Oct.2015-Mar.2021 funded by Japan Science and Technology Agency

Secure Data Sharing and Distribution Platform for Integrated Big Data Utilization

- Handling all data with encryption -

Group Members

Waseda University Hayato YAMANA
Institute of Information Security Atsuhiro GOTO
Ochanomizu University Masato OGUCHI
Kogakuin University Saneyasu YAMAGUCHI
The University of Electro-Communications Takahiko SHINTANI
Meiji Pharmaceutical University Tamotsu NOGUCHI

Brief Introduction of our Project

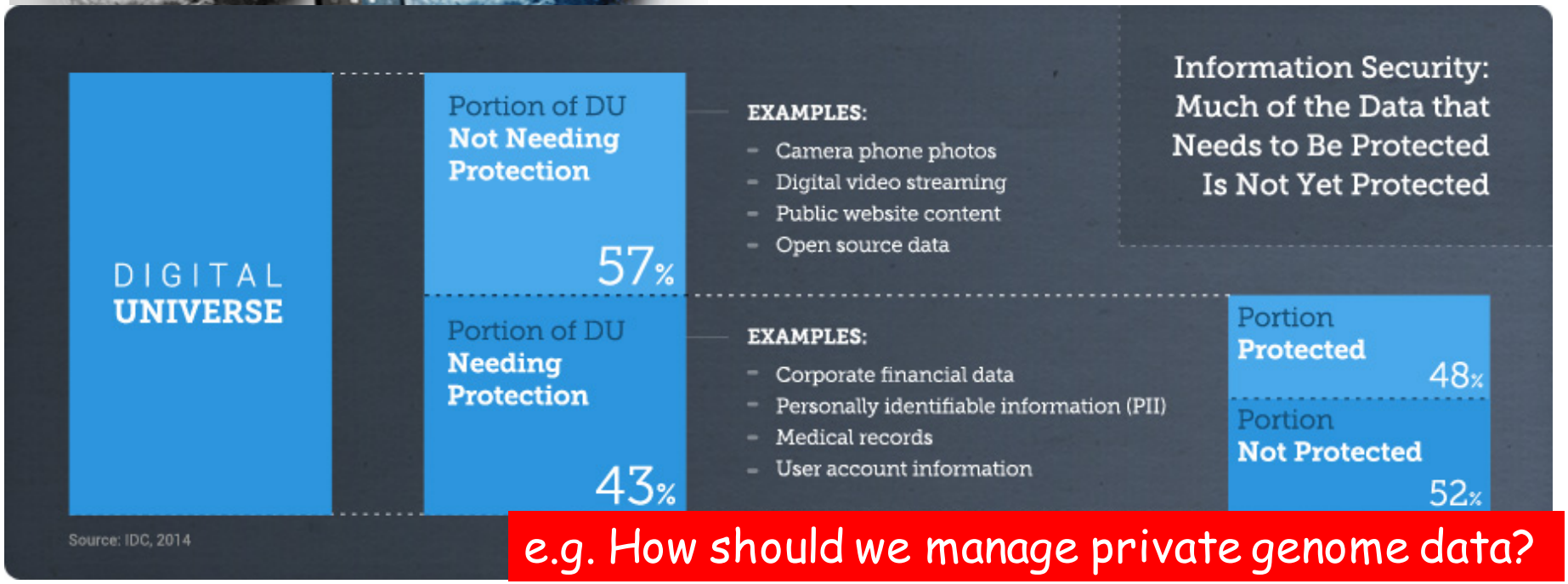
1. Research Background
2. Objective
3. Research Goal
4. Research Strategy
5. Experiment
6. Schedule
7. Progress in 2015FY

1. Research Background

SD² Platform for Integrated Big Data Utilization

At least 40% of it requires some level of security, from privacy protection to full-encryption 'lockdown.' …

Also unfortunately, the amount needing protection will grow …



e.g. How should we manage private genome data?

(*) http://www.emc.com/leadership/digital-universe/2014iview/

1. Research Background

SD² Platform for Integrated Big Data Utilization

- Anonymization
  - Attribute Linkage Model
    - **k-anonymity, l-diversity, t-closen**
  - Probabilistic Model
    - **differential privacy**

LIMITATION

Limitation of Anonymization

Link Attack

William was governor of Ma[illegible]etts and his medical record[illegible] in the GIC data. Governor W[illegible]a in Cambridge. Accordi[illegible]e Cambridge Voter list, six p[illegible]d his particular birth date; [illegible]ee of them were men; and, he [illegible]he only one in his 5-digit ZIP code.

