遺伝情報と医療情報のアウトソーシング型プライバシ保護統計解析

佐久間 淳

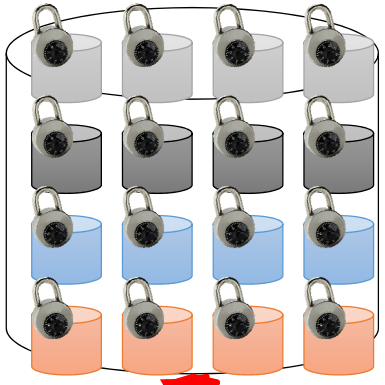
筑波大学/JST CREST/理研AIP

Joint work with
陸文傑, 川崎将平, 中澤貴明, 西出隆志 (筑波)
國廣昇, 津田宏治、竹内聖悟(東大)
竹内一郎、松井孝太、花田博幸、高田敏行(名工大)
山田芳司、安河内 彦輝(三重大)

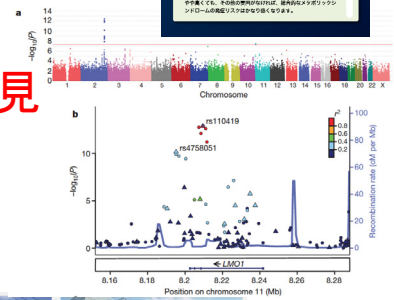
目標:孤立small dataをBIG DATAに

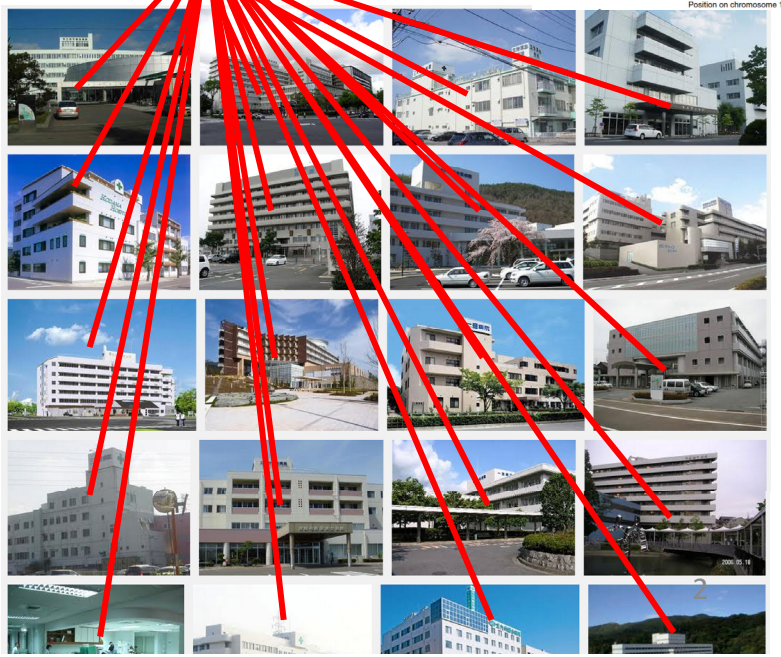
孤立したSmall data

新サービス



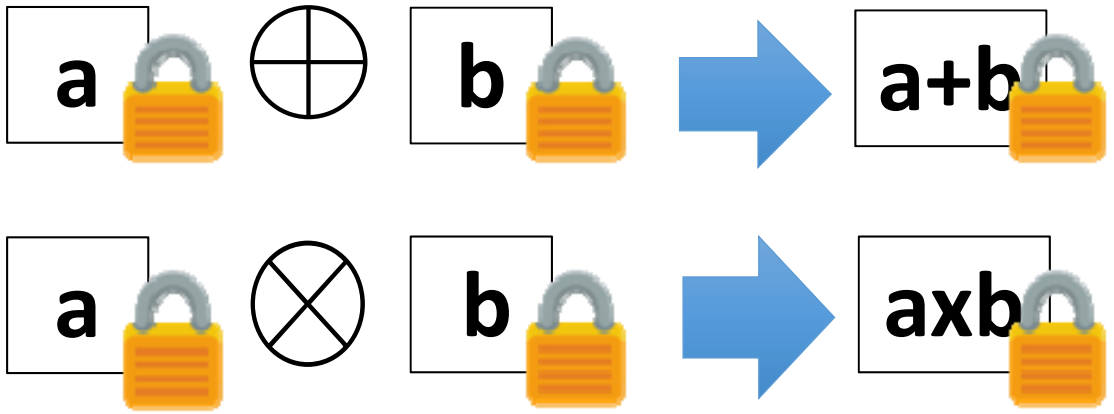
新発見





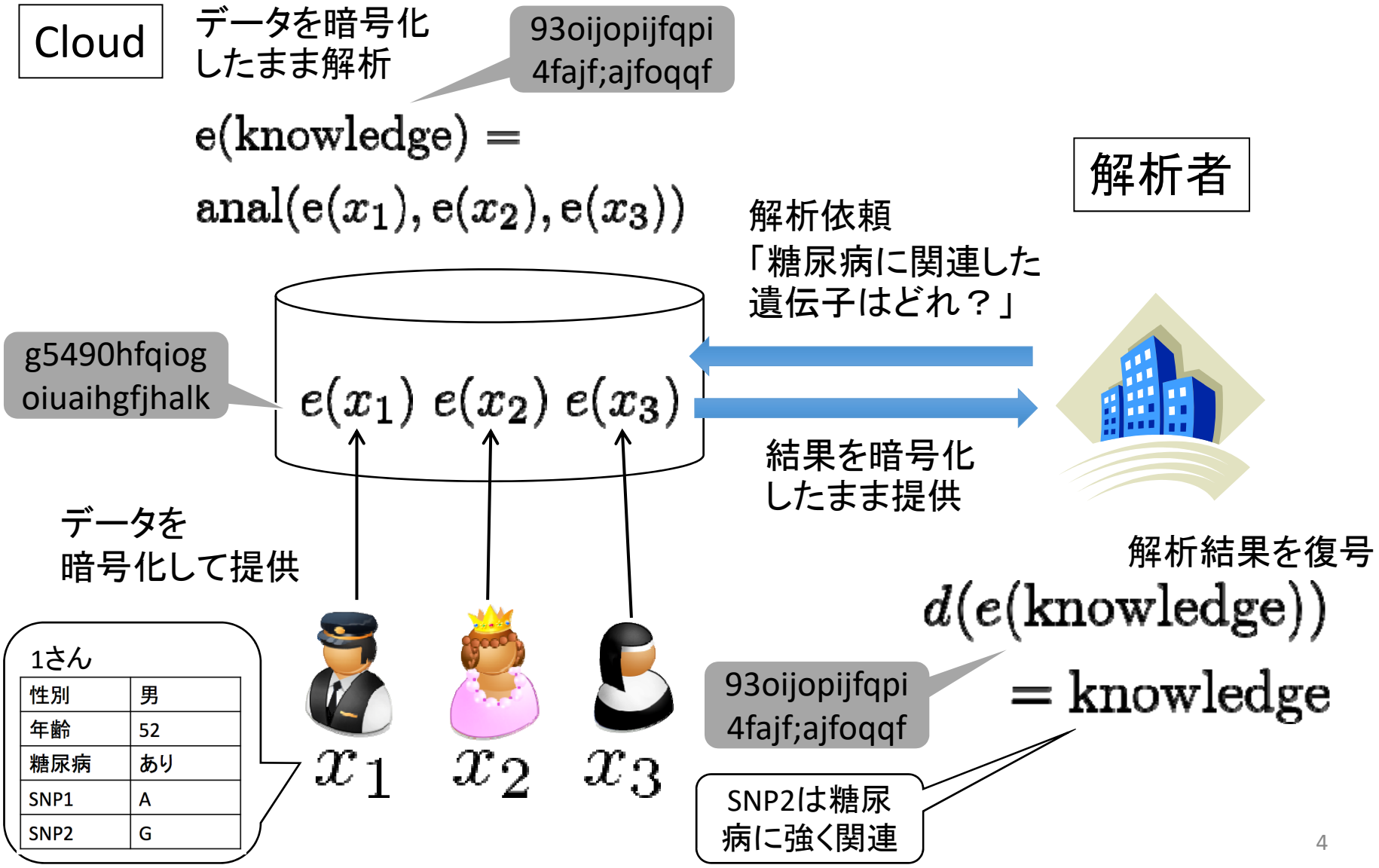
完全準同型暗号

- データを暗号化したまま加算・乗算
- 原理的には任意の演算を「データを見せずに」評価可能

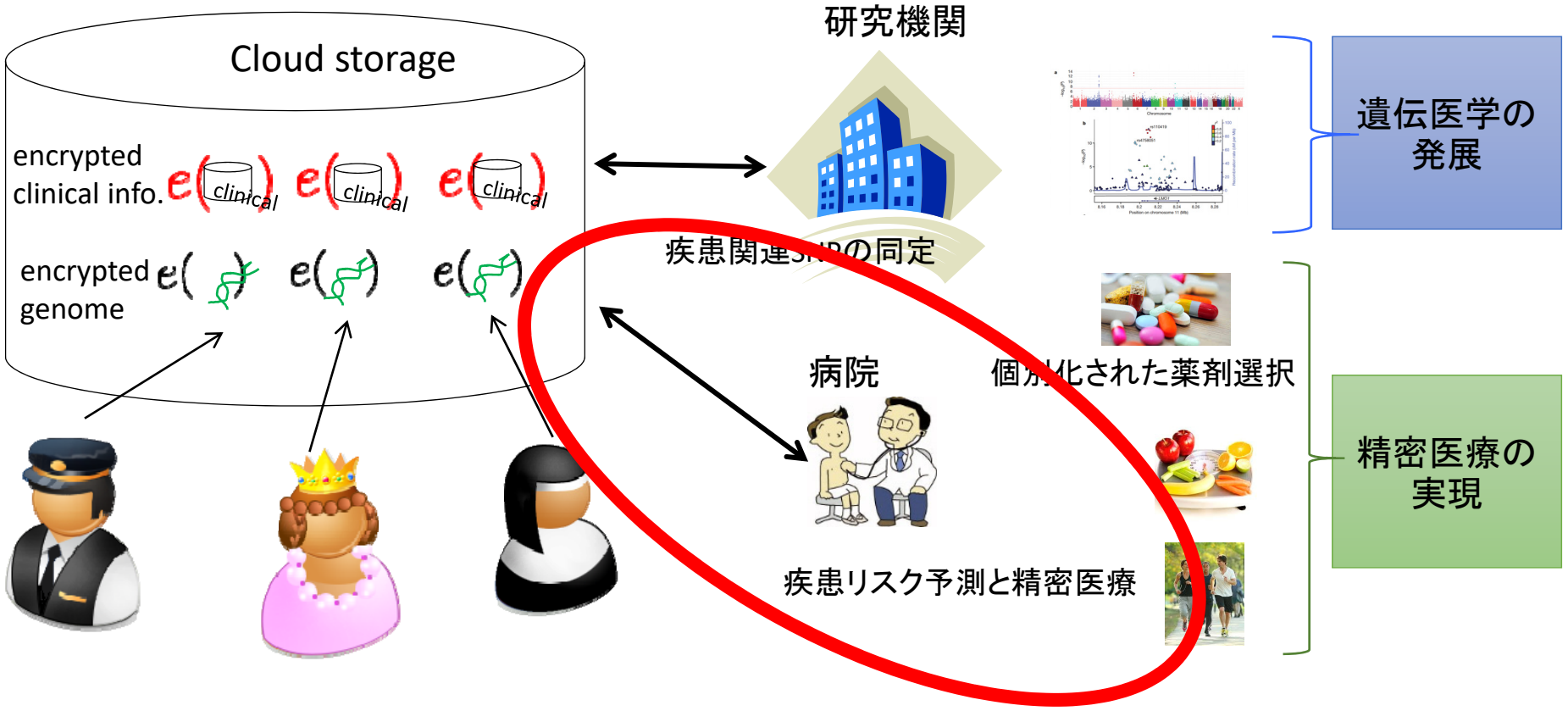


Gentry Craig, “A fully homomorphic encryption scheme”, Doctoral dissertation, Stanford University, 2009

基盤技術: 秘密計算・準同型暗号



CRESTにおける研究開発:暗号化個人ゲノム利用基盤



- 個人ゲノムは解析は高コストだが生涯変化しない
- 準同型暗号化個人ゲノムをクラウド上に蓄積
- 様々な機関から様々な用途で安全に利用できるように個人ゲノム解析のサービス提供

従来の生活習慣病個別化予防システム

血液検査や生活習慣情報を利用しない
