資料５－別添５

（提案５）

（案）

提言

ビッグデータ時代に対応する人材の育成

平成２６年（２０１４年）〇月〇日

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

この提言は、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審議結果を取りまとめ公表するものである。

委 員 長	北川源四郎	（第三部会員）	大学共同利用機関法人情報・システム研究機構・機構長
副委員長	安達　淳	（連携会員）	大学共同利用機関法人情報・システム研究機構 国立情報学研究所教授
幹　　事	樋口　知之	（連携会員）	大学共同利用機関法人情報･システム研究機構 統計数理研究所　教授（所長兼務）
幹　　事	鷲尾　隆	（連携会員）	大阪大学産業科学研究所教授
	相田美砂子	（連携会員）	広島大学大学院理学研究科教授
	市川　晴久	（連携会員）	電気通信大学人間コミュニケーション学科教授
	今井　桂子	（連携会員）	中央大学理工学部教授
	加藤　直樹	（連携会員）	京都大学大学院工学研究科教授
	狩野　裕	（連携会員）	大阪大学大学院基礎工学研究科教授
	喜多　泰代	（連携会員）	独立行政法人産業技術総合研究所主任研究員
	酒井　英昭	（連携会員）	京都大学大学院情報学研究科教授
	下條　真司	（連携会員）	大阪大学サイバーメディアセンター教授
	高野　明彦	（連携会員）	大学共同利用機関法人情報・システム研究機構 国立情報学研究所教授
	高安　秀樹	（連携会員）	株式会社ソニーコンピュータサイエンス研究所 シニアリサーチャー
	椿　　広計	（連携会員）	大学共同利用機関法人情報･システム研究機構 統計数理研究所教授
	徳山　豪	（連携会員）	東北大学大学院情報科学研究科教授
	中島　秀之	（連携会員）	公立はこだて未来大学学長
	西田　豊明	（連携会員）	京都大学大学院情報学研究科教授
	松本　裕治	（連携会員）	奈良先端科学技術大学院大学教授
	宮野　悟	（連携会員）	東京大学医科学研究所教授
	山西　健司	（連携会員）	東京大学大学院情報理工学系研究科教授

本提言の作成に当たっては、以下の方にご協力いただいた。

坂田　一郎	東京大学大学院工学系研究科教授
矢田　勝俊	関西大学商学部教授

本提言の作成に当たっては、以下の職員が事務を担当した。

事務局	盛田　謙二	参事官(審議第二担当)
	齋田　豊	参事官(審議第二担当)付参事官補佐
	沖山　清観	参事官(審議第二担当)付審議専門職（平成 26 年 6 月 30 日まで）
	加藤　美峰	参事官(審議第二担当)付審議専門職付（平成 26 年 5 月 1 日から）

要　　旨

１　作成の背景

情報通信技術の飛躍的進歩によって、多くの学術研究分野や社会において時々刻々ビッグデータが蓄積しつつある。ビッグデータには膨大な知識や潜在的価値が埋蔵されているため、その有効活用が今後の学術や産業発展の鍵となっており、激しい国際競争が始まっている。

当分科会ではビッグデータ活用のために今や喫緊の課題となっている、データ中心科学の確立と学術分野への普及・定着のための方策の検討を行ってきたが、その中で新しい科学のための新しい人材、すなわちデータサイエンティストの育成が特に重要であることを認識した。データサイエンティストの育成は、既に海外では急速に進められており、我が国においても直ちに着手しなければ、学術研究や産業界におけるビッグデータ活用において大きく立ち遅れる恐れがある。

そこで、当分科会ではビッグデータ時代に対応した人材育成の在り方を中心に検討を行い、本提言を作成した。

２　現状及び問題点

米国ではビッグデータ研究開発イニシアティブによって戦略的な研究投資を開始しているが、EU でも第 7 次研究枠組み計画（FP7）の後継プログラムの中で、ビッグデータ関連のプログラムに対し 2020 年までに巨額の研究投資を決定している。このような政策決定を受けて、米国及び英国では 2013 年夏以降の半年の間に、ビッグデータあるいはデータサイエンスの研究所が相次いで設立されている。

その一方で、ビッグデータの活用に携わる研究者の不足は深刻である。McKinsey Global Institute のレポートによれば 2020 年には、データサイエンティストが 14 万-19 万人不足すると推定されている。このような状況を背景に、欧米やアジア諸国では統計学などのビッグデータに関連する教育組織や学位授与数が急速に増加している。

このように諸外国がデータサイエンティストを急速に増加させている中で、我が国だけが逆に減少していることは、今後の我が国の科学技術研究発展及び産業におけるイノベーションにむけて重大な問題であり、早急な対応が必要である。

３　提言等の内容

提言１　データ中心科学を専門とする教育組織の設置

高等教育においては、データ中心科学を専門とする学科、専攻あるいは教育プログラムの設置が必要である。この教育組織においては、ビッグデータ解析の３要素技術を中心とする分野横断型の学問を主専攻、（複数の）領域科学を副専攻とするシステムを採用し、T 型・Π型人材の育成を目指すべきである。

現時点では、データサイエンスを系統的に学習できる組織や場が極めて少ないので、当面の対応策として、副専攻や副専攻プログラムの開設や、普及が進みつつある大学のオンライン教材や民間の講習会等を利用してまずは裾野レベルから大量にデータサイエンティストを育成することが適当と考えられる。

提言 2　基幹的研究組織内に恒久的なデータ解析部門を設置する

領域の特性を十分に考慮しながら、効果的にデータ中心科学を推進するため、ライフサイエンス、医学・疫学、天文学、高エネルギー物理学、地球科学、環境科学、材料科学、安全性など、データ解析が重要な役割を果たす研究領域の基幹研究所内に部署横断的なデータ解析専門の恒久的な研究部門を設置すべきである。

また、ビッグデータの利活用を効果的に進めるには、学術組織、研究開発機関、企業のマネジメント層を中心とするデータリテラシーを向上させ、データサイエンティストを利用する側の意識改革を行うことも欠かせない。

提言 3　日本版インサイト・プログラムの早急な設置

アカデミアの人材と産業界が要求する人材の乖離を埋めるために、日本版インサイト・プログラムを早急に設置し、既卒の博士研究員の再教育を実施し、データサイエンティストを育成すべきである。データサイエンティスト育成においては，現実の課題への挑戦と異分野交流の経験が不可欠であることから，このプログラムでは、領域科学の縦型研究者あるいは情報学、統計科学、数理科学等の横断型科学の研究者にビッグデータ解析法の習得とビッグデータプロジェクトの体験をさせ、T 型、Π型人材を育成する。

このプログラムで育成する人材は、ポスドクに大きな付加価値を付け進路転換のチャンスを与えることから、我が国の発展の担い手となるばかりでなく、現在、我が国で深刻な問題となっているいわゆる“ポスドク就職難民”問題の解決に貢献することが期待できる。

提言 4　データサイエンティストの資格の制定

データサイエンティストは、ビッグデータを扱う専門職として以上に、ビッグデータ時代の科学技術研究及び産業界のイノベーションを先導するトップタレントとして今後の我が国発展の鍵ともなる重要な役割を果たすことになる。従って、データサイエンティスト人材の質保証の観点からデータサイエンティスト資格の制定が望ましい。但し、資格のあり方や、その付与の仕組みなどついては、データサイエンティストの業務内容が多岐にわたり、求められるレベルもさまざまであることから今後慎重に検討すべきである。

目　　次

１　はじめに

2012 年３月、オバマ米国大統領はビッグデータ研究開発イニシアティブを発表し、巨大で複雑なデジタルデータから知識と洞察を得るための能力向上のために、２億ドル規模の継続的研究投資によりビッグデータの収集、管理の最新技術構築を行うことを明らかにした[48]。さらに、このイニシアティブによって、過去の連邦政府の投資の結果スーパーコンピュータとインターネットが飛躍的に発展したと同様に、科学的発見、環境・生命医学研究、教育及び国家安全保障におけるビッグデータ活用が一変するであろうと指摘している。情報社会の進展に伴って生じた大きな可能性と挑戦すべき課題は 21 世紀当初から徐々に顕在化していたが、これを契機にアカデミアでも産業界でもビッグデータが一躍脚光を浴び、ビッグデータ元年ともいえる状況が到来した。

(1)　ビッグデータの定義、特性、発生源

ビッグデータとは、「市販されているデータベース管理ツールや従来のデータ処理アプリケーションで処理することが困難なほど巨大で複雑なデータ集合の集積物を表す用語」とされているが、その本質はデータ量そのものの大きさよりも、むしろ、あらゆる情報を取り入れようとする大規模性にある。その結果、ビッグデータは、様々な形式、構造、計測頻度、精度、非定常性などを伴った多様で不均質なものとなる。

