Secure Computation on Datasets

Steve Lu
Stealth Software Technologies, Inc.

Rafail Ostrovsky
UCLA

Special Topics on Privacy and Public Auditability Event #2
NIST
April 19, 2021

The Paradox of Privacy

- The value of data is in using it
- Data is often private
- Can parties compute on joint data toward a common goal while maintaining privacy?

Variety of Examples

- Passenger Manifest vs No-Fly list

- Reproducibility of scientific experiments on private data
- Genetic analysis without revealing the algorithm or my genes

- Health care providers sharing patient data
- Longitudinal studies on social, economic, educational data

Variety of Examples

- Logistics
- Fraud Detection
- Private Machine Learning

- This leads to a general question…

Can we compute on *private* datasets?

Outline

- MPC Review
- New Research
- Looking Ahead
- Conclusion

Outline

- MPC Review
- New Research
- Looking Ahead
- Conclusion

Features of MPC

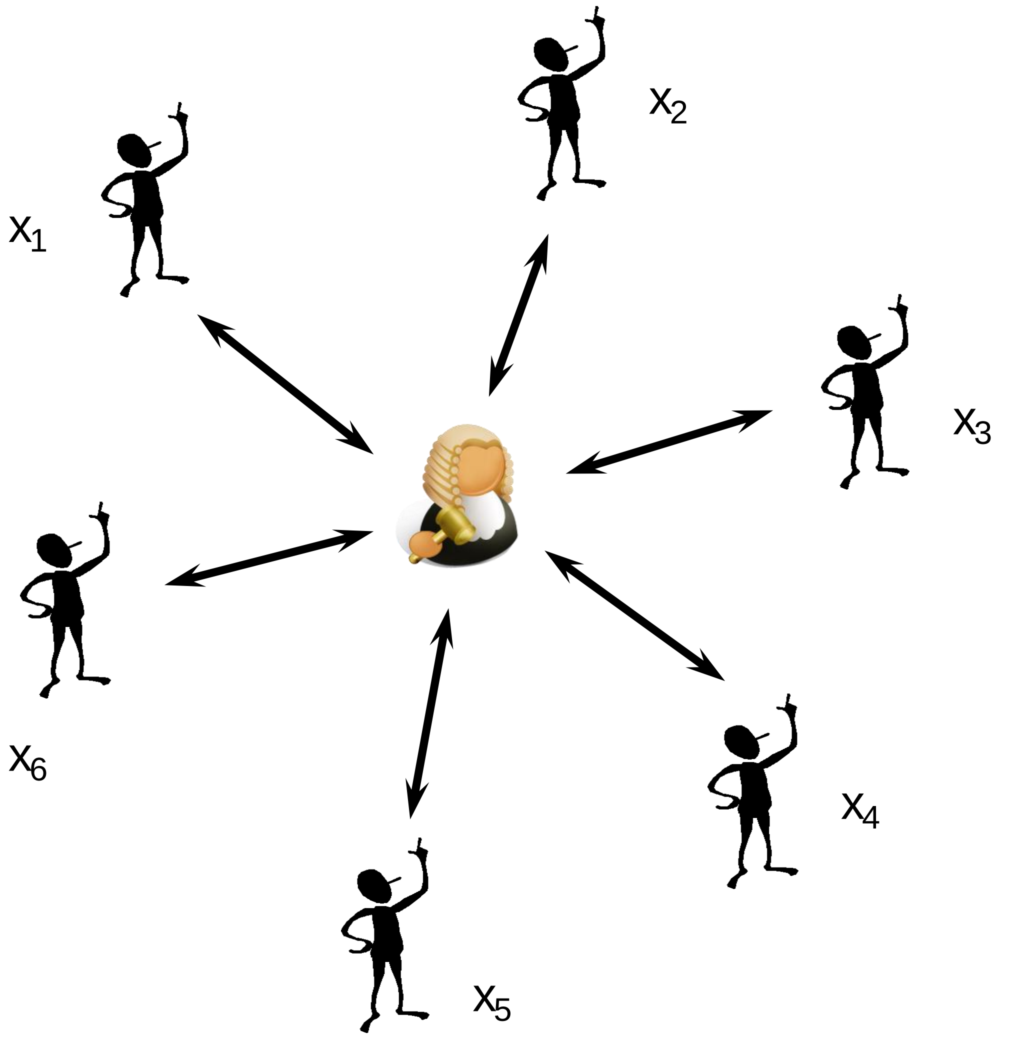
- Protocol for parties to interact to obtain only the output of prescribed function
- Doesn't release a "fuzzed" corpus of data
- Highly controlled release
- Virtual enclave
- Replaces honest broker

