

VOL.2
2021

長崎大学 情報データ科学部

NEWS LETTER

TOPICS

自然言語処理

はたらく細胞

データサイエンス

PCR検査

スパース

サイバーフィジカルシステム

開かれた情報データ科学部をめざして

情報データ科学部 副学部長 喜安 千弥

情報データ科学部よりニュースレター第2号をお届けします。学部のあるさまざまな活動を皆様に直接お伝えするために、直近のコロキウム講演内容をご紹介するほか、PBL科目の目玉である「実社会課題解決プロジェクト」の活動状況や、オープンキャンパスの開催結果をお知らせします。

本学部は、おかげさまで創設2年目となって新たな1年生を迎え、一段と活気を増しています。6月には待望の情報データ科学部棟が完成し、データサイエンス分野の先生方の研究スペースが整いました。いまだにコロナ禍による不自由を余儀なくされているところですが、7月には感染拡大の合間をぬってオープンキャンパスを対面で開催し、県内外から多くの高校生を迎えて、情報科学・データ科学の魅力を知っていただくことができました。

毎月オンラインで開催しているコロキウムは、私どもの研究活動を皆様に知っていただく貴重な機会となっています。ぜひ一度、コロキウムに直接ご来聴ください。また、本ニュースレターで講演の概要をご覧になって興味を持たれた方や、より深く知りたいと思われた方、お手元のニーズを解決する可能性を感じられた方は、ぜひ私どもにお声をおかけください。引き続き、世の中のニーズを敏感に感じ取り、世界に貢献する成果を生み出すべく、開かれた情報データ科学部をめざして参ります。

本ニュースレターが、皆様と本学部をつなぐ新たなきっかけとなることを願っています。



長崎大学
NAGASAKI UNIVERSITY

第4回 情報データ科学部コロキウム

＜新型コロナとデータサイエンス～データで読み解く・分かる社会の動き～＞

日 時 2021年3月17日(水) 17:00～18:00

場 所 オンライン開催(Zoom)

参加者 77名

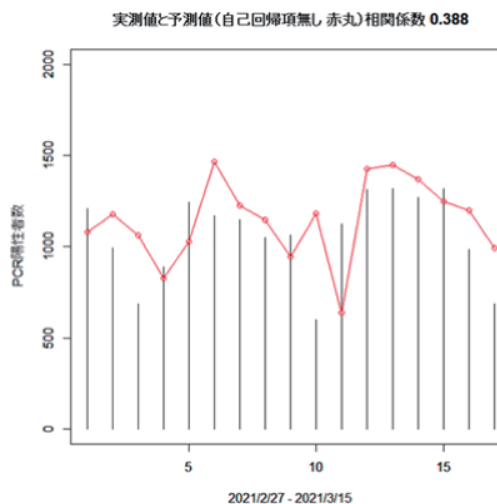
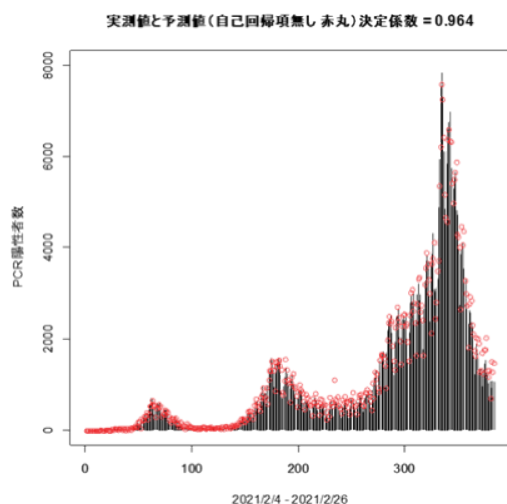
内容	講演者
講演1 日本全国におけるPCR検査陽性者数の時系列解析	西井 龍映
講演2 原因と結果のデータサイエンス：データがないときの分析方法	高橋 将宜

日本全国におけるPCR検査陽性者数の時系列解析

長崎大学 情報データ科学部 学部長 西井 龍映

厚生労働省が公開しているCOVID-19関連データを用いて、PCR検査陽性者数を目的変数とし、検査数、要入院者数、回復者数、死亡者数を説明変数とする時系列モデルを考察した。一週間の周期性を考慮するため、一週間前までの時間遅れを持つ目的変数および説明変数に基づく回帰モデル(ARXモデル)を用いた。また政策評価のため、非常事態宣言1 および 同2、GoToTravelの期間、また曜日効果をダミー変数として説明変数に追加した。モデル推定のための教師データは2020/2/18から2021/2/26まで、評価のためのテストデータは2021/2/27から2021/3/15とした。

