ゼロから学ぶ DS 入門講座

データサイエンス

著
阿部晋也

株式会社 コガク
とおとうみ出版

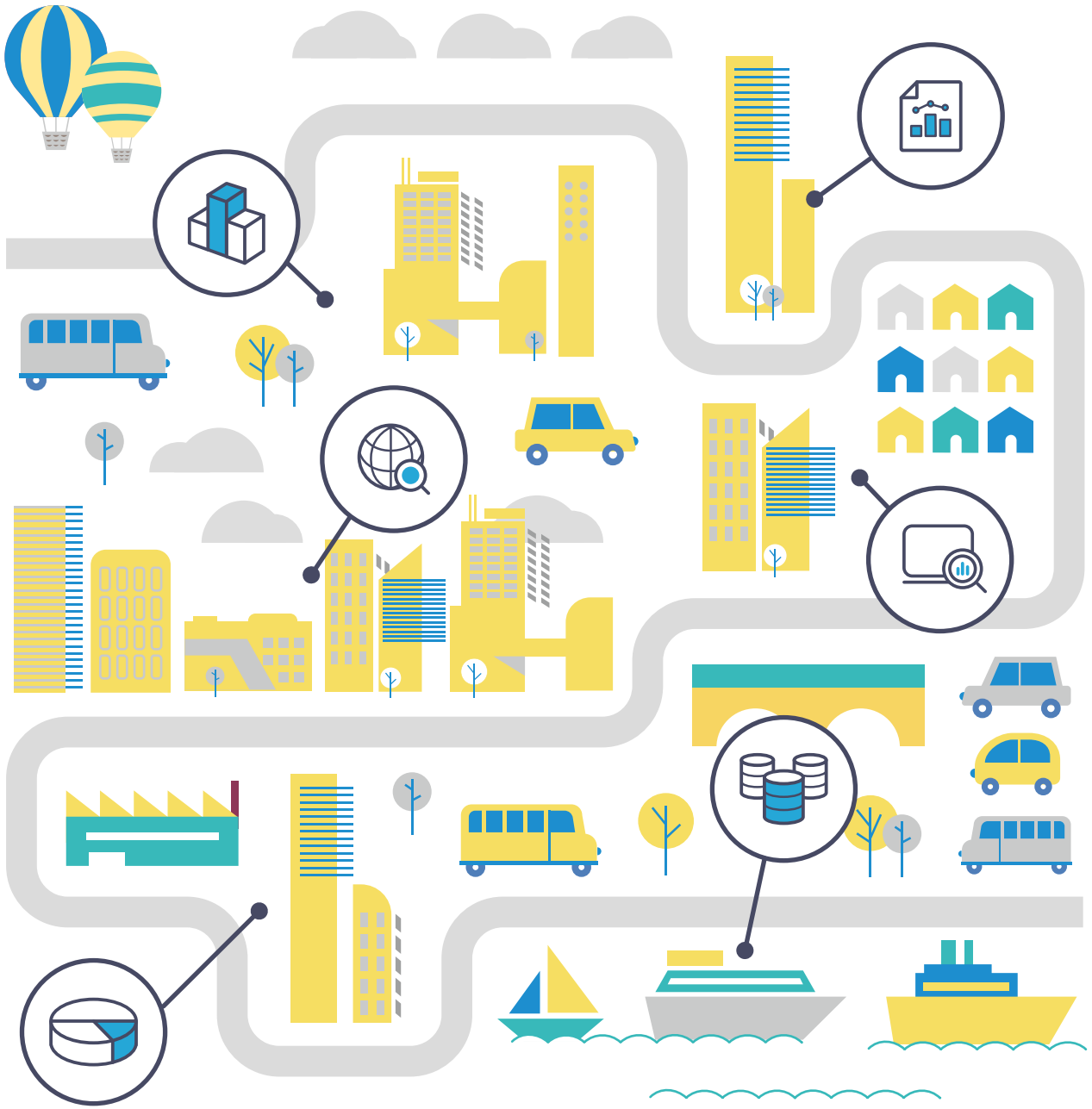
Data Science

はじめに

IoTの浸透、インターネットやクラウドの普及によりあらゆるもののデータ化はさらに加速しています。私たちの日常生活をデータなしで語ることはできません。その一方で、データは正しく扱い読み解かなければ、誤った結果にもつながりかねません。「データサイエンス」という言葉が昨今とくに強く叫ばれるようになったのはごく自然な現代の流れといえるでしょう。