従来の科学研究では、明確な目的のために厳密に設計されたデータを取得し解析することが基本であったが、インターネットやデータベース等の飛躍的発展にともなって、他の目的で得られたデータや特定の目的を持たずに収集された多数の雑多なデータをも統合・利用し、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。

科学研究では、測定・観測機器、ネットワーク及びスパコンの発展により、ゲノミクス、気象学、地球環境、天文学、高エネルギー物理学、シミュレーションなどの領域で日々大量かつ大規模な観測データと計算データを蓄積している。

一方、人間・社会においても、情報通信技術の発展により経済活動から日々の生活に至るまであらゆる人間の活動が精細に捕捉されデジタル化されるようになりつつある[6]。実際、センサー技術の向上と低価格化により、インターネット（Web、ソーシャルメディア)、センサー（モバイル、車載等のプローブ、防災、スマートグリッド)、リモートセンシング、POS、 RFID タグ、ビデオサーベイランスなどによって、科学研究においてと同様に、日々大量で大規模なデータが蓄積されている。

このように、あらゆる研究活動過程や人間の活動を精細に記録しデジタル化した結果がビッグデータである。ビッグデータは大きな価値創造の可能性を内包しているが、その多くは構造化されていない上に、その価値密度は低く、形式、観測頻度、精度、非定常性など様々な意味で不均質であり、さらに逆説的であるが、１サンプル（事例）あたり、数多くの項目が同時計測される超多変量データのサンプル集団が成すデータ空間は、

ほとんど常にスパース（疎）である。

古典的な情報処理においては、明確な目的のために収集・管理された入力データを計算アルゴリズムによって処理し、ユーザの必要な情報を出力する事が主体であった。しかしながら、ゲノム科学やWebマイニングなどに代表されるように、自然現象や生命活動あるいは社会が日常的に発信する情報をセンサリング、あるいはモニタリングにより、記録した膨大なデータの活用結果が社会や経済、あるいは市民生活へ強い影響を持つ時代となった。そこでは情報処理の主体は、これらのデータの深い理解と適切なモデリングにより、情報あるいは知識を取り出すことに移りつつあり、データ取得技術、データ活用技術、機械学習、統計的モデリング等、新しい情報処理を中核としたデータ中心科学が非常に重要になってきている。

人間は、日常生活のセンシングによって得られた膨大なデータを、無意識に活用して学習し、ある種のパターン認識によって動作や判断を行う。それを、計算機上で実現する人工知能の開発は従来困難であったが、そのような知能情報処理が、データ中心科学によって可能となりつつある。

例えば、ロボットの動作制御においては、大量のシミュレーションデータを生成・学習し、柔軟かつ自在な制御を行うことが可能になり、また、過去の専門棋士の棋譜データという非常に優れたデータを活用して学習することにより、コンピュータ将棋の現在の発展がある。

このように、データに関する科学をとりまく環境は、その概念も含めて過去20年間に大幅に変化し、データを中心に研究を行うことが、多くの分野で重要になっている。

(2)　ビッグデータが拓く世界

ビッグデータ解析は、最先端の巨大実験・観測装置を使う高エネルギー物理学や天文学などに

おいて不可欠なものとなっている。しかしながら、生命科学や地球環境科学などのように第一原理が全体的に適用困難な領域、あるいは多階層性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では、データに基づきモデルを構築する帰納法的研究アプローチが不可欠である。また、これらの領域では、シミュレーションにおいても支配方程式に立脚した物理モデルだけによる結果では限界があり、実世界に関する知見を得るためには、シミュレーションモデルに基づく解析結果とデータからの情報を統合するデータ同化が欠かせない[22]。情報検索、可視化、機械学習、構造発見、離散系最適化技術、統計的モデリングなどの、ビッグデータから深い知識を獲得するためのデータ解析基盤は、今後の科学・技術進歩の鍵となる最も重要な研究インフラのひとつである。

一方、ビッグデータは社会システムのイノベーションにも大きなインパクトをもたらしつつある。下記の①－⑧が示唆するように、ビッグデータの活用の仕方と対象領域の組み合わせによって、多くの活用例があるが、主要な活用の仕方としては、集団から個へのサービスの転換、オフライン計算からオンライン計算への転換、データ駆動型産業

の実現、スマート化、稀少事象の発見などがある。

①　個人化サービス・データ駆動型産業の創出

20 世紀の産業は、科学技術の発展により得られた知識を応用し、大量生産・大量消費による効率化を実現してきた。しかし、ビッグデータの登場によって、個々人の特性やニーズに合わせた医療、教育、情報提供など、テーラーメード型のサービス提供を可能にしつつある。これは効率最大化から満足度最大化への転換ともいえる。マーケティング、サプライチェーン効率化、リスク管理などにおいては、集団に代わってデータに基づいて個への対応に基づく意思決定が行われるようになりつつある。

また、グーグルのビジネス戦略に代表されるように、地球規模で大規模かつ網羅的に収集・蓄積したデータを活用する新しいサービスが創出され、短期間にネットビジネスがグローバルな産業に成長している。このようにビッグデータは今後の産業のあり方や人間の生活自体を激変させつつある。

ビッグデータの利用による高い産業創出効果に関する認識の下に、政府レベルでも平成 24 年 7 月に首相官邸が主導する政策会議:高度情報通信ネットワーク社会推進戦略本部（IT 戦略本部）において、電子行政オープンデータ戦略が策定された[9]。そこでは、「公共データは国民共有の財産であるという認識の下、公共データの活用を促進するための取組に速やかに着手し、それを広く展開することにより、国民生活の向上、企業活動の活性化等を図り、我が国の社会経済全体の発展に寄与する」ことの重要性が強調されている。

このような動きの中で、現在、総務省は、「公共データの活用促進、すなわち「オープンデータ」の推進により、行政の透明性・信頼性の向上、国民参加・官民協働の推進、経済の活性化・行政の効率化が三位一体で進むことが期待されている」として、情報通信(ICT 政策)の一環で「オープンデータ戦略の推進」を掲げている[11]。また、国土交通省観光庁も、「観光地域づくりを通じた地域の活性化を図るためには、来訪者が地域に何を求めているのかを把握した上で、より来訪者のニーズに合致した取組を実施していくことが重要」との認識に立ち、「GPS 機能により許諾を得て蓄積される「位置情報」を活用することにより、観光地における来訪者の行動・動態について調査・分析し、その結果を地域の取組に反映していくことを可能とする手法を構築」を検討するワーキンググループを立ち上げている[4]。

②　１次産業・２次産業の効率化

人類の歴史とともに長い歴史を持つ１次産業や前世紀の日本の経済的な発展を支えた２次産業においても、ビッグデータは産業の形を大きく変えつつある。

多数のセンサーデータを活用した農業は、天候等に左右される影響を軽減する安定した食糧生産を可能とする[2]。最先端の半導体工場では、数千台の製造装置から得られる膨大な数の変数のデータを分析することによって製品の質を高め、不良の発生を

抑制している。この結果、世界最高水準の製品評価及び莫大な利益が得られる。数十万もの部品から構成される自動車の製造においてもサプライチェーンのビッグデータを管理することでオンデマンド型の無駄のない製造を可能としている。新しい材料物質の開発においても、材料データベースや、知識（経験）ベース、及び膨大な実験結果などを集積したビッグデータを利用して、効率的な実験デザインを探る動きも急速である。このように、ビッグデータは、１次産業・２次産業においても品質と効率性を高める競争力の要として認識されている。

③　医療・保健におけるビッグデータ活用

ゲノム配列を高速に読むシーケンス技術の発展は、あらゆる分野の中でこの 10 年間で最もすすんだ技術と言っても過言でない。たった一人のゲノム配列を読む日米欧などの国際協力プロジェクトは 1990 年から 13 年を要したが、今や個人のゲノム配列を調べる費用は数 10 万円程度に下がっており、血液検査の感覚で、１万円ほどの経費で全ゲノムを調べる時代が、もう目の前である。

ゲノム配列の特異的変化がガンや難病の主原因となっていることは広く一般にも認知されてきたが、昨年のハリウッド女優の予防的乳房切除のニュースは、ゲノム検査と医療行為が直結した段階にまで科学技術が進んでいることに、多くの人々を驚かせた。ゲノム配列や環境因子と生活習慣病の関係を網羅的に調査する大規模プロジェクトも各国で進行中であり、個人の健康・医療にかかわるあらゆる情報をクラウド上で統合化する技術的基盤は、ほぼ整いつつある。

これにより、テーラーメード投薬や治療などによる個々の QOL（生活の質）の向上はもちろん、国家財政的観点からも効果的かつ効率的な医療費支援策の立案が可能となる。またデータ駆動型臨床試験では、実験計画、データ取得、分析、報告書作成、結果の公開・伝達の一連の流れからなる統合的な臨床試験によって、安全かつ有効な治療薬の開発をより短期的に実現しようとしている。

④　社会インフラのスマート化

交通システム、電力供給システム、ビル管理などにおいては、大量に配置したセンサーからのビッグデータの活用により、社会インフラのスマート化が行われようとしている[47]。

より積極的な試みとしては、人間・社会活動を模倣する計算モデルに基づく社会システムの知能化・スマート化がある。都市内の人流、交通流をセンサリングし、モデル化し、それを使って少し先の状態をシミュレートすることによって交通システムのリアルタイム効率化（バスやタクシーのデマンド運行、カーナビの経路指示最適化、電車遅延時の効率的復旧など）が可能になる。

⑤ データに基づく意思決定・政策決定

公共投資や観光政策、環境対策などにおいては、データに基づく科学的政策決定を行うことが模索されている。従来は、専門家を集めた委員会などで、専門家が過去に収集した事例にもとづいて政策を議論・決定する、いわゆるファクトベースの政策決定が中心であったが、ビッグデータに基づいて、実際のデータにより一層裏付けられた根拠に基づく政策の議論・決定を行う「アクチャルベース・エビデンスベース」のアプローチが急速に普及してきている。

さらに、新しい調査報道の手法としてデータ駆動型ジャーナリズムも提唱されている。オープンデータをオープンソースツールで分析することにより、データの発見、フィルタ、可視化、出版、配布、評価の一貫したプロセスを構築し、今まで見つけられなかった事象や規則性を見出し、消費者、ビジネスマン、政策当局、政治家等の意思決定・政策決定に役立てることを目指している[40]。

昨年行われたプロ棋士とコンピュータ将棋の対戦結果は、経験と勘に基づく専門技能対データ解析の象徴的出来事として、産業革命期に行われた鉄道馬車と蒸気機関車の競争の歴史的再現ともいえる。

米国の人気クイズ番組 Jeopardy!の歴代チャンピオン２人を破った IBM のコンピュータシステム Watson は、辞書、Wikipedia、ニュース記事など約２億ページの大規模言語データから有用な情報を抽出し、数千台のコンピュータによる並列処理によって、ファクト型の質問応答が人間のパーフォーマンスを凌駕しうることを示した[35]。Watson はその後、医療分野に応用され、腫瘍学に関する医学専門誌や臨床試験のビッグデータからの学習により、エビデンスベースの治療法を医師に提供できるようになっている[38]。東大入試に機械学習に基づく人工頭脳システムが挑戦中であるのも同様である。

[29]には、ワインのヴィンテージ予測、判決予測、取引業者評価、保険料の層別、人事採用、スポーツ選手スカウティングなど多くの事例で、単純な回帰分析ですら専門家の判断よりもよい結果を残しているという驚くべき事実が示されているが、ビッグデータの解析はこの科学的意思決定の可能性をさらに高めるものと期待できる。

⑥ 稀少事象の発見とリスクの検知

ビッグデータ解析の意義は、対象の平均的特性や顕在化した関係を精密に推定することよりも、サンプリングでは見逃してしまう稀な事象や隠れた関係性を発見することにある。これによって、故障や災害の事前予測、列車等の運行安全の確保、金融リスク管理などの実現が期待される。

また、マーケティングでロングテールと言われるように、イベント生起の頻度的には稀少だが、優良顧客のようなリターンとしては大きな価値・利益につながる事象の発見は大きなイノベーションの源泉ともなり得る。

機械学習、特徴抽出、可視化、統計学、探索的データ解析の方法を、通信ログやユ

ビキタスセンサー、監視カメラ画像データなどのビッグデータに適用することにより、様々なリスクを検知し、情報セキュリティの確保や不正発見を行うことが考えられている。既に部分的には、事故や災害の早期の検出やテロや犯罪の捜査にも活用され、大都市が東京オリンピックのような国際的大規模イベントの招致を実現する上では、安全で快適な都市空間の必要条件になりつつある。

高度成長期に集中的に建設されたトンネル、橋梁、ダムなどの大型社会インフラの老朽化も問題となっており、センサーシステムの設置と得られるビッグデータの解析による、経済効率的に国民の命を守るインフラ管理策の考案も急がねばならない。

さらには、防災センサーネットワークデータ、GPS データ等の統合により、崖崩れ、地盤変化、地殻変動、火山噴火等の自然災害リスクの早期検出も試みられている。保健面においても、パンデミックの恐れがある感染症流行の早期把握や流行防止対策など、ソーシャルメディアの積極的利用や医療現場のクラウドへの一体化により、ビッグデータは国家的リスクの回避に新しい方法を提供しつつある。

⑦ 災害時対応

既に東日本大震災時に一部利用されたように、モバイル情報やカーナビ情報などの位置情報を、個人情報保護の観点も考慮しつつ緊急時には避難、救助、ロジスティックスなどの支援に活用することが考えられている。

実際 2011 年の東日本大震災の際には、災害地に向かった個々の車のカーナビがシステムセンターに発信した「通行実績情報」を平均速度情報とともに地図上に集積し、「通れた道マップ」として Web 上で日々更新公開することで、後続のボランティアに有用な情報を提供することが行われ，それ以降、災害時情報のひとつの形態として定着している[18]。また、災害直後のみではなくその後の中長期に亘る企業復興支援や避難者支援などにも、災害前後の企業取引情報、避難者携帯位置情報などのビッグデータ利用の有効性が議論されている[44]。

災害時のような突発的な事象においては、あらかじめシナリオ（解法モデル）を十分に用意できない問題に対して、時々刻々と得られる空間的に不均質なデータをリアルタイムに処理していくことが重要となる。東日本大震災東京電力福島第一原子力発電所事故の例では、リアルタイムに入手される災害現場の様々な情報と、問題解決の可能性を持つ技術・専門家に関する膨大な蓄積情報（バックグラウンド情報）をうまく結び付け、より良い対策を果断する必要性が浮き彫りとなった。

⑧ 人文科学におけるデータ活用

計量文献学や歴史典籍のデータベース化のように、文学、絵画、言語や文化遺産などの人間の営み全てと人間に関する現象について、データに基づく科学的な方法によって新しい発見や新しい解釈を生み出されるようになっている。遺跡のレーザー３Ｄ計測、ミイラの CT スキャン、ゲノム解析による家系同定、絵画の X 線撮影、画材物質

の先端計測による年代同定や真贋判定、重要文化財の３Ｄプリンタによる高精度レプリカを用いた実証実験など、先端計測技術で得られたデータの解析が人文科学に驚くべき新世界を開きつつある。

現在のところ、芸術や文化の研究におけるデータは、一般に自然科学領域におけるビッグデータほどは大量ではない。しかし、芸術や文化の研究においてはデータに基づき、これまで行われなかったデータ駆動型の科学的方法が適用されるようになっている。また、今後は上記のようなデータが、解像度、詳細度の急激な向上と多くの事例に関する横断的蓄積によって、ビッグデータ化していく可能性は大きく、データ中心科学の方法論適用の必要性が増大すると見込まれる。

(3)　ビッグデータ活用に必要な課題

過去半世紀にわたって、計算速度及び記憶容量などの情報処理の技術革新は、ムーアの法則と呼ばれる経験則に沿って５年で約 10 倍の速度で進められてきた。

しかしながら、近年のビッグデータは、次世代シーケンサーの登場によってゲノム解読の速度が５年間で１万倍程度増加したように、ムーアの法則をはるかに超える速度で増大している。また、同じく情報通信ネットワークの一層の発展により、例えばカーナビがネット接続を通じて飛躍的に膨大な情報を収集・利用することが可能になりつつある。

これは、単に情報処理機器性能の量的発展のみによってはビッグデータには対応できないことを明らかに示している。従って、従来の情報処理能力を超える、一般には価値密度の低いビッグデータの処理のためには、データ取得現場での１次処理を行うストリーム計算、並べ替えや探索を桁違いのスピードで実現するデータベース技術や計算アルゴリズムなど、データ処理技術の革新が不可欠である。

しかし、ビッグデータの活用に必要な技術革新はそれだけに止まらない。大規模・不均質性を特徴とするビッグデータからの知識獲得のためには、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とそのための人材育成が必要である。

これまでの科学研究は長い歴史を持つ経験科学と理論科学の方法論に支えられてきたが、20 世紀後半にはシミュレーションを主とした計算科学が確立し、複雑な非線形系や多粒子系の挙動の理解や予測技術が飛躍的に進展し、気象予測や車体の形状設計やエンジン設計などで大きな成果を挙げた。そして現在、ビッグデータの到来により、データの持つ情報を最大限活用して研究を推進するデータ中心科学の重要性が明らかになっている。実験科学と理論科学がそれぞれ、研究者の個人的才覚に依拠した帰納的方法と演繹的方法と考えられるのに対して、計算科学は計算機が拓いた新しい演繹的方法、データ中心科学は情報通信技術と大規模データが可能にする新しい帰納的方法と位置づけることができる。

個人情報の問題も、ビッグデータの利用において避けては通れない課題である。ビッグデータは健康・医療（テーラーメード医療・創薬、臨床試験）、マーケティング、教

育、観光、情報提供などにおける革新的個人化サービスの実現などのように、社会における大きなイノベーションの可能性を秘めているが、その実現のためには個人情報の利用に関する社会的合意と法制度の整備が不可欠である。 その一方で、個人情報秘匿化技術、暗号化したままの情報検索・モデリング技術、統計処理など個人情報の保護と有効利用を両立させるための技術開発や、ユーザの個人情報開示とユーザの得る利得をバランスさせ、より積極的な個人情報活用を促進させる個人情報保護活用基盤の構築なども重要となる。

2　海外の動向

(1)　統計学、データサイエンス、e-サイエンス

データ駆動型の科学的方法の動きは、大規模データに力点を置かない形では、1966年に「自然とデータ利用の科学」を目指す datalogy として欧州で提案され[43]、1969年には同名の学科がコペンハーゲン大学に設置されている。また、1977 年にプリンストン大学の J.Tukey によって提唱された「探索的データ解析」は、モデルありきを前提とするのではなく、データの示唆する情報を多面的に捉えるという、解析初期の段階を重視したものであった[51]。

データから自動的に有用な知識を発見する技術は、1970 年代から機械学習として米国を中心に研究されてきたが、1990 年代後半からは、データの大量性や多様性に注目しデータマイニング技術として発達してきた。広義のデータマイニング技術は機械学習に加え、データの収集、管理、運用などのデータ処理の一連の流れを含む総合技術とされている[3]。

データ科学（Data Science）の名前が明示的に使われたのは、1992 年に開催された第２回日仏データ解析セミナーのサブタイトルが最初と思われる[19]。その後、米国の J.Wu は 1998 年に“Statistics = Data Science ?”という講演を行って、統計学はデータ科学に転換すべきことを提言している[54]。2002 年には国際科学会議の科学技術データ委員会（CODATA）は Data Science Journal を発刊した。なお、CODATA に関しては日本学術会議情報学委員会の下に対応する分科会が設置されており、またこの雑誌は JST の J-Stage によって運営されている。

欧州では 1999 年英国の科学技術局長官 J.Taylor が e-サイエンスを提唱し、計算機技術によって、研究計画、実験、データ収集・分析、成果の普及、研究全過程の長期保存と活用を一体的に進めることによって、素粒子物理学、地球科学、生命科学、社会シミュレーションなどの分野における先端科学研究を推進した[50]。

また、e-サイエンスによる科学的発見を支援するために、WLCG（Worldwide LHC Computing Grid)、European Grid Infrastructure や天文学、素粒子物理学、計算生物学、ゲノミクス、分子動力学、材料科学、計算機科学、ナノ技術などを対象とする Open Science Grid のようなグリッド計算基盤、クラウド基盤が開発されている。

英国は2000年から2009年にかけて2.6億ポンド以上を投入し、Grid計算の応用、Grid Middleware開発及び各地のe-Science Centerの支援を行った[42]。

一方、米国では、大規模データに関連する研究開発プログラムは、10年前頃から開始されている。NSFの優先領域のひとつの数理科学では2004年から「mathematical and statistical challenges posed by large data sets」が重要課題として取り上げられ[45]、情報学関連ではCDI（Cyber-enabled Discovery and Innovation、2007-2011年度）[46]、CPS（Cyber-Physical Systems、2009年度-)[47]の研究開発プログラムが実施されている。また、Materials Genome Initiative（2011年11月）においては計算ツール、実験ツール、数値データを材料イノベーション基盤とし、実験データやノウハウを蓄積したビッグデータを利活用するプロジェクトに大きな予算を投下している[52]。

(2) ビッグデータ元年

このように、米国では個別分野においてビッグデータに関連する取組がなされてきたが、2011年に科学技術に関する大統領諮問委員会（PCAST）が、連邦政府はビッグデータ技術への投資が少ないと結論づけたことに対応し、科学技術政策局（OSTP）が2012年３月29日にビッグデータイニシアティブを発表した。このイニシアティブの下で６機関（NSF、NIH、DOD、DARPA、DOE、USGS）が総額２億ドルを投資し、データへのアクセス、体系化、知見を集める技術を改善、強化するとしている[48]。これを契機にアカデミアでも産業界でもビッグデータが一躍脚光を浴び、ビッグデータ元年ともいえる状況が到来した。

欧州、アジアにおいても、ビッグデータに対する研究投資を実施しており、既に激しい国際競争が始まっている。欧州では第７次研究枠組計画（通称FP7)、BIG（Big Data Public Private Forum）の事業を予算総額約30億ユーロで2012年９月（26ヶ月プログラム）から開始した[30]。さらにFP7の後継プログラムであるHorizon 2020においては、2014-2020年にビッグデータ関連事業に９億ユーロ以上を投入するとしている[34]。

このような動きをうけて、米国及び英国の大学では昨年秋以降、多数のBig Data あるいはData Scienceの研究所が設立されている[7]。また、中国では情報資源を共有するためのセンターを設置し、収集したデータの相互の関係付けのためにメタデータの付与や自動分類等の技術開発を行っている。さらに、韓国ではビッグデータを含む研究データの共有とデータ中心科学を推進するNational Scientific Data Centerを2013年から構築することとなっている[49]。

(3) ビジネス界との連携

産業界でも、ビッグデータに関して急速に関心が高まり、様々な形でビッグデータの活用によるイノベーションを実現するための人材育成の必要性が強調されている。

IBM Almaden研究所において2008年５月に開催された恒例のシンポジウムInnovating with InformationでGoogle チーフエコノミストのH.Varianは「今後10年間で最も魅

力的な職業は statistician である」という講演を行っている[53]。また、McKinsey Global Institute のレポート[41] (2011 年６月､以下 MGI レポートと呼ぶ)では､医療、小売業、公共サービス、製造業、個人位置情報を用いたサービスの 5 分野を例として、統計学と機械学習の知識とスキルを持った「Deep Analytical Talent」が今後、14-19 万人必要になると予測している。Harvard Business Review(2012 年 10 月号)には「データサイエンティスト:21 世紀のもっとも魅力的な職業」という記事が掲載されている[33]。

アメリカ労働統計局の職業展望便覧によると修士レベルの Statistician は 2012 年では 27,600 人、2022 年には 27%増加し 34,900 人になると予測されている[31]。これらの報告や記事では、実際には Statistician、Deep Analytical Talent あるいは Data Scientist などと呼ばれているが、大略趣旨は同じと考えられることから本稿では以下データサイエンティストと呼ぶことにする。

データ中心科学の方法[37]やデータサイエンティスト重視を裏付ける動きとしては、IBM が 2009 年に SPSS を買収し、SAP は Business Objects(2008)と Sybase(2010)を買収している。買収された企業はいずれも、統計処理あるいはデータ解析、データに基づく意思決定支援を専門とする企業である。

また､米国ではアカデミアの人材と産業界の求める人材の乖離を解消す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 (Insight Data Science Fellows Program) が開始されている[39]。これはシリコンバレーの主要な IT、SNS 企業 30 社以上が協力しているもので､ポスドク､博士課程大学院生を対象とする６週間の夏季短期人材養成 (2014 年からは夏期と冬季の２回) 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。

通常のインターンシップと異なり、企業の研究者がプログラム研修所に出向いて、プロジェクトベースの自学自習をアドバイスし､必要な情報処理技術やビジネスセンスを習得させようとするものである。プロジェクトの内容は各参加者が考え、ビッグデータは公開されているものだけを使う。

興味深いことは、このプログラムの主要な人材ソースとして、計算機科学の学位取得者よりはむしろ物理学のようなハードサイエンティストを想定していることである。ここから伺われることは、数理的知識、計算スキル、データから知識を獲得する訓練を既に受けた学位取得者に情報処理技術を習得させるのがデータサイエンティスト育成の早道と判断していることである。この点は、今後の情報学の人材育成においても考慮すべき重要な点を含んでいる。2013 年までこのプログラムの修了者は 100%協力企業に高給で採用されている[39]。このプログラムで採用された逆インターンシップの方法は、研修当初から特定企業の特殊性に漬からないことから､汎用性を特徴とするデータサイエンティストの育成には有効と考えられる。また、プロジェクトの内容は異なっていたとしても､データを取り扱う手法に関する知識習得のために同じ分野の人間は同じ場所を共有させる点も示唆に富む[22]。

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以上の大学に統計学科が整備されており、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。韓国でも統計学科・応用統計学科など統計関連の学科が 50 以上設置され[49]、データサイエンティスト関連の人材育成が進んでいる。

３　我が国の現状と人材育成に関する課題

(1)　プロジェクト

我が国では統計科学の分野においては、伝統的な数理統計学の方法に対して、問題の本質を把握し、実験や調査を計画してデータを獲得し、モデリングを経て、対象の理解、予測、意思決定を行う一連のプロセスを重視する「統計数理」の立場が戦後の早い時期から確立していた［20]。このような背景の下で、1992 年の日仏データ解析セミナーにおいて「データサイエンス」が初めて明示的に用いられ[36]、1996 年に神戸で開催された IFCS（国際分類学会）を経て、日本発の用語は国際的に広まることになった。

また、このような動きとは別に、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た。文部科学省の特定領域研究では、大規模情報からの知識獲得の方法に関する「発見科学」が 1998-2001 年度、情報洪水時代に向けた「アクティブマイニング」が 2001-2004 年度、情報爆発時代に向けた新しい IT 基盤技術に関する「情報爆発」が 2005-2009 年度に、また、あらゆる局面で必要な情報を解析できる情報基盤を実現しようとする経済産業省プロジェクト「情報大航海」が 2007-2009 年度に実施されている。

JST では文部科学省の戦略目標の下で、さきがけ「知の創生と情報社会」が 2008-2013 年度に実施され、CREST・さきがけ複合領域「ビッグデータ統合利活用のための次世代基盤技術の創出・体系化」及び CREST「科学的発見・社会的課題解決に向けた各分野のビッグデータ利活用推進のための次世代アプリケーション技術の創出・高度化」が 2013 年度から開始されている。また、文部科学省委託事業「数学・数理科学と諸科学・産業との協働によるイノベーション創出のための研究促進プログラム」もビッグデータへの挑戦を想定したものとなっている。日本学術会議においても情報学委員会では、第 21 期及び第 22 期において、国際サイエンスデータ分科会、E-サイエンス分科会、大量実データの利活用基盤分科会、E-サイエンス・データ中心科学分科会を設置し、ビッグデータに関連した研究方法論や基盤構築の在り方に関する検討を行ってきた。

日本学術会議が取りまとめたマスタープラン 2014 においては、「アカデミック・ビッグデータ活用研究拠点の形成」と「複雑データからのディープナレッジ発見計画」が採択されている[17]。前者は、ビッグデータの活用のためのデータ中心科学を確立するために、データ基盤整備、モデリング・解析基盤整備、人材育成の三位一体の事業を推進するもので重点大型研究として採択されている。一方、後者は大量データが内包する

複雑さや発掘すべき知識の質の問題の重要性を指摘し、複雑な関係同士や不均一・非一様なデータ同士の性質の隠ぺいによるデータ解析の質低下の克服やデータの背後の潜在的な関係性、階層性、因果性、ダイナミクス、変化、予兆等の深い知識の導出が標榜されている。

総合科学技術会議による講評「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」では、ビッグデータ関連施策は「ICT 分野の動きの速さを考慮し、諸外国の状況も見据えつつ、状況に応じて目標の前倒しも視野に入れ推進すべきである。」という評価を得ている[10]。

また、我が国における企業によるデータ解析企業買収の動きとしては、NTT データが統計・データマイニング、数理計画、科学技術計算、知識工学を基盤技術とする数理システムを買収している。

(2) 人材育成プログラム

人材育成に関しては、2008 年度から文部科学省の産学連携による人材育成事業「プロセスイノベーター育成プログラムの開発」において高度な統計推論、データマイニングに関する知識と社会科学の素養に基づき、ビジネスプロセスを科学的かつ実践的に解明できる人材育成が行われてきたが[27]、2013 年度からは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[24]。近年の産業界を中心とした人材育成の動きとしては、データサイエンティスト協会やデータサイエンスコンソーシアムの設立などがある。また、日本統計学会は RSS（王立統計協会）と連携して統計検定を開始している。

但し、少し古い統計ではあるが、MGI レポートの表 36[41]によれば、2008 年の Deep Analytical Talent（統計学、機械学習、データマイニング、最適化、OR、データ解析等に相当）を専門とする人の数は、アメリカ 24730 人、中国 17410 人、イギリス 8340 人に対して、日本は 3400 人と著しく少ない。さらに 2004 年－2008 年の間の変化でも、海外ではアメリカ 3.9%、中国 10.4%、 ロシア 12.8%、イギリス 2.5%、イタリア 18.9%と増加しているのに対して、唯一日本だけが -5.3%と減少している。

このように、欧米諸国や中国等がデータサイエンティスト関連の人材を増加させつつある中で、我が国だけが実際には減少していることは今後の我が国の科学技術研究発展及び産業におけるイノベーションにおいて重大な障害となりかねない。特にその中でも我が国の統計学の特異性が顕著である[22]。統計学部あるいは統計学科、生物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。既に統計学の大学院教育の在り方に関しては 1983 年に日本学術会議 勧告『統計学の大学院研究教育体制の改善について』[16]が出されているが、これまでに実現したのは、1988 年に設置された総合研究大学院大学

複合科学研究科統計科学専攻の１か所にすぎない。

４　ビッグデータ活用に必要な要素技術と人材育成

J.Wu は IFCS の動きを受け、1998 年の講演の中で、統計学は、今後、データ収集、モデリング、データ解析、問題解決、意思決定の一貫したプロセスとして取り組む、データサイエンスとして発展すべきである、と主張した[54]。また、2001 年に W.S.Cleveland はデータサイエンスを統計学と先端計算技術が融合した独立した学問領域として確立すべきであると指摘している[32]。本節では、データ中心科学の確立のために何が必要か、またビッグデータの活用を推進するために必要な人材の育成について検討する。

(1)　データ中心科学の要素技術

MGI レポートに示されているように、ビッグデータ活用の３大要素技術はビッグデータ処理技術、データ可視化、データ解析法である[41]。データ中心科学を確立し、ビッグデータからの価値創造を実現するためには、これらの要素技術の革新が不可欠である。

ビッグデータ処理技術は、ペタバイト級の散在するデータを処理するために必要な分散処理・格納、並列処理、HPC、ストリーミング計算（オンライン処理、圧縮センシング、サーベイランス・センサー、先進的フィルタリング技術)、巨大データベース、リンケージ技術（情報統合、例：医療保険データと自然環境データ、高解像度白黒画像と低解像度カラーデータ)、クラウド計算、信号処理などの技術である。

データ可視化は膨大な高次元データや計算結果を人間が把握できるようにするための技術であり、次元圧縮、特徴抽出や画像処理などを含む。

データ解析法はビッグデータからの深い知識（Deep Knowledge）獲得のために不可欠な方法であり、関連研究分野としては統計学、機械学習、データマイニング、統計的モデリング、ベイズ推論、テキスト検索、情報検索、Web 情報解析、自然言語処理、画像認識・理解、パターン認識、データ解析、情報抽出、最適化などの方法がある。特に、ビッグデータ解析のためには、スパースモデリング、データ同化法、インピュテーション技術（内挿・外挿、不完全データ・異常値の処理)、時空間センシング、変化解析、新 NP 問題の解決、高次元空間の構造探索とモデル化、異種情報統合による個人化技術（製品・医療サービスなどの個人化技術、テーラーメード化)、社会情報ネットワークにおける知識発見、隠れた関係の検出、特異性の発見、因果推論の実現などの課題がある。

このようなビッグデータのための解析要素技術の実現のためには、統計学と画像処理のように従来の要素技術間の連携・融合を積極的に行っていく必要がある。

(2)　データサイエンティストの要件

MGI レポートではデータ中心科学の要素技術を駆使して、諸科学分野での発見や社会

における知識創造・意思決定あるいは産業イノベーションを担うデータサイエンティスト（Deep Analytical Talent）が必要であり、今後 14-19 万人の需要があることを指摘している[41]。 このようなデータサイエンティストの要件としては、ビッグデータ活用に必要な３つの要素技術（ビッグデータ処理技術、データ可視化、データ解析法）に習熟していることが必要である[22]。しかしながら、データ中心科学の実践においては、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することになる。従って、データサイエンティストの育成においては、３つの要素技術と当該領域の問題に習熟させるだけでなく以下のような能力も同時に伸ばす教育が必要である。

- 戦略立案能力、問題発掘・企画能力、問題解決能力
- データ収集能力
- データの裏にある真実を見抜き、関連するデータを見出す力
- キュレーション能力（データの選択、前処理、クレンジング）
- データ分析結果の業務や事業への実装能力
- 異分野研究者・事業者との連携能力

本稿では、これらの能力を合わせてデータリテラシーと呼ぶが、このようなデータリテラシーを備えた研究者がデータサイエンティストである。但し、すべての項目を備えることは相当困難であり、そのようなデータサイエンティストは正確に言えばスーパーデータサイエンティストと呼ぶべきであろう。一般には、自分の弱点部分を補強してくれるデータサイエンティストと協業し、チーム全体として大きな力を発揮することが求められる。このようにデータリテラシーは多面的であり、多様なプロジェクトすべてに個人や単独のチームで対応できるとは限らない。単独のチームで対応できない場合には、必要とされるスキルセットを正しく特定し、新しくデータサイエンティストを採用するなり、アウトソーシングすることが欠かせない。その状況を考慮すると、データサイエンティストとしても自分の持つスキルセットを明確にできるほうが、人材のミスマッチングを防ぐ意味で大きな効果がある。従って、スキルセットを認証する仕組み、つまり資格認定の実施が急がれる。

