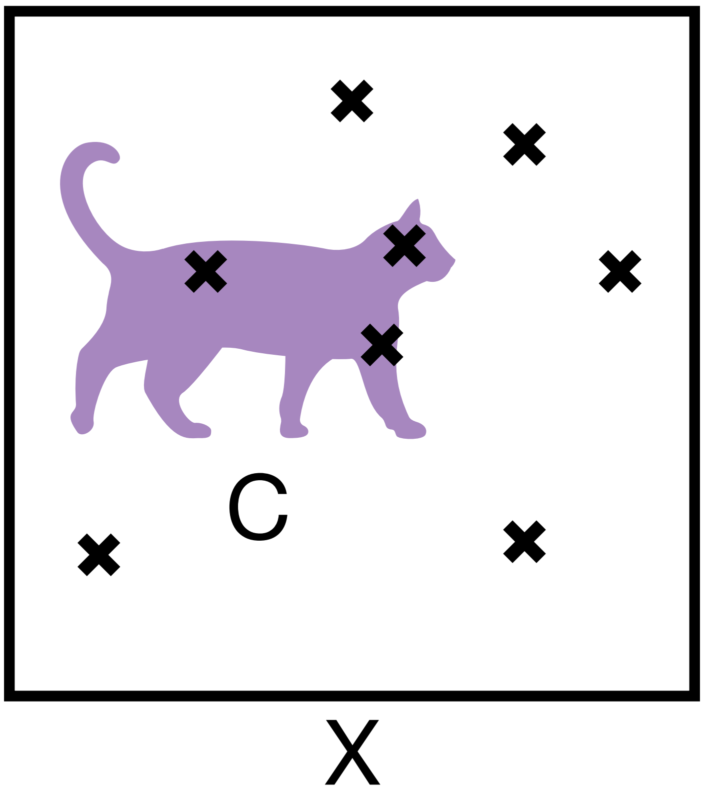
1. Count #accesses to estimate **f**(value)

How many queries to get estimate sufficient for ε approx. DR?

More work needed to break symmetry. See paper for details

Estimating a Probability

Set X with probability distribution D.
Let $C \subseteq X$ be a set.

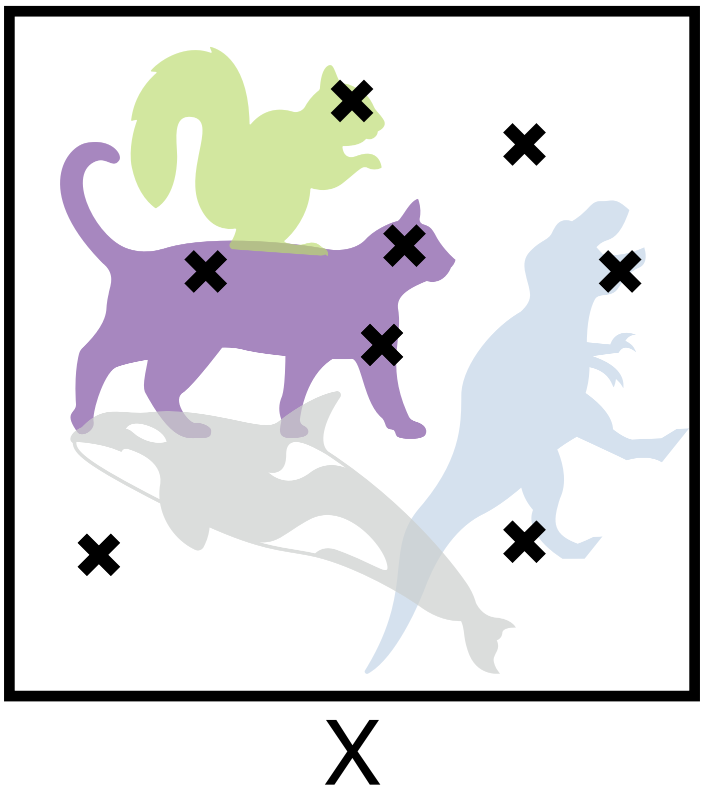


$$\Pr(C) \approx \frac{\#\text{points in } C}{\#\text{points total}}$$

Sample complexity:
to measure $\Pr(C)$ within ε,
you need $O(1/\varepsilon^2)$ samples.

Estimating a Set of Probabilities

Now: set of sets $\mathcal{C}$.
Goal: estimate all sets' probabilities *simultaneously*.

