データサイエンスと公的統計
ミクロデータ利活用

「官民オープンデータ利活用の動向及び人材育成の取組」
共同研究集会

2021年11月18日

北村行伸
立正大学データサイエンス学部

データの時代(1)

- 21世紀はデータの時代と言われています。確かに、インターネットやスマートフォン上を行き交う情報量は膨大な規模に上っています。
- そのデータの時代を担う学問として期待されているのがデータサイエンスです。データサイエンスとは、様々な種類のデータを蓄積し、分析することで、データから問題を発見し、データを用いて、問題を解決していくことを目指す学問です。
- では、そのデータとはどのようなものでしょうか？今集められている多くのデータはデジタル信号化されており、電波に乗って1日24時間365日世界中の空間を飛び回っているものです。これを、特定のデータベースに格納して、解析に資する形に加工し、それを統計処理、画像処理、音声処理することがデータサイエンスの仕事になります。

データの時代(2)

- スマホやインターネットで行き交うデジタル化されたデータ量は、指数関数的に増加してきており、人類が過去4000年ぐらいの間に蓄積してきた情報量に匹敵するデータがほんの数か月で蓄積されるようになり、そのスピードはさらに加速しています。言うまでもありませんが、これらのソーシャル(デジタル)データには、個人の位置情報やSNSのつぶやき、電力使用量やネット取引で購入した服のサイズまで含まれています。
- ワイガンド(2017, pp.14-15)によれば、ソーシャル・データは18か月ごとに倍増しています。2000年に丸1年かけて生み出されたデータと同じ量が2017年には1日で生み出され、2020年には1時間以内で生み出されるようになったということです。

データの時代(3)

- こうやって生み出されたソーシャル・データと人類が過去4000年間に貴重な資料として保管してきたものを単純にバイト数で比較していいものでしょうか。
- もちろん、歴史の評価を経て厳選されて残ってきた古典や文化遺産と、インターネット上でのつぶやきやそれほど重要ではない個人の記録等の価値は比較にならないことは確かです。
- しかし、指数関数的に増加するソーシャル・データの中から、価値のある情報とそれ以外の情報(フェイクニュースを含む)を選別することは、それほど簡単なことではありません。これこそがデータサイエンスに課された最大の課題なのです。

データの時代(4)

- また、世の中ではデータサイエンスにおけるデータを石油や金の採掘に例えて、何か特別に価値のある資源を見つける学問であるかのように論じる向きもありますが、データに含まれる情報は、特にそれ以前の人間行動や企業行動から変わった訳ではありません。ただ、これまでは記録されていなかった情報もデータとして蓄積されるようになったというだけで、データの質や内容が格段に向上したというものではないのです。
- 芸術や音楽、学問の世界で、本質的な発見が相次いだり、飛躍的な発想の転換が行われている訳でもありません。
- どちらかと言えば、より多くのノイズあるいはゴミの増加の中から、少しでも使えそうな情報を獲得すべく努力しているというのが適切な実態のように思います(人工知能の代替)。

データの時代(5)

- とは言え、経済活動の中で、データの利活用を軸にしたグーグル、アップル、フェイスブック、アマゾン(GAFA)などの、いわゆるプラットフォーム・ビジネスが世界経済を席巻していることも事実です。
- 世界の経済に占めるGAFA＋M(マイクロソフト)の影響力は、中規模の先進国よりはるかに大きくなっています。
- これらの企業は私が言うようにゴミの中から再生資源を探してきた訳ではありません。むしろ、より積極的に、消費者・利用者の情報を組織的に紐づけして集め、それをマーケッティングやマッチングに利用してきたのです。今我々が出来ることは、日々蓄積されるソーシャル・データあるいはビッグデータによって問題を発見し、その問題を理解することであるとGAFAは教えてくれているのです。

データの時代(6)

