Computing Over Distributed Sensitive Data

Monday December 11, 2017

HARVARD
John A. Paulson
School of Engineering
and Applied Sciences

Georgetown
University

Sensitive Distributed Data

- Lots of data collected about individuals
  - Government agencies, banks, hospitals, research institutions, ...
- Many organizations collecting similar data
  - Or data about similar populations.

- Sharing this data would be great!
  - Benefits: social, scientific, business, security, ...
  - E.g., each hospital has small dataset; collectively could obtain statistically significant results
- But: much data contains sensitive personal details
  - Data sharing curbed for ethical, legal, or business reasons

Sensitive Distributed Data

- Lots of data collected about individuals
  - Government agencies, banks, hospitals, research instit
- Many
  - Or d
- Sharir
  - Bene
  - E.g., uld obtain statistically significant results
- But: much data contains sensitive personal details
  - Data sharing curbed for ethical, legal, or business reasons

This project:

Design protocols to allow sharing the computation (and not the data) between entities

Metadata

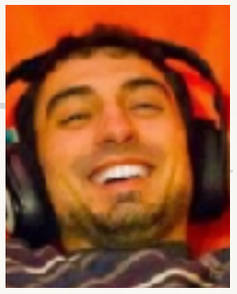
- NSF Secure and Trustworthy Cyberspace (SaTC) Large
- 4 years (May 2016–April 2019)
- PIs
  - Stephen Chong, James Honaker, Salil Vadhan
    - Harvard Center for Research on Computation and Society (CRCS)
  - Marco Gaboardi
    - University at Buffalo
  - Kobbi Nissim
    - Georgetown
  - Or Sheffet (collaborator)
    - University of Alberta

