【ＤＲニュース・025】：ビッグデータの活用事例・統計学とデータ解析の違い＆ＡＩ関連

2016年10月11日発信

昨今話題のビッグデータの活用事例を通して、どんな観点からビッグデータを解析したら良いのか？
また、ビッグデータが集まるプラットフォームで何が出来るのか？　どんなスキルが必要となるのか？

急速な技術革新により、可能となるビッグデータの解析・分析について、探求して、考えてみよう。

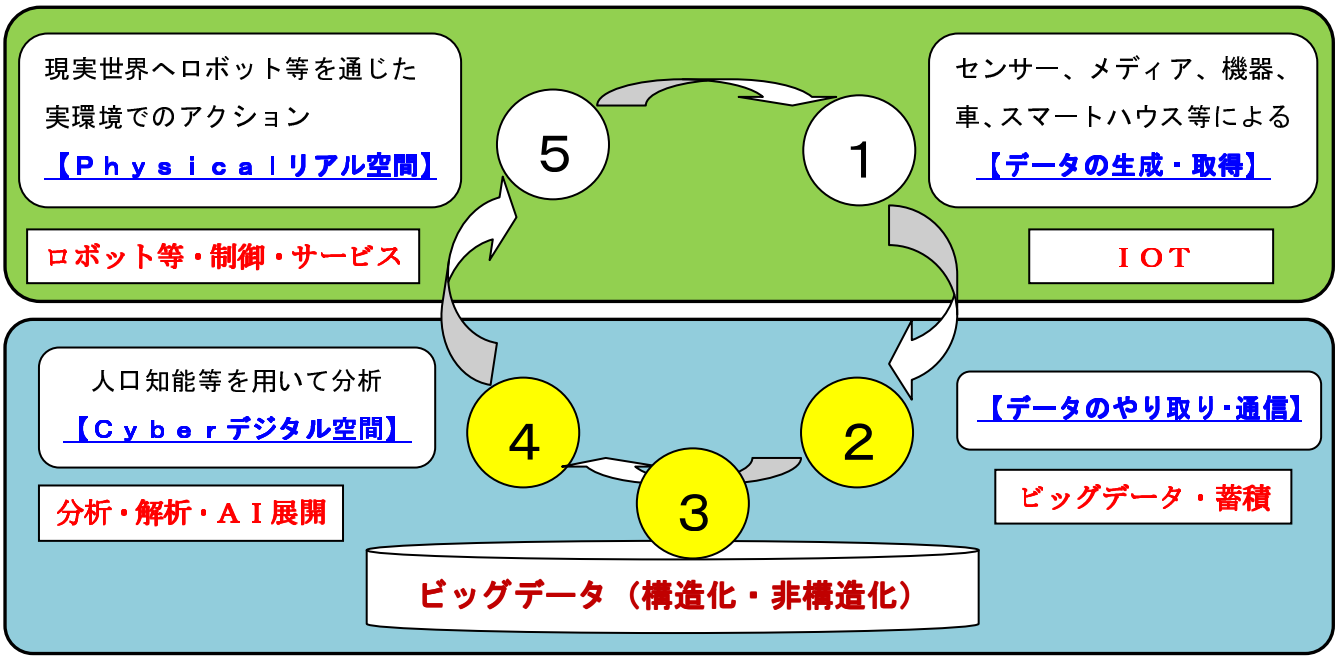
1.　ビッグデータとは

ビッグデータとは、一般的には「**３つのＶ**」で、その特徴を説明することが多く、具体的には、「**Ｖｏｌｕｍｅ（多量性）**」、「**Ｖａｒｉｅｔｙ（多様性）**」、「**Ｖｅｌｏｃｉｔｙ（流動性）**」の特徴を持ったデータのことを指します。その他に**Value（価値）**をつけた**4V**で定義されることもあります。

多量性はデータの総量、多様性はデータの種類、流動性はデータが生成されるスピードを示します。

近年、スマートフォンやタブレット、ＳＮＳなどのソーシャル・メディア、Ｍ２Ｍ通信の普及に伴い世界中で生成・蓄積されているデータ量は急増しています。また、扱われるデータの種類は、従来からの販売や在庫などに関する数値や文字列のデータ（**構造化データ**）のみならず、ツイッターなどの代表されるテキストデータ、センサーやカメラから得られる位置情報やセンサーデータ、音声、動画、クリックストリーム（ウェブサイト内で訪問者がどのページを歩き回ったのかを示す履歴）などのデータ（**非構造化データ**）が増加しており、**ビッグデータの「８割」が非構造化データが占めています**。

● **急速な技術革新により、大量データの所得、分析、実行の循環が可能に**



2. ビッグデータの活用事例

これまで、大量データの収集、分析などは、時間と費用の面で限界がありました。しかし、ＩＣＴ（情報通信技術）の進歩によって、非構造化データのデータベース管理システムの実用化、大量データの「**分散・リアルタイム処理手法**」の普及などが進むことで、大量で多種多用なデータの生成・収集・分析が可能となりました。これに伴い、様々な業界でビッグデータを活用し、新たなビジネスを創り出そうとする流れが生まれています。

（1） 震災・ビッグデータ

2011 年 3 月に発生した東日本大震災では、当時記録された携帯電話やカーナビの位置情報、ツイッターなどの震災に関する膨大な電子データ（いわゆる“震災ビッグデータ”。）が残されており、これらを活用することで、震災の全貌を明らかにしようとする動きがあります。

地震発生後の人々の避難行動、被災地における復興の遅れや帰宅困難者を生んだ首都圏での混乱の要因が明らかになるなど、産官学が連携し震災ビッグデータを防災などに活かす取り組みが進められています。

震災ビッグデータによって、地震発生後に一旦避難した被災者が再び家族や知り合いを心配して浸水地域に戻り、命を落としたことなどが明らかになりました。

（2） 小売・ビッグデータ

① **リコメンド・・・**利用者の属性や行動・購買履歴データを基に、最適な商品を推奨する手法。（リコメンド機能とは；通販サイトでは、「この商品を買った人はこんな商品も買っています。」といった表示（レコメンド）をよく目にします。これは、“レコメンド機能”と呼ばれるもので、ウェブサイトへのアクセス履歴などの膨大な情報を取集し、利用者と類似した商品やカテゴリに関心を持っている他の利用者を関連づけ、グループ化し、類似した利用者がよく見ているが、利用者がまだ見ていない商品を表示させる仕組みです）

・・・このため、レコメンド機能は、利用者が今まで存在を知らなかった商品に気づくきっかけとなり、利用者の潜在的なニーズの引き出しに貢献しています。

② **行動ターゲティング広告・・・**ウェブサイトの閲覧履歴や EC サイト（商品やサービスをウェブサイト上で販売するサイト）上での購買分析などを蓄積し、利用者の趣味・嗜好を分析した上で、利用者ごとにインターネット広告を出し分けるサービス。

③ **顧客離反分析・・・**携帯電話会社や通販会社、保険会社、レンタル DVD 会社など、契約による会員制で商品やサービスを提供する会社では、過去の顧客データや退会データに基づき、サービスを退会しそうな顧客（離反）を予測。

（3） 交通・ビッグデータ

① **渋滞予測・・・**実際に自動車が走行した位置や車速などの情報から、道路の渋滞情報などの道路交通情報を提供。

② **テレマティクスサービス・・・**通信機器やGPS機能を備えた車載機を車両に搭載し、走行距離や走行時間、運行速度、位置情報などの運行状況をリアルタイムに取集・分析することで、車両管理や安全運転支援、エコドライブなどのサービスを提供。

（4） 農業・ビッグデータ

① **農業ＩＣＴ・・・**ビニールハウス内に設置したセンサーがM2M通信機器を通じて、温度や湿度、日照時間、二酸化炭素など作物の育成環境のデータをモニタリング。農家はスマートフォンやタブレット端末でデータを随時参照して、作物の品質確保と生産業務の効率化に活用。

（5） 製造・ビッグデータ

① **故障予測・・・**コピー機や複合機などのハードウェア機器に取り付けられた各種センサーから、紙詰まりなどのエラー情報や機器の利用履歴、消耗品の劣化状態などのデータを取集・分析することで、故障などのトラブルの予兆を検出。

（6） 金融・ビッグデータ

① **不正検出・・・**クレジットカードの膨大な利用履歴データを分析することで得られる顧客ごとに不正利用を示唆するパターンから、オンラインでの不正モニタリングなどの可否判定を行う。

② **株式市場の予測・・・**ツイッター上のツイートを解析することで、市場の動向を予測。

（7） 健康・ビッグデータ

① **医療/風邪の流行予測・・・**ツイッター上の風邪に関するツイートを検出し、風邪をひいている可能性が高いユーザーを集計。そして、風邪に関するツイートの増減と気温や湿度などの気象データとの関連性を分析することで、週間天気予報と組み合わせて風邪の流行を予測。