- グーグルやアマゾンなどのプラットフォーマーがやっていることは、広告の打ち方や消費者の需要喚起の方法に関して、様々な試行錯誤を繰り返しながら、新たなデータを得て、それを解析することで、消費者行動を根本的に見直しているということです。
- 一流のプラットフォーマーが付加価値を生み出しているのは、ただ単にデータを収集することではなく、顧客を取り込み、その顧客が、プラットフォーマーが提供するインセンティブや仕掛けにどのように反応するかをよく見定めて、広告をタイミングよく出すことで、売り上げを伸ばすという一連の試行の結果として理解すべきです。(社会実験を行っている)

データの時代(7)

- そのための多角的な研究開発の取り組みは、一流大学や高等研究機関を凌駕するものであり、データサイエンスの最先端の研究者・技術者がそこに集められているといっても過言ではありません。
- よく日本のマスコミが、日本でもプラットフォーマーを輩出しなければならないという論陣を張ることがありますが、日本で世界に通用するデータサイエンティストがどれだけ育っているのかをご存じでしょうか？失われた30年の遠因
- 政府・民間企業・大学すべてにおいて、データサイエンティスト人材が不足しており、その育成が急務となっているところです。多くの大学がデータサイエンス系の大学院・学部・学科を創設しているところです。これらの大学間での連携も重要だと認識しております。

データサイエンスの歴史的位置づけ(1)

- 100年に1回程度起こる、社会経済全体を変革させるような動きを産業革命と呼び、現在進行中のデジタル化・AIを軸としたものを第4次産業革命(個人的には20世紀末から始まった情報革命の延長として捉えれば、第3次産業革命期と呼んでもいいと思う)と呼んでいます。
- データサイエンスが興隆している背景には、現在進行中の産業革命があり、それへの対応として、デジタル・プラットフォーマーの出現があり、デジタル・トランスフォーメーション(DX)が求められているということです。
- 従来の産業革命でも、その時点でのプラットフォーマーやトランスフォーメーションが求められていました。それぞれの産業革命のコア技術をより早く、より効率的に取り入れた企業がその産業革命の中心的担い手になってきました。逆に、新技術の取り入れに遅れた企業は、市場からの退出を余儀なくされたのです。

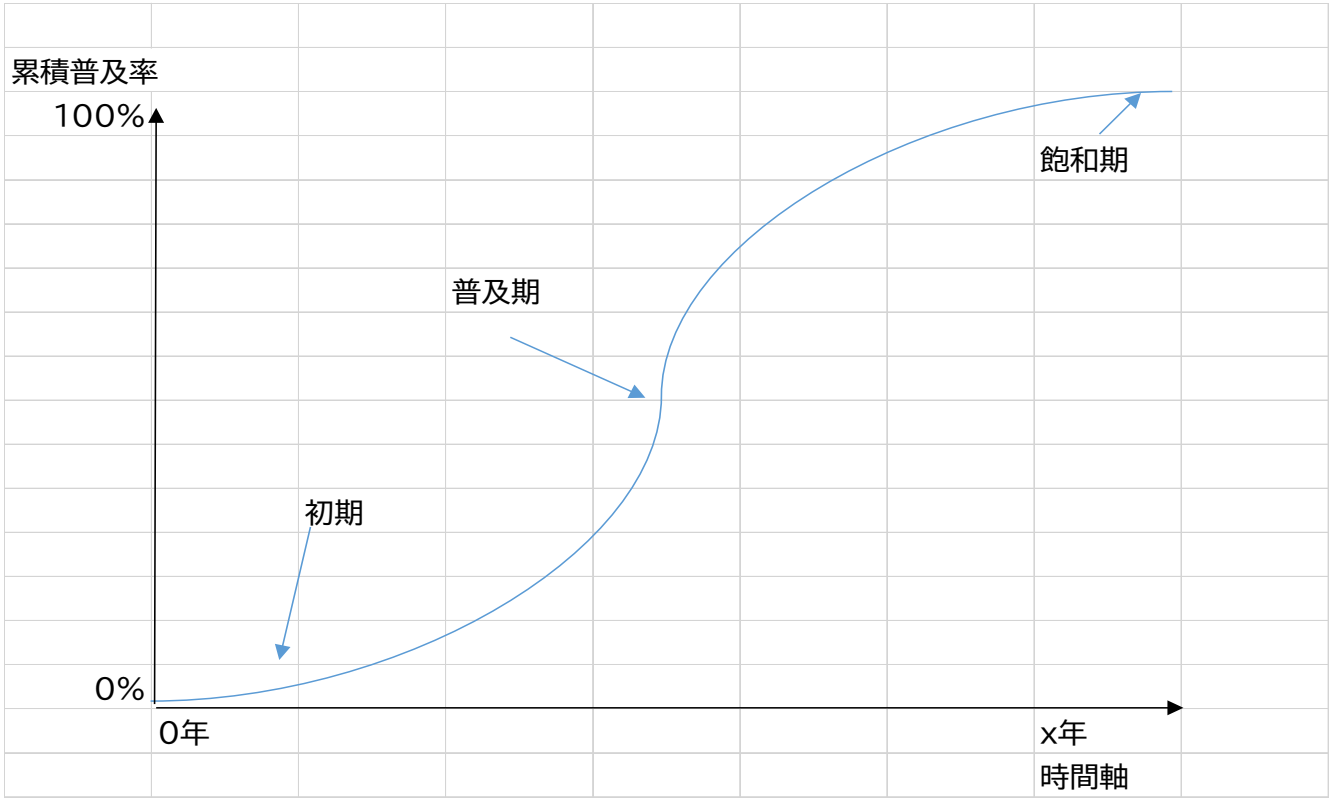
データサイエンスの歴史的位置づけ(2)

