UNIL | Université de Lausanne

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

CHUV

MedCo: Enabling Privacy-Conscious Exploration of (Encrypted) Distributed Clinical and Genomic Data

Jean Louis Raisaro, Juan Ramón Troncoso-Pastoriza, Mickaël Misbach, João Sá Sousa, Sylvain Pradervand, Edoardo Missiaglia, Olivier Michielin, Bryan Ford and Jean-Pierre Hubaux.

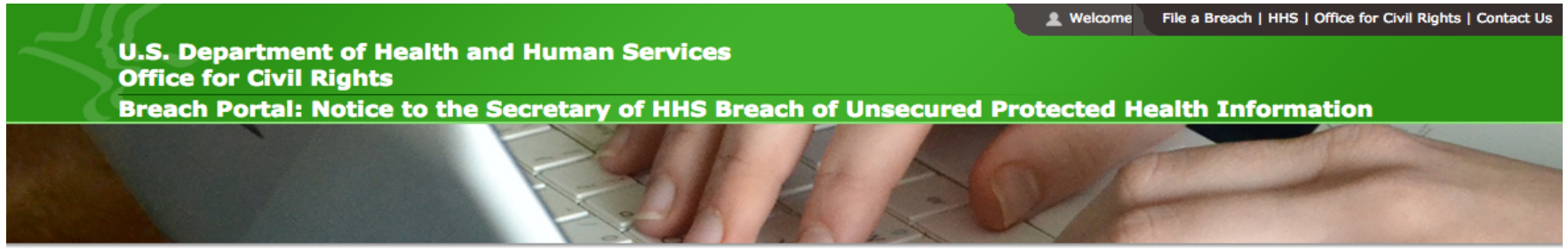
Contact: jean.raisaro@epfl.ch

October 5th – 6th , 2017
Paris, France

Growing Concern: Medical Data Breaches

Around 2 declared breaches per week, each affecting 500+ people

https://ocrportal.hhs.gov/ocr/breach/breach_report.jsf



Under Investigation | Archive | Help for Consumers

As required by section 13402(e)(4) of the HITECH Act, the Secretary must post a list of breaches of unsecured protected health information affecting 500 or more individuals. The following breaches have been reported to the Secretary:

Cases Currently Under Investigation

This page lists all breaches reported within the last 24 months that are currently under investigation by the Office for Civil Rights.

Show Advanced Options

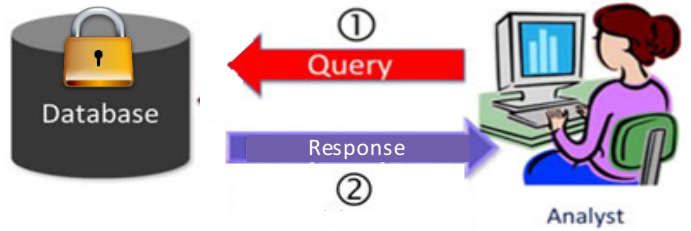
Breach Report Results

Expand All	Name of Covered Entity	State	Covered Entity Type	Individuals Affected	Breach Submission Date	Type of Breach	Location of Breached Information
	Mercy Family Medicine	CO	Healthcare Provider	2069	08/16/2017	Loss	Other Portable Electronic Device
	MJHS Home Care	NY	Healthcare Provider	6000	08/11/2017	Hacking/IT Incident	Email
	Pacific Alliance Medical Center	CA	Healthcare Provider	266123	08/10/2017	Hacking/IT Incident	Network Server
	MDeverywhere, Inc.	TX	Business Associate	1396	08/10/2017	Unauthorized Access/Disclosure	Other

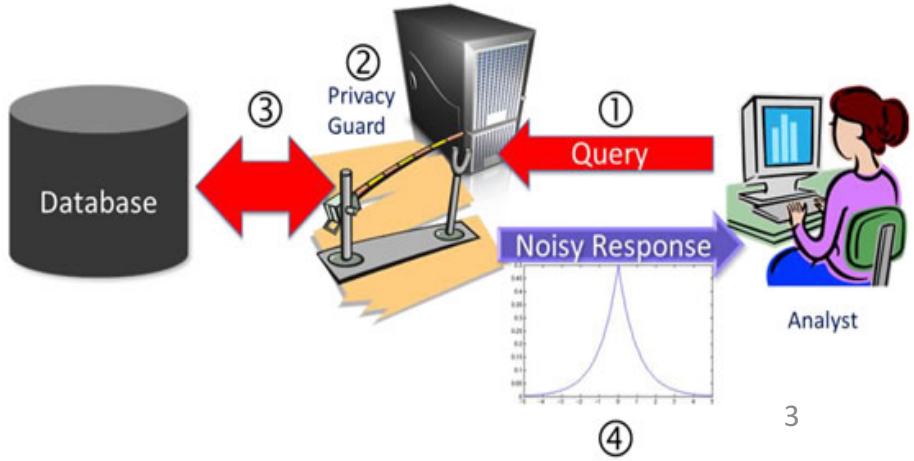
Privacy-Enhancing Technologies

Two main approaches:

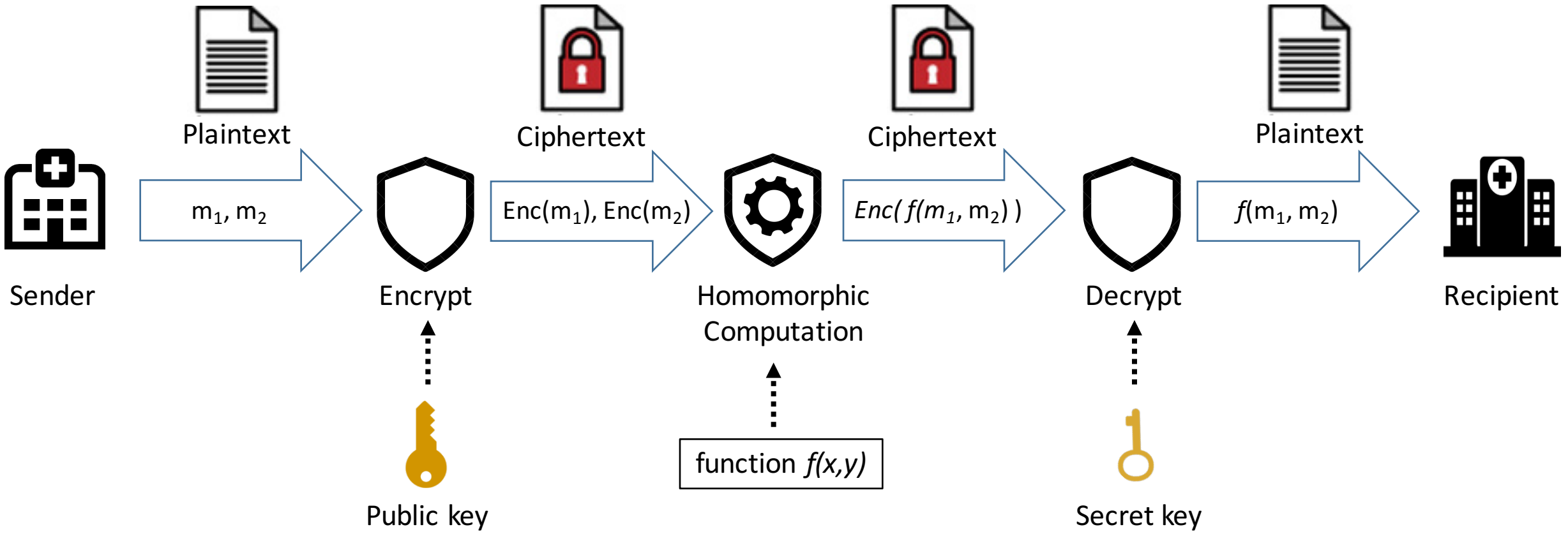
- Protect the data themselves:
  Use of **cryptography**
  - Symmetric / asymmetric encryption
  - Property-preserving encryption
  - (Partially) homomorphic encryption
  - …



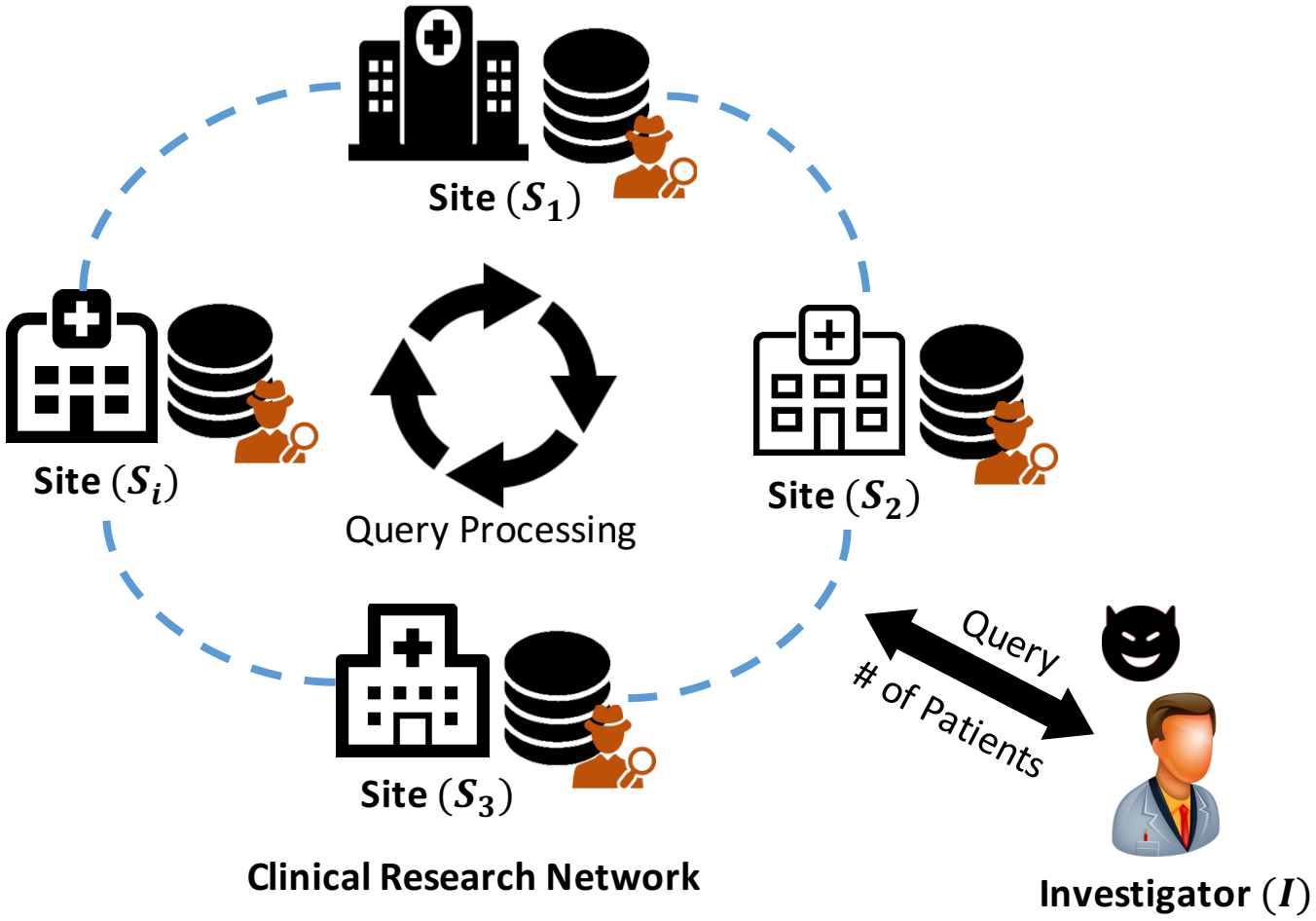
- Avoid that responses leak "too much" information:
  Provide only **global** (e.g., statistical) **results**
  - K-anonymity, l-diversity, t-closeness
  - Differential privacy
  - For genomics, see "Homer attack" and subsequent ones



Crypto 101 – Homomorphic Encryption



System and Threat Models



Honest-but-curious adversary:

- honestly follows the protocol
- tries to infer sensitive data from the different steps of the protocol

Malicious-but-covert adversary:

- can tamper with the protocol
- tries to infer sensitive data from the query end-result

What are the main concerns?

- **Loss of data confidentiality** due to illegitimate access to the data
  - External (hacker) or internal (insider) attacker stealing the data

➔ Standard encryption can protect data ONLY at rest or in transit BUT NOT during processing (e.g., in the memory)

- **Patient re-identification** due to legitimate access to the data
  - Malicious users performing "smart" data requests in order to re-identify patients in a specific dataset (e.g., patients with HIV)

➔ De-identification or anonymization is ineffective with genomic data

Main Requirements

Functionality:

COUNT(patients**)/SELECT(**patients**)**

FROM database

WHERE * **AND/OR** *

GROUP BY *

* represents any possible concepts in the otology

Security/Privacy:

- Protection of data confidentiality at rest, in transit and **during computation**
- no single point of failure
- only the investigator can obtain the query end-result
- (optional) unlinkability
- (optional) differential privacy

MedCo: Combining the Best of Both Worlds

Biomedical Informatics:

- Data model from *i2b2*

  Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, Kohane I. *Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2)*. Journal of the American Medical Informatics Association. 2010 Mar 1;17(2):124-30.

