

Data Centric Science Research Common 018-029

未来へつながる データサイエンス

データ中心科学 リサーチコモンズ研究紹介

北川源四郎

松井知子・吉田 亮

宮尾祐介

佐藤 薫

樋口知之

山下智志

植竹 淳

新井紀子

丸山 宏

武田英明

持橋大地

城石俊彦

栗木 哲・藤澤洋徳

未来へつながる データサイエンス

データ中心科学 リサーチコモンズ研究紹介

未来へつながるデータサイエンス データ中心科学 リサーチコモンズ研究紹介 Research View 2015 – 2016
編集: 情報・システム研究機構 URA(広報)
取材・文: 池谷瑠絵
写真: 水谷 充、ERIC、北岡 稔章
ブックデザイン: 畑 友理恵 + 大崎奏矢
発行: 大学共同利用機関法人 情報・システム研究機構 Inter-University Research Institute Corporation Research Organization of Information and Systems
発行日: 2016年2月22日 第1刷 2016年4月20日 第2刷
印刷: シナノ印刷株式会社 ISBN978-4-909638-05-2 C0040

はじめに 北川源四郎 p.4	Research View 2015-2016
このブックレットのみかた p.6	松井知子・吉田 亮 p.8
	宮尾祐介 p.12
	佐藤 薫 p.16
	樋口知之 p.20
	山下智志 p.24
	植竹 淳 p.28
	新井紀子 p.32
	丸山 宏×第3回「IT×災害」会議 実行委員会 p.36
	武田英明 p.40
	持橋大地 p.44
	城石俊彦 p.48
	栗木 哲・藤澤洋徳 p.52
	プロジェクト紹介 p.56
	研究者紹介 p.80
	編集後記 池谷瑠絵 p.88

データサイエンスの時代を先導する役割を目指して

情報通信技術の飛躍的発展によって、研究分野や社会において複雑・多様なデータが爆発的に増加し、その有効活用が今後の研究発展や産業におけるイノベーションの鍵となっています。産業革命が人間から多くの肉体労働を解放したのと同様に、ビッグデータの登場は人間の知的労働を大きく変えるものと予見されており、人類の社会はいま、新たな歴史的転換点に直面しています。やがて、データ中心の社会が出現し、すべての研究はデータサイエンス化するといっても過言ではありません。ただし、それを実現するためには、研究基盤、科学的方法論および人材育成のすべてについて格段の変革が必要です。

実際、ビッグデータは、膨大である反面、一般にその価値密度は低く、複雑・多様なデータから重要な情報や知識を取り出すことはそれほど簡単ではありません。この課題を解決するためには、科学的方法論の観点から言えば、従来の実験科学と理論科学、さらには近年急速に発達したシミュレーションを中心とする計算科学に加えて、大規模データからの知識獲得のための第4の科学——すなわちデータサイエンス——の確立が不可欠です。一方、研究基盤の観点からは、データと知識の共有・統合・活用による知識発展を目指すオープンサイエンスを視野に入れたデータ共有基盤、およびビッグデータからの知識獲得のためのデータ解析基盤の構築が欠かせません。そしてデータサイエンスを推進するためには、データサイエンティストの育成も大きな課題です。この新しい人材には、データ解析、データ可視化、ビッグデータ処理技術などの要素技術と領域分野に習熟するだけでなく、問題の本質を見抜いて課題解決する力や研究倫理やセキュリティの能力も求められます。

このような認識に基づき、情報・システム研究機構では、これまで3年間にわたって「データ中心科学 リサーチコモンズ」事業を推進してまいりました。この経験と実績を活かし、2016年度から始まる法人第3期においては、データサイエンス共同利用基盤施設を設立し、機構が強みを持つ生命科学分野、地球環境分野、人間社会分野を中心にデータ共有支援事業を発展させるとともに、ゲノムデータ解析支援やデータ融合計算支援などのデータ解析支援事業を推進し、我が国全体のデータサイエンス化に貢献したいと考えています。

リサーチコモンズを発展させ組織化したこの施設は、他の研究所と同様に運営会議による開かれた運営体制を持ち、人事評価も人事自体もデータサイエンティストの仕事が正当に評価され、データサイエンティストがキャリアパスを形成できるような組織にする必要があります。データサイエンスの時代においては、データを中心に新たな研究が過去の仕事や他の領域とつながり、そして人類の知識が発展していくことが想定されます。論文の形ひとつをとっても、それが可能になるような形へと変わっていかなくてはなりません。リサーチコモンズの経験を活かしつつ、新たな決意のもと、当機構はデータサイエンスの時代を先導する役割を目指してまいります。

大学共同利用機関法人
情報・システム研究機構
機構長 北川 源四郎



凡例 このブックレットのみかた

記事の通し番号• Research View 018

記事タイトル• **データから新事実を
探し出すための可視化。**

この研究者が紹介します• 統計数理研究所・教授
プロジェクトディレクター
松井知子

研究プロジェクト名または研究データベース名• **Project**
イメージデータ解析

同プロジェクトまたはデータベースに属する他の
記事で、関連項目がある場合は、ページ番号を
示します。2015年刊の冊子の該当ページを示し
たものもあります。•

..... Research View014 p.60>



Research View 018

データから新事実を 探り出すための可視化。

統計数理研究所・教授
プロジェクトディレクター

松井知子

統計数理研究所・准教授

吉田 亮

Project

イメージデータ解析

Research View 018

INTRODUCTION



可視化技術のノウハウを 蓄積しよう

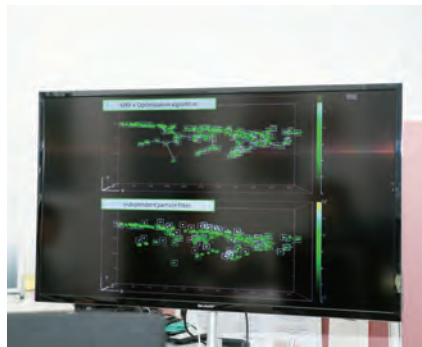
とにかくいろいろなデータを使って複雑な現象の可視化ということが、私どものプロジェクトの狙っているところです。このプロジェクトでは生物系、神経系、医療関係、それから人間社会というように、異なる分野の研究者たちが集まっていますので、まずはそれぞれ独自に可視化の技術を開発して、そこから互いに利用できるノウハウを蓄積しようと考えています。「イメージデータ解析」の重要な役割のひとつは、大量のデータを人が理解できるようなかたちにして、そこから人の目で新しい事実やしくみを推測できるようにすることだと思います。つまり人が見て分かなければ、それをもとに新しいアイデアを考え出すことは難しいと思います。「イメージデータ解析」は、このように新しい知見を発見しやすくして科学の進展に役立ちたいと思っているプロジェクトです。その具体例として、さっそく吉田准教授の研究をご紹介します。（松井）

生物、人間社会、宇宙の大量のさまざまなデータが駆使できるようになった現在。しかしデータが増えれば増えるほど、また時間的、空間的に変化したり、いろいろな情報を合わせ持つようになればなるほど、それらの中から重要な情報を抽出する難しさと手間は増していきます。さまざまなデータから、人間が見て直感的に重要な情報を取り出せるようなかたちにしていこうという「イメージデータ解析」プロジェクト。いま世界中で活発な研究が進められているこの分野に取り組む、プロジェクトディレクターの松井知子教授とサブリーダを務める吉田准教授にお聞きしました。

線虫の全中枢神経系の 活動状態を可視化する

私が扱っているのはライフサイエンスデータで、具体的には線虫の神経回路です。線虫は分子生物学において非常に重要なモデル生物で、神経系を構成する神経細胞（ニューロン）302個のつながりが解剖学的にすべて分かっています。記憶、学習、意思決定などを行う線虫の中枢神経系は、人間でいえば脳にあたります。私たちは、生きた線虫の全神経細胞の活動状態をリアルタイムで計測できるイメージングの技術を作り、現在、最大180個の神経細胞の情報が取れるようになりました。線虫に臭いなどの刺激を与えると、線虫の中にあるニューロン間で情報のやりとりが起きます。この様子を、ニューロンの核内に存在するカルシウムイオン濃度の変化を測定して可視化し、同時に共焦点顕微鏡の立体動画で捉えます。そして画像に含まれる輝度情報を取得し、個々のニューロンの動きをできるだけ見失うことなく、正確に追跡していきます。特に、線虫の頭部には多く

のニューロンが密集しているため、これまで線虫が不規則に動くときと正確な追跡が非常に困難になるという問題点がありました。そこで私たちはベイズ統計



このイメージ解析プロジェクトは、JSTのプログラムCRESTの研究課題『神経系まるごとの観測データに基づく神経回路の動作特性の抽出(代表者:飯野雄一(東京大学))』の一環として、東大・九大の計測グループ、神経回路のシミュレーション・モデルをつくる茨城大のグループと連携して行われている。このうち吉田准教授が、データ解析とモデル統合を担当する。「モデル統合では、モデルとデータを合わせて、データの見えない部分にある状態を推定するという逆問題を解く課題に取り組んでいます」という。



学のアプローチに基づき、密集する多数のニューロンの動きを高い精度で追跡するトラッキング手法を開発しました。(写真内、上段の解析結果)。この技術により、刺激に対する神経系のシステム応答を直接観察できるようになり、神経回路の動作原理についていくつかの新しい事実が見つかってきました。(吉田)

バイオイメージインフォマティクスの革新

ライフサイエンスのイメージング技術は過去10年で急速に進歩しました。細胞や分子のイメージング技術が発展し、生物や医学の研究において静止画や動画データの解析手法が欠かせないものとなり、機械学習やコンピュータビジョン、統計科学の先進技術が急速に導入されています。このような研究分野は、バイオイメージインフォマティクスと呼ばれ、データサイエンスや機械学習の人たちがどんどんこの分野に参入している状況です。データから何ができるかを考えるには、統計やデータサイエンスのセンサが非常に重要な役割を担います。統計科学の独自

Research View 018



の視点から問題に対するユニークな切り口を発見し、ライフサイエンスの新しい方法論を提案していくことが私たちの仕事です。(吉田)

隠れた情報としての気持ちを 探るソフトウェア

目に見えないたくさんのデータを人間の目で分かるかたちに変えることからもう一歩踏み込んで、今見えているデータを生み出すさらに元にある心理を探る技術にも注目しています。たとえば人間の動きには、気持ちが明るいとスキップしたり、気持ちが沈んでいれば静かに歩いたりというように、その裏に何らかの気持ちがあってそのような動きになっているということが多くですね。そこで歩き方の特徴から、いわば隠れた情報としての気持ちを統計的に推定し、より確実に動きを識別するようなソフトを作りました。今後、このソフトをいろいろな識別問題に適用して、それぞれのデータの背後にある隠れた情報を探っていききたいと考えています。(松井)



サブリーダーを務める吉田准教授(写真 左)が大学時代に在籍したのは、なんと経済学部。「経済学部で統計をずっとやってきて、今はライフサイエンスです」という。一方、松井教授は情報科学から人工知能・機械学習へ。しかしその根本に、共通して「ある統計的な規則に従って学習してやれば、自動的に入力と出力の関係が出てくるのが面白い」という関心を持つ。この大量データから入出力の関係を自動推定する「モデル選択」技術は、コンピュータの発達に伴ってますます高度化し、研究が活発化しているという。

▶プロジェクト概要
<http://rc.rois.ac.jp/research/a01/>



Research View 019

人間のしわざには 類型がある。

国立情報学研究所・准教授
プロジェクトディレクター
宮尾祐介

Project

メタ知識構造解析

Research View 019

INTRODUCTION



人は勝手に文の間に 因果関係を読む

人は、同じ意味を伝えるにもいろいろな言い方をします。例えば「大雨になる」と「強い雨が降る」は表現は違うけれども同じ意味を表しています。このように言葉の表層は違っても意味は同じ／違うということを機械に理解させる研究をずっとしていたところ、それだけではうまく表せない意味があることに気づいたんです。例えば「外へ出かけてごはんを食べる」という場合、「ごはんを食べる」は「外へ出かけて」の目的です。ところが言葉の表面だけ見ても、目的であることは直接書かれていません。「ごはんを食べることを目的として」と書いてあるわけではない。「雨が降って地面がぬれている」という表現も、文の構造としては文が2つ並んでいるだけなのに、人間は勝手にその間に「雨が降って→地面がぬれている」という因果関係を読んでしまいます。文の間にある

人類が日々生み出すデータの多くは「テキスト」という形式で蓄積されています。機械が読む「プログラミング言語」に対して、私たち人間が使う言葉を「自然言語」といいますが、自然言語で書かれた例えば新聞記事や学術論文等のテキストから、その波及効果や成果を客観的・定量的に把握する方法は、今のところありません。そこで、私たちがやりとりする言語活動を一段掘り下げたところにある論理や因果関係といった「メタ知識構造」を探りだそうというのが「メタ知識構造解析」プロジェクト。語や文の連なりの関係性だけでなく論理的な構造を取り出して計算する、国立情報学研究所 宮尾祐介准教授の「深い解析」を核に、解析手法を横断的に駆使する自然言語処理のプロジェクトをご紹介します。

このような目的・理由・原因・結果・手段のような関係性は、自然言語の中には多く含まれているのですが、今まであまり研究されてきませんでした。これをどう整理し、どう解析すればいいかを明らかにしようというのが、このプロジェクトのきっかけです。

さまざまな立場を持ち寄っての議論

私たちの研究は、実際のデータを見て、データが示す現象を観察するところから始まります。言語学が一般に、何らかのテーマに沿ってピックアップしたデータを分析するのに対して、私たちの自然言語処理では、基本的に網羅的にデータを解析し、頻度の高い現象から始めて対象範囲を徐々に広げていくアプローチをとります。プロジェクトの飯田龍主任研究員(情報通信研究機構)は、数千〜2万ぐらいの新聞の文章を対象に、その間にどういう表現があるという関係になりやすいかを分析しています。

戸次大介准教授(お茶の水女子大学)のグループは、実例を1個1個見ながら、例えば原因・結果といった関係をどういう論理式に書くことができるかという研究を進めています。この式は、表層的にはもちろん言語によって異なりますが、関係性の意味を表現できればおそらく言語に依存しないため、日本語での解析が進化したら英語などにも適用する予定です。また持橋大地准教授(統計数理研究所)は、恣意的な要素を一切排除して、基本的にはすべて統計的に解析できるという立場で研究を進めています。「深い解析」という私の立場とは違いますが、月1回のミーティングでも、いろいろ議論すべきことがあります。ちなみに戸次准教授と持橋准教授は、自然言語処理の方法論としてまさに対極にあり、そんな2人が一緒に議論する「奇跡的」なプロジェクトだとも言えるでしょう。

ツリー構造からグラフ構造へ

このような解析をもとに、私のグループでは情報科学の論文を対象に、「深い解析」によって内容を理解し、欲しい情報を獲得することに取り組んでいます。



情報科学の分野では、例えば手のひらに何かを映写してデバイスとして使う新技術などのように、世の中のさまざまな事物に、新しい使われ方や役割が与えられるケースが多いんですね。

そこでやはりこの場合も、論文の中に目的・手段・結果といった関係性がどのように記述されているかを形式化するのが効果的なのです。そしてその関係性は、文の構造解析で使われる、動詞句、名詞句などに枝分かれしていく木構造よりも、むしろ一般的なグラフ構造になっています。つまり、文章というデータの背後にもうひとつの構造を予測し、そこにグラフ構造を発見するという問題になっていることがわかってきました。データの分析を通じてどういうグラフになるかという問題を定義したら、いろいろな手法を適用して繰り返し分析し、自動解析の精度を

Research View027 p.44

Research View 019



上げていきます。現在は、これを賢い論文検索などのアプリケーションとして使おうという段階に入ってきています。

人間の「論理的思考」には二通りある？

人間はよく「この文章は論理的だ」と言いますが、何をもって「論理的」かは、実はまだ誰にもわかっていません。ふだんあまり意識されていませんが、この「論理的」は、数学の一部としての論理学とは全く別のものであります。むしろ、今このプロジェクトで注目している理由・手段・原因・結果などがつながることを、人は「論理的」だと考えている可能性があります。例えば「風が吹けば桶屋が儲かる」の論旨にはギャップがありますが、それを埋めるのは語の関係性というよりも、「風が吹くと砂が舞うものだ」といった一般的知識のようなものです。人間はふだん何ステップぐらい埋めているのか？——このような研究は未開拓だと言えるでしょう。

言語の使用、中でも論理的な思考は、他の動物にはない人類だけが持つ最高の知であると、長い間考えられてきた。「しかし」と宮尾准教授は言う。「言語データをよく見ると、単に人間にとって便利な関係とそうでない関係があるだけかもしれませんが。つまり手段、理由、原因などは人間が生きていくために必要な関係だからこそ重要だと考えられているのかもしれない。もし宇宙のどこかにこのような情報が要らない世界があったら、たぶん全く違う情報のやりとりのしかたがあるのではないかな……。人間の言語処理はたまたま進化の過程で獲得したものであり、今私たちは、それを後から発見しようとしているのだ、と考えています」。

▶ プロジェクト概要
<http://rc.rois.ac.jp/research/a02/>
 ▶ 研究室ホームページ
<http://kmcs.nii.ac.jp/mylab/>



Research View 020

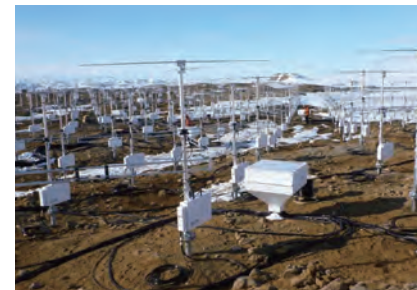
データ中心科学の 花ひらく冒険。

東京大学・教授
プログラムディレクター
佐藤 薫

Database
地球環境データ

Research View 020

INTRODUCTION



PANSYレーダー

「PANSY」という花の名前を冠した、南極昭和基地大型大気レーダーのプロジェクト。そもそもは2000年1月、極地研に着任してきたばかりの研究者が「南極でこんなデータが取れたら、きっと素晴らしい」と夢を語ったことから始まりました。同年5月には、

理論にたえる大量データを 生み出すようになった観測機

地球物理学は、たとえば素粒子物理学などのように、高度な数学を駆使して現象を理論的に突き詰め、非常に緻密な実験によって理論を検証していく分野とはかなり違います。かつては1、2回の観測から得られたごく少ない情報だけで議論が進められることもありましたが、しかしPANSYは、「理論にたえるデータが出せる」観測機なんです。鉛直分解能150メートル・時間分解能1分の高解像度を持ち、データを見ただけで「何だろう、これは？」と思わせるような、大気の様子が浮かび上がってくる。仮定を置くことなく理論と結び付き、新しい発見が生み出せるような、非常に大量なデータなんです。

当時最先端の京都大学MUレーダーを手本に、その設計者や極地研のメンバーが集まって、核となるプランが一気に作り上げられたといいます。そして2002年には、朝日新聞の一面や科学雑誌『Nature』の取材記事として報道されるなど、世界的にも大きな注目を集めました。しかし実際にプロジェクトが動き始めたのは、発案から約10年後。そして2011年、念願の初観測が実現しました。ところが、そのあと南極は記録的な大雪に見舞われます。しかも2011～12年、南極観測船「しらせ」は2年連続の接岸断念という非常事態に……それでも「PANSYは面白い」と言う、東京大学の佐藤薫教授。佐藤教授こそ、2000年に新任助教授としてPANSYプロジェクトを発案し、2005年からはその代表として研究を推進してきた、その人です。リサーチコモンズ「地球環境データ」プロジェクト(2013年度～)に関わる話題を中心に、佐藤教授にお聞きしました。