What MPC *Isn't*

- No fuzzing or noise
  - Outputs are exact
- Distinct from other privacy mechanisms
  - k-anonymity (Sweeney 2002)
  - Differential Privacy (Dwork 2006)
  - These technologies can be combined with MPC
- Output Inference and Probing Inputs

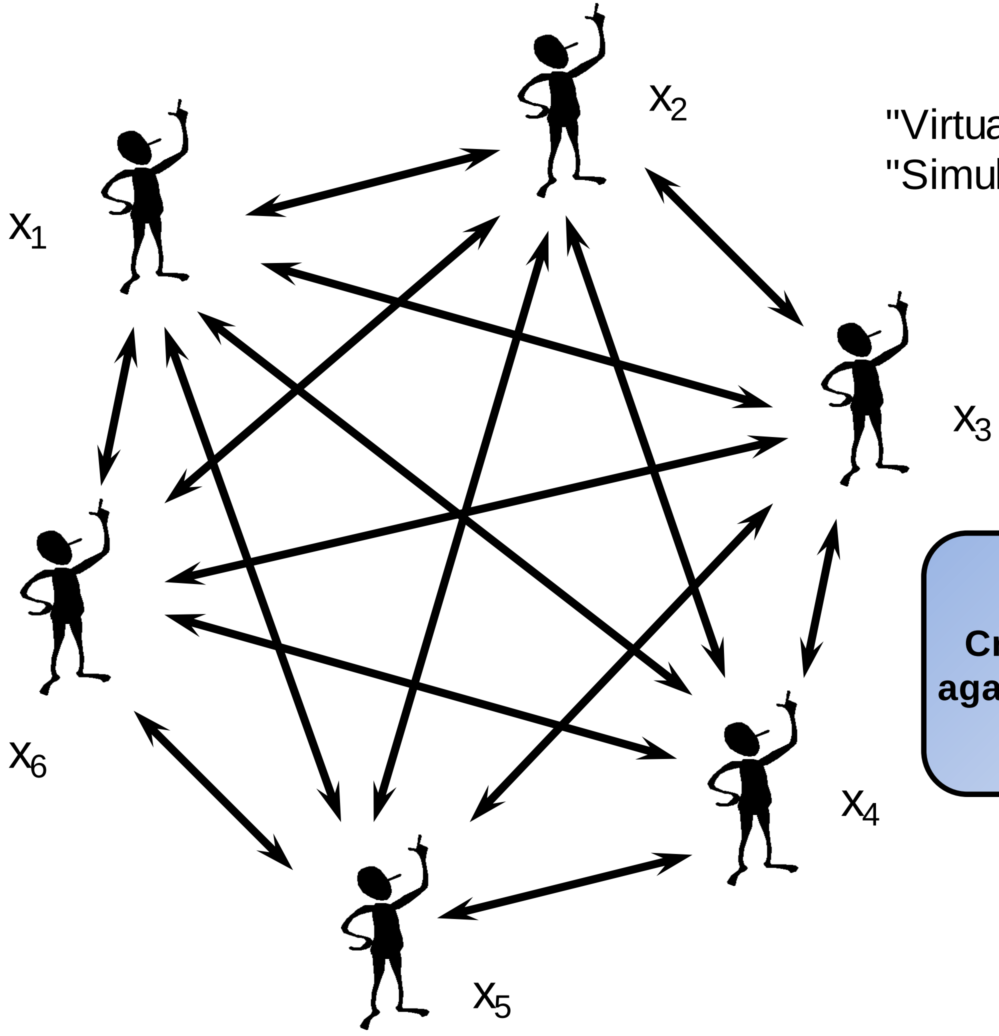
Secure Multiparty Computation (MPC)

