科学技術動向研究

米国政府のビッグデータへの取り組み

野村　稔
客員研究官

1 はじめに

現在、ビッグデータをめぐっての研究開発が、産業界・アカデミア・各国政府によって盛んに進められている。ビッグデータに必ずしも明確な定義はないが、巨大なデジタルデータの総称である。データから有意な情報を抽出することは従来からIT技術が得意としてきた機能であり、それだけを見れば特に新しい話ではない。しかし、生み出されるデータの量・速度・種類などに桁違いに大きな変化が起きている。この変化に追随するためにデータの蓄積方法や処理方法における対応が盛んに進められており、さらに今までとは異なる新しい動きも始まっている。それは、従来は、蓄積されているのに活用されなかった膨大なデータの中から、有意な情報を抽出し、新たな「価値の創出」を図ろうとする研究開発の動きであり、この動きこそ、ビッグデータが大きな関心を集める理由である。

米国政府は、2012年3月にビッグデータの利活用を目的とした研究開発イニシアティブを発表した。オバマ政権の科学技術政策には、具体的に推進する5つのイニシアティブがあり[1)]、ビッグデータはそのひとつとして位置づけられている。ビッグデータのイニシアティブの推進のために、6つの政府機関が2億US$以上を投じて、大規模なデジタルデータの取り扱いに必要とされる技術の向上が図られる。最も興味ある視点は、ビッグデータを、インターネットと同等のインパクトを世の中に与えうるものとみなしていることである。すなわち、ビッグデータは様々な領域に非常に大きい影響を与えるものとしてとらえられている。

ビッグデータへの関心は、米国に限ったものではない。欧州ではEUプロジェクトが科学研究コミュニティでのデータ増大の課題に対して欧州全体としての解を見出そうとしている。

本紙では、2章でビッグデータとは何かを紹介し、3章で米国政府で発足した研究開発イニシアティブの内容を紹介し、4章で特に注目される諸点を探る。

なお、日本におけるビッグデータの議論においては、個人情報やセキュリティ問題などの解決すべき課題も多く挙げられている。「法を整備し企業がビッグデータの利活用に萎縮することのない状況を作れば画期的なアイディアが創出する」[2)]との意見もある。しかし、ビッグデータが世界で大きく注目される背景には、自由にアクセスしうる情報がすでに爆発的に増加しているという状況がある。したがって、セキュリティ等の課題は他書にゆずり、本稿では主にビッグデータのもたらすであろう可能性について、主に米国におけるポジティブな動きを中心に伝えていくことにする。

注1：日本では2005年以降、文部科学省の「情報爆発時代に向けた新しいIT基盤技術の研究（info-plosion）」、経済産業省による「情報大航海プロジェクト」、独立行政法人日本学術振興会による最先端研究開発支援プログラム（FIRST）の「超巨大データベース時代に向けた最高速データベースエンジンの開発と当該エンジンを核とする戦略的社会サービスの実証・評価」などの公的支援がある。最近の国家戦略会議・総合科学技術会議・文部科学省の資料には、以下のような記載が見られる。

メニューへ戻る

- 2012年7月30日の「国家戦略会議」での決定を経て7月31日に閣議決定された、「日本再生戦略　～フロンティアを拓き、「共創の国」へ～」[3]の「Ⅳ．日本再生のための具体策」の第2章、科学技術イノベーション・情報通信戦略の「重点施策：情報通信技術の徹底的活用と強固な情報通信基盤の確立」に、「情報通信技術の進展に伴い収集等が可能となった多種多量データ（ビッグデータ）の利活用や情報通信技術を活用した異分野融合等、官民が保有するデータの利活用促進を図る」とある。
- 2012年7月30日に開催された第103回総合科学技術会議本会議の参考資料1-2参考2「平成25年度　重点施策パッケージの重点化課題・取組」[4]に重点化取組のひとつとして「大規模情報（ビッグデータ）の利活用の基盤技術の開発・標準化・普及促進」がある。
- 2012年7月5日に開催された文部科学省の情報科学技術委員会（第77回）の資料2「ビッグデータ時代におけるアカデミアの挑戦～アカデミッククラウドに関する検討会　提言～」[5]に「ビッグデータの持つ可能性を最大化するため、データ科学の高度化に資する情報科学技術分野の研究開発やアカデミッククラウド環境構築のためのシステム研究等のビッグデータに関する研究開発、研究開発法人等におけるビッグデータ活用モデルの構築に関する事業について、分野間連携、国際連携、人材育成の観点に十分留意しつつ、早急に開始する必要がある」とある。

2 ビッグデータとは何か

2-1 ビッグデータとは

ビッグデータとは、必ずしも明確な定義はないが、巨大なデジタルデータの総称である。ここでいうデータとは、どこか一箇所に集められたデータだけではなく、ソーシャル・ネットワーキング・サービス（SNS）などの普及に伴って巨大化したWeb情報、インターネット上に蓄積される大量の写真や動画、センサーが検出し送出した膨大な「モノ」からの情報、スーパーコンピュータなどで生成される巨大な数値データなど様々な分野の様々な種類のデータが具体例として挙げられる。そのデータは、量的に既存の技術では管理できないほどに増え、そして複雑化している。

ビッグデータは、文書・画像・センサーデータなどのようなデータが大半を占めている。FacebookやTwitterなどのSNSの利用拡大に加え、大容量の映像データのサイトへの投稿が増えており、日々、ネット上で急増しているからである。また、あらゆる「モノ」をWebにつなぎネットワーク化するという考え方である「モノのインターネット」（Internet of Things：IOT）の具体化と進展があり、これもデータの急増を招いている。

