特集

社会調査に携わる人のためのビッグデータ入門

「ビッグデータ」は「AI」と並んで近年の情報技術の発展を象徴する言葉である。ビッグデータに関する解説書をひもとけば，ビッグデータ分析の威力を鮮やかに印象づける事例が綺羅星のごとく並んでいる。しかし社会調査に携わる人（研究・実務で社会調査を実施し，そのデータを利用する人）にとって，ビッグデータは必ずしも身近な存在ではないかもしれない。ビッグデータに積極的に手を出さず，いままで通りの調査・分析を行うぶんにはビッグデータと関わり合うことはまずないからだ。それでは，社会調査とビッグデータの関係はどうなっているのだろう。社会調査を利用する研究領域において，ビッグデータはどのように役に立つのだろうか。

そこで本特集では，「社会調査のことはある程度わかっているけど，ビッグデータのことはよくわからない」という読者を念頭に，ビッグデータとは何か，社会調査を利用する研究・実務領域においてビッグデータがどのように使えるのか・役に立つのかを，5名の専門家に解説していただいた。これまでの『社会と調査』の特集とはやや毛色の異なるテーマであるが，読者のお役にたてば幸いである。

（神林博史）

特集論文

1

ビッグデータと社会調査
特集「社会調査に携わる人のためのビッグデータ入門」について

神林博史
東北学院大学教養学部 教授

1 はじめに

ビッグデータは，ここ数年の間に自然科学・社会科学を問わず科学研究において幅広く注目を集めるようになった。一般向け解説書も数多く出版され，マスメディアでも「ビッグデータ」を冠したテレビ番組や情報コーナーが登場するなど[1)]，ビッグデータは学術の世界の枠を超えた流行語となった（最近は「AI」にとって代わられてしまったが）。

ビッグデータの解説書をひもとけば，情報科学，マーケティング，医療などの領域におけるビッグデータ分析の成功例の数々が紹介されており，読者はそのインパクトに圧倒される。しかし，そうした印象的な（しかし断片的な）事例を脇に置いて考えたとき，社会調査データを利用する研究領域にとってビッグデータがどのような可能性を有しているのかは，ビッグデータに詳しくない者にはなかなか見えにくい。

既存のビッグデータ解説の中には，従来のデータ収集法やデータ分析法に対して挑発的な見解を表明したものがある。たとえばMayer-Schönberger and Cukier（2013）によれば，ビッグデータはデータ分析の世界に3つの重要な変化をもたらすとされる。第一に，「すべてのデータを扱う」（「N＝全部」が実現する）ので抽出標本の必要性が低下する。第二に，データが巨大になると統計量の精度の重要性が低下するので「量さえあれば精度は重要ではない」。第三に，相関があるという事実の実用上の重要性が増すので「因果関係ではなく相関関係が重要になる」。これらの主張を額面通りに受け取るならば，従来型の社会調査データ（というより社会調査そのもの）はビッグデータにとって代わられるかのように思えてしまう。読者の中にも，そのような漠然とした不安を抱いた方がおられるかもしれない。

もちろん，これら3つの主張に反論することは難しくない（たとえば仁平・藤田 2017，および本特集の瀧川論文参照）。そして冷静に考えるならば，従来型の社会調査データとビッグデータとの関係は必ずしも競合的ではなく，むしろ相補的である（佐藤，2017）。それゆえ，社会調査データがビッグデータに駆逐される暗黒の（?）未来を心配する必要はなさそうだ。

2 本特集のねらい

ビッグデータと社会調査データが競合しないのだとしたら，社会調査あるいは社会調査データの分析に携わる者はビッグデータとどのようにつきあっていけばよいのだろうか。最も気楽なのは，従来型の社会調査とデータ分析の枠組の中にとどまり，ビッグデータはその道の専門家に任せておくことだろう。しかし，社会調査

に携わる研究者・実務家がビッグデータと無縁のまま調査・研究を続けることが好ましいとも思えない。詳しくは本特集の諸論文（特に瀧川論文）をお読みいただきたいが，ビッグデータは従来の社会調査では測定が困難だった（それゆえ分析も困難だった）人間行動や社会現象の分析を可能にしてくれる。言い換えると，ビッグデータは学術的・実務的な問題関心の検証可能性を拡張してくれるのである。ならば，それを利用しない手はないはずだ。社会調査に携わる人こそ，ビッグデータのことをよく知っておく必要がある。

以上のような問題意識から，本特集では「社会調査のことはある程度わかっているけど，ビッグデータのことはよくわからない」という読者を念頭に，(1) ビッグデータとはそもそも何なのか，(2) ビッグデータは従来の社会調査データとどのような関係にあるのか，(3) これまで社会調査データを利用してきた研究領域にとってビッグデータはどのような有用性をもつのか，(4) ビッグデータを実際に収集・分析するにはどうすればいいのか，といった素朴な疑問について大まかな答えを示すことを目指した[2)]。

3 特集論文の紹介

本特集の企画担当者（筆者）は量的社会調査およびそのデータ分析を専門とする社会学者だが，これまでビッグデータの収集・分析を行った経験はない。ビッグデータにはそれなりに興味があるので，解説書や論文を散発的に読んできたという程度の門外漢である。本特集が想定する「社会調査のことはある程度わかっているけど，ビッグデータのことはよくわからない読者」というのは，実は企画担当者自身のことに他ならない。『社会と調査』の読者には企画担当者と似たような立ち位置の方が多いだろうと勝手に想像して（そうでない読者の皆様にはお詫びいたします），本特集を構想した次第である。このように，企画担当者はビッグデータの素人であるが，特集論文の執筆者陣はビッグデータ研究の最前線で活躍する優秀な方ばかりである。企画担当者の無理なお願いに応えて，素晴らしい論文を寄稿していただいた。以下，各論文を簡単に紹介しよう。

論文2「ビッグデータとは何か」（笹原和俊）は，ビッグデータの概説である。ビッグデータの定義，研究史，ビッグデータの用途などがコンパクトにまとめられており，ビッグデータに詳しくない読者には特に有用だろう。既存の解説ではあまり言及されることのなかった社会調査との関連についても一節が設けられており，ビッグデータが社会調査データを補完・補強する役割を果たしうることが説明されている。

論文3「社会学におけるビッグデータ分析の可能性」（瀧川裕貴）は，ビッグデータ分析を援用した社会学的研究のレビューである。本論文の前半では，(1) ビッグデータは社会学的研究にどのような意義を持つか，(2) 社会学者はビッグデータの持つ特性のどこに注目すべきか，の二点が議論される。後半では，ビッグデータを用いた社会学的研究の具体例が紹介される。詳しくは論文をお読みいただきたいが，ビッグデータ分析が社会学理論の検証可能性を大きく広げてくれることを実感できるだろう。なお，本論文に興味を持った読者は，同じ著者による計算社会科学（これもビッグデータ分析の一種）のレビュー論文（瀧川，2018）を併せて読むことをお勧めしたい。

論文4「データジャーナリズムの歩みと可能性」（奥山晶二郎）では，メディアにおけるビッグデータ分析の用途の1つとしてのデータジャーナリズムが解説される。データジャーナリズムとは，伝統的なジャーナリズムの情報源に加えて，ビッグデータを含む様々なデジタル情報を用いるジャーナリズムのことである。テレビニュースなどでもSNS上の書き込み等を情報源とする報道はすでによく見かけるようになったが，それ以外にも様々な情報がデジタルジャーナリ

ズムでは利用される。マスメディアの現場においても,ビッグデータが伝統的な情報源を駆逐・代替するのではなく,補完・補強する形で利用されていることがわかるだろう。

論文5「ビッグデータによって変わる未来の公的統計」(水野貴之) では,公的統計におけるビッグデータ利用の可能性が解説される。公的統計とビッグデータの関係は大まかに,(1) 既存の公的統計の作成にビッグデータを利用する,(2) ビッグデータを利用して新たな公的統計を作る,という2つの面がある。本論文ではこの両面について,国内外の最新の動向が紹介されている。公的統計の領域においても,ビッグデータがその可能性を拡大してくれることが示される。

論文6「ビッグデータ分析の実践にむけて」(榊剛史) は,ビッグデータ分析を実際にどう行うのかを解説する。ビッグデータ分析のための環境の構築,データの取得・収集,データの加工・前処理,分析,可視化,という一連のプロセスが解説されると共に,サンプルデータを用いて読者が実際に分析を行えるよう配慮されている。社会調査データの分析がある程度できる読者であれば,(環境構築さえクリアすれば) 大きな問題なくビッグデータ分析を追体験できるはずである。ビッグデータに興味はあれども触れる機会のなかった読者は,ぜひ挑戦していただきたい。

以上の論文を読んでいただければ,社会調査データを利用する研究領域においても,ビッグデータがその視野と可能性を大きく広げてくれることを実感できるだろう。多くの読者がビッグデータに関心を持ち,自身の研究・業務に生かしていただければ幸いである。

【謝辞】

瀧川裕貴先生 (東北大学) には,特集論文の執筆のみならず,本特集の準備段階から助言と協力をいただきました。計算社会科学研究会 (https://css-japan.com/) 主査の遠藤薫先生 (学習院大学) には,同研究会メンバーを中心に特集執筆陣を組みたいという企画担当者の提案をご快諾いただき,そのお蔭で本特集を形にすることができました。日野愛郎先生 (早稲田大学) には,特集論文執筆者の奥山晶二郎先生をご紹介いただくと共に執筆依頼を仲介していただきました。以上の皆様に深く感謝いたします。

注

1) たとえば,NHK スペシャル『震災ビッグデータ』(2013年3月3日,2013年9月8日,2014年3月2日の3回シリーズ)。この番組の詳細については阿部 (2014) を参照。

2) ビッグデータと社会調査の関係については,ここに挙げた4点以外にも重要な課題がある。たとえば「社会調査でビッグデータを作る」「ビッグデータ収集の方法論を用いて社会調査を行う」「ビッグデータの収集・分析における倫理的問題」などである。企画担当者の能力不足および紙幅の都合から,残念ながら本特集ではこれらの問題は扱わない。これらの課題については,改めて特集が組まれることを期待したい。

文献

阿部博史, 2014,「『震災ビッグデータ』から見えてきた東日本大震災の姿」『放送メディア研究』11:271-289。

Mayer-Schönberger, V. and Cukier, K., 2013, *Big Data—A Revolution That Will Transform How We Live, Work, and Think*, Houghton Mifflin Harcourt. (斎藤栄一郎訳, 2013,『ビッグデータの正体:情報の産業革命が世界のすべてを変える』講談社。)

仁平典宏・藤田真文, 2017,「特集『テキストマイニングをめぐる方法論とメタ方法論』によせて」『社会学評論』68 (3):326-333。

佐藤俊樹, 2017,「データを計量する　社会を推論する ―『新たな』手法が見せる社会科学と社会」『社会学評論』68 (3):404-423。

瀧川裕貴, 2018,「社会学との関係から見た計算社会科学の現状と課題」『理論と方法』33 (1):132-148。

特集論文

2

ビッグデータとは何か

笹原和俊
名古屋大学大学院情報学研究科講師，科学技術振興機構さきがけ研究者(兼任)

1 はじめに

ノーベル賞物理学者フィリップ・ウォレン・アンダーソンの言葉に，“More is different”というものがある。「量が増えると質的な変化が起きる」という意味である。これは，大規模データが価値に転化するというビッグデータの考え方にも通じる言葉である。

ビッグデータとは，人間と様々な情報機器との相互作用から日々生み出される膨大な情報の集合体である。モバイル，ソーシャル，クラウドという要素をもつ情報技術の発展によってデータが集積されるようになり，それを分析することからもたらされる知見や価値は，私たちの生活やビジネス，社会のあり方をも変え始めている。近年では，IoT（モノのインターネット）やAI（人工知能）が普及し，その傾向にますます拍車がかかっている。

ビッグデータという言葉が使われ始めたのは2010年頃である。2010年のThe Economistの記事ではビッグデータという言葉は使われていないものの，現在のビッグデータに通ずる考え方が「データ洪水（the data deluge）」という言葉で紹介されている（Cukier, 2010）。また，2011年にマッキンゼー・グローバル・インスティテュートが発行した報告書では，ビックデータは「典型的なデータベースソフトウェアの格納，管理，分析能力を超えるサイズのデータセット」と紹介されている（McKinsey Global Institute, 2011）。日本では，総務省が取りまとめている「情報通信白書」平成24年版にビッグデータという言葉が掲載され，そこでは「事業に役立つ知見を導出するためのデータ」と紹介されている（総務省,2012）。

黎明期にはバズワードと揶揄されたビッグデータだが，登場から10年目の節目が見えつつある現在はすっかり社会的に定着し，イノベーションを起こすために必要不可欠な資源となっている。本稿では，ビッグデータとは何かについて様々な角度から再考する。

2 ビッグデータとはどのようなデータか

ビッグデータには明確な定義はないが，ビッグデータと呼ばれるデータ群は三つのVを頭文字にもつ英語の特徴で説明されることが多い。米国の調査会社ガートナーが報告書の中でそのように説明したのが最初で，ビッグデータの「3Vモデル」とも呼ばれる（Beyer and Laney, 2012）。

一つ目のVは「Volume（大容量）」である。つまり，これまでのツールでは取り扱えないほどデータの容量が大きいということである。現時点ではテラバイト以上の容量のデータを指すことが多いが（テラは10の12乗），データを蓄積・処理する技術の発展とともに，ビッグデータと呼ぶにふさわしいデータの容量は今後さらに増大すると考えられる。その理由として，スマー

トフォンやクラウドサービスの普及,センサを搭載したIoT機器の急増によって,生成されるデータが爆発的に増加していることが挙げられる。

二つ目のVは「Variety(多様性)」である。つまり,データの種類やそれを扱うための形式が多様化したということである。例えば,商品の取引データや顧客データ,SNSのテキストや画像や動画,スマートフォンなどに組み込まれたGPS(Global Positioning System:全地球測位システム)の位置情報,IoT機器のセンサ情報など,生成されるデータの種類が増大した。GPSなどのように,あらかじめ決められたフォーマットで保存することができるデータを「構造化データ」,SNSのデータのように決まったフォーマットのないデータを「非構造化データ」という。従来型のデータベースはリレーショナルデータベースと呼ばれるもので,あらかじめフォーマットの決まったテーブルにデータを格納し,それらの集合を関連づけてデータを管理する方式のため,非構造化データを取り扱うことが難しかった。しかし,大規模な非構造化データを蓄積・処理する技術が発達し(後述),現在は多様なデータを分析することが可能になった。

三つ目のVは「Velocity(速度)」である。つまり,データの発生頻度や更新頻度が増大したということである。オンラインショッピングにおけるユーザのクリックのデータ,コンビニエンスストアで24時間発生するPOS(Point of Sales)データ,全国の道路に設置された渋滞検知センサ,交通系ICカードの乗車履歴,ツイッターの投稿やフェイスブックの「いいね!」など,現在,データは24時間365日ほぼ途切れることなく生成されている。このように時々刻々と発生するデータを処理することは一昔前までは不可能だったが,分散処理やストリーミング処理の技術の発達によって,リアルタイムにデータを処理することが可能になった。

近年ではビッグデータの3Vモデルに,さらに二つのVを特徴として付け加えることがある。IBMはビッグデータの四つ目の特徴として,「Veracity(正確性)」を取り入れ,データの矛盾や曖昧さによる不確実性を排除して,信頼できるデータにもとづく意思決定が重要だと主張している。さらに,ビッグデータの五つ目の特徴として「Value(価値)」が入ることもある。これは,実社会における人間行動やIoTなどを利用した社会活動が生み出すデータを活用して価値を生み出すことの重要性を強調している。

3V(あるいは5V)の特徴をもつビッグデータを価値に転換できるようになった理由に,データの大容量性,多様性,速度の問題を解決する情報技術が登場したことがあげられる。例えば,ビッグデータ処理のためのソフトウェアフレームワークにNoSQLやHadoopがある。あらかじめ決められたデータ構造しか扱えないリレーショナルデータベースとは異なり,NoSQLはデータの一貫性を緩くし,データ構造を単純化することで,データ管理の柔軟性や拡張性を可能にした。Hadoopは,2004年にGoogleが発表したMapReduceという分散データ処理の方法(Dean and Ghemawat, 2008)を実装したオープンソースのソフトウェアフレームワークで,大量の非構造化ファイルを高速に処理すること可能にした。

ここまでは,主に狭い意味でのビッグデータの特徴の説明である。広い意味のビッグデータには,機械学習や統計解析などのデータ分析やそれを行う人材や組織なども含まれる(図1)。

3　ビッグデータ小史

コンピュータやインターネットなどの情報技術の進歩とともに,データの生産量はかつてないほど飛躍的に増大している。特に,ビッグデータの歴史を振り返る上でインターネットの登場以後のモバイル,ソーシャル,クラウドの発展は重要な位置を占める。

ティム・バーナーズ=リーによって最初のウェブサイトが作られたのは1991年で,この頃はま

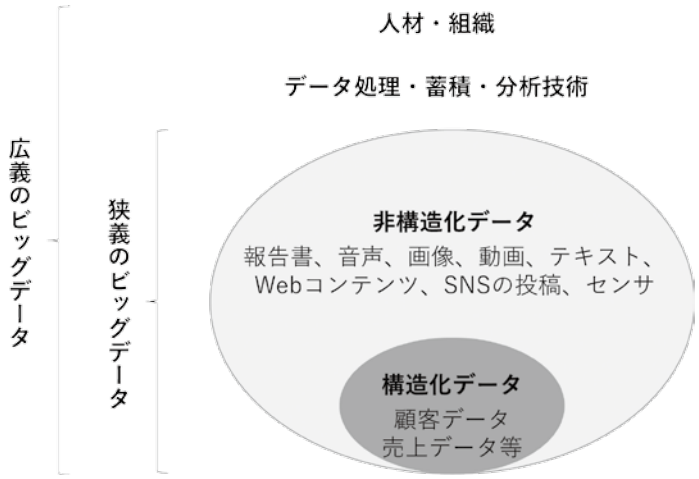


図1　広義と狭義のビッグデータ
（総務省「情報流通・蓄積量の計測手法の検討に係る調査研究」（平成25年）を参考に改変）

だ一部の限られたユーザだけが情報発信できる時代だった。2000年代になるとパソコン，次いでスマートフォンが普及し，ブログやSNS，動画共有サイトやチャットアプリなどが一般に広く使われるようになり，いわゆる「ウェブ2.0」ムーブメントが起こった。ウェブ2.0とはティム・オライリーが提唱した概念で，旧来のウェブにおける情報の流れが送り手から受け手へという一方的なものだったのに対し，誰もが自由に情報を発信・受信できるようになった新しいウェブの利用状態のことを指す。ウェブのソーシャル化である。

モバイルの代名詞ともいえるスマートフォンは，2010年には約10%だった保有率が2016年には70%以上に達し，ほとんどの人がスマートフォンをもつ時代になった（総務省，2017）。スマートフォンにはGPSセンサ，磁気センサ，加速度センサ，ジャイロセンサなど様々なセンサが搭載されている。これはつまり，大半の人がセンサ付きの小型コンピュータを常時持ち歩いているということに等しい。自宅や職場にあるパソコンだけならばインターネットの利用は限られるが，スマートフォンの登場で屋外や移動中の利用が増え，GPSをはじめとする様々なセンサから時々刻々とデータが発生するようになった。

ソーシャルメディアの普及もデータの増大と多様化に大きく関係している。代表的なSNSであるツイッターでは1日に5億以上の投稿（Twitter, 2014），フェイスブックでは1日に47億件以上の投稿が共有されている（Facebook, 2013）。SNSの特徴はリアルタイム性が高く，社会的つながり（フォロー関係）の中を情報が伝搬し，それが連鎖することである。SNSから生み出されるデータはテキストだけでなく，画像や動画を含む非構造化データである。先述のような新しい情報技術の登場で，大規模で多様性に富むSNSのソーシャルデータを分析して，マーケティングや新商品・新サービスの開発に活用することが可能になった。

