ビッグデータにおける学術と研究の動向と方向

徳山　豪

ビッグデータはいわゆるバズワードであるが，その社会的な認知度や影響力は非常に強い．筆者は理論計算機科学を専門とするが，データマイニングへの応用を先駆的に行った関係で，ビッグデータ関係のさまざまなプロジェクトに関与して（させられて？）いる．本解説では，それらの俯瞰を行うとともに，OR に関係する数理科学や計算理論などとの関連を述べ，今後日本が世界の中でビッグデータ研究でどのような先導性をもつべきかという課題について，議論の話題を提供したいと思う．

キーワード：ビッグデータ，学術動向，プロジェクト紹介

1. AI とビッグデータ

1.1 ビッグデータに対する社会の期待

AI（人工知能）とビッグデータは，世の中のほとんどの人が知っていて興味をもっている事項である．その一般社会における注目度は半端ではない．その証拠に，安倍晋三首相は第 197 回国会の所信表明演説（10 月 24 日）において，冒頭，本庶教授のノーベル賞受賞への祝辞の後で，このように述べている．

> 世界は，今，かつてないスピードで，変化しています．この，わずか五年余りの間に，人工知能は急速な進歩を遂げ，様々な分野で人間の能力を凌駕しようとしています．膨大なデジタルデータが，世界を瞬時に駆け巡り，全く新しい価値を生み出す時代となりました．次の五年，いや三年もあれば，世界は，私たちが今想像もできない進化を遂げるに違いない．そうした時代にあって，私たちもまた，これまでの「常識」を打ち破らなければなりません．私たち自身の手で，今こそ，新しい日本の国創りをスタートする時であります．強い日本．それを創るのは，他の誰でもありません．私たち自身です．激動する世界を，そのど真ん中でリードする日本を創り上げる．次の三年間，私はその先頭に立つ決意です．私たちの子や孫の世代のために，希望にあふれ，誇りある日本を，皆さん，共に，切り拓いていこうではありませんか．

異例の力の入れ方であり，さらに同演説の後半では，IoT，ロボット，人工知能，ビッグデータの活用を阻む規制や制度の大胆な改革が宣言される．

この安倍首相の言葉が，現在一般市民の感じている感覚に近いのではないだろうか．データは価値を生む．人工知能はすぐにでも人間を凌駕する．AI やビッグデータなどの情報科学技術分野で世界最先端に立たないと日本の将来は暗いので，活性化しないといけない．

人工知能が人間を凌駕するというのは，囲碁や将棋において AI ソフトが専門棋士を凌駕し，またクイズ Jeopardy で IBM のワトソン応答システムが人間チャンピオンを破る状況で，安倍首相の言葉のように，分野によってはすでに事実である．データが価値を生んでいるのは，アマゾンやグーグルなどがデータ利活用を主な動力源にして超巨大企業になっていることを見れば，これも疑いようのない事実である．われわれ研究者には，「データを自由自在に扱って，計り知れない価値を生む」ような技術が期待されている．

筆者も「集まってくるデータの活用のアイデアがほしい」とか「ビッグデータ・AI でプロジェクトに参画して何か貢献してほしい」などと他分野の研究者などに言われたりする．これには誤解があって，まず筆者は理論計算機科学をメインとする研究者であり，データ解析の実務家ではない．それは置いておいても，「なんでもいいからデータなら解析すれば何か出る」というのは大きな間違いである．

こういう誤解を含む言葉を耳にすると，1985 年に公開された映画「バック・トゥ・ザ・フューチャー」のラストシーンを思い出す．めでたく過去から現在に帰ってきた後，ドク博士が「ちょっと 30 年後（2015 年）に行ってくる」といって出かけ，慌てた様子で戻ってきて，「君たちの子孫に問題が起きたから一緒に未来に行こう」とマーティーとロレインに告げるのだが，「燃料を補給しなくちゃ」となる．タイムマシンの燃料はプルトニウムだが，ドクは，「30 年後はこうなのさ」と

