

Sharing Knowledge without Sharing Data

Platforms for resolving the false dichotomy between privacy and utility of information

Azer Bestavros

Computer Science Department
Hariri Institute for Computing
Boston University



Massachusetts Juvenile Justice Policy and Data Board
October 10, 2019

(Yao's) Millionaires' Problem

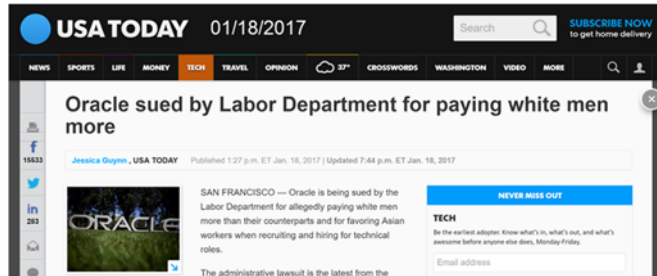
Want to know who is wealthier



Can we reveal the answer without revealing the inputs – not even to an app?

The Labor Department Question

Want to know if companies like Google/Oracle are paying white men more

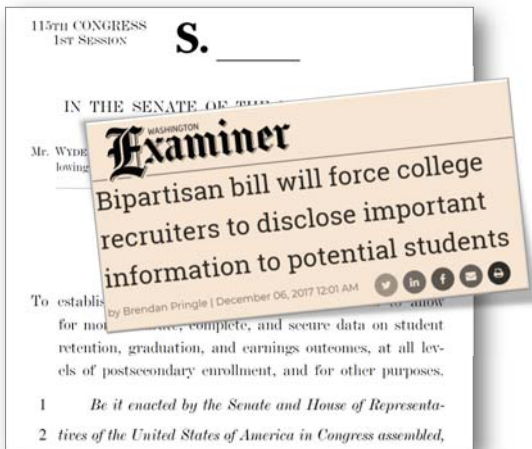


“In a statement, Google said it balked at turning over the private information of employees.”

Can DOL prove (non)compliance without access to sensitive employee records?

The Right to Know Before You Go Question

Want to enable cost-benefit analysis of higher education across colleges and majors




ID	College	Degree	...

ID	Income	...

The Massachusetts Child Advocacy Question

Want to measure educational success for various juvenile cohorts in a DCF database



Leadership Forum
Massachusetts Child Welfare and Juvenile Justice Leadership Forum
Results Statement
10.30.18



Program Population, Program Result
Program Population: All system-involved young people, and those at risk of system involvement in the Commonwealth regardless of race, ethnicity, gender (individual and expression), sexual orientation in the commonwealth...



Program Result: ... are treated fairly, educated, and in permanent, safe, and stable homes with a caring, responsible adult and grow up to be happy, healthy, self-sufficient contributing members of society.

Better off Performance Measures:

1. Equitable increase in Educational Success
Target Population: young people ages 0-22 who are in custody of DCF and/or have a delinquency complaint issued.


Targets: A. Equitable increase school stability,
B. Equitable decrease chronic absenteeism,
C. Equitable decrease school exclusion,
D. Equitable increase the number of children in early education, and
E. Equitable increase academic achievement by:
1. Increasing 3rd and 4th grade literacy,
2. Decreasing dropout rates, and
3. Increasing grade completion

ID	Status	...

ID	Success	...


 The Rafik B. Hariri Institute for Computing and Computational Science & Engineering

Azer Bestavros: Sharing Knowledge without Sharing Data -- On the false choice between privacy and utility of information


5

The answer to all these questions is **YES**

We can derive knowledge (K) from data (x_1, x_2, x_3, \dots) without requiring owners of the data to share it or to trust anything other than mathematics under some assumptions about threats



$$K = f(\text{TOP SECRET CONFIDENTIAL TOP SECRET}, \text{TOP SECRET CONFIDENTIAL TOP SECRET}, \text{TOP SECRET CONFIDENTIAL TOP SECRET}, \dots)$$

 The Rafik B. Hariri Institute for Computing and Computational Science & Engineering

Azer Bestavros: Sharing Knowledge without Sharing Data -- On the false choice between privacy and utility of information

