2021年度の活動紹介（公的ミクロ事業）

公的ミクロデータにおけるデータ構造化とその利用に関する諸問題

社会データ構造化センター / 統計数理研究所
山下智志

第6回社会データ構造化シンポジウム
2022年2月18日

公的ミクロデータグループの課題

統計数理研究所　リスク解析戦略研究センター　データ基盤プロジェクト

1. 政府統計など公的ミクロデータの学術利用の促進

1-a　オンサイト拠点の運営と実験
1-b　公的統計ミクロデータ研究コンソーシアムの運営・管理
1-c　総務省、統計センターなど、政府への業務支援、技術供与
1-d　公的ミクロデータ利用の海外事例調査

2. ミクロデータ活用のための基礎技術の開発

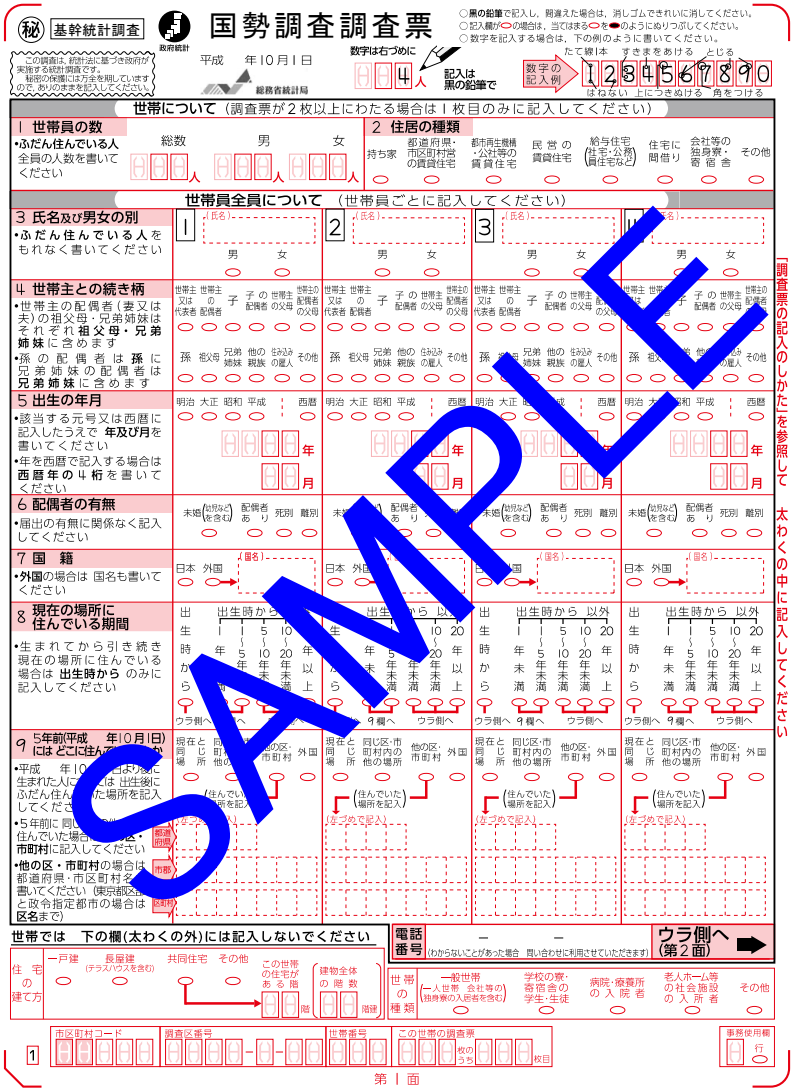
2-a　プライバシー保護のための統計的セキュリティ
2-b　匿名加工データの作成技術開発

3. ミクロデータ構造化のための統計的技術開発

3-a　欠損値補間およびデータ結合のための統計的方法論
3-b　複数のデータベースに対する統計的マッチング
3-c　法人ミクロデータの収集とリスク分析
3-d　精度と規模が極端に違うデータベースの結合と利用
　　その社会実験（不動産データ）

公的ミクロデータとは？

- 国が行う統計調査で作成される調査客体単位のデータ



オンサイトで利用可能な調査一覧

（令和２年11月現在）

（内閣府）
- 企業行動に関するアンケート調査
- 青少年のインターネット利用環境実態調査

（総務省）
- 通信利用動向調査
- 国勢調査
- 住宅・土地統計調査
- 就業構造基本調査
- 個人企業経済調査
- 労働力調査
- 科学技術研究調査
- 家計調査
- 全国消費実態調査
- 社会生活基本調査
- 経済センサス-基礎調査
- 経済センサス-活動調査
- 家計消費状況調査
- サービス産業動向調査
- 小売物価統計調査

（財務省）
- 法人企業統計調査

（文部科学省）
- 学校基本調査
- 学校教員統計調査

（厚生労働省）
- 賃金構造基本統計調査
- 人口動態調査
- 就労条件総合調査
- 薬事工業生産動態統計調査
- 医薬品・医療機器産業実態調査
- 国民健康・栄養調査

（経済産業省）
- 工業統計調査
- 経済産業省企業活動基本調査
- 外資系企業動向調査
- 情報通信業基本調査
- 経済産業省生産動態統計調査
- 商業統計調査
- 商業動態統計調査
- 特定サービス産業動態統計調査
- 特定サービス産業実態調査
- スポットＬＮＧ価格調査
- 工場立地動向調査
- 容器包装利用・製造等実態調査
- エネルギー消費統計調査
- 経済センサス-活動調査

（経済産業省（続き））
- 石油製品需給動態統計調査
- ガス事業生産動態統計調査
- 経済産業省特定業種石油等消費統計調査
- 知的財産活動調査
- 模倣被害実態調査
- 中小企業実態基本調査
- 海外事業活動基本調査
- 海外現地法人四半期調査
- 情報処理実態調査

（環境省）
- 産業廃棄物排出・処理状況調査
- 環境にやさしい企業行動調査
- 水質汚濁物質排出量総合調査
- 環境経済観測調査
- 食品廃棄物等の発生抑制及び再生利用の促進の取組に係る実態調査
- 家庭からの二酸化炭素排出量の推計に係る実態調査　試験調査
- 家庭部門のCO2排出実態統計調査

７府省　５６調査

ミクロデータ（調査票情報）とは？

総務省など国の行政機関で実施した統計調査の結果は、ホームページ（政府統計の総合窓口 e-Stat）等を通じて広く一般の方にご利用いただいています。

このような調査結果の提供に加え、公益性のある学術研究等にご活用いただくため、**調査対象の秘密の保護を図った上**で、**世帯単位や事業所単位といった集計する前の個票形式のデータ**を提供しています。

この個票形式のデータを**ミクロデータ（調査票情報）**と言います。

ミクロデータ（調査票情報）を用いることで、研究者の方々は、より自由で多様な分析を行うことが可能となるため、新たな発見につながることが期待されます。

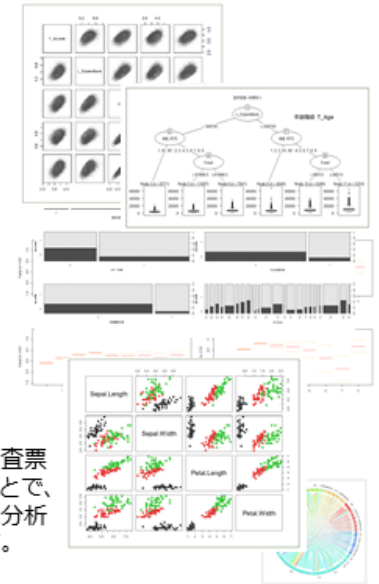
行政機関による集計・公表

行政機関は、ミクロデータ（調査票情報）を集計して、調査結果を作成しています。
調査結果は、「政府統計の総合窓口（e-Stat）」等を通じて公表・提供しています。
https://www.e-stat.go.jp/

ミクロデータ（調査票情報）のイメージ

Weight	Y_Income	L_Expendi	Food	Housing	LFW	Furniture	Clothes	Health	Transport	Education
895.2667	3917	201649	47756	16028	9652	6702	8088	726	21546	0
895.2667	6675	166381	34054	7416	26313	17062	6989	7637	20773	0
895.2667	6706	259736	84501	1927	10082	6741	5090	11015	53372	0
895.2667	2790	114511	41664	730	22358	5413	1205	5049	17411	0
895.2667	2577	193505	56981	3779	28747	4812	4243	751	16435	0
895.2667	3452	152109	34924	3418	8131	4164	6970	4247	47698	0
895.2667	3233	136900	49956	203	15429	3659	22843	4365	8684	0
895.2667	9252	192439	68882	2832	23042	2598	5714	2052	37006	0
895.2667	2359	138415	53591	753	13072	5140	1786	5416	11593	0
895.2667	2059	79179	32853	14134	7977	3017	2364	1607	7630	0
895.2667	2324	243835	60528	28118	16392	4652	10759	37060	17666	0
895.2667	4524	241539	104433	5253	40637	7711	13833	3853	43003	0
895.2667	4415	207854	95504	7687	13801	19702	7239	3939	9547	0
895.2667	4162	185110	59798	565	12146	14552	12301	2486	26726	0
895.2667	16647	219935	81572	3704	21164	4944	10282	6573	7103	0
977.1795	6760	176625	43112	5013	11780	1697	3008	4836	12295	848
977.1795	6614	130803	32336	3149	14723	21002	12075	3447	24273	140
977.1795	6849	395294	66755	9524	19962	19777	10598	7398	56370	80
977.1795	6813	284803	86655	13479	15121	22966	1134	3649	84535	551
977.1795	6595	260459	46614	1646	15044	6415	16207	3709	76074	69

