

# AI時代のプライバシー保護 —技術的解決策とは—

早稲田大学 情報理工学科  
教授



山名早人

[yamana@waseda.jp](mailto:yamana@waseda.jp)

<http://www.yama.info.waseda.ac.jp/>

謝辞 本資料の中で紹介する内容の一部は、2015-2021年に実施したJST CREST「ビッグデータ統合利用のためのセキュアなコンテンツ共有・流通基盤の構築」の成果であり、研究室メンバーに感謝する。 <http://www.yama.info.waseda.ac.jp/crest/>

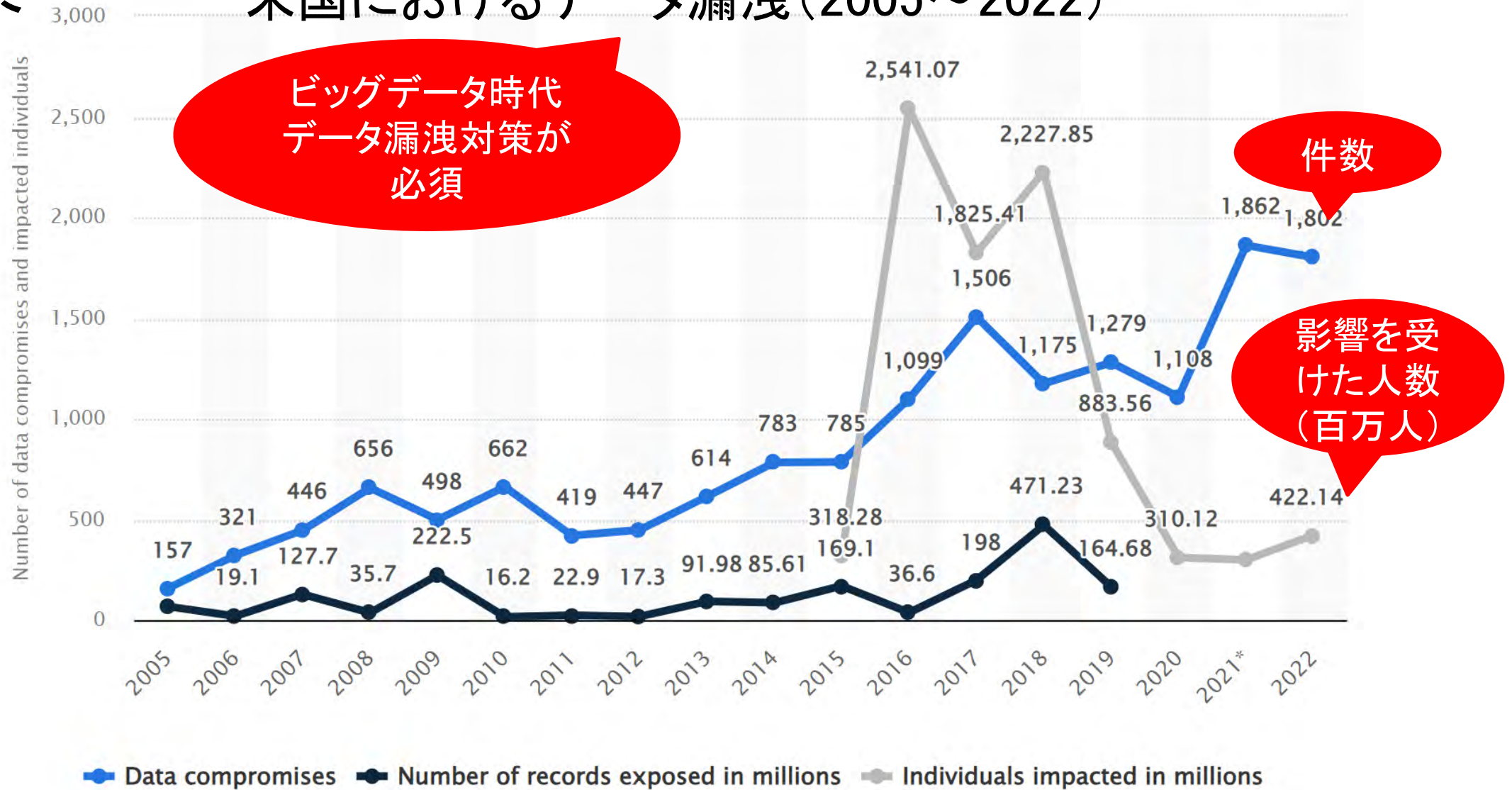
# 目次

1. 背景 - 不正アクセス他
2. 大規模言語モデル(LLM)の基本的な仕組み
3. これからのデータ漏洩防止技術
4. 完全準同型暗号(FHE)
5. FHE関連研究と今後
6. おわりに

# 1. 背景

# 1. 背景

## 米国におけるデータ漏洩(2005~2022)



出典: <https://www.statista.com/statistics/273550/data-breaches-recorded-in-the-united-states-by-number-of-breaches-and-records-exposed/>

# 1. 背景

## 不正アクセス

2020年 第一四半期  
任天堂は、30万件のアカウント情報が漏洩したと報告（名前、電子メールアドレス、生年月日、国籍）

著作権保護のため、以下URLより参照ください。

**データ漏洩の危険性**

# 1. 背景

## 不正アクセス

歴史上最も大規模なデータ漏洩  
2013年 Yahoo! のアカウント  
30億件が影響

著作権保護のため、以下URLより  
参照ください。

データ漏洩の危険性

出典: <https://www.nytimes.com/2017/10/03/technology/yahoo-hack-3-billion-users.html>

# 1. 背景（ある翻訳サービスの事例）

AI学習用データから漏洩

AI翻訳結果に、原文（英語）に存在しない、  
個人名が表示された事例  
（学習データに含まれていたと思われる）

著作権保護のため、原文の英語を隠しています

とはいえ、この[ ] [ ]にととてもよく合っており、興味深い議論につ  
ながるでしょう

その他のコメン  
関連する仕事と[ ]このような種類のおとり効果があるのかを明確にしてくださ  
い。レコメン  
システムの文脈では、最も自然に発生するのは妥協効果だと思わ  
れます。

コメント [ ] [ ]（総合研究所）この論文は、デコイ効果に関連する概  
念を導入する際に、あまり正確に書かれていませんでした。序章はかなりゆるく書  
かれています。予備的な部分にはもう少し詳しいことが書かれていますが、少し繰  
り返しが多いように感じます。

データ漏洩の危険性

# 1. 背景（事例:ChatGPT） ※GPT: Generative Pre-trained Transformer

## ■ ChatGPT（OpenAI）のプライバシーポリシー（2023.6.23版）

**お客様が提供する個人情報**：当社は、お客様が本サービスを利用するためにアカウントを作成した場合、又は当社と連絡を取る場合において、以下のとおり個人情報を取得します。

- アカウント情報：お客様が当社でアカウントを作成する場合、当社は、お客様の氏名、連絡先情報、アカウント認証情報、支払いカード情報、及び取引履歴を含む、お客様のアカウントに関連する情報（以下、総称して「アカウント情報」といいます。）を取得します。
- ユーザーコンテンツ：お客様が本サービスを利用する際、当社は、お客様が本サービスに提供する入力情報、ファイル、又はフィードバックに含まれる個人情報（以下、「コンテンツ」といいます。）を取得します。

**入力データ漏洩の危険性**

出典：<https://openai.com/ja/policies/privacy-policy>



# 1. 背景（事例:ChatGPT）

**統計情報又は個人が特定されない情報**：当社は、お客様を特定することができないようにするために個人情報を集計又は非識別化し、本サービスの有効性の分析、本サービスの改善及び機能の追加、調査の遂行及びその他の同様の目的のためにこれらの情報を利用することがあります。また、当社は、随時、本サービスのユーザーの一般的な行動及び特性を分析し、一般的なユーザー統計のような統計情報を第三者と共有し、そのような統計情報を公表し、又はそのような統計情報を一般に利用可能にすることがあります。当社は、本サービス、クッキー及び本プライバシーポリシーに記載されているその他の手段により、統計情報を収集する場合があります。当社は、非識別化された情報を匿名又は非識別化された形式で維持又は利用し、法律で要求されない限り、当該情報によって再度個人が識別されるような試みを行いません。

上述のとおり、当社は、ChatGPTを動かすモデルをトレーニングするためなど、本サービスを改善するために、お客様から提供されたコンテンツを利用することがあります。当社のモデルをトレーニングするためにお客様のコンテンツを利用することにつきオプトアウトする方法については、こちらを参照してください。

# 1. 背景（事例:ChatGPT）

ChatGPTのトレーニングに個人情報は使われるのでしょうか？

インターネット上の大量のデータは人に関係するものであるため、当社の学習情報には、意図せずに個人情報が含まれる場合があります。当社は、当社のモデルのトレーニングのために個人情報を積極的に求めることはしません。

当社は、モデルが言語について学び、言語を理解し反応する方法を学ぶためにのみ、学習情報を使用します。当社は、学習情報に含まれるいかなる個人情報も、個人に関するプロフィールを構築するため、当該個人に連絡したり、宣伝したり、何かを販売しようとするため、又は情報そのものを販売するために使用することはありませんし、今後もそのようなことはありません。

当社のモデルは、名前や住所といったものが言語や文章の中でどのように適合するかを理解したり、著名人や公人について学習したりするために、個人情報から学習を行う場合があります。これにより、当社のモデルは、より関連性の高い回答を提供できるようになります。

