ビッグデータの正体　ビクター・マイヤー＝ショーンベルガー

検索語でインフルエンザ流行検知

人々がネットでインフルエンザ情報を探すときは、「咳の薬」や「解熱剤」といったキーワードで検索するはず、とグーグルのチームは推測した。しかし、それが何かはわからないし、そんなことにいちいち注意を払うようなシステムに設計されているわけでもなかった。グーグルのシステムは、各検索語の使用頻度と、インフルエンザ感染の時間的・空間的な広がりとの間の相関関係の有無を見ていただけだ。グーグルは、合計 4 億 5000 万にも上る膨大な数式モデルを使って検索語を分析し、CDC が提供している 2007 年、2008 年の実際のインフルエンザ症例とグーグルの予測を比較検討した。そこで彼らは大変なことに気付く。特定の検索語 45 個と、ある数式モデルを組み合わせたとき、グーグルの予測と公式データの間に高い相関関係が見られたのだ。

つまり、CDC と同じようにグーグルも**インフルエンザがどこで流行しているのか特定できる**ことになる。両者に決定的な違いがあったとすれば、グーグルは 1～2 週遅れではなく、**ほぼリアルタイムに特定できた**点だ。

ビッグデータの時代には、暮らし方から世界との付き合い方まで問われることになる。特に顕著なのは、相関関係が単純になる結果、社会が因果関係を求めなくなる点だ。**「結論」さえわかれば、「理由」はいらないのである。**過去何百年も続いてきた科学的な慣行が覆され、判断の拠り所や現実の捉え方について、これまでの常識に疑問を突きつけられるのだ。

例えば、膨大な電子カルテのデータから、「オレンジジュースとアスピリンの組み合わせで癌が治る」ことが言えるなら、正確な理由はどうあれ、この組み合わせが癌に効くという事実のほうがはるかに重要となる。航空運賃の決まり方など詳しく知らなくても、航空券の買い時さえわかれば財布にやさしい。それで十分だ。ビッグデータの世界では、ある現象の理由を何が何でも知る必要はない。**データがすべてを物語っているからだ。**

データ収集前に仮説をいくつも考え、これを検証する分析作業に追われる必要もない。データに語らせるだけで、想像もしなかったような関係性が浮かび上がってくる。

目の前で起こっている現象のほんの一部だけでなく、大部分あるいは全体を取り込んだ包括的なデータ集合が手に入るなら、個々の測定値の良し悪しにいちいち悩む必要もない。

だから製造の現場では、**無線センサーが工場に次々に導入**されている。石油大手 BP のチェリーポイント**製油所**（米ワシントン州ブレイン）では、無線センサーが施設の至るところに張り巡らされ、この見えない網の目がリアルタイムに大量のデータを取り込んでい

る。高温になる装置や電気系統が密集する過酷な環境だけに、測定値が狂い、データは乱雑化しやすい。しかし、膨大なデータが生み出されているため、多少の不良データが発生しても問題にはならない。配管の内圧は一定間隔ではなくほぼ連続的に測定しているから、原油の種類によって配管の腐食が早いかどうかもわかる。**従来ならそうした変化は特定できなかったから、事前対策の打ちようもなかった。**

毎月、米国労働統計局からはインフレ率の計算に使われる**消費者物価指数**（CPI）が発表されている。投資家や企業に欠かせない数字だ。米連邦準備制度理事会では、これを基に利上げ・利下げを決定している。企業もインフレ率に応じて賃上げを決める。連邦政府は消費者物価指数を基に社会保障給付額や国債の利回りを調整する。

その基となる消費者物価指数の算出に当たり、労働統計局は、全米 90 都市の小売店や企業を対象に、数百人もの職員が日々、電話、ファックス、直接訪問による聞き取り調査を実施している。トマトの価格からタクシー料金まで実に 8 万点の価格が報告書としてまとめられる。報告書作成にかかる費用は年 2 億 5000 万ドル（**約 250 億円**）ほどだ。大規模な調査にしては、データはきれいで、きちんと整理されている。