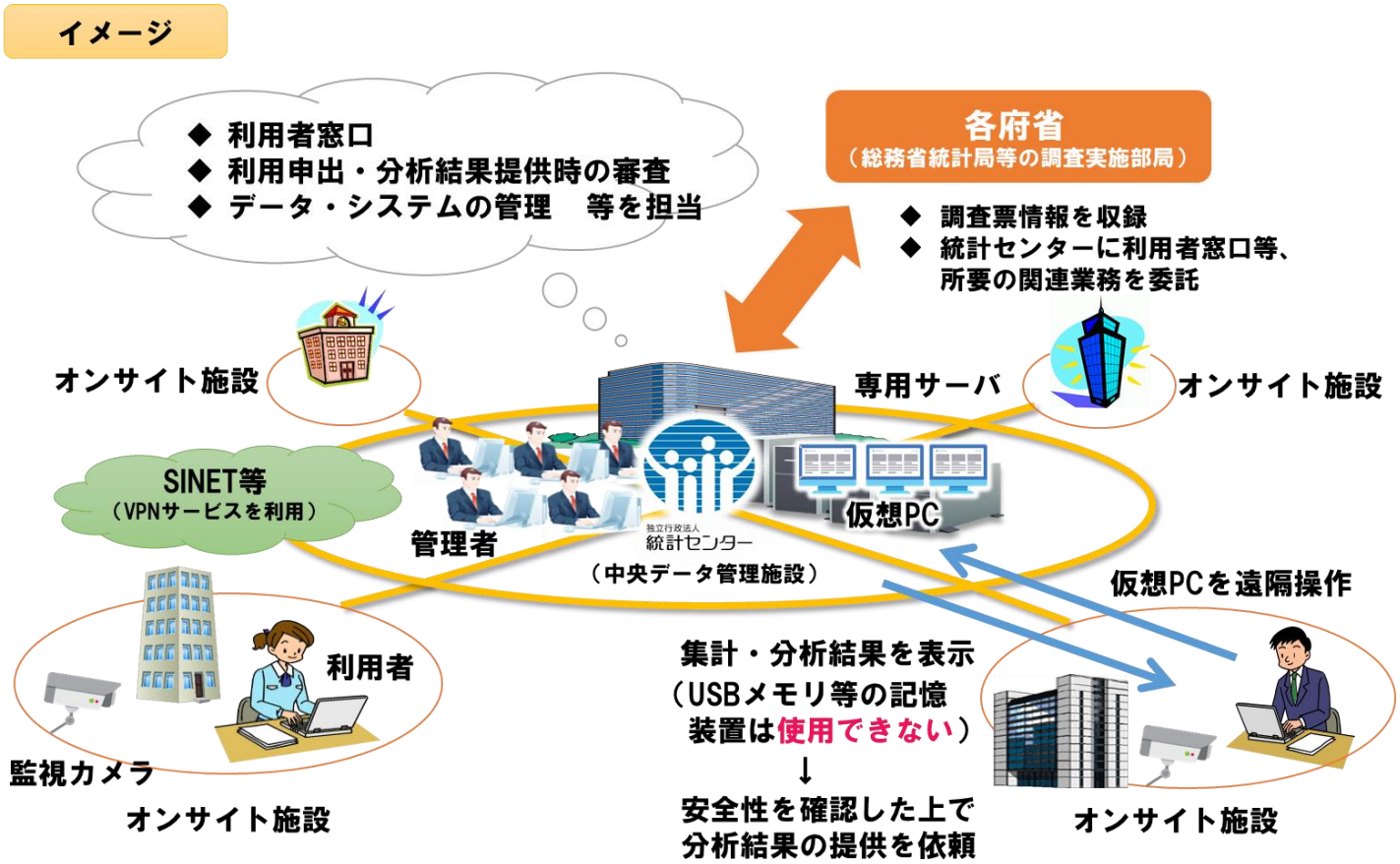
※データは擬似データです



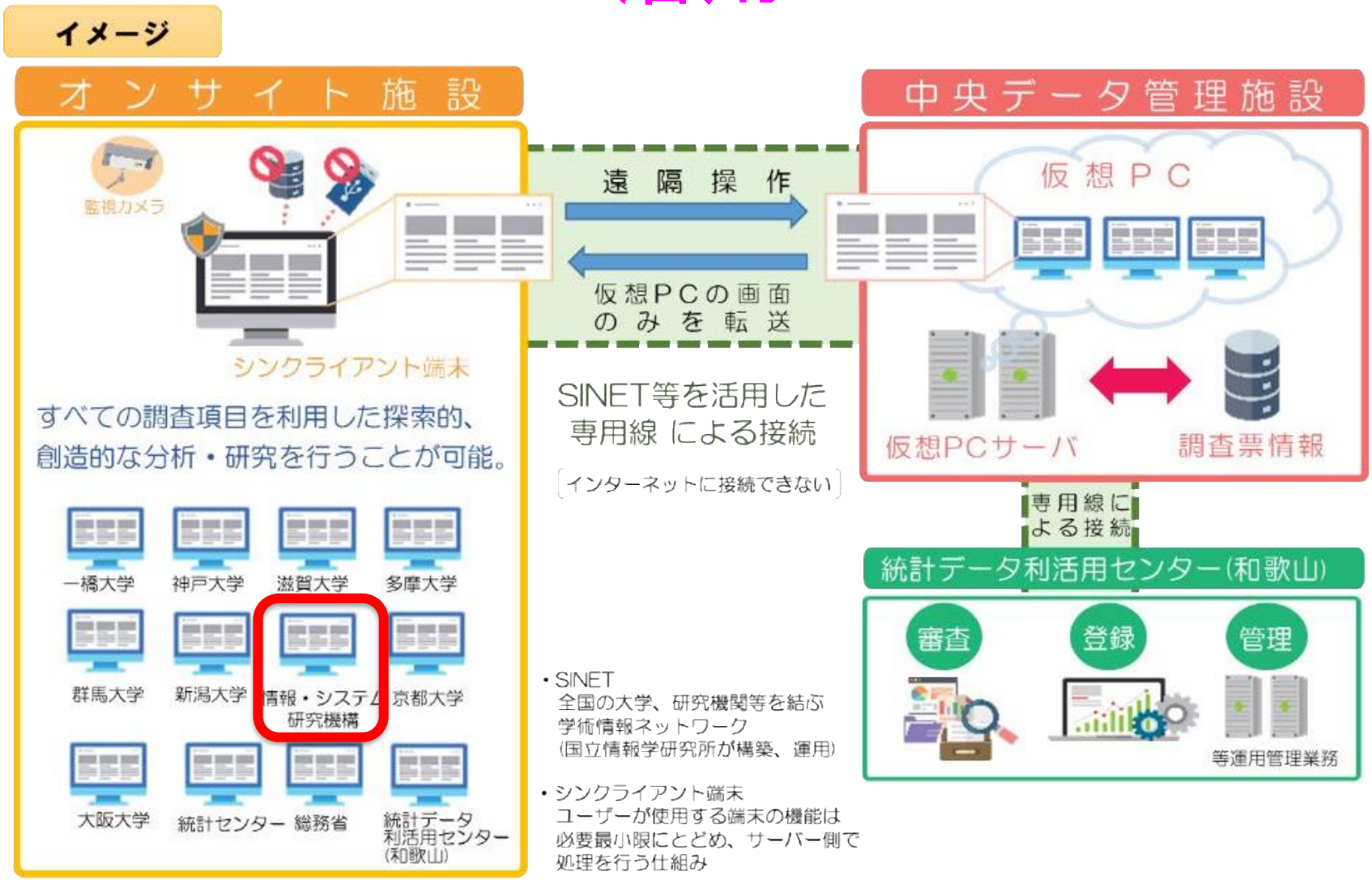
ミクロデータ（調査票情報）を用いることで、より自由で多様な分析が可能になります。

https://www.e-stat.go.jp/microdata/micro

オンサイト利用全体の概要



オンサイト施設におけるミクロデータの活用



（統計データ利活用センター高部勲様のオンサイト説明資料より抜粋）

公的統計ミクロデータ
研究コンソーシアムの活動

公的統計ミクロデータ研究コンソーシアム

- 公的統計ミクロデータの研究利用（二次的利用）を促進するための環境の整備するために2017年に設立
  - 利用手続きの煩雑さ、利用機会の少なさにより、少数の研究者に利用が限定
- 公的統計ミクロデータの研究利用促進に係る学官産連携の推進
- 公的統計ミクロデータ分析の普及・啓発
- オンサイト利用によるミクロデータの研究利用の推進

