De-identification and Encryption Technologies to Protect Personal Information

● Koichi Ito ● Jun Kogure ● Takeshi Shimoyama ● Hiroshi Tsuda

The volumes of data in society are anticipated to further increase for reasons such as expansion of big data analysis and future development of the Internet of Things (IoT). There are high hopes that it will be possible to utilize the data gathered in these processes as information linked with individuals (personal information) and create new businesses. Meanwhile, there have been a series of cases of large-scale leakage of personal information due to cyber attacks and internal fraud, along with discontinuance of services that did not pay sufficient attention to privacy. In addition, legal regulations are being tightened as seen in Japan's "My Number (Individual Number)" system, the amendment to Japan's Personal Information Protection Act, and the EU's General Data Protection Regulation. Fujitsu Laboratories has been developing technologies for de-identification and encryption of personal information in accordance with these measures and regulations. This paper describes the de-identification technologies such as *k*-anonymization and encryption technologies such as homomorphic encryption developed by Fujitsu Laboratories for safe handling of personal information.

1. Introduction

The volume of data in society is expected to increase dramatically from here on thanks to the appearance of new enterprises focused on the analysis of big data and the ongoing evolution of the Internet of Things (IoT). In this environment, the use of data in business will become increasingly important. In particular, the use of personal information (personal information such as location information, behavior history, and purchase history) is highly anticipated in fields such as healthcare, finance, and retail sales. At the same time, personal information can also include sensitive information that, if handled carelessly, could severely impair a company's reputation and business operations.

A prime example is the attempt by East Japan Railway (JR East) to sell de-identified passenger-riding history from its Suica (prepaid e-money card) train pass system in July 2013.[1)] Many users voiced their opinions of JR East's actions, saying, for example, "They are selling our data for their own benefit," "The monitoring of passengers' behavior leaves a bad feeling," and "Isn't riding history also personal information?" Such backlash effectively forced JR East to suspend this service offering.

This incident prompted the IT Strategy Headquarters within Japan's Cabinet Office to establish a "Personal Information Review Working Group" to clarify the rules on handling personal information. The result was a draft proposal entitled "Outline of System Reform Concerning the Utilization of Personal Data" formulated in June 2014. The plan was to use this proposal as a basis for amending the Personal Information Protection Act in 2015 and putting it into effect in 2017.

The amended Act has many features. Of these, a major one is that personal information processed into "de-identified information" may be provided to third parties without the consent of the individuals involved under certain restrictions (e.g., re-identification of individuals from the provided data is prohibited). Of course, de-identified information means data that has been processed so that a specific individual cannot be identified. As a result, interest is growing in "de-identification technology" as a means of achieving such processing. This technology is expected to become increasingly important in the years to come.

In this paper, we survey trends in laws and regulations covering personal information in three key regions of the world—the EU, the U.S., and Japan—and describe Japan's Amended Personal Information Protection Act and its key points. We also explain de-identification technology from the viewpoint of achieving both privacy protection and effective use of data and introduce the de-identification technology researched and developed at Fujitsu Laboratories. Finally, we introduce homomorphic encryption technology, also from Fujitsu Laboratories, for fields that cannot be supported solely by de-identification technology.

2. International trends

The following describes trends in laws and regulations covering personal information in the EU, the U.S., and Japan.

1) EU

The "EU Data Protection Directive" was adopted in 1995 to regulate the protection of personal information and to serve as a basis for related legislation in each of the EU member countries. However, the lack of compatibility in regulations among the member countries placed a great burden on business conducted by multinational companies. To rectify this problem, the "EU General Data Protection Regulation" was proposed in January 2012 as a comprehensive regulation covering the entire EU with the aim of passing this proposal through the European Parliament in 2015 and enacting it in 2017. The definition of personal information here is broad (including biological information, location information, and IP addresses). A violation can incur a fine of up to 5% of sales, imposed by data protection agencies of EU member states (supervisory authorities). Regulations have also been established on the transfer of data overseas. All in all, the General Data Protection Regulation provides for rigorous protection of personal information.

2) U.S.

Although there is no unified law covering the protection of personal information, the Federal Trade Commission (FTC) mediates on individual cases of data usage that raise privacy concerns. For example, the Netflix Prize, given for the most accurate method of making movie recommendations on the basis of DVD rental history was discontinued in 2010 after the FTC intervened in the case.[2] Although the U.S. has generally placed more importance on data usage than privacy protection, the submittal of the Consumer Privacy Bill of Rights Act[3] by the White House in March 2015 signals a shift toward more privacy protection.

