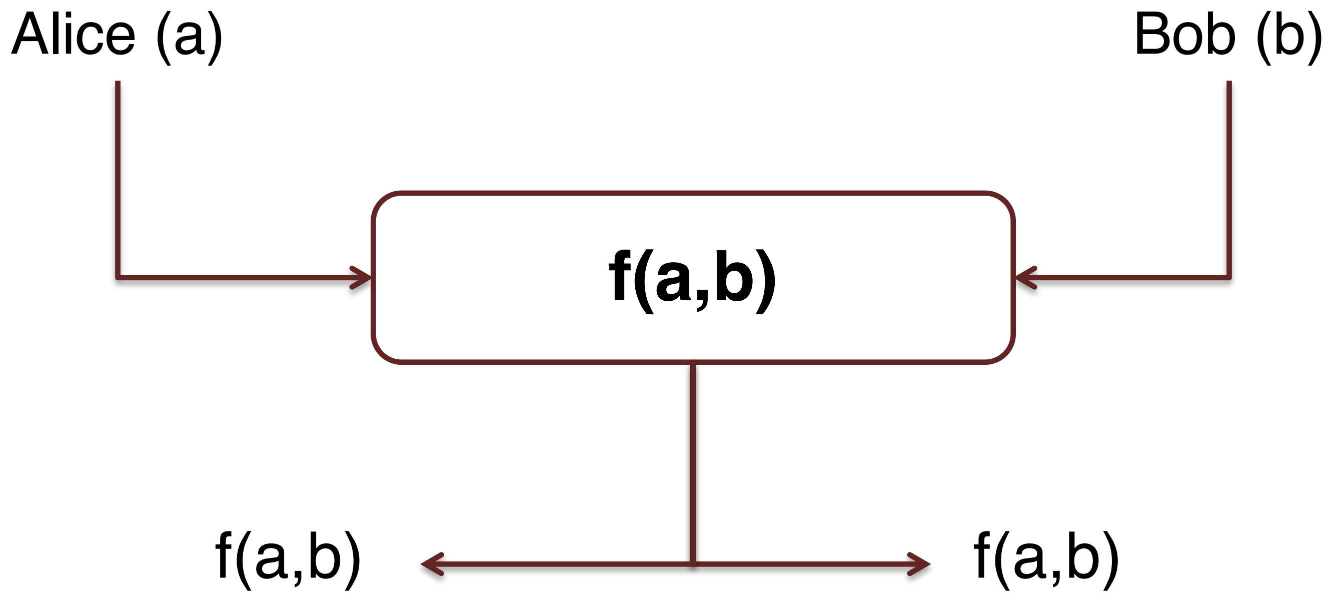
Cryptographic Protocols for Privacy-Preserving Genomic Testing: Tools and Applications

Emiliano De Cristofaro

University College London (UCL)
https://emilianodc.com

Secure Multiparty Computation (SMC)



How to Implement SMC?