説明変数はベイズ型情報量規準BICを用いて逐次的に選択した。選ばれた説明変数により、非常事態宣言2とGoToTravelの期間でモデルが変化していることがわかり、また弱い曜日効果も検出された。下図の左、右はそれぞれ教師データ、テストデータの予測を表し、実測値(黒)とARXモデルによる予測値(赤)の相関係数はそれぞれ0.9984、0.9455と十分な説明力があることが示された。



原因と結果のデータサイエンス： データがないときの分析方法

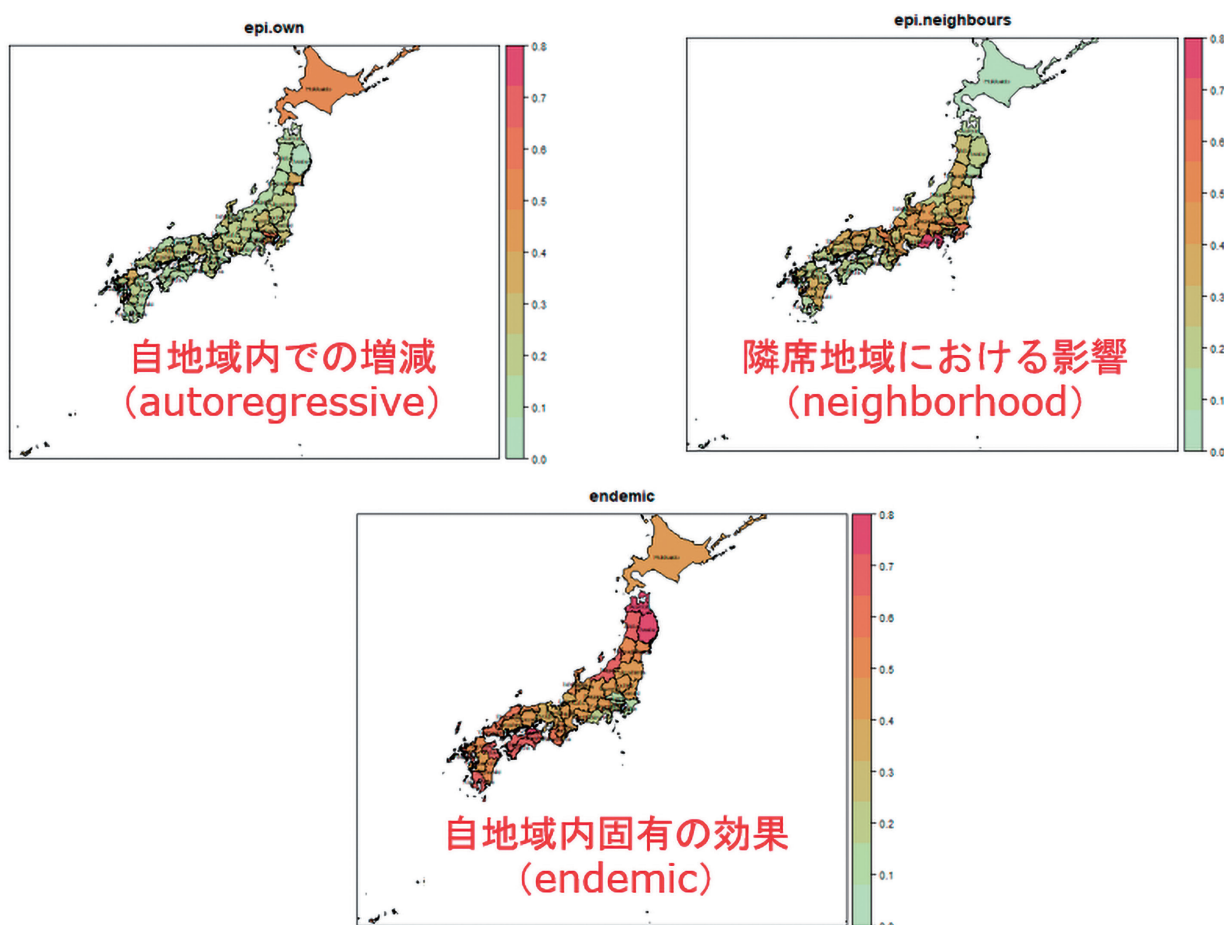
長崎大学 情報データ科学部 准教授 高橋 将宜

本講演では、新型コロナに関する共同研究(共同研究者：植木先生・一藤先生・梅津先生)、欠測データ解析の概論、統計的因果推論の概論に関する話題を提供した。

最初に、時空間計数データモデリングを利用して、新型コロナ感染者数を可視化するために開発した教育用アプリを紹介した。このアプリでは、自地域内での増減、隣席地域における影響、自地域内固有の効果の三要素に分解して、日本地図や時系列プロットとして可視化できる。