本書では「データサイエンスって何?」という基本から、データの見方・種類・可視化などの基本的な考え方、データの活用手法など、データサイエンスの概要を解説します。難解な数式は用いず、あくまでデータサイエンスのイメージが掴める内容に留めることで入門レベルの学習者が取り組みやすいようにしています。

データリテラシーを身につけ正しくデータを読み解き、新たな価値を創造するための第一歩を踏み出しましょう!



目次

はじめに …… 2
本書の使い方 …… 6

第1章 データサイエンスのこと知っていますか？

1-1 私たちの日常生活にあふれているデータ …… 12
1-2 データサイエンスとは …… 16
1-3 データサイエンティストの仕事 …… 18
まとめ 第1章の振り返り …… 24

第2章 データの見方

2-1 平均というトラップ～正しい視点で見る …… 28
2-2 分散を考える …… 30
2-3 幅広い視点でみる …… 36
まとめ 第2章の振り返り …… 38

第3章 データの種類

3-1 量的データと質的データ …… 42
3-2 構造化データと非構造化データ …… 46
3-3 外れ値や異常値 …… 50
まとめ 第3章の振り返り …… 54

第4章 データの可視化

4-1 可視化の意味 …… 58
4-2 量的データの表し方 …… 60
4-3 質的データの表し方 …… 62
4-4 分布の表し方 …… 64
4-5 時系列データの表し方 …… 66
まとめ 第4章の振り返り …… 68

第5章 データの活用

5-1 回帰分析 …… 72
5-2 クラスタリング …… 78
まとめ 第5章の振り返り …… 80

第6章 データを制する者が DX を制す

6-1 AI とデータサイエンス …… 84
6-2 ビッグデータとデータサイエンス …… 87
6-3 業界で求められているデータリテラシー …… 90
まとめ 第6章の振り返り …… 95

ワークシート …… 96
解答一覧 …… 98
索引 …… 102

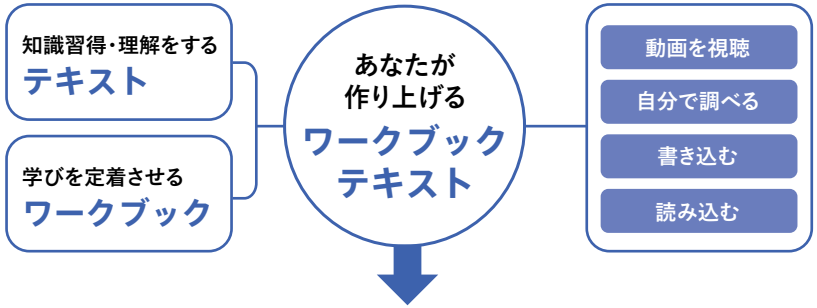
本書の使い方

本書はeラーニング「ゼロから学ぶデータサイエンス入門講座」(※)の講義をリアルに紙面上で再現しました。講義の臨場感を感じられるよう、できるだけ講師の言葉をそのままにお届けしています。

本書の特長として、「ワークブックテキスト」としてお使い頂けます。基本的な知識習得・理解をするための「テキスト」としての側面と、手を動かしてその学びを定着させる「ワークブック」としての二つの側面を持っています。ただ読むだけのテキストではなく、解いたら終わりのワークブックでもなく、その両方の機能を備えたのが「ワークブックテキスト」です。