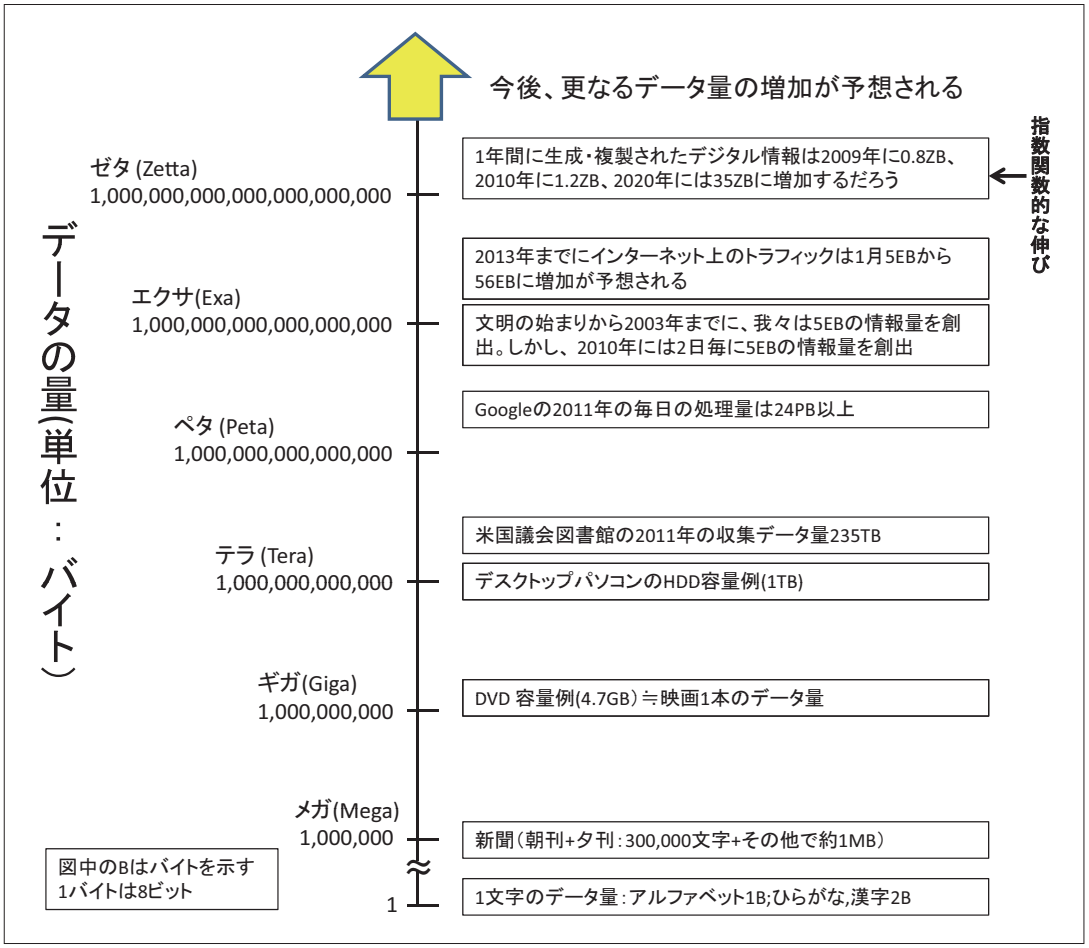
図表1に、データ量の増加状況を各種資料から抜粋して示す。ここで縦軸はデータ量を対数表示している。すでに、10の21乗というゼタバイトレベルのデータ量が現れてきており、それもここ数年は、指数関数的に増加している。そして、さらにこの勢いは継続しそうである。

2-2 データ規模増大の背景

データ規模が急速に膨大になってきた背景としては、Webデータの収集が以前に比べはるかに容易になったこと、デバイスからのデータ収集（携帯電話などからのデータ収集）やモノからのデータ収集が容易になったこと、そして大量データを扱える蓄積･処理技術が高度になったことが挙げられる。

Web情報の収集を例として採り上げると、サーチエンジンが必要とするWeb情報を収集したデータベース（DB）の作成は、1994年の後半ごろまでは主に人間による作業に依存していた。しかし、Webの世界的な普及によりその限界が見えてきた。この打開策として登場したのがクローラというプログラムであり、これによりWeb上の文書や画像などが周期的に取得され、自動的にDBとして収集されるようになった[11]。デバイスや「モノ」からのデータ収集に必要なセンサーと通信機能の小型化・低価格化の進展もデータ収集を容易にした。例えば、3軸加速度センサーは、チップの大きさが2000年の10 mm^2から2010年の2～3 mm^2以下へと小型化し、平均販売価格は2000年の約240円以上から2010年の約56円程度へ低価格化した。また、センサーにより収集等した

図表1　データ量の増加状況



出典：参考文献[6~10]を基に科学技術動向研究センターにて作成

データを送信する通信モジュールの低価格化が進展し、契約者数も増加した[12]。

データの蓄積・処理という面に関しては、巨大なデータセットが分散処理環境上の様々なコンピュータ上に分散して格納されている状況下でデータ処理を並列的に行う技術として、最近Hadoopの利用が大きく取り上げられている。これは、Google社のMapReduce[13]の仕組みをベースに作られたオープンソースであり、現実的な利用局面における種々の問題を解消すべく、多くの商用版が利用可能となっている。一方でリレーショナルデータベース（RDB）は20年以上もの最適化コンパイラ技術の蓄積があり、現時点においてはHadoopとRDBは併存状況にある。

大量のデータの収集、蓄積・処理が可能になりつつある現在は、そのデータからいかにして価値を生み出し、新産業の創出や社会課題の解決に繋げるかが鍵となって来ている。ビッグデータが注目される最も大きな理由はここにある。

2-3

ビッグデータの解析に求められるもの

ビッグデータの共通した特徴は、主に、量（多量性）、速度（リアルタイム性）、種類（多種性）であるとされている[12]。

多量性については、もしデータが大きいことが困るだけなら、サンプリングによって小さくして扱えばいいが、それでは結局、一部しか見ることができない、または重要なものを落とすかもしれないという懸念がある[14]。大きなデータの集合の中から、特徴的なパターンを発見したり、データの集合をある特徴のグループに分割したりすることでデータから知識を発掘する処理として、データマイニングがある。例えば、コンビニエンスストアの棚配列で、こう並べればより儲かるというパターンを見出せるといった例を考えるとわかりやすい。ビッグデータには、多量であるがゆえに、よくある特徴的なパターンと共に、希なパターンも含まれているはずで、むしろ、この希なパターンを発見することがビジネスでは求められるもののひとつであろう。

また、多量なデータには別のポテンシャルも内在している。今までの物理では一般的に、まず観察し、内在する法則を「式」に落とし込んで一般化し、この式を使うことで物理現象を再現してきた。例えば飛行機の場合、まず流体の

メニューへ戻る

式でシミュレーションを行い、動きを解析した。しかし、さらに高速になったときには、その式が成り立たなくなる。ビッグデータの解析では、「式」に落とし込む方法とは異なる方法で知識を抽出するとも言える。ただし、そのためにはより多くのデータがなければならない。同じパターンが見つけられる程度にデータが無ければならないからである。

