## 3.3 ビッグデータ

ビッグデータの研究開発で目指すのは、膨大なデータの収集・解析、実世界現象の精緻でリアルタイムな把握・予測により、さまざまな社会課題を解決し、安全・安心で生産性の高い社会を実現することである。人間の手に負えない大規模複雑な社会課題が深刻化しており、その解決手段としてビッグデータに大きな期待が寄せられている。

これを支える技術は、ビッグデータの収集・蓄積→分析→活用という流れに沿った基礎技術、活用基盤、重点応用、制度設計という塊で捉え、以下の八つの研究開発領域を設定した（図 3-3-1）。

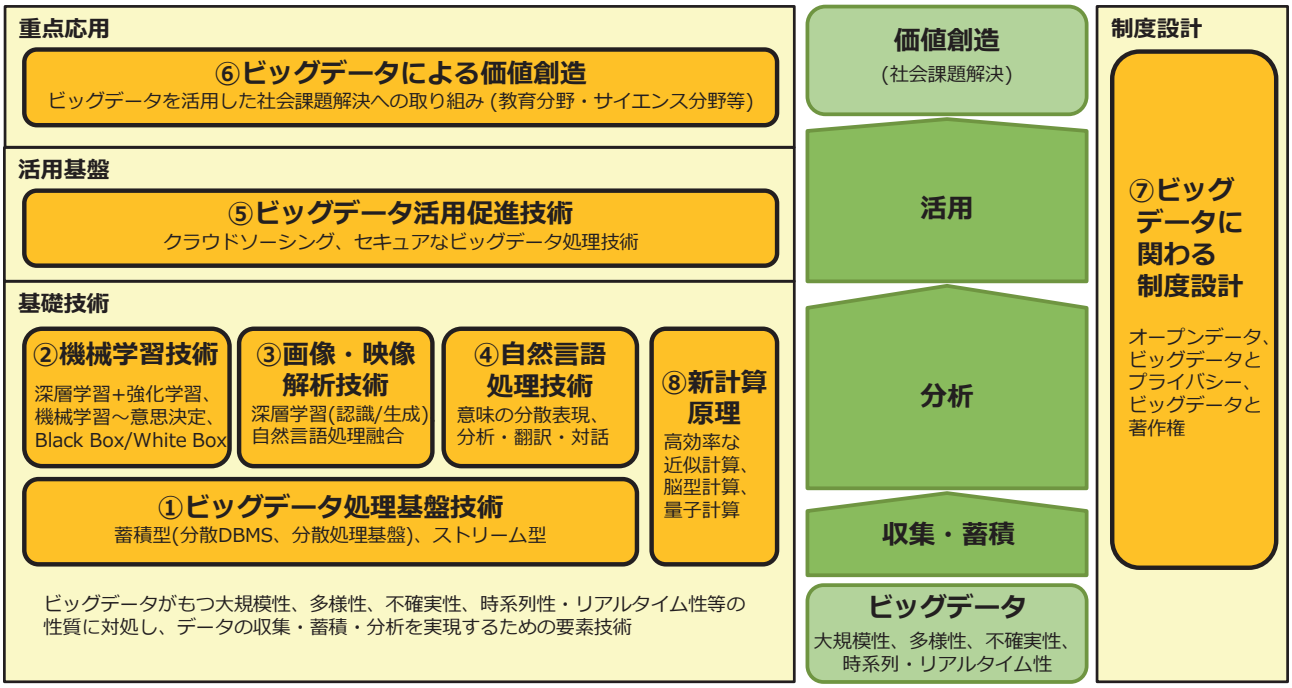


図 3-3-1　ビッグデータの俯瞰図

①ビッグデータ処理基盤技術

大量の計算機あるいはハイエンド計算機を活用し、ビッグデータを高速・高効率に処理するソフトウエア技術である。大規模性、多様性、不確実性、時系列性・リアルタイム性を備えたビッグデータを扱うさまざまな応用において共通的に必要とされる。

②機械学習技術

データの背後に潜む規則性や特異性を発見することにより、人間と同程度あるいはそれ以上の学習能力をコンピューターで実現しようとする技術である。事象や対象物についてその観測データに基づく判別・分類、予測、異常検知等が可能になる。

③画像・映像解析技術

カメラ等で撮影された画像や映像をコンピューターで解析して、その画像･映像の内容、つまり、そこに写っているもの（人や物体、文字等）の位置・カテゴリあるいは風景・場所・状況等を認識する技術である。

④自然言語処理技術

コンピューターで自然言語（人間が日常の意思疎通のために用いる自然発生的な記号体

系）を処理する技術である。自然言語で表現された大量テキスト情報の活用の促進・支援や、自然言語によるコミュニケーションの促進・支援に活用される。

⑤ビッグデータ活用促進技術

ビッグデータの基礎技術である①～④以外に、ビッグデータをさまざまな課題解決（価値創造）に活用していく上で共通的に必要になる技術である。特にクラウドソーシングとセキュアなビッグデータ処理技術が重要である。

⑥ビッグデータによる価値創造

ビッグデータに基づく実世界現象の精緻でリアルタイムな把握・予測による課題解決、安全・安心で生産性の高い社会に向けた価値創造のための技術である。特に、科学技術政策の面から、教育やサイエンスにおけるビッグデータ活用に着目する必要がある。

⑦ビッグデータに関わる制度設計

ビッグデータの流通・活用を支える仕組み・制度を取り上げる。特に、最小限の制約のみで誰でも自由に利用・加工・再配布ができるオープンデータと、データに関わる権利確保の面からプライバシー保護と著作権に注意が必要である。

⑧新計算原理

実世界に大規模分散するビッグデータを低消費電力で高速・高効率に処理する技術である。①の発展形にとどまらず、②～④のアルゴリズムの進化と密に関わる基本計算原理への取り組みであり、高効率な近似計算・脳型計算・量子計算等を含む。

これらビッグデータの研究開発領域は、②～④の領域を中心に人工知能（AI）技術との関わりが深く、各国とも国家資金が重点投入され、産業界での競争が激化している。

米国は、Google、Facebook、Microsoft 等の巨大 IT 企業が積極的な投資をしていることに加えて、Airbnb、Uber 等、次々とベンチャー企業が誕生し、成長している。技術・資金だけでなく、その支配下にあるデータ量の面でも、米国が強いポジションを取っている。

トップ国際会議の論文は、量・質とも米国が世界をリードしているが（例えば機械学習分野の ICML は 2016 年の採択論文のうち半数近くが米国発の論文であった）、中国の追い上げが目立つ。採択論文数で 2 位に浮上しただけでなく、深層学習技術に関わる大規模画像認識性能の競技会 ILSVRC において、2015 年に MSRA が 1 位、2016 年は 1 位が中国公安部第三研究所、3 位が香港中文大学と、中国勢の躍進が著しい。産業界も、Baidu の検索エンジン、テンセントのチャットサービス、アリババの EC サービス等、盛んな研究開発投資、大規模データの囲い込みを進めており、勢いがある。

欧州は、セマンティック Web 研究の層が厚く、プライバシー保護を含む制度設計への取り組みに特徴がある。加えて、Google が買収した英国の DeepMind は AlphaGo で囲碁の世界トッププロに勝利する等、最先端の深層学習技術で世界中から注目され、存在感を高めている。

日本は、米国・中国のようなネットサービスのビッグデータよりも、実世界ビッグデータの活用に基づく実社会の課題解決（ソリューション）への取り組みに特徴がある。タスク特化の画像認識で世界トップ性能を出してきた画像認識技術、ロボット制御等で先行する深層強化学習技術・転移学習技術、機械学習に基づく意思決定問題の解法、非順序型実行原理に基づく超高性能 DB エンジン等、社会課題領域で迅速かつシャープに性能を出す AI 技術・実世界ビッグデータ解析技術に強みを有する。

### 3.3.1　ビッグデータ処理基盤技術

#### (1) 研究開発領域の簡潔な説明

ビッグデータ処理基盤技術は、大規模なデータ処理を高速・高効率に実行するためのソフトウエア技術であり、大量の計算機あるいはハイエンド計算機を活用することで、データ分析･解析処理の効率的な実行を実現する。ビッグデータは大規模性に加えて、多様性、不確実性、時系列性・リアルタイム性等の性質を備えており、このような性質に対処してビッグデータを高速・高効率に処理する技術は、さまざまなビッグデータ応用において共通的に必要とされる。

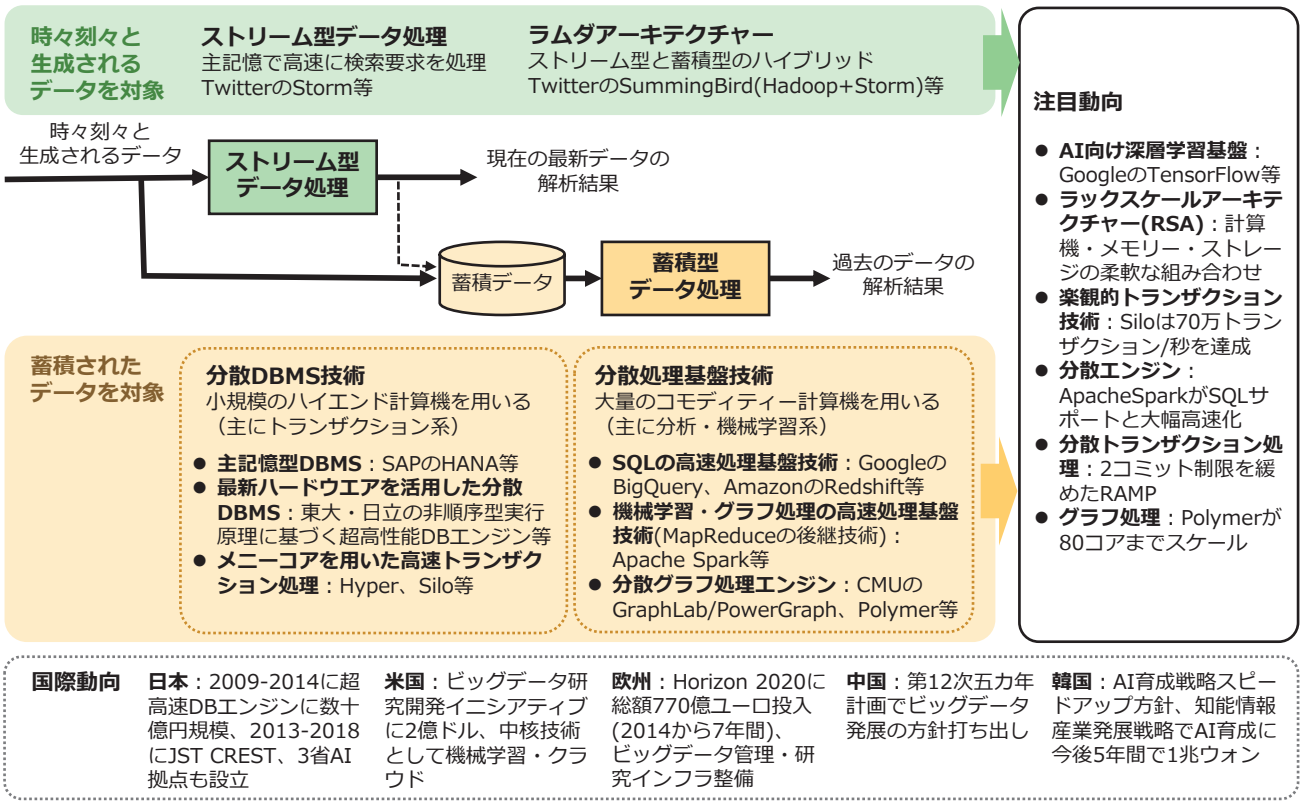


図 3-3-2　領域俯瞰：ビッグデータ処理基盤技術

#### (2) 研究開発領域の詳細な説明と国内外の動向

IDC Japan の 2016 年 6 月の報告[1)]によれば、2015 年の国内ビッグデータテクノロジー / サービス市場は、前年比 32.3 ポイント増の高い成長率を記録して市場規模は 947 億 7600 万円に到達し、今後の国内ビッグデータテクノロジー / サービス市場は 2020 年に 2889 億 4500 万円、2015 年～ 2020 年の年間平均成長率は 25.0% になると予測している。

ビッグデータを活用したさまざまなサービスを実現するには、ビッグデータを格納して分析・解析プログラムを高速・高効率に実行できる処理基盤が必要である。実際、クラウド環境では、エクサバイトスケールのデータ規模、100 万台を超えるサーバ群による大規模データ処理を実現できるような、ソフトウエア、ハードウエアから成るビッグデータ処理基盤が構築・運用されている。計算機資源を低価格で大量に調達することで、スケールメリットを生かした低価格なクラウドサービスを利用者に提供するため、このような技術が活用されている。

このようなビッグデータ処理基盤技術は、処理の対象となるデータが、蓄積されたデータであるか、あるいは時々刻々と生成されるデータであるかの二つの観点から、「蓄積型データ処理」と「ストリーム型データ処理」という技術領域に分類することができる。言い換えると、蓄積型データ処理は過去のデータに対する処理であり、ストリーム型データ処理は現在の最新データに対する処理である。

［蓄積型データ処理技術］

蓄積型データ処理技術はさらに、小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システム（分散 DBMS）技術」と、分散 DBMS の機能を簡略化して数千台を超える大量のコモディティ計算機を用いる「分散処理基盤技術」（NoSQL、MapReduce 等がその代表）に分類される。

前者の「分散 DBMS 技術」には Oracle、IBM、Microsoft 等のデータベース（DB）製品や、SAP の HANA に代表される主記憶型の DBMS がある。最先端の技術としては、大量の主記憶・メニーコア・不揮発性メモリ・高速ネットワーク等の最新ハードウエアを活用した分散 DBMS の研究がなされている。例えば、東京大学と日立製作所が共同開発・商用化した非順序型実行原理に基づく超高性能 DB エンジン [2] があげられる。これは、参照中心の分析系 DB に対して、検索命令を細分化し、ハードディスクレベルで実行順序を入れ替えることでハードディスクの性能を最大限に引き出す技術である。一方、大学においては、参照と更新系の両方の処理を行うトランザクション系 DB の分野で、メニーコア上でトランザクションを高速処理する DB エンジンの研究が盛んになっている。その中心的なシステムとして Hyper [3] や Silo [4] があげられる。Hyper はベンチャー起業として独立後、2016 年にデータ分析大手企業の Tableau 社に買収されて話題となった [5]。

後者の「分散処理基盤技術」に関しては、米国の Google、Amazon、Facebook、Twitter 等の Web 系の企業が技術を先導しており、各社のサービスの専用用途に応じて開発が進められている。DB 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 がクラウドサービスとして提供している BigQuery や Amazon の Redshift がある。BigQuery は、数億レコードを数秒で検索することが可能で、数千台のディスクを同時使用していると推測される。Redshift は、ディスクアクセスを並列化すると同時に、分析対象のデータのみをディスクアクセスするカラム指向技術とその軽量圧縮技術を使うことで高速処理を実現している。一方、大学においては、機械学習処理やグラフ処理を高速化する基盤技術の研究が進められており、MapReduce の後継となる技術が多数提案されている。特に、機械学習や行列計算における繰り返し型の分析処理に着目し、その中間結果を主記憶に保持することでディスクアクセスコストを削減した Apache Spark が成長を続けている [6]。日本発の技術としては機械学習を処理する Hivemall が Apache foundation のインキュベータープロジェクトに 2016 年 9 月に認定された [7]。グラフ処理は人・モノ・場所などの関係情報から知識を発見する技術であり、分散グラフ処理エンジンの代表的例としては、CMU の GraphLab/PowerGraph、マイクロソフト研究所の Graph Engine [8]、メニーコア環境におけるグラフ分析の代表的技術としては Polymer [9] があげられる。

［ストリーム型データ処理技術］

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型データ処理技術がある。ストリーム型データ処理技術は、米国や日本の DBMS ベンダから製品化されており、産業利用が進んでいる。Web 系企業では、Twitter がストリーム処理エンジン Storm を開発し、2011 年に公開している。

近年では、蓄積型データ処理とストリーム型データ処理をハイブリッドに組み合わせる技術が出現してきている。これは「ラムダアーキテクチャー」と呼ばれる。この類の技術は古くからあり歴史は長いが（2003 年に論文発表された PSoup[10] が代表例）、2014 年には Twitter から Hadoop と Storm を組み合わせた SummingBird が発表され、オープンソースソフトウエア（OSS）として公開された [11]。

［国際動向］

国家施策を中心に各国の動向をまとめる。

日本では、人工知能（AI）研究拠点として文部科学省の理化学研究所革新知能統合研究センター、経済産業省の産業技術総合研究所人工知能研究センター、総務省の情報通信研究機構脳情報通信融合研究センターおよびデータ駆動知能システム研究センターが活動している。特にビッグデータ処理基盤の観点では、2009 年～ 2014 年の期間に最高速 DB エンジンの開発をテーマとして、数十億規模の予算による DB エンジン技術および機械学習技術の研究開発が実施された。2013 年～ 2018 年の期間、JST CREST の「ビッグデータ統合利活用のための次世代基盤技術の創出・体系化」領域で研究開発が進められている。

米国では、2012 年 3 月に科学技術政策局（OSTP）がビッグデータ研究開発イニシアティブを発表、政府として戦略的に取り組む姿勢を明確にしている。その中核技術として、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あげられている。代表的な成功プロジェクトとしてスタンフォード大を中心とした Apache Spark があげられる。

欧州における主な取り組みとしては、Horizon2020 の中で、ビッグデータの管理・研究インフラの整備、データのオープンアクセスが計画されている。

中国では、ビッグデータ産業の発展を奨励・推進するため、「情報消費の促進による内需拡大に関する意見」、「ソフトウエアと情報技術サービス業の『第 12 次五カ年計画』」等の政策・計画の中で、ビッグデータ発展の方針を打ち出した。さらに、全国情報技術標準化技術委員会を組織し、ビッグデータ標準化の需要分析を行い、標準体系の枠組み研究および関連の標準研究を実施、国際標準化機構にビッグデータ研究案を提出している [12]。

韓国では、2012 年にビッグデータ・マスタープラン、2015 年にビッグデータ活用拡大対策を発表し、2017 年までに 97 のビッグデータ活用事業を進めるとしている [13]。

### (3) 注目動向

Google では、検索サービスやクラウドサービスに必要な基盤技術はほぼ完成した状況にあると考えられ、現在は画像認識、音声認識、多言語翻訳等の Google が提供する AI 系サービスを支える深層学習の基盤技術（TensorFlow）の開発が目立っている [14]。同様

に Microsoft 研究所では、特に対話を中心とした技術に注力しており、その基盤技術として深層学習エンジンを OSS で公開している [15]。

計算機ハードウエア関連では、計算機・メモリ・ストレージを柔軟に組み合わせ可能なラックスケールアーキテクチャー（RSA）が主流になっている。

（2）で取り上げたビッグデータ処理基盤技術もさらに発展している。メニーコア上でトランザクションを高速処理するエンジンの研究が進み、近年は楽観的トランザクション技術 [i] に関する研究が盛んである。代表技術の Silo[4] は、論理クロックに基づく非集中処理機構を導入することで、70 万トランザクション / 秒という高い性能を達成した。FOEDUS は、主記憶と不揮発性メモリの併用に適したトランザクション処理システムを実現している [16]。分散エンジンに関しては、Apache Spark が発展しており、2016 年 7 月に 2.0 版が正式リリースされ、SQL サポートと大幅な高速化が達成された [6]。分散トランザクション処理に関しては、2 層コミットの制約を緩めた RAMP（Read Atomic Multi-Partition）が提案され、1 層によるコミットを実現し [17]。グラフ処理に関しては、Polymer がメニーコアでのデータレイアウトを工夫し、さらにデータの並行制御を効率化する工夫により、80 コアまでスケールする性能を達成した [9]。

### （4）科学技術的課題

ビッグデータ処理基盤技術の動向をまとめると、蓄積型データ処理技術における①トランザクション系の処理エンジンである分散 DBMS 技術と②ビッグデータを分析・機械学習するエンジンである分散処理基盤技術、③センサーなどから時々刻々と生成される最新データを処理するストリーム型データ処理技術がある。社会的なニーズから見ると、①は銀行業務や EC サイトなどで利用されており、現在の IT 社会の根幹を成すものである。大量の主記憶・メニーコア・不揮発性メモリ・高速ネットワーク等の最新ハードウエアを活用することで、より大規模なデータを高速・省電力で処理できる技術を目指した技術開発が求められる。②はメディアの自動認識・自動翻訳といった近年のビッグデータ・AI の実現に必要不可欠な技術である。従来培った大規模データ処理技術を応用して、新たなデータ処理アーキテクチャーを含めた技術の発展が見込まれる。また、③は、IoT の普及がもたらす実世界のリアルタイムセンシングデータを分析するシステムを実現する技術であり、新たな応用の発展に従って技術の発展が見込まれる。

### （5）政策的課題

ビッグデータ処理基盤技術の研究は、最新ハードウエアの利用や、リアルなビッグデータ処理で生じる課題への取り組みが重要である。前者に関しては、最新ハードウエアを提供している Intel や HP における研究開発活動が大学における活動よりも有利で、後者に関しては、リアルなビッグデータを利用したビジネスを運営している Web 系や通信系の

i）　楽観的トランザクションは、通常の（悲観的）トランザクションと比べて、ロックよりもトランザクションの並列実行を優先し、複数トランザクションの競合頻度が低いケースに高速性能を発揮する技術である。近年、CPU の高性能化に伴ってトランザクション処理が短時間で実行できるようになって競合頻度が下がったため、注目されている。

企業が有利な傾向にある。日本の大学がビッグデータ処理基盤技術の分野で活躍するには、産学連携の推進等、このような場・環境との接点を一層強化することが不可欠である。現在の日本は、他国同様ビッグデータ処理分野に大型国家予算が投資されている状況にある一方、その基盤技術の研究開発を遂行できる人材、特に学生が少ないことが、日本の競争力を低下させている大きな原因の 1 つと考えられる。この状況を改善するには、まず大学教育においてデータ処理の基盤技術やプログラミングの教育を強化するとともに、特に研究開発を牽引する博士学生・ポスドクの数を倍増させる必要がある。そのためには、博士・ポスドク後の安定雇用のシナリオを充実させなければならない。

### (6) キーワード

並列分散データベース管理システム（DBMS）、ストリーム型データ処理、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 | ○ | → | ・最新ハードウエアを用いた DBMS の研究 [18),19)]、大規模なグラフデータ処理の高速アルゴリズム [20),21)] 等、トップ会議で多数採択されている。 |
| | 応用研究・開発 | ○ | ↗ | ・Hivemall が Apache foundation のインキュベータープロジェクトに 2016 年 9 月認定された。<br>・Apache Hadoop や Spark へのコミッタ輩出等、日本からビッグデータ処理基盤の OSS への貢献が伸びている [22)]。 |
| 米国 | 基礎研究 | ◎ | → | ・大学・企業の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。最新ハードウエアを用いた基盤研究では楽観的トランザクションをベースとした Silo（Google Scholar で引用数 98）、分散トランザクション処理に関しては RAMP（Google Scholar で引用数 50）が影響を与えている。 |
| | 応用研究・開発 | ◎ | → | ・分散処理基盤技術の取り組みは、Web 系企業が OSS 開発を先導しており、米国が圧倒している。<br>・OSS 開発コミュニティでは大学と連携し、基礎研究成果を取り込む土壌も備えているのも強みである。Spark コミュニティは UC Berkeley 発で急成長を遂げている。 |
| 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 等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。Hyper は最新ハードウエアを活用して分析系処理とトランザクション系処理を両立するエンジンの代表格である。 |
| | 応用研究・開発 | ○ | ↗ | ・SAP の HANA 等のカラム指向インメモリ DBMS やストリーム型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 | ◎ | → | ・大学が中心となり、基礎研究で多くの成果を上げている。難関国際会議採録も米・欧に次いで多い。グラフデータ処理等のアルゴリズム系が中心で、分散処理基盤技術の高速化の取り組みもある。Polymer はメニーコアでデータレイアウトを工夫し、OS レベル競合を回避することで 80 コアまでスケールする高性能を達成 [9)]。 |
| | 応用研究・開発 | ○ | → | ・Microsoft はグラフエンジン [8)] や深層学習エンジン [15)] 等の取り組みを行っている。 |

| | | | | |
|---|---|---|---|---|
| 韓国 | 基礎研究 | ○ | → | ・フラッシュメモリを利用した DBMS の高速化等の高速 DBMS に関する取り組みや、グラフエンジンに関する取り組み [23] 等がある。 |
| | 応用研究・開発 | △ | → | ・ビッグデータ基盤技術に関して目立った活動はない。 |

（注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
（注２） 現状 ※わが国の現状を基準にした評価ではなく、CRDS の調査・見解による評価である。
◎ 特に顕著な活動・成果が見えている、 ○ 顕著な活動・成果が見えている
△ 顕著な活動・成果が見えていない、 × 活動・成果がほとんど見えていない
（注３） トレンド
↗：上昇傾向、 →：現状維持、 ↘：下降傾向

## (8) 参考文献

1）IDC, 国内ビッグデータテクノロジー / サービス市場予測を発表 ,
http://www.idcjapan.co.jp/Press/Current/20160613Apr.html
2）早水悠登 , 合田和生 , 喜連川優 , アウトオブオーダ型クエリ実行に基づくプラグイン可能なデータベースエンジン加速機構 , 情報処理学会論文誌 データベース 7 (2),pp.104-116,2014.
3）Alfons Kemper, Thomas Neumann, HyPer, A hybrid OLTP&OLAP main memory database system based on virtual memory snapshots, Proceedings of ICDE, 2011.
4）Stephen Tu, Wenting Zheng, Eddie Kohler, Barbara Liskov, Samuel Madden, Speedy transactions in multicore in-memory databases, Proceedings of SOSP, 2013.
5）TechCrunch,BI サービスの Tableau がドイツの HyPer を買収してビッグデータ分析を高速化 ,
http://jp.techcrunch.com/2016/03/11/20160310tableau-scores-advanced-database-tech-with-acquisition-of-german-startup-hyper/
6）Spark Release 2.0.0,
https://spark.apache.org/releases/spark-release-2-0-0.html
7）Apache Hivemall, https://hivemall.incubator.apache.org/
8）Microsoft, Graph Engine, https://www.graphengine.io/
9）Kaiyuan Zhang, Rong Chen, Haibo Chen, A NUMA-aware Graph-structured Analytics Framework, Proceedings of PPoPP, 2015.
10）Sirish Chandrasekaran, Michael J. Franklin, PSoup, A System for Streaming Queries over Streaming Data, VLDB Journal, 12（2）, 2003.
11）Oscar Boykin, Sam Ritchie, Ian O'Connell, Jimmy Lin, Summingbird, A Framework for Integrating Batch and Online MapReduce computations, PVLDB 7（13）, 2014.
12）SciencePortal China, 中国のビッグデータ発展の現状と趨勢 ,
https://www.spc.jst.go.jp/hottopics/1507/r1507_cheng2.html
13）ICT global trend, アルファ碁ショックで人工知能（AI）開発戦略がスピードアップ ,
https://www.fmmc.or.jp/ictg/country/korea.html
14）Jeff Dean, Building Machine Learning Systems that Understand, SIGMOD

tutorial, 2016.
15）Microsoft, The Microsoft Cognitive Toolkit, http://cntk.ai
16）Hideaki Kimura, FOEDUS: OLTP Engine for a Thousand Cores and NVRAM. Proceedings of ACM SIGMOD, 2015.
17）Peter Bailis, Alan Fekete, Ali Ghodsi, Joseph M. Hellerstein, Ion Stoica: Scalable Atomic Visibility with RAMP Transactions. Proceedings of ACM SIGMOD, 2014.
18）Hiroshi Inoue, Kenjiro Taura, SIMD- and Cache-Friendly Algorithm for Sorting an Array of Structures. PVLDB 8（11）, 2015.
19）Takashi Horikawa, Latch-free data structures for DBMS: Design, Implementation, and Evaluation. Proceedings of ACM SIGMOD, 2013.
20）Naoto Ohsaka, Takuya Akiba, Yuichi Yoshida, Ken-ichi Kawarabayashi, Dynamic Influence Analysis in Evolving Networks. PVLDB 9（12）, 2016.
21）Hiroaki Shiokawa, Yasuhiro Fujiwara, Makoto Onizuka, SCAN++: Efficient Algorithm for Finding Clusters, Hubs and Outliers on Large-scale Graphs. PVLDB 8（11）, 2015.
22）ITPro,NTT/NTT データから 3 氏が Hadoop や関連プロジェクトの「コミッタ」に就任 , http://itpro.nikkeibp.co.jp/atcl/news/14/121902330/
23）Hyeonji Kim, Juneyoung Lee, Sourav S. Bhowmick, Wook-Shin Han, Jeong-Hoon Lee, Seongyun Ko, Moath H. A. Jarrah, DUALSIM: Parallel Subgraph Enumeration in a Massive Graph on a Single Machine. Proceedings of ACM SIGMOD, 2016.