（8） セキュリティ・ビッグデータ

① **不審者監視サービス・・・**カメラ映像からリアルタイムに人物の顔を検出し、その特徴を自動的にデータベースに登録することで、不審者を検索するサービス。

（９） 【2016 年度版】ビッグデータ活用事例　　（売上げ向上・コスト削減方法とは）

① **ダイドウドリンコ・・・アイトラッキング分析と購買データの組み合わせで売上げが前年比１．２%増**（ダイドードリンコは“Z の法則”に従い、主力シリーズ「ブレンドシリーズ」を左上に配置していました。しかし、自動販売機にアイ【eye:目】トラッキングを取り付けて調査したところ、自動販売機に限っては下段に視線が集まることが分かった）

② **TRUE&CO・・・自分の体ににあったブラをオンライン購入できるシステムを開発**
（過去の顧客の注文と返品データを分析することで、メーカーによるサイズのばらつきなどを数値化し、オンラインで、自分の体にフィットするブラジャーを購入できるシステムを開発しました）

③ **スシロー・・・皿に IC タグをとりつけ、レーンに流れる寿司の鮮度や売上状況を管理し売上向上**（どの店で、いつどんな寿司がレーンに流されいつ食べられたのか、どのテーブルでいつどんな商品が注文されたのかなどのデータを毎年 10 億件以上蓄積することで、需要を予測し、レーンに流すネタや量をコントロールしています）

④ **ＧＥＯ・ゲオ・・・王道的なビッグデータのクラスタリングでテコ入れ**（会員向けアプリをリニューアルすることでビッグデータを取得し、他社のネット通販や VOD（ビデオ・オン・デマンド）などの攻勢に立ち向かっています。具体的なデータの利用方法としては、会員を「趣味別」及び「売上貢献別」にクラスタリングすることで、趣味に応じたクーポンの発行やメールを送付し売上の向上を測ったり、新作 DVD の仕入れを最適化しています）

⑤ **ローソン・・・売上 31 位のほろにがショコラブランを売り続ける理由**（ポンタカードの導入により、ビッグデータの分析が進んでいます。分析の結果、例えばほろにがショコラブランが「1 割のヘビーユーザーが 6 割の売り上げを占めている」と分かりました。その分析結果をもとに、リピート率の高いほろにがショコラブランは、今も継続的に販売されています）

⑥ **大阪ガス・・・コールセンターの依頼内容から修理に必要な部品を割り出す**
（過去数百万件にわたる修理履歴や機器の型番データを保有しています。また、コールセンターに寄せられる給湯器などの修理依頼の内容も同時に蓄積しています。これらの情報を組み合わせることで、ケースごとに必要となる部品を自動的に割り出すことに成功しました）

⑦ **城崎温泉・・・観光客のニーズをつかみ売上増**（携帯電話やスマートフォンをお財布代わりに使えるシステムを導入することで、観光客の利用履歴を蓄積し、定量的な分析を行いました。何時頃に観光客が多いか、人の組み合わせは親子連れが多いのか、男女ペアが多いのか、また、どこの外湯が一番人気なのかなどを分析することで、より効果の高い施策を実施したり、温泉街の街づくりやサービス、広報の方法などの改善につながりました）

⑧ **株式会社開園システム・・・タクシー乗務員用アプリで機会獲得＆業務効率化**

（タクシー業界向けにビッグデータを活用したアプリを提供しています。GPS で蓄積された過去の乗車位置を、月・曜日・時間帯別に地図上に表示したり、リアルタイムの実写位置を表示したりすることで、どこで顧客が増えているのかを把握します）

⑨ **カルフォルニア州オークランド・・・犯罪データを蓄積して、未然に予防**

（全米屈指の犯罪都市で、観光客や、市民が犯罪に巻き込まれないための注意を呼びかけるシステムにビッグデータを活用しています。犯罪の種類別（殺人・強盗・から泥酔まで様々）、日にち別、時間別に、フィルタが可能で色別に確認することができ、一見してその日その時に危険な場所を特定できるため、利用者は危険を避けることが可能となります）

⑩ **楽天・・・レコメンドだけでなくランキングの更新頻度とジャンルの細分化で売上向上**

（楽天は更新頻度の短縮と、ジャンルの細分化を試みて大きな成果をあげました。これはビッグデータを分析することで、ランキング頻度が高いほど売上は増加し、ジャンルが細かいほど全体の売上があがるという結果に基づいた改善施策です）

⑪ **ホームセンター・・・従業員の配置を調整して売り上げ 15%アップ**

（売り上げデータと従業員の行動データや、商品の陳列データを蓄積したところ、顧客単価の高いスポットの特定に成功しました。そしてそのスポットに従業員を重点配備したところ、売り上げが 15%もアップしたという、まさに予測通りの結果となったそうです）

⑫ **コールセンター・・・休憩中のスタッフ同士の雑談を増やして売り上げ 27%アップ**

（受注率の異なるコールセンターのスタッフにセンサーを取り付けてデータを検証したところ、受注率の高いコールセンターのスタッフの方が低いコールセンターのスタッフよりも休憩中の活動が活発だということが判明しました。また休憩中にスーパーバイザーがスタッフに声をかけていくとスタッフの雑談が盛り上がるということもわかりました）

⑬ **ヤクルト社・・・自社商品による顧客の奪い合いを解消して売り上げ 20%増加**

（ヤクルト社の商品は 1 つのカテゴリに 150 点も存在し、店頭で顧客を奪い合っていました。またその組み合わせを分析し最適化しようにも、俗人的に作成されたスプレッドシートが社内に分散していました。そこでそれらのデータを一元管理し分析したところ、15 本パックと 7 本パックは購入する顧客層が異なるため、並べて販売すると両方の売り上げが増加するということを発見しました）

⑭ アンデルセン・・・データから製造量を決定し売り上げ増加

（広島県でパンの製造・販売事業者であるアンデルセン社では元々、各店舗の店長が自身の経験から製造量を決定していました。そこでアンデルセン社は販売履歴と来店客数を関連付けて分析し、商品の売れ行きパターンを予測しました）

⑮ 石川県羽咋市・・・農業で人工衛星の画像データを活用して収益アップ

（地場の民間企業とともに人工衛星の画像データから米の味を計るシステムを開発しました。そして画像からタンパク質の含有量を測定することで、一般的に美味いと言われる量のタンパク質を含有した米を安定して収穫できるようになり、低タンパク米をブランド化して販売したところ、収益も増加したとのことです）

⑯ おまけ；データを扱う上での注意点

◆ **定量だけでなく定性的なデータも見る必要がある**（データを活用する上で定性的なデータも非常に重要な役割を担います。数字だけを分析していても結局何が起こっているのかを正確に把握することは困難。定性的なデータを調べるには行動観察が効果的です）

◆ **相関関係ではなく因果関係が重要**（さまざまなデータの変動からその変動の原因を探る際、複数の事象の「相関関係」を探るのではなく、「因果関係」を見出すことが重要です。一見相関性があるデータも擬似相関である可能性があるので、目的と仮説をもって検証を行い、因果関係を見出しましょう）

◆ **限られたビッグデータには限りがある**（商業ビル周りだけのビッグデータで出来る洞察には限りがあり、都市計画、商流、物流までも解析することで、初めて意味のある洞察が出来るようになります）

ビッグデータは、様々な業界で売上増、コスト削減、業務効率化などの目的のために活用されています。使い方次第で絶大な効果を発揮するビッグデータですが、数字ばかり見ていると一見相関性があるように見える擬似相関などに騙されてしまう可能性があるので、分析には十分な注意が必要です。

様々なビッグデータを活用する際には、目的や仮説を持ち（擬似相関に騙されることなく）因果関係を見極め、成果につながる施策に繋げて行きましょう。

（１０） ビッグデータの活用範囲と解析・分析

ビッグデータをどのように活用し、どのような価値を生み出すかは、まさにアイデアやノウハウまたは事業戦略そのものであり、企業が今後考えていかなければならない課題だと思います。
例に挙げた活用方法以外にも、表にあげたような活用はすでに考えられています。

これらの活用例には、既存のシステムで行われてきたものもありますが、

これら従来のシステムとビッグデータを活用したシステムとの差は、**使用するデータの種類にあると言えます。**

「ビッグデータ」が単なるバズワード（定義や意味が曖昧な用語）になるのか、それともあなたの企業の武器となるかは、データ活用をどのように考えるかということにかかっていると思います。つまり、**データを資産として考え、分析という活用を行うか否かということです。**

「ビッグデータ」の活用には、必ずしも決まった形や１つの正解があるわけではありません。
ただ、経営環境が目まぐるしく変わる現在、これまでと同じようなデータの使い方や意思決定のスタイルを継続していて良いのかということは、考える必要があることは確かだと思います。

そこから、あなたの「ビッグデータ」への分析の道は始まります。

マーケティングに「**知見**」を活かすために、従来のように「**勘と経験と度胸**」によって経営戦略やマーケティング戦略の立案、新商品開発などの意思決定を実施するのではなく、
分析とは、何かしらの目的の元で収集されたデータに対し、統計などさまざまなツールを用いて**知見**を得ていく作業ですが、雑多に集められたデータをいくら高度な手法で分析しても、有用な**知見**はその中のごくわずかであることが多いものです。

また、企業の分析担当者は日々、データ分析にたどり着く前の「前処理」に多くの時間を割かれ、肝心の分析内容の精査やレポーティング、そこからのアクションを検討する時間を取れていないのが現状ではないでしょうか?