出典：<https://help.openai.com/en/articles/8055555-chatgpt>と当社の言語モデルはどのように開発されたか

## 2. 大規模言語モデルの基本的な仕組み

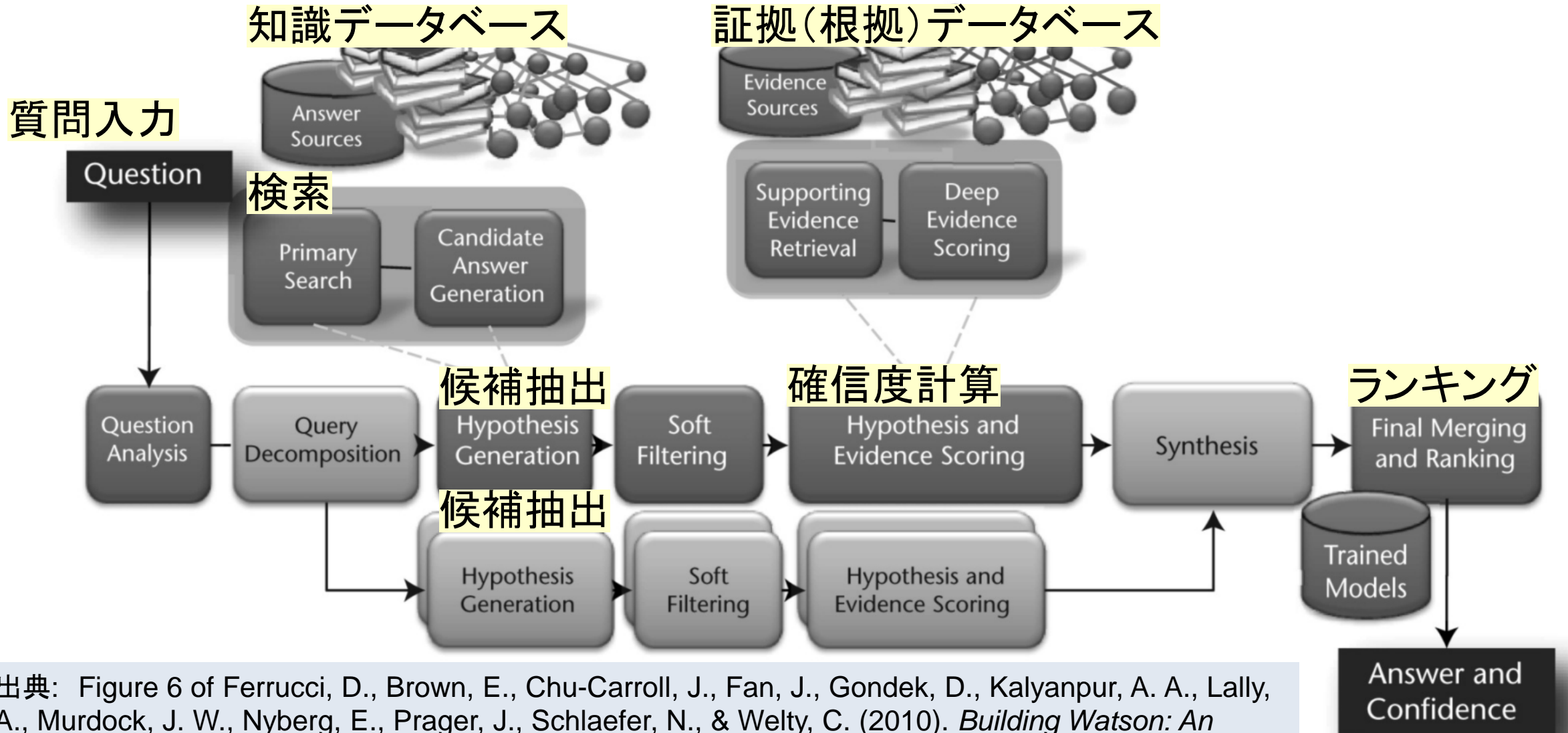
## 2. 1 大規模言語モデル(LLM)登場前 (DeepQA(Watson))

- 2011年2月 IBM Watson (DeepQA)が米国で有名なクイズ番組「*Jeopardy!*」で歴代チャンピオンを破る。



出典: [https://commons.wikimedia.org/wiki/File:IBM\\_Watson\\_w\\_Jeopardy.jpg](https://commons.wikimedia.org/wiki/File:IBM_Watson_w_Jeopardy.jpg)

## 2. 1 大規模言語モデル(LLM)登場前(DeepQA(Watson))



出典: Figure 6 of Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A. A., Lally, A., Murdock, J. W., Nyberg, E., Prager, J., Schlaefel, N., & Welty, C. (2010). *Building Watson: An Overview of the DeepQA Project*. *AI Magazine*, 31(3), 59-79. <https://doi.org/10.1609/aimag.v31i3.2303>

## 2. 2 LLMの基本的な仕組み(概要)

### 基本的な仕組み

※膨大な文章を学習(1,750GB/ChatGPT3.5≒新聞1万年分のテキスト)

雨にも負けず 風にも負けず

雨にも負けず 風にも負けず

←形態素(品詞)に分割 (mono-gram)

雨にも負けず 風にも負けず

←2品詞毎にスライド (bi-gram)

長いほど精度向上

←n品詞毎にスライド (n-gram)

例

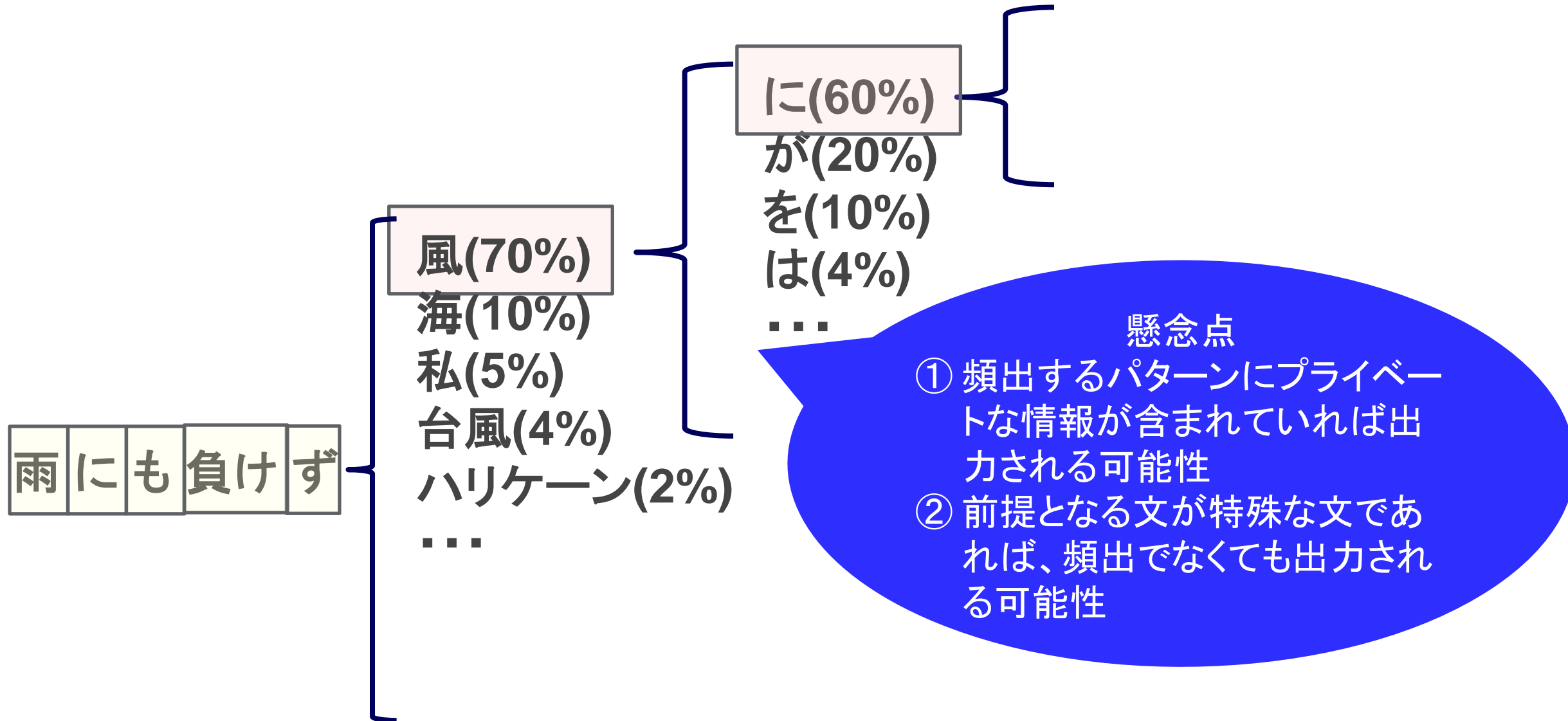
雨にも負けず

が入力された時、次に出現する単語を予測

? ? ? ?