6

Azer Bestavros, Boston University

3

April 9, 2013

WOMEN'S WORKFORCE COUNCIL

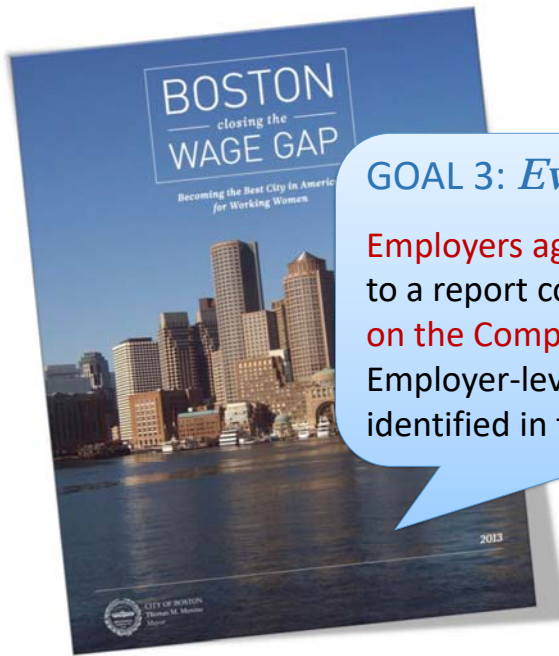
The Women's Workforce Council was established by Mayor Thomas M. Menino on April 9th, 2013— known nationwide as Equal Pay Day. The day marks how far into 2013 women need to work to earn what men earned in 2012. The first of its kind in the country, the Council's mission is to help transform Boston into the best city in the country for working women.



Meeting with Mayor Menino @ BU, July 31, 2014



Published October 31, 2013



100% TALENT The Boston Women's Compact

GOAL 3: Evaluating Success

Employers agree to ... contribute data to a report compiled by a third party on the Compact's success to date. Employer-level data would not be identified in the report.



BOSTON
closing the
WAGE GAP
Becoming the Best City in America
for Working Women

100% TALENT
The Boston Women's Compact

GOAL 3: Evaluating Success
Employers agree to ... contribute data to a **report compiled by a third party** on the Compact's success to date. **Employer-level data would not be identified** in the report.

STATE STREET, EMC², Raytheon, MassMutual FINANCIAL GROUP, BUFFOLK, Putnam INVESTMENTS, Core.com, nationalgrid, MASSACHUSETTS TECHNOLOGY COLLABORATIVE, EVERSURCE ENERGY, Abt, MASFAST, WENTWORTH Institute of Technology, Tech Networks of Boston, WHEELLOCK, The Boston Foundation, Boston Children's Hospital, ROXBURY TECHNOLOGY, BENTLEY, JLL, MASSACHUSETTS WOMEN'S POLITICAL CAUCUS, CHARLES ENERGY CONSULTING

BU The Rafik B. Hariri Institute for Computing and Computational Science & Engineering

Azer Bestavros: Sharing Knowledge without Sharing Data -- On the false choice between privacy and utility of information 9

April 14, 2015

BU Today In the World
Computational Thinking Breaks a Logjam
Hariri Institute helps address Boston's male female pay gap

Business
Mayor Walsh pushes to gather data on gender wage gap

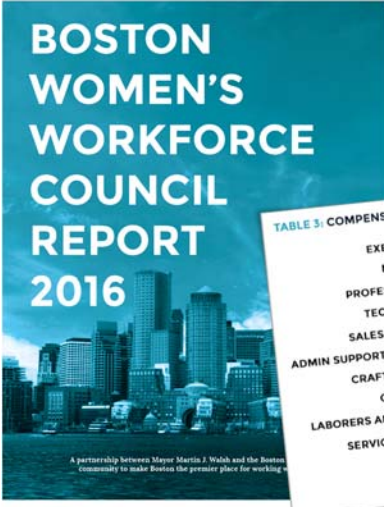
True data A + random mask A = masked data A
True data B + random mask B = masked data B
masked data A + masked data B = masked aggregate data
masked aggregate data + random mask A + random mask B = true aggregate data

Public-key Encrypted Storage only Trusted Party has key; no one else (including the Server) can read the content of this data

BU The Rafik B. Hariri Institute for Computing and Computational Science & Engineering

Azer Bestavros: Sharing Knowledge without Sharing Data -- On the false choice between privacy and utility of information 10

January 5, 2017



BOSTON WOMEN'S WORKFORCE COUNCIL REPORT 2016

A partnership between Mayor Martin J. Walsh and the Boston community to make Boston the premier place for working women.

"We collected data regarding 112,600 employees, which represents 11% of the Greater Boston workforce and almost \$11 billion in annual earnings."

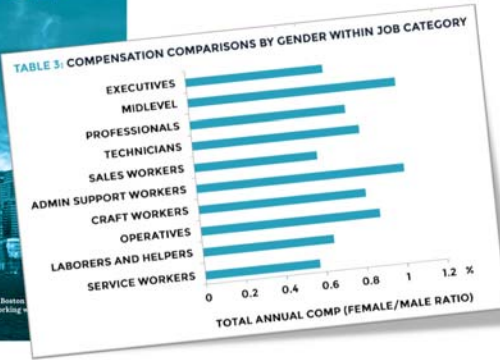



TABLE 3: COMPENSATION COMPARISONS BY GENDER WITHIN JOB CATEGORY

Job Category	Total Annual Comp (Female/Male Ratio)
EXECUTIVES	0.85
MIDLEVEL	0.75
PROFESSIONALS	0.70
TECHNICIANS	0.65
SALES WORKERS	0.60
ADMIN SUPPORT WORKERS	0.55
CRAFT WORKERS	0.50
OPERATIVES	0.45
LABORERS AND HELPERS	0.40
SERVICE WORKERS	0.35

Even in liberal Boston, there's a gender wage gap

By **Katie Johnston**
GLOBE STAFF JANUARY 05, 2017

Working women in Greater Boston make 77 cents on the dollar compared to men — a gender wage gap that echoes the national average — according to a report released Thursday by the Boston Women's Workforce Council.




