文系のためのデータサイエンス

DX時代の歩き方

大塚佳臣 著

日科技連

はじめに

ICT（情報通信技術）の発展に伴い、膨大なデータが世の中に溢れるようになり、さまざまな分野でデータの利活用が急速に進められています。データは「新しい石油」、「新しい通貨」と表現されることもあるなど、データサイエンスの活用は、エンジニアだけでなく、すべてのビジネスパーソンにとって必要不可欠になってきています。

本書は、主に文系の大学生や高校生を対象として、データサイエンスの概念や手法、その活用方法を紹介することで、データとデジタル技術で社会の変革をめざすDX時代を生きていくうえで誰もが身につけるべき発想や考え方を学ぶことを目的としています。

データサイエンスというと理系というイメージが強く、文系には関係がない、ハードルが高いという印象があるかもしれませんが、データを活用して意思決定するというごく普通の営みを、カタカナでカッコよく表現しているだけです。少し違うところがあるとすると、単なる意思決定のためだけでなく、新しい価値を創造するという概念が加わっていることです。これらは文系・理系を問わず必要な能力です。本書では、DX時代を生きるうえで必要となる、データを使って意思決定をする、新しい価値を創造するという能力を身につけるために必要な考え方やスキルを紹介していきます。

また、データサイエンスというと、統計解析やAI、プログラミングをイメージする人が多いと思います。本書では、それらの理論の詳細や具体的な分析方法は解説していません。本書を読んで、データサイエンスに強い興味をもって、理論を深く学びたい、実際に分析をしてみたいと思うようになったら、他の専門書を紐解いて欲しいと思います。

［主に大学の先生方へ］

データサイエンスという概念が脚光を浴びるようになり、政府は2025年までにすべての大学・高専生がデータサイエンスの初級レベルを習得する目標を掲げており、各大学でデータサイエンス科目の設置が進められています。そのような動きを受け、データサイエンスに関する教科書も多く出版されるようになりましたが、その多くは数学や統計学の知識を必要としており、文系学生にはハードルが高くなっています。そこで、本書では、数学や統計学の知識をもたない学生でも、データサイエンスの概念や活用方法を理解できる内容としました。本書は、主に大学生のボリュームゾーンを占める私立文系の学生を対象とした、教養（初級レベル）としてのデータサイエンスの授業の運営にご活用いただくことを想定した構成になっています。

［謝辞］

本書の執筆にあたり、講義での質問を通じて、内容のわかりにくいところ・わかりやすいところ、知りたいことに関するフィードバックを寄せていただいた東洋大学の履修生に感謝いたします。また、日科技連出版社の鈴木兄宏氏、石田新氏には、企画の段階からご尽力いただきましたことにお礼申し上げます。

2023年1月

大　塚　佳　臣

目　次

【コラム】

★本書のデータ・分析手順の解説のダウンロード方法

本書で使用しているデータならびに分析手順の解説は、日科技連出版社のWeb サイト（https://www.juse-p.co.jp/）からダウンロードできます。実際に分析もしてみたい方はご活用ください。また、データサイエンス関係の授業や演習を担当されている先生方も例題としてご活用ください。

ID：[redacted]
パスワード：[redacted]

★動作環境

上記データは、Windows 版 Excel がインストールされているパソコンでの利用を対象としています。任意の環境で動作することを保証しているわけではありません。

★免責事項

著者、および、出版社のいずれも、上記データを利用した際に生じた損害についての責任、ならびに、サポート義務を負うものではありません。

第2章
データサイエンスのデザイン

2.1　データサイエンスのワークフロー

データをもとに、新しい価値を創造するうえで役に立つ法則を見つけるための実際の作業手順(ワークフロー)は、図 2.1 のようになります。

最初に、データを集めます。販売データ、国や自治体などが整備している統計データのようにすでに集まっているデータもあれば、温度計や圧力計のようなセンサーで集めているデータ、アンケート調査によって集めたデータもあります。また、SNS で発信されている文字もデータです。