入力に沿った人間らしい出力が得られる

## 2. 2 LLMの基本的な仕組み(概要)





## 2. 2 LLMの基本的な仕組み(学習)

学習の仕組み(基本):

次の単語を予測(自己回帰)

雨にも負けず  雪にも夏の暑さにも負けぬ

マスク部分を予測(穴埋め)

雨にも負けず (?) 夏の (?) にも負けぬ  
 雪にも  暑さ

LLMでは、自己回帰、穴埋め等により  
事前学習を行っている。

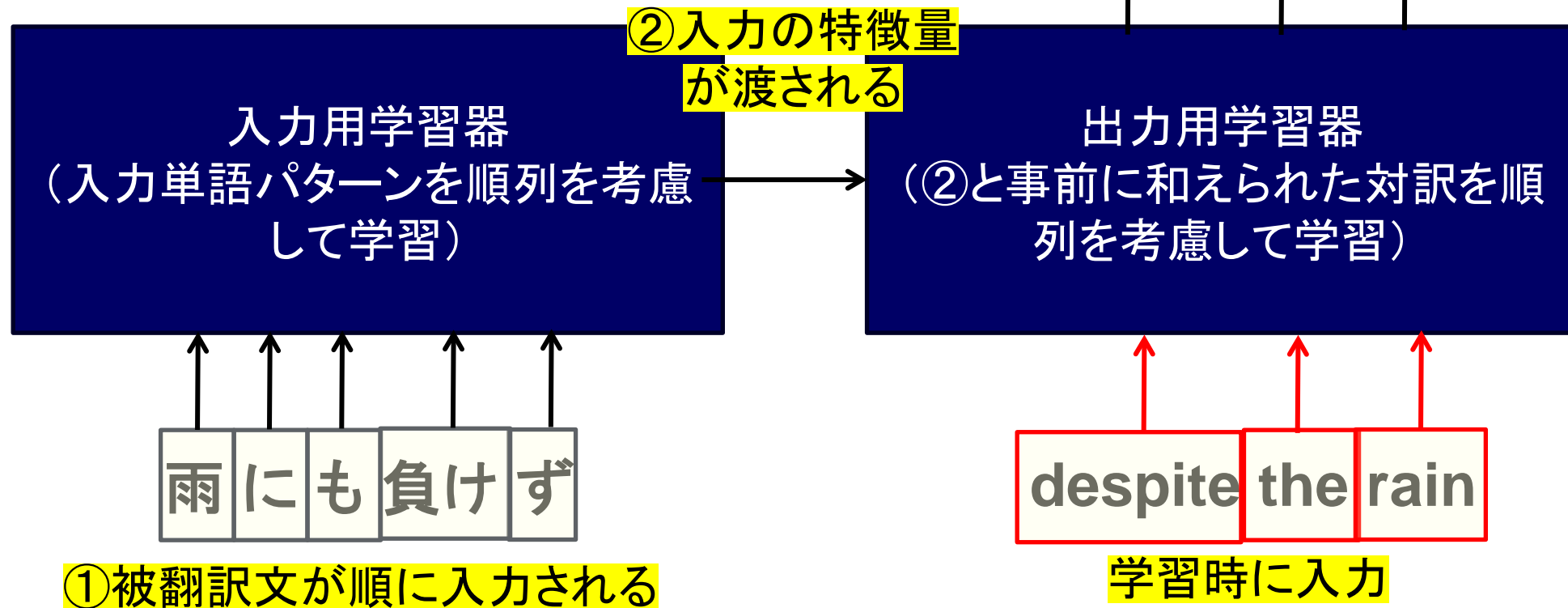


## 2. 2 LLMの基本的な仕組み(翻訳の事例)

学習の仕組み(例: 翻訳)

シーケンスtoシーケンスの学習

入力と出力の関係を学習させる



## 2. 2 LLMの基本的な仕組み(多様な質問への対応)

学習の仕組み(汎化)

プロンプトを使った学習

日本語を英語に  
翻訳してください。

+

雨にも負けず

LLM

②学習した結果を出力

despite the rain

①プロンプト(指示文)と一緒に  
入力する

プロンプトと一緒に学習する  
ことで、様々な質問への回答  
が可能に

## 2. 3 LLMにおけるプライバシー注意点

### プライバシーに関連した問題

- ① 超大規模な文章から、様々な単語をパターンとして学習

学習に利用された単語の羅列は、LLM利用時、出力される可能性がある

プライバシーに関わる単語(単語のシーケンス)を  
学習させない(入力しない)!

- ② 利用者が入力したクエリから解答を生成(サーチエンジン等も同様)

