ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Schools of {Computer and Communication Sciences} and {Life Sciences}

Erman Ayday, Jean Louis Raisaro, Paul J. McLaren, Jacques Fellay and Jean-Pierre Hubaux

Privacy-Enhancing Technologies for Disease Risk Tests Based on Genomic and Non-Genomic Data

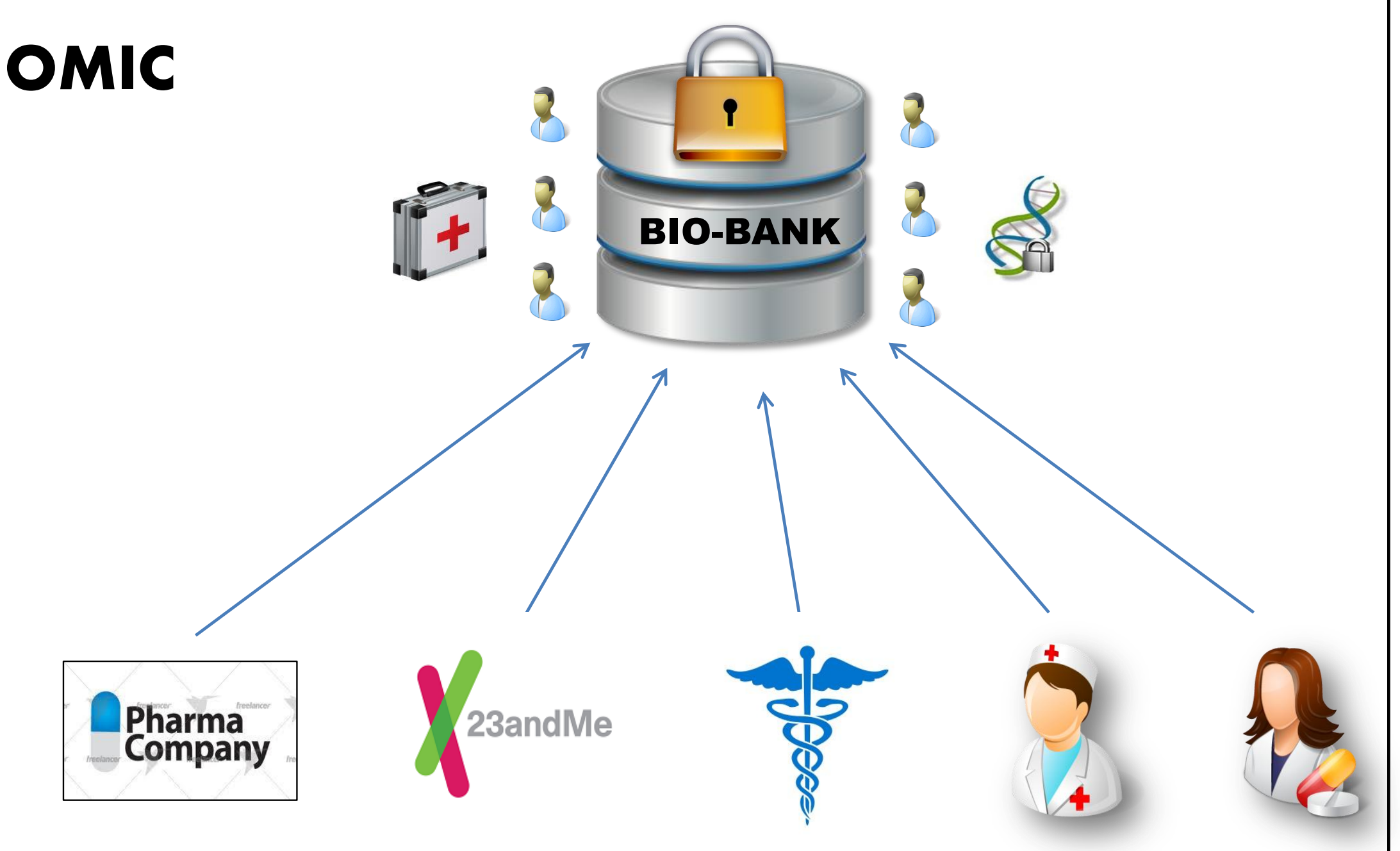
Sophia Genetics

1. Motivations

- **Genomic data provides opportunities for substantial improvements in diagnosis and preventive medicine.**
- **Individual's predisposition to disease depends on genomic variations.**
- **Non-genomic attributes of individuals also contribute significantly to their disease risks.**

PRIVACY THREATS DUE TO GENOMIC INFORMATION LEAKAGE:

- Revelation of predisposition to diseases, ethnicity, paternity, filiation, etc.
- Genetic discrimination.
- Denial of access to health insurance, mortgage, education, and employment.