高解像の観測・モデル・理論を むすぶデータ中心科学

一方、大気現象を解くためには「モデル」が使われます。大気物理学では、たとえば天気予報にもそのためのモデルがあって、スーパーコンピュータを使って大規模なシミュレーションが行われています。このようなシミュレーションのデータと、観測が生み出すデータは、どちらも大量なため直接比較することができません。さらに、このような高精度なデータから解明が期待される現象には、その礎となる気象学的な理論自体も作らなければならない場合がほとんどです。これは、大量データから法則を発見するのとは違って、第一原理としての運動方程式から組み立てていく、いわば“紙とペンでつくる”新理論であり、それぞれのデータに適用することでその解析を大きく進展させることができます。この観測データ、シミュレーション、新理論の3つが揃ってこそ、



高解像度観測時代のサイエンスは面白くなる。そして、この3つの中心にあって、データ解析を担うのが「データ中心科学」です。私たちの研究室では今、この発展を目指しています。

解明が待たれる 大気の小さな構造「重力波」

ところで2006年4月頃、私たちは海洋研究開発機構の研究者とも共同でスパコン「地球シミュレーター」を使って、PANSYの仕様に合わせて全球3年分のシミュレーションを始めました。計算には2年を要しましたが、鉛直の広がりや80キロメートル・鉛直解像度が300メートルの仮想大気空間で繰り広げられる大気のような階層構造を再現したシミュレーションは、少なくとも7～8年間、世界一でした。中でも私たちが知りたかったことのひとつに「重力波」と呼ばれる、大気の中にある非常に小さな現象があります。「気象ノイズ」と呼ばれていたのですが、80年代の半ばにこの効果を計算に入れたら天気予報の

精度が格段に向上し、その重要性が広く知られるようになりました。私たちはこれをシミュレーションデータの解析によって詳しく調べていたのですが、実は予想外の特徴も見えてきて、この重力波の伝播経路には集中する場所があり、それがなんと極域、まさに昭和基地の真上に位置することがわかったのです。

極域だけに見られる 美しい2つの雲のサイエンス

観測に適した南緯69.0度、東経39.6度に建設されたPANSYは、2015年3月からは、1045本のアンテナによるフルシステムで観測を行っています。省電力に徹してつくったため、連続した大規模観測データが長期間にわたって取れるレーダーであることも、世界的に際だった特徴のひとつです。PANSYを使って解明を進める重要テーマには重力波の他、オゾンホール、南極特有のカタバ風、宇宙からのエネルギーや物質流入に関わるオーロラ、極域固有の2つの雲

などがあります。雲の1つは、冬の成層圏約20kmの高さにできる「極成層圏雲」です。ピンクやオレンジ色に輝く美しい雲ですが、その表面でフロン起源の物質を化学変化させオゾン層を破壊する恐ろしい作用があり、この雲の量がオゾンホールの大きさを決めています。もう1つは最も高い約90kmの上空で夏に発生する「極中間圏雲(夜光雲)」です。1885年以前の記録がないことから「人間がつくった雲」と言われ、雲の生成量が敏感に温度に反応して気候変動を映し出すことがわかっています。PANSYを中心に、昭和基地にあるさまざまな光学・電波(中波、ミリ波、レーザー等)観測をつないで、たとえば重力波の持つ風と温度、鉛直構造と水平構造をそれぞれ同時に捉えるなど、南極大気現象の多角的理解を進めています。



2003年に南極昭和基地で行われた直径32m、高さ6μmの大気球によるオゾンホール観測。2種類の測定器をつないで地上から高さ40kmまでのオゾン濃度を高解像度で測る。定期的に行われているゴム気球観測と比べると大がかりだが、より高い高度まで観測できて、オゾン層全体をとらえられるのが特徴だ。(写真右端が佐藤薫)

北極と南極をつなぐ、 観測機をモデルでつなぐ

南極における大気現象の解明は、グローバルな大気循環の出発点や終着点になっている点で重要であることはもちろんですが、南極の研究で北極の大気が理解できることもあるばかりか、北極と南極の大気状態は連動していることなども徐々にわかってきています。そこで今、取り組み始めている新しい夢は、

地球上にあるいくつかの大型大気レーダー拠点をつないで同時観測しよう、ということなんです。非常に小さな現象を捉えようというわけですから、局所的なものであって、同時観測しても意味がないのではないかと……と最初は思っていました。しかし2006年に開発された私たちの全球モデルを土台にして、もしこれにリアルな地球の観測データをつなぎ、地上から中間圏までの大気状態を再現したら、画期的なシミュレーションになることは間違いありません。ちなみに北極の成層圏で、数日で50度ぐらい気温が上がることのある「突然昇温」という現象が知られています。起きる頻度は3年に2回ぐらいですが、いったん起これば中緯度の状態も変わり、赤道も変わり、南極も変わります。このようなダイナミックな大気循環を世界同時の高解像度観測によって捉え、地球規模の気候変動機構の解明に貢献したいと考えています。



「私たちは、南極の厳しさにすごく苦労したんですけども、自然はすごく豊かで、逆に助けられてもいるんです。たとえばPANSYは、フルシステムでなければ中間圏の観測はできない仕様です。しかし、しらせが2度も接岸せず、5分の1のシステムで観測しなければならなかったため、何も受信できないのを承知でビームを向けたところ、なんとデータが取れたのです。極域では、太陽風の活動などによって高エネルギーの電子や陽子が降り込み、中間圏の大気が電離するため、これが原因ではないかと考えられます。もう15年もPANSYに取り組んでいますが、落ち込んだり、舞い上がったりの繰り返して、本当に飽きないプロジェクトです。」

▶ データベース概要

<http://rc.rois.ac.jp/database/d01/>

▶ プロジェクトウェブサイト

<http://pansy.eps.s.u-tokyo.ac.jp/>



Research View 021

すごいチーム作りを しくみで支えるために。

統計数理研究所・所長
前プロジェクトディレクター
樋口知之

Project

データ同化・シミュレーション支援技術

Research View 021

INTRODUCTION



最先端研究を支えるチームワーク

ビッグデータ時代といわれる現代のサイエンスの進め方では、チームワークなしに最先端研究を行うことはできません。今や、1人でやる最先端なんて、ちょっとあり得ない。そのような時代には、「チーム作り」が最先端研究を推し進める、非常に重要な力になります。まとめる力、企画する力といったいろんな役割を担う人材が必要であり、キャリアパスの多様化も進めていかなければなりません。たとえばアメリカでは「サイエンティフィック・プログラマー」という職種があって、外部資金などによる研究プロジェクト単位で採用され、プロフェッショナルな優れた仕事をしている。雇用の仕方の改善ひとつで、研究はすごく進むのだという一例ですね。

属人的発展から組織的人材育成へ

リサーチコモンズの前身にあたる第1期の新領域融

新領域融合研究センター発足当時から、現リサーチコモンズの「データ同化・シミュレーション支援技術」に連なるプロジェクトを牽引してきた樋口知之教授。現在は統計数理研究所所長として、「確率に関する数理及びその応用の研究」を担う日本で唯一の研究拠点の絶えざる自己改革に取り組んでいます。その核心にある共同研究・共同利用の促進や、データサイエンティストなど人材育成の取り組みは、データ中心科学を推進するリサーチコモンズとも共通のもの。また近年は、最先端研究の“質”を支えるしくみづくりとして、統数研を拠点とした国際的な研究ネットワーク「NOE(Network Of Excellence)」も拡充しています。ますます「データ中心」がキーワードになるこれからの科学プロジェクトにおいて、研究機関が持つどんなしくみが推進力になるのか？
—— お話は、樋口知之所長(統計数理研究所)です。

合研究を振り返ると、当時4つの研究所が初めて一緒になって、まったく手探りの状態から体制作りが始まったと記憶しています。苦労もあった代わりに、走り出してからチームワークには手作り感もあった。自分の研究力にどれだけの価値があるかはわからないけれども、分野を越えるのが面白いからチャレンジしてみようという人たちが集まって、頑張ったんですね。このような融合研究のやり方を引き継いだリサーチコモンズ事業は、プロジェクト・マネジメント面でスムーズになった一方、反省点もあるように思います。つまり次の第3期中期目標期間以降は後継者を見つけてくるようなやり方ではなくて、やはり融合的な人材を生み出せる体制を、意識して作っていかねばならないのです。

統数研の役割をもっと明確に、 もっと高度に。

しかも機構が設立された10年前とは異なり、わが国



うな工夫をしていく。2番目は、統計的に考えることができる人材を育てる「統計思考力育成事業」の強化。そして3番目は、統計数理をグローバルにネットワーク化していくことです。たとえばここ1、2年で、統数研は海外機関とのMOU(Memorandum of Understanding: 学術交流の覚書)等を多数締結しています。

変動し続ける対象を 予測・解析・制御・最適化する

と近隣諸国との摩擦は増す一方です。現代の複雑な社会におけるさまざまな現象や問題と結び付いて、われわれは統計的に予測し、リスクなどの数理解析をする。さらに制御の方法を考えて、政策に役立てたり、最適化手法の改善によって限られたリソースを活かしていく。この役割をよりよく果たしていくために、統数研はいま3つの目標を掲げています。1番目は公募型共同研究の機能強化です。チーム作りの中に、あらかじめ結果が予想できないような面白いもの、もっとチャレンジなものを拾い上げるよ

統計学は現在、たとえば薬剤の有効性や安全性の基準、ファイナンスのリスク判断、工場の品質管理などさまざまな場面で、データ扱いの基礎にあって合理的な判断の根拠を提供しています。しかし統計とは、それだけではありません。私の考える「データ同化」のチャレンジのひとつは、フォワード(前向き)推論としてのシミュレーション計算とバックワード(後ろ向き)推論としてのデータ解析をつなぐこと。もう1つは、静止・固定した事象ではなく、常に変動し続けるダイナミックな現象を対象として予測・解析・



Research View 021



制御することです。現代の科学が対象とする複雑な現象には、必ず何らかの不確実性があり、これを避けることはできません。このような私の「統計」観から展望すると、現在の機械学習には2つのタイプがあり、ひとつはベイズ推論を含め、人間の脳が行う情報処理に近づいていくとするディープラーニングを含む人工知能。もう1つは、極限まで数理に立脚して、これまでの膨大なデータからはけっして生み出せないようなひらめきを導く可能性のあるものです。「データ同化・シミュレーション支援技術」でも、この両方に取り組んでいきたいと考えています。

データ中心科学をイノベーション に活かす鍵

先日、インペリアル・カレッジ・ロンドンのデータサイエンス研究所所長であるGuo教授と会話する機会がありました。その時彼が、データ中心科学を特徴付ける4つのキーワードがあると言ったんです。それはIntelligence, Integration, Interdisciplinarity, and Interactionであると。私は膝を打ちました

「ポストドクという後進をどうするか……今、みなさん少しずつ変えようとしているけれど、若い人たちを育てることを怠った分野は、必ずつぶれます」統計数理研究所にて。

——この4つの「I」こそが、現代のイノベーションを生むもの、つまり「フォーアイ、フォーアイ(4Is for Innovation)」だということなんです。まず知性なり頭脳(Intelligence)があって、しかしブレークスルーを生み出すには、それらが個々バラバラにあってはだめで、強力に学際的(Interdisciplinary)に統合(Integration)していかなければならない。最後のインタラクション(Interaction)も、非常に重要な視点だと思いますね。それは、人と人のインタラクションでもいいし、人とマシンのインタラクションでもいい。このようにここ10年間で明らかになってきたことを、意識的にしくみの中に組み込み、育成する。データ中心科学から、より社会と結び付いた成果が生まれてくる時代が始まっています。

▶プロジェクト概要

<http://rc.rois.ac.jp/research/a03/>

▶プロジェクトウェブサイト

<http://tric.rois.ac.jp/math/>



Research View 022

社会データが生む知見を、 よりよい社会へ実装する。

統計数理研究所・教授
山下智志

Database

人間・社会データ

Research View 022

INTRODUCTION



社会データを社会に活かす道筋

個人や法人の信頼性に関わるような金融データは、当然のことながら非常に厳しく管理されなければなりません。研究者としては、データを集めるにしても、誰かにご提供いただくにしても、まず自分が信頼されなければ何も始まらないんですね。いかに信頼されるか、いかによい人間関係を構築していくか——そんな課題を持つ研究に長く携わってきました。金融機関の方にデータをいただくならば、研究者としてどう貢献できるのか。たとえば銀行などでも昨今、かなり高度な統計に取り組んでいますが、われわれはそれよりも圧倒的にいいものができなければなりません。また、そのツールを活用してコンサルティングの要請に応じてこそ、新たな信頼関係を築くことができます。統数研に聞けば、どんなデータも揃うんだという認識を持っていたくことも重要です。金融機関やデータベンダー

リサーチcommonsの「社会コミュニケーション」と「人間・社会データ」にまたがる研究を推進する、統計数理研究所 リスク解析戦略研究センター長の山下智志教授。従来、企業の財務データだけに頼っていた貸倒予測を、担保や保証などを統合的にデータベース化して、実際のリスク算出に役立てたり、さらに巨大財務データベース中の欠けた値を補完する手法の開発や、政府間融資のリスク(ソブリンデフォルトリスク)の格付システム開発など、広範囲に収集した信頼度の高い金融データとその解析に多くの実績があります。プロジェクトでの取り組みについて、山下智志教授をご紹介します。

等のデータに加え、総務省の事業所統計

を始めとした各種官公庁データも
整った体制を、これからいっそう強化していく計画です。

賃貸不動産30年の収益を予測する

銀行ローンの中に、アパートを建てる大家さんにお金を貸す「アパートローン」という商品があります。銀行融資のうち10%という少くない比率を占めるのですが、他の融資と異なり、リスク評価モデルが全く存在しません。というのも、賃貸物件のどれが空いていて、どれだけ収益が上がっているかといったことはなかなか調査できないため、客観的なデータが存在しないんですね。そこでわれわれのリサーチcommonsのプロジェクトでは、賃貸不動産のデータベース作りから地道に取り組み、正確な賃貸住宅の収益予想を導く研究に取り組んでいます。



統計数理研究所・国立情報学研究所、そしてデータベンダー・不動産鑑定事務所・銀行など5者の協力体制で開発を進めています。

ウェブデータ×サーベイデータ×解析手法

まずプロジェクトの一藤裕特任研究員(国立情報学研究所)が開発した手法で、ウェブに公開されてい

る賃貸不動産の空室／占室情報を、平米数や間取りなどを含めて10日に1回入手し、蓄積します。膨大に集めることができる反面、ウェブに掲載されている情報は、実際の状況を正確に反映していない面もあります。そこでとても質の高いデータを、3カ月に1回、別途収集します。これは不動産鑑定士が実際に不動産を見に行き、エキスパート・ジャッジメントを下すという高精度な調査になっています。この2種類のデータから、空室が3カ月後に埋まる確率、部屋が3カ月後に空室になる確率をそれぞれ導く統計モデルを作ります——この3つのプロセスが、他には真似の出来ない、われわれのプロジェクトの大きな特徴になっていると言ってよいでしょう。この手法の開発をさらに進めて、20年後、30年後の空室予測につなげていきたいと考えています。

Research View005 (2014-2015) p.27



Research View 022



データ提供者と利用者を結びつけるモデリング

僕が一番実現したいと思う研究のかたちは、このように非常に多くのデータ提供者・利用者がいて、それぞれのニーズとシーズをうまく融合するというものです。なぜなら、そのような研究は、必ず世の中にインパクトを与え、研究成果を社会還元できる。このことは僕の研究への強いインセンティブになっています。長い時間がかかりますけれどね。利用開始後にフィードバックをもらって、改良バージョンを作るといったことにも取り組んでいます。工学的分野では製品が動けばわかるし、医療分野なら治ればいいのかもしれない。しかしわれわれの分野では「従来モデルより10%精度アップしました」と言っても「都合のいいデータだけでやっているんじゃないか？」という人々の疑念を払えなければ、その技術が使われることはないでしょう。そういったところで、やはり人間がキーなんです。

「大学時代、高速道路に表示する〈〇〇まで〇分〉という数字を計算するのが、私の修士論文のテーマでした。道路を水路に見立て、水理学のモデルを使うことで、不慮の事故にも対応できる精度の高い予測ができました」という山下教授。それ以来、一貫して「世の中で使われない研究はしない」という。

▶データベース概要

<http://rc.rois.ac.jp/database/d03/>



Research View 023

赤道直下の氷河で 見つかった新しい生態系。

新領域融合研究センター／
国立極地研究所・特任研究員
植竹 淳

Project

地球・環境システム

Research View 023

INTRODUCTION



氷河上一面に広がる 小さな黒いコケの塊

ルウェンゾリ山は、アフリカの他の2山と比べて降水量が多く植生がとても豊かです。頂上までは徒歩で約5日の行程で、登り始めは熱帯特有のジャングルだった植生が、登るにつれてコケと地衣類に覆われた雲霧林、アフリカの高山にしか生えない奇妙なかたちをした植物へと変化していきます。標高が4,500メートルを超えると、次第に植生がまばらな岩場になり、ついにスタンレープラトーと呼ばれる氷河が見えてきます。この氷河は長さ1キロメートル足らずで、氷河としてはかなり小さいものですが、この山地では最大の規模です。実際に訪れてみると、氷河の上にはこれまでに見たこともない厚さ1センチメートルほどの真っ黒な塊が一面に広がっており、その異様に圧倒されました。手にとってみると、これらが繊維状の細胞が絡み合っている構造で

赤道直下の熱帯にも、氷河を頂く山々があります。標高5,895メートルのキリマンジャロ山(タンザニア)、5,199メートルのケニア山(ケニア)、そしてウガンダとコンゴ民主共和国の国境に位置する標高5109メートルのルウェンゾリ山です。ルウェンゾリ山では1906年、初めての氷河観測が行われ、山頂付近で6.5キロメートル平方の氷河が確認されました。しかしこの約100年の間に氷河は減り続け、2020年には消滅するだろうとの予測も発表されています。この消滅の危機に瀕する山頂の氷河で、コケ無性芽の集合体を発見し、その成果を2014年11月にプレスリリース発表した植竹 淳特任研究員(新領域融合研究センター／国立極地研究所)。今回は、その最先端の観測と成果を紹介します。

あることがわかり、日本に持ち帰って詳しく調べるようになったのです。当初は顕微鏡で見たかたちが微生物のような単細胞であることから、マリモのような緑藻かと考えていました。しかし遺伝子解析の結果、これらが実はコケであることがわかりました。微生物かと思っていた細胞は、コケのライフサイクルの一部である『無性芽(gemmae)』と呼ばれるステージにあたることから、この塊を「氷河コケ無性芽集合体(glacial moss gemmae aggregation: GMGA)」と名付けることになりました。しかし……ちょうどチキンナゲットのようなかたちなので、個人的には「氷河ナゲット」と呼んでいます。(笑)

氷河上の生態系が 緑の大地を準備する

氷河ナゲットは、コケが光合成をして有機物を作り、それを内部でいろいろ



な物質に変換したり、分解したりする微生物が共生する「集合体」で、氷河ナゲットのひとつひとつで1つの生態系が完結しているといえます。また氷河ナゲットは、微生物が作り出す腐植などの物質により真っ黒い色をしています。黒い色は光をよく吸収するため、コケが繁殖すればするほど氷河の氷は溶けていきます。観測の結果、氷河は年間約2〜3メートルずつ後退していることがわかりました。今まで氷河に覆われていた岩の上には、たくさんの氷河ナゲット



