資料3

公的統計ミクロデータ二次的利用推進のための取り組み

情報・システム研究機構　統計数理研究所
山下智志

CONTENTS

取り組みの背景（1）

- 我が国では主要先進国に比べて、社会・経済に関する実証研究が非常に遅れており、状況の改善が急務である
  - 研究や高等教育の場で**利用可能なミクロデータが限られている**
  - 研究者にミクロデータ実証分析に関する経験や研究力量が決定的に不足している

- 人文社会科学分野の国際的なジャーナルでは、**ミクロデータを用いた実証分析が主流**
  - 実証分析なしに論文が掲載されることはほぼ不可能
  - 我が国の研究者は、海外のミクロデータに依存するなど著しく不利な状況にある

- 欧米、オーストラリア、韓国などでは政府の公的統計のミクロデータの研究の利用環境が整えられ、研究の知見に基づいて**Evidence-Based Policy Making**が実践されている

取り組みの背景（2）

これらの課題の解決には、国の公的統計調査のミクロデータを、個人の秘密を厳守しつつ、実証分析を有効に活用する体制の確立が必要

- 我が国研究者の研究力量向上
- 人文社会科学分野の国際競争力の向上
- 国民生活の向上
- 社会経済の発展に資する政策科学研究の促進
- エビデンスに基づく科学的な施策の立案・評価の推進

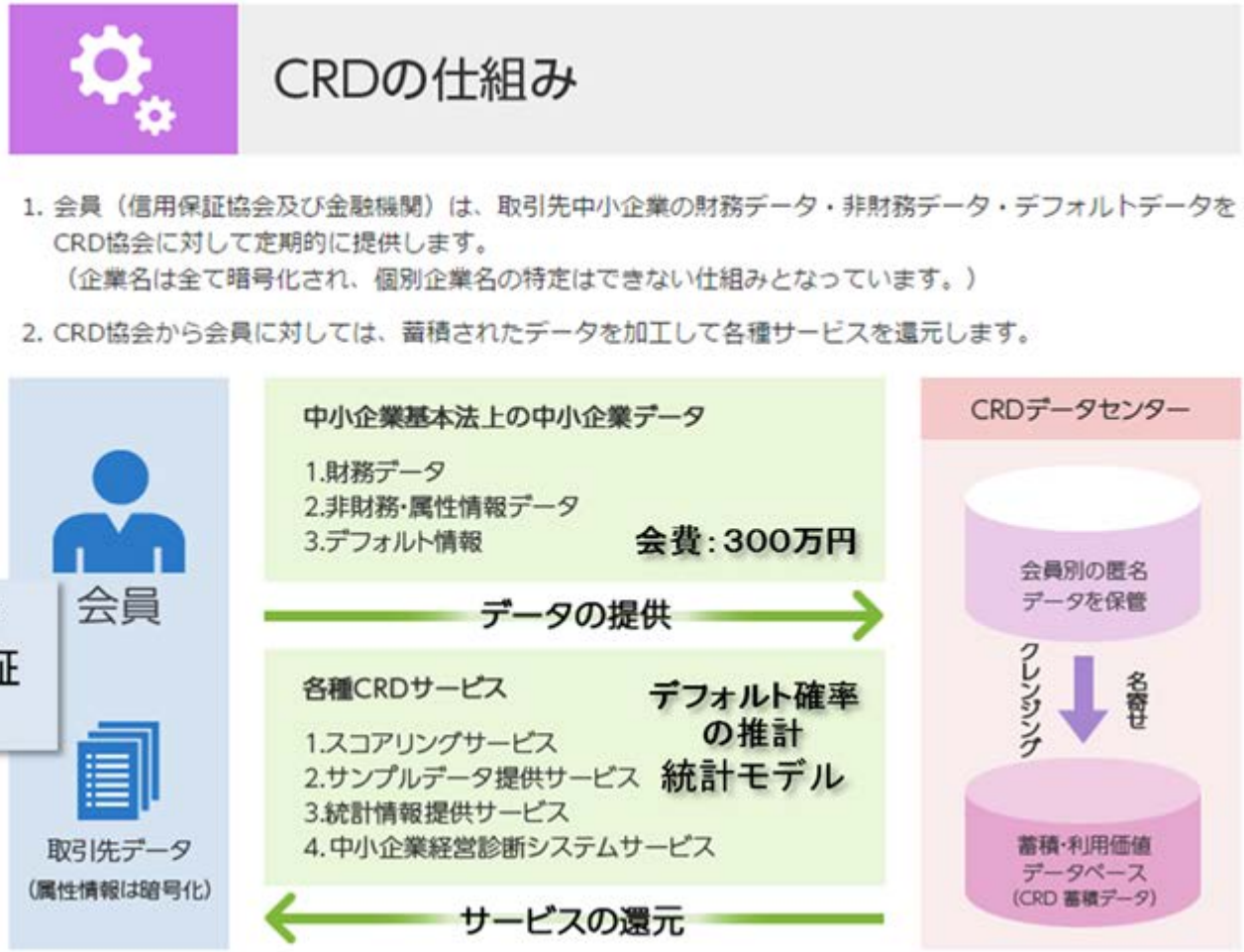
平成19年の統計法の改正に伴い、**公的統計のミクロデータ（個票）の二次的利用**（研究・高等教育利用）が認められることとなり、我が国でも、集計データではなく、大規模ミクロデータに基づく人間・社会政策科学研究を推進する機運が高まっている