3) Japan

Despite the enactment of the Personal Information Protection Act in April 2005, the definition of personal information and rules governing its use were still vague, and there was no organization for supervising personal information on the whole.

The above trends in laws and regulations in the EU, U.S., and Japan are compared in **Table 1**. A common trend among them is the strengthening of laws and regulations concerning privacy and efforts at achieving a system in which users can entrust their personal information to organizations without worry.

3. Japan's Amended Personal Information Protection Act and key points

In this section, we describe the key points of Japan's Amended Personal Information Protection Act.[4] A comparison of this act before and after the revision is summarized in **Table 2**.

The main differences are summarized below.

1) Definition of personal information

In addition to the existing definition of personal information, the amended act adds two definitions: "individual identifying codes," which include information on physical features of individuals, and "sensitive information," which includes race, beliefs, and other personal information. In contrast to ordinary personal information, individual identifying codes may not be provided to third parties through de-identification processing, while sensitive information requires an opt in (prior consent) when that information is obtained or provided.

2) Purpose of use

With respect to a change in purpose of use, the word "substantially" in the existing phrase "substantially related to the original purpose of use" was deleted. However, the extent to which this stipulation has actually been eased is unclear.

3) Business operators handling personal information

Limiting of the act to "business operators possessing personal information of 5,000 persons or more" was abolished. Small-scale business operators not

previously covered by the act are now included.

4) Provision of data to third parties

Publically disclosing that de-identified information (the result of processing information to remove any data that could identify a specific individual) will be provided to third parties enables an opt-out option to be provided (third-party provision without individual consent). However, the party to which the data is provided is obligated to prohibit the re-identification of individuals from that data.

5) Provision of data to overseas parties

Before the revision, no particular restrictions were specified for cross-border operations, but the amended act requires an opt-in option to provide data to other countries or overseas organizations recognized as having at least a certain level of data protection.

6) Supervisory authority

As a result of this revision, a Personal Information Protection Commission has been launched as a supervisory authority. It is responsible for creating standards governing the preparation of de-identified information on a field-by-field basis, conducting on-site inspections, and offering advice and guidance to business operators handling personal information.

7) Penalties

The amended law adds new penalties such as a criminal penalty for providing "a personal information database, etc." for the purpose of obtaining a wrongful gain. Amid an overall tightening of regulations, attention has been focused on the provision of de-identified information to third parties. In particular, provision of such information to third parties without individual

Table 1
Comparison of laws and regulations concerning personal data in the EU, U.S., and Japan.

	EU Data Protection Directive (EU)	Consumer Privacy Bill of Rights Act (U.S.)	Amended Personal Information Protection Act (Japan)
Definition of personal data	Includes biological information, location information, and IP addresses	Includes biological information, IP addresses, and insurance/vehicle ID numbers	Includes biological information
Change in purpose of use	Cannot change afterward	A subsequent change requires opt in by explicit consent	Change allowed "within the scope that the new purpose of use is reasonably determined to be related to the original purpose of use"
Sensitive information	Sensitive personal data: religion, political beliefs, race, etc.	Disparate impact: race, religion, sexual orientation, etc.	Sensitive information: race, beliefs, social status, medical history, victim of crime, criminal record, prior history, etc.
Opt out	Not allowed, in principle	Not allowed, in principle (except under FTC approval)	Not allowed for sensitive information and overseas data transfer (opt in, in principle)
Provision to third parties without consent of individual	Not allowed, in principle (individual consent, in principle)	De-identified data is outside the scope of this act	• De-identified information (descriptions that can be used to identify individuals have been removed) may be provided. • Two types of obligations: Disclosure of data provision to third parties and prohibition of re-identification of individuals from provided data
Scope of use	Except in cases of public benefit or valid profit by concerned parties, scope of use may not exceed that when obtaining consent	Use and analysis of out-of-context personal data allowed only under FTC inspection and approval	Three types: opt in, opt out (not allowed for sensitive information, overseas data transfers, etc.), de-identified information
Privacy evaluation	Privacy Impact Assessment (PIA), etc.	Periodic privacy assessments	None in particular
Right of self control of information	Yes (right to rectify, right to erase, right to object)	Yes (right to withdraw consent, right to erase)	Yes (right to erase, right to disclose information, right to rectify, right to suspend provision)
Overseas data transfer	Restrictions exist	Restrictions exist	Restrictions exist
Supervisory authority	European Data Protection Board	State attorneys general, FTC	Personal Information Protection Commission (reorganization of Specific Personal Information Protection Commission)

Table 2
Comparison of Japan's Personal Information Protection Act before and after revision.

