3.10 ビッグデータ

計算機やセンサー装置が低価格化し、性能が向上するに従って、膨大なデータが急速に生成され格納できる環境が整いつつある。例えば、Facebook 社で扱うデータ量は日々600 テラバイトの規模で増加しており、全体で 300 ペタバイトのデータを管理している [1)]。また、ハードディスクの価格は 1 テラバイト当たり数千円程度であり低価格化が進んでいる。大量のデータを生成・格納できる環境が整う一方で、データを活用する観点においては、コンピュータ将棋や IBM 社の質問応答システム Watson などに代表されるように、膨大なデータ（将棋の棋譜の記録やクイズの過去問、Wikipedia や新聞等のデータなど）を活用することで、高度な分析処理が可能なシステムが実現されている。このように急速に生成・管理される大量データはビッグデータと称され社会的に注目されている。

ビッグデータと一口に言っても、それを構成するデータは様々な種類がある。

各種のネット接続端末から上がって来るデータは、今までコンピュータが扱ってきた企業内の業務データのように構造化されたデータとは違い、SNS のテキスト、画像・音声・動画データ、IC カードや RFID 等の各種センサーで検知され送信されるデータなど、非構造化データがかなりの割合を占める。技術の進歩は、これらのデータを収集・蓄積・分析し、大量で多種多様なデータを扱うことを可能にしている。

ビッグデータは、大量かつ多種・多様なデータを、許容できる時間内に効率的に収集・蓄積・処理・分析し、活用するための技術 [2)] といえる。量的に増大する様々な種類のデジタルデータに埋もれている未知の知識や洞察を抽出し、研究やビジネスに活用したいという期待がある。しかしながら、活用したいデータが、従来のログや SNS データなどのテキストデータから、画像、音声、動画などの、サイズのより大きなデータや、IoT などの進展に伴うリアルタイムのセンサーデータに変わってきており、現状の技術レベルではビッグデータ活用の要件を満たすには不十分であるため、ビッグデータに対応した新たな研究開発が必要である。

また、データを活用するという点においては、データの収集、データの解析結果の表示、解析結果の活用の際に生じる著作権の問題、個人情報の取得、パーソナルデータを匿名化することによって利活用をはかりたいという要望とプライバシー保護の要請をいかにして調和させつつ活用を図るかという個人情報保護の問題、のふたつが発生する。これらは、技術開発だけでは解決することが難しく、対応する法律の整備、運用において対処する必要があり、他国も含めた法律問題などの調査研究が必要である。

本俯瞰区分では、様々な分野のビッグデータの統合解析を可能にするデータ処理基盤、必要な知識を効率的に取り出すための技術、ビッグデータが広く活用されるための社会受容やステークホルダ同士の連携を促進するための技術と社会的仕組みを対象とする。このような観点から、ビッグデータに対応した研究開発領域として 4 つの基盤技術、5 つの応用分野、著作権法や個人情報保護法など 2 つの法律面を含む合計 11 のテーマを取り上げた。基盤技術は、ビッグデータの収集・蓄積・処理を効率的に実行する技術、ビッグデータを分析して有用な知見を引き出す技術、それらの収集や分析にクラウドソーシングを効率的に活用する技術、データの処理において秘密性を担保できる技術について、現状と今後必要となる課題をまとめた。応用分野は、すでに大量のデータ蓄積がありビッグデータとしての活用が始ま

っている分野から、IT メディア、ゲノム、教育分野、社会インフラ、オープンデータを選択し、分野ごとの取り組み状況を説明する。

基盤技術：

(1) ビッグデータ基盤技術
(2) ビッグデータ解析技術
(3) クラウドソーシング
(4) プライバシー保持マイニング技術

応用分野：

(5) IT メディア分野におけるビッグデータ
(6) ライフサイエンス分野におけるビッグデータ
(7) 教育とビッグデータ
(8) 社会インフラとビッグデータ
(9) オープンデータ

法律面：

(10) 著作権とビッグデータ
(11) プライバシーとビッグデータ

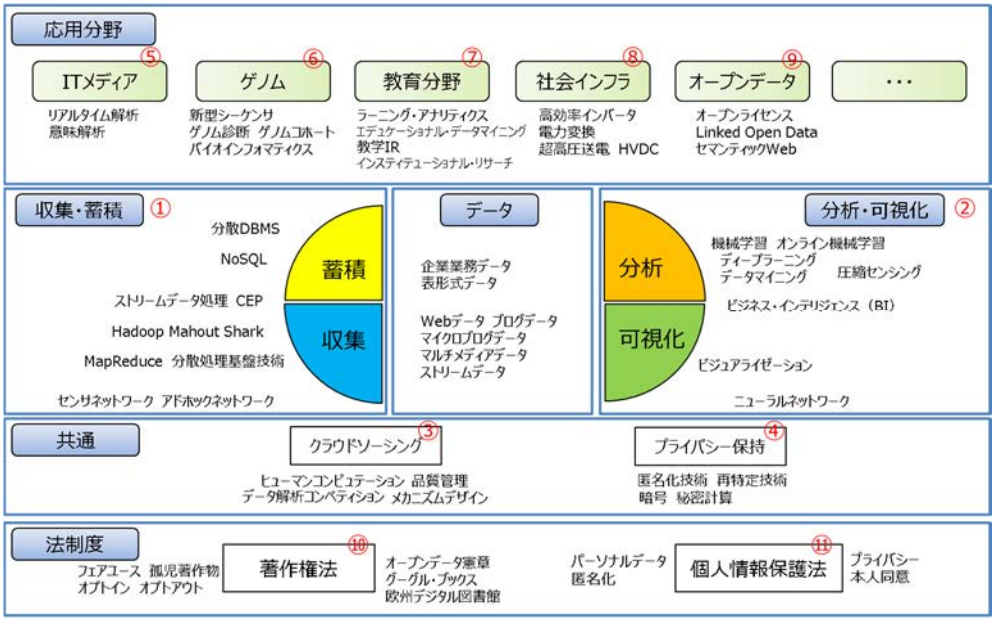

図 3.10.1 ビッグデータの俯瞰図

1) How Facebook Compresses Its 300 PB Data Warehouse, http://www.enterprisetech.com/2014/04/11/facebook-compresses-300-pb-data-warehouse
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～, 2012

3.10.1　ビッグデータ基盤技術

（1）研究開発領域名

ビッグデータ基盤技術

（2）研究開発領域の簡潔な説明

大量なデータを管理・分析するためのベースとなる基盤技術に関する研究開発

（3）研究開発領域の詳細な説明と国内外の動向

ビッグデータの活用において、実際のビジネス業界では、会員カードを活用して消費者の購買履歴を取得・分析することで、商品の在庫管理を最適化してコストダウンを図り、また消費者が求める新たな商品を開発することで利益を上げる例が挙げられる。特に Web スケールの大量のビッグデータを扱う事業者は、Warehouse-Scale Computers（WSCs）[1]と称されるように、数 1000 台規模の計算機をデータセンターに設置し、WSCs の環境に適したソフトウエアおよびハードウエアから成るビッグデータ基盤を構築することで、大規模なデータ処理を実現している。このようなビッグデータ基盤に関する技術は、処理の対象となるデータが蓄積されたデータであるか／時々刻々と生成されるデータであるかの観点から、2 つの技術領域に分類することができる。

蓄積されたデータを対象とする領域: 技術を細分化すると、数 10 台規模のハイエンドな計算機を用いる分散データベース管理技術（分散 DBMS）と、分散 DBMS の機能を簡略化して 1000 台を超える大量のコモディティ計算機を用いる NoSQL および MapReduce に代表される分散処理基盤技術がある。前者の分散 DBMS については Oracle、IBM、Microsoft 社などの米国企業が商用化を主導している。また低価格が進む主記憶を活用した分散 DBMS として、欧州の SAP 社や米国のスタートアップである VoltDB 社の製品が注目を集めている。研究フェーズの取り組みにおいては、大量の主記憶・メニーコア・SSD・高速ネットワークなどの最新ハードウエアを活用した分散 DBMS の研究が特に欧州の大学で顕著であり、日本においてもディスク装置における非順序型実行原理に基づく超高性能データベースエンジンが開発され、データ分析系の性能ベンチマークにおいて世界最高性能を達成している。

一方、後者の分散処理基盤技術に関しては、米国の Google、Facebook、Twitter 社などの Web 系の企業が技術開発を先導しており、各社の用途に応じて研究開発を進めている。データベースの問い合わせ言語である SQL を高速に処理する基盤技術としては、Google 社がクラウドサービスとして提供している BigQuery や Amazon 社の Redshift がある。BigQuery では、数 1000 台のディスクを同時に利用することで、数億レコードを数秒で検索処理することが可能である。大学においては、主に機械学習処理を高速化する基盤技術に関する研究が進められており、MapReduce の後継となる基盤技術が多く提案されている。例えば、主記憶を活用した Spark や、グラフデータ処理の専用エンジンとして GraphLab が研究開発されている。大規模グラフデータの高速化アルゴリズムについては、香港・韓国・日本などアジア勢が研究をリードしている。また、計算機が低価格化するに従って開発コストが重要視されつつあり、開発コストを削減するため、抽象化言語から最適な実行コー

ドを生成するシステムとして Hive や Flink などが出現している。

今後の研究開発の動向としては、最新ハードウエアを利用した分散 DBMS 系の技術が分散処理基盤技術に融合され、さらなる高速化と低コスト化の観点で技術が発展するとともにデータ構造や応用ごとに専用化された基盤技術が今後も発展するものと考えられる。

時々刻々と生成されるデータを対象とする領域: 時々刻々と生成されるデータを対象とする技術として、一定量の限られた主記憶を利用して高速に検索要求を処理するストリームデータ処理技術がある。ストリームデータ処理は DBMS とは対称的な技術であり、DBMS が事前に登録された過去のデータに対して入力される検索要求を処理するのに対して、ストリーム処理では事前に登録された検索要求に対して入力されるデータストリームを処理する。ストリームデータ処理技術の商用化については、米国および日本の DBMS ベンダから製品が出ており産業での利用が進みつつある。Web 系の企業においても、Twitter 社はストリーム処理エンジンである Storm を開発して公開している。研究フェーズの取り組みについては、蓄積されたデータ処理と時々刻々と生成されるデータの処理を統合的に処理する技術が出現しつつある。またストリームデータを対象とした機械学習エンジンとして日本で開発された Jubatus があり、新たに映像認識の応用向けに Deep Learning の機能開発が開始されている。

今後は M2M 等に代表されるようにストリームを生成するセンサー装置が普及することで、ストリームデータ処理技術の応用が増加して、応用ごとにさらに技術が発展するものと考えられる。

国際比較: 国際比較については、技術の各論については上述した通りであるが、国家施策という観点では、日本では H21-26 年の期間に最高速データベースエンジンの開発をテーマとして、数十億規模の予算を獲得してデータベースエンジン技術および機械学習技術の研究開発を実施してきた。米国・中国・韓国でも H24 年度から具体的な応用先と連携したビッグデータ基盤技術に対して莫大な予算が投資されており、例えば米国では科学技術政策局が、ビッグデータ研究･開発イニシアチブを発表して、ビッグデータ活用に向けて 2 億ドル以上の投資をする戦略を打ち出している [2]。EU におけるビッグデータに関する主な取組としては、Horizon2020（2014 年から 7 年間の計画で総額 770 億ユーロを投じる計画）においてビッグデータの管理・研究のインフラの整備・データのオープンアクセスについて計画している [3]。

産業主導の動向については、Web 系の企業を中心として米国が先行している。欧州では製品開発では SAP 社が強く、システム開発は米国の企業が欧州に進出している。分散 DBMS については欧州の大学と研究機関が世界に先駆けて先進的な成果をあげている。中国では E コマースの企業であるアリババ社がニューヨーク市場で上場して、NY 市場、時価総額 25 兆円の資金を調達したことが注目を集めている。中国・韓国については、ビッグデータ基盤技術の中でもアルゴリズム系の研究開発が強い。

（４）科学技術的・政策的課題

・ビッグデータ基盤技術に関する研究分野では、実データおよび大規模な計算機環境を有している企業が研究開発を進める上で大学より有利な状況にある。欧米ではデータのオープン化が進んでおり、例えばAmazon社の Public Data Sets では50億ページのWebアーカイブやゲノムプロジェクトデータが公開されていて学術の進展に貢献している。対比的に日本はパーソナルデータの議論が途上であることも影響してデータの利用がまだ進んでおらず、大学でビッグデータを利用した研究促進を阻む1要因となっていると考えられる。米国では一部の医療データを病院が公開することを義務づけている州もあり、同様の制度を日本でも行政主導で進めビッグデータが共有できる環境を構築することが重要である。

・ビッグデータ基盤の領域を含めて IT の技術領域では、優れたシステムを開発して容易に世界に公開することが容易であり、実際に日本からもトレジャーデータ社がスタートアップとして起業し、ガートナー社にCool Vendors in Big Data、2014 として認定されるに至っている。しかし、米国と比較して日本では技術主導のスタートアップ企業は少なく、産業をけん引する力が不足している問題がある。この問題を改善するためには、1)学部生レベルの講義においてOSやデータベースなどのシステム技術を設計して実装する能力を育成する講義を増やす、2)大学院レベルの学生向けに海外企業でインターンシップを受けられる支援プログラムを増やす、3)大学に勤めながら企業と兼業がしやすい制度を大学に導入する等が考えられる。

（５）注目動向（新たな知見や新技術の創出、大規模プロジェクトの動向など）

［新たな技術動向］

・大量の主記憶・メニーコア・SSD・高速ネットワーク・GPUやFPGAなどの最新ハードウエアを活用した分散 DBMS やビッグデータ基盤技術が、産業界を含めて注目されている。

・SNS などの多様なサービスが増加しスマートフォン端末が普及することで多様な情報が急速に増加し、人・物・場所といった多様な情報のつながりを表現・分析するのに適したグラフ構造を扱うデータベースと分析アルゴリズムが世界的に研究されている。

・ビッグデータの分析工程を効率化するため、ビッグデータ基盤と分析結果の可視化の連携技術や分析の試行錯誤を自動化する技術などが、近年増加しつつある。

［注目すべき国家プロジェクト］

・米国では科学技術政策局（OSTP）が平成24年3月29日、ビッグデータ研究・開発イニシアチブ（Big Data Research and Development Initiative）を発表して、政府として2億ドルの投資を行い戦略的に取り組む姿勢を明確にしている。中核技術として同イニシアチブでは、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等が挙げられている。

・欧州では2010年5月に策定された「欧州のためのデジタルアジェンダ」において、3億ユーロの予算をかけ、2011年から5年計画のFI-PPP（次世代インターネット官民連携）プログラムが実施されている。FI-PPP プログラムでは、インターネット技術との強い統合を通じ、交通、医療またはエネルギー等の公共サービスのインフラと業務プロセスの競

争力強化と、それらのアプリケーションの出現の支援を目的としている。
・中国では 2011 年 3 月に「国民経済・社会発展第 12 次 5 カ年計画綱要」が採択されており、2011～2015 年の重点技術としてクラウドコンピューティングや Internet of Things などが盛り込まれている。
・韓国では 2009 年 9 月に「IT コリア未来戦略」を発表しており、その中で成長が有望視される 10 サービスを集中育成するとしている。この政策の中には M2M 技術などのストリームデータ処理に関係する項目も含まれている。

（6）キーワード

分散データベース管理システム(分散 DBMS)、ストリームデータ処理、NoSQL、MapReduce、分散処理基盤技術、クラウドコンピューティング

（７）国際比較

国・地域	フェーズ	現状	トレンド	各国の状況、評価の際に参考にした根拠など
日本	基礎研究	○	↗	・非順序型実行原理に基づく超高性能データベースエンジンはベンチマーク結果で世界トップの結果を達成 ・大規模なグラフデータ処理の高速アルゴリズムは先端的である ・分散基盤技術については、分析処理の高速化について一定の取り組みがある。
	応用研究・開発	○	→	・ストリームデータを対象とした機械学習エンジンとして日本で開発された Jubatus があり、映像認識の応用向けにDeep Learning の機能開発が開始された ・ストリームデータ処理については、多くの企業でプロダクトが開発され、検証が実施されている。代表的なものでは、NECや富士通がデータセンターの省電力化や交通渋滞予測などのトライアルを実施したと報告がある。
	産業化	○	→	・大手の情報・通信系企業からクラウドサービスが提供されている ・分散基盤技術についてはWeb系の企業やSI系企業を中心に開発事例が増えている。 ・ストリーム処理については、日立・NEC・富士通は自社開発のストリーム処理エンジンの販売・SI事業を行っている
米国	基礎研究	◎	→	・米国の大学・企業における基礎研究レベルは高く、アルゴリズム系とシステム系の両面で世界をリードしている。 ・分散DBMS・分散処理基盤技術については、高速な分散トランザクション処理、分散処理における処理の最適化（ロードバランス・負荷予測・ワークフロープランの最適実行計画・応用ごとの最適な分散処理）が特徴的である。 ・ストリームデータ処理系については不確実性の導入などのストリーム処理の高度化（CEP的なパターンマッチオペレーターの導入、不確実ストリームへの対応、各種マイニングアルゴリズムの応用など）などがある。
	応用研究・開発	◎	→	・分散処理基盤技術の取り組みについては、Web系の企業がOSS開発を先導しており米国が圧倒している。OSS開発コミュニティでは大学との連携を行っていて、基礎研究成果を取り込む土壌も備えている強みもある。 ・ストリームデータ処理についても、企業が中心となりOSS版のストリーム処理エンジンが開発されている。代表的なプロダクトとしてTwitter社(Storm)、Yahoo(S4)、EsperTech(Esper)、Streambase、HStreaming、Drools Fusionなどがある。
	産業化	◎	↗	・クラウドサービスについて、特にAmazon社が世界市場で活躍している。Google は超高速なSQLサービスであるBigQueryを投入して巻き返しを図っている。 ・分散処理基盤技術の取り組みについては、応用研究・開発と同様に米国が優位な立場にある。 ・ストリーム処理については、Microsoft社(StreamInsight)やIBM社(InforSphere streams)、Oracle社、Progress社(Apama)、TIBCO社ではストリーム処理エンジンの販売や、それらを利用したSI事業を展開している。

欧州	基礎研究	◎	↗	・大量の主記憶・メニーコア・SSD・高速ネットワーク・GPUやFPGAなどの最新ハードウエアを活用した分散DBMSが産業界を含め注目されている。 ・分散処理基盤技術については、分散処理の最適化（データ配置・インデックス構築・テーブル結合最適化）の取り組みがある。 ・ストリームデータ処理については処理モデルや高性能化の研究で成果を挙げている。特にFPGAや分散並列環境を用いた高性能化についての取り組みや、地理情報のストリーム処理などが特徴的である。米国と共同で研究開発を行う場合もあり、ストリーム処理エンジンの開発において多くの先導的な成果を生み出している。
	応用研究・開発	△	→	・米国やアジアに比較して商業規模が小さい欧州では、目立った活動はない。
	産業化	○	→	・SAPのHANAなどのカラム指向インメモリDBMSやストリームデータ処理エンジンの取り組みが目立っている。
中国	基礎研究	◎	→	・大学が中心となって基礎研究において多く成果を挙げている。難関国際会議への採録も米国・欧州に次いで3番目に多い。グラフデータ処理などのアルゴリズム系が中心であり、分散処理基盤技術の高速化の取り組みもある。
	応用研究・開発	○	→	・IBM中国研究所ではハイブリッドクラウドやクラウド間でのマイグレーションの取り組みがある。
	産業化	○	↗	・クラウドサービスについては、Baiduなど国内企業の他、日本企業も各社中国に進出している。 ・Baidu社でNoSQLを利用した開発事例があり、Eコマースの企業であるアリババ社がニューヨーク市場で上場して、NY市場、時価総額25兆円の資金を調達した
韓国	基礎研究	○	→	・大学が中心となって基礎研究において成果を挙げている。フラッシュメモリを利用したDBMSの高速化などの超高速DBMSの取り組みが盛んである。RFIDやセンサーネットワーク上におけるストリームデータ処理の高性能化、高度化に関する研究もある。
	応用研究・開発	△	→	・ビッグデータ基盤技術に関しては目立った活動はない。
	産業化	○	→	・Amazon に続いて、新たにGoogle、Microsoft が相次いで韓国のクラウドサービス市場に参入し、国内企業としては LG CNS 社がデータセンター等を運用している。

（註 1）フェーズ
基礎研究フェーズ ：大学・国研などでの基礎研究のレベル
応用研究・開発フェーズ ：研究・技術開発（プロトタイプの開発含む）のレベル
産業化フェーズ ：量産技術・製品展開力のレベル

（註 2）現状
※我が国の現状を基準にした相対評価ではなく、絶対評価である。
◎：他国に比べて顕著な活動・成果が見えている、 ○：ある程度の活動・成果が見えている、
△：他国に比べて顕著な活動・成果が見えていない、×：特筆すべき活動・成果が見えていない

（註 3）トレンド
↗ ：上昇傾向、 →：現状維持、 ↘：下降傾向

（８）引用資料

1) The Datacenter as a Computer: An Introduction to the Design of Warehouse-Scale Machines, Second edition, Luiz André Barroso (Google), Jimmy Clidaras (Google), Urs Hölzle (Google), Morgan & Craypool Publishers.
2) 総務省 H25 年度版 情報通信白書, http://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h25/pdf
3) Horizon 2020 and the Big Data issue http://www.scienceonthenet.eu/content/article/giacomo-destro/horizon-2020-and-big-data-issue/january-2014

３.10.２　ビッグデータ解析技術

（１）研究開発領域名

ビッグデータ解析技術

（２）研究開発領域の簡潔な説明

データの背後に潜む規則性や特異性を学習することにより、人間と同程度あるいはそれ以上の学習能力の実現を目指す。大量、高次元、多様なデータからいかに効率よく学習を行うかが重要な研究課題であり、理論的な性能保証の追求、および、実用的なアルゴリズム開発の両面から、当該分野は非常に盛んに研究されている。

（３）研究開発領域の詳細な説明と国内外の動向

機械学習(Machine Learning)とよばれるデータからの学習に関する研究は、古くは人間の脳の学習機能をコンピュータで実現しようという 1950 年代の研究に遡る。パーセプトロン(Perceptron) [1]とよばれる単純なニューラルネットワークモデルが提案され、任意の線形分離関数を学習できることから 1960 年代に一大ブームを巻き起こした。しかし、単純なパーセプトロンでは XOR のような入り組んだ関数を学習できないことが明らかになり [2]、1970 年代には関連する研究は下火になった。

しかし、1980 年代に入って、バックプロパゲーション(Backpropagation) [3]とよばれる、階層構造を持つニューラルネットワークモデルの勾配学習アルゴリズムが提案され、第２次ニューラルネットワークブームが到来した。ニューラルネットワークは、画像や音声の認識、ロボットの制御など、様々な問題に適用され、優れた性能を示した。しかし一方では、モデルの階層性に起因する非凸性(Non-convexity)のため、学習に非常に時間がかかり、一般に大域的な最適解を求めることができないという弱点が明らかになった。実際、複雑なニューラルネットワークをうまく学習するためには、学習パラメータの初期値をうまくチューニングするためのノウハウや、学習を高速に実行するための様々なヒューリスティックが必要であり、信頼性の高い学習システムを構築することは極めて困難であった。

そこで 1990 年代に入り、階層性を持たないサポートベクトルマシン(Support Vector Machine)とよばれるカーネル学習器が提案された [4]。サポートベクトルマシンの学習は凸最適化として定式化されるため、容易に大域的な最適解を求める事ができる。その後、大規模データに対する超高速学習アルゴリズムなどの開発を経て、21 世紀初頭にはカーネル法を中心とする機械学習の一大ブームが到来した。カーネル法は音声・画像・自然言語などの処理や、生命情報の解析など、様々な分野で優れた性能を発揮した。しかし、カーネル法の性能を十分に引き出すためには、特徴抽出に対応するカーネル関数をうまく設計する必要があり、さらなる学習性能の向上のために、特徴抽出も自動化したいというニーズが高まってきた。その後、カーネル関数の学習法が盛んに研究されているが、いまだ決定打に欠ける状況である。

このようなカーネル法の全盛期の 2006 年に、多層ニューラルネットワークの新しい学習アルゴリズムが発表された [5]。これは、教師なし手法によって事前学習(Pre-training)した単層ニューラルネットワークを順々に積み上げていく(Layerwise Training)ことによって多層ニューラルネットワークを構築するという手法であり、特徴抽出を自動化する新たな可能

性を切り開いた。また、この流れとは独立に、カーネル法が全盛だった 1980 年代後半～1990 年代前半にも、畳み込みニューラルネットワーク(Convolutional Neural Network) [6]とよばれる特殊な構造を持つ多層ニューラルネットワークが画像認識の分野で研究されていた。これは、1970 年代後半に提案されていたネオコグニトロン(Neo-cognitoron) [7]とよばれる人間の脳の視覚野をモデル化したニューラルネットワークと本質的に同等である。カーネル法が一層の学習器であるのに対して、ニューラルネットワークは多層構造を持つ。そのことから、これらのニューラルネットワーク技術は、ディープラーニング(Deep Learning)とよばれるようになった。

ディープラーニングの技術は、2010 年代に入ってから画像認識や音声認識のコンテストで既存手法を大幅に上回る世界最高性能を達成するに至った。現在は第 3 次ニューラルネットワークブームの真っただ中であり、ドロップアウト(Dropout) [8]とよばれる特殊な学習技術が提案されるなど、当該分野は大きく発展しつつある。

国際的には、現在のディープラーニングブームの火付け役であるトロント大学の Geoffrey Hinton 教授、モントリオール大学の Yoshua Bengio 教授、ニューヨーク大学の Yann LeCun 教授、スタンフォード大学の Andrew Ng 教授らなどの北米の大学教員がディープラーニングの基礎技術開発の中心的な役割を担っている。産業界でも北米の企業がディープラーニング技術の産業応用をリードしている。特に、Ng 教授と Google 社が 2012 年 6 月に発表したディープラーニングによる画像認識の研究成果 [9]は、ニューヨーク・タイムズ紙に取り上げられるなど、世界的な注目を集めた。