とくやま　たけし
関西学院大学理工学部
tokuyama@dais.is.tohoku.ac.jp

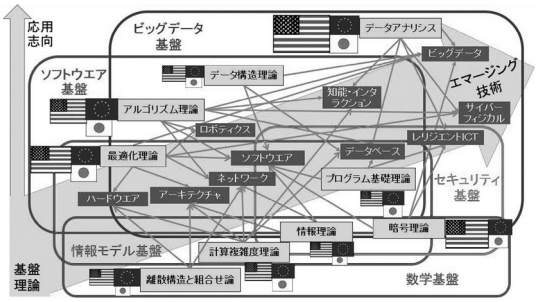

図 1　計算理論と IT の関連の俯瞰図

言って，その辺のゴミをタイムマシンに流し込む．

ゴミリサイクルはここまで実用化できていないが，ビッグデータに関しては，「使えないとゴミでしかないデータでも，機械学習に放り込めば，何らかの価値を生んでくれる」というような感覚のほうが企業経営者などにも現実に多いのである．

1.2　OR と AI・ビッグデータ

一方で，世界がかつてないスピードで変化し，その原動力が情報技術であることは疑いがない．図 1 は，筆者が 2014 年初頭に作成した，計算理論の諸分野と IT 分野の関係を示した俯瞰図であるが，ここにはビッグデータはあるが AI の記述がない．深層学習のブレーク前であり，AI がここまで急速に普及するとは予想できなかったのである．そして，今やビッグデータと AI 分野において，アメリカに拮抗する勢力になりつつある中国も記載されていない．たった 5 年でここまで大きい変化があり，その社会へのインパクトは巨大である．しかし，この俯瞰図にあるように，ビッグデータには計算理論や IT 技術のほとんどが必要とされ，オペレーションズ・リサーチ (OR) が長年築いてきた数理や最適化を基盤としたデータ解析や分析手法は必須の要件である．

AI やビッグデータに最も期待されているのは，「正しい選択をするための手段」としての役割である．コロンビア大学のシーナ・アイエンガー教授は選択の科学の権威であり，数年前に NHK コロンビア白熱教室で素晴らしい講義をされていた．盲目のインド系移民である彼女は，「正しい選択は情報から得る」ということを実際にデータからの情報取得を実演して示していたが，現代では，ビッグデータから情報を取り，そこから AI を使って選択肢を絞って選択をするというのは当たり前になりつつある．IBM ワトソンや AlphaGO が注目を浴びたのは，それらが，「正しい選択をする」ことに対して，ドメインを絞れば人間よりはるかに正確であることを実証したからであり，これを実社会における選択に活用できたら素晴らしいと皆が色めきだったのである．特に，AlphaGo の進化型である AlphaGo Zero では，人間が作った学習データは全く使わないので，「人間の考えのもつ愚かさ」をさえ排除できるような期待さえ抱かせる．

もちろん，正しい選択は昔からの人間にとっての最重要課題であり，そのための情報収集や情報活用は世界の歴史を変えてきた．一例として織田信長は桶狭間の戦いにおいて，敵軍の進軍ルートと領内の地理情報，天候情報を用いてワンチャンスを捉える奇跡的な選択を行う．さらに浅井軍の寝返りを，妹のお市の方から送られた「両端を縛られた小豆の袋」というだけの情報から気づいて，即座に逃げ出す決断をする．まさに歴史を変える「情報による選択」の天才である．こんな天才的な判断や選択を AI が自動的に行い，万人が享受できるのならば，これは素晴らしいことである．

一方で，アイエンガー教授は正しい選択には普遍性はないと述べている．環境や宗教などの個人の背景によって正しい選択は変わるのであり，臨機応変な情報活用は一般的なデータから得るのは容易なことではなく，書物や資料からの学習だけではとてもできない．

三国志を見ても，学識に頼って状況を見誤った選択ミスは，「泣いて馬謖を切る」というような大きな影響を与える．つまり，IBM ワトソン や AlphaGo のシステムを実社会応用しても，臨機応変な判断ができるとは言えない．囲碁のような単純なルールの下で明確な目的がある場合や，ロボットの動作や自動運転のように瞬時の行動選択が必要な場合は全自動化が望ましいが，通常の人間生活での活用では，AI が提示した選択肢を人間が活用する，決定支援システムとして用いるのが現状では妥当と考えられる方針である．実際，ワトソン は（IBM の創始者の名前でもあるのだが）補佐役で，ホームズではないのである．

