03

情報・システム研究機構ブックレット

SCIENCE REPORT

013-018

データサイエンスでここが変わる。

大学共同利用機関法人
情報・システム研究機構

03

情報・システム研究機構ブックレット

SCIENCE REPORT 013-018

データサイエンスでここが変わる。

大学共同利用機関法人
情報・システム研究機構

『サイエンスリポート』について

『サイエンスリポート（https://sr.rois.ac.jp/）』は、
大学共同利用機関法人情報・システム研究機構が
運営するウェブサイトです。

2016年12月公開以来、
当機構の大学共同利用機関法人としての役割を踏まえ、
自機構のみならず他機構および
広く大学・研究機関等の話題を採り上げ、
広く一般・マスコミの方々へ向けて、
学術の成果や取り組みに関する情報を提供しています。

その中心的なコンテンツとして、ウェブサイトと同名の連載記事
を定期的に（月1回程度）掲載し、
継続的に新しい記事を追加しています。
先端的な取り組みをわかりやすく紹介し、
人々の疑問に科学者コミュニティが答えるような構成で、
学術の役割と活動を広くご紹介しています。

本書は、その記事が身近にお手にとれるよう、
冊子にまとめました。

目次

SCIENCE REPORT

013-018

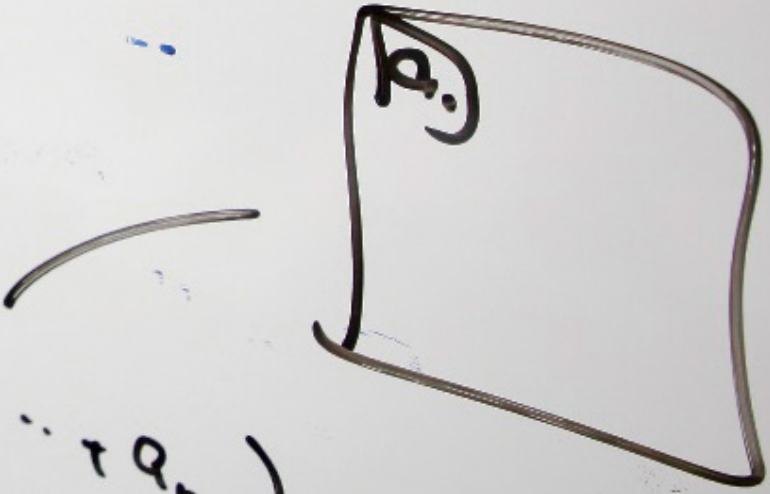
データサイエンスでここが変わる。

$f + g + h > 0$)

interpolant

φ_1

φ_2



$\text{abs}(a_1 + \ldots + a_n) + \text{abs}(\ldots)) + \ldots$

SCIENCE REPORT 013

ものづくりには、データと計算の力が利く。

2017年、情報・システム研究機構は、機構を構成する2研究所から、7月に統計数理研究所 ものづくりデータ科学研究センター、12月に国立情報学研究所システム設計数理国際研究センターというものづくりに関わるセンターをそれぞれ開設した。アメリカ、ドイツをはじめ欧米・アジア諸国でデータサイエンスや人工知能（AI）を採り入れて、ものづくりのあり方を変えようという国家レベルの成長戦略の活発な動きを踏まえたものだ。日本が得意なものづくりの世界に、進化を続けるコンピュータの計算能力やビッグデータの情報力を、どうしたらうまく結びつけることができるか？ ――データサイエンスや新しい数理を開拓し、ものづくりのイノベーションへ向けて産学が協働するフロンティアをお伝えしよう。

答える人

吉田 亮准教授

［統計数理研究所］

よしだ・りょう。2004年総合研究大学院大学統計科学専攻修了、博士（学術）。東京大学医科学研究所ヒトゲノム解析センター特任助教、情報・システム研究機構 統計数理研究所助教を経て、2011年より現職。2017年7月より、同研究所ものづくりデータ科学研究センター センター長。データサイエンスの解析手法の開発に取り組み、神経回路などの生命動態に適用した研究や、新規材料の探索などの物質・材料開発に応用した共同研究等で知られる。国立研究開発法人 物質・材料研究機構 特別研究員を兼任。

蓮尾一郎准教授

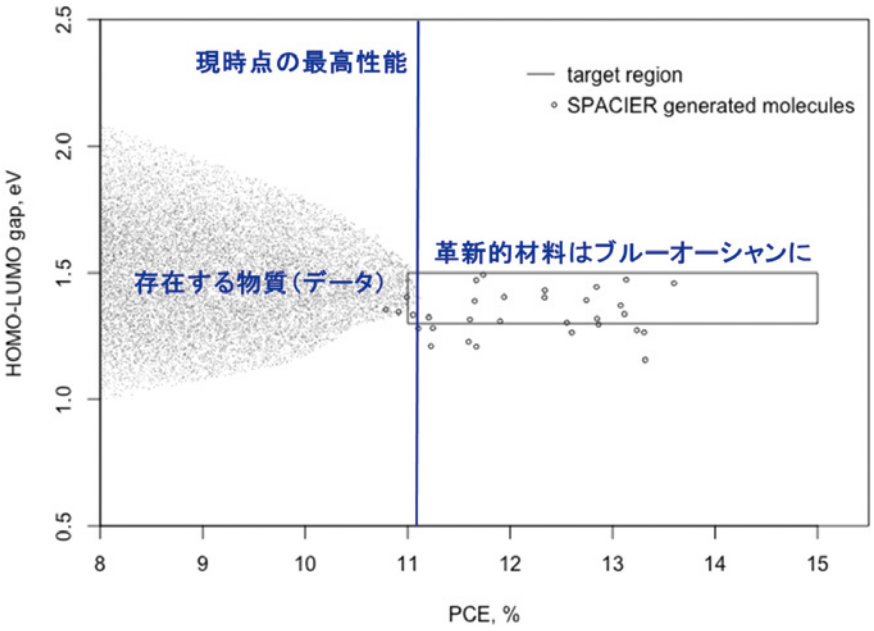
［国立情報学研究所］

はすお・いちろう。2002年東京大学卒業、2008年学術博士（計算機科学、オランダ ナイメーヘン・ラドバウド大学）。東京大学 大学院情報理工学系研究科准教授、京都大学 数理解析研究所客員准教授等を経て、2017年4月より現職、同11月より同研究所システム設計数理国際研究センター センター長。専門は理論計算機科学、特にシステム検証、プログラミング言語理論、物理情報システム、情報科学における数学的構造に関心を持つ。2016年10月よりJST ERATO 蓮尾メタ数理システムデザインプロジェクト研究総括。

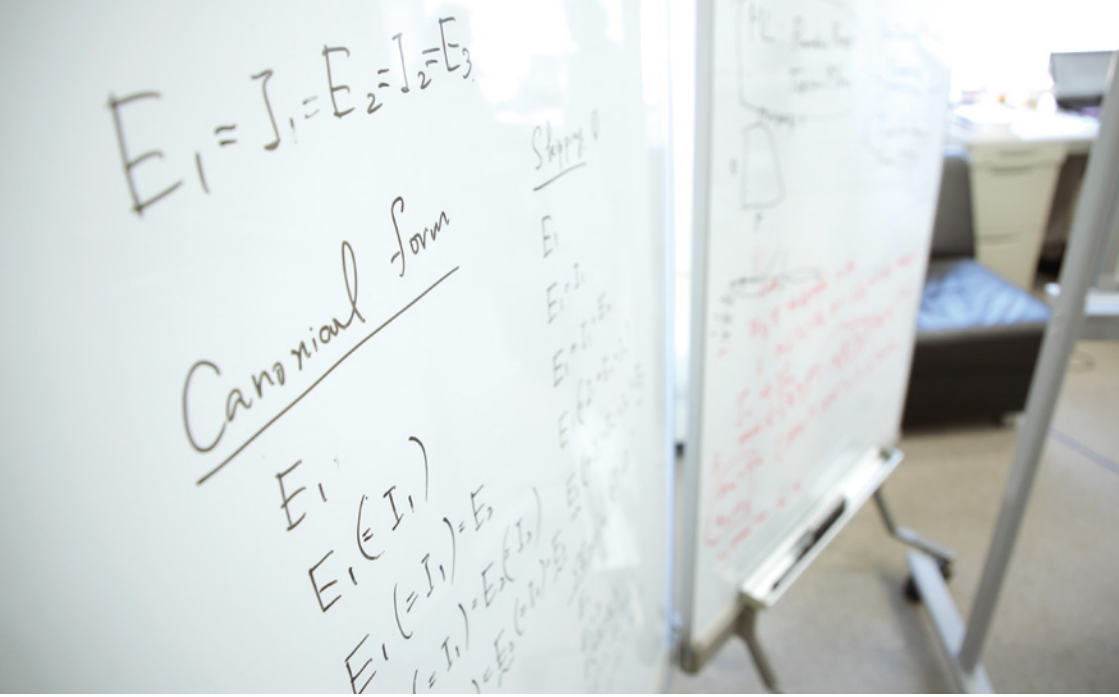
マテリアルズ・インフォマティクスの潮流

統計数理研究所 ものづくりデータ科学研究センター長を務める吉田亮准教授は、学際領域におけるデータサイエンスのスペシャリストだ。これまで生物等のさまざまな複雑な現象の解明に取り組み、近年は物質・材料に研究のターゲットを定める。「マテリアルズ・インフォマティクス」と呼ばれるこの分野は、2011年米国オバマ大統領（当時）が主導した「マテリアルズ・ゲノム・イニシアチブ」以来、データサイエンスや人工知能の技術がもたらす製造業へのインパクトのゆくえが、世界的な注目を集めてきた。

「下のグラフは有機太陽電池の性能を示しています。点の集まりは現在存在する物質を示しており、縦軸に示したパワー変換効率の最高性能は、今のところ約11%です。一方、グラフの右方、既存物質のデータが存在しない未踏領域がわ