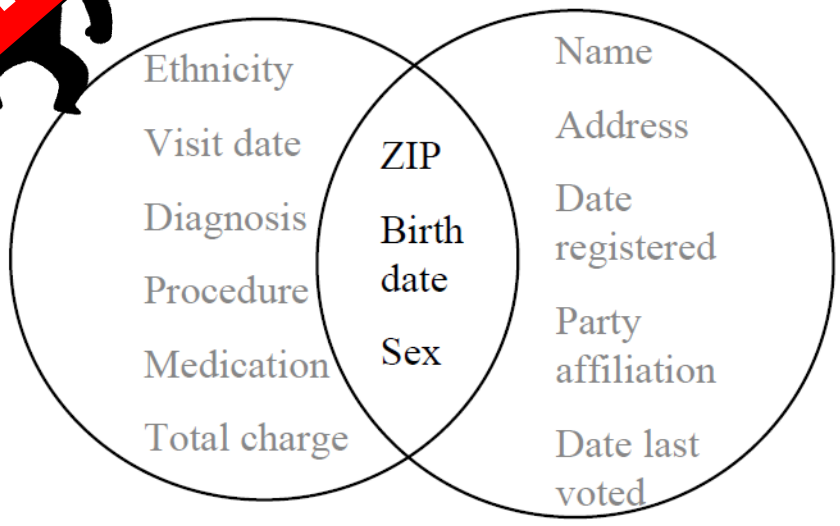


Figure 1 Linking to re-identify data

2. OBJECTIVE

OUR APPROACH
IS NOT ANONYMIZATION

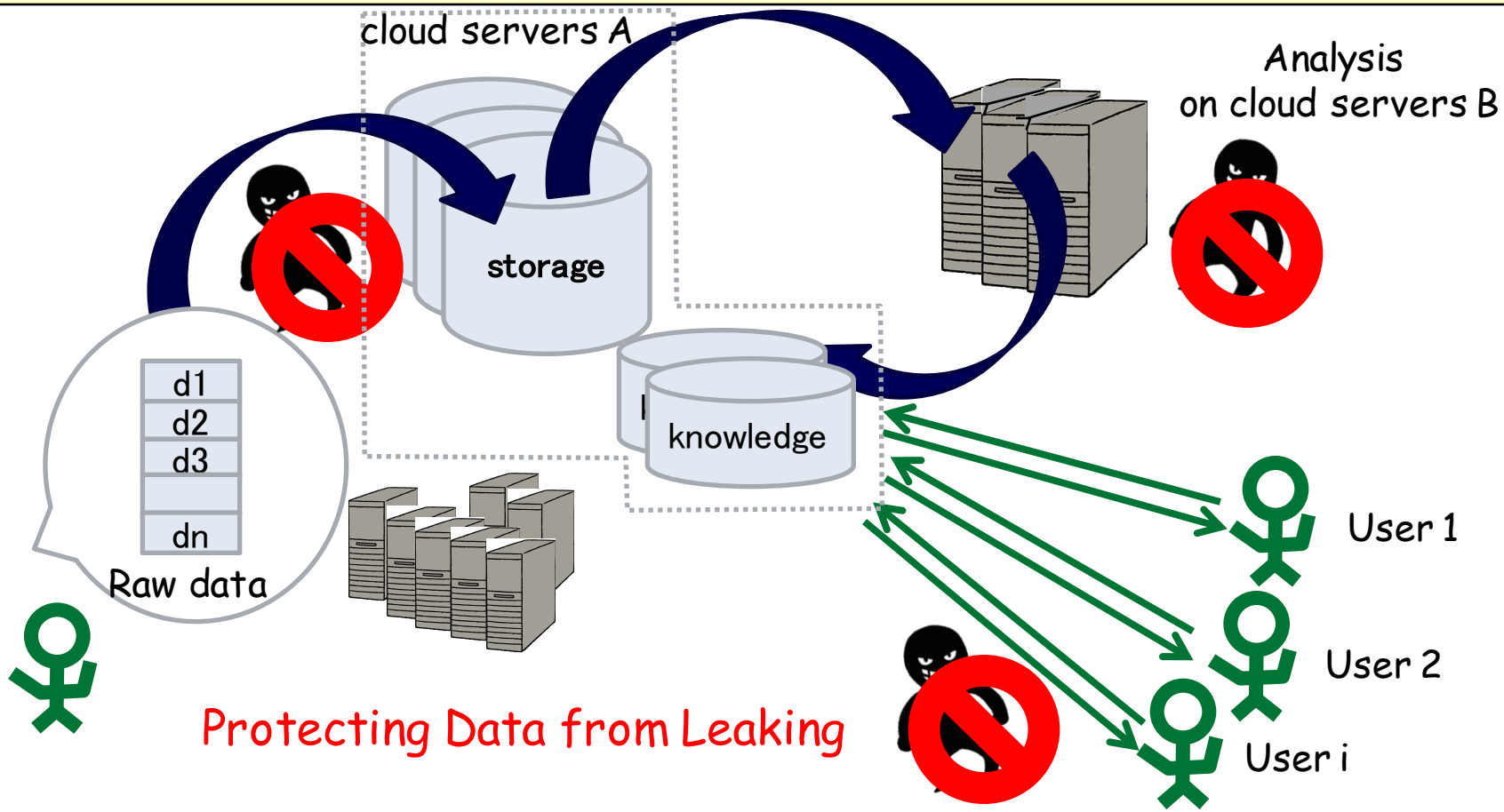
NEW POINT

OUR APPROACH
IS
HANDLING ALL DATA
WITH ENCRYPTION 
THROUGHOUT DATA LIFE CYCLE

YOU CAN ADOPT
ANONYMIZATION, BESIDES.