次に、欠測データの根本的な問題点として、そのままでは統計解析ができないこと、欠測値を含む行を取り除く方法では分析結果に偏りが生じることを指摘し、代入法によって偏りが是正されることを解説した。また、欠測値を推定したことによる誤差と根本的な誤差の2種類を考慮できる手法として、多重代入法の概略を説明した。

最後に、観察研究における統計的因果推論として、共分散分析と傾向スコアを紹介し、傾向スコアの利点を解説した。また、事前分布を仮定することで、多重代入法によって個体因果効果を推定できる可能性があることにも言及した。



第5回 情報データ科学部コロキウム

<食料と細胞とデータサイエンス>

日 時 2021年4月21日(水) 17:00~18:00

場 所 オンライン開催(Zoom)

参加者 52名

内容	講演者
講演1 食料生産とサイバーフィジカルシステム	持田 恵一
講演2 はたらく細胞データサイエンス	松本 拡高

食料生産とサイバーフィジカルシステム

長崎大学 情報データ科学部 教授 持田 恵一

サイバーフィジカルシステム(Cyber Physical System, CPS)とは、現実世界(Physical System)から得られる様々なデータを、サイバー空間(Cyber System)で解析して得られた結果や知見を、現実世界にフィードバックすることで最適化や課題の解決を実現しようとするシステムである。サイバーフィジカルシステムは、センサーやIT技術の発展により、多様な分野で利用され始めている。

本発表では、バイオ分野でのサイバーフィジカルシステムの利用について、特に作物の生産性の向上を目指した、育種分野での利用に焦点を絞って解説した。作物の収量に関連する農業形質も遺伝的多様性と環境との相互作用で決定される。近年、ゲノム解析技術が各段に進展し、多様な作物で遺伝的多様性の情報の収集が可能になった。また、植物の成長過程での変化を経時的に計測する様々なセンサーやコンピュータ・ビジョンによる画像解析技術が開発されつつある。農業形質を定量化し遺伝要因や環境要因との関係性をモデル化することで、農業形質の改良に有用な遺伝要因の探索する取り組みの具体例を示した。作物の遺伝的変化に加えて、精密農業等による営農技術の高度化も作物の生産性の向上に有用であること、その背景にあるセンサーや情報処理技術についても紹介した。

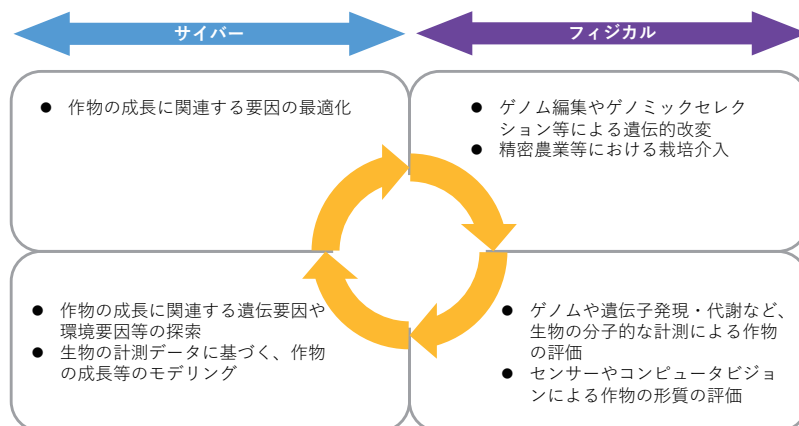


図1. 作物の改良におけるサイバーフィジカルシステム

はたらく細胞データサイエンス

長崎大学 情報データ科学部 准教授 松本 拡高

ゲノム解読技術などの実験技術や計算機の発展に伴い、大規模な生命情報データから生命現象の解明を目指すバイオインフォマティクス研究の発展が著しい。

本発表では、遺伝子の発現量を調べるRNA-seqのデータ解析の話題を、長崎大学内での共同研究とともに紹介した。まず、統計解析を用いて疾患関連遺伝子を同定する基本的なアイデアを解説し(図1)、その上で、iPS細胞などを用いて観測困難な脳の遺伝子発現を擬似的に計測できることや(図2)、ゲノム編集技術を用いてデータサイエンスでは困難な因果推論を実験的に特定できることを紹介した(図3)。さらに、1細胞解像度で遺伝子発現を計測する1細胞RNA-seqの解析技術も解説した(図4)。その他に、血液を「診る」技術であるリキッドバイオプシーに関し、細胞の形態や内容物を用いた疾患研究を一部紹介した。