上図は、有機太陽電池の新材料の探索結果。グラフ右の黒点のない場所に欲しい機能を示す四角形が描かれている。



れわれのターゲットです。われわれの仕事は、この位置に、現在の最高性能である11%を大きく上回る新しい性能・性質を持つ材料を発見することです」と吉田准教授は言う。「材料科学は今まで、長年の研究の積み重ねによって物質の分布をゆっくりと拡大していましたが、われわれは今、データサイエンスの最先端の技術を駆使することで、この分布を一気に拡大させることができるんですね。これが科学や産業を加速させると考えられる理由です」。

データサイエンスの本質は、「データを集め、データが持つパターンを機械に読み取らせ、認識させること」だと、吉田准教授は言う。「理論は要らないんです。データの中に暗黙に入っているパターンや理論を解析によってあぶり出していく。そして読み解いたパターンの「逆問題」を解くことで、所望の機能に必要な構造を持つ仮想物質を、コンピュータの中に作り出すことができます。これを材料の研究者に提案し、実際に作れれば、新材料の発見につながります」。

「外挿」で、データサイエンスの限界を突破する

一方、データサイエンスは一般に、どのくらい科学の発達に貢献してきたのだろうか？「少なくとも、これまでの私の研究では、せいぜい5〜10%程度ではないでしょうか。歴史を振り返っても、科学技術の頂点はいつの時代も実験か理論であって、データサイエンスがノーベル賞級の大発見に決定的な役割を果たした事例は今のところない」のだそうだ。

吉田准教授は言う。「近年のデータサイエンスの変化の1つは、たとえば機械が画像や音楽を生成したり、アニメのキャラクターをデザインしたりというように、創造的な問題を解く方向へパラダイムシフトしていることです。しかしどんな応用分野であれ、データサイエンスにはデータがある領域しか予測ができないという限界がある。この限界を突破して、人の知性や経験を大きく上回る機械を作るためには、内部にデータを作り出すしくみが必須です」。

そこで、吉田准教授のグループは『SPACIER』というアルゴリズムを開発した。「『SPACIER』は、コンピュータの中に仮想的な実験室を作り、実験計画法とコンピュータシミュレーションで、外挿（現在データがない）領域にデータを作り出します。さらにこの逆問題を解くことによって、機械が新しい予測性能を獲得し、これまであった材料の分布から少し外に出ることができる。このような外挿性の獲得を何度も繰り返すことで、未踏領域に到達するアルゴリズムです。原理的には無限にデータを生産できるため、既存の領域を大きく超えることができます。また工業品の構造設計など、さまざまな分野に適用可能な汎用性も備えています」。

データサイエンスはパートナーシップが重要な科学

機械によるデータの生産といえば、2017年、自分自身と対局することで強くなるAlphaGo Zero（アルファ・ゴ・ゼロ）でも話題になったが、「データを内部で生産しながら機械に「超創造性」を獲得させるという試みは、まだ始まったばかり。技術的な課題も多く、社会実装が本格化するにはもう少し時間がかかるかもしれません」と吉田准教授は言う。「少なくとも現在のものづくりの世界では、コンピュータの中の実験だけでは不十分です。むしろ実際の実験、理論、計算による研究開発と、データサイエンスのアルゴリズムの組み合わせによっていかに実現のシナリオを描くかが本質なんですね」。

中期的にデータサイエンスと実際の実験・理論を循環させ、機械をどんどん賢くして、新しい材料を発見する。データサイエンスによるものづくりへのイン

パクトを「どう説明したら？」の問いに、「言葉ではだめ」と吉田准教授は言った。「実証するんですよ。ここ数年のうちに、産業界の研究開発の最前線に行って、強力なパートナーシップの下で実際にモノができることを社会に発信する」。センターでは今年度6社、来年度以降はより多くの企業と連携して、データサイエンティストの育成を含めた産学協働を計画している。

吉田准教授は、産学連携の「場」のデザインにも積極的だ。プロジェクトでの協働を通じて、企業が優れたデータサイエンティスト人材を獲得できるしくみも構築する。

クルマの自動運転が品質保証の世界を変える

「ちょうどいい時期にプロジェクトが始まったと感じている」と言うのは、ERATO 蓮尾メタ数理システムデザインプロジェクトの研究総括を務める、国立情報学研究所 システム設計数理国際研究センター長の蓮尾一郎准教授である。「自動運転の実用化に代表されるように、品質保証を巡って、今、製造業の方がいわば未知の領域へ踏み出しています。これまで蓄積した経験的なノウハウだけでは十分ではなくなってくるのが目に見えているのだが、産業界でもどうしたらいいかわからないし、学術界でも誰が何を担うべきかはっきりしていない」。

背景にあるのは、ものづくりにおける大きな変化だ。「工業製品はもともと機械だったので、力学に則り、制御理論の成果を使って安定化させるという手法が基礎になっていますが、今やほとんどの製品にコンピュータ制御が入っているんですね。単純な機械の多くは線形システムなので、たとえば入力を2倍にすると効果も2倍になるというように、そのふるまいがある程度予想しやすいのですが、自動車など大規模で複雑なシステムのコンピュータ制御ではそうはいきません」。

「圏論」で、品質保証のガイドを構築する

安全な製品であることをどう説明し、どう保証するのか。そこでプロジェクトが用いるのは、論理学と、蓮尾准教授が特に専門としてきた「圏論（Category Theory）」である。「圏論とは、代数学から生まれた構造記述のための数学の言葉で、ものとものの関係を抽出して抽象化するという使い道があります。同じく関係性を記述するグラフ理論とは異なり、現象そのものを抽象化するのではなくて、現象について語る理論について語り、これを抽象化する「メタ」な言葉であるところに特徴があります」。

ひょっとするとこの圏論、1940年代、数学の代数的構造を元にフランスに起こった構造主義人類学とも関連がありそうな視点である。「クロード・レヴィ＝ストロースの貢献の一つは、さまざまな共同体に共通する構造を抽出したことだと理解しています。情報科学が従来相手にしてきた情報システムと、それに物理系が加わった工業製品のような物理情報システム――この2つは一見違うけれども代数的な目で見れば一緒なところもたくさんある、ということなんです」。むしろこのような圏論の強みに、物理情報システムの品質保証という応用を見つけた点が、プロジェクトを世界的にもチャレンジングな取り組みにしていると言えるだろう。

情報システムにおける「形式手法」を拡張する

一方、情報システムにおいては、以前から無限ループに陥ることはないか、脆弱性はないかといった検証が欠かせない。針の穴ほどのわずかな間違いが、ロケットの打ち上げに不具合や、コンピュータの四則演算にエラーを生んだ例もある

ことから、ソフトウェアやハードウェアが思った通りに動いているかどうかを数学的に検証する「形式手法」が比較的普及しているという。ところが情報システムにおける形式手法を、物理情報システムである実際の工業製品にいかに展開するかは自明でないため、使われている例はまだ多くない。「未開の領域に踏み出すにあたって、数学的な積み上げが、確かなガイドになります。われわれは圏論を使って形式手法を拡張し、問題の本質を数学の言葉で書いたテンプレートのようなものを充実させていきます。そしてこれらをいわば「ひきだし」に蓄えて、現場の技術者の方々と対話しながら具体的な課題に合った証明をつくっていく。またこの過程では、近年発達を遂げている人工知能や機械学習の技術が活用できます。実際の製造の現場においては、今だったらエンジニアの方が知識と長年の経験に基づいて、だいたいこのあたりをテストしておけば大丈夫だろう……と検証しているプロセスの時間と労力が、大いに短縮できるはずです。このような例を、5年半のプロジェクトが終了するまでに、5例作ることを目標にしています」と蓮尾准教授は言う。既に数社との間で対話が始まっているそうだ。

公開日：2018/01/10

JST ERATO 蓮尾メタ数理システムデザインプロジェクトは、都内の地下鉄の駅に直結したビルの一室にオフィスを構える。手法の一般化の担い手と応用の担い手が集まって、日々議論を重ねている。

SCIENCE REPORT 014

データを発掘し、新たな歴史を記述する。

私たちはどこから来たのか？ ――歴史は、いつも人類の想像力をかきたててきた。人類のよりリアルな過去の姿を求めて、現代の考古学や歴史学では、科学的な計測・解析技術が広く用いられている。またデータがアナログからデジタルへ移行してからは、その利用方法も大きく広がり、さらにビッグデータ時代を迎えた昨今、大量データの高度な解析や、AI・機械学習などの手法も含めた、新しい「歴史を書く」という作業が試みられつつある。今回は、歴史などの人文学と、ゲノミクスや情報学が連携して取り組むデータサイエンスの挑戦についてお伝えしよう。

答える人

斎藤成也教授

［国立遺伝学研究所］

さいとう・なるや。1979年、東京大学理学部生物学科人類学課程卒。1986年テキサス大学ヒューストン校生物学医学大学院修了。Ph.D.（テキサス大学（米国））、博士（理学）（東京大学）。人間の進化に注目し、さまざまな系統独自の進化をゲノムデータの大規模比較により解析する。1987年に系統樹を作成する「近隣結合法」を提案した博士論文は5万3千件以上引用され、現在も被引用数を更新中。著書に『核DNA解析でたどる日本人の源流（2017年）』他多数。総合研究大学院大学遺伝学専攻教授、東京大学生物科学専攻教授を兼任。

北本朝展センター長

［情報・システム研究機構］

きたもと・あさのぶ。1997年、東京大学工学系研究科電子工学専攻修了。博士（工学）。現在、情報・システム研究機構 データサイエンス共同利用基盤施設 人文学オープンデータ共同利用センターセンター長、国立情報学研究所 コンテンツ科学研究系 准教授、総合研究大学院大学 情報学専攻 准教授。画像データの分析を中心に、人文科学、地球科学、防災などの幅広い分野で、データ駆動型のサイエンスを展開する。オープンサイエンスの展開に向けた超学際的研究コラボレーションにも興味を持つ。

貴重な出土品にドリルで穴を開けて

生物のゲノム進化を専門とし、特に現代人の進化、そしてヒトにいたる霊長類と哺乳類の進化に焦点をあてた研究を展開する国立遺伝学研究所の斎藤成也教授。2016年には福島県・三貫地（さんがじ）貝塚から出土した人骨からゲノムを解析し、縄文人が中国や東南アジアの人々とは遺伝的に大きく異なることを明らかにした。解析の元となるサンプルは、東大総合研究博物館所蔵の男女2体の頭骨の中にある奥歯に、ドリルで穴を開けて取得したという。「人類進化を知るには、人文系の情報が非常に重要です。日本の考古学が長い年月をかけて日本各地で発掘した史料が、既にたいへん豊富に集められているんですね。われわれは、それらの史料からデータを獲る探検に出かける。博物館がフィールドです」。