トが取り残されて乾燥しており、その状態のまま数年を経た場所では、氷河ナゲットだったものの上に他の種類のコケや草などが生えていました。つまり氷河の上で、氷河ナゲットによって高山の土壌の元になるような物質が作られていたことになります。このような氷河と氷河後退域の生態系について、ルウェンゾリ山だけでなく、北極域でも検証を進め始めているところです。

グリーンランドの海岸線を観測する

北極圏にあるグリーンランドを衛星写真で見ると、その表面の多くが氷に覆われていることがわかります。この大きな氷床と陸地の境界線を実際に訪れてみると、氷河ナゲットよりもさらに細かい、直径約1ミリメートルの黒い粒に覆われていることがわかります。これらは「クリオコナイト粒」と呼ばれるもので、極限環境でも多く見られるシアノバクテリアという微生物が絡まってできています。ここでも冷たいと

Research View 023



ころを好む“黒い”生物の集合体が、氷河の溶解を促進しているのです。しかもこのような場所は、現在グリーンランド氷床の端で年々広がっており、その面積の大きさから溶けた水による海面上昇などへの影響も懸念されています。

大気の中に含まれる微生物をチェックする

ルウェンゾリ山でも、グリーンランドでも、氷河上の生態系を理解するだけでなく、そこに何が入り出しているのかを調べることも、今後重要なポイントになると考えています。その謎をとくひとつのキーワードが風による微生物の移動です。人や動物が運ぶのとは違って、風は面的かつ長距離に物質を運ぶことができるため、生態系に関する影響がとても大きいことが予想されます。そこで大気を採取して、その中にどのような微生物がどのくらいいるのか、遺伝子解析によって解明していく準備を進めています。この観測は極地だけでなく、東京などの都市部でもま

「消滅する恐れのある高山の氷河の重要性については、いろんな見方があります。例えば生物多様性の側面からは、今のうちに氷河の上に住む微生物のアーカイブを作っておくべきです。生態系があるということは、私たちの生活に役に立つものを作ってくれる生物がいるかもしれないことを示して、それらは氷河を持つ国の資源ともなりえます。また乾燥している地域(中国の乾燥地、ボリビア、チリ)では、生活・工業用水を氷河の雪解け水に頼っているため、水源として真剣に調査が進められています。」(遺伝子解析室にて)

だ行われていないため、まずは研究所の屋上に機器を設置して、テストを行っているところです。

- ▶ プロジェクト概要
<http://rc.rois.ac.jp/research/b01/>
- ▶ プロジェクトウェブサイト
<http://tric.rois.ac.jp/earth/>



Research View 024

人工知能は可読な ビッグデータを待っている。

国立情報学研究所・教授
プロジェクトディレクター
新井紀子

Project

e-サイエンス基盤技術

Research View 024

INTRODUCTION



情報・システム研究機構発足直後の融合研究で「サイエンス2.0」として研究開発が始まった researchmap。研究者の卵や研究支援者にも門戸を開き、共同研究を加速し、大学共同利用の基盤となるようなシステムとして2009年に公開され、以降も

融合研究のための プラットフォーム

researchmapという企画を機構の融合研究として始めた時、最初に私に招待講演を依頼くださったのが「NSF Tokyo Regional Office(米国 The National Science Foundationの東京オフィス)」だったんです。ところが講演したところ、みんなホームページを持っているのに何でもこういうものが必要なのかといったご意見が多く、全然理解されなかった。ところが程なくオバマ政権になって科学技術政策のための科学という方針が打ち出され、NSFには政策のための科学の大規模プロジェクトが作られました。そのディレクターだった Julia Lane 氏が来日したときに、researchmapのデモをご覧になって、これが自分の欲しかったものだとおっしゃったんです。現在でもまだ国全体をカバーする規模の統一的な研究者総覧が出せる researchmap のようなものを持っているのは、本当に限られた国しかないと言えます。

改良が続けられてきました。現在24万以上の登録を数える日本の研究者総覧に成長し、去る2015年9月16日には、一橋大学一橋講堂(東京都千代田区)にて、科学技術振興機構、情報・システム研究機構主催、大学IRコンソーシアム後援により、researchmapシンポジウム「自らの強み・特色を知るために～これからの大学のIRとは?～」も開催されました。「一人一人の研究者がいかにグローバルに活躍するかということだけでなく、日本の研究を国際的に競争力あるかたちでポジショニングしていくために必要なのが researchmap」と、プロジェクトディレクターの新井紀子教授(国立情報学研究所)。その研究開発が、リサーチcommons研究プロジェクト「e-サイエンス基盤技術」として進められています。(researchmapについてはこちらから <http://researchmap.jp/>)

世界に先駆けてこういうものを日本が作ったことは、ぜひ強調しておきたいと思います。

大学のリポジトリを 「機械可読」にする

どんな研究があるかを知るには、まず論文データです。多くの大学や研究者は、論文のリストを提出することが重要だと思っていられる。しかし、それは紙の時代の連想なんです。研究は論文の本数で測るものでもないし、実際関心がある研究者が本文をクリックして、その後に読むなり引用するなりといった一連のアクションが伴わないと、その研究は加速していきません。この意味で、リンクが張られていない「黒い」データはダメだ、と私は言っています。学術プラットフォームは、参照することによって連想が広がるとか、ああ、これが自分が求めている情報だな、というふうに誘導されていくような「プッシュ型」のデータベースであるべきです。この点、



researchmapは現在JSTさんが運用していますから、今後たとえば競争的資金に関するJSTのデータベースとリンクする、といったことが実現できれば、研究者がresearchmapに自分の研究情報を蓄積する大きな動機になるでしょう。

科学技術政策・研究戦略策定のためのデータ基盤

researchmapは、日本が研究とイノベーションを推進する戦略に役立つ基盤データを提供していきたいと考えています。ある科学技術政策を行ってそれが本当によかったのか——たとえば交付金を減らして競争的資金を増やしたらどうなるのか？ またこのことによって期限付き雇用の若手研究者が増加していますが、この人材がどこへ行くのか？ この雇用を把握することによって、元の政策を評価するといったことです。現在はこのような評価の基になるデータがないんですね。あるいはこのようなデータを大学にフィードバックし、各大学でさらに教務のデータを加えることで、大学の強み・弱みを分析する。今大学が迫られている「体質改善」を、最も望ましいかたちで実現するといったことにも、ぜひresearchmapをご活用いただきたいと思います。その際、人間が見て分析しやすいような情報の可視化技術にも、私たちは取り組んでいきたいと考えています。



Research View 024



進化する「プッシュ型」で いっそう使いやすく

researchmapはこれまでも背後でデータベースを駆使して、研究者の方が手間をかけずに論文や科研費情報などの業績を入力できるという便利な特徴を備えていました。このような使いごちの背後にある研究開発を、サブリーダの武田先生、相澤先生との協力体制の下で発展させています。これからは、たとえば共著論文を共著者の一人が入力したら「これはあなたの論文ではないですか？」と通知する機能や、同一人物・同一論文を同定する「名寄せ」などの機能強化を行います。機械が入力したものは薄く、本人が入力したものは濃く表示して入力支援するようなイメージのインターフェースを、来年度ぐらいからご提供できるのではないかと考えています。

「オープンサイエンスへ向けて、何よりも重要なのは、研究データをどう共有していくかだと思います。しかしながら、データによって共有しようと思っても読めない、機械が情報を取得することができない、といったことはよくあるんですね。このように人にしかわからないビッグデータが大量にあっても、データ利用は進みません。言い換えれば、人工知能は機械可読なビッグデータを待っているということです。ログインしたら、あなたはこういう情報が必要でしょうか、こういう研究者を欲しているでしょうかといったことが、人工知能の技術を使って、自然と推薦できる状態になるというふうなふうに思っています。」

(researchmapシンポジウムの会場にて)

- ▶ プロジェクト概要
<http://rc.rois.ac.jp/research/a04/>
- ▶ プロジェクトウェブサイト
<http://trc.rois.ac.jp/science3/>

Research View026 p.40



Research View 025

レジリエンスの知恵を 役立てるために。

統計数理研究所・教授
プロジェクトディレクター

丸山 宏
×第3回「IT×災害」会議実行委員会

Project

システムズ・レジリエンス

Research View 025

INTRODUCTION



防災減災を考えるには 人のネットワークが大切

今回のイベントでは、学術の立場からの提案として「システムズ・レジリエンス」プロジェクトの概要をスピーチ、ポスター、トークセッション等のいろいろなかたちでご紹介しました。また参加されたみなさんからもさまざまな立場からの報告や提言があって、その中でやはりよく話題にのぼったのが人的ネットワークの重要性だったと思うんです。それ、実は、僕らが考えているレジリエンスの本質にとても近いところにあって、多様性や利他性がうまく機能している一例として捉えることができると思います。今回の共催は、プロジェクトスタート当初の目標のひとつであったレジリエンスのタクソノミー(分類学)を、具体的な例に沿って検証する機会として大変有意義だったと考えています。

世界のレジリエンス研究が 集まった湘南会議

一方、世界の研究者たちは、レジリエンスというテーマにどうアプローチしているのか? —— というと、まず挙げられるのは、生態系や環境系のシステムを

本年度で4年間のプロジェクトを終える「システムズ・レジリエンス」。去る2015年11月21日には、現場の災害対応を考える人たちのコミュニティとの交流の機会として、第3回「IT×災害」会議を共催しました。主催の「IT×災害」会議実行委員会も、当プロジェクトと同様に、東日本大震災をきっかけに発足したコミュニティとのこと。「レジリエンス」がまさに生きたテーマであることを、実感させる機会でもありました。今回はこのイベントの様子を交えながら、「システムズ・レジリエンス」プロジェクトを総括します。お話は、プロジェクトリーダーの丸山宏教授(統計数理研究所)です。



「IT×災害」会議は、それぞれの仕事やバックグラウンドを持ちながら、ITや災害対応を考える人たちのコミュニティ。2013年秋に、アンカンファレンス形式による第1回を開催してから、いくつかのプロジェクトが派生し、成長してきたという。東日本大震災後のボランティアをきっかけに、復興庁の非常勤を務める小和田香さんは、第1回から実行委員を務め、「現場に行くのではない支援がある」と、情報を集約するウェブサイト「減災インフォ」を運営する。「平時も含めた、命を守るための情報発信に特化して活動しています。自治体・市民・地域メディアに一次情報の発信を呼びかけたうえで、情報の信頼度を高めるための発信方法やキュレーションのしかたを遠隔からサポートする。そのために人的ネットワークと情報系の両方からアプローチしています。」

Research View001 (2014-2015) p.08

対象としたレジリエンスの研究です。それから、システムの振る舞いを解析・モデル化するシステム・ダイナミクスの研究もさかんです。また地域別に見ると、危機意識の高さなどを背景にシンガポールやフィンランドなどが比較的早くから取り組んでいるといった特徴も見えてきます。われわれはこのようなレジリエンスに関する研究の世界的なネットワークを構築して、2015年2月にレジリエンスをテーマにした湘南会議を開催しました。6日間にわたる議論でさまざまな進展があり、中でも大きな「気づき」となったのは、脅威が現実になってあるシステムを救おうという時、そのシステム「だけ」を救うことは

研究者として、災害やリスクに強い社会を作るためのソーシャルデータやオープンデータの分析等を手がける村上明子さんも、今回の実行委員のひとり。災害発生時にいち早く現場に入り情報支援を行う「先遣隊」である「IT DART(代表理事:及川卓也氏、他)」の立ち上げにも携わった。IT×災害会議には第2回から参加し、「今回のテーマは“つながり×ひろげる”。学術の世界でレジリエンスをテーマとしたこのプロジェクトはまさにぴったりだと思い、共催を提案しました」。なおイベントでは「全日本芋煮会同好会」による炊き出し訓練も行われ、好天候の下「つながり」の輪が広がった。



どうしてもできない——言い換えると、ひとたび危機が起こったらそれまで対象としていたシステムの縁(バウンダリー)を変更しないといけない、ということでした。たとえばもし地球規模の脅威があったら、太陽系もある程度一緒に救わなければ地球を救えない。このことを境界からの溢出効果という意味で「バウンダリー・リーケージ(Boundary Leakage)」と呼んだりします。

タイの洪水における口コミ効果を解明する

「システムズ・レジリエンス」プロジェクトは、このような分類学に基づく帰納的アプローチと、システムの数理モデルから出発する演繹的なアプローチという2つの方法を、同時進行させるデザインになっています。具体的な研究成果を2つ採り上げると、まず1つに、レジリエンスは人々のムードにかなり依存しているという現象を取り扱う、岡田仁志准教授(国立情報学研究所)の研究です。人間社会には、人々

Research View 025



丸山教授は「技術者・研究者と社会貢献」のセッションでも講演。職業としての科学者はここ100年の歴史しかないことや、オープンサイエンスの新しい流れなどを紹介した。セッションを終えて「災害と言った途端にわれわれの頭の中には地震、津波、台風といった災害が思い浮かぶけれども、パンデミックのほうはるかに大きな脅威かも知れない」と、丸山教授。

が大丈夫だと思えば比較的安定するし、逆に誰も信じないムードになってくるとどんどん事態が混乱するといった傾向があります。そこでタイの洪水災害を例にムードが悪くなる傾向をモデル化し、みんなが信じられない等と言い始めた時期の特定などを、今ツイッターのログ等から検証しているところです。

遺伝子からわかった 収獲逓減則と多様性の相関

もうひとつは、生物の進化において、収獲逓減則を入れるとシステムが多様になるという明石裕教授(国立遺伝学研究所)の研究成果です。収獲逓減則があるシステムでは、有利な構成要素を横軸に、その効用を縦軸にグラフを描くと、最初は右上がりに伸長し、途中から横ばいに近くなるような軌道を描きます。実際の遺伝学の研究では、個体に有利な突然変異を横軸に、環境への適合性を縦軸にとります。ある程度まで有利な条件が積み上がってくれば、それ以上有利な条件が増えてもたいては変わらない、というのが収獲逓減則の利いたシステムです。実際にも、ダーウィンの法則に従って、私たちは100%最適な者しか生き残らないのではなく、多様な生命にあふれる

世界に生きていますよね？ 収獲逓減則を利かせることによって、有利な遺伝子を適度に備えた中間層が多様に存在するようなシステムをつくることができる——これはいま企業などが持続的発展のために積極的に取り組んでいる人材の多様化施策にもヒントになる発見と言えるのではないのでしょうか。

▶プロジェクト概要
<http://rc.rois.ac.jp/research/b04/>
 ▶プロジェクトウェブサイト
<http://trc.rois.ac.jp/resilience/>



Research View 026

ウェブデータは オープンを目指す。

国立情報学研究所・教授
武田英明

Project

e-サイエンス基盤技術

Research View 026

INTRODUCTION



人間の知識へ、 工学的なアプローチで近づきたい

AIの分野は今、第3次ブームを迎えています。機械が学習するためにはやはりその前提として知識が必要です。ところで知識とは何でしょうか？……仮にブリタニカ百科事典を人間の知識源の代表例とすれば、近所のレストラン情報や列車の運行表のような身近な知識もありますね。そもそも情報を集めようがなかった時代には、そんな世界の隅々にわたる些末な事柄まで知り得るとは誰も考えなかった。ちなみに近年話題になっているIBMの人工知能ワトソンも、まさに巨大な知識ベースという昔からの夢を実現したものだと言えるでしょう。しかし今や、ある種の哲学的な知識よりも、むしろ生活の中にある些末な知識のほうが重要だろうと私は考えています。ブログやSNSなどを含めたこのような知識に対して、工学的なアプローチで近づきたいというのが、私の研究です。

ウェブ以前から人間の持つ知識に注目し、人類の巨大な知識をいかに機械が使えるようにするかに取り組んできた武田英明教授(国立情報学研究所)。まだ人々を惹きつけることはなかった1990年代前半のウェブに大きな衝撃を受けて「未来の知識はここにある」と、人工知能(AI)の研究からウェブによる知識獲得へ、いち早くテーマを移したといいます。約6年前に「学術世界の情報共有を進歩させよう」という新井紀子教授(国立情報学研究所)の講演に賛同し、セマンティックウェブと呼ばれる技術を利用した学術情報流通基盤構築を提案したのが、リサーチcommons「e-サイエンス基盤技術」プロジェクトへの参加のきっかけでした。

ブレークスルーとなった「DBpedia」

ではこのようなウェブ上の知識がどのような形式で記述されていれば、機械と人がともに活用できるのでしょうか？そこで、ウェブの生みの親であるティム・バーナーズ＝リーが提唱したのが、ネット上の情報を再利用可能な意味ネットワークにするための技術「セマンティックウェブ」でした。これは主語・述語・目的語からなるシンプルな構造化言語「RDF(Resource Description Framework)」を標準言語とし、この言語によって書かれたデータは「LOD(Linked Open Data)」と呼ばれます。その名の通り、ネット上に「オープン」に置かれ、他のデータと「つながる」ことができるのがLODの大きな特徴です。たとえば「来月10日東京で行われるイベント」といった文に含まれる「いつ」「どこで」「何を」などの意味を機械が理解し、「来月のイベント情報」といった他のデータベースともスムーズにつなぐことができるわけです。ウェブがこの新しい記述方法へ移行するひとつの契機が、ウィキペディアのRDFデータである「DBpedia」だったと言えるでしょう。いったん機



械が理解できるデータが作られれば、人手を介さず RDF で知識を出力できるため、たとえばウィキペディアの知識を自社のデータベース上でシームレスに活用するといったことが簡単に行えます。実は、人類の知識をいかに集めるかという問題には、すでに答えが出ているのです。

つながるデータを作る LODAC プロジェクト

とはいえ LOD は、ウェブ全体のデータ量からみると 1% にも満たないと言われています。そこでわれわれの目的である学術情報の共有を実現するためには、われわれの知識を共有するしくみを準備しなければなりません。そこでまず DBpedia の日本語版「DBpedia Japanese」を、プロジェクトの中で運用しています。たとえばスポーツのカテゴリに「相撲」がなかったのを加えるというようにオントロジーの拡充を図ったり、検索などで入力される文字列と

ウィキペディアの項目をつなぐ「エンティティ・リンク」を構築して、利便性を向上させたりしています。また日本の美術館・博物館 100 館以上の収蔵作品情報を集めて、LOD によるデータベースも作りました。専門家による同一作者の特定を行うなどして、串刺し検索等ができる「LODAC Museum」というサイトを運用しています。さらに生物の分野でも、分子生物学・生態学といった領域の違い、植物・動物・微生物という対象の違いなどによってばらばらの状態に置かれていたデータベースの統合を行い、ウェブサイト「LODAC Species」を公開しました。また日本特有の和名という問題も、世界の情報とつないで最大の生物和名データベースを構築して同サイト上で提供しています。この際、苦労したのは種の名称で、新しい科学的発見によって学名が変更になること等がわかったため、知識の変化を書くオントロジーを作って学名の変遷がたどれるような仕組みを提案しました。

Research View 026

オープンでつながる 科学データのこれから

研究データを作って、論文は投稿したけれども、その後データは共有されないままになっている……それじゃ駄目なんだ、と今、学術の“文化”が変わろうとしています。実はこれ、生物学に限らず科学全般に広がる問題であって、データ標準である LOD を使って、研究データを将来にわたって使えるように公開することの重要性は、世界的にも広く認識され、研究データのオープン化を加速しています。ちなみに当りサーチ commons の DBCLS の取り組みも、同じ方向にあると言えるでしょう。また個々の論文や研究者をウェブ上で一意に認識できるようにするための識別子として、データの置き場所が変わると変わってしまう「URI」に代わって、「DOI」を付与しようという活動も始まっています。さらに論文そのものも——ひょっとするとすでに大半の論文が、文章を書くのではなく、データを構造的な情報として提示すればよいのかも知れません。実際、多くの学術分野で日々生まれる論文を人間が読み切れないため、いかに適切な論文を探すかという課題が生じています。LOD データであれば、誰もが、より欲しい情報に的確にアクセスできるようになる——それはみんなにとってよいことだと言えるでしょうね。