ただし、データが出そろうころには、すでに数週間も古い内容になっている。2008 年の金融危機で明らかになったように、致命的な遅れになりかねない。政策決定者はいち早くインフレ率の数字をつかんで的確な対策を打ち出す必要があるが、従来のような標本抽出と精度第一主義の手法では、そうもいかない。

整然と定義
された情報の階層構造を前提とするレコードや事前定義フィールドといった古い原則とは決別したのだ。データベースを利用する際の共通言語としては、長らく **SQL**（構造化照会言語）が主役だった。いかにも融通の利かなそうな名称だ。だが、最近では、「**NoSQL**」なるものへと移行が進んでいる。事前定義のレコード構造を使わないデータベースだ。

種類やサイズがばらばらのデータにも寛大で、検索も問題ない。構造が乱雑でも構わない代わりに、データベースの設計上、処理能力や記憶容量はこれまで以上に必要となる。もっとも、記憶装置や処理能力のコストが劇的に安くなっていることを考えれば、決して呑めない条件ではない。

膨大なデータの処理に強いシステムの一例が、グーグルが開発したアルゴリズムである MapReduce 技術だが、これのオープンソース版に当たるのが Hadoop で、ずいぶん人気を博している。実はその人気も、乱雑さへのシフトが背景にある。

Hadoop では、データを小さなまとまりに分け、別のコンピュータに移す。機械はいつか壊れるという前提に立っているから、内部の機能にはすべて予備が用意されている。また、データは雑で、きれいに整理されていないことも前提にしている。実際、データが大

きすぎて、処理前に整理などできないとわかっているのだ。

従来のデータ分析では、初めにデータの「抽出・変換・挿入」と（英語の頭文字を取って「ETL」とも）呼ばれる操作で前処理されてから、ようやく分析者のもとにデータが送られてくる。ところが Hadoop では、そんな厄介な操作は不要だ。そもそも、データ量がとてつもなく膨大なので、データを分析者のもとに送ること自体、不可能と割り切っている。だから逆に分析者がデータのある場所に（ネットワーク上で）出向いて分析処理を実行するのである。

このようにして得られた結果は、**リレーショナルデータベースほどの精度はない**。宇宙ロケットを打ち上げたり、銀行口座明細を保証したりするほどの信頼性はない。だが、それほど重要度が高くない用途はたくさんあり、超高精度の答えが不要であれば、圧倒的に高速に作業を実行できる。例えば、顧客を細分化して、特別な販売キャンペーンを告知するといった用途だ。クレジットカード会社の Visa では、2 年分の記録に相当する約 730 億件の取引の処理にこれまで**1 カ月**かかっていたが、Hadoop を使って同じ処理を実行したところ、**わずか 13 分**で完了した。本格的な会計処理に使うのは難しいかもしれないが、少々の誤りが許容される案件であれば、並外れた能力を発揮する。

次にアマゾンが悩んだのは、提示する内容だ。専属の書評委員による書評か、それともコンピュータがはじき出した顧客別のおすすめやベストセラーリストか。極論すれば、書評委員の言葉を信じるか、蓄積されたクリックの〝声〟を信じるかだ。

文字どおり人間とマウスの戦いだ。

リンデンは、書評委員の推薦から販売につながったケースとコンピュータの推薦から販売につながったケースとを比較した。差は歴然だった。データから導き出したコンテンツのほうが、100 倍も大きな売上げを生み出していたのだ。

ある客がヘミングウェイを読んだ後になぜ F・スコット・フィッツジェラルドを買いたいと思ったのか、コンピュータは知る由もない。が、それは重要ではなかった。ともかく売れたことは事実である。やがて人間の手による書評がオンラインで公開されるたびに、書評委員らに正確な売上げデータが突きつけられた。ついに書評チームは解散を余儀なくされる。リンデンは「書評チームが負けたことはとても残念だった。しかしデータは嘘をつかない。コストも非常に高かった」と語る。