上記の能力の中でも，最後の項目（連携能力）が相対的に重要であり，その資質の大きな構成要素はコミュニケーション能力である。一方、スーパーデータサイエンティストは、ビッグデータ時代においてはデータ中心科学の専門職の域を超えたトップタレントとして、各分野においてリーダーとして活躍することが期待できる。今後のグローバル大企業においては最高経営層に必ず一人は求められる人材像といっても過言ではないであろう。

⑶　データサイエンティストの育成方法

データサイエンティストは、データ中心科学の方法を駆使して、諸科学分野や社会の課題を解決することを要請されている。従って、データサイエンティストの育成にあた

っては、ビッグデータ解析のための要素技術をマスターするとともに、領域分野の知識と経験も必要なことから、方法論（横型の知識と経験）及び領域（縦型の知識と経験）を熟知したＴ型、Π型人材の育成が不可欠となる。これを実現するためには、情報処理、機械学習、統計数理などの横断型の方法論を主専攻とし、（複数の）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、インサイト・プログラムのように領域科学の博士号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有効と考えられる[22]。

異分野領域をつなぐ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ついては長年、多くの努力が成されてきたが、いまだ成功と言える方法が確立していない。そもそもスケール化する方策を求めること自体が適切かどうか疑問であり、地道で泥臭いやり方が着実のようである。例えば、日本学術振興会が行なっている二国間先端科学シンポジウムのように、全く専門分野の異なる若手研究者が、特定のテーマについて泊まり込みで集中的討議することは有効である[15]。情報・システム研究機構が行なっている若手クロストーク事業は、異分野の研究者が数人集まって１チームを構成し、架空の共同研究テーマについて構想を練り、最後にチーム相互に批評しあう合宿型集会である[5]。統計数理研究所の統計思考院では、外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[12]。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[14]。これらの例は、いずれも現段階では地道ではあるが、育成の規模をスケール化できる要素的アイデアが含まれている。

⑷　データサイエンティスト育成の効果と活用の体制

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における困難の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マスターし、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、さらに研究コーディネーションができる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して活躍できるばかりでなく、分野横断型の融合研究をリードする人材となることが期待される[22][23]。

このような人材を積極的に活用できる組織体制を同時に整備することも極めて重要である。比較的容易に実現できるものとしては、データサイエンティストを抱えるプロジェクトを多数同時に実施した経験のある研究機関内に、恒久的なデータ解析部門を設置することが考えられる。これまでは、時限のプロジェクト期間内にポスドクレベルのデータサイエンティストを非常勤職員として雇用することで、人材育成の観点ではやや場当たり的に対応してきた。データ解析専門部署に属する常勤の研究者が各プロジェクトにエフォート管理のもと参加する体制にすれば、データサイエンティストは常時、複

数のプロジェクトに参加することになり、分野間の知識移転が自然に実現される。また、データサイエンティスト側からの新分野開拓につながる提案もしやすくなる。

もちろんこのようなプロフェッショナル集団を抱え込む部署をつくることで新たに抱える不安材料もある。インハウス部署は時間の経過と共に、新しいプロジェクトの提案や機関全体目標の変化に対して必ず保守的になり、結果として組織内で孤立した抵抗組織となる傾向があるのもその一例である。そのリスクを軽減するため、プロジェクト内でデータ解析の部分を切り出し、その部分をデータ中心科学の専門機関にアウトソーシングする方策もあり得る。特に、ビッグデータの利活用にかかわる国家レベルの大きなプロジェクトにおいては、インハウスで対処するよりもアウトソーシングするほうが効果的であると思われる。但し、日本においては、データ中心科学の専門機関と言える機関が情報・システム研究機構及び傘下の研究所を除いて無い状況に近いことは、その実施において十分認識しておかねばならない。

データサイエンティストの育成においては、データ中心科学の要素技術を一定レベル習得したものに現場の具体的な問題に触れさせ、オンザジョブトレーニングを実施するのが効果的であることは、諸外国の育成先行実績例をみても明らかである。この現場主義自体は、文部科学省が実施している「数学協働プログラム」[13]や、トップタレントの大学院生の教育プログラムであるリーディング大学院でも同様の認識である。一方、既に現場において問題を特定しているが、データ中心科学の要素技術の未習得のものに、それらを習得してもらうことで問題解決につなげる道筋もあり得る。言わば、逆インターンシッププログラムの実施である。この考え自体は特に新しいものではなく、大学等での客員教員として民間の方々を招聘する形態はその狙いを一部踏襲していると言える。但し従来の客員制度は、受け入れ研究室や講座を超えるような展開性をもった柔軟な制度とは言えず、広範なデータ中心科学を習得するためには、データ中心科学を専門とする機関に逆インターンシップ制度を設けることが合目的であろう。

科学技術創造立国を目指す我が国は、これまで多くの国費を投入し博士号取得者の量産に取り組んできたが[26]、就職の受け皿となる大学や公的研究機関のポストを増やさなかったために、現在では毎年 6,000 人以上の博士号取得者が正規の職につけていない（文部科学省『平成 25 年度学校基本調査（確定値）』[25]の 11 ページの表 6，博士課程修了者の就職率 65.8%、理学系の正規の職員への就職率は 38.3%、図 12 および図13.）。このいわゆる“ポスドク就職難民”問題[21]の解決の鍵は、産業界の要求する人材を育成し、企業への就職増をいかに実現するかにあるが、汎化能力に富むデータサイエンティストは当該研究者の異分野や産業界への進出をも容易にすることから、日本版インサイト・プログラムはポスドクに付加価値を付け進路転換のチャンスを与え、産業界のイノベーションの担い手となるとともに、ポスドク就職難民問題の解決に貢献することが期待できる[22]。

５　提言

ビッグデータ活用に不可欠な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今後の科学技術研究の推進においても、また産業界のイノベーションにおいてもなくてはならぬ存在であり、特に科学技術創造立国を目指す我が国においては今後の発展の鍵となる[26]。しかしながら、これまでも横断型の科学技術の担い手の重要性は認識されても、そのキャリアパスの形成が困難という問題が解決できず、横断型技術発展の障害となってきた。以下ではこの点も考慮して、前節で示した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提言を行う。提言１は文部科学省と大学関係者、提言２は関係各省と研究機関および企業トップ、提言３は文部科学省と産業界，提言４は各省庁と関連学会に向けたものである。

提言１　データ中心科学を専門とする教育組織の設置

高等教育においては、データ中心科学を専門とする学科、専攻あるいは教育プログラムの設置が必要である。この教育組織においては、ビッグデータ解析の３要素技術を中心とする分野横断型の学問を主専攻、（複数の）領域科学を副専攻とするシステムを採用し、T型・Π型人材の育成を目指すべきである。この教育組織では、領域科学を主専攻とする学生に対しても、データ中心科学を副専攻として採用し教育する、比重を逆転させた教育を行うことも効果的と考えられる。

現時点では、データサイエンスを系統的に学習できる組織や場が極めて少ないので、当面の対応策として、副専攻プログラムの開設や、普及が進みつつある大学のオンライン教材や民間の講習会等を利用してまずは裾野レベルから大量にデータサイエンティストを育成することが適当と考えられる。また、データ活用で先行している企業群に学生を派遣するインターンシップの推進も有効と考えられる。国はそれらの策の実現にむけて積極的に支援すべきである。

提言２　基幹的研究組織内に恒久的なデータ解析部門を設置する

領域の特性を十分に考慮しながら、効果的にデータ中心科学を推進するために、ライフサイエンス、医学・疫学、天文学、高エネルギー物理学、地球科学、環境科学、材料科学、安全性など、データ解析が重要な役割を果たす研究領域の基幹研究所内に部署横断的なデータ解析専門の恒久的な研究部門を設置する。

但し、分野点在型だけでは応用分野のデータの特性に固着した方法論の研究開発に偏りかねない。データ中心科学の深い理解を促す系統的教育を行う教育研究機関があれば、異分野への転向や新分野の開拓に積極的に取り組める人材を量的にも多く育成できる。ビッグデータの利活用にかかわる国家レベルの大きなプロジェクトは、データを取得する基幹研究所が、このデータ中心科学の専門機関と機関レベルで連携して担うのが適当である。産業界の例で言えば、機関内に部門を設置する形態は、社内に事業部門横断的な専門組織

