Privacy-preserving Information Sharing: Tools and Applications

Emiliano De Cristofaro

University College London

https://emilianodc.com

Prologue

Privacy-Enhancing Technologies (PETs):

- Increase privacy of users, groups, and/or organizations

PETs often respond to privacy threats

- Protect personally identifiable information
- Support anonymous communications
- Privacy-respecting data processing

Another angle: privacy as an enabler

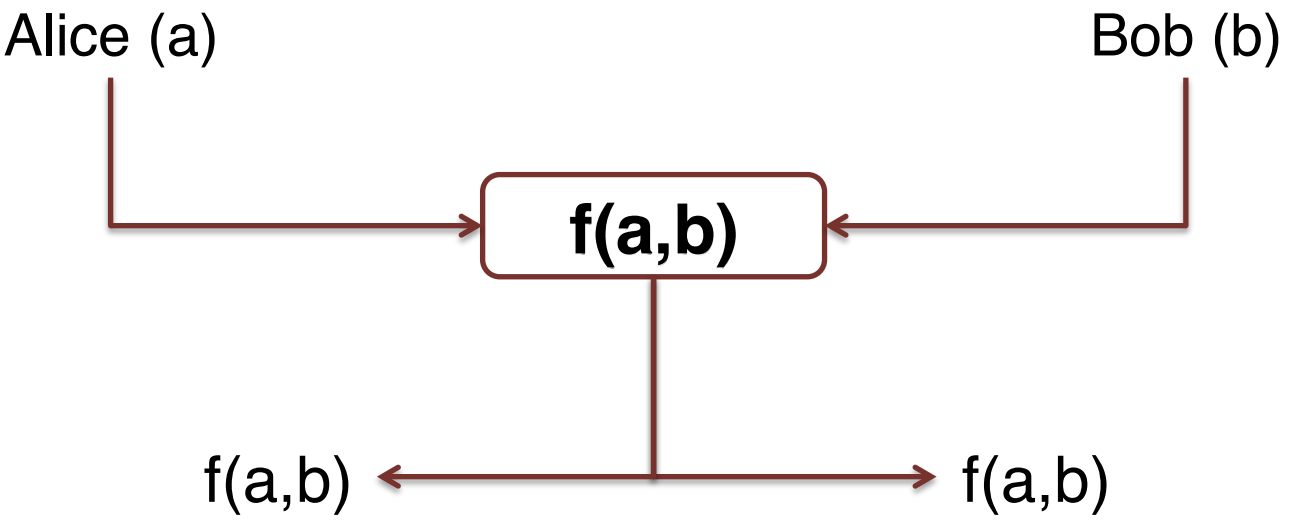
- Actively enabling scenarios otherwise impossible w/o clear privacy guarantees

Sharing Information w/ Privacy

When parties with limited mutual trust willing or required to share information

Only the **required minimum** amount of information should be disclosed in the process

Secure Computation

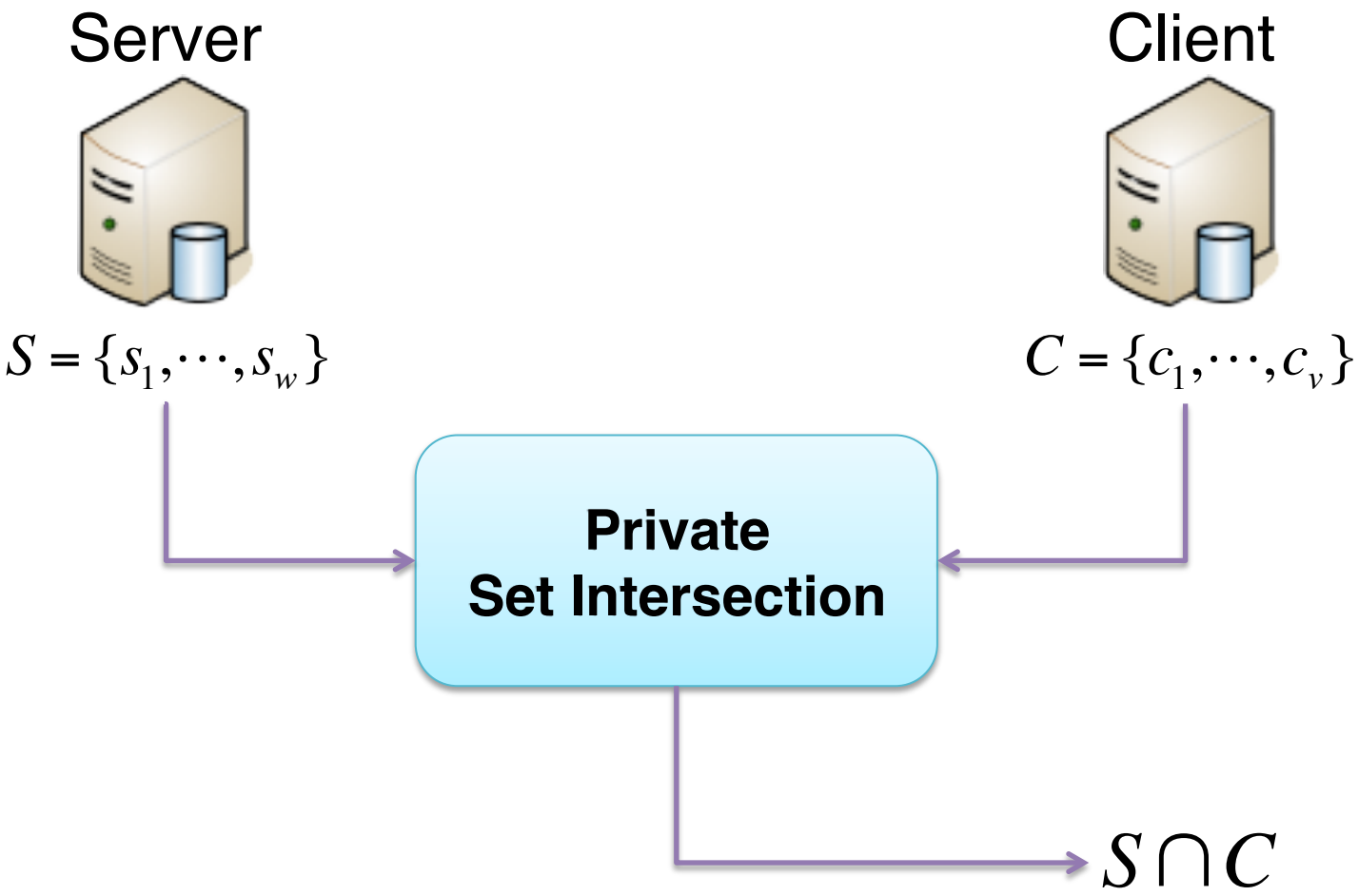


Map information sharing to $f(\cdot,\cdot)$?

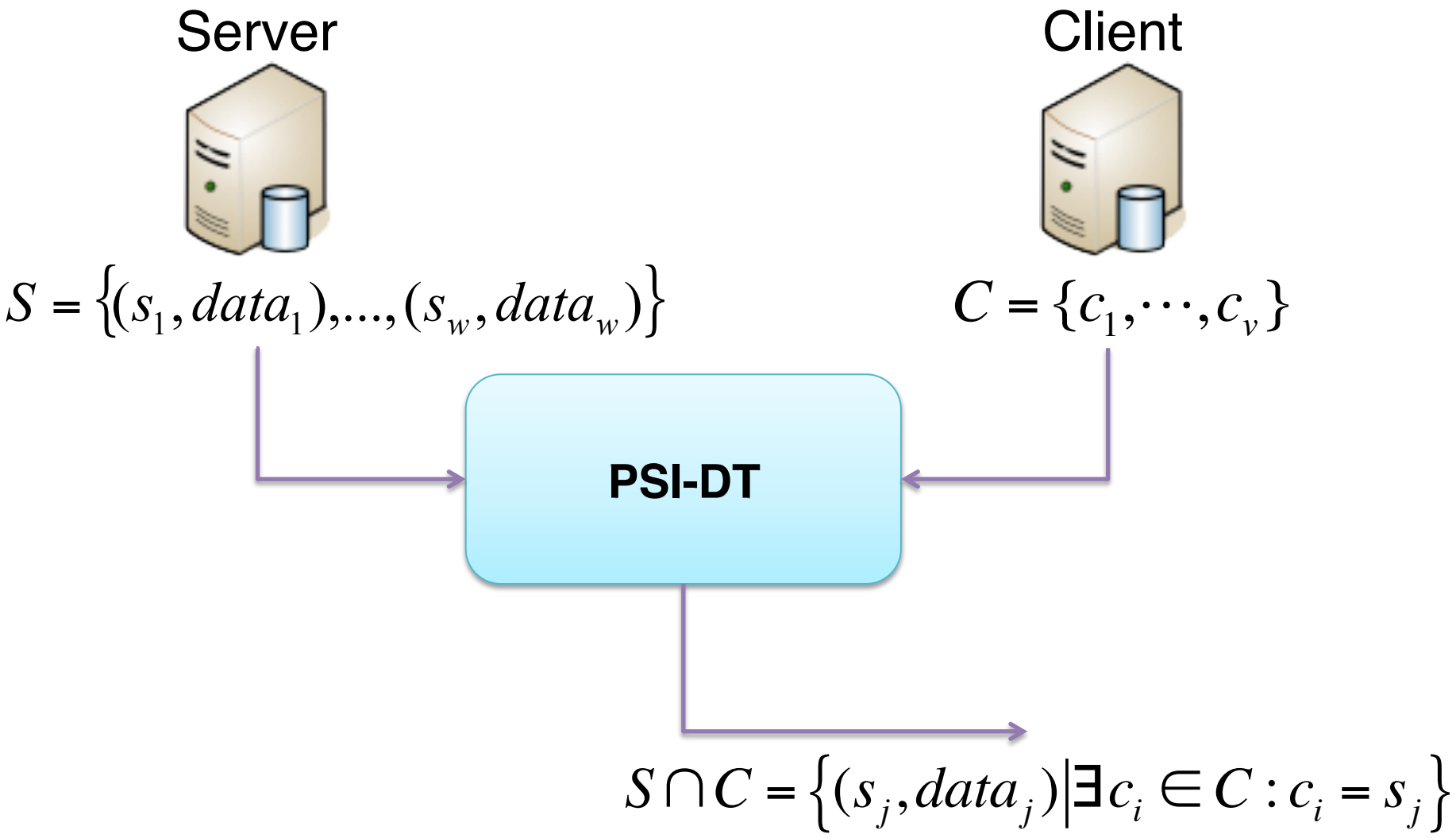
Realize secure $f(\cdot,\cdot)$ efficiently?