現在、アマゾンの売上げ全体の 3 分の 1 は、この「おすすめ」とパーソナル化のシステムから生み出されているという。アマゾンの台頭で大型書店や大型 CD・レコード店が廃業に追い込まれただけでなく、顔の見えるサービスが強みと自負していた地域密着型の書店も次々に店を閉めている。それほどまでにリンデンの技術はオンライン販売の世界に革命をもたらした。

欧米では、結婚や出産の際に自分が希望する贈り物を記載して友人らに公開する「おねだりリスト」（「ギフトレジストリー」とか「**ウィッシュリスト**」などと呼ばれ、ショップがリストを預かる）という習わしが昔からある。友人らはショップに出向いてリストを眺め、

自分の予算で買えそうな贈り物を選ぶ。購入された商品はショップがリストから消去する仕組みである。

長年にわたって経済学者や政治学者は**幸福と所得との間に直接的な相関があると考えていた**。平均的な人間であれば、所得が増えると幸福感が高まるというわけだ。しかし、データをグラフ化し、統計ツールで線形の相関が見られた部分には、もう少し複雑な力学が働いていることがわかった。年収が一定の水準以下の場合、所得が増えるたびに幸福度が大きく高まるが、**一定以上の年収になると所得が増えても幸福度はあまり変化しなかった**のである。つまり統計分析では直線グラフになると予測されたが、実際にグラフに描けば曲線の非線形になるのだ。

これは政策立案者にとって大きな発見だった。線形の相関ならば、全体の幸福度を高めるためには、全員の所得を増やすことに意味があった。ところが、非**線形の関係が見つかってしまった以上は、貧困層の所得増に特化すればよい**ことになる。そのほうが費用対効果も大きくなることは、データがはっきり示している。

ビッグデータは、物事の理解の仕方や考え方を大きく変える。スモールデータの時代には、物事の仕組みについて仮説を立ててから、データを収集・分析して検証していた。これからは、仮説ではなく、豊富なデータを使って理解を深めることになる。そして仮説主導型からデータ主導型へと重心が移れば、理論不要論が飛び出しても不思議ではない。

今や位置を特定できるシステムは多数あり、GPS はそのうちの 1 つにすぎない。すでに中国や欧州ではライバルとなる衛星システム計画が進行中だ。GPS は屋内や高層ビルの谷間ではうまく機能しないが、携帯電話基地局や無線 LAN スポットの信号強度の三角測量で位置特定の精度がさらに高まっている。グーグルやアップル、マイクロソフトといった企業が GPS を補完する独自の位置情報システムの開発に取り組んできたのも、そのためだ。グーグルのストリートビューは、街の写真を撮影する際に**近隣から電波が漏れ出ている Wi-Fi ルーターの情報も収集**していたし、**iPhone は位置情報と Wi-Fi データを収集してアップルにこっそり送る**スパイまがいの機能**が入っていた**（グーグルのアンドロイドやマイクロソフトの携帯向け OS も同様のデータを収集していた）。

1990 年代末、インターネットは急に荒れ始め、居心地の悪い環境に変わっていく。原因は「**スパムボット**」と呼ばれるメールアドレス検索ロボットだ。ロボットがオンライン掲示板などに勝手に入り込み、メールアドレスらしき情報をかき集めていく迷惑行為だ。

2000 年、大学を卒業したばかりの**ルイス・フォン・アーン**という若者が問題解決の

アイデアを温めていた。要は「ロボットではなく人間であることを利用者自身に証明させればいい」わけだ。そこで、フォン・アーンは、人間には簡単でも機械には難しいこととは何かを考えた。

やがて入会時やログイン時に、ぐにゃつと曲がって読みにくい文字列をユーザーに入力させるというアイデアを思いつく。これは実に効果的だ。人間なら確実に読み取って正解を入力できるが、コンピュータにはお手上げなのだ。ヤフーはこの手法を採用したところ、たった一晩でスパムボット問題は解決に向かった。フォン・アーンはこの技術を**「キャプチャ」**と命名した。それから 5 年後には、毎日 2 億件近い文字列が入力されるまでになった。

