統計ミクロデータ利活用の意義：
-*経済的統計から統計的経営への転換*-

独立行政法人　統計センター
椿　広計

所属組織：(独)統計センター
旧総理府統計局製表部

- 1871年　太政官正院**政表課**設置（総務省統計局・（独）統計センター前身）
- 1881年　太政官統計院設置の建議と設置：初代院長　大隈重信
  - **現在ノ国勢ヲ詳明セザレバ政府則チ施政ノ便ヲ失フ**
  - **過去施政ノ結果ヲ鑑照セザレバ政府其政策ノ利弊ヲ知ルニ由ナシ**
- 1885年　内閣統計局設置
- **1920年　第1回国勢調査実施**
- **1946年　標本調査開始（家計調査・労働力調査）：Deming博士らが指導**
- 1947年　**統計法施行**・総務庁統計局・事業所統計調査開始
- 1949年　総理府統計局**製表部**
- 1984年　**総務庁統計センター**，2001年**総務省統計センター**
- 2003年　**独立行政法人統計センター**
- 2007年　**新統計法施行**
- 2008年　**政府統計共同利用システム運用開始**
- 2009年　**第1回経済センサス実施**
- 2017年　統計改革の開始：エビデンスに基づく政策立案：オンサイト拠点ミクロデータ分析
- 2018年　**統計法・統計センター法改正？**

Contents

- はじめに：EBPM(Evidence Based Policy Making)の時代？
  - マクロデータ利活用基盤の高度化
- 統計ミクロ分析データ分析の基盤
- 経済的統計から統計的経営へ
  - ビッグデータ時代で変わらないこと，変わること
- おわりに：ビッグデータ時代？
  - 組織内に残すべきモデリングの知と外注可能な機械学習

はじめに
Evidence Based Policy Making の時代？

マクロデータ利活用基盤の高度化

私見：公的統計調査は情報という形の税徴収

- 国勢調査など基幹統計調査に協力しないとどのようになるか？
  - 統計法には罰則規定がある
  - 国民の義務の一つ
- 租庸調＋「報」
  - 税を金銭で支払う
  - 税を労働で支払う（兵役：戦前は義務）
  - 税を物品で支払う（今は無い）
  - **税を情報で支払う：報告(情報には明らかな経済的価値がある)**
    - 統計以外に，届け出義務を課し徴集する情報も多数ある
    - **国をEvidenceに基づいて運営するのに必要な個人情報，法人情報**
      - 個人情報保護の枠外
      - 目的外の利用を公益性のある統計研究等にこれまでは制限
- 「報」を適切に政策利用できないのは逆に国家の怠慢
  - Evidence Based Policy Makingの必要性

データ中心政策科学への期待
Risk and Evidence Based Policy Making

- 2004/11：OECD　第1回国際フォーラム（Palermo）
  ”Statistics, Knowledge and Policy: Key Indictors to Inform　Decision Making”
  - 経済・社会・環境**政策の質評価指標の設定**に基づく**パフォーマンスの定量的評価**
    （**KPI**☚Balanced Score Card (Kaplan)☚方針管理(コマツの旗管理)）
- **2006/10/26：Scientific Advice, Risk and Evidence Based Policy Making**
  - 英国下院科学技術委員会
    - 勧告14: **政策立案者の要件：科学的方法，ピアレビュー，証拠の様々な種類の関連性の役割と重要性，それらを評価する方法に対する基本的理解**
    - 勧告15: **証拠の利用と分析の専門的技量**の強調.
    - **効果的政策決定の前提条件：**
      - 科学的方法＋多様な証拠の解釈
        ➡科学的インプットと分析に基づく十全に説明された要請が**総合職である高級官僚**の中で展開
    - 勧告16: **総合職と共に科学的力量を持つ専門職**の価値
      - 経済的知識や法案作成能力同様，**科学的リテラシーが政策決定には必要**.
    - 勧告50: **透明性**：**政策決定に用いられる全ての証拠とそれをどのように使ったかの公表**

Precautionary Principle:
「予防原則」の代わりにRisk Basedの強調

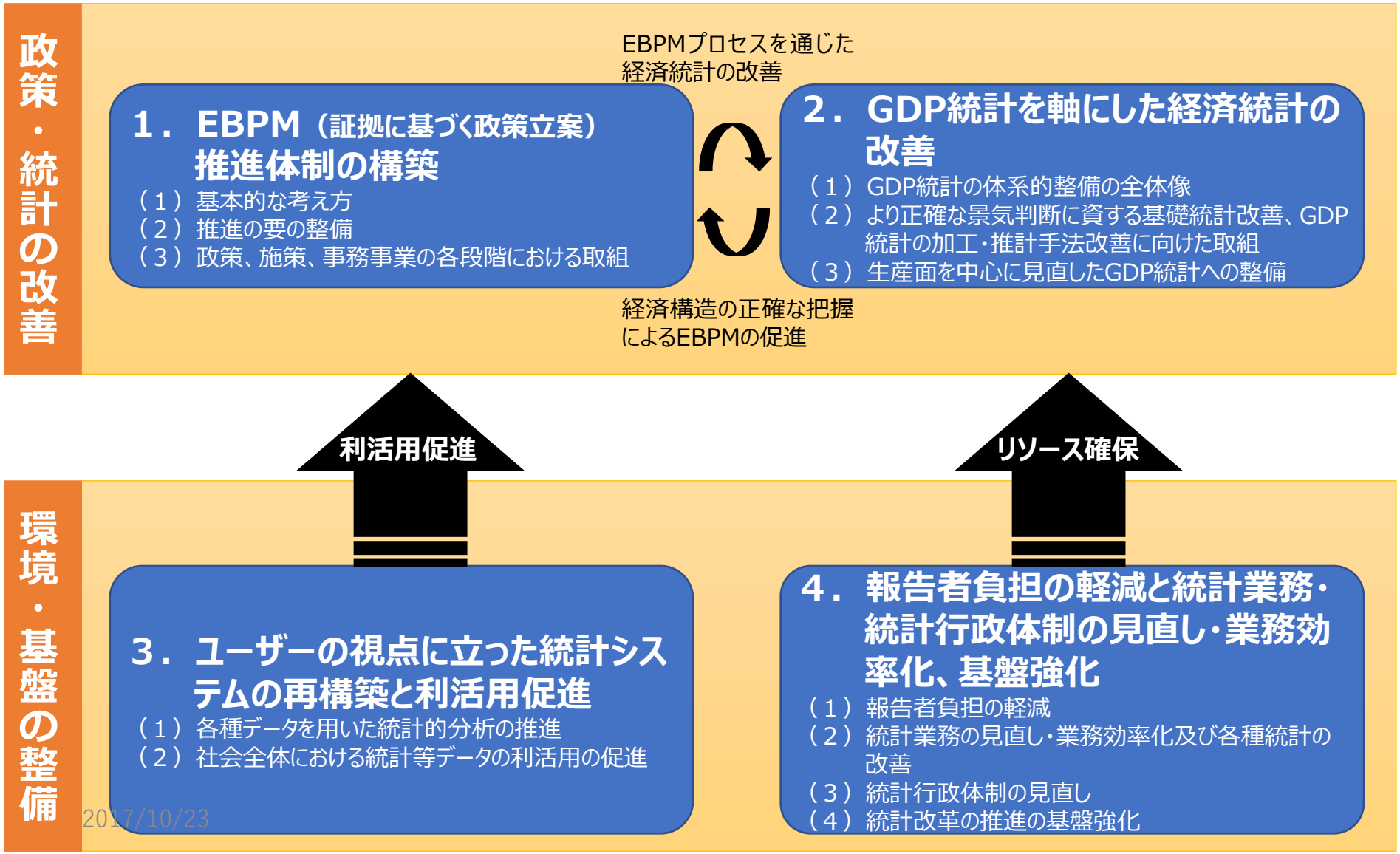
- 勧告60：「予防原則」を政策ガイダンスから外し，
  代わりに正しい予防アプローチの適用
  - リスクとベネフィット，不確かさの一貫した説明
  - 予防原則を科学的知見が不確かなときに，
    徹底したリスク分析に替わる政策決定原理として用いるべきではない
  - 翌年勧告でも確認
- 勧告61：科学的不確かさの下での
  リスクマネジメント理論の実用化と
  意思決定過程の効果的コミュニケーション
  - 参考文献
    - https://publications.parliament.uk/pa/cm200506/cmselect/cmsctech/900/900-i.pdf

わが国公的統計分野の歩み
2009年統計法全面改訂

- 旧統計法（1947年）法の目的
  - 第一条 この法律は、**統計の真実性**を確保し、統計調査の重複を除き、統計の体系を整備 し、及び統計制度の改善発達を図ることを目的とする。
- 新統計法第一条
  - この法律は、**公的統計が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**であることにかんがみ、公的統計の作成及び提供に関し基本となる事項を定めることにより、公的統計の体系的かつ効率的な整備及びその有用性の確保を図り、もって**国民経済の健全な発展及び国民生活の向上に寄与**することを目的とする。

統計改革推進会議最終取りまとめの全体構成（イメージ）[内閣官房2017年5月25日]

以下3スライド：内閣官房最終とりまとめ参考資料より抜粋
http://www.kantei.go.jp/jp/singi/toukeikaikaku/pdf/saishu_sankou.pdf

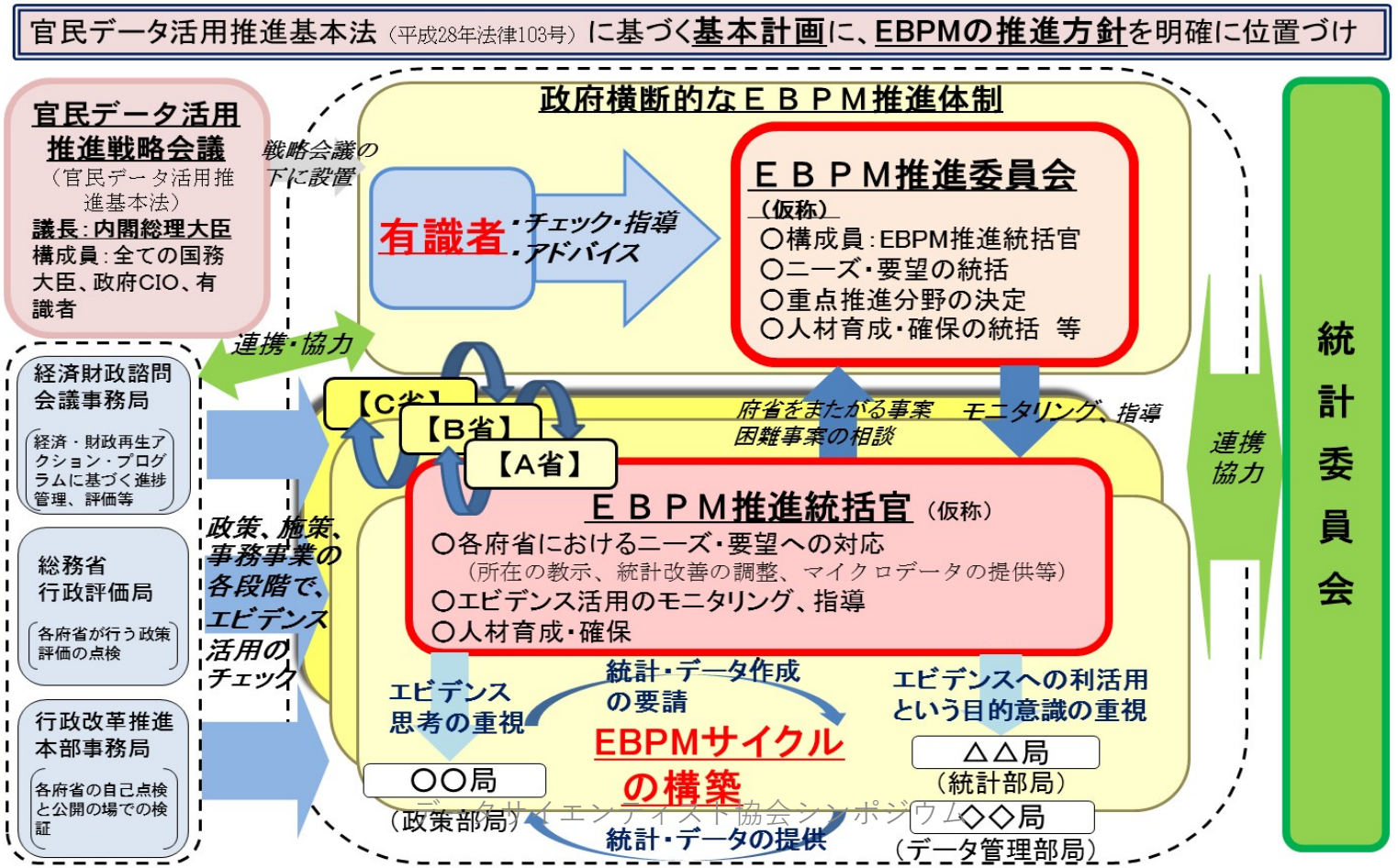


エビデンス・ベースト・ポリシー・メイキング

EBPM（証拠に基づく政策立案）推進体制の構築

内閣官房
2017/05/25

- EBPM（証拠に基づく政策立案）を推進する体制を政府内に構築
- これにより、政策部局による統計・データの利活用と統計部局によるニーズを反映した統計・データの改善が連動する「EBPMサイクル」を確立



