Privacy-Preserving Database Systems:

Balancing Privacy and Utility for Query Execution

Johes Bater

Tufts University | School of Engineering

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

Organizations con… romise user data

List of data breaches

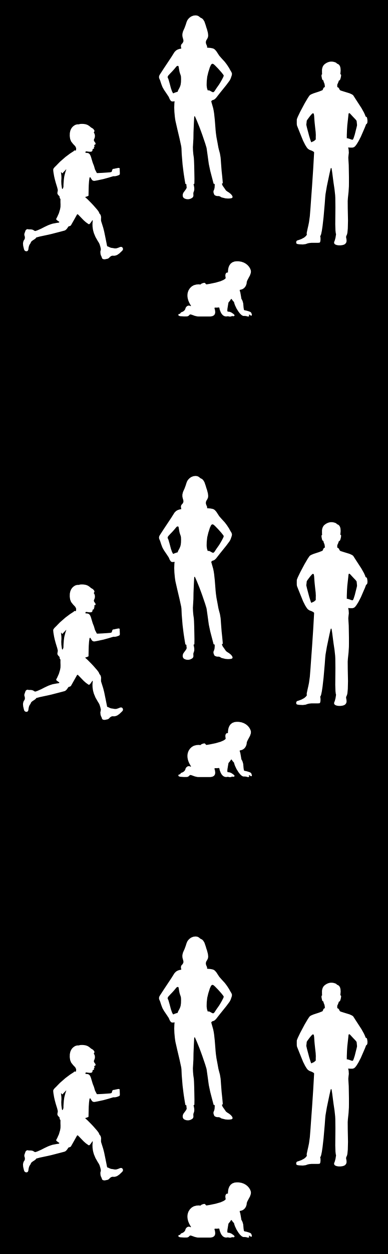
From Wikipedia, the free encyclopedia

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.

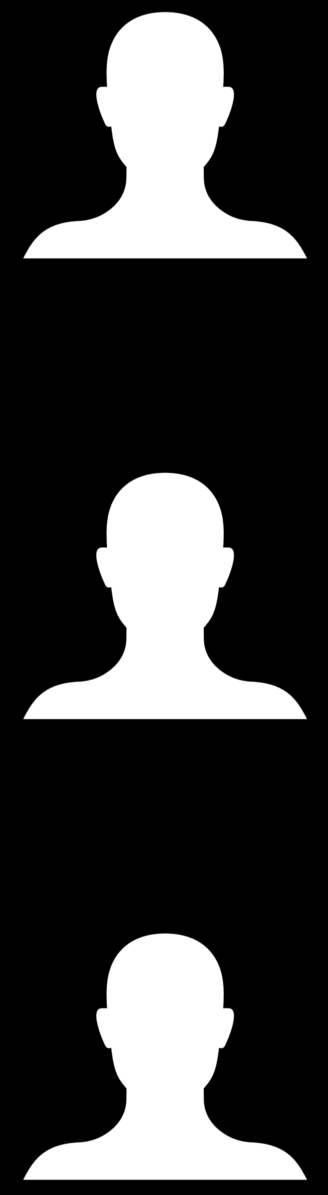
This is a list of **data breaches**, using data compiled from various sources, including press reports, government news releases, and mainstream news articles. The list includes those involving the theft or compromise of 30,000 or more records, although many smaller breaches occur continually. Breaches 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

Query Execution with an Untrusted Server

What about encrypted execution?

Information leaks even if computation is encrypted!

Information Leakage Side Channels

- Data Ingestion: Can reveal when events occur on the data owner
- Query Execution: Can reveal the exact data values

Focus of today's talk

- View Materialization: Can reveal how data changes over time
- Indexing: Can reveal the exact data distribution
- And many more

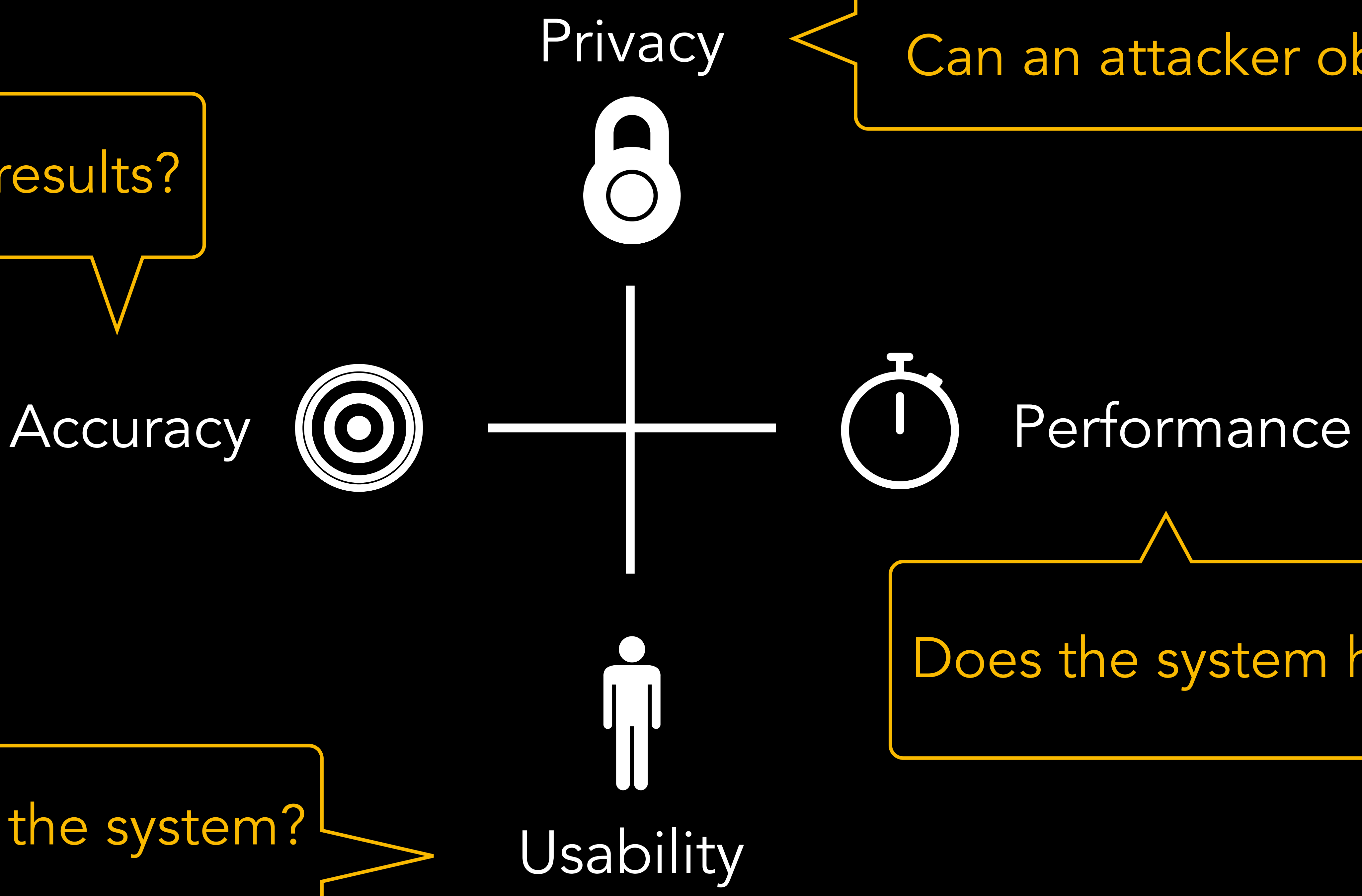
Any data dependent operation can leak information!

We need to ensure privacy while maintaining utility

System-Building Challenges

Can an attacker obtain sensitive user data?

Does the system provide accurate results?

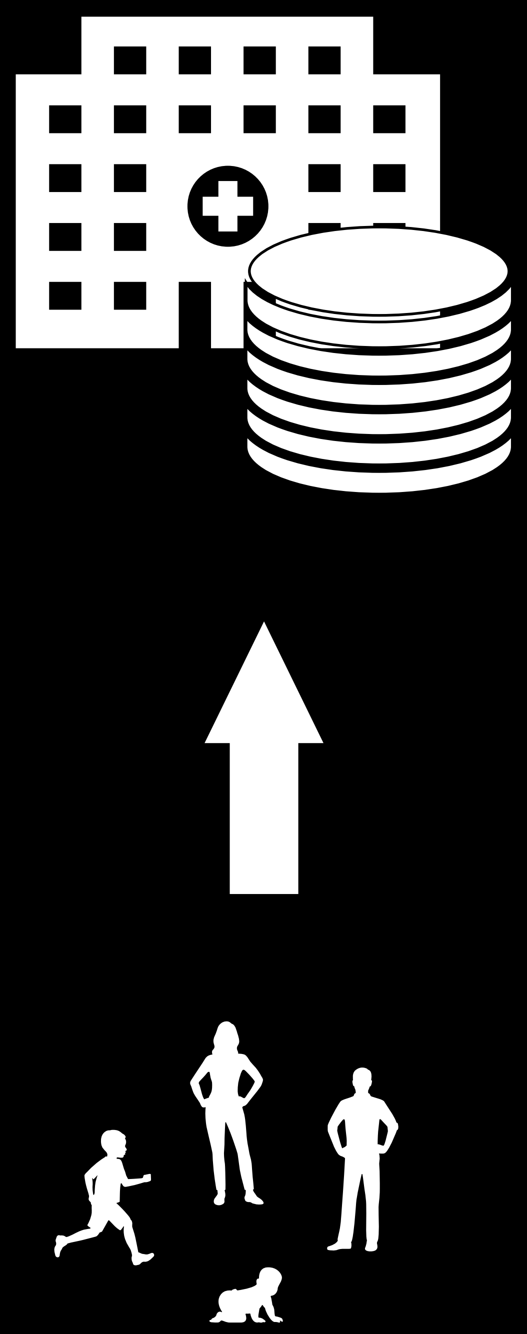


Does the system have acceptable execution time?

Can users understand and use the system?

Building a Private Data Federation

Example: Clinical Data



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

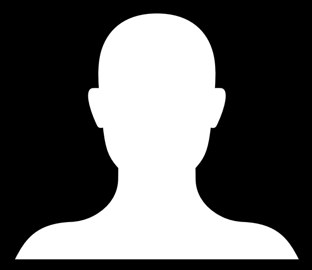
Example: Clinical Data

For this project, we partnered with HealthLNK, a Chicago-based consortium of healthcare sites that agree to share their data for research.

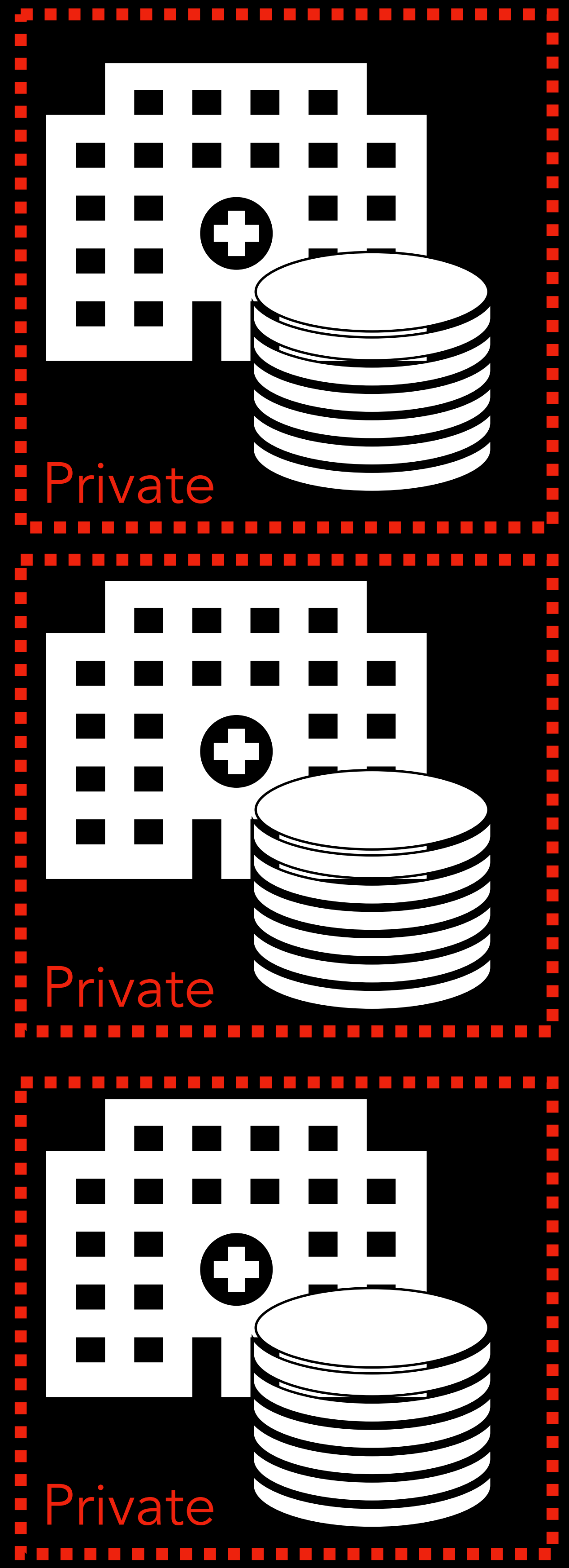
This project is part of a pilot study at three Chicago-area hospital networks used to identify patient populations that are potentially under-treated for hypertension.

Example: Clinical Data

How many diagnoses
of rare disease X occurred?