- Interoperability layer from *SHRINE*

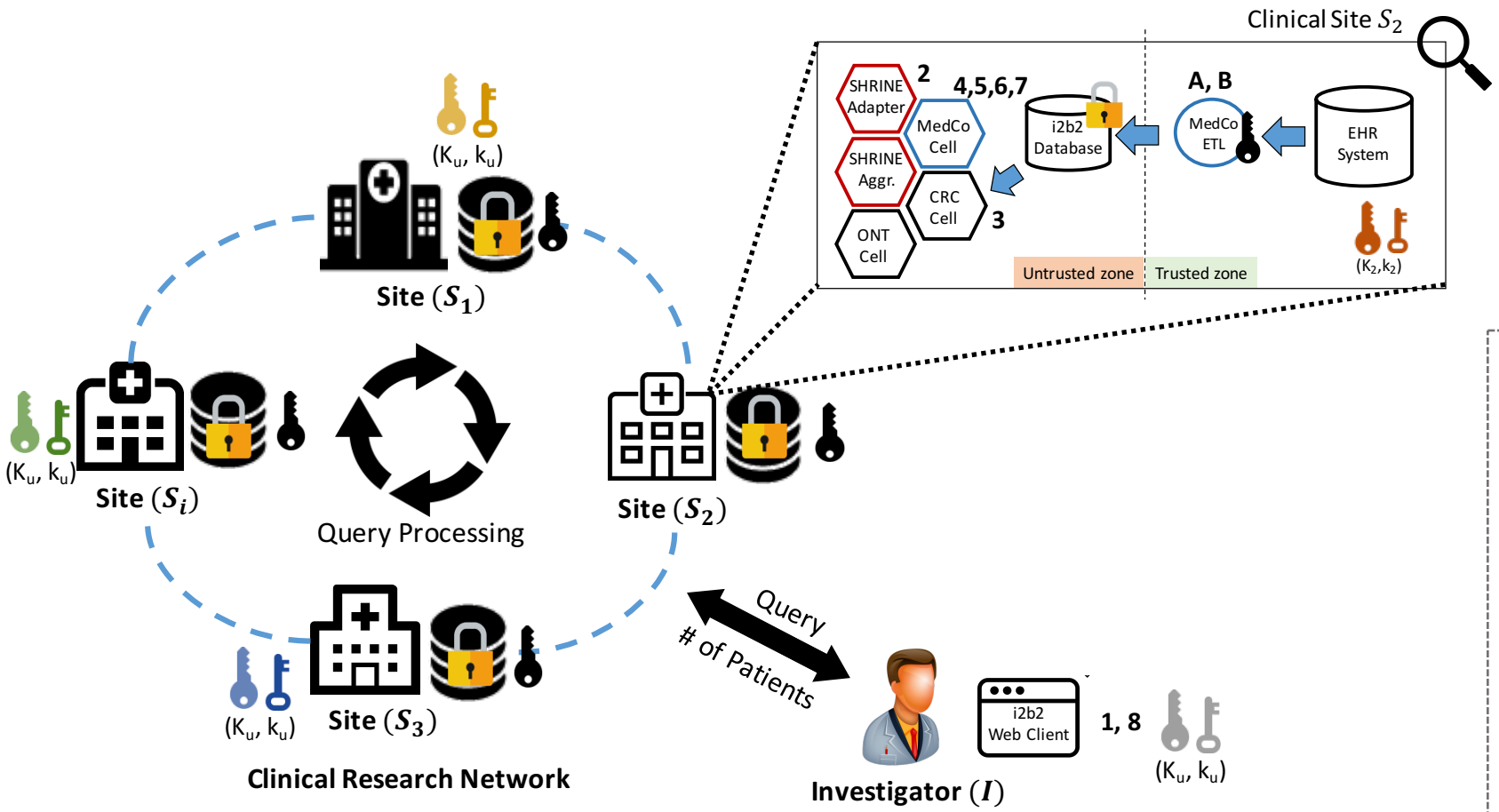
  McMurry AJ, Murphy SN, MacFadden D, Weber G, Simons WW, Orechia J, Bickel J, Wattanasin N, Gilbert C, Trevvett P, Churchill S. *SHRINE: enabling nationally scalable multi-site disease studies*. PloS one. 2013 Mar 7;8(3):e55811.

IT Privacy and Security:

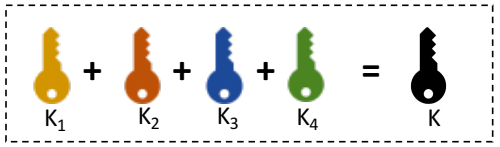
- Privacy-preserving distributed protocols from *UnLynx*

  Froelicher, D., Egger, P., Sousa, J.S., **Raisaro, J.L.,** Huang, Z., Mouchet, C., Ford, B. and Hubaux, J.P., 2017. UnLynx: A Decentralized System for Privacy-Conscious Data Sharing. In *Proceedings on Privacy Enhancing Technologies* (Vol. 4, pp. 152-170).

MedCo: Core Architecture & Protocol



Initialization phase



MedCo Protocol:

A, B) ETL & Encryption Phase

1) (user) Query Generation
2) (local) Query Analysis
3) (local) Query Processing
4) (local) Result Aggregation
5) (local) Result Obfuscation
6) (distributed) Results Shuffling
7) (distributed) Results Re-Encryption
8) (user) Result Decryption

i2b2 SHRINE MedCo

Tests on Clinical Oncology Use Case

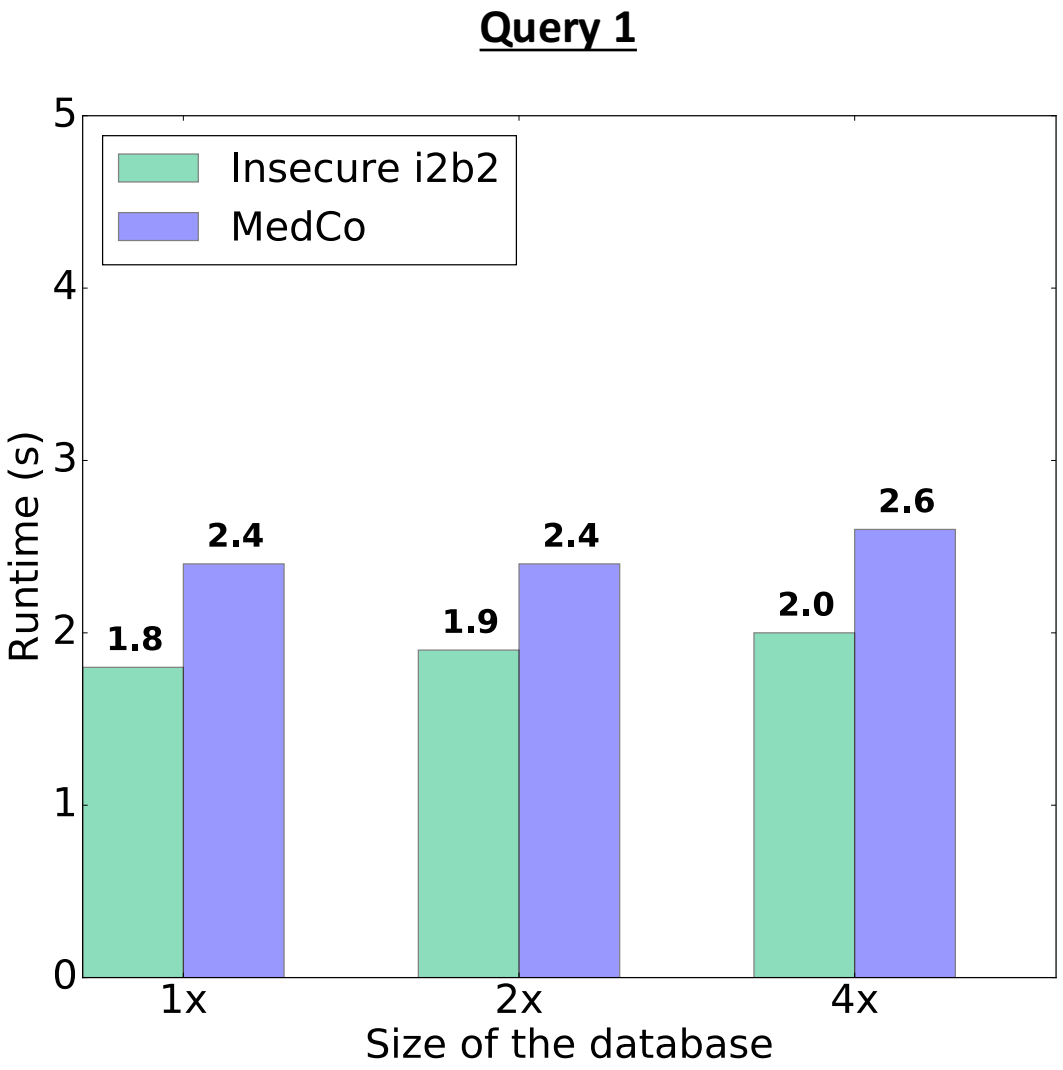
Public Data from cBioPortal

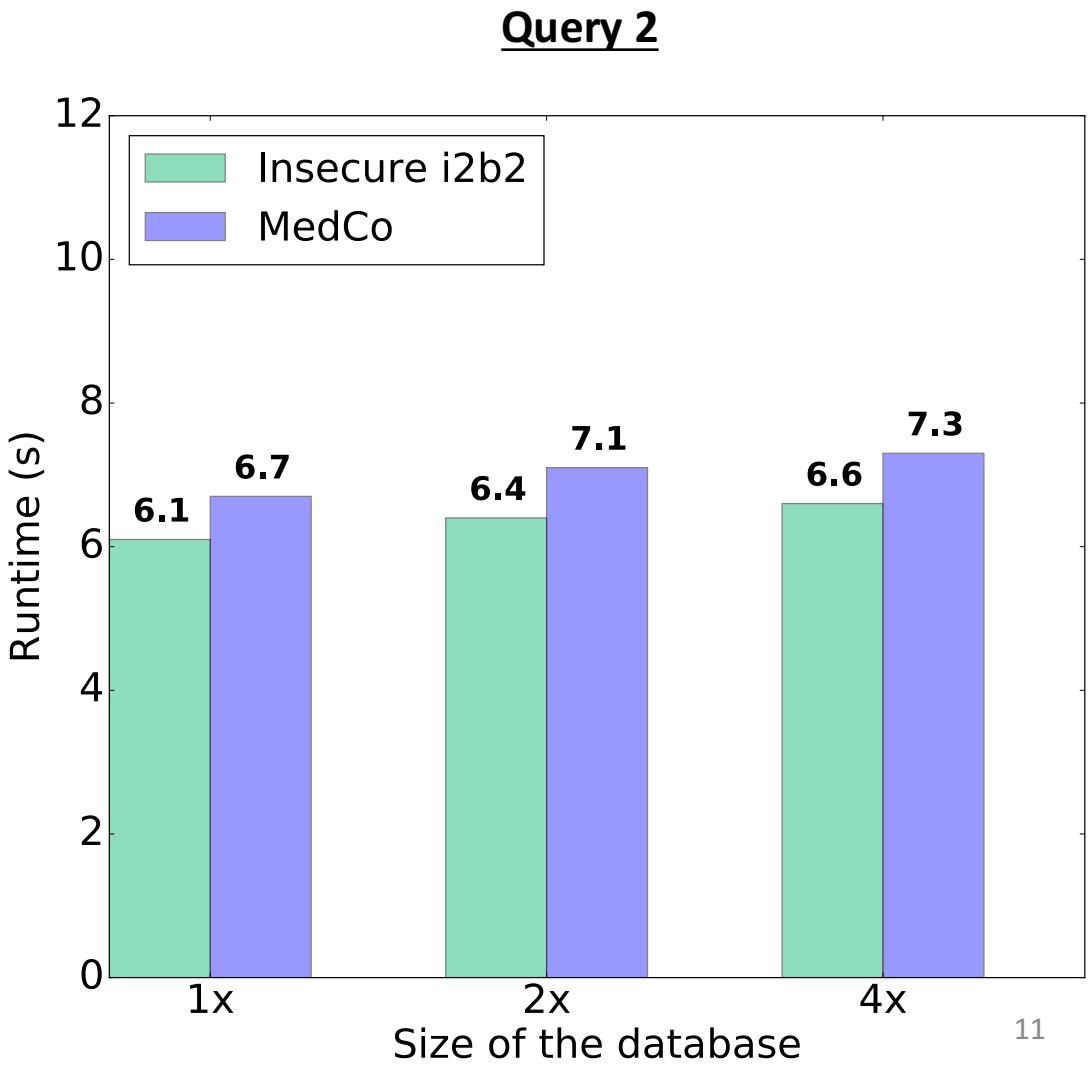
- 121 patients with 9 clinical attributes and 1,978 mutations on average per site
- **Query 1**: "*Number of patients with skin cutaneous melanoma AND a mutation in BRAF gene affecting the protein at position 600*."
  ➔ (2 clinical attributes, 4 mutations)
- **Query 2:** "*Number of patients skin cutaneous melanoma AND a mutation in BRAF gene AND a mutation in (PTEN OR CDKN2A OR MAP2K1 OR MAP2K2 genes)*"
  ➔ (2 clinical attributes, 77 mutations)

Hardware and Software Setting

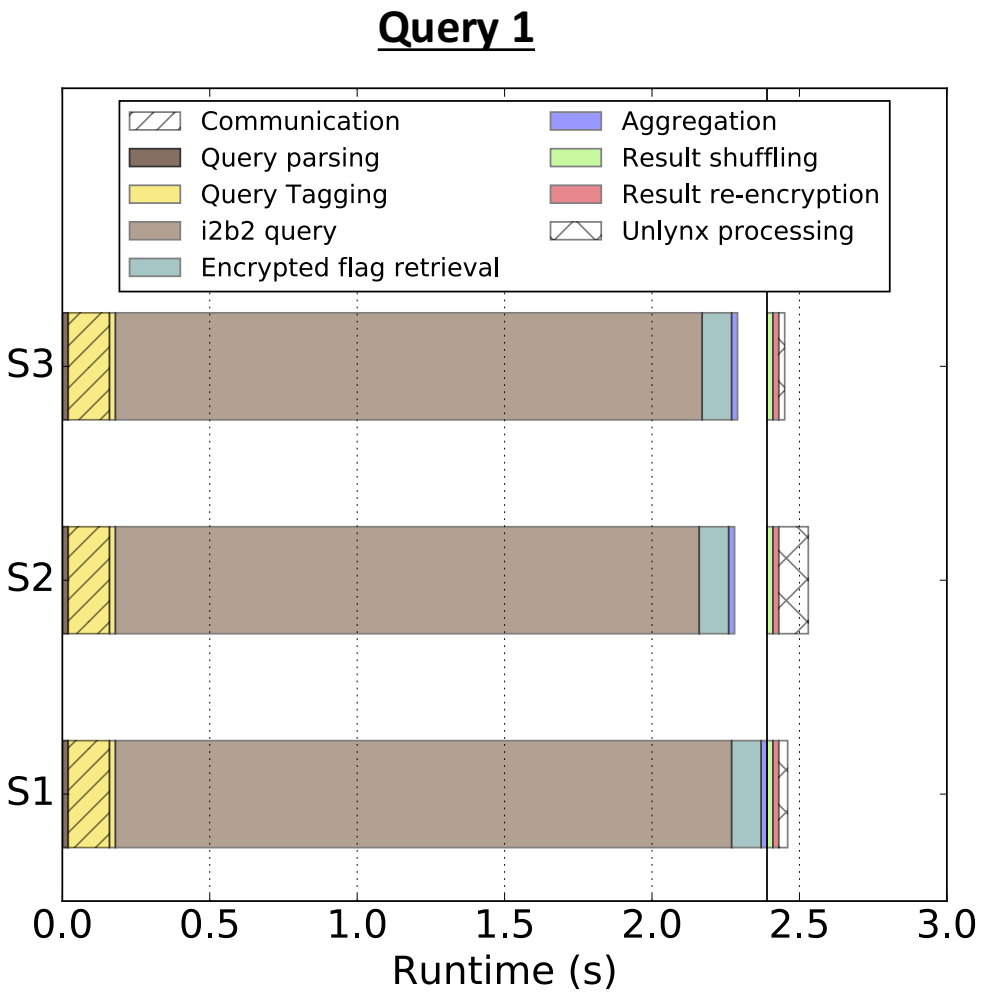
- 3 servers: 2.5GHz Intel Xeon E5-2680 v3 CPUs with 12 cores
- memory: 256GB of RAM
- network: 10 Gbps link
- OS: Ubuntu
- crypto: ElGamal encryption on Ed25519 elliptic curve with 128 bit security
- database: PostgreSQL
- deployment technology: Docker