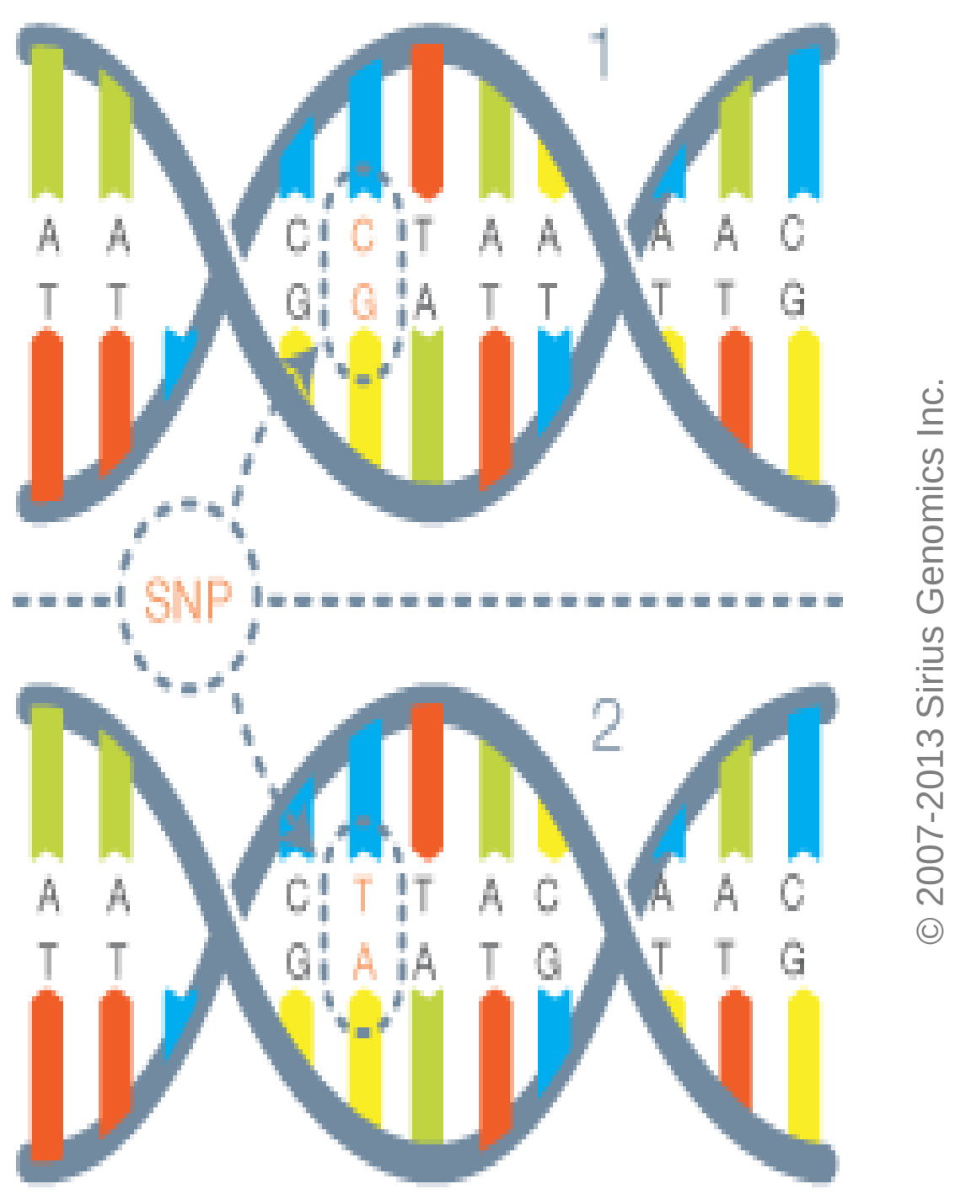
GOALS:

- Protect the privacy of patients' genomic data and non-genomic data on a centralized bio-bank.
- Allow different health stakeholders to access only to the medical data they need (or they are authorized for).
- Allow different health stakeholders to perform some computations on the encrypted data in a privacy-preserving fashion in a reasonable time.

2. Genomic Background

The human genome has approximately 3 billion letters.

