数理解析研究所講究録
第2021巻 2017年 114-119

ビッグデータ時代の統計学[1]

情報・システム研究機構　北川 源四郎
Genshiro Kitagawa
Research Organization of Information and Systems

統計学はこれまで科学的研究の方法論としても，また社会における意思決定の基盤としても重要な役割を果たしてきた．しかしながら，近年の情報通信技術や計測技術の急激な進展にともなって，学術分野や社会において大量・大規模なデータが集積し，ビッグデータ時代が到来し，その有効活用が研究や社会の飛躍的発展の鍵となっている．科学の文法を標榜する統計学やその教育の今後の在り方を考えるにあたっては，ビッグデータが統計学や科学技術に及ぼす影響は避けては通れない．本稿では，この問題に焦点を絞って検討することにする．

1．　ビッグデータと第4の科学：データ中心科学

2012年3月，オバマ大統領がビッグデータ研究開発イニシアティブ[3]を発表し，一般社会においてもビッグデータが一躍脚光を浴びることとなった．従来の科学研究では，目的のために厳密に設計され取得されたデータに基づき解析や検証がおこなわれてきたが，現在ではあらゆる研究過程や人間活動を記録しデジタル化して得られた多種多様なデータを利用して，従来は考えられなかった科学的発見，予測・知識獲得あるいは価値創造が実現できるようになりつつある．　ただし，ビッグデータには，大きな価値が潜在していることが期待されるとはいえ，その多くは構造化されていない上に，価値密度は低く不均一であり，さらに逆説的であるが大規模データの多くはスパースである．ここに，データの大量さに止まらないビッグデータ解析の困難さと統計学の新しい役割が見出される．

ビッグデータ解析は，高エネルギー物理学や天文学のような最先端の実験・観測科学においては今や不可欠なものとなっているが，むしろ生命科学，地球環境科学や人間・社会科学のように第一原理モデルが適用できない領域や，多階層や超多数の要素からなる複雑なシステムを対象とする領域において，それ以上に活用の場が広がりつつある．特に，人間・社会においては，個人化サービスやデータ駆動型産業の創出，1次産業・2次産業の効率化，テーラーメイド医療・保健の実現，社会インフラのスマート化，データに基づく意

[1] 本稿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の提言「ビッグデータ時代に対応する人材育成」[1]およびその作成過程の議論に基づき作成したものである．

思決定・政策決定，稀少事象の発見とリスクの検知，災害時オンライン対応など，様々な形でイノベーションを起こしつつある．

その一方で，ビッグデータの登場は科学的研究方法自体の変革を迫っている．過去 50 年間，計算機のメモリや計算速度は 5 年間で約 10 倍増加するというムーアの法則と呼ばれる経験則にそって技術革新が進められてきた．しかし，次世代シーケンサーの登場によってゲノム解読速度が 5 年間で約 1 万倍に増加したように，世界に蓄積するデータ量はムーアの法則をはるかに超える速度で急激に増大し続けている．したがって，ビッグデータの処理のためには，ストリーム計算など，巨大なデータに対応するためのデータ処理技術の革新は不可欠ではあるが，それだけでは対応することは不可能で，ビッグデータ時代にふさわしいデータ駆動型の研究方法論とそのための研究基盤の確立が必要である．

20 世紀の科学研究は，実験科学と理論科学のふたつの方法論に支えられてきたが，前世紀後半にはシミュレーションを中心とする計算科学が確立し，気象予測，流体設計，ゲノム創薬などの分野で様々な成果が得られようとしている．そして今や大規模データの登場により，第 4 の科学ともいうべきデータ中心科学（データ科学とも呼ばれる）の確立を目指すべき時に来ている[4]．理論科学と理論科学がそれぞれ研究者の才覚に依拠した帰納的方法と演繹的方法なのに対して，計算科学とデータ中心科学は計算機(Cyber)が拓いた新しい演繹的（モデル駆動型）方法と帰納的（データ駆動型）方法と位置づけることができる（図１）．２１世紀の科学はこれらの４つの方法論をバランスよく駆使することによって発展していくことができると考えられる．

図１　第４の科学（データ中心科学）の位置づけ

２．　データ駆動型の研究パラダイムと課題

我が国ではデータ駆動型の科学的方法論の嚆矢として，「データによって現象を理解する」という統計数理の立場が戦後の早い時期から確立していたが，その後，「データの科学」および「統計的モデリング」の二つの流れが形成され，1996 年に東京で開催された IFCS（国際分類学会）を経て，データ科学（Data Science）は国際的な流れに繋がっていく．我が国では，ビッグデータに関連する研究プロジェクトも比較的早くから開始され，1998 年以降，特定領域研究の「発見科学」，「アクティブマイニング」や「情報爆発」など一部は欧米に先行して開始された．また，JST でも 2008 年以降，さきがけ，CREST のプログラムが