	Current Personal Information Protection Act	Amended Personal Information Protection Act
Personal information	• Information that can be used to identify a specific individual by name, date of birth, or other descriptions • Information that can identify a specific individual by comparison with other information	Same as on the left
Individual identifying codes	None	• Codes representing physical features of individuals (fingerprints, facial recognition data) • Codes attached to services, purchases, or documents (passport number, driver's license number, mobile phone number, etc.)
Sensitive information	No stipulations	• Defined as race, beliefs, social status, medical history, victim of crime, criminal record, prior history, etc. • Opt in, in principle
Purpose of use	May be changed "within the scope that the new purpose of use is reasonably determined to be substantially related to the original purpose of use"	May be changed "within the scope that the new purpose of use is reasonably determined to be related to the original purpose of use"
Operators handling personal information	Business operators possessing personal information of 5,000 persons or more	No limitation
Third-party provision	Prior individual consent, in principle exceptions: opt out, joint use, consigned work	De-identified information (which removes any descriptions that could be used to identify an individual) may be provided to third parties provided that the provider publically discloses such third-party provision and the party receiving the data prohibits re-identification.
Overseas provision	Same as domestic provision	Conditions on overseas provision exist
Supervisory authority	Various ministers in charge	Personal Information Protection Commission (reorganization of Specific Personal Information Protection Commission)
Penalties	Administrative penalties levied against business operators	Criminal penalty established for unauthorized database provision (employees)

consent has been approved together with an obligation to publically disclose such third-party provision and an obligation of the party receiving the data to prohibit the re-identification of individuals from that data. The idea behind this revision is to promote the smooth use of personal information.

4. De-identification technology

A number of techniques exist for processing personal information into de-identified information. The Personal Information Protection Commission plans to formulate standards for de-identification techniques on a field-by-field basis.

We must therefore wait for decisions on techniques for generating de-identified information, but at the same time, we consider that technology for processing data to prevent identification of specific individuals (de-identification technology) to be a strong candidate. There has been much research to date on de-identification technology, and a variety of techniques have been developed,[5] each having advantages and disadvantages. In this section, we introduce "pseudonymization" and "*k*-anonymization" as two techniques representative of de-identification technology. We also describe Fujitsu Laboratories' work on these techniques and introduce the Fujitsu NESTGate solution, which incorporates Fujitsu Laboratories' de-identification technology.

1) Pseudonymization

As shown in **Figure 1**, replacing real names with aliases (such as temporary IDs unrelated to the real names) prevents individuals from being identified. One advantage of this technique is the ability to track some aspect of a particular person. For example, it would enable the state of a person's health even under an alias to be tracked over time. A disadvantage, however, is that a combination of attributes such as gender and age that could indirectly identify an individual could be used to identify the individual corresponding to a particular record. For example, a person who knows that

the individual corresponding to "male, 103 years old" in the table of Figure 1 is Mr. Tanaka would be able to identify which record corresponds to Mr. Tanaka.

2) *k*-anonymization

As described above, the use of aliases leaves open the possibility of identifying the individual corresponding to a particular record. In contrast, "*k*-anonymization" avoids this problem by defining attributes that could be indirectly used to identify an individual as quasi-identifiers (QIs) and processes data so that there are at least *k* individuals having the same combination of QI values. This prevents an individual corresponding to a particular record from being identified. An example of *k*-anonymization with *k*=2 is shown in **Figure 2**. Here, age and gender are defined as QIs, and there are at least two individuals in the table having the same combination of values for those attributes. As a result, the record corresponding to Mr. Tanaka cannot be identified.

3) De-identification technology from Fujitsu Laboratories

Fujitsu Laboratories has developed an information gateway (GW)[6] and a *k*-anonymization library as technology for achieving de-identification.