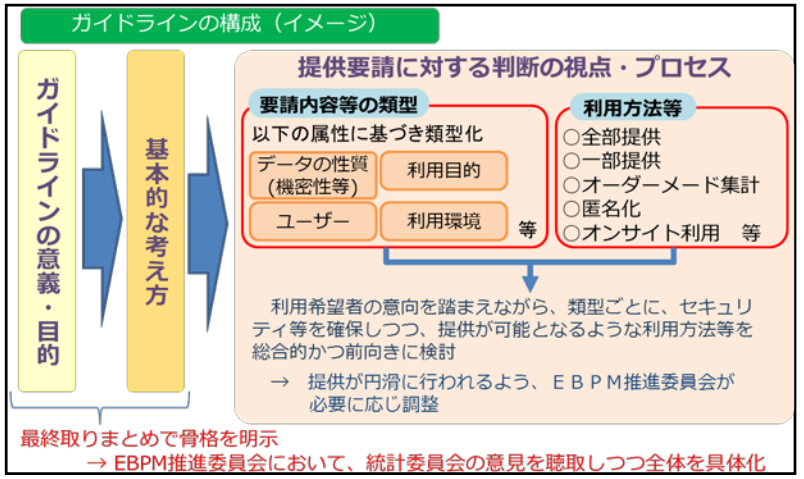
ユーザーの視点に立った統計システムの再構築と利活用促進

内閣官房
2017/05/25

各種データの利活用推進のための統計関係法制等の見直し・整備

- 利活用が限定的な統計の個票データや、利活用規定の不十分な行政記録情報や地方自治体・民間が保有する各種データについて、**セキュリティを確保しつつ、利活用を促進**するための**統計関係法制等の見直し・整備**

【具体的な取組（例）】
- 地方自治体・民間が保有する各種データへの各府省による**提供要請**や提供された**データの保護**、統計委員会を通じた各府省と地方自治体・民間の**あっせん**等の仕組の整備
- 各府省が、未提供のデータ等の提供要請を受けた場合に、データの性質や利用目的等の類型に応じ、**提供の判断を適切に行うためのガイドライン**の策定



社会全体における統計等データの利活用の促進

- EBPM推進委員会が、統計等データの**ユーザーからの提案募集**を実施
- 各府省のEBPM推進統括官の下、外部からの統計等データの**問合せや要望への対応**のための体制を整備
- 統計等データの利活用のための**基盤の整備**

【具体的な取組（例）】
- **e-Stat（政府統計の総合窓口）** への行政記録情報の検索機能追加など、抜本的な機能強化
- **オンサイト施設**（p.12参照）の整備の推進
- 一般の人が利用できる**匿名データ（匿名化した個票データ）**の提供
- **行政記録情報**の標準化・電子化

独立行政法人
統計センター
統計を
つくる・
活かす・
支える
政府中央統計機関の一翼を担う
ビッグデータ時代の支援活動
「統計を活かす」ミクロデータの研究利活用

オンサイト拠点試行事業開始
2017/01

データ利活用センター2018/04
和歌山市に設置

一般利用可能なミクロデータ開発

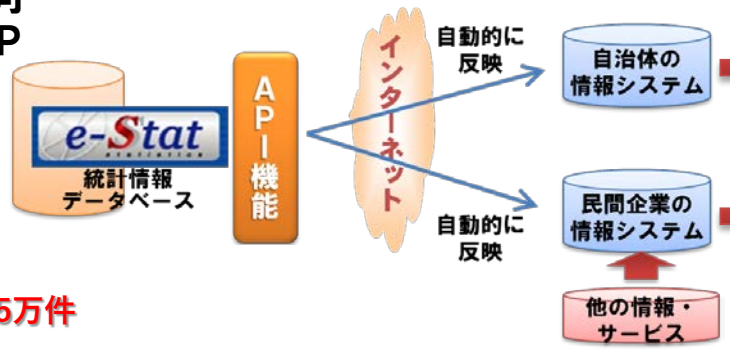
統計におけるオープンデータの高度化

統計データの提供方法を高度化し、新たな付加価値を創造するサービスや革新的な事業の創出などを支援する取り組みを、総務省統計局と連携し行っており、政府が取り組んでいるオープンデータの推進を先導。

ＡＰＩ機能による統計データの提供

2014.10.31から運用開始

統計データを機械判読可能な形式で提供するＡＰＩ機能（Application Programming Interface）を提供中



利用登録者数5816名
APIリクエスト件数：7145万件
2016年12月31日現在

活用例１：利用者の情報システムにe-Statのデータを自動的に反映

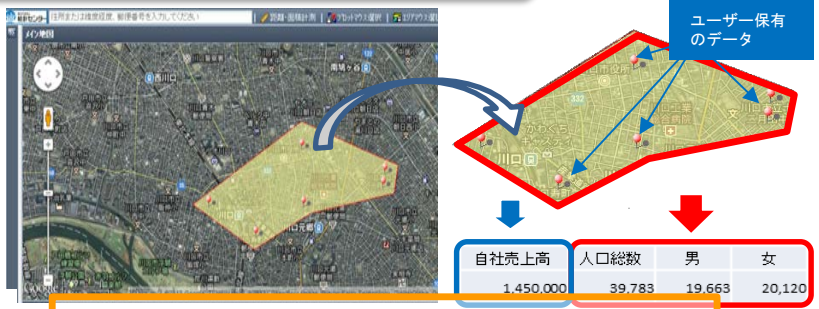
活用例２：ユーザー保有やインターネット上のデータ等と連動させた高度な統計データ分析

