

## ② データサイエンス分野

データサイエンス分野では、収集されたデータから科学的あるいは社会的に有益な新しい知見を取り出すデータ解析法を研究します。この分野のアプローチは、人手に頼らず、コンピュータを用いて自動的にデータ解析できるため、2つの利点(1)大規模なデータ(いわゆるビックデータ)を扱えること、(2)客観的な視点を持って結論を導けること(特定の偏った考え方に結論が依存されず、常識を覆す発見の可能性が高まる)があります。計測の自動化や高度化にともないデータ取得コストが下がっているために、様々な研究分野や実世界応用分野において期待されている分野となります。以下では、この分野の特徴を述べ、そして、私が最近取り組んでいる物質材料科学のための情報学(マテリアルズ・インフォマティクス)を例に紹介します。



中心研究者  
工学部電気電子・情報工学科  
准教授 志賀 元紀

データサイエンス分野(あるいは情報科学分野)の中にも、様々な研究分野があります。例えば、データ発生機構における確率的な振る舞いを踏まえてデータ解析結果を評価する統計学、観測されたデータから規則を発見したり未知の事象を予測したりする機械学習(あるいはデータマイニング)、システムの調節・パラメータ学習の効率化を追求する最適化理論などがあります。また、ビックデータを扱うためには、計算機を効率的に処理するためのアルゴリズム論・データ圧縮・簡潔データ表現などの理論も必要となります。このように、データサイエンス分野は様々な数学や基礎理論の上に成り立っていますが、そのために、応用分野の研究者や技術者が基礎となる数学を全てを理解して使いこなすのは容易ではありません。最近流行りの人工知能(AI: Artificial Intelligence)のシステムを構築するためには、大規模なデータをコン

ピュータに入力してデータから目的のタスクを達成するために必要な規則をシステムに学習させる必要があります。最近、人工知能システムを構築するための無料パッケージが公開されていますが、正しく使用するためには、ある程度の理解と専門家のお墨付きが必要になるかと考えています。そのための支援・技術交流のために、本研究センターの意義があると考えています。

データサイエンス分野の研究者は、汎用データに役立つ手法だけでなく、個々のデータに特化して最高の性能を発揮する手法の研究もしています。最近の私の研究では、物質・材料科学のデータを対象とした情報学(マテリアルズ・インフォマティクス)を精力的に研究しています。以下に、2つの研究事例を紹介します。

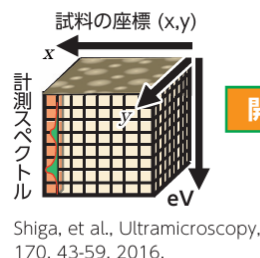
### 研究例A

#### 網羅的な波形(スペクトル)データのクラスタ解析・予測法

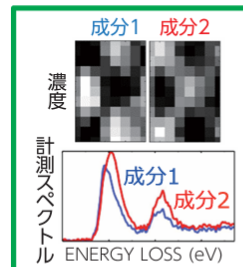
様々な実験条件・位置で計測された多数のスペクトルの波形形状を用いて、実験条件や位置をグループ分けする手法を開発しました。これによって、各計測点に含まれる化学成分などを同定でき、材料の構造を自動的に抽出できます。新たに作成した試料を客観的に(実験者の主観を除いて)評価したい場合、また、膨大な数のデータを自動処理するのに向いています。

#### 顕微分光データから自動的な物質構造の同定

##### 特性スペクトルイメージ



##### 本手法の出力(原子配置)



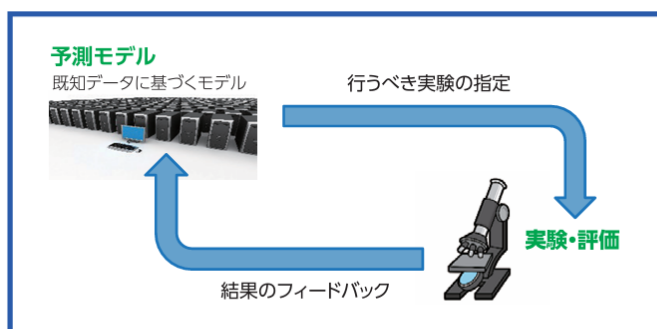
Shiga, et al., Ultramicroscopy, 170, 43-59, 2016.

### 研究例B

#### 新規材料探索のための効率的な実験計画

所望の物性・材料特性を持つように化学組成を最適化するための手法を開発しています。一度の実験で最適なものを発見することは難しいので、実験を反復する回数を減らすことを目指しています。最適なものを得られる確率を少しでも増やすことを目指しており、ベイズ最適化などの機械学習技術を用います。

上記の応用分野以外に、一般に、医療(ゲノム医療、創薬)や自動運転などの様々な応用が研究・実用化されています。この分野の研究ではアイデアや発想がとても大事です。データを抽象的に捉え、既存手法から新しい応用を生み出すだけでも十分に面白いです。私が数年前に研究を始めたマテリアルズインフォマティクスも、始めた当時は日本であまり知られていませんでしたが、現在では産業界も非常に注目し投資して



いる研究分野となっています。各研究者が蓄積したデータが原石であれば、データサイエンス分野の手法によって輝く宝石になることは十分ありえると考えています。そうした原石を探すのも本センターの重要なミッションと考えています。