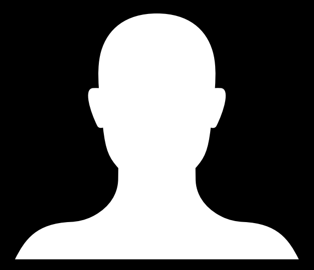


Researcher



Example: Clinical Data

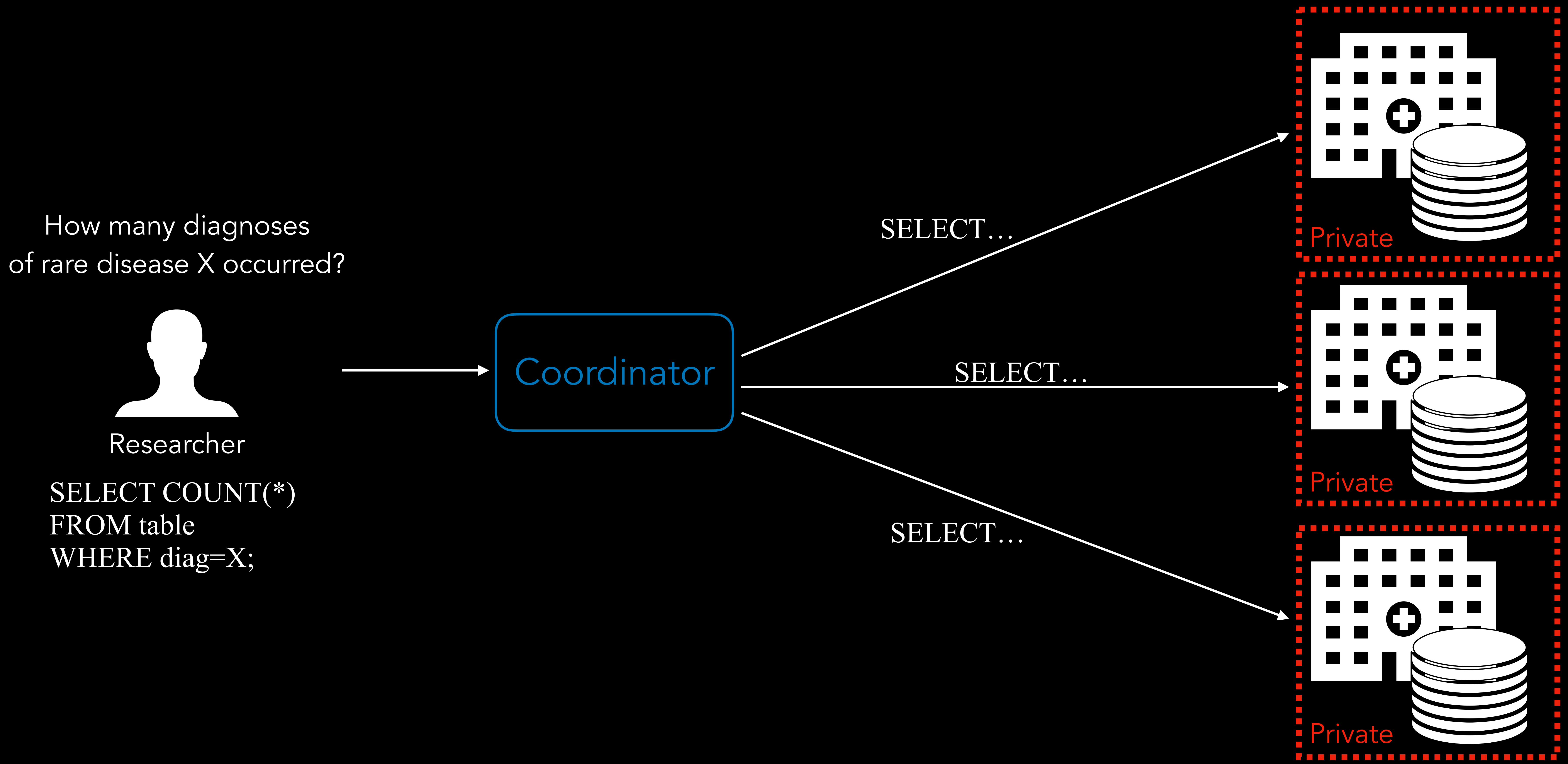
How many diagnoses
of rare disease X occurred?



