Internet world
ネット時代に生きる

櫻井 哲朗

第16回
データサイエンティスト

データ分析力を駆使して「価値ある情報」生み出す

寒さも一段落してきたかなと思えるようになってきた今日この頃みなさんいかがお過ごしですか。もう今年も1年の4分の1が終わろうとしています。残り4分の3、元旦にたてた「今年こそは真面目に生きる」という今年の目標も最早遥か忘却のかなたに追いやってしまいました。

心理学者のヘルマン・エビングハウスが提唱した忘却曲線もビックリのど忘れっぷりです。ちなみに、ヘルマン・エビングハウスの忘却曲線とは、子音・母音・子音からなる無意味な音節を記憶し、その再生率を調べたものです。その再生率は図1のような曲線を描いて、どんどん低下していきます。

この図の横軸は完全に憶えた時点を原点として時刻（分）に対数をとったもので、再生率は20分後には58％つまり半分に、1日後には26％つまり4分の1に低下していきます。つまり人は忘れる生きものであるということを数値として詳しく表しています。しかし、ここでの記憶は学問のような体系だった知識などとは異なる無意味な音節の記憶の結果であるため、全ての記憶がこの通りに忘れていかれるわけではありません。

「きのこ」or「たけのこ」

前回、バレンタインの話に触れましたが、みなさん、結果はいかがでしたでしょうか。チョコレートといえば、みなさんは「きのこ派」ですか、それとも「たけのこ派」ですか。そうです、かの有名な「きのこ・たけのこ戦争」のことです……と突然言われても何のことだか分からないかと思います。明治ミルクチョコレートでお馴染みの「きのこの山」、「たけのこの里」という2つのお菓子のことです。このどちらが好みかということで長らく論争が続いていました。

ある人はチョコレート部分とビスケット部分を分離した「きのこの山」こそ究極の形という一方、ある人はチョコレート部分とビスケット部分をブレンドした「たけのこの里」こそまさに至高の形と両者一方もゆずらない血で血を洗う？抗争の歴史がありました。その長らく続く問題に1つの決着がつきました。

「レシレコ」の軍配は

それは、ITmediaというIT系のニュースを集めたサイトで取り上げられた1つの記事でした。詳しくは記事[1]を見ていただきたいのですが、簡単に書かせていただ

きますとブレインパッドから提供されているスマートフォン向けアプリ「ReceReco（レシレコ）」という家計簿アプリからの２０１３年１月～11月末までのデータを集計したところ、「たけのこの里」に軍配があがったそうです。

また前回紹介したGoogleトレンドを使い、ほぼ同期間の「きのこの山」、「たけのこの里」の検索数を調べてみました。その結果が図2になります。これからは「きのこの山」のほうが検索数が多いという逆の結果となりました。もちろん、実際の購買と検索数という違いがありますので、これらの結果を同列に扱うことはできません。

ここで注目していただきたいのは「たけのこの里」が「きのこの山」よりも多く買われているという点ではなく、これらの分析を専門の仕事として請け負っているデータサイエンティストという職業があることです。

ちなみに、ここで登場してきた「ReceReco（レシレコ）」というアプリ、これはレシートをカメラで撮影することで画像データからコンピュータで扱えるデータに自動的に変換してくれるアプリですが、これを提供している会社はソフトウェア系の会社ではなくブレインパッドと呼ばれるデータマイニングを中心としたビジネスを展開している会社です。最近、データを扱う業種が注目の的となっています。そこで、今回はデータサイエンティストという新しい職業に焦点を当てていきたいと思います。

最もセクシーな新職業
ビッグデータは原石

この連載で紹介してきた内容の中にSNS（ソーシャル・ネットワーキング・サービス）、クラウド、ビッグデータがあります。これらは個別に成り立っているわけではなく、相互に関係しながら形作られています。実際、SNSの投稿やクラウドコンピューティングを使ったクラウドサービスなどによって日々大量のデータが生み出され、そのような高頻度で大量かつ多様なデータをビッグデータと呼んでいます。ビッグデータを扱う上でコンピュータは欠かせません。