1. Garbled Circuits

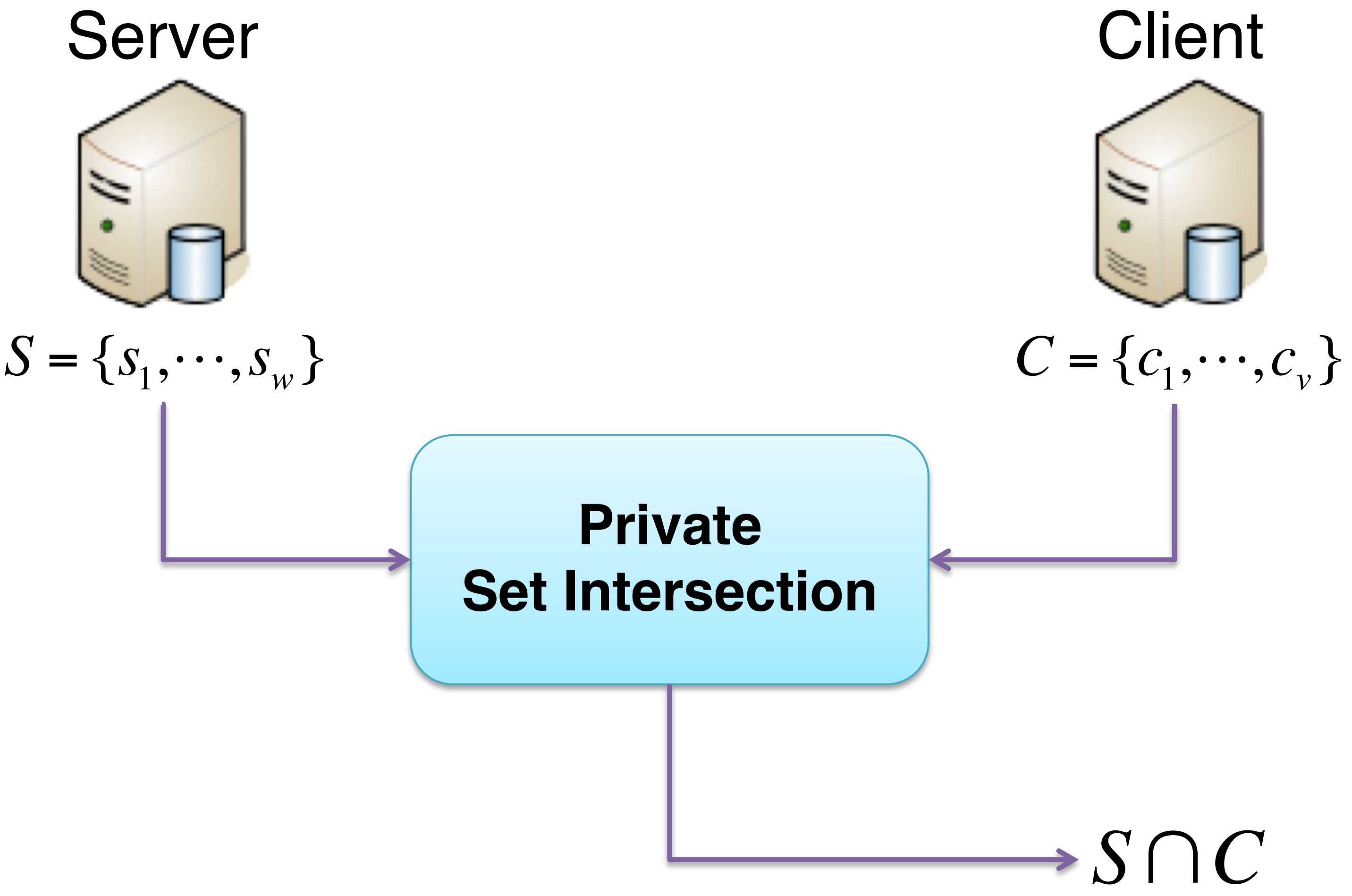
Sender prepares a “garbled” circuit and sends it to the receiver, who obliviously evaluates the circuit, learning the encodings corresponding to both his and the senders output

2. Special-Purpose Protocols

Implement one specific function (and only that)

Usually based on public-key crypto properties
[Have you ever heard of homomorphic encryption?]

Private Set Intersection (PSI)



Private Set Intersection?

FBI (Domestic suspect terrorists) and **CIA** (Foreign suspect terrorists)

