Techtonic 2018

-

Thu . Nov 15

-

SAMSUNG SDS Tower
West Campus B1F
Magellan Hall /Pascal Hall

SAMSUNG SDS
Realize your vision

Privacy Preserving Data Mining!

데이터가 돈이 되는 세상

삼성SDS 윤효진 프로
서울대학교 박사과정 한규형

Techtonic 2018

Agenda

- 프라이버시 관련 동향
- 사업활용방안
- 동형암호 – HeaAn 기술소개
- PoC 수행 결과

프라이버시 관련 문제, 법/규제, 그리고 기술

프라이버시 관련 동향

들어가며

Data = Money

프라이버시 위험

페이스북 네트워크 해킹...
사용자 5천만명 개인정보 위험

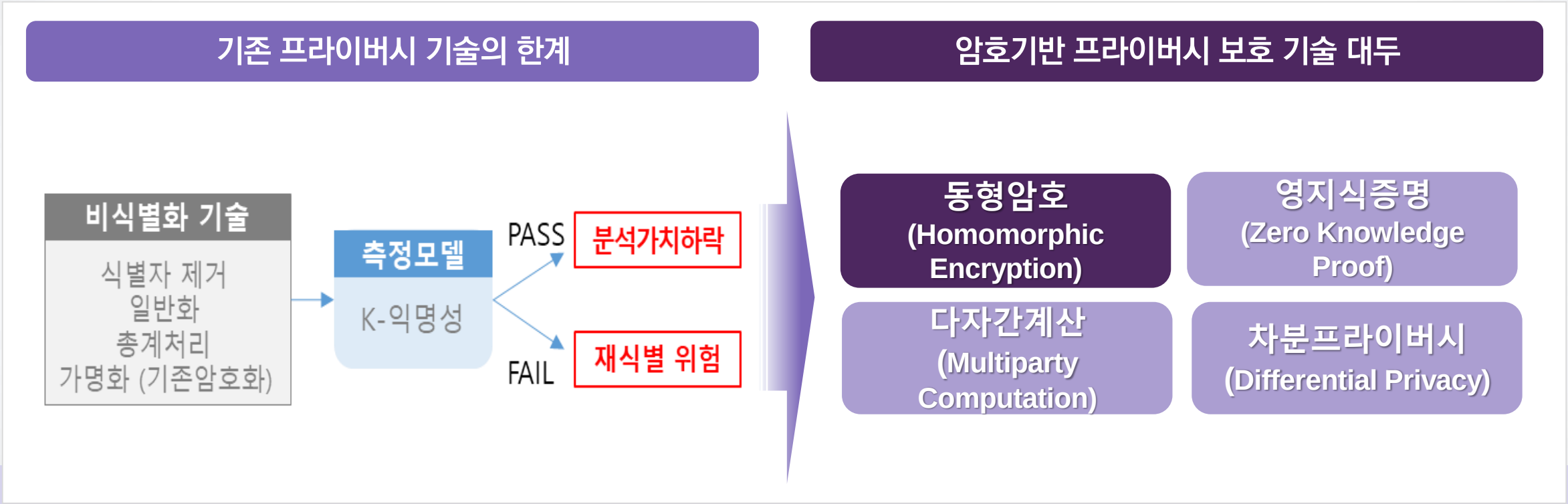
프라이버시 관련 법/규제

매출의 4%까지 벌금이 부과됨

데이터 중심의 세상, **프라이버시 보호 기술**은 더 이상 **선택이 아니라 필수**입니다.

프라이버시 보호기술 변화

기존 프라이버시 보호 기술의 한계로 **암호 기술 기반의 프라이버시 보호 기술** 연구 및 사업적용 요구 증가

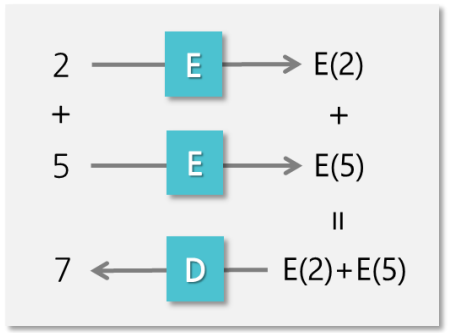


Google, IBM, Microsoft, Cisco, Ant Financial 등 글로벌 기업과 스타트업들이
금융, 신용, 의료, 클라우드, 마케팅 등에 암호기반 프라이버시 기술 적극적 도입

