SCALABLE PRIVACY-PRESERVING DISTRIBUTED EXTREMELY RANDOMIZED TREES FOR STRUCTURED DATA WITH MULTIPLE COLLUDING PARTIES

AMIN AMINIFAR[1], FAZLE RABBI[1,2], AND YNGVE LAMO[1]

[1] WESTERN NORWAY UNIVERSITY OF APPLIED SCIENCES

[2] UNIVERSITY OF BERGEN

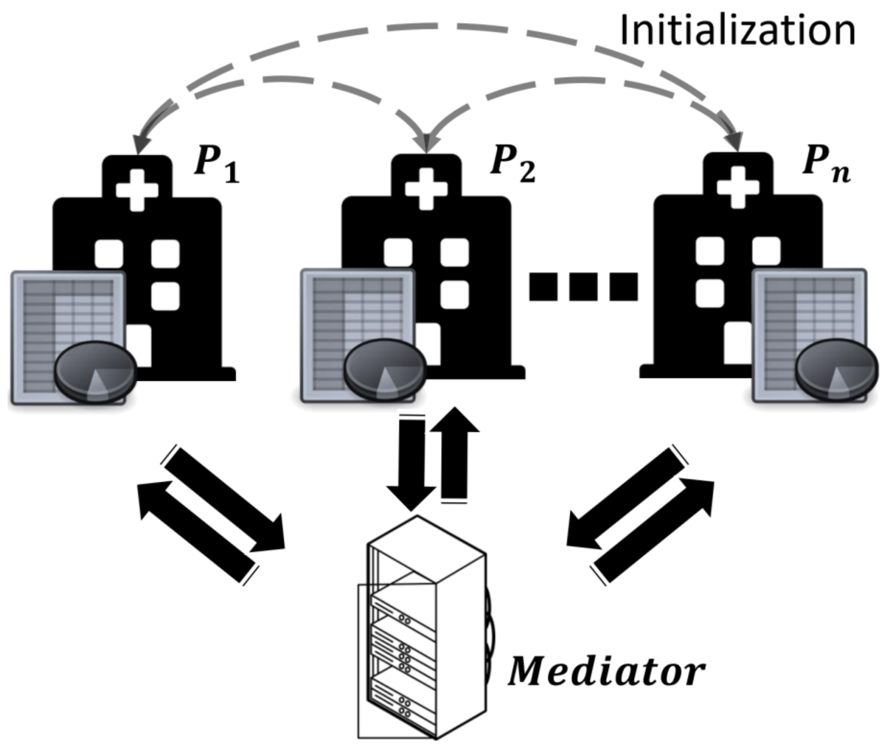
IEEE ICASSP 2021

Outline

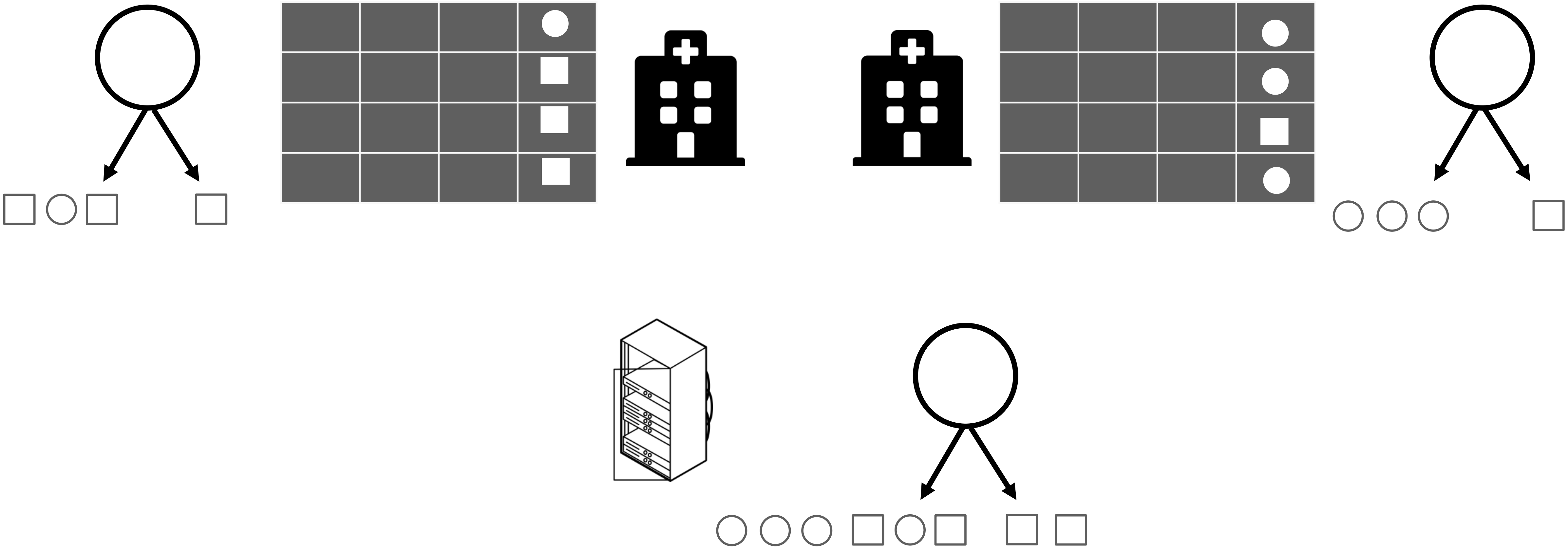
- Problem
- Distributed Extremely Randomized Trees
