ビッグデータ社会の到来

小林 啓倫

Agile Media Network

「ビッグデータ(Big Data)」が私たちの生活をどう変えるのか？
世界的に活用が始まった「ビッグデータ」の可能性がわかる。

悲しいことに、最近では無意味な情報というものがすっかり影をひそめてしまった。

――オスカー・ワイルド

はじめに

2009年7月。ミニブログという新たな世界を拓いたツイッター（Twitter）社から社内文書が流出し、複数のメディアに送られるという事件が発生した。その際に公開された文書の中には、次のようなビジョンが掲げられていた。

10億人のユーザーがいれば、ツイッターは地球の鼓動になる。

この言葉を聞いて、「何を大それたことを」と感じた人は少なくないだろう。

確かにユーザー数が10億人にまで達すれば、ビジネス的には無視出来ない存在になる。しかしツイッター上で交わされているメッセージといえば、今日のお昼や晩ご飯の話、今から観ようとしている映画や、買ったばかりの服の話など、他愛もないものばかりだ。それが地球の鼓動を目指すなど、馬鹿げた話にも程がある――

それから2年。残念ながらツイッターのアクティブユーザー数は1億人にとどまっているが、ユーザー数8億人を誇るフェイスブック（Facebook）など、ソーシャルメディア全体で見れば「10億」という数字も荒唐無稽なものではなくなった。その結果、何が起きているだろうか？

2010年10月、米国で驚くべき研究論文が発表された。インディアナ大学のヨハン・ボーレン准教授らによると、ツイッター上の書き込み（ツイート）を分析して予測システムに組み込むことで、将来の株価を86.7パーセントの精度で予想することができるというのである。

彼らは2008年2月から12月までの期間、270万人のユーザーによって投稿された980万件のツイートを収集し、さらにその中から感情を吐露しているものを抽出して分析を行った。その結果、「平穏」の感情を示す尺度が3～4日後のダウ・ジョーンズ工業株価平均の動きと近いことを発見し、株価データから株式市場予測を行う機械学習アルゴリズムに組み込んだところ、先に述べた精度を達成することに成功した。

研究の発表後、ボーレン准教授のもとに問い合わせが殺到。実際にツイッター等のソーシャルメディアの動きを株価予測に役立てるヘッジファンドまで登場している。

いまや予測されようとしているのは株価に留まらない。映画の興行成績や選挙での投票行動、さらに国民の健康状況に至るまで、様々な分野でツイート分析を行おうという動きが出ている。確かにソーシャルメディア上には、今夜食べたラーメンの話など、一見すると無意味なデータが少なくない。しかし無数のデータを集め、その全体像を捉えることによって、まさしく「地球の鼓動」が見出されようとしているのである。

これまでには取り扱うのが困難だった、非常に大容量のデータを分析し、その中に隠されている「情報」を見出す――それが「ビッグデータ（Big Data）」の発想だ。IT関係者の間では数年前から使われるようになっていた言葉だが、2011年10月に米ガートナーが発表した「2012年の戦略的テクノロジー（これから企業に大きな影響を与える可能性を持つ技術）トップ10」にも含められるなど、ビジネス界からも急速に注目される存在になってきている。

また2011年8月には、HP社がエンタープライズ検索技術などを提供するオートノミー社を103億ドルで買収。同じく10月には、オラクル社がビジネス・インテリジェンス用ソフトを提供するエンデカ社を7.5億ドル（推定）で買収するなど、ビッグデータに関連する企業買収の動きも活発になってきているところだ。

確かに今までにない情報を手に出来るのであれば、企業がこぞって取り組もうとしているのも当然だろう。しかし隠されていた事実が明らかになるということは、個人にとって好ましい事態だけを生むとは限らない。また新しい仕組みに対応することを通じて、個人の生活から社会全体に至るまで、様々な側面が姿を変えて行くことも考えられる。

ビッグデータを扱うことで、どのような情報が手に入るようになり、そこからどのような行動が取れるようになるのだろうか。またどのようなリスクが生まれ、私たちはどう対処して行くべきなのだろうか。いま到来しつつある「ビッグデータ社会」の可能性を概観してみることにしよう。

ビッグデータの3要素

データを集め、分析し、隠されていた情報を得る。それは何も、最近になって始められた手法ではない。例えばストーンヘンジやピラミッドなど、世界各地に残る紀元前数千年前の遺跡から、古代人が天体の運行に関する正確な情報を有していたことが分かっている。彼らは何十年も辛抱強く天体観測を行い、わずかなデータを蓄積して分析することで、こうした知識を手に入れていたのだろう。

さらにギリシャ・アンティキティラ島の沖からは、紀元前150から100年前に制作されたと考えられる「天体運行を計算するための機械」まで発見されている。データを集め、利用可能にするという行為には人類と同じくらいの歴史があるのだ。

ではなぜ今、「ビッグデータ」という新しい言葉が登場してきたのか。そこには3つの要素が存在している。

最初の要素は、「ビッグ」という単語が示す通り「データ量の増加」である。米調査会社のIDCは、2010年から15年にかけてのわずか5年間で、世界全体のデジタルデータ総量が約1.2ゼタバイトから約7.9ゼタバイトへと6倍以上に増加すると予測している[*1]。1ゼタバイトは地球上にある砂浜の砂粒の数と言われていることを考えれば、これがいかに膨大な量か分かるだろう。

こうした状況を背景に、企業によって扱われるデータ量も飛躍的に増加している。例えばグーグルが処理するデータ量は毎時約1ペタバイト（100万ギガバイト）と言われており、またデータウェアハウスの提供を行っている米テラデータ社によれば、同社の顧客であるイーベイ社のデータウェアハウスは84ペタバイトにまで達している。

2つめの要素は、データ形式の変化だ。同じくIDCの予測によれば、デジタルデータの中で非構造化データが占める割合は、将来的に9割以上になるとされている。非構造化データとは、ブログ等への書き込みや画像・映像など、従来型のデータベースには格納できないデータのこと。単にデータの容量が増えるだけでなく、その種類も多様化するわけだ。

そして3つめの要素が、処理速度の圧倒的な向上である。

大容量データの処理はこれまでも取り組まれてきたのだが、通常はバッチ処理で、長い時間をかけて処理されてきた。しかし技術革新により、従来は考えられなかったほどの短時間で処理を完了することが可能になってきている。さらにリアルタイムで分析結果を返す、という事例も珍しくない。

こうした3つの要素が含まれていること――つまり非構造化データを含む大容量のデータを、高速で処理することが、従来のデータ分析と「ビッグデータ」の違いと言えるだろう。

ただし現状を見ると、非構造化データが含まれていない、単なる大容量データの高速処理が「ビッグデータ」として紹介されている例も少なくない。またどの程度のデータ量が処理されれば「ビッグ」データなのかは、今後の技術革新によって変化して行くことが予想される。定義があいまいになるのは新しい技術の常であり、ユーザーには個々の事例を注視する姿勢が求められるだろう。

データ爆発を支える技術

ビッグデータを最も身近に感じる瞬間。多くの人にとって、それはiPodなどのデジタルオーディオ端末で音楽を持ち歩くことかもしれない。「今日はどのテープ（あるいはCDやMD）を持って行こうか」などと悩んだのは遙か昔の話で、いまや数百曲、数千曲という単位で好きな音楽を持ち歩くことができる。しかも端末を購入するのに必要なお金も、数千円程度で十分だ。

米マッキンゼー社の計算によれば、いま世界中にある音楽をデジタルデータで保存しようとした場合、必要な記憶媒体にかかるコストはたったの600ドルである。また1ギガバイトを保存するのに必要なコストは、2005年の時点では約19ドルだったのに対し、2015年には約0.7ドルと30分の1近くに下落すると予測されている。[*2]こうした技術革新を背景に、一般の個人や企業でも膨大なデジタルデータを保有することが可能になっており、ビッグデータというコンセプトを促す一因となっている。