数理と計算機の力を用いて決定支援をするための学術基盤が OR であったはずである．OR においては，基本的には正しい選択をするためにデータや情報を整理し，人間の判断を補助するための道具立てをする．そのために統計解析や数理計画法やゲーム理論などの数理的な手法が基盤となる．

AI やビッグデータでも深層学習や強化学習，カーネル学習などの数理的な手法を基盤にしているのだが，数理的に容易に定式化できない問題，たとえば画像の認識や自然言語の理解，囲碁の局面判断などを学習を用いて自動的にモデル化することで，従来の OR での数理の範疇を超えたデータ解析を行える．

実際，「この画像は犬に似ているか？」，あるいは「囲碁でこの局面はどのくらい黒が優勢か？」というような問題を数理的に定式化するのは困難であり，例え数理的に定式化できても非常に複雑になる．また，機械翻訳やツイート解析などでは，人間が長い時間で地域ごとに培ってきた言語や習慣，主義主張などが絡むので，これも素直に数理定式化できるものではない．

このような差異はあるが，AI やビッグデータにおいて世界を先導する研究を行うために，OR で培った研究力を活かすことはとても重要である．

2. 世界を先導する日本のビッグデータ研究

世界を先導するには，既存技法の改良と平凡なデータを用いた応用研究だけではだめなことは明らかである．中国や米国の研究者の数と研究費を考えると，日本が対抗するためにはオリジナルな技法の開拓とその実用化は必須である．残念なことに，脳神経回路網に関しては，パーセプトロンや甘利の理論研究など日本人の貢献が大きいのにかかわらず，深層学習の実証と実用化はトロント大学の研究者によって行われた．

その一方で，理論を基盤として開発された独自の技法が，日本のビッグデータ研究を支えている．図 2 は，計算理論関係のプロジェクトで，ビッグデータ技法につながるもの（筆者が関わったもの中心で，プロジェクト名は略称）の時間的経緯と関係を簡単に書いたものである．これらのプロジェクトに言及しながら，日本のもつ最先端技術をいくつか紹介しよう．

2.1 劣モジュラ最適化

ビッグデータにおいて重要な作業はデータの類別，すなわちクラスタリングである．これは数理的には集合の分割であり，さまざまな数理的なモデルや手法が知られている．集合分割において，最小記述長やグラフのカット，相互情報量，さらには経済的な指標などを目的関数にすると，数理計画問題としてさまざまな定式化が生じるが，それらを包括し，効率よく（専門的には多項式時間で）解けるぎりぎりの定式化として劣モジュラ最適化が深く研究されている．

数学的に書くと抽象的なのだが，集合 U の部分集合全体の族の上の実数値関数 F を考える．このとき，任意の $A \subset B$ と任意の U の元 x に対して

$$F(A \cup \{x\}) - F(A) \geq F(B \cup \{x\}) - F(B)$$

が成立するとき，F を劣モジュラ関数と呼ぶ．直観的には，ある要素 $\{x\}$ を加えて関数値を上げるには，元の集合が小さいほうが効果が高いということで，人間の実生活（人員配置や経済効果など）での現象を自然に数理化していると思える．実際，学習データを増やすことによる学習効果も劣モジュラ性をもつ．

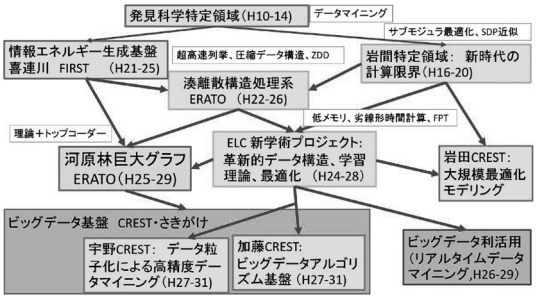

図 2 計算理論のプロジェクトの流れ

劣モジュラ関数は，Edmonds や Lovász らとともに，伊理，藤重などによって 1970 年代から研究され，その後，図 2 にある特定領域研究「新時代の計算限界」や ELC 特定領域研究などで研究され，室田，岩田などが中心になって発展させて，日本がハンガリーなどとともに研究の一大拠点であり続けている．