斎藤教授はさらに2017年、出雲に住む現代人のゲノム解析によって、従来2段階と考えられていた日本古代における大陸からの人々の渡来が、3段階だったことを示した。「三貫地貝塚では縄文人のゲノムを調べたわけですが、今度はこれを現代人のゲノムと比較することによって、ゲノムが似通った集団の分布にパター

ンがあることが見えてきたわけですね。すると、これまで知られていた渡来モデルが示す集団分布と、ずれがあることがわかったのです」。

このような遺伝学による発掘が、日本中でどんどん始まっている、と斎藤教授は言う。「遺伝学はようやく追い付いてきたんですよ。縄文時代のヒト集団のゲノムを、多くの地域で面的に調べていくことによって、これからさらに詳しいことが分かってくるでしょう」。

歴史記述とデータサイエンスの不可分な関係

このようにして、これまで人文学が蓄積してきた史料から、新しいデータを引き出す。ところが、「いや、データなんて排泄物みたいなもの」と斎藤教授は笑う。「計測する、デジタルファイルになる、それでおしまい。むしろ南方熊楠が書き残した“心・事・物”のベン図が参考になる」のだそうだ。「人間の“心”の働きと、自然界＝“物”の世界が重なった部分に“事”と書いてある。“物”を把握しようとする人間の“心”の中に“事”が生じる。“事”とは情報であり、データです」。

「生物学がすごく発達していろんなことが分かってきたと言われるけれども、実は分かっていないことだらけ。しかしゲノムは、遺伝情報という“事”でありながら、同時にアデニン（A）、グアニン（G）、シトシン（C）、チミン（T）という物質に1対1対応

しているので、非常に“物”＝本当の客観的な現象に近い。これは素晴らしい」。

斎藤教授が推進する「進化ゲノム学（Evolutionary Genomics）」は、これに時間軸が加わって、進化の過程でいかにゲノムが変化してきたかを問う分野だ。「僕らはゲノムを足がかりにして進化という自然現象を知りたい。人間を含めた世界に何があったのか、どう変化したのか、データによって現象をつぶさに記述することこそ重要なのです。この取り組みをデータサイエンスと呼ぶとすれば、それはデータという記述によって、歴史を証明することだと私は考えています」。

強者が生き残るのなら、なぜ生命は絶滅する？

広く知られる生命の進化の法則は、環境に対してより強い突然変異遺伝子を持つ者がより多くの子孫を増やすことによって、これまでの遺伝子と置き換わっていくとするダーウィンの自然選択（自然淘汰）説だ。「すべての生物がそれぞれの環境にすでに適応しているなんていうのはまったくの幻想ですし、自然淘汰でよりよい方が残るなら、基本的に種は絶滅しないはずです。ところが生命の歴史は、絶滅ばっかり。強い者の遺伝子が生き残ってきたというのはうそだ

2018年2月26日には機構合同シンポジウム「人文知による情報と知の体系化～異分野融合で何をつくるか～」（一橋講堂 千代田区一ツ橋2-1-2、学術総合センター内）が開催され、斎藤成也教授は「ヒトゲノム情報の革命がもたらした日本列島人史研究の新展開」と題する講演を行った。

ということです」と、斎藤教授は問いかける。

というのも、学術においては現在、突然変異が生じて遺伝子の塩基配列が変化しても、子孫を残す比率は従来の遺伝子と同様であって「淘汰上中立」であるという、木村資生の「中立進化説」が定説となっているからだ。「突然変異が生じて遺伝子の塩基配列が変化しても、大体のものは悪くなって消えてしまう。残るのは現状維持です。これをダーウィンの「正の自然淘汰」に対して、「負の自然淘汰」といいます。現在では、遺伝子変化の大部分が、この淘汰上中立な突然変異が長いあいだに蓄積していって進化が生じたものであることがわかっています」。

自然・社会・人文データで江戸時代を再現する

情報学の手法を用いてさまざまなデータを統合し、人文学などの様々な目的に利活用できるデータセットやツールを公開している、情報・システム研究機構データサイエンス共同利用基盤施設の人文学オープンデータ共同利用センター（CODH）北本朝展センター長。また2018年3月12日には、古文書に由来する地震学、気候学、天文学などのデータを多角的な視点で統合解析する手法を探る

CODHセミナー『歴史ビッグデータ〜過去の記録の統合解析に向けた古文書データ化の挑戦〜』を開催した。

「ビッグデータ時代の今ならTwitterなどに出来事を投稿するでしょうが、かつてはそれが古文書や古記録に記述され、今も残っているんですね。西東京の旧家が300年前から書き残している日記に天気の記録があったり、長野県の諏訪湖では、氷結した湖面が割れてせり上がり道のように見える「御神渡（おみわたり）」という神事があって、この公式記録が約600年も残されていたり、京都の神社では桜の開花日が長年にわたって記録されていたりします」。北本センター長が注目するのは、中でも比較的史料の多い江戸時代だという。

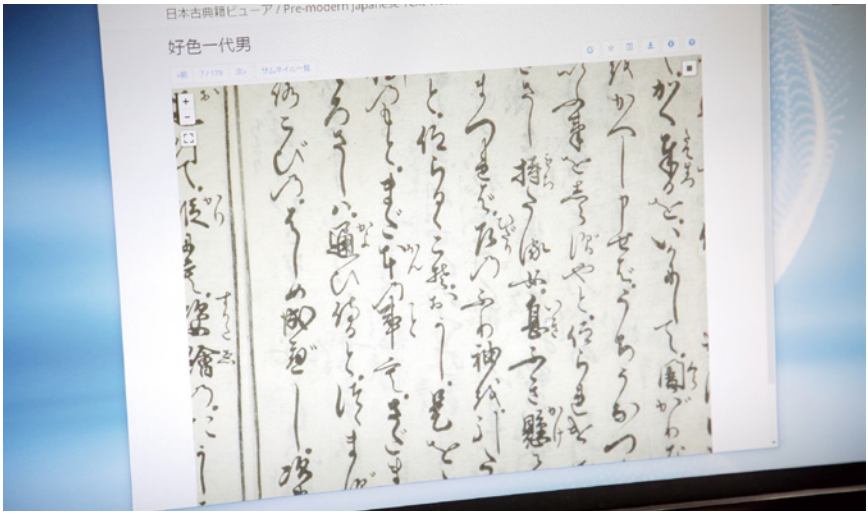
今回のイベントでも、気象の記録から古気候を再現したり、地震などの災害に関する記録を発掘したり、近世の市場変動や地下水管理などと自然現象を結びつけて解析したりする研究者達を集めて、人文情報学のコミュニティ形成を目指す。このように過去の「ビッグデータ」、すなわち過去の大規模な記録の網羅的な解析から歴史を再構成するという研究は、例えばベネチアの約1,000年分もの公文書などを分析して過去のベネチアを再現する「ベニス・タイム・マシーン」プロジェクトなど、世界的な潮流でもあるのだそうだ。「歴史的な記録を統合化

し再利用していこうという研究活動はこれまでも行われてきましたが、そうした研究成果はバラバラに散らばっているのが現状です。各地に残る記録を読み解き、デジタル化し、統合し、共有化するという複雑な作業を、みんなで協力して進められるような基盤を構築するのがわれわれの役割だろうと思っています」。

ミュージアムやライブラリの画像公開標準化とオープンサイエンス

CODHでは、ミュージアムやライブラリなどの画像配信方式として国際的な標準化が進む「IIIF（トリプルアイエフ）」の日本国内におけるコミュニティ活動推進にも取り組む。IIIFには国立国会図書館、大英図書館、フランス国立図書館、EU欧州委員会の電子図書館ポータルサイト「ヨーロピアナ」、オックスフォードをはじめ世界の大学が参加しており、現在約3億5,000万件以上の画像データが公開されているという。CODHでは「IIIF Curation Viewer」を開発・公開し、この活動に貢献している。

国文学研究資料館の関係機関が公開する古典籍15点の画像データから切り取った、くずし字3,999文字種・字形データ403,242文字の「日本古典籍字形データセット」も公開している。「機械学習（AI）によるくずし字認識ではルビの処理が難しい」と北本センター長。



「ウェブサイトで作品を見て、そこに描かれている人物の顔など注目する部分を切り取ることで、スクラップブックのようにお気に入り画像を蓄積・保存する仕組みがあります。以前なら、実際に美術館に行き、カメラやコピー機で複写し、はさみで切って、ノリで貼って……と行っていた作業が、この仕組みを使うと何百倍、何千倍も速くできます」と、北本センター長は言う。手軽に集めて分析できれば、発見の機会も増える。「顔だけを集めたキュレーションを作ってみたのですが、それで顔を比較してみると、実は別の絵巻にとてもよく似た描き方が見つかりました。絵巻は文字の部分と絵画の部分があるのですが、絵画の部分については絵師に外注する工房システムが存在していたらしいので、工房では顔のテンプレートを見ながら描いていた可能性もある。……現代のマンガとも共通するような技法が使われていたのかもしれませんね」。

データを集積し、公開し、そのプロセスを標準化することで、研究の検証や解釈の共有を飛躍的に改善することもできる。「人文学研究に貢献するだけでなく、それをどう変えるかがわれわれの重要なミッションです」。これまでは専門家の脳内に蓄えられていた素材や知識が、他者と共有して再利用できるようになる。「誰でもウェブ上の画像を使って、キュレーションができる。そんな市民参加の

オープンサイエンスも推進したいと考えています」。

人文学と情報学の交点から見えること

2017年には、江戸料理レシピデータセットの第一弾として、1795年刊行の『万宝料理秘密箱』に掲載されている卵料理のレシピを公開した。料理レシピの人気サイト「クックパッド」にも公開され、広く話題となった成果だ。しかし、このことは同時に「データが真に利活用されるためには、データを単に公開するだけでは不十分という大きな教訓を残した」と、北本センター長は言う。