開発支援情報も提供中

地図による小地域分析（jSTAT MAP）

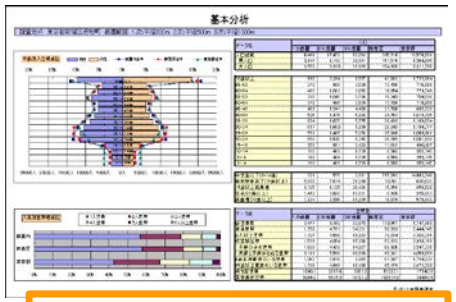
2015.1.20から運用開始

任意に指定したエリアによる集計や利用者が保有するデータの取り込み集計する機能などを提供



利用登録者数　19960名
ログイン件数：34万件
2016年12月31日現在

活用例１：任意に指定したエリアによる集計や、利用者が保有するデータと統計データを組み合わせ、集計結果を地図上で視覚的に把握可能



活用例２：選択したエリアの年齢構成等の基本的な分析結果のレポート作成

タブレット用アプリも提供中

地図による小地域分析（jSTAT MAP）　ニーズに沿った小地域分析が可能です！

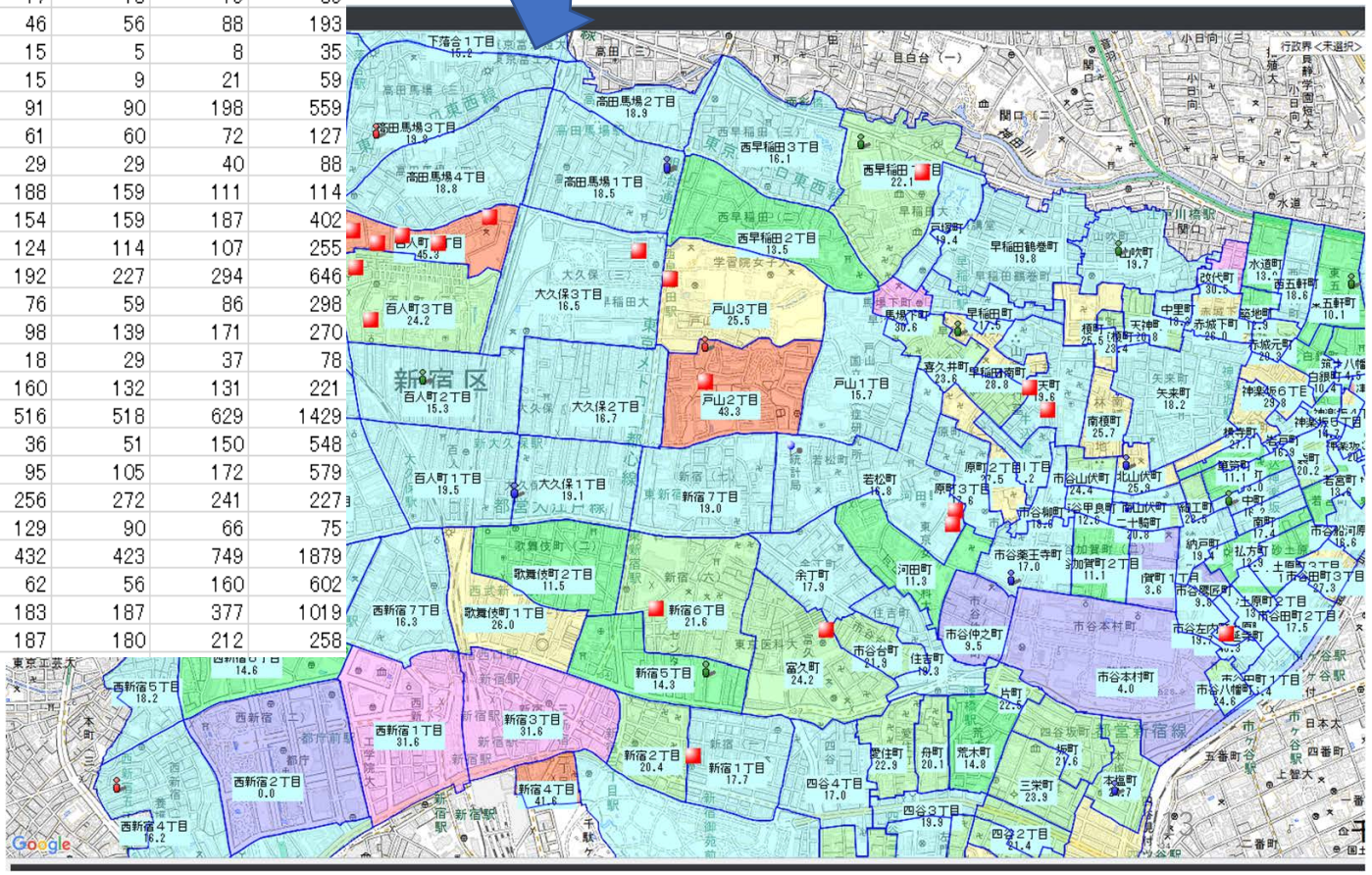
統計表は数字が並べられている状態・・

平成27年国勢調査　小地域集計　（総務省統計局）
第3表　年齢(5歳階級)，男女別人口，総年齢及び平均年齢(外国人一特掲)－町丁・字等

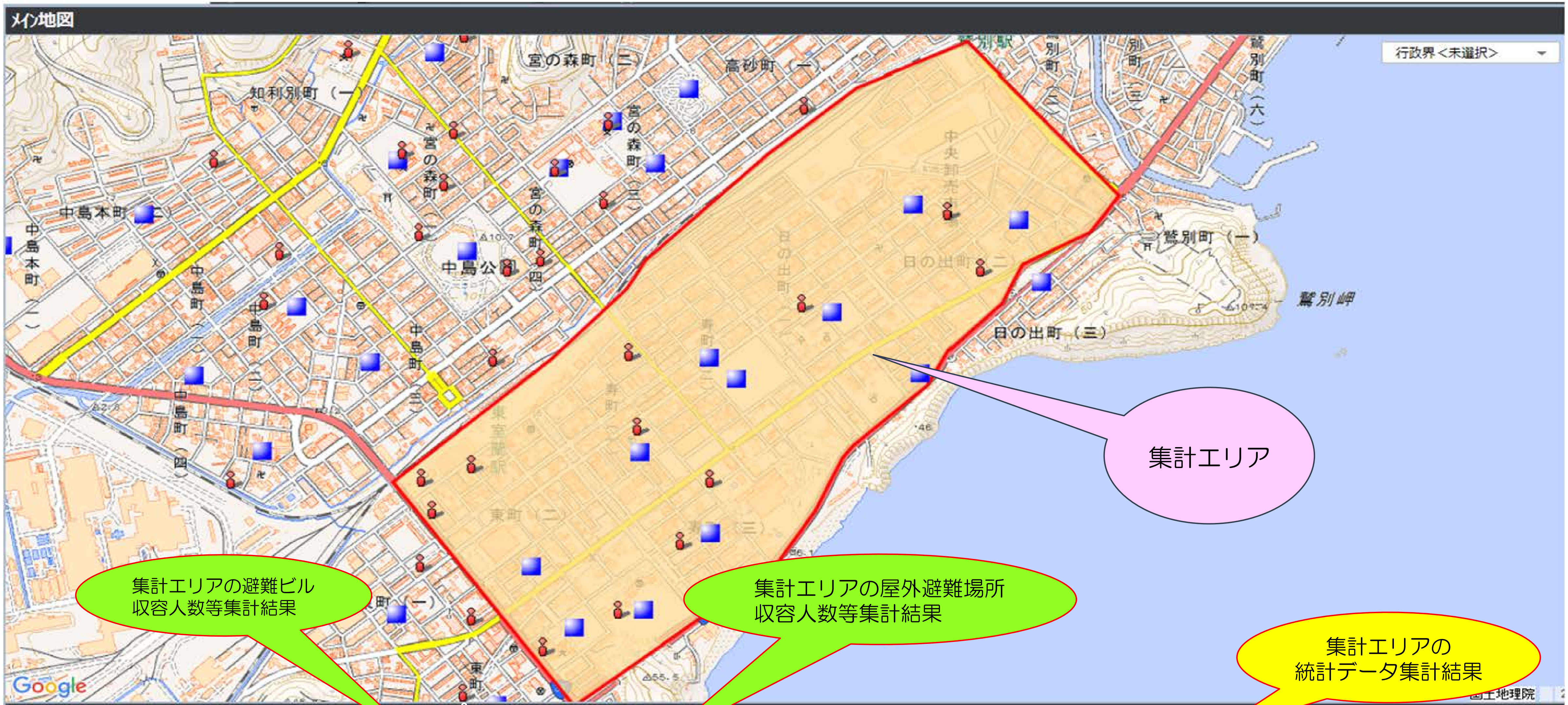
1) 日本人・外国人の別「不詳」を含む。
2) 無国籍及び国名「不詳」を含む。
総数(男女別)

市区町村コ	町丁字コー	地域識別番	都道府県名	市区町村名	大字・町名	字・丁目名	総数(年齢	0～4歳	5～9歳	10～14歳	15～19歳	20～24歳
13104	680	2	東京都	新宿区	喜久井町		1953	56	52	58	93	163
13104	690	2	東京都	新宿区	築地町		558	25	13	10	14	32
13104	700	2	東京都	新宿区	弁天町		3380	104	80	92	124	235
13104	710	2	東京都	新宿区	中里町		707	40	17	13	19	35
13104	720	2	東京都	新宿区	山吹町		3151	80	46	56	88	193
13104	730	2	東京都	新宿区	改代町		571	34	15	5	8	35
13104	740	2	東京都	新宿区	水道町		895	16	15	9	21	59
13104	750	2	東京都	新宿区	早稲田鶴巻町		5298	137	91	90	198	559
13104	760	2	東京都	新宿区	住吉町		2586	63	61	60	72	127
13104	770	2	東京都	新宿区	市谷台町		1301	45	29	29	40	88
13104	780	2	東京都	新宿区	河田町		2905	167	188	159	111	114
13104	790	2	東京都	新宿区	若松町		5443	206	154	159	187	402
13104	800	2	東京都	新宿区	余丁町		3904	160	124	114	107	255
13104	810	2	東京都	新宿区	戸山		9480	127	192	227	294	646
13104	81001	3	東京都	新宿区	戸山	1丁目	2590	65	76	59	86	298
13104	81002	3	東京都	新宿区	戸山	2丁目	5940	45	98	139	171	270
13104	81003	3	東京都	新宿区	戸山	3丁目	950	17	18	29	37	78
13104	820	2	東京都	新宿区	富久町		5729	291	160	132	131	221
13104	830	2	東京都	新宿区	百人町		17668	488	516	518	629	1429
13104	83001	3	東京都	新宿区	百人町	1丁目	4443	48	36	51	150	548
13104	83002	3	東京都	新宿区	百人町	2丁目	5004	115	95	105	172	579
13104	83003	3	東京都	新宿区	百人町	3丁目	5407	219	256	272	241	227
13104	83004	3	東京都	新宿区	百人町	4丁目	2814	106	129	90	66	75
13104	840	2	東京都	新宿区	大久保		16925	463	432	423	749	1879
13104	84001	3	東京都	新宿区	大久保	1丁目	4402	70	62	56	160	602
13104	84002	3	東京都	新宿区	大久保	2丁目	8442	214	183	187	377	1019
13104	84003	3	東京都	新宿区	大久保	3丁目	4081	179	180	212	258	

地図に載せることにより
様々なことが見えてきます！



★利用者の保有するデータを取り込み集計する機能（任意エリアでの集計ープロットとの比較）



150219室蘭市避難施設（1件）　プロット　平成27年国勢調査小地域集計

エリア（按分集計）（合計）	ﾌﾟﾛｯﾄ数[150219室 13	収容人数[150219 1472	ﾌﾟﾛｯﾄ数[150219室 11	収容人数[150219 17287	人口総数 6555	総数０～４歳 255	総数５～９歳 255	総数６５歳以上 1711	総数７５歳以上 832
任意エリア	13	1472	11	17287	6555	255	255	1711	832

《jSTAT MAPの照会先》統計センターGIS担当　E-Mail: statGIS_supportdesk@nstac.go.jp

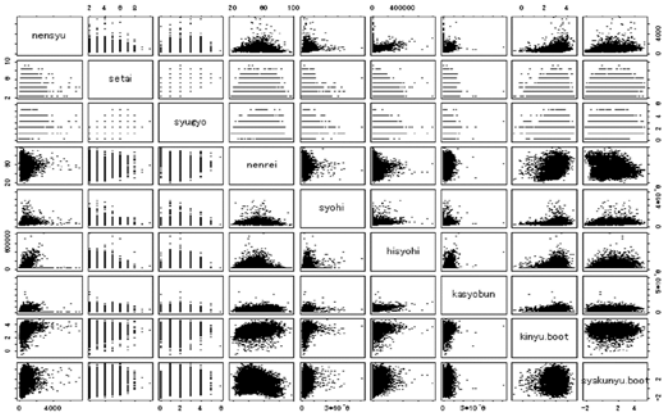
データ出典：室蘭市HP

オンサイト拠点への途

2017年1月試行運用開始：4拠点をつなぐ

人間・社会分野研究者の公的統計ミクロデータ利活用のニーズ

- 国際競争力を低下させる日本の社会科学研究
  - 1980年代欧米で大きな変化
    - 自国公的統計ミクロ（個票）データの研究利用が可能に
    - 経済・社会研究の中心が集計データからミクロデータ分析に
  - 日本の経済学者の危機感
    - 科研費特定領域:1996-1999
    - **統計情報活用のフロンティアの拡大の総括的研究:**
      - **松田芳郎（一橋大）**
      - 13研究班の要請を集約：5省庁17調査「目的外使用申請」
      - データ処理センター活動→現行の**オーダーメード集計**制度
      - リサンプリングデータの作成→現行の**匿名データ提供**制度
      - イミテーションデータの作成→現行の**教育用擬似データ**制度
      - データ処理結果および処理要求をデータの秘密保持を保ったうえで各研究班相互を計算機ネットワーク化して計算結果等の情報を流通させるシステムを開発
        →今後の**リモートアクセス**制度
        - 2000年：日本評論社：講座ミクロデータ分析シリーズ4巻により啓発
  - 新統計法下のミクロデータ研究利用の先駆け研究



統計データの二次（研究）利用促進に関する経緯

- 2007年総務省政策統括官室：統計データの二次利用促進に関する研究会
  - 廣松毅（東京大学，現情報セキュリティ大学院大学）座長
  - 統計法改正に向けた二次利用のスタイルと展開
    - *オーダーメード集計・匿名データ・疑似ミクロデータ*の提供
    - 利用目的としての公益性
- 2009年統計法公布後の研究会活動
  ➡内閣府統計委員会（現総務省統計委員会）へのインプット
  - ミクロデータ提供方法：各国制度比較と**日本の特異性**
    - **個人情報・法人情報が付随したデータをセキュアな環境を持たない研究者が管理**する可能性
  - オンサイト拠点設置➡**ミクロデータの全情報の探索的・創造的モデリング**を可能にする
    - 目的外申請で個票をセキュアな監視環境下で分析➡各拠点ごとの人員・設備整備にかなりなコスト
  - リモートアクセス型オンサイト拠点での分析ネットワーク形成
    - オンサイト拠点にはデータを置かず，**中央機関で一括監視・管理**
    - **利用計画事前審査**から**分析結果持出し審査**へ

「公的統計の整備に関する基本計画」への組み込み

- 2014年1月31日内閣府統計委員会答申
- 2014年3月25日閣議決定
  - 調査票情報等の提供及び活用については、
    セキュリティレベルや調査票情報等の匿名性の程度に応じた利用形態ごとの特性,
    諸外国における取組状況等を総合的に勘案した上法制度上の整理を含め、
    以下の取組を行う
    - オーダーメード集計における利用条件の緩和に向けた検討
    - 調査票情報の提供におけるリモートアクセスを含むオンサイト利用やプログラム送付型集計・分析の実現に向けた整理・ 検討
    - 匿名データの作成及び提供における提供対象統計調査の種類や年次 の追加等による
      サービスの充実
  - 効率性及び利便性の観点から、
    政府一体として一元的な取組を推進

