国立国会図書館 調査及び立法考査局

Research and Legislative Reference Bureau
National Diet Library

論題 Title	第 1 部　データ活用技術・データの扱い方（統計学・情報学等）の動向
他言語論題 Title in other language	Part1 Trends in data-utilization and handling data (statistics, informatics, etc.)
著者／所属 Author(s)	樋口　知之（HIGUCHI Tomoyuki）／情報・システム研究機構統計数理研究所　所長　ほか
書名 Title of Book	データ活用社会を支えるインフラ：科学技術に関する調査プロジェクト報告書 （Infrastructure for Data-Driven Society）
シリーズ Series	調査資料 2017-6　（Research Materials 2017-6）
編集 Editor	国立国会図書館 調査及び立法考査局
発行 Publisher	国立国会図書館
刊行日 Issue Date	2018-03-30
ページ Pages	1-38
ISBN	978-4-87582-815-0
本文の言語 Language	日本語（Japanese）
キーワード keywords	—
摘要 Abstract	ビッグデータがもたらす科学技術全般への影響や、その解析技術を解説する。また、人文科学への適用などビッグデータ活用による新しい可能性を示唆する。

第1部　データ活用技術・データの扱い方（統計学・情報学等）の動向

統計数理研究所　所長　樋口　知之

ビッグデータの登場は、科学技術ばかりか私たちの日常生活にまで広範囲にそして奥深く変革を巻き起こしている。第1部では、そのビッグデータの利活用を支える解析技術に焦点を当てる。まず、ビッグデータがもたらした科学技術全般への影響を方法論とその接点の視点で概説する。キーワードはデータ中心科学[1]とオープンサイエンスである。次に、ビッグデータ解析技術の基盤となっている統計的機械学習を理解するため、その要素技術を俯瞰（ふかん）した後に、ベイズ統計及びその周辺技術を解説する。あわせて、ビッグデータ活用に欠かせない匿名化技術を取り上げる。最後に、ビッグデータの利活用で大きなチャンスが期待できる人文科学における動向に触れることで、今後のビッグデータ活用による新しい可能性を示唆する。

Ⅰ　科学研究のデータサイエンス化

情報・システム研究機構　機構長補佐　戦略企画本部副本部長　丹羽　邦彦

1　科学研究の歴史

科学研究は長い歴史を持つが、その方法論に着目してみると4つの段階を踏んで進化してきたことが認められる[2]。科学の黎明期は専ら自然現象を観測して経験を積み重ねる「経験科学」であった。17世紀辺りからは、数学を駆使して自然現象を理論的なモデルで表現する「理論科学」の手法が発展した。これにはヨハネス・ケプラー（Johannes Kepler, 1571-1630）、アイザック・ニュートン（Sir Isaac Newton, 1643-1727）、ジェームズ・クラーク・マクスウェル（James Clerk Maxwell, 1831-1879）などが大きな足跡を残した。20世紀後半になると、計算機が発達したことにより、コンピュータシミュレーションを用いて複雑な現象を解析する「計算科学」が生まれ、理論だけでは複雑すぎて対処できない現象の理解が急速に進んだ。さらに現在は情報通信技術（ICT）が飛躍的に発展し、それによって可能となった大規模かつ多様な「ビッグデータ」が利用できる時代が到来しており、それを駆使して研究を推進する「データサイエンス」が重要になってきている。これを「第4のパラダイム」と呼ぶこともある。

2　何がデータサイエンスを可能にしたか

データサイエンスを可能にしたものとして、技術的には、測定機器、通信（インターネット）、情報処理の飛躍的発展と低廉化が挙げられる。例えば、ゲノム解析に用いられるシーケンサーは遺伝子の塩基配列を読み出す装置であるが、2000年に米国で「次世代シーケンサー」が登場したことで多数の塩基配列を高速に読み出せるようになり、ゲノム解析の速度が5年間で約

＊　本稿におけるインターネット情報の最終アクセス日は、平成30（2018）年1月17日である。
(1)　データ駆動科学、データ集約科学などの用語も使用されている。
(2)　Tony Hey et al., eds., *The Fourth Paradigm: Data-Intensive Scientific Discovery,* Redmond: Microsoft Research, 2009, p.xviii. Microsoft Website〈https://www.microsoft.com/en-us/research/publication/fourth-paradigm-data-intensive-scientific-discovery/〉

『データ活用社会を支えるインフラ―科学技術に関する調査プロジェクト2017報告書―』
（調査資料2017-6）国立国会図書館調査及び立法考査局, 2018.

1万倍に増加した。これは、半導体の技術革新の速さを表現するためによく使われる「ムーアの法則」（5年で約10倍）をはるかに超える速度である。このような技術進歩は塩基配列の読み取りに必要な期間と費用を劇的に縮小することを可能とした。例えば、2003年に終了した国際プロジェクト「ヒトゲノム計画」では、ヒトゲノム解読に13年、30億ドルを費やしたのに対して、現在では数日間、1,000ドルほどの費用で解読可能となっており(3)、遺伝学の研究やがん治療の研究に大きな進展をもたらしている。

一方、政策的には2012年3月に米国のオバマ政権が「ビッグデータ研究開発イニシアティブ」（Big Data Research and Development Initiative）(4)を発表し、ビッグデータの利活用を目的とした研究開発に2億ドルを投じることを公表した(5)ことがきっかけとなり、各国でビッグデータ及びデータサイエンスに関する政策的な動きが加速した。米国では、国防総省、国土安全保障省、エネルギー省、国立衛生研究所（National Institutes of Health: NIH）、国立科学財団（National Science Foundation: NSF）など多岐にわたる連邦政府機関において、ビッグデータ関連の研究開発が推進されている。

欧州連合（European Union: EU）では、中長期の科学技術・イノベーション戦略「ホライゾン2020」（Horizon 2020）(6)の中で、データ活用に関する研究開発が実施されている。我が国においては、平成24（2012）年10月、総務省・文部科学省・経済産業省の三省合同で提案した「ビッグデータによる新産業・イノベーションの創出に向けた基盤整備」が、内閣府総合科学技術・イノベーション会議の審議を経て重点施策パッケージとして選定され(7)、各省で施策が展開されてきた。平成28（2016）年1月に閣議決定された「第5期科学技術基本計画」(8)においても、具体的な強化策が提言されている(9)。

3　データサイエンスの事例

あらゆる学問分野においてデータの重要性は増大している。例えば、自然科学では生命科学、物理学、化学、天文学、地球環境学など、社会科学では金融工学、経済学など、人文科学では言語学、考古学などが典型例であるが、これら以外にもほとんどの分野で研究のデータサイエンス化が進展している。以下に代表的な例として、地球観測、生命科学、人文科学におけるデータサイエンスの事例を紹介する。

（1）地球観測

地球規模で発生している温暖化、環境劣化、異常気象、生態系の破壊、資源の枯渇などの諸

---

(3)　詳しくは、大下淳一「次世代シーケンサーとは」2016.4.5. 日経デジタルヘルスウェブサイト〈http://techon.nikkeibp.co.jp/atcl/word/15/327920/040400013/?ST=health〉; 日本学術会議情報学委員会 E-サイエンス・データ中心科学分科会「ビッグデータ時代に対応する人材の育成」2014.9.11, p.4.〈http://www.scj.go.jp/ja/info/kohyo/pdf/kohyo-22-t198-2.pdf〉を参照。

(4)　Office of Science and Technology Policy Executive Office of the President, “Obama Administration Unveils “Big Data” Initiative: Announces $200 Million in New R&D Investments,” March 29, 2012.〈https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf〉

(5)　詳しくは、日本学術会議情報学委員会 E-サイエンス・データ中心科学分科会　前掲注(3), p.9. を参照。

(6)　“Horizon 2020.” European Commission Website〈https://ec.europa.eu/programmes/horizon2020/〉　ホライゾン2020は欧州連合（European Union: EU）による欧州全体の研究力向上を目的とした研究・イノベーション枠組み計画で、2014年から2020年までの7年間に総額800億ユーロ（約11兆円）が投じられる。

(7)　科学技術政策担当大臣・総合科学技術会議有識者議員「平成25年度科学技術関係予算 重点施策パッケージの特定について」2012.10.25, p.10. 内閣府ウェブサイト〈http://www8.cao.go.jp/cstp/budget/h25package.pdf〉

(8)　「第5期科学技術基本計画」（平成28年1月22日閣議決定）内閣府ウェブサイト〈http://www8.cao.go.jp/cstp/kihonkeikaku/5honbun.pdf〉

(9)　同計画の「第2章（3）②基盤技術の戦略的強化」を参照（同上, pp.13-15）。

問題に対応するには、まず地球規模での自然観測を行い、その観測データを基に地球環境を理解、予測し、対策を立案するための情報提供が求められる。

このような要請に応えるため、我が国ではデータ統合・解析システム（Data Integration and Analysis System: DIAS）[10]が文部科学省のプロジェクトとして平成18（2006）年にスタートしている。平成23（2011）年度からの第2期における高度化・拡張を経て、平成28（2016）年度からは第3期としてシステム構築及び運用が開始されている。DIASは、地球規模及び各地域の観測で得られたデータを収集、蓄積、統合、解析するとともに、社会経済情報などと統合し、気候、水、農業など多くの分野での施策立案に資する情報提供を目指している。また、全球地球観測システム（Global Earth Observation System of Systems: GEOSS）[11]に参加する世界各国のデータセンターとの接続を通じて気候変動等の地球規模課題の解決に資する国際貢献も果たしている。

DIASは超大容量データのアーカイブと解析及びシミュレーションを行うため、2015年10月時点で、合計25ペタバイト（1ペタバイト＝百万ギガバイト）の超大容量ストレージと解析ツールを備える[12]。また、各地のデータセンターやスーパーコンピュータ保有機関と学術情報ネットワーク（SINET）[13]で結ばれ、高速データ転送が可能である。

（2）生命科学

近年の生命科学の目覚ましい発展の原動力となっているのは、塩基配列データから得られる知識である。塩基配列データは現在までの生物進化を直接示す記録であり、これを蓄積・処理・解析することによって生命情報学（バイオインフォマティクス）と呼ばれる新しい研究領域を切り開き発展させることが可能となっている。しかも国際塩基配列データベース（International Nucleotide Sequence Databases: INSD）と呼ばれる国際的な仕組みによって、世界中の研究者がこのデータベースに塩基配列データを登録し、目的や国籍によらず閲覧、利用することが可能になっている。

INSDに登録された塩基配列数は、平成18（2006）年9月に6100万件、平成23（2011）年9月に1億4200万件、平成28（2016）年9月には1億9600万件と急速に伸びており[14]、この分野の研究の進展ぶりを表している。

（3）国文学

大学共同利用機関法人人間文化研究機構国文学研究資料館は、日本の歴史的典籍30万点をデジタル化し、関連情報を付与（タグ付け）したデータを公開して、国内外の大学・研究機関と連携した国際共同研究を推進する大型プロジェクト「日本語の歴史的典籍の国際共同研究ネットワーク構築計画」を実施している[15]。本プロジェクトは多くの機関との連携を活用していることが大きな特徴である。例えば情報・システム研究機構データサイエンス共同利用基

⑽　DIASデータ・統合解析システムウェブサイト〈http://www.diasjp.net/〉
⑾　「我が国における地球観測の推進：GEOSSの概要」文部科学省ウェブサイト〈http://www.mext.go.jp/a_menu/kaihatu/kankyou/suishin/detail/1284744.htm〉
⑿　詳しくは、「DIASの機能」DIASデータ・統合解析システムウェブサイト〈http://www.diasjp.net/about/system/〉を参照。
⒀　学術情報ネットワークSINET5ウェブサイト〈https://www.sinet.ad.jp/〉
⒁　詳しくは、「The Number of Entries by Contributors to DDBJ Release」2016.12.21. DDBJウェブサイト〈http://www.ddbj.nig.ac.jp/breakdown_stats/prop_ent_old-e.html〉を参照。
⒂　「日本語の歴史的典籍の国際共同研究ネットワーク構築計画（略称：歴史的典籍NW事業）」国文学研究資料館ウェブサイト〈https://www.nijl.ac.jp/pages/cijproject/〉

盤施設人文学オープンデータ共同利用センターは、国文学研究資料館がデジタル化した古典籍を再利用しやすい形式に変換し、オープンデータとして一括配布する役割を担っている[16]。平成29（2017）年12月時点で「源氏物語」、「徒然草」など1,767点の古典籍データとして、画像データ（約33万コマ）及び書誌データ、作品紹介、翻刻テキストデータなどを公開している[17]。データをオープン化することにより、国内はもとより、海外研究者もデータを活用して研究することが可能となり、国際共同研究も進展することが期待できる。またいわゆる書誌情報にとどまらず画像データを取り込むことによって、データサイエンスの新たな研究領域（例えばくずし字の解読など）を開拓している点も大きな特徴であり、人文科学におけるデータサイエンスの一事例である。

4　データサイエンスの課題

このようにデータサイエンスの重要性が増す一方で、検討課題も指摘されている。総合科学技術・イノベーション会議議員及び外部有識者から構成される「国際的動向を踏まえたオープンサイエンスに関する検討会」は、平成27（2015）年3月に「我が国におけるオープンサイエンス推進のあり方について」[18]を取りまとめ、以下のような課題を指摘した上で、長期的視点から検討に取り組むべきと述べている。この報告書では、これらの課題はオープンサイエンスに関する課題として挙げられているが、データサイエンスの課題にもなり得るものである。

①論文、研究データの公開・共有化
②研究データの保存
③保存すべきデータ及び保存期間等
④研究データの技術的な品質の評価等
⑤研究者に対するインセンティブ等
⑥データ駆動型の研究をサポートするサービスを企画、開発、運用する人材の確保

Ⅱ　オープンサイエンス・オープンデータ

国立情報学研究所准教授　北本　朝展

1　はじめに

「オープン」が多義的な言葉であるのと同様、オープンサイエンスは多様な意味を持ち、多様な活動を内包する概念である。サイエンスをよりオープンにするという方向性を共有する多くの活動が同時に展開しているが、何をオープンにするかという点では個々の活動によって違いがある。また個々の活動の中でも、なぜオープン化するかという動機については、複数の動機が混在している場合もあろう。つまりオープンサイエンスとは、何をなぜオープン化するかという点では多様性がある種々の活動を、オープン化するという共通の方向性で束ねた「傘」

---

⑯　国文学研究資料館「のぞき見しません?・・・古典の世界!旧本古典籍データセット700点 大公開」2016.11.10.〈https://www.nijl.ac.jp/pages/cijproject/images/20161110_news.pdf〉
⑰　「日本古典籍データセット」人文学オープンデータ共同利用センターウェブサイト〈http://codh.rois.ac.jp/pmjt/〉
⑱　国際的動向を踏まえたオープンサイエンスに関する検討会「我が国におけるオープンサイエンス推進のあり方について―サイエンスの新たな飛躍の時代の幕開け―」2015.3.30, pp.21-22. 内閣府ウェブサイト〈http://www8.cao.go.jp/cstp/sonota/openscience/150330_openscience_2.pdf〉

のような概念である。ゆえにオープンサイエンスに関するコミュニケーションでは、どの意味でのオープン化なのかを明確にしなければ、話がかみ合わないことになる。そこでオープンサイエンスについては、「同床異夢」の状況を踏まえた上で全体像を把握することが重要である。以下では、このようなオープンサイエンスの全体像を描き、その活動の源流をたどることで、オープンサイエンスに関する議論を整理する。