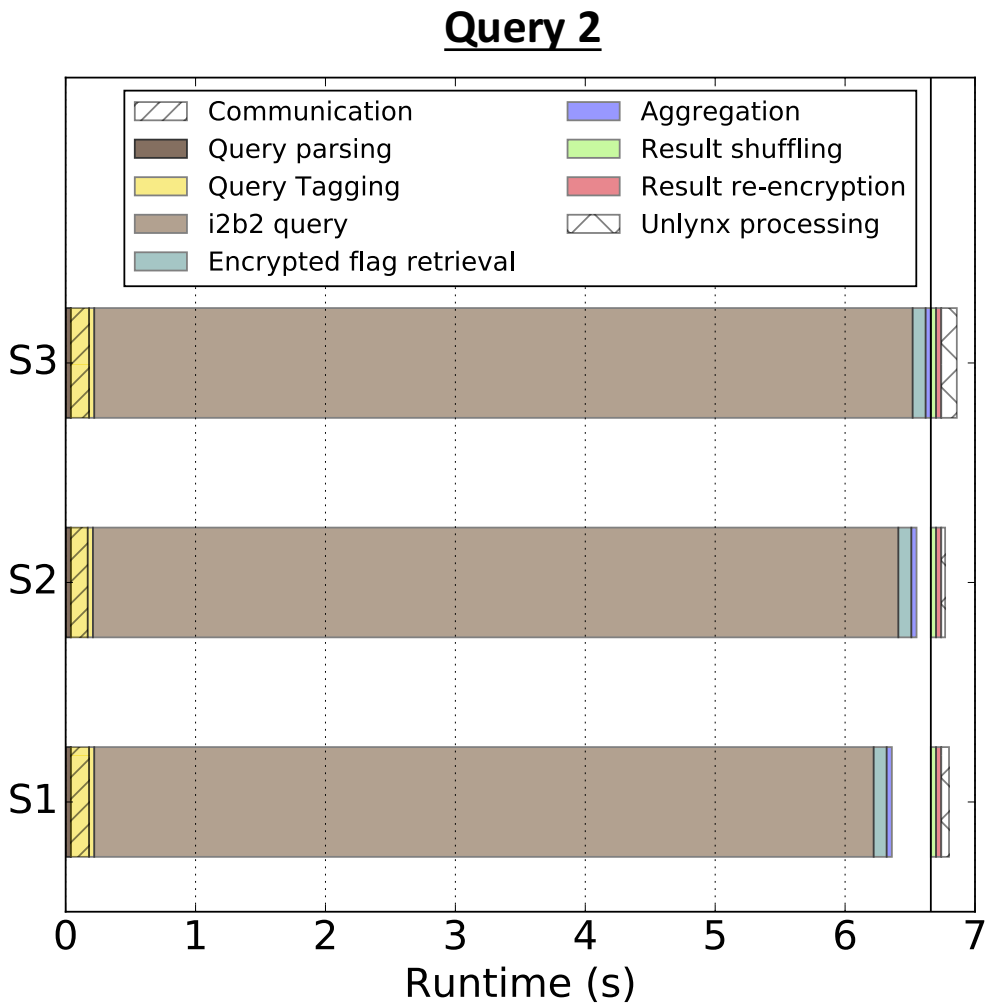
Performance Results: Query Runtime vs. Database Size



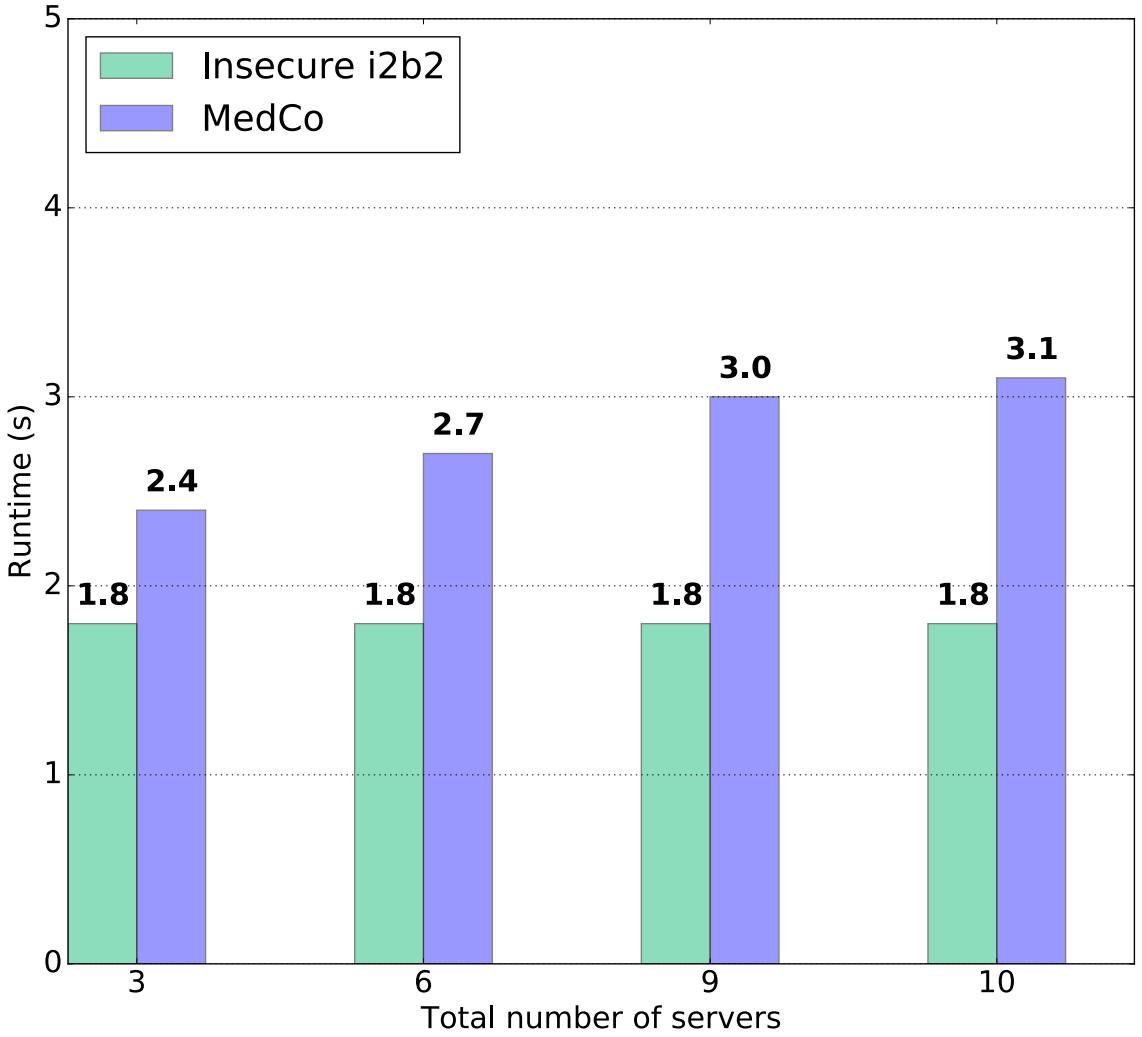


Performance Results: Query-Workflow Breakdown

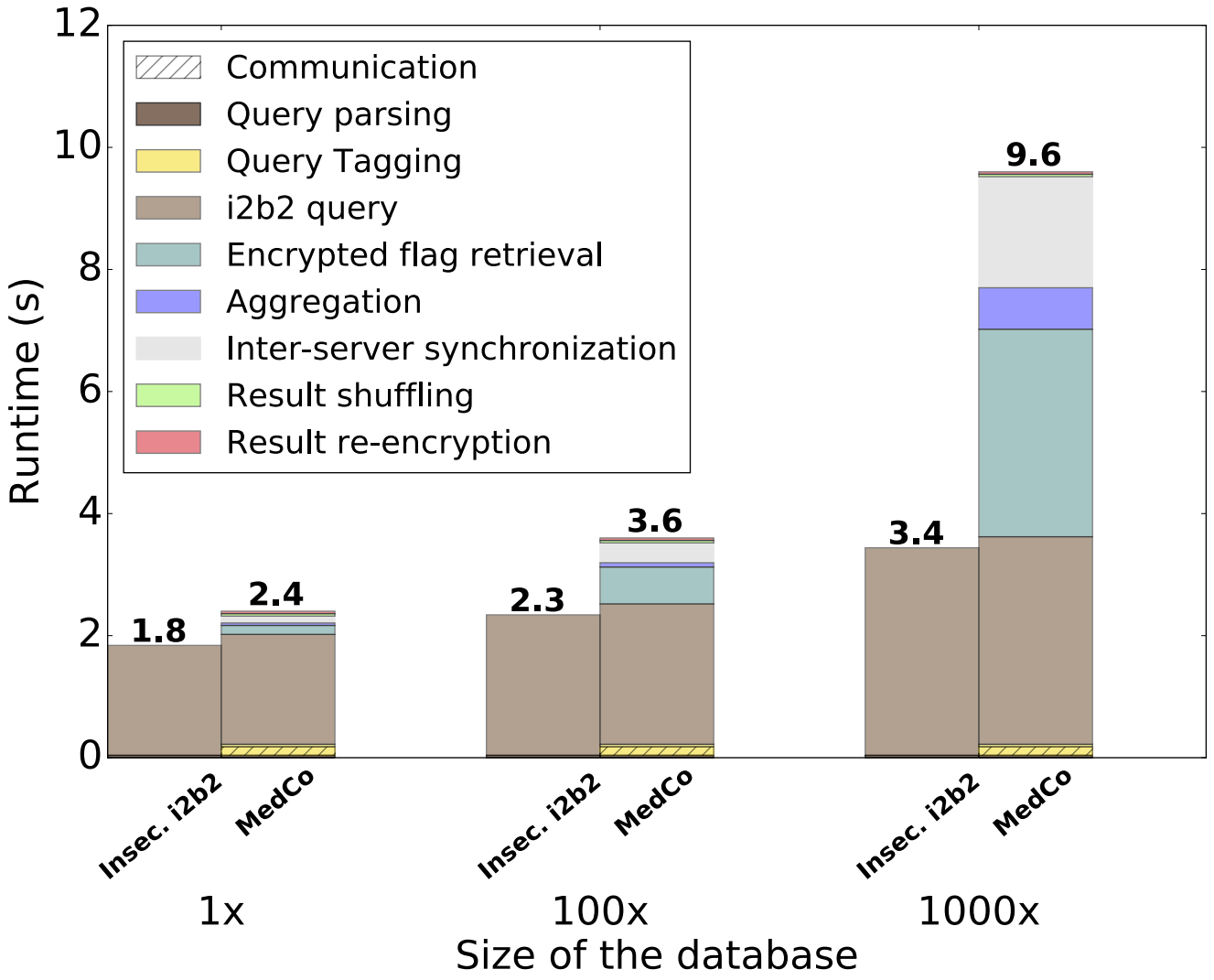




Performance Results: Query Runtime vs. Number of Sites



Performance Results: Query Time vs. Database Size



1000x:

- 40k patients per site
- 80M observations per site

Encryption flag retrieval time: it depends on the number of patients satisfying the query (16k patients in the 1000x case)

Conclusion

We have introduced MedCo:

- **First privacy-preserving system** for exploration of distributed clinical and genomic data
- Based on **collective and homomorphic encryption** and (optionally) differential privacy
- **End-to-end data protection**
- Potential to enable to share clinical data beyond HIPAA/GDPR "limited data set" and genomic data
- **Low overhead** with respect to the unprotected version based on i2b2 & SHRINE
- Easy installation on top of existing i2b2 instances thanks to ***Docker* technology**

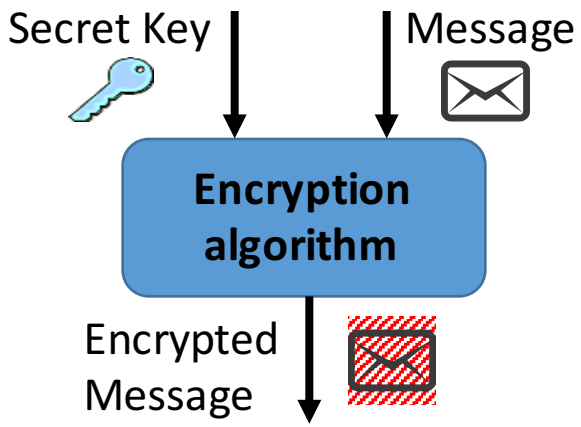
For More Details:

- **Come to see us at our poster!** More on protection of medical and genomic data @http://lca.epfl.ch/projects/genomic-privacy/
- **Full paper** and oral presentation of MedCo at GenoPri'17, co-located with GA4GH Annual Meeting (Orlando Oct. 15th) http://www.genopri.org/

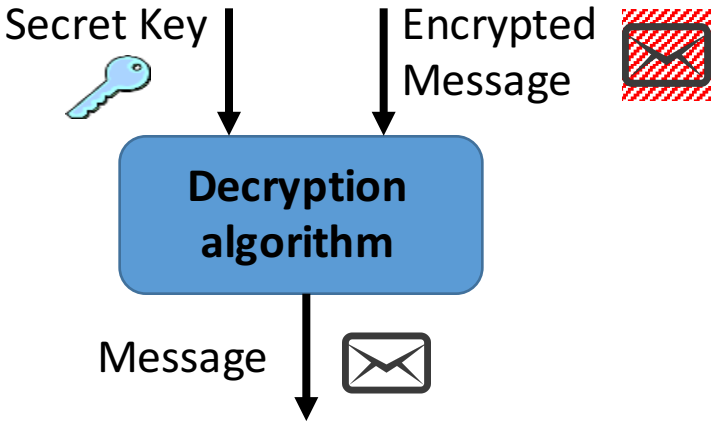
Backup Slides

Crypto 101 – Symmetric Vs. Asymmetric Cryptography

- Symmetric encryption
  - Fast
  - Problem: How to agree on the same secret key?