DBpedia Japanese
<http://ja.dbpedia.org/>
 LODAC Museum / LODAC Species
<http://lod.ac/about>
 ジャパンリンクセンター (JaLC)
<https://japanlinkcenter.org/top/>
 ORCID
<http://orcid.org/>
 ▶ プロジェクト概要
<http://rc.rois.ac.jp/research/a04/>
 ▶ プロジェクトウェブサイト
<http://tric.rois.ac.jp/science3/>

ソーシャルメディアのデータ等を調べていると、知識がどのように出来てくるかというところに、すごく興味を引かれますね。その中には、後世ずっと残っていくような知識もきつと含まれている……。現実の知識を扱う研究としては、あまり大きな仮説よりも、むしろ今、目の前にある知識、情報なり、データなりをしっかりと見るのが重要だと思っています。研究のかたわら、DOI 識別子の登録機関「ジャパンリンクセンター (JaLC)」の委員長、また学術論文の著者 ID を統合的に管理しようというプロジェクト「ORCID」の理事も務める。

Research View 014 (2014-2015) p.60



Research View 027

統計的思考が 自然言語処理を変えていく。

統計数理研究所・准教授
持橋大地

Project

メタ知識構造解析

Research View 027

INTRODUCTION



モデリングを担う統計解析

僕は、単語それぞれがどのような確率を持って使われているかを一般的に説明できる、言語モデルについて研究しています。これまで人間が営々として記述してきた知識を、実際に人がどのように使っているかというデータから導きたい。この「使われ方から導く」というのが、まさに統計的な方法ということですね。統計的自然言語処理の学者は、自然言語をモデル化する知見を持っているけれど、これは非常に新しい学問分野なので、何をすべきかについては伝統のある言語学も参考にして、われわれは問題を解く方法を提案する。するとこれまで手作業でやっていたアノテーションのような作業に「自動的にできるんじゃないの?」という方法が示せたり、また言葉の論理性だけに頼っては解くことができなかった現象にも迫れたりします。

人と機械が共生する以上、人間が使う言葉を研究する「自然言語処理」という分野を欠かすことはできません。そもそも言語は、人間にとって思考の道具であり、他の動物にはない大きな特徴と捉えられてきました。その根幹にあるものは論理でありながら、機械のためのプログラミング言語とは異なり、論理だけでは説明できない難しさを持っています。どうしたら人間が使う言葉の「意味」をモデル化できるのか?——学術における自然言語処理は、その背後で統計学、離散数学、最適化問題など、さまざまな数学を駆使しながら発達してきました。なかでも確率・統計の考え方にに基づき、観測された言葉の集合だけから機械が学習する「教師なし学習」のアプローチで知られるのが、
「メタ知識構造解析」
プロジェクトの持橋大地准教授(統計数理研究所)。今回は、その研究の最前線をご紹介します。

基礎研究ならではの コラボレーションの広がり

自然言語を解析するためのこうした統計的な方法は、実は言語以外にも、言語と共通性のある他のデータにも使うことのできる基礎的なものです。そこで現在、自然言語処理だけでなく、他の分野への適用や応用化などを含めた共同研究がいくつか同時進行しています。たとえば国立国語研究所との共同研究では、言語のもうひとつの重要な要素である「音声」に注目し、言語モデルの新たな基礎になるような研究を進めています。また言語処理で研究開発した「階層ベイズ言語モデル」や「隠れマルコフモデル」等と呼ばれる統計的手法を音楽や楽譜解析に応用した「音楽情報処理」、音響解析に活用した音響モデル、さらにロボットの動作を時系列に沿ってモデル化するロボティクスへの応用研究なども進行中です。



「たたずまい」という言葉を 論理は表現できない

共同研究においては、しかし、たとえば言語学者や社会学者がまったく統計を知らないまま「データだけ送りますから、分析をお願いしますね」といったことでは意味がないだろう、と僕は考えています。



方法とは、おまけのようなものではなくて、新しい知見が生まれる土台のようなものだからです。最終的にはやはり自分で両方できるようにすることを目指す、その第一歩になるのが共同研究だと思います。というのも僕自身が、実は文系から理系へ進学した経歴を持っているんです。院生時代、強く思っていたことのひとつは、「たたずまい」という言葉を論理は絶対に表現できないということでした。「たたずまい」という言葉の意味を、その言葉自体の中へいくら深く追っていても、意味を取り出すことはできません。意味はむしろ、この語が他の語とどのような関係にあるのか、外からどう観察されるかを測ることによって定義されます。実際、統計・確率の方法を使えば「たたずまい」を表現できるのです。

Research View 027



宮尾祐介プロジェクトディレクター との出会い

リサーチコモンズの「メタ知識構造解析」でも、このような自分の確率の方法や知識が役立てられればと考え、宮尾准教授(国立情報学研究所)とプロジェクトとして初めて一緒に研究しています。つい最近も「ところで」という語は、話題を切り替える力を持っていますが、その力を前後の言葉の分布を比較することによって測ってはどうか? といった提案をしたところでした。宮尾准教授は、言葉の論理性の要である構文解析で知られる方ですが、論理といってもある種のメタ論理体系を構築することによって、集合とその下位集合のような論理的な包含関係を考えるなど、イノベーティブな取り扱い方に特徴があります。統計から見ると、このような集合論は表現しにくいので、とても興味のあるところですね。人の言語の使用をすべて統計モデルから導けるのか、それともどこかに人手を介した方がよいのか、まさにそういった点をプロジェクトでいっそう明らかにしていきたいと考えています。

2016年2月刊行予定の『岩波データサイエンス』第2巻 自然言語処理特集では、オーガナイザーを務める。執筆陣には、宮尾准教授の名前も。自然言語処理は近年、社会インフラとしての利用を見据えて、大学だけでなく企業等でも開発に取り組む研究者・技術者が増えつつある。「自然言語処理に興味を持っている人々へ向けて、学術的な研究の最前線でどんなことが行われているかをわかりやすく伝えようというのが狙いです」。

▶ プロジェクト概要

<http://rc.rois.ac.jp/research/a02/>



Research View 028

データが語る“豆ぶち鼠” の大きな帰還。

国立遺伝学研究所 副所長・教授
城石俊彦

Project

遺伝機能システム

Research View 028

INTRODUCTION



それはメンデルの法則から始まった

われわれが取り組む遺伝学という分野は、メンデルの遺伝の法則に始まったと言ってよいでしょう。単に子が親に似るというのではなく「何らかの単位化された物質が親の性質を決めている」というまったく新しい考え方は、1866年の論文発表当時すぐには認められず、1900年になってようやく3人の学者によって再発見されました。この時、メンデルの理論を検証するために行われた追試のひとつが、今回の話の主役である「豆ぶち」鼠ことファンシーマウスと、ヨーロッパ産マウスとの交配実験でした。ファンシーマウス「JF1」は小さな体に白黒のぶちという特徴を持つ鼠で、日本の文献をひもとくと、江戸時代の天明年間に大阪で出版された『珍玩鼠育草(1787年)』の中に絵入りで紹介されています。また当時のイギ

実験に理想的な遺伝的特質を持つことに加え、ショウジョウバエとならんで約100年もの歴史を持つ実験用マウス「ムス・ムスculus」(*Mus musculus*)」。なかでも世界のラボで最も広く使われているのが、黒っぽい栗毛を持つC57BL/6という系統です。この基準系統は、20世紀初頭のアメリカで樹立されたもので、ヨーロッパ産のマウスに由来するとされながらも、実は日本産マウスのルーツを持っているらしい、と長い間考えられてきました。しかしヒトゲノムが解読された21世紀初頭、バイオ分野だけでなく科学全体に大きな変化が訪れます。2002年12月、6カ国からなる国際研究チームによって、ヒトに次ぐ2番目の哺乳類としてマウス基準系統の全ゲノムが解読されると、そのルーツにも新たな光が当たることになりました。30余年に及びこの問題に取り組んできた国立遺伝学研究所と、その一端を担うリサーチコモンズの「遺伝機能システム」プロジェクトの成果を、国立遺伝学研究所 城石俊彦教授をご紹介します。

リスの文献などから、この愛玩用マウスを知ったイギリス人の貿易商が横浜から持ち帰り、19世紀の後半に本国の愛好家に渡ったことがわかっています。当時のJF1マウスは、内耳の三半規管の異常のために方向感覚がなく、くるくると回ったため「ワルツを踊る鼠(Japanese waltzing mice)」と呼ばれていました。

デンマークから帰還した「豆ぶち」

ところが1987年のこと、デンマークの首都コペンハーゲンのペットショップで、偶然にもこの「豆ぶち」あるいはファンシーマウスが、実際に見つかったのです！ この時点では、まだ本当に日本産マウスであるかどうかはわかりませんでした。が、国立遺伝学研究所の故・森脇和郎教授がこの

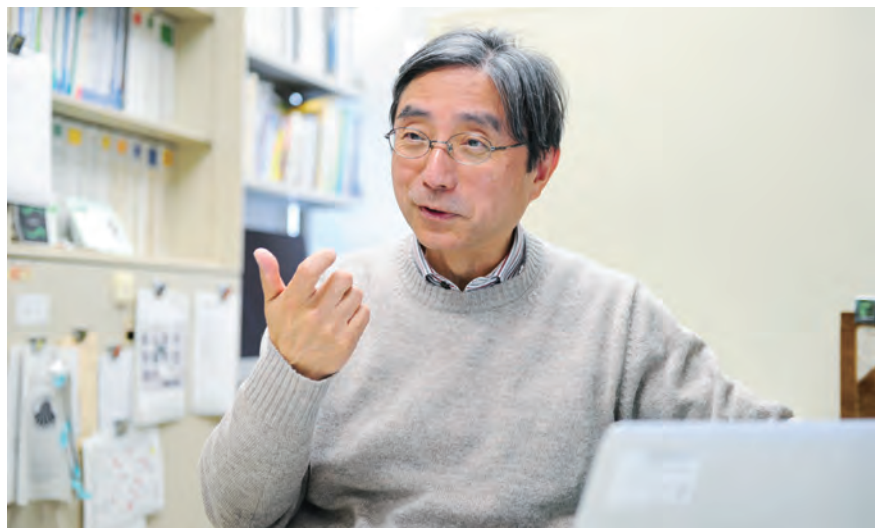


マウスを日本に持ち帰り、長年の努力の末に豆ぶち鼠「JF1/Ms」系統を樹立することに成功しました。また研究所では同じ頃、日本各地の野生鼠の系統育成にも取り組んでおり、そのうちのひとつ、三島産の「MSM/Ms」の系統樹立にも成功しました。

最先端研究を「データ中心」の解析で競う時代へ

一方、1953年に発表されたDNA二重らせん構造と技術の進歩によって、20世紀末には分子生物学・分

子遺伝学が発展し、多くのゲノムプロジェクトが生まれました。特にヒトゲノム計画をきっかけに、ゲノムプロジェクトで作成された塩基配列データは論文発表の前に全データを公開して研究者が自由に利用できるようにするという、研究コミュニティの世界的な合意が形成されました。生命科学の分野では、大量の情報を解析して“データに語らせる”「データ中心」の方法が不可欠な時代に入ったと言ってもよいでしょう。2002年に公開されたマウス基準系統のゲノムデータからは、哺乳類の中でもマウスが遺伝的にヒトにとても近いことが明らかになり、特に医療分野等における実験用マウスの重要性が改めて認識されました。国立遺伝学研究所を中心とする日本のチームは、2008年に系統樹立した日本産マウス2系統の全ゲノムの解読を終了し、そのデータを世界へ向けて公開しました。その後2011年には、英国サンガー研究所から西欧産亜種「WSB/EiJ」の全ゲノムも公開され、ヨーロッパ産と日本産のマウス系統の全ゲノム情報の比較解析から、実験用マウスのルーツを探ることが可能になったのです。



Research View 028



国立遺伝学研究所にある野生マウスの標本室。世界各地から捕獲して集められた剥製が、引きだしっぱいに並ぶ。「ここにある標本は、すべてゲノムDNAが調整済み。毛色や形態の計測データも揃っています」と城石俊彦教授。

ムス・ムスクルス100年の旅路

「遺伝機能システム」プロジェクトでは、われわれは日本産マウス2系統とヨーロッパ産亜種の全ゲノムを、ベイズ推計という統計の手法を用いて基準系統と詳細に比較し、基準系統の遺伝的な起源を明らかにしました。まず基準系統には日本産マウス由来の遺伝子が6~7%含まれており、これが基準系統に2種類のゲノムが混じった「モザイク構造」を与えていたことがわかりました。そして基準系統に混じている日本産マウスの遺伝子は、三島産マウスよりもファンシーマウスJF1の特徴によく一致していました。これらの結果を総合して、われわれは——日本の愛好家に育てられていた「豆ぶち」が英国の愛好家に渡り、ヨーロッパ産マウスと混血し、この系統がアメリカへ渡って基準系統となったことはほぼ間違いない——という結論を得ました。この成果は2013年にGenome Res誌に発表されています。



データ公開サイト『国立遺伝学研究所マウスゲノムデータベース(NIG_MoG)』の運営を担う、国立遺伝学研究所の高田豊行助教。エネルギー代謝に関連した遺伝子の探索と機能解析が専門だが、「ふだんラボで仕事をしている研究者が、なるべく直感的に使えるよう工夫しています。少しでも多くの研究に役立ててもらえれば」と、日々改良に励む。たとえばトップページでは、染色体の図表にカーソルを合わせるだけで、ゲノム番号と位置情報(座標)が表示され、希望するゲノム領域を探索できる。遺伝子の検索結果としては、塩基多型が一覧できるテーブルが表示され、テーブル内のリンクからさらに他のデータベースへ飛び、関連するタンパク質やアミノ酸置換情報等も参照できる。「現在、ゲノム多型情報と他の表現型情報をリンクさせる機能強化を進めています。遺伝子機能の解明を加速できればと考えています」。

▶プロジェクト概要
<http://rc.rois.ac.jp/research/b02/>
 ▶プロジェクトウェブサイト
<http://tric.rois.ac.jp/life/>



Research View 029

ゲノム×統計学の NEXT STEP。

統計数理研究所・教授

栗木 哲（写真左）

統計数理研究所・教授

藤澤 洋徳（写真右）

Project

遺伝機能システム

Research View 029

INTRODUCTION



いま注目の 「セレクトティブ・インファレンス」

今回の研究集会はプロジェクトのまとめの報告会として開催するもので、プロジェクトに関わってきた仲間がこれまでに培ってきた統計的手法を、ゲノム解析に関わるコミュニティに紹介しようという趣旨です。そしてこの他にもうひとつ、この機会を活用して、この分野に関わる別の新しい動向をご紹介いただこうと、プロジェクト外の研究者にも声をかけました。招待講演のテーマは、機械学習の中の新しい潮流である「セレクトティブ・インファレンス」です。私がかねてから専門としていた多重検定という統計的手法が別の文脈で展開され、しかも近年、非常に発展していると聞き、たいへん楽しみにしてきました。（栗木）

セレクトティブ・インファレンスについては、われわれ統計の立場から見て、機械学習がここまでやるのか、という驚きがありました。というのは、データマイニングのように、データからの発見という志向を持つ機械学習に対して、統計は得られた結果に本当に再現性があるのかという、方法が担うことができる信頼性をとても重視するからです。そして多重検定

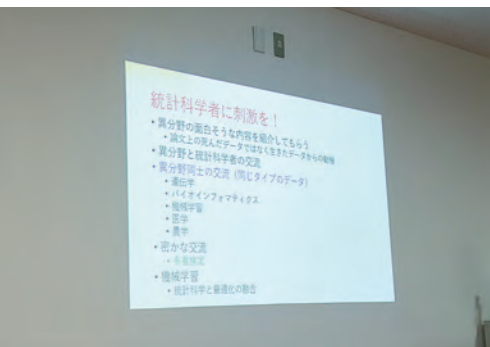
2016年3月、3年間のリサーチコモンズ事業が終了します。そのプロジェクトのひとつである「遺伝機能システム」は、しかし、当機構設立以来の融合研究を継承し、併せて約10年にわたって発展してきました。そこで去る1月25日、その成果を報告する研究集会「遺伝学と統計学における数理とモデリング」が、東京・六本木にある政策研究大学院大学にて開催されました。ゲノムを中心とした遺伝学のデータを解析する統計学にとって、「遺伝機能システム」はどんなプロジェクトであり、今後どんな展開へとつながっていくのか——研究集会の開催について、開催責任者でもある統計数理研究所の栗木哲教授・藤澤洋徳教授をご紹介します。

とは、実はこのような統計の性質をととても色濃く持つ手法のひとつなのです。（藤澤）

数学が「難しすぎる」と 「簡単過ぎる」の狭間で

融合研究を始めた10年前は、新しい研究者ネットワークが急に目の前に開け、すごく刺激的でした。この人たちは何を知っているんだろうと、お互いに触手を伸ばし合ったし、私もゲノムに接して、なんというデータの宝庫だろうと魅力を感じました。一方、最近ではみんなすっかり顔なじみになって、論文の投稿やレフリーへの回答等に関する細かな相談を受けることが多いですね。（栗木）

Research View002 (2014-2015) p.12



研究自体は面白いのですが、一番困ったのは、論文が通りにくいことでした。既存の研究分野には、それに対応した科学雑誌があるので、ある程度こういうものを書けば通るという暗黙の理解があります。ところが新しい研究には専門のジャーナルがないし、しかも融合研究ではたとえば遺伝系の雑誌に出せば数学が難し過ぎると言われ、統計の雑誌に出すと数学が簡単過ぎると言われます(笑)。ちょうど中間にあたるバイオインフォマティクス系の雑誌にも投稿しましたが、それでも苦労しました。(藤澤)



生殖隔離を起こす イネ遺伝子型の多重検定

本プロジェクトで私と藤澤教授、倉田のり教授(国立遺伝学研究所)、春島嘉章融合プロジェクト研究員とが行った解析のひとつに、イネの遺伝子型を大量に集めたデータの多重検定問題があります。イネの染色体は全部で12本あり、そのうちのある特定の場所の遺伝子とまた別の特定の遺伝子の組み合わせによって、なかなか生き残らないイネがあることが知られています。純粋種を掛け合わせて樹立された系統を考え、このような個体が減数分裂の段階で「交差」したりするような現象を含めてモデリングしていくと、ある染色体とある染色体の組み合わせの個体数というのが計算できます。ところが実際に実験を行うと「ずれ」がある——これが、遺伝子が不適合な場合に生殖隔離を起こすという有名な「ドブジャンスキー・ミュラー(Dobzhansky-Muller)モデル」で、この原因遺伝子を特定することが遺伝学的に重要だというのが、そもそもの問題でした。そこで私たちはチューブ法を用いて、1,000個掛ける1,000個の遺伝子型の

Research View 029



Research View009 (2014-2015) p.40

組み合わせで理論値と実測値を比較する、全55万個の検定を行いました。それぞれの組み合わせのピークを算出し、それが有意なピークであるかどうかの判定に関する論文を発表しました。(栗木)