表1　産業革命の推移

	期間	主たる内容	主要国
第1次産業革命	18世紀末－19世紀前半	蒸気機関・繊維産業	イギリス・ドイツ・フランス
第2次産業革命	19世紀末－20世紀前半	電力・内燃機関・鉄道・電信電話・自動車	アメリカ・イギリス・ドイツ
第3次産業革命	20世紀後半ー21世紀前半	コンピュータ・インターネットなどの情報機器	アメリカ・日本・中国
第4次産業革命	21世紀前半	デジタル化・AI	アメリカ・中国・インド

データサイエンスの歴史的位置づけ(3)

図1　新技術・新商品の普及パターン



データサイエンスの歴史的位置づけ(4)

- 図1の新技術・新商品の普及パターンは、どの技術、どの商品でも似たような形状(ロジスティック曲線で近似)を示すが、その普及期に達するまでのスピードが近年早くなってきています。例えば、電気が解明されたのは1827年のオームの法則に始まり、1831年のファラデーの電磁誘導現象の発見、1873年のマックスウェルの『電気と磁気』の発表、エジソンの白熱電球の発明があった1879年以後、電気の応用技術は開発され続けたのですが、1900年でも電気モーターで動くアメリカの工場は5%以下でした。アメリカで電化された工場が本格的に稼働し始めるのは1920年代以後でした。エジソンの発明から40年を経て、初めて電気という新たな技術に合わせて工場の統治のあり方を変え、ビジネスモデルを変化させることで電気の真価が発揮されるようになりました。

データサイエンスの歴史的位置づけ(5)

- 電気の普及に比べれば、自動車の普及、テレビの普及はもっと速いスピードで起こりました。パソコンや携帯電話の普及はさらに早く、FacebookやLineは、それよりもさらに早く広がっていきました。
- データサイエンスが直面している、技術進歩のスピードとそれを応用したアプリケーションの普及は、はるかに急速であり、それを供給し続ける企業サイドは相当な人的・物的資源を投入しなければならないというのが実態です。
- では、このデータの時代のデータサイエンスの応用は、これまでよりスムースにしかも高速に実装されていくでしょうか。恐らく、データの利用可能性とデータサイエンス人材の適切な供給が鍵を握るのだと思います。

データサイエンスと公的統計(1)

- データの時代のデータとは、先にお話ししたようにデジタル化され、電波として発信されたソーシャル・データが中心になります。しかし、それらのデータを統計学的に把握するためには、従来の標本理論からの知見を利用せざるを得ません。
- とりわけ社会の母集団情報について知りたいときには、民間のソーシャル・データではなく、政府が定期的に調査収集している国勢調査や経済センサスなどの公的統計を利用するしかありません。言うまでもありませんが、公的統計は、その調査にかける資金や人材が民間企業や大学で行う調査とは比較にならないほど大きく、歴史的な継続性もあります。そういう意味では、公的統計はデータサイエンスの時代にも統計の重要なインフラストラクチャーでありつづけるものです。