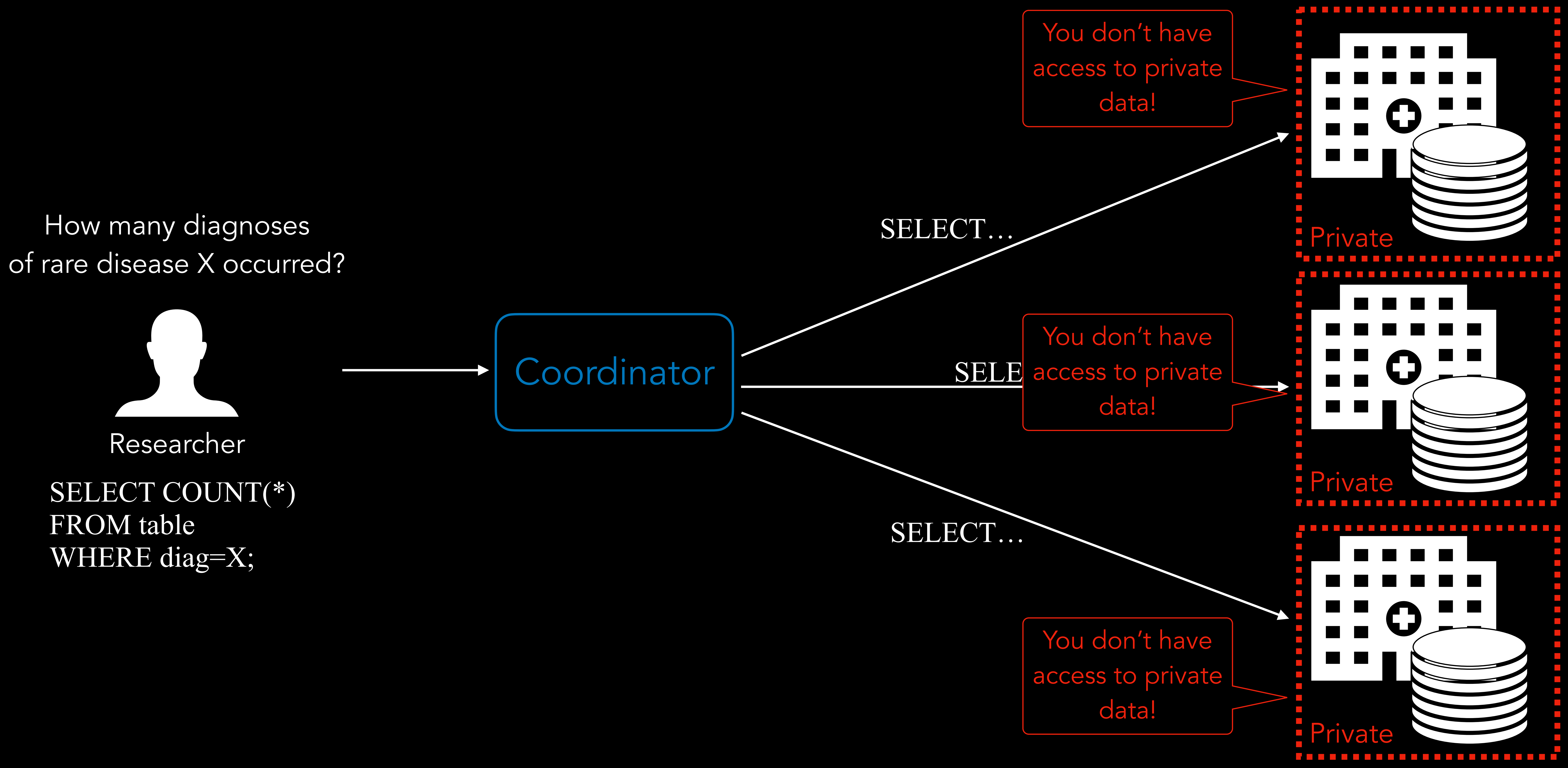
Researcher

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

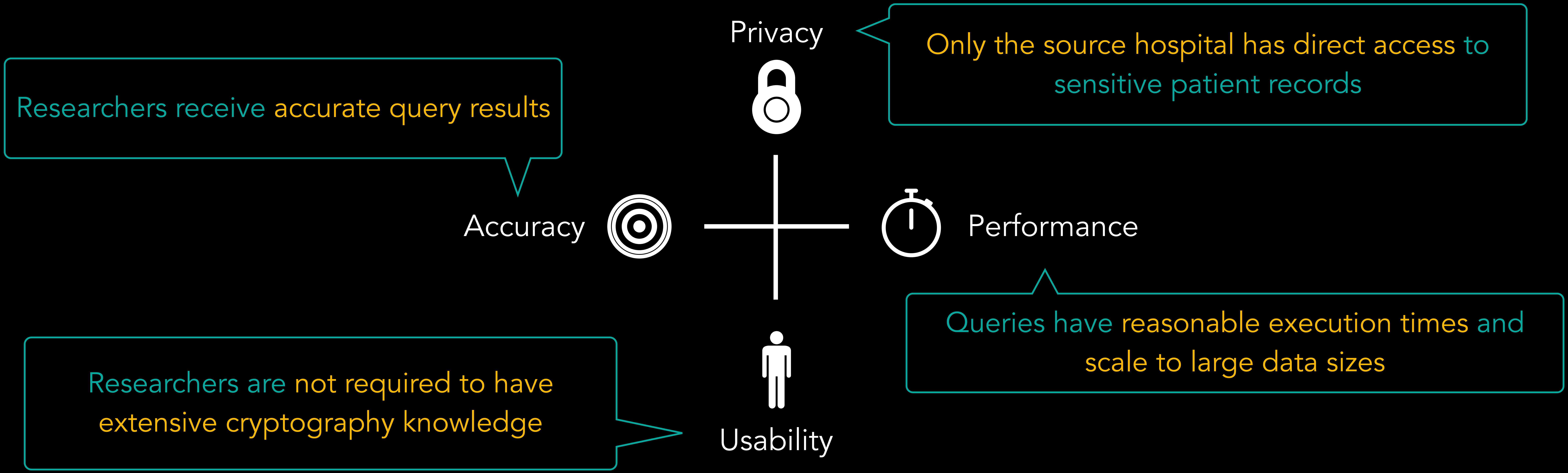
Example: Clinical Data



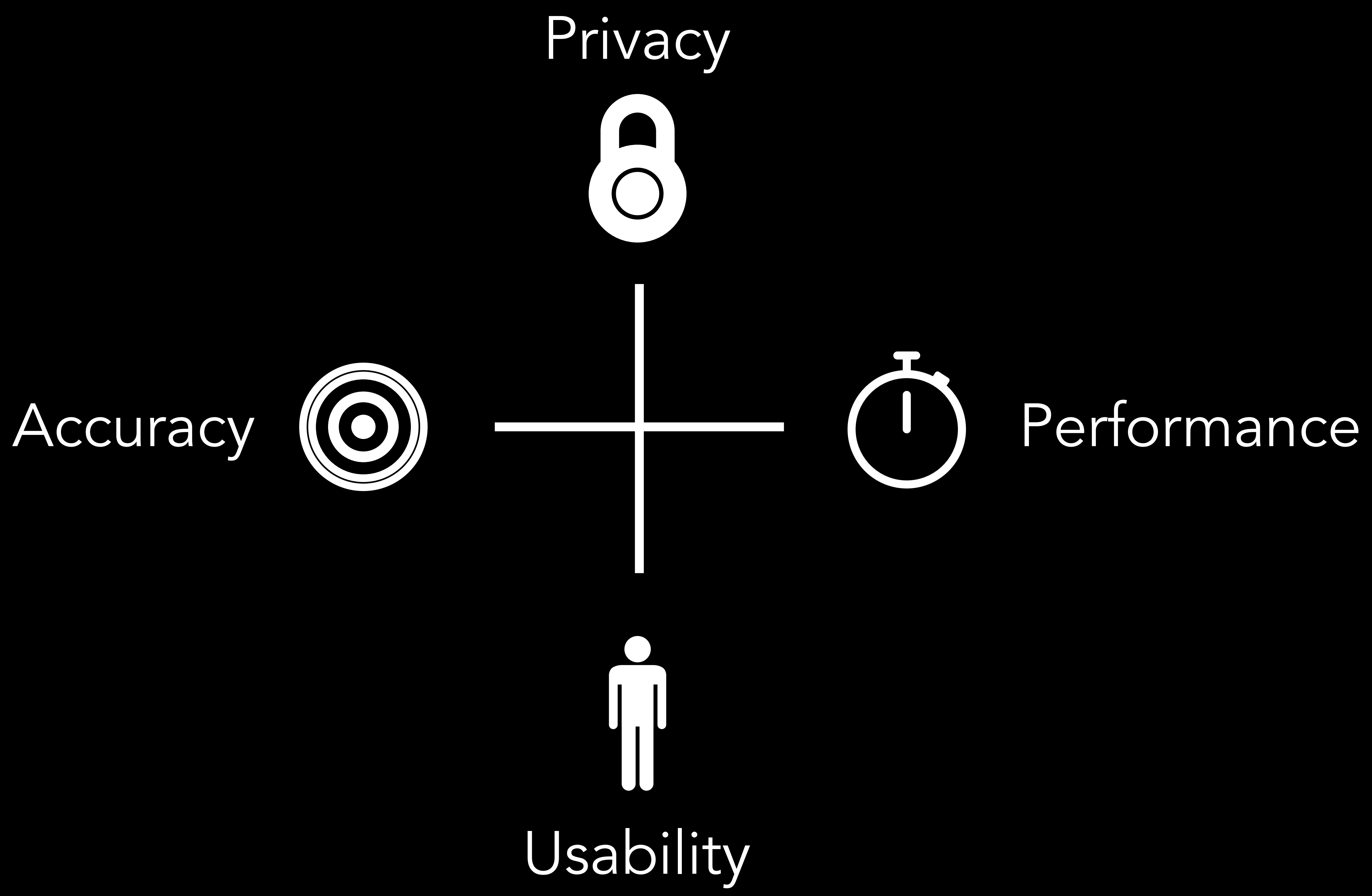
Example: Clinical Data



Private Data Federation Requirements



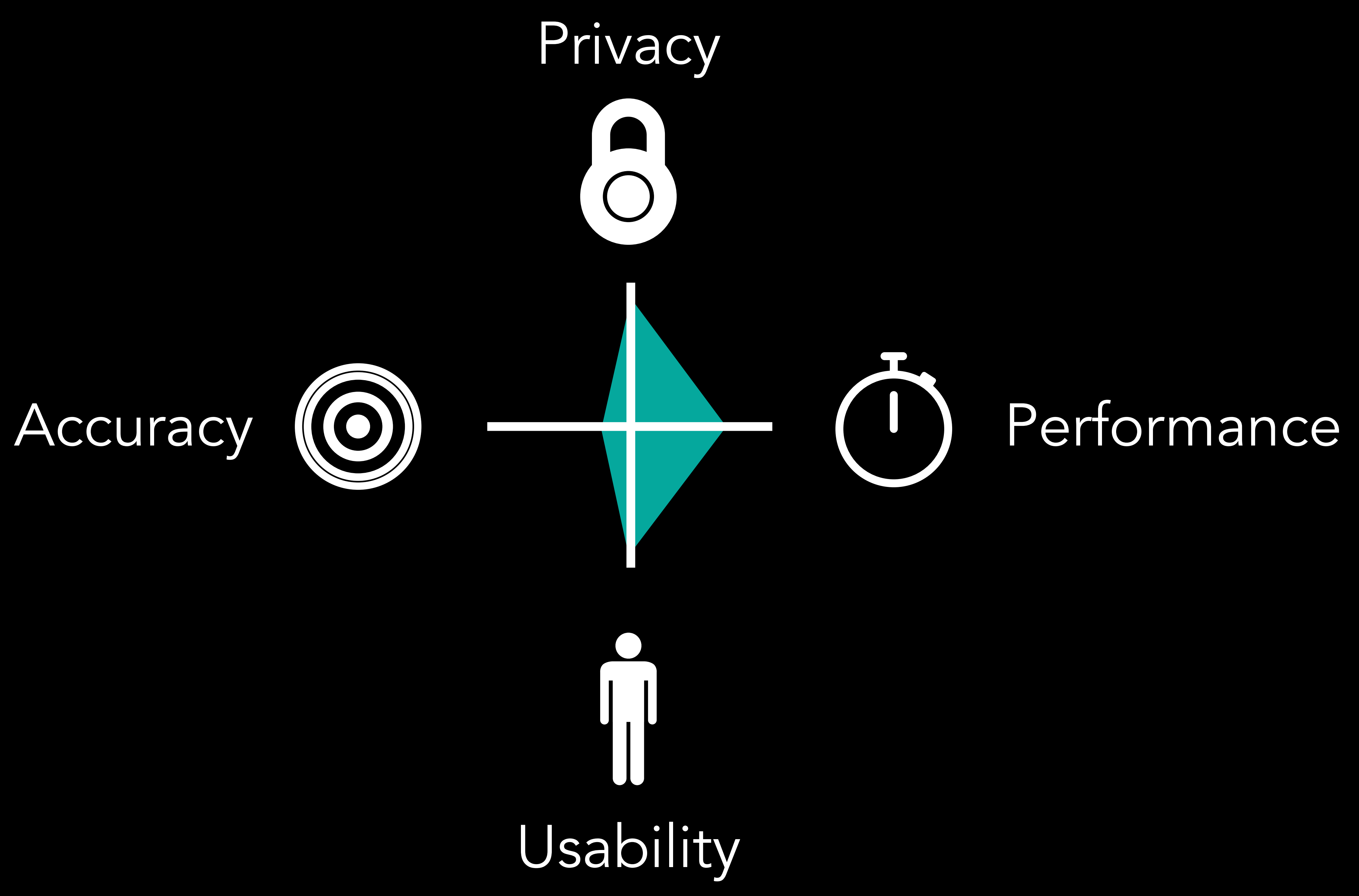
Building Blocks



Differential Privacy (DP)

Secure Multiparty Computation (MPC)

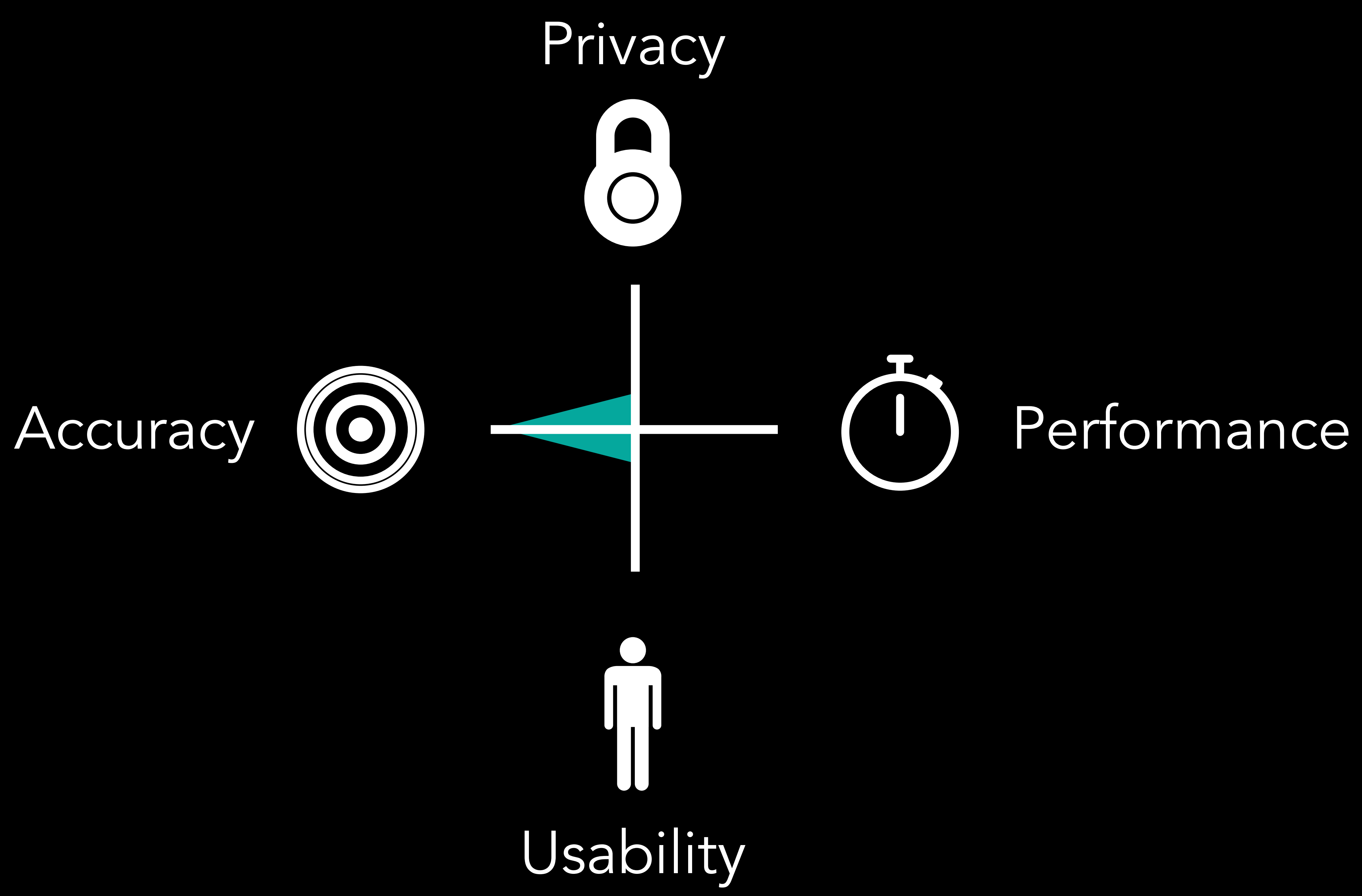
Building Blocks



Differential Privacy (DP)

Protect sensitive patient records by adding privacy-preserving noise

Building Blocks



Secure Multiparty Computation (MPC)

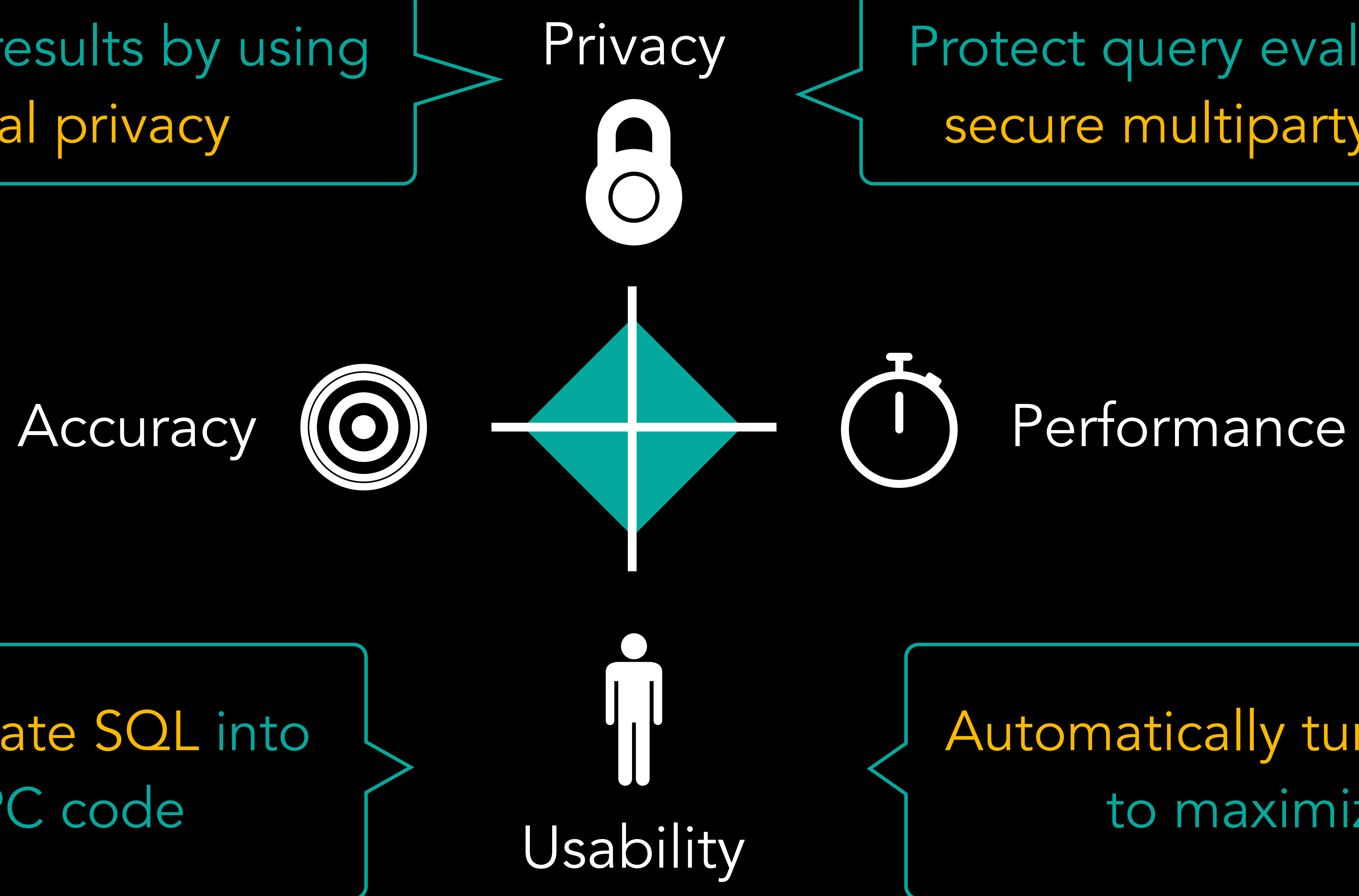
Protect sensitive patient records by using encrypted execution

Private Data Federation

Protect query results by using differential privacy

Protect query evaluation by using secure multiparty computation

Use secure multiparty computation to minimize noise



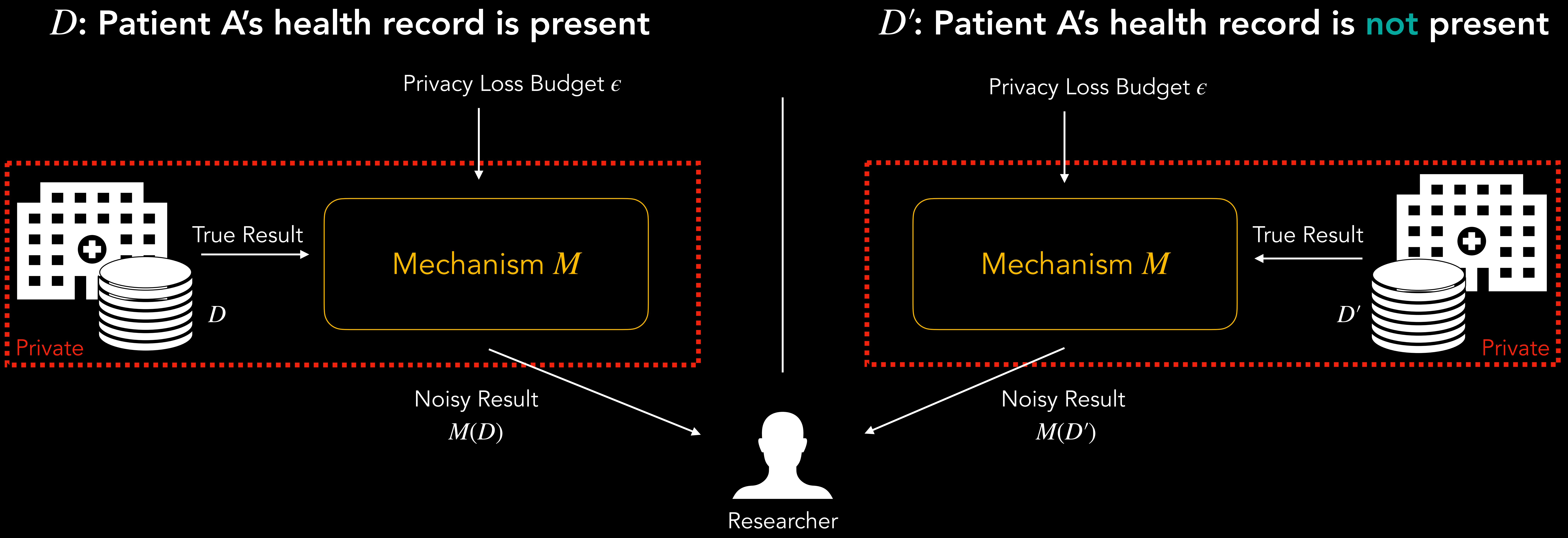
Use differential privacy to minimize computation

