ビッグデータの始まりと終焉

北海道大学 情報基盤センター
先端データ科学研究室　水田 正弘

それは、ある学生との会話からはじまった。

S 君「修論(修士論文)でビッグデータをやりたいんですけど・・・」

M 先生（自分でもよくわからないので、誤魔化すために）「ビッグデータって、ただデータが大きいだけでしょ。そんなんじゃ修論なんかになる訳ないじゃない」。

S 君（めげずに）「僕が就職したい NTT 関係の会社ではビッグデータできる人、欲しいって言っていました」

M 先生（就職が絡むと弱気になる）「じゃ、とにかく次のゼミでビッグデータを調べてきなさい」

統計学の専門家である M 先生、ビッグデータを「大きなデータ」と理解しており、それほど興味はわきませんでした。でも、しかたなく、図書館や Web でいろいろ調べてみました。数学セミナー10 月号に書かれている記事によると、ビッグデータの始まりは、2007 年ころで、2011 年 5 月にでたマッキンゼーリポートから本格的にはじまったことがわかりました。マッキンゼーリポートは英語で 100 ページを超えるもので、簡単にダウンロードできることもわかりました。「600 ドルで世界中の全ての音楽を記録できるハードディスクを購入できる」などびっくりすることが書いてありました。ついでにブラウジングしてみると、「これから 10 年間で最もセクシーな職業は統計家」と書いていました。統計学の専門家である M 先生は、その発言が 2009 年になされていることを素早く確認し、2018 年までの安心感を得ました。M 先生「なるほど、統計家は安泰だ」。ふと考える M 先生「でも、統計関係の学会でビック[1]データの研究発表は多くないな・・・」。

１．ビッグデータって何？

ビッグデータを大きなデータと理解してしまうと、本質を見失うことになります。情報技術の発展は目覚ましいものであり、だれでも、毎月、数千円出せば、インターネットは使い放題で、ついでに、スマホ（ご存知、スマートフォン）には GPS やカメラ、クレジットカードの機能が当然のようについており、ワンクリックで毎日の食事を写真にとって Web Page で公開します。ある人が何を食べようと、多くの人にとって全く価値のないつまらない情報です。しかし、それが、全国各地から数百万人の人のデータが集まると、話は変わってきます。一昔前でしたら、そんな「つまらない情報」でも、収集・保存するコストは莫大でした。しかし、今では、サラリーマンの小遣い(毎月平均 3 万 8457 円, 2013 年 6

[1] ビッ**グ**データを、ビッ**ク**データと言う人が多い…。ビックカメラの bic は、big を語源とする英語の方言らしい。

月新生銀行調べ)レベルでも、世界中の音楽の半分を保存できるハードディスクを購入できます。

データの収集や保存方法が安価になることで、**「従来のアプローチでは対応できないデータ」**が大量発生するようになりました。実は、これがマッキンゼーリポートにおけるビッグデータの定義です。すなわち、「＊TB 以上のデータをビッグデータと呼ぶ」とは定義していません。従来のアプローチでは対応できないかどうかは、データの容量だけではなく、利用目的やコンピュータ環境に依存します。例えば、ノートパソコンで作業をしている大学生と、アメリカ合衆国国勢調査局で働く統計家では、あるデータがビッグデータであるかどうかが変わることになります。

ビッグデータの本には、ビッグデータの特徴を 3 つの V で表すと書かれていることが多いです。すなわち、(1) Volume (容量)、(2) Velocity (更新頻度)、(3) Variety (多様性)です。これらは、「従来のアプローチでは対応できないデータ」をビッグデータの定義とする場合、対応できなくなる要因と考えると分かりやすいです。一般に、Volume (容量)が大きくなると対応が困難になるのは自然なことです。しかし、単に数値データの算術平均を求めるだけであれば、その数値データのサイズはほとんど問題になりません。そのデータがどの程度、頻繁に発生または変化するか (更新頻度)、複雑なデータかどうか (多様性)によって対応できるデータのサイズが大きくかわります。さらに、実用上、大きな観点は、リアルタイムで結果を要求しているか、それとも、時間的余裕があるかです。3 つの V をビッグデータの定義とする記事もありますが、これは、ビッグデータの特徴と理解する方が正確だと思います。

２．ビッグデータにお金の匂いが・・・

ビッグデータを使うと金集めができることの好例は、2012 年 11 月のアメリカ合衆国選挙です。再選を目指すオバマ大統領は、約 200 名の専門チームを使い、ネット選挙に対応しました。特に、注目すべき点は、「ビッグデータ」を解析して、オバマを支持する 40、50 代の女性の多くが人気俳優ジョージ・クルーニーのファンであることを見つけ出し、彼を囲む会費 4 万ドルのパーティを開催するとともに、1 口 3 ドルのネット献金をした人から抽選で数名を無料招待しました。これにより、イベント全体の収入は 1500 万ドルに達しました。オバマ陣営がネット選挙対策に投入した金額が 1000 万ドルといわれていますので、それ以上の収入になった計算です。