### 3.3.2　機械学習技術

#### (1) 研究開発領域の簡潔な説明

機械学習（Machine Learning）は、データの背後に潜む規則性や特異性を発見することにより、人間と同程度あるいはそれ以上の学習能力をコンピューターで実現しようとする技術である。これにより、事象や対象物について、その観測データに基づく判別･分類、予測、異常検知等が可能になる。情報爆発・ビッグデータの時代と言われる今日、さまざまな事象や対象物について大量の観測データが得られるようになり、機械学習技術は幅広い分野に応用されている。例えば、画像認識、音声認識、医療診断、文書分類、スパムメール検出、広告配信、商品推薦、囲碁・将棋等のゲームソフト、商品・電力等の需要予測、与信、不正行為の検知、設備・部品の劣化診断、ロボット制御、車の自動運転、等々。現在の第 3 次人工知能（AI）ブームをけん引しているのは機械学習技術の進化だとも言われる。

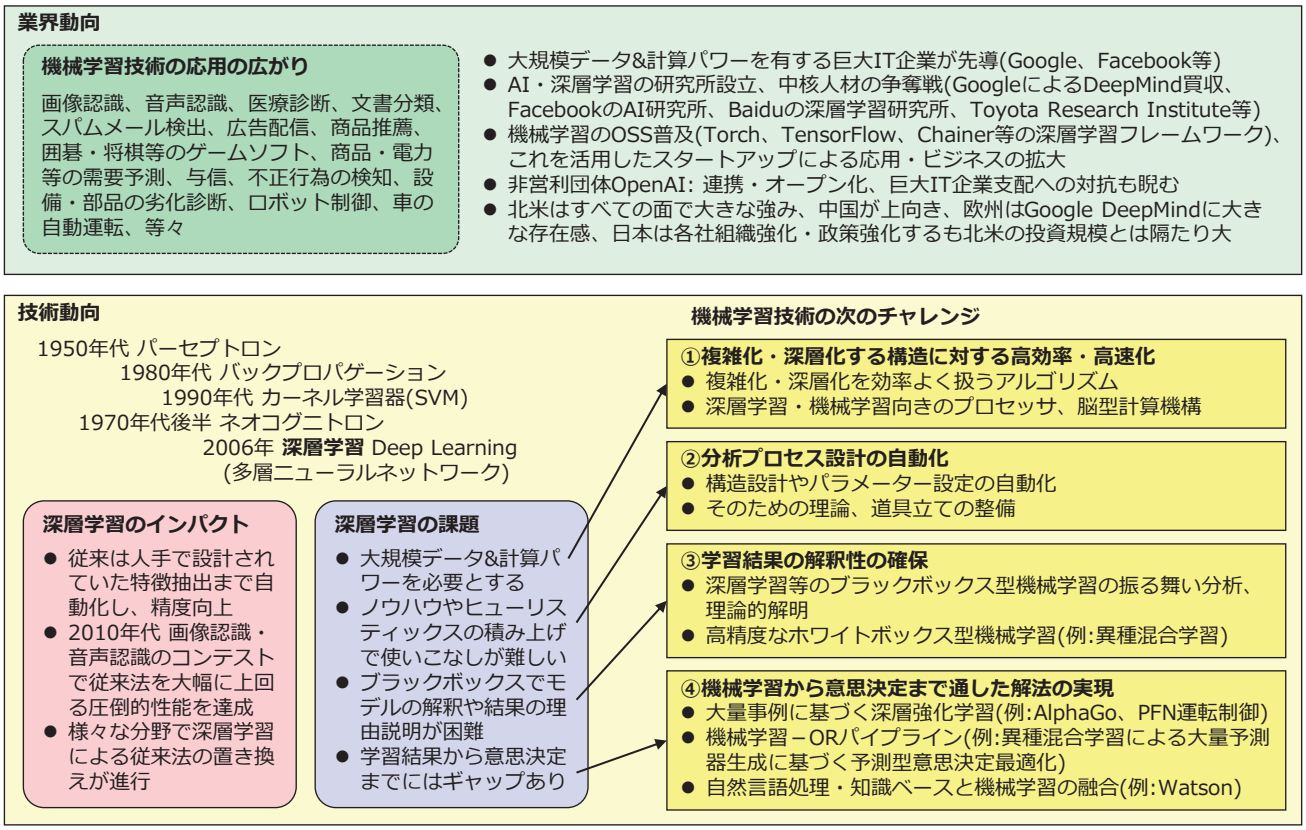


図 3-3-3　領域俯瞰：機械学習技術

#### (2) 研究開発領域の詳細な説明と国内外の動向

［機械学習研究の発展概観］

機械学習の研究は、古くは人間の脳の学習機能をコンピューターで実現しようという 1950 年代の研究にさかのぼる。パーセプトロン [1] と呼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[3]と呼ばれる、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、ロボットの制御等、さまざまな問題に適用され、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、学習パラメーター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さまざま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:SVM）と呼ばれるカーネル学習器が提案された[4]。SVM 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ことができる。その後、大規模データに対する超高速学習アルゴリズム等の開発を経て、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・自然言語等の処理や、生命情報の解析等、さまざま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るようになったが、いまだ決定打に欠ける状況である。

カーネル法の全盛期の真っただ中であった 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、教師なし学習手法によって事前学習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、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:CNN）[6]と呼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独立して研究されていた。これは、1970 年代後半に提案されていたネオコグニトロン[7]と呼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、深層学習（Deep Learning）と呼ばれるようになった。

深層学習の技術は、2010 年代に入ってから画像認識や音声認識のコンテスト（ILSVRC 等）[8]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（同時に第 3 次 AI ブーム）の真っただ中であり、当該分野は大きく発展しつつある。なお、機械学習・深層学習の画像認識への適用については第 3.3.3 節「画像・映像解析技術」、自然言語処理への適用については第 3.3.4 節「自然言語処理技術」で述べている。

［最近の業界動向］

国際的には、現在の深層学習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 教授らの北米の大学教員が深層学習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深層学習技術の産業応用をリードしている。特に、Ng 教授と Google が 2012 年 6 月に発表した深層学習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等、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 に買収され、Hinton 教授も Google の Distinguished Researcher に就任した。Bengio 教授と LeCun 教授は International Conference on Learning Representations（ICLR）という深層学習に特化した国際会議を 2013 年に開始し、LeCun 教授は 2013 年 12 月に Facebook に新設された AI 研究所の所長に就任した。Baidu は 2013 年 1 月に深層学習研究所を設立し Ng 教授が 2014 年 5 月に Chief Scientist に就任した。Google は 2014 年 1 月に DeepMind というロンドンの深層学習のベンチャー企業を買収した。Google DeepMind は、深層学習を用いた強化学習の研究で顕著な成果を挙げており、2016 年 2 月には Google DeepMind の囲碁ソフト AlphaGo が韓国のプロ囲碁棋士に 4 勝 1 敗の成績を収めた[10]。2015 年 12 月には、SpaceX や Tesla Motors の CEO である Elon Musk 氏らを中心として、人工知能の研究を行う非営利団体 OpenAI を設立した。2016 年 1 月には、トヨタ自動車が人工知能研究会社 Toyota Research Institute（TRI）を米国に設立した。2016 年 9 月には、Facebook、Amazon、Google、IBM、Microsoft が AI 研究で提携すると発表した。

一方、欧州では、英国、ドイツ、フランス、スイス、イタリア、スペイン等の大学、研究機関にて基礎研究が行われている。アジアでは、中国の Baidu が深層学習の研究開発をリードしている。

国内では、人工知能学会、電子情報通信学会等で深層学習に関する研究発表が盛んになっている。産業界でも、Preferred Networks（PFN）、ドワンゴ人工知能研究所等[11]で深層学習の先端的な研究や適用事例があるほか、富士通の ZINRAI や日立の H、NEC の NEC the WISE 等、国内各社が深層学習を含む AI 技術ブランドを立ち上げて AI の産業応用を加速している。

なお、機械学習の産業応用のトピックは、第 3.3.6 節「ビッグデータによる価値創造」でも取り上げている。

### (3) 注目動向

[深層学習の性能向上]

深層学習の性能を向上させようと、さまざまな改良案が提案されている。その中でも、深層残差ネットワーク（Deep Residual Network）と呼ばれる迂回路をもつネットワーク構造は、階層を深くしても効率よく学習が行えるという特徴があり、従来の 8 倍程度の深さである 152 もの層を持つ深層残差ネットワークを用いたアルゴリズムが、2015 年の画像認識コンテストで圧勝した[8]。

深層ニューラルネットワークは学習に膨大な時間がかかるため、計算を高速化する研究は、アルゴリズムだけでなくハードウエアの観点からも行われている。NVIDIA は、グラフィックス処理プロセッサ（Graphics Processing Unit:GPU）を応用し、深層学習を

高速に行える GPU を開発している [12]。Google は Tensor Processing Unit と呼ばれる深層学習に適したプロセッサを開発し、GPU と比べて 10 倍の速さで学習が行えると発表した [13]。これらの新しいハードウエアでは、従来のように浮動小数点計算を倍精度で行わず、単精度や半精度で行うことによって高速な学習を実現している。

このように、近年は、深層学習の性能をさらに向上させるべく、データ量は巨大化し、モデルも複雑化・深層化されているため、学習時間は増加の一途をたどっている。そのため、深層学習の計算効率を改善するためのアルゴリズムやハードウエアの研究が最も重要な研究テーマの一つとして、学術界・産業界において盛んに行われている。

［機械学習・深層学習の産業界への普及］

深層学習を含む機械学習の最先端の研究開発が進められている一方で、その基本的なアルゴリズムは、オープンソースソフトウエア（OSS）として提供され、また、さまざまな商用プラットフォームにも実装され、容易に使い始められる市場環境になっている。OSS として提供されている深層学習のフレームワークとしては、Torch（Facebook 等）、TensorFlow（Google）、Chainer（PFN）、MXNet（CMU、Amazon）等がよく知られている [14]。また、企業がデータや問題をネット上で公開して、世界中の多数の人々に解かせる場（Kaggle 等）も生まれ、さまざまなデータ分析問題で解法・精度が広く競われるようになってきた。

このように道具立てとしては、機械学習・深層学習のコモディティ―化が進んでいるが、その一方で、これらの道具を使いこなせる人材（いわゆるデータサイエンティスト）は依然として不足している。深層学習はその構造設計やパラメーター設定等にノウハウや試行錯誤が必要な状況であり、道具があっても使いこなすのは必ずしも容易ではない。データと目的を与えれば結果が自動で出てくるような、簡単に使える道具・フレームワークがあれば、データサイエンティスト不足を補えて、産業応用の拡大を加速できる。目的や手法に限定はあるものの、このような設計自動化の試みは出始めている [15),16]。

また、深層学習は大量データと大規模計算パワーを要求するため、具体的な問題への適用・応用まで含めた技術発展をリードしているのは、Google、Facebook 等のネット系巨大 IT 企業である。それ以外のプレーヤーは、特に北米では、主に理論的研究に取り組む大学と、上述のコモディティ―化した道具立てを活用した応用・事業で市場に参入するスタートアップ企業という構図になりつつある。ただし、前述の非営利団体 OpenAI は、出資者メンバー連携による技術の開発とオープン化を進めており、一部の巨大 IT 企業による AI 技術の寡占化の構図に対抗する動きと見ることもできる。

［意思決定問題への取り組み］

機械学習に関わり、重要性が増しているのは、AI による意思決定問題への取り組みである。機械学習はさまざまなリアル問題に適用されるようになったが、機械学習による判別・分類や予測は、それだけでは必ずしも問題を解決するまでに至らないことも多い。例えば、スパムメールフィルターという応用シーンならば、メールをスパムか否かを判別すればそれで問題解決になるが、機器の劣化検知や資源の需要予測といった応用シーンでは、検知された劣化にどう対処するか、予測された需要に対して資源供給をどうコントロールするかというアクションまで決定（意思決定）しないと問題解決にならない。こうした機

械学習から意思決定まで通した解法として、以下のような取り組みが注目される。

まず、深層強化学習を用いるアプローチがある。強化学習（Reinforcement Learning）は、学習主体が、ある状態で、あるアクションを実行すると、ある報酬が得られるタイプの問題を扱う機械学習アルゴリズムである。より多くの報酬が得られるようにアクションを決定する意思決定方策を、アクション実行と報酬の受け取りを重ねながら学習していく。この強化学習では、ある状態で、あるアクションを実行することの良さを表す評価関数を求める必要があるが、この評価関数や方策を深層学習によって学習するのが深層強化学習である。Google DeepMind は深層強化学習を用いて、Atari ゲームのプレイ方法をほぼゼロから学習して人間よりも高いスコアを出したり、囲碁で世界トッププロに勝利したり（AlphaGo）と注目を浴びた[10]。また、PFN は CES2016 で「ぶつからない車」、CEATEC2016 でドローン制御等のデモンストレーションを展示し、ロボット等の制御への適用・実用化を進めている[17]。このアプローチは、大量の事例を容易に集められるタイプの問題、あるいは、AlphaGo がコンピューター同士の対戦を繰り返して学習したようにシミュレーションで大量事例を生成できるタイプの問題に適している。例えば、ゲーム、広告、検索、アルゴリズムトレード等があげられる。

また、上記とは異なるタイプの意思決定問題に対するアプローチとして、機械学習に基づく予測型意思決定最適化がある。深層強化学習が適するのは、大量の事例が集められる / 作れるタイプの問題、つまり、意思決定を繰り返し試すことが容易なタイプの問題である。一方、古典的なオペレーションズリサーチ（OR）で扱われているようなタイプの意思決定問題は、意思決定で失敗したときのダメージが大きく、意思決定を繰り返し試すことは難しい。例えば、小売業における商品価格設定の戦略策定や、上水道システムにおける配水制御計画等の数理最適化問題がこれに該当する。古典的な OR 問題では、データが静的であることを仮定してきたが、機械学習によって大量の予測（外れる可能性もある不確かな情報）を動的に生成することができるようになり、機械学習からの大量出力に基づく OR 問題と捉えることで新たな発展が生まれている。この具体的な取り組みとして、異種混合学習[18]による大量の予測器生成に基づく予測型意思決定最適化[19]という、機械学習 -OR パイプラインのフレームワークがあげられる。

もう一つは自然言語処理・知識ベースと機械学習の融合によるアプローチである。IBM Watson 質問応答システム[20]がその代表である。他にも、Google のメール返信文の自動生成機能「スマートリプライ」[21]でも、リカレントニューラルネットワークによる機械学習が自然言語処理と組み合わせて使われている。

### （4）科学技術的課題

深層学習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、さまざま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、深層学習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、さまざまな統計的理論解析が行われてきた[22]。一方、深層学習は非凸最適化学習手法であり[i]、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。これに関しては、例えば、ニューラルネットワークの学習で見つかる解が条件さえそろえば最適解に近いと予想される等[23]、ここ数年で理論的解析が進みつつある。

深層学習を含むニューラルネットワーク系の機械学習は、学習した規則性（モデル）を人間が理解できる形で示すことができない（判別・分類や予測の結果について理由を説明できない）ブラックボックス型である。深層学習・機械学習が社会のさまざまなシーンに適用されていくなかで、ブラックボックス型であることの問題が多く指摘され、Transparency（透明性）や Accountability（説明責任）が AI 開発原則における重要な要件としてあげられるようになってきた[24),25]。例えば、産業応用において通常は、モデル作成者（データサイエンティスト）とモデル利用者（ビジネス部門）が異なり、モデル作成者は予測結果や根拠に対するユーザーからの質問に答えることを要求される。また、医療のように人命・健康に関わるシーンでは、機械学習で導かれた結果と医学的な所見との整合性を確認でき、正しい判断のもとで利用することが求められる。さらに、公的機関での利用においては、人種・性別等によって不平等が生じないモデルであることを担保しなくてはならない。

そこで、深層学習のブラックボックス性に解釈性を与えるため、モデルがどういう特徴に反応しているかを調べて、その結果から逆にモデルが何を見ているのかを推定するアプローチや、生成モデルと認識モデルを同時に作るアプローチ等が試みられている[26]。一方、決定木や線形回帰は解釈性を持つホワイトボックス型の機械学習方式だが、精度に限界があった。これに対して因子化漸近ベイズ推論（異種混合学習）[18]という新しい理論が提案され、ホワイトボックス型のアプローチでありながら、高い精度と解釈性を両立させ得ることが示された。機械学習が社会のさまざまなシーンに浸透していくなかで、ここで述べたような解釈性への取り組みはますます重要性を増している。

また、ハードウエアに関しては、脳型の計算機構を持つ新しいプロセッサの研究も行われている[27),28]。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。

一方、応用分野によっては、ビッグデータを収集するために非常に多大なコストがかかってしまう、あるいは、原理的にビッグデータを収集できないこともある。このような場面では、限られた情報から機械学習を行う必要があり、汎化能力を持たせるための基礎研究の重要性が再認識されつつある。

こうした機械学習の基礎研究は、Neural Information Processing Systems Conference（NIPS）、International Conference on Machine Learning（ICML）等の国際会議で主に議論されている。これらの国際会議は規模が爆発的に大きくなっており、NIPS は、2013 年が 1200 人、2014 年が 2400 人、2015 年が 3800 人、2016 年は 5000 人規模の参加者があった。ICML も参加者数は、2013 年が 900 人、2014 年が 1200 人、2015 年が

i)　非凸最適化学習が可能になった主要因として、1) ReLU 関数等を活性化関数に使うことで深い層であっても勾配が発散・減衰することがなくなった、2) RMSProp、Adam 等の新しい学習則により学習が停滞するプラトーの問題を回避できるようになった、3) GPU 等の高速な演算装置の登場により現実的な時間で学習できるようになった等があげられる。

1600 人、2016 年が 3000 人と増加の一途をたどっており、第 3 次ニューラルネットワークブームの象徴となっている。

### (5) 政策的課題

国内では、文部科学省による理化学研究所革新知能統合研究センター、経済産業省による産業技術総合研究所人工知能研究センター、総務省による情報通信研究機構脳情報通信融合研究センターおよびデータ駆動知能システム研究センターが人工知能の研究開発を行っている。政府は、2016 年 4 月に人工知能技術戦略会議を設立し、その中に設置された研究連携会議が上記 3 省を通じた基盤技術の研究開発を、産業連携会議が産業界を通じた社会応用を議論するという枠組みが作られ、具体的な研究開発のロードマップ、および、重点応用分野の選定が急ピッチで行われている。しかし、3 省間のデマケーション等の制約があり、さらなる制度の柔軟化が求められる。

国際的には、Google、Facebook、Amazon、Microsoft 等の巨大 IT 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、テニュアの大学教授までをも大量に囲い込もうと躍起になっており、人材獲得競争が熾烈になっている。近年は、金融系や自動車系の企業もこの競争に参入しており、人材不足が深刻な問題となりつつある。

機械学習は米国が強いが、中国やインドは、トップ人材を組織的に米国に送り、彼らが本国に戻って活躍するという流れを作っている。また、中国やインドはシリコンバレーの人材と技術を使いこなすために、拠点を構えている。

さらに、AI・機械学習はアルゴリズムを適切なソフトウエアとして実装してこそ威力を発揮する。日本は人材育成において、理論・アルゴリズムの基礎研究に加えて、ソフトウエア開発力においても強化施策が望まれる。

また、機械学習で扱うデータや学習結果（学習済みモデル）等に関わる知的財産権の扱いやプライバシー問題は、産業応用における重要課題であり、法制度やガイドラインの整備が早急に望まれる。詳しくは第 3.3.7 節「ビッグデータに関わる制度設計」で述べる。

### (6) キーワード

機械学習、ニューラルネットワーク、深層学習、強化学習、異種混合学習、意思決定、ブラックボックス問題

### (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 | ○ | ↗ | 電子情報通信学会 IBISML 研究会は 700 人を超え、人工知能学会は会員数 4000 人、全国大会参加者数 1500 人を超える等、AI・機械学習のコミュニティは育っている。しかし、国際トップ会議（ICML、NIPS、AAAI 等）に採択される数は限られている。理化学研究所革新知能統合研究センターが機械学習の基礎研究強化を打ち出し、JST CREST、ERATO も含めて上向き。 |

| | | | | |
|---|---|---|---|---|
| 日本 | 応用研究・開発 | ○ | ↗ | NEC、富士通、日立、パナソニック、NTT、Yahoo Japan!、楽天、リクルート等が AI 分野に積極的な技術開発投資を行っている。特に深層学習に関しては、PFN がロボット制御で Amazon Picking Challenge 2016 に初出場して 2 位（1 位と同スコア）、トヨタやファナック等の大手との提携による事業強化、ライフサイエンス分野への AI 適用等も進めている。 |
| 米国 | 基礎研究 | ◎ | ↗ | 大学、企業とも機械学習の研究を非常に盛んに行っており、規模、質ともに世界をリードしている。例えば、機械学習のトップレベル国際会議の一つ ICML では、2016 年の採択論文のうち半数近くが米国発の論文であった。 |
| | 応用研究・開発 | ◎ | ↗ | 既に巨大 IT 企業となった Google、Facebook、Microsoft 等が AI・機械学習に積極的投資をしていることに加えて、Airbnb、Uber 等 AI 技術を活用したベンチャー企業が次々と誕生し、国際的に成功を収めている。例えば、Google の収益のほとんどを占める自動広告配信（年間売り上げ 6 兆円）は機械学習技術によるものであり、戦略メッセージを Mobile First から AI First へと切り替えた。DeepMind をはじめ AI・深層学習のベンチャー買収も積極的に実行している。トヨタが AI 研究開発のために設立した TRI も 1000 億円規模の予算を投入すると発表された。 |
| 欧州 | 基礎研究 | ○ | → | 英国、ドイツ、フランス、スイス、イタリア、スペイン等の大学や研究機関にて機械学習の基礎研究が盛んに行われている。 |
| | 応用研究・開発 | ○ | ↗ | ロンドンの Google DeepMind、ベルリンの Amazon Machine Learning 等、北米の企業の欧州支社が中心となり、応用研究開発を行っている。特に DeepMind が基礎・応用の両面で存在感を増している。ICML や NIPS 等での採択率もトップクラスである。 |
| 中国 | 基礎研究 | ○ | ↗ | 機械学習の主要な国際会議である ICML を 2014 年に北京でホストする等、当該分野の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ある。清華大学・MSRA（Microsoft Research Asia）等を中心に、機械学習の国際会議での中国からの採択数が伸びている（国別で米国に次ぐ）。 |
| | 応用研究・開発 | ○ | ↗ | Baidu、Huawei Noah’s Ark Lab、MSRA、Horizon Robotics 等、企業による応用研究開発が活発に進められている。 |
| 韓国 | 基礎研究 | △ | → | ソウル大学、KAIST、POSTECH 等の主要大学にて関連の研究は行われているが、国際的に顕著なものは多くない。 |
| | 応用研究・開発 | △ | ↗ | 韓国の大企業が共同で出資して設立した民間の研究所である知能情報技術研究院（AIRI）が 2016 年 10 月に始まり、応用研究のトレンドは上昇しつつある。また、Samsung 等も応用研究開発に力を入れている。 |

（注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
（注２） 現状 ※わが国の現状を基準にした評価ではなく、CRDS の調査・見解による評価である。
◎ 特に顕著な活動・成果が見えている、 ○ 顕著な活動・成果が見えている
△ 顕著な活動・成果が見えていない、 × 活動・成果がほとんど見えていない
（注３） トレンド
↗：上昇傾向、 →：現状維持、 ↘：下降傾向



## (8) 参考文献

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）ImageNet, http://image-net.org/
9）RBB Today, 2012 年 6 月 27 日号 ,
http://www.rbbtoday.com/article/2012/06/27/90985.html
10）Wikipedia AlphaGo 対李世ドル ,
https://ja.wikipedia.org/wiki/AlphaGo%E5%AF%BE%E6%9D%8E%E4%B8%96%E3%83%89%E3%83%AB
11）特集「国内人工知能研究拠点紹介」, 人工知能学会誌 , 31（3）～（5）, 2016.
12）NVIDIA,
https://images.nvidia.com/content/APAC/events/deep-learning-day-2016-jp/NVIDIA-DeepLearning-Intro.pdf
13）日経コンピュータ ,
http://itpro.nikkeibp.co.jp/atcl/ncd/14/457163/052001464/
14）Wikipedia, Comparison of deep learning software,
https://en.wikipedia.org/wiki/Comparison_of_deep_learning_software
15）三菱電機 , ディープラーニングの自動設計アルゴリズム , 2016.10.7
http://www.mitsubishielectric.co.jp/news/2016/1007.html?cid=rss
16）NEC, 予測分析自動化技術 , 2016.12.15.
http://jpn.nec.com/press/201612/20161215_06.html
17）Preferred Networks,
https://www.preferred-networks.jp/ja/news/ceatec2016
18）Fujimaki, R. and Morinaga, S.: Factorized Asymptotic Bayesian Inference for Mixture Modeling, Proceedings of AISTATS, 2012.
19）藤巻遼平、村岡優輔 , 伊藤伸志 , 矢部顕大 , 予測から意思決定へ ～予測型意思決定最適化～ ,NEC 技報 69（1）,2016.
http://jpn.nec.com/techrep/journal/g16/n01/pdf/160115.pdf
20）IBM Corp., “This is Watson”, Special Issue of IBM Journal of Research and Development, Issue 3-4, May-June 2012.
21）Google Research Blog, 2015.11.3
https://research.googleblog.com/2015/11/computer-respond-to-this-email.html
22）Vapnik, V. N., Statistical Learning Theory, Wiley, 1998.
23）Kawaguchi, K.: Deep Learning without Poor Local Minima, Proceedings of NIPS, 2016.