2　オープンアクセス

オープンサイエンスに関して、最も歴史が長いオープン化活動が「オープンアクセス」であり、学術出版物に対するアクセスのオープン化を目標に掲げている。

(1) オープンアクセスを推進する動機

オープンアクセスを推進する動機は、第一に出資者からの要求である。国民の税金で行われた研究の成果が納税者に還元されていないという観点からの議論に加えて、近年では欧米の一部の民間財団においても人類への貢献に寄与するオープンアクセスという観点からの意識が高まっており、米国のビル＆メリンダ・ゲイツ財団のように、同財団が助成する研究成果の出版にオープンアクセスを必須要件とする（オープンアクセスでない出版物への投稿を研究成果と認めない）など、オープンアクセスの推進により積極的に取り組む財団も出てきている[19]。第二に学術出版物の中でも特に学術雑誌の購読料が高騰しているため、大学図書館等の学術雑誌購読者が契約を継続するのが困難になる（シリアルズクライシス）という費用の観点からの議論である。

これら2点の問題は、学術雑誌の電子化が進み、出版社によるアクセス制限が強まるにつれて深刻化しており、ドイツなどでは一国の学術界がまとまって出版社と交渉するなど、大きな転換期を迎えている[20]。

(2) オープンアクセスを実現するモデル

こうしたオープンアクセスを実現するために、主に3つのモデルが考案されている。

第一が「グリーンオープンアクセス」である。これは出版者の認める条件に従って著者が自身の出版物を公開する場所を自ら選択するモデルであり、公開場所としては著者が所属する研究機関が運営する機関リポジトリが用いられることが多い。著者にとっては低価格で公開できる利点があるものの、出版者による公開からある程度の期間を置かなければならない場合が多いため、即時性や網羅性に問題がある。

第二が「ゴールドオープンアクセス」である。これは出版物の著者が、出版に関わる諸費用を出版者に支払うことでアクセスをオープン化するモデルであり、出版と同時にオープン化することも可能になるというメリットがある。しかし、従来は読者から徴収していた費用を著者

---

⑲　“Gates Foundation research can’t be published in top journals,” 2017.1.13. Nature Website〈https://www.nature.com/news/gates-foundation-research-can-t-be-published-in-top-journals-1.21299〉では、ビル＆メリンダ・ゲイツ財団の助成を受けた研究成果が、オープンアクセスでないジャーナルでは出版できなくなった状況を紹介している。日本語の概要は、「ビル＆メリンダ・ゲイツ財団のオープンアクセス方針が正式発効　助成を受けた論文はNature、Science等のトップジャーナルに掲載できない状態に」2017.1.17. 国立国会図書館カレントアウェアネス・ポータルウェブサイト〈http://current.ndl.go.jp/node/33272〉を参照。

⑳　「ドイツ・DEALプロジェクトとElsevier社による全国規模でのライセンス契約交渉が決裂」2016.12.16. 国立国会図書館カレントアウェアネス・ポータルウェブサイト〈http://current.ndl.go.jp/node/33123〉

から徴収するという転換が生じると、出版物の品質を高く保ってより多くの読者を獲得するよりも、出版物の品質を下げてより多くの著者を獲得するほうが出版者の利益が増える構造となるため、品質管理のインセンティブがゆがんでしまう場合がある。

第三に「プレプリントサーバ」である。従来の出版が査読や編集を価値の源泉としていたのに対し、査読や編集の大部分を省略し、その代わりに即時性と低コストを実現するモデルである。人工知能（AI）で用いられる機械学習などの先端研究分野では、迅速な研究成果共有へのニーズが高まっており、プレプリントサーバが学術出版の方法を根本的に変革しつつある。しかし、公開される出版物が玉石混交となり品質管理の問題が生じることから、プレプリントサーバで公開された後で査読を行う方式も生まれている。

3　オープンデータ

（1）オープンデータと透明性

オープンデータは、もともとガバメントデータ（行政機関が保有するデータ）のオープン化（オープンガバメント）に源流がある。情報公開に基づく政治や行政の透明性を現代のデータサイエンスによって検証するには、機械可読形式の生データのオープン化が必要であるとの考え方が広まった[21]。中でもWorld Wide Webを発明したティム・バーナーズ＝リーが2010年2月に行ったTEDの講演「The year open data went worldwide」[22]であり、そこで「raw data now!」として政府や研究者や機関に生データの公開を呼び掛けたことが、その後の動きを後押しするきっかけとなった。

一方、サイエンス分野においては、後述のように昔から研究データを共有する文化が存在した分野もあるが、それが特にオープンデータとして注目を高めた背景には、研究の再現性に対する疑義の声がある。論文に書いてある手法を使っても同じ結果が再現できないという研究の再現性の危機に対して、論文の結論だけではなく論文の主張の根拠（エビデンス）となるデータもオープン化して他者が検証可能とすることで、研究不正を防ぐべきではないかという議論が起こってきた[23]。我が国では平成26年8月26日に「研究活動における不正行為への対応等に関するガイドライン」が定められるなど、この問題は近年では学術界全体の課題として認識されている。

つまり、オープンガバメントにおいてもオープンサイエンスにおいても、何らかの政策や主張のエビデンスをオープン化することで、より透明性が高く再現可能な環境を実現しようという方向性は同じである。そして日本では、大きな研究不正が相次いだため、エビデンスの公開だけではなく研究の途中経過のデータなども長期保存し、不正を検証できるようにする方向で議論が進んでいる[24]。

⑵1　Open Data Handbook〈http://opendatahandbook.org/〉はオープンデータの基礎を解説するものであり、その中ではデータの公開方法についても言及がある。

⑵2　Tim Berners-Lee, “The year open data went worldwide,” February 2010. TED Website〈https://www.ted.com/talks/tim_berners_lee_the_year_open_data_went_worldwide〉

⑵3　この議論は世界的にも様々な機関で行われているが、例えば日本では、国際的動向を踏まえたオープンサイエンスに関する検討会　前掲注⑴8, p.12に、「論文、研究データの公開は、研究不正を回避する意味でも重要」との記述がある。

⑵4　同上, p.21に研究データの保存の仕組みが今後の検討課題として挙げられている。

（2）オープンデータと利活用

データをオープン化するもう1つの目的は、データを誰でも使いやすくしてデータの利活用を促進することにある。サイエンスの分野によっては、こうしたデータ共有が規範として根付いている。

例えば高エネルギー物理学や天文学のように、限られた大規模装置でしか最先端の観測ができない分野では、それを使って得られたデータを一定の期限内に必ず公開しなければならないルールが決められている場合がある[25]。また、地球科学のように地球全体の分析を目的とする分野では、地球全体のデータを共有して分析することが研究の前提条件となる[26]。さらに最近のAI分野の研究においては、データセットを共有して手法の比較を行うことが標準的な方法として確立されており、オープンデータとクローズドデータの使い分けが戦略的に行われている。

このように、データの公開や共有という面からオープンサイエンスを推進する動きは、様々な分野で始まっている。

４　組織のオープン化と参加の拡大

オープンサイエンスでは組織のオープン化、すなわちより多くの人を研究に参加させるという意味でのオープン化も重要である。

（1）市民科学

市民が科学に参加するという活動は市民科学（シチズンサイエンス）と呼ばれ、生態学や天文学などの分野では以前から行われてきた。それは、科学者だけでは地理的に広域のデータを網羅的に収集することができないため、市民の協力を得ながらデータを集め、それを科学者が分析することで研究成果を得るというものであったが、最近では分析そのものにも市民に参加してもらうことで、市民が科学的知識や体験を得ることに価値を置く活動も増えている。歴史学などでは各地域の郷土史を研究する「在野」の研究者がこれまでも多く存在した。しかしそれに加えてネット技術の発達により、大学や研究機関に属する研究者という「狭義のアカデミア」の外での研究活動を可視化、共有化することが容易となりつつあり、オープンサイエンスの裾野が拡大する機運が高まっている。

（2）オープンイノベーション

このようにアカデミアの内外の人々が協働しながら新しいアイデアや成果を生み出すことへのニーズが高まる中で、それを実現するための新しい方法論に対する関心も高まっている。例えば「ハッカソン」（プログラムやウェブサービスの開発）や「アイデアソン」（アイデア出しの議論）のように、不特定多数の人間を集めたオープンなイベントを通じてこれまでにない知恵やアイデアを得る方法や、コンペティション型研究[27]のようにある問題に対するソリューションを

⑵⑸　例えば国立天文台のすばる望遠鏡で観測したデータは、観測者が18か月間は独占的に利用できるが、それ以後はデータアーカイブでオープン化される。“Subaru Open Use Policy.” Subaru Telescope Website〈https://subarutelescope.org/Observing/Proposals/Submit/policy.html〉

⑵⑹　例えば気象学の分野では国際的なデータ共有が気象予測の前提となることから、世界気象機関（World Meteorological Organization: WMO）の全球通信システム（Global Telecommunication System: GTS）は1963年から始まっている。

⑵⑺　機械学習分野におけるImageNet Large Scale Visual Recognition Competition（ILSVRC）では、共通のデータセットを用いた物体認識技術のコンペティションを開催しており、そこでディープラーニング（深層学習）の圧倒的に高い性能が注目を集めてその後広く使われるようになったという事例がある。

不特定多数の参加者に競わせて比較する方法など、様々な新しい研究方法が考案されている。

こうした研究方法は、企業が関わる活動ではオープンイノベーションと呼ばれることが多い。企業内の人員だけで考えたアイデアよりも、外部の協力を得て考えたアイデアの方が、これまでにない斬新な製品を生み出せることを示す成功事例が幾つか報告されている[28]。

例えばAI分野では、ディープラーニング（深層学習）[29]の最先端ソフトウェアをオープンソースで公開するという動きが広まっており、オープンソースの世界で企業間の競争が起きている。多額のコストを投入して作成したソフトウェアであっても、それをオープン化することでユーザーが増加し、様々なニーズに合わせたツールが開発され、それがソフトウェア環境の利便性を向上させるという好循環のメリットが大きいためである。ただしそこには、ソフトウェアをオープン化したとしても、より価値が高いデータをクローズドにしておけば競争力を十分に保てるとの計算も働いている。つまり、何をオープンにして何をクローズドにするかという戦略が、企業の競争力と長期的な成長を左右するのである。

(3) 超学際（トランスディシプリナリ）研究

(1)や(2)のように様々な背景を持つ人々がサイエンスに参加する状況を超学際（トランスディシプリナリ）と呼ぶ。学際（マルチディシプリナリ）とは、既存の学問分野を越えて協働することを意味するが、超学際とは学問という壁そのものを乗り越え、アカデミアとその外の人々とが共通の目的を持って創造することを意味する。市民科学やオープンイノベーションもアカデミアの壁を乗り越えるという意味では超学際の実例と位置付けることができる。さらに行政との協働を通してデータ駆動型の政策決定を進めていくことへの期待も大きい。オープンサイエンスは、サイエンスの広報活動（アウトリーチ）とは異なり、アカデミア内外の人々が共に何かを作り上げていくという指向を持っている。

5　オープンサイエンスと研究データ基盤

(1) 研究データ基盤の持続可能性

オープンサイエンスを支える研究データ基盤に関する動きも加速しつつある。ネットワーク、ストレージ、コンピュータの3要素の強化が研究の基盤強化に必要であることは改めて論じるまでもないが、研究データ基盤の構築においては持続可能性という観点も重要となる。オープンデータをサイエンスで利用可能とすることは、そのデータを長期的に安定して公開する基盤の存在が前提になるからである。

例えば、論文の根拠となるデータが5年程度で消えてしまっては、次の世代の研究者はその正しさを検証することができないため、システムの更新などがあってもデータを持続できる長期的な基盤が求められる。そのためには研究データ基盤の運用体制や運用資金を含めた長期的な視野が必要になってくる。研究データにおける長期的な視野とは、理想的には半永久的な持続可能性を考えるべきであるが、現実には研究データの10年保存ルールなどを踏まえ、少なくとも10年程度は稼動する基盤について考えることになる。

---

[28] 例えば国立研究開発法人新エネルギー・産業技術総合開発機構（NEDO）は「オープンイノベーション白書　初版」2016.7.〈http://www.nedo.go.jp/library/open_innovation_hakusyo.html〉にて、オープンイノベーションの推進事例を多数紹介している。

[29] 第Ⅲ章「ビッグデータ活用に係る要素技術」で詳説する。

（2）長期的なアクセシビリティの確保（FAIR データ）

データを長期的に安定して公開する基盤を構築する上では、データへの長期的なアクセシビリティを確保することが重要な課題である。たとえデータへのアクセス方法を URL で明示しても、データの公開場所の移転によってその URL は無効になることから、この方法では長期的なアクセシビリティを確保することはできない。ゆえにデータ基盤で公開したデータについては、所在が不明になったり消滅したりすることを防ぐための基盤が必要となる。

この基盤としては「デジタルオブジェクト識別子」（Digital Object Identifier: DOI）という仕組みが広く使われている。これは、世界的に通用する永続的な識別子をデータに付与することで、短い文字列でその所在を特定可能とするものである。DOI が付与されたデータは、たとえアクセス先が移転したとしても、DOI の管理者である DOI 財団に新しいアクセス先を登録する仕組みがある。このように DOI はアクセス先の変更にも強いことから、研究の根拠となるデータを引用する際に DOI を明記すれば、データへの長期的なアクセスが実現できることになる。さらに、データの説明（メタデータ）をデータに付与することで、キーワード検索などによるアクセスも可能となる。

このように DOI とメタデータによって、データへの長期的なアクセシビリティを確保することは、研究データ基盤の基本的な機能と考えられている[30]。こうした研究データ基盤の要件を整理したものとして「FAIR データ」が共通認識となりつつある[31]。これは、Findable（探せる）、Accessible（アクセスできる）、Interoperable（相互運用性がある）、Reusable（再利用できる）という 4 つの要件を示したものである。

（3）データ論文とデータ引用

次に問題となるのはデータ公開のインセンティブである。そこでは研究成果と研究評価の関係が深く関わってくる。従来、研究者の業績（研究成果）は「査読論文」が中心となっているが、近年はデータの公開そのものを業績とするための「データ論文」という新しい論文フォーマットが誕生するなど、「論文」の概念が拡張しつつある。データ論文とは、データから得られる研究成果に関する論文ではなく、データの品質とその根拠となるプロセスなど、データに関する重要事項を説明する論文であるが、データから得られる成果を論文中に記述してはならず、その新規性も評価基準としないなど、内容と評価の両面において従来の論文とは異なる。

こうしたデータ論文なども活用して、データが永続的な識別子によって特定可能となれば、学術出版の基本である「引用」の仕組みを研究評価に活用できる。例えばデータの引用回数に基づき、そのデータを公開した研究者による研究のインパクト（研究評価）を計量することも可能になる。こうした方法でデータ公開のインセンティブを高めるには、研究者がデータを引用し、データを公開した研究者を評価する文化を拡大していくことが重要な課題となる。

⑶⓪　国際的動向を踏まえたオープンサイエンスに関する検討会　前掲注⑱, p.8 に、オープン研究データに関する原則として、持続的な識別子フレームワーク（DOI はその実例）、記述メタデータ標準の採用といった項目が挙げられている。

⑶①　例えば欧州委員会は、2016 年 7 月に発行した “H2020 Programme: Guidelines on FAIR Data Management in Horizon 2020, Version 3.0,” 26 July 2016. European Commission Website〈http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf〉にて、データ管理に FAIR データ原則を取り入れるガイドラインを示している。

（4）信頼できる研究データリポジトリ

データを長期的かつ安定的に公開するためには、しっかりしたデータ公開の受け皿が必要である。その受け皿となる研究データリポジトリは、研究データを受け入れ、保存し、公開するための機能を備えるべきである。しかし、必ずしも全ての研究データリポジトリが高い基準を満たしているわけではないため、信頼できる研究データリポジトリかどうかを、データを預ける研究者やその他の人々が判断するための仕組みが必要となる。そのための仕組みが「リポジトリ認証」と呼ばれる。例えば「CoreTrustSeal」[32]は、研究データリポジトリの信頼性を申請書類に基づき専門家が査読し、基準を満たしていると認められるリポジトリに証明書を発行するサービスである。