公的統計ミクロデータ研究コンソーシアム

総務省
政策
統括官室

総務省
統計局

（独）統計
センター

総務省
統計研究
研修所

評議会
議長
情報システム・研究機構
藤井 良一

運営委員会
委員長
統計数理研究所
南 和宏

副委員長
中央大学
伊藤伸介

事務局

プラットフォーム
分科会

データ構造化
分科会

利用普及促進
分科会

研究代表者：椿 広計 統計数理研究所 名誉教授
研究分担者：山下 智志、南 和宏、岡本 基

統計数理研究所 | 一橋大学 | 青山学院大学 | 慶應義塾大学

科学研究費補助金 基盤研究(A)
「政府統計ミクロデータの構造化と
研究利用プラットフォームの形成」

情報システム・
研究機構
データサイエンス
共同利用基盤施設

国際ミクロ統計
データベース
山下 智志、岡本 基、
馬場 康維 名誉教授

（公財）
統計情報
研究開発
センター

評議会　評議委員名簿

- 議長　藤井 良一　情報･システム研究機構 機構長
- 川崎 茂　日本大学経済学部 特任教授
- 北村 行伸　立正大学経済学部 教授
- 玄田 有史　東京大学社会経済研究所 教授
- 佐和 隆光　公益財団法人国際高等研究所 副所長
- 椿 広計　統計数理研究所 所長
- 岡部 寿男　京都大学学術情報メディアセンター 教授
- 松林 洋一　神戸大学経済経営研究所 研究科長
- 杉山 学　群馬大学社会情報学部 教授
- 南 和宏　統計数理研究所 教授
- 山下 智志　統計数理研究所 副所長/教授
- 渡邉 聡 広島大学 高等教育研究センター 教授

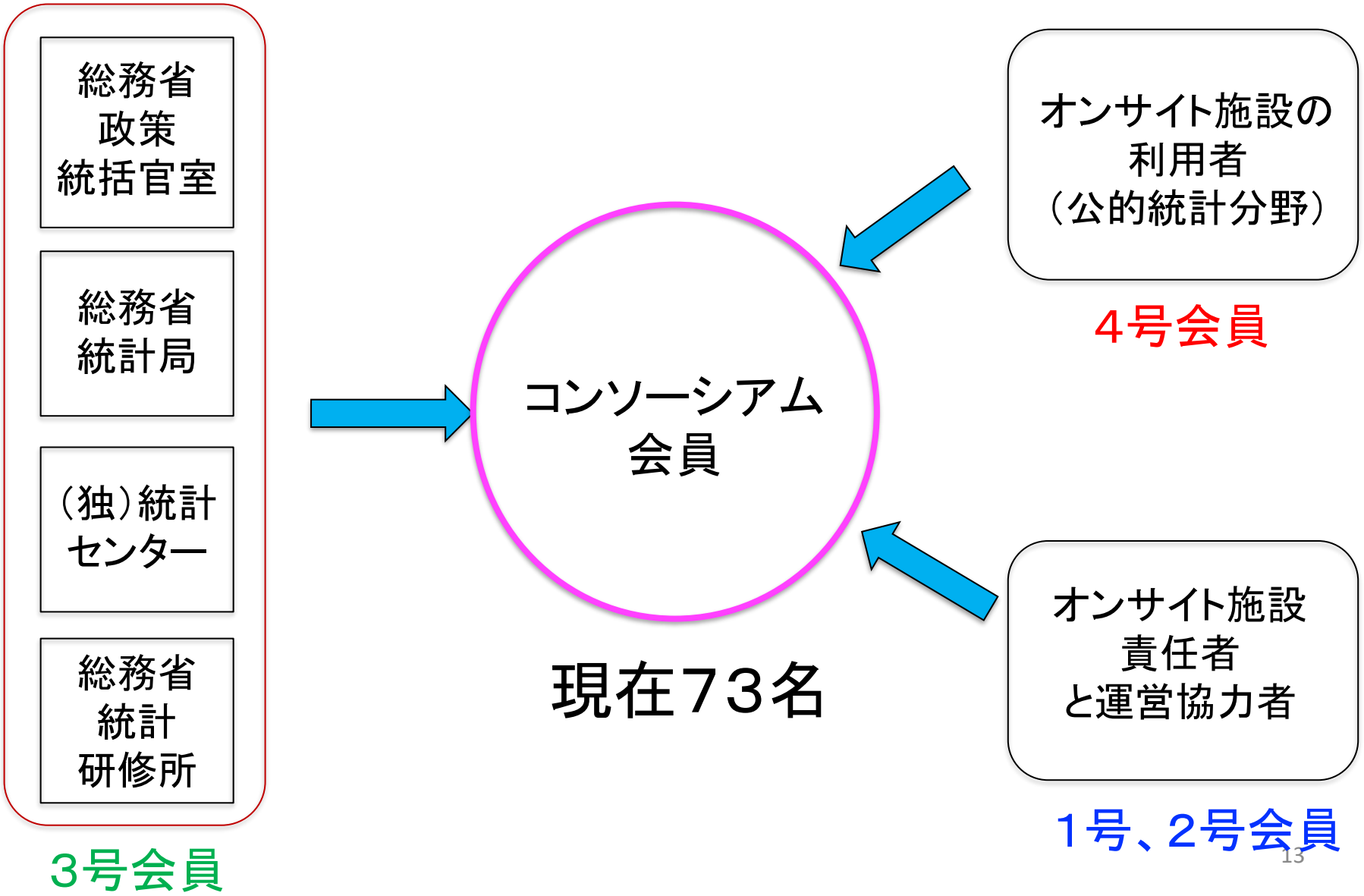
運営委員会　委員一覧

- 委員長
  南 和宏　統計数理研究所 教授
- 副委員長
  伊藤 伸介　中央大学経済学部 教授

委員

- 山下 智志　統計数理研究所 副所長/教授
- 岡本 基　情報・システム研究機構戦略企画本部
  主任URA/特任准教授
- 稲垣 好展　総務省政策統括官(統計基準担当)室 参事官
- 佐藤 紀明　総務省 統計局 統計調査部 調査企画課 課長
- 高部 勲　総務省 統計局 統計データ利活用センター センター長
- 三神 均（独）統計センター 情報技術センター 統計情報提供課 課長

コンソーシアム会員


総務省
政策
統括官室
総務省
統計局
（独）統計
センター
総務省
統計
研修所
3号会員
コンソーシアム
会員
現在73名
オンサイト施設の
利用者
（公的統計分野）
4号会員
オンサイト施設
責任者
と運営協力者
1号、2号会員

シンポジウムの概要
(2021年11月19日（金）開催）

「デジタルの日」協賛イベント

【プログラム】

＜午前の部　10:00-11:30＞

公的統計のリモートアクセス型オンサイト利用に関するチュートリアル

司会：伊藤 伸介（中央大学）

開会挨拶

南　和宏（統計数理研究所）

「統計データ利活用センターにおけるオンサイト利用推進の取組」

赤谷 俊彦（総務省統計局・独立行政法人統計センター 統計データ利活用センター）

「オンサイト利用による分析結果等の安全性確認における注意と事例」

阿部 穂日（独立行政法人統計センター 統計情報提供課）

「公的統計匿名データを利用したデータサイエンス講義のための取組み」

白川 清美（立正大学 データサイエンス学部）

＜午後の部　13:30-17:00＞

公的統計ミクロデータ研究コンソーシアムシンポジウム

総合司会：南 和宏（統計数理研究所）

開会挨拶

藤井 良一（情報・システム研究機構）

【第１セッション「公的統計と統計教育」】司会：岡本 基（統計数理研究所）

「政府における統計人材の確保・育成に向けた取組」

稲垣 好展（総務省 統計局 調査企画課）

「データサイエンス教育と公的ミクロデータに期待する役割」

渡辺 美智子（立正大学 データサイエンス学部）

「統計エキスパート人材育成プロジェクトの推進」

千野 雅人（統計数理研究所）

「SSDSE を中心とした統計教育への取組み」

山下 雅代 （独立行政法人統計センター）

【第２セッション「統計データの高度利用に関する研究」】司会：山下 智志（統計数理研究所）

「Synthetic Data の考え方に基づく疑似的なミクロデータ作成の可能性」

高部 勲（立正大学 データサイエンス学部）

「公的統計データの匿名化に関する海外の動向とわが国における課題」

伊藤 伸介（中央大学 経済学部）

「公的ミクロデータに対する k-匿名化加工の検討」

南 和宏（統計数理研究所）

閉会挨拶

椿 広計（統計数理研究所）

「デジタルの日」特設ページ (http://jmodc.org/digitalday.html)

はじめに

公的統計ミクロデータ研究コンソーシアムは、国が実施する公的調査で収集された個人や企業に関する 公的ミクロデータを学術研究に利用するために、様々な活動を行っています。

このページでは、人に優しいデジタル化の実現を目指す 「デジタルの日」 の趣旨に賛同し、このコンソーシアムの活動をできるだけ分かりやすく紹介したいと思います。

なぜコンソーシアムが設立されたか？

我が国では、社会の状況を把握するために国勢調査、家計調査など、 さまざまな公的調査を定期的に実施しています。 調査の結果から、調査対象である個人・企業の単位で編集されたミクロデータ（調査票情報）が作成されます。 このミクロデータは国の行政機関において政策立案を行う際の基礎資料にして利用されます。
またミクロデータを集計した統計情報は、我が国の社会経済情勢を多様な観点から示す有益な情報と位置づけられ、 国民全体で共有するため