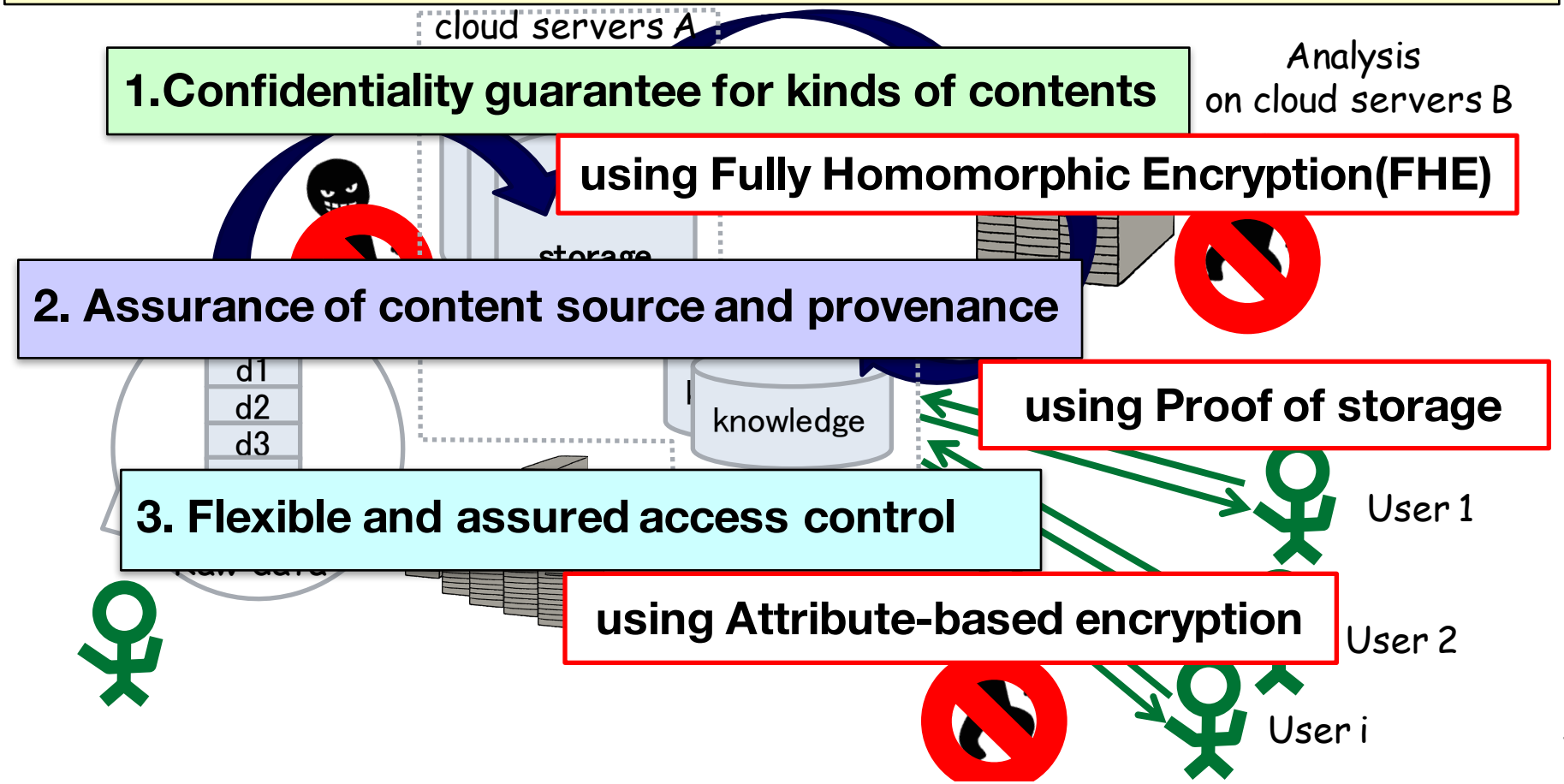
3. Research Goal

HANDLING ALL DATA
WITH ENCRYPTION
THROUGHOUT DATA LIFE CYCLE