24）Fairness, Accountability, and Transparency in Machine Learning, http://www.fatml.org/
25）ITpro, ニュース解説（2017.1.16), 総務省が AI 開発ガイドライン作成へ , http://itpro.nikkeibp.co.jp/atcl/column/14/346926/011200764/
26）Park, D.H., Hendricks, L.A., Akata, Z., Schiele, B., Darrell, T., Rohrbach, M.: Attentive Explanations: Justifying Decisions and Pointing to the Evidence, https://arxiv.org/abs/1612.04757
27）http://www.ibm.com/smarterplanet/jp/ja/brainpower/
28）http://news.mynavi.jp/news/2016/09/05/204/

### 3.3.3　画像・映像解析技術

#### (1) 研究開発領域の簡潔な説明

カメラ等で撮影された画像や映像をコンピューターで解析して、その画像･映像の内容、つまり、そこに写っているもの（人や物体、文字等）の位置・カテゴリあるいは風景・場所・状況等を認識する技術である。産業の効率化、社会の安全安心、生活の快適性に役立つものであり、具体的な応用先としては、例えば、文字認識、医療画像による診断支援、監視カメラからの不審者・不審行動や異常状況の検知、インターネット上の画像検索・動画像検索、製品の検査、顔や指紋からの個人認証、スポーツ画像解析、ロボットビジョンや自動車の自動運転、動作認識によるヒューマンインターフェースデバイス等があげられる。

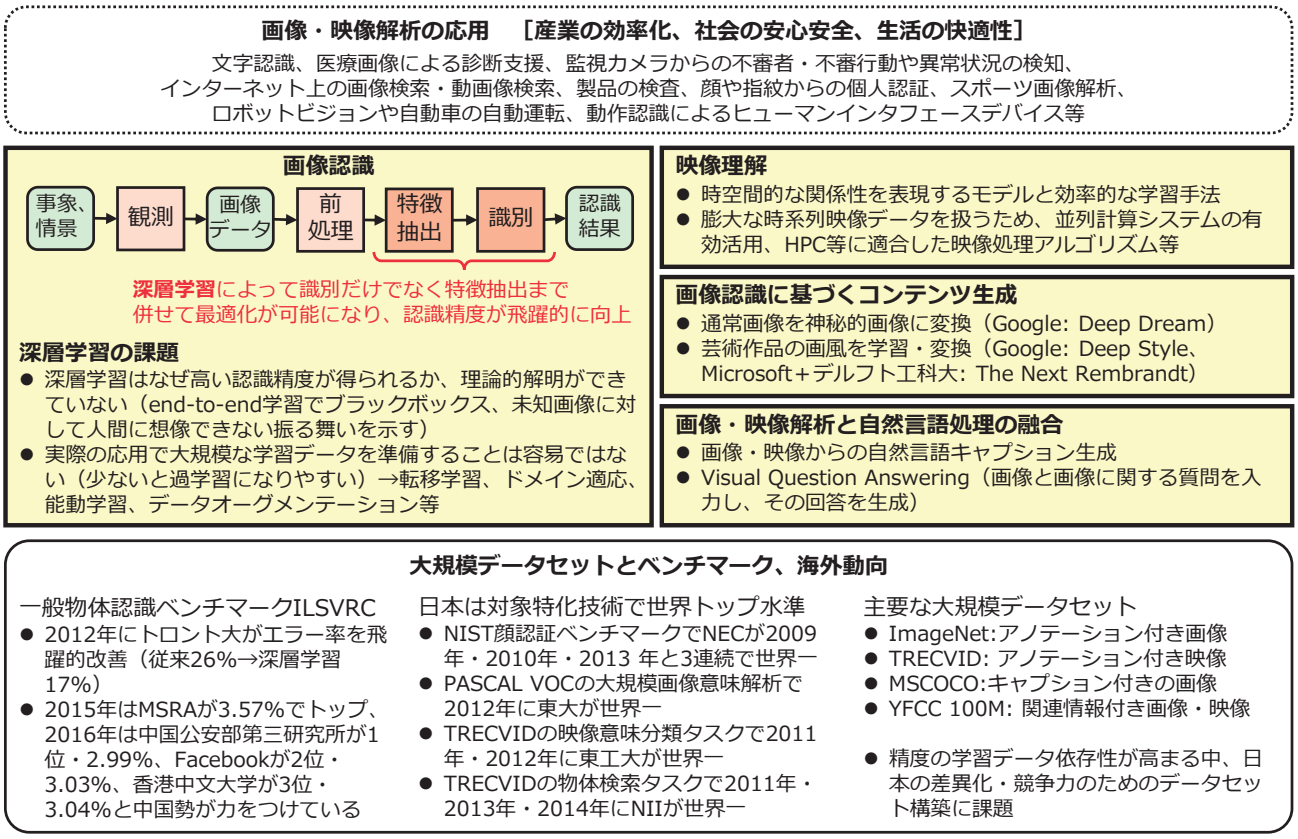


図 3-3-4　領域俯瞰 : 画像・映像解析技術

#### (2) 研究開発領域の詳細な説明と国内外の動向

［画像・映像のパターン認識］

画像・映像解析はパターン認識を画像・映像に適用するのが基本的な実現形になる。そのパターン認識処理は、観測、前処理、特徴抽出、識別から構成される。

観測は、実世界の事象を処理可能なデータに変換する処理である。画像・映像のデジタルデータは可視光カメラから得るのが一般的であるが、カメラから被写体までの距離情報を表す深度画像（デプス画像）､赤外線カメラ画像､マルチスペクトル画像等も用いられる。実世界は 3 次元立体であるが、カメラで撮影されるデータは 2 次元平面のため、被写体の姿勢変動や照明変動の影響で、被写体の見えが大きく変化する。また、隠れ（オクルー

ジョン）によって情報の欠落も生じる。これらの変動を受けにくいセンシングをどう設計するかが、観測の課題である。文字認識やウェハ欠陥検査等は、被写体自身が 2 次元平面のため、姿勢変動や照明変動の影響を受けにくく、産業を効率化する技術として早期に実用化されている。一方、情景中の 3 次元オブジェクトは、姿勢や照明の変動を受けやすいため課題も多い。そのため、エラー発生にシビアな車の自動運転では、前方障害物のセンシングにレーザーレンジセンサー等の測域センサーも用いられている。

前処理は、以降の処理にかかる演算量を軽減するための処理であり、具体的にはデータの正規化やノイズの除去が行われる。画像からの対象領域の切り出しや白色化と呼ばれる信号の大きさの正規化、対象の姿勢の正規化、不明瞭な領域の S/N 比（信号雑音比）を改善する鮮鋭化やデヘイズ処理､あるいは画像の解像度を上げる超解像処理等も含まれる。これら画質を改善する処理は、視認性を高めることを目的に利用されることもある。

特徴抽出は、前処理後の画像・映像から識別に有効な特徴を抽出する処理である。主に局所フィルターを用いて、エッジやコーナー等の画像特徴が抽出されることが多い。処理の高速化、あるいは次元の呪いと呼ばれる認識精度の低下を抑えるために、得られた多数の特徴を少数個に低次元化する処理も含まれる。具体的には、識別に有効な特徴の組み合わせを選ぶ特徴選択や、識別に有効な特徴に作り直す線形あるいは非線形な特徴変換が行われる。

識別は、特徴抽出で得られた特徴をもとに、あらかじめ設定したクラス（あるいはカテゴリと呼ぶ）に分類する処理である。特徴抽出で得られた多数の特徴値を多次元特徴ベクトルとみなし、クラスごとにあらかじめ設定した代表ベクトルとの内積や距離に基づいて識別する方法や、特徴抽出で得られた特徴点の集合と、クラスごとにあらかじめ設定した特徴点の集合のマッチングに基づいて識別する方法等がある。クラスは目的に応じて人間が設定するものであり、例えば人物と車両を識別する場合は、それぞれが一つのクラスとして設定される。一方、顔認証では、人物一人一人を識別する必要があるため、それぞれが一つのクラスとして設定される。

[深層学習を中心とした最近の動向]

パターン認識で高い精度を実現するには、観測から識別までの全体処理を最適設計することが重要である。観測と前処理は、専門家がこれまでの経験に基づき、目的に応じて設計している。識別は、計算機処理の発展とともに機械学習の技術開発が進み、テンプレートマッチングと呼ばれる単純な手法から機械学習で自動設計する手法に既に移行した。一方、特徴抽出は、従来専門家が経験に基づき設計するのが一般的であったが、2010 年頃からの深層学習 [1)] の進展に伴い、特徴抽出と識別を合わせて機械学習で自動設計できるようになってきた。

深層学習（Deep Learning）とは、深い層構造からなるニューラルネットモデル Deep Neural Network（DNN）の学習方法である。2010 年頃までは AutoEncoder や Restricted Boltzmann Machine といった教師なし学習に基づく手法が研究されていたが、現在では後述する CNN をベースとした教師あり学習に基づく手法が主流となっている。深層学習が登場した背景として、1）インターネットや SNS の普及によってインターネット上に大量の画像がアップロードされ、かつクローリング等で容易にダウンロードできる

ようになったこと、2）クラウドソーシングによって安価に画像の正解付けができるようになったこと、3）GPGPU を含む計算機性能の向上により、ネットワーク構造の決定に必要な多くの試行錯誤が可能になったこと、があげられる。

現在主流となっているモデルは、福島邦彦が 1979 年に発表したネオコグニトロン[2)]がもとになっている。畳み込み層（S 層）とプーリング層（C 層）の組を複数積み重ねることで、パターンの局所変動に頑健になることが示されていた。それを誤差逆伝播学習法で最適化できるようにしたのが、1989 年に Yann LeCun が発表した畳み込みニューラルネットワーク Convolutional Neural Network（CNN）である。MNIST 等の手書き数字データベースで有効性が示されていたが、ネットワーク構造が深くなるほど伝播される誤差が小さくなり、学習が進まなくなる問題が指摘されていた。トロント大の Geoffrey Hinton 教授らは、誤差が深い層まで伝播するように活性化関数や正則化を工夫することでこの問題を解決し[3)]、CNN をさらに深くしたニューラルネット構造を用いて、2012 年に一般物体認識のベンチマークテスト Large Scale Visual Recognition Challenge（ILSVRC）[4)]でトップを獲得した。Hinton 教授らは、従来手法がエラー率 26% だったのに対してエラー率 17% と、深層学習の適用によって一気に約 10% もの飛躍的な精度向上を達成したことが、画像認識・機械学習の研究者らに衝撃を与えた。ILSVRC を舞台として現在も深層学習の改良が続いているが[5)-7)]、試行錯誤的な改良に終始しており、なぜ深層学習で高い認識精度が得られるかは、理論的にはまだ解明されていない。

深層学習の技術開発を先導しているのは、Google、Microsoft、Facebook 等の米国の巨大 IT 企業であり、膨大な画像と潤沢な計算リソースを保有していることが、その背景にある。また、深層学習がブームになった後、トロント大の Hinton 教授は 2013 年から Google 兼務、ニューヨーク大の LeCun 教授は 2013 年から Facebook 兼務になる等、人材の囲い込みが今も進んでいる。米国のコンピュータービジョンとパターン認識に関する著名な会議 CVPR で 2016 年のベストペーパーに選ばれた Deep Residual Network[5)] を提案した Kaiming He は、2015 年の ILSVRC でトップを獲得した時は MSRA（Microsoft Research Asia）に所属していたが、直後に Facebook に移ったことが話題になった。

中国の Baidu（百度）は、2013 年にシリコンバレーに Institute of Deep Learning を設立し、Google の AI プロジェクトの元責任者である Andrew Ng をトップに迎える等、深層学習に力を入れている。深層学習を使った自動運転技術の開発では NVIDIA とも提携し、2018 年末までに自動運転タクシーを実用化するという[8)]。既に米カリフォルニア州の認可を受け、同州で走行テストを始めている。その背景には、中国の AI 技術に対する国策がある。中国政府は、2018 年までに同国の AI 技術者を世界と同水準にし、中堅企業を育成して、2025 年までに AI 産業で世界をリードするという目標を掲げている。

日本では、産業の効率化として古くから文字認識の研究が進められ、郵便番号読み取り区分機を世界で初めて実用化したのを始め、マンモグラフィ等の医用画像スクリーニング、半導体製造の際のウェハやフォトマスクの欠陥検査等が実用化されている。バイオメトリクス認証としては、虹彩認証、指紋認証、静脈認証、顔認証が実用化されている。静脈認証は世界で初めて日本国内の銀行で採用され、顔認証は 2001 年に発生した米国同時多発テロ以降のボーダーコントロール強化策として、世界各国に導入され始めている。日本の画像認識技術が世界トップ水準にあることは、国際的ベンチマークテストの成績でも示さ

れている。NIST の顔認証ベンチマークテストでは、NEC が 2009 年・2010 年・2013 年と 3 連続で世界第 1 位を獲得している [9]。ImageNet[10] データを用いた PASCAL VOC[11] における大規模画像意味解析コンペティションで東京大学（原田達也教授）のチームが 2012 年に世界第 1 位の認識精度を達成、TRECVID[12] の映像意味分類タスク（SIN）で東京工業大学（篠田浩一教授）のチームが 2011 年・2012 年に世界第 1 位の認識精度を達成、TRECVID の物体検索タスク（INS）で国立情報学研究所（佐藤真一教授）のチームが 2011 年・2013 年・2014 年に世界第 1 位の検索精度を達成している。

### (3) 注目動向

［ILSVRC、一般物体認識］

深層学習の一般物体認識ベンチマークとしてよく知られる ILSVRC の 2016 年の結果が公開され [4]、上位 5 位までに正解が含まれないエラー率は、中国公安部第 3 研究所が 2.99% で 1 位となった。2 位の Facebook が 3.03%、3 位の香港中文大学が 3.04% で、2015 年のトップが MSRA の 3.57% だったことを考えると、本テストに対する精度は、ほぼ飽和状態とみられる。2015 年は、MSRA が Deep Residual Network と呼ぶ画期的なアイデアを提案してトップを獲得したが、今回は技術的には既存モデルの統合であり、面白みはないものの、中国は着実に力をつけてきていることが注目される。中国の Baidu も有力とみられていたが、ルール違反を犯したため、今回は参加が禁じられている [13]。

［画像・映像解析と自然言語処理の融合］

注目分野として、画像や映像と自然言語処理の融合がある。画像や映像に対して、その内容を自然言語で記述したキャプション（例えば「赤い服を着た女性が街中で電話をしている」等）を生成するタスクがその一例である [14]。このタスクを実施するためには、静止画像における ImageNet のような高品位のデータセットが必要となるが、画像キャプション生成では MSCOCO[15] がよく利用される。また、画像と自然言語処理を融合した新しい問題設定として Visual Question Answering（VQA）[16] がある。これは、計算機に画像と画像に関する質問を入力し、この質問に回答する知的システムを開発する課題である。

［画像認識に基づくコンテンツ生成］

画像認識によってセマンティクスに近い圧縮された情報が得られるが、それにグラフィックス系を融合し、実世界情報を生成するコンテンツ生成技術が進展している。2015 年に Google が Deep Dream[17] と呼ばれるシステムを開発し、大きな話題となった。Deep Dream は、通常の画像を夢に出てくるような見たこともない神秘的な画像に変換するシステムである。さらに Google は Deep Style と呼ばれる入力画像の画風を変換するシステムを開発し、人工知能が製作する芸術作品としてメディア等で取り上げられている。Microsoft とデルフト工科大学は Rembrandt（レンブラント）の画風を学習し、3D プリンターを利用して新作を描く The Next Rembrandt プロジェクト [18] を行っている。

### (4) 科学技術的課題

[深層学習の課題]

画像認識で高い精度を達成し、現在最強力な手法と考えられる深層学習だが、いくつかの課題が指摘されている。（参考：深層学習とその課題については第 3.3.2 節「機械学習技術」でも論じている）

まず、深層学習は、100 層以上にも及ぶ多層のニューラルネットを用い、パラメータ数が多いことから過学習になりやすい問題を内在している。過学習を軽減し、高い認識精度を得るには、大規模な学習データが必要である。しかしながら、実際の応用で大規模な学習データを準備することは容易ではない。学習データを十分に集められないケースには、1）データは大量にあるが、正解ラベルがないため学習に使えない、2）正解ラベルが付いた類似のデータは大量にあるが、認識したいタスクとは異なるため使えない、3）正常データは大量にあるが、異常データは極めて少ない、4）そもそもデータ自体が少ない、等がある。

1）に対しては、いかに低コストで正解ラベルを付与するかが課題であり、クラウドソーシングを利用して人件費を削減するやり方の他、精度改善に有効な一部のデータだけに効率良く正解ラベルを付与する能動学習や、一部のラベルありデータを手掛かりにラベルなしデータも学習させる、半教師あり学習が求められる。2）に対しては、大量の類似データで学習した DNN を初期値として、別タスクでファインチューニングする転移学習、類似データの分布を別タスクのデータ分布に合わせることで類似データを学習に流用するドメイン適応等が求められる。3）に対しては、二つのクラスのサンプル数のインバランスを考慮した学習や、異常データを正常データの外れ値とみなして検出する手法が求められる。4）に対しては、物理シミュレーション等でデータを生成する方法や、パターン変動に関する先見知識を反映してデータを生成する、データオーグメンテーションと呼ばれる方法等が求められる。

また、深層学習はなぜ高い認識精度が得られるか、理論的解明ができていない。深層学習は大量の学習データに基づく end-to-end 学習、つまり、入力と望ましい出力の対からそれらの間の関係性だけを学習する方法で、それを実現するメカニズムはブラックボックスである。画像の意味分類が end-to-end でできるようになったが、これは画像の意味を認識したことになるのかという疑問が生じ、その解明も兼ねて、画像からのキャプション生成や質問応答の研究が進んだ。しかし、これらにおいても end-to-end の方式が高い精度を示し、結局、ブラックボックスであることは変わっていない。さらに、これと表裏一体の問題として、学習時とは異なる未知画像に対して、人間には想像できない振る舞いを示すという問題も明らかになっている。砂の嵐のような奇妙な画像なのにペンギン等と認識されてしまう画像、どう見てもパンダなのにテナガザルと断言されてしまう画像等が生成可能であることが指摘されている [19),20)]。

[映像理解の課題]

静止画像認識はここ数年で大きな飛躍が見られたが、映像については現在まだその途上にある。実映像に対して時空間的な関係性を適切に表現するモデル構築、効率的な学習手法、これらを支える理論基盤の多くに課題が残されている。また、時系列の映像データは

量が膨大となるため、ストレージや計算パワーが大量に必要であり、並列計算システムの有効活用や、HPC（High-Performance Computing）等に適合した映像処理アルゴリズムの構築も重要な課題となる。

映像の内容理解・検索の技術発展には NIST の TRECVID[12] が大きく貢献してきた。また、ImageNet のような大規模で高品位なデータセットは映像理解においても重要であることから、さまざまな動画データセットの開発が進められている。例えば、4800 クラス 800 万本 50 万時間の動画からなる YouTube-8M[21]があげられる。この規模のデータセットになると大学レベルでの実現は困難であり、企業・政府等の力が必要となる。

### (5) 政策的課題

［大規模データセット構築とベンチマークテスト］

この研究分野では、大規模な共通データセットを構築し、同じ土俵で適正に競争することが技術発展に大きく貢献してきた。一般物体認識ベンチマークテスト ILSVRC では、上位 5 位までに正解が含まれないエラー率が、2011 年に 25.77% だったのが 2016 年には 2.99% に改善された。NIST の顔認証ベンチマークテストでは、他人受理率を 0.1% とした時の本人棄却率が、1993 年に 79% だったのが 2002 年に 20%、2010 年には 0.3% にまで改善された。なお、ILSVRC では 1000 クラスに対して 120 万枚を学習し、10 万枚で評価される。NIST の顔認証ベンチマークテストでは、160 万枚に及ぶ顔画像で評価されている。企業においては、大規模データセットは技術競争力の源泉として内部に蓄積してきたが、最近、Yahoo! Research の研究者らが、画像・映像合計 1 億にも及ぶデータセット YFCC 100M[22] をさまざまな関連情報付きで公開した。オープイノベーションの流れと推測される。

日本国内でも、かつては電子技術総合研究所（現、産業技術総合研究所）が ETL9 と呼ばれる JIS 第 1 水準手書き漢字データセットを公開し[23]、特に大学を中心に国内の手書き漢字認識に関する技術開発が大いに進展した経緯がある。しかし、クラス当たり 200 枚の画像であり、今となっては極めて小規模なものである。NIST の顔認証ベンチマークテストで日本企業が上位成績を収めている背景には、企業各社が独自に構築した顔画像データセットがある。ILSVRC で米国ネット企業が好成績を収めているのも、大量画像データ収集において優位な立場にいることが要因であろう。したがって、この分野で日本の技術力を高めるためには、日本独自の大規模データセットを国が主導して構築し、大学・国研、企業も含めて適正な技術競争ができる土壌を造ることが最も有効である。特に、インターネット上では容易に得られない実世界の画像・映像データにフォーカスし、例えば RGB-D のような深度情報も含めたデータセット構築が、今後の技術競争力を高める上で重要であろう。また、公共の場でのデータ収集や社会実証行う際の規制緩和も、検討すべき政策の一つである。

［長期視点での研究育成、人材施策］

昨今の画像・映像解析は深層学習一色だが、かつてのニューラルネットワークブームのあと 2012 年頃まで世界の第一線で活躍する画像認識研究者において深層学習の専門家は

ごく少数派だった。深層学習の成功は少数の研究者の長期間にわたる地道な努力が実を結んだものであって、流行に便乗することで成功したわけでは決してない。今後また全く新しい計算機アーキテクチャの出現により、深層学習とは異なる新規手法が生み出され、時代遅れとみなされた手法が再度注目される可能性も十分あり得る。そのために、すぐに役立つ技術だけではなく、10 年後、20 年後に役立つ可能性のある研究を支えていく制度が重要である。

一方で、深層学習を用いた実世界の画像・映像解析は、実社会のさまざまな課題解決に結びつき得る。複雑な実社会の課題分析から技術的・理論的にチャレンジングな研究ターゲットが見いだされ得る。このような課題解決では、画像・映像解析技術だけでなく、社会課題の分析や、音声・自然言語処理やロボティクスとの融合といったより広く複合的なアプローチのできる人材育成が求められる。このような境界領域研究の奨励や博士課程学生への支援制度、その後のキャリアパスの確保等も課題である。さらに、この分野ではグローバルな人材獲得競争が激化しているが、グローバル競争力の面で日本の給与体系は弱く、海外からの優秀な研究者・技術者の招聘による底上げが難しく、日本からの人材流出も懸念される。

## (6) キーワード

深層学習、画像認識、映像理解、マルチメディア、大規模データ、メディア処理、パターン認識、検索、識別、認証、意味理解

## (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 | ○ | ↗ | NEC、オムロン、富士通等の企業で取り組まれているバイオメトリクス認証（顔・静脈等）を中心には世界的に競争力がある。大学や国研では一般画像の認識等の研究が進められている。ILSVRC、TRECVID では東大、東工大、国立情報学研究所等による世界的にも競争力のある技術が開発されている。産総研、理研や多くの大学で AI 研究センターが立ち上がり、基礎研究を支援する体制が整いつつあるが、現状、国際会議での存在感は中国ほど大きくない。 |
| | 応用研究・開発 | ○ | → | 顔認証では NEC が NIST ベンチマークで 3 回続けてトップを獲得し、世界的にも大きな存在感を示している。オムロンの OKAO Vision[24] は世界で 5 億ライセンスを出荷、世界トップレベルのシェアを持つ。静脈認証も日立や富士通が銀行向けに実用化し、国内大手電機メーカーの技術力は世界的に見ても高いものがある。 |
| 米国 | 基礎研究 | ◎ | → | Stanford、UC Berkley、CMU 等の大学で基礎研究が盛んに行われている。大学のみならず巨大 IT 企業 Google、Facebook、Microsoft 等の研究所でも世界中の有能な研究者を集め、基礎研究を推進、この分野を主導。また、基礎研究に必要なデータセットの多くが米国の大学・巨大 IT 企業によって公開されている。研究すべきタスクの設定や研究コミュニティーへの情報発信等でも中心的な役割を果たしている。例えば Stanford 大が一般物体認識向け大規模データベース ImageNet 構築し、深層学習のベンチマークテスト ILSVRC を主導。NIST が顔認証・指紋認証を初めとしてデータセット構築とベンチマークテストを継続的に実施。 |

| | | | | |
|---|---|---|---|---|
| 米国 | 応用研究・開発 | ◎ | ↗ | Google、Microsoft、Facebook 等の巨大 IT 企業では有能な技術者を全世界から集め、応用研究や開発が盛んに行われ、大学との連携も盛んに行われている。大学でも起業を目指した応用研究や開発も数多く実施されている。深層学習を利用した画像・映像処理のためのツール群を無償で公開し、画像・映像処理用のサービスをクラウド上で実施する等、他企業の応用研究のサポートも実施している。 |
| 欧州 | 基礎研究 | ◎ | ↗ | Oxford 大、ETH、Amsterdam 大、INRIA、Max Planck 等に優秀な研究者が多数在籍、基礎研究力が高い。Google DeepMind、Facebook AI Research、Qualcomm 等の企業の研究部門での基礎研究もインパクトのある成果をあげている。PASCAL VOC プロジェクト（2005 ～ 2012 年）が画像認識研究の底上げに大きく貢献した。ImageCLEF、MediaEval でデータセットを公開、行うべきタスク設定を行っている。 |
| | 応用研究・開発 | ○ | ↗ | FP7 に続く Horizon 2020 が 2014 年に開始され産官学連携のプロジェクトや応用研究・開発が推進されている。上述の企業の研究部門は、応用研究でもインパクトある成果をあげている。DeepMind は、2014 年に Google に買収された。 |
| 中国 | 基礎研究 | ○ | ↗ | MSRA が ILSVRC 2015 の多部門でトップ成績を出し、注目を浴びた。北京大・清華大・香港中文大等はトップ会議におけるプレゼンスを高めている。欧米の有力研究者と連携で基礎研究の底上げを行っている例が多く見られる。欧米留学が多いことも研究レベルの底上げに寄与。 |
| | 応用研究・開発 | ○ | ↗ | MSRA、Baidu、アリババグループ等が応用研究・開発向けのソフトウエア公開・サービス提供を始めている（Baidu の PaddlePaddle を提供、アリババグループでは画像・映像処理も行える AI システム ET をクラウド上で実施）。 |
| 韓国 | 基礎研究 | ○ | → | トップ会議において KAIST・ソウル大学校等のプレゼンスがある。画像・映像圧縮、コンピュータービジョン分野では顕著な成果がある。 |
| | 応用研究・開発 | △ | ↗ | サムスン等での応用研究はあるが、日本企業ほどのプレゼンス・影響力はない。韓国政府が人工知能を国家戦略プロジェクトにすえて、2026 年までに AI 専門企業を 1000 社、専門人材を 3600 人育成する計画がある。 |

（注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
（注２） 現状　※わが国の現状を基準にした評価ではなく、CRDS の調査・見解による評価である。
◎　特に顕著な活動・成果が見えている、　○　顕著な活動・成果が見えている
△　顕著な活動・成果が見えていない、　×　活動・成果がほとんど見えていない
（注３） トレンド
↗：上昇傾向、　→：現状維持、　↘：下降傾向

## (8) 参考文献

1）麻生英樹 , 安田宗樹 , 前田新一 , 岡野原大輔 , 岡谷貴之 , 久保陽太郎 , ボレガラ・ダヌシカ（人工知能学会監修 , 神嶌敏弘編),「深層学習―Deep Learning―」, 近代科学社 ,2015.
2）福島邦彦 , 位置ずれに影響されないパターン認識機構の神経回路モデル―ネオコグニトロン―, 電子通信学会誌 J62-A（10),pp. 658-665,1979.
3）A. Krizhevsky, I. Sutskever, and G. E. Hinton, ImageNet Classification with Deep Convolutional Neural Networks, NIPS, 2012.
4）ILSVRC, http://www.image-net.org/challenges/LSVRC/
5）K. He, X. Zhang, S. Ren, and J. Sun, Deep Residual Learning for Image Recognition, CVPR, 2016.
6）Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf, DeepFace: Closing the Gap to Human-Level Performance in Face Verification, CVPR, 2014.
7）Florian Schroff, Dmitry Kalenichenko, and James Philbin, FaceNet: A Unified

Embedding for Face Recognition and Clustering, CVPR, 2015.
8）https://blogs.nvidia.com/blog/2016/08/31/baidu-nvidia-worlds-first-map-to-car-platform-self-driving-cars/
9）http://jpn.nec.com/rd/research/DataAcquition/business.html
10）ImageNet, http://image-net.org/
11）PASCAL VOC　http://host.robots.ox.ac.uk/pascal/VOC/
12）TRECVID, http://trecvid.nist.gov/
13）http://ameblo.jp/karurosu2013/entry-12039873013.html
14）A. Karpathy and L. Fei-Fei, Deep Visual-Semantic Alignments for Generating Image Descriptions, CVPR, 2015.
15）MSCOCO, http://mscoco.org/
16）S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, VQA: Visual Question Answering, ICCV, 2015.
17）Deep Dream, http://deepdreamgenerator.com/
18）The Next Rembrandt, https://www.nextrembrandt.com/
19）Nguyen, Yosinski, and Clune, Deep Neural Networks are Easily Fooled, CVPR, 2015.
20）Goodfellow, Shlens, and Szegedy, Explaining and Harnessing Adversarial Examples, ICLR, 2015.
21）YouTube-8M, https://research.google.com/youtube8m/
22）YFCC 100M, http://www.yfcc100m.org/
23）ETL 文字データベース ,http://etlcdb.db.aist.go.jp/?lang=ja
24）OKAO Vision, http://plus-sensing.omron.co.jp/technology/

### 3.3.4　自然言語処理技術

#### (1) 研究開発領域の簡潔な説明

自然言語は人間が日常の意思疎通のために用いる自然発生的な記号体系である。自然言語処理技術は、コンピューターによって自然言語を処理する技術であり、その主な応用分野には、(A) 自然言語で表現された大量テキスト情報の活用を促進・支援する情報検索・情報抽出・情報分類・質問応答や、(B) 自然言語によるコミュニケーションを促進・支援する機械翻訳・対話システム等がある。