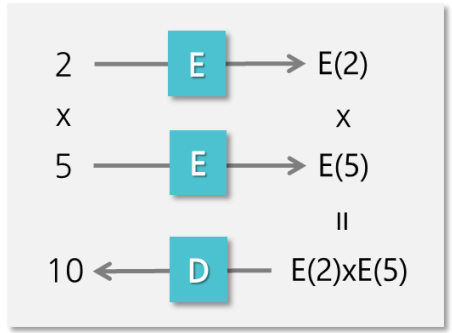
동형암호란

동형암호란 데이터의 손실/유출 없이 **암호화된 상태에서 연산 및 분석이 가능하도록 지원**하는 기술

덧셈 보존



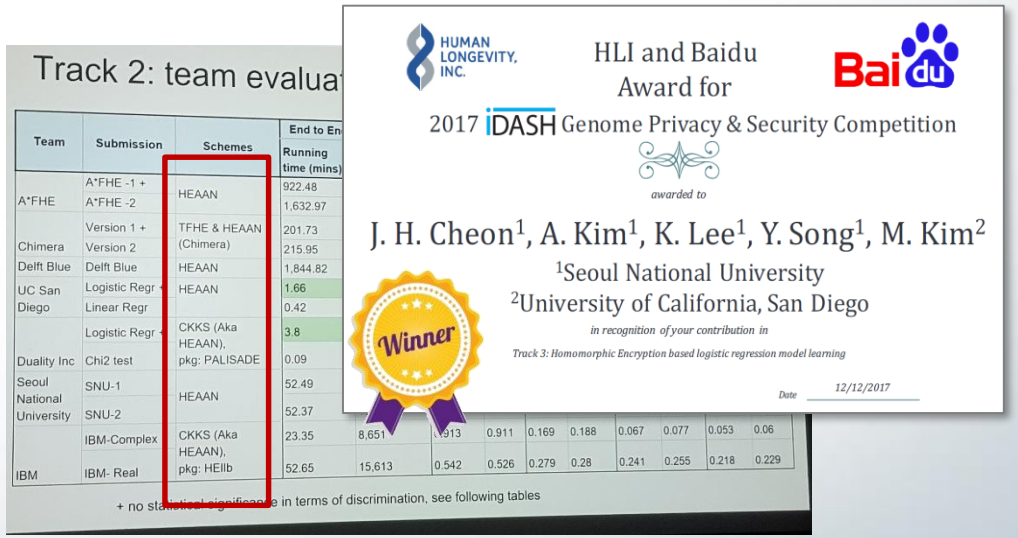
곱셈 보존



- **동형암호는 덧셈/곱셈을 보존**
  암호문의 연산 결과를 복호화 하면 평문의 계산 결과와 동일
- **모든 산술연산 지원**
  근사 계산을 통해 모든 산술연산 지원 가능
- **암호화된 상태에서 분석/처리**
  산술 연산으로 이루어진 머신러닝 또는 딥러닝 등에 적용 가능

Samsung SDS는 세계 최고의 동형암호기술 (HeaAn) 확보를 위해 서울대학교 암호랩과 협업

- **학계 검증 완료**
  Asiacrypt 2017, Eurocrypt 2018, SAC 2018 등 Toptier 암호학회 발표를 통해 학계 검증 완료
- **암호화된 분석지원 검증 완료**
  2017 iDash 우승
  2018 iDash 모든 예선통과자 HeaAn 사용