初期にはマトロイドやネットワークフローの抽象的な一般化であり，応用例も少なく，純粋な理論研究の観があった．それが 10 年少し前から画像処理におけるグラフカットと呼ばれる切り出しへの利用で注目され，今では岩田による「大規模複雑システムの最適モデリング手法の構築」CREST プロジェクトに代表されるように，機械学習においてなくてはならない手法となっている．そして，日本での研究の先進性を強みに，河原林巨大グラフ ERATO [1] で育った多くの若手研究者が，劣モジュラ最適化を用いた機械学習でのさまざまな最先端成果をトップコンファレンスで発表し，わが国の研究の強みになっている [2].

2.2 簡潔データ構造と圧縮列挙

ビッグデータという言葉にはいろいろ誤解があり，必ずしもデータサイズが大きいということではない．一方で巨大なデータの取り扱いが重要な事項であることは確かである．多くの場合，そのままでは巨大なデータでも工夫すれば小さく格納できる．たとえば，30 億の遺伝子対をもつヒトゲノムを 1 万人分もつデータは，単純には 30 兆の遺伝子対，つまり 60 兆ビットのデータ量であるが，標準的なヒトゲノムとの差異のみを格納すると，100 分の 1 以下にデータ量は減らせる．さらに，Sadakane and Grossi によって開発された簡潔データ構造 (Succinct Data Structure [3]) を利用すると，検索のための索引構造を含めて通常のパソコンのメモリに格納して，自由に高速検索することができる [4].

データに離散的な構造，あるいは隠れた生成規則がある場合に，さらに超絶的なデータ圧縮を行う可能性をもつデータ構造として，湊によって提案されたZDD (Zero Supressed Decision Diagram) がある [5]．これはVLSIの設計などで論理関数を表現するときに広く使われるBDD (Binary Decision Diagram) の亜種と考えられるが，特に集合族を表現するときにはBDDよりはるかに効果的なことが多い．たとえば組合せ最適化における解集合などは典型的な例であり，n 行 n 列のグリッドグラフのハミルトンパスの集合族ならば，全体集合は辺集合であり $2n(n-1)$ の辺をもち，ハミルトンパスは，そのうちの n^2-1 本の辺からなる集合でパスとなっているものたちである．この集合族のサイズは巨大であり，$n=16$ で 2×10^{48} 程度で，とても列挙して記憶できるサイズではない（ヨタバイトはたった 10^{24} バイトである）．ZDDはこれを144,759,636頂点の決定ダイアグラムとして表示して圧縮列挙し，パスの総数をはじめとして，さまざまな最適化の応答や集合演算を，ダイアグラムのサイズに比例した高速計算で答えてくれる．実に 10^{40} 倍を超える圧縮率である．

湊が統括したERATO湊離散構造処理系プロジェクト [6, 7] では，ZDDを基盤にして，圧縮データ列挙技術の発展とビッグデータ応用を幅広く研究して若手研究者を育成し，日本におけるビッグデータ研究の大きな強みとなっている．たとえばデータマイニングにおけるアイテム集合列挙は，データマイニングの発祥であるAggrawalらのAprioriアルゴリズム以来の重要な問題であるが，属性数の多いデータだと，列挙する集合族の候補が巨大になり，それをメモリに載せてさまざまな集合演算を効率的に行うにはZDDが欠かせない．

また，スマートグリッドや電力網などの制御や，化学物質の構造列挙など，ビッグデータ処理での活用の場は広い．さらに，組合せを列挙することにより統計や確率計算を精密に行うことができ，実験の再現性を測る統計指標である p 値の正確な計算は，津田によるCRESTプロジェクト「離散構造統計学の創出と癌科学への展開」の基盤技術となっている．

2.3 性質検査による超高速アルゴリズム

データのサイズ n がいくら大きくても，n に依存しない定数時間で計算できればこわくない．

統計量の計算は定数時間で行うことができる．すなわち，大数の法則によって，データサイズに依存せずに，ある程度大きいサンプル上での検定を行えば，平均値，偏差値などの統計量は高い精度で求まり，データ全体を俯瞰したモデルが作れる．たとえば身長の平均値がサンプルで170 cmなら，高い確率で全体の平均値は168 cm以上172 cm以下であり，180 cm以上の人は全体の1割以下であるという具合である．