これらのデータはさまざまな形式で蓄積されますので、分析可能な形に加工する必要があります。分析が可能な形式に加工できたら、分析を行います。分析、というと難しいことをしている印象を抱きがちですが、それを単純に集計して表やグラフにするだけで役に立つ法則が得られることも数多くあります。これも立派な分析です。もし、これだけでは役に立つ法則が得られない場合は、統計解析や機械学習と呼ばれる手法を使って法則を見つけます。

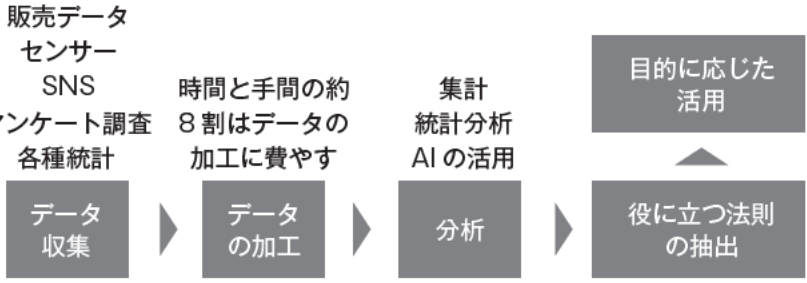

図 2.1　データサイエンスのワークフロー

実際の作業プロセスはこのようになりますが、こんな疑問が湧いてきたのではないでしょうか。「どういうデータを集めて、どういう分析をしたら役に立つ法則が見つかるのだろうか」と。

AI(人工知能)という言葉や概念が一般的に知られるようになり、AIにデータを入れたら欲しい結果を自動的に出してくれる、というイメージがあるかもしれません。実際にデータを使って新しい価値を創造しようとしたとき、データを闇雲に分析しても役に立つ法則が見つかるわけではありません。新しい価値を創造するためには、ゴール(創りたい価値)を設定し、ゴールを見据えて作業を行っていく必要があります。データサイエンスのデザインの概念を図2.2に示します。データサイエンスを使って新しい価値を創造するには、まず、どういう価値を創りたいのかを設定します。次にその価値を手にするうえで、どういう法則が見つけられればよいのかを明確にします。そして、その法則を見つけるにはどういう分析をすればそれが得られるか、そのために必要なデータは何かを考えたうえで、データを集めます。

データが集まったら、分析可能な形式に加工して、いよいよ分析します。分析すればすぐに役立つ法則が得られるとは限りません。試行錯誤しながら分析を繰り返して、最初に設定した価値の創造に資する、役に立つ法則を抽出します。

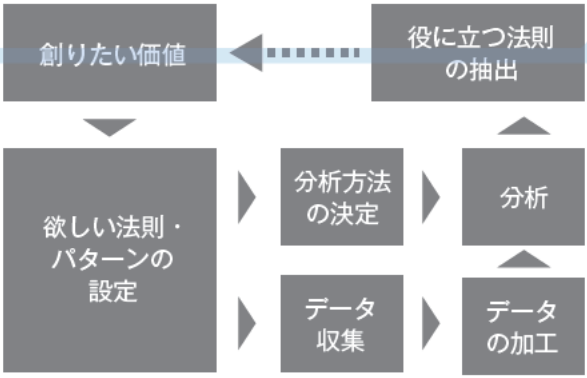

図2.2　データサイエンスのデザインの概念

2.2　データサイエンスを料理に喩える

ここでは、データサイエンスの実践における各プロセスを料理に喩えて説明します。

(1)　何を満たしたいのか【創りたい価値の設定】

料理は何のためにするのでしょうか？　多くの人は、食べるもの・食べたいものをつくるために料理をすると思います。では、何か料理をしようと思ったとき、何を満たすためにつくるのでしょうか？　おいしいものを食べたい、お腹いっぱい食べたい、健康になりたい、安く済ませたい、早く食べたい、リッチな気持ちになりたい、みんなで楽しく食べたい、などさまざまな目的があると思います。これらが料理によって創りたい価値になります。データサイエンスにおいても、最初に創りたい価値を設定しましょう。