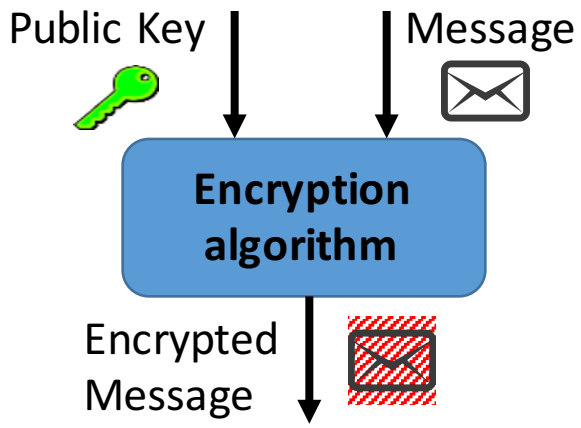
Encryption



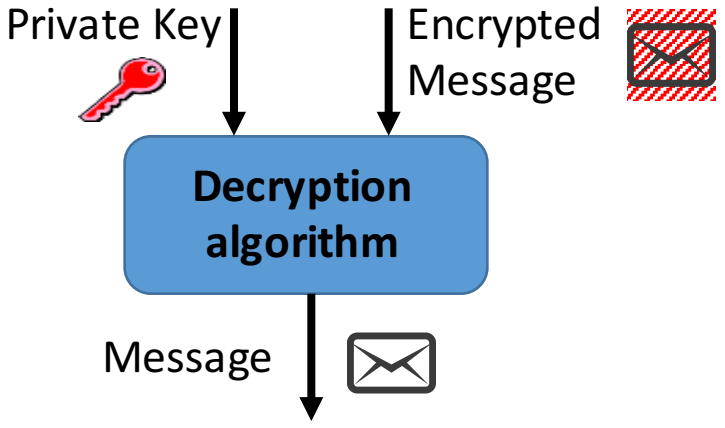
Decryption

Crypto 101 – Symmetric Vs. Asymmetric Cryptography

- Asymmetric encryption
  - Anyone can encrypt. Only the holder of the private key can decrypt.
  - Slow (compared to symmetric encryption)



Encryption



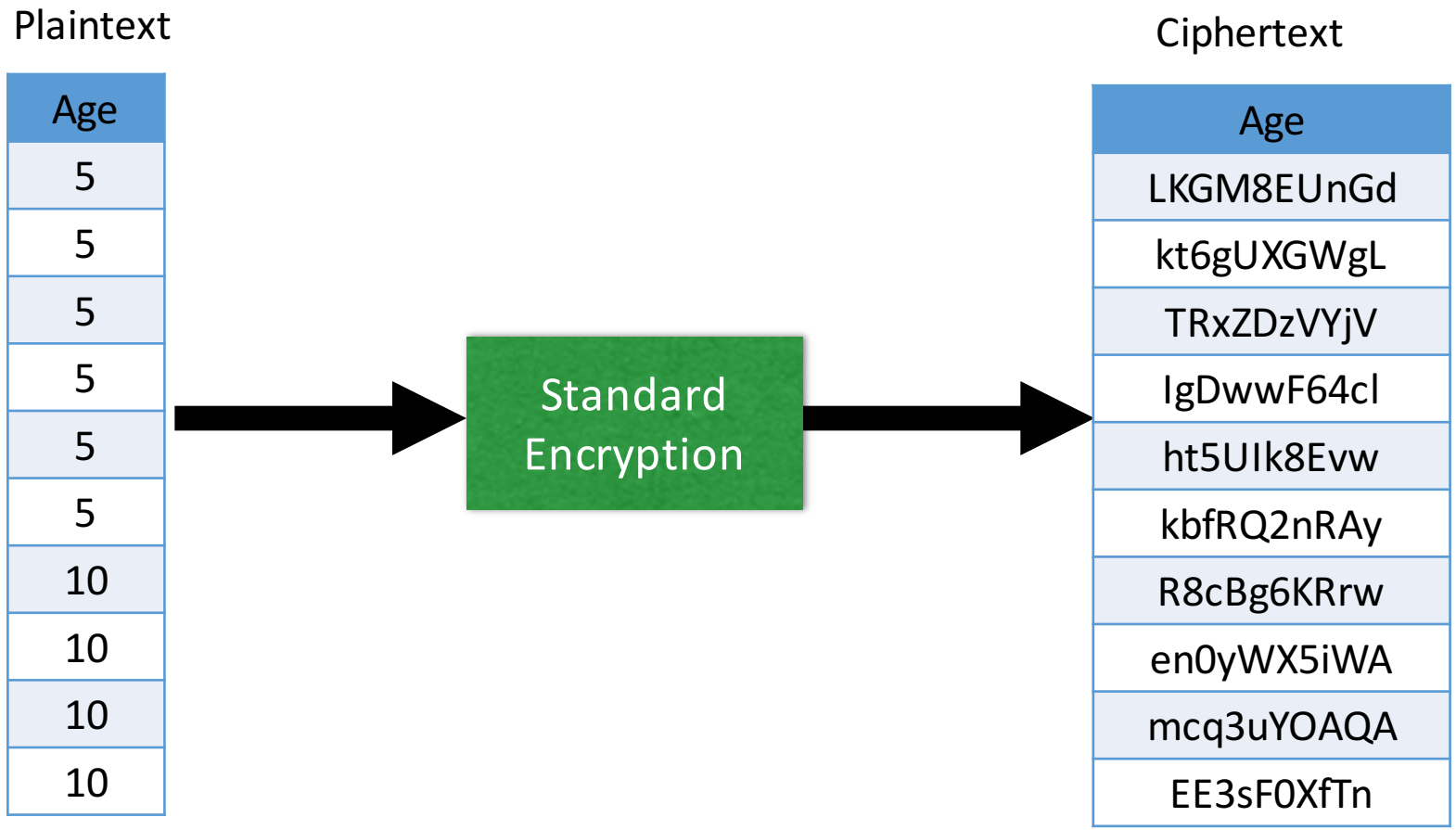
Decryption

Important application: digital signatures

Crypto 101 – Randomness in Standard Encryption

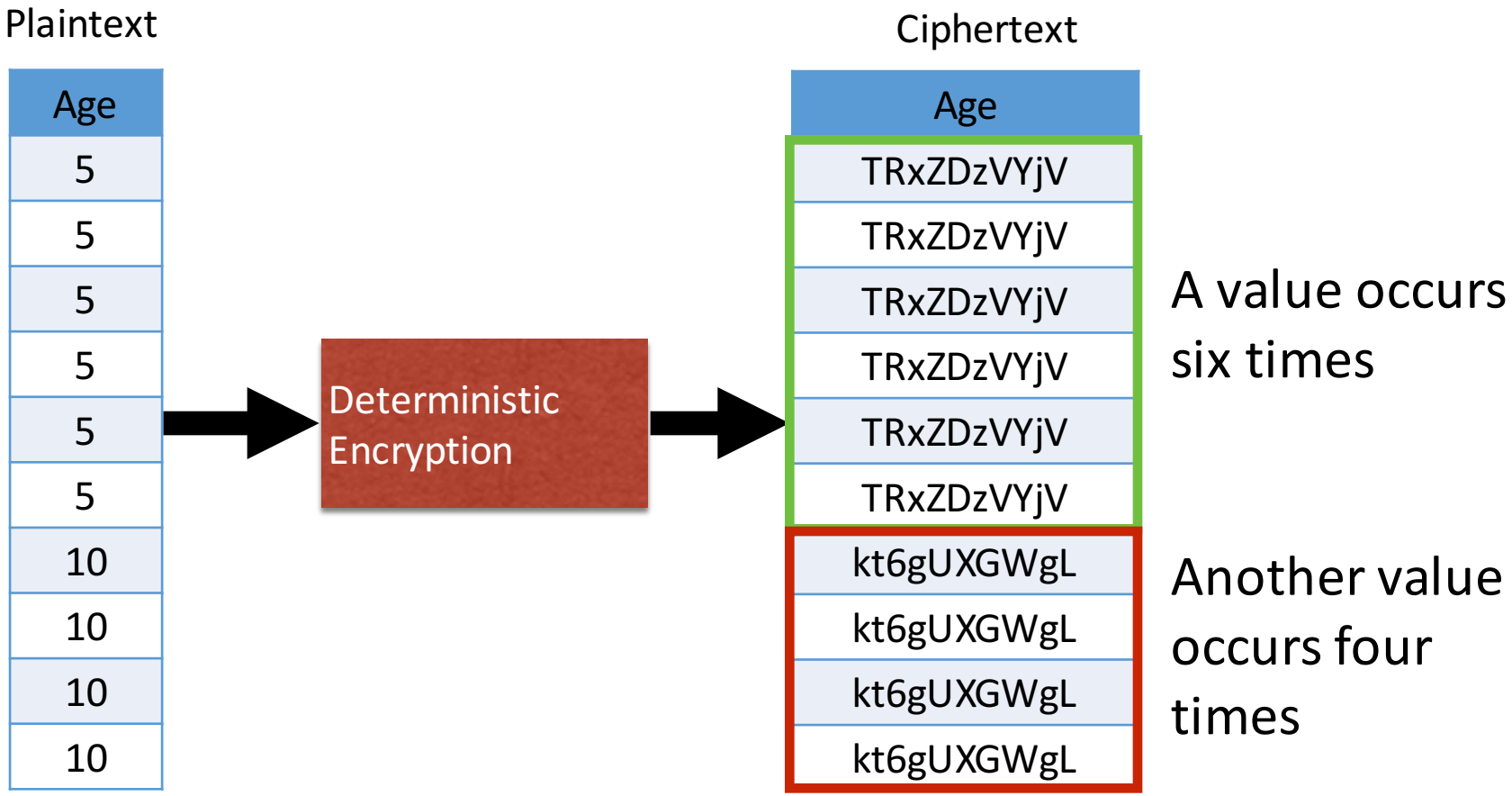
Standard encryption is **not** property-preserving

Semantically secure encryption leaks no partial information about the message

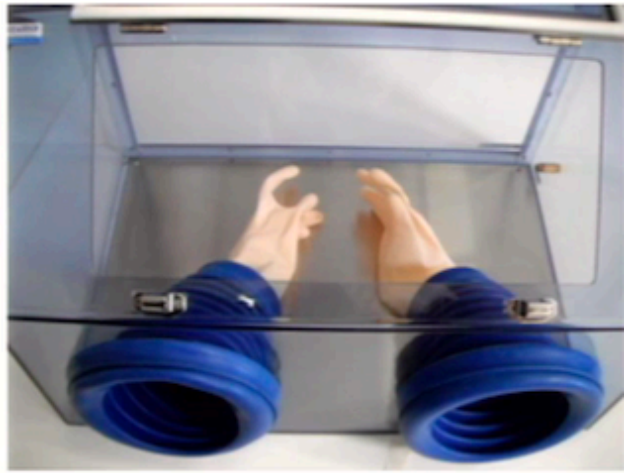


Crypto 101 – Property-Preserving Encryption (PPE)

Deterministic encryption (preserves and leaks equality)



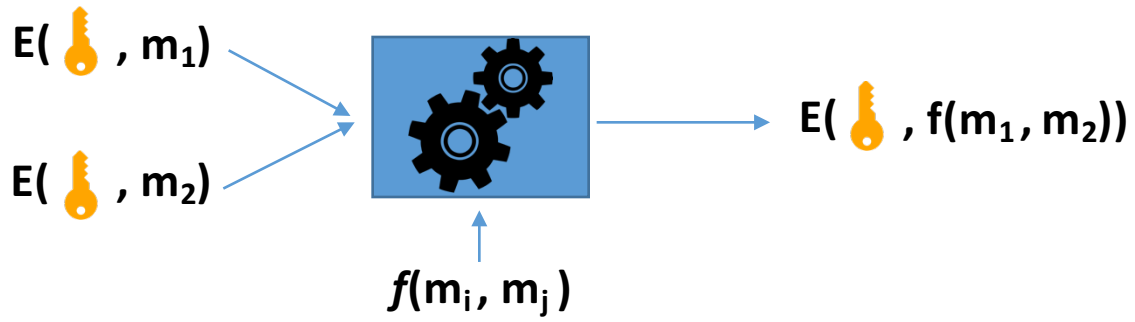
Crypto 101 – Homomorphic Encryption



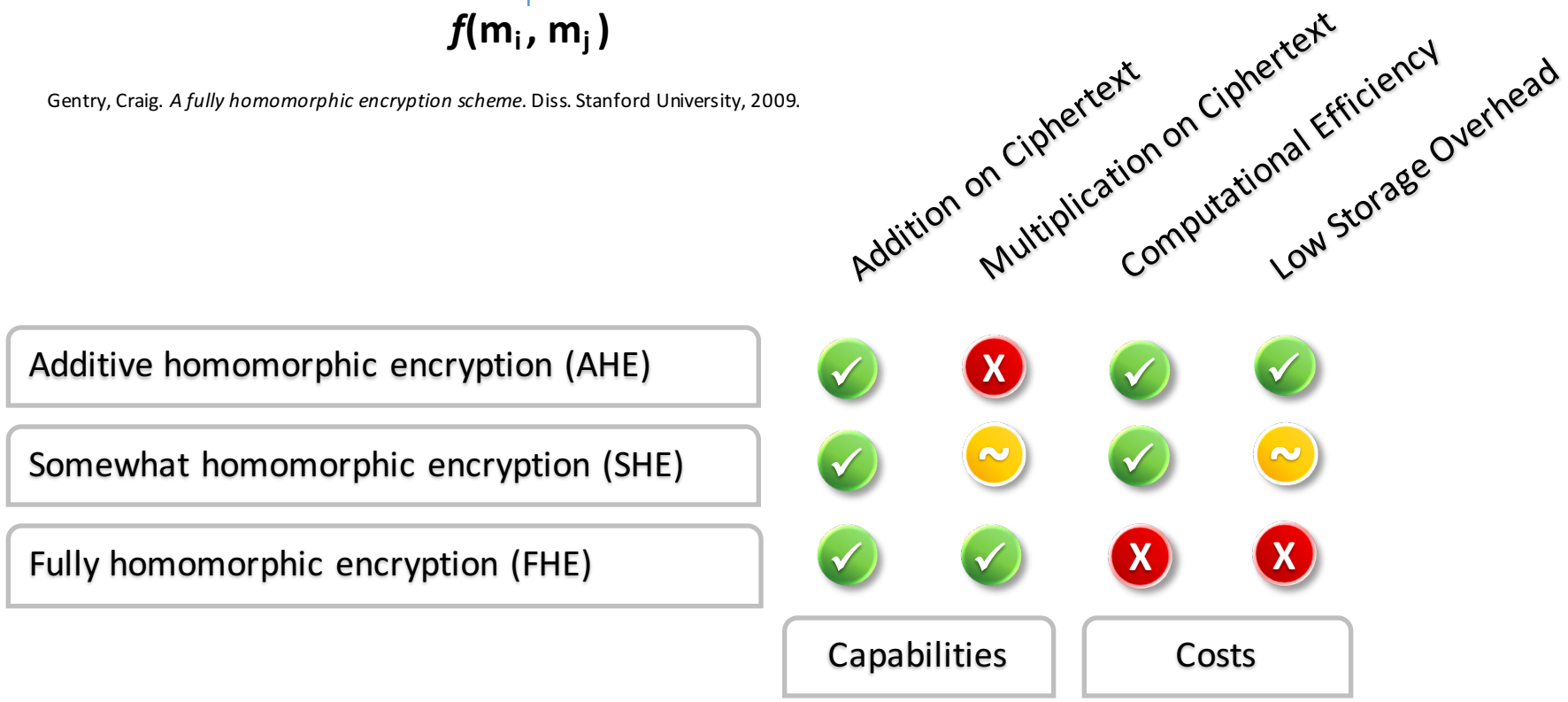
1. Put your gold in a locked box.
2. Keep the key.
3. Let your jeweler work on it through a glove box.
4. Unlock the box when the jeweler is done!