Metadata

Victor Balcer

Thomas Brawner

Steve Chong

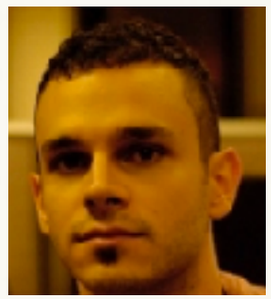


Gian Pietro Farina

Marco Gaboardi

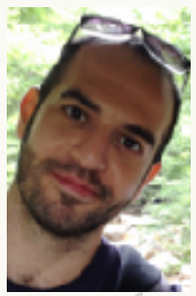
Anitha Gollamudi

James Honaker



Georgios Kellaris

Kobbi Nissim

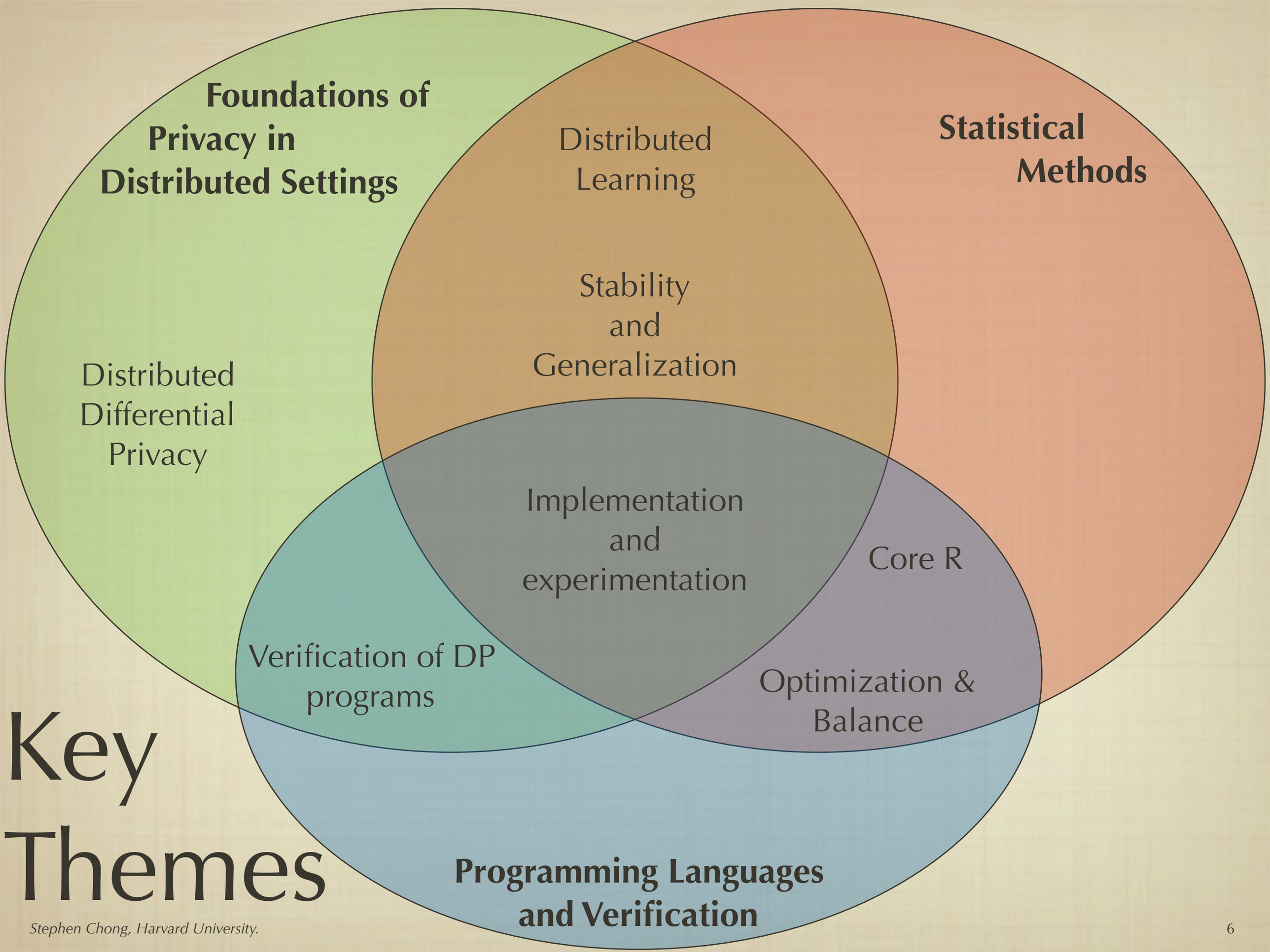


Or Sheffet

Uri Stemmer

Salil Vadhan

Key Themes


Foundations of Privacy in Distributed Settings
Distributed Differential Privacy
Verification of DP programs
Distributed Learning
Stability and Generalization
Implementation and experimentation
Statistical Methods
Core R
Optimization & Balance
Programming Languages and Verification

Specific Research Areas

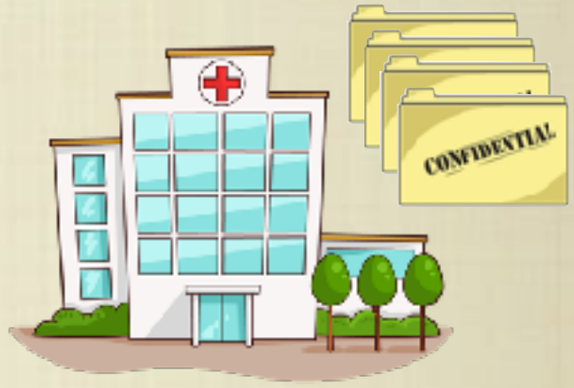
- Distributed Differential Privacy
  (Nissim, Vadhan, Sheffet)
  - Develop techniques for constructing protocols that apply differentially private computations to distributed data.
- Distributed Learning and Statistical Inference
  (Honaker, Nissim, Vadhan, Sheffet)
  - Develop techniques to perform statistical inference and machine learning computations over distributed sensitive data.
- Verification of Differentially Private Computation
  (Chong, Gaboardi, Nissim, Vadhan)
  - Develop automated and semi-automated techniques for ensuring programs are differentially private.

Specific Research Areas ctd.

- Verified Estimation and Balancing
  (Chong, Gaboardi, Honaker)
  - Use verified objective functions on distributed datasets, such as likelihood functions for statistical analysis, and balance functions for causal analysis, to create informationally efficient statistical estimates without needing to merge datasets or share raw data.
- Stability and Generalization
  (Honaker, Nissim, Vadhan, Sheffet)
  - Research the statistical stability and generalization provided by differential privacy and how these properties can be used in statistical analysis.
- Implementation and Experimentation
  (all PIs)
  - Validate our techniques by implementation and experimentation.