ニュースレターの創刊（3月発行予定）

創刊の挨拶

公的統計ミクロデータ研究コンソーシアム評議会　議長

藤井　良一

公的統計ミクロデータ研究コンソーシアム評議会の議長を務めます情報・システム研究機構の藤井良一です。このたびのNewsLetterの創刊にあたり、一言ご挨拶を申し上げます。

公的統計は、国民にとって合理的な意思決定を行うための基盤となる重要な情報です。そして、公的統計作成の基礎となるミクロデータを分析することにより、我が国の実情を多様な観点から客観的に評価するEBPM（エビデンスに基づく政策立案）の実現が期待されています。しかしながら、我が国においてミクロデータは価値の高いデータであるにもかかわらず、その価値を十分に引き出せていませんでした。

そのような状況を踏まえ、本コンソーシアムは公的統計ミクロデータの利用推進を目的に2016年に設立され、産官学のメンバーが緊密に連携して研究利用の環境整備の推進に取り組んでまいりました。特にミクロデータの利用拠点である「オンサイト施設」を全国展開する活動に注力し、当初４拠点だったオンサイト施設は現在14拠点に拡大するなど、着実な成果が得られています。

また利用機会の少なさに伴う研究スキルの不足などにより、ミクロデータの利用が少数の研究者に留まっている状況を改善するため、ミクロデータ利用の普及活動に広く取り組んでまいりました。統計行政の最新動向、ミクロデータの利用事例を紹介するシンポジウムを毎年開催し、今年度はコンソーシアム・ホームページをリニューアルし、コンソーシアム概要の紹介記事、ミクロデータ利用の具体的ノウハウに関するチュートリアル動画の掲載等、コンテンツの大幅な拡充を行いました。このような啓発活動の結果、本コンソーシアムの会員数も着実に増加しています。現在もコンソーシアムの評議会、運営委員会では、公的統計の利用に関する課題、施策を活発に議論しており、ミクロデータの利用制度改善に関する提言、ミクロデータ利用者に対する実践的な内容のサービス提供を計画しています。

このたび、今後さらに活動領域を広げる本コンソーシアムの情報を定期的にお伝えするため、NewsLetterを発行することと致しました。今後もコンソーシアムの活動の成果、得られた知見などを取り上げ、公的ミクロデータをより広く活用していただくために有意義な情報発信をしてまいりたいと思います。今後とも本コンソーシアムへの皆様のご支援・ご協力を賜りますことをお願い申し上げます。

創刊インタビュー（座談会）も実施

座談会

公的ミクロデータ二次利用の道のりと活用促進への期待

公的統計ミクロデータ研究コンソーシアムの南和宏運営委員長と、コンソーシアム設立に深く関わった統計数理研究所の椿広計所長、山下智志副所長、岡本基主任URAの4人が、当初のいきさつを振り返るとともに、将来の展望を語り合いました。

30年越しの官学ネットワークがコンソーシアムに結実

――みなさんが公的統計と関わるようになったのはいつ頃からですか？

南　和宏　統計数理研究所　教授

南　「私は2016年5月から統計センターの非常勤研究員となり、公的ミクロデータのオンサイト利用における安全性審査をするための基準づくりを担

2年前に初代委員長の山下先生から引き継ぎました。」

椿　広計　統計数理研究所　所長

椿　「私の場合は1990年に、統計審議会調査技術開発部会の専門委員を命じられたのが最初でした。統計審議会は、統計行政の整備・改善について提言を行う総務庁長官の諮問機関で、現在の総務省統計委員会の前身です。

た。公的統計の分野の人たちと付き合うようになったのは、そこからですね。」

南　「なるほど、その頃からの交流がコンソーシアムで実を結んだのですね。」

岡本　基　情報・システム研究機構
主任URA／特任准教授

岡本　「私は2010年に、情報・システム研究機構（ROIS）の新領域融

それから10年以上ずっと公的ミクロデータに関わっていまして、2016年にコンソーシアムが立ち上がってからは、事務局の実働を担ってきました。」

山下智志　統計数理研究所　副所長／教授

山下　「私が公的ミクロデータのプロジェクトに参加した理由の一つは、それまでの自分の研究のスキームが活かせると思ったからです。

私の専門はファイナンス関係の統計学で、全国160万社の企業の財務諸表データを集めて解析に使えるようにクレンジングしたデータベースの構築・社会実装を手掛けました。それ以外にも、国際協力銀行（JABIC）や国際協力機構（JICA）といった政府機関のデータを使った研究の経験もあったことから、公的ミクロデータについても財務諸表データのように共有できれば面白い、と思っていました。

――新しくできるセンターで、以前から温めていた企画を実現しようと思われたわけですね？

山下　「最初に『公的ミクロデータをデータベース化し、利活用を実現することをセンターの目玉としよう』というアイデアを出されたのは椿所長でした。私はそれをスキームに落とし込む役割を担いました。

2015年の夏から半年ほどかけてROISや文部科学省など各方面と折衝を続け、北川源四郎機構長（当時）を議長とするコンソーシアムを設立しました。データサイエンス共同利用基盤施設の正式な事業と位置づけたことで、総務省の協力も得られましたし、統数研の存在意義をアピールする結果にもなったと思います。」

統計法の改正で二次利用への道が一気に拓けた

――コンソーシアムの設立以前、公的ミクロデータの利用はどんな状況だったのでしょうか。

椿　「欧米では公的ミクロデータが経済学の研究に利活用できるのに、日本ではまったくできないというのが実情でした。日本の経済学の先生方は、1995年頃からこのことに問題意識を持っていました。

コンソーシアム・ウェブサイト
(http://jmodc.org)

チュートリアル動画の公開

- シンポジウムのチュートリアル講演の動画を公開(3月公開予定)

●「統計データ利活用センターにおけるオンサイト利用推進の取組」
赤谷 俊彦（総務省統計局・独立行政法人統計センター 統計データ利活用センター）

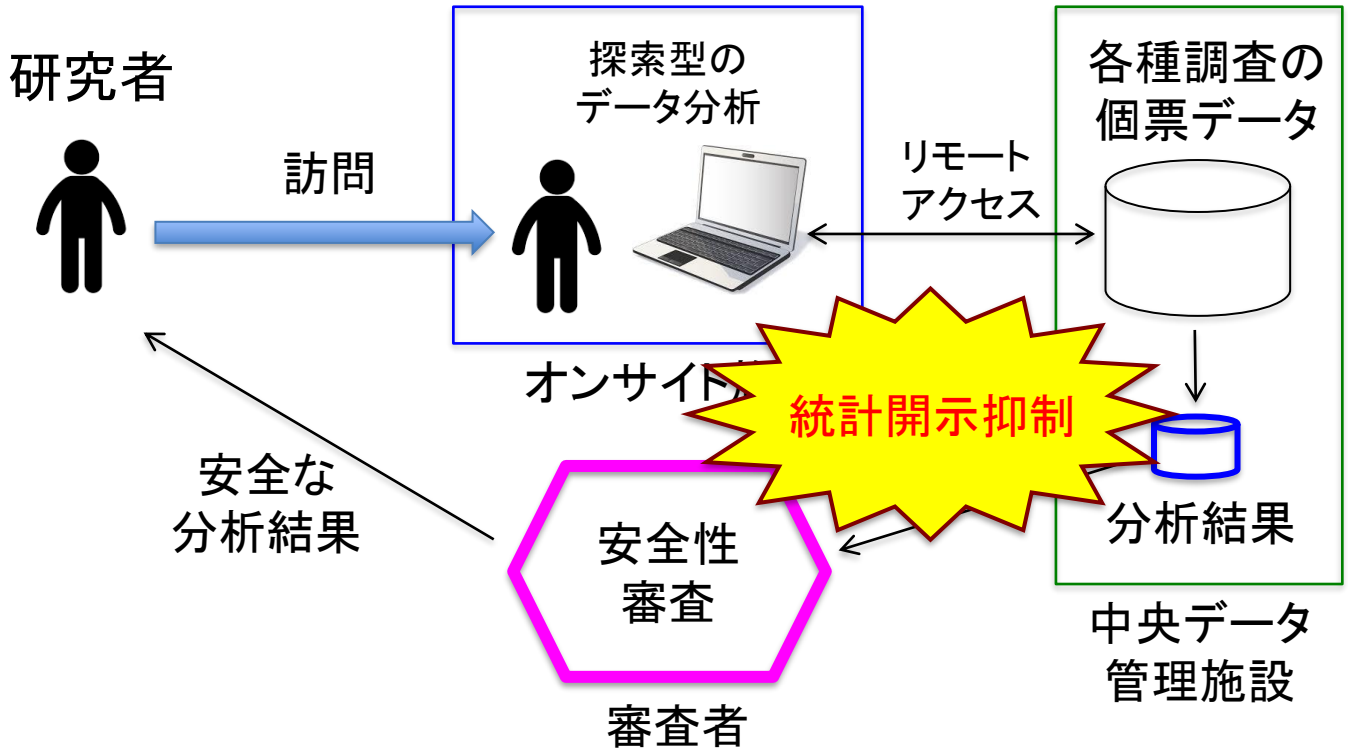
●「オンサイト利用による分析結果等の安全性確認における注意と事例」
阿部 穂日（独立行政法人統計センター 統計情報提供課）

