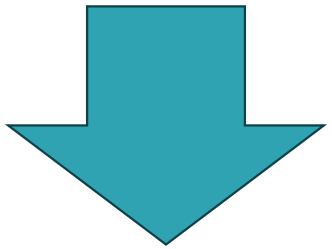
組織横断での安全な統計情報の作成における安全性要件について

2025/10/28

NTT ドコモ／京都橘大学
寺田 雅之

おはなししたいこと

複数組織にまたがるデータに基づき、**安全な統計情報**を作成するためには、**どのような技術的な要件**が求められるか？



(たった) **2 つの安全性要件**に帰着できる

適切な技術を**適切に組み合わせ**れば、現実的に**構成可能**

(なんでも良いから **PETs** を使えばそれで **OK**...というわけではない)

つまり、これを安全に実現するには...

A 社のデータ

なまえ	すきなくだもの
P さん	りんご
Q さん	りんご
R さん	みかん
S さん	みかん
T さん	ドリアン

B 社のデータ

なまえ	すきなどうぶつ
P さん	いぬ
Q さん	ねこ
S さん	ねこ
T さん	いぬ
U さん	くじら

お互いに中身を知ることなしに...

すきなどうぶつ

すきなくだもの	いぬ	ねこ	くじら	...	らいおん
りんご	123人	212人	16人	...	97人
みかん	38人	66人	11人	...	43人
ドリアン	13人	6人	5人	...	7人
:	:	:	:	...	:
すいか	41人	30人	9人	...	19人

こんな表を作る。
(安全な統計情報)

この 2 つの安全性要件を満たす必要がある

A 社のデータ

なまえ	すきなくだもの
P さん	りんご
Q さん	りんご
R さん	みかん
S さん	みかん
T さん	ドリアン

B 社のデータ

なまえ	すきなどうぶつ
P さん	いぬ
Q さん	ねこ
S さん	ねこ
T さん	いぬ
U さん	くじら

お互いに中身を知ることなしに…

■要件 1

自らの不正がない限り、
出力される統計情報以外に、
自らのデータに関する情報が漏洩しないこと。

すきなどうぶつ

すきなくだもの	いぬ	ねこ	くじら	…	らいおん
りんご	1				97人
みかん					43人
ドリアン					7人
:					:
すいか	41人	30人	9人	…	19人

■要件 2

出力されるデータは、
適切にプライバシーが保護された統計情報であること。

こんな表を作る。
(安全な統計情報)