Automatically translate SQL into executable MPC code

Automatically tune privacy parameters to maximize performance

Private Data Federation

Differential Privacy

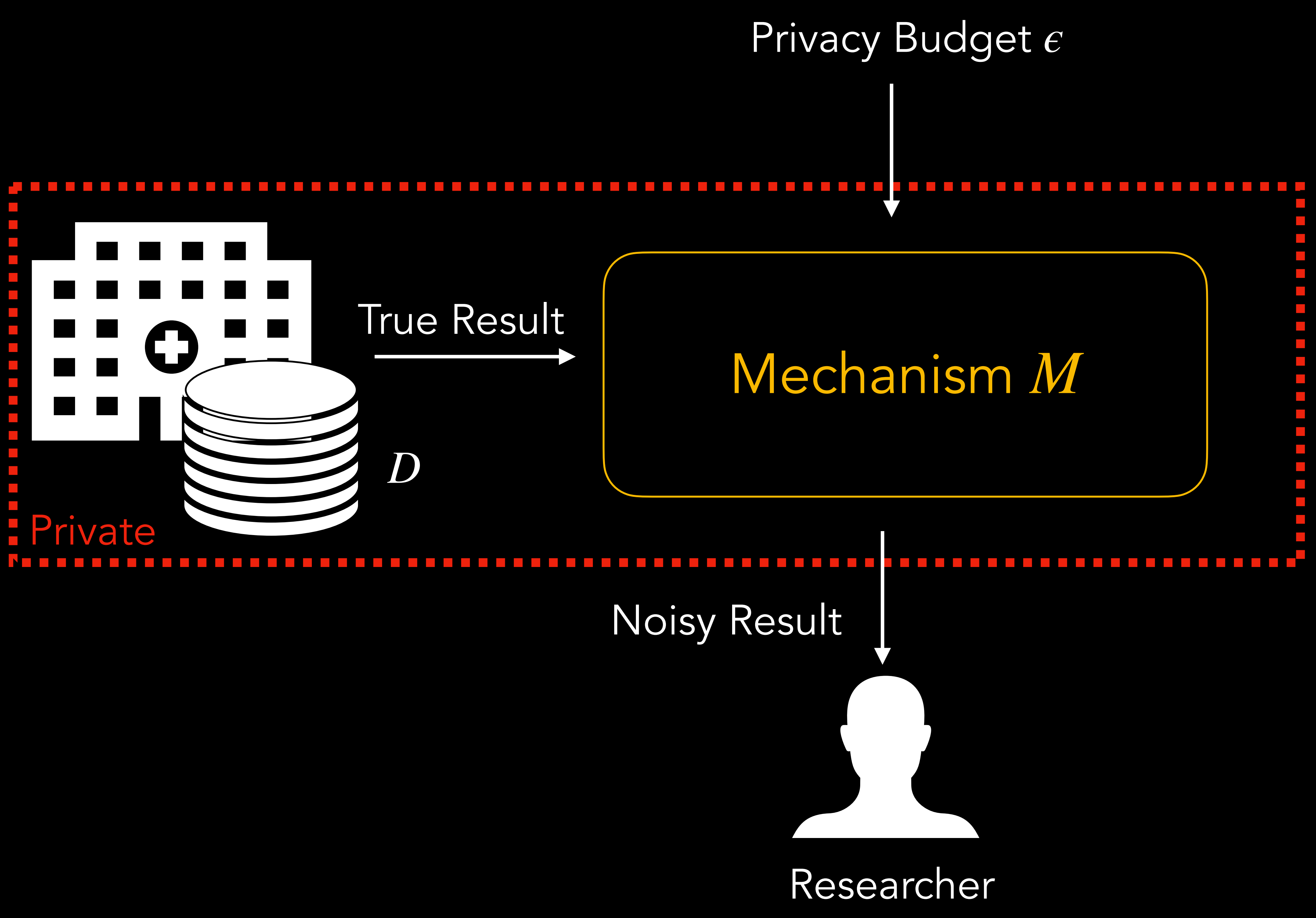


M satisfies differential privacy if for any two neighboring databases D and D'

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

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

Differential Privacy

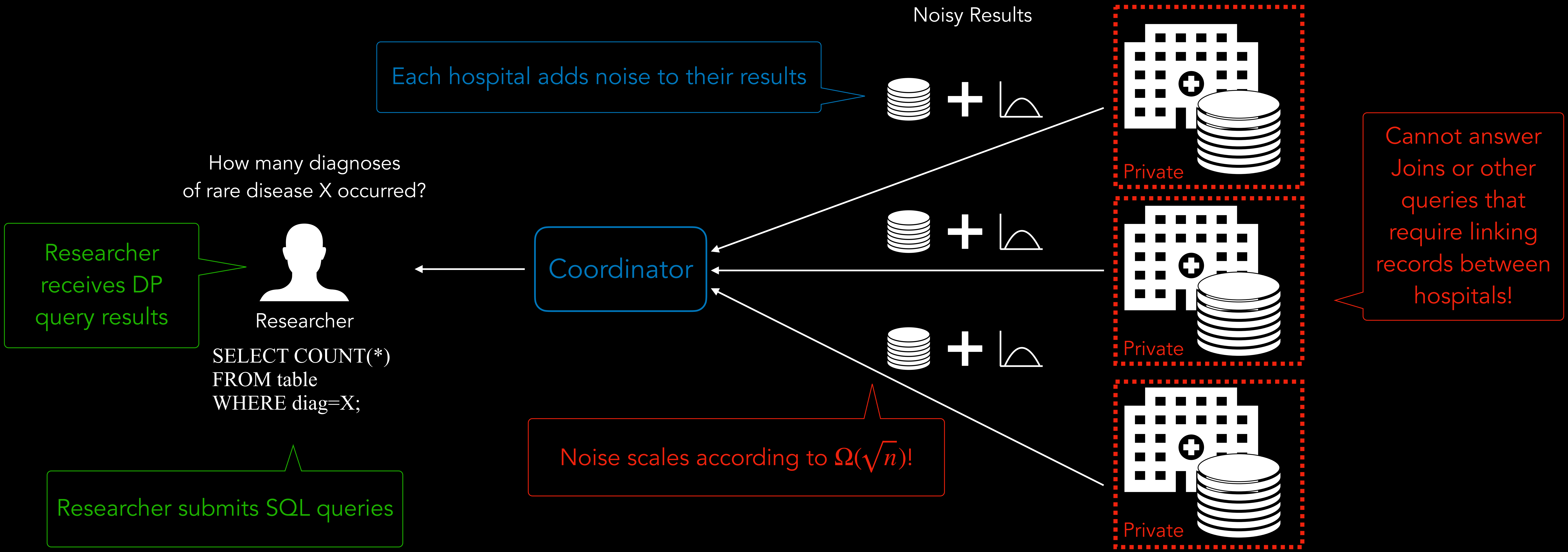


Accuracy-Privacy Trade-off
Adds noise to query results to hide contributions of individual users

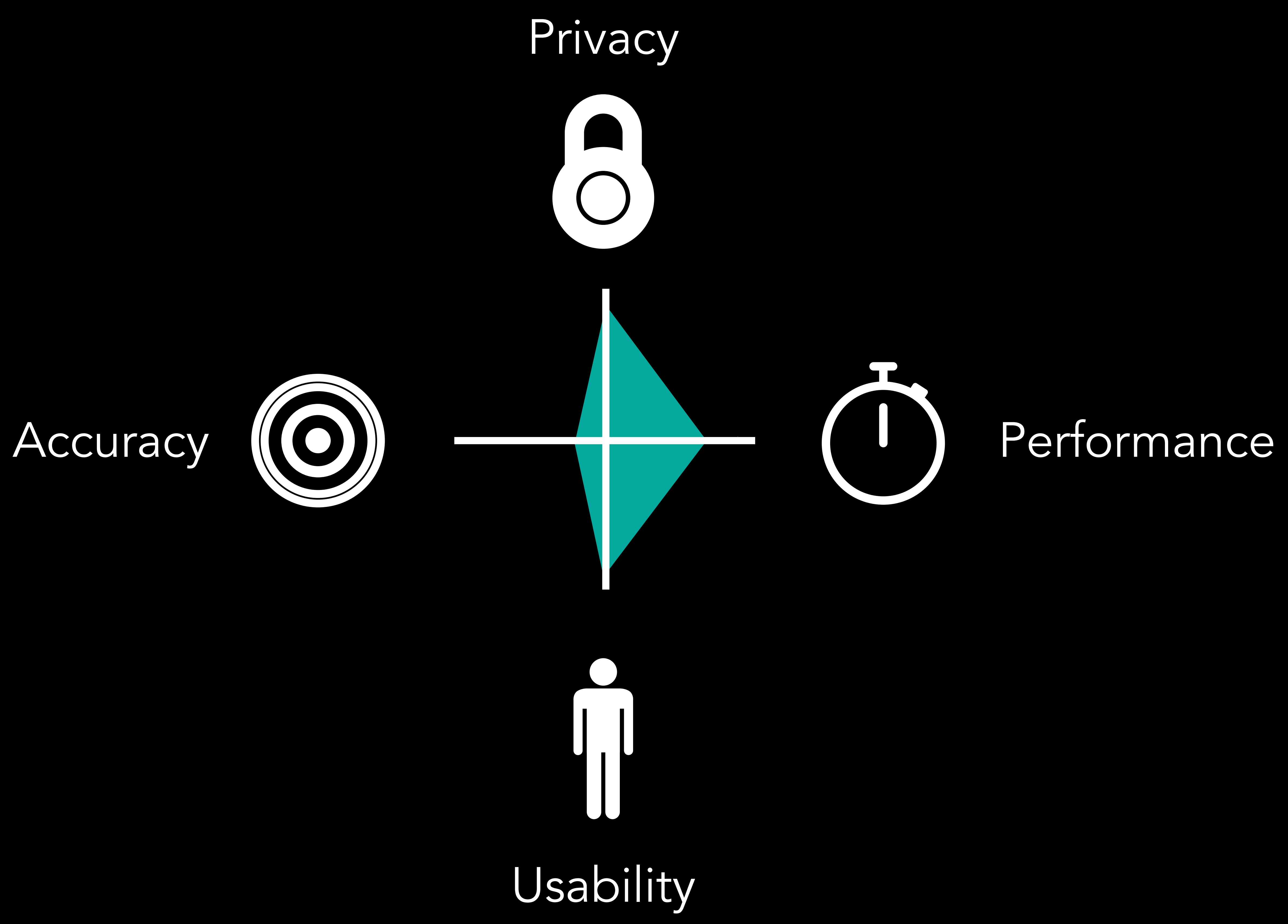
Quantifies Information Leakage
Bounds cumulative privacy loss according to a privacy loss budget

Utilized in Existing Applications
Used by organizations such as US Census, Apple, Google, etc.