現行公的統計データ研究利用（二次利用）の種類と利用要件の緩和
➡2018年度以降の統計法改正にも注目

利用形態	根拠	利用できる者	利用目的
調査票情報の二次利用	法第32条	調査を実施した各府省等（行政機関、独法等）自身が利用する場合	統計の作成 統計的研究 調査名簿の作成
調査票情報・ミクロデータ提供 探索的分析	法第33条第１号	公的機関（行政機関等＋会計検査院、地方独法等）が利用する場合	
	法第33条第２号 リモートアクセス型オンサイト利用を主流に	公的機関が委託又は共同して調査研究を行う者 公的機関が公募の方法により補助する調査研究を行う者 行政機関等（行政機関＋地方公共団体、独法等）が**政策の企画・立案、実施又は評価に有用であると認める統計の作成**等を行う者	統計の作成 **統計的研究**
オーダーメード集計	法第34条	一般の者（民間も含む） ※企業活動の一環としての研究も可 ・学術研究等の目的に限定 ・研究成果の公表義務	研究 高等教育
匿名データ	法第36条		

遠い未来か？

オンサイト拠点以外でのミクロデータ活用環境の研究：秘密計算・秘密分散

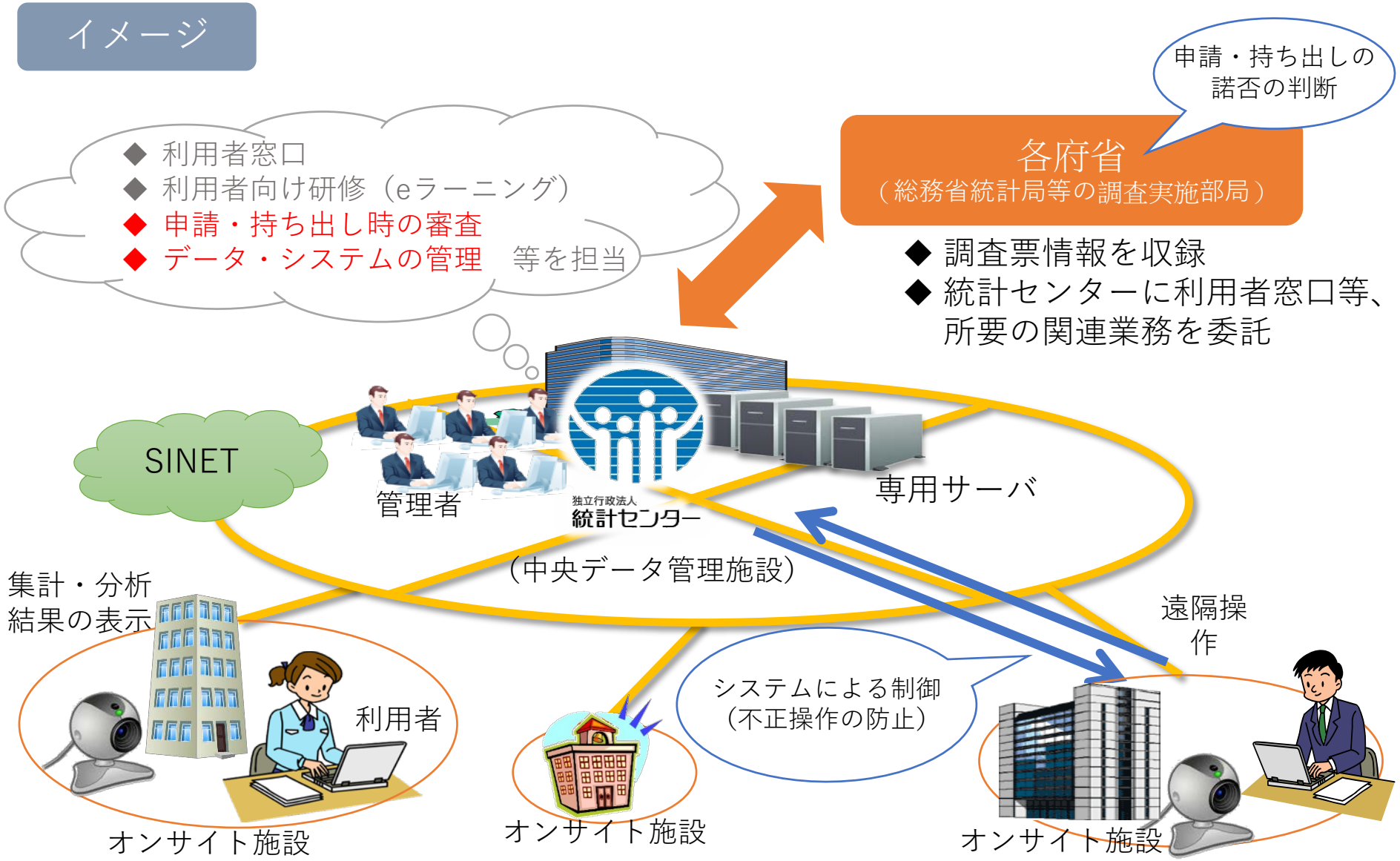
近い将来か？

利用範囲拡大か？？

PUF開発！

※オーダーメード集計及び匿名データの提供は有料サービス：無料の教育用一般ミクロデータも開発（擬似データ）

リモートアクセスを活用したオンサイト利用



リモートアクセス型 オンサイト拠点とこれまでのDVDデータ提供との比較

DVD提供方式

- **共通条件**
  - 公益性の高い学術研究
    - 科研費・大学共同利用機関公募型研究・地方公共団体のための研究
  - 公的統計ミクロデータあるいは自身保有のデータとのリンケージ可能
- **DVD独自条件**
  - 事前の利用環境審査：分析環境のセキュリティ審査
  - 自身の指定した場で分析可能
  - 詳細な分析事前計画提出：集計表のイメージ
  - 分析計画に必要な変数のみDVDで提供
  - 計画通りの分析と公表
  - 分析後にデータ廃棄

リモートアクセス型オンサイト方式

- 事前の審査の大幅簡易化→最終・中間生成物の持ち出し審査
- 全変数を用いた自由な探索的分析：事前の変数絞り込み不要：全分析経過・試行錯誤・ミクロデータは施設内では目視可能
- 認可を受けたオンサイト施設での仮想PC利用：分析ログを確認可能
- 場所の制約・時間の制約（オンサイト拠点運営組織に依存）
- **研究中の秘密計算オンサイト**
  - 場所や時間の自由疎
  - データ自体は眺められない

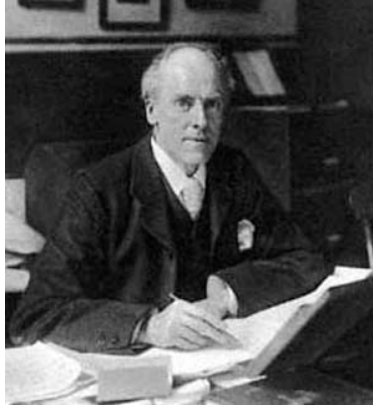
直近の取組み

- 2016年3月　（独）統計センター
  - 機能検証施設設置
- 2016年3月：情報・システム研究機構公的統計ミクロデータ研究コンソーシアム設立
  - 2016年4月：事務局：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
  - 2016年8月：コンソーシアム評議会
  - 一橋大学，神戸大学，滋賀大学，京都大学，広島大学，群馬大学
- 2017年1月：統計センター
  - 機能検証＋運用管理施設設置
  - 分析結果持ち出し基準の運用と改善作業
  - **一橋大、神戸大オンサイト試行拠点との接続**
  - 試行運用の開始
- 2017年10月現在 2か所オンサイト施設設置試行運用追加
  - 情報・システム研究機構、滋賀大
  - 総務省基幹統計調査利用可能
- 2018年1月から本格運用
- 2018年中に設置個所を10施設程度に拡大
- 2018年4月：**和歌山市に総務省統計局統計データ利活用センター**設置
- 2020年1月：システム更新
  - 設置拠点拡大
- 2019年～：調査票情報の順次拡大
  - 経済産業省参画予定，厚生労働省は？
- 将来
  - 研究室オンサイト（秘密計算）？
    - 一橋大学での試行実験開始

経済統計から統計的経営へ

ビッグデータ時代で変わること変わらないこと

科学の文法(K. Pearson, 1892)
プロセスに基づく認識科学の定義



https://en.wikipedia.org/wiki/Karl_Pearson

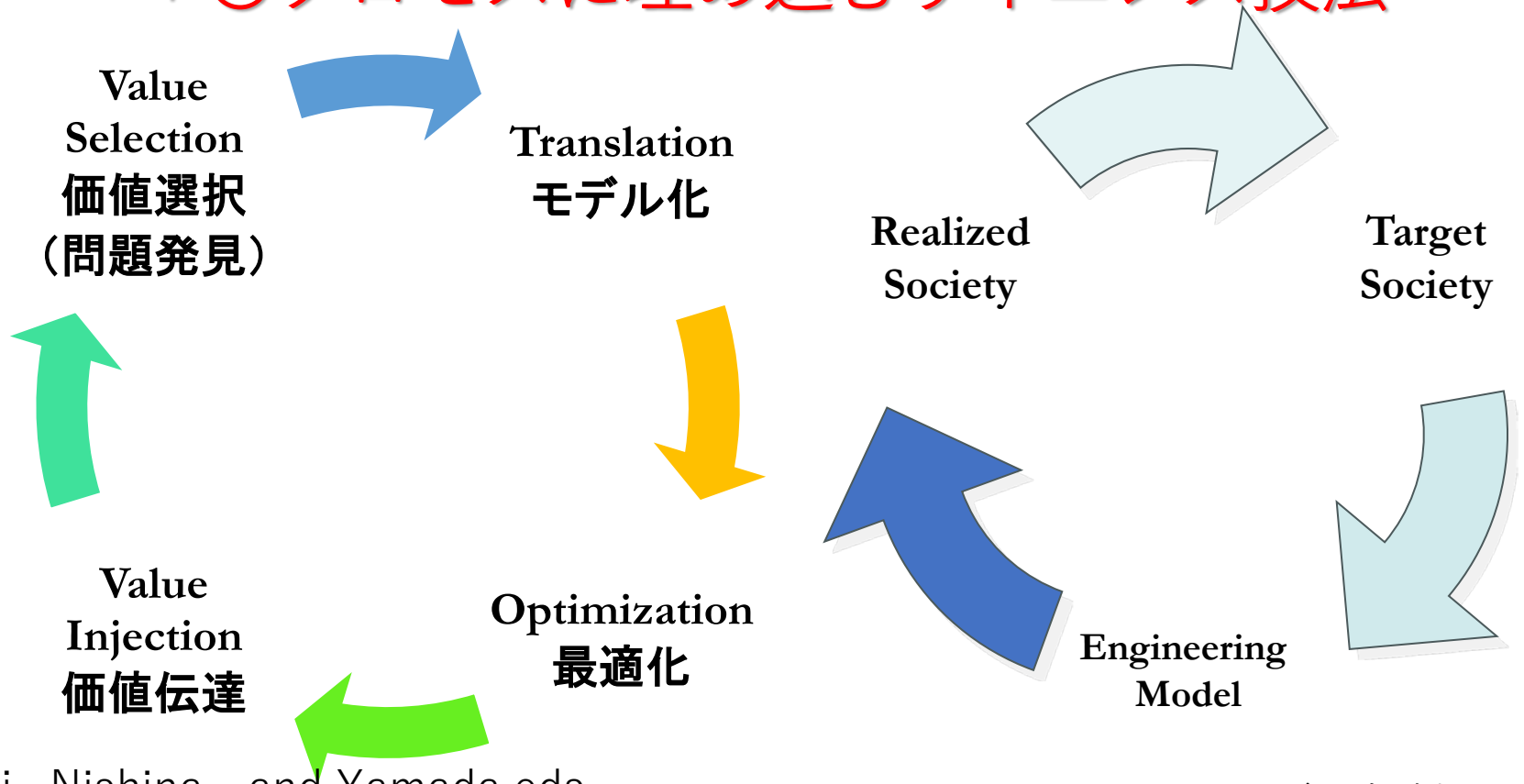
- *科学の適用範囲（Scope）*
  - *科学を特徴づけるのは対象ではなく，そのプロセスである.*
  - *あらゆる知的方法を用いて真実を確かめること*
- *科学的プロセス*
  - ***分解：Analysis**; 事実の周到な分類と事実間の関連性と順序の観察*
  - ***総合：Synthesis**; 創造的想像に支援された科学的法則の発見*
  - ***妥当性検証：Validation**; “自己批判と全ての人が等しく妥当性を認めるか否かを検証*
- *科学的プロセス支援技法*
  - *ピアソンの統計的方法開発*
    - *視覚に訴える：ヒストグラム，散布図*
    - *指標化：標準偏差，相関係数*
    - *妥当性検証：モデルの適合度検定*
- *産業界の指導：ギネスビールのゴセットへの指導*

認識のための科学から設計・デザインのための科学へ！

「設計科学の文法」＝必要な「情報循環」構築の標準シナリオ

文法＝　◎プロセスモデル（マネジメントの基本）

＋〇プロセスに埋め込むサイエンス技法



Tsubaki, Nishina and Yamada eds.
The Grammar of Technology Development
Springer, 2008.