The Rafik B. Hariri Institute for Computing and Computational Science & Engineering

Azer Bestavros: Sharing Knowledge without Sharing Data -- On the false choice between privacy and utility of information

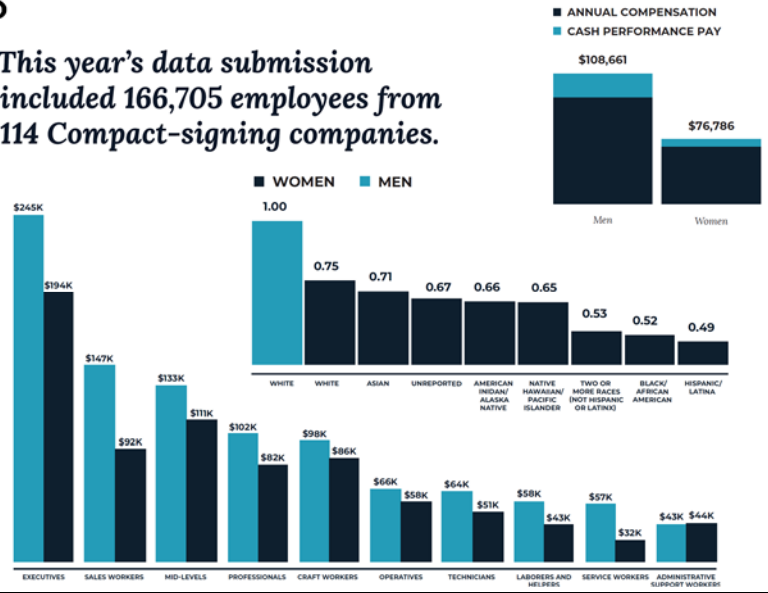
14

January 31, 2018



BOSTON WOMEN'S WORKFORCE COUNCIL REPORT 2017


"This year's data submission included 166,705 employees from 114 Compact-signing companies."



ANNUAL COMPENSATION (Dark Blue) | **CASH PERFORMANCE PAY** (Light Blue)

Men: \$108,661 | Women: \$76,786

Job Category	Women Ratio	Men Ratio
EXECUTIVES	1.00	0.75
SALES WORKERS	0.75	0.71
MID-LEVELS	0.71	0.67
PROFESSIONALS	0.67	0.66
CRAFT WORKERS	0.66	0.65
OPERATIVES	0.65	0.53
TECHNICIANS	0.53	0.52
LABORERS AND HELPERS	0.52	0.49
SERVICE WORKERS	0.49	0.49
ADMINISTRATIVE SUPPORT WORKERS	0.49	0.49



The Rafik B. Hariri Institute for Computing and Computational Science & Engineering

Azer Bestavros: Sharing Knowledge without Sharing Data -- On the false choice between privacy and utility of information

15

“This [is] the first time actual wage data has been reported both anonymously and voluntarily. This is a groundbreaking moment in tackling the gender gap.”
Mayor Marty Walsh



2014



2017

“[MPC] has never been used for public good. Here, we’re beginning to show how to use this sophisticated computer science research for public programs.”
BWWC co-chair Evelyn Murphy



2015

The Boston Globe

The congresswoman, who had signed onto a bill addressing income disparity between men and women, was impressed by the relevance he outlined. “It’s linking it back for the members of Congress,” Clark said. “Nobody would think, oh, the Paycheck Fairness Act, how is that tied into NSF funding?”





The Rafik B. Hariri Institute for Computing and Computational Science & Engineering

Azer Bestavros: Sharing Knowledge without Sharing Data -- On the false choice between privacy and utility of information

16

Multi-Party Computation (MPC)

What is it?

- Given multiple parties p_1, p_2, \dots, p_n each with private data x_1, x_2, \dots, x_n
- Parties engage in computing a function $f(x_1, x_2, \dots, x_n)$
- Nothing is revealed about the inputs beyond what the output of f reveals
- What f leaks is an orthogonal question, e.g., the realm of “differential privacy”

State of the Art

- Theory known since 1979, with Shamir’s “How to share a secret”
 - Frameworks and libraries increasingly available over the last few years ...
 - Experience with real use cases at scale is limited
 - Deployments are not easily portable
- ← We are changing that

← We are changing that

What is Multi-Party Computation (MPC)?


Boston University
 Published on Feb 19, 2019
 



The Rafik B. Hariri Institute for Computing and Computational Science & Engineering

Azer Bestavros: Sharing Knowledge without Sharing Data -- On the false choice between privacy and utility of information



18

How Does it Work?

Data Owners



Analyst


CITY of BOSTON


Happy to help you, but there is no way I am going to tell anybody what my salary is.

Happy to help you, but there is no way I am going to tell anybody what my salary is.