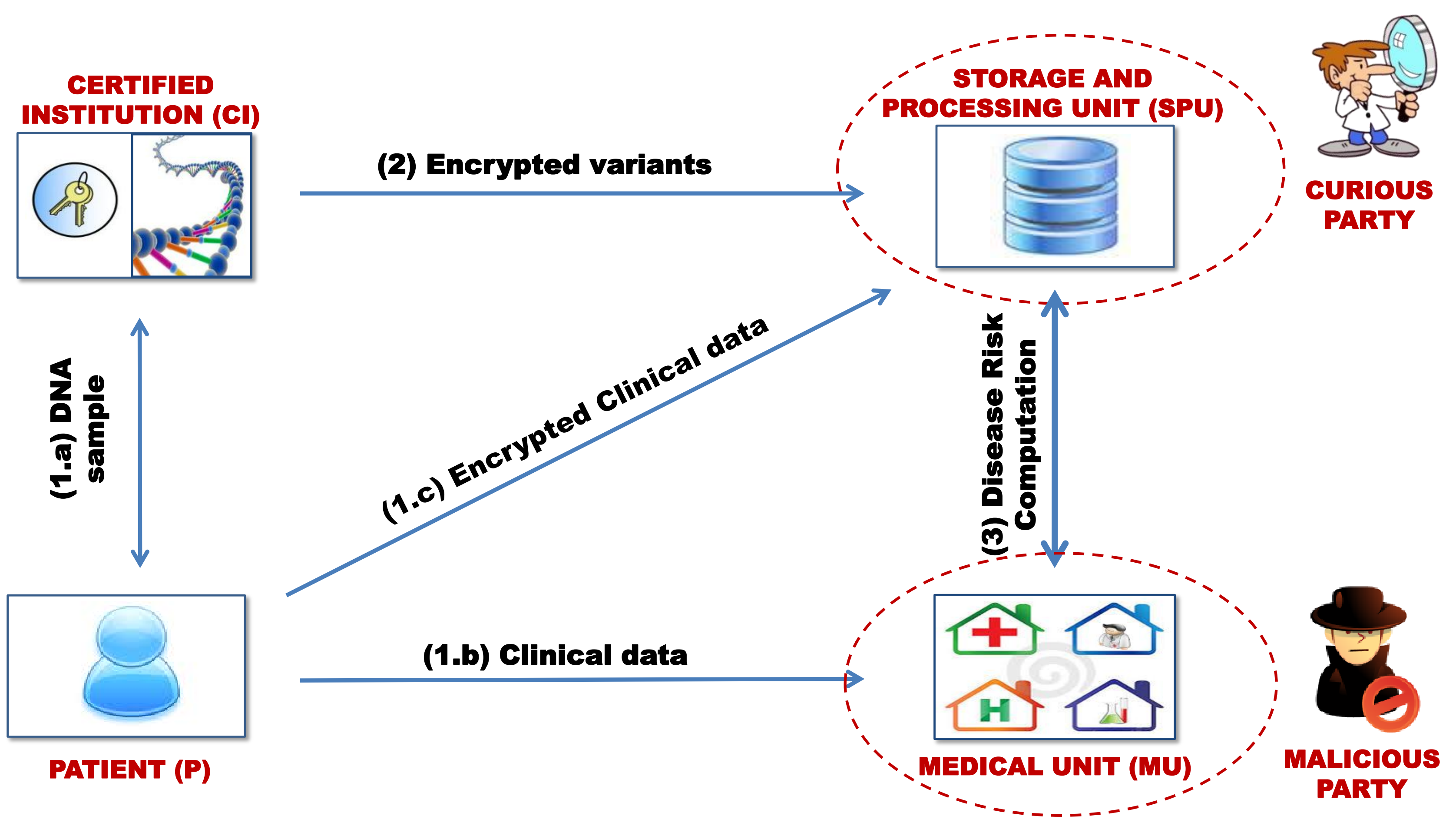
- ➢ Single Nucleotide Polymorphisms (SNPs): DNA variations, occurring when a single nucleotide differs between members of the same species.
- ➢ Potential nucleotides for a SNP position are called alleles.
- ➢ A disease risk test is done by analyzing particular SNPs along with other non-genomic risk factors.



© 2007-2013 Sirius Genomics Inc.

- ➢ Each SNP contributes to the disease risk in a different amount.
- ➢ 40 million approved SNPs in the human population.
- ➢ Each patient carries around 4 million SNPs out of 40 million – real SNPs of the patient.
- ➢ 75 real SNPs enable the attacker to identify a person.