ただ大量のデータを持っていても、分析できなければ意味が無い。そしていくらハードウェアの性能が上がっているからといって、単体のマシンだけで処理できるデータ量には限界がある。そこで近年、複数のマシンに分けて処理を実行する「分散並列コンピューティング」が追求されるようになってきており、関連技術が次々に実用化されている。

その代表例が、マップリデュース（MapReduce）やハドゥープ（Hadoop）と呼ばれる大規模分散処理のフレームワークだ。ハドゥープはオープンソースとして公開されており、アマゾンやフェイスブック、楽天などの大手企業で採用され、既に多くのユーザーが恩恵を受けている。また分散データベースのHBase、分散ファイルシステムのGFS（Google File System）、オープンソースの統計解析向けプログラミング言語「R」など、関連技術が続々と登場している状況である。

また間接的にビッグデータを支えている技術として、クラウドコンピューティングの登場も無視できない。

いくらデータの保存、処理に革新的な技術が生まれているといっても、ある程度のハードウェアが必要な状況には変わりがない。今後ビッグデータに継続的に力を入れて行くという覚悟を決めた大企業ならば話は別だが、一般企業でいきなりサーバやストレージを買うと決断するのは難しいだろう。またビッグデータ関連技術はまだ登場して間もなく、十分なスキルを持った技術者が確保できない恐れもある。

しかしクラウドを利用すれば、誰でも必要な処理能力を必要な時に、必要な分だけ手に入れることができる。例えば米ニューヨークタイムズ紙は、1851年から1922年までの同紙のアーカイブ、40万ファイル以上のイメージ画像をPDFファイル化するために、米アマゾン社のクラウドサービス「Amazon EC2」を利用。仮想マシン100台分の処理能力をレンタルし、わずか24時間でタスクを完了させることができたそうである。[*3]

クラウドコンピューティングを提供する企業の側でも、ビッグデータ関連サービスを需要喚起の施策として捉えるようになってきている。クラウドとビッグデータが同じ文脈で語られることも多くなって行くだろう。

ソーシャルメディアとスマートフォンの定着

浴槽を用意すれば自然に湯でいっぱいになるわけではないように、サーバとストレージを用意すればどこからか膨大なデータがやってくるわけではない。そこにはデータの源泉が必要になるが、ビッグデータにおいて期待されている源の1つがソーシャルメディアである。

冒頭で紹介したように、いまや主要なソーシャルメディアの利用者数は数億人という単位に達している。そして「誰かとコミュニケーションしたい」という人間の根本的な欲求が刺激されることで、ソーシャルメディア上では加速度的にデータが増えている状況だ。

かつてフェイスブックでデータ分析チームを指揮し、現在は米クラウデラ社の主任研究員を務めるジェフ・ハマーバッカーは、フェイスブックの「ウォール」（個人ページ上の掲示板的機能）に対する書き込み量が、ブログ全体の10倍にも達すると推定されると語っている[*4]。またユーチューブ上には、1分間に約24時間分の長さの映像がアップロードされているそうだ。前述したIDCの予測でも、将来的に全デジタルデータの75パーセントが個人によって生成されると考えられており、「個人」がデータの源泉として最も重要な存在になってゆくことだろう。

さらにソーシャルメディア上に寄せられているのは、テキストデータばかりではない。スマートフォンの普及により、従来は難しかった種類のデータまで手軽に扱われるようになってきている。

日本では「携帯電話で写真を撮影・共有する」という発想は以前からお馴染みのものだったが、海外でもスマートフォンの登場によって定着しつつある。写真共有サイトのフリッカーでは、アップロードされた写真の撮影に使われた機種を集計、グラフで公開しているのだが、2008年にはアップルのアイフォーンが1位の座を獲得した。またグーグルが開設したソーシャルメディア「Google+」のアンドロイド携帯電話向けアプリには、撮影した画像・映像を自動的にピカサ（グーグルの写真共有サービス）に送信する「インスタントアップロード」機能が設けられているが、今後はこうした「撮影して即ウェブ公開」という行動もスマートフォンによって加速されて行くことだろう。

またGPSを通じて得られる位置情報も、スマートフォンによって取り扱いが容易になったデータの1つだ。位置情報系サービス（あるいは既存ウェブサービス内の位置情報系機能）の利用は急速に一般化しつつあり、ソーシャルメディア上に新たなデータをもたらす要因となっている。そのほか音声や振動、端末の向きなど、様々なデータがスマートフォンを通じて取得・集約されるようになってきている。

M2Mとセンサーの普及

もう1つ、ビッグデータの源泉として注目しなければならないのがM2Mの拡大である。

M2Mとは「Machine to Machine」を略したもので、文字通り機械と機械の間でデータのやり取りを行うことを意味している。携帯電話で会話することも機械と機械の通信と言えるが、M2Mとは特に、様々な機器類が接続して自動で動作することを示す言葉だ。

例えばIBMは、「スマートマンホール」という一風変わった機器を発表している。自らを通過した水量を自動で検知し、データを送信してくれるというマンホールだ。これを街中に配置しておけば、送られてくるデータを分析することで水量を把握、大雨による水溢れを防止するといった対応が可能になるわけである。

こうした「スマート」な機器が、今後社会の中に普及して行くと予測されている。携帯通信事業者の業界団体であるGSMAの調査によれば、無線ネットワークに接続している機器（携帯電話を含む）の数は、2011年時点で約90億台。これが2020年には、約245億台へと増加すると考えられている。このうちM2M関連の機器については、2011年の約25億台から2020年の約126億台へと、10年弱で5倍に成長する。

またM2Mの分野では、ETSI（欧州電気通信標準化機構）や米国のTIA（電気通信工業会）などといった団体によって、標準化が進められていることにも注目だ。これまでM2Mでは、ベンダーが独自のインターフェースに基づいて製品を提供してきたため、ベンダーが異なる製品間で相互接続性を実現するのが難しかった。標準化が進められればM2Mシステムの普及やコストの下落も進み、それだけM2M由来のデータ量も増加すると考えられるのである。

さらにセンサー技術の進歩と普及により、M2Mによって収集・送信されるデータの種類も多様化している。

例えば交通情報の分野では、以前から道路に埋め込んだセンサーで交通量を把握するという取り組みが行われてきたが、最近では自動車から様々なデータ（位置情報や走行状態など）を集め、それを分析して渋滞などが推測できるようになっている。2011年の東日本大震災やその後の台風被害においては、各社が持つデータを統合して、どの道路が走行できる状態なのかを把握するために役立てられた。

またNTTドコモは携帯電話の基地局を利用して「環境センサーネットワーク」の整備を進めている。これは気温や雨量、雷といった気象データ、さらに花粉の飛散量といったデータを収集できる各種センサーを基地局に設置、データ収集を行うというもので、2011年度中に4000局まで対象局を拡大することが予定されている。さながら「スマート基地局」といったところだろうか。

カリフォルニア大学サンディエゴ校のロジャー・ボーン教授は、「機械によって生成され、機械によって利用される情報の量は、他のあらゆる情報よりも速いペースで増加して行くだろう」と予測している。[*5]さらに人間と違い、誤ったデータを送ってしまうことも、疲れて休んでしまうこともないM2M＋センサーという組み合わせは、ビッグデータにおいても重要な役割を演じることになるだろう。

集約されるデータ

19世紀のアメリカ。海軍士官であったマシュー・フォンテーン・モーリーは、公務中の事故で足が不自由となり、やむなく現場から後方支援にまわることとなった。

しかしこの事態は、米軍にとって幸運な出来事だったと言えるかもしれない。海に出られなくなったモーリーは、代わりに船舶から無数の航海日誌（ログ）を収集。そこから統計分析を行うことで、潮の流れや風向きなどの情報を割り出すという方法を確立したのである。さらに彼は1853年にブリュッセルで行われた国際会議において、気象観測資料の報告形式の統一など、国際的なデータ収集の仕組みを提案。米海軍天文台に各国からの報告が寄せられるようになり、分析を実施して結果を世界に公表するという活動が行われることとなった。