→生活習慣病のリスク予測として不十分

多くは文献値を利用
→日本人に適合しない

静的なリスク評価
→生活習慣改善の動機付けとして不十分

検査希望者

遺伝因子の解析

SNPs解析

リスク予測

	予測(%)
心筋梗塞	10
脳梗塞	5
脳出血	3

遺伝因子によるリスク予測

	遺伝子型	相対危険度
SNP 1	*TT*	1.3
SNP 2	*CG*	1.2
SNP 3	*AT*	0.8
SNP 4	*GG*	1.4
SNP 5	*CC*	0.7
SNP 6	*GT*	1.2
SNP 7	*TT*	1.4
SNP 8	*GG*	1.3
SNP 9	*CG*	0.9
・・・		
・・・		
SNP 100	*CT*	1.2

遺伝子検査機関

日本人のための生活習慣病個別化予防システム

検査希望者

臨床・遺伝情報の名寄せ

家族歴

問診
身体検査
血液検査
尿検査
生理検査
X線検査
MRI/MRA

SNPs解析

病院外への
臨床データ提供

検査機関外への
遺伝情報提供

総合評価によるリスク予測

	予測確率(%)
心筋梗塞	10
脳梗塞	5
脳出血	3

リスク予測

肥満	1.7
高血圧	2.3
脂質異常症	2.1
糖尿病	2.5
慢性腎臓病	1.4

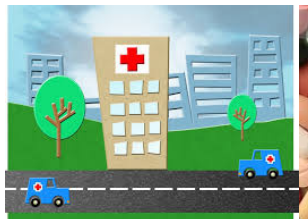
病院

個人にとって最適な
疾患予防法の実践
・食事療法・禁煙
・高血圧・糖尿病・脂質異常症
などの治療

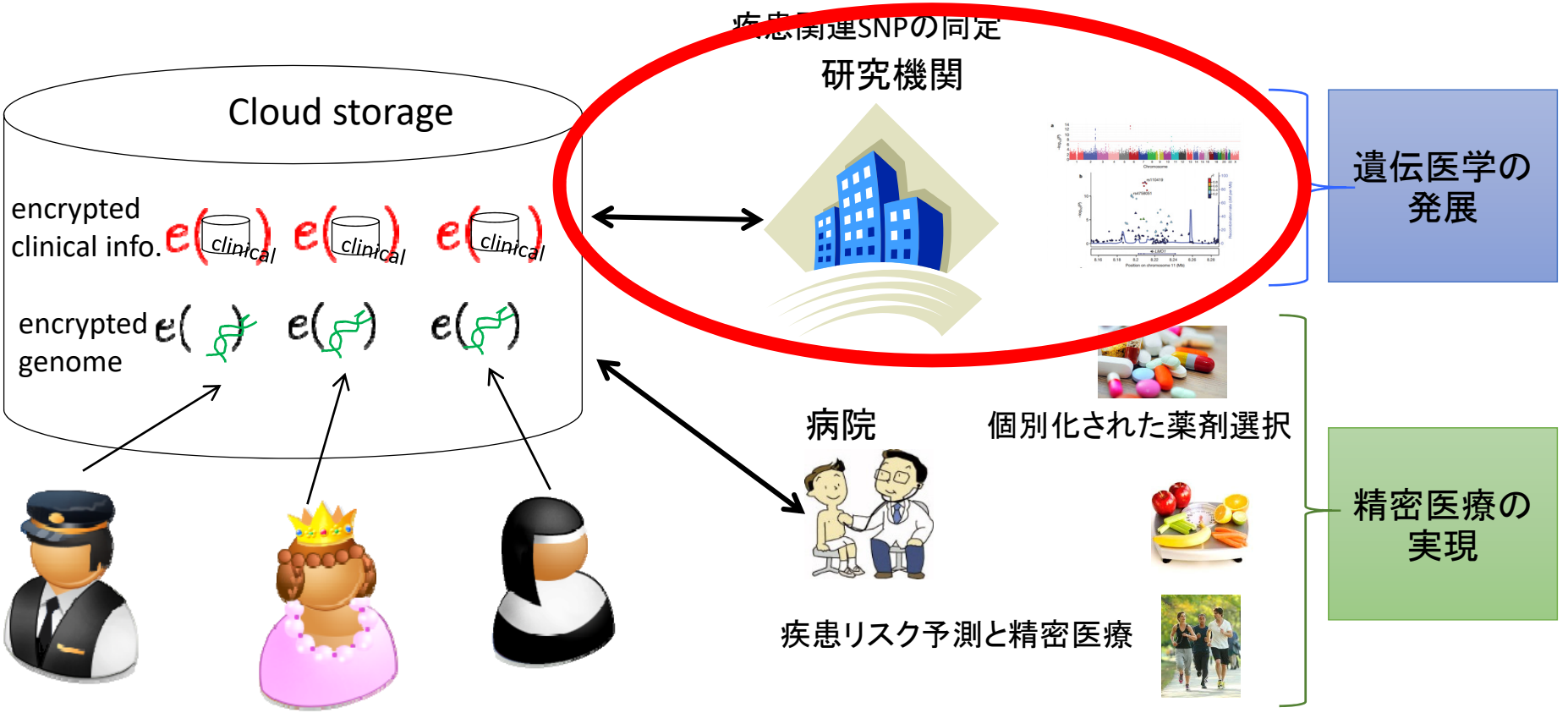
遺伝

SNP 2		
SNP 3	*AT*	0.8
SNP 4	*GG*	1.4
SNP 5	*CC*	0.7
SNP 6	*GT*	1.2
SNP 7	*TT*	1.4
SNP 8	*GG*	1.3
SNP 9	*CG*	0.9
・・・		
・・・		
SNP 100	*CT*	1.2

