Privacy-Preserving Medical Data Sharing

SAIL Team

Motivation

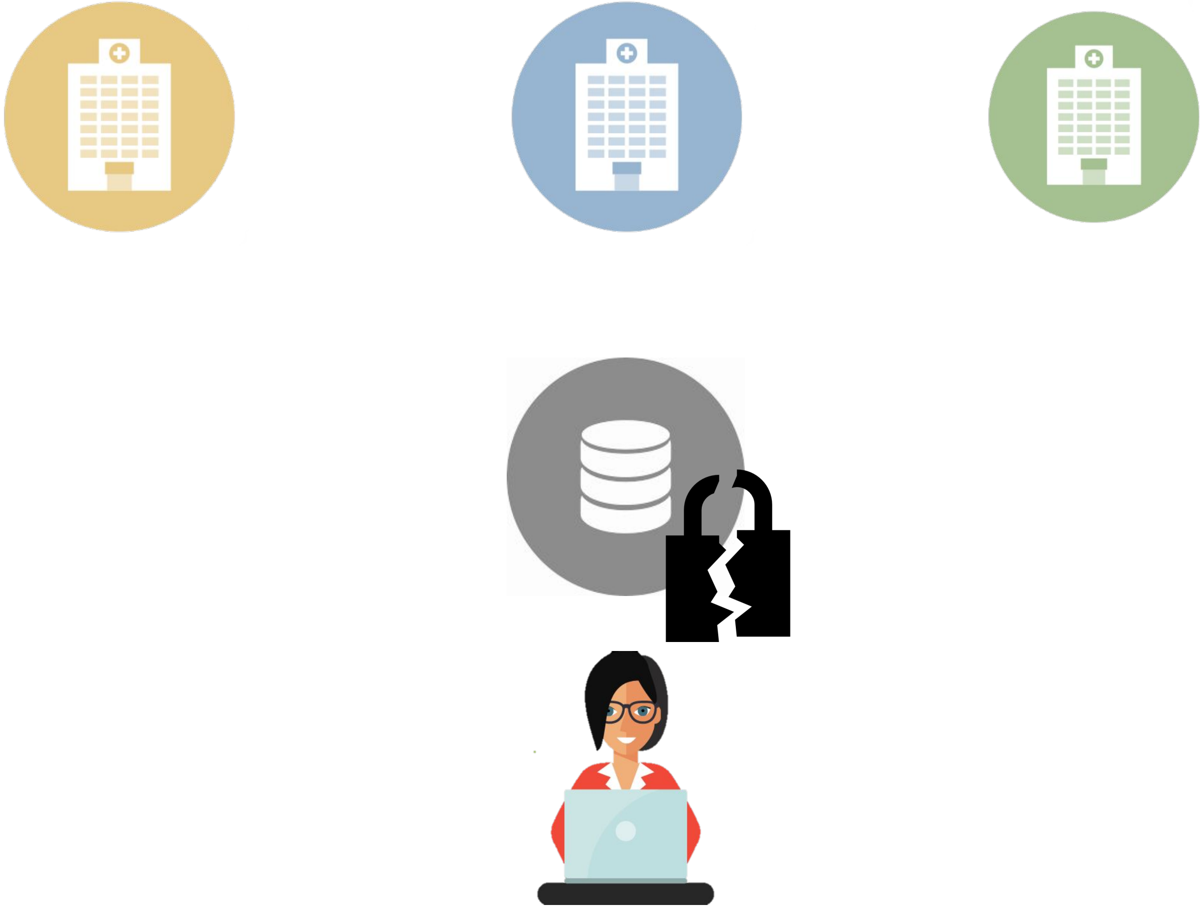
Current de-identification methods do not offer enough privacy protection for small populations, such as those in clinical trials for rare diseases. When a population is small, even when “personal” data is removed, an individual may still be identifiable.

