ビッグデータと人工知能の近未来

筑波総研株式会社　顧問（筑波大学名誉教授）　高　木　英　明

今、あなたの買いたい物は**アマゾン**（Amazon）が教えてくれ、あなたは誰の友達かを**フェイスブック**（Facebook）が教えてくれ、あなたの評判は**グーグル**（Google）が教えてくれる。そのうち、外出する時に傘を持っていくべきかどうかを傘立てが教えてくれ、夕方までに歩くべき歩数を腕時計が決め、体脂肪を考慮した食材を冷蔵庫が発注するようになるだろう。あなたは学力に合った大学に現役で合格し、マッチングされた良き伴侶を得て、最適な年齢で子供を持ち、可もなく不可もなく職歴をこなした後、充実した終末期医療を受けて長寿を全うする。これが、あなたから発信されるすべての情報が自覚の有無にかかわらず収集され、あなたを取り巻く環境情報と照合されて、あなたに**リコメンド**（推奨）される最適の人生である。また、あなたが経営する会社では、精度の高い顧客管理に基づく需要予測、ロボットによる無駄のない生産工場、自動化された経理システムで、人事管理に煩わされることなく、商品の製造販売ができる。このようなことが、近い将来、ヒト・モノ・企業・政府・社会から出るあらゆるデータを集めた**ビッグデータ**（big data）を**人工知能**（Artificial Intelligence, **AI**）が処理することにより、起こり得ると考えられている。人工知能は「超スマート社会を実現する技術」か、「大きなお世話」か、はたまた「人間の感動や尊厳を脅かす悪魔の技術」であろうか？

本稿では、ビッグデータと人工知能について、大きな期待とともに漠然とした不安も抱いている読者を想定し、近未来における人工知能に関する最近の話題を整理して、**秋の夜長**に供したい。

1. ビッグデータとは？

最近になってビッグデータの効用が現実になった技術的な理由は、大量のデータを瞬時に処理できるようになったハードウェア技術（電子工学）とソフトウェア技術（情報科学）の進歩であるが、根本的な理由は、文字どおり「桁違い」の量のデジタルデータが生み出されるようになった社会の変化である。ビッグデータとは、単に「大量のデータ」を意味するのではない。その特徴は次の**4つのV**で示される。

・**大量性**（Volume）：どのくらい大量のデータをビッグデータと呼ぶのかについて、一定の数があるわけではない。各時代において、その時代のコンピュータでは簡単に扱えないほどのデータ量がビッグデータと呼ばれてきた。調査会社IDCの2012年12月の発表によれば、2012年までに作成または複製されたデータ量は2.8ゼッタバイトであり、2020年には40ゼッタバイト（世界の全人口76億人の1人当たり5,250ギガバイト）に拡大すると見られている（データ量の単位については表1を参照）。多くのデータは、ビジネス分野、医療分野、センサー等から生成されると予想されている。

■表1　データ量の単位

1 MB（メガバイト）	10^{6}（百万）
1 GB（ギガバイト）	10^{9}（十億）
1 TB（テラバイト）	10^{12}（一兆）
1 PB（ペタバイト）	10^{15}（千兆）
1 EB（エクサバイト）	10^{18}（百京）
1 ZB（ゼッタバイト）	10^{21}（十垓）
1 YB（ヨタバイト）	10^{24}（一秭）

・**高速性**（Velocity）：データが生成・収集・処理される速さと頻度のこと。24時間絶え間ない金融取り引き、高速道路のETCゲート通過、フェイスブックのコメント、監視カメラ等、多くのデータは発生すると同時に処理されるリアルタイム性が要求される。日本の理化学研究所（神戸市）に2012年に完成したスーパーコンピュータ「京（けい）」の名前の由来は、1秒間に1京（10^{16}）回の浮動小数点演算（10ペタフロップス、70億人が1秒に1回計算すると17日間かかる計算量）という計算能力である。

・**多様性**（Variety）：1980年代になって、それ以前は数字と文字しか扱うことができなかったコンピュータが音声や画像を処理するようになったとき、**マルチメディア**と呼ばれた。ビッグデータで言う「多様性」は、その意味ではない。従来は、文字・数字とマルチメディアデータが混在していても、その形式と内容は規定のものであり、それらは**構造化データ**である。ビッグデータには、例えば、ヒトがツイッターに投稿する自由意見や画像のように、文字や符号の意味や区切りや長さが処理前に分からないものや、人体に取り付けたセンサーやスマホが出す位置情報のように機械が勝手に出す情報などの**非構造化データ**が含まれ、それらに対してエクセルの表計算やSQL（構造的照会言語）による検索を行うことはできない。世界中のデータの約80％は非構造化データであると言われている。非構造化データを今のコンピュータで処理するためには構造化しなければならない。グーグル社がその技術ハドゥープ（Hadoop）を開発して、ビッグデータの処理が可能になった。

