ビッグデータとそれに基づくアナリティクスは、この先の十年で、ほとんどすべての産業とビジネスを変えるだろう。ビッグデータを早期に活用し始めた企業や、そこに属するすべての人たちは、大きな競争力をすでに得ていると言える。ちょうどスモールデータ時代の初期に分析力で勝負した人々が競争相手をぐっと引き離したように、会社や企業にとって、いまがビッグデータで勝負するチャンスなのだ。

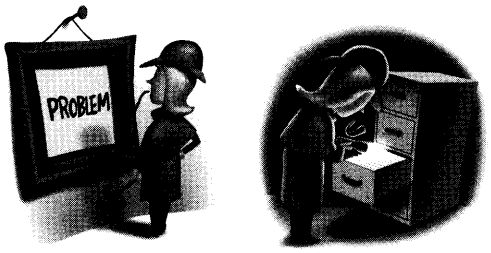
ビッグデータは、ユビキタスコンピューティングやデータ計測収集技術の発展に伴い、ますますそのカを発揮するだろう。間もなく、センサーとマイクロプロセッサーがいたる所に設置されるようになるだろう。ほとんどあらゆる機械や電子機器に、人やモノの動作、位置、あるいは状態を示すデータの痕跡が残される。こうした装置とそのユーザーは、インターネットを介して交信し合い、それがさらなる膨大なデータの源となる。このすべてが、他のメディア、例えば無線あるいは有線電話、ケーブル、衛星などからのデータと合わされば、データの未来は、さらに広がっていくだろう。

これらすべてのデータが利用できるようになると、**企業や組織のほとんどすべての業務が、ビッグデータと密接な関係をもつようになる**。製造業では、すでにほとんどの機械がマイクロプロセッサーを１個または複数搭載するようになっていて、急速にビッグデータ環境に移行しつつある。無数の顧客のタッチポイントやクリックストリームといったデータが存在する消費者マーケテイングの分野も、ビッグデータの問題に直面している。グーグルは自社の無人自動車を「ビッグデータ・プロジェクト」と呼んでさえいる（左記「ビッグデータとは何か？」を参照)。

図表１ 定量分析の３つの段階と６つのステップ

第１段階　問題のフレーミング

1 問題認識　　2 過去の知見のレビュー

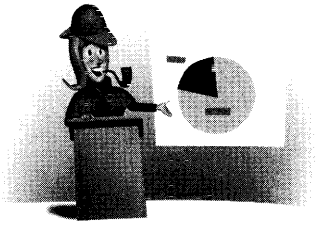


第２段階　問題の解決

3 モデル化　　4 データ収集　　5 データ分析

第３段階　結果の説明と実行

6 結果の説明と実行



何を決めたいのかを明らかにする

例えば、アナリティクスサービス会社ファースト・アナリティクスのマイク・トンプソンは、問題認識の段階で顧客と行なった会議を振り返った。クライアントは外食チェーン店を運営していて、**分析の主な焦点は商品（食事メニュー）の収益性にあると思い込んでいた**。顧客の幹部は、ファースト・アナリティクスに、それぞれのメニューがどれくらい利益を上げているかを調べてほしいと依頼した。

そこで、トンプソンは何を決めたいのかに焦点をあて、「では、その収益性分析の**結果が得られたとして、みなさんはどのような決定を下すつもりですか**」と顧客に尋ねた。長い沈黙があった。一人の幹部が、「主に、どの料理をメニューに入れ続けるべきかを決めることではないかな」と言った。すると、別の幹部が「うちのチェーンでは過去 20 年間に、外したメニューは一つもありませんよ」と指摘した。さらに議論を重ねた後、顧客のチームは分析の焦点は収益性よりむしろメニューの価格設定になるだろうという結論に達した。「設立以来、うちの会社では価格をいろいろと変えてきたのでね」と幹部社員の一人が述べた。

例えば、マイクロソフトは**ビング検索**エンジンを使わせるために、精力的に「オファー」を出してくる（ビングは無料で使用できる。マイクロソフトはあなたにビングを使わせたがっているのだ）。オファーにつられてあなたはビングを使ってみる。そして、ビングの検索バーをブラウザーに表示して、ビングの特別な機能を試めすようになる。オファーは、あなたのクッキーなどのソースから割り出したさまざまな要素（住所、年齢、性別、最近のネットの使い方）に基づいて**カスタマイズされている**。