データサイエンスと公的統計(2)

- 具体的に民間が集めたソーシャル・データについている個人・家計属性と国勢調査などの母集団情報を重ね合わせることで、さまざまなデータが、より重層的に分析可能になり、また、ソーシャル・データの持っている標本バイアスも明らかになってきます。
- 公的統計の関係者は、やはり公的統計が統計の主流であって、民間ソーシャル・データは補完的なものであると考えがちです。ただ、公的統計の調査方法自体が、時代の変化の中で見直しを迫られていることも事実です。例えば、統計調査の回収率(とりわけ、特定属性のサンプル捕捉の困難)の低さや回答者負担意識の高まりなどは無視できない問題です。

データサイエンスと公的統計(3)

- また、少子高齢化社会の中で、統計調査に対して十分な予算・資源が配分されにくくなってきていることも事実です。これらの問題に対する一つの解決策は、行政記録情報の積極的な統計への利活用です。政府は行政上の理由から、国民・企業に多くの個人情報の届け出を義務づけ、資金の給付・納付を受付、受給資格者に各種の社会給付を行っています。これらの情報を統計情報として利用すれば、統計調査を別途行わずとも、必要な情報がかなりの程度集められることが知られています。
- 現在は、まだ、行政記録情報を統計として利用するということが全面的には行われていませんが、徐々にその道を開いていくことが期待されます。その際、データベースの構築方法、データの保存方法などに関してデータサイエンスの最新の知見(暗号化や可視化)が生かされることが期待されています。

データサイエンスと公的統計(4)

- 公的統計の利活用を進めて官民の活動を活発化するということは当然のことですが、民間データのオープン化も重要な課題です。実際に、公的統計でも、スーパーマーケットや家電量販店のPOSデータを活用して、情報を補完しており、その方向での民間データの利活用は徐々に進んでいます。
- 民間データの強みは、IoTを利用した、人流・物流などのリアルタイムの活動が把握可能になるということにあります。公的統計が構造的な情報を把握するのに長けているのに対して、コロナ禍での人流データなどのフローの統計把握には民間データが圧倒的な強みを発揮しました。
- このような民間データサイドからも、本当に企業のコア・ビジネスになるデータ以外のもの(過去のコア・データも含む)については社会に対してデータをオープン化し、データインフラを構築していっていただければと思います。

データサイエンスの近未来

- データサイエンスは現在進行中の産業革命を推進していくうえで、中核となる技術や思想を提供する学問です。
- ただ、元になっている学問が統計学や数学、情報工学や社会科学と広範に広がっていて、これまでそれらを総合的に教えるデータサイエンス学部のような場所がありませんでした。
- 日本はこれから急速にこの分野の研究やアプローチにキャッチアップし、有為な人材を社会に供給していく必要があります。
- 数学や工学など日本の比較優位のある分野を中心に、データサイエンスの標準化は進んでいくでしょう。広い意味での人類社会への貢献（より良き社会の建設）という観点から、我が国のデータサイエンスを育てていきたいと考えています。

参考文献(1)

- 馬田隆明(2021)『未来を実装する』、英治出版
- 北村行伸(2021)「データと経済学の近未来像」、『経済セミナー』2021年4・5月号、No.719., pp.22-27.
- クスマノ、マイケル・A他(2020)『プラットフォームビジネス』、有斐閣
- グレアム、ポール(2005)『ハッカーと画家』、オーム社
- 此本臣吾(監修)(2018)『デジタル資本主義』、東洋経済新報社
- フレイ、カール・B(2020)『テクノロジーの世界経済史』、日経BP社
- 國部克彦・玉置久・菊池誠(編)(2021)『価値創造の考え方』、日本評論社
- ワイガンド・アレックス(2017)『アマゾノミクス　データ・サイエンティストはこう考える』、文藝春秋社