データを増やそうとした場合、自ら記録を取ったり様々なセンサー類を配置したりする以外に、既に存在するデータを集約するという方法がある。モーリーが取ったのはまさにそんな行動で、最終的には国の壁を越えてデータの集約に成功したわけだ。そして今日でも、利用可能なデータ量を増やすために、データの標準化やアクセスの容易化といった取り組みが行われている。

例えばウェブの世界では、「ウェブ2.0」という言葉が囁かれるようになった2005年の後半頃から、APIという概念が一般化するようになった。APIは「Application Programming Interface」の略で、これが提供されているプログラムは、別のプログラムから操作することが可能になる。仮にそのプログラムが独自のデータを保有していた場合、APIを通じてデータを取得することもできるのだ。

冒頭で紹介したように、ツイッターが社会分析のリソースとして注目されるようになっているが、その一因としてツイッターが様々なAPIを公開していることが挙げられる。一定の制限はあるが、大量のツイートを分析して情報を手にしたいと思う者は誰でも、すぐにAPIを通じてデータを手に入れることができるのである。

また政府や研究機関などの公的組織から、これまで様々な理由でアクセスが難しかったデータを、ウェブ上等で積極的に公開するという動きも生まれている。

2008年に就任したオバマ大統領は、就任以来「開かれた政府（オープンガバメント）」を掲げ、政府に関係するデータおよび情報システムについて様々な改革を行っている。2009年5月には、各種政府機関が保有するデータを生のままで公開するサイト「Data.gov」を開設、一般市民でも400種類以上とされる多種多様な統計データにアクセスできるようになった。こうしたデータ公開は州や市のレベルでも進んでおり、「公共機関のデータを利用して面白いサービスを創造するコンテスト」まで各地で開催される状況となっている。

またデータの集約は、同じ種類の中だけで行われていれば良いというものではない。多種多様なデータを組み合わせて「ビッグデータ」を実現することからも、様々な価値が生まれようとしている。

2006年9月、米ウィスコンシン州で大腸菌O157による集団感染が発生し、数週間のうちに患者の数が150人近くまで達するという事件が起きている。米疾病対策センター（CDC）は統計的分析から袋詰めホウレン草が感染源ではないかと疑い、最終的にその疑いが正しかったことが証明さ

れるのだが、調査の上で重要だったのは情報共有ネットワークの存在であったことを、ニューヨーク大学非常勤教授のカイザー・ファングが著書”Numbers Rule Your World”（邦題『ヤバい経済学』）で指摘している。全米各地の公衆衛生研究所の情報を1つにまとめる「パルスネット」、腸管疾患感染に携わる疫学者が情報共有を行っている「アウトブレイクネット」、州の公衆衛生部門のネットワークである「フードネット」など、様々な情報インフラによって様々な種類の情報が集められた結果、感染源の特定という非常に困難なタスクが達成されたのである。

　こうした多種多様なデータを一括で処理できるようになることも、「ビッグデータ」というアプローチの特徴の1つだ。面白いのはデータが組み合わされることで、一見無意味な情報までが重要な存在になってくるという点である。

　例えば調査会社のニールセンでは、ツイートからテレビ番組に対する反応を把握するために、ツイッターから得られる情報だけでなく、社会的背景による話し言葉・書き言葉の変化に関するデータを活用している。こうしたクセからユーザーの個人的な情報（どこに住んでいる何歳くらいの人物で、男女どちらなのか等）を割り出すことで、「この番組は若い女性には人気だが、年配の男性からは批判的だ」といった傾向を導き出せるというわけだ。このように様々なレベルでのデータにアクセスできるようになることで、ビッグデータの役割はさらに重要になって行くことだろう。

*****

　技術的・社会的要因を背景に、大量に取得・分析することが可能になったデジタルデータ。そこから何が生まれようとしてのか、次章で見て行くこととしよう。

注釈

１．IDC, ‘The 2011 Digital Universe Study: Extracting Value from Chaos,’
http://www.emc.com/collateral/demos/microsites/emc-digital-universe-2011/index.htm

２．McKinsey Global Institute, May 2011, ‘Big Data: The Next Frontier for Innovation, Competition, and Productivity’

３．ITpro、「『マルチ・クラウド』という選択肢」、2008年9月24日
、http://itpro.nikkeibp.co.jp/article/OPINION/20080922/315232/

４．ジェフ・ハマーバッカー、『ビューティフルデータ』、オライリージャパン、2011年。

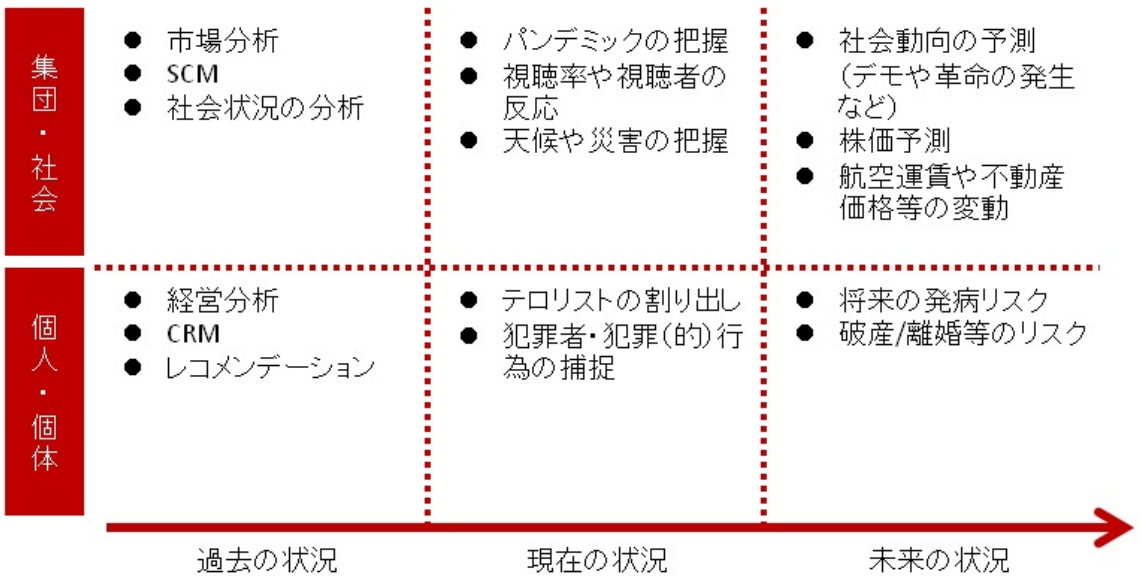
５．The Economist, February 25, 2010, ‘All Too Much’

映画化されたデータ分析

本書の公開とほぼ同じくして、一本の映画が日本で封切られる。名前は『マネーボール』。野球に関する様々なデータを統計分析し、そこから戦略を考え出そうというアプローチ「セイバーメトリクス」を一躍世に知らしめた同名著の映画化である。主人公ビリー・ビーンを演じるのはブラッド・ピット。彼の活躍する姿は恐らく、映画の感動と同時に、データ分析が拓く可能性を広く印象づけるものとなるだろう。

セイバーメトリクスの発想は野球を超え、近年ではバスケットボールなど他のスポーツや、映画制作といったエンターテイメントの分野にまで応用されるようになってきている。しかし、こうした領域でデータ分析が行われるのはもはや驚きではないだろう。冒頭の例のように株価の傾向を予測するために使われる場合もあれば、たった一人の犯罪者を捉えたり、カンニングを防止するために活用されたりする場合もある。処理スピードの速さを活かして、これまでは社会全体の動きをリアルタイムで把握する、などというのも重要な利用法だ。

ここではビッグデータが「見ようとしているもの」に着目し、「個人と集団のどちらを見るのか」と「過去・現在・未来のいつを見るのか」という2つの軸から利用法を整理してみたい。



社会インフラとビッグデータ

データから過去の状況を把握すること。程度の差はあれ、企業で働いている人々であれば誰しも経験している行為だろう。様々な市場分析や競合他社の分析、あるいは自社の経営分析などが毎日のように行われ、サプライチェーン管理（SCM）や顧客管理（CRM）など、関係する情報システムも日々高度なものになっている。