を設けるものであり、一方、データ中心科学の専門機関は、ビッグデータアナリティクス業務を請け負うコンサルティング会社のようなものである。

また、ビッグデータの利活用を効果的にすすめるには、過去に自動車産業において品質管理の重要性を徹底させるために、企業トップまでを含めて教育活動を徹底させた企業が大きな成果を挙げたように、ビッグデータの活用のためには、学術組織、研究開発機関、企業（IT、品質管理、アナリティクス、金融・証券）のマネジメント層を中心とするデータリテラシーを向上させ、データサイエンティストを利用する側の意識改革を行うことも欠かせない。

提言 3　日本版インサイト・プログラムの早急な設置

アカデミアの人材と産業界が要求する人材の乖離を埋めるために、日本版インサイト・プログラムを早急に設置し、既卒の博士研究員の再教育を実施し、データサイエンティストを育成すべきである。データサイエンティスト育成においては，現実の課題への挑戦と異分野交流の経験が不可欠であることから，このプログラムでは、物理科学、生命科学等の領域科学の縦型研究者あるいは情報学、統計科学、数理科学等の横断型科学の研究者に半年から１年間の長期間、ビッグデータ解析法の習得とビッグデータプロジェクトの体験をさせ、データサイエンティスト（T 型、Π型人材）を育成する。これによってアカデミアの人材と産業界の要求の乖離を埋め、ビッグデータ活用に関するアカデミア及び産業界からの要請に応えることができるようになる。

いうまでもなく、このプログラムは産業界との密接な連携のもとで実施する必要がある。但し、この事業は汎用性を目指した人材育成になるので、通常のインターンシップよりはオープンな体制が必要であり、またアメリカのインサイト・プログラムと異なり、ポスドク・インターンシップ制度などの国主導のプログラムで実施するのが効果的かつ現実的と考えられる。

日本版インサイト・プログラムは、アカデミアの人材と産業界の要求する人材との溝を埋め、ポスドクに大きな付加価値を付け進路転換のチャンスを与えることから、我が国の発展の担い手となるばかりでなく、現在、我が国で深刻な問題となっているいわゆる“ポスドク就職難民”問題の解決に貢献することが期待できる。

提言 4　データサイエンティストの資格の制定

データサイエンティストは、ビッグデータを扱う専門職として以上に、ビッグデータ時代の科学技術研究及び産業界のイノベーションを先導するトップタレントとして今後の我が国発展の鍵ともなる重要な役割を果たすことになる。従って、データサイエンティスト人材の質保証の観点からデータサイエンティスト資格の制定が望ましい。

民間においては、昨年の夏以降、同様の趣旨で、スキル標準などを業界主導で構築する動きが急である。その背景には、データサイエンティストに関する明確な定義がないため、人材に期待される役割とスキルセットのミスマッチにより、業務依頼者と担当者の双方に

とって不満足かつ残念な状況が頻発していることがある。

データサイエンティストのスキル・知識を定義し、評価制度を整えることは、データサイエンティストの業務遂行能力の正当な判定を可能にし、その結果、ビッグデータ関連市場の健全な発展が期待できる。また、データサイエンティスト本人にとっても、資格制度の存在は自分の能力を高める大きな動機付けとなる。

既に臨床試験においては、生物統計家や試験統計家が必ず責任もって参画することが国際的なルールとなっているため、そのことが統計学を学ぶ学生にとって大きな学習意欲になっている。データ解析が重要な役割を果たすそれ以外の分野の国家認定においても、臨床試験と同様にデータサイエンティストの関与を義務づけることも考慮すべきである。但し、資格のあり方や、その付与の仕組みなどついては、データサイエンティストの業務内容が多岐にわたり、求められるレベルもさまざまであることから今後慎重に検討すべきである。

＜用語の説明＞

ビッグデータ

近年の ICT（情報通信技術)、特にセンサーの飛躍的発展によって、地球物理、気象、地震、天文、生命科学、マーケティング、ファイナンスなど多くの研究分野や社会で出現した大量・大規模のデータ。ある事象に関する非常に多数の多種多様な要因データが得られるようになったことが本質で、これによって大きな可能性が開けた反面、従来の方法では解決できない新しい課題を生み出している。

データ中心科学

ICT の急速な発展に伴って利用可能となった大規模・大量データ（ビッグデータ）を活用した、研究や技術・サービス開発のための科学的方法論。ICT をフルに活用した、帰納的（データ駆動型）な方法と位置づけられる。

第４の科学、第４のパラダイム

従来の科学研究を駆動した実験科学、理論科学に対し、20 世紀後半にシミュレーションを中心とした計算科学が確立したが、21 世紀に入って ICT 技術の飛躍的進歩やビッグデータの出現によって確立しようとしている新しい科学研究の方法論（データ中心科学)。第３の科学（第３のパラダイム）とも呼ばれる計算科学に続く第４の科学（第４のパラダイム）ともいわれる。

ｅ-サイエンス

計算機技術やそれに基づくインフラの上で、一連の探求から成果発表までを行う科学的方法論の総称である。そこでは、探求を行うために必要な種々の準備や実験、データ収集、成果の普及、探求の過程で生成されるあらゆる資料データの長期保管やアクセスが、計算機技術を用いて実現される。より具体的な形態としては、計算機を用いた科学的データベース、データモデリング、シミュレーション解析、デジタル実験室、電子的実験ノート、論文・報告書作成、電子的成果発行など、あらゆる電子化された科学的研究活動を指す。

Ｔ型・Π型人材

統計科学、数理科学、情報科学など横断的基幹科学技術の専門力量を有する研究者の中で、研究対象を明確に有する学術分野に対する専門性も単数ないしは複数有し、それら固有学術分野でも活躍可能となる人材。固有学術分野での横断的科学技術の利用の知の構築の過程の中で、革新的な横断的基幹科学技術開発に繋がることも多い。

データマイニング

大量のデータが表す全体的ないし部分的な傾向、規則性、パターンを、計算機上の種々のアルゴリズムを用いて解析、把握する技術の総称である。特にその中でも、人間にとって有用な知識を見いだすという目的に重点を置く技術を知識発見技術と呼ぶ場合が多い。このような解析のために、統計数理、離散数学、データベース、情報検索、機械学習、パターン認識、人工知能など、幅広い分野の情報処理理論・技術が用いられる。

機械学習

分類や予測など目的とするタスクに関して、与えられたデータを用いて精度などの性能評価指標を高めるための計算機実装可能な方法及びアルゴリズムの総称。現在では、音声認識、画像認識、テキストマイニング，自然言語処理をはじめとして、広範な領域で利用されている。

データ同化

複雑なシステムのシミュレーションにおいて、仮定された物理モデルのパラメータ、初期値、境界条件などを大規模な観測データを用いて、修正・改良する方法。 現在では遺伝子ネットワーク推定など、第１原理モデルの利用が困難な他の領域でも利用されている。

ムーアの法則

コンピュータを構成する大規模集積回路(LSI)技術の発達において、単位面積あたりに含まれるトランジスタ数（集積度）の増加率の長期的傾向の経験的規則性に関する予測である。インテルの創業者ムーアは 1975 年に今後２年毎に２倍になるだろうと述べたと言われるが、一般には 18 か月ごとに２倍（５年で約１０倍）になる傾向を指す場合が多い。

RFID (Radio Frequency IDentifier)

種々の情報を書き込み可能で、かつ特定周波数の電磁波を用いてその情報を外部読み取り可能な大きさ数ミリ以下の集積回路チップである。元々は複数の電子素子で構成される電子回路基板であったが、近年では小型化、１チップ化されている。また自らは電源を持たず、外部至近距離から照射される電磁波よって動作するパッシブタイプが主流となっている。RFID を商品や動物などに埋め込みあるいは貼り付けることによって、各々の個体認識や管理を電子的に容易に行う目的で多用されている。

＜参考文献＞

[1] 科学技術基本計画 (平成 23 年８月 19 日閣議決定).

[2] 金間大介，野村稔，農業をめぐる IT 化の動き ―データ収集, 処理, クラウドサービスの適用事例を中心に―, 科学技術動向，142, 13-18, (2014): http://data.nistep.go.jp/dspace/handle/11035/2473

[3] 研究開発の俯瞰報告書　電子通信分野 248-262 (2013).