We want to compute $F(x_1,...,x_6)$



Secure Multiparty Computation (MPC)

We want to compute $F(x_1,...,x_6)$



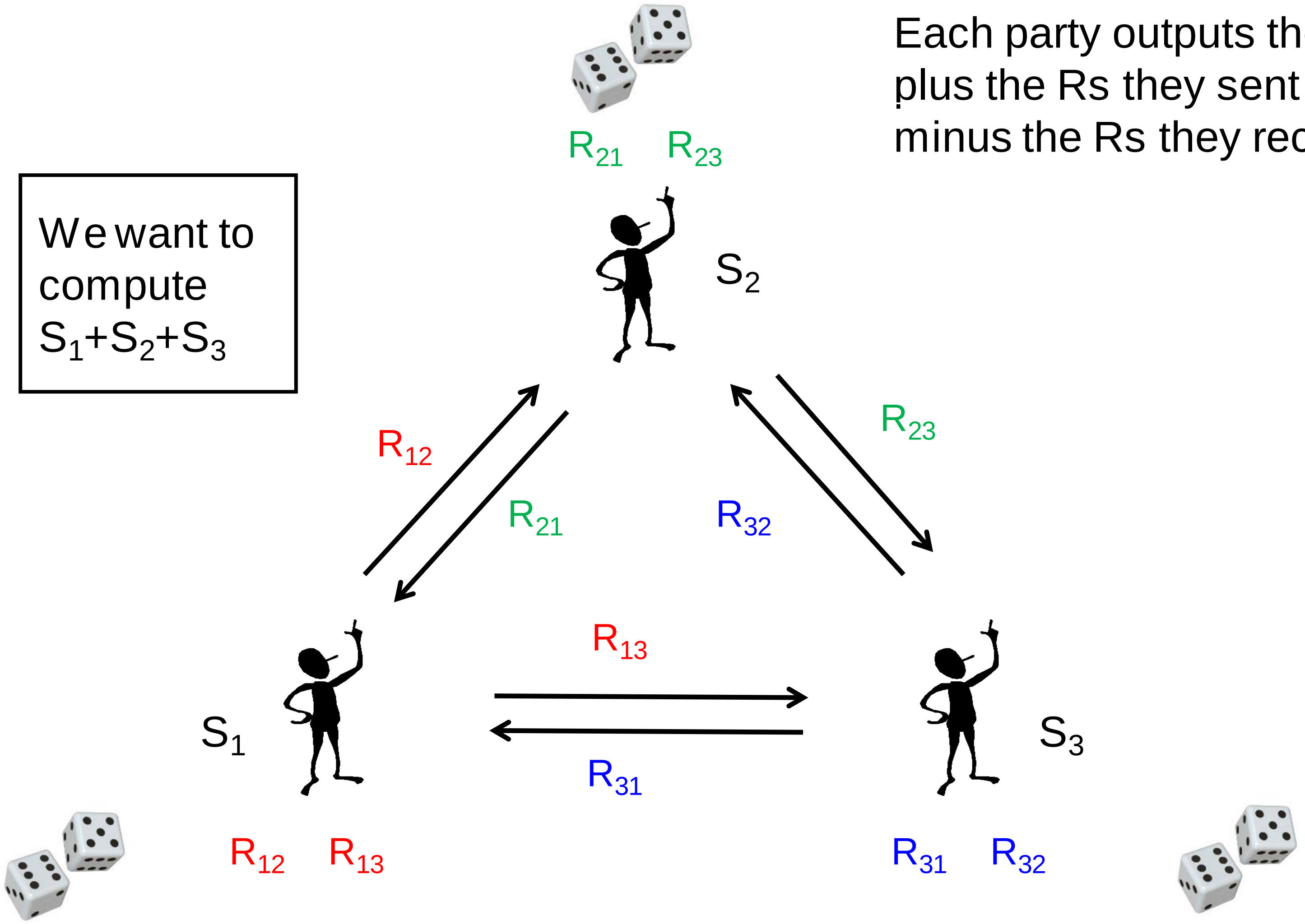
"Virtual Enclave"
"Simulated Honest Broker"

Crypto provides robustness against collusion and cheating!