平成23年からは厚生労働省のナショナルレセプトデータベースの研究利用も開始された

データ基盤整備、ならびに、政府情報の研究者利用の推進に資する環境を整備

取り組みの背景（3）

情報・システム研究機構　統計数理研究所における
秘匿性ミクロデータ相互利用プラットフォーム開発事例



取り組みの背景（4）

平成19年統計法改正により、公的統計データの二次的利用（研究・高等教育利用）開放

新しい条項の追加

- 第34条　委託による統計の作成等（オーダーメード集計）
- 第36条　匿名データの提供
  - 学術研究の発展に資すると認める場合
  - 高等教育の発展に資すると認められる場合

独立行政法人統計センターによるオーダーメード集計、匿名データ提供の開始

平成22年7月、機構－統計センターと連携協力協定締結

- 統計数理研究所内に「新領域融合研究センター統計数理研究所・オンサイト解析室」（当時）を設置
- 同年9月よりサテライト機関としての業務を開始

調印式で握手を交わす堀田凱樹 情報・システム研究機構 機構長（当時）と戸谷好秀 統計センター 理事長（当時）

CONTENTS

オンサイト施設の整備・運用

統計法**第33条に基づく調査票情報の提供（目的外利用）**を受けるためには、一定のセキュリティ要件を満たす必要がある

- データ分析を実施するエリアの入退管理
- 施錠可能なキャビネット等にデータを保管
- 外部ネットワークへの接続禁止

全ての研究者が要件を整えられるわけではない

- 共同研究室を割り当てられている若手研究者
- 施設の都合上、セキュリティを強化することができない組織に属する研究者　など

→条件を満たすことができない研究者にとって
データ利用申請は困難

研究者のデータアクセシビリティを高めるためには、セキュリティ要件を満たす施設の供給が不可欠

オンサイト施設の整備・運用

情報セキュリティが確保された施設と使用を厳重に管理することが可能な仕組みを整備し、提供することにより、33条の適切かつ円滑な運用、調査票情報の適正管理、行政機関等の事務の負担軽減及び手続の迅速化に寄与

高度なセキュリティ環境が整った分析拠点を研究者に提供し

データ利用環境整備の負担を軽減

統計センターとの連係協力協定に基づき、情報・システム研究機構でも統計数理研究所内にオンサイト施設を設置し整備と運用を開始

- 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
  社会データ構造化センター オンサイト解析室

オンサイト利用に係る施設基準（試行運用版）について（概要）

１　運営・管理体制について

・オンサイト施設管理責任者や、オンサイト施設管理運用窓口を置いていること　など

２　オンサイト施設について

・施設管理責任者が認めた者や施設利用者以外の立入りを制限し、機密情報を安全に利用するための措置を講じた施設であること
・オンサイト施設への入退室に際し、氏名、所属、日付、時刻の記録を行う措置が講じられていること
・施設利用者に対し、施設利用許可証を貸与し、施設利用時において常に見やすいところに着用させる措置が講じられていること
・利用者端末が複数ある場合、それぞれをパーテーションで区切る等の措置が講じられていること
・統計センターが指定する回線（SINET）により、利用者端末から中央データ管理施設へのアクセス可能なネットワークが構築されていること
・利用者端末及び機器等について、定期的にメンテナンスを行い、正常な状態を維持する体制が整えられていること
・オンサイト施設内に、インターネットにアクセス可能なパーソナルコンピュータの設置を行う場合は、大学のセキュリティポリシー等に準拠したものであれば設置可能とし、利用規約等を定め、利用者端末から離れた場所に設置されていること　など

３　利用者端末について

・利用者端末の設置は、3台までとすること。なお、利用者端末を増設したい場合は、統計センターと協議すること
・リモートアクセスのために使用する起動用USBメモリの盗難や紛失を防止するため、管理を徹底すること
・利用者端末の盗難、外部への持ち出しを防止する措置を講じること