そのため、これらのデータは電子化され、言い換えればビッグデータは全て数値データに変換されます。たしかにビッグデータの中には画像データやテキストデータなどが含まれていますが、それらは全て原始的には0と1の数列に分解されますのでコンピュータで扱うデータは数値データと見なすことができます。

この膨大な数字のデータをそのまま持っていたのでは意味が、そして価値がありません。月並みな言い方ではありますが、つまりビッグデータとはダイヤモンドの原石のようなものなのです。ここでダイヤモンドの原石と断定せずに、わざわざ〝ような〟と含みをおいたのはビッグデータの中には磨いても特に有用な結果を見つけられないデータである場合もあるため、

図1

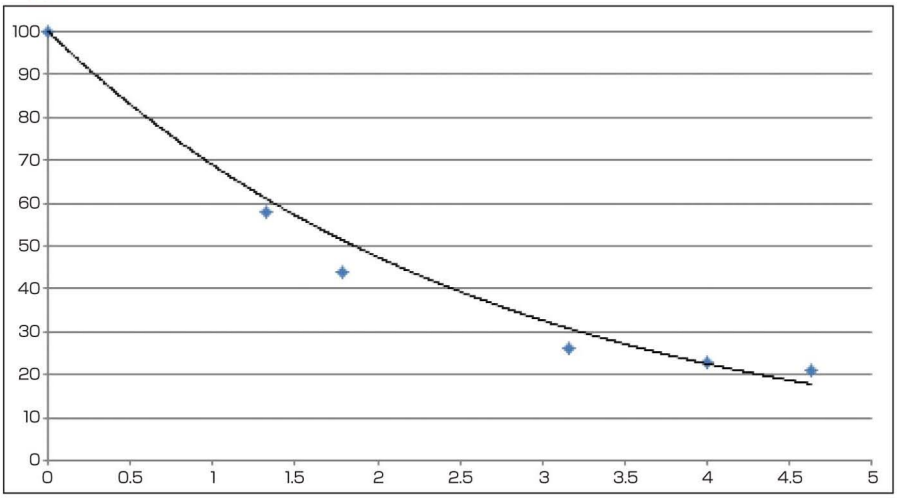
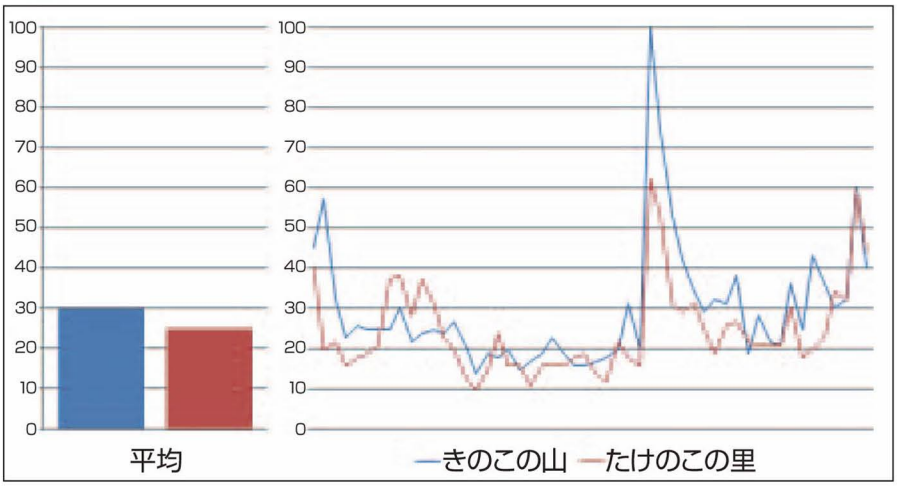