その選挙の少し前、2012 年 3 月 29 日に、オバマ大統領はビッグデータプロジェクトに 2 億ドルを投資すると発表しました。また、我が国においても、2012 年 7 月 30 日に「日本再生戦略」においてビッグデータで約 10 兆円規模の関連市場の創出を目標に掲げました。それに前後して、情報関連の大手企業の多くは、ビッグデータに対応できる組織改編をしました。例えば、NTT データは、2012 年 2 月 15 日に数理システムというデータ解析のプロ集団の会社を子会社化して、2016 年度末までの累積で新たに 100 億円規模のビジネス創

出を目指すと発表しました(その後、目標は2015年度までに200億円になったようです…)。
ビッグデータを活用することで、売り上げを急激にアップできた例が多数、公表されています。今回の講演でもいくつか紹介いたします。

３．統計学、データサイエンス、データアナリスト

ビッグデータが利用可能になった背景には、ネットワーク、記憶装置、センサーなどをはじめとする情報技術の発展があります。つまり、クラウドコンピュータや Hadoop と呼ばれるソフトウェアなどの利用、それらの改良により、すばらしい情報環境が提供されるようになりました。次の課題は、刻々と収集される大量なデータをどのように活用するかです。

データの可視化や、比較的単純な集計により、有効な戦略が見つかることもないとは言えません。しかし、多くの場合、データの山に飲み込まれてしまいます。また、ある戦略が見つかったとしても、より良い戦略の存在を見逃す可能性が高いと言えます。多くの可能性を検証することが肝要です。そのために、データに関する知識、情報技術、そして統計学を活用しましょう。統計学には、コンピュータが発明される前からデータと対峙してきた歴史があります。単に歴史があるのではなく、常に新しい道具を取り込んできました。19 世紀には確率論、20 世紀に入ってはコンピュータを使いこなし、そして現在は情報環境全体を道具としています。また、統計学を含むデータサイエンスへの展開を目指す専門家もいます。

ビッグデータを扱うデータアナリストは、上記の「データに関する知識」、「情報技術」、「統計学」をうまく使いこなすこと、さらには、コミュニケーション能力が重要になります。これらの全ての能力を有する人材は非常に少ないのが現状ですが、社会的ニーズはあまりにも大きいと言えます。

４．Death of Big Data

ビジネスの世界でも、研究の世界でもビッグデータを活用することが成功の鍵です。とにかく可能な限り大量なデータを集めて、解析ツールに入力すればいいのです。もはや「ビッグデータ」を無視してビジネスを進めることは無謀な時代となりました…と締めくくりたい誘惑にかられます。でも、本当にデータは多いほどよいのでしょうか？すべてのデータを使うのが得策でしょうか？

少し昔の話になりますが、人工知能に関する優秀な研究者と議論したことを思い出します。データは全て使うべきであると先生は主張されていました。数式を使って説明しても前提条件などが空回りして議論が進みませんでした。今なら、歴史的に有名な例を紹介するのにと反省しております。1936 年のアメリカ大統領選挙の予測で、「リテラリー・ダイジェスト」（The Literary Digest）という総合週刊誌は、200 万人以上を対象から回収した調査結果を基に共和党のランドン候補が 57%の得票を得て当選することを予想しました。そ

れに対して、前年に世論調査の業界に参入したばかりのジョージ・ギャラップが率いる「アメリカ世論研究所」（the American Institute of Public Opinion）は、わずか 3000 人という少ない対象者からの回答を基にルーズベルト候補が 54%の得票を得て当選することを予想しました。ご存知の通り、ルーズベルトが 60.2%を得票して当選しました。これは、単にデータ数が大きければよいとの主張に対する反例となります。データの質がデータの数を凌駕しました。

バズワード(はやり言葉)の最終段階ではよくあるパターンですが、Death of Big Data と主張している人もでてきました[2]。Big Data が大切なのではなく、Any Data(すべてのデータ)が大切であるとの主張です。同意したいと思います。ビッグデータのブームで分かったことは、「従来のアプローチでは対応できないデータ」にオドオドするのではなく、データ主導型で真摯にデータを収集・保存・解析することが大切であることです。

言うまでもなくビッグデータの活用は重要です。しかし、本当に重要なのは、データの活用であり、「ビッグ」の部分ではありません。情報科学や経営学の素養を有する統計学の専門家、または、統計学の素養を有する情報科学や経営学の専門家、広くはデータを重視する専門家が大活躍する時代になったと言えます。

S 君は、修論としてビッグデータ関係の研究を完成させました。また、無事に NTT グループの会社に就職できたようです。でも、修論の題目には「ビッグデータ」との言葉が入っていません。１０年後に、「私の修士論文はビッグデータです」と言ったら爆笑されるのが心配だったようです。彼の判断は正しかったでしょうか？

参考文献

水田正弘・南 弘征(2013) ビッグデータと統計解析, 2013年度統計関連学会連合大会チュートリアルセミナー

水田正弘(2013) ビッグデータとは何か, 数学セミナー10月号、23-27ページ

John M. Chambers (著) 垂水共之、水田正弘、山本義郎、越智義道、森 裕一(翻訳), データによるプログラミング ーデータ解析言語Sによる新しいプログラミング, 森北出版, 2002 (Programming with Data).

[2] http://www.forbes.com/sites/ciocentral/2012/10/04/the-death-of-big-data/