I would love to know the difference between your salaries. Can you please share them with me so that I may figure that out?

I have a solution for you!





The Rafik B. Hariri Institute for Computing and Computational Science & Engineering


Azer Bestavros: Sharing Knowledge without Sharing Data -- On the false choice between privacy and utility of information

19

How Does it Work?




\$9




\$7

=




10

+




11

=




4


+



4

Each one of the two servers has one share of each secret






The Rafik B. Hariri Institute for Computing and Computational Science & Engineering

Azer Bestavros: Sharing Knowledge without Sharing Data -- On the false choice between privacy and utility of information


20

How Does it Work?




10

-




4




11

-




4


Compute on shares to get secret-shared result...




6



8






The Rafik B. Hariri Institute for Computing and Computational Science & Engineering

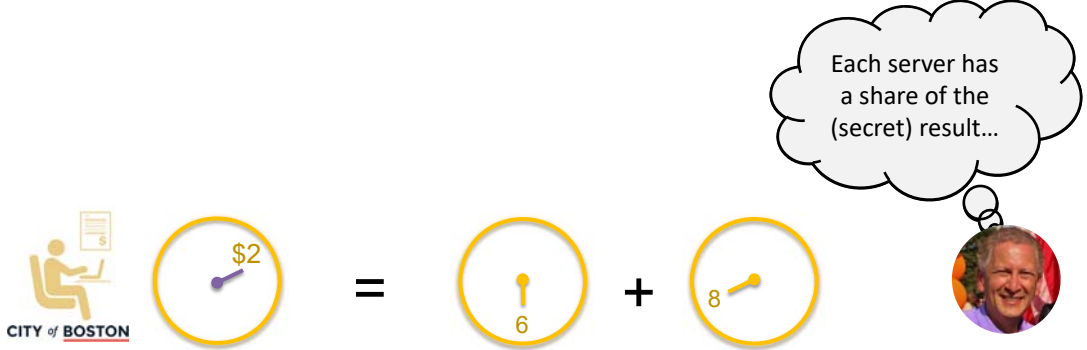
Azer Bestavros: Sharing Knowledge without Sharing Data -- On the false choice between privacy and utility of information


21

How Does it Work?



Service Providers



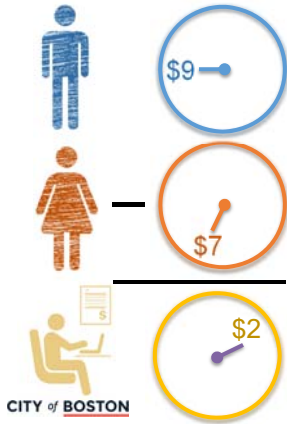


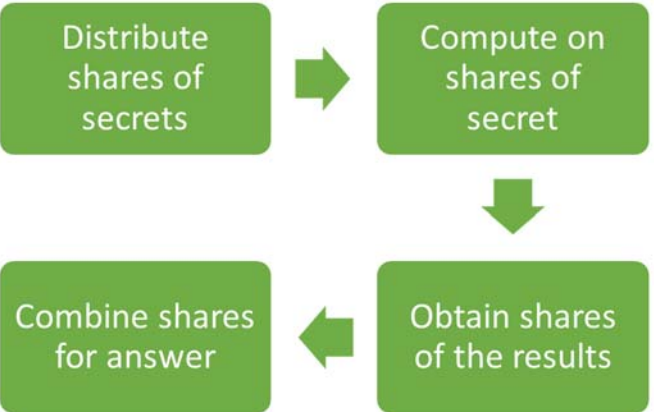
The Rafik B. Hariri Institute for Computing and Computational Science & Engineering


Azer Bestavros: Sharing Knowledge without Sharing Data -- On the false choice between privacy and utility of information

22

How does it work?





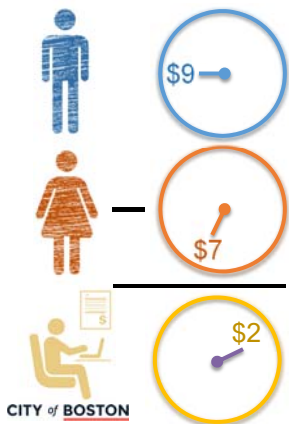


The Rafik B. Hariri Institute for Computing and Computational Science & Engineering

Azer Bestavros: Sharing Knowledge without Sharing Data -- On the false choice between privacy and utility of information