そんなのあたりまえ？

それぞれまじめに検討してみると、
実はそれほどナイーブには実現できません

技術的にしっかりと考えなくちゃいけないこともそれなりに多かったりします

安全性要件 1 について

自らの不正がない限り、出力される統計情報以外に、
自らのデータに関する情報が漏洩しないこと。

あたりまえに思えるかもですが…
実は「自らの不正がない限り」という条件がとてもだいじ

- 自分のデータを、自らの不正で暴露するのは基本的に止められない
- もしもプライバシーが漏洩してしまったときの「責任の所在」を明確にするためにも重要

安全性要件 1 が満たされない例

どちらか片方に (「匿名化」した) データを渡して計算する

もちろんダメ

自らが不正をしなくても、相手の不正などにより自らが保有するデータに関するプライバシー (などの、出力となる統計情報以外の情報) が暴露されうる

第三者に両方ともデータを渡して計算してもらう

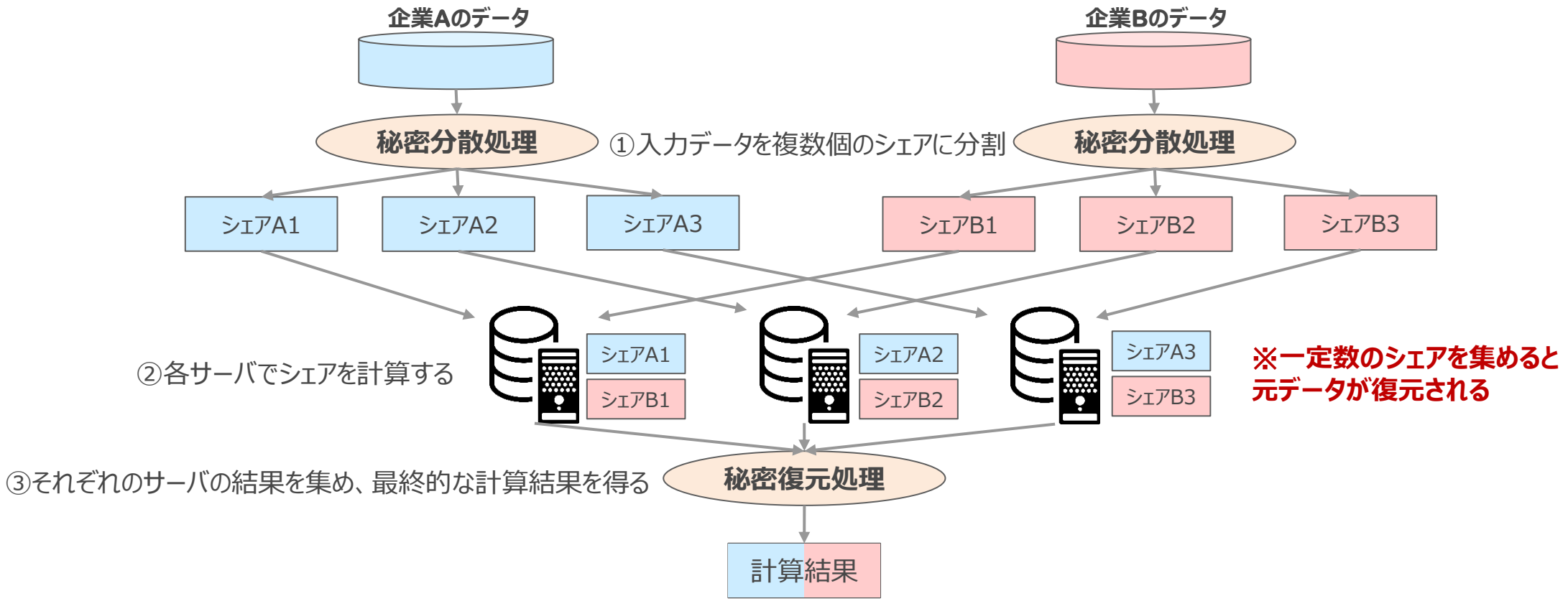
自らが不正をしなくても、その第三者の不正などにより (以下同文)

(k-out of-n) 秘密分散に基づく秘密計算

(PETs の一種である) 「秘密計算」を使えばそれで大丈夫？

秘密分散に基づく秘密計算

入力データを複数の「シェア」に分割し、複数サーバで計算を行い、その結果を集めて計算結果を得る。
単一のシェアからは元データの情報は一切わからないため、各サーバからのデータ漏洩は防止される。
ただし、一定数のシェア (2-out-of-3 秘密分散の場合、3つのうち 2つ) を集めると元データが復元される。