あなたがマイクロソフト・パスポートに申し込みをすれば、マイクロソフトはあなたに関するさらに詳しい情報を手に入れ、もっと効果的に**あなたというターゲットに絞ったオファーを提供できるようになる**。マイクロソフトは、インフォ・エビファニー・インタラクション・アドバイザーというソフトウェアを利用し、あなたがオファーをクリックした瞬間に、あなたの受信トレイにターゲットメールを送れる。この間ほんの 200 ミリ秒ほどしかかからない。**マイクロソフトは、この仕組みは成約率を上げるのに非常に効果的だと述べている**。

グーグルでは、従業員のどのような特徴が、その人の仕事ぶりの予測因子になるかを知りたいと考えた。大学の成績や面接の出来といった、**当初グーグルが重要と考えていた要素は、仕事ぶりの予測因子としては弱いことがわかった**。どんな要素が重要かわからなかったので、グーグル上層部は従業員を対象に 300 問のアンケート調査を行った。グーグルの人事担当部長のラズロ・ボックはこう述べている。「私たちは、非常に大きな網を投げたかったんです。ここの廊下を歩いていると、よく犬に出くわすことがあります。もしかすると、犬を飼っている人は会社にとって有益なパーソナリティ特性を持っているのかもしれない」(7)

職場にペットを連れて来ることは予測因子とならなかったが、グーグルはいくつか思いがけない発見をした。例えば、**求職者が何らかの世界記録または国内記録を樹立していたり、非営利団体かクラブを創設したりしていた場合、優れた業績を挙げていた**。グーグルは現在 0、オンライン就職面接でこうした経験の有無を聞くことにしている。

問題認識ステップの初期には、幅広く考えることが重要だが、このステップの終わりまでには、**問題を明確な文章にできるようにしておきたい**。そして、その文章には分析したい要因や変数の具体的な定義を含めること。その理由は、**定量的に物事を考える場合は、物事がどう定義されるかによって大きな違いが生まれるから**だ。例えば、あなたはテレビ局の幹部で、視聴者がどのチャンネルを見ているかを知りたがっているとしよう。二人の分析コンサルタントがそれに対する提案を持ってくる。あなたは興味半分で両者の結果を比較するために二人とも雇うことにする。

一人目のコンサルタントは、1 週間毎日、実際に見たチャンネルと番組を視聴者に（オンラインか紙で）記録してもらうという方法を提案する。二人目は、調査に応じてくれた人々に過去数か月間によく見たチャンネルを思い出してランク付けしてもらうという案を出していた。どちらも母集団を代表する調査のサンプル集団をうまく設定した。

この二人のコンサルタントは、同じ問題を解決しようとしているが、彼らの出す結果は大きく異なるものになりそうだ。視聴者に実際に見たチャンネルと番組を記録してもらう提案をしたコンサルタントは、より正確な結果を得る可能性が高いが、記録がめんどうなためサンプル集団の調査への参加率は低くなるだろう　（ニールセン・メディア・リサーチは継続的にチャンネルと番組をモニターしているが、離脱率は 50 パーセントである）。このコンサルタントのもう一つの問題は、視聴パターンは季節や特定の週の番組編成によって大きな影響を受けるかもしれないことである。

二人目のコンサルタントの調査はそれほど正確とはいえないが、より広い期間をカバーしているため、季節による影響を受けにくい。重要なのは、**この二つの調査結果はたぶん、すり合わせるのが難しいほど違ったものになる**だろうということだ。だからこそ、この問題認識ステップにおいては、何を調べたいかを明確にすることが大事なのである。

過去の知見をレビューする

問題が認識されたら、次は、その問題に関連する過去のすべての知見を調べるステップだ。このステップは、まだ、分析の第 1 段階（問題のフレーミング）の中にある。先人たちの取り組みを調べれば、分析者も意思決定者も、これまでにこの種の問題はどのように構造化されてきたかを知り、それを別の視点から概念化する方法についても知ることができるからだ。過去の知見をレビューすると、分析者は、問題認識ステップの大幅な修正につながるような**気づきに出合うことが多い**。

4 年、ナチスのドイツ空軍はロンドンの市民に恐怖を与え始めた。その後数か月にわたって、3172 発の V－2 ロケットが連合国の標的に向けて発射されたが、そのうちの 1358 発が上空からロンドンを襲い、およそ 7250 人の軍人と民間人を死に至らしめたのである。