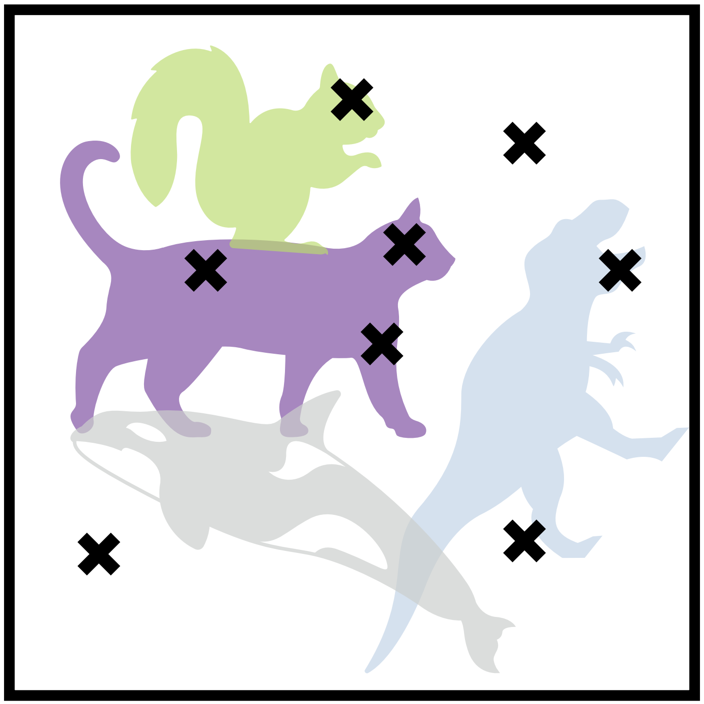


The set of samples drawn from X is an **ε-sample** iff for **all** C in $\mathcal{C}$:

$$\left| \Pr(C) - \frac{\#\text{points in } C}{\#\text{points total}} \right| \leq \epsilon$$

The ε-sample Theorem

How many points do we need to draw to get an ε-sample w.h.p.?



X

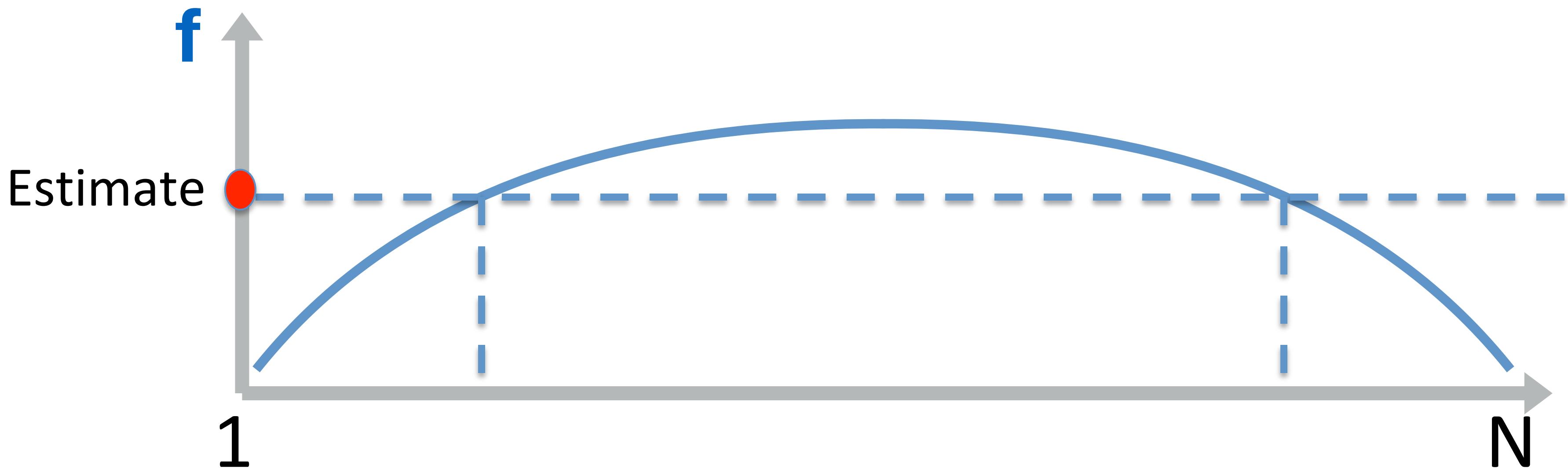
V & C 1971:

If $\mathcal{C}$ has **VC dimension** d, then the number of points to get an ε-sample whp is

$$O\left(\frac{d}{\epsilon^2}\log\frac{d}{\epsilon}\right).$$

Does not depend on $|\mathcal{C}|$*!*

GeneralizedKKNO



Idea: for each record…

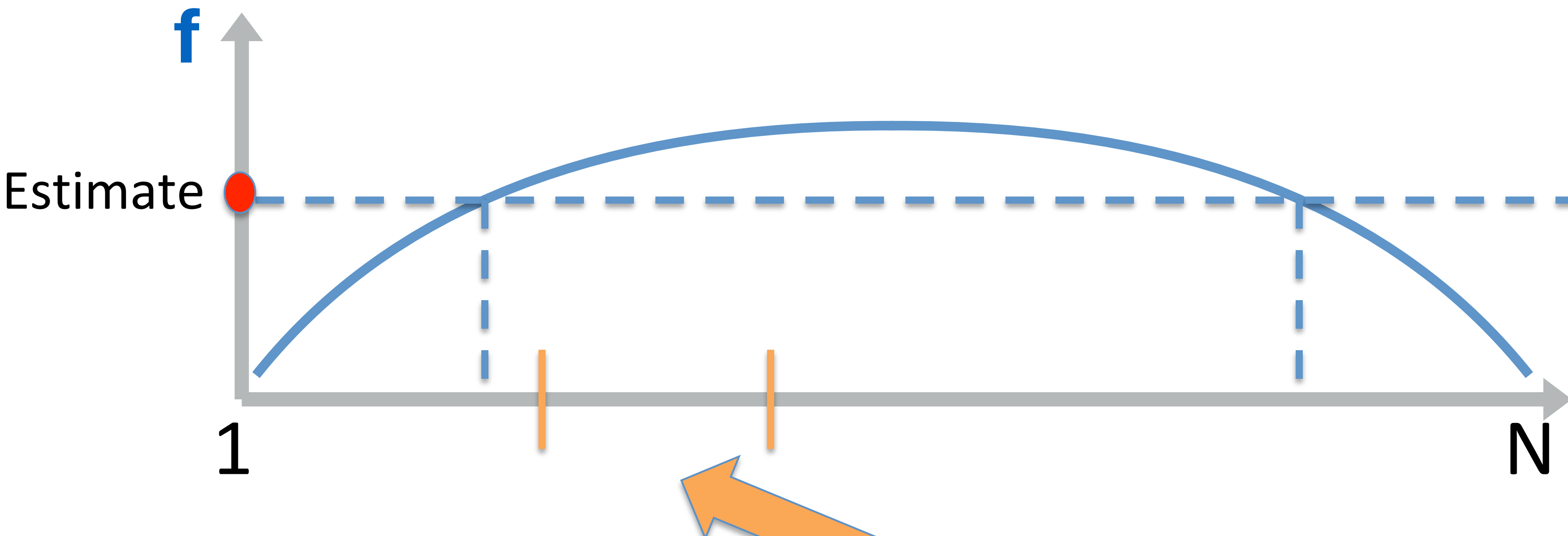
1. Count #accesses to estimate **f**(value)
2. Find value by "inverting" **f** estimate

This is an ε-sample!

X = range queries $\mathcal{C}$ ={{range queries ∋ x}: x ∈ [1,N]} VC dim. = 2

Can we get rid of squaring?

GeneralizedKKNO



Assume there exists at least one record in [N/8, 3N/8].