遺伝子
検査機関



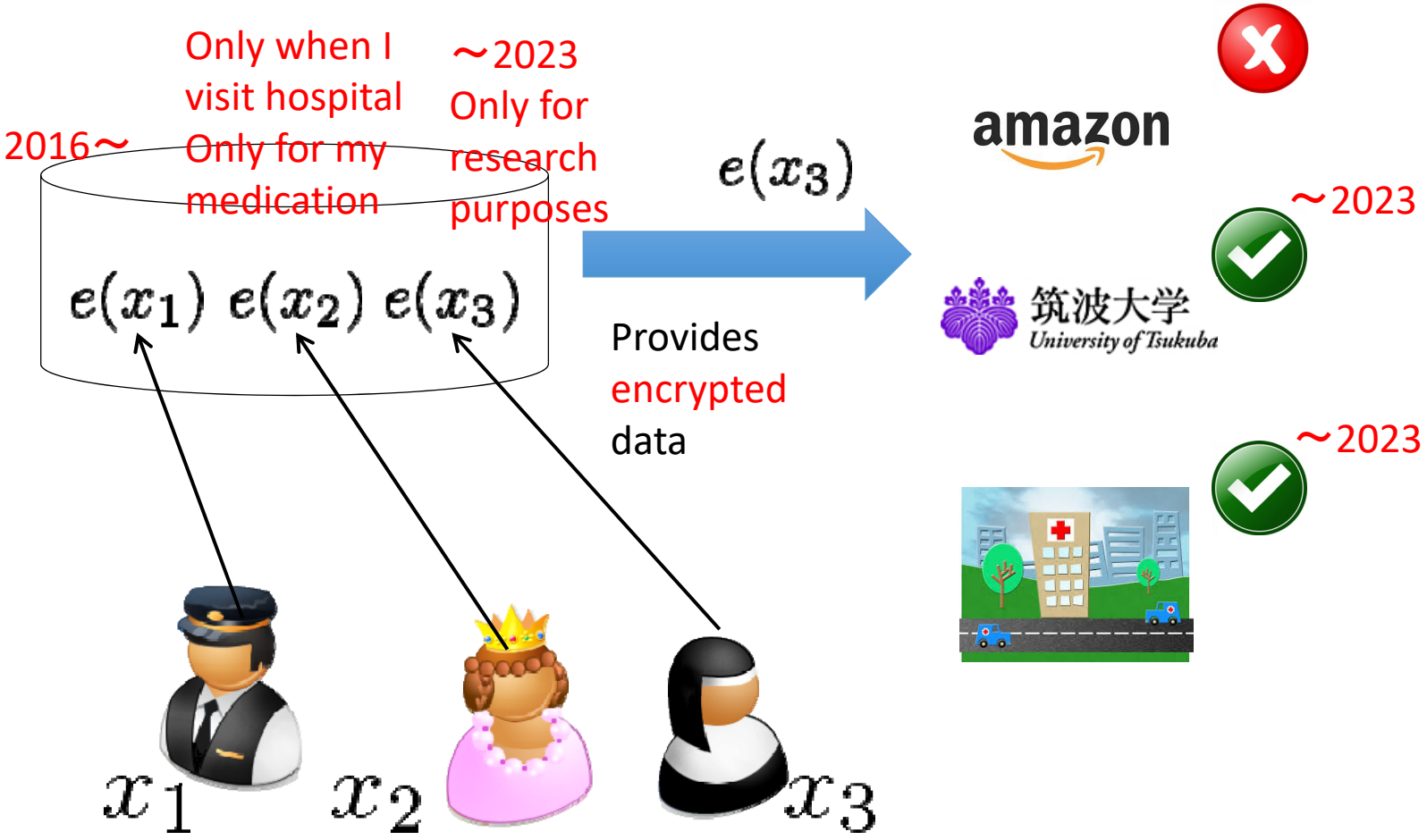
暗号化個人ゲノム利用基盤



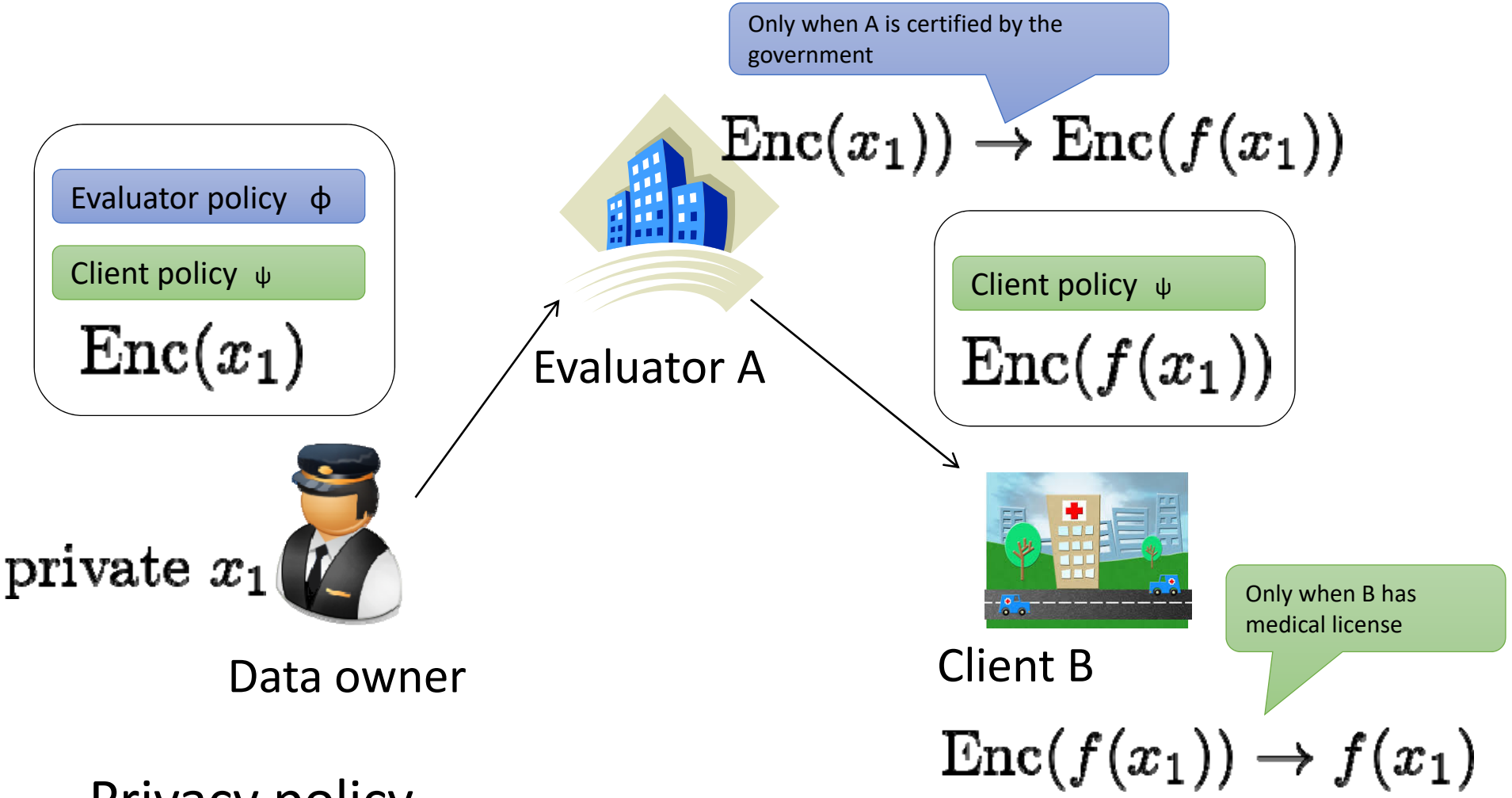
- 個人ゲノムは解析は高コストだが生涯変化しない
- 準同型暗号化個人ゲノムをクラウド上に蓄積
- 様々な機関から様々な用途で安全に利用できるように個人ゲノム解析のサービス提供

Secure Function Evaluation with Privacy Policy Enforcement

Sakuma, Lu, Nishide, Kunihiro, under review.



Secure function evaluation with privacy policy enforcement



Privacy policy

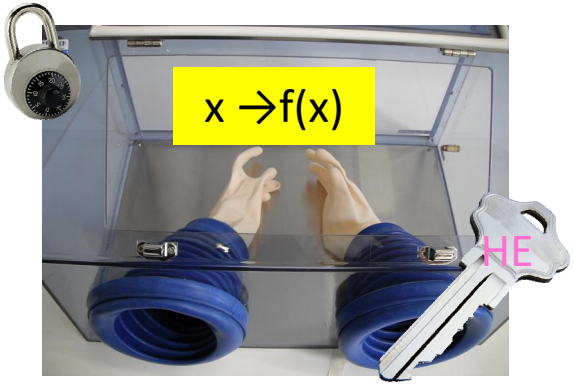
- Evaluator policy defines entities who are allowed to f(x)
- Client policy defines entities who are allowed to obtain f(x)

Intuition of policy enforcement

Symmetric key enc. (SKE)

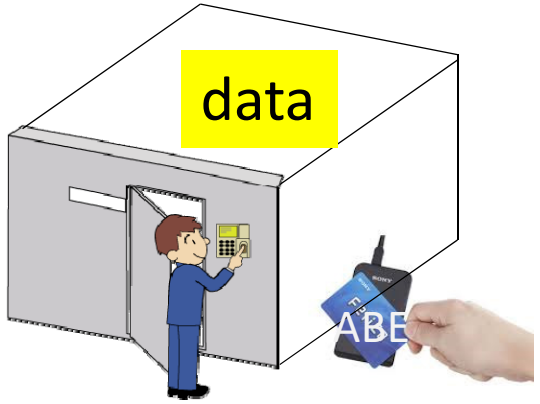
Those who have the key can see the data

Homomorphic enc. (HE)



Those who do have the key cannot see the data, but can analyze the data

Attribute-based enc. (ABE)