Average Income Example

Each party outputs their salary
plus the Rs they sent
minus the Rs they received

We want to compute $S_1+S_2+S_3$



Average Income Example

Each party outputs their salary
plus the Rs they sent
minus the Rs they received

$$X_1 = S_1 + (R_{12}+R_{13}) - (R_{21}+R_{31})$$

$$X_2 = S_2 + (R_{21}+R_{23}) - (R_{12}+R_{32})$$

$$X_3 = S_3 + (R_{31}+R_{32}) - (R_{13}+R_{23})$$

(total)

$$X_1+X_2+X_3 = S_1+S_2+S_3$$

MPC Models

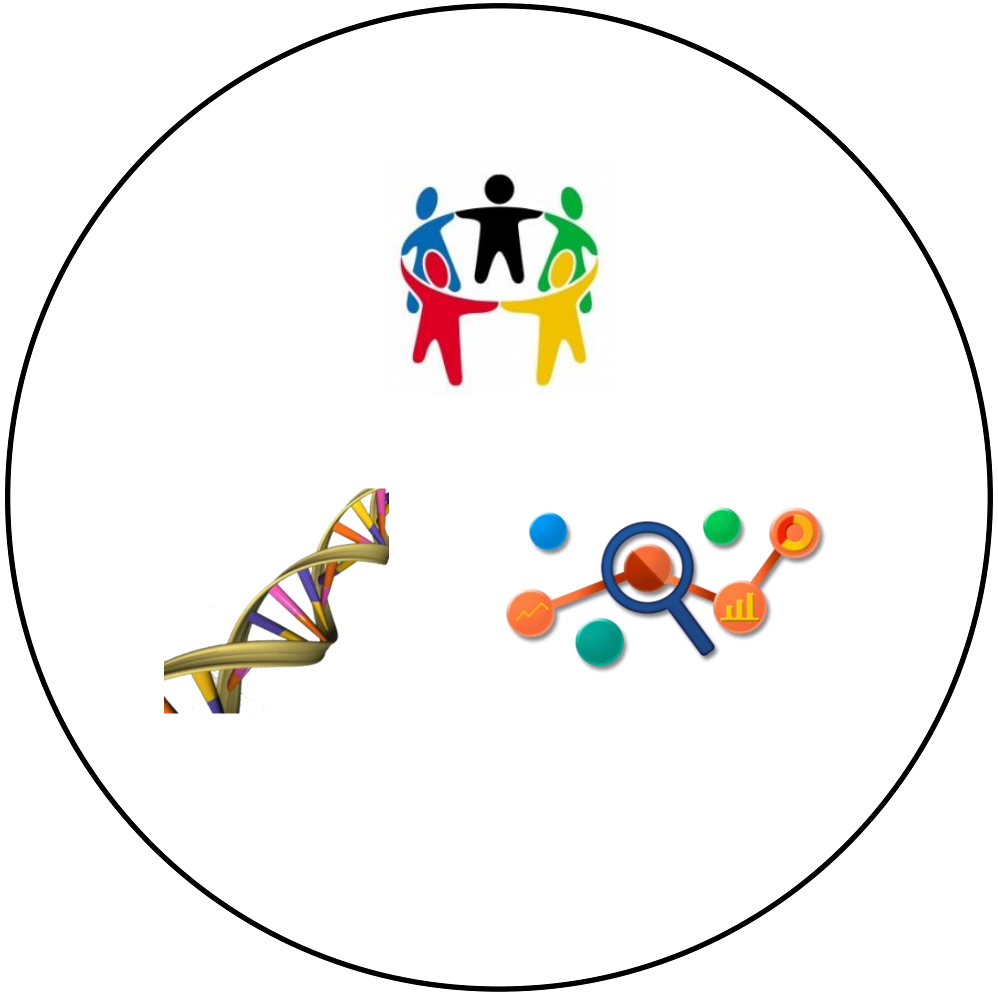
- Collusion threshold: no collusion, fractional, all-but-one
- Adversarial Behavior: Honest-but-curious, covert, malicious, honest-looking
- Forward secrecy/refreshing: Static, Mobile/Proactive
- Types of computation: circuits (math formulas), RAM (database/programs)
- Setup: None, Correlated Randomness, Physically Unclonable Functions (PUFs)