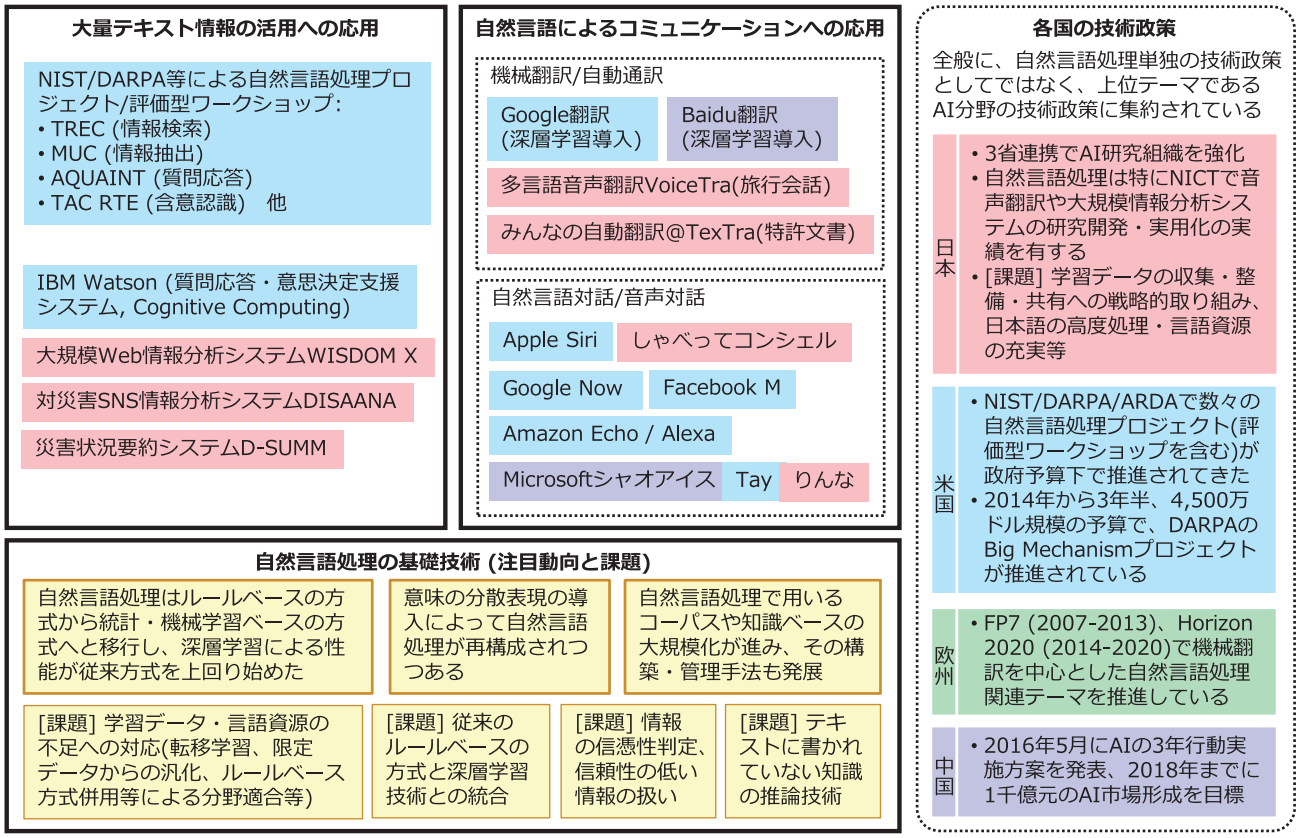


図 3-3-5　領域俯瞰 : 自然言語処理技術

#### (2) 研究開発領域の詳細な説明と国内外の動向

自然言語処理技術において共通的に必要とされる基礎技術はコンピューターによる自然言語理解であり、形態素解析、構文解析、文脈・意味解析（語義の曖昧性解消や照応解析を含む）というステップで、より深い理解の実現が試みられてきた。そのアプローチは、黎明期の 1950 年代から 1990 年代頃まで、人間が記述した辞書・文法を用いるルールベース方式が主流だった。しかし、近年は大量のテキストデータ（テキストのビッグデータ）が利用可能になったことや、機械学習技術が大幅に進化したことから、統計的な方式、機械学習を活用した方式に主流が移っている。

[応用分野（A）大量テキスト情報の活用の促進・支援]

主な応用分野（A）（B）に目を向けると、まず、(A) 自然言語で表現された大量テキスト情報の活用がさまざまなシーンに広がっている。

テキストのビッグデータは、上で述べたような自然言語処理技術を改良・精緻化するための訓練・学習データとして使われる側面もあるが、本来、人々の知識源として役立ち得るものである。しかし、いまやその量が人間の扱える規模を超えてしまっている。例えば、膨大な科学技術論文や Web 等に書かれた専門的知識のすべてをカバーすることは、専門家・研究者でももはや不可能に近い。また、災害時にバースト的に流通する多様な災害関連テキスト（SNS、ニュース、報告書等）を精査・統合して適切なアクションを決定することも、時間制約・人資源制約のある状況では極めて困難である。その結果、新たな科学的発見のチャンスの見落としや、不十分な状況判断によるアクションミス等が起きてしまう。キーワード検索による絞り込みは、テキストのビッグデータを前にして広く利用されている手段であるが、上記例のようなケースを解決することはできない。

この問題に対して、テキストに記載されている情報を構造化した形（例えば、いつ、どこで、誰が、何をしたか）で取り出す情報抽出技術や、テキスト間の同義性や含意関係（文 A から文 B が推論できるような関係）を認識する含意関係認識技術等、テキストの内容解釈に踏み込む研究が盛んになった。情報抽出技術については 1980 年代終わり頃、含意認識技術については 2000 年代半ば頃から評価型ワークショップも開催され [1]、テキストの内容解釈に関わる基礎技術は着実に進展してきている。

近年は、大量テキスト情報を知識として活用する具体的応用システムが開発され、テキストの内容解釈に踏み込んだ技術の有効性・実用性が示されるようになった。IBM の Watson[2] は、大量テキスト情報を知識源として自然言語で書かれた質問に回答する技術を中核とし、米国の人気クイズ番組「Jeopardy!」で人間のクイズ王に勝利するというグランドチャレンジを成功させ、大きく注目された。Google の Knowledge Vault[3] は、大規模 Web 情報から事実を抽出して知識ベース化している。日本では、2016 年の熊本地震の際に政府でも活用された対災害 SNS 情報分析システム DISAANA、その拡張版である災害状況要約システム D-SUMM、大規模な Web 情報をもとに質問に回答する WISDOM X 等が、情報通信研究機構（NICT）から公開されている。

［応用分野（B）自然言語によるコミュニケーションの促進・支援］

自然言語によるコミュニケーションの分野では、機械翻訳（あるいは自動通訳）と自然言語対話（あるいは音声対話）が二大応用である。いずれも「夢の技術」として古くから技術的挑戦が続けられているが、最近は実際の場面で使われる技術になってきた。

機械翻訳の方式は、ルールベースの方式から、大規模な対訳コーパス（元言語のテキストとターゲット言語のテキストを対にしたもの）と機械学習技術を用いた統計的機械翻訳へシフトした。さらに、検索サービス大手の Google や Baidu は、対訳コーパスからの深層学習による end-to-end 翻訳のアプローチを、インターネット上で公開している翻訳サービスに導入し始めている [4),5)]。

また、分野を限って高品質な対訳コーパスを大量に収集し、そこから学習すると非常に高品質な翻訳システムを構築できることが知られている。日本では例えば、旅行会話にフォーカスしたスマートフォン版多言語音声翻訳アプリ VoiceTra が NICT で開発され、多数の民間企業、スポーツイベントでの案内や病院診察での外国人対応等に使われ始めている。2020 年の東京オリンピックに向けた総務省のグローバルコミュニケーション計画

では外国人観光客対応の活用も計画されている。NICT と特許庁が共同開発した特許文書翻訳システムは「みんなの自動翻訳 @TexTra」として公開され、特許庁では日常的に活用されている。

一方、自然言語対話のシステムは、長年にわたってプロトタイプの域を出ることはなかったが、2011 年に Apple の Siri、NTT ドコモの「しゃべってコンシェル」等が出現してから状況が変わった。これは音声認識技術の進歩だけでなく、スマートフォンの普及によるところが大きい。さらに、Amazon Echo は家庭向けの音声対話アシスタント内蔵スピーカーとも呼ばれるような形態で大きく注目され、今後、対話機能を備えた新しいタイプの製品が生まれていく可能性を示した。

技術的には、Siri の初期のモデルや「しゃべってコンシェル」等はプログラマーが記載したルールと従来型の機械学習の組み合わせであったが、近年は深層学習の活用が始まっている。Microsoft の「りんな」等には深層学習による発話選択が組み込まれている。また、そもそもユーザーへの発話をあらかじめ準備するのではなく、発話内容も深層学習・強化学習によって獲得する枠組みの研究も始まっている。しかし、発話内容のコントロールが難しく、場合によっては不適切な対話を行ってしまう問題があり [6]、新たな枠組みが待たれる。また、応用分野（A）で扱われているような膨大な量の情報・知識を対話に取り込むことも、期待はされているが、いまだ研究は初歩の段階にある。

［海外動向］

技術政策面を概観すると、現在は自然言語処理単独の技術政策としてではなく、その上位テーマである人工知能（AI）分野の技術政策に集約されているといえる。米国の政策は自然言語処理分野でも際立っている。もともとアメリカ国立標準技術研究所（NIST）による情報検索・質問応答技術、国防高等研究計画局（DARPA）による情報抽出技術や文章理解、高等研究開発局（ARDA）による質問応答技術の研究開発等、多くのプロジェクト（評価型ワークショップを含む）[1] が政府予算によって推進されてきた。2014 年からは DARPA の Big Mechanism プロジェクトが 4500 万ドル規模の予算、3 年半の予定で推進されており、技術論文から因果関係の情報を抽出し、そのモデル化を行う技術が研究されている。がんにかかわる因果関係モデルを論文から獲得するという研究であるが、このような長期的な研究テーマに先行投資する政策は高く評価できる。

欧州では EU の第 7 次フレームワークプログラム（FP7）が 2007 年から 2013 年まで実施され、言語解析ツールの相互運用や機械翻訳のプロジェクトが work programme として推進されていた。その後継となる Horizon 2020 プロジェクト（2014 ～ 2020）でも機械翻訳、対話、テキスト分析、文生成といったテーマが選ばれている。中国は 2016 年 5 月に AI に係る 3 年行動実施方案を発表し、2018 年までに AI 分野で 1000 億元レベルの市場形成を目標としている。自然言語処理はビッグデータ処理のための基盤技術として位置付けられている。

ビジネス面を見ると、自然言語処理は企業のサービス差異化のための中核技術として注目されるようになっており、2015 年はグローバルなベンチャー・キャピタル投資総額 550 億ドルの約 5％（27.5 億ドル）が AI 分野への投資であったと見積もられている。Google、Amazon、Microsoft、Facebook 等の巨大企業は各社が年間 10 億ドル以上の研

究開発投資を行っており、これらの各社が有するインターネット上のテキスト情報コンテンツをもとに新たなサービスを創出する意欲は非常に旺盛である。そのような中でも対話技術をサービス構築の基盤技術としてベンチャー企業や開発者のエコシステムを形成する動きが顕著である。IBM による Cognea 社の買収と、その後の Watson Developer Cloud 上での Dialog API の公開、Microsoft による Conversations as a Platform（CaaP）の構想と Cognitive Services の公開、ほかにも Amazon Alexa（Amazon Echo にも搭載）、Apple Siri、Google Now、Facebook M 等の対話アシスタントやチャットボットが今後の大きなアプリケーションあるいは多様なアプリケーションに共通する汎用ユーザーインターフェースとして急速に成長している。

### (3) 注目動向

ビジネス・応用面や技術政策面の動向は（2）で言及しているので、ここでは、特に技術面から最近の注目動向を三つあげる。

［深層学習による性能向上］

既に書いた通り、自然言語処理は 1990 年代まではルールベースの方式が主流だったが、それ以降、機械学習技術の適用が進んでいる。適用される機械学習技術は、ナイーブベイズに始まり、2010 年頃には SVM（Support Vector Machine）が主流となったが、ここ数年は深層学習の適用が盛んである [7),8)]。ただし、画像認識・音声認識の分野では、深層学習の適用によって華々しい性能向上が見られたのに対して、自然言語処理の分野では簡単にそうはならなかった。

しかし、まさにこの 1 年くらいのタイミングで、自然言語処理の分野でも、深層学習を活用した方式が従来方式の性能を上回り始めた。形態素解析・品詞タギング、パージング、機械翻訳等のいずれにおいても、従来方式の性能を上回り始めた。ただし、画像認識・音声認識等の分野と比較すると、深層学習による性能向上のゲインはまだ小さい。また、その性能向上の要因は、何か技術的な飛躍があったというよりは、自然言語処理分野で深層学習の使いこなしが進んだという面が強い。

［意味の分散表現］

深層学習は end-to-end の結果を出すのが基本であり、自然言語処理において重要な問題と考えられてきた意味の計算モデルはブラックボックスになってしまう。この問題に対して、最近一躍注目されているのが、word2vec に代表される意味の分散表現 [9)-11)] である。分散表現は、文脈類似性に基づいて、単語の意味ベクトルを学習したものである。従来よく使われていた bag-of-words とは異なり、分散表現は固定長ベクトル（数百次元程度）の形をとり、ベクトル演算によって単語や文の意味の合成・分解が可能である。

この分散表現の導入によって自然言語処理が再構成されつつある。つまり、自然言語処理のプロセスを分解して、単語や意味の要素表現を分散表現に置き換える動きが進んでいる。分散表現は、ニューラルネットへの入力形式としても適しており、自然言語の解析・変換に用いる深層学習に組み込むのも容易である。シソーラスや辞書にも分散表現の学習

が取り入れられ、意味選択や QA の問題等も分散表現の演算として扱われるようになりつつある。

［大規模コーパス・大規模知識ベース］

ビッグデータの時代と言われるように、自然言語処理で用いるコーパスや知識ベースも大規模化が進んできたが、そのための管理技術も発展している。

まず、その構築手法としてクラウドソーシングが活用されるようになった。クラウドソーシングについては第 3.3.5 節「ビッグデータ活用促進技術」で詳しく触れるが、さまざまな分野・タスクで使われるようになってきた。一つ一つのタスクで見ると非常に安価ではあるが、作業者による品質のばらつきは避けられない問題に対しても、複数の作業者の結果を統計処理するとか、作業者を評価・選別するとか、管理手法が整備されてきている。また、さまざまな文例の中から文脈を理解しないと判定できないものを見つける等、クラウドソーシングの適用が特に有効なケースもわかってきた。

次に、QA や推論等を支える知識ベースの大規模化と有効活用も進んでいるが、その知識ベースの効率的構築技術や品質管理技術も発展している。知識の関係を補完する技術や、知識の足りない部分を Web・テキストから拡充する技術等も整備されてきている。

### (4) 科学技術的課題

・現在の自然言語処理は機械学習ベースが主流となっているため、学習データや情報源として利用可能な言語資源が存在しない分野では、研究開発や社会実装の大きな進展が望めない。このため、豊富なデータが利用可能な分野からの転移学習や、限られた学習データからの汎化能力、ルールベース的手法との併用等による分野適合手法のより一層の推進が急務である。この点は、理化学研究所革新知能統合研究センター（AIP）においても重要な研究テーマとして言及されている。

・従来のルールベースの自然言語処理技術と深層学習技術との統合も重要課題である。自然言語処理のタスク全体を end-to-end の深層学習で解くよりも、入力の一部の素性抽出に従来手法を適用する等、適切な組み合わせによってさらなる精度向上が期待できる。最近の研究では、要約タスクにおいて単純な単語の分散表現のみでなく品詞・固有表現のクラス名等の導入が精度向上に寄与したという報告[12]、読解タスクにおいて文脈共起・係り受け等の情報を用いた分類手法とのアンサンブルによって性能改善したという報告[13]、形態素解析エンジン JUMAN が深層学習採用（インターフェースは従来のまま）により構文解析誤りを約 40% 減少できたという報告等がある。また、深層学習による画像へのキャプション付与のように、従来の自然言語処理手法ではほとんど対応できていなかった問題の解決の糸口が見つかりつつあることも注目に値する。

・Web や SNS の情報を扱う上で常に問題となるのは情報の信ぴょう性である。東日本大震災ではインターネット上に大量のデマが出回った。これに関連し、信頼性の低い情報を検出する手法に関する研究も既に始まっているが[14]、いまだ決め手となる手法等は出現しておらず、今後さらなる展開が待たれる。

・知識の活用においては推論技術も重要である。テキストに書かれた情報を抽出するだけ

でなく、テキストに書かれていない知識を推論によって求める技術への取り組みも始まっている。生命科学において有力な仮説を提示して研究を加速するケース[15]や、Web 上の情報から導かれた仮説が著名な科学ジャーナルの内容を先取りするケース[16]等が報告されている。

## (5) 政策的課題

・深層学習で高い性能を得るためには大量の学習データが必要である。日本として、どのような種類の学習データに重点を置くか、大量テキストデータの収集・整備・共有の仕組みも含めて、戦略的な取り組みが必要である。その収集から学習の実行まで、高い効率・生産性で回す深層学習ファクトリーのような姿を目指した取り組みも必要である。そのためには、専門性の高い人材の育成・確保や、大規模な計算機環境の構築・整備も考える必要がある。また、フェアユースの導入も含めた著作権法等の法制度の整備も急務であるが、これに関しては第 3.3.7 節「ビッグデータに関わる制度設計」で論じる。

・自然言語処理技術に特有の課題として、言語別に技術が細分化されやすいという問題がある。ネット系巨大 IT 企業では、そのソフトウエア・サービス提供のプラットフォームを公開し、グローバルビジネスのための開発者コミュニティー形成や基盤構築を戦略的に推進している。そこでは言語独立・共通的な技術、英語および多言語化しやすい技術が優先されやすい。これを意識し、日本としては、深層学習等の最先端・共通的技術分野への投資は AI 全般の研究開発活動の一環として重視しつつ、英語主導の自然言語処理研究ではフォーカスされない日本語の高度な処理や言語資源の充実を図ることで、国内の先進的事例と将来研究に不可欠な学習データを蓄積し、広く利用可能にしていくことが重要である。

## (6) キーワード

自然言語処理、人工知能、機械学習、深層学習、分散意味表現、質問応答、機械翻訳、対話、言語理解、情報分析、情報抽出、知識獲得

## (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 | ○ | → | 言語処理学会を中心に基礎研究に取り組まれ、ACL・EMNLP 等の国際的トップカンファレンスでも一定数の発表がなされている。言語資源や研究利用可能なコンテンツの整備が不足しているとともに、自然言語処理単独の基礎研究のための政府プロジェクトがなく、AI 全般の研究振興のなかで相対的に埋没する危険がある。 |
| | 応用研究・開発 | ○ | ↗ | NICT が音声翻訳システム VoiceTra、大規模情報分析システム WISDOM X や災害関連情報分析システム DISAANA､D-SUMM を開発･公開（一部は商用化）。NEC、NTT ドコモ、日本 IBM、Softbank、トヨタ等の民間企業でも、含意認識技術、Watson 技術（日本語版）の他、対話系の「しゃべってコンシェル」、Pepper、Kirobo mini 等の実用化が進んでいる。 |

| | | | | |
|---|---|---|---|---|
| 米国 | 基礎研究 | ◎ | → | スタンフォード大学、CMU、MIT、ペンシルバニア大学、ワシントン大学等の有力大学のみならず、Google 等の有力企業はそれぞれ強力な研究チームを持っており、基礎研究を遂行している。ACL、EMNLP 等の有力な国際会議等での発表の多くは米国発である。 |
| | 応用研究・開発 | ◎ | ↗ | 上述した基礎研究が大学教員の移籍等によって、比較的短期間に Google、Microsoft、IBM、Facebook、Amazon 等の応用研究・開発に回るサイクルが確立している。Amazon の 125 億ドルを筆頭に Google、Microsoft、Apple 等 1 億ドル規模の研究開発費を投入している企業が多数存在する（2015 年時点）。自然言語処理を含む AI 全般において極めて広範囲の実用化プロジェクトが進行している [17]。 |
| 欧州 | 基礎研究 | ○ | → | EU の第 7 次フレームワークプログラム（FP7）が 2013 年で終了、現在は後継プロジェクトとして Horizon 2020（2014 ～ 2020）が約 800 億ユーロの予算規模で進行中。多国語を公式言語として使用する環境のため、機械翻訳への投資が伝統的に大きいが、突出した研究は少ない。 |
| | 応用研究・開発 | ○ | → | グローバル企業の研究所が存在し、一定のアクテビティーはあるが、米国主導による産業化の側面が強い。一方で欧州言語間の機械翻訳の精度は、人間による翻訳プロセスにおける「下訳」として使えるレベルにある [18]。 |
| 中国 | 基礎研究 | ○ | ↗ | 北京大学・清華大学等の有力大学や Microsoft Research Asia、Baidu 等の民間企業の研究所を中心に基礎研究が進められている。ACL、WWW 等のトップコンファレンスでも中国発の論文が多数発表されている。2016 年 5 月に国家発展改革委員会等が合同でインターネット +AI の 3 カ年行動実施方案を発表し、2018 年までに 1000 億元レベルの市場規模形成を目標とすることを示した。自然言語処理はビッグデータ、マルチメディア情報処理や脳型コンピューターと並んでコアの技術として認定されている。 |
| | 応用研究・開発 | ○ | ↗ | Microsoft のチャットボット「シャオアイス」、検索エンジン最大手である Baidu、チャットサービスのテンセント、EC 分野のアリババ等、盛んな研究開発投資と、サービスを通して得られる巨大データに背景にした AI 分野全般の実用化能力は脅威である。 |
| 韓国 | 基礎研究 | △ | → | KAIST、ETRI、KISTI 等の有力大学、国研を中心に基礎研究が進められている。ただ、ACL 等のトップコンファレンスでの韓国発の発表は減少している。 |
| | 応用研究・開発 | ○ | ↗ | 政府が Naver、サムソン、SK Telecom、Hyundai 等と Artificial Intelligence Research Institute を設立予定。Naver 等によるチャットボットや機械翻訳のサービスの開始が相次ぐ。 |

（注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
（注２） 現状　※わが国の現状を基準にした評価ではなく、CRDS の調査・見解による評価である。
◎　特に顕著な活動・成果が見えている、　○　顕著な活動・成果が見えている
△　顕著な活動・成果が見えていない、　×　活動・成果がほとんど見えていない
（注３） トレンド
↗：上昇傾向、　→：現状維持、　↘：下降傾向



## (8) 参考文献

1）NIST, The Retrieval Group, http://www-nlpir.nist.gov/
2）IBM Corp., This is Watson, Special Issue of IBM Journal of Research and Development, Issue 3-4, May-June 2012
3）Xin Luna Dong, et al., Knowledge Vault: A Web-Scale Approach to Probabilistic Knowledge Fusion, Proceedings of KDD 2014, pp. 601-610.
4）Zhongjun He., Baidu Translate: Research and Products, Proceedings of ACL 2015, pp.61-62.
5）Yonghui Wu et al., Google's Neural Machine Translation System: Bridging the

Gap between Human and Machine Translation.
https://arxiv.org/abs/1609.08144
6）The Telegraph, Microsft ‘deeply sorry’ after AI becomes ‘Hitler-loving sex robot’, 26 March 2016,
http://www.telegraph.co.uk/technology/2016/03/26/microsoft-deeply-sorry-after-ai-becomes-hitler-loving-sex-robot/
7）Manning, C. D., Computational linguistics and deep learning. Computational Linguistics, 41（4）, 2015.
8）Luong, M.-T., Pham, H., Manning, C. D., Effective Approaches to Attention-based Neural Machine Translation, Proceedings of EMNLP 2015, pp. 1412-1421.
9）岡崎直観 , 言語処理における分散表現学習のフロンティア , 人工知能 31（2）, pp. 189-201,2016.
10）Nickel, M., Murphy, K., Tresp, V., and Gabrilovich, E., A Review of Relational Machine Learning for Knowledge Graphs, Proceedings of IEEE 104（1）, pp.11-33, 2016.
11）Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J., Distributed Representations of Words and Phrases and Their Compositionality, Proceedings of NIPS 2013, pp. 3111-3119.
12）Nallapati, R. et al., Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond, Proceedings of CoNLL 2016.
13）Chen, D. et al, A Thorough Examination of the CNN/Daily Mail Reading Comprehension Task, Proceedings of ACL 2016.
14）Carlos Castillo, et al., Information credibility on twitter, Proceedings of WWW 2011, pp. 675-684.
15）W. Scott Spangler, et al., Automated hypothesis generation based on mining scientific literature, Proceedings of KDD 2014, pp.1877-1886.
16）Chikara Hashimoto, et al., Toward Future Scenario Generation: Extracting Event Causality Exploiting Semantic Relation, Context, and Association Features, Proceedings of ACL 2014, pp. 987-997.
17）JETRO, 米国における人工知能に関する取り組みの現状 ,
http://www.ipa.go.jp/files/000044196.pdf
18）MT@Work 2015, Streaming Service of the European Commission,
https://webcast.ec.europa.eu/machine-translation

### 3.3.5　ビッグデータ活用促進技術

#### （1）研究開発領域の簡潔な説明

ここまでの節では、ビッグデータの内容を深く解析する技術、および、それを高速に処理する技術を中心に取り上げてきたが、本節では、それら以外に、ビッグデータをさまざまな課題解決（価値創造）に活用していく上で共通的に必要になる技術を「ビッグデータ活用促進技術」と呼び、特にクラウドソーシングとセキュアなビッグデータ処理技術を取り上げる。

クラウドソーシング（Crowdsourcing）[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。人工知能（AI）技術だけでは解決困難な課題に対しては、人間の力を一部組み合わせ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[2),3)] が有効である。クラウドソーシングは、このヒューマン・コンピュテーションのプラットフォームとして注目されており、ビッグデータ領域にもその活用が広がっている。

また、ビッグデータとして集まる情報には、個人の生活や行動にまつわる情報等、その利用に十分な配慮が必要なものが含まれる。したがって、ビッグデータの活用による価値創造を促進すると同時に、プライバシーを保護し、情報を保護するセキュアなビッグデータ処理が求められる。この保護と活用のトレードオフ問題に対して、法制度を含めた制度設計面からのアプローチは第3.3.7節「ビッグデータに関わる制度設計」で述べるものとし、以下では技術的なアプローチにフォーカスする。

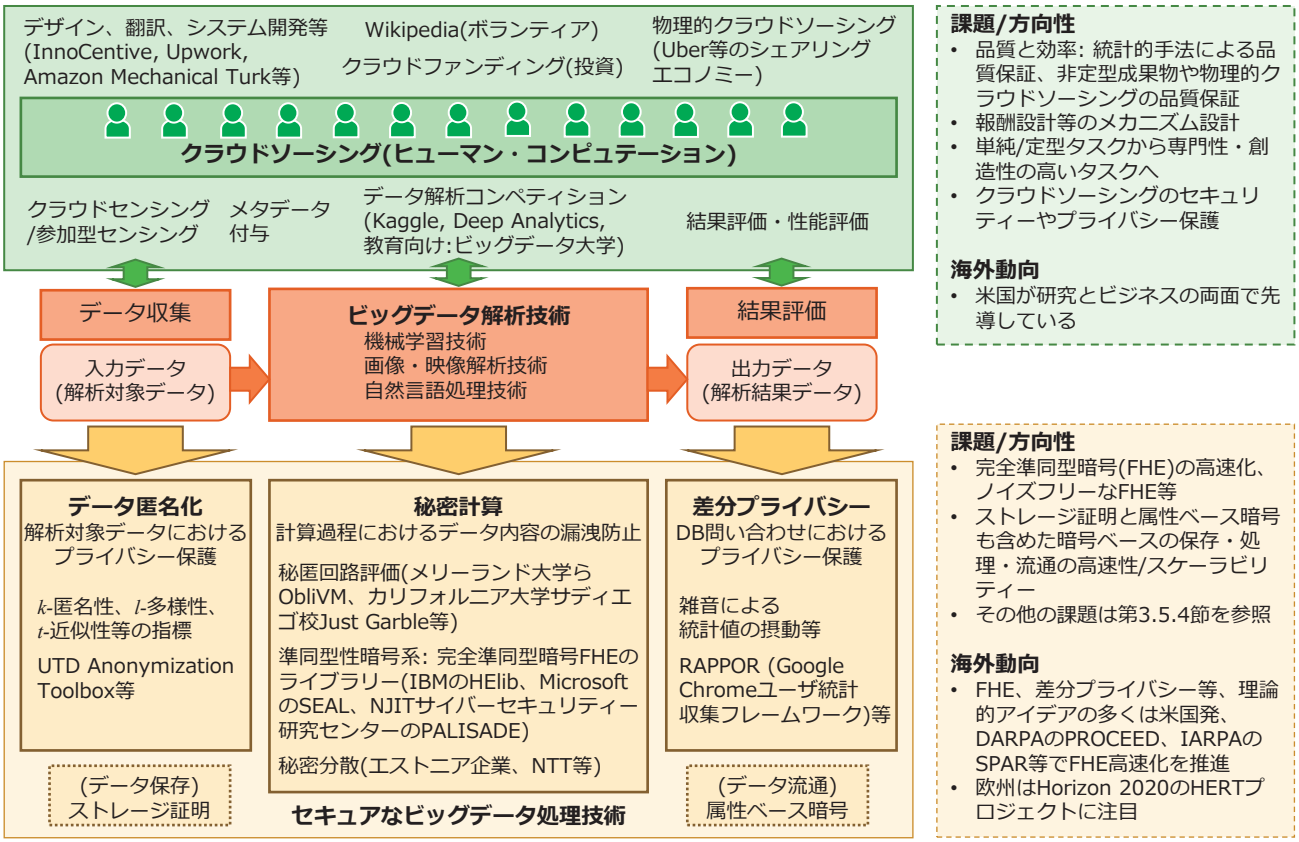


図 3-3-6　領域俯瞰：ビッグデータ活用促進技術

## (2) 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］

クラウドソーシングは、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、場所や時間にとらわれない新しい働き方として注目され、さまざまな分野・目的に急速に広がっている。その実施形態は、報酬の有無や雇用方式の違い等からさまざまなバリエーションがある。InnoCentive、Upwork、Amazon Mechanical Turk 等、オンライン労働力を募っての課題解決は、デザイン、翻訳、システム開発等に広く活用されている。ボランティアベースのものもあり、米国航空宇宙局（NASA）や米国国防高等研究計画局（DARPA）等の公的機関での活用のほか、国内では震災時の安否情報や医療情報の構造化への活用事例がある。不特定多数の人間が編集に関わる Wikipedia や、プロジェクトへの投資を一般に募るクラウドファンディングも、クラウドソーシングの一種とみなすことができる。情報処理技術分野ではコンピューターだけで完全な解決が困難な問題（例えば自然言語の理解・翻訳、画像理解、データベースにおける SQL 評価や参照解決等）に有効である。ヒューマンインターフェースや検索等の性能評価では、従来は少数の人間による小規模評価が主流だったが、クラウドソーシングを用いることで大規模評価が可能になった。