また，計算論的機械学習ではVC次元という尺度を用いて概念の複雑度を測り，定数VC次元をもつ概念の学習は定数個のサンプルデータから学習できる．

このような思想をさらに拡張し，グラフなどでのさまざまな計算問題を考えてみよう．たとえば，「n 頂点のグラフ G が平面グラフですか？」という設問に答えるのは，定数時間では直感的には到底不可能に思えるのだが，少し定式化を変え，次の二択問題を考える，

(1) G が性質Aを満たす
(2) G から ϵn 個の頂点と，接続する辺を取り除いても性質Aを満たさない（ϵ は0.01のような定数）

これを性質検査 (Property Testing) と呼ぶ．ちなみに，(1) でも (2) でもない場合は，どちらを答えても構わない．性質Aが「平面グラフである」として言い換えると，(2) が答えられたら，グラフは平面的ではない．(1) が答えられたら，グラフから ϵn 個以下の頂点を取り除くと平面的にできるのである．

ここで，定数時間で検査するとは，n に依存せず，ϵ にのみ依存する計算時間で上記の二択に答えられるということを意味し，サンプルの選び方に依存するので，確率2/3以上で検査結果が正しいことを保証するのである．

グラフにおける性質検査は限られた性質以外は困難と思われていたが，伊藤は文献 [8] において，「グラフが階層的スケールフリー性をもつとき，任意のグラフ性質（グラフ同型で不変な性質）は定数時間で性質検査できる」ことを証明した．

Webや社会ネットワークは階層的スケールフリー性をもつので，上記の成果は強力であり，サンプリングによって統計のように全体を俯瞰するモデルの構築が理論的には可能なのである．性質検査は実装して実用化するにはさらなる改良が必要ではある．しかしながら，将来的にビッグデータの処理はこのような高度な計算理論的な手法で，データサイズに依存せずに超高速に行えるという期待ができる．性質検査を含め，データ全体を読むのよりも速い時間での計算，すなわち劣線形時間計算は，ビッグデータ研究では大きな課題であり，加藤CRESTプロジェクト「ビッグデータ時代に向けた革新的アルゴリズム基盤」の中心テーマである．これも日本が強みをもつ研究分野と言えよう．

2.4 情報統計力学と量子アニーリング

物理学においては 1 モルで約 6×10^{23} 個の粒子をもつ系を考えねばならないが，これを離散系として見るとあまりに巨大であり，統計力学的に捉えることは常識である．ノーベル物理学賞受賞者の Anderson による “More is different” という標語があるが，これがビッグデータにおいても通用するという思想で，物理学で培われた手法を適用することが有力である．

日本においては，平成 14–17 年度の特定領域研究「確率的情報処理への統計力学的アプローチ」（代表：田中和之）と，平成 18–21 年度の特定領域研究「情報統計力学の深化と展開」（代表：樺島祥介）での研究重点化により，独自の強みをもつ研究分野であり，機械学習におけるスパースモデリングに展開され，また，上述の加藤 CREST でも統計物理学的な視点によるデータモデリングが大きなテーマになっている．

さらに，西森により提唱された量子アニーリングは，統計力学的なイジングモデルでのエネルギー関数の最適化を量子計算によって超高速に解くシステムであり，D-Wave 社が量子アニーリングマシンを実用化して，大きな注目を集めている．NP 問題はすべて理論的にはイジングモデルのエネルギー関数の最適化に還元できるため，将来的には，現在困難と思われている問題のほとんどが量子アニーリングで解けるかもしれない．情報統計力学や量子アニーリングにおける日本の研究力は高く，文献 [9, 10] などをお読みいただきたい．