リアルタイム性と多種性に関しては、より多くのデバイスやより多くの「モノ」からのデータ収集が可能になることで、データがリアルタイムに入力され、収集されることになる。したがって、即時処理によって、出力やフィードバックをすることが重視されることになろう。時々刻々と到来するデータを自動分析し、さらに判断し、迅速な意思決定に結びつけるストリーム・コンピューティング[15]がその一例と言える。

また別の観点であるが、データ量の増大とともにデータ中の曖昧さや不明確さが増加している。データ分析では、様々な不明確なデータを斟酌する必要がある[16]。ビッグデータ解析のひとつの要点となりうる。

2-4

ビッグデータ解析の萌芽事例

ここでは、ビッグデータ解析の萌芽事例を幾つか示し、どのような価値が生み出されているのかを見る。

2-4-1　家庭内の健康管理

日本で行われた情報大航海プロジェクトの研究のひとつに、センサーを活用したホームケアの実証研究があった。報告によれば、「血糖値をモニターするセンサー、運動量を測定する加速度センサーを用いて、糖尿病患者に継続的に自宅で血糖値を計測してもらい、その値に応じて、運動を促したり、食事量を抑えるといった、行動を促すメッセージを適切なタイミングで被験者に提供する。このメッセージを情報薬と呼び、情報薬を提供した時期においては、被験者の血糖値の上昇を抑制することに成功し、行動を促す情報を適切なタイミングで与えることが薬と同等の有効性を発揮することを実証した」[17]とされている。この研究は、個人の健康管理に焦点があてられているが、より広域での公共的な健康管理に発展していく可能性があるだろう。

2-4-2　精度の高い翻訳

大量なデータを活用することで翻訳の精度を向上できる例としてGoogle翻訳（Google Translate）がある。Google翻訳は、64種類の言語を瞬時に翻訳できる無料の翻訳サービスであり、Google翻訳が対応している言語同士であれば、どの組み合わせでも単語・文・ウェブページのいずれの単位でも自動翻訳できる。これは、従来型の自動翻訳のように特定の辞書や文法ルールを使った方法ではない。Google翻訳の方法についての記述を抜粋すると、「Google翻訳による訳文生成では、最適と思われる訳文を生成するために、何億もの文書からパターンを探しだす。既に人間の翻訳者によって翻訳された文書からパターンを検出することで、どのような訳文が適切かを考えて推定する仕組みになっている。このように大量のテキストからパターンを探す処理を「統計的機械翻訳」という。訳文は機械で生成される」[18]とある。これは、より大量のデータがあれば、より有意な結果が得られるという例である。

2-4-3　道路交通情報

定常時はもちろんであるが、非常時にも有効な道路交通情報の例がある。2011年の東北太平洋沖地震の直後、GPSデータを活用した道路交通情報が提供され、支援物資の輸送など物流効率化で威力を発揮した。個々の自動車が実際に走行した位置や走行速度などの情報であるプローブ情報が用いられている。ITS Japanが、民間4社が匿名かつ統計的に収集した通行実績情報を使用してプローブ統合交通情報として地図上に示し、同じ地図上に国土地理院が作成した「東北地方道路規制情報　災害情報集約マップ」情報をもとに通行止情報を反映させた。これは、官民連携により、被災地での通行実績・通行止情報をタイムリーに提供できた例である[19]。

2-4-4　防災対策の研究

2012年6月5日に平野博文文部科学大臣と米国NSFのSubra Suresh長官が日米間における災害研究協力の重要性に対して合意した。基本的合意内容は、災害に対する堅牢性（ロバストネス）および回復力（レジリアンス）の強化に関し、ビッグデータを通したコンピュータ科学・工学・社会科学・地球科学といった幅広い分野の研究協力および支援である。期待できる研究分野の具体例として、以下が挙げられている[20]。

○災害から得られた大量のデータを活用して、分析・モデリング・計算分析的能力やハザード確率モデルなどのアプリケーションの高度化を行うこと
○情報技術のレジリアンスや応答力を改良し、即時意思決定に必須である、リアルタイムなデータセンシング・可視化・分析・予測を可能にすること
○緊急時の準備と対応に関し、多様な学問分野、エンドユーザからの入力、全ての情報源からの

大量データなど、それぞれの知見を統合すること

2-4-5 産業界でのソリューション開発

センサーデータを活用した事例は、これまでにも産業界で多く散見される。例えば、「橋梁モニタリング」[21]、「農産物の生産・管理の見える化や生産工程の改善」[22]、「コンテキストアウェアネス技術」[23]を利用したサービス、エネルギー管理システム（EMS）などである。

また、サイトでの購買履歴情報やSNS情報を活用した商品・サービスの「おすすめ」や販売支援、GPSを活用した位置に直結した情報提供サービスなどがある。

これらはビッグデータの技術の向上に伴って、ますます発展および変化すると想定され、産業界では、もう一歩進んだソリューション開発に期待している。

3 米国政府によるビッグデータイニシアティブ

オバマ政権の科学技術政策では、5つのイニシアティブが示されており、ビッグデータは、そのうちのひとつである[1]。本章では、この米国政府による「ビッグデータイニシアティブ」の内容を紹介する。

3-1 科学技術政策室（OSTP）によるBig Data Initiativeの提示

OSTPは、ビッグデータの利活用を目的とした研究開発イニシアティブの内容を発表し[24]、このために新規に2億US＄以上を投じるとしている。大規模で複雑なデジタルデータから知識や洞察を引き出す能力を高め、国家の喫緊の課題解決に役立てることを目標としている。まずは、6つの政府機関（NSF、NIH、DOD、DARPA、DOE、USGS）が、ビッグデータを取り扱うためのツールや技術の向上に向けた研究投資を行う。Dr. John P. Holdren大統領科学顧問・大統領府科学技術政局長は、「過去の情報技術の研究開発の政府投資がスーパーコンピューティングとインターネットの創造に劇的な進歩をもたらしたのと同様に、このイニシアティブは、科学的発見・環境や生命医学研究・教育・国家安全保障の向上へ向けてビッグデータを使用するため、我々の能力を変容させる（transform）」[24]と語っている。

