業界横断の安全なデータ活用に基づく社会課題解決の試み

2024/2/15

NTT ドコモ　寺田 雅之

自己紹介

寺田 雅之 (てらだ まさゆき)

NTTドコモ クロステック開発部 担当部長・セキュリティプリンシパル ← 本務
滋賀大学 データサイエンス・AIイノベーション研究推進センター 特任教授
総務省 統計研究研修所 特任教授
情報処理学会 フェロー、理事　など

主に、統計データを…

- 作る - **大規模データからの統計作成** (モバイル空間統計)
- 使う - **統計の社会予測への活用** (AI渋滞予知)
- 守る - **統計のプライバシー保護** (差分プライバシー/**秘匿クロス統計**)

ための技術の創出と、その**研究・実用化**に取り組んでいます。

研究内容の紹介 (1/3)

統計データを「つくる」技術

「モバイル空間統計」

統計データを「つかう」技術

「AI 渋滞予知」

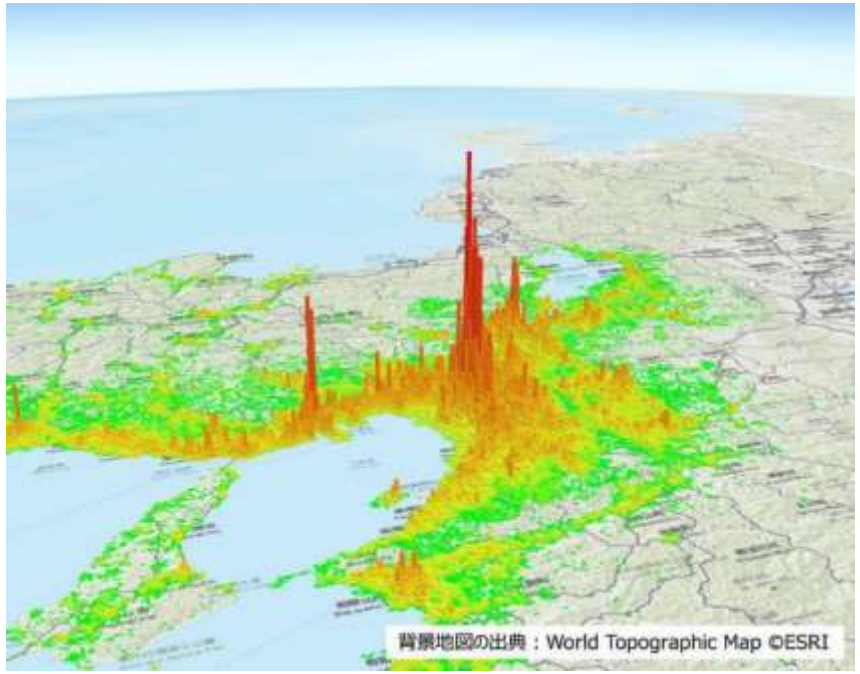
統計データを「まもる」技術

「差分プライバシー / 秘匿クロス統計」

モバイル空間統計

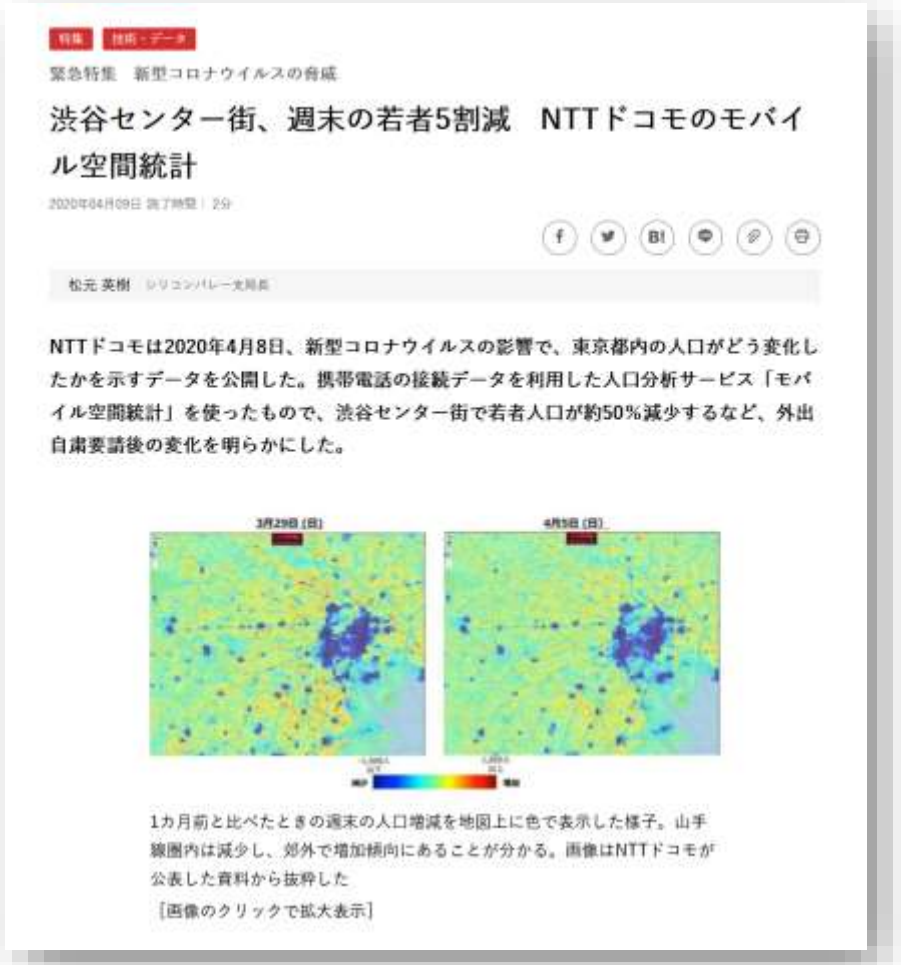
「携帯電話がつながるしくみ」を支える、**携帯電話ネットワークの運用データ**に基づいて、
どこにどういう人が何人いて、それが**時間ごとにどう変化**しているかを、
日本全国で継続的に把握できるようにする統計情報です。

日本におけるビッグデータの活用の先駆けとして2013年3月に**商用での提供**を開始し、
2020年1月からは**「1時間前の人口」がわかるリアルタイム版を提供**しています。



新型コロナウイルス (COVID-19) 対策で広く活用

「リアルタイム版モバイル空間統計」の商用サービス開始（2020年1月）直後に**新型コロナウイルス**の感染拡大が社会問題化し、国・自治体やマスメディアなどにより、**感染拡大抑止のための政策立案**や、**「三密」回避に向けた行動変容の啓発**、**人々の行動変化に関する報道**などに広く活用いただきました。



緊急特集　新型コロナウイルスの脅威

渋谷センター街、週末の若者5割減　NTTドコモのモバイル空間統計

松元 英樹

NTTドコモは2020年4月8日、新型コロナウイルスの影響で、東京都内の人口がどう変化したかを示すデータを公開した。携帯電話の接続データを利用した人口分析サービス「モバイル空間統計」を使ったもので、渋谷センター街で若者人口が約50%減少するなど、外出自粛要請後の変化を明らかにした。

1カ月前と比べたときの週末の人口増減を地図上に色で表示した様子。山手線圏内は減少し、郊外で増加傾向にあることが分かる。画像はNTTドコモが公表した資料から抜粋した

［画像のクリックで拡大表示］

（出典: 日経クロストレンド）

神奈川県

ホーム > くらし・安全・環境 > 防災と安全 > 災害情報・危機管理 > 新型コロナウイルス感染症発生後の人口流動分析

新型コロナウイルス感染症発生後の人口流動分析

掲載日：2020年6月30日

新型コロナウイルス感染症発生後の人口流動分析

（6月30日　更新）

6月29日（月曜日）15時時点の県内13地点の人口流動の分析を株式会社ドコモ・インサイトマーケティングの「モバイル空間統計」を活用して数値化しました。

緊急事態宣言解除後も、「密閉」、「密集」、「密接」を避ける行動や人と人との距離の2mの確保、咳エチケット、手洗い・うがい・手指消毒など引き続き、感染症防止対策の御協力をお願いします。

6月29日（月曜日）15時時点と各時点での比較

前日との比較：6月28日15時時点
事態宣言前：4月7日15時時点
感染拡大以前：1月18日から2月14日の休日の平均データ

（出典: 神奈川県）

さらなる応用領域の拡大に向けた機能拡充

DX 化の進展により高度化しつつある社会ニーズに応えるため、推計可能な統計のバリエーションをさらに拡充し、さまざまな社外パートナーとの連携を通じ、その有用性の検証を進めています。

「リアルタイム異常検出」の実現と災害対応への適用性検証

防災科研 NIED

モバイル空間統計

×

避難場所

「隠れ避難所」の検出など

「リアルタイム移動人口統計」の実現と交通計画への適用性検証

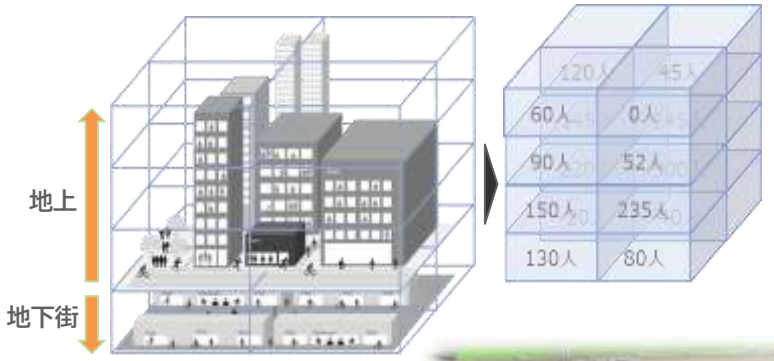
長崎県 NAGASAKI PREFECTURE

国立大学法人 長崎大学 NAGASAKI UNIVERSITY

「三次元人口統計」の実現と防災分野への適用性検証

デジタル庁 Digital Agency

NTT DATA



G7デジタル・技術大臣会合
(2023年4月29～30日) で展示・紹介

研究内容の紹介 (2/3)

統計データを「つくる」技術

「モバイル空間統計」

統計データを「つかう」技術

「AI 渋滞予知」

統計データを「まもる」技術

「差分プライバシー / 秘匿クロス統計」

人々の集まり方がわかると、何ができるようになるか？

雨が降ると…

しばらくして…

川の水位が上がる

降雨の状況から
水位の上昇や
堤防の越水などが
予測できる

人が集まると…

しばらくして…

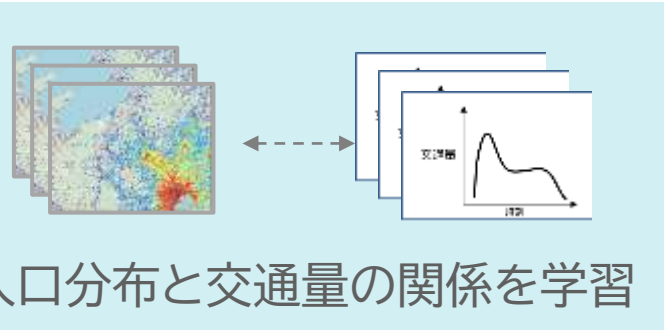
渋滞したり、売上が増えたり、
感染者が増えたりする

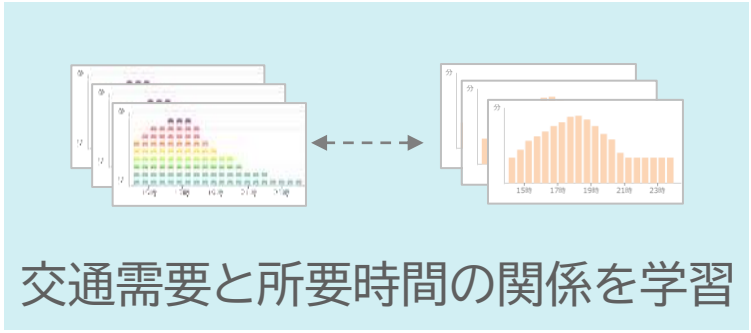
人口の推移から
交通渋滞の発生や
店舗売上の変動や
感染症の拡大などが
予測できる？