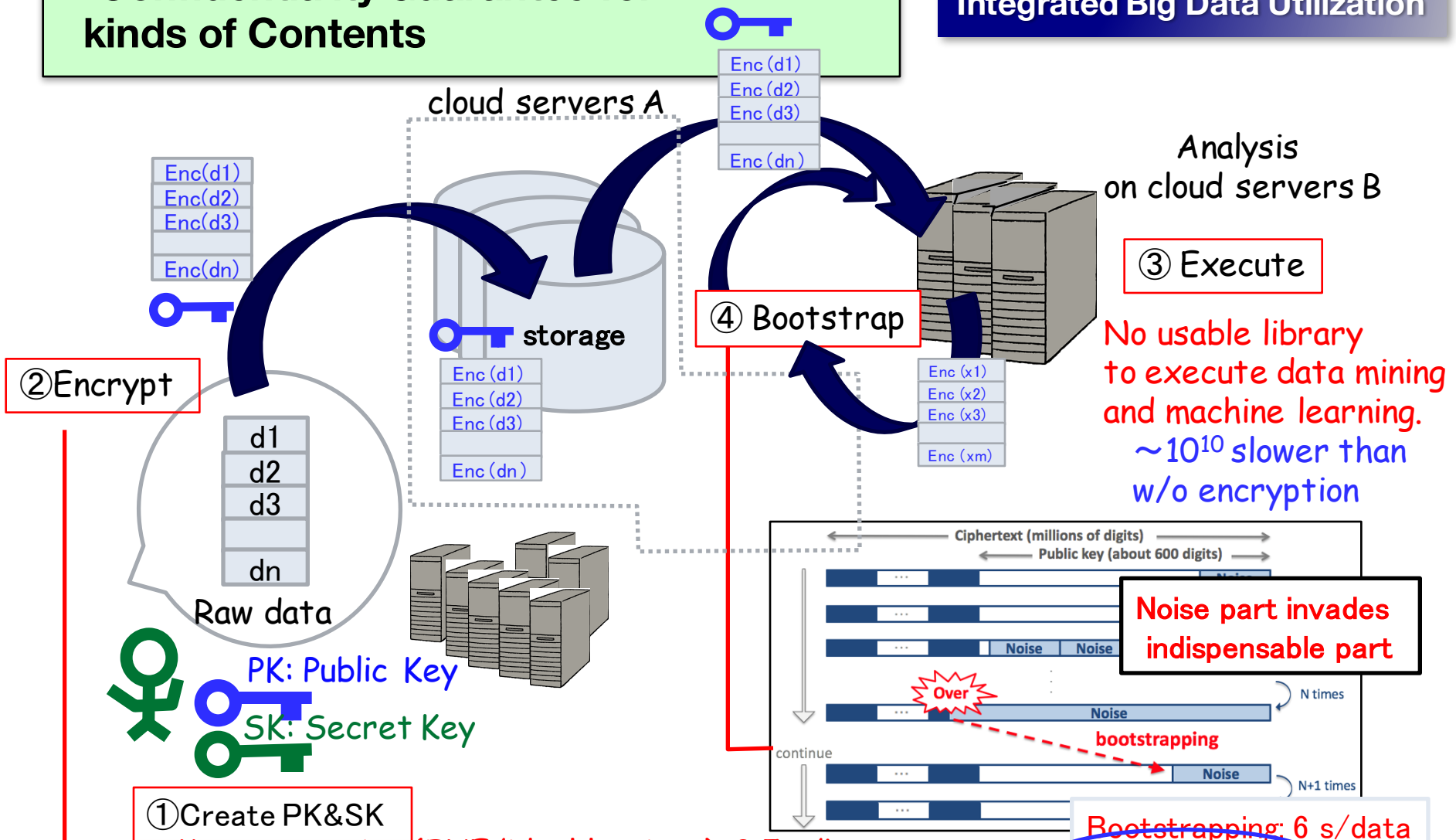
3. Research Goal

HANDLING ALL DATA
WITH ENCRYPTION
