特集「データサイエンスを支えるツール-方法論」に当たって

椿　広計
（大学共同利用機関法人情報・システム研究機構統計数理研究所長）

今月号の特集で扱うのは、近年勃興してきたデータサイエンスを支えるツールや方法論である。そもそも、データサイエンスをどう捉えるのか、統計科学と何が異なるのかについては、国内外でも様々な説がある。2019年に、IT分野で発行した国際規格ISO/IEC 20546“Information technology-Big data-Overview and vocabulary”（https://www.iso.org/obp/ui#iso:std:iso-iec:20546:ed-1:v1:en）では、「データサイエンス」が「発見と呼ばれる仮説と仮説検証のプロセスを通じて、データからアクションに繋がる知識を抽出する」と定義されている。

アメリカ合衆国労働統計局では、これまでも職業分類で数学職としての15-2021の数学家、あるいは15-2041も統計家を集計しており、2021年現在、それぞれ2,000名、34,200名が従事している。

“Occupational Outlook Handbook（以下、OOH）”というWEBページには、統計家などの専門職がどのような業界や地域に何名従事しているか、その俸給はどの程度かなど興味深い情報が掲載されている。労働統計局は、統計家を数学の理論と技術を適用して、ビジネス、工学、科学及びその他の分野の実問題を解決する専門職としている。

このOOHに、2023年２月14日に、数学職としてのデータサイエンティスト（職業分類15-2051）の掲載が開始された（https://www.bls.gov/ooh/math/data-scientists.htm（visited February 14, 2023））。2021年現在就労者数は、113,300名である。

統計家とデータサイエンティストは、何を行う専門職かという記述はOOH上では次のように類似している。

1．統計家：特定の質問や問題に回答するために必要なデータを決定
　データサイエンティスト:プロジェクトで利用可能な有用なデータを決定
2．統計家：データ収集のための調査、実験を設計
　データサイエンティスト：データの収集、分類、分析
3．統計家：データ分析の数理モデル・統計モデルを開発
　データサイエンティスト:アルゴリズムとモデルの作成、検証、テスト更新
4．統計家：統計ソフトウェアを使用してデータ分析し、ビジネスの意思決定を支援する可視化資料を作成
　データサイエンティスト：データの可視化ソフトウェアを用いて検討結果を提示
5．統計家：データを解釈し、専門家・非専門家に分析結果を伝達
　データサイエンティスト:データ分析に基づいてステークホルダーにビジネスを提言

筆者も統計家に求められているこの種のミッションについては理解しており、1980年代から急速に実用化が進んだ統計ソフトウェアなどを活用した統計的問題解決に関心を持ち続けてき

た。しかし、組織全体にデータサイエンスのアーキテクチャーが組み込まれる時代の進展を目のあたりにして、データアナリティクス側面の周辺テーマが組織の価値創生に大きな役割を果たすことを感じるようになった。特に注目したのは、アメリカの人気高ランクのデータサイエンティスト育成専門職修士コースの１つであるUC Berkleyの社会人向けプログラムが強調した５フェイズからなるデータサイエンスライフサイクルというプロセスである。

１．Capture
　データ取得、データ入力、信号抽出、データ抽出
２．Maintain
　データウェアハウジング、データクレンジング、データステージング、データプロセッシング、データアーキテクチャー
３．Process
　データマイニング、クラスタリング／分類、データモデリング、データ要約
４．Analyze
　探索と検証、予測分析、回帰、テキストマイニング、定性的分析
５．Communicate
　データレポート、データの可視化、ビジネスインテリジェンス、意思決定

これら各フェイズで、ビジネスなどの価値を最大化するための教育体系が構成されている（https://ischoolonline.berkeley.edu/data-science/what-is-data-science/）。これがプロフェッショナルとしてのデータサイエンティスト育成プログラムで何を教育するかを俯瞰した青写真に思われる。

今回の特集の目的は、これらデータサイエンスのライフサイクルに関連する活動を展開している専門家に寄稿頂くことで、日本の産官学が何を学習し、実装しなければならないかのヒントを得ることである。

多摩大学の久保田貴文氏には、テキストマイニングを前提としたネット上からのデータ収集（WEB Scraping）の政策活用実践例を紹介いただいた。これまでとは異なったデータ取得技術や分析技術、可視化技術が可能になってきていることを実感できるのではないかと思う。

統計研究研修所の和田かず美氏には、公的統計の世界におけるMaintainプロセス、特にデータクレンジング等を支援するツールの最近の進展を紹介いただいている。様々な標準ツールが開発され共有されている。これは公的統計に限らず、取得データの質に問題がある場合には有用な技術となろう。

Data Robot社の伊地知晋平氏には、データサイエンスライフサイクル自体を非専門家対象に支援するツールの動向を紹介いただいている。これは、ProcessやAnalyzeだけでなく、データ自体の質に関わるMaintainをも含んでおり、今後、データサイエンティスト自体の社会的役割を変貌させる可能性もある。人材育成で後れをとっている日本の近未来戦略で参考にすべき情報と考える。

デロイトトーマツコンサルティング社の福井健一氏には、組織の価値実現に資するデータ群をどのように蓄積・マネジメントするかといったデータサイエンスを推進する組織が必要とする戦略を論じていただいた。まさに、Maintainステージに必要なデータアーキテクチャーに対する基本的方法を論じていただいた。ぜひ、産官のトップマネジメントに今後きちんと考えていただきたいテーマである。

統計家はこれまで、ProcessやAnalyze支援ツールは意識してきた。本特集では、データサイエンスのライフサイクルの比較的上流も含めて、今後必要となる考え方が示唆された。特集に寄稿頂いた方々に深甚の謝意を捧げる。