24

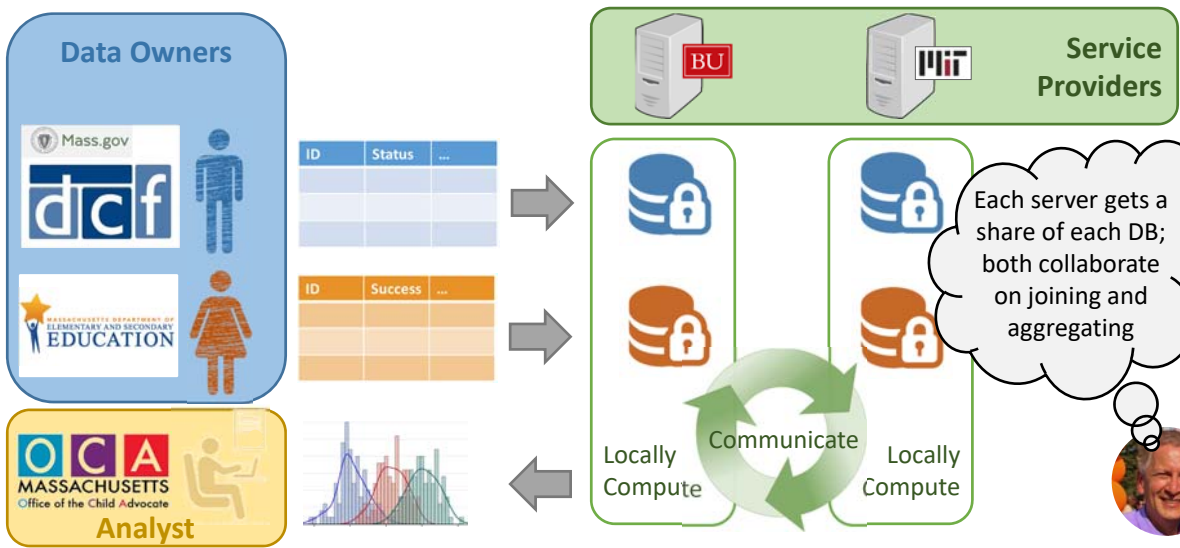
How Does it Work?



In a nutshell, MPC is the collaborative analysis of multiple silo-ed data sets that are never communicated nor trusted to any central authority or database

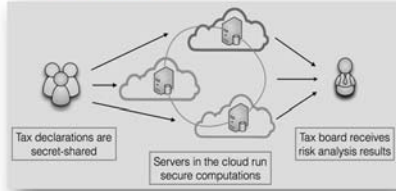


Outsourced MPC Architecture

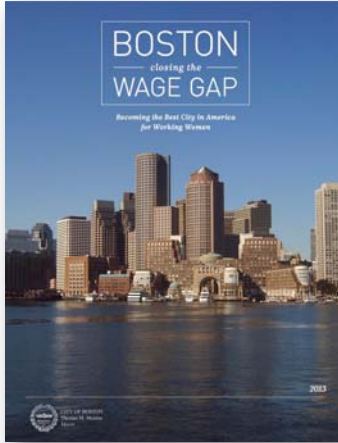


Is anybody else using MPC?

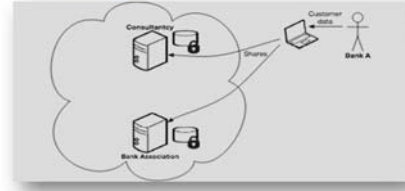
Cybernetica / VAT tax audits



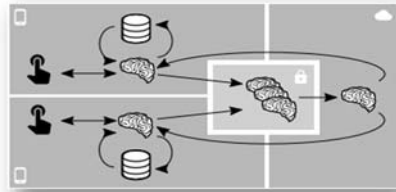
BU / Pay equity in Boston



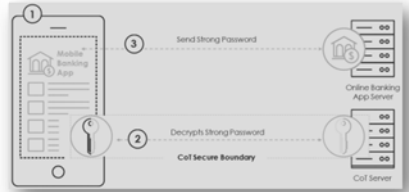
Partisia / Rate credit of farmers



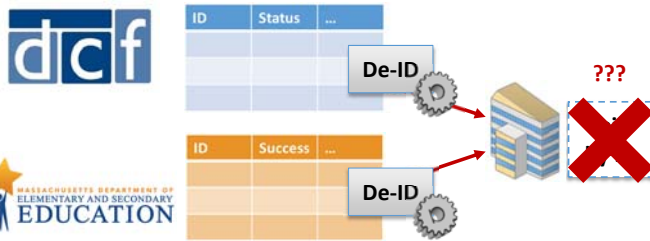
Google / Federated machine learning



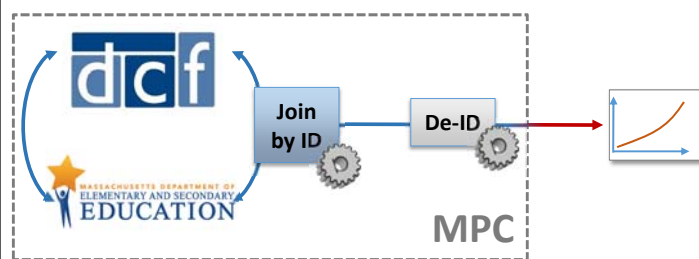
Unbound / Protect cryptographic keys



Why use MPC as opposed to anonymization?



- With de-identification, privacy protections are applied too early in analysis pipeline



- With MPC, data is never shared prematurely
- MPC ⇒ detailed analysis that is privacy compatible

Isn't MPC the same as Blockchain?