さらに近年はIoTの発達によって，人間だけでなくあらゆるモノが情報を発信するようになり，データがクラウドに集約されるようになった。米国のEMCコーポレーションと調査会社IDCが2014年に実施した調査によると，全世界的に生み出されるデータの量は，2013年には4.4ゼタバイトだったが2020年には44ゼタバイトに達すると予想されている（ゼタは10の21乗）（EMC, 2014）。

ここまで述べたように，インターネット誕生

後に登場したモバイル,ソーシャル,クラウドという情報技術は人間を巻き込む形で発展し,私たちの生活や仕事のスタイルを大きく変化させた。私たちのオンライン上での行動履歴,意見や感情,写真や動画,IoT機器が常時発信するセンサ情報など世界で生み出されるデータは爆発的に増加し,それらがクラウドに蓄積されるようになった。その結果として誕生する,人やモノが発するこの世界に関する情報の断片の集積がビッグデータである。今後も様々な情報機器がインターネットを通じてつながり,私たちの行動や社会に関するデータは累積していく。そのような現実世界の大規模な電子的痕跡をAIでリアルタイムに処理・分析し,その結果を人や社会にフィードバックしたり,みんなで共有したりすることでビッグデータの価値は高まっていく。

内閣府はこのようなデータ主導型社会の未来として「Society 5.0」というコンセプトを提唱している。Society 5.0とは,「サイバー空間（仮想空間）とフィジカル空間（現実空間）を高度に融合させたシステムにより,経済発展と社会的課題の解決を両立する,人間中心の社会」だと定義されている。現在の社会（Society 4.0）では知識や情報が共有されず,分野横断的な連携が不十分であるという反省に基づき,ビッグデータと情報技術でこの問題を克服しようという発想から生まれた考え方である。Society 5.0は,IoTで全ての人とモノがつながり,様々な知識や情報が十分に共有されることで,今までにない新たな価値を生み出し,それによって少子高齢化,地方の過疎化,貧富の格差など,現在社会が抱えている様々な問題を克服することを目標に掲げている。楽観論だという批判的な声もあるが,データを循環させることに重きを置いた「ウェルビーイング・エコシステム」を志向している点は新しい。

4　ビッグデータはどのように使われているか

ビッグデータには,個人レベルから地球規模まで様々なスケールのデータがある。基本的には図2に示したプロセスで,ビッグデータを知見や予測や価値に変換する。以下では,ビッグデータのスケールの違いに着目し,活用例を紹介する。

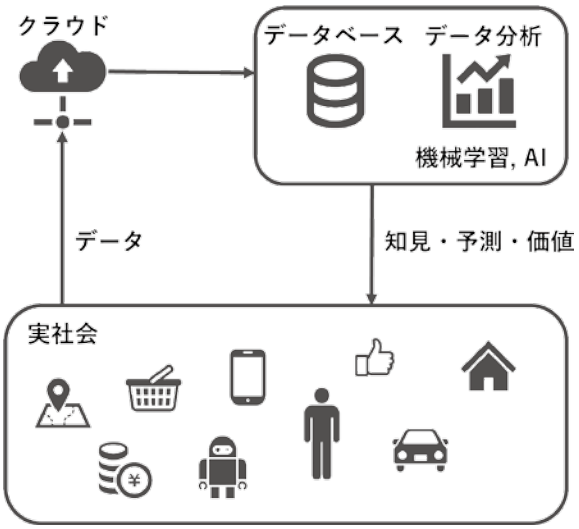


図2　ビッグデータから価値を生み出すプロセス

個人スケールのビッグデータといえば，身体に装着可能なウェアラブル・センサや日常生活に欠かせないスマートフォンから時々刻々と生み出されるデータが代表的である。それらのデータは私たちの感情や居場所，習慣や購買などの日常生活の電子的記録である。例えば，フィットビットやアップルウォッチは，移動歩数や消費カロリーなどの身体活動を簡便かつ継続的にモニターすることで，健康目標の達成や運動習慣の改善をサポートする。スマートフォンにもGPSセンサや加速度センサをはじめ様々なセンサが搭載されており，それを利用したアプリケーションはこれまでには不可能だったような「ライフ・ログ」を簡便かつ継続的に取ることを可能にした。そのような個人が生み出すビッグデータからは，利用者の気づきを促し，行動を改善させるようなサービスが生み出されている。

ウェアラブル・センサとそれが生み出すビッグデータは，集団レベルでの生産性の向上にも役立てられる。例えば，マサチューセッツ工科大学メディアラボのアレックス・ペントランドのグループが開発した「ソシオメータ」と呼ばれる電子バッチは，顔の表情や体の動き，声などに表れる無意識的な行動，すなわち，偽ることの困難な「正直シグナル」を記録する装置である（Pentland, 2008）。同グループは実社会の様々な場面で「正直シグナル」を計測して，それが将来の行動の予測に有効であることを明らかにした。そして，ソシオメータのビッグデータが優れた集団的意思決定や賢い組織・社会作りにどのように役立てられるかに関する洞察を与えた。日立が開発した「ビジネス顕微鏡」は，個々人の身体運動のパターンから集団の幸福感と相関する組織活性度を計測する装置で，そこから生み出されるデータを組織の生産性向上に活用することを目的としている（矢野，2014）。

インターネットの検索記録も有用性の高いビッグデータである。例えば，グーグルの研究グループは検索行動のビッグデータがインフルエンザの流行の予測に有効であることを示し，大きな話題となった（Ginsberg et al, 2009）。同グループは，米国人がインターネット検索する際に入力した検索語のうち上位5,000万語を抽出し，疾病予防管理センターの2003年から2008年までの5年間のインフルエンザの流行に関するデータとの相関を調査した。その結果，特定の45個の検索語がインフルエンザの流行と高い相関を示すことがわかった。この知見を利用して，同研究チームはグーグル・フルー・トレンズ（Google Flu Trends）というシステムを作り，疾病予防管理センターでは1週間かかるものを，このシステムだと1日の遅れでインフルエンザの流行を予測できることを示した（現在は公開終了）。

ツイッターやフェイスブックなどのSNSのデータは，私たちがどのように考え，行動し，どのようにつながりあっているのかに関する地球規模の貴重なデータを提供する。コーネル大学のゴールドとメイシーは，世界84カ国の2,400万人が2年間に発した計5億900万件のツイートに含まれる感情語の時間変化を分析し，人間の気分の変動には周期性があることを明らかにした（Golder and Macy, 2011）。ツイッターの集団気分は株価予測などへも応用されている。インディアナ大学のヨハン・ボレンらは感情に関するツイートを分析し，集団気分の平穏さの度合いが3日後のダウ・ジョーンズ工業株価平均の動きを87.6%の精度で予測できたと報告している（Bollen, 2011）。

ケンブリッジ大学のマイケル・コジンスキー（現スタンフォード大学）らは，約58,000人のボランティアに個人属性に関するアンケートを答えてもらい，許可を得た上でフェイスブック上の記事のどれに「いいね!」したのかに関するデータを収集し，個人の趣味や属性をどの程度推定できるかを調査した。その結果，「いいね!」のデータだけから，そのユーザが男か女かのような基本的な属性だけでなく，同性愛者かどうか，薬物を使用するかどうかなどの個人情報も予測でき

ることがわかった（Kosinski et al, 2013）。このようなビッグデータから個人属性を推定する技術はターゲティング広告などに応用されている。一方で，この技術はフェイクニュースや政治的プロパガンダにも悪用されていることは注意しておきたい。

個人から地球スケールまで，今やビッグデータは現実をモニターするだけでなく，予測し変化させる力をもつ。ビッグデータは諸刃の剣でもある。

5　ビッグデータと社会調査

ビッグデータの利活用に関する考え方はコンピュータの普及以前からあり，国勢調査や経済統計なども広義にはビッグデータの範疇に入る。国が国内の実情を把握するために全国民を対象に実施する国勢調査は，紀元前3000年のエジプトや中国でも行われていたことが知られており，ローマ帝国時代に行われた国勢調査は今でも記録が残っている。1890年に米国で行われた国勢調査ではパンチカード式のシステムが初めて導入され，これは3Vの意味における最初のビッグデータだと言える。

社会調査を行う場合，まず調査目的を決め，国勢調査などの全数調査でなければ，その目的を検討するのに必要な限られた人数のサンプルを対象とし，観察や質問票などの方法で調査するのが通常の流れである。コンピュータやインターネットが発達した現在，このような社会調査をオンラインで行い，サンプルの規模を拡大することも可能になったが，既存の社会調査を大規模化することとビッグデータによるアプローチとは質的に異なる。

ビッグデータのアプローチでは，質問票などの方法ではとれなかったようなデータが取れるようになる。スマートフォンの大量の通話記録からは人間関係が浮かび上がり，SNSへ投稿されるデータからは人々の心理状況や購買意識などが明らかになる。これらは人々の自発的な行動の記録であり，質問票で「0がリベラル，10が保守的として，あなたの政治的な立場を答えて下さい」などのように回答を制約してしまうやり方とは異なる。高密度な時系列ビッグデータを用いると，人間集団の社会的ダイナミクスを観察・定量化することも可能になる。また，産業，人口，観光，農業などの官民のビッグデータを集約し組み合わせることで，「地域経済分析システム（RESAS）」のように，産業構造や人口動態，人の流れなどのマクロな情報を可視化し，地域経済に関する鳥瞰図を得ることが可能になる。このような情報はこれまでの社会調査では取得が困難だったものである。

しかし，ビッグデータがこれまでの標本抽出や調査票による作業に置き換えられるというわけではない。むしろ，ビッグデータは既存の社会調査を補完したり，補強したりするものである。2015年に開催された計算社会科学の国際会議IC^{2}S^{2}でコーネル大学のマイケル・メーシーは，社会科学におけるビッグデータの意義について，「これまでは不可能だったことが計測できるようになったことだ」と述べている。これまでは計測不可能だったこととは，端的に言えば人間行動や社会現象の要素である。

ビッグデータを活用した新しい経済指標を作り出そうという動きもある。2017年，経済産業省は，ツイッターの投稿などのビッグデータとAIを活用した新たな経済指標サイト「ビッグデータスタッツ」を試験的に公開した（現在は閉鎖）。これまでの経済統計では，データの収集から公表までに時間がかかり，現在のおおよその経済動向を迅速に把握することが困難だった。しかし，リアルタイム性の高いSNSのデータをAIでうまく分析すれば，この問題が解決できる可能性がある。

6 結論

本稿では,ビッグデータという言葉が世に出てから10年の節目が見えてきた現在において,「ビッグデータとは何か」について改めて様々な観点から見てきた。ビッグデータが暮らしや産業,社会科学そのものを根本的に変える可能性を秘めていることは疑いようのない事実だが,一方で,ビッグデータの問題点やリスクも徐々に明らかになってきている。最後に,ビッグデータの問題点を整理し,今後の展望を述べる。

ビッグデータの黎明期には,「N＝全部」や「答えがわかれば,理由はいらない」などのキャッチーな表現がビッグデータの特性としてよく使われた。これはつまり,「正確で厳密だが量が少ないデータより,乱雑でも膨大なデータの方が実用的メリットが大きく,役立つ相関関係（因果関係ではなく）を見つけることが重要である」という極端な態度のことを言っている。技術雑誌WIREDの編集長を務めたクリス・アンダーソンも「理論の終焉」と題した記事の中で,「膨大な量のデータと応用数学があらゆるツールに取って代わる世界が始まる。言語学も社会学も,人間行動に関する理論は全て駆逐される」と,挑戦的な発言をしている（Anderson, 2008）。しかし,これは明らかに言い過ぎである。

大規模で多様性に満ちたデータを取得・処理できるようになったことは大きな進歩だが,ビッグデータには「バイアス」などの問題があり,目先の相関だけを追っていると大きな間違いを起こす可能性がある。例えば,ビッグデータの成功例として取り上げられることの多いグーグル・フルー・トレンズ（先述）だが,これには後日談がある。フルー・トレンズが発表された後,英語以外の言語でもサービスが開始されたり,デング熱の流行にも同じやり方が応用されたりと展開は順調に見えた。しかし,2014年にフルー・トレンズがインフルエンザの流行を2倍近く多く見積もっていたことを示す論文が発表されると（Lazer et al., 2014）,ビッグデータにおいて安易に相関のみに頼ることの危険性を指摘する報告が次々となされ,検索語のみに基づく方法の限界が明らかになった。同年,グーグルはフルー・トレンズおよび関連サービスを終了している。検索データを疾病予測に用いるというアイデアは斬新であり貴重な試みだが,ビッグデータに基づくデータ主導型推論の難しさを示すエピソードである。

ビッグデータの寡占も大きな問題である。「21世紀の石油」と形容されることもある個人データに関しては,GAFAと呼ばれるITの巨大企業（グーグル,アップル,フェイスブック,アマゾン）が国境を超えて貪欲にデータを収集し,その規模は他社の追随を許さない状況になっている。4社は豊富な個人データを分析することで,自社の商品やサービスを効果的に販売し,ターゲティング広告などで巨額の収益を得ている。個人データは集まれば集まるほど価値が増し,他社との差は開く一方である。2018年7月に開催されたG20財務相・中央銀行総裁会議では,GAFAらIT企業への課税が議論されるなど,ビッグデータの寡占の問題は今後の課題の一つである。

ビッグデータ利活用の促進を考える上で,オープンデータの整備やプライバシー保護も重要な課題である。ビッグデータは組み合わせることでさらに価値が高まり,新しいビジネスやサービスにつながるため,行政や企業が抱えるデータを誰もが自由に使えるようにする取り組みは重要である。現在,日本には「DATA GO JP（http://www.data.go.jp/）」や「LinkData.org（http://linkdata.org/）」などのオープンデータのポータルサイトがあるが,海外と比べるとオープンデータの整備は遅れている。一方,ビッグデータにはプライバシーの問題もある。データを匿名処理したとしても,複数のデータを組み合わせると高い精度で個人を特定できてしまうため,オープンデータの整備を進めるためにも,より安全

なプライバシー保護技術の開発が必要である。

今後，これらの問題点やリスクを克服しつつ，ビッグデータを使いこなす新しい方法の開発やビッグデータの新しい用途の発見がますます重要になってくる。それは，イノベーションの促進にとっても社会科学の発展にとっても大きな鍵となる。

文献

Cukier, K., 2010, "The data deluge: Businesses, governments and society are only starting to tap its vast potential", *The Economist* (https://www.economist.com/leaders/2010/02/25/the-data-deluge/)

McKinsey Global Institute, 2011, *Big data: The next frontier for innovation, competition, and productivity* (https://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation/)

総務省，2012，『情報通信白書平成 24年版』(http://www.soumu.go.jp/johotsusintokei/whitepaper/h24.html)

Beyer, M. A. and Laney, D., 2012, "The Importance of 'Big Data': A Definition", *Gartner*.

Dean, J. and Ghemawat, S., 2008, "MapReduce: Simplified data processing on large clusters", *Communications of the ACM*, 51 (1): 107-113.

総務省，2017，『情報通信白書平成 29年版』(http://www.soumu.go.jp/johotsusintokei/whitepaper/h29.html)

Twitter, 2014, https://blog.twitter.com/official/en_us/a/2014/the-2014-yearontwitter.html

Facebook, 2013, https://www.facebook.com/FacebookSingapore/posts/563468333703369

Pentland, A., 2008, *Honest Signals: How They Shape Our World*, MIT Press.

矢野和男，2014，『データの見えざる手：ウエアラブルセンサが明かす人間・組織・社会の法則』，草思社。

EMC, 2014, *The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things* (https://www.emc.com/collateral/analyst-reports/idc-digital-universe-2014.pdf)

Ginsberg, J. et al, 2009, "Detecting influenza epidemics using search engine query data", *Nature*, 457: 1012-1014.

Golder, S. A. and Macy, M. W., 2011, "Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures", *Science*, 333: 1878-1881.

Bollen, J., Mao, H. and Zeng, X., 2011, "Twitter mood predicts the stock market", *Journal of Computational Science*, 2: 1-8.

Kosinski, M., Stillwell, D. and Graepel, T., 2013, "Private traits and attributes are predictable from digital records of human behavior", *Proceedings of the National Academy of Sciences*, 110: 5802–5805.

Anderson, C., 2008, *The End of Theory: The Data Deluge Makes the Scientific Method Obsolete* (https://www.wired.com/2008/06/pb-theory/)

Lazer, D., Kennedy, R., King, G. and Vespignani, A., 2014, "The Parable of Google Flu: Traps in Big Data Analysis", *Science*, 343: 1203-1205.

特集論文

3

社会学における ビッグデータ分析の可能性

瀧川裕貴
東北大学災害科学国際研究所 助教

1 はじめに

近年,社会学を始め,政治学,経済学,心理学その他の社会科学全般において,いわゆるビッグデータを用いた研究が急速に普及している。その背景としてまず挙げられるのがインターネットの普及に伴う社会生活のデジタル化である。電子メールはもちろんのこと,ソーシャルネットワーキングサイトを通じたコミュニケーションや議論,様々な商取引,オンライン上の百科事典や予測市場での集合知の活用等,われわれの社会生活の多くはいまやウェブ上で行われている。このようなウェブ上での活動はリアルタイムでデジタルに記録され,「デジタルフットプリント」ないし「デジタルトレース」と呼ばれる膨大なビッグデータを残すことになる(Golder and Macy 2014)。また,ウェブ上の社会行動のみならず,スマートフォンやIoT(モノのインターネット)の普及によって,物理的な行動もますますデジタルに記録されるようになりつつある。例えば,スマートフォンの位置情報サービスを利用した際には移動履歴がデジタル記録として残される可能性がある。

さらにビッグデータを用いた研究を考えるときには,政府行政記録のビッグデータも無視すべきではない。例えば,政府が保管しているデジタル化された納税記録を用いることで,所得不平等や世代間移動について従来とは異なる規模と詳細さで研究を行うことができる。政府行政記録自体は古くから社会科学者の主要なデータソースの一つであったが,デジタル化や大規模化に伴い,従来型の分析戦略や研究戦略は大幅な見直しを迫られつつある。

また,デジタル化された文学作品や議事録,裁判記録その他の歴史資料も重要なビッグデータの一類型である。日本の事例で言えば,著作権の消滅した本などを電子化して誰もがアクセスできるようにした青空文庫や第1回の国会から衆参両院の本会議および委員会の議事を記録している国会会議録が著名である。

以上を背景として,ビッグデータを活用した社会科学的研究には多くの注目が集まっている。他方で,ビッグデータの社会研究への応用をめぐっては過度の期待と過度の疑いが寄せられがちだということに注意する必要がある。過度の期待といえば,一時期もてはやされたビッグデータ万能論がある。例えば,評論家のC.Andersonはかつて,ビッグデータの活用により,理論はもはや不要になると述べた(Anderson, 2008)。現在ではさすがにこのような楽観論を述べる研究者はいないが,ビッグデータのご利益を過度に述べ立てる評論やTV番組はいまでも世の中にあふれている。他方,このような世の中の風潮に反発してか,とくに伝統的な社会科学者の中にはビッグデータに「慎重な」態度をとる者も

多い（例えばGoldthorpe, 2015）。彼らの懸念には理解できる面もあるが，やはりビッグデータが社会科学の発展に貢献する潜在的な可能性を無視して，慎重論に流れすぎるのも問題である。

本稿では，ビッグデータの限界にも十分な注意をはらいつつ，社会科学的な研究を進める上でビッグデータのもつ潜在的な可能性について強調したい。

2 ビッグデータと社会学的研究

伝統的な社会学者のなかにはビッグデータに懐疑的なスタンスを取る人もいる。その理由の一部は，特に日本でのこれまでのビッグデータの紹介のされ方が，「社会学研究でいかに活用できるか」，という観点を欠いたものだったからだといえよう。例えば，比較的早い段階で翻訳されたマイヤー＝ショーンベルガーとクキエの『ビッグデータの正体』(Mayer-Schönberger and Cukier 2013=2013)では，ビッグデータがもたらす3つの変化として次が挙げられている。