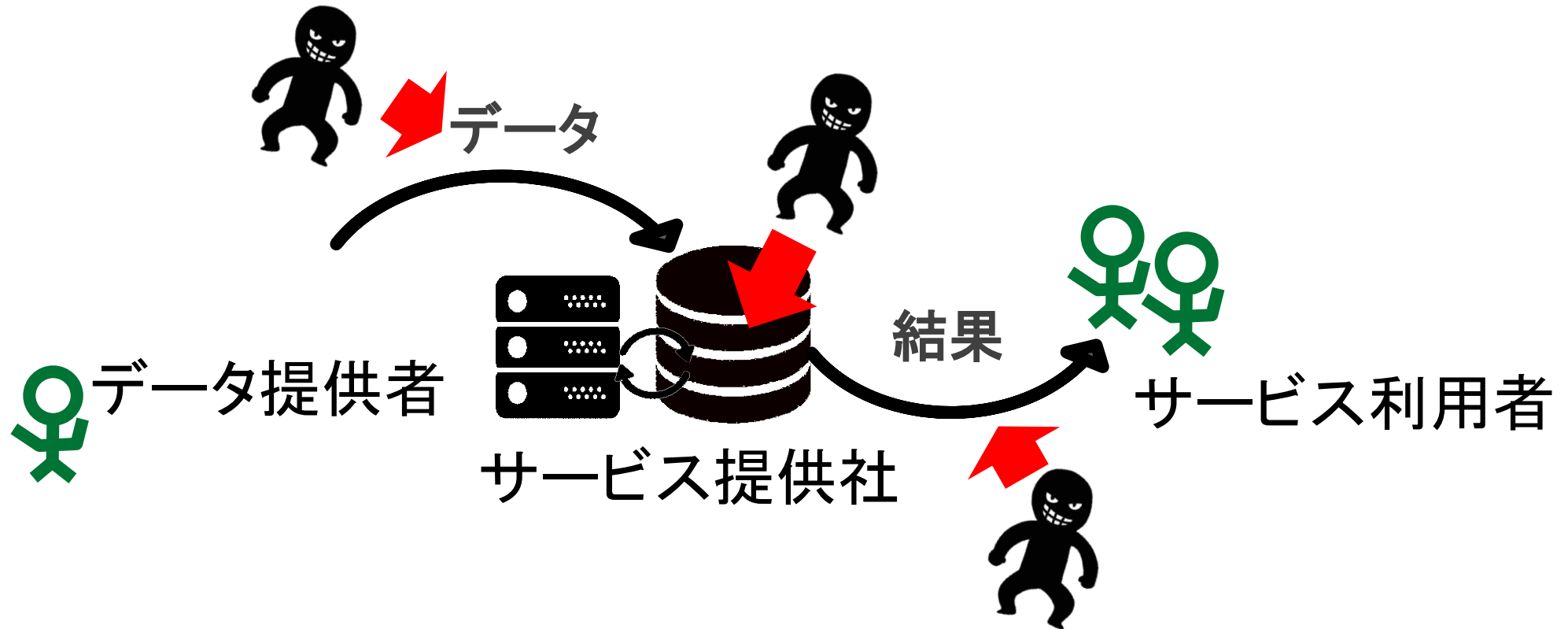
利用者が入力した文章はサービス事業者に漏洩している

利用者が「何をしようとしているのか」が  
サービス事業者には筒抜け

### 3. これからのデータ漏洩防止技術 —オンラインサービス事業者を例にとって—

### 3.1 どこで何が漏れるか？

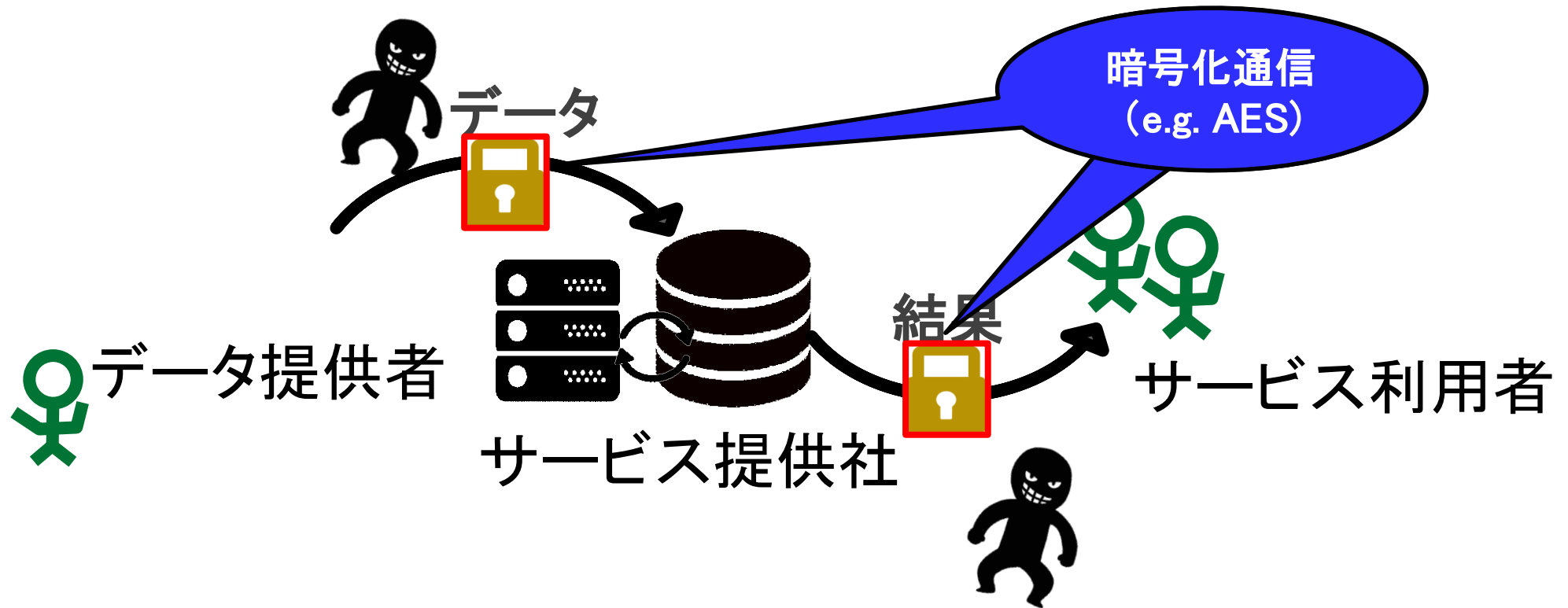
多くのデータはクラウドに保存され、クラウドで処理(LLMを含む)される時代へ



データ漏洩の危険性

## 3. 2 現在のデータ保護対象（一般的な保護）

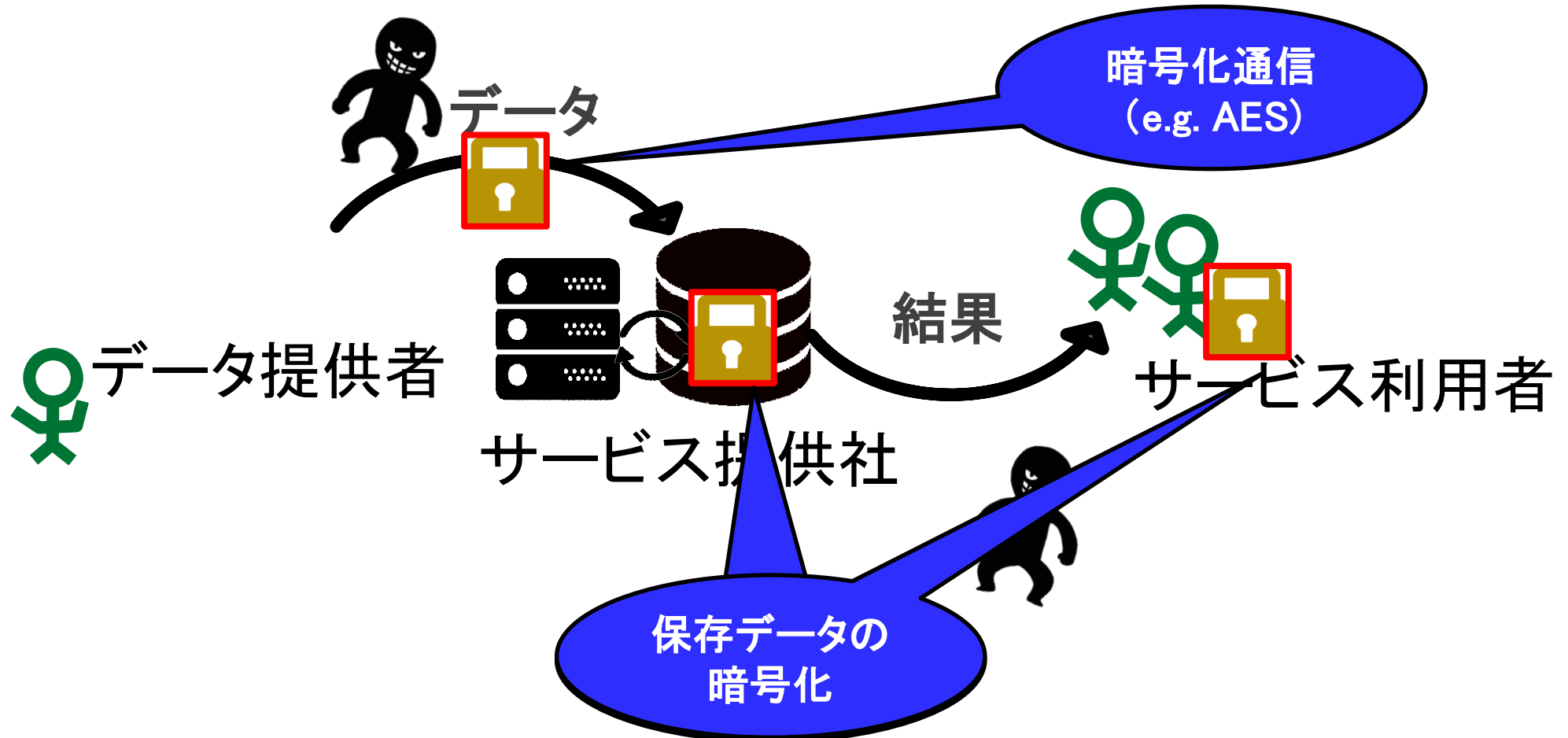
### よくあるデータ漏洩対策－通信路の暗号化（1）



データ漏洩の危険性

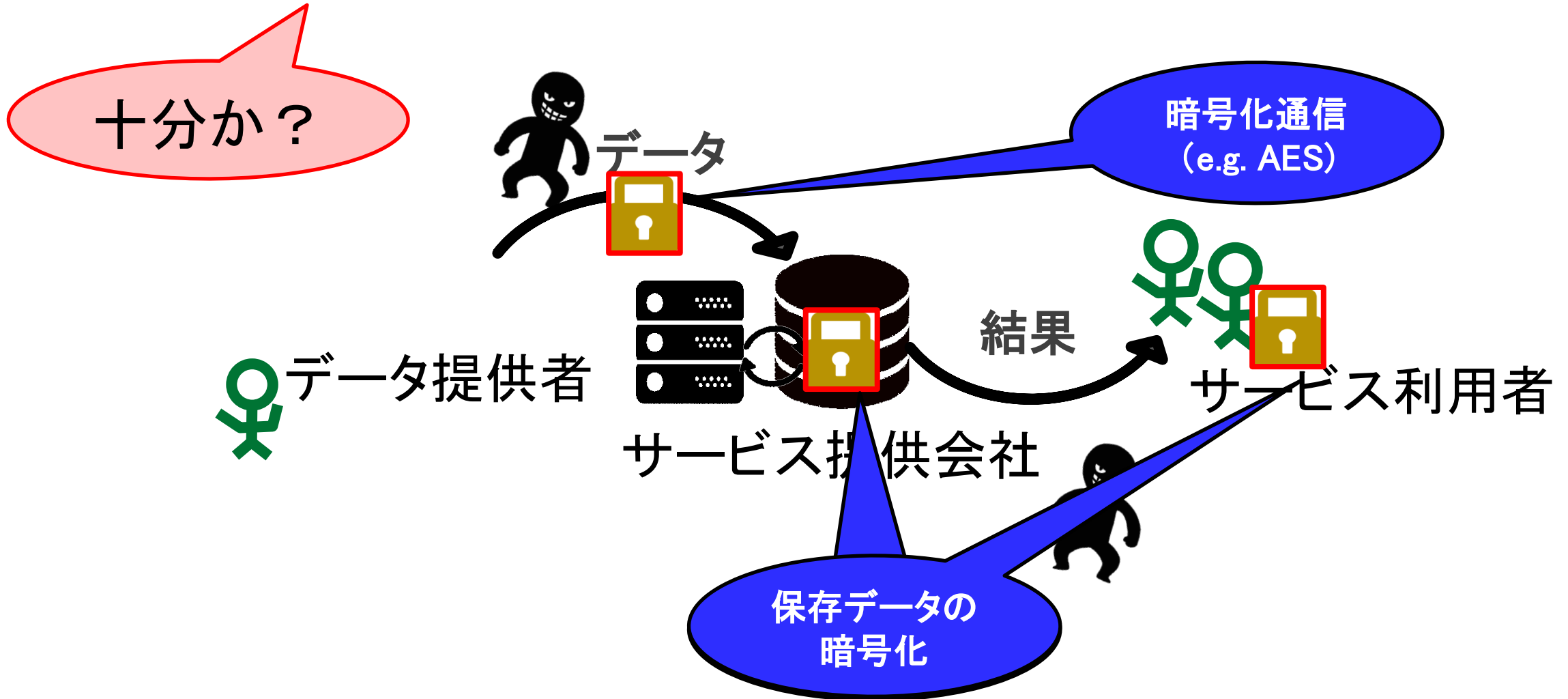
### 3. 2 現在のデータ保護対象(高度な保護)