ビッグデータ領域においてもクラウドソーシングは極めて有用である。まず、効果的で効率的なビッグデータ収集のため、人間の知覚能力や判断を活用したクラウドセンシングや参加型センシングと呼ばれるアプローチ [4] がとられている。一方、ビッグデータの解析に際しては、機械学習アルゴリズムに与える正解データ等、メタデータの準備が大きなボトルネックとなる。この作業には通常、多大な人手が必要で、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になった。ただし、クラウドソーシングによって収集されたデータやメタデータは質のばらつきが大きい。そのため、これをいかに抑えて有効な解析やモデル構築を行うかについて、盛んに研究が進められている [5]。

さらに、ビッグデータ解析そのものにおいても、クラウドソーシングは極めて有効である。その典型例は、共通のデータを用いて参加者がその予測精度を競うデータ解析コンペティションである。解析したいデータを提供するスポンサーから勝者に賞金が支払われる賞金付き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。そのプラットフォームとして最もよく知られたものが Kaggle（kaggle.com）であり、世界中から 10 万人以上が参加している。国内でも同種のプラットフォームとして Deep Analytics（deepanalytics.jp）が運用されている。コンペティションの形式をとることによって、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。現在のプラットフォームはほぼ予測モデリングに特化したものであるが、探索的なデータ解析を対象とした試みも出てきており、今後の発展が見込まれる。

データサイエンティストの育成は喫緊の課題だと言われ、近年、国内外のさまざまな教育機関がそのための教育課程の提供を開始している。データ解析コンペティションは、実際のデータ解析を通じてデータ解析技術を身に付けられる教育的効果もあると認識されつつあり、Kaggle でもコンペティションを大学の講義等で利用できるサービスが提供され

ている。国内でも同様の試みとしてビッグデータ大学（universityofbigdata.net）がある。

［セキュアなビッグデータ処理技術］

ビッグデータ処理において、プライバシー保護、情報保護を確保する上で、技術的な面から押さえるべきポイントがいくつかある。解析対象データにおけるプライバシー保護、データベース問い合わせにおけるプライバシー保護、計算過程におけるデータ内容の漏えい防止である。以下、そのおのおのに対する技術的対策として、①データ匿名化、②差分プライバシー、③秘密計算（秘匿計算ともいう）を簡単に紹介する。なお、これらの詳細は、セキュリティー区分の第 3.5.4 節「プライバシー情報の保護と利活用」にも説明がある。

①データ匿名化

匿名化の処理は、ISO/TS 25237（Health informatics – Pseudonymisation）の中で「データとデータ主体（所有者）との間の相関を取り除くプロセス」と定義されている。個人データの収集において、姓名等の識別子を削除しただけでは、上記の相関は完全には取り除けず、他の属性情報・履歴情報を束ねて見ることで個人が特定され得るリスクがある。プライバシー保護のため、このような個人特定のリスクを定式化し、低減するための考え方として *k*- 匿名性 [6] がよく知られている。具体的には、表形式データについて、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 k 個以上存在している状態が、*k*- 匿名性が成立した状態である。データの正確性は犠牲になるが、個人データを改変することで、*k*- 匿名性を成立させ、個人特定を困難にする。その後、*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等が進められている（*l*- 多様性、*t*- 近似性等）。

日本では 2015 年に個人情報保護法が改正され、匿名加工情報の取り扱いスキームが新設された。特定の個人を識別できないように個人情報を加工したものを匿名加工情報と呼び、その加工方法を定め、事業者による公表等、取り扱いの規律を設けている。*k*- 匿名性をはじめとする匿名化技術は実務的に重要な技術になるものと考えられる。

②差分プライバシー

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。母集団数が大きい場合には正しいが、「X 町 Y 町目在住の 60 代独身女性の平均所得」（例えば母集団数 5 人）のように母集団数が少ない場合は、統計値の開示が個別の個人所得の推測をある程度可能にしてしまうリスクがある。

差分プライバシーは、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 [7]。差分プライバシーでは以下のような状況を想定する。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、データ収集者に対して、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）を発し、その応答値を得る。このとき「その応答値から、対象データベースに、ある個人が含まれているか否かを判定できる」ことをリスクと捉え、

このリスクを低減するために、統計値を雑音によって摂動させた上で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。

差分プライバシーは、その安全性が攻撃者の背景知識によらないため、*k*- 匿名性等に比べて、より強力な安全性を保証できる。Google Chrome のユーザー統計収集フレームワーク RAPPOR（randomized aggregatable privacy-preserving ordinal response）[8] 等で実用化事例がある。

③秘密計算

秘密計算（Secure Multi-party Computation）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データを所持しているときに、その秘密データを互いに他者に開示することなく、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。その実現方式は、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、という 3 種類が知られている。

秘匿回路評価は紛失通信を用いた秘密計算法である [9]。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、汎用性が高い。しかし、大規模な入力を引数に取る関数の評価に膨大な時間がかかり、実用性が低かった。近年は高速実装の研究が進み、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。メリーランド大学らのグループによる Java 系コードで記述された計算を秘匿回路に変換して簡便に秘匿回路評価が可能な開発環境 ObliVM[10]、カリフォルニア大学サンディエゴ校らのグループで開発された秘匿回路開発フレームワーク Just Garble[11]（プログラムを論理回路として実装する必要がある）等がある。

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、これらを暗号化したまま復号することなく（秘密鍵の知識なしに）、二項演算を行ったとき、もとの整数に対する加法・乗法等の演算結果の暗号値を得ることができるという性質を持った暗号系である。加法または乗法のいずれ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 年に加法と乗法の両方で同時に準同型性を満たす暗号「完全準同型暗号」（Fully Homomorphic Encryption: FHE）が発表 [12] されたのがブレイクスルーとなり、研究が進展した。FHE は計算時間や空間計算量が膨大であるとされていたが、さまざまな高速化テクニックが盛んに研究され、ライブラリーも公開されるようになった。FHE を実現するライブラリーには、IBM の HElib、マイクロソフトの SEAL（Simple Encrypted Arithmetic Library）、DARPA SAFEWARE プログラム（2015 年～）で NJIT サイバーセキュリティー研究センターが開発中の PALISADE（2017 年前半に正式リリース）がある。FHE は秘匿回路評価と比較すると、算術演算をベースとするため、数値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おり、比較的単純な統計計算であれば実用的な時間で動作する例が報告されつつある。

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。秘密分散では、秘密情報を複数の情報の断片（シェアと呼ぶ）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。これを用いた秘密計算は、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、これらの秘密情報を入力としたあるクラスの計算はランダムシェア同士の演算で実現し得ることを利用して実現する。実現可能な関数のクラスは限定されるが、暗号演算や暗号プロトコルの実行を含まないので一般に計算効率がよい。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見られる[13)]。NTT は 1 万件のソートを 10 秒、100 万件の乗算を約 1 秒で処理可能な高速なライブラリーを開発し、医療統計処理に関する実証実験を実施した[14)]。

### (3) 注目動向

[クラウドソーシング]

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、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された。それ以降、ビッグデータ領域でのクラウドソーシングの活用が広がり、標準的な選択肢の一つと考えられるようになってきた。研究領域としても、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 HCOMP が 2013 年に発足した。これは毎年開催されており、本研究分野の中心的コミュニティーとしての役割を果たしている。

個別では、Microsoft Research は目立っている。企業の研究所でこの分野に力を入れるところが増えているが、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。また、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 プロジェクト（duolingo.com）を開始し、その今後の動向が注目されている。

さらには、近年、Uber に代表されるシェアリングエコノミーが注目されているが、これもある種の物理的クラウドソーシングと考えることができる。今後、物理世界も視野に入れたクラウドソーシングは一層広まっていくと思われる。

[セキュアなビッグデータ処理技術]

米国では、2012 年のビッグデータ研究開発イニシアチブ設立以降、大統領科学技術諮問委員会（PCAST）による 2014 年 5 月の「ビッグデータとプライバシー」[15)]、米国科学技術政策局ネットワークおよび情報技術研究開発プログラム（NITRD）による 2016 年 2 月の「連邦サイバーセキュリティー研究戦略」[16)] や 2016 年 5 月の「ビッグデータ研究開発戦略」[17)]、また、2014 年 9 月の「NSF Workshop on Big Data Security and Privacy」[18)] 等、本分野の研究開発に対して活発な検討が行われている。秘密計算アプローチの重要性も認識されており、DARPA やインテリジェンス高等研究計画活動（IARPA）等で、完全準同型暗号（FHE:Fully Homomorphic Encryption）の高速化研究も推進されている。

米国 DARPA 予算で実施されている PROCEED（Programming Computation on Encrypted Data、2011 ～ 2016、予算 20MUS$）プロジェクトは、「完全準同型暗号を用いた秘密計算が実用化できないのは、通常計算に比較して 10 桁遅くなるためであること」を問題として挙げ、これを 10 万倍高速化することを目指したものである。IBM、MIT、スタンフォード大、UCLA、テキサス大、バージニア大等により実施されており、FHE を実行する上で欠かせないブートストラップの高速化に加え、属性ベース暗号、量子暗号等の理論研究を含めこれまでに 130 件の論文が発表されている。

米国 IARPA 予算により実施されている SPAR（Security and Privacy Assurance Research、2011 ～終了年不明）プロジェクトは、FHE を用いた効率的なデータベース検索のプロトタイピング（最終的には通常検索の 8 倍の実行時間以内に抑える）を目指し、理論研究を含めた研究が実施されている。IBM、コロンビア大、MIT 等が参画しており MySQL に対する検索とほぼ同等（空間計算量は 7 倍）の性能を達成している。また関連する成果として、これまでに 62 件の論文がある。

企業では、FHE を提案し GitHub 上で HElib を提供している IBM ワトソン研究所の他、フランスに拠点を置く CryptoExperts 社の動きが注目される。Google Scholar での検索によれば、CryptoExperts 社は、これまでに 58 件の準同型暗号に関連する論文を発表すると共に、GitHub において、SFHE（演算回数を一定数に制限し高速化を図った FHE）のライブラリーを公開している。また、CryptoExperts 社が中心となり HERT プロジェクト（FHE を用いた革新的暗号技術の研究開発が目的、2015 ～ 2017、heat-project.eu）が、Horizon2020 の枠組みの中で実施されている（予算 :41MEUR）。

一方、日本では、匿名加工情報のスキームを含む個人情報保護法の改正が 2015 年に実施されたものの、秘密計算・秘匿化に特化した記述はない。秘匿化に関するプロジェクトとしては、JST 戦略的創造研究推進事業 CREST の「ビッグデータ統合利活用のための次世代基盤技術の創出・体系化」（2013 ～）において、三つの関連プロジェクト「自己情報コントロール機構を持つプライバシ保護データ収集・解析基盤の構築と個別化医療・ゲノム疫学への展開」（佐久間淳：筑波大）、「ビッグデータ統合利活用促進のためのセキュリティ基盤技術の体系化」（宮地充子：北陸先端大）、「ビッグデータ統合利用のためのセキュアなコンテンツ共有・流通基盤の構築」（山名早人：早稲田大）が推進されている。また、情報通信研究機構（NICT）において、NICT が開発した準同型暗号 SPHERE をベースにロジスティック回帰分析を暗号化したまま行う技術開発が行われている [19]。なお、秘密計算に関わる日本からの特許出願数は米国・欧州に比較して多い。2011 年までの出願人上位には、三菱電機、日本電気、日本電信電話等がある [20]。

### （4）科学技術的課題

［クラウドソーシング］

・アウトソーシング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。そのため、そ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は、クラウドソーシングにおける重要課題である [21]。成果物の品質は仕事を受けた人間の能力ややる気に

大きく依存するため、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が必要である。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の仕組み等が用いられる。また、文章（記事作成や翻訳）、デザイン、プログラム等の複雑な構造あるいは非定型の成果物を求める仕事が多くなっているが、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。さらに、前述のシェアリングエコノミーのような物理的クラウドソーシングにおける品質保証も今後の課題といえる。一方、より積極的に、ワーカーから情報を正しく引き出したり、タスクを適切に実施させたりするための報酬設計等のメカニズム設計も重要な課題である。

・クラウドソーシングは多数の人的資源に同時にアクセスできるが、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と比べてはるかに遅いため、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。
・従来は、実行にそれほど深い専門知識や創造性を要しないクラウドソーシングが主流であったが、今後、より専門性が高く、創造性を必要とする課題をどのようにクラウドソーシングで実現していくかは、クラウドソーシングの利用可能性を広げる意味で大きな課題である。
・クラウドソーシングによるデータ収集では、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で利用するケースや、医療や健康の向上を目的に利用するケース等では、セキュリティーとプライバシーの問題が実用化への大きな壁となる。また、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場合、（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ちゅうちょ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である。

［セキュアなビッグデータ処理技術］

・第 3.5.4 節「プライバシー情報の保護と利活用」でも同様の技術的課題を述べているので、ここでは（3）注目動向で取り上げた FHE の課題を述べる。
・FHE では、暗号文にノイズを含めることにより、暗号強度を高めている。しかし、このノイズがあるが故に複数回演算を繰り返すとノイズが巨大化し復号できなくなる。このため、ノイズを小さくするためのブートストラップを定期的に適用しなければならず高速化においてのボトルネックとなっている。しかし、2016 年 9 月に実施された ESORICS2016 国際会議では、ノイズフリーな FHE が提案されており、日進月歩で進む暗号理論研究を正確に把握した上での実用化に向けた高速化が必要不可欠と考える。
・FHE により暗号化されたデータに対する演算を実現するためには、計算途中で「値を参照できない」ことを前提としたアルゴリズムを考える必要がある。つまり、暗号化さ

れたデータに対して大小比較等に代表される条件分岐を適用できない。従来の多くのアルゴリズムが枝刈りによって計算量を削減するという基本原理の上に成り立っていることを考えると、実用化にあたっては従来のさまざまなアルゴリズムを見直す必要がある。

・FHE によるデータ処理だけでなく、データ保存のためのストレージ証明、データ流通のための属性ベース暗号についても、空間計算量、時間計算量の面で、ビッグデータにそのまま適用するのが困難な状況である。データを暗号化して保存・処理・流通を可能にする暗号体系について、理論とコンピュータアーキテクチャの両面から実用に向けてのプロトタイピングを進める必要があり、理論研究者と HPC 研究者との協業が求められる。

### (5) 政策的課題

・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、必ずしもこれまでの倫理的・社会的規範に沿っているとは言えない部分がある。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等の問題がある。技術的な課題解決と同時に制度的な整備も必要である。

・プライバシー保護・情報保護に関わる政策的課題は、第 3.5.4 節「プライバシー情報の保護と利活用」および第 3.3.7 節「ビッグデータに関わる制度設計」に記載した。

### (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、データ解析コンペティション、プライバシー保護、秘密計算、完全準同型暗号、準同型暗号

### (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 | ○ | ↗ | クラウドソーシングは、米国に後れをとる形だが、国内の研究コミュニティー（クラウドソーシング研究会等）が育ってきており、トップレベルの研究も生まれている。<br>暗号理論は企業・大学・国研とも高いレベルにある。秘密計算関連の特許数は日本がトップ[20]。プライバシー保護･システムセキュリティーの基礎研究はトップ国際会議で発表数が少ない。 |
| | 応用研究・開発 | ○ | ↗ | クラウドソーシングは、汎用・特化型それぞれにおいてさまざまなサービスが登場している。<br>秘密計算の応用技術は、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等が知られている。 |
| 米国 | 基礎研究 | ◎ | ↗ | クラウドソーシング、ヒューマン・コンピュテーションを先導しており、初期からクラウドソーシングの研究・活用が盛んで、さまざまな分野ら広がり、定着してきた。<br>暗号分野、データベース分野、データマイニング分野のどの研究領域でも、理論的アイデアの多くは米国の大学・企業の研究者から提案されている（差分プライバシー、完全準同型暗号等）。 |

| | | | | |
|---|---|---|---|---|
| 米国 | 応用研究・開発 | ◎ | → | クラウドソーシングは、大学と産業界の距離が近く、研究者自身が新しいビジネスを手掛ける等（von Ahn 等）、さまざまな先進的試みがある。<br>完全準同型暗号を提案した IBM ワトソン研究所をはじめ、ObliVM、JustGarble 等、秘密計算への取り組みでも先行している。 |
| 欧州 | 基礎研究 | ○ | → | クラウドソーシングでは、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。<br>プライバシー保護に関わる法制度や EU プロジェクトが積極的に進められており、それらを支える基礎研究には実績がある。 |
| | 応用研究・開発 | ○ | → | 基礎研究とほぼ同様。完全準同型暗号等を用いた革新的暗号技術の研究開発への取り組みとして、Horizon 2020 の HERT プロジェクトは注目（CryptExperts 社が中心）。 |
| 中国 | 基礎研究 | ○ | → | クラウドソーシング関連で Microsoft Research Asia 等から着実に研究成果が出ている以外、特に目立った活動は見られない。 |
| | 応用研究・開発 | △ | → | 第 13 次 5 カ年計画（2016 ～ 2020）でビッグデータ産業促進がうたわれている。2016 年国家自然科学基金の重点プログラムの一つとして暗号化システム [13] があがっているものの詳細不明。 |
| 韓国 | 基礎研究 | △ | → | 暗号分野で基礎研究・標準化活動がある以外（例えば 2012 年以降、準同型暗号を用いたプロジェクトが 9 件実施 [16] されている）、特に目立った活動は見られない。 |
| | 応用研究・開発 | △ | → | ビッグデータ産業発展戦略（2013 年）、ビッグデータ個人情報保護ガイドライン（2014 年）[14] 等の整備が進んでいるものの、技術開発では特に目立った活動は見られない。 |

（注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
（注２） 現状　※わが国の現状を基準にした評価ではなく、CRDS の調査・見解による評価である。
◎　特に顕著な活動・成果が見えている、　○　顕著な活動・成果が見えている
△　顕著な活動・成果が見えていない、　×　活動・成果がほとんど見えていない
（注３） トレンド
↗：上昇傾向、　→：現状維持、　↘：下降傾向

## (8) 参考文献

1）Howe, J., The Rise of Crowdsourcing, Wired Magazine, 2006.
2）Law, E., Von Ahn, L., Human Computation, Morgan & Claypool Publishers, 2011.
3）鹿島久嗣 , 馬場雪乃 , 小山聡 ,「ヒューマンコンピュテーションとクラウドソーシング」, 講談社 ,2016.
4）Ganti, R. K., Ye, F., Lei, H., Mobile Crowdsensing: Current State and Future Challenges, IEEE Communications Magazine 49（11）, pp. 32-39, 2011
5）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）,pp. 381-388,2012.
6）Sweeney, L., *k*-Anonymity: A Model for Protecting Privacy, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10（5）, pp. 557-570, 2002.
7）Dwork, C., McSherry, F., Nissim, K., Smith, A. D., Calibrating Noise to Sensitivity in Private Data Analysis. TCC 2006, pp. 265-284.
8）Erlingsson, Ú., Pihur, V., Korolova, A., RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response, ACM Conference on Computer and Communications Security 2014, pp. 1054-1067.
9）Chi-Chih Yao, A., How to Generate and Exchange Secrets（Extended Abstract）.

FOCS 1986, pp. 162-167.
10）Liu, C., Wang, X. S., Nayak, K., Huang, Y., Shi, E., ObliVM: A Programming Framework for Secure Computation, IEEE Symposium on Security and Privacy 2015, pp. 359-376.
11）Bellare, M., Hoang, V. T., Keelveedhi, S., Rogaway, P., Efficient Garbling from a Fixed-Key Blockcipher. IEEE Symposium on Security and Privacy 2013, pp. 478-492.
12）Gentry, C., Fully Homomorphic Encryption Using Ideal Lattices, STOC 2009, pp. 169-178.
13）Bogdanov, D., Laur, S., Willemson, J., Sharemind: A Framework for Fast Privacy-Preserving Computations, ESORICS 2008, pp. 192-206.
14）NTT, http://www.ntt.co.jp/news2012/1202/120214a.html
15）The White House, PCAST Report on Big Data and Privacy: A Technological Perspective,
https://obamawhitehouse.archives.gov/the-press-office/2015/11/16/fact-sheet-pcast-report-big-data-and-privacy-technological-perspective
16）NITRD, Federal Cybersecurity Research and Development,
https://www.nitrd.gov/cybersecurity/publications/2016_Federal_Cybersecurity_Research_and_Development_Strategic_Plan.pdf
17）NITRD, The Federal Big Data Research and Development Strategic Plan,
https://www.nitrd.gov/PUBS/bigdatardstrategicplan.pdf
18）NSF, Workshop Report Big Data Security and Privacy,
http://csi.utdallas.edu/events/NSF/NSF-workhop-Big-Data-SP-Feb9-2015_FINAL.pdf
19）暗号化したままデータを分類できるビッグデータ向け解析技術を開発 ,
http://www.nict.go.jp/press/2016/01/14-1.html
20）平成 25 年度特許出願技術動向調査報告書 ビッグデータ分析技術 , 特許庁 ,
http://www.jpo.go.jp/shiryou/pdf/gidou-houkoku/25_bigdata.pdf
21）Kittur, A. et al., The Future of Crowd Work, CSCW 2013.

### 3.3.6　ビッグデータによる価値創造

#### （1）研究開発領域の簡潔な説明

世界はボーダーレス化し、さまざまなものが相互に影響し合い、社会システムは大規模化・複雑化・複合化している。そのため複雑化した社会問題の解決や社会システムの最適設計等は人間の手に負えないものになりつつある。これらの深刻化する問題に対する技術面からのアプローチとして期待されるのがビッグデータである。ビッグデータの収集・解析は、（もはや人間には困難になった）実世界・実社会の現象や活動の精緻でリアルタイムな把握と予測を可能にする。そして、これは上述の問題解決や最適設計等のための大きな手掛かりとなり、安心安全で生産性の高い社会に向けたさまざまな価値創造につながる。本節では、このようなビッグデータによる価値創造について、さまざまな分野での取り組みを概観する。

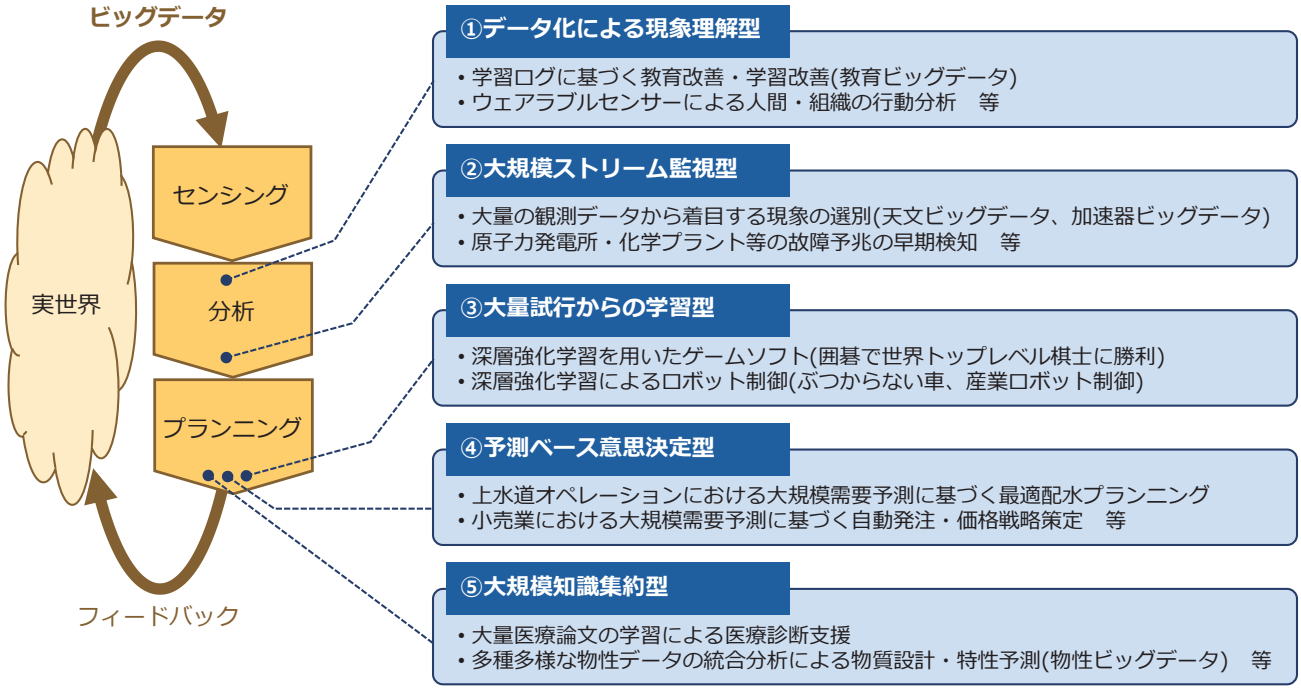


図 3-3-7　領域俯瞰：ビッグデータによる価値創造

#### （2）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」という言葉が使われ始めたのは 2010 年頃からであるが、既に社会・市場に浸透し、さまざまな分野・業種に応用が広がっている。当初は、Google 等のサーチエンジンにおける検索連動型広告や、Amazon 等の EC サイトにおける商品レコメンデーションのように、インターネット上のサービスに集まるビッグデータを売り上げ向上に活用するのが中心であった。しかし、現在は、実世界・実社会から集まるビッグデータを活用した社会課題解決へと広がってきており [1)]、その社会的価値はますます高まっている。

分野・業種ごとの事例をあげると際限がないので、ここでは技術的な視点からビッグデータ活用による価値創造 / 課題解決のタイプ分けを試みる。ビッグデータの活用フローは、実世界からデータを取得・収集するセンシング、そのデータの分析、分析結果に基づいてアクション（実世界へのフィードバック）を計画するプランニングから成る。以下、この

フローに沿って、図 3-3-7 に示すような五つのタイプがあることを説明し、各タイプの代表的事例も紹介する。

①データ化による現象理解型

従来は定量的な把握ができていなかった行動や事象について、ビッグデータを取得・収集する仕組みを作ることで、その精緻な状況把握や定量的な分析が可能になる。例えば、九州大学のラーニングアナリティクスセンターでは、みつば（M2B:Moodle/Mahara/BookLooper）システムから取得した学生の学習ログを分析することで、復習よりも予習が成績の高さに関連すること、成績上位者は複数の情報源をリンクさせて学習していること等を見いだし、教育・学習の改善に結びつけている[2)]。また、日立製作所では、HBD センサー（Human-oriented Big Data、旧称ビジネス顕微鏡）と呼ぶウエアラブルセンサーから取得した人間行動ログを分析することで、コールセンターや大規模店舗等のビジネス事例において、個人や組織の業績向上につながる（意外な）要因を発見している[3)]。なお、一つめの事例に関連した教育ビッグデータ分野の動向については、後述の（3）でも取り上げる。

②大規模ストリーム監視型

①は従来データを取得できていなかった現場に、新たにビッグデータを取得する仕組みを入れることで、価値を生み出した事例である。一方、この②は、もともとデータを取得する仕組みをもっているのだが、データ量が膨大になってしまい、従来の仕組みでは処理しきれなくなってしまった事例である。例えば、監視カメラが多数設置されるようになると、同時並行で多数の監視カメラ映像を見続けねばならなくなり、人間では見落としが避けられなくなる。画像認識技術を用いて異常事態や不審者・不審行動を検知することで、見落としを減らすことができ、大規模監視・広域監視が可能になる。

観測天文学や素粒子実験物理学の分野では、観測機器から膨大な量の観測データが得られるが、その膨大なデータ中に多種多様な現象が混在しており、研究・分析のターゲットとなる現象を選別することさえ容易なことではない。これに対して、大型加速器実験の観測データから機械学習技術を使ってヒッグス粒子の現象を特定する精度を競う「ヒッグス粒子機械学習チャレンジ」[4)] のような試みがある。さらに、観測天文学（天文ビッグデータ）の分野における取り組みについても、後述の（3）で取り上げる。

また、原子力発電所・化学プラント等で多数配置されたセンサーからのデータを監視し、故障・異常の予兆をできる限り早期に検知しようとする試み[5)] も②の事例である。故障・異常のケースを機械学習し、それに該当するケースを検知するのがアプローチの一つだが、故障・異常のケースは一般に発生が少なく学習しにくい。そのため、逆に正常時のモデルを学習し、それから外れるケースを故障・異常として検出するのがもう一つのアプローチである。

③大量試行からの学習型

①②はセンシング・分析までであるが、③④⑤は課題解決にさらに踏み込んでアクションのプランニングまで行う。コンピューターが課題を検出するだけでなく、解決策まで提

案するので、生み出される価値はより大きなものになる。③④⑤は分析からプランニングまでのアプローチに違いがある。その技術面も含む詳細は第3.3.2節「機械学習技術」のパートで説明している。