THROUGHOUT DATA LIFE CYCLE



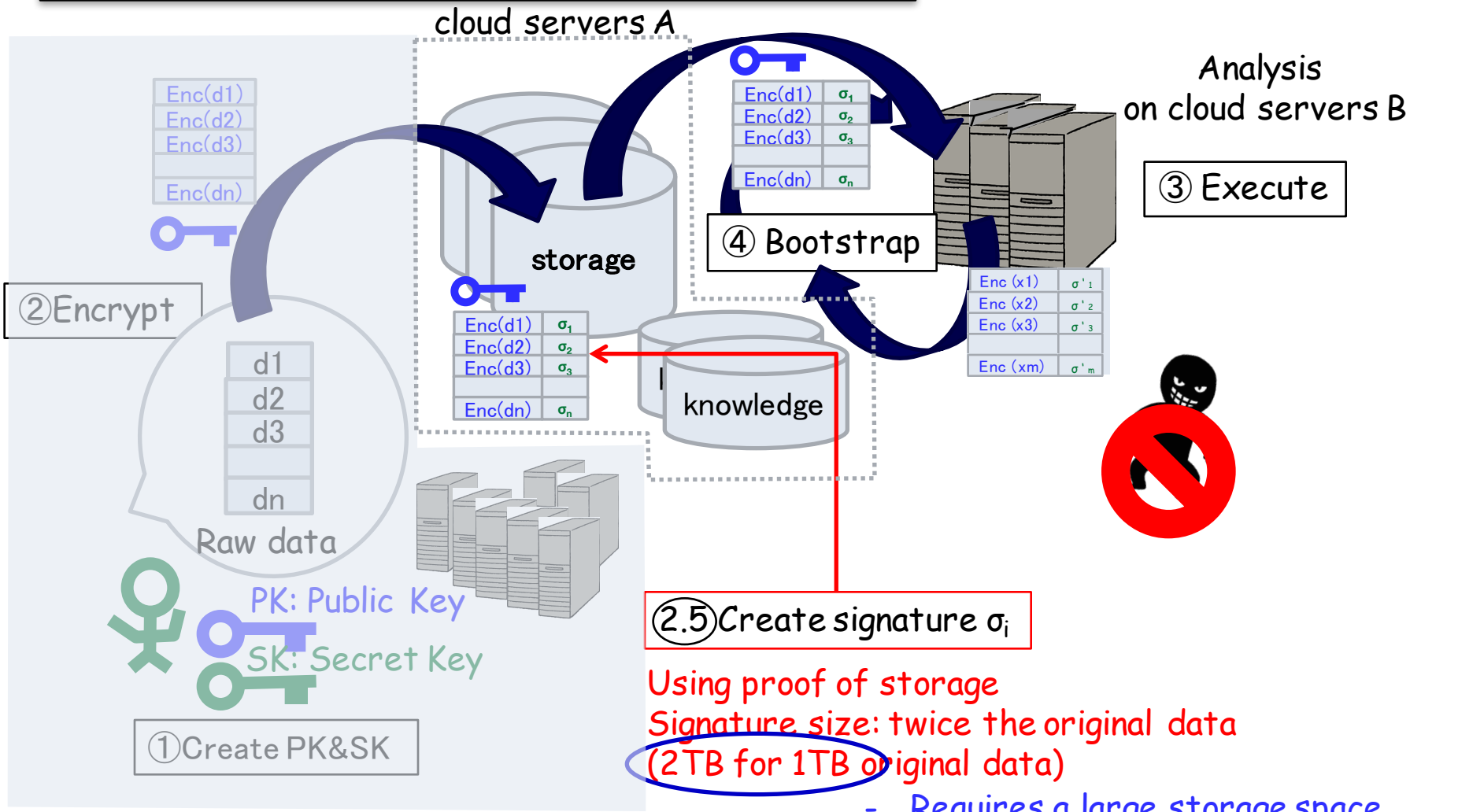
1.Confidentiality Guarantee for kinds of Contents