Example: Hospital Data

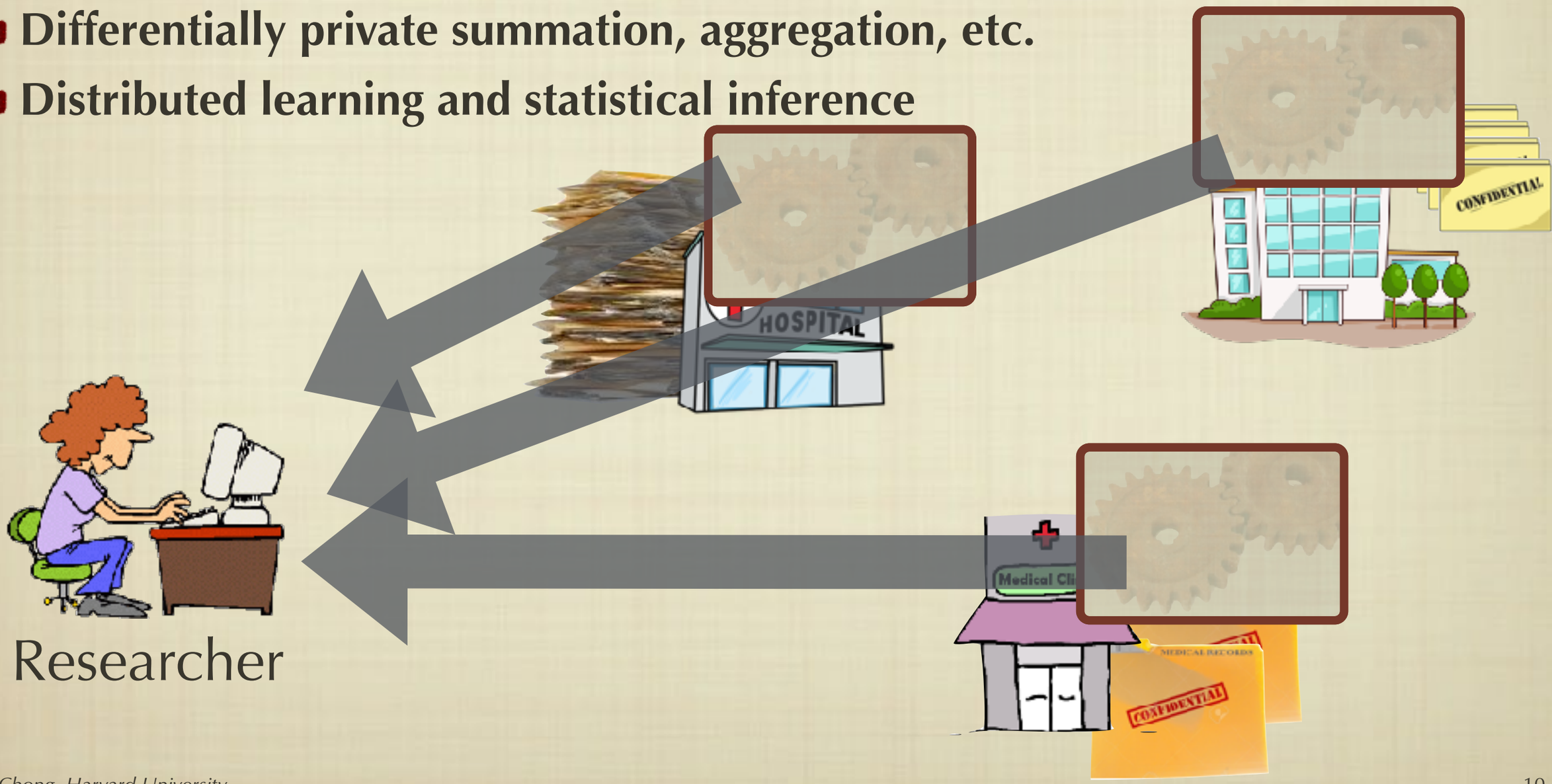
- Many hospitals, each with patient data
- Researcher would like to analyze data
- Hospitals cannot send data directly to researcher