・**正確性**（Veracity）：コンピュータが少量のデータしか扱えなかった時代には、例えば世論調査やアンケート調査のように、統計学を応用して、膨大な母集団からの**標本抽出**により母集団の特性を推定・仮説検定した。ビッグデータ技術により、すべての個体についてデータを集める**全数解析**を行うことができるようになった。その結果、「各データが曖昧でも、膨大な量が集まれば精度が高まる」という原理が実証された。

さらに、近代以降の科学技術は「原因があるから結果がある」という**因果律**を追求してきたが、ビッグデータにより、ソリューションにはデータ間の**相関関係**があれば十分となり、「答が分かれば理由は不要」という時代が来ている。

ビッグデータ処理による衝撃的な発見例として、以下のようなものが有名である。

・1990年頃、プリンストン大学のオーリー・アッシェンフェルター教授は、ボルドーワインの評価額が、収穫前の冬の降雨量、育成期の平均気温、収穫期の降雨量を説明変数とする回帰分析を使って予測できることを示し、ワイン鑑定の専門家たちと大論争になった。

・アメリカの大手スーパー「ウォルマート」では、顧客の購買データを分析して、おむつを購入した人は缶ビールも同時に買う傾向を発見した。

・グーグル社では、ソーシャルメディアに現れた検索語の頻度分析から、医学・医療データの調査もせずに、各地でのインフルエンザの流行を公的医療機関よりもはるかに速く、高い精度で予測することができた。

・『ヤバい経済学』の著者スティーヴン・レヴィット教授は、大相撲の千秋楽に7勝7敗と8勝6敗で対戦した2人の力士の対戦における7勝7敗の力士の勝率79.6%は、同じ力士どうしの全対戦での勝率48.7%よりも異常に高いことを指摘した。

以上のように、ビッグデータを特徴づける4つのVのうち、大量性と高速性は、コンピュータ技術のこれまでの発展の延長線上にあるが、多様性と正確性は、別の次元への展開である。集めただけではゴミの山であるビッグデータを分析して、そこに宝を見つけることを**データマイニング**（データの山の採掘）という

2. 第3次AIブームの到来

最近、グーグル社の自動走行車やアルファ碁など、耳目を奪う応用成果が華々しい人工知能であるが、人工知能の研究は、これまで一直線に発展してきたのではない。以下のように、2回のブームと「冬の時代」を経て、現在は「第3次AIブーム」を迎えている。ブームのときは、

世の中が「人工知能がもうすぐ出来る」と浮かれ、企業の投資が殺到し、多額の国家予算も投入されるが、期待された成果は出ずにブームが去り、その後に訪れる冬の時代にブレイクスルーが起きて、次のブームが始まるという循環が繰り返された。

・**第1次AIブーム**（1950年代後半～1960年代）

「世界最初の電子式コンピュータ」と言われるENIACが完成した1946年から僅か10年後の1956年夏にアメリカ東部ニューハンプシャー州のダートマス大学で開催されたブレインストーミングが、AI研究の嚆矢である。その会議において、単に計算をするだけでなく、人間のように考える（起こり得る場合の探索や推論の手順を実行する）コンピュータを「人工知能」と呼ぶことにした。その会議の参加者がリーダーとしてその後のAI研究を牽引することになった。しかし、パズルはともかく、現実の課題を解くことはできず、ブームは下火になって、1970年代に第1次冬の時代を迎えた。

・**第2次AIブーム**（1980年代）

1980年代になると、専門分野の知識を蓄えたエキスパートシステムが人工知能研究の本命になった。そのために専門家から聞き出したIF-THEN-ELSE（もし○○なら□□、そうでなければ△△）型の推論規則をコンピュータに数多く蓄える。例えば、スタンフォード大学で開発された「マイシン」は、500個の推論規則を備え、血液疾患の患者に問診を繰り返すことで、感染した細菌を特定し、抗生物質を処方するという専門医の診断を肩代わりすることに成功した。

しかし、専門家から聴取する規則の数は膨大であり、全てが整合的とも限らない。また、人間にとっては常識的な諸概念間の関係は無数にあり、とても書き切れるものではないことが認識された。機械翻訳でも、それに由来する文脈の不確定さを解消できなかった。こうして1990年代にはこのブームも終わって、1995年頃には、再び冬の時代に入った。

日本の通商産業省（現在の経済産業省）が主導した国家プロジェクト「第五世代コンピュータ」（1982年から10年間）はこの時期に咲いたあだ花である。電子技術総合研究所（現在の産業技術総合研究所の一部）を中心に（財）新世代コンピュータ開発機構（ICOT）が設立され、並列推論マシンが開発されたが、当時はAI型の応用ソフトがなく、このプロジェクトが産業に直接的影響を与えることはなかった。

なお、2011年にアメリカのクイズ番組「ジョパディ！」でクイズ王に勝って有名になったIBMの質問応答システム「ワトソン」は、大量の知識（テキストデータ）を蓄えて構築したビッグデータの威力を印象づけたが、「意味ネットワーク」という第2次ブームの技術を使っている。