オンサイト利用に係る施設基準（試行運用版）について（概要）（続き）

４　施設利用者への規制及び監視措置について

・パーソナルコンピュータ、カメラ、レコーダ等の記録機器類、無線LAN端末、携帯電話等の通信機器類、その他これらに類する機器について、施設利用者に対し、オンサイト施設内への持込み禁止等の措置が講じられていること
・監視カメラ及びカメラ映像を記録する機器を設置するなど、オンサイト施設内における施設利用者の行動を監視できるようにすること　など

５　報告・検査措置について

・施設管理責任者等の名簿、施設内の利用者端末、その他の機器等の構成及び配置について、統計センターに提出すること
・定期的にオンサイト施設の利用状況を統計センターへ報告すること。また、統計センターが施設の利用状況について報告を求めた場合は、速やかに求めに応じること
・施設管理者は、統計センターの求めに応じ、オンサイト施設の検査を受けられる体制をとること

６　その他

・上記で定めることのほか、オンサイト施設におけるセキュリティの確保に努めること
・オンサイト利用に障害等が発生した場合には、その対応を行うこと
・調査票情報の漏洩等の事故や不適切な利用等の事象が生じたとき又はその疑いが生じたときは、統計センターと共同で原因究明等に当たること
・施設の利用時間は、統計センターの運用体制を踏まえ、施設管理者が決定すること。決定した利用時間は、統計センターへ報告すること

統計数理研究所　オンサイト解析室の概要

強固なセキュリティにより、データを保護

- 管理側でデータの持ち出しをコントロール
  - PCへの外部メディア接続の全面制限
  - データ保存用端末による集中管理
  - 外部ネットワークからの遮断による
    ネットワーク経由のデータ流出防止

- 監視カメラによる常時監視
- パーティションにより区切られた作業環境
  - 意図するor意図しない窃視の防止

統計法第33条による公的統計の目的外利用におけるセキュリティ要件を満たした環境

- 統計センターによる検査を受け、認可

最終生成物データを持ち出す際には、解析室、データ提供者による内容チェックを受ける必要あり。原則として即日持ち出しは出来ない。

本年8月、立川新棟に移転・拡充

統計数理研究所　オンサイト解析室の利用可能システム

解析用端末:2台

- OS：Microsoft Windows7 Professional SP1 64bit
- CPU：Intel Xeon E5-2630 v3 (2.4GHz 8Core)
- GPU：NVIDIA GeForce GTX980 (GDDR5 4GB)
- メモリー：DDR4-SDRAM 32GB (8GB×4) (ECC/PC-17000)
- ドライブ：SSD SATA 6GB/s 512GB×2 (RAID1)

光学ドライブ無し、USBポート使用不可

- データは、室内ネットワークを通じてデータ保存用NASにアクセスし、利用する。
- **最終生成物データを持ち出す際には、解析室、データ提供者による内容チェックを受ける必要あり。原則として即日持ち出しは出来ない。**

使用可能なソフトウェア

- IBM SPSS Statistics 24 (日本語版)
  Statistical Base, Advanced Statistics, Regression, Custom Tables, Missing Values, Categories, Amos
- R (日本語版) + RStudio
- Microsoft Office 2016 Professional (64bit) (日本語版)
- Adobe Acrobat DC Pro (日本語版)

追加可能なソフトウェア

- SAS 9.4 (32bit/64bit)