まずはテキストを読んで基礎学習を進め大まかな流れを掴んでください。次に、本文随所に設けられた空欄を埋めるワークとして、キーワードを自分自身で書き込んでください。「ゼロから学ぶデータサイエンス入門講座」のeラーニング動画(※)を視聴して講師の説明を聞き取る方法もありますし、自分自身で書籍やネット等を使い調べる方法もあるでしょう。空欄に当てはまる正解は巻末に掲載していますので、答え合わせをしてみてください。空欄が埋まり文章が完成されれば、完全版のテキストが出来上がります。完成したテキストを繰り返し読むことでさらに理解を深めてください。

自分自身で読んだり、見聞きしたり、調べたり、そして手を動かして書いて、さらには正解かどうかを確認したり…と、様々な感覚を用いて「体験」することで、ただ読むだけでは得られない学習効果を得ることができるでしょう。テキストを「あなたが作り上げる」、そんなイメージで取り組んでみてください。



確実に身に付く!!

※ eラーニングの申込先…https://www.cogaku.co.jp

Data Science

第1章

データサイエンスのこと知っていますか？

Data Science

第1章

皆さんこんにちは。ゼロから学ぶデータサイエンス入門講座をお送りします。講師を担当させていただきます阿部晋也と申します。どうぞよろしくお願いいたします。

1章では、「データサイエンスのこと知っていますか？」として、最近このデータ分析だとかデータサイエンスという言葉が少しずつ出てきたかなというふうに思いますけれども、それについて皆さんと一緒に考えていきたいと思います。

1章のポイントですけれども、日常生活には様々なデータがあふれています。データサイエンスの重要性について確認をしていきたいと思います。また、データサイエンスを職業とするデータサイエンティストのスキルだとか役割について確認をしていきましょう。

POINT

- 日常生活にはデータがあふれ、データサイエンスの重要性が高まっています
- データサイエンティストのスキルや役割を理解しましょう

1-1

私たちの日常生活にあふれているデータ

私達の日常生活というのは様々なデータがあふれています。コンピューターだとかスマートフォンでいろんなデータをやりとりしていると思いますよね。SNSでいろいろ発信しているという人もいるかもしれないですしExcelか何か使ってデータ分析しているという人もいるかもしれませんね。

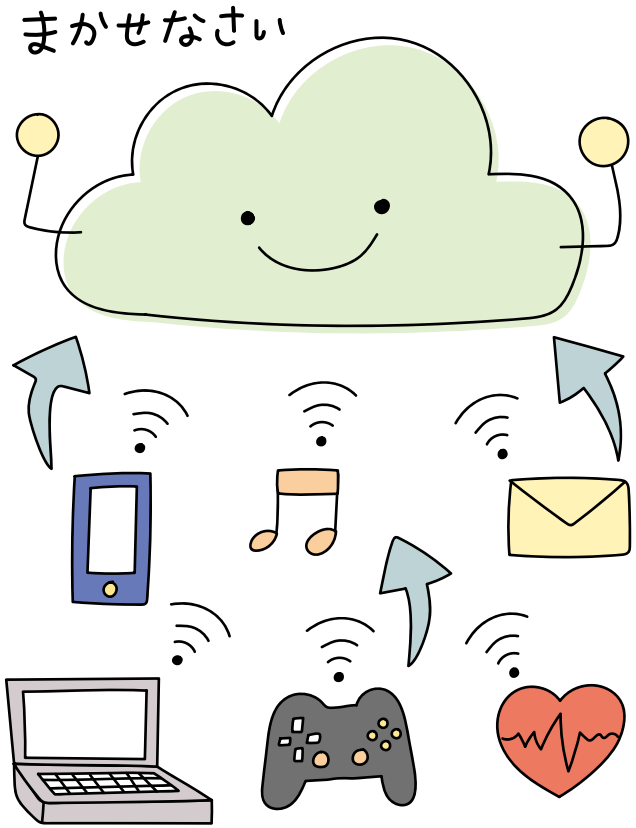
また、ネットワークも昔に比べるとすごく高速な通信ができるようになってきました。さらに、人々の行動の履歴だとか購入の履歴、様々なデータというのが蓄積されているというような形で、落ち着いて考えてみると「そう言われればいろんなデータがあふれているな」といったところがわかるかなと思います。