ロンドンへの大空襲のさなか、多くの対空監視員が、爆弾の着弾点はいくつかの場所に集中していると断言した。そこで、**イギリス政府はドイツがロケットの標的を定めることができるのか、それともただ無作為に撃っているだけなのかを知りたい**と考えた。もしもドイツがでたらめに撃ってくるだけならば、さまざまな防護設備を国中に配備すれば、国を守るのにかなり役立つだろう。しかし、ドイツが爆弾で標的を狙えるなら、イギリスは実に手強い相手と戦っていることになる。防護設備の配備くらいでは、国民を守ることは難しい。イギリス政府は、この疑問を解決するための調査を統計学者の R・D・クラークに依頼した。クラークは、過去の知見をレビューし、それに基づき簡単な統計検定を行った。

クラークは、**爆弾の着弾地点の分布の分析にポアソン分布を使用できる**ことに気づいた。ある現象が既知の出現確率に従い発生するならば、ポアソン分布はその現象が一定の時間、一定の面積、または一定の体積の中で何回起こりそうかという確率を教えてくれる。ポアソン分布を用いるには、現象発生の平均回数さえわかっていればよい。もしも爆弾が無作為に投下されているなら、ある特定の狭い面積に投下される爆弾の数はポアソン分布に従う。例えば、単位面横あたり投下される爆弾の平均個数が 1 個の場合、爆弾がまったく当たらない確率も、ちょうど 1 個の爆弾が当たる確率も、ちょうど 2 個の確率も、ちょうど 3 個の確率も、4 個以上の爆弾が当たる確率についても、ポアソンの式を用いれば簡単に計算できる。

クラークは、一定の面積単位に区切られたエリアごとの着弾数を調べるために、サウスロン

ドン地区を面積 4 分の 1 平方キロメートルの正方形 576 個に分割して、着弾数 0 個の正方形の数、着弾数 1 個の正方形の数、着弾数 2 個の正方形の数、着弾数 3 個の正方形の数といったように数えていった。もしロケット弾がランダムに着弾しているなら、ある正方形の着弾数が 0 個、1 個、2 個、3 個…である確率はポアソン分布に従うだろう。

実際のデータはポアソン分布に驚くほどよくフィットし、ドイツがロケットの標的を定めているという仮説はまったく支持されなかった。

定量分析もしくは統計解析用
・ＩＢＭ　ＳＰＳＳ
・R（オープンソースソフトウェアパッケージ）
・ＳＡＳ

データ分析について詳しい人は専門用語を使って話すことが好きだ。分析に使った統計手法について説明したり、回帰係数について説明したり、決定係数（データの変化量のうち、回帰モデルで説明される割合。第 3 章を参照）を説明することが好きだ。そのため、つい、聞き手もこうしたことについて聞きたがっていると思い込んでしまう。しかし、これは非常によくないことだ。ほとんどの聞き手は専門用語を理解できない。インターコンチネンタルホテル・グループの分析専門家の一人が言った。「誰も決定係数になんて興味がないんですよ」。

夫婦円満カップル
落ち着きがあって親密で、互いを支え合い、仲睦まじい関係を共有している。**個人単位の経験よりも夫婦で共有できる経験を好む。**

彼らが開発した数学的モデルは、結婚が持続しそうな二つのタイプの安定したカップル（夫婦円満カップルと回避型カップル）と、二つのタイプの不安定なカップル（敵対的カップルと敵対的・無関心カップル）の違いを明らかにすることができた。一触即発型カップルは、結婚生活は不安定ながら、結婚は持続すると予測された。

そして、研究に参加した 700 組のカップルが、1 年から 2 年の間隔で 12 年にわたり追跡調査された。その結果、**マレーとゴットマンの公式は 94 パーセントの精度で離婚率を予測した**ことがわかった。

このモデルは、ゴットマンとマレーらの「結婚の数字――動的非宣型モデル（The Mathematics of　Marriage: Dynamic Nonlinear Models）」というタイトルの本で発表された。

1938 年、物理学者のフランク・ベンフォードは、ニューカムよりもはるかに大量なデータの中に、同じパターンを発見した。

ベンフォードは、2 万 229 ものデータセットを分析した。そのデータセットの中には、河川の面積、野球の統計、雑誌記事の数、米国科学者名鑑に載っている最初の 342 人の住所などが含まれていた。各データセットは無関係であるにもかかわらず、各データセットに含まれる数値データの最上位桁に現れる数字（1 ～9）の頻度は、ページのすり切れた対数表が示していたのと同じパターンだったのである。この最上位桁に現れる数字（1～9）の頻度パターンは、ベンフォードを讃えて、**ベンフォードの法則**と名づけられた。ベンフォードの法則は、現実世界の多くの状況に適用できることがわかっている。