図2

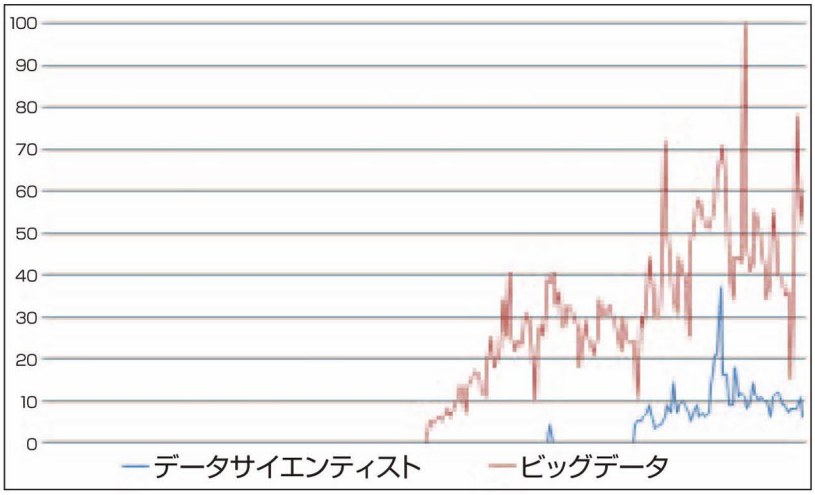


このような書き方をさせていただきました。

データを磨く研ぎ師

ダイヤモンドの原石のようなものを磨く研ぎ師がデータサイエンティストという職業になります。データはそのままでは価値はなく、データは分析してこそ意味があります。そこに、どのような価値を見い出すかはデータサイエンティストの分析力にかかっています。

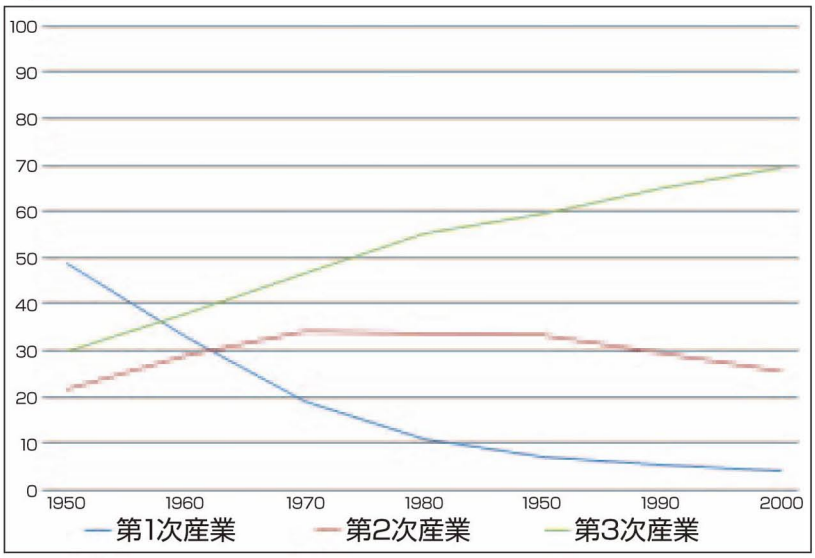
図3



このようにビッグデータを分析するデータサイエンティストが各種業界の注目を集めています。

世界的なコンサルティング企業であるマッキンゼー・アンド・カンパニー(McKinsey & Company)が発行している雑誌マッキンゼー・クオータリー(McKinsey Quarterly)の2009年の記事[2]の中で、Googleのチーフ・エコノミストであるHal Varian氏が次のように述べています。

図4



"I keep saying the sexy job in the next ten years will be statisticians."