２．ミクロデータ活用のための基礎技術の開発

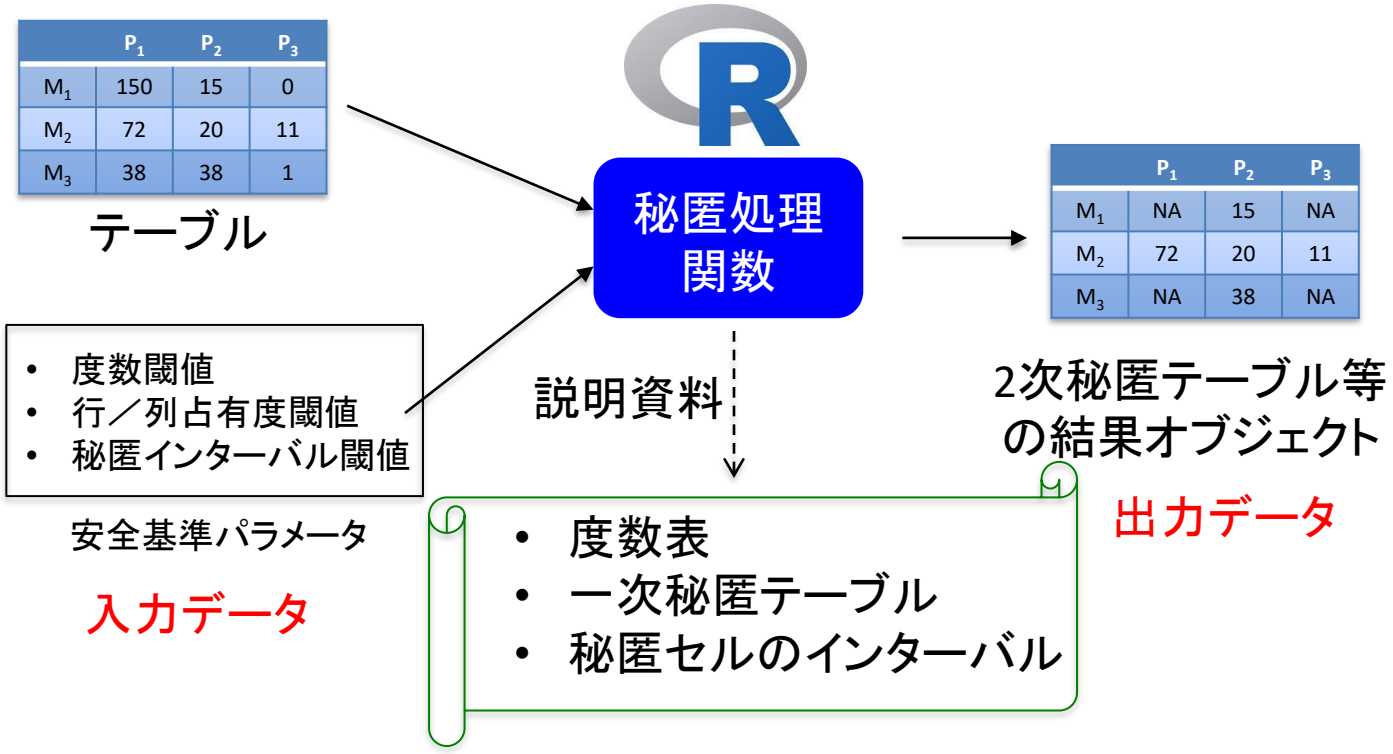
2-a　プライバシー保護のための統計的セキュリティ

2-b　匿名加工データの作成技術開発

オンサイト利用におけるプライバシー保護と安全性審査



Rによる説明資料も作成する秘匿処理ツールの開発



- 今後はオンサイト施設への展開を目指す
- 一部機能は、「簡易集計システム」で既に提供

公的ミクロデータへの k-匿名化技術の適用の検討

- k-匿名化は匿名データのレコード識別を防ぐ有望な安全性指標
- これまでの公的ミクロデータの匿名データにおける地域情報の公開は非常に保守的（k=数十万）であった
- 今年度はがん登録情報を題材に、より詳細な地域情報の公開について検討した

3. ミクロデータ構造化のための統計的技術開発

3-a　欠損値補間およびデータ結合のための統計的方法論
3-b　複数のデータベースに対する統計的マッチング
3-c　法人ミクロデータの収集とリスク分析
3-d　精度と規模が極端に違うデータベースの結合と利用
　　その社会実験（不動産データ）

データ構造化の目的と手法

データの供給を受けた研究者がストレスなく分析を開始できるよう、
データベースを供給段階で整えておく必要がある

・欠損値補間

経済統計、医療統計を中心に研究成果が多い。
経済（Hot Deck, Cold Deck）と医療（ICE, Knn, MICE）とは系統がやや異なる
単純は「削除」「平均値補間」は少なくなりつつある。

・異常値補正

バイアスに関係する研究は多いが、個別セルに対する研究成果は少ない。
3シグマ折り返し処理、数値情報の順位化、数値情報のカテゴリー化などの
経験的処理が一般的である。
一般的にモデルパラメータに与える影響が大きい

・データ結合、リレーション、データ・リンケージ

いわゆる名寄せ。
情報が複数のデータベースに分散している場合、必須の処理
特定フィールドの完全一致から、確率一致性へ進化

・テキストデータの数量化

いろいろなミクロデータと構造化の関係

官学データ利用（政府データの学術利用）

・オンサイト拠点の全国展開によるデータ・アクセシビリティの向上
　　→オンサイト拠点設置・運営のための費用負担
・公的ミクロデータ研究コンソーシアム活動による広報活動と利用者ニーズの把握
・データ構造化よるクレンジングされたデータの供給、データ・リライアビリティの向上
　　→データ構造化研究の推進
・データ利用のための人材育成、情報提供

学学データ利用（研究活動データの学術相互利用）
・社会データアーカイブズの開発と利用（東大、慶應大などの実績）
　　→データ提供者のメリット、利用者フレンドリーなインターフェースなどが課題

民学データ利用（民間データの学術利用）
・民間との共同研究などでデータを入手可能
　　→知財契約、守秘義務契約などのコンプライアンス、セキュリティ体制など
　　なれていない研究者にはハードルが高い

今後の展開（2）

民官データ利用（民間データの政府利用）
→所轄官庁が企業よりデータを取得することは一部行われている（例　銀行法24条）
　ただし省庁間で利用する仕組みはこれから？
　cf. 銀行法24条のデータ提出先は内閣総理大臣であるが実質的には金融庁が独占利用に近い状態

民民データ利用（民間データの民官利用）
　→コンプライアンスが整理できた業界からデータの共有化は進められている。
　（データベンダーの存在）

学民データ利用、学官データ利用については学学のアーカイブズを通して可能
→アーカイブズの改良が必須？

・統計改革推進会議
官官データ利用の早期実現、
官学データ利用については早急推進、
官民データ利用については重要課題として積極的に検討

・骨太の方針
官民が保有するデータの徹底した利活用を図るべく、新しい社会インフラとなるデータ利活用基盤を構築する。「官民ラウンドテーブル」等を通じた公共データのオープン化、安心してデータ流通を促進させるための法制度整備等を進める。

データベース結合が前提とする条件

データオーバーフローの時代
→人的資源、計算資源を超えたデータの規模と種類
→Webデータ、テキストデータなど非構造化データの存在
→複数データベースにまたがった情報

データから目的にあった情報を抽出するために
→複数のデータベースに存在する情報を正確に抽出する

例；企業データの結合

●政府データ：　無借金企業を含めたセンサス
労務情報など特定の重要フィールドあり
調査フリークエンシーが長い
個人事業主については情報が欠落

●銀行データ：　債務者のみのデータ（例CRD協会：160万社）
財務データが正確で広範囲
毎年情報が更新
個人事業主のデータもあり

日本全体の企業を対象とした分析・政策

無借金企業に対する貸出

より多いフィールドからの情報抽出

入手した法人ミクロデータ

【法人データ】

① 高度信用リスク統合データベース：

地銀5行（滋賀、伊予、群馬、北陸、八十二）の
統合与信データベース・貸倒損失データベースの作成とモデル化
→担保・保証・毀損情報をデータとして持つ世界初の統合データベース