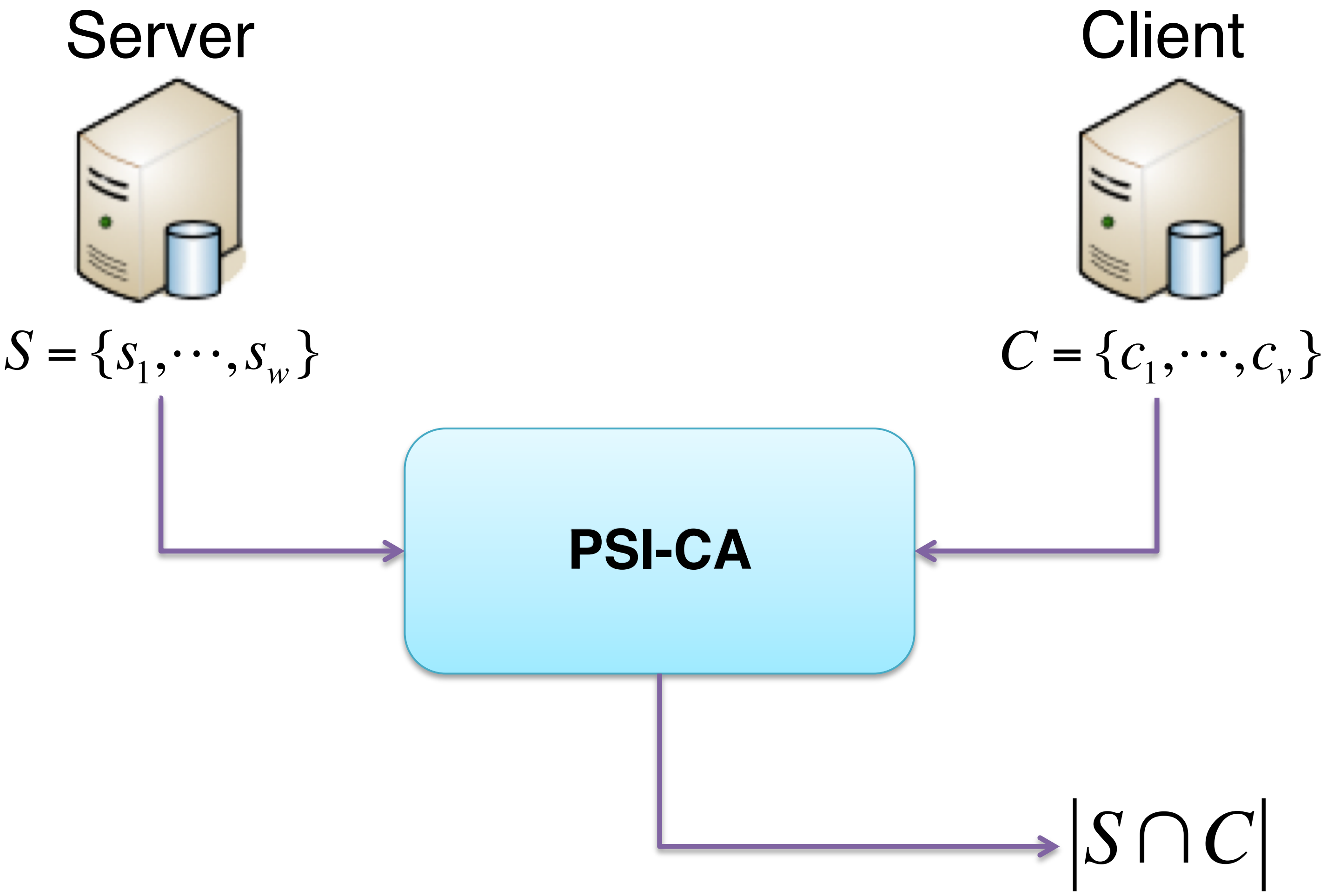
Find out whether any suspect is in common

IRS (Tax Evaders) and **Swiss Bank** (Customers)