・・・

ただ、フォン・アーンはすっきりしなかった。毎日何百万人もの人々に無意味な文字の羅列を入力させても、そのままゴミになるだけだ。どうせなら人々の入力行為を有意義に利用できないかと考えたのである。もっと生産的な方法を模索していたところ、キャプチャの後継技術を思いつく。その名も「**リキャプチャ**」である。

リキャプチャでは、2 つの単語を入力させる。1 つめは従来のキャプチャと同じで、本当に人間が利用しているかどうかを確かめるための単語だが、2 つめの文字列にちょっとした細工が仕掛けてある。この単語、コンピュータによる文書読み取り（OCR）業務でうまく読み取れなかった文字列なのだ。OCR では、原稿の文字にかすれがあったり、滲んでいたりすると、読み取りミスが発生する。**コンピュータでお手上げだった単語を拾ってきて、リキャプチャで人間に読み取り作業をさせれば、正しい読み取りが可能なうえに、人間かどうかの認証もできるとあって一石二鳥**。おまけに作業報酬もタダと来ている。

その価値は計り知れない。実際に入力作業に人を雇えばとてつもないコストがかかる。

マイクロソフトでは、ここ加年ほど同社の主力ワープロ「MS ワード」用に強力なスペルチェッカーの開発に取り組んできた。ユーザーが文字列を入力すると、内蔵辞書と照らし合わせる仕組みで、辞書は随時更新される。ただし、辞書に載っている単語が正しい単語ということになるから、単語の変化形であっても、辞書に載っていないものはミススペルと判断し、正しい綴りに訂正しようとする。また、辞書の編纂・更新には膨大な手間がかかるため、MS ワードのスペルチェッカーは主要言語だけに対応していた。それでも、開発と保守には莫大なコストがかかる。

一方、**グーグルが提供するスペルチェッカー**は、まず間違いなく最高の完成度を誇る。しかも、基本的にあらゆる言語に対応している。常に改良され、新語も追加されている。実は、世界中の人々が検索エンジンを毎日使用した副産物なのだ。だから「iPad」の入力ミスも含まれているし、「Obamacare」（オバマケア＝オバマ政権が進める医療保険制度改革の通称）のような新語も当然ある。

さらに言えば、グーグルのスペルチェッカーは開発費もかかっていない。グーグルの検索エンジンでは、毎日 30 億件の検索が実行されている。その中に含まれるミススペルを再利用しているのである。

どういうことかというと、巧みなフィードバックの仕組みが用意されていて、ミススペルしても「ユーザーが本来入力したかった単語は何か」をシステムに教え込む仕組みができている。ご丁寧にもユーザー自身が正しい答えをグーグルに「指導」することさえある。それは、検索結果の表示画面の上部に現れる「もしかして〇〇〇？」の質問だ（〇〇〇は正しいと思われる語句）。ここで正しい候補をクリックすると、その語句で再検索される。仮にクリックしてもらえなくても、その後ユーザーが実際に移動した先のサイトがわかれば、正しいスペルを確認できる。そのサイトについて、正しいスペルとミススペルのそれぞれの相関を取れば、どちらが正しいか一発でわかるからだ。グーグルのスペルチェック機能は絶えず改良されているから、正しい検索語句を入力しなくても、グーグルが適当に処理してくれるほどだ（最近は「もしかして」機能が強化され、正しい語句で検索した結果を最初から表示することもある）。