② CRD協会データ、日本の企業の過半を把握

財務とデフォルト情報を格納した巨大データ（）のクレンジング。
データ元（全国都道府県信用保証協会、比較的小さい金融機関）

③ 公的ミクロデータ：オンサイト拠点

総務省、経済産業省の法人統計データ

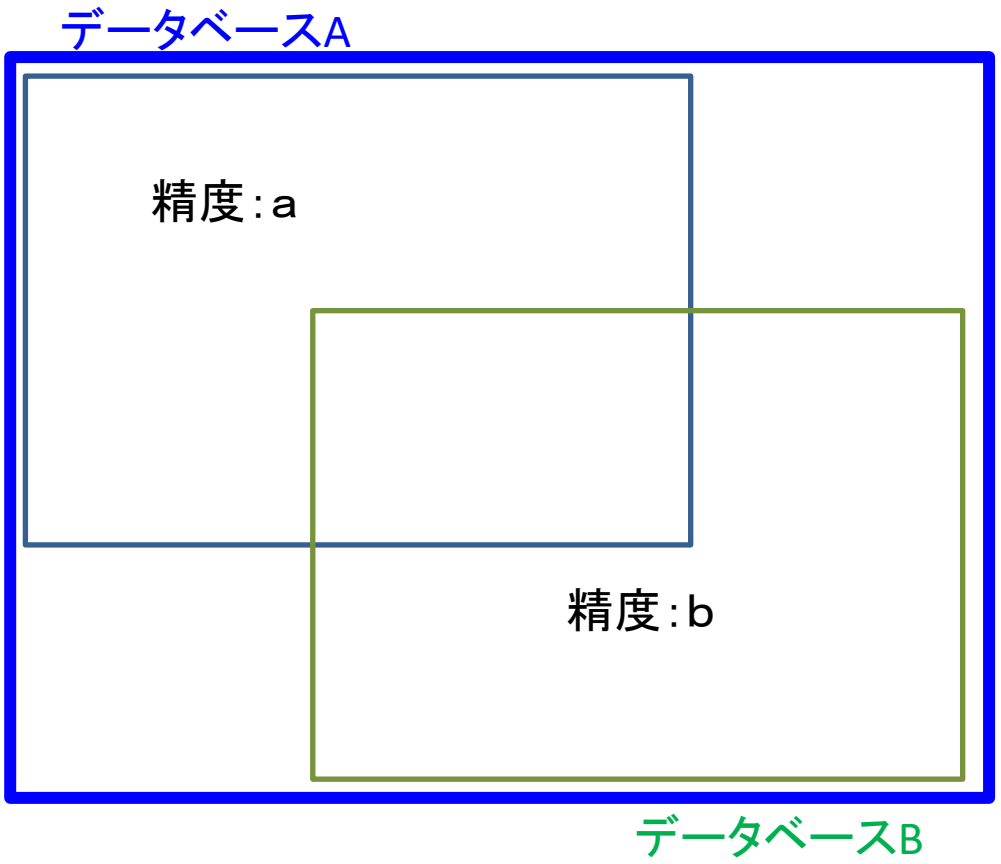
④商用民間信用データ

帝国データバンク 3県×3年

【その他】

④ 銀行の口座情報（勘定系データ）
⑤ ソブリンデフォルトリスクの計測と格付システムの開発（JBIC、JICA）
国際協力銀行（JBIC）、 国際協力機構（JICA）の海外融資データ
2014年開発開始、2015年からモニタリング
⑥ 賃貸不動産定点観測データ
⑦ 賃貸不動産ネットクローリングデータ
⑧ 物件情報の巨大データベースの作成 → 空室率予測モデルの作成

事例　2つのデータベースの結合と欠損･重複



データベース（A+B)