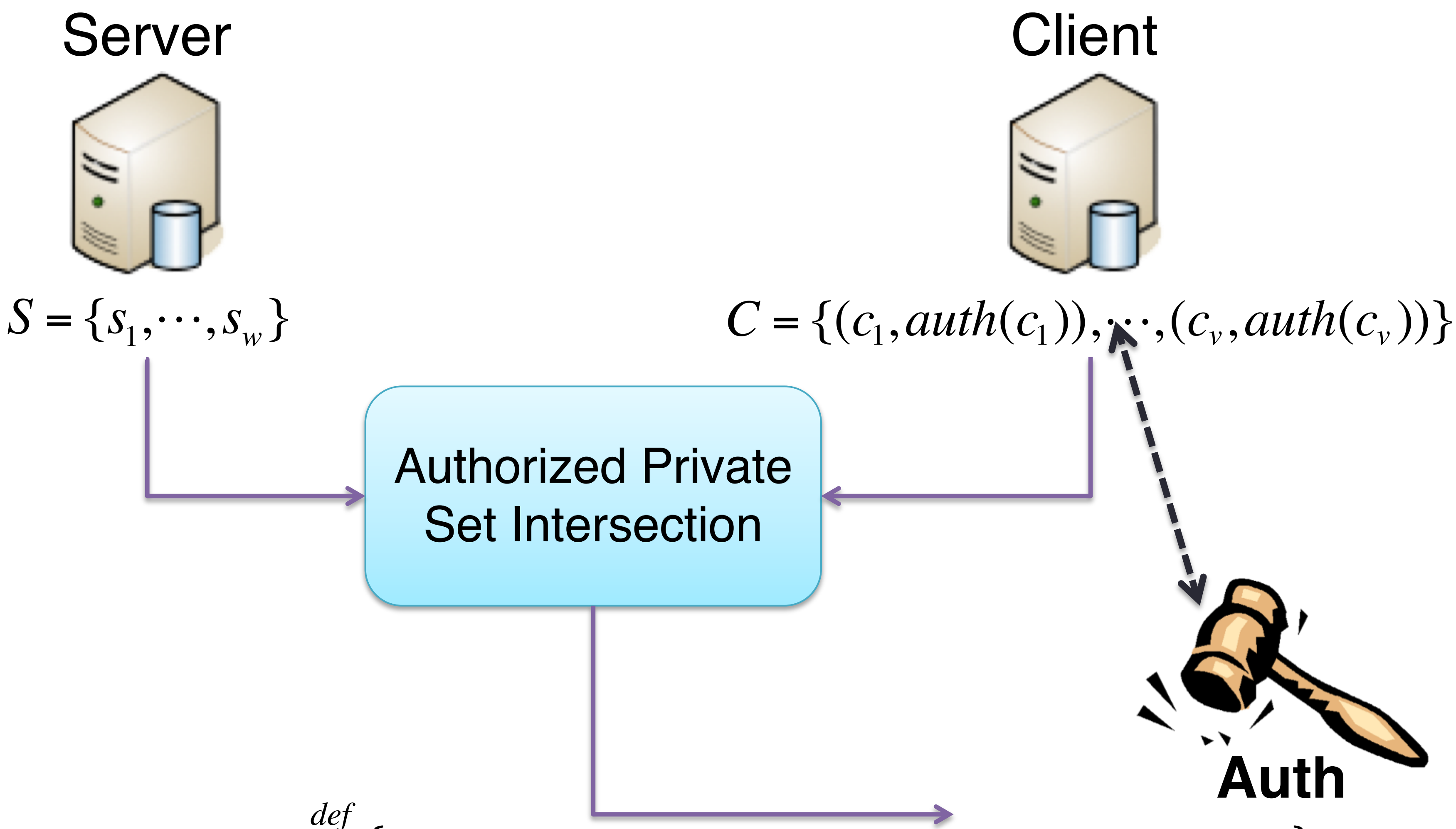
Discover if tax evaders have accounts at foreign banks

And more!

Private Set Intersection Cardinality (PSI-CA)



Authorized Private Set Intersection (APSI)



$$S \cap C \stackrel{def}{=} \left\{ s_j \in S \,\middle|\, \exists c_i \in C : c_i = s_j \wedge auth(c_i) \text{ is valid} \right\}$$

