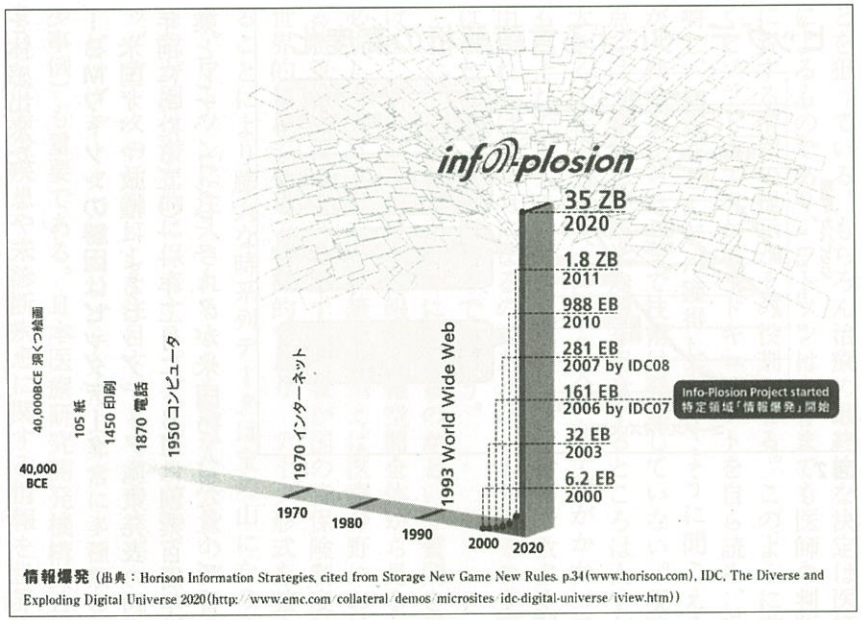
ビッグデータ

喜連川 優

ビッグデータと情報爆発

二〇一二年三月米国は予算額二百Mドルのビッグデータ・イニシアティブを発表した[1]。以降、ビッグデータなる言葉は大変判り易い表現であることもあり、瞬く間に世界中に報道され日本においても多くの人々が耳にするに至った。大量のデータが科学を、産業を、そして社会を変革するというメッセージは新鮮であった。若干手前味噌になるやもしれないが、我が国では二〇〇四年、筆者らは文部科学省科学研究費特定領域研究に「情報爆発（略称）」なる研究を申請し、幸いにも採択され二〇〇五年より五年半に亘る大型研究を推進した[2]。**図1**に示すように、とりわけ二十一世紀に入り人類が創出する情報が爆発的に増大している。この情報爆発なる現象は人類が初めて遭遇するものであり、爆発する情報を人々に如何なる新たな機会を与え、如何なる問題を生むかを研究者が同定し、該機会の活用と該問題の克服に挑戦することが必須と考えた次第である。皮肉にも情報爆発なるプロジェクトが終了して一年後に登場した言葉がビッグデータであった。情報とデータという言葉の違いがあるものの、情報もデータもいずれも爆発的に増えており、二つの用語はいずれも「大量データ・情報の積極的活用による



情報爆発（出典：Horison Information Strategies, cited from Storage New Game New Rules. p.34(www.horison.com), IDC, The Diverse and Exploding Digital Universe 2020(http://www.emc.com/collateral/demos/microsites/idc-digital-universe/iview.htm)）

図1

価値創出」を指すと言える。

膨大なデータが注目されるに至った背景には、ウェブやソーシャルメディアにより膨大なテキスト、画像、映像、音声情報がサイバー空間上で生み出され、加えて、多様なセンサー技術並びにそのシステム化技術の進歩によってあらゆるものが可観測となり、膨大なセンサーデータが作りだされるようになったことが挙げられる。またストレージコストが大幅に低廉化し大量のデータが保存可能になり、更にクラウド技術の進展により膨大なデータの処理も可能となったことの恩恵も大きい。

情報爆発時代における言語処理

情報爆発プロジェクトを推進していた当時は検索エンジンに関する研究が注目されており、テキスト情報爆発に対する言語処理は重要な課題の一つであった。従前は新聞記事からテキストコーパスを作成していたが概ね二千万文程度であったものを、新たに登場したウェブに着目し、その規模を百五十億文以上にする挑戦を行ったところ、飛躍的に言語処理の精度が向上した。**図2**に京都大学の黒橋教授の研究成果を示す。アルゴリズムを精緻化することはもちろん大切であるが、学習するデータ量を増やすことの圧倒的な価値観