このイニシアティブでは、次の点が研究開発の目的として挙げられている。

○大量なデータの収集・蓄積・保存・管理・分析、そして共有のために必要となる最先端の革新的技術を前進させること
○それらの技術を、科学工学における発見の速さの加速・国家安全保障の強化・教育と学習の変容のために利用すること
○ビッグデータ技術の開発とその使用に必要とされる労働力を増強すること

このイニシアティブは、2011年に、「連邦政府はビッグデータに関する技術への投資が低い」と結論付けた科学技術に関する大統領評議会（President's Council of Advisors on Science and Technology）の勧告に対応したものと述べられている[24]。

米国の各政府機関では、すでにビッグデータに関わる様々な取り組みを開始している。OSTPは、2012年3月29日の研究開発イニシアティブの発表と同じ日に、ビッグデータの「ファクトシート[25]」をリリースした。OSTPは、この資料で、政府機関のミッションの遂行と、科学的発見によってイノベーションを推し進める「ビッグデータ革命」の課題に対して、現在進行中である政府関連プログラムをハイライトとして示しており、多くのプログラムが挙げられている。図表2にファクトシートに記載された機関とプログラム数を示す。

以下では、6つの政府機関の施策を紹介する。（その一部は上記ファクトシートにも記載されている）

図表2 ファクトシートに記載された機関とプログラム数

機関		プログラム数
Department of Defense (DoD)	国防総省	10
Department of Homeland Security (DHS)	国土安全保障省	1
Department of Energy (DoE)	エネルギー省	12
Department of Veterans Affairs (VA)	退役軍人省	9
Health and Human Services (HHS)	保健社会福祉省	5
Food and Drug Administration (FDA)	食品医薬品局	1
National Archives and Records Administration (NARA)	米国国立公文書館	1
National Aeronautics and Space Administration (NASA)	航空宇宙局	7
National Institutes of Health (NIH)	国立衛生研究所	23
National Science Foundation (NSF)	全米科学財団	16
National Security Agency (NSA)	国家安全保障局	3
United States Geological Survey (USGS)	米国地質調査所	1

出典：参考文献[25]を基に科学技術動向研究センターにて作成

メニューへ戻る

3-1-1　国立科学財団（NSF）と国立衛生研究所（NIH）の共同サポート

NSFとNIHでは、ビッグデータの科学工学の進展に向けた中核技術の研究開発が行われる。具体的には、大規模・多種類のデータセットの管理・分析・可視化・有用な情報抽出の手段となる中核の科学技術の進展をNSFとNIHが共同でサポートするため、“ビッグデータ”という名称の公募（solicitation）を行う。目的は、科学的な発見を加速し、他の方法では実現できない新しい調査領域を創出することである。NIHは、この募集の中で、特に、分子・細胞・電気生理学・化学・動作・疫学・臨床・健康や病気に関係するデータセットのイメージングなどに関心を抱いている。

3-1-2　国立科学財団（NSF）

NSFは、上記のビッグデータ公募による基礎研究への継続的なフォーカスに加え、データから知識を引き出す新しい方法、データを管理し、キュレート（下記、注2参照）し、コミュニティへ提供するインフラストラクチャ、教育や人材開発へ、新アプローチを含めた総合的で長期的な戦略を表明している。具体的には、まず、以下が行われる。

○次世代のデータ科学者や工学者を養成するために、研究大学に学際的な大学院プログラムを開発するように奨励する
○データを情報に変える3つの強力なアプローチである、「機械学習」「クラウド（Cloud）コンピューティング」「クラウド（Crowd）ソーシング」を統合する研究に対し、カルフォルニア大学バークレイ校を拠点とするプロジェクトに1000万US$をファンディングする
○地球科学者が、地球についての情報へアクセスし、分析し、情報共有できるシステムである“EarthCube”を支援するための第1回目の助成金を供与する
○大学生に向けて、複雑なデータに対してグラフィカルな可視化手法を使用できるように、トレーニングとサポートを行う研究グループへ200万US$を支給する
○タンパク質の構造や生物学的パスウェイを究明する統計学者と生物学者からなる重点研究グループをサポートするために140万US$を提供する
○「ビッグデータがどのように教育と学習を変えるか」を研究する学際的研究者を招集する

NSFのSubra Suresh長官は、「米国の科学者は、この新しい「データドリブン革命」によって生じた機会をしっかりととらえて欲しい。現在行っている研究は、新しい事業のための地ならしとなり、数10年先の米国の競争力の基盤強化につながるだろう」と述べている。

NSFは、さらに、異なった研究コミュニティ間で、データを利用可能とするメカニズム・政策・統治構造を開発するための科学研究プロジェクトを早期に起こすことも計画している[26)]。

3-1-3　国防総省（DoD）

DoDは、Data to Decisionsイニシアティブと名付け、各プログラムを開始している。ビッグデータを「大きな賭け」と位置づけ（“placing a big bet on big data”）、DoDの各部門間のシリーズにつながったプログラムに対して年間2.5億US$（新規研究プロジェクトには6000万US$を割当）を投入するとしている。各プログラムとしては、以下が挙げられている。

○新しい方法で大量のデータを利用し、自ら操作して意思決定ができる完全な自律的システムを作るため、センシング・知覚・意思決定支援などの要素を結びつける
○戦闘員や分析者を支援し、オペレーションを高度にサポートできるように状況認識機能を改善し、例えば、分析者が任意の言語のテキストから情報を引出すための能力を100倍改善することを目指す。また、分析者が観察可能な、オブジェクト数・活動数・イベント数を同様の規模で改善する

DoDは、これらの要求に適合

注2：キュレートまたはキュレーションの意味
ビッグデータからの価値創出という意味で、日本語になりにくいのだが、キュレートまたはキュレーションという言葉が大きな意味を持つと考えられる。
キュレーションは、ここではインターネット上の情報を収集しまとめること、または収集した情報を分類し、つなぎ合わせて新しい価値を持たせて共有することである。
本来、キュレーション（curation）は、「情報などを集め、整理し、新しい視点から価値を加えて、その情報を他者と共有する」といった意味で、その動詞形がキュレート（curate）である。
（http://kotobank.jp/word/%E3%82%AD%E3%83%A5%E3%83%AC%E3%83%BC%E3%82%B7%E3%83%A7%E3%83%B3 から）
（http://www.nttpc.co.jp/yougo/%E3%82%AD%E3%83%A5%E3%83%AC%E3%83%BC%E3%83%88.html から）