また、Hinton 教授らが起こした DNNResearch というベンチャー企業は 2013 年 3 月に Google 社に買収 [10]され、Hinton 教授も Google 社の Distinguished Researcher に就任した。Bengio 教授と LeCun 教授は International Conference on Learning Representations というディープラーニングに特化した国際会議 [11]を 2013 年に開始し、LeCun 教授は 2013 年 12 月に Facebook 社に新設された AI 研究所の所長に就任 [12]した。Baidu 社は 2013 年 1 月に Deep Learning 研究所 [13]を設立し Ng 教授が 2014 年 5 月に Chief Scientist に就任した。Yahoo! 社は 2013 年 10 月に LookFlow という Deep Learning のベンチャー企業を買収し [14] 、Google 社は 2014 年 1 月に Deep Mind というロンドンのディープラーニングのベンチャー企業を買収 [15] した。

一方、ヨーロッパでは北米ほど盛んにディープラーニングが研究されているわけではないが、イギリス、ドイツ、フランス、スイス、イタリア、スペインなどの大学、研究機関にて基礎研究が行われている。特に、スイスの IDSIA 人工知能研究所の Jürgen Schmidhuber 教授のグループ [16]が活発に学術研究を行っている。アジアでは、中国の Baidu 社がディープラーニングの研究開発をリードしているが、大学、研究機関やその他の企業では、ではこれまでのところ目立った活動は行われていない。

国内では、2013 年 11 月の情報論的学習理論ワークショップ [17]にてディープラーニングに特化した招待セッションが組まれ、400 名もの参加者を集めるなど、ディープラーニングに対する注目が高まりつつある。産業界でもディープラーニング技術の活用が進みつつあり、ＮＴＴドコモ社が音声認識サービス「しゃべってコンシェル」[18]にディープラーニング技術を用いた。

（４）科学技術的・政策的課題

ディープラーニングはすでに音声や画像の認識の分野で従来法を寄せ付けない圧倒的な性能を達成しているが、これらの成果は、様々なノウハウやヒューリスティックの積み上げ、および、大量のデータと計算リソースによって実現されている。学習の性能を保証する理論の構築、および、ビッグデータを効率よく処理するための超高速学習技術の開発が、ディープラーニングの科学技術的な重要研究課題である。

カーネル法に代表される従来の凸最適化学習手法では、最適解の一意性が保証されることを用いて、様々な統計的理論解析が行われてきた[19]。一方、ディープラーニングは非凸最適化学習手法であり、そもそも得られた解が何らかの誤差を最小にしている保証すらない。そのため、既存の統計解析手法が適用できず、新しい数学的な道具立ての登場が望まれている。

高速学習技術に関しては、多数の CPU コアを用いた分散学習技術や GPU を用いた並列計算技術が盛んに研究されている[20]。また、ディープラーニングを直接実行するハードウエアを開発しようという研究も行われている[21]。システムレベルとアルゴリズムレベルの研究を融合した高速化技術のさらなる発展が期待されている。

政策的な課題としては、産学官による基礎理論から実世界応用までを広範に含む大型プロジェクトの登場が切望される。国際的には、前述した Google、 Facebook、 Amazon、 Yahoo!、 Baidu などの大企業が、機械学習を専門とする博士学生、ポスドク研究員、さらには、テニュアの大学教員までをも大量に囲い込もうと躍起になっている。

（５）注目動向（新たな知見や新技術の創出、大規模プロジェクトの動向など）

多層ニューラルネットワークのような複雑なモデルでは、学習データへの過適合(Overfitting)が起こりやすい。カーネル法などの従来の学習法では、正則化項(Regularizer)とよばれる罰則項を学習の目的関数に追加することによって過適合を軽減させることができたが、多層ニューラルネットワークではさらに強力な過適合回避技術が必要だと言われている。そのような背景のもと、ドロップアウト(Dropout) [8]とよばれる特殊な学習技術が近年提案された。これは、学習時にニューラルネットワークのユニットをある一定の確率でランダムに取り除くという手法であり、強い正則化効果が得られる。また、ドロップアウト法を用いて学習したニューラルネットワークのテスト時には、出力を定数倍することによって整合性をとるが、これによって多数の微妙に異なるモデルのアンサンブル(Ensemble)を取る効果があり、ディープラーニングのさらなる性能向上につながっていると言われている。また、従来のニューラルネットワークではシグモイド(Sigmoidal Function)関数とよばれる、ヒトのニューロンの動作を模した滑らかな関数を活性化関数として用いていたが、それを修正線形関数(Rectified Linear Function)という簡単な関数に変更することにより、学習性能を低下させること無く、学習スピードを大幅に向上させられることが示された。

大規模プロジェクトの動向としては、上述したように国際的な巨大 IT 企業が巨額の資金をディープラーニング研究に投じている。例えば、Google 社は、社員数わずか 75 名の、産業的な意味で特段の業績を持たない小さなベンチャー企業 Deep Mind 社の買収に、5 億ドル以上も投じたと言われている[15]。

ディープラーニングに関連する最新イベント情報、技術情報、ソフトウエアなどが、deeplearning.net に掲載されている[21]。

（６）キーワード

機械学習、ニューラルネットワーク、ディープラーニング、データマイニング

（７）国際比較

国・地域	フェーズ	現状	トレンド	各国の状況、評価の際に参考にした根拠など
日本	基礎研究	△	→	・機械学習の基礎研究全般に関しては、電子情報通信学会の情報論的学習と機械学習(IBISML)研究会[22]が国内の学術界を主導しているが、これまでのところ、ディープラーニングに関する研究発表はわずかである。
	応用研究・開発	○	↗	・画像処理、音声認識、自然言語処理などを専門とする研究者、エンジニアがディープラーニング技術に対する強い興味を示している。これまでのところ、北米で開発された技術の勉強会や、それらの技術の実証実験などが盛んに行われている。
	産業化	○	↗	・海外の企業の成功を受け、同様の技術を自社に取り込もうと、多くの企業がディープラーニング技術の習得に力を入れている。 ・ＮＴＴドコモが音声認識サービス「しゃべってコンシェル」にてディープラーニング技術を実用化[18]。 ・デンソーアイティーラボラトリが歩行者認識技術を開発[23]。
米国	基礎研究	◎	↗	・元来、機械学習の研究が非常に盛んであり、これまでのディープラーニングに関するほぼすべての技術は、米国とカナダの大学や企業から発信されている。 ・もともとは、Neural Information Processing Systems [24] やInternational Conference on Machine Learning [25]という機械学習全般の国際会議にてディープラーニングの研究情報が発信されていたが、2013年よりInternational Conference on Learning Representations [11]というディープラーニングに特化した国際会議が開始され、今後も北米が中心となってディープラーニングの基礎研究がけん引されていくものと考えられる。
	応用研究・開発	◎	↗	・Facebook社がディープラーニングを用いたほぼ人間レベルの顔認識技術DeepFace [26]を発表。 ・Microsoft社が画像認識技術Project Adam [27]を発表。
	産業化	◎	↗	・Google社が画像検索にディープラーニングを採用[28]。 ・Apple社がSiriの音声認識にディープラーニングを採用[29]。
欧州	基礎研究	○	↗	・イギリス、ドイツ、フランス、スイス、イタリア、スペインなどの大学や研究機関にて機械学習の研究が行われており、ディープラーニングの基礎研究も活発に行われつつある[16]。
	応用研究・開発	○	↗	・Google社、Amazon社などのヨーロッパのブランチが、ディープラーニングの応用研究を行っている。
	産業化	○	↗	・欧州に特化した産業化が行われているわけではないが、北米の企業が欧州にも進出している。
中国	基礎研究	○	↗	・機械学習の主要な国際会議であるInternational Conference on Machine Learning [25]を2014年に北京でホストするなど、研究者人口が爆発的に増加している。北米とのコラボレーションも活発で、アジアの機械学習研究をけん引しつつある。ディープラーニングの研究も盛んになりつつある。
	応用研究・開発	○	↗	・Baidu社にて、ディープラーニングに関する応用研究が活発に行われている[13]。
	産業化	○	↗	・Baidu社を中心として、ディープラーニングの産業化が行われていく模様である。
韓国	基礎研究	×	→	・ソウル大学、KAIST、POSTECHなどの主要大学にて機械学習に関する研究は行われているが、ディープラーニングに関する基礎研究はほとんど行われていない。
	応用研究・開発	△	↗	・Samsung社がディープラーニングに関する研究開発に力を入れつつある[30]。
	産業化	△	→	・今のところ、韓国発のディープラーニング産業は見当たらない。

（註 1）フェーズ
基礎研究フェーズ ：大学・国研などでの基礎研究のレベル
応用研究・開発フェーズ ：研究・技術開発（プロトタイプの開発含む）のレベル
産業化フェーズ ：量産技術・製品展開力のレベル
（註 2）現状
※我が国の現状を基準にした相対評価ではなく、絶対評価である。
◎：他国に比べて顕著な活動・成果が見えている、 ○：ある程度の活動・成果が見えている、
△：他国に比べて顕著な活動・成果が見えていない、×：特筆すべき活動・成果が見えていない
（註 3）トレンド
↗ ：上昇傾向、 →：現状維持、 ↘：下降傾向

（８）引用資料

1) Rosenblatt, F. “The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain”. Psychological Review 65 (6): 386-408, 1958.
2) Minsky, M. & Papert, S. Perceptron, Cambridge, MA: MIT Press. 1969
3) Rumelhart, D. E., Hinton, G. E., & Williams, R. J. “Learning representations by back-propagating errors”. Nature 323 (6088): 533–536, 1986.
4) Boser, B. E., Guyon, I. M., & Vapnik, V. N. “A training algorithm for optimal margin classifiers”. In Proceedings of the fifth annual workshop on Computational learning theory, 144-152, 1992.
5) Hinton, G. E. and Salakhutdinov, R. R. “Reducing the dimensionality of data with neural networks”. Science, 313(5786), 504-507, 2006.
6) LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. “Backpropagation applied to handwritten zip code recognition”. Neural Computation, 1(4):541-551, 1989.
7) Fukushima, K. “Neocognitron: A Self-organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position”. Biological Cybernetics 36 (4): 193–202, 1980.
8) Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R., “Dropout: A Simple Way to Prevent Neural Networks from Overfitting”, Journal of Machine Learning Research, 15, 1929-1958, 2014.
9) RBB Today, 2012 年 6 月 27 日号,
http://www.rbbtoday.com/article/2012/06/27/90985.html
10) Cnet Japan, 2013 年 3 月 13 日号,
http://japan.cnet.com/news/business/35029415/
11) International Conference on Learning Representations 2013,
https://sites.google.com/site/representationlearning2013/
12) IT media ニュース, 2013 年 12 月 10 日号,
http://www.itmedia.co.jp/news/articles/1312/10/news076.html
13) Wired, 2013 年 4 月 16 日号,
http://wired.jp/2013/04/16/baidu-research-lab/
14) IT Pro, 2013 年 10 月 24 日号,
http://itpro.nikkeibp.co.jp/article/NEWS/20131024/513342/

15) Techcrunch, 2014 年 1 月 28 日,
http://jp.techcrunch.com/2014/01/28/20140127why-google-bought-deepmind/
16) IDSIA 人工知能研究所, Prof. Jürgen Schmidhuber,
http://www.idsia.ch/~juergen/
17) 第 16 回情報論的学習理論ワークショップ(IBIS2013),
http://ibisml.org/ibis2013/
18) 朝日新聞 DIGITAL, ２０１４年６月７日,
http://www.asahi.com/articles/ASG666TGTG66UHBI02G.html
19) Vapnik, V. N., Statistical Learning Theory, Wiley, 1998.
20) ディープラーニングに関する情報サイト,
http://deeplearning.net/
21) Techcrunch, 2014 年 4 月 22 日号,
http://jp.techcrunch.com/2014/04/22/20140421nervana/
22) 電子情報通信学会　情報論的学習理論と機械学習研究会
http://ibisml.org/
23) Car Watch, 2014 年 4 月 17 日号,
http://car.watch.impress.co.jp/docs/news/20140417_644713.html
24) Neural Information Processing Systems,
http://nips.cc/
25) International Conference on Machine Learning,
http://machinelearning.org/icml.html
26) ニューズウィーク日本版, 2014 年 3 月 31 日,
http://www.newsweekjapan.jp/stories/business/2014/03/post-3230.php
27) IT media ニュース, 2014 年 7 月 15 日,
http://www.itmedia.co.jp/news/articles/1407/15/news039.html
28) NTT データ 技術＆レポート,2013 年 11 月 7 日
http://www.nttdata.com/jp/ja/insights/trend_keyword/2013110701.html
29)日本経済新聞 2014 年８月 24 日号
http://www.nikkei.com/money/invest-
ment/toushiyougo.aspx?g=DGXLASFZ21H13_25082014000000
30) サムソン社ホームページ,
http://job.samsung.ru/MainPage/Vacations/Software-Engineer-for-Object-Recognition-by-Us-
ing-.aspx

3.10.3　クラウドソーシング

（1）研究開発領域名

クラウドソーシング

（2）研究開発領域の簡潔な説明

クラウドソーシングを利用して不特定多数の人間の力を合わせて問題解決を行うという考え方が広く受け入れられつつある。 計算機科学の技術はその実現において本質的な役割を果たす一方、現在の計算機科学において解決困難な課題を人間の力をかりて解決するヒューマンコンピュテーションのアプローチも注目されている。ビッグデータ領域においてクラウドソーシングの利用はデータ収集や解析など様々な場面で活躍が見込まれる。

（3）研究開発領域の詳細な説明と国内外の動向

クラウドソーシング [1]は「(インターネットを通じて）不特定多数の人に仕事を依頼すること、もしくはその仕組み」を意味する比較的新しい考え方であり、米 Wired 誌の寄稿編集者ジェフ・ハウ氏によって名づけられた。この言葉は企業内の業務の一部を外部に委託する「アウトソーシング」に由来したものであり、これが素性の知れた特定の相手に対して仕事を依頼するのに対して、クラウドソーシングは不特定多数の相手に依頼する点が特徴である。クラウドソーシングの実施形態には報酬の有無や雇用方式の違いなど様々なバリエーションがあり、実施の目的もビジネス、科学、社会への貢献など多様である。

自社での解決が困難な課題を広く一般に公開し革新的な解決策を募る InnoCentive や、oDesk、 Amazon Mechanical Turk 等のオンライン労働力市場の出現により、クラウドソーシングはデザイン・翻訳・システム開発などあらゆる分野に適用され、様々なビジネスが生み出されている。仕事の発注側にとっては必要に応じた労働力調達の手段として、働き手にとっては場所や時間に囚われない新しい働き方として急速に拡大しつつあり、労働市場のグローバル化のけん引力としてその動向に注目が集まっている。一方、公共の目的をもったボランティアベースのものも多く登場しており、NASA や DARPA 等の公的機関もクラウドソーシングを積極的に活用している。国内でも先の震災では安否情報や医療情報の構造化にその力を発揮した。Wikipedia もまた不特定多数の人間がその編集に関わるという意味で、公共の目的をもったクラウドソーシングであるとみることができる。

情報処理技術分野、特に知能関連分野においてクラウドソーシングの利用は増加している。例えば自然言語処理における言語理解や翻訳、コンピュータービジョンにおける画像理解、データベースにおける SQL 評価や参照解決などのように完全な解決が困難な問題において、人間の助けを借りてこれを行うアプローチがしばしばとられる。これらは人間を計算資源の一部として用いることで、コンピューターと人間の両方を適切に組み合わせて問題解決をはかることからヒューマンコンピュテーション [2]と呼ばれており、従来の計算機を中心とするパラダイムに変革を起こしつつある。ヒューマンコンピュテーションでは人間に繰り返し作業を依頼することによって計算を進めることになるが、その労働力の供給源としてクラウドソーシングは有力な手段となる。

ヒューマンインターフェースや検索などのシステムの性能評価においてもクラウドソーシングは用いられる。従来は少数の人間による小規模な評価が主流であったものが、クラウド

ソーシングを用いることで大規模な評価が可能になった。同じことはアンケート等のサーベイにも当てはまる。

その他のクラウドソーシングの利用方法として非常に重要なもののひとつがデータ収集である。データ駆動システムの構築にあたり最も重要なことのひとつがデータ量の確保である。効果的で効率的なデータ収集のために人間の知覚能力や判断をデータ収集に用いるクラウドセンシングあるいは参加型センシング [3]と呼ばれるアプローチがとられている。一方、機械学習などのデータ解析技術の利用にあたり大きなボトルネックとなるのがメタデータの収集、特に機械学習アルゴリズムに与える正解データの収集である。データに対して意味を付与するメタデータの獲得には多くの場合人間の手が必要であり、高い金銭的・時間的コストが生じる。クラウドソーシングを利用することでそのコストを低減させることが可能となる。クラウドソーシングによって収集されたデータ・メタデータは質のばらつきが大きいため、これをいかに抑えて有効な解析やモデル構築を行うかは現在盛んに研究が進んでいる分野のひとつである [4]。

データ解析そのものにおいてもクラウドソーシングは利用される。最も典型的な例はいわゆるデータ解析コンペティションと呼ばれるものであり、共通のデータを用いて参加者がその予測精度を競うものである。コンペティションの中には解析したいデータを提供するスポンサーによって勝者に対して賞金が支払われるものもあるが、最近ではこのような賞金付きのコンペティションを組織的に運営する「データサイエンティストのクラウドソーシング」が出現している。最も良く知られたものが Kaggle[5]であり、世界中から 10 万人以上が参加している。コンペティションの形式をとることにより少数精鋭でデータ解析に取り組むよりも、短期間で高精度のモデルが得られることが実証されている。

2012 年に米国政府が打ち出したビッグデータ研究開発イニシアチブの中では、クラウドソーシングは機械学習やクラウドコンピューティングと並び注力すべき情報技術分野としてその名が挙がっており、今後ますますその重要性は高くなると思われる。

（4）科学技術的・政策的課題

クラウドソーシングにおける重要な技術的課題 [6]のひとつは、クラウドソーシングの駆動力が不確定要素の大きい人間であることに起因する品質と効率の保証の問題である。クラウドソーシングで得られる成果物の品質は、仕事を行う人間の能力ややる気に依存して大きく左右されるため「同一のタスクを複数の働き手に割り当て多数決をとる」冗長化による品質保証などの仕組みが必要であり、統計的手法を用いた品質保証や非均一な質のデータの解析手法は重要な課題である。例えば、クラウドソーシングでやり取りされる仕事の多くを占めるのが文章（記事作成や翻訳）、デザイン、プログラムなどの複雑な構造あるいは非定型の成果物を求めるものであるが、これら非定型成果物を伴うクラウドソーシングの品質保証は重要な未解決課題である。一方、結果の処理という受け身の対応にとどまらず、ワーカーから情報を正しく引き出し、あるいはタスクを適切に実施させるための仕組みを設計するメカニズム設計も重要な課題である。

多数の人的資源に同時にアクセスできるクラウドソーシングであるが、そのキャパシティーは無尽蔵でなく、必要な時に必要な質をもった必要な量の労働力が安定して得られる保証はない。また、人間の処理速度は機械のそれと比して遥かに遅いため、可能な限り自動化す

るのが望ましい。人的資源の動的な調達、タスクの最適な割り当て、作業フローの状況に応じた制御、可能な部分については機械学習による置き換えを図る等、積極的な効率化の方策が必要である。

別の観点からは、クラウドソーシングを利用してデータ収集を行う際、企業や個人が積極的にデータ供出するインセンティブを与えることが重要である。例えば、企業においてクラウドソーシングを自社の主要ビジネスにおいて利用することを、あるいは医療や健康の向上といった目的に利用するといったことを考えた時、実用化への大きな壁となるのがセキュリティーとプライバシーの問題である。クラウドソーシングを通じて人に作業を依頼する際には、具体的な依頼内容をクラウドソーシングワーカーに対して開示する必要があるが、これはすなわち（しばしば不特定多数の）ワーカーに対してタスク情報を公開することになり、タスク情報を通じた情報漏えいリスクが懸念される。このことが企業・行政がクラウドソーシングの本格的な利用に躊躇する一因となっている。クラウドソーシングにおけるセキュリティーやプライバシーを扱った研究は少なく、今後一層の研究開発が必要となるだろう。

クラウドソーシングは比較的新しい考え方・ビジネスモデルであるため、必ずしも現在の倫理的・社会的規範に沿っているとは言えない部分も多い。例えばしばしば指摘されるように、マイクロタスク型のクラウドソーシングでは経済的格差を利用してワーカーが極めて低賃金での作業を強いられたり、報酬が支払われなかったりなどの問題がある。技術的な課題解決と同時に制度的な整備も必要となるであろう。

（５）注目動向（新たな知見や新技術の創出、大規模プロジェクトの動向など）

クラウドソーシングとこれを利用した計算パラダイムであるヒューマンコンピュテーション分野は近年急激な成長をとげる新しい研究領域であり、これらの研究を対象としたコミュニティが形成されつつある。特に 2013 年からはクラウドソーシングとヒューマンコンピュテーションをテーマとした国際会議である HCOMP[7]が開催されており、本研究分野の中心的コミュニティとしての役割を果たしている。

情報技術関連企業の研究所でもこの分野の研究に力を入れる所は多いが、特に Microsoft Research はこの分野に注力しており、毎年多くの重要な研究成果を発表している。

ヒューマンコンピュテーションの概念を唱えた CMU の von Ahn 氏は、クラウドソーシングを言語学習に利用する DuoLingo[8]プロジェクトを開始し、その今後の動向が注目されている。

（６）キーワード

クラウドソーシング、ヒューマンコンピュテーション、品質管理、メカニズムデザイン、セキュリティー、プライバシー、データ解析コンペティション、

（７）国際比較

国・地域	フェーズ	現状	トレンド	各国の状況、評価の際に参考にした根拠など
日本	基礎研究	○	↗	・米国に後れをとる形であるが、国内での研究コミュニティ[9]が育ってきており、トップレベルの研究も生まれている。
	応用研究・開発	○	↗	・企業における応用研究も増えている。大学と企業の結びつきも強まりつつある。
	産業化	○	↗	・国内クラウドソーシング企業は成長しており、研究成果の産業利用も見られる。
米国	基礎研究	◎	↗	・MechanicalTurk等のプラットフォームを利用したクラウドソーシング研究が初期から盛んである。ヒューマンコンピュテーションの考え方も米国で誕生したこともあり、分野をけん引している。
	応用研究・開発	◎	↗	・大学と産業界の距離が近く、両者が密接に連携をとりながら、様々な先進的試みを行っている。
	産業化	◎	↗	・産業的な問題意識が研究に直接結びついている。von Ahn氏に代表されるように、研究者自身がビジネスを手掛けるケースもある。
欧州	基礎研究	○	↗	・欧州独自の目立った動きは見られないが、Microsoft Research Cambridgeを中心に多くの研究成果がある。
	応用研究・開発	△	↗	・欧州独自の目立った動きは見られないが、トレンドとしては上向き。
	産業化	△	↗	・欧州独自の目立った動きは見られないが、トレンドとしては上向き。
中国	基礎研究	○	↗	・Microsoft Research Asiaや有力大学から着実に研究成果が出ている。
	応用研究・開発	△	↗	・特に目立った動きはないが、早晩頭角を現すと予想される。
	産業化	△	↗	・特に目立った動きはないが、早晩頭角を現すと予想される。
韓国	基礎研究	△	→	・特に目立った動きはない。
	応用研究・開発	△	→	・特に目立った動きはない。
	産業化	△	→	・特に目立った動きはない。

（註 1）フェーズ
基礎研究フェーズ ：大学・国研などでの基礎研究のレベル
応用研究・開発フェーズ ：研究・技術開発（プロトタイプの開発含む）のレベル
産業化フェーズ ：量産技術・製品展開力のレベル
（註 2）現状
※我が国の現状を基準にした相対評価ではなく、絶対評価である。
◎：他国に比べて顕著な活動・成果が見えている、 ○：ある程度の活動・成果が見えている、
△：他国に比べて顕著な活動・成果が見えていない、×：特筆すべき活動・成果が見えていない
（註 3）トレンド
↗ ：上昇傾向、 →：現状維持、 ↘：下降傾向

（８）引用資料

1) Howe, J. The Rise of Crowdsourcing. Wired Magazine, 2006.
2) Law, E. & Von Ahn, L. Human Computation.Morgan & Claypool Publishers, 2011.
3) Ganti, R. K., Ye, F., & Lei, H. Mobile crowdsensing: current state and future challenges. IEEE Communications Magazine, 49(11):32-39. 2011
4) 鹿島久嗣, 梶野洸. クラウドソーシングと機械学習. 人工知能学会誌, 27(4): 381-388, 2012.
5) Kaggle.
http://www.kaggle.com/
6) Kittur, A. et al. The future of crowd work. In Proceedings of the 2013 Conference on Computer Supported Cooperative Work (CSCW), 2013.
7) Conference on Human Computation & Crowdsourcing (HCOMP)
http://www.humancomputation.com
8) DuoLingo.
https://www.duolingo.com/
9) クラウドソーシング研究会.
https://sites.google.com/site/crowdsourcingresearch/

３.10.４　プライバシー保持マイニング技術

（１）研究開発領域名

プライバシー保持マイニング技術

（２）研究開発領域の簡潔な説明

個人データ活用において求められるプライバシー保護技術の研究開発

（３）研究開発領域の詳細な説明と国内外の動向

［背景と意義］

様々なオンラインサービスの発展とともに、個人の生活や行動にまつわる情報が収集されつつある。大規模・高解像・多様な個人に関わるこのような情報は、ビッグデータ利活用の中心的な課題の一つであるが、このような個人情報の利用においてはプライバシー保護上の配慮が欠かせない。個人情報・パーソナル情報の利活用と保護の問題は、我が国の法制度、諸外国との法制度との整合性、社会倫理、産業応用上の価値など、技術上の問題にとどまらない様々な側面を総合的に考慮する必要がある。

