データサイエンス普及の隘路

（株）データサイエンスコンソーシアム，慶應義塾大学　　柴田 里程

1　ギャップ

データサイエンスという言葉だけが一人歩きしている面もあることは否めない現状で，データサイエンスの普及を議論するのは少々躊躇する面もあるが，統計学という古い殻を脱ぎ捨てて少しでも先へ進もうとする流れは大いに歓迎すべきであろう．もちろん，何でもかんでもデータサイエンスといってしまうのも危険である．すでに統計学が経験しているように，世の中での認識とのギャップが大きくなりすぎれば，一つのディシプリンとしての存在価値すら失われかねない．特に，データはいまやだれでもが目にし手にする身近な存在だけに，ほとんどの人がデータにわざわざ科学が必要とは思っていないこともギャップを大きくしている．ブームも去った今，データサイエンスといわれてもピンとこないのが普通の感覚であろう．特に「データにもとづく客観的な判断」がまだまだ未熟である日本の社会ではなおさらである．

2　データサイエンス

データサイエンスを定義するには，まずデータとは何かをはっきりさせておく必要がある．辞書によれば，データとは「数値，記号で表した推論の根拠となるもの」である．この定義のポイントは「推論」という目的が含まれている点で，これが無ければ単なる「情報」でしかない．その上で，データサイエンスは「データに関するなぜを追求するサイエンス」が定義となる．すると「データエンジニアリング」との違いは明らかである．サイエンスが「なぜ」を追求するのに対してエンジニアリングは「目標を達成するために有効な何らかの方法を開発し実装する」ことで，本質的に異なるパラダイムである．データサイエンスの定義をはっきりさせず，ただ漫然とデータエンジニアリングも含めてデータサイエンスと呼んでしまうと，データエンジニアリングはデータサイエンスを基盤として成立しているという関係性や役割分担が曖昧になってしまい，その進歩を妨げることになりはしないだろうか．

3　データサイエンスの実践

データサイエンス実践にあたっては，まず「データの総体的な理解」を基本に据えることが欠かせない．特段の推論という目的なしに単にサマリーを作ったり，逆に特定の目的にむかってデータをつまみ食いするだけなら，なにもデータサイエンスは必要ない．「データを的確にとらえ理解する」ことの助けとなるのがデータサイエンスであり，「どのようにしたらデータから新たな価値を発見できるか」その指針を与えるのがデータサイエンスであるからである．統計的推測論は確率論の枠組みに収まるようなデータだけを対象とし，記述統計学はもっぱらさまざまなサマリー手法の開発に専念してきた．しかし，D.R.Cox の ”Statistics survives but not necessarily by statistician” を持ち出すまでもなく，いまや看板の架け替えではすまない，時代の変化に合わせた自己変革を統計学者に迫っている．データの流れに広がる未開の原野を積極的に開拓することで，変量の概念やデータ型の再定義，正規化や欠損値，外れ値の扱いなど，これまでの統計学ではカバーしきれない概念から始まる数多くの課題が潜んでいることが明らかになる．「統計学は方法の学問」などと言ってこれまでの枠組みに閉じこもるのではなく，より広い世界にチャレンジする時なのではないだろうか，研究でも教育でも．

参考文献　柴田里程著『データ分析とデータサイエンス』, 2015, 近代科学社