「人文学データの面白さは、データの背後に潜む歴史や文化とのつながりを深めていける点にあるのではないでしょうか。私は、単に人文学データを使って情報学の研究を進めるだけでなく、人文学データの中身を理解することで人文学的にも新しい知識を得たいと思っています。誰かからもらったデータをブラックボックスに放り込んで答えを得るだけの研究をやっていると、高度なツールのオープン化に伴って情報学者の活躍の場はどんどん狭くなってしまうかもしれません。データをどう取得するか、データをどんな知識を得るために使うかなど、データを取り巻く環境に目を向ければ、新しい可能性が見つかるのではないでしょうか。」

公開日：2018/02/13

准教授を務める国立情報学研究所（千代田区一ツ橋）にて。

SCIENCE REPORT 015

宇宙の始まりを観測と計算で導き出す。

現代宇宙論では、現在、宇宙はその誕生から約138億年が経過したと考えられている。ビッグバンで始まった宇宙は超高温・高密度の状態で膨張を続けながら、どんどん冷却されていった。そして約38万年後に「宇宙の晴れ上がり」が起こって光が直進できるようになり、1億年後には重力によって集められたガス雲の中から最初の星が輝き始めたという——。そもそも天文学は「本当にただ無目的に星を観測して、宇宙の姿をジーッと見ていたら面白いものが見つかったという、むしろデータサイエンスに近いところから始まっている」という、東京大学・カブリ数物連携宇宙研究機構の吉田直紀教授。「例えば銀河系がどら焼き型だというのは、もう200年以上も前に発見されたのですが、なかなか頭で考えて出て来るような話でもない」。かつてない高解像度のデータが大量に得られるようになった21世紀現在、これらを駆使して、宇宙の始まりはどう解明されつつあるのだろうか？

答える人

吉田直紀教授

[東京大学・カブリ数物連携宇宙研究機構]

よしだ・なおき。1996年、東京大学工学部航空宇宙工学科卒業、理学博士（ミュンヘン大学）。米国ハーバード大学天文学科、名古屋大学を経て、2012年より東京大学教授。専門は宇宙論・宇宙物理学。宇宙のダークマターとブラックホールの謎に迫るべく研究を行う。2014年よりカブリ数物連携宇宙研究機構 特任教授を兼任。

池田思朗教授

[統計数理研究所]

いけだ・しろう。1996年、東京大学博士課程修了、博士（工学）。統計的多変量解析手法の1つである「独立成分分析」を基にした音信号の分離や、雑音の多い計測データ解析などにむけた幅広い信号処理、解析手法の開発に取り組む。推定対象の信号に零が多いという仮定を用いる疎性モデリングにより、天文や物理計測の改善法などを手がける。

森井幹雄特任助教

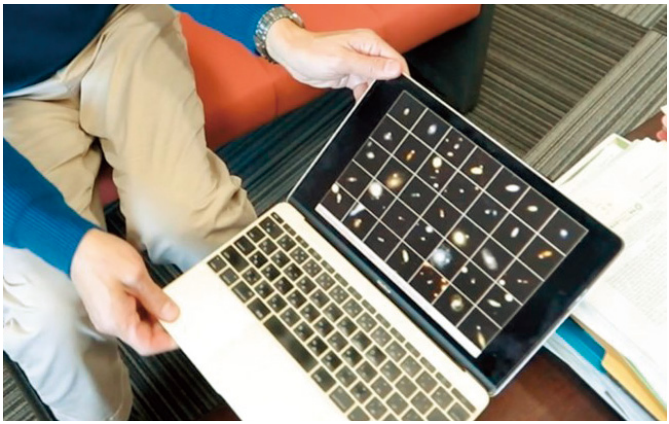
[統計数理研究所]

もりい・みきお。森井幹雄特任助教の専門は、観測天文学。

どうしたら宇宙の始まりを知ることができるのか？

吉田直紀教授は、5年間にわたってすばる望遠鏡の新型カメラを使用して宇宙を観測し、得られた大量の画像を解析するプロジェクトに取り組んでいる。集める画像データは、およそ1ペタバイトにもなる。「1つの画像の中にだいたい数千個の銀河が写っています。もやっとした雲のように見えても、ズームインしていくと実は1個1個の星にまで分解できる高い解像度を持っています。かなり広い領域の宇宙で起こっているいろんな出来事をカメラに収めることができるんです」と吉田教授は言う。「宇宙は静的な、いつも変わらぬ姿をしていると思いがちですが、昨日撮った画像と今日撮った画像を比べると結構変化していて、すばる望遠鏡の観測データを解析すると、星が爆発したとわかる箇所が一晩にだいたい100個ぐらい見つかります。1つ1つ拡大すれば人間の目でもわかるのですが、なにしろ膨大な数の天体が映っているので、コンピュータでそれらの画像の差分を検出し、宇宙の変化を捉えています」。

取り組みの背景にあるのは、2011年にノーベル物理学賞を受賞したソール・パールムッターら3人の天文学者の発見だ。1998年、彼らは遠くの超新星爆発を調べることによって、「加速膨張」すなわち現在の宇宙は膨張し続けていることを示した。吉田教授がターゲットとする超新星も同じIa型で、観測で見つかる超



解析によって特定されたIa型超新星。星の明るさやその変化とともにデータベースに収められている。

新星の約半分を占めるという。「最近の人工知能の画像解析能力は結構すごくて、画像からIa型を見分けることができるんですね。このツールづくりに、統数研の池田教授に協力いただいています」。

Ia型超新星を高い精度で見つけ出す

「われわれの課題は、Ia型と思われる突発天体を確実に見つけることです」と、統計数理研究所の池田思朗教授。森井幹雄特任助教によれば「Ia型超新星は太陽質量の1.4倍の白色矮星が爆発してできるもので、どれも星自体の明るさがほぼ同じ」ものなのだそうだ。「見かけの明るさから距離を見積もることができ、スペクトルを観測することによって星が遠ざかる速度がわかります。距離と速度を合わせれば宇宙がどれくらいの速さで膨張しているかを推定することができます」と、森井特任助教。「また宇宙の遠くでは、もともと考えられていた速度よりも速く膨張しているというのが加速膨張」だという。

ところが、「99％は偽物なんです」と池田教授。「時刻の異なる二つの画像の差から超新星を見つけるのですが、うまく差が取れなかったとか、全然違うもののほうがずっと多く見つかります。これだというIa型の候補は1％未満しか含まれません。そうしたIa型の候補を選んで、別の望遠鏡で追観測します。追観測では時間をかけてスペクトルを観測し、位置、速度と併せて星のプロフィールがほぼすべて明らかになります」。現在、高い精度で検出できており、一晩におよそ10～20個程度が超新星と特定されているのだそうだ。

ところで、池田教授の専門分野は宇宙ではなく、統計である。「何らかのノイズが含まれるような観測結果がある時に、その背後にある対象の本当の姿を推定する――これが統計の仕事ですね。このうち特に対象が物理的な信号であるような場合を信号処理と呼びます。いろんな信号処理の方法を使わないと何も見えないデータはたくさんあって、天文データではすばる望遠鏡のような可視光データの他、電波望遠鏡が受け取ったデータの解析にも取り組んでいます。計測機器の発達によって、これまでにないデータが得られるようになった現在、信号処理は広がりのある活発な学問になっている」と池田教授は言う。

東京都立川市にある統計数理研究所・国立極地研究所共同棟。

ダークマターの分布地図を作る

一方、吉田教授は、研究のもう1つの大きな柱として、現在「重力レンズ」という現象にも注目している。「宇宙のあちらこちらに、歪んだ形の銀河があることが知られています。それらの銀河は実際に歪んだ形をしているのではなく、そのように“見える”だけなのです。実は、その見かけの歪み具合を逆算すると、銀河の周辺には見えないけれども何か大量の物質があるということが分かります」。この「見えない何か」こそ、ダークマターと呼ばれるものだ。宇宙の構成物は現在、4％が通常の元素、22％がダークマター、そして残りの74％がダークエネルギーだと考えられている。「ダークマターの3次元的な分布を広範囲に調べて、いわば宇宙の地図を作っています」。

宇宙の大規模構造については、1900年代後半頃から世界的に研究が進展しているが、「重力だけを考慮した計算からダークマターの分布を見ていくことでも、宇宙についての理論がだんだん洗練されていく」と吉田教授は言う。「本当にごく微細な差を検出し、宇宙観測から得た分布と一致するためにはダークマターにどんな性質が必要なのか、物質の密度はどのくらいなのかなどを導くことで、理論やモデルを判定するんですね。しかし物質分布を導くというのは、出てきた結果から原因を探るというすごく難しい問題なので、ここでも統計数理研究所の知識を結集しているところです」。

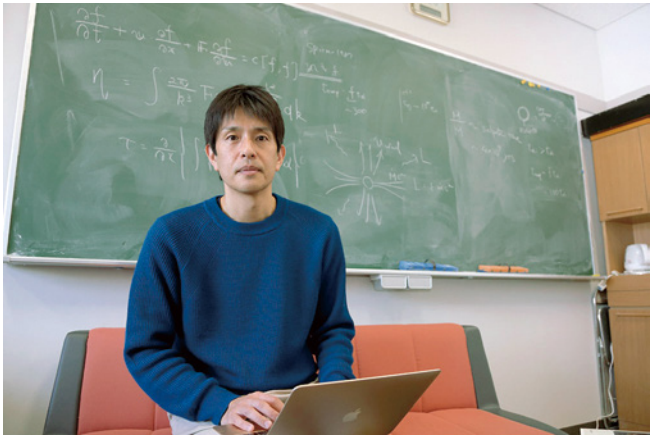
吉田教授が指さしている顔のような画像にある円弧状のものが「歪み」の一例。
強い重力が直線に進むはずの光を曲げ、遠い銀河が曲がっているように観察される。

