Building Useful Systems That Protect People and Their Data

Johes Bater

Tufts University | School of Engineering
Computer Science

Organizations collect, store, and process user data to produce valuable insights

Organizations con … romise user data

List of data breaches

From Wikipedia, the free encyclopedia

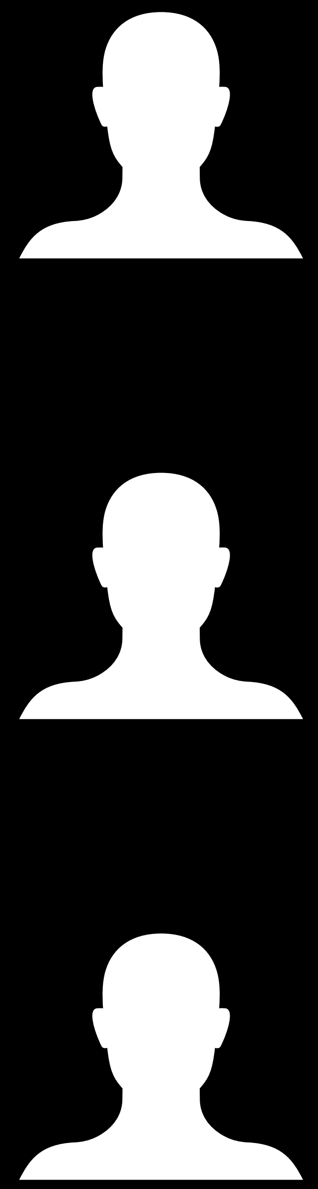
For broader coverage of this topic, see Data breach.
For broader coverage of this topic, see List of security hacking incidents.
This is a dynamic list and may never be able to satisfy particular standards for completeness. You can help by adding missing items with reliable sources.

This is a list of **data breaches**, using data compiled from various sources, including press reports, government news releases, and mainstream news articles. The list includes those involving the theft or compromise of 30,000 or more records, although many smaller breaches occur continually. Breaches of large organizations where the number of records is still unknown are also listed. In addition, the various methods used in the breaches are listed, with hacking being the most common.

Most breaches occur in North America. It is estimated that the average cost of a data breach will be over $150 million by 2020, with the global annual cost forecast to be $2.1 trillion.[1][2] As a result of data breaches, it is estimated that in first half of 2018 alone, about 4.5 billion records were exposed.[3] In 2019, a collection of 2.7 billion identity records, consisting of 774 million unique email addresses and 21 million unique passwords, was posted on the web for sale.[4]

Entity	Year	Records	Organization type	Method	Sources
Yahoo	2013	3,000,000,000	web	hacked	[391][392]
First American Corporation	2019	885,000,000	financial service company	poor security	[152]
Facebook	2019	540,000,000	social network	poor security	[145][146]
Marriott International	2018	500,000,000	hotel	hacked	[232]
Yahoo	2014	500,000,000	web	hacked	[393][394][395][396][397]
Friend Finder Networks	2016	412,214,295	web	poor security / hacked	[156][157]
Exactis	2018	340,000,000	data broker	poor security	[133]
Airtel	2019	320,000,000	telecommunications	poor security	[18]
Truecaller	2019	299,055,000	Telephone directory	unknown	[337][338]
MongoDB	2019	275,000,000	tech	poor security	[246]
Wattpad	2020	270,000,000	web	hacked	[380]
Facebook	2019	267,000,000	social network	poor security	[148][149]
Microsoft	2019	250,000,000	tech	data exposed by misconfiguration	[238]
MongoDB	2019	202,000,000	tech	poor security	[245]
Unknown	2020	201,000,000	personal and demographic data about residents and their properties of US	Poor security	[161]
Instagram	2020	200,000,000	social network	poor security	[199]
Unknown agency (believed to be tied to United States Census Bureau)	2020	200,000,000	financial	accidentally published	[404]
Zynga	2019	173,000,000	social network	hacked	[402][403]
Equifax	2017	163,119,000	financial, credit reporting	poor security	[127][128]
Massive American business hack including 7-Eleven and Nasdaq	2012	160,000,000	financial	hacked	[234]
Adobe Systems Incorporated	2013	152,000,000	tech	hacked	[10]
Under Armour	2018	150,000,000	Consumer Goods	hacked	[354]
eBay	2014	145,000,000	web	hacked	[120]
Canva	2019	140,000,000	web	hacked	[67][68][69]
Heartland	2009	130,000,000	financial	hacked	[187][188]
Tetrad	2020	120,000,000	market analysis	poor security	[329]

Users



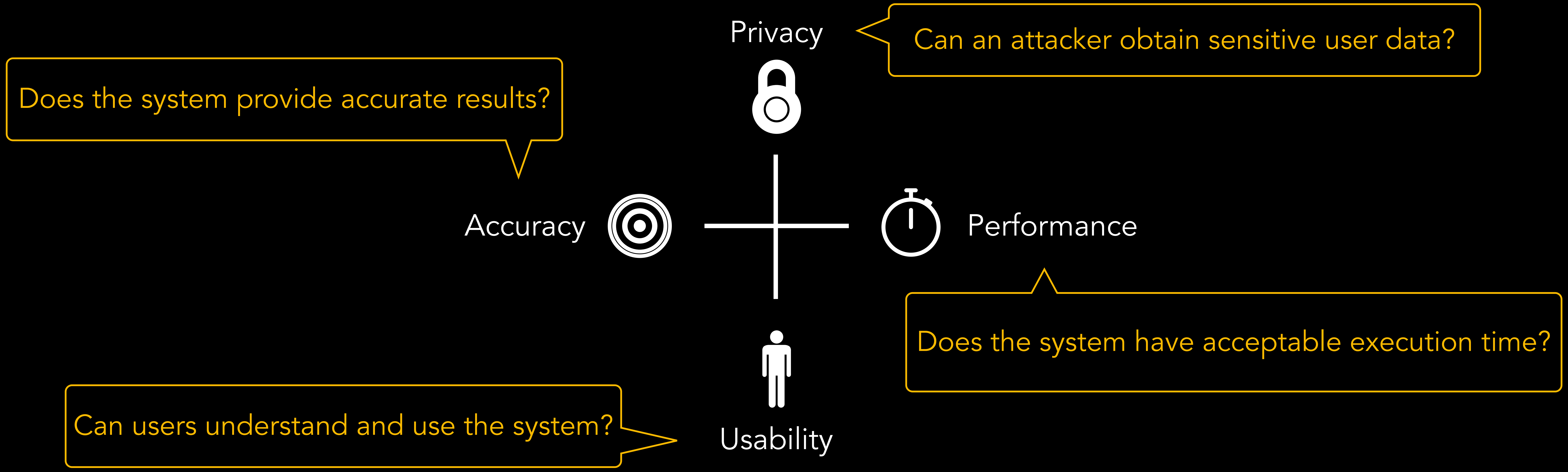
Clients

during computation

released results

Systems must ensure privacy while maintaining utility

System-Building Challenges



Selected Research

Ensure end-to-end protection of sensitive data

Minimize user intervention to simplify system usage

Private Data Federations

Efficient SQL Queries for Private Data Federations
- SMCQL (VLDB ’17)
- Shrinkwrap (VLDB ’18)

Privacy-Preserving Approximate Query Processing
- SAQE (VLDB ’19)

Optimize utility while preserving privacy

Privacy for Growing Data

Secure Growing Databases in the Untrusted Cloud
- DP-Sync (SIGMOD ’21)
- IncShrink (under revision @ SIGMOD ’22)

Countering Cache Side Channel Attacks in Web Browsers

Privacy in Real World Systems

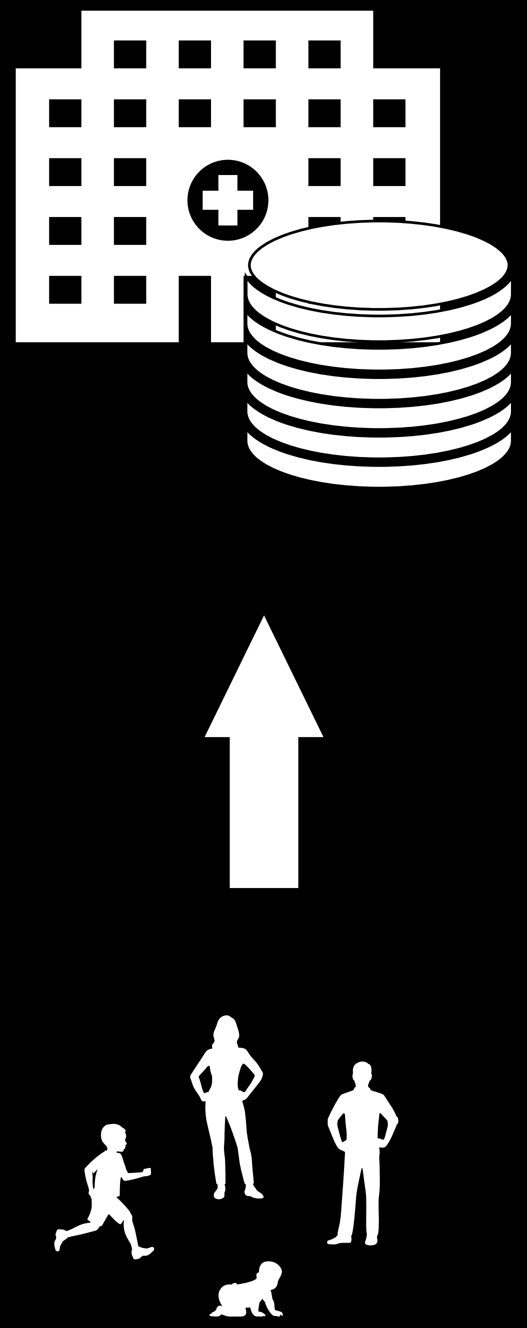
Visualizing Privacy-Utility Trade-offs in Differential Privacy
- ViP (PETS ’22)

Private Contact Summary Aggregation for Covid-19

Enable expert configuration by non-experts

Building a Private Data Federation

Example: Clinical Data



glucose	sex	diag	…..
120	M	blues	…..
80	F	cdiff	…..
100	M	X	…..

Example: Clinical Data

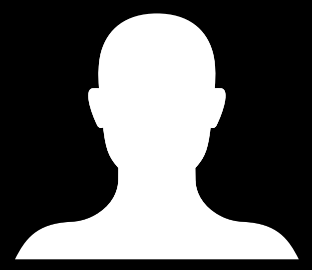
A Clinical Research Network (CRN) is a consortium of healthcare sites that agree to share their data for research.

For this project, we partnered with HealthLNK, a Chicago-based CRN, that wants to make their data available to researchers.

This project is part of a pilot study at three Chicago-area hospital networks used to identify patient populations that are potentially under-treated for hypertension.

Example: Clinical Data

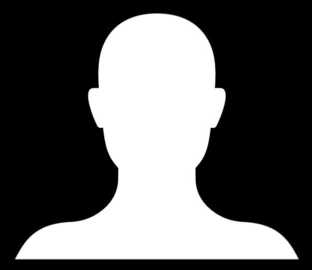
How many diagnoses
of rare disease X occurred?



Researcher

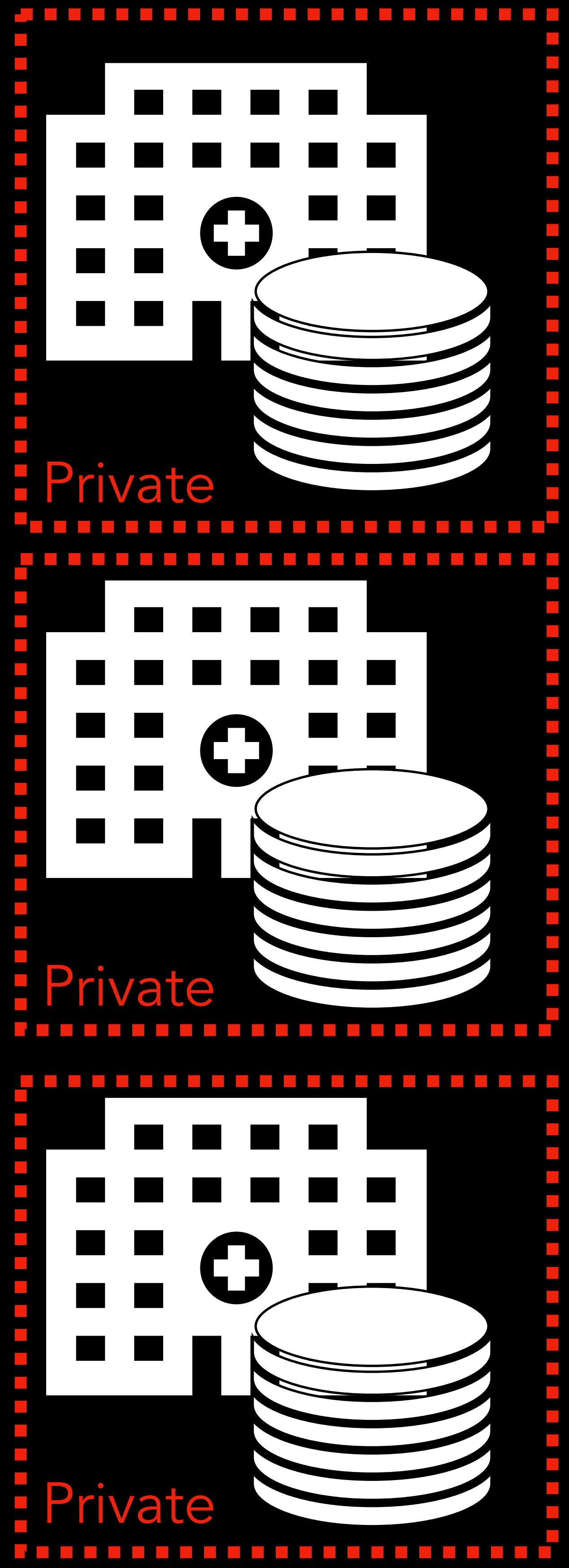
Example: Clinical Data

How many diagnoses
of rare disease X occurred?

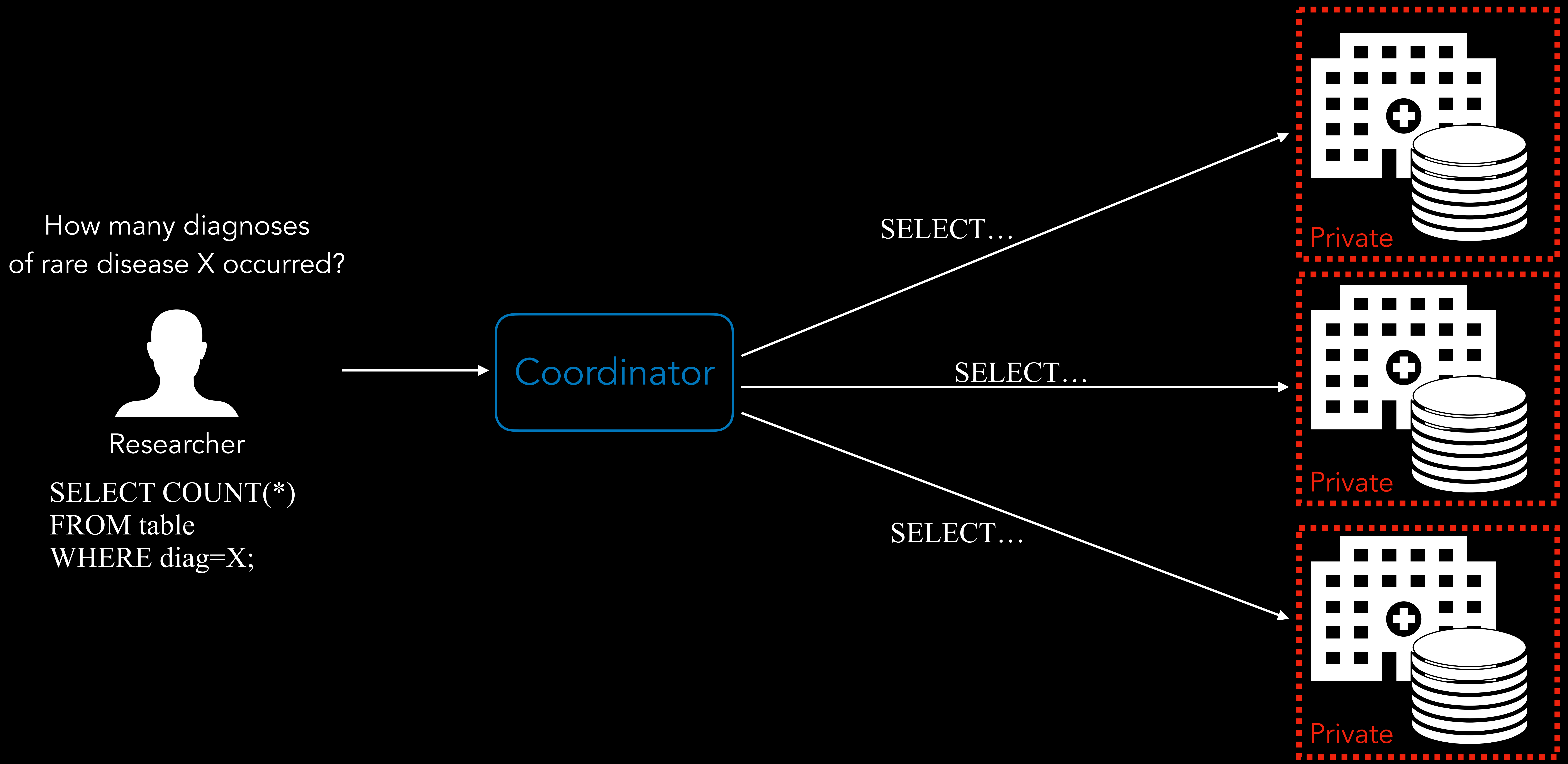


Researcher

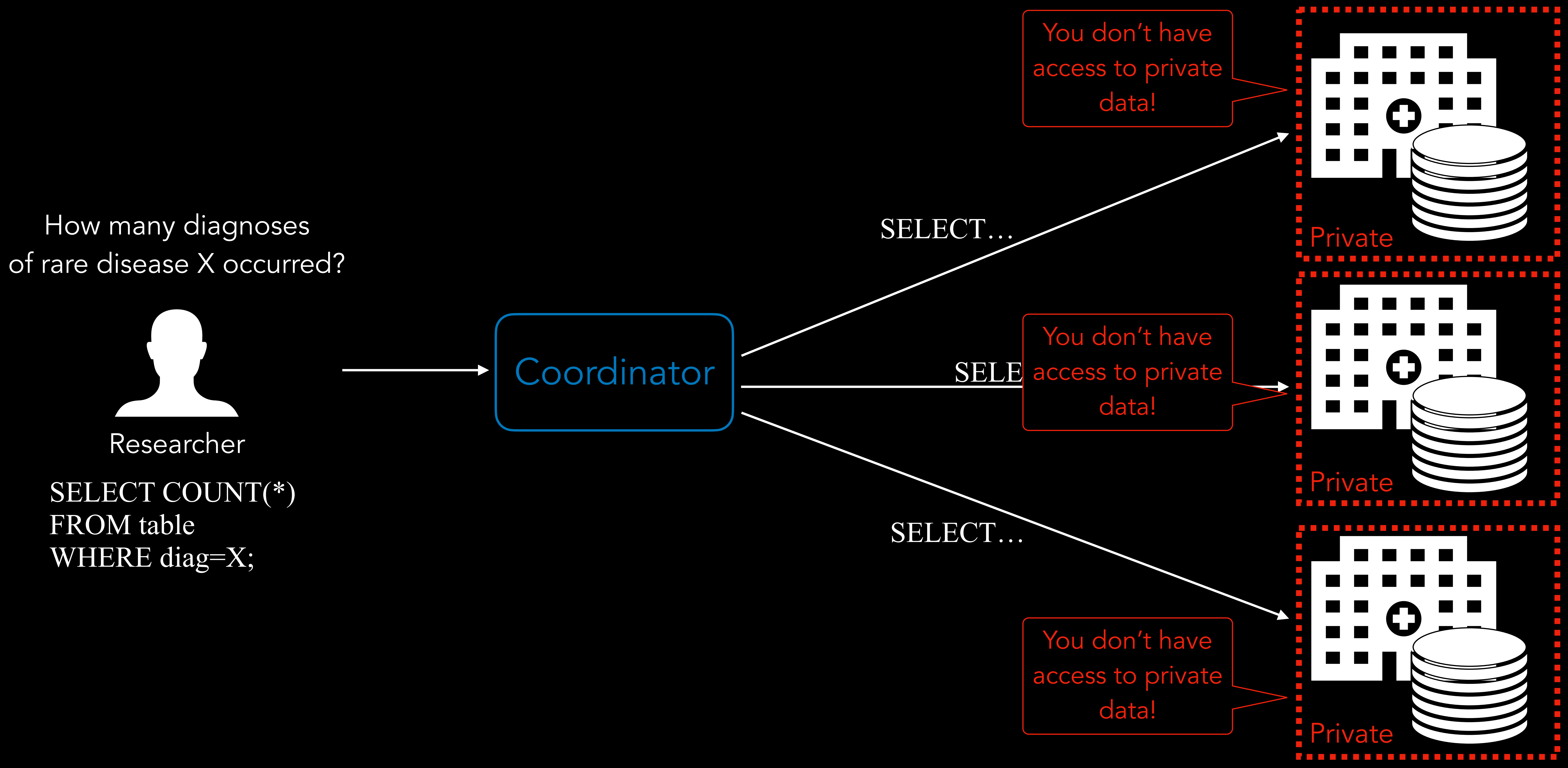
SELECT COUNT(*)
FROM table
WHERE diag=X;