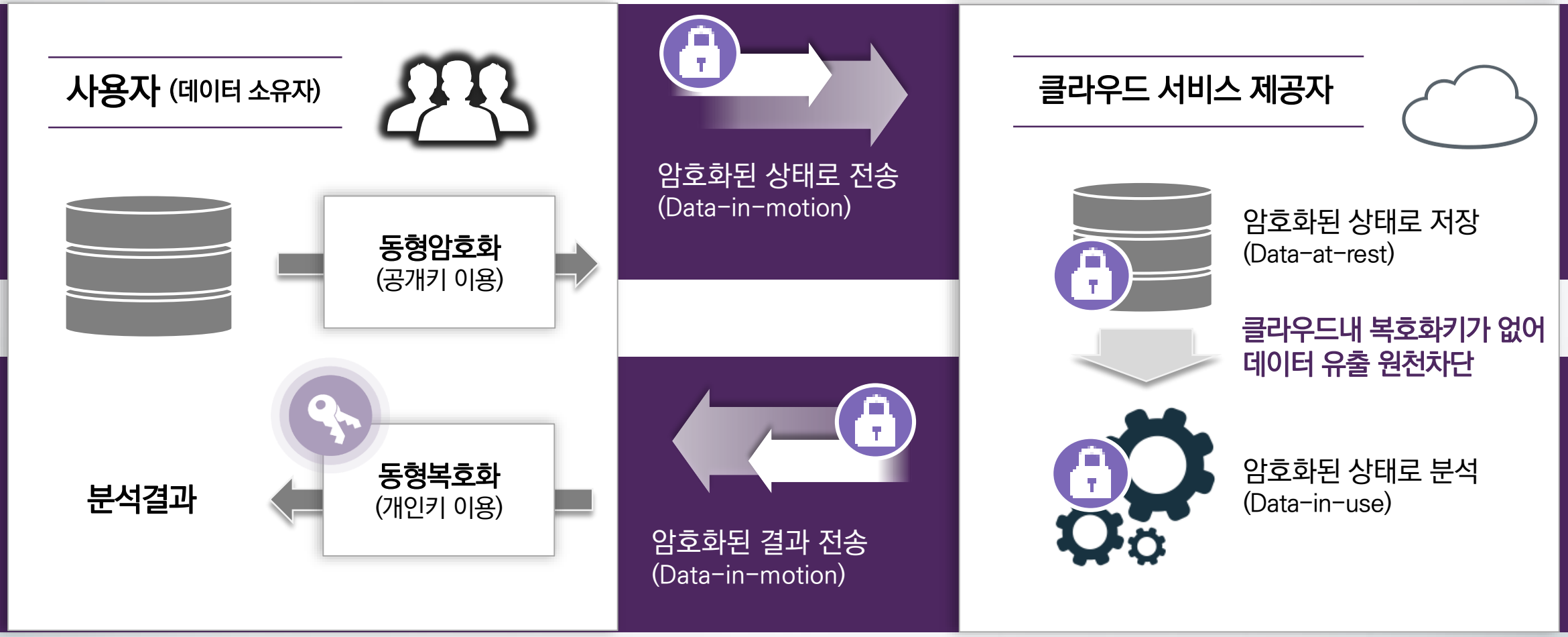


프라이버시 보호 기술 적용 Use Case 발굴

동형암호 사업활용방안

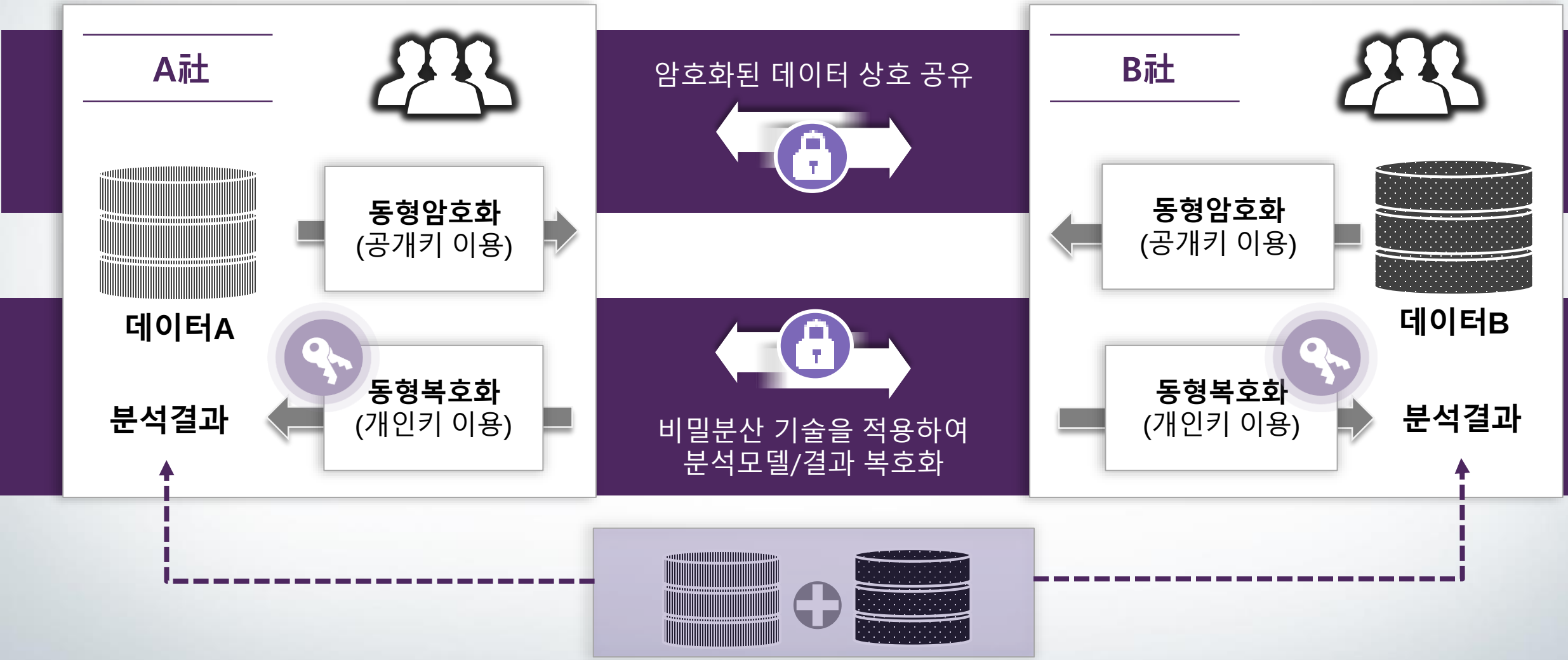
신뢰할 수 없는 환경에서 암호화된 상태의 데이터 분석 지원

사용자는 데이터 유출 걱정 없이 퍼블릭 클라우드 등의 외부 분석 서비스를 이용할 수 있는 환경 제공



분산된 데이터에 대해 암호화된 상태의 결합분석 지원

상대 또는 제 3자로의 데이터 유출 걱정없이 더 정밀한 분석모델 도출 가능



분산된 데이터에 대해 암호화된 상태의 결합분석 지원

상대 또는 제 3자로의 데이터 유출 걱정없이 더 정밀한 분석모델 도출 가능

데이터 융합 결합 예시

속성 결합: DB의 열 증가

A

ID	속성 A
1	A(1)
2	A(2)
...	...
n	A(n)

B

ID	속성 B
1	B(1)
2	B(2)
...	...
n	B(n)

암호화된 상태로 공유

ID	속성 A	속성 B
1	Enc(A(1))	Enc(B(1))
2	Enc(A(2))	Enc(B(2))
...	...	...
n	Enc(A(n))	Enc(B(n))

속성 A와 B 전체에 대한 분석 결과

암호화된 상태로 처리

$$f(Enc(A(i), Enc(B(j)) = Enc(f(A(i), B(i)) = \mathrm{Enc}(\mathrm{Result})$$

사례 결합: DB의 행 증가

A

ID	속성 A	속성 B
1	A(1)	B(1)
2	A(2)	B(2)
...	...	...
n	A(n)	B(n)

B

ID	속성 A	속성 B
n+1	A(n+1)	B(n+1)
n+2	A(n+2)	B(n+2)
...	...	...
n+m	A(n+m)	B(n+m)

암호화된 상태로 공유

ID	속성 A	속성 B
1	Enc(A(1))	Enc(B(1))
2	Enc(A(2))	Enc(B(2))
...	...	...
n	Enc(A(n))	Enc(B(n))
...	...	...
n+m	Enc(A(n+m))	Enc(B(n+m))