===== Web =====
judgement
【名】〈英〉 ＝ judgment
==========

会計基準に基づく価値（63 億ドル）と当初の時価総額（1040 億ドル）はあまりにかけ離れている。その差は何か。実は、企業価値は「帳簿上の純資産額」（大半が有形固定資産の価値）で判断する方法が広く普及しているため、真の価値が適切に反映されていないのだ。一方、時価総額は企業が株式市場で調達した額だ。事実、**帳簿上の純資産額と「時価総額」の差は年々広がっている**。2000 年には米国上院が財務報告規則の近代化に向けた公聴会を開催したほどだ。何しろ現行の規則は 1930 年代に作られたもので、当時は情報ビジネスなど皆無に近い時代だった。この問題の影響は、企業財務にとどまらない。企業価値を適正に評価できないと、ビジネス上のリスクや市場の乱高下につながる。
純資産額と時価総額の差額部分は「無形固定資産」に当たる。米国では 1980 年代半ばに上場企業全体の時価総額の 40%ほどが無形固定資産だつたが、2000 年代の幕開けごろには 4 分の 3 を占めるまでに比重が大きくなった。無形固定資産には、ブランドや人材、戦略、有形ではないが、形式的に財務会計の対象となる資産がすべて含まれる。
最近では企業が保有・活用しているデータも徐々に無形資産と考える傾向が強まっている。

100 年前は基本的な計算能力が必須だったように、これからは数学や統計学、それにプログラミングとネットワークのちょっとした知識が　〝現代の読み書きそろばん〟　になる。

多くの企業にとってビッグデータが競争力の源泉になるにつれて、産業界全体の構造が大きく変わろうとしている。しかし、その見返りの大きさは企業ごとに違う。しかも、勝者と言えるのは大手と小規模の企業で、**その割を食うのが、中間にいるプレイヤー**だ。

アマゾンやグーグルといった超大手は今後も伸びていく。しかし工業化時代と違って、物理的な規模が競争優位になるわけではない。データセンターのような巨大なインフラは重要ではあるが、最も重要な条件ではない。デジタル情報を保管する記憶装置やデータ処理能力はリースで安く利用でき、インフラの追加もあっという間だから、コンピュータの処理能力や記憶装置の量は必要に応じて自由に調整できる。かつては固定費だったものが変動費に変わるため、これまで大手だけが享受してきたインフラの規模の優位性は薄れていく。

規模は依然として重要だが、規模の種類がシフトしている。大切なのは、データの規模だ。つまり**大量のデータを保有**すれば、そこから手軽に得られるものも増える。

米国の
将官のうち、戦死者数が戦争の進捗状況を計る尺度として有効だったとする回答はわずか2%にとどまった。また、およそ3分の2の将官は、数字が誇張されることなど日常茶飯事だったと指摘する。「デタラメ。まったく意味なし」、「多くは真っ赤な嘘。マクナマラのような連中が驚くほど食いついてくれるので、大半の部隊はひどく水増しして書いていた」など生々しい回答が並ぶ。

フォード工場の従業員がエンジン部品を川に捨ててごまかしていたように、戦場の下士官たちもまた上司に優秀な数字を報告していたのだ。つまりは、**上司が聞きたがっている数字を伝えたまで**だ。マクナマラらは上がってくる数字に疑いを抱くこともなく、ただただ盲信した。きれいなオフィスで、表に星印を書き込むことで戦況を理解した気になっていたのだろう。

ベトナム戦争中の米軍によるデータの使用、濫用、誤用は、情報が乏しいスモールデータ時代のいい教訓となった。世界がビッグデータ時代に向かっている今、まさに心に細田めておきたい教訓だ。

2009年、グーグルのトップデザイナーだったダグラス・ボウマンは、こうした何でも数値化する会社の体質に嫌気がさして辞表を叩きつける。「最近も線幅は3ピクセルがいいのか、4ピクセルかで議論したばかりだよ。意見があるなら、それが正しいことを証明しろと必ず迫られる。そういう環境ではやっていけない」とブログで辞任の経緯を報告。「**エンジニアがたくさんいる会社は、エンジニアリングの発想で問題を解決しようとする。どんな判断も、すべて単純な論理の問題に還元されてしまう。そうやって出てきたデータを頼りにあらゆる決定が下されていて、会社は麻痺している**」と批判したのだった。

アルゴリズミスト