このように、生命科学研究のさらなる推進には多様な技術を活用することが求められ、幅広い分野の知見をキャッチアップすることが重要となっている。特に、生命科学分野のデータサイエンスであるバイオインフォマティクス研究では、何が観測可能で、データから何をどこまで解くべきかを正しく認識することが重要であることを説いた。

図1

疾患群と対照群の遺伝子発現比較と要因候補の同定

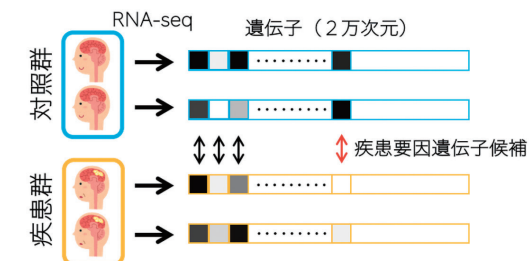


図2

iPS細胞を用いた脳の擬似的な遺伝子発現量計測

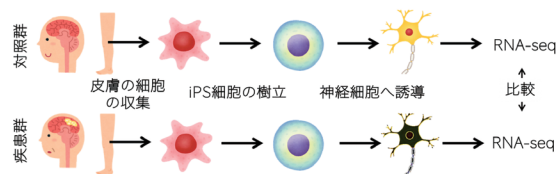


図3

ゲノム編集を用いた疾患要因遺伝子の特定

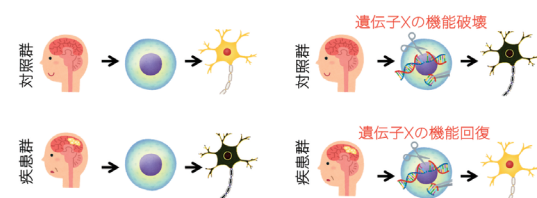
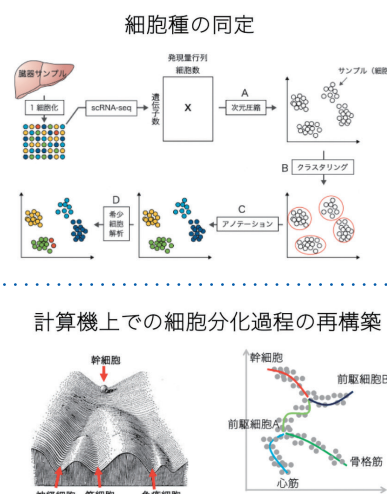


図4

1細胞遺伝子発現解析におけるデータサイエンス



第6回 情報データ科学部コロキウム ＜ビッグデータとデータサイエンス＞

日 時 2021年5月19日(水) 17:00～18:00

場 所 オンライン開催(Zoom)

参加者 44名

内容	講演者
講演1 スパース性に基づくデータサイエンス	梅津 佑太
講演2 自然言語処理における生成タスクの検証	鈴木 郁美

スパース性に基づくデータサイエンス

長崎大学 情報データ科学部 准教授 梅津 佑太

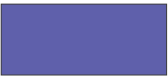
データを解析する際に興味のある特徴は(大規模なデータから見ると)少数である、と考えるのは自然な発想であろう。例えば、遺伝子発現量解析では、ある疾患に影響を与える遺伝子は少数である。また、時系列解析などで、構造が変化する時点もデータから見ると少数である。データサイエンスでは、行列やベクトルとして与えられるデータやパラメータの多くの成分がゼロであるときに「スパース」であるという。したがって、興味のある特徴はこれらが非ゼロである部分に対応している。

本コロキウムでは、モデル選択やスクリーニングのような、スパース性が背後に現れるようなデータ解析手法をいくつか紹介した。また、スパース性と関連のある話題としてpost-selection inferenceについて述べた。データが少なかった時代の古典的な統計解析とは大きく異なる解析ツールが求められる昨今のデータサイエンスでは、モデル選択のような仮説の探索と、選択された仮説の検証を一つのデータから行うことが求められることもあり、post-selection inferenceは今後ますます重要な役割を果たすことが期待される。

ビッグデータ



- Tall Data
 - Huge sample & Moderate dimension
 - ▶ 単一細胞解析, SNS, IoT, ...



