日本統計学会誌
第30巻，第3号，2000年
327頁～331頁

データサイエンスのすすめ

柴 田 里 程*

Fascinating Data Science

Ritei Shibata*

本来は，パラダイムとは何かをはっきりさせておかないと，議論がかみあわないと思いますが，吉村［1］に従ってとりあえずおおざっぱに一つの学問分野ととらえておきます．

典型的にはその分野のパラダイムは教科書や学術雑誌がその内容を表していますし，学会も表しているわけです．このどれをとっても統計学の表すパラダイムはすでに古臭くなっていることは，すでにシンポジューム［1］の何人ものパネリストが指摘している通りです．したがって吉村［1］の問題提起である「パラダイムの変化は起こっているのか」に対する答は「イエス，とっくの昔に」です．ただしここでのパラダイムとは教科書とか学会というような表層的なパラダイムではなく，底流としてのパラダイムの変化です．

「変化が必要なのか不要なのか」という問題提起もあったわけですが，パラダイムの変化というのは我々が変化させるのかさせないのか，というレベルの問題ではないのですから，そのような議論は不毛です．クーンのいう科学革命とは「どのパラダイムも成熟期をむかえれば次のパラダイムにとって代わられる」ということです．スクラップアンドビルドが不可避的に進むということですから，古いパラダイムにしがみついてその流れを止めようとしても所詮無理な話です．それよりはなぜこのようなパラダイムの変化が起きたのか，また起きているのかをよく認識し，われわれの研究も教育も，さらには学会もさっさと次のパラダイムに乗り移る，つまり脱皮することを考えたほうがよっぽど生産的です．

少々挑戦的に言うならば，「統計学のパラダイムは消滅する」というのが小生の認識です．これは，もちろん統計学のパラダイムとはなんぞや，ということに大きく依存してくるわけですが，ここではいまの日本の教科書や学会のありように反映されている統計学，というパラダイムであると考えて下さい．

統計学の有用性は疑う余地がない
社会は統計学を必要としている
統計教育に力をいれて後継者を育てなければならない
統計学の需要は潜在的に大きい
統計学を必要とする新たな分野が生まれてきているではないか
統計学を知らないと困るだろう
統計学を知らない人の誤用が目に余る

* 慶應義塾大学理工学部　〒223-8522 神奈川県横浜市港北区日吉 3-14-1
E-mail : shibata@math.keio.ac.jp

などという自画自賛やおだてにのっているうちに，他の学問分野や実社会からは

統計学を知らなくたって別に困らない
結局，ゴミの部分に異常な興味を持つ学問分野で，結果にたいして違いがなくてもなんやかんやとうるさいことをいうだけじゃないか
何か新しい発見につながるの？
必要になったらちょっと勉強して適当なソフトウエアだけ拝借すればいい

とさげすまれ，どんどんその活躍の場を失っていっているというのが現状，特に日本の現状ではないでしょうか．統計学に「科」という一文字加えて「統計科学」とすればすこしは見直してもらえるのではないかというのは甘い期待に終わった気がします．科学というからには，なにが対象かをはっきりさせなければならなかったはずなのに，あたかも統計自身がその対象であるような曖昧なネーミングは，脱皮を遅らせただけだったような気がしますし，各実質科学への拡散を加速しただけであったような気もします．

このような現状に対して，すでに何人かのパネリストからかいろいろな改善の方策が示されました．統計学自身は消滅しても各分野ごとに発見を支援する科学として生き残ればいい．あるいは持株会社として生き残ればいいんだという話もあったのですが，小生はそう思いません．すでに底流では統計学の次のパラダイムが生まれつつあるからです．それを小生は名付けてデータを対象とする科学，データサイエンスと呼んでいます．

いうまでもなく統計学はデータを対象とする学問として出発しました．そして，数学とりわけ確率論を導入して近代化をはかりました．しかし自然の成り行きとしてファンシーな理論のほうが研究者の目を引き，本来データを対象とする学問で，しかもかなり泥臭い学問であるということは忘れられがちになり，かたや数学の一部門，かたやメソドロジーの単なる集積と分裂してしまっているのが現状ではないでしょうか．物理が物理現象を対象にし，化学が化学現象を対象にして一つのパラダイムを構成し世の中で広く認められているのとと同じように，世の中の情報化の動きは情報の一つの具体的な姿であるデータの科学というパラダイムの成立を後押ししてくれています．こう思って私は随分前からデータサイエンスを提唱してきたのですが，もちろんこのパラダイムが広く認められるためには数々の壁を乗り越える必要があります．