その他のソフトウェア利用については応相談

オンサイト利用施設整備の意義

高度なセキュリティ環境・分析用端末を自前で整える必要がなく、
データ利用者の負担を軽減

あらかじめセキュリティ要件が満たされている環境を利用してもらう
ことで、セキュリティ審査負担・利用者管理負担、漏えいリスクを軽減

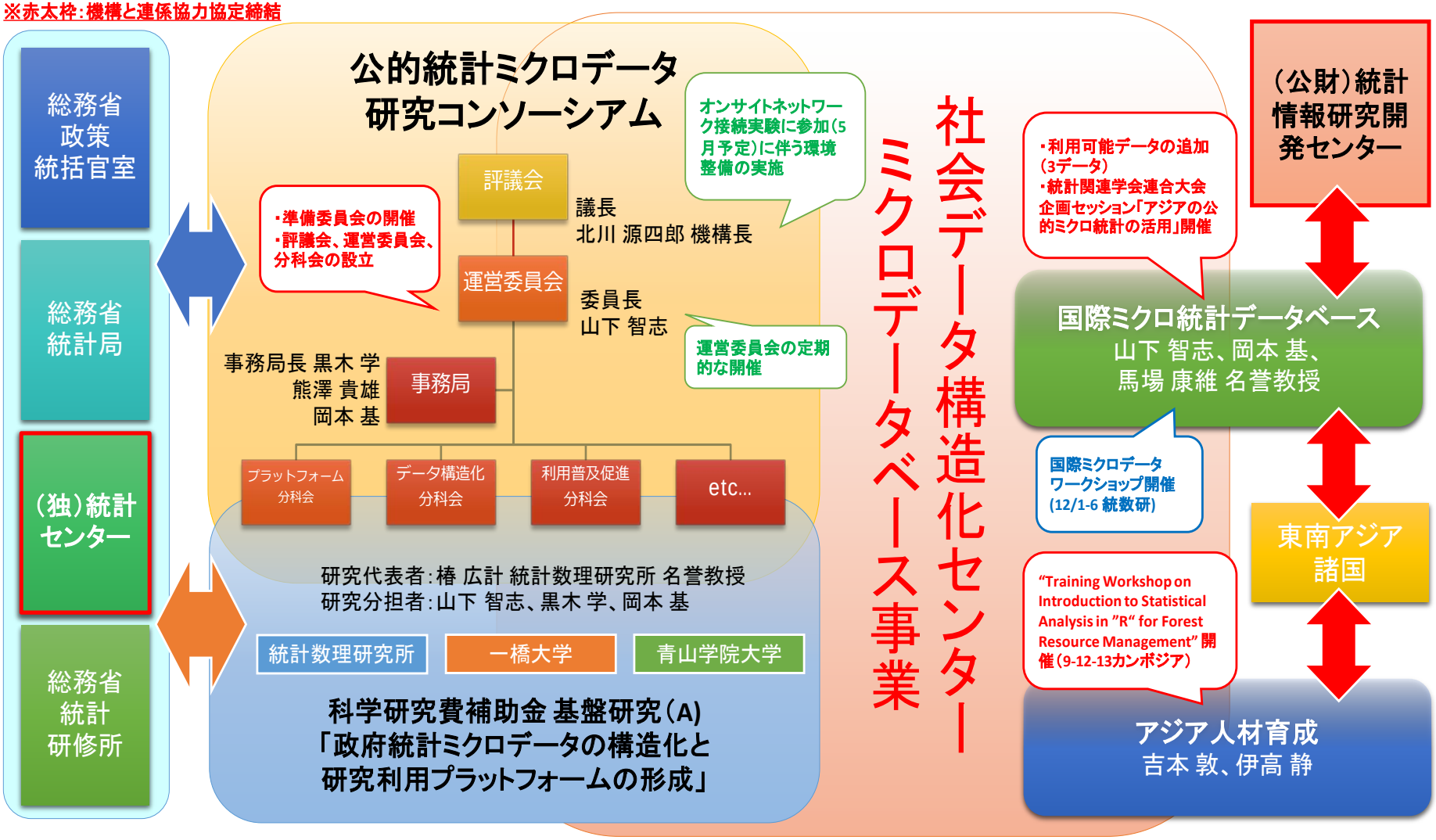
これまで環境を整えられなかった研究者が
データ利用の機会を得られる可能性を提供

公的統計データの二次的利用を促進

CONTENTS

ミクロデータ事業の運営体制
普及活動と必要技術の開発

※赤太枠:機構と連係協力協定締結



公的統計ミクロデータ研究コンソーシアムの設立

公的統計の利活用を推進めるために、全国の研究者に向けて、利用しやすい環境を整える必要性

- 必要なセキュリティを確保した「オンサイト施設」を**全国的に整備**し、潜在的な研究の裾野を広げる

学官産が協力して、公的統計の二次的利用、オンサイトネットワーク整備の支援と意義を広く周知し、諸課題の検討と解決にあたる「公的統計ミクロデータ研究コンソーシアム」を設立