MPC

- **Strong confidentiality:** data encoded and split in a privacy-preserving way
- **Strong integrity:** distributed general purpose data analysis + incentives for long-term accuracy and stability
- **Good availability:** tolerates adversarial behaviors to a point, and fails safely

Blockchain

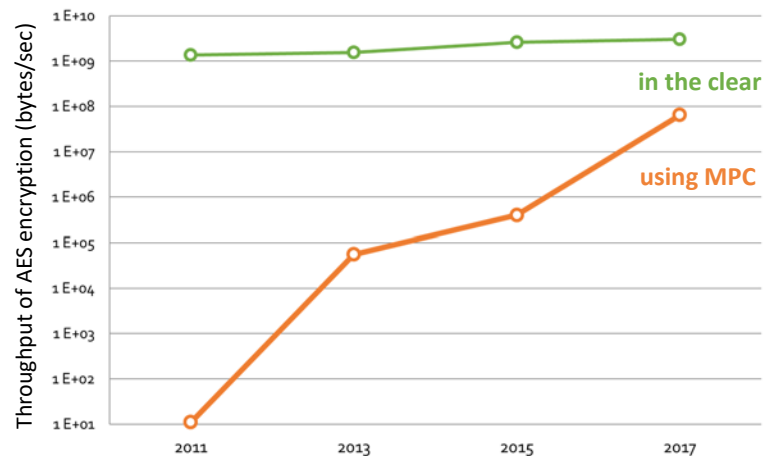
- **Poor confidentiality:** data copied in the clear across the internet
- **Strong integrity:** distributed general purpose data analysis + incentives for long-term accuracy and stability
- **Strong availability:** persists through attacks from the distributed servers

How about MPC's performance?

Benchmark: Outsourced Encryption

How fast could two parties jointly encrypt a secretly-shared message via MPC compared to doing it in the clear (which means trusting cloud with private message and key)?

→ Performance shouldn't be an impediment to deploying and using MPC.

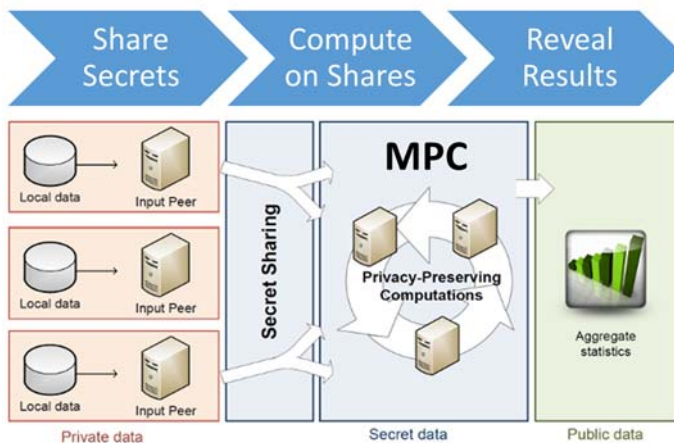


How commoditized is the technology?

We're getting there! Secret Sharing and MPC as a Service

$$f(x) = s + r_1x^1 + r_2x^2 + \dots + r_ix^i + \dots$$

$$\begin{aligned} s_1 &= f(1) \\ s_2 &= f(2) \\ s_3 &= f(3) \\ \dots &= \dots \\ s_i &= f(i) \\ \dots &= \dots \end{aligned}$$



Is the technology proprietary?

BU Open-source MPC Libraries

JIFF: JavaScript Implementation of Federated Functionalities

Library for building web-based applications using secure multi-party computation

<https://github.com/multiparty/jiff>

Web-MPC

JavaScript application for user-friendly privacy-preserving web-based data aggregation

<https://github.com/multiparty/web-mpc>

Conclave Workflow Manager

Compiler that optimizes relational queries to be executed under MPC by factoring it into (1) scalable, local, cleartext processing workflows using backends such as Apache Spark, and (2) isolated MPC workflows that utilize existing MPC backend frameworks

<https://github.com/multiparty/conclave>

How easy is it to use in practice?

Familiar (same workflow)

A screenshot of a data table with a complex header structure. The columns are grouped by race/ethnicity: Hispanic or Latinx, White, Black/African American, Native Hawaiian or Pacific Islander, Asian, American Indian/Alaska Native, and Two or More Races (Not Hispanic or Latinx). Each group has sub-columns for Female and Male. The rows list job categories like Executive/Senior Level Officials and Managers, etc.

Reviewable (quality control)

A screenshot of a simplified data table. The columns are grouped by race/ethnicity: Hispanic or Latinx, White, and Black/African American. Each group has sub-columns for Female and Male. The 'Executive/Senior Level Officials and Managers' row shows a value of 100000 in the Female column, with a yellow warning box that says 'Warning: Data is too big' and 'Are you sure this value is correct?'.