Those who have a specified attribute can see the data

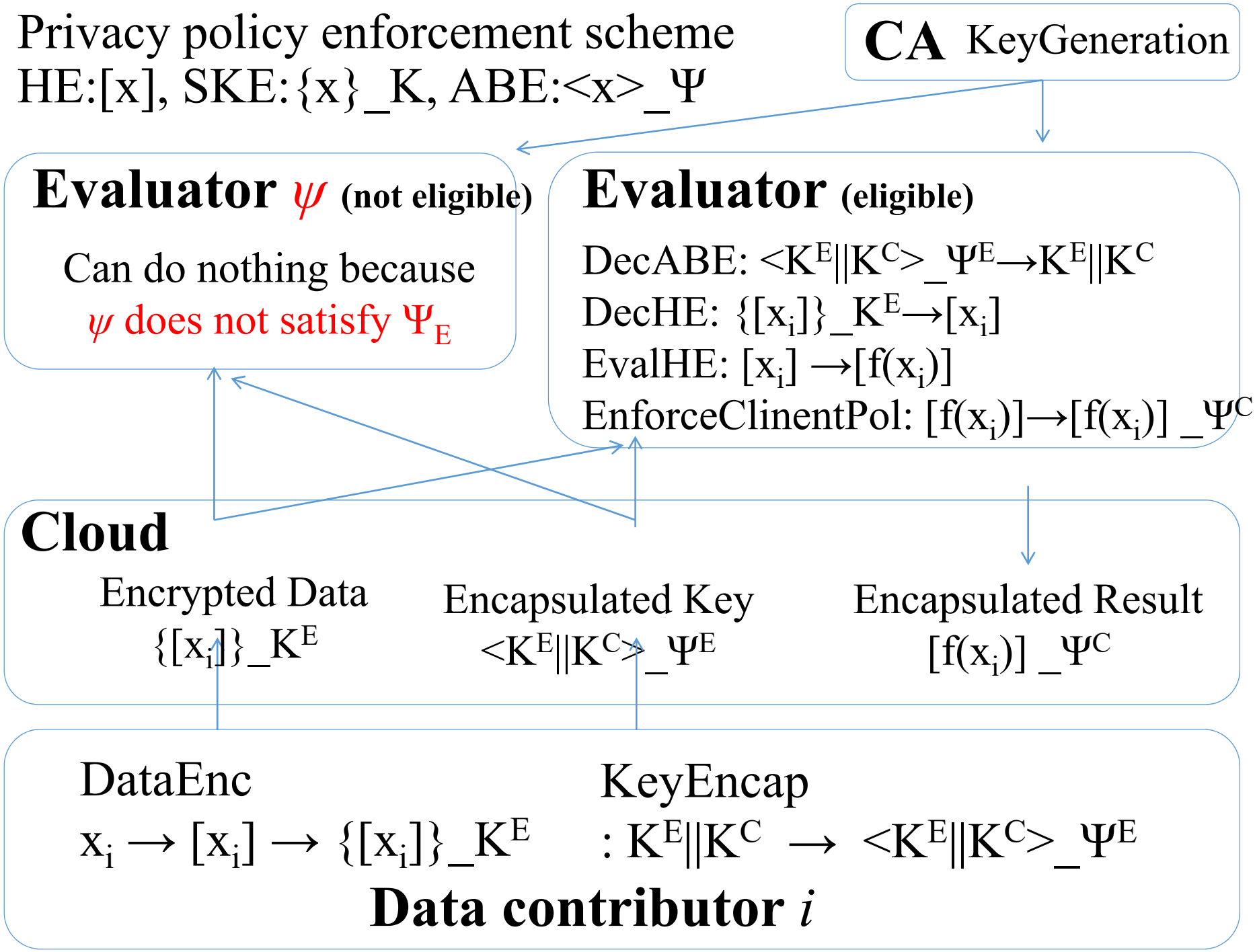
Privacy policy enforcement

Data encryption

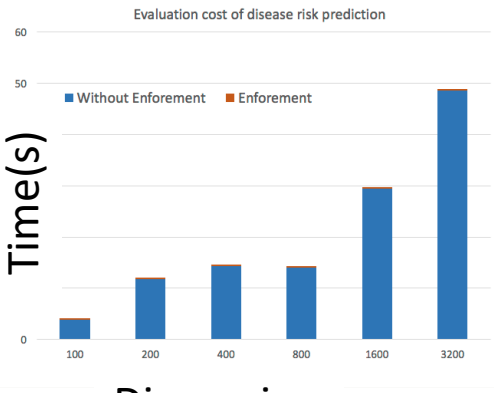
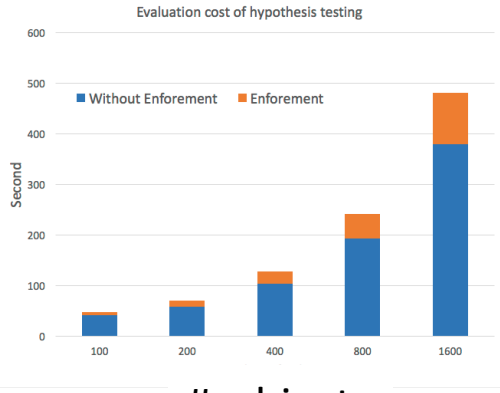
Those who have the key can analyze the data

Encapsulated key

Those who have a specified attribute can obtain the key

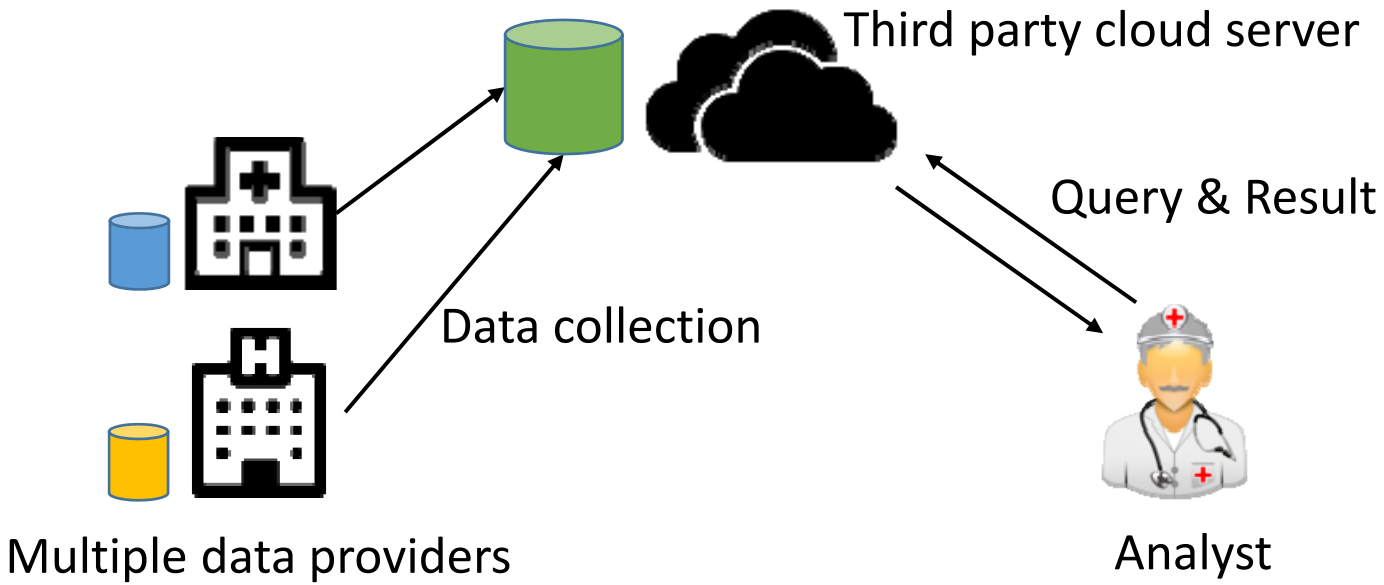

Privacy policy enforcement scheme
HE:[x], SKE:{x}_K, ABE:<x>_Ψ
CA KeyGeneration
Evaluator ψ (not eligible)
Can do nothing because
ψ does not satisfy Ψ_E
Evaluator (eligible)
DecABE: <K^E||K^C>_Ψ^E→K^E||K^C
DecHE: {[x_i]}_K^E→[x_i]
EvalHE: [x_i] →[f(x_i)]
EnforceClinentPol: [f(x_i)]→[f(x_i)] _Ψ^C
Cloud
Encrypted Data
{[x_i]}_K^E
Encapsulated Key
<K^E||K^C>_Ψ^E
Encapsulated Result
[f(x_i)] _Ψ^C
DataEnc
x_i → [x_i] → {[x_i]}_K^E
KeyEncap
: K^E||K^C → <K^E||K^C>_Ψ^E
Data contributor i

Experimental results

	Disease risk prediction	SVM classification	Chi-square test
w/policy enforcement  wo/policy enforcement			
task	$\mathbf{risk} = w^T x$	$f(\boldsymbol{x}) = \sum_{\boldsymbol{x}_i^{SV} \in SV} \alpha_i k(\boldsymbol{x}, \boldsymbol{x}_i^{SV})$	$\chi^2 = \sum_{i=1}^{4} \frac{(o_i - e_i)^2}{e_i}$
HE	Paillier	Ring-LWE	Ring-LWE
overhead	7%<	1.4%<	20%<

Privacy policy enforcement is not that costly!