ISO 16355シリーズ：新製品・市技術開発のための統計・統計関連技法

品質経営の文法：1951年以降日本（JUSE/QCRG)で確立
デミング・石川のマネジメント・プロセスモデル＋支援技法
Japanese PDCA：日常管理(Control)と改善行動(Improvement)の合体

改善・改革の標準シナリオ

以下，8スライド：一般社団法人日本品質管理学会監修総務省研修「データに基づく問題解決」資料より抜粋

問題解決の標準シナリオ
（問題解決型QCストーリー）

- テーマ（問題）の選定
- 現状の把握と目標の設定
- 要因の解析
- 対策の立案
- 効果の確認
- 標準化と必要組織への展開

課題達成の標準シナリオ
（課題達成型QCストーリー）

- 経営課題の確認
- 課題の明確化と目標の設定
- 方策の立案
- 成功シナリオの追求
- 成功シナリオの実施
- 効果の確認
- 標準化と必要組織への展開

問題発見のための統計的方法

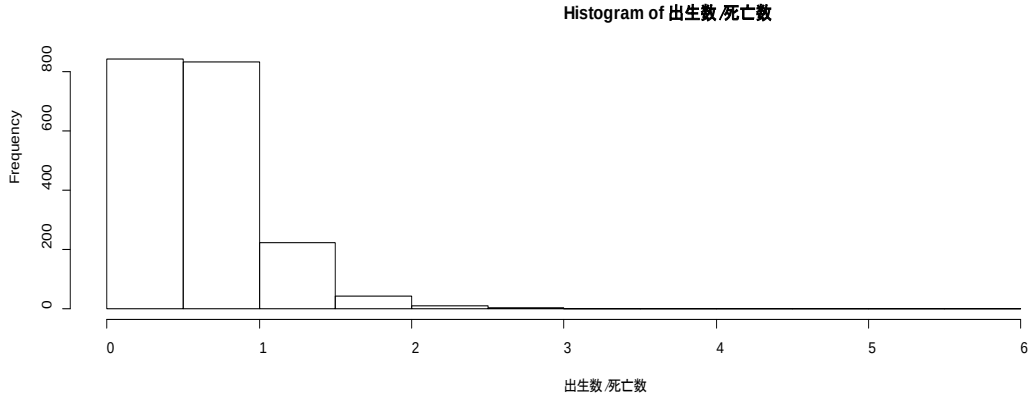
- 問題の発見は**異常検知**
  - 常態ではない状況（異常）の発見
    - **空間的（地域的）異常**
      - この地域が特に高い又は低い
    - 時系列的・時間的に異常
      - この地域では、ある値が、この時点で急上昇した
        - 守口の花火大会でPM2,5急上昇
        - 1998年日本の自殺率急上昇
    - 属性的異常（要因分析に関連）
      - この属性を有する世帯の値が高い
        - 単身世帯の自殺率が非常に高い
    - 関係性の異常（要因分析に関連）
      - これだけ予算を投下しているのにこの地域では効果が出ない
  - **異常には良い異常もある**
    - 悪い異常
      - 自殺のホットスポット（多発地域）
    - 良い異常
      - 自殺のクールスポット

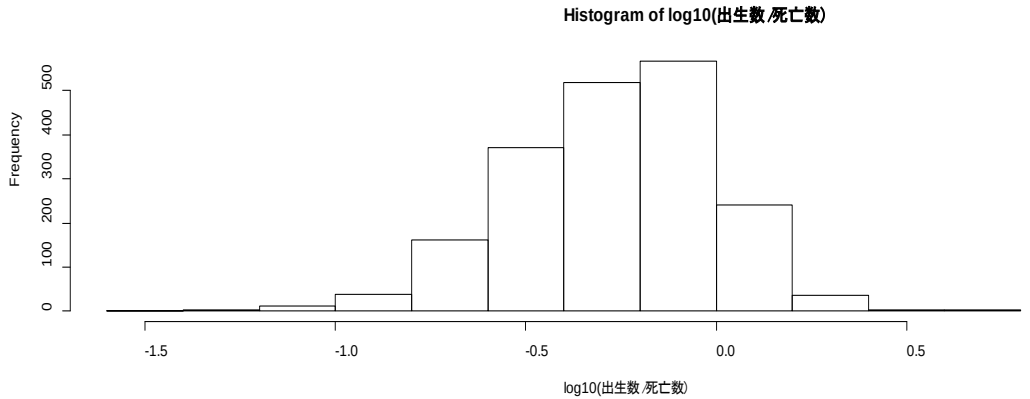
- 改善活動
  - **問題の解決＝異常原因への対策**
    - 悪い異常の原因は除去
    - 良い異常の原因を標準にする
- 改善活動の効果とは？
  - 常態自体の平均が適正化
  - 常態自体のばらつきが減少

地域的異常の検知とQC七つ道具

- 地域横断的データの収集
  - データをヒストグラムに表現
    - 明らかに集団から外れている異常地域の摘出
      - 人口0地域（福島第一原発周辺）
    - 対称で一山の分布になる場合
      - 山の両裾野の部分の地域異常候補として分析
    - 対称で幾つかの山がある場合
      - 要因分析に進め！
    - J字型、逆J字型の分布
      - データを適切に変換して対称一山分布を導く

社会・人口統計体系：全国市町村（n=1957）の出生者数/死亡者数のヒストグラム
下段は、出生者数/死亡者数を常用対数変換





異常検知の方法：
上位・下位少数例に学ぶ

- 出生数が死亡数より多い良い異常地域
  - 一例：出生数/死亡数のベスト20
    (上位約1%)自治体を抽出

利島村	**小笠原村**	**御蔵島村**
北大東村	南風原町	長久手市
与那原町	栗東市	粕屋町
豊見城市	宜野湾市	中央区
菊陽町	*野々市市*	浦添市
西原町	新宮町	中原区
みよし市	和光市	

    - 太字：島部、下線：沖縄県、
      斜字：ベットタウン化
    - その他：長久手市：日本一若い街

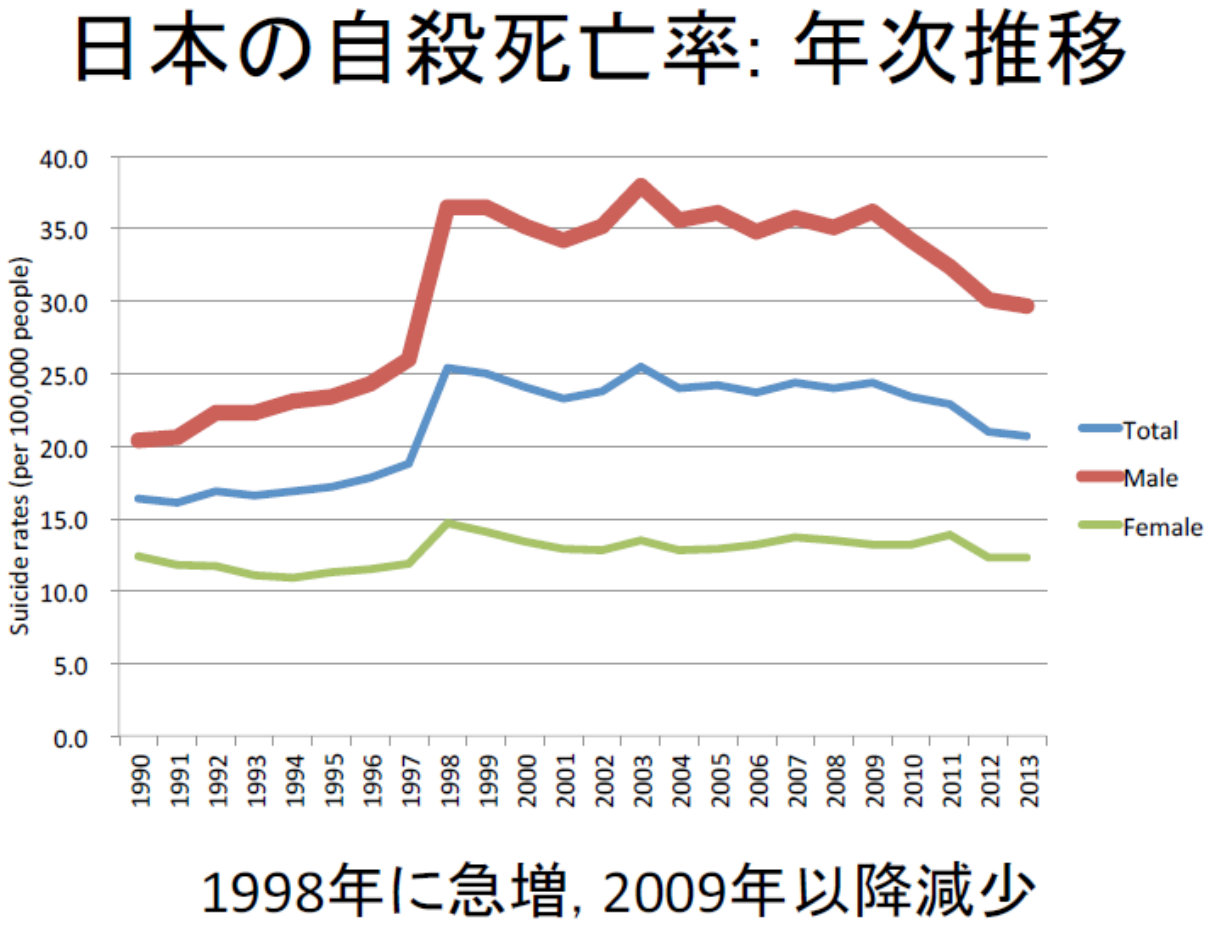
- 出生数が死亡数より少ない異常地域
  - 一例：出生数/死亡数の下位20
    (下位約1%)自治体を抽出

黒滝村	上北山村	川上村
奥多摩町	上小阿仁村	七ケ宿町
早川町（最小人口町）	南牧村	今別町
平谷村	東吉野村	馬路村
東洋町	大豊町	金山町
伊根町	外ヶ浜町	根羽村
夕張市	西伊豆町	