- Wide Data
 - Moderate sample & Huge dimension
 - ▶ マイクロアレイ, 画像, ...

スパース

- **スパース** (Sparse, 疎)
 - 【形】(物・人間・動物・植物などの存在量が少なくて) まばらな, わずかな, 希薄な¹
 - ↔ **デンス** (Dense, 密) 【形】ぎっしり詰まった, 隙間がない²
- データから見ると興味のある特徴はスパース
 - グラフの隣接関係, 画像の輪郭
 - 外れ値, 変化点
 - ある現象に寄与する変数

¹<https://eovf.alc.co.jp/search?q=sparse>
²<https://eovf.alc.co.jp/search?q=dense>

自然言語処理における生成タスクの検証

長崎大学 情報データ科学部 准教授 鈴木 郁美

近年、自然言語処理において、深層学習を用いた言語モデルの研究が盛んに行われ、様々な自然言語処理のタスクでテキストの自動生成が着目されている。深層学習により生成されたテキストが実用的であるためには、自動生成されたテキストが正しい文法に従っていることが重要である。英語の場合、最近の報告では、深層学習により学習された言語モデルは正しい文法を捉える能力を示すことが報告された (Marvin and Linzen, 2018)。

一方、日本語の文法能力についての検証は行われる必要がある。我々は、日本語におけるターゲット評価方法の確立を目的とした研究を行っている。今回は特に、日本語格文法における格助詞と動詞の関係に注目した。評価のためのコーパス自動作成、深層学習に評価を反映させるための損失関数を設計することで、文法能力を我々が確立したターゲット評価法により評価を行うのみならず、文法能力を獲得した深層学習のモデル構築を行った。

自然文
閉路 を 持た ない 無向 グラフ を 森 と いう。

変更後
閉路 \$より 持た ない 無向 グラフ を 森 と いう。

動詞「持つ」が取らない格助詞「より」に変更

動詞「いう」は全ての格助詞を取りうるので変更なし

2020年度 実社会課題解決プロジェクト 活動の総括

前期の活動 初年次セミナーと連動し、長崎の地域課題に対する解決方略を提案

観光PR、離島の積極的活用、若者世代の人口流出、坂道の多い環境、道路の整備などの地域課題を21グループに分かれた学生らが設定し、それぞれ議論を重ねた後に解決に向けた方略を提案した。SNSの利用やスマートフォンアプリの開発のみならず「聖地巡礼」のように、アニメ・映画など文化的な背景を踏まえた提案もあり柔軟なアイデアは注目に値した。



後期の活動 参画事業者と取り組む課題解決(8つの課題に22グループを配置)

前期成果発表に基づいてマッチングが成立した継続課題4つと、新たに事業者から出していただいた4つの計8課題に分かれて取り組んだ。希望調査によりグループを再編成したことで、チームビルディングに苦労したグループもいくつかあった。一方で、アンケート調査やインタビュー調査の実施と分析をしたり、画像認識のシステムのプロトタイピングをして作成したプログラムでの実験を行ったグループがあるなど、短期間ながら成果を示すことができたグループもあった。



2021年度 長崎大学オープンキャンパス

日時 2021年7月17日(土) 10:00～16:00

場所 長崎大学文教キャンパス
(オンライン実況視聴あり)

2020年4月に新設された「情報データ科学部」では、入学を検討している学生さんや、新たな学部に興味をもつ方々へ、学部の雰囲気から研究への取り組みまで幅広く体験いただけるオープンキャンパスを開催した。情報科学やデータ科学そのもの、またそれらを活かすあらゆる学問領域に興味を持つ多様な皆様のご参加を期待している。

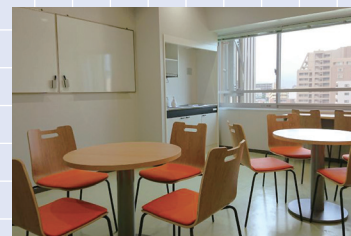


INFORMATION

6月11日(金)
情報データ科学部棟 完成!



情報データ科学部 玄関



リフレッシュスペース



長崎大学 総合生産科学域事務部西地区事務課 総務第一係(情報系)
〒852-8521 長崎市文教町1-14 TEL.095-800-4101

詳しくはHPへ!

YouTubeチャンネル



情報データ科学部

HP <https://www.idsci.nagasaki-u.ac.jp/>