③はアクションを繰り返しながら、より良いプランニング方策を学習していくもので、主に深層強化学習技術が用いられている。2016 年に囲碁で世界トップレベル棋士に勝利した AlphaGo[6),7)] は大きな話題となった。また、Preferred Networks（PFN）が CES2016 でデモンストレーション（トヨタ・NTT と共同展示）した「ぶつからない車」[8)] もこの応用事例である。学習過程において、最初、車はうまく走れず、他の車等にぶつかりそうになりながら、トライを繰り返し、徐々にぶつからずに速く走る方策を学習していく。PFN はこの技術を適用した産業用ロボット制御にも取り組んでおり、2016 年の Amazon Picking Challenge の Pick タスク（棚からアイテムを取り出す）では 2 位（1 位と同スコア）に入った[8)]。

④予測ベース意思決定型

アクションのプランニングを実現するための、③とは異なるアプローチとして、オペレーションズ・リサーチ（OR）に機械学習技術による予測を組み合わせる解法がある。その代表的な事例として、NEC の上水道オペレーション向けの大規模需要予測に基づく最適配水プランニングや、小売業向けの大規模需要予測に基づく自動発注・価格戦略策定等があげられる。膨大な数の予測対象点（需要家・商品等）のそれぞれに対して、機械学習技術による需要予測を行い、全体最適なアクションプランを生成する。ここでの予測結果は、誤差（外れるリスク）や複雑な相互依存関係を持っているため、従来の OR 手法で扱うのが困難である（組み合わせ爆発が起きる、リスクを考慮できず精度が低下する等）。この問題に対して、予測型意思決定最適化[9)] という新しい手法が開発された。

また、予測の方法として、過去データから機械学習によって求めた帰納的なモデルを用いる方法以外に、扱う問題の種類によっては、物理モデルに基づくシミュレーションを用いる方法もある。

⑤大規模知識集約型

大量の生データ / 観測データから現象を捉えようとする④までのアプローチとは異なり、ビッグデータを知識として大量に蓄積・構造化し、それらを組み合わせた判断・推論からアクションを導くアプローチが、ここで述べる⑤である。第 3.3.4 節「自然言語処理技術」と関連が深い。

このアプローチの代表は IBM の質問応答・意思決定支援システム Watson[10)] である。Watson は米国の人気クイズ番組「Jeopardy!」で人間のクイズ王に勝利するというグランドチャレンジを成功させ、大きく注目された。このときの Watson は Wikipedia をはじめ、この番組で出題される問題に関わる類いのさまざまな情報源から大量の知識を集約し、それらを活用して回答を導出した。その後、情報源を問題に応じて収集・構築することで、さまざまな分野に適用を広げている。その一つ、大量医療論文の学習による医療診断支援では、医療機関での治療で回復が芳しくなかった患者について、Watson による病気タイプの判定が的中し、治療方法を変えたことが回復につながったことが報じられた[11)]。

医療論文の数は膨大、かつ、ますます増加しており、人間がそのすべてを読んで記憶することはもはや不可能になっていることから、この⑤のタイプの必要性が増している。同様に、災害時等にバースト的に流通する大量の情報を即時に理解し、対処することも人間の限られた能力では難しい。情報通信研究機構（NICT）が開発した対災害 SNS 情報分析システム DISAANA（disaana.jp）は、災害時のツイートを分析・集約することで、この問題解決を支援する。

また、物質系材料研究の分野では、多種多様な物性データの統合分析による物質設計・特性予測への取り組み（物性ビッグデータ）が進められており、後述の（3）で取り上げる。

### (3) 注目動向

ビッグデータは（2）でも述べたようにさまざまな産業応用が広がっている。激しい企業間競争の中でこれはいっそう加速・拡大されていくにちがいない。この状況において科学技術政策の視点から、むしろアカデミア分野への応用に注目し、ここでは、教育、天文、物性という 3 分野でのビッグデータ活用を取り上げる。

［教育ビッグデータ］

教育分野におけるビッグデータ活用は、オンライン教育、対面講義、在宅学習等のさまざまな教育・学習の場面で、多種多様なデータを適切に収集・分析することで、教育や学習の効果を高めることを狙っている。データの具体的な例としては、授業への出席や小テスト、教材のページめくり、学生の心拍情報等があげられる。

このようなアプローチが注目されるようになった背景には、①情報機器・情報通信技術の普及により、授業内外を問わず、日常的に PC や携帯端末を利用した学習が可能となり、学習履歴データの収集が容易となったこと、②ソフトウエア面でも e-Learning、e- ポートフォリオ、履修情報、シラバス等の学内の教育情報システム、MOOCs（Massive Open Online Courses）等の教育機関の枠組みを越えたオンライン教育システムの利用が拡大し、そこでの学習履歴データが大量に蓄積されるようになったこと、③環境の変化・多様化に即した教育の質の向上、きめ細かな教育・学習支援の必要性が高まってきたこと等があげられる。

このような教育ビッグデータ活用への関心の高まりから 2008 年から EDM（Educational Data Mining）、2010 年から LAK（Learning Analytics and Knowledge）という国際会議が開催されるようになった。それぞれは、機械学習・知識発見アルゴリズムの応用に重点を置いた学会 IEDMS（International Educational Data Mining Society）と、解析結果を教育現場に応用することに重点を置いた学会 SoLAR（Society for Learning Analytics Research）の設立に至った。日本においては、教育ビッグデータの分析に関する学会として、2015 年に学習分析学会が設立された。また、2016 年 2 月には、九州大学がラーニングアナリティクスセンターを設置し、学内の教育ビッグデータの分析と、それに基づく教育・学習のサポートに取り組んでいる [2)]。

米国では、LA（Learning Analytics）を含む学習環境に関するコンソーシアム Unizin が 2014 年に設立され、現在では参加大学数は 10 を超える。Unizin では参加機関に対し、

教材などのコンテンツからソフトウエアやプラットフォーム、さらには学習分析に至るまでをサービスとして提供している。産業界においては、近年 EdTech と呼ばれるセクターが形成され活発な動きを見せている。有名大学と組み MOOCs のプラットフォームを展開している Coursera、誰でも講義をアップロードして売買できるマーケットプレイスを提供する Udemy、教育向けリアルタイムコラボレーションシステムを開発する Kramer 等のスタートアップ企業が生まれている。米国政府の Small Business Innovation Research 助成金から、2015 年には 900 万ドル以上が教育分野に投資された [12]。さらに、大企業も参入し始め、Google は 2014 年に Google Classroom を、それを追う形で 2016 年に Microsoft や Amazon も同様のプラットフォームを立ち上げた。メール、カレンダー、ストレージ等を備えた Google Apps や Office 365 等の既存サービスをベースに、コミュニケーション、課題の提出・採点、行事のスケジュール管理等の機能を提供し、紙ベースだったこれまでの教室活動の置き換えを狙っている。また、Microsoft は Minecraft を買収する等、Game-based Learning にも力を入れている。

欧州では、LA に関するコミュニティーをベースにしたプロジェクト LACE（Learning Analytics Community Exchange）[13]や、OWL（Open World Learning）[14]、TeSLA（Adaptive Trust-based E-assessment System for Learning）[15] 等のプロジェクトが数多く発足している。欧州ではプライバシー保護に関わる法整備等が進んでいるが、教育データのプライバシー保護を提唱した論文が LAK 2016 で Best Paper Award を受賞し、注目された。

システムの相互運用性を高めるための標準化も進められている。IMS Global Learning Consortium による試験問題の交換フォーマット仕様 Question & Test Interoperability、LA のためのフレームワーク Caliper Analytics[16]。より汎用的な標準仕様 Experience API（xAPI）[17]、国際標準化のためのワーキンググループ ISO/IEC JTC 1/SC 36 等があげられる。

［天文ビッグデータ］

天文ビッグデータは、宇宙における諸現象（宇宙の誕生と進化、銀河形成、星・惑星形成、宇宙生命等）を、観測的手法および理論的手法によって解明する科学領域である。観測的手法では、望遠鏡や検出装置を用いて研究対象天体からの電磁波（電波～可視光～ガンマ線）、粒子（ニュートリノ、宇宙線粒子等）、重力波のデータを取得し、コンピューターで分析する。理論的研究では、スーパーコンピューター等を用いたシミュレーションも盛んである。

望遠鏡等の観測装置の性能やデータ転送バンド幅の向上は著しく、国内外で次世代観測装置の建設が複数進行している。取得予定データの規模も格段に大きくペタバイト（PB、1015 バイト）を優に超える。(a) 地上からの観測および (b) 衛星からの観測が取り組まれている。

（a）地上からの観測装置としては、国立天文台は、すばる望遠鏡の次世代広視野カメラ Hyper SuprimeCAM[18]、ALMA（Atacama Large Millimeter/ submillimeter Array、日米欧の共同）[19] からそれぞれ年間約 200TB（テラバイト、1012）のデータを生み出している。海外では、米国を中心に、可視光や赤外線領域のサーベイ型望遠鏡 LSST（Large Synoptic Survey Telescope）[20] や Pan-STARRS（Panoramic Survey Telescope & Rapid Response System）[21] の建設や運用が始まっている。他にも大規模可視光・赤外線望遠

鏡として、直径 30m の TMT（Thirty Meter Telescope、日本・米国・カナダ・インド・中国が参加）[22]、直径 40m クラスの E-ELT（European Extremely Large Telescope）[23] をはじめ、各地で計画が進行している。

特に LSST は、米国 NSF の decadal survey でもトップにランクされ、2020 年の稼働開始に向けて建設が始まっている。その生み出すデータ総量は 450PB と見込まれており、ペタフロップス級のスーパーコンピューターによって迅速に処理されて分散データベースに格納された後、即時公開されて天文学研究に利用される。プロジェクトリーダー Tony Tyson 氏は「LSST は望遠鏡ではなく、データ供給マシンである」と述べ、天文学における研究パラダイムの刷新（データ活用型天文学へ）をもくろんでいる。

さらに、電波領域では、欧州 LOFAR（Low Frequency Array）[24] と豪州を中心とした SKA（Square Kilometre Array、第 1 期分は 2023 年までに建設）[25] は電波望遠鏡の概念を大変革しつつある。これらは、地理的に分散した複数の電波望遠鏡からの信号を光ファイバーで伝送し、高性能スーパーコンピューターで実時間処理をすることで複数天体を同時観測するプロジェクトである。SKA は毎秒 1PB のデータを生産すると見込まれている。その得られた観測データから、検出天体に関するカタログが作成され、広く天文学研究に利用される予定である。現存の電波望遠鏡では、アンテナ・受信機等のハードウエア性能が望遠鏡の性能を決していたのに対し、LOFAR や SKA では信号処理技術を含むソフトウエアがプロジェクトの成功のカギとなっている。このため、例えば豪州では、西オーストラリア大学で SKA をターゲットとしたデータ管理やデータ処理技術開発のために世界から優秀な人材を集め、また、スウィンバーン大学では巨大データの可視化や機械学習を通じた知識発見に向けた研究開発を推進している。これらの大規模データを扱うプロジェクトでは、その予算のほぼ半分がデータ管理、データ処理等に関わるソフトウエア開発に充てられていることも注目に値する。

（b）衛星からの観測は、地球の大気に吸収されるガンマ線・エックス線・赤外線の一部は地上での観測が不可能または困難なため必要になる。国内では、宇宙航空研究開発機構宇宙科学研究所（JAXA/ISAS）が天文観測衛星を打ち上げ、研究者にデータを提供している。海外では、米国 NASA や欧州宇宙機関（ESA）が大きなプロジェクトを組んで天文観測衛星を打ち上げ、運用している。多くの場合、これらの主要宇宙機関に JAXA も含めた他国の宇宙機関（独 DLR、カナダ宇宙機関等）が国際協力する形をとっている。一般的に、衛星で取得されたデータの容量は、地上観測に比べて非常に少ない。これは地上へのデータ伝送帯域幅が限定されているためで、典型的には、一つの衛星プロジェクト全体で数百 GB 程度である。しかし、ESA が打ち上げた位置天文観測衛星である Gaia のデータは TB を超えるようになった。

天文データをアーカイブして公開するデータセンターとして代表的なものは、国内の国立天文台天文データセンターや宇宙航空研究開発機構宇宙科学研究所の C-SODA、仏国のストラスブール天文データセンター（CDS）、欧州南天天文台（ESO）、米国の赤外線処理解析センター（IPAC）、カナダ天文データセンター（CADC）等である。仏国の CDS は、世界中の天文データや関連情報（天文系雑誌、雑誌掲載のデータ等）を収集・公開することに特化している点が特徴である。国立天文台天文データセンターは、観測データのアーカイブ・公開だけでなく、データ解析用の計算サーバー群の提供やバーチャル天

文台の開発・運用も行っている。バーチャル天文台は、世界各地の天文台やデータセンターが保有・公開しているデータアーカイブを、標準プロトコル（国際バーチャル天文台連盟が定めたもの）を介して相互利用できるようにしたものである。このプロトコルに基づくアプリケーションも多数提供され、天文学者が天文ビッグデータを活用した研究が進めやすい環境が整えられている。

［物性ビッグデータ］

ビッグデータの活用の仕方として、大量の生の観測データから規則性を発見し、事象を理解しようとするものが主流になっているが、物質系材料研究におけるビッグデータ（ここでは物性ビッグデータと呼ぶ）の活用の仕方は、それとはやや異なる。物性ビッグデータの活用では、大型装置によって観測・測定データを大量に集めるというよりも、さまざまな物質・材料に関する知見・データ・技術等を集約し、横断的な活用（interoperability）を可能にすることに重点がある。先人が蓄積してきた材料科学に関する知見や技術のデジタル化、従来型の材料科学のデータの収集・統合等によって、それを実現する。より詳細には、①質・信頼性・精度の均一性を担保したデータをメタデータとともに収集し、②データ群を扱う計算ソフトウエアやツールやアプリケーション、実験で得た測定・観測データ、論文とそのデータマイニング結果等を集約し、③これらからデータの関連性・傾向や物質の特性を分析し、新しい材料特性を予測し、その予測に基づいた実験・計算を提案することや、④データ科学や機械学習・人工知能技術を用いたプロセス全体の再構成・最適化によって、既知の積み上げでは想定し得ない新たな解を導出することを狙っている。これらによって、材料開発プロセスを短縮し、開発投資を最大化し、日本社会の多様な需要に応えることが期待されている[26)]。

このような物性ビッグデータのアプローチは、これまでの物質・材料科学にデータ科学を融合させたデータ駆動型材料研究とも呼ばれるものである。現在、日本では、このアプローチによる三つの大型プロジェクトが進行中である。

（1）JST イノベーションハブ構築支援事業「情報統合型物質・材料開発イニシアティブ（$MI^2I$：Materials research by Information Integration Initiative）」（2015 年～）
（2）内閣府戦略的イノベーション創造プログラム（SIP）「革新的構造材料」（2014 ～ 2018 年、研究機関 10・大学 37・企業 31）のうちのマテリアルズインテグレーション（MI）
（3）JST 戦略的創造研究推進事業さきがけ「理論・実験・計算科学とデータ科学が連携・融合した先進的マテリアルズインフォマティクスのための基盤技術の構築」（2015 年～）

従来の研究手法が物質の組成や結晶構造を入力して物性を評価するものであるのに対して、（1）は望ましい物性・機能を持つ物質を設計するという逆問題を解くことに挑戦している。そのためには、データから物性値を正確に予測する技術と、それをもとに膨大な可能性の中から望みの物性や機能を持つ物質・材料を発見する技術が必要である。産学官の協働体制・オープンイノベーションにつながるハブ拠点化、新物質・材料開発手法のパッケージ化・システム化、より広範囲の物質・材料系への発展・普及、組織的な人材育成、産業界とアカデミアの両者が活用できるデータベース構築等により、データ駆動型材料研

究のデータプラットフォームを目指している。このデータプラットフォームを試用するコンソーシアムには、既に 40 社以上の企業会員、約 100 名の活動員が登録されている。

（2）のマテリアルズインテグレーション[27]は、当分野に蓄積された多くの材料の知やデータと計算科学を組み合わせて、材料の組織・特性・性能を予測するシステムを目指している。材料科学の理論、実験、解析、シミュレーション、データベース等の科学技術をツール群として統合し、材料や部材の研究開発時間を短縮する。

これらに類似する取り組みに ICME（Integrated Computational Materials Engineering）がある。米国では 2000 年頃から国防総省の研究助成による基礎検討がなされ、さまざまなシミュレーションソフトウエアをマルチスケールシミュレーターとして統合して、材料開発に活用する試みが進められてきた。2011 年 6 月には MGI（Materials Genome Initiatives）が発足、近年に航空・軍事産業等に活用の例が出始めている。欧州では第 7 次欧州研究開発フレームワーク（FP7）において、ICME の実現に向けたワーキンググループが設置されている[27]。

### （4）科学技術的課題

［教育ビッグデータ］

・デジタル教科書を用いたテーラーメード教育への取り組みは緊急性が高い。2020 年に全国の小中高にデジタル教科書を導入する計画があるが、デジタル教科書を用いた学習ログの蓄積・分析はあまり検討されていない。この学習ログを全国規模で蓄積・分析することで、多様性・個別性に対応したテーラーメードの個別教育が実現し得る。韓国・台湾等のアジア地域で既にデジタル教科書の導入が進められているが、ログの蓄積・分析はまだであり、日本がこの分野で大きくリードできるチャンスにもなる。
・全国規模の教育ビッグデータの蓄積・分析を行う教育用クラウド情報基盤の開発が望まれる。デジタル教科書と同時に全国導入することで、長期間の学習ログが全国規模で蓄積可能になる。プライバシーを保護した上で、これを研究者間で共有・分析することによって、教育学・心理学等の教育関連分野における研究を促進できる。そのため、まずは情報通信設備の整った大学等の高等教育機関から導入し、その効果検証後、初等中等教育に展開するのがよい。

［天文ビッグデータ］

・天文ビッグデータプロジェクトに共通する技術的課題は、数百 PB ものデータをいかに格納し、効率的に取り出し、高速かつ自動的に処理して科学的成果に結びつけるかに集約される。
・データ格納法は効率的なストレージ使用のために必須であり、全ての生データを格納せずに処理済みデータ（派生するカタログデータを含む）のみを格納・蓄積・公開する方法も試みられている。データの性質やアクセス頻度に応じて格納先（SSD、HDD、磁気テープ等）を選択する階層的ストレージや、分散した階層的ストレージに対応したデータフォーマットの研究も必須である。
・データ生産率が高いため、広帯域ネットワークを通じた伝送技術と超広帯域入出力性能

を付加したスーパーコンピューターを組み合わせる技術の開発が急務である。また、人手を介さず高い信頼度でデータを自動処理するためのロバストなソフトウエアシステムや、超巨大データ内から宇宙に関する知見を見いだすための可視化や機械学習のソフトウエア開発が、天文ビッグデータを活用する上で極めて重要となっている。

［物性ビッグデータ］

・生データからの規則性発見・現象理解ではなく、選別・整理された知見・技術としてのデータ集約のアプローチをとっていることから、集約としてのデータベースの質の担保が重要である。さまざまな研究機関に蓄積されてきた実験･測定データの統一的な収集、統一的なメタデータの設計・付与のための効率的な仕掛け作りが必要である。また、企業にとっては競争力・技術力の源泉になるデータであるため、企業からのデータ集約の仕組みも難しい課題である。
・物性・材料分野の知や経験に機械学習・AI 技術も駆使し、望ましい物性・機能を持つ物質を設計するという逆問題は挑戦的な技術課題である。

### (5) 政策的課題

［教育ビッグデータ］

・エビデンスやデータに基づく政策決定と評価の推進が望まれる。海外では e-Learning や MOOCs などのオンライン講義の利用が盛んで、そこで蓄積された教育ビッグデータを分析する LA 研究が多い。しかし、日本はオンライン講義がまだ盛んでないため、小中高・大学等での対面型講義からログを蓄積するのがよい。これが教育の質保証と継続的な改善につながり、さらに各教育機関に蓄積されたデータを統合分析することで、全国にわたる教育ビッグデータの蓄積と､それをエビデンスとした国の教育政策の提案･評価が可能になる。また、教育効果を評価するには大規模で長期的な追跡調査が必要であるため、長期的な大型研究の予算化が求められる。米国・欧州では、長期の大型研究費の下で研究が推進され、米国 Unizin や欧州 LACE 等のコミュニティーも形成されている。中国・韓国でも、近年注目が非常に高まり、近く予算化されると思われ、日本においても競争力のある研究推進・コミュニティー作りが望まれる。
・教育ビッグデータの蓄積には、各教育機関におけるインターネットの整備や一人一台のコンピューターの利用が不可欠である。さらには、ICT インフラ整備だけでなく、それを活用した教育効果・学習効果の高い授業の設計・導入等、その実践のための教員研修・養成も重要である。また、実践結果のデータやエビデンスを共有する枠組み、継続的参照・分析を可能にする学習履歴フォーマットの標準化も重要である。
・教育ビッグデータの分析を行う教育データサイエンティストの育成も重要である（大学学部課程や大学院専攻コースの創設等)。

［天文ビッグデータ］

・米国 NSF は科学的成果を最大化するため、大型プロジェクトに対して原則データ公開を求めてきた。その結果､例えば SDSS カタログを用いた論文数は約 3000 件にもなる。日本でも同様の施策がビッグデータ推進に有用と考える。

・データ公開に関しては、国立天文台を中心にバーチャル天文台システムが構築され、その公開データによる研究成果も多数出ているものの、国際共同・国際競争の環境において人員・予算とも十分とは言えない。データ収集・公開に加えて、関連ソフトウエア技術開発や情報技術活用（分散データ管理、分散 DB、自動データ処理、統計的データ処理、機械学習による知識発見等）、日本が高いレベルにある情報学・統計科学やスーパーコンピューター技術との連携等のための責任と競争力のある体制構築・投資が望まれる。
・天文分野で創出するデータ量は、国内だけでも数十 PB 規模となっている。これを格納するストレージの予算が膨大となるとともに、ストレージの更新に極めて長い時間を要するようになった（〜数ヶ月）。過去のデータは廃棄できないため、格納量は増大の一途をたどっている。この問題を解決するため、個別の研究機関がストレージをそれぞれで調達・運用するのではなく、国が学術分野全体で利用できるクラウドストレージを用意して、各研究機関に利用させる制度に変えていくべきである（参考：カナダの COMPUTE Canada[28]）。
・この分野の研究スタイルは、従来「ものづくり」（ハードウエア重視）的なスタイルだったが、天文ビッグデータ研究では、むしろソフトウエアが重要である。そのための専門人材の育成・処遇も重要課題である。

［物性ビッグデータ］
・日本での取り組み開始は、米国 MGI に 5 年の後れをとった。とはいえ、米国・欧州もさまざまな試行錯誤を展開しているフェーズであり、データ戦略・データ外交を含めて日本の道筋を描いて取り組むことが重要になっている。例えば、米国 MGI の後追いではなく、日本が得意とする職人芸・知を踏まえつつ、計算科学・機械学習の活用によって人間の限界を超えるアプローチ等が考えられる。

## (6) キーワード

価値創造、課題解決、機械学習、深層強化学習、予測型意思決定最適化、教育ビッグデータ、天文ビッグデータ、物性ビッグデータ、ラーニング･アナリティクス、エデュケーショナル・データマイニング、バーチャル天文台、マテリアルズインテグレーション、情報統合型物質・材料開発、データサイエンティスト

## (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 | ○ | ↗ | 政府系投資・研究組織が強化されつつある（例えば、JST の CREST・さきがけ等における AI・ビッグデータ関連の活動等）。企業におけるこの分野の研究開発投資も増加している。 |
| | 応用研究・開発 | ○ | ↗ | 市場における存在感の面では、先行している米国との大きな差があるが、深層強化学習や予測型意思決定最適化等を用いた先進的な応用事例が見られ、伸びてきている。教育・サイエンス分野においても、M2B システムやバーチャル天文台等が注目される。 |

| | | | | |
|---|---|---|---|---|
| 米国 | 基礎研究 | ◎ | ↗ | ビジネス分野での先進事例でリードしているだけでなく、教育・サイエンス分野の基礎的な投資（大規模望遠鏡建設等）も続けられており、存在感は大きい。 |
| | 応用研究・開発 | ◎ | ↗ | インターネット系の巨大 IT 企業（Google、Facebook、Amazon 等）が市場を先導し、研究開発や先端技術応用においても大きな存在感がある。また、教育・サイエンス分野でも MOOCs 活用や Edtech ベンチャー等、活気がある。 |
| 欧州 | 基礎研究 | ○ | ↗ | 英国の DeepMind（Google が買収）が深層学習の最先端技術と応用で注目されている。 |
| | 応用研究・開発 | ◎ | ↗ | DeepMind が AlphaGo 他、先進的応用で大きな存在感を示している。教育・サイエンス分野においては、膨大な天文ビッグデータ処理を行う LOFAR、ラーニング・アナリティクスの Open University 等、世界をリードする取り組みが見られる。 |
| 中国 | 基礎研究 | ○ | ↗ | 深層学習技術に関する急速な追い上げが見られ、潜在的に大量データ収集が進めやすい状況にあるため、活用面でも勢いが出てきている。教育・サイエンス分野での活用にも基礎的な投資がされつつある。 |
| | 応用研究・開発 | ○ | → | 応用の広がりが出てきているものの、独自性の面では弱い。 |
| 韓国 | 基礎研究 | △ | → | 特筆すべき活動は見られない。 |
| | 応用研究・開発 | △ | → | 特筆すべき活動は見られない。オンライン教育は普及しており、教育分野の関心は高い。 |

（注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
（注２） 現状　※わが国の現状を基準にした評価ではなく、CRDS の調査・見解による評価である。
◎　特に顕著な活動・成果が見えている、　○　顕著な活動・成果が見えている
△　顕著な活動・成果が見えていない、　×　活動・成果がほとんど見えていない
（注３） トレンド
↗：上昇傾向、　→：現状維持、　↘：下降傾向

## (8) 参考文献

1）喜連川優 , ビッグデータ , 學士會会報 918,2016.
2）緒方広明 , 殷成久 , 毛利考佑 , 大井京 , 島田敬士 , 大久保文哉 , 山田政寛 , 小島健太郎 , 教育ビッグデータの利活用に向けた学習ログの蓄積と分析 , 教育システム情報学会誌 33（2),pp. 58-66,2016.
3）矢野和男 , 渡邊純一郎 , 佐藤信夫 , 森脇紀彦 , ビッグデータの見えざる手 ―ビジネスや社会現象は科学的にコントロールできるか―, 日立評論 95（6-7),pp. 432-433,2013.
http://www.hitachihyoron.com/jp/pdf/2013/06_07/2013_06_07_03.pdf
4）Cowan, G., Germain, C., Guyon, I., Kegl, B., Rousseau, D., The Higgs Boson Machine Learning Challenge, JMLR Workshop and Conference Proceedings（42), 2014.
http://jmlr.csail.mit.edu/proceedings/papers/v42/
5）棗田昌尚 , 落合勝博 , 朝倉敬喜 , 林司 , インバリアント分析技術の大規模物理システムへの適用―原子力発電所の監視への適用を例に― , 情報処理学会デジタルプラクティス論文誌 6（3),2015.
6）Silver, D., et al., Mastering the Game of Go with Deep Neural Networks and Tree Search, Nature 529, pp. 484-489, 28 January 2016.
7）AlphaGo 対李世ドル , Wikipedia,

https://ja.wikipedia.org/wiki/AlphaGo%E5%AF%BE%E6%9D%8E%E4%B8%96%E3%83%89%E3%83%AB
8）奥田遼介 ,Deep Learning を用いたロボット制御 , 第 9 回科学技術におけるロボット教育シンポジウム ,2016.
http://www.slideshare.net/ryokuta/deep-learning-64339091
9）藤巻遼平、村岡優輔 , 伊藤伸志 , 矢部顕大 , 予測から意思決定へ ～予測型意思決定最適化～ ,NEC 技報 69（1),2016.
http://jpn.nec.com/techrep/journal/g16/n01/pdf/160115.pdf
10）IBM Corp., This is Watson, Special Issue of IBM Journal of Research and Development, Issue 3-4, May-June 2012.
11）「AI、がん治療法助言 白血病のタイプ見抜く」, 日本経済新聞 ,2016/8/4.
http://www.nikkei.com/article/DGXLZO05697850U6A800C1000000/
12）http://ies.ed.gov/sbir/2015awards.asp
13）LACE（Learning Analytics Community Exchange),
http://www.laceproject.eu
14）OWL（Open World Learning),
http://www.open.ac.uk/iet/main/research-innovation/research-projects/owl
15）TeSLA（Adaptive Trust-based E-assessment System for Learning),
http://www.open.ac.uk/iet/main/research-innovation/research-projects/adaptive-trust-based-e-assessment-system-learning-tesla
16）https://www.imsglobal.org/specifications.html
17）https://www.adlnet.gov/adl-research/performance-tracking-analysis/experience-api/
18）国立天文台すばる望遠鏡の超広視野カメラ Hyper Suprime-Cam,
http://www.naoj.org/Projects/HSC/j_index.html
19）The Atacama Large Millimeter/submillimeter Array（ALMA),
http://alma.mtk.nao.ac.jp/j
20）The Large Synoptic Survey Telescope（LSST),
http://www.lsst.org/lsst/
21）The Panoramic Survey Telescope & Rapid Response System（Pan-STARRS),
http://pan-starrs.ifa.hawaii.edu/public/
22）国立天文台 The Thirty Meter Telescope（TMT）推進室 ,
http://tmt.nao.ac.jp/
23）The European Extremely Large Telescope（E-ELT),
http://www.eso.org/public/teles-instr/e-elt.html
24）The LOw Frequency Array（LOFAR), http://www.lofar.org/
25）The Square Kilometre Array（SKA), http://www.skatelescope.org/
26）Borgman, C. L., Big Data, Little Data, No Data, MIT Press, 2015.
27）小関敏彦 , 材料データとマテリアルズインテグレーション , 情報管理 59（3),2016.
28）COMPUTE Canada, https://www.computecanada.ca/