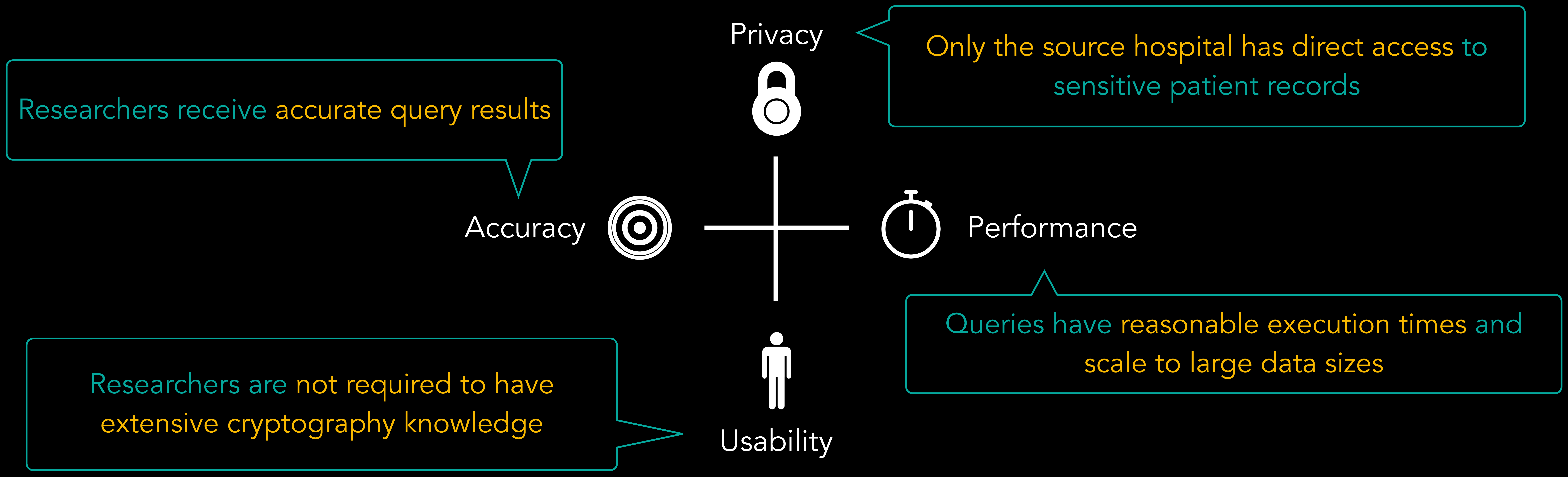
Example: Clinical Data



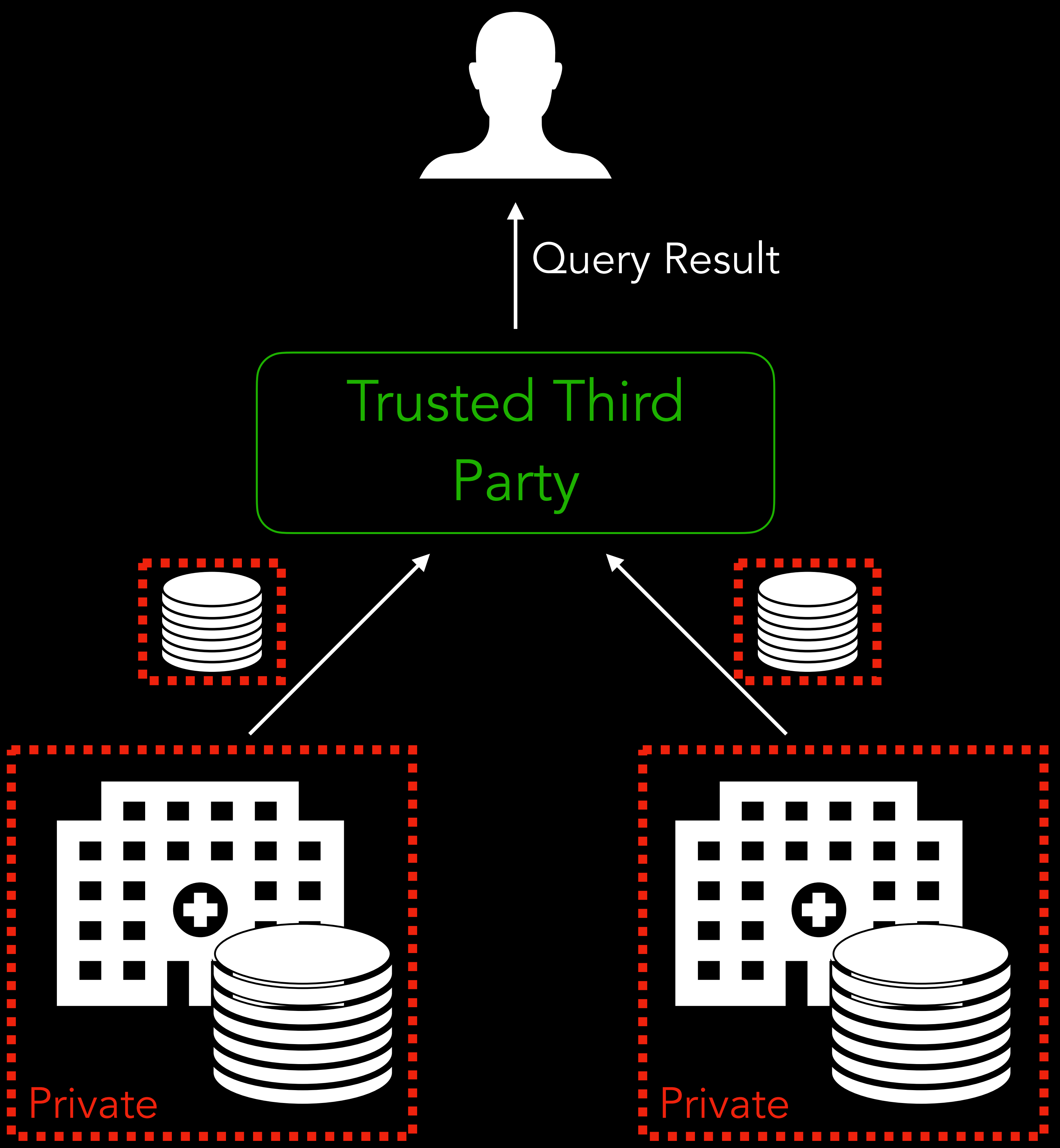
Example: Clinical Data



Private Data Federation Requirements



Potential Solution: Trusted Third Party



Trusted by All Parties

Allowed to see all records from all parties

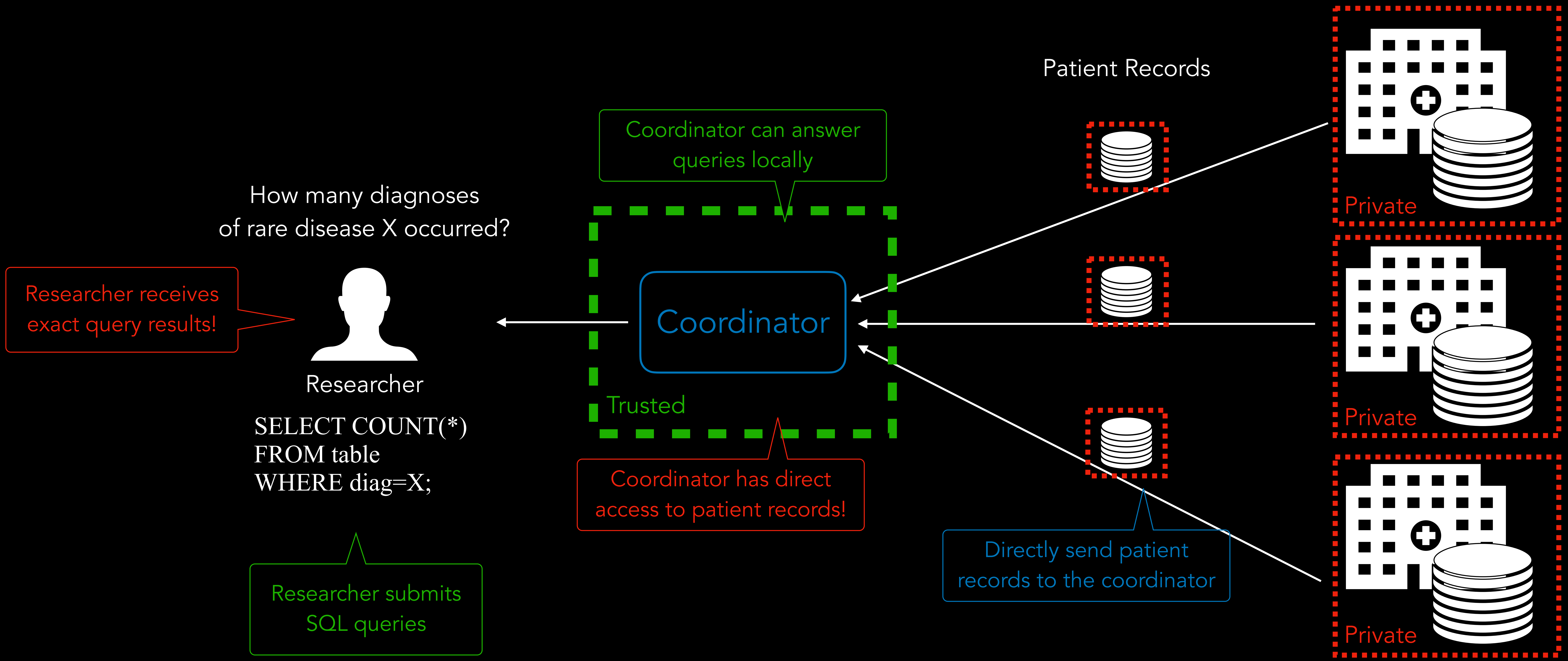
Local Storage

Collects and stores all records locally

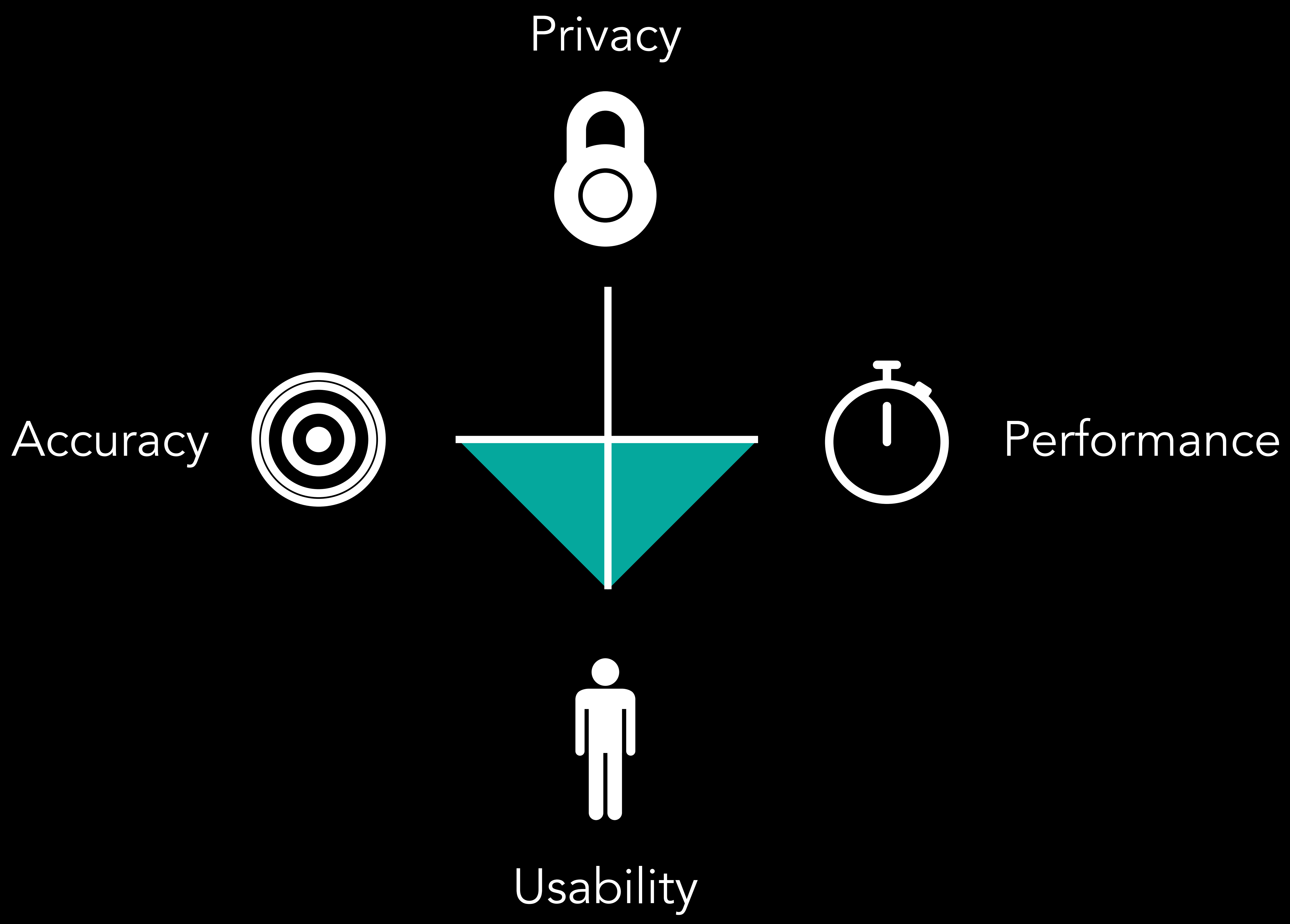
Local Computation

Executes all received queries without additional communication

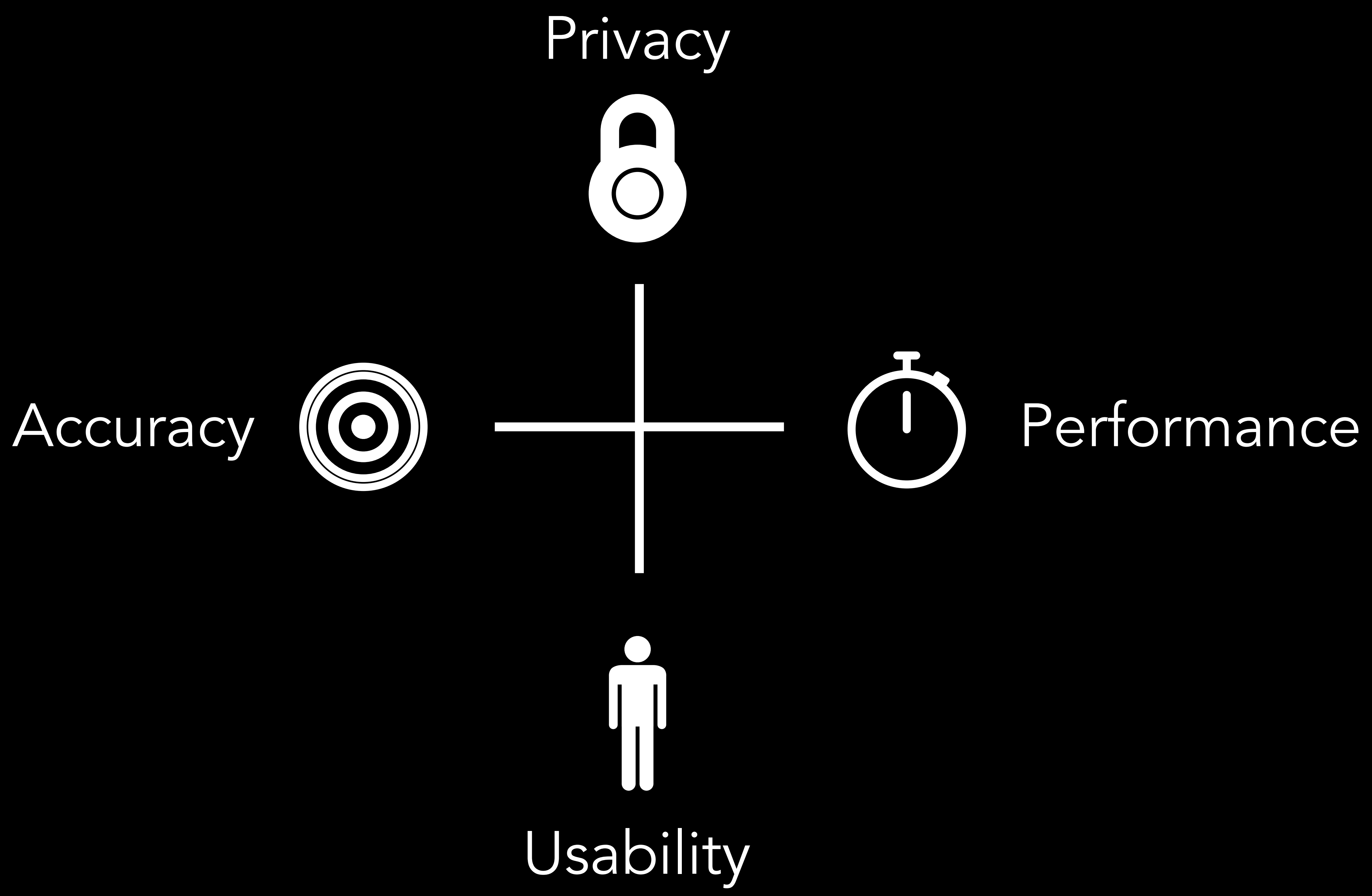
Potential Solution: Trusted Third Party



Potential Solution: Trusted Third Party



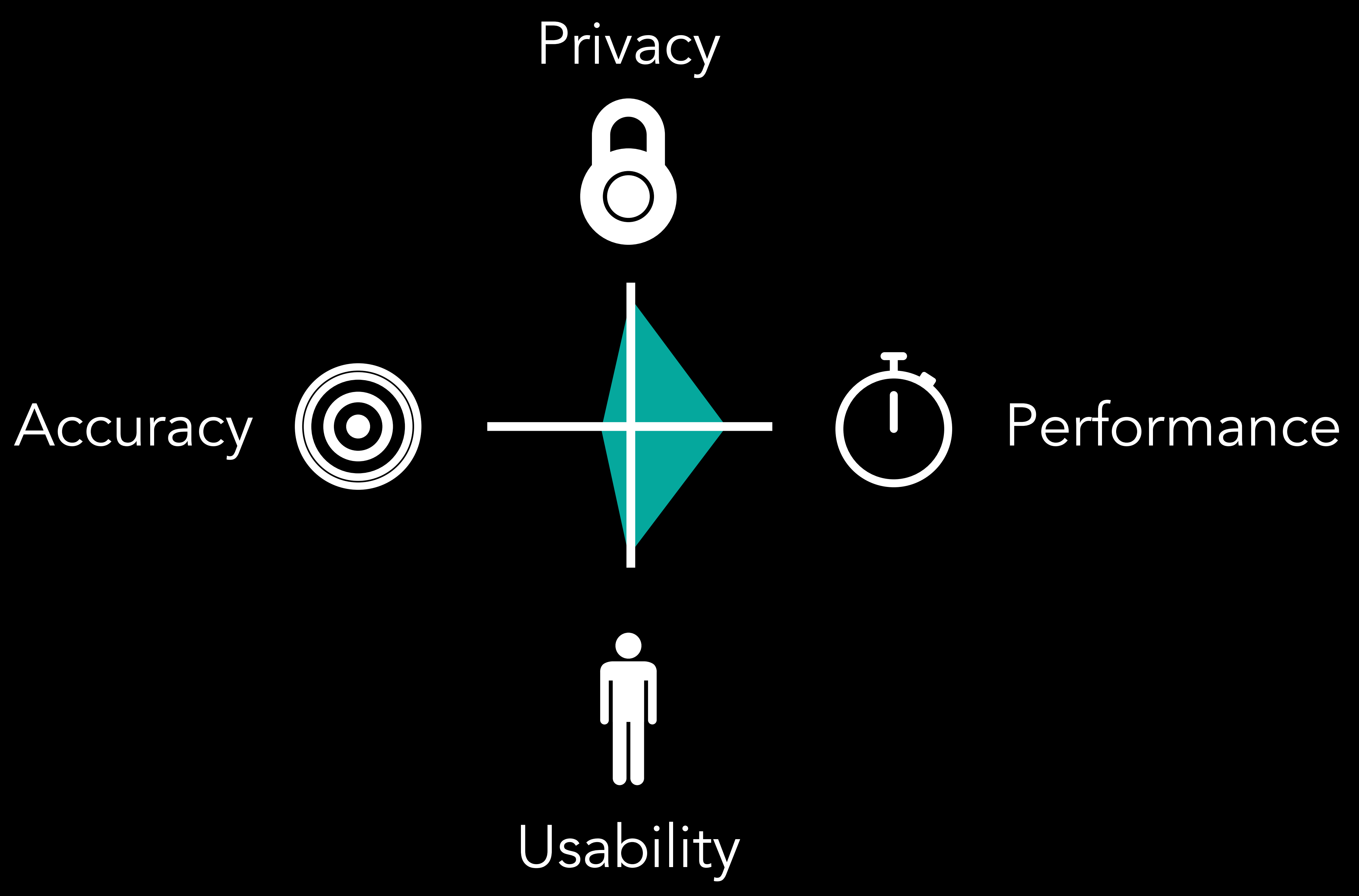
Building Blocks



Differential Privacy (DP)

Secure Multiparty Computation (MPC)

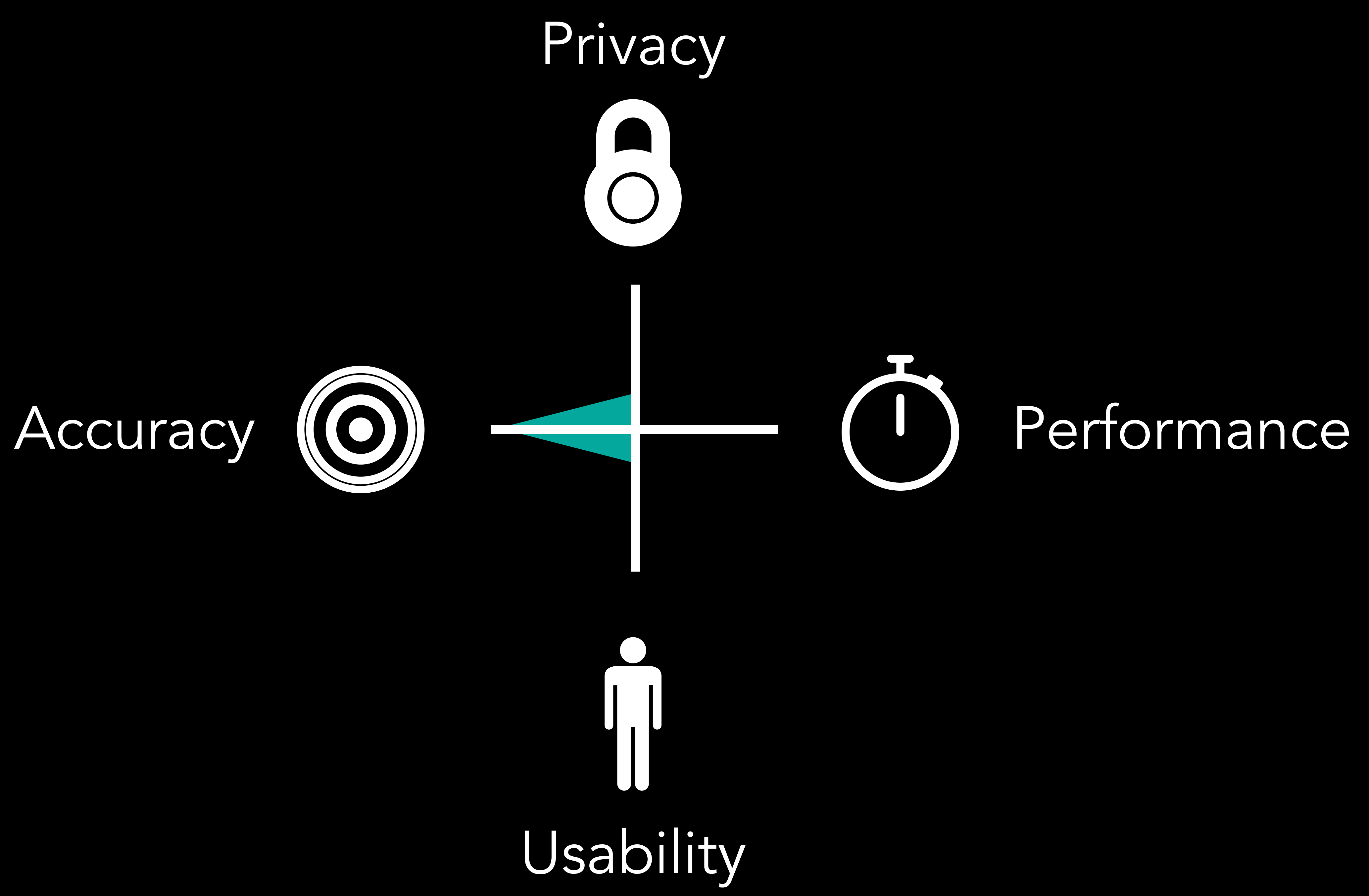
Building Blocks



Differential Privacy (DP)

Protect sensitive patient records by adding privacy-preserving noise

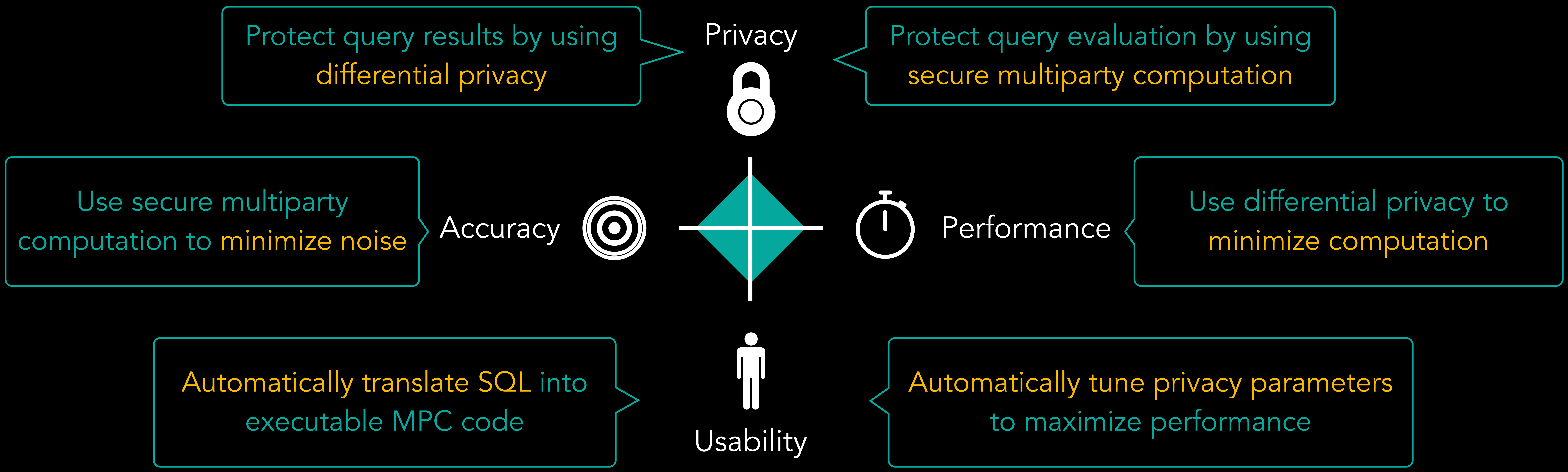
Building Blocks



Secure Multiparty Computation (MPC)