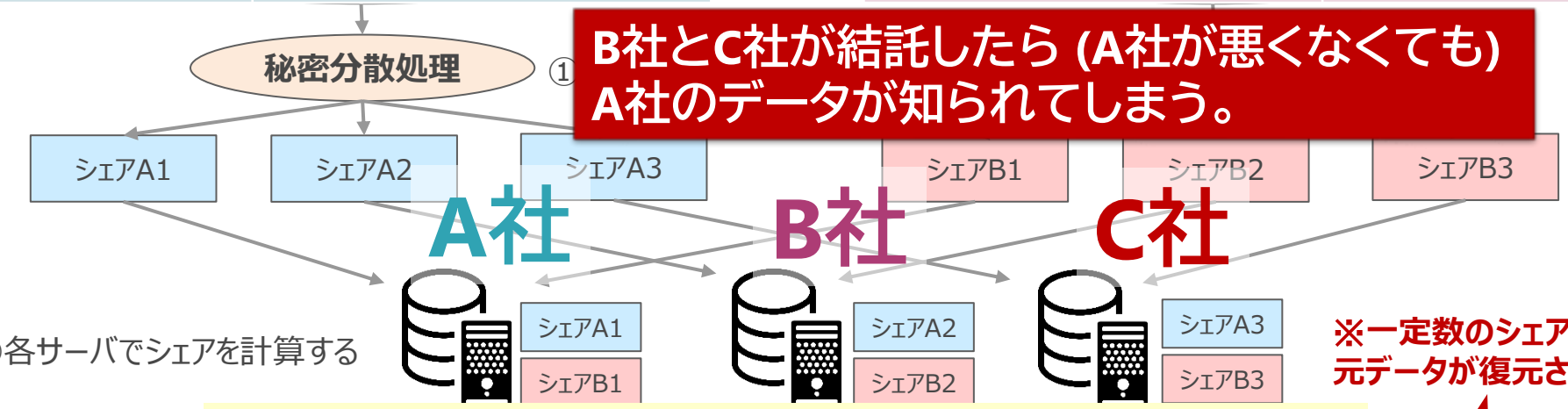
こうすれば良い？

A 社のデータ

なまえ	すきなくだもの
123	りんご
456	りんご
789	みかん
abc	みかん
def	ドリアン

B 社のデータ

なまえ	すきなどうぶつ
def	いぬ
123	いぬ
abc	ねこ
456	ねこ
0ab	くじら



B社とC社が結託したら (A社が悪くなくても)
A社のデータが知られてしまう。

②各サーバでシェアを計算する

③それぞれのサーバの結果を集

※一定数のシェアを集めると
元データが復元される

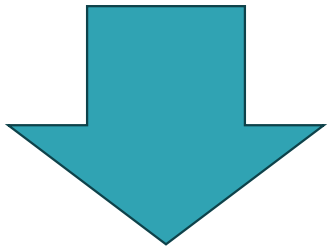
惜しいけどダメ

要件1:「自らの不正がない限り、出力される統計情報以外に、自らのデータに関する情報が漏洩しないこと」を満たさない。

ここが問題

「秘密計算」ならなんでも良いわけではない

「**自らの不正がない限り**、出力される統計情報以外に、自らのデータに関する情報が漏洩しないこと」という条件を満たすためには、他者の**結託による攻撃を許してはならない**。