「次の10年で魅力的(セクシー)な仕事は統計学者であるだろう」

昨年から注目度増す

つまり、今後はデータ分析の仕事にスポットライトが当たると言いたかったのだと思います。また実際、「データサイエンティスト」という言葉に関してGoogleでの検索数の推移を見てみますと図3のようになっています。比較対象として「ビッグデータ」という単語もあわせて見てみましょう。これより、先に「ビッグデータ」という言葉が検索されていることがわかります。

だいたい2012年以降から各企業がビッグデータの分析に着手したことなどを新聞などのメディアを通して発表しはじめました。それに続くように2013年つまり昨年頃から「データサイエンティスト」という言葉が検索されはじめ徐々に世の中に広まってきています。また昨年の7月にはNHKで放送中のクローズアップ現代でも特集されています。その詳しい内容などは参考文献[3]のHPで見ることができます。このように、データサイエンティストという職業に注目が集まってきています。

職業人口の観点から情報関係者の細分化

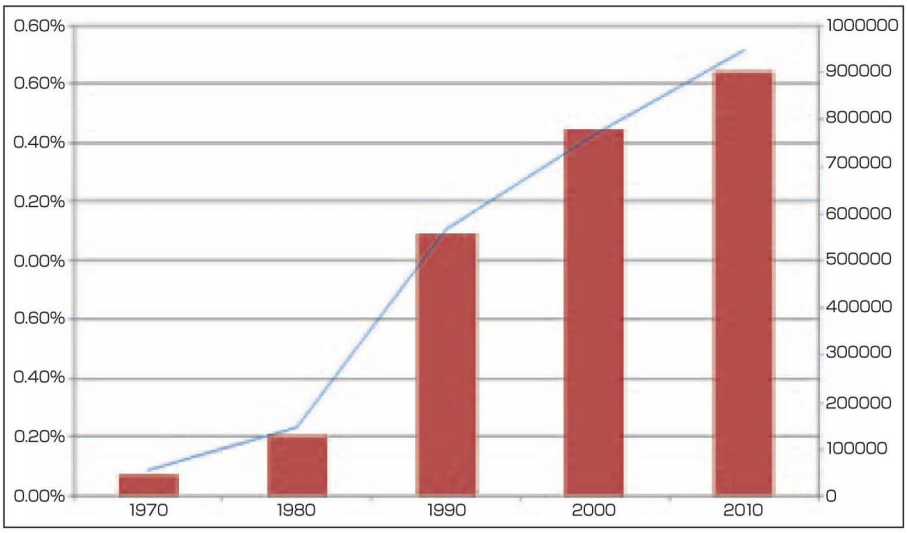
では次に総務省統計局で公開されている統計データを使って日本における職業人口の変遷を通してデータサイエンティストという職業に迫ってみたいと思います。

まずは図4をご覧下さい。こちらは職業を三部門、第1次産業、

さくらいてつろう

中央大学大学院理工学研究科を卒業し、専攻は統計学。コンピュータなどによって計測される大量のデータをまとめる多変量解析の研究。現在は、諏訪東京理科大学共通教育センター講師。東京都出身、30歳。

図5



第2次産業、第3次産業に大別した場合の各種就業人口の比率の推移となっております。大雑把に分類しますと第1次産業は農業など、第2次産業は生産業、第3次産業はサービス業です。

２０００年から職業の分類項目が変更されていますので完全な比較というわけではありませんが、第3次産業の増加が顕著であることがわかります。これは、みなさんが小学校や中学校の社会で習った事柄と一致するかと思います。この第3次産業の中にコンピュータ関連の職業が多く含まれています。

また職業小分類別の情報処理技術者に注目すると図5のような推移となります。図5では、折れ線が比率を表し左側の軸が目盛りとなります。また縦棒が就業人口の実数を表しており右側の軸が目盛りとなります。

先ほどもふれましたが２００年から職業の分類項目が変更されており、２０００年以前は情報処理技術者となっていた項目が２００５年にはシステムエンジニア、プログラマーに細分化され、さらに２０１０年にはシステムコンサルタント・設計者、ソフトウェア作成者、その他の情報処理・通信技術者といった形に分類されるようになりました。

そこで図では、これらを合計した値を使って集計しております。この図からも見て取れるように情報処理技術者は１９９０年を境に急激に増加していることがわかります。実際、１９８０年代以降においてインターネット誕生や世界や日本においてもビジネス向けのパソコンが発売され、オフィスなどに普及していき家庭などにはワープロが普及していきました。このような背景によりオフィスのオートメーション化の発展とともに情報処理技術者の需要が拡大し、１９９０年の就業人口の急増にいたったのだと考えられます。