第一に，このパラダイムの参加者がみなデータのプロとしての意識を持つ必要があります．実験科学などではそれほど構造の入り組んだデータを扱わないことが多いわけですが，それでも点過程データを無理やり時系列データとして扱ったり，強引に結果に結び付けるような解析が一つのプロトコルとして定着してしまっているようなところもある．また，その他の科学では頭の中だけで考えた理論を裏付けるためだけにデータを利用しているようなこともある．ところがかなり入り組んだ構造のデータもうまく分解しモデル化して見せる，頭の中だけで考えた理論がそう都合よくはなりたっていないことをデータから実証して見せる．そういうことを積み重ねていってはじめてひろく認められるパラダイムになるんだと思います．データ解析の手伝いをしてくれる便利屋さんではないんだということを常に気をつける必要があると思います．これがデータのプロという意味でもあります．

データのプロというからには，経験を抽象化し蓄積し必要な理論を作り上げていく必要があります．そして臨床の医者のように我々はデータのプロなんだから結果については責任をもつとそこまでいかないと認めてもらえない．このような総体がデータサイエンスであると考えています．

ちょっと注意していただきたいのは情報とデータとの違いです．データは具体的なものですが，情報というのは便利な言葉ですが非常にあいまいで，たとえば情報科学という分野名が使われはじめてもう何十年になるわけですが，依然としてその内容はあいまいで，いまだに一つのディシプリンあるいはパラダイムとしては成立していないのが現状です．

それに対してデータというのは具体的な存在ですから，これを中心に据え対象とする科学というのはきちっとしています．もう一つ注意して頂きたいのは抽象化の大切さです．抽象化することによって違う分野の話でもアナロジーが通用するようになるわけで，データというあらゆる分野にまたがるものを対象とする以上，避けて通れないことです．この意味でもさきの持ち株会社方式はこのパラダイムには合いません．

それから名前の話ですが，データサイエンスといっても，統計学とどうせ似たことをやるのだからいままでのなじみのある統計学でいいじゃないか，という議論が当然起きてくるわけですが，新しい酒を古い革袋にいれてはならないという諺があります．データサイエンスというパラダイムには，データのランダムな部分だけにこだわらない，データのプロとして独自の視点で他の科学にも積極的に発言して行くんだという，これまでの統計学とはかなり質的に異なる部分があるわけですから，古い革袋に入れたらたちまち酸っぱくなってしまいます．

吉村［1］の次の質問である「統計学の次のパラダイムは何か」ということですが，小生が底流としてデータサイエンスが次のパラダイムであると考えていることは先に述べた通りです．しかしそれを表面化し一つの学問分野として確立するのはそんなにたやすいことではないと思います．新しい酒にふさわしい新しい革袋をつくらなければならないわけですから，統計学を統計科学と呼びかえたときのように簡単にはいきません．へたをすれば，せっかくの新しい底流をとらえ損なって，もとも子もなくなります．

まず他の科学との関係があります．さきほど述べたこととも関係しますが，他の科学とデータサイエンスとの関係をはっきりさせておかなければなりません．これをはっきりさせておかないと，すでに日常的にデータを扱っている諸科学からは，単なる便利やさんとみなされるか，下手すればじゃまと言われてしまいます．ですからそれに対して抽象化し蓄積したさまざまな分野にわたる経験を武器に，独自の視点で発言をしていくという立場をいつも鮮明にする必要があります．それぞれの分野にとらわれていたのではわからない，データの特性なり構造が，データサイエンスの目でみれば新たな視点が開けるんだ，ということを積極的に示していかなければならないということです．

つぎにインフラの確立ということがあります．もちろんこれまでの統計学が蓄積してきた財産を継承する必要はありますがそれだけでは不十分です．データとそれに対する技術を研究者みなで共有していく基盤が必要です．現在，小生が科学研究費の助成を受けて遂行しているプロジェクト「D&D の実用化」もこのような基盤整備の一つと思っています．

データの取得化からモデル化の段階まで一貫した流れとしてとらえ，それを一定のルールに則って記述し，それを計算機ネットワーク上に実現することができれば，共有の基盤は格段に充実するのではないでしょうか．

ところが最近，ある会合でデータを使うのは奴隷の仕事であると見なされている分野もあると言われたんですね．びっくりしたのですが，こういう意識が残っている限り，一方，理論は理論，研究している人が楽しければそれでいいという変に覚めた考え方が残っている限り，新しいパラダイムを受け入れる革袋にはなり得ないのだろうと思います．

それからもう一つは日本特有のことでコンサルティングにあまり重きをおかれていないという現状も変える必要があります．御存知のように日本以外の大学に行けば統計研究者の仕事の

うちコンサルティングが非常に大きなウエートを占めている。そのへん先程，刈屋さんがアメリカの大学は変わり身が早いと指摘されたことにも大きく関係していると思います。

しかし，アメリカのように極端に振れるのもどうかと思います。2年前に香港でスタンフォードのOlkin教授にあったとき，データサイエンスのパラダイムについて話したところ，その場ではあまり納得しなかったのですが，翌朝，小生のところにわざわざやってきて，とてもいい考えで全面的に賛同するといっていました。つまり今のアメリカの動きは研究資金などに動かされたものであって，けっして底流を見据えたものではなかったということが分かります。