を体感出来た。

IBMワトソンの勝因はビッグデータ

米国オバマ政権によるビッグデータ施策発表の約一年前にあたる二〇一一年二月にIBMは創業百周年事業として、コンピュータが米国の人気クイズ番組

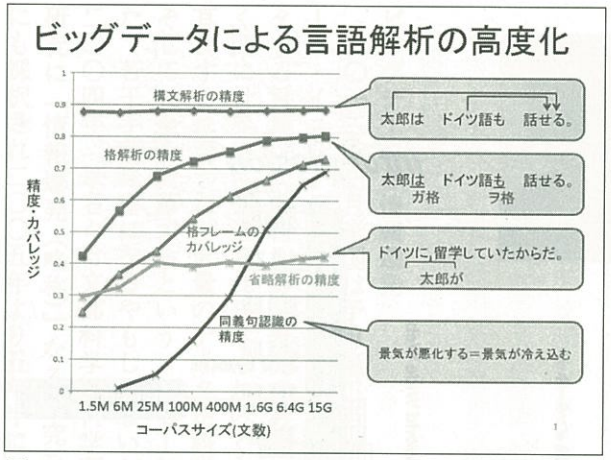


図2

「Jeopardy!」においてチャンピオンに挑戦するというイベントを企画し、ワトソンなる質問応答システムソフトウエアが見事にチャンピオンを打ち負かした[3]。クイズの多様な問題に答えるには多くの常識の源泉となる情報を獲得する必要がある。当時ウィキペディア等のソーシャルメディアが急速に発展しつつあり、サイバー空間上にある多様な情報が利用された。推論機構の貢献も大きいが、この膨大な情報の活用がワトソンのパワーの源泉であり、これが無ければ勝利は無かったと言える。人工知能（AI）ブーム再来の火付け役になったワトソンは、一人の人間には覚えきれない知識を持つ超人にも見え、大きな驚きを与えた。

このようにビッグデータは、膨大なデータの活用が圧倒的なパワーになり得ることを表出したメッセージと言える。

医療とビッグデータ

クイズに勝利した後、IBMは対象を医療分野に焦点を当て事業を展開している[4]。多様な知見や治療指針が学術メディアに発表され、その数は今や膨大であり、多忙を極める医師が丁寧に全てを読むことは極めて困難であろう。爆発的に増える臨床論文、カルテ、治験等の情報をワトソンに注入し、医療現場で利用するこ

とを狙っている。もちろん治療の最終的な決定は医師によるものであり、ワトソンはあくまでも医師の判断に資する情報の提示がその役割となる。このように書くとソフトウエアが電子ドキュメントを自ら読み、咀嚼し、知識をどんどん獲得していくように聞こえるが、残念ながらそこまで技術は発達していない。現時点では、整理して医学知識を用意するところは人手が大きく介在する。人手による大きな手間がかかっても、それでも、このソリューションが世界で数多く利用されるほど強力になるのであれば、十分にその手間は報われることは明らかであろう。

よく、日本はグーグルに勝てるのかという質問を受ける。グーグルが持つ情報は情報空間全体から見ると必ずしも大きくはない。筆者は例えば医療分野における勝算は少なくないと感ずる。我が国の皆保険制度は世界的にみて極めて特徴的であり、データ形式を整えることにより膨大な時系列データは宝の山になりうる。ワトソンに注入されるよりも遙かに大量のデータを我が国は潜在的に保有する。グーグルが持てないビッグデータの活用にこそ注目すべきであろう。

ビッグデータではロングテイル（非常に多種類の希少事例）も重要である。日本医療研究開発機構（AMED）は希少疾患や未診断疾患に関する情報を世界的に集約することにより、早期診断に貢献するとともに治療法確立の加速に取り組んでいる。日本だけでは少ない症例も、世界が協調して情報を提供することにより症例数を増やし、疾患解明に大きく資することが期待出来る。もちろんコモンディジーズに対してもその多様性の明確化に活用出来よう。

医療は一般市民にとっても身近な課題であり、国家財政の観点からも重要であることから、ビッグデータへの期待が大きい。医療費削減の究極的課題として、そもそもどのような経緯で病気になるのかという問いにも迫れるかもしれない。多種多様な生活様態を補足する比較的安価なセンサーが利用可能になり、病気に至る以前の段階での日々の生活の情報を採取することにより、その解明にも一手が打てるかもしれないという夢も、それほど非現実的ではない状況になりつつある。筆者は情報学を専門としており医学は門外漢であるため、正確性を欠いているかとは思うが、遺伝子のみならず多様なビッグデータ解析の医療分野への展開は世界的に注目されている。