社会構造の大変化か

２０００年頃になるとＩＴ革命または情報革命という言葉と共にさらに需要と職種が拡大していきます。情報革命と聞くと少し大げさな言い方かと思われる方もいらっしゃるかもしれません。人類の技術の大きな発展としてよく取り上げられるものとして18世紀の産業革命があり、それ以前には農業革命が起こったとされています。それぞれ、社会構造に変化を及ぼすような大きな変革でした。

では、インターネットの誕生やコンピュータの普及による発展は我々の社会構造を大きく変えるような変化でしたでしょうか。これらの情報通信技術の発展により便利になり、多くの分野でこれらの技術が使われています。たしかに我々のなかでは社会構造まで変化は捉えることができていませんが、それは我々が流れの中にいるので緩やかな変化に気づけていないだけなのかもしれません。これは情報革命がまだ進行中であり、現在次の段階に到っていると考えることができます。

２０００年代の変化をインターネット環境やパソコンの普及といったインフラ整備として捉えると、これからの変化はしっかりした土台の上に大きな家を建てるようなものとして見ることができるかもしれません。後に、この時代を振り返ったときに情報革命となった大きなターニングポイントになっているかもしれません。そういった意味で我々は時代の節目を目撃した生き証人だともいえるのかもしれませんね。

仕事はどう展開

多角的なアプローチ

次にデータサイエンティストという職業の中身に注目してみたい

と思います。データ分析を使ったシステムの1つとしてネットショッピングでお馴染みになったオススメシステムがあります。例えば、amazonでは買い物をすると、「この商品を買った人は以下の商品を買っています」という形で他の商品を推薦してきます。これはレコメンダシステムとも呼ばれていて、利用者の好みを分析することで利用者にあった商品をオススメするシステムになっています。

　このとき、好みを分析するのが焦点となってきます。アンケートを行うことで利用者の好みや買い物をした商品を分析したり、買い物をしなくてもクリックした商品を分析することで利用者の好みを判定していきます。そのため、システム的には自分と他人ではオススメされるものが違ってきます。

　マニアックなものを買うとそちら側の商品をオススメされる機会が増えてきたりもします。このように好みの分析がすむと次に必要になってくるのが、どのようなものをオススメ商品とするかということです。このとき、よく使われる手法の1つとして他の利用者との違いを測ることにより近い利用者の買い物状況によってオススメ商品を決定するという方法があります。

　このようなシステムによって、例えば好きなマンガの作者が描いている違う作品を見つけることができたり、好きな映画の俳優や監督が参加している映像作品を知ることができたりします。つまり利用者は自分の知ることのできなかったかもしれない商品を知ることができ、また販売する企業としては新たなビジネスチャンスの機会を得ることができます。

「経験」をシステム化

　書籍「会社を変える分析の力」[4]で紹介されている大阪ガスの取り組みも興味深いです。著者の河本薫さんは同社の情報通信部ビジネスアナリシスセンターの所長であり実際にデータ分析をされた方でもあります。同センターの取り組みとして給湯器の修理の効率化があります。これは過去の修理データを分析することで故障原因の予測、さらにそこから持って行く修理部品を予測するシステムを構築していきました。

　同書では、いままで現場で培ってきた勘（K）と経験（K）と度胸（D）のKKDを否定するのではなく、それを裏付けるような分析をすることを提案しています。たしかに、いわゆるベテランと呼ばれる人たちは長年の経験とある種のひらめきのような勘により意志決定をしています。これらの判断をシステム化することによって誰でもベテランの経験と勘を使えるようになります。