- Information GW

When using a service such as a cloud provided by an outside organization, an information GW enables data to be safely exchanged with an outside organization or service, as shown in **Figure 3**. Specifically, the information GW anonymizes outgoing data so that the outside user is unaware of the original data, and it restores the de-identified data back to its original

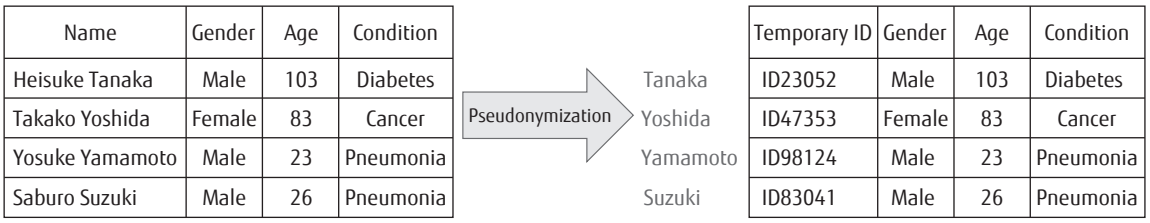


Figure 1
Example of pseudonymization.

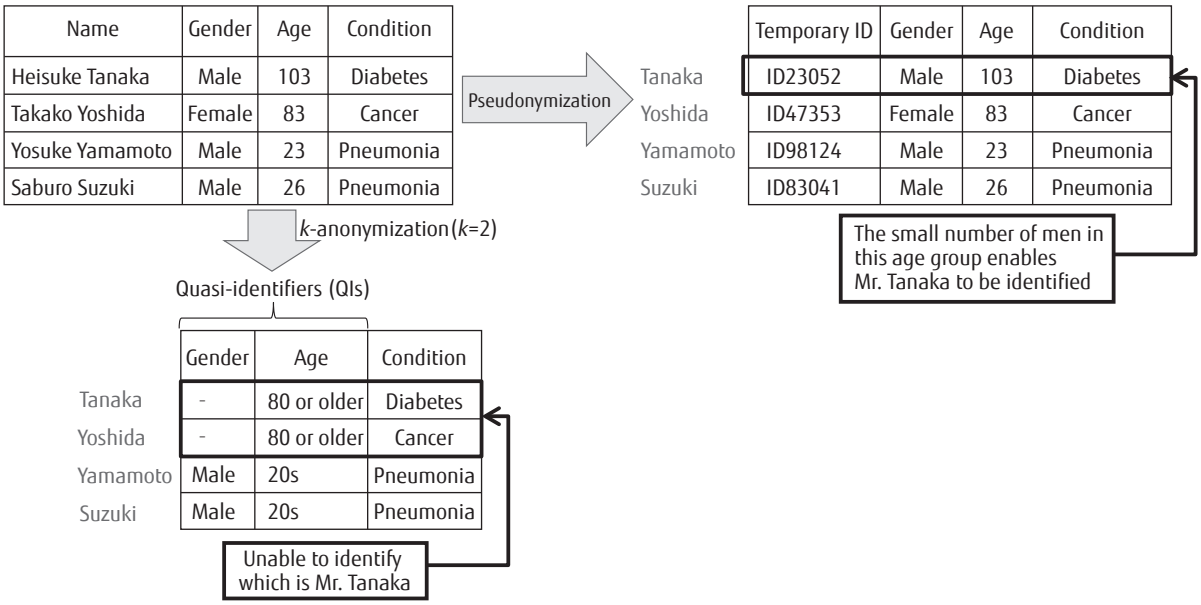


Figure 2
Example of *k*-anonymization (*k*=2).

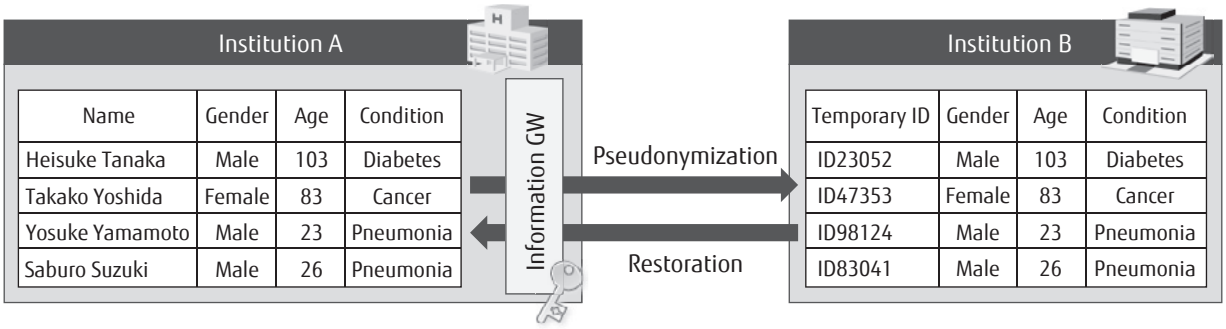


Figure 3
Implementation of information GW.