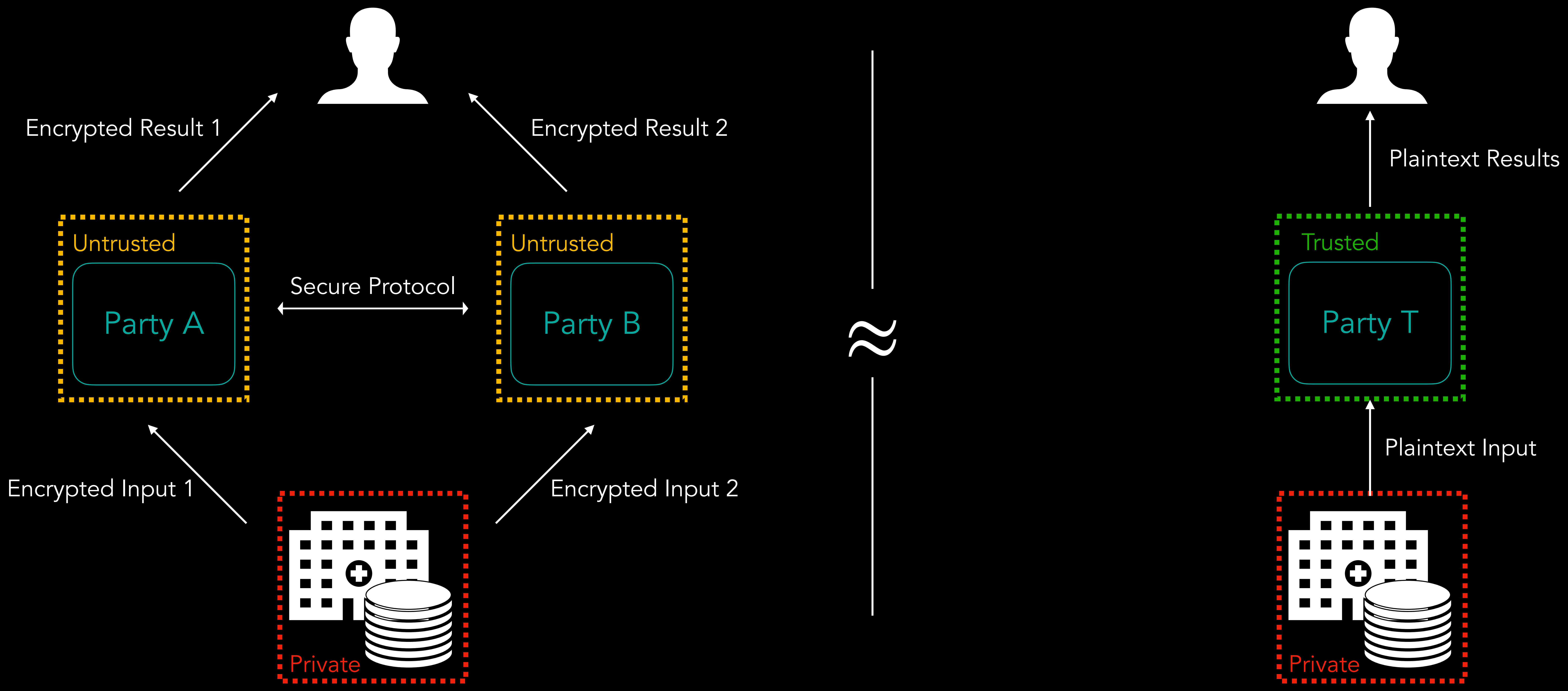
Differential Privacy



Differential Privacy

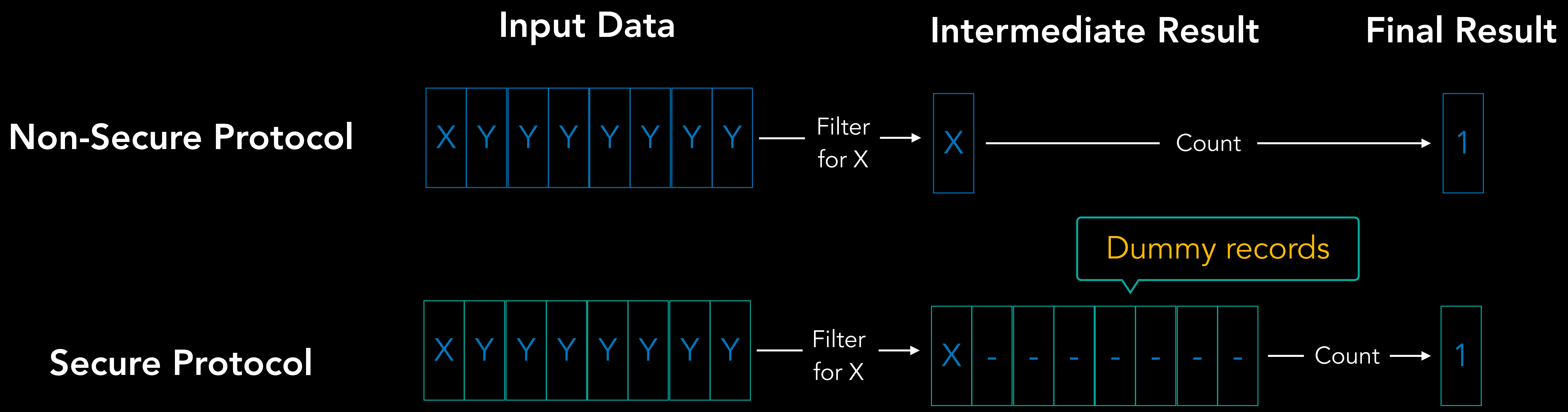


Secure Multiparty Computation



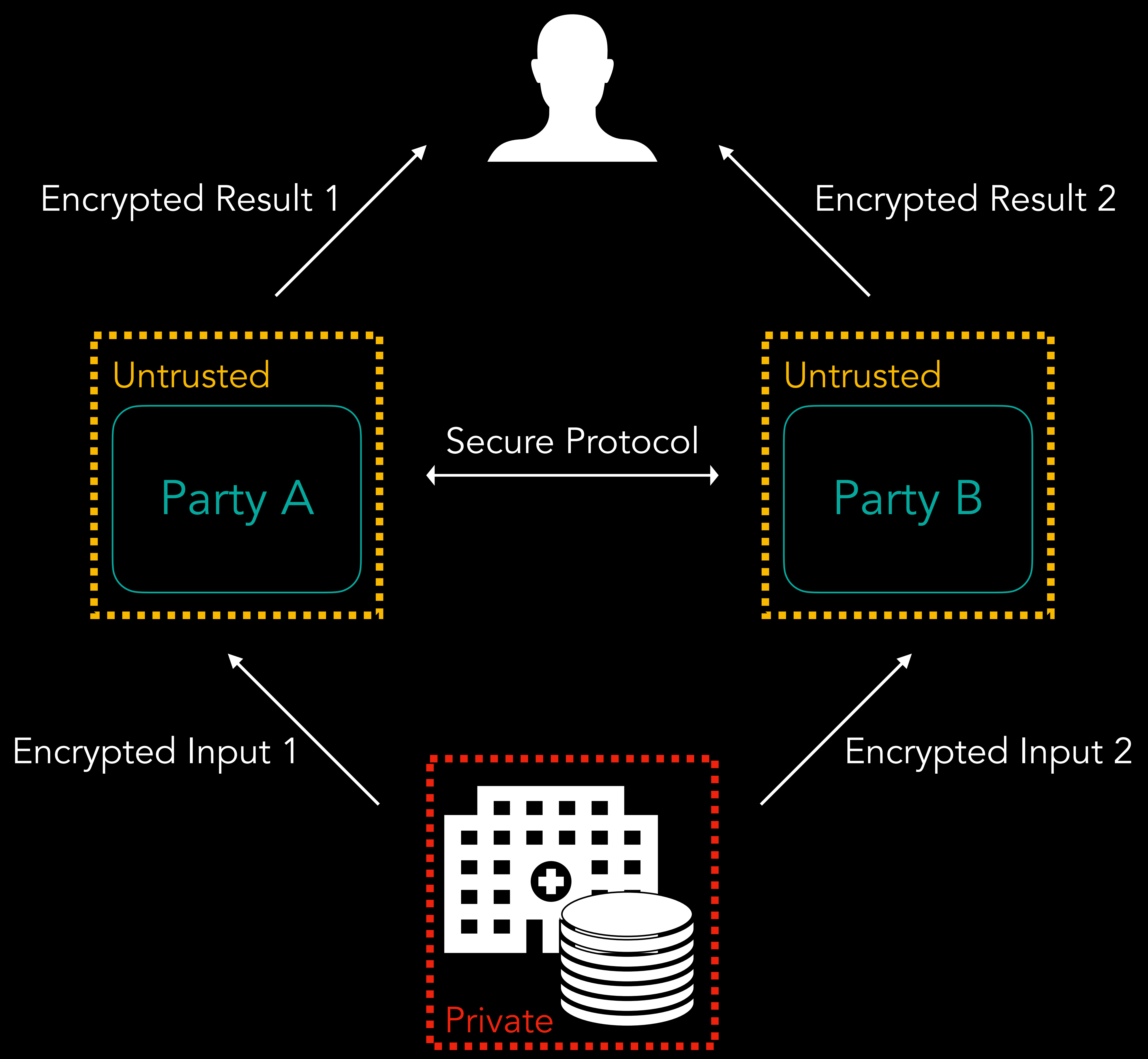
* Assumes non-collusion between parties A and B

Secure Multiparty Computation



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

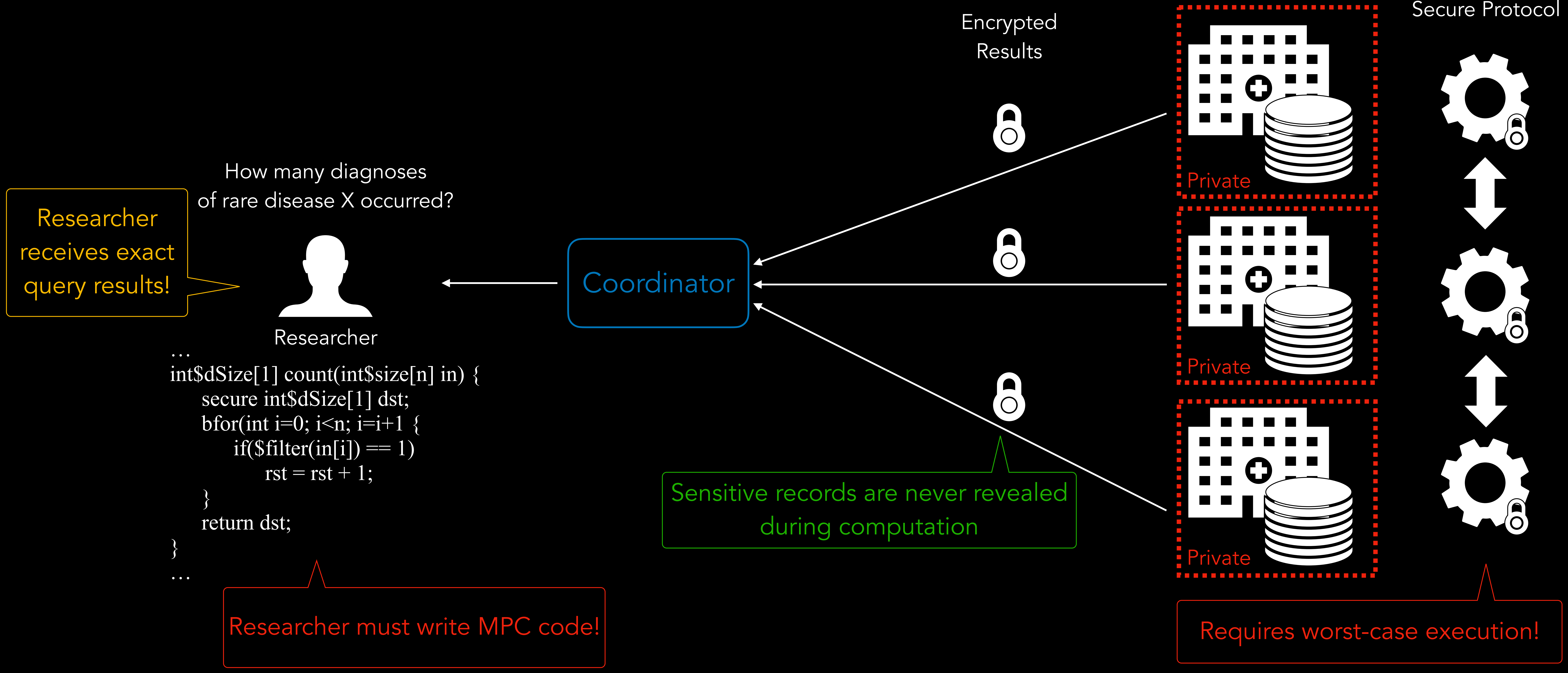
Computing parties evaluate queries without seeing records in plaintext

Exact Query Results

Final recipient reconstructs exact answer using encrypted results

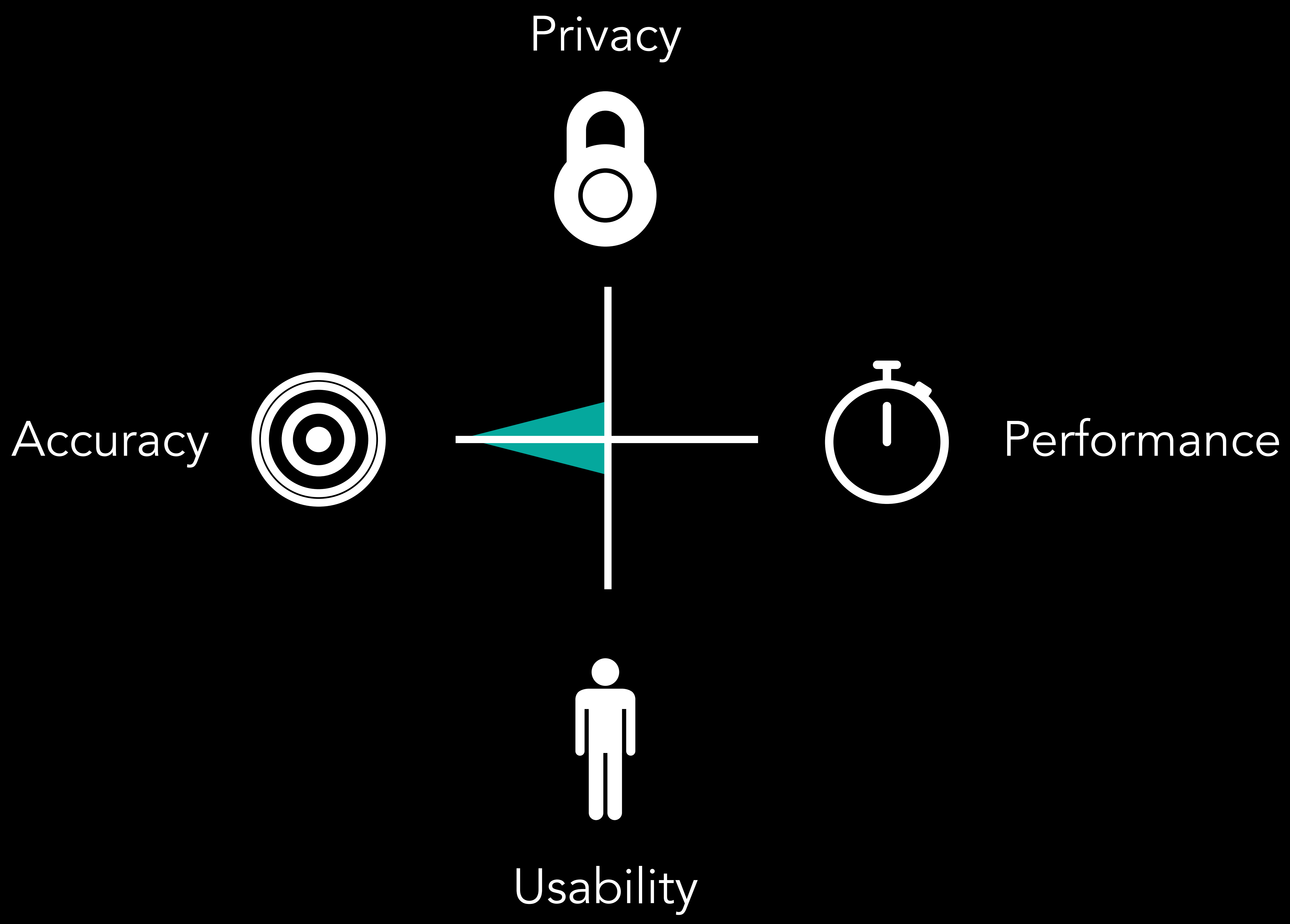
* Assumes non-collusion between parties A and B