Setting

1. Data providers: hospitals, clinics, etc.

2. Data sharing platforms: Vivli

3. Analysts: researchers, companies, etc.



Concrete Problems

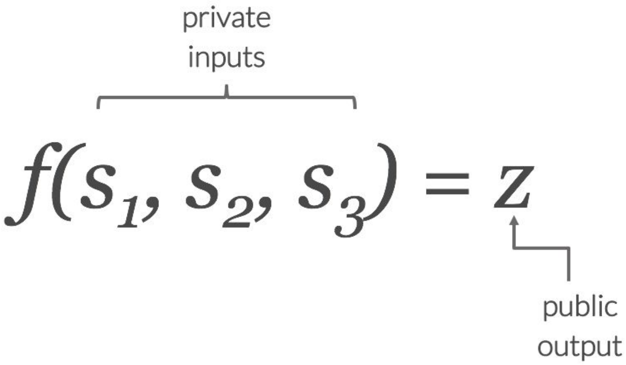
1. De-anonymizing or randomizing data set at the input provider harms utility.
    a. May make linking between different datasets impossible.
    b. Randomization is amplified when the analytic is run.

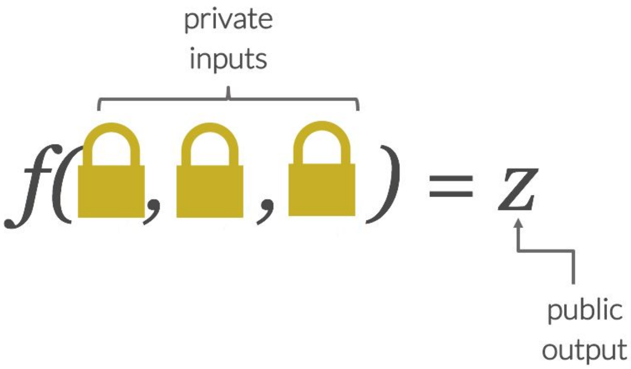
2. However, small datasets are highly privacy sensitive, and cannot be provided to platforms in the clear.

3. Output of analytics run on small datasets can leak a lot of information.

Cryptographic Primitives

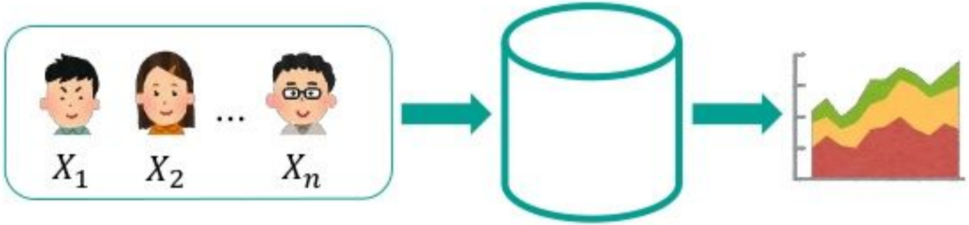
- Secure Multi-Party Computation:
  - Computes a value using private inputs *without revealing* those private inputs, and then reveals a public output
    - Ex. Calculating the average age of everyone in the room without revealing our ages to one another.





Cryptographic Primitives

- Differential Privacy:
  - Given two nearly identical datasets, where one has an individual's data and one does not, add noise to the data so that the same statistical query on each dataset will produce a similar result.
    - Ex. After calculating the average age in the room, someone new walks in, we then calculate the average age again, and we get a similar result.