宇宙は本当の分布を教えてくれない

吉田教授はこの観測・解析と並行して、宇宙の始まりから星や銀河がどのように生まれてきたのかを再現するコンピュータ・シミュレーションに取り組んでいる。「宇宙は本当の分布を教えてくれないので、われわれは観測から最も確からしい答えを出します。一方シミュレーションは、コンピュータの中に正確な物質分布を作り出すことができる」と吉田教授は言う。理論に基づいたコンピュータ・シミュレーションを行うことは、解析方法を検証し、その精度を上げることにもつながる。「宇宙全体の進化や、宇宙全体の構成物が分かってくるのがゴールですね。そこのためには人工知能も使うし、コンピュータ・シミュレーションも使います」。

考えていくための手がかりとして「宇宙では、この地球とは全然違うことが起きているんだと言ったら何も先に進まない」と吉田教授は言う。「むしろ考えの拠りどころは地球上にあって、身の回りで確かめられたことを使って宇宙を理解できるかどうかですね。今、私たちが知っていることを総動員しても理解できない場合は、何か足りないのか？　何かまだ知らない現象が起こっているのか？と推測できる。どう考えても新しい要素を足さざるを得ないというところまで来て、初めて発見につながります」。

公開日：2018/03/12



東京大学本郷キャンパスにある吉田教授の研究室にて。

$$\frac{\partial f}{\partial x} + F_1 \frac{\partial f}{\partial v} = c[f, f']$$

$$\frac{2\pi v}{k^3} F_1(k, w)\, e^{-\frac{k^2}{2\sigma^2}}\, dk$$

$$\frac{\partial}{\partial x}\int \frac{G(q, x)}{\sqrt{x^2 + 3t}}\, dx^3 - \alpha$$

SCIENCE REPORT 016

スポーツを統計の知で応援しよう。

貧乏球団が、データの力で一躍「強豪」チームに！　――野球のプレーに客観的な指標を与えたり、データを集めることでチームの戦略に結びつけたりする「セイバーメトリクス（SABRmetrics）」が広く注目され始めたのは、1980年代頃のことだ。ますます進化する大リーグのデータ解析に触発され、スポーツのための統計を日本に広めようと、2009年、日本統計学会にスポーツデータ分科会が設立された。設立以来のメンバーである田村義保教授（統計数理研究所）、酒折文武准教授（中央大学）、竹内光悦教授（実践女子大学）が中心となって始めた「スポーツデータ解析コンペティション」は、今回で第7回を数える。スポーツの発展に、統計はどう貢献できるのか？　またどんな人材が、ゲームを変えていくのだろうか？

答える人

田村義保特任教授

［統計数理研究所］

たむら・よしやす。兵庫高校出身、理学博士（東京工業大学）。1986年統計数理研究所助教授、1997年同教授を経て、2018年より現職。統計的手法の研究開発と、その実社会へのさまざまな応用に取り組む。同研究所の大学共同利用機関としての役割から、シミュレーションに不可欠な物理乱数発生装置の研究開発でも知られる。

答える人

酒折文武准教授

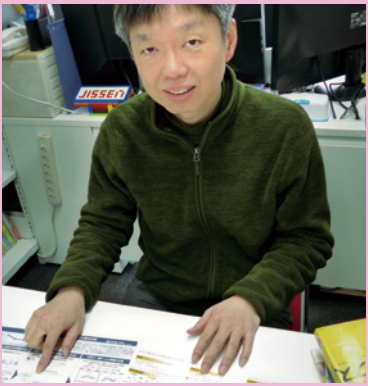
［中央大学］

さかおり・ふみたけ。中央大学理工学部数学科准教授、博士（理学）。専門は統計的モデリング、計算機統計学、スポーツ統計科学、統計教育。立教大学社会学部助手、中央大学理工学部専任講師などを経て、2009年より現職。統計数理研究所客員准教授を兼任。

竹内光悦教授

［実践女子大学］

たけうち・あきのぶ。実践女子大学人間社会研究科教授、情報センター長、博士（理学）。専門は、統計科学、行動計量学、統計教育。日本統計学会統計教育委員会の長として、長く統計的思考力を育成する統計教育に取り組み、またデータサイエンス教育への情報端末の活用、学外イベントを活用したアクティブ・ラーニング、統計教育へのゲーミフィケーションの導入、公的統計を利用した授業開発等を推進。



ただ単にすべてのスポーツが好きだから

2018年3月19日、東京都立川市にある統計数理研究所にて、日本統計学会、情報・システム研究機構統計数理研究所等が主催するシンポジウム「スポーツアナリティクスと統計科学（第7回スポーツデータ解析コンペティション受賞者講演会）」が開催された。そもそもこのコンペは、株式会社データスタジアム提供によるデータを用い、野球、サッカー、バスケットボールの3つの種目から選んで解析して、「分析」「インフォグラフィック」「中等教育」という3つの部門ごとに研究成果を競うものだ。前年の9月末に応募が締め切られ、2018年1月に最優秀賞をはじめ各賞の発表が行われた（中等教育部門は別スケジュールで実施）。本シンポジウムは、その表彰式を兼ね、コンペの締めくくりとして行われたものだ。

野球が好きで、大学院生の時はキャッチャーをしていたという統計数理研究所の田村義保特任教授は、「開始当初は、データサイエンティスト協会もないし、

データサイエンスという言葉もなかった」と振り返る。「スポーツを対象に統計を駆使しようと思ったのは、すべてのスポーツが好きだから。僕だけでなく、みんなスポーツには興味があるでしょう？ つまり、スポーツはデータ解析の練習に最適なんです」。参加者も年々増加し、今年の分析部門には61チームの研究成果が寄せられたという。

また種目別には「野球よりもサッカーの解析のほうがいい研究が多かった印象がある」と言う。野球は1球ごとのデータしか提供されないが、サッカーはプレイデータに加え、ボールの軌跡や1つ1つのボールのタッチデータを記録したトラッキングデータがあるため、組み合わせて解析することができるためだ。「つまり、サッカーのデータの質が上がった」と田村教授。2016年のコンペでは、ラグビーを採り上げるなど、種目の幅も広げているという。難しい統計でも、スポーツとなるとなんだかわくわくして、参加したくなってくる……どうもスポーツには、そんな不思議な力があるようだ。「うれしいことに、このコンペで2回も発表したことのある学生が、日本のプロ野球選手になったという例もあるんです。スポーツデータの解析ができる人材のニーズは高く、今や、すべてのプロ野球チームが欲しがっていると言っていいでしょう」。

スポーツのデータに触ってみるという意義

田村教授と並んで第1回のコンペから中心的に活動してきた中央大学の酒折文武准教授も、「そのスポーツを知っていれば、分析した結果から何を読み取ったらいいかわかる。その意味で、スポーツは教育的なコンテンツとして、非常に優れている」と言う。「特にスポーツ好きの人にとっては、最高のテーマなのではないでしょうか」。

コンペを通じて、とにかくデータに触れ、統計的な思考力を鍛えて欲しい――そんな主催者たちの思いは、しかし、本当に実現しているのだろうか？ 「やはり非常に成果があるというのが、率直な感想です」と酒折准教授。「何度も参加しているチームが、どんどん上達しているんです。どういう視点で、どういう

第7回スポーツデータ解析コンペティションの結果は日本統計学会 スポーツ統計分科会のウェブサイト（http://estat.sci.kagoshima-u.ac.jp/sports/）で公開されている。

切り口で、何を伝えたいのか。他のチームの優れた研究発表を見ることによって、統計的な考え方が洗練されてくるのだと実感します」。

審査では、まず分析テーマが魅力的であるかが問われる。「結論が意外だったとか、実用性が高い研究も評価されます。つまり、なぜその分析をしなければならないかということが大事」と酒折准教授。「次に分析手法が適切か、工夫が見られたかどうかを審査します。分析手法やその使い方に独得な点や新規性があると加点の対象になります」。与えられている変数をそのまま使うのではなく、加工したり新たなデータを取得したりするのも評価されるそうだ。

そして何より「誰に向かって提案しているのか。選手やコーチに伝えたいことなのか、あるいはチームの運営戦略なのか、それとも、そのスポーツの中継を面白くしようというメディアの目線なのかが重要」なのだという。「それがあれば、成果をどうプレゼンテーションするかも自ずと決まってきます」。

ビヨンド2020のスポーツ界へ向けて

コンペは整理されたデータセットが提供された中での競争だが、「ほとんどのスポーツは、僕らから見るとデータが充実していない、取得すらされていない状況にあります。その中でスポーツ統計がどれだけ貢献できるかは、ひとつのチャレンジ」と、酒折准教授は言う。「例えば柔道のような対戦競技の試合の動画があったら、そこからいかに分析データを生み出していくかも今後の課題だと思います。しかしそれは形式が整っていない、いわゆる“非構造化データ”であるため、どうアプローチすべきかは自明ではありません。統計学の知識だけではなく、もっと情報学的な進展も必要になってきますね」。

大リーグをはじめバスケットボール、サッカー、アメリカンフットボール等々のプロスポーツが盛んな国々では、「実は、スポーツ統計専門の科学雑誌がたくさんあって、かなり研究が進んでいる」のだそうだ。「日本でも2020年以降のスポーツ振興へ向けて、スポーツチームのコンサルティングを行う会社が活動し

シンポジウムで、データの説明をする酒折准教授。インフォグラフィック部門については、今回応募が少なかったこともあり、「何を伝えるべきかを考えて工夫してほしい」との講評を述べた。

始めています。また2014年には日本スポーツアナリスト協会（JSAA）が設立されて、年々大規模なカンファレンスを開催しており、スポーツアナリストの情報共有も活発化しています。このような状況から、10年後には、かなりゲームチェンジが起こってくることが予想されますね」。

日本人の誰もがデータで語れるように

一方、中等教育部門は、3回目のコンペから加わった部門で、参加者は研究成果をポスターにまとめて応募する。第5回にあたる今回は、全国から全65もの力作が寄せられた。実践女子大学の竹内光悦教授は、開設以来、本部門を担う。「日本人の数学の能力は、国際比較などで、その高さが認められています。でも統計について調べてみると、中国、アメリカ、韓国、ニュージランド、イギリス、オーストラリアといった、つまり日本以外の国々では、その教育がもっと進んでいる」。そこで文系・理系問わず、市民が必要な能力として統計を入れていこうと中等教育の学習指導の見直しが進められているが、竹内教授は近年、その活動にも注力してきた。「平均という概念ひとつをとっても、少し前まで日本では小・中学生の段階で、データを代表する値として中央値や最頻値を学んでいなかった。それで大人になって、国際社会における日常的なビジネスの会話についていけるのか？」と、強い危機感を訴える。