3. ビッグデータのプロジェクト

3.1 河原林 ERATO におけるチーム編成

前述の河原林ビッググラフ ERATO では，理論計算機科学で開発されたさまざまな先端的なアルゴリズム技法をビッグデータ分野へ導入し，また逆にビッグデータ利活用から発生した問題を理論的に解明することで研究の活性化を行い，図 3 にあるように，4 年間でのビッグデータ関連のトップコンファレンスでの論文で，それ以前に比べて 60 本の増加，比率にすると 13 倍を達成している．このほかに理論系のトップコンファレンスでも 40 本ほどの論文を発表している．

この成功には，先端的なアルゴリズム技法の活用に加えて，研究チームの今までにない構成が大きな原動力になっている．すなわち，従来の分野に閉じた研究体制ではなく，理論研究者でプログラミングもできる若手人材とトップコーダー（主にプログラミングコンテストの常連である学生を研究員として雇用）を混成したチームでの研究と論文作成を行い，さらに現場のデータ応用に詳しい企業研究者を含めた討論やサーベイを常時行うことで問題の探索と研究の動向を常に最新にするシステムを導入し，それを PI が統括し，討論に必ず参画しリードするシステムである．国立情報学研究所の環境とデータホルダの立場も活かしており，この研究体制は学ぶべきものだと考えている．

分野	論文採択数 [ERATO前]		論文採択数 [ERATO中]
データベース (SIGMOD, VLDB, ICDE)	0	→	**8**
機械学習等 (ICML, NIPS, AAAI, IJCAI)	4	→	**41**
データマイニング (KDD, WWW)	1	→	**13**
その他 (LICS, INFOCOM,ACL)	0	→	**3**
合計	5	→	**65**

図 3　河原林 ERATO の論文発表件数

3.2 データ利活用の経験と苦労

筆者自身のデータ利活用プロジェクトの経験としては，IBM 東京基礎研究所在籍中の 1996〜1997 年に IPA の創造的ソフトウエア事業から受託した「知識獲得機能付き関係データベースの開発」に遡る．このときのチームは計算理論の専門家である筆者と，データマイニングの総本山であったスタンフォード大学で学位を取った森下真一（現東京大学教授），トップレベルのプログラミング能力と理論基盤を有する若手研究者である福田剛志（現 IBM 東京基礎研究所所長），森本康彦（現広島大学教授）で，上述の河原林 ERATO プロジェクトでのチーム編成を先取りする素晴らしいものであり，データマイニングの発祥直後であるために新規手法を開拓しながら実用システムを構築する先導的な研究成果を挙げ，非常に実りのあるものであった．

一方で，近年は技法的にビッグデータ研究は成熟しつつあり，既存手法の組合せによって実用システムを構築するほうが独自手法開拓に勝ることも多い．したがって，最先端の研究と利活用プロジェクトへの貢献を，うまくバランスを取りながら統括していく必要がある．そして，現実のデータ解析における経験としては，非常に大きな労力をデータの整理や作成に費やす必要があるという現実がある．

実社会ビッグデータ利活用のためのデータ統合・解析技術の研究開発 [11] では，地図情報データ，交通データ，ツイートデータの 3 種のデータの融合解析を担当したが，ここで最も多大な労力を割いたのは，ツイートに書かれた地名や施設の候補を実際の地図上の候補にマップするためのコーパスの作成である．たとえば「八幡宮」というときには，文脈や関連情報によって場所を特定する必要があり，江ノ島と関連すれば鶴岡八