Researcher

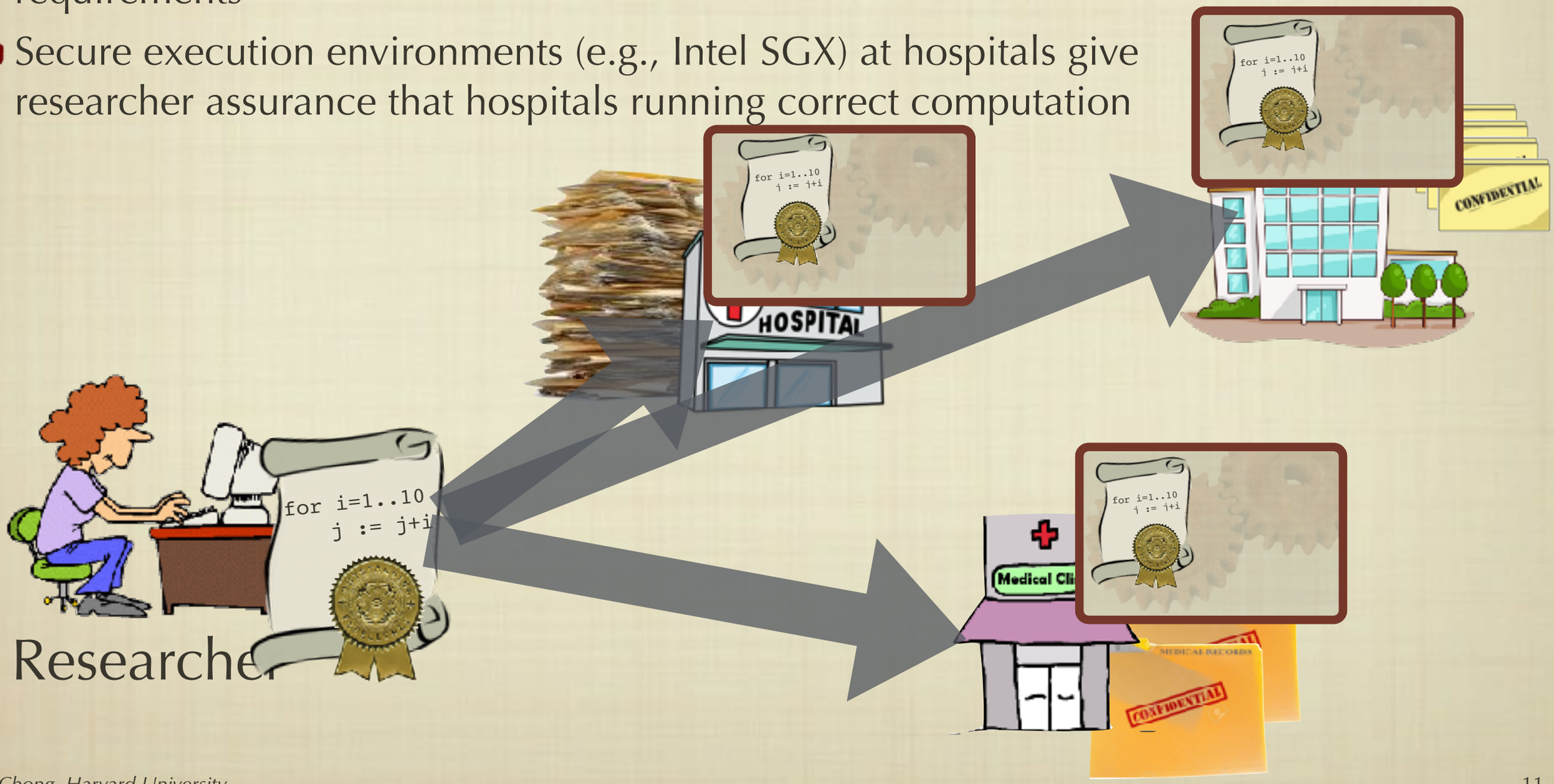
Example: Hospital Data

- Hospitals can individually run computation, send results to researcher
- What computations?
  - **Differentially private summation, aggregation, etc.**
  - **Distributed learning and statistical inference**