1. ビッグデータは限りなくすべてのデータを扱う。
2. 量さえあれば精度は重要ではない。
3. 因果関係ではなく相関関係が重要になる。

著者らはおそらく確信犯的に挑発的な主張をしている。だが，まともな社会学者ならばこれらの主張に反論することは容易だ。例えば，「量さえあれば精度は重要ではない」というのは，ランダムではない系統的バイアスが存在するデータでは全く当てはまらないし，社会現象の因果メカニズムを解明するという社会学の立場からいえば，因果関係の解明を放棄する態度は受け入れられないだろう。

しかし，いうまでもなく，ビッグデータに関するこのような見解は一部の論者のものに過ぎない。社会学におけるビッグデータの意義についてより注意深くかつ建設的に議論した論考もさまざまに存在するし(Bail 2014; Evans and Aceves 2016; Golder and Macy 2014; McFarland et al. 2016)，何よりAmerican Sociological ReviewやAmerican Journal of Sociology等の一流誌にはビッグデータを有意義に活用した社会学的研究がしばしば掲載され，しかもその数も年々増えている（そのいくつかについては後に紹介する）。

このような状況を鑑みると，いま必要なのは，ビッグデータに対する過度の楽観論でも過度の反発でもなく，「ビッグデータのどの特徴を，いかにして社会学的研究に活用することができるのか」に関する注意深い検討であろう。本稿では，このような立場からまずはビッグデータの特徴を一般的にまとめたうえで，とくにビッグデータの活用という点で意義深い社会学的研究のいくつかを紹介し，社会学におけるビッグデータ分析の可能性を検討することにする。

3 ビッグデータの特徴と課題

「はじめに」で述べたように，ビッグデータといってもそのソースに応じて，様々に異なるビッグデータが存在する。大まかには，次の三つに区別できる。

1. ソーシャルメディアやIoTのセンサなどの記録したデジタルトレースデータ
2. 政府や公的機関の記録をデジタル化した政府行政記録
3. 文学，学術論文，議会等の議事録，裁判記録などの大規模テクストデータ

もちろん，この分類は便宜的なもので，相互に排他的ではないし，網羅的でもない。例えば，分析技術が進めば，大規模な音声・映像データも重要なデータソースになりうる。

社会学的研究にとってのビッグデータという観点から考えると，ビッグデータのいかなる特徴に特に注目するべきだろうか。本稿では，社会学的研究の目標が，社会関係や相互行為の因果メカニズムの解明にあるとの前提にたつ。ビッグデータがメカニズムについての理論を不要とするという一部の論者の議論とは逆に，因果メ

カニズムの解明にこそビッグデータを利用すべきだと主張する。このような立場から，いかにしてビッグデータの社会学的活用可能性を特徴づけることができるだろうか。

ここでは，以下の4つの特徴がとりわけ注目されるべきであると論じたい（以下の議論は瀧川，2018も参照。ビッグデータの特徴についてのより包括的議論についてはSalganik，2017を参照）。

1. 社会的・関係的データ

ビッグデータの重要な特徴の1つは，人々の関係性や相互行為を記録している点にある。M.Weber（1921, 訳, 1987）の社会学の定義を引くまでもなく，社会学の基本的関心は，ある人の行為がいかなる因果メカニズムを通じて，他の人の行為を引き起こすかという点にある（ある制度が行為を引き起こすこと，またはある行為がある制度を生み出すことなども社会学の関心対象に含まれる）。例えば，ある人の「傘をさす」行為が別の人の「傘をさす」行為を引き起こすとき，これを社会的行為と呼び，社会学の研究対象となる。

このような社会的行為のメカニズムの解明こそが社会学の主たる目的であったにもかかわらず，サーベイデータのような従来の量的データでは，直接に社会的行為を扱うことは難しかった。質問紙には，人々の相互行為は記録されないし，交流や付き合いの様子や頻度を聞くにしても，それらは回顧的に尋ねるほかなく，かつその回答は主観的なものとなる。そのため，社会学の理論の検証は，理論的に想定された因果メカニズムのありそうな含意を，いくつもの推論を重ねた上で，サーベイデータで測定可能なデータに翻訳して検証する，という間接的な方法をとらざるを得なかった。

これに対して，ビッグデータには社会的行為が直接に記録されている。例えば，ソーシャルメディア上での誰かの発言がその後，他の人々にどのような反応を引き起こしたのかを調べることができる。もちろん，ある人の行為が他の人の後続する行為を因果的に引き起こしたと厳密に述べるためには，いくつかの問題をクリアする必要があり，それほど容易ではない。だが，サーベイデータに比べれば，因果メカニズムを直接的な方法で検討できる素材ははるかに揃っている。

2. 時系列データ

社会学にとって有用な第２の特徴として，ビッグデータは多くの場合，時系列的に記録されているという点が挙げられる。このことは因果メカニズムの解明にとって非常に役立つ。因果メカニズムとは，原因から結果に至るダイナミックなプロセスのことだからである。例えば，社会学的研究において，ある制度が導入されたり，ある出来事が生じたことで，人々の振る舞いがどのように変化したのかを検討するとき，ここで問われているのは，制度や出来事が人々の振る舞いを因果的に形作るというダイナミックなプロセスである。

関連する特徴をさらに挙げるならば，ビッグデータは「常時オンalways-on」（Salganik, 2017）に記録されることが多い。常時オンであるとは，人々の振る舞いがそのつど，意図して記録する必要もなく，常時データとして収集されているということだ。近年では，サーベイ調査でもパネル調査として複数時点にわたるデータ収集が行われることがある。だが，常時オンのデータの場合には，そうしたパネルデータでも難しい，予期せぬ出来事のメカニズムを解明することが可能となる。この常時オンという特徴を用いて，例えば，ある種の制度変更や自然災害のような出来事が発生したことによって，人々の振る舞いがどのように変化するのかを調べることができる。

3. 異質性の探索

3番目の有用な特徴として，ビッグデータは大

規模であるがゆえに，異質な因果効果を調べることができる，という点を挙げておこう。ここでいう異質性とは，ある因果関係が，時間や場所，関係する人々の特徴などによって，異なる効果をもつことをいう。例えば，親の地位達成の程度が子の地位達成に因果的に影響を与えていることが分かったとする。しかし，このような因果的効果は，公的教育の質の低い地域でのみみられるということも同時に分かったとすれば，親の地位が子の地位達成に与える因果メカニズムについてさらに深い知見を得ることができる。

Weberの比較宗教社会学に典型的なように，普遍的に成り立つと考えられていた社会「法則」に対して時間や場所の限定をつけるというのが社会学の伝統的なスタンスであった。この営みは因果効果の異質性という仕方で再定式化することができる。重要なのは，時間や場所による限定は普遍的な因果メカニズムの追求を否定するものではなく，むしろより「深化」させるものだということである（cf. Morgan and Winship, 2014）。

このように，異質性は，因果メカニズムについての理解をより深めるためには必須のものといえる。にもかかわらず，従来のサンプルサーベイでは，サイズが比較的限られているために，異質性に十分に注意を払うことができなかった。

4. 全数性・網羅性

最後にビッグデータの全数性・網羅性を社会学的研究にとって有用な特徴として挙げておきたい。ただし，この点については取扱いに注意が必要である。ビッグデータ，特にデジタルトレースから得られたビッグデータは，通常，サンプリングを経たものではないし，会議録などの大規模テクストデータは，やじなどを除外するという意味での選択性を別にすれば，国会などでの議論を網羅的にカバーしている。しかし，全数性という特徴は結局のところ，母集団として何を想定するかに相対的である。

例えば，Twitterにおける原発を巡る議論に関わるツイートを母集団とするならば，それらを網羅的に収集し，原発に関するツイートを母集団とした「全数調査」を遂行することは不可能ではない。しかし，より広く，「公共空間で行われた原発に関する議論」を母集団として想定すれば，これらのツイートは，Twitterというソーシャルメディアで限られた人だけが参加した非常に限定的なサンプルとなる。

以上のことに留意した上でなおやはり，ある母集団について全数調査やそれに近いサンプルサイズを確保できることは，社会現象のメカニズムを明らかにする上で大きな強みになるといっておきたい。というのは，適切に想定された母集団を考える限り，データの選択性という問題を相対的に回避することができるからである。例えば，どのような資源動員戦略を採用した社会運動が成功しやすいかを調査したいとしよう。従来の調査法では，時間と場所を限定したとしてもなお，発生した社会運動すべてを調べることは難しい。そこで，成功して存続した社会運動を中心に調査することになりやすい。その場合，こうした社会運動の多数がある特定の資源動員戦略を採用していたからといって，その戦略が運動の成功を因果的に導いた，ということはできない。それは，例えば，成功とは何も関係なく，単にその時点で広く普及していて，成功した運動も失敗した運動もともに採用していた戦略であっただけかもしれない。これに対して，網羅的なビッグデータを用いれば，成功した運動も失敗した運動もともに分析でき，データの選択性に由来するバイアスを回避できる。このような単純な例だけでなく，データの選択性は様々な仕方で，因果メカニズムを推定する際の障害となりうる。

4 ビッグデータを用いた社会学的研究の事例

以上の一般的な議論をふまえた上で，本節で

は,ビッグデータを適切に利用することで,社会学上の重要な問題に取り組み,新たな知見の提出に成功した,と考えられる模範的な研究をいくつか紹介したい。

密度の高い社会構造の規範維持メカニズム

ソーシャル・キャピタルの理論では,密度の高い社会構造が社会的・集合的目標を促進するメカニズムの解明が主たる課題の一つとなっている。例えば,J.Coleman (1988,訳,2006) によると,アメリカのカトリックスクールの生徒の退学率が低いのは,親同士が互いに知り合いであり,そのことによって生徒に規範やルールを効果的に守らせることができるからである。親同士が互いに知り合いであるというような密度の高い社会構造はソーシャル・キャピタル論では,結束型ソーシャル・キャピタルと呼ばれ,多数の研究の主たる関心事となっている。

ソーシャル・キャピタルの理論は,ある特定の社会的関係がある帰結をもたらすダイナミックな因果的プロセスを理論化するという意味で典型的に社会学的な理論でありながら,従来のデータでは,社会関係とダイナミックな因果的プロセスを捉えることが難しいため,直接的に検証することが非常に困難であった。例えば,コールマンの提案したメカニズムは次のような要素からなる。

1. ネットワークの密度が高いと規範の侵害が減少する。
2. 規範の侵害が防止されるのは,第三者が規範侵害を処罰するからである。
3. 規範侵害の処罰が可能なのは,さらに別の他者が処罰者に報酬を与えるからである。

しかし従来のデータで検証できるのはせいぜい,ネットワークの密度が高いと規範の侵害度合いが低いという関係を示すことくらいで,この関係が2と3のメカニズムによって成立しているというのは理論的推測にとどまっていた。

このような状況において,ウィキペディア編集データを用いて,2と3のメカニズムそれ自体を直接検証しようとしたのがM.J.PiskorskiとA. Gorbatâiの研究 (Piskorski and Gorbatâi, 2017) である。ウィキペディア編集データは,従来のデータにはない特徴を備えている。第1にウィキペディアの編集者たちの行為と相互行為が記録されている。その中には,「規範の侵害」,「処罰」,「報酬の付与」として解釈可能な行為も含まれる。第2に,編集者たちの行為がすべて時系列的に記録されている。第3に,記事を一緒に編集したという関係を社会関係とみなし,かつ関係への埋め込まれの程度を密度とすることで,密度の多寡のある社会構造のデータを得ることができる。こうして,彼らは,密度の高い社会構造における規範維持の因果メカニズムの直接的検証にまで踏み込むことができた。

彼らが規範の侵害としたのは,他人の編集した記事を勝手にもとに戻す”undo”という行為であり,これに対する処罰がundoを取り消す行為である。さらに第3者によるundoを取り消す行為は,もともとundoをされた側にとっては自分で取り消すよりも正当性を得やすく,それゆえ彼らにとっての報酬となる。この操作化にもとづいて彼らが分析の結果,見出したのは次のメカニズムであった。

1. 社会構造の密度が高いと規範侵害の程度が低くなること。
2. 社会構造の密度が高いと規範侵害に対する処罰が生じやすいということ。
3. 社会構造の密度が高いと処罰をした他者への報酬が生じやすいこと。

このように,PiskorskiとGorbatâiは関係的・時系列的データであるというビッグデータの特性を活かすことで,ソーシャル・キャピタル,とくに結束的ソーシャル・キャピタルが規範の維持に寄与するメカニズムに関する理論を直接検証することに成功した,といえる。もちろん,この研究に問題がないわけではない。例えば,処罰や報酬付与行動の概念がウィキペディア上の

行為によって適切に測定できているかどうかは問題となるだろう。これはSalganikが「データの不完全性」と呼んだ問題で,サーベイの質問項目とは異なり,「生」の社会行動を相手にするビッグデータ分析に特有の問題である(Salganik, 2017)。これに関して万能の解決策は存在しないが,同一の理論概念を異なる操作化・測定による複数のデータを用いて多角的に検証していくことが,筋の通った方策であろう。

制度・出来事から行為へのフィードバックメカニズム

社会学の研究では,ある法や制度が人々の振る舞いにどのように影響するかを検討することが非常に多い。プロテスタンティズムが人々の禁欲行動に与える影響を述べたWeberの古典的な研究はいうまでもなく,例えば,ある種の性差別的な集合的表象が人々のマイクロな差別行動を促すとするフェミニズムの議論も,そのような因果メカニズムを暗黙に仮定していると考えることができる。

しかしながら,従来の研究では,制度から行為への因果関係を,定量的に示すことは難しく,もっぱら理論的推測にとどまるか,あるいは少数のケースを取り上げた質的研究を行うほかなかった。量的な研究についていえば,従来のクロスセクショナルなサーベイデータは,制度が作られた後に行為や態度が変容するという時間的プロセスを分析することが難しいため,制度と行為の関係の定量的な分析には大きな限界があった。

しかし,ビッグデータ,特に常時オンの性質をもつソーシャルメディアのデータを用いることによって,ある制度が人々の態度や行動に引き起こした変化を定量的に把握することが可能になった。ここで紹介したいのは,反移民法的な性格をもつアリゾナ州のSB1070という法が人々の移民に対する態度や行動にいかにして影響を与えたかを,Twitterデータを用いて検証したR. D. Flores(2017)の研究である。

彼が具体的に検討したのは,法のもつ表出的・シンボル的効果である。法は,例えば,人がしてよいこと・悪いことを定める社会規範に影響を与えることで,人の行動を変える効果をもつと一般に議論されている。だが,この効果についての経験的証拠は多くない。そこで反移民的性格をもつアリゾナ州のSB1070が制定された前と後で,Twitterに現れる人々の態度と行動の変化を分析しようというのがFloresの研究戦略となる。この戦略は,人々の行動が常時記録されているソーシャルメディアの常時オンの性質を利用することで可能になっているということに再度注意を促しておきたい。

具体的な分析戦略は大きく分けて次の2つの部分からなる。第1は,Twitterから人々の態度や行動を測定するためのセンチメント分析という方法論である。具体的には,語にポジティブな感情とネガティブな感情を割り当てた辞書を用いて,ツイートの表すセンチメントを特定するのである。例えば,”I hate immigrants.”のような文は移民に対するネガティブな感情を示すと想定される。第2に,同じ時期のネバダ州の移民に対するツイートを比較対象として設定する。そうすることで,アリゾナ州でのSB1070の制定(これをX日とする)前後での移民関連ツイートの差分differenceとネバタ州でのX日の前後での移民関連ツイートの差分の間の差分を検討することができる。これは差分の差分法Difference in differencesと呼ばれ,因果推論を行うための標準的な方法論の一つである。

分析の結果,反移民法的性格を持つSB1070はメキシコ・ヒスパニック系の移民に対するセンチメントに負の因果的効果をもったと結論されている。ただし,これは態度変化というよりも,新ユーザーの参加とツイート数の増加によるものであった。

ある出来事が人々の差別意識・行動にどのような因果的効果をもつのかを調べた別の研究として,Legewie (2016)も挙げておこう。彼の研

究は，行政記録のビッグデータを用いた研究である。具体的には，警官が通行人を呼び止めて所持品検査をするstop and friskという慣行の大規模時系列データを用いて，黒人によるとみられる警官銃撃事件の発生が警官の黒人に対する差別行動を因果的に引き起こしていることを示した。この研究は，差別意識・行動の形成や促進において，特定の出来事が果たす因果的役割について新たな証拠を提示するものといえる。

社会移動の詳細なメカニズム

世代間社会移動の分析は伝統的に社会学の中心的な研究対象であり，膨大な知見が蓄積されている。しかしながら，近代化が進むにつれて世代間の社会移動が「開放的」となるとした近代化理論が衰退した後，世代間の社会移動を阻害したり促進したりするメカニズムについての確立した理論はいまだない。パス解析を用いた地位達成モデルは，地位の世代間伝達メカニズムに関する一定の知見を蓄積しているが，問題もある。第1に，従来の地位達成モデルは，因果推論の観点からはその妥当性を強く疑われている。つまり，因果メカニズムを明らかにするというよりは，変数間の関連を記述するものにとどまっている，と批判されている。第2に，地位達成モデルは，個人を取り巻く社会環境や構造よりも，例えば本人学歴などのような個人的要因に注目して，地位達成を説明する傾向がある。これは，近代化のような構造的要因によって社会移動の変化を説明しようとした伝統的な社会移動論の関心からはずれている。

では，社会移動論に対してビッグデータに基づく研究は新たな寄与をすることができるだろうか。ここでは，経済学の立場から大規模政府行政記録（納税記録）を用いて社会移動の研究を行ったChettyらの研究（Chetty et al.，2014）を取り上げたい（以下，Salganik, 2017における解説も参考にしている）。

Chettyらの研究においてもっとも重要なのは，ビッグデータを用いることで社会移動の地域的異質性を明らかにした点にある。地域レベルの異質性の発見は，地域という構造レベルにおいて社会移動を促進したり，阻害したりする因果メカニズムの探求につながる。その意味で従来データでは明らかにならなかった構造レベルの因果メカニズムを解明する可能性がある。

彼らの研究では，社会移動の程度は，親と子それぞれの全体の所得分布に占めるランクの間の相関として操作化される。この測定を用いると，線形モデルで親と子の所得ランクの相関がよく近似でき，しかもロバストだからである。この指標を基礎にして，社会移動の程度を，CZ(Community Zone)と呼ばれる地域ごとに測定すると，地域により大きな相違があることがわかった。例えば，ユタ州ソルトレイクシティやカリフォルニア州サンノゼなどは高い上昇移動率をもつ一方で，「ラストベルト」に位置するインディアナ州インディアナポリスは低い上昇移動率で特徴づけられる。

このような地域の間の異質性はそれらの地域の特徴によってどの程度，「説明」できるだろうか。厳密な因果分析ではないが，Chettyらは OLS回帰などの方法により，当該地域の居住隔離の程度やソーシャル・キャピタルの量，学校教育の質や家族の安定性（シングルペアレントが少ないこと）などが社会移動を促進したり阻害したりする社会構造的要因ではないかと示唆している。ここで重要なのは，これらが個人レベルの要因ではなく，社会構造レベルの要因だということだ。例えば，シングルペアレントの子どもが，上昇移動をしにくいというだけでなく，シングルペアレントの多い，家族の安定性を欠く地域では，他の子どもも上昇移動の機会を阻害されるということが示唆されているのである。事実，続く研究（Chetty and Hendren, 2018）では，Chettyらは子どもの地域移動の機会を利用した準実験デザインを用いて，地域の因果効果の存在を実証している。

相互行為の量的分析

言語的,非言語的相互行為やコミュニケーションの分析は，社会学の根本に位置するものであり，E.Goffmanによる分析やエスノメソドロジー・会話分析による緻密な研究の蓄積が存在する。しかしながら一部の例外を除いて，これらの研究では通常，定量的な分析が志向されていない。他方，社会学の流れとはほぼ独立に,ビッグデータと計算的手法を用いて言語的・非言語的相互行為を定量的に分析する流れも近年，急速に増えてきている（例えば，Pentland, 2012など）。