多くの統計家や会計士たちは、ベンフォードの法則を、詐欺、横領、脱税、不正会計などを発見するためのシンプルだが強力なツールであると信じている。なぜ、ベンフォードの法則を使えば不正を発見できるか、それはとてもシンプルだ。だれかがデータセットをねつ造したなら、*たぶんべ*

ンフォードの法則から予測される分布には従わないだろう。だから、すべてのデータの最上位桁の数字を調べて、1～9 の各数字の出現頻度を調べ、それとベンフォードの法則からと予測される出現頻度とを比較すれば、ねつ造されたデータか本当のデータ化を容易に判断できる。

最上位桁の数字	1	2	3	4	5	6	7	8	9
各数字が最上位桁に来る確率（%）	30.1	17.6	12.5	9.7	7.9	6.7	5.8	5.1	4.6

確率を理解することは、単に誕生パーティーのためだけでなく、さまざまなことを正しく理解するのに役に立つ。確率を理解していないと、株式市場の変動はランダムウォークであり（すなわち株価の変化には識別できるパターンも傾向もない）、また、投資家は、2～3 年は儲け続けたとしても、何れ必ず大損失をこうむるということがわからない。
「平均回帰現象」という現象も理解できない。それはつまり、**あなたの収入が平均よりかなり上なら、あなたの子どもの収入はあなたの収入よりも少ない可能性が高い**ということだ。

----------Ｗｅｂ----------
「平均回帰」という統計学用語がある。１回目の試験結果が「特別に良かった」あるいは逆に「特別に悪かった」事象において、２回目の試験結果は１回目の試験結果よりも、１回目全体の平均値に近くなる傾向が強い現象のことをいう。1887 年にスイートピーの種子の重さを観察していてこの現象を見出したイギリスの人類学者フランシスコ・ゴルトンは、人間の親子についても、**背の高い父親を持つ子供は必ず背が高いとは限らず、むしろ子孫はだんだん平均的な身長に近づいてゆく現象**を指摘した。
---------- ---------- ---------- ----------

するとサマーズはこう答えたという。「週に一度会社に行って、数学的トレーディングモデルをつくっている分析専門家たちの机のまわりを歩くんだよ。彼らにモデルの土台になっている仮定は何か、そして、どんな状況でそうした仮定が無効になるかを聞く。彼らから明確な答えを得られないことがどれほど頻繁にあるかを知ったらきみは驚くだろうね」。伝えられるところによれば、この仕事のためにサマーズには 500 万ドルが支払われているそうだが、本当ならば、この仕事は価値のあるものと考えられているに違いない。
あなたもラリー・サマーズのように行動できる。だれかにモデルを見せられたとき、**モデルの土台になっている仮定は何か、そして、どんな状況でそうした仮定が無効になるか**をきいてみよう。できる人と見られることうけあいだ。

メルク社のような大手製薬会社で営業チームの最適な規模を、分析的に判断するのは難しい。新製品が定期的に発売されれば、販売員の需要と必要性は増すが、既存の製品の特許が切れれば、販売員の需要と必要性は減小する。新製品がどれくらい売れるかについては、当然ながら過去の記録がないため、**どのくらいの営業チームが必要なのか**を正確に予測する方法を編み出すのは不可能だ。

----------Ｗｅｂ----------

真実を見抜く分析力

ビジネスエリートは知っているデータ活用の基礎知識

2014/07/18

山端 宏実＝日経情報ストラテジー

企業人向けデータ分析教本

著者は「分析力を武器とする企業」というベストセラーの著作もある米バブソン大学のトーマス・ダベンポート教授。ビジネスパーソン向けに、データ分析プロジェクトを通して分析結果を意思決定に役立てるための知見・ノウハウを盛り込んだ。

本書では「問題のフレーミング」「問題の解決」「結果の説明と実行」という 3 フェーズに分けて、分析プロセスを解説する。例えば、データ分析の結果を関係者に説明し、実行する段階では、「科学捜査型ストーリー」や「マッドサイエンティスト型ストーリー」といった 6 つの方法論を分かりやすく紹介している。これ 1 冊でデータ分析プロジェクトのイロハはつかめる。

真実を見抜く分析力

トーマス・H・ダベンポート／キム・ジノ 著

河本 薫 監修

古川 奈々子 訳

日経 BP 社発行

1944 円 (税込)

http://itpro.nikkeibp.co.jp/article/COLUMN/20140702/568502/?top_tl1