ワンポイント・ブックレビュー

竹村彰通著『データサイエンス入門』岩波新書（2018年）

データサイエンティストといわれたとき、どのような人をイメージするだろうか。本書によれば、「データを処理し分析し、データから価値を引き出すことのできる人材」がデータサイエンティストとなる。スマートフォンの急速な普及も含め、ビッグデータ時代が加速し、「21世紀の石油」といわれるまでに情報やデータが氾濫している今だからこそ、そのような人材が必要になってきている。ただし、世界的にみれば、日本のデータサイエンティスト育成は遅れている。著者は2017年４月から日本で初めての滋賀大学のデータサイエンス学部の学部長としてその設立に携わっており、日本におけるデータサイエンスの第一人者であるといえるが、日本の大学でこの分野を研究できるところはまだ少ないのが現状である。

そもそもデータサイエンスは、データ処理、データ分析、価値創造の３要素を組み合わせて、何らかの意味のある情報や関連を見出すための学問・研究である。本書はこの入門として、「Ⅰ　ビッグデータの時代」「Ⅱ　データとは何か」「Ⅲ　データに語らせるー発見の科学へ向けたスキルー」の３つの領域について書かれている。

まず、「Ⅰ　ビッグデータの時代」では、データサイエンスの登場から今までが、統計学の流れを踏まえながら簡潔に説明されている。そして「Ⅱ　データとは何か」ではデータの作られ方から処理の仕方、情報の引き出し方や読み方といったところまで、さまざまな例を出しながら検討されており、本書の主眼というべき領域である。なかでも面白いと思ったのは、認知心理学や社会心理学における「確証バイアス」と「後知恵バイアス」について論じている部分である。「確証バイアス」とはいわば結論ありきでデータを読んでしまうこと、「後知恵バイアス」とは物事が起きてから予測可能だったと考えること（「そうなると思った！」）なのだが、これらを避けるために事前（データを取る前）にどのような項目についてデータを取り、どの項目を評価するか、検討することが重要だという。一方で、ビッグデータの多くはこのような事前検討ができないデータ・情報であるから、専門家であるデータサイエンティストがデータ・情報から適正に価値を引き出し、具体的な行動につなげる意思決定を手助けすることが必要だと指摘されている。最後の「Ⅲ　データに語らせる」は、スキルの学び方や使えるソフトウェアの紹介などがされている。

著者がもともと数理統計の研究者であることから、統計に関する説明、それを背景とするデータの読み方、分析手法といった部分の内容は非常にわかりやすく、面白い。一方で、コンピュータ関連の内容においては、入門ということでそれほど深く入り込まず、表面的なものとなっているせいか、コンピュータに関する知識がないと少し読みにくい部分もある。ただ、入門書としてデータサイエンスに触れるには手ごろな一冊であると思う。（加藤　健志）