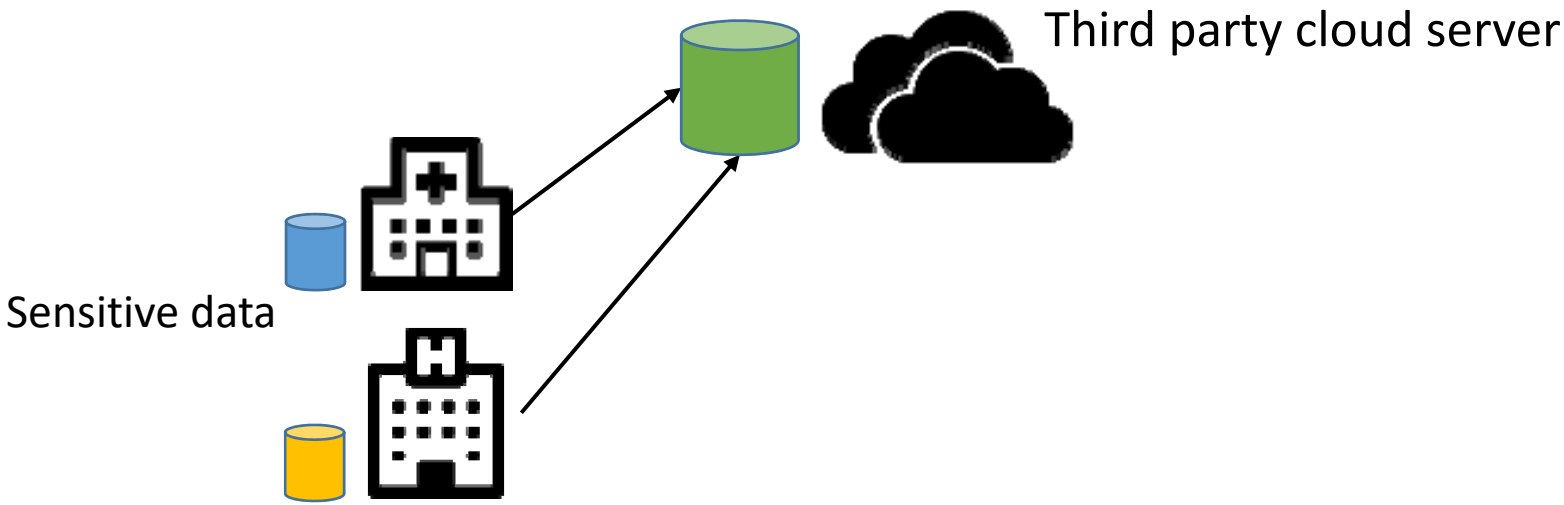
Statistical Analysis on the Cloud



Cloud computing is useful for statistical analysis

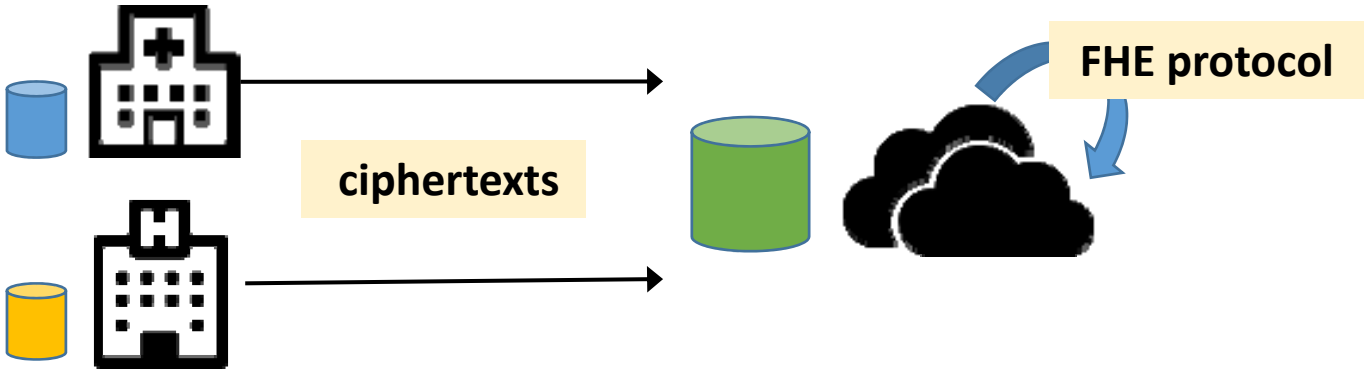
- Gather distributed data, and reduce hardware cost.
- Minimal interactions between data providers and the cloud.
- The cloud does most of the work for the analyst.

Cloud Computing with Sensitive Data



- Using outside cloud servers raises privacy concerns.
  - E.g, medical records, federal data.
- We want to calculate statistics on the cloud while keeping the data secret.

FHE on the Cloud Environment



- Less development cost
  - Suitable to outsourcing
- But might be inefficient in practice
  - Encrypt bits one by one.
  - 1~10 ms per evaluation.
  - 1~10 megabytes per ciphertext.

Gentry et al. *Homomorphic Evaluation of the AES Circuit*. 2012.

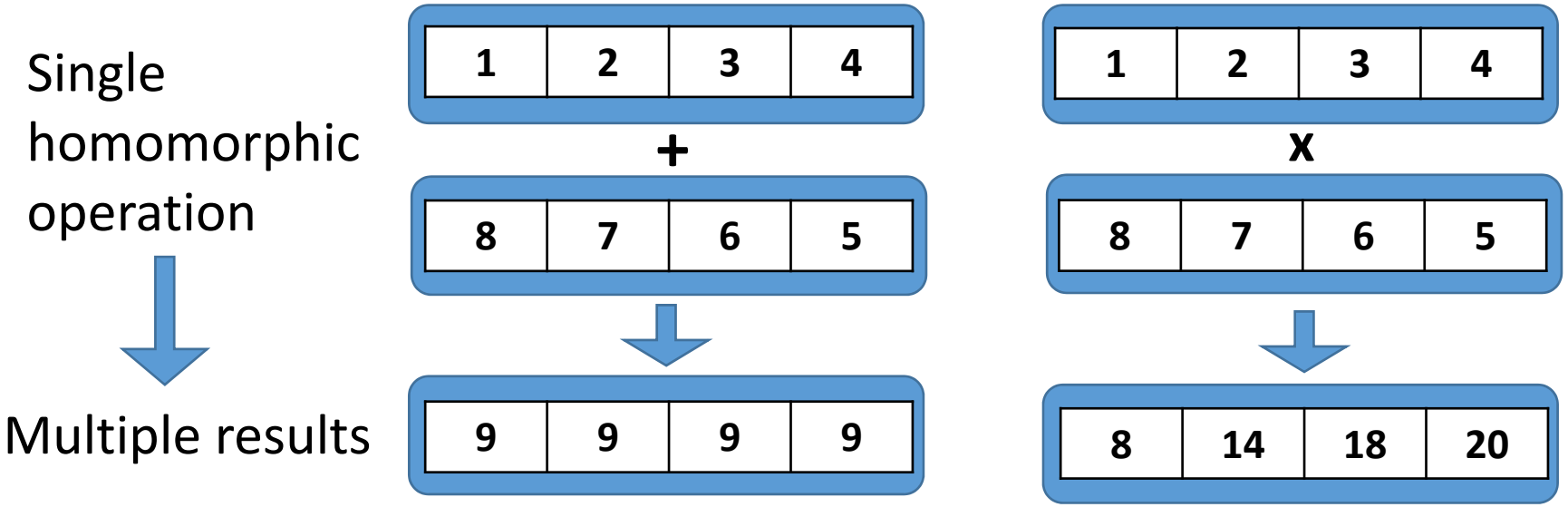
Using Fully Homomorphic Encryption for Statistical Analysis of Categorical, Ordinal and Numerical Data

NDSS'17 Lu, Kawasaki, Sakuma

- Two new efficient FHE-based primitives:
  - *Matrix Operations*
  - *Batch Greater-than*
- Secure statistical protocols:
  - histogram (count),
  - order of counts,
  - contingency table (with cell-suppression),
  - percentile,
  - principal component analysis (PCA),
  - linear regression.
- Source codes: https://github.com/fionser/CODA

Preliminaries: Packing (Batching)

- Enable to encrypt and process **vectors** at no extra cost.

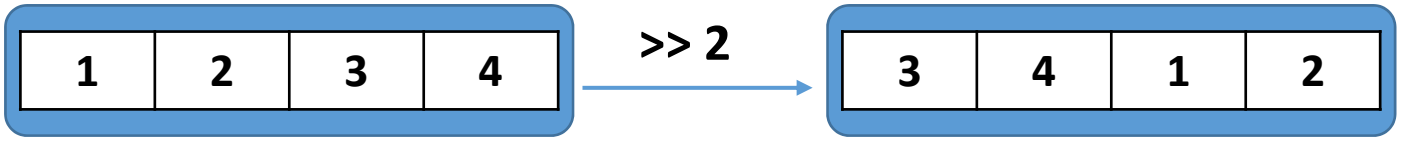


  - Fewer ciphertexts
  - Faster computation

N.P. Smart et al. *Fully homomorphic SIMD operations*. 2011.

Preliminaries: Slot Manipulation

Rotate slots of the encrypted vector.



Replicate a specific slot.

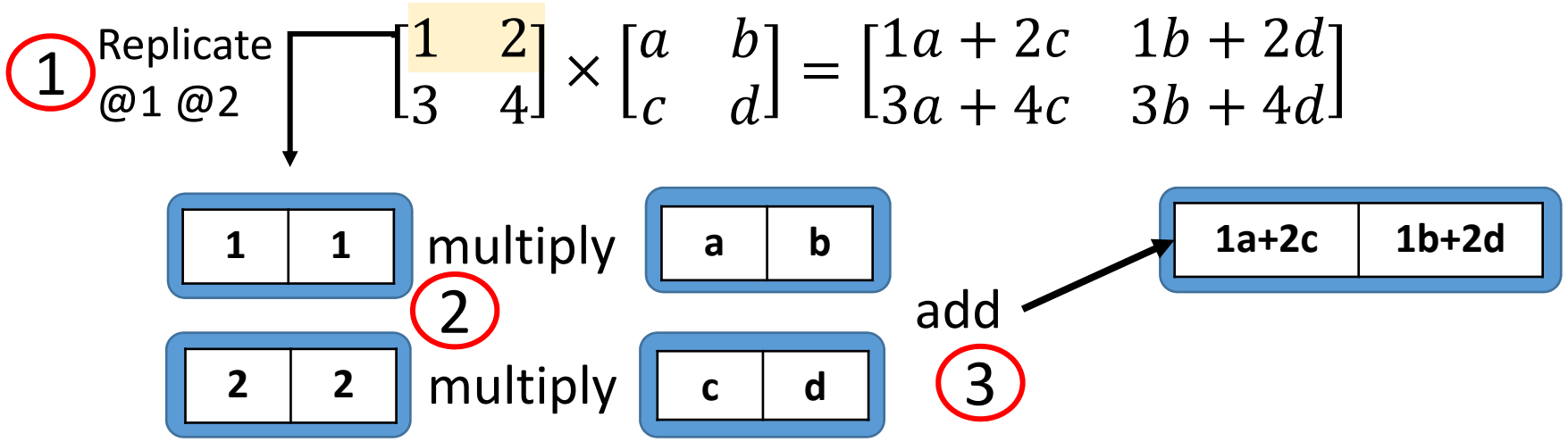
Halevi et al. *Algorithms in Helib*. 2014.

Proposed Matrix Primitive

- Used for adding & multiplying encrypted matrices
- Encrypt each row separately by packing.
  - Row-wise encryption.
  - Horizontally partitioned data
- Efficient and layout consistent.
  - $O(N^2)$ homomorphic operations.

