

Ensuring Data Protection: Technical Methods

Rebecca Wright

Rutgers University

www.cs.rutgers.edu/~rebecca.wright

Engaging Data Forum

Cambridge, MA, USA

October 12-13, 2009

Data Privacy and Data Use

- provide sufficient privacy as well as sufficient utility.
- What “sufficient”, “privacy”, and “utility” mean depends on context (or your point of view).
- The strongest notions of utility (being able to do anything you want with some data) do conflict with the strongest notions of privacy (not being able to learn anything about any individual), unless interactions are assumed to take place in a vacuum.

Personally Identifiable Information

- The concept of “personally identifiable information” (PII) is not robust in the face of today’s realities.
- Any interesting and relatively accurate data about someone can be personally identifiable if you have enough of it and appropriate auxiliary information.
- In today’s data landscape, both of these are often available.
- Examples: Sweeney’s work, AOL web search data [NYT06], Netflix challenge data [NS08], social network reidentification [BDK07], ...

Differential Privacy

- Introduced by Dwork et al. in 2006.
- Has proven useful for obtaining good utility and rigorous privacy guarantees in many cases (and growing).
- Separates privacy as information about an individual. Anything else can potentially be utility.
- Roughly speaking, an algorithm (for answering queries or for providing published data) is differentially privacy if the risk an individual user incurs to her privacy by participating in a database is essentially the same if she does not.

Differential Privacy

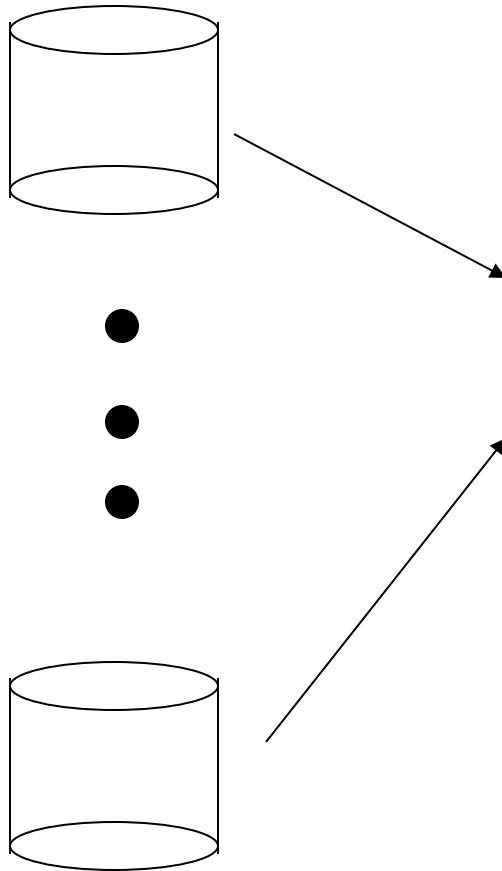
- Can be achieved via adding carefully chosen noise, in some cases little enough to still provide good utility for useful tasks. [DMNS06]
- Works for mediated interaction and for published data.
- Negatives (perhaps unavoidable):
 - Privacy degrades with the total number of queries made.
 - Published data utility guarantees are for specific purposes only.
- Practical examples: differentially private versions of the main Netflix recommender contenders [MM09], an SQL-like programming interface that provides differential privacy [McS09].

Secure Multiparty Computation

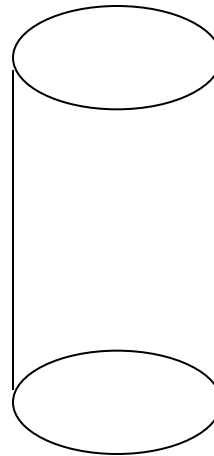
- Multiple data holders collaborate to compute a joint function of their data without needing to share their data with each other or any other party. [Yao86, BGW88, ...]
- In contrast to differential privacy, secure multiparty computation talks about the privacy of the computation process, not the privacy implication of the results.
- Both may be useful, perhaps in combination.