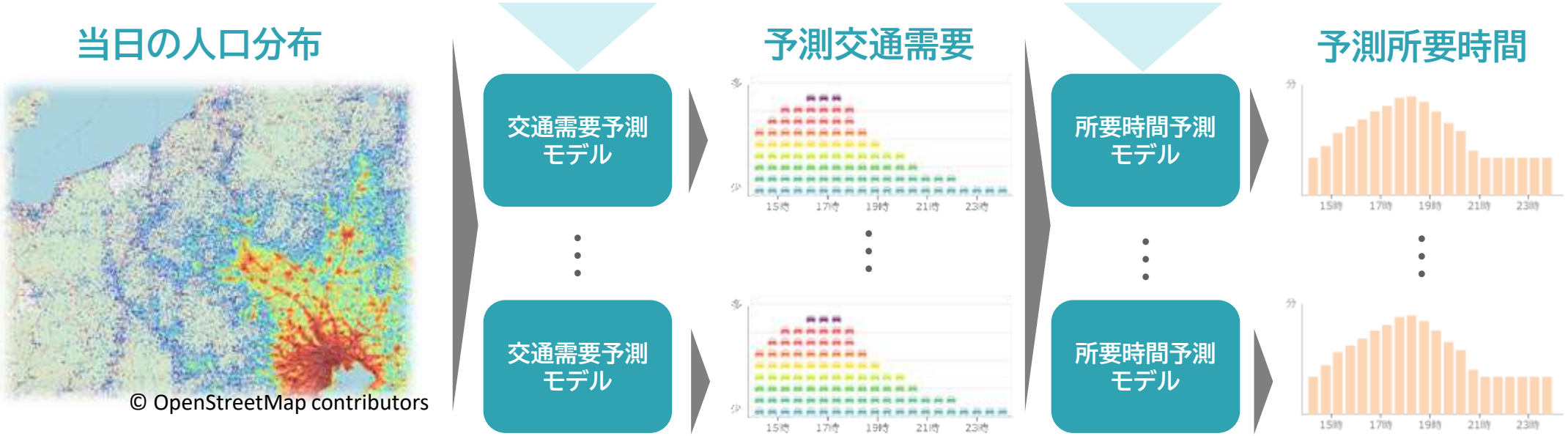
人々の集まり方がわかれば、社会の「未来」が予測できる？

AI渋滞予知：「いま」の人口から「未来」の渋滞を予測する

モバイル空間統計から得られる「いま」の人出の状況に基づいて、数時間先の「未来」における渋滞の発生を**実用的な精度で予測**する技術です。
実際の人出の多さに基づく信頼性が高い予測情報を提供することで**ドライバーが渋滞を回避**しやすくなり、さらに交通が分散されることにより**渋滞そのものが消失・緩和**することなどが期待されます。

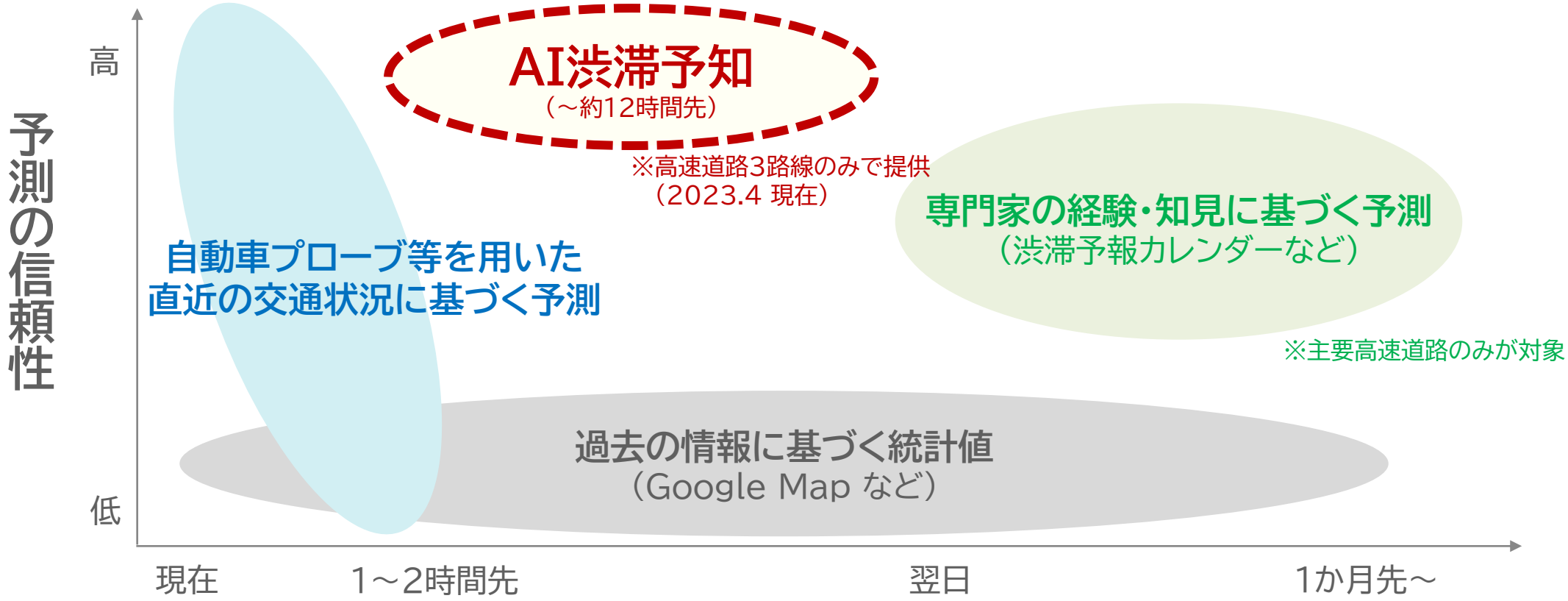






AI渋滞予知の位置づけ

その日の「これからの行動」を決めるために重要となる**「数時間後」の渋滞**を高い信頼性で予測することができる、(おそらく) **世界で初めて**の技術です。



関越自動車道・東京湾アクアラインでサービス提供中

東京湾アクアライン上り線 (川崎方面) と、**関越自動車道**上り線 (沼田～練馬間) の
渋滞予測情報を、NEXCO東日本の道路情報サイト「ドラぷら」で**毎日配信**しています。

予測所要時間を表示

予測交通需要を表示

京葉道路でも正式提供に向けて共同実証中

「AI渋滞予知」で検索！

研究内容の紹介 (3/3)