研究データリポジトリがオープンデータの受け皿となるには認証などの仕組みを活用し、「信頼できるリポジトリ」として世界に広く認識される必要がある[33]。しかし、日本国内に信頼できるリポジトリが確立していないと、データの保存先として海外の研究データリポジトリが選ばれるという「海外流出」を招きかねない。これは、日本国内における研究基盤の整備を遅らせるだけでなく、安定的なデータ提供を海外の研究データリポジトリの方針に依存させることにもなる。我が国の研究資産でもある研究データを自らの方針によって管理、保存、公開するための信頼できる仕組み作りが課題である。

（5）永続的識別子の拡大

論文やデータへのDOI付与が進展すると、研究に関わるその他の要素にもDOIのような永続的な識別子を付与して特定可能とし、研究に関する様々な情報を可視化して定量的に評価することが可能になる。

例えば、研究者を識別するIDである「ORCID」（Open Research and Contributor Identifier）を活用することで、誰がどのような研究成果を生み出したかを集計できるようになる。同様に、研究資金のIDを付与すればどの研究資金によって、また組織のIDを付与すればどの組織によって、どのような研究成果が生み出されたかなどを集計できる。

このように研究に関わるあらゆる要素をIDで接続することにより、研究業績の数量化が促進され、研究評価の透明性向上が期待される。ただし、数量化がどのような方針の下で行われているかという点を抜きにして数値の大小を比較することは危険であり、不適切な数量化によって研究評価がゆがめられる危険性にも配慮することが必要である。例えば、研究分野ごとに平均的な論文数や引用数が異なることはよく知られた事実であり、そうした補正を抜きに数値を単純に比較しても、意味のある評価を導くことは困難である。

6　欧米及び国際的動向

最後にオープンサイエンスに関連する欧米の動向と国際的な動向について触れておく。

32　CoreTrustSealとは、ICSU（国際科学会議）のWorld Data System（WDS）及びオランダDANS（Data Archiving and Networked Services）のData Seal of Approval（DSA）というリポジトリ認証を統合することで誕生した、国際的なリポジトリ認証の仕組みである（CoreTrustSeal Website〈https://www.coretrustseal.org/〉）。CoreTrustSealのほかに、ISO 16363等リポジトリ認証の国際標準なども存在するが、これらは審査基準がより厳しいため認証を取得したリポジトリの数は少なく、現在のところCoreTrustSealがリポジトリ認証の仕組みとしては最も普及している。

33　信頼できるリポジトリのリストについては、先述のCoreTrustSealに加えて、簡単な審査だけでメタデータを登録できるre3data〈https://www.re3data.org/〉などがあり、研究データリポジトリの事例を探すのに便利である。

（1）欧州

欧州では、「欧州オープンサイエンスクラウド計画」（European Open Science Cloud: EOSC）[34]というプロジェクトが立ち上がり、オープンサイエンスを旗印に研究基盤の構築にまい進している。これはEUの科学技術・イノベーション戦略である「Horizon 2020」[35]から多額の支援を受けて2020年までに実現を目指す研究基盤であり、「データ文化とFAIRデータ」、「研究データサービスとアーキテクチャ」、「ガバナンスと資金」という3つの項目に対してEOSCの目標を宣言した文書が2017年10月に公開された[36]。その最初の項目は「データ文化」であり、研究データが研究成果として重要であることを認め、研究の過程において適切に組織化（キュレーション）され、データの利用が長期的に持続するというデータ文化を実現するには、研究文化の変化も不可欠であるとしている。そのほか、オープンアクセスやインセンティブなどにも配慮することで、サイエンスの方法そのものの変革に向けた包括的なアプローチを掲げている。

（2）米国

米国では、非営利組織「Center for Open Science」（COS）[37]の活動が目を引く。COSでは、「再現可能な研究」を目的として、「Open Science Framework」（OSF）という研究基盤により、研究プロジェクトの管理、アクセス制御、作業工程の強化、成果発信の拡大、信頼できるリポジトリの構築などのサービスを研究者や研究機関向けに提供している。

2013年のCOS設立時には数名であった従業員が、約4年間で50名に増加しており、オープンサイエンスに関する活動の拡大傾向を示している。米国ではこのように企業的な性格を持つ非営利組織によりオープンサイエンスが進められている。この点は欧州とは異なる1つの特徴となっている。

（3）国際組織

最後に国際的な活動として、経済協力開発機構（OECD）と研究データ同盟（Research Data Alliance: RDA）について触れておく。OECDの科学技術政策委員会は以前からオープンデータに熱心であったが、2015年10月に「Making Open Science a Reality」と題するレポート[38]を発行するなど、オープンサイエンスの普及を後押ししている。

一方、RDAは研究データに関する諸活動が集結したコミュニティであり、欧州委員会や全米科学財団、オーストラリア政府などから一部の資金を受けているものの、特定の国や組織には属さない国際的な運営を行っている。オープンサイエンスに関わる成功事例の共有や標準化を進めており、研究者だけではなくライブラリアンやエンジニア、そして政策決定者や資金提供者など多様な人々が議論を続けながら、研究データに関する諸活動を進めている点に特徴がある。平成28（2016）年3月には東京で第7回RDA総会が開催され、この分野に対する国内の関心を高めることにつながった。その結果、同年6月にオープンサイエンスや研究データに

⑶④　“European Open Science Cloud.” European Commission Website〈https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud〉

⑶⑤　前掲注⑹を参照。

⑶⑥　European Open Science Cloud, “EOSC Declaration,” 26 October 2017. European Commission Website〈https://ec.europa.eu/research/openscience/pdf/eosc_declaration.pdf〉

⑶⑦　Center for Open Science Website〈https://cos.io/〉

⑶⑧　OECD, “Making Open Science a Reality,” OECD Science, Technology and Industry Policy Papers, No.25, Paris: OECD Publishing, 2015.〈http://dx.doi.org/10.1787/5jrs2f963zs1-en〉

関する議論を進める国内のコミュニティとして「研究データ利活用協議会」[39]が設立された。ここには、データマネージメントに関わる研究者、データサイエンスの研究者、ライブラリやミュージアムで実務に携わる専門家、研究資金提供機関の関係者など多様な人々が集まり、RDAと同様にデータを取り巻く多様な人々が集まり議論する場となっている。

Ⅲ　ビッグデータ活用に係る要素技術

1　データ解析と機械学習

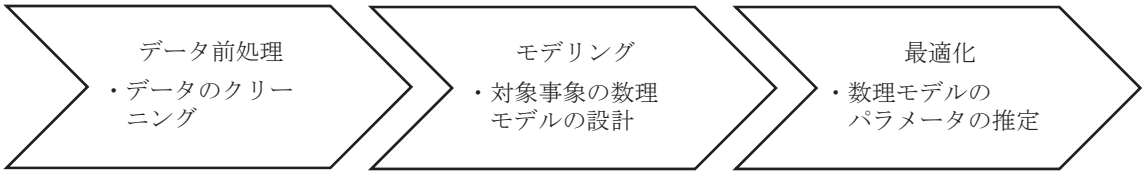
統計数理研究所教授　松井　知子

(1) はじめに

一般に、ビッグデータは、データ量（Volume）、データの種類（Variety）、データの生成・処理速度（Velocity）、データの変動性（Variability）、データの正確性（Veracity）の5つのVで特徴付けられる[40]。例えば「データの量」がどのくらいあれば「ビッグデータ」なのか、明確な答えはない。しかし、数十や数百ではなく、数十万、数百万以上になるであろう。ビッグデータが利用できるようになると、データ解析の方法も変化し、機械学習と呼ばれる手法が中心的に用いられるようになりつつある。本節ではこの機械学習の手続とその例について紹介する。

そもそもデータ解析は、解析の対象とする事象とそれに関連する観測データに基づき、「その事象を分類する」、「その事象の近未来を予測する」などの目的を達成するために実施される。機械学習は、これらを達成するための仕組み（モデル）をデータから自動的に獲得する技術である。一般に、機械学習の処理はデータ前処理、モデリング、最適化の3つに分けられる（図1）。データ前処理では、観測データに含まれるノイズや外れ値の除去や、欠損値の補完など（データのクリーニング）を行う。モデリングでは、データ解析の目的に合わせて、対象事象を表現するための計算式（数理モデル）を設計する。最適化では、数理モデルのパラメータ（モデルを記述するための変数）を前処理済みのデータを用いて推定する。

図1　機械学習の流れ



（出典）筆者作成。

(2) ビッグデータによるモデリングの変革

従来のデータ解析について、「ある工場で生産されるボルト」を例として説明する。データ解析の目的を「新たに生産されるボルトの長さを予測する」とする。観測データはボルトの長

---

[39] 研究データ利活用協議会ウェブサイト〈https://japanlinkcenter.org/rduf/〉

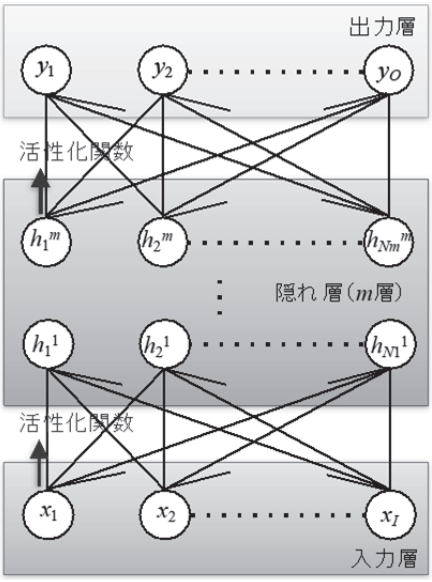
[40] 5つのVについての詳細はMartin Hilbert, “Big Data for Development: A Review of Promises and Challenges,” *Development Policy Review,* vol.34, January 2016, pp.135-174; “What is Big Data?” Villanova University Website〈https://www.villanovau.com/resources/bi/what-is-big-data/〉

さの測定値である。ボルトの長さの分布を表現するための数理モデルとしては、正規分布モデル[41]などが考えられる。この数理モデルの設計に必要なパラメータ（正規分布モデルの場合は平均と分散）は、ボルトの長さの測定値データから解析的に求めることができる。ところで、どのくらいのデータ量があれば平均と分散を精度良く計算できるであろうか。経験則として（データが独立に同一の分布に従う場合）、正規分布モデルのパラメータ（平均と分散）の計算には30個以上のデータがあればよいとされている[42]。

ところがビッグデータを利用できるようになり、特にモデリングの考え方に変化が現れている。データが数十や数百しかない時には、上記の正規分布モデルのように、パラメータ計算（推定）の精度を考慮して、パラメータ数が少ない（正規分布モデルでは2つ）数理モデルを用いざるを得なかった。しかし近年、数十万、数百万以上の大量のビッグデータを利用できるようになり、数理モデルのパラメータ数を少なくする工夫は余り必要なくなっている。

そこで機械学習の一手法であるディープラーニング（深層学習）[43]という方法が注目されている。ディープラーニングでは、図2に示すように入力層と出力層、それらに挟まれた複数の隠れ層により構成されるディープニューラルネットワークと呼ばれる数理モデルが用いられる。入力層、隠れ層は数百、数千のノード（図2の丸形）を持ち、各ノードは1つ上の層の多くのノードにリンクされる。この数理モデルのパラメータは各リンクを通して伝達される各ノードの出力値を決める係数（事前に設定する関数を用いて計算される。）に該当するため、膨大となる[44]。ビッグデータは、この膨大な数のパラメータの推定を可能にする。

図2　ディープニューラルネットワーク



（出典）筆者作成。

⑷1　平均値の付近に集積するようなデータの分布で表したモデル。平均と分散の2つのパラメータで表すことができる。

⑷2　詳しくは、Sheldon M. Ross, *Introduction to Probability and Statistics for Engineers and Scientists, 5th Edition,* Academic Press, 2014 の “Chapter 6. Distributions of Sampling Statistics” を参照。

⑷3　詳しくは、人工知能学会監修 , 神嶌敏弘編『深層学習』近代科学社, 2015 を参照。簡単には、4層以上の多層のニューラルネットワーク（脳機能に見られるいくつかの特性を計算機上のシミュレーションによって表現することを目指した数学モデル）による機械学習の一手法。

⑷4　パラメータの数は、入力層から隠れ層、出力層までの各層について、「その各層のノード数」×「1つ上の層のノード数」を計算し、これらを合計したものとなる。例えば、3つの隠れ層を持つディープニューラルネットワークを想定し、各層のノード数が千で、各ノードが1つ上の層の全てのノードにリンクされる場合、パラメータ数は400万個（=千ノード×千ノード×4層間）に達する。

ディープラーニングでは、初めに対象とする事象を、膨大なパラメータを持つディープニューラルネットワークで表しておき（モデリング）、最適化の過程でデータを用いて、その事象を表すのに不必要なパラメータがあればその値をゼロとし、実質的に計算しなくて済むように推定する。言わば、パラメータはデータにより取捨選択される。このように、ディープラーニングでは「データに基づいてモデル（のパラメータ）を自動選択する」という考え方が採られる。

このように、従来はデータが少ないため、最適化の精度との兼合いから「パラメータをできるだけ節約したモデルを工夫して設計する」という考え方であったが、ビッグデータの登場によりモデリングの考え方に変革がもたらされたのである。

以下では、機械学習におけるデータ前処理、モデリング、最適化について、それぞれ概説する。

(3) データ前処理

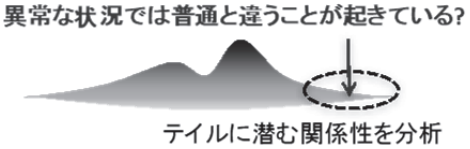
(i) データクリーニング

ビッグデータは多様なデータで構成され、ノイズ、外れ値や欠損値を含む場合が多いため、前処理によりデータをクリーニングすることが重要となる。ノイズや外れ値を除去するための代表的な方法の1つに「フィルタリング」がある。この方法は音声･画像情報処理を始め、様々な分野で研究開発・実用化されている。ノイズや外れ値に関する事前の知識に基づいて、ノイズや外れ値と判断される部分を除去するフィルタを設計し、これを用いてデータをフィルタリングする。

欠損値の処理は、定期的に観測されるべきデータについてある時刻のデータがない場合や、複数項目を持つデータについてある項目のデータが欠けている場合などに必要となる。欠損値に関する事前の知識に基づいて、その値は前後の平均値やスプライン曲線[45]上の値などで補完されることが多い。このようにデータクリーニングでは対象事象に関する事前の知識が不可欠であり、ビッグデータを扱う分野での専門性が強く求められる。

ここでビッグデータ解析の優れた点の1つとして、異常事象の解析について触れる。例えば、気象、地震、金融取引などを対象事象とする場合、降水量や気温、震度、商品価格に関するデータ分布の平均的特性を推定することよりも、それらの異常（豪雨、熱波、大地震、価格暴落など）に相当する「データ分布の裾（テイル）」を検知し、何が起きているのかを分析することが重要となる（図3）。少量データでは、データ分布のテイルまで十分に表す数理モデルを用いることはできないが、ビッグデータがあれば、データ分布の長いテイルも数理モデルに取り込むことができる。なお、異常事象の解析を目的とする場合には、データクリーニングにおいてノイズや外れ値として誤って削除しないようにすることが重要である。