Quantify information disclosure from output of $f(\cdot,\cdot)$?

Private Set Intersection (PSI)

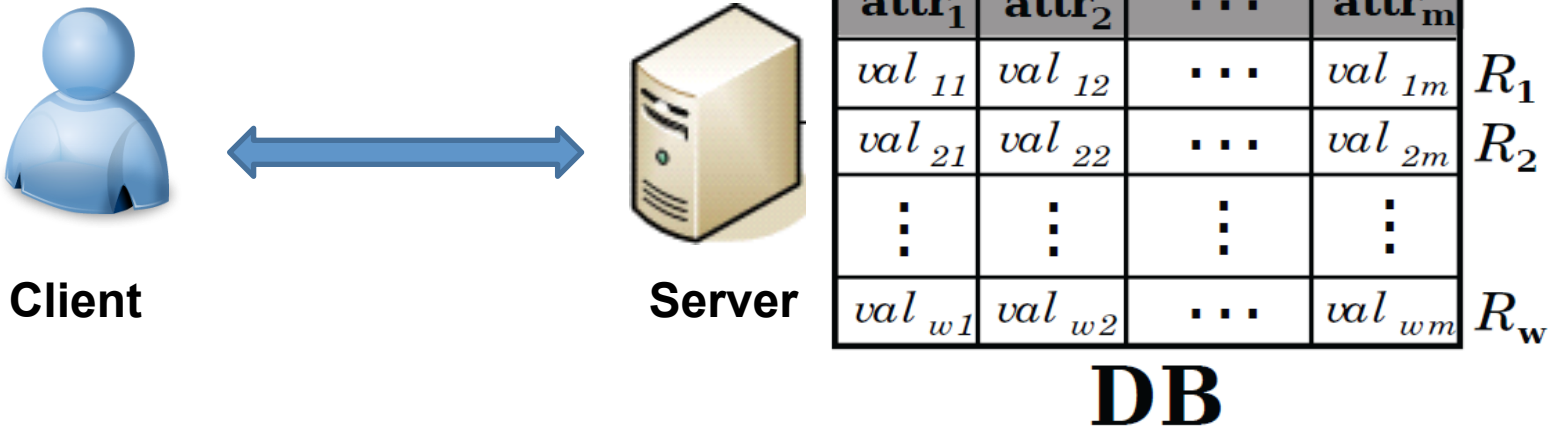


PSI w/ Data Transfer (PSI-DT)

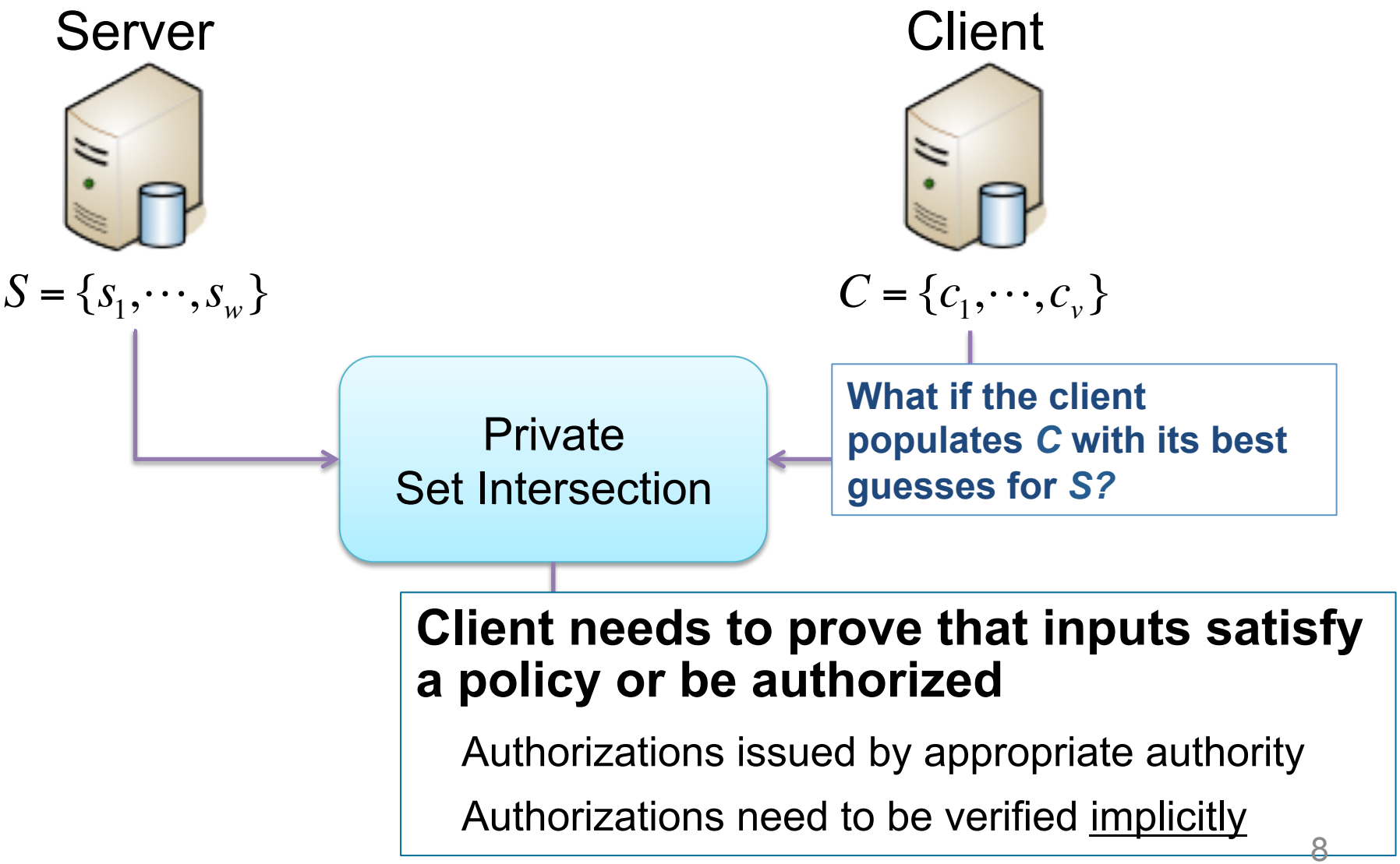


PSI w/ Data Transfer

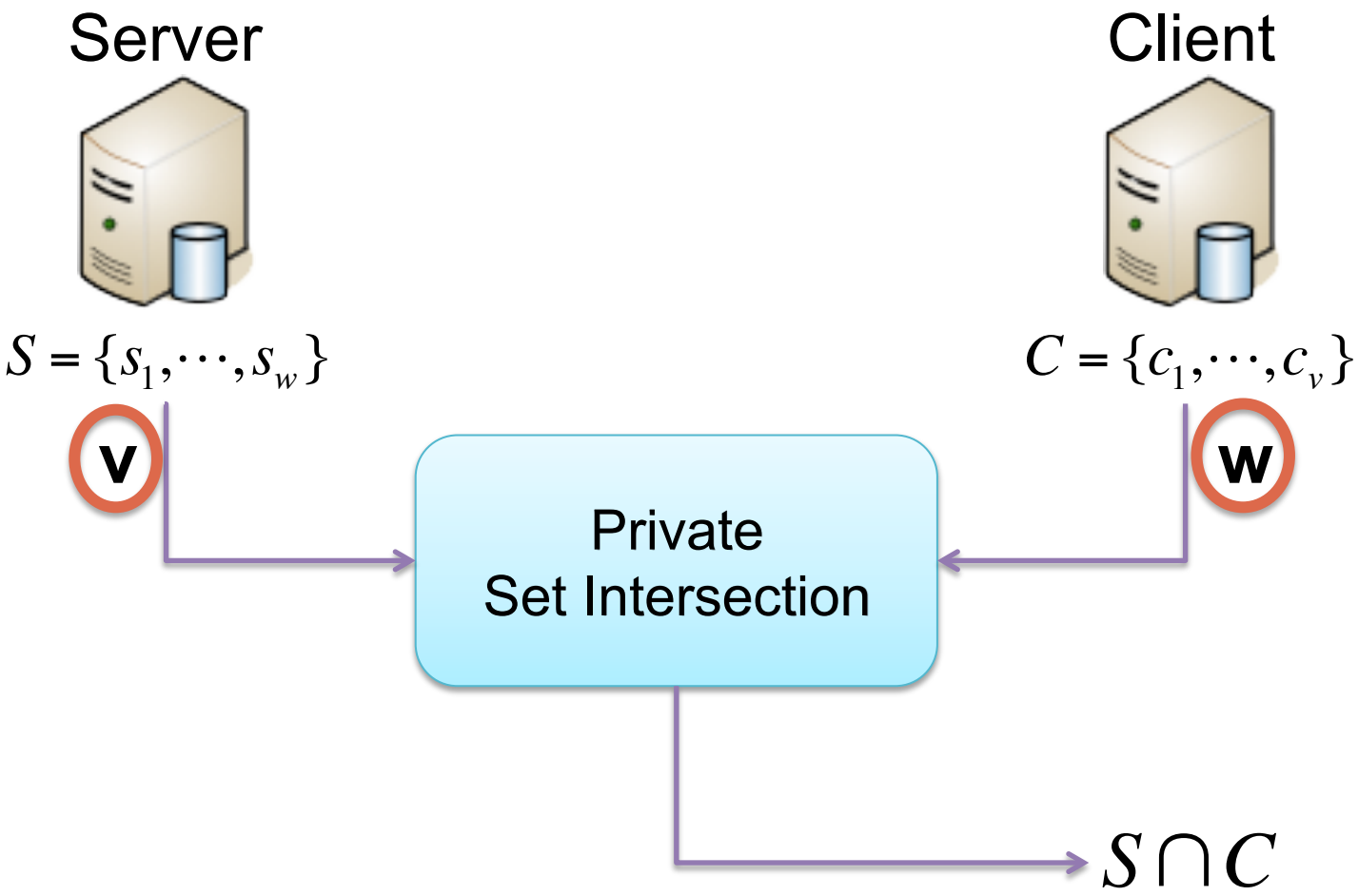
$$\text{SELECT * FROM DB WHERE } (attr_1^* = val_1^* \text{ OR } \cdots \text{ OR } attr_v^* = val_v^*)$$