・**第3次AIブーム**（2010年以降）

1990年代にパソコンが接続されたインターネット上にワールドワイドウェブが構築され、画面上に使い勝手の良いブラウザが登場して、大量のデータが流通し始めた。ウェブ上で一般消費者が情報の発信源となった2005年頃を**ウェブ2.0**という。2010年頃からスマートフォン（スマホ）が急速に普及し、フェイスブックやツイッターに数億人が書き込むようになった。最近ではモノに付けられたセンサーもネットにつながる**モノのインターネット**（Internet of Things, **IoT**）の時代になり、人類が有するデータ量が飛躍的に増大している。

ビッグデータを背景として始まった第3次AIブームの基礎技術は**機械学習**（マシンラーニング）である。機械学習とは、膨大な事例を集めておいて、起こる頻度が高い場合を「機械的に」に当てはめる方法である。例えば、翻訳においては、文法や意味構造を考えずに、訳語として最も多く使われた例を対応させる。医療では、患者の症状に対して、最も多くの症例が当てはまる病名を診断する。そのような事例適応を重ねる度に、人間の脳の神経回路を真似て作った**ニューラルネットワーク**の特性を強化する「学習」により、コンピュータに知識が蓄えられる。初期の機械学習は、人間が適用分野を決めて、判定の訓練データを与えていた**教師あり学習**であった。今では、コンピュータが自ら与えられ

たデータに内在する構造と特徴量を抽出する**教師なし学習**が一部実現している。これが多層構造のニューラルネットワークを用いることから、**深層学習**（ディープラーニング）と呼ばれる現在の最先端AIである。こうなると、人間はもはやAIが示す解答の理由を知ることはできない。昨年10月に世界最強と言われる韓国の李世乭九段に勝った**アルファ碁**のソフトウェアが深層学習から繰り出す手は、既に人間が理解できない域にあると言われる。

意外と（と言うか、今となっては、寧ろ自然の理に適って）、AIは人間の幼児の脳の発達に似た成長をしていると思われる。「三つ子の魂百まで」と言われるように、幼児期の知能の発達はすさまじい。このとき、幼児は推論規則を1つずつ教えてもらっているのではなく、五感を通して周囲から無数の範例を吸収し、脳細胞が発育している。幼児が意味も分からずに脳に蓄えた膨大な範例がいつしか知能と感性に変わる過程は実に不思議である。深層学習もこのプロセスをなぞっているのだが、人間とは別の方法で抽出する特徴量から作られる知能が、人間の知能とは似ても似つかぬ代物になるのは必然である。

3. シンギュラリティ（技術的特異点）は起こるか？

人工知能がどんどん発達して、クイズ王や囲碁・将棋の名人に勝ったり、東大入試に合格したり、自動運転タクシーで病院に行けるようになっても、すべて想定内の進歩と言える。これまでも、多くの科学技術は人間の能力を凌駕してきたが、人類はそれを利用して（武器への転用と葛藤しながらも）繁栄を続けてきた。人工知能やロボットは確かにすばらしく、人間はそれらも取り入れて、スマート社会を謳歌することができるだろうと思っていたら、実はそうではないかもしれないという話が出てきた。

初期の機械学習の時代までの分野特化型の「狭い人工知能」に対し、深層学習がその可能性を開いた**汎用人工知能**に対する関心と恐れが高まっている。アメリカの発明家・未来学者のレイ・カーツワイル氏は、人工知能が自分の能力を超える人工知能を自ら生み出せるようになる臨界点として、**シンギュラリティ**（技術的特異点）の到来を予言した。このとき、人工知能は人間の制御から解放され、独自の進化を始める。彼は、半導体の集積度が18カ月で2倍になるという**ムーアの法則**等を援用して、「技術革新の成果は次世代の進歩にフィードバックされるので、技術は加速度的に進歩する」という**収穫加速の法則**を提唱し、シンギュラリティが起きる時期を**2045年**と予測している。これは遠い未来ではない。読者の大半は、存命どころか、まだ定年退職前であろう（その職業がまだ存在していればの話だが）。

但し、シンギュラリティが起きるのは21世紀後半と言う人も、起きないだろうと言う人も、また、起こしてはいけないと主張する人もいる。深層学習はまだ海のものとも山のものともつかない。テレビ、パソコン、インターネット等は、アイデアが社会を変革した技術になるまで30～40年を要した。60年経っても実用化できないAIや原子力発電は人類にとって何か本質的な欠陥を孕んだ技術であるのかもしれない。

シンギュラリティが起きる時期はともかく、その後の世界がどうなるかは誰も予測できない。イギリスの宇宙物理学者スティーブン・ホーキング博士は「人工知能の成功は人類の歴史において最高の出来事になり得るだろう。しかし、もし危機をいかにして避けるかを知ることができなければ、残念ながら人類の歴史の最後の出来事になってしまうかもしれない」（堀浩一東京大学大学院教授の訳）と警告している。われわれの子や孫たちのために、AIにせよ原発にせよ、超先端技術の開発には人類の叡智と良心を信頼するしかない。

参考文献

1．V. M. ショーンベルガー・K. 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、2013年.

2．松尾豊『人工知能は人間を超えるか：ディープラーニングの先にあるもの』、KADOKAWA、2015年.