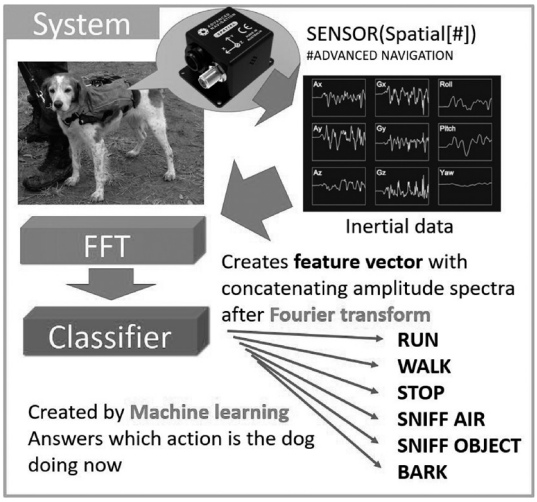

図 4　災害救助犬のセンサデータ解析

幡宮であろう．Wikipedia などにもリンクするが，かなりの部分を新しく構築せざるを得なかった．

これは企業などでのビッグデータ利活用でも同様であり，既存や公共のビッグデータだけを使うのではなかなかよいソリューションには結び付かない．

もう一つ苦労した例として，災害救助犬にセンサーを配備する「サイバー救助犬」のデータ解析がある（図 4）．災害救助犬は瓦礫の下や建物内部などの目視できない場所をも探索し，その挙動をセンサーデータから推測することが求められている．サイバー救助犬には数百グラムの装備しか装着できず，リアルタイムでは胴体に着けた加速度センサーの波形データしか送信できない．そこで，加速度センサーのデータから犬の挙動を推測することが課題であり，まずは基本として「歩く」「走る」「止まる」「吠える」「空中のにおいを嗅ぐ」「物体のにおいを嗅ぐ」などの行動を推測したい．特に物体のにおいを嗅ぐ行動は，被災者やその遺留物の発見の可能性があるので重要である．

救助犬は 2 頭で，使える時間も機会も限られており，データ解析目的の新規実験は行えない．そのため，データは少なく，目的に即して収集されてはいない．

「歩く」「走る」「止まる」は，加速度センサーデータは積分すれば犬の軌跡がわかるため，通常の最適化で分類可能であり，95％程度の確実度で判断できる．

一方で，吠える，においを嗅ぐなどは，人間が見ても加速度センサーデータからは判断できない．学習に必要な教師付きデータを作成するためには，訓練全体を捉えた定点カメラの動画が頼りである．インターネット上などでの犬画像とキャプションを用いて典型的な犬の行動と画像の対応を学習することも考えられるが，においを嗅ぐなどの行動では困難である．やむを得ず，画像を学生が見てキャプションを付ける作業を行った．多大な手間をかけてデータを準備して機械学習で分類を行って，「物体のにおいを嗅ぐ」という動作の推測の確実度は 75％である．

このように，データの準備と整理が重要であり，実データの解析を行うには覚悟が必要である．

3.3　ビックデータ基盤「さきがけ」から

ビッグデータ基盤（喜連川総括，正式名は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）においては，アドバイザの立場でそれぞれのプロジェクトを見させていただいた．CREST はすでにいくつか紹介しているので，「さきがけ」において，冒頭の安倍首相の演説に答えるような，データを活かして新しい価値を生み出す成果を二つ紹介する．

3.3.1　人腸内環境ビッグデータ

ビッグデータ基盤では「データさきがけ」を設定している．これはビッグデータ解析技術の新規性をもたなくても，重要なビッグデータを構築する立場にある研究者のプロジェクト提案を受け入れるものである．本プロジェクトは「データさきがけ」の一つであり，研究者は東京工業大学山田拓司准教授である．

ヒトの腸内のバクテリア分布は非常に巨大なデータであり，これが健康や疾病に関する情報をもっている．このデータに診療データや生活データなどを統合したデータベースを構築し，そこから新しい医療の構築を目指す．慶応大学，国立がんセンター，東京工業大学の 3 機関の連携によって，現状で 2,000 名，将来的には 5,000 名の患者に対して 2,738 項目をもつデータベースを構築する．特筆すべきなのはがん患者のデータが多数を占めており，疾病と腸内環境の関係に対する多くの新発見が与えられた．腸内環境によるがんマーカーはその一つであり，これによって，検便によって早期のがんの発見やがんのステージの推測が可能である．また，創薬に関する知見も導かれ，山田氏の主導で「日本人腸内環境とその産業応用プラットフォーム」というコンソーシアムが立ち上がり，産業化への道筋が開かれた．