3. Proposed Framework


CERTIFIED INSTITUTION (CI)
(2) Encrypted variants
STORAGE AND PROCESSING UNIT (SPU)
CURIOUS PARTY
(1.a) DNA sample
(1.c) Encrypted Clinical data
(3) Disease Risk Computation
PATIENT (P)
(1.b) Clinical data
MEDICAL UNIT (MU)
MALICIOUS PARTY

4. Disease risk test

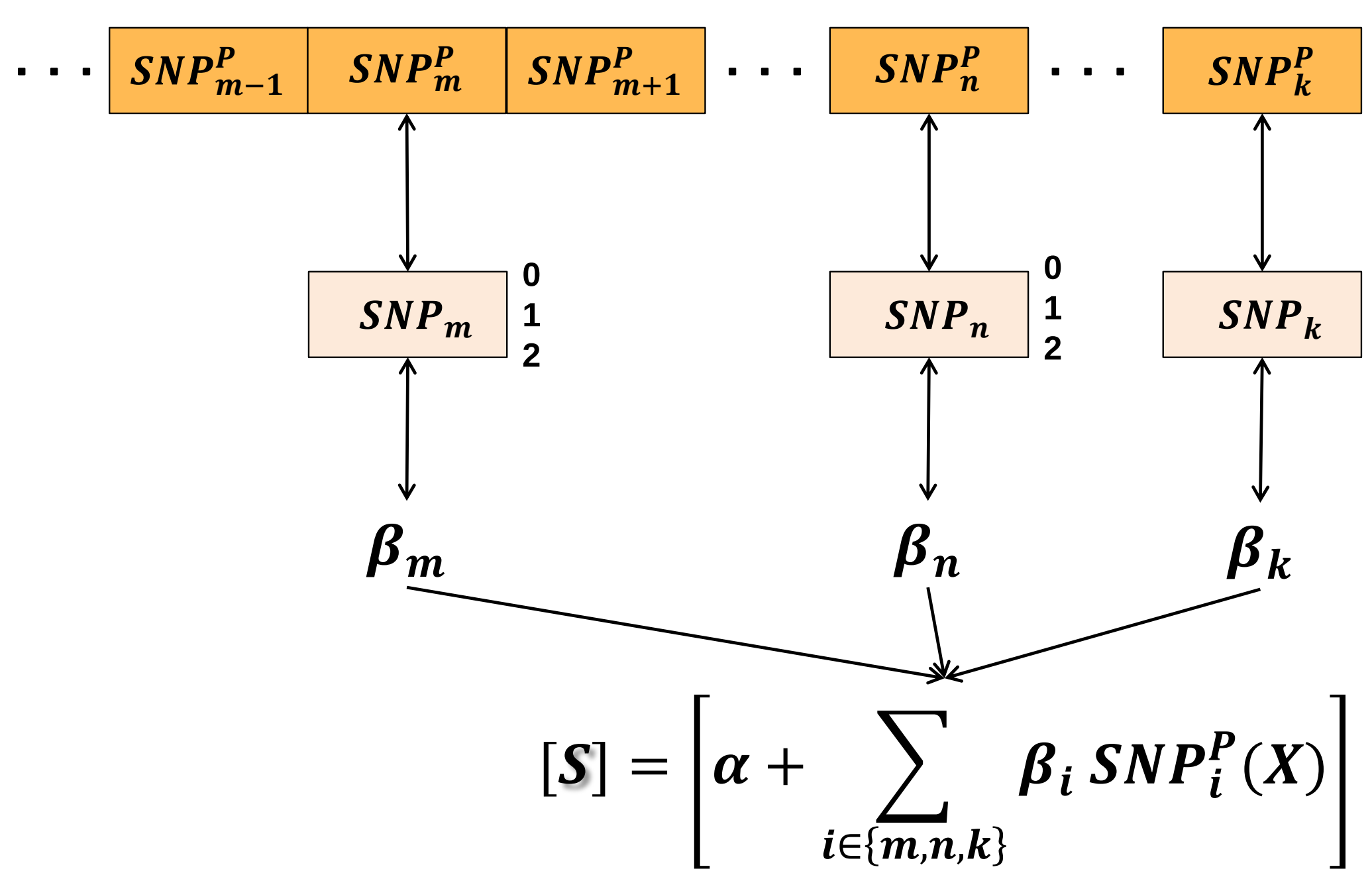
A. GENETIC RISK COMPUTATION THROUGH A PRIVATE LOGISTIC REGRESSION MODEL

Patient (P)'s SNPs:

Markers for disease X $SNP_i^P \in \{0, 1, 2\}$:

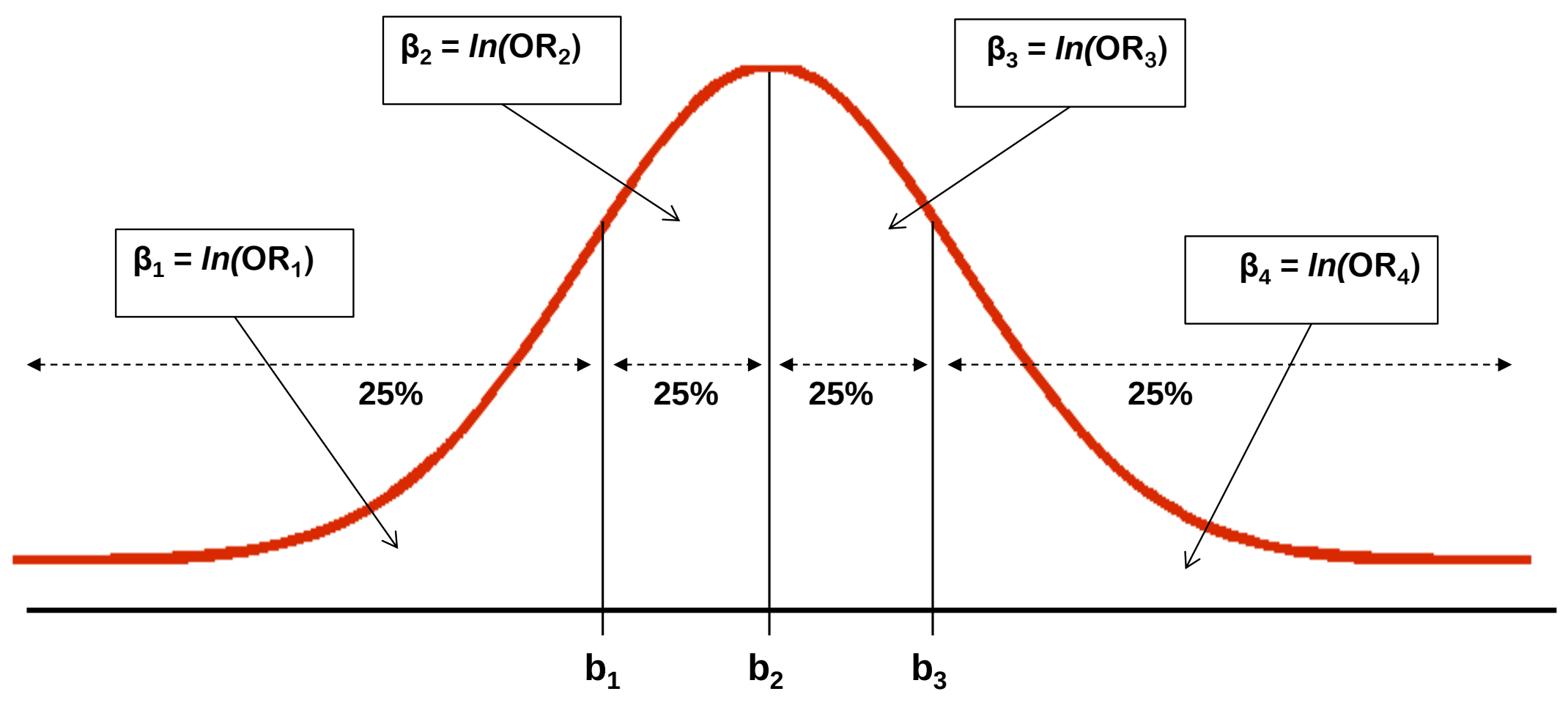
Contributions of markers $\beta_i = ln(OR_i)$:

P's genetic risk for disease X:


$\cdots$ SNP_{m-1}^P SNP_m^P SNP_{m+1}^P $\cdots$ SNP_n^P $\cdots$ SNP_k^P $\cdots$
SNP_m 0 1 2
SNP_n 0 1 2
SNP_k 0 1 2
β_m
β_n
β_k

$$[\boldsymbol{S}] = \left[\alpha + \sum_{i \in \{m,n,k\}} \beta_i \, SNP_i^P(X)\right]$$

B. GENETIC RISK CATEGORIZATION



- For clinical use explanatory variables like the genetic risk should be categorized based on their risk group.
- A private preserving comparison algorithm between **SPU** and **MU** allows to compare encrypted values.

- Let $[G(\boldsymbol{S}, b_i)]$ be the encrypted result of the comparison between $\boldsymbol{S}$ and b_i thus the encrypted genetic regression coefficient $[\beta_G]$ can be computed as follows:

$G(\boldsymbol{S}, b_i)] = 1 \leftrightarrow \boldsymbol{S} \geq b_i$

$G(\boldsymbol{S}, b_i) = 0 \leftrightarrow \boldsymbol{S} < b_i$

$$[\beta_G] = \left[\beta_1\big(1 - G(\boldsymbol{S}, b_1)\big) + \sum_{i=2}^{(k-1)} \beta_i\big(G(\boldsymbol{S}, b_{i-1}) - G(\boldsymbol{S}, b_i)\big) + \beta_k C(\boldsymbol{S}, b_{k-1})\right]$$