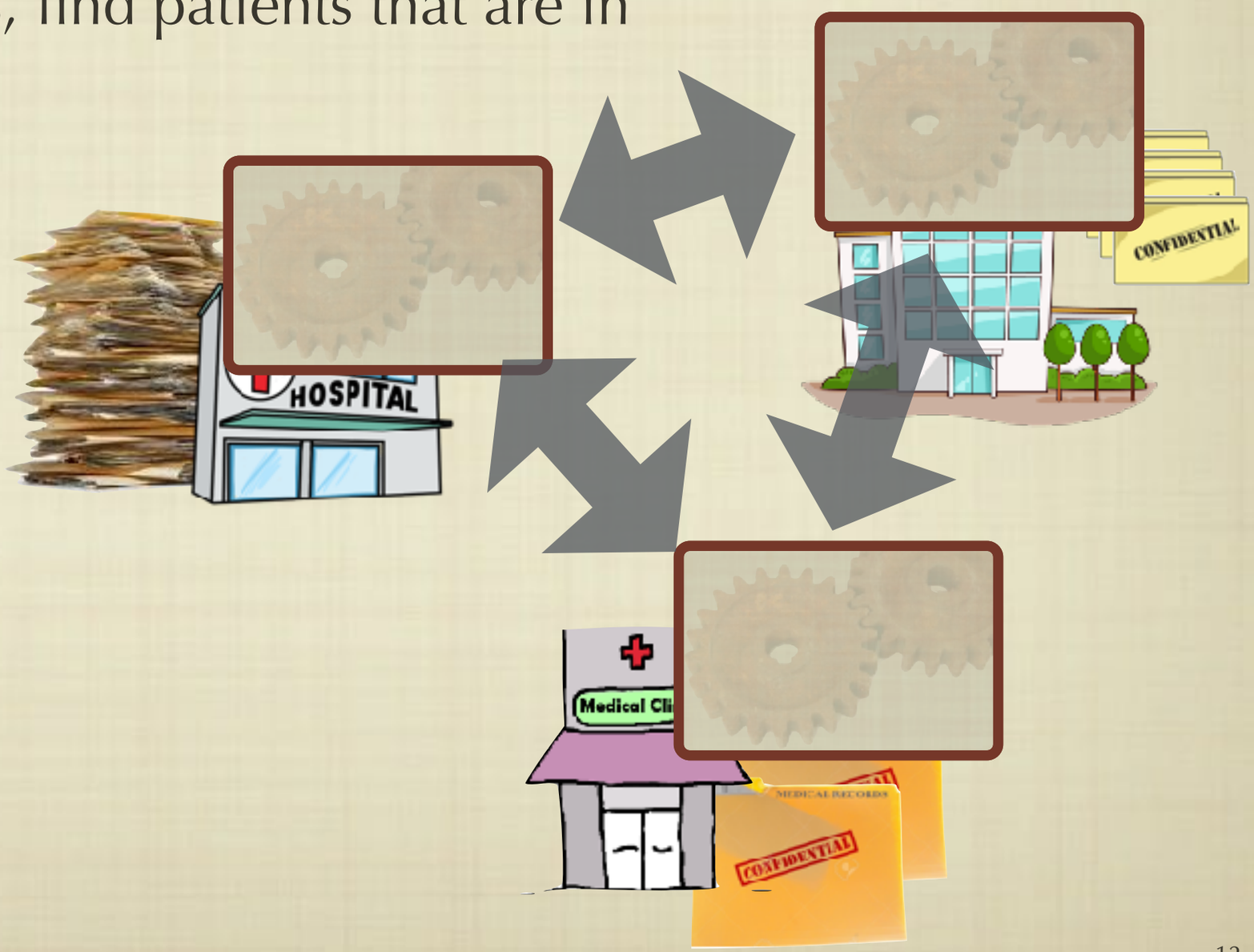
Example: Hospital Data

- How can the hospitals trust the computations?
- Researcher sends **verified computation** sent to hospitals
  - Hospitals check that computation respects their security/privacy requirements
  - Secure execution environments (e.g., Intel SGX) at hospitals give researcher assurance that hospitals running correct computation



Example: Hospital Data

- Why not just use meta-analysis?
  - Combine results of many small studies to get more robust results
- **Verified Estimation and Balancing**: allow hospitals to balance study participants, find patients that are in multiple datasets, ...



