ビッグデータで何が変わるか？

経済研究部　研究員　高山 武士
takayama@nli-research.co.jp

ここ数年で、情報を得るためにウェブサイトを検索する機会が増えた。エコノミストとして電子データを解析する機会も多い。最近は「ビッグデータ」という言葉も良く聞くようになり、改めて大量の情報に囲まれていることを認識させられる。

ビッグデータという言葉には明確な定義はないが、「３Ｖ」によって特徴付けされることが多い。これは、「量(Volume)」「多様性(Variety)」「速度(Velocity)」の頭文字で、つまり、「高速で多様かつ大量に生み出される」データのことである。グーグルの検索欄には、この瞬間も多くの人が様々な言葉を入力しているだろうし、ツイッター上には、様々な内容のツイート（つぶやき）が投稿されている。ツイートは世界の多様な言語でなされ、画像がアップロードされることもある。まさに、高速で多様かつ大量のデータ（ここでは検索ワードやツイート）が生み出されており、３Ｖのイメージに合致する。大規模なだけでなく、多様であり高頻度で生み出されていることも大きな特徴だ。スマートフォンから取得される位置情報データや、防犯カメラで記録される人間の顔（表情）や動きのデータもこうした３Ｖデータの代表と言える。

（今までの統計学とビッグデータ）

ビッグデータという言葉は良く聞くようになったが、データをどこからがビッグか明確に区切ることはできない。突如として、画期的な分析手法が生まれた訳でもない。そのため、言葉だけが流行っている感じもするが、今まで非現実的（非常識）と考えられていた分析手法に脚光が当たっているという側面も確かにあるだろう。

ビクター・マイヤー＝ショーンベルガー、ケネス・クキエは、ビッグデータ分析が進んだことで、①できるだけ全データを使う、②精度は重要ではなくなる、③因果関係ではなく相関関係が重視される、といった変化（いわば、パラダイム・シフト）が起きていると指摘する[1]。

筆者は、「法則を確かめる道具」としてのデータ分析から「法則を発見する道具」としてのデータ分析に（再び）注目が集まっているように感じる[2]。上記の③に近いが、因果関係と相関関係の違いというよりは、データを分析する前に、あらかじめ強い仮説（仮定）を設けているか、そうでないかという違いが大きいように思う。

具体的に言えば、前者には、「仮説検定」と呼ばれる伝統的な統計の手法がある。これは、読んで字のごとく「仮説」が正しいか調べるために、データを使って検定（検証）する方法である[3]。例えば、「ある書店がキャンペーンをしたとき売上がどれだけ増えるのか(あるいは、キャンペーンは売上増に貢献するのか)」という事柄を確かめるために、一定期間の売上データを収集して検証するような例が挙げられる。この例では「キャンペーン（原因）⇒売上増（結果）」という仮説（因果関係の想定）があって、それをデータによって確かめている。経済学の分野では、経済理論を、現実のデータを用いて検証する実証分析と呼ばれる研究もさかんである。

一方、あらかじめ強い仮説（仮定）を設けない後者のアプローチは、「データマイニング」と呼ばれる。こちらはデータからお宝となる事実を「発掘（マイニング：mining)」する作業である。このアプローチは、とにかく沢山のモノの中から関係があるものを手当たり次第に探す。その結果、「理由は分からないけど、紙おむつとお酒が一緒に買われることが多いことが判明した」などという事実が発掘される。もちろん、結果的に因果関係などなんらかの関係が推測できることもあるが、とにかく沢山のモノの中からなんらか関係があるものを手当たり次第に探す点に特徴があると言える。

感覚的に言えば、同じツイートと株価の関係を調べるにしても、「ポジティブ（楽観的）なツイートが多いときには株価が上がるだろう」と仮説を設定してデータを調査するのは仮説検定、「株価が上がるとき

---

[1] ビクター・マイヤー＝ショーンベルガー、ケネス・クキエ、斎藤栄一郎訳「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）

[2] 「再び」と書いたのは、昔からどちらの側面もあるためである（昔も、データの集計や単純化で新しい知見を得ていた)。ただ、ビッグデータが扱えるようになって、違う側面から新しい発見ができる可能性が広がり、再び「法則を発見するための道具」として注目を浴びているのだろう。