### 3.3.7　ビッグデータに関わる制度設計

#### (1) 研究開発領域の簡潔な説明

ビッグデータの流通・活用を支える仕組み・制度を取り上げる。まず、データの流通・活用を促進する仕組みとして、最小限の制約のみで誰でも自由に利用・加工・再配布ができるオープンデータに着目する。その一方で、何でもオープンデータ化できるわけではない。データに関わる権利は確保されるべきであり、権利が守られていることで、データの流通・活用が進められるという面もある。これを支える法制度として、ビッグデータと関係の深いプライバシー保護と著作権にも着目する。

| | データ流通・活用を促進する仕組み | データに関わる権利確保のための法制度 | |
|---|---|---|---|
| | **オープンデータ** | **プライバシー保護** | **著作権** |
| 動向 | オープンガバメント／オープンソースソフト／オープンアクセス → オープンデータ（3つの流れが収斂）<br>米data.gov<br>英data.gov.uk<br>Open Data Institute（ODI）<br>セマンティックWeb<br>Linked Open Data (LOD)<br>Linked Data<br>政策 ↔ 技術（W3Cでの標準化）<br>コミュニティ（Open Knowledge International (OKI)） | 日本<br>2003 個人情報保護関連5法<br>2015 個人情報保護法改正(匿名加工情報制度等)<br>欧州<br>1995 データ保護指令<br>2016 一般データ保護規則(各国に直接適用、罰則強化)<br>米国<br>1974 プライバシー法<br>(包括法なし、セクトラル方式)<br>2015 消費者プライバシー権利章典法案 | 元データ → ビッグデータ解析 → 解析結果<br>● 元データの著作権問題<br>創作性の有無と権利制限規定<br>● 解析結果の著作権問題<br>1)成果物が元データの著作権を侵害しないか<br>2)成果物に著作権が発生するか<br>● 解析結果の活用に伴う著作権問題<br>● 国・自治体データの公開時(オープンデータ化)の著作権：日米で異なる<br>● 人工知能による創作物の著作権問題 |
| 課題 | ● 政策としてのオープンデータ戦略<br>● 先駆的な適用事例からのフィードバックによる次世代技術開発<br>● 人工知能の知識源としての活用対応 | ● パーソナルデータ取得時の本人同意の形骸化の問題<br>● パーソナルデータ匿名化時のプライバシー保護の問題 | ● デジタル・ネット時代に即した改革(フェアユース規定、人工知能を介した創作物の著作権等)<br>● 巨大プラットフォーマーによる支配 |
| 国際比較 | [欧]セマンティックWeb・LOD研究の本拠地。OKI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。<br>[米]政策主導、OSS連携に強み。<br>[日]遅れているが立ち上がりつつある。 | [欧]削除権・忘れられる権利の導入や企業への罰則強化等を含む一般データ保護規則、匿名化時のプライバシー保護問題等の取り組み等で先行。<br>[米]民間部門を規制する包括的な個人情報保護法はなく、企業の自主規制に委ねられている。著作権法はフェアユース規定で迅速対応を確保。<br>[日]個人情報保護法改正で事業促進。フェアユース規定には未対応。 | |

図 3-3-8　領域俯瞰：ビッグデータに関わる制度設計

#### (2) 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］

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され、現在の形に至っている。その一つは、政策としてのオープンガバメント（開かれた政府）の流れを受けたものである。米国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、米国政府がオープンガバメントを進めると宣言した。このための政策は主に透明性（Transparency）、参加（Participation）、協同（Collaboration）という三つから成る。このうち透明性の実現には、政府保有の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。実際、米国政府は data.gov をというサイトを立ち上げ、そこでさまざまなデータを公開していった[1)]。

二つめはオープンソースソフトウエア（OSS）の流れを受けたものである。OSS の活動は、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、技術者にも基

本的な考え方として広く受け入れられてきた。この考え方の延長として、ソフトウエアだけでなくデータもオープンになるべき、データは誰でも使えて再利用できるべきと考えられるようになった。また、OSS に関わる人々がオープンデータの活動にも関わるというケースもよく見られる。

三つめは学術界で近年広まってきたオープンアクセス（OA）の流れを受けたものである。OA は、学術成果は社会で広く共有すべき、有料でしか読めない論文では成果の共有ができないので、自由に読めるオープンな形で論文を公開すべきという考え方である。有名なものとして、2002 年に発表されたブダペスト・オープン・アクセス・イニシアチブがある。学術論文の OA は、研究成果を文献（論文）として無料で自由に閲覧できる方向に進んでいることに加え、近年のデータ科学の盛り上がりも受け、データとしても公開・利用を可能にする方向（オープンデータに通じる）に向かっている。

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になっている。以下、これを政策、技術、コミュニティーという 3 面から簡単にまとめる。

①政策

米国と英国が先行している。米国は最初に政府のオープンデータポータル data.gov を公開した。英国が data.gov.uk を開設したのはそのあとだが、英国は国を挙げて積極的にオープンデータへの取り組みを進めている。特に Open Data Institute（ODI）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。欧州は EU が 2000 年前半から公共セクター情報の再利用に取り組んでおり、取り組みの歴史は長い。

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」を発表した。2015 年に決定された世界最先端 IT 国家創造宣言にも、オープンデータは目指すべき社会・姿を実現するための取り組みの一つとして含められた。2016 年には官民データ活用推進基本法において、国および地方公共団体等におけるデータ活用の推進が規定された。

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 OSS コミュニティーにおける開発と標準化の三つに大別されるが、相互に関係し合っている。アカデミア研究は主にセマンティック Web の分野で進められている。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 関連で、RDF（Resource Description Framework）、RDFS（RDF Schema）、OWL（Web Ontology Language）等の記述言語、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）、最近では、オープンデータに多くみられる表形式を適切に扱うための技術（“CSV on the Web”）等の技術が開発された。これらの多くは World Wide Web Consortium（W3C）に標準として提案され、採択されている。LOD は、セマンティック Web 分野で開発・標準化された技術による、Web 上のデータを公開・利用する方式あるいは公開されたデータセットであり、従来の Web が「文書の Web」であるのに対して「データの Web」とも言われる [2)]。

LOD 研究は欧州の研究グループ（中でもドイツ）が特出している。FP7 および後継の Horizon 2020 では多くのセマンティック Web 関連のプロジェクトが進められている。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・利用等）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等がこのグループおよびその周囲で作られている。なお、2015年あたりから、Wikipediaのデータ版といえるWikidataをLODとして使うことも増えてきた。

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ーベースで行われている。特に英国に本拠をもつOpen Knowledge International（OKI、旧称Open Knowledge Foundation:OKF）はさまざま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等がその例である。

③コミュニティー

このような政府の動きや技術開発をつなぐのに重要な役割を果たしているのがコミュニティーである。主に非営利組織を母体に技術者や市民が募り、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。英国では先にあげたOKIが組織としては著名である。日本でもOpen Knowledge Japan（OKJP、旧称Open Knowledge Foundation Japan:OKFJ）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、コンペティション等を企画・運営し、オープンデータを広める活動を行っている。

［ビッグデータとプライバシー］

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、移動履歴等がある。

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。日本では、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、個人情報保護法の改正に向けた議論が行われ、2015年に改正個人情報保護法が成立している[3)]。この改正では、特定の個人を識別することができないように個人情報を加工した匿名加工情報の新設、要配慮個人情報の導入、個人情報保護委員会の設立等、データを取得した個人の同意なしでデータを流通・利用活するための新しい枠組みが創設された。

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、2016年に一般データ保護規則が成立している[4)]。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、「削除権」ないし「忘れられる権利」の導入、罰則の強化等の点が特徴になっている。

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、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。さらに、その後、2015 年に、この権利章典を基にした「消費者プライバシー権利章典法案」が公開されている [5]。

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 [6]。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、いかにして分かりやすい表示・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である。この課題については、日本でも経済産業省等によってさまざまな取り組みがなされている（これに関しては後述する）。

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、日本では、「パーソナルデータに関する検討会」等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、英国の ICO が 2012 年に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「匿名化技術に関する意見書」を公表している [7]。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、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・再識別技術が発達してき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。さらに、ISO/IEC JTC1 SC27/WG5 では、De-identification 技術に関する国際規格の検討が進められている。

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。

［ビッグデータと著作権］

ビッグデータに関わる著作権問題は、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。

①「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）等がこれにあたる。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要らない。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
・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 等のサービス・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

②「解析結果と著作権」の問題

成果物が元データの著作権を侵害しないかと、成果物に著作権が発生するかという二つの問題がある。

・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。

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、著作権等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見られる。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等から、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。

前述した国・自治体のオープンデータに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。日本では政府情報は法令や裁判所の判決等は著作権の対象にはならないが（12 条）、白書等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するオープンデータ政策を推進するにあたっても英米に後れをとっている。

### (3) 注目動向

オープンデータ関連は、その次世代技術として LOD が広く関心を集めており、アーリーアダプターとしての適用事例が複数分野で見られるようになってきた[2)]。

・LOD に関する技術は 2014 年頃までに一通りの技術標準が作られ、現在は、これらの技術を用いたシステム、アプリケーション、サービスが作られるようになった段階である。アーリーアダプターとしての適用事例は、クロスドメイン関連（前述の DBpedia 等）、図書館関連、バイオサイエンス関連、政府データ関連で先導的なものが立ち上がっている。さらに、これらの適用経験から不足する機能が明らかになり、そのフィードバックを受けた新たな技術要素が第 2 弾の標準化として検討され始めている。
・このような技術適用を支援するように、商用を含む LOD 公開支援サービス等が登場し、技術的知識がなくても保有するデータを LOD として簡単に公開できる環境が整いつつある。LinkData.org の簡便な LOD 化ツール、jig.jp 社のオープンデータプラットフォーム（有料サービスを 2014 年 6 月開始）、インフォラウンジ社の Datashelf 等があげられる。

ビッグデータのプライバシー問題は、前述の 2 つの観点を中心に、さまざまな委員会・プロジェクトにおいて検討されている。
・第一の観点、パーソナルデータ取得時にいかにして分かりやすい表示・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問題は、経済産業省の IT 融合フォーラム・パーソナルデータ WG（2012 年～ 2013 年）で基本的な問題提起がなされた [10]。この WG では、分かりやすい表示とするために求められる要素や具体的な手法等が示された。これを引き継ぐ形で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、2014 年 3 月に事前相談評価・評価基準書が公表された [11]。これらの流れを受けた国際標準化への取り組みも進められており、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 [12]。これを ISO/IEC JTC1 SC27/WG5 において国際規格化する提案を日本から行い、現在、標準化活動が継続されている。
・第二の観点、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は、まず医療分野において、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」で議論がなされた。2013 年に規制改革会議創業等 WG での議論や、世界最先端 IT 国家創造宣言の公表を経て、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。この検討会は 2014 年にも引き続き実施され、その技術検討 WG で匿名化技術に関する詳細な検討がなされたことは注目される。これらの検討を踏まえて、2015 年改正個人情報保護法において匿名加工情報に関する規律が導入された。同改正法では、個人情報取扱事業者が匿名加工情報を作成する際の義務や、匿名加工情報取扱事業者が匿名加工情報を第三者に提供する際の義務等が規定されている。ただし、具体的な加工方法等、個人情報保護委員会規則に委ねられている部分も多い。
・プライバシー・個人情報の保護が不十分なサービスを公開してしまい、消費者の反発や社会的批判を受ける事態に陥るのを防止するには、新しい情報システムやサービスの導入や実施の前に、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）の実施が有効である。PIA は我が国ではまだあまりなじみがないが、米国、カナダ、英国、オーストラリア等の諸外国では既に実施されている。

この PIA は現在、ISO/IEC29134 という名称で、ISO/IEC JTC1 SC27/WG5 において国際標準化の作業が進められている [13)]。

著作権関連では、人工知能を使って書いた小説が 2016 年の星新一賞の 1 次審査を通過したという話題もあり、人工知能（AI）を介した創作物の著作権をどう考えるべきかが現実の問題として検討され始めた。

・内閣に設置された知的財産戦略本部の次世代知財システム検討委員会は、2016 年 4 月に報告書を発表した。その第 3 章「新たな情報財の創出と知財システム」で取り上げた主要課題の一つめが「人工知能によって生み出される創作物と知財制度」である。
・この報告書によると、「現在の知財制度上、人工知能が生成した生成物は、人工知能を人間が道具として利用して創作をしていると評価される場合には権利が発生しうる。他方で、人間の関与が創作的寄与と言えず、人工知能が自律的に生成したと評価される場合には、生成物がコンテンツであれ技術情報であれ、権利の対象にならない」というのが一般的な解釈とされる。しかし、これらは創作過程の違いであって、その結果の創作物においては外見上見分けることができない。しかも、AI は膨大な数の創作物（の外見を持つもの）を高速に生成し得る。一方、AI は創作において、新しい手法や文化・イノベーションを生み出す可能性もあり、新しい時代に知財制度はどう対応していくべきか、方向性が論じられている。
・その検討において、膨大なデータを収集・保有し、AI 技術の開発にも積極的に投資している海外の巨大プラットフォーマーの存在感は大きい。上記報告書の第 3 章では、当面、具体的に進めていくことが考えられる事項を 3 点に整理しているが、その 2 点目で「制作ができるような人工知能の構築において重要なビッグデータの収集・活用に優位性を有するプラットフォーマーについて、ビジネスモデルの実態把握等を含め、その影響力について調査分析を行う」としている。

### （4）科学技術的課題

・オープンデータの次世代技術として位置付けられる LOD は、第 1 弾の標準化を経てアーリーアダプターによる適用事例を通した次の技術発展段階に入っている。日本としてもキャッチアップしていく必要がある。
・第 3 次 AI ブームに対するビッグデータの貢献は大きいが、深層学習等のブラックボックス的な機械学習の学習データとしてビッグデータが使われる形と、IBM Watson のような自然言語処理・質問応答の知識としてビッグデータが使われる形がある。特に後者はオープンデータ・LOD と容易に結びつくものであり、応用領域として広がる可能性が高い。
・プライバシー保護に関わる科学技術的な側面や課題は、本俯瞰報告書のセキュリティー区分の第 3.5.4 節「プライバシー情報の保護と利活用」が詳しい。また、ビッグデータ区分の第 3.3.5 節「ビッグデータ活用促進技術」でも触れている。

### (5) 政策的課題

・オープンデータ戦略は、米国のオープンガバメントに代表されるように、政府の戦略・政策との結びつきが強い。先行する米国・英国等の動きを踏まえつつ、日本のオープンデータ戦略・政策を鮮明化し、推進したい。英国の ODI は各国にノードという形で関連組織を増やしており、日本では大阪、アジアでは他にはソウルにノードがある。ODI の動き・影響にも注意が必要である。

・パーソナルデータの匿名化の課題については、前述の通り海外でも活発な議論がある。日本国内での議論は、これらの海外動向を十分に検討できていない面があり、今後も海外の最新動向を注視し、それらを十分に踏まえて、検討に反映していくことが重要である。

・プライバシーポリシーの確認・同意が形骸化している問題に対しては、ユーザーの実質的な本人同意が得られるよう、ポリシー内容を分かりやすく簡潔に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等による表示が検討されている。しかし、さまざまな表示方法等が提案されているのが実情で、ルールの明確化や標準化が課題である。

・「知的財産推進計画 2016」は四つの柱の一つに「第 4 次産業革命時代の知財イノベーションの推進」を掲げ、今後取り組むべき施策の一つに「デジタル・ネットワーク化に対応した次世代知財システムの構築」を挙げて以下の提案をしている。
「ネットワーク時代の著作物の利用への対応の必要性に鑑み、新たなイノベーションへの柔軟な対応と日本発の魅力的なコンテンツの継続的創出に資する観点から柔軟性のある権利制限規定について検討し、必要な措置を講ずる。」

・柔軟性のある権利制限規定の代表例は米著作権法が定めるフェアユース規定、つまり、フェア（公正）な利用であれば著作権者の許諾なしに著作物の利用を認める規定である [14)]。知財本部は過去にも「知的財産推進計画 2009」で「権利制限の一般規定（日本版フェアユース）の導入」を提案、文化庁で検討したが、権利者の反対などで骨抜きにされた。こうした経緯からか、今回は「柔軟性のある権利制限規定」という言葉に置き換えているが、提案を受けて、文化庁が文化審議会著作権分科会にワーキングチーム設けて検討中である。

・人工知能を活用したビジネスの特性を踏まえて、学習済みモデルに着目した知的財産権の議論がある [15)]。著作権に限らず、ビジネスモデル・契約も含めた考え方の整理や制度整備が必要と思われ、産業構造審議会でも検討が進められている。

### (6) キーワード

オープンデータ、Linked Open Data、Linked Data、セマンティック Web、パーソナルデータ、プライバシー、個人情報保護法、一般データ保護規則、本人同意、匿名化、著作権、権利制限規定、フェアユース

## (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 年に個人情報保護関連 5 法が整備された。2015 年にはパーソナルデータに関する検討会での議論を経て個人情報保護法が改正された。しかし、フェアユース問題等、全般に動きが慎重で時代変化への対応が遅れ気味の面がある。<br>セマンティック Web 研究は一通りあるものの、オントロジー研究以外は相対的に弱い。最近は関心が高まり、コミュニティーが広がっている。 |
| | 応用研究・開発 | ○ | ↗ | 2015 年の個人情報保護法改正で匿名加工情報の制度が導入され、事業者が個人情報を匿名化すれば本人の同意なしに個人情報を利用できるようになり、ビッグデータビジネスが展開しやすくなった。2016 年に行政機関個人情報保護法と独立行政法人等個人情報保護法も改正され、非識別加工情報の提供制度も導入された。これも行政機関等に眠る個人情報を利活用しやすくし、ビッグデータビジネスの追い風となる。 |
| 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。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す。<br>AI・Web 研究に強みがあり、興味深い研究も出てくるが、全体としての動きにはなっていない。 |
| | 応用研究・開発 | ◎ | → | 2012 年に消費者プライバシー権利章典が提案され､2015 年には消費者プライバシー権利章典法案が公開された。民間部門を規制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り、さまざまな分野で自主規制のためのガイドラインの制定等、産業活性を図りつつルール形成の努力が行われている。<br>行政データ専用サイト data.gov を世界で初めて開設するなどオープンガバントの取り組みが進んでいる。OSS コミュニティーとの連携も強み。 |
| 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。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。<br>セマンティック Web 研究の本拠地で、EU プロジェクトで層の厚い研究が進められている。 |
| | 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、2016 年に一般データ保護規則が成立した。削除権 / 忘れられる権利の導入や、企業に対する罰則の強化等が特徴になっている。<br>欧州委員会は 2016 年に European Data Portal を公開。英国データポータル data.gov.uk も充実。OKF 主導によるオープンデータ関連のソフトウエアやサービスが多数生み出されている。 |
| 中国 | 基礎研究 | △ | → | 2006 年に個人情報保護法草案が公表されたが、まだ成立していない。消費者権益保護法、電信およびインターネットユーザー個人情報保護規定等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が、包括的な個人情報保護法は存在しないため、13 億人のビッグデータの産業での活用は進む可能性がある。 |
| | 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」を定めた。2008 年に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、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を設けている。 |
| 韓国 | 基礎研究 | ○ | ↗ | 米国類似のセクトラル方式だったが、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施行された。2015 年には個人情報保護法が改正され、法定損害賠償制度や懲罰的損害賠償制度等が導入された。 |
| | 応用研究・開発 | ○ | ↗ | 未来創造科学部と情報化振興院が、2013 年に民間企業・大学とも連携して実証実験等に取り組むビッグデータ分析活用センターを開設した。2014 年には韓国放送通信委員会がビッグデータ個人情報保護ガイドラインを公表、個人情報の識別化事例集も公表された。位置情報の保護および利用等に関する法律の存在も注目される。 |

（注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

（注２） 現状　※わが国の現状を基準にした評価ではなく、CRDS の調査・見解による評価である。
◎　特に顕著な活動・成果が見えている、　○　顕著な活動・成果が見えている
△　顕著な活動・成果が見えていない、　×　活動・成果がほとんど見えていない

（注３） トレンド
↗：上昇傾向、　→：現状維持、　↘：下降傾向

## (8) 参考文献

1）JETRO/IPA, 米国オープンデータの動向調査 ,2013 年 3 月 ,
http://www.ipa.go.jp/files/000033718.pdf
2）武田英明 , リンクト・オープン・データの原理原則と最近の進歩 , 情報処理 57（7),「リンクト・オープン・データの利活用」特集号 ,pp. 588-593,2016.
3）個人情報の保護に関する法律及び行政手続における特定の個人を識別するための番号の利用等に関する法律の一部を改正する法律 ,2015 年 9 月 3 日成立 , 同月 9 日公布 .
http://www.kantei.go.jp/jp/singi/it2/pd/info_h270909.html
4）Regulation（EU）2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC（General Data Protection Regulation）
http://eur-lex.europa.eu/eli/reg/2016/679/oj
5）Administration Discussion Draft: Consumer Privacy Bill of Rights Act
https://obamawhitehouse.archives.gov/sites/default/files/omb/legislative/letters/cpbr-act-of-2015-discussion-draft.pdf
6）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 ,pp. 269-279,2014.
7）Opinion 05/2014 on Anonymisation Techniques,
http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf
8）Big Data and Privacy: A Technological Perspective,
https://obamawhitehouse.archives.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）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
11）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
12）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
13）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),pp. 33-56,2014.
14）城所岩生 ,「フェアユースが経済を救う ーデジタル覇権戦争に負けない著作権法」, インプレス R&D,2016.
15）江村克己 , 人工知能の活用と共有経済の進展から考察するこれからの知的財産 , 知財研フォーラム（107),pp. 30-37,2016.

### 3.3.8　新計算原理

#### (1) 研究開発領域の簡潔な説明

森羅万象のデジタル化が誘う超ビッグデータ時代に資する新しいコンピューティングパラダイムとしての計算原理のこと。既存の計算原理の性能限界を突破するための「ポスト・ムーア」時代を見据えた研究開発領域である。

#### (2) 研究開発領域の詳細な説明と国内外の動向

システム・情報科学技術のトレンドとして、2 章において「社会に浸透する AI、Big Data」を取り上げて説明したように、AI やビッグデータが社会に浸透する今日、ビッグデータ解析と機械学習を組み合わせた新しいサービスやアプリケーションの普及が急激に始まっている。このような計算ニーズが極めて大きくなっていることを背景として、AI 処理を高速化する専用チップ（AI チップ）の開発が世界中で始まっている。森羅万象のデジタル化データの内、スマートデバイス（エッジコンピューティングデバイス）中でローカルに処理できるものは、組み込み AI を使ったエッジコンピューティングにより瞬時に高速処理し、難しく複雑なものはクラウド AI で知的情報処理することにより森羅万象に潜む叡智を掘り起こさなければならない。いわばスマートデバイスは人の五感を凌ぐ超知覚であり、ハイパフォーマンスコンピューター（HPC）は超ビッグデータ時代における叡智の炬火（かがり火）となる超知能（森羅万象解析エンジン）である必要がある。

しかしながら、超ビッグデータ時代においては、既存の計算原理の性能では太刀打ちできないのは明白であり、このような新時代に資する「新計算原理」の必要性が急速に高まっている。しかも、超ビッグデータ時代の到来とともに、「ポスト・ムーア」時代が到来する。したがって、「新計算原理」を実行する今後のコンピューティングデバイス・システムの高性能化・低消費電力化には、限界を迎える微細加工技術の進展による半導体の高集積化を活用したプロセッサーの処理能力向上だけでなく、3 次元積層メモリーや不揮発性メモリー・光ネットワーク・各種省電力技術・新原理デバイス、さらにはそれらを活かすシステムソフトウエアやアプリケーションの進化を誘うコ・デザイン開発が必要不可欠である。

以下では、森羅万象解析エンジンであるハイパフォーマンスコンピューター（HPC）の現状を概観するとともに、HPC と AI と超ビッグデータとの必然的統合に焦点を絞って、「新計算原理」の観点から解説する。

［ハイパフォーマンスコンピューティング（HPC）の現状俯瞰］

2020 年代においてエクサ（$10^{18}$）フロップス（1 EFLOPS: 毎秒 100 京回の浮動小数点演算能力）を達成するエクサ級スーパーコンピューターの実現に向けて、日米欧中の各国がしのぎを削っている。主に、小型化すればするほど高速かつ省電力になるデナードスケーリングの終焉により、必要な並列性が億単位とも予想され、その際に消費電力を約 20 ～ 30 MW に抑えるには 33 ～ 50 GFLOPS/W と、現時点の世界最高値を 5 倍程度改善しなくてはならず、2008 年に達成されたペタ（$10^{15}$）フロップス（PFLOPS）よりはるかに困難である。以下では、HPC 研究の現状を概観する。なお、それらの技術的詳細は、2015 年度版俯瞰報告書（情報科学技術分野）[1] に記載されているので、ここでは割愛する。

HPC 研究としては、(a) 超高性能計算能力を有するスーパーコンピューターのハードウエアおよびソフトウエアのシステム技術、および (b) スーパーコンピューターが有する計算能力を科学や産業などに活用するアプリケーション技術の研究開発、さらに (a) (b) 全体のコ・デザインを挙げることができる。2015 ～ 2020 年にかけて、上記 (a) のシステム技術に関して各国はエクサへの橋渡しとして数百ペタフロップスに至る中間的な性能のマシンを計画しており、数十万～数百万のプロセッサーを内包するハードウエアを、厳しい電力制約化で実現するために、メニーコア・ヘテロジニアス型アーキテクチャーや、積層メモリー・不揮発性メモリー、超高速光ネットワーク、高効率冷却など、IT における最先端のデバイスやハードウエアをインターネットデータセンター（IDC: Internet Data Center）等に先行して採用する傾向が高まっている。加え、巨大な電力を消費する大規模システムの高度な信頼性の達成、数億の並列性の有効なプログラミングなど、大規模スパコンの各種システムソフトウエア技術の研究開発にも多くの投資がなされている。上記 (b) のアプリケーション技術に関しては、新たな数値的な手法やアルゴリズムにより、現行の数十 PFLOPS のトップスパコンにて数百万から 1000 万以上の並列性を有効に活かしたアプリケーション群が多く出現しているが、エクサへ到達する事の技術的な困難さから、アプリケーション・システムソフトウエア・ハードウエアの共同の研究開発「コ・デザイン」の重要さが強調され、各種プロジェクトに取り入れられている。

世界のスパコンの性能の上位 500 位までのコンピューターを LINPACK ベンチマーク（密行列の大規模連立一次方程式を解くベンチマーク）でランキングし、半年ごとに更新される TOP500 リスト[2] では、2016 年 11 月版で、中国国立スーパーコンピューターセンター（江蘇省無錫市）の「神威太湖之光（Sunway TaihuLight）」が 2016 年 6 月の初登場首位に引き続き、ダントツで首位の座を守った（93.0 PFLOPS）。2013 年 6 月から 2016 年 6 月まで首位の座を守ってきた中国防衛大学の「天河 2 号（Tianhe-2）」は、Sunway TaihuLight に比べて約 1/3 倍程度の 33.9 PFLOPS で連続世界 2 位であり、日本の「Oakforest-PACS」（東大・筑波大 :13.5 PFLOPS）は 6 位、「京」は 7 位（理研 :10.5 PFLOPS）であった。なお、Tianhe-2 のプロセッサーは米 Intel 製であるのに対して、驚くべきことに Sunway TaihuLight のプロセッサーも接続回路もすべて中国純製であった。このプロセッサーは、中国の国家並列計算機工程技術研究中心（NRCPC）が独自開発した 260 コア高性能プロセッサー「SW26010」であり、Sunway TaihuLight ではこのプロセッサーを搭載したシステムを 40960 ノードで構築することで、合計 10649600 コアという大規模システムを構築している。