Researcher

Current Research: Remote Secure Storage (Nissim)

- How can organizations securely store sensitive data in the cloud, and issue queries to retrieve records?
- Existing schemes:
  - In theory, strong cryptographic tools may be used (ORAM, FHE, searchable encryption,…)
  - Practically, for efficiency use weakened encryption primitives (e.g., deterministic encryption, order preserving encryption)
  - Privacy and security of theoretical and practical schemes not well understood; Recent attacks on practical implementations
- Our work:
  - Generic reconstruction attacks that apply to all existing (theoretical and practical) implementations (CCS 2017)
  - New model of differentially-private storage
    - Combines ORAM with differential privacy to provably subvert attacks (in submission)
- Planned:
  - Further develop theory
  - Implement and test constructions, examine use of SGX technology.

Current Research: Balance (Honaker)

- *Balance* measures how close datasets are to a randomized controlled trial
  - Different definitions of balance have different notions of what's important for randomization
- Traditional approach: each data owner finds balanced subset of own dataset, publishes causal estimates (unbiased, but small *N*)
- Our approach:
  - 1. Data owners collectively compute balance, choose subsets to achieve balance globally
  - 2. Data owners publish causal estimates from own biased dataset (biased, but larger *N*)
  - 3. Combined result is balanced
- E.g., one owner has only control observations; one has only treatment observations
  - Traditional approach: neither publishes; Our approach: good estimation of treatment effect?
- Experimenting with different balance functions in this shared setting
  - Trade off between sample size gains and computational demands
  - Understand privacy implications of balance functions
  - How to verify computation of balance function
  - Doing analogous work for optimization of parametric statistical models
    - regression, logit, maximum likelihood, etc., over distributed data

Current Research: Verification of DP Programs (Gaboardi)

- Enhancing existing program verification methods for DP
  - Better semantic foundations by using ideas from metric spaces
  - Probabilistic coupling as a formal verification reasoning principle
  - Developing support for recent privacy notions like *concentrated differential privacy*
- Improving existing tools for verifying differential privacy
  - Adapting tools to handle probabilistic inference and learning techniques
  - Adding better support for reasoning about two runs of a program (as required by the DP definition)