図3　データ分布のテイル



（出典）筆者作成。

---

(45)　複数の節点を結ぶ滑らかな曲線。

（ii）データ研磨

ビッグデータ解析において、データの細部ではなく、ある程度明らかな特徴を共有する「グループ」を抽出することで、その全体像を知ろうとする時に利用される手法である。ビッグデータにはノイズが含まれるため、そのまま解析するとグループ構造が見えにくい。このため、まずノイズの少ないデータがどのようなものであるかを定義し、これに基づいて元のノイズ混じりのデータをノイズのないデータへと変換することでデータの揺らぎを消す（データ研磨）[46]。このデータ研磨により、グループに属するデータと属さないデータの境目が明確化され、グループ構造が理解しやすくなり、解析精度の向上が見込まれる。データ研磨はビッグデータを理解し、主観的に価値を創造していく上で有効な手法であると考えられる。

（4）モデリング

対象事象の数理モデルは、分類や判別、回帰、予測などデータ解析の目的に合わせて設計される。ここではディープニューラルネットワーク、サポートベクターマシン、状態空間モデルを例として紹介する。

（i）ディープニューラルネットワーク

上述のようにディープニューラルネットワークは、大量のパラメータを内包する、表現力に優れた数理モデルである。特に音声・画像情報処理の分野では、ここ数年、ディープニューラルネットワークは非常に高い性能を示している。そのネットワーク構成には様々なバリエーションがある。

例えば、画像情報処理では「畳み込みニューラルネットワーク」[47]の採用により、一般物体認識の精度が改善した[48]。音声情報処理では「Long Short-Term Memory」（LSTM）[49]が利用されることが多い。現在、スマートフォン等のほとんどは、音声認識にディープニューラルネットワークを活用しており、多くの人が認識性能の飛躍的向上を実感している[50]。最近では、金融分野においてディープニューラルネットワークを利用して、トレーダーのノウハウをデータから自動的に学習させる試みが注目されている[51]。

現在、関連ソフトウェアは数多く公開されている[52]。その多くは、大量のパラメータ計算をこなすために、「Graphics Processing Unit」（GPU）[53]が不可欠となっている。なお、ディー

---

⑯　詳しくは、大河原克行「データの揺らぎを消す「データ研磨」で、ビッグデータをさらに有用に―NII―」2014.8.21. クラウド Watch ウェブサイト〈http://cloud.watch.impress.co.jp/docs/news/662928.html〉; 宇野毅明ほか「データ研磨によるクリーク列挙クラスタリング」『情報処理学会研究報告』Vol.2014-AL-146 No.2, 2014.1, pp.1-8 を参照。

⑰　ディープニューラルネットワークの一種。識別部位の位置にかかわらず画像認識できるように画像を分割して取り込んだデータを統合（畳み込み）して処理する方法。詳細は、「数学知識もいらないゼロからのニューラルネットワーク入門」2017.4.15. TechCrunch Japan Website〈http://jp.techcrunch.com/2017/04/15/20170413neural-networks-made-easy/〉（原文 : Ophir Ianz, “Neural networks made easy,” 2017.4.13.〈https://techcrunch.com/2017/04/13/neural-networks-made-easy/〉）; 人工知能学会監修, 神嶌編　前掲注⑬, pp.19-21; Yann LeCun et al., “Deep learning,” *Nature,* Vol.521, May 2015, pp.439-440 を参照。

⑱　Alex Krizhevsky et al., “ImageNet Classification with Deep Convolutional Neural Networks,” F.Pereira et al., eds., *Advances in Neural Information Processing Systems* 25 (NIPS2012), 2012.

⑲　時系列データを扱うことができるディープニューラルネットワーク。詳しくは、人工知能学会監修 , 神嶌編　前掲注⑬, pp.214-217.

⑳　詳しくは、篠田浩一「音声言語処理における深層学習―総説―」『日本音響学会誌』Vol. 73 No.1, 2017.1, pp.25-30.

㉑　例えば、Saijel Kishan「ヘッジファンド、究極のトレーダー脳に接近か―AI 技術の深層学習―」『Bloomberg』2017.3.28.〈https://www.bloomberg.co.jp/news/articles/2017-03-27/ONH9SB6JIJUO01〉

㉒　関連ソフトウェアは、Deep Learning Website〈http://deeplearning.net〉などで紹介されている。

㉓　3DCG（3D グラフィックス）を描画する際に必要な計算処理を受け持つ半導体チップ。

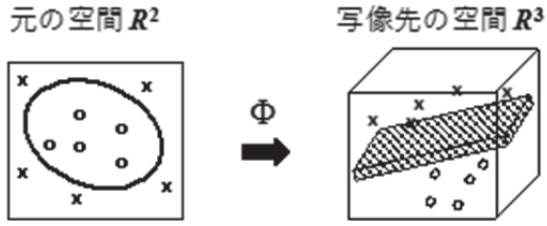
プニューラルネットワークの理論はまだ十分に成熟しておらず、推定されたパラメータの解釈やネットワークの理論的理解は難しい。ネットワークを構成する際は、対象事象や目的、データに応じて、経験に基づく調整が必要である。

（ii）サポートベクターマシン

サポートベクターマシンはディープニューラルネットワークに先んじて、1990 年代後半からパターン認識 (54) に幅広く適用されている数理モデルであり、カーネル法 (55) と呼ばれる手法の 1 つである。カーネル法では、次に説明するように、複雑なデータ（非線形かつ高次元のデータ）を容易に解析することができる。あるデータ（○と×）の特徴を認識するため何らかの方法で両者を区別する場合（図 4）、図 4 左のように平面上（2 次元空間）にあるデータ（○と×）は、線形分離（1 本の直線で分離）ができない。

そこで 2 次元空間の各データを適当な方法 (56) により 3 次元空間に変換（写像）すると、図 4 右中央の斜線で示した平面（線形面）で分離できるようになる。この境界となる平面の数式は容易に求めることができる。つまり、複雑なパターン認識の問題を、次元の高い空間へ写像することによってシンプルな形（線形）として解析できるようになるのである。なお、ディープニューラルネットワークにおいて、ノードの出力を計算する際にも、実はこれに相当する計算が行われており、複雑なパターン認識の問題を扱うことができる。

図４　データの３次元空間への写像



（出典）筆者作成。

サポートベクターマシンは、データのうち境界平面付近のサンプル（サポートベクター）だけを用いてデータを区分する境界平面を求める手法である。境界平面はサポートベクターの中心に位置するように設定する。サポートベクターだけを用いて計算するので、データを効率的に処理でき、計算コストを抑えることができる。なお、公開されている関連ソフトウェアの数は多い (57)。

（iii）状態空間モデル

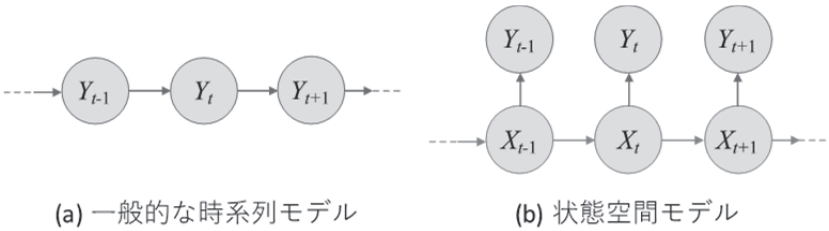
状態空間モデル (58) は、対象事象の動的な特徴を捉えるための数理モデルであり、これまで音声や映像などの時系列データ解析に広く利用されてきた。図 5 は一般的な時系列モデル (a) と状態空間モデル (b) の考え方を比較したものである。前者は観測データの動的な変化をその

---

(54)　文字や音声などの入力情報をパターンとして蓄え、新しい入力情報と照らし合わせて、どのような情報かを認識すること。
(55)　詳しくは、福水健次『カーネル法入門―正定値カーネルによるデータ解析―』朝倉書店, 2010 を参照。
(56)　例えば、$(x_1, x_2) \rightarrow (x_1^2, \sqrt{2}x_1x_2, x_2^2)$ などの変換が考えられる。
(57)　例えば、「LIBSVM」臺灣大學資訊網路與多媒體研究所ウェブサイト〈http://www.csie.ntu.edu.tw/~cjlin/libsvm/〉; "SVMlight." Cornell CIS Website〈http://www.cs.cornell.edu/people/tj/svm_light/index.html〉
(58)　状態空間モデルについて詳しくは、岩波データサイエンス刊行委員会編『岩波データサイエンス Vol. 6』岩波書店, 2017; 樋口知之『予測にいかす統計モデリングの基本―ベイズ統計入門から応用まで―』講談社, 2011 の 2 文献を参照。

まま捉えようとするが、後者では状態変数（X）という「潜在的な変数」を導入して、状態が動的に変化することにより観測データが変化していくと考える。

図５　一般的な時系列モデル vs 状態空間モデル
（Y：観測データ、X：状態変数）



（出典）筆者作成。

例えば、降水量予報において、降水量の観測データだけからその日々の変化をモデル化するのは難しい。気圧等の気象の状態が変化することにより雨が降ると考えた方が、物理的な現象によく合致しているし、より正確な降水量予報が実現できる。

状態変数は、最適化に用いられる「Expectation-Maximization（EM）アルゴリズム」などによりデータから推定できるため、直接観測できなくてもよい。例えば音声認識では、LSTM が用いられる以前は状態空間モデルが一般的に用いられていた。その場合、観測データは発声された音信号、状態変数はそのための発声器官の動きを表すと考えられる。発声器官の動きは直接観測できない潜在的な変数であるが、状態空間モデルを導入することにより、より高精度な認識を実現した。

状態空間モデルは、観測モデルと状態遷移モデル（システムモデルとも呼ばれる。）の２つのモデルにより構成される。観測モデルは各時刻とそのとき観測されるデータの関係（図５の X → Y）を表し、状態遷移モデルは状態遷移の関係（図５の Xt-1 → Xt → Xt+1）を表す。公開されている状態空間モデルの関連ソフトウェアとして、「Hidden Markov Model Toolkit」（HTK）[59] はよく知られている。

（5）最適化

最適化では、データから数理モデルのパラメータを推定する。本節では代表的なアルゴリズムである最尤（ゆう）推定と、近年ビッグデータ解析に関連して注目されているスパース推定について紹介する。

(i) 最尤推定

ビッグデータ解析では、「１　データ解析と機械学習」の冒頭で述べた「5 つの V」（データ量、種類、処理速度、変動性、正確性）が異なるデータを同時に解析する必要がある場合がある。例えば、津波や地震などの災害のデータ解析では被害状況、人流、物流などのデータ、映像のデータ解析では画像、音、テキストデータを、金融のデータ解析では取引状況に加え、社会状況などのデータも同時に解析する必要があるであろう。

5 つの V が異なるデータを同時に扱うための代表的な方法の 1 つとして、対象事象の発生機

---

[59] “What is HTK?” HTK3 Website〈http://htk.eng.cam.ac.uk〉

構を [0,1] の確率に基づく統計的な数理モデルで表す方法が挙げられる。統計的な数理モデルでは、そのモデルに対する観測データの尤度（ゆうど）（モデルと観測データが当てはまる度合い）を計算することができる。最尤推定のアルゴリズムでは、この尤度が最大となるようにモデルのパラメータを推定していく。

最尤推定の手法の1つに「EMアルゴリズム」があり、特に上述した状態空間モデルの状態変数のように、直接観測できない潜在的な変数を含む数理モデルに対して広く利用されている[60]。

（ii）スパース推定

ビッグデータ解析ではデータは大量にあるけれども意味のある情報は少ない場合がよく見られる。そのようなデータは「スパースなデータ」と呼ばれ、その解析に役立つのがスパース推定である。上述したように、近年、ディープラーニングの普及とともにデータ解析の考え方は「データに基づいてモデルを自動選択する」に変化している。スパースなデータに関しては、意味のある少ない情報をうまく捉えるようモデルを選択する必要がある。

スパース推定の代表的な手法の1つに「L1正則化」（線形回帰モデルに利用した場合にはLasso[61]と呼ばれる。）という手法がある[62]。上述のディープニューラルネットワークやサポートベクターマシンの最適化においても、このL1正則化がしばしば利用されている。

（6）まとめ

本節ではビッグデータ解析でよく用いられる機械学習のデータ前処理、モデリング、最適化の3つの処理について概説した。その中で、機械学習における重要な考え方として「データに基づいてモデルを自動選択する」、「データを写像して線形問題として解く」、「潜在的な変数を導入する」などを紹介した。各種の関連ソフトウェアが公開され、ビッグデータ解析は急速に進展しているが、我が国発のソフトウェアはほとんどないのが実状である。

２　ベイズ統計

統計数理研究所准教授　吉田　亮

（1）本節の概要

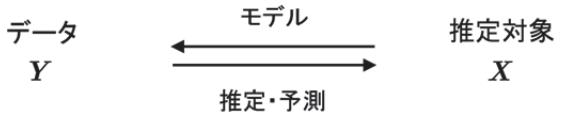
ベイズ統計は、トーマス・ベイズ（Thomas Bayes, 1702-1761）により示された確率の公式であるベイズの定理を基礎とする統計学の体系である。推定の対象であるXからデータYへの順方向の予測モデルを構築し、未知のXに関する事前情報を表す確率分布（事前分布）を組み合わせ、ベイズの定理と呼ばれる条件付確率の反転公式を用いて事後分布と呼ばれるYからXへの逆方向の予測モデルを導く。この事後分布に基づきXについて確率的な推論を展開する。これがベイズ統計の手続である（図6）。

60　詳しくは、渡辺美智子「EMアルゴリズム」北川源四郎・竹村彰通編『21世紀の統計科学〈Vol.III〉数理・計算の統計科学　増補HP版』東京大学出版会, 2012, pp.222-256. 統計科学のための電子図書システムウェブサイト〈http://park.itc.u-tokyo.ac.jp/atstat/jss75shunen/Vol3.pdf〉

61　Least absolute shrinkage and selection operator

62　L1正則化手法の詳細は、Trevor Hastie et al., *Statistical Learning with Sparsity: The Lasso and Generalizations*, Chapman and Hall/CRC, 2015を参照。

図６　ベイズの定理に基づく統計的推測



ベイズの定理に基づきXを推測

$$P(X \mid Y) \propto P(Y \mid X) \times P(X)$$

（事後分布 / 尤度 / 事前分布）

【尤度】モデル（$X \rightarrow Y$）を用いてデータを予測．データYに対するXの適合度

【事前分布】Xに対する先見知識を確率分布として表現

【事後分布】データYを観測したもとでXの確からしさ

※ $\propto$ は比例記号

（出典）筆者作成。

いわゆる頻度主義（後述）に基づく古典的な統計学と比較して、不確実性や確率に対する解釈、モデリングの柔軟性など、実用面の利便性が高いことから、ベイズ統計は現代のデータ科学のメインストリームに躍り出ることとなった。さらに、広範な普及に向けた最大の障壁であった事後分布の近似計算の問題に対し、1980 年代後半から 1990 年代を中心にマルコフ連鎖モンテカルロ法[63]や変分ベイズ法[64]などの優れた汎用手法が開発されたことで、ベイズ統計のコモディティ化が急速に進んだ。

ベイズ統計は、近年はディープラーニング[65]の隆盛と合流しながら、ビッグデータ分析や人工知能技術の基盤として大きな役割を果たしている。一方、ベイズ統計のモデリングの柔軟性は利点であると同時に、柔軟さゆえの汎用性の低さという問題があり、研究開発の高コスト化が課題として残されていた。しかしながら、近年、「WinBUGS」[66]のような汎用ソフトウェアが登場したことで、ベイズ統計のコモディティ化が更に進むこととなった。

（2）頻度主義とベイズ統計の対立の歴史

頻度主義に基づくいわゆる古典統計学は、20 世紀初頭から中期にロナルド・フィッシャー（Ronald A. Fisher, 1890-1962）、イェジ・ネイマン（Jerzy Neyman, 1894-1981）、カール・ピアソン（Karl Pearson, 1857-1936）らにより確立された。現在、大学等で学ぶ統計学（検定、推定など）は、古典統計学の体系が大半を占める。