#### よくあるデータ漏洩対策—通信路の暗号化(2)



### 3. 2 現在のデータ保護対象(高度な保護)

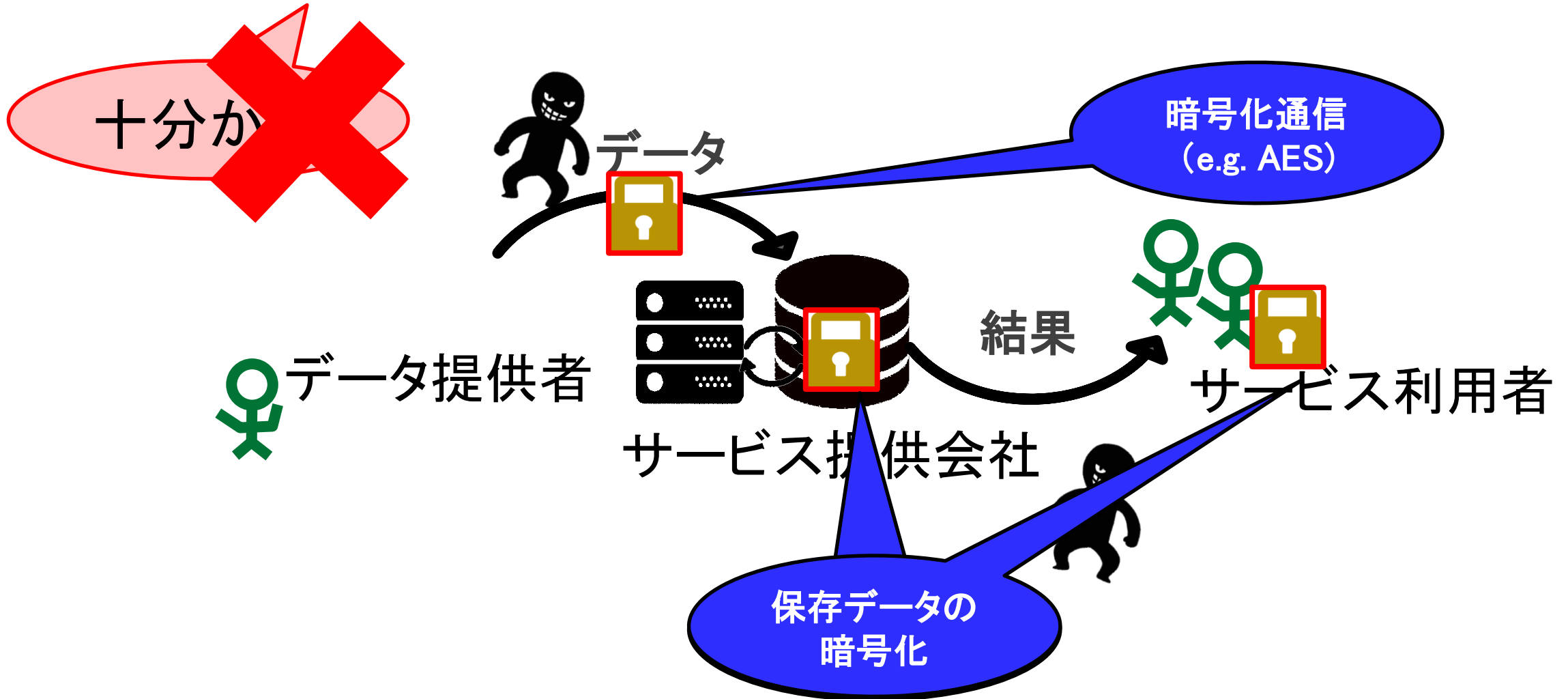
#### よくあるデータ漏洩対策—通信路の暗号化(2)



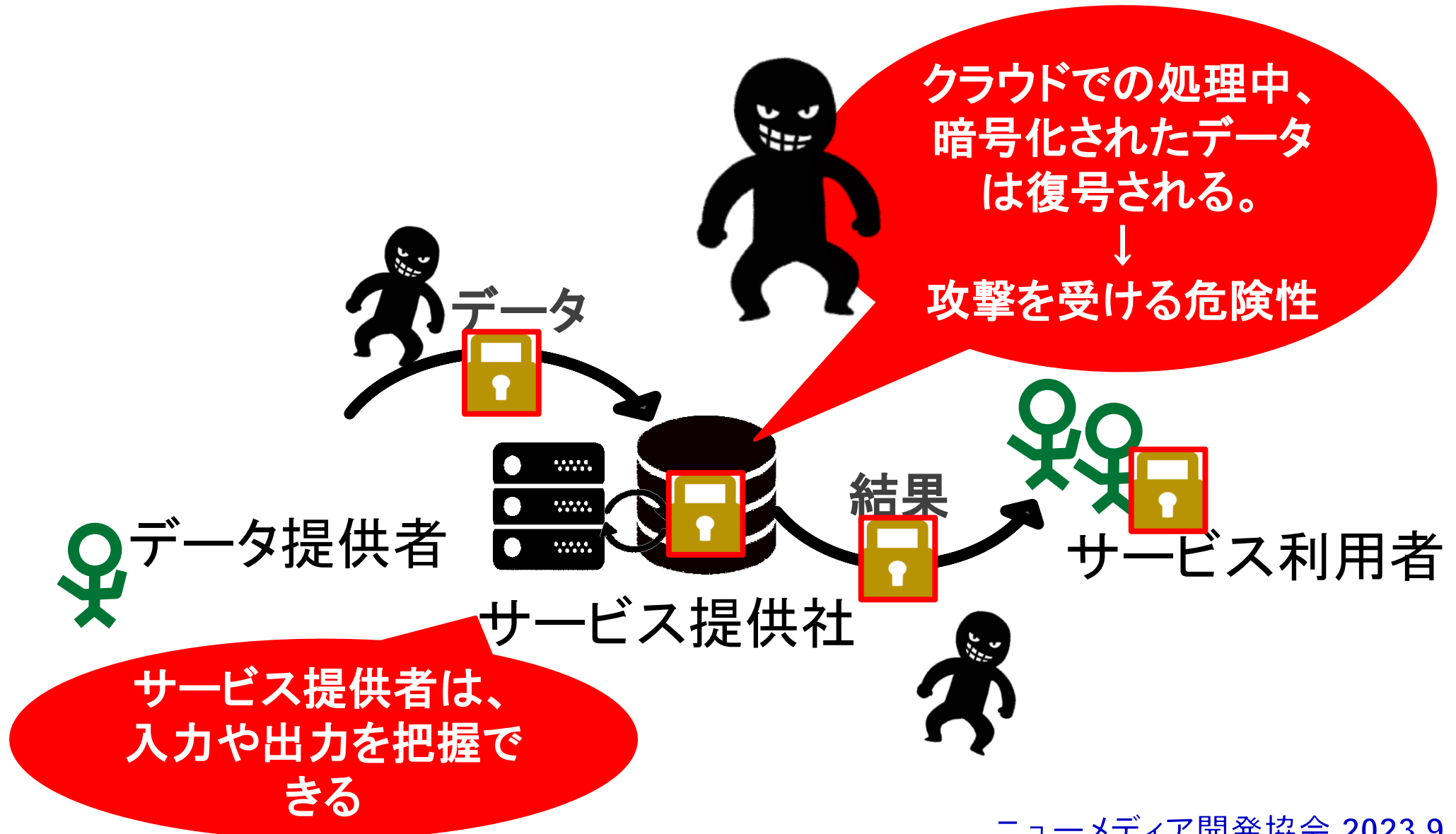


### 3. 2 現在のデータ保護対象(高度な保護)

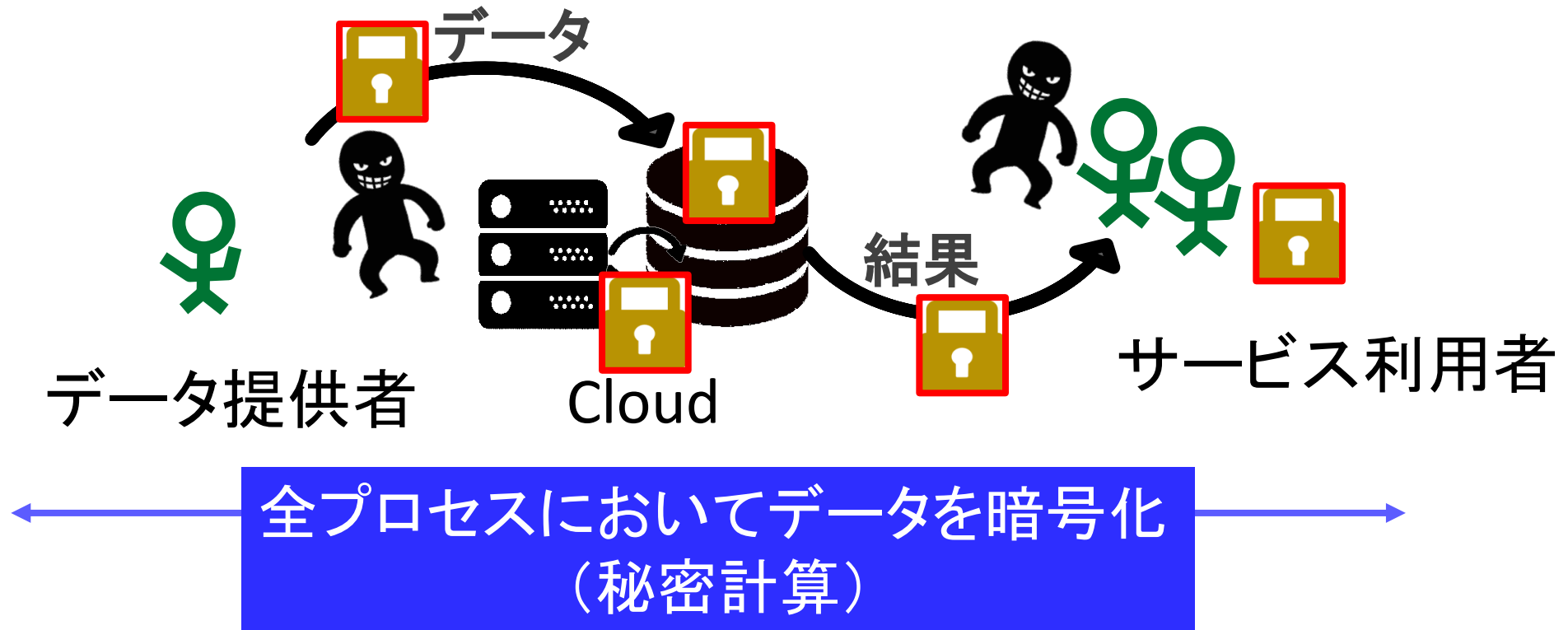
#### よくあるデータ漏洩対策—通信路の暗号化(2)