統計データを「つくる」技術

「モバイル空間統計」

統計データを「つかう」技術

「AI 渋滞予知」

いちばん地味だけど
いちばん大事

統計データを「まもる」技術

「差分プライバシー / 秘匿クロス統計」

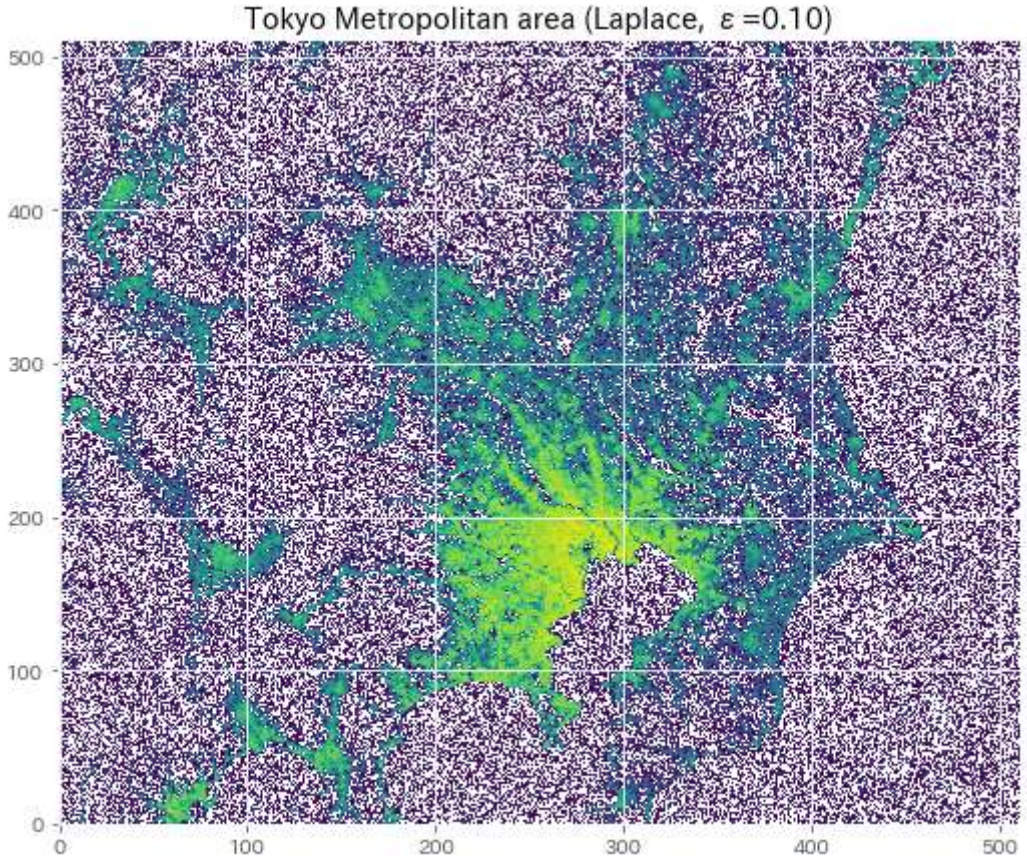
なぜならば

健全なデータ活用は、
適切なプライバシー保護なしには
実現できない。

大規模集計データへの差分プライバシーの適用

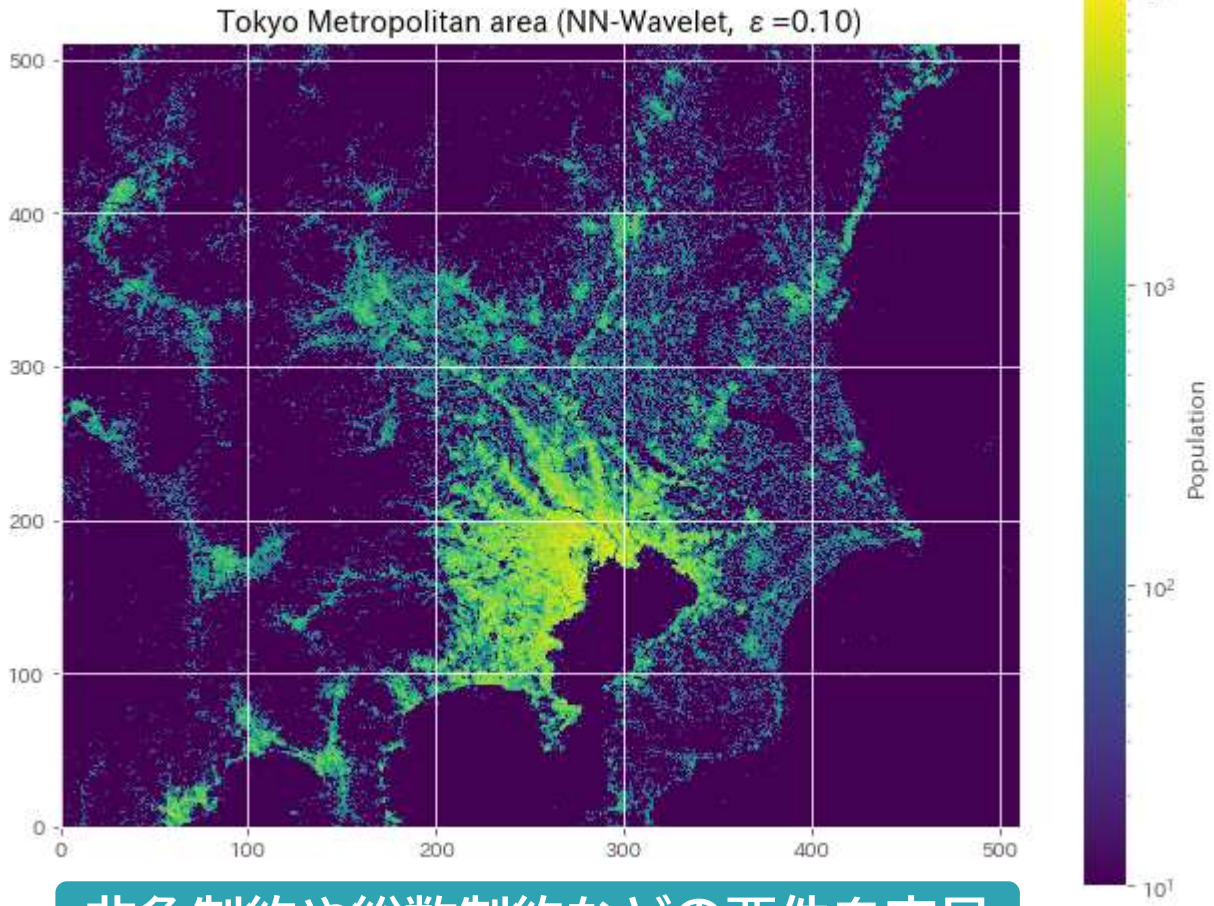
局所性保存写像である**Morton順序写像**や、大域-局所に情報を分解する**Wavelet変換**などを組み合わせ、非負制約・総数制約の逸脱など統計としての**有用性を毀損する要因を排除**しつつ、**大規模なデータに対しても高速**に**差分プライバシーに基づく安全性**を与えます。

Laplace メカニズムの適用結果



負の値を表す「白抜き」が全体に拡がる

提案手法の適用結果



非負制約や総数制約などの要件を充足

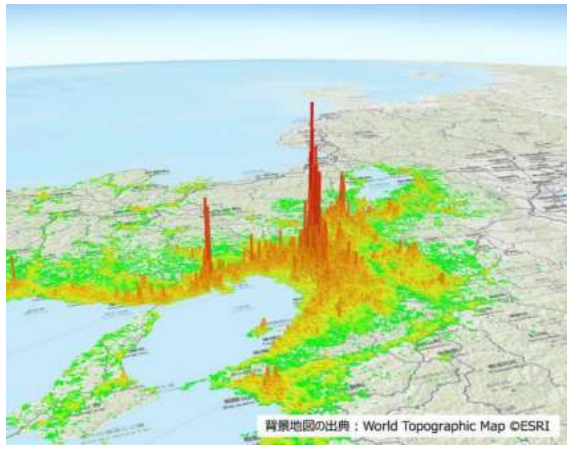
「安全なデータ」に基づく社会価値の創出

本研究により得られた各種の成果・知見は、モバイル空間統計のプライバシー保護に活用され、**モバイル空間統計のリアルタイム化を実現する上での鍵**となりました。

また、「AI渋滞予知」などの**モバイル空間統計と AI 技術を組み合わせた応用**の実現を通じ、**安全なデータ**を用いて**有益な社会価値**を創出できること（**安全性と有用性の両立可能性**）を示す取り組みを進めています。

差分プライバシー技術に基づき
リアルタイムに安全性を保証

携帯電話ネットワークの運用データ
(そのままでは使えないデータ)



モバイル空間統計
(誰でも使える安全な統計情報)

「AI渋滞予知」などの応用
(有益な社会価値の創出)

秘密計算技術との組み合わせによる、さらなる応用


JAPAN AIRLINES　JALCARD　docomo

トピックス

2022年10月20日
日本航空株式会社
株式会社ジャルカード
株式会社 NTT ドコモ

JAL、JAL カード、ドコモが、顧客体験価値向上と社会課題の解決に向けて、
「秘匿クロス統計技術」を用いた企業横断でのデータ活用の実証実験を開始
～各社が保有するデータを相互に開示せず作成した統計情報を活用する国内初の取り組み～

日本
株式会
株式会
JAL、
（以下
本実
本技術
状態
機械
ことで
国内
一部

JAPAN AIRLINES　docomo

国内線航空券の予約データの搭乗に関する情報

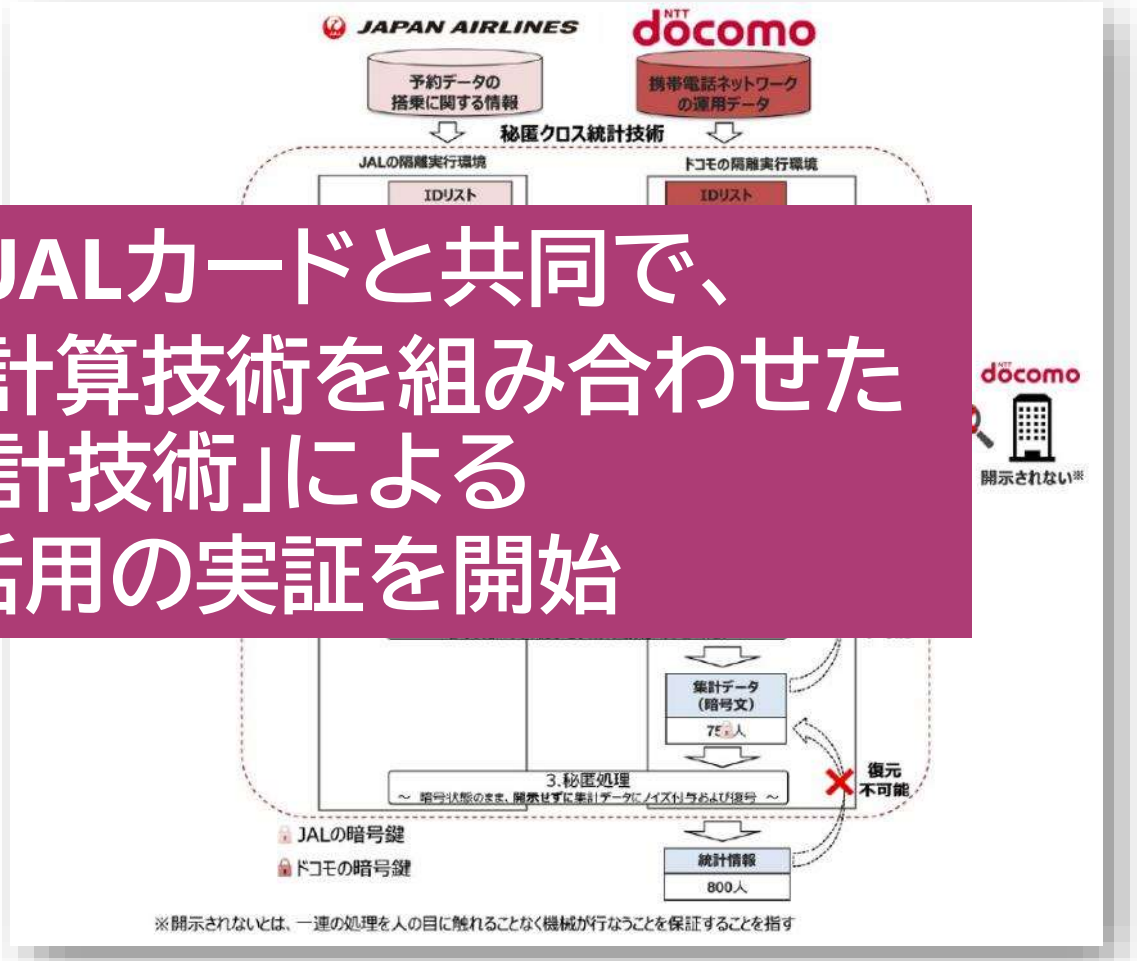
携帯電話ネットワークの運用データ

秘匿クロス統計技術により、
個人を識別できない状態に加工したうえで、
各社が保有するデータを開示せずに統計情報を作成

空港ご到着前の各時点でのお客さまの移動状況に関する統計情報

	便出発前日			便出発60分前			便出発40分前			便出発20分前		
	居住地域	その他	空港周辺	居住地域	その他	空港周辺	居住地域	その他	空港周辺	居住地域	その他	空港周辺
11月の午前便の搭乗客	2,898	2,001	101	1,610	3,085	305	252	1,206	3,542	79	86	4,835
11月の午後便の搭乗客	3,898	2,901	201	2,610	3,985	405	352	2,106	4,542	60	87	5,855

※回は一例であり、単位は人です。



2022年10月に JAL・JALカードと共同で、
差分プライバシーと秘密計算技術を組み合わせた
「秘匿クロス統計技術」による
企業横断データ活用の実証を開始

今日の本題になります。


JAPAN AIRLINES　JALCARD　NTT docomo

トピックス

2022年10月20日
日本航空株式会社
株式会社ジャルカード
株式会社 NTT ドコモ

JAL、JAL カード、ドコモが、顧客体験価値向上と社会課題の解決に向けて、「秘匿クロス統計技術」を用いた企業横断でのデータ活用の実証実験を開始
～各社が保有するデータを相互に開示せず作成した統計情報を活用する国内初の取り組み～

JAPAN AIRLINES　docomo

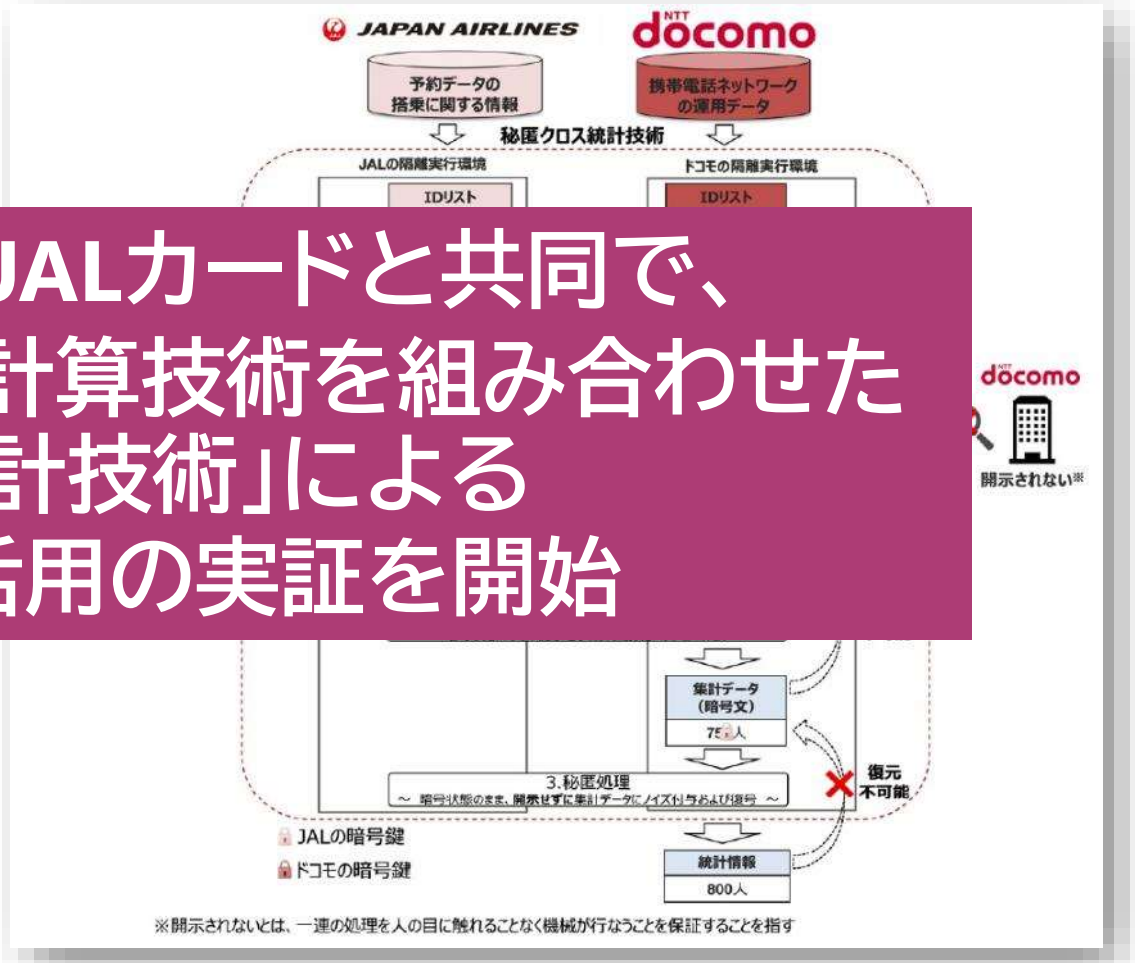
国内線航空券の予約データの搭乗に関する情報　　携帯電話ネットワークの運用データ

秘匿クロス統計技術により、個人を識別できない状態に加工したうえで、各社が保有するデータを開示せずに統計情報を作成

空港ご到着前の各時点でのお客さまの移動状況に関する統計情報

	便出発前日			便出発60分前			便出発40分前			便出発20分前		
	居住地域	その他	空港周辺	居住地域	その他	空港周辺	居住地域	その他	空港周辺	居住地域	その他	空港周辺
11月の午前便の搭乗客	2,898	2,001	101	1,610	3,085	305	252	1,206	3,542	79	86	4,835
11月の午後便の搭乗客	3,898	2,901	201	2,610	3,985	405	352	2,106	4,542	60	87	5,855

※回は一例であり、単位は人です。

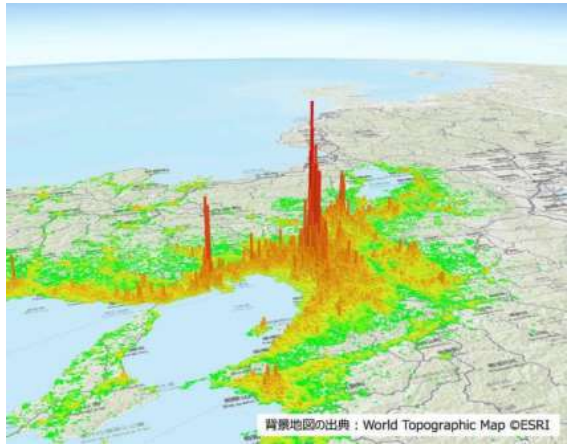


2022年10月に JAL・JALカードと共同で、差分プライバシーと秘密計算技術を組み合わせた「秘匿クロス統計技術」による企業横断データ活用の実証を開始

「モバイル空間統計」でやってきたこと

差分プライバシー技術に基づき
リアルタイムに安全性を保証

携帯電話ネットワークの運用データ
(守らなくてはいけないデータ)

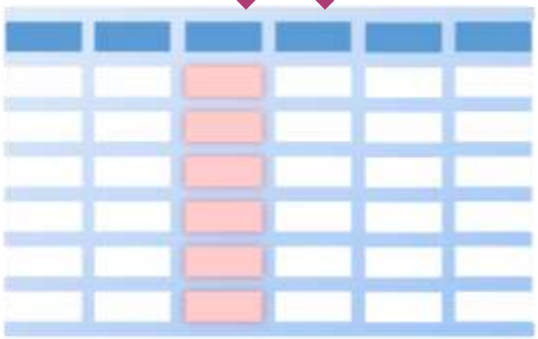


モバイル空間統計
(誰でも使える安全な統計情報)

「AI渋滞予知」などの応用
(有益な社会価値の創出)

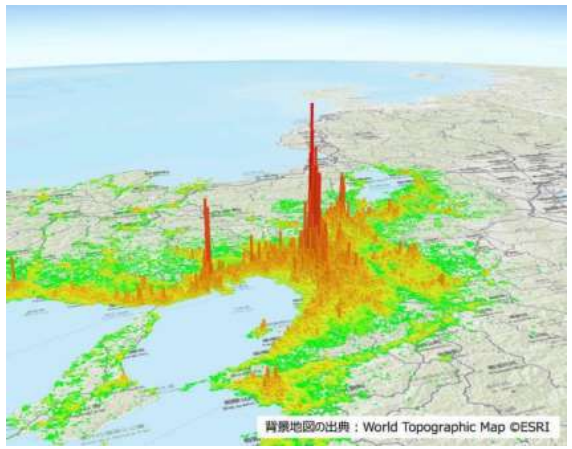
「秘匿クロス統計」でやりたいこと

×



異なる組織がそれぞれ保有する
複数のデータ

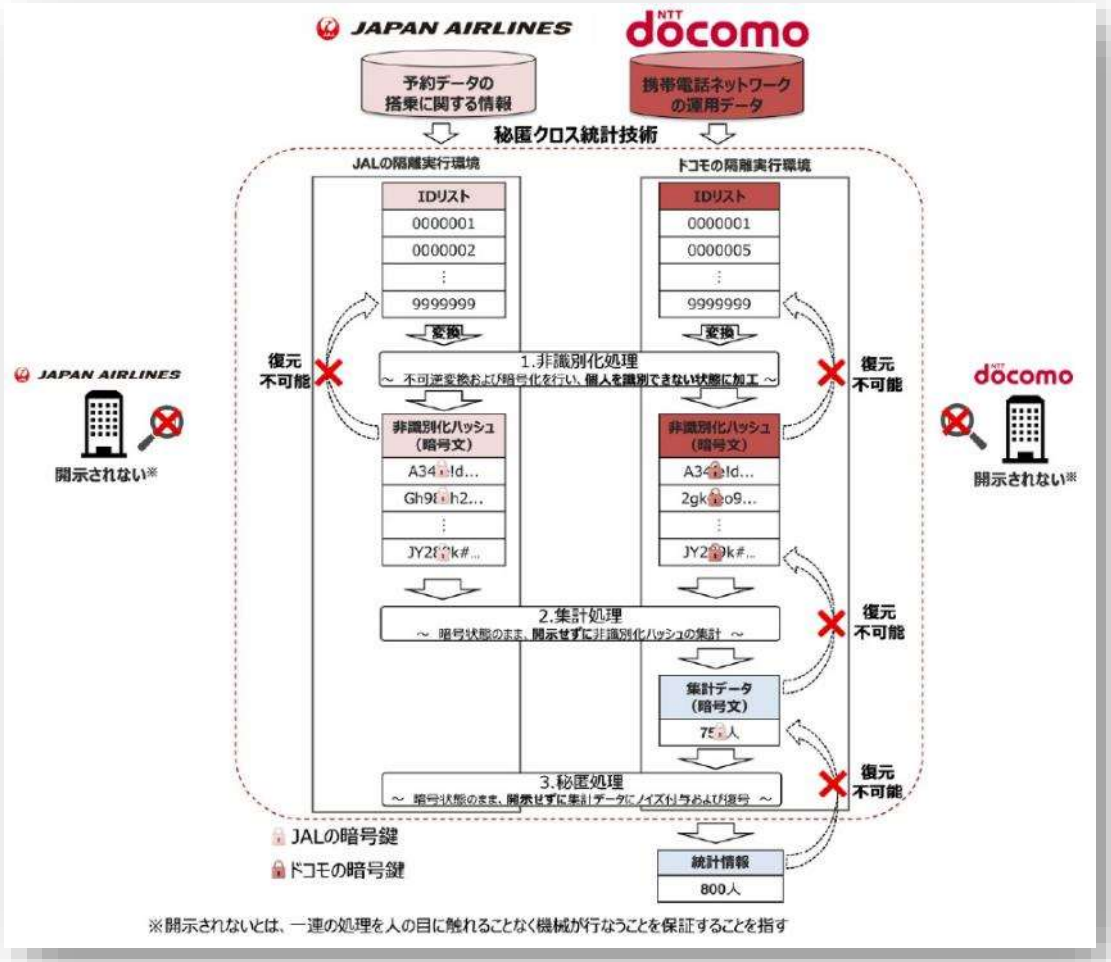
どうやって実現する？



(いままで作れなかった)
安全な統計情報

(いままでできなかった)
有益な社会価値の創出

秘匿クロス統計の概要



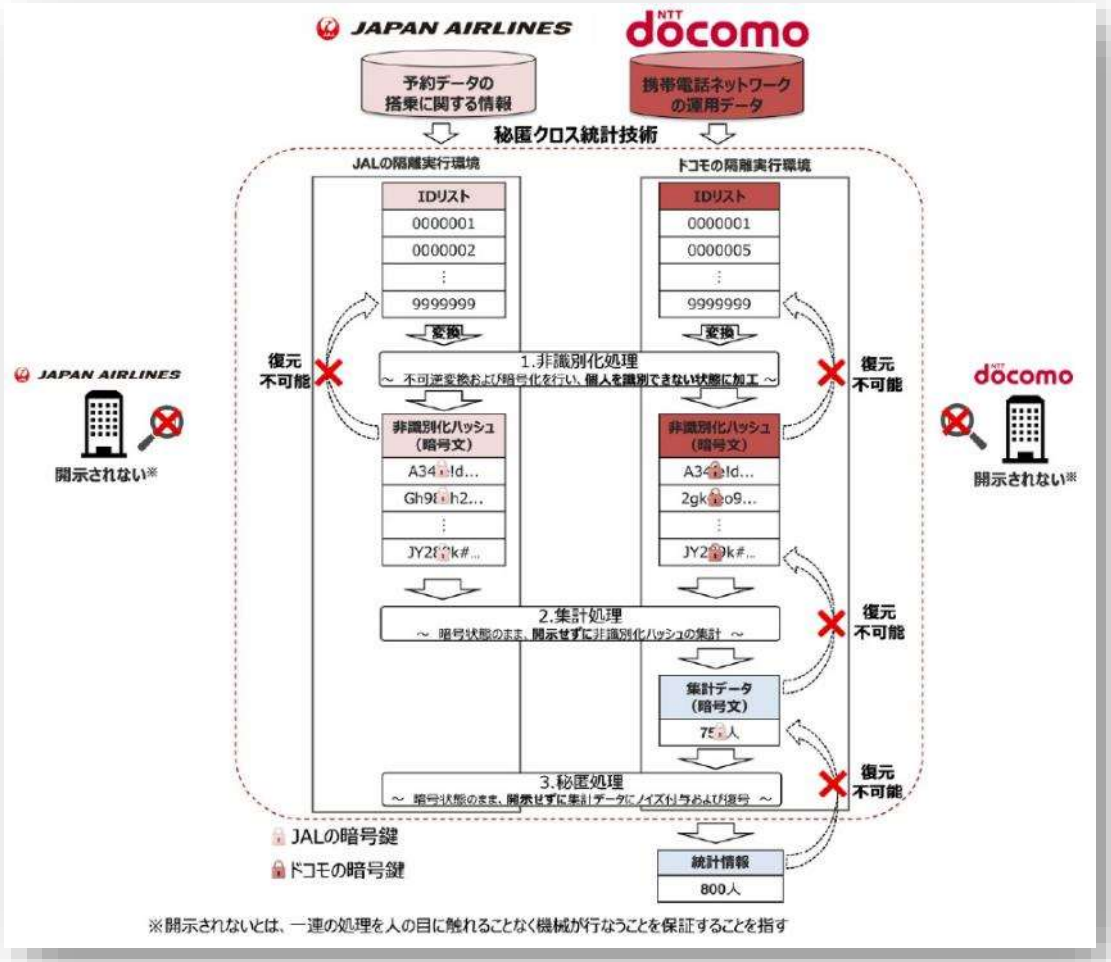
A社が保有するデータと、
B社が保有するデータに基づき、

1. **非識別化処理**
2. **集計処理**
3. **秘匿処理**

の3段階のプロセスを経て、

プライバシーが保護された安全な統計情報 (クロス集計表) を出力する技術 (NTT社会研と協力して開発)