統計学者に、もっと刺激を！

ゲノムデータは、非常に巨大であり、それまでの統計的手法では対応しにくいことも多く、また統計学に大きな意識の変化を要求しました。このことは私自身がこれまで関心を持ってきた、データの中心から著しく外れているため通常は除外する「外れ値」と、観測データの欠損である「欠測値」についても同様でした。テクノロジーが進化すれば精度が上がり「外れ値」はあまり出て来なくなるだろうと思っていたのですが、新しいものこそ不安定な部分を持っており、むしろ外れ値は増えたりすることも多いのです。するとやはり自動化したり、新しい統計的手法をつくり出そうとモチベー

ションが湧いてきます。それからプロジェクトを通じて、欠けのある生々しい科学データに接したことも、大きな動機づけになりました。このような刺激こそ、新しい研究の源ではないでしょうか。(藤澤)



写真左から城石俊彦副所長、高田豊行助教、プロジェクトディレクターの倉田のり教授(以上、国立遺伝学研究所)、土谷隆教授(政策研究大学院大学、共同開催責任者)、栗木哲教授、藤澤洋徳教授(以上、統計数理研究所)。

▶プロジェクト概要

<http://rc.rois.ac.jp/research/b02/>

▶プロジェクトウェブサイト

<http://trc.rois.ac.jp/life/>

Project

イメージデータ解析

複雑な現象のイメージデータを高度に解析する技術をつくります。

生物、人間社会、宇宙などの大量のデータをさまざまな組み合わせで利用できるようになったビッグデータ時代——「イメージデータ解析」プロジェクトは、現代の複雑な現象から生みだされる膨大なデータから、人間が重要な情報を直感的に取り出せるようにしようという挑戦であり、世界的にも活発な研究分野だと言えるでしょう。

本プロジェクトは、生物・医療関係、人間社会にまたがる3つの異なる分野の研究者たちが集まって、それぞれサブテーマを持って活動してきました。まずは各分野で実際に生み出されたイメージデータを取り扱い、独自にデータを高度に解析する可視化の技術を開発します。さらにその成果を汎用的なイメージデータ解析ツールとして公開し、他の基礎分野のさまざまな問題に適用したり、応用分野に適用してさまざまな意思決定支援に役立てたりできるような基盤技術を確立します。

≫ バイオイメージング・インフォマティクス (サブテーマ1)

バイオイメージ・インフォマティクスとは、近年バイオインフォマティクスの新しい分野として急速に活発化している研究分野です。バイオイメージ・インフォマティクスは応用分野が広く、分子生物学の基礎研究から、医療における創薬・診断・治療まで多岐にわたります。サブテーマ1ではこのうち、バイオイメージの特徴抽出、分画、レジストレーション、分類、可視化等の統計的な方法論構築に取り組んできました。

Project director

松井 知子〔統計数理研究所〕



具体的な成果として、4Dライブセルイメージングのデータを対象とした細胞検出、セルトラッキング、イメージレジストレーションという3つのアルゴリズムを開発し、ソフトウェア「SPF-CellTracker」として公開しました。また線虫Cエレガンスの中枢神経系のカルシウムイオンイメージに開発手法を適用し、神経回路の動作特性に関する研究を発展させました。

≫ オプティカルイメージングデータを用いた脳の動作原理の解明 (サブテーマ2)

脳の動作原理の解明には、脳に蛍光色素を導入して、その流れの軌跡を画像で辿るといった画像解析が欠かせません。しかしこれまでは、たとえばカルシウムイメージングデータ画像を獲得しようとしても、蛍光色素が不均一に拡散するために適切な画像が得られないなど、数多くの問題点がありました。

サブテーマ2では、カルシウム指示薬である「Oregon green BAPTA-1」を利用した蛍光ライブイメージングデータの各ピクセルに対応する時系列の分散値のマップ、局所細胞外電位として計測される呼吸活

動と蛍光ライブイメージングデータの各ピクセルに対応する時系列に時間依存相互相関解析を施して得られる相関マップ、そして「Red Fluorescent Protein, RFP」の強度マップにより、アストロサイトと吸気性のニューロンを区別する方法を開発しました。これらの解析方法の開発によって、大脳の賦活部位の推定の信頼度が向上することが期待されます。

≫ 人間・社会の視覚情報データ解析 (サブテーマ3)

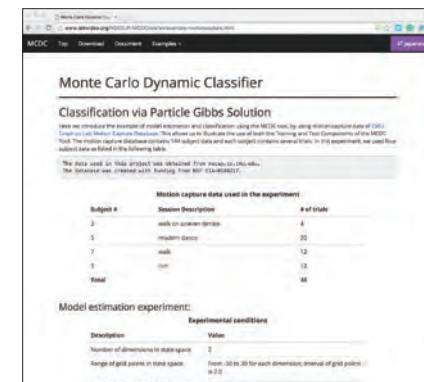
人間・社会に関わるサブテーマ3では、ケアを必要とする人の増加や老人の孤立化などの社会問題に応えるような研究テーマとして、人や物の外的な特徴である映像データをトラッキングすることで、潜在的に重要な(動的)因子を推定し、イベント検出・予測

を行うできる研究開発を行いました。2014年、これをソフトウェア「Monte Carlo Dynamic Classifier, MCDC」として公開しました。

MCDCは、ガウス過程状態空間モデルに基づいて構成されています。公開後も、複数サンプルからの学習、学習時に潜在的因子の情報が利用できる場合への対応、潜在的因子モデルが利用できる場合への対応、いろいろなノイズモデルの利用、アルゴリズムの高速化について検討し、汎用的ツールとして拡張・整備しました。モーションキャプチャーのデータ(CMU Graphics Lab Motion Capture Database)を用いて識別性能を評価する実験を行い、妥当な推定結果が得られることを確認しています。



サブテーマ1より「SPF-CellTracker」
https://github.com/ohirose/spf



サブテーマ3より「Monte Carlo Dynamic Classifier, MCDC」
http://www.ismvideo.org/NDE0UP/MCDCtool/en/index.html

Project

メタ知識構造解析

人が生み出すテキストからまとまった意味のある知識を自動獲得します。

私たちの社会的な活動のさまざまな成果は、そのほとんどがテキストデータというかたちで大量に蓄積されています。特に知識を生み出す分野である学術の世界では、多様な研究成果が学術論文や特許として、日々大量に公開されています。しかしながら、これらの成果やその波及効果を客観的かつ定量的に把握・活用する方法は、今のところ存在しません。

本プロジェクトでは、テキストの中で明示的・非明示的に表現された因果関係、理由、目的などの「メタ知識構造」に注目し、これに基づいて構造化された知識を自然言語テキストから自動抽出する手法について研究を行っています。言語学、自然言語処理、コーパスアノテーション、統計学と専門分野の異なる共同研究者が集まり、分野ごとに責任を持って研究開発を進め、定期的なミーティングにおいて情報交換と議論をしながら推進してきました。

メタ知識構造の研究開発項目

開発にあたり、われわれはメタ知識構造の2つの性質に着目しました。1つは言語的な性質で、ある関係(例えば理由)を示す明示的な言語表現(手がかり表現)を利用します。もう1つは統計的特徴で、ある概念(例えば理由になりやすい概念)や概念間関係(因果関係になりやすい2つの概念)の統計的・確率的分布が、テキストを横断して共通することを利用します。これら2つの性質はそれぞれ不十分かつ相補的であるため、最終的にはこれらを統合して、自動抽出手法を開発します。そこで本プロジェクトのアウトラ

Project director

宮尾 祐介 [国立情報学研究所]



インとして、以下のような研究開発項目を挙げることができま。

- メタ知識構造の定式化
- 学習・評価データとしてアノテーションコーパスの構築
- 言語的手がかりに基づく自動抽出手法の開発
- 統計的特徴に基づく自動抽出手法の開発
- 言語的・統計的手法を統合した自動抽出手法の開発

アノテーションコーパスの構築

学術論文を解析すると、原因、結果、手段、目的、理由、状態変化……といったメタ知識構造が、文章の至るところに現れることがわかります。これらは知識を記述する際に用いられる汎用的関係性であり、知識という情報の根幹を成すといっても過言ではないでしょう。われわれはまず、関連する研究の調査を行い、次年度以降の研究に必要不可欠なリソースとなる理論やデータの整備を進めました。

まず、実際に公開されている情報処理学会論文誌アブストラクト230件、ACM, ACL のアブストラクト

300件を分析し、これに基づいてメタ知識構造を形式的に表現するための枠組みを構築し、コーパスのアノテーション方法(スキーマ)を策定しました。なお本プロジェクトでは、メタ知識構造が明確な学術論文や特許文書などのテキストを対象としてスタートしていますが、将来的にはより一般的なテキストデータを対象とすることを視野に入れています。

言語理論に基づくアノテーションスキーマの策定

策定したスキーマに従ってアノテーションコーパスを構築する一方、言語的手がかりに基づく自動抽出手法の開発の一環として、SDRT 理論を参考にした、談話関係に対して形式的意味記述を与えるアノテーション方法を構築しました。実験を行ったところアノテータ間の一致度が低いことから、今後スキーマの再検討を行う予定です。

統計的手法の開発と検索プロトタイプシステム

次に、言語的手がかりに基づく手法と統計的特徴に基づく手法を統合した「構造予測モデル」を用いて、メタ構造の自動抽出を行いました。構造を自動認識できる精度は、100アブストラクトを用いた10分割交差検定によって行いました。またこのような機能を提供するアプリケーションとして「検索プロトタイプシステム」をウェブ上で公開しました。このプロトタイプシステムでは、「目的」や「結果」といった関係を区別して、論文を検索することができます。例えば、同じ「機械学習」に関する研究でも、「機械学習を応用した研究」と、「機械学習を改良した研究」を区別することができます。

しかしながら、たとえば並列して述べられている2つの記述が因果関係を持つような文章では、因果関係があるにもかかわらず、その関係がテキスト上に明示されません。このような非明示的な関係や、受け受けのあいまい性などは、メタ知識構造を記述するに際して、残る課題と行うことができるでしょう。

京大コーパスからSVM で学習した形態素解析器を用いて朝日新聞を解析したところ精度は98.5%であった

We trained an SVM-based morphological analyzer on the Kyoto University Corpus and analyzed the Asahi Shimbun with 98.5% accuracy.

「メタ知識構造」を形式的に表現するアノテーションコーパス



検索プロトタイプシステム

Project

データ同化・シミュレーション支援技術

シミュレーションと大量データ解析を高度に一体化します。

ビッグデータ時代のデータ解析は、単純にデータ量が増加しただけでなく、多くの異なるタイプの発生プロセスからのデータを扱わなくてはならない点に大きな特徴があります。複雑な大規模数値シミュレーションと大容量観測データを融合させる計算基盤技術として、主に気象学などで目覚ましい発展を遂げた「データ同化」は、ひとこと言えば、シミュレーションモデルとデータ解析を結び付ける統計手法。ビッグデータに有効な方法として、近年注目を集めているもののひとつです。

本プロジェクトは、この手法をいっそう高度化・一般化し、地球科学や生命科学などに応用・拡大して、具体的な問題解決に貢献しています。分野を超えた共同研究を推進する「数理・計算」「モデリング」「データデザイン」の3チームから、主な研究成果をご紹介します。

≫ 数理・計算チーム

インフルエンザ定点観測に基づく感染症予測の開発：感染症の拡大はどのように起こり、どうしたら緩和・阻止することができるのか？ 確率的に変動する結合定数を通じて地区間相互作用を取り入れた「連結 SIR モデル」に、47都道府県における週ごとのインフルエンザ定点当感染者数をデータ同化しました。

破碎イベントで発生する破片デブリのモデリング：年々増加し続けている宇宙空間中の不要な人工物であるスペースデブリ(宇宙ごみ)は、宇宙空間の安

Project director

中野 純司〔統計数理研究所〕



全かつ持続的な利用のために、その対策が欠かせません。これにはまず、宇宙における爆発や衝突などの破碎イベントによって発生するメカニズムの正しい理解が重要です。既存のNASA標準破碎モデルから一歩踏み込み、破片の放出方向まで一致するより正確なモデルの開発に取り組んでいます。

その他：逐次ベイズモデルに基づく南極氷床コアの年代推定、内部磁気圏統合モデルへのデータ同化システムの開発、国際宇宙ステーションの共鳴散乱光画像データに基づくプラズマ圏プラズマ密度分布の推定ほか。

≫ モデリングチーム

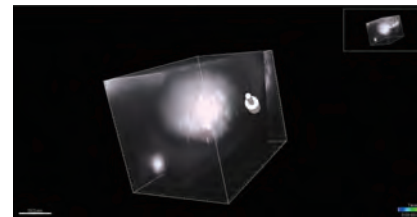
グローバル磁気流体力学(MHD)シミュレーションモデル:SuperDARNレーダーネットワークから得られる電離圏プラズマ対流速度データについてモデルパラメータ感受性を検証し、さらに電離圏電場

ポテンシャル分布の経験値モデルとの比較検証を行いました。

南極大型大気レーダー(PANSY)データを用いた高次情報処理手法の構築:PANSYを含む世界の主要なレーダーから送られてくる観測データと推定値を蓄積し、この統計・解析を通じて理論の妥当性や推定手法の精度検証等に取り組みました。物理量の大气乱流強度とエネルギー消散率の正確な推定、直接測定データのグローバルモデルへの同化スキームのアップデート、GPGPU計算機を用いたレーダー観測シミュレーションによる小型プロファイラーの推定バイアス問題の検証など。

≫ データデザインチーム

マウスの体節形成現象のモデル化とパラメータ推定:分節形成に関与するタンパク質であるTbx6、Ripply2、Mesp2を免疫蛍光染色で検出し、これらの蛍光画像を簡便かつnon-arbitraryに定量化する方法(GBIQ法)を確立しました。Tbx6の分解が分節境界の確立に必須であり、Ripply2あるいはMesp2のどちらが実際にこの分解に関わっているのかを一般化加法モデルで明らかにしました。これらを元にタ

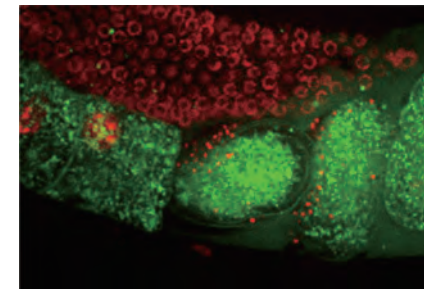


細胞内顆粒の流動立体像を捉えた高速3次元撮影

ンパク質の分解と生成に関わるパラメータを推定し、さらにRipply2の野生型と変異型のパラメータ推定から実際の観察データと数理モデルの同化に挑戦しています。

生殖細胞分化におけるRNA制御モデルの構築とパラメータ推定:RNA-seqとRIP-seqの統合により、Nanos2の標的RNAの候補を導出し、生殖巣での実際の発現の有無を調べます。またNanos2の転写から翻訳・分解までの過程を微分方程式でモデル化し、Nanos2の標的mRNAを推定します。そして、推定された標的遺伝子についてRNA-seqのデータを参照し、Nanos2の標的に特徴的な遺伝子構造があるのか否か、隠れマルコフモデルを援用して探索しました。

その他:細胞核内の染色体の動きから染色体物性を推定する「線虫の染色体構造」、「線虫の細胞質流動」ほか。またデータデザインチームでは、発生・細胞生物学分野においてこのような課題を持つ国内外の研究者に、問題解決する統計手法の普及、実験生物学と最先端の統計学を結びつける共同研究の拡大、および人材育成を行いました。



染色体細胞内顆粒を捉えた多色タイムラプス撮影

Project

e-サイエンス基盤技術

異分野の知と人を共有し、
共同研究促進の基盤を構築します。

爆発的な勢いで増加しつつある研究資源と研究情報……これを研究者が理解し、分野固有の文化や空間的制約に妨げられることなく分野横断的な研究が促進されるよう機械が支援する「サイエンス3.0基盤」構築プロジェクトです。より具体的には研究資源・研究情報を機械可読化するために、大規模データマイニング技術・検索技術、既存学術データベースと連携するためのセマンティックウェブ技術、異種データベースをつなぐ情報リンケージ技術、オントロジー技術等の研究開発を推進します。

本研究の実施にあたっては、プロジェクトを以下の3つのサブテーマに分け、これらの要素技術を統合した上で、データ構造の可視化・探索支援技術を盛り込み、研究者にサービスするための基盤の研究開発を行って、実際に大学・国内主要研究グループ・学会等に提供しながら、実証的に研究を推進しました。なお本研究の実施に先立って、日本全国の研究者を対象とした研究者向けサイエンス2.0基盤サービス「researchmap」を公開。平成23年度にはReaDとの統合を果たし、24万人の研究者が参加する世界最大級の研究者コミュニティへと飛躍的に成長しています。

研究者の情報探索を支援するための基盤技術(サブテーマ1)

本サブプロジェクトの目的は、分野や言語を横断して論文を推薦するシステムの実現手法について引き続き検討を進め、実証的なシステムを構築することにあります。具体的には、トピックモデルに基づき

Project director

新井紀子〔国立情報学研究所〕



構築した日英が混在する意味空間を用いて、言語横断論文推薦機能を実現し、最新の国際会議論文を対象にデモシステムを立ち上げて有効性を検証しました。また、深層学習や機械翻訳による関連情報の検索手法の開発にも取り組み、きめの細かな情報推薦を可能としています。さらに、関連研究者の研究支援のために、2013年度に公開したXML文書の解析ツール PlaneTextの改良および機能拡張も行いました。

LOD 技術による学術データの共有・再利用の促進(サブテーマ2)

学術データをより容易かつ高度に共有・再利用するための基盤として、Linked Open Data (LOD) 技術を用いたデータハブ構築を行いました。DBpedia Japaneseといった一般的な概念に関するハブと、生物種名に関するハブといった特定分野におけるハブの両方の構築に取り組んでいます。またLODによ

る関係性に関わるデータマイニングや関係性そのものの可視化も重要な課題であり、このような作業を通じて、データのいっそうの活用促進を図っています。さらに国立極地研究所やDBLCSとも連携を深めてデータ基盤を拡充し、またリサーチコモンズ「地球環境データ」との連携により、科学データにおけるメタデータのLOD化とその活用にも取り組みました。

融合研究を加速するための

情報共有クラウドサービスの確立(サブテーマ3)

現在運用中のResearchmapは、日本の研究者24万人の登録を集めた、世界最大級の研究者データ



サイエンス2.0基盤サービス「researchmap」

ベースとしての意義と役割を担うようになってきました。Researchmapのデータベースには数百に及ぶ項目で分類された、数千万に及ぶアイテムが格納されており、このような情報基盤から知的探索を行うことは、研究者自身や科学の進むべき方針策定にあたっても有意義であることは間違いありません。そこでこのような知的探索を支援する技術として、この上でサブテーマ1・2の要素技術に基づき、名寄せ・探索・推薦を柔軟に行うための実証研究を行いました。



Linked Open Data (LOD)の代表例である「DBpedia」概念図(サブテーマ2)



名寄せ・探索・推薦を柔軟に行うための実証研究から(サブテーマ3)

歴史を蔵する極地の稀少なデータから、
地球生命システムを解明します。

本プロジェクトは、情報・システム研究機構の4研究所に加え、北海道大学、筑波大学、千葉大、東京大学、日本海洋技術研究機構(JAMSTEC)、東京工業大学、玉川大学、京都大学、京都府立大学、広島大学等、国内外の研究機関との連携体制の下で進められました。また国立極地研究所は、ドームふじ氷床コア・コンソーシアム(ICC)、南極昭和基地、北極スバルバル、日本ニーオルスン基地の利用等について、北海道大学低温科学研究所、北見工業大学、JAMSTEC、アラスカ大学国際北極研究センター等と、共同研究のMOUを交わしています。