Private Personal Genomic Tests

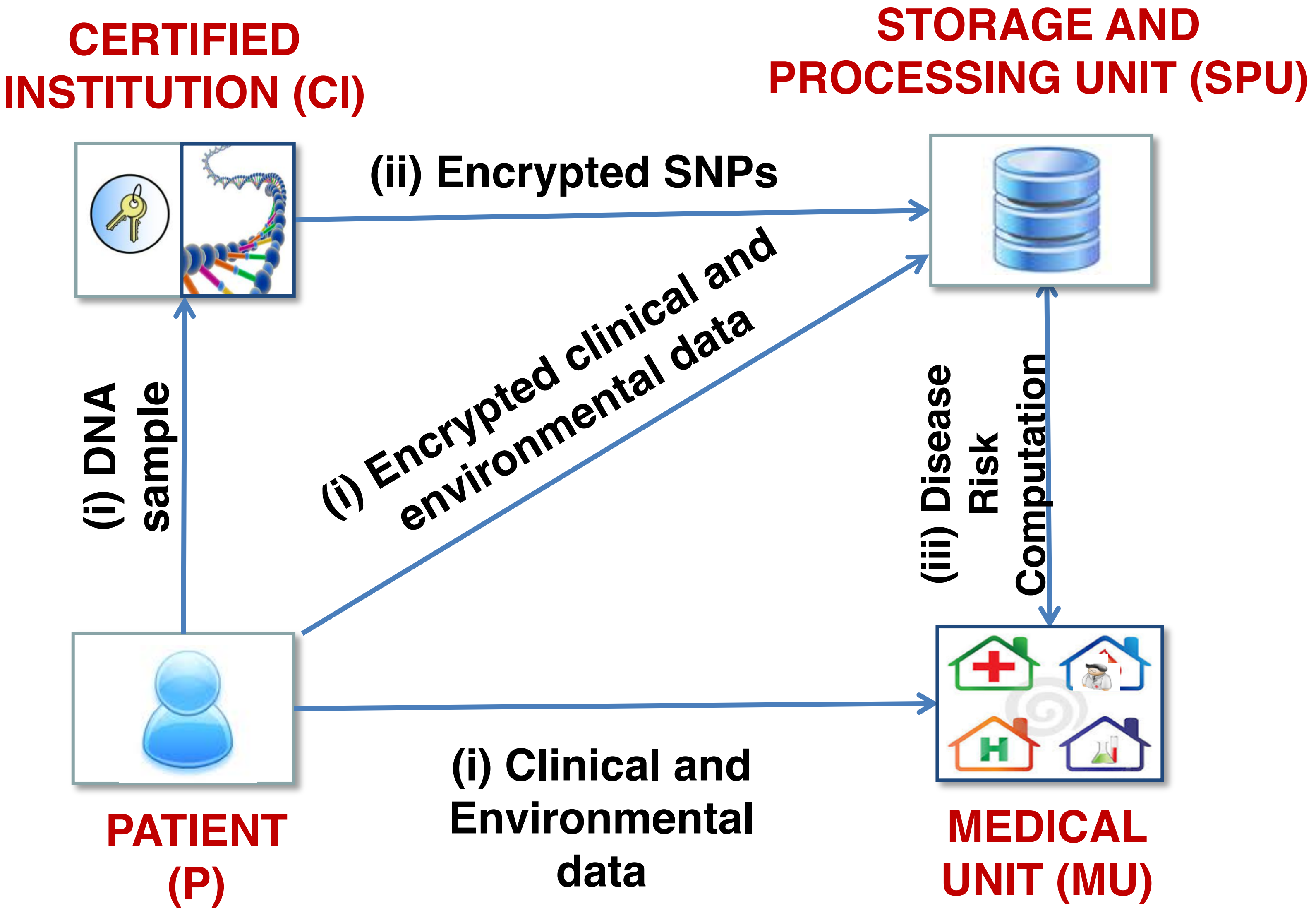
Individuals retain control of their sequenced genome

Allow doctors/labs to run genetics tests, but:

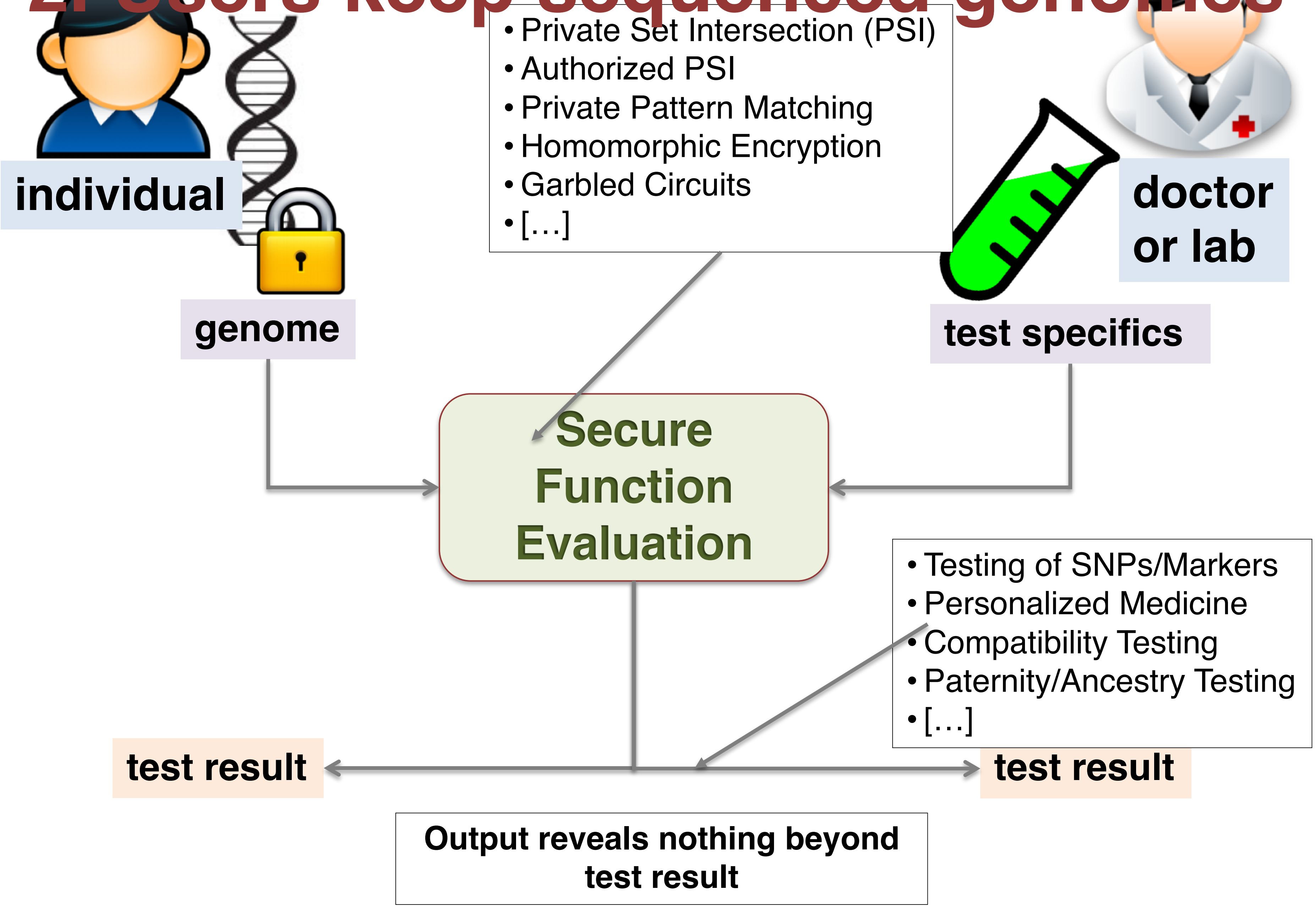
1. Genome never disclosed, only test output is
2. Pharmas can keep test specifics confidential

… two main approaches …

1. Using Semi-Trusted Parties



2. Users keep sequenced genomes



2. Users keep sequenced genomes

Baldi et al. (CCS'11)

Privacy-preserving version of a few genetic tests, based on private set operations

Paternity test, Personalized Medicine, Compatibility Tests

(First work to consider fully sequenced genomes)

De Cristofaro et al. (WPES'12), extends the above

Framework and prototype deployment on **Android**

Adds Ancestry/Genealogy Testing

Genetic Paternity Test

A Strawman Approach for Paternity Test:

- On average, ~99.5% of any two human genomes are identical
- Parents and children have even more similar genomes
- Compare candidate's genome with that of the alleged child:
  - **Test positive if percentage of matching nucleotides is > 99.5 + τ**

First-Attempt Privacy-Preserving Protocol:

- Use an appropriate secure two-party protocol for the comparison
- PROs: High-accuracy and error resilience
- CONs: Performance not promising (3 billion symbols in input)
  - In our experiments, computation takes a few days

Genetic Paternity Test

Wait a minute!