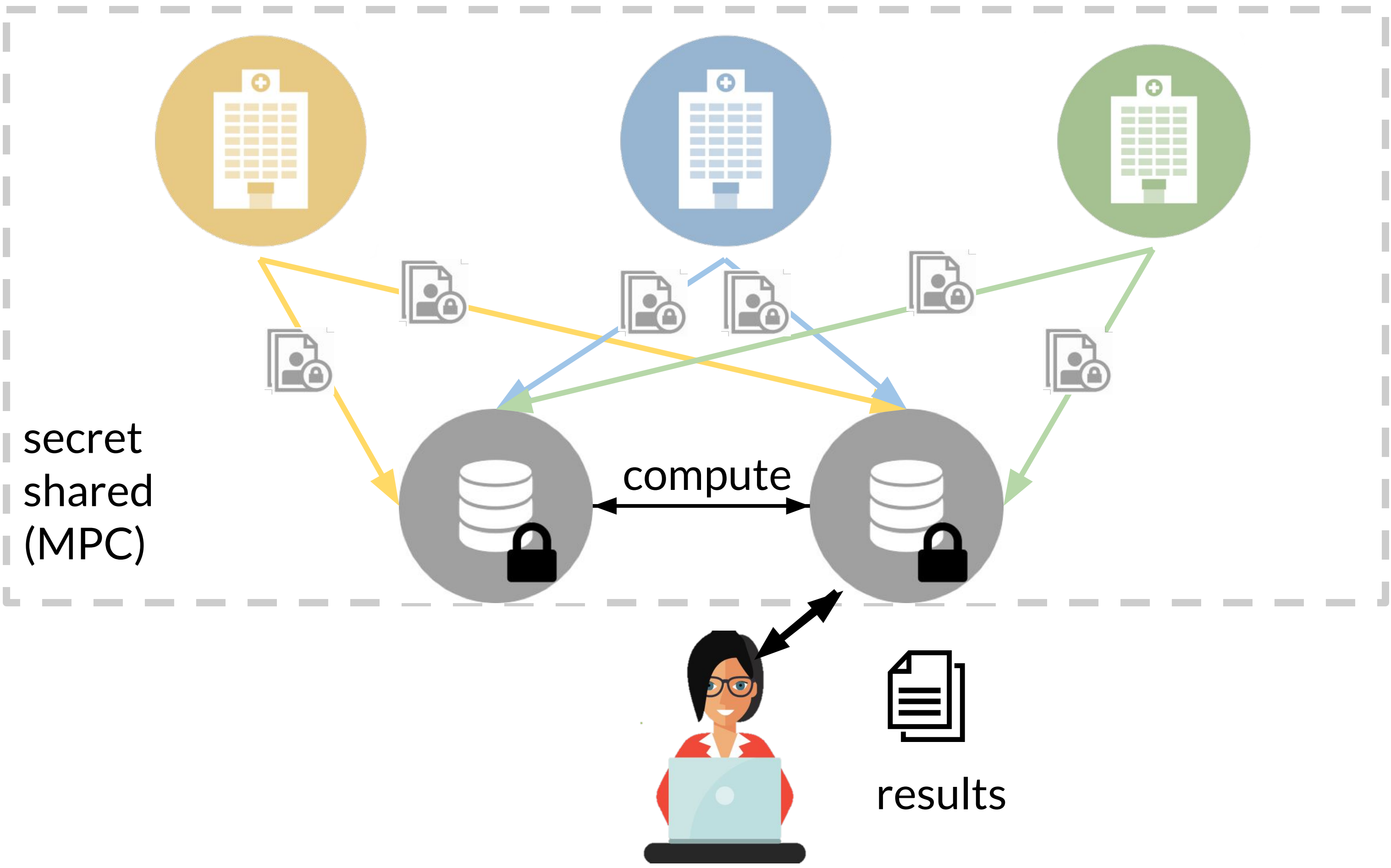


Solution

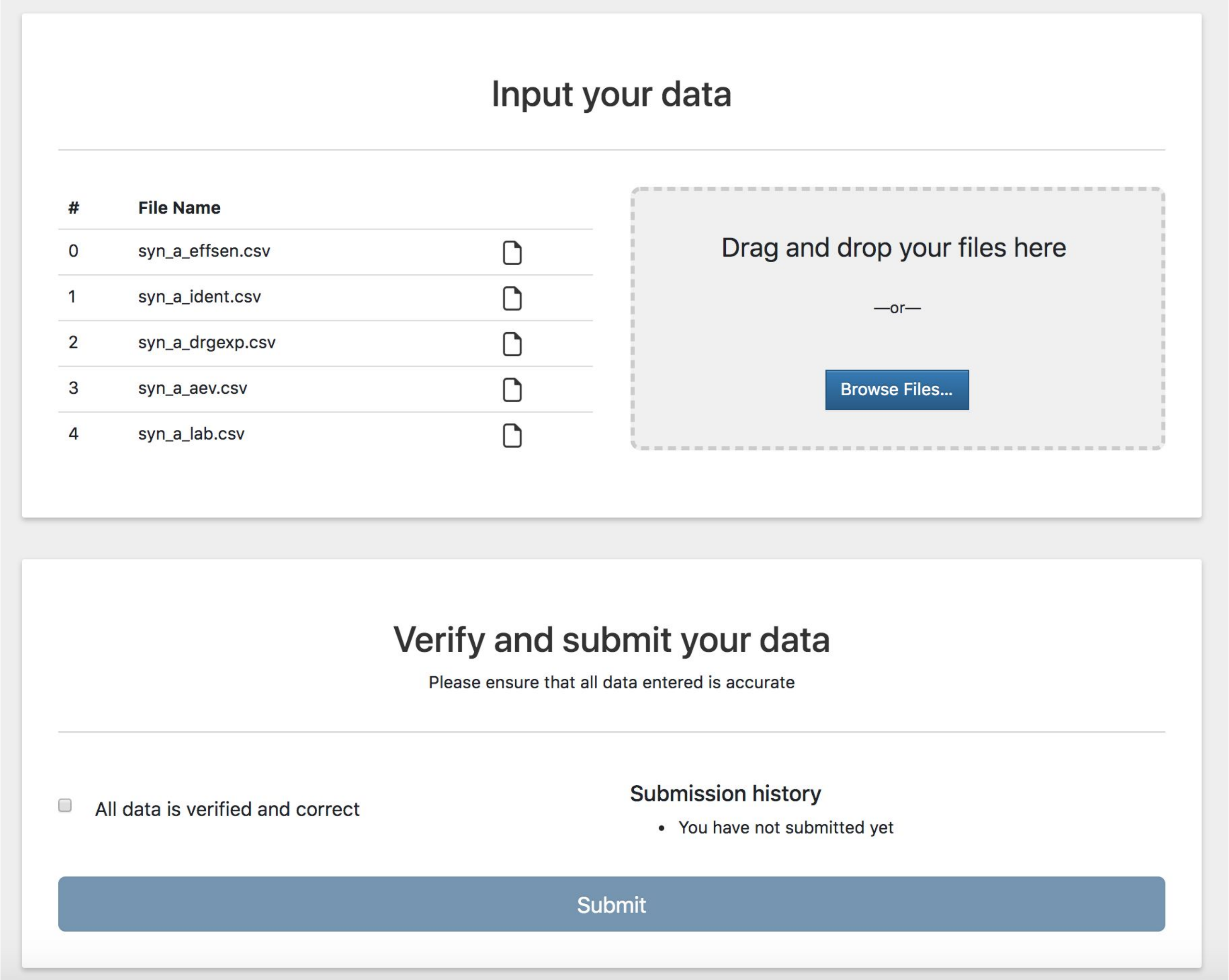
We propose a combination of secure multiparty computing and differential privacy protocols, served through a web-based interface, to protect private data.

Our service follows a basic model:

1) Data providers secret share their data with a federated data sharing platform.
2) Analysts provide their desired analytics to the federated data sharing platform.
3) The platform compute the analytics using MPC, may randomize the output with DP, and reveals the answers to the analyst.