### 3. 3 現在のデータ保護対象（高度な保護）の弱点



### 3.4 対策



## 4. 完全準同型暗号

## 4. 1 完全準同型暗号 (FHE: Fully Homomorphic Encryption)

- 任意の「掛け算」「足し算」が暗号のままできる(耐量子暗号)

$$\begin{aligned}x + y &= \text{復号}(\text{暗号}(x) + \text{暗号}(y)) \\x \times y &= \text{復号}(\text{暗号}(x) \times \text{暗号}(y))\end{aligned}$$

- 2009年 IBM Watson研究所 Craig Gentry氏により提案 (Stanford大Ph.D論文)

### ■ 特徴

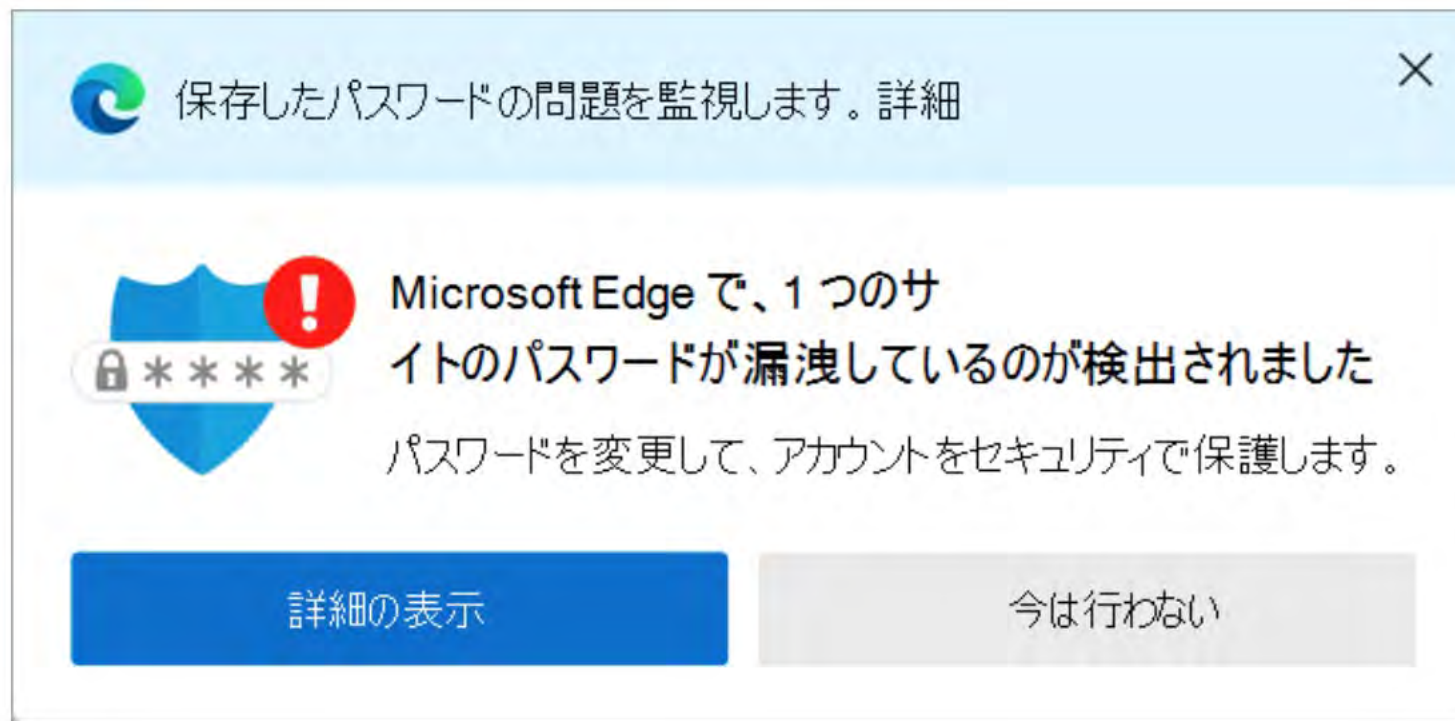
- **利点**: 暗号のまま加算、乗算可
- **利点**: 同一の値を暗号化しても同一の暗号とはならない(ノイズを含むため)
- **欠点**: 平文に比較して6桁~8桁実行速度が遅い(暗号文が大きい(数MB))
- **欠点**: 条件分岐が利用できない
- **欠点**: 計算中にノイズが大きくなるので計算中にノイズリセットが必要(数秒)

## 4. 2 主要なFHEライブラリ

- Helib 2.1 (IBM Research) released in 2013
  - BGV(RLWE version), CKKS
  - Bootstrapping 対応
- SEAL 4.1.1 (Microsoft) released in 2018
  - BFV, CKKS
  - Bootstrapping 未対応
- PALISADE 1.11 (NJIT) released in 2017
  - BGV, BFV, CKKS, FHEW, TFHE
  - Bootstrapping 対応
- OpenFHE 1.1.1 (OpenFHE community) released in 2022
  - BGV, BFV, CKKS, TFHE
  - Bootstrapping 対応

## 4. 3 既に実用化されているサービス

- Webブラウザでのパスワード漏洩チェックで実用化(Microsoft, Google)
- ブラウザ保存パスワードを暗号化しサーバに送信、暗号のまま漏洩チェック