3.3.2　タイムドメイン宇宙観測用動画データの高速逐次処理法の開発

東京大学酒向重行助教の提案プロジェクトである．

酒向氏は，Tomoe Gozen という名前の新しい宇宙観測カメラを構築（さきがけとは別資金）しており，そこで得られる一晩で 30 テラバイトの動画データの解

析で，宇宙全体での事象，たとえば超新星や中性子星の変化や，見えないほどかすかな流星などを常時監視する．データは保持できないので数時間で消去されるが，特徴のある変化を抽出・圧縮して共同研究機関に送り，ほかの観測データと突き合わせて宇宙の事象を探る．解析の手法自体は既存手法をベースに工夫されているものだが，動画データ解析とデータ圧縮の威力が存分に発揮され，すでに宇宙ニュートリノの放射元天体同定や中性子星合体からの重力波天体が放つ光の初観測，はやぶさ 2 のスイングバイの観測などの成果が新聞紙上をにぎわし，社会からの注目を得ている．

4. まとめ

社会にインパクトのあるビッグデータ研究を行おうと思うと，「さきがけ」の例にあるように，自ら特徴のあるデータを収集し，構造化することが重要であり，そのために非常に大きな労力が必要になる．自分で収集できない場合は，高いデータ解析技術を武器に，データをもつ研究者とタッグを組むべきである．

例えて言えば歴史に残るデータ解析の実社会応用は，コロンブスの新大陸発見のようなものであり，それを支える造船技術や航海術の研究は基礎技術になる．

筆者自身はアルゴリズム理論の研究者として AI やビッグデータを見ており，基礎技術側である．一方で学生はビッグデータや AI の活用を志向するコロンブス側にあるので，積極的に応用研究を勧めている．

研究者個人として無闇にビッグデータや AI というトレンドに乗って，国際的な研究競争の激戦区の真っ只中に飛び込む必要はない．しかしながら，ビッグデータや AI に関する多数の高度な技術者や若手研究者の育成は国の浮沈をかけた使命であり，このトレンドに無関心であってはならない．また，脳が行っている情報処理がどのようなものかは皆目わかっておらず，そのモデルとしても AI とビッグデータの研究は「知の探究」として興味深いことは確かである．

さらに，アルゴリズムという広い目で見ると，アルゴリズムなしに社会生活は成り立たず，アルゴリズムの進歩が社会のさまざまな課題解決に大きな革新をもたらすことは歴史が示す事実である．政府が唱える Society5.0 の目標とする持続可能な社会構築の実現のために不可欠な要素として，ビッグデータの利活用を含めて，アルゴリズムの研究に社会の期待は非常に大きい．

参考文献

[1] ERATO 河原林巨大グラフプロジェクト，http://www.jst.go.jp/erato/kawarabayashi/index.html（2019 年 3 月 1 日閲覧）

[2] 河原吉伸，永野清仁，『劣モジュラ関数と機械学習』，講談社，2015.

[3] K. Sadakane and R. Grossi, “Squeezing succinct data structures into entropy bounds,” In *Proceedings of the 17th Annual ACM-SIAM Symposium on Discrete Algorithm*, pp. 1230–1239, 2006.

[4] D. Arroyuelo, R. Cánovas, G. Navarro and K. Sadakane, “Succinct trees in practice,” In *Proceedings of the Meeting on Algorithm Engineering & Expermiments*, pp. 84–97, 2010.

[5] S. Minato, “Techniques of BDD/ZDD: Brief history and recent activity,” *IEICE Transactions on Information and Systems*, **E96-D**, pp. 1419–1429, 2013.

[6] ERATO 湊離散構造処理系プロジェクト，https://www.jst.go.jp/erato/minato/（2019 年 3 月 1 日閲覧）

[7] 北海道大学大学院情報科学研究科　基盤 (S) 離散構造処理系プロジェクト，https://www-erato.ist.hokudai.ac.jp/（2019 年 3 月 1 日閲覧）

[8] H. Ito, “Every property is testable on a natural class of scale-free multigraphs,” In *Proceedings of 24th Annual European Symposium on Algorithms*, 51:1-51:12, 2016.

[9] 西森秀稔，大関真之，『量子アニーリングの基礎』，共立出版，2018.

[10] 片岡駿，大関真之，安田宗樹，田中和之，『画像処理の統計モデリング―確率的グラフィカルモデルとスパースモデリングからのアプローチ―』，共立出版，2018.

[11] 実社会ビッグデータ利活用のためのデータ統合・解析技術の研究開発，http://bigdata.kde.cs.tsukuba.ac.jp/（2019 年 3 月 1 日閲覧）