    - 中山間地を擁する自治体が多い
    - 孤立しやすい自治体
    - 下線部は海岸も擁する

時間的異常検知による問題発見

- 時系列グラフを描く
  - 地域内・県レベル・全国集計
    - 様々なグラフが考えられる
  - 急変化点を異常現象として抽出
    - 要因分析につなげる
      - 1998年急増は経済的要因？
        - 都市域50台男性自殺急増
      - 経済回復後も高水準に推移
        - 2007年自殺総合対策大綱
          - うつ病対策から働き方の見直しなど社会的要因（失業、倒産、多重債務、長時間労働等）への対策

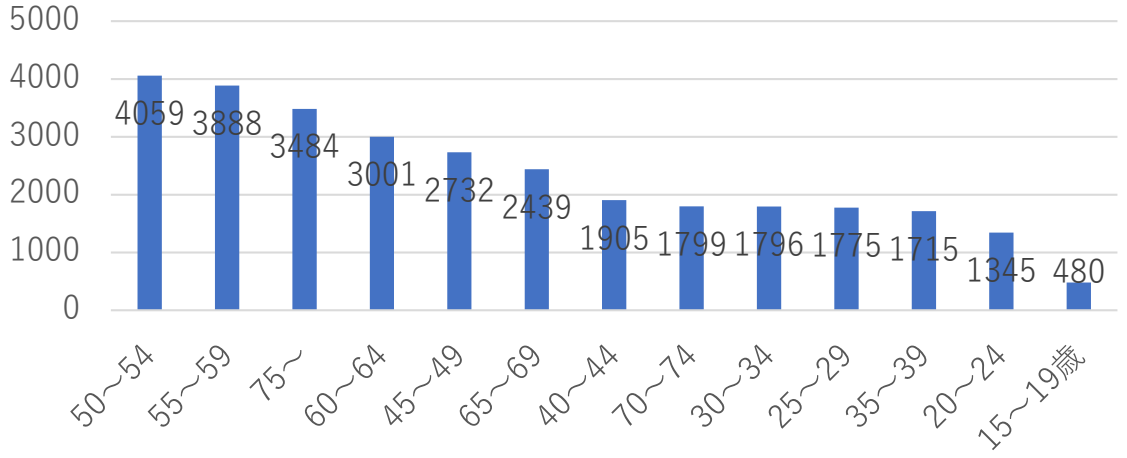


厚生労働省：人口動態統計より作成

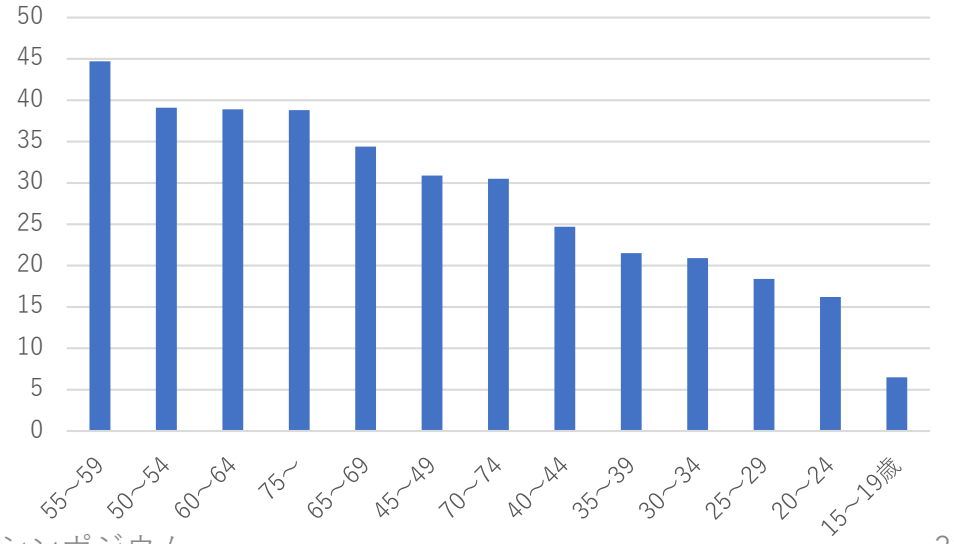
属性別問題発見：
現状把握分析を進めることで
重点となる問題の絞り込み

- パレート分析：重点指向
  - 属性別問題事象件数の把握
  - 当該問題がある属性に集中していないかを検討
  - 集中度が高い属性に対して、要因分析を進め、原因を追及、改善策立案につなげる
- 性・年令・家族類型
  - どの属性に対して対策を打つのが効果的かを調べる
    - 50-60歳ないしは50歳以上の自殺

厚生労働省人口動態統計特殊報告：自殺死亡統計
平成12年度自殺者数年令別集計



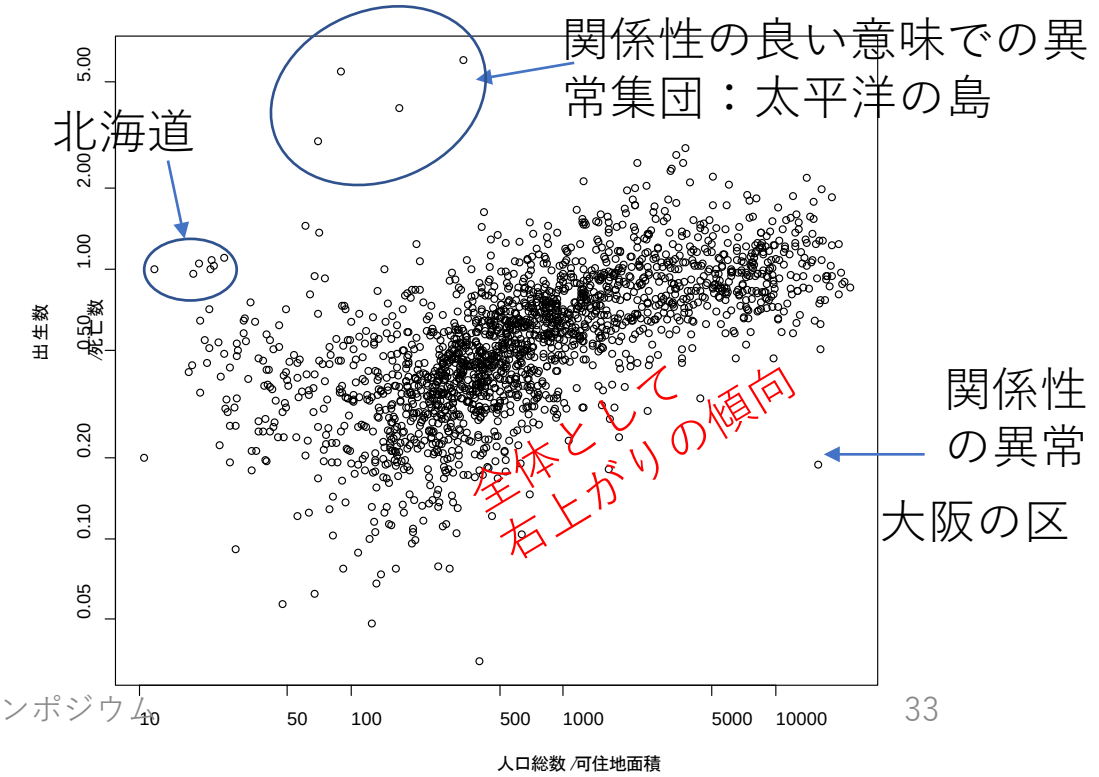
平成12年度自殺死亡率（10万人当たり）



要因の分析と統計的方法の役割
定性的な要因分析から定量的要因分析へ

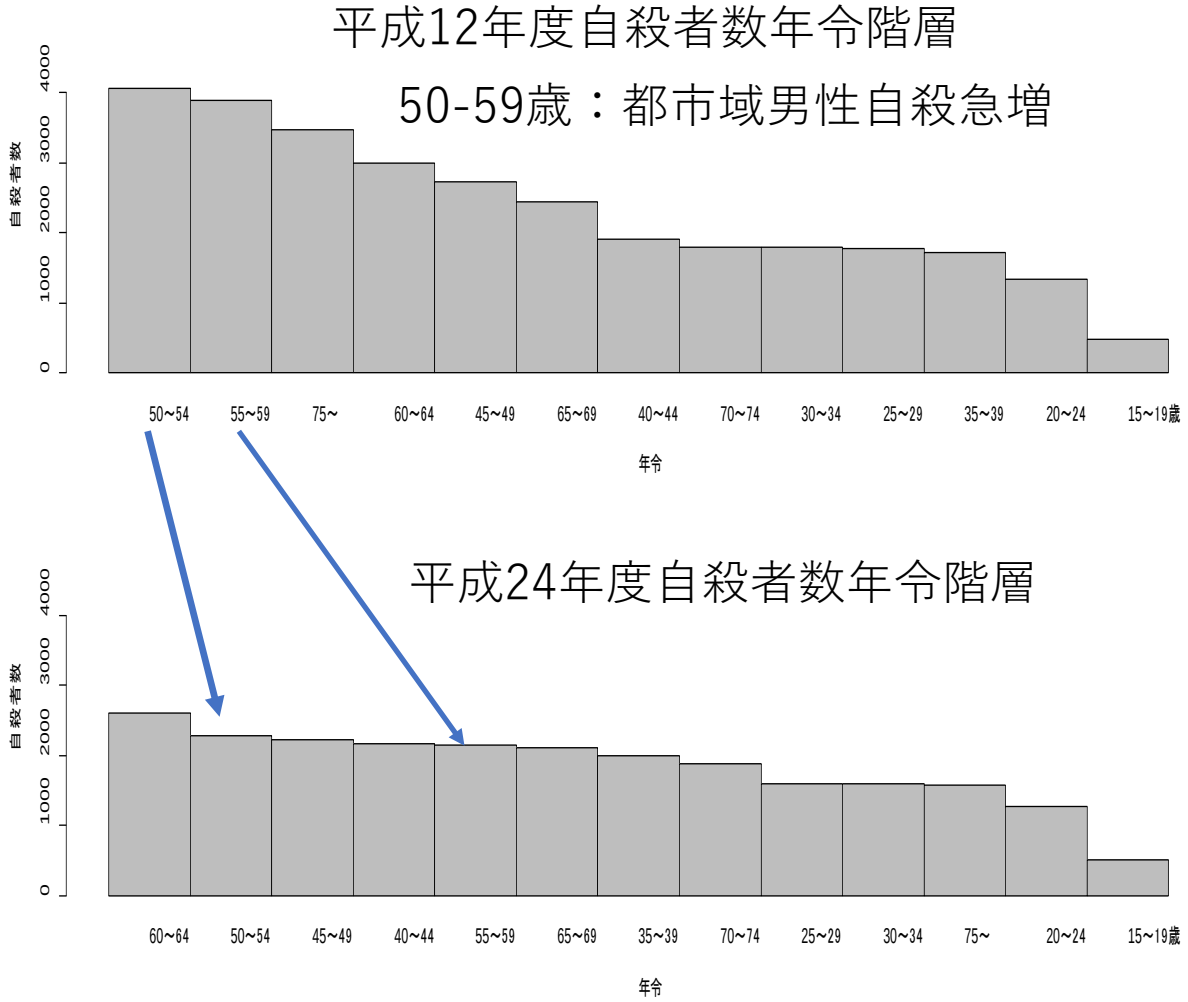
- 定性的要因分析
  - 特性要因図の活用
    - 問題に影響を与えると考えられる原因候補を議論して網羅
    - エキスパートの経験と勘の集約
      - 重要と考えられる原因候補（要因）を抽出
- **定量的要因分析**
  - 散布図（あるいは層別）の活用
    - 原因候補と結果との対応のあるデータの取得
    - 重要と考えられた要因を横軸に、問題と考えられた特性を縦軸にした散布図で関係性を検証