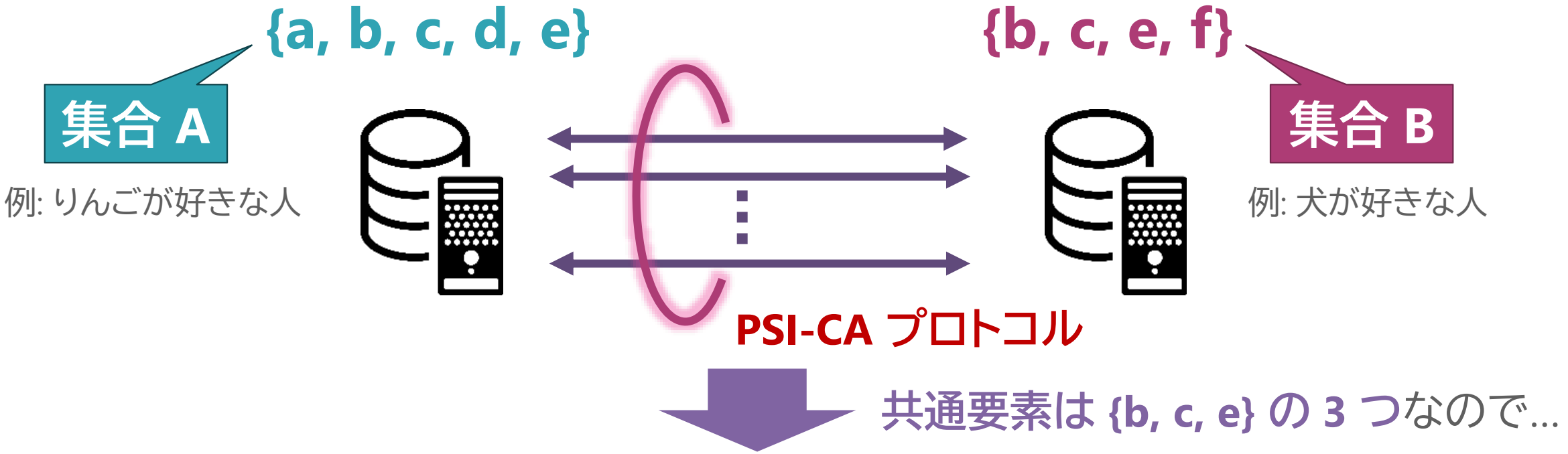
このような技術は存在するか？

(たとえば) **秘匿共通集合濃度計算 (PSI-CA)**

秘匿共通集合濃度計算 (Private Set Intersection Cardinality, PSI-CA)

集合Aと集合Bがそれぞれ異なる計算機で保持されているとき、**互いに集合の内容を明かさずに、共通集合の要素数 |A∩B| のみを計算して出力する**、秘密計算の一種。

一般に、準同型暗号などを用いた**二者間のプロトコル**として実現される。



それぞれの集計区分ごとにこれを繰り返すと...

A 社のデータ

なまえ	すきなくだもの
P さん	りんご
Q さん	りんご
R さん	みかん
S さん	みかん
T さん	ドリアン

B 社のデータ

なまえ	すきなどうぶつ
P さん	いぬ
Q さん	ねこ
S さん	ねこ
T さん	いぬ
U さん	くじら

二者間のみで互いにデータを開かすことなく
以下のような集計表を作成できる
(本当に逐次的に繰り返すと効率が悪いが、実際にはもっと効率的なプロトコルがある)

すきなどうぶつ

すきなくだもの	いぬ	ねこ	くじら	…	らいおん
りんご	123人	212人	16人	…	97人
みかん	38人	66人	11人	…	43人
ドリアン	13人	6人	5人	…	7人
:	:	:	:	…	:
すいか	41人	30人	9人	…	19人

安全性要件 2 について

出力されるデータは、適切にプライバシーが保護された
統計情報であること。(※「統計情報」の定義には深入りしません)

「集計」したら**自動的に安全な統計情報になるわけではない**
⇒ 統計的開示制御 (SDC) を適切に適用すれば実現できるが…

安全な統計情報になる前の (SDC の適用前の) 集計データが
出力されない (誰にも見られない) ことを**保証する**必要がある

安全性要件 2 を満たすためには

(PSI-CA などによる) 集計結果を平文で出力したら NG

それがプライバシーが保護された安全な統計情報とは限らない

その平文を得た人が故意や過失によりプライバシーを漏洩したら？

(SDC 適用後の) 安全な統計情報のみが出力として得られるようにする必要がある

たとえば PSI-CA の適用結果は (そのままでは復号できない) 暗号文として出力し、(差分プライバシーなど) プライバシーが数理的に保証された SDC 技術を暗号文のまま適用してから復号するなど

まとめと今後の展望 (1/2)

複数組織にまたがるデータに基づき、安全な統計情報を作成するために必要となる安全性要件を提起

要件1: 自らの不正がない限り、出力される統計情報以外に、自らのデータに関する情報が漏洩しないこと

要件2: 出力されるデータは、適切にプライバシーが保護された統計情報であること

あたり前のようにも思えるが、まじめに考えると技術的にしっかりと検討すべきことは多い

まとめと今後の展望 (2/2)

これを本当に「あたり前」にするために技術に求められること

※私見をたっぷり含みます

安全性要件 1 を満たす、さらに効率的な秘密計算技術の確立

以前に比べるとだいぶ実用的になってきたが、それでも大規模データに適用することを考えるとさらなる高速化や負荷軽減が求められる

安全性要件 2 を満たすための SDC 適用技術の高度化

暗号文のまま SDC を適用するのはけっこう計算が大変 (ボトルネック)

単純な Laplace メカニズムを適用するだけでもひと苦労

より高度な (複雑な) メカニズムを、スケーラブルに適用可能にするには？