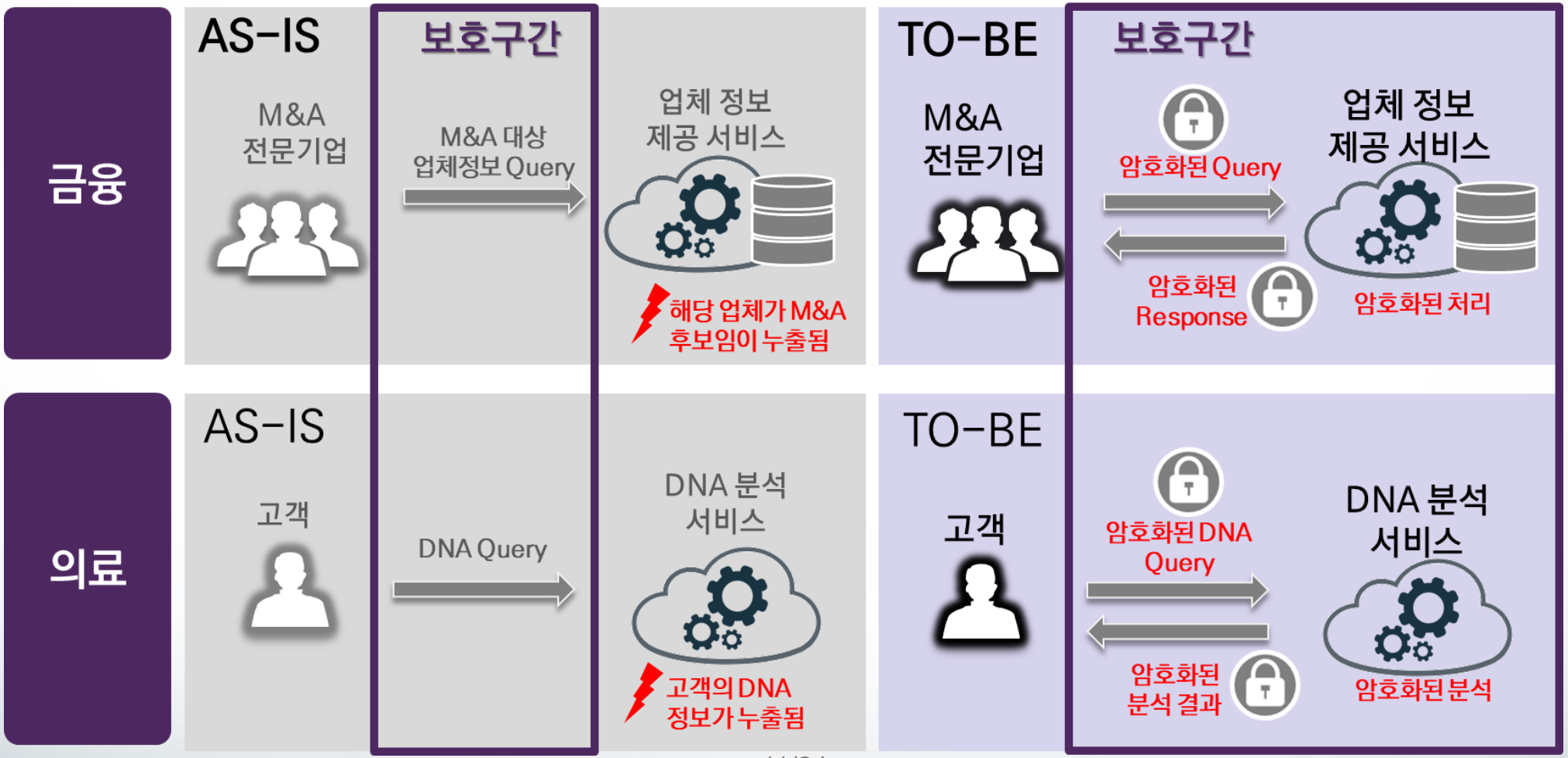
전체 사례 (1~n+m)의 분석 결과

암호화된 상태로 처리

$$f_{1\ to\ m+n}(Enc(A(i), Enc(B(j)) = Enc(f(A(i), B(i)) = \mathrm{Enc}(\mathrm{Result})$$

고객의 질의/응답내용 보호

고객은 질의/응답 내용을 서비스 제공자를 비롯한 어느 누구에게도 유출시키지 않고 원하는 서비스를 이용할 수 있는 Private Information Retrieval (PIR) 지원



암호화된 실수 간의 효율적인 연산 기술

HeaAn (혜안, 慧眼)

HEAAN (혜안 慧眼)

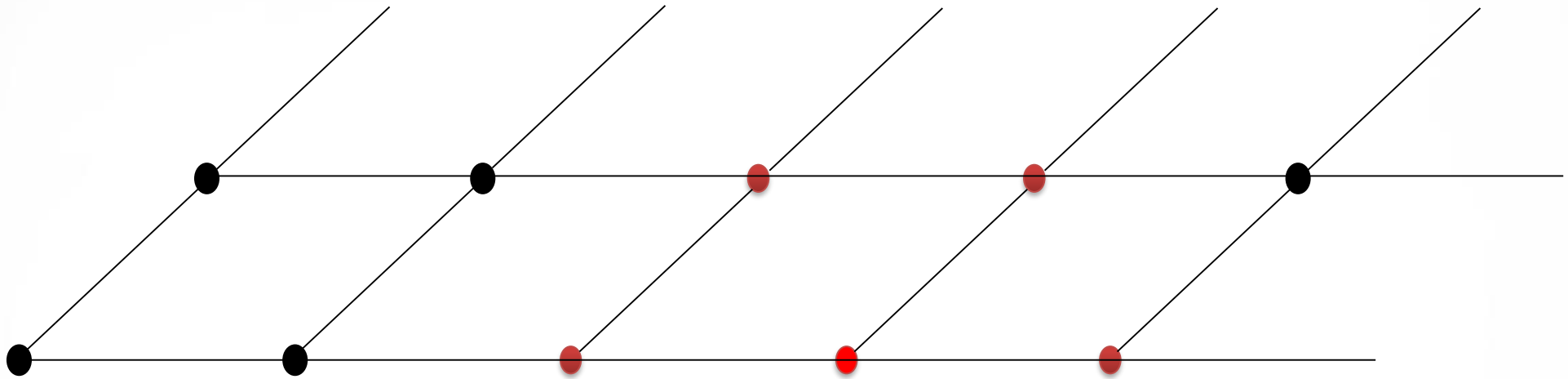
- 실제 응용에서 필요로 하는 실수 계산을 암호화된 상태로 수행하는 최초의 암호 시스템