では，社会学的な立場による相互行為分析にとって,ビッグデータはどのように利用できるだろうか。この点について考えるために，社会学者D. McFarlandらによる言語的・非言語的コミュニケーションの双方を検討した研究(McFarland et al. 2013)を紹介しよう。彼らの研究目的は，社会的絆の形成に関する社会学理論，とくにE.Durkheim, GoffmanそしてR.Collinsの相互行為儀礼理論を，新たなデータを用いて検証することにある。この理論のポイントは，社会的絆の形成について，当事者たちの属性等の他に，相互行為における経験そのものが重要であるということ,そこでの経験とは,様々な感情の高まりや相互行為のシンクロなどであるということである。

彼らの調査対象は，順番に多数の異性と出会い，交際相手を探すスピードデーティングというイベントである。彼らは，私立大学の大学院生向けのスピードデーティングイベントを複数開催し,これをオーディオで記録した。このオーディオは書き起こされ，また音声情報それ自体も解析されている。さらに，彼らは相手に対してクリック（気が合う）を感じたか,デートする意志はあるか,等を別途,質問紙でたずねている。

理論上の被説明変数である社会的絆の形成は，このデータでは,クリック度やデート意志の有無として操作化されている。これに対して，絆の形成を説明する相互行為的特徴は，言語的・非言語的コミュニケーションの音声・テクストデータから，機械学習・自然言語処理等を用いて抽出されている。例えば，相互行為理論では，行為のシンクロは対象を共有する場合に生じやすいとされているが，これは”I” や”You” といった代名詞の使用として操作化される。また，状況への関わりもシンクロにとって重要だが，これは　”I mean” や”You know” など関わりを示す話法によって操作化可能である。

クリック度（やデート意志）を従属変数に，個人属性と様々なテクスト的・音声的特徴を説明変数とした分析の結果として次が報告されている。まず,クリック度については個人属性の説明力が最も高いが,コミュニケーション的，相互行為的特徴もその半分程度の説明力をもつ。これは，相互行為儀礼理論の一定程度の妥当性を確認するものといえよう。さらに，とりわけ女性において，対象の共有や状況への関わりがクリックと強く関連することが明らかになった。男女を比べると，一般に女性の方が，個人属性よりも相互行為的特徴により社会的絆を形成する傾向が強いようだ。

相互行為のビッグデータによる分析はどのような新しい知見をもたらすだろうか。McFarlandらの研究を踏まえるならば，誤差のともなう相互行為間の関係の析出には，大規模データに基づく統計的解析が威力を発揮するだろうということだ。例えば，恋愛関係，また，より一般には社会的絆の形成には多数の要因が介在するため，相互行為上のある特徴が絆形成に寄与したかどうかは,ケーススタディでは見えづらい。このような場合にも，行為間の隠された因果関係をビッグデータ分析が解明する可能性がある（とはいえ，McFarlandらの研究は厳密には因果分析ではない）。

資源動員にフレイムがもたらす因果効果

最後に，社会運動論におけるビッグデータの応用事例を検討しよう。社会運動論の中心的問

いは,社会運動はいかにして動員を可能にし,どのように集合的目標を達成しているか,という点にある。ここでも,様々な資源の動員のあり方と社会運動の成功との間の因果関係を説明するという意味で因果メカニズムの探求が行われている。

社会運動における動員の方法については,資源動員論というアプローチのもとで,様々な資源に着目した分析が行われているが,なかでも運動に対する文化的な意味づけも決定的な要素の1つと考えられる。つまり,運動がどのように文化的に意味づけられるのかによってその動員力は大きく変化するのである。これは伝統的にGoffmanのフレイム分析を社会運動に援用する形で行われてきた(Benford and Snow, 2000)。しかし従来の研究の大半は,成功した運動についてのケーススタディであり,そこには特定の文化的意味づけのゆえに運動が成功した可能性だけでなく,単に成功した運動がある特定の意味づけを採用していただけであるという「従属変数による選択性バイアス」の問題が常につきまとう。理想的には,成功した運動だけでなく,失敗した運動もすべて調査対象に含める必要があるが,従来の方法ではこれは困難である。

Bail (2016)はフェイスブックアプリを用いた大量データ取得という方法によってこの問題を解決しようとした野心的な研究である。Bailが調査対象としたのは,フェイスブックにファンページをもつ臓器移植アドボカシー組織である。これらを母集団と設定すれば,フェイスブックを通じた調査という手続きをふむことで,成功,失敗の有無にかかわらず,ほぼすべての組織を目標サンプル(その数は79組織)とすることができる(最終的な回収率は59.5%)。アプリによるデータ収集は次のようになされる。まず組織が調査に協力し,アプリの利用を許可すると,ページの視聴回数やシェア回数などファンページ管理者のみが有する様々な情報にアクセスできるようになる。加えて,統制変数として用いるための組織の特徴等の簡単なサーベイ調査もアプリによって行うことができる。

彼らの理論的問いは,フレイミングが社会運動の動員にいかに効果をもつか,という点にある。具体的には,フレイミングと動員との関係について,文化収容力という概念に基づく独自の理論を提唱している。この理論では,フレイミングの包摂するトピックの多様性が鍵となる。具体的には,多様なトピックによるフレイミングは一定程度までは多くの異なる関心の人びとの注意を引く点で有効であるが,多様性がある点を超えると,かえって(認知的な複雑性や一貫しない印象を与えるなどの理由で)有効性が低下すると予想される。トピックの測定は,構造トピックモデルというトピックモデルの一種によって行われる。このモデルで測定されたトピックの多様性が説明変数であり,フェイスブックでのファンページへの「いいね!」の回数やシェアなどのユーザーの関心が従属変数となる。

分析の結果,組織の資金力や組織戦略などを統制しても,トピックの多様性とユーザーの関心は逆U字カーブの関係にあることが報告されている。つまり,文化収容力の理論の予測するとおり,トピックの多様性は一定程度まで効果的であり,その後効果を減ずるということが確認されたのである。ここでのポイントは,フレイミングの中程度の多様性が動員に対して効果的であるという因果メカニズムの発見は,成功した社会運動と失敗した社会運動の双方をできるだけ網羅的に調べることによって,得られたものだということである。単に成功した社会運動のフレイミングの仕方を調べて,そこに中程度の多様性があったという結果だけでは,このような主張をすることはできない。「失敗した運動も同程度に多様性を保っていたにもかかわらず失敗していた」という可能性を排除できないからである。

5 結論

以上，本稿では，社会学にとって有用なビッグデータの特徴を，1.社会性・関係性，2.時系列，3.異質性，4.全数性，の4つにまとめた。そのうえで，ソーシャル・キャピタル，制度による行為形成，社会移動，相互行為分析，社会運動論など，社会学にとって重要なトピックに対していかにしてビッグデータ分析が貢献できるかを，実際の研究事例に即して検討してきた。

これらから明らかになったのは，社会現象の因果メカニズムの解明という点に関して，ビッグデータが新たなブレイクスルーをもたらしうるということであった。もちろんこのことは従来のデータ，サーベイデータや諸々の質的研究のデータが不要になるということを全く意味しない。例えば，最後に紹介した研究にあるように，ビッグデータと別のデータ（この場合，サーベイデータ）を組み合わせることは，単独のデータだけを用いるよりも，メカニズムの解明にとってはるかに有効たりうる。当然ながら，それぞれのデータにはそれぞれの特徴があり，研究目的に応じて向き・不向きがある。ビッグデータの無批判な礼賛は問題であるが，逆に，ビッグデータの短所にのみ目を向け，この新たなデータソースをことさら無視し，ブレイクスルーの可能性を閉ざしてしまうこともまた愚かしい。ビッグデータの特徴を十分に理解し，社会現象の因果メカニズムの解明に向けた効果的な活用の戦略を十分に組み立てることで，私たちは社会学的研究を大幅に前進させることができるのだ。

謝辞

本研究はJSPS科研費 JP16K04027,JP16H03698の助成を受けたものです。本研究の草稿に有益なコメントをいただいた常松淳先生（日本大学），大林真也先生（青山学院大学）に感謝いたします。

文献

Anderson, C., 2008, “The End of Theory: The Data Deluge Makes the Scientific Method Obsolete”, *Wired*, 23 June 2008.

Bail, C. A., 2014, “The Cultural Environment: Measuring Culture with Big Data”, *Theory and Society*, 43(3-4): 465-482.

————, 2016, “Cultural Carrying Capacity: Organ Donation Advocacy, Discursive Framing, and Social Media Engagement”, *Social Science and Medicine*, 165: 280-288.

Benford, R. D. and Snow, D. A., 2000, “Framing Processes and Social Movements: An Overview and Assessment”, *Annual Review of Sociology*, 26(1): 611-639.

Chetty, R., Hendren, N., Kline, P. and Saez, E., 2014, “Where is the Land of Opportunity? The geography of Intergenerational Mobility in the United States”, *The Quarterly Journal of Economics*, 129(4): 1553-1623.

Chetty, R. and Hendren, N., 2018, “The Impacts of Neighborhoods on Intergenerational Mobility I: Childhood Exposure Effects”, *The Quarterly Journal of Economics*, 133(3): 1107-1162.

Coleman, J. S., 1988, “Social Capital in the Creation of Human Capital”, *American Journal of Sociology*, 94: S95-S120.（金光淳訳, 2006,「人的資本の形成における社会関係資本」野沢慎司編・監訳『リーディングスネットワーク論――家族・コミュニティ・社会関係資本』勁草書房。）

Evans, J. A. and Aceves, P., 2016, “Machine Translation: Mining Text for Social Theory”, *Annual Review of Sociology*, 42: 21-50.

Flores, R. D., 2017, “Do Anti-immigrant Laws Shape Public Sentiment? A Study of Arizona’s SB 1070 Using Twitter Data”, *American Journal of Sociology*, 123 (2): 333-384.

Golder, S. A. and Macy, M. W., 2014, “Digital Footprints: Opportunities and Challenges for Online Social Research”, *Annual Review of Sociology*, 40: 129-152.

Goldthorpe, J. H., 2015, *Sociology as a Population Science*, Cambridge University Press.

Legewie, J., 2016, "Racial Profiling and Use of Force in Police Stops: How Local Events Trigger Periods of Increased Discrimination", *American Journal of Sociology*, 122 (2): 379-424.

Mayer-Schönberger, V. and Cukier, K., 2013, *Big Data–A Revolution That Will Transform How We Live, Work, and Think*, Houghton Mifflin Harcourt. (斎藤栄一郎訳, 2013,『ビッグデータの正体：情報の産業革命が世界のすべてを変える』講談社。)

McFarland, D. A., Jurafsky, D. and Rawlings, C., 2013, "Making the Connection: Social Bonding in Courtship Situations", *American Journal of Sociology*, 118 (6): 1596-1649.

McFarland, D. A., Lewis, K. and Goldberg, A., 2016, "Sociology in the Era of Big Data: The Ascent of Forensic Social Science", *The American Sociologist*, 47 (1): 12-35.

Morgan, S. L. and Winship, C., 2014, *Counterfactuals and Causal Inference 2nd. ed.*, Cambridge University Press.

Pentland, A., 2012, "The New Science of Building Great Teams", *Harvard Business Review*, 90 (4): 60-69.

Piskorski, M. J. and Gorbatâi, A.D., 2017, "Testing Coleman's Social-norm Enforcement Mechanism: Evidence from Wikipedia", *American Journal of Sociology*, 122 (4): 1183-1222.

Salganik, M. J., 2017, *Bit by Bit: Social Research in the Digital Age*, Princeton University Press.

瀧川裕貴, 2018,「社会学との関係から見た計算社会科学の現状と課題」『理論と方法』33 (1)：132-148。

Weber, M., 1921, *Soziologische Grundbegriffe,* Tübingen: J. C. B. Mohr. (阿閉吉男・内藤莞爾訳, 1987,『社会学の基礎概念』恒星社厚生閣。)

特集論文

4

データジャーナリズムの歩みと可能性

奥山晶二郎
朝日新聞withnews編集長

現代社会にとって欠かせない様々なデジタル技術は，ビッグデータと呼ばれる膨大なデータの出現と共に発展してきた。今日の人工知能の発展は，ウィキペディアのようなデジタル上の巨大な情報群がまずあり，そこに言語処理や機械学習の技術が加わったことで生まれた。そして今，SNSでの発信内容や検索単語の動向などは，新しい「世論」として注目され，世の中の動向を知る貴重な手がかりになっている。地図やグラフを使ったデジタル上の表現の進化は情報伝達の場面で新たな可能性を広げている。「ダボス会議」を主催する世界経済フォーラムは，2018年12月に発表した報告書で「デジタル環境」がすでに「自然環境」と同じ存在感を持っていると指摘している[1)]。

国家と企業と市民にとって，切っても切り離せない存在になったビッグデータに対して，ジャーナリズムも，従来の取材対象と同じ姿勢で取り組まなければならない時代になっている。本稿では，データジャーナリズムについて，筆者が新聞社のデジタル部門に所属する新聞記者として関わってきた経験を元に，従来の取材手法との違いやこれまでの事例，フェイクニュースが問題化する現在において，中立公正を担保できる取材手法の可能性などについて，以下の5項目の切り口で考察する。

1 「データジャーナリズム」の定義

本題に入る前に，データを取り巻く環境によってデータジャーナリズムの定義自体，大きく変わり続けていることを確認しなければならない。

2011年にデータジャーナリズムの意義や手法についてウェブ上で公開された「The Data Journalism Handbook」（Gray et al, 2012）では，データジャーナリズムの特徴について，従来の取材手法に加えるものとして，「現在利用可能な壮大なスケールと範囲に及ぶデジタル情報を組み合わせることで開ける新たな可能性」にあると述べている[2)]。

事実，この論文が発表された2011年以降も，あらゆるモノをインターネットでつなぐ「IoT」と呼ばれる技術の普及に伴い，取得できるデータの種類は広がっている。日本では2020年に導入予定の超高速・大容量の「第5世代移動通信方式（5G）」が浸透すれば，動画などの表現方法は一変するだろう。「The Data Journalism Handbook」が用いる「可能性」という表現自体が，データジャーナリズムの特徴を表していると言える。

上記の前提に立った上で考えた場合において，データジャーナリズムの特性は，取材の元となったデータの公開範囲について，従来の取材

との違いがでることが多い。報道機関が取材で得た情報は原則，記事として発信されるもの以外，外部に出ることはない。報道目的以外の使用は憲法が保障する「報道の自由」を侵しかねない行為で，過去，テレビ局が撮影した映像を証拠として提出させるか否かが問われた訴訟では「報道・取材の自由や当事者以外のプライバシーが侵害される恐れが高い」として最高裁が証拠提出を認めない判断を示している[3]。

従来の取材が報道機関の厳密な管理下にある情報を土台にしているのに対し，データジャーナリズムは公開情報を扱うことが多い。記事の評価は，発見されたデータそのものよりも，データの解釈や，従来の取材手法で得た関係者の証言などを含む異なる種類の情報を組み合わせることによって明らかになる事実に重きが置かれやすい。

一方，世論調査や，政治家，自治体へのアンケートなど報道機関が独自に入手したデータの場合，全ての情報を公開することは原則ない。企業の内部に蓄えられたデータについて取材などを通じて入手した場合も，記事化された情報以外を公開することは原則なく，公開情報か否かでデータジャーナリズムを定義することができないことも留意しなければならない。

データの公開範囲に加えて，データジャーナリズムはデジタル空間を意識した形で発表されることが多い。従来の紙中心の報道では，データを元にした記事であっても，資料の一部を撮影した写真や図表をつけることはあるものの，発信手段は紙面に印刷されたテキストが中心で，デジタルの発信の場合も，文字情報が主役になりがちだった。データジャーナリズムの場合，膨大なデータのすべてを表現することに主眼が置かれることが多く，データの見せ方自体もコンテンツの価値になり，様々な表現方法が生み出されている。データジャーナリズムにおいて，報道機関はデータの発見や解釈だけでなく，ビジュアライゼーションと呼ばれる，データの可視化の手腕も問われていると言える。

一方，扱うデータの種類によってはテキスト中心の報道であってもデータジャーナリズムとしての価値を発揮できる場合はある。2018年7月の九州北部豪雨で「#救助」というハッシュタグのついたツイートが実際の救助活動に結びついたかどうかを調べた際には，公開情報である4万2,750件の投稿を分析しているが，紙面やデジタルに発信された記事においても一部の投稿しか紹介していない[4]。

2 データジャーナリズムの歩み

データジャーナリズムの歴史にとって，デジタル技術の進化は切っても切り離せない。日本で本格的にブロードバンド（高速・大容量通信）のサービスが始まった2000年以降，インターネット上で提供されるサービスが次々と生み出されていった[5]。2008年には日本でツイッターのサービスが始まり，2年後の2010年には日本からつぶやく人の割合は最大の米国に次ぐ15%に達したと報道されている[6]。送受信される通信量は急増しており，米シスコシステムズによると，世界全体の月間トラフィックは2008年に10エクサバイト（1ギガバイトの10億倍）だったのが2016年には96エクサバイトになっており，デジタル空間に膨大な情報が発信されるようになった[7]。デジタル空間でやりとりされるデータのすべてが公開されているわけではないが，データ量の増加は活用できる情報の拡大にもつながっている。

インターネットが進化する一方で，従来の報道機関の立ち位置も変化してきた。国内の新聞発行部数は2000年5,370万部だったのに対し，2010年には5,000万部を割り，4,932万部となっている[8]。紙面の購読者が減るなか，デジタル発信への取り組みは強化され，2010年3月に日本経済新聞が有料電子版の事業を開始[9]。2011年5月には朝日新聞が有料電子版である「朝日新

聞デジタル」をスタートさせている[10]。

メディアをめぐる外部環境の変化から，国内の報道機関がデジタルへの取り組みを本格化させていくなか，データジャーナリズムは新しい発信手段として注目された。なかでも日本のメディアが参考にしたのが，英語圏のメディアの事例だった。

英紙ガーディアンが2011年12月に公開した「England riots: was poverty a factor?[11]」は，2011年にイングランドで起こった暴動について，当時のキャメロン首相の「貧困は原因ではない」という発言を逮捕者の住所や貧困地域のデータを組み合わせて検証。摘発者の約6割が英国の最貧困地域の居住者だったことを明らかにした[12]。コンテンツは，公開情報を元に，複数のデータを組み合わせる「マッシュアップ」という手法を用いている。分析の際には，グーグルのスプレッドシート（表作成ソフト）と，地図上にデータを表示するソフトであるフュージョンテーブルでデータを整理し，グーグルマップに落とし込むなど，主なツールとして無料のサービスを活用している。公開情報や無料で使えるデジタル上のツールを駆使している点は，データジャーナリズムの特徴をとらえているといえる。

2012年9月には米ニューヨーク・タイムズ電子版が民主・共和両党大会での議員の言葉を可視化した「At the National Conventions, the Words They Used」を公開した[13]。議員によって発言された言葉を円で分類し，出現回数に比例して円の大きさを変え，円の中を2色に分けて面積比によって党による違いも表現した。単語の出現回数を可視化させる手法は，その後のデータジャーナリズムでも多用されていく。

2012年12月，同じく米ニューヨーク・タイムズ電子版が公開した雪崩事故の特集記事「スノーフォール」は，デジタル上の新たな表現手法として注目された。上から下に画面をスクロールさせると，動画や写真，コンピューターグラフィックスが次々と現れる表現は「イマーシブコンテンツ」と呼ばれ，優れたジャーナリズム作品に贈られるピュリツァー賞を受賞した[14]。

2013年2月には，ロイターが8カ月かけて取材した中国の政治指導者の関係性をビジュアライズした「CONNECTED CHINA」を公開。紙面には載せきれない膨大な情報であっても，デジタル上ならユーザーに伝えられる可能性があることを示した[15]。

3　データジャーナリズムの実践例

日本の報道機関がデータジャーナリズムに取り組むきっかけの一つになったのが2011年3月に起こった東日本大震災だった。国内ではインターネットが本格的に普及した後に起こった初めての大災害であったことから，ツイッターなどソーシャルメディアでの発信や，グーグルなどインターネット事業者が安否確認サービスを提供するなどの動きが生まれた。報道機関もツイッターを使い，交通や避難所に関する情報，食料，給水，帰宅困難者の受け入れ場所などについて，新聞紙面の締め切り時間に縛られず発信した[16]。デジタル空間に発信された誰もがアクセスできる情報が，災害という人の生命に関わる状況で活用される場面が生まれていた。