そんな中で、ビッグデータの解析・分析について、いろいろな観点から調査して見ました。

3. 大量データの分散処理技術　　【分散処理技術の進歩が、データ解析に欠かせない】

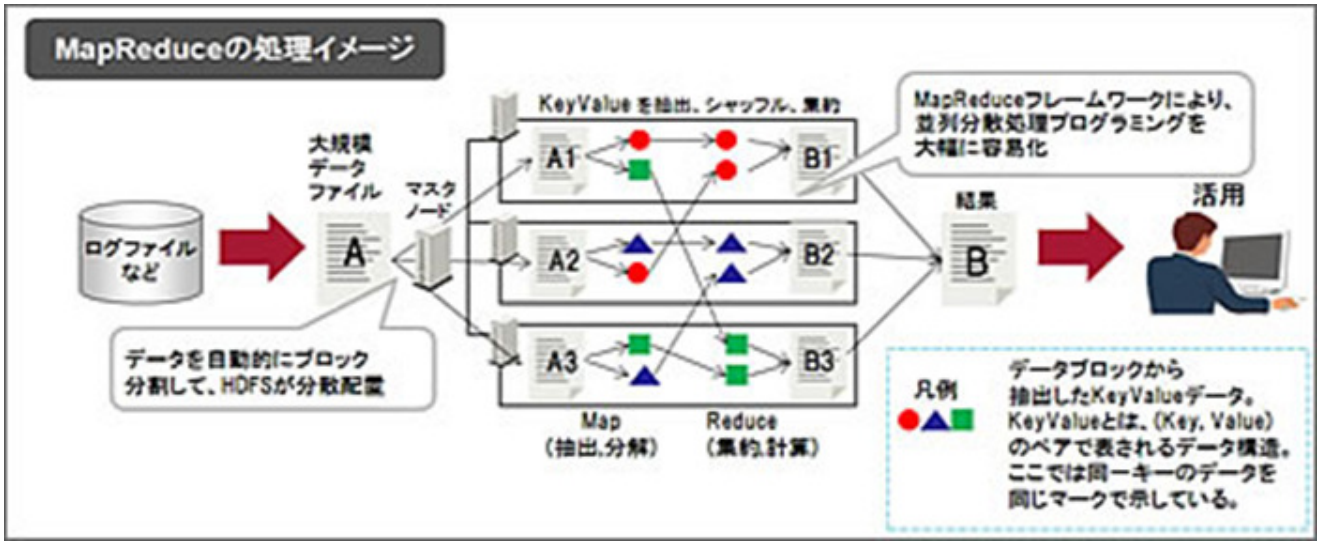
近年、高速なインターネット回線の普及や各種センサー技術の進展などで、企業や社会インフラにおけるデータが爆発的に増大しており、大量データを短時間/リアルタイムに処理することで、新しい価値が生まれています。

そのための技術として、Hadoop などの**並列分散処理基盤**や、**バッチジョブ分散処理**、**ストリームデータ処理技術**に注目が集まる **大量データ分散処理技術の特徴と適用分野**を探って見ましょう。

（1） Ｈａｄｏｏｐ

Hadoop は、**OSS (Open Source Software)** の並列分散処理基盤および分散ファイルシステムから構成されます。ソフトウェアフレームワークで **JAVA の言語**で書かれています。

- オープンソースソフトウェアコミュニティ Apache Software Foundation にて開発・公開されている、大量データを効率的に分散処理するためのソフトウェアの基盤です。
- サーバを大量に並べ、並列処理を行うことにより、従来 **RDBMS** で扱うことが難しかった大量データのバッチ処理を高速化することが可能となる。
- 主なソフトウェアコンポーネント
  - **MapReduce** ： 長時間かかる処理を複数のマシンに分散させるフレームワーク。
  - **HDFS** (Hadoop Distributed File System) ： 複数のサーバの **HDD** (hard disk drive) を 1 つの巨大なボリュームに見せる分散ファイルシステム。



※【Hadoop の使い方まとめ（2016 年 5 月版）から抜粋；(Hadoop はバッチ処理だけではない)

Hadoop が登場して 10 年が経ち、その間に Hadoop とそのエコシステムも誰も予想できない程、大きく進化してきた。当初バッチ処理専用と言われていた Hadoop も、**今や SQL エンジンや機械学習など様々なアプリケーションを動作させることができる汎用基盤となっている**。

（２） バッチジョブ分散処理

バッチジョブを分散して１つまたは複数のサーバ内で並列実行することにより、バッチジョブ全体の処理時間を短縮して高速化するための機能を提供します。

（３） ストリームデータ処理

リアルタイムなデータの活用を行い、インメモリ処理と差分計算処理によって、大量データを高速処理、SQL ライクなスクリプト言語(CQL)で分析シナリオを記述可能なため開発が容易。

「**ストリームデータ処理技術**」は、大量発生する実データを逐次(ちくじ)に時系列処理(じけいれつしょり)する技術です。データ発生時に、あらかじめ登録したシナリオにしたがって集計・分析に必要なデータを抽出し、データ処理を行います。その際、分析対象データをメモリー上で処理する「**インメモリデータ処理技術**」により、高速なデータ処理を実現しています。これらの技術によって、大量データを高速に、かつリアルタイムに処理できます。

大量データを活用するためには、新技術への理解を深め、どの新技術を適用すべきか適用効果の検証などをふまえ判断し、適切なシステムの導入形態を見極めることが重要なポイントとなります。

4. 統計学とは

データとは「何らかの目的のために取得されたまとまった数値や符号の集合体」ですが、それらの集合体を漠然と見ても、そこからは何も得ることはできません。
データの数を数えたり、平均を出したり、傾向を見たり、分類をしたりと、何らかの手を加えることによって、初めてデータの性質や意味を知ることができ、活用することができるのです。

（１） 統計学(とうけいがく)の基礎(きそ)

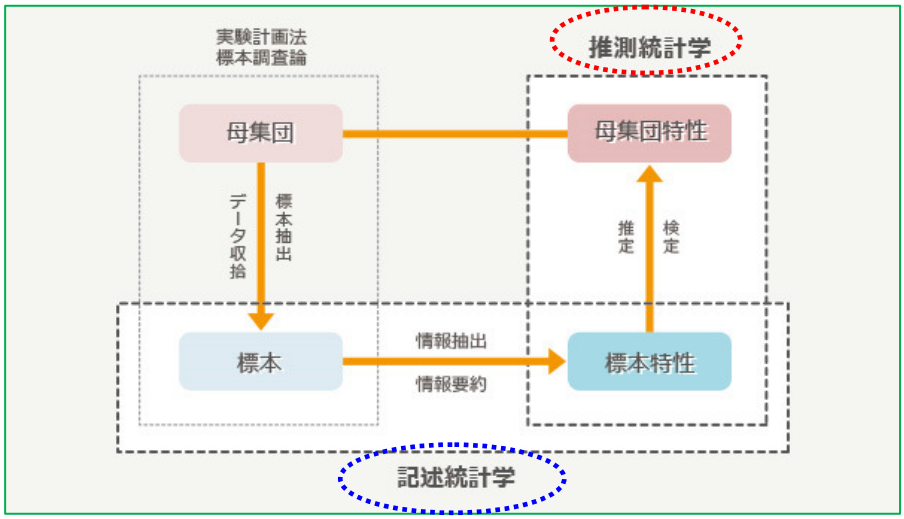
ある程度の数のデータには、必ずバラツキ（不確実性(ふかくじつせい)）が伴います。もし、ある学校のテストの点数が全員同じであったら、平均点や順位、偏差値(へんさち)を出すことに全く意味はありません。一年中天気や気温が一定であったとしたら、天気予報は要らないし、気温をグラフに描く必要もないのです。しかし、実際には、学年やクラスによって点数は異なりますし、地域や日時によって天気も気温もばらつきます。それゆえ、クラス別の平均点や気温のグラフなどを描いて、**クラスの特性を把握したり、明日の気温の予測をしたりします**。

統計学とは、ある程度以上の数の**バラツキのあるデータの性質を調べたり**、大きなデータ（母集団）から一部を抜き取って、**その抜き取ったデータ（標本(ひょうほん)）の性質を調べることで、・・・・・元の大きなデータの性質を推測したりするための方法論を体系化したものです**。

（２） 統計学の体系

統計学には、大きく分けて「**２種類**」あります。

- あるデータを集めて、表やグラフを作り、平均や傾向を見ることでデータの特徴を把握するという統計を「**記述統計**」といいます。
- 一方、母集団からサンプルを抜き取って、そのサンプルの特性から母集団の特性を推測し、それが正しいかどうかを検定する統計を「**推測統計**」といいます。



（３） 統計学とデータ分析

ビッグデータの登場で統計学が注目を集めている。理由は、統計学を駆使してビッグデータを分析することで、経営戦略やマーケティング戦略の立案、新商品・新サービスの開発などで大きな成果が得られることがわかってきたからです