Outline

- MPC Review
- New Research
- Looking Ahead
- Conclusion

MPC in the Real World

- Taulbee Survey (Feigenbaum et al. 2004)
- Sugar Beet Auction (Bogetoft et al. 2009)
- Boston University Wage Study (Lapets et al. 2015)
- Estonian Ministry of Economic Affairs
  - Statistics on Estonian Companies (Bogdanov, Talviste, Willemson 2012)
  - Statistics on Tax and Education data (Bogdanov et al. 2014)
- Secure Conjunction Analysis (Hemenway et al. 2016)
  - Numerical analysis on 200k+ operations on floating point approximations
- Google Private Join and Compute (Ion et al. 2019)
- The list goes on...



Interdisciplinary Efforts

- Financial Sector
  - Abbe, Amir, Lo (2012) in The American Economic Review
  - Flood et al. (2013) Federal Reserve Bank of Cleveland
- Biomedicine & Healthcare
  - iDASH project
  - U. Michigan
- DARPA
  - PROgramming Computation on EncrypEd Data (PROCEED)
  - Brandeis
  - Securing Information for Encrypted Verification and Evaluation (SIEVE)
  - ...

One Story

- Two Sloan-funded workshops hosted at ICPSR in 2015
- Cryptographers + data scientists from health, education, finance, etc.
- Use MPC for data science!

MPC Toolkit

- Descriptive Statistics
  - Means, (co)variances, crosstabs
- Multiple Linear/Logistic Regression
- Survival Analysis
- …
- <u>Many of these exist</u> in research works, Cybernetica's RMind, etc.
- *How to move from crypto to deployment for the benefit of the sciences?*

Secure Analytics For Reticent Non-consolidated databases (SAFRN)

- MPC platform for data science and analytics
- Different architecture from existing solutions
- Start with easy queries with focus on strong baseline of scalability, compatibility
- Recently completed work, joint with ICPSR, funded by Laura and John Arnold Foundation (now Arnold Ventures)

Existing Paradigms

- Secure Enclave (security in transit and at rest), e.g.
  - Centralized database collects encrypted data that it can decrypt
  - Stores it in an encrypted database that it can decrypt
  - Queries are ran on semi-decrypted data

Existing Paradigms

- Crypto trend (secure computation on end-to-end encrypted data), e.g.
  - Centralized database collects encrypted data that it cannot decrypt
  - Stores it in an encrypted database that it cannot decrypt
  - Queries are ran on encrypted data, results can only be decrypted by recipient
  - This is pretty good, but…

Existing Paradigms

- Crypto trend (secure compu … end encrypted data), e.g.
  - Centralized database collects encrypted data that it cannot decrypt
  - Stores it in an encrypted database … decrypt
  - Queries are ran on encrypted data, results can only be decrypted by recipient
  - This is pretty good, but…

> Who would play this role? Recently ISRG has helped serve as this party, but difficult in general

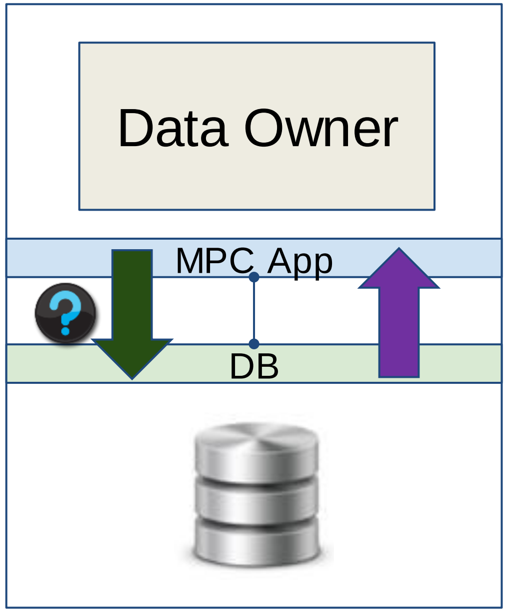
> For Big Data, this might be costly to maintain