した、イノベーションを加速するために、数か月にわたり、ビッグデータに関して、懸賞付きのオープンコンテストを連続的に実施することにしている。

3-1-4 国防総省国防高等研究計画局（DARPA）

DARPAでは、半構造化データ（例えば、表・リレーショナル・カテゴリ・メタデータなど）や非構造化データ（例えば、テキスト文書・通信文のトラフィックなど）の両方から成る大量のデータを解析するための、計算手法やソフトウェアツールを開発する。今後4年にわたって毎年約2500万US $を投入予定の「XDATAプログラム」を開始している。主な研究課題は次のとおりである。

○分散データストア内の不完全なデータを処理するスケーラブルアルゴリズムの開発
○多様なミッションに応じて迅速にカスタマイズ可能なビジュアルリーズニングを容易にする人間とコンピュータ間の効果的なインタラクションツールの作成

XDATAプログラムは、柔軟なソフトウェア開発を可能とするためオープンソースのソフトウェアキットをサポートする。

この大規模投資が予定されているXDATAプログラムについては少し詳しく紹介する[27]。

まず基本的に、DARPAは、DoDや他の省庁機関と連動して使用ケースや運用コンセプトを開発するとしている。これは、技術開発が、運用サポートの専門知識をもつエンドユーザのニーズによっても先導されるからである。ライブラリ・API・コードがユーザのフィードバックによって洗練されていく「development-in-process」というソフトウェア開発モデルを採り、このプログラム期間を通してDoDや他の省庁機関から選定された人が確保される。

このプログラムでは、データを処理し可視化するために、高速で、スケーラブルで、かつ効果的な方法を開発することが重要とされており、単にデータの取得や変換をサポートするだけでなく、高速な検索や分析も可能にすることが要求されている。

このプログラムは、以下の4つの技術領域（Technical Areas：TAs）から構成されている。

TA1：スケーラブルな分析とデータ処理技術
TA2：可視化ユーザインターフェイス技術
TA3：ソフトウェア統合研究
TA4：評価

このプログラムのため､ワシントンDCにアジャイルでかつ共同作業によるソフトウェア開発・統合・テスト／評価を促進するための施設（technology integration facility）を設けることが予定されている。この施設でユーザとのインタラクションや使用ケースの開発が行われる。

3-1-5 国立衛生研究所（NIH）

NIHには、前記のNSFと共同の中核技術開発の他に、クラウドコンピューティング（以降、クラウドと略記する）上で利用可能な1,000ゲノムプロジェクト（下記、注3参照）の推進がある[28]。すでに、Amazon.com, Inc（以降、Amazon社とする）との共同により1000ゲノムプロジェクトで生成された世界最大規模の人の遺伝的変異に関するデータセットが、Amazon Web Services（AWS）クラウド上で自由に利用可能である。このデータサイズは200テラバイトであり、テキストでは1600万ファイルキャビネット分、標準DVDでは30,000枚以上のデータ量に相当する。現在の1000ゲノムプロジェクトのデータセットは、ビッグデータの典型例であると言える。その量はさらに膨大化しているため、それらを最大限に利用できるコンピューティング能力を持つ研究者が現時点ではまだほとんどいない状況である。AWSは、1000ゲノムプロジェクトをホスティングしており、研究者は、無料で公的に利用可能なデータセットにアクセスでき、自分が使用する計算サービスだけの費用を支払うだけでよくなる。

3-1-6 エネルギー省（DoE）

DoEは、5年間で2500万US $をファンディングする一部として、Scientific Discovery Through Advanced Computing（SciDAC）プログラムを通してSDAV研究所（Scalable Data Management, Analysis and Visualization Institute）[29]を設立する。SDAV研究所は、ローレンスバークレイ国立研究所がリードする形で、6つの国立研究所と7つの大学の専門知識をとりまとめる。そのゴールは、科学者が、データ管理や可視化を容易に行えるような新しく改良されたツールを開発することである。DoEでは、所有するスーパーコンピュータ上で動作するシミュレーションの規模や複雑さが増加してきており､このような新ツールの必要性が増大している。

計算規模の大幅な増加につれて、シミュレーションによって生成されるデータは、規模、複雑さが数桁も増加しており、この傾向は継続すると想定される。しかし、コンピューティングユーザは、データの管理や解析のため

注3：1000ゲノムプロジェクトは、2008年に開始したpublic-privateコンソーシアムであり、世界中の26 Populationsからの2,600人以上のゲノム変異（variation）の詳細マップの作成を目的としている。

メニューへ戻る

に、テラFLOPS（Floating point number Operations Per Second）時代に最適とされた時代遅れのツールを使用して知識発見を試みなければならないという状況に直面している。これらの課題に対応できる新しい技術やツールも存在するが、科学者はそうしたツールを知らないか、ツールの使用に不慣れであるか、または適切な計算機施設にそうしたツールが導入されていないなどが現状である。

SDAV研究所は、かかる課題に対処するために、データの管理・解析・可視化の3領域における技術的なソリューションを開発・配備し、その使用を通して各分野の科学者を支援することを目標としている。

3-1-7　米国地質調査所（USGS）

USGSは、地球システム科学に向けたビッグデータを対象とする。すでに、John Wesley Powell Center for Analysis and Synthesisを通して助成金を出している。このセンターは、科学者に対して、詳細な分析が行えるような場所と時間・最新のコンピューティング能力・巨大なデータセットの意味理解に有益なコラボレーションツールを提供することで、地球システム科学における革新的な思考を生じさせることを目指している。地球システム科学でのビッグデータプロジェクトによって、気候変動・地震の再発率・次世代の生物学的指標などの研究が進むと考えられている。

3-2

ビッグデータ上級運営グループの設置

これまでの米国政府における情報通信技術の研究開発は、国家科学技術会議（NSTC）が策定したネットワーキングおよび情報技術研究開発（NITRD：Networking and Information Technology Research and Development）プログラムに基づいて行われている。NITRDには15の政府機関が参加し、8の個別研究分野PCA（Program Component Areas）と、各機関が連携すべき優先課題を扱う4つの上級運営グループ（Senior Steering Group：SSG）がある。各省庁で実施される個々のプログラムは省庁間で連携を取りながら実施される。毎年、NITRD管轄プログラムの計画と予算に関してのBluebook（Supplement to the President's Budget）が発行されている。