Secure Multiparty Computation

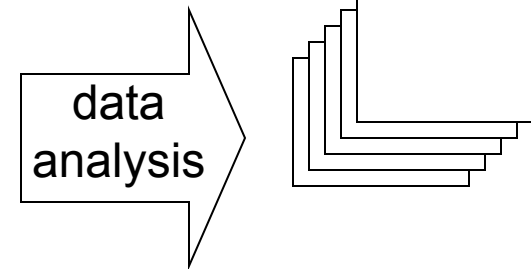
Multiple Data Sources



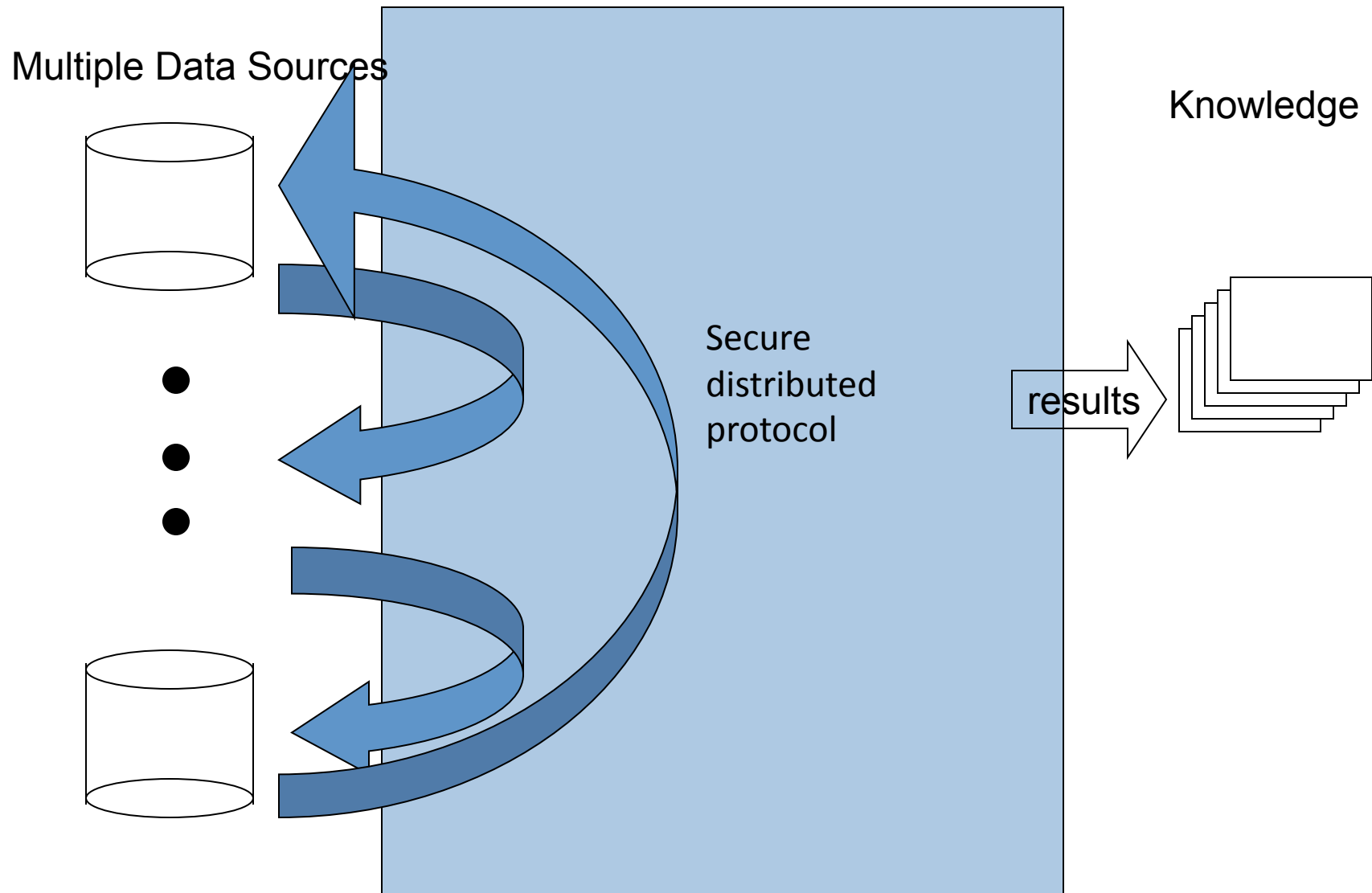
Combined data



Knowledge



Secure Multiparty Computation



Summary

- There are many reasons to use data and many associated benefits that can be derived.
- We need a better understanding of inherent tradeoffs between utility and privacy goals.
- Where applicable, differential privacy seems the best approach we currently have for providing rigorous privacy guarantees.
- Privacy is a social concept, not a technical one. Technology and policy must work together.