この日常生活のデータというのはいろいろ役に立っていて、私達の行動のアシストをしてくれているんですね。

図1.1　Twitterのつぶやきを分析して運営の参考に

例えば通信会社ですけれども、Twitterで投稿されたつぶやきというのを分析して運営の参考にしていたりするわけですね。図1.1で例えば「Xエリアは全然通話できなくなっちゃったよ」だとか、「Xエリアにいるけれどもネットは全然駄目だ」というようなつぶやきがいくつかあるわけですね。そこからそのデータというのを直接拾うというようなことは勿論するんですけれども、加工して、例えば「Xエリア」という言葉を出したりだとか、「通話不良」という言葉を出したり、また「ネットに繋がらない」というようなキーワードを出して、そこから「Xエリアはどうも何か悪いようだ」、「Xエリアの基地局を改善した方が繋がりやすくなるんじゃないか」といったことを確認していくわけなんですね。

こういうような事例のようにインターネットだとかクラウドコンピューティング、またはいろんなセンサーとインターネットが繋がるいわゆる（①　　）といったものが浸透したことによって、データを素早く大量にいろいろ収集することができるようになりました。その中でデータサイエンスの重要性が高まってきました。



1-2

データサイエンスとは

データサイエンスというのはデータとサイエンスをくっつけたものというような形で、両方とものいろんな性質が含まれているということになります。

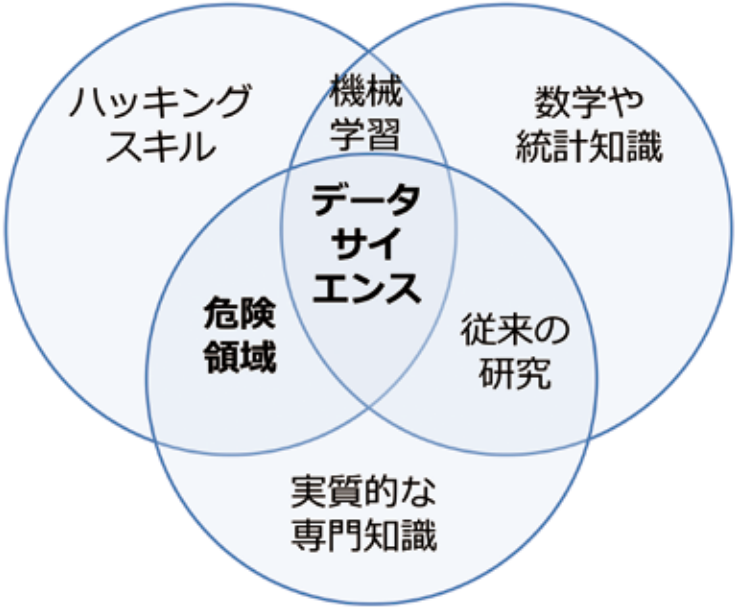


図1.2　ドリュー・コンウェイのベン図

有名なものにドリュー・コンウェイのベン図ということで、図1.2のようないろんな要素があるといったものを表した図があるんですけれども、例えばこの方は二つだけを組み合わせた場合最悪データサイエンスではなくて危険になる可能性があるという意味が含まれていると、こういうふうに言ってるわけです。つまりデータの（②　　）だとかという嘘となる可能性があるというふうに言っているわけなんですね。

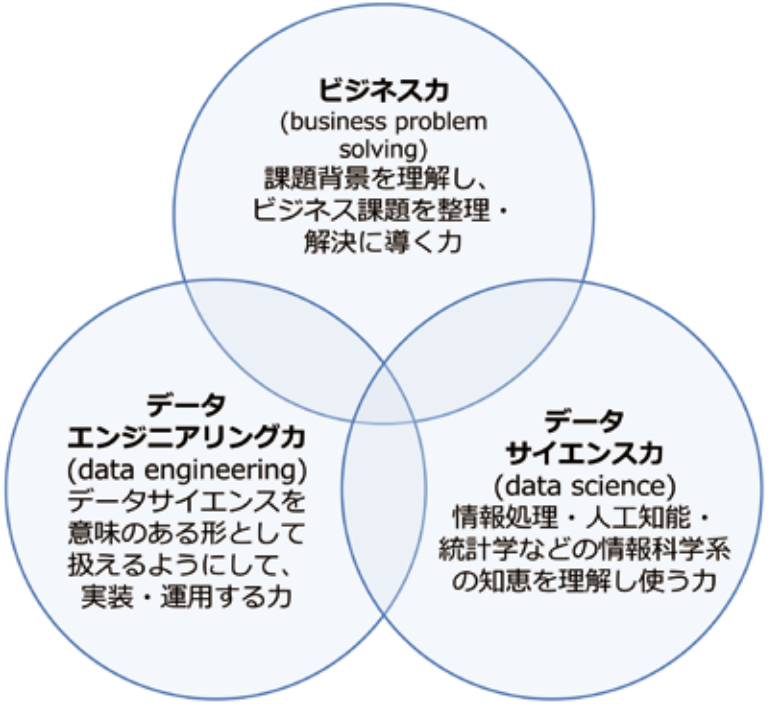
ですから単純にデータ分析をするというわけではなくて、ちゃんとしたことをやらないとデータなんか作っちゃうというような危険性もあるということを言っているわけなんですね。その中でデータサイエンティストという職業といったものが生まれてきています。

1-3

データサイエンティストの仕事

ちょっと間違えれば数学とかできる人なんか捏造とかになってしまうので、そうではなくてデータサイエンス自体に総合的な取り組みというのが必要になってきます。データ自体の本質を理解して、(③　　)するためにいろんなスキルが必要になってくる、その中で、データサイエンティストというような職業が注目されています。

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のスキルセットといったものが定義されていますのでそれを確認してみましょう。(図1.3)



【出典】一般社団法人データサイエンティスト協会・独立行政法人情報処理推進機構「データサイエンティストのためのスキルチェックリスト/タスクリスト概説(第二版)」

図1.3　データサイエンティストのスキルセット

データサイエンティストのスキルセットではどういったものがあるかですけれども、まずビジネス力というものが必要です。つまり課題の背景を理解する必要性があります。

もう一つはデータエンジニアリング力ですね。データサイエンスを意味がある形に捉えていく（意味のある形として扱えるようにして実装・運用する）。

あとはデータサイエンス力。情報処理だとか人工知能だとか、こういうような知識を統合する。

こういったような三つの要素が必要だというふうに言われています。

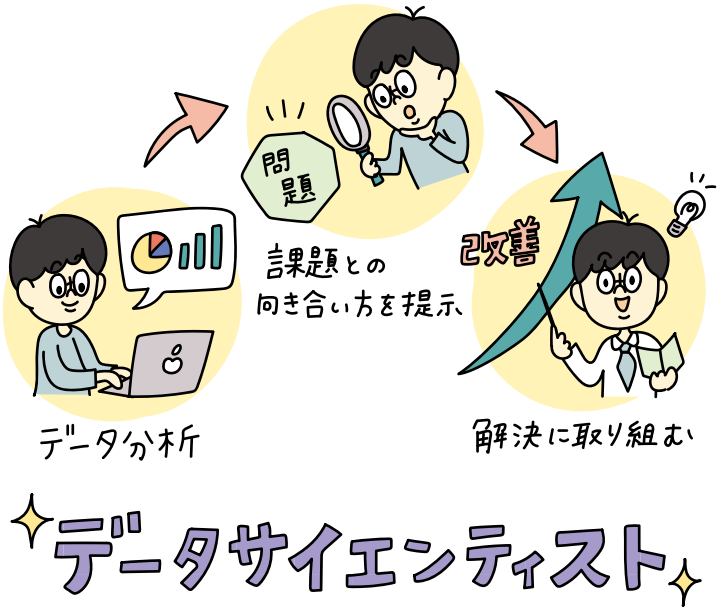
その中でデータサイエンティストの役割といったところを確認していきたいと思います。