ビッグデータ時代を迎えた現在では、プライバシーの概念は二面的に捉えられる。個人情報を継続的に取得する国家や企業側からは、個人情報の流通と深い活用を望んでいる。一方で、情報を取得される個人は、詳細な個人情報の無制限の流通や利用に関する不安感を強めつつある。個人に関する情報の利活用と保護は、本質的に矛盾した要請であり、両立は困難である。データプライバシー保護技術は、この矛盾した要請をトレードオフの関係として捉え、理論的・技術的に健全な「落としどころ」を確立する手段と理解することができる。データプライバシー保護は、「個人データ利活用のシナリオ」と「個人データ保護の方法」を両輪として同時に考察する必要がある。

［これまでの取り組み］

(データ匿名化)

個人データの収集においては、目的とする個人データ(例：購買履歴、Web 閲覧履歴、移動履歴)とともに、その個人に関する情報(例：年齢、性別、居住地、職業)などが同時に収集されることが多い。このような個人データの収集者が保持する個人データの集合が、収集者以外の第三者に提供された際に、第三者がある個人データを特定の個人のデータと結びつけることを特定と呼ぶ。個人データの集合が第三者に適用されたときに、このような特定が発生する確率をプライバシー上のリスクと捉え、個人データの集合を第三者に提供する前に適切にデータ内容を変更することによってこのような特定が発生する確率を低めるk-匿名化の概念が 2002 年に Sweeney によって提案された [1]。

k-匿名性とはデータの正確性を犠牲にして個人データを改変し特定を困難にする手法である。具体的には、 表形式データについては、個人データの属性値の組み合わせが同じであるデータが、個人データ集合中に少なくともk(> 1)個存在している状態として定義される。

アカデミアでは、2000 年以降、主にデータ工学やデータマイニング分野においてデータ匿名化について多くの研究が発表された。具体的には、先に述べたk-匿名性を基礎概念として、匿名化対象を表形式データからグラフや時系列データに拡張する研究や、k-匿名性モデ

ルにおいて十分にプライバシーを保護できない状況下におけるより強力な匿名性定義の研究などが提案された。ただし、2014 年現在ではトップ国際会議においては匿名化技術についての発表はほとんど見られず、データ匿名化の研究がアカデミアにおいて活発に行われているとは言えない。k-匿名性をベースとしたプライバシーモデルにおいては、特定を試みる者(攻撃者)が持つ背景知識に依存して、攻撃者に漏えいする情報が変化することから、特定のリスクに関して一般的な理論的保証を与えることが困難であることがその理由の一つである。一方、k-匿名性の概念は技術者以外にも直感的な理解が容易であることから、我が国における個人情報保護法改正に向けたデータプライバシー保護における技術的検討においても k-匿名性の概念が広く検討されたことは注目に値する。

(差分プライバシー)

個人情報や機密情報を含むデータベースを利活用するための一方式として、データベース問い合わせにおけるプライバシー保護についての理論とモデルが整備されつつある。差分プライバシーとは、データベース問い合わせとその応答に関わる情報漏えいを理論的に評価し、その対応策を与える技術である。

差分プライバシーが想定する状況は以下の通りである。データ収集者は、多数の個人に関するデータをデータベースに保存している。また収集者以外の第三者が、データ収集者に、そのデータベースの内容について統計的なクエリ(例：ある数値属性に関する平均値や、ある離散属性値の組み合わせにマッチするレコード数のカウント)を問い合わせ、その応答値を得る。差分プライバシーでは、「その応答値から、問い合わせ対象としたデータベースに、ある個人が含まれているか否かが判定可能である」ことをリスクと捉え、このリスクを低めるために、統計値を雑音によって摂動させ、開示するアプローチを取る。より具体的には、差分プライバシーの理論は与えられた母集団数と対象クエリについて、その応答値に加えるべき雑音の分散の上限を理論的に与える。

差分プライバシーの問題はデータベースに対するクエリとその応答で定式化されるが、必ずしもデータベースとの対話的操作を介しなくとも、個人に関する情報についての統計値の開示におけるプライバシー保護の問題は、差分プライバシーの問題として解釈することができる。例えば、ある地方自治体に在住する住人の平均所得や、所得の頻度分布などの統計量の開示などが該当する。

一般に、統計的クエリの応答値の開示は個人情報の開示とは見なされにくい。例えば、母集団数十万人を対象とした平均所得が特定の個人の所得の開示を引き起こすことはないであろうと考えられるためである。この直感は、母集団数が大きい場合には正しいと言えるが、「X 町 Y 丁目在住の 60 代独身女性の平均所得」といったように、母集団数が少ない場合の統計値の開示(例えば母集団数 5 人)は、個別の個人の所得の推測をある程度可能にすることから、その公開にはプライバシー上の配慮が必要であろう。差分プライバシーは、このような統計量の開示におけるプライバシー上のリスクを、母集団数や対象とする統計的クエリについて、定量化することができる。

差分プライバシーの研究は 2006 年の Dwork らの研究 [2)]に端を発し、以降、理論計算科学、暗号理論、データ工学、機械学習等、多数の理論分野にわたり、主に理論的アプローチの下で複合的に発展しつつある。現在では、上記複数分野において、分野を代表する一流の

国際会議において、コンスタントに数本ずつ研究発表が見られ、徐々に増加している。差分プライバシーのモデルは、暗号理論における標準的な安全性定義として用いられている「識別不可能性」の議論を下敷きとしていることから、従来から積み上げられてきた理論的な安全性解析との親和性が高い。また、その安全性が攻撃者(差分プライバシーにおいてはクエリを発行する者)の背景知識によらないことから、k-匿名性等に比べより現実的な状況下における安全性を保証できる。ただし、差分プライバシーを達成するためには、クエリ応答値に非常に分散の大きい確率分布から得た雑音による摂動を与えることが必要な場合があることから、実用に耐えないケースも多く、現実的な状況において差分プライバシーの保証と実用性のバランスを取るための研究上の試みが現在も続いている。

(秘密計算)

秘密計算とは複数の者が互いに相手と共有することができない秘密の入力を所持しているときに、その秘密入力を互いに他者に開示することなく、これらの秘密データを入力に取る関数を評価し、その出力をいずれかの者あるいは第三者のみに報告する分散計算の総称である。

秘密計算はこれまでに大きく分けて、秘匿回路評価、準同型性暗号系による暗号プロトコル、秘密分散に基づく秘密計算、の三種類の実現方式が知られており、それぞれの方式には計算効率性や計算モデルの自由度の観点において、長所と短所がある。

秘匿回路評価(Yao's garbled circuit)とはブーリアン回路の秘密計算を構成要素とした秘密計算の技法であり、1986 年に Yao によりその基礎が築かれた [3]。

秘匿回路評価は、 紛失送信とよばれる暗号理論上の要素技術が利用可能であることを前提として設計されている。紛失通信とは、送信者が複数項目からなるリストを保持し、受信者がそのリストのインデクスを保持しているときに、送信者に受信者が持つインデクスを与えることなく、受信者が送信者からリスト中のそのインデクスに対応する値を得るプロトコルである。秘匿回路評価は、論理回路で記述可能な任意の関数を秘密計算として実装することが可能であり、汎用性の高い手法である。大規模な入力を引数に取る関数の評価には膨大な時間がかかるため、データマイニングのようなデータ解析の用途には不向きであると考えられていたが、近年は高速実装の技術が進み、文字列操作等の離散的な処理については実用性が高まりつつある。秘密計算の歴史は長く、 Yao による発表からすでに四半世紀が経つが、主に暗号理論・システムセキュリティー分野の国際会議ではコンスタントに研究発表が続いている。秘匿回路評価の実装系は米国・イスラエルなどの大学によるものが著名である [4][5]。

準同型性暗号系とは、二つの整数値の平文に対応する二つの暗号値について、これを暗号化したままである二項演算を行うことで、その整数に対して加法や乗法などの算術演算の結果を暗号化した暗号値を、復号なしで(秘密鍵の知識なしに)得ることができる性質を持つ暗号系である。加法あるいは乗法のどちらかに対応した準同型性暗号系は以前から知られていたが、2009 年には Gentry[6]によって加法と乗法の両方について準同型性を同時に有する暗号（完全準同型暗号）が発表された。これがブレイクスルーとなり、いくつかの完全準同型性暗号を実現する方式が続いて発表された。完全準同型暗号は計算時間や空間計算量が膨大

であることが知られており、現在その削減の努力が続いているが、大規模データ解析における実用性は現段階では低く、もう一段のブレイクスルーが必要であるといえる。これに付随し、加法と乗法の両方をある有限な回数を上限として許す準同型性暗号系を somewhat 準同型暗号とよび、時間計算量や空間計算量において、より実用性の高い方式の研究開発も近年増えている。準同型性暗号系に基づく秘密計算は、算術演算を基に構成されるため、秘匿回路評価との比較でいえば、数値を入力に取るデータ解析や、行列演算をベースとするアルゴリズムに適しているといえる。

準同形暗号に基づく秘密計算は主に 2000 年代からセキュリティー分野、データ工学分野、データマイニング分野等で研究されたが、Gentry らによる完全準同型性暗号の考案が発端となって、完全準同型性暗号自体の開発が現在暗号理論分野で非常に活発に行われ、今後、そのデータ解析応用が増加すると考えられる。

暗号系を用いない秘密計算手法として、秘密分散に基づく秘密計算が知られている [7]。秘密分散では、秘密情報を複数のシェアと呼ばれる情報の断片に分散させる。単独のシェアからは秘密情報を復元することはできないが、複数のシェアを集めることで秘密情報を復元することができる。秘密分散に基づく秘密計算は、複数の秘密情報をランダムシェアとして複数の者に分散させ、ランダムシェア同士の演算によって、所望の計算を行い、これらの秘密情報を入力とするあるクラスの計算は、ランダムシェア同士の演算から実現可能であることを利用して秘密計算を実現する。この方式は、暗号演算や暗号プロトコルの実行を含まないため、一般に計算効率的である。ただし実現可能な関数のクラスは限定される。また、秘密分散を基礎とする方式であることから、3 以上の互いに独立な計算主体の存在を仮定する必要がある。このことは、秘密計算をシステムとして実装した場合には、安全性が保証された多数の計算資源の運用が要求されることを意味している。

秘密分散に基づく秘密計算は、近年の暗号理論の国際会議などにおけるプレゼンスは高くないが、エストニアの企業から開発環境が提供されており、実用化を目指した動きもみられる [8]。

［今後必要となる取り組み］

(高機能暗号)

関数型暗号や完全準同型性暗号などの高機能暗号の基礎研究は、近年発展の速度が速く、今後も継続した取り組みが必要である。高機能暗号に関連する研究は理論の構築に比して実装技術やライブラリの構築がこれに追いついていないため、その整備が必要である。

(秘密計算)

秘密計算を利用したプライバシーデータの活用については、準同型暗号や Garbled circuit の活用については多くの研究例がすでにあるが、高機能暗号の活用はこれまでにほとんどない。データ解析と暗号理論分野の両方に精通した研究者・技術者が少ないことがその理由の一つであると考えられ、分野間の交流等人材開発が必要である。また「ビッグデータ解析」と「個人データのセキュリティー・プライバシー保護」を個別的な研究開発課題でなく、両者を両輪としたソリューション型の研究開発を進める必要がある。

(差分プライバシー)

差分プライバシーは理論計算科学分野ではすでにデファクトスタンダードのプライバシーモデルとして定着した感がある。実用上、今後普及するか否かは現段階では明確ではないが、この分野に精通した研究者・技術者は諸外国に比べ日本には極めて少なく、少なくとも研究者レベルではこの分野の専門家の養成が必要であろう。

（４）科学技術的・政策的課題

(ソリューション開発)

ビッグデータ活用の機運が高まる中、プライバシー保護の重要性も同様に強く認識されつつある。データ工学やデータマイニング・機械学習に代表される「データ活用技術」および暗号理論・システムセキュリティーに代表される「情報セキュリティー」等の個別の要素技術においては我が国には厚い蓄積があるが、個人データ活用におけるプライバシー保護は、法制度との整合性を維持しつつ両技術を適切に組み合わせる複合的なソリューション開発の技術が必要である。このような複雑な課題に取り組むことができる組織・人材の育成は科学技術・政策上の課題であろう。

(技術の進化に柔軟な法制度整備)

米国はサービス利用者の理解を得つつ個人情報を深く解析し、これをサービスの高度化に活用するとともに利益を生み出す源泉とするサービスのあり方を構築してきた。一方で我が国は、法制度が現実の個人データ活用に追いついていなかったこともあり、そのようなサービスが大きく発展したとは言えない。デバイス・通信技術の発展や、新種のサービスの普及に伴い、プライバシー保護のあるべき姿は急速に変容していくともに、個人情報活用に対する個人の受容度も変化していく。技術の進歩に柔軟に対応できる法制度の整備も重要な課題である。

（５）注目動向（新たな知見や新技術の創出、大規模プロジェクトの動向など）

・(米国のプライバシー保護政策) ホワイトハウスは 2014 年 5 月にビッグデータ解析におけるプライバシー保護・差別の撤廃・自己情報のコントロール等を主題とした報告書”Big Data: Seizing Opportunities、 Preserving Values”を公表した。この報告書は、2012 年２月に公表された オバマ大統領による「消費者プライバシー権利章典」(A Consumer Privacy Bill of Rights) が定める消費者(データ提供者たる個人)の権利保護の方針を明確にしつつも、プライバシーデータの収集・流通・活用が生み出す知識の経済的価値・公共の利益を強く認識し、データに基づくイノベーションの促進を目指す提言となっている。

・(欧州の EU データ保護指令) 2014 年 3 月に EU データ保護指令が欧州議会によって可決された。これには、データ元の個人から削除の請求があった場合にはその削除を義務づける「忘れられる権利」が明文化されている。Google は、スペインにて個人情報を含む Web ページへのリンクを検索結果から削除する義務があるという「忘れられる権利」を認めた判決を受け、リンク削除に応じた。

・(日本の個人情報保護法改正) 高度情報通信ネットワーク社会推進戦略本部は、個人情報

保護法の改正にむけ、2014 年 6 月に、データを取得した個人の同意無しでデータを流通・利活用するための新しい枠組みの創設や、プライバシーの保護のためのデータの加工方法などについて判断を示す第三者機関の設立などを含めた「パーソナルデータの利活用に関する 制度改正大綱」を示した。

・中国では「第１２次五カ年計画」（2011 年～2015 年の五カ年）において、セキュリティーが重要分野として挙げられている
・韓国では 2009 年 9 月に「IT コリア未来戦略」を発表しており、その中で世界最高の情報保護対応センターの構築が挙げられている

（６）キーワード

プライバシー保護、暗号、秘密計算、ビッグデータ解析、データマイニング

（７）国際比較

国・地域	フェーズ	現状	トレンド	各国の状況、評価の際に参考にした根拠など
日本	基礎研究	◎	↗	・暗号理論に関する基礎研究は、企業、大学、国立研究所ともに高いレベルにある。国際的にも競争力があり、優れた成果を挙げている ・プライバシー保護技術・システムセキュリティーについての基礎研究は、トップ国際会議等での発表は数が多くなく、国際的に競争力があるとは言えない
	応用研究・開発	◎	↗	・秘密計算の応用技術については、産総研・筑波大による化合物検索のプロトタイプ構築や、NTTによる疫学の実証実験など、応用研究のアクティビティが高い。
	産業化	○	→	・各種暗号アルゴリズムの国際標準の保有（ISO/IECなど）
米国	基礎研究	◎	↗	・暗号理論分野・データベース分野・データマイニング分野、いずれの研究領域においても、多くの理論的アイデアはほとんど米国の大学・企業の研究者から提案されている(差分プライバシー、完全準同型暗号など)
	応用研究・開発	◎	→	・マサチューセッツ工科大学のCryptDBをはじめとする、暗号のデータベース分野への応用 ・バージニア大の高速秘密計算の開発環境FastGCを初めとする秘密計算研究
	産業化	◎	→	・NISTによる暗号規格の実質的な世界標準化 ・民間企業による高度な個人履歴情報の活用
欧州	基礎研究	○	→	・暗号理論に関する基礎研究は、企業、大学、国立研究所ともに高いレベルにある。国際的にも競争力があり、優れた成果を挙げている ・プライバシー保護技術についての基礎研究は、匿名化技術などにおいては複数名著名な研究者が活動している
	応用研究・開発	○	→	・オランダ統計局による匿名化ツールARGUSの開発
	産業化	○	→	・各種暗号アルゴリズムの国際標準の保有（ISO/IECなど）
中国	基礎研究	○	→	・秘匿回路評価の提案者を排出するなど学者を生み出しているが、活躍の場は米国などが主である
	応用研究・開発	○	↗	・データ工学分野を中心に、複数名のデータプライバシー研究者が活動している
	産業化	◎	↗	・国策として4G携帯電話用の暗号アルゴリズムの国産化を目指し開発と標準化を進めている
韓国	基礎研究	○	→	・各種の暗号アルゴリズムの基礎的な研究を行い、国際標準に提案活動を行っている
	応用研究・開発	△	→	・特に目立った活動は見られない
	産業化	○	→	・各種暗号アルゴリズムの国際標準の保有（ISO/IECなど）

（註 1）フェーズ
基礎研究フェーズ ：大学・国研などでの基礎研究のレベル
応用研究・開発フェーズ ：研究・技術開発（プロトタイプの開発含む）のレベル
産業化フェーズ ：量産技術・製品展開力のレベル

（註 2）現状
※我が国の現状を基準にした相対評価ではなく、絶対評価である。
◎：他国に比べて顕著な活動・成果が見えている、 ○：ある程度の活動・成果が見えている、
△：他国に比べて顕著な活動・成果が見えていない、×：特筆すべき活動・成果が見えていない

（註 3）トレンド
↗ ：上昇傾向、 →：現状維持、 ↘：下降傾向

（８）引用資料

1) Sweeney, Latanya. "k-anonymity: A model for protecting privacy." International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10.05 (2002): 557-570.
2) Dwork, Cynthia, et al. "Calibrating noise to sensitivity in private data analysis." Theory of Cryptography. Springer Berlin Heidelberg, 2006. 265-284.
3) Yao, Andrew Chi-Chih. "How to generate and exchange secrets." Foundations of Computer Science, 1986., 27th Annual Symposium on. IEEE, 1986.
4) Malkhi, Dahlia, et al. "Fairplay-Secure Two-Party Computation System." USENIX Security Symposium. Vol. 4. 2004.
5) Huang, Yan, et al. "Faster Secure Two-Party Computation Using Garbled Circuits." USENIX Security Symposium. Vol. 201. No. 1. 2011.
6) Gentry, Craig. "Fully homomorphic encryption using ideal lattices." STOC. Vol. 9. 2009.
7) Bogdanov, Dan, Sven Laur, and Jan Willemson. "Sharemind: A framework for fast privacy-preserving computations." Computer Security-ESORICS 2008. Springer Berlin Heidelberg, 2008. 192-206.
8) Bogdanov, Dan, Sven Laur, and Jan Willemson. "Sharemind: A framework for fast privacy-preserving computations." Computer Security-ESORICS 2008. Springer Berlin Heidelberg, 2008. 192-206.

3.10.5　IT メディア分野におけるビッグデータ

（1）研究開発領域名

IT メディア分野におけるビッグデータ

（2）研究開発領域の簡潔な説明

IT 技術を用いて蓄積・伝送が可能な膨大なデジタルデータ（数値、テキスト、音声、画像、映像など）に共通する性質を抽出し整理することで、新たな情報を得ることを目的としている。例えば、画像認識、映像解析、メタデータ付与などの基礎研究から、社会分析に至るまで幅広い研究開発がある。

（3）研究開発領域の詳細な説明と国内外の動向

IT メディアの種類は、金融・国勢などの統計情報を主とした数値情報、各種文書などに代表されるテキスト情報、音響・画像・映像などのマルチメディア情報に大別できる。本研究開発領域では、こうした IT メディア情報に対するビッグデータ処理の観点から、ビッグデータ処理に特化した解析技術について述べる。

数値およびテキスト情報については、十分な研究開発が行われ、技術的に成熟している。ただし、テキストについては、Web 情報、特に blog や microblog にみられるような、話し言葉に近い、あるいは短すぎて従来の自然言語処理が機能しないテキストについてはまだ解析が困難である。また、マルチメディア情報については、単一話者・背景無音の音声認識や、文字認識、顔検出・認識など、一部の例外を除いては、その解析は極めて困難である。一方で、blog や microblog は一般市民の意見や日々の生活の情報から社会動向や風俗を、マルチメディア情報は実世界の情報をそのまま表現できるという利点があり、その効果的な解析技術の実現が強く要望されている。しかしこうした情報は、実世界・社会・生活により近いが故に、多様性が高く、解析が困難であり、その研究開発はまだ発展途上にある。現在最も有望視されているのは、実際の解析対象と同等の特性・多様性を有する大規模な研究開発用データセットを構築し、機械学習などの技術により頑健性の高い解析技術を研究開発していく手法である。

国内外の状況としては、IT メディア分野の基礎研究では、NSF や IARPA などからの資金援助、ならびに Google、Microsoft、Facebook、Yahoo!など、IT メディアの対象そのものを大量に有する企業の存在にも支えられた米国が、研究開発の先導ならびに実際の推進の両面で世界をリードしている。欧州では、Horizon 2020 において、EU Open Data portal のさらなる整備と共に、IT メディアを対象とした研究開発が複数国家間かつ産学官を含む複数のパートナー間の連携によって進んでいる。

日本においては、研究の立ち上がりは早く、例えばマルチメディア分野では米国とともに最も早くから当該分野の研究を開始した国の一つであった。加えて、日本語コンテンツという特殊性、microblog の中で最も利用者の多い Twitter に占める日本語の割合が英語に次いで第二位であるという事実、映像コンテンツ制作のプロフェッショナルとしての NHK が世界第二位の規模の放送事業者であるという事実 （1 位は BBC） など、IT メディアにおける存在感からは、日本が当該研究分野を主導する一翼を担うべき立場にあると考えられる。しかし、米国においてはフェアユース規定により研究用データセットの構築・配布がなされ

ているのに対し、日本では著作権・肖像権・個人情報保護など様々な制約からこれができず、研究開発においても主導的な役割を果たすことができないでいる。ただし、基礎研究自体は盛んに行われている。

中国・韓国においては、オープンデータ構築が徐々に進んではいるものの、研究用データセットの構築・配布や研究開発の先導という動きは見えない。しかし、優秀な留学生や欧米で実績を上げた研究者を多数自国に迎え、加えて国内の大学などと欧米の有力研究拠点との共同研究が多く見られ、有力ジャーナルやトップ国際会議などにおける存在感を急速に増しており、一部の分野ではすでに日本を凌駕していると考えられる。

応用研究・開発においても、層の厚い米国が先導的立場にある。日本においても、産学官の研究などで盛んに研究開発が行われており、日本語であるなど特殊性の高い IT メディアについては特に盛んである。産業化においては、Google、Microsoft、Facebook、Yahoo!などを擁する米国が圧倒的に強い。

（４）科学技術的・政策的課題

・大規模データの収集と配布の重要性: 多様性を有する IT メディアの解析の研究開発には、実際の対象と同等の特性や多様性を反映するだけの十分大規模な研究開発用データの収集と配布が重要となる。特に、インターネットから容易に収集できないデータについては、そのニーズが高い。さらに、研究開発の効果的な推進のために、解析結果として望ましい正解データが付与されている必要がある。こうしたデータを多くの研究者が共通して用いることで、研究成果の比較が可能となる。米国ではフェアユース規定によりデータ公開が広く行われているが、我が国においては、著作権・肖像権・個人情報保護法などの問題により難しくなっており、本分野の研究開発推進において障壁となっている。

・研究開発のマイルストーンの設定: IT メディア分野の研究開発の特段の推進のためには、研究開発用データセットの公開とあわせて、適切な研究開発マイルストーンを設定・発信し、研究開発活動を戦略的に振興することが重要である。米国では TREC/TRECVID や、NSF/IARPA などのプロジェクトでこうした戦略が多くみられ、欧州においても Image CLEF や MediaEval などの例がみられる。日本でも研究開発の適切なマイルストーンの独自の設定が重要となると考えられる。

・政策的な研究開発の底上げ: IT メディア分野の産業においては、特に研究開発から展開までの速さ、世界的な展開の容易さなどから、一ないしはごく少数の企業のみが「勝ち残る」ケースが多い。こうした「勝者」が米国に集中している状況では、米国以外では当該分野の研究開発が、自然発生的に世界的にも存在感を示すほど進展することは考えにくい。これを打破するためには、欧州における FP7 や Horizon 2020 での ICT 分野への取り組みの一定の成功を見るに、日本でも政策的な研究開発の底上げが重要と思われる。

（５）注目動向（新たな知見や新技術の創出、大規模プロジェクトの動向など）

・欧州では、FP7 に続く Horizon 2020（2014～2020）の枠組みを通じて EU Open Data portal のさらなる整備を進めている。2015 年予算では「ICT 16-2015:Big data – research」において、Web、マルチメディアを含むビッグデータを対象とした革新的解析技術（データストラクチャー、アルゴリズム、ソフトウェアアーキテクチャー、最適化、言語理解を

含む）に関する研究開発が公募される（予算総額37MEUR）[1]。特に注目すべきは、多種多様なデータストリーム（多言語、マルチモーダル）に対するリアルタイム解析技術について、公募文書中で特に明示している点である。