Protect sensitive patient records by using encrypted execution

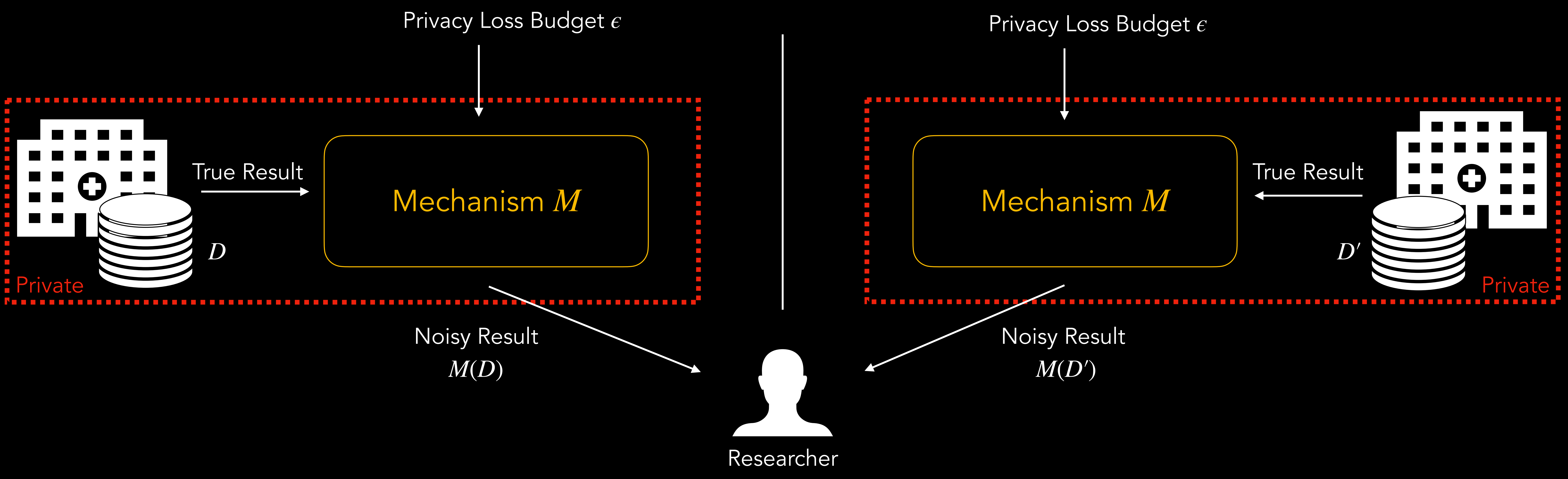
Private Data Federation



Private Data Federation

Differential Privacy

D: Patient A's health record is present | **D': Patient A's health record is not present**



M satisfies differential privacy if for any two neighboring databases D and D'

$$Pr[M(D) \in O] \leq e^{\epsilon} Pr[M(D') \in O],$$

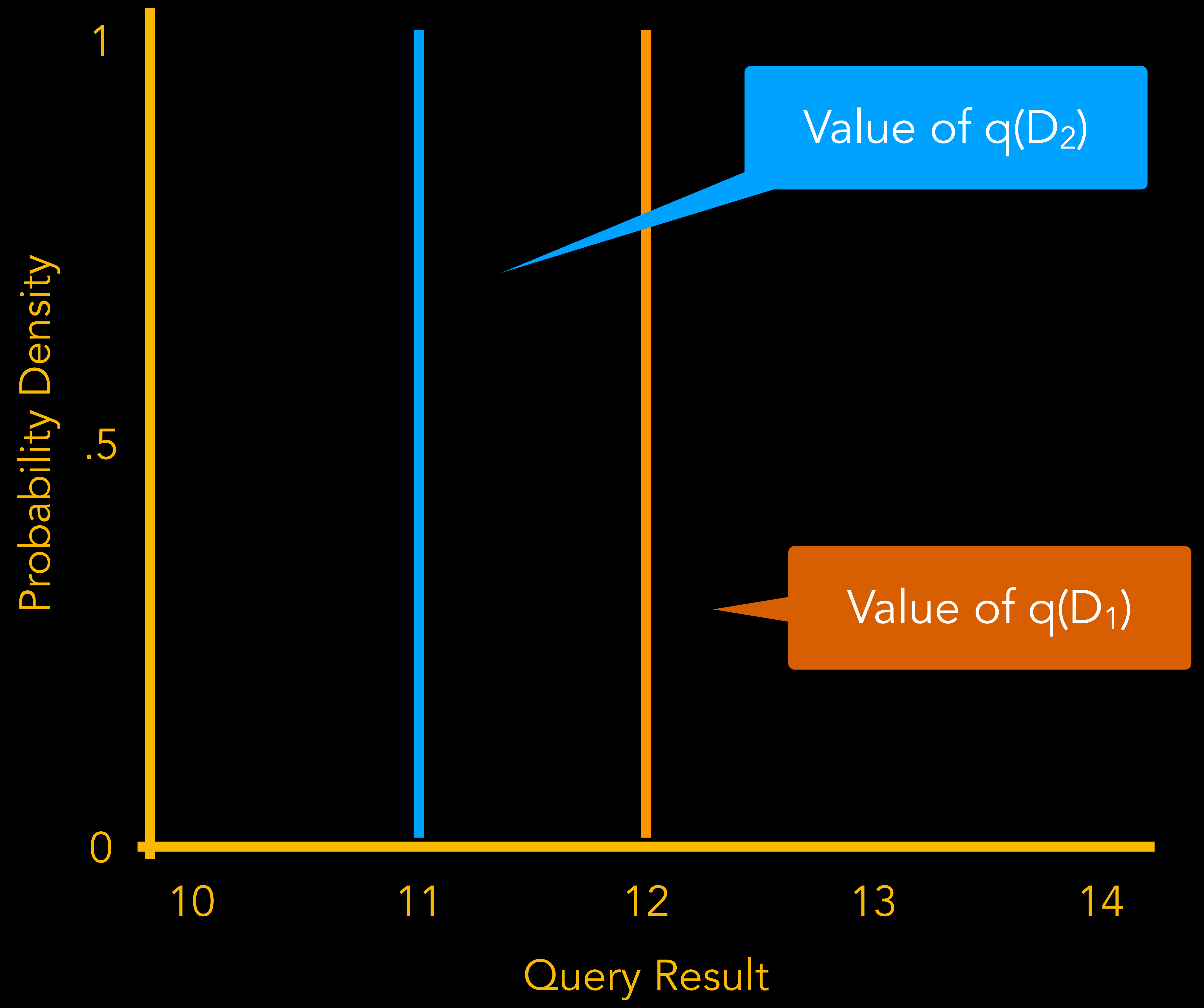
$O \subseteq \mathrm{O}$ where O is the universe of all possible results and ϵ is the privacy loss budget

Deterministic Mechanism

Assume there is a mechanism A takes in a query q and a database D, then returns the true result q(D).

Furthermore, there is a database D_1 contains Alice's sensitive information and a database D_2 that does not.

If the true result is 12 with Alice and 11 without Alice, the plot will look like the figure to the right.

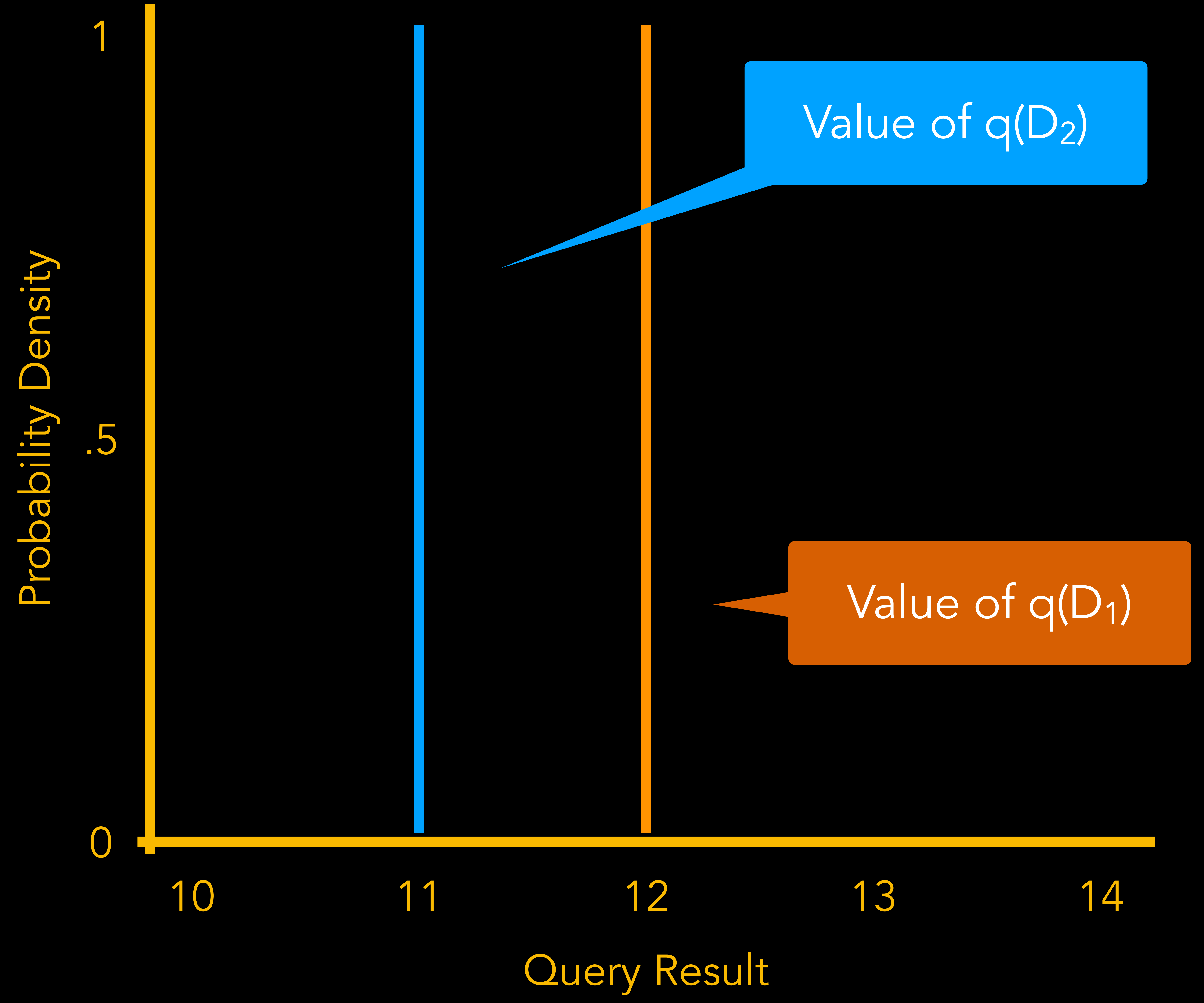


Deterministic Mechanism

Question: Does the mechanism satisfy differential privacy?

No, because Alice's presence or absence can be deduced with 100% accuracy. An analyst with enough background knowledge could deduce Alice's sensitive information.

$$Pr[A(D) = 12] > e^{\epsilon} Pr[A(D') = 12]$$



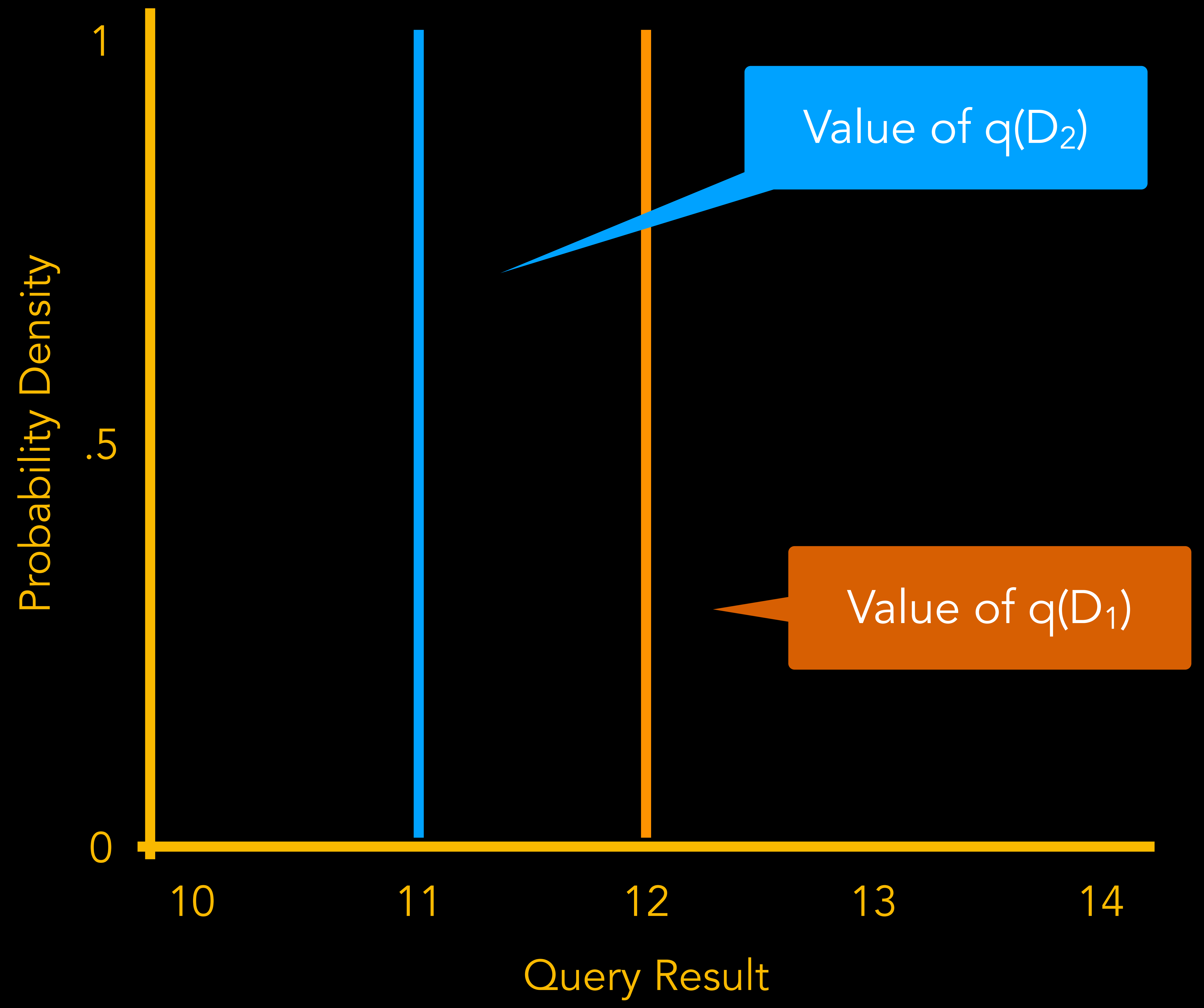
Deterministic Mechanism

Is this privacy-preserving?

No, because Alice's presence or absence can potentially be deduced with 100% accuracy.

Is this useful?

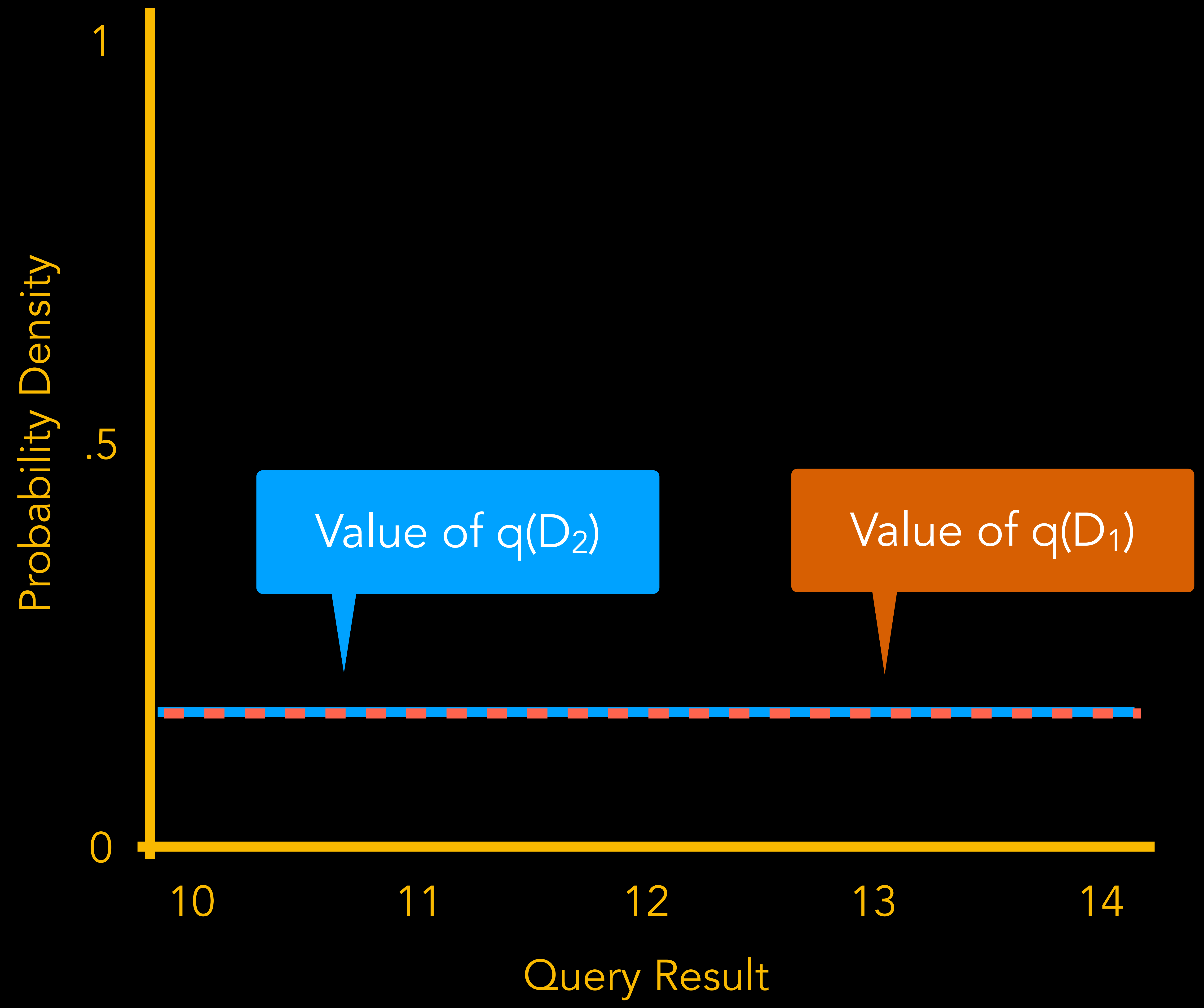
Yes, because the true result of the query is always returned.



Uniform Mechanism

Now assume that mechanism A takes in a query q and a database D, then returns a value drawn from a uniform distribution centered on the true value.

If the true result is 12 with Alice and 11 without Alice, the plot will look like the figure to the right.

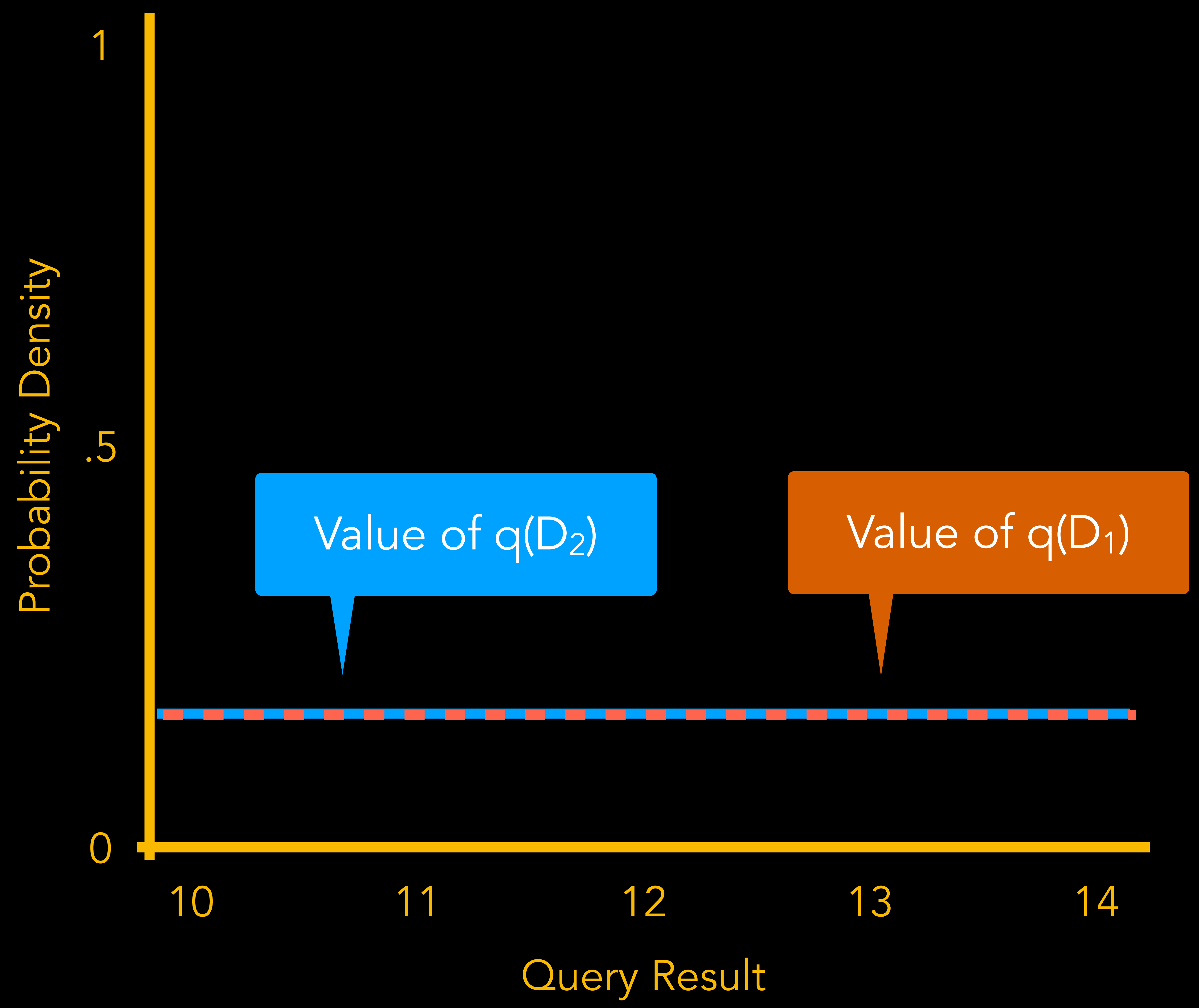


Uniform Mechanism

Question: Does the mechanism satisfy differential privacy?

Yes, because Alice's presence or absence cannot be deduced with 100% accuracy even by an analyst that knew all other records except Alice's information.

$$Pr[A(D) = o] = Pr[A(D') = o]$$



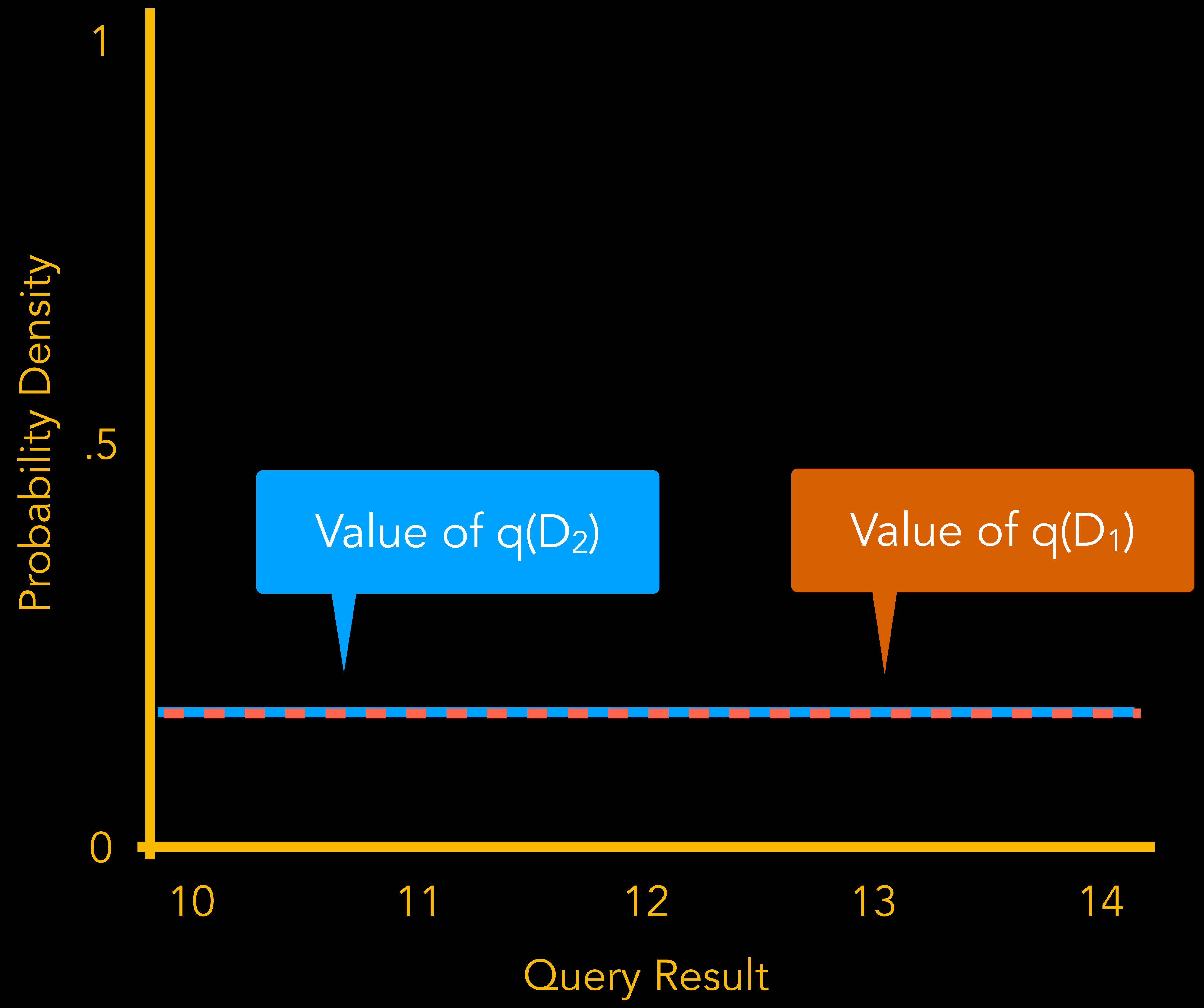
Uniform Mechanism

Is this privacy-preserving?