第7回スポーツデータ解析コンペティション中等教育部門の結果は、
日本統計学会 スポーツ統計分科会のウェブサイト「中等教育部門の結果発表ページ」
（http://estat.sci.kagoshima-u.ac.jp/cse/sports.htm）にて公開。

さて、今回のスポーツ統計だが、竹内教授は「社会調査のデータと違って、正規分布に近いようなきれいなデータが多い」と指摘する。「また、勝つというゴールが見やすい――しかし、なぜうまくいったのか、どの要因が一番効いているのかは、分かりませんね？　そこを科学的に突き詰めていく、その考え方を統計を通じて学んでいって欲しいのです。例えばバスケットなら、シュート率が悪い原因は何か、勝つチームの傾向、負けるチームの傾向は何か……というように、いろいろな統計手法を使って見ていきます。さらに、今見えているものは表面的なものだから、別の分析によって新しい関係性が見つかるのではないかといった気付きも重要です」。

地元のバスケットチームを強くしようという香川県立観音寺第一高等学校チームのポスターは、まさにその好例だ。まず特性要因をまとめて図にし、主成分分析を使った解析やその検証などを行い、最後に提言を行う――この一連の内容がぎっしり詰まったポスターで、最優秀賞を受賞した。「理系に進むにしても、文系でマーケティングや戦略立案へ行くにしても、その基礎にはすべて統計がある。このスポーツコンペの役割は、データで語れる人を育成することにあると考えています」。

公開日：2018/04/12

SCIENCE REPORT 017

データ解析力で日本の医療を支援する。

2017年11月、国立情報学研究所（NII）医療ビッグデータ研究センターが、2018年4月、統計数理研究所（ISM）医療健康データ科学研究センターが、それぞれ設立された。情報・システム研究機構に属するこの2研究所はそれぞれIT、統計の知見を結集するだけでなく、大学共同利用機関として、他分野との共同研究を促進する役割をも任ずる。両研究所が担うデータ基盤、そして昨今「AI（人工知能）」に象徴されるデータ解析力は、本連載のこれまでの回でも見てきたように、ますますイノベーションや科学の発展の駆動力となっており、中でも私たちに身近な医療は、連携によって拓かれる新たな道に期待がかかる。データサイエンスは日本の医療にどれだけ貢献できるのか、研究開発の狙いはどこにあるのか、2つのセンターの活動を紹介しよう。

答える人

佐藤真一教授

［国立情報学研究所］

さとう・しんいち。1987年東京大学卒、工学博士（東京大学）。専門は、画像・映像解析に基づく検索・知識発見。1995～1997年米国カーネギーメロン大客員研究員としてInformedia映像ディジタルライブラリの研究に従事。2000年国立情報学研究所助教授、2004年より現職。

答える人

伊藤陽一教授

[統計数理研究所]

いとう・よういち。東京大学卒、保健学博士（東京大学）。専門は生物統計学。ゲノム解析、臨床試験のデザインに関わる統計解析コンサルティングの専門家。2009年より新薬を審査する医薬品医療機器総合機構（PMDA）の専門委員を務める。近年は臨床試験データの管理改善を効率の観点から評価するという新しい研究領域に挑む。

野間久史准教授

[統計数理研究所]

のま・ひさし。九州大学卒。博士（社会健康医学）。専門は医療統計学・公衆衛生学。京都大学大学院医学研究科で医療統計学を専攻した後、統計数理研究所助教等を経て現職。「専門の医療統計学を通じ医学・医療へ貢献したい」と研究に教育に奔走する。

ITのエキスパートがつくるチーム

東京・一ツ橋にあるNIIは、日本で唯一ITだけを総合的に研究する研究機関だ。2017年から日本医療研究開発機構（AMED）「医療のデジタル革命実現プロジェクト」のパートナーとなり、日本消化器内視鏡学会、日本病理学会、日本医学放射線学会、日本眼科学会の4つの学会とともに医療健康分野のICT基盤をつくり、画像解析、深層学習、AI（人工知能）などを開発する課題に取り組む。

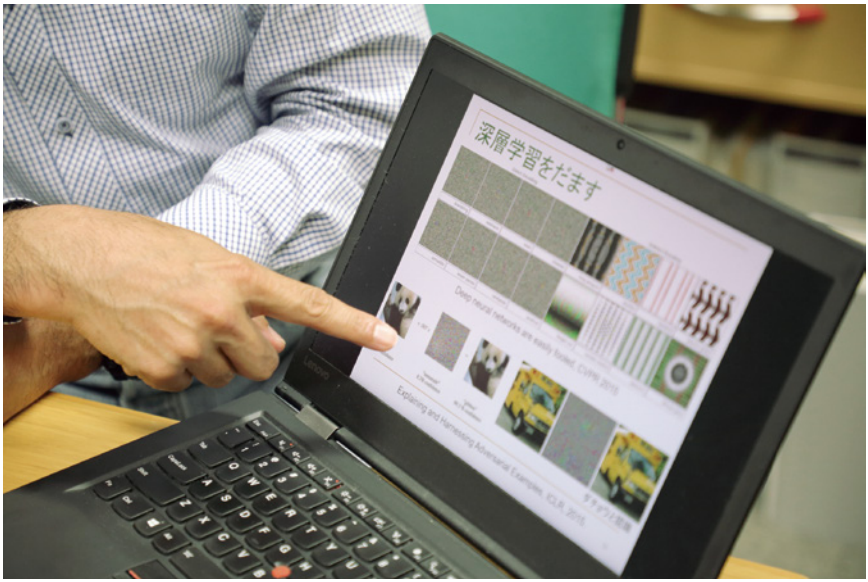
このミッションを担う医療ビッグデータ研究センターには、NIIだけでなく東大、名大、九大等から選りすぐりの画像解析研究者が参加する。メンバーは、若いポスドク研究員などを含めて約10名。ただし、そのほとんどが医療画像解析の経験は浅いという。「私は画像解析や深層学習の研究者で、これまで医療はやってこなかった」と言うのは、センター長の佐藤真一教授だ。「けれども、ある意味、できる。つまり、画像を見て僕らは判断できないのだけれど、僕らがトレーニングしたAIの一種であるニューラルネットワークは判断するんです」。

これまで医療画像の解析に取り組んでいた科学者たちは、自分の中にある医学の知識をいかに画像処理に反映させるかを目指してプログラミングしていた。「潮目が変わったのは、機械に大量の学習データを食わせれば何らかの結果が出てくるようになったから」と佐藤教授は言う。「共同研究がスタートしてまだ1年と数カ月ですが、特に眼底画像を使った糖尿病や緑内障の診断では、ほとんど工夫をしていないような画像解析の人工ニューラルネットワークで、既に高精度な識別ができています。ところがこれ、医師が行うと難しい診断なのだそうです」。

画像解析・深層学習の「第三の波」

ところで佐藤教授の言う潮目の変化とは、いつ頃なのだろうか？ 「2012年に画期的な機械学習アルゴリズムが現れ、神経細胞（ニューロン）網をシミュレートして学習や判断を行う手法により、これまでをはるかに凌駕する性能を達成できたことが大きいですね」。歴史的に振り返ると、人工ニューラルネットワークの研究はこれまで大きく3段階で発達してきたと佐藤教授は言う。「第一の波は、脳の神経回路を模したニューラルネットワークの登場です。事実上1層しかないネットワークで、任意の線形関数が学習できることが証明され、入力さえ与えれば学習できる枠組みができました。第二の波がいわゆるニューロファジーです。深層学習は2〜3層と少し複雑になりました。そして先ほどの2012年に、コンピュータの性能向上やビッグデータを背景に大量データを使って学習しようという第三の波が到来したのです」。

また、その年トロント大のヒントン教授らが発表した、画像を1,000種類の物体に見分ける論文は、どのくらいの量のデータを使って機械が学習し、どの程度の認識精度で見分けられるのか、画像認識というゲームの中でひとつの大きな到達点となった。「これまではどの時点で機械が何をどのように見るべきかは人間によって与えられていたのですが、ニューラルネットワークはすべて白紙で、全部学習によって（機械が自らが）決めていくところに大きな特徴があります。たとえば1,000種類の物体の画像をとにかく大量に入れて学習させると、未知の



高い精度で画像を認識できるシステムが「だまされる」ケースがあることも、専門家の間では広く知られている。「例えばパンダの画像に慎重に生成したノイズ状の画像を加えてやると、99.3％の確信度で機械はGibbon（テナガザル）だという」と佐藤教授。「想定外のことが起こった時に深層学習がどう振る舞うか、いろいろとテストして理解していく必要がある」。

画像が来た時に非常に精度よく判断してくれる。残念ながら、機械が何を勉強したのかはわからないのですが、人間が昔一生懸命考えてやっていたことを簡単に陵駕する性能が出てくるのです」。当時ヒントン教授らが用いた深層学習は8層だったが、現在では「100、200といった数の層が用いられるのが普通」だと言う。

深層学習は「理論解析が30年前から本質的に進んでいない」という問題はあるものの、専門家の間では人工ニューラルネットワークに何がどこまでできるのかが経験的に知られるようになり、特に近年、画像領域において圧倒的に高い性能が達成されるようになってきた。NII医療ビッグデータ研究センターでは、時宜を得てこの画像認識精度を医療ビッグデータのデータ解析に使おうという狙いがある。

医師たちをいかに支援するか？

現在、佐藤教授らが取り組む医療画像では、それぞれの画像を医師がどう判断したかという「正解」のデータ作りも進められている。しかし「例えばイヌやネコなら誰でも識別できますが、医療画像はお医者様でないと判断できないので、お金に代えられない高価さがある」。そこで「能動学習のアルゴリズムを活用し、どうしても機械がここは医師の判断を仰ぎたいというポイントを絞って、要所要所で正解を付けてもらう」という方法によって、医師の労力を数分の1にする設計を進めているという。「お医者様の忙しさを解消しようとしているのに、データを作るにはその方々にご協力いただかなければならない。それが今、最も悩みの種ですね」。