- **解決したい課題への向き合い方を提示**する
- **データに対する向き合い方を提示**する
- **アルゴリズムの選定**や**モデルの構築**
- データの**分析結果に対する評価**
- **新しい技術や解法**へ取り組む

図1.4　データサイエンティストの役割

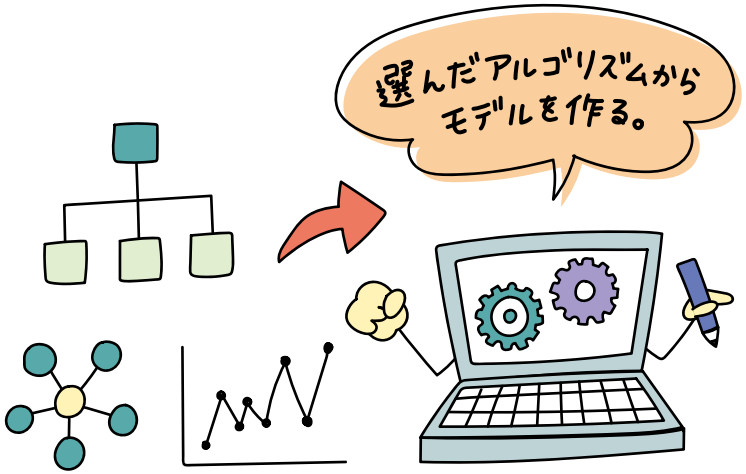
データサイエンティストにはどのような役割があるかということなんですが、例えば単純にデータを統計学的に何かやればいいでしょうという話ではないんですよ。データサイエンティストというのは問題解決をして、それが評価に繋がっていくわけなんですね。したがって「（④　　）」、こういったことが大事になってきます。

また、データに対してどういうふうに加工したらいいかどうか（向き合い方）といったところもチェックをしていくということになってきます。



また、図1.4にアルゴリズムの選定やモデルの構築と書いてありますけれども、アルゴリズムというのはここでは分析の方法というふうに考えていただければ結構です。分析の方法にはどういった方法がいいのかを選定をしていく。例えば何か予測をするというようなことであれば、一つの予測するための仕組み＝ロジック、このロジックのことをここではモデルというふうに呼んでいるんですけれども、そのモデルを作っていくといったところが仕事になります。

さらには分析結果に対しての評価というようなことになってくるわけなんですね。



研究分野でもデータサイエンティストというような職業はあるわけなんですけれども、データ分析の研究になってくると新しい技術に取り組んでいくといったところが必要になってくるんですが、実務では先ほど申し上げたように課題解決がポイントになってくるので、データを分析していって何かこういうような指針があるんじゃないかといったところを出すことがポイントになってくるわけですね。

「そういうような見方があったのか」と思ってもらえることに、データサイエンティストの価値があるというようなことになります。これは今回の講座でも一つポイントになってくるところで、単純にデータを分析すればいいというわけではなくて、何度も申し上げますが、どういうふうに使っていくのか・どういうような問題が解決できるのかがポイントになります。

MEMO

1章のまとめですけれども、日常生活ではデータがあふれていてデータサイエンスの重要性が高まっています。それはなぜか、なぜそうなったかというと、インターネットだとか（⑤　　）の技術を使ってデータを素早く入手できるようになったからです。一方で、データを正確に解釈する必要性というのも出てきています。

また、データサイエンティストのスキルや役割を確認していきましたが、これは単なる統計学ではないですよということ、これが本当に大切です。「統計ですよね？」で終わってしまうと駄目なんですね。なのでビジネスでは「意味のあるデータと捉えるためのスキル」というふうに考えていただきたいと思います。

第1章の振り返り

日常生活にはデータがあふれ、データサイエンスの重要性が高まっています

インターネットやクラウドなどが普及し、データをすばやく入手できるようになりました。
一方、データを正確に解釈する必要性もでてきました。

データサイエンティストのスキルや役割を理解しましょう

単なる統計学ではありません。ビジネスや意味があるデータと捉えるためのスキルも必要です。

Data Science

第2章 データの見方