秘匿クロス統計の安全性要件



■要件1

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。

■要件2

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。

もう少しひらたく言うと…

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

お互いに中身を知ることなしに…

■要件2

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					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1

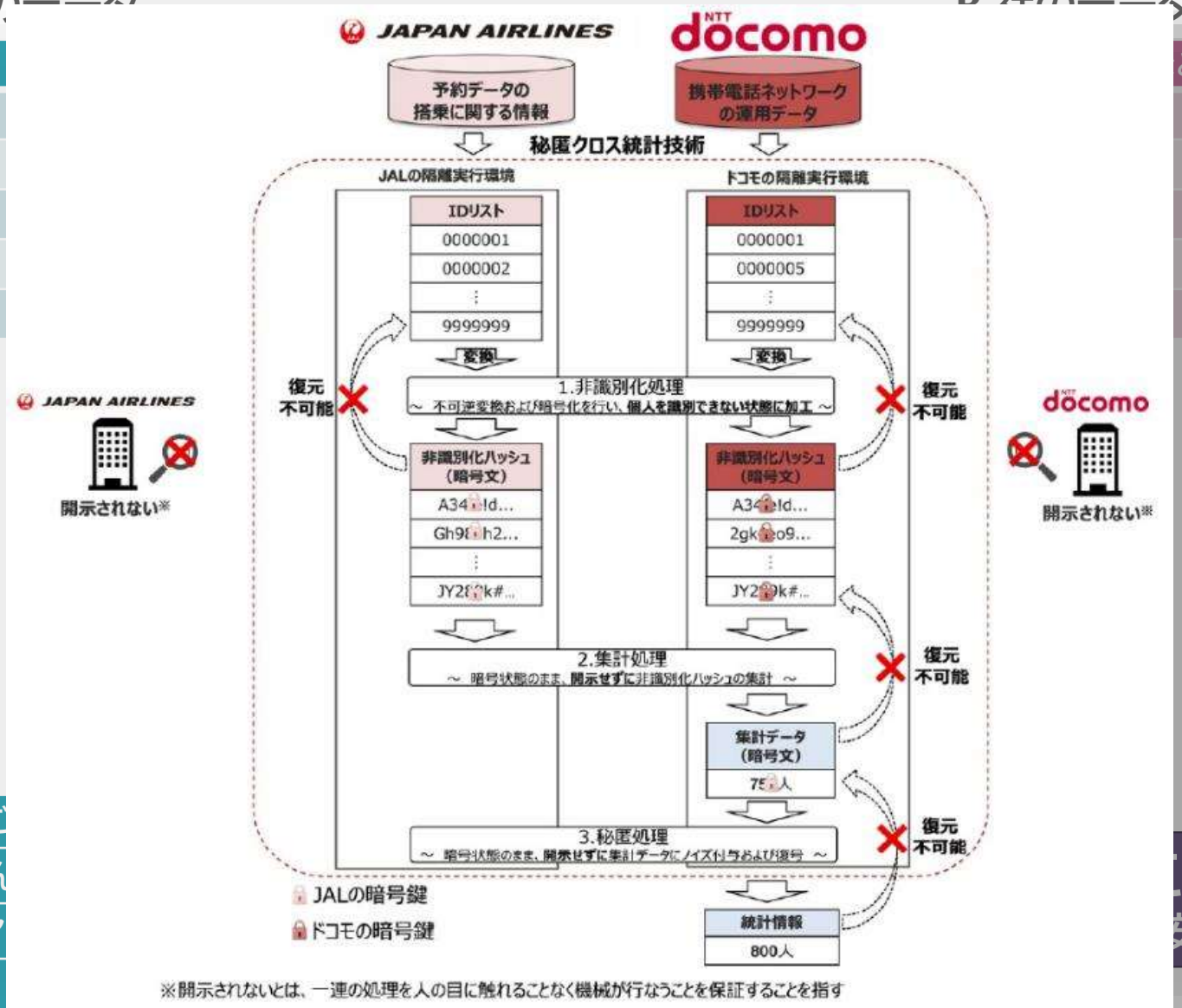
出力されるデータは、
適切にプライバシーが保護された統計情報であること。

こんな表を作る。
(安全なクロス集計表)

なんでこんな (めんどくさい) ことしなくちゃいけないの？

A 社のデータ

なまえ
P さん
Q さん
R さん
S さん
T さん



B 社のデータ

どうぶつ

すきなくだもの

りんご	みかん	ドリア	:	すいか
				41人
				30人
				9人
				…
				19人

んな表を作る。
安全なクロス集計表)

なんでこんな (めんどくさい) ことしなくちゃいけないの？

というあたりが
今日のお題になります。

やりたいこと (再掲)

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。
(安全なクロス集計表)

片方にデータを渡してもよければ簡単

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

A社にデータを全部わたしてしまう

なまえ	すきなくだもの	すきなどうぶつ
P さん	りんご	いぬ
Q さん	りんご	ねこ
S さん	みかん	ねこ
T さん	ドリアン	いぬ

この表を集計すれば
欲しかった集計表ができるが…

B社のデータ (みんなの「すきなどうぶつ」) が
A社に知られてしまう。

もちろんダメ

要件2:「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。

「匿名化」すれば OK か？

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

共通のソルト乱数を使って「なまえ」をハッシュ化

ソルト乱数

共通のソルト乱数を使って「なまえ」をハッシュ化

※順序もシャッフルしておく

「くじら」が好きな人は、Uさん一人しかいない
→「0ab は Uさん」と特定できるので削除する

ハッ	
123	りんご
456	りんご
789	みかん
abc	みかん
def	ドリアン

ハッシュ値	すきなどうぶつ
def	いぬ
123	いぬ
abc	ねこ
456	ねこ
~~0ab~~	~~くじら~~

「匿名化」すれば OK か？

ハッシュ値	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

ハッシュ値	すきなどうぶつ
def	いぬ
123	いぬ
abc	ねこ
456	ねこ
~~0ab~~	~~くじら~~

A社に「匿名化」したデータをわたす

ハッシュ値	すきなくだもの	すきなどうぶつ
123	りんご	いぬ
456	りんご	ねこ
abc	みかん	ねこ
def	ドリアン	いぬ

この表を集計すれば
欲しかった集計表ができるが…

「ドリアン」が好きな人は、T さん一人しかいない
→「T さんはいぬが好き」と知られてしまう

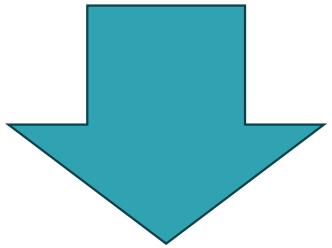
B社のデータ (T さんの「すきなどうぶつ」) が A社に知られてしまう。

これもダメ

要件2:「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。

「匿名化」だけではどうやっても難しい

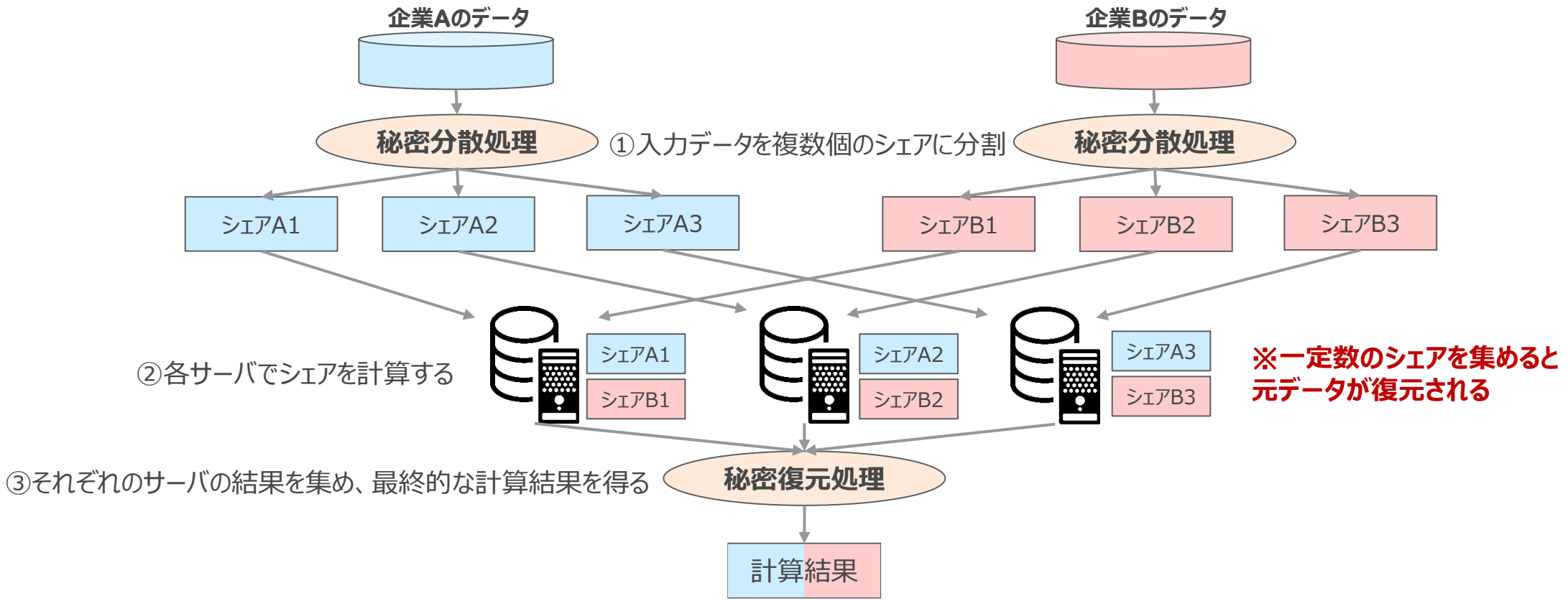
もう少し「凝った」匿名化をすることもできるが、いずれにしろ
共通 ID を持たせたデータを「平文のままで」渡してしまうと、
特定個人の識別やプライバシーの暴露をまぬがれない。