Yes, because no information is leaked about Alice

Is this useful?

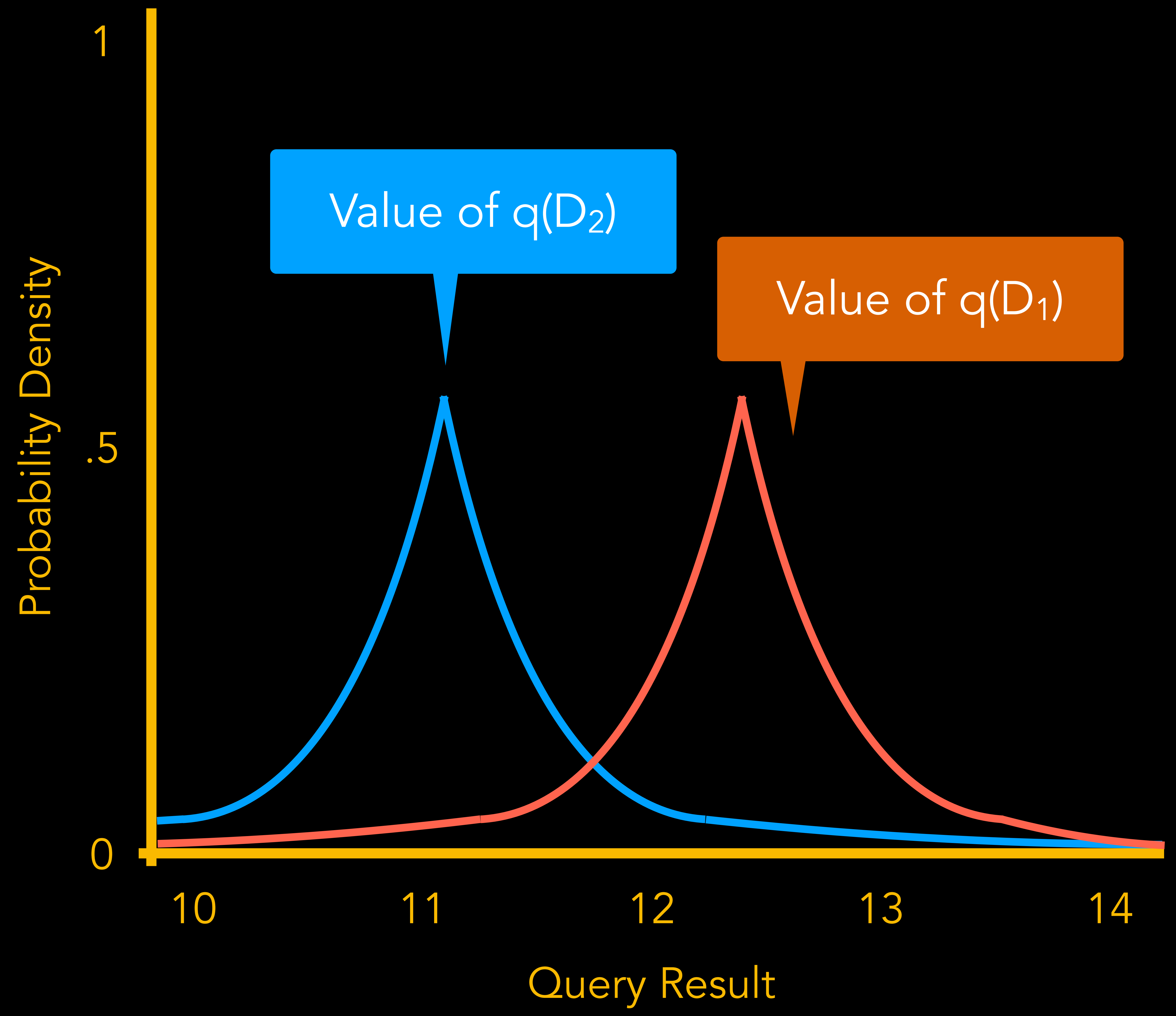
No, because the query result is not tied to the database contents



Randomized (or Noisy) Mechanism

Now assume that mechanism A takes in a query q and a database D, then returns a value drawn from a Laplace distribution centered on the true value.

If the true result is 12 with Alice and 11 without Alice, the plot will look like the figure to the right.

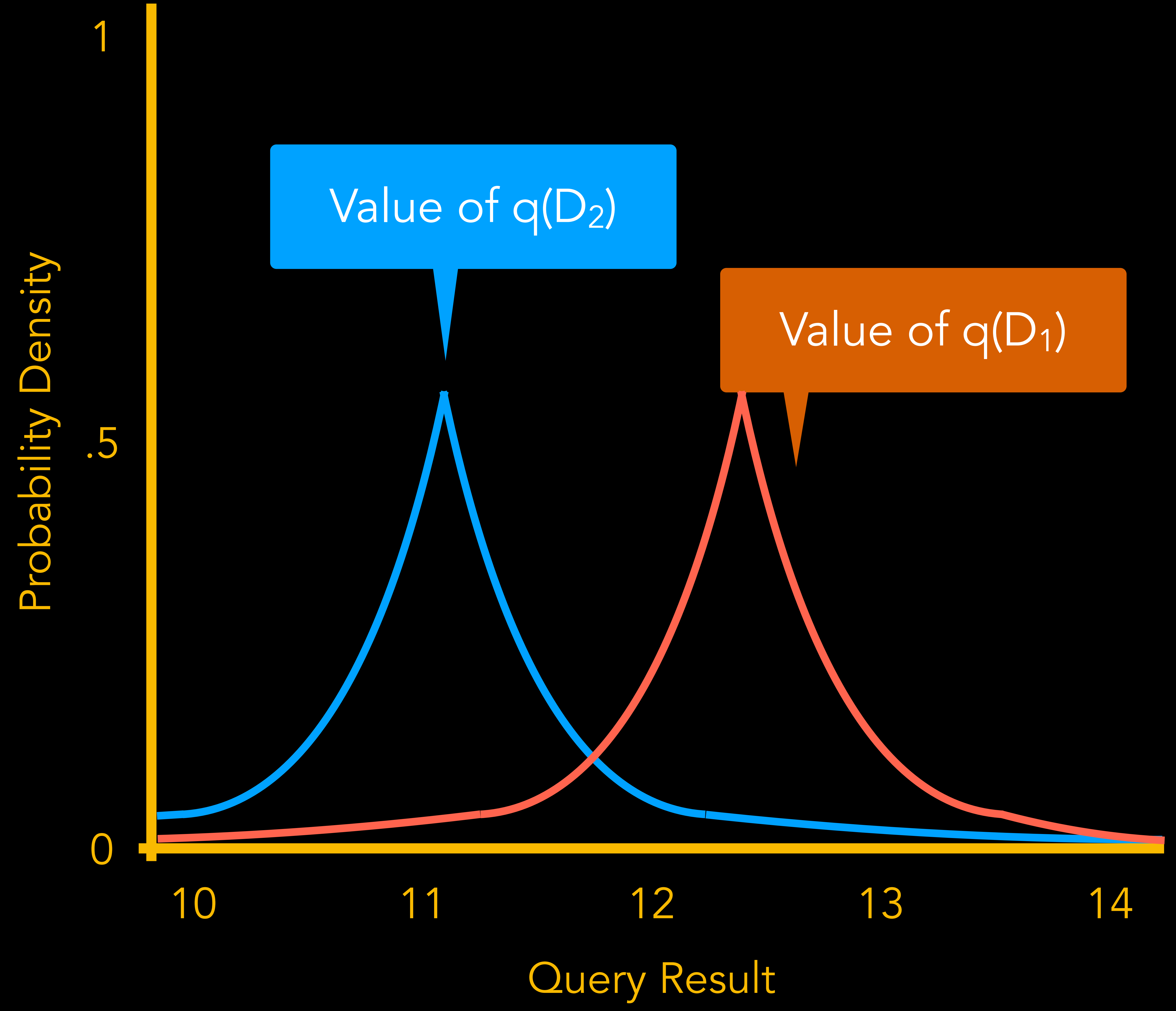


Randomized (or Noisy) Mechanism

Question: Does the mechanism satisfy differential privacy?

Yes, because Alice's presence or absence cannot be deduced with 100% accuracy even by an analyst that knew all other records except Alice's information.

$$Pr[A(D) = o] \leq e^{\epsilon} Pr[A(D') = o]$$



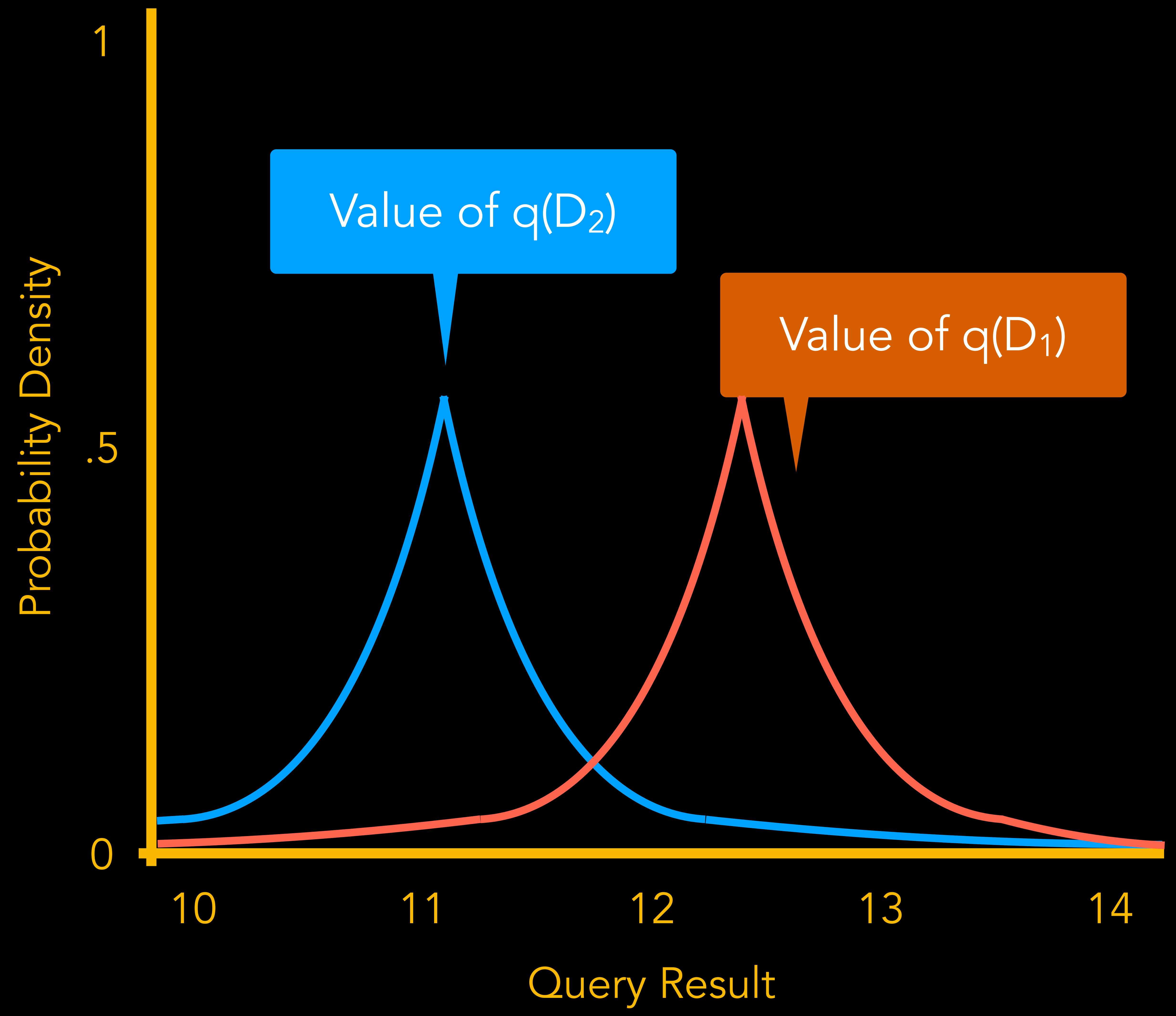
Randomized (or Noisy) Mechanism

Is this privacy-preserving?

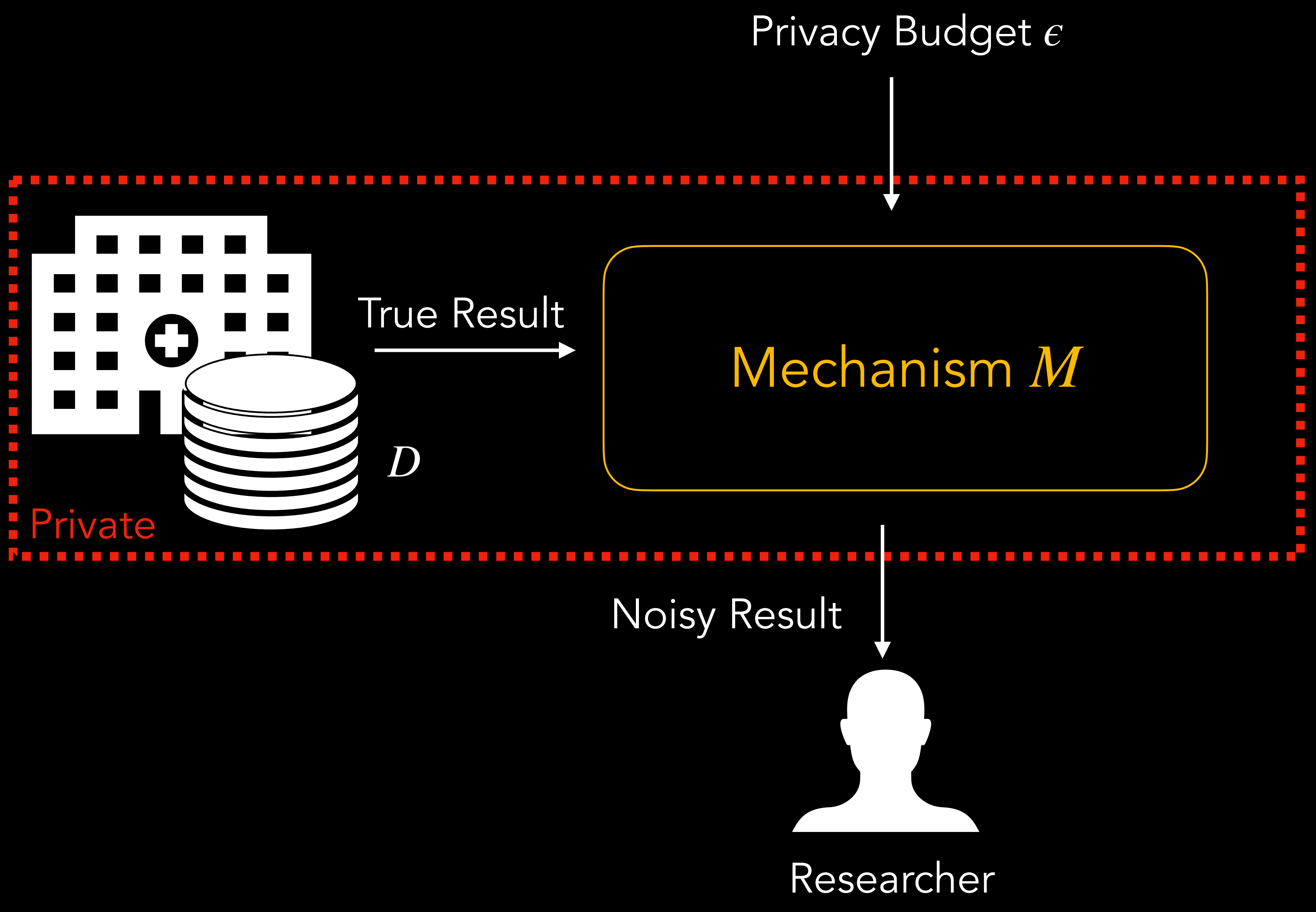
Yes, but only if not a large number of queries are evaluated

Is this useful?

Yes, because the query result is tied to the database contents



Differential Privacy



Accuracy-Privacy Trade-off

Adds noise to query results to hide contributions of individual users

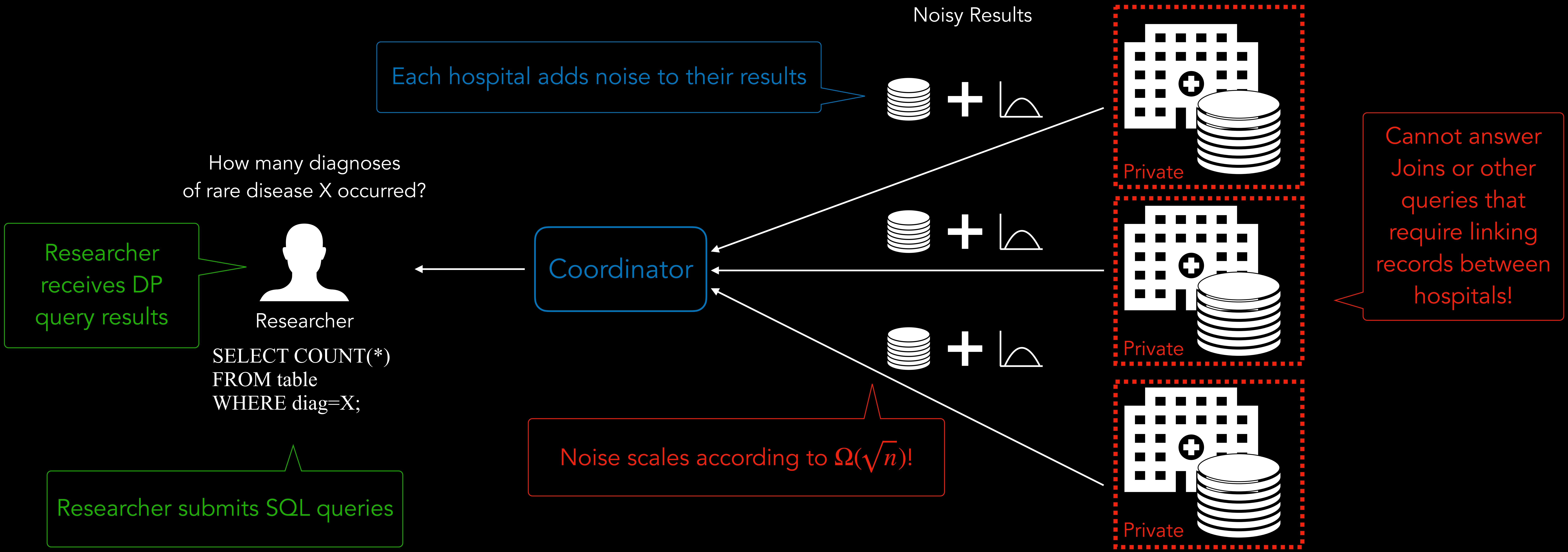
Quantifies Information Leakage

Bounds cumulative privacy loss according to a privacy loss budget

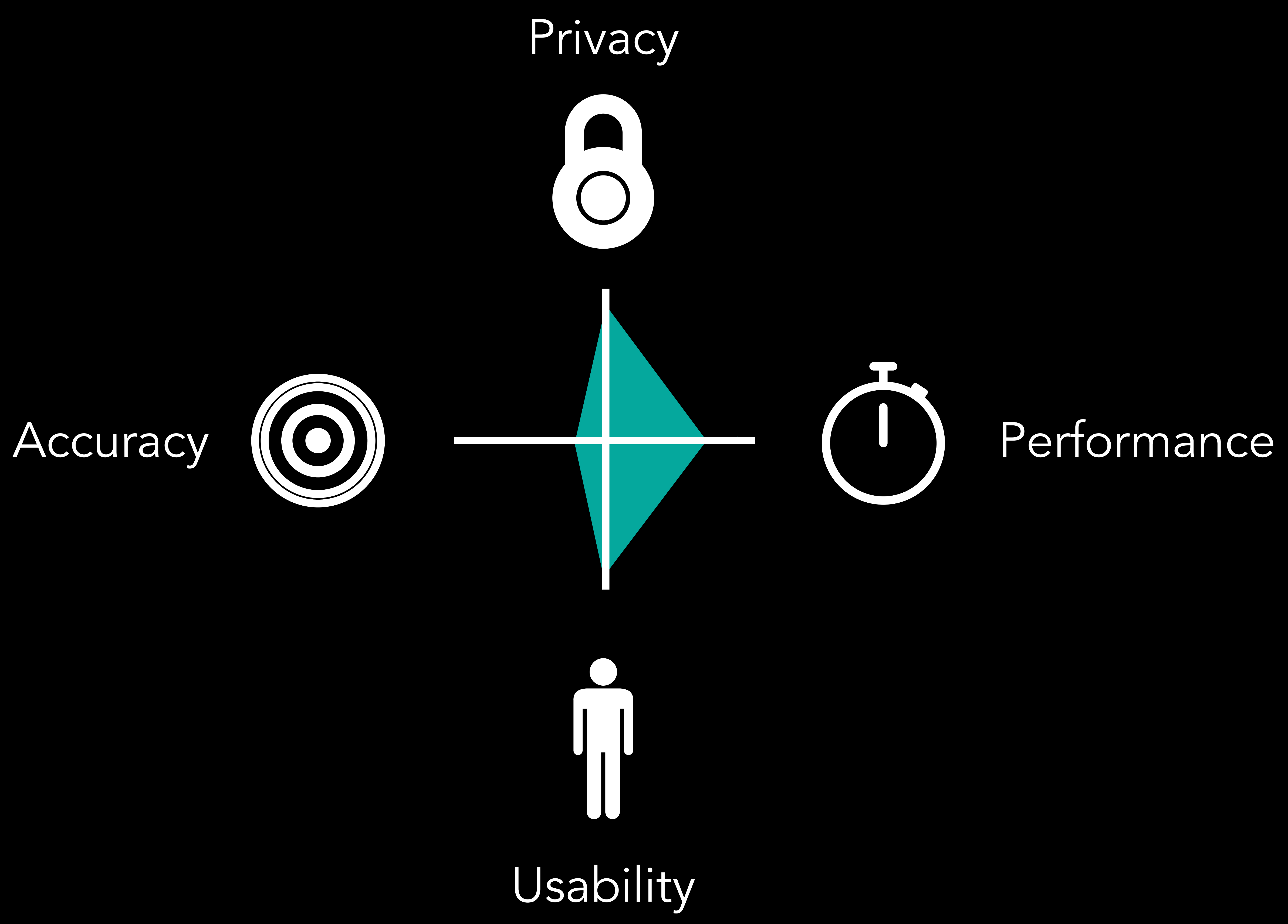
Utilized in Existing Applications

Used by organizations such as US Census, Apple, Google, etc.