一方でビッグデータの技術を用いることで、これまでは巨大すぎて扱うのが困難だった存在も分析の対象にしようという取り組みが始まっている。その代表が、エネルギーなど社会インフラの世界だ。

東日本大震災以降、日本でも自然エネルギーの活用に注目が集まっているが、風力発電機の市場でトップシェアを握る企業の1つに、デンマークのヴェスタス社がある。2011年10月、彼らはIBMのビッグデータ分析システムを導入し、風力タービンの設置場所決定に役立てることを発表した。

自然エネルギーを利用した発電には「出力が一定ではない」という最大の弱点がある。風力発電の場合も、タービンが設置される場所の地形や気象条件等によって発電量は大きく変化する。それを正確に、しかも機器の設置前に把握するのは困難であり、日本でもつくば市と早稲田大学の間で告訴騒ぎにまで発展したことが記憶に新しい。ヴェスタスはビッグデータを活用することで、判断の精度を上げようとしているわけだ。

同社のデータを構成するのは、天候や潮汐に関するデータ、地理情報、センサーデータ等々で、合計で数ペタバイトにも達している。当然ながら分析はこれまでも行われてきたのだが、以前は数週間を要していたデータ解析が1時間以内で完了できるようになるとのこと。さらに同社では、今後4年間で気象データはますます複雑化し、20ペタバイトを超えるのではないかと予測している。

電力に関するデータ分析が貢献しているのは、発電の領域だけではない。ICT技術を使い、発電所から一般家庭に至るまでの電力網全体の状況を把握、電力利用の最適化を図ろうという「スマートグリッド」や、その中で特に需要家の側に注目し、建築物内での最適化を実現する「スマートハウス」「スマートビル」などといった概念が実用化に向けて動き出している。

さらにその延長線上には、電力以外の様々な社会インフラ、例えばガス・水道・交通などの状況を含めた社会全体を把握しようという「スマートシティ」「スマートコミュニティ」などの概念も生まれてきている。当然ながら、こうしたシステムの実用化がもたらすのは新たなビッグデータの登場だ。例えばオラクルでは、発電施設や送配電網に設置されるセンサー、あるいは家庭に設置されるスマートメーター（データ処理能力と通信能力を備えた次世代型電力検針器）などの機器類の増加により、電力会社が取り扱うデータ量は将来的に800テラバイト以上になると予測している。

その需要を狙い、IBMやオラクルを始めとした主要なIT企業が既に行動を始めており、ビッグデータ活用の主要な市場の1つとして期待されている。またNTTドコモが携帯電話の使用状況データを都市計画に役立てようとしているなど、これまでは関係の薄かったデータを役立てようという

動きも少なくない。社会インフラの改善は世界各国での主要テーマであるだけに、今後はヴェスタス社のような事例を数多く耳にするようになるだろう。

高度化するレコメンデーション

過去の傾向を割り出すことで恩恵を受けられるのは、社会という大きな単位の中だけではない。個人のレベルでも様々な価値を手にすることができるようになっているが、その代表例の1つはユーザーの趣味・嗜好に基づく情報のフィルタリング、いわゆる「レコメンデーション」の分野だろう。

例えば「過去の行動履歴から商品をおすすめする」という発想をいち早く具体化したアマゾン社では、既にレコメンデーション経由での売上が全体の約30パーセントを構成している。フェイスブックはサイト内でのユーザーの行動を分析し、エッジランク（EdgeRank）という仕組みで情報を整理、ユーザーが知りたがっていると考えられる情報を上位に表示するほか、表示される広告もユーザーの嗜好や属性に合うよう調整がなされている。

もちろんこうしたレコメンデーションは、サービス側が売上やサイト滞在時間を増やすために行っているという側面もあるが、反面で本来の目的、つまり「ユーザーが欲しいと思っているもの」を提供しているのも事実だ。アマゾンで最近の興味に一致する本があることを教えてもらったり、アイチューンズ（iTunes）の「ジーニアス・プレイリスト」（世界のアイチューンズユーザーが作成したプレイリストの情報を集め、それを元に相性のいい曲を推薦する機能）を通じて思いがけない一曲を知ったり、などという体験をした人も多いだろう。

本や音楽ももちろん大切な存在だが、より重要なもの、例えば将来の伴侶を選ぶという場合はどうだろうか？

英紙フィナンシャル・タイムズは大手デートサイト「マッチ・ドットコム」に関する取材を行い、「マッチ・ドットコムの裏側（Inside Match.com）[*6]」という記事を発表している。マッチ・ドットコムは交際相手を探すためのサービスだが、「出会い系」などという軟派なイメージとは異なり、テラバイト級のデータを駆使するIT企業であることを記事は指摘。2005年から現在までででサイト上で交わされたメールの数は、実に12億通を数えるそうだ。

マッチ・ドットコムではより相思相愛のカップルを誕生させるために、「シナプス」と呼ばれる新しいアルゴリズムを開発。会員が自ら希望する条件（年齢層や外見など）を考慮するだけでなく、「彼らがサイト内でどのように行動しているか」をデータ化して蓄積、そこから読み取れる情報も加味して候補者を探すという仕組みを作り上げている。

まるでアマゾンかフェイスブックの話を聞いているようだが、なぜ自己申告した条件だけでなく、サイト内の行動まで把握する必要があるのか。それは自らの行動を振り返ってみれば、説明するまでもない話かもしれない。自分はロングヘアの方が好みだと思っていたのに、実際にはショートカットの人物の方が気になってしまう――そんな無意識の行為をサイト上で繰り返すうちに、システムの側で分析を行い、隠れた（しかし本心に近い）望みを明らかにしてくれるというわけだ。

こうした分析アルゴリズム等のおかげでマッチ・ドットコムの評価は上昇、有料会員数は1年間で30パーセントの増加を見せている。

他にも医療の分野では、過去のヘルスケアデータから患者個人にとって「オススメ」な治療法

や薬を割り出す、などといった試みが始まっている。過去の状況を分析するだけでも、これからの人生を左右するような決断を下す上で、大きなヒントを手にすることができるのである。

危機的状況の把握

生涯の伴侶を探すというのは一生の決断であり、その時その時の感情で判断するといった類の話ではない。しかし一部の分野では、「いま何が起きているか」というリアルタイム性が何にも増して要求される。そんな状況も、ビッグデータが真価を発揮する場面だ。

地球規模での気候変動が起きているのかどうか、専門家の間でも意見が分かれているところだが、ここ数年異常気象による被害が目立っていることは事実だろう。それは天候だけでなく、少子高齢化や社会インフラの老朽化といった社会環境の側面にも原因はあるものの、いずれにせよ被害を防止・低減させる努力が求められている。

中でも都市部で問題になっているのが、ゲリラ豪雨（局所的・突発的な集中豪雨）の発生だ。ゲリラ豪雨は予測が困難な一方で、一度発生すると地下街への浸水をもたらすなど、被害は決して無視できない。そこで様々な対応が進められているが、気象情報サービス大手のウェザーニュース社では、同社に登録した「市民レポーター」から寄せられるリアルタイムの情報を予想に役立てるという試みを行っている。

現在登録されている市民レポーターの数は20万人以上。彼らはGPS付きの携帯電話を使い、現状の報告と位置情報を合わせてウェザーニュース側に送信。それを同社で分析するという仕組みだが、未登録のユーザーから寄せられるものも合わせると、1日に平均で20万件以上の報告が届くそうである。[*7]

こうした大量の情報を、気象庁の観測データなどと組み合わせて予測を行ったところ、ゲリラ豪雨の発生を7割超の確率で事前予測することに成功。現在同社では、他の気象予報の判断基準としても「市民レポーター」からの報告を採用している。

リアルタイムでの状況把握が重要になるのは、気象情報の分野だけではない。その他の自然災害の状況や、感染症の流行、事件・事故の状況なども同様の対応が必要になるが、これらの分野でもビッグデータが活用されて始めている。