Authorized Private Set Intersection



Size-Hiding Private Set Intersection



Special-purpose PSI

[DT10]: scales efficiently to very large sets

- First protocol with linear complexities and fast crypto

[DKT10]: extends to arbitrarily malicious adversaries

- Works also for Authorized Private Set Intersection

[DJLLT11]: PSI-based database querying

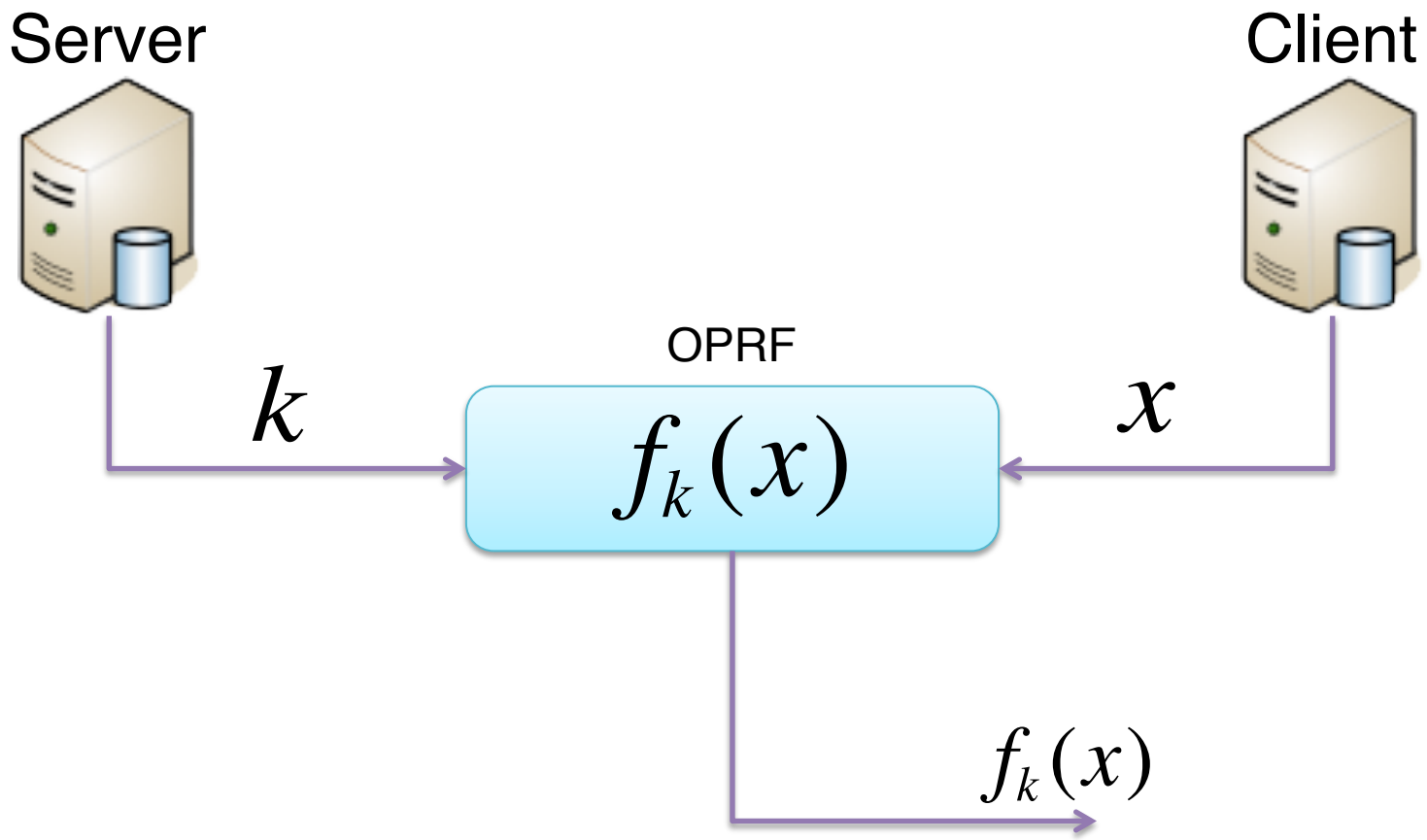
- Won IARPA APP challenge, basis for IARPA SPAR

[DT12]: optimized toolkit for PSI

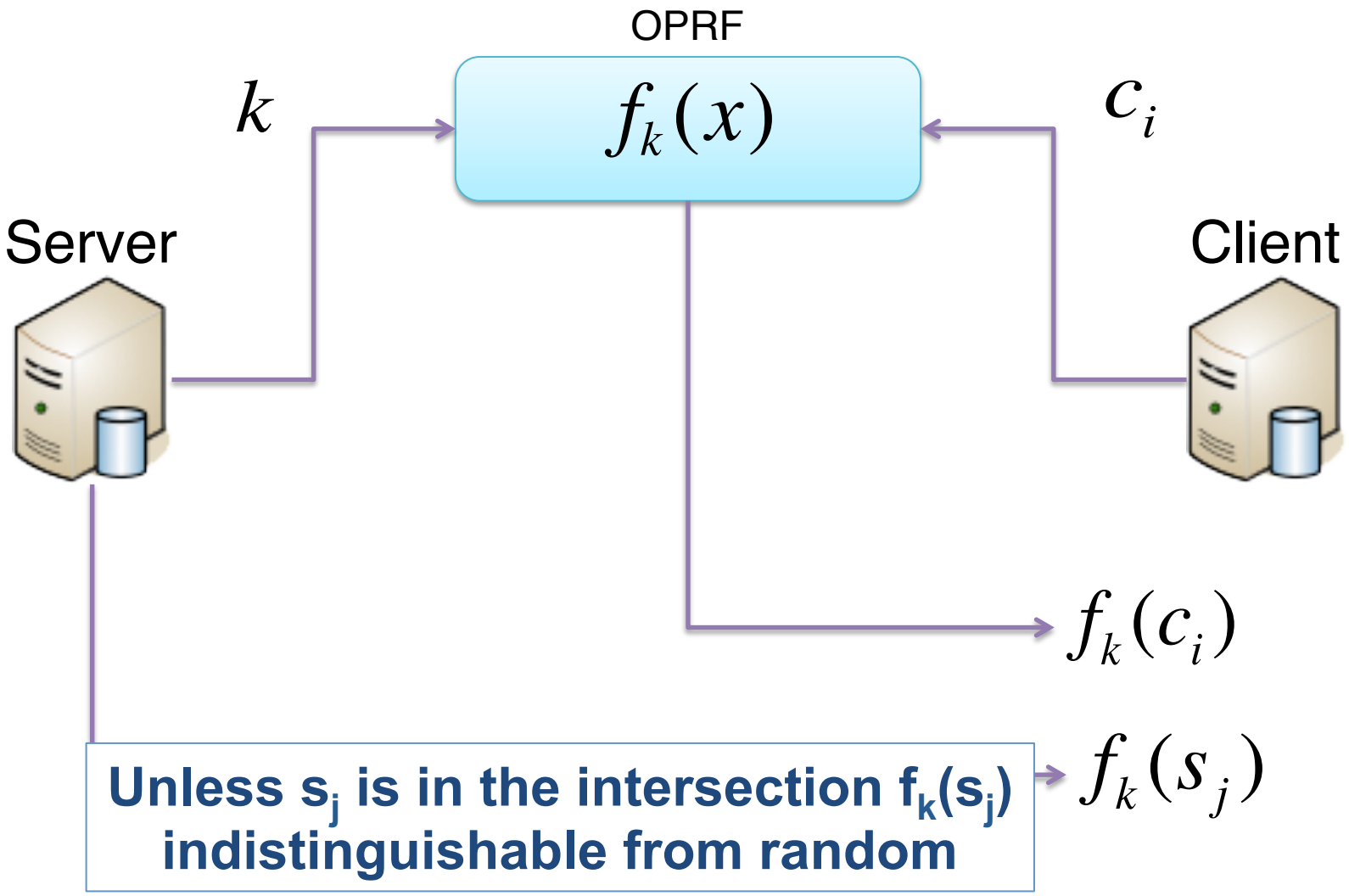
- Privately intersect sets – 2,000 items/sec