教育というのが吉村[1]の最後の質問ですけれども，教育に限らず我々がエキサイティングに研究を続けていく上で，やっぱり「データっておもしろいな，不思議だな」と思えることが第一です。いろいろわからないことがわかってくるという意味だけでなく，わからないことがよけいわからなくなるという意味でもおもしろいということです。奥が深い，そう簡単ではないことは，最近の学生さんはあまり好きじゃないみたいですけれども，すぐわかってしまうようなことに対する興味は長続きしません。

よく教科書なんかで単純なデータで何とか統計学をわからせようと苦心している例がありますが，小生の経験からするとその効果は疑問です。教えている方がおもしろいのであって学生はそれほどおもしろいと思っていないことの方が多いのではないでしょうか。

小生が実際に授業で使っているのは例えば気象庁発表の最近10年間の日本での気象観測データすべてとか地震データすべてです。このくらいになれば彼らも膨大でとても直接眺めてもなにも摑めないということはすぐ分かります。その上で，データサイエンスのこれまでの蓄積に則って解析しモデル化すればいろんなことがわかることを示し，また計算機を使って実際にやらせてみれば，いかに有用な学問かということを実感させられます。

それと，学生には夢をあたえることも必要です。社会に出て大いに役立つんだ，具体的にいえば金儲けにもなるんだといったことも含んでいます。もちろん，データのプロとして尊敬されるんだということを感じさせることも必要です。これはわれわれ教育者自身がそう思うようにならなければ話は始まりません。

例えば私が引っ越したときに隣の人に「何がご専門なんですか」と聞かれていつも答えに困るのですね。「数学です」と答えていれば相手は安心する。下手に「統計」と答えようものなら，こちらが当惑させられるような反応がいくつも返ってくる。これは皆さんも経験されたことがあると思います。残念ながら統計というのは一般社会では，統計表をつくったり図をつくったりする退屈な仕事という認識が一般的ですから，それが学問としてどれだけの広がりをもっているのか分かるはずはありません。これはいままでの高校，大学でのおざなりな統計教育のつけといってもよいかも知れません。一方，理科系の人からは確率論の応用分野としてしか認識されていない，というきらいもあります。

こんなわけで小生は，最近では自分の専門を「データサイエンス」ということにしています。これだと，どこまで深く理解されているかは分かりませんが，すくなくとも誤解されるようなことはなくなりました。名前は本質的ではないという意見もありますが，すくなくとも教育に関しては「名は体をあらわす」と思ったほうがよいと確信しています。

最後に，もう少しデータサイエンスの具体的な姿をお見せして，小生の話を終えたいと思います。

ここ数年，統計学会でデータサイエンスのセッションをオーガナイズしてきましたが，柳川(堯)氏から「じゃあ，きちんとしたシリーズ本でも作ったらどうか，そしたら皆も少しはイメ

データサイエンス シリーズ

- データリテラシー
- データサンプリング
- データマイニング
- データモデリング
- データ学習アルゴリズム
- 空間データモデリング
- モデルヴァリデーション
- 地球環境データ
- 環境と健康データ
- 医学データ
- ファイナンスデータ
- スポーツデータ

ージが湧くだろう」と指摘され，昨年やっと重い腰をあげて企画した「データサイエンス」のシリーズが，いろいろな方の御協力を得て，2001年春より全12巻のシリーズ本として共立出版から発行される予定になりました．(別紙参照.)

このシリーズではデータの流れにそってデータサイエンスの基本を示そうとしています．一番最初はデータリテラシー，これは新しい言葉ですがデータに関する基本的な概念，それから標本調査と実験計画を統一的にとらえるデータサンプリング，データマイニング，データモデリング，データ学習アルゴリズム，空間データモデリング，モデルヴァリデーション．ここまでが一つの流れで，あと実際の応用分野として地球環境データ，医学データ，ファイナンスデータ，スポーツデータと続きます．このへんから私の意図しているデータサイエンスとはどんなものかを読みとっていただければ幸です．

補足：「データサイエンスというのは日本語で言えば資料科学なのに，なぜわざわざデータサイエンスといって，資料科学と言わないのか」という質問があります．しかし，英和辞典を引いたとき，「データ」に対する訳として「資料」という言葉はでてきません．事実，知識，情報，覚え書きといった訳が並んでいるだけです．データに対してはデータしかないんです．日本にはデータに相当する言葉がなかったんではないでしょうか．名前をつけるときは世の中に認められやすい名前をつけた方がいいわけで，データのあとは科学よりはサイエンスの方がいいだろう，しかも中黒があるのは古臭いということでデータサイエンスという名前を使っています．これは小生の専売特許じゃありませんからご自由にお使いください．

参 考 文 献

[1] 吉村 功(1999) 21世紀に向けての統計科学：シンポジュームへの問題提起，日本統計学会誌，**29**, 357-362.