2008年11月、グーグルは「Google Flu Trends」という一風変わった名前のサービスをオープンさせた。これは米国内でのインフルエンザの流行状況（現在は日本国内版[*8]も存在）をグラフで示してくれるというものだが、その仕組みがユニークだ。インフルエンザが身の回りで発生するようになると、人々は関連情報や予防策を知るために、グーグルの検索窓に関連語句を入力する。それがごく少数の行動であれば意味はないが、グーグル検索を利用しているのは何千何百万人という人々だ。グーグルは「インフルエンザ関連のトピックを検索するユーザー数と、実際にインフルエンザの症状を示す患者数の間に密接な関連性がある[*9]」ことを発見、それを可視化するサービスを開発したのである。

またペンシルバニア州立大学の研究者らは、インフルエンザの予防接種を人々がどの程度受けているかを把握するため、ツイッターを分析するという研究を行っている。

彼らは2009年8月から2010年1月にかけ、インフルエンザ予防接種について語っている477,768件のツイートを収集。予防接種に対する態度を「肯定的」「否定的」「中立/無関係」の3つに分類すると共に、位置情報が含まれているツイートから、全米のどの地域で摂取率が高いかを把握す

るという試みを行った。現時点ではリアルタイムでの解析結果ではないが、研究者等はこうした技術を確立することで、将来的に「予防接種の呼びかけが必要な地域」や「今後必要となるワクチンの量」といった情報を把握することが可能になるだろうと期待している。

こうした危機的状況の把握は、あくまでも社会全体に対して行われているものだ。しかし分析の方向性を変えることで、個人が取っている行為をいち早く捕捉することも可能になる。そうしたアプローチが重要になる世界といえば、もちろん犯罪対策の分野だ。

金融機関やクレジットカード会社は、犯罪対策のために早い段階から大量データ分析を活用してきた業界の一つである。例えば一般的な（犯罪ではない）利用パターンを分析しておき、そこから逸脱するような行動パターンが見られた場合には、すぐにアラートが上がるといったシステムが既に各所で導入されている。ちなみに同じ発想は、CBT（Computer Based Testing、コンピュータを使って各種試験を実施すること）に応用され、カンニングなどの不正行為を検知するという試みに役立てられている。

金融機関では利用者の行動を「お金」という数値で把握することが可能なため、ある意味でこうした分析が導入しやすかったとも言えるだろう。ただ最近では、より正確な判断を行うため、外部に存在する多種多様なデータを参照しようという動きが見られる。

ここでも重要になるのが、ソーシャルメディアの存在だ。例えば米国の銀行では、融資の依頼を受けた場合に、相手にどの程度の返済能力があるかを判断する要素の1つとしてフェイスブック上の書き込みを考慮に入れるという行為を始めている。現時点では「犯罪者を見つける」という段階ではないが、例えば豪華な食事の写真をアップする頻度が増えてきたらアラートを上げる（良い傾向か悪い傾向かはアルゴリズム次第だが）などといった仕組みへと発展させることも可能だろう。

同様に既に普及している取り組みとしては、コンテンツ共有サイトにおける違法行為対策が挙げられる。

前述のように、ネット上で様々なコンテンツを共有することはごく当たり前の行為となり、ユーチューブ上にはわずか1分間で24時間分の映像がアップロードされるような状況だ。誰かが全ての映像を確認し、著作権に違反するものや、ポルノなどの違法コンテンツが公開されていないかチェックすることは到底不可能である。そこでデータ分析の出番というわけだが、例えばユーチューブにも採用されている米オーディブル・マジック社の技術では、オリジナル映像の特徴を「指紋」として用意しておき、それに合致する映像が無いかを解析するという処理が行われている。他にも顔認識技術で著名人が写っているかどうかを確認する・肌色の多さでポルノの可能性を疑うなど、様々な解析技術が登場してきている。

ただし違法アップロードを行う人々の側でも、こうした対策の存在を認識した上で、それを出し抜くような行動が行われている。対抗するためにはより複雑な処理を、より大量のデータに対して実行することが求められ、その分ビッグデータ技術に寄せられる期待も高まることだろう。

革命はツイートされるのか？

2009年6月。大勢の若いイラン人達がテヘランの街頭に繰り出し、直前に行われた大統領選挙（現職のアハマディネジャド大統領が当選）の最中に不正行為があったとして、大規模な抗議活動を始めた。その様子はツイッターを通じて世界に発信され、さらにデモ参加者の一人（ネダという名の若い女性）が射殺されるというショッキングなシーンがユーチューブ上で公開されるに至り、緊張が一気に高まった。

その後反政府活動は制圧され、現在に至るまでアハマディネジャド大統領の退陣は実現されていない。しかしイランの外でこの状況を眺めていた人々の一部には、「ソーシャルメディアが民主主義を促すのではないか」という期待感（あるいは恐怖感）が高まることとなった。中でもアトランティック誌のアンドリュー・サリバン記者は、「ツイートされる革命」と題した記事を公開[*10]、ツイッターが（体制側に利用されている）既存メディアをバイパスして、抗議活動を組織化するのに役立てられていると指摘、大きな反響を呼んだ。

実際にツイッター、あるいはソーシャルメディア全般が民主主義を促すツールであったのかどうか、あるいは今後そうなる潜在性を秘めているのかどうかは、専門家の間でも意見が分かれるところだ。しかし2011年に相次いで発生したチュニジア革命・エジプト革命、その後に発生した英国の暴動、そして米国のデモ活動「ウォール街を選挙せよ（Occupy Wall Street）」などで各種ウェブサービスが活用されたことにより、少なくとも抗議活動がソーシャルメディア上でも展開されるという状況は珍しいものではなくなった。

そこで生まれてきたのが、ソーシャルメディアの解析を通じて未来の重大事件を予測するという発想だ。米国防総省の先端研究プロジェクト（Intelligence Advanced Research Projects Activity, IARPA）は「オープンソース・インディケーター（OSI）」と名付けられたプロジェクトを2011年8月に発表。これは検索エンジンで使われる検索語句や、ブログ・ミニブログ上の書き込み、さらには街中の監視カメラやウィキペディアの編集状況などといったデータまでを統合して、革命など重要な社会的事件の発想を予測しようという内容である。

同プログラムに対しては批判も多く、どこまでの精度で予測ができるのか、そもそもソーシャルメディアと過去の革命の結びつきは小さかったのではないかといった疑問が呈されている。しかし株価の予測など、限定された分野ではビッグデータの「予測能力」が示されるようになってきており、決して荒唐無稽な試みとは言えないだろう。

具体的に将来予測をビジネスにしている企業も存在する。土壌や天候などのデータを解析することを通じて、農作物の生育に関する保険を販売する「クライメート・コーポレーション」である。元グーグル社員によって2006年に設立された同社は、公的機関から提供される無料のデータ（温度や降雨量、土壌、農作物の収穫量など）14テラバイト分を集約し、トウモロコシや麦、大豆等の発育量予測を実施。農業関係者に対して情報提供を行うと共に、悪天候などによる被害を補う「天候保険」を販売している。ちなみにデータの処理には、アマゾンのクラウドサービスを利用しているそうだ。

またマイクロソフトが提供している「ビング・トラベル（Bing Travel）[*11]」の中には、航空運

賃予測という機能がある。これはかつて「フェアキャスト（FareCast）」として公開されていたサービスで、2,250億件にものぼる航空関連情報を分析し、今後の航空運賃の変動を予測してくれる（参照するデータの中には、燃料価格や大規模なイベントの発生など、航空券の需要/供給に関する要因も含まれる）。それによって特定の航空券をいま買うのが得なのか、それとも出発ギリギリまでねばった方が良いのかを判断できるというわけだ。

　他にも特定の病気に罹るリスクを算出する、あるいは離婚や破産に至るリスクを算出するといった形で、個人の未来を予測するためにビッグデータを応用しようという動きが登場している。未来予測はビッグデータの最終目標であり、現在は精度が低くても、チャレンジングな取り組みが次々と生まれてくることだろう。