事務局を社会データ構造化センターに設置

平成28年3月29日、
一橋講堂にて設立記念シンポジウムを開催

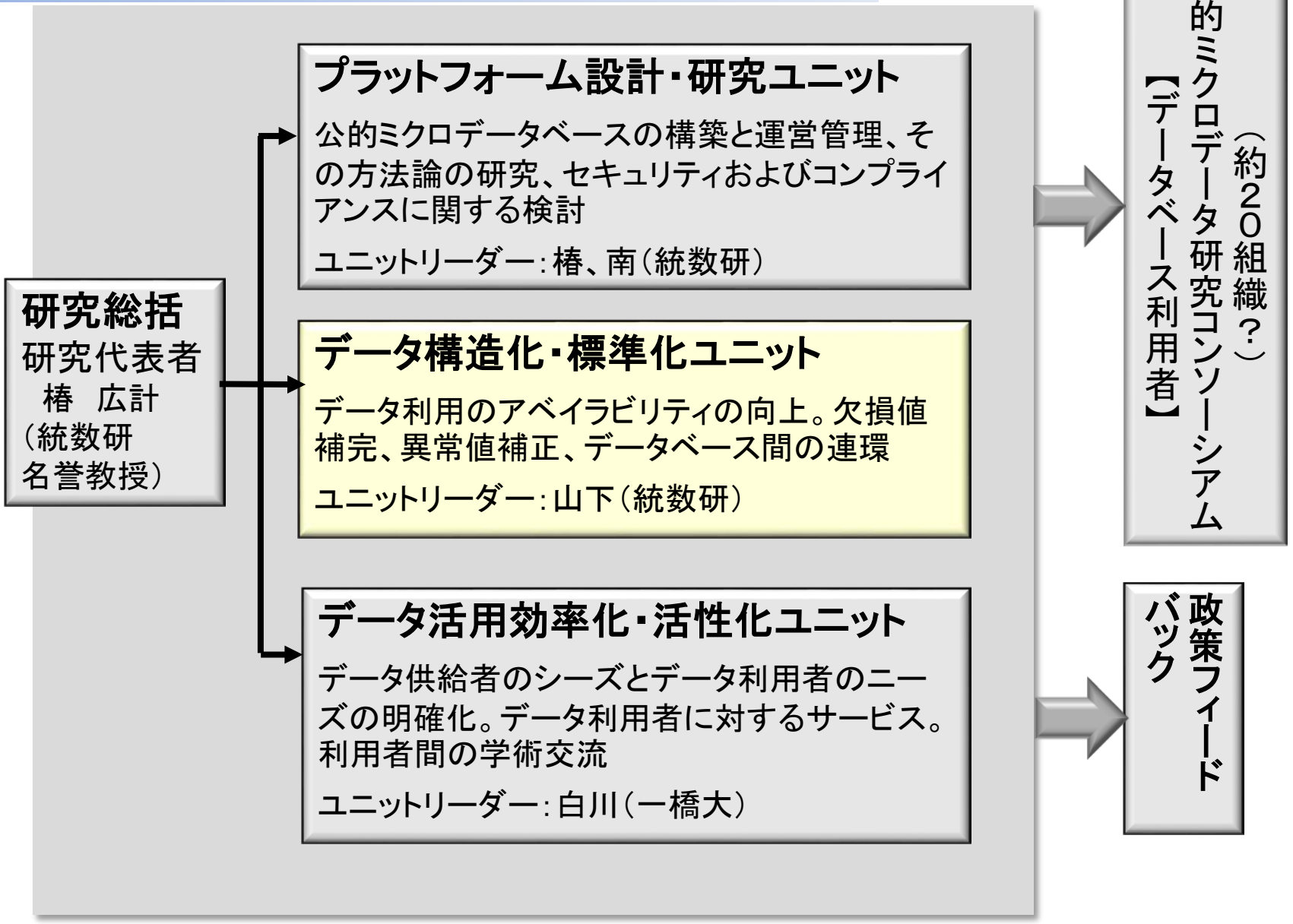
平成28年8月28日、第1回評議会開催

…

本年9月より、会員募集開始（予定）

設立記念シンポジウムの様子

データ活用のための研究体制



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研究を開始できるよう、
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

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

・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

CONTENTS

今後の展開（1）

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（民間データの民間利用）
　→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。