Accessible (comprehensible)



Transparent (open source)

A screenshot of a GitHub repository page for 'data-aggregator / client / script / ssCreate.js'. It shows the branch 'master', a commit by 'frederickjansen' with the message 'Update to use native pki and forge instead of jsencrypt', 4 contributors, and file statistics: Executable File, 322 lines (288 sloc), 11.6 KB.

A screenshot of the Boston Women's Workforce Council website. The page title is 'Boston Women's Workforce Council' with the subtitle '100% Talent Data Submission'. It features logos for the council and Boston University. Below is a table titled 'Number Of Employees' with a header structure similar to the first table, but with a simplified column layout: Hispanic or Latinx, White, Black/African American, Native Hawaiian or Pacific Islander, Asian, American Indian/Alaska Native, Two or More Races (Not Hispanic or Latinx), and Unreported. Each race group has Female and Male sub-columns. The rows list job categories like Executive/Senior Level Officials and Managers, etc.

Total Annual Compensation (Dollars)

	Hispanic or Latinx		White		Black/African American		Native Hawaiian or Pacific Islander		Asian		American Indian/Alaska Native		Two or More Races (Not Hispanic or Latinx)		Unreported	
	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
Executive/Senior Level Officials and Managers																
First/Mid-Level Officials and Managers																
Professionals																
Technicians																
Sales Workers																
Administrative Support Workers																
Craft Workers																
Operatives																
Laborers and Helpers																
Service Workers																

BU The Rafik B. Hariri Institute for Computing and Computational Science & Engineering

Total Annual Cash Performance Pay (Dollars)

	Hispanic or Latinx		White		Black/African American		Native Hawaiian or Pacific Islander		Asian		American Indian/Alaska Native		Two or More Races (Not Hispanic or Latinx)		Unreported	
	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
Executive/Senior Level Officials and Managers																
First/Mid-Level Officials and Managers																
Professionals																
Technicians																

Total Length of Service (Months)

	Hispanic or Latinx		White		Black/African American		Native Hawaiian or Pacific Islander		Asian		American Indian/Alaska Native		Two or More Races (Not Hispanic or Latinx)		Unreported	
	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male	Female	Male
Executive/Senior Level Officials and Managers																
First/Mid-Level Officials and Managers																
Professionals																
Technicians																

Contribute

BU The Rafik B. Hariri Institute for Computing and Computational Science & Engineering

What other apps have you considered?

Collective Intelligence in Competitive Settings

- Banking and Finance: Multi-institutional systemic risk assessment
- Data Markets: Valuation of marginal utility of data products
- Plausible Deniability: Analytics over possibly “toxic” data
- Information Brokerage: Business and marketing Intelligence
- E-Commerce: Analytics over segmented proprietary data assets
- Sharing Economy: Personalization across multiple service providers

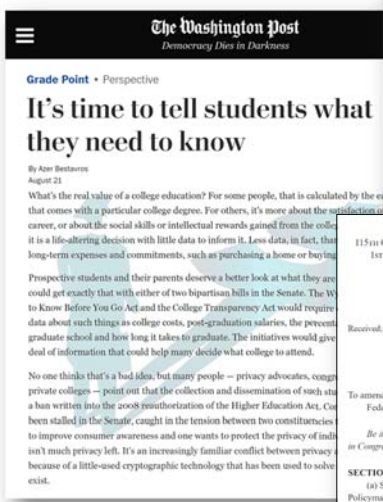
Social Good in Public Settings

- Privacy-preserving census and surveys
- Basic and applied research in healthcare, education, sociology, ...
- Multiagency analytics for evidence-based policy making
- Transparency for corporate and government operations
- Compliance testing/reporting for trade associations
- Private/fair Reporting of sexual harrasement/abuse in workplace



November, 2017 ++

"We are excited to see legislation promoting the use of multi-party computation (MPC) in formulating sound public policy. Boston University's successful collaboration with the City of Boston and the Boston Women's Workforce Council brought this technology into practice to maintain data privacy while gaining insight into an important societal issue -- potential wage inequality in private industry. Such applications demonstrate that MPC can bring enormous value to policymakers at all levels of government."
 -- Azer Bestavros (on behalf of the team from BU)



October 8, 2019

HOT OFF THE PRESS

Datavant Collaborates with Boston University to Pioneer the Use of Multi-Party Computing for Sharing Insights from Health Data

Oct 08, 2019, 09:00 ET

BOSTON and SAN FRANCISCO, Oct. 8, 2019 /PRNewswire/ -- Datavant, the leader in helping healthcare organizations connect and share health data, today announced that it will be a founding member of the Boston University (BU) Data Privacy Collaborative, which aims to bring privacy-preserving technologies such as multi-party computing (MPC) to bear on real-world applications. Through the partnership, Datavant intends to design MPC-enabled approaches for sharing insights from health data drawn from multiple sources.

