オンサイト施設における個票データの利用報告と利便性向上のための提案

千葉亮太（一橋大学）
白川清美（一橋大学）

本日の発表内容

- オンサイト施設の試行運用の利用報告
- 秘密計算システムに関する研究

【報告】オンサイト施設の試行運用

- 「申請から結果持ち出し」までの作業

1. 全国消費実態調査の独自集計を平成２６年個票データを用いて集計
2. 独自集計：全国消費実態調査を用いた世帯主の年齢各歳別の家計収支
   (http://rcisss.ier.hit-u.ac.jp/Japanese/database/special.html)
3. 前提条件
   – 平成２１年集計用プログラムを使用
     ➢平成２６年集計のために一部プログラムを修正

3

調査情報提供の申請

1. 申請項目の一部が削除
   – 利用場所・データ管理
2. 利用項目と統計表が不要
   – 調査票を参照、項目の列挙不要
     ※全変数の申請が容易
   – 作成する統計表の事前作成は不要
     ※従前は事前審査の統計表の審査が厳しく、申請に時間が掛かった

✓ 申請時に掛かる労力・コストが減少
✓ 研究室がない研究者には朗報

※平成２６年の申請は、平成２１年の申請書を基に作成した。

Ｈ21申請項目		Ｈ26申請項目	
1	統計調査の名称	1	統計調査の名称
2	利用目的	2	利用目的
3	利用者の範囲	3	利用者の範囲
4	オンサイト利用の有無	4	利用するオンサイト施設
5	利用する調査票情報の名称及び範囲	5	利用する調査票情報
6	利用する調査事項	6	利用する情報
7	利用方法	7	利用方法
8	利用期間	8	利用期間
9	利用場所、利用環境、保管場所及び管理方法		
10	結果の公表方法及び公表時期	9	結果の公表方法及び公表時期
11	転写書類等の利用後の処置		
12	著作権	10	著作権
13	連絡先（事務担当者）	11	連絡先（事務担当者）

4

事前準備

1. 環境整備（必要なSWのインストール）
   - Officeなどのソフトは事前に導入済み
   - 導入されていないSWのインストールを依頼
     - ※ライセンスの確認・手順書の用意も必要
     - ※オフラインでのインストールが可能なものに限る
   - データ（プログラム）の持ち込みを依頼
2. 動作確認等
   - 提供環境で準備されていたSASが起動せず、調整を要した
   - Visual Studioのコンポーネントを追加で依頼

✓メディア送付・導入・確認作業の時間

5

分析作業

- 演算時間（出力されたログ比較）

	平成21年 ログ				平成26年 ログ				増加割合(%)	
	2人以上世帯		単身世帯		2人以上世帯		単身世帯			
	52716行	649変数	4343行	649変数	51768行	639変数	4654行	639変数	2人以上	単身
DATA処理	処理時間 3.62 秒		処理時間 1.40 秒		処理時間 21.55 秒		処理時間 3.82 秒		595	273
	CPU時間 3.21 秒		CPU時間 1.03 秒		CPU時間 5.79 秒		CPU時間 0.78 秒		180	76
DATASETS処理	処理時間 1.41 秒		処理時間 0.29 秒		処理時間 0.55秒		処理時間 0.37 秒		39	128
	CPU時間 0.70 秒		CPU時間 0.21 秒		CPU時間 0.10 秒		CPU時間 0.16秒		14	76
1表作成 (DATA処理)	処理時間 1.17 秒		処理時間 0.26 秒		処理時間 27.73 秒		処理時間 2.51 秒		2370	965
	CPU時間 1.09 秒		CPU時間 0.18 秒		CPU時間 2.94 秒		CPU時間 0.33 秒		270	183
1表作成 (TABULATE処理)	処理時間 1.15 秒		処理時間 0.14 秒		処理時間 12.12 秒		処理時間 3.16 秒		1054	2243
	CPU時間 0.85 秒		CPU時間 0.09 秒		CPU時間 5.33 秒		CPU時間 2.34 秒		627	2600
sas7bdatサイズ	282996kb		23688kb		295436kb		26869kb			

※平成21年のSASのバージョンはSAS 9.4　平成26年はSAS University Edition
※プログラムは、レイアウト位置を修正したのみ
✓ 集計処理はローカル環境よりオンサイト環境の方が時間が掛かる

6

分析作業２

1. 情報収集・利用が不便である
   - インターネット等からの検索
   - 標準以外のパッケージ・ソースコードの利用等
2. オンサイト施設内のみの作業
   - プログラムの修正等、結果持ち出し前の論文作成作業
3. SAS University Editionの使用感
   - ブラウザベースのため、表示を多くすると重い

