論文

国際競争へ向けたビッグデータの正しい理解と定義

Accurate Understanding and Its Definition of BIGDATA for the Global Competition

岡　村　久　和（亜細亜大学都市創造学部　教授）

Hisakazu OKAMURA（Professor of Urban Innovation, Asia University）

〔要旨 / Abstract〕

日本におけるビッグデータの理解は、ビッグや、データと言う今や日本語化してしまった外来語からの連想により、国際社会や国際ビジネス社会と大きくかい離していると言わざるを得ない。ここでは、その正しい理解をし、国際社会が何を狙っているのかを正確に理解する。

We cannot avoid saying that the knowledge and understanding of BIGDATA in Japan is far apart from both the global society and global business world, due to the English words "Big and Data" which have almost become Japanese words meaning differently from their original. We will try understanding its correct meaning and the purpose of global society with that.

日本語になってしまったビッグとデータ

BIG は大きいと中学の英語の授業で習います。この翻訳が日本人の頭の中にこびりついているのです。では LARGE はどうなのでしょうか？

さすれば BIGDATA は、大量のデータなのか？ラージデータではいけないのでしょうか？

この大きな問題が日本の社会を間違った方向に導き、ビッグデータを取り巻く国際競争の中で日本の立場を悪くしていると言わざるを得ないのです。

この誤解によるマイナスは単にビジネス界の事だけだは無く、多くの研究者やソフトウェアを中心とする技術開発者にも、国際社会で勝てない努力を強いていると言えるかも知れません。

ソフトウェア技術者の話を出したのは、もう一つの意図があります。ビッグとラージの誤解とは別に、ビッグデータを構成する後ろの部分の単語、データについても、日本の社会では誤解があるのです。それはビッグデータが IT 用語であると理解する誤解、なのです。確かにデータと聞くとなんとなく IT 用語の様に感じる人も多いと思いますが、これも外来語翻訳のトリックなのです。データは日本語では、数値だったり情報と考えますが、果たしてこの数値や情報と言う言葉は IT 用語でしょうか？

データとカタカナで書くと IT 用語、数値や情報と書くと一般的に使う言葉に聞こえますね。

今回はこの日本社会のビッグデータのビッグとデータ、二種類の誤解について解説していきたいと思います。

例えば下の絵にあるように、私たちはお金を使って物を売ったり買ったりしている様に誤解していますが、実際の所、給与の振り込みも、アマゾンでの通販もデータをやり取りして、行われているのです。データと言う言葉の広い意味を理解する事が非常に重要と考えられます。

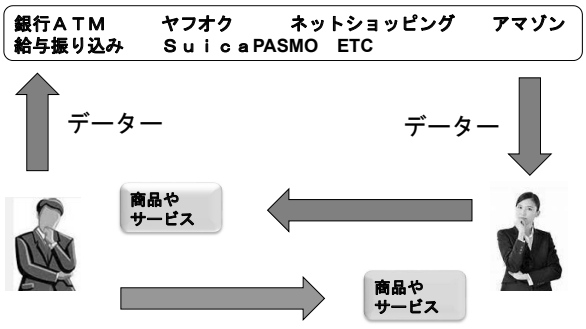

図１　実際には無いお金、しかしデータはある

ビッグの意味

ビッグデータのビッグを単純に大きいと訳して大量のデータを扱う手法と説明される事が多いのですが、実は本当の意味するところは大量のデータばかりではありません。多くのメディアもここを誤解して使っていることがあります。ビッグデータとは猛烈なスピードで飛んでくるデータ、正しくないデータ、無意味なデータ、文字や映像などのバラバラなかたちなどのとてもつかみきれない多様性をもった、そしてビジネスに使うデータを意味します。単に大量データを意味する単語ではありません。ビッグはラージとは違います。ビッグと言う英単語にはそもそも実態のつかめない大きさとか、物差しで計測できない大きさの様な意味があります。この誤解が日本のビジネスに少なからず影響を出しているのです。
無理やり新しい訳語をつけるとすると、ビッグは得体の知れない大きなもの、ラージは大サイズとなりますでしょうか。

データの意味

またデータとは、日本語でも使う情報や数値の事を言うのです。従って、ビッグデータとは、IT 用語ではなく、主にビジネスの用語です、つまり、ビジネスにおける得体の知れない大きな情報や数値と言う事になる訳です。

いかがでしょうか、ビッグデータを大量のデータと訳すのとは、少し違うと思います。

国際ビジネスチャンスを失う和製ガラパゴス英語

和製英語には一見それと気づかない物が多くあります。本来は恐ろしい“復讐”という意味の“リベンジ：Revenge”と言う単語は、“再挑戦”と言う和製英語で使われています。もう十数年前、あるスポーツ選手が再挑戦したいと言う時に間違って使ったのが広まってしまったと記憶しています。

日本で使っている分には良いのですが、海外でのビジネス現場でこれを使うと大変な事になります。品質向上に再挑戦するとか、顧客満足度第一位に再チャレンジするなどと良く言いますね。「今年は惜しくも品質コンテスト２位でしたが来年は再チャレンジされますか？」と言う質問を受けたとします。それに対して「はい是非リベンジしたいと思います」と言う事は「はい是非私たちは復讐し殺してしまおうと思っています（We will revenge）」と言う意味になります。とはなんとなく再挑戦するわけだは無く、ある特定の人を頭に描きながら殺してしまう勢いで復讐するというニュアンスを持った言葉なのです。こんな事をビジネスやスポーツで言ったら、間違いなくおかしな人だと恐れられるでしょう。

実は多くのビジネス流行語にも、日本で独自の翻訳をされ、和製ガラパゴス英語の育った言葉が意外と多く時にはビジネスまで阻害してしまう事があります。

定着していく誤解

ビッグデータと言う言葉に関しても同じことが起きています。はやり始めた頃は、それほど大きな実害は感じませんでしたが、本来はビジネス戦略の言葉なのにデータという言葉が原因で IT 用語としてだんだん取り上げられるようになりました。日が経つにつれビッグデータは主に大量データと解釈されて行きました。最近では国外とは別の大量データと言う理解のまま、どんどん普及が進んでいるように感じます。外資系大手 IT 企業が言葉としての位置づけをきちんと説かずに、大量データを扱えるサーバーやネットワークの宣伝の中で使ってしまったのもその一因と思います。アメリカやヨーロッパの企業や社会ではビッグとラージの違いをきちんと理解しているので問題ないのですが、そのビジネスを日本に持ち込むときにきちんとした定義を社会に説明せず安易にカタカナでビッグデータと使った事がその要因でしょう。現にスマートと言う言葉は以前は痩せてすらっとしている人を掲揚する言葉でしたが、スマートシティと言う言葉が入って来た時に、比較的丁寧に意味を説明し、賢いと言う訳語を充てました。さらに、日本のビジネス社会もスマートとは何だろうと真剣な議論を繰り返した為に、まあ比較的正しく日本社会で訳され理解されている様に思います。

ところが、ことビッグデータに関しては一般企業やビジネスマンの間にもビッグデータは大量データとして広がってしまったので、これを大量データをうま

く利用する事と捉える人が増えてしまいました。当然、その戦略は国際的な意味合いから遠ざかり、日本独自のガラパゴス英語文化の一つとなって行ったのです。

一方で、世界の人々には、日本人と同じようにビッグデータを大量のデータと誤解している人も数多くいます。しかし、日本の様に多くの人々が誤解している状態とは違います。

本来の“BIG”の意味

日本では中学一年生の英語でBIGを習います。“大きい”と訳したと思います。その直後今度は“Large”を習います。これも”大きい“と訳しました。

さて、BigとLargeどう違うのでしょうか。私は英語の先生に質問した覚えがありますが、“両方ともほとんど同じ（大きい）という意味で、その後につく言葉によって変えるんだ”と言うのがその時の答えでした。

これが根本的な間違いなのです。これがビッグデータを使ったビジネスの誤解の根本原因でしょう。

実はBIGとLarge　の意味は全く違うのです。人間の感覚の話ですので、日本語の訳だけで考えれば“大きい”で良いのかもしれませんが、言葉は人間の感覚や感情を表す道具であり、違いがあるからこそ言葉も違うのです。日本語で同じでも全く同じ意味の外国の言葉など無い。単語が違えば微妙に意味は違うはず。ここが重要です。

例えば偉大な人、小柄でも世の中で重要な人、アメリカアップルの創始者スティーブジョブスなどは“BIG-MAN”と言われます。偉大な人、と言うようなニュアンスですが、私達とは根本的に違う掴みどころの無い大きな人と言うニュアンスなのです。“LargeMan”と言いませんね。

“俺はいつかBIGになる！”とは日本語でも言いますが“俺はいつかラージになる！”とは言いません。

洋品店で気に入ったTシャツを買う時や、ハンバーガーショップでコーラのサイズを聞かれる時”Lサイズの長袖”とか“コーラはLサイズ”と表現しますが“ビッグの長袖シャツ”とか“コーラのビッグ”とは言いません。

それでもビッグサイズのポップコーンやビッグバーガーなんて使う時は、途轍もなく、すごく大きいと言う意味を込めていると考えてください。

“ラージ”と言う単語は、物理的に大きい、物差しで測れる、サイズであってそれ以外の価値を全く説明していない“大きい”と言う意味なのです。一方で“BIG”と言う言葉は、“つかみどころの無い大きさ”“自分より明らかに大きいがその大きさが分からない”“物差しなのか、体重計なのか、長いのか、高いのか、太いのか、重いのか計測する道具も選べない大きさ”と言うような意味を持っています。

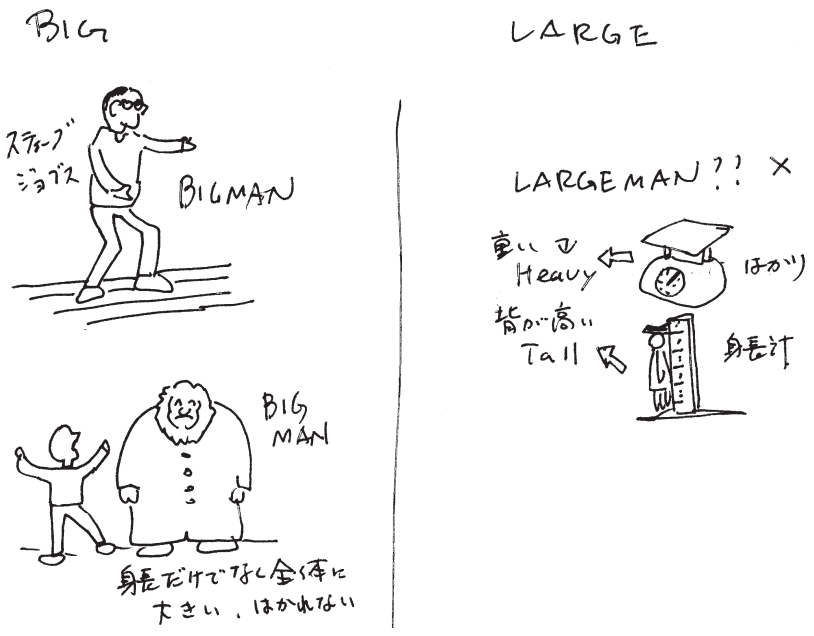

図2　ビッグとラージ　その1

お分かりと思いますが、ビッグデータのビッグとはこう言う意味を持って表現された言葉なのです。つまり“よくわからない大きさ”といったニュアンスです。

ではビッグデータとは何？

これまでに説明した様にビッグという言葉が何か物差しでは測れない大きな物、逆に物差しや計りで計測できる大きな物をラージと表現する事は理解できたと思います。

ではビッグデータとはそもそも何を意味するのでしょうか。答えは至って簡単です。物差しや計りで計測できない、数字できちんと表せないデータとは何でしょう。途轍もなく大量のデータでも数字や計りで表せればビッグデータでは無くラージ・ボリューム・オブ・データつまり大量のデータと呼びます。

ビッグデータはこのようなデータの量であるとか、サイズを意味していません。ビジネスの判断に必要な様々な形をした、様々な性格を持った、様々な種類のデータを意味し、それらを駆使してビジネスの効率や効果を上げる事が出来る情報と、その情報の使い方の事を表すのです。

ビッグデータの目的

図４は米IBMが顧客に取ったアンケートの結果ですが、その半分がその目標を顧客中心の情報を取るためとしています。

ビジネスにおいて、顧客中心の情報を欲しいとは、どういった状態でしょうか？　もちろん、製品のデザインや、顧客に好まれる機能設計から、サービスの改善や物流など、顧客に関する情報を取得したいという目標はビジネス活動のほとんどのシーンに登場します。

しかし、それらのほとんどの場合、顧客情報を持って行う事は、意思決定なのです。小さな意思決定から、企業レベルの意思決定まで、実に多くのケースで顧客情報が利用されます。

つまりビッグデータとは意思決定を行うためのツールなのです。

ただ、単純に大量データであると言う定義とはかなり違う事がわかると思います。

図３　ビッグとラージ　その２

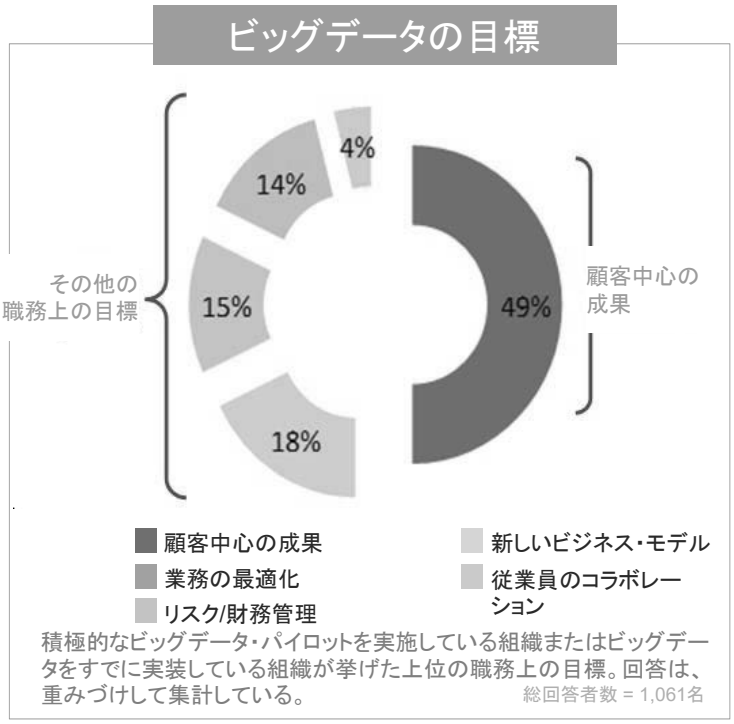

図４　ビッグデータの目標

いい加減でも良い加減な顧客情報

余りにもいい加減な情報だとわかっていても、まあまあ当たらずとも遠からずと言う事は良くありますね。いい加減な情報であっても大量の情報を分類してみると何となくその傾向が分かる事もありますね。

顧客の考えている事などは、流行や趣味嗜好に影響されて、日々変わっていくと考えられます。ところが、日本で誤解している大量データでさえ、それを取得するとなんとなく傾向値が出たり、なんとなく流行っている物の方向性などが推測できます。大量のデータを使った統計にも、もちろんツールとしての強い使い道があります。後に述べますが、大量データもビッグデータの重要な要素の一つなのです。

ところが、統計数値や大量データでは出てこない情報も世の中には数多く存在します。地域別の売り上げ集計の様な、大量データで出てこない情報とは、良い加減な情報なのかもしれません。人間の思考はすべてエクセルの表に記述できません。私たちは、なんとなく好きとか、なんかいやだとか、その本人でもうまく定義できない感覚を持っています。この人間特有の良い加減の思いや、趣味嗜好は統計データには出てきませんが、人の行動には表れてきます。

人の行動とは、言葉や、活動、の事です。顧客情報を取得するには、この顧客と言う人間の言葉や活動を把握する必要があるのです。この良い加減のデータをビジネスで使えたらと言う発想がビッグデータ活用戦略の重要な一側面です。

では一体、言葉や活動の情報なんて取れるのでしょうか？　その代表と言える物がSNSです。Facebook、Twitter、ブログなどの大量の媒体には、個人が数多くこの良い感じの言葉を載せています。SNSには、旅行先の情報や行ったお店の情報も載っていますが、そこに挙げられているデジタル写真の多くはGPSの位置情報が埋め込まれています。もちろんタイムスタンプと言う、作成時刻も付随しています。これがビッグデータの重要な対象なのです。

間違った情報はどうするのか

ビッグデータでSNSを考えるとき、その中にある大量の間違ったデータの存在に気が付きます。人は、物を思うとき、自分でも気が付かないうちに、間違った記述をします。また、一度書いたブログの内容を次の日に打ち消すなんてことも普通に行われます。そう考

えると、SNS などはデータとして信用できないのではないか？　そんな疑問も湧いてきます。

実はこの事は一般の統計でも起きる事なのです。ただ統計情報の場合、その大量のデータ取得の理由もあり、間違った情報は極力排除され、信用するに足りる情報が使われます。例えば100人のうち１人だけが極端に違う数値を出した場合などは、残りの99人で考えたりしますね。ビッグデータの場合は、この間違った情報排除の方法が若干違うのです。

ビッグデータの一つの要素として大量データがありますので、一般統計の様に大量データの中の異質なデータを排除する事も行います。しかし、ビッグデータとは多種多様の観点からデータを見ていくので、数の理論だけで排除の判断をしなくても良いのです。ブログや SNS には GPS などの位置データが含まれるケースがあります。ある統計で特定の観光地の顧客動向をビッグデータを使って把握し、意思決定をしようとした時、そのデータに GPS 位置情報や、作成時間情報が入っていれば、対象の観光地以外から入力した情報や、他の帆ととは違う日時に入力した情報は、明らかに対象では無い事がわかります。

この例は非常に簡単ですが、実際には日時、場所のみならず、IP アドレスや防犯カメラ映像などなど、映像、画像、傾向、推移、など、取得できる限りのデータを相互にかけ合わせてその信ぴょう性を測ります。これは人が人の言葉を信じる時に使うやり方、アルゴリズムに非常に近い方法論だと思います。只一点違うのは、データの対象が途轍もなく大きく、広域で文字や映像など様々な媒体から得られるものであるという事でしょう。

ビッグデータはコンピューター用語では無い

こういう説明をすると、なるほどコンピューターうごの進化でビッグデータを使える様になったのだから、ビッグデータを活用する為には IT をしっかり勉強しなければならないと思う人が日本には大変多いのも事実です。書店に出かけビッグデータ関連の本を見ると分かります。書店ではビッグデータ関連は IT やコンピューターの棚に置かれる事が多く、実際その内容もデータの解析手法やソフトウェア、データベースに関連した物があまりにも多いのです。

私の所属する亜細亜大学都市創造学部では、学部の設立を考えカリキュラムを設計している過程でこの点だけには早い時期に気を付けて作業を続けました。ビッグデータ活用やビッグデータ活用実習などと言う授業名からは、多くの人がコンピューターを使ったソフトウェアの操作実習のイメージを持ちます。

実際のビッグデータ活用とは、前述した様々なデータをスケートボードのメーカーなら何に利用できるか、洋服屋さんなら何に使えるのか、レストランチェーンならどの食材仕入れに使える情報を得られるのか、これらがビッグデータの活用なのです。

ビッグデータはコンピューター用語などでは無いのです。それどころか、子前述の図にあるように経営者はコンピューターの事はおろか何のデータがどこにあるかなど細かいデータについて考える行為の優先順位はとても低いのです。経営者であれ部門長であれさらに営業マンや製造マンであれ、自分の仕事として「この情報があったらビジネスはもっとうまく回る」と言う考えをしっかり持てるかどうかがビッグデータの活用では最も大事な事なのです。

何が欲しいのかを明確にすることであって、どんなデータを得られるのか、どうやったら得られるのかは二の次で良いのです。

ビッグデータの生まれた背景と国際社会の理解するその特徴

ここからは、ビジネス実行者としても最低限知っておくべきビッグデータの生まれた背景や国際的なビジネス社会で理解されている技術的な特徴の話も入れて行こうと思います。

ビッグデータはその性格からどんどん複雑性を増し、どんどん業務に特化した戦略的な使い方が進化してきています。そう考えるとビッグデータの専門家に必要な知識とは業界知識、業務知識、そして情報の所在が予測し活用の方法を発想出来る能力なのです。

欧米ではと良く言われますが、ことビッグデータに関しては欧米諸国はもちろんアジア各国、インドやアフリカでもこの認識はほぼ正しく理解されています。なぜか日本だけこの解釈が間違ってされている事が多いのです

それではそもそものお話に入って行きます

ビッグデータの起源、３つのＶと５つのＶ

さてここでそもそもビッグデータとはいつ誰が言い出したのかそんな話題に入って行こうと思います。

まず2001年に META グループで定義されたビッグデータの定義について説明しましょう。この META グループの定義ではビッグデータは当然“よくわからないビッグなデータ”とされその表現に３つのＶが使われました。この３つのＶが提唱された瞬間がビッグデータの言葉の誕生と考えて良いと思います。

その３つのＶとは、まず第一に volume（量：データの量），そして二番目が velocity（速さ：データが入ったり出たりするスピード）最後は variety（多様性：データの範囲、種類、源泉）．と言われました。

前述した様に、日本での誤解のほとんどが最初の VOLUME と言う単語がボリュームつまり大量としてまた BIG がビッグで大きいとなり、日本の中で外来語でありながらもほとんど日本語になってしまった事にこの誤解は起因しています。“ビッグデータは大量のデータ”と解釈されてしまっている原因です。このあたり少し補足したいと思います。

もし大量のデータをビッグデータとするならば、英語的には“ラージボリュームオブデータ Large Volume of Data”となるべきです。データの大きな量という事です。英文法的に見てこれをラージデータと出来ない理由もあります。Big は曖昧な形容詞なので Big Data で文法的にも正しいですが、Large は“大”なので、大きな“サイズ”、大“量”などと何が“大”なのか、正確には単位を付ける必要があります。Large Volume, Large amount，Large Number と言った具合です。

ビッグデータの３Ｖでは、途轍も無い大量で、追いかけられないほど速いスピードで動き、文字や映像の違いだけでなく出所も種類もまちまちなデータをビッグデータと定義しています。

この３Ｖに対し、近年では、Variability（可変：データが変わっていってしまう事）と Veracity（真実性：信用できるデータかどうか）が加わって５Ｖとも言われています。

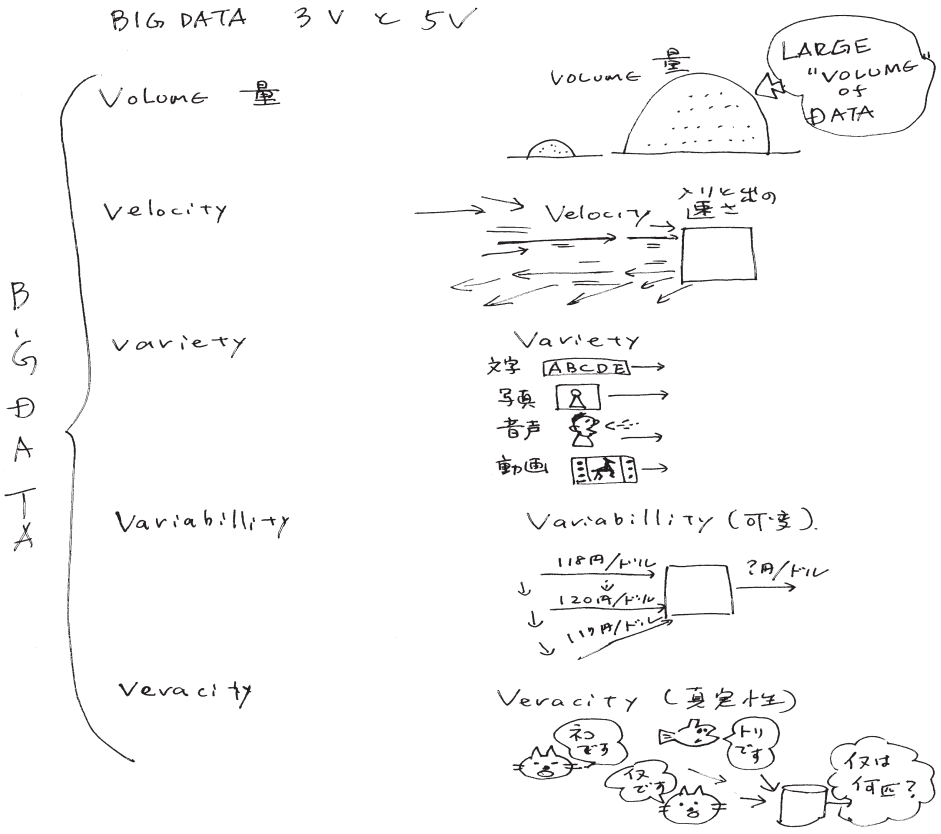

図５　ビッグデータ５つのＶ

ビッグデータ、具体的なその特徴から来る定義

1）Volume（量）

IT 古くはコンピューターと言われた時代には、その典型的な構成は計算する本体とそこから延びるケーブルその先にぶら下がる端末とキーボードでした。ほとんどのデータは人の手によってキーボードから入力されました。１分間に30語180文字のタイプが可能な人が100人いるとデータが１分間に3000語18000文字のデータが入ると言う計算になります。非常に簡略化して計算しますが、これを60分×８時間で計算すると、18,000文字×60分×８で864万文字です。一文字２バイトと考えると1728万バイトつまり17メガバイトです。iPhone６の最低記憶容量は16GB つまり16万メガバイトですからこの計算で例にあげた100人の手入力者が入力できるデータの9411倍です。

昔のコンピューターは端末が100台ついているような物でも iPhone の9000分の１の情報しかなかったと言う事になります。それでも企業はそのデータを元にビジネスを行いました。

当時もコンピューターは専用回線や電話線で接続されていましたが、基本的には一つ一つ相手を特定しての接続でした。線ですね。インターネットの構造や歴史を語るつもりではありませんが、インターネットのインターとは複数の物のつながりを意味し、ネットとは網の事です。インターネットとは網と網が繋がったと言うような言葉です。インターネットの普及でコンピューターは網の目の様につながって行きました。この時点で人の入力に頼っていたデータは相関関係を持つようになり、データとデータは結合や構造化をされデータの量も増えて行きました。この時点では、データはまだダイナミックに動かず、入力されたデータはどこかに蓄えられ、使う時にそれらは取り出されて処理されました。従ってこの時に必要だった仕組みは、検索の仕組みと大きなデータベースだったのです。ビジネス上何か重要な相関関係を見たいとか、地区ごと、日にちごと、性別毎の製品の売り上げを見たいなどという需要にはリレーショナルデータベースがあれば事足りました。リレーショナルデータベースとはデータを順番に並べて置くだけでは無く、データ同士に関係性を記述して大きな記憶装置にデータを効率よく格納する仕組みでした。あくまでも静的に格納されたデータを正確に高速に検索する仕組みと整理されデータを存在情報を持って収納する仕組みが必要でした。これらさえあればこの時代のデータベースを高性能で管理しビジネスへの貢献には十分でした。

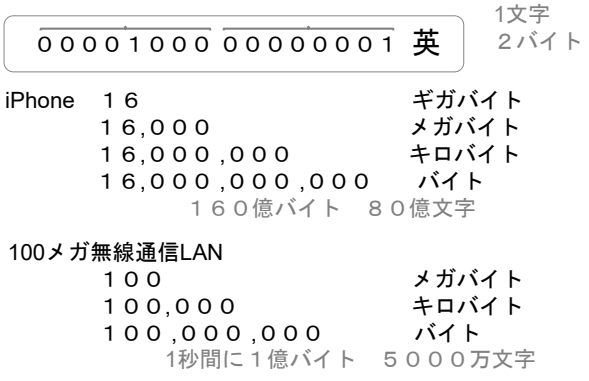

図６　ポケットの IPhone に収まった80億文字

データの量が人間の手入力量と同じで、いい加減なデータもほとんど入って来ませんでした。さらにすべて手入力なのでデータの入力もゆっくりとした物でした。

図６を見てもらうとわかりますが、今 iPhone を始めとするいわゆるスマホの中には80億文字の情報があり、家庭内 LAN の100メガ通信とは、一秒間に5000万文字を送ると言う事なのです。このスマホがほとんどの国民のポケットやカバンにあり、この通信速度でやり取りをしている訳です。大量データと一言で言ってもこれまでの大量データとは文字通り桁が違う世界なのです。これがビッグデータで言う大量データなのです。

Volume 量と IOT

IOT もビッグデータを生み出した大きな原因です、それまで人が入力をしていたデータの量を凌駕する大容量を機械が勝手に入力する様になりました。

人間の入力とは全く関係の無い防犯カメラ映像や、交通信号や照明、自動車や電車、工場の生産設備、携帯電話など、使っている人さえ知らないうちにデータをやり取りしている機械がデータを勝手に作り、勝手にインターネットを使ってデータを送ったり受けたりするように人間によって進化させられてきました。

昔、企業が使うコンピューター増え、映像や画像でデータの量が増加した事によってデータ量が増えたと

されました。現在では機械や機器から人が直接情報を取りたいと望む事で技術を著しく進化させ、とうとうそれらがインターネットを使える様に技術が進歩したのです。

把握しにくい種類や量のデータが表れてきます。バラバラで信頼のおけないデータなどが飛び交う中でいかにして正しい情報だけを把握すべきか“と言うニーズも生まれてきます。

ビッグデータの誕生原因の一つに IOT

大量のデータを使って何か統計的な特徴を引き出す事は、現在のコンピューター技術ではさほど困難ではありません。量が多いのですから、データベースに格納するディスク装置の量や、データを検索したり並べ替えたりするコンピューターの機械的計算能力だけがあれば良いのです。

図にあるようにビッグデータには量ばかりでなく沢山の捉えづらい特徴がある理由の一つに IOT の普及があります。これまで人間が入力してたデータに加え機械が自動的に送ってくるデータも格段に増えています。量が格段に伸びた事が問題であり、それを解決すればビジネスに役立つのであれば、大型のコンピューターや処理の早いコンピューターを導入すればそれだけで良かったのですが、工場や現場などで勝手に働いている機械のデータまでも利用しようと人間は考えるわけです。

機械が持っているデータを利用しようとすると課題はデータ量ばかりでは無くなります。大体言葉で情報が上がって来ないのですから、そのデータの方法から根本的に考えなければなりません。データをうまく集められたって人間の数万倍数百万倍と言う情報を吐きだす機械データをどうやって見極めるかだってそう簡単な話ではありません。

ここで出て来たテクノロジーが IOT であり、この IOT が出て来たのでさらにビッグデータが複雑化していったのです。

機械がインターネットにつながったばかりでは無くつなげて利用しようと言う人間の意図が基本にあるのでこの IOT はどんどん進化し複雑化し現在では機械のデータを機械が受けて判断し対策を打つところまで来ています。

音声データも文字情報？

この図を見ていただきたい。これは、内閣官房で個人情報保護法の新法検討で私が警鐘を鳴らしたチャートの一つですが。人の話し声、いわゆる音声データも大きな大量データ創成の原因となっているのです。今の音声認識の技術は革新的に進んでいますが、その認識と言う言葉が誤解を招き、認識した後にその文章が大量保存され利用されている事に目を向けるべきなのです。

例えば、図８にありますが、これは実際のコールセンターでのオペレーターと顧客の会話を即時テキスト化。つまり文字化し、そしてその場で傾向分析をしてマネージャーが確認している画面です。コールセンターにおける、オペレーターと顧客のやり取りなどは多くの企業で音声認識を行い、その後分析や戦略利用に使われています。

一人のオペレーターが一秒間に２文字話すとすると、１時間7200文字です。100名のオペレーターが一時間話す文字数72万文字と言うことになります。SNS のデータに比べ途轍もない量ですね。それらが日々蓄積され利用されるのがビッグデータの世界です。

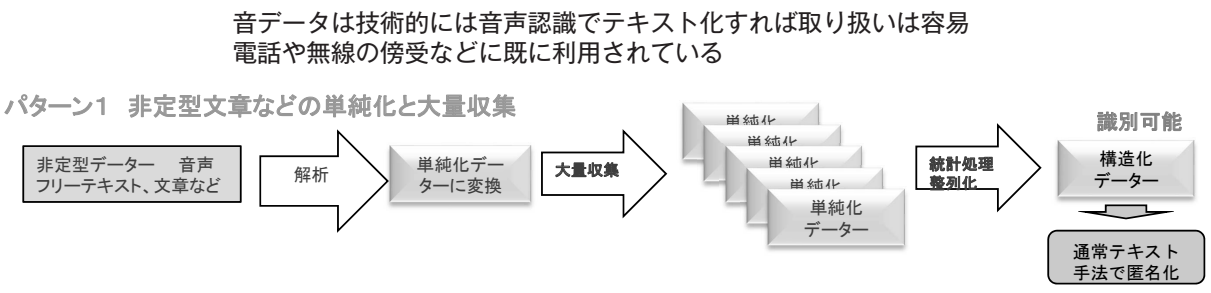

図７　音声データはテキスト化（文字化）される

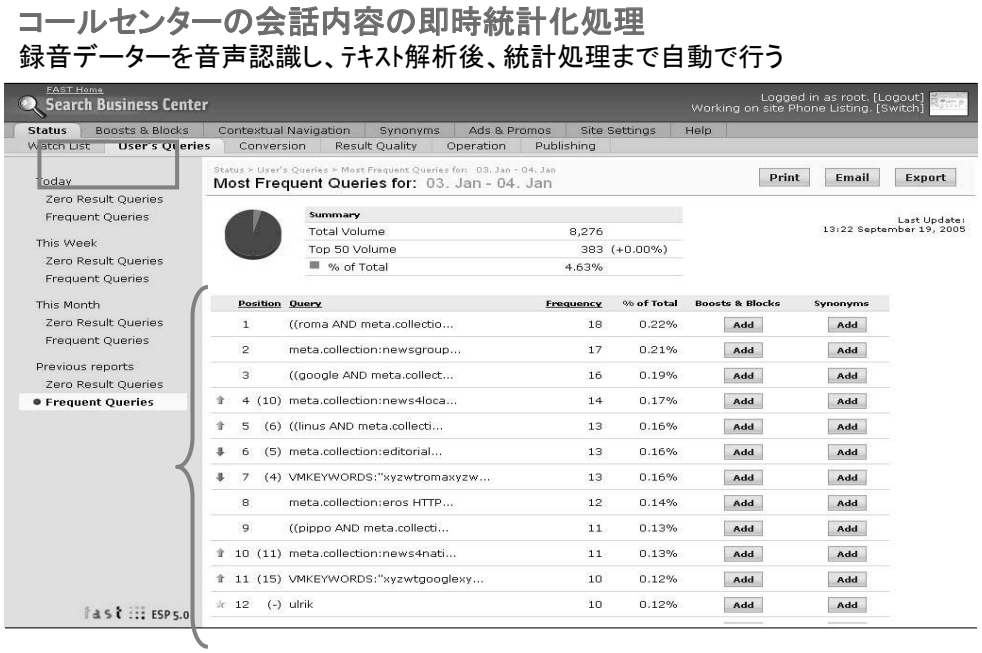

図８　実際にコールセンター会話をテキスト化している例

２）Velocity（速さ）

Velocity、データの速さとはいったい何の事でしょうか。ネットワークを流れるデータのスピードや、ディスクに存在するやコンピューター内部のデータのやり取り処理のスピードも速さと言います。

例えばネットワークですが、まず回線に流れる文字の量の話をします。銀行にATMが設置されてオンラインシステムが一般的になって行った頃、ネットワークに流れるデータのスピードは9600bpsつまり毎秒9800ビット一秒間に612文字が一般的でした。

今、家庭用のLANや光ファイバーでを使った家庭用のインターネット回線のスピードは100メガビットなどと言いますね。一秒間に625万文字の伝送スピードを意味します。昔のざっと１万倍です。それに昔はコンピューターの回線はポイントトゥポイントと言って、一つの場所から一つの場所にそれぞれ一本ずつ繋いでいました。今はインターネットをはじめとして複数の場所から複数の場所にネットが繋がっていますね。こんなに急激に増えた大量のデータが四方八方に流れているのです。IOTによって機械が勝手にどんどんデータを排出するようになりそのデータが昔の銀行ATMの１万倍の速さで世界中縦横無尽に飛び交っているのです。これがコンピューターシステムが進化し情報を高速で流通させることが出来るようになったと速さというインフラストラクチャーの側面です。

千分の一秒単位で発生するデータ

その一方で、データが発生してから流通するまでの“速さ”にもビッグデータの特徴があります。

株式市場を考えてみましょう。かつては仲買人が立ち売人と顔を合わせてセリの様に株を売買していました。証券会社では電話などで入ってくる株価情報をそれぞれの店舗にあった黒板に担当者がせっせと書き続けていました。

一般客はその黒板の数字を見てから担当窓口に行き手続きをして株券を買い、その紙の株券を持ち帰ってその株価の上下を毎日新聞で見ていたわけです。

この株の売買の仕組みが、まずシステムオンライン処理化されました。このオンライン化と言う表現でIT化されたまたはITのシステムなのだと感じる人も多いとのですが、株価のオンライン化とはセリの仲買人と売人とのやり取りをシステムで代行し、各証券会社の営業所に電話で伝達していた株価を伝送し、黒板に書いていた数字を端末に出し、株券を印刷するのを止め電子化しましたが、取引の判断は相変わらず人間が

行っていて、その取引の流れもなんら変わっていませんでした。

株のオンライン化とは、鉛筆とソロバンで計算していた経理事務をパソコンのエクセルに変えたような物で、仕事のやり方は根本的には変わっていませんでした。

さて、現在の株の取引はどうでしょう。株価を見て株の売買判断を行うのもシステムで行う様になりました。株を買いたい人が株を買う判断の条件をあらかじめシステムに覚えさせておきます。例えばA社の株が100円になったら1000株買うと言った具合に変化しました。オンラインで世界中の株に関して売買が出来るようになりました。

これだけではありません。例えば小麦粉の値上がりで製粉会社やその先のケーキ会社の仕入れ価格が変わり、販売価格にそれらが反映されます。そうすると当然売上高にも影響がでるので株価も変わります。もっと遡ると小麦粉の価格は産地の天候に大きな影響を受けるわけです。天候の変化をさらに遡ると、高気圧低気圧、海水温の上昇など地球規模の気象変化がそれに影響している事が分かります。

こう考えると、エルニーニョ現象による南米の海水温の上昇が最後にはケーキ会社の株価に影響を出すと言う事になります。

ケーキ会社の株で高い利益を得たい人は当然この海水温上昇の元となる地球規模の現象まで継続して監視したくなります。

今の株に関する情報の仕組みはここまでを追いかけているのです。例えば、アメリカの小麦の産地の温度が１度上がったら、すぐに株を買いたい人に警告メールを飛ばす、そんな仕組みが世界中で動いています。実際にはこんな単純な条件での警告では無く、温度や場所、為替、戦争、など本当に様々な条件を掛け合わした検討結果も株式取引の中では飛び交っているのです。

これらの情報は人間が手入力している訳では無く、様々なセンサーや機械機器にインターネットが接続され、複雑な条件の元に情報をマイクロセカンド（線分の１秒）の単位で送り出しているのです。これも典型的なIOTなのですが、この高速で発信されるデータの発信のスピードもビッグデータの速さの一つです。

株の投資家が多くの情報で自動判断されて株を買うとすると、それは実際にはコンピューターが株を買った事になります。このコンピューターが買い注文を出すと株価が変わり、別のコンピューターを使っている人が自動的に株の売注文を出します。これはあたかも対戦ゲームをコンピューター同士が勝手に行い人間が対戦していない様に見えます。

コンピューターはマイクロセカンドで判断しますので売りも買いもその先の売りも買いも大変なスピードで行われます。

さて結果はどうなるでしょう。どんなに早くても数秒はかかっていた人の判断による株式売買の時間は千分の一秒を大きく切るスピードにまで上がって行きます。データの発生そのものや流通の開始スピードにも革命的な変革が訪れました。

３）Variety（多様性）

初期のコンピューターシステムはいわゆる文字で情報処理をしていました。その後のシステムではYoutubeやインスタグラムなど画像や映像や音声もふんだんに使われるようになってきました。この事はデータの多様性と言う側面では大変大きなインパクトでした。

ここで言うビッグデータの多様性はコンピューターが元々理解できる文字信号から機械が発する電波信号、Youtubeやインスタグラムなどの映像信号、Suicaなどの交通ICカード、などなどの情報を多様なデータの入手先と考えるわけです。データが多様化している、または多様化したデータを扱わなければならないと言う事です。多様化しながらもデータの量やいい加減さスピードなども同時にその特徴として重なるわけですから、それらの利用を可能にする技術の進歩が求められます。

一方では多様化し、一見あまり関係ないと思われるデータも何とか利用しようと言いう発想も求められるわけです。今では当たり前になりましたが、ツイッターのつぶやき情報を大量に集め構造化して市場の動きを見ますが、この利用方法の発想をする事もビッグデータ利用の大きなビジネス技術の一つと考えられます。

もう少し、多様化した情報の利用例を紹介しましょう。身近にある例としては銀行の印影のチェックもその一つです。昔は人間の目で確認作業を行っており、かなり最近までシステム化されなかった一つです。これは多分に画像の認識制度への信頼性の課題と、重要な判断は最後人間が行わなければならないという銀行法や銀行の商習慣によるものでした。身分証明書として使われる物も画像データ処理の対象になってきています。マイナンバーカードやパスポートなど画像と電子チップを合わせて収納し相互に偽造防止のセキュリティや使用履歴の保存を行う例も増えてきました。一枚一枚は個人が使っていますがそれらが集まると高度なビッグデータになっています。これらはネットを介して蓄積され保管され利用されます。いまでは日本の街ので大変多く見かけるようになった防犯カメラですが、かつてはすべて人が目で確認し不審者の発見などに使っていました。ところが今は段々人の目での確認作業が減り、反対にシステムが画像や映像解析技術を使って自動監視する事が多くなってきました。これらも一見監視システムに見えますが、画像処理された情報が車の流れであれば交通に、車種まで捉えれば自動車会社の販売調査にと様々なビッグデータとしての利用が可能になります。ニューオーリーンズ市を始めとする幾つかのアメリカ合衆国の都市では地図上に犯罪履歴をアイコンで表示しクリックすると犯罪の種類や発生時期、被害者の名前までも表示しています。これも一見地図上の情報表示システムの様に見えますが実はその逆です。犯罪情報は発生時期、被害者名、発生場所、理由など単純な犯罪データベースだけでなく複数の管理部門で持っている情報を合わせる事によって完成されます。また公開できる物できない物、間違った情報、古すぎる情報など情報の公開にあたっての性格もそれぞれ違います。そもそもそれらの多様化した情報を犯罪捜査や犯罪抑止の為にビッグデータ技術を使って集めて利用する事が本来のビッグデータ利用目的でしたが、その膨大なデータを地図の上に張り付けて表示する事を発想し実行する事によって視覚的に観光客や住民に危険な地域への警告を発信しているのです。危険と表示された地域の住民や企業は防犯への意識を高め街を良くしようと言う気持ちがはたらくそうですが、土地の価値が下がるので反対意見を持つ不動産業者も多いと聞きます。

日本では高速道や幹線道路にはＮシステムと言う警察の監視システムで走行中の車のナンバーをすべて読み取り同時に運転者の写真を撮影しナンバーと運転者を合わせて情報保管しています。これはもちろん犯罪捜査にも使われますが、同じ車がある距離を何分かかって通過したかと言うデータを取得する事によって渋滞の予測やドライバーへの各種警告などにも利用されています。

駅の改札でピッっとなる IC チップ付きの交通カードですがこれにも以前のマグネットなどとは違い、駅の改札の機械が瞬間的に電気を交通カードに送り、交通カードは瞬間的に受けた電気を使って自分の持っている情報を改札機に返すわけです。この改札機と交通カードで行われる双方向での会話で改札が開いたり残額を表示します。この情報が改札機を通して鉄道会社に上がりさらに統計や管理に使われるわけです。さらに交通カードにクレジットカード機能がついていれば、さらにその情報はクレジットカード会社に渡されデータはどんどん流通して行きます。

この様に、データの性格を示す３つのＶや５つのＶばかりでは無く、情報を入手する仕組み、それを転送する仕組みなど、物理的形態から利用の頻度、相互利用される方法まで様々な多様性が生まれて来ています。文字を一つ一つ手でコンピューター入力していた頃と現在とでは社会システムにも大きな違いがある事が理解できると思います。

４）Variability（可変：データが変わっていってしまうと言う事）

前述の３つのＶと５つのＶの他に、IBM が提唱していたビッグデータの特徴は４種類４Ｖと呼ばれています。これにはこの Variability が入っていませんが、この Variability もビッグデータの重要な特徴です。良く日本語では可変などと訳されますが、英語から感じ取られるそのまま意味を考えると、可変という訳では無く、データがだんだん変化して行ってしまう可能性のばらつきと言う意味も持っていると思います。

一つのデータもたくさん集まれば別の特徴が出てきます。取得した時期は同じでもデータの母集団の数に

よってデータが変わって行ってしまいます。選挙速報で見られる当選予測なども分かりやすい例です。出口調査と言って投票所の出口でランダムに投票した候補者の名前を聞き取りそこから特別な計算式を用い、さらに地域や人口密度などなど多くの条件を付加して高度な計算を行い結果的に有権者のほんの１ﾊﾟｰｾﾝﾄ足らずの情報であっても当選の可能性が高いと判断します。また出口調査よりもより高度な方法としては開票速報があります。これもほんの少しの開票結果から全体を推測すると言う高度なビッグデータ統計解析の例と思います。選挙に出馬している候補者のみならず所属する党も同様の分析を行い次の選挙へのビッグデータとして利用している事も広く知られています。典型的な変わって行く可能性のあるデータですね。

データが時間の経過によって変わって行く事も世の中に多く存在します。取得した時期が古くなれは新しい情報と一緒に比較はできなくなります。逆にデータそのものは正しくても周りの環境が変化する事でそのデータの位置づけが代わる事もあります。ある地域の対象消費者データが時代と共に変わって行き製品への需要動向が変化する事は容易に想像できます。ここで分かりやすい例としてはある地域の消費者の好みの情報を利用するケースが当てはまるでしょう。消費者の好みと情報を大規模に綿密に取っていても、時間が経てば対象の消費者の年齢も上がります。子供が生まれたり、転入転出があったり住民情報の基本はどんどん変わって行きます。

消費者の情報がほとんど変わらないような短期間であっても、近くで町が統合されたり駅が出来たり災害があったり環境が変化する事によって消費者の好みが急激に変わってしまう事もあります。

企業側にも同様な事象が起こりえます。原材料や流通ルートが急に変化してしまいこれまでと同じ内容のサービスが続けられなる事があります。例えばその地域の所得と需要に合わせた食品を提供していたレストランチェーンが野菜の高騰でターゲット層に合せたメニューを提供できなくなる場合などがその例です。

せっかく取得した消費者に関するビッグデータをうまく業務に結び付けられなくなってしまうデータの変化です。この場合は原材料価格を決めるビッグデータや広域に料理の提供価格を決定するためのビッグデータにも大きな影響が出ている事になります。

データ自身やその位置づけが変わっていってしまう性格をビジネスとしてしっかり見極められないとビッグデータの適正な利用が出来ていないという事になるわけです。

5）Veracity（真実性：信用できるか？）

初期のコンピューターシステムはデータの発生源も入力者も限られていました。もちろん入力ミスはありましたが受注データなどを意図的に捜査して入力したりする事はありませんでした。いい加減なデータ、信用できないデータは基本的に全く入力されませんでした。

現在のデータ入力の目的は、データを入力する事と言う考え方と同時に、データを集めると言う見方もします。別の観点ですと、そこにあるデータを取って来ると言う行為もその大きな目的になっています。

それまで入力がすべての発生源であったデータに集めて来た、取って来た、入れてみた、書いてみた、描いてみた、消した、変えたなどの七色の性格と定義を持った物が増えている訳です。こうした現在の情報はそもそもの目的がデータ入力用でない事や、正しいデータ入力の入り口が業務用やFacebook、オンラインショッピングとバラエティに富んでおり、当然その正確性や責任については全く保証される物ではありません。データの信頼性や真実性に多くの疑問が残ったままのケースも多く存在します。これまで説いてきました様にデータが一つのルールと目的で集められて並べられたものだとすると、ビッグデータは様々多種多様のデータであると言えます。例えば個人の住所が記録された大量のデータと複数駅の改札の利用者数、さらには一つの駅の利用者データについて住所や利用駅などで共通性を探し出し、住民と乗降駅の関係を作った場合でも、別のデータが原因でその正確性が無くなってしまう事もあります。この例などでは異常気象や駅構内工事などの理由により通勤通学者の行動が変われば大幅にその相関関係が狂います。他にも見込まれたスタジアムの入場者が台風で大幅に変わった時もそうですね。そもそも特定日のスタジアムの入場者をビッグデータをたくさん使って予測したデータ自身が天候の変化に加え電車の事故が加わった原因で全く使えな

くなる事もあります。入場者の予想数を元にスタジアムや飲食店交通などのビジネスを組んだ人たちから見ると、これはとても信頼できないデータという評価になってしまいます。SNS などを筆頭にデータに嘘を承知で入力する人もそれを是とする仕組みも数多く存在しますので、真実性とはビッグデータの大変大きな難しさの要素であると言う事が出来ます。

ビッグデータを使って都市を考えるなどと言いますが実際はビッグデータを賢く集めることで賢いシミュレーションや分析が行われ、考えるチームが最適な方法論を考えて初めて戦略が生まれるのです。つまり考える人の為にビッグデータは容易されるという事です。

スマートシティにおけるビッグデータとはビッグデータをいかに賢く集めるかという技術と、いかにそれらの技術を基に高度な分析や考察を行い最良な判断を下すという作業の流れを言うのです。

決して“大量のデータ“ではない訳です

ビッグデータの実際の利用

ここまで日本に入ってくる外来ビジネス用語についてその本質を解説してきました。ここからはいよいよ本当のビッグデータの使い方を考えていきましょう。

例えば、ある製品を拡販したいと言うときにビッグデータを使って営業現場の効率を上げたり生産現場の改善をしたりと言う戦略はあまり馴染みません。営業部門自身がどのような市場でどう売れているか見ながら戦略を作り活動を高品質な物にする事があります。これらをビッグデータの利用による市場開拓や営業戦略の作成などと言う人も多いのですが、これは単純なデータの利用であってビッグデータの利用とは言えません。

ビッグデータは前述の様に、様々な性格を持ったデータです。営業マンの効率や顧客開拓の方針策定に使われる例えば地域別売上推移とか営業マン別の代金回収比率、そのほかの何とか別何々比率や経緯などのデータのほとんどはビッグデータの活用とは言えないでしょう。上記の営業マン別、時期別などの情報のほとんどは既存の情報システムに基礎データとして存在するはずです。それらは単純なプログラミングによる既存データ分析行為です。

営業関連で言うならばビッグデータを活用するのはたぶんもう少し上層部であったり、徹底した管理を目指す部門などでないかと考えられます。

自社で持っている情報に、数千と言うお客様の会社の株や投資の推移の情報をぶつけ、かつ社会的情報例えばその商品が若いファミリー向けなら待機児童の都市別推移や部品の調達に関連するプラスチック原材料と為替情報を合わせた情報と、営業マンの過去の成績データからわかるその実力データを掛け合わせて今後の予測と人的リソースの配置を見ます。このような事が行われるのであれば、十分にビッグデータの利用と言える訳です。ビッグデータは大量データだけでは無いのです。たとえ数兆件のデータであってもそのデータが既存の社内データベースに存在するなら、または社内システムから集められるのであればビッグデータではありません。

この意見に反論のある方も多いとは思いますが、最も重要な視点はこれらの利用法やビジネスでの戦略利用が日本以外では活発に行われていて、日本ではほとんど行われていないと言う事実なのです。重要な人物の発言でニューヨークやロンドンの株価はすぐに上下します。ニュースでは誰々の景気予測見通しの難しさ発言で、株価が下がったなどとよく聞きます。

このニュースからは多くの投資家が画面を見ながら自分で判断して大量の株の売り買いをしている様に見えます。ところがアメリカばかりでなく多くの国の株式投資家は実は人間ではありません。そのほとんどがIT のコンピューターシステムなのです。もちろんコンピューターが勝手に判断して株を売買しているわけではありませんが、人間が指定した情報の収集と分析方法に基づいて、人間が指定した売買方法に基づいて売り買いが行われているのです。

天候で穀物の相場は変わりますが、小麦一つにしても世界中のピンポイントの農場の天候を見て、小麦の出来を推測します。小麦の出来は当然製パン業のビジネスには直接的な影響が出ます。こういった日々変わる情報を集めて来たり、整理して分析したものを人間が判断材料として使うわけです。製パン行の株を持っている人は世界各地の天候の変化が気になる訳です。それらの途轍もない天候の情報を集められたとしても、その情報を理解して判断する事は人間にはできません。

世界各国100ヵ所小麦主要農場の温度推移が毎分報告されても人間には何の判断もできません。そこにはこれらの情報を人間にわかるように加工するという自動処理も含まれます。自動処理と言ってもその処理の方法は人間が指定するのです。簡単な例ですが温度の差が前日の同じ時間よりも５度以上変わった地域について報告する。と言った指示をして株式投資家は条件を設定します。ビッグデータですから小麦農場の天気ばかりではなく他の株や為替や物流に影響する政変などの情報も広く収集します。

このように、判断基準は人間が作っているのでニュースでは誰かの発言で株価が動いた様に言いますが実際の高速処理は人間の設定した複雑な条件を守ったシステムが自動で株の売買を行っているのです。

ビッグデータがこのように運用されて株価や為替相場はあたかも自動ロボットで動いているわけです。株の動きは、現場の営業にとってはそれほど神経をとがらせる程の情報ではありませんが、彼らを統括し営業の人数や投資を決めている部門からすると大変重要な情報なのです。営業統括部門は彼らのお客様がどういう状況にあるか今後顧客の業界がどこに進むかどのぐらい利益を生むのかを予想するために顧客の株価情報や顧客業界全体の情報は非常に重要な情報なのです。小麦の相場が製パン業を心配させ、製パン業の株価が製パン業を担当する企業の営業戦略策定に影響を出すわけです。

一方ではその製パン企業自身にとっては自社の株の動きですから重要です。自社の株がこうしたロボットで操作されている事を知っている会社はさらにその株価や資金調達や為替の情報や物流情報、各種相場情報も必要になります。

そこでまたビッグデータと呼ばれる多岐に渡る情報が欲しくなるのです。

こうしてビッグデータの探索と利用はどんどん広がっているのです。

しかし！　繰り替えしになりますが、このような連鎖するビッグデータ利用はあまり日本では積極的に行われていません。最も大きな理由がビッグデータは大量データであるという翻訳の誤解です。日本の企業経営者や役員のほとんどは大量のデータを使うべきなんだと誤解しデータの準備や提出を情報システム部門やリサーチ会社に“ビッグデータを活用せよ”と指示していますが、全くその活用方法は欧米とは別なのです。その間にアメリカやイギリス、ドイツ、オーストリアなどではビッグデータを含むビジネスツールの戦略的な利用が日に日に改善され拡大され、ビジネス戦略に大きな影響を出しているのです。

もちろん製パン業だけでは無く、ホテルやレストラン業、教育産業、交通などありとあらゆる業種で争った利用競争が行われているのです。それらの進化の理由は非常に簡単で“こう見てみたい、この観点で比べてみたい”と言う経営者の好奇心なのです。経営者の好奇心は企業の中核となる仮説です。仮説がビッグデータの利用でどんどん実証され、またはどんどん否定されれば経営者はどんどん賢くなりその次の手の正確さは上がっていくわけです。

さてその海外の経営者と“ビッグデータは大量データ”と言い続けている日本の経営者が今同じ国際舞台で戦っているわけです。丸腰とは言いませんが同じ武器を持って互角に戦っているかと言うといささかの疑問を持たざるを得ません。

おわりに

お分かりになったと思いますがビッグデータは決してIT用語でもなく、大量データでもなく、さらに現場の改善システムでもないのです。ビッグデータとは企業戦略の根幹をなす仮説検証の巨大なツールですから、現在の利用実態や今後の利用方法予測は業界自身の現在と未来の実態把握とビジネスの仮説そのものになる訳です。事業の大小や職種にかかわらず、是非発想の変換をされて強い戦略を瞬時に建てられるようになる事を願っています。また、ITや統計関連の研究者や企業でソリューション製品を開発している技術者も早くこのビッグとラージの違いを正確に理解し、日本の発展にもっと役立てて欲しいと願っています。