互いに相手のデータを見ることなく、計算の結果だけを得る
ような技術 = 秘密計算の導入が必要

秘密計算の例: 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。



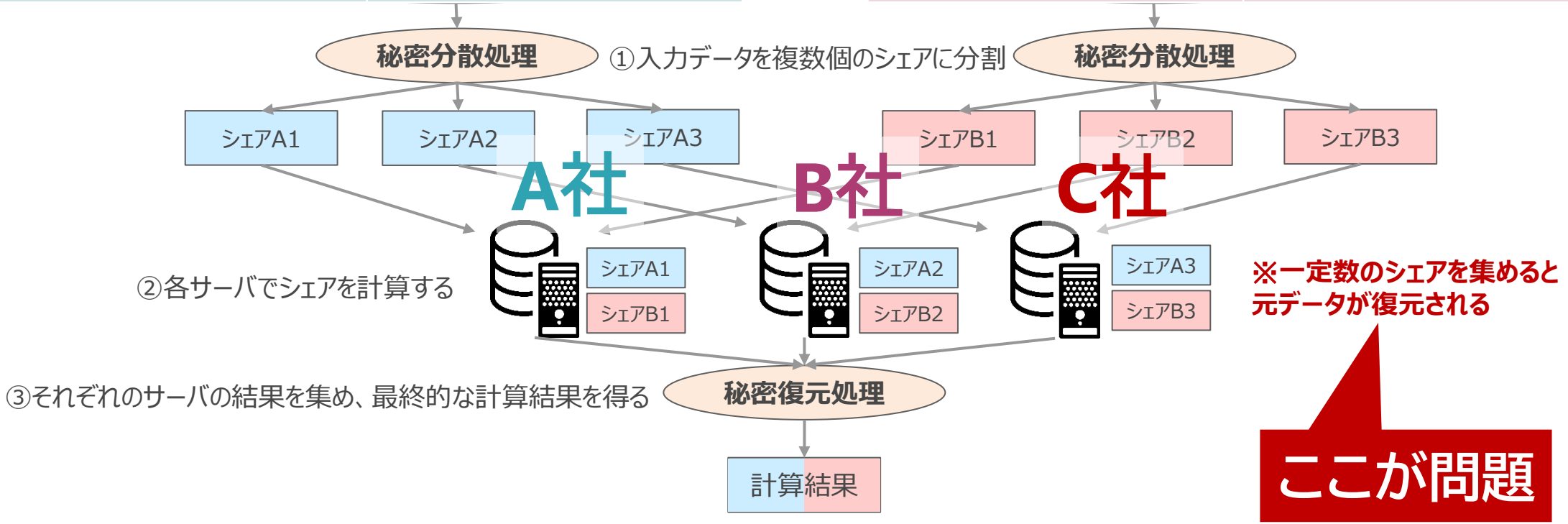
こうすれば良い？

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



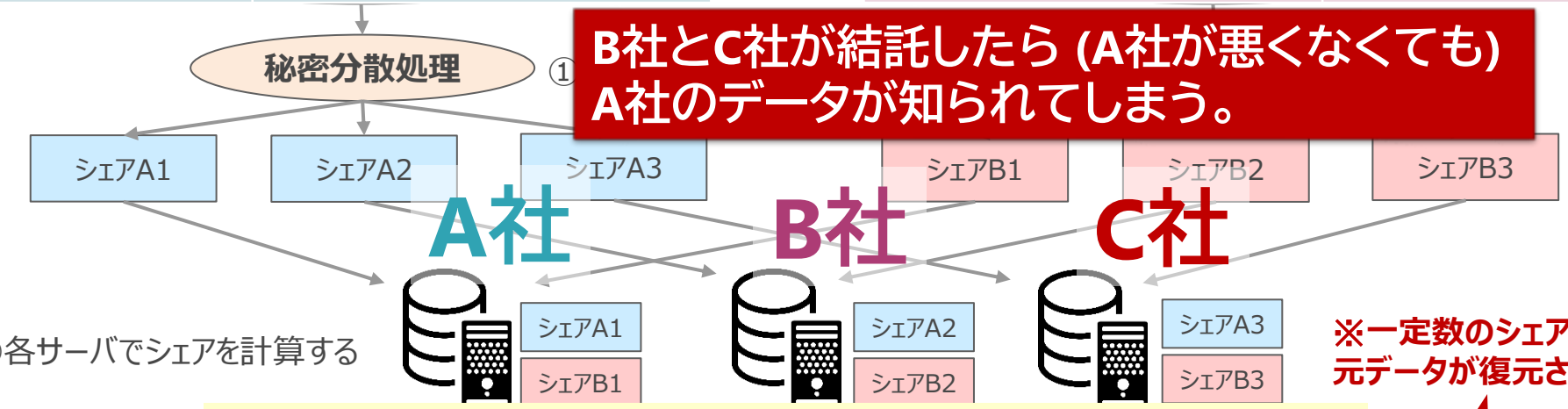
こうすれば良い？

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

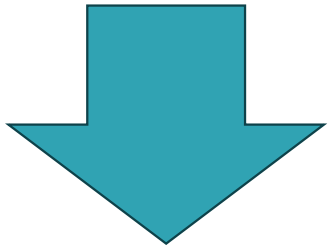
惜しいけどダメ

要件2:「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、**結託による攻撃を許してはならない**。

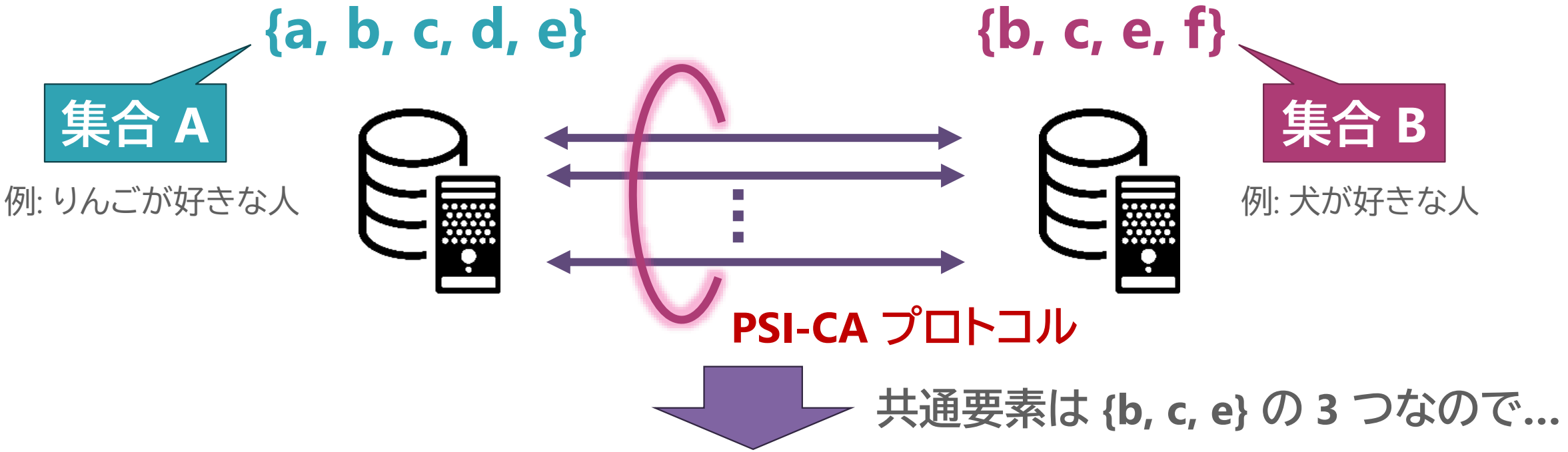


二者間に閉じて (結託なしに) 安全に集計ができる技術
= **秘匿共通集合濃度計算 (PSI-CA)** に着目

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。



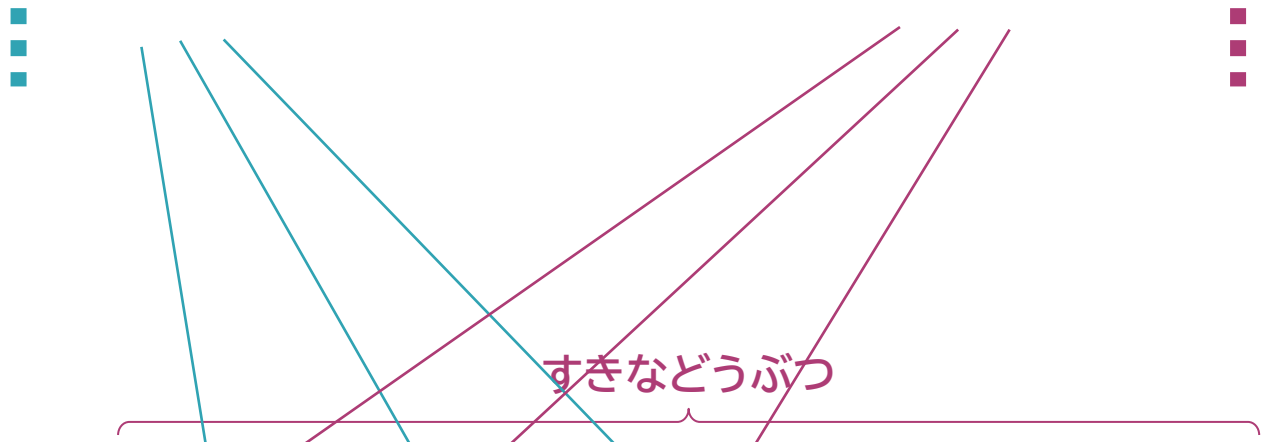
それぞれの集計区分ごとにこれを繰り返すと...

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



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

こんな表ができる
(クロス集計表)

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

要件2:「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、
(ようやく) 満たすことができた。

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

こんな表ができる
(クロス集計表)

実際にはもっと効率的な方法で計算しています。

A 社のデータ

なまえ	すきなくだもの
P	
C	
R	
S	
T	

データ

すきなどうぶつ
いぬ
ねこ
ねこ
いぬ
くじら

具体的には
こんな感じ

要件2:「自らの不正がな
自らのデータに関する情
(ようやく) 満たす

服以外に

Protocol 2 秘匿共通集合プロトコルによるクロス集計 [9]

Input: T'_A（組織 A），T'_B（組織 B）
Output: $Enc_A(T_C)$（組織 B）

1: 組織 A は ID'_A を秘密鍵 a で暗号化を行い，$(ID'_A)^a$ を得る．同様に組織 B も ID'_B，ID_D を秘密鍵 b でハッシュ化を行い，$(ID'_B)^b$，$(ID'_D)^b$ を得る．
2: 組織 A が組織 B に $(ID'_A)^a$ を送る．
3: 組織 B が $(ID'_A)^a$ を秘密鍵 b でハッシュ化を行い，$(ID'_A)^{ab}$ を得る．組織 B は $(ID'_A)^{ab}$ をシャッフルする．
4: 組織 B は組織 A に $(ID'_A)^{ab}$，$(ID'_B)^b$，$(ID'_D)^b$ を送る．
5: 組織 A は秘密鍵 a と $(ID'_A)^{ab}$ から $(ID'_A)^b$ を計算する．
6: 組織 A は $(ID'_A)^b$ と $(ID'_B)^b$ から ID'_B に含まれ ID'_A に含まれないような ID の個数 k を得る．
7: 組織 A は $(ID'_D)^b$ から k 個の要素を選ぶ．
8: 組織 A は $(ID'_B)^b$ の要素のうち，$(ID'_A)^b$ に属さない k 個を前ステップで選んだ $(ID'_D)^b$ の要素で置き換え，$(ID'_{(B|D)})^b$ とする．
9: 組織 A は秘密鍵 c で $(ID'_{(B|D)})^b$ をハッシュ化し，$(ID_{(B|D)'})^{bc}$ を得る．
10: 組織 A は $T'_{A|D} := T_A||T_D$ を計算する．ただし，$T'_D := ((id_{D,1},0),...,(id_{D,k},0))$ とする．
11: 組織 A は $T'_{A|D}$ をシャッフルする．
12: 組織 A は $ID'_{A|D}$ を秘密鍵 c でハッシュ化，$X_{A|D,i}$ を秘密鍵 A で暗号化を行い，$Enc(T'_{A|D})$ を得る．
13: 組織 A は組織 B に $Enc_A(T'_{A|D})$，$(ID'_{B|D})^{bc}$ を送る．
14: 組織 B は $(ID'_{B|D})^{bc}$ を秘密鍵 b でハッシュ化を行い，$(ID'_{B|D})^c$ を得る．
15: 組織 B は $Enc_A(T'_{A|D})$ と $(ID'_{B|D})^c$ から共通要素列 $Enc_A(T'_{A\cap B|D})$ を得る．
16: 組織 B は $Enc_A(X'_{A\cap B|D,i}, X'_{B,i})_{i=1,...,n}$ を得る．
17: 組織 B は自身の持つ各レコード $X_{B,i}$ に基づき，対応する $Enc_A X'^{A\cap B|D}_{A,i}$ をカテゴリ分けする．
18: 組織 B は各カテゴリ毎に対応する $Enc_A X'^{A\cap B|D}_{A,i}$ を総和し，集計テーブル $Enc_A(T_C)$ を得る．