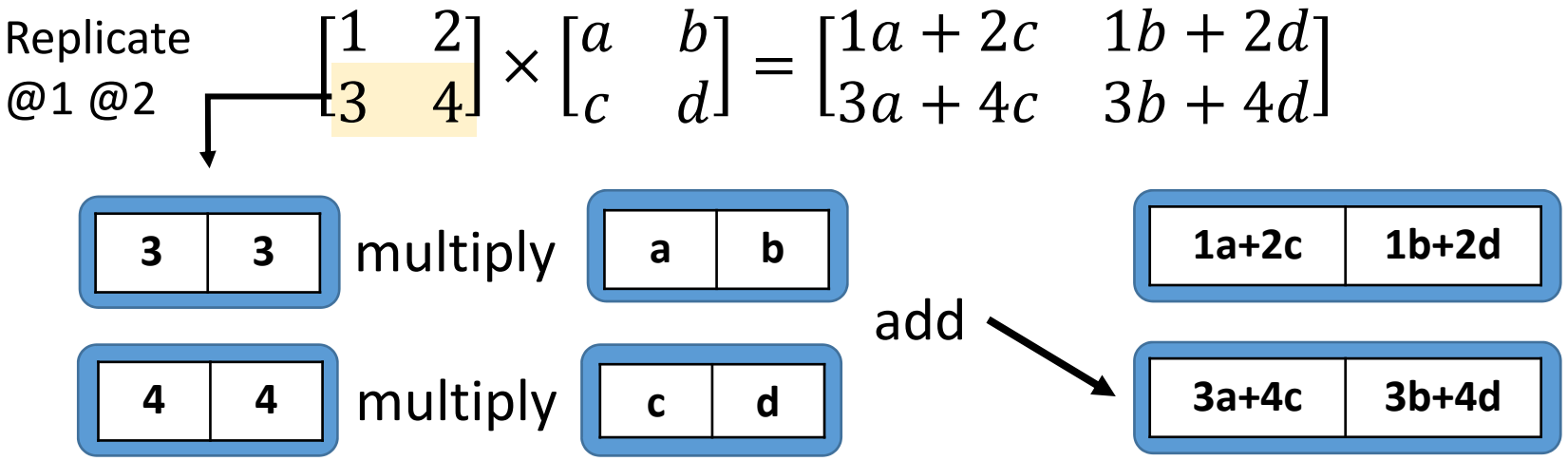
Matrix Multiplication[1/2]

- Encrypt the matrix row by row with packing.



Matrix Multiplication[1/2]

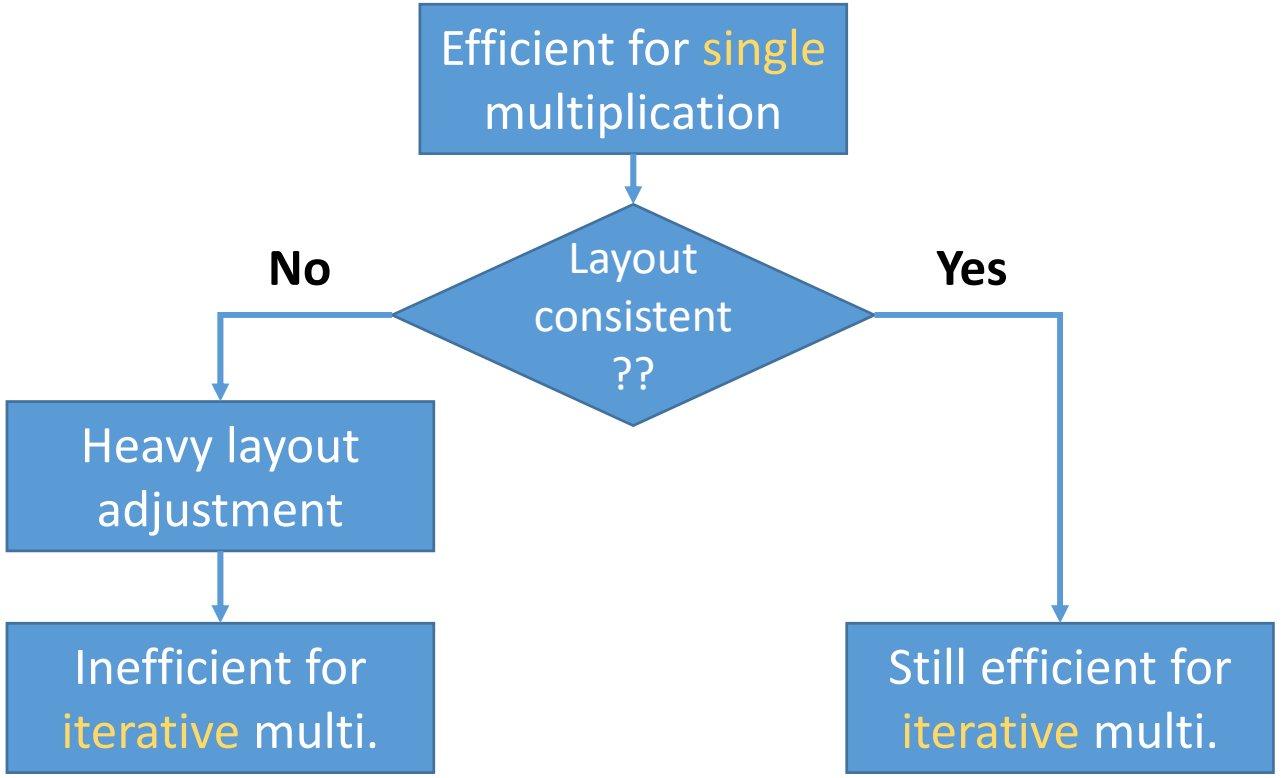
- Encrypt the matrix row by row with packing.



- N^2 replications, multiplications and additions
  - $O(N^2)$ complexity compared to $O(N^3)$ (no packing).
- Also row-wisely encrypted resulting matrix.

Matrix Multiplication[2/2]

- Layout consistency is important for developing efficient statistical protocols.
  - Statistical algorithms need iterative matrix multiplications

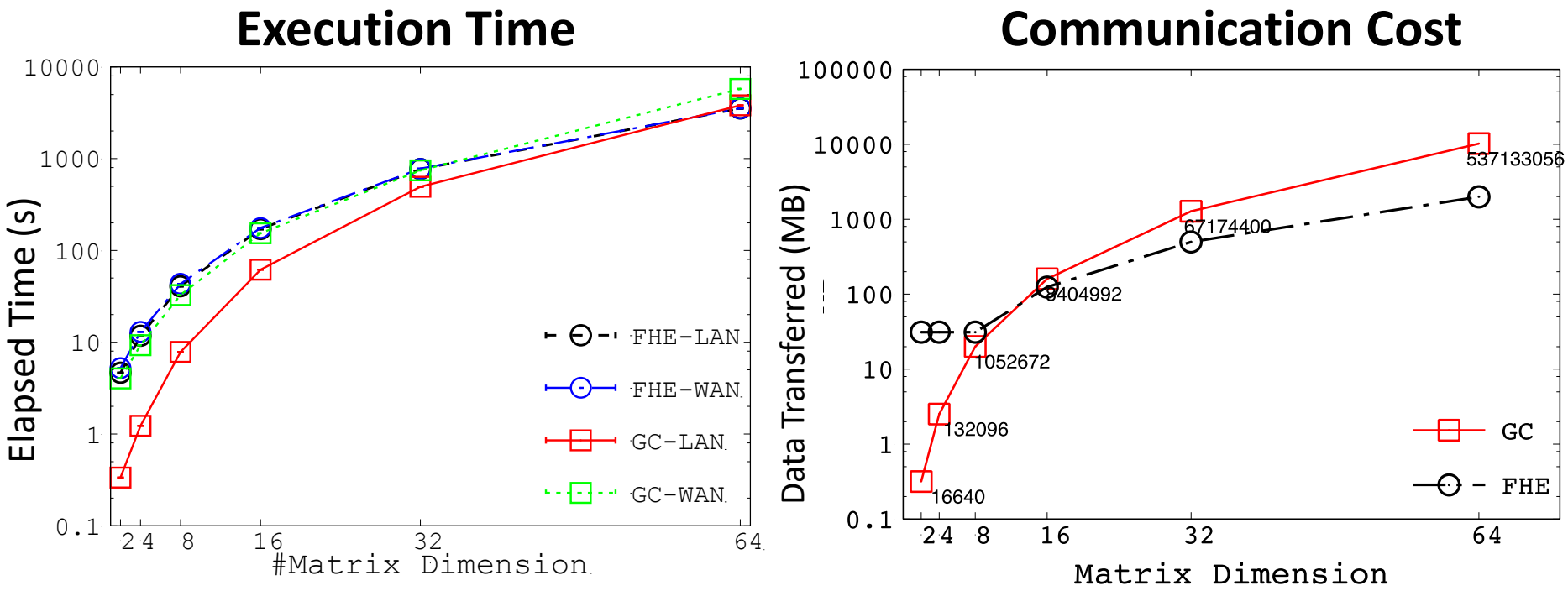


Experimental Settings of Matrix Primitive

- Implementations:
  - FHE: HElib (C++ based)
  - GC : ObliVM (java based)
- Evaluated on 32-bit integers
- Networks:
  - LAN (about 88 Mbps)
  - WAN (about 48 Mbps)

HElib. https://github.com/shaih/HElib.
Liu et al. *ObliVM: A programming framework for secure computation*. 2015.

Evaluation of Matrix Primitive



- When do iterative multiplications, FHE-based primitive can offer better performance.
  - Save communication cost between each iteration

Greater-than (GT) Primitive

$$\mathrm{GT}\big(\mathrm{e}(x), e(y)\big) \rightarrow e(x >^{?} y) \text{ s.t. } 0 \leq x, y \leq \mathrm{D}$$

- [Golle06] based on Paillier cryptosystem:

$$if\ x > y\ then\ \exists k \in [1, D] \rightarrow x - y - k = 0$$

- Combination with packing gives great improvements:

$$e([x, \dots, x]) - e([y, \dots, y]) - [1, 2, \dots, D] \rightarrow e(\boldsymbol{\eta})$$

Replicated D times

  - $0 \in \boldsymbol{\eta} \Leftrightarrow x > y$ (i.e., decryption is needed)
  - Complexity from D to $\lceil \mathrm{D}/\ell \rceil$.