ビッグデータを創る

ビッグデータは身近な課題の解決にも有効である。筆者らは内閣府最先端研究開発支援プログラム（FI

ＲＳＴ）において、医療現場における看護師の方々を如何に支援可能かという課題に挑戦した。そもそも看護師の方々にとって最も手間のかかる作業は何かを理解すべく、センサーを付けて頂き、膨大な活動データを取得した。機械学習技術により、二十五の行動種を九〇％程度の精度で識別することが可能となった。その結果、ＩＴ研究者としては大変皮肉な結果として、看護師の方々はＰＣへのデータ入力に最も時間を費やしていることが判明した[5]。機器からＰＣに直接データを送るＭ２Ｍ技術を利用することにより大幅に効率化することが出来ると共に、人手入力によるエラーも回避出来る。職場の改善に大きく寄与出来る手応えを得ることが出来た。当たり前かもしれないが、ビッグデータの研究において重要なポイントの一つは、やりたいことに必要なデータが必ずしも揃っているわけではないことである。データが無ければ、データを創る必要がある。どんなビッグデータを創るかというデザインが重要と言える。

車のセンサーが生み出すビッグデータも価値を生む。例えば、地図上は一本のまっすぐな道にも拘わらず、多くのドライバーがブレーキを強く踏む場所があることがセンサーデータの解析から明らかになった。何らかの原因があるはずと思い更に調べてみると、ドライバーの視点からは見えづらい出入口等が突然視界に入る状況であった[6]。ハインリッヒの法則によれば、一つの大きな事故の陰には三百の火種があると言う。車が生み出す膨大なセンサーデータから潜在的に事故の危険性のある場所を同定し、それらをカーナビ表示に導入すれば事故件数を減らすことが出来るかもしれない。ドライバーが協力し合い、なるべく多くの車からの情報を集めることにより道路のカバレッジは上がり、時間依存性など精度の向上が期待される。筆者は社会便益を向上するためのビッグデータ構築に、我が国が世界をリードして積極的に取り組むべきと考えている。

ビッグデータ時代の課題

一方で課題も山積している。プライバシーは最も重要な視点の一つである。教育におけるビッグデータに関する巨大プロジェクトが該観点から頓挫した事例もある。個人のデータの取り扱いは極めて重要であり、多様な匿名化手法、秘匿計算が研究されている。データが大きくなればなるほど、全てのデータの質を担保することは困難になる。そもそも、データが誰のものなのか、データの所有権がはっきりとしない場合も多い。執筆時点で、ディープラーニングを駆使したＡＩが碁のチャンピオンに勝利したニュースが話題となっている。大量の棋譜データからの学習に加え、膨大な回数の自己対局結果を用いた学習を組み合わせ、従来に比べ圧倒的な強さを実現した。この延長線上としてＡＩを利用してかなり高度なコンテンツを容易に作れるようになる可能性が高い。その際の著作権のあり方は難解であり、内閣府知的財産戦略本部でも議論されている。

多くの研究者にとってデータは極めて貴重であるが故に、これまでデータを共有する動機付けが無かったのも事実である。現在、オープンサイエンスの潮流の中で、リサーチデータの共有がＧ７で大きく注目されている。論文と同様に、利用したデータを引用することによってデータ提供者へ謝意を表す習慣を醸成することが考えられている。論文よりも多くの引用数がデータに対して生まれるかもしれない。データ共有への仕組みづくりが大切である。データジャーナルも発刊されつつある。

ＩＴの歴史の中で、データの役割の重要性が強く認識される時代に突入したことは間違いない。一過性のブームではなく、今後ビッグデータ技術は更なる発展が期待される。

(注)
1) https://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf
2) 喜連川優　情報爆発のこれまでとこれから　電子情報通信学会会誌 Vol.94(8) pp.662-666　２０１１年
3) http://www.ibm.com/smarterplanet/jp/ja/ibmwatson/quiz/
4) http://www.ibm.com/smarterplanet/us/en/ibmwatson/watson-oncology.html
5) Sozo Inoue, Naonori Ueda, Yasunobu Nohara, and Naoki Nakashima: Mobile activity recognition for a whole day: recognizing real nursing activities with big dataset. UBICOMP'15, proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp1269-1280
6) Masahiko Itoh, Daisaku Yokoyama, Masashi Toyoda, and Masaru Kitsuregawa: Visual interface for exploring caution spots from vehicle recorder big data. IEEE Big Data 2015, pp776-784

（東京大学生産技術研究所教授・国立情報学研究所所長・東大・工博・工・昭53）