いくつか実施され，現在に至っている．

一方，欧州では 1966 年には P. Naur により datalogy が提案されている．また，米国ではプリンストン大学の J. Tukey(1977)によって解析初期の段階を重視した「探索的データ解析」が提唱され，これが後に ATT による S 言語およびその後の R 言語の開発に繋がっていった．その後，欧州では 1999 年に e-サイエンスが提唱され，研究の計画，実験，データ収集，解析，出版，成果の普及までの研究の全過程を一体的に進めることによって先端科学研究が推進されてきた．また米国では，NSF の数理科学では 2004 年から巨大データの問題が重要課題となり，情報学関連では CDI(Cyber-enabled Discovery and Innovation)，CPS(Cyber- Physical Systems)の研究プログラムが実施されている．2012 年にはビッグデータ研究開発イニシアティブ[3]により国家プロジェクトとしてのビッグデータ研究開発がスタートして現在に至っている．

産業界においては近年，特にビッグデータに関連する人材育成に関して急速に関心が高まっており IBM Almaden 研究所のシンポジウム（2008 年），McKinsey Global Institute のレポート（2011 年[6]），Harvard Business Review(2012 年)で取り上げられ，データサイエンティストや統計研究者の重要性が指摘されている．また，産業界の求めるデータサイエンティストを育成するために，2012 年からはインサイト・プログラム(Insight Data Science Fellow Program [5])が開始されている．これはシリコンバレーの主要な IT，SNS 企業 30 社以上が協力して実施しているもので，ポスドク，院生を対象とする 6 週間の短期人材養成によってトップタレントを養成することを目的としている．

データサイエンティストや統計専門職の育成は，近隣のアジア諸国でも積極的に行われている．中国では 150 以上の統計学科が整備され，年間 2 万人以上の広義の統計学修了生が育成されている．韓国でも 50 以上の統計学科・応用統計学科が設置されている．

これに対して，人材育成に関して我が国は，ようやく 2013 年度から文部科学省の次世代 IT 基盤構築のための研究開発事業の一環としてデータサイエンティスト育成ネットワークの形成が開始されたところである．このように，ビッグデータの研究は，我が国ではむしろ海外に先行して開始されたが，統計教育やデータ中心科学の確立に向けた組織的取組およびその推進に必要なデータサイエンティストの育成おいては後塵を拝しているのが現実である．特に統計学科等を数多く設置している欧米諸国あるいは極東諸国と異なって，日本ではこれまで専門の統計学科を設置せずに各応用分野での具体的課題に取り組ませる中で統計科学の専門家を育成する分野点在方式をとってきたが，異分野への転向，新分野開拓，分野間知識移転のためには，今後はむしろ集中化し抽象度を上げた専門的教育が必要と考えられる．

3． ビッグデータ活用に必要な要素技術と人材育成

MGI レポート[6]にも示されているように，ビッグデータ活用のために必要な主要な要素技術はデータ解析法，データ可視化，ビッグデータ処理技術である．データ解析法はビッグデータからの深い知識獲得のための方法であり，統計数理，機械学習，情報検索，自然言語処理，最適化などの方法が主要な役割を果たす．特にビッグデータ活用においては，明確なモデルが先験的に存在しない分野における知識獲得や意思決定・政策決定が今後ますます重要になることを考えると，統計的モデリングやベイズ推論を最も重要な方法と位置づけるべきである．データ可視化は，次元圧縮，特徴抽出，パターン認識など，膨大な高次元データそのものや解析結果を人間が的確に把握できるようにするための技術である．ビッグデータ処理技術は，分散処理，並列処理，ストリーミング計算など現在でもペタバイト級の散在する多様なデータを処理するために必要な情報処理技術である．したがって，今後の統計科学の人材育成にあたっては，機械学習，自然言語処理，最適化，情報処理技術などの統計科学の境界領域の分野を積極的に取り込み，従来の数理統計学よりもスパンを広げた教育を行うことが必要である．

近年，国内の大学や研究機関で，研究不正にかかわる重大な事件が発生している．特に，ビッグデータの活用やデータ駆動型の研究にあたっては，研究倫理の確立が不可欠であり，これ無くしてはかえって国民の信頼を失うことになりかねない．したがって，ビッグデータ活用を目指す人材育成にあたっては，データ取得やデータの取り扱いにおける研究倫理を徹底することが必要である．