［超ビッグデータと AI と HPC との必然的統合について］

ビッグデータ処理（特に大規模グラフ解析）に関するスーパーコンピューターの国際的な性能ランキング Graph500 リスト[3] においては、グラフの幅優先探索（1 秒間にグラフのたどった枝の数、Traversed Edges Per Second: TEPS）という複雑な計算を行う速度で評価され、2016 年 11 月版で、1 位「京」（理研 :38621 GTEPS）、2 位「Sunway TaihuLight」（中国 :23756 GTEPS）、3 位「Sequoia」（米ローレンスリバモア国立研究所 :23751 GTEPS）、4 位「Mira」（米アルゴンヌ国立研究所 :14982 GTEPS）であった。これらは、このようなデータ集約型の計算は今後市場拡大が予測されているビッグデータ

市場に HPC を適用するためには重要な指標である。また、産業利用など実際のアプリケーションで用いられる共役勾配法の処理速度の国際的なランキング HPCG（High Performance Conjugate Gradient）においては、2016 年 11 月版では、1 位「京」（理研 :602 TFLOPS）、2 位「Tianhe-2」（中国 :580 TFLOPS）、3 位「Oakforest-PACS」（東大・筑波大 :385 TFLOPS）、4 位「Sunway TaihuLight」（中国 :371 TFLOPS）であった。これは、TOP500 では「京」よりも上位で 9 倍以上のスコアを持つ「Sunway TaihuLight」などを上回っており、CPU の演算性能のみならずメモリーやネットワークも含めたシステム全体としての性能バランスを重視して設計開発された京の方が産業利用など実際のアプリケーションを非常に効率よく処理できることを示している。

スーパーコンピューターの電力効率を競うランキング Green500 リスト [4] の 2016 年 6 月版では、国内スパコン開発ベンチャーの PEZY/ExaScaler が開発したスパコン「Shoubu（菖蒲）」と「Satsuki（皐月）」が 1 ～ 2 位を独占した。ここで、2 位の「Satsuki（皐月）」は研究室の 3 $m^2$ のスペースに設置されており、TOP500 にランキングされた世界初のオフィス設置のスパコンであった。なお、Green500 は TOP500 に入っていることが前提であり、3 位には TOP500 を制した「Sunway TaihuLight」が入り、その消費電力性能も 2 位に僅差で迫る勢いであった。そして、Green500 の 2016 年 11 月版では、NVIDIA の最新アーキテクチャー「Pascal」をベースとした GPU「P100」を用いた NVIDIA 社内部向けシステム「DGX SATURNV」が 1 位（9.5 GFLOPS/W）、同じく GPU「P100」を使うスイス国立スーパーコンピューティングセンターの「Piz Daint」が 2 位（7.5 GFLOPS/W）となった [5]。そして、前回 1 位の「菖蒲」は 3 位に、前回 4 位の「Sunway TaihuLight」は 4 位に後退した。なお、今回の Green500 では、PEZY/ExaScaler は新たな値を登録しておらず、3 位になった菖蒲のスコアは前回のものであった。6 位以下では、日本最速の東大・筑波大の「Oakforest-PACS」が 6 位、京大の「Camphor 2」が 9 位に入るなど、日本勢が健闘した。また、ドイツの富士通テクノロジー・ソリューションズ本社に設置された富士通のサーバー機「QPACE3」も 5 位に入った。

近年の IDC におけるクラウドやビッグデータ基盤の出現により、HPC もそれらへの対応や、IDC インフラとの統合化の流れも始まっている。クラウドに関しては米国 DoE や NSF の試験的な導入を初めとして、Amazon HPC Instance[6] や Penguin Computing[7] が商用クラウド HPC サービスを開始している。また、各国のエクサの計画にはビッグデータも重要な項目として記載されている。ハードやソフトのシステムレベルの対応も要求され、シミュレーションが発生する莫大なデータの処理や、一般 IDC では扱えない規模や複雑さのビッグデータを処理するための基盤として期待されている。よって、森羅万象のデジタル化が劇的に進む超ビッグデータ時代においては、スーパーコンピューターが叡智を掘り起こすために必要不可欠なエンジンとなるためには、主に科学技術計算向けのスパコン用チップと深層学習向けの AI チップとの間で、要求性能が大きく乖離してはならない。なぜならば、脳におけるシナプスの発火 [8] という脳神経回路を模した演算処理（深層学習）は究極的には 1 ビット演算であるのに対して、科学技術計算では、現在の倍精度から 4 倍精度、8 倍精度と、求められる演算精度が高まる見通しであるからである。したがって、超ビッグデータ時代においては、演算精度を落として近似計算することも必要である。PEZY Computing を率いる齊藤元章 [9] によれば、演算精度を落とせば、演算コ

アを小さくできるので、今後登場する 7 nm（ナノメートル）プロセスであれば、1 チップに 4 ビット演算のコアを 100 万個搭載でき、現行の GPU の 1000 倍かそれ以上の性能をもった AI チップが実現できる見込みがある。

### (3) 注目動向

［HPC 関連国家戦略について］

我が国では「フラッグシップ 2020」[17] という名称で、ポスト「京」コンピューターの開発が 2014 年度より開始され、2020 年度までに「京」の約 100 倍のアプリケーション実効性能をもつスーパーコンピューターの実現を目指している。また、2016 年 11 月、産業技術総合研究所が、深層学習の演算能力で世界一を狙って「人工知能処理向け大規模・省電力クラウド基盤（AI Bridging Cloud Infrastructure: ABCI）」の開発に乗り出した。深層学習に特化した演算性能で 130 PFLOPS を目指している [18]。

しかしながら、次に述べるように、米国、欧州、中国の動向から楽観視できないのが現状である。バラク・オバマは、米国大統領時代の 2015 年 7 月 29 日、スーパーコンピューターの研究開発を推進する新たな大統領令を発令した。この大統領令は、世界初となるエクサ級スーパーコンピューターの研究開発に専念する「国家戦略コンピューティングイニシアチブ（National Strategic Computing Initiative: NSCI）」を立ち上げ、HPC 分野における将来的な米国の地位強化を目指すものである [19]。このイニシアチブは、米国政府の下、エネルギー省、国防総省、国立科学財団の協力体制で、産学官共同により推進されており、ハード・ソフトのコ・デザインにより、アプリケーションが現行の約 100 倍で実行可能なエクサスケールシステムの開発が進められている。一方、欧州では、2016 ～ 2018 年に Human Brain Project [20] で 50 PFLOPS のスーパーコンピューターを設置する計画が進んでおり、Horizon2020 プロジェクト（2014 年～ 2020 年）の中で、エクサスケール技術の研究開発が進んでいる。また、中国でも、第 13 次 5 カ年計画（2016-2020）において、中国の威信をかけて 2 台の 100 PFLOPS 級スーパーコンピューターシステムの技術研究開発が進められている [21]。

［脳型コンピューティングのための AI チップの潮流について］

脳型コンピューティングのための AI チップの研究開発には、二つの異なる潮流がある。一方は、脳科学の知見に基づいて神経回路の働きを半導体で再現する「ニューロモルフィックチップ」を目指す流れであり、他方は、深層学習の演算を加速させる「深層学習アクセラレーター」を目指す流れである [22]。

前者の脳型コンピューティングに関しては、2016 年 3 月末、米国ローレンスリバモア国立研究所は、IBM リサーチと協力して、IBM のニューロシナプティックコンピューターチップ「TrueNorth」を使って、脳からヒントを得た深層学習のための極低エネルギー効率スーパーコンピューティングプラットフォームの開発を開始した [23]。また、日本では、2016 年 9 月、日本電気と東京大学は、日本の競争力強化に向け戦略的パートナーシップに基づく総合的な産学協創として、「フューチャー AI 戦略協定」を結び、「ブレインモルフィック AI 技術」の早期実現を目指して、脳を模した専用のアナログ回路技術に関する

研究を開始している[24]。

後者の脳型コンピューティングに関しては、深層学習の演算を加速させる「深層学習アクセラレーター」を目指す流れである。この開発競争を先導しているのは NVIDIA であり、前述したように、Green500 の 2016 年 11 月版[4] では、NVIDIA の最新アーキテクチャー「Pascal」をベースとした GPU「P100」を用いた NVIDIA 社内部向けシステム「DGX SATURNV」が 1 位（9.5 GFLOPS/W）を取り、深層学習用スーパーコンピューターの開発にも着手している。また、Google は、機械学習ライブラリ「TensorFlow」向けにカスタマイズした深層学習専用プロセッサー「Tensor Processing Unit」（TPU）を独自開発し、同社のデータセンターで過去 1 年間にわたって使用しているとともに、韓国のプロ棋士イ・セドル氏に勝利した囲碁 AI の「AlphaGo」にも使用していた[25]。驚くべきことに TPU は、深層学習のために開発した ASIC（Application Specific Integrated Circuit、特定用途向け IC）で、GPU（Graphic Processing Unit）や FPGA（Field Programmable Gate Array）といった深層学習処理に使用する他の技術と比較して、消費電力当たりの性能は 10 倍である。TPU は機械学習向けにカスタマイズされているため、演算精度をそれほど必要ではなく、1 演算あたりのトランジスタ数も少なくて済むのが利点であり、消費電力も小さい。したがって、機械学習に最適化された、1 ワットあたりのパフォーマンスが極めて高い AI チップと言える。

［Microsoft StationQ の挑戦］

Microsoft は、IBM、Google や D-wave とは全く違った量子コンピューター戦略をとっており、「新計算原理」に基づくトポロジカル量子ビットを使ったスケーラブルな量子コンピューターの開発に取りくんでいる[26]。

Microsoft は、2016 年 11 月、革新的なトポロジカル量子コンピューターを世に送り出すためのプロジェクトを管理する最高責任者として Todd Holmdahl を任命した[27]。彼は、これまで、Xbox、Kinect モーション・コントローラ、および、HoloLens 拡張現実（AR）ヘッドセットの開発を導いてきた人物であり、今回、従来の量子コンピューターをロバスト性で凌ぐ可能性をもったトポロジカル量子コンピューターを製品化することを使命とした役職に就いたと言っても過言ではない。Microsoft は、世界中の共同研究者と一緒に量子コンピューターを構築するための取り組みを積極的に行っており、研究コンソーシアム的な研究施設 StationQ を世界各国に建設し、トポロジカル量子コンピューティングに関する研究を行っている。

Microsoft は、カリフォルニア大学サンタバーバラ校キャンパスに StationQ Santa Barbara という研究施設を先導的に開設し、トポロジカル量子コンピューティングに関する先導的な基礎研究を行ってきた。StationQ Santa Barbara を率いるのが、数学者 Michael Freedman であり、トポロジーにおける難問とされるポアンカレ予想が 4 次元において成立することを証明し 1986 年にフィールズ賞を受賞した人物である。ここでは、数学者、物理学者、およびコンピューター科学者が共同で、2016 年のノーベル物理学賞受賞テーマの新概念が開いた扉の先に見いだされた物質のトポロジカル相がどのように、頑強でスケーラブルなトポロジカル量子コンピューター・アーキテクチャを構築するために使えるかを理解することを目的として、世界中の大学の研究者とも協力しながら基礎研

究を進めている。また、Microsoft レドモンド本社の敷地にある StationQ Redmond では、量子アーキテクチャーと量子計算に関する研究を行っており、実世界量子アルゴリズムの開発とそれらの意味を理解することに専念するとともに、スケーラブルでフォールトトレラントな量子コンピューターのアルゴリズムやプログラムを作るための包括的なソフトウエア・アーキテクチャをデザインすることに打ち込んでいる。

Microsoft は、世界中の共同研究者と一緒にトポロジカル量子コンピューターを構築するための取り組みを積極的に行っており、キーとなる量子技術を有する世界トップクラスの有力な教授をその立場を維持したまま、戦略的に雇い入れている [26]。例えば、オランダのデルフト工科大学教授 Leo Kouwenhoven、デンマークのコペンハーゲン大学ニールス・ボーア研究所教授 Charles Marcus、スイスのチューリッヒ工科大学教授 Matthias Troyer、豪シドニー大学教授 David Reilly 教授の 4 氏が、各大学の教授の立場を維持したまま Microsoft に戦略的に雇い入れられたのである。Microsoft は、各大学にそれぞれ Station Q Delft、Station Q Copenhagen、Station Q Zurich、Station Q Sydney を開設し、プロジェクト最高責任者 Todd Holmdahl のもと、トポロジカル量子コンピューター研究開発に全 Microsoft StationQ 総力上げてこれまで以上に積極的に取り組む姿勢を見せている [26),27)]。

### （4）科学技術的課題

［ポスト・ムーア時代の到来と新計算原理について］

コンピューティングにおける進歩は科学技術が関係するほとんどすべての領域を変容してきた。とりわけ、新しい発見の探究はコンピューティングのすべての領域を横断したイノベーションをもたらしている。現在、我々は入手可能なデータの Volume（量）、Velocity（速度）、Variety（種類）から洞察を得るための能力の限界に直面しており、コンピューティングにおける共生的な進歩を通して対処せざるを得ない根本的な課題を提起している。前例のない複雑さのデータの本質を理解し洞察を得る能力は、森羅万象のデジタル化が劇的に進む超ビッグデータ時代に資する「新計算原理」によって大いに高められ、超ビッグデータ時代における叡智の光として進む道を正しく照らす炬火（かがり火）となるに違いない。

しかしながら、超ビッグデータ時代の到来とともに、「ポスト・ムーア」時代が到来する。したがって、「ポスト・ムーア」時代を見据えた「新計算原理」デバイスの基礎研究も必須である。当然このような総合的な開発法は、従来の微細化のみに頼るスパコン開発法と比べてはるかにコストがかかり、他のスーパーコンピューターや IDC サーバーから組み込みに至る他の IT エコシステムに対する技術的なレバレッジによる開発コストの分散が総合的なコスト削減、および迅速な進化の鍵となる。その点、他のマーケットにレバレッジが期待できず、かつ国際的にシェアが伸びない我が国のスパコンに特化した開発法は圧倒的に不利になると言える。今後はアプリケーションやシステムソフトウエアの汎用性を担保するハードウエア開発を目指すと共に、コ・デザインのためには国際共同開発を含めた体制を確保することが重要となる。この点、米 DoE と文部科学省間のエクサスケールスーパーコンピューターのシステムソフトウエアに関する日米協力取極 [10] が締結されて

おり、その具体化が今後望まれる。

「ポスト・ムーア」時代を見据えたとき、相対的なデータ移動を可能とするアーキテクチャーとして、チップ間に光配線を用いてアルゴリズムを光で実行することが必要になってくる。また、相対的な計算量の伸びに対して、半導体パッケージや帯域幅の伸びはよくないことを鑑みると、相対的な計算量を減らすには、例えば、間違ってもいいから精度を許容する近似計算をして相対的な計算量を減らすことができるような計算方法を取り入れる必要がある。また、リアルタイム性を、今のビッグデータ処理に求めると、従来型のデータベースでは間に合わない。精緻なマイニングをすると計算量が増えてしまう。FLOPS が上がらなくなったら、どうするのか !? そのためには、ニューロモルフィックコンピューティングや量子コンピューティング等を取り入れて、計算量を減らすことも考えられる。そのためには、新しいアーキテクチャーを設計しなければならない。

マクロ・スコピックな観点から科学技術的・政策的課題を鑑みると、多様性のあるインフラ整備が大切であるが、プレーヤーが少ないのが現状である。また、ミクロ・スコピックな観点からは、やはり「新計算原理」に基づいたデバイスの深化が必須である。IT の研究分野は、応用分野であると思われているが、ロングタームの基礎研究が重要であり、必要不可欠である。企業が研究開発をしているからショートタームで物事を考えがちであるが、すぐに役立つと考えるのは間違いであり、肝に銘じるべきである。しかも、応用研究がもたらした社会実装の成果でもうかったお金を、米国企業（例えば、Microsoft や Google）のように基礎研究に回すべきである。

「新計算原理」は、世の中を変革する可能性が高い。このようなコンピューティングの変革期に、日本から新しいプレーヤーが出てくる可能性がある。そのためには、たくさんの種をまく必要がある。芽が出たら、集中投資すればよい。「ムーアの法則」の終焉が来るのは分かっているが、そのためにも、基礎研究が必要不可欠である。十数年後に待ち受ける「ポスト・ムーア」時代に向けた革命のための基礎研究を急がなければならない。

### (5) 政策的課題

[量子コンピューティングに関する先導的基礎研究からみた人材育成戦略]

2015 年度版俯瞰報告書（情報科学技術分野）[1] の「量子コンピューティングデバイス」の章において、先導的基礎研究からみた人材育成戦略が記載されている。これは、全く「新計算原理」にも当てはまることであり、肝に銘じなければならないことである。以下に、引用文献をアップデートして抜粋する（量子コンピューティングデバイスに関する技術的俯瞰内容は、上記俯瞰報告書[1] を参照のこと）。

ムーアの法則に頼るシリコンチップの高性能化が曲がり角にさしかかり、また、最新のチップはすべて海外に外注生産という状況を鑑みると、我が国は新しい発想に基づく素子のデザインと知財化に注力すべきフェーズを迎えている。量子アニーリングというコンセプト[11),12)] が日本で発明され、米国の研究者がそれに少しだけひねりを利かせて広めることに成功し、そこに価値を見いだしたカナダの D-wave 社[12),13)] がやはり日本で発明された超伝導量子ビット[14)]、磁束量子パラメトロン[15)] といった手法を組み合わせて製品化に成功した過程を解析する事が重要であろう。1980 年代に電子立国として君臨した日本は、

米国で発明されたトランジスタ、酸化膜、プレーナー型接合、集積回路技術を引き継ぎ、高度な製造技術で向上させた一方、時には人まねと非難された歴史がある。量子アニーリング装置の要素技術の多くは日本の発明ということで大きく前進したが、製品化は北米が先んじた。コヒーレントイジング装置[16]も我が国の発明であるだけに、これからが産学共同の腕の見せ所である。いずれにせよ高度な基礎科学的成果をどのように組み合わせるかがイノベーションである。よって世界初の製品化を一途の目標として高度な基礎科学的成果を理解し、組み合わせることができる人材の育成が急務である。世界に通用する工学博士の育成といってよいであろう。そのような人材育成は、我が国の大学で実行するものと、一人でも多くの日本人を世界に飛び立たせる 2 本立てで実行すべきであり、そのような政策の企画・実行が期待される。

## (6) キーワード

超ビッグデータ、ムーアの法則、ポスト・ムーア時代、森羅万象のデジタル化、エッジコンピューティング、高性能アプリケーション、量子コンピューティング、ニューロモルフィックコンピューティング、最適化、グラフ問題、深層学習、人工知能、メニーコア、ヘテロジニアス型アーキテクチャー、低消費電力システム、冷却技術、光ネットワーク、Approximate Computing（積極的な近似計算）

## (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 | ○ | → | ・スーパーコンピューターに関する基礎研究は産官学で行われているが、米国には質・量とも劣る。<br>・ImPACT「量子人工脳を量子ネットワークでつなぐ高度知識社会基盤の実現」の一部で量子最適化装置の開発が進められている。 |
| | 応用研究・開発 | ○ | ↗ | ・京コンピューターが稼働し、対応する超並列アプリケーションの開発が進展し、ポスト京の開発もスタートしている。<br>・深層学習の演算能力で世界一を狙って「人工知能処理向け大規模・省電力クラウド基盤（AI Bridging Cloud Infrastructure: ABCI）」の開発に乗り出した[18]。<br>・スーパーコンピューターベンチャー企業の PEZY Computing が低消費電力スーパーコンピューター開発に取り組んでいる。また、AI ベンチャー企業の Preferred Infrastructure が、深層学習の研究開発を行っている。<br>・富士通は、従来技術を深化させたトランジスタを用いて、組み合わせ最適化問題を 1 万倍速く解く技術を開発した[28]。東芝は、脳型 AI チップをフラッシュ派生技術で省電力化させることに成功した[29]。<br>・産業力の低下から企業における、量子コンピューティングデバイスに関する基礎研究が不振である。 |
| 米国 | 基礎研究 | ◎ | → | ・DoE を中心にスーパーコンピューターに関する研究費の手当がされ、大学・メーカーにも手厚く研究費が渡り、多くの研究者がスーパーコンピューターに関する基礎研究に携わっている。<br>・脳型および量子コンピューティング研究者の層が厚く、様々な基礎研究へ支援策が講じられている。 |

| | | | | |
|---|---|---|---|---|
| | 応用研究・開発 | ◎ | → | ・HPC 用チップは、Top500 において圧倒的なシェアを握っており、それらを利用したアプリケーション開発、システムソフトウエア開発、ハードウエア開発技術も、米国が最先端を走っている。<br>・IBM が、ニューロシナプティックコンピューターチップ「TrueNorth」を開発している。<br>・Google、IBM、および Microsoft といった主要企業が量子コンピューターの開発に挑戦している。 |
| 欧州 | 基礎研究 | ○ | ↗ | ・Horizon2020 で欧州でもビッグデータ等と共に予算措置がなされ、高い性能を有するシステムも各所に設置、アプリケーションやシステムソフトウエア中心に競争力の高い研究がなされている。<br>・英国、オランダ、ドイツ、デンマークなどで量子コンピューティングに関する基礎研究に対する多額の支援が行われている。 |
| | 応用研究・開発 | △ | ↗ | ・一部の特殊なスーパーコンピューターを除き、独自のスーパーコンピューターのシステム研究は弱いが、ARM プロセッサーや Extoll などを含む一部の技術を中心に、新たな開発の試みが試みられている。<br>・脳型および量子コンピューティングデバイスの企業による研究開発が見受けられない。 |
| 中国 | 基礎研究 | △ | ↗ | ・基礎研究はハード・システムソフトウエア・アプリケーションの全領域でまだ日米欧に比べて弱いが、急速に力をつけつつある。<br>・量子コンピューティングに関しては、まだ研究が始まったばかりで、せいぜい 10 年程度しか蓄積がなく、他国に比べて大きく出遅れている。 |
| | 応用研究・開発 | ◎ | ↗ | ・Tianhe シリーズだけでなく、米国技術を買った上で追加開発にて独自開発した Shenwei など、大規模スーパーコンピューターを構築する力は日米に匹敵する。比較して大規模アプリケーション技術はまだ弱いが、最適化・運用技術は急速に向上しつつある。スーパーコンピューター TOP500 でダントツの首位をキープした Sunway TaihuLight のプロセッサーも接続回路もすべて中国純製であった。このプロセッサーは、中国の国家並列計算機工程技術研究中心（NRCPC）が独自開発した 260 コア高性能プロセッサー「SW26010」である。<br>・2015 年 7 月、中国科学院と Alibaba は、「中国科学院―Alibaba 量子計算実験室」を共同設立し、量子情報科学研究を開始した。 |
| 韓国 | 基礎研究 | × | → | ・スーパーコンピューターの基礎研究は見るべきものがほとんどない。<br>・韓国科学技術院（KAIST）が、アナログ回路を使うアプローチを採用して、深層学習向け AI チップを開発している。<br>・米国・欧州で学んだ研究者を活発にリクルートし、量子コンピューティングに関する基礎研究の裾野を広げている。 |
| | 応用研究・開発 | △ | → | ・スーパーコンピューターの計算能力は進展しておらず、応用技術の進展もない。<br>・韓国サムスン電子は、英国 AI チップ開発スタートアップ企業 Graphcore に 34 億ドルを出資した。<br>・量子コンピューティングに関する研究開発が見受けられない。 |

（注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
（注２） 現状 ※わが国の現状を基準にした評価ではなく、CRDS の調査・見解による評価である。
◎ 特に顕著な活動・成果が見えている、 ○ 顕著な活動・成果が見えている
△ 顕著な活動・成果が見えていない、 × 活動・成果がほとんど見えていない
（注３） トレンド
↗：上昇傾向、 →：現状維持、 ↘：下降傾向



## (8) 参考文献

1）JST-CRDS, 2015 年度版俯瞰報告書（情報科学技術分野）
https://www.jst.go.jp/crds/pdf/2015/FR/CRDS-FY2015-FR-04.pdf（閲覧日 2017-2-8）
2）TOP 500 Supercomputer Sites: http;//www.top500.org/（閲覧日 2017-2-8）

3）The Graph 500 List: http://www.graph500.org/（閲覧日 2017-2-8）
4）The Green500 List: http://www.green500.org/（閲覧日 2017-2-8）
5）Tiffany Trader, “Nvidia Sees Bright Future for AI Supercomputing”（Nov. 23, 2016）
https://www.hpcwire.com/2016/11/23/nvidia-bright-future-ai-supercomputing/
（閲覧日 2017-2-8）
6）Amazon AWS HPC instance: http://aws.amazon.com/hpc/（閲覧日 2017-2-8）
7）Penguin Computing: http://www.penguincomputing.com/（閲覧日 2017-2-8）
8）甘利俊一（監修), 深井朋樹（編集), 脳の計算論（東京大学出版会 , 2009）.
9）齊藤元章 , エクサスケールの衝撃（PHP 研究所 , 2014）.
10）総合科学技術・イノベーション会議 , “スーパーコンピュータに関する協力取極”
（評価専門調査会会議資料）（閲覧日 2017-2-8）
http://www8.cao.go.jp/cstp/tyousakai/hyouka/haihu109/siryo1-2_part15.pdf
11）T. Kadowaki and H. Nishimori, “Quantum annealing in the transverse Ising model,” Physical Review E 58, 5355（1998）.
12）西森秀稔 , 大関真之 , 量子コンピュータが人工知能を加速する（日経 BP, 2016）.
13）J. Condliffe, “Can a Powerful New Quantum Computer Convince the Skeptics?” MIT Technology Review（January 24, 2017）:
https://www.technologyreview.com/s/603438/（閲覧日 2017-2-8）
14）Y. Nakamura, Y. A. Pashkin, and J. S. Tsai, “Coherent control of macroscopic quantum states in a single-Cooper-pair box” Nature 398, 786（1999）.
15）JST-ERATO 後藤磁束量子情報プロジェクト（総括責任者・後藤英一）:
https://www.jst.go.jp/erato/research_area/completed/gjrj_PJ.html（閲覧日 2017-2-8）
16）K. Takata, S. Utsunomiya, and Y. Yamamoto, “Transient time of an Ising machine based on injection-locked laser network,” New Journal of Physics 14, 013052（2012）.
17）理化学研究所 計算科学研究機構 , “フラッグシップ 2020 プロジェクト”
http://www.aics.riken.jp/jp/overview/post-kcomputer/（閲覧日 2017-2-8）
18）National Institute of Advanced Industrial Science and Technology（AIST）, “AI Bridging Cloud Infrastructure（ABCI）”
http://www.itri.aist.go.jp/events/sc2016/pdf/P06-ABCI.pdf（閲覧日 2017-2-8）
19）The National Science Foundation（NSF）, “The National Strategic Computing Initiative（NSCI）”:
https://nsf.gov/about/（閲覧日 2017-2-8）
20）Human Brain Project: http://www.humanbrainproject.eu/（閲覧日 2017-2-8）
21）JST Science Portal China, “第 13 次五カ年計画、中国の技術革新計画が明らかに”
http://www.spc.jst.go.jp/news/160704/topic_4_03.html（閲覧日 2017-2-8）
22）浅川直輝 , “東芝の「脳を模倣した」AI チップ、フラッシュ派生技術で省電力化 ,
日経コンピュータ ITpro ブログ（2017 年 2 月 1 日）
http://itpro.nikkeibp.co.jp/atcl/column/17/013000015/013100003/（閲覧日 2017-2-8）

23）Lawrence Livermore National Laboratory（LLNL）news（March 29, 2016）, “Lawrence Livermore and IBM collaborate to build new brain-inspired supercomputer”
https://www.llnl.gov/news/lawrence-livermore-and-ibm-collaborate-build-new-brain-inspired-supercomputer（閲覧日 2017-2-8）
24）日本電気プレスリリース（2016 年 9 月 2 日）, “NEC と東京大学、日本の競争力強化に向け戦略的パートナーシップに基づく総合的な産学協創を開始”
http://jpn.nec.com/press/201609/20160902_01.html（閲覧日 2017-2-8）
25）Wired.com（October 28, 2016）, “How AI Is Shaking Up the Chip Market”
https://www.wired.com/2016/10/ai-changing-market-computer-chips/（閲覧日 2017-2-8）
26）Microsoft Research, Station Q（Worldwide consortium for the advancement of topological quantum computation）:
https://stationq.microsoft.com/（閲覧日 2017-2-8）
27）Microsoft Official Blog, “Microsoft doubles down on quantum computing bet”
https://blogs.microsoft.com/next/2016/11/20/microsoft-doubles-quantum-computing-bet/（閲覧日 2017-2-8）
28）富士通研究所ニュースリリース（2016 年 10 月 20 日）, “量子コンピュータを実用性で超える新アーキテクチャーを開発”：
http://pr.fujitsu.com/jp/news/2016/10/20-1.html（閲覧日 2017-2-8）
29）東芝ニュースリリース（2016 年 11 月 7 日）, “ディープラーニング（深層学習）を低消費電力で実現する脳型プロセッサーを開発”
https://toshiba.semicon-storage.com/jp/company/news/news-topics/2016/11/micro-20161107-1.html（閲覧日 2017-2-8）