Idea: for each record…

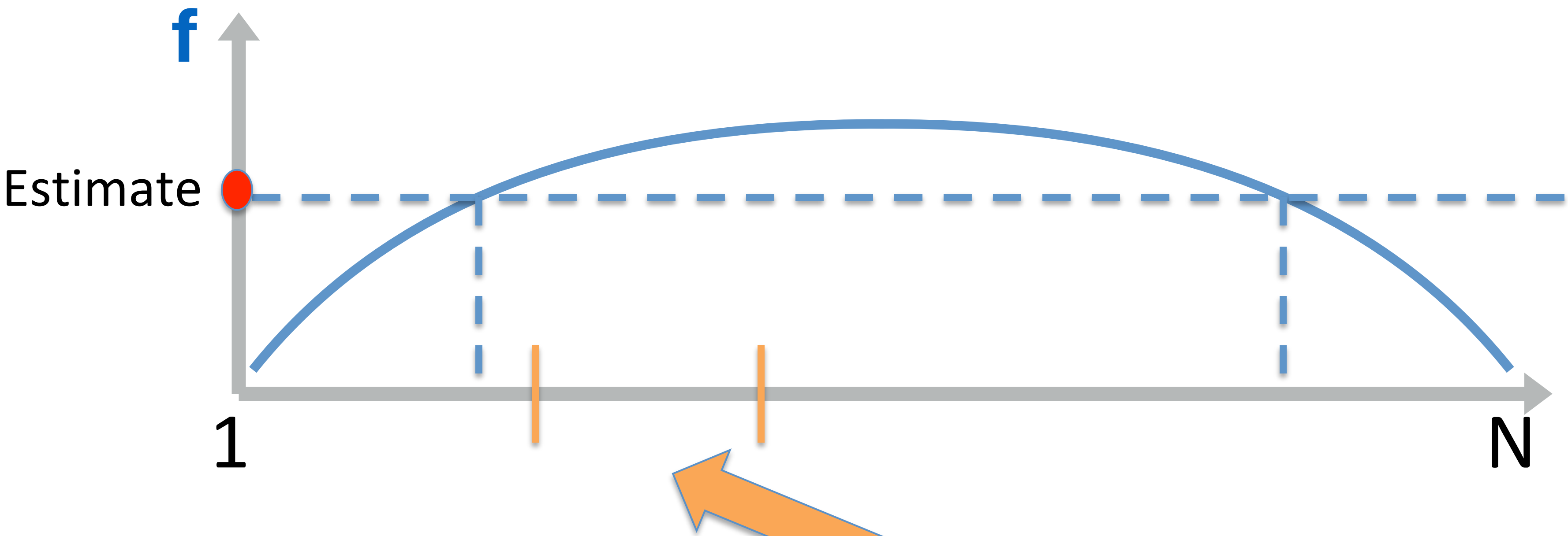
1. Count #accesses to estimate **f**(value)
2. Find value by "inverting" **f** estimate

This is an ε-sample!

X = range queries $\mathcal{C}$ ={{range queries ∋ x}: x ∈ [1,N]} VC dim. = 2

We need O(ε^{-4} log ε^{-1}) queries (inverting **f** adds a square)

ApproxValue



Idea: for each record…

1. Count #accesses to estimate **f**(value)
2. Find value by "inverting" **f** estimate

Assume there exists at least one record in [N/8, 3N/8].

This is an ε-sample!

X = range queries $\mathcal{C}$ ={{range queries ∋ x}: x ∈ [1,N]} VC dim. = 2

We need $O(\varepsilon^{-2} \log \varepsilon^{-1})$ queries! More complex attack – see paper

DR For Range Queries: Our Work

Three attacks:

	Full DR	Lower Bound
GeneralizedKKNO: $O(\varepsilon^{-4} \log \varepsilon^{-1})$ for approx. DR	$O(N^4 \log N)$	$\Omega(\varepsilon^{-4})$
ApproxValue: $O(\varepsilon^{-2} \log \varepsilon^{-1})$ approx. DR*	$O(N^2 \log N)$	$\Omega(\varepsilon^{-2})$
ApproxOrder: $O(\varepsilon^{-1} \log \varepsilon^{-1})$ for approx. *order* rec*	$O(N \log N)$	$\Omega(\varepsilon^{-1} \log \varepsilon^{-1})$

With DB distribution info, get approx. DR

Require iid uniform queries, adversary knows query distribution. What can we do without making these assumptions?

DR For Range Queries: Our Work

Three attacks:

	Full DR	Lower Bound
GeneralizedKKNO: $O(\varepsilon^{-4} \log \varepsilon^{-1})$ for approx. DR	$O(N^4 \log N)$	$\Omega(\varepsilon^{-4})$
ApproxValue: $O(\varepsilon^{-2} \log \varepsilon^{-1})$ approx. DR*	$O(N^2 \log N)$	$\Omega(\varepsilon^{-2})$
ApproxOrder: $O(\varepsilon^{-1} \log \varepsilon^{-1})$ for approx. *order* rec*	$O(N \log N)$	$\Omega(\varepsilon^{-1} \log \varepsilon^{-1})$