●一部のレコードが共通

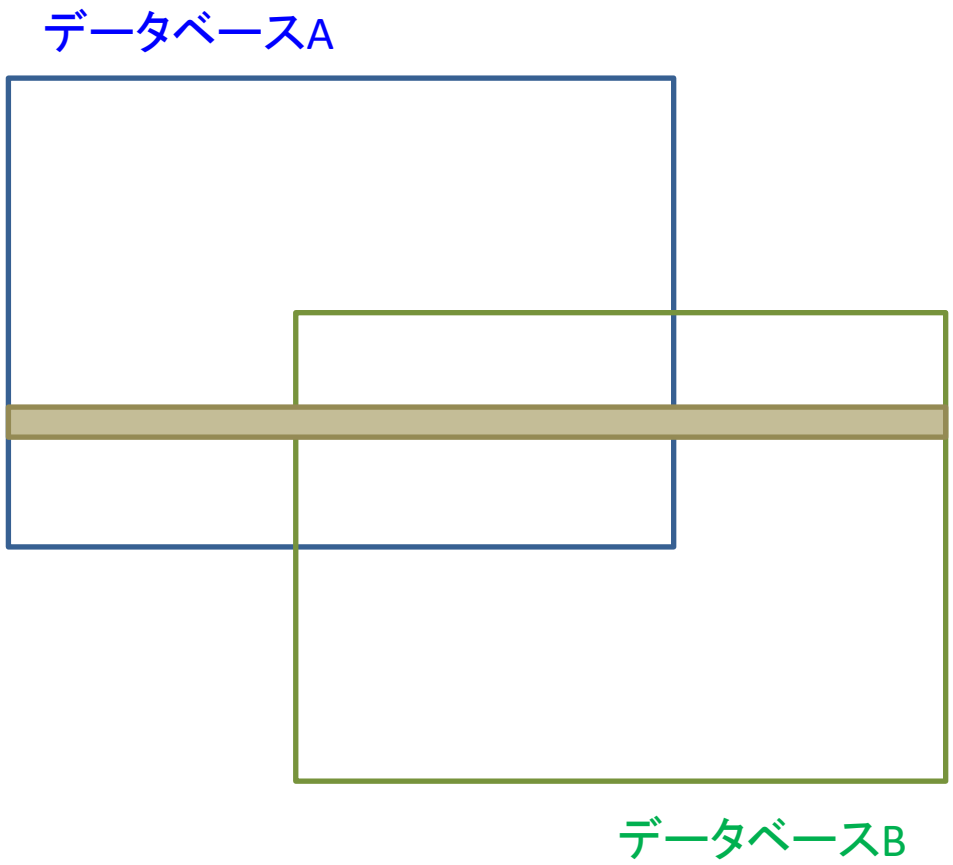
全部共通もしくは共通レコードがなしの場合は少し問題が単純

●一部のフィールドが共通

全部共通もしくは共通フィールドがなしの場合は少し問題が単純

●データベースの精度は同じでは無い

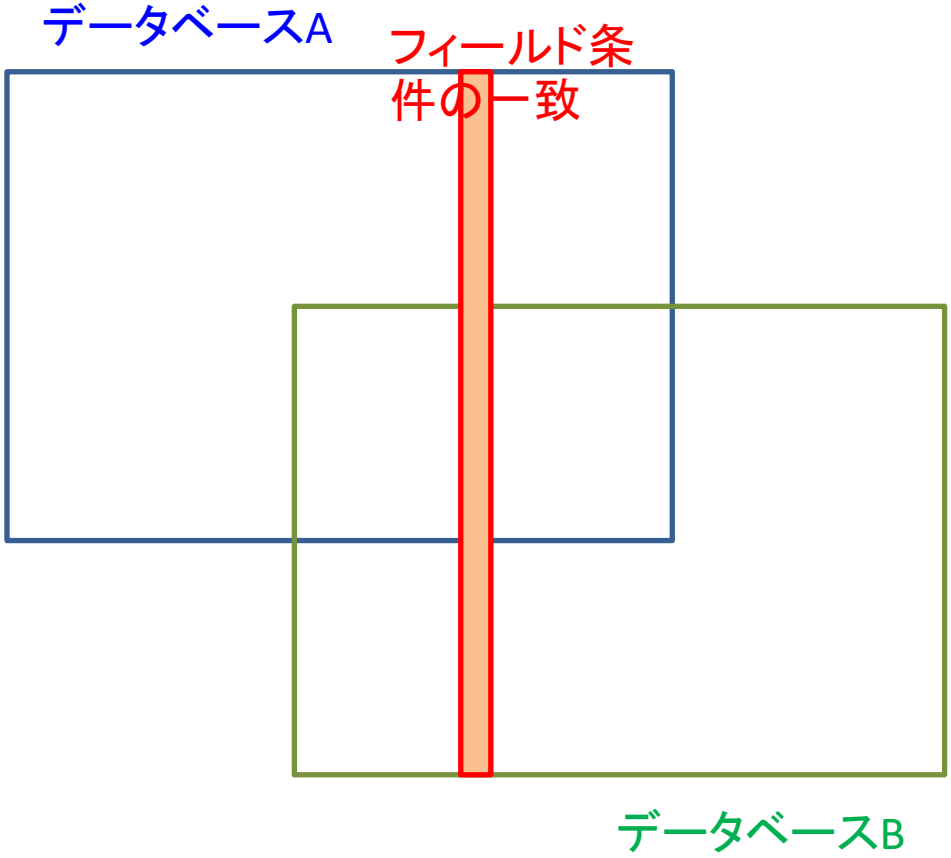
名寄せ・マッチングの問題



●部分フィールドの完全マッチングが一般的
→表記の揺れ、フィールド定義の違いなどで名寄せできない
→秘匿性情報では完全マッチングが禁止されていることが多い

●統計的マッチングの導入
→高部発表

データベース結合が前提とする条件



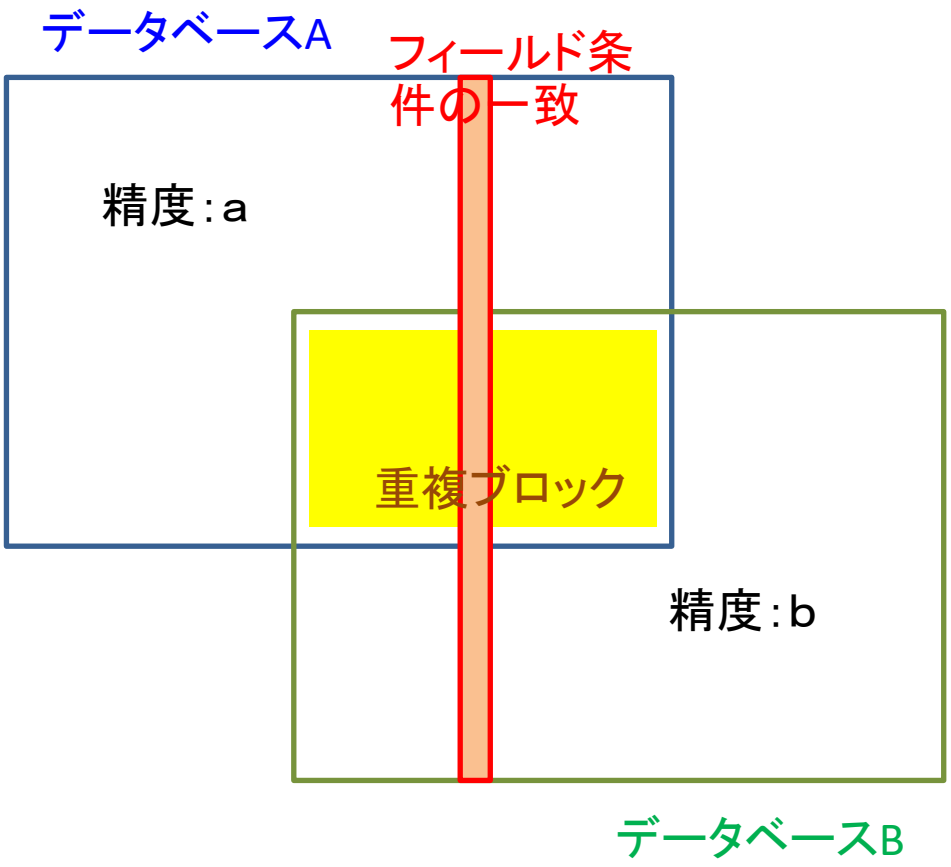
同じ指標に見えるが、データによって微妙に定義が異なるかもしれない

例：雇用者数・従業員数
・社長、役員は入れる？
・バイトは？週何時間以上？
・観測時点は年末？年度末？

↓

少しでも違えば、違うフィールドとすべきか？
小さい違いなら1つのフィールドとして扱うか？
　→その場合、どのように合成するか？

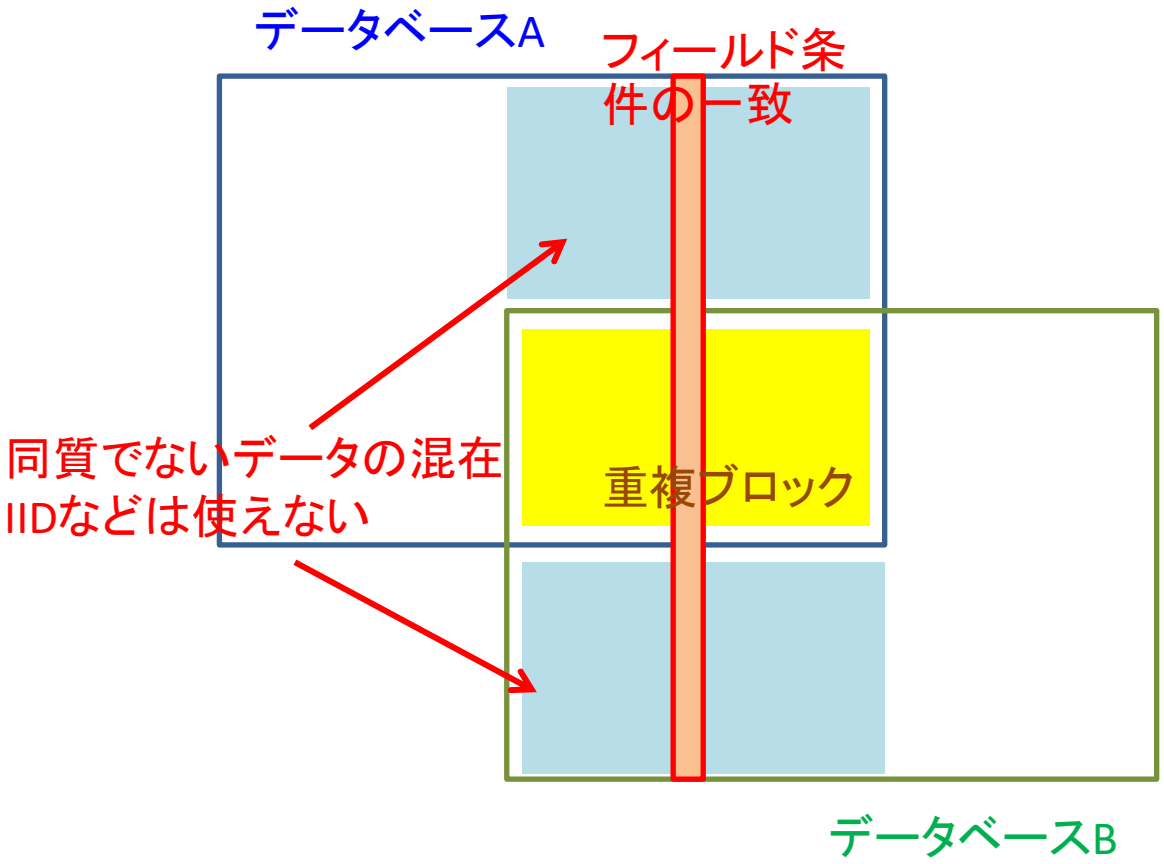
重複ブロック



フィールドが同じと判断された場合、
データが重なっているところについては、2つのデータを1つにする必要がある。

・高精度のデータをつかう？
・精度に応じた案分？

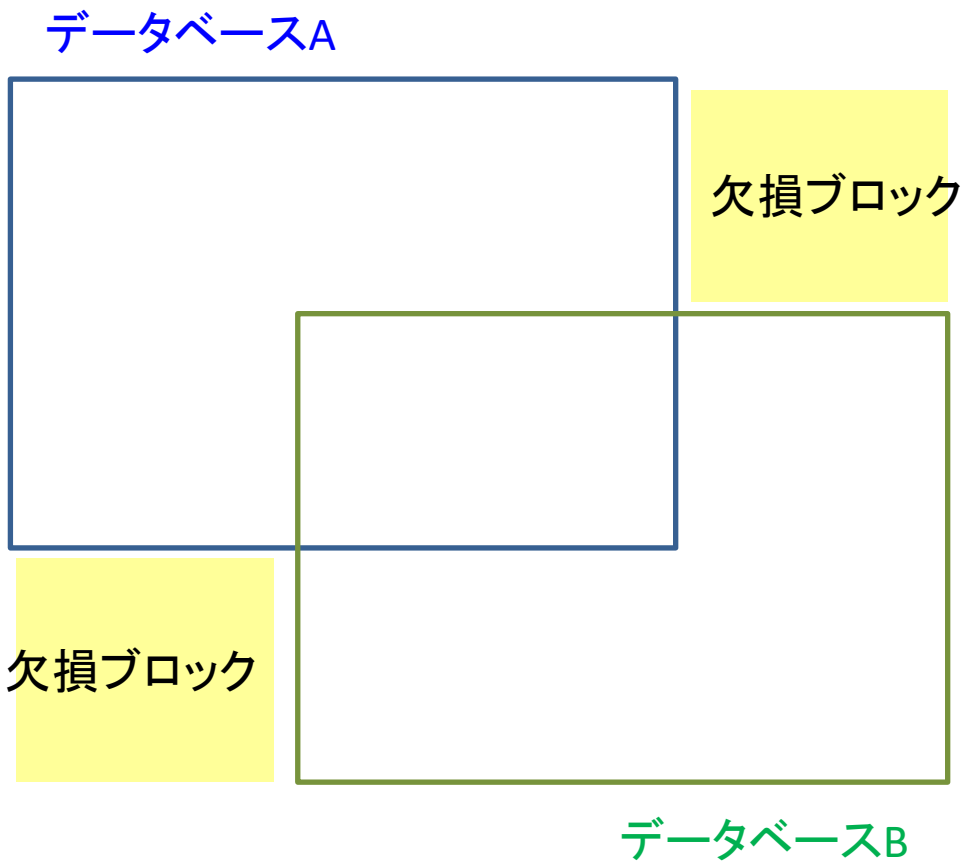
同質でないデータが混在するフィールド



合成されたフィールドにおいては、精度の違うデータが混在することになる。

同一分布からのサンプリングとはいえないため、統計分析にいろいろな制約ができる。