Current Research: Core R

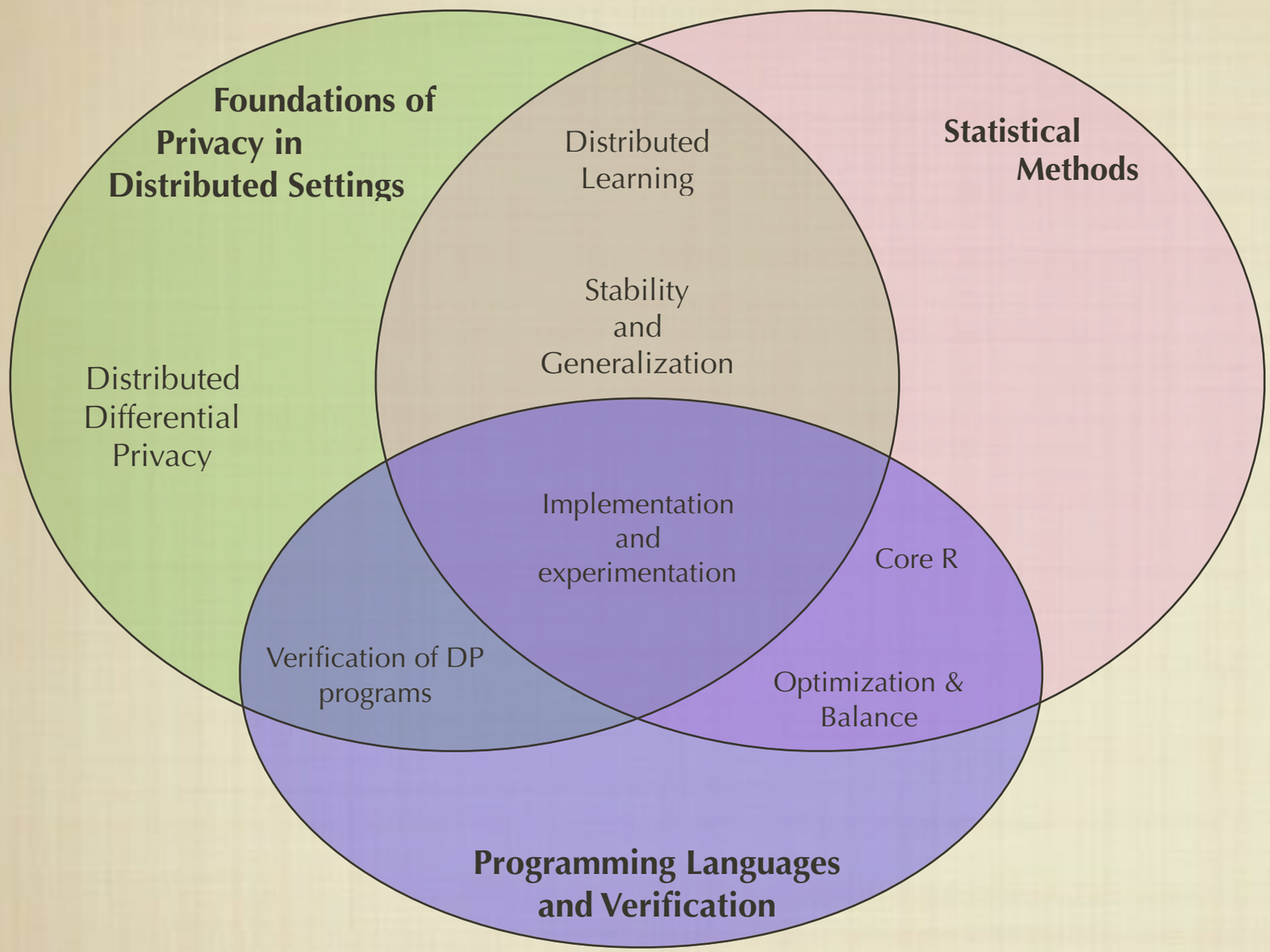
(Gaboardi, Chong)

- Developed a subset of the R programming language that is amenable to formal reasoning (i.e., verification)
  - Have specification of language and interpreter
- Working on translation of Core R to existing tools for verifying differential privacy
  - Extending Core R with primitives for differential privacy
  - Designing domain-specific verification conditions for the translated language

Current Research: Language Support for SGX (Chong)

- Bridge gap between secure execution environments (e.g., Intel SGX) and strong language-level information security guarantees
- Working on providing Haskell-based language runtime in SGX
- Enable expression and verification of security/ privacy guarantees of programs

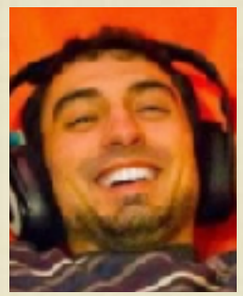
Computing Over Distributed Sensitive Data




Balcer


Brawner

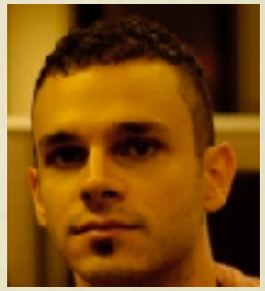

Chong


Farina

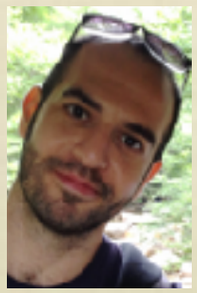

Gaboardi


Gollamudi


Honaker


Kellaris


Nissim


Sheffet


Stemmer


Vadhan

- Questions?
- More info at:

https://privacytools.seas.harvard.edu/computing-over-distributed-sensitive-data