すきなくだもの	いぬ	ねこ			
りんご	123人				
みかん	38人				
ドリアン	13人				
:	:				
すいか	41人	30人	9人	...	19人

こんな表ができる
(クロス集計表)

めでたし、めでたし…？？

「要件2」は満たせたけど…

「要件1」は？

要件1:
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

「適切にプライバシーが保護された統計情報」って？

要件1:
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

	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

集計しただけじゃダメなの？

「適切にプライバシーが保護された統計情報」って？

要件1:
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

	いぬ	ねこ	くじら	…	らいおん
りんご	123人	212人	16人	…	97人

単に集計しただけだと、安全とは言えない

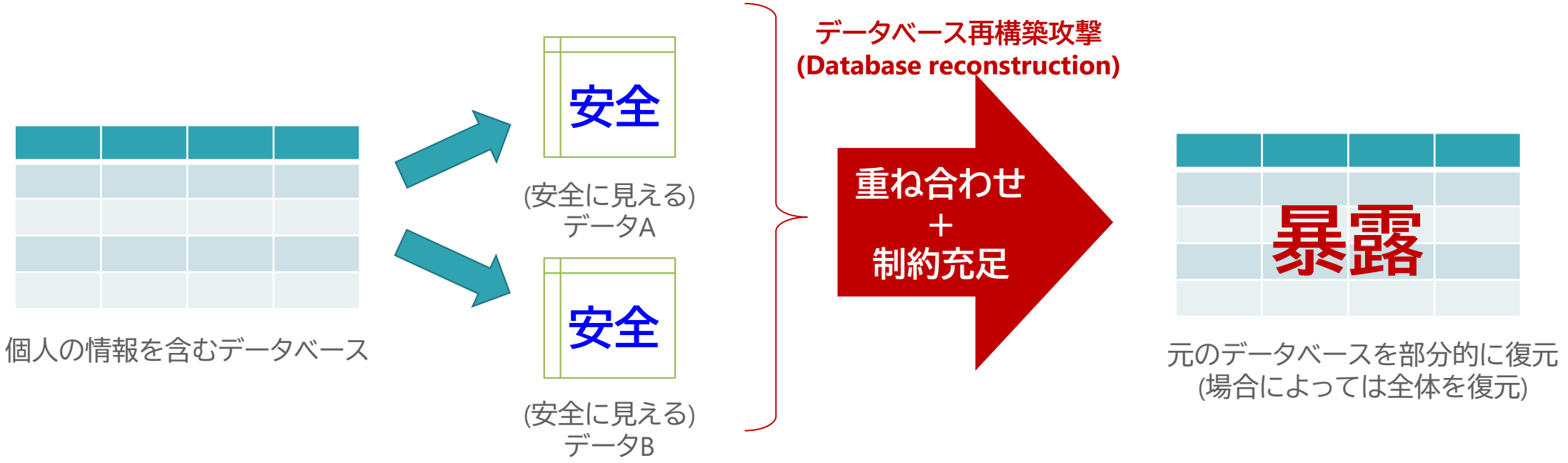
たとえば「データベース再構築攻撃」と呼ばれるこわい攻撃が簡単に成立する

集計しただけじゃダメなの？

データベース再構築攻撃 (Database reconstruction)

あるデータベースから生成された (一見して安全に見える)
データを重ね合わせることによって制約充足問題を構築し、

その問題を解いて元のデータベースを復元することにより、
データに含まれる**個人のプライバシーを暴露**する攻撃。



参考: アメリカ国勢調査局による再構築攻撃の適用実験

米国の2020年国勢調査 (Census) の実施にあたり、2010年国勢調査の集計表に**再構築攻撃を実際に適用**してみたところ…

- 再構築攻撃の適用により、**米国民の 46% (約 1.44 億人) について、居住ブロック、性別、年代、人種、民族が復元**された。
  - 年齢に 1 歳の誤差を許すと 71% が復元された。
- その結果を一般に入手可能な市販データと照合することによって、**約 5,200 万人分のレコードが再識別 (個人特定)** された。
  - これは米国民の約 17% に相当。

「理論上のリスクから対策が求められる課題に変化した」として 2020 Census からの差分プライバシーの導入を決定。

元データを「いじれる」状況だともっと簡単に攻撃できる

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

この数字に着目
(123人)

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

安全に見える

一人消してみると...

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	
S さん	
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
⋮	

Pさん (犬がすき) のレコードを消す

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	**122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

これも安全に見えるけど...

一人消してみると…

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	ドリアン
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

Pさん (犬がすき) のレコードを消す

数字の変化に着目
(123人→122人)

P さんの好きなくだものが「りんご」だと B 社に知られてしまう

要件1:「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」を満たせない

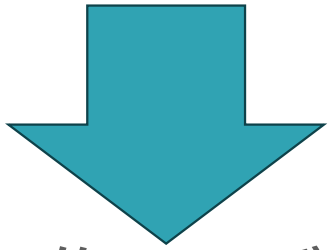
すきなくだもの	いぬ	ねこ	くじら	…	らいおん
りんご	122人	212人	16人	…	97人
みかん	38人	66人	11人	…	43人
ドリアン	13人	6人	5人	…	7人
:	:	:	:	…	:
すいか	41人	30人	9人	…	19人

これも安全に見えるけど…

単に「集計しただけ」だと安全とは言えない

計算の過程が安全だったとしても、集計結果からプライバシーが漏れるリスクがある。

特に、**データベース再構築攻撃のリスク**に対処する必要



(広く知られている中では) 唯一、データベース再構築攻撃に対処可能とされる**「差分プライバシー」を適用**

(復号前の集計結果 (の暗号文) に対して適用する必要があることに注意)

差分プライバシー (differential privacy) とは

任意の隣接したデータベース D1 と D2 ($D_1, D_2 \in \mathcal{D}$) に対し、ランダム化関数 $\mathcal{A}: \mathcal{D} \to \mathcal{R}$ が下記を満たすとき、$\mathcal{A}$ は ε-

ただし、 部分集合である

今回はパスします

さすがに時間が足りないので...

$$\Pr[\mathcal{A}(D_1) \in S] \leq e^{\varepsilon} \times \Pr[\mathcal{A}(D_2) \in S]$$

「差分プライバシー」で検索すると、いろいろ解説が見つかると思います。
(私の解説記事もいくつかトップページにありますのでよろしければ)

まとめと補足 (1/2)

秘匿クロス統計は、異なる組織のデータを安全に融合して、それぞれのデータ単体からでは得られないような知見を得るために開発された技術です。

これを安全に実現するためには、

- **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 (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)**

の 2 つの安全性要件を満たす必要があります。

まとめと補足 (2/2)

これらの安全性要件を満たすために、秘匿クロス統計では、

1. **非識別化処理** (入力データの個人識別性を除去する)
2. **集計処理** (秘匿照合集計を用いてクロス集計表を作成する)
3. **秘匿処理** (暗号状態のまま差分プライバシーに基づいて集計結果のプライバシーを保護する)

の三段階処理を実施しています。

ただし、たとえば Semi-honest 安全性と Malicious 安全性のギャップに対処するための TEE の適用や、TEE 間の安全な相互認証、差分プライバシーの適用による有用性劣化の対処、プライバシー損失予算 (ε) の節約や配分の最適化など、安全性や実用性を確保するための細々とした話は他にもいろいろありますが、**基本的なポイントは今日お話ししたところ**になります。

つまり、これを「安全に」実現するために、

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

お互いに見せあうことなしに…
(非識別化処理 → 集計処理 → 秘匿処理)

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。
(安全なクロス集計表)

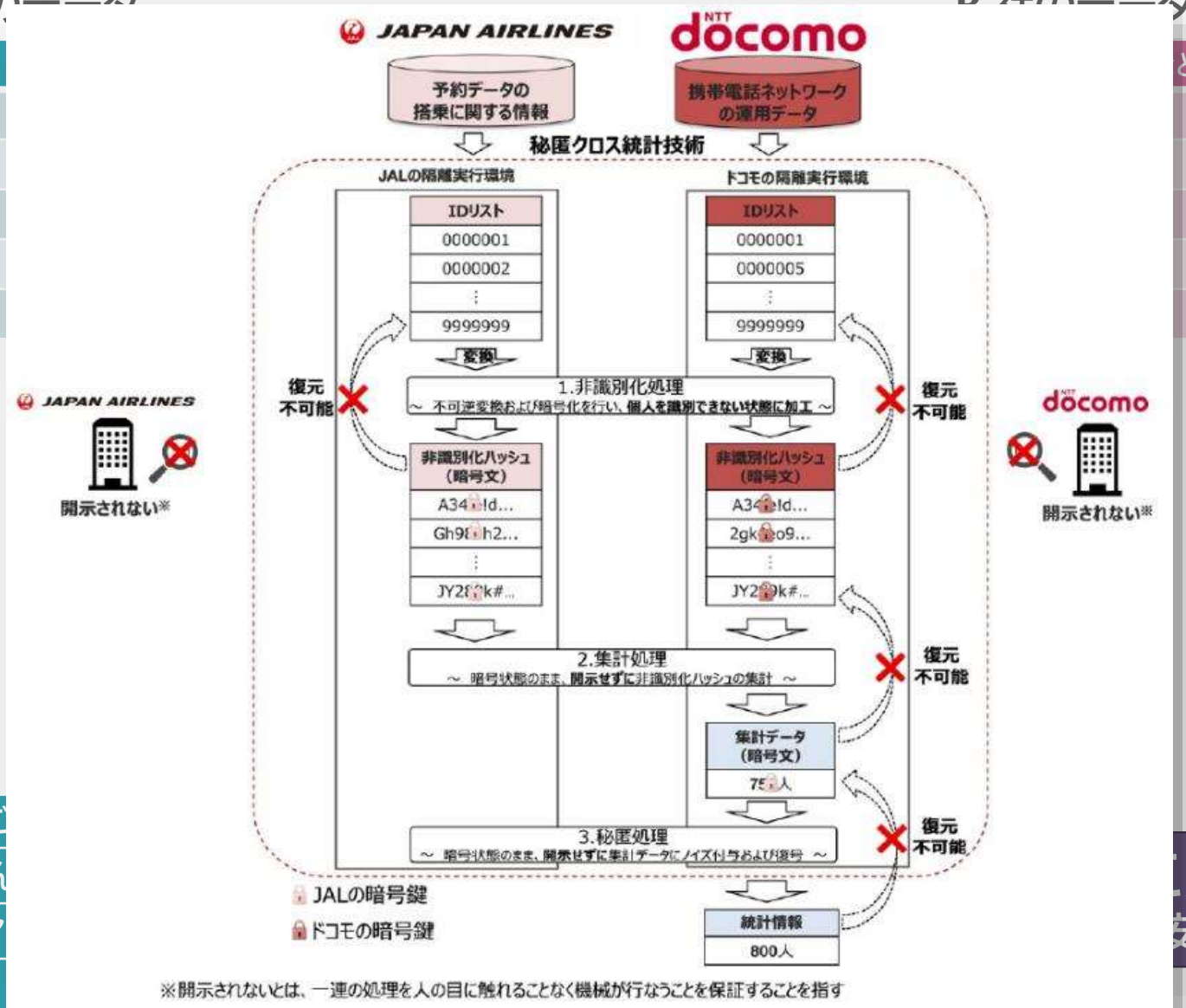
こういう「めんどくさい」ことをしているのが秘匿クロス統計

A 社のデータ

なまえ
P さん
Q さん
R さん
S さん
T さん

B 社のデータ

どうぶつ



すきなくだもの

りんご					
みかん					
ドリア					
:					
すいか	41人	30人	9人	…	19人

んな表を作る。
安全なクロス集計表)