*****

　米国では特定の映画に対するソーシャルメディア上での反応が、最終的な興行成績を予測する上で有効であるとの研究結果が出てきており、こうした反応をまとめるウェブサイトなども登場してきている。果たして映画『マネーボール』に対してはどのような会話が飛び交い、興行成績とどのように関係するのか、そちらの分析にも注目が必要かもしれない。

注釈

６．http://www.ft.com/intl/cms/s/2/f31cae04-b8ca-11e0-8206-00144feabdc0.html

７．「＜予測する＞データの増加と精度向上で革命」、Tech-On!、
http://techon.nikkeibp.co.jp/article/FEATURE/20111018/199452/

８．「Googleインフルトレンド｜日本」、
http://www.google.org/flutrends/intl/ja/jp/#JP

９．「Googleインフルトレンド｜インフルトレンドの仕組み」、
http://www.google.org/flutrends/intl/ja/about/how.html

１０．Andrew Sullivan, ‘The Revolution Will Be Twittered,’ The Atlantic,
http://www.theatlantic.com/daily-dish/archive/2009/06/the-revolution-will-be-twittered/200478/

１１．http://www.bing.com/travel/

1930年の「夢の技術」

米国の歴史家ジョセフ・コーンは、1920年代から30年代にかけて、ある技術に対して「民主主義や自由・平等を広め、戦争や暴力の世界を過去のものにするだろう」という評価がなされていたことを指摘している。さて、この夢のような技術とはいったい何なのだろうか？

その答えは「航空機」である。航空機によって空間上の隔たりを超えることが容易になり、異文化間の交流が促される結果、相互理解や民主主義が進むだろうというわけだ。

しかし航空機という技術がその後に何をもたらしたのか、21世紀に住む私たちはよく知っている。本格的な兵器としての立場を確立した航空機は、第二次世界大戦における大空襲や原子爆弾の投下、ベトナム戦争における化学兵器の使用といった悲惨な出来事に関与することとなった。あらゆる技術は中立ではあり得ず、利用者は善と悪、両方の顔に直面することを覚悟しなければならない。

それでは「ビッグデータ」にはどのようなリスクがあり、私たちにはどのような課題が突きつけられているのだろうか。もちろんビッグデータは航空機のように、社会に対して物理的なダメージを与えるわけではない。しかし情報の力によって、個人の生活に対して実害をもたらすことは十分に可能である。新しい技術の到来を喜びつつも、その負の側面にも対応する姿勢が必要だろう。

複雑化するプライバシー侵害

2010年7月、ニューヨークタイムズ紙のウェブサイト「マルチメディア」コーナーに1枚の地図が掲載された[*12]。ニューヨーク・マンハッタン島とその周辺を対象にしたその地図には、小さな青丸が無数に配置され、カーソルを合わせると詳細情報がポップアップされるという仕掛けになっている。

実はこの地図、ニューヨークにおける職務質問の実態を示したもの。ニューヨーク市警本部から提供されたデータを集計し、丸の配置で職務質問が行われた場所を、丸の大きさで行われた回数を表している。どの地域で職務質問を受けることが多いかということは、ある面では警察の活動状況を、また別の面では特定の地域の危険度を示したものと言えるだろう。

しかし最も重要なのは、ポップアップされる詳細情報の中身だ。そこには職務質問の実数と同時に、各地域における白人と非白人の居住者割合、さらに職務質問対象者の白人と非白人の割合が掲載されていた。つまり「非白人の居住者が少ないのに非白人に対する職務質問が多い」ということになれば、警察官が何らかの偏見を抱いていることを示すわけである。地図上ではそれ以上ドリルダウン（情報の詳細化）はできないようになっていたが、市警本部では担当した警察官のレベルまで把握していることは想像に難くない。意図的なものかは別にして、蓄積されたデータはこのような隠れた傾向まで暴いてしまうのである。

このような手法は「データジャーナリズム」として、報道機関に新たな調査報道を可能にするものとして注目されているが、同じアプローチを個人の秘密を暴くことに使われる恐れは否定できない。しかも思いもよらぬデータから個人的な情報が明らかになるという点で、従来にも増して防止や対応が困難であるという性質も持っている。

最近ドイツのミュンスター応用科学大学から、興味深い研究結果が発表された。スマートメーターによって採集された電力消費データを分析することで、家庭内で視聴されているテレビ番組やDVDなどのコンテンツを判別することが可能というのである。まだ研究レベルの可能性でしかないが、実用化が進めば電力会社が正確無比の視聴率情報を握る企業へと変貌を遂げるかもしれない。

また国内で行われている「スマートグリッド」や「スマートハウス」実証実験においても、収集される生活関連データで個人の行動をある程度把握できることが明らかになっている。さらにジョージア工科大学からは、スマートフォン経由で収集される音や振動といったデータを解析することで、近くにあるキーボードで何がタイプされているのかを把握する技術まで発表された。人々はよもや普通に生活しているだけで個人的情報が漏れているなどとは考えもしないだろうから、この種の「情報漏洩」に対抗するためには、消費者の側で自ら学んで行くという姿勢が要求されるだろう。

もう一つ、ビッグデータ時代のプライバシーを考える上で注意しなければならないのは、データの組み合わせで秘密が暴かれるという状況である。従来も「名寄せ」という行為により、断片的なデータが集約されることでプライバシーが侵害されるリスクはあった。しかしビッグデータによって多種多様な情報が同時に処理可能になることで、新しい形でのデータ集約リスクが生ま

れつつある。

例えば前述のニールセンの事例において、仮に同社が「人間の言葉に現れるクセ」というデータを得ることができず、単にツイートを分析しているだけだったとしよう。そのような状況では、ユーザーの性別や年齢層・住所などを推測することは非常に困難なはずだ。

また冒頭のニューヨークタイムズの例でも、住民構成というデータと付き合わせなければ、警察官たちの偏見をあぶり出すことはできない。つまり単独ではリスクのない情報であっても、それが集められて統合されることで、プライバシー侵害の危険性が飛躍的に高まるのである。そのような状況が発生しているのを把握することは、先ほどの「意外なデータから個人情報が漏れる」というケースにも増して困難な作業になるだろう。

そう考えると、「もはやあらゆるデータがプライバシー侵害を後押しする危険性を秘めるようになった」と言わざるを得ないのではないだろうか。プライバシー問題の専門家で、ジョージワシントン大学ロースクールの教授を務めるダニエル・ソロブは、この状況を環境汚染に喩えている。つまり爆発事故のように突然プライバシーが破壊されるのではなく、徐々に汚染物質が蓄積され、気付いた時には貴重なものが失われているというわけだ。

監視社会のリスク

2008年1月。練馬区立の図書館で、本の貸し出し履歴を職員が一定期間参照できるシステムが導入されたことを朝日新聞が報じている。履歴参照を可能にした理由は、貸し出した本が破損されるというケースが相次ぎ、誰が破損したかでトラブルになるケースが増えていたため。蔵書を借りた直近2人の利用者番号を確認できる（最長13週間まで履歴を保存）ようにすることで、余計なトラブルを防ごうというわけだ。

これがなぜニュースになるのか。実は貸し出し履歴を保存することは、個人の思想・信条の自由を侵すことにつながるという批判があり、日本図書館協会は個人情報保護などに関する基準で「返却後は速やかに消去しなければならない」と定めていたのである。またデータが第三者の手に流れ、悪用されるリスクが生まれるという批判もなされた。

しかし改めて考え直す必要もなく、いまや「自分が何を目にしたか・耳にしたか」の履歴はありとあらゆる場所に残されている。図書館の貸し出し履歴よりも、オンライン書店での発注履歴を見られる方が嫌だという人も多いだろう。音楽や映像に関しても、アイチューンズやユーチューブ、動画ストリーミングサイトのユーストリームに至るまで、いくらでも個人の嗜好を後追いできるサイトが存在する。