[ADT11]: size-hiding PSI

Oblivious Pseudo-Random Functions



OPRF-based PSI



OPRF from Blind-RSA Signatures

RSA Signatures: $(N = p \cdot q, e), d \qquad e \cdot d \equiv 1 \bmod (p-1)(q-1)$

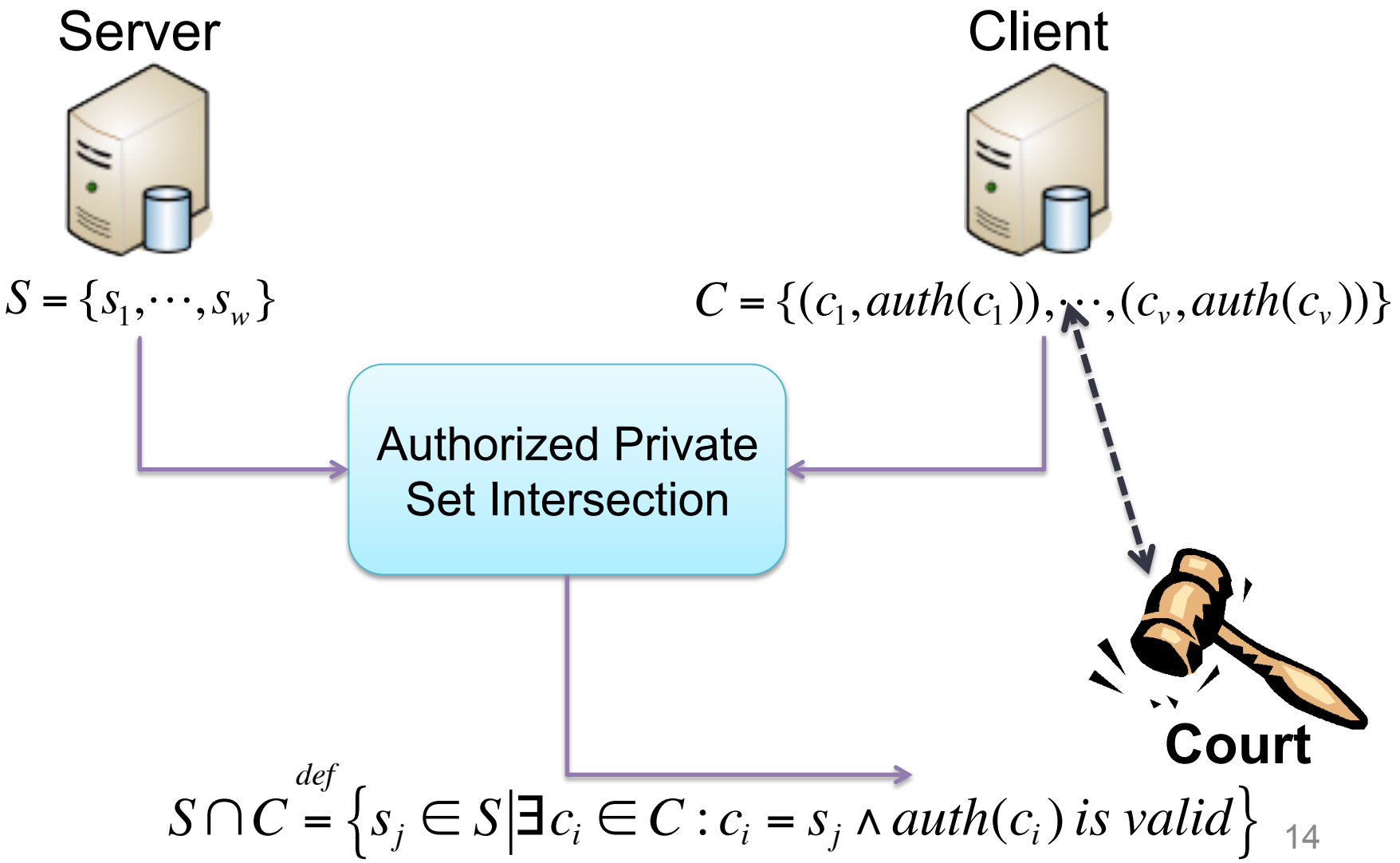
$$Sig_d(x) = H(x)^d \bmod N,$$

$$Ver(Sig(x), x) = 1 \Leftrightarrow Sig(x)^e = H(x) \bmod N$$

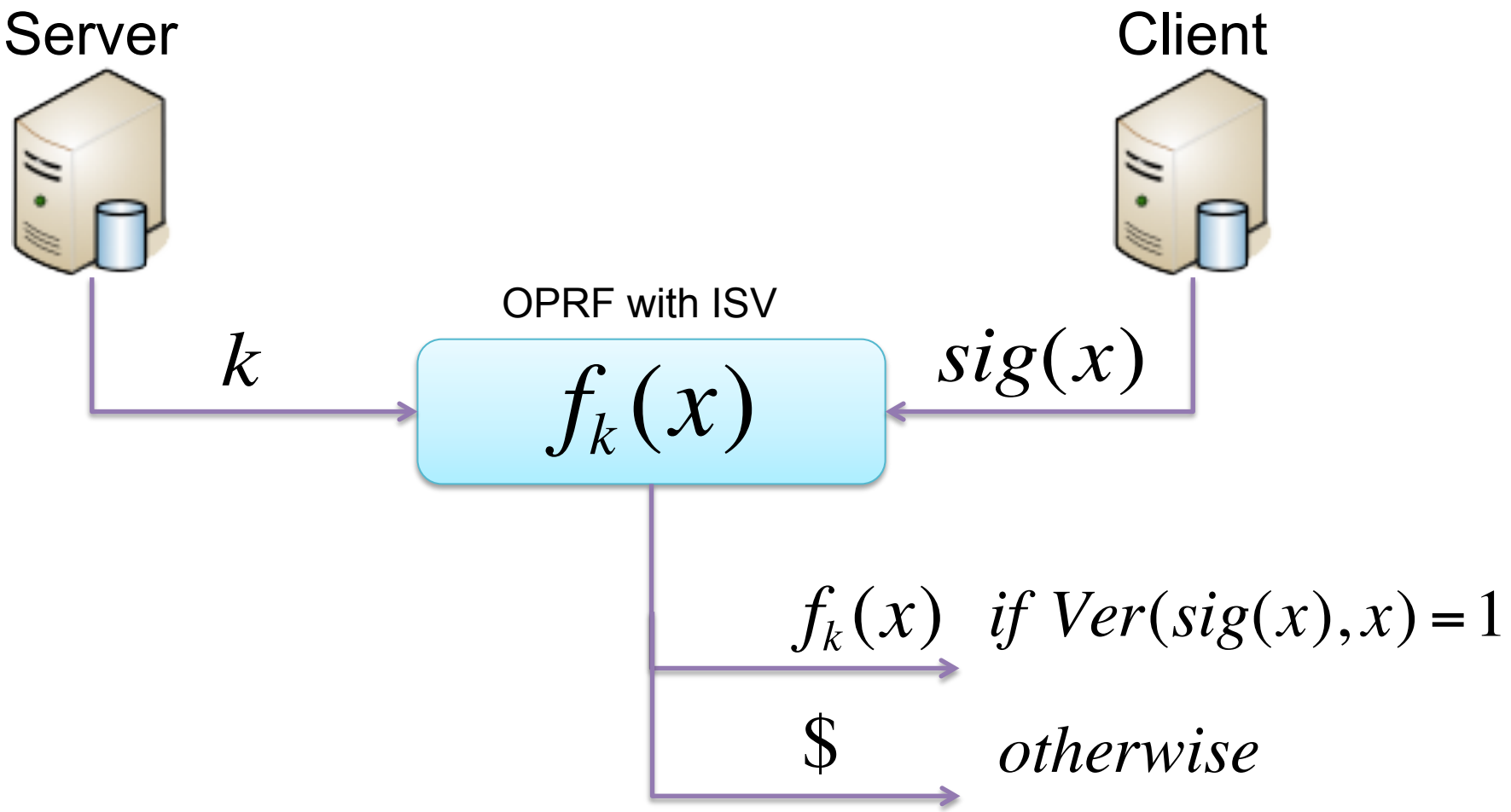
PRF: $\boxed{f_d(x) = H(sig_d(x))}$ (H one way function)

Server (d)		Client (x)
	$a = H(x) \cdot r^e$ ←	$r \in Z_N$
	$b = a^d$ →	$sig_d(x) = b / r$
	$(= H(x)^d r^{ed})$	$f_d(x) = H(sig_d(x))$

Authorized Private Set Intersection (APSI)