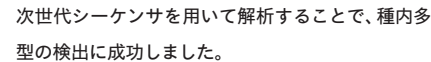
氷河や氷床のコア解析によって、氷に閉じ込められてきた過去100万年間の地球生命史を明らかにしよ

うというテーマで研究が進められてきました。具体的成果としては、氷床中のカルシウムイオン濃度の高精度測定により、氷期開始期などにおける二酸化炭素濃度の減少と気温の低下のメカニズムがいつそう明らかになりました。また古い氷床コア試料からメタゲノムライブラリーを構築し、微生物群の系統推定、遺伝子群の機能解析を目的としたパイプラインの構築等を行いました。これらを南極氷床下の試料に適用して解読したところ、シアノバクテリアおよびアーキアなどが検出され、ドームふじ基地から得られた南極氷床下試料には、多様な微生物群が存在していることが明らかになりました。

また、全ゲノム増幅法の改良による氷河中の花粉一粒ずつの DNA 分析法の検討を重ね、100年前のサンプルからでも種同定可能なレベルの解析手法をほぼ確立しました。さらに、増幅された DNA 断片を

Project director

本山 秀明〔国立極地研究所〕



➤ 極限環境における生物多様性とそのパターン
(サブテーマ2)

南極沿岸域の水床、氷河、湖沼生態系から見た地球環境の変遷と、そこに見られる極限生物の多様性及び分布パターンの解明を目指して研究を進めてきました。南極の極貧栄養湖に繁茂するコケ坊主については、コケ坊主生態系における物質循環構造の全貌解明へ向けて、既着手の16S/18S rRNA遺伝子解析に加え、物質生産・物質循環に関与する遺伝子群の大規模解析を行いました。また南極産地衣類に共存する微生物相を16SrRNAを用いて解析し、多様な共存細菌相を検出しました。

一方、沿岸域の水床末端・水床上・水床下などの境界領域を、水床を取り巻く自由水環境と位置づけ、そこに存在する生物圏を探索しました。南極ばかりでなく、アフリカ、ウガンダの高山山頂付近に見られる熱帯水河上という特殊な環境下に生育する生物集塊に注目し、これらが蘚類やネノウエノアカゴケの仮根と無性芽からなる、まったく新しい水河上の生物現象であることを明らかにしました。

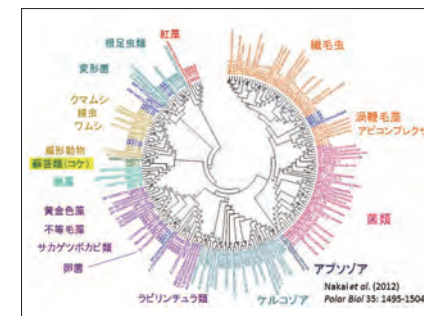


南極の極貧栄養湖に繁茂するコケ坊主

➤ 極限生物の環境適応メカニズムと進化
(サブテーマ3)

コケ坊主生物圏を対象に、生物の極限環境への適応メカニズムの解明を目指しました。コケ坊主共生生物相については、コケ坊主本体及びその直下、周辺の藻類堆積物についても解析を進め、南極湖沼底のコケ坊主生態系の全容解明を行いました。このメタゲノム情報を詳細に解析することで、物質生産・物質循環に関与する機能遺伝子群を同定したほか、コケの内生菌として分離した細菌および真菌の機能遺伝子の解析も実施しました。またコケ坊主から分離した細菌の12株についてはゲノム完全配列を決定し、生育至適温度とGC組成の対応等の解析から、細菌の種分化と適応進化過程を明らかにしました。同時に系統的に離れた細菌種間での大規模な遺伝子の水平伝播が明らかになりました。

南極線虫を対象とした大規模RNA解析、および全ゲノム解析を行い、乾燥・凍結耐性遺伝子やトレハロース代謝遺伝子などのストレス耐性遺伝子の同定に成功しました。また南極クマムシのゲノム解析を進めています。



コケ坊主内の真核生物多様性

Project

遺伝機能システム

表現型と遺伝的ネットワークの
相関を解く新手法を目指します。

ヒトゲノム解読計画以来、生命科学は大量のゲノム情報を基盤とするデータ駆動型へとパラダイムシフトし、近年の超並列大規模シーケンサの登場は世界的にもその勢いを加速させています。そこで本プロジェクトでは、最新のゲノムテクノロジーを駆使して系統的にゲノム・遺伝子関連情報の大規模生産を行う一方、多元的な表現型データを系統的に収集し、統計学に基づくその解析方法の研究開発を通じて、多数の遺伝因子の高次連関から形成される生物多様性をシステムとして理解することを目指しました。

本プロジェクトでは遺伝学と統計科学者それぞれの研究コミュニティが邂逅し、3つのサブテーマを設けて、データ中心科学ならではの新しい原理や解釈を提案できるよう融合研究を推進してきました。

- 次世代シーケンサによるゲノム関連情報の大規模生産とその情報解析手法の開発
- 大量ゲノム関連データと多元的な生物表現型多様性データの統合による遺伝的相関構造抽出のための統計手法の開発と最適化
- 大量で多元的なデータの情報・統計手法を適用したゲノム機能と遺伝的ネットワーク抽出

≫ 次世代シーケンサによるゲノム情報生産と解析 (サブテーマ1)

最新のゲノムテクノロジーを駆使して年あたりペタバイト級の超大容量ゲノム・遺伝子関連データを系

Project director

倉田 のり〔国立遺伝学研究所〕



統的に生産する研究開発に取り組みました。データ処理をクラウド型で行う解析パイプライン開発を進め、ユーザの解析システムと DNAPod の登録システムを統合しオルソログ遺伝子領域で多型差を計算するシステム間多型比較ツール「OrthologAln」を作成。またシステム間多型のデータベースの生物種数を拡大し、既存イネ約 700 系統に加えてトウモロコシ 404 系統とソルガム 66 系統の一塩基多型と挿入・欠失多型の注釈データベースを構築しました。さらに重要な細胞機能であるゲノム分配メカニズム解析については、セントロメアができるために必要なヒストン修飾(H4K20me1)を明らかにしました。平行して、ゲノム機能調節メカニズムであるメチル化とトランスポゾン転移機構解析について、シロイヌナズナを研究材料に全ゲノムマッピングおよびDNAメチル化プロファイリングを進め、セントロメア近傍にトランスポゾンが集積する原因を解明しました。

≫ 遺伝的相関構造抽出の統計手法の開発と最適化 (サブテーマ2)

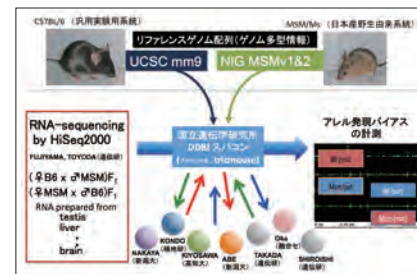
遺伝子型および表現型ゲノムデータの解析における多重比較、検定、モデル選択といった統計的推測問題について、解析手法の開発とその有効性の検証を行なうとともに、具体的な問題を解決しました。また適宜Rなどのツールとして公開しました。たとえばコンソミックマウスのQTL解析においては、Lasso 型の回帰分析を用いてQTLの存在範囲を推定する方法を開発しました。また複数の形質を同時に比較する多重検定問題においては、見せかけの相関を局外パラメータの推定量に用いて多重性を調整する手法を確立し、その数値的特性を明らかにして最適な疑似相関の用い方を模索するなどの成果を挙げています。

さらに遺伝データ解析でしばしば用いられている主成分分析を、スパース推定の枠組みで定式化しました。この手法は通常別々に行われる、高次元説明変数の次元圧縮と予測式の構築を、一段階で効率的に行なうこと、連続変数だけでなく離散変数への拡張も可能にします。この他、コンソミックマウスの行動状態の自動推定アルゴリズムの精度向上と攻撃行動の推定、コンソミックマウス成長曲線の同時信頼区間と多重比較による解析、マウス超音波発声データの解析などにおいて、生命現象の解明に役立つ新たな統計手法を提案しました。

≫ ゲノム機能と遺伝的ネットワーク抽出 (サブテーマ3)

マウス、イネ、ゼブラフィッシュ、ショウジョウバエなどの多様な遺伝資源研究素材を用いて、サブテーマ1が生み出す大量データを活用し、サブテーマ2による統計的な検証を経て、生命・遺伝現象に関するメカニズムの新たな手法での解明を進めました。たとえばマウスでは、複数のマウス実験および交配システムを用いて収集・解析した多様な形質と多様な系統の遺伝子型のデータをカタログ化し、データベースを作成し、表現型差に関連した遺伝子群の発現差とその発現調節エレメントとの関連性の解明を進めました。また野生由来近交系マウス8系統を基に、高い遺伝的分解能を持つ野生由来ヘテロジニアスストックを独自に作製し、この集団を用いたゲノムと表現型の相関解析を行う手法を確立しました。

野生イネを用いた研究では、表現型と遺伝型との相関解析のための多様な形質(表現型や代謝産物)の反復調査と、e-QTL解析のための実験交配による固定化集団の作成を終了し、アソシエーション解析手法の改良を行い、遺伝型対表現型の相関の高効率抽出を行いました。一方ゼブラフィッシュでは、網羅的な遺伝子機能情報・表現型情報の抽出を基に、国立情報学研究所との共同研究により特定の神経活動を可視化するイメージングシステムを開発し、神経ネットワーク解析を行ないました。この他、新しい遺伝子破壊法を用いて、ターゲット遺伝子を欠失させたヌル変異体を体系的に作出しました。



SNP 情報に基づいたアレル発現バイアス解析プラットフォーム

Project

社会コミュニケーション

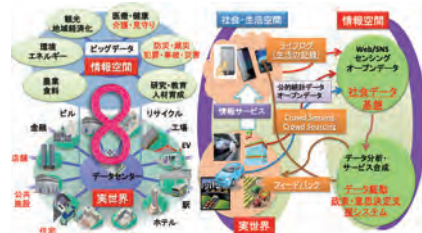
現代の複雑なシステムの問題を見据え、人間・社会科学と融合した新手法を創成します。

ICTの発達によりあらゆる情報機器・端末がネットワークへ接続されて流通し、いつでも、誰もが、どこからでも情報にアクセスできるようになりました。この結果、情報空間と実世界が統合した「サイバー・フィジカル融合社会」が形成されつつあります。このような社会では、情報やデータの力によって社会問題の解決や新たな価値創成を行うことが期待されています。このためには、人間・社会の挙動を感じて、人やモノを制御する情報・サービスを合成し、迅速かつタイムリーにフィードバックする技術的・社会的仕組みが求められていると言えるでしょう。そこで本プロジェクトは、観光・地方創生、防災・減災、医療・健康、教育・ネット学習等、人間・社会のさまざまな課題に関わる、合理的な政策決定や意思決定を支援できるような新しい学問領域「データ中心政策科学」の創成を目指しました。

本プロジェクトは、サイバー空間とリアル空間が統合する融合社会のソーシャル・データに注目し、データベース「人間・社会データ」との有機的連携により研究を推進してきました。観光・地方創生、防災・減災、医療・健康、教育・ネット学習等の政策課題に取り組む自治体(京都市、山梨県、長野県、高知県、大崎市、長崎県、長崎市、仙台市、横浜市、松本市)を情報・システムの側面から支援し、また大学(高知大学、山梨大学、和歌山大学、信州大学、同志社大学、岡山大学、長崎大学、福岡工業大学、宮崎大学)や民間事業者と連携してデータ駆動政策決定システム科学を推進

Project director

曾根原 登〔国立情報学研究所〕



情報空間と実世界が統合した「サイバー・フィジカル融合社会」

しました。このような産官学協働研究の成果は、情報通信研究機構(NICT)の公募型研究開発「ソーシャル・ビッグデータ活用・基盤技術の研究開発」の受託や、第13回ドコモ・モバイル・サイエンス賞の先端技術部門優秀賞、社会科学部門奨励賞の受賞など、高い評価を得ています。

>> 観光・地域経済活性化及び

防災・減災政策決定支援システム

地域自治体などの公共サービスは、科学的根拠データに基づいた合理的な政策決定が望ましいことは

言うまでもありません。しかし従来型の社会調査では、実時間の制御ができないことや、近年の個人情報保護意識の高まりから調査協力が得にくいなどの問題が生じていました。そこでウェブの中にある社会データを収集・分析し、必要な情報を市民に配信し、平時・災害時を問わず持続的運用可能なモデル開発を行いました。

具体的には宿泊設備や交通システムなどWebソーシャル・ビッグデータをオンラインで収集・分析し、スマートフォンなどを用いて、市民・社会に実時間でフィードバックするWebデータ駆動ICT政策決定支援システムを研究開発し、大学参加型による自治体と企業が連携した実証実験を通して社会に実装しました。

>> 学習ライフログデータ駆動型学習支援システム

教育分野はこれまで基本的に集団を対象としてきましたが、ICT技術やネット社会の進展に伴い、電子教科書の導入、MOOCsをはじめとするオンライン教育やeラーニングによる学習、反転授業など新しい教育パラダイムが学習環境を大きく変えつつあります。学習者の学習活動に関する多様な学習・教育データが、大量かつ容易に蓄積され、利用可能になっていく中で、学習者の学習関連データをもとに教育者や学習者に対して、どう教育・学習支援を行うかを明らかにすることは、非常に重要な課題だと言えるでしょう。そこで学習者に適合した個人化教育(personalized learning)へ向けて、学習・教育活動によって生成される「学習ライフログ・ビッグデータ」を活用した、個々の学習者に最適な学習支援の方策を解明する研究を推進しました。

具体的には、Webテストで得られた学習・評価データを認知診断モデルにもとづいて分析し、個々人の学習プロセス・成果を推定します。同時にデータマイニング・可視化手法等を駆使して、学習者の特性や学習スタイル、進度、学習者間コミュニケーション

ンなど、学習改善に資する情報を抽出する方法を確立します。さらに認知診断テストと学習ライフログの両方から得られる結果を融合させることで、学習者個々人のニーズ、学習特性、学習状況に合わせて最適化した学習支援の方法を確立しました。

>> 政策科学・経営科学に資する

意思決定・コミュニケーションプロセス

意思決定・コミュニケーションプロセスに関わる情報循環の高度化実践:リサーチcommons「人間・社会データ」で検討する「意思決定・コミュニケーションプロセスに関わる情報循環の高度化と標準化」の研究成果を、人間行動の不確定性を配慮した政策、投資、スポーツ・ビジネスゲームなどの効率的かつ頑健なシナリオ決定プロセスに応用しました。

自殺予防対策等に関する独自統計整備と予防政策支援:国立精神・神経医療研究センターと共に、引き続き内閣府の次期自殺統計作成に協力し、関連して自殺統計データのモデリングに関する共同研究を推進しました。さまざまな分野の自殺研究者を横断する国内外の共同研究集会を主催し、また内閣府自殺予防センターの立案する政策・評価に資するわが国自殺要因の可視化作業などを推進しました。

安全規準など策定のための数理的方法整備とその食品安全分野などへの実装:食品安全基準、環境基準の策定の基礎となる数理的风险管理技術を整備するとともに、特性値の測定の不確かさ、特性値の重篤なエンドポイントへの量反応関係を勘案した俯瞰的风险管理政策決定様式についての研究を行い、また食品・環境分野の社会実装へ向けて活動しました。

Project

システムズ・レジリエンス

「レジリエンスなシステムをつくる」
知の体系化を目指します。

レジリエンスとは、環境の大きな変化に対して、一時的に機能を失ったとしても柔軟に回復できる能力を指す言葉です。これは私たちの社会が持続可能になるために重要な考え方であり、2011年の東日本大震災では、そのようなシステムを設計・運用するための知識体系の必要性が改めて認識されました。「システムズ・レジリエンス」は、科学者のネットワークを通して、これに答えるプロジェクトです。

レジリエンスの概念は、生態学や心理学などで知られ、その後計算機システムや工学、社会学などでも注目されるものとなっています。このため、本プロジェクトにおいては、情報・システム研究機構の新領域融合センターの仕組みを最大限活用した、真に領域横断的な研究を行ってきました。「レジリエンス」というテーマを通じた海外研究者との交流も積極的に展開し、2014年11月には沖縄において X-Center Global Network 国際ワークショップに参加し、2015年2月には6日間にわたる国際ワークショップ「湘南会議」を開催しました。またドイツ・フライブルグ大学との連携協定も締結しました。

≫ レジリエンスの統合戦略

さまざまな文脈で用いられるレジリエントなシステムに共通な性質として「レジリエンスの戦略」を統合し、科学的意思決定に結びつけるための理論の構築を行いました。レジリエンスの基本戦略は、主に冗長性、多様性、適応性の3つであり、それぞれにコストがかかります。また平時・非常時・回復期など

Project director

丸山 宏〔統計数理研究所〕



の時間的局面において考えることが重要なポイントであることも知られています。具体例な成果としてわれわれは、ステークホルダーから見て、レジリエントなシステムがどのように認知されるかを定量的に評価する進化型マルチエージェントモデルを開発しました。また緊急時における戦略の1つである分離(isolation)に関して、電力網という具体的な対象を想定して効率的な電力網の分離を行うアルゴリズムを構築しました。

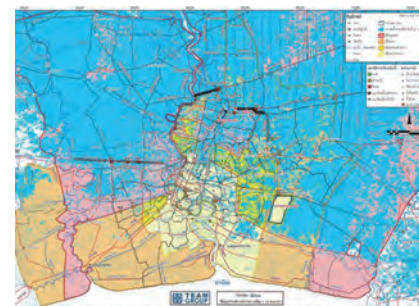
≫ 生物・生態系におけるレジリエンス

集団遺伝学、生態系、生物学における強靱(robust)性とは何か、そのメカニズムについて多様性とレジリエンスの指標を定義し、それらの間の関係を支配する法則を導く研究を行いました。具体的には集団遺伝学におけるWeak Selectionの理論を中心に、どのようなメカニズムが遺伝子の多様性と種の保存に有利に働くかを検討し、凹型適応関数によって理論的に説明できることを示しました。またこれを検証するために祖先型配列の正確な推定法を開発し、

キイロショウジョウバエの近縁種を用いてデータ収集を行いました。

≫ レジリエンスの計算モデル

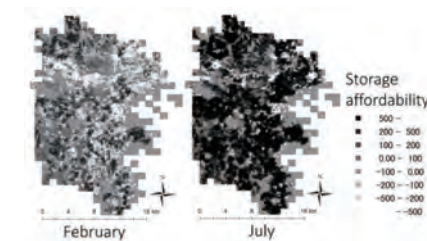
レジリエンスの計算モデルを定義し、レジリエントなシステムの理論的性質を解明する研究を行いました。中でもシステムの現在の状態とその環境を数理的に記述し、環境をシステムに対する制約として捉えた「制約充足問題」として解く、新規性の高い「システムズレジリエンス(SR)モデル」を構築することができました。さらにこのSRモデルを元に、システムが満たすべきさまざまな評価尺度に対して多目的最適化による手法について研究し、さらにダイナミックシステムに関する推論・学習方式について検討しました。