筆者は2012年3月，首都大学東京の渡邉英徳准教授（現東京大学教授）の研究室などが東日本大震災発生直後に作成した「東日本大震災アーカイブ」と連携したコンテンツ作りに参加した[17]。「東日本大震災アーカイブ」はデジタル上の地球儀「グーグルアース[18]」上に東日本大震災関連の写真や動画などを実際に撮影された場所の緯度経度に表示させるというもの。筆者は，被災地で取材をする記者たちが得た現地住民の震災に関する証言を「東日本大震災アーカイブ」でも発信するための許諾の調整や，デジタル上で表現できる素材の加工などを

担当した。住民の証言は紙面上ではマス目上のレイアウトで配置されていたが，「東日本大震災アーカイブ」では，住民が住んでいた場所に証言が表示され，ユーザーが自分でデジタル上の地球儀の縮尺や視点を変えながら読むことができるようにした。

東日本大震災を巡っては，2012年9月に「東日本大震災ビッグデータワークショップ」の取り組みも生まれている。グーグルやツイッター，朝日新聞社，NHKなど8社が，それぞれの組織にあるデータを提供し合い，危機対応や事前の対策について検証した。震災発生から1週間の記事や検索キーワード，カーナビの走行情報などのデータを研究機関や企業に提供するというものだった[19]。

2013年3月，筆者らは検索エンジン大手ヤフーのデータを用いて，東日本大震災から2年間のデジタル空間の状況を伝えた。「被災地」と「求人」の二つの単語を組み合わせて検索する人は，岩手・宮城・福島の被災3県より，秋田・青森の両県の方が多いことが明らかになった。分析したデータと現地での取材とを組み合わせた結果，検索傾向の背景には，両県の有効求人倍率の低さなどの影響があることがわかった[20]。

災害報道と並行して試みたのは，候補者関連の膨大な情報が集まる選挙報道だった。筆者らは2012年12月に投開票された衆院選の選挙期間中に発信されたツイッターの投稿を分析する企画「ビリオメディア」を発表した。当時，日本でもツイッターのユーザー数は1千万人を超えていたことから（野村総合研究所, 2013），ツイートデータはデジタル上に発信された人々の考えを知る重要な取材源になると考えた。

衆院選に立候補者を立てた12政党の名前や主な略称が含まれるほぼすべてのつぶやき計約460万件を抽出した結果，「自民」を含むものは計150万7,281件あり，「民主」の99万9,903件の約1.5倍に達していたことがわかった。選挙は，自民党が294議席（公示前118議席），民主党は57議席（公示前230議席）という自民党の圧勝という結果に終わっていた[21]。記事では，自民圧勝という選挙結果の一端が，ツイッターでの政党名の出現回数にも現れていたことを，政治学の研究者の分析とともに伝えた[22]。ツイッターの分析では，ビッグデータ分析を手がけるプラスアルファ・コンサルティング社の協力を得た。データは，企業がマーケティングなどに使うシステムを用いて抽出した。政党名などにはデジタル上で使われる特有の言い換えがあることから，テストデータを抽出しながら単語の整理を進めた。例えば「共産党」の場合は「中国共産党」は除外するなどの規則を設定し，検索の手がかりとなる辞書セットを作成。報道用として扱っても問題のない精度まで高めた。

筆者らは，2013年7月にあった参院選でもツイッター分析を実施した。選挙関連のツイートの数は参院選前では，衆院選前に比べ約125万7千件減っていたことから，デジタル上で選挙への関心が薄れている現れであると伝えた。

ツイッターの分析においては，報道機関が選挙のたびに実施する「世論調査」と違い，取材者があらかじめ設定していない問いに対する答えが出るところに大きな特徴があった。2013年7月の参院選に関するツイッターを分析した結果，当時，大きな争点とは考えられていなかった「児童ポルノ禁止法案」が熱心に議論されていることがわかった。調査には，東北大学の乾健太郎教授と岡崎直観准教授（現東京工業大学教授）の協力を得た。調査に参加した岡崎氏はツイッター分析について「世論調査では見つけられない，思ってもいない動きを浮かび上がらせることができる」と述べている[23]。

同じく2013年7月の参院選では，東京大学の松尾豊特任准教授の協力を得て候補者のツイッターアカウントのフォロワーについても調査した。フォロワーのフォロワーである「孫フォロワー」の多さから，候補者のツイートがリ

ツイートされた率と「孫フォロワー」の数を掛け合わせた「拡散力」を割り出した。候補者のフォロワー同士の関係性も調査し，ある候補者のフォロワーが，別の候補者をフォローしている結びつきを辿った結果，公明党は党内部か自民党とのつながりがほとんどで，他党への広がりが見えないことなどがわかった[24]。

4　データジャーナリズムの課題

2012年から本格的にデータジャーナリズムに取り組んで以降，筆者らは継続的にコンテンツを発表してきた。

2013年12月には，岩手県大槌町に駐在していた朝日新聞の記者が保存していた新聞紙と一緒に配達される折り込みチラシ5,551枚について，東日本大震災発生から1,000日の間に枚数や内容がどのように変化したのかを伝える「チラシでたどる震災1000日」を発表した。津波によって壊滅的な被害を受けた大槌町では，新聞配達が再開された後も長く折り込みチラシは止まっていた。震災から77日目にあたる2011年5月26日に再開した際に配られた折り込みチラシは1枚だけで，町外の重機リース会社によるものだった。デジタル上の特集サイトでは上から下にスクロールすることでチラシが日を追うことに積み重なっていく様子を表現。手書きで営業再開を伝える自動車整備工場のチラシや，町の中心施設だった大型スーパーの再開を知らせるチラシなどを，現地で起きた震災関連の出来事に連動する形で体験してもらうことで，復興の度合いと震災によって起こる様々な問題の関係性を可視化した[25]。

2014年3月には，政府が指定した2015年度までの「集中復興期間」に予算化される25兆円の使い道などを分析する「お金でたどる震災3年」を発表した。被災3県の決算情報から，単年度で終わらない大型事業の増加と工事業者の不足による入札の不成立によって自治体の基金の残高が急増している状況を地図とグラフを組あわせて可視化した[26]。

2014年7月，筆者らはスマートフォン向けの新たなウェブメディア「withnews」を立ち上げた。「withnews」は，「Yahoo! News」など外部のプラットフォームで新聞社の記事を読んでもらうことを柱にしている。「withnews」立ち上げの背景には，データジャーナリズムのコンテンツを手がける中で見えた課題があった。特に大きかったのが，労力に見合うだけの効果を数字として出しにくいことだった。

データジャーナリズムを発表する際には，数ヶ月から半年以上かけて様々な工夫を施したページを準備することが少なくない。通常の取材よりも労力をかけた企画の場合，紙面上では1面など目立つ位置に配置することで，取材成果に見合う効果を生み出すことが期待できる。一方，デジタルの世界は，自社の記事であっても新聞社の裁量で記事の流通は制御できず，読者の目に触れるまでには，ポータルサイトやSNSなど，新聞社以外の事業者，個人の関与が必要になる。ロイターの調査によると，日本のインターネットユーザーの51%は「Yahoo! News」経由で記事に接触しているのに対し，朝日新聞（Asahi Shimbun online）は9%にとどまっている[27]。デジタル上のニュースコンテンツの流通を考える上では「Yahoo! News」のようなポータルサイトに適した形式で発信することが重視されるが，データの見せ方などをこだわったページの多くは，自社サイトでしか動作できないため，多くの人の目に触れるポータルサイトには直接配信できない。デジタルの世界で重要な流通面での課題は，データジャーナリズムの成果を認知させる場面で大きな障壁になっている。

人材の面でも課題は少なくない。データジャーナリズムにおいて，記者は情報を発掘するだけでなく，様々な専門分野の人間をまとめるディレクターとしての役割が求められる。

ビッグデータの分析には自然言語処理，機械学習，ビジュアライゼーションにおけるHTMLの技術などの知識が欠かせない。新聞社内部に専門家がいなければ，大学や企業など，外部との連携が必要になる。研究者の専門分野と報道との相性を判断する知見がなければ，連携相手を見つけることができない。取材の際には，膨大なデータを分析する調査方法の妥当性も考えなければならない。投票行動に影響を与える選挙報道の場合，論文としての正確さとは別に，当事者である立候補者や有権者である受け手に読者として納得してもらう配慮を手当てしなければならない。分母となるデータはサンプル調査で分析可能な場合でも，あえて全データを対象にするよう判断も必要になる。

一連の作業で必要な技術と報道を組み合わせた人材の育成基盤は，ほとんどの新聞社で整っておらず，コンテンツが生まれるか否かは記者個人の力量と担当分野との相性に左右されることが多い。人事異動などで担当が変わると，データジャーナリズムの取り組みを継続できなくなることも少なくない。

現在，継続的にデータジャーナリズムに取り組んでいる報道機関は，2015年5月から「日経ビジュアルデータ」として企画を発表している日経新聞以外，目立った存在がいないのが実情だ[28]。

取材源であるデータ取得でも課題が生まれている。デジタル上には膨大な情報が存在するものの，統計指標など分析に適した形である構造化されたデータとして使えるものは一部にとどまる。ツイッターの投稿文など，自然言語で発信された非構造化データを扱おうとすると，単純な単語の出現回数でしか測れないことが多い。ネガティブ，ポジティブの極性判定まで踏み込むには，日本語の複雑な係り受けの法則を見極めなければならず，報道目的に扱うにはハードルが高いのが現状だ。IoTの発展により企業内部に蓄えられたデータも増えているが，個人情報保護との兼ね合いから外部への提供は限定的になることが多い。筆者らは2013年7月の参院選や，2016年4月の熊本地震などにおいて，検索エンジン大手のヤフーからデータの提供を受けているが，データジャーナリズムの成果として報じる際には，絶対値ではなく相対値を伝えている[29]。

5 データジャーナリズムの可能性

検索サービスやSNSなどが浸透したことで，読者は自分が好む情報を簡単に入手できるようになったが，偏った考えが固定化し異なる意見を排除する社会の分断化を生むことにもつながっている[30]。既存の報道機関を介さず，ツイッターやフェイスブック，インスタグラムやブログなど様々なプラットフォームからメッセージを発信する政治家や芸能人が現れ，読者の意思決定に影響力のあるインフルエンサーと呼ばれる存在も生まれている。

一方で，公益財団法人新聞通信調査会の調査によると「新聞への信頼感が低くなった理由」として，「特定の勢力に偏った報道をしているから」と答えた人の割合は年々上昇しており，2016年度調査で29.7%だったのが，2018年度調査では46.7%になっている。

2016年の米大統領選で問題となったフェイクニュースの背景には，インターネット上に様々な情報があふれるようになった変化に加え，メディアに対する不信感があることも報道機関は受け止めなければならない状況になっている[31]。

筆者は，報道機関をとりまく環境が大きく変わる中，客観的なデータを中心に記事を組み立てるデータジャーナリズムには，メディア不信を払拭させる可能性があると考える。

記事本文で使用する部分しか表に出さなかった従来の報道に比べ，公開情報を元にしたデータジャーナリズムによって生まれる記事は，反

証する機会を担保することができる。記事に対して批判的な意見が出たとしても，客観的データと一緒に議論をすることで，足りないデータの再調査が提案されたり，データに対する記事とは異なる解釈が提示されたりといった，共創的な動きが生まれる可能性がある。

朝日新聞社と東京大学谷口将紀研究室は2003年から衆院議員選挙と参院議員選挙時に候補者に対してアンケート調査を続けている。紙面上で調査結果の分析記事を掲載するとともに，2012年の衆院選からはデジタル上での独自コンテンツの制作にも本格的に取り組んでいる[32]。調査結果はデジタル上で公開しており，無料でダウンロードできるようになっている[33]。

メディア不信の高まりと同時に，既存の報道機関に接する人も減少している。NHK放送文化研究所による「2015年 国民生活時間調査」の「新聞の行為者率」では，1995年に「国民全体」の52%が平日に新聞（電子版を含む）を読んでいたのに対し，2015年には33%にまで減少している。テレビも1995年に92%だったのが，2015年は85%となっている。対して2005年から調査を始めた「趣味・娯楽・教養のインターネット」で2005年にインターネットを利用していた「国民全体」の数字は13%だったのが，2015年は23%に増加している（NHK放送文化研究所, 2016）。

既存メディアの読者離れ，視聴者離れに対して，ニューヨーク市立大学大学院のジェフ・ジャービス教授は，啓発的な性格が強かったジャーナリズムの役割に，サービス産業としての価値が求められていると指摘する（Jarvis, 2014）。発信手段が限られていた時代，物理的制限から，報道機関が個々のユーザーに合わせて作り方を変えることは難しく，最大公約数としての重要な情報を精査して届けることに専念していた。一方，デジタルサービスの多くは，検索サービスに代表されるように，個々のユーザーの趣味趣向に細かく応える形で発展してきた。データの網羅性に特徴があるデータジャーナリズムなら，公開された膨大な情報から個々の読者の興味関心のあるものだけを届けることができ，サービスとしての価値を報道に生み出すことができる。

朝日新聞が，全国的に不足する待機児童問題の現状を伝えるために制作した「待機児童問題「見える化」プロジェクト」では，「あなたの街の待機児童」という項目を設け自治体ごとの待機児童数や保育所数，定員，利用児童数を可視化した。関心のある自治体をページから選ぶことで，待機児童の実態を知ることができるようになっている[34]。細かなデータも扱いやすい形で公開できれば，ジャーナリズムとサービス，両方の機能を果たすことができる。

同じく朝日新聞が制作した「揺れやすい地盤」は，防災科学技術研究所の「地震ハザードステーション」のデータを用いて，任意の住所を入力することで「揺れやすさの目安（表層地盤増幅率）」を数値化したものが表示されるようになっている[35]。大災害に至らない規模の地震速報であっても，人々の関心が高まる機会に防災への意識を高めてもらうため，地震情報を伝える記事の中に関連情報として特集サイトへのリンクをつけている。

2016年4月，朝日新聞は，南ドイツ新聞と非営利の報道機関「国際調査報道ジャーナリスト連合」（ICIJ）が入手したカリブ海の英領バージン諸島などのタックスヘイブン（租税回避地）に設立された21万余の法人に関する電子ファイル「パナマ文書」についての分析記事を報道した。本稿の冒頭で紹介した「The Data Journalism Handbook」は，2011年の「Mozilla Festival」というエンジニアやデザイナーなどが集まるカンファレンスの中で生まれており[36]，日本語訳も，国内のメディア関連の研究者や報道機関の記者の有志によって手がけられている[37]。「パナマ文書」や「The Data Journalism Handbook」の取り組みは，データが

媒介となり，報道機関同士，あるいは報道機関と研究者らとの間に連携が生まれ，新しい報道の形につながる可能性があることを示している。

個人の趣味趣向が細分化し，思想信条が分断化しつつある現代社会において，データジャーナリズムには社会のハブ役になり得る価値があると考える。最後に，本稿がきっかけとなり報道機関と研究者，エンジニアらとの連携が進むことを期待している。

文献

Gray, J., Bounegru, L. and Chambers, L. (eds.), 2012, *The Data Journalism Handbook*, O'Reilly Media.

Jarvis, J., 2014, *Geeks Bearing Gifts: Imagining New Futures for News*, New York: CUNY Journalism Press. (夏目大訳, 2016,『デジタル・ジャーナリズムは稼げるか ― メディアの未来戦略』東洋経済新報社。)

NHK放送文化研究所 (編), 2016,『データブック国民生活時間調査2015』NHK出版。

野村総合研究所, 2013,『ソーシャルメディア利用実態』(http://www.soumu.go.jp/main_content/000208354.pdf)

注

1)“World Economic Forum Report Addresses Crisis of Trust, Slowing Growth in Our Digital World.”https://www.weforum.org/press/2018/12/world-economic-forum-report-addresses-crisis-of-trust-slowing-growth-in-our-digital-world

2) データ・ジャーナリズム・ハンドブック日本語翻訳プロジェクトによる日本語版がウェブ上で無料公開されている。(http://datajournalismjp.github.io/handbook/)

3)『朝日新聞』2017.07.28朝刊「警官制圧で死亡，映像提出「不要」　鹿児島の事件，最高裁」

4)『朝日新聞』2017.11.04朝刊「救助要請224件，拡散92578件に　警察受理は1件 … 受け皿作り課題」

5)『朝日新聞』2004.01.15朝刊「ADSL利用1000万件突破　昨年末，総務省まとめ」

6)『朝日新聞』2010.05.24夕刊「(メディア激変:37) 発祥の地から:4　つぶやき大国日本へ」

7)『日本経済新聞』2017.10.24朝刊「ネット投資の行方は ―― 米シスコシステムズ日本法人社長鈴木みゆき氏，中小の伸びしろ大きく（経済観測）」

8) 日本新聞協会「新聞の発行部数と世帯数の推移」(https://www.pressnet.or.jp/data/circulation/circulation01.php)

9)『日本経済新聞』2010.03.23朝刊「日経電子版創刊しました，多彩なコンテンツ，機能も充実」

10)『朝日新聞』2011.05.19朝刊「<お知らせ>創刊，朝日新聞デジタル　電子新聞，新しい形」

11)https://www.theguardian.com/news/datablog/2011/aug/16/riots-poverty-map-suspects

12)『朝日新聞』2012.05.14夕刊「(データジャーナリズムの世界:1) 暴動の背景，あぶり出す」

13) https://archive.nytimes.com/www.nytimes.com/interactive/2012/09/06/us/politics/convention-word-counts.html

14)『朝日新聞』2013.10.06朝刊「(ワールドけいざい) 米の新聞，変革に挑む　読者に合わせ『パッケージ化』」

15) http://china.fathom.info/

16)『朝日新聞』2011.10.15朝刊「被災者の『言いたいこと』をストレートに　震災と原発事故　新聞週間特集」

17)http://shinsai.mapping.jp/index_jp.html

18) 現在は別のシステム「cesium」を使用

19)『朝日新聞』2012.09.13朝刊「東日本大震災当時，情報はどう流れた　研究者ら検証」

20)『朝日新聞』2013.03.13朝刊「(ビリオメディア) ネットの関心，動く　東日本大震災，つぶやき・検索この2年」

21)『朝日新聞』2012.12.18朝刊「党派別当選者の内訳　衆院選」

22)『朝日新聞』2012.12.22朝刊「(ビリオメディア) 自民に『期待』『嫌だ』つぶやき白熱　衆院選, ツイッター分析」

23)『朝日新聞』2013.07.26朝刊「(ビリオメディア) つながる力, 次こそ真価　参院選」

24)『朝日新聞』2013.07.03朝刊「(ビリオメディア) 「原発」つぶやき続く　衆院選から参院選へ, ツイッターを分析」

25)http://www.asahi.com/shinsai_fukkou/otsuchiad/

26)http://www.asahi.com/shinsai_fukkou/3nen/

27)Reuters Institute Digital News Report http://www.digitalnewsreport.org/survey/2018/japan-2018/

28)『日本経済新聞』2015.05.17朝刊「特集 ―― 日本経済新聞電子版, 電子版トップページ刷新, 紙面連動企画もスタート」

29)『朝日新聞』2016.09.13朝刊「熊本地震, そのとき検索したのは　ヤフー分析」

30)『朝日新聞』2018.05.10朝刊「(耕論) 揺らぐ言論の土台　スマイリーキクチさん, アーロン・シャロックマンさん, 上井靖さん」

31)『朝日新聞』2017.05.23朝刊「メディア不信の時代に　第6回メディアフォーラム」

32)http://www.asahi.com/senkyo/sousenkyo46/asahitodai/

33) 東京大学大学院法学政治学研究科谷口将紀研究室:東大谷口研・朝日調査ウェブサイト　http://www.masaki.j.u-tokyo.ac.jp/utas/utasindex.html

34)http://www.asahi.com/special/taikijido/

35)http://www.asahi.com/special/saigai_jiban/

36)http://datajournalismjp.github.io/handbook/%E5%89%8D%E4%BB%98%E3%81%91_0.html

37)http://datajournalismjp.github.io/releases/20161118.html

特集論文

5

ビッグデータによって変わる未来の公的統計

水野貴之
国立情報学研究所 准教授

1 はじめに

公的統計は社会・経済に対する人間ドックのような働きをしている。現在の社会・経済の状況だけではなく，どこに問題があるのか，問題に対する対策の効果はでているのか，将来はどのようになるのかということを把握するために必要不可欠な情報である。我々が，健康診断ではなく人間ドックに行って，身体のありとあらゆる箇所を調べてもらうにはお金がかかるし，検査結果が出るまでにも時間がかかる。これが国全体の調査であるから，どれほどの検査項目があるのか，どれほどのお金がかかるのか，どれほどの調査期間が必要であるのか，想像し難い。先進国では投資家から高頻度で低遅延の統計が求められるし，途上国では厳しい予算制約の中で統計作成が求められる。このような問題を解決するために注目を集めているのが，民間が保有するビッグデータ（以降，ビッグデータと呼ぶ）である。ビッグデータを活用することにより，(1) 新たな統計の作成，(2) 従来の公的統計よりも精細な情報提供，(3) 公的統計の早期化（公表周期の短縮化），(4) 公的統計のナウ・キャスティング，(5) 公的統計の精度向上，(6) 公的統計の報告者の負担軽減，(7) 公的統計の製作コストの削減と効率性向上が期待できる（Struijs, 2016）。

IMFによると，ビッグデータは，個人生成型のソーシャルネットワーク情報と，取引仲介型の商業取引データ，デバイス生成型のIoT活用によるデータの３つに分類できる（Hammer et al., 2017）。ソーシャルネットワーク情報には，FacebookやTwitter，LinkedInなどのSNS情報，インターネット検索エンジンでの検索情報が該当し，これらを活用して失業率のナウキャスティングなどが各国でおこなわれている（Antenucci et al.,2014; Bean, 2016; D’ Amuri. and Marcucci, 2012）。

商業取引データには，商取引記録，銀行決済，民間企業のウェブサイト情報などがある。小売店のPOSデータやネットスーパーでの商品価格を活用することで，物価の把握が効率的に精度良く即時性を持ってできることが知られている（Cavallo and Rigobon, 2016; 東大日次物価指数プロジェクト, 2014～）。

IoT活用によるデータは，携帯位置情報（GPS）や自動車走行データなどが該当する。近年，この種のデータの伸びが特に顕著であり（総務省情報通信国際戦略局情報通信経済室, 2014），様々な領域の専門家から，これらのデータによる景気動向把握の可能性が指摘されている（NTTデータ経営研究所, 2015）。2017年，ドイツやデンマークでは移民局の職員が亡命希望者の携帯電話からデータを抽出できるように法律を改正した。把握の難しい移民・難民に関する統計での活用も期待されている。

ビッグデータは万能ではなく，公的統計への

活用には注意しなければならない点がある。従来の公的統計の算出に使われるデータは，予めサンプリング等が調査設計されたデータであるのに対して，ビッグデータは一般社会の経済社会活動の副産物として生成・蓄積されたデータであり，本来の利用目的に起因して，ノイズ（SNSの大半は意味のないおしゃべり）や，バイアス（Instagramは女性ユーザーが多い），持続性（Twitter社は2015年に位置情報取得を2段階承認に変更したために詳細な位置情報ツイートが激減），データ形式（あるPOSデータではラム酒がウィスキー分類），真実性（SNSの居住地設定が火星）という問題がある（別所，2018）。従って，公的統計に活用する場合には，データ供給元の知見と協力が欠かせない。そのためには，ビッグデータを社会インフラ・公共財として価値観共有をおこなうこと，データ供給元へのインセンティブ設計が重要である。

以降の章では，まず2章で，我が国の省庁における既存の公的統計の速報性の向上やコスト削減に向けたビッグデータ活用の取組について紹介する。3章では，海外事例について紹介する。たとえば税収の少ない小国では，ビッグデータが積極的に活用されている。4章では，ビッグデータによる新たな公的統計の可能性について指摘し，5章では，ビッグデータによる公的統計作成の課題について述べ，まとめとする。

2　日本の動向

2.1　背景

現在の公的統計におけるビッグデータの活用の検討は，第Ⅲ期公的統計基本計画（2018年度）にもとづいて進められている。ここに至る我が国の背景としては，2014年の「世界最先端IT国家創造宣言」・「日本再興戦略」でビッグデータ活用による新サービスの創出を謳い，2015年10月の経済財政諮問会議における麻生財務大臣の「経済情勢を的確に把握するためには，GDPを推計するもととなる基礎統計の充実に努める必要がある」との発言のもと，骨太方針2016，骨太方針2017を経て，「証拠に基づく政策立案」（Evidence-Based Policy Making）の推進が決まった。これを支える基盤として公的統計を強化すべく，ICTの進展により生成・収集・蓄積等が可能・容易になる多種多様のビッグデータの活用が，現在の公的統計基本計画で閣議決定された（総務省，2018）。

このような背景のもと，公的統計改革がスタートした。公的統計は，景気サイクルや物価の上昇などを捉えるために過去の指数との比較が必要である。それには調査条件を同様にする必要があるが，時代の流れとともに変化する社会構成には対応できない。ビッグデータを活用することで，既存統計で把握できていない経済活動の把握に努めることが期待されている。また，POSデータ等を積極活用することで，調査から統計公表までのタイムラグや，公表周期を短くすることが望まれている。さらに，国連が定めた持続可能な開発目標（SDGs）等の統計のグローバル対応や，地方自治体とも共働した地方経済の動向把握なども進められている。しかし，ビッグデータは企業活動の副産物であるために，あらゆる面で標準化や統一化がなされていないという点で，利活用に向けての障壁となっている。そこで，各省庁では，企業等からのデータ提供のあり方，データの品質の確保，専門人材の育成等について産官学連携で取り組んでいる。以降，我が国における各省庁及び公的統計に関連する民間の取組について紹介する。

2.2　取引仲介型の商業取引データの検討状況

我が国におけるビッグデータの公的統計への活用の中心は，商業取引データ，特に小売店から収集されたPOSデータの活用である。総務省では，消費者物価指数の算出において，製品サイクルが極めて短いパソコンやカメラの価格に

ついてPOSデータを利用している。また，2018年からは，各社のPOSデータを順次活用し，消費動向をマクロ・ミクロの両面から捉える速報性の高い消費指標（参考値）：消費動向指数（CTI）の公表を始めている。内閣府においても，POSデータによる経済情勢の変化の早期把握に向けた取組が概算要求の資料等から確認できる。民間等での取組については，東大日次物価指数プロジェクトがよく知られており，スーパーマーケットの日々のPOSデータを使うことで日用品の価格と販売量を把握し，月次単位の公的物価指数CPIを日次で補間することが可能であるとの報告がある（東大日次物価指数プロジェクト，2014〜）。速報性を高めることで，東日本大震災後の物価上昇，タバコ増税前後の駆け込み需要と反動など，公的統計よりタイムリーに動向を把握することが可能である。

他の商業取引データの活用事例では，内閣官房及び経済産業省が地域経済分析システム（RESAS）において，企業間取引データを活用した地域経済を支える中核企業の分析プラットフォームを2015年度から提供している（経済産業省, 2015）。内閣府では，国民経済計算において各種企業の有価証券報告書を利用し，景気動向指数においては日経商品指数・東証株価指数・長期国債利回り・中小企業の売上見通しを利用している。また，世界各国の週次の資金流出入データや日次のM&Aデータを活用して，危機発生時における海外経済リスクの点検が検討されている。国土交通省でも，設備工事業に係る受注高調査において，関係する工業協会からデータの提供を受けている。労働市場に関する分析事例もある。総務省のワーキングペーパーでは，民間の人材紹介サービス企業が保有する「転職時の賃金変動状況」の情報から，公的統計では補足できない新たな雇用の動向把握や，雇用動向調査の先行指標を作成できる可能性を指摘している（高田他, 2018）。

2.3 デバイス生成型のIoTデータの検討状況

橋脚に取り付けられた振動センサーや，風力発電塔に付けられた風力計，家庭にあるスマートセンサーなど，近年，このようなIoTデバイスが爆発的に増加している。各省庁ではIoTデバイスから得られるデータの公的統計への活用の検討が始まっている。国土交通省では，交通・運輸関連のビッグデータのなかに，公共交通ICカード，車両感知器，携帯電話・スマートフォンの位置情報，カーナビの位置情報，速度情報，燃費情報，自動車の操作情報といったIoTデータが多いことから，いち早く活用に取り組んでいる。交通量調査は概ね5年に1回，全国24,000区間において人海戦術で観測をしており，予算面から高頻度に調査することはできない。そこで，2010年の調査では，民間事業者等が収集した車の位置情報であるプローブデータを活用することで，調査にかかるコスト軽減を図った（NTTデータ経営研究所, 2015）。さらに，2015年度においては，プローブデータにより被災状況の収集・分析がおこなわれた。2018年の東京都市圏のパーソントリップ調査では，域内の人々の移動を推計するために携帯電話基地局データ（位置情報データ）が利用された（国土交通省都市局都市計画課都市計画調査室, 2018）。これらプローブデータや位置情報データは，個人情報保護の観点から，匿名化処理前のマスターデータを国土交通省も扱うことができない。個人ごとの移動に着目した交通調査をおこなうには，どのように個人の同意を取るのかという問題が残っている。

2.4　ソーシャルネットワーク情報の検討状況

３章で紹介する諸外国における活用状況に比べて，我が国では活用事例が少ない。社会全体を反映した公的統計を作るためには，サンプリングバイアスの修正が必要だが，そのバイアスの修正に必要な属性情報をSNSデータや検索履歴データからは読み取りにくいという問題があ

るためだと思われる。それでも，経済・社会の大まかな傾向を読み取ることは可能であり，例えば，日本銀行レポート・調査論文では，東日本大震災前後のサービス消費（旅行）についての分析がおこなわれており，旅行関連の検索データが旅行取扱額のナウキャスティングにおいて有益な情報を有していると報告している（白木他，2013）。

3 海外の動向

3.1 背景

ビッグデータの公的統計への活用に絞って背景を述べていく。昨今，先進各国では公的統計のもととなる世論調査の回答率が下落している。また，グローバル化により新興国や途上国の経済状況を知りたいという要望があるが，予算制約が厳しい中，調査項目や頻度を改善することが難しい。このような背景により多くの国で公的統計改革の機運が高まり，2009年より国連において，ビッグデータの公的統計への活用に関する検討が開始された。そこでは，ビッグデータの活用は雇用，経済，人口についてリアルタイムの状況を把握することに役立つこと，国際的な協力を通じて活用技術を新興国や途上国に移転すること，公的統計への利用にはデータの信用性が課題であることが述べられた（United Nations Economic and Social Council, 2014）。同じくOECDでも，2012年の統計委員会では，政策形成過程にビッグデータが有する可能性や公的統計にとっての課題が議論され，2015年に，ビッグデータの増殖と公的統計及び公的機関に対する影響の予備分析報告書が公開された（Reimsbach-Kounatze, 2015）。

このような国際的なトレンドのもとで，経済規模の大きな国では，日本と同様に公的統計の多様化・精緻化・速報性を高めるためにビッグデータの活用が検討されている。より積極的に活用が進んでいるのが経済的に小さな国々で，オランダやエストニアなどが該当する。これらの国では，ビッグデータの活用には信頼性の面では不安があるが，経済や人口などの傾向を捉えられ，報告者負担や調査費用を削減できることに大きな利点を感じている。このような海外の利活用状況について，順に紹介する。

3.2 取引仲介型の商業取引データの検討状況

物価に関する新興国も含めたナウキャスティング事例として，しばしば実例に挙がるのは，米国MITの研究者が中心となって実施している「ビリオン・プライス・プロジェクト」である（Cavallo and Rigobon, 2016）。このプロジェクトでは，世界70カ国以上の約900のインターネット小売業者の商品価格を日々収集し，それをもとに各国の価格指数を作成している。2008年のリーマン・ブラザーズの破綻のさいには，公的統計の公表前に，小売り各社の販売価格引き下げを察知した。このようなオンライン市場の価格データやPOSデータを用いた物価の把握は，英国統計局，オランダ統計局，スウェーデン統計局，ルクセンブルク統計局，そして我が国の統計局でも，検討や試験的な運用がおこなわれている。

雇用統計に関してもビッグデータの活用が検討されている。毎月第１週目の金曜日に，米国労働省により前月の雇用統計が発表され，世界の金融市場に大きな影響を与えている。米国における民間就業者の約2割の給与計算業務を実施している給与計算アウトソーシング会社Automatic Data Processing（ADP）は，その給与データを活用して雇用統計の2営業日前に米国ADP雇用統計を発表している。この統計は先行指数として注目されている。連邦準備制度理事会のワーキングペーパーでは，その給与データを活用することで，公的統計を補間する週次指数や給与支払いの頻度指数の開発，速報値に対する改定幅の予測が可能であると分析している（Cajner et al,2018）。JP Morgan Chase

Instituteは，同行顧客2,800万人以上の口座記録からランダムに100万人を匿名化して抽出し，米国における賃金格差を計測した（Farrell and Greig, 2016）。給与水準が低い若年層の労働者において賃金格差が高いことや，米国西部でより強い格差を観測した。これら以外にも，求人広告をもとに地域単位での雇用のミスマッチ分析（Rothwell, 2012）や，米国カンファレンスボードによる求人広告数からの求人広告指数（HWOL）作成などが存在する。

3.3 デバイス生成型のIoTデータの検討状況

海に囲まれた我が国に対して内陸国では，隣国間の人の往来が激しく，国境管理にコストが掛かっている。エストニア中央銀行では，国境に係る公的統計の活用に代えて，携帯電話のSIMカード情報を活用することで旅行収支を推計している（Hammer et al, 2017）。欧州各国の移民局では，ドイツとデンマークが亡命希望者の携帯電話からデータを抽出できるよう，2017年に法を改正した。また，ベルギーとオーストリアでも同様の法案が提出されており，英国とノルウェーでは数年前から亡命希望者が所有するデバイスの調査がおこなわれている。

環境に関するセンサー関連のデータも利活用が進んでいる。ブラジルの水資源機関では，水道ネットワーク上にある22,000ものセンサーから水流，降雨，水質等を把握し，公的統計に組み込んでいる。コロンビアでは，衛星画像データを農業統計の作成に活用する試験的な取組が始まっている。また，カナダでも，スマートメーターで電力消費量を把握して既存統計に活用することを2014年から検討している。

3.4 ソーシャルネットワーク情報の検討状況

景気動向を捉える動きとしては，2016年にビッグデータ統計センターを設立し，ビッグデータの公的統計活用の先駆者を目指しているオランダで，統計局はツイート内容のセンチメント分析の結果が消費者信頼感指数と高い相関を持つことを明らかにし，これを活用して統計調査のサンプルサイズや調査頻度の削減が可能であることを示した。欧州中央銀行でも同様の分析がおこなわれており，SNSから数値化した消費者のセンチメントが消費者信頼感指数と同様な傾向を示しているとともに，１週間程度先行していることが示された。この関係性が安定的なものであれば，これまで以上の頻度で指数を公表することが可能である（Piet and Puts, 2014）。

雇用統計に関しても商業取引データと同様に検討が進んでいる。各国の既存の雇用統計では職種や必要スキル，勤務地情報，肩書き，職歴などに関する情報量が少なく，企業が求める労働者のスキルニーズの変化を捉えることが難しい。この問題を解決すべく，Eurostatや各国統計局からなる欧州統計システムのESSnet Big Dataプロジェクト等では，ウェブ求人情報の活用に取り組んでいる（Mandel and Scherer, 2015; Swier, 2016）。同じく，米国大統領経済諮問委員会では，職歴やスキル等を共有するビジネスSNSであるLinkedInを利用する米国内の５千万人以上の登録者情報をもとにして，詳細な雇用の把握が可能であることを指摘している（Bean, 2016; Council of Economic Advisors, 2012）。一方，失業に関しては，失業した人々による失業に関するSNSでの書き込みや，ネットでの職探しの検索履歴を利用して，失業率を早期予測する研究がおこなわれている。ミシガン大学では，失業に関するツイートをもとに作成した失業インデックス（The University of Michigan Social Media Job Loss Index）を作成し，イタリア中央銀行のスタッフらはGoogleにおける「jobs」の検索件数を調査した（Antenucci et al., 2014; D’Amuri and Marcucci, 2012）。これらは失業率と強い相関があるだけではなく，フィラデルフィア連銀が実施するエコノミスト予測（SPF）をアウトパフォームする結果を示す。

高頻度に失業率を推定することで，自然災害や政府閉鎖などの特定イベントに関わる影響分析が可能になる。

物価や家計に与える不動産価格の影響は大きい。Auction.comは，新築レポートやGoogleでの「Home appraisers in Irvine」等の検索件数などを活用して，不動産購入に関する公的統計を予測する不動産業界初のナウキャストReal Estate Nowcastを開発した（Auction.com, 2014）。また，インドネシアやアルメニアの中央銀行は，不動産仲介業者のオンライン情報を用いての住宅価格調査に着手している。

4　新たな公的統計の可能性

ビッグデータの利用により，既存の公的統計が補足できないような現象の把握が可能になってきた。ここでは，新たな公的統計の可能性について紹介する。

先進国に比べて，新興国や途上国では公的統計の整備が遅れている。このような国の経済状況（GDP）を把握するために，人工衛星から捉えた夜間光量を利用する試みがおこなわれている（Henderson, et al, 2012）。近年，負債を抱える国などでは，銀行が経営破綻するという噂や不確実な情報，デマが引き金となり，預金者が預金・貯金・掛け金を取り戻そうとして金融機関に押し寄せる事例が発生している。こうした預金取り付け（bank run）を，Googleにおける「預金保険（deposit insurance）」等の検索件数から予測する取り組みを，ドイツ連銀がおこなっている（Weber, et al. 2017）。我が国でも，総務省が大手通信キャリアの携帯の位置情報データを活用して，公的統計では把握できないテレワーク・デイの効果を検証した（総務省, 2017）。

我々のプロジェクトでも，公的統計では把握できない様々な属性の流動人口推計をおこなっている[1)]。SNSのフォロワー関係等から利用者の属性（AKB48アカウントのフォロワー＝AKB48ファン）を推計することが可能である。そして，位置情報付きのSNSからは20代30代の昼間の移動や居住エリアが推定できる。さらに，ウェアラブル活動量計からは，各地の利用者の睡眠率について日中サイクルが把握できる。居住エリアで睡眠を取ると仮定すると，これらのデータから特定の属性を持つ20代30代の流動人口が推定できる。図1は，秋葉原におけるAKB48ファンの流動人口を表す。AKB48ファンは平日の夕方に増加する特徴を持っていることがわかる。このように公的統計では把握できないニッチな需要にも，ビッグデータは応えることができる。例えば，震災後にPTSD症状が発現した（＝地震に関するストレス性の高い内容をSNSに投稿する）人々の人口分布とその時間変化を観測することも可能であろう。

5　ビッグデータによる公的統計の課題

統計を作る目的で集められてはいないビッグデータを公的統計に活用する上での課題を示す。ビッグデータは玉石混交のデータであり，構築される指数の信頼性の面からもノイズの除去は必須である。例えば，イタリア中央銀行は，失業率予測のための「jobs」の検索件数の利用において，「Steve Jobs」などを取り除いている（D’Amuri and Marcucci, 2012）。機械的に取り除く統一した手法を構築することが今後必要になってくる。民間各社では，データのコード体系を，その時々の自社のみのビジネスに最適化している。そのため，合併や業務提携後などでは関連会社間でデータの接続の不具合がしばしば発生する。公的統計として，各社のビッグデータを繋げるには，政府主導での共通コードの普及が必要である。例えば，2015年から始まった法人番号は，表記揺れの多い企業の名寄せに活躍している。公的統計では，過去と現在

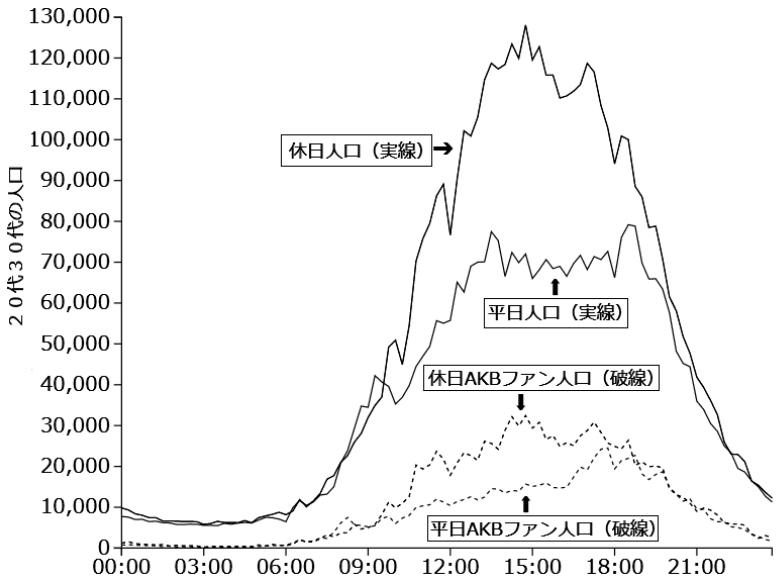


図1　秋葉原における20代30代の総人口とAKB48ファンの人口

との比較が可能であることを求められるために，膨大なデータの保持と同じ精度で継続的に実施することができるのかという問題を解決しなければならない（国土交通省都市局都市計画課都市計画調査室，2018）。そのためには，収集されたビッグデータの情報管理体制の整備や，特定企業，業界の枠を超えた社会インフラ・公共財としての価値観共有，情報提供者へのインセンティブ設計が必要である（NTTデータ経営研究所, 2015）。ビッグデータの分析の技術革新は日進月歩で進んでおり，公的統計の推計精度も日々向上している。推計精度が高いがブラックボックスと言われる深層学習でさえ，昨今は推計根拠の解釈可能性を上げる技術（Guidotti et al, 2018）が生まれており，将来的には統計発表に際して説明責任が果たせるようになると考えられる。技術が過渡期である現在，精度向上，説明責任面からも，ビッグデータ分析の公的統計活用に標準を合わせた人材育成が必要である。彼らが，新しい公的統計を担っていくことであろう。

【謝辞】

本稿は，総務省「ビッグデータ等の利活用推進に関する産官学協議のための連携会議」，内閣府大臣官房「公的統計におけるビッグデータの活用に関する調査研究」で得た公開情報などに基づいて水野の私見で作成している。本稿の成果の一部は科学研究費補助金17H05123, 17KT0034, 16H02872の支援を受けている。

注

1) 水野貴之, 2017,「人流ビッグデータによる地球規模の課題解決のための情報学と社会科学の融合基盤構築」JSTさきがけネットワーク’17採択課題。

文献

Antenucci, D., Cafarella, M., Levenstein, M., Re., C & Shapiro, M. ,2014, “Using Social Media to Measure Labor Market Flows”, *NBER Working Paper*, No. w20010.

Auction.com, 2014, “Auction.com Real Estate Nowcast”, FALL 2014 White Paper.

Bean, C., 2016, “Independent Review of UK Economic Statistics”, Cabinet Office and HM Treasury.

別所英実, 2018,「民間ビッグデータを統計として活用するためには，何が必要か：諸外国の取組事例の紹介と日本における課題の整理」『総務省統計委員会担当室ワークングペーパー』: 2018-WP01 (http://www.soumu.go.jp/main_content/000561917.pdf) (平成 30年 12月 21日閲覧)。

Cajner,T, et al.,2018, “Using Payroll Processor Microdata to Measure Aggregate Labor Market Activity”, *Finance and Economics Discussion Series*, Washington: Board of Governors of the Federal Reserve System, 2018-005.

Cavallo, A. and Rigobon, R., 2016, “The Billion Prices Project: Using Online Prices for Measurement and Research”, *NBER Working Paper*, 22111.

Council of Economic Advisors, 2012, “Economic Report of the President”, 187-188.

D’Amuri, F. and Marcucci, J., 2012, “The predictive power of Google searches in forecasting unemployment”, *Temi di discussione (Economic working papers), Bank of Italy, Economic Research and International Relations Area*, No. 891.

Farrell, D. and Greig, F., 2016, “Paychecks, paydays, and the online platform economy”, JP Morgan Chase Institute.

Guidotti, R. et al., 2018, “A Survey of Methods for Explaining Black Box Models”, *ACM Computing Surveys*, 51 (5): Article 93.

Hammer, C., et al., 2017, “Big Data: Potential, Challenges, and Statistical Implications”, *IMF Staff Discussion Note*.

Henderson, J. V., Storeygard, A. and Weil, D. N., 2012, “Measuring Economic Growth from Outer Space”, *American Economic Review*, 102 (2): 994-1028.

東大日次物価指数プロジェクト, http://www.cmdlab.co.jp/price_u-tokyo/ (平成30年12月21日閲覧)。

経済産業省, 2015,「地域経済分析システム (RESAS(リーサス))の提供を開始しました」(http://www.meti.go.jp/press/2015/04/20150421001/20150421001.html) (平成30年12月21日閲覧)。

国土交通省都市局都市計画課都市計画調査室, 2018,「総合都市交通体系調査におけるビッグデータ活用の手引き」(http://www.mlit.go.jp/common/001241230.pdf) (平成30年12月21日閲覧)。

Mandel, M. and Scherer, J., 2015, “A Low-Cost and Flexible Approach for Tracking Job and Economic Activity Related to Innovative Technologies”, *Nesta Working Paper*, 15/11.

NTTデータ経営研究所, 2015,「公的統計におけるビッグデータの活用に関する調査研究」平成26年度内閣府大臣官房統計委員会担当室請負調査. (http://www.soumu.go.jp/main_content/000422923.pdf) (平成 30年 12月 21日閲覧)。

Piet, J.H. Daas and Marco J.H. Puts, 2014, “Social media sentiment and consumer confidence”, *ECB Statistics Paper Series*, No. 5.

Reimsbach-Kounatze, C., 2015, “The Proliferation of Big data and Implications for Official Statistics and Statistical Agencies: A Preliminary Analysis”, *OECD Digital Economy Papers*, No.245.

Rothwell, J., 2012, “Education, Job Openings, and Unemployment in Metropolitan America”, Brookings Institution.

白木紀行, 松村浩平, 松本梓, 2013,「景気判断における検索データの利用可能性」日本銀行レポート・調査論文。(https://www.boj.or.jp/research/brp/ron_2013/ron130130a.htm/) (平成 30年 12月21日閲覧)。

総務省, 2017,「モバイルビッグデータを活用したテレワーク・デイの効果検証」(http://www.soumu.go.jp/menu_news/s-news/01ryutsu02_02000185.html) (平成30年12月21日閲覧)。

————, 2018,「公的統計の整備に関する基本的な計画」(http://www.soumu.go.jp/toukei_toukatsu/index/seido/12.htm) (平成30年12月21日閲覧)。

総務省情報通信国際戦略局情報通信経済室, 2014,「ビッグデータ時代における情報量の計測に係る調査研究報告書」(http://www.soumu.go.jp/johotsusintokei/linkdata/h26_05_houkoku.pdf) (平成30年12月21日閲覧)。

Struijs, P., 2016, “Big Data for official statistics”, Basque Statistics Office.

Swier, N., 2016, “WP1 - Webscraping for Job Vacancy Statistics”, *Eurostat Social Statistics Conference 2016*.

高田悠矢, 別所英実, 五十嵐盛仁, 2018,「労働市場の民間ビッグデータ:経済統計としての活用可能性」『総務省統計委員会担当室ワーキングペーパー』:2018-WP02(http://www.soumu.go.jp/main_content/000562980.pdf)(平成30年12月21日閲覧)。

United Nations Economic and Social Council, 2014, “Big data and modernization of statistical systems”, E/CN.3/2014/11.

Weber, P., Fecht, F. and Thum, S., 2017, “Capturing depositors’ expectations with Google data”, *IFC Bulletin*, Bank for International Settlements, No.44.

特集論文

6

ビッグデータ分析の実践にむけて

榊 剛史
株式会社ホットリンク　開発本部R&D部 部長

1 はじめに

「ビッグデータの時代」と言われ始めてから久しい。最近では,バズワードとしては,「AI（人工知能）」に取って代わられた感は強いが，学術界・産業界問わず,ビッグデータを活用することは単なる流行としての話題から，広く使われるアプローチとして定着してきたと言っても過言ではないだろう。このようにビッグデータ分析が普及した現在では，以前とは比較にならないくらい容易にビッグデータの分析を可能にするツール・環境が整いつつある。本特集ではここまでビッグデータの概要と用途にふれてきたが,本論文では,それらのツールを用いてビッグデータ分析を実践する方法について説明していきたい。

具体的には,PCとウェブブラウザのみでビッグデータ分析を実践する方法を紹介する。PCやプログラミング言語に詳しくない方でも極力わかりやすく書いたつもりだが，分かりにくい点もあるかと思うので，そこはご容赦いただきたい。

ところで,ビッグデータ分析というが，どのくらいの規模を超えるとビッグデータと呼ばれるのであろうか？このあたりは諸説あり，また個人個人によってもわかれるため，一概に定義することは難しい。ただ，著者の主観的な基準としては,だいたいサンプルサイズ（標本の大きさ）が数千くらいのオーダーまでは通常のデータ分析であり，それを超えて数万のオーダーになるとビッグデータであると考えている。もう少し大ざっぱに言えば,「Excelで扱えない規模のデータ」をビッグデータと考えて頂ければわかりやすいかも知れない。環境にもよるが，普段使うPCで数十万行のCSVファイルをExcelで開くのが難しいのは,容易にご想像いただけるだろう。

ビッグデータと同時期に出てきた概念として,「オープンデータ」というものがある。行政に関するデータや研究に用いるデータを誰もが利用可能な形で公開・提供することで,それらのデータの利活用により行政業務の効率化や科学技術の進歩を促進させるコンセプトである。「ビッグデータ」という言葉が使われる際に,「オープンデータ」の意味合いが含まれることも少なく無いので,そこは留意していただきたい。

2 ビッグデータ分析の流れ

全体の流れ

ここでは，まずビッグデータ分析の大まかな流れについて述べる。ビッグデータ分析の流れは大きく分けて下記の4段階（5段階）に分かれる。

0.（環境の構築）

データ分析ツールを利用できる環境を構築する（一度構築してしまえば，二度目以降は不要な作業であるため,括弧書きで記載している）。

1.データの取得・収集

自前での作成や企業からの提供，公開データの取得などにより必要とするデータを取得・収集する

2.データの加工・前処理

次にツールで分析可能な状態にするために，獲得したデータに加工や前処理を加える

3.分析

実際にツールを用いてデータの分析を行う

4.可視化

分析を通じて得られた結果を大勢が理解可能な形で可視化する

通常，研究やビジネスにおける「分析」と呼ばれるプロセスは３，４を指すことが多い。実際，読者の方にもそのようなイメージを持っておられる方が多いであろう。しかし，実際，ビッグデータ分析においては，１，２に分析にかかる時間の50%,多い場合では70%を占めることがある。わかりやすいように，Pythonで分析対象データが読み込まれた状態から主成分分析を実行するためのサンプルコードを提示する。

```
pca = PCA()
pca.fit(dataset)
```

ご覧の通り，わずか2行である。このように，RやPythonなど，現在主流で使われているビッグデータ分析ツールにおいては，主要な統計分析や機械学習の手法は1,2行で実行可能なものがほとんどである。可視化も同様である。

そこで，今回では３，４のみならず，０，１，２も含めてデータ分析の手順を解説する。ただし，１についてはアプローチがかなり多様であり，本誌面では十分な説明するのが難しいため，代表的なアプローチと参考資料の紹介にとどめる。

2-0. 環境の構築

環境の構築は，慣れない人にとってはある意味一番時間を要するところであり，かつつまずきやすいステップである。本記事ではPythonのJupyter Notebookを使うが，それ以外のツールについても紹介する。

・Python関連

–Python：

いわゆるスクリプト言語と呼ばれるプログラミング言語である。近年では，機械学習，とりわけ深層学習の開発するためのプログラミング言語としてはデファクトスタンダードとなっている。数値計算のためのライブラリも充実しており，プログラミングに慣れ親しんでいる人間が，統計分析・機械学習を行うには最初に選ぶべき選択肢であると言える。各自の環境にもよるが，かなり大規模なデータにも対応できる。ただし，分析環境の構築には，一定の知識が必要であり，導入の障壁が高い。なおデータ分析に必須なライブラリは下記の通りである。

- NumPy：数値計算用のライブラリ
- SciPy：科学計算用のライブラリ
- scikit-learn：統計分析・機械学習用のライブラリ
- pandas：データ分析支援機能のライブラリ

–Anaconda[1)]：

データ分析でよく使われるPythonライブラリをまとめてインストールしてくれるPython環境構築ツールである。自分のPCやサーバにPythonによるデータ分析環境を構築する際は，Anacondaをインストールするのが一番簡便な方法である。

–Jupyter Notebook：

「ノートブック」と呼ばれる形式で，ウェブブラウザ上でPythonのコード（プログラム）を対話的に実行できるツールである[2)]。Pythonでデータ分析を行う場合，開発したプログラムを毎回最初から最後まで実行する必要があるため，試行錯誤を繰り返していくデータ分析とは相性が悪いが，Jupyter Notebookを使う事で，この問題を解消することができる。可視化も容易に行うことができる。ただし，メモリを大量に使用する

ため,大規模なデータを扱うことは難しい。また,Jupyter Notebookを実行する環境を構築する必要がある。最も簡便な方法は前述のAnacondaをインストールすることである。

–クラウドNotebook環境：

Jupyter Notebookを実行できる環境を提供するクラウドサービス。本記事執筆現在では,Google Colaboratory[3]　と Microsoft Azure Notebooks[4]の２つが提供されている。環境構築が不要であり,すぐにデータ分析が始められる点が優れている。使用メモリやデータ保持期間,連続使用時間に制限などがある。

・R関連

–R言語：

統計解析のために開発されたプログラミング言語である。Pythonよりも統計分析のアルゴリズムが充実しており[5],また可視化に優れているのが特徴である。Windows/Mac/Linux共に環境構築が容易であり,取っつきやすい言語であると言える。ただし,数千万ケースを超えるような大規模データは扱うことが難しい場合もある。

–Exploratory[6]／EZR[7]：SPSSなどの従来の統計分析ツールのように,マウスによるユーザインターフェースを通して,Rを使ったデータ分析を行うことができるツール。筆者は使ったことがないため,概要の紹介にとどめるが,プログラミング言語に抵抗がある方でも,これらのツールを使うことで,従来の統計分析ツールと同様の簡易な操作でデータ分析を行うことができる。

本稿では,環境の構築しやすさから,ウェブブラウザとクラウドNotebook環境を用いて実践を行う。

なお,著者がよく行う分析の流れは,以下の通りである。まず,自前の環境にJupyter Notebook環境を構築し,その上で分析対象の一部を抽出した小規模データで試行錯誤して分析の流れを決める。その流れをコード化し,最後にそのコードをJupyter Notebookを使わずにPythonで実行する,という手順である。こうすることで,試行錯誤は対話的に行いつつ,最後に分析をデータ全体に高速に適用することができる。ただ,数十万～数百万レコード程度の分析であれば,Jupyter Notebookのみで十分に分析可能であると思われるため,本記事読者の方々にはそちらをお勧めする。

備考

環境構築には一定の知識が必要と述べたが,LinuxベースのMacではウェブ等の記事を参考にしながら比較的簡単に環境を構築することができる。一方,WindowsではPythonの実行環境をなかなか整備することが難しい。前述の通り,一番確実なのは,Anaondaをインストールすることである。また,PythonはPython2.XとPython3.Xで大きな違いがあるが,基本的にはPython3.Xの方を使っておく方が望ましい(Python2.X系の方にしか存在しないパッケージなどもあるが,主要な数値計算パッケージであるNumPy/SciPyがPython2.Xへの対応を2020年に終了する)

2-1. データの取得・収集

データ収集のステップでは,文字通り様々な情報源からデータを収集する。例えばウェブ上からデータをクロールする,企業からデータ全体の提供を受ける,などである。近年,研究におけるデータセットの取得方法は大きく分けて,下記の４つに分けられると考えられる。

・企業からのデータ提供
・公開データセットの取得
・公開情報の自動収集
・自身でのデータセットの構築

ただ,これらのアプローチについて詳細に説明をすると本記事では紙面が足りないため割愛する。ただし参考資料として東京大学鳥海准教授による資料「計算社会科学におけるWebマイニング」(鳥海,2018)を紹介しておく。ぜひこ

ちらの資料をご参照されたい（p.41以降がデータ取得に関する話である）。

また研究に活用可能なデータセットが公開されているウェブサイトを下記に紹介する。

・情報学研究データリポジトリ[8)]

国立情報学研究所が研究用に提供しているデータセット。民間企業が提供するデータが多いのが特徴。

・Stanford Network Analysis Project[9)]

Stanford大学の社会ネットワーク分析研究室によるデータセット。社会ネットワークのデータだけでなく，様々なウェブに関するデータが公開されている。

・Kaggle[10)]

最も著名なデータ分析コンペサービスによるデータセット。

・SIGNATE[11)]

様々なデータ分析コンペを開催しているSIGNATE社によるデータセット。

・ICWSM Dataset Sharing Service[12)]

ソーシャルメディアに関する国際会議ICWSMにより公開されている学術研究用データセット。

2-2. データの加工・前処理

データの加工・前処理のステップでは分析ツールにデータをインポートする前にそれに適した形にデータを整形するプロセスのことである。このステップは大きく3つに分けることができるが すべてを適用する必要はなく，手元にあるデータの状況にあわせて取捨選択していただきたい。

・読み込み可能なデータ形式への加工・変換
・欠損値／重複ケース／異常値／カテゴリカルデータの扱い・数値変換について
・ノイズとシグナルの分離

a.読み込み可能なデータ形式への変換

多くのデータビッグデータ分析ツールは，表形式でデータを読み込むこととなっている。表形式のデータとは分かりやすく言えばエクセルのデータのことである。ただしほとんどのツールではCSV(Comma Separated Values)やTSV(Tab Separated Values)のファイル形式でデータを読み込みに対応しているので，まずは手元のデータをこれらの形式に変換する必要がある。企業からデータ提供を受ける場合やオープンデータサイトからデータをダウンロードする場合など，データ保有者から直接データを手に入れる場合は，Excel/CSV/TSVなどの形式でデータを得られることが多い。一方で，ウェブ上からのデータの自動収集など，公開データを収集する場合はその限りではない。また，見落としがちな点であるが，日本語文字を扱う場合，文字コードをUTF-8に変換する必要があることに注意されたい。

b.欠損値／重複ケース／異常値／カテゴリカルデータの扱い・数値変換について

ビッグデータの場合，分析の目的のために集められたデータを使う場合よりも，別の目的で収集されたデータを分析者が使いたい目的に利用する場合が多い。このような場合往々にして，データが欠けていたり，余分なデータや異常なデータが含まれていたりする場合がある。そのため必要なデータを補完したり，不要なデータを除去してやる必要がある。

・欠損値への対応

欠損値がある場合は，データや分析目的によるが，欠損値を全体の平均値，前後の値の平均値で補完する，そのままサンプル自体を除去する，などの対応が考えられる。例えば，ある企業の株価や気温のような時系列データが1日だけ欠損していれば，欠けた時間帯の前後の平均値で補完することが望ましい。また，数十万アイテムに対する購買データで値が欠損しているものであれば，そのケースは除去（無視）してしまっても問題がないと考えられる。

・重複ケースへの対応

あきらかに同じであると分かるケースが2つ以上ある場合は，それを除去するのが一般的な対

応である。

・異常値への対応

異常値についても，欠損値と同様の対応が考えられる。明らかな異常値であれば無視すれば良いし，無視することが難しい順序性があるデータ（時系列データ等）であれば，前後の値の平均値で補完することが望ましい。

