Secure Multiparty Linear and Logistic Regression Based on Homomorphic Encryption

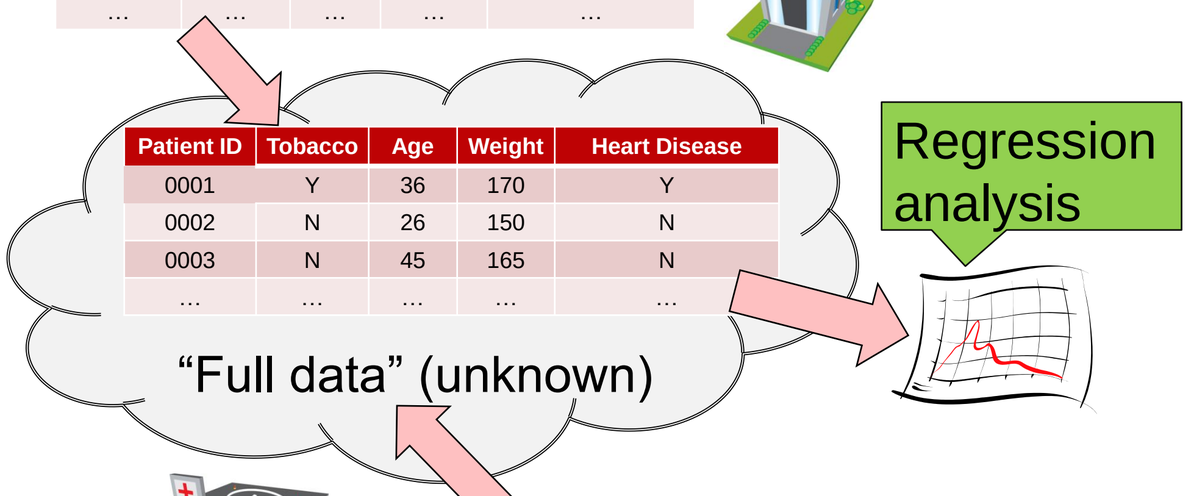
Rob Hall (MLD), Stephen Fienberg (Statistics Dept.) and Yuval Nardi (Technion)

Privacy Preserving Data Mining

- Several parties have data on a common set of entities, but each party's data is incomplete:

Patient ID	Tobacco	Age	Weight	Heart Disease
0001	?	36	170	?
0002	N	26	150	?
0003	N	45	165	?
…	…	…	…	…

Party 1



Patient ID	Tobacco	Age	Weight	Heart Disease
0001	Y	36	170	Y
0002	N	26	150	N
0003	N	45	165	N
…	…	…	…	…

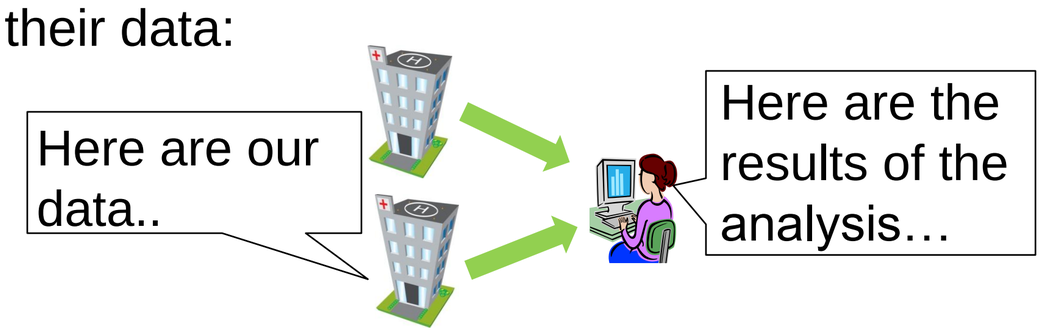
Party 2

Patient ID	Tobacco	Age	Weight	Heart Disease
0001	Y	?	170	Y
0002	?	26	150	N
0003	?	?	165	N
…	…	…	…	…

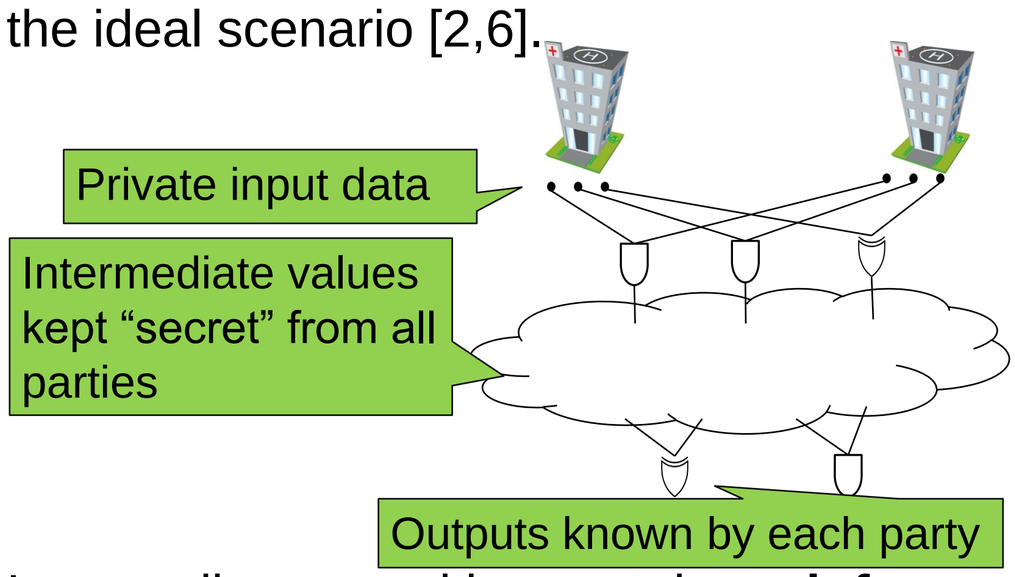
- Each party's data is **private**, and the parties are **unwilling to share** their data.
- We do regression on the unknown, full data matrix, without requiring the parties to reveal their private data.