古典統計学とベイズ統計学の本質的な違いの 1 つは、確率に対する考え方にある。古典統計学では、事象が起こる頻度の極限を確率とみなす。例えば、サイコロを振り、各々の目が出現する相対頻度を数え、試行を無限回繰り返したときの極限の値を真の確率とする。最尤推定という頻度主義に基づく推定では、有限回の試行から得られる各々の目の相対頻度が推定値となる。

⑹3　マルコフ連鎖モンテカルロ法は、確率分布から標本を生成する手法である。詳細は、伊庭幸人ほか『計算統計 2』（統計科学のフロンティア 12）岩波書店, 2005 を参照。

⑹4　変分ベイズ法は、確率分布の近似手法である。詳細は、中島伸一『変分ベイズ学習』（機械学習プロフェッショナルシリーズ）講談社, 2016 を参照。

⑹5　ディープラーニングについては、人工知能学会監修, 神嶌編　前掲注⑷3を参照。

⑹6　ケンブリッジ大学のウェブサイトでダウンロードできる。“WinBUGS.” MRC Biostatistics Unit Website〈https://www.mrc-bsu.cam.ac.uk/software/bugs/the-bugs-project-winbugs/〉

これに対してベイズ統計では、基本的に確率は人間の主観的な信念を数値化したもの（主観確率）と考える。六面サイコロの場合、多くの人は各目の出現確率は同様に確からしいと考える。このときの主観確率は1/6である。これを事前確率又は事前分布という。そして、サイコロを振り、データ（各目の出現頻度）を集積しながら尤度（データへの適合度）を計算し、試行ごとに自己の主観確率を補正していく。この補正された主観確率が事後分布に相当する。

頻度主義の推定では、データ数が少ない局面において相対頻度に大きな揺らぎが生じるため、推定値は1/6から大きく外れることもある。これに対し、ベイズ統計の推定では、先見的知識を積極的に活用することで、データ数が少ない状況でも1/6に近い合理的な推定値を得ることができる。未来を予測する際、必要な情報が足りなければ人間は何らかの信念に基づいて予測を行うしかない。そのような状況において、事前に適切な主観確率を設定できるのであれば、情報の不足を補うために推論過程に積極的に活用すべきという思想がベイズ統計の背後に存在する。

頻度主義者によるベイズ統計への批判の矛先は、客観性が求められるサイエンスの世界に主観や信念を持ち込むことの非合理性にある[67]。事前分布の違いで推定結果が変わるベイズ統計の体系には、科学的手法としての致命的な欠陥があるという主張である。一方、頻度主義の統計学に完全な客観性があるかと言えば、決してそうではない。頻度主義にせよ、ベイズ統計にせよ、統計的推測を行うにはモデルが必要になる。モデルは人間の直観や経験、信念に立脚したものである。例えば、データは正規分布に従うというモデリングは、人間が定める仮定である。ニュートン力学の方程式も経験に基づく現象の模倣に過ぎない。したがって、ベイズ統計は非客観的であるが、頻度主義の統計学は客観的であるという主張に正当性はないと言える。

しかしながら、科学の方法論としてベイズ統計を活用するには、事前分布の選択に対する何らかの指針が必要であることに変わりはない。基本的にベイズ統計の事前分布の選択根拠は導かれた予測モデルの有用性に委ねられる。有用性の尺度の1つは、未来のデータに対する予測性能（汎化能力）である。現在は不完全なデータしかないが、事前情報を組み合わせて何とか有用な予測モデルを導きたい。事前分布の選択には恣意性があるが、実際の運用で有用性が認められれば、それで十分と考える。ベイズ統計の背後に存在するこのようなプラグマティズムが現代の科学や産業が志向するデータ駆動型アプローチと合流したことが、ベイズ統計の普及を後押しした側面もある。

(3) ベイズ統計の隆盛：その背景

(i) データの量的・質的な変化

ベイズ統計をメインストリームに押し上げた社会的要因の1つに、データの巨大化や多角化の流れが挙げられる。例えば、ゲノム情報のデータ解析では、サンプル数（例えば、検体数）がモデルに含まれる変数の個数（例えば、遺伝子数）に比べて圧倒的に少ない状況に頻繁に遭遇する。このような場合、推定対象のパラメータ数に対してデータの数が不足するため、パラメータの推定値等が一意に定まらない。これを不良設定問題という。古典統計学の枠組みでは、一般にこのようなデータから推定値を導くことはできない。

67　詳しくは、シャロン・バーチュ・マグレイン（冨永星訳）『異端の統計学ベイズ』草思社, 2013.（原書名：Sharon Bertsch McGrayne, *The Theory That Would Not Die: How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines, and Emerged Triumphant from Two Centuries of Controversy,* 2012.）

これに対してベイズ統計では、物理法則や経験的に獲得した確からしいパラメータの値等を事前分布として推論過程に取り込むことでデータ量の不足を補い、不良設定問題を解消する。また、ウェブデータやセンサー情報のような経時的に蓄積される大量データや一括処理が困難な巨大データを解析する際に、ベイズ統計に自然に備わる大量データの分割処理及び学習の機能が有効性を発揮する[68]。このようなデータの量的・質的な変化に柔軟に対応する上でベイズ統計は有用なソリューションを提供してきた。

（ii）モデリングの柔軟性

現代の統計学が対象とする自然・社会現象の多くは極めて複雑であり、現象を模倣するには柔軟なモデリング・スキームが求められる。例えば、全球規模の気象予測では、データ数よりも圧倒的に数が多い状態変数を含む物理モデルが用いられる[69]。材料開発では、設計パラメータの数は候補物質の構造多様性で決まる。例えば、低分子有機化合物の場合、10 の 60 乗個以上の候補物質が存在すると言われている[70]。このような膨大な数のパラメータと材料機能の関係を記述するには、柔軟性の高い統計モデルが必要になる。

一般に複雑なモデルを用いるとパラメータ数とデータ数の間にアンバランスが生じるため、上述のような不良設定問題を解消する必要が生じる。この問題においても同様に、ベイズ統計の事前分布を活用した情報補完が有効な解決手段となり得る。

（iii）不確かさの解釈容易性

統計学において、パラメータの推定値や予測の不確かさを定量的に評価することは極めて重要である。古典統計学の欠点の 1 つに、不確かさの定量における解釈の不自然さがある。

例えば、パラメータの推定値の信頼区間を求める際、一般に我々が知りたいことは信頼区間に真の値が含まれる確率であるが、頻度主義に基づく統計学の信頼区間はそのような意味を持たない。頻度論の 95% 信頼区間は、現在のデータと同じ条件で 100 個のデータセットを取得した際、各データセットから得られた 100 個の信頼区間のうち 95 個は真値を含むということを意味する。このような定義は、データ解析の実務家に極めて複雑な解釈を要求する。

これに対し、ベイズ統計の 95% 信頼区間はシンプルに真の値が含まれる確率が 95% ということを意味する。このような確率の解釈容易性もベイズ統計の利点である。不確かさの定量と意思決定の問題を対象とする領域（例えば、品質工学や防災等）でベイズ統計が広く普及した背景には、確率に対する解釈の容易性が要因の 1 つとして考えられる。

（iv）事後分布の近似計算法の発展

ベイズ統計の基本的な形は、1960 年前後にデニス・リンドリー（Dennis V. Lindley, 1923-2013）、レナード・サベッジ（Leonard J. Savage, 1917-1971）、エイブラハム・ウォルド（Abraham Wald, 1902-1950）らによって構築された。しかしながら、ベイズ統計の普及が本格化した時期は 1990 年代以降であり、そこに至るまで約 30 年という年月を要した。

複雑なモデルを用いるベイズ統計のデータ解析では、事後分布や予測値を求める際に非常に複雑な計算が必要になることが多い。1990 年代以降に高性能な汎用計算機が利用できるようになり、これを機に計算手法の研究が活発化・進展したことで、ベイズ統計はようやく実用的

---

[68]　詳しくは、C. M. ビショップ（元田浩ほか監訳）『パターン認識と機械学習―ベイズ理論による統計的予測―　上・下』丸善出版, 2012.（原書名 : Christopher M. Bishop, *Pattern Recognition and Machine Learning,* 2011）

[69]　参考文献として、樋口知之ほか『データ同化入門―次世代のシミュレーション技術―』（シリーズ予測と発見の科学 6）朝倉書店, 2011 がある。

[70]　Christopher M. Dobson, “Chemical space and biology,” *Nature,* Vol.432, December 2004, pp.824-828.

な統計学の体系になり得た。現在では様々な事後分布の近似計算法[71]が確立されている。

（v）オープンソースソフトウェアの登場

事後分布の計算スキームは解析者が設計した独自のモデルや事前分布に依存するため、汎用プログラムの開発が困難であるという問題があった。したがって、ベイズ統計で使用するプログラムは問題特化型になる傾向が強く、汎用性の低さゆえ、一般ユーザーにまで浸透するのに多くの時間を要することとなった。しかしながら、2000年代に入り、データ科学の世界に数多くのソフトウェア[72]が出現し、この流れの中で前述のWinBUGSのような汎用ソフトウェアが登場したことでベイズ統計のコモディティ化が急速に進行した。

（4）広がる応用領域

UQ（Uncertainty Quantification）と呼ばれる学問体系が存在する[73]。モデルには、数学的記述の誤りやデータ計測の誤差等、予測の不確かさをもたらす様々な要因が内在している。UQの目的は、不確かさの要因を特定し、不確かさの大きさを同定し、予測に基づく意思決定に対する不確かさの影響度を合理的に評価することである。

ベイズ統計の不確かさの解釈容易性がUQとの親和性をもたらし、ベイジアンUQ[74]と呼ばれる方法論の体系が構築されるに至った。地震工学や気象予測における流体解析、都市の輸送経路ネットワークの最適設計、ものづくり、自然災害のリスク評価など、ベイジアンUQは広範な問題に応用されている。

（i）機械学習やディープラーニングとの合流点

機械学習の手法は元来非確率的な設計思想から生まれたものが多く、そのような手法には基本的に確率論的な不確かさの評価を行うすべがない。非確率的な手法の例としては、ニューラルネットワークやサポートベクターマシンが挙げられる[75]。データ科学の歴史において、非確率的な思想から誕生し、後になり確率的な手法に拡張されたものが数多く存在する[76]。このような学術的展開から生まれた方法論を総じて、「統計的機械学習」と称することがある。

確率的な手法への拡張は、不確かさの評価のほかに、従来の機械学習の主なタスクであったデータの認識を行う「識別モデル」から、データの生成規則を学習して何かを生み出す「生成

⑺1　逐次モンテカルロ法、マルコフ連鎖モンテカルロ法、変分ベイズ法、Approximate Bayesian Computation（ABC）等が挙げられる。ABCは、確率分布から標本を生成する手法である。詳細は、Simon Tavaré et al., “Inferring Coalescence Times From DNA Sequence Data,” *Genetics,* vol.145 no.2, February 1997, pp.505-518.〈http://www.genetics.org/content/genetics/145/2/505.full.pdf〉を参照。

⑺2　例えば、R言語やPythonにより実装されたソフトウェア。前者は、The Comprehensive R Archive Network Website〈https://cran.r-project.org/〉で、後者は、Python Website〈https://www.python.org/〉で配信されている。

⑺3　UQについて詳細は、Ralph C. Smith, *Uncertainty Quantification: Theory, Implementation, and Applications,* Philadelphia: Society for Industrial and Applied Mathematics, 2014を参照。

⑺4　ベイジアンUQについては，Andrew M. Stuart, “Inverse problems: a Bayesian perspective,” ACTA Numerica, Volume 19, May 2010, pp.451-559を参照。

⑺5　詳しくは、第Ⅲ章「1（4）モデリング」の（i）及び（ii）を参照。また、非確率的な手法の例については、Trevor Hastie ほか（杉山将ほか監訳）『統計的学習の基礎―データマイニング・推論・予測―』共立出版, 2014.（原書名：Trevor Hastie et al., *The Elements of Statistical Learning: Data Mining, Inference, and Prediction,* Second Edition, 2009）を参照。

⑺6　ニューラルネットワークを確率的に動作する生成モデルに書き換えたのは、1980年代にジェフリー・ヒントン（Geoffrey Hinton）らによって提唱されたボルツマンマシンが最初である。ボルツマンマシンの提唱について詳しくは、人工知能学会監修, 神嶌編　前掲注⑷3を参照。

モデル」[77]への転換をもたらす[78]。例えば、所望の特性を有する画像・音声・自然言語等を自動的に生成・編集・変換・復元する問題などに、生成モデルを適用することができる。近年隆盛を極めるディープラーニングの研究でも同様に、生成モデルへの拡張が着実に進行している[79]。

(ii) シミュレーションとベイズ統計学

データサイエンスとシミュレーション科学の境界領域に、データ同化と呼ばれるデータ統合型シミュレーション技術の体系が存在する[80]。データ同化では、物理法則に基づき設計された数理モデルを観測データに適合させ、逆問題を解くことにより、直接的に観測できないモデル内の物理変数などを推定する。ベイズ統計は、データ同化の方法論と計算技術の礎を担っている（詳しくは、「3　データ同化」を参照）。

近年、シミュレーションから生成されたデータを対象とするデータサイエンスに注目が集まっている[81]。これは、シミュレーションモデルに似せた統計モデルを構築し、大規模シミュレーションの計算を統計モデルに代替させることで計算コストの大幅な削減を図るものである。この統計モデルの構築は、当該シミュレーションモデルによる実際の入力値と出力値を基にして行われるが、多くの場合、入出力データを取得するには、本来であれば大規模シミュレーションを実施する必要がある。そこで、実験計画法[82]を用いて最適な入力値の選択を行い、シミュレーション回数の抑制と推定精度の改善を図る。実験計画法では、ベイズ統計の考えに基づいて推定値の不確かさを評価し、その不確かさの下で最適な意思決定（入力値の選択など）を行う（詳しくは、「4　エミュレーション」を参照）。

このような解析手法がもたらす科学的・産業的価値は極めて大きい。例えば、材料開発では、候補材料の特性を評価するためのシミュレーション[83]に膨大なコストを要するため、従来は極めて少数の候補材料のみを評価の対象とせざるを得なかった。仮にシミュレーションモデルを統計モデルで代替できれば、例えば、数億個という規模の有機化合物を対象にした網羅的な評価を速やかに実行できるようになる。このような研究は、マテリアルズインフォマティクス[84]と呼ばれ、急速に進展している。

⑺　特に近年は、変分自己符号化器（Variational Autoencoder: VAE）や敵対的生成ネットワーク（Generative Adversarial Networks: GAN）と呼ばれるディープラーニングに基づいた生成モデルに注目が集まっている。VAEについては、Carl Doersch, “Tutorial on Variational Autoencoders,” arXiv:1606.05908v2, August 2016.〈https://arxiv.org/pdf/1606.05908.pdf〉; GANについては、Ian J. Goodfellow et al., “Generative Adversarial Nets,” arXiv:1406.2661, June 2014.〈https://arxiv.org/pdf/1406.2661.pdf〉を参照。

⑻　やや厳密性に欠く説明になるが、生成モデルにベイズの定理を適用することで、コンピュータを用いて仮想的なデータを作り出すことができる。

⑼　機械学習の最大の国際会議の1つである「Conference on Neural Information Processing Systems」（NIPS）では、「Bayesian Deep Learning」という特別ワークショップが継続的に開催されている。“NIPS 2017 Workshop.” Bayesian Deep Learning Website〈http://bayesiandeeplearning.org/〉