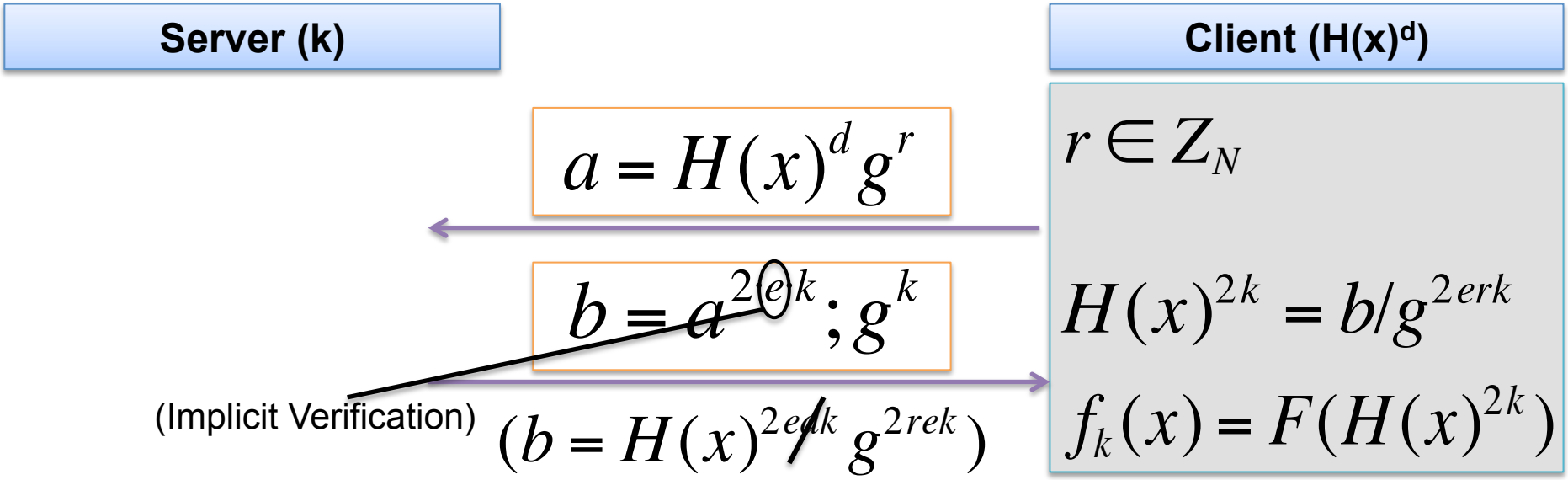
OPRF w/ Implicit Signature Verification



A simple OPRF-like with ISV

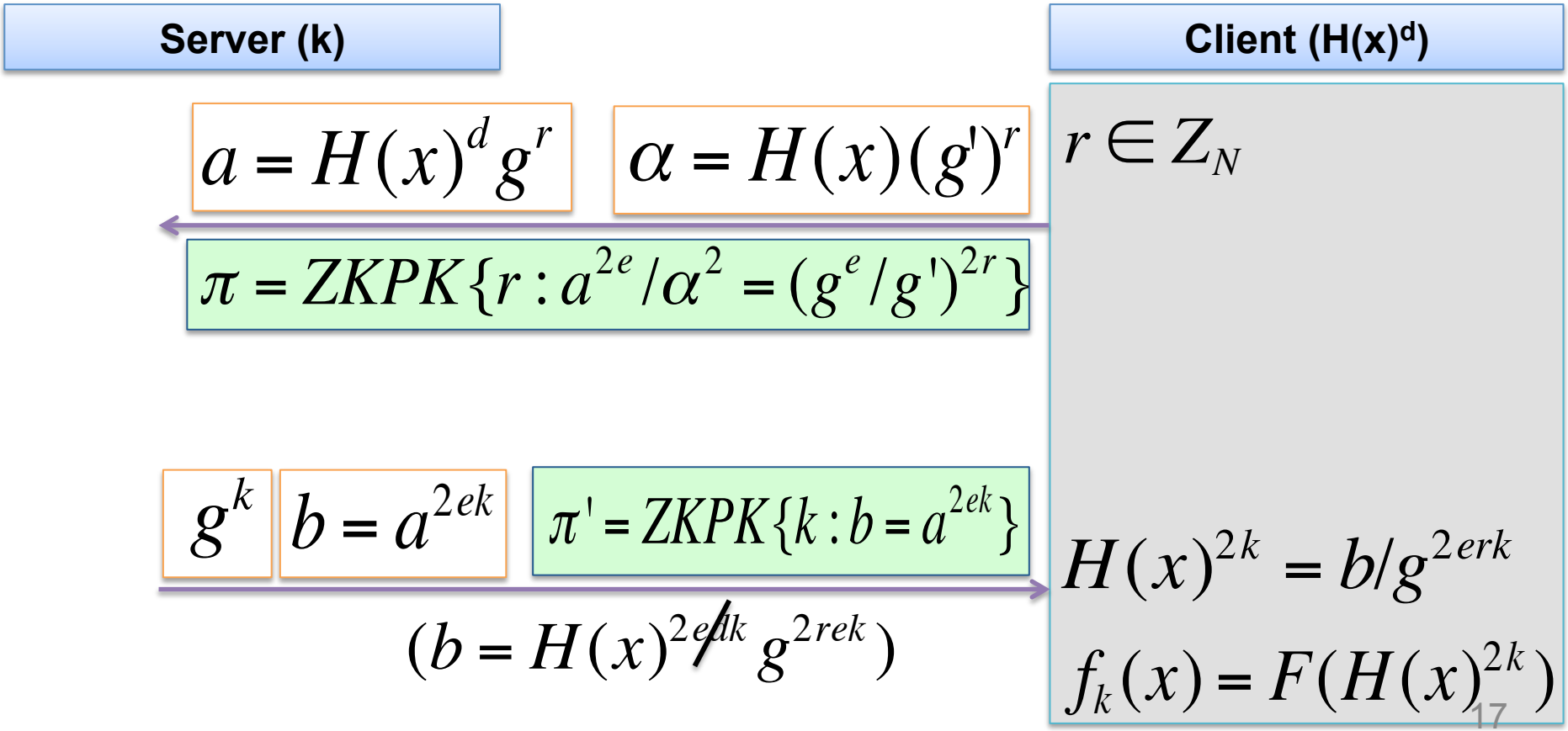
Court issues authorizations: $Sig(x) = H(x)^d \bmod N$

OPRF: $f_k(x) = F(H(x)^{2k} \bmod N)$



OPRF with ISV – Malicious Security

OPRF: $f_k(x) = F(H(x)^{2k})$



Other Building Blocks

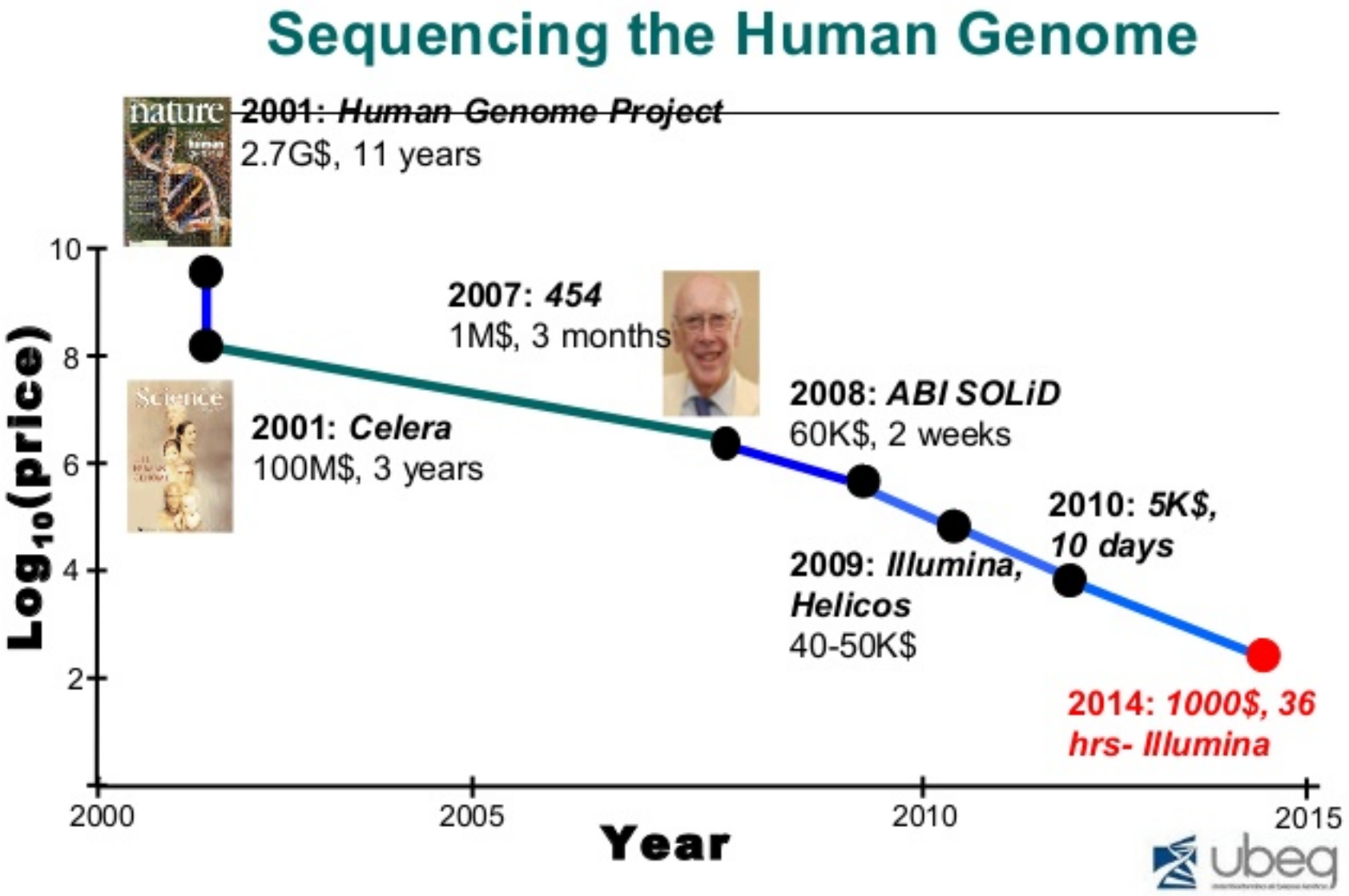
[DGT12]: Private Set Intersection Cardinality-only

[BDG12]: Private Sample Set Similarity

[DFT13]: Private Substring/Pattern Matching

Cool! So what? ☺

Genomics…


Sequencing the Human Genome
nature
2001: Human Genome Project
2.7G$, 11 years
2007: 454
1M$, 3 months
Science
2001: Celera
100M$, 3 years
2008: ABI SOLiD
60K$, 2 weeks
2010: 5K$, 10 days
2009: Illumina, Helicos
40-50K$
2014: 1000$, 36 hrs- Illumina
Log10(price)
10
8
6
4
2
2000
2005
2010
2015
Year
ubeq