[3] サンプル（標本）から全体（母集団）の性質を推定する方法（統計学的推定）を用いて検証している。仮説があれば、それを検証するためのランダム（無作為）なサンプル（標本）を取り出すことができ、そこから母集団の性質を知ることができることがポイントである。

に投稿されているツイートの特徴はなんだろうかと内容、言語、単語数、地域など様々なことを調べていく」方法はデータマイニングと言えるだろう。

前者のようにデータを「仮説を検証する」ために使うならば、母集団が大きくてもビッグデータを使う必要はない[4]。ただ、データを「法則を発見する」ために使うならば、ビッグデータはできるだけ大きい方が良い。発掘しがいがあり、思い至らなかった知見が得られる期待も高まる[5]。

これまで、こうした「法則の発見」は、熟練労働者の経験などに頼ってきた部分が大きい。しかし、これらの知識や法則の発見というプロセスに、データの力が役立つようになっている。実際、検索ワード、ツイート、位置情報データ、顔や動きのデータなどから、お宝を発掘する動きは進んでいる。

（人間の役割とコンピュータの役割）

もうひとつ注目すべきは、データを解析するコンピュータやアルゴリズムの発達である。処理能力の向上だけでなく、機械学習（ベイズ推定）など、３Ｖのデータと相性が良い分析手法・アルゴリズムが発達し、コンピュータが代替できる領域が広がっている。情報の「蓄積や集計」は機械が最も得意とするところだが、それだけではない。

例えば、検索サイトでは、大量のウェブサイトの中から適切と思われるサイトを上位に表示してくれる。この場合、情報の「取捨選択」はほとんど機械がしてくれている。オンラインショッピングで人の購買履歴を学習し、その人が好みそうな商品をオススメすることなども機械が行うようになった。また、受け取ったメールを迷惑メールか否かの判断も機械がしてくれる。これらは、情報の「解釈」を機械がしてくれる例と言えるだろう。迷惑メールか否かの分類は、マークシート方式のテストを正解と不正解に分類するのとは違って、明確な分類基準を設けることは難しい。しかし、人間がメールの内容を読み「これは迷惑メールだろう」と判断するように、機械が迷惑メールとはどういうものかを学習し、分類してくれるのである。

こうした機械による判断は、多くのデータ（購買履歴やメール）を学習させればさせるほど性能が上がる。将棋ソフトは、まだ短時間で

[4] 例えば、西内啓「統計学が最強の学問である」（ダイヤモンド社）を参照。
[5] 仮説検定では、仮説が「間違っている」（棄却される）か「そうとは言い切れないか」のどちらかの結論が得られるのに対し、データマイニングでは必ずしもお宝となる関係が見つかるとは限らない。そのため、サンプルを用いないでできるだけ３Ｖのデータをつかうデータマイニングは発掘コストが高く、費用対効果が悪くなりやすい。しかし、現在はこの発掘コストの低下が進んでいる。

の全数探索は出来ないけれど、大量の局面（良い手・悪い手）を学習することで、どの手が最も良い手かを判断しており、最近は、将棋ソフトがプロ棋士に勝利できるほどの実力をつけていることで話題にもなった。道徳的な議論を棚上げすれば、人事評価や就職（採用）などの相性判断はすぐに機械化できるだろう（男女の出会いも…）。

これまで、コンピュータはデータを「集計や蓄積」することが主流だったけれど、いまは、コンピュータがデータから「判断や評価」を出来るようになっている。それだけ、人間の行ってきた仕事を機械にまかせられるようになったとも言える。

今後、ビッグデータの流行を経て、さらに機械の存在感は大きくなるだろう。それだけ人間の役割は少なくなるかもしれない。

ただ、それは人間がより「人間らしい」仕事に注力できる時代なのかもしれない。そういった「人間らしさ」とはなんだろうかと考えつつ、ビッグデータの今後に注目している。

高山　武士

たかやま たけし
ニッセイ基礎研究所
経済研究部　研究員

●
東京工業大学理学部卒。06 年日本生命保険相互会社入社。日本経済研究センターへの派遣、米国カンファレンスボードへの派遣を経て、11 年ニッセイ基礎研究所入社。