複製・再配布などの二次利用はご遠慮ください

第1章 データ駆動型科学における計測インフォマティクスの意義

東京科学大学
安藤 康伸

はじめに

研究、開発の現場では、分光法をはじめとする計測技術の高度化に伴って日々の実験から得られるデータの量は増え続けており、1台の装置が1日に出力するスペクトルの数は数年前と比べて大きく増加した。一方で、得られたデータを前処理し、有用な情報へと変換する作業の多くは、依然として研究者個人の経験と勘に依存したままで、計測装置の進歩にデータを解析し活用するための体制が追いついていない。これが多くの研究・開発現場に共通する状況だろう。

このような状況を打開する手段として注目されているのが、計測から得られるデータに情報科学的な手法を組み合わせて活用する「計測インフォマティクス」である。計測インフォマティクスは、マテリアルズ・インフォマティクス（Materials Informatics：MI）をはじめとするデータ駆動型科学の一分野として位置付けられ、計測データの蓄積、前処理、統計的機械学習による解析までを一連の流れとして扱う点に特徴がある。

本章では、計測インフォマティクスがなぜ求められているのかを整理した上で、本書の第2章から第4章で扱うデータ構築・人材育成、前処理と解析環境、統計的機械学習、そしてスペクトルデータ解析事例が、それぞれどのように位置付けられるのかを概観する。後続の各章を読み進めるための見取り図として活用していただきたい。

SAMPLE

1 計測インフォマティクスとは

1.1 データ駆動型科学における位置付け

マテリアルズ・インフォマティクスは、材料の組成、構造、物性に関するデータを情報科学的な手法で解析し、新材料の探索や物性予測に役立てる研究分野として発展してきた。その成果の代表例として固体電解質の分野で知られているのが、液体電解質に匹敵する、最も高いリチウムイオン伝導度を示す物質であるLGPS（$Li_{10}GeP_2S_{12}$）にまつわるエピソードである。LGPSは、東京工業大学の菅野了次教授のグループが2011年に発見した。この発見を受けて、MITのGerbrand Cederらのグループは、第一原理計算によってLGPSのイオン伝導機構を明らかにするとともに[1)]、Geをより安価なSiやSnなどに置換した類似の物質群を網羅的に計算することで、有望な候補物質

複製・再配布などの二次利用はご遠慮ください

を短期間で予測した[2]。この一連の研究にはサムスン先端技術研究所（SAIT）も参画しており、実験による発見を起点に、計算科学的な探索によって関連物質群の可能性を広げた事例として知られている。

同様の枠組みは他の分野にも見られる。京都大学の田中 功教授のグループは、第一原理計算とベイズ最適化を組み合わせた探索により、熱伝導率の低い新規化合物を予測した[3]。また、統計数理研究所の吉田 亮教授らのグループは、分子設計アルゴリズムとベイズ最適化を組み合わせた探索的な高分子設計により、高い熱伝導率を持つ高分子材料を見いだしている[4]。こうした成果の多くは、計算科学によって生成されたデータを基として研究・開発されたものである。

一方で、実験、計測から得られるデータを対象とした情報科学の活用は、計算科学の分野に比べて発展が遅れていた。理由の1つは、計測データが装置や測定条件に依存したノイズや誤差を含んでおり、そのままでは機械学習の入力として扱いにくい点にある。計測インフォマティクスは、こうした実験計測データ特有の性質を踏まえた上で、データの蓄積から前処理、モデリング、解析までを体系立てて扱う分野として、マテリアルズ・インフォマティクスの一角を占めるようになってきた。

1.2　計測現場が抱える課題

計測インフォマティクスが求められる背景には、計測現場が抱えるいくつかの課題がある。第一に、スペクトルのピーク位置や強度を読み取る作業の多くが、担当者の経験に依存している点である。実際、同一のデータを複数の研究者が解析しても、得られるパラメータにばらつきが生じることは珍しくない。第二に、計測装置の高度化やロボット技術による試料作製、計測の自動化が進んだことで、人手による解析が追いつかない量のデータが出力可能になった点である。そして第三に、得られたデータやその解析結果が個人の端末やノートにとどまり、組織として蓄積・再利用されにくい点である。

第一の課題、すなわち属人性が解析結果に影響する例として、粉末X線回折パターンの解析に用いられるリートベルト解析が挙げられる。リートベルト解析では、結晶構造に関わるパラメータに加えて、バックグラウンド形状やピーク形状を表す設定パラメータを数十個程度扱う必要があり、熟練者であっても徐々にパラメータ空間を絞り込みながら最適化を進めることが多い。そのため、同じデータを解析しても、担当者ごとに推定されるパラメータがばらついてしまう。尾崎嘉彦らは、ベイズ最適化を用いて設定パラメータの探索を自動化する方法を報告し、熟練者による解析結果と同程度、あるいはそれを上回る構造を見いだせることを示した[5]。担当者の経験に依存していた作業も、高度な最適化パッケージを導入することで、再現可能な手続きに置き換えられることを示した例である。

計測現場が抱えるこれらの課題に対応するためには、計測データを前処理・解析す

SAMPLE

複製・再配布などの二次利用はご遠慮ください

るための手法だけでなく、データを組織として蓄積し、担当者が入れ替わっても解析のノウハウが引き継がれる仕組みが欠かせない。次項以降では、この2つの側面、すなわちデータ構築・人材育成と、前処理・解析のための統計的機械学習について、本書の構成に沿って概観する。

2　データ駆動型研究開発を支える基盤:データ構築と人材育成

2.1　データ蓄積環境の整備とデータベース構築

計測インフォマティクスを実践するための出発点は、計測データを解析可能な形で蓄積すること、すなわちデータベースの構築である。多くの研究・開発現場では、計測装置から出力されたデータがそのままファイルサーバーや個人のパソコンに保存されており、後から横断的に検索・活用することが難しい状態にある。せっかく取得したデータも、これでは活用可能なデータであるとは言いにくい。測定条件や試料情報などのメタデータとともにデータを整備してこそ、過去の計測データを機械学習の学習データとして再利用できるようになるのである。

データベースの構築に当たっては、CSVやJSONといったフラットファイル形式で個別に保存する方法と、複数のテーブルを関連付けて管理する関係データベースを用いる方法があり、扱うデータの量や検索、分析の要件に応じて使い分ける必要がある。また、こうしたデータ蓄積を支援する仕組みとして、近年は民間企業によるSaaS（Software as a Service）事業も展開されている。代表的な例がトヨタ自動車の開発した「WAVEBASE」[6]であり、研究データの蓄積、活用を支援する「Randeft」[7]のようなクラウドサービスも登場している。第2章では、このような取り組みも含めて、R&Dにおけるデータ蓄積環境の整備とデータベース構築について詳しく解説する。

2.2　人材の育成と仕組み化

データベースを整備しても、それを活用できる人材がいなければ意味をなさない。計測インフォマティクスの実践には、計測技術に関する知識と、統計学や機械学習に関する知識の両方が必要になる。もちろん、全ての研究者が情報科学の専門家になる必要はない。しかし、計測と情報の両方を見渡して課題を設定できる人材は少なく、多くの現場で人材の不足が課題となっている。

人材育成には、個々の研究者への教育に加えて、解析のノウハウを個人にとどめず組織の資産として蓄積する「仕組み化」が欠かせない。第2章では、人材の育成と仕組み化、技術の蓄積について、具体的な取り組みと併せて紹介する。

SAMPLE