SAFRN Design

- Data is not required to be centralized
  - No centralized server or servers (maybe only for massive # of parties)
  - Each organization can use their existing database system
  - Local data can change as fast and often as they like, with no impact to others

SAFRN Design

- Orchestrate and federate crypto and DB tasks
  - Asynchronous point-to-point design
  - Analyses are translated into plain database queries
  - Defines a secure edge-compute manifest which allows DBs to process queries before sending out encrypted intermediate data for secure computation



SAFRN Design

- Does not require new specially encrypted databases
  - Let the databases do the databasing
  - Parties can use their own existing data storage solution (text, JSON, CSV, XML, Excel, SQL, NoSQL, ...), just use a small query plugin adapter (ODBC)

SAFRN Preliminary Demo

- Collaboration with ICPSR to design synthetic database, data, and queries
  - Inspired by several real-world needs
- Analyst: Wants to build a report for the public good, linking group data to income data
- Income: Contains (CaseID, Income) pairs
- Group1,2,3: Contains (CaseID, Attrib_A={1,2}, Group_X={1,2,3}, Attrib_B={1,2,3}) tuples

Sample Source Data

CaseID	Income	Attrib_A	Group_X	Attrib_B
5144502	1258	1	2	1
5072643	2872	2	1	3
7784607	1436	2	3	1
141444	1369	2	2	1
2136566	5093	1	1	1
8610663	499	2	2	2
486581	2803	2	1	2
1111017	311	2	3	1
5091884	1275	1	2	1

Computations

- Very simple computations to start with
- Frequency/Crosstabs
  - E.g. tabulate Attrib_A across all 3 Groups without revealing individual counts
- Means
  - Compute average Income categorized by (Attrib_A, Attrib_B)
- Also some not-so-simple computations
  - Secure Regression
  - Higher order moments

Example: Average Income by Group_X and Attrib_A

Average Income by Group and Attrib_A				
Group_X				
Attrib_A	1	2	3	Total
1	???	???	???	???
2	???	???	???	???
Total	???	???	???	???

Approach

- Lots of cryptographic and engineering questions arise even with simple computations!
- Frequency/Crosstabs
  - Secure sums
- Means
  - Secure shared-output private intersection-sums between Income and each Group
    - Can leverage private set intersection solutions for summation
  - Use secure sum to gather shares
  - Secure division

Data Divided Among Databases Linked by CaseID

Group 1 Database

CaseID	Attrib_A	Group_X	Attrib_B
5072643	2	1	3
2136566	1	1	1
486581	2	1	2
6909778	2	1	2
2567912	1	1	1
7567352	1	1	1
9720209	2	1	1
1537271	2	1	1

Income Database

CaseID	Income
5144502	1258
5072643	2872
7784607	1436
141444	1369
2136566	5093
8610663	499
486581	2803
1111017	311