ビッグデータに関しての上級運営グループ「Big Data（BD SSG）」は、2011年の早期に設置されており[30]、ビッグデータも省庁連携で推進すべき分野であるとの位置づけである。

このBD SSGは、連邦政府で行われているビッグデータの研究開発活動を調べ、調整の場を提供し、ビッグデータに関するイニシアティブの全体目標がどのようであるべきかを確認するために設置されている。この背景には、データ量が指数関数的に増加するに伴い、データの保存・アクセス・普及・使いやすさに関する懸念が増していることがある。そして、自動解析技術・データマイニング・機械学習・プライバシー・データベースの相互運用性などにおける研究が、多くの機関で進行中であり、それらを調査することで、今後、ビッグデータがどのように科学を進展させることができるかを確認することに役立てるとしている。

BD SSGの活動としては、この領域のイニシアティブがいかにあるべきかを定義するために専門家を招集して意見を求めること、既存の技術プロジェクトばかりでなく、教育サービス・コンテスト・民間セクターのイノベーションを利用するファンディングメカニズムなどを確認することが記されている。また、その機能としては、連邦政府全体の現在の活動に関する情報収集、イニシアティブがもつべき目標に対するビジョンの作成、政府内での議論を支援する適切な資料の開発、そして、現在の投資とリソースを活用する実装戦略の開発などが記されている。

4　米国のイニシアティブにおいて特に注目すべき点

この章では、3章で述べた、米国政府のビッグデータへの取り組みにおいて特に注目すべき点を挙げる。図表3は、3章の各機関の主な研究支援のポイントを研究対象に分けて示す。以下の議論はこの図表を参照しながら進める。

4-1

ビッグデータのとらえ方

Dr. John P. Holdren大統領科学顧問・大統領府科学技術政局長は、ビッグデータをインターネットと同じぐらいのインパクトを世の中に与えるかもしれないものとみなしている。すなわち、その影響領域が社会にも科学技術全般においても非常に大きいものとしてとらえられている。ビッグデー

図表3　米国の各機関のビッグデータに関する研究支援のポイントと研究対象

機関	技術開発	人材育成	データ共用	推進施策（施設提供、研究者招集など）
NSFとNIH	ビッグデータの管理、解析、可視化、情報抽出のための中核技術			
NSF	・機械学習、クラウドコンピューティング、クラウドソーシング（Crowd sourcing)を統合しデータから情報抽出 (1000万ドル)	データ科学者・工学者養成のための大学院プログラム開発の奨励 タンパク質の構造や生物学的パスウェイを究明する統計学者と生物学者からなる重点研究グループへの支援 （140万ドル） 可視化技術の習得などを支援し、ビッグデータの取り扱いを学部学生に教育する活動に助成 （200万ドル）	地球科学者への地球のデータへのアクセス、解析、共有支援 “EarthCube”	ビッグデータで教育と学習をいかに変革するかについて研究する学際的研究者の招集
DoD	・ビッグデータを活用した完全自律的システムに向けて、センシング、知覚、意思決定支援を統合 ・戦闘員や分析者を支援する状況認識機能向上（任意言語のテキストから情報を引出す能力を100倍改善。観察可能な、オブジェクト数、活動数、イベント数を改善） （年間2.5億ドル：内6000万ドルは新規プロジェクト）			ビッグデータによるイノベーションの加速に向け、数か月に渡り、オープンな懸賞付きコンテストを実施
DARPA	ビッグデータを解析する計算手法やソフトウェアツールの開発（XDATAプログラム） ・分散データストアの不完全データ処理用のスケーラブルなアルゴリズム開発 ・多様なミッションに応じた迅速にカスタマイズ可能な人とコンピュータのインタラクションツール ・オープンソースのソフトウェアキットの提供による柔軟なソフトウェア開発を指向 （4年間に毎年約2500万ドル）			アジャイルで且つ共同作業によるソフトウェア開発・統合・テスト/評価を促進する施設(technology integration facility)を設置
NIH			人ゲノム情報にアマゾンのAWSクラウド上で無料でアクセス可能	
DoE	・SDAV研究所を設立し、スーパーコンピュータでのデータの管理、分析、可視化ツールを開発、展開 （5年間で2500万ドル）			
USGS				地球、環境、気候などの解析が可能な場所・時間・計算能力の提供

出典：参考文献[24]を基に科学技術動向研究センターにて編集

タに関する米国のとらえ方を、文部科学省の「ビッグデータ時代におけるアカデミアの挑戦～アカデミッククラウドに関する検討会提言～」[5]では以下のように記している。

「米国のビッグデータイニシアティブ構想では、今後の重要な技術課題を強力に解決していく姿勢が示されている。ビッグデータを、スーパーコンピュータやインターネットと並んだ重要な分野ととらえており、大量データの核となる技術を向上させることで、安全保障、教育の改革、人材育成を実施することや、NSFがデータサイエンティスト育成のための大学院コースや、機械学習、クラウドコンピューティング、クラウドソーシングを行うことを技術課題ととらえている」

4-2 技術開発における可視化技術の重視

図表3で最も多いのが可視化であることが分かる。つまり、ビッグデータから価値を創出するためには、多くの課題を総合的に解決していくことが望まれているが、その中で最大の技術ポイントを挙げるとしたら可視化ではないかと考えられる。また、解析と可視化が密接な関係になっており、可視化に結びつかない処理は価値の創出につながりにくいと考えられる。可視化して「知を抽出」し、その後の「価値の創出」への行動に結びつけることが重要であり、ここがその源泉であると言える（図表4）。