⑽　樋口ほか　前掲注⑹

⑾　例えば、岩波データサイエンス刊行委員会編　前掲注⑸の「小特集　シミュレーションとデータサイエンス」を参照。

⑿　解析手法として、機械学習の分野で提唱された能動学習やベイズ最適化、空間統計学に由来するクリギング法等が挙げられる。使用目的や設計思想に多少の違いはあるが、いずれも実験計画法の一種と言うことができ、基本原理には共通点が多い。詳しくは、Eric Brochu et al., “*A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning,*” arXiv:1012.2599, December 2010.〈https://arxiv.org/pdf/1012.2599.pdf〉

⒀　シミュレーションにおいては、第一原理計算や分子動力学シミュレーション等が多用されてきた。

⒁　物質･材料科学とデータ科学の新しい学際領域。吉田亮「物性研究におけるデータ科学活用の現状と展望」『機能材料』36(9), 2016.9, pp.23-29を参照。

3　データ同化

統計数理研究所准教授　上野　玄太

(1) はじめに

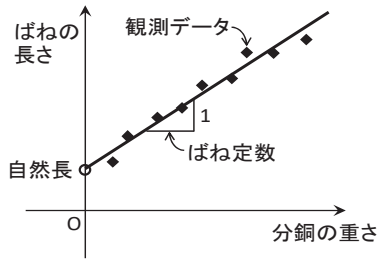
自然科学や工学はもとより、経済学などにおいても、複雑な現象の予測や解析の手段としてシミュレーションが有効な手段として知られている。シミュレーションでは、知りたい対象（変数）の時間変化を表現するような方程式を立て、その解がどのように変化していくかをコンピュータで精密に計算する。対象が風速のように物理的なものであれば、運動方程式の解を求めることで、今後の風速や圧力がどのように変化していくか、すなわち天気予報につなげていくことができる。シミュレーションは、数学的な厳密解を求めることができない方程式においても高精度で解が得られる非常に便利な方法である。

シミュレーションでは、変数のある時点の値（これを初期値という。）を決めると、立てた方程式に従ってその後は自動的に将来の変化を計算することができる。しかし、実際にシミュレーションを行うと、計算結果は思うように現実を再現できないことが多い。その原因はシミュレーションの設定の不備にある。計算結果の良し悪しは、初期値や方程式の妥当性によるが、多くの場合、初期値は不正確であり、方程式は不完全である。データ同化とは、そういったシミュレーションの不備を、観測データで補い、より正確な予測を可能にする方法である。現実の観測データを積極的に活用してシミュレーションの初期値や方程式を改善するのである。

(2) データ同化とは

データ同化とは、端的に言えば観測データにシミュレーションを当てはめる操作である。例として、ばねに重さの異なる分銅をつるし、重さとばねの長さの関係を調べる実験を挙げよう。つるした各分銅に対してばねの長さを点でプロットし、点の集団の中央を通るような直線を引く。この直線の傾きの逆数がばね定数、切片がばねの自然長として、このばねを適切に表現する値が得られる（図7）。

図7　分銅の重さとばねの長さの関係を調べる実験



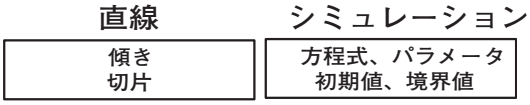
（出典）筆者作成。

この例をデータ同化の操作に対応させると、図8に示すように、直線がシミュレーションに、傾きと切片は方程式と初期値にそれぞれ対応する。方程式によっては、特性係数（パラメータ）(85)

(85)　例えば、粘性係数（流体の粘り気の強さを表現する係数。水は小さい値、蜂蜜は大きい値となる。）や拡散係数（物質の密度や濃度が空間的に拡散する過程を表現する方程式において拡散の速さを表現する係数）。

や境界値[86]を含むが、それぞれ直線の傾きと切片に対応する。これらのシミュレーションの要素（方程式、パラメータ、初期値、境界値）を調整して、計算結果が観測データの中央を通るようにすることで、観測データを反映した「調整済み」のシミュレーションができ上がる。これをデータ同化モデルと呼び、もとのシミュレーションよりも精度の高い予測が可能になる。

図８　直線とシミュレーションの構成要素の対応



（出典）筆者作成。

（3）データ同化の目的

データ同化は、不完全であるシミュレーションの欠点を補うものであり、次のような多くの目的に用いられる[87]。

・シミュレーションを用いた予測を行うための最適な初期値を求めること[88]。
・シミュレーションを行うときの最適な境界値を求めること。
・シミュレーションで使用されるパラメータの表現方法や最適値を求めること。
・データ同化モデルの計算結果として得たデータセット（再解析データ）を用いて対象を理解すること。
・観測データの誤差の評価、各観測の重要性を調べることにより、効果的な観測システムを検討すること。

（4）データ同化の方法

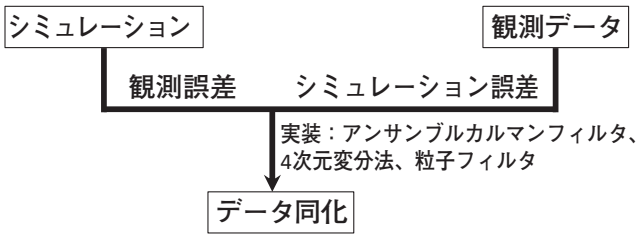
既に述べたように、データ同化は観測データを用いてシミュレーションを改善する方法である。古くはそれまでのシミュレーション結果を捨てて観測データと置き換えて計算するという素朴なものであった。ところが、変数の値としては完璧であるはずの観測データと入れ替えるからといって、必ずしも予測の精度が高くなるわけではない。シミュレーションの構成要素（図8）は現実の現象を表現するには力不足であり、また観測データには誤差が含まれるからである。そのため、一度に置き換えずに複数の時間ステップに重みを分けて徐々に置き換える方法、そして現在主流であるベイズ的なアプローチ（シミュレーションを基にした事前分布、シミュレーション変数と観測データの関係を表す尤度を組み合わせ、事後分布を得るアプローチ）へと変遷を遂げてきたのである[89]。

図９に、データ同化の仕組みを示す。シミュレーションと観測データが左右の皿に載ったてんびんを想定した図である。右の皿は観測データを指し、右腕の長さはシミュレーションの誤

86　時間変化に加え空間変化も再現するシミュレーションにおいては、計算対象とする空間の境界位置における変数の値を決めておく必要があり、これを境界値と呼ぶ。初期値は時間変化の出発点であるが、境界値は空間変化の出発点に当たる。
87　データ同化の目的については、蒲地政文ほか「熱帯太平洋での気候変動に関連した海洋データ同化の最近の発展」『統計数理』54(2), 2006, pp.223-225.〈http://www.ism.ac.jp/editsec/toukei/pdf/54-2-223.pdf〉を参考にした。
88　天気予報ではこの目的のためにデータ同化が利用されている。
89　Ichiro Fukumori, “Chapter 5 Data Assimilation by Models,” *International Geophysics*, Volume 69, 2001, pp.245-251.

差[90]の大きさを示す。同様に、左の皿はシミュレーションを指し、左腕の長さは観測誤差の大きさを示す。つまり、一方の誤差が他方の誤差より大きい場合、他方の腕が長くなり、その結果、誤差の少ない他方の情報がデータ同化の結果により強い影響を及ぼすことになる。データ同化ではこの関係に基づき、データ同化による予測の精度が高くなるように両者の誤差の大きさや相関関係を調整するのが標準的な考え方となっている。

図９　データ同化の仕組み



（出典）筆者作成。

てんびんの両皿と両腕が定まるとデータ同化の実行に移る。実装（プログラミング）においては、計算コストやシミュレーションの複雑さなどに応じて、幾つかの方法から選択される[91]。

（5）おわりに

データ同化によって得られる再解析データは、科学研究はもちろん、実用的にも広く利用されている[92]。将来の効率的な観測システムの設計[93]にも役立つ技術である。

４　エミュレーション

統計数理研究所准教授　中野　慎也

実用的な計算機シミュレーションは、1回の実行にかなりの時間を要する。高精度のシミュレーションになると、最新のスーパーコンピュータを用いても数日以上の計算時間が掛かるものも少なくない。さらに、不確実性を含む実問題にシミュレーションを応用する際には、様々な設定で何度も繰り返しシミュレーションを実行する必要が生じる。

例えば、地球温暖化予測を行う場合、今後の二酸化炭素排出量を設定してシミュレーションを行う必要があるが、現時点で将来の二酸化炭素排出量を正確に知ることはできない。そこで、

⑼⓪　正確には、図8に示す構成要素の誤差。

⑼①　実装が容易な「アンサンブルカルマンフィルタ」（多数の初期値や境界値に基づく多数のシミュレーション結果（アンサンブル）を基にデータ同化を行う手法の1つ）、超大規模なシミュレーションを扱える「4次元変分法」（初期値や境界値を未知数とし、観測データと整合的なシミュレーション結果を与える最適値を探索するデータ同化手法の1つ）、一般的な（非線形性の強い）事象に強い「粒子フィルタ」（アンサンブルカルマンフィルタと同じく、多数のシミュレーション結果（ここでは粒子と呼ぶ。）を基にデータ同化を行う手法の1つ）などが代表的である。詳しくは、樋口ほか　前掲注⑹⑼, p.14を参照。

⑼②　詳細は、気象庁「JRA-55: 気象庁55年長期再解析」JRA projectウェブサイト〈http://jra.kishou.go.jp/JRA-55/index_ja.html〉を参照。具体例として、農林漁業分野での基礎的な気象データベースとしての利用、ロケット打ち上げ時の大気参照データとしての利用が挙げられる（大野木和敏「長期再解析JRA-25」『天気』54(9), 2007.9, p.776.〈http://www.metsoc.jp/tenki/pdf/2007/2007_09_0013.pdf〉を参照）。

⑼③　例えば、淡路敏之ほか『データ同化―観測・実験とモデルを融合するイノベーション―』京都大学学術出版会, 2009, pp.242-245では、具体例として、北太平洋の塩分濃度の予測には、特定の海域としてオホーツク海・ベーリング海及び黒潮域に観測器を配置することがコストパフォーマンスの点で優れることが記されている。

二酸化炭素排出量の不確かさが予測にどの程度影響するのかを評価するため、様々な異なる設定でシミュレーションを実行することになる。シミュレーションを用いて機械やシステムの設計を行う場合においても、システム内の不確かな箇所を微調整しながらシミュレーションを繰り返し実行する作業はどうしても避けられない。

このように、不確かさを含む問題において、シミュレーションを用いた研究・開発には多大な時間が必要となる。そこで、限られた回数のシミュレーション結果を基に、与えられた設定・入力に対するシミュレーションの出力を予測する簡単なモデルを構築し、これを用いて不確実性の評価や設計の調整を行うことが考えられるようになった。このようなモデルを「統計的エミュレータ」又は単に「エミュレータ」と呼んでいる。

エミュレータは、シミュレーションの入力と出力の関係を統計的な手法で学習し、これによって、シミュレーションの挙動を模倣する。エミュレータを使えば、様々な入力に対するシミュレーションの出力を短時間で予測できるため、不確かさの評価や、リスク評価、システム設計などを効率的に行うことができる。また、統計的な手法に頼るのではなく、精度は低いが高速で実行できるシミュレーションとエミュレータの手法を組み合わせて、高精度シミュレーションの挙動を模倣させる手法も提案されている[(94)]。エミュレータは、上述の気候変動予測[(95)]、環境評価[(96)]、疫学[(97)]などの分野において、不確かさの評価、リスク評価などに活用されるようになっているほか、機械設計への応用[(98)]などの研究もされており、シミュレーションを用いたUQを行う有力な手法の1つとなっている。このように、不確かさやリスクの評価を行う際には、単なるシミュレーションだけではなく、エミュレータのような統計的手法を活用することも今後ますます重要になると考えられる。

5　秘匿処理技術（匿名化）

統計数理研究所准教授　南　和宏

(1) はじめに

近年、インターネット上での購買活動や、個人間の情報発信を可視化するソーシャルメディア、ネットワーク上でユーザーのデータを管理するクラウドサービスによる情報共有の機会が飛躍的に増えており、また現実社会における人、モノの活動もモノのインターネット（Internet of Things: IoT）を用いた様々なセンシング技術を通してデジタル化され、大量に蓄積されている。このような「ビッグデータ」を様々な手法で分析することで、我々の社会生活、行動に関する新しい知見が続々と明らかになっている。また「個人情報の保護に関する法律」（平成15年法

---

(94) Mark C. Kennedy and Anthony O'Hagan, "Predicting the Output from a Complex Computer Code when Fast Approximations are Available," *Biometrika,* Vol.87 No.1, March 2000, pp.1-13.

(95) Jonathan Rougier et al., "Analyzing the Climate Sensitivity of the HadSM3 Climate Model Using Ensembles from Different but Related Experiments," *Journal of Climate,* Vol.22 No.13, July 2009, pp.3540-3557.〈http://journals.ametsoc.org/doi/pdf/10.1175/2008JCLI2533.1〉

(96) Peter C. Young and Marco Ratto, "A unified approach to environmental systems modeling," *Stochastic Environmental Research and Risk Assessment,* Vol. 23 Issue 7, October 2009, pp.1037-1057.

(97) Ioannis Andrianakis et al., "Bayesian History Matching of Complex Infectious Disease Models Using Emulation: A Tutorial and a Case Study on HIV in Uganda," *PLOS Computational Biology,* Vol.11 Issue 1, e1003968, January 2015, pp.1-18.〈http://journals.plos.org/ploscompbiol/article/file?id=10.1371/journal.pcbi.1003968&type=printable〉

(98) Robert B. Gramacy and Herbert K. H. Lee, "Bayesian Treed Gaussian Process Models With an Application to Computer Modeling," *Journal of the American Statistical Association,* Vol.103 No.483, September 2008, pp.1119-1130.