- Secure Multi-Party Computation for Privacy-Preserving Distributed ERT
- Efficient Handling of Large-Scale Data
- Evaluation
- Conclusion

Problem

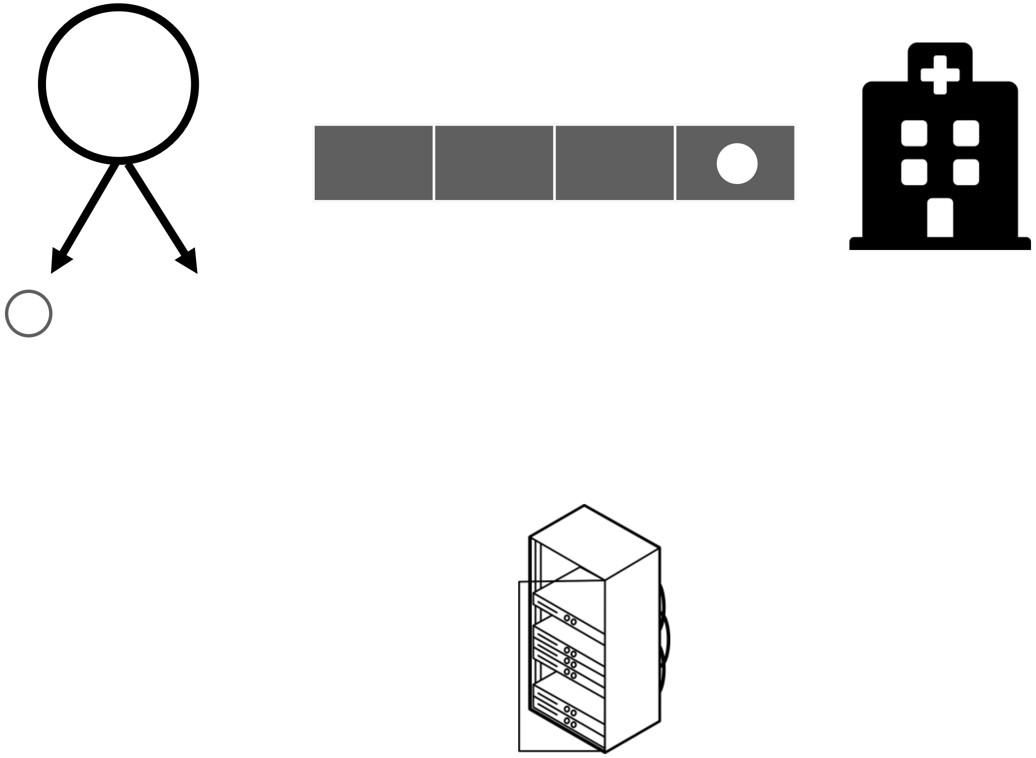
- Learning classification models from data distributed over multiple parties
- Without sharing of the raw healthcare information, due to privacy and legal concerns
- Horizontally partitioned structured data