form when that data is needed for inside use. Using aliases as a de-identification technique enables pseudonymization processing to be performed in batch at the information GW, thereby minimizing changes to the existing system.

- *k*-anonymization library

Fujitsu Laboratories has successfully developed an original *k*-anonymization library having memory-saving, high-speed, and safety features. Referring back to Figure 2, we see that *k*-anonymization groups together closely related records to prevent independent processing of individual records. In other words, *k*-anonymization requires that input data be read into main memory for processing. However, if the amount of input data is excessively large, storing all of it in main memory will not be possible, and processing speed will suffer as a result. To solve this problem, Fujitsu Laboratories developed an original *k*-anonymization algorithm that can make the amount of data to be read into main memory smaller than the total amount of input data (about 1/10–1/100 depending on the type of data). This algorithm enables high-speed *k*-anonymization processing to be achieved even on economical servers and personal computers not having a great amount of on-board memory.

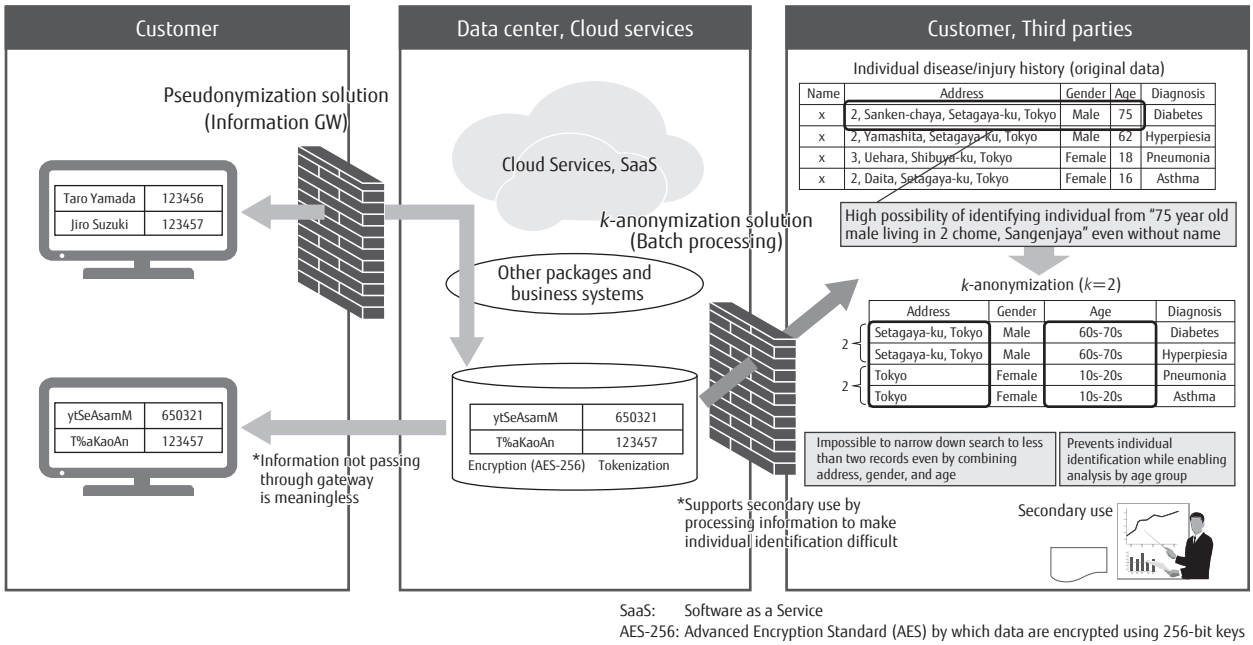
This library also supports de-identification functions more advanced than *k*-anonymization. For example, in the result of the *k*-anonymization process shown in Figure 2, it is impossible to identify whether Mr. Yamamoto's record is the 3rd or 4th record, but it is possible to infer his condition to be pneumonia. The library therefore includes a function called "*k*-presence secrecy,"[7] which provides better privacy protection than *k*-anonymization. This function can prevent the leakage of personal information that cannot be prevented by the *k*-anonymization process.