Courtesy Kristin Lauter, Microsoft Research

Crypto 101 – Homomorphic Encryption



Gentry, Craig. *A fully homomorphic encryption scheme*. Diss. Stanford University, 2009.



Differential Privacy: informal definition

Output is similar whether any single individual's record is included in the database or not

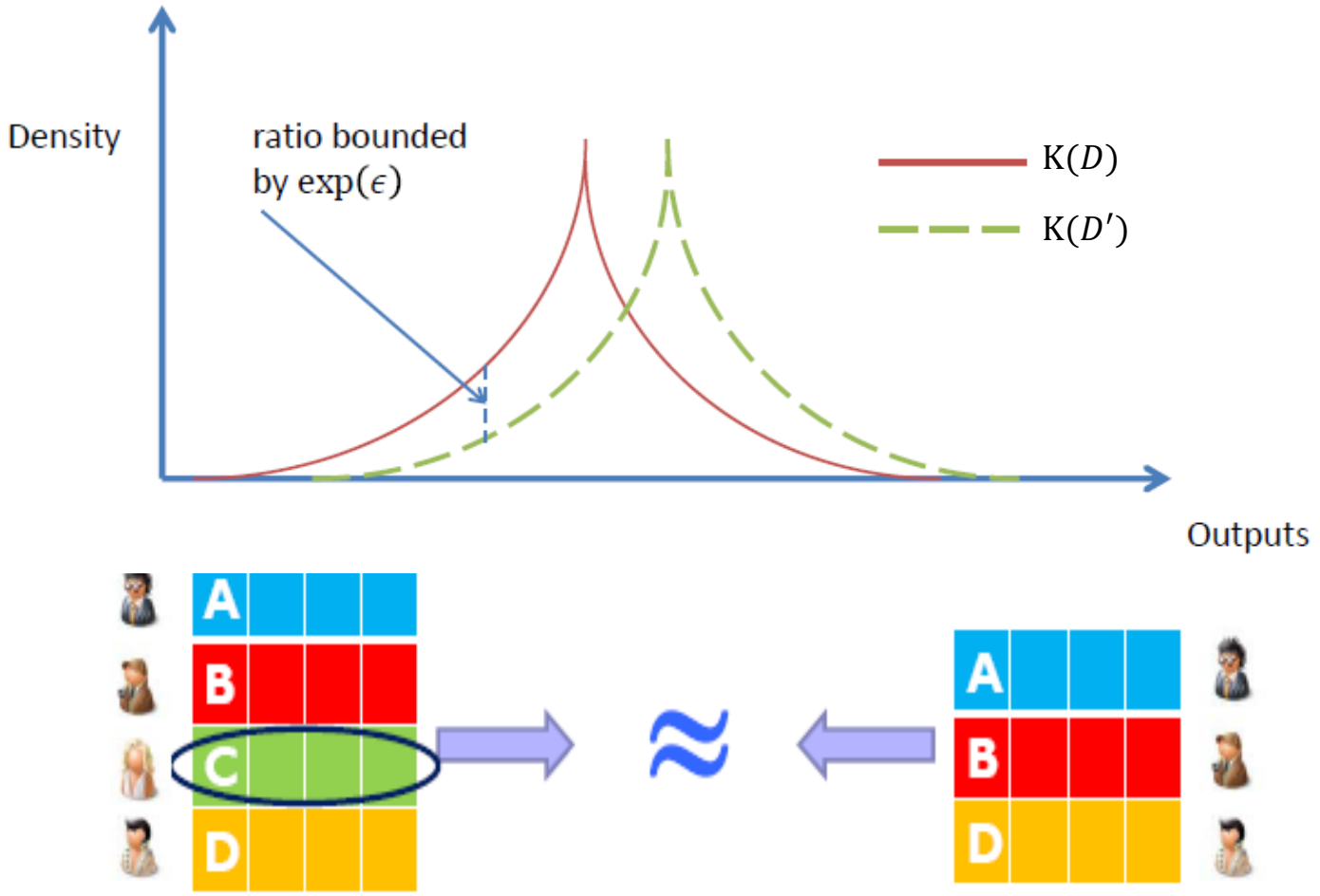
C's inclusion of her record in the computation does not make her *significantly* worse off

If there is already some risk of revealing a secret of C by combining auxiliary information and something learned from DB, then that risk is still there but not *significantly* increased by C's participation in the database

Differential Privacy: formal definition

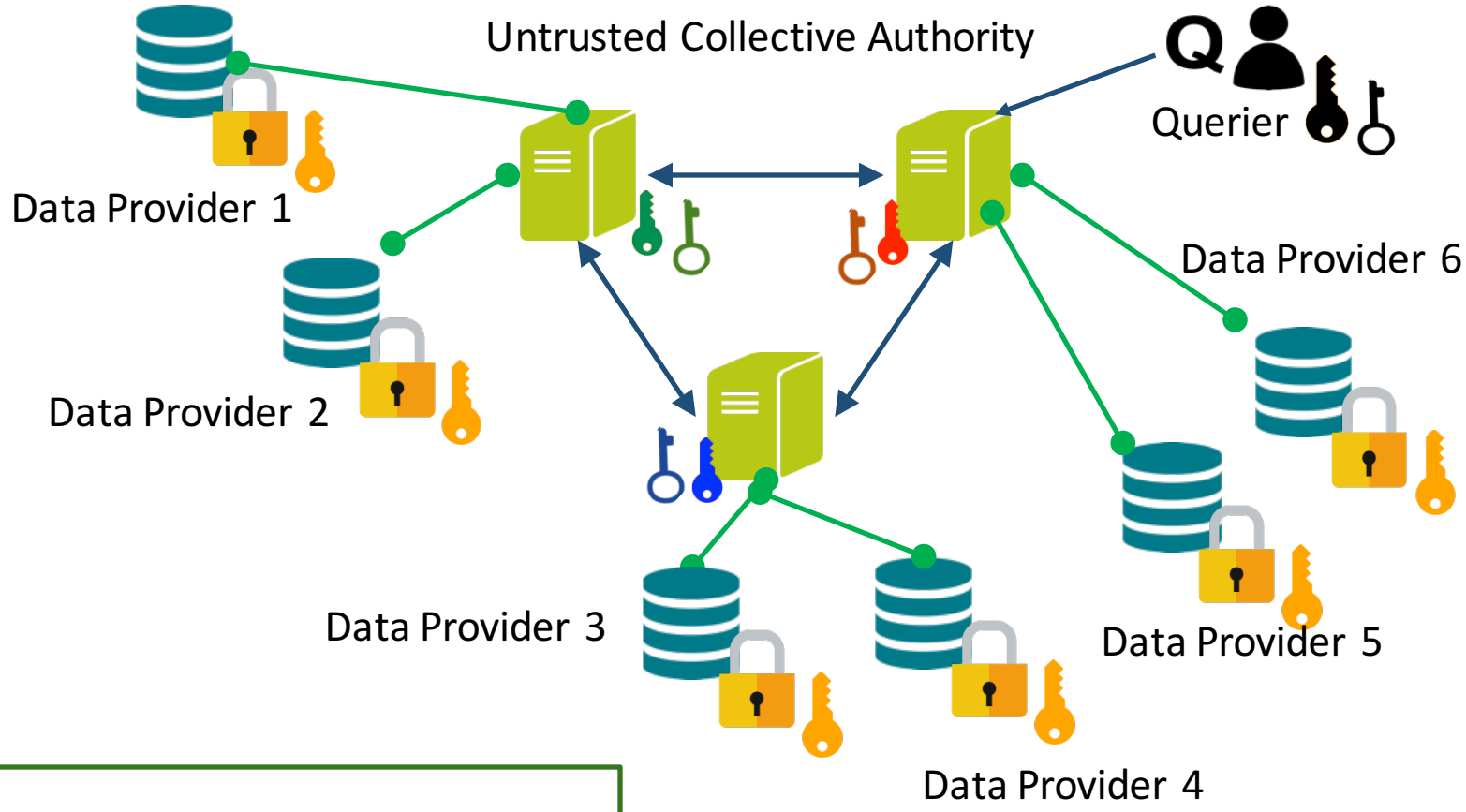
Definition (Differential privacy). A randomized mechanism K is ϵ-differential private if, for all data sets D and D' which differ in at most one individual and for any $t \in \mathbb{R}$ (output space),

$$\frac{\Pr(\mathrm{K}(D) = t)}{\Pr(\mathrm{K}(D') = t)} \leq e^{\epsilon}$$



UnLynx: Framework For Privacy-Conscious Data Sharing

- Trust is shared across a group of servers forming a collective authority
- They collaborate together to generate a collective encryption key
- The collective encryption key is used to encrypt the data and can be compromised only if all servers are compromised



UnLynx: Privacy-Preserving Distributed Protocols

- Distributed Deterministic Tag (DDT) Protocol:

$Det(m_1, \text{key}) \neq Det(m_2, \text{key})$

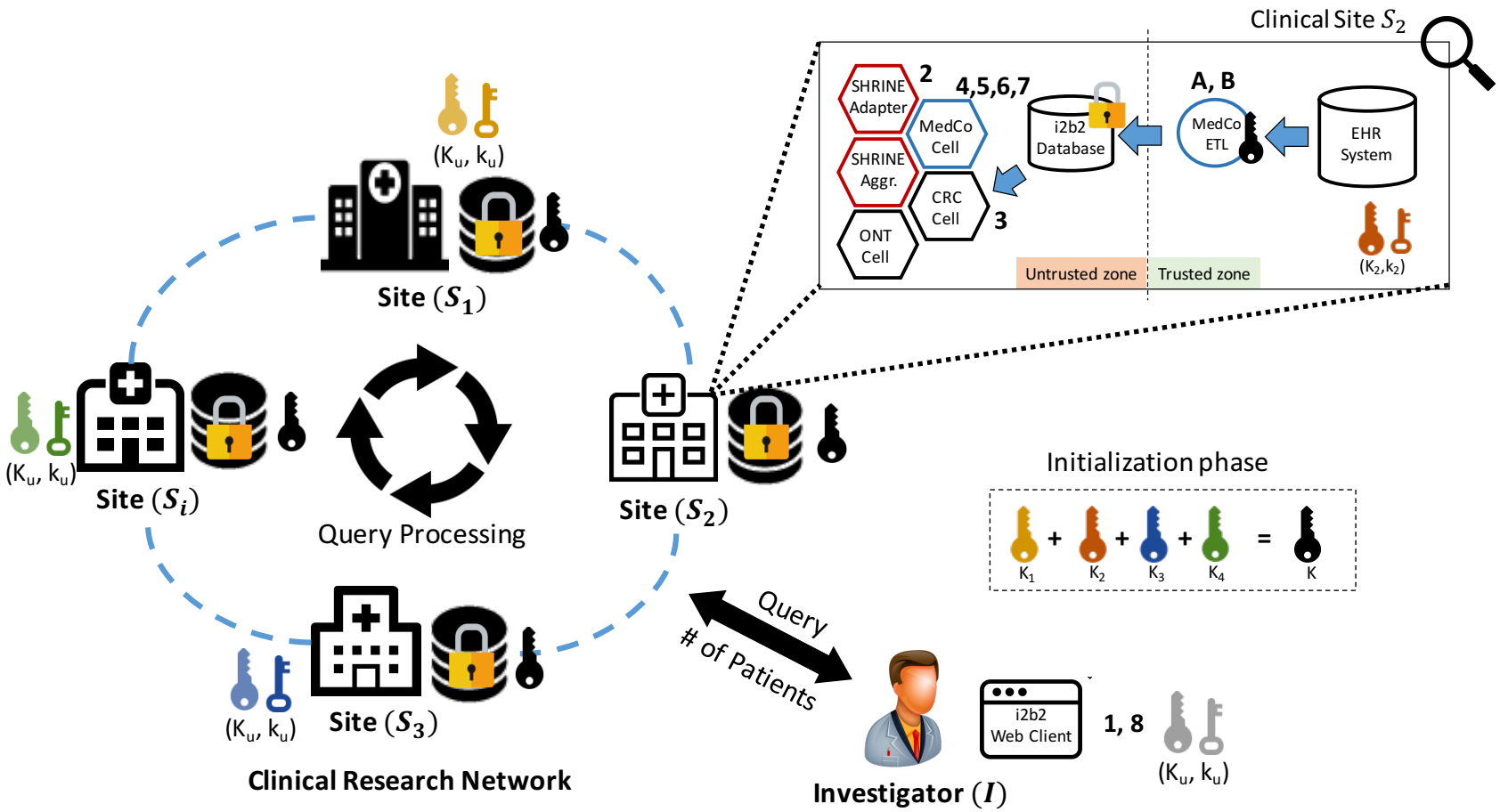
$E(m_1, \text{key}) \approx E(m_1, \text{key})$

- Distributed Verifiable Shuffling (DVS) Protocol:

- Distributed Key Switching (DKS) Protocol:

$D(E(m, \text{key}), \text{key}) = m$

MedCo: Core Architecture & Protocol



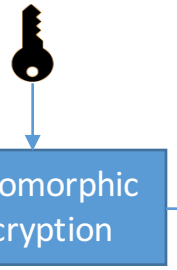
A, B) ETL & Encryption Phase

1) (user) Query Generation
2) (local) Query Analysis
3) (local) Query Processing
4) (local) Result Aggregation
5) (local) Result Obfuscation
6) (distributed) Results Shuffling
7) (distributed) Results Re-Encryption
8) (user) Result Decryption

i2b2 SHRINE MedCo

MedCo: ETL & Encryption Phase

Observation Fact	
Patient_Num	Concept_CD
0000001	code_1
0000001	code_2
0000001	code_3
0000002	code_1
0000002	code_4
0000003	code_2
0000003	code_3