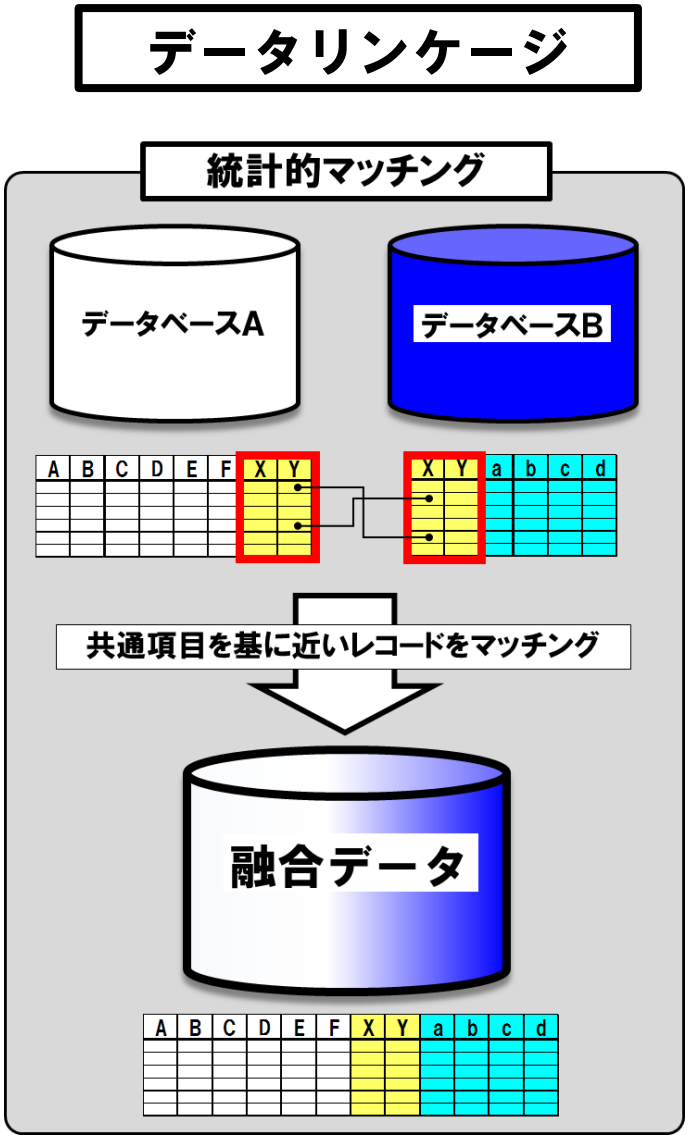
CONTENTS

データリンケージ、リレーションとは？

- 複数のデータベースをレコード単位で結合(Matching)することにより、豊富な情報を持つ単一のデータベースを構築する方法。

- 新たな調査やデータ収集を行うことなく有益な情報を持つデータベースを構築することが可能。

マーケティング・サイエンスの分野では「データ融合」(Data Fusion) とも呼ばれており、POSデータと消費者アンケートデータを統合することにより、消費者行動に関する詳細な研究・分析が行われている(星野(2009))



特定フィールドの完全一致から、統計的一致性評価へ

これまで
特定フィールド（名称、住所、設立年など）の完全一致
　→　名称・住所表記のぶれなどから、不完全一致の対応アルゴリズムはある
完全一致しないデータは基本的にリレーションしない

これから
類似度を指標化し、ある閾値を超えたものはリレーションを行う
　→　統計的マッチング

０．統計的マッチングの際の前提条件

（1）「条件付き独立性」

・$[X,Z]$ と $[Y,Z]$ をマッチングする場合（共変量はZ）、XとYに関するZの条件付き独立性（CIA: Conditional Independence Assumption）が成立していることが前提。

（2）「強く無視できる割り当て」

・どちらの群に割り当てられるかは共変量の値に依存し、従属変数による割り当ての影響はあくまで「共変量と従属変数の関係」を通じてのみ間接的に依存している。

高部勲（2017）より抜粋

傾向スコア（propensity score）マッチング

共変量マッチングはすべてのデータペアの距離を計算する必要があり、
データ量が多いと計算負荷が現実的でない。

→ より簡便的なマッチングを考える（傾向スコア）

医療の死亡、
金融の倒産など
外的な2値変数が存在

傾向スコアの定義は安定していない。
2群判別のロジットZスコアの１次元で定義しているものが多い。

汎用データベースには
用いることが困難

(1) **最近隣法（nearest neighbor matching）**
- 傾向スコアでみて、最も距離の近いレコードとマッチングする

(2) **キャリパーマッチング**
- ある特定の距離以上になるときは、マッチングしない。
- マッチングできないレコードが生ずる可能性がある。

(3) **カーネルマッチング**
- 対象群のすべてのレコードの値を、カーネルの重みで利用する。
- 処置群と対象群を１対１でマッチングする際に生じる問題点（マッチングできないレコードが生じる可能性等）に対処するためにヘックマンが提案。

$$\hat{y}_{i0} = \frac{\sum_{j=1}^{N}(1 - Z_j)K_{ij}y_{j0}}{\sum_{j=1}^{N}(1 - Z_j)K_{ij}}$$