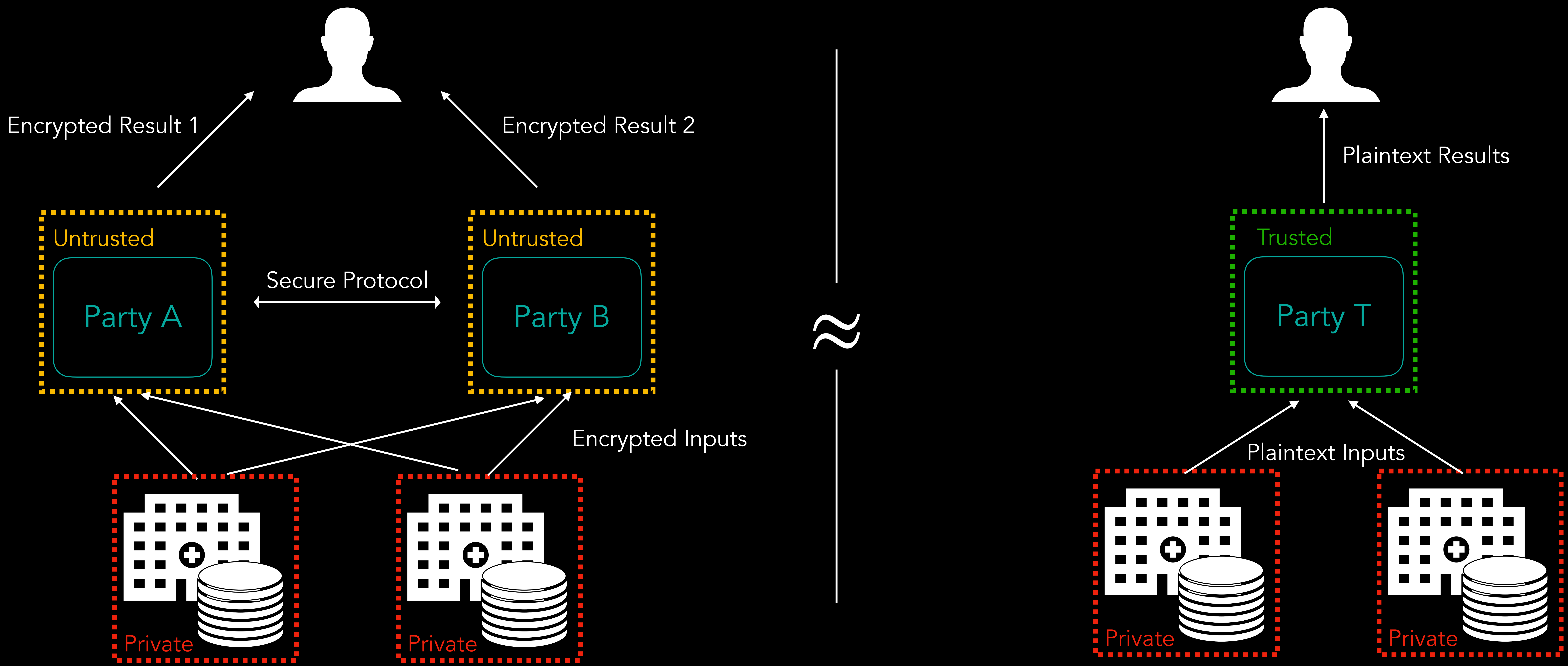
Differential Privacy



Differential Privacy



Secure Multiparty Computation



* Assumes non-collusion between parties A and B

Secure Multi-party Computation

Does Alice have more money than Bob?

$f(x, y)$

- Can see own data: x
- Can see result: $f(x, y)$
- Cannot see other user's data: y

- Can see own data: y
- Can see result: $f(x, y)$
- Cannot see other user's data: x

Secure Multi-party Computation (MPC)

Trustworthy Charlie

How trustworthy is Charlie?

x = $100

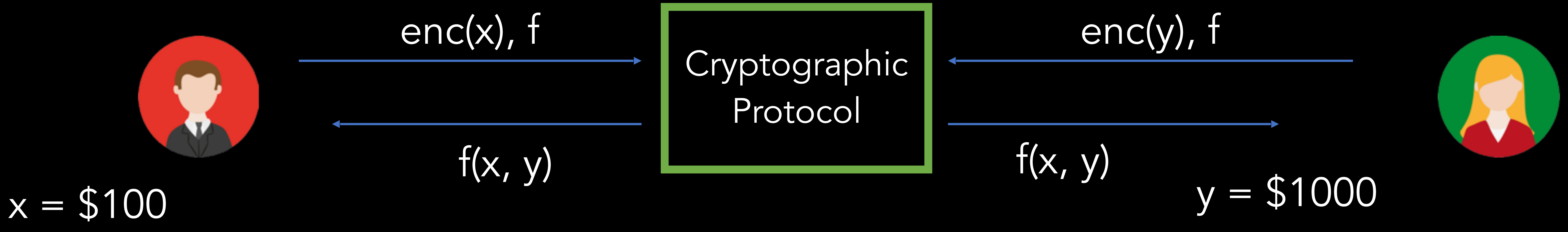
- Can see own data: x
- Can see result: f(x, y)
- Cannot see other user data: y
- Honestly reports x

y = $1000

- Can see own data: y
- Can see result: f(x, y)
- Cannot see other user data: x
- Honestly reports y

f(x,y) = Is y > x?

Secure Multi-party Computation (MPC)

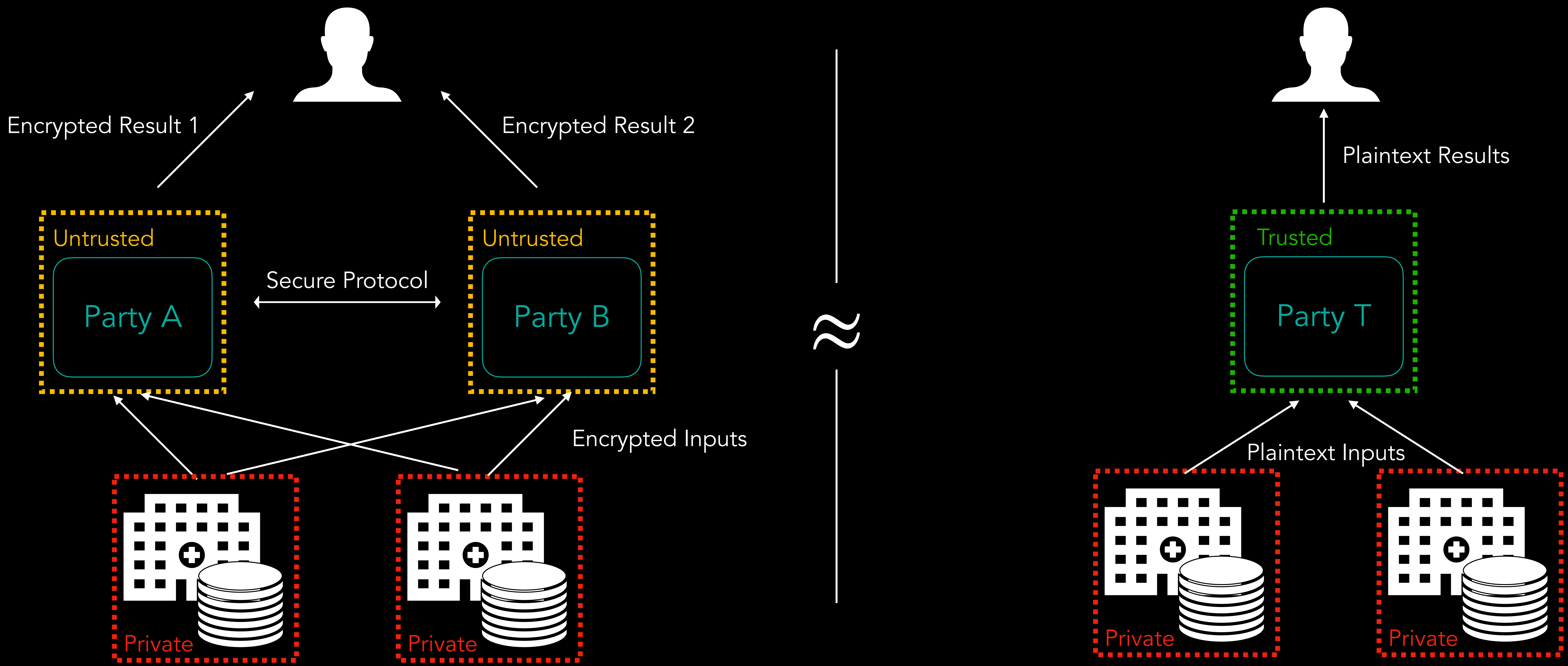


- Can see own data: x
- Can see result: f(x, y)
- Cannot see other user data: y
- Honestly follows protocol

- Can see own data: y
- Can see result: f(x, y)
- Cannot see other user data: x
- Honestly follows protocol

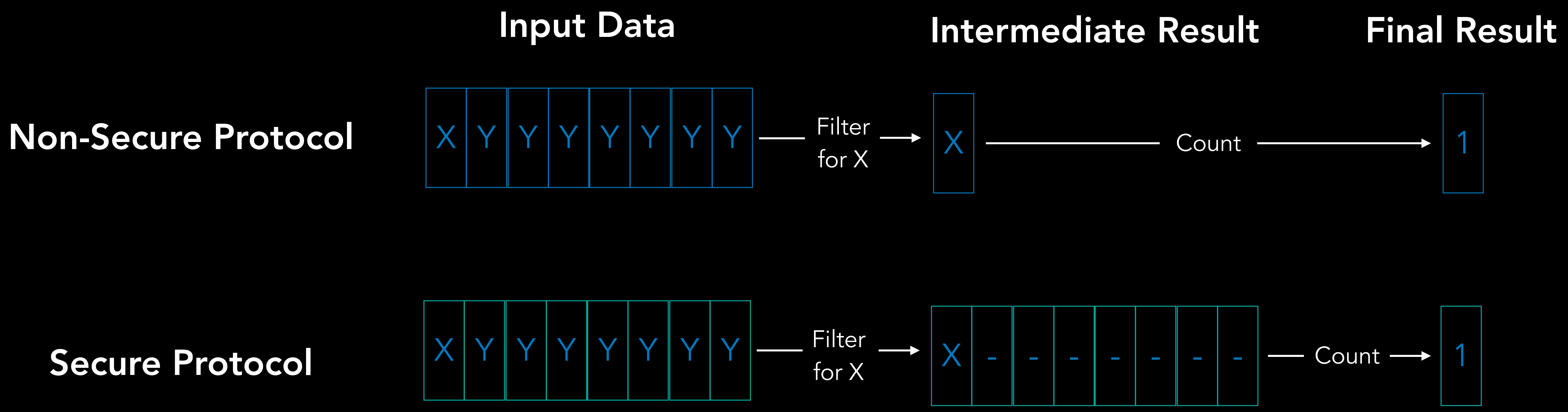
enc(x) = “encrypted” version of x

Secure Multiparty Computation



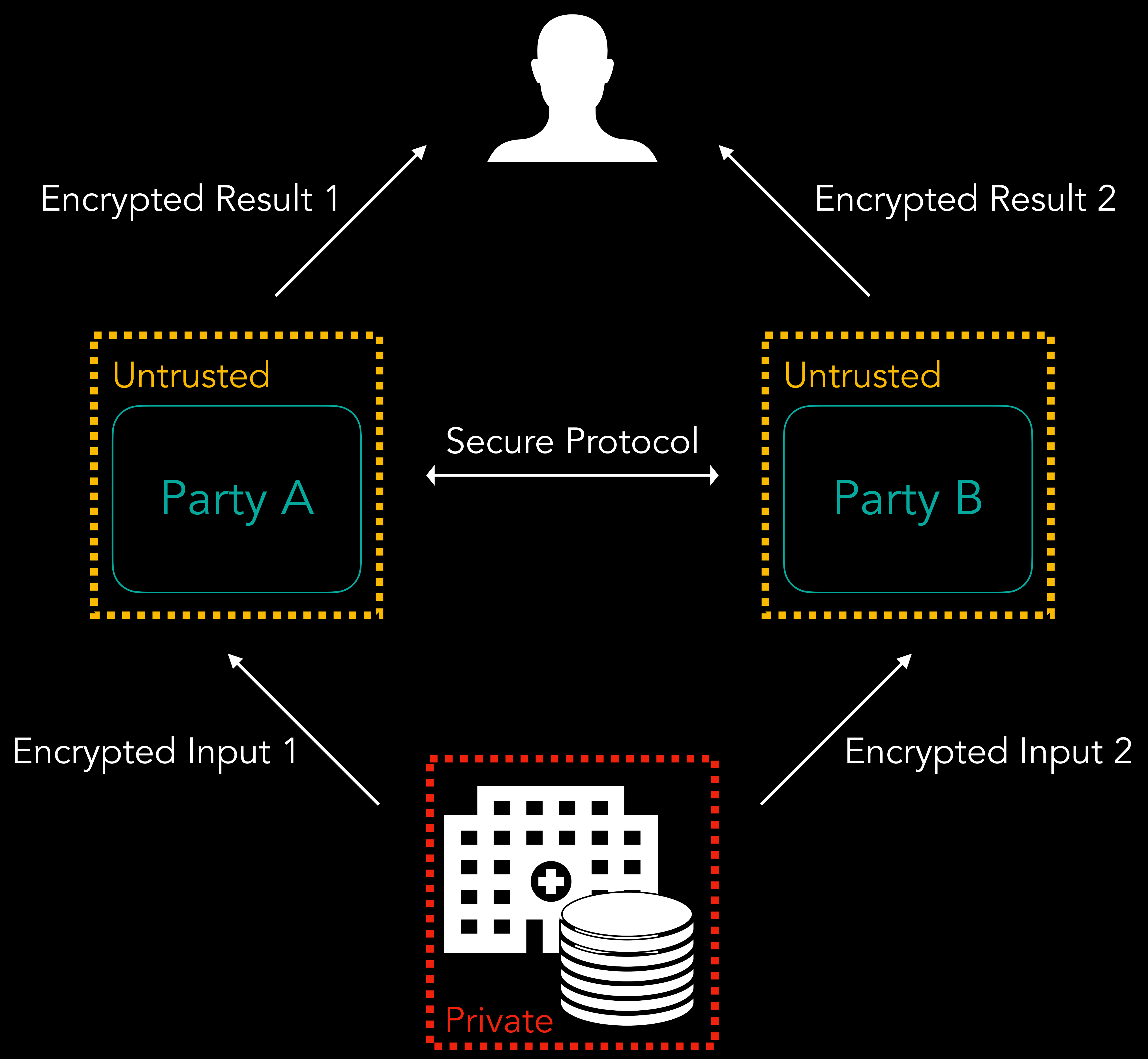
* Assumes non-collusion between parties A and B

Oblivious Execution



Secure Multiparty Computation requires worst-case execution to protect data during execution

Secure Multiparty Computation



Privacy-Performance Trade-off

Requires worst-case query execution during computation

End-to-End Encryption

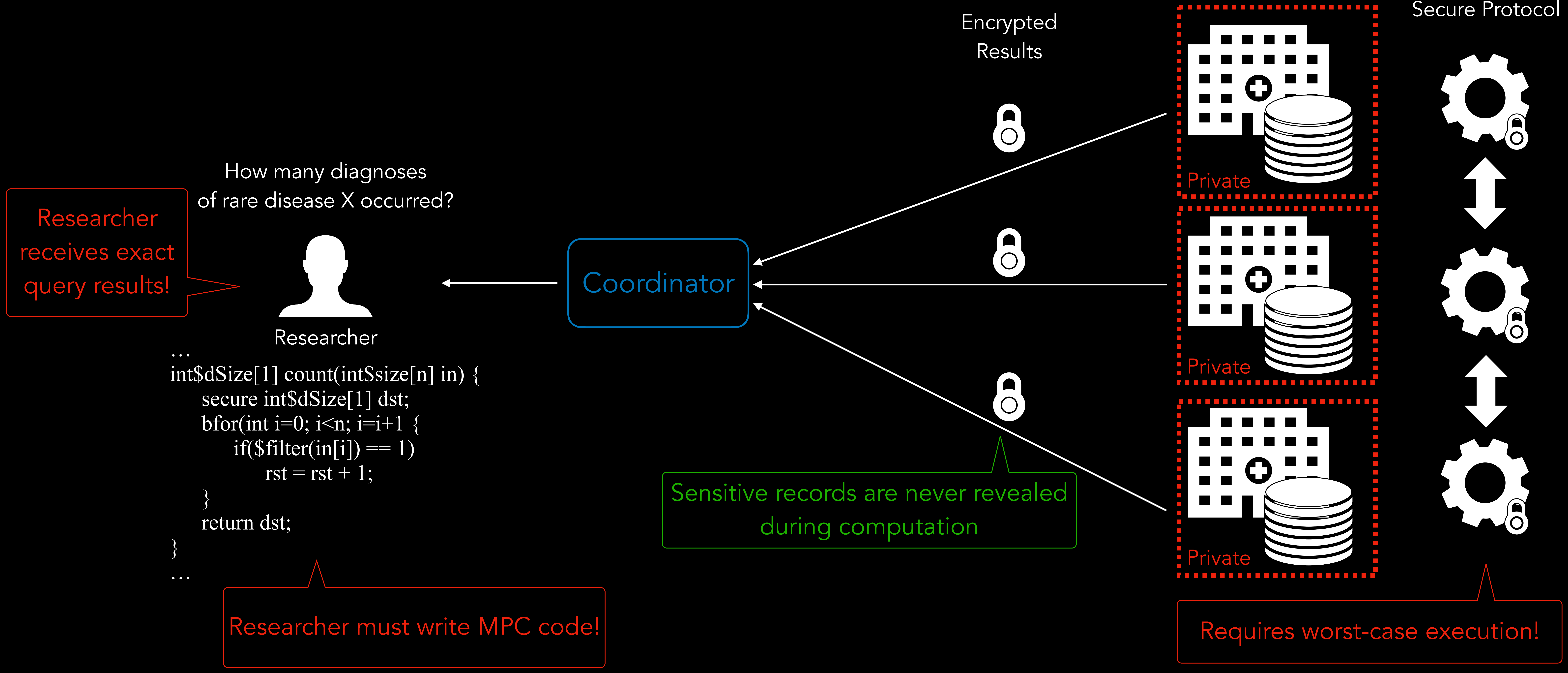
Computing parties evaluate queries without seeing records in plaintext

Exact Query Results

Final recipient reconstructs exact answer using encrypted results

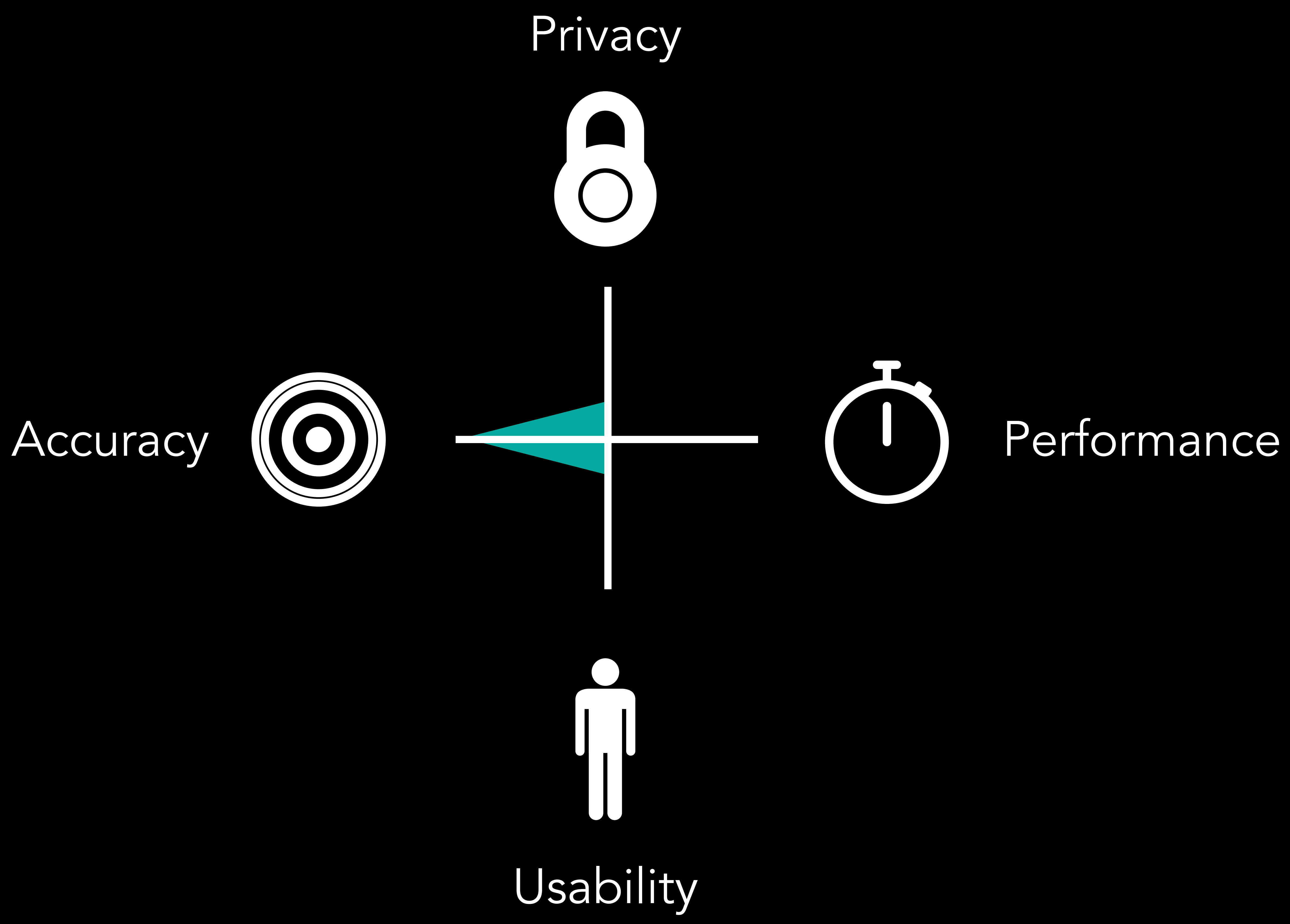
* Assumes non-collusion between parties A and B

Secure Multiparty Computation



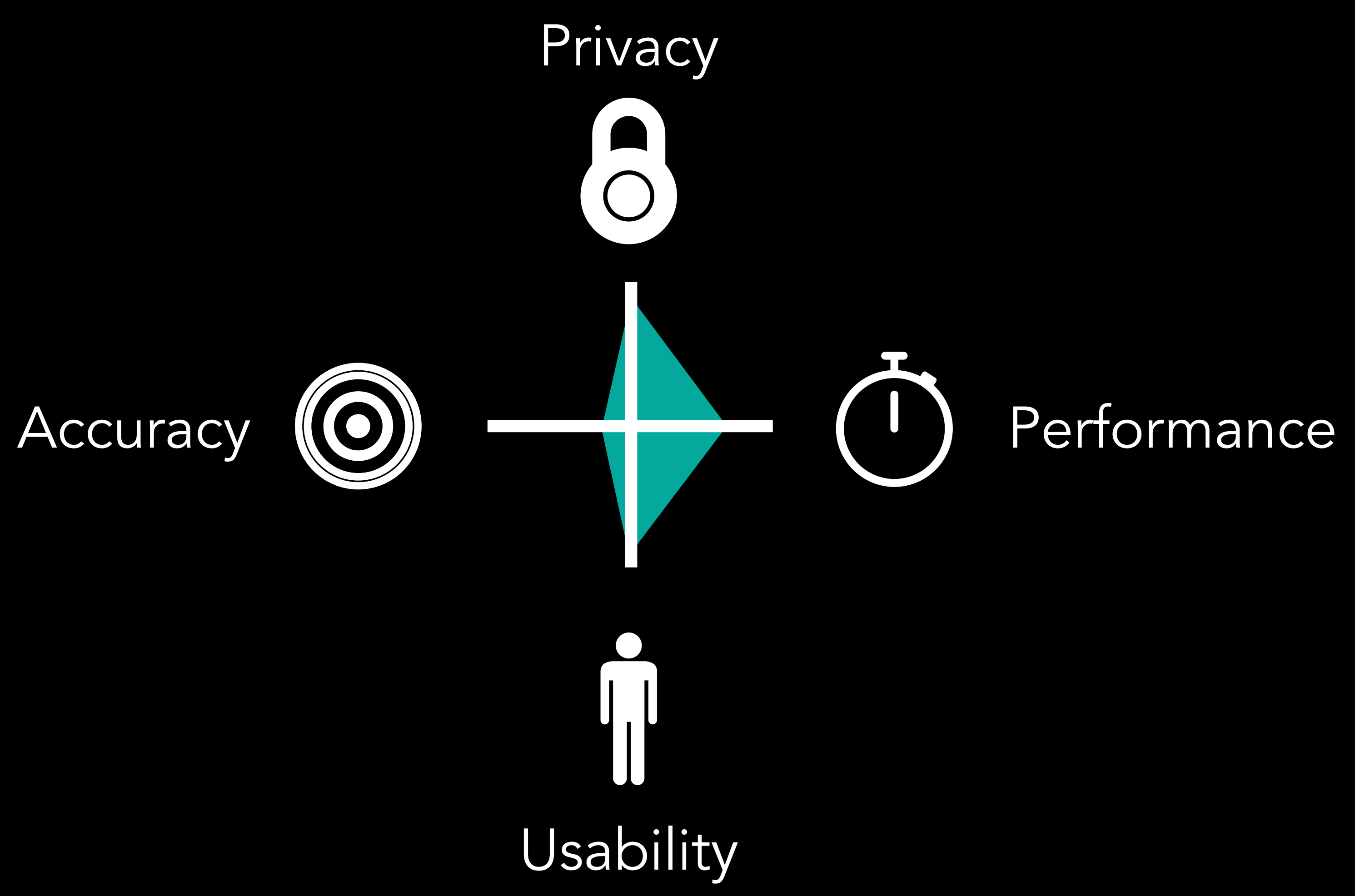

```
…
int$dSize[1] count(int$size[n] in) {
    secure int$dSize[1] dst;
    bfor(int i=0; i<n; i=i+1 {
        if($filter(in[i]) == 1)
            rst = rst + 1;
    }
    return dst;
}
…
```