・変数のダミー化

多様な機械学習手法を適用するためには，数値ではないデータ（主に文字列データやカテゴリカルデータ）は，数字に変換して扱う必要がある。具体的には数字ではないデータを「0」と「1」だけの変数に変換する「ダミー化」を行うのが一般的である。例えば曜日データであれば，「月曜日を1，その他の曜日を0とした変数」「火曜日を1，その他の曜日を0とした変数」のように，7種類のダミー変数を作成する必要がある。

・数値変換

回帰分析や主成分分析において，各変数の係数を比較する場合には比較可能なように数値を変換する場合がある。代表的な変換方法としては，標準化(standardization)，正規化 (normalization) または最大最小スケーリング (min-max scaling) などがあげられる

c.ノイズとシグナルの分離

前述のように，ビッグデータの場合，分析の目的のために集められたデータを使う場合よりも，別の目的で収集されたデータを分析者が使いたい目的に利用する場合が多い。そのため，分析に必要なシグナル以外に様々なノイズが含まれてしまうことが多い。例えば，筆者の所属する会社ではソーシャルメディアデータ（Twitter/ブログ）を扱っているが，これらを用いて社会の動向を分析したい場合，機械で生成されたスパム投稿はノイズとして除去する必要がある。

このようにビッグデータ分析においては，必要に応じてシグナルとノイズを分離し，シグナルのみを選択的に抽出してやる必要がある。

2-3. 分析

最初に言及した通り，本記事はある程度，統計分析の知識を持つ方を対象としているため，分析自体については多くを割かない。基本的には，ビッグデータ分析においても用いる手法は，これまでの統計分析手法と大きくは変わらない。ただし，あまり計算量が大きい手法は現実的な時間で終わらないことがあるため，注意が必要である。実際には，まず小規模データで分析方法の策定と分析時間の測定を行い，策定した分析手法が速度的に大規模データに適用できるかを検証することが望ましい。

2-4. 可視化

ビッグデータ解析において，可視化は重要なプロセスである。なぜなら，データの規模が大きすぎるため，人間が全てのデータを確認することが困難だからである。だからこそ，データが持つ特徴や分析の結果を，わかりやすい形で可視化することが求められる。 幸い，データの性質によってどのような可視化を用いるべきかをまとめている人が多くいるため，それらの知見を参照されたい。「可視化 チートシート」「data visualization cheat sheet」などのキーワードでウェブ検索することをお薦めする。また，R -Studio が公開している可視化に関する文書に，目的別の可視化方法がよくまとまっている（R-Studio，2018）。

3 実データによるビッグデータ分析の実践

本節では，前節で紹介したビッグデータ分析の流れに基づいて，実際にビッグデータ分析を実践する。本節，特に３－１．以降では，教材ノートブックを傍らに見ながらお読み頂くことを想定している。また教材ノートブックのコードは上から順番に実行して頂くことを想定している（途中から実行するとうまく動作しない可能性がある）。

3-0. 環境構築と準備

a.利用するデータと分析

今回は,2章で紹介したSIGNATEのサイトから, 学習用として公開されているデータセット「【練習問題】国勢調査からの収入予測」を利用する。データ規模が1万6千サンプルと大きすぎないサイズであること, 他のデータと比べて社会調査寄りの内容であることが選定した理由である。(https://signate.jp/competitions/107)

今回は, 本データを用いて,「18歳〜65歳の就業者において, 収入に影響を与えている項目を明らかにする」ことを目的とした分析を行う。

b.環境構築の方法

本記事では, Google Colaboratoryを利用する方法と自身のPCにJupyter Notebook環境を構築する手法を紹介する。

・Google Colaboratory

Google Colaboratoryを利用する手順は下記の通り。Googleアカウントが無くても利用することはできるが, Googleアカウントと連携する方が便利である。

1. 下記のリンクにアクセスする
   https://colab.research.google.com/
2. 図1のように「Python3の新しいノートブック」をクリックする

・AnacondaとJupyterNotebook

自分のPCにJupyter Notebookの環境を構築するには, Anacondaをインストールするのが, 簡単かつ確実である。Windows/Mac/Linux用のインストーラーが公式サイトにて提供されている。インストール方法は公式ウェブサイトをご確認頂きたい[13]。ただし, すでにPythonを導入している場合, Pythonが競合し, うまく起動しない場合が多い。またJupyter Notebookを起動するには,一定のPC知識が必要である。詳しくは下記のページを参照されたい。

〈Windowsの場合〉
https://pythondatascience.plavox.info/pythonの開発環境/jupyter-notebookを使ってみよう。
〈Macの場合〉
https://qiita.com/heimaru1231/items/b71aa638e71a535cd790

c. 教材の利用方法

教材の利用方法について説明する。なお, Google ColaboratoryはChromeもしくはFirefoxでしか開けないことにご留意頂きたい。
教材ノートブックのページ:https://bit.ly/asr22_2019

・Google Colaboratoryを利用する場合
—ウェブブラウザで教材ノートブックにアクセスする
—図2のように「Playgroundで開く」をクリックする

図1　Google Colaboratoryの起動

図2　教材ノートブックへのアクセス

—教材ノートブックが実行可能になる
Googleアカウントがある場合は，図3のように「Google Driveに保存」を選択すると，自分でノートブックを編集することができる。

・ローカルPCのJupyter Notebookを利用する場合
—ウェブブラウザで教材ノートブックにアクセスする
—教材ノートブックのメニューから「.ipynbをダウンロード」をクリックしてノートブックをダウンロードする
—ローカルPCのJupyter Notebookでダウンロードしたノートブックを開く

・注意点
—前述のとおり，教材ノートブックのコードは上から順番に実行して頂くことを想定している（途中から実行するとうまく動作しない可能性がある）。
—Google Colaboratoryには下記のような時間的制限がある。時間的制限を超えた場合，変数の値などがリセットされるため，最初からコードを実行する必要がある
・新しいノートブックを起動してから12時間
・ノートブックのセッションが切れて[14]から90分

【用語と基本操作】

・ノートブック
Jupyter Notebookでのプログラムファイル。****.ipynbという名前で表記される。なおGoogle Colaboratoryと通常のJupyter Notebookではデザインが異なる（機能はほぼ同じである）

・セル
コード（プログラム）やメモを記述する単位。ブラウザで表示されているノートブックの実行ボタンをクリックすると，今選択されているセルのコードが実行される。

図4にGoogle Colaboratoryの画面の簡単な説明を示す。

3-1. データの取得・収集

前述したように，今回は公開されているデータセットを用いるため，取得・収集に特別な手間は不要である。下記URLから，train.tsvとtest.tsvをダウンロードすればよい。なお，データセットのダウンロードにはSIGNATEへのユーザ登録（無料）が必要となる。また今回，test.tsvは複数ファイルアップロードの例を示すためのみに用い，分析には用いない。
https://signate.jp/competitions/107/data

図3　教材ノートブックのダウンロード

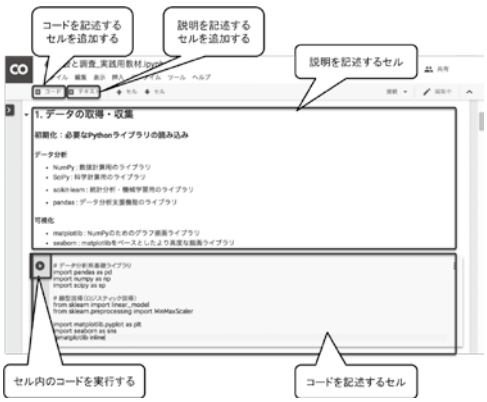


図4　Google Colaboratory画面の説明

3-2. データの加工・前処理

読み込み可能なデータ形式への加工・変換

「3-1. データの取得・収集」の手順と同様に，特別な手間は不要である。ただし，ノートブックへのファイルの読み込みがわかりにくいため，下記で説明する。

・Google Colaboratoryの場合

下記のコードを実行すると，教材ノートブックの説明にあるようにファイルをアップロードするボタンが表示される。ここからファイルを選択して，アップロードすれば良い。手間を避けるため，複数のファイルを選択し，アップロードすることが望ましい。また，Google Driveから読み込む方法もある。

・ローカルPCの場合

下記のコードのように，ローカルPC内のファイルの場所を設定すれば，ファイルを読み込むことができる。

```
from google.colab import files
list_files=files.upload()
traindata = pd.read_table("train.tsv")
testdata = pd.read_table("test.tsv")
```

pd.read_tableを実行することで，TSVファイルが，Python上にデータが「データフレーム」と呼ばれる形式で読み込まれる。データフレームについて詳細な説明は省くが，図５のように各行が一つのケース，各列が一つの変数を表す表形式のデータである。いわゆるExcelの表と同等のものと考えて頂いて差し支えない。Pythonの場合，通常，表形式のデータ（リスト）は全ての列は同じ種類の変数（整数なら整数のみ，文字列なら文字列のみ）でなければならないが，データフレームは，各列ごとに異なる変数の種類をとることができる。

なお，教材では，データの概要を表示する方法を紹介している。

重複ケースの除去

次に重複する行がある場合には，それをデータから除去する。教材ノートブックのコードを実行すれば，各列ごとに重複するケースがあるかどうかを表示できる。今回のデータについては，重複ケースは存在しないため，重複ケースの除去は行わない。

欠損値の対応

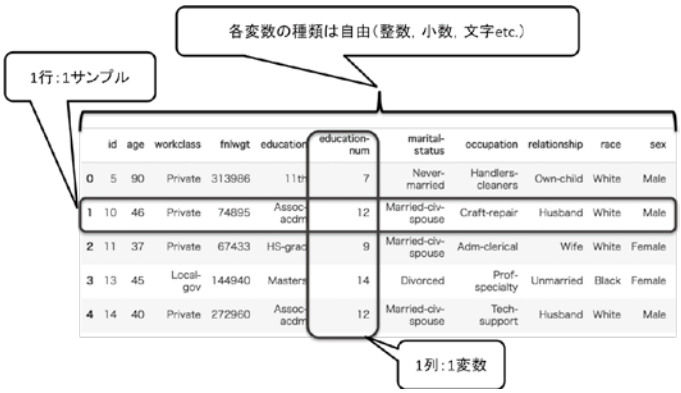
次に欠損値への対応を行う。欠損値への代表的な対処方法は下記の通り。

・リストワイズ法：欠損ケースを除去

・ペアワイズ法：相関係数など2変数を用いて計算を行う際に，対象の変数が 欠損している場合に計算対象から除外

・平均値代入法：欠損を持つ変数の平均値を補完

・回帰代入法：欠損を持つ変数の値を回帰式をもとに補完



	id	age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex
0	5	90	Private	313986	11th	7	Never-married	Handlers-cleaners	Own-child	White	Male
1	10	46	Private	74895	Assoc-acdm	12	Married-civ-spouse	Craft-repair	Husband	White	Male
2	11	37	Private	67433	HS-grad	9	Married-civ-spouse	Adm-clerical	Wife	White	Female
3	13	45	Local-gov	144940	Masters	14	Divorced	Prof-specialty	Unmarried	Black	Female
4	14	40	Private	272960	Assoc-acdm	12	Married-civ-spouse	Tech-support	Husband	White	Male

図５　データフレームについて

小規模なデータであれば，統計的な有意性を担保するために，サンプルサイズは減らさないように平均値代入法や回帰代入法で欠損値を補完することが望ましい。

しかし，ビッグデータのように，欠損値があるケースの数に対して全体のサンプルサイズが十分に大きければ，リストワイズ法を採用してしまうのが，一番簡便かつ妥当である。つまり，欠損値があるケースは無視してしまう，ということである。

教材ではリストワイズ法を採用している（ただし，結果的に欠損値がないデータであったため，ケースは削除されていない）。

ノイズとシグナルの分離

次にノイズとシグナルの分離を行う。ビックデータ分析の場合，分析の目的から考えて，明らかに対象外となる変数やケースをノイズとして除去する。逆に対象となることが明らかな情報，対象となるかが不明確な情報は除去しない。

今回の分析の目的は，「18歳～65歳の就業者において，収入に影響を与える要素を明らかにする」ことが目的であるため，教材は「age」が18～65の範囲にあるサンプルのみを抽出している。

数値変換

分析手順にあるように，構築した予測モデルにおける各説明変数の係数を比較するため，数値変数の変換を行う。ここでは，各変数が-1から1の範囲に収まるように最大最小スケーリング法を適用する。

分析と可視化

ここまでで分析の準備が完了である。今回は，二つの分析と一つの可視化を行う。被説明変数と説明変数について，「ロジスティック回帰」「ロジスティック曲線の可視化」「データの特徴を俯瞰するための可視化」を行う。

ロジスティック回帰

今回のデータでは，年収に関する変数は，年収が$50,000より高いか否かの二値で表現されている。そこで，二値分類の予測問題で良く用いられるロジスティック回帰を用いて，目的変数Y（年収が$50,000を超えていると1，$50,000以下だと0）についての予測モデルを構築した後，各説明変数の係数を比較することで，年収に影響を与える変数を明らかにする。ロジスティック回帰は線形回帰族の一つの手法であるため，係数を直接比較できることが特徴の一つとなる。学習した予測モデルの各係数上位10件は図6の通り。

各変数の係数を比較すると，「capital-gain」

```
### ロジスティック回帰の結果表示###############################
係数が正の説明変数を上位10件を表示
    Coefficients                    Name
3     13.131905             capital-gain
5      2.293780           hours-per-week
4      1.951208             capital-loss
0      1.707415                      age
2      1.541477            education-num
33     1.353812  marital-status_Married-civ-spouse
32     1.279285  marital-status_Married-AF-spouse
29     1.116407     education_Prof-school
25     1.108882      education_Doctorate
1      0.904625                   fnlwgt
係数が負の説明変数を下位10件を表示
    Coefficients                    Name
64    -1.832960              sex_Female
21    -1.303655           education_9th
35    -1.250064  marital-status_Never-married
62    -1.206949              race_Other
55    -1.205751  relationship_Other-relative
56    -1.145008    relationship_Own-child
100   -1.131936     native-country_South
37    -1.131532    marital-status_Widowed
36    -1.118419  marital-status_Separated
46    -1.086645  occupation_Other-service
```

図6　予測モデルの係数上位10件

「hours-per-week」「capital-loss」「age」「education-num」が,目的変数に対して相対的に大きな正の影響を及ぼしていることがわかる。また,「sex_Female」「education_9th」「marital-status_Never-married」が目的変数に対して相対的に大きな負の影響を及ぼしていることがわかる。(「race_Other」「relationship_Other-relative」は「その他」にあたる項目のため,ここでは省略する)

ロジスティック曲線の可視化

次に,得られた予測モデルから,正の影響を及ぼしていた5つの変数「capital-gain」「hours-per-week」「capital-loss」「age」「education-num」を合成した変数を用いて,ロジスティック曲線を可視化する。結果は図7のようになる。境界付近では,うまく分類できていないサンプルもあるが,Y=0,1について,大まかにはうまく分類できていることがわかる。

データの性質を俯瞰するための可視化

付加的ではあるが,データの性質を俯瞰するために,データセットの数値変数について,図8,9のように相関行列の表示およびヒストグラム・散布図の混合グラフの可視化を行った。各ケースの色は,「薄い色の●」が年収$50,000以下,「濃い色の＋」が年収$50,000超を表している。

相関行列および散布図から,数値型の説明変数同士にはほとんど相関が無いことがわかる。またヒストグラムと散布図から,「hours-per-week」「age」「education-num」は,年収の二値分類には,全体的になだらかな相関性があることが推測される。一方,「capital-loss」「capital-gain」については,少数サンプルの影響により係数が大きくなっていることが推測される。

4 おわりに

本稿では,実行可能な教材を用いて,ウェブブラウザ環境によるビッグデータ分析の流れについて説明した。ごく簡単な事例ではあったが,皆様の想像以上に簡単にビッグデータ分析を実行できることをご理解頂けたのではないかと思う。幸い,当該分野については,書籍もさることながらウェブ上にも大量の解説記事が掲載されており,またMOOC(大規模オンライン講義)なども充実している。そのため,きっかけと時間さえあれば個人でもビッグデータ分析の学習を進めることができる。本記事が,そのようなきっかけの一助となれば幸いである。

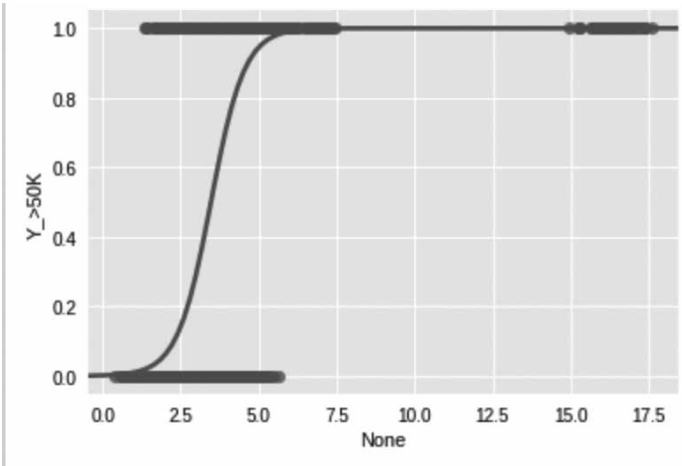


図7　ロジスティック曲線の可視化

	age	fnlwgt	education-num	capital-loss	capital-gain	hours-per-week
age	1.000000	-0.071269	0.039734	0.061454	0.083618	0.154290
fnlwgt	-0.071269	1.000000	-0.054013	-0.004881	0.001942	-0.026091
education-num	0.039734	-0.054013	1.000000	0.081099	0.127128	0.124606
capital-loss	0.061454	-0.004881	0.081099	1.000000	-0.031710	0.059918
capital-gain	0.083618	0.001942	0.127128	-0.031710	1.000000	0.083786
hours-per-week	0.154290	-0.026091	0.124606	0.059918	0.083786	1.000000

図 8　各変数間の相関行列

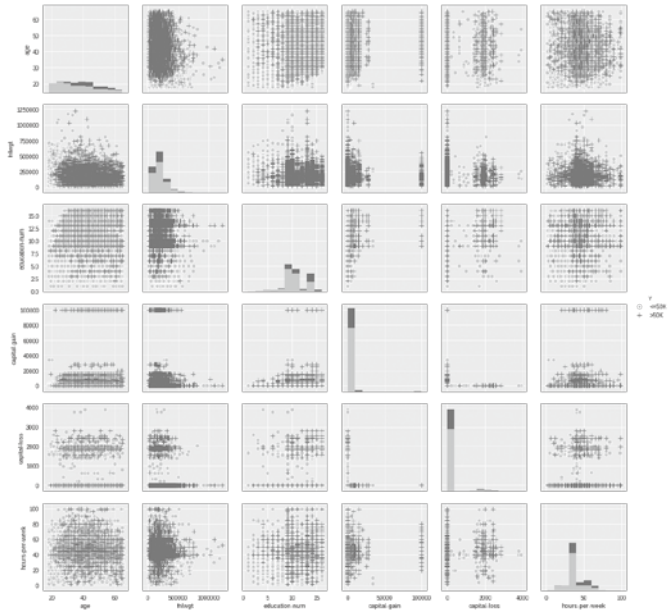


図 9　各変数間の混合グラフ

注

1) https://www.python.jp/install/anaconda/index.html

2) 以前はiPython Notebookと呼ばれていた。

3) https://colab.research.google.com/

4) https://notebooks.azure.com/

5) 2018年11月現在。

6) https://ja.exploratory.io

7) http://www.jichi.ac.jp/saitama-sct/SaitamaHP.files/statmed.html

8) https://www.nii.ac.jp/dsc/idr/index.html

9) http://snap.stanford.edu/data/index.html

10) https://www.kaggle.com/datasets

11) https://signate.jp/competitions

12) https://www.icwsm.org/2018/datasets/datasets/

13) https://www.anaconda.com/

14) ブラウザを閉じたり,インターネット接続が切れた場合に,セッションが切れる。

文献

R-Studio, 2018, “Data Visualization with ggplot2 Cheat Sheet.” (https://www.rstudio.com/wp-content/uploads/2015/04/ggplot2-cheatsheet.pdf)

鳥海不二夫, 2018,「計算社会科学におけるWebマイニング」(https://www.slideshare.net/toritorix/web-100905271)