1/05/2011 @ 4:57PM | 30,076 views

The First Child Saved By DNA Sequencing

+ Comment Now + Follow Comments

Genetic Gamble

New Approaches to Fighting Cancer

PART ONE
A Race to Leukemia's Source

PART TWO
Promise and Heartbreak

In Treatment for Leukemia, Glimpses of the Future

LETTER

doi:10.1038/nature13394

Genome sequencing identifies major causes of severe intellectual disability

Christian Gilissen[1]*, Jayne Y. Hehir-Kwa[1]*, Djie Tjwan Thung[1], Maartje van de Vorst[1], Bregje W. M. van Bon[1], Marjolein H. Willemsen[1], Michael Kwint[1], Irene M. Janssen[1], Alexander Hoischen[1], Annette Schenck[1], Richard Leach[2], Robert Klein[2], Rick Tearle[2], Tan Bo[1,3], Rolph Pfundt[1], Helger G. Yntema[1], Bert B. A. de Vries[1], Tjitske Kleefstra[1], Han G. Brunner[1,4]*, Lisenka E. L. M. Vissers[1]* & Joris A. Veltman[1,4]*


MAY 27, 2013
THE
ANGELINA
EFFECT
Angelina Jolie's double mastectomy puts genetic testing in the spotlight. What her choice reveals about calculating risk, cost and peace of mind
BY JEFFREY KLUGER & ALICE PARK
time.com
Time

health overview

Print my health overview | Share my health results

Show results for [redacted]

See new and recently updated reports »

23andWe Discoveries were made possible by 23andMe members who took surveys.

Disease Risks (114, 2 locked reports)

Elevated Risks	Your Risk	Average Risk
Psoriasis	22.4%	11.4%
Celiac Disease	0.5%	0.1%
Bipolar Disorder	0.2%	0.1%
Primary Biliary Cirrhosis	0.10%	0.08%
Scleroderma (Limited Cutaneous Type)	0.08%	0.07%

See all 114 risk reports...

Carrier Status (27, 1 locked report)

Hemochromatosis	Variant Present
Alpha-1 Antitrypsin Deficiency	Variant Absent
Bloom's Syndrome	Variant Absent
Canavan Disease	Variant Absent
Congenital Disorder of Glycosylation Type 1a (PMM2-CDG) new	Variant Absent
Cystic Fibrosis	Variant Absent
Familial Dysautonomia	Variant Absent
Factor XI Deficiency	Variant Absent

See all 27 carrier status...

Traits (52)

Alcohol Flush Reaction	Does Not Flush
Bitter Taste Perception	Can Taste
Earwax Type	Wet
Eye Color	Likely Blue
Hair Curl	Slightly Curlier Hair on Average

See all 52 traits...

Drug Response (20)

Warfarin (Coumadin®) Sensitivity	Increased
Abacavir Hypersensitivity	Typical
Alcohol Consumption, Smoking and Risk of Esophageal Cancer	Typical
Clopidogrel (Plavix®) Efficacy	Typical
Fluorouracil Toxicity	Typical

See all 20 drug response...

The genotyping services of 23andMe are performed in LabCorp's CLIA-certified laboratory. The tests have not been cleared or approved by the FDA but have been analytically validated according to CLIA standards. The information on this page is intended for research and educational purposes only, and is not for diagnostic use.

Genetic Ethnicity

Southern European	37%
West African	20%
British Isles	13%
Native South American	9%
Finnish/Volga-Ural	9%
Eastern European	6%
Uncertain	6%

The Bad News

Sensitivity of human genome:

- **Uniquely** identifies an individual
- Discloses ethnicity, disease predispositions (including mental)
- Progress aggravates fears of discrimination
- **Once leaked, it cannot be "revoked"**

De-identification and obfuscation are not effective

More info:

[ADHT13] Chills and Thrills of Whole Genome Sequencing. IEEE Computer Magazine.

Secure Genomics?

Privacy:

Individuals remain in control of their genome

Allow doctors/clinicians/labs to run genomic tests, while disclosing the required minimum amount of information, i.e.:

(1) Individuals don't disclose their entire genome

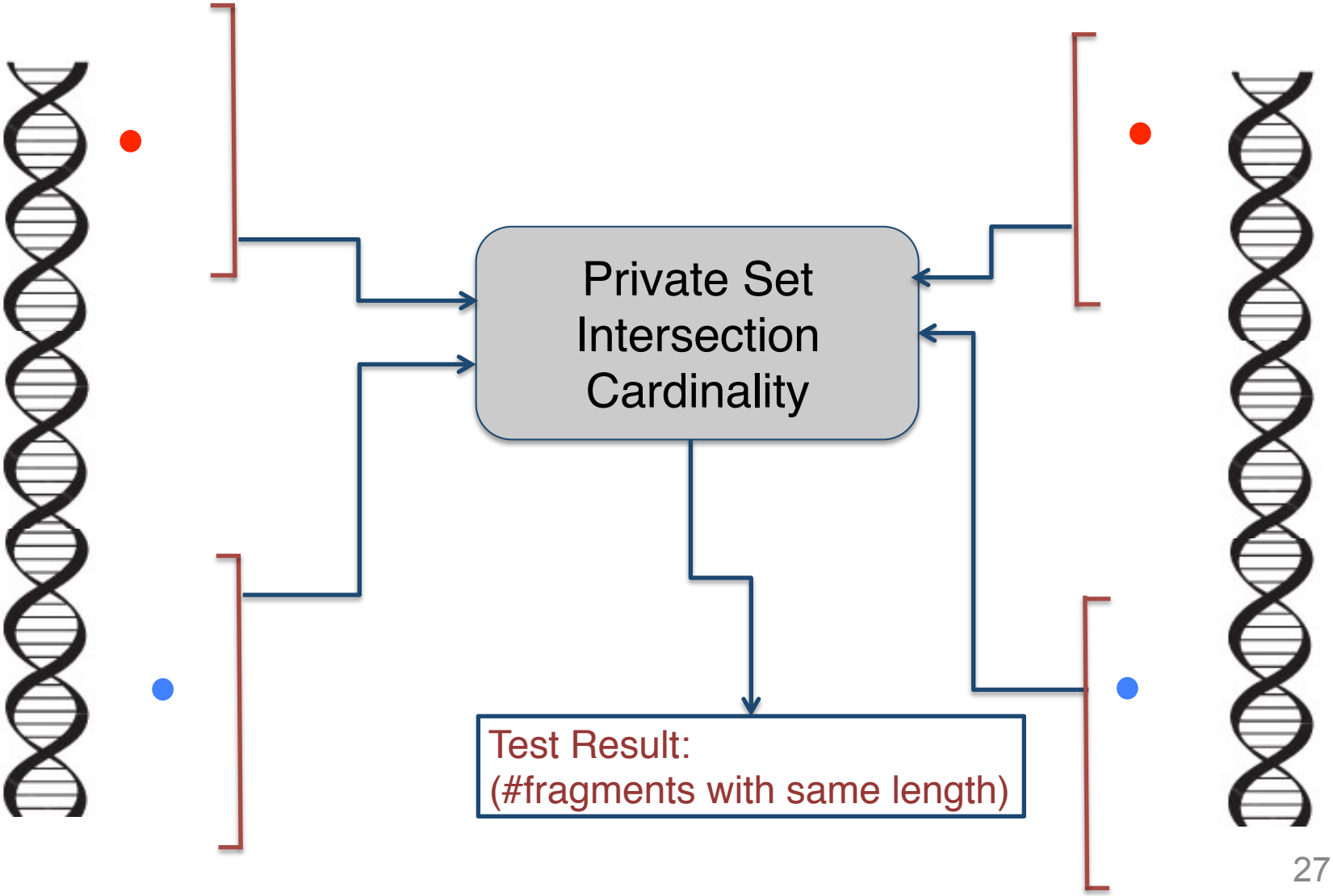
(2) Testing facilities keep test specifics ("secret sauce") confidential

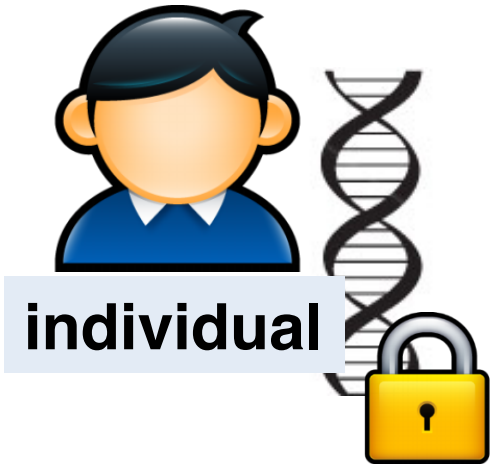
[BBDGT11]: Secure genomics via *-PSI

Most personalized medicine tests in < 1 second

Works on Android too

Private RFLP-based Paternity Test





individual

genome

- Private Set Intersection (PSI)
- Authorized PSI
- Private Pattern Matching
- Homomorphic Encryption
- Garbled Circtuis
- […]

doctor or lab

test specifics

Secure Function Evaluation

- Paternity/Ancestry Testing
- Testing of SNPs/Markers
- Compatibility Testing
- Disease Predisposition […]

test result

test result

Output reveals nothing beyond test result

Open Problems

Where do we store genomes?

- Encryption can't guarantee security past 30-50 yrs
- Reliability and availability issues?

Cryptography

- Efficiency overhead
- Data representation assumptions
- How much understanding required from users?