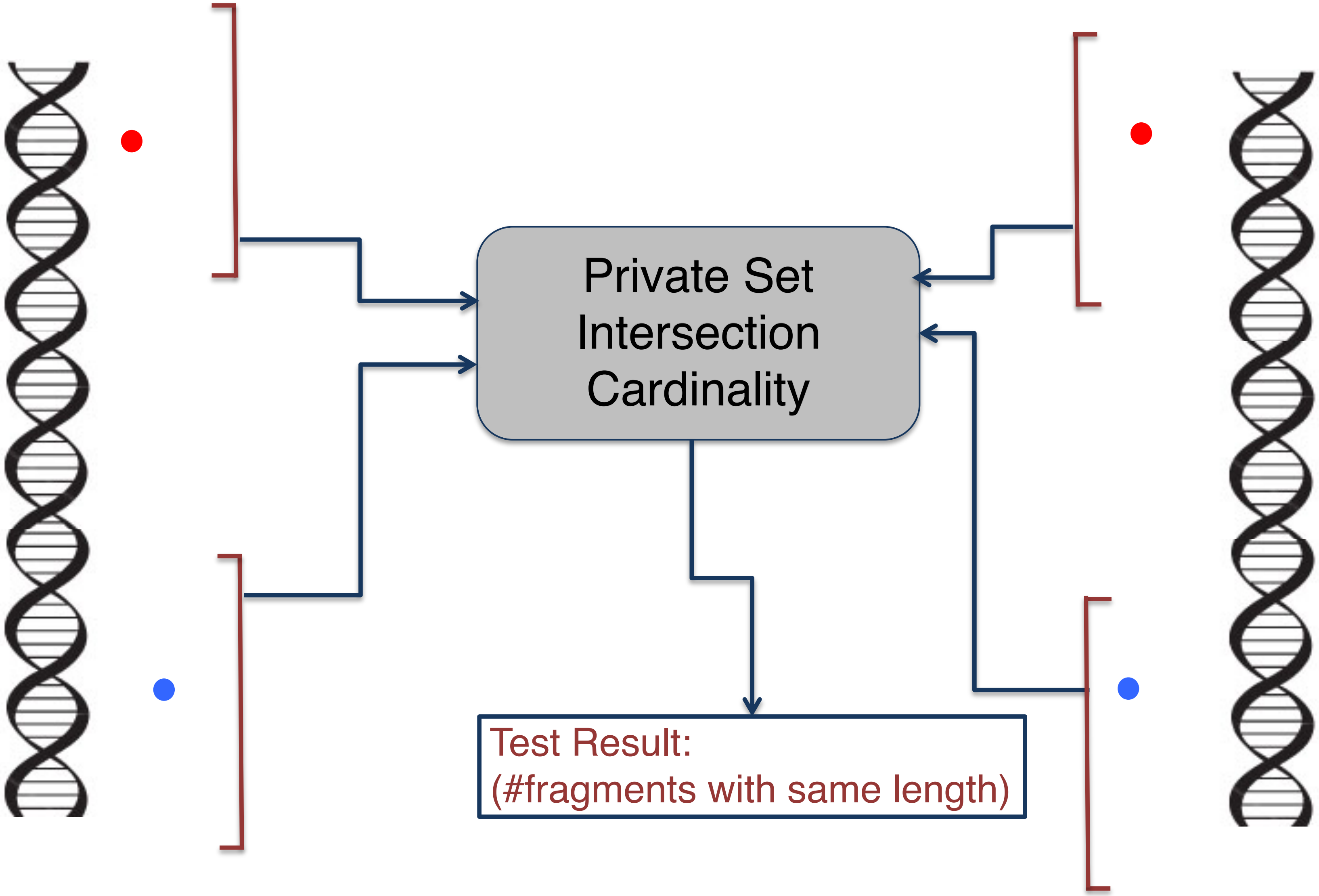
~99.5% of any two human genomes are identical

Why don't we compare *only* the remaining 0.5%?

We can compare by counting how many

But… We don't know (yet) where *exactly* this 0.5% occur!