PoC 수행 결과

- 암호화된 데이터를 이용해서 회귀분석을 수행
- 현실적인 데이터 크기를 이용한 최초의 결과 (약 40만건)
- 약 16시간에 0.8 이상의 AUROC를 가지는 모델을 얻음

“HEAAN ?
(혜안 慧眼)” 암호화된 실수 간의 효율적인 계산하는 기술

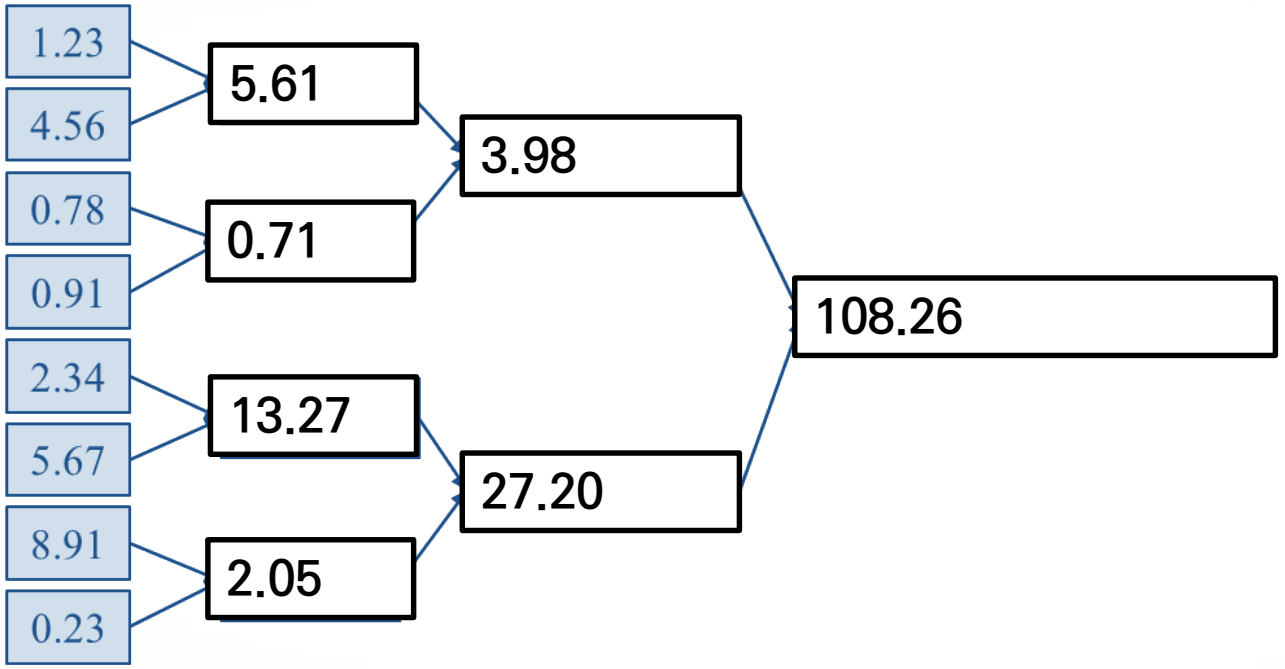
격자기반 동형암호 (Lattice Based Homomorphic Encryption)



$$< \vec{c}_1, \overrightarrow{sk} > \simeq m_1, < \vec{c}_2, \overrightarrow{sk} > \simeq m_2$$
$$< \vec{c}_1 \otimes \vec{c}_2, \overrightarrow{sk} \otimes \overrightarrow{sk} > \simeq m_1 m_2$$