データ分析は、何らかの目的を持って行なわれます。従って、分析を始めるにあたっては、出てきた結果が目的できなものであるかどうかの正しい判断が求められます。
そのために以下の**“3 つを理解”**をしておく必要があります。

① **分析しようとする問題そのものについての理解**・・・分析者は、なぜ分析をするかという理由やその問題点について、正しい現状認識をしていなくては意味がありません。

② **分析手法についての理解**・・・統計ソフトやデータマイニングのソフトを用いれば、どんなデータを入れてもそれなりの結果がクリックするだけで出ます。
しかし、正しく分析手法を理解していないと、明らかに間違ったデータ処理をしていたり、相応しくない手法で分析をしたりしていても、その誤りに気づきません。

③ **分析結果に対する判断力**・・・分析結果が予想通りであればアクションを大きく変更する必要はありません。しかし、予想外の結果が出た場合は、それに対してどのようなアクションを取るべきかの判断が非常に重要になります。

データの取得方法や処理方法が間違っていたのかもしれないし、分析手法が間違っていたのかもしれない。そもそもの仮説が間違っていたということになり、そこでは方向転換を余儀なくされることもあるでしょう。
想定外の結果が出たとき、柔軟に頭を働かせて様々な可能性を考えるべきだと思います。

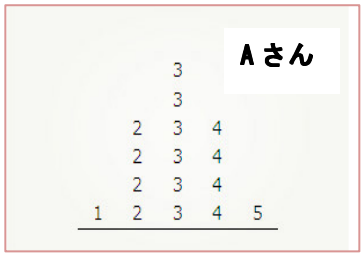
(3-1) 基本統計量（きほんとうけいりょう）

A さんと B さん二人の大学の成績です。これは、このデータを一見しただけでは、それぞれがどのような成績で、**どのような差があるかはよくわかりません**。

	経営戦略	社会心理学	流通論	財務会計論	管理会計論	基礎統計学	計量経済学	国際金融論	多変量解析	不動産基礎	確率論入門	証券化技術	都市計画論	企業法務	著作権	企業倫理
Aさん	2	3	3	3	4	3	4	1	2	2	2	4	4	3	3	5
Bさん	2	2	1	2	4	2	1	2	2	3	3	1	4	5	3	2

① **度数分布図**

二人の成績がどんな分布をしているのか？　チェックする必要があります。



② **基本統計量の算出**

エクセルの分析ツールを利用すれば、右記のような結果が簡単に出力されます。

基本統計量を見ると、平均は B さんのほうが悪く、形の歪みを表す歪度が、左右対称の A さんはゼロなのに対し B さんは 0.77 なので、歪んだ形だということが分かります。

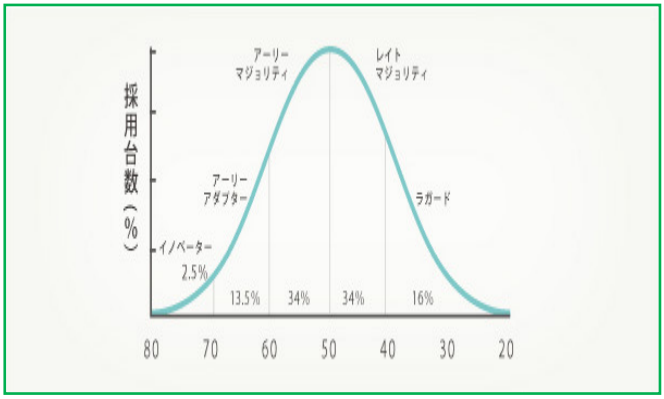
Aさん		Bさん	
平均	3.00	平均	2.44
標準誤差	0.26	標準誤差	0.29
中央値（メジアン）	3.00	中央値（メジアン）	2.00
最頻値（モード）	3.00	最頻値（モード）	2.00
標準偏差	1.03	標準偏差	1.15
分散	1.07	分散	1.33
尖度	-0.21	尖度	0.15
歪度	0.00	歪度	0.77
範囲	4	範囲	4
最小	1	最小	1
最大	5	最大	5
合計	48	合計	39
標本数	16	標本数	16

（3-2）確立分布（正規分布）

なぜ、ここで正規分布の話をするかというと、ビジネスに限らず、我々の生活のなかでも正規分布を利用しているものがたくさんあり、また自然界においても正規分布にあてはまる現象がたくさんあるからです。

① ﾏｰｹｯﾃｨﾝｸﾞにおける正規分布の活用１

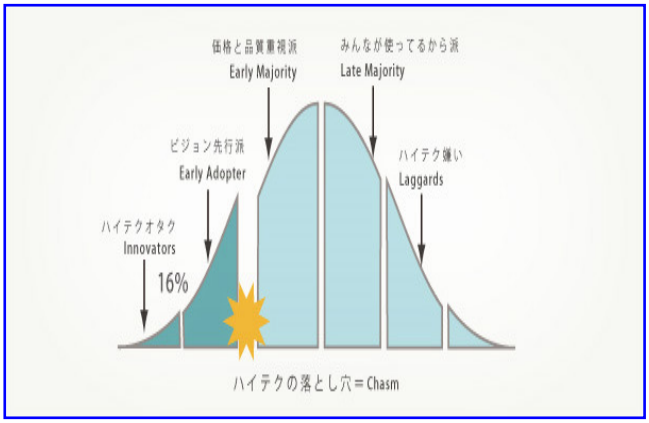
正規分布は様々なビジネスシーンでも使われます・・・「**イノベーター理論**」というのは、商品の普及を説明するもので、消費者商品が発売されてから購入までの特徴を５つのタイプに分類したものです。



この数字は何なのでしょうか。
まさにこれは「**偏差値**」と同じ、イノベーターというのは偏差値でいえば 70 以上ということであり、1000 人いれば 23～25 人くらいの確率で出現すると考えると分かりやすいでしょう。

② ﾏｰｹｯﾃｨﾝｸﾞにおける正規分布の活用２

「**キャムズ理論**」というものがあり、初期市場と主流市場の間には、深い裂け目があり、特にハイテク商品において顕著で、多くの企業がこの裂け目を超えられずに、失敗しているというものです。



つまり、新商品採用に対する**偏差値**が６０までの人には採用されるが、６０未満の人には採用されないという意味なのです。

（3-3）単変量解析（たんへんりょうかいせき）と多変量解析（たへんりょうかいせき）

① 単変量解析

単変量解析は、ひとつの対象にデータが 1 つしかないデータを扱います。
例えば、ある人の通信簿のデータなどです。また、ある科目の成績や平均点の時系列データもデータは 1 つなので、単変量といえるでしょう。
後者は時間というもう一つの指標がありますので、正確には 2 変量なのかもしれませんが、時間の進み方は一定と考えれば単変量として考えてもよいのではないかと考えています。

② 多変量解析

多変量解析とは、多くの情報（変数に関するデータ）を、分析者の仮説に基づいて関連性を明確にする統計的方法のことですが、もっと簡単にいえば、「複雑なことをわかりやすくすること」です。

例えば、ある商品に対して様々な評価や結果があります。　売上高や利益率もそうですが、顧客満足度や商品特性など、その商品に関する評価データは、すべて何らかの原因があって作り上げられるものです。

③ 多変量解析の手順

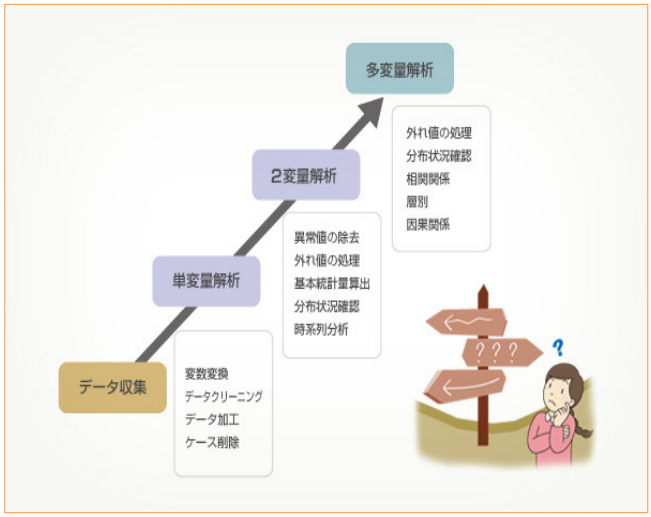
収集したデータは、必要に応じて対数変換、1/0 変換等をすることがあります。

また、ゴミ・ノイズデータがないか等を確認しクリーニングや加工などをして整えます。

その後、**単変量解析**、**2 変量解析を経て**、**多変量解析に進みます**。多変量解析の結果が思わしくない場合、単変量解析に戻って、再度 2 変量解析、多変量解析に進むこともあります。

④ 単変量解析と２変量解析

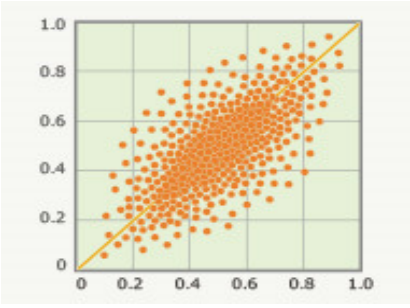
多変量は単変量をたくさん集めたものですから、単変量解析を理解していないとよい結果が出ないことになります。まずは単変量、2 変量解析を充分行なうことが重要です。