ここで問題になるのは、データ収集を行っている主体が公的機関であるという点だ。例えば政府や警察機関が個人の思想を把握できるようになれば、それによる弾圧という危険性が生まれかねない。もちろん公的機関が民間組織のデータを強制的に手に入れる可能性も考えられるが、同じ組織内であればよりデータの流用が行いやすいだろう。

どこぞの強権国家ならいざ知らず、日本を始めとした先進国でそんな監視が行われるわけがない、と思うだろうか。実際にそのリスクは小さいかもしれないが、念のため最近グーグルが公表したデータを頭の片隅に入れて置いた方が良いだろう。

2011年10月、グーグルは世界各国の政府から寄せられたデータ提供の要請をまとめ、同社が透明性維持の一環としてまとめている「トランスペアレンシー・レポート[*13]」の一部として公開した。それによると、2011年上半期中には、26の国および地域の政府から、合計で1万5640件の個人情報開示要請が行われていた。ちなみに日本政府から寄せられた個人情報開示要請は75件で、昨年に比べて増加傾向にあり、このうち87%の要請にグーグルは答えたそうである。ウェブ上の行動履歴を政府が入手するという事態は、日本においても決して無縁な話ではない。

また問題なのは、こうした「監視社会」が到来するリスクがあると感じられるようになるだけで、様々な場面におけるデータ品質が悪化しかねないという点である。

例えば図書館の貸し出し履歴が政府に流用されるということになれば、漠然とした不安感を覚え、図書館に近づかなくなるという住民が出てくるかもしれない。また娯楽的な作品、あるいは逆に社会問題を扱った作品の貸し出しは控え、書店でひっそりと購入する、という行動を取ることも考えられるだろう。どちらも貸し出しパターンを歪め、本来であれば正しく行えていたはずの「利用傾向から考えた蔵書構成改善策」を不完全なものとしてしまう。

むしろ危険な書物が読まれなくなるといった「抑止効果」は歓迎すべきであり、多少のデータ

が取れなくなるのは致し方ないという意見もあるかもしれない。しかし一部の人間の不参加が、仕組み全体を破壊してしまう場合もある。例えばスマートメーターによる監視が不安視されてしまえば、導入に反対する住民が増加し、「社会全体の最適化」という最終的な目標を達成することが困難になるだろう。

　今後はどんな些細なデータであっても、それを扱う企業や組織は「データをどこまで公開・流用する意図があるのか」を明確にし、意図せぬ不安感の増大に先手を打っておく必要が生まれると考えられる。グーグルのように「トランスペアレンシー・レポート」を定期的に公表する、という企業も増えてくるかもしれない。

一方で、政府によるデータ活用がより巧妙に行われるという可能性についてはどうだろうか。人々の行動がリアルタイムで把握できるようになれば、意図せぬ行動を取る人々を弾圧せずとも、彼らを巧みに望む方向へと連れて行くことができるだろう。

フェイスブックなどのソーシャルメディアにゲームを提供している企業、ジンガ（Zynga）。彼らが運営する「ソーシャルゲーム」の中にはユーザー数が一億人を突破するものも現れ、ジンガは彼らの膨大な行動履歴を解析し、より良いゲーム作りに役立てている。日々新たに生み出されるデータの量は、実に15テラバイト。実際にジンガは統計分析の専門家や数学系の知識を持つ社員を多数有し、ゲーム会社ではなくテクノロジー企業といった趣だ。

ユーザー行動の分析を通じてユーザーの引き留めを行っているのは、何もジンガだけというわけではない。例えばイアン・エアーズの” Super Crunchers”（邦題『その数学が戦略を決める』）には、ラスベガスのカジノが利用者の行動分析を行い、さらに年齢や平均年収といった様々なデータと統合して「顧客がお金をすっても楽しいと感じ、また来店してくれるのはいくらまでか」を予測、この損失額数値を「痛みポイント」と呼んでいる話が紹介されている。

実はこうした「参加していることをユーザーに楽しく感じてもらうための仕組み」はゲーミフィケーションと呼ばれ、最近ではゲーム以外の分野への応用が進んでいる。例えば夏休みのラジオ体操に出席するとカードにシールがもらえて、それを集めたいがために毎朝早起きして学校に出かけて行くというのも、ごく初歩的なゲーミフィケーションと言えるだろう。それを高度化し、さらにジンガやカジノのように行動履歴把握・データ分析と結びつけることで、ユーザーが知らず知らずのうちに参加してしまう状態をつくり出す企業が増えているのだ。また企業だけでなく、社会問題を解決するためにゲーミフィケーションのテクニックを活用する組織や個人も現れている。

仮に納税者の「痛みポイント」に相当するものがどの程度なのか、データ分析を通じて微妙なバランスを探り当てることができれば、政府は「不必要な」諸手当を廃止することや「低すぎる」税率を上げることが容易に実施できるようになるだろう。それが誰もが幸福になれるユートピアなのか、知らず知らずのうちに支配されているという映画『マトリックス』的なディストピアなのかは一概には判断できない。しかしビッグデータで社会全体からのフィードバックを即座に得られるようになれば、こうしたゲーム化への道へと進むことも可能であることを意識しておく必要があるだろう。

ゼンメルワイスの手洗い運動

マッキンゼーはビッグデータをテーマにした調査報告書の中で、ビッグデータ活用により、米国のヘルスケア分野では年間3000億ドルに相当する価値が創出されるだろうと予想している。同様にヨーロッパの公共セクターが受ける恩恵は年間2500億ユーロと予想され、これはギリシャのGDP以上の額だ。

これだけのお宝が眠っているとあれば、企業はすぐさまビッグデータに飛びついて実用化し、前述のような様々なリスクを世に放ってしまうだろう。そうなる前に早く対応しなければ――

確かにその恐れもあるが、一方で筆者は、それほど早急に企業が動かない可能性も高いと考えている。その理由は、「データを基盤とした行動」という発想をすんなりと受け入れられる人ばかりではないという点にある。

19世紀ハンガリーの医師・ゼンメルワイスは、病院内や医療関係者を清潔に保つことで院内感染を予防するという概念を打ち立てた人物として知られている。ところが現代では当たり前のように聞こえるこの発想も、ゼンメルワイスが提唱した当時は広く受け入れられていたものではなかった。

分娩の際に生じた傷口に細菌が入り、妊産婦に高い発熱をもたらす産褥熱（さんじょくねつ）という病がある。現在は細菌研究や消毒法が進歩したこともあり、それほど深刻な病気ではないが、細菌の存在が知られていなかったゼンメルワイスの時代には、妊産婦死亡の大きな原因となっていた。

ある時ゼンメルワイスは、同じ大学の産婦人科の中でも産褥熱の発生に大な差があることを発見し、何らかの微粒子が原因となっているのではないかという仮説を立てた。そして医師たちに手洗いを徹底するよう働きかけたところ、産褥熱の発生を激減させることに成功したのである。しかしゼンメルワイスの説は猛反発を受け、当時の学会は彼の論文を否定。彼は失意のうちに死んでいったと言われている。

一節によれば、彼の主張が受け入れられなかったのは「妊産婦が死亡する原因をつくっていたのは当の医者である」という事実を突きつけるものであったからだと言われている。いずれにせよ、いくらデータがより良い行動を示唆していたとしても、それが心理的・精度的に受け入れられる環境がなければ行動には結びつかないのだ。

残念ながら同じ警告は、現代の企業に対しても言うことができる。「今までこうしてきた」「それは前例がない」「部外者には分からない」などといった態度によって、データ分析の結果が拒絶されるという話は珍しくない。米バブソン大学のトーマス・ダベンポート教授は、技術的な整備を行うと同時に、社内環境の整備を行うことがデータ分析という行動を企業内で成功させる上で欠かせないと主張している。

またビッグデータは大容量データの収集・蓄積・分析が必要になることから、ある程度の大企業の方が有利になる可能性があるが、一方で大企業ほど過去の成功体験の束縛・組織の「サイロ化」といった傾向が存在しやすい。意外にビッグデータ活用で最初に成功を収めるのは、体力の面では不利なものの、様々なしがらみのない新興企業なのかもしれない。