With DB distribution info, get approx. DR

Reveal order without no assumptions on query distribution. See paper for details

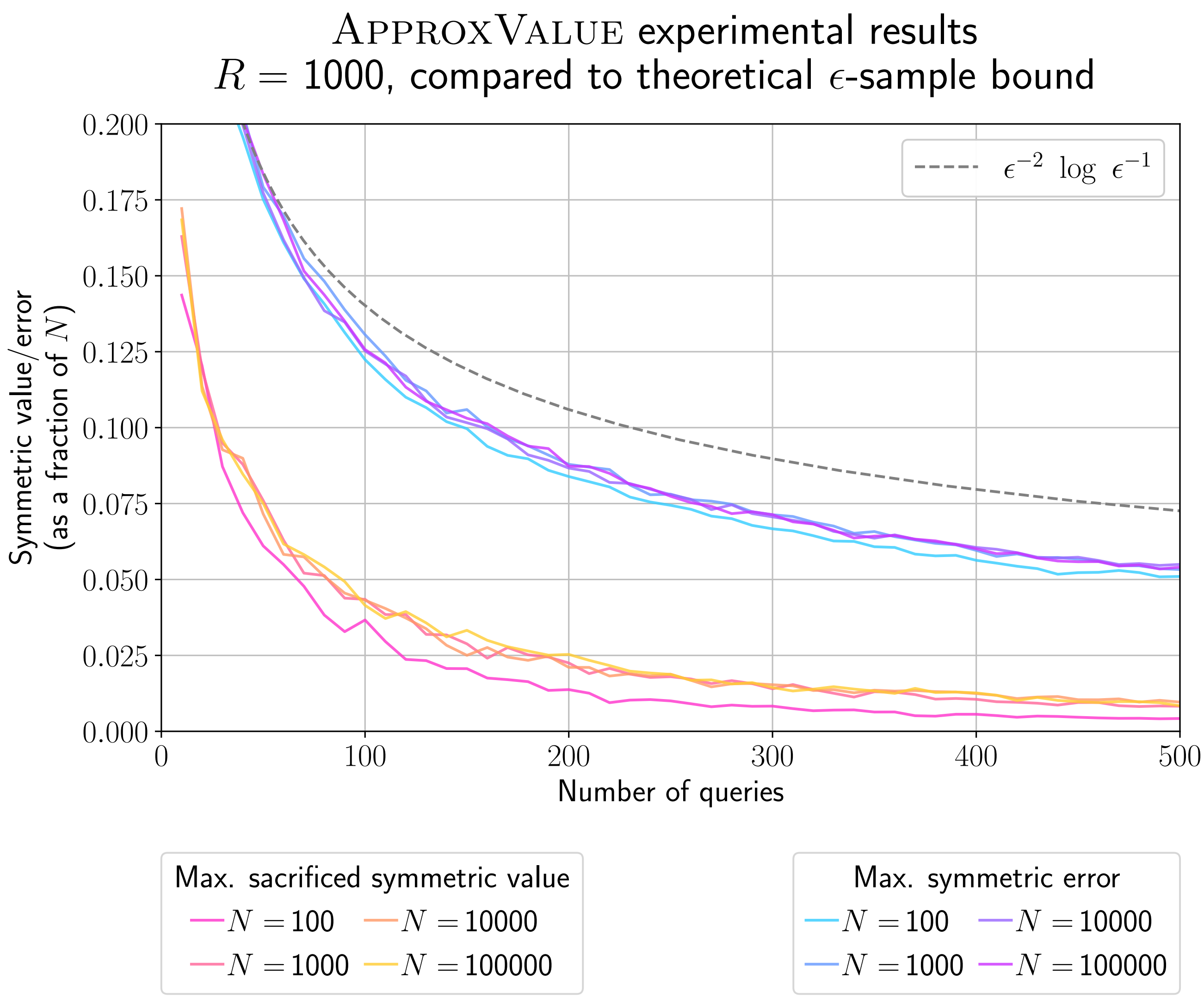
Conclusion

- Enabling insight: access pattern leakage is a binary classification
  Use statistical learning theory (SLT) to build and analyze attacks
- New DR attacks on range queries
  Generalize and improve [KKNO], [LMP] with SLT + PQ trees
  On real data: with only 50 queries, predict salaries to 2% error
- Generic reduction from DR with known queries to PAC learning
- Give "minimal" attack for all query types via ε-nets
  Instantiate with first DR attack for prefix queries
- First general lower bound on #queries needed for DR

Full version: https://eprint.iacr.org/2019/011

Thanks for listening!
Any questions?

Attack Simulation



Effective constants are ~ 1!

DR As Learning a Binary Classifier

This formulation is not specific to range queries!

$$\underbrace{\phantom{\{\{\text{range queries} \ni x\}: x \in [1,N]\}}}_{}$$

Record values are binary classifiers

X = ~~range~~ queries $\quad \mathcal{C} = \{\{\cancel{\text{range}}\ \text{queries} \ni x\}: x \in [1,N]\}$

Approximately learning classifier => approximate DR