Distributed Extremely Randomized Trees

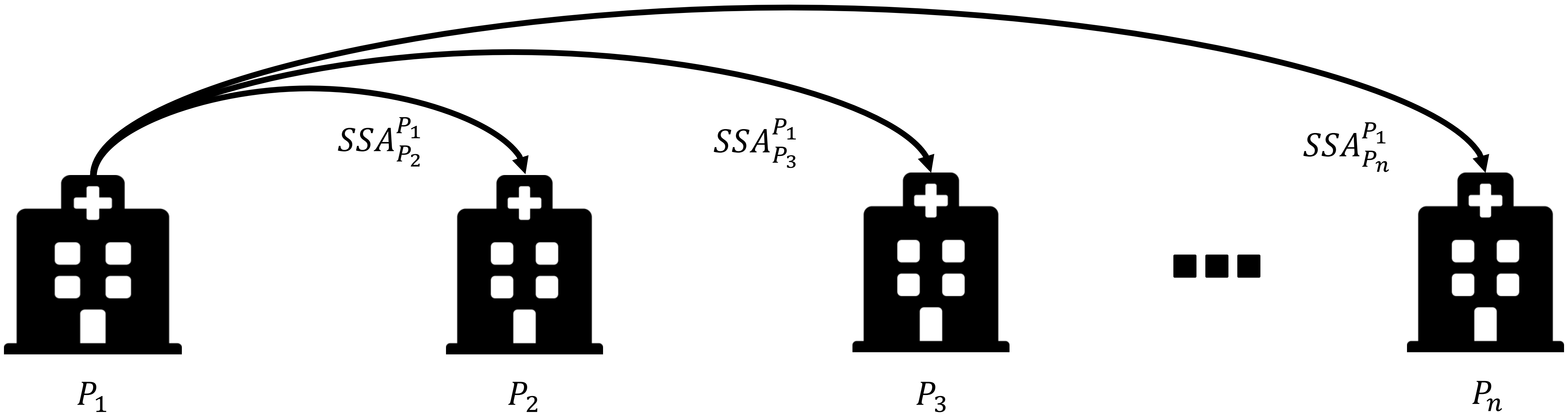


Secure Multi-Party Computation for Privacy-Preserving Distributed ERT

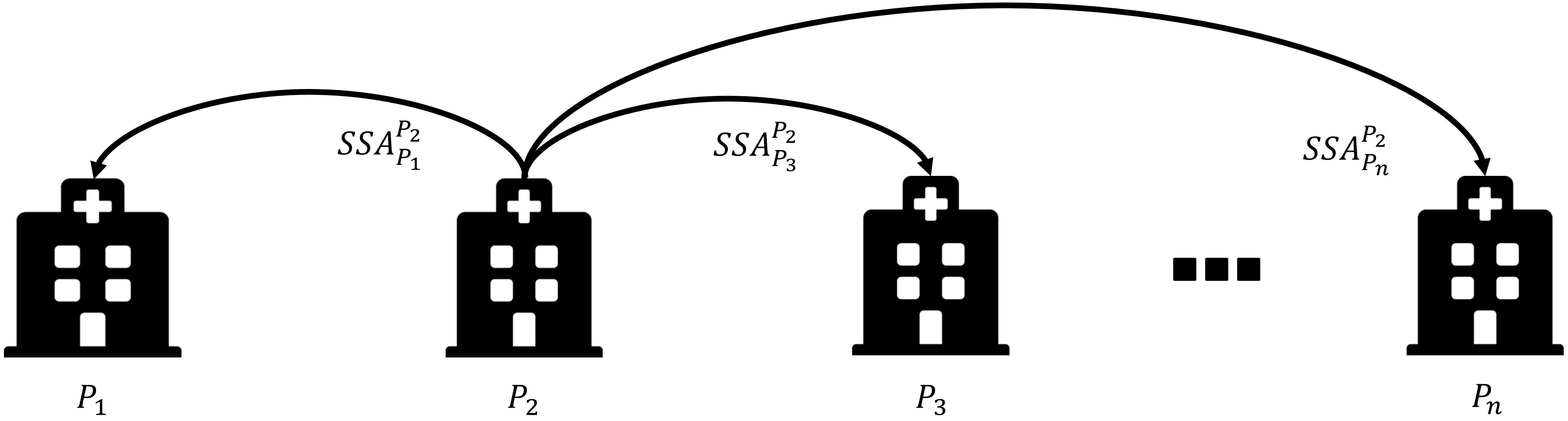