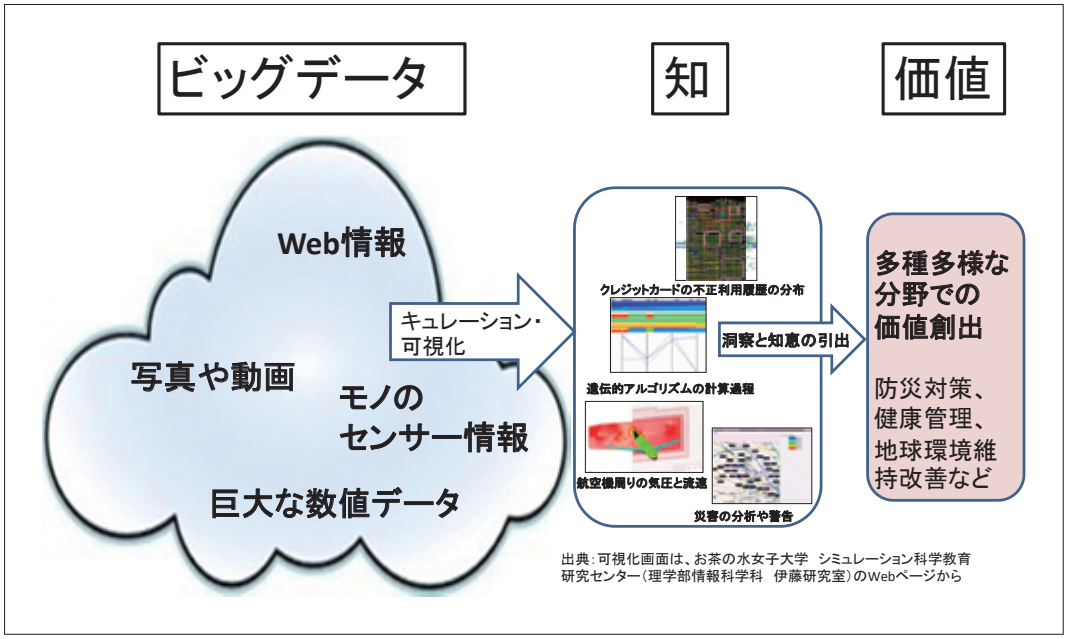
NSFとNIH、DARPA、NIH、DoEの研究テーマには可視化が盛り込まれており、その重要性を認識していることがわかる。

4-3 クラウドコンピューティングとの関係

クラウドについては、既報の「科学技術動向、2010年6月号」に詳しいが、ビッグデータの推進の背景のひとつにクラウドの普及がある。例えばNIHのプロジェクトはAmazon社のクラウドで使えるとある。クラウドで使えることの意味は、まず研究者が容易にかつ軽負担で使えることだろう。NIHの1000ゲノムプロジェクトの記述には、「AWSは1000ゲノムプロジェクトをホスティングしており、研究者は、無料で公

メニューへ戻る

図表 4　ビッグデータからの価値の創出



科学技術動向研究センターにて作成

的に利用可能なデータセットにアクセスでき、自分が使用する計算サービスに対して費用を支払うだけでよい」とある。

言うまでもなくビッグデータは大量のデータである。その処理のためには、大量のディスクと大量のコンピュータが必要になる。データも並列の入出力が行われないと、時間がかかり過ぎてしまう。複数のコンピュータにより複数のディスクへの並列アクセスができる環境を提供することで、入出力の短縮ができる。クラウドコンピューティングは、大量のディスクと大量のコンピュータを共に適える手段として、ビッグデータには欠かせない要素となっている。

4-4

人材育成への配慮

ビッグデータから価値を見出そうとすれば、数学的・統計学的・法学的の知識やビジネス管理における知識も必要であり、その人材育成が重要になる。NSF の記述の中には、データ科学者・工学者育成の大学院プログラム開発、統計学者と生物学者からなる研究グループへの支援、可視化技術の習得などを支援し学部大学生を育成する研究に助成、などのプログラムの準備が含まれている。

ビッグデータに限らないが、欧米のプロジェクトにはこうした人材育成までを包含しているものが多い。欧州の例であるが、全欧州の研究者に向けてハイエンドスーパーコンピュータを提供する PRACE プロジェクト[31] や FP7 の中のプロジェクトでは、人材育成（あるいは教育）と広報活動の専門家を用意している。

4-5

産業界および大学の積極的参画

OSTP の政策担当副ディレクターである Tom Kalil 氏は、ビッグデータの取り組みに対して民間企業や大学の積極的参画を呼びかけている。同氏は、「ビッグデータが生み出す機会を最大限に活用するために、政府に対する協力を業界企業・研究大学・非営利機関に呼びかけたい」とし、「政府だけでこれを推進することはできないのは明らかである。オバマ大統領の言う”全員が総力を挙げる（all hands on deck）”取り組みが必要」と語っている[32]。現に例えば NIH の 1000 ゲノムプロジェクトでは Amazon 社が協力している。

研究開発における問題のひとつは、プロジェクトが終わると研究が途絶えてしまい、研究成果が社会に生きる可能性が減ることだろう。社会活用まで達するよう、研究開発を持続および進展させていけるような仕組み作りが必要である。

4-6

データ共用の促進

ビッグデータイニシアティブでは、データの共用を促している。例えば、NIH では、1000 ゲノムプロジェクトで生成されたデータセットを AWS クラウド上で自由にアクセス可能としている。また、NSF では、科学者が、だれでも地球に関するデータへアクセスし、分析し、情報共有できる“EarthCube”をサポートしている。

米国では、政府機関が保有する

情報・データを入手できるサイトData.gov[33]があり、2009年5月21日に開設されている。これは、研究者コミュニティのためのデータ共用とは異なる目的があるが、このサイトでは、統計データの集計結果を公表すると共に、その基となった生データも、データをツールで加工した形式でも利用可能であり、それらのデータは、さらに自由に加工や分析することができる。

欧州にも、EUプロジェクトのEUDAT[34](European Data Infrastructure)があり、科学研究コミュニティでのデータ増大に対する課題に対して、欧州としてのソリューションを提供することを目的に活動している。したがって、米国も欧州もデータを共有することで効率よく研究を行えることをめざし、そのためのインフラストラクチャとツールの構築を指向することは、共通していると言えるだろう。

4–7 その他の注目点

DoDは、戦場での戦闘員や分析者を支援する状況認識機能の向上を挙げており、その内容として、任意の言語のテキストから情報を引出すための能力を100倍改善することを目指すとある。軍事的な用途であるらしいが、将来、社会へ展開されることもありえるため、その時に起こる劇的な変化を想定すると注目すべき研究開発かもしれない。

またDARPAが、分散データストアの不完全データ処理用のアルゴリズムや、迅速にカスタマイズ可能な人とコンピュータのインタラクションツールを開発する点も興味深い。

5 おわりに

ビッグデータとは、明確な定義はないが巨大なデジタルデータの総称であり、既存の技術では管理できないほどに量的に増え、そして複雑化しているデータである。そしてそこから価値を創出する動きが活発化している。