しかし近い将来、画像解析によって医師を支援するシステムが回るようになれば「見落としを少なくしたり、全く独立な意見としてセカンドオピニオンを提示したり、また簡単な症例は機械が判断して医師たちは本当に難しい症例の診断に注力したりといった支援が考えられます」と佐藤教授は言う。「もしも経験を積んだ医師とほぼ同じような判断をする人工知能アルゴリズムができたら、離島等の必要な場所へ配布できるのも夢ですね。しかもプログラムなので100でも

1万でもコピーできます」。

医学部にデータサイエンティストを育成する

一方、来る2019年に75周年を迎える歴史ある統計数理研究所（ISM）には、このほど医療健康データ科学研究センターが発足した。医療統計のエキスパートを結集して、来る医療・健康データの利活用時代に備え、さまざまなデータ解析基盤の整備を目指す。「日本全国に約60の大学医学部がありますが、これらの学部で決定的に不足しているデータサイエンティストの養成や、医療統計のリテラシー向上のための教育をミッションとしているところも、われわれのセンターの大きな特徴です」と伊藤センター長は言う。

年間4本の医療統計教育コース、5つの公開講座等が既にスタートしている他、2018年5月28日には、当センターの設立記念シンポジウムも開催された。シン

「統計数理研究所 医療健康データ科学研究センター設立記念シンポジウム」は2018年5月28日（月）14:00～18:10　秋葉原コンベンションホール（東京・千代田区）にて開催された。写真は伊藤陽一センター長（左）、野間久史副センター長（右）。

ポジウムを担当する野間副センター長は、「その前の公開イベントでは、思いがけず高名な医学の先生にお越しいただき、医療・健康分野における統計への関心の高まりを肌で感じた」という。また発足に先立ち、統計教育推進のための「健康科学研究ネットワーク」形成を呼びかけたところ、全国の医大、病院など約70の機関が「わずかな時間で集まった」という。

新薬の臨床試験を統計的に厳密に評価する

もともとライフサイエンスの分野は、統計とは切っても切れない関係にある。「生物統計の重要性は広く認知されていますし、新薬を開発する時に、その有効性や安全性をきちんと評価する臨床試験は、統計が不可欠となる典型例」と野間副センター長。その臨床試験評価のエキスパートとして知られる伊藤センター長は、実験を設計する統計コンサルタントとしても長いキャリアを持つ。

「ヒトは生物学的に非常に複雑なシステムを持っているので、マウス、イヌ、人間にかなり近いサルなどで実験して安全だという結果が得られていた薬剤でも、ヒトに投与したらまったく予想外の副作用が起こったという事例が、歴史上に

「小さい頃体が弱くて、医者になりたいと思っていた」という伊藤センター長。2009年より務める医薬品医療機器総合機構（PMDA）専門委員は、PMDAの考え方が妥当かどうかを外部の視点でチェックする役割を担う。「新薬審査の質の向上に貢献できる重要な仕事と考えています」。

いくつかあります。新薬が本当に治療に効果があるのか、ヒトに投与して安全なのか、実験や臨床試験から得られるデータの不確実性を統計学的に厳密に評価した上で、科学的評価を行わなくてはなりません。ここに統計の技術が要求されるのです」（伊藤センター長）。

「臨床実験のデザインも日進月歩で新しい方法が生まれていますから、キャッチアップしながら新しい方法論を生み出していかなければなりません。この点はビッグデータの解析とは対照的で完全に計画し尽くして証明する作業であり、また極めて専門性の高い分野になっています」（伊藤センター長）。

健康リスクとなる環境要因どう測るのか

もうひとつ、集団を対象として疾病の発生原因や予防などを研究する疫学（epidemiology）という分野でも、統計が大きな力を発揮する。疫学は19世紀イギリスにさかのぼり、コレラ等の感染症の予防や原因究明に関する学問として始まったが、現在社会の疾病構造の多様化とも関連して、むしろ医療と健康全

2018年5月、医療健康データ科学研究センターにおける研究プロジェクトの最新の成果として、臨床試験のエビデンスの統合解析「メタアナリシス」の新しい統計手法のプレスリリースも公開された。「メタアナリシス」とは、「過去に行われた臨床試験のエビデンスを統合し、その有効性・安全性を総合的に評価するもの」と話す、野間副センター長。

般を対象とした研究へと変化してきた。大学院時代からこの研究に取り組んできた野間副センター長は、「人を対象とした観察研究に基づいて、たとえば喫煙、環境ホルモン、大気汚染への曝露といった要因の健康への影響を科学的に計るもの」と解説する。「しかし実験とは違って、健康に有害な要因はコントロールすることができません。このような対象からどうデータを取るのか、他の要因が関連してはいないか、本当に因果関係があるのか、正確に推定するための統計的手法を構築します」。

中でも専門家たちが統計的な視点を注ぐのは、真実と測定されたデータの間にあるバイアス（偏り）だ。「これを評価するのが非常に難しい」と野間副センター長。「問題設定上、100％正確な評価はまず不可能な領域なんです。でも公衆衛生的には何か手を打たなければいけない。そこで何が有効かを判断するにも、実はまた統計が必要なんですね」。

伊藤センター長は「疫学のほうが歴史が長いが、近年、その方法論が非常に精緻化されてきている」と見る。「私の専門である臨床試験とは本来出自が異なるのですが、疫学の中にある統計と、臨床試験をはじめとする生物統計が出会って、今、方法論が活発に共進化している状況にあると考えています」。

統計手法を基礎研究として融合的に発展させる

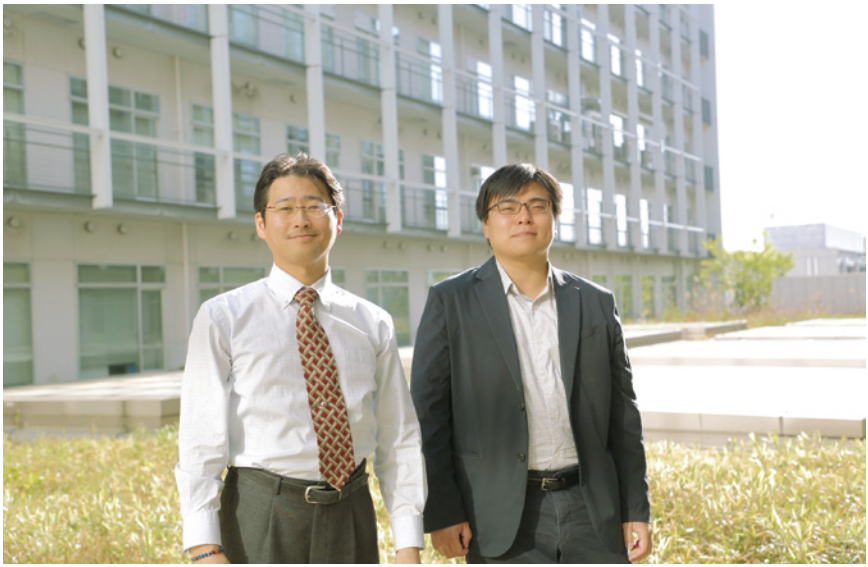
伊藤センター長の言う方法論的な展開は、センターが担う6つの研究プロジェクトの1つである「臨床研究・臨床試験とエビデンス統合の方法論プロジェクト」で進められている。ちなみに、このプロジェクトの具体的なテーマの1つに、2015年に米国でオバマ前大統領が一般教書演説で発表した精密医療（precision medicine）の実現をめざしたビッグデータ解析がある。野間副センター長は言う。「例えばサリドマイドは、1950～60年代に世界的な薬害を引き起こした物質ですが、血管新生を阻害する点に逆に着目して、1999年に米国で多発性骨髄腫へのエビデンスが認められ、2008年には日本でも抗がん剤として承認されるなど、世界的に再評価が進んでいます。また、新しいデータサイエンスの手法を用い

た解析を通して、その薬効にどうやら個人差があるらしいこともわかってきました」。

「患者さんから切除したがん細胞から、遺伝情報や分子レベルの情報を網羅的に測定すると、数百万次元以上の規模の大規模データが得られます。これらをオミックスデータといい、この10、20年の間に、その測定技術は非常に発達してきました。このデータと薬効の関連を新しい統計手法で分析すると、サリドマイドを投与した患者グループは、そうでないグループと比べて予後がいい。また特定の分子の発現パターンがある患者さんとそうでない患者さんでは薬の効き目が違う……といったことが詳細に評価できます」（野間副センター長）。

この他、「機械学習の技術を遺伝子データの探索的な解析等に使ったり、古典的な統計の理論的な進展をさらに発展させるなど、ビッグデータ利活用の広がり呼応したデータ解析の基礎研究を積極的にやっていきます」と、伊藤センター長。「私たちにとって、今ある集団を完全に記述するモデルよりも、新しい患者さんをうまく予測できるほうがいいモデル。これを構築することこそ喫緊の課題であり、また面白いテーマと言えます」。

公開日：2018/05/10



SCIENCE REPORT 018

ビッグデータ時代、その先を展望する。

日々大量に生成されるデータが企業、大学などで、広く利活用されるようになったビッグデータ時代。今回は、ITと統計数理の最先端をそれぞれ担う、情報・システム研究機構の2人の所長の対話により、このような時代のサイエンス、技術、教育をリポートする。データを資源として機械自ら判断する「AI（人工知能）」、データを駆使して現代の複雑な問題に解決を見出そうとする「データサイエンス」、そして加速的に発達するコンピュータがいつか人間の能力を越えるであろう特異点「シンギュラリティ」といった、まさにホットな科学技術のキーワードを交え、今起こりつつある変化を展望する。

答える人

喜連川優所長

［国立情報学研究所］

きつれがわ・まさる。2013年より情報・システム研究機構 国立情報学研究所所長、ならびに東京大学生産技術研究所教授。1983年東京大学卒、工学博士。専門はデータベース工学。情報学を代表する研究プロジェクトを率い、また非順序実行方式による高速データベースエンジンの開発、巨大なデータ量を持つ地球環境統合データベースの運用などで知られる。情報処理学会前会長、日本学術会議情報学委員会第23期委員長。