✓分析中に必要となったデータ・情報をオンサイト環境で利用する方法

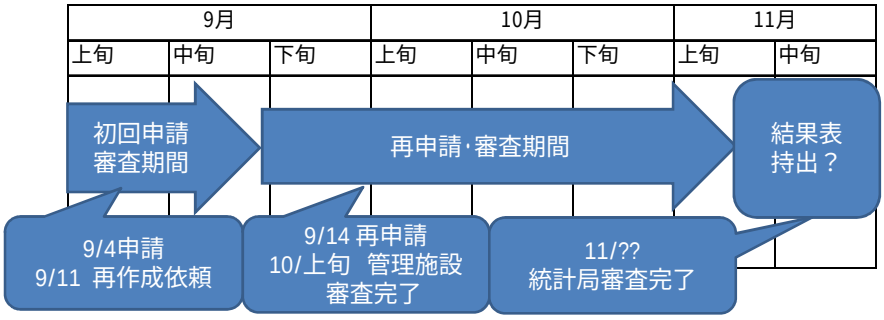
✓作業領域が狭い（1280×1024）

✓データ処理時の性能は十分か

7

結果の持ち出し審査

1. 中央データ管理施設の審査及び統計局の審査
   - 今回の審査状況



✓結果表が手元に戻るまで2ヶ月以上（全4表）

✓審査がいつ終わる（手元に来る）のか分からない

結果の持ち出し審査

2. 持ち出し申請書の項目
   - チェックリスト形式
     - 10以上の調査客体から算出した値
     - 最大値・最小値・グラフは不可　　等
3. 分析結果審査資料
   - 個票データを用いた審査表の作成
     - 各セルの加重なしの度数及びその構成比
     - 各セルにおいて最も大きく寄与する調査客体の占める割合

✓ 項目の内容は持出用資料の作成の参考

✓ 結果表・審査表の修正指摘時にファイルがない

9

その他

1. セキュリティ・利用要件の担保
   - 更新やインターネット不接続等への対応
   - 不正使用への対応
2. 時間・場所等の制約
   - 研究室（国立キャンパス）とオンサイト施設（小平キャンパス）が遠い場合・開設時間等への対応
3. 持ち込み・持ち出し
   - 符号表などの事前準備のための資料
   - 即時性・経費等への対応
4. オンサイトだけには限らない課題
   - 利用期限が過ぎた査読審査等への対応

10

オンサイト施設の試行運用のまとめ

1．申請は簡易に、ただし結果持ち出しに時間が掛かる
  - 分析目的・利用者の要件などは従前同様
  - 様々なデータ分析が可能
  - 論文等の投稿には余裕が必要

2．個票データの適正利用可能
  - 従前は結果表の様式の審査、今回は分析の結果を踏まえた審査
  - 第三者（中央データ管理施設・統計局等）が成果物を審査

3．分析環境に慣れが必要
  - 利用者の環境とは異なるため、事前準備が必要
  - データ・PG等の送受に時間が掛かる
  - 大規模なデータ集計に対応可能か否か（スペック等）

✓ オンサイト施設のみの利用を標準にする方が安全面ではよい
  - オンサイト・磁気媒体利用の選択可能であれば、磁気媒体の方が使い勝手がよい
  - オンサイト施設が近くにないと不便
  - オンサイト施設の初期投資・維持費用（施設要件等の遵守も大変）

11

これからの個票データ利用に向けて

- 媒体提供における課題
  - 安全面における課題　等
- オンサイト施設の継続的な改善
  - 利用者からの意見などからの改善
  - 法制度の変更で利用要件の緩和　等