Secure Multi-Party Computation for Privacy-Preserving Distributed ERT

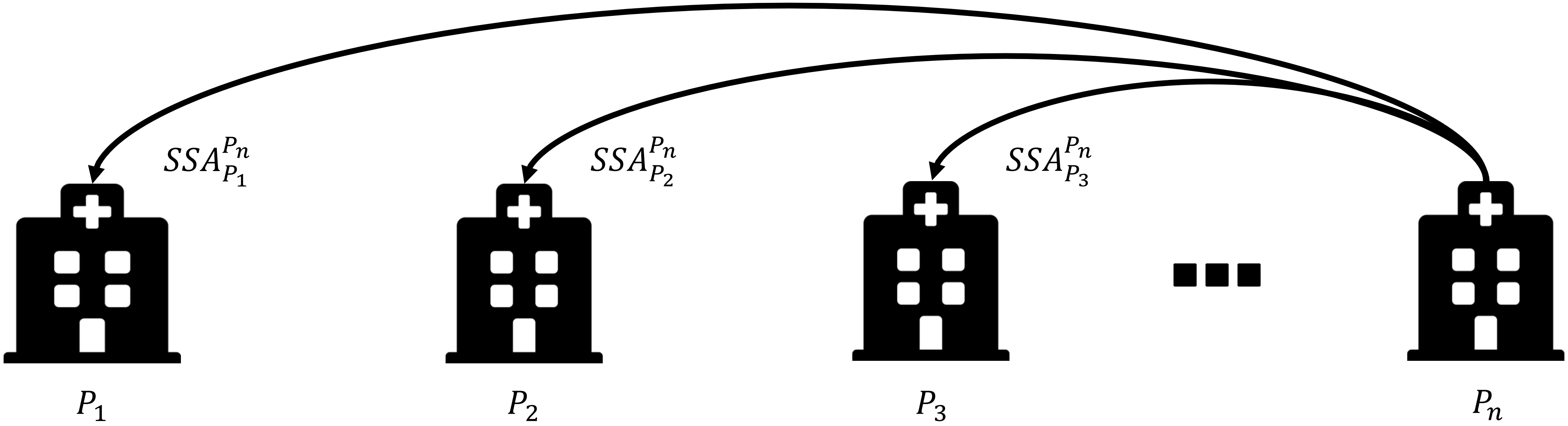
Each data holder party sends personal random seeds to all data holder parties



