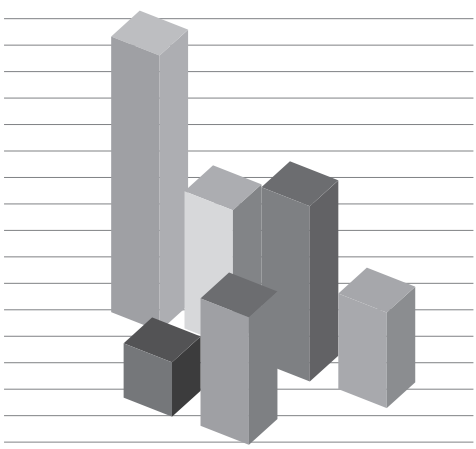
Excel で学ぶデータサイエンス入門講座

No.1

基　礎　編

監　修／株式会社 ROX
執　筆／福井　高志



コガク

目　次

第１週
ビッグデータとデータの前処理

【学習のポイント】

今週は，ビッグデータの特性についての学習から始めます。データについて理解を深めることは，今般キーワードとなっている IoT や AI を理解することにも繋がります。アクセス可能なデータが爆発的に増加していること，またその流れを受けてデータの重要性がいかに増しているかを実感してください。

後半は，データの質の話です。データが多ければ多いほどいいという訳ではありません。データ分析の質はデータの質によって大きく左右されます。データ分析の前段階として，データをすぐに分析可能な状態にしておくことは，非常に重要なプロセスです。一般的には軽視されがちなこの前処理というプロセスですが，データ分析の実務においては，時には全作業時間の８割を占めることもあります。大切な前処理のプロセスの手順を，しっかりと習得してください。

1.1　データサイエンティスト / データサイエンスとは？

1.1.1　データサイエンティストとは？

データ分析といえば，昔から基本のビジネススキルの一つであり，そのスキルを身につけようと多くの人が取り組んできました。更に，最近になってデータサイエンティストが新しい職業として認知されるとともに，ハーバード・ビジネス・レビュー（2013 年 2 月号）では「データサイエンティストほど素敵な仕事はない」と取り上げられるなど，高い注目を集めています。

一般社団法人データサイエンティスト協会では，データサイエンティストを「データサイエンティストとは，データサイエンス力，データエンジニアリング力をベースにデータから価値を創出し，ビジネス課題に答えを出すプロフェッショナル」と定義しており，データサイエンティストに求められるスキルセットとして以下の 3 つの力を挙げています。

1. ビジネス力（business problem solving）
   課題背景を理解した上で，ビジネス課題を整理し，解決する力
2. データサイエンス力（data science）
   情報処理，人工知能，統計学などの情報科学系の知恵を理解し，使う力
3. データエンジニアリング力（data engineering）
   データサイエンスを意味のある形に使えるようにし，実装，運用できるようにする力

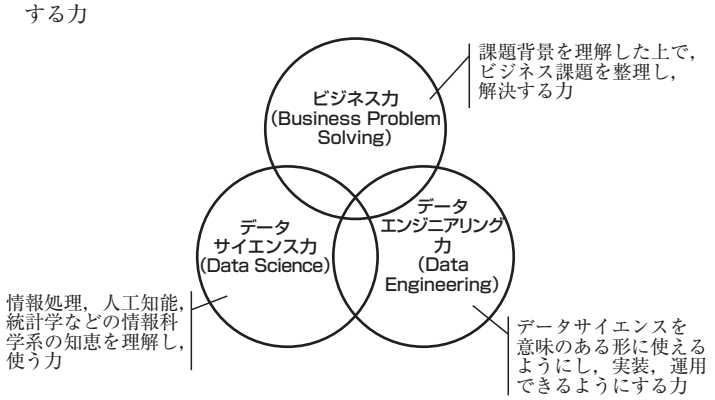


図 1.1　データサイエンティストに求められる 3 つの力
（出典：一般社団法人データサイエンティスト協会プレスリリース 2014 年 12 月 10 日」「データサイエンティスト協会，データサイエンティストのミッション，スキルセット，定義，スキルレベルを発表」図 1：データサイエンティストに求められるスキルセット）

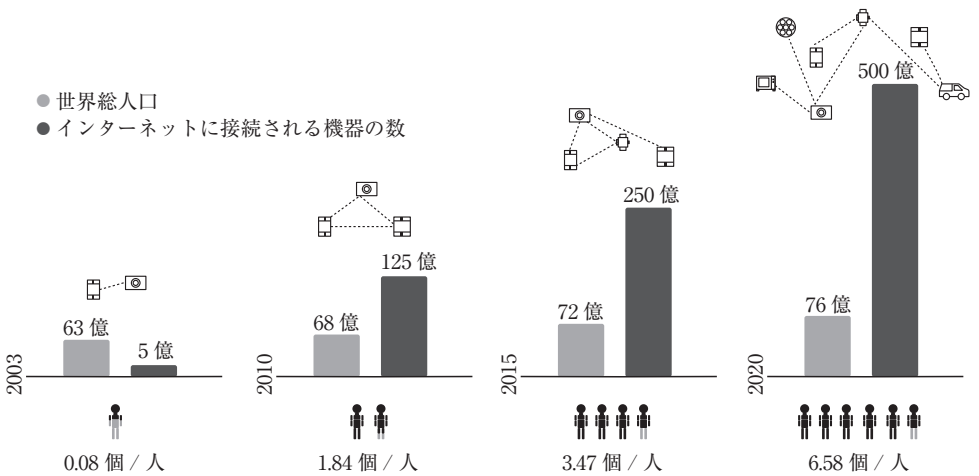
本講座は，データサイエンス力としての統計学，データエンジニアリング力として Microsoft Excel をツールとしたデータ分析力を身につけることを目標とした入門講座です。