✓オンサイト利用をより便利に利用する手法の一つとして…

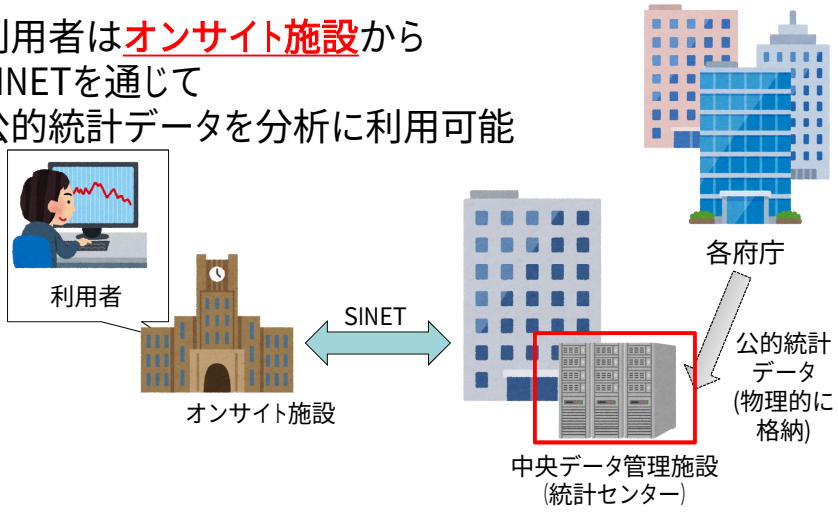
12

秘密計算システムに関する研究

国立大学法人 一橋大学 経済研究所
NTTセキュアプラットフォーム研究所

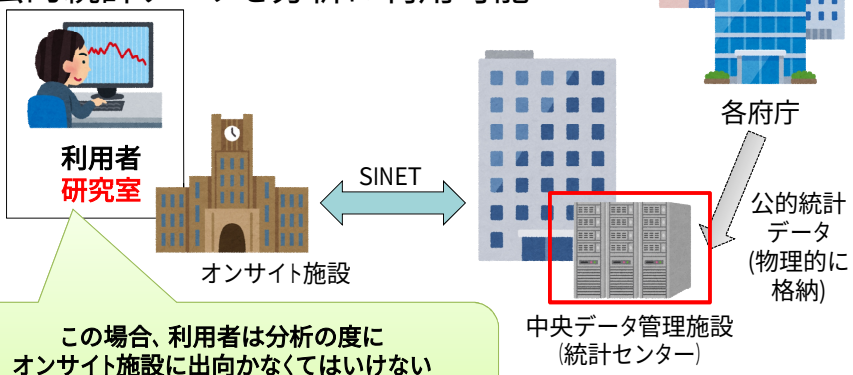
公的統計データのオンサイトの利便性の向上

利用者は**オンサイト施設**から
SINETを通じて
公的統計データを分析に利用可能



公的統計データのリモートアクセス利用

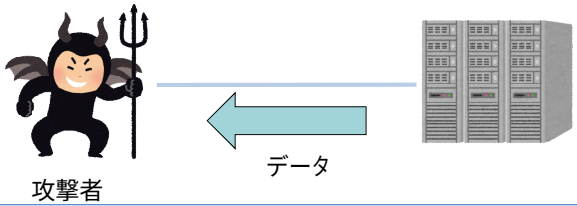
利用者はリモートアクセス施設から
SINETを通じて
公的統計データを分析に利用可能



この場合、利用者は分析の度に
オンサイト施設に出向かなくてはいけない
→利便性向上のために利用者の拠点からの
分析を実現することはできないか？

利用者拠点からのアクセス方法
→インターネット経由でのアクセス

公的統計データ分析者の利便性は向上する反面
不正アクセス等によるデータ漏えいのリスクが増加



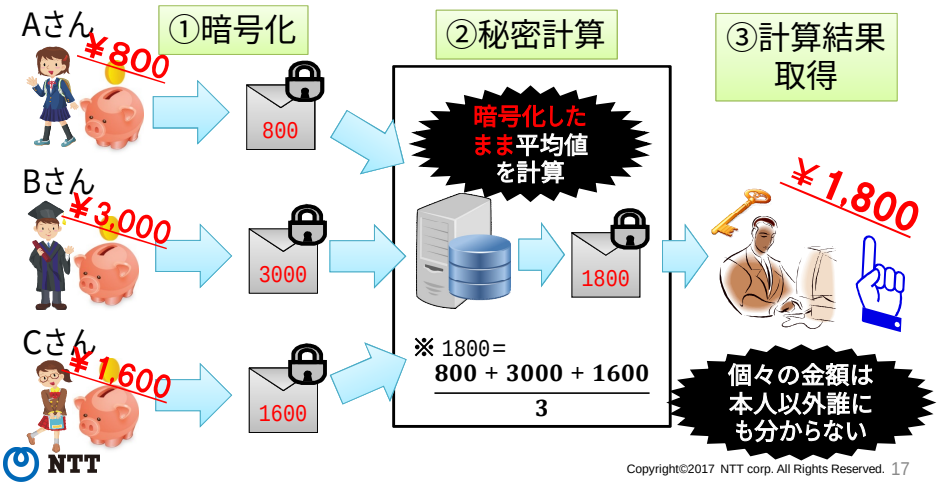
データ漏えいリスクに対する対策として、
「データの漏洩防止」「漏洩時のデータ保護」を考える

→データを秘匿化したまま処理する事ができる
「秘密計算技術」を適用することで
“公的統計データの安全性を確保しながら”
“分析者拠点からの公的統計データの分析”を実現する