Secure Multiparty Computation

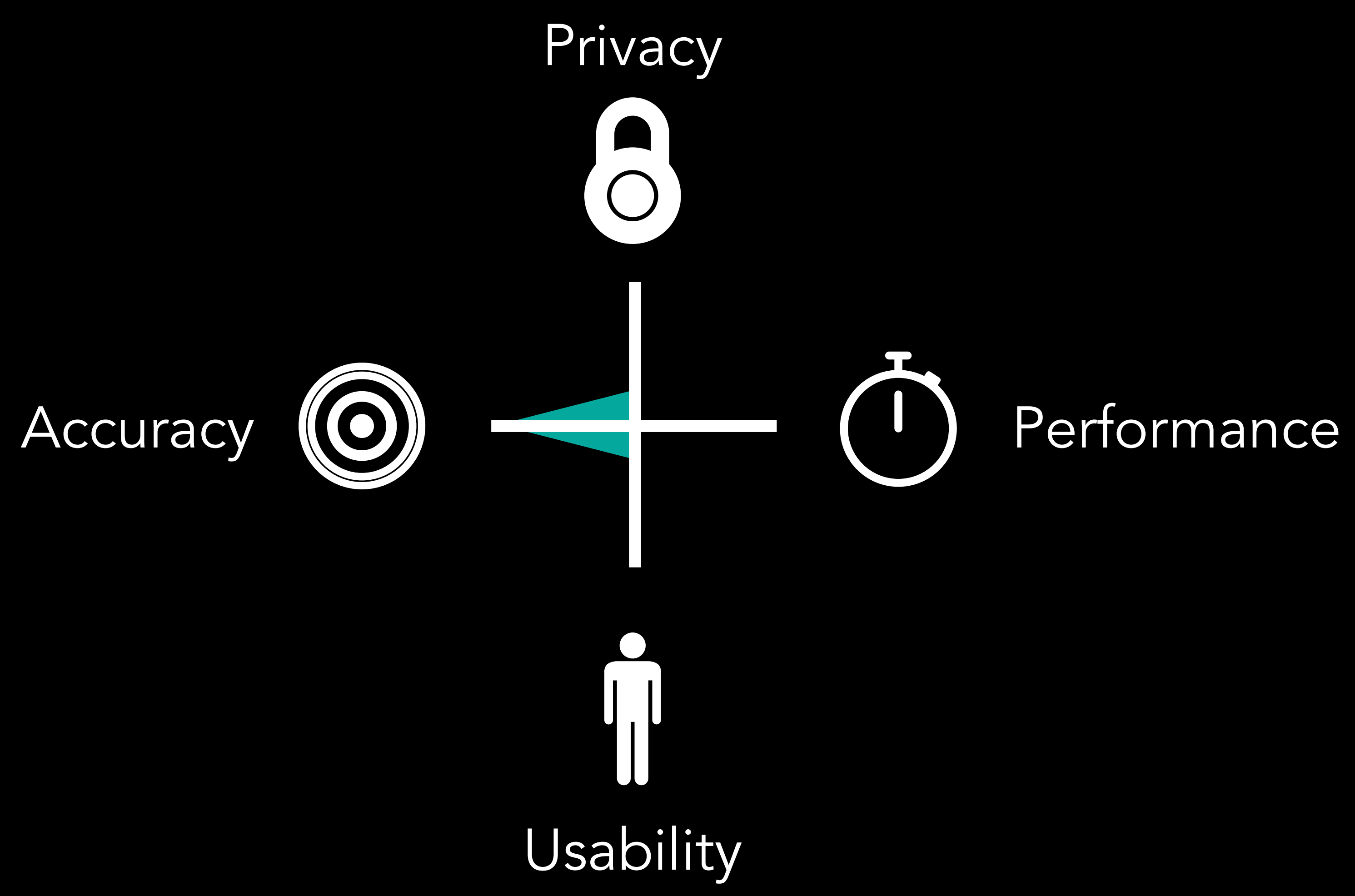


Building Blocks

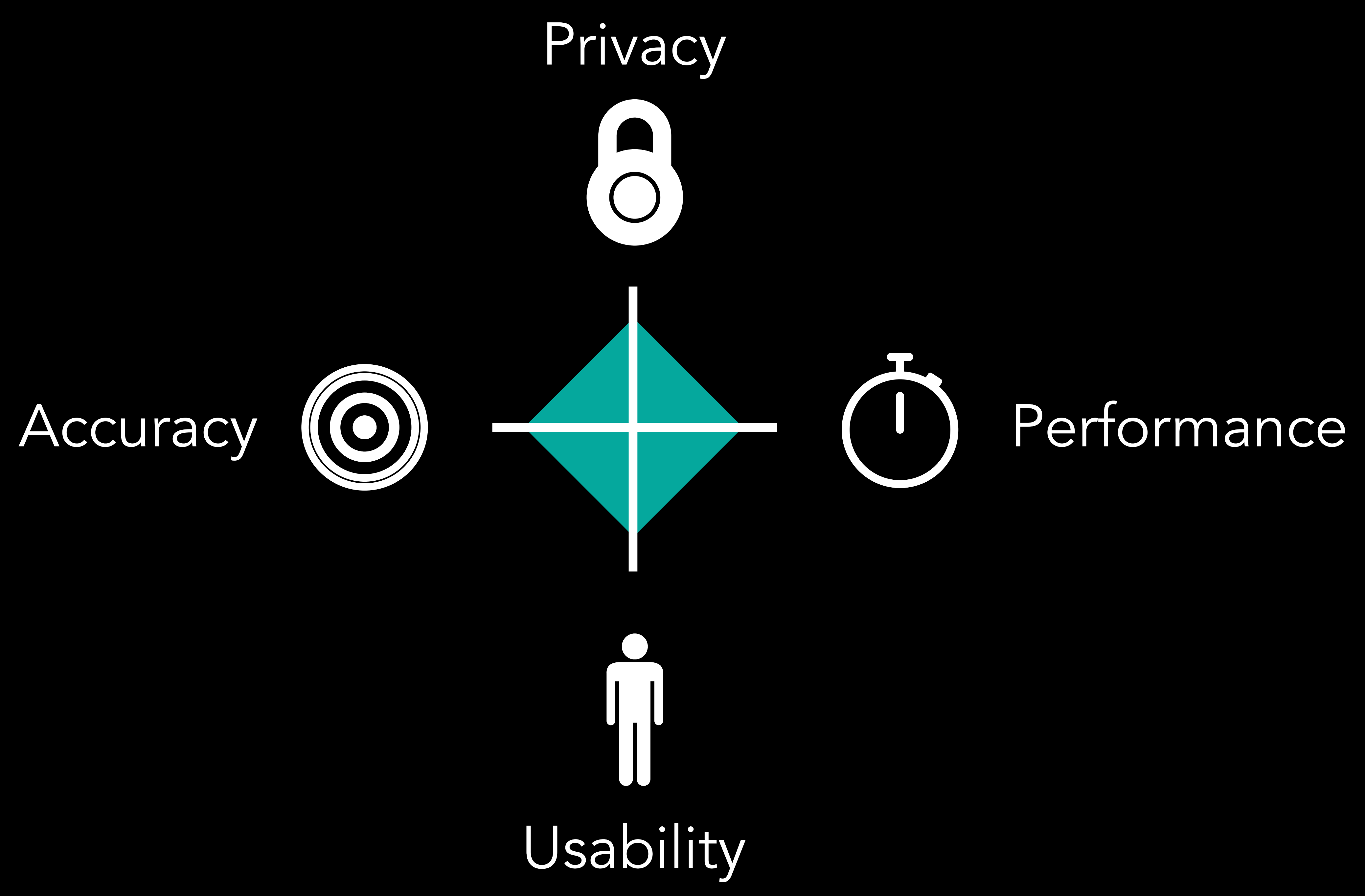
Differential Privacy



Secure Multiparty Computation



Private Data Federation



SQL Query Interface

Allows users to submit SQL queries to a single unified interface

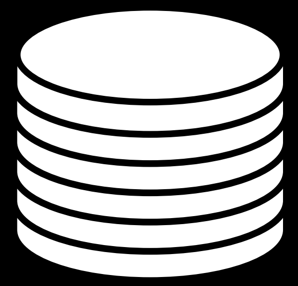
Secure Query Evaluation

Optimizes secure multiparty computation for query evaluation

Differentially-Private Guarantees

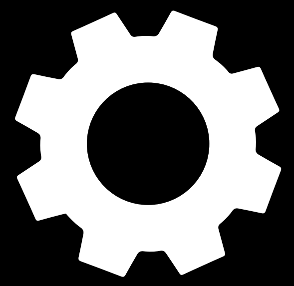
Provides differentially-private guarantees for query results

Privacy Challenges



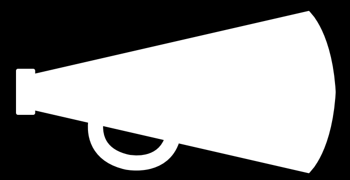
Data Storage

Can an attacker directly access private data?



Data Computation

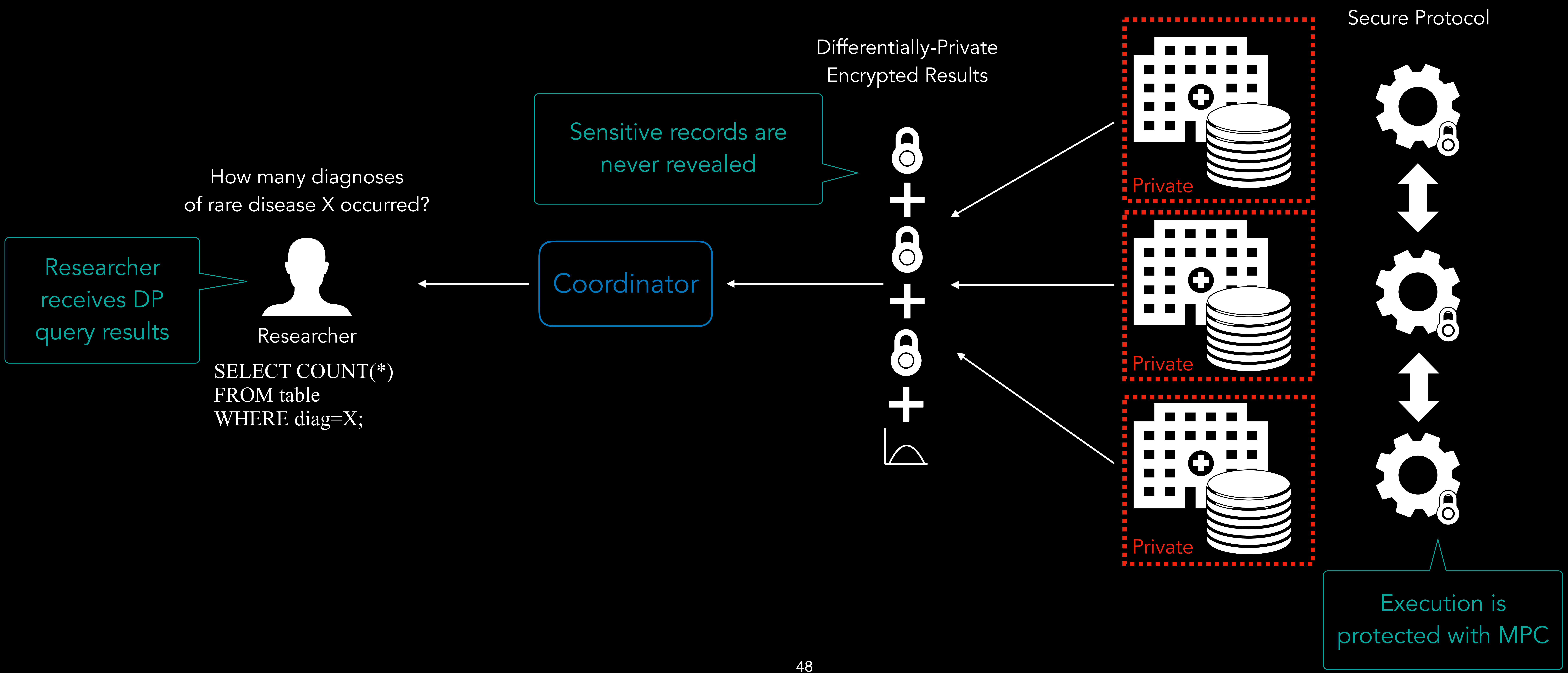
Can an attacker reconstruct private data by measuring computation?



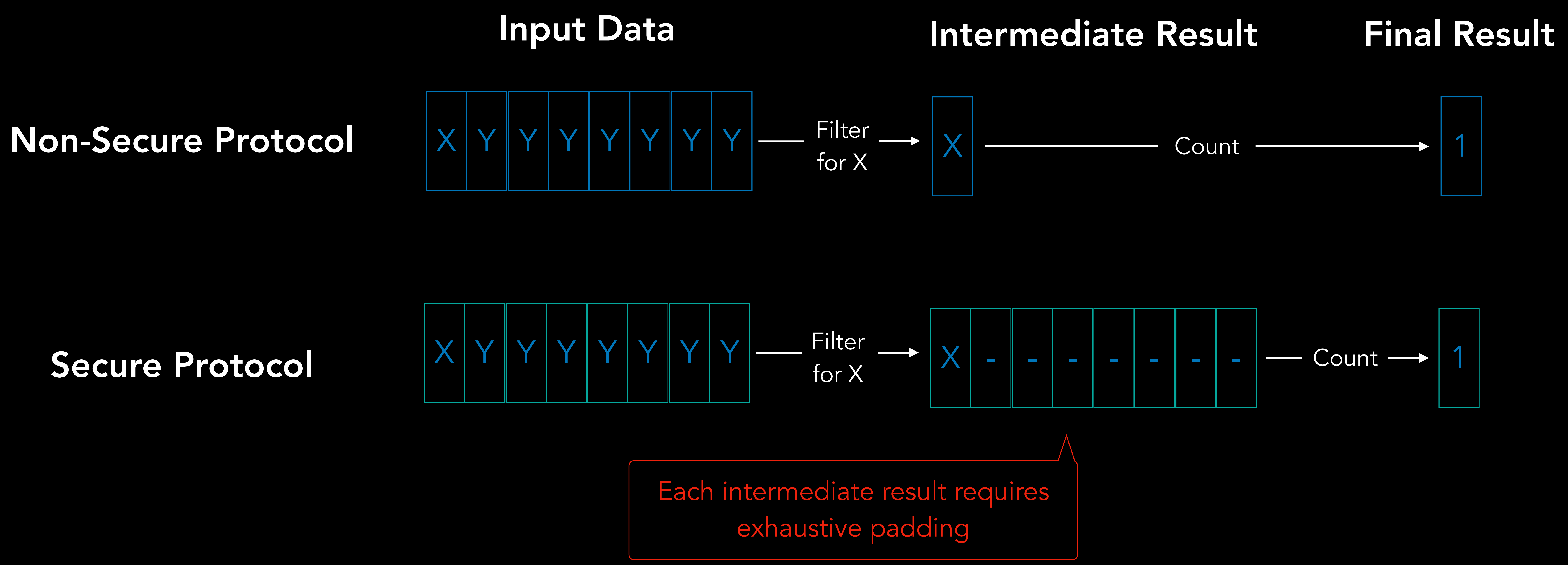
Data Release

Can an attacker reconstruct private data from published results?

Privacy Challenges

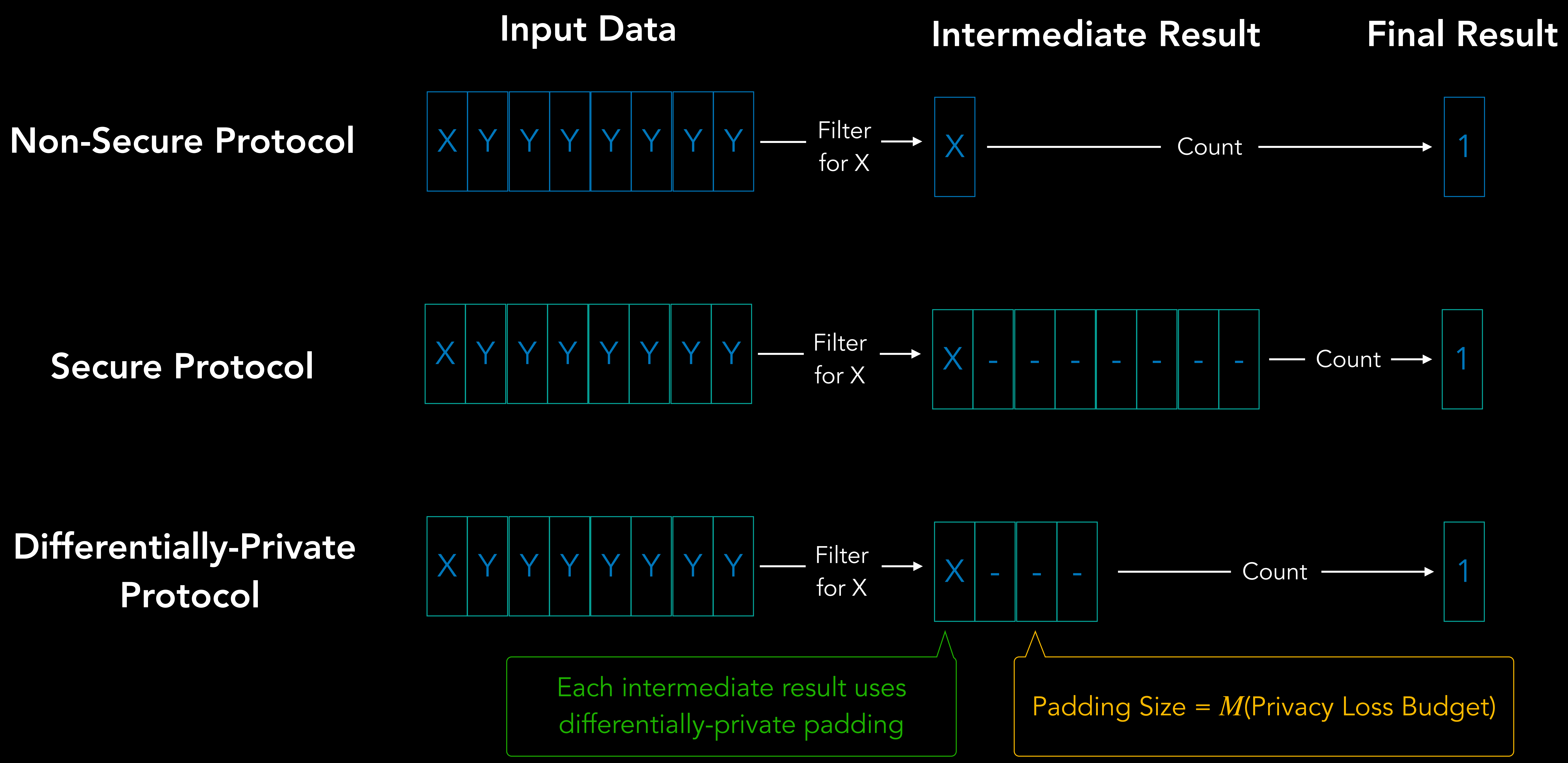


Performance Challenge



Secure Multiparty Computation requires worst-case execution to protect data during execution

Performance Challenge



Usability Challenges

SQL to Secure Code Translation

How do users write C-style code for MPC?

Privacy Budget Allocation

How do users split the privacy loss budget across query operators?

Usability Challenges

```
int$dSize[m*n] join(int$lSize[m] lhs, int$rSize[n] rhs) {
      int$dSize[m*n] dst;
      int dstIdx = 0;

      for(int i = 0; i < m; i=i+1) {
          int$lSize l = lhs[i];
          for(int j = 0; j < n; j=j+1) {
              int$rSize r = rhs[j];
              if($filter(l, r) == 1) {
                dst[dstIdx] = $project;
                dstIdx = dstIdx + 1;
              }
          }
       }
    return dst;
}
```

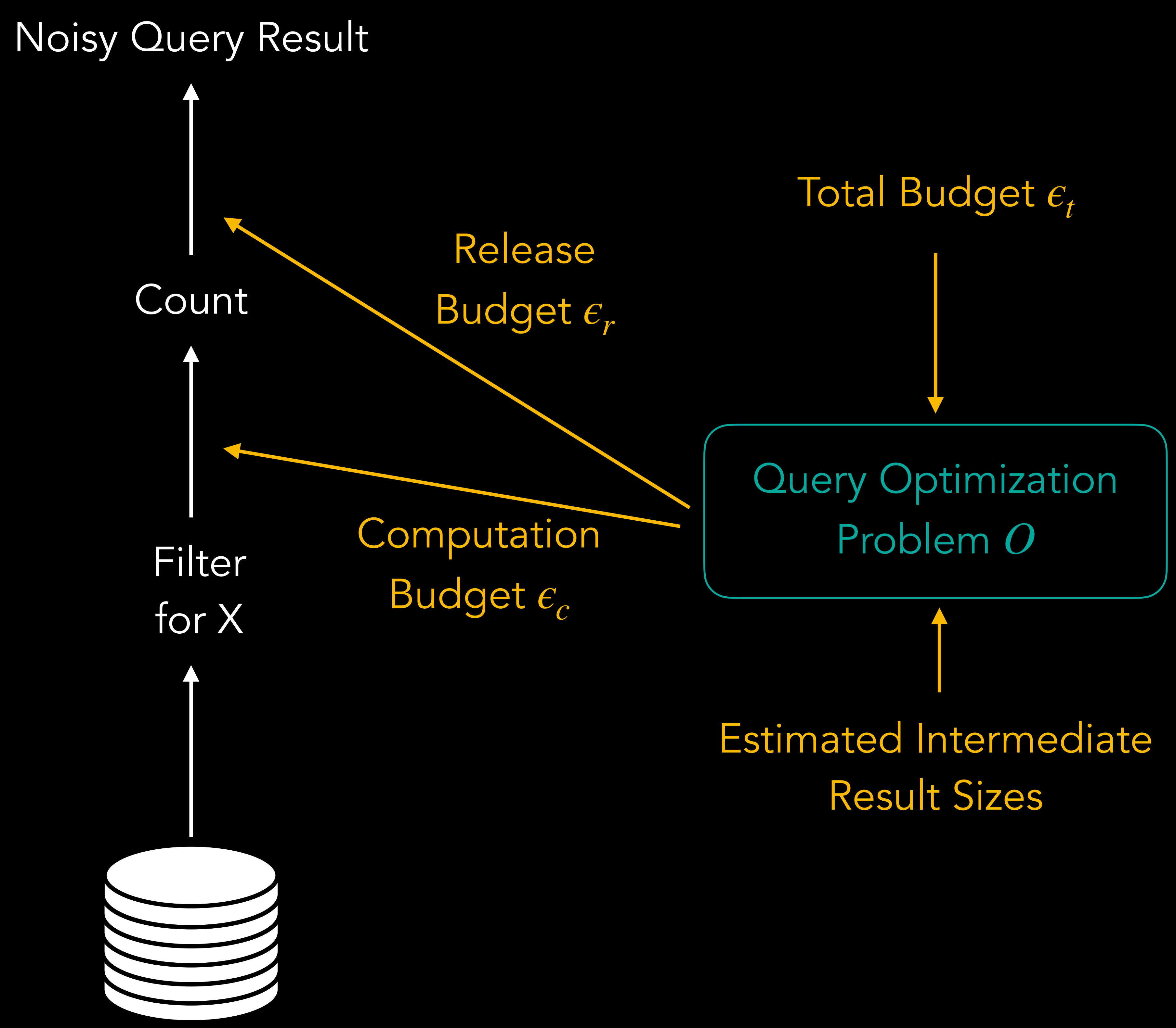
SQL to Secure Code Translation

Automatically converts SQL to secure code at codegen and runtime

Privacy Budget Allocation

How do users split the privacy loss budget across query operators?