　同センターの分析はこれだけにはとどまらず、より業務を効率化するためにはどうするかに注目し、当直シフトの自動化や修理体制の見直しなども行っています。

目指すのは「最適化」

　DeNAのMobageやグリー株式会社のGREEなどのSNSで展開されているソーシャルゲームでもデータ分析を駆使することによって、より使いやすいシステムに修正・改善を繰り返しています。実際、ゲームのユーザーインタフェースつまりゲームの操作画面を変更した際に利用者からの不満や参加率の低下などにより前の操作

画面に戻すといった対応をとる場合もあります。これらも利用者からの報告や操作などをデータ分析することによって見えてきます。

これ以外にも多くの分析事例がありますが、データサイエンティストがデータ分析を通して目指すものの1つの目標として最適化があります。最適化の形には色々ありますが、顧客の情報推薦システムの最適化により新たなビジネスチャンスの機会を得たり、業務の効率化という最適化を行うことでよりスピーディーな業務を展開できたり、顧客情報から意見を吸い上げることでより満足度の高い最適化されたシステムを構築することができたりします。

どんな能力が必要か「技術」と「数理的知識」

では先ほどのような最適化を実現するためには、どのような能力が必要なのか考えてみたいと思います。筆者が思うデータサイエンティストに求められる能力は図6にあるような「数理的知識」、「ITの技術」、「現場の知識」の3つだと考えています。

なぜ必要なのか説明したいと思います。まず「数理的知識」は分析する際に客観的な結果を提示するために必要となる能力です。例えば、データを分析するのに統計学などを使ったり、予測を行うために機械学習やシステムの最適化のために数理計画法などを使ったりします。

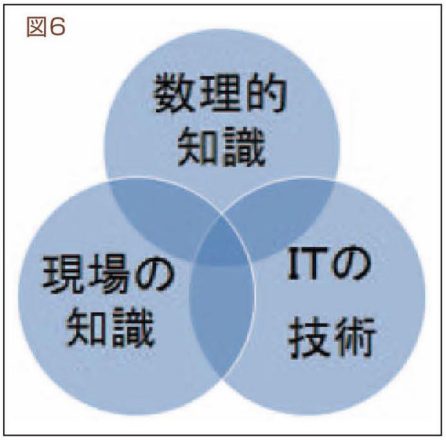


図6

「ITの技術」は分析の入口と出口に必要となってきます。データ分析をするためには、そもそもデータが必要となってきます。このとき、一般的にデータは多いほど精密な分析をすることができます。そのため効率的なデータ取得を行う必要があり、このときに必要となるのがネットを駆使したデータ取得だったり独自端末を使ったものだったりします。

分析した結果を提示するためにも、個々の結果に対応するため、また即時性を保つためにもコンピュータによる自動化が必要となってきます。

欠かせぬ「現場の知識」

最後にあげた「現場の知識」について説明したいと思います。データを扱う際にそれらの数字がどのような意味を持っているのかを知らなければ分析した結果に意味を見い出すことはできません。そのためにも「現場の知識」を知っておく必要があります。せっかく分析したのに当たり前のことだったというような結果にならないために必要な知識だと筆者は考えています。

このようにデータ活用の重要性を受けて、2005年頃に日本統計学会の統計教育委員会から文科省に対して統計教育推進の提言が行われました。また最近ではコンピュータリテラシーというコンピュータの操作などの能力を持っていることを指す言葉から派生してデータリテラシーというデータを正しく読み取る能力を持っていることを指す言葉なども出てきています。

2014年のセンター試験の数学Ⅱ・Bの試験における選択問題では統計に関する問題も出題されています。このように、現在ではデータを分析するということは一部の人しか行っていないですが、コンピュータを使ってメールを書いたりホームページを見たりするように一般の人たちも簡単な分析ならできるような時代がくるかもしれません。

参考文献

[1]http://www.itmedia.co.jp/news/articles/1402/04/news016.html

[2]http://www.mckinsey.com/insights/innovation/hal_varian_on_how_the_web_challenges_managers

[3]http://www.nhk.or.jp/gendai/kiroku/detail02_3375_all.html

[4] 河本薫著「会社を変える分析の力」講談社、2013