Private RFLP-based Paternity Test



Personalized Medicine (PM)

Drugs designed for patients' genetic features

Associating drugs with a unique genetic fingerprint

Max effectiveness for patients with matching genome

Test drug's "genetic fingerprint" against patient's genome

Examples:

tmpt gene – relevant to leukemia

(1) G->C mutation in pos. 238 of gene's c-DNA, or (2) G->A mutation in pos. 460 and one A->G is pos. 419 cause the *tpmt* disorder (relevant for leukemia patients)

hla-B gene – relevant to HIV treatment

One G->T mutation (known as *hla-B*5701* allelic variant) is associated with extreme sensitivity to abacavir (HIV drug)

Reducing P³MT to APSI

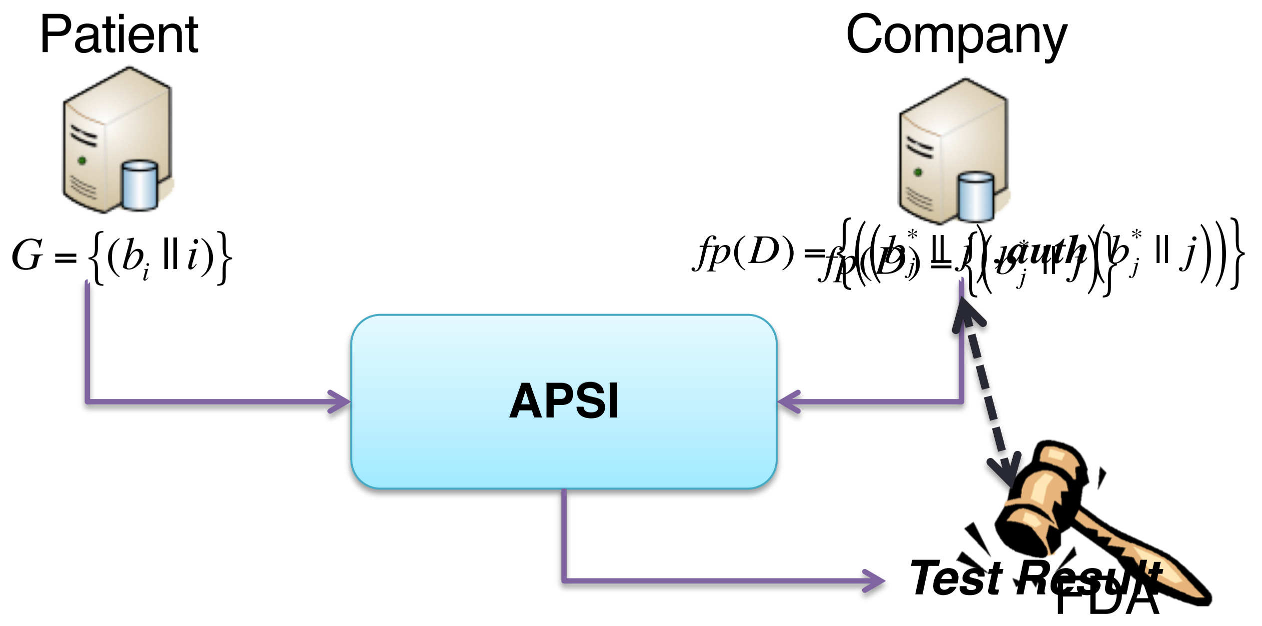
Intuition:

FDA = Court, Pharma = *Client*, Patient = *Server*

Patient's private input set: $G = \{(b_i \| i) \mid b_i \in \{A, C, G, T\}\}_{i=1}^{3 \cdot 10^9}$

Pharmaceutical company's input set: $fp(D) = \{(b_j^* \| j)\}$

- Each item in $fp(D)$ needs to be authorized by FDA



Other Areas 1/

Secure computation for data sharing

Homomorphic encryption for computation outsourcing

Honey encryption for long-term storage

Beyond Crypto

Differential privacy

Adding noise to a dataset with the goal of supporting statistical queries while preserving the privacy of the users whose information is contained in the dataset

Examples:

Computing number/location of SNPs associated to disease

Significance/correlation between a SNP and a disease

Open Problems

Where do we store genomes?

Encryption can’t guarantee security past 30-50 yrs

Reliability and availability issues?

Challenges with Crypto

Efficiency overhead

Dealing with sequencing errors

How much understanding required from users?

Thank you!

Special thanks to

E. Ayday, P. Baldi, R. Baronio, G. Danezis, S. Faber,
P. Gasti, J-P. Hubaux, A. Mittos, B. Malin, B. Oprisanu, G. Tsudik