Secure Multi-Party Computation for Privacy-Preserving Distributed ERT



Secure Multi-Party Computation for Privacy-Preserving Distributed ERT

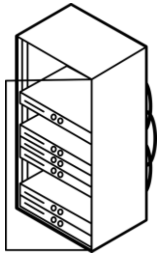


Secure Multi-Party Computation for Privacy-Preserving Distributed ERT

Send	$SSA_{P_2}^{P_1}, \ldots, SSA_{P_n}^{P_1}$	$SSA_{P_1}^{P_2}, \ldots, SSA_{P_n}^{P_2}$	$SSA_{P_1}^{P_3}, \ldots, SSA_{P_n}^{P_3}$	$SSA_{P_1}^{P_n}, \ldots, SSA_{P_{n-1}}^{P_n}$
Receive	$SSA_{P_1}^{P_2}, \ldots, SSA_{P_1}^{P_n}$	$SSA_{P_2}^{P_1}, \ldots, SSA_{P_2}^{P_n}$	$SSA_{P_3}^{P_1}, \ldots, SSA_{P_3}^{P_n}$	$SSA_{P_n}^{P_1}, \ldots, SSA_{P_n}^{P_{n-1}}$

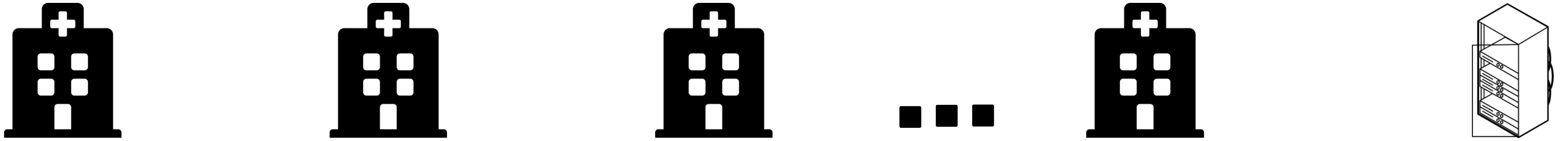
Secure Multi-Party Computation for Privacy-Preserving Distributed ERT

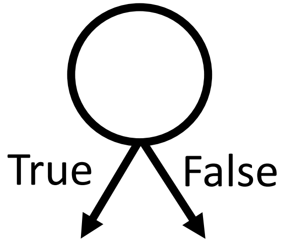
P_1 P_2 P_3 P_n



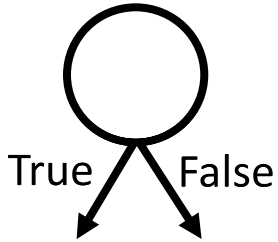
$$rnd\_sum^{P_1}_{others} + (secret\_val^{P_1} - rnd\_sum^{P_1}_{self}) +$$
$$rnd\_sum^{P_2}_{others} + (secret\_val^{P_2} - rnd\_sum^{P_2}_{self}) +$$
$$\vdots$$
$$rnd\_sum^{P_n}_{others} + (secret\_val^{P_n} - rnd\_sum^{P_n}_{self}) = Sum$$