では、単変量解析とはどのようなものなのでしょうか。 2 変量というのは、“**身長と体重**”のように、1 つの対象に 2 種類のデータがあることをいいます。

単変量解析は、ひとつの対象にデータが 1 つしかないので、”**身長**”のデータしかないということです。身長のデータしかないということは、クラス全員の身長のデータがあるという場合と、ある人の身長の時系列データがあるという場合があります。

⑤ 相関関係



2 変量では、身長と体重、気温とアイスキャンディーの売り上げなど、2 つの変数を扱います。

身長と体重には関係がありそうですが、同じ身長でも太っている人も痩せている人もいるので、身長がわかったからといって、体重を知ることはできません。

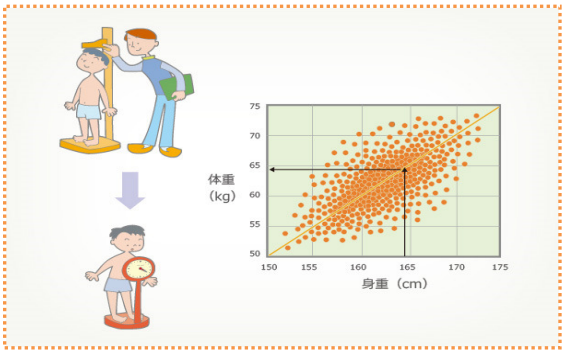
ただ、ある程度のばらつきの中には入っているだろうことは推測できますので、そのばらつきが少ないことを相関が高いといいます。

⑥ 多変量解析の手法

多変量解析を行なう目的としては、大きく分けて「**予測**」と「**要約**」の 2 つがあります。
例えば、広告クリエイティブの最適化は、複数のコンテンツの組み合わせパターンからクリック率を予測するモデルを使っています。購買データから顧客をいくつかのクラスターに分類するには、要約の手法を使っています。

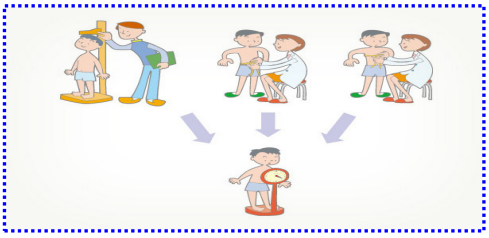
A) 予測の手法

予測の手法は、複数の変数から何らかの結果を予測するものですが、因果関係明確化の手法ともいわれます。どういう原因を作れば、欲しい結果が得られるかということを知るためにも使われます。



原因側のデータを「**説明変数**」、結果側の変数を「**目的変数**」といいます。結果は原因によって決まる、つまり結果は原因に従属しているという意味で、目的変数を「**従属変数**」、また原因は独立しているので、説明変数を「**独立変数**」ということもあります。

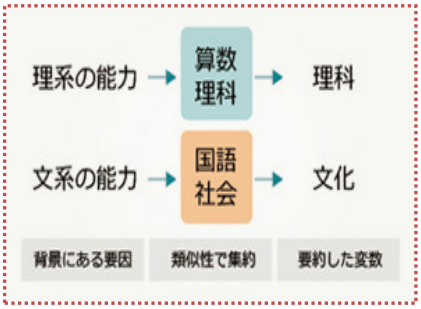
予測の手法で最も簡単なのは、**直線回帰**でしょう。身長から体重を予測するようなものですが、さらにここに腹囲や胸囲のデータが加われば、より精度が上がるでしょう。
さらにその人の食生活や運動量などの変数を加えれば、**体重が何によって決まるかの因果関係も明らかになりますし、予測の精度もさらに上がるわけです。**



B) 要約の手法

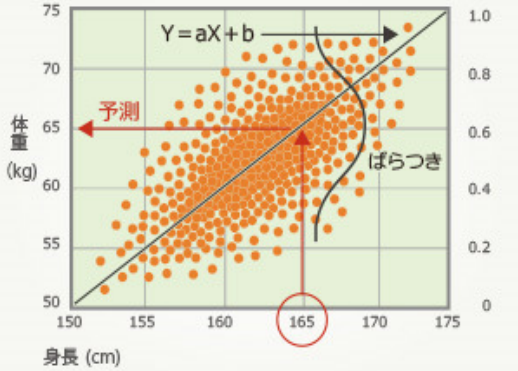
要約の手法は複数の変数を新しい変数に要約する、すなわち多くの変数を少ない変数で説明する手法です。類似関係明確化の手法ともいわれます。

例えば、算数、理科、国語、社会の 4 科目のテスト結果から、算数と理科は似ている、国語と社会は似ているということがわかればそれぞれをまとめ、
その背後に理系能力と文系能力があるというように、
・・・4 科目を 2 つの能力に要約したことになります。
今まで 4 教科のテストをやっていたものを、
・・・理系のテストと文系のテストに集約することが可能になるわけです。



⑦ 単回帰分析

回帰分析のうち、単回帰分析というのは
１つの目的変数を１つの説明変数で予測するもので、その２変量の間の関係性を**Y＝aX＋b**という**一次方程式の形で表します**。
a（傾き）と**b（Y切片）**がわかれば、
X（身長）から**Y（体重）**を**予測すること**ができるわけです。



⑧ 重回帰分析

単回帰分析が、１つの目的変数を１つの説明変数で予測したのに対し、重回帰分析は１つの目的変数を複数の説明変数で予測しようというものです。
多変量解析の目的のところで述べた、身長から体重を予測するのが”単回帰分析”で、身長と腹囲と胸囲から体重を予測するのが”重回帰分析”です。
式で表すと以下のようになります。

単回帰分析　$Y=aX+b$

重回帰分析　$Y=b_1X_1+b_2X_2+b_3X_3+b_4X_4\cdots+b_0$

No.	体重(kg)	身長(cm)	腹囲(cm)	胸囲(cm)
1	79.8	177.5	88.4	96.7
2	58.0	171.0	71.9	85.8
3	56.2	167.6	64.9	78.8
4	53.4	173.5	66.7	83.2
5	59.0	169.3	72.9	86.9
6	70.2	180.9	77.9	92.3
7	64.6	178.9	69.2	89.9
8	65.0	174.4	72.3	86.1
9	64.8	178.5	75.2	87.7
10	68.0	177.8	78.0	88.1
11	68.4	180.9	75.1	88.2

※ 河内まき子・持丸正明 2005　AIST人体寸法データベース, 産業技術総合研究所H16PRO 287.

体重・身長・腹囲・胸囲の
【人体寸法データ】

回帰統計	
重相関 R	0.9352474
重決定 R2	0.8746877
補正 R2	0.86594498
標準誤差	2.32023507
観測数	47

●エクセルの「分析ツール」から「回帰分析」を用いると結果が簡単に出力されます。

分散分析表

	自由度	変動	分散	散比	有意 F
回帰	3	1615.81543	538.605	100.047565	2.020E-19
残差	43	231.490103	5.383		
合計	46	1847.30553			

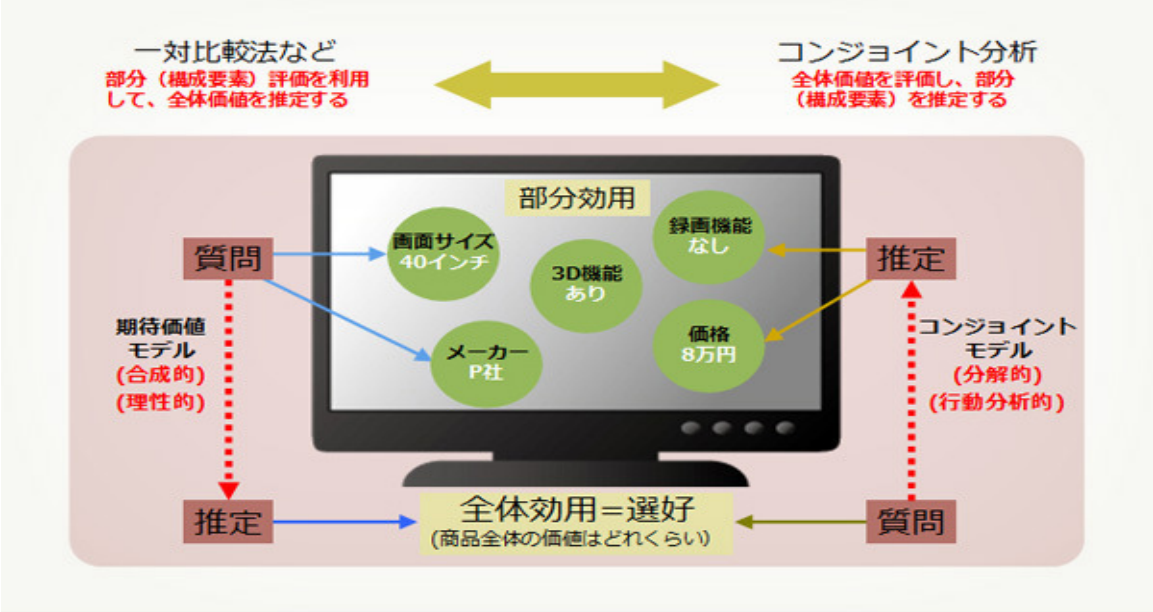
	係数	標準誤差	t	P-値	下限 95%	上限 95%
切片	-62.123638	11.547792	-5.3796984	2.8948E-06	-85.41198	-38.835297
身長	0.2233	0.0792	2.8187	0.0073	0.0635	0.3831
腹囲	0.6706	0.1280	5.2395	0.0000	0.4125	0.9287
胸囲	0.4300	0.1763	2.4399	0.0189	0.0746	0.7855