※傾向スコアについては、主に経済学の分野で、企業データに対する適用事例がある。

※あるイベントが起こった地域の企業と、起こっていない地域の類似企業をマッチングし、その差をみることで、イベントの影響について分析する事例がほとんど。

多項ロジットモデルに基づく統計的マッチング

- 多項ロジットモデルの枠組みを統計的マッチングの問題に適用】

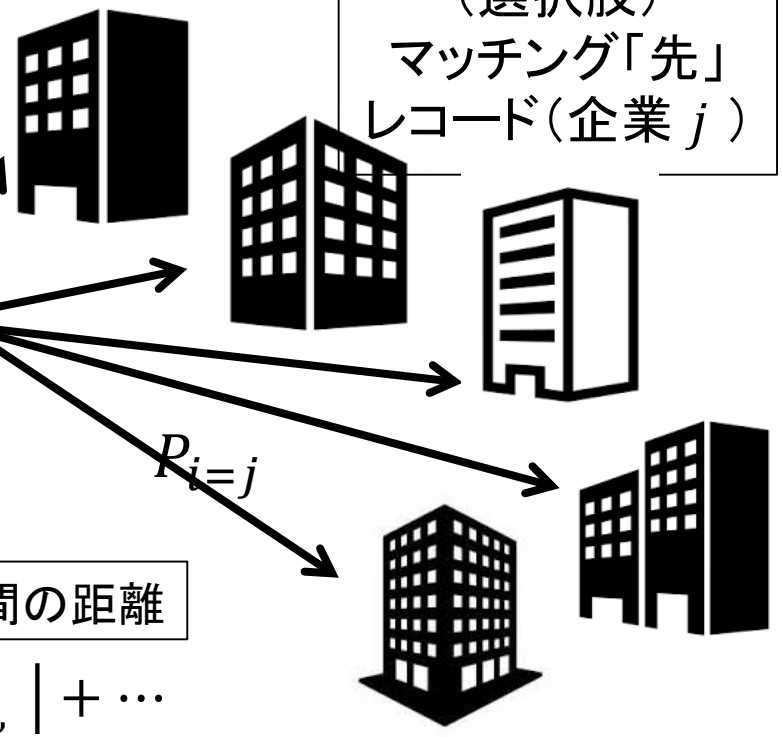
Cf.マーケティングにおける消費者選択確立推計のモデル

マッチング「元」
レコード（企業 i ）

$$P_{i=j} = \frac{exp(D_{ij})}{\sum_{j=1}^{N} exp(D_{ij})}$$

マッチング確率

（選択肢）
マッチング「先」
レコード（企業 j ）



レコード（企業）間の距離

$$D_{ij} = \beta_1 \left| X_{i,} - X_{j,} \right| + \beta_2 \left| X_{i,} - X_{j,} \right| + \cdots$$

資本金　資本金　従業者数　従業者数

マッチング確率により、レコード間リレーションを作る。
　リレーション作成ルールは様々：
　　例）レコード間で最高確率のペアをマッチング
　　→残ったレコード間で最高確率のペアをマッチング→繰り返す…
　　→ある一定の確率ではマッチングさせない

リレーション実験例

〇マッチング元：

「帝国データバンク」データ（平成24年２月分）

商品名：COSMOS Ⅱ　企業概要ファイル・レイアウトC

産業分類は、「日本標準産業分類」を基に類似のコードを付与。

地域コード(都道府県・市区町村)は政府統計のものを使用。

⇒ 質的データ項目の内容は類似している

〇マッチング先：

「平成24年経済センサス‐活動調査」 ミクロデータ（※）

（※統計法第33条による二次的利用の制度に基づき提供を受けたもの。

民間企業データベースと公的ミクロ企業データベースがくっつくかの実験

距離の計算のイメージ

		帝国データバンク【マッチング元】i			
		TDB企業１	TDB企業２	…	TDB企業M
経済センサス【マッチング先】j	EC企業１	dist(1, 1)	dist(1, 2)		dist(1, M)
	EC企業２	dist(2, 1)	dist(2, 2)	…	距離(2, M)
	⋮	⋮	⋮		⋮
	EC企業N	dist(N, 1)	dist(N, 2)	…	

民間データは主に借入をしている企業のサンプルデータ、
政府データはセンサスデータ、

お互いに独自の変数があり、統合することによって情報量が増大する。

$$P_{ij} = \frac{exp(D_{ij})}{\sum_{j=1}^{N} exp(D_{ij})}$$ 最尤法で推計

$$D_{ij} = \beta_1 \left| X_{i,\text{資本金}} - X_{j,\text{資本金}} \right| + \beta_2 \left| X_{i,\text{従業者数}} - X_{j,\text{従業者数}} \right| + \cdots$$

$$LL = log\left[\prod_{i,j} P_{ij}^{\delta_{ij}}\right] = \sum_{i,j} \delta_{ij}\, log[P_{ij}]$$