Efficient Handling of Large-Scale Data



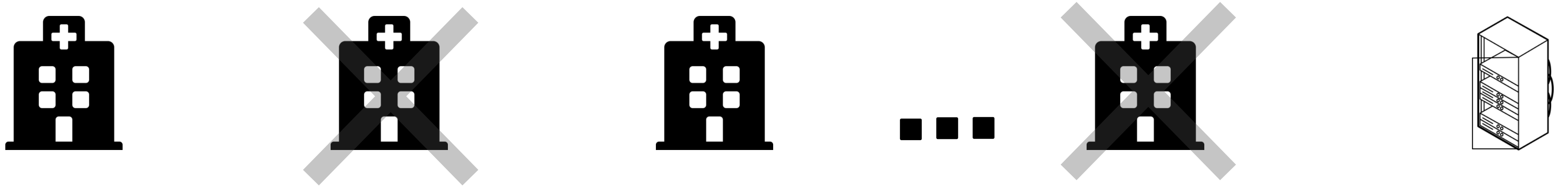


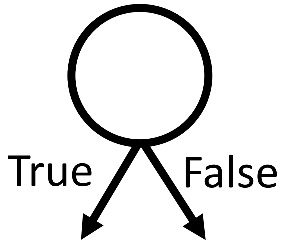
$$\begin{cases}\text{True: } [0,1] \\ \text{False: } [1,1]\end{cases} + \begin{cases}\text{True: } [2,1] \\ \text{False: } [1,0]\end{cases} + \begin{cases}\text{True: } [2,1] \\ \text{False: } [3,1]\end{cases} + \begin{cases}\text{True: } [0,4] \\ \text{False: } [1,1]\end{cases} = \begin{cases}\text{True: } [4,7] \\ \text{False: } [6,3]\end{cases}$$



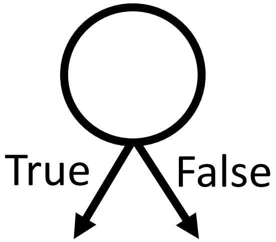
$$\begin{cases}\text{True: } [0,2] \\ \text{False: } [1,0]\end{cases} + \begin{cases}\text{True: } [0,1] \\ \text{False: } [3,0]\end{cases} + \begin{cases}\text{True: } [1,1] \\ \text{False: } [4,1]\end{cases} + \begin{cases}\text{True: } [1,5] \\ \text{False: } [0,0]\end{cases} = \begin{cases}\text{True: } [2,9] \\ \text{False: } [8,1]\end{cases}$$

Efficient Handling of Large-Scale Data





$$\begin{cases}\text{True: } [0,1] \\ \text{False: } [1,1]\end{cases} + \cancel{\begin{cases}\text{True: } [2,1] \\ \text{False: } [1,0]\end{cases}} + \begin{cases}\text{True: } [2,1] \\ \text{False: } [3,1]\end{cases} + \cancel{\begin{cases}\text{True: } [0,4] \\ \text{False: } [1,1]\end{cases}} = \begin{cases}\text{True: } [2,2] \\ \text{False: } [4,2]\end{cases}$$



$$\begin{cases}\text{True: } [0,2] \\ \text{False: } [1,0]\end{cases} + \cancel{\begin{cases}\text{True: } [0,1] \\ \text{False: } [3,0]\end{cases}} + \begin{cases}\text{True: } [1,1] \\ \text{False: } [4,1]\end{cases} + \cancel{\begin{cases}\text{True: } [1,5] \\ \text{False: } [0,0]\end{cases}} = \begin{cases}\text{True: } [1,3] \\ \text{False: } [5,1]\end{cases}$$

