ビッグデータ時代の統計学

情報・システム研究機構
北川 源四郎

統計学は科学的研究の方法論として，また社会における意思決定の基盤としての役割を果たしてきた．しかしながら，情報通信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ている．統計学やその教育の在り方を考えるにあたっては，ビッグデータが統計学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を発表し，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たデータが解析されていたが，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雑多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や予測・知識獲得が実現できるようになりつつある． ビッグデータは，大きな価値を内包しているが，多くは構造化されていない上に，その価値密度は低く，不均一・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ある．

ビッグデータ解析は今や最先端の実験・観測科学において不可欠となっているが，生命科学や地球環境科学のように第一原理が適用できない領域や多階層や超多数の要素の複雑なシステムを対象とする領域では，それ以上に重要になっている．特に，人間社会では，データ駆動型サービス・産業の創出，社会インフラのスマート化，データに基づく意思決定・政策決定，希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の分野でイノベーションを起こしつつある．

ビッグデータはムーアの法則をはるかに超える速度で増大しており，ストリーム計算など，データ工学の革新は不可欠である．しかし，それ以上に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の確立が必要である．20 世紀の科学研究は経験科学と理論科学の方法論に支えられてきたが，前世紀後半には計算科学が確立し，いまや第 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．経験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方法と帰納的方法と位置づけることができる．

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．欧州では 1966 年には P. Naur により datalogy が提案されている．また，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「発見科学」,「アクティブマイニング」,「情報爆発」など一部は欧米に先行して開始された．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ている．一方,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．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,CPS 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いる.

産業界においては，特にビッグデータに関連する人材育成に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)， Harvard Business Review(2012 年)で取り上げられ，データサイエンティスト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 年からインサイト・プログラムが開始されている．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 週間の夏季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 万人以上の広義のデータサイエンティストが育成されている.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.

一方，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，日本は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，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

ビッグデータ解析の３大要素技術はビッグデータ工学，データ可視化，データ解析法である．ビッグデータ工学は，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り，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．データ解析法はビッグデータからの Deep Knowledge 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．

データサイエンティストの要件としては，ビッグデータ解析に必要な 3 大要素技術の習得は当然であるが，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，データサイエンティスト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．これを実現するためには，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も考えられる．

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りうる．さらに，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．データ中心科学の確立のために必要なデータサイエンティストを育成し，社会に定着させるための具体的な方策を考えていく必要がある．