NTT

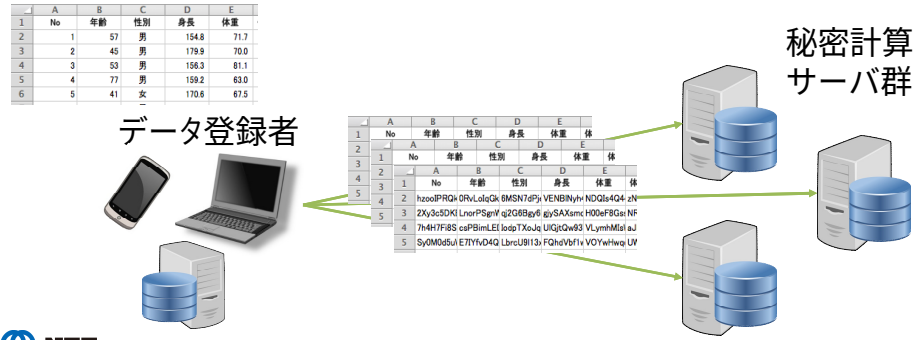
秘密計算とは

データを**暗号化したまま**各種計算ができる技術
- 3人の平均貯金額を求める場合の処理イメージ



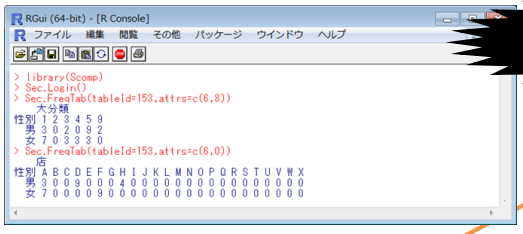
NTT秘密計算システム：データ登録

- CSV形式のデータファイルに対応
- Windows / Linux PC から登録可
  - iOS 対応版も開発中
- 「秘密分散」技術を使ってデータをセル単位で暗号化
  - 3つのデータに分割し、別々のサーバに送信



NTT秘密計算システム：データ分析

- フリーの統計ソフト‘R’から秘密計算で統計処理を実行可能
  - データ登録者が所望の統計処理をリクエスト
  - 秘密計算サーバ間で通信しながら処理
  - データ分析者に統計処理結果の暗号化データが送られる→復元



秘密計算サーバには、
計算中も含め、暗号化されたデータしか存在しない

秘密計算
サーバ群

データ分析者

分析者の端末には
機微なデータが一切
入らない

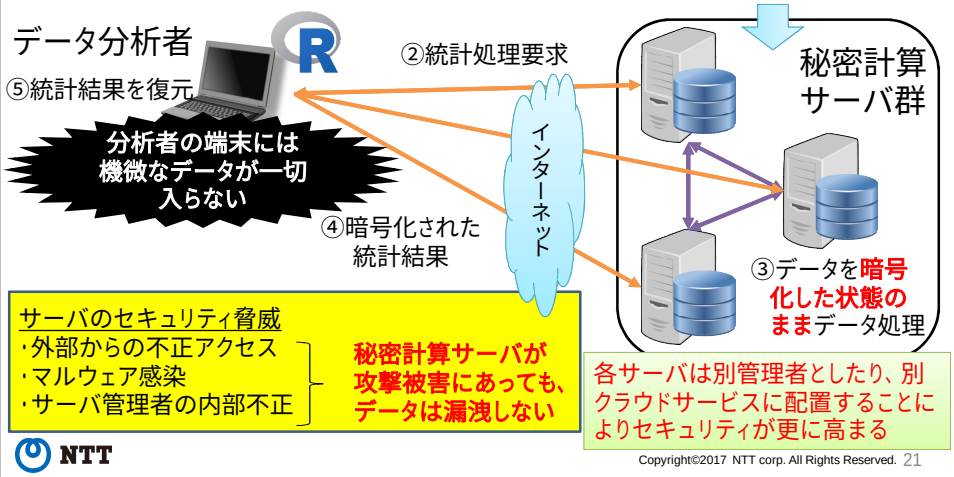
NTT秘密計算システム：データ分析

‘R’のコマンド関数として実行できる秘密計算　(現状版)

- 基本統計量
  - 平均、分散
- 集計
  - 単純集計、二重クロス集計、三重クロス集計 (カテゴリ値のみ)
- 補助演算
  - 条件フィルタ (等号判定、大小比較)
    - 使用例：65歳男性の血糖値を抽出
  - 二属性積和
  - 件数取得
  - ランダムシャッフル
- 上記演算結果をRスクリプトで加工し新たなコマンドを定義
  - 共分散、t検定、Kaplan-Meier法
  - 線形単回帰、線形重回帰、主成分スコアの算出、多項式回帰、主成分分析