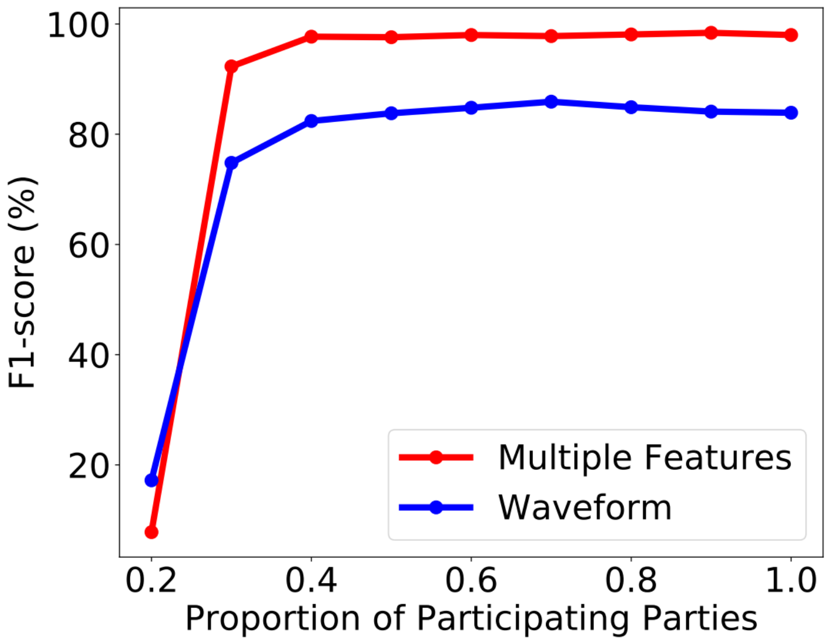
Evaluation

- Criteria of evaluation for privacy-preserving data mining approaches
  - Classification performance, overhead, and privacy

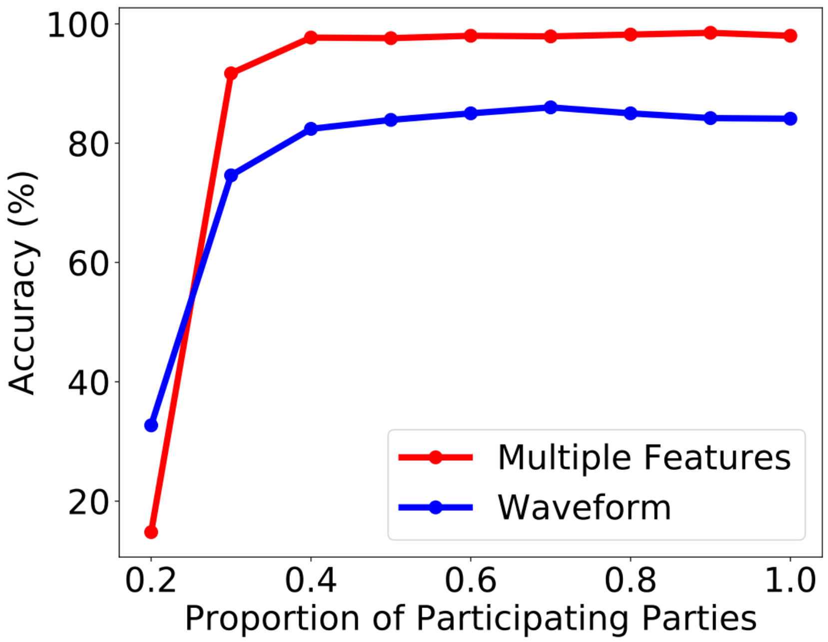
Table 1: Scalability and privacy comparison against existing techniques

Approach	Party	Communication (N is the number of parties)			Min Number of Colluding Parties
		Send	Receive	Total (All N parties)	
Distributed ERT	All	1	1	$2N$	1
k-PPD-ERT	Data Holders	1	0	$2(N-1)$	$k+1$ ($k < N$)
	Mediator	0	$N-1$		
Shamir [31]	k-1 Parties	N	$N-1$	$2(N^2-N+k-1)$	k ($k < N$)
	One Party	$N-1$	$N+k-2$		
	The Rest	$N-1$	$N-1$		