- xLiMe[2]は、様々なメディアデータ（テキスト、音声、映像）から複合的な知識抽出をリアルタイムで行うとするものであり、FP7のプログラムとして2016年10月までの3年間で44MEURが支出される。その他、FP7では、CUbRIK[3]、MICO[4]などマルチメディア検索システムのための基盤構築が進められている。
- EU 個人データ保護規則案第17条「忘れられる権利」に関連し、FP7においてForgetItプロジェクト[5]が2013年から開始され、ITメディアデータの保存をさらに進めるためのForgetting modelの研究が開始されている。
- Chorus+[6]は、FP7 の援助による、多数のプロジェクトを携えた巨大プロジェクトであり2012年末にプロジェクト自体は終了した。しかし、avmediasearch.euのサイトを通じ、多くの成果にアクセスすることが可能となっている。
- フランス国立視聴覚研究所 （INA）[7]では、フランス国内のラジオ・テレビ放送を原則すべて蓄積し続けており、商用・研究用に提供されている。オランダ視聴覚研究所（Netherlands Institute for Sound and Vision）[8]でも、オランダ国内のテレビ放送を含む視聴覚情報を蓄積し、研究用の提供を行っている。
- 米国では、2014年5月のPCASTによる報告「ビッグデータとプライバシー」[9]において、ビッグデータにおける各種メディアデータ（ソーシャルメディア、画像、映像などを含む）の解析が新ビジネスを生み出すのは勿論のこと、プライバシー保護のための研究が重要であることが指摘されている。
- NSFの援助により、2013年から5年計画でBrownDogプロジェクト[10]が10MUS$の予算でイリノイ大学を中心に進められている。画像、動画、音声などのあらゆるインターネット上のBorn Digitalデータに対してタグを付与し利用可能にすることを目指している。
- ImageNet [11]は、米国スタンフォード大学Li Fei-Fei教授らが構築している画像意味解析用のデータセットであり、2.1 万の概念数、1.4 千万の画像を含む。規模、品質ともに群を抜き、同分野の研究、ならびに評価のデファクトスタンダードとなっている。同大学の他、AmazonやGoogleのサポートを受けている。
- 米国 Internet Archive [12]は、1996年に設立された非営利のデジタル図書館であり、世界中の Web 情報を定期的に収集し、検索などのサービスを提供している。研究などの用途にデータセットとしての一般公開は行われていない。
- 米国標準技術局 （NIST） では、TREC[13]ならびにTRECVID[14]プロジェクトにより、情報検索ならびに映像解析・検索プロジェクトの推進を図っており、大規模データセットの構築・配布、研究マイルストーンの策定を行い、世界中の研究グループを主導している。
- 米国 IARPA ALADDIN プロジェクト[15]では、2011年から5年計画で、インターネット上の映像を含む任意の映像中の動作などのイベント情報を高精度で検出する技術の実現に取り組んでいる。IARPA 主導のプロジェクトであることから、国家安全上重要な課題として認識されていると考えられる。
- 韓国では、情報通信研究振興院から先端融合コンテンツ技術開発の公募が 2014 年に実施（総額230億ウォン）[16]され、コンテンツ、プラットフォーム、ネットワーク、機器の融

合を目指したコンテンツに関わる研究開発が行われている。2013 年度の同公募では、韓国データベース振興院を中心としてビッグデータコンテンツサービスのためのマルチメディアコンテンツ内オブジェクトデータ抽出に関わる研究[17]が進んでいる（4.34 億ウォン）。

・文部科学省受託研究「多メディア Web 解析基盤の構築および社会分析ソフトウエアの開発」[18]は、2009 年から 4 年間のプロジェクトであり、東大においてアジア域最大の 200 億 URL に及ぶ Web 時系列アーカイブ、国立情報学研究所において 20 万時間に及ぶ放送映像アーカイブを実現しているが、著作権などの問題のため、内部利用にとどまっており、研究コミュニティ一般への公開には至っていない。

（６）キーワード

Web、blog、microblog、マルチメディア、内容解析、意味解析

（７）国際比較

国・地域	フェーズ	現状	トレンド	各国の状況、評価の際に参考にした根拠など
日本	基礎研究	○	→	・データセット構築・公開は、著作権、肖像権・個人情報などの問題により、研究用途に広く公開されている例はまれである。特に画像・映像を主としたデータセットはほとんど存在しない。 ・再配布まで可能な権利を有した機関が配布する場合には、データセットが公開されている場合がある。毎日新聞社は研究用に毎日新聞テキストデータを配布しており、日本語自然言語処理の発展に顕著に貢献した。 ・blogやTwitterなどのデータは、産学で大規模に収集されているが、著作権問題等により一般には公開されない状況が続いている。 ・blog、Twitterを対象とした解析、画像・映像の解析も産学官において盛んに研究されている。 ・テキスト情報検索を主対象とした共同研究プロジェクトNTCIRでは、大量のテキスト情報を研究グループに配布しているが、研究覚書を取り交わし配布するなど、特段の配慮が必要となっている。 ・研究会等の発表件数の推移から、基礎研究は引き続き活発に行われていることがわかる。また、日本語自然言語処理や日本語音声認識については、実用レベルの技術が広く利用可能となっている。
	応用研究・開発	○	→	・基礎研究の成果を利用し、日本語という特殊性に基づく独自の応用研究、国内のニーズに対応した応用研究・開発が盛んに行われている。
	産業化	○	→	・応用研究・開発と同様であり、日本語という特殊性や国内の特有のニーズに対応した産業は国内に多く存在する。
米国	基礎研究	◎	→	・研究の戦略的な振興のため、研究用データセット構築・配布と、研究すべきタスクの設定など、世界でも中心的な役割を果たしている。NIST、LDCなどがその中核的な機関となっている。 ・NSFやDARPAは、上記機関とも呼応し、米国内での基礎研究を特段にサポートしており、顕著な研究成果を上げている。 ・米国立議会図書館によるTwitterデータを含むウェブデータ収蔵、IARPA ALADDINプロジェクトにおけるインターネット上の映像のイベント解析など、より挑戦的なITメディア解析研究への政府主導の動きがみられる。 ・BrownDogプロジェクトが目標とするBorn Digitalデータに対するタグ付与は、あらゆるインターネット上のメディアコンテンツの利用推進を促すものであり、NSFはその重要性を認識していると思われる。
	応用研究・開発	◎	→	・Google、Facebook、Microsoftなどの米国内の巨大ITメディア企業でも研究部門を有し盛んに応用研究・開発を行っている他、大学とこうした企業との連携も盛んであり、加えて大学において、起業を目指した応用研究・開発も極めて盛んである。
	産業化	◎	→	・今後の研究開発において重要となる、ITメディアにおけるビッグデータを実際に保有する企業は、Google、Facebook、Microsoftなど、ほとんど米国に集中している。基礎研究、応用研究・開発の層の厚さとも相まって、今後ともITメディア研究の産業化においても米国は主導的な役割を果たすと思われる。
欧州	基礎研究	◎	→	・Horison 2020の「Leadership in Enabling and Industrial Technologies」の項目では、ITメディア分野の基礎研究として、EU Open Data portalへのデータの集約、および、それらのデータを活用するプラットフォームの研究開発の公募が盛んである。 ・FP7において、ITメディア分野のパーソナルデータ、組織データの保存をより進めるための忘却に基づくForgetItプロジェクトが2013年から開始されていることからも分かるように、時代のニーズに沿った基礎研究にも力点が置かれている。 ・Image CLEF、PASCAL VOC、MediaEvalでは、実際の研究用マルチメディアデータセットを構築し、研究コミュニティに公開している。

欧州	応用研究・開発	○	↗	・Horison 2020の「Social Challenge」では、高度道路交通システムのさらなる高度化や警察等の取り締まり機関における証拠抽出において、インターネットやソーシャルメディア上のデータ活用を含む公募がある。 ・同基礎研究公募においても、プラットフォームの研究開発に重きが置かれ、SMEにおける利用促進を狙っている。
	産業化	○	→	・産学官連携コンソーシアムとしてのQuaeroにおける参画企業の積極的な活動やExaleadの起業などの例もある。しかし、世界的に顕著な存在感を示すまでには至っていない。
中国	基礎研究	○	→	・オープンデータの収集、公開がOpendataChinaで開始されているものの、データ設置の構築・配布やタスク設定による行うべき研究の方向性の提示例はほとんどない。 ・研究は極めて盛んに行われている。特にマルチメディア分野においては、中国の大学は、特に近年顕著に能力を上げてきており、トップ会議にも採択論文が散見されるようになっている。 ・中国の大学において、米国や欧州の有力研究者との連携を図り、研究能力の飛躍的発展を実現でいる例が多く見られる。
	応用研究・開発	△	→	・基礎研究から応用に結びついた例はまだ多くは見られないが、基礎研究の進展を見るに、今後は応用研究・開発の進展も見込まれる。
	産業化	○	→	・Baidu、Sina Weiboなどの起業例もみられるが、世界的に存在感を示すまでには至っていない。 ・言語の特殊性や国内の特有のニーズに対応した産業は国内に多く存在する。
韓国	基礎研究	○	→	・Government 3.0(2013.6)の一環として、オープンデータの収集が積極的に行われているものの、データセットの構築・配布、タスク設定による行うべき研究の方向性の提示例はほとんどない。 ・先端融合・複合コンテンツ技術開発事業（情報通信研究振興院）において、ITメディアを含むコンテンツに関わる研究が開始されている。 ・データベース、コンピュータービジョン （画像・映像解析） などの分野では顕著な成果が見えている。一方、画像・映像の検索や応用を主としたマルチメディア分野の研究はあまり盛んではない。
	応用研究・開発	△	→	・基礎研究から応用に結びついた例はまだ多くは見られないが、基礎研究の進展を見るに、今後は応用研究・開発の進展も見込まれる。
	産業化	○	→	・オープンデータ利用の成功事例として、バス情報を利用したサービス構築などがある。 ・Naverの起業などの例もみられるが、世界的に存在感を示すまでには至っていない。

（註 1）フェーズ
基礎研究フェーズ ：大学・国研などでの基礎研究のレベル
応用研究・開発フェーズ ：研究・技術開発（プロトタイプの開発含む）のレベル
産業化フェーズ ：量産技術・製品展開力のレベル

（註 2）現状
※我が国の現状を基準にした相対評価ではなく、絶対評価である。
◎：他国に比べて顕著な活動・成果が見えている、 ○：ある程度の活動・成果が見えている、
△：他国に比べて顕著な活動・成果が見えていない、×：特筆すべき活動・成果が見えていない

（註 3）トレンド
↗ ：上昇傾向、 →：現状維持、 ↘：下降傾向

（８）引用資料

1）HORIZON 2020 WORK PROGRAMME 2014-2015
http://ec.europa.eu/research/participants/portal/doc/call/h2020/h2020-ict-2015.zip
2）xLiMe – crossLingual crossMedia knowledge extraction
http://www.xlime.eu/
3）CUbRIK
http://www.cubrikproject.eu/
4）MICO: Media in Context
http://www.mico-project.eu/
5）ForgetIT
http://www.forgetit-project.eu/
6）Chorus+
http://avmediasearch.eu/
7）フランス国立視聴覚研究所 （INA）
http://www.ina.fr/
8）オランダ視聴覚研究所 （Netherlands Institute for Sound and Vision）
http://www.beeldengeluid.nl/
9）Report to the President: BIG DATA and PROVACY
http://www.whitehouse.gov/sites/default/files/micro-sites/ostp/PCAST/pcast_big_data_and_privacy_-_may_2014.pdf
10）CIF21 DIBBs: Brown Dog
https://opensource.ncsa.illinois.edu/confluence/display/BD/CIF21+DIBBs:+Brown+Dog
11）ImageNet
http://www.image-net.org/
12）Internet Archive
http://archive.org/
13）TREC
http://trec.nist.gov/
14）TRECVID
http://trecvid.nist.gov/
15）IARPA ALADDIN プロジェクト
http://www.iarpa.gov/index.php/research-programs/aladdin-video
16）先端融合コンテンツ技術の開発
http://www.nipa.kr/biz/biz.it?bizId=00602&menuNo=815&boardId=info
17）2013 年度コンテンツ産業技術支援事業（指定公募）2 次発表
http://iacf.kw.ac.kr/index.php?mid=iBizNotice&document_srl=2839&listStyle=viewer&page=22
18）文部科学省受託研究「多メディア Web 解析基盤の構築及び社会分析ソフトウエアの開発」
http://www.mext.go.jp/b_menu/shingi/gijyutu/gijyutu2/006/shiryo/__icsFiles/afieldfile/2011/08/31/1310172_01.pdf

3.10.6　ライフサイエンス分野におけるビッグデータ

（1）研究開発領域名

ライフサイエンス分野におけるビッグデータ

（2）研究開発領域の簡潔な説明

ライフサイエンス分野においては、これまでは主に、純粋な生物学的興味から、ある生物種（例えば、ヒト）の動作原理に関する研究が主であった。しかし、シークエンサ（DNA配列解読装置）の発達により、個別のヒト・患部・細胞の違いを解析することが可能となった。このような研究には、多数の対象から、大量のデータを取ることが必要であり、その情報学的・統計的解析の必要性が高まっている。

（3）研究開発領域の詳細な説明と国内外の動向

2000 年代前半にヒトゲノムが解明されたことにより、生物学の方法論は、かなり情報学的手法を駆使したものに生まれ変わったが、それでも、扱うデータの量は限られており、最先端の統計学・機械学習の手法が用いられることも、あまりなかった。疾病の原因として、一塩基変異(SNP)のような単純な原因が簡単にみつかるだろうという楽観論に基づけば、基本的には、病例が数例あれば、特に複雑なデータ解析をすることなく原因を発見し、それに基づき治療法を開発することができるだろう。しかし、現実はより複雑である。例えば、同じ人のがんであっても、違う部位を取ってゲノムを比較すると差異が存在し(intra-tumor heterogeneity)、どれか一つに効く薬が完成したとしても、他の種類のがんが生き残ってしまい、命を救う事はできない [1)]。同時に、抗がん剤の効き目、副作用は、投与される人のゲノム変異に大きく影響される。従って、がん制圧の具体的な姿というのは、非常に多数の薬剤を用意し、それを、患者、患部のゲノム変異に基づいてカクテルを作成し、薬剤耐性に考慮しながら時系列で変化させていくという形にならざるを得ない。このような医療を個別化医療(personalized medicine)、あるいは、精密医療(precision medicine)と呼ぶ [2)]。現在、HIV の治療においては、このような医療がほぼ実現されており、HIV の体内での変異をシーケンサでモニタリングし、適切な投薬を行う事で、AIDS を発症することなく生存することができる [3)]。このような医療を、生活習慣病・精神病・がんなどにおいて実現することが、これからの課題である。このように複雑な医療を実現するには、データ解析、また、治療の設計に情報学の優れたアルゴリズムが欠かせない。個別化医療では、これまでの生物学の原理理解に向けたアルゴリズムの延長線上にはない、全く異なる高度なアルゴリズムが必要である。

米国 NIH が 2012 年に開始したプロジェクト Big Data to Knowledge (BD2K)は、このような生物学の情報化の流れを最も良く表している [4)]。このプロジェクトは、2020 年までに 6.56 億ドルを投じるという非常に大規模なものであり、13 カ所の研究機関に研究センターを立ち上げるなど、従来から強い米国のバイオインフォマティクスの力をさらに盤石にする効果があると考えられる。このプロジェクトでは、生物情報学を、これまでのような生物学のおまけではなく、主役として位置づけている。つまり、この巨額の資金は、生物学実験に使われるのではなく、もっぱら、情報学的な目的、すなわち、データベース作成・新手法開発・人材育成などに用いられるという点で、これまでの ENCODE[5)]、1000 Genomes[6)]など

の大型プロジェクトとは異なる。

2015 年には、オバマ政権が Precision Medicine Initiative を発表した [7]。これは、2016 年だけで 2.15 億ドルを投じて個別化医療を実現しようというプロジェクトである。その内の 1.3 億ドルは NIH によって、100 万人のゲノム情報をシークエンシングするために使われる。0.7 億ドルは、National Cancer Institute (NCI)に配分され、がんのドライバー変異を含む遺伝的原因を発見するために用いられる。残りの 0.5 億ドルは、情報集約のための予算とされている。このような巨大な国家プロジェクトだけではなく、個別化医療のスタートアップとしては、Patrick Soon-Shiong による NantHealth[8]、Craig Venter の Human Longevity[9]、 あるいは、Regeneron[10]が活発に活動しており、このようなデータを Pfizer などの製薬企業が利用することで、創薬や疾病治療のあり方が、根本的に変わる可能性がある。これに対し、日本では、特筆すべき動きはなく、彼我の差がさらに拡大することが懸念される。

このように、ビッグデータによる科学への投資が拡大する一方、論文に出版された結果に誤っているものが多いという指摘もなされている。例えば、Ioannidis らによる Lancet の論文に詳しいが、製薬会社による追試の結果、ほとんどの論文に再現性がなかったという信じられない結果もある [11]。STAP 細胞に関する騒動などで、一般社会にも明らかになった通り、生物学においては、非常に意外な現象が報告されても、理論の積み上げが未熟なため、すぐに真偽を明らかにすることができない。それをいいことに、自らに都合のいいストーリーを考えて、それに合うように実験を進めるといった望ましくない態度で研究を行う傾向がある。本来であれば、自らに都合のいい仮説はできるだけ疑ってかかり、石橋をたたいてわたる態度が重要であるが、現状では全く逆と言わざるを得ない。例えば、実験を行う際には、期待される効果を事前設定し、それに見合うサンプルサイズをあらかじめ決めておかなければ、統計検定は意味を成さないのであるが、そのような手順は無視されるのが一般的である。このままでは、生命科学への巨額投資が、誤りだらけの論文を生み出すだけで、全く無駄になってしまう恐れすら考えられる。米国でも、研究の再現性の問題は深刻に受け止められている [12]。その対策として、全てのデータとプロトコルを、信頼できるタイムスタンプをつけた状態で、公共のレポジトリにアップロードさせることや、論文の著者でないオブザーバーの研究への関与を義務づけること、また、再現された論文の著者を表彰する仕組みなどが提案されている。再現性問題に関しては、これまでの統計学が低次元データを仮定しており、現在の状況に追いついていないという問題も指摘されているので、新たな数理的アイデアによる革新的な統計手法が求められている。

中国 BGI(Beijing Genome Institute)[13]は、純粋な公的研究機関ではない。大量のシークエンサと、大勢の若い技術者(大半が 20 代)を背景に、自らも研究を行う一方、世界中の研究機関・企業から、サンプルを集めて、商業的にシークエンシングサービスを行っている。価格競争力も高く、日本をはじめとする各国に支社も存在し、人気を集めている。中国の基礎的な生物学のレベルはまだ高くないが、請負仕事を大量にこなすことによって、ノウハウが蓄積され、今後は急速にレベルが向上するのではないかと予想される。これは、半導体において、台湾メーカーが、海外の仕事を請け負うことを通して結果的に世界トップに上り詰めた過程と類似している。

個人の細胞から DNA を抽出し、その変異を解析することによって、生活習慣病などのリ

スクを評価する商用サービス（ゲノム診断）が世界中で一般化している。最大手は、米国の23andMe[14]であるが、日本でも、DeNA[15]、 Yahoo![16]などの企業が参入し、一つの産業として成立しつつある。それ自体は喜ばしいことであるが、多数の個人ゲノムが一企業に蓄積されることは、プライバシー保護の意味ではあまり好ましいことではない。匿名化・秘密計算などの情報技術を生かして、ゲノム診断や個別化医療におけるプライバシー保護技術を確立させることが求められる。

（４）科学技術的・政策的課題

・大量のデータを解析し、信頼性ある結果を得ることができるアルゴリズムの開発が急務である。非常に多くの変数が観測可能になると、そこから疾病の原因となる変数を選び出す際に、誤発見の確率が高くなるという問題がある。論文の再現性問題を無視し続けていると、ビッグデータによる科学そのものの存立基盤を脅かしかねない。
・統計的なデータ解析結果には、どうしても誤りが含まれるため、それらをスムーズに検証実験に移行させる枠組みが必要である。現状では、各研究者の恣意的な判断に基づいている。
・個別化医療に不可欠な個人データの利用を促進するために、プライバシー保護技術の開発と普及が必要である。
・日本では、バイオインフォマティクス人材の絶対数が不足しているため、人材育成への取り組みが必要である。

（５）注目動向（新たな知見や新技術の創出、大規模プロジェクトの動向など）

・米国 Precision Medicine Initiative、 BD2K は、最も注目を集めているプロジェクトである。
・欧州では、域内でのデータを流通されるため、European Life Sciences Infrastructure for Biological Information （ELIXIR）[17]がスタートしている。
・日本では、東北メディカルメガバンク[18]などのコホート研究がスタートしている。産出されるビッグデータがどのように生かされるか注目である。
・ゲノムのプライバシーに関しては、JST CREST「ビッグデータ基盤」[19]において研究が行われている。

（６）キーワード

新型シーケンサ、ゲノム診断、ゲノムコホート、バイオインフォマティクス、プライバシー保護

（７）国際比較

国・地域	フェーズ	現状	トレンド	各国の状況、評価の際に参考にした根拠など
日本	基礎研究	○	↗	・iPS細胞の発見や、それに基づく再生医療の研究など、ライフサイエンスそのもののアクティビティは高い。ただし、情報学に詳しい人材の不足から、データ解析のレベルは高いとはいえない。この事実が、ライフサイエンス研究の質そのものを悪化させるかどうかは、まだ不明である。
	応用研究・開発	△	→	・欧米に比べて、大規模な研究プロジェクトが少なく、人材が不足している。これは、数十年にわたって投資を怠ってきたことが原因であり、すぐに好転するとは考えられない。長期的な視点に立って、トレンドを好転させる努力が必要である。
	産業化	△	→	・欧米に比べると低調といわざるを得ない。
米国	基礎研究	◎	↗	・すべてに関して世界をリードしている。基礎的な生物学、シークエンサなどの測定機器、情報学に関してすべて圧倒的な力を持つ。NCBI[20]には世界中の情報が集まる。
	応用研究・開発	◎	↗	・現状では世界一の力を持つ。Googleが、Google genomics[21]というクラウドプラットフォームを開始した。
	産業化	◎	↗	・米国の強みは、新技術が開発されるとすぐにベンチャーによって事業化される点である。ゲノム診断に関しても23andMeを始めとする多くの企業がある。
欧州	基礎研究	◎	↗	・Wellcome Trust Sanger Institute[22]、EBI[23]、ドイツのMax Planck[24]などが強力なゲノム関係のプロジェクトを推進している。基礎的な生物学のレベルも高い。
	応用研究・開発	○	↗	・イギリスにおいては、50万人規模のUK Biobank[25]が進行中である。また、スウェーデンでも、同様に50万人規模のLifeGene[26]プロジェクトが立ち上がりつつある。
	産業化	○	→	・DecodeMe[27]などのゲノム診断サービスが存在する。ただ、米国に比較すると広がりとしては小さい。
中国	基礎研究	△	↗	・前述のBGIを中心に、ハイインパクトジャーナルに多くの論文を出版している。ただし、基礎生物学のレベルは高くない。
	応用研究・開発	○	↗	・コホート研究では、UK Biobankと同規模のChina Kadoorie Biobank[28]がスタートしている。
	産業化	△	↗	・BGIによるゲノム解析の受託解析サービスが世界的に展開されている。
韓国	基礎研究	△	↗	・今のところ特筆すべき点はない。
	応用研究・開発	△	↗	・今のところ大きな動きはない。
	産業化	△	↗	・特筆すべき点はない。

（註 1）フェーズ
基礎研究フェーズ ：大学・国研などでの基礎研究のレベル
応用研究・開発フェーズ ：研究・技術開発（プロトタイプの開発含む）のレベル
産業化フェーズ ：量産技術・製品展開力のレベル

（註 2）現状
※我が国の現状を基準にした相対評価ではなく、絶対評価である。
◎：他国に比べて顕著な活動・成果が見えている、 ○：ある程度の活動・成果が見えている、
△：他国に比べて顕著な活動・成果が見えていない、×：特筆すべき活動・成果が見えていない

（註 3）トレンド
↗ ：上昇傾向、 →：現状維持、 ↘：下降傾向

（８）引用資料

1) http://www.ncbi.nlm.nih.gov/pubmed/22469128
2) http://www.ncbi.nlm.nih.gov/pubmed/23361103
3) http://www.ncbi.nlm.nih.gov/pubmed/22673150
4) http://bd2k.nih.gov/
5) http://www.genome.gov/encode/
6) http://www.1000genomes.org/
7) http://www.whitehouse.gov/blog/2015/01/30/precision-medicine-initiative-data-driven-treatments-unique-your-own-body
8) http://nanthealth.com/
9) http://www.humanlongevity.com/
10) http://www.regeneron.com/
11) http://www.ncbi.nlm.nih.gov/pubmed/24411645
12) http://www.nature.com/news/policy-nih-plans-to-enhance-reproducibility-1.14586
13) http://www.genomics.cn/
14) https://www.23andme.com/
15) https://mycode.jp/
16) http://medical.yahoo.co.jp/hdl/gene/
17) http://www.elixir-europe.org/
18) http://www.megabank.tohoku.ac.jp/
19) http://www.jst.go.jp/kisoken/crest/research_area/ongoing/bunyah25-6.html
20) http://www.ncbi.nlm.nih.gov/
21) https://cloud.google.com/genomics/
22) https://www.sanger.ac.uk/
23) http://www.ebi.ac.uk/
24) http://www.molgen.mpg.de/
25) http://www.ukbiobank.ac.uk/
26) https://www.lifegene.se/
27) http://www.decode.com/
28) http://www.ckbiobank.org/

3.10.7　教育とビッグデータ

（1）研究開発領域名

教育とビッグデータ

（2）研究開発領域の簡潔な説明

学習者の学びの向上、あるいは教育機関や教育行政における意思決定のために、教育に関わる様々なデータを組み合わせて行うデータ解析に関わる研究開発を指す。

ビッグデータという用語は、学習マネジメントシステム（LMS）やオンライン教育、特に大規模公開オンライン講座（MOOCs）により蓄積される学習ログやその他のライフログ等の、大規模データの有する可能性への期待から特に用いられるようになったが、広く、各種教育統計や学生の属性等管理情報なども含む多様な教育データの組み合わせを指す場合もある。