今回採り上げた米国政府の「ビッグデータ(Big Data)」イニシアティブの内容は、従来から実施されているプロジェクトも包括して述べられており、率直なところ、色々なものの寄せ集めという感もある。しかし、NITRDでの動きにも見られるように、省庁間連携がとられ、ビッグデータというテーマに向けて総合的に推進する姿勢が強く感じられる。これは、オバマ政権の科学技術政策の特徴である「政策の包括的な枠組みの中に、大学や民間企業、さらには一般の人々をより積極的に参加させることにより、大きな政策効果をもたらそうという試み」「ひとつの理念の下で、多様な連邦政府の施策を包含することによりイニシアティブを形成する」[1]を実現しようとする、ひとつの例ともいえる。

研究開発の取り組みにおいて、最も興味ある点は、「ビッグデータのとらえ方」である。インターネットと同様、新たなパラダイム創出に寄与しうる科学技術であるとみなし、様々な領域に非常に大きい影響を与えるものととらえている。あわせて、「可視化技術の重視」「クラウドコンピューティングとの関係」「人材育成への配慮」「産業界および大学の積極的参画」「データ共用の促進」なども注目すべき点である。共同作業を促進するための施設や計算パワーの提供などの推進施策への配慮も見える。こうしたイニシアティブ主導での研究開発成果は、数年または数十年後に社会に普及・浸透し、大きなイノベーションを起こすとも考えられるため、研究動向には今後も注視すべきであろう。

ビッグデータ研究開発の強化へ向けた世界共通の方向性やその課題の難度等を考えると、今後の研究開発にはグローバル連携も必須になろう。2012年6月には文部科学大臣と米国NSFの長官が日米間における災害研究協力の重要性に対して合意しており、ビッグデータにも期待している。例えば、NSFは、この合意をSAVI(Science Across Virtual Institutes)アワードのひとつである「GLOBAL RESEARCH ON APPLYING IT TO SUPPORT EFFECTIVE DISASTER MANAGEMENT (GRAIT-DM)」に「MEXT and NSF Collaborate on Big Data and Disaster Research」を設けている[35]。日本でもビッグデータ研究を推進させるために、こうしたグローバルな研究協力において、顕著な価値を生み出していくことが望ましい。

謝辞

本稿の執筆にあたり、東京大学の喜連川優教授(東京大学地球観測データ統融合連携研究機構長・東大生研戦略情報融合国際研究センター長)および早稲田大学理工学術院の山名早人教授から多くの情報と助言を頂きました。この場を借りて、厚く御礼申し上げます。

メニューへ戻る

参考文献

1）「緊縮財政下における米国の科学技術政策：2012 年 AAAS 科学技術政策年次フォーラム報告」、科学技術動向 2012 年 7・8 月号
2）http://www.mizuho-ir.co.jp/publication/navis/017/pdf/navis017_07.pdf
3）http://www.npu.go.jp/policy/pdf/20120731/20120731.pdf
4）http://www8.cao.go.jp/cstp/siryo/haihu103/sanko2.pdf
5）「ビッグデータ時代におけるアカデミアの挑戦～アカデミッククラウドに関する検討会　提言～」、情報科学技術委員会（第 77 回）資料 2
6）http://gigaom.files.wordpress.com/2010/05/2010-digital-universe-iview_5-4-10.pdf
7）http://idcdocserv.com/1142
8）http://www.i-cio.com/features/august-2010/eric-schmidt-exabytes-of-data
9）A Vision for Exascale（Mark Seager,Intel）2012.6 ISC12
10）Big data:The next frontier for innovation, competition, and productivity 、McKinsey Global Institute 2011.5
11）http://ja.wikipedia.org/wiki/%E3%82%AF%E3%83%AD%E3%83%BC%E3%83%A9
12）「ビッグデータの活用の在り方について」、情報通信審議会　ICT 基本戦略ボード　ビッグデータの活用に関するアドホックグループ取りまとめ、2012 年 5 月 17 日
13）http://research.google.com/archive/mapreduce.html
14）宇野毅明、「ビッグデータ高速処理に向けた計算理論的アプローチ」、情報処理学会連続セミナー 2012　ビッグデータとスマートな社会、2012 年 6 月 25 日
15）http://www-06.ibm.com/ibm/jp/provision/no65/pdf/65_article2.pdf
16）http://www.slideshare.net/IBMDK/global-technology-outlook-2012-booklet
17）喜連川優、「第 4 のメディアが作り出すビッグデータの時代」（ProVISION Winter2012 No.72 p13-14）
18）http://translate.google.com/about/intl/ja_ALL/
19）http://www.its-jp.org/saigai/
20）http://www8.cao.go.jp/cstp/tyousakai/innovation/ict/4kai/siryo4.pdf
21）櫻井保志、「データストリームのためのマイニング技術とその応用」、情報処理学会連続セミナー 2012　ビッグデータとスマートな社会、2012 年 6 月 25 日
22）http://jpn.nec.com/press/201207/20120713_03.html
23）「ユビキタスネット社会のコンテキストアウェアネス技術研究の動向と課題」、科学技術動向 2007 年 8 月号
24）http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf
25）http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_fact_sheet_final.pdf
26）http://www.nsf.gov/news/news_summ.jsp?cntn_id=123607
27）Broad Agency Announcement XDATA　DARPA-BAA-12-38
28）http://www.1000genomes.org
29）http://sdav-scidac.org/report.html
30）http://www.nitrd.gov/Index.aspx
31）http://www.prace-project.eu/?lang=en
32）http://www.whitehouse.gov/blog/2012/03/29/big-data-big-deal
33）http://www.data.gov/
34）http://www.eudat.eu/
35）http://www.nsf.gov/news/special_reports/savi/awards.jsp#grait-dm

執筆者プロフィール

野村　稔

科学技術動向研究センター　客員研究官

http://www.nistep.go.jp/index-j.html

企業にてコンピュータ設計用 CAD の研究開発、ハイ・パーフォーマンス・コンピューティング領域、ユビキタス領域のビジネス開発に従事後、現職。スーパーコンピュータ、LSI 設計技術等の科学技術動向に興味を持つ。現在、科学技術イノベーション政策における「政策のための科学」に関する研究に従事し、研究開発のもたらす社会的・経済的効果の定量化・可視化に取り組んでいる。

メニューへ戻る