Usability Challenges



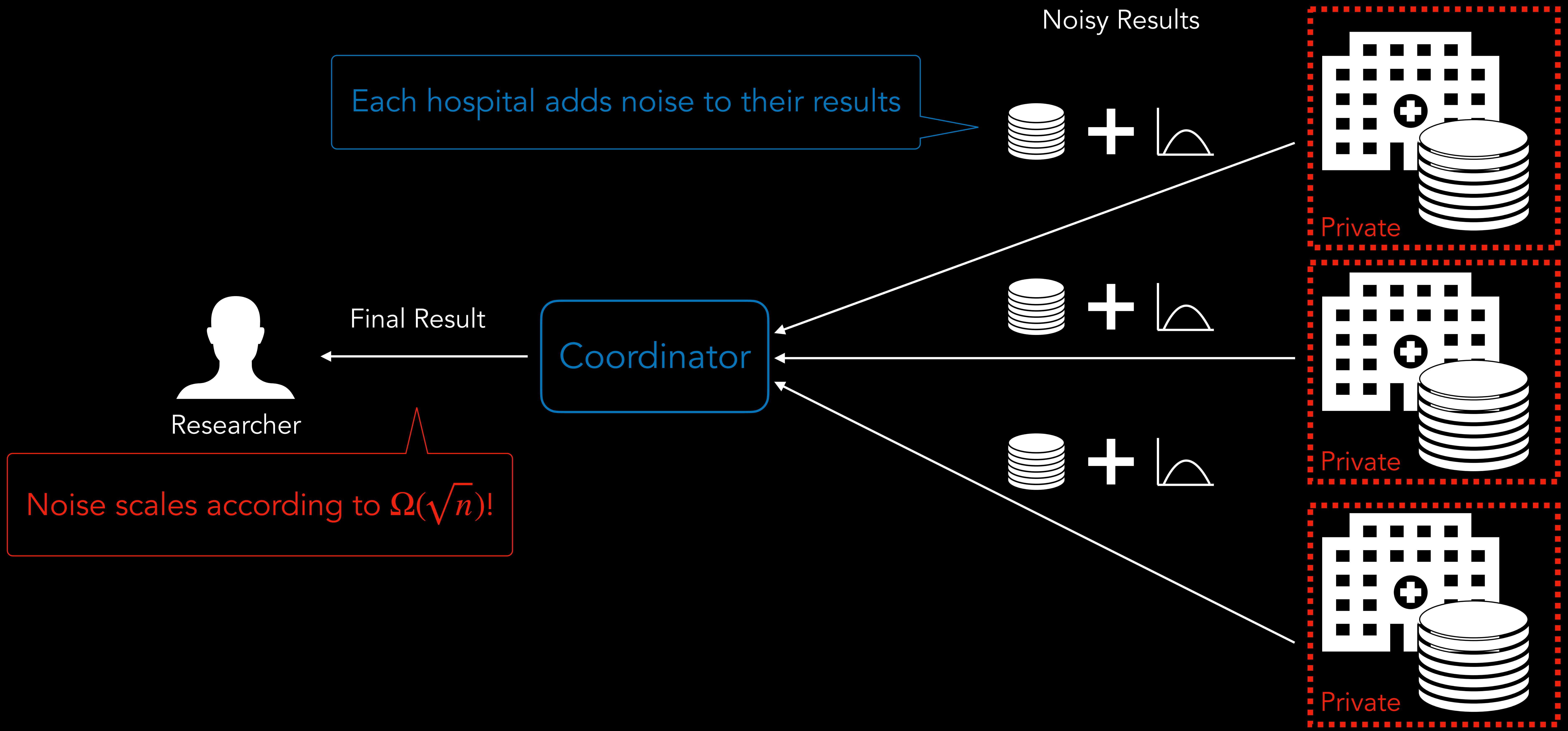
SQL to Secure Code Translation

Automatically converts SQL to secure code at codegen and runtime

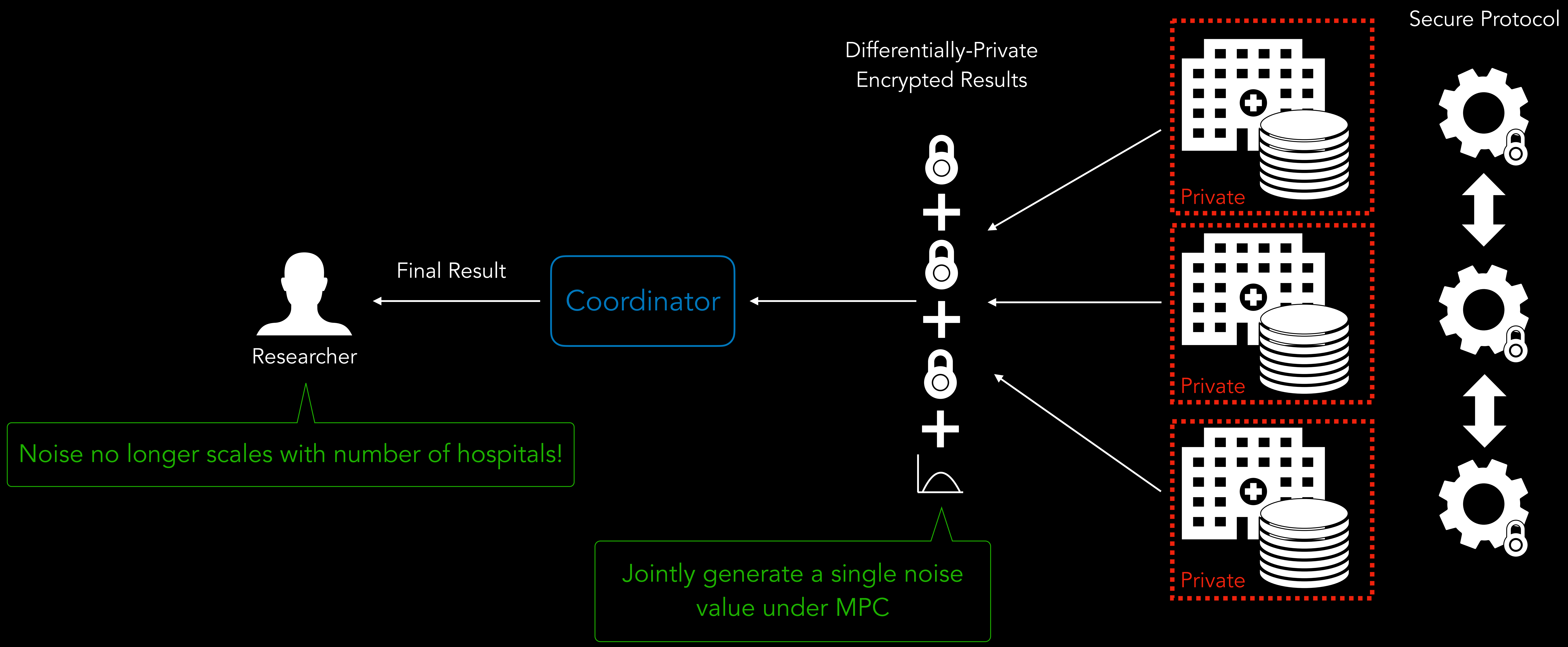
Privacy Budget Allocation

Optimal allocation of a privacy loss budget without user intervention

Accuracy Challenge



Accuracy Challenge

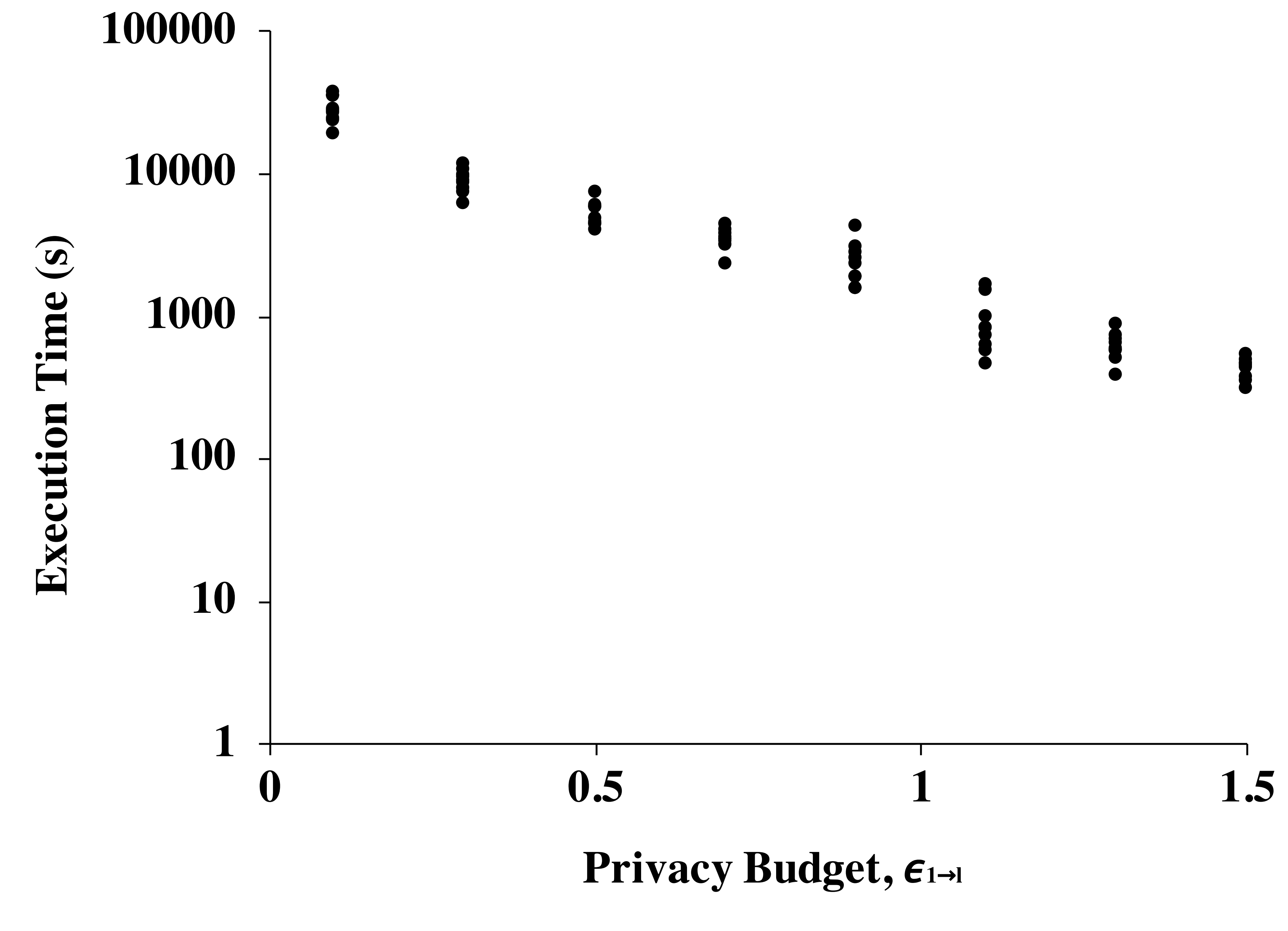


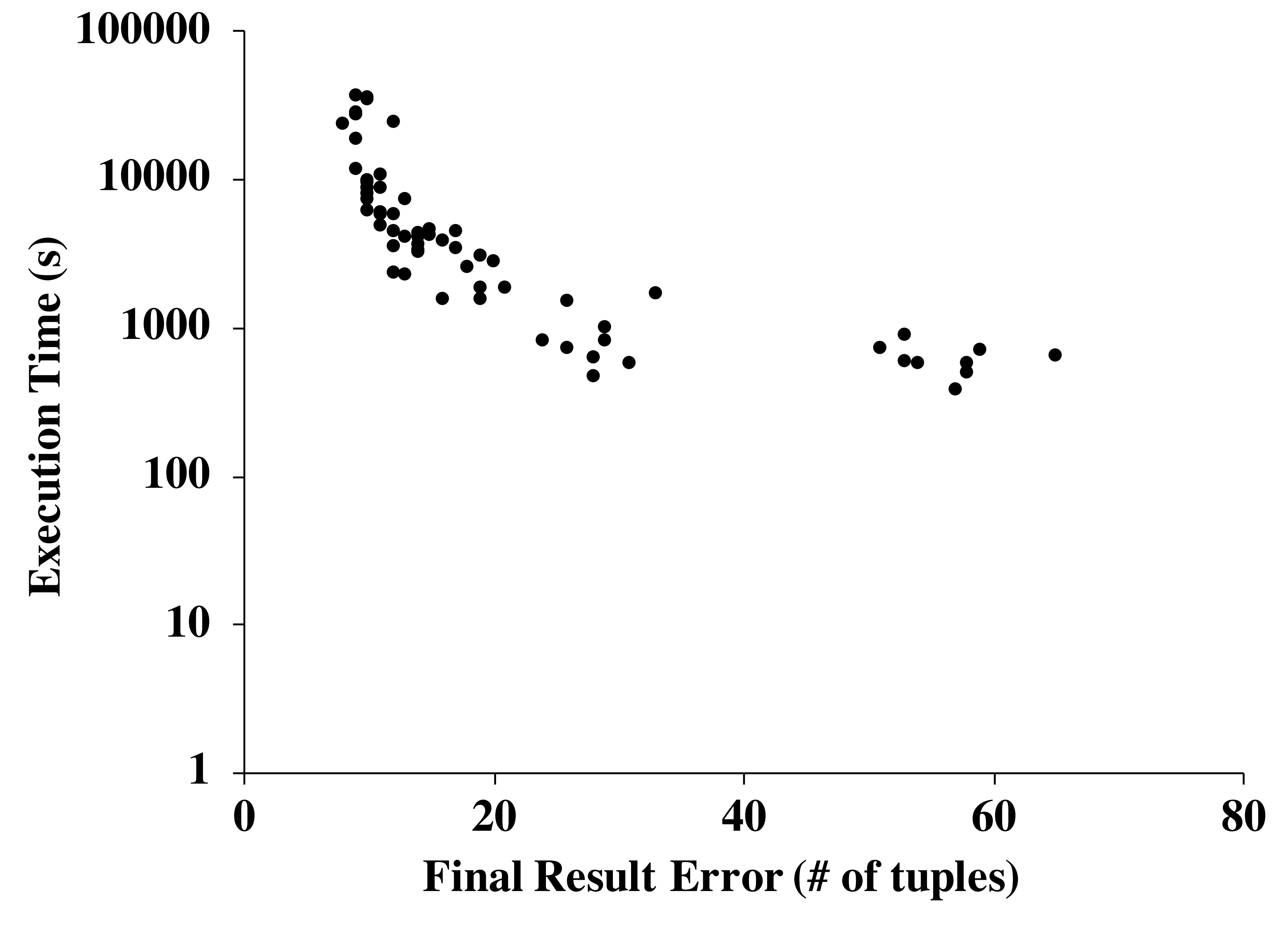
Private Data Federation

Experimental Results

- Ran experiments using one year of data from a Chicago-area hospital
- Source data size of ~500,000 patient records (15 GB)
- Synthetic data size of 750 GB
- Used benchmark queries provided by medical researcher

Performance Trade-offs



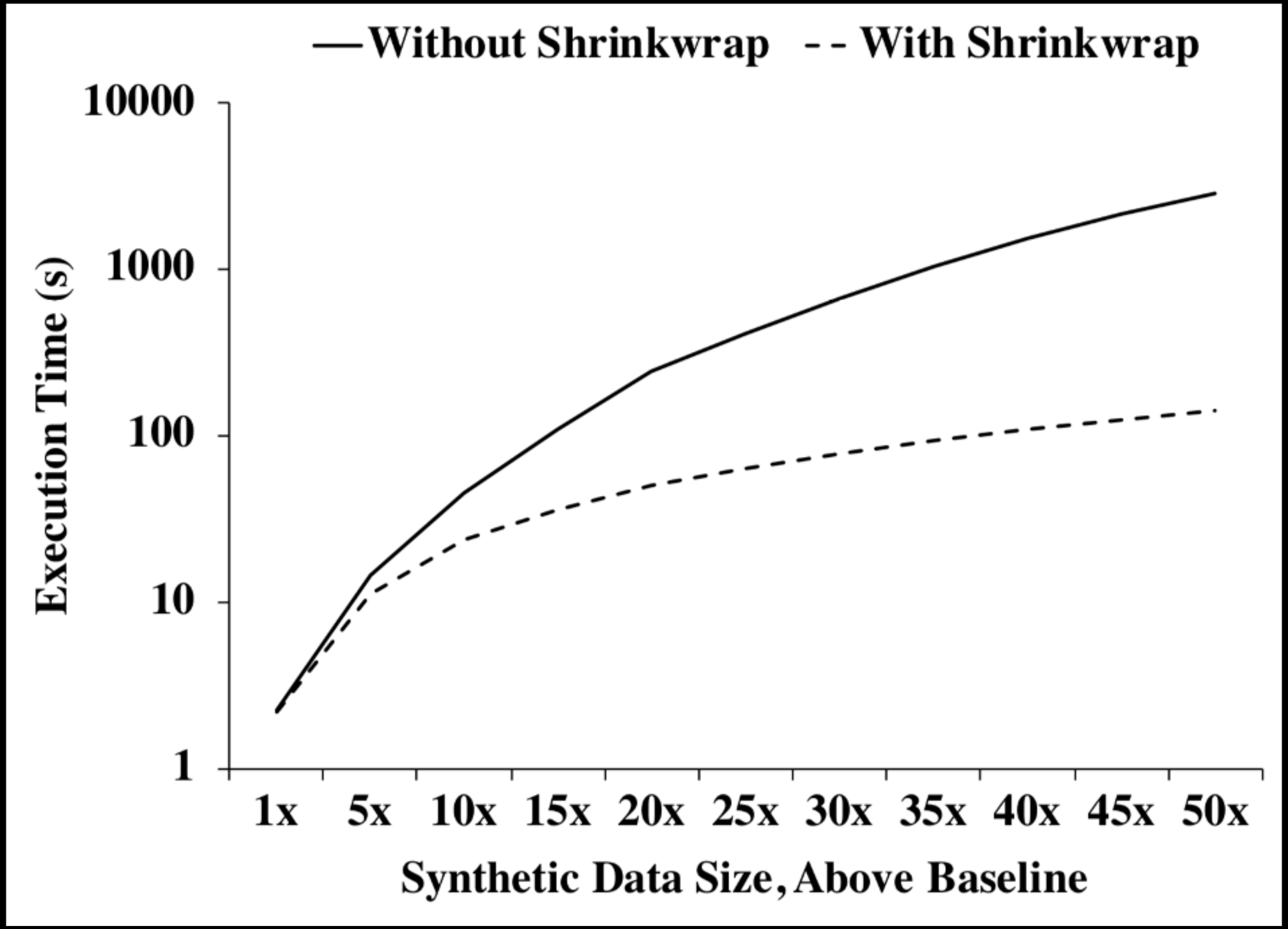


$\epsilon = 0.5$, $\delta = 1 \times 10^{-5}$

Lower Privacy, Higher Performance

Higher Accuracy, Lower Performance

Scaling with Data Size



$\epsilon = 0.5$, $\delta = 1 \times 10^{-5}$

More Data, More Speed Up!

Private Data Federation

Data release privacy with differential privacy

Privacy

Data computation privacy with MPC

Higher accuracy by using MPC to compute differentially private noise

Accuracy

Performance

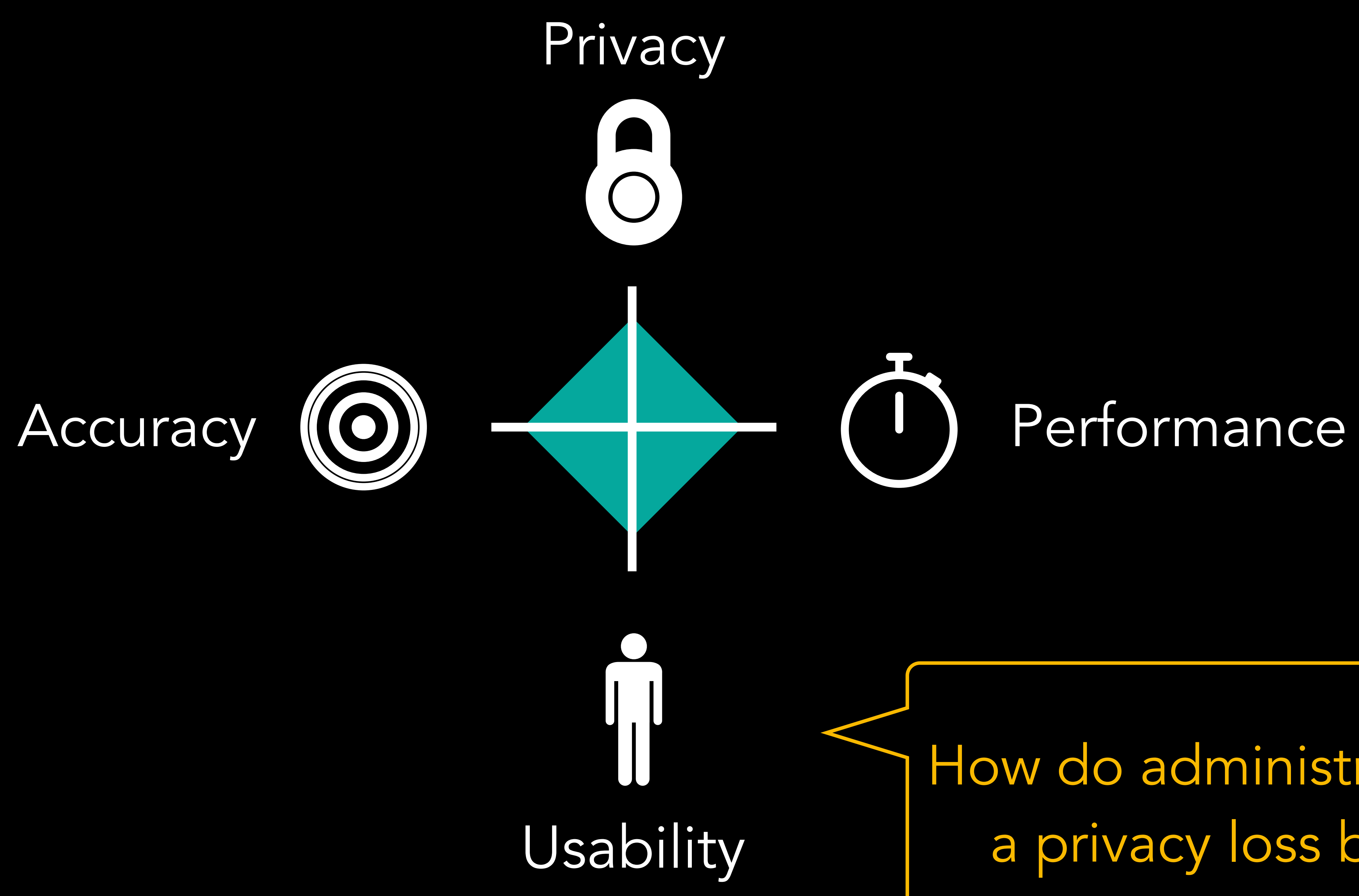
Optimized performance by using differential privacy to improve MPC

Automatic SQL to MPC translation through code generation

Usability

Automatic privacy loss budget usage through query optimization

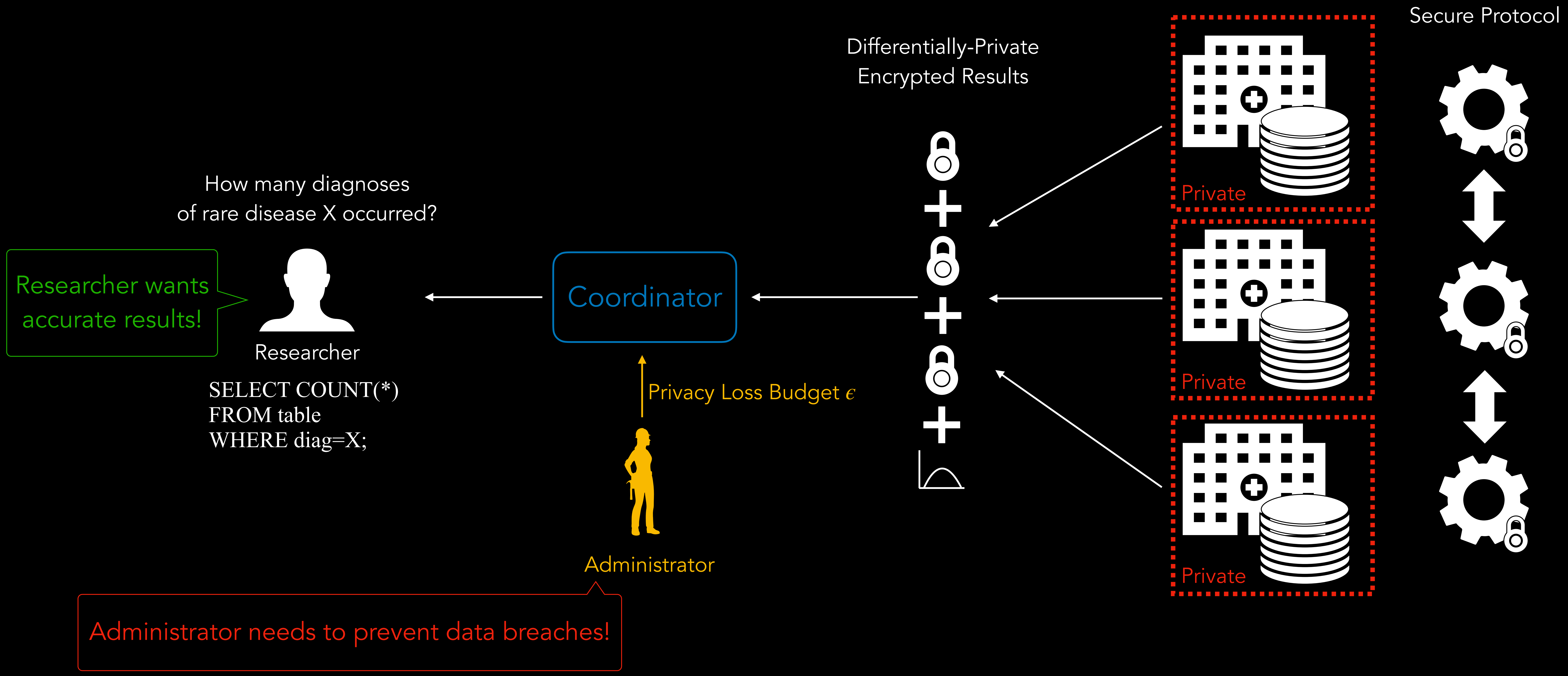
Private Data Federation



How do administrators pick a privacy loss budget?