律第 57 号）が平成 27 年に改正され、様々な組織、企業が所有するビッグデータの組織を越えた流通を実現する法的な道筋もできた。

しかしビッグデータは、個人や組織の活動に関する機密情報を含むため、2 次利用の前に個人識別情報を取り除く秘匿処理（以下「匿名化」）を行う必要がある。ただし、匿名化は、元データから個人の氏名、ID 等の識別子を削除するといった単純な作業ではない。実際、2006 年に米国のインターネットサービス会社 AOL が約 65 万人分のユーザーの検索ログを匿名化した上で公開した際、ジョージア州在住の女性が特定される事件[99]が起き、データ公開によるプライバシー侵害のリスクが広く一般に認知された。

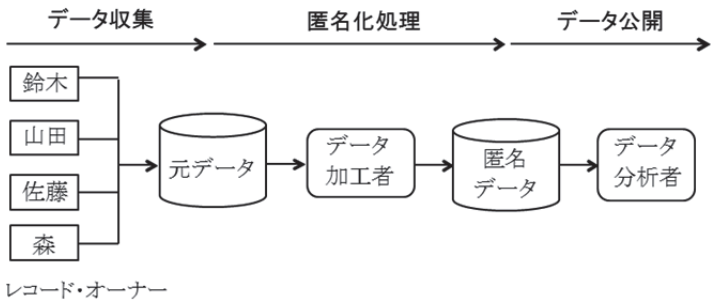
本節では、匿名化データを作成する際に注意すべき情報漏えいシナリオを示し、そのようなリスクに対処する代表的な匿名化技術を解説する。

（2）匿名化技術の目的

図 10 に匿名化技術の利用形態を示す。データ収集時にデータ加工者は各個人の属性情報を収集する。「年齢」、「性別」といった一般的属性に加え、「商品の購入」、「位置情報」といった個人の行動も属性となり得る。ここでは表形式のデータを想定し、各個人の属性値は割り当てられた行レコードに格納されるとする。データ公開時には、元データを管理するデータ加工者がプライバシー保護のための匿名化等の処理を行い、不特定多数のデータ分析者に向けて匿名化されたデータセットを公開する。データ分析者による様々なデータ分析のニーズに答えるため、匿名化処理を行うデータ加工者は可能な限り元データに近いデータセットを公開することが求められる。

セキュリティの観点から、匿名化処理を行うデータ加工者は信頼できる主体とし、情報を提供する個人のプライベートな情報が漏えいしないように適切な処理をすると仮定する。それに対し、データ分析者は悪意を持っている可能性があり、その場合、入手した匿名化されたデータセットから個人の機密情報を取得又は推測しようとする。ただし、どのデータ分析者が悪意を持つかをデータ公開時に判別することは不可能であり、信頼できる分析者にのみ情報を提供する手法（暗号化やアクセスコントロール）は利用できない。匿名化におけるセキュリティ上の課題は、データ分析者のニーズにできるだけ応えつつ、悪意のあるデータ分析者への機密情報の漏えいを防ぐことにあるといえる。

図 10　匿名化技術の利用モデル



（出典）筆者作成。

---

[99] Michael Barbaro and Tom Zeller Jr., “A Face is Exposed for AOL Searcher No. 4417749,” Aug. 9, 2006, New York Times Website〈http://www.nytimes.com/2006/08/09/technology/09aol.html〉

（3）匿名データの情報漏えいリスク

匿名化処理の第一歩は個人を識別する名前、住所等（識別子情報）の削除であるが、これで十分であろうか。表1[100]に示す医療データの場合、名前の列を削除することで表2のように名前と病名の関連は分断され、プライバシー保護が達成できたように見える。しかし、他の公開された様々なデータも参照され得ることを忘れてはならない。

例えば米国では、表3のような投票者リストが一般に入手可能である。もし悪意のある者（攻撃者）が、表3の「高橋三郎」が表2の匿名化された医療データにも含まれることを知っていれば、「職業、性別、年齢」の3つの属性を照合することで、表2の3番目のレコードと一致することを発見し、「高橋三郎」の病名が「エイズ」であることを突き止めてしまう。このように複数のデータセットに共通する属性を照合することで、互いのレコードを関連付ける行為は「レコードリンク攻撃」と呼ばれる。

表1　医療データの例

名前	職業	性別	年齢	病名
鈴木太郎	技術者	男	35	肝炎
木村二郎	技術者	男	35	ねんざ
高橋三郎	弁護士	男	38	エイズ
田中優子	作家	女	30	インフルエンザ
上田聡子	作家	女	30	エイズ
岡本英子	ダンサー	女	30	エイズ
中村和子	ダンサー	女	30	エイズ

（出典）筆者作成。

表2　識別子を削除した医療データ

職業	性別	年齢	病名
技術者	男	35	肝炎
技術者	男	35	ねんざ
弁護士	男	38	エイズ
作家	女	30	インフルエンザ
作家	女	30	エイズ
ダンサー	女	30	エイズ
ダンサー	女	30	エイズ

（出典）筆者作成。

表3　投票者リストの例

名前	職業	性別	年齢
鈴木太郎	技術者	男	35
木村二郎	技術者	男	35
高橋三郎	弁護士	男	38
田中優子	作家	女	30
上田聡子	作家	女	30
岡本英子	ダンサー	女	30
中村和子	ダンサー	女	30

（出典）筆者作成。

1997年、米国マサチューセッツ州で医療データが公開された際、このようなレコードリンク攻撃による個人情報の漏えい問題が起きた[101]。名前、社会保障番号、住所、電話番号等個人を特定する属性は取り除かれたにも関わらず、生年月日、郵便番号、性別といった情報を州の投票者リストと突き合わせることで当時の州知事の病名が特定されたのである。当時、マサチューセッツ州では、生年月日と郵便番号の組合せで97%の住民の特定が可能であった。

---

(100)　表中の氏名は実在しない架空のものである。以下同。

(101)　Daniel C. Barth-Jones, “The ‘Re-Identification’ of Governor William Weld’s Medical Information: A Critical Re-Examination of Health Data Identification Risks and Privacy Protections, Then and Now,” July 2012. 〈https://iapp.org/media/pdf/knowledge_center/Re-Identification_of_Welds_Medical_Information.pdf〉

レコードリンク攻撃が成立しない場合でも個人の機密情報が漏えいする危険性はある。表3の「岡本英子」は、表2の6番目及び7番目のレコードと3つの属性において一致するが、2つのレコードのどちらが「岡本英子」に該当するかは分からない。しかし、いずれも「病名」が「エイズ」であるため、機密情報が漏えいしてしまう。このように特定の個人とその機密属性を関連付ける行為は「属性リンク攻撃」と呼ばれる。

（4）*k*-匿名化

（3）で述べたレコードリンク攻撃への対策として、1998年、ピエランジェラ・サマラティ（Pierangela Samarati）らは「*k*-匿名化」と呼ばれる概念を提案した[102]。*k*-匿名化は直感的に理解しやすく、またこれを行うための効率的なデータ加工手法が開発されており、現在に至るまで代表的な匿名化技術として用いられている。

k-匿名化は、攻撃者が標的となる個人の属性情報を既に入手しており、その中から候補となるレコードを絞り込むことができると仮定した場合、このような攻撃者が候補となるレコードを*k*個以下に絞り込めないことを保証する。

k-匿名化では攻撃者が利用する情報を明確に規定している。公開されるデータセットには、図11に示す4種類の属性情報が含まれる。「識別子」は名前や米国の社会保障番号のように直接個人を特定する情報、「準識別子」は住所等、間接的に個人を特定する情報、「機密情報」は収入や病名といった個人のプライバシーに関する情報であり、これら3つに当てはまらない属性が「非機密情報」となる。*k*-匿名化では攻撃者が各個人の準識別子に属する属性情報を全て知っていると仮定する。これは既に流通しているデータセットが何であるか妥当な仮定を設けることが困難であり、最悪の場合を想定する必要があるためである。

図11　レコード属性の分類

識別子	準識別子	機密情報	非機密情報

（出典）筆者作成。

k-匿名化のためのデータ加工は以下の手順で行われる。まず、データセットの各レコードから識別子に属するデータを削除する。次に、準識別子のデータに対して一般化の処理を行い、一般化された準識別子データの組合せで決まる各グループが必ず*k*個以上のレコードを含むようにする。一般化とは、属性情報を抽象度の高いより一般的な情報に置き換える処理である。例えば、「技術者」という職業をより一般的な「専門職」といった名称に置き換える。

準識別子に分類される属性値の一般化により、レコードリンク攻撃の候補となるレコード数を*k*個未満に絞り込むことを防止できる。つまり攻撃者が全ての個人の準識別子の属性情報の値を知っていたとしても、*1/k*以下の確率でしか標的とする個人のレコードを特定することができない。加えて、*k*-匿名化は機密情報及び非機密情報の属性値に関しては、一般化も含め、情報を修正する必要がないという利点を有する。

---

[102] Pierangela Samarati and Latanya Sweeney, “Generalizing Data to Provide Anonymity when Disclosing Information (abstract),” Proceedings of the Seventeenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, Seattle: ACM Press, 1998, p.188.

表4は表1の医療データを *3*-匿名化したものである。この場合の準識別子は「職業、性別、年齢」の3属性であり、7つのレコードは、3属性が同じ組合せとなる3つのレコードと4つのレコードの2つのグループに分割される。例えば、攻撃者が準識別子について、「専門職、男、[35-40]」のグループまで絞り込んでも、その中のどのレコードが標的とする個人のものかは分からない。

表４　*3*-匿名化された医療データ

職業	性別	年齢	病名
専門職	男	[35-40]	肝炎
専門職	男	[35-40]	ねんざ
専門職	男	[35-40]	エイズ
芸術家	女	[30-35]	インフルエンザ
芸術家	女	[30-35]	エイズ
芸術家	女	[30-35]	エイズ
芸術家	女	[30-35]	エイズ

（出典）筆者作成。

k-匿名化データを作成する場合、パラメータ *k* の選択が重要である。*k* 値を大きくするほど、個人のレコードが識別されるリスクは減少する。その一方、準識別子の属性情報は大幅に一般化されてしまい、データの有用性も損なわれてしまう。このため、*k*-匿名化によるリスク低減を果たしつつ、匿名化による情報損失を最小限に抑える最適化手法が長年にわたり活発に研究されてきた[103]。

また、準識別子の適切な選択も重要である。*k*-匿名化による機密情報の保護は、あくまで「データ加工者が準識別子を適切に選択した」という前提で成立する。原則として、攻撃者が知る可能性のある属性情報は全て準識別子に分類する必要がある。もしこれを誤って機密情報に分類した場合、攻撃者は、準識別子の各属性が同じ組合せとなる *k* 個のレコードに対し、誤って「機密情報」に分類され、一般化処理されなかった属性の値の差異に着目することにより、更に絞り込むことが可能になってしまう。

（5）*k*-匿名化の派生指標

既に述べたように、*k*-匿名化は、標的とする個人のレコードの候補が *k* 個未満に絞り込まれないことを保証する。しかし、候補となった各レコードの機密情報が偶然同じになった場合、それ以上候補を絞り込まずとも機密情報の推定が可能になってしまう（属性リンク攻撃）。例えば、表5は「職業、性別、年齢」の準識別子に関して *3*-匿名化された医療データであるが、グループの全てのレコードが機密情報として同一の病名「エイズ」を持つため、このグループに絞り込まれた個人の病名が推定される。

---

[103] Latanya Sweeney, "*k*-anonymity: a model for protecting privacy," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems,* Vol.10 No.5, 2002, pp.557-570; Kristen LeFevre et al., "Incognito: Efficient Full-Domain *K*-Anonymity," *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data, SIGMOD '05,* New York: ACM, 2005, pp.49–60.

表５　病名が同一の *3*- 匿名化された医療データ

職業	性別	年齢	病名
芸術家	女	[30-35]	エイズ
芸術家	女	[30-35]	エイズ
芸術家	女	[30-35]	エイズ

（出典）筆者作成。

表６　*3*- 多様化された医療データ

職業	性別	年齢	病名
芸術家	女	[30-35]	エイズ
芸術家	女	[30-35]	はしか
芸術家	女	[30-35]	胃潰瘍

（出典）筆者作成。

この問題を克服するため、アシュウィン・マカナバヤラ（Ashwin Machanavajjhala）らは「*l*- 多様化」という概念を提案した[104]。*l*- 多様化は、準識別子の各属性値が同じ組合せとなるグループの各レコードが少なくとも *l* 個の異なる機密情報のデータを持つようにすることである。これは各グループに *l* 個以上のレコードがあることを意味するので、自動的に *l*- 匿名化の要件を満たす。表６の医療データは *3*- 多様化され（同時に *3*- 匿名化の要件も満たしている）、同じグループに３つの異なる病名が含まれているため、病名が推定されない。つまり、*3*- 多様化は *3*- 匿名化よりも安全であるといえる。

しかし例えば、患者の病名の 95% がインフルエンザで 5% がエイズという医療データがあり、その中のある準識別子のグループにおいて、病名の割合がそれぞれ 50% と 50% となったとする。このグループは *2*- 多様化の要件を満たしているが、それでも病名がエイズであることは、全体で見た場合（5%）よりもはるかに高い 50% の確率で推定されてしまう。このような問題を解決するために李佇輝（Ninghui Li）らは「*t*- 近似性」と呼ばれる概念を提案している[105]。*t*- 近似性は、機密情報について全体の頻度分布と各グループでの頻度分布の類似性を定量化し、その差を一定の値 *t* 以下であることを要求するものである[106]。

（6）多次元データの匿名化

購買履歴や位置情報といった行動履歴データは、莫大な数の属性を持つ多次元データである一方で、外観識別性が高く一般的にはこれを準識別子として取り扱う必要がある。つまりこうしたデータには莫大な数の準識別子が含まれるため、各属性の組合せが類似するレコードを複数見つけることは通常難しい。

このような多次元データに対して *k*- 匿名化のための一般化処理を行うと、著しくデータの有用性が劣化してしまう[107]。多次元データの匿名化に関する既存研究では、攻撃者が知り得る準識別子の数に上限を設けるなどによって対応しているが、依然として根本的な解決策を提示した研究が見当たらないのが現状である。

104　Ashwin Machanavajjhala et al., “*ℓ*-Diversity: Privacy Beyond *k*-Anonymity,” *ACM Transactions on Knowledge Discovery from Data (TKDD),* Vol.1 No.1, March 2007.

105　Ninghui Li et al., “*t*-Closeness: Privacy Beyond *k*-Anonymity and *ℓ*-Diversity,” *2007 IEEE 23rd International Conference on Data Engineering*, IEEE, 2007, pp.106-115.

106　ただし、*t*- 近似性自体にはレコードリンク攻撃を防ぐ要件は入っておらず、*k*- 匿名化と組み合わせる必要がある。また *t*- 近似性にはデータの有用性を著しく低下させる欠点があり、プライバシーとの両立が課題となる場合が多い。

107　この問題は「次元の呪い」と呼ばれている。Charu C. Aggarwal, “On *k*-Anonymity and the Curse of Dimensionality,” *Proceedings of the 31st International Conference on Very Large Data Bases,* Trondheim, Norway: 2005, pp.901–909. 〈http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.60.3155&rep=rep1&type=pdf〉

Ⅳ　人文科学におけるデータ活用

国立情報学研究所准教授　北本　朝展

1　はじめに

人文科学は幅広い学問領域を含むが、その基本となるのは「読む」という行為である。例えば書籍を読む、美術を「読む」といった行為を通して、人類の知的資産に込められた情報の意味を解釈し、それを体系化して知識を積み上げていくとともに、その価値に対する判断を下していく。従来は「読む」行為が人間に限定されていたため、テキストを人間が詳細に読み解き、それに注釈をつけることが人文科学の基本的な研究スタイルであり、記憶した多くのテキストから関連性を発見する知的訓練が研究の基礎的能力となっていた。

こうした精読（close reading）に対して、近年では遠読（distant reading）という考え方が広まってきた。これはテキストをそのまま読むのではなく、情報技術を用いて定量的に分析するなど、全体像を俯瞰（ふかん）できるように表現又は可視化し、部分を見るだけでは分からない構造や関係性を解釈するという読み方である。実際に使われた（しばしば大規模な）言語データの集合であるテキストコーパスを分析するコーパス言語学などの研究は従来から盛んに行われてきたが、情報技術の発展に伴ってテキストの背後に隠れる構造や関連性を新しい方法で解釈できるようになれば、人文科学におけるデータ活用に大きなインパクトを与えると考えられる。

2　人文科学と画像データ

人文科学におけるデータは、文字情報であるテキストが中心的な存在ではあるが、近年は他のデータも重要性を増している。例えば画像データである。人文科学の研究対象をデジタル化する場合、その成果物は画像データとなることが多いため、画像データはテキストデータと並んで人文科学におけるデジタルアーカイブの中心的な存在である。

このデータを検索することを考えると、最もニーズが大きいのはやはりテキスト検索であり、画像を自動分析してテキストに変換する光学的文字認識（Optical Character Reader: OCR）への期待は大きい。OCR の研究には長い歴史があり、現代の印刷文字については高い精度を達成できているものの、手書き文字やくずし字となると実用的な精度を達成するソフトウェアの構築はいまだに困難であり、例えばくずし字 OCR ソフトウェアは対象を単一文字認識に限定するなど、難易度を下げることで実用性を高めている[108]。ただし、近年はディープラーニングなどの機械学習技術が発達しているため、文字認識についても精度向上への期待が高まっている。

人工知能を用いた画像の自動分析や自動タグ付けなどについては、人文科学分野への適用はまだ始まったばかりである。ディープラーニングツールの普及に伴って、データさえそろえれば人文科学者でも基本的な処理を試せる時代になってきた。ただし人文科学においては、技術的な性能評価よりもリサーチクエスチョン（研究における基本的な問い）に対する知見が重要なため、機械学習がどのような新しい知見を生み出したかを批判的に見る視点も欠かせない。絵画を 1 点ずつ詳細に見るのではなく、多数の絵画の分析から時代のトレンドを捉えるなど、精読と遠読のアプローチを使い分けてデータを俯瞰（ふかん）的に見る研究が今後は増えていく

---

[108]　山本純子･大澤留次郎「古典籍翻刻の省力化―くずし字を含む新方式 OCR 技術の開発―」『情報管理』58 巻 11 号, 2016.2, pp.819-827.