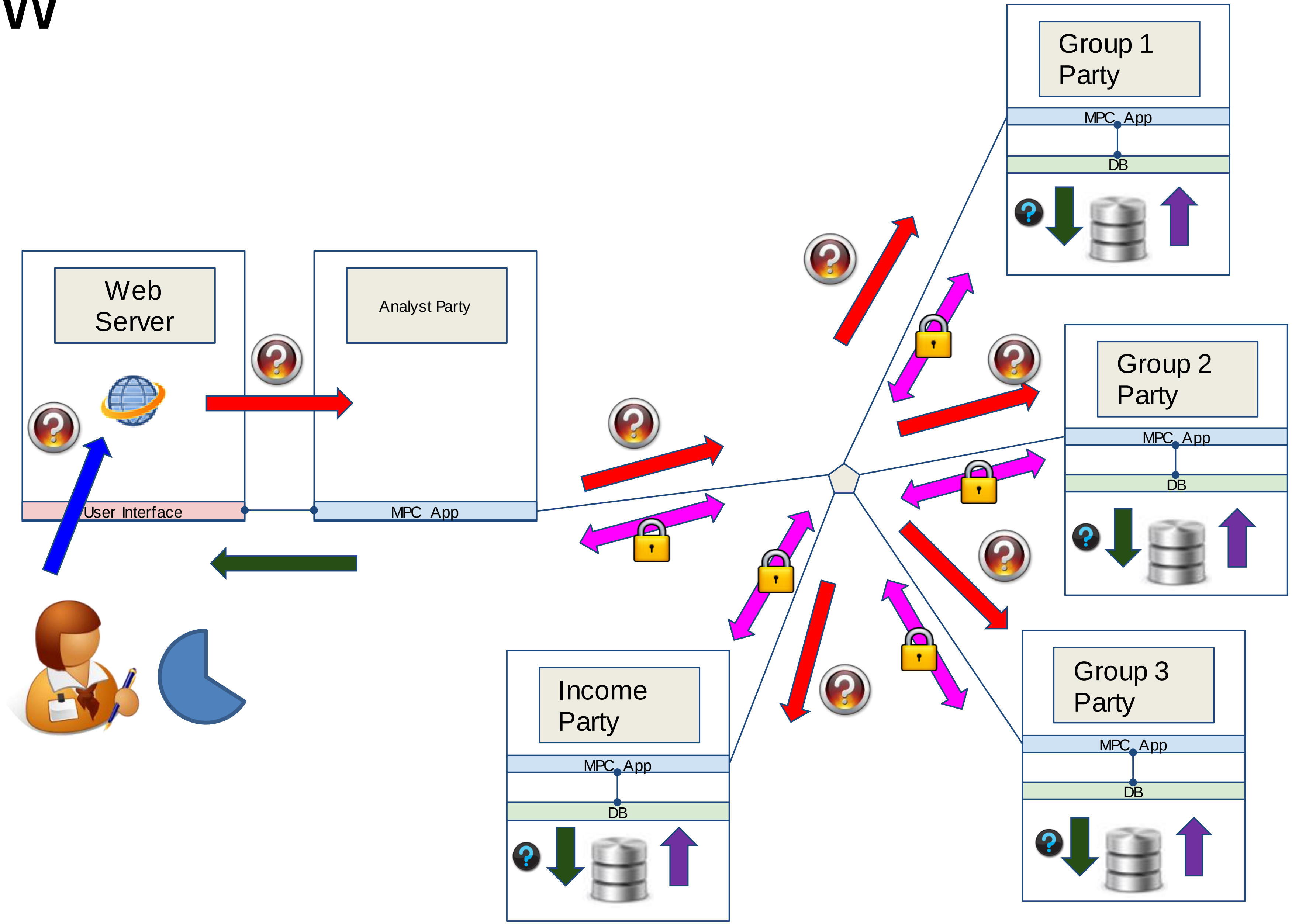
Group 2 Database

CaseID	Attrib_A	Group_X	Attrib_B
5144502	1	2	1
141444	2	2	1
8610663	2	2	2
5091884	1	2	1
9242299	1	2	1
2087684	1	2	1
1247904	1	2	1
2293696	1	2	1

Group 3 Database

CaseID	Attrib_A	Group_X	Attrib_B
7784607	2	3	1
1111017	2	3	1
455424	2	3	2
7863237	2	3	2
2791144	1	3	1
3122300	2	3	1
9589976	2	3	3
5043970	1	3	1

Flow



Web Server
User Interface
Analyst Party
MPC App
Group 1 Party
MPC App
DB
Group 2 Party
MPC App
DB
Group 3 Party
MPC App
DB
Income Party
MPC App
DB

Example: Average Income by Group_X and Attrib_A

Average Income by Group and Attrib_A				
	Group_X			
Attrib_A	1	2	3	Total
1	$2,696	$3,110	$2,110	$2,754
2	$2,552	$2,436	$1,514	$2,106
Total	$2,657	$2,685	$1,621	$2,408