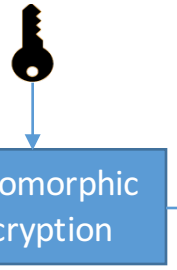
Encrypted Observation Fact	
Patient_Num	Concept_CD
0000001	griabfiyqeg
0000001	fgeiwbgohg
0000001	code_3
0000002	bcbuyrigodf
0000002	code_4
0000003	rrrrreuyubbu
0000003	code_3

COUNT(DISTINCT(patient_Num))
FROM Encrypted Observation Fact
WHERE Concept_CD = ?
AND/OR Concept_CD = ?
…

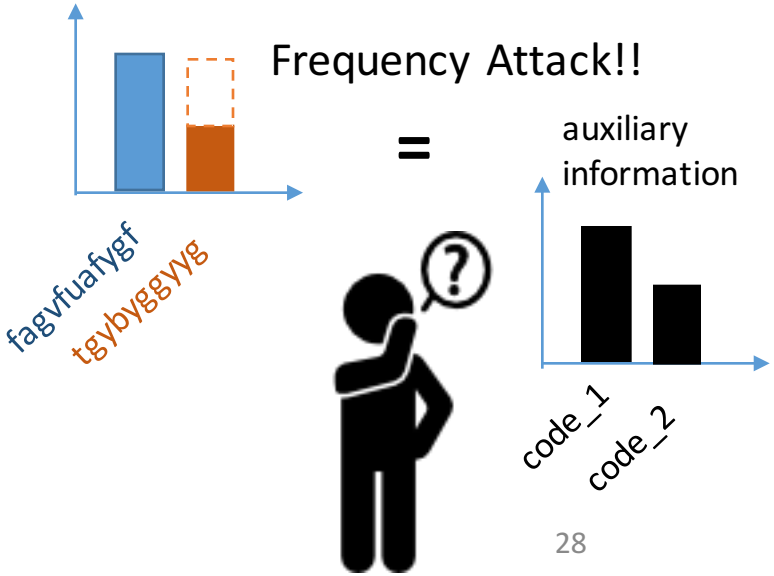
Encrypted Observation Fact	
Patient_Num	Concept_CD
0000001	fagvfuafygf
0000001	tgybyggyyg
0000001	code_3
0000002	fagvfuafygf
0000002	code_4
0000003	tgybyggyyg
0000003	code_3

- sensitive
- not sensitive

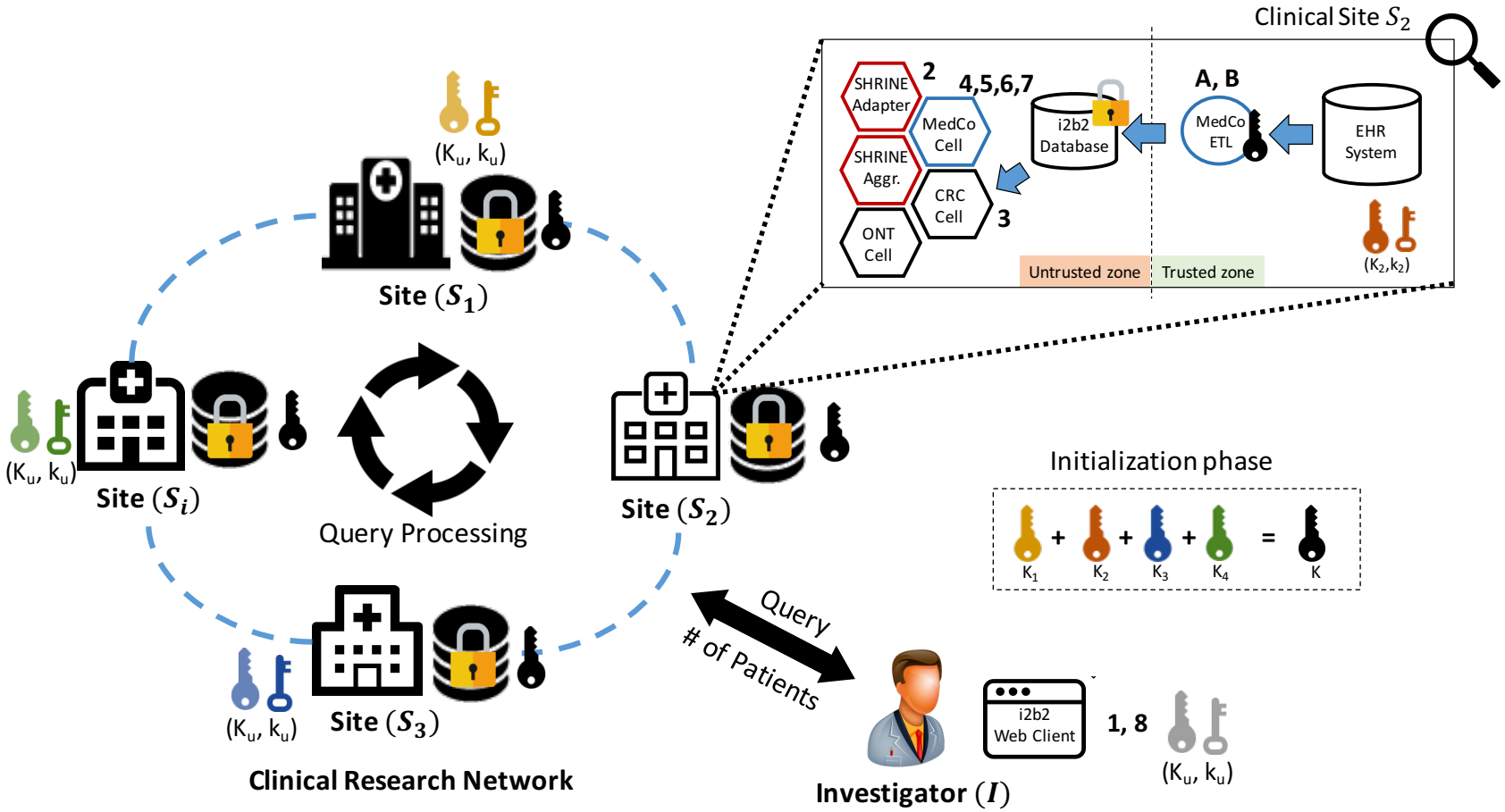
Patient Dimension	
Patient_Num	isReal
0000001	1
0000002	1
0000003	0



Patient Dimension	
Patient_Num	isReal
0000001	rteiqugfhb
0000002	vbwiygw=
0000003	fhbfg72=g



MedCo: Core Architecture & Protocol



1) (user) Query Generation:

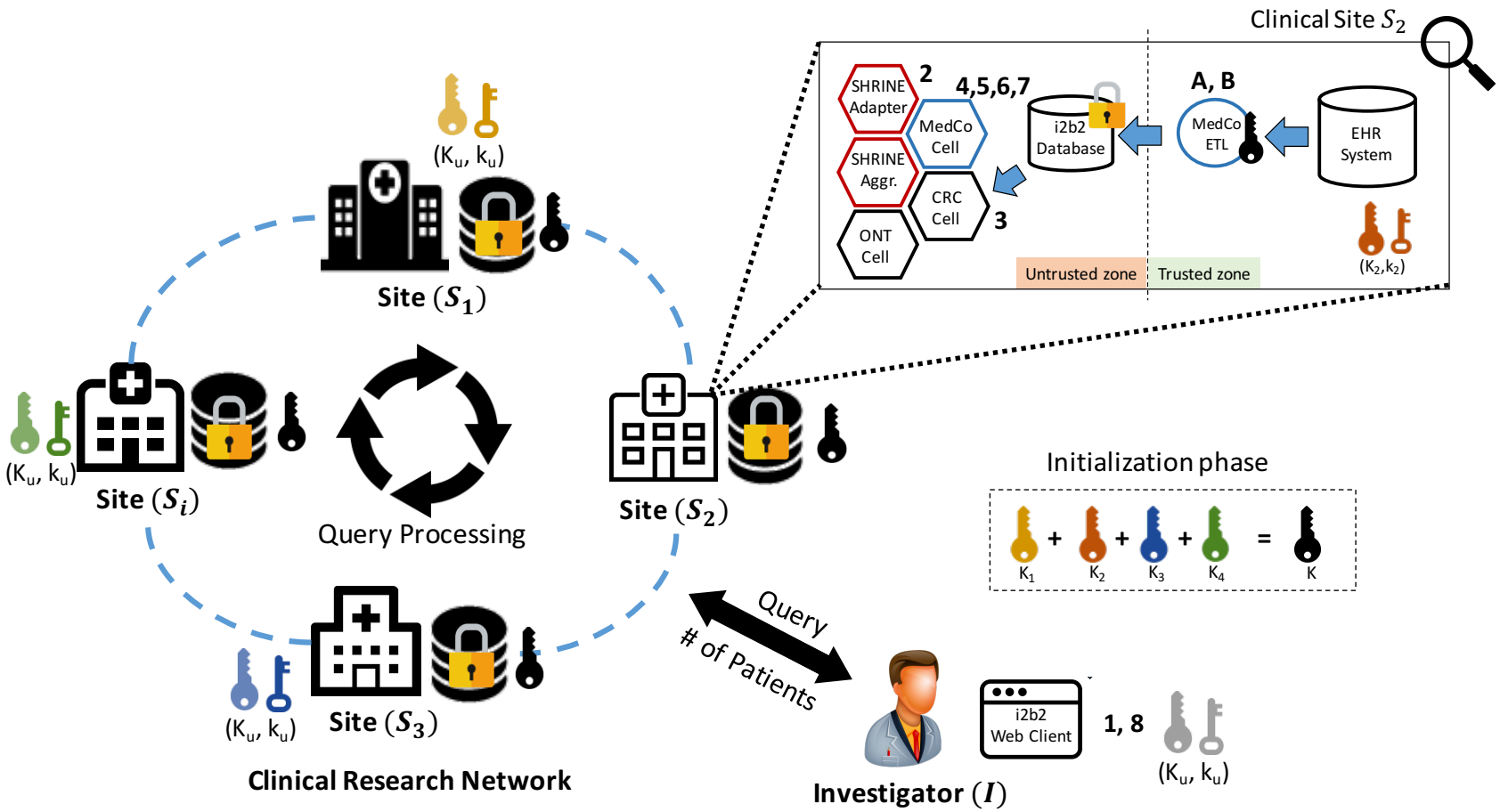
COUNT(DISTINCT(patient_Num))
FROM Encrypted Observation Fact
WHERE Concept_CD = code_2
AND Concept_CD = code_3
…

Homomorphic Encryption

COUNT(DISTINCT(patient_Num))
FROM Encrypted Observation Fact
WHERE Concept_CD = hf78e2ib78ffg
AND Concept_CD = code_3
…

i2b2 SHRINE MedCo

MedCo: Core Architecture & Protocol



2) (local) Query Analysis:

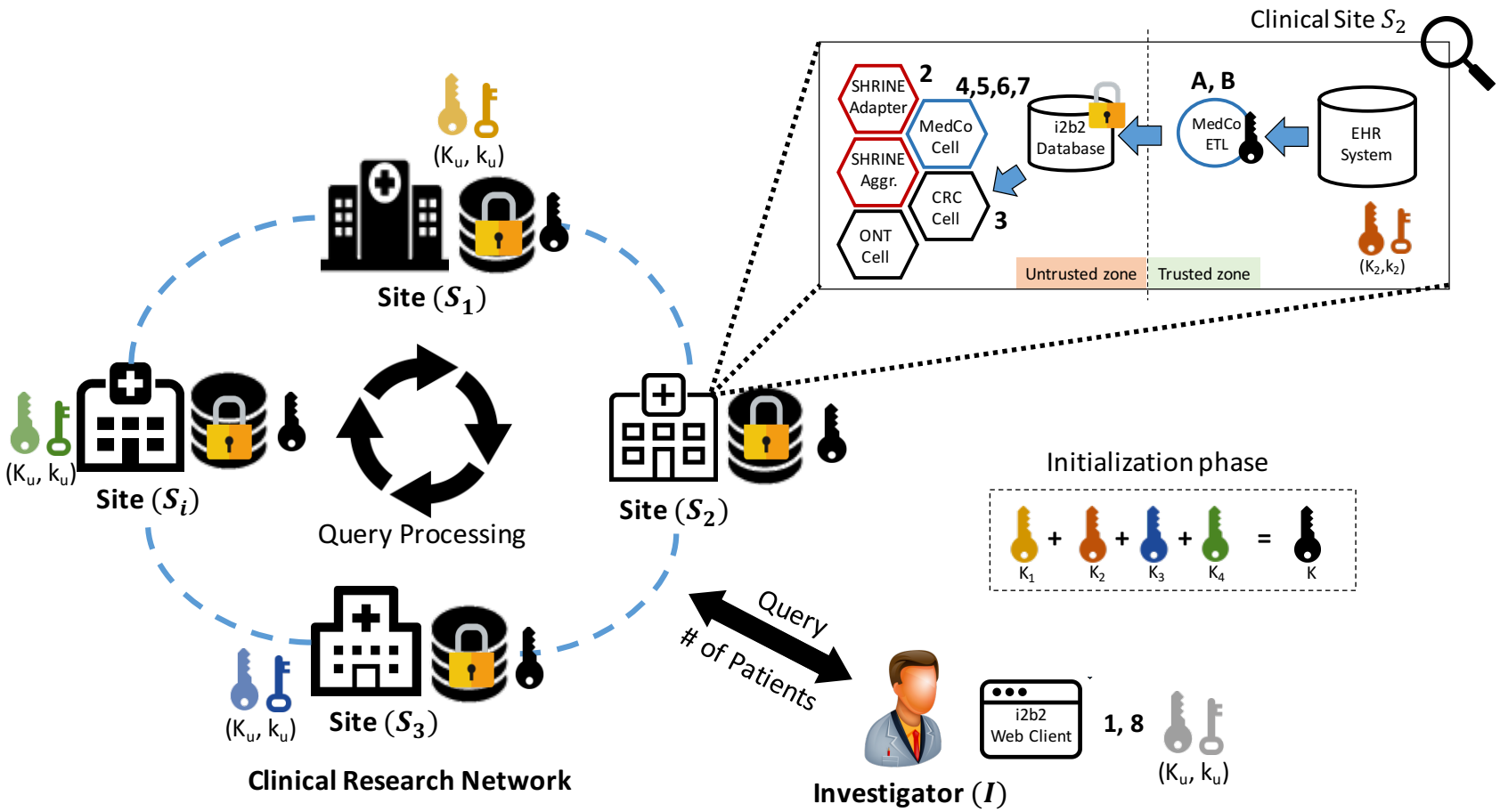
COUNT(DISTINCT(patient_Num))
FROM Encrypted Observation Fact
WHERE Concept_CD = hf78e2ib78ffg
AND Concept_CD = code_3
…