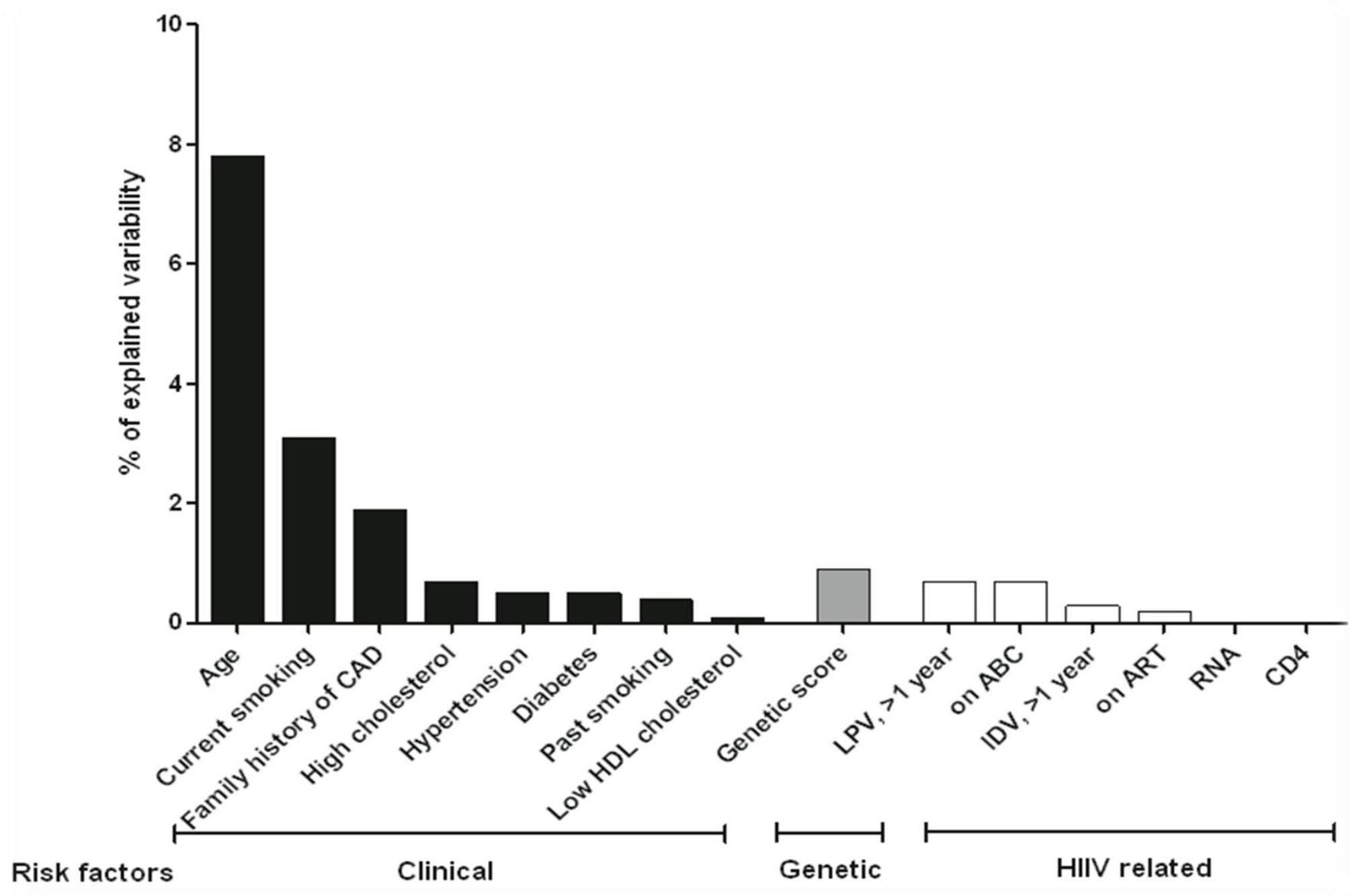
C. FINAL DISEASE RISK COMPUTATION

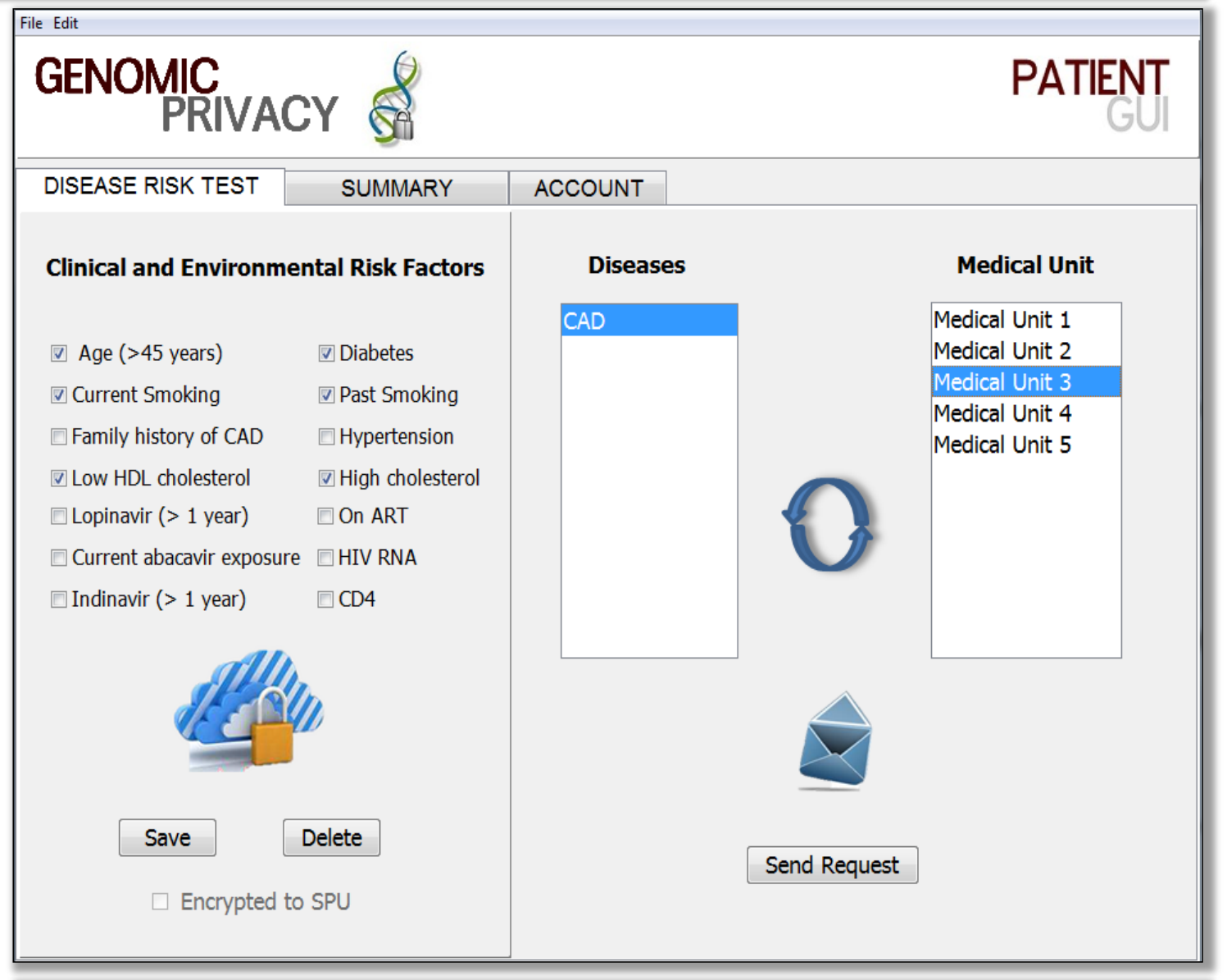
- Let $\mathbb{N} = \{[N_1], [N_2], \dots, [N_m]\}$ be the encrypted non-genomic attributes of the patients, the final disease risk is computed as follows:

$$[\beta_f] = \left[\beta_0 + \beta_G + \sum_{i=1}^{m} \bar{\beta}_i N_i\right] \Longrightarrow P(disease) = \frac{e^{\beta_f}}{1 + e^{\beta_f}}$$

5. Evaluation on Real Data

- Intel Core i7-2620M CPU with 2.70 GHz processor.
- Size of the security parameter: 4096 bits.
- Real SNP profiles from 1000 Genomes Project.
- Coronary artery disease (CAD) risk factors (23 SNPs, 14 non-genomic factors).
- Java implementation.