- LLを最小にするパラメータβ_1等を求める。

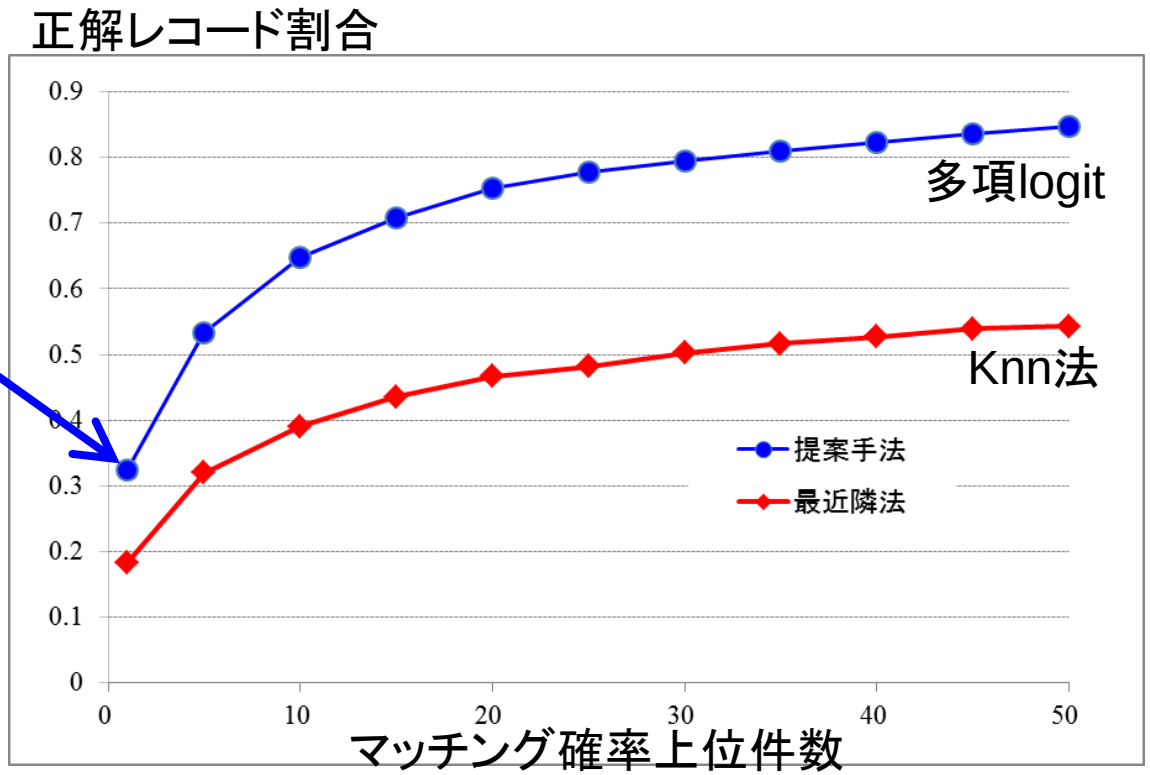
⇒ウエイトを統計的に推定することが可能。

リレーション実験例

2つのデータでマッチングが確認できたものだけで、
正解リレーションデータベースをつくり、
モデルでリレーションを再現できるかを実験。

正解割合はマッチング元から見て、正しいマッチング先を見つけた割合。

計算法の改良により1位マッチング正解率は0.55程度に上昇



社会実装を意識したリレーション実験の例
低質ビッグデータと高質少量データの融合法（アパートローンリスク）

10日ごと●●県の全数データを自動ダウンロード
→約2年分のデータ集積済み

社会実装のための運営会議を定期的（3ヶ月ごと）に開催

【モデル開発側】
統計数理研究所、国立情報学研究所、長崎大学
社会データ構造化センター

【モデル実装側】
滋賀銀行、CRD協会

CRD協会のデータ提供システムにより全都道県庁に対して開発モデルを提供を行うことで合意
（すでに全都道府県庁がCRDの会員である）

低質ビッグデータ

Web賃貸住宅募集データ（パネル）

空室→占室モデル
（OVモデル）
2項ロジットモデル

高質少量データ

3ヶ月ごと
滋賀県
の4000軒
のアパートを
パネル調査

7回調査済み
↓
アパートの経年変化を観測可能に

鑑定士によるサーベイデータ（実地調査・パネル）

占室→空室モデル
（VOモデル）
2項ロジットモデル

銀行の融資情報

部屋の初期状態
$S_{O,V}(0)$

個別部屋の状態遷移行列

前|後	占室	空室
占室	$P_{O\to O}$	$P_{O\to V}$
空室	$P_{V\to O}$	$P_{V\to V}$

t期後の状態
$P_O(t)$,$P_V(t)$の推定

・個別部屋のT年後までの収益シュミレーション
・棟全体のポートフォリオ評価
・アパートローンの貸倒確率の計量化

(1)占室→空室モデル
上記サーベイのパネルデータを元に，現在空室である部屋が翌期までに占室になる確率を2項ロジットモデルを最尤法により推計する

$$P_{i,t} = \frac{1}{1+\exp Z_{i,t}^{P}}$$

$$Z_{i,t}^{P} = \alpha^{P} + \sum_{j} \beta_{j}^{P} x_{i,j,t}$$

$$L^{*}(\alpha^{*P}, \beta_{j}^{*P}) = \arg\max_{\alpha^{P},\beta_{j}^{P}} \prod_{i,t} P_{i,t}^{\delta_{i,t}^{P}} (1-P_{i,t})^{1-\delta_{i,t}^{P}}$$

占空モデルの開発
試作版開発済み

占空モデルの予測力評価
一般的な信用リスクモデルに遜色ないレベルを達成