- 出生数/死亡数に関係する要因
  - 地域コミュニケーション密度の代用特性：可住地人口密度との散布図
    - 両対数プロット



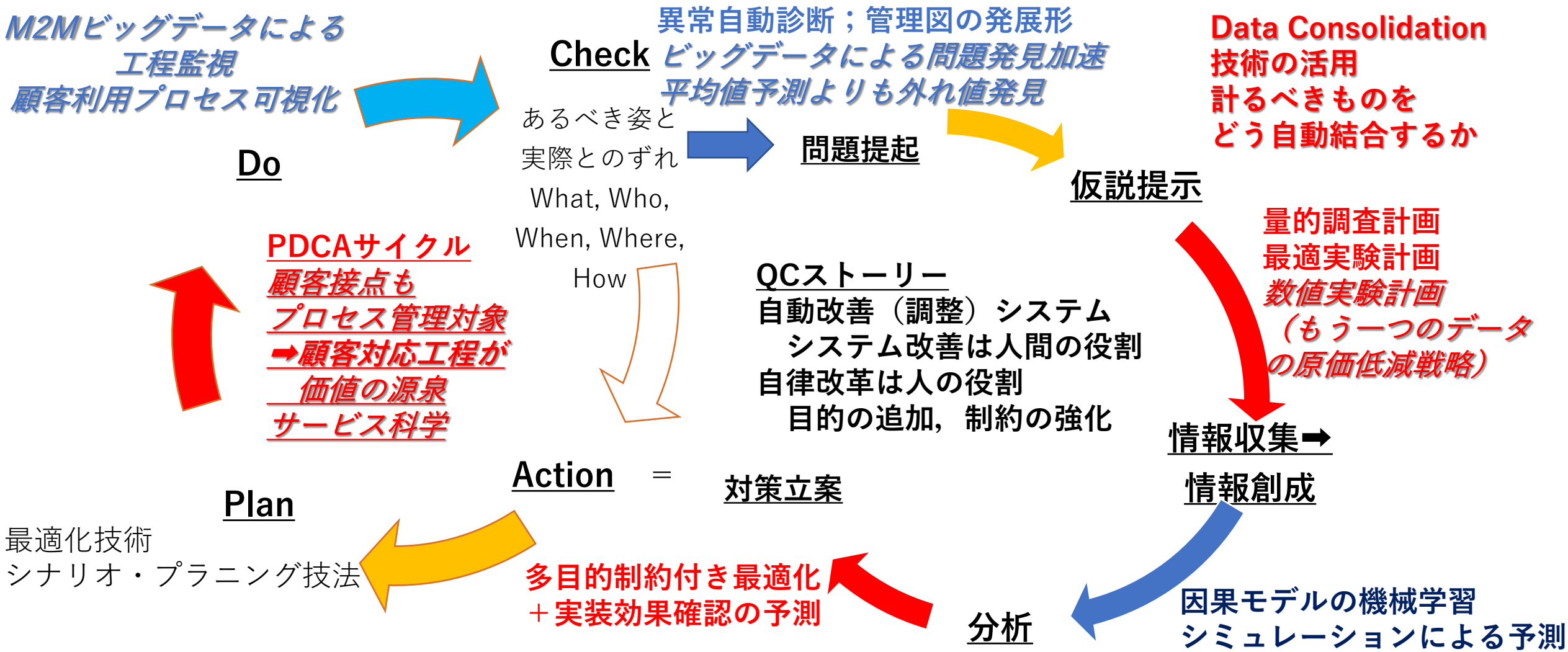
効果確認のための統計的方法

- ヒストグラム、パレート図、時系列グラフなどの、対策前、対策後の姿を比較
  - 時間的に自然増が期待される状況では、対策の効果か否かを慎重に検討
- **介入的な比較実験**で関係検証
  - 施策実施群vs非実施群
    - 新医薬品の許認可
    - 社会実験
      - ただしコストは高くつく
      - 倫理的問題
      - 公平性の問題



地域的にはまだ増加地域・増加年令区分も存在

デミング・石川モデルへのビッグデータ×AIの配置
挑戦：デミング・石川モデルのSHINKAの方向性



ミクロデータ探索的データ解析で変わると期待されること

2016年7月8日　　於：和歌山県民文化会館
ミクロデータ拠点形成試行実験報告資料より

ミクロデータ分析事例紹介
―介護に関わる生活時間の分析―

岡檀（和歌山県立医科大学講師（当時），現統計数理研究所）
椿　広計

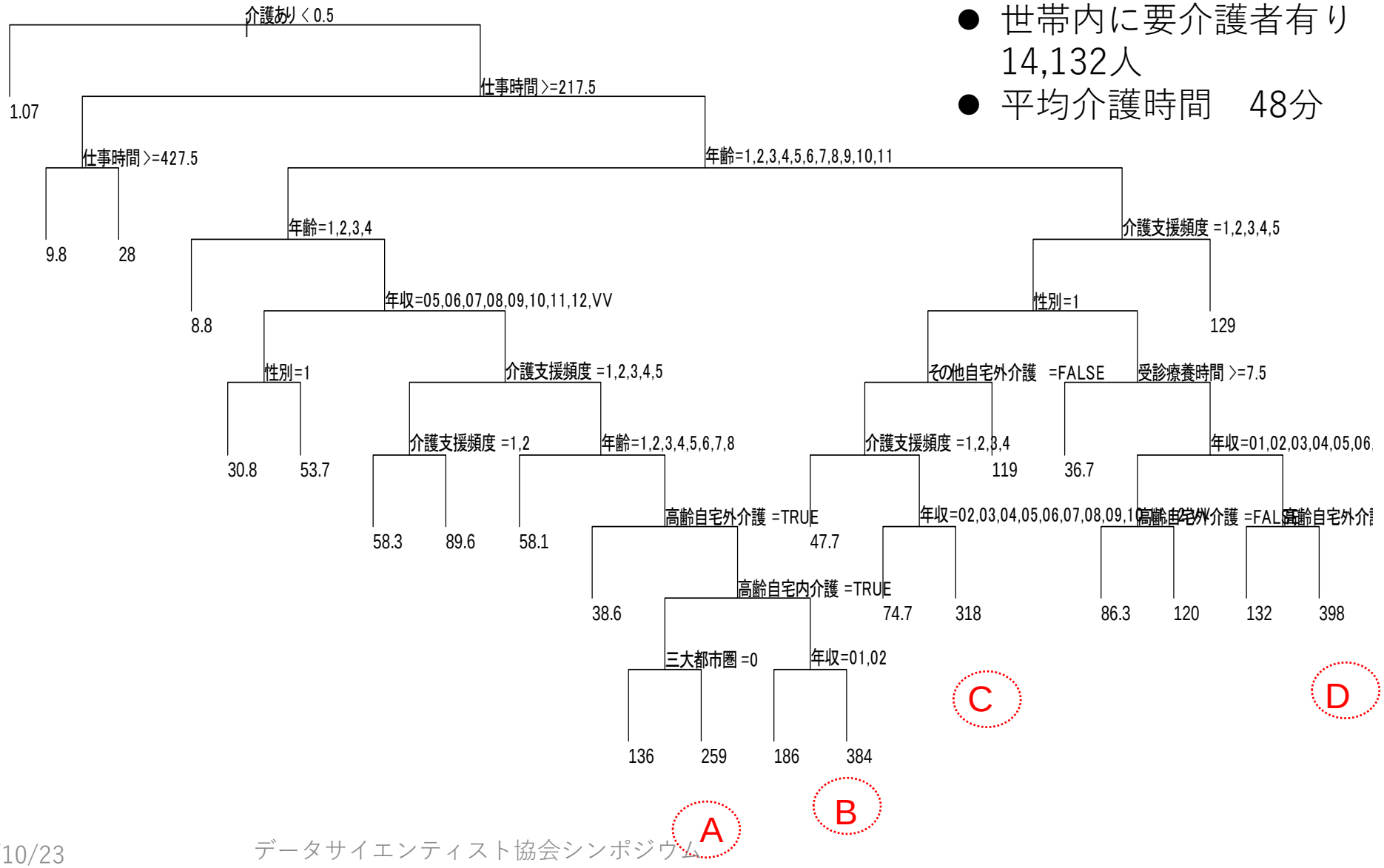
総務省社会生活基本調査

- 国民の生活行動とその時間配分について調査
- 全国の世帯から無作為に選定した、10歳以上の世帯員を対象とする
- 5年ごとに実施
- 平成18年度調査の生活時間編27万レコード、生活行動編14万レコード

社会生活基本調査データを用いた介護負担分析

- 調査結果の集計からたんに総数や平均値を参照していたのでは実態把握に限界があるが、ミクロデータを分析することにより、どのような状況下の者に重い介護負担がかかっているのかを探索できる➡問題の発見
- 介護うつ、虐待など深刻化　社会問題
- インフォーマルケア(家族等が無償で実施するケア)のコストの顕在化
- インフォーマルケア実施者の就学・就業機会の損失や心身の健康被害により、社会が負担するコストの把握

介護時間を増やす要因の探索



- 世帯内に要介護者有り 14,132人
- 平均介護時間　48分

期待される統計的行政の展開

- より手厚い支援の優先順位を検討
- インフォーマルケアコスト：社会が負うコストの過小評価是正
- 危険因子のみならず予防因子の抽出
    - 類似の状況にありながらも、よりストレスの小さい事例が参考
- 他情報の利用・他ミクロデータのリンケージ
    - 地理情報（原ミクロデータには存在）
    - 心身の健康度（厚生労働省国民生活基礎調査）

おわりに
ビッグデータ時代？

組織内に残すべきモデリングの知と外注可能な機械学習

ビッグデータ時代とは何か：価値＝便益―コスト
―部のデータと分析の価格破壊

- **（データ×機械学習）コンサルタント産業創生の時代**
  - **データ計測とモデリングの原価低減➡ビジネスとして成立**
    - データサイエンス人的資源への投資に大きな意味
  - **役割分担を明確に意識すべき時代**
    - 社内エキスパート⇒固有技術＋管理技術で一般解（非ブラックボックス知見）
    - **新産業に外注⇒管理技術（ブラックボックスも可）で迅速なSolution**
- モノづくり分野ではIndustry 4.0**➡自律分散システム産業の創生**
  - システム内での要素間のデータの自動的やりとり
    - **データの構造的結合とその自動モデリング➡確率的ロボティクス産業**
      - **機械学習、意思決定理論と固有知識（計測技術・制御技術等）の結合体**
      - それに基づく分散自律的高速意思決定を実現するシステム

データのもたらす経済的便益

- 部品Aの寸法：Aが規格内か否かを判断できる便益:*計測の基本価値*
  - 不良品を市場に出すことを回避する経済的価値D
  - 部品ロットA_i (i=1,…,n)の全ての寸法：ほぼn倍: nD
- 寸法に影響を与えると考えられる原因Bとの対データ（B,A)の便益
  - 一個一変数の追加では殆ど価値が増えない: D
- 集合データ（B_i, A_i)，i=1,…,nの便益：*データ分析による付加価値*
  - *安価なBを上手く選べば*，将来のロットに対して原因Bを管理することで結果Aを管理することが可能となる
  - 予測に基づく意思決定・制御の経済的価値
    - nD+N(将来生産)×C
- IoTその種の原因や結果が織り交ざったデータ集合体
  ネットワーク型データの結合体の分析がもたらす便益!?