Outline

- MPC Review
- New Research
- Looking Ahead
- Conclusion

Lessons Learned

- We ran tests with ITS from the University of Michigan

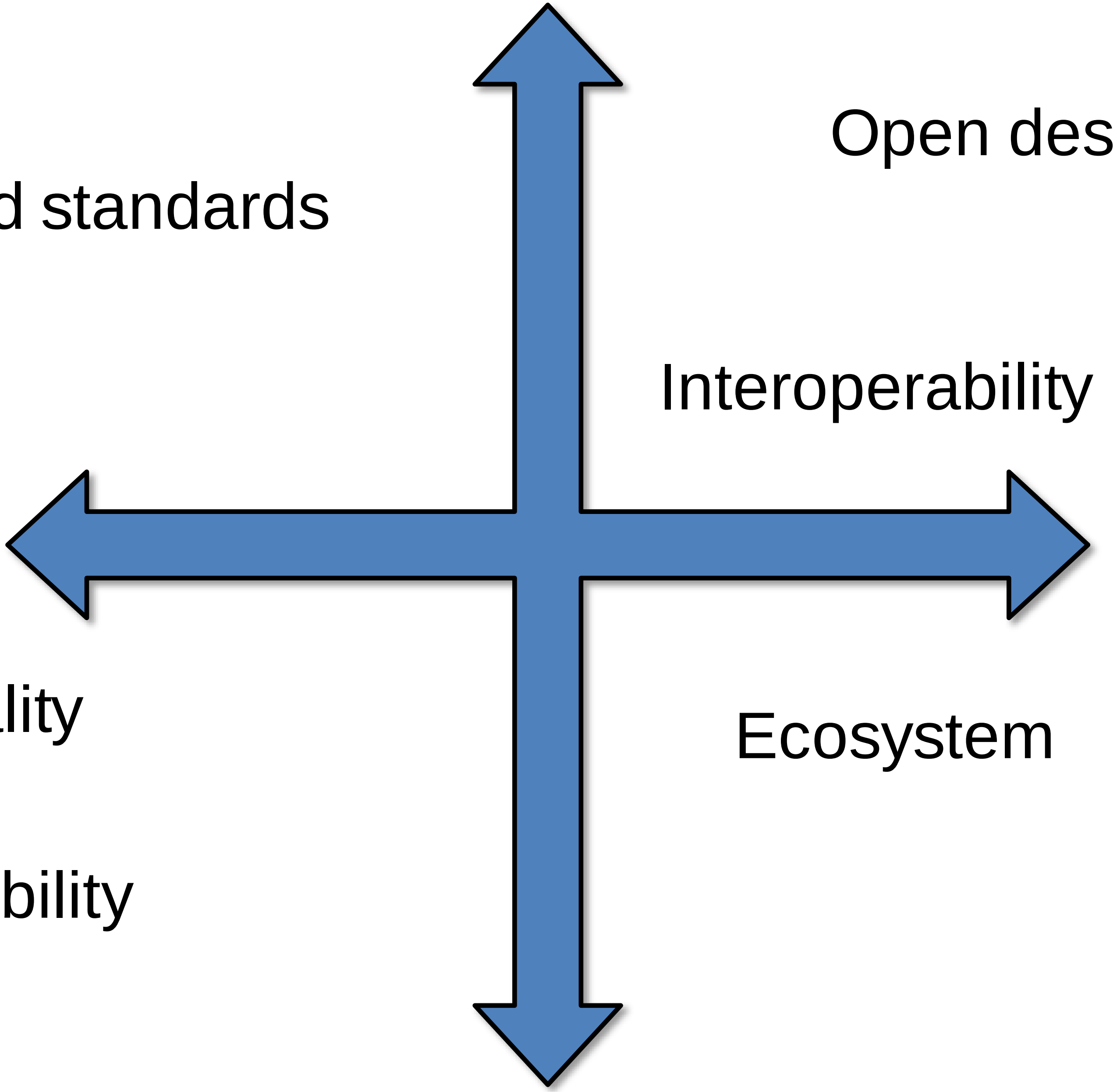
Getting the crypto right is just the first step

Deployment

Usability

Future Work

- Better deployment support
- Make it easier to use
- Enhance capabilities
  - Other analytics
  - Better support for different database plugins
  - Language for expressing computations



Open design, open source
Help build standards
Interoperability
Functionality
Ecosystem
Usability
Support

Future Work

- Outreach to various communities
- Applications to problems faced by data scientists

Outline

- MPC Review
- New Research
- Looking Ahead
- Conclusion

Conclusion

- Introduction into secure multiparty computation
- Presented a new approach: SAFRN
- Hope to see continued growth in this area

Thank you!