[4] 国土交通省観光庁ホームページ：ＧＰＳを利用した観光行動の調査分析に関するワーキンググループ http://www.mlit.go.jp/kankocho/news04_000066.html

[5] 情報・システム研究機構 クロストーク： http://rois.ac.jp/topics/index.html

[6] 情報通信白書 第３節 ビッグデータの活用が促す成長の可能性: http://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h25/pdf/n1300000.pdf

[7] 設置された研究組織の Web サイト

NIH, Big Data Centers of Excellence (July 2013)
http://www.nih.gov/news/health/jul2013/nih-22.htm

University of Virginia, Big Data Institute (Sept. 2013)
http://dsi.virginia.edu/

University of Rochester, Institute for Data Science (Oct. 2013)
http://www.rochester.edu/data-science/

University of California Berkeley, Berkley Institute for Data Science (Nov. 2013)
http://vcresearch.berkeley.edu/datascience

University College London, UCL Big Data Institute (Dec. 2013)
http://www.ucl.ac.uk/news/news-articles/1213/UCL_Elsevier_partnership_181213

Columbia University, Institute for Data Science and Engineering (Nov. 2013)
http://idse.columbia.edu/

[9] 首相官邸ホームページ：電子行政オープンデータ戦略 http://www.kantei.go.jp/jp/singi/it2/pdf/120704_riryou2.html

[10] 総合科学技術会議「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」http://www8.cao.go.jp/cstp/budget/h25package.pdf

[11] 総務省 HP：「ICT 利活用の促進」内の「オープンデータ戦略の推進」について http://www.soumu.go.jp/menu_seisaku/ictseisaku/ictriyou/opendata/index.html

[12] 統計数理研究所 統計思考院： http://www.ism.ac.jp/shikoin/index.html

[13] 統計数理研究所『数学協働プログラム』: http://coop-math.ism.ac.jp/

[14] 東北大学 原子分子材料科学高等研究機構 https://research.wpi-aimr.tohoku.ac.jp/jpn/spotlight/777

[15] 日本学術振興会 二国間先端科学シンポジウム： http://www.jsps.go.jp/j-bilat/fos/index.html

[16] 日本学術会議 勧告, 統計学の大学院研究教育体制の改善について, 1983 年 11 月

[17] 日本学術会議　提言『第 22 期学術の大型研究計画に関するマスタープラン（マスタープラン 2014）』, 2014 年 3 月
http://www.scj.go.jp/ja/info/kohyo/pdf/kohyo-22-t188-1.pdf
[18] 野田 五十樹 : 災害救助支援のための情報共有プラットフォーム ーデータ仲介による情報システム連携ー, 産総研シンセシオロジー5 巻 2 号, 113-125, 2012 年 5 月
[19] 林知己夫, 第 2 回日仏データ解析セミナー ーデータの科学とその応用ー, 計算機統計学, 第 5 巻第 2 号 197-199 (1992).
[20] 林知己夫, 回想＜統数研の創成期＞, 統計数理研究所 50 年のあゆみ (1994).
[21] 濱中淳子, 「ポスドク就職難民問題～解決のための処方箋は何か～」, 大学と学生, 2008 年 7 月.
http://www.mext.go.jp/b_menu/toukei/chousa01/kihon/kekka/k_detail/1329235.htm
[22] 樋口知之, データ・サイエンティストがビッグデータで私たちの未来を創る, 情報管理, 56 巻 1 号, 2-11, 2013 年 4 月
[23] 丸山宏, ビッグデータの時代に一番欠けているのは人財である, Harvard Business Review, 特集 ビッグデータ競争時代 2013 年 1 月
[24] 文部科学省委託事業　ビッグデータ利活用によるイノベーション人材育成ネットワークの形成『データサイエンティスト育成ネットワークの形成』
http://datascientist.ism.ac.jp/index.html
[25] 文部科学省 『平成 25 年度学校基本調査（確定値）』:
http://www.mext.go.jp/component/b_menu/other/__icsFiles/afieldfile/2014/01/29/1342607_1_1.pdf
[26] 文部科学省, 第 4 期科学技術基本計画, 科学技術政策（2011 年 8 月閣議決定）
http://www8.cao.go.jp/cstp/kihonkeikaku/kihon4.Html
[27] 文部科学省 平成 20 年度「産学連携による人材育成事業」プロセスイノベーター育成プログラムの開発（関西大学）
http://www.scj.go.jp/ja/info/kohyo/pdf/kohyo-22-t188-1-1-3.pdf
[28] 文部科学省 平成 25 年度戦略目標『分野を超えたビッグデータ利活用により新たな知識や洞察を得るための革新的な情報技術及びそれらを支える数理的手法の創出・高度化・体系化』第 7 節
http://www.mext.go.jp/b_menu/houdou/25/03/attach/1331492.htm
[29] I. Ayres, Super Crunchers, Why thinking-by-numbers is the new way to be smart. 邦訳, イアン・エアーズ著, 山形浩生訳, その数学が戦略を決める, 文春文庫.
[30] Big Data Public Private Forum (BIG):
http://www.slideshare.net/edwardcurry/big-data-public-private-forum-big-european-data-forum-2013
[31] Bureau of Labor Statistics, Employment Projections:
http://data.bls.gov/projections/occupationProj

[32] W. S. Cleveland, Data Science: An Action Plan for Expanding the Technical Areas of the Field of Statistics. ISI Review, 69: 21-26 (2001).

[33] T.H. Davenport and D.J. Patil, Data Scientist: The Sexiest Job of the 21st Century, Harvard Business Review, 2012 年 10 月

[34] EU funding opportunities for Big Data in Horizon 2020, http://www.uni-gr.eu/fileadmin/NEWS/VERANSTALTUNGEN/2013/Luxinnovation_Meisch_Hijazi_Horizon2020_Big_Cata_uniGR_final_v2.pdf

[35] D. Ferrucci, et al, "Building Watson: An Overview of the DeepQA Project, " AI Magazine, 59-79 (2010).

[36] C. Hayashi, What is data science? -Fundamental concept and heuristics examples, IFCS-96, Abstracts, Col.1, 53-56.

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

[38] IBM Watson Hard At Work: New Breakthroughs Transform Quality Care for Patients: http://www-03.ibm.com/press/us/en/pressrelease/40335.wss

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

[40] M. Lorenz, "Data driven journalism: What is there to learn?" IJ-7 Innovation Journalism Conference, Stanford (2010).

[41] 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

[42] National e-Science Centre: http://www.nesc.ac.uk

[43] P. Naur. "The science of datalogy". Communications of the ACM 9 (7): 485. doi:10.1145/365719.366510 (1966).

[44] NHK ホームページ:震災ビッグデータ http://www.nhk.or.jp/datajournalism/

[45] NSF Priority Areas: https://www.nsf.gov/od/lpa/news/publicat/nsf04009/cross/priority.htm

[46] NSF, Cyber-enabled Discovery and Innovation (CDI): http://www.nsf.gov/crssprgm/cdi/

[47] NSF. Cyber-Physical Systems (CPS): https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=503286

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

[49] Study in Korea run by the Government of Korea :
http://www.studyinkorea.go.kr/en/sub/overseas_info/request/universityList.do
[50] J. Taylor, e-Science definition: http://www.e-science.clrc.ac.uk.
[51] J.W. Tukey, Exploratory Data Analysis, Addison-Wesley (1977).
[52] White House, Materials Genome Initiative: http://www.whitehouse.gov/mgi
[53] H. Varian, Keynote Presentation at the 2008 Almaden Institute:
http://www.youtube.com/watch?v=D4FQsYTbLoI
[54] C. F. J. Wu, "Statistics = Data Science", 1998 P.C. Mahalanobis Memorial Lectures.

＜参考資料＞情報学委員会Ｅ-サイエンス・データ中心科学分科会審議経過

平成２３年

１２月２８日　Ｅ-サイエンス・データ中心科学分科会（第１回）

・委員長、副委員長、幹事の選出
・各委員からの報告
・今後の活動方針について

平成２４年

３月９日　Ｅ-サイエンス・データ中心科学分科会（第２回）

・「学術の大型施設計画・大規模研究計画」の動向について
・情報学委員会で議論を深めるべき議題について
　サイエンスの中における情報学、人材育成
・情報学分野のアウトリーチ活動について
・東日本大震災に関連した活動について

１２月２８日　Ｅ-サイエンス・データ中心科学分科会（第３回）

・マスタープラン策定に関する分科会の方針について
・ビッグデータの学術研究のあり方について
・今後の活動について

平成２５年

１２月２６日　Ｅ-サイエンス・データ中心科学分科会（第４回）

・今後の分科会実施必要事項・日程確認
・参考人講演　東京大学大学院工学系研究科 坂田 一郎 教授
「データ中心科学を活かす社会システム構築と人材育成戦略」
・ビッグデータ時代の人材育成に関する提言内容に関する審議
・今後の活動について

平成２６年

１２月２６日～２月２５日　　メール審議

２月２６日　Ｅ-サイエンス・データ中心科学分科会（第５回）

・分科会報告または提言案の策定
・今後の分科会実施必要事項・日程確認

２月２７日～３月３０日　　メール審議

３月３１日　Ｅ-サイエンス・データ中心科学分科会（第６回）

・分科会提言案の検討・承認
・日本学術会議第２２期学術の大型研究計画
に関するマスタープラン提言に関する説明

○月○日　日本学術会議幹事会（第○回）

・情報学委員会地球・E−サイエンス・データ中心科学分科会提言「ビッグデータ時代に対応する人材の育成」について承認