DDT

SELECT(DISTINCT(patient_Num))
FROM Encrypted Observation Fact
WHERE Concept_CD = tgybyggyyg
AND Concept_CD = code_3
…

i2b2 SHRINE MedCo

MedCo: Core Architecture & Protocol



3) (local) Query Processing:

SELECT(DISTINCT(patient_Num))
FROM Encrypted Observation Fact
WHERE Concept_CD = tgybyggyyg
AND Concept_CD = code_3
…

Encrypted Observation Fact

Patient_Num	Concept_CD
0000001	fagvfuafygf
0000001	tgybyggyyg
0000001	code_3
0000002	fagvfuafygf
0000002	code_4
0000003	tgybyggyyg
0000003	code_3

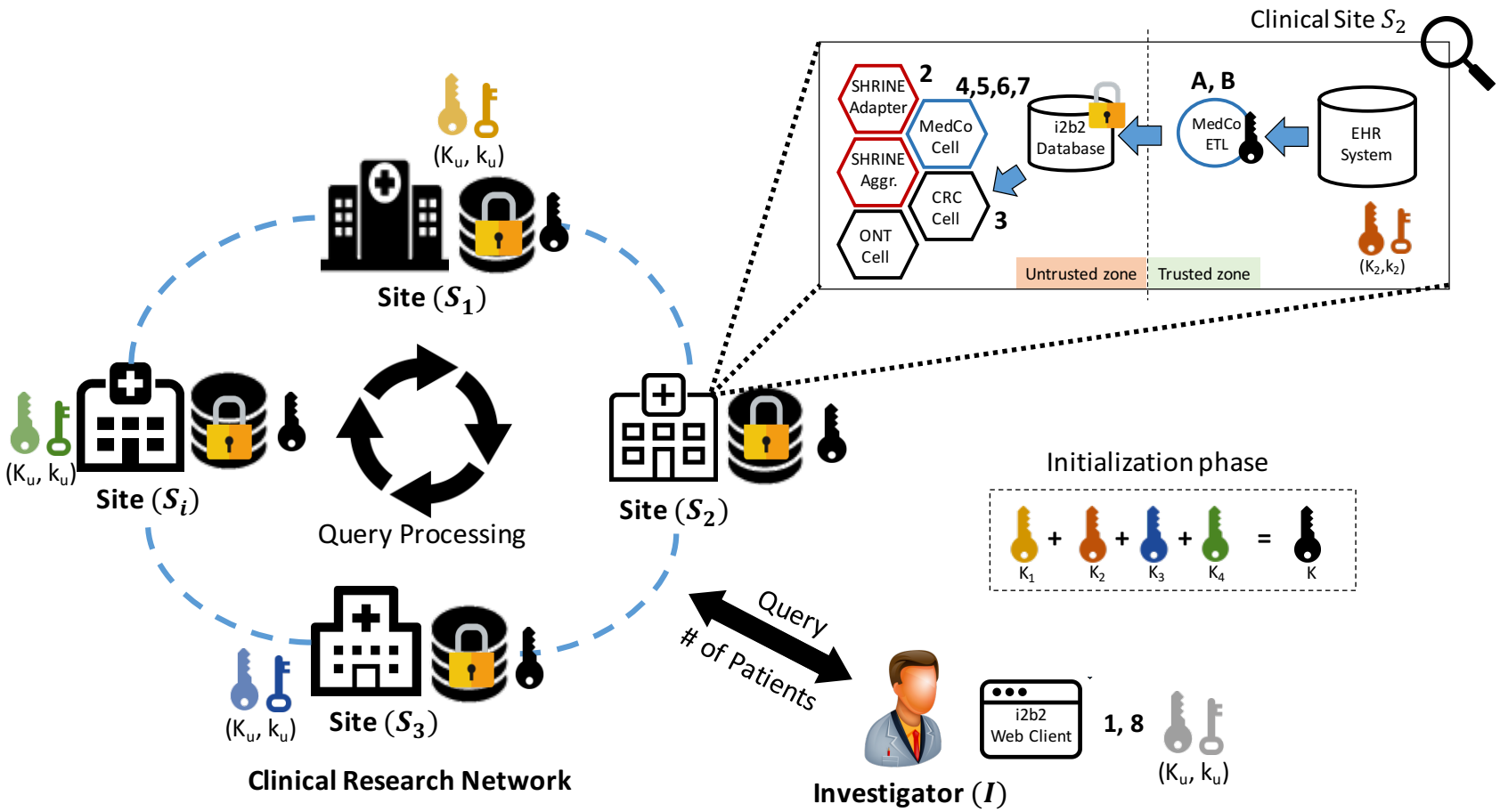
Patient Dimension

Patient_Num	isReal
0000001	rteiqugfhb
0000002	vbwiygw=
0000003	fhbfg72=g

0000001	rteiqugfhb
0000003	fhbfg72=g

i2b2 SHRINE MedCo

MedCo: Core Architecture & Protocol



4) (local) Result Aggregation:

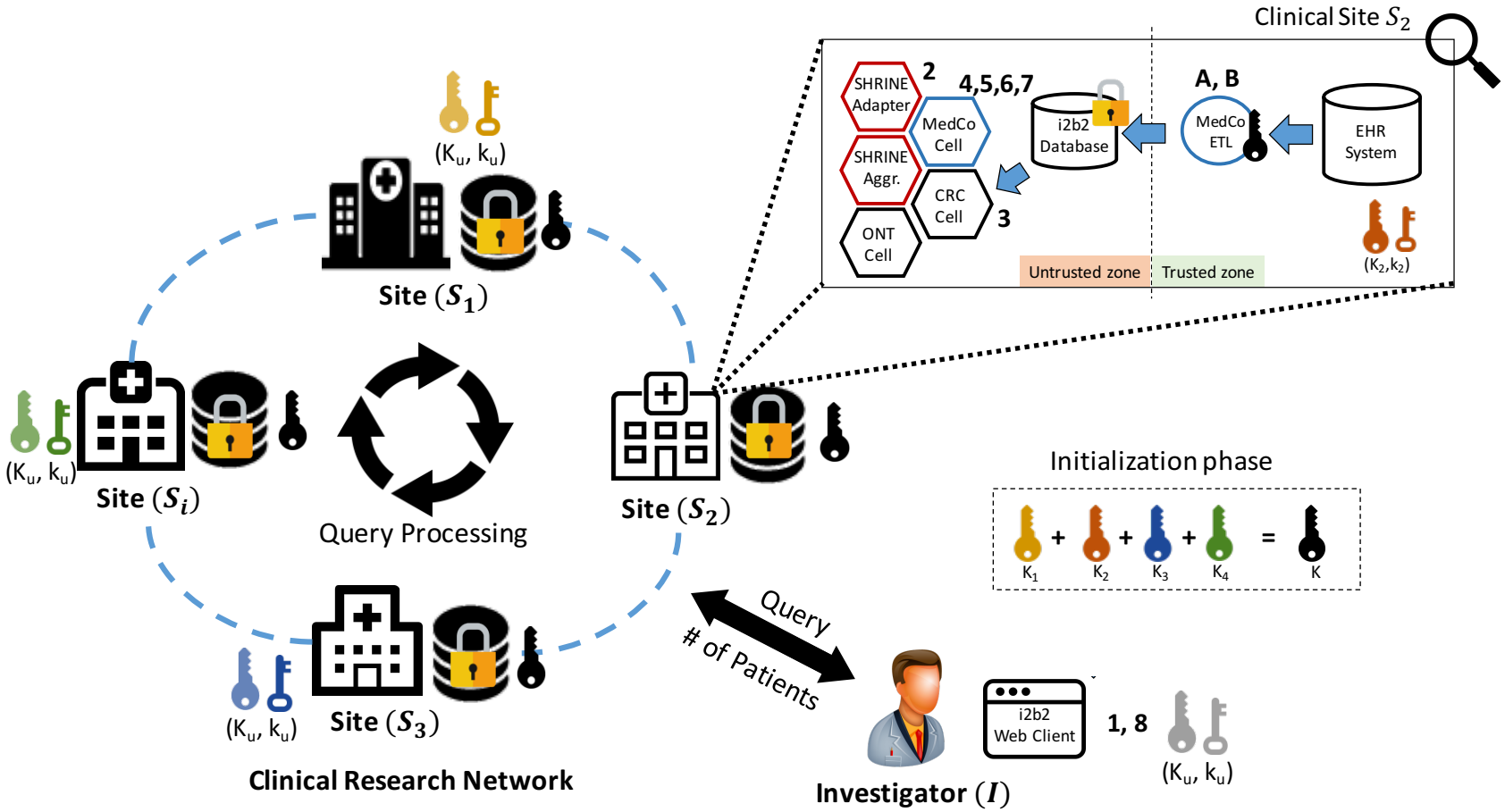
0000001	rteiqugfhb = E(1)
0000003	fhbfg72=g = E(0)

Homomorphic Addition

Result	yfgqebfiqy = E(1 + 0)

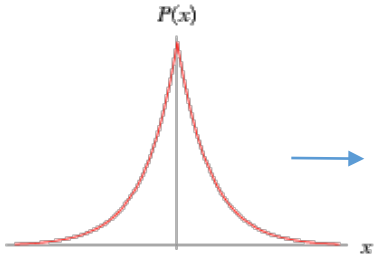
i2b2 SHRINE MedCo

MedCo: Core Architecture & Protocol



5) (local) Result Obfuscation:

Result	yfgqebfiqy = E(1 + 0)



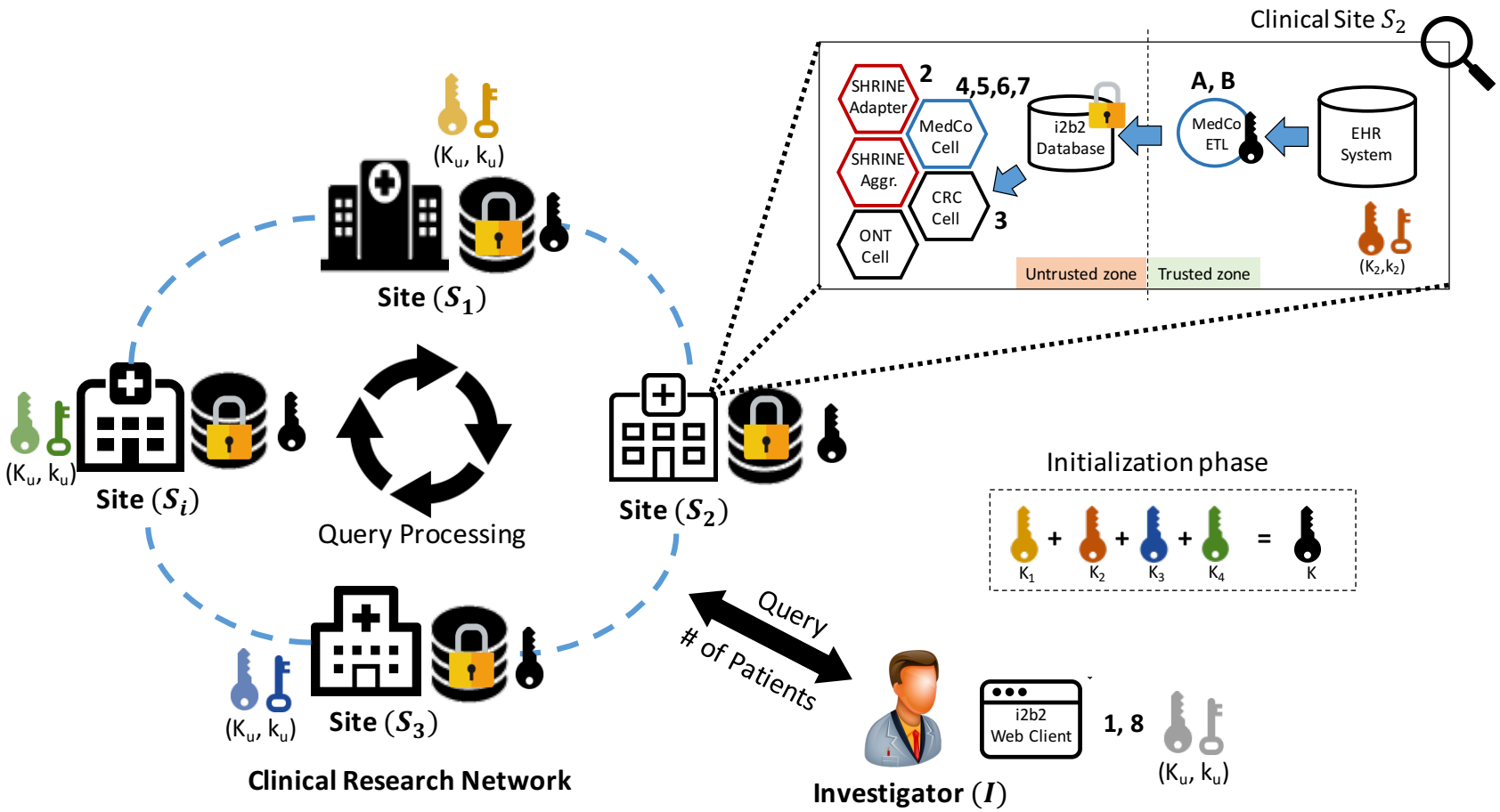
random noise x from Laplacian distribution

Homomorphic Addition

Obfuscated Result	fiqubgfwbi = E(1 + 0 + x)

i2b2 SHRINE MedCo

MedCo: Core Architecture & Protocol



6) (distributed) Results Shuffling:

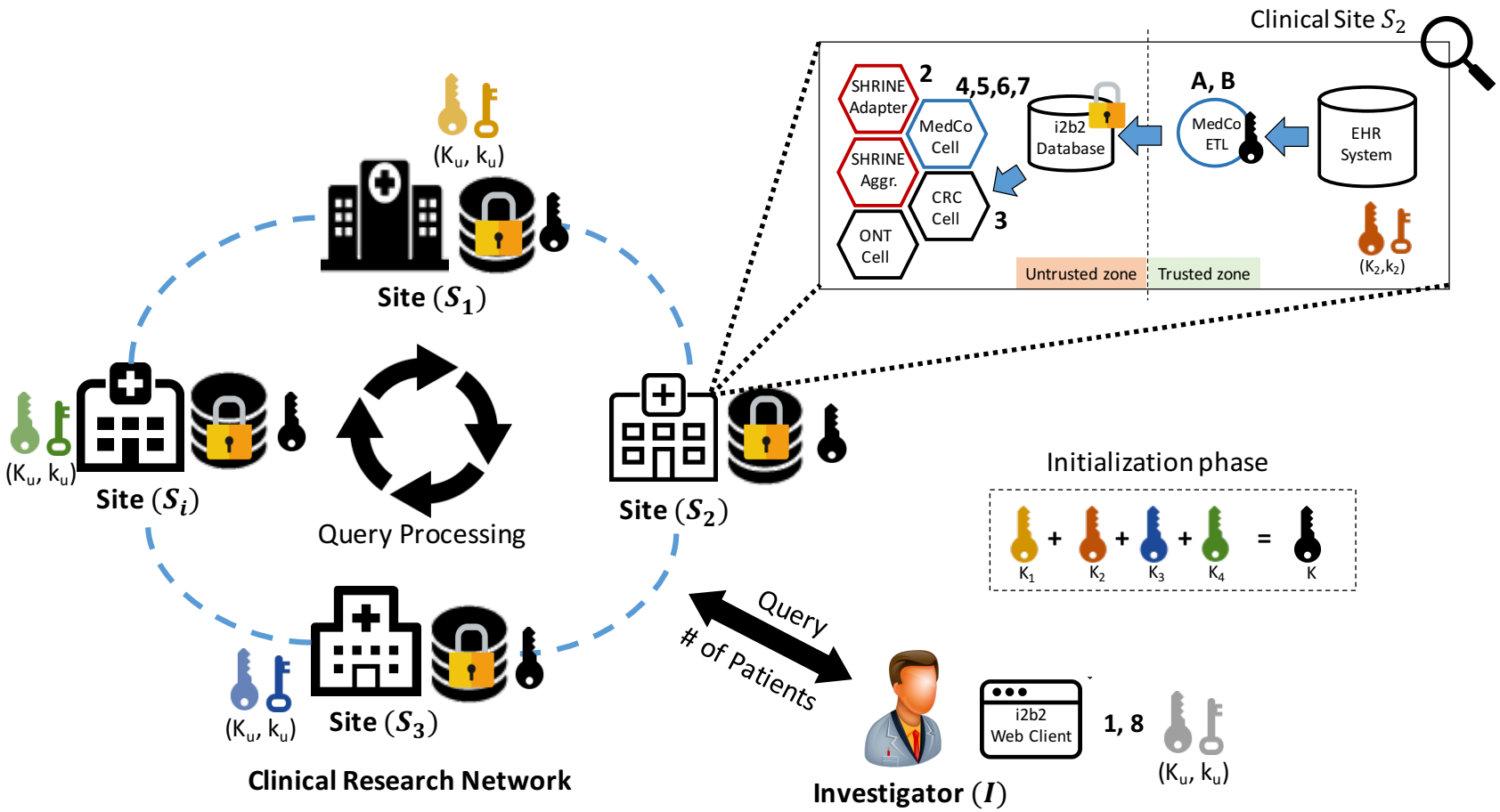
$Result_1$	yfgqebfiqy = E(1)
$Result_2$	f3rngi3rng= E(5)
$Result_3$	nfeingbrbd= E(10)
$Result_4$	vnnvnvugin= E(2)

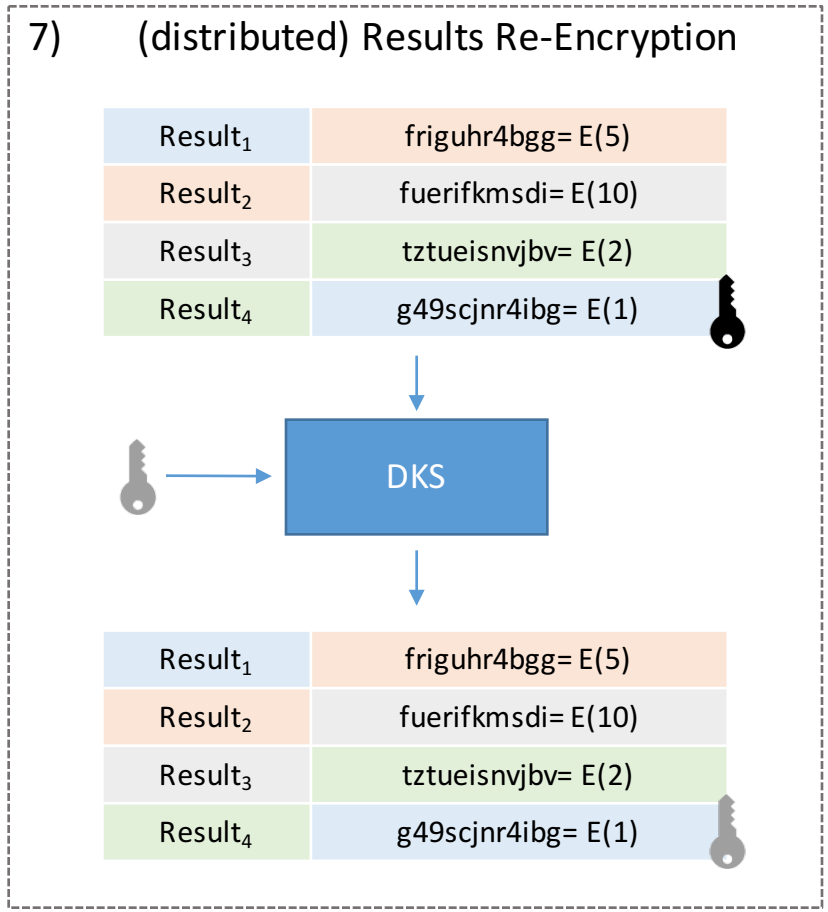
DVS

$Result_1$	friguhr4bgg= E(5)
$Result_2$	fuerifkmsdi= E(10)
$Result_3$	tztueisnvjbv= E(2)
$Result_4$	g49scjnr4ibg= E(1)

i2b2 SHRINE MedCo

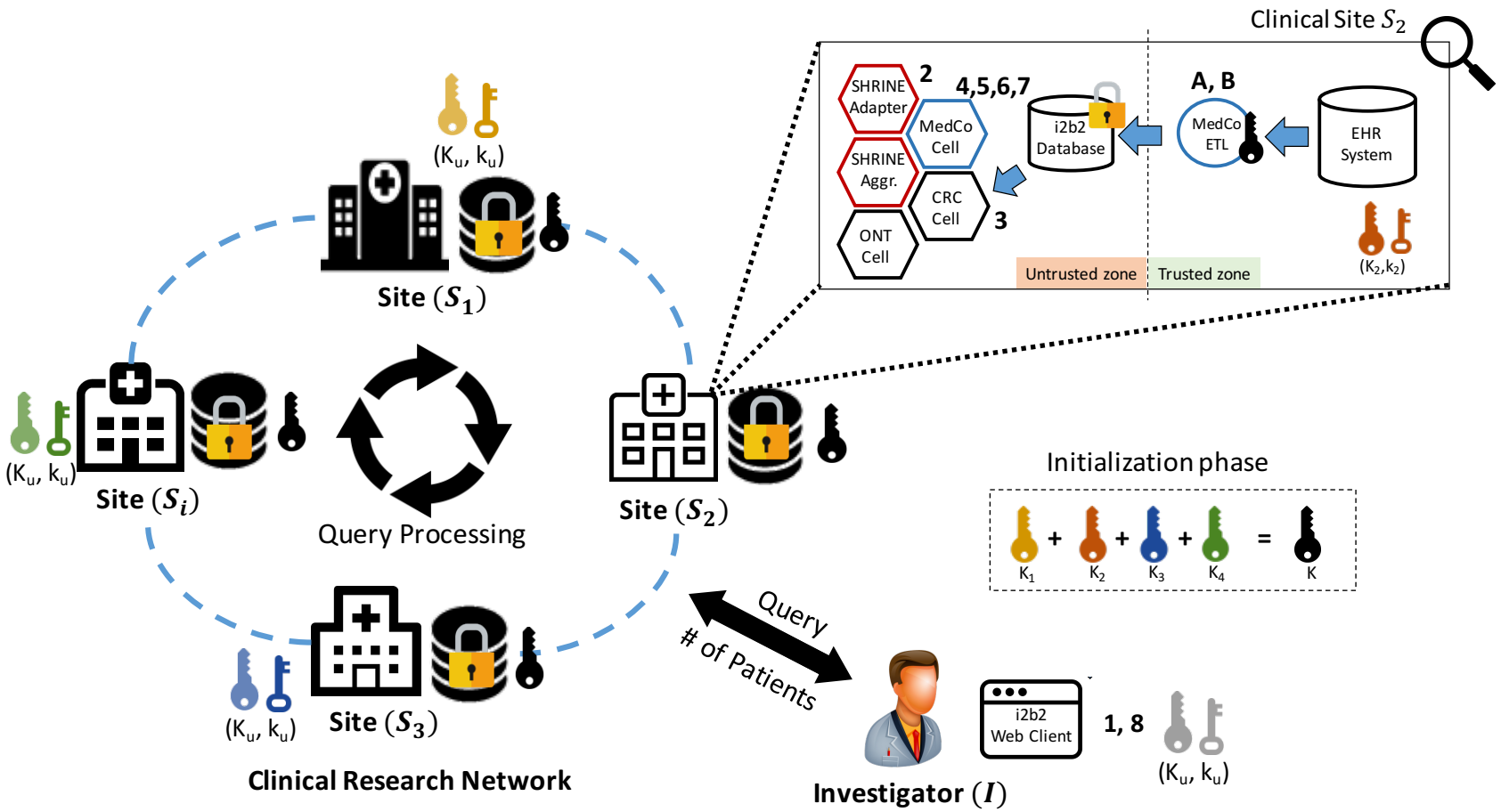
MedCo: Core Architecture & Protocol





i2b2 SHRINE MedCo

MedCo: Core Architecture & Protocol



8) (user) Results Decryption

$Result_1$	friguhr4bgg= E(5)
$Result_2$	fuerifkmsdi= E(10)
$Result_3$	tztueisnvjbv= E(2)
$Result_4$	g49scjnr4ibg= E(1)

Decryption

$Result_1$	5
$Result_2$	10
$Result_3$	2
$Result_4$	1

i2b2 SHRINE MedCo

Main Requirements Are Satisfied

Functionality:

COUNT(patients**)/SELECT(**patients**)**

FROM database

WHERE * AND/OR *

GROUP BY *

* represents any possible concepts in the otology

Security/Privacy:

- Protection of data confidentiality at rest, in transit and **during computation**
- no single point of failure
- only the investigator can obtain the query end-result
- (optional) unlinkability
- (optional) differential privacy

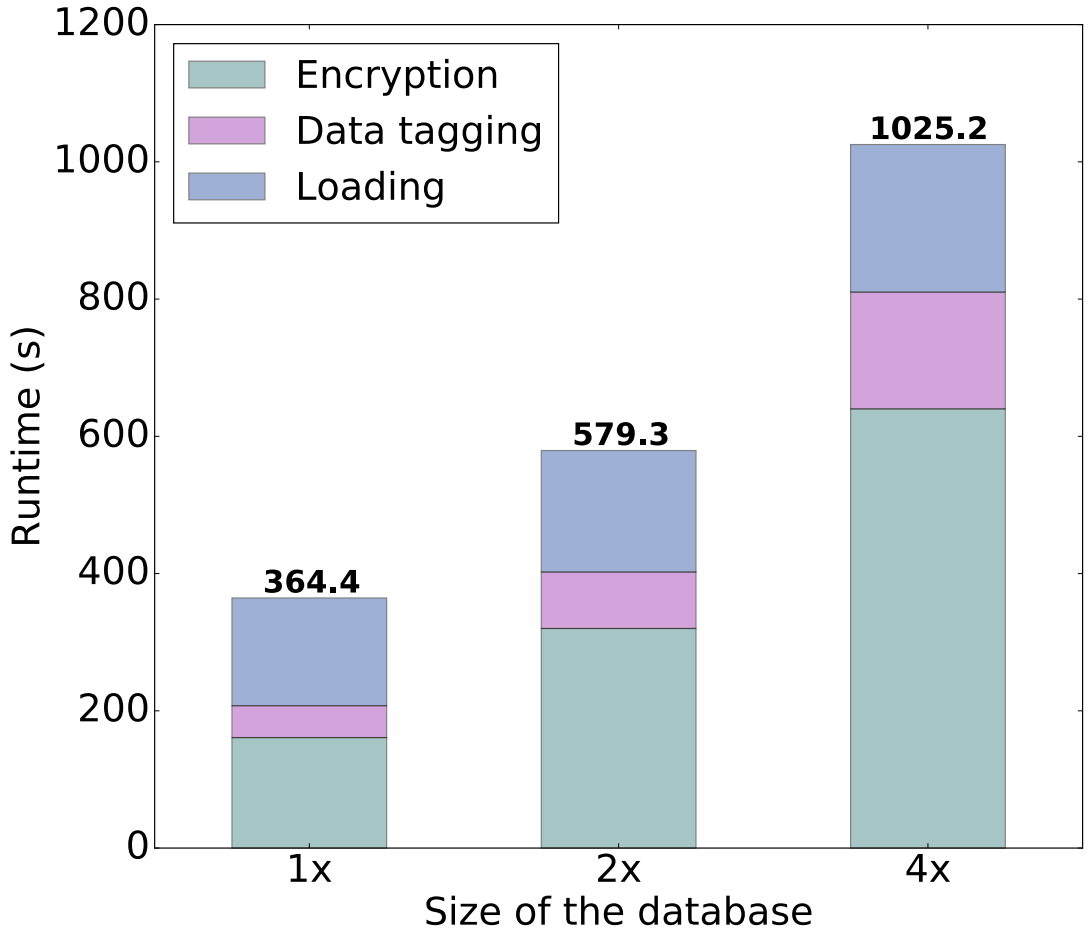
MedCo+: Security Extensions

- **Query Protection**
  - Data are stored homomorphically encrypted (not tagged)
  - Data are deterministically tagged (with DDT protocol) at each query with a fresh secret to ensure **query unlikability** and **query confidentiality**

```
COUNT(DISTINCT(patient_Num))
FROM Encrypted Observation Fact
WHERE Concept_CD = ?
AND/OR Concept_CD = ?
...
```

- **Malicious Clinical Sites**
  - Zero-knowledge proofs of computation can be generated at each step of the protocol to ensure **verifiability**
  - Malicious clinical sites can be identified and excluded from the system

Performance Results: ETL Time vs. Database Size



- **Encryption overhead**: Ontology Dimension is 4x larger
- **Dummy data overhead** 3.6x (highly dependent on the data)