Secure Multiparty Computation

- Each party should only learns the regression estimates and *whatever is implied by them*.
- Ideally, the parties could trust a 3rd party with their data:



- Using cryptography, the parties perform a protocol, with the same impact on privacy as the ideal scenario [2,6].



- Intermediate quantities contain **no information** about the private data *beyond that contained in the output.*

Protocol For Regression

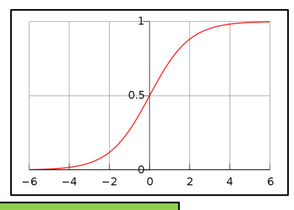
Main Ideas

- Any computation may be performed but the construction of [6] is inefficient.
- Computations consisting of **addition** and **multiplication** are **efficient** (see e.g., [5,7]).
- Compute regression parameters via Newton-Raphson method:

$$\beta_{t+1} = \beta_t - (\nabla^2 \ell)^{-1} \nabla \ell$$

Matrix inversion

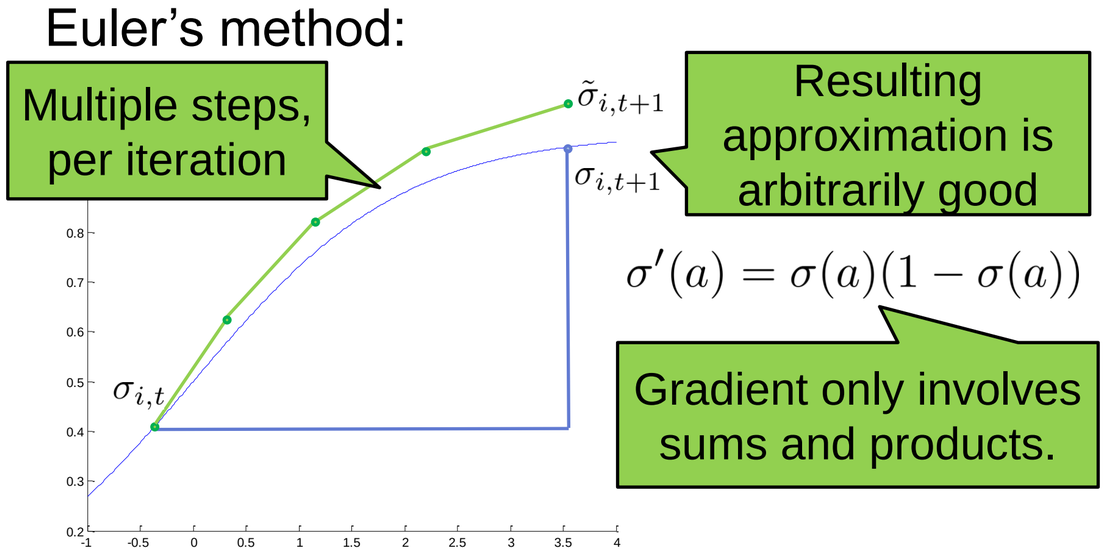
Non-linear (logistic)



- Matrix inversion reduces to sums and products (applying Newton's method):

$$A_0 = \epsilon I, \;\; A_{t+1} = 2A_t - A_t^2 M$$

- Logistic function may be approximated by Euler's method:



$$\sigma'(a) = \sigma(a)(1 - \sigma(a))$$

Gradient only involves sums and products.

Implementation

- Implemented in C using GMP to handle operations on 1024 bit long numbers.
- Secure multiplications take ~0.1s.
- Much slower than non-secure methods.
- Appropriate for moderate size problems.
- May easily be parallelized to take advantage of multiple machines.

Ongoing Work

- Securely computing link functions of other GLMs.
- Investigating which goodness of fit statistics compromise privacy.
- Secure **record linkage** (when no unique ID for the records is known to all parties)
- Comparison with the alternative: transform each party's data and share it [1].

MACHINE LEARNING DEPARTMENT
Carnegie Mellon University
Statistics Department

References

[1] Dwork, C (2008) . Differential Privacy, a Survey of Results. In TAMC.

[2] Fienberg, S. Hall, R and Nardi, Y. (2011) Achieving Both Valid and Secure Logistic Regression Analysis on Aggregated Data from Different Private Sources. (In submission).

[3] Goldreich, O (2004). Foundations of Cryptography: Volume 2, Basic Applications. Cambridge University Press.

[4] Hall, R. Fienberg, S. (2010) Privacy Preserving Record Linkage. In PSD 2010.

[5] Hall, R. Fienberg, S. Nardi, Y. (2011) Secure Multiple Linear Regression Based on Homomorphic Encryption. (In submission)

[6] Yao, A. C. (1982) Protocols for secure computations. In FOCS.

[7] Vaidya et. Al. (2006). Privacy Preserving Data Mining. Springer.