(2)　何をつくるのか【欲しい法則の設定】

そのときの気分に応じて、お腹いっぱい食べたい、おいしいものを食べたい、といった目的を設定したら、次に、何をつくるかを決めると思います。お腹いっぱい食べたいと思ったら、カツ丼やカレーライスなどが頭に浮かぶかもしれませんが、お刺身は頭に浮かばないのではないでしょうか？　このように、目的に応じてつくるものを決めると思います。これが欲しい法則にあたります。データサイエンスにおいても、創りたい価値を設定したら、それを実現できる法則を設定しましょう。

(3)　必要な食材を集める【データ収集】

例えばトンカツをつくると決めたら、食材を揃えますね。卵、塩、ソースはあるけど豚肉はない、パン粉がない、あれ小麦粉がない、油がない、キャベツの千切りも欲しい、味噌汁もつくりたい、と足りない食材をお店に買いに行くことになります。データサイエンスにおいては、データが食材にあたり、欲し

い法則（つくる料理）に必要となるデータ（食材）を集めることになります。

(4)　調理方法を考える【分析方法の決定】

食材が揃ったらさっそく調理、といきたいところですが、「ところでトンカツはどうやってつくるんだっけ？」とならないよう、調理の前に段取りを確認しなければなりません。肉に下味をつけて、衣をつけて、揚げて…、という手順（レシピ）を先に整理しておく必要があります。また、揚げ物をするための鍋と、きっちりやりたい人、やったことがない人は、油温計の準備も必要かもしれません。

データサイエンスにおいては、分析が調理にあたりますが、分析（調理）の前に分析方法（段取り、必要な道具、調理方法）を事前に決めておく必要があります。

(5)　下ごしらえをする【データの加工】

トンカツを揚げる前に、食材の下ごしらえをします。肉を筋切りして、塩・コショウを振って、小麦粉をまぶして、溶き卵に浸けて、パン粉をつけます。キャベツの千切りも食べたければ、キャベツから、葉っぱを１枚ずつ外して千切り、あるいは半分・４分の１にカットして千切りにします。このように、買ってきた食材はすぐに調理に回せるわけでなく、下ごしらえが必要で、多くの場合、かなり時間と手間がかかります。データサイエンスにおいても同様で、データ（食材）を分析（調理）する前に、分析可能な形式にデータを加工（下ごしらえ）することが必要で、ここに多くの時間と手間を費やすことになります。

(6)　調理する【分析】

食材の調達から下ごしらえを経て、やっとトンカツを揚げるという調理に入ることができます。トンカツは油で揚げます。もしも、ガスコンロの魚焼きトレーで焼いてしまったら、トンカツになりません。データサイエンスにおいても、つくるもの（欲しい法則）に合わせて、調理方法（分析方法）を選んで調理

索　引

著者紹介

大塚 佳臣（おおつか よしおみ）

1969年生まれ。東京工業大学工学部有機材料工学科卒業。東京大学大学院工学系研究科都市工学専攻博士課程修了。博士(工学)、技術士(衛生工学部門、総合技術監理部門)。

1994年4月よりエンジニアリング会社にて、廃棄物リサイクルシステムの開発に従事。

2010年4月より東洋大学総合情報学部総合情報学科准教授。アンケート調査データ、統計データなどを活用した住民・消費者の環境心理、環境マーケティングの研究に従事。

現在、東洋大学総合情報学部総合情報学科教授。

文系のためのデータサイエンス
DX時代の歩き方

2023年2月23日　第1刷発行

著　者　大塚　佳臣
発行人　戸羽　節文

発行所　株式会社 日科技連出版社
〒151-0051　東京都渋谷区千駄ケ谷5-15-5
DSビル
電　話　出版　03-5379-1244
営業　03-5379-1238

検　印
省　略

Printed in Japan

印刷・製本　壮光舎印刷

ISBN 978-4-8171-9769-6

URL https://www.juse-p.co.jp/