Evaluation

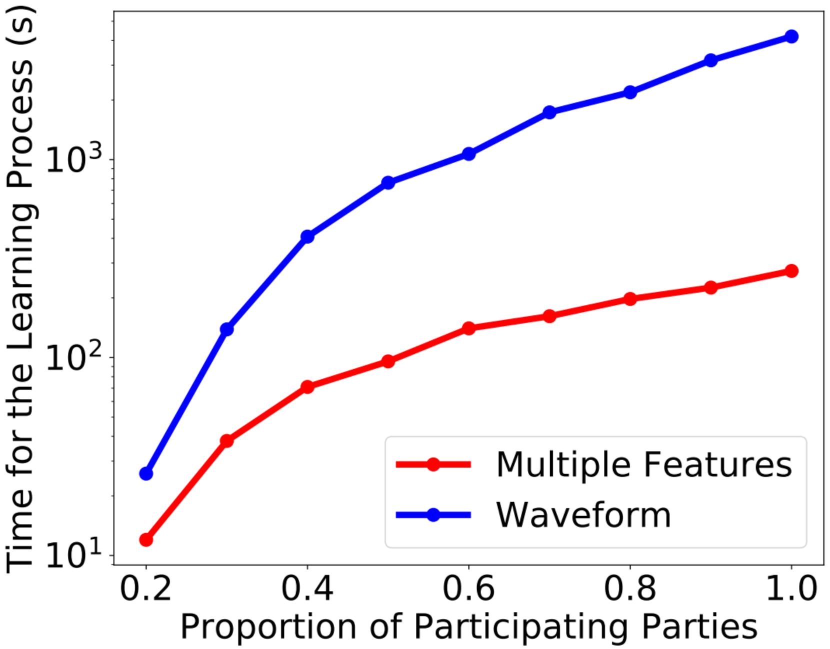


100
80
60
40
20
F1-score (%)
0.2
0.4
0.6
0.8
1.0
Proportion of Participating Parties
Multiple Features
Waveform

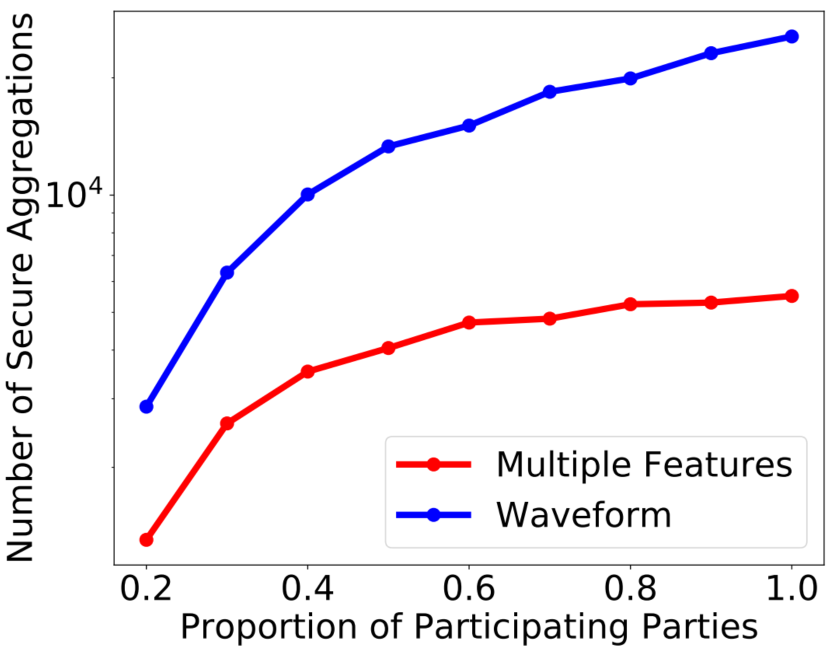


100
80
60
40
20
Accuracy (%)
0.2
0.4
0.6
0.8
1.0
Proportion of Participating Parties
Multiple Features
Waveform

Evaluation



Time for the Learning Process (s)
10^3
10^2
10^1
Multiple Features
Waveform
0.2
0.4
0.6
0.8
1.0
Proportion of Participating Parties



Number of Secure Aggregations
10^4
Multiple Features
Waveform
0.2
0.4
0.6
0.8
1.0
Proportion of Participating Parties

Conclusion

- k-PPD-ERT is an extension of ERT algorithm learning classification models when data is distributed.
- The secure multi-party computation technique for k-PPD-ERT is resilient to the collusion of up to k data holder parties.
- The secure multi-party computation technique for k-PPD-ERT is efficient with respect to the communication overhead.
- Limited participation of data holder parties at every round of the learning process decreases the overhead without any noticeable loss in the learning performance.