Sign in Get started

Advancing the Frontier of Privacy-Preserving Technology in Healthcare

Aneesh Kulkarni Follow

Oct 8 · 2 min read

Datavant's mission is to connect the world's health data, and make it accessible for valuable research and analytics. The traditional approach of aggregating data creates a direct tradeoff between individual privacy and analytical capability. With health data, this isn't good enough. The stakes are high — data-driven decisions can save lives, just as data breaches can ruin them. The only way we can achieve our mission at scale is by simultaneously increasing the connectivity of health data, and also the privacy of individual health records. This will demand more from our technology.

This morning, we announced Datavant's collaboration with Boston University (PRNewswire) to pioneer the use of multi-party computing to share insights from health data. Cutting-edge privacy-preserving technologies such as Secure Multiparty Computation (SMC), Homomorphic Encryption, and Differential Privacy enable the utilization of data to be safer than ever before. SMC and homomorphic encryption enable secure

The Rafik B. Hariri Institute for Computing and Computational Science & Engineering

Azer Bestavros: Sharing Knowledge without Sharing Data -- On the false choice between privacy and utility of information

41

How does MPC impact regulation/disclosure?

Status Quo: Disclose extent of use and/or regulate what data is made accessible to whom (e.g., HIPAA and FERPA). Otherwise trust NDAs!

Implications:

- Allow new uses that are consistent with multiple regulations
- Allow new uses beyond the artificiality of restricting access
- Forces lawmakers to think about the purpose of regulation
- Enables purposeful transparency with implications on auditing...

The Rafik B. Hariri Institute for Computing and Computational Science & Engineering

Azer Bestavros: Sharing Knowledge without Sharing Data -- On the false choice between privacy and utility of information

47

What are the legal aspects of using MPC?

MPC



Disclosure: Is release of sensitive data?

- FERPA prohibits "improper disclosure of personally identifiable information derived from education records without the prior written authorization for any use or disclosure for treatment, payment or health care services not otherwise required by the Privacy Rule."

MPC side-steps disclosure issues because it allows entities not authorized to view the data to compute over it.

Disclaimer: I am not a lawyer, but I talk to BU School of Law colleagues

- **Use: Are subjects informed?**

- This is about adherence to the terms of service (e.g., as Facebook) are held accountable if they deviate from the terms of use that they develop for using/sharing of data.



What are the liability considerations for MPC?

MPC



Collective Trust:

- A key assumption is that not all parties will collude (and if they do, they are not the party). Disclosure could be violated if service providers collude.
- MPC is about distributing computation over multiple (independent) service providers, thus disclosing the data to any single party.

More service providers implies better security and better privacy!

- **Privacy requires Security:**

- A key assumption is that not all service providers are honest and not all are an adversary. MPC does not obviate the need for security because it is harder for an adversary to collude over multiple (independent) service providers.



Takeaway: We can have it both ways

We can derive knowledge (K) from data (x_1, x_2, x_3, \dots) without requiring owners of the data to share it or to trust anything other than mathematics under some assumptions about threats

$$K = f(\text{TOP SECRET CONFIDENTIAL FOR SECURITY}, \text{TOP SECRET CONFIDENTIAL FOR SECURITY}, \text{TOP SECRET CONFIDENTIAL FOR SECURITY}, \dots)$$

When it comes to data and computation over data, we need to rethink our notions of ownership, custody, jurisdiction, sharing, disclosure, liability, and introduce new ones such as collusion.

The Path Forward: Our Answers

How can we balance the need for transparency and exploration with fairness and sensitivity to users?

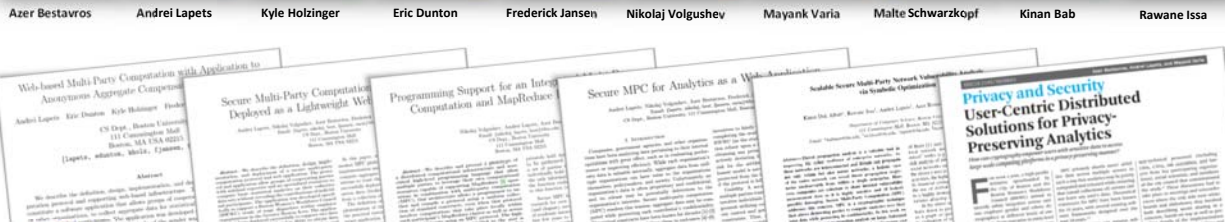
Fix the narrative! Deconstruct the false dichotomy between utility and privacy of data, and between transparency and confidentiality.

How do we ensure that individuals and communities can trust these systems?

Empathize! Adapt technology to how people work as opposed to adapt people to how technology works. Do not change the workflow!

Acknowledgments: It takes a village!

www.multiparty.org



BU The Rafik B. Hariri Institute for Computing and Computational Science & Engineering

Azer Bestavros: Sharing Knowledge without Sharing Data -- On the false choice between privacy and utility of information

53

leveraging the computational perspective

BOSTON UNIVERSITY

Hariri Institute for Computing

"Leveraging the Computational Perspective in a Data-Driven World for a Better Society"

Website: www.bu.edu/HIC
 Twitter: @BU_Computing
 Facebook: BUcomputing

54