Secure Multiparty Computation



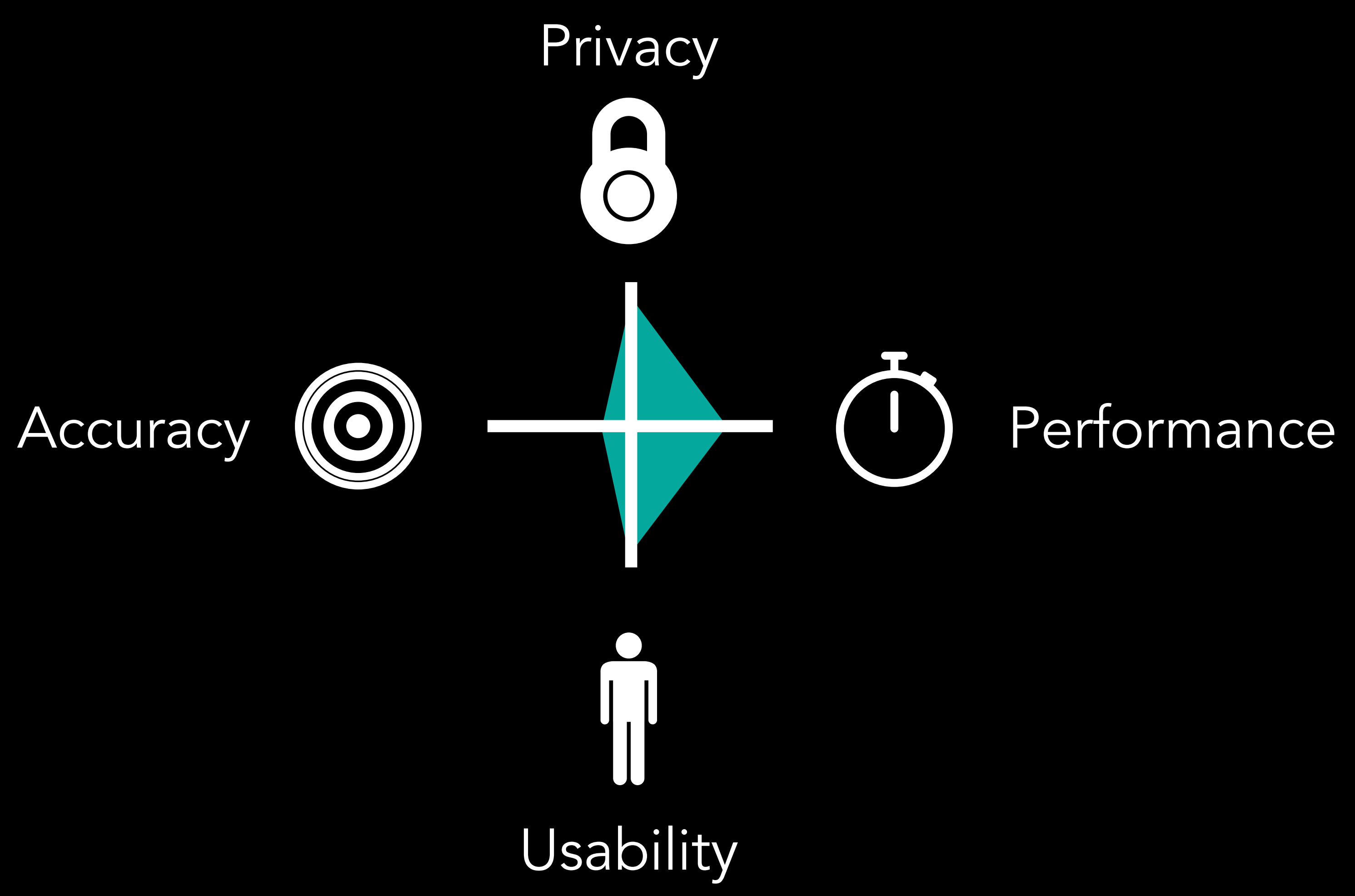

```
…
int$dSize[1] count(int$size[n] in) {
    secure int$dSize[1] dst;
    bfor(int i=0; i<n; i=i+1 {
        if($filter(in[i]) == 1)
            rst = rst + 1;
    }
    return dst;
}
…
```

Secure Multiparty Computation

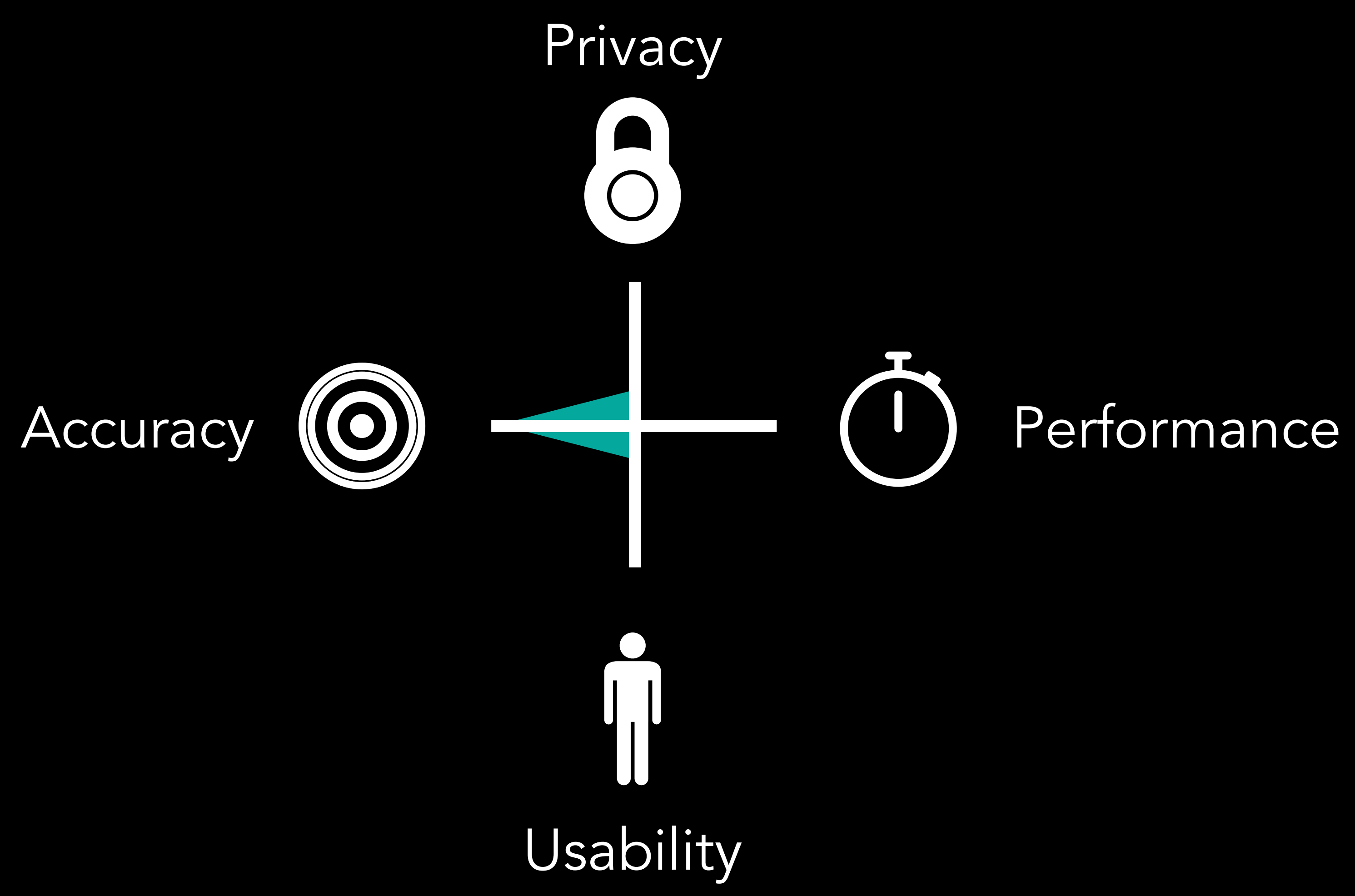


Building Blocks

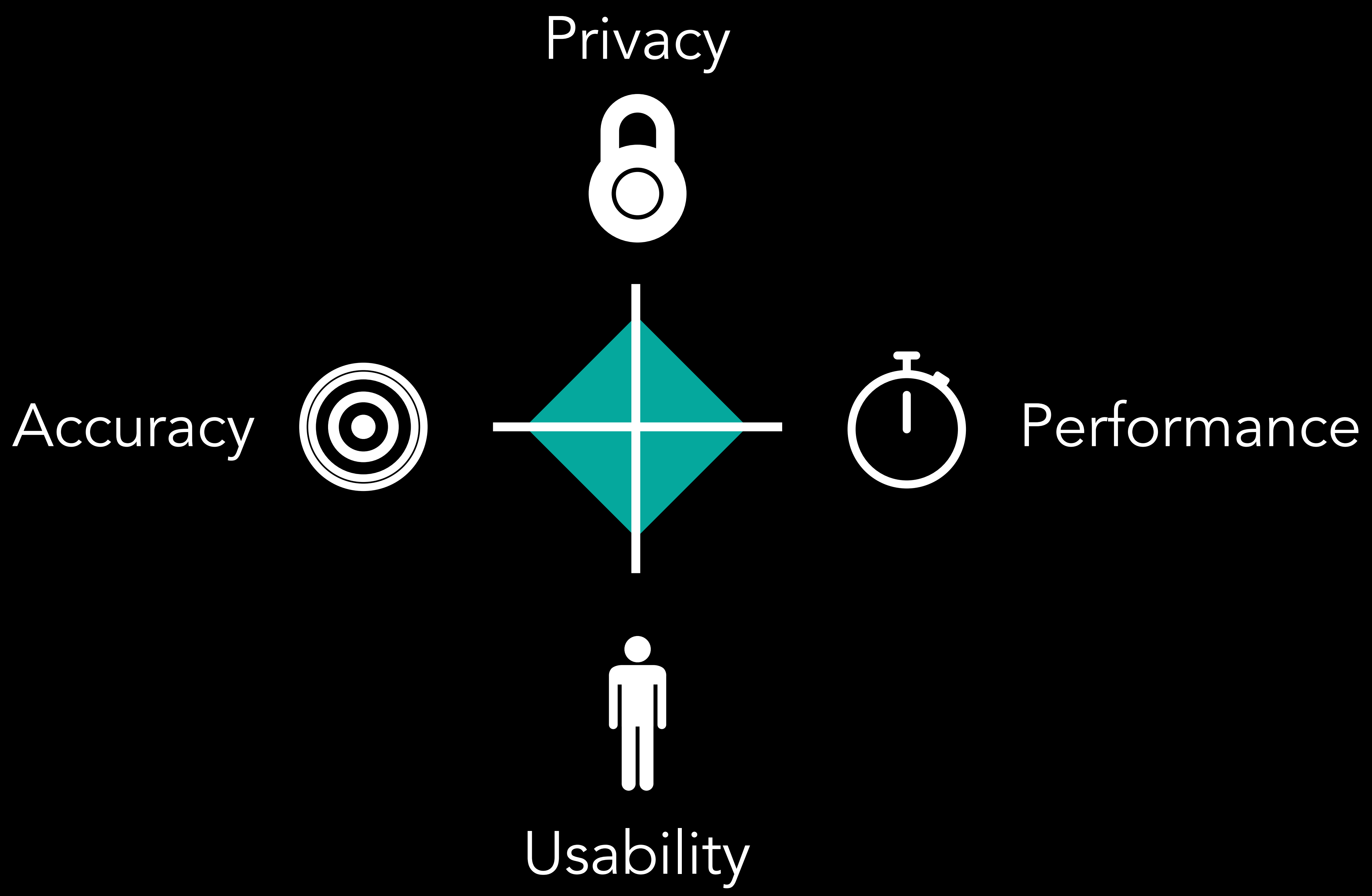
Differential Privacy



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?

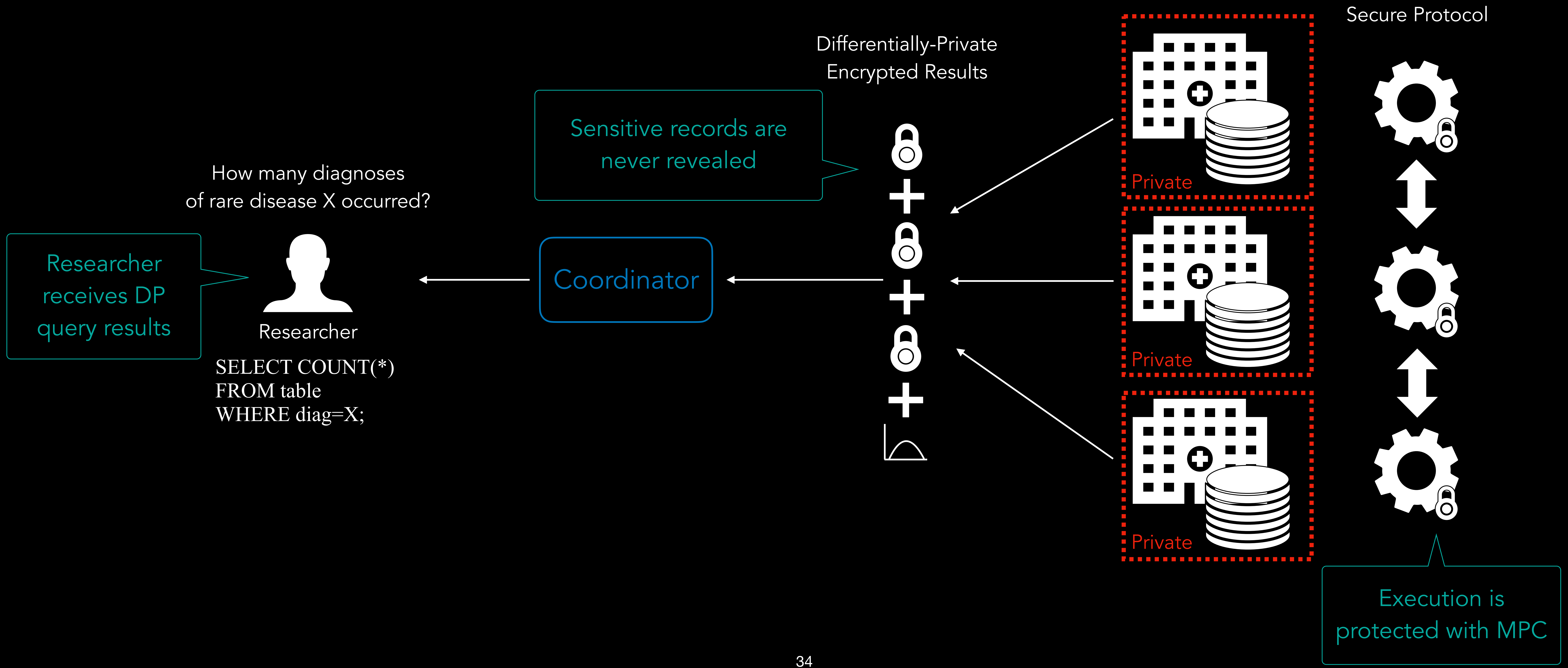
Data Computation

Can an attacker reconstruct private data by measuring computation?