⑨ コンジョイント分析

コンジョイント分析とは、最適な商品コンセプトを決定するための代表的な多変量解析を用いた分析方法で、個別の要素を評価するのではなく、商品全体の評価（全体効用値）することで、個々の要素の購買に影響する度合い（部分効用値）を算出する手法です。

店頭に 2 種類しかなかったとします。一方は 3D はないが価格が安い、もう一方は 3D はあるが価格が高い。

それに対し、コンジョイントモデルは、いくつかの商品を提示して全体としてどれが一番欲しいかを質問し、その結果から各属性の重要度や水準の効用値を算出しようとするものです。期待価値モデルが合成的で合理的だといわれるのに対し、コンジョイントモデルは分解的で行動分析といわれます。



このような場合、どちらがより欲しいかを考えたときに、「商品 A は、3D がないことは-30 点、しかし価格が安いことは＋20 点なので合計 90 点」。「商品 B は価格が高いことが-20 点なので 80 点」。この結果から「100 点ではないけれど、どちらがよいかといえば、商品 A だ」。という結論を出すことができます。

このように個々の要素や機能（属性）の有無や値（水準）の良し悪しを質問し、その合計から全体の価値を定量的に計ろうというモデルを“**期待価値モデル**”といいます。　　期待価値モデルの場合、3D 機能と録画機能ではどちらを優先するかなどの重要度を質問することもあります。

（4） データマイニング

データマイニングとは、**ＣＲＭ** すなわち「**顧客１人ひとりの深い理解に基づく企業と顧客の長期的かつ良好な関係を形成する手法＆戦略**」を強力にサポートする**テクノロジーです**。

より具体的には、企業が収集する大量のデータを分析し、有用なパターンやルールを発見し、アルゴリズムによって、マーケティング活動を支援する統計的手法やツールの集合体のこと。

よく、データマイニングと統計解析の違いを比較することがあります。

- データマイニングは**知識発見**で、
- 統計解析は**仮説検証**であると言われます。

・・・**はたして本当にそうなのでしょうか？**

データマイニング	統計解析
データ量が多い 知識発見	データ量が少ない 仮説検証

確かに、統計解析が扱うデータ量は比較的少なく、データマイニングのほうが多いでしょう。また、データマイニングには知識発見の要素もありますが、データを入れれば何らかの知識が自動的に発見できるものではありません。 **データマイニングには 2 種類ある**、すなわち**知識発見**だけではなく、統計解析と同じように、**仮説検証**もあることを認識しておく必要があります。

① 仮説検証（目的志向）的データマイニング

（●推定、把握（量的変数） ●分類、抽出（質的変数） ●将来の予測）

仮説検証的データマイニングの中で、「**推定、把握**」というのは、例えばどのエリアでどのような商品がどの程度売れているのかといった、量的数値を推定したり把握したりするものです。

「**分類、抽出**」というのは、そのエリア別に売れている商品や商品カテゴリーを抽出し、 分類､整理して分析するものです｡この 2 つは正しい現状認識をするという目的で使うものです。

「**予測**」は現状ではなく、 将来の売上高や売れ筋商品などを何らかのモデルを作って予測することをいいます。

② 知識発見（探索）的データマイニング

（●アソシエーションルール策定 ●クラスタリング）

知識発見的データマイニングの「**アソシエーションルール策定**」は、同時に何と何が買われているかなどを探索的に知ることで、 例えば､この商品を買った人にはこの商品をお薦めしようというレコメンデーションに活用します。

「**クラスタリング**」は、 購買動向などから似たような人をグループ化し、グループ毎に最適な施策を打とうというものです。この 2 つは､目的変数がないので、 多変量解析でいうところの要約の手法に当たりますが、分析の目的がないわけではありません。

③ データマイニングが解決する課題

課題を、商品についてと、顧客について、に分けてみました。どの課題も、２つのデータマイニング分類のどれかに当てはまります。

これらのマーケティングの課題を解決するのが、データマイニングの究極の目的です。

A) 商品について知りたいこと

- **どこでどの商品がどれくらい売れているのだろうか？・・・推定・把握**
- **自社の商品はどのように分類すればよいのだろうか？・・・分類・抽出**
- **今後力を入れるべき重点商品はどの商品なのだろうか？・・分類・抽出**
- **この商品は将来どの程度売れるのだろうか？・・・・・・・・将来の予測**
- **どの商品とどの商品が一緒に買われているのだろうか？・・ｱｿｼｴｰｼｮﾝ策定**
- **新商品の評判はどうだったのだろうか？・・・・・・・・・・ｸﾞﾙｰﾌﾟの特性を知る**

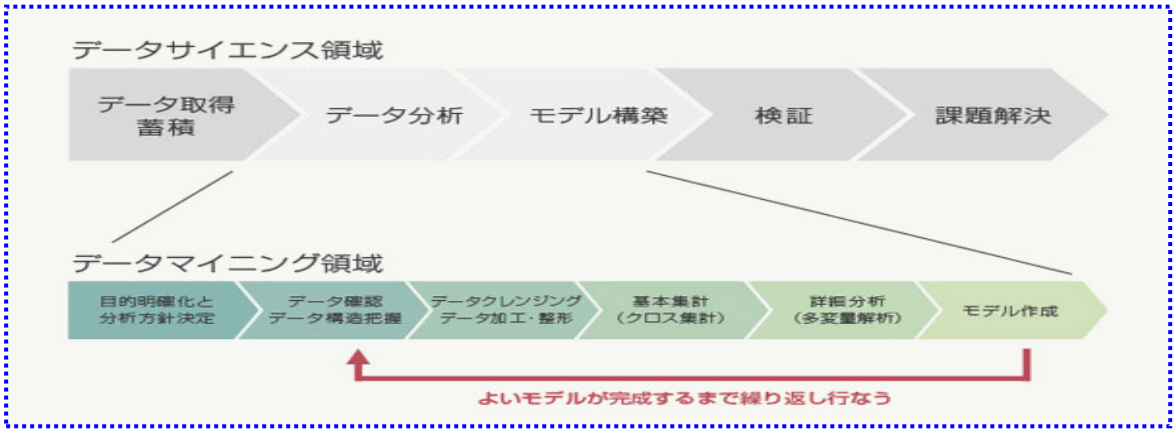
B) 顧客について知りたいこと

- **この顧客はどんな商品を買ってくれているのだろうか？・・推定・把握**
- **優良な顧客、離反しかけている顧客は誰なんだろうか？・・分類・抽出**
- **この商品は将来どの顧客が買ってくれるのだろうか？・・・将来の予測**
- **どの顧客クラスターにはどの商品を薦めればよいのか？・・クラスタリング**
- **自社の顧客は性年代別、地域別にどんな人なのだろうか？・・ｸﾞﾙｰﾌﾟの特性を推測**

④ データマイニングとデータサイエンス

◆**データマイニングとは**、扱うデータは整形されておらずノイズも多い、混沌としたものです。これらのデータをいかに科学的アプローチによって、課題解決につなげるかが重要。

◆**データサイエンスは**、データの取得、蓄積、解析、 モデル構築、検証、課題解決までを一気通貫で科学することが求められ、 主にこのステップの中のモデル構築までを主な守備範囲としています。



（５） ビッグデータ時代を切り開く知識と人材

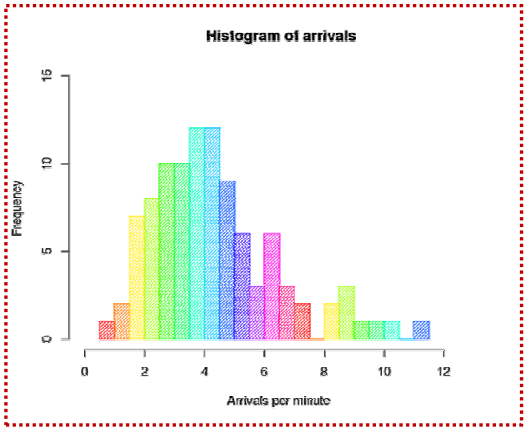
近年、不確実性の時代を迎え、急速な情報技術の進化があいまって、バラツキのある大量のデータ（ビッグデータ）を収集、分析し、意思決定に活かすことが、企業経営に必須だという考えが台頭し、2012 年 3 月、米国のオバマ大統領がビッグデータ研究開発イニシアティブを発表し、統計学が一躍脚光を浴びるようになって来ました。