SD² Platform for Integrated Big Data Utilization



2. Assurance of content source and provenance

SD² Platform for Integrated Big Data Utilization

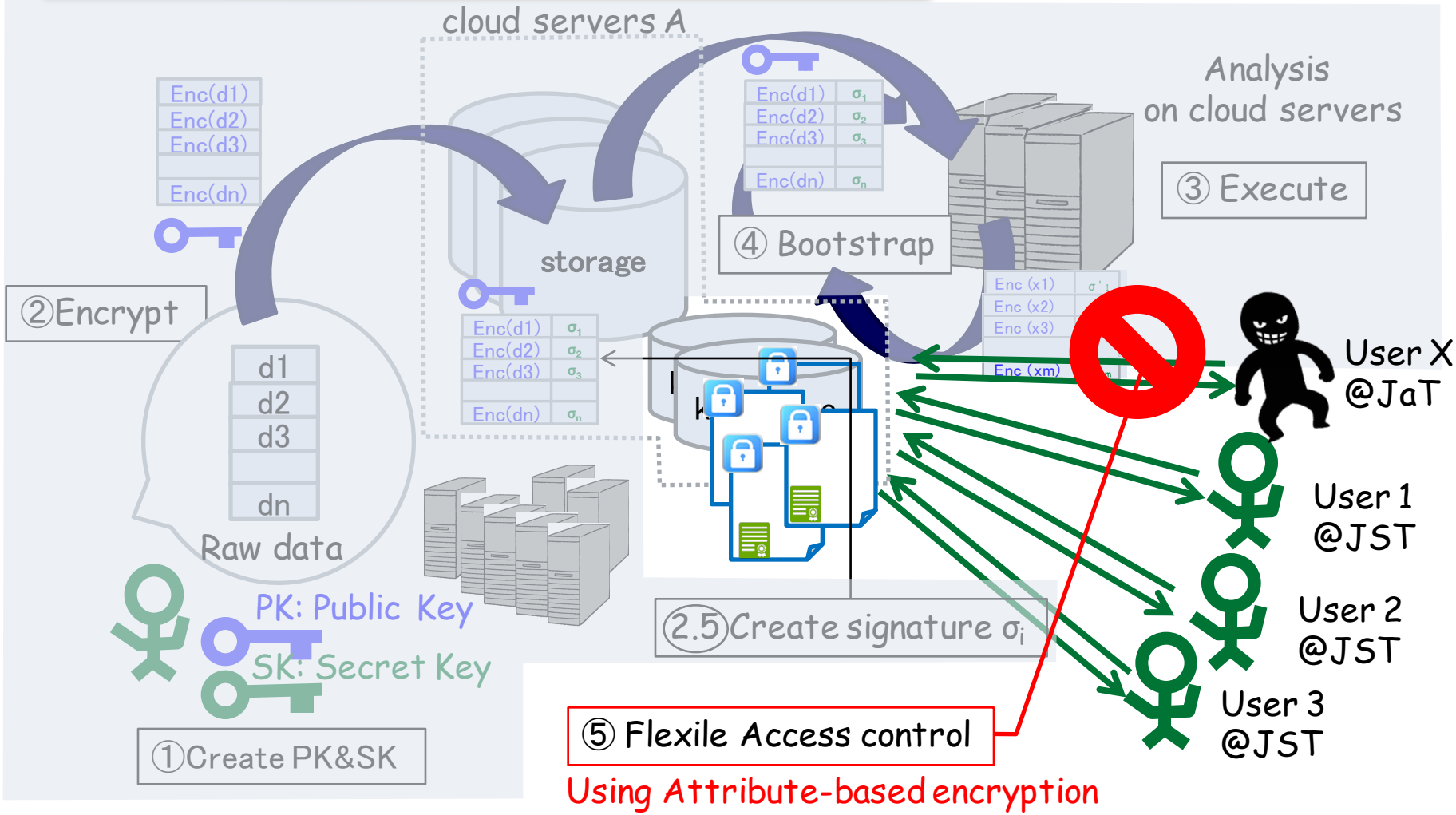


Using proof of storage
Signature size: twice the original data
(2TB for 1TB original data)

- Requires a large storage space
- Re-tagcreation is required depending on a kind of calculation

3. Flexible and assured access control

SD² Platform for Integrated Big Data Utilization



Using Attribute-based encryption
$10^2 \sim 10^3$ speedup is indispensable

- handling "numeric number" as it is, not as character

3. Research Goal

SD² Platform for Integrated Big Data Utilization

BASELINE

CURRENT FHE,
PROOF OF STORAGE,
ATTRIBUTE-BASED ENCRYPTION

GOAL

1,000 TIMES FASTER THAN CURRENT ENCRYPTION METHODS

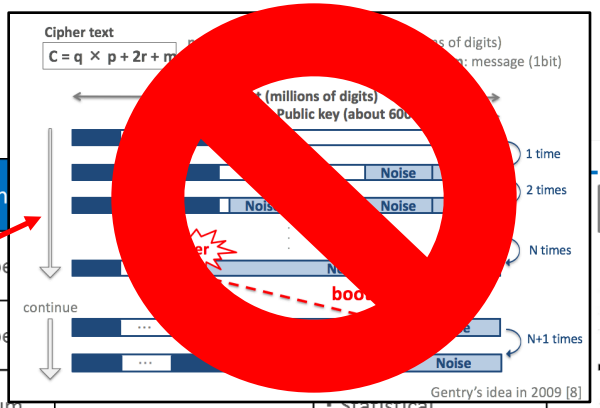
TO SHOW THE EFFECTIVENESS OF OUR PLATFORM WITH EXPERIMENTAL DEMONSTRATION

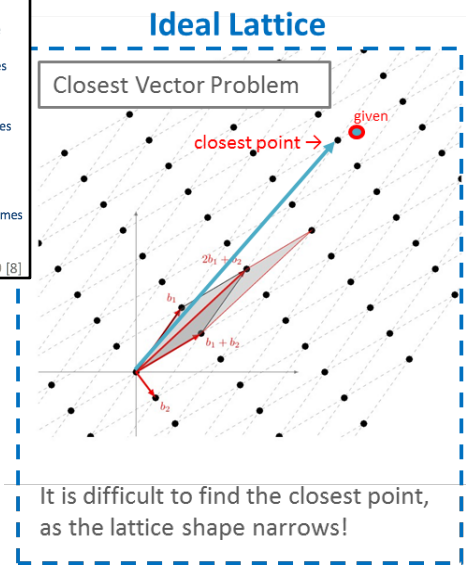
4. Research Strategy

- Parallelizaion
  - (1) For FHE, adopt "Ideal Lattice" whose basic operation is "matrix calculations," to parallelize

- Escape Bootstrapping as possible as we can
  - (2) If SWHE is applicable at some execution, use it

Types	Crypto operation	Efficien		
HE	Only Addition	High Spe		
	Only Multiplication	High Spe		
SWHE (Somewhat HE)	Multiplication for Once + Addition	Medium Speed	Pairing-based	▪ Statistical Computation & Data Analytics
	Multiplication for a few times + Addition	Medium Speed	(i) Ideal Lattices-based (ii) Integer-based (iii) LWE-based Ring-LWE-based	▪ Biometrics Authentication
FHE (Fully HE)	Addition + Multiplication for N times	Very Low Speed		Complex Computation (e.g. Machine Learning)





SD² Platform for
Integrated Big Data Utilization

4. Research Strategy

- Off-load Engine/Stream Processing/Migration
  - Parallelization & adopt FPGA OUR ORIGINAL
  - Strem-processing called Queue Linker platform OUR ORIGINAL
  - Inter-cloud migration OUR ORIGINAL
- I/O tuning / optimization OUR ORIGINAL
- Cache unfriendly tuning of workload
  - Effective use of "memory hierarchy"