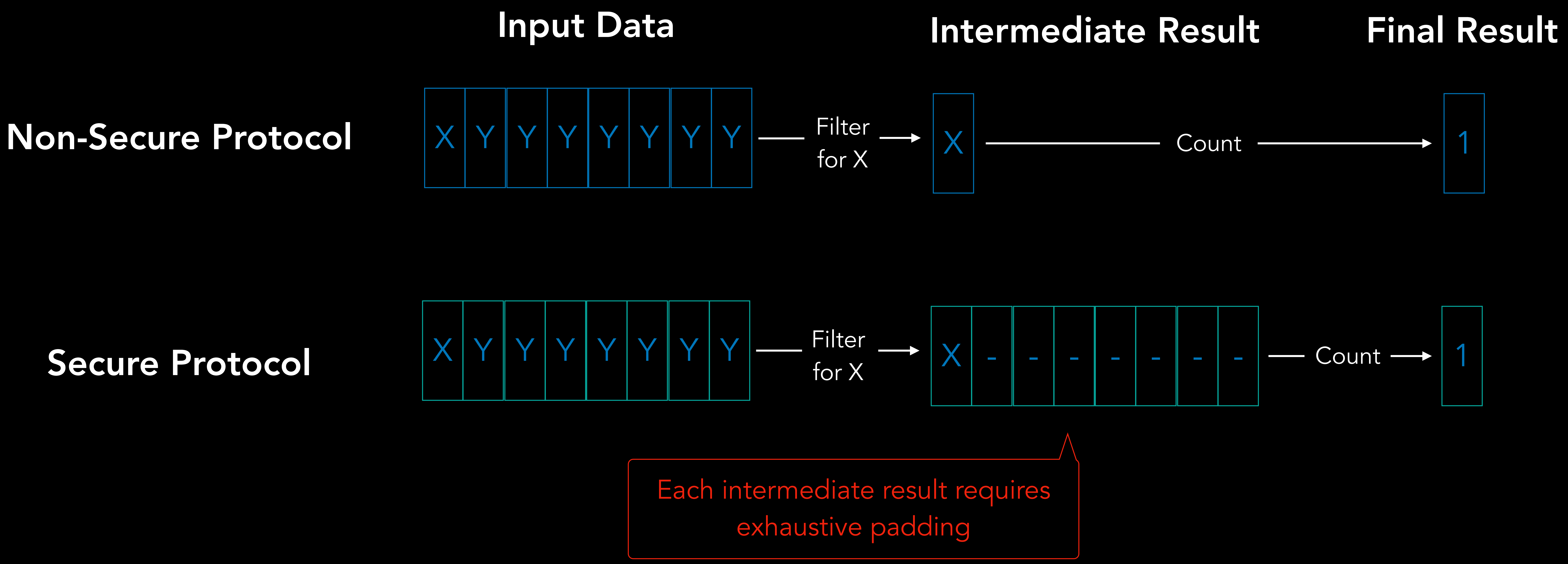
Data Release

Can an attacker reconstruct private data from published results?