（3）研究開発領域の詳細な説明と国内外の動向

a)　研究開発領域の詳細な説明

「教育とビッグデータ」という研究開発分野には、二つないし三つのルーツがある。

一つは、認知科学や学習科学、人工知能、機械学習等の分野にルーツをもつ「エデュケーショナル・データマイニング（EDM)」である。LMS や各種の e-ラーニング・システム、学習端末等を通して学習者の学習状況に関わるデータが大規模に集まるようになり、2000 年頃から各種の国際会議でワークショップが開催され、2008 年には初の EDM 国際研究集会（モントリオール）の開催、2009 年には初の EDM の学術雑誌（”Journal of Educational Data Mining”）の創刊、2011 年には初の学会設立（”International Educational Data Mining Society (IEDMS)”）につながった。

EDM とほぼ同様の研究開発領域ではあるが、解析結果を応用することに、より重点を置いた「ラーニング・アナリティクス（LA)」という分野がある。こちらは、ICT の活用推進を通して高等教育を発展させることを目的とする米国の業界団体「EDUCAUSE」が、重要な技術トレンドを紹介するために毎年発刊する”Horizon Report”の 2011 年版（HR2011）に取り上げてから急速に発展し、同分野の学会（”Society for Learning Analytics Research (SoLAR)”）と 2011 年からの会合開催（” Learning Analytics and Knowledge Conferences (LAK)”)、そして 2014 年からの学術雑誌（” Journal of Learning Analytics”）の発刊につながった。

EDM と LA の研究開発内容はほぼ同様であるが、その違いを説明することも試みられている [1)2)]。EDM は人工知能や機械学習をルーツとすることから、人が介入することのない機械学習により、学習者の学習モデルを見いだすことに重きがあるのに対して、LA は教育工学や学習デザイン、LMS 等の大規模 e-ラーニング・システムを基盤として、教育者や学習者への教育・学習支援や、教育機関や教育行政における意思決定の支援などの応用的側面に重きがあるとされる。しかし両者ともカバーする領域は同じであり、EDM においても学習支援や意思決定支援といった応用が試みられるようになっており、一方で LA においても各種の支援情報を提供するために、学びのプロセスに関する理解の精緻化に取り組むという動きがあり、両者歩み寄っていると言える。SoLAR の発起人である G. Siemens は、両者の歩

み寄りと連携を提唱している[1)]。

一方、EDM と LA とは別に、G. Siemens は LA と対比させるかたちで、”Academic Analytics (AA)”という概念を提示している。”Open Learning Analytics”を呼びかけた SoLAR のパンフレット[3)]によると、LA が学習者や教育者を支援するのに対して、AA は機関レベルや行政レベルの執行部の意思決定を支援する。また、そのようなことから、LA は学習者の学習ログやライフログなど、学習プロセスレベルのミクロなデータを扱うのに対して、AA は教育統計や学習者の属性等管理情報など、よりマクロなデータを扱うとしている。なお、AA は狭義には、ビジネスインテリジェンス（BI）を機関レベルの意思決定に用いる際の解析を指すこともある。

AA は日本において（そしておそらく米国を含む他国においても）十分浸透している用語ではなく、データ解析等を専門とする情報科学分野等の研究者の用いる用語である。これに対して、高等教育機関においては伝統的に、「インスティテューショナル・リサーチ（IR）」という用語を使用している。IR とは、情報収集やデータ解析等を通じて執行部等の意思決定を支援するための機関レベルの調査・研究活動を指し、米国では 1960 年代から取り組まれ、米国の各高等教育機関等においては IR のための専門部署が常設されている場合が多い。伝統的には、大学の財務に直結する入学者管理等や、外部への説明責任を問われる評価報告書の作成を中心に、財務や企画関連部署内に置かれていたが、近年は高等教育のマス化に伴うラーニング・アウトカム重視の傾向から、在学生一人一人の学習等を分析し、学習支援をすることも行われている。

日本においては 2008 年ごろから IR の重要性が中教審等において指摘[4)]されるようになり、各高等教育機関において学内に IR 担当や IR の機能を有する部署の設置、あるいは設置の検討が進められている。しかし日本の大学においては大学運営が事務の各担当部署にて行われてきていたことから、データを一つの部署に集約し、これを意思決定に用いるという体制を敷くことが容易ではなく、日本の大学における IR は、高等教育における教育重視の傾向を受け、学生の学習支援を主眼とするいわゆる「教学 IR」を中心に展開されている場合が多い。これは学内に設置された教育開発センター（Center for Teaching and Learning (CTL)）などで行われている。

AA や IR は、教育機関の機関レベルの意思決定を主眼とし、伝統的に教育統計や学習者の属性等管理情報などのマクロなデータを用いていたが、高等教育における世界的な教育重視の流れから「教学 IR」への注目が高まり、学生の科目の履修状況や成績など、学生一人一人を対象とした分析と学習支援が行われ、LA や EDM などへの歩み寄りが見られる。一方で、LA や EDM においても、学習ログ等の学習プロセスそのものに関わるデータだけでなく、学生の属性等管理情報が解析精度を上げる上で極めて重要なファクターとなることが確認されていることから、よりマクロなデータも用いることが行われており、両者歩み寄っていると言える[1)]。

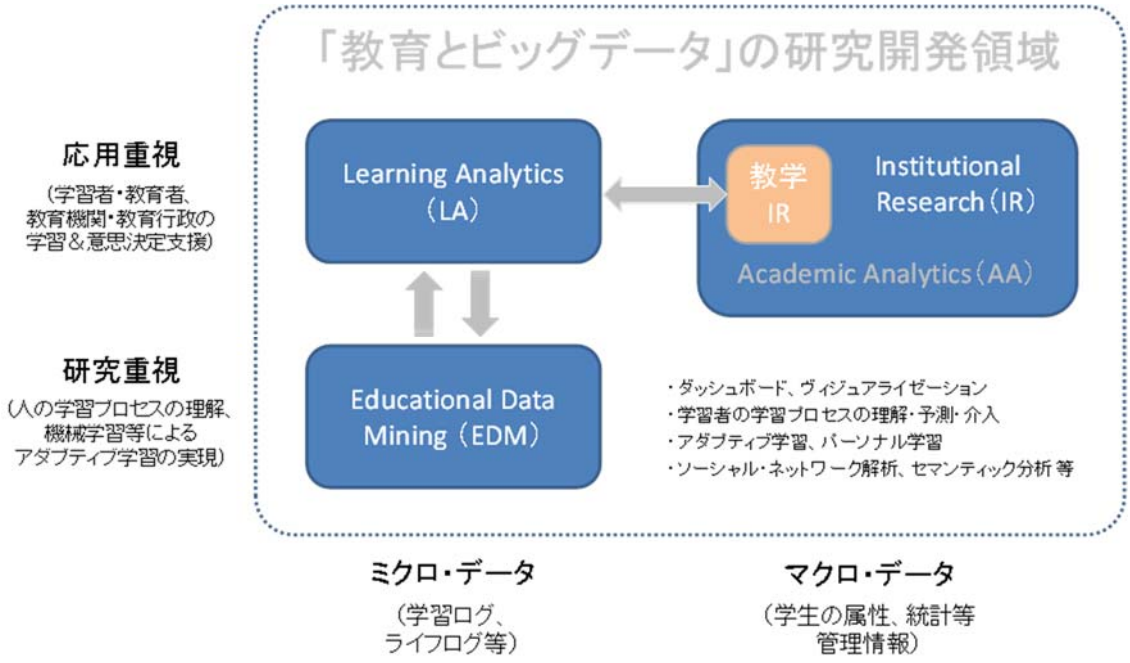

図 3.10.2 「教育とビッグデータ」の研究開発領域

なお、EDM や LA などの研究開発分野が出現したのは、LMS や各種の e-ラーニング・システムを通した学習者の学習ログや学習者の位置情報等ライフログなどの、学習状況に関わるデータが大規模に集まるようになったこと、ビッグデータへの一般的な関心が高まったこと、専門家でなくとも簡易にデータ解析や解析結果のヴィジュアライゼーションを可能とするビジネス・インテリジェンス（BI）などが普及だしたことなどが背景にある[3)5)6)]。一方、大規模公開オンライン講座（MOOCs）が世界的にブレークしたのは LA が提唱された翌年の 2012 年であり、MOOCs は必ずしも EDM や LA 形成の直接的な要因ではない。他方、MOOCs は同一のオンライン教育モジュールを数万名の学習者が学び、同じ条件下の均質なデータを大規模に提供する媒体となったことから、EDM や LA の絶好の解析対象となり、同分野の研究が飛躍的に伸びる材料を提供したと言える[7)8)]。

b) 国内外の動向

全般に機運は盛り上がっており[9)10)]、EDUCAUSE の HR2011 で LA が紹介されたときは 5 年以内にブレークするとあったものが、HR2012 では 2-3 年以内となり、HR2014 では 1 年以内となるなど、進展のスピードも感じさせる。ユネスコでは 2012 年 11 月に LA を紹介する政策文書をリリースし[6)]、米国では教育省が EDM と LA の動向に関する調査報告書を 2012 年 10 月にリリースしている[11)]。EU では「(5)注目動向」で詳しく述べるように、第 7 次研究枠組み計画（FP7）のもと、LA や EDM に関連して既に 3 つのプロジェクトが走っている[12)13)14)]。

国別で見ると、北米が大きく世界をリードしており、これに次いで欧州が研究開発を行っている。欧州では、オープン・ユニバーシティ（OU）が比較的普及しており、特にスペインと英国の OU が成功していると言われ、このためか、EDM をけん引しているのはスペイン・コルドバ大学の S. Ventura と C. Romero[15)]、LA をけん引しているのは英・OU の教育工学研究所（”Institute of Educational Technology”）の R. Ferguson である[16)]。

アジア諸国においては、台湾やシンガポール、香港、そして韓国などが各種 e-ラーニン

グ・プログラムの開発・提供という観点では極めて進んでいるが、これら学習データを解析し、学習支援や意思決定支援に役立てるという点ではそれほど多くの取り組みは行われていない模様である。ただし、これら諸国はデータ解析対象となるデータがこれらシステムから取得可能であるという意味で、これから本分野の研究が飛躍的に進む土壌を有している。他方、日本については、LMS の利用も含め、e-ラーニング等が伝統的なキャンパスを有する大学において普及していないこともあり、このデータ解析となるとさらにおぼつかない。

米国は、いくつかの要因により、この分野の研究開発が活発かつ産業化も進んでいる。第一には、オンライン教育や LMS など、デジタル時代における学習環境がすでに普及している。米国ではインターネット普及以前から、州立大学が州の農業や工業の発展のために遠隔教育を提供する機能を有しており、これが技術の発展とともにインターネットを媒体としたオンライン教育も提供するようになっている。こうしたノウハウやインフラが存在したことが、キャンパスで提供される正規の教育課程においても e-ラーニングが早い段階から導入される環境条件につながった。米国の大学では一科目以上オンラインで受講している学生の比率は 2011 年ですでに 32%に達し [17]、遠隔教育を受講する生徒のいる高校の割合は 96%にものぼる [18]。

第二に、米国では高等教育のマス化、ユニバーサル化が世界に先駆けて進み、学力面も含め多様な学生を高等教育で受け入れていることから、ドロップアウト等の問題が深刻で、学生一人一人への学習支援に対する必要性が高い。しかし高等教育財政の枯渇から、こうしたきめ細かい学習支援を提供するための人件費の予算は割けないため、電子システムで自動的に学生の学習支援を行う eAdvising などへの需要が生まれた。結果として、Knewton などの教育のパーソナル化やアダプティブ学習に対応可能な製品が民間において開発されている。アリゾナ州立大学は全米でも特に先駆的な大学で、来る人口増に備え、オンライン教育を全学的に展開し、eAdvising 等の学習支援も全学的に提供している（「注目動向(5)」に詳説）。

更に、2007 年のリーマンショック以降、高等教育予算の削減に伴う授業料の高騰、それと同時に提供科目数の縮小とそれに伴う在籍期間の長期化により、学生の借り入れる学生ローンが 10 万ドル以上で卒業後も返済しきれないほどと報道され [19]、大きな社会問題となり、さらに同時期に MOOCs が一大旋風を巻き起こしたことから、オンラインのみの無償もしくは安価な高等教育の提供に社会および行政から注目が集まった。オンラインのみの教育はドロップアウトの危険性が対面授業以上に高く、学生一人一人に対応したアダプティブ学習やパーソナル学習等を実装した eAdvising 等の必要性がさらに高い。

こうした社会・経済的要因なども背景とするため、米国は現在研究開発面だけでなく、産業化や実用化の面でもこの分野で世界断トツ 1 位のフロントランナーである。

なお、カナダは、カナダのオープン・ユニバーシティであるアタバスカ大学技術拡張知識研究所（Technology Enhanced Knowledge Research Institute (TEKRI)）に 2013 年 12 月まで在籍していた G. Siemens が、デジタル環境における学習ネットワークや技術、解析、ビジュアル化、オープン性、組織効率性などに関わる世界的な第一人者であり、「デジタル時代における学習理論：コネクティビズム」を提唱、LA の学会である SoLAR の発起人ともなっていたことから、（同国の国土が広く遠隔教育がもともと盛んであったこともあるが）、この分野で注目すべき存在であった。他方、G. Siemens が 2013 年 12 月にテキサス大学オ

ースティン校の学習革新とネットワーク知識研究ラボ（Learning Innovation and Networked Knowledge Research Lab（LINK Lab））に移籍したことから、LA の発信源がテキサスに移る可能性はある。

（４）科学技術的・政策的課題

（一般的な課題）

- **学習支援のリアルタイム性：**2011 年になって LA が大きく取り上げられるようになったのは、e-ラーニングや LMS 等を通じて学習者の学習ログや学習端末等を通じたライフログが大規模に取得可能となり、リアルタイムの学習支援も可能となってきたことによる。一方、リアルタイムで質の高い適切な学習支援をするためのデータの蓄積や解析はまだ開発途上であり、現在機械学習による方法等も含め、多様な研究開発が進められている [5)20)]。
- **データの質、解析に加えるデータセットの選定、重み付けの難しさ：**この分野では、「取得できたデータをもとに分析するのか、それとも分析したい事項をベースに分析するのか」が問題として指摘されるほど、解析対象とするデータの質やデータの取捨選択が問題となっている [6)16)]。オンライン教育や LMS 等から学習ログ、学習端末や ID カード等からの位置情報等ライフログなどの膨大なデータが取得可能とはなったが、それら解析対象となるデータが真の意味で「学習プロセスの理解」につながる情報を内在していないのであれば、解析をいくら精緻化しても、存在しない情報は引き出せない。学習者がクリックをしたり、講義ビデオを再生したりしていても、真の学びにつながっているかはその行為だけからは推し量ることはできないのである。さらに、LA ではこうした学習ログやライフログの解析だけでなく、学習者の属性等管理情報やその他の情報（学習者のその日の体調や性格、生活リズム等）も合わせて解析することが解析の精度を上げる上で重要とされているが、これら付帯情報をどこまで入れるべきか、またそれぞれをどのような重みでいれるべきかなども問題となっている。
- **データの取得可能性と紐付けの難しさの問題：**教育面のビッグデータの解析では、複数のデータを組み合わせることに大きな期待が寄せられている [11)16)]。例えば、科目の履修状況や成績だけでなく、学習者の属性情報や入試の点数、出身校における成績や家庭環境、趣味等も解析に含めた方が解析精度は上がる。しかし、これらのデータを教育機関が保有していない場合もある上、取得可能あるいは保有していたとしても、これらが別々の部署やデータベースで管理されていて、これらを統合できない、あるいは紐付けができないことが多いことも指摘されている。
- **プライバシーと倫理の問題：**プライバシーとデータ使用に関わる倫理の問題も強く指摘されている [5)6)11)]。こうしたデータを取得すること自体がプライバシーの侵害であるという考え方があるのはもちろんのことであるが、仮に在学期間中はそれらデータを用いてきめの細かい学習支援を行ったとしても、学習者が卒業後にどの程度の期間、それらデータや解析結果を保有し、かつ利用に供するかの問題がある。保管するデータ容量の問題という以上に、こうした過去の学習記録が人生を通じてつきまとうことに対して疑義が呈されている。ある教育課程で、たまたま健康状態や人間関係等により学業面で振るわなかったとしても、次の教育課程に進んだり、就職したりしたときに、気分を切り替えリセットして、新しい人生に新しい気持ちで挑むという権利を人は有するべきであるし、それがあるから

こそ発展もありうるのに、過去の事実が常につきまとうことが人生に大きくマイナスに作用する危険性が危惧されている。また、そこまでいかなかったとしても、学習支援情報を学習者に全て知らせるべきであるか、また担当教員にどこまで知らせるべきかについても、慎重を要するという意見が出ている。他方、プライバシーや倫理を危惧する余り、データの提供を渋ると、適切な学習支援を得られず、教育格差が拡大する危険性も指摘されている。

・**学習支援を自動化して行うこと自体の適切性に関する問題**：LA は学習者の学習支援に役立つことが重要な目標となっているが、コンピューターでいくつかのパターンに応じて機械的に学習支援を行うことが果たして適切であるかについても疑義が呈されている。数学に弱い学生が心理学を専攻する際、必須である統計学の単位を落とす確率が高いことを過去の類似の学生のデータから割り出し、同学生の適性から心理学以外の例えば英文学や異文化理解などの専攻を勧めるといったことがなされているが、それが当該学生にとって長い目で見たときに本当に最適な進路であったのか、また人類にとってそのようにパターン化された行動を生み出すことが良いのかといったことが問題となっている。一方で米国のように、ドロップアウト率が高く、4 年以内に卒業する大学生は 39%、6 年以内に卒業する大学生は 59.2%（社会人学生等を除く、いわゆる大学入学適齢の 2006 年度入学者の、それぞれ 2010 年度、2012 年度卒業時の統計 [21]）という事態を改善することが喫緊の課題の場合、人手を割けないのであればコンピューターによるパターン化されたアドバイスでも良い結果を生むことも事実である。

（日本特有の問題）

・**e-ラーニングの普及不足による解析対象データの不足**：国内においては、e-ラーニングやオンライン教育が十分に浸透していない。統計上はオンラインで授業を行っている大学の割合は 39.3%、LMS を導入している大学は 57.2%であるが（2012 年）[22]、これは全学に一科目でも導入されていればカウントされるため、大きく水増しされていると推測される。運用が活発な大学においても、語学等一般教育の限定的な科目や ICT 教育等が好きな教員が実践している場合が多く、組織的な取り組みは皆無である。
日本ではこのように、LA の解析対象となるデータを取得できるインフラの整備・利用が大きく遅れているため、この分野の研究開発を伸ばすためにはまず e-ラーニング等オンライン教育を拡大することが大前提である。同時に日本はプライバシー等の問題から、教育機関が LA 研究者と協力することを躊躇うことが多く、これも解析対象となるデータが不足する要因となっている。

・**研究者、周辺産業の発達不足**：e-ラーニング等が全般にそれほど普及していないこともあり、これに関わる研究者や周辺産業の発達が他の先進諸国と比較して少ない。LA 等の研究者は一般に教育工学あるいは人工知能等の分野から取り組むことが多く、こうした基盤となる分野が十分に発達しないと、これをベースにデータ解析する LA 等にまで至らない。

・**研究者と教育提供者の断絶**：国内における LA 等の研究開発は解析対象となるデータの取得の難しさから、研究者が自身の提供する科目において e-ラーニング等に取り組み、そこで得たデータを解析することが多く、研究開発のスケールが総じて小さい。また教育提供主体の協力が得られた場合においても、LA 研究者の提供する解析結果が技術的すぎてそ

の意味するところが教育提供者に十分に理解されず、また逆に、教育提供側のニーズを研究者が十分にくみ取れず、教育や学習の改善につながらないことが多いことが指摘されている。e-ラーニングに関する研究開発が極めて活発な台湾などにおいては、数多くの研究者が地元の初等中等教育機関の協力を得て、共同でe-ラーニング・システムを開発したり、改善の工夫を行ったりしており、優れた研究開発につながる土壌が形成されている。

（５）注目動向（新たな知見や新技術の創出、大規模プロジェクトの動向など）

・**Next Generation Learning Challenges：** Bill & Melinda Gates Foundationが、特に生徒の大学入学準備を図るための、テクノロジーを通じた教育革新について2011年に1000万ドルの資金を提供し、さらにWilliam and Flora Hewlett FoundationやEli and Edythe Broad Foundation も追加の資金提供をしている。米国はこうした巨大財団が、連邦政府とともに、全米の課題解決に乗り出すケースが近年多く、巨大財団の政策形成に関わる影響力の強さが指摘されている。

このイニシアチブにおいて LA も重点分野の一つとされており、初等中等教育段階ではInnovate EDU Inc.が、高等教育段階ではRio Salado CollegeとBryn Mawr Collegeが資金を得ている[23]。Rio Salado Collegeはアリゾナ州にあるコミュニティ・カレッジで、広域の学生を対象として遠隔教育が以前から展開しており、そうしたインフラをもとに今回のプロジェクトでは特に、低所得者の卒業率向上をターゲットとしたLAの研究開発を行い、実用化を進めている。なお、Rio Salado Collegeと次に説明するアリゾナ州立大学は近接している。

・**アリゾナ州立大学：**アリゾナ州では2000-2025年の間に人口が86%増加することが予想されており[24]、アリゾナ州立大学では”A New American University”という包括的な全学戦略を打ち立て、想定される高等教育人口の拡大に対応しようとしている[25]。産学連携等を通じた人員・施設の拡充等も想定されているが、それだけでは対応しきれないことから、教育面についてはオンライン教育を全学的に整備する方向性を打ち出している。世界最大規模の教育サービス企業であるピアソン社と協力し、同社に入学者募集から入学者管理、学習管理、学習支援等を全て任せ[26]、2009年に1200名程度であった受講生を2014年には1万名に拡大し、2015年には2.5万名達成見込み、2020年には10万名を受け入れ予定である。

高等教育人口の拡大に伴う学生が多様化によりドロップアウト率が現状以上に悪化することも懸念されるため、アリゾナ州立大学ではeAdvisingという仕組みを導入し[27]、履修科目の選定等も含め、過去の類似の学生のパターンから、どの科目であれば卒業に至る可能性が高いかを割り出しレコメンドするといった仕組み等も導入した。またリメディアル教育についてもアダプティブ・テクノロジーを用いてパーソナライズド学習を可能とするKnewton 社と協力し、数学やその他の科目について学習モジュールを開発、学生に提供している[28][29]。eAdvisingおよびこの先進的なリメディアル教育を通じて、ドロップアウト率の改善や、4 年より多くの期間を要していた卒業までの期間の短縮も見られた。アリゾナ州立大学は、オンライン教育の全学的導入も含めたこれらの取り組みにより、全米で最も先駆的な大学とされている。

・**スタンフォード大学 LyticsLab：**複数の学部基礎科目について、認知科学を用いた無償の

オンライン教材を高等教育機関向け開発した Open Learning Initiative をカーネギーメロン大学で率いた Candace Thille を 2013 年に引き抜き、大学院ゼミや研究会などを多数開催している。周辺シリコンバレーの IT 企業も参加しており、共同研究も活発に行われている。LA については Bill & Melinda Gates Foundation から資金提供を得て、2012-2014 年にかけて特別の WG（”The Learning Analytics Workgroup (LAW)”）が立ち上がった。LA に関わる米国の有識者 40 名前後の参加を得て、ワークショップやサマースクール（”Learning Analytics Summer Institute (LASI)”）を開催したり、LA に関わる各種の白書を作成したりした [30]。

スタンフォード大学の Lytics Lab で特筆すべきなのは、その研究開発内容だけでなく、解析対象となるデータの入手可能性である。この LyticsLab はスタンフォード大学のオンライン担当副学長室（”Vice Provost for Online Learning Office（VPOL Office)”）のもとにある。VPOL はスタンフォード大学におけるオンライン教育モジュールや MOOC の開発・提供を一手に担い、2011 年から 2014 年 5 月現在までにすでに 246 科目も提供をしている [31]。ここで蓄積されたデータは VPOL の管轄下にあるため、Lytics Lab の研究者や大学院生等がこれらデータを用いて研究をしたい場合は、VPOL が利用許諾や権利関係等の仲介をし、データ利用の便宜を図ることができる。この分野で優れた研究開発を行う上で肝要なのは質の高いデータを大規模に入手できることであり、スタンフォード大学ではそのような環境条件が整っていると言える。

- **edX 上の MOOCs「Data、 Analytics and Learning」**：注目動向というほどではないが、LA の第一人者で SoLAR の発起人である G. Siemens 他が LA 等の分野について解説する MOOCs。G. Siemens がカナダ・アタバスカ大学からテキサス大学アーリントン校に移籍したため、テキサス大学により提供されている。MOOCs は数万名の受講者、数千名の修了者を得ることが多いため、宣伝効果は大きく、ここで説明された体系が世界の流れとなる可能性はある。
- **欧州**：EU では第 7 次研究枠組み計画（FP7）のもと、LA や EDM に関連して既に 3 つのプロジェクトが走っている。１）LA と EDM の研究開発や実践を共有するためのコミュニティ形成のためのプロジェクト「The Learning Analytics Community Exchange（LACE）[12]」（2014.1-）、２）LA の解析用ツールを学習者および教育者に提供しようとする「LEA's BOX: A Learning Analytics Toolbox[13]」（2014.3-）、３）企業等の従業員が過去のパフォーマンスと自己学習の成果をもとに、現在直面している課題をリアルタイムに創造力をもって解決できるためのツールを提供しようとする「Mirror: Reflective Learning at Work[14]」である。
- **台湾・デジタル・ラーニング・イニシアチブ**：データ解析に特化している訳ではないが、台湾はデジタル学習に国策として力を入れ、初等中等教育から高等教育までを包含する「デジタル学習イニシアチブ」が教育部主導で推進されている。これはネットワークインフラだけでなく、「教育クラウド」やクラウド上に乗るアプリケーション「MOOCs」「モバイル学習」「デジタル読書」「デジタルチューター」「4G モバイルアプリケーション」なども含めて整備するものである。台湾の国立師範大学が e-ラーニングで進んでいることから、初等中等教育における教師のデジタル・リテラシーも高く、独自にソフト開発が可能であったり、e-ラーニング等の研究開発を行う大学の研究者との連携も進んでいたりする。

データ解析という観点では、「教育クラウド」を用いてアプリケーションを教育機関間で共有することが想定されており、解析対象となりうるデータが蓄積されつつある。またこの分野の研究者が国立中央大学ほか多くの大学にまたがって多数存在することから、台湾はこの分野で大きく伸びる可能性がある。

（６）キーワード

ラーニング・アナリティクス（LA）、エデュケーショナル・データマイニング（EDM）、教学 IR、インスティテューショナル・リサーチ（IR）、ビジネス・インテリジェンス（BI）、ヴィジュアライゼーション

（７）国際比較

以下の国際比較表は、本分野に関わる各種の取り組み事例とともに、LA および EDM に関わる論文サーベイにより行った。図 3.10.3、図 3.10.4 は九州大学情報基盤研究開発センター廣川研究室の検索システムにより得た、学術論文にみる本分野の俯瞰図である。キーワードとともに、国および著者の関係性をマインドマップで示してある。なお、検索に用いた論文は SCOPUS に含まれる LA および EDM に関わる 2200 件の論文である（一部他のデータベースより補完した論文あり）。

図 3.10.3 LA および EDM 分野における学術論文の KW および国の関係性と論文数

（出典）九州大学情報基盤研究開発センター廣川研究室の検索システムによる
使用データ：SCOPUS 等に含まれる LA および EDM に関係する論文 2200 件

図 3.10.4 LA および EDM 分野における学術論文の KW および著者の関係性と論文数

出典）九州大学情報基盤研究開発センター廣川研究室の検索システムによる
使用データ：SCOPUS 等に含まれる LA および EDM に関係する論文 2200 件

国・地域	フェーズ	現状	トレンド	各国の状況、評価の際に参考にした根拠など
日本	基礎研究	△	↗	・2004-2014年にこの分野で投稿された論文数は53件あり、世界第10位である。ただし、米国が827件でダントツ一位で、その後にスペイン（163）、イギリス（163）、ドイツ（116）、カナダ（114）、豪州（102）が続き、少し離れて中国（84）、インド（80）、そしてまた少し離れてオランダ（54）、日本（53）、台湾（50）という順番である。 ・3件以上論文を執筆している研究者が7名で、うち5名が同じ研究グループで基本的にEDMの分野である。その他の研究者については、それぞれ個人ベースで個別のテーマを追っている。 ・ただし、この分野をテーマとした特別セッションが学会等において組まれるなど、興味関心の芽生えは見られる[32]。また、この分野の動向を紹介する試みは行われている[33][34]。 ・なお日本は、もともとe-ラーニングやオンライン教育等の実践や研究が他国に比べて立ち遅れているなか、そこから取得/解析できるデータが少なく、この分野が発展する障害となっている。
	応用研究・開発	△	↗	・（同上）
	産業化	◎	→	・この分野を高等教育に限定すると、産業化はほとんど行われていないに等しい。オンライン教育を積極活用する株式会社大学のBBT大学やサイバー大学は、学生の履修や学習状況確認のためのシステムを導入しているが、これら大学以外ではキャリア教育やジェネリック・スキルの開発、リメディアル教育、そしてLMS等の提供において民間企業が若干、入り込みつつある程度である。 ・一方、この分野を初等中等教育に拡大すると、受験産業が発展している日本においては、生徒一人一人の学力判定や志望校への合格確率の算出、適切な志望校の提示等など、研究開発だけでなく実用化も含め、既に成熟しているとも言える。
米国	基礎研究	◎	↗	・すべての面で世界をリード。この分野をけん引するG. SiemensやR.S. Bakerもいる一方、LAやEDMの要素技術を研究・開発する研究者も充実している。 ・研究グループとしても、Open Learning Initiativeで有名なカーネギーメロン大学、同大学からC. Thilleを引き抜きリティクス・ラボを擁するスタンフォード大学、社会と技術の関係を探るということを研究の中心におき、教育面の研究ではR.S. J.d. Bakerを擁するWorcester Polytechnic Institute の Department of Social Science and Policy Studiesなど充実している。MOOCsにより、edXを設立したMITやハーバード大学も、MOOCsデータのビジュアル化やこれを用いた教育・学習支援に乗り出している[35]。
	応用研究・開発	◎	↗	・（同上）。 ・米国における深刻なドロップアウト対策のために、研究開発という意味ではなくても、全学的にLAやEDM、IR等を用いた学生の学習支援に乗り出す大学は多く、アリゾナ州立大学などは全米でも有名である。また、こうしたドロップアウト率削減対策についてオバマ大統領も積極的に乗り出しており、ゲイツ財団やルミナ財団等も多額の資金を提供している。
	産業化	◎	↗	・アダプティブ学習あるいはパーソナライズド学習のためのプラットフォームを提供するKnewton社は有名である。 ・その他、BlackboardやCheggなどのLMSプラットフォーム等を提供する企業もこれに乗り出している。Pearson社も、学生のエンロールメント・マネジメントから始まり、学習支援、卒業までの一連のサービスを手がけ、このなかで技術開発やテクノロジーを用いたサポートの提供も行っている。

欧州	基礎研究	○	↗	・欧州では、スペイン（163）、イギリス（163）、ドイツ（116）、オランダ（54）などの論文数が多い。特にEDMの分野をけん引してきたRomero C.とVentura S.の業績が突出し、次にLAの分野をけん引するR. Fergusonの業績が突出する。
	応用研究・開発	○	↗	・EUにおいて、１）コミュニティ形成のため「The Learning Analytics Community Exchange（LACE）[12]」、２）LA解析用ツールを提供する「LEA's BOX: A Learning Analytics Toolbox、[13]」、３）企業等向けの学習ツール「Mirror: Reflective Learning at Work[14]」などのプロジェクトを2014年から開始している。
	産業化			・（詳細は不明） ・ただし、EUがスポンサーする企業等向けの学習ツール「Mirror: Reflective Learning at Work[14]」において複数の協力企業が名乗りを上げている[36]。
中国	基礎研究	△	↗	・LAとEDMに投稿された論文数という観点では、日本の1.6倍程度であるが、2011年から論文数の急成長が見られ、かつ2013年以降はそれまではデータマイニング等情報科学の技術的論文が主流であったのに対して、e-ラーニング・システムで取得されたデータ等を用いた、より教育の現場に近い論文が見られるようになっており、これからの伸びを感じさせる。 ・特別に論文数の多い研究者あるいは研究グループは存在せず、論文数3件以下の研究者が100名以上存在するような状況である。
	応用研究・開発	△	↗	・（同上）
	産業化		↗	・詳細は不明であるが、中国においてビッグデータに向けての注目は大きく、教育面も含め大きなイベントが開催されるなど、これに向けての動きは活発化している模様である[37]。
韓国	基礎研究	△	↗	・意外なことではあるが、LAとEDMに投稿された論文数を見ると、日本は2004-2014年に53件あるのに対して、韓国は17件しかない。韓国はe-ラーニングやサイバー大学なども活発で、国立・韓国教育学術情報院（KERIS）なども設置されているが、e-ラーニングの提供は活発でも、そのデータの活用やオンライン上の学習支援という観点ではそれほど進んでいない模様である。 ・米国中心のジャーナルに投稿がなされていないということも考えられるが、現地の研究者にヒアリングしたところ、実際、それほど研究はなされていない模様であった。
	応用研究・開発	△	↗	・（同上） ・ただし、韓国は日本と違い、オンライン教育やLMS等が普及しているため、データ解析を開始すれと一気に伸びる可能性はある。 ・また、ソウル大学内には2014年4月にビッグデータ研究所が設置され、これは様々な分野に対応できるように学内横断的に各学部から教員が兼任で担当をし、一方で研究資金を政府や産業から、それぞれのニーズに対応するかたちで得るという仕組みを取っているため、研究開発をする体制は整備されつつある。同所長のサン・キュン・チャ教授は教育分野のビッグデータについても、2014.10の訪韓時に高い関心を示していた。
	産業化			・（詳細不明）

（註 1）フェーズ
基礎研究フェーズ ：大学・国研などでの基礎研究のレベル
応用研究・開発フェーズ ：研究・技術開発（プロトタイプの開発含む）のレベル
産業化フェーズ ：量産技術・製品展開力のレベル

（註 2）現状
※我が国の現状を基準にした相対評価ではなく、絶対評価である。
◎：他国に比べて顕著な活動・成果が見えている、 ○：ある程度の活動・成果が見えている、
△：他国に比べて顕著な活動・成果が見えていない、×：特筆すべき活動・成果が見えていない

（註 3）トレンド
↗ ：上昇傾向、 →：現状維持、 ↘：下降傾向

（８）引用資料

1) Siemens, G., R. S. Baker (2012) “Learning Analytics and Educational Data Mining: Towards Communication and Collaboration”
http://www.columbia.edu/~rsb2162/LAKs%20reformatting%20v2.pdf
2) Baker, R., G. Siemens “Educational Data Mining and Learning Analytics”
http://www.columbia.edu/~rsb2162/BakerSiemensHandbook2013.pdf
3) SoLAR (2011) “Open Learning Analytics: an integrated & modularized platform”
http://solaresearch.org/OpenLearningAnalytics.pdf
4) 中央教育審議会（2008)「学士課程教育の構築に向けて（審議のまとめ)」
http://www.mext.go.jp/component/b_menu/shingi/toushin/__icsFiles/afieldfile/2013/05/13/1212958_001.pdf
5) Picciano A.（2012)” The Evolution of Big Data and Learning Analytics in American Higher Education”
http://files.eric.ed.gov/fulltext/EJ982669.pdf
6) UNESCO Institute for Information Technologies in Education (2012) “Policy Brief: Learning Analytics”
http://iite.unesco.org/pics/publications/en/files/3214711.pdf
7) Eisenberg, M. and Fischer, G.（2014） “MOOCs: A Perspective from the Learning Sciences" in J. L. Polman et al. (Eds.), Learning and Becoming in Practice: 11th International Conference of the Learning Sciences (ICLS), Boulder, pp. 190-197
http://l3d.cs.colorado.edu/~gerhard/papers/2014/ICLS-MOOCS.pdf
8) Siemens, G., Dillenbourg, P., et al.（2014） "Where Are the Learning Sciences in the MOOC Debate?" in J. L. Polman et al. (Eds.), Learning and Becoming in Practice: 11th International Conference of the Learning Sciences (ICLS), Boulder, pp. 15-17
http://l3d.cs.colorado.edu/~gerhard/papers/2014/ICLS-panel.pdf
9) Chatti, M.A., et al. (2012) ” A reference model for learning analytics”, Journal International Journal of Technology Enhanced Learning, vol 4 issues (5/6), pp. 318-331
http://dl.acm.org/citation.cfm?id=2434498
10) Hirsh, L., DELTA (2013) ” Learning Analytics - Annotated Bibliography”
http://delta.ncsu.edu/assets/Learning_analytics_annotated_bibliography.pdf
11) U.S. Department of Education, Office of Educational Technology (2012) “Enhancing Teaching and Learning Through Educational Data Mining and Learning Analytics: An Issue Brief”
12) EU（2014)” The Learning Analytics Community Exchange（LACE)”
http://www.laceproject.eu/lace/
13) EU（2014)” LEA's BOX: A Learning Analytics Toolbox”
http://leas-box.eu/
14) EU（2014)” Mirror: Reflective Learning at Work”
http://www.mirror-project.eu/aboutus/about-mirror
15) Romero, C. S. Ventura (2007) “Educational data mining: A survey from 1995 to 2005”,

Expert Systems with Applications, Volume 33, Issue 1, July 2007, pp. 135-146
http://www.sciencedirect.com/science/article/pii/S0957417406001266

16) Ferguson, R. (2012) "Learning analytics: drivers, developments and challenges", International Journal of Technology Enhanced Learning 4 (5/6) pp. 304-317
http://oro.open.ac.uk/36374/1/IJTEL40501_Ferguson%20Jan%202013.pdf

17) Online Learning Consortium（2012)" Changing Course: Ten Years of Tracking Online Education in the United States"
http://onlinelearningconsortium.org/publications/survey/changing_course_2012

18) National Center for Education Statistics（2011)" Distance Education Courses for Public Elementary and Secondary School Students: 2009-10"
http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2012008

19) Andrew Martin and Andrew W. Lehren (2012) "A Generation Hobbled by the Soaring Cost of College," New York Times,
http://www.nytimes.com/2012/05/13/business/student-loans-weighing-down-a-generation-with-heavy-debt.html?pagewanted=all&_r=o

20) Ifenthaler, D. et al (2014) "−Introduction to the Inaugural Special Issue of Technology, Knowledge and Learning", Tech Know Learn vol 19, pp. 121-126
http://link.springer.com/article/10.1007%2Fs10758-014-9228-2

21) National Center for Education Statistics（2013)" Table 326.10. Graduation rates of first-time, full-time bachelor's degree-seeking students at 4-year postsecondary institutions, by race/ethnicity, time to completion, sex, and control of institution: Selected cohort entry years, 1996 through 2006"
http://nces.ed.gov/programs/digest/d13/tables/dt13_326.10.asp

22) 平成 25 年度文部科学省先導的大学界各推進委託事業「高等教育機関等における ICT の利活用に関する調査研究」
http://www.mext.go.jp/a_menu/koutou/itaku/1347642.htm

23) Next Generation Learning Challenges "Learning Analytics"
http://nextgenlearning.org/topics/learning-analytics

24) Arizona Board of Regents（2008） "2020 VISION: The Arizona University System Long-term Strategic Plan 2008-2020"
http://usenate.asu.edu/files/ABOR_2020.pdf

25) Arizona State University（2008） "A New American University"
http://newamericanuniversity.asu.edu/docs/NAU_Dec10.pdf

26) Buzz Feed News（2014)" The New American University: Massive, Online, And Corporate-Backed"
http://www.buzzfeed.com/mollyhensleyclancy/the-new-american-university-massive-online-and-corporate-bac#27mjga3

27) The Chronicle of Higher Education（2012)" College Degrees, Designed by the Numbers"
http://chronicle.com/article/College-Degrees-Designed-by/132945/

28) Inside Higher Ed（2013)" The New Intelligence"

https://www.insidehighered.com/news/2013/01/25/arizona-st-and-knewtons-grand-experiment-adaptive-learning

29) Campus Technology（2014）” The Great Adaptive Learning Experiment”
http://campustechnology.com/Articles/2014/04/16/The-Great-Adaptive-Learning-Experiment.aspx?Page=1

30) Lytics Lab（2014） “LAW Report”
http://lytics.stanford.edu/law-report/

31) Stanford Online（2014）” Stanford Online 2013 in Review−Harnessing New Technologies and Methods to Advance Teaching and Learning at Stanford and Beyond”
http://web.stanford.edu/dept/vpol/vpol-files/2013_Report/Stanford_Online_2013_In_Review.pdf

32) 日本教育工学会第 29 回全国大会（2013）「課題研究 K03-2-103 教育・学習支援システムにおける Learning Analytics 的アプローチ」
https://www.jset.gr.jp/taikai29/program/program_session.php?tp=K

33) 阿部圭一（2014）「Column：Learning Analytics とは」情報処理学会『ぺた語義』第 36 回
http://www.ipsj.or.jp/magazine/9faeag0000005al5-att/5505.pdf

34) 安武公一（2012）「ライフログの教育活用における海外動向 −Learning Analytics and Knowledge (LAK) 2012 −Learning Analytics and Knowledge (LAK) 2012 報告−」サイエンティフィック・システム研究会 2012 年度教育環境分科会 第 2 回会合
https://www.ssken.gr.jp/MAINSITE/download/newsletter/2012/20121024-edu-2/lecture-02/SSKEN_edu-2012-2_yasutake_paper.pdf

35) MIT News（2014）” MIT and Harvard release working papers on open online courses”http://newsoffice.mit.edu/2014/mit-and-harvard-release-working-papers-on-open-online-courses-0121

36) EU（2014）” Mirror: Reflective Learning at Work, Consortium Partners”
http://www.mirror-project.eu/aboutus/consortium-partners

37) 新浪教育（2013）「新浪 2013 教育盛典：中国教育进入大数据时代」
http://edu.sina.com.cn/l/2013-11-28/1731236680.shtml

3.10.8　社会インフラとビッグデータ

（1）研究開発領域名

社会インフラとビッグデータ

（2）研究開発領域の簡潔な説明

エネルギー、交通、都市、医療、金融など社会インフラを、IT を前提として再設計する動きが加速している。ここでキーになるのは、大量のセンサーから発生するビッグデータの解析であり、このために必要な技術・手法をカバーする。センサーやネットワークなど汎用の基盤技術から、分野特有の応用技術、また社会受容のための文理融合研究など、幅広い研究開発領域となる。

（3）研究開発領域の詳細な説明と国内外の動向

［背景と意義］

エネルギー、交通、通信、都市、医療、金融、教育、社会保障制度など現在の社会システムのほとんどは、IT やインターネットが普及する前にデザインされたものである。IT が普及し始めてからほぼ 60 年の間に、電力の検針と支払い、鉄道や航空機の予約、オンラインバンキングなどに IT が取り入れられてきたが、これらは基本的に今までのデザインを踏襲しながら、そのプロセスの一部を IT によって効率化したものと言える。これらは、より複雑化し、グローバル化によってスケールする社会において、必要なものであり、これからも多くの IT 化が進むだろう。しかしながら、これらの社会システムを、ビッグデータを前提として再設計するならば、はるかに効率が良く、安心・安全で、社会の要請により応えるシステムになる可能性があるだろう [1]。社会システムを根本から変革する可能性のあるという意味で、本研究開発領域には、社会的に大きな意味があり、多くのステークホルダを巻き込んだ、議論が必要である。

技術の観点から社会システムのビッグデータ化を見ると、CPS (Cyber-Physical Systems)[2]あるいは IoT（Internet of Things）のような言葉で特徴づけられる。社会のあらゆる要素にセンサーが埋め込まれ、情報の流れがスムーズになることによって、より効率的で適切な判断ができるようになる。CPS の世界では、センサーからのリアルタイムのデータが大量に発生する。このため、CPS における IT は、必然的にビッグデータ処理に主眼を置いたものになる。

［これまでの取り組み］

1. スマート・グリッド

電力網は、IT を導入することによって効率の大きな向上が期待できる社会インフラである。特に、今後風力・太陽光などの自然エネルギーが電力グリッドに接続されるようになると、これらの不安定な電力ソースを無駄なく利用することは、複雑な最適化と制御の問題になり、ビッグデータの役割が大きい。例えば、デンマークで行われている EDISON プロジェクトでは、風力発電機で生成された余剰電力を一時的に電気自動車に蓄えることによって、生成された電力利用の効率化を図ろうとしている。

2. 交通システム

道路交通は、IT の利用によって、社会制度を含めた大きな変革が期待できる分野である。ロンドンや、ストックホルムなどで行われている道路課金の仕組みは、個別の車両が課金エリアに入ったことをセンサーで検出できるようになって、初めて可能になった。もし、個々の車両の行動の把握とその予測は、リアルタイムなビッグデータ処理であり、これが実現できれば、きめ細かい課金の調整など、柔軟な交通政策を実現できる。

3. スマートシティー

アブダビ首長国で行われている、Masdar プロジェクトは、人口 9 万人の都市をゼロから設計することによって、二酸化炭素を排出しない街をつくろうという壮大な試みである。都市には様々な活動があり、それらを総体的にコントロールできなければ、都市全体のグリーン化はできない。人々の移動や活動に基づく交通・エネルギーの予測を行い、また、気象など、非常に多様なソースからのビッグデータを統合し、全体状況を把握し、需要や生産の将来を予測し、最適なエネルギーの生成・分配計画を立てる必要がある。

ロンドンやニューヨーク、シカゴなどは、ビッグデータを用いて犯罪捜査や犯罪防止に役立てている。防犯カメラなどの大量のセンサーデータと、犯罪の発生データから犯罪を予測し未然に防止したり、センサーデータに基づいてテロリストを早期に発見したりする試みである。

4. センサーとしての人々の活動

社会インフラのセンサーとしては、人間活動に基づくものがあってもよい。国内の事例で面白いのは、コマツ製作所の KOMTRAX である。これは、コマツ製作所の建機に内蔵されている GPS とセンサーから、統合的な車両管理を行うシステムである。このセンサーデータは、単に建機の車両管理に使えるだけでなく、全世界 10 万台を超える建機からの信号を統合することによって得られたビッグデータにより、世界各地域の経済活動指標を得るなど、新たな利用も模索されている。

また、発熱などの検索キーワードからインフルエンザの流行を検出する Google の Flu Trend や、メンバーからの通報によって、きめ細かい気象情報を集めて天気予報に利用する、ウェザーニュース、スマートフォンなどモバイル機器の位置情報を利用して人の混雑状況を予測する Skyhook 社の SpotRank なども、広い意味でのビッグデータに基づく社会インフラと言えるだろう。

［今後必要となる取り組み］

社会インフラは多くの要素技術の統合であり、個別にすべてを網羅することはできないので、特に必要と思われる 4 点の技術について述べる。

1. デバイス・通信・位置情報

ビッグデータに基づく社会インフラには、まずデータを集める仕組みが必要である。これには、センサーデバイス技術とそのセンサーからデータを送る通信技術を含む。センサーは、道路、橋梁やビルなど固定的な構造物に設置されるもの、自動車やウェアラブルセンサーのように移動体に伴うものがある。末端のセンサーで使われる通信は、多くの場合、アドホックネットワークなどの無線通信である。これも非常に低消費電力でかつ低コストなものが求められる。移動体に設置されるセンサーでは、特に位置情報に関するセンサー技術が重要で

ある。自動車等では、GPS が普及しているが、建物内や地下街にいる人の位置情報の把握などについては、まだまだ十分とは言いがたい。

2. アクティブセンサーとしてのロボティクス

ロボティクスを、アクチュエータを持ったセンサープラットフォームと考えると、このアクチュエータを利用して現実世界に影響を与えながら観測することで、パッシブなセンサーでは得られない情報を得ることができる。例えば、交通においては、交通信号のタイミングを変えてドライバーの反応を計測することにより、個々の車両のより精緻なモデル化・予測ができ、それによって、より精密な交通の把握が可能になる。Web マーケティングなどでは、A/B テスティング、品質管理などでは実験計画として知られている手法である。ロボティクスにおいては、周囲の状況を把握するのに実際に動いてみて、あるいは触ってみるということが一般的に行われているが、センサーネットワークにおいてはまだ普及しているとはいえない。

3. オンライン機械学習

社会インフラから発生するビッグデータで特徴的であるのは、それらのほとんどすべてが時系列データであり、それをオンラインで処理しなければならない点である。電力や交通など多くの場合、リアルタイムのセンサーデータに基づいて予測モデルを常にアップデートし、将来の予測を行わなければならない。新たなデータに基づいて逐次的にモデルを更新する技術には、オンライン学習やデータ同化があり、これらの分野の進歩は目覚ましい。これらの技術が上述のアクティブセンサーと結びついて、動的に最適な観測計画を立てられるようになるのは、これからであろう。

4. アーキテクチャーとプログラミングモデル

IT の応用分野と見た時に、社会インフラのビッグデータはユニークなチャレンジとなる。センサーからは大量のデータが発生するが、それらは必ずしも信頼できるものではない。センサーの取り付け不良、経年変化、環境の変化などあらゆる理由でノイズの乗ったものになる [3]。また、センサーだけでなく通信などの障害により欠測値が出ることも多い。さらに、センサーからの生データをすべてクラウドなどのサーバーに送ることは、データ量の点で現実的でない。すなわち、ほとんどのデータがネットワークのエッジで蓄積・処理されるエッジ・ヘビー・データ [4]のアーキテクチャーが必要とされるだろう。

（4）科学技術的・政策的課題

社会インフラに関する議論は、自治体などの政府関係者、電力会社などの事業者、それに利用者たる市民や企業などの多くを巻き込んだ議論にならざるを得ない。このため、科学技術的課題、政策的課題は、どちらも包括的な議論の必要なものである。

1. 分野横断・システム思考・ライフサイクル

社会システムは単独の技術や手法だけで実現するものではない。多くの要素技術を統合し、継続的に運用していく包括的な視点が必要となる。このため、ビッグデータに基づく社会インフラの推進のためには、情報分野の研究者だけでなく、交通・都市・医療などの個別分野の研究者、また社会インフラの受益者に与える影響などを考えると、社会科学の研究者との分野横断的な研究開発体制を取らざるを得ない。

また、我が国においては、個別技術は重視するがシステム思考に関してはその重要性が十

分に理解されているとは言い難い。特に、社会インフラにおいては個別技術の優位性ではなく、システム全体のバランスのとれた設計をすることが肝要である。このため、システム科学の研究を推進すると同時に、システム的思考のできる研究者・開発者の育成が急務である。

さらに、社会インフラは設計されてから何十年も利用されるのが普通であるので、その運用・ライフサイクルに関する手法の研究開発が欠かせない。社会インフラがビッグデータ化するということは、社会インフラがソフトウエア化するということであり、そこではシステムの仕様が刻々と変化していくということでもある。今までの設計・製造と運用が分離された「ものづくり」パラダイムではなく、DevOps など IT で培われた継続的運用のパラダイムを社会インフラに持ち込む必要がある。

2. オープンデータとプライバシー保護

政策・制度的な観点からは、社会インフラでビッグデータを利用することの最大の困難は、データの利活用をいかに促進するかである。特に、プライバシーに関する議論は十分に行わなければならない。

（５）注目動向（新たな知見や新技術の創出、大規模プロジェクトの動向など）

［新たな技術動向］

1. ディープラーニング

機械学習の分野では、コンピュータの処理速度の向上によって初めて可能になった多段ニューラルネットワークによる機械学習手法「ディープラーニング」[5]が注目を浴びている。ディープラーニングの応用は特に画像理解の分野でめざましく、センサーから得られた画像から、そこに写っている車や人、その他の物体を認識することが高精度で行えるようになっている。

2. ストリーミングビッグデータ向け処理基盤

社会インフラに埋め込まれたセンサーから生成されるビッグデータは必然的にストリーミングデータであり、これらのほとんどはオンライン処理されなければならない。このための処理基盤として、CEP (Complex Event Processing)と呼ばれるイベント駆動型のプログラミングモデルがある。オープンソースでは、Apache S4 があるほか、IBM、Oracle なども CEP 製品を出荷している。

CEP はルールに基づく演繹的処理を行うものだが、帰納的な処理を行う基盤として、上述のオンライン機械学習がある。分散オンライン機械学習を行う日本発のプラットフォームとして、オープンソースの Jubatus[6]がある。

［注目すべきプロジェクト］

NEDO の IT 融合プロジェクト

ビッグデータ処理に基づく社会インフラとして注目すべきプロジェクトが、自動運転に関するものである。Google の自動運転車を始め、各国の自動車会社が力を入れている。

（６）キーワード

センサーネットワーク、アドホックネットワーク、ディープラーニング、圧縮センシング、オンライン機械学習、CEP、高効率インバーター、電力変換、超高圧送電、HVDC

（７）国際比較

国・地域	フェーズ	現状	トレンド	各国の状況、評価の際に参考にした根拠など
日本	基礎研究	○	→	大規模データベース、機械学習、データ同化などの基礎研究において、世界レベルの研究がある。
	応用研究・開発	○	→	オンライン分散機械学習フレームワークJubatusをいち早くオープンソースで世に出しており、この分野では一定の存在感がある。NEDOのIT融合プロジェクトは2年で打ち切りになった。柏の葉キャンパスシティプロジェクトなどの実証実験は行われているが、注目度はあまり高くない。
	産業化	△	→	日立製作所などが積極的だが、大きなビジネスとしてはまだ育っていない。オープンデータに関しては流山市などの一部の自治体が取組[7]を始めているが、まだ大きな流れにはなっていない。国レベルでは、総務省のe-Statがあるが、統計データであり、個別データの利用促進はまだ緒についたばかりである。
米国	基礎研究	◎	→	機械学習、 CPSなど広い分野で底力のある基礎研究を続けていて、層は厚い。ディープラーニングや圧縮センシングなどの新しいアイデアも米国発である。
	応用研究・開発	◎	→	CPS、ビッグデータではそれぞれNSFが巨額の研究資金を投じている。Googleは自動運転車を通してそのビッグデータの知見を社会インフラに適用しはじめている。
	産業化	◎	→	GEのIndustrial Internetなど、インフラ系産業がビッグデータに乗り出している。また、CISCO、 IBM、 Intel、 OracleなどIT系ベンダも、IoTに力を入れていて、CEPなどのプロダクトをいち早く商業化している。
欧州	基礎研究	◎	→	
	応用研究・開発	○	↗	FP7
	産業化	◎	→	ABB、 シーメンスなどのインフラ系産業が強い
中国	基礎研究	○	↗	学会でのプレゼンスは急速に増えている。
	応用研究・開発	○	↗	北京市の人流データ(Microsoft) 吉林市、ビッグデータを活用した渋滞予測・信号制御シミュレーションの実証実験(NTT Data)
	産業化	△	↗	インフラ系産業の成長はまだこれから。したがって、そこにおけるビッグデータ利用にも時間がかかるものと思われる。
韓国	基礎研究	△	↗	
	応用研究・開発	△	↗	松都（ソンド）市におけるスマートシティーの取組
	産業化	△	→	住民登録番号が普及していて、個人のマッチングがとりやすい。スマホは競争力あるが、インフラ系は？

（註 1）フェーズ
基礎研究フェーズ ：大学・国研などでの基礎研究のレベル
応用研究・開発フェーズ ：研究・技術開発（プロトタイプの開発含む）のレベル
産業化フェーズ ：量産技術・製品展開力のレベル

（註 2）現状
※我が国の現状を基準にした相対評価ではなく、絶対評価である。
◎：他国に比べて顕著な活動・成果が見えている、 ○：ある程度の活動・成果が見えている、
△：他国に比べて顕著な活動・成果が見えていない、×：特筆すべき活動・成果が見えていない

（註 3）トレンド
↗ ：上昇傾向、 →：現状維持、 ↘：下降傾向

（８）引用資料

1) McKinsey Global Institute, Big data: The next frontier for innovation, competition, and productivity, http://www.mckinsey.com/insights/mgi/research/technology_and_innovation/big_data_the_next_frontier_for_innovation, 2011.
2) The National Science Foundation, "Cyber-physical systems (CPS): Program Announcements & Information," http://www.nsf.gov/publications/pub_summ.jsp?ods_key=nsf08611, 2008.
3) Candès, Emmanuel J.; Romberg, Justin K.; Tao, Terence (2006). "Stable signal recovery from incomplete and inaccurate measurements". Communications on Pure and Applied Mathematics 59 (8): 1207.
4) Daisuke Okanohara, Shohei Hido, Nobuyuki Kubota, Yuya Unno, and Hiroshi Maruyama, “Krill: An Architecture for Edge-Heavy Data,” Third Workshop on Architectures and Systems for Big Data, Tel Aviv, June, 2013.
5) Bengio, Yoshua (2009). "Learning Deep Architectures for AI". Foundations and Trends in Machine Learning 2 (1).
6) http://jubat.us/en/
7) 柏の葉スマートシティ, http://www.kashiwanoha-smartcity.com/

３.10.９　オープンデータ

（１）研究開発領域名

オープンデータ

（２）研究開発領域の簡潔な説明

オープンデータとは、最小限の制約のみで誰でも自由に利用、加工、再配布ができるデータのことである。インターネットの普及に伴い、各分野でデータの利活用が行われるようになっているが、特にインターネットとの親和性が高いデータ公開のあり方として前述のような定義がなされている。オープンデータにおける公開の方法は、政策面、法律・社会制度面、技術面において整備される必要があり、それぞれの分野において解決すべき課題があり、これが研究テーマである。政策面では公的組織や学術組織におけるオープンデータの位置づけや制度化が課題である。法律・社会制度面では、オープンデータのためのライセンスの確立が課題である。技術面では、再利用性を高める機械可読フォーマットやその処理、また活用のプラットフォームなどが課題である。

（３）研究開発領域の詳細な説明と国内外の動向

オープンデータという概念はいくつかの異なる領域での活動が収斂する形で現在ある。ひとつは政策としてのオープンガバメント（開かれた政府）を実現するための 1 方策として現れている。アメリカのオバマ大統領は電子政府のあり方として 2009 年 12 月に Open Government Directive を発表して、アメリカ政府がオープンガバメントを進めることを宣言した。このオープンガバメントは主に透明性（Transparency）、参加（Participation）、協同（Collaboration）の３つの政策からなる。この中で特に透明性の実現には政府が盛っている情報をできるだけ生のまま公開することが必要であると考え、これをオープンデータとして実現するとした。Web2.0 を提唱したティム・オライリーはこの考えを単に Web が我々の生活世界が変わるだけでなく、政府も同じように変わるべきだととらえ、Goverment2.0 を提唱したことと軌を一にしており、彼は Government 2.0 と読んだ。実際、アメリカ政府は data.gov をというサイトを立ち上げ、そこで様々なデータを公開していった。これは政策面から定義されるオープンデータである。

一方、オープンソースソフトウエアは Linux を初めとする数々の社会的に重要なソフトウエアを生み出し、技術者において基本的な考え方として広く受け入れられてきた。この考え方の延長にソフトウエアだけでなくデータもオープンになるべきだと考えるようになった。データはだれでも使えて再利用できるべきだと考える。これは技術から定義されるオープンデータである。また、オープンソースソフトウエアに関わる人々が実際にオープンデータの活動にも関わることもできてきた。

これとは別に学術界では学術成果は社会で広く共有すべきであるというオープンアクセス（OA）という考え方が近年広まってきた。これは有料でしか読めない論文では学術成果は共有できないので、自由に読めるようなオープンな形で論文を公表すべきだという考え方である。有名なものとしてはブダペスト・オープン・アクセス・イニシアチブで、これは 2002 年に発表されたもので、学術文献をインターネット上で自由に利用できるようにすることを目的とした宣言である。学術論文は特にこの OA の方向で、無料で自由に閲覧できる

という方向に進んでいる。さらには近年のデータ科学の発展に伴い、研究成果は単に文献としての論文ではなく、データとして公開・利用するという方向に向かっており、オープンデータと同じ方向にある。

現在のオープンデータに関わる活動は研究から社会まで多様なセクターの人を巻き込んだ複合的な活動である。これを俯瞰するのは容易ではないが、ここでは政策、技術、コミュニティの３つの視点からまとめる。

（３－１）政策

アメリカおよびイギリスが先行している。アメリカは最初に政府のオープンデータポータル data.gov を公開している。イギリスはそのあとに data.gov.uk を開設している。ただし、そのあとイギリスは国を挙げてオープンデータに積極的に取り組んでいる。特に Open Data Institute を産官で設立して、国内外でのオープンデータを推進している。ヨーロッパは EU が公共セクター情報(Public Sector Information)の再利用について 2000 年前半から取り組んでおり、歴史は長い。日本は 2012 年に内閣官房 IT 戦略本部が「電子行政オープンデータ戦略」を発表している。

（３－２）技術

オープンデータに関する技術開発はアカデミア研究とオープンソースコミュニティにおける開発と標準化の３つに大別されるが、相互に関係し合っている。アカデミア研究では主にセマンティック Web の分野で行われている。特に高可用性に有用な Linked Open Data (LOD)/Linked Data はこの分野である。LOD に関係する技術としては、RDF(Resource Description Framework)、 RDFS(RDF Schema)、 OWL(Web Ontology Language)といった記述言語、問い合わせ言語 SPARQL、RDF データベースマネジメントシステム（triple store）などがあげられる。これらは多くは World Wide Web Consortium (W3C)に標準として提案、採択されている。

LOD に関しては欧州の研究グループが特出している。中でもドイツの研究者、研究グループが活発である。EU の IST 7th Framework では多くのセマンティック Web 関連のプロジェクトがあり、研究開発を進めている。重要なものでは、LOD プロジェクトや LOD2 プロジェクト [1)]がある。LOD2 プロジェクトでは LOD のライフサイクル（生成から蓄積、利用など）を支える技術・ソフトウエアを開発している。Wikipedia を LOD 化した DBpedia やその利用ツール DBepdia Spotlight、Silk などがこのグループおよびその周囲で作られている。

オープンデータに関わるソフトウエア開発はほとんどがコミュニティベースで行われている。特にイギリスに本拠をもつ Open Knowledge Foundation (OKF)は様々なオープンデータ関連のソフトウエア開発の中心となっている。例えば、データカタログサイトのソフトウエア CKAN や税金の可視化の openspending.org などがその例である。

（３－３）コミュニティ

このような政府の動きおよび技術開発をつなぐのに重要なのがコミュニティである。主に非営利組織を母体に技術者や市民を募り、オープンデータ化の推進や利用促進の活動を行っている。イギリスでは先にあげた OKF が組織としては著名である。日本でも OKF Japan やリンクト・オープン・データ・イニシアチブ、Linkdata といった組織が活動をしている。例えば、これらはハッカソン、アイデアソン、コンペティションなどを企画、運営し、オー

プンデータを広める活動をしている。

（４）科学技術的・政策的課題

・各国において政府自身のオープンデータ政策が大きな要素となる。よいオープンデータ政策がオープンデータに関わる研究活動を活性化しうるし、また逆によい研究活動は国のオープンデータ政策を左右する。例えば、イギリスは Tim Berners-Lee がオープンデータ政策に関わっており、ODI 設立に至り、また ODI がオープンデータ政策に影響を与えている。
・Linked Open Data はまだ early adaptor 的位置づけでそれほど一般化していないが、オープンデータの技術として期待されている（他に対抗的技術はない）。欧州では引き続き研究投資は続けられる。ここにどうキャッチアップするかが課題である。

（５）注目動向（新たな知見や新技術の創出、大規模プロジェクトの動向など）

・RDA(Research Data Alliance)は科学におけるデータ共有がテーマであるが、オープンデータとも強く関わる。RDA の動きに注目する必要がある。
・イギリスの ODI は各国に Node という形で関連組織を増やしている。日本では大阪、アジアでは他にはソウルにあり、ODI の影響に注意が必要である。
・IBM の Watson も Web 上のデータを使っており、オープンデータと近い位置にある。今後、つながってくることが予想される。

（６）キーワード

オープンフォーマット、オープンライセンス、Linked Open Data、Linked Data、セマンティック Web

（７）国際比較

国・地域	フェーズ	現状	トレンド	各国の状況、評価の際に参考にした根拠など
日本	基礎研究	△	→	セマンティックWeb研究は一通りあるが、オントロジー研究以外は、相対的に弱い。
	応用研究・開発	△	↗	これまでは散発的であったが、多くの人が興味を持ち始めている。
	産業化	△	↗	コミュニティ活動は盛んになっている。これからは実用化事例も期待できる。
米国	基礎研究	○	→	人工知能研究やWeb研究の強みがあるので、興味深い研究もでてくるが、全体としての動きにはなっていない。
	応用研究・開発	○	→	オープンソースのコミュニティとの連携が強み。
	産業化	◎	↗	ビジネス化の例がいくつかできつつある。
欧州	基礎研究	◎	↗	セマンティックWeb分野の研究の本拠。EUのISTプロジェクトで多くの研究がなされている。
	応用研究・開発	◎	↗	OKF主導によるソフトウエアやサービスが多数生み出されている EuropeanaのLOD版といった公的セクターがオープンデータ化を産官学で進めている
	産業化	◎	↗	英国のODIを中心に産業化がなされている。
中国	基礎研究	△	→	セマンティックWebの研究は少しある。
	応用研究・開発	×	→	政治体制的にオープンデータはそぐわない
	産業化	×	→	政治体制的にオープンデータはそぐわない
韓国	基礎研究	△	→	セマンティックWebの研究は少しあるが、特出するものはない。
	応用研究・開発	△	→	特段の活動は見受けられない。
	産業化	△	↗	ソウル市の例など実用的な展開があるが、まだ点的広がり。

（註 1）フェーズ
基礎研究フェーズ ：大学・国研などでの基礎研究のレベル
応用研究・開発フェーズ ：研究・技術開発（プロトタイプの開発含む）のレベル
産業化フェーズ ：量産技術・製品展開力のレベル
（註 2）現状
※我が国の現状を基準にした相対評価ではなく、絶対評価である。
◎：他国に比べて顕著な活動・成果が見えている、 ○：ある程度の活動・成果が見えている、
△：他国に比べて顕著な活動・成果が見えていない、×：特筆すべき活動・成果が見えていない
（註 3）トレンド
↗ ：上昇傾向、 →：現状維持、 ↘：下降傾向

（８）引用資料

1）http://lod2.eu/

3.10.10　著作権とビッグデータ

（1）研究開発領域名

著作権とビッグデータ

（2）研究開発領域の簡潔な説明

ビッグデータの収集や解析結果の表示に伴う著作権問題

（3）研究開発領域の詳細な説明と国内外の動向

ビッグデータ解析に伴う問題は、①ビッグデータ収集時に問題となる「元データと著作権」の問題と　②解析結果を表示する時に問題となる「解析結果と著作権」の問題に二分される。以下、条文は特にことわらないかぎり日本の著作権法の条文である。

ビッグデータを収集する際に問題となる元データと著作権の問題

・元データに創作性がなければ、著作物にあたらないので、著作権はない（2 条 1 項 1 号、6 条)。このため許諾を得ずに利用しても著作権上の問題は発生しない。行動履歴（ライフログ）などがこれにあたる。創作性があれば、著作権があるため、原則として許諾が必要となる。ただし、著作権法は許諾を得ずに利用できる権利制限規定を列挙している。この権利制限規定に該当すれば許諾はいらない。ビッグデータに関連する権利制限規定としては以下の規定があげられる。
・検索サービスのための複製等（47 条の 6)：検索サービスを提供するにあたり、検索エンジンがウェブサイトの情報を収集するために行う複製等
・情報解析のための複製等（47 条の 7)：電子計算機による情報解析を行うための記録媒体への記録等 [1)]
・情報通信技術を利用した情報提供の準備に必要な情報処理のための利用（47 条の 9)：例えば SNS などのサービス・プロバイダーがユーザーの投稿したコンテンツを配信する際、サーバー上で分散処理するために行う複製等

ビッグデータ解析結果を表示する際に問題となる解析結果と著作権の問題

ここでも二つの側面がある。①成果物が元データの著作権を侵害しないかという問題と②成果物に著作権が発生するかという問題である。

・成果物が元データの著作権を侵害しないかは、成果物が元データの表現を利用する際に問題になる。元データに創作性があれば、著作権が発生するため、権利制限規定で認められている「引用」にあたらないかぎり著作権者の許諾を得る必要がある。引用は、「公正な慣行に合致するものであり、かつ、報道、批評、研究その他の引用の目的上正当な範囲内で行われるものでなければならない。」(32 条①項)。
・成果物に著作権が発生するかについては、編集著作物に該当すれば著作権が発生する。編集著作物は「編集物（データベースに該当するものを除く）でその素材の選択又は配列によって創作性を有するもの」(12 条)。編集物のうちデータベースについては、「情報の選択又は体系的な構成によって創作性を有するものは、データベースの著作物として著作権が認められる」(12 条の 2　①項)。創作性がなければ著作物ではないので、著作権は認め

られない。具体例として裁判例を紹介すると、京都大学学位論文事件 [2]で、「実験結果等のデータ自体は、事実又はアイデアであって、著作物ではない以上、そのようなデータを一般的な手法に基づき表現したのみのグラフは、多少の表現の幅はあり得るものであっても、なお、著作物としての創作性を有しないものと解すべきである」とされた（知財高判平成 17.5.25）。

・以上はビッグデータ解析に伴う著作権問題だが、解析結果の活用に伴う著作権問題もある。著作権で保護されるには何らかの創作性が必要で、単なる事実では著作権は発生しないが、企業が解析結果に対して、著作権などを理由に契約上の権利を主張して囲い込みを図る傾向がみられる。著作権法がわかりにくいこと、企業の法令遵守・コンプライアンス意識が高いことなどから、日本ではこうした「疑似著作権」の主張がまかり通りやすい土壌があることもビッグデータ解析結果の活用を妨げている。

（4）科学技術的・政策的課題

・「知のビッグデータ」ともいえる図書館・公文書館・博物館・美術館の収蔵品をデジタル化して、保存するデジタル・アーカイブが欧米で進んでいる。デジタル・アーカイブ化する際の大きな障害が孤児著作物（Orphan Works）問題である。孤児著作物は権利者の死亡などにより、著作権者が不明の著作物。欧州連合は 2012 年に、「孤児著作物の利用に関する指令」を採択した。この指令が画期的なのは、著作権者を探す努力をして見つからない場合、非営利目的であれば使えるところにある。著作物を使用する際は著作権者の許諾を得る、つまりオプトイン（原則禁止）が著作権法の大原則である。これをひっくり返して、著作権者が反対しないかぎり、つまりオプトアウトしないかぎり使用を認める（原則自由とする）わけである

・日本でも孤児著作物の二次利用をしやすくするため、文化庁に供託金を支払い、文化庁長官が著作者に代わって許諾する制定制度がある。この制度をより使いやすくするための改正が 2009 年にも行われたが、その後の年間裁定件数も 20～30 件にすぎない。オプトインの限界ともいえる。

・米国では孤児著作物を利用しやすくする法案が 2000 年代に二度にわたって議会に提案されたが、成立に至らなかった。その間隙をついたのが、私企業のグーグルである。グーグルは 2004 年に書籍検索サービスを発表した。出版社や図書館から書籍を提供してもらった書籍をデジタル化し、全文を検索して、利用者の興味にあった書籍を見つけ出すグーグル・ブックスとよばれるこのサービスに対して、2005 年に全米作家協会などが著作権侵害訴訟を提起した。グーグルは書籍をスキャンして、検索サービスのデータベースを作成するが、これを著作権者の同意なしに行うことは、著作権者の持つ複製権を侵害すると主張した。検索結果は「抜粋（スニペット）表示」とよばれ、ウェブ検索同様、検索語を含む数行だけしか表示されない。グーグルはこの抜粋表示は米国著作権法に定めるフェアユースであると反論した。

・著作権法は著作物の利用と保護のバランスを図ることを目的とした法律である（1 条）。著作物の利用には著作権者の許諾を要求して保護する一方、許諾がなくても使用できる権利制限規定を設けて利用に配意している。わが国の著作権法はこの権利制限規定を「私的使用のための複製」（30 条）、「引用」（32 条）など個別具体的に列挙しているが、米国は

使用する目的がフェア（公正）であれば、許諾なしの使用を認める権利制限の一般規定として、「フェアユース」規定（米著作権法 107 条）を置いている。フェアな使用であるかどうかは、使用目的や使用される著作物の市場に与える影響（市場を奪わないかどうか）などの４要素を総合的に考慮して判定するよう定めている。

・2008 年に全米作家組合とグーグルは和解案を発表した。和解案は当初、全世界の著作権者を対象にしたため、日本の出版業界にも電子書籍の黒船騒ぎが起きた。その後、米国以外の対象国を英国および旧英領諸国に絞ったため、日本は対象外となった。その修正和解案も 2011 年に裁判所が承認しなかったため、訴訟に復帰した。ニューヨークの連邦地裁は 2013 年にグーグルのフェアユースの主張を認める判決を下した [3]。まだ地裁段階だが、グーグルに書籍を提供した図書館が訴えられた訴訟では、控裁でもフェアユースが認められた。このようにグーグルはフェアユース規定を使ってオプトアウトを実現し、孤児著作物問題を解決しつつあるといえる。

・英語圏の書籍を多量にデジタル化するグーグル・ブックスは、英語文化の世界支配にもつながる。これに脅威を抱いた欧州は、グーグル・ブックスの発表直後にフランスのジャンヌネー国立図書館長がこの問題を指摘。それに応えてシラク大統領が 5 カ国の首脳に呼びかけ、欧州委員会がデジタル・ライブラリー計画を策定、2008 年に欧州デジタル図書館（Europeana）を一般公開し、現在までに 3000 万冊以上をデジタル化した。2012 年には「孤児著作物の利用に関する指令」[4]を採択。この指令は上記のとおり、オプトアウトの発想を採用している点でグーグル・ブックスと共通するが、書籍以外の著作物（画像、動画、音楽など）も対象にしている点ではグーグル・ブックスを上回る「知のビッグデータ」であり、孤児著作物の利用に関しても世界をリードしている。このように米国の一民間企業のプロジェクトにすぎないグーグル・ブックスが、欧州に与えたインパクトは大きかった。

・日本でも必要の都度、権利制限規定を個別に追加する方式では法改正に時間がかかり、著作物の利用形態が急激に変化するデジタル・ネット時代に追いつけないとの認識が 2000 年代後半に高まった。このため、知財本部は知財計画 2008 で「包括的な権利制限規定」の導入を提言、これを受けて文化庁が検討した。検討結果をまとめた 2011 年の文化審議会著作権分科会報告書 [5]は、3 類型の利用行為を権利制限の一般規定による権利制限の対象とするよう提言した。3 類型の中にはビッグデータに関連する「ネットワーク上で複製等を不可避的に伴う情報ネットワーク産業のサービス開発・提供行為のような『著作物の表現を享受しない利用』」のような類型も含まれていた。

・報告書の内容は内閣法制局の審査を経て条文化され、2012 年の法改正に盛り込まれ、「情報通信技術を利用した情報提供の利用」（47 条 9、上記（３）参照）を含む二つの条文に落とし込まれた。その過程で「著作物の表現を享受しない利用」という表現も削除された。この文言が生きていれば日本版フェアユース規定とよぶにふさわしい規定が誕生したわけだが、落とされてしまったために「権利制限の一般規定」というより、これまでの法改正でも追加してきた「個別の権利制限規定」と大差ない改正に終わってしまった。

・文化審議会著作権分科会は 2009 年度報告書 [6]で検索エンジンの法制上の課題について検討、権利制限を講ずることが適切であると報告した。これを受けた 2009 年の法改正で、検索サービスのための複製等（47 条の 6）（上記（３）参照）の権利制限規定が追加され

た。同報告書はコンピューター・プログラムのリバース・エンジニアリングについても、相互運用性の確保や障害の発見等の一定の目的のための調査・解析についての権利制限措置の必要性については概ね意見の一致が見られたと報告した。ところが、2009 年の法改正には盛り込まれず、その後、2 度にわたる法改正でも実現していない。

・2008 年に文化庁が実施した関係者からのヒアリングでは、リバース・エンジニアリングの法的扱いが不明であることを認知している企業の中には、コンプライアンスの観点から、海外で解析している企業もあり、わが国に情報セキュリティー技術者のスキルが高まらず、世界の情報セキュリティー・ビジネスの中で、日本が比較劣位に置かれる一つの要因にもなっているとの指摘もあった。米国では 1990 年代はじめの 2 件の控裁判決でリバース・エンジニアリングのフェアユースが認められた。その米国に 20 年遅れてもいまだにリバース・エンジニアリングを合法化できないようでは、秒進分歩の IT の世界ではビジネス化を断念せよと宣告されるようなものである。

・かつて検索エンジンも、フェアユース規定のないわが国では、著作権侵害の恐れを回避するため、事前に検索するウェブサイトの了解を取る「オプトイン方式」を採用した。これに対して米国では、自分のウェブサイトを検索されたくない場合には、その旨を意思表示すれば、検索を技術的に回避する手段を企業側が用意する「オプトアウト方式」で対応した。検索サービスは情報の網羅性、包括性が命であるだけに、オプトインしたサイトしか検索できないサービスとオプトアウトしないかぎり、検索可能なサービスとの差は決定的で、日本も 2009 年の上記法改正で個別権利制限規定を追加し、検索エンジンを合法化したが、時すでに遅し。日本の著作権法の適用を逃れるため米国内にサーバーを置き、日本にサービスを提供した米国勢に日本市場まで制圧されてしまった。