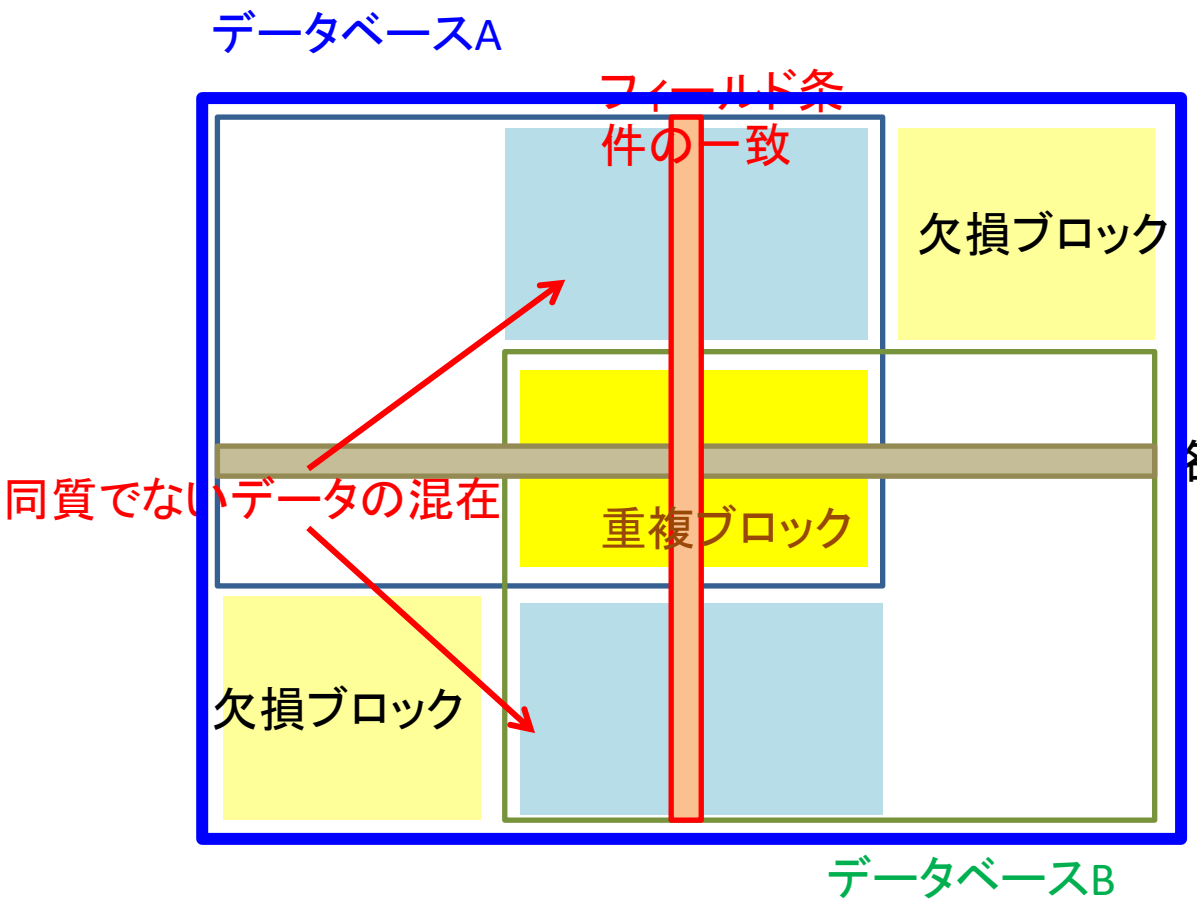
欠損ブロック



どちらのデータベースにも情報が存在しない部分がある。

基本的に欠損値補間問題であるが、MAR（Missing at Random）ではありえない。

データベース結合における統計的問題の種類

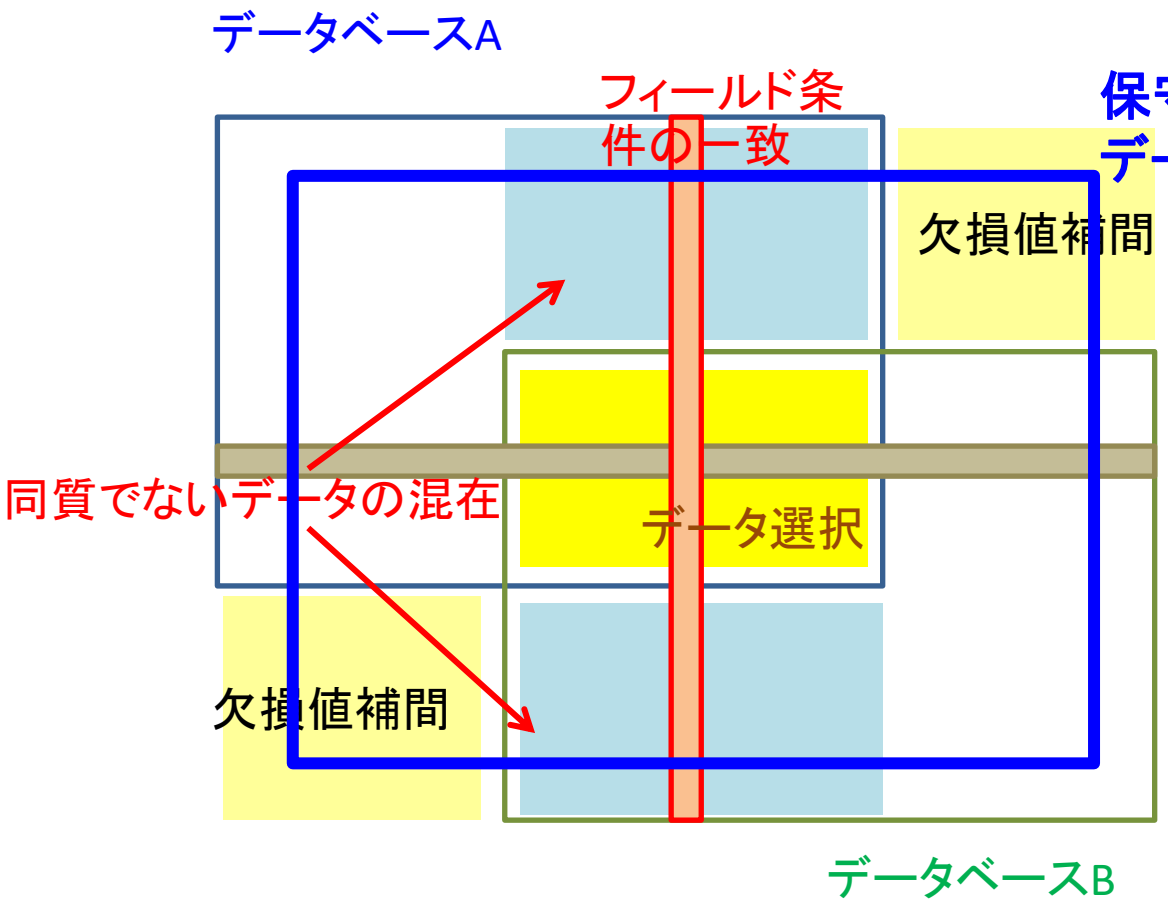


これらの問題が解決されれば、データ結合が完成

↓

全体を考えた事例はみあたらない？

データ縮約について

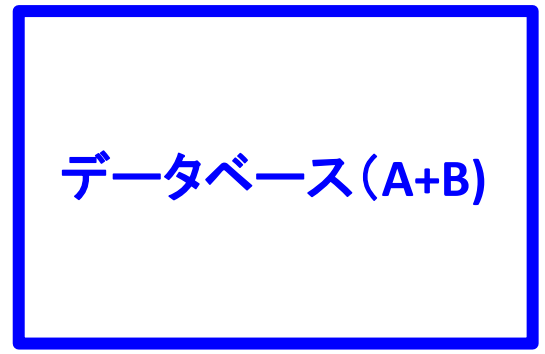


結合されたデータの精度に自信が無い場合、一部のデータを捨てるという方法もある。

↓

捨て方に応じたセレクションバイアスが生じる

目的感の違いによるデータベース結合の2つのケース

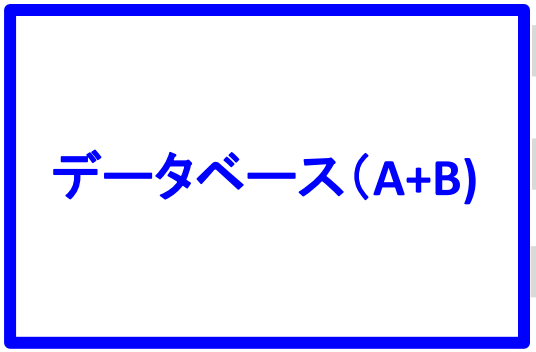


問題・モデル

特定の問題・モデルのためにデータ結合を行うケース

例：倒産確率推定

予測精度（正答率など）を目的関数として最適化を行うことによって、合成のためのパラメータを得ることができる



問題・モデル

問題・モデル

問題・モデル

不特定多数の問題・モデルのためにデータ結合を行うケース

例：オンサイト施設におけるデータ供給

明確な目的関数が存在しないため、問題の定義づけが難しい。

まとめ

・データ過剰の時代の分析のあり方を探る

・質の高いデータ（公的ミクロデータなど）の学術普及

・プライバシー、秘匿性に注目したデータハンドリング

・ミクロデータ分析のためのデータ構造化理論を構築する

⇒社会的環境や協力者に恵まれ、順調に事業が進んでいる