Privacy Challenges

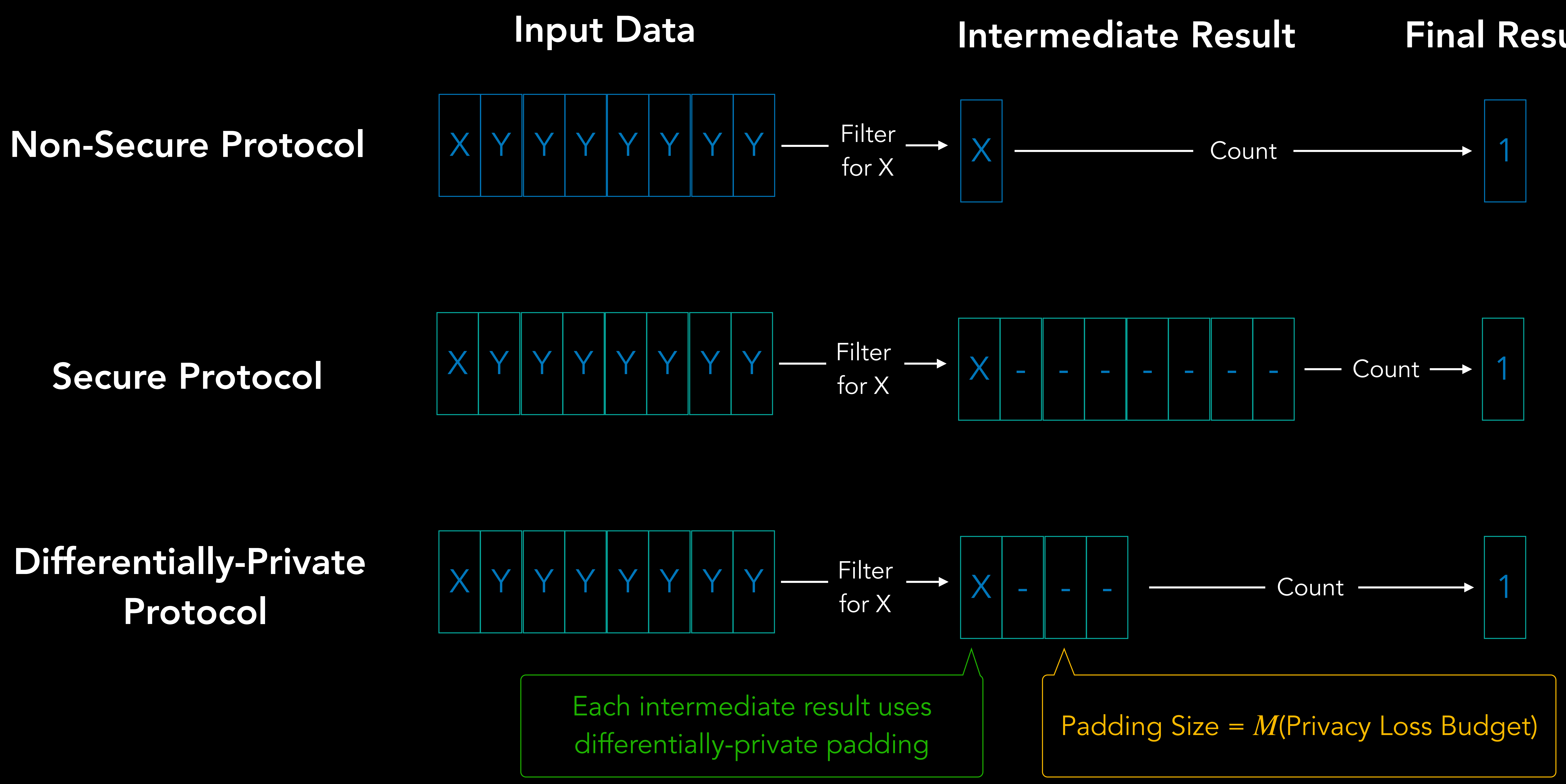


Performance Challenge



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;
}
```

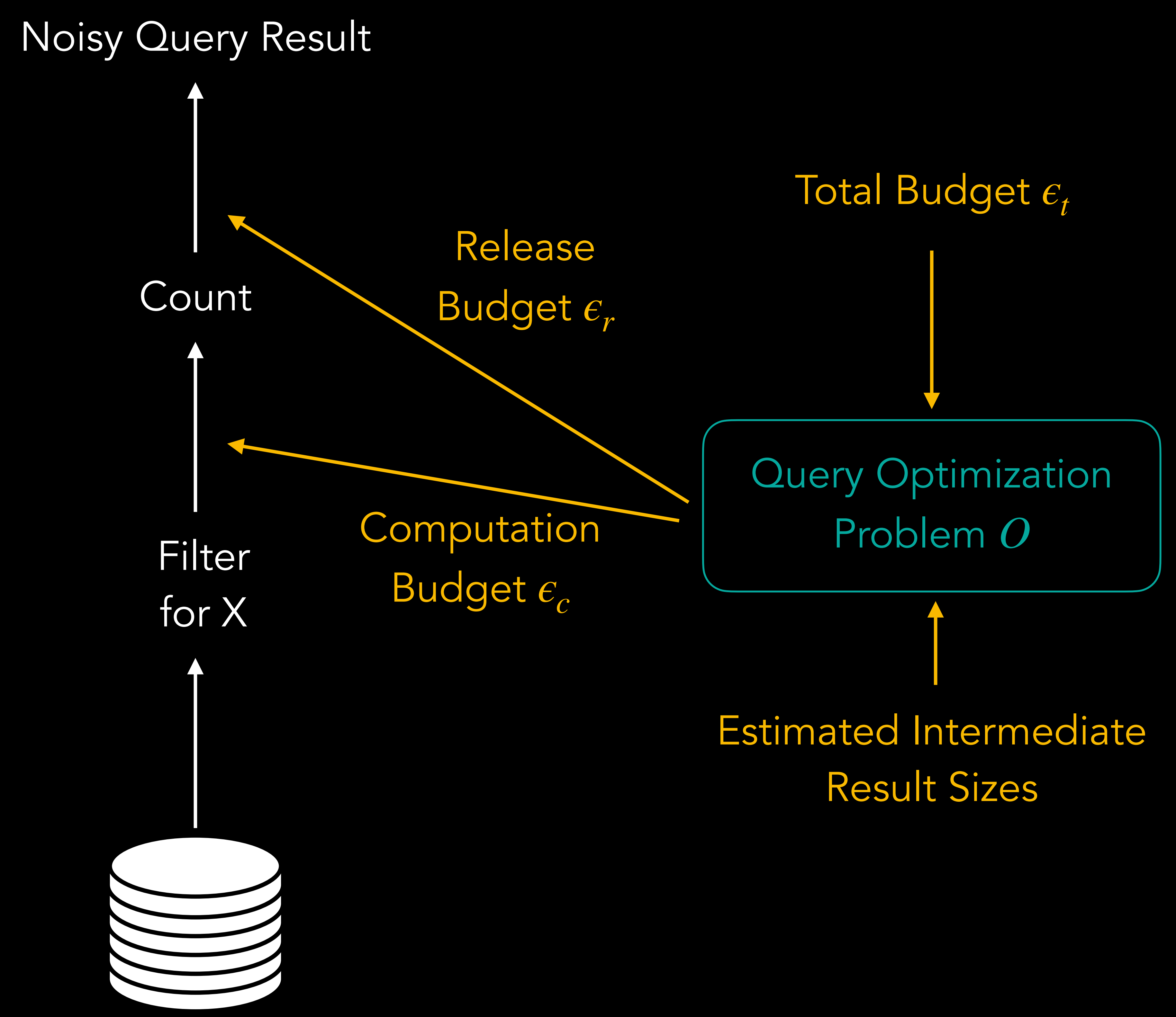
SQL to Secure Code Translation

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



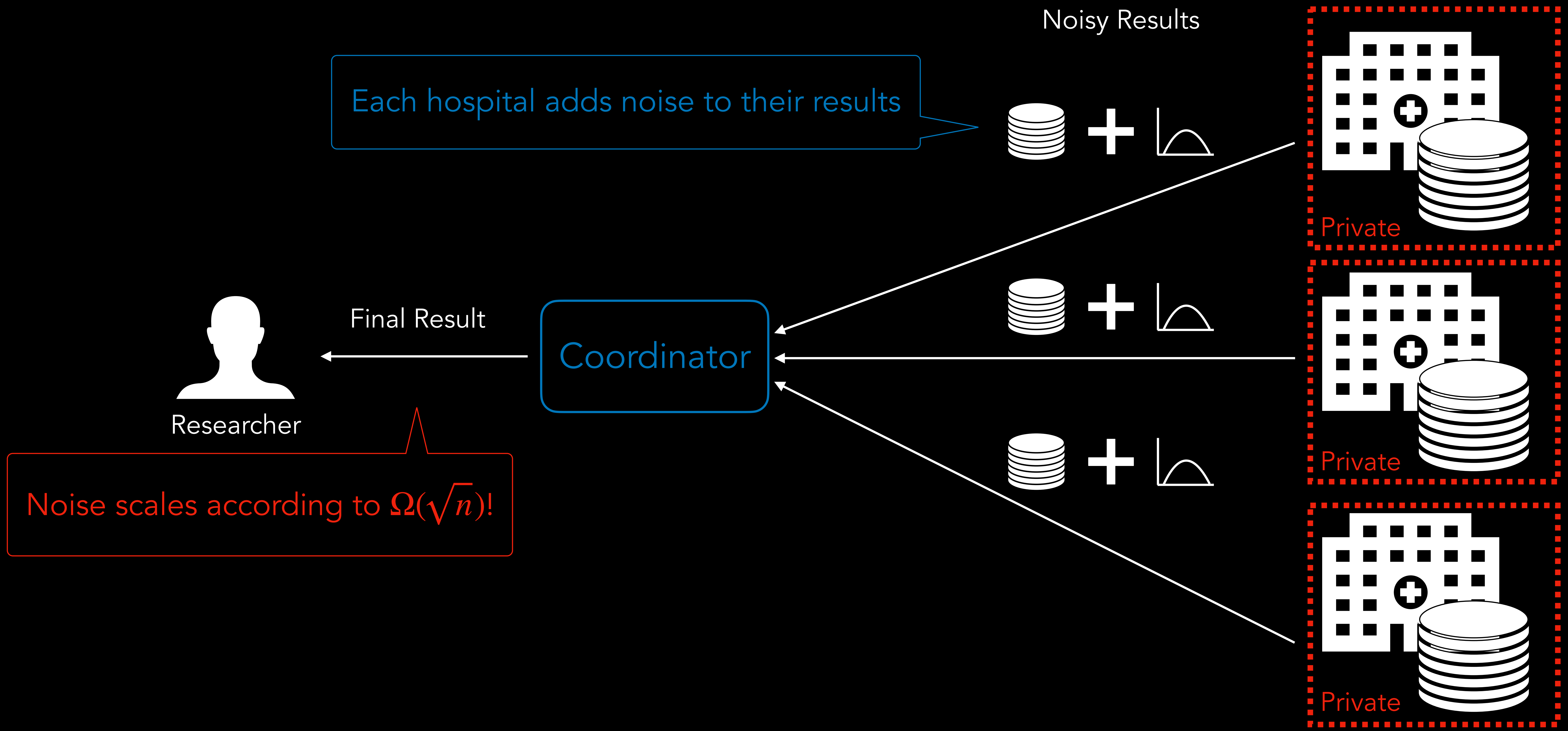
SQL to Secure Code Translation

Automatically converts SQL to secure code at codegen and runtime

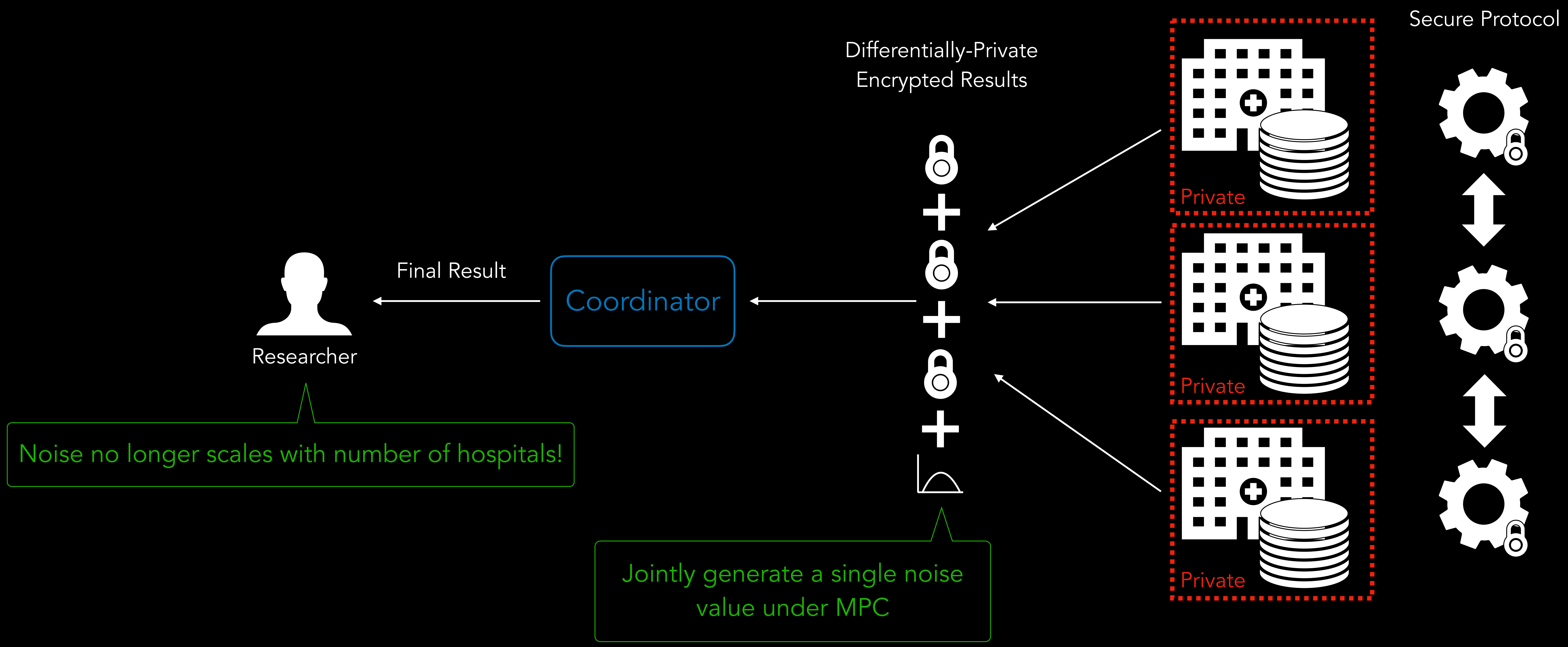
Privacy Budget Allocation

Optimal allocation of a privacy loss budget without user intervention

Accuracy Challenge



Accuracy Challenge



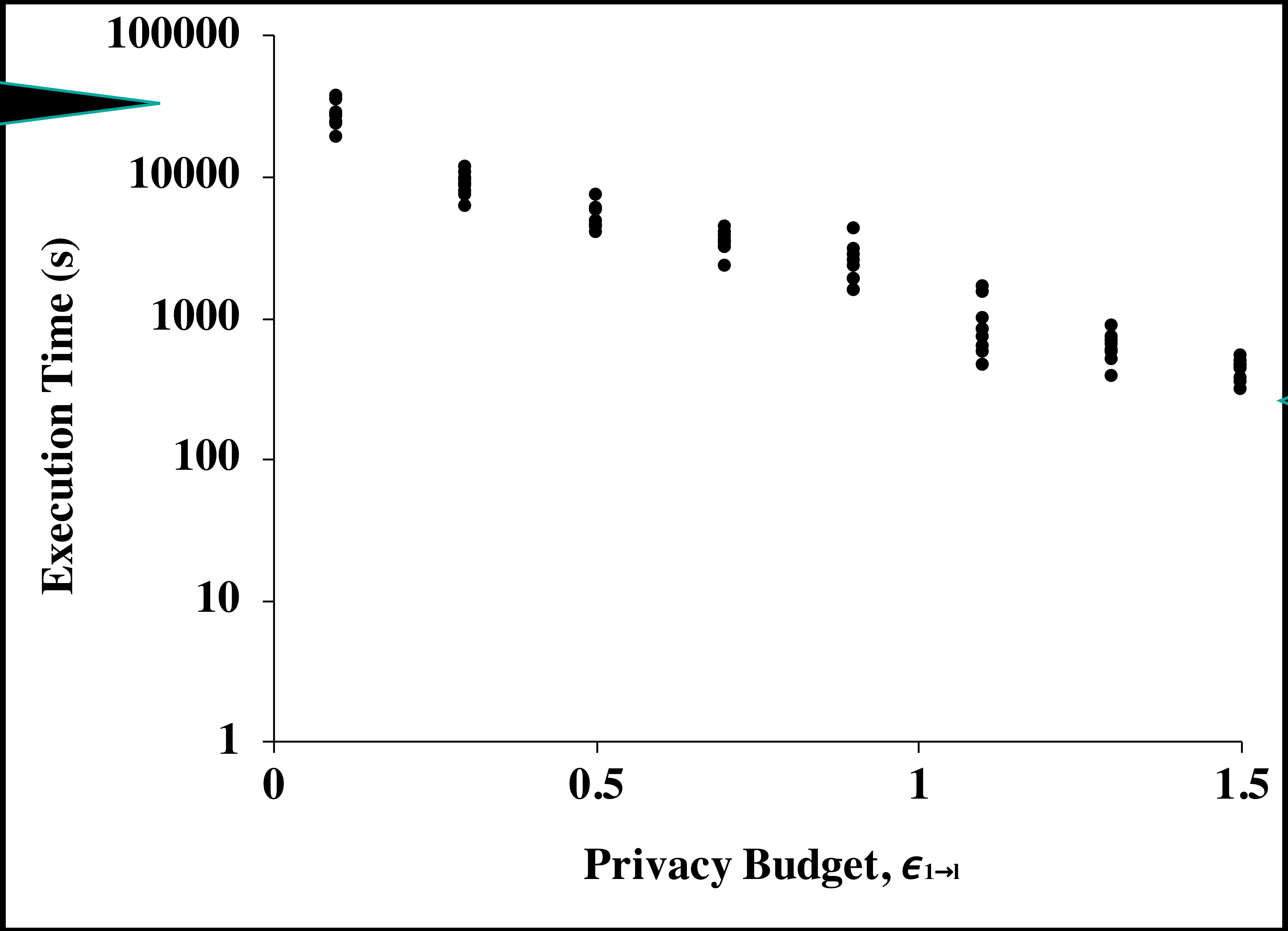
Private Data Federation

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 HealthLNK medical researchers

Privacy-Performance Trade-off

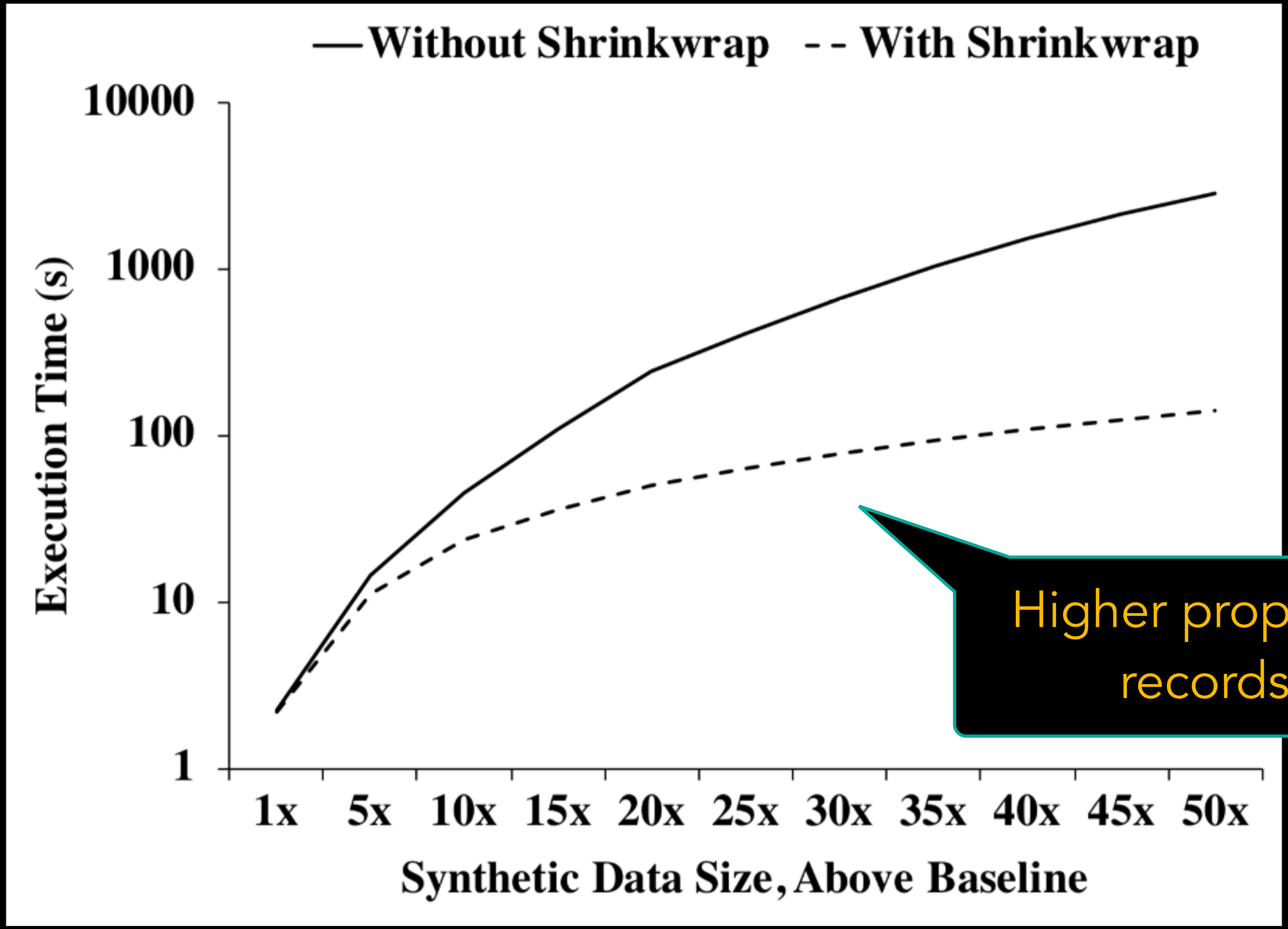
~15 hours without optimization



~15 minutes with optimization

Lower Privacy, Higher Performance

Scaling with Data Size



Baseline = 15 GB, ε = 0.5, δ = 1 x 10^{-5},

More Data, More Speed Up!

Higher proportion of dummy records in the input

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

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

Privacy-Preserving Systems:

Balancing Privacy and Utility for Query Execution

Johes Bater

Tufts University | School of Engineering

My Research

Private Data Federations

Efficient SQL Queries for Private Data Federations

SMCQL (VLDB '17)

Shrinkwrap (VLDB '18)

Privacy-Preserving Approximate Query Processing

SAQE (VLDB '19)

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

Ensure end-to-end protection of sensitive data

Minimize user intervention to simplify system usage

Optimize utility while preserving privacy

Enable expert configuration by non-experts