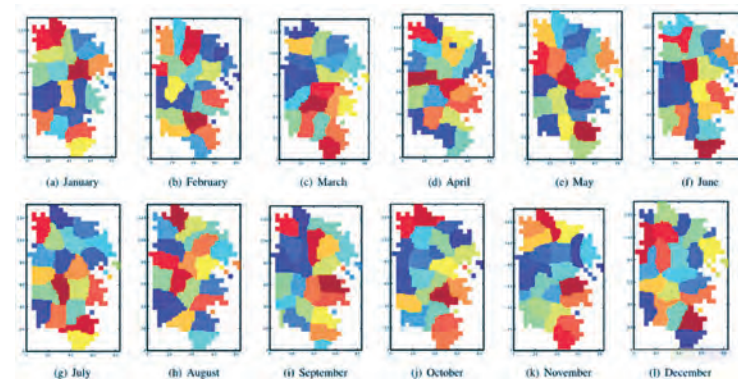
タイ・バンコクにおける洪水マップ(2011)

≫ 社会システム・コミュニティにおける
レジリエンス

レジリエントな社会制度枠組みを設計するための、ケーススタディや計量的手法を用いた研究を行いました。トップダウンな法規制に着目する従来の法制度研究と比して、ICT社会における災害時等に有効な、当事者の自律的なルール形成に着目したアプローチを採り、この共同規制の概念に基づいて、グローバルに流通する分散型仮想通貨をはじめとした新たな経済活動におけるルール形成のあり方、社会における緊急災害時の個人情報共有についての制度的ルール形成のあり方という2つの事象について研究を行いました。また、タイにおける2011年の洪水を例に、市民というステークホルダーが災害をどのように認識しているのか、Twitterデータの解析によって分析しました。



横浜市における電力供給。災害時の最適な分離の計算に利用する



我々のアルゴリズムによって分離された電力網

Project

地球環境データ

PANSY レーダーの

貴重な観測データを解析・配信する
共同利用を確立します。

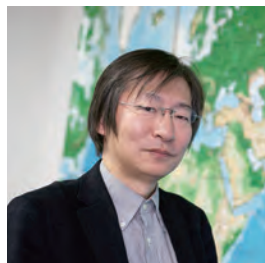
南極昭和基地に建設され、その上空の運動とプラズマパラメータを地上から高度500kmまで高精度で観測する、南極昭和基地大型大気レーダー(PANSY)は、2011年に定常観測が開始され、2014年度からは全群フルシステムで稼動しています(PANSY プロジェクト代表: 東京大学・佐藤薫教授)。極域観測で得られた貴重な地球環境データを、人類の資産としてアーカイブし、利用に供するという大きな責務に向けて、レーダー操作およびデータ解析・配信を担うプロジェクトです。

PANSY から送られてくるデータは、衛星などからは得られない大気の実際の状態を、高度域を含む広いエリアにわたって観測した高精細データであり、全球大気モデル発展の基礎になるものとして、学術コミュニティから待望されています。そこで本プロジェクトは国立極地研究所・東京大学・京都大学が中心となって推進し、国内外の大学・研究機関と連携した共同利用・共同研究を展開しています。北海道大学、東北大学、電気通信大学、信州大学、名古屋大学、九州大学、琉球大学、気象研究所、国立環境研究所、海洋研究開発機構、宇宙航空研究開発機構、情報通信研究機構等に加え、国際科学会議(ICSU)の下でのSCAR(Scientific Committee on Antarctic Research)、SCOSTEP(Scientific Committee on Solar-Terrestrial Physics)、COSPAR(Committee on Space Research)、WCPR(World Climate Research Programme)等の国際コミュニティとも

Project director

データベース事業・部門責任者

中村卓司〔国立極地研究所〕



国際研究集会等を通じて交流を図り、併せて若手育成に取り組んでいます。なおPANSYプロジェクトは、その開発の功績に対して2014年度文部科学大臣表彰を、2015年には海洋立国推進功労者表彰(内閣総理大臣賞)を授与されました。

南極昭和基地大型大気レーダー(PANSY)の

主な特徴:

- 1,045基のアクティブ・フェーズド・アレイ・アンテナ
- 対流圏・成層圏・中間層の風、電離圏のプラズマ精密観測
- 現象の3次元構造を捕捉可能グローバルなエネルギー収支の定量化や大気結合の解明
- 内外学術コミュニティとの連携による気候変動・温暖化の将来予測モデルの高精化

≫ PANSY データを活かす

「地球環境データ基盤整備」

PANSY データ解析センターとして、南極にあるPANSY レーダーの運用を国内からサポートしています。定常観測に加え、2015年度にはPANSY 主導で12カ国が参加する国際共同観測キャンペーンも実施しました。またレーダー本体を制御する「リアルタイムデータ解析ソフトウェア」の機能を向上して、PANSY の55群のアクティブアレイを活かした多チャネル観測を稼動させました。データ同化の手法を用いた解析により、風速高精度推定および将来的には乱流強度高精度推定等が行えます。この他、不要電波の除去による高品質化、流星など新たな観測に応える高機能化など、さまざまなシステム拡充も図りました。さらに地形・地表や海水との相互作用が注目される、高度1500メートル以下の地上付近の連続観測システムの開発にも着手しました。



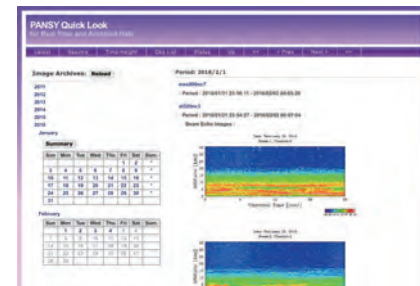
南極昭和基地大型大気レーダー
(Program of the Antarctic Syowa MST/IS(PANSY) radar)

≫ データ利用・データサイエンスを促進する

「科学データの公開と共有」

PANSY のシステムを南極初の大型大気レーダーデータを所望していた研究者だけでなく、データ・サイエンティストや天文・気象等の隣接分野研究者・市民が幅広く利用できるオープン・サイエンス基盤へと拡充していく努力が地球環境データ部門(PANSY データ解析センター)で行われています。たとえば2014年より地上～下部成層圏の6時間値の風速データを蓄積したデータベースを公開し、国内外からのアクセスを得てきました。また時間・高度断面データの追加や、国内研究者用ミラーサーバーの設置などの改良も行いました。

南極で大量に取得される地球環境のビックデータを効率よく国内に転送し、ユーザーに展開していく仕事は、地道ながら数々の技術チャレンジを克服してのみ可能なやりがいのある仕事です。今後もデータで人類の資産を形成するというような大きな視点で取り組んで行きたいと考えています。



対流圏・成層圏エコー強度の時間・高度断面のクイックルック画像(2016/02/01の24時間)。クイックルックシステムではリアルタイムでのエコー強度プロットのほか、過去の観測日時を選択して、観測高度(対流圏・成層圏と中間圏)、観測量(エコー強度と風速)などについて日/週/月ごとの時間高度断面を表示することが可能。

Project

ライフサイエンス統合データベースセンター (DBCLS)

ライフサイエンスデータの統合化&国際標準化に取り組みます。

1990年代後半から急速に進展したゲノム科学により、生命科学に関わる情報がさまざまな形式で大量に生成・蓄積されるようになりました。全国の大学、研究機関などにばらばらに存在するデータベースの保全と共有化を通じて生命科学分野の知識をまるごと有効活用するためのデータベース統合化の最新技術・研究開発とサービス提供を行う専門機関として、2007年にライフサイエンス統合データベースセンター(Database Center for Life Science, DBCLS)が設立されました。「統合データベース」に関する将来的なビジョンを国際的な研究コミュニティと共有し、その標準化に深く関わりながら、基盤技術の開発を継続的に行っています。

DBCLSは、科学技術振興機構(JST)「ライフサイエンスデータベース統合推進事業」(2011年度より)を実施するJSTのバイオサイエンスデータベースセンター(NBDC)と共同し、また理化学研究所、産業技術総合研究所、京都大学化学研究所、東京大学をはじめとする多くの機関と連携して、生命科学分野のデータ統合化に必要な基盤技術を中心とした研究開発を行っています。

ライフサイエンス統合データベースセンター(DBCLS)の主な事業内容:

- データベース統合化のための基盤技術開発(RDF化の推進)

Project director

小原雄治

[DBCLS]



- インターネットを活用した高度な検索を可能にする技術の開発
- ハイスループットシーケンスを中心とする大規模データの利用技術開発
- アノテーション、キュレーションの支援技術開発
- 国際開発者会議「BioHackathon」の開催
- 生命科学分野の研究者を対象にした日本語による情報提供

>> RDFデータの流通を促進する

RDFポータル構築

わが国では数多くのデータベースが存在しているものの、独自の形式や語彙で記述され、孤立してい

るものが多い現状にあります。DBCLSでは設立の早期からRDF化によるデータベース統合とその技術開発を推進してきましたが、データベースのRDF化ガイドライン作成を経て、2015年にはRDF形式による生命科学分野データベースを1ヶ所に集約した日本初のポータルサイト「NBDC RDFポータル」を構築、NBDCから公開しました。収載RDFデータセットは随時追加していく予定です。近年、世界の主要な生命科学研究拠点である欧州バイオインフォマティクス研究所(EBI)、米国立生物工学情報センター(NCBI)、スイスバイオインフォマティクス研究所(SIB)、スタンフォード大学等から、さまざまなデータベースが、相互参照や連結が容易なRDF形式で配布される潮流がありますが、DBCLSの活動は次に記す国際標準化の動きとともに、この潮流の先端のひとつとなっています。

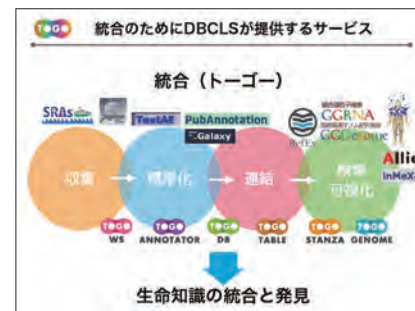
>> 国際的RDF仕様を策定する「BioHackathon」

世界的な生命科学分野のデータベース開発コミュニティにおいて、国際標準となる仕様を検討・策定するハッカソンの会「国際版BioHackathon」を年1回開催してきました。またこの成果を国内の現場のデータベース開発者・研

究者に引き継ぎ、RDF関連技術の普及と向上を促すために、定期的に「国内版BioHackathon」や「SPARQLthon」を行うなど、国内外の生命科学分野RDFデータの開発コミュニティ運営に主導的な役割を果たしてきました。

>> RDFデータ利用へ向けた研究開発

データベースをRDF化する努力と並行して、既存のリレーショナルDBに蓄積されたデータを、RDFと同様に検索できる「D2RQ Mapper」を開発しました。利用者はリレーショナルDBへも、RDFの問い合わせ言語であるSPARQLを用いてシームレスに問い合わせることができるようになります。また、RDFデータを利用したアプリケーション開発のための開発者向けフレームワーク「TogoStanza」も提供しました。これを視覚化ライブラリが利用できるJavaScriptによる開発が行えるように改造したり、利用を促すポータルサイト「togostanza.org」の公開などを進めるなど、開発環境を向上させました。さらにRDFデータで実装されたゲノムデータベース「TogoGenome」は、2015年度には新規に125の真核生物のゲノム情報を追加し、RDFデータ数16.8億トリプルへ拡張しました。



統合のためにDBCLSが提供するサービス概念図



「NBDC RDFポータル」

Project

人間・社会データ

政策にも活かせるサイバーフィジカル
融合社会のデータベースを構築します。

人間・社会の複雑な状況や行動などを反映する多種多様なデータが時々刻々と集積される、ビッグデータ時代。国勢調査などの公的統計データ、自治体オープンデータ、宿泊施設など Web ソーシャル・データ、無線アクセスのロケーション・データ、ネット教育における学習ライフログ・データ等を収集し、学術・研究・教育目的で利用可能な「リアルタイム人間・社会データ共同利用基盤」基盤を整備します。

本データベース基盤整備は、ソーシャル・ビッグデータ基盤の構築とデータ駆動の政策決定支援システムの研究開発へ向けて、プロジェクト「社会コミュニケーション」との有機的連携により推進しています。国立情報学研究所と統計数理研究所を中核拠点として、大学群（一橋大学、京都大学、広島大学、北海道大学、山梨大学、信州大学、同志社大学、大阪大学、東北大学、九州大学、広島大学、岡山大学、福岡工業大学、長崎大学、島根大学、山口大学医学部、高知大学医学部、宮崎大学医学部等）と独立行政法人統計センター、自治体、産業界と連携しています。そしてこの共同利用基盤を活用し、山梨県 情報産業振興、長野県 松本市商工部、京都市 産業観光局観光MICE推進、大崎市 ICT街づくり、高知県 高知医療再生機構、仙台市 総務企画局情報政策部、高知県健康政策部医療政策・へき地医療支援、広島市 経済観光局雇用推進、長崎県・長崎市 観光政策等の自治体・産業界と連携し、大学参加型の「データ駆動社会イノベーション」を推進しています。

Project director

曾根原 登〔国立情報学研究所〕



≫ Web/SNS データ駆動政策科学のための高度人間・社会データ基盤の整備

観光など地域経済活性化や防災・減災政策などの社会問題解決に役立ち、科学的意思決定を支援するようなデータとして、Web/SNS/モバイル空間のソーシャル・ビッグデータを対象としたデータ基盤を構築してきました。このデータの利活用として、発信者としての市民に注目し、Facebook など SNS の行動履歴・行動分析するための SNS アプリケーション基盤の構築や、観光・防災政策決定支援に必要なデータの抽出方法を検討しました。また自治体等のステイクホルダーとしての市民から、質の高い意見を収集するオンライン・リアルタイムデータ収集システム開発にも取り組みました。

≫ データ駆動型教育システムのための学習ライフログデータ収集基盤

情報空間上に大量に蓄積される学習ライフログから、データマイニング・可視化手法等によって、学習者の特性や学習スタイル・進度・学習者間コミュニケーション

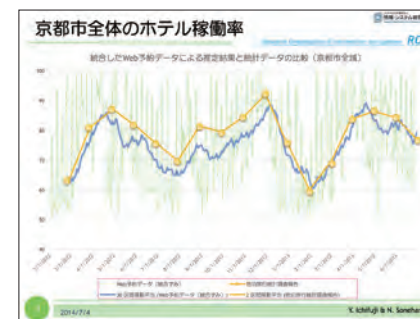
ンなど、学習改善に資する情報を抽出する方法を検討しました。また Web 上の認知診断テストから個人の学習プロセス・成果を推定し、この2つの方法を融合させることで、個人に最適化した学習支援の方法の確立にも取り組んでいます。

≫ オープンなソーシャル・ビックデータ連携基盤

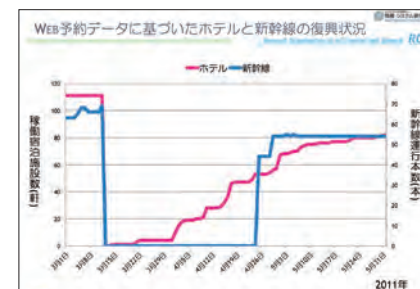
災害時に特定の地域に対して、被災者のプライバシー情報を積極的に開示するなど、時間軸・空間軸を制限した人間・社会データのプライバシー情報保護活用基盤を構築しています。具体的には、利用者が個人情報取り扱いを条件付きで事前承諾するような仕組みとしての「ID データコモンズ」の考え方や、個人・民間・行政の保有する個人情報を連携させて危機に強い人間・社会データ基盤を実現を目指す点に特徴があります。

≫ 人間・社会データ収集・利用加速の基盤整備

高品質な政府公的統計が利用できる、2011年度統計センター認可の政府公的統計「オンサイト解析室」を整備しています。また産学官の共同研究に資する人間・社会データベースとして、まず一般社団法人 CRD 協会の保有する企業財務データを基に「中小企業財務データベース」の構築を行っています。次に産官学における情報管理・利用・検索のためのデータベースを考え、一般社団法人日本規格協会の協力の下、この産学研究コミュニティの活性化を促す SNS システム「VCP-NET (Value Creation Process Network)」の運用を開始しました。さらに島根大学と共同で、レセプト等からなる健康科学データに関わる情報基盤構築を目指しています。一方、人間行動の不確定性を配慮した政策・投資・スポーツなどのシナリオ決定プロセスや、集団間コミュニケーションに関わる決定理論等の共同研究を推進し、国際規格 ISO TC69 SC8 における意思決定支援プロセスの標準化を進めています。



新たなリアルタイム観光・宿泊統計(京都市)。1ヶ月単位で測定する宿泊旅行統計調査報告(オレンジ線)と、30日間平均を日ごとに表示する推定稼働率(青線)



ソーシャル・ビッグデータによる宿泊・交通システムのレジリエンス可視化(仙台市)。本研究は、2015年度モバイル・サイエンス賞(社会科学)を受賞した

Project

データ中心ケミストリ

未知の化学反応を創発する
データケミストリをデザインします。

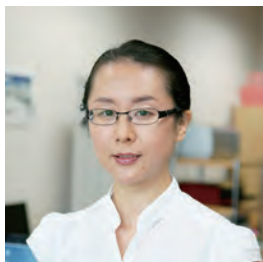
現在までに存在が確認されている化学物質は1億種類を超え、年間数10万～100万種オーダーで加速度的に増え続けています。しかし理論的には存在し得るものの、いまだ人類が手にしていない化学物質種数は、これをはるかに凌駕することが明らかになりつつあります。そこで本プロジェクトでは、まずこの革新をもたらす物質を「埋蔵分子」と名付け、埋蔵分子とこれらの供給方法である化学反応経路を量子化学の理論に基づいて自動的に探索・発掘し、データケミストリの視点から研究基盤の整備に向けて研究を進めています。

本プロジェクトは国立情報学研究所、統計数理研究所、東北大学、量子化学探索研究所、京都大学、東京大学、京都産業大学、スイス連邦工科大学に所属する研究者と共同で行ってきました。データベース構築を進めるとともに、データの可視化と解析のためのソフトウェアの開発、化学への応用研究、ならびにデータ構築の高速・自動化も、プロジェクトの重要な柱です。応用研究では、まだ実験的には発見されていないが、理論的には存在することが予測される「埋蔵分子」や「埋蔵反応経路」の発掘、グローバル反応マップデータのデータマイニングによる理論的な化学反応の分類と体系化などを目指しています。

≫「データ中心ケミストリ」というデータ基盤
量子力学をベースに電子の動きまで考慮できる「量子化学計算」は、化学反応過程も含む物質の精密な構造と電子的性質を予測することができます。これ

Project director

佐藤寛子〔国立情報学研究所〕



に基づき、分子のポテンシャル曲面を自動探索する画期的な手法を用いて得られるグローバル化学反応経路マップのデータを蓄積し、有用な化学物質を創ることを目的とする化学研究において重要なプロセスである「分子設計」と「合成・反応経路設計」に資する理論的データ基盤を構築します。

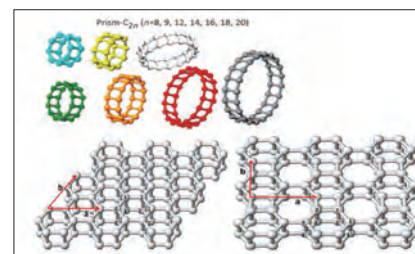
グローバル反応マップは、ポテンシャル曲面の極小点に位置する安定に存在する構造(平衡構造)、鞍点に位置する反応が起こる際に経路する最もエネルギーの高い構造(遷移状態)、結合が解離した状態(解離チャンネル)のそれぞれの構造が化学反応経路によってネットワークのように連結されたもので、ポテンシャルエネルギーや電子状態に関する物理化学パラメータのデータも包含されています。