さらに，ビッグデータ活用に携わる研究者の要件としては，ビッグデータ活用に必要な3要素技術の習得，研究倫理の確立は当然として，現実の課題を解決するためには，問題の本質の把握，定式化，データ取得，分析，知識獲得，課題解決の全過程に関与できる全人的能力が必要である．このように，今後の統計科学研究者はビッグデータ解析のための要素技術とともに，領域分野の知識と経験，問題発掘能力，コミュニケーション能力も必要なことから，方法論と領域研究を熟知したＴ型，Π型人材としての育成が不可欠となる．

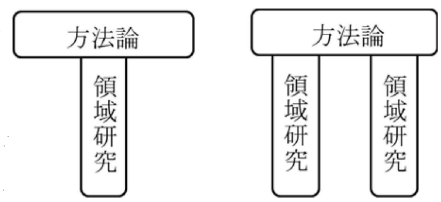

図2　T型の人材とΠ型の人材

異分野交流のために欠かせないコミュニケーション能力の育成方法に関しては，長年多くの努力がなされてきたが．未だ成功と言える方法は確立されていない．ただし，現時点では地道ではあるが，既にいくつかの試みは開始されている．統計数理研究所の統計思考院では、外部から持ち込まれた共同研究の課題に対し、豊富な知識と経験を持つシニアの

特命教授が、博士号を取得したばかりの領域を専門とする若手ポスドクにメンターとしてアドバイスし、いっしょに課題解決に臨んでいる。東北大学原子分子材料科学高等研究機構では、材料科学と数学の架け橋を担当するインターフェースユニットを設け、異分野はもちろん、実験家と理論家の間の交流促進に機能している。

このように，ビッグデータを活用できデータ中心科学の担い手ともなる新しいタイプの統計学の研究者（データサイエンティスト）を育成するためには，統計数理，数理科学，機械学習，情報処理などの横断型の方法論を主専攻とし，領域分野を副専攻とする教育組織・プログラムの編成が必要になる．また逆に，領域科学の博士取得者にビッグデータ処理・解析技術を取得させる方法が有効と考えられる．

4． データサイエンティスト育成の効果

第 4 の科学の担い手となるデータサイエンティストは，過度に細分化し融合研究が困難な現在の科学技術研究の局面打開の切り札となることが期待される．また，抽象度の高い方法論を取得し，領域研究者とコミュニケーションができる知識と能力を備えたデータサイエンティストは研究ネットワークのハブとして分野間の知識移転や新分野開拓の担い手となることが期待される．さらに，数理科学研究者のもつ汎化能力は当該研究者の異分野や産業界への転向をも容易にすることから，産業界からの要請やポスドク問題解決へ向けての貢献も期待できる．

このように，データサイエンティストは分野横断型の研究が要求されるビッグデータ時代の科学技術研究の推進に不可欠なだけでなく，科学技術創造立国を目指す我が国の発展の鍵でもある．データ中心科学の担うべきデータサイエンティストの育成にあたっては，統計学がその中心となるにしても，その果たすべき役割は従来の統計学の枠に収まるものでないことは明らかである．今後は，統計学自体の革新を目指すとともに，データサイエンティスト育成を目標として，その育成のための具体的方法を更に検討していく必要があると考えられる．

参考文献

[1] 日本学術会議提言「ビッグデータ時代に対応する人材育成」，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，2014 年 9 月 11 日
http://www.scj.go.jp/ja/info/kohyo/pdf/kohyo-22-t198-2.pdf

[2] 日本学術会議提言「ビッグデータ時代における統計科学教育・研究の推進について」，日本学術会議数理科学委員会数理統計学分科会，2014 年 8 月 8 日

http://www.scj.go.jp/ja/info/kohyo/pdf/kohyo-22-t197-1.pdf

[3] Obama Big Data Research and Development Initiative.
http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release.pdf

[4] T. Hay, S. Tansley and K. Tolle, eds., The Fourth Paradigm: Data-Intensive Scientific Discovery, Microsoft Research (2009)
http://research.microsoft.com/en-us/collaboration/fourthparadigm/

[5] Insight Data Science Fellows Program. http://insightdatascience.com/

[6] J. Manyika, M. Chui, J. Bughin, B. Brown, R. Dobbs, C. Roxburgh and A.H.Byers, Big Data The next frontier for innovation, competition, and productivity, McKinsey Global Institute, (2011)
http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation