データをビジネスに活用する実践アナリティクス

<第 1 回>　ビッグデータとアナリティクス

梶山 昌之
株式会社ワイハット

1.　はじめに

今から 1 年ほど前には「ビッグデータ」という言葉がブームとなり，メディアでも盛んに取り上げられていました.

しかし，ビッグデータに対する期待が大き過ぎ，ビッグデータの分析が，いつも期待通りの結果をもたらすものではないということが明らかになるにつれ，しだいに，話題になることが少なくなりました.

しかしながら，ビッグデータのブームによりデータ解析およびその処理技術の重要性が認識された結果，ビックデータの処理技術は，現時点でも着実に進歩している状況です.

ただし，以前のようにビッグデータのみに過度な期待を抱くのではなく，身近にあるデータにも目を向けこれらを活用することが求められている点が異なります.

この連載では，企業または組織に蓄積されたデータ，およびインターネットで得られるデータの活用を表す用語として「アナリティクス」を用います.

2.　家族に関する俗説

ここで，アナリティクスの本題に入る前に，アナリティクスの本質とも関係するちょっと面白い話を紹介します.

この話は，医学統計学者であるスティーブン・センン氏の書著「生と死のパラドックス」[1]で紹介されているものですが，ここでは，図を加えて，わかり易く解説します.

今，テーブルには一人の男性がいて，「ブラウンさんには二人の子がいます．そのうち少なくとも一人は男の子です．もう一人の子が女の子である確率はいくつでしょう.」と聞いています（図 1）.皆さんは，どのように考えるでしょうか.

ほとんどの人は「男女はほとんど同じ確率で生まれるのだから 女の子である確率は 1/2 に違いない.」と考えます.

図 1　家族に関する俗説

続いて男性は「男の子だったら 5 万円払います．女の子だったら 4 万円いただきます.」という賭けをもちかけます.

図 2　賭けた方が得？

このような賭けを持ちかけられた場合，皆さんは賭けますか．もし，女の子である確率が 1/2 という判断が正しいのであれば，男の子である確率も 1/2 であり，もらえるほうの金額が 1 万円多いのですから，賭けに参加した方が得です（図 2）.

もちろん，1 回限りの賭けでは，負けることもありますが，このような賭けを何回も繰り返すと

すれば，得をすることは間違いありません．

ところが，答えは 2/3 になります（図 3）．何故なら，二人の子供のうち一人は男の子であるという情報が与えられていますので，上の子と下の 4 つの組み合わせの各事象のうち，二人とも女の子という事象は計算から除外されます．残った 3 つの事象のうち一人が女子である事象は 2 通りなので，2/3 になるという訳です（図 4）．

図 3　もう一人が女の子である確率

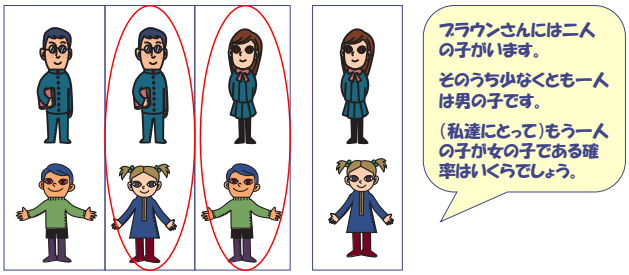


図 4　もう一人が女の子である確率は 2/3

もう一人が女の子である確率が 2/3 であることを認めた場合，この賭けに参加することによってどのくらいの利益（または損失）が見込めるかを計算しましょう．この値は「期待値」と呼ばれます．

この場合は，もう一人が男の子である場合と女の子である場合のいずれになりますが，男の子の場合または女の子の場合で得られる利益（または損失)に，それぞれの発生確率を掛けた値を合計して，期待値は -1 万円 (=5*1/3-4*2/3)，すなわち 1 万円の損失となってしまうことがわかります．

もちろん，一回だけ賭けに参加した場合は，勝つこともあるでしょうが，何度も賭けに参加すると，平均して一回あたり 1 万円の損失になるという意味です．

従って，「賭けには参加してはいけない！」ということがお分かりいただけたでしょうか．

ところが，この話の面白いところは，このような説明が終わった後にも，直観的に納得できない方が少なからずいらっしゃるという事です．何故でしょうか．

この問いは「ブラウン氏には二人の子供がいます．そのうち一人は男の子です．(その子にとって,) もう一人が女の子である確率はいくつでしょう．」と解釈することもできます（図 5）．

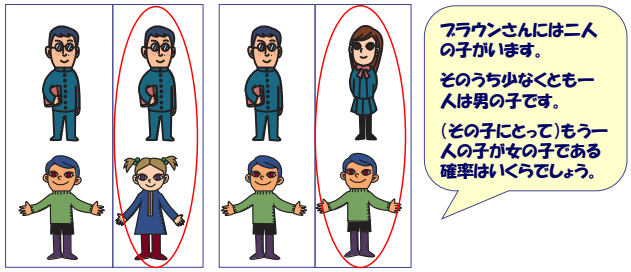


図 5　もう一人が女の子である確率は 1/2

最初に 1/2 と答えた人は，「そのうち少なくとも一人は男の子です．」と聞いた瞬間に自分自身がその子になった気分になり，無意識でそのように考えてしまったのかもしれません．

以上のように考えると，問題自体に曖昧な部分があり，1/2 という答えも誤りとは言えなくなってしまいます．

この例は，確率の計算では，多くの人が錯覚に陥り異なった結論を出してしまうこともあることを示しています．

現場で使用されている指標は正しく現状を把握し，適切な計算方式で算出されているでしょうか．見なおしてみる必要があるかもしれません．

3.　ビッグデータとは？

車の自動運転は 2020 年には実用化するということが話題になっています．車が歩行者の動きや他の車の走行を予測し，自動的に目的地に移動する技術は，大量のデータを収集して，リアルタイムに判断する必要がありますので，ビッグデータの活用の形態の一つです．

しかし，単に大量のデータであることだけが，ビッグデータの特徴ではありません．単に大量のデータということであれば，ビッグデータという言葉が登場するずっと前から企業や組織にも存在していました．

ビッグデータという言葉が登場した背景には，CPU の処理速度の増大，ストレージの高密度化，分散処理，パターン認識，音声認識といった技術

の進化があります.

これらの技術により，いままで活用できなかった，大量のテキスト，画像，音声なども分析の対象とすることができるようになったということがポイントです.

例えば，警視庁のデータベースには現時点で約1040 万人の指紋データが蓄積されていますが，画像イメージの状態で蓄積されているだけでは，容疑者の指紋照合をリアルタイムで行うということはできません．このようなデータは非構造化データと呼ばれます.

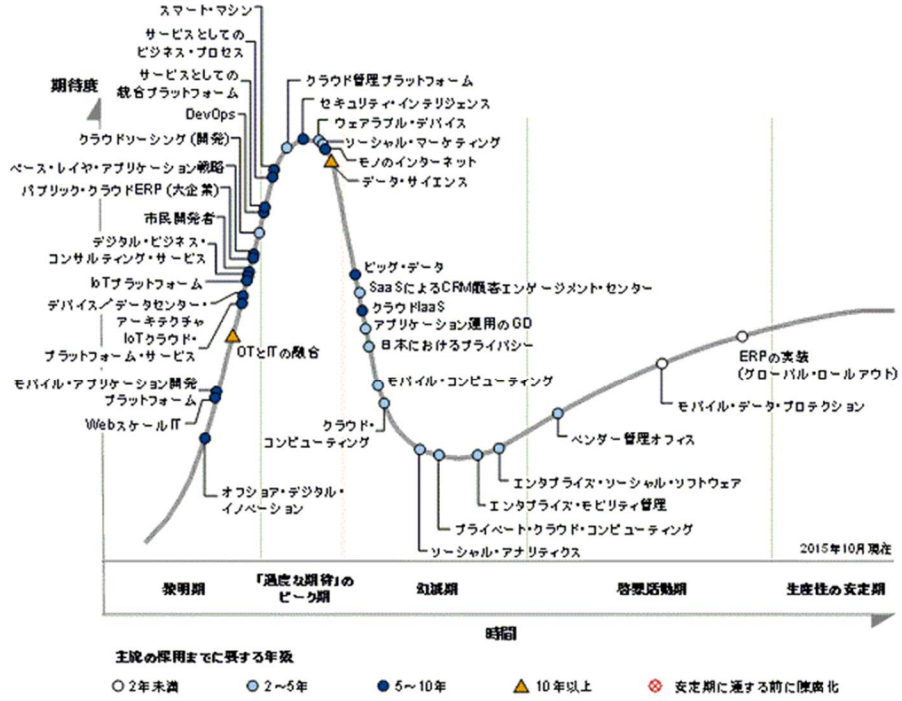
指紋には特徴点がいくつかありますが，この位置関係を数値データに変換して初めて活用が可能になります.

変換されたデータは，Excel またはデータベースのテーブルで表現できるデータとなっており，既存のデータ解析の手法で分析可能です．このようなデータを構造化データと呼びます（図 6）.

重要なポイントとして「非構造化データはそのままでは分析できない」ということです．これを構造化データに変換するための前処理の技術が進化したことが，ビッグデータが注目されることになった一つの理由です.

図 6　非構造化データは分析できない

ビックデータを特徴づけるもう一つの要素として，絶え間なく流れ込んでくるデータを対象とするということがあります. 例えば amazon で買い物をされた経験がある方は多いと思いますが，商品の注文データは直ちに収集され，購入者が次に購入する可能性がある商品を予測するために使用されます.



出典：ガートナー（2015年10月）

図 7　日本におけるテクノロジのハイプ・サイクル：2015

この様な仕組みを実現するためには，Hdoop/MapReduce に代表される分散処理の技術と，あいまいな情報から予測を行うためのベイズの定理を活用した予測手法の活用が必要になります.

以上の様に，ビッグデータの特徴として 3 つのポイントがあることがわかりました.

IT 分野の助言およびコンサルティングを行っている企業であるガートナーは，高ボリューム（Volume)，高速度（Velocity)，高バラエティ（Variety) の頭文字を取って, これを 3V モデルと呼んでいます.

はじめに，ビッグデータは 1 年ほど前にブームとなり，その後，次第に話題になることが少なくなったと述べましたが，現在はどのような状況にあるのでしょうか.

新技術全般に当てはまる傾向として，その技術が初めて登場した時点から人々の注目を集め話題となり，その後，過度な期待に応えることができず，一旦は幻滅期に移行するが，技術の利点と効用が正しく認識された後は着実な進展を遂げる次期（啓蒙活動期）に移行するという理論があります.

これをハイプ・サイクルと呼びますが，ハイプ（hype）とは「誇大広告」という意味です.

ガートナーは, 2014 年 10 月および 2015 年 10 月に, 各種の新技術がどの位置づけとなるかを分析した結果を発表しています[2].

この図で, ビックデータに着目すると, 2014 年 10 月の時点で「過度な期待」のピーク期にありましたが, 2015 年 10 月の時点では幻滅期に移行していると言えます（図 7）.

しかし, これはビッグデータやデータの利活用に価値がなくなるという意味ではなく, 現在も着実に研究と活用が進んでおり, 今後は啓蒙活動期に移行すると考えられます.

データ解析そのものは，昔からあった技術ですが，ビッグデータも分析の対象とすることができるようになり, ますます活用の場を広げています.

次回は，データ解析と統計解析の違いやデータ解析とデータマイニングとの違いなどについて解説の予定ですのでご期待ください.

参考文献

[1] スティーヴン・セン，松浦俊輔 訳: 「確率と統計のパラドックス」，青土社，ISBN4-7917-6164-2, 2005.

[2] ガートナー: 「ガートナー,「日本におけるテクノロジのハイプ・サイクル：2015 年」を発表」, 2015 年 プレス・リリース, 2015 年 10 月 27 日, http://www.gartner.co.jp/press/html/pr20151027-01.html.