≫ 埋蔵分子発掘ツール「RMapViewer」の

公開と拡張

このグローバル化学反応経路マップから人が埋蔵分子・反応を発掘するためのツール「RMapViewer」を、2014年7月に無償公開し、随時改良を行ってきました。最新バージョンには反応マップのグラフィカル表示、分子構造や化学反応経路の可視化、マップ内の最適経路の探索をはじめグローバル化学反応経路マップの解析のために必要な基本機能が搭載されています。さらに探索計算高速化のための分散化、入力ファイル生成から探索計算、データ蓄積までのプロセスの自動化についても検討を進めています。

また2015年度には、より汎用的な言語であるSCALAに移植を行い、並行して機能強化も実施しました。これをオープンソースとして公開後、さらに分子構造による検索やタブレットによる描画入力などの機能の開発と実装の作業を進める計画です。なおRMapViewer活用促進のため、本ソフトウェアを利用して作成したさまざまな化学反応アニメーションを、YouTubeを通じて公開しました。

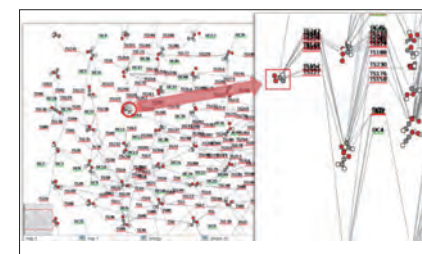


「埋蔵分子」発見！

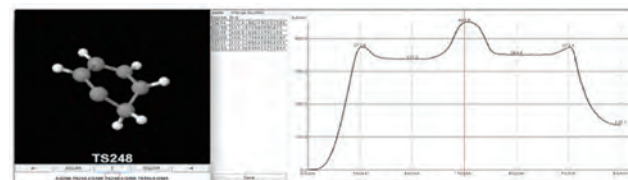
≫ 応用研究で相次ぐ「埋蔵分子発見！」

グローバル反応マップデータの解析をきっかけに、埋蔵分子と呼んでよいと思われる超原子価分子の候補構造が発掘され、また新規炭素クラスターファミリーの存在が理論的に予測されました。また応用研究として、RMapViewerの経路検索機能を用いて糖の立体配座探索を行ったところ、2種類の椅子型配座の最安定構造間に複数の最小エネルギー経路の存在が発見されるなど、埋蔵分子・埋蔵反応経路の発掘に関する新しい成果が相次いでいます。

これらの研究を継続するとともに新規分子系・反応系にも着手し、また実験化学者との協働により、理論的に存在が予測される分子合成の実証実験についても検討を始めています。なおグローバル反応マップデータは、これらの応用研究によっても蓄積されるほか、系統的な分子式リストに沿って順次追加するなどして、継続的にデータ基盤拡充を図っています。



C₂H₂O₂の反応マップ:全体マップ図(左)と、エネルギープロファイル図(右)



化学反応経路に沿ったムービー表示:
ベンゼンからフルベンへの最小エネルギー経路



Research View 018

Research View 018

北川 源四郎

新領域融合研究センター センター長

松井 知子

イメージデータ解析PD

吉田 亮

イメージデータ解析

大学共同利用機関法人 情報・システム研究機構 機構長。1971年東京大学理学部数学科卒、1974年、東京大学博士課程から統計数理研究所研究員となる。同研究所助教、教授を経て、2002～2011年統計数理研究所長。理学博士（1983九州大学）。専門は時系列解析、情報量統計学、計算統計学。1996年の逐次モンテカルロ法に代表される理論研究に加え、統計的手法を駆使した地震波の自動処理、船舶のオートパイロット、経済時系列の季節調整法、列車安全のための強風予測法の開発などにも携わった。2011年より現職。日本学術会議会員、総合研究大学院大学名誉教授。

統計数理研究所 モデリング研究系 統計的機械学習研究センター教授。博士（工学、東工大）。専門は統計的機械学習、音声・画像情報処理。1986年東京工業大学卒、1988年同大学院修士課程修了。同年NTT（株）に入社し、音声研究に従事。1998年よりATR研究所に出向。2003年より統計数理研究所准教授、2008年より同教授。2010年には同副所長、2011～2014年には同研究主幹を務める。2001年1～6月米ルーセント・テクノロジー社ベル研究所客員研究員他、海外の最先端的研究拠点で研鑽を積む。IEEE、電子情報通信学会、日本音響学会、日本統計学会各会員。

統計数理研究所 モデリング研究系・データ同化研究開発センター准教授。総合研究大学院大学 複合科学研究科 統計科学専攻 博士課程修了、2004年博士（統計科学）。データサイエンスを方法論面の駆動力とし、生命科学や物質・材料科学の分野で研究活動を展開。国立研究開発法人 物質材料研究機構招聘研究員、総合研究大学院大学 複合科学研究科 統計科学専攻准教授を兼任。JST-CREST生命動態領域「神経系まるごとの観測データに基づく神経回路の動作特性の解明」、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」に参画。



Research View 019



Research View 020



Research View 021

宮尾 祐介

メタ知識構造解析PD

佐藤 薫

地球環境データ

樋口 知之

データ同化・シミュレーション支援技術

国立情報学研究所 コンテンツ科学研究系准教授。博士。東京大学大学院情報学環助手、情報理工学系研究科助教等を経て、2010年4月から現職。専門は自然言語処理、中でも文法理論に基づく深い構文解析の研究で知られる。言語の構造の言語学的・数理的モデルを追究する一方、それを実用アプリケーションで実証する研究にも取り組む。大規模テキストを利用して文法を獲得する、高速かつ高精度な構文解析器「Enju」他。2012年文部科学大臣表彰若手科学者賞、2015年日本学術振興会賞、2014年6月情報処理学会長尾真記念特別賞他受賞多数。

東京大学 大学院理学系研究科地球惑星科学専攻教授。専門は大気力学、中層大気科学。1986年東京大学大学院理学系研究科修士課程修了、1991年理学博士（京都大学）。京都大学助手等を経て、1999年国立極地研究所助教授、2005年より現職。高分解能観測・高精度モデル・理論を同時かつ密接に取り扱い、データ解析を主軸としたデータ中心科学に取り組む。極地研と共同して、対流圏から電離圏までをカバーする大型大気レーダー「PANSY」の建設と観測を推進。2005年からPANSYプロジェクト代表、リサーチコモンズ「地球環境データ」プロジェクト プログラムディレクター。

統計数理研究所 所長・教授。1984年東京大学 理学部地球物理学科卒、1989年理学博士。1989年統計数理研究所助手、2002年同教授、2004年同副所長を経て、2011年より現職。専門はベイジアンモデリング。フォワード（前向き）推論としてのシミュレーション計算とバックワード（後ろ向き）推論としてのデータ解析をつなぎ、常に変動する現象の予測・解析・制御を目指す。データサイエンティスト育成に加え、現代科学が対象とする複雑な現象の不確実性に挑むものとしての「統計」観に立ち、「データ同化」手法の他分野への普及啓発活動にも取り組む。



Research View 022

山下 智志

データ同化・シミュレーション支援技術

統計数理研究所 データ科学研究系教授。1987年京都大学 工学部交通土木工学科卒、1989年同大学院工学系研究科応用システム専攻。博士(工学、京都大学)。安田信託銀行 投資研究部、熊本大学 工学部助手を経て、1994年統計数理研究所助手、1997年同助教授、その後助教授・准教授を経て、2011年同教授。2013年統計数理研究所 リスク解析戦略研究センター センター長。1998年マサチューセッツ工科大学 客員准教授、2001～2011年金融庁金融研究研修センター特別研究官を併任。



Research View 023

植竹 淳

地球・環境システム

情報・システム研究機構 新領域融合研究センター 融合プロジェクト特任研究員。理学博士。専門は雪氷微生物学、環境動態解析。2010年から現職。2014年にアメリカ・ウガンダの高山、ルウェンゾリ山頂付近の熱帯氷河において雪氷生物試料の採取と観測を行い、世界で初めて氷河上に蘚類の無性芽からなる集合体を発見。熱帯高山、南極、グリーンランド等の世界中の極限環境を舞台に数々のフィールド調査を重ねつつ、培養株を用いた生理生態学的実験や分子系統解析を行っている。



Research View 024

新井 紀子

e-サイエンス基盤技術PD

国立情報学研究所 社会共有知研究センター センター長・教授。一橋大学法学部およびイリノイ大学卒業、イリノイ大学大学院数学科修了。博士(理学)。専門は数理論理学(証明論)・知識共有・協調学習・数学教育。2001年より、教育機関・公共機関向けの情報共有基盤システム NetCommons を開発。2009年より学術研究情報の循環型情報活用基盤システム Researchmap 開発に携わる。2011年より人工知能プロジェクト「ロボットは東大に入れるか」プロジェクトディレクターを務める。著書多数。



Research View 025

丸山 宏

システムズ・レジリエンスPD

統計数理研究所 モデリング研究系 サービス科学研究センター センター長・教授。1981年東京工業大学理学部情報科学科卒、博士(工学、京都大学)。専門分野は自然言語処理、XML、Web サービス、情報セキュリティ。1983年日本アイ・ビー・エム(株)入社、2009年までの26年間に在籍し、東京基礎研究所長や執行役員を務める。巨大なグローバル企業で、コンピューター言語XMLの研究・開発と標準化など最先端の研究・開発に携わった。1997～2000年、東京工業大学情報理工学研究科 客員教授等を経て、2011年から現職。



Research View 026

武田 英明

e-サイエンス基盤技術

国立情報学研究所 情報学ブリンシプル研究系教授。総合研究大学院大学複合科学研究科情報学専攻教授。1991年東京大学工学系研究科修了。工学博士。ノルウェー工科大学、奈良先端科学技術大学院大学を経て、2006年より現職。2005～2008年東京大学客員教授、2006～2010年国立情報学研究所 学術コンテンツサービス研究開発センター センター長。専門はWeb情報学、人工知能、設計学。ジャパンリンクセンター(JaLC) 共同運営委員会委員長、ORCID (Open Researcher and Contributor ID) の理事、NPO 法人リンクト・オープン・データ・イニシアティブ理事長も務める。



Research View 027

持橋 大地

メタ知識構造解析

統計数理研究所数理・推論研究系准教授。1993年東京大学入学、1998年奈良先端科学技術大学院大学 情報科学研究科 博士課程へ進学。2005年同大学院修了、博士(理学)。ATR 音声言語コミュニケーション研究所、NTT コミュニケーション科学基礎研究所を経て、2011年より統計数理研究所准教授。専門は、統計的自然言語処理、特にベイズ統計に基づく生成モデルと意味的处理、および連続系との接続に取り組む。2015年 日本学術振興会 グローバル学術情報センター分析研究員、2015年 産業技術総合研究所 人工知能研究センター客員研究員他を兼務。



Research View 028

城石 俊彦
遺伝機能システム

国立遺伝学研究所副所長・教授。1981年東北大学大学院理学研究科博士課程修了(理博)。1984年国立遺伝学研究所助手、1998年同教授。1999年同研究所系統生物研究センター センター長、2013年より現職。専門は哺乳類遺伝学。マウスの形態形成機構やエネルギー代謝などの遺伝制御をテーマに、ゲノム解析や遺伝解析を用いた研究を推進する。1999～2008年、理研ゲノム科学総合研究センターのプロジェクトディレクターを兼任し、大規模マウス突然変異開発プロジェクトを統括。またリサーチコモンズでは「遺伝機能システム」の初代プロジェクトディレクターを務めた。



Research View 029

栗木 哲
遺伝機能システム

統計数理研究所 数理・推論研究系研究主幹・教授。専門は 統計科学・統計数学。1984年東京大学大学院工学系研究科卒。日本電気株式会社、東京大学助手、信州大学助教授を経て、1995年より統計数理研究所助教授、2005年同教授。総合研究大学院大学教授併任。多重比較法、チューブ法等の統計手法で知られる。『Annals of the Institute of Statistical Mathematic, AISM(統数研編集、Springer刊)General Sectionの編集主幹を務めるなど、数理・統計を通じた国際的な研究者との交流に加え、分野を超えた学際的な研究にも力を注ぐ。



Research View 029

藤澤 洋徳
遺伝機能システム

統計数理研究所 数理・推論研究系教授。博士(理学)。東京工業大学大学院 情報理工学研究科助手、統計数理研究所助教授を経て、2013年より現職。外れ値に強い統計手法、中でも外れ値の割合が大きい場合でも潜在バイアスを小さくする手法で知られる。2008～2015年『Annals of the Institute of Statistical Mathematics (AISM)』のCo-Editor。2014年から応用統計学会理事。2015年から文部科学省委託事業「数学・数理科学と諸科学・産業との協働によるイノベーション創出のための研究促進プログラム」実施担当。



中野 純司
シミュレーション支援技術PD

統計数理研究所 モデリング研究系研究主幹・教授。理学博士(東京工業大学 1988)。日本統計学会理事長(2015-2016年度)。専門は計算機統計学、時系列解析。特にインターネット、並列計算、対話的グラフィックスなどを利用する統計解析システム、大規模データ解析に取り組む。1978年東京大学工学部 計数工学科 卒業、1980年同大学修士課程修了。1980年徳島大学工業短期大学 生産機械工学科助手、1989年埼玉大学大学院政策科学研究科 講師、1992年一橋大学経済学部 助教授を経て、1998年統計数理研究所教授。



本山 秀明
地球・環境システムPD

国立極地研究所教授、気水圏研究グループ長、先進プロジェクト研究グループ長。理学博士(北海道大学1988)。第31次日本南極地域観測隊員以来、南極観測隊への参加は11回に及ぶ。1998年同研究所助教授、2006年教授。氷床コアを軸にした地球環境変動研究や雪氷圏での水循環などに独自の研究成果がある。南極大陸ドームふじ基地では降雪や積雪の通年観測を実施したり3035m深の氷床コア掘削に成功した。北極での氷河調査も多い。これらを基に、現在から過去数十万年までの気候・環境変動の復元や、その変動機構についての研究を進めている。



倉田 のり
遺伝機能システムPD

国立遺伝学研究所 系統生物研究センター センター長・教授。九州大学理学部より農学研究科へ(農学博士)。専門は、植物生殖機構、ゲノム多様性、植物育種遺伝学。三菱生命科学研究所、藤田保健衛生大学助手・講師を経て、1991年より農水省生物資源研究所主任研究官、1996年国立遺伝学研究所助教授、2003年同教授、2009～2013年副所長。イネの染色体研究から始まり、酵母、ショウジョウバエ、ヒト、マウスなど様々な種を対象に分子・細胞生物学研究に携わる。1991年からイネゲノム研究に戻り、以後多様性・進化・発生の複合的研究を展開し、Oryza属23種の解明に挑む。



曾根原 登
人間・社会データPD

国立情報学研究所 情報社会相関研究系研究主幹・教授。工学博士（信州大学1994）。専門は情報流通システム工学、デジタルメディア工学。特に近年はサイバー空間とリアル空間が統合する融合社会のソーシャル・データに注目した研究開発を展開する。1978年日本電信電話公社入社。国際電気通信基礎技術研究所、ATR視聴覚機構研究所、NTTヒューマンインタフェース研究所、NTTサイバーソリューション研究所を経て、2002～2004年東京工業大学大学院理工学研究科 連携講座教授、2004年より現職。



中村 卓司
地球環境データ データベース事業・部門責任者

国立極地研究所副所長、宙空圏研究グループ教授。1984年京都大学工学部卒、博士（工学、京都大学1992）。南極昭和基地大型大気レーダー（PANSY）をはじめとするレーダーを用いた観測を専門とし、極域中層・超高層大気の構造・変動および諸現象の解明を推進する一方、レーダーから得られた貴重な極域大気データの解析・共同利用を推進する。1989～1998年京都大学超高層電波研究センター助手、2003年米国コロラド大学航空宇宙工学科客員教授等を経て、2009年より同研究所教授、2012年より現職。



小原 雄治
ライフサイエンス統合データベースセンター(DBCLS)PD

DBCLSセンター長、国立遺伝学研究所特任教授。専門はゲノム科学。1951年京都生まれ。1974年京都大学理学部生物物理卒業後、大学院は名古屋大学理学部分子生物・岡崎令治教授の元へ進学した。1980年名古屋大学理学部分子生物助手、1985年頃から大腸菌ゲノムの物理地図作成に取り組む。1988年から2年間、英国ケンブリッジMRC分子生物学研究所で客員研究員として、モデル生物「線虫（C.elegans）」に親しむ。国立遺伝学研究所遺伝情報研究センター助教授を経て、1996年教授。2004～2012年国立遺伝学研究所第8代所長を務めた。



佐藤 寛子
データ中心ケミストリPD

情報・システム研究機構 新領域融合研究センター特任准教授。博士（理学、お茶の水女子大学）。基礎学問としての化学へ情報学的にアプローチするケモインフォマティクス、計算化学、データケミストリが専門。1996～1998年理化学研究所 基礎科学特別研究員等を経て、2000年国立情報学研究所助手、2002～2015年准教授。2015年7月より現職。2007～2008年スイス連邦工科大学（ETH）客員教授、2010年よりチューリッヒ大学ゲストサイエンティスト。1998年より理化学研究所客員研究員、2013年より量子化学探索研究所理事を兼務。

「データ中心科学 リサーチコモンズ事業」は、データサイエンスの確立へ向けて、ビックデータ時代における学術の研究基盤となるデータ基盤整備、モデリング解析基盤整備、人材育成を3つの柱として、大学共同利用機関法人情報・システム研究機構が推進してきた研究事業です(2013～2015年度)。プロジェクトディレクターを中心に、8つのプロジェクトと4つのデータベースプロジェクトの計12の研究が推進されており、その多くが、2004年の設立以来の融合研究を継承・発展させたものとなっています。

さて、この研究成果を発信する「リサーチコモンズ ウェブサイト(<http://rc.rois.ac.jp/>)」を、約2年前に公開し、運営を担当してきました。なかでも「Research View」という、インタビュー記事の形式を借りて研究成果を発信する連載枠を設け、ほぼ月1回の頻度で記事を公開してきました。2014年4月～2015年3月掲載分を、昨年度『データに語らせる科学』の冊子にまとめ、このたび、ここに2015年4月～2016年2月掲載および3月掲載予定分を1冊に収載しました。ちなみに、本ページにも参照ページを記す紙の上の「リンク」を記載していますが、本冊子の記事が前年の記事と関連がある場合には、その参照ページも記すことにしました。またタイトルの『未来へつなげるデータサイエンス』には、3年前と比べ、「データサイエンティスト」の存在や、育成の必要性などからいっそう身近になった「データサイエンス」という言葉を用いて、その未来への展望を込めました。

本冊子の制作にあたりましては、北川源四郎機構長をはじめ、樋口知之統計数理研究所所長、城石俊彦国立遺伝学研究所副所長にご尽力いただきましたことを、御礼申し上げます。またわれわれの取材にご協力くださいましたイベント開催関係者の皆様、そして調査中の隊員に代わってご対応いただいた国立極地研究所 伊村智教授、武田先生のインタビューの折、撮影にご協力いただいた国立情報学研究所の皆様、そして誠にお忙しい中ご協力賜りました東京大学佐藤薫先生ならびに全プロジェクトディレクターの先生方に、改めて御礼申し上げます。

なお今回も、池谷が原稿を書き起こして先生方にチェックいただく共著のスタイルでまとめてまいりました。個人的にはこの連載を通じて、これからの、英語圏を含むより多様な記者・一般読者が想定される学術広報へ向けて、これまで書いてきた日本語の文体を大きく変える契機となったことを記しておきます。

2016年2月

大学共同利用機関法人
情報・システム研究機構 URA(広報)
池谷瑠絵