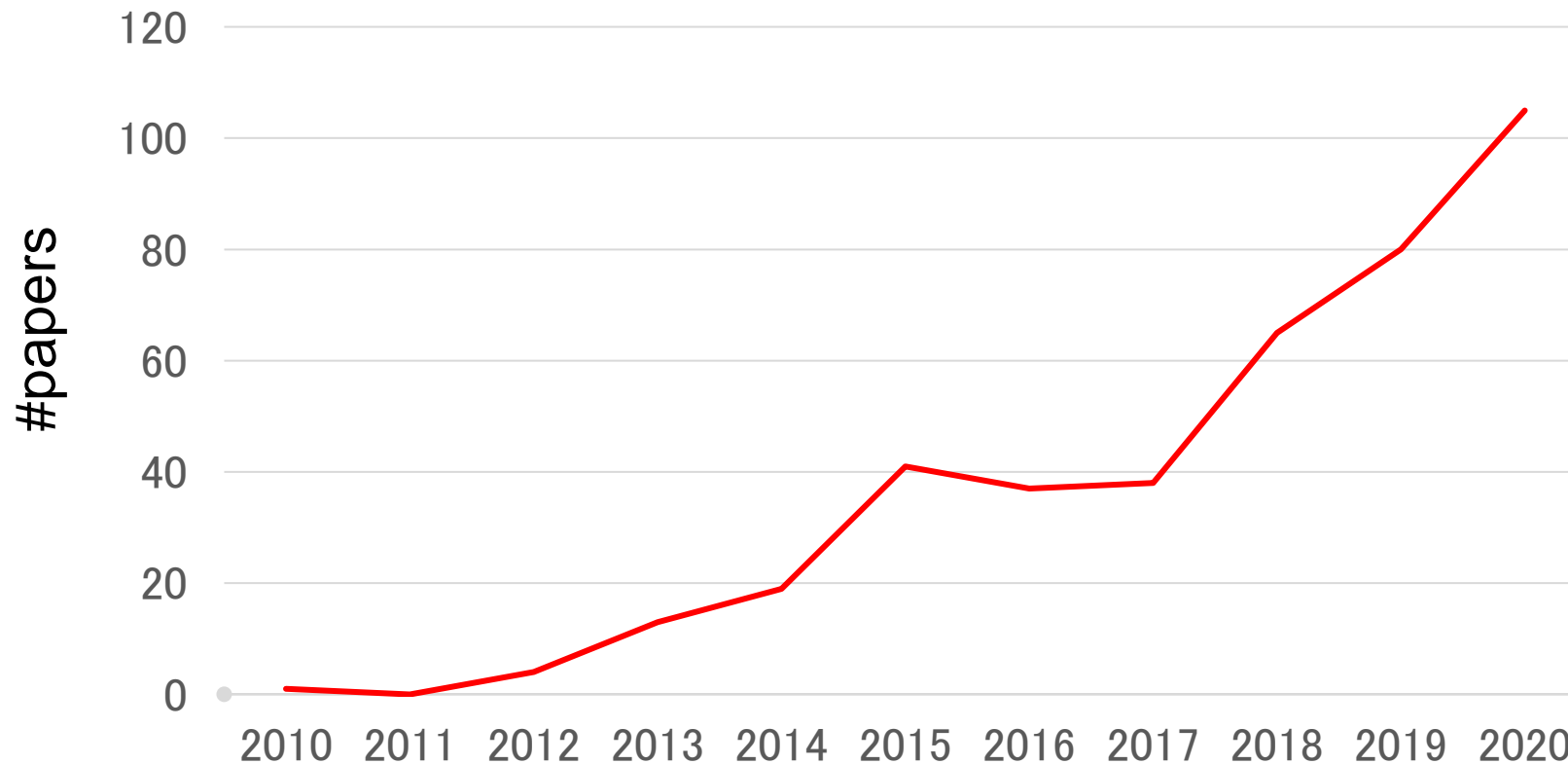
## 5. FHE関連研究と今後



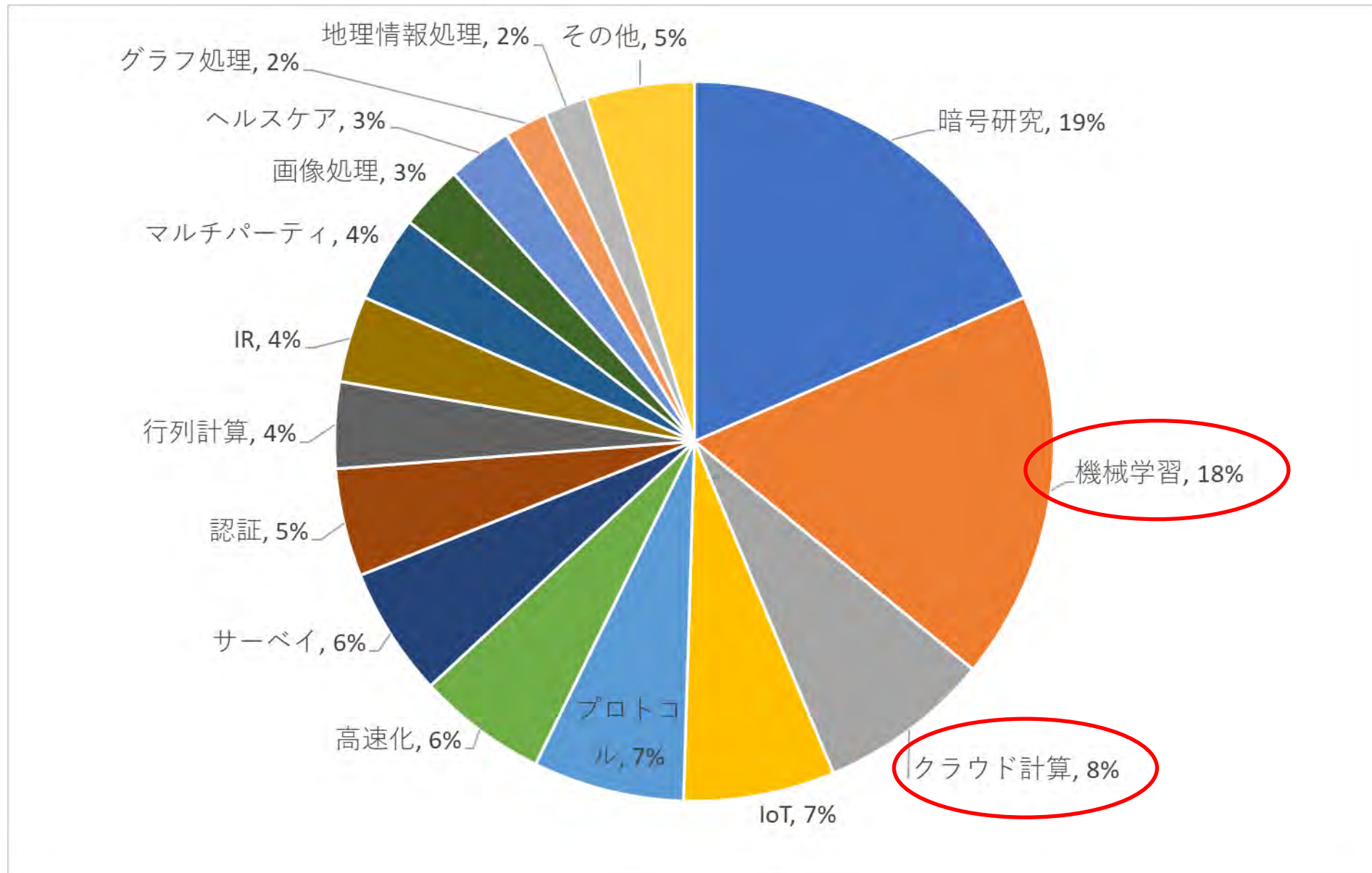
# 5. 1 FHE関連研究

## ■ FHE関連論文数の推移

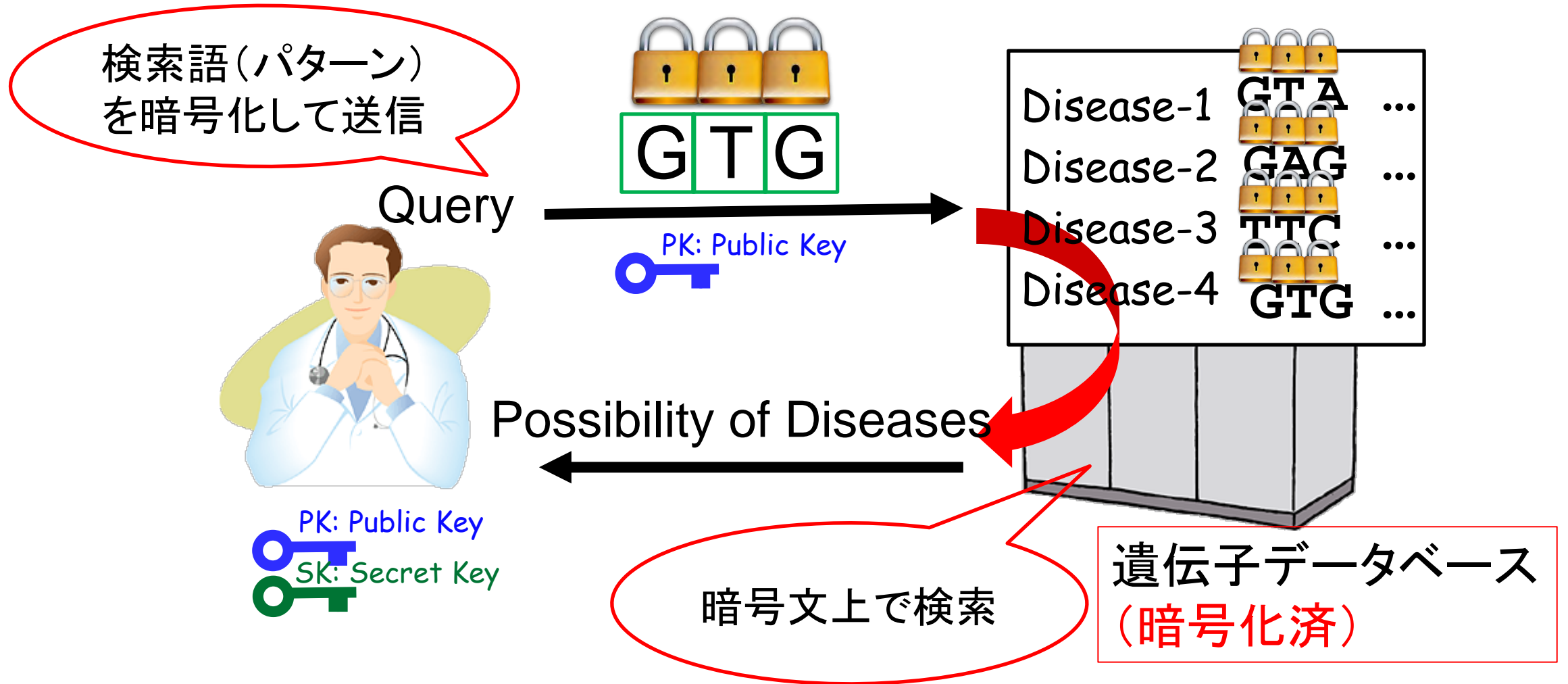
「Fully Homomorphic Encryption」で検索 (Web of Scienceを利用)