- NESTGate de-identification solutions

Fujitsu provides a product called NESTGate as a de-identification solution incorporating the above information GW and *k*-anonymization library. NESTGate enables a user to perform various types of anonymization processing including the pseudonymization, *k*-anonymization, and processing conforming to statistical methods guidelines.[8] A conceptual diagram depicting the application of NESTGate is shown in **Figure 4**. Passing data through "pseudonymization solution (information GW)" in NESTGate enables automatic pseudonymization processing in accordance with prescribed policies. In addition, using "*k*-anonymization solution (batch processing)" makes it possible to execute high-speed *k*-anonymization library functions developed by Fujitsu Laboratories and to perform de-identification processing in accordance with statistical methods guidelines. For details on the NESTGate product, please refer to the article "NESTGate for Personal Data Protection by *k*-anonymization Technology" elsewhere in this issue.

5. Homomorphic encryption

As described above, de-identification technology processes the data itself and reduces the amount of information so that a specific individual cannot be identified. This, however, may have a negative effect on the accuracy of data analysis. In genome analysis, for example, anonymizing the genome data prevents those data from being analyzed. In response to this dilemma, Fujitsu Laboratories is moving forward with the development of "homomorphic encryption" as a



SaaS: Software as a Service
AES-256: Advanced Encryption Standard (AES) by which data are encrypted using 256-bit keys

Figure 4
NESTGate anonymization technology solution.

privacy protection technology that can be applied to such fields.

Homomorphic encryption enables mathematical operations such as addition and multiplication to be performed on encrypted data, that is, while input data is still in its concealed state. Theoretically speaking, combining such operations should enable any processing to be performed on encrypted data.

Denoting the data that encrypts x as Enc(x), the basic properties of homomorphic encryption can be summarized by the following expressions.

$$\begin{cases} \mathrm{Enc}(x)+\mathrm{Enc}(y)=\mathrm{Enc}(x+y) \\ \mathrm{Enc}(x)\times\mathrm{Enc}(y)=\mathrm{Enc}(x\times y) \end{cases}$$

Homomorphic encryption can therefore be used to calculate the similarity (Hamming distance) between two sets of vector data in their encrypted form. It also enables concealed searching of character strings, that is, the searching of data and search keys in their encrypted form. These features and the fact that search results can also be concealed means that homomorphic encryption shows promise for application to genetic-information searches in the research and development of new drugs, for example, in which searching by any keyword or combination of keywords is desirable and the occurrence of search hits needs to be concealed.

The development of homomorphic encryption was triggered by the announcement of fully homomorphic encryption by IBM in 2009. At that time, complex encryption processing, large encryption size, and the huge amount of calculation involved in concealment processing made homomorphic encryption impractical. Later, however, Microsoft proposed a method that simplified the generation of parameters, making homomorphic encryption easier to use, and Fujitsu Laboratories developed a parallel processing method at the algorithm level that improved processing performance by several thousand times. These and other improvements to homomorphic encryption have been made throughout the world.

A conceptual diagram of searchable encryption is shown in **Figure 5**. In this example, base sequences of patients are stored in their encrypted form in the cloud. A pharmaceutical company, for example, can then search for patients having a particular partial sequence by passing that partial sequence after encryption to the cloud. In contrast, the use of homomorphic encryption makes it possible to execute search processing

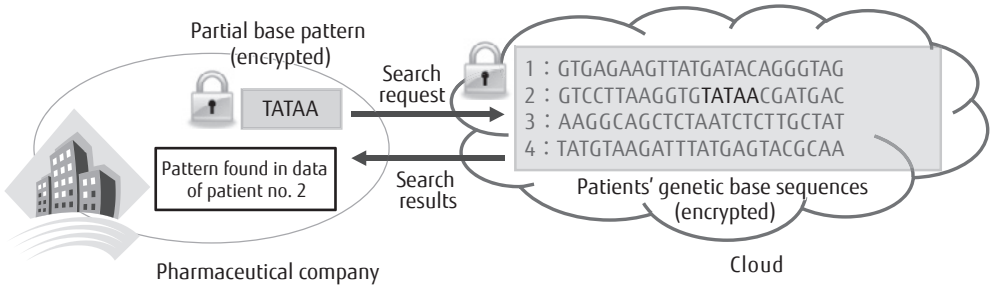


Figure 5
Searchable encryption technology using homomorphic encryption.