① ビッグデータ時代を切り開く、五つの基礎統計学

ビッグデータ解析の３大要素技術は、ビッグデータ工学，データ可視化，データ解析法であると言えます。そんな中で基礎となる統計手法（ヒストグラム、平均値と中央値、ランダムサンプリング、正規分布、回帰分析等々）から時代に必要な能力を抽出して見よう。

◆ **ヒストグラム**

実際のデータを視覚的にわかりやすくグラフ化したものです。
棒が長ければ長いほど、そこにデータが沢山あると言うことです。



◆ **平均値と中央値**

平均値とはすべてのデータの平均です。
しかし、**平均値には弱点があります**。

よくある喩え話に、貧困層の集まる居酒屋に**ビル・ゲイツ**が入ってきたら、平均所得は一気にあがるけれど、もともといた貧困層の人びとが豊かになったわけではない、というものがあります。（平均値はどれかひとつ、極端な値を示すデータが混入しただけで、一気に値が上昇あるいは下落してしまう。それに対して中央値は、貧困層もビル・ゲイツも含めたデータを集める際、所得の低い順からデータを並べたとき中央にくる値を示します）

これは常に中央に来る値を示してくれるので、平均値を求めるときのようにブレがなく時系列比較できます。欧米の新聞では、徐々にこちらの値が使われ始めているといいます。

◆ **ランダムサンプリング（無作為抽出）**

選挙の際の「**出口調査**」というものがあります。この手法のメリットは、**少ないサンプル数で結構な精度の高いデータの解析結果を得ることが出来ます**。

◆ **正規分布**・・・経営学の分野にも応用され、**消費者の行動の図としても活用されています**。

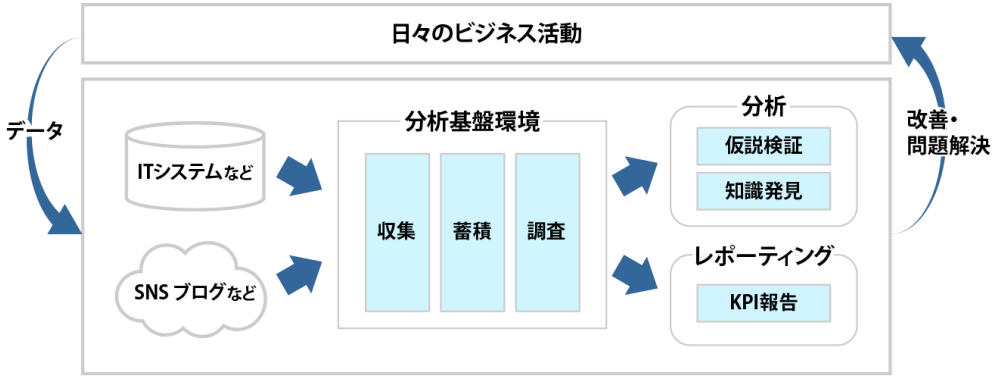
◆ **回帰分析**・・回帰分析によって、たとえば年ごとのワインのおいしさやあるいはどういった年代の顧客がどういった商品を買うかということを**定量的に分析することができます**。

ビッグデータを活用したいのならば、統計学がどういったものかをきちんと概説だけでも理解しておく必要があります。また、データの可視化は、次元圧縮、特徴抽出、パターン認識など、膨大な高次元データそのものや解析結果を人間が的確に把握できるようにするための技術の習得が欠かせません。

統計数理・数理科学・機械学習・情報処理などの横断型の方法論から、それぞれの領域分野に熟知した知識と経験が要求されます。

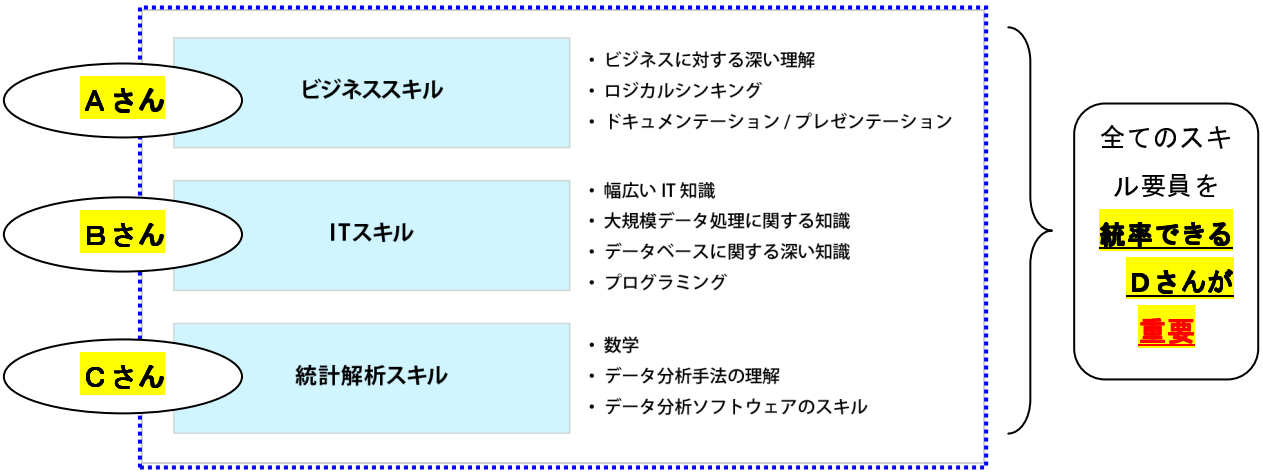
② データサイエンティストの育成と必要なスキル

◆ **データサイエンティスト**は、ビッグデータ解析のための要素技術とともに、領域分野の知識と経験、問題発掘能力、コミュニケーション能力も必要なことから、**方法論と領域研究の双方に熟知した人材の確保と育成が不可欠となってきています。**



◆ **データサイエンティストに必要なスキル領域**

・ビジネススキル　・ＩＴスキル　・統計解析スキル



現実の課題を解決するためには、問題の本質の把握、定式化、データ取得、分析、知識獲得、課題解決の全過程に関与できる全人的能力が必要となります・・・**個人で全てのスキルを網羅するのは容易ではない。複数の人材を有効活用し統率できる能力の人材が重要となる。**

5. 統計学とビッグデータ解析の違い

統計学とビッグデータ解析は似ているようで、実は考え方としては対極にあったりします。
（本来、統計分析の学問というのは、いかに少ないサンプリング数で全体を捉えるかを追求したもの）

（１） 対極的な違いについて簡単な視点

- **統計学は・・・何らかの問題の要因を見つけだして改善しないといけないような場合**
  - **一部のサンプルだけを分析して、全体のデータを推測する。**
  - **サンプルデータは可能な限り正確で、整理されていなければいけない。**
  - **分析結果がなぜそうなったのか、理由をはっきりさせることにこだわる。**

- **ビッグデータ解析・・セールスとかマーケッティングでより多くモノを売りたい場合**
  - **一部のサンプルではなく、全てのデータを分析する。**
  - **サンプルはちょっとくらい乱雑でもかまわない。データの質より量を重視する。**
  - **分析結果がなぜそうなったのか、理由はどうでもよい。**

サイトのアクセス解析とかアプリのダウンロード分析で役立ちそうなのは、ビッグデータ解析です。なお、アマゾンがオススメ本を表示させるのは、ビッグデータ解析による賜物ものです。

ただし、結局は商品の魅力とか、インフルエンサー（人々の消費行動に影響を与える人）に口コミされるとか、マーケティングは、そういう要素の方が大きいと思われるので、あまりデータ分析や解析だけで判断するのは、難しいかも知れません。　また、本当に大量データが必要なのか？ここでは、目的さえ明確であれば、必ずしも大量データはいりません。

（２） ビジネススキルのロジカルシンキングとは

データ分析の目的の多くは、ビジネス上の問題解決やビジネスの継続的な改善です。
それらを円滑に進める技術として、**ロジカルシンキング（論理的思考力）は非常に大切です**。

例えば、「12 月のアクティブユーザーの数が 11 月比で 20%減少した理由はなぜか」という問いを考えてみましょう。やみくもにデータを抽出し、「女性ユーザーの数が 20%減少していたからです」という結論を出したとしても、誰も納得してくれないでしょう。

というのは、上記の結論では「新規ユーザーが減っているのでは？」や「30 代のユーザーが減っているのでは？」という**他の仮説を検証できていないからです**。
このような間違いを犯さないために、ロジックツリーや MECE（漏れがなくダブりもない）といった**ロジカルシンキングのアプローチを利用しながら、論理的に結論を導くことが重要となります**。

6. “人工知能・ＡＩ”とは!? ビッグデータとの関連性

人工知能には、基本的に“推論”と“学習”という 2 つの機能が搭載されており、両者を組み合わせることで様々なことを実現しています。

（1） 人工知能のフェーズ

現在、世界に存在する人工知能は、その成熟度によって以下の「**４段階に分類**」されます。

① フェーズ 1：簡単な制御プログラム

近年、冷蔵庫やエアコンに人工知能を搭載した「スマート家電」を市場で見かけることが多くなりました。エアコンで言えば室温、湿気、エネルギー消費量などのデータを収集し自動運転を行うといったものです。しかし実はこれ、人工知能というよりは、システム工学などの分野であり、マーケティング的に人工知能と呼んでいるに過ぎません。