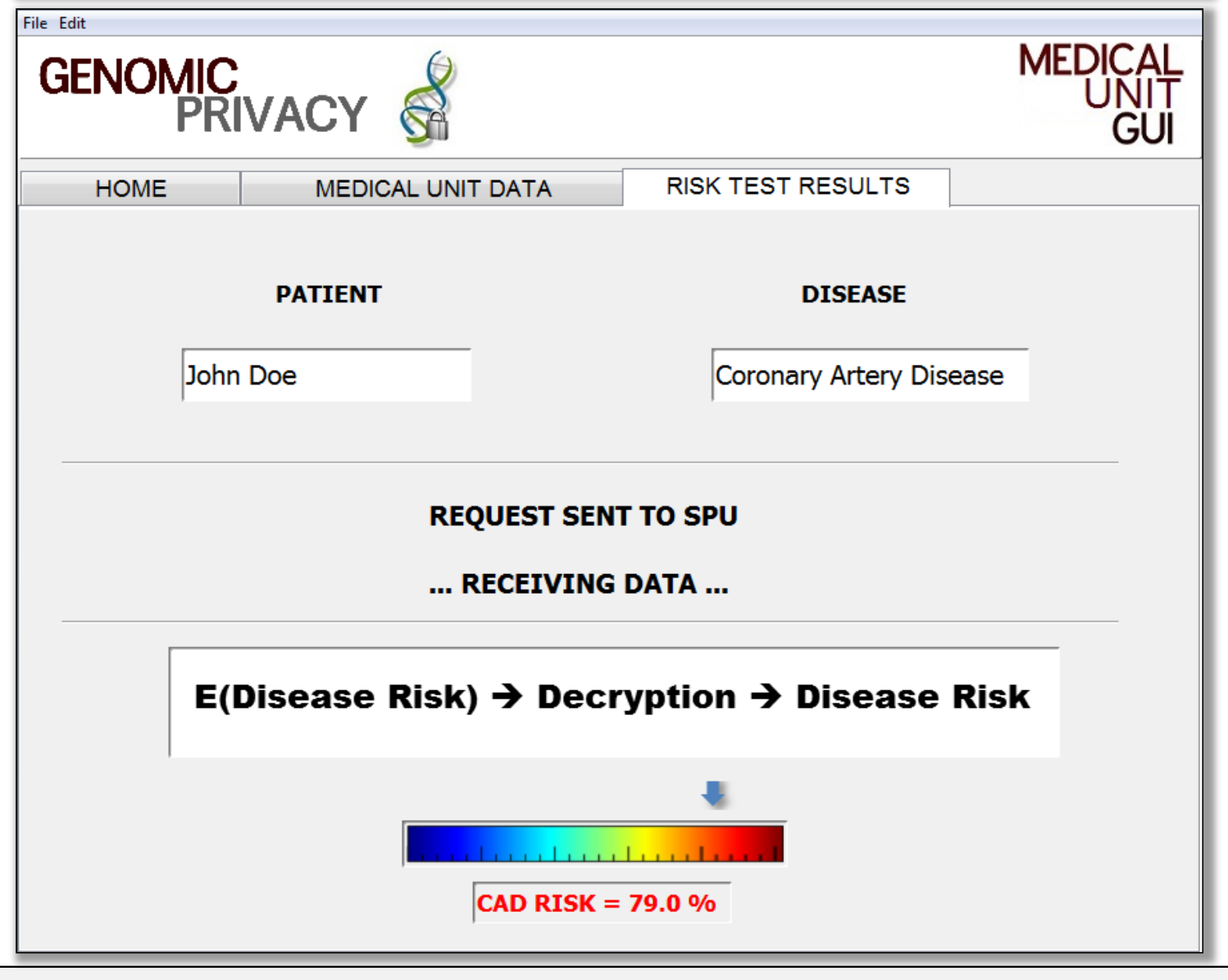
Clinical and Environmental Risk Factors

☑ Age (>45 years) ☑ Diabetes
☑ Current Smoking ☑ Past Smoking
☐ Family history of CAD ☐ Hypertension
☑ Low HDL cholesterol ☑ High cholesterol
☐ Lopinavir (> 1 year) ☐ On ART
☐ Current abacavir exposure ☐ HIV RNA
☐ Indinavir (> 1 year) ☐ CD4

☐ Encrypted to SPU

Diseases: CAD

Medical Unit: Medical Unit 1, Medical Unit 2, Medical Unit 3 (selected), Medical Unit 4, Medical Unit 5



PATIENT: John Doe

DISEASE: Coronary Artery Disease

REQUEST SENT TO SPU

... RECEIVING DATA ...

E(Disease Risk) → Decryption → Disease Risk

CAD RISK = 79.0 %

Complexity of the Proposed System				
Encryption	**Storage**	**Computation of disease risk**		
380 ms./attribute (with pre-computed values: 0.168 ms./attribute)	51.2 GB per patient	Computation of the genetic risk	Privacy-preserving integer comparison	Computation of the final risk
		230 sec (23 SNPs)	3.390 sec (3 comparisons)	140 sec (14 environmental factors)
		Total: 373.432 sec		

E. Ayday, J. L.Raisaro, P. J. McLaren, J. Fellay, and J.-P. Hubaux. ***Privacy-Preserving Computation of Disease Risk by Using Genomic, Clinical, and Environmental Data***. *USENIX Security Workshop on Health Information Technologies (HealthTech '13), Aug. 2013.*

Web link: http://lca.epfl.ch/projects/genomic-privacy/