・ビッグデータの宝庫といえる国や自治体の持つデータをネットで公開し、利用できるようにするオープンデータについても著作権の制約がある。米国では連邦政府が作成した著作物には原則として著作権は発生しない（米著作権法 105 条）。国民の税金を使って作った著作物は国民のモノという考え方に根ざしている。日本では政府情報の中でも法令や裁判所の判決などは著作権の対象にはならないが（12 条）、白書などの著作権は政府にある。政府も国税を使って作成したものは国有財産であるとの考え方から、契約で著作権を国に帰属させようとしてきた。このため、宝の山といわれる行政データを活用してオープンデータ政策を推進するにあたっても、専用サイトで膨大なデータを公開している米国（後記（5）参照）などに大きく水をあけられている。

・知財計画 2008 が最初に包括的権利制限規定の検討を提言してから 6 年経過した。この間、デジタル・アーカイブ、リバース・エンジニアリング、オープンデータなどの取り組みで欧米に遅れを取っていることが判明するなど、その後の情勢も急激に変化しているため、こうした問題の有効な解決策でもあるフェアユース規定の導入を再考すべきである。

・著作権法とともにビッグデータ活用の鍵を握る個人情報保護法については改正の動きが具体化しつつある。

（５）注目動向（新たな知見や新技術の創出、大規模プロジェクトの動向など）

・2013 年の Ｇ８で「オープンデータ憲章」が発表される [7]など、各国のオープンデータに対する取り組みが活発化しているが、欧米の取り組みは早かった。EU は 2003 年に「公的機関の情報の再利用に関する指令」を公表。2013 年にはその改訂版 [8]を、2014 年にはガイドラインを発表した。米国では 2008 年の大統領選でオープンガバメントを公約に掲げたオバマ大統領が、その一環として就任直後からオープンデータに取り組み、行政データ公開専用サイトの「data.gov」サイトを世界で最初に立ち上げた [9]。

・日本では 2012 年に IT 総合戦略本部が公表した「電子行政オープンデータ戦略」で、公共データの活用促進について定め、関係省庁の取り組みが活発化した。2013 年には「世界最先端 IT 国家創造宣言」[10]を閣議決定し、2020 年までに「世界最高水準の IT 利活用社会を実現する」ことを目標にした。IT 総合戦略本部も「電子行政オープンデータ推進のためのロードマップ」[11]を発表した。これに従って、行政データ公開専用サイト「data.go.jp」が 2014 年 10 月から本格的にサービスを開始した。

（６）キーワード

権利制限規定、情報解析、編集著作物、データベースの著作物、デジタル・アーカイブ、フェアユース、孤児著作物、オプトイン、オプトアウト、裁定制度、グーグル・ブックス、欧州デジタル図書館（Europeana）、孤児著作物の利用に関する EU 指令、オープンデータ憲章、オープンガバメント、世界最先端 IT 国家創造宣言

（７）国際比較

国・地域	フェーズ	現状	トレンド	各国の状況、評価の際に参考にした根拠など
日本	基礎研究	○	→	・情報解析のための複製等を認める個別の権利制限規定がある。
	応用研究・開発	○	↗	・2013年に「世界最先端IT国家創造宣言」を閣議決定、2020年までに「世界最高水準のIT利活用社会を実現する」ことを目標に掲げた。
	産業化	○	↗	・上記により、自治体などのオープンデータへの取り組みが活発化、ビジネスへの波及が期待される。
米国	基礎研究	◎	→	・フェアユース規定、連邦政府が作成した著作物には原則として著作権を認めない規定などがある。
	応用研究・開発	◎	↗	・オバマ政権は2012年にビッグデータ研究開発イニシアチブを発表、研究機関に総額2億ドル（220億円）の補助金を投入。 ・オープンデータについては2008年の大統領選でオープンガバメントを公約に掲げたオバマ大統領が、その一環として就任直後から取り組み、行政データ公開専用サイトの「data.gov」サイトを世界で最初に立ち上げた。
	産業化	◎	↗	・「data.gov」を活用したビジネスが多数出現。 ・グーグル・ブックスも書籍ビッグデータの活用例。
欧州	基礎研究	◎	→	・EU は2003年に「公的機関の情報の再利用に関する指令」を公表。2012年には「孤児著作物の利用に関する指令」を採択。 ・英国は2014年の著作権法改正で情報解析のための個別権利制限規定を導入。
	応用研究・開発	◎	↗	・EUはIT業界が設立した非営利団体ビッグデータバリュー協会とともにPublic Private Partnership（PPP）を2015年に立ち上げ、2020年までに民間のビッグデータ関連投資に25億ユーロ（3400億円）を拠出予定。 ・英国は2005年に「公共データの再利用に関する規則を制定。
	産業化	○	↗	・上記により、ビッグデータ活用ビジネスの本格化が期待される。
中国	基礎研究	○	→	・フェアユース規定も情報解析のための個別の権利制限規定もないが、フェアユースについては検討した判例がある。ただし、認めた判例は今のところない。
	応用研究・開発	○	↗	・2012年に中国共産党第18回大会で「工業化・情報化・都市化・農業の近代化」の推進の方針が示され、ビッグデータやオープンデータの利活用が取り上げられた。 ・これを受け、2013年に工業・情報化部によって「情報化及び工業化の深度融合プロジェクト・アクションプラン（2013～2018年）」が発表された[12)]。
	産業化	○	↗	・上記により、ビッグデータ活用ビジネスの本格化が期待される。
韓国	基礎研究	◎	→	・2011年の著作権法改正でフェアユース規定を導入。
	応用研究・開発	◎	↗	・2012年に国家情報化戦略委員会が「スマート国家具現のためのビッグデータ・マスタープラン」をまとめ、公共分野主導でビッグデータ活用を促進する方針を発表。 ・同プラン実施のための予算は2016年までに官民合わせて5000億ウォン（520億円）と発表され、これを受け、2013年から次々と各省がビッグデータ戦略を策定。
	産業化	○	↗	・上記により、ビッグデータ活用ビジネスの本格化が期待される。

（註１）フェーズ
基礎研究フェーズ ：大学・国研などでの基礎研究のレベル
応用研究・開発フェーズ ：研究・技術開発（プロトタイプの開発含む）のレベル
産業化フェーズ ：量産技術・製品展開力のレベル
（註２）現状
※我が国の現状を基準にした相対評価ではなく、絶対評価である。

◎：他国に比べて顕著な活動・成果が見えている、　○：ある程度の活動・成果が見えている、
△：他国に比べて顕著な活動・成果が見えていない、×：特筆すべき活動・成果が見えていない
（註 3）トレンド
↗ ：上昇傾向、　→：現状維持、　↘：下降傾向

（８）引用資料

1）末吉亙「情報解析と著作権」
https://www.jstage.jst.go.jp/article/johokanri/55/6/55_434/_article/cited-by/-char/ja/
2）安東 奈穂子「著作権法のもとでの情報解析」
http://ci.nii.ac.jp/naid/110007731101
3）グーグルの書籍検索サービス合法判決でますます拡大する日米格差
（その 1）http://agora-web.jp/archives/1569233.html
（その 2）http://agora-web.jp/archives/1570001.html
4）EU Orphan Works Directive
http://ec.europa.eu/internal_market/copyright/orphan_works/index_en.htm
5）文化審議会著作権分科会報告書（2011 年 1 月）
http://www.bunka.go.jp/chosakuken/singikai/bunkakai/33/pdf/shiryo_4_2.pdf
6）文化審議会著作権分科会報告書（2009 年 1 月）
http://www.bunka.go.jp/chosakuken/pdf/21_houkaisei_houkokusho.pdf
7）Ｇ８サミットにおけるオープンデータに関する合意事項
http://www.kantei.go.jp/jp/singi/it2/densi/dai4/sankou8.pdf
8）Directive 2013/37/EU of the European Parliament and of the Council of 26 June 2013 amending Directive 2003/98/EC on the re-use of public sector information.
http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32013L0037&rid=1
9）米国オープンデータの動向調査
http://www.ipa.go.jp/files/000033718.pdf
10）世界最先端 IT 国家創造宣言（2014 年 6 月　改訂版）
http://www.kantei.go.jp/jp/singi/it2/kettei/pdf/20140624/siryou2.pdf
11）平成 26 年版　情報通信白書
http://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h26/pdf/index.html
12）上原伸一・馬鉄「中国著作権法第３次改正案（2014 年 6 月 06 日公表）中日対訳検討表」(発行元：あみのさん,2014 年)

3.10.11　ビッグデータとプライバシー

（1）研究開発領域名

ビッグデータとプライバシー

（2）研究開発領域の簡潔な説明

ビッグデータに含まれるパーソナルデータの利活用とプライバシー・個人情報の保護

（3）研究開発領域の詳細な説明と国内外の動向

ビッグデータの中でも、個人に関する情報ないしパーソナルデータを活用する場合には、プライバシー・個人情報保護に関する課題が発生する。そのようなパーソナルデータとしては、例えば、大量のネット上の閲覧履歴、購買履歴、アプリケーションの利用履歴、書き込み内容や、現実空間における位置情報、移動履歴などがある。

ここでは、まず、プライバシー・個人情報保護法制の基本的な動向について述べる。我が国では、2003 年に個人情報保護法を含む個人情報保護関連 5 法が制定された。最近では、2013 年から IT 総合戦略本部の「パーソナルデータに関する検討会」が中心になって [1]、個人情報保護法の改正に向けた議論が行われている。また、EU では、1995 年に制定されたデータ保護指令が存在するが、2012 年以降、一般データ保護規則の制定に向けた検討が進められるようになっている [2]。データ保護指令は、各国に一定の法律の制定を義務付けているが、指令が各国に直接適用されるわけではない。それに対し、一般データ保護規則は、各国に直接適用される点で異なっている。そのほか、データ保護規則は、「忘れられる権利」ないし「消去される権利」の導入、罰則の強化などの点が特徴になっている。そして、米国では、公的部門については、1974 年のプライバシー法が存在するものの、民間部門に関する包括法は存在していない。そこでは、個別領域ごとに個別法が存在するにすぎず、いわゆるセクトラル方式がとられている。もっとも、米国では、2012 年に「ネットワーク化された世界における消費者データプライバシー」という政策大綱が公表され [3]、その中で「消費者プライバシー権利章典」が提案されている点が注目される。

ビッグデータのプライバシー問題については、主として、以下のような観点から議論がなされている [4]。第一に、パーソナルデータを取得する際の本人同意が形骸化してしまっているため、いかにして分かり易い表示・説明をすることによって、実質的な本人同意を実現するかということである。多くの場合、プライバシーポリシーは、長文で難解な文章になっているため、多くの消費者が中身を全く読まずに盲目的に同意してしまっているというのが実態になっており、本人同意が形骸化してしまっている。この課題については、我が国でも経済産業省などによって様々な取り組みがなされているが、その点は後述することにする。

第二に、ビッグデータに含まれるパーソナルデータを匿名化することによって利活用をはかりたいという要望とプライバシー保護の要請をいかにして調和させるかということである。この点については、我が国では、「パーソナルデータに関する検討会」などで議論されてきたが、海外でも活発な議論がなされている。EU では、イギリスの ICO が 2012 年に匿名化に関する行動規範を策定し [5]、2014 年には、29 条作業部会が「匿名化技術に関する意見書」を公表している [6]。また、米国では、2014 年にホワイトハウスが、「ビッグデータ：機会の獲得、価値の保持」と題する報告書を公表し [7]、それと同時に、大統領科学技術諮問委員会

（PCAST）が、「ビッグデータとプライバシー：技術的視点から」という報告書を公表している [8]。これらの報告書でも匿名化技術についての記述があるが、再特定技術が発達しているため完全な匿名化が難しいことにも触れられている。

なお、ビッグデータの利用については、上記二つの課題のほかにも、いわゆるプロファイリングの問題も指摘されている。2013 年のデータ保護プライバシーコミッショナー国際会議では、プロファイリングを実施する際の条件を定めた「プロファイリングに関する決議」が採択された [9]。また、EU の一般データ保護規則案でも、プロファイリングを規制する条文が盛り込まれている。

（４）科学技術的・政策的課題

・パーソナルデータをユーザーから取得する際には、取得する情報項目や利用目的などを記載したプライバシーポリシーを明示したうえで本人同意を取得することが望ましい。しかし、現状のプライバシーポリシーは、長文で難解なものになっているため、多くのユーザーは中身を全く読まずに盲目的に同意してしまっており、本人同意が形骸化してしまっている。

そこで、実質的な本人同意をいかにして実現するかが課題となっているが、これを実現するためには、ユーザーに対して分かり易く簡潔にポリシーの内容を提示する必要がある。そのために、食品表示ラベルに類似した「情報共有標準ラベル」や「アイコン」などによる表示が検討されるようになっている。しかし、様々な表示方法などが提案されているため、ルールの明確化や標準化などの課題が残されている。

・ビッグデータに含まれるパーソナルデータを匿名化することによって利活用をはかりたいという要望が産業界を中心に出されているが、匿名化に関するルールが不明確なために利活用が進まない状況になっている。また、ポール・オームなどによって指摘されているように [10]、匿名化された情報とネット上に散在する情報などを照合することによって再特定化を行う再特定技術が発達しているため、完全な匿名化を行うことが技術的に難しいということが問題になっている。そこで、再特定行為を法的に禁止するなどの政策が提案されているが、このような政策に対しても、そもそも再特定行為を検知することが難しいため規制が難しいといった問題が残されている。

（５）注目動向（新たな知見や新技術の創出、大規模プロジェクトの動向など）

・パーソナルデータの利用について、いかにして分かり易い表示・説明をすることによって実質的な本人同意を実現するかという課題については、総務省の「パーソナルデータの利用・流通に関する研究会」の活動も関係するが、直接関係するものとしては、経済産業省の一連のプロジェクトがある。

まず、2012 年から 2013 年に行われた経済産業省の「IT 融合フォーラム・パーソナルデータワーキンググループ」によって、基本的な問題提起がなされた [11]。同 WG では、消費者と事業者の間の信頼関係の構築が重要であるという観点から、分かり易い表示とするために求められる要素や、具体的な手法などが検討された。具体的手法としては、「平易で簡潔な表示」、「ラベルによる一覧表示」、「アイコンによる一覧表示」があげられている。そして、この活動を引き継ぐ形で、2013 年から 2014 年にかけて、パーソナルデータの利

活用に関する事前相談評価の試行が実施された。これは、消費者に対する分かり易い表示・説明を中心に、消費者との信頼関係構築のための取り組みについて、事業者の相談に乗り、評価しようとする試みである。その成果として、「事前相談評価・評価基準書」が公表されている[12]。さらに、最近では、これらの活動の流れを受けて、国際標準化に向けた取り組みが開始している。2014 年に、経済産業省内に検討委員会が設置され、「消費者向けオンラインサービスにおける通知と同意・選択に関するガイドライン」が策定された[13]。今後、このガイドラインを ISO/IEC JTC1 SC27WG5 において国際規格化することを目指した活動が行われることになっている。

・ビッグデータに含まれるパーソナルデータを匿名化し利活用をはかる際のプライバシー保護の問題については、国内でも様々な委員会やプロジェクトにおいて検討されてきた。まず、医療分野については、2012 年に「社会保障分野サブワーキンググループ及び医療機関等における個人情報保護のあり方に関する検討会の合同開催」において議論がなされた[14]。そして、2013 年には、規制改革会議「創業等ワーキンググループ」での議論や[15]、「世界最先端 IT 国家創造宣言」の公表を経て[16]、IT 総合戦略本部の「パーソナルデータに関する検討会」において、本格的な検討が開始された。同検討会は、2014 年にも引き続き実施されたが、特に同検討会のもとに設置された「技術検討 WG」において、匿名化技術に関する詳細な検討がなされたことが注目される。

また、匿名化の課題については、前述したように海外でも活発な議論が行われている。すなわち、ICO の匿名化に関する行動規範、29 条作業部会の意見書、ホワイトハウスや PCAST の報告書などである。我が国の議論では、これらの海外動向が十分に検討されていないところがあるが、今後も海外の最新動向に注視し、それらを十分に踏まえたうえで、検討を行うことが重要である。

・国内外で、ビッグデータを活用した新しい情報サービスが展開されるようになっているが、中には、プライバシー・個人情報の保護が不十分であるとして、消費者からの反発を受けたり、社会的批判を受けたりしているものもある。このような事態を防止するためには、新しい情報サービスや情報システムを導入する前に、それがプライバシーに対してどのような影響を与えるのかを評価するプライバシー影響評価（PIA）を実施することが有効である。PIA は我が国では、まだあまり馴染みがないが、アメリカ、カナダ、イギリス、オーストラリアなどの諸外国ではすでに実施されているものである。この PIA については、現在、ISO/IEC JTC1 SC27WG5 において国際標準化の作業が進められている[17]。

（６）キーワード

パーソナルデータ、プライバシー、個人情報保護法、本人同意、匿名化技術、再特定技術、一般データ保護規則、消費者プライバシー権利章典

（７）国際比較

国・地域	フェーズ	現状	トレンド	各国の状況、評価の際に参考にした根拠など
日本	基礎研究	○	→	・2003年に個人情報保護法を含む個人情報保護関連5法が整備された。最近では、「パーソナルデータに関する検討会」における議論を経て、個人情報保護法の改正に向けた検討が進められている。
	応用研究・開発	○	→	・パーソナルデータを利活用する際の消費者に対する分かり易い表示・説明については、経済産業省による一連のプロジェクトがある。IT融合フォーラム・パーソナルデータWG、パーソナルデータの利活用に関する事前相談評価試行、国際標準化に向けた活動などである。 ・匿名化の問題については、「パーソナルデータに関する検討会」が中心となって議論が進められてきた。2014年には、「パーソナルデータの利活用に関する制度改正大綱」が公表された[18]。そこでは、パーソナルデータの利活用を促進するための枠組みとして、「個人特定性低減データ」に関する規律が提案されている。
	産業化	○	→	・日本情報経済社会推進協会（JIPDEC）、次世代パーソナルサービス推進コンソーシアム、インターネット広告推進協議会（JIAA）などが産業界を巻き込む形で様々な活動を行っている。
米国	基礎研究	○	→	・民間部門を規制する包括的な個人情報保護法は存在せず、いくつかの個別法が存在するにすぎない。基本的には企業の自主規制に委ねられている。 ・消費者のプライバシー保護については、FTC（連邦取引委員会）が重要な役割を果たしている。FTCは、企業によるプライバシー保護に向けた自主規制を促進しつつ、プライバシーポリシーに対する違反があった場合は、FTC法5条による制裁を科すことによって、実効性を担保している。
	応用研究・開発	◎	↗	・2012年に「ネットワーク化された世界における消費者データプライバシー」という政策大綱が公表され、その中で「消費者プライバシー権利章典」が提案されている。 ・2014年に、ホワイトハウス「ビッグデータ：機会の獲得、価値の保持」、PCAST「ビッグデータとプライバシー：技術的視点から」が公表されている。これらの報告書は、匿名化について言及しているが、再特定技術が発達しているため完全な匿名化が難しいことにも触れられている。
	産業化	○	↗	・米国では、民間部門に関する包括的な個人情報保護法が存在しないため、企業による自主規制が重要になる。そのため、様々な分野において、自主規制のためのガイドラインの制定など、ルール形成の努力が行われている。
欧州	基礎研究	◎	↗	・1995年にデータ保護指令が成立し、これに基づいて各国が個人データ保護に関する法制度を整備している。また、2002年には、電子通信の分野について、e-プライバシー指令が成立している。この指令では、通信の秘密保持、Cookieやロケーションデータを利用する際のオプトインによる本人同意取得などが定められている。
	応用研究・開発	◎	↗	・2012年に、データ保護指令を大幅に改正することを目的とする一般データ保護規則提案が公表されている。この一般データ保護規則案では、「忘れられる権利」ないし「消去される権利」の導入や、企業に対する罰則の強化などが特徴になっている。 ・パーソナルデータを匿名化する際のプライバシー保護の問題については様々な動きがある。2012年にICOが匿名化に関する行動規範を策定し、2014年には29条作業部会が意見書を公表している。同意見書は、様々な匿名化技術を取り上げて検討を加えているが、いずれの匿名化技術にも限界があるので、特定の匿名化技術に依存しないことを推奨している。
	産業化	○	→	・イギリスでは、2011年以降、midataと呼ばれるプロジェクトが進行している。このプロジェクトは、消費者が民間企業の持つ自己に関するデータに自由にアクセスしコントロールできるようにすることを目指すものである。このプロジェクトには、情報通信、金融、エネルギーなど様々な分野の企業が参加している。

中国	基礎研究	△	→	・中国では、2006年に個人情報保護法草案（「中華人民共和国個人情報保護法（専門家意見版）及び立法研究報告書」）が公表されたが、現在までのところ成立していないようである。そのため、包括的な個人情報保護法は存在しないが、「消費者権益保護法」、「電信及びインターネットユーザー個人情報保護規定」などの法規において個人情報保護に関する規定が定められている[19]。
	応用研究・開発	△	→	・2012年に、国家標準化委員会は、「公共及び商業用サービスに関わる情報システムの個人情報保護のガイドライン」を定めている。
	産業化	○	↗	・2008年より大連ソフトウエア産業協会が、個人情報保護の水準を認証する個人情報保護評価制度（PIPA）の運用を開始している。そして、このPIPAと日本のプライバシー・マーク制度の間で相互承認プログラムが行われている。
韓国	基礎研究	○	↗	・韓国では、かつては、米国類似のセクトラル方式がとられていたが、2011年に、民間部門と公的部門の両方を対象とする包括的な個人情報保護法が制定され、施行された。
	応用研究・開発	○	→	・ビッグデータの活用については、大量のユーザーに関する位置情報の活用とそれに伴うプライバシー保護の課題が取り上げられることがあるが、韓国では「位置情報の保護及び利用等に関する法律」が存在することが注目される。
	産業化	○	→	・韓国では、未来創造科学部と情報化振興院（NIA）が、2013年に「ビッグデータ分析活用センター」を開設した。同センターは、民間企業や大学とも連携しながら実証実験などの様々なプロジェクトを進めている[20]。また、2014年には、民間企業がビッグデータを活用する際の指針として「個人情報の識別化事例集」が公表されている。

（註 1）フェーズ
基礎研究フェーズ ：大学・国研などでの基礎研究のレベル
応用研究・開発フェーズ ：研究・技術開発（プロトタイプの開発含む）のレベル
産業化フェーズ ：量産技術・製品展開力のレベル

（註 2）現状
※我が国の現状を基準にした相対評価ではなく、絶対評価である。
◎：他国に比べて顕著な活動・成果が見えている、 ○：ある程度の活動・成果が見えている、
△：他国に比べて顕著な活動・成果が見えていない、×：特筆すべき活動・成果が見えていない

（註 3）トレンド
↗ ：上昇傾向、 →：現状維持、 ↘：下降傾向

（８）引用資料

1) パーソナルデータに関する検討会
http://www.kantei.go.jp/jp/singi/it2/pd/index.html

2) Proposal for a Regulation of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data and on the free movement of such data （General Data Protection Regulation）
http://ec.europa.eu/justice/data-protection/document/review2012/com_2012_11_en.pdf

3) Consumer Data Privacy in a Networked World: A Framework for Protecting Privacy and Promoting Innovation in the Global Digital Economy
http://www.whitehouse.gov/sites/default/files/privacy-final.pdf

4) 村上康二郎：ビッグデータ時代におけるプライバシー・個人情報の保護と法的問題点（ビッグデータ・マネジメント, 2014, 269-279 所収）

5) Anonymisation: Managing Data Protection Risk Code of Practice
http://ico.org.uk/for_organisations/data_protection/topic_guides/~/media/documents/library/Data_Protection/Practical_application/anonymisation-codev2.pdf

6) Opinion 05/2014 on Anonymisation Techniques
http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf
7) Big Data: Seizing Opportunities, Preserving Values
http://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_5.1.14_final_print.pdf
8) Big Data and Privacy: A Technological Perspective
http://www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_big_data_and_privacy_-_may_2014.pdf
9) Resolution on Profiling, adopted on 24 Sep. 2013
https://privacyconference2013.org/web/pageFiles/kcfinder/files/2.%20Profiling%20resolution%20EN%281%29.pdf
10) Paul Ohm: Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization, UCLA Law Review, Vol.57, 1701-1777
11) パーソナルデータ利活用の基盤となる消費者と事業者の信頼関係の構築に向けて
http://www.meti.go.jp/press/2013/05/20130510002/20130510002-2.pdf
12) パーソナルデータ利活用ビジネスの促進に向けた,消費者向け情報提供・説明の充実のための「評価基準」と「事前相談評価」のあり方について
http://www.meti.go.jp/press/2013/03/20140326001/20140326001-2.pdf
13) 消費者向けオンラインサービスにおける通知と同意・選択に関するガイドライン
http://www.meti.go.jp/press/2014/10/20141017002/20141017002a.pdf
14) 社会保障分野サブワーキンググループ及び医療機関等における個人情報保護のあり方に関する検討会の合同開催
http://www.mhlw.go.jp/stf/shingi/2r9852000000ai9a.html#shingi129272
15) 規制改革会議・創業等ワーキンググループ
http://www8.cao.go.jp/kisei-kaikaku/kaigi/publication/p_index.html
16) 世界最先端 IT 国家創造宣言について
http://www.kantei.go.jp/jp/singi/it2/kettei/pdf/20130614/siryou1.pdf
17) 村上康二郎：プライバシー影響評価（PIA）に関する国際的動向と我が国における課題, 情報ネットワーク・ローレビュー13(2), 2014, 33-56
18) パーソナルデータの利活用に関する制度改正大綱
http://www.kantei.go.jp/jp/singi/it2/info/h260625_siryou2.pdf
19) TMI 中国最新法令情報―(2014 年 1 月号)―
http://www.tmi.gr.jp/global/legal_info/china/2014/january-2.html
20) 韓国 IT 事情第 29 回『ビッグデータ活用政策の動向』
http://www.jpc-net.jp/cisi/mailmag/m220_pa5.html