	Latency(clock)	Bandwidth
Registers	1	
L1 cache	4+	330GB/s
L2 cache	11+	220GB/s
L3 cache	24+	110GB/s
DRAM	200–400	10–50GB/s
SSD	350,000	200MB/s
HDD	35,000,000+	600MB/s

10^7

Adopting a mechanism to bridge the gap.
Use Memory Appliance to bridge the gap between SSD and HDD
NEW CHALLENGE

- Data Mining Library based on FHE

5. Experiment

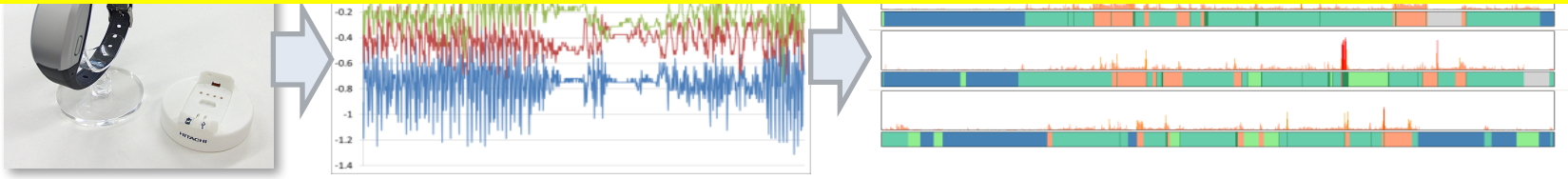
SD² Platform for Integrated Big Data Utilization

Experimental demonstration

→ show the effectiveness of our platform

- **Life Log Analysis (sensor data)**
  - Gathering hundreds of thousands users data (raw 1TB data)

> Proof of Storage
> verifiable delegation of computation



- **Drug Adverse Analysis (text data)**
  - Gathering over 2 million users' drug

> Proof of Storage
> Secure multiparty computation with fully homomorphic encryption
> verifiable delegation of computation
> attribute based encryption

6. Schedule

SD² Platform for Integrated Big Data Utilization

	2015	2016	2017	2018	2019	2020
Outlook			Mid-term evaluation			Final Evaluation
Legal Study		Legal Coordination			Guideline	
Encryption		●Attribute-based encryption ●fully homomorphic encryption ●Proof of storage Computer Architecture Friendly algorithm		Improvement based on LS		
Hierarchy of Storage ・ Parallel & Distributed Computing		Pre-fetch I/O optimization FPGA Platform Const. at Waseda&Ochanomizu		Cont. of Inter-Cloud		
Experi-menal		Life log Over 1,000 logs drug adverse Several drugstores		Improvement Use of platform Over 10,000 logs Use of platform 20 drugstores		

20 times faster

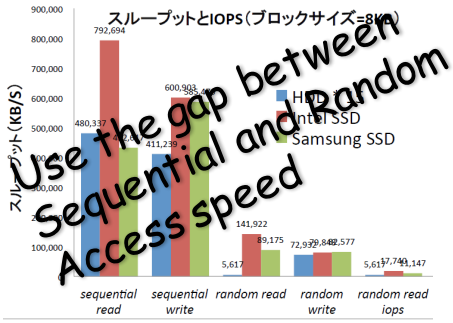
20 times faster

Over 30 times faster

Over 30 times faster

Parallelizing by using "Ideal lattice" base encryption

× = Over 10^3 faster (practical use level)



Use the gap between Sequential and Random Access speed

7. PROGRESS IN 2015FY

SD² Platform for Integrated Big Data Utilization

- Legal Study
  - Studied possible data transfer and analysis under the provision of 2015 Japanese amendment of Act on the protection of personal Information.
- Encryption Algorithm
  - Proposed a theory of FHE for real numbers called FHE4FX.
  - It enables Homomorphic Greater-Than-bit computation.
- Implementation
  - Implemented "Apriori algorithm," 10 times faster than the state-of-the-art method by adopting packing with HElib.
- Platform
  - Analyzed I/O performance where data are on outer/inner zone of platter with large scale data access.
  - Prepared our Cloud Platform between Waseda Univ. and Ochanomizu Univ.

THANK YOU