ビッグデータとデータ中心科学

田村　義保　　モデリング研究系　教授

データ解析や情報処理の分野の周辺で、最近、非常に耳にする「言葉」として次のようなものを挙げることができる。

- ビッグデータ
- アナリティクス
- データ・サイエンティスト
- データ中心科学
- クラウドコンピューティング
- ・・・・最強の学問・・・・
- ・・・・sexy job・・・

世の中には「ビッグデータ」と呼ばれているデータがあふれかえっており、それを解析するための方法の総称がアナリティクスであり、解析の実施者がデータ・サイエンティストであり、解析のためのプラットフォームがクラウドコンピューティング環境である。ビッグデータが決してバズワードではなく、ビッグデータ解析に成功すればビジネス等で勝者になることができる可能性もあると考えている。上述した他の「言葉」についてふれておく。データ中心科学は情報・システム研究機構の北川源四郎機構長が提唱している科学の第4のパラダイムであり、情報・システム研究機構のホームページ（http://rc.rois.ac.jp/rc/message.html）に、次のような機構長からのメッセージが掲載されている。

> そして今、膨大なデータに何を語らせることができるのか、第4の科学ともいわれる「データ中心科学」の確立が急がれています。

第1から第3の科学は、それぞれ、実験科学、理論科学、計算科学としている。4つの科学を車の車輪に見立て、すべてが協働して科学が発展していくということを主張されている。データをもっと顧みるべきであるという意見には田村も賛成するが、ニュートン以前の科学も「データ中心科学」であったように思える。このことから、第4の科学ではなく、第0の科学と呼んでもよいのではないかと思っている。ヨハネス・ケプラーはティコ・ブラーエの惑星に関する観測結果から有名な三法則を発見している。ルネ・デカルトやエドモンド・ハレーは貿易風の可視化を行っているし、1800年代初頭に提唱された天気図は、天気予報を可能にしている。北川機構長が提唱する「データ中心科学」は、より大規模データを対象としており、また、解析手法もより洗練されたものを想定したものなので、似て非なるものかもしれない。しかし、ビッグデータ解析のツールとして、最も重要なものの一つである可視化の活用により、対象の有する真実を知ろうとする姿勢はニュートン以前からあったと言うことである。ニュートンにこだわっているのは、ニュートン以前と以後で科学は完全に変わったと考えていることによる。

「・・・・最強の学問・・・・」というのは、皆様も良くご存じの平成25年のベストセラーである「統計学が最強の学問である」（西内啓、ダイヤモンド社、2013）のタイトルの一部を抜き取ったものである。著者の西内啓氏は自らを統計家と称しているが、統計学が重要であることを社会に発信された功績は大きいと考える。ダイヤモンド社のホームページの内容紹介は下記の通りである。

> 統計学はどのような議論や理屈も関係なく、一定数のデータさえあれば最適な回答が出せる。そうした効能により旧来から自然科学で活用されてきたが、近年で はITの発達と結びつき、あらゆる学問、ビジネスへの影響力を強めている。こうした点から本書では統計学を「最強の学問」と位置付け、その魅力と可能性を 伝えていく。

ビッグデータの特徴を示す4V

・容量（Volume）　・種類（Variety）　・頻度・スピード（Velocity）　・正確さ（Veracity）

の中で、一般的には、Volumeが一番重要な概念であると考えられている。 最後のVeracityは、最近、付け加えられたものであり、それまでは、3Vであった。最初から生データが「正確」であった方がよいが、クレンジング（あるいはクリーニング）という手続きにより正確にした後、解析すべきであると考えている。

ビッグデータについて概説したつもりである。あまりにも、概論すぎると感じられているかもしれない。ビッグ、ビッグと騒がれているが、10年ほど前のデータマイニングとはどう違うの、従来の統計的データ解析とはどう違うのという疑問をお持ちの方は多いと思う。田村は、「大きさ」だけに引っ張られる必要は無いと考える。これまでの、ビッグデータの成功例として発表されているもののほとんどは、適当なサンプリングをして、一部だけで解析しても同じような結果になっていると思っている。しかし、ビッグデータの一部には、全数を解析しないと、対象の性質が分からないものもあると思う。このようなデータを解析するための手法は、まだ、見つかってないかもしれないが、計算機環境等は整いつつあるという意味で、ビッグデータ時代があるのだと考える。「まだ、見つかっていない解析法」を見つけるのが統計の研究者の役割であると考える。”big”という言葉を使わずに、“any”や”fast”という言葉が使われることもある。データを解析するのに数理的すぎる方法ではなく適切な統計的な手法でデータ解析する、データ中心科学的な思考が重要である。