と考えられる。

3　人文科学と地理データ

人文科学でのデータ活用において伝統的に重要な地位を占めてきたのが地理データである。特に歴史学や考古学においては、どこで何が起こったか、又はどこで何が発見されたかという情報が重要な役割を果たしてきた。地理情報システム（Geographic Information System: GIS）を活用して情報を地図上にマッピングする手法は従来から用いられてきたが、オープンソースによって導入コストが低下し、クラウドサービスによって共有コストも低下したため、研究での利用がより一般化する流れが生じている。また、マッピングしたデータの空間分布を調べることで、過去の社会の特徴や現代の人の動きを追跡するなどの多彩な研究が進行中である。さらに近年はモバイルアプリの活用が地理情報収集の可能性を広げている。

しかしマッピングには人文科学データ特有の困難が生じる場合もある。例えば緯度経度が付与されていないデータに対しては、住所から緯度経度への変換（ジオコーディング）や、地図の位置合わせ（ジオリファレンス）といった処理が必要となるが、そこで歴史的な地名のデータベースなどが必要となる場合があり、その構築には多大なコストを要する。したがって基盤的なデータについては、オープンデータとして共有する文化が広がることが望まれる。

4　人文科学と動画データ・3次元データ

動画データや3次元データなども活用が広がっている。動画データは舞踊など動きの記録に必要なだけでなく、映画やアニメーション、現代アートなど、映像として残された文化遺産の保存にも必要である。特に映画ではフィルムの劣化が進んでおり、できるだけ早くデジタル化して長期保存する必要があるものの、データ量が膨大なために費用負担が重いという課題がある。また、アニメーションは日本の現代文化を象徴する文化遺産となり得るが、デジタル化して活用するには権利処理が難しいという課題がある。

一方、3次元データについては、レーザーを用いた精密な3次元計測や、多数の視点から撮影した写真を組み合わせた簡易3次元計測などが、考古学やミュージアムなどで広く活用されている。考古学では遺跡の3次元計測にドローン（Unmanned Aerial Vehicle: UAV）なども用いられており、今後は3次元データが増加する見込みである。また、3次元データにより3Dプリンタで複製した文化遺産に触れる展示や、バーチャルリアリティを使って楽しめる展示など、3次元データはエンターテインメントへの利用可能性が高いと考えられる。ただし3次元計測ができない過去の遺跡の復元となると、膨大な手作業と専門家による考証などが必要となり、3次元データの構築に大きなコストを要する場合がある点には注意が必要である。その他、医療機器の技術などを応用して作成した3次元断面データは、物体内部を輪切りにして調べる場合に適しており、考古学などで活用されている。

5　デジタル・ヒューマニティーズ

データを用いた人文科学研究は、一般にデジタル・ヒューマニティーズ（Digital Humanities: DH）と呼ばれる。日本における一般的な呼称である「人文情報学」では人文科学と情報学が並列しているが、デジタル・ヒューマニティーズという呼称は「情報学的ツールを活用した人文科学研究」との立場がより明確である。

DHの目的は、人文科学のリサーチクエスチョンに答えることであり、どんなデータを対象とし、いかに情報学的ツールをうまく使いこなすかが研究の重要なポイントとなる。そして研究データを共有する際に「FAIRデータ」の原則が重要となるのはⅡ章の「5（2）長期的なアクセシビリティの確保（FAIRデータ）」で述べたとおりであるが、特に相互運用性（interoperable）の観点から人文科学研究に大きなインパクトを与え得る2つの標準を紹介したい。

（1）Text Encoding Initiative（TEI）

「Text Encoding Initiative」（TEI）[109]は1987年に始まったコンソーシアムであり、マークアップ言語（Extensible Markup Language: XML）[110]を用いてテキストの意味内容をタグ付けするためのガイドラインを作成している。TEIはXMLを利用しているため、XML編集ツールやXML変換ツールなど、XMLのために開発された多くのツールを活用することができ、テキストの生成から分析に及ぶDHの研究を支えている。ただし、ガイドラインは大部なドキュメントであり、その考え方になじむには時間がかかることから、人文学者が誰でも使えるツールにはまだなっていない。

（2）International Image Interoperability Framework（IIIF）

「International Image Interoperability Framework」（IIIF）[111]は、画像データを共有するための標準である。技術的にはリンクトデータ（Linked Data）[112]技術を利用しており、書籍などを構成する画像やそれに付随するメタデータに統一資源識別子（URI）を付与し、それらを意味的にリンクする方法で全体を記述している。そしてLinked DataをJSON形式[113]で記述するためのJSON-LDフォーマット[114]を用いて、現代のウェブ技術として広く普及するJSONを活用した相互運用性を確保するなど、2010年以降のウェブ技術のトレンドを踏まえた仕様となっている。

標準の改良と普及に取り組むIIIFコンソーシアムには、大英図書館やフランス国立図書館などの大規模図書館、スタンフォード大学やオックスフォード大学などの著名大学、ゲティ財団やウェルカム財団などの有力財団が、我が国からも東京大学次世代人文学開発センターと京都大学図書館が正式メンバーとして加盟している[115]。

最初の標準仕様が提案されたのは2012年であり比較的新しいが、画像配信のニーズの高まりに伴って世界中で急速に普及しつつあり、既に3億5千枚以上の画像がIIIFのサイトを経由してアクセス可能となっている[116]。画像へのアクセスには、オープンソースソフトウェア

⑽⑼　Text Encoding Initiative Website〈http://www.tei-c.org/index.xml〉

⑽⑽　TEIにおけるマークアップはテキストの意味として、テキストの構造である段落や文章、文字などを明示できるほか、名前や日付、場所などの固有名詞や、マークアップの信頼性などを記述することも可能になっている。記述方法に関する長大なガイドラインは「P5: Guidelines for Electronic Text Encoding and Interchange」としてTEIウェブサイトに公開されている。

⑽⑾　International Image Interoperability Framework Website〈http://iiif.io/〉

⑽⑿　リンクトデータとは「結合したデータ」という意味であり、複数の機関から公開されるデータの項目の意味やIDを揃えておくことで、多くのデータを連結して活用するための相互運用性技術を指す。

⑽⒀　JSON形式とは、ウェブサイトでよく利用されるJavaScriptで読み書きしやすいデータ形式であるが、近年はJavaScriptに限定せずウェブサービス間などのデータ交換に広く使われている。以前は同じ目的にXMLが使われていたが、XMLよりも軽量で扱いやすいことから現在はJSONが主流である。

⑽⒁　JSON for Linking Data Website〈https://json-ld.org/〉

⑽⒂　IIIF Consortium Website〈http://iiif.io/community/consortium/〉

⑽⒃　Sheila Rabun, “Community Snapshot,” *IIIF Community News Letter,* Volume 2 Issue 1,13 Nov 2017.〈http://iiif.io/news/2017/11/13/newsletter/〉

を用いるが、アクセス方法を標準化し、異なるサイトが公開する画像を同じ方法で閲覧できるようになれば利便性が更に高まる。多くのミュージアムや大学がIIIFでの画像配信に切り替えるにつれて人文学者がIIIFを目にする機会が増えつつあり、画像配信及び画像を利用した研究に今後ますます利用されることが期待される。

6　欧米及び日本の動向

人文科学分野でもデータが大規模化するにつれて、データ活用のためのデータ基盤の重要性が高まりつつある。

(1) 欧州

EUは人文科学に関する様々なプラットフォームの構築に力を入れており、幾つかのプロジェクトが動いている。その中でも「Digital Research Infrastructure for Arts and Humanities」(DARIAH)[117]と「Common Language Resources and Technology Infrastructure」(CLARIN)[118]は中心的な存在である。

DARIAHは、人文科学研究に必要なデジタル研究基盤を構築するためのネットワークである。人文科学研究におけるデータ活用の問題は、多くの人文科学研究者がデータ及びデジタルツールの使い方や作り方を知らないという点にある。一方、情報学者も人文科学者のニーズを知らないため、両者が密接に協力しながらデジタルツールを構築する環境が必要となる。

DARIAHは人的ネットワークを構築し、デジタルツールを開発し、データを共有し、ノウハウを伝える場を提供する。例えば、ある研究グループが小規模な人文科学のデジタルツールを開発したとする。そのツールを限られた研究者が使う分には問題ないが、更に多くの研究者に提供したいと考えた場合、そのツールは様々な利用状況に対応していないかもしれないし、多人数が使用する環境では安定的に動作しないかもしれない。ツールの改良に必要な予算を国内で確保できる保証はなく、その永続性も保証できないという問題もある。そこでEUのDARIAHにおいて、大規模かつ長期的なサービスに耐えるようツールを改良して利用に供することで、国を越えた人文科学研究を支援しようとするものである。

一方CLARINは、デジタル化された言語資源（電子テキストを集めたコーパスやプログラム処理に適した辞書など）をあらゆる分野（特に人文科学と社会科学）の研究者、学生や市民にシングルサインオンアクセス（1回の認証手続で様々なサービスにアクセスできる仕組み）で提供する研究基盤である。実際のサービスは各国が運営するデータセンターなどが担当し、国ごとに異なる方法で長期的かつ安定的な研究資源の提供を目指すが[119]、CLARINはこうした欧州各国の取組の間で、情報共有や相互運用性の調整を行うための枠組みである。

その他のEUによる取組としては、「ヨーロピアナ」(Europeana)[120]が著名である。ヨーロピアナは、欧州各国の文化遺産保存機関と協力してデジタルデータへのアクセスを提供するポータルサイトである。文化遺産に関するメタデータを収集して形式を標準化することで、5300万件以上の文化遺産が検索できる。また、検索だけでなく「ウェブ展示」を通してテーマごとに厳選された文化遺産にアクセスできる。

---

(117)　DARIAH EU Website〈https://www.dariah.eu/〉
(118)　CLARIN ERIC Website〈https://www.clarin.eu/〉
(119)　オランダではDARIAHとCLARINという2つの基盤を統合したCLARIAH〈https://www.clariah.nl〉が構築されているが、こうした具体的な運営方法は各国に任されている。
(120)　Europeana Collections Website〈https://www.europeana.eu/portal/en〉

（2）米国

米国における事例として学術機関等が共同で運営する「ハーティトラスト」（HathiTrust）[121]を挙げる。これは780万冊の書籍（55億ページ）のデジタル化データを提供するプロジェクトである。米国グーグル社が開始した大規模書籍デジタル化プロジェクトである「Google ブックス」から提供されたデータがその大半を占めるが、非営利団体として書籍・映画・ソフトウェア・音楽・ウェブサイト等のアーカイブを構築する「インターネットアーカイブ」（Internet Archive）から提供されたデータ等も含まれている[122]。

ハーティトラストのデータを用いた各種サービスも立ち上がっている。「bookworm」はテキストデータから特定の単語を抽出し、その出現頻度を時系列で可視化するサービスである[123]。単語の出現頻度は文化を反映するため、人文科学における文化の調査に大きなインパクトを与え得る[124]。

（3）東アジアと日本

中国や台湾、韓国などでは国家的なプロジェクトで主要な古典テキストの電子化を進めており、その多くは既にオンラインでアクセス可能となっている[125]。これまで電子化は人海戦術で進められてきたが、中国語は個別の文字が独立しており日本語のような複数文字が連続するくずし字の問題がないことから、古典中国語のOCRの方が古典日本語のOCRよりも難易度が低く、OCRにより電子化を進めるプロジェクトも進んでいる[126]。このような状況の中で、日本は東アジアの中でもデータ公開が遅れており、文化資源のオープン化において存在感を示すことができていない。例えば、米国の大学図書館司書などからは、デジタル資料公開の遅れが米国における日本研究の退潮傾向につながっているとの危機感が表明されている[127]。

DHが進展する米国では、きちんとしたデータが豊富に使えるかどうかが研究対象を選ぶ際の1つの基準となることから、データが充実していないと研究対象として選ばれにくい傾向が生じている。ゆえに、東アジアのどこかの国に関する研究を行う場合、データ公開が遅れている我が国よりも、データ公開が進む中国、台湾、韓国の方が研究対象に選ばれる可能性が高くなる[128]。このことは日本研究の分野で活躍する研究者の減少を招き、世界における我が国の存在感を低下させることになる。こうした事態を防ぐには、国内だけではなく国外の研究者にも活用されるよう、人文科学やその周辺分野におけるデータ基盤の強化が求められる。

そうした中、我が国でも大規模なプロジェクトとして、国文学研究資料館が中心となって進める「日本語の歴史的典籍の国際共同研究ネットワーク構築計画」が平成26（2014）年に始まっ

⑿⑴　HathiTrust Digital Library Website〈https://www.hathitrust.org/〉

⑿⑵　“Ingest Checklist.” HathiTrust Website〈https://www.hathitrust.org/ingest_checklist〉では、Google Content、Internet Archive-digitized Content、Non-Google Contentの3種に分類した記述がある。

⑿⑶　bookworm: HathiTrust Website〈https://bookworm.htrc.illinois.edu/develop/〉

⑿⑷　日本語に関しては、OCRによる文字認識の精度が低い課題に加え、テキストを単語に分割する処理にも課題があり、まだ「bookworm」のようなサービスは存在しない。

⑿⑸　例えば、台湾の中央研究院は6億7千万字以上の中国古典を電子テキストとして提供している。中央研究院「漢籍電子文献資料庫」〈http://hanji.sinica.edu.tw/〉

⑿⑹　Chinese Text Project〈https://ctext.org/〉

⑿⑺　例えば、江上敏哲『本棚の中のニッポン―海外の日本図書館と日本研究―』笠間書院, 2012, p.117には「Nippon Invisible」として、米国から見た我が国のデジタルデータ公開遅れに対する危機感が述べられている。

⑿⑻　世界における我が国の経済的地位の低下による関心の減少という構造的要因も無視できないが、ここでの議論は「インターネットで検索できないものは存在しないのと同じ」という感覚を反映したものである。

た[129]。このプロジェクトは、我が国の古典籍30万点をデジタル化し、画像データを誰でもアクセスできる形で公開するだけでなく、古典籍データを活用した様々な共同研究を立ち上げることで、国際的な研究ネットワークを形成するという目標を掲げている。日本文化に関してこれだけの大規模なデータが公開されることは画期的であり、国内外の日本研究者からの期待も大きい。

さらに、くずし字のOCRを最新の人工知能技術を使って開発するコンテスト[130]がスタートするなど、古典籍の画像データ及びそこから派生した字形データセット等の大規模データの公開は、情報学分野にも新たな研究への刺激を与えている。大規模データに人工知能技術を適用することで人文科学に新しい可能性が広がり、国内外で過去から現在に至る日本文化への理解が深まるといった、社会的価値が生まれることが期待される。

⑿⑼　前掲注⑿⑸
⑿⑽　「くずし字チャレンジ!」人文学オープンデータ共同利用センターウェブサイト〈http://codh.rois.ac.jp/old-char-challenge/〉