1.1.2　データサイエンス進展の背景

上記で紹介したデータサイエンティストに求められる力は，従来のデータ分析において求められていた力と比べても，特段の目新しさはありません。では，なぜ今般，データサイエンティストやデータサイエンスなどの言葉がキーワード化しているのでしょうか？その背景の 1 つとして，データの取得が容易となり，分析対象となるデータの量が爆発的に増えていることが挙げられます。データの取得が容易になった要因は色々ありますが，ここではインターネットの浸透と IoT の進展，この 2 つを取り上げます。

まず第 1 の要因はインターネットの浸透です。インターネット環境が普及したことで，私たちがアクセスできる情報量は爆発的に増加しました。今日では，ウェブサイトに関する知見を持たない人でも，ブログなどを通じて情報を発信することができます。やり取りされる情報も，テキストから始まり音声，画像，映像と，データ量は段々と大きくなっており，YouTube ユーザーのアップロードする動画の時間の総計は，2015 年時点で 1 分あたり 300 時間とも言われています 。

第 2 の要因は IoT の進展です。IoT とは，Internet of Things（モノのインターネット）のことで，あらゆるモノがインターネットに繋がるという「概念」を指します。従来，インターネットに繋がるものと言えばパソコンや携帯電話などで，その利用は画面を介したネットサーフィンやメールなどがほとんどでした。それが，今ではメガネや靴などの身近な製品から，工場内の機械，さらにはスマートシティと呼ばれるインフラレベルまでインターネットに繋がるものが出てきており，インターネットに接続される機器の数は急速に増加を続けています。なお，図 1.2 の人型下部の数字は一人当たりの IoT 機器個数です。



（出典：Jigsaw Academy Pvt Ltd.（https://www.jigsawacademy.com/iot/）
“Connected Devices per Person”を基に執筆者が作成）

図 1.2　IoT 機器の数

また，データの取得が容易になったことで，取得されるデータの量だけでなく種類も大きく増えました。製品の利用状況や機械の稼働状況，購買行動からインターネットの検索履歴まで，ありとあらゆるデータが日々蓄積されています。

さらには，データの蓄積コストが下がったことも見逃せません。ただデータを貯めておくだけでも，そこにかかるコストは無視できないものですが，クラウドコンピューディングの登場により，そのようなコストも大きく低下しました。

このようにして，これまでには考えられなかった多種多量のデータが取得され，蓄積されるようになりました。これらのデータは，ビッグデータと呼ばれることがあります。

1.2　ビッグデータとは？

1.2.1　ビッグデータの定義

ビッグデータとは，その名の通り巨大なデータ群のことですが，「典型的なデータベースソフトウェアが把握し，蓄積し，運用し，分析できる能力を超えたサイズ」などと定義されることが一般的です。どの程度の量のデータをビッグデータというのか，と言った明確な定義があるわけではありませんが，多くの場合，量的には数十テラバイトから数ペタバイトとされます。

一方で，巨大なだけでなく，データが複雑であることも，ビッグデータの特徴としてよく挙げられます。この文脈でよく使われるのが，データを構造化データと非構造化データに大別する分類です。構造化データは，一言で言えば整理がしやすいデータです。Excel や CSV など，行と列の二次元に並んでいるデータがその代表です。逆に非構造化データとは，整理の難しいデータで，規則的な区切りがなく二次元の表形式では表しにくいデータです。非構造化データの代表的な例は，画像や動画，文章などのデータです。この中間で，データ内に規則的な区切りはあるが二次元化しにくい半構造化データもあり，json ファイルや HTML タグ形式が該当します。

従来，データと言えば扱いやすい構造化データがほとんどでしたが，先に述べたようなデータ取得の容易化，取得経路の多岐化によって取得されるようになったデータには非構造化データも多く，これらを総称してビッグデータと呼ぶ傾向があります。

1.2.2　ビッグデータの特性

ビッグデータの質的な定義，つまりビッグデータが持ちうる特性としては，ガートナーが提唱した次の 3 つの V（Variety，Volume，Velocity）が有名です。

（1）Variety

Variety は，データの多様性です。上で見た通り，データはテキストや音声，画像など様々な情報かつ様々なファイル形式です。