Golle. *A private stable matching algorithm*. 2006.

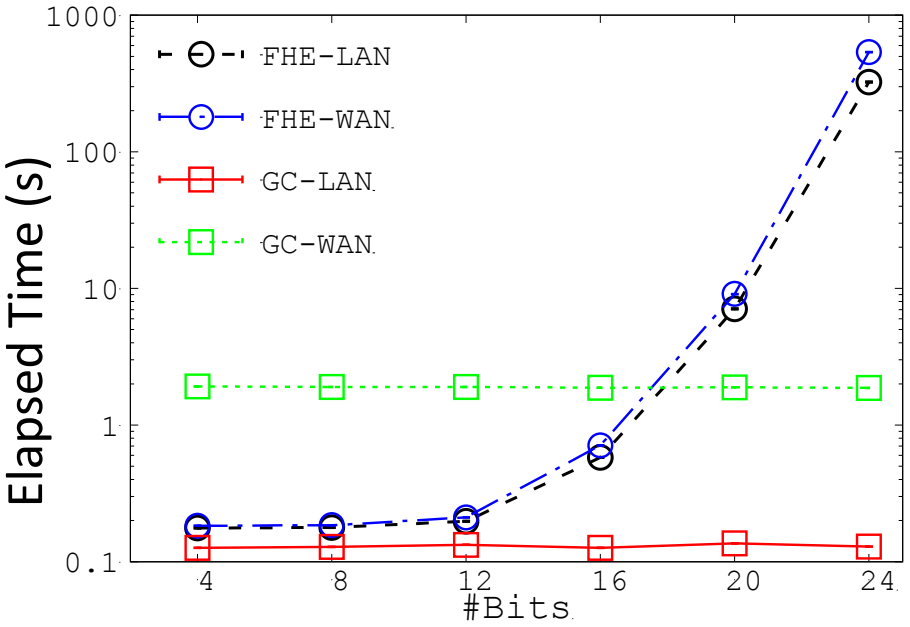
Experimental Settings for GT Primitive

- Implementations:
  - FHE: HElib (C++ based)
  - GC : ObliVM (java based)
- Domain $D = 2^4 \sim 2^{24}$
- Number of slots $\ell \approx 1700$.
- Networks:
  - LAN (about 88 Mbps)
  - WAN (about 48 Mbps)

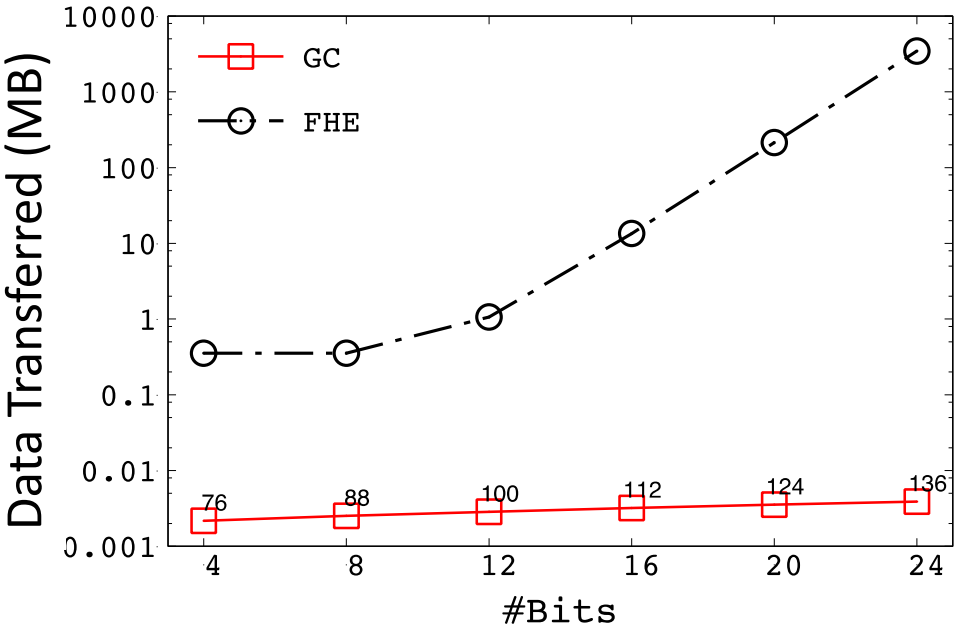
HElib. https://github.com/shaih/HElib.
Liu et al. *ObliVM: A programming framework for secure computation*. 2015.

Evaluation of Greater-than Primitive

Execution Time



Communication Cost



Works for small domains, which is enough for ordinal statistics.

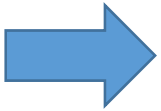
Secure Statistical Protocols

- Contingency table with cell-suppression protocol:
  - Use the greater-than primitive.
  - One round protocol between cloud and analyst.
- Linear regression protocol:
  - Use the matrix primitive.
  - Two rounds protocol.
  - Use a plaintext precision expansion technique (discuss it latter).

Contingency Table

Gender	Smoke
Male	Smoker
Female	Non-smoker
Male	Non-Smoker

Categorical data



$K_1 = 2$

$K_2 = 2$

	Smoker	Non-smoker
Male	1	1
Female	0	1

Contingency Table

- Indicator encoding:

  Male → [1, 0], Female → [0, 1]

  Smoker → [1, 0], Non-smoker → [0, 1]

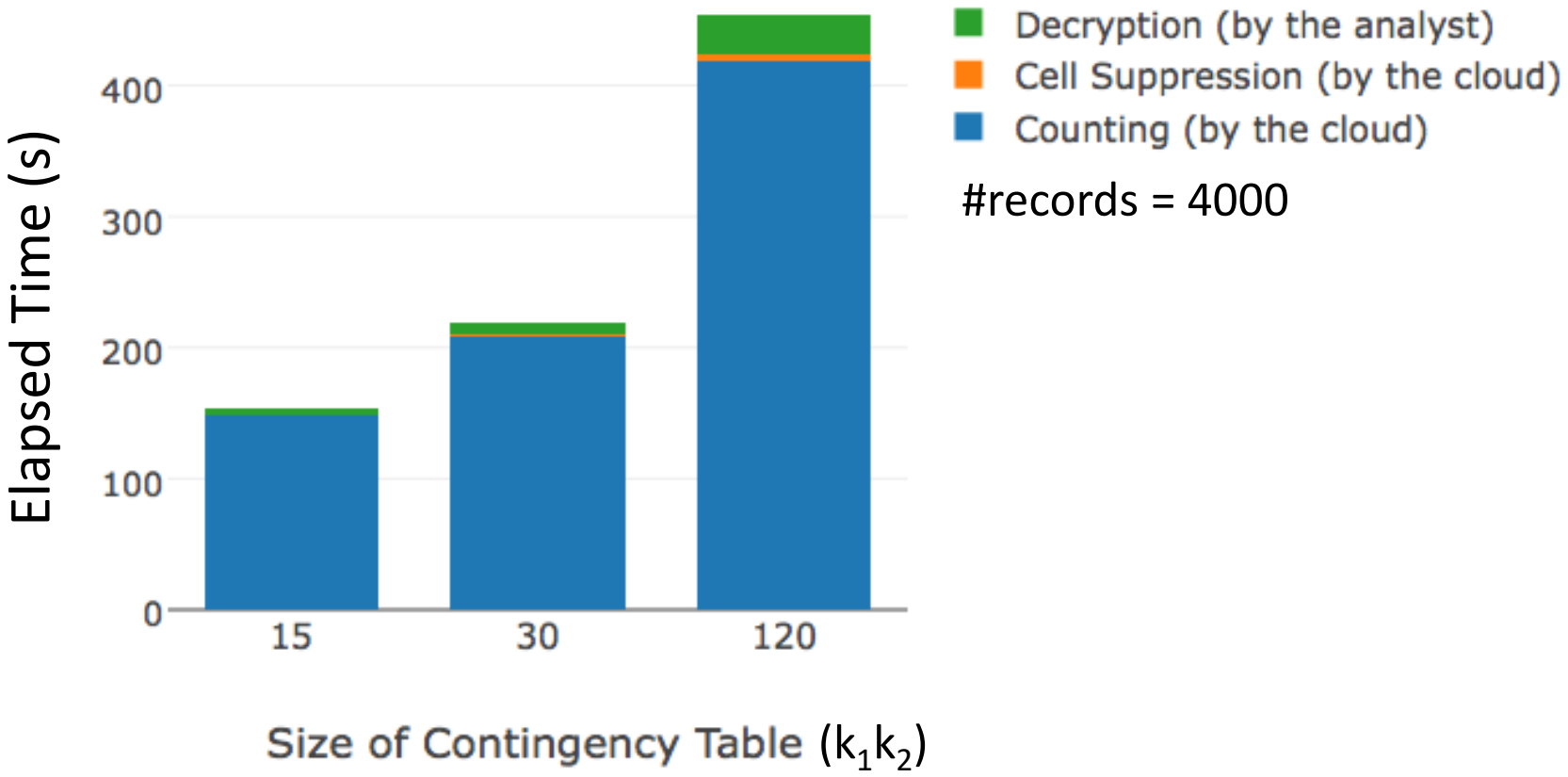
- Basic Idea: **multiply & rotate**

  $[a_1, a_2]$ x $[b_1, b_2]$ counts Male-Smoker, and Female-Nonsmoker

  $[a_1, a_2]$ x $([b_1, b_2]>>1) = [a_1, a_2]$ x $[b_2, b_1]$ gives the other two counts.

- Improvement with no extra preprocessing
  - $O(\max(k_1,k_2)) \Rightarrow O(\log k_1 k_2)$.