2022年10月からのスムーズな航空輸送に向けた実証に加え、

JAPAN AIRLINES　JALCARD　docomo

トピックス

2022年10月20日
日本航空株式会社
株式会社ジャルカード
株式会社 NTT ドコモ

JAL、JAL カード、ドコモが、顧客体験価値向上と社会課題の解決に向けて、「秘匿クロス統計技術」を用いた企業横断でのデータ活用の実証実験を開始

～各社が保有するデータを相互に開示せず作成した統計情報を活用する国内初の取り組み～

日本
株式会
株式会
JAL、
（以下

本実
本技術
状態
機械
ことで
国内
一部

JAPAN AIRLINES　docomo

国内線航空券の予約データの搭乗に関する情報

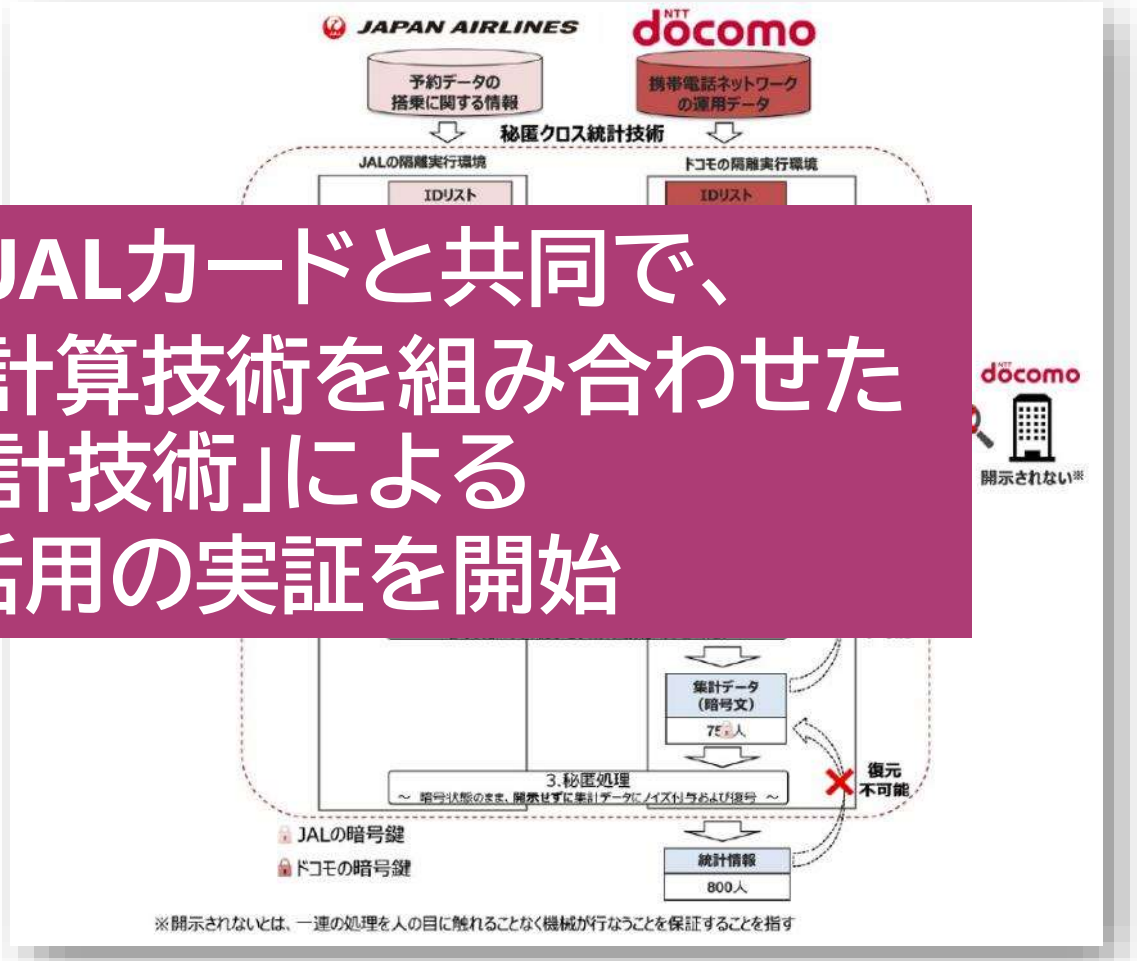
携帯電話ネットワークの運用データ

秘匿クロス統計技術により、
個人を識別できない状態に加工したうえで、
各社が保有するデータを開示せずに統計情報を作成

空港ご到着前の各時点でのお客さまの移動状況に関する統計情報

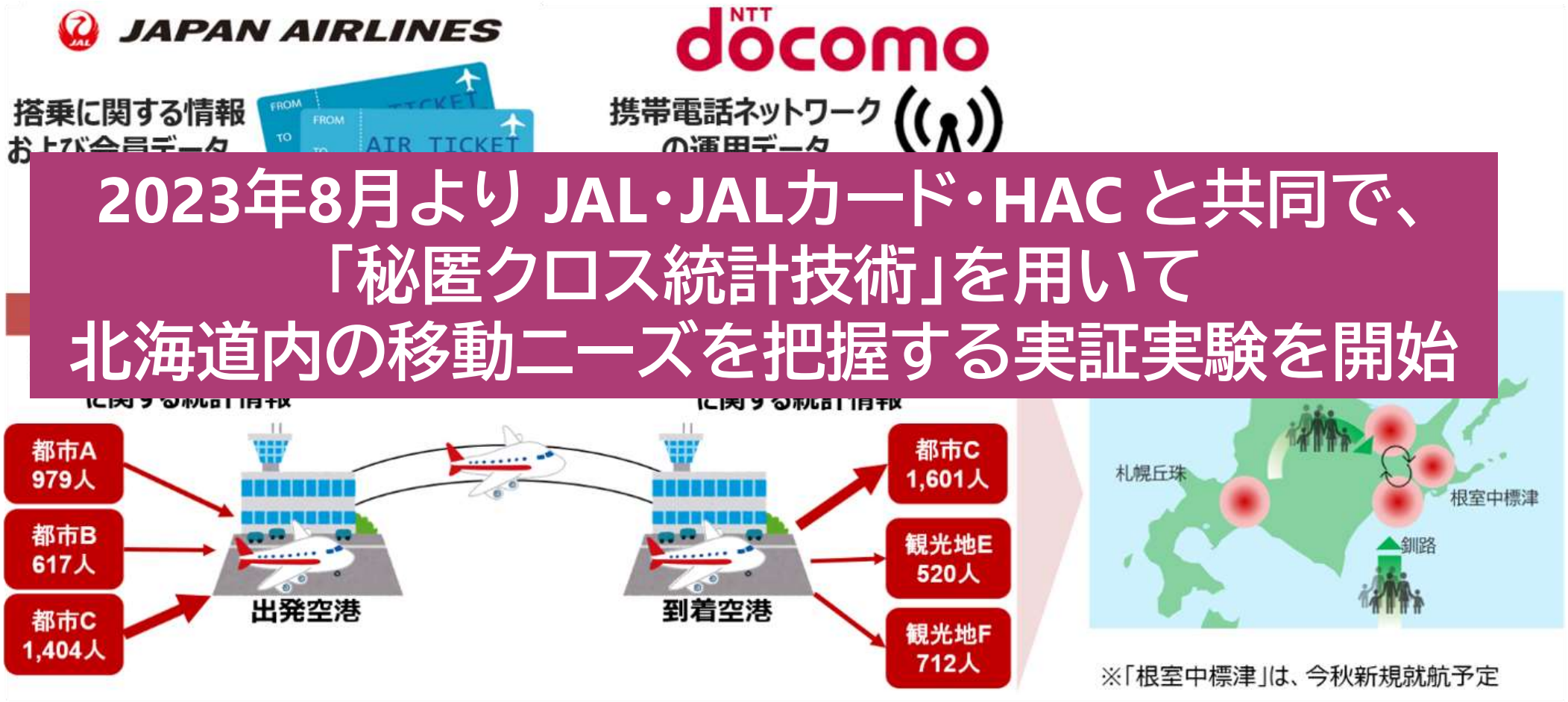
	便出発前日			便出発60分前			便出発40分前			便出発20分前		
	居住地域	その他	空港周辺	居住地域	その他	空港周辺	居住地域	その他	空港周辺	居住地域	その他	空港周辺
11月の午前便の搭乗客	2,898	2,001	101	1,610	3,085	305	252	1,206	3,542	79	86	4,835
11月の午後便の搭乗客	3,898	2,901	201	2,610	3,985	405	352	2,106	4,542	60	87	5,855

※回は一例であり、単位は人です。



**2022年10月に JAL・JALカードと共同で、
差分プライバシーと秘密計算技術を組み合わせた
「秘匿クロス統計技術」による
企業横断データ活用の実証を開始**

2023年8月からは北海道の地域活性化に向けた実証を開始



今後の展望

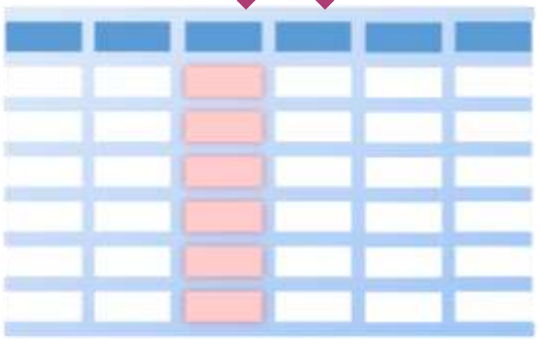
秘匿クロス統計は、**いままでできなかった組織間データ連携**を、**プライバシー保護技術を組み合わせることにより可能とする**技術です。

まだその活用は始まったばかりであり、これからさらに実績を重ねるとともに、**新たなデータ分析の方法論**も確立していく必要があると考えられます。

※トライアンドエラーに基づく既存のデータ分析フレームワークが使い物にならない

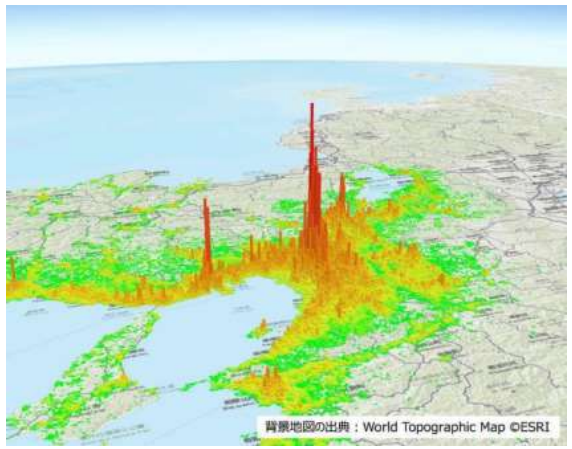
今後、これらの技術改良に加え、**本技術に基づく社会課題解決**の有用性検証をさらに進めることにより…

この世界を「あたりまえ」に。



異なる組織がそれぞれ保有する
複数のデータ

**いわゆる「名ばかり同意」に依存しない、
安全性が保証された組織間データ活用**



(いままで作れなかった)
安全な統計情報

(いままでできなかった)
有益な社会価値の創出