while not only protecting the base data of patients but also preventing a third-party from learning about the partial base sequence of concern to a pharmaceutical company. This is important because such a partial base sequence can also be classified as confidential information. Fujitsu Laboratories will continue its R&D efforts in protecting information with the aim of improving processing performance and ease of use, expanding usage scenarios, and achieving practical systems.

6. Conclusion

In this paper, we discussed trends in legal systems governing the use of personal information, described Fujitsu Laboratories' de-identification technology for processing personal information into de-identified information as stipulated in Japan's Amended Personal Information Protection Act, and described homomorphic encryption technology for application to fields that cannot be supported solely by de-identification technology. Fujitsu's NESTGate de-identification technology product enables an enterprise to introduce de-identification technology for safe use of personal information with minimal revision to existing systems.

Going forward, we plan to use actual data to evaluate de-identification and homomorphic encryption techniques for achieving an optimal balance between privacy protection and data use depending on the field and purpose of data use. We are committed to researching and developing privacy protection technologies for achieving the safe use of personal information in accordance with both Japanese and overseas legal systems.

References

1) The Japan News by The Yomiuri Shimbun ChuoOnline: Approve or Disapprove: Selling Suica Travel Records. *http://www.yomiuri.co.jp/adv/chuo/dy/opinion/20140303.html*
2) New York Times: Netflix Cancels Contest After Concerns Are Raised About Privacy. March 12, 2010. *http://www.nytimes.com/2010/03/13/technology/13netflix.html*
3) The White House: Administration Discussion Draft: Consumer Privacy Bill of Rights Act of 2015. March 2015. *https://www.whitehouse.gov/sites/default/files/omb/legislative/letters/cpbr-act-of-2015-discussion-draft.pdf*
4) Cabinet Secretariat Bill Submitted to Diet (189th Ordinary Session): "Law Amending the Act on the Protection of Personal Information and the Act on the Use of Numbers to Identify a Specific Individual in the Administrative Procedure." March 10, 2015 (in Japanese). *http://www.cas.go.jp/jp/houan/150310/siryou4.pdf*
5) WP216: Opinion 05 2014 on Anonymisation Techniques. April 10, 2014. *http://www.cnpd.public.lu/fr/publications/groupe-art29/wp216_en.pdf*
6) H. Tsuda et al.: Inter-Cloud Data Security for Secure Cloud-Based Business Collaborations. FUJITSU Sci. Tech. J., Vol. 48, No. 2, pp. 169–176 (April 2012). *http://www.fujitsu.com/downloads/MAG/vol48-2/paper10.pdf*
7) Y. Yamaoka et al.: *k*-Presence Secrecy: Practical Privacy Model as Extension of *k*-Anonymity. SCIS2015—32nd Symposium on Cryptography and Information Security, 3C4-3, January 2015 (in Japanese).
8) Ministry of Internal Affairs and Communications: Guidelines for the Preparation and Provision of Anonymous Data. March 28, 2011 (in Japanese). *http://www.soumu.go.jp/main_content/000398971.pdf*
9) Y. Morisawa et al.: NESTGate—Realizing Personal Data Protection with *k*-anonymization Technology. FUJITSU Sci. Tech. J., Vol. 52, No. 3 (2016). *http://www.fujitsu.com/global/documents/about/resources/publications/fstj/archives/vol52-3/paper06.pdf*

Koichi Ito
Fujitsu Laboratories Ltd.
Mr. Ito is currently engaged in the research and development of de-identification technology.

Jun Kogure
Fujitsu Laboratories Ltd.
Mr. Kogure is currently engaged in the research and development of digital currency and encryption technology.

Takeshi Shimoyama
Fujitsu Laboratories Ltd.
Mr. Shimoyama is currently engaged in the research of encryption technology.

Hiroshi Tsuda
Fujitsu Laboratories Ltd.
Mr. Tsuda is currently engaged in the research and development of countermeasures to information leakage.