Contingency Table Performance Evaluation

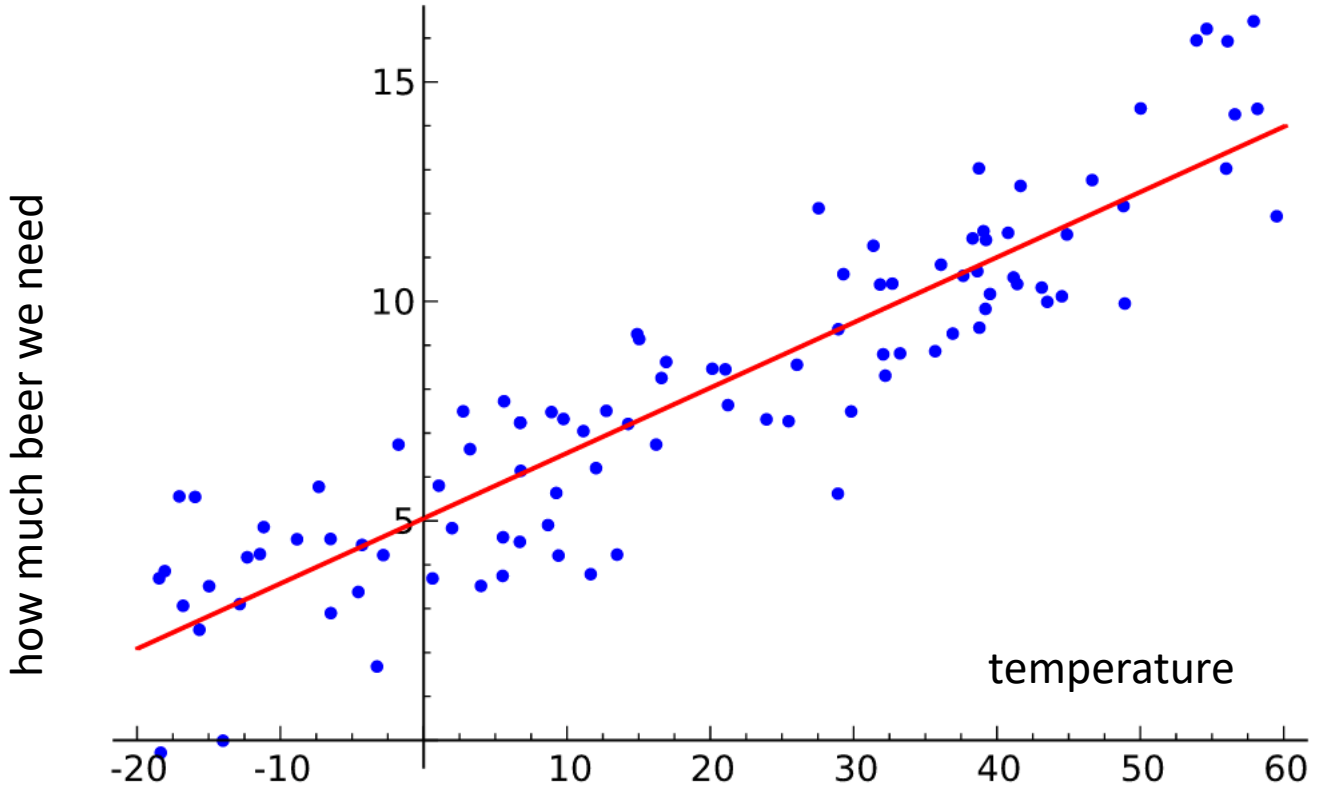


- Complexity increases logarithmically with the table sizes.
- Most of the work (>90%) done by the cloud.

Liner regression

- From data $\{(\boldsymbol{x}_i, yi)\}_i$, computes a model $\boldsymbol{w}$ s.t.

$$\boldsymbol{w} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$$



How to solve LR over FHE

- The inversion of an encrypted matrix.

Division-free Matrix Inversion $(\boldsymbol{Q}, \lambda)$:

set $\boldsymbol{A}^{(1)} = \boldsymbol{Q}, \boldsymbol{R}^{(1)} = \boldsymbol{I}, a^{(1)} = \lambda$, and *iterate*

$$\boldsymbol{R}^{(t+1)} = 2a^{(t)}\boldsymbol{R}^{(t)} - \boldsymbol{R}^{(t)}\boldsymbol{A}^{(t)}$$

$$\boldsymbol{A}^{(t+1)} = 2a^{(t)}\boldsymbol{A}^{(t)} - \boldsymbol{A}^{(t)}\boldsymbol{A}^{(t)}$$

$$a^{(t+1)} = a^{(t)}a^{(t)}$$

Layout consistency leads to efficient iterative protocols.

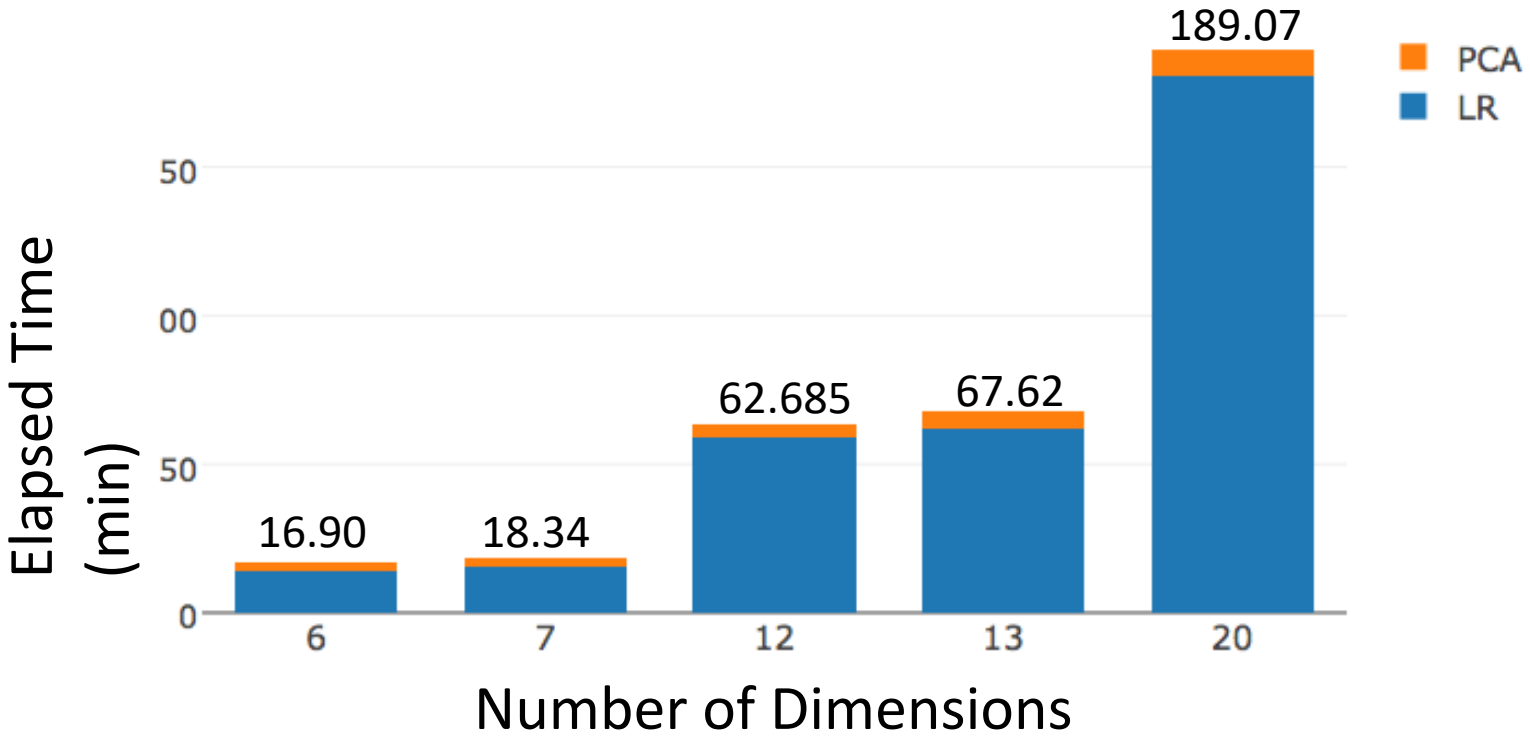
[Guo06] $\boldsymbol{R}^{(t)}$ gives a good *approximation* to $\lambda^{2^t}\boldsymbol{Q}^{-1}$ if λ is close to largest eigenvalue of $\boldsymbol{Q}$ (use PCA to compute λ).

Guo et al. *A Schur-Newton method for the matrix pth root and its inverse*. 2006.

Plaintext Precision Expansion (PPE)

- Division-free algorithms introduce large integers. (λ^{2^t})
  - But the current FHE library allows at most 60-bit integers.
- Allows division-free algorithms without changing the FHE library.
- Uses *K* different FHE parameters (each *b*-bit < 60)
  - Achieves an equivalent *Kb*-bit parameter.
  - Increases the time by *K* times, but naturally parallelizable.
- Direct application of the Chinese Remainder Theorem.

Experiments: Linear Regression



- Negligible decryption time (less than 2 s).
- 20x faster than previous FHE solution [Wu et al. 12]
  - 5 dimensions (400+ mins).
- Good scalability (reduced execution using more cores).

Summary

- Secure statistical analysis in the cloud with multiple data providers.
- Two primitives
  - Matrix operation and greater-than
- Two protocols.
  - Contingency table and linear regression.
- Encoding and packing can improve FHE's balance between generality and efficiency.

The Gap we see

- Epidemiology
  - Carefully design the objective of analysis
  - Collect a limited amount of well-controlled data
  - We evaluate whether or not the result is “significant”
- Big data analysis
  - Collect a huge amount of not well-controlled data
  - Objective of analysis is explorative
  - We can find a “good finding” if we are lucky
- How can we have “significant” finding with big data approaches?