secret
shared
(MPC)
compute
results

Screenshots from Demo



Input your data

#	File Name	
0	syn_a_effsen.csv	🗋
1	syn_a_ident.csv	🗋
2	syn_a_drgexp.csv	🗋
3	syn_a_aev.csv	🗋
4	syn_a_lab.csv	🗋

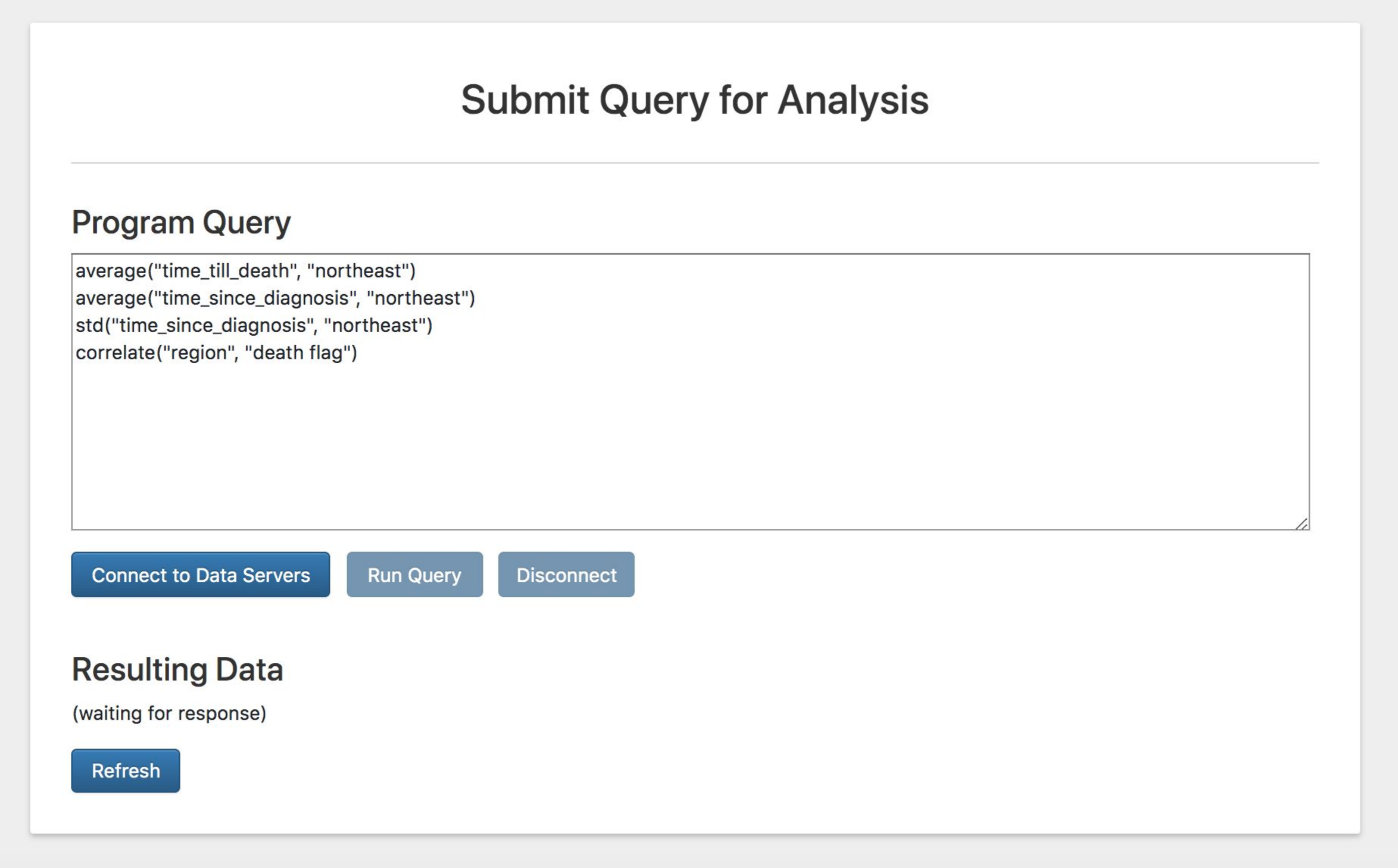
Verify and submit your data

Please ensure that all data entered is accurate

□ All data is verified and correct

Submission history

- You have not submitted yet



Submit Query for Analysis

Program Query

```
average("time_till_death", "northeast")
average("time_since_diagnosis", "northeast")
std("time_since_diagnosis", "northeast")
correlate("region", "death flag")
```

Resulting Data

(waiting for response)

Open-Source Code:
github.com/multiparty/datathon