- 격자(lattice)에서 짧은 원소를 찾는 문제 = NP-hard
- 동형암호의 암호문에서 평문의 정보를 얻는 방법 -> 격자에서 짧은 원소를 찾는 방법
- ”동형” 성질은 격자의 구조적 특성으로 인해서 얻어짐

HeaAn (혜안 慧眼) (Homomorphic Encryption for Arithmetic of Approximate Numbers)

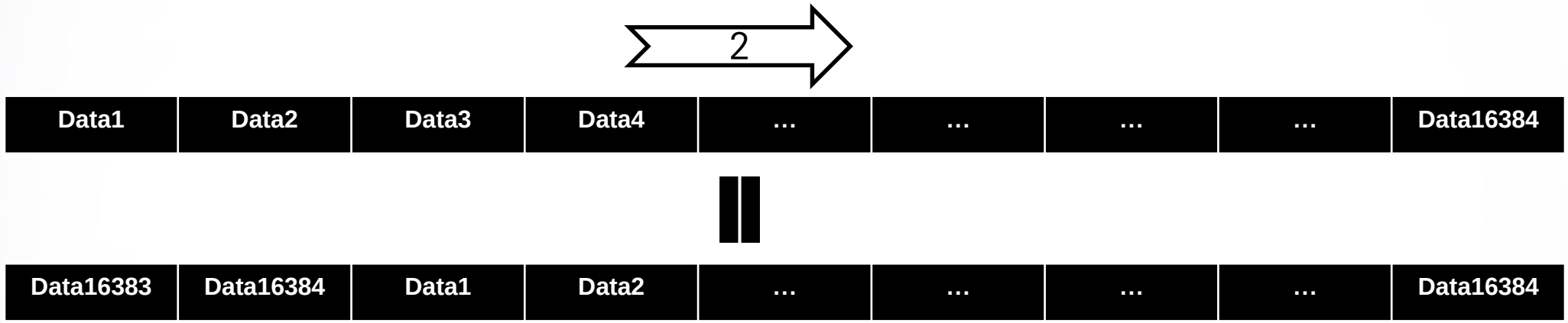


1. 기존 동형암호의 문제점: 정확한 정수연산을 암호화된 상태에서 수행
2. HeaAn: 암호화된 상태에서 아래 자릿수를 버리는 **Rescaling**이라는 기능 지원

HeaAn (혜안 慧眼) (Homomorphic Encryption for Arithmetic of Approximate Numbers)

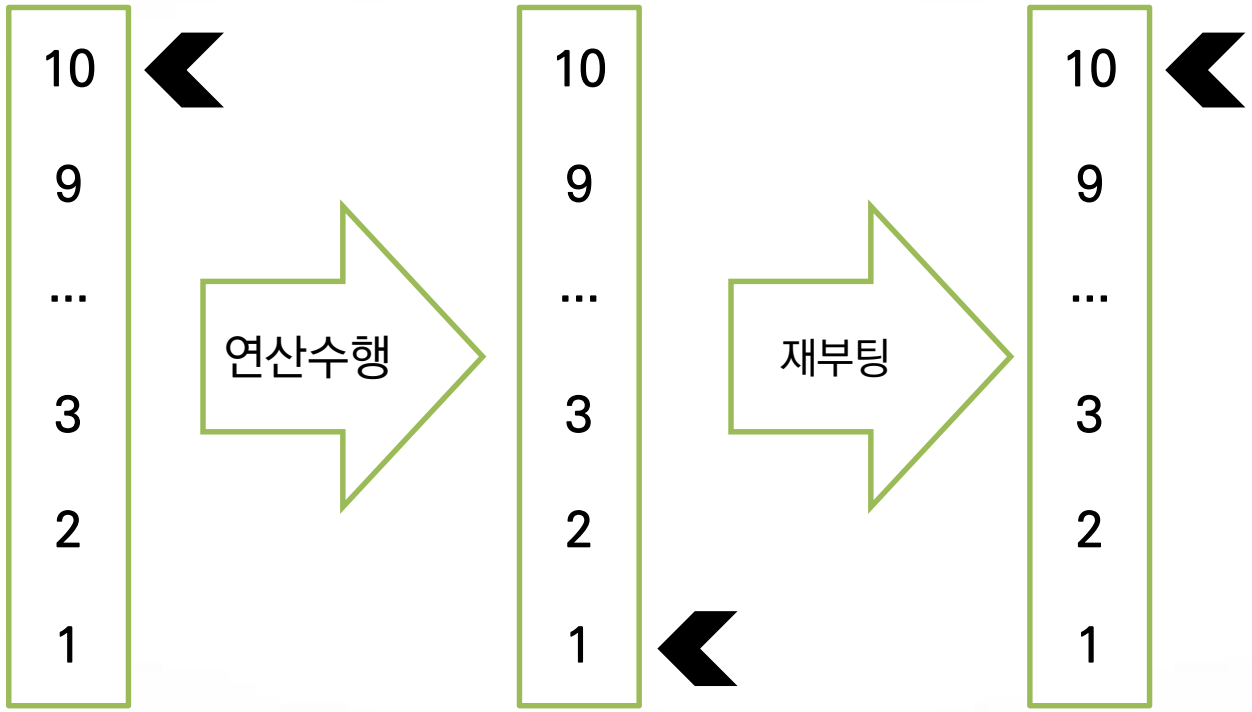
- **Single instruction, multiple data (SIMD) 연산의 지원**
- Multiple Data management is also possible!!