Collaborative Anomaly Detection

Anomaly detection is hard

- Suspicious activities deliberately mimic normal behavior
- But, malevolent actors often use same resources

Wouldn't it be better if organizations collabo

- It's a w

> "It is the policy of the United States Government to increase the volume, timelines, and quality of cyber threat information shared with U.S. private sector entities so that these entities may better protect and defend themselves against cyber attacks."
>
> Barack Obama
> 2013 State of the Union Address

Problems with Collaborations

Trust

Will others leak my data?

Legal Liability

Will I be sued for sharing customer data?

Will others find me negligible?

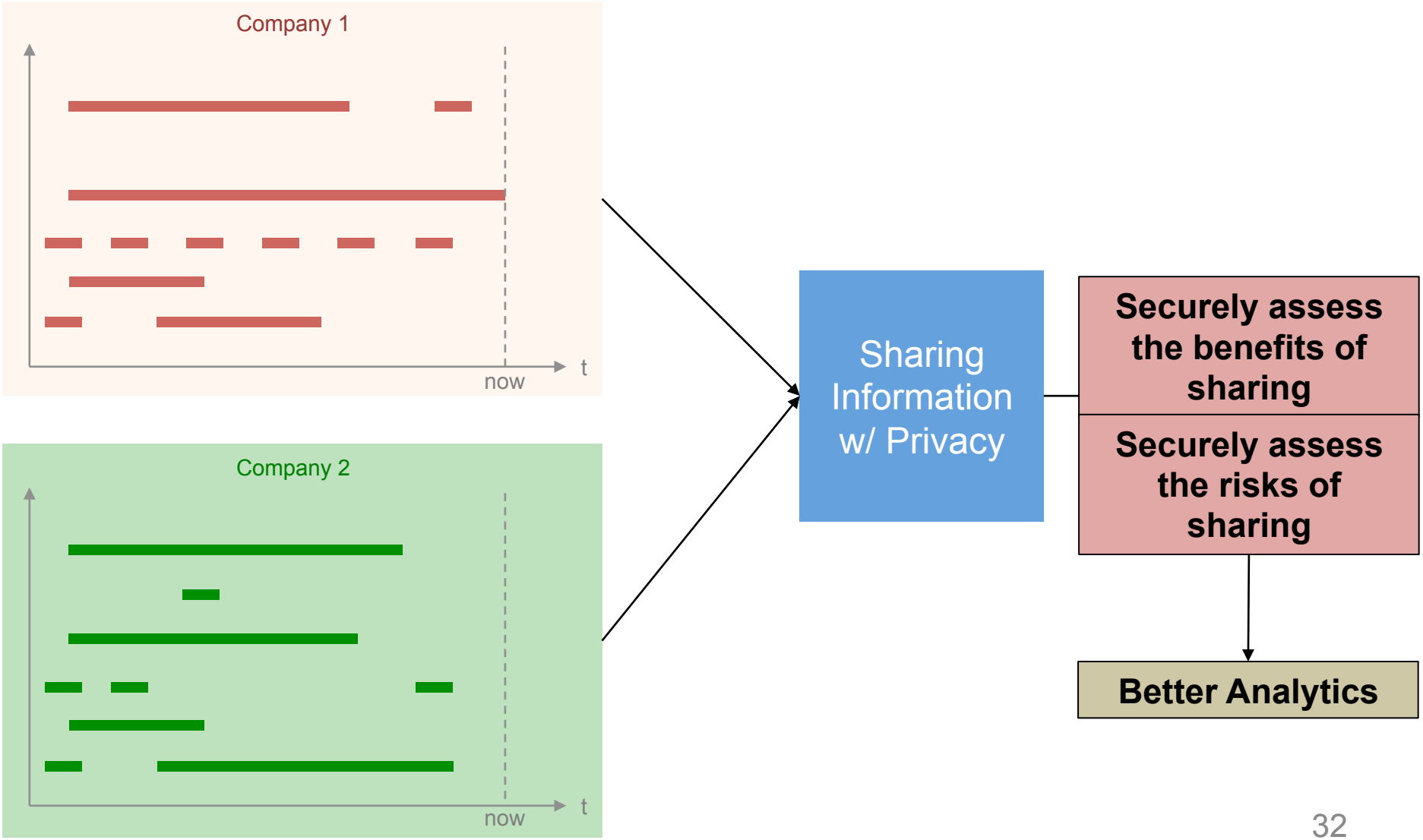
Competitive concerns

Will my competitors outperform me?

Shared data quality

Will data be reliable?

Solution Intuition [FDB15]



Training Machine Learning Models

The Big Data "Hype"

Large-scale collection of contextual information often essential to gather statistics, train machine learning models, and extract knowledge from data

Doing so privately…

Efficient Private Statistics [MDD16]

Real-world problems:

1. Recommender systems for online streaming services
2. Statistics about mass transport movements
3. Traffic statistics for the Tor Network

Available tools for computing private statistics are impractical for large streams collection

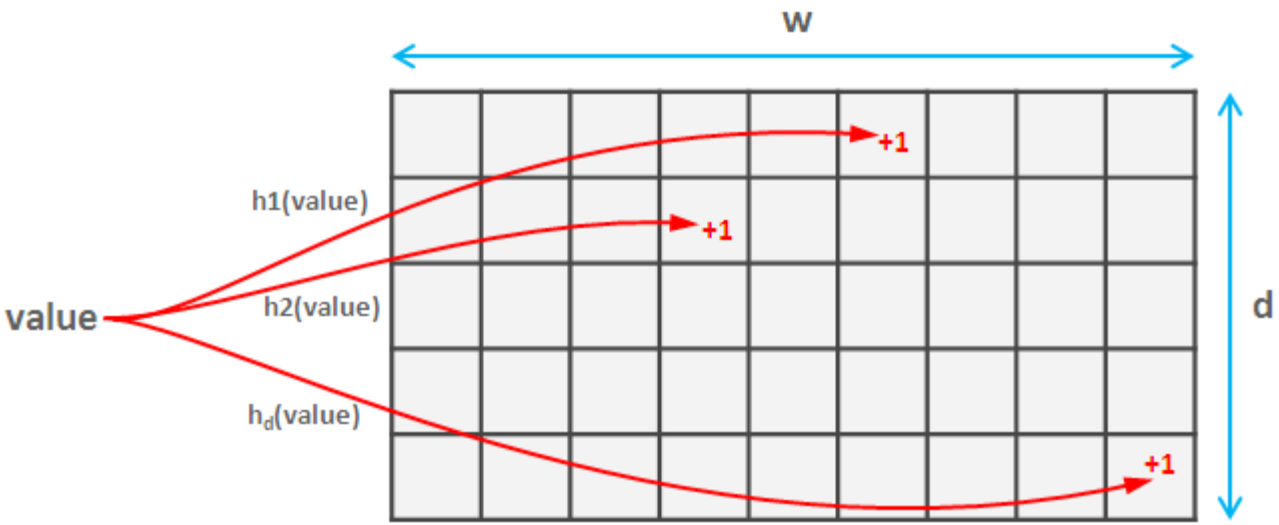
Intuition: Approximate statistics are acceptable in some cases?

Preliminaries: Count-Min Sketch

An estimate of an item's frequency in a stream

Mapping a stream of values (of length T) into a matrix of size O(logT)

The sum of two sketches results in the sketch of the union of the two data streams



ItemKNN Recommender Systems

Predict favorite TV programs based on their own ratings and those of “similar” users

Consider N users, M programs and binary ratings

Build a co-views matrix C, where C_{ab} is the number of views for the pair of programs (a,b)

Compute the Similarity Matrix $\{Sim\}_{ab} = \frac{C_{ab}}{\sqrt{C_a \cdot C_b}}$

Identify K-Neighbors based on the Similarity Matrix

Private Recommender System

We build a global matrix of co-views for training ItemKNN in a privacy-friendly way by relying on:

- Private data aggregation based on [Kursawe et al. 2011]
- Count-Min Sketch to reduce overhead

System Model

- Users (in groups)
- Tally Server (e.g, the BBC)

Security & Implementation

Security

In the honest-but-curious model under the CDH assumption

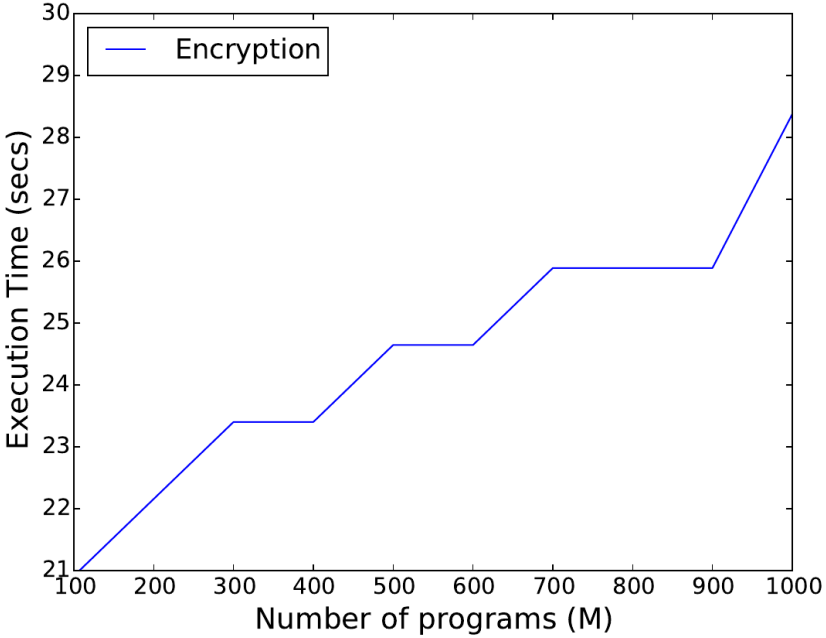
Prototype implementation:

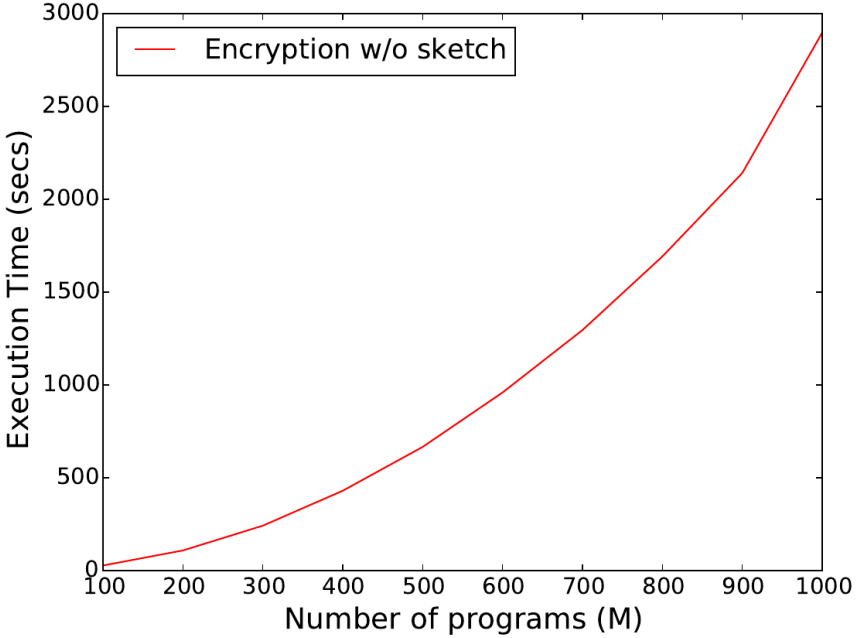
Tally as a Node.js web server

Users run in the browser or as a mobile cross-platform application (Apache Cordova)

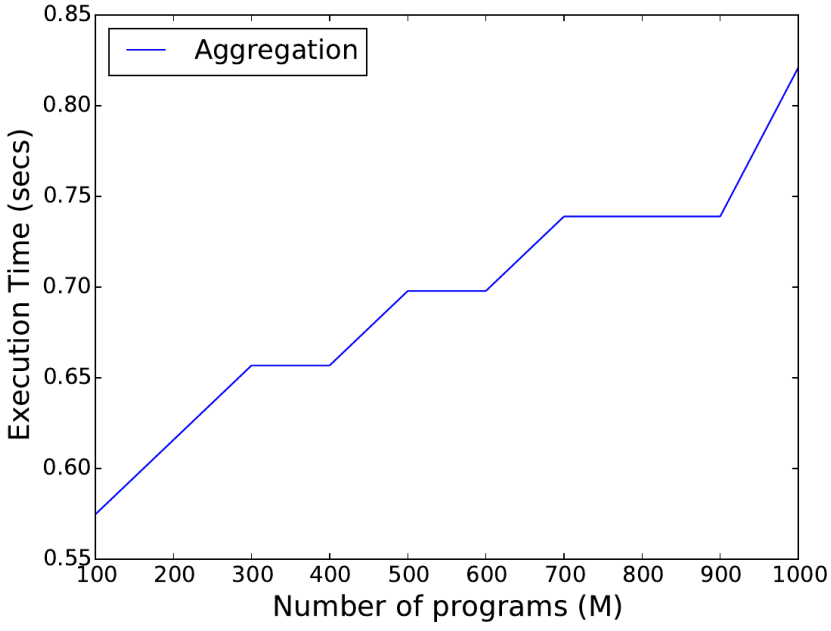
Transparency, ease of use, ease of deployment

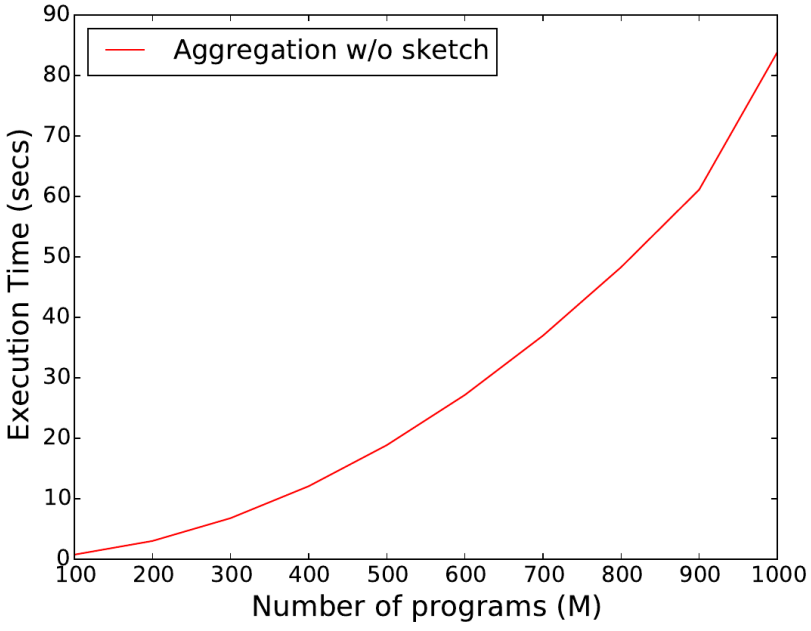
User side



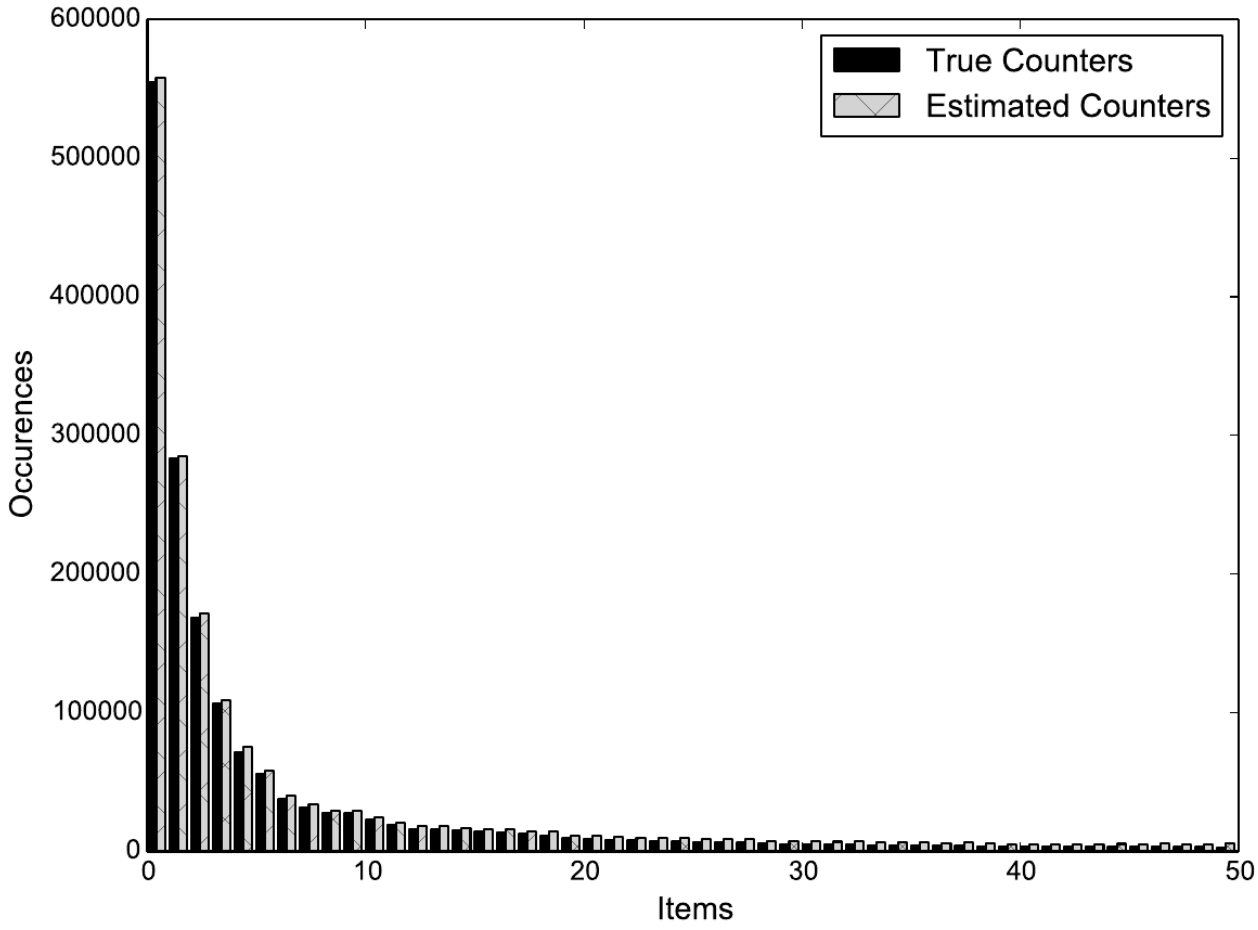


Server side





Accuracy


600000
500000
400000
300000
200000
100000
0
Occurences
0
10
20
30
40
50
Items
True Counters
Estimated Counters

The Road Ahead…

This slide is intentionally left blank

Shameless Advertising

UCL MSc in Information Security

http://www.cs.ucl.ac.uk/admissions/msc_isec/

Several PhD positions in security/privacy

http://sec.cs.ucl.ac.uk

https://privacyus.cs.ucl.ac.uk

Several post-doc positions in security/privacy

Talk to me – https://emilianodc.com