② フェーズ 2：スタンダードな人工知能

先に紹介した“**推論**”と“**学習**”を基本的に備えているのがこの人工知能であり、大量のデータベースを所持していることから、時に人間を超える振る舞いをするものが存在します。
代表的なのがコンピュータチェス(コンピュータが指すチェス)であり、IBM が人工知能を搭載して開発した「**Deep Blue(ディープブルー)**」は、2007 年にチェス世界王者から白星を獲得しています。

③ フェーズ 3：機会学習を搭載した人工知能

まずこの代表例を挙げると、Google や Yahoo!などの検索エンジンが該当します。
サンプルデータをもとに学習していき、ルールや知識を独自につける人工知能です。
検索エンジンではアルゴリズムをもとにコンテンツの評価を行い「良いコンテンツ」を検索結果上位に、そして「悪いコンテンツ」の順位を下げていきます。

④ フェーズ 4：ディープラーニングを搭載した人工知能

人間がりんごを目にして「これはりんごだ」と認識するのは当たり前にことですが、従来の人工知能には不可能な領域でした。これまではりんごの色や形といったメタデータを人間がインプットしなければならなかったのです。

しかしディープラーニングでは、例えば大量のりんごの画像を読み込ませることによって機械が徐々にりんごを認識していきます。「りんごは赤い」「りんごは丸い」といった知識を次第に付けていくのです。

ちなみに、近年最も注目されている人工知能はこのディープラーニングを搭載したものであり、主に音声認識や画像認識といった分野で活用されています。

（2） ビッグデータと人工知能の関連性

なぜビッグデータ活用が進む中で、人工知能というキーワードが頻出しているのか？
それは、ビッグデータ解析のため、人工知能を用いることに徐々に期待が高まっているからです。

- ビッグデータはその名の通り、膨大かつ高頻度で更新されているデータ群を指します。
- これまでのビッグデータ解析では人が解析ツールを用いて、抽出・加工・分析・レポート・情報化するのが当たり前でした。このうち解析ツールが担っている領域は抽出～レポートまでです。

つまり、ビッグデータを最終的に有用な情報へと変換するためには、人の知能がなければ成しえなかったのです。しかし近年になり、この環境に変化が起きつつあります。

- 先に紹介したディープラーニングの登場で、徐々に知識を蓄積し学習し、自ら意思決定を下す人工知能が増加しています。
- センサーなどから取得したデータを取り込み蓄積し、データを構造化した上で最適な次の戦略を導き出す人工知能が既に登場しているのです。

もちろん、未だ人間の知能のように柔軟性があるわけではなく、実用的になるのはもう少し未来の話です。また、IoT（Internet of Things）というもう一つにキーワードもビッグデータと人工知能に深く結びついています。

- IoT で重要なのは、センサーから取得した膨大なデータ（ビッグデータ）を瞬時に処理し、ユーザーに最適なフィードバックを返すことです。

- 例えばスマホアプリで自宅の鍵をロックできる「スマートロック」なら、鍵を開けた時間や誰が開けたかといった情報をスマホでユーザーに通知するのがフィードバックとなります。スマートロックでは少々簡単な処理ですが、複雑になるにつれ人工知能が活躍します。

IoT 製品に対するユーザーの満足度を向上するためには、ビッグデータをリアルタイムで解析しつつ分析や学習を行い、その時々で常に最適なフィードバックを返すことです。

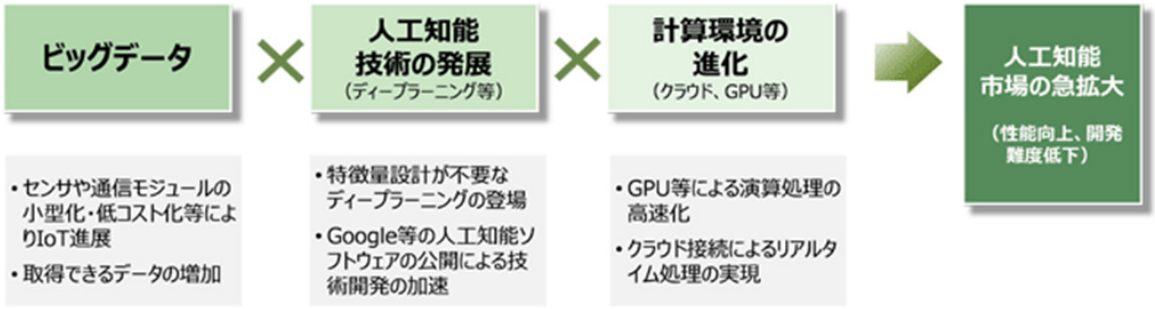
人工知能は“推論（すいろん）”や“学習（がくしゅう）”、そして“ディープラーニング”といった機能でこれを可能にします。だからこそ、現在ビッグデータと共に人工知能というキーワードが急上昇しています。

また、人工知能に何よりも大切なのがデータ処理を行う基盤です。
ビッグデータという膨大かつリアルタイム性の高いデータを処理するのですから、当然高性能かつ高品質なデータ処理プラットフォームが必要となります。

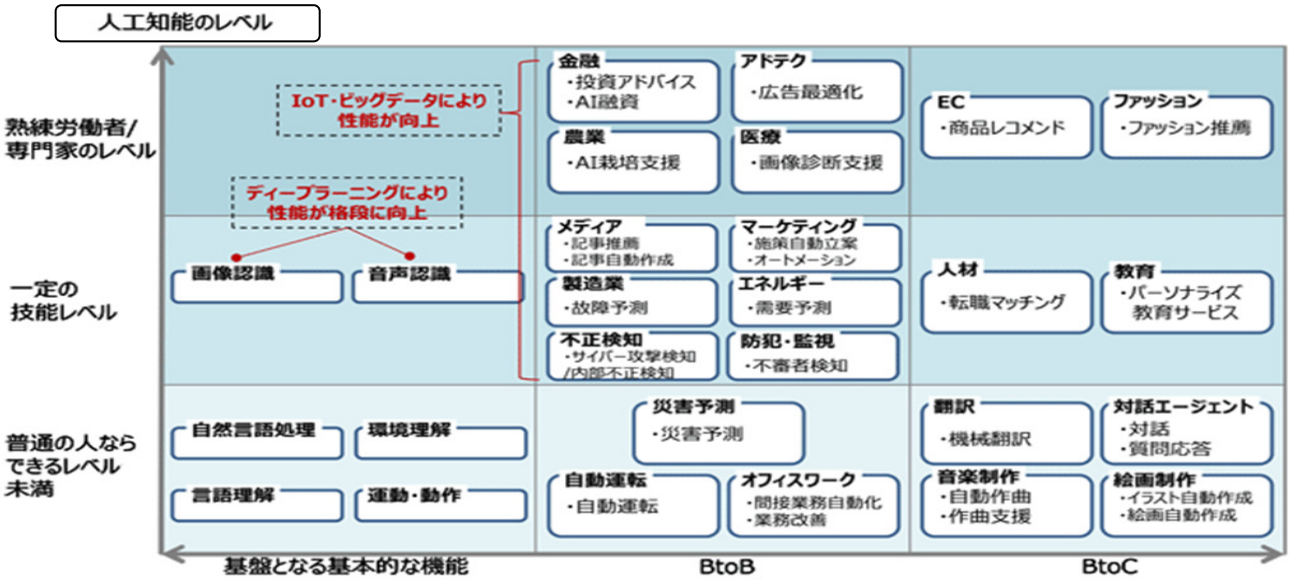
（3）　ビッグデータと人工知能市場の急拡大

人工知能がここにきて急速に進展している背景には 3 つのブレークスルーがある。

- まず、人工知能が判断を学ぶにあたってお手本となる**膨大なビッグデータの収集が可能になった**。これはセンサーや通信モジュールの小型化・低コスト化に伴う **IoT の進展がもたらした**。
- 次に、**画期的な人工知能アルゴリズム「ディープラーニング」の登場により技術開発が加速したこと**。この技術の登場が人工知能業界の空気を変えた。Google や Facebook など世界の大手 IT 企業が相次いで人工知能に莫大な投資を行い、大きな注目を集めている。
- そして、クラウド含め計算環境が進化し、ビッグデータに対してディープラーニングなどの**人工知能アルゴリズムを適用できる環境が整ったこと**。



- 人工知能のレベルは、**農業**における栽培ノウハウの AI 化や**医療**における画像診断支援のような**熟練労働者・専門家に匹敵するレベル**、Web メディアにおける記事の自動生成のような専門家には及ばないが**一定の技能レベルを提供できるレベル**、そして、対話エージェントのような**普通の人間のレベルにもう一歩到達していないレベルの 3 つに分類した**。



【ＤＲニュース・002】ＩＯＴプラットフォームビジネスの拡大でも話題にしたように、
人工知能技術の発展やクラウド接続によるリアルタイム計算環境の進化により、
プラットフォームに集まった「ビッグデータ情報」を解析できる時代が到来しています。
ビッグデータの活用事例の理解を深めて、皆さんの知識を出し合って、時代を謳歌（おうか）しよう。