データのもたらす経済的価値＝意思決定の改善価値

ビジネススクールがマネジメント・サイエンスで教えること

- 不確かな将来：Y vs 現在しなければならない意思決定：X
- 意思決定Xを行ったとき将来がYとなった場合の損失：L（X, Y）
  - 将来現象Yが，完全予測された場合：損失関数を最小化する最適決定：X(Y)*
- 期待損失関数＝リスク関数：$R(X) = E_Y[L(X, Y)]$
  - 実際に可能なのはリスクを最小化する決定：X**
  - 完全予測の経済価値 = 期待機会損失 = R(X**) – E[L(X(Y)*, Y)] :
- Yの原因系データ：　Z が計測
  - Yの最適予測
    - **初級：回帰分析，中級：機械学習による予測，上級：機械学習を解釈した統計モデル**
  - Zを観測した条件でのリスク関数の変貌： R(X|Z=z)
- 条件付リスクを最小化する決定：X***(Z) ：リスクマネジメント・サイエンス
  - 最適予測の 期待機会損失：　E[R(X***(Z)] - E[L(X(Y)*, Y)]
- Δ ＝ Zによる最適予測の機会損失ー完全予測の機会損失 ≧0
  - 原因系現象のデータを知ることの経済的価値Δ＝データの価格上限
    - Bernard W. Taylor (2009), *Introduction to Management Science*, 10th ed., Global Edition, Pearson Education (US).
    - データにどの程度まで投資可能かを明確に議論

ビッグデータ時代：古典統計家の憂鬱

- データは集めるものから集まるものとなったのか？
  - ごみも積もれば山となるのか？
    - **ビッグデータでも意図的に目的をもって集めるべきではないのか？？**
      - QRIS・コマツ・数値実験計画などは良い事例
- 古典的統計家はデータの料理人：味付けは予測・発見
  - よい材料があれば，良い料理ができる：古典的統計家受難の時代
  - **安くても沢山材料があれば，良い料理ができるのか？？**
    - 昔は：Garbage In, Garbage Outと教えていたのに
- 統計的実験計画ないしは田口流品質工学との意識のずれ
  - データは創るもの！
  - 味付けは最適化
  - 良い材料を少しだけ創れば，良い料理はできる

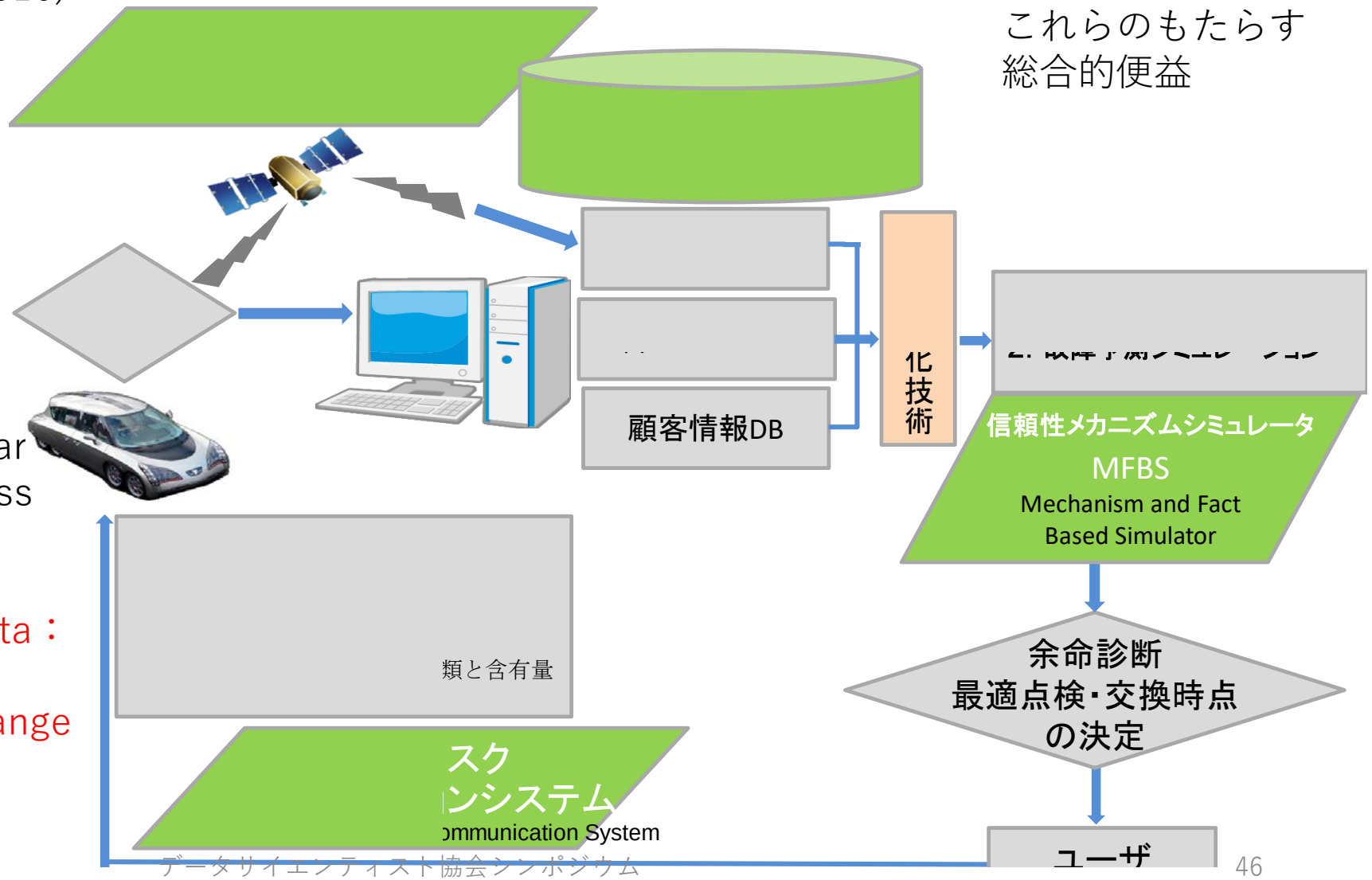
講演者（統計家）周辺のビッグデータへの期待
特にQuality Managementに関わる動向

電気通信大学
鈴木和幸日本品質管理学会長
（当時）のQRIS構想

伊藤，鈴木，椿，田中，横山他(2010)
第２期 信頼性安全性
計画研究会報告 第１報
- 次世代信頼性・安全性
情報システムによる
未然防止への取り組み -,
日本品質管理学会研究発表会

Edgar Dietrich　(2014)
Big Data – Industry 4.0　-Quality,
ICQ 2014 Pre-Conference Seminar
ISO TC 69 SC4 “Statistical Process Management 議長”
COE, QDAS
Format for Structured Quality Data：
標準化
Systems communicate and exchange data with one another
Data consolidation
Automated Analysis
Answers in Real Time

これらのもたらす
総合的便益



ビッグデータと統計的機械学習の誤解

- **AI＝Machine Learning**は画期的な予測理論：No
  - 予測に用いる統計的機械学習➡回帰分析・層別分析の自動化(自動的学習)
    - **回帰予測**とは何か？➡予測に
    - 用いる情報を与えたときの被予測情報の条件付平均値予測
      - Y=f(x1,x2,…,xp)
    - 良い**層別**とは何か？➡群内のばらつきを最小化
    - 最適意思決定とは何か➡**損失関数**を平均的に最小化
    - 機械学習が採用している幾つかの原理とは何か？➡単純だが利用者が意識すべき
- ビッグデータから**Nowcastや行動の可視化**は完璧になる：Yes
- 既存ビッグデータからの**将来予測**や行動選択は完璧になる：No
  - どんなモデルを用いても将来の不確かさを超えた予測は不可能
- **行動とその結果の評価モデル（損失関数・価値関数）が存在**する場合
  - **既存ビッグデータではなく自律的実験データを創造的に大量生成・分析**することで
    - **人工知能が自律的にPDCAサイクルを回す**ことができれば
  - **行動選択は既存の最適行動を上回るパフォーマンス**となる；**Yes**
    - **数値シミュレーション＋実験計画**で新たなデータを創ることで，革新的技術が生まれる可能性

ブラックボックスとしての機械学習をどう教えるか
機械学習の幾つかの主要原理
なぜ機械はデータで賢くなるのかを認識すべき

- 複数単純**予測値の最適結合**（個人の知より集団の知）戦術
  - **Bartlettの因子得点推定，田口のT法が基本原理➡Random Forest**
- 深層学習（ニューラルネット系統計的機械学習）の戦術
  - **交互作用の消去**と**非線形主効果**との合わせ技で任意関数近似：　$4XY = (X+Y)^2 - (X-Y)^2$
    - 特性値を上手に選択せよ：**因子モデル，PLS➡合成指標の最適探索**
    - 層別を上手く行え：**潜在クラスモデル➡**層別境界の自動探索：**CART, ニューラルネット**
    - 線形関数近似可能な非線形関数の利用➡ボルツマンマシン：温度パラメータが高ければステップ関数も近似可能
  - **予測モデルの最適単純化（パラメータ節約）：高次因子モデル➡深層学習**
- 予測誤差を改善する予測ロジックの追加戦術
  - 回帰分析改善のPDCAサイクルを回している
    - 田口の実験的回帰分析や**回帰診断➡Boosting**
- 最適パフォーマンスを持つ予測方式探索に資する計算機パワー
  - **モデル選択➡価値関数クロス・バリデーション評価**などの**自動最適化**
  - 分類・層別の**自動最適化**：Classification and Regression Tree
  - 超高次元内での分類の**自動最適化**：カーネルトリック（Support Vector Machine）
- **モデル選択➡正則化＝ベイズモデル事前分布の最適化**
  - ちょっとないがしろにされている逆問題（制御問題）の正則化

マネジメントプロセス支援技法の「機能の分類」解決プロセスに手法を配置し文法を確立

- 大藤・黒河（2014）知の巡りをよくする手法の連携活用 ー価値創生プロセスのデザイン，日本規格協会

	状態	メカニズム	シナリオ	ターゲット	業務
整理	SWOT	QFD	BSC		工程図
分析・評価	回帰分析 コビッグデータ 機械学習	パラメータ 設計 ➡数値実験活用	AHP		流れ線図
創出	潜在構造分析 コビッグデータ 機械学習		SWOT	業務機能展開	工程FMEA
選択			実験計画法 ➡数値実験活用	なぜなぜ分析	チェックシート
管理・保障	管理要管理図 コビッグデータ 機械学習				Gant Chart

手法の階層性を意識した**組織内教育**への希望

同一機能のマネジメントサイエンス技法の階層に関する意識付け

回帰分析（状態の関係を分析する方法）を一例として

- 初級者用
  - 入力制約が少ない，出力システムが理解でき，中程度のパフォーマンス
    - Classification and Regression Tree：層別は品質管理の基本
- 中級者用
  - 入力に加工を行い，出力システムは理解でき，中程度のパフォーマンス
    - セミパラメトリック回帰モデル（交互作用や非線形性を考慮）
      - 散布図に直観で線を引く
    - 通常の回帰モデル」
      - 散布図に式を当てはめる
- 上級者用
  - メカニズムに関する解釈を含むモデルでベストパフォーマンス
    - プロの統計家が技術者との協働で作る非線形パラメトリック予測モデル
- ベンチマーク用
  - 入力制約が少ない，システムは分からないが，ベストパフォーマンス
    - Random Forest（チューニング簡単）
    - 深層学習（完全ブラックボックス，チューニングは結構大変）

おわりのおわりに

- データサイエンティストのCompetence Dictionary確立を目指す
  データサイエンティスト協会の取組みに敬意
- データサイエンティストが,
  マネジメント力量を持つことは確かに本質的
- データサイエンス技法群の機能は，解決プロセス,
  改革プロセスの情報循環のどこに位置づけられるのか
  - 前工程から頂戴する情報のあるべき姿は？
  - 後工程に引き継ぐ情報のあるべき姿は？
- 産業界は勿論，今後開始されるEBPMにおいても
  行政データサイエンティストの役割は重大
- 組織的人財育成と確保にオールジャパンで取り組む必要

ご清聴
ありがとうございました