想定ユースケース①

オンライン統計分析のサーバセキュリティ強化



サーバのセキュリティ脅威
・外部からの不正アクセス
・マルウェア感染
・サーバ管理者の内部不正

秘密計算サーバが攻撃被害にあっても、データは漏洩しない

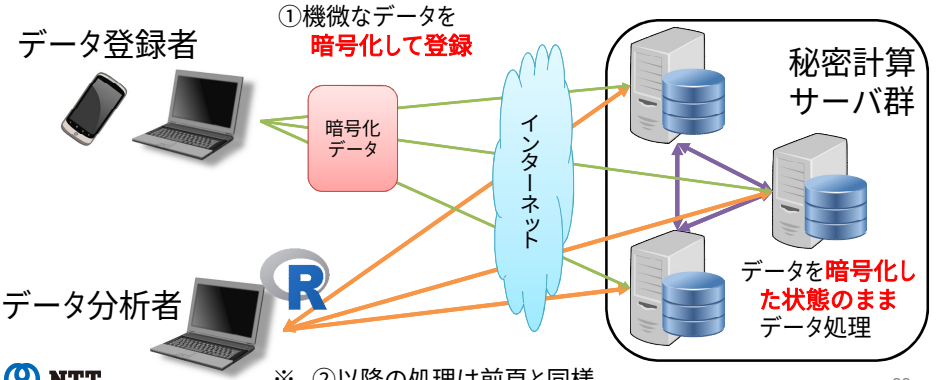
各サーバは別管理者としたり、別クラウドサービスに配置することによりセキュリティが更に高まる

NTT

想定ユースケース②

機微なデータをより収集しやすくする仕組みの提供

スマホやセンサデバイス等に格納された機微なデータを暗号化してから登録
→データを統合（Append）して秘密計算を実行し、統計データを取得

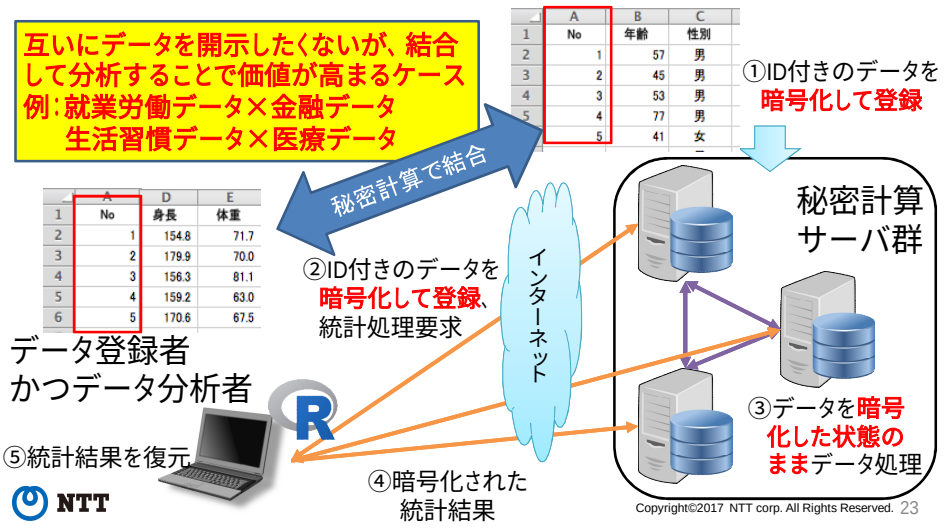


※　②以降の処理は前頁と同様

NTT

想定ユースケース③

ID付きのデータを安全に結合（Join）してクロス分析



 23

まとめ

- オンサイトをより便利とするための方法
  - ✓分析方法に応じた利用方法の提案
    - ➢オンサイト（実データを用いた分析）
      - –データ全体の分析・分析の方針を決定
      - –個票データ自体の確認が必要な分析
      - –持ち出し審査のための審査表作成
    - ➢オフサイト（暗号化されたデータを用いた分析）
      - –分析方法・方針が定まった分析
      - –個票データを確認せずとも可能な分析
      - –結果の取得のみを行う

24

ご清聴ありがとうございました。

ミクロデータの利用に関するご質問等は，
すべて下記のメールアドレスにお願いします。

micro＠ier.hit-u.ac.jp

窓口開設期間･時間

4月1日～2月末日(土日，祝日，年末年始の期間，
その他お知らせで事前に周知する期間を除く)
午前の部　10:00～12:00
午後の部　13:00～17:00

25