HeaAn (혜안 慧眼) (Homomorphic Encryption for Arithmetic of Approximate Numbers)



- Single instruction, multiple data (SIMD) 연산의 지원
- **Multiple Data management is also possible!!**

HeaAn (혜안 慧眼) (Homomorphic Encryption for Arithmetic of Approximate Numbers)



- HeaAn은 제한된 depth (or degree)의 연산만 암호화된 상태로 수행
- 재부팅기법은 이러한 한계를 제거하여 무제한의 연산을 수행하게 해줌

HeaAn (혜안 慧眼) (Homomorphic Encryption for Arithmetic of Approximate Numbers)

	Data size	Enc(ms)	Dec(ms)	연산시간
add	16384	110 ms	40 ms	8.6 ms
mult	16384			250 ms
average	16384			1.8 s
variance	16384			2.7 s

Table. HEAAN을 이용한 기본적인 연산의 암호화된 수행시간

- SIMD 길이 16384의 더하기, 곱하기 연산 시간
- 16384개의 실수에 대한 평균, 분산 계산 연산 시간

의료, 신용데이터에 HeaAn 적용 분석

PoC 수행 결과

암호화된 데이터를 이용한 회귀분석 (Homomorphic Logistic Regression)

[병원]
- 데이터 소유
- 개인정보 보호를 위해서 데이터를 외부로 반출할 수가 없음
- 데이터를 이용한 분석을 수행하고 싶지만, 기술 및 도구가 부족

[기업]
- 데이터 분석 기술과 도구를 가지고 있음
- 분석과 도구를 이용할 데이터를 구하기가 힘듦

소스코드 url: https://github.com/HanKyoohyung/HELR

암호화된 데이터를 이용한 회귀분석 (Homomorphic Logistic Regression)

[1] 소유한 데이터의 암호화

[2] 암호화된 데이터 전송

[기업]
- 데이터 분석 기술과 도구를 가지고 있음
- 분석과 도구를 이용할 데이터를 구하기가 힘듦

- 평문 데이터의 크기: 1088×24 (MIMIC) / 16428×28(SEER) / 400000×200(신용정보)
- 암호화된 데이터의 크기: 16 MB / 256 MB / 39GB
- 암호화 소요시간: 0.3 초 / 1 초 / 300초 (16개 Thread 사용)

암호화된 데이터를 이용한 회귀분석 (Homomorphic Logistic Regression)

- 실제 신용정보 데이터를 이용한 PoC 진행
  - 매우 큰 데이터 (40만 × 200)
  - 현장에서 사용할 수 있는 수준의 performance
  - 라이브러리의 개선 / GPU와 같은 하드웨어의 적용을 통해 단축 가능할 것으로 예측

[1

이터 전송

[3] 암호화된 데이터를 이용한 회귀분석 수행
[4] 결과로 암호화된 회귀분석 모델 획득

[기업]
- 데이터 분석 기술과 도구를 가지고 있음
- 분석과 도구를 이용할 데이터를 구하기가 힘듦

마치며

금융 의료 제조 공공 마케팅

“Privacy Preserving”

Brightics AI
AI Analytics Platform

Nexledger
Enterprise Blockchain Platform

Brightics IoT
IoT Platform for Big Data

Q & A

Thank you