## 5.2 研究分野

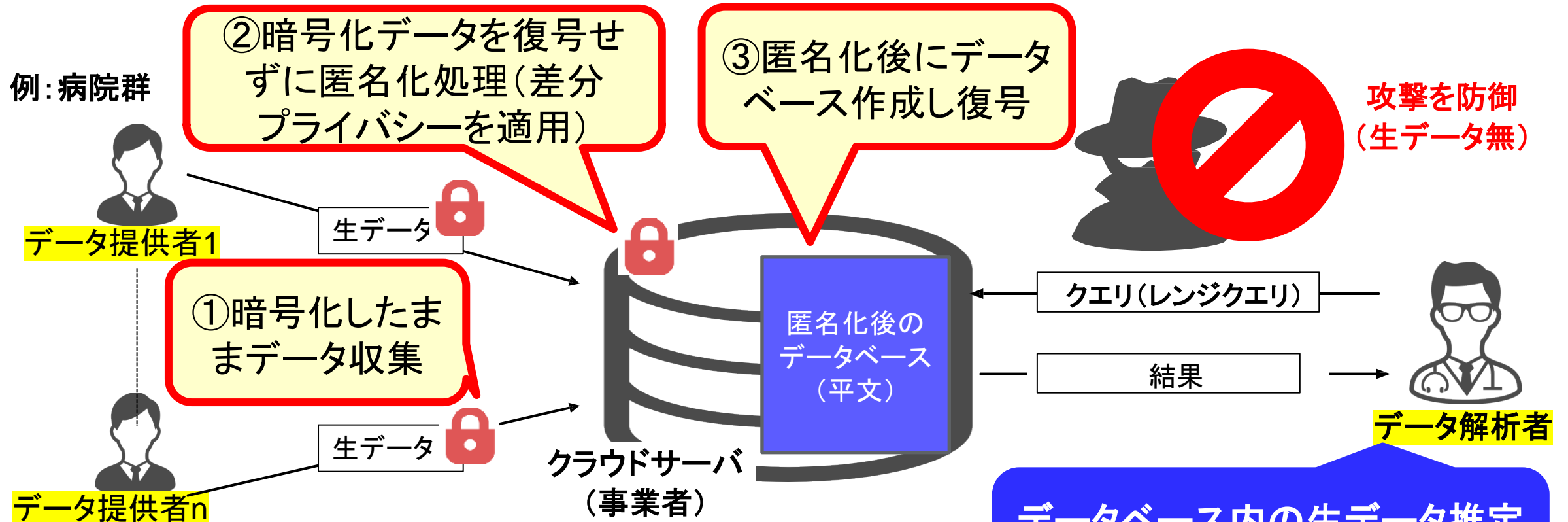


# 5.3 応用事例 (遺伝子をキーとした症例検索)



## 5.4 研究事例(1) - 山名研究室(2023)

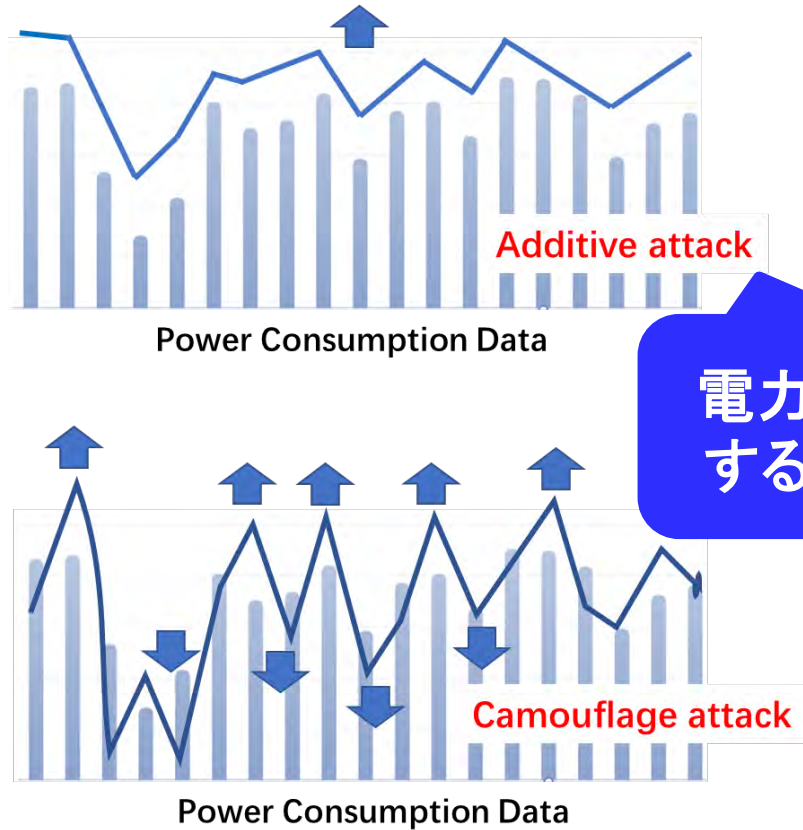
### ■ 暗号のままデータ統合し匿名化する技術[1]



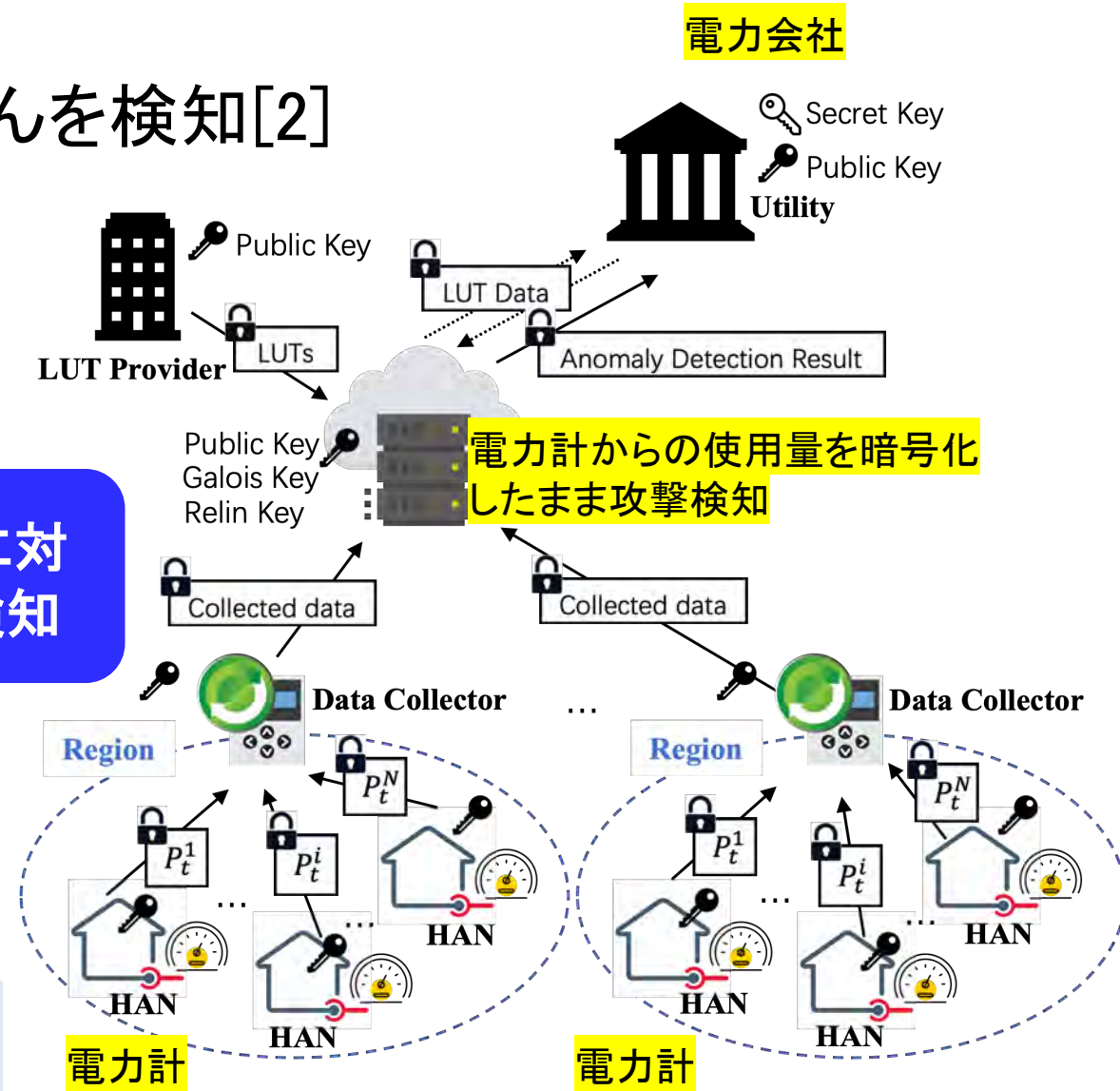
[1] S.Ushiyama, T.Takahashi, M.Kudo, H.Yamana, "Acceleration of a Differentially Private Partitioning Algorithm over Homomorphic Encryption," *IPSJ Transactions on Databases (TOD)*, 16(3), pp. 1-16 (2023-07-21)

# 5. 4 研究事例(2) - 山名研究室(2022)

## ■ 電力使用量を暗号化して収集・改ざんを検知[2]



電力使用量に対する攻撃を検知



[2] R. Li, S. Bhattacharjee, S. K. Das, and H. Yamana, Look-Up Table based FHE System for Privacy Preserving Anomaly Detection in Smart Grids, Proc. of 8th International Conference on Smart Computing (SMARTCOMP2022), pp.1-8 (June 2022)

## 6. おわりに

## 6. おわりに

- 攻撃事例・LLMを用いた生成系AI等でのプライバシー問題を概観
- 対応策：完全準同型暗号による秘密計算を紹介
- 研究動向を紹介

### 秘密計算の利用拡大は、今後5年で様々な分野に浸透か

- 自社・自国データをそのままでは公開したくないが、結果はほしい
  - マネーロンダリングの各国間での共有・解析・予測(金融庁ヒアリング)
  - 素材メーカー(化学、樹脂、ゴム)間でのデータ統合・解析
- オンラインサービスでの消費者保護
  - 生成系AI、検索エンジンなど