Visualizing Privacy Trade-offs

Private Data Federation



Visualizing Privacy

Researchers want to release computed statistics

I need to prevent data breaches due to data releases

Administrator

How do I trade-off between accuracy and risk?

System Challenges

Relating the Privacy Loss Budget to Accuracy

Can non-expert administrators understand the relationship between accuracy and the privacy loss budget?

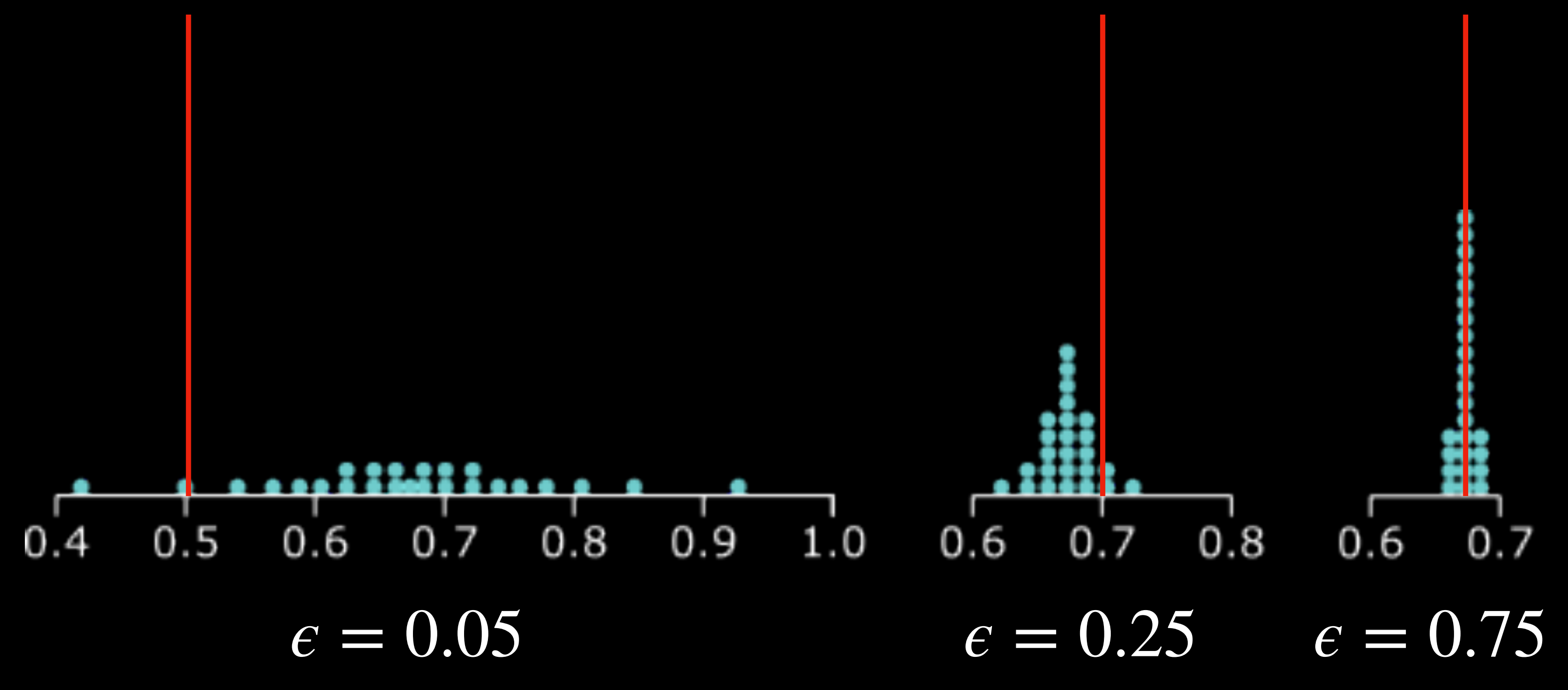
Relating the Privacy Loss Budget to Risk

Can non-expert administrators understand the relationship between risk and the privacy loss budget?

Choosing a Privacy Loss Budget

Can non-expert administrators pick the right privacy loss budget for their desired goals?

Relating Privacy Loss Budget to Accuracy



Visualizing Probability Distributions

Quantile dot plots with hypothetical outcomes visually describe DP mechanisms

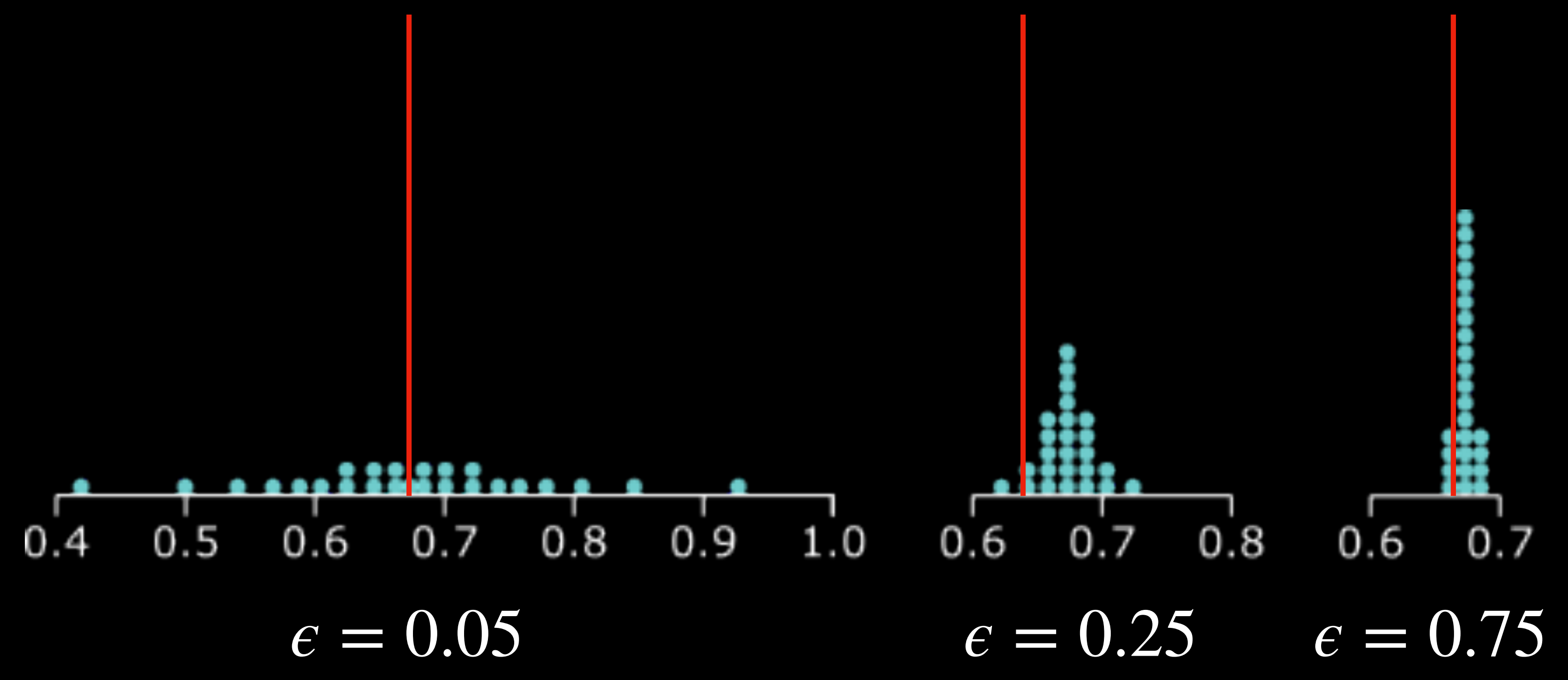
Linking Privacy Loss Budget to Accuracy

A selected privacy loss budget visually corresponds to a specific accuracy level

Intuition for Non-Experts

Administrators do not require expert knowledge to understand trade-offs

Relating Privacy Loss Budget to Accuracy



Visualizing Probability Distributions

Quantile dot plots with hypothetical outcomes visually describe DP mechanisms

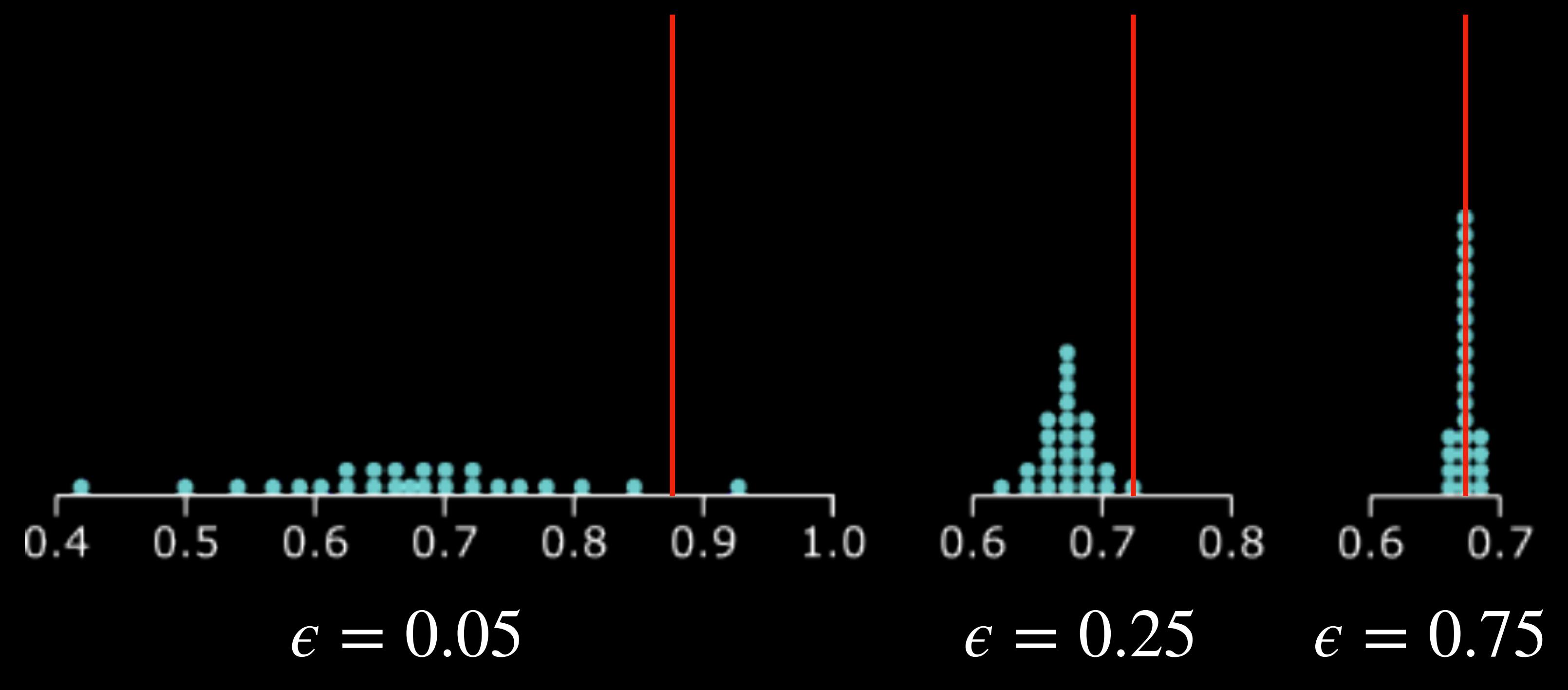
Linking Privacy Budget to Accuracy

A selected privacy loss budget visually corresponds to a specific accuracy level

Intuition for Non-Experts

Administrators do not require expert knowledge to understand trade-offs

Relating Privacy Loss Budget to Accuracy



Visualizing Probability Distributions

Quantile dot plots with hypothetical outcomes visually describe DP mechanisms

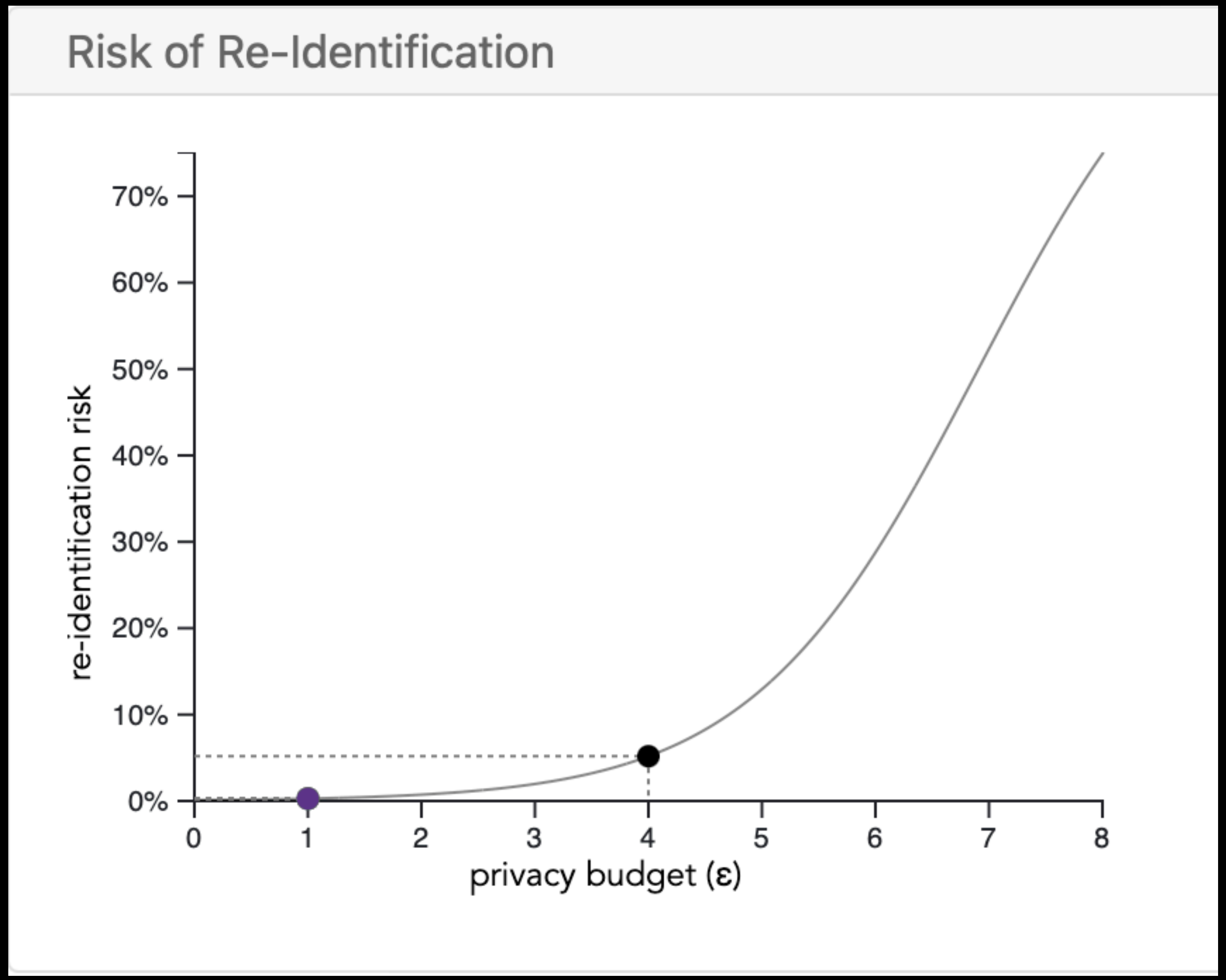
Linking Privacy Budget to Accuracy

A selected privacy loss budget visually corresponds to a specific accuracy level

Intuition for Non-Experts

Administrators do not require expert knowledge to understand trade-offs

Relating Privacy Loss Budget to Risk



Visualizing (one of many) Attack Models

Graph shows how risk changes as a function of the privacy loss budget

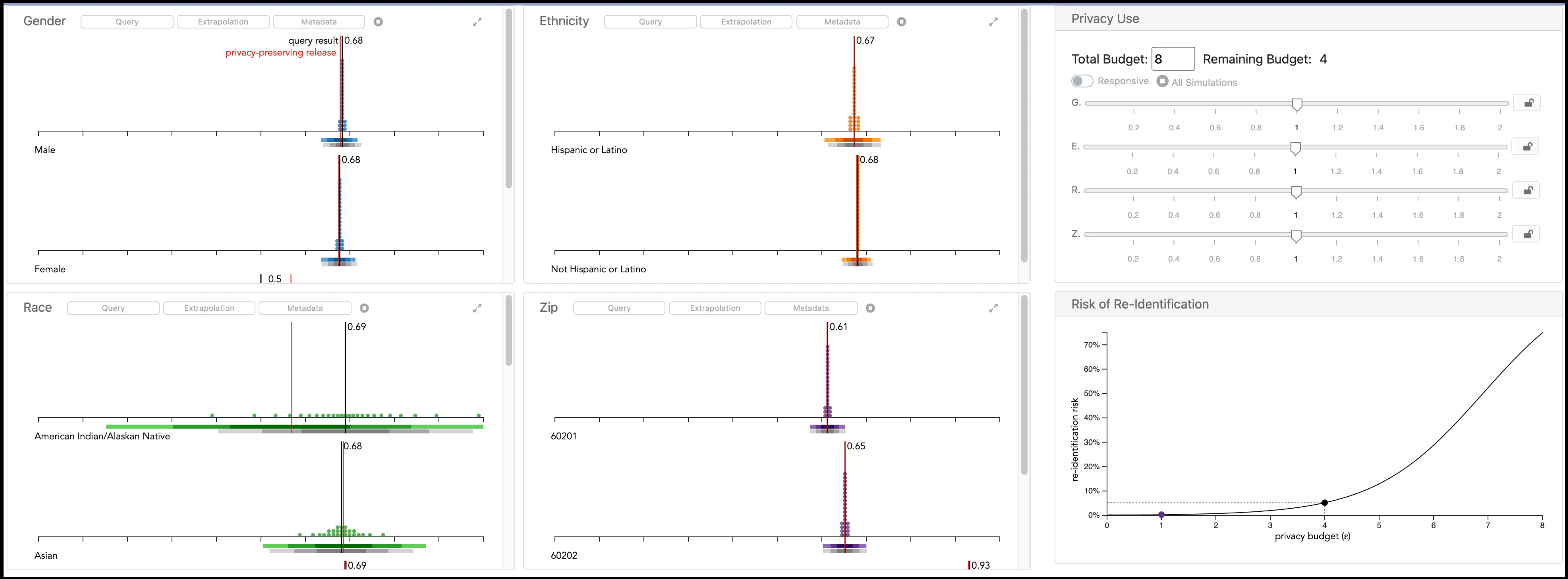
Linking Privacy Budget to Risk

A selected privacy loss budget visually corresponds to a specific risk level

Intuition for Non-Experts

Administrators do not require expert knowledge to understand trade-offs

Choosing a Privacy Budget



Qualitative User Study

- Interviewed 22 researchers
- Researchers worked with sensitive data, but unfamiliar with differential privacy
- Provided a 5-minute video tutorial on differential privacy
- Created a spreadsheet version of the interface as a control
- Compared the performance of researchers between interfaces
- Tasks were split into two versions and researchers were alternated on which interface was seen first

Example User Study Tasks

CDF Judgment

- At privacy loss budget = x, what is the probability that the privacy-preserving release for the A subgroup will be greater than y?

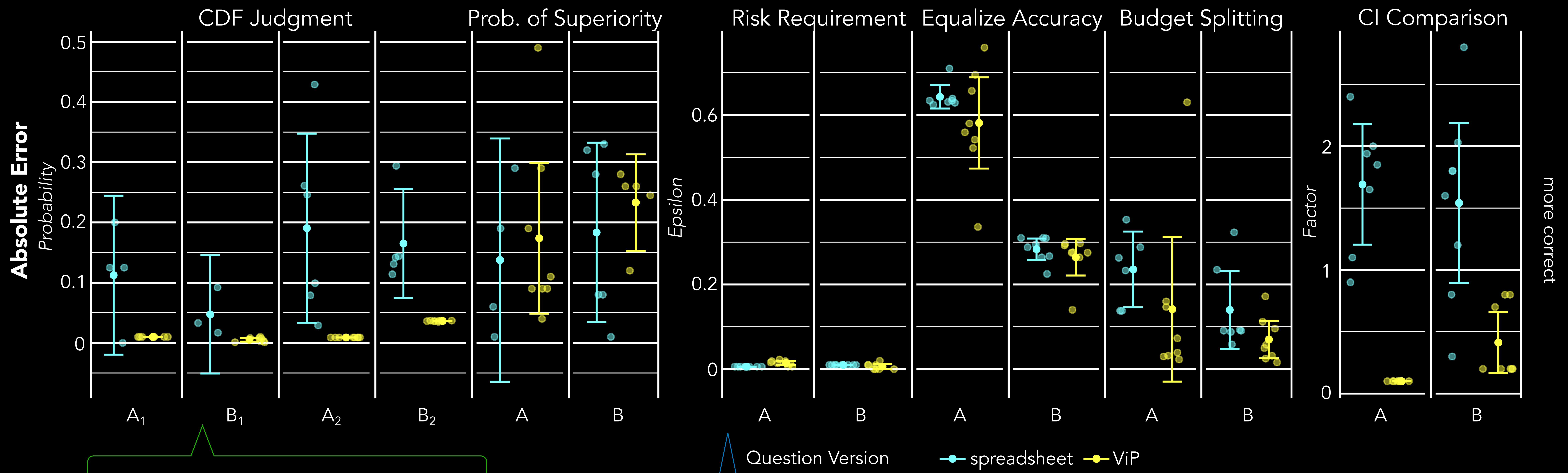
Probability of Superiority

- At privacy loss budget = x, estimate the probability that the release for the A subgroup will be greater than the release for the B subgroup.

Risk Requirement

- What value for the privacy loss budget is needed to achieve a risk less than or equal to X?

Study Results



Visualization improves participant answers for judgment questions!

Visualization does not improve answers for requirement questions

Study Results

"If I'm increasing a budget, and it's a privacy budget, it's counterintuitive to me. I would think the higher the budget the more you're spending on privacy, the lower your re-identification risk. It's easy to figure out once you start sliding it but I guess the first thing I thought is I'm increasing a budget, I should be spending more, which would mean increasing my re-identification risk"

Study Results

"I imagine many researchers are really tight about their estimates, and in health in particular it's so often you barely find any significance in the first place that, I mean in my work—and I work with a lot of data—and even then significance is not that easy to come by"

Study Results

"The dynamic aspect was the most useful, in other words literally watching where the release would fall and how often it would fall and how often it would fall outside a range… how often the query value would literally be outside the confidence interval of the release"

Study Results

Risk Awareness

- Participants reported that the interface made them more cognizant of risk when working with sensitive user data

Understanding Uncertainty

- Participants reported that the interface let them understand how accuracy changes as a function of the chosen DP mechanism

Trade-off Intuition

- Participants reported that the interface gave them an intuition about the utility vs risk trade-off and allowed them to make quick mental calculations

Visualizing Privacy Trade-offs

Relating the Privacy Loss Budget to Accuracy

Uncertainty visualization gives users an intuition about privacy mechanism accuracy

Relating the Privacy Loss Budget to Risk

Risk visualization pushes users to carefully consider risk implications of data release

Choosing a Privacy Loss Budget

Users develop an intuition about the privacy vs utility trade-off through interactive interface controls

Summary

Summary

Protect people and their data

Use DP and MPC to protect sensitive data from end-to-end

Build useful systems

Combine DP and MPC to optimize the privacy vs utility trade-off

Minimize user intervention

Automatically translate MPC code and allocate DP privacy loss budget

Allow non-experts to use the system

Interactive interface that gives intuitive understanding of privacy vs utility trade-offs

Building Useful Systems That Protect People and Their Data

Johes Bater

Tufts University | School of Engineering Computer Science