人材不足の懸念

さらに問題視されているのが、誰がビッグデータ導入を推進して行くのかという点だ。同じく米マッキンゼーの調査によると、今後米国内だけでも、統計学に関する専門家が14万から20万人、さらにデータ指向の管理者が150万人不足すると予想されている[*14]。

いったい企業の中で、どのようなスキルを持つ人々が今後必要になるのだろうか。

クラウデラ社のジェフ・ハマーバッカーは、かつて所属したフェイスブック内の解析チームの典型的な一日は、「多段階の処理パイプラインをPythonで下記、仮説検定を設計し、統計ソフトウェアRを用いてデータサンプルの回帰分析を行い、Hadoopで大量のデータを扱う製品やサービスのアルゴリズムを設計して実装し、分析結果を明瞭かつ簡潔な方法で組織の他のメンバーと話し合う」という状況であったことを語っている[*15]。さらに彼は、こうしたタスクを遂行するスキルを備えた人物を表す言葉として、「データサイエンティスト」という肩書きをつくり出している。

こうした「データサイエンティスト」的技術者に加えて必要になるのが、ビジネス面からデータの価値を把握し、その扱いに関する戦略を練ることができる人物だろう。

ビッグデータの時代には、データの価値を把握することが非常に困難になる。一見無駄のように見えるデータも、新たな解析技術を導入すること、社内に隠れているデータと組み合わせること、あるいは外部から別のデータを取り寄せることで宝の山に変わるかもしれない。さらにどうしても自社内で活用できないデータであっても、他社に高値で提供することができる可能性が残されている（そのためにデータの標準化を進めておくという判断も要求されるだろう）。こうした可能性を正しく把握し、数値に変換できなければ、ビッグデータに対する投資を正当化することはできない。

米メタウェブ社に所属するプログラマーであるトビー・セガランは、編著『ビューティフルデータ』の中で、アマゾンの元チーフサイエンティストであるアンドレアス・ウェイゲンが発した言葉として、「このデータセットにはどんなすごい技術が使えますかとみんなよく聞くけど、手に入る最良のデータセットは何なのかを考えるべきだよ[*16]」というアドバイスを紹介している。技術面からビッグデータを支える専門家を揃える一方で、データそのものの準備に頭をひねる人物を確保しておくこと。ビッグデータに取り組もうとする企業は、この2つが要求されて行くに違いない。

さらに両者を仲介し、全体的な視野からビッグデータを統括する役職として、CDO（Chief Data Officer）的な存在が一般化するようになるかもしれない。あるいは企業トップに立つ人物は誰でも、ビッグデータに関する素養を身につけておく必要がある、などという時代が来る可能性もあるはずだ。

*****

米国の連邦政府には、政府としての「CTO」という役職が設置されている。それだけテクノ

ロジー、特に情報技術が社会に果たす役割が大きくなっているというわけだ。であれば近い将来、政府内に「政府CDO」という役割、つまり国家としてビッグデータとどう向き合うのかを考える責任者が置かれる社会が到来することも、あながち絵空事とは言えないだろう。

注釈

１２． New York Times, July 11, 2010, ‘Stop, Question and Frisk in New York Neighborhoods,’ http://www.nytimes.com/interactive/2010/07/11/nyregion/20100711-stop-and-frisk.html

１３． http://www.google.com/transparencyreport/

１４． McKinsey Global Institute, May 2011, ‘Big Data: The Next Frontier for Innovation, Competition, and Productivity’

１５．ジェフ・ハマーバッカー、『ビューティフルデータ』、オライリージャパン、2011年。

１６． Financial Times, October

おわりに

米ガートナーのシニア・バイスプレジデント、ピーター・ソンダーガードは「情報は21世紀における石油であり、分析はエンジンのようなものだ」と述べている[*17]。まさに情報は重要な資源であり、それを出来る限り有効活用すること、言い換えれば燃費が良くて馬力もある「エンジン」を用意することで、社会を一段と加速することができるだろう。またビッグデータ活用は社会全体に対してだけでなく、個人にも直接的な恩恵をもたらす可能性があることを見てきた。

しかし同じく解説してきたように、そこには様々なリスクが存在する。他の技術とは異なり、ビッグデータの場合には「何らかの情報漏洩が起きている」「不利益を被っている」などといったことをユーザーに気付かせぬまま悪用できるようになる恐れもあるのだ。石油と内燃機関が20世紀の社会を大きく効率化し、その一方で交通事故や環境破壊などの不利益をもたらしたように、ビッグデータもコントロールを誤ってしまえば、深刻な被害が発生しかねない。

しかしデメリットがあるからと言って、全てのメリットを放棄してしまう必要はないと筆者は考える。確かに不利益を被る可能性は否定できないが、より良い新薬の開発や社会インフラの整備など、従来の方法では難しかった価値を生み出す、しかも多くの人々に提供するということが実現される可能性があるのだ。

現在の状況を見ると、まだビッグデータの活用が技術的に始まったばかりであり、制度面・文化面での整備はこれからだと言えるだろう。安全性を維持した上で、さらなるデータの生成・共有を促し、データの流動性を高めるためには、技術よりもむしろ環境面での進化の方が重要になるかもしれない。

残念ながらビッグデータに関係する情報技術という点では、日本は欧米に先を越されてしまっているが、その活用という点では決して出遅れてはいない。モバイル端末や高速回線の普及、情報リテラシーを持つユーザーの存在など、他国に勝っている面も多い。ビッグデータの活用に向けた環境面でのロールモデルとして、世界に発信して行くチャンスも十分にあるのではないだろうか。

しかし日本社会はとかく「石橋を叩いて渡る」姿勢を取りたがる。だが私たちが相手にしているのは、これまでに無かった状況であり、全てを計画してから行動に移すことは不可能だろう。

開発問題の専門家、オーウェン・バーダーは、途上国支援を成功させるためには「より良い世界を描こうとしてはならない。そうではなく、より良いフィードバックが得られる仕組みを創らねばならない」と述べている。恐らく私たちにいま求められているのも、「完璧なビッグデータ・ビジネスモデル」や「完璧なビッグデータ社会」を描くことではなく、それが何をもたらしているかをすぐに把握できるような体制をつくり、トライ＆エラーで進んで行くことではないだろうか。

*****

本書に関するディスカッション、また関連事例のフォローを、フェイスブック上で行ってい

ます。ご興味のある方は、以下のアドレスからご確認下さい。

■フェイスブックページ「ビッグデータ社会の到来｜The Arrival of Big Data Society」
http://www.facebook.com/BigDataSociety

17, 2011, “IT chiefs must ‘lead from the front’,’
http://www.ft.com/intl/cms/s/0/829408e2-f8c8-11e0-ad8f-00144feab49a.html#axzz1bl78jrg8

小林 啓倫 （こばやし あきひと）

　株式会社日立コンサルティングの経営コンサルタント。1973年東京生まれ。筑波大学大学院卒。国内SI企業、外資系コンサルティングファーム、米国でのMBA留学等を経て、2005年より現職。新規事業の立ち上げ支援や市場調査、クラウド事業などを担当している。またPolar Bear Blog（http://akihitok.typepad.jp）、シロクマ日報（http://blogs.itmedia.co.jp/akihito）の2つのブログを運営するブロガーでもある。
著書に『災害とソーシャルメディア』『リアルタイムウェブ―「なう」の時代』（ともにマイナビ新書）、訳書に『「ツイッター」でビジネスが変わる！』（ディスカヴァー・トゥエンティワン）など多数。

POLAR BEAR BLOG
http://akihitok.typepad.jp/

奥付

ビッグデータ社会の到来

http://p.booklog.jp/book/38557

著者 : akihitok
著者プロフィール：http://p.booklog.jp/users/akihitok/profile

感想はこちらのコメントへ
http://p.booklog.jp/book/38557

ブクログのパブー本棚へ入れる
http://booklog.jp/puboo/book/38557

電子書籍プラットフォーム：ブクログのパブー（http://p.booklog.jp/）
運営会社：株式会社paperboy&co.