樋口知之所長

［統計数理研究所］

ひぐち・ともゆき。1984年東京大学卒、1989年同博士課程修了後、統計数理研究所入所。2011年より情報・システム研究機構 統計数理研究所所長。専門はベイジアンモデリング。現実の問題に即した統計的モデリング、シミュレーション計算と大量データをつなぐデータ同化の研究で知られる。数多くの融合研究を推進し、また研究者の育成にも尽力。情報・システム研究機構理事。データサイエンティスト協会顧問。

ビッグでないデータの解析方法を設計・開発する

まずはビッグデータにかかわるサイエンスについて、その概況をお伝えしよう。最先端では何が起こっているのか、またどんな開発が注目を集めているのだろうか？

喜連川：今一番面白いのは、スモールデータをどうやってビッグにするかという課題です。例えば日本に台風が襲来するのは年間およそ10回ですから、10年経ってもたった100個しか学習するデータがありません。その100個でどうやって勝負をするのか。つまり本質的にデータが少ない領域でどうするかが、今一番大きなチャレンジだと思います。

樋口：レア・イベント（稀な事象）と言って、そういうところこそリスク分析やイノベーションの卵になる。統計数理には伝統的に、数の少ない実験でも仮説が検証できるようにする「実験計画」の手法がありますが、今はむしろ実験、観測・観察をもっとアクティブに、一体化したようなシステム作りが重要になってきていますね。

喜連川：例えば梶田隆章教授（東京大学）は、非常に少ない実験データを解釈することでニュートリノに質量があることを発見して、ノーベル賞を受賞された。この「解釈」というのは非常に特別な解析であって、よくあるAIのパターン学習とは全然次元が違うわけです。もちろん現象そのものは物理則に従っているので、一歩一歩精緻に検証していくことができます。でも、世の中の現象はなかなかそうはいきません。例えば、われわれの最近のテーマの1つに稀少疾患があります。糖尿病、高血圧などの患者が100万人もいるような疾患であれば、現在はもうかなり解けるんですが、日本に1人、イギリスに1人といった、世界にほんの少ししかない病気をいかに解くのか。このような課題には、やはり人間の英知を結集して、いろんなことを考えながらやるわけです。

樋口：物質・材料のデータサイエンスでは、ある機能を持った物質の分子結合や構造はわかっても、実際に作るのは難しくて、そこがボトルネックになっています。仮想実験のシミュレーションを走らせるとか、何らかのパターンを得ようとかいろいろな試みがありますが、今のところこのプロセスではまだ経験値のほうが優位という印象です（笑）。

喜連川：つまり、実際の作り方を発見するために、構造を探索した計算手法をどう拡張すればいいのかが、自明でないんですね。また自然言語処理においても、大量のデータから深層学習させてできることは限られてきています。では次にどんな切り札を入れるのかというフェーズに入ってきていると思います。

データサイエンスで解けるもの、解けないもの

気象のような地球環境から、素粒子物理学、医療分野、材料開発まで……このようにデータサイエンスは今、とても幅広いシーンで必要とされるようになってきた。研究領域にとどまらず、商業的な利用も活発で、イノベーションの引き金という期待もかかる――。

喜連川：科学者の立場から言うと、データサイエンスで解ける領域がある一方で、先ほどのように解けない問題もある。学術としてはそのような限界も示していく必要がありますね。

樋口：ええ。私はふだん「内挿・外挿」という概念で説明しています。データサイエンスは、基本的に既に獲得したデータの範囲内で推論する、つまり内挿です。一方、データの範囲外にある事象を予想しようというのが外挿です。内挿の手法がどんどん高度化して、AIも大きく発達してきたわけですが、ではAIは外挿にどう応えていくのか？　これは非常に足りない部分であって、現在のデータサイエンスのある種の限界です。それからデータサイエンスではいろいろな相関関係を導けるけれども、因果に関してはまだほとんど…。

喜連川：無理ですね。

樋口：ええ。外挿というのは言わば「予想外」の事象ですから、例えば地震がどの瞬間にどこで起きるか当てるといったことは、少なくともデータサイエンスで、あるいは今われわれの手にあるシミュレーションではできない。しかし観測網の豊富なデータや高速計算を駆使して、だいぶ定量的なリスク評価ができる時代にはなってきたんです。

これからの社会で大事な素養とは？

さて、ビッグデータの未来はこれから、どういう社会になっていくのだろうか。今度は発達する技術を受け入れる側の人間と、その教育について訊ねた――。

喜連川：科学技術の進歩が過去に比べると非常に速くなっていて――「収穫加速の法則」と言いますが――、やがて技術的特異点（シンギュラリティ）を迎えると言われていますね。ここで一番難しいのは、人間がそれをどう咀嚼するか、社会の変化の中でどう受け入れていくか、そのスピードが遅いということだと思うんです。例えばネット上の著作権法の改正をするのにも、気が遠くなるほど時間がかかる感じがします。社会に有益な技術を採用しても、人間が腑に落ちないままでは、社会がどんどん混迷していくでしょう。人の理解のプロセスを何とか加速化できないかと考えます。

さかのぼれば、文明を大変換あるいは破滅させるのは、いつも最後の最後、1年で言えば大晦日の1秒ぐらいの時間で起こっています。その現象はまさに想定外の、何とも言えない妙な人工物が引き金となって、戦争になり、不安が広がり……というシナリオだったのだと、現時点からは見えてくる。すると今、技術だけが、人間が判断する間もないうちにどんどん発展しているとすれば、近未来に大きな大変化が見込まれるような、まさに、そういうことが起こっているんじゃないかと……。

樋口：全く同感ですね。どう人を教育するか……。ちなみに、小・中・高校におけるプログラミング教育については、どうお考えですか？

喜連川：日本では日本語をまず勉強します。ちょっとしてから英語を、大学に入ると第2外国語を学びます。プログラミング言語は、少なくとも日本語以外の言語と同じぐらいのレベルの市民権を、与えられるべきではないかと思います。例えば英語を話せると世界中のかなり多くの人々と話が出来、相手をより深く理解でき、それはまた人生をすごく豊かにすることもできますね。プログラミング言語は、自分が思ったものを実際の「モノ」にすることができます。ITの講義の最初の授業で、いつも「コンピュータが入っていない人工物を持って来い」という問いかけをしているのですが、今、コンピュータが入っていない物はほとんどないんです。そんな中で、プログラミングができなけ

れば何も作れない、あるいは会話に加われないということも出て来るでしょう。この意味で、プログラミング言語は人間にとって非常に基礎的な素養の1つだし、自分の思ったものを作る能力を若い頃から育てることは重要だと私は考えています。

樋口：身の回りを見れば、こんなにもデータに取り囲まれた世界に私たちは生きています。自分の手でデータを分析してみて、解釈して、明日の生活に役立てるといったことを、小さい頃から体験して欲しいですね。統数研では、中・高校生を対象に「統数研データサイエンス・ハイスクール」を実施しています。機械学習や統計の凝った手法を習う機会のように思われるかもしれませんが、狙いはそうじゃないんです。データサイエンスの一番大切なところは、実は着眼点や課題設定であることを、学んでもらいたいと思って取り組んでいます。

公開日：2018/06/11

大学共同利用機関法人
情報・システム研究機構について

平成16年、すでに大学共同利用機関として活動していた国立極地研究所、国立情報学研究所、統計数理研究所、国立遺伝学研究所の4研究所が結集し、大学共同利用機関法人情報・システム研究機構が誕生しました。全国の大学等の研究者コミュニティと連携して、極域科学、情報学、統計数理、遺伝学についての国際水準の総合研究を推進する中核的研究機関を担うとともに、21世紀の重要な課題である生命、地球、人間、社会など複雑な現象に関する問題を、情報とシステムという視点から捉え直すことによって、新たな研究パラダイムの構築及び新分野の開拓を目指しています。

平成28年度、当機構は 4研究所に「横串」を貫く組織改革を行い、「データサイエンス共同利用基盤施設」を設置しました。これにより、データ共有支援、データ解析支援、データサイエンティスト育成の取り組みを一層強化し、社会のイノベーションにつながるデータ駆動型科学の発展を推進しています。そして研究者コミュニティの要請に応える共同利用・共同研究により、大学等における研究の発展に貢献するとともに、産業界との連携や、市民が参加するオープンサイエンスも進めています。また総合研究大学院大学の基盤機関として、もうひとつの重要な使命である人材育成にも取り組んでいます。

情報・システム研究機構は、各研究所の学理の追究に基づき、データサイエンス時代の新しい研究パラダイム構築を通じて、現代の課題解決や超スマート社会構築等の社会の要請に応えてまいります。皆様の一層のご支援、ご協力を心よりお願いいたします。

機構長　藤井良一
2019年3月

『情報・システム研究機構ブックレット』について

本機構は、新たな研究パラダイムの構築と新分野の開拓を推進し、
また大学共同利用機関法人として、大学等の研究の発展に貢献しています。
『情報・システム研究機構ブックレット』は、
その研究と貢献をわかりやすく紹介していく、シリーズ小冊子です。

情報・システム研究機構ブックレット 3
SCIENCE REPORT 013-018
データサイエンスでここが変わる。

著　者　　大学共同利用機関法人 情報・システム研究機構
監　修　　樋口知之（統計数理研究所 所長）
取材・文　池谷瑠絵（情報・システム研究機構 URA広報）
写　真　　飯島雄二（Science Report013-018）、池谷瑠絵（p37下）
デザイン　ヤマノ印刷株式会社
シリーズデザイン　hata design
発　行　　大学共同利用機関法人 情報・システム研究機構
〒105-0001　東京都港区虎ノ門4丁目3番13号
ヒューリック神谷町ビル2階
TEL：03-6402-6200　FAX：03-3431-3070
http://www.rois.ac.jp/
発行日　　2019年3月15日

ISBN978-4-909638-10-6 C0040
Printed in Japan

大学共同利用機関法人
情報・システム研究機構
Research Organization of Information and Systems

ISBN978-4-909638-08-3 C0040