

無断使用をお断りします。日科技連出版社

# 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

はじめに	iii
<b>第1章 データサイエンスとは</b>	1
1.1 データサイエンスとは	1
1.2 情報とデータ	2
1.3 データサイエンスがめざすところ	3
1.4 DX時代に求められるスキル	4
1.5 DXとは	6
1.6 なぜデータ“サイエンス”なのか	7
<b>第2章 データサイエンスのデザイン</b>	9
2.1 データサイエンスのワークフロー	9
2.2 データサイエンスを料理に喩える	11
2.3 各プロセスにおけるポイント	14
2.4 データサイエンスのデザインのポイント	18
<b>第3章 データマイニングとは</b>	21
3.1 データマイニングの定義	21
3.2 データ分析のアプローチの変遷	22
3.3 なぜデータマイニングなのか： ビッグデータに対する統計解析手法の限界	23
3.4 新しい分析手法(機械学習)は万能か？	24
3.5 統計解析と機械学習の役割	25
3.6 データマイニングとの付き合い方	27
<b>第4章 データの見方・見せ方</b>	29
4.1 データの見方・見せ方の基本的な考え方	29
4.2 度数データの見方・見せ方	31
4.3 数値データの見方・見せ方	38
4.4 データの見方・見せ方の戦略	40

4.5	やっではいけないこと、やるべきではないこと	47
<b>第5章</b>	<b>データ分析と検定</b>	51
5.1	統計的仮説検定とは	51
5.2	帰無仮説と対立仮説の考え方	53
5.3	度数データの検定	54
5.4	平均値の検定	55
5.5	どういうときに検定を行うのか	57
<b>第6章</b>	<b>データ同士の関連を知る</b>	61
6.1	「関連」とは	61
6.2	相関分析とは	62
6.3	疑似相関	65
<b>第7章</b>	<b>影響度を知る</b>	69
7.1	回帰分析とは	69
7.2	重回帰分析とは	70
7.3	偏回帰係数について	72
7.4	世界のデータを扱うときの工夫	73
7.5	重回帰分析に関するまとめ	77
<b>第8章</b>	<b>因果関係を知る</b>	79
8.1	構造方程式モデリングとは	80
8.2	モデルの作成方法	81
8.3	モデルの改善	84
8.4	システム全体で「小型軽量であること」の効果を考える	86
8.5	因果関係を活用した根本的な問題解決 (システムズ・アプローチ)	88
<b>第9章</b>	<b>データの収集と整備</b>	91
9.1	情報のデータ化の歴史	91
9.2	データの収集方法	92
9.3	データの整理・加工	97
9.4	データの整理と加工の難しさ	99

<b>第10章 AI・機械学習とは</b> .....	103
10.1 AI・機械学習の定義.....	103
10.2 機械学習の手順.....	105
10.3 機械学習の手法.....	106
10.4 どのようなときにどの手法を使うのか.....	112
<b>第11章 AIの活用事例</b> .....	115
11.1 スポーツのハイライト映像の自動生成：ATP/WTA.....	115
11.2 店舗オペレーションの改善：(株)あきんどスシロー.....	116
11.3 野菜の市場価格予測：(株)ファームシップ.....	118
11.4 採用選考の効率化：ソフトバンク(株).....	119
11.5 AIの活用にあたって.....	121
<b>第12章 データサイエンスとビジネス</b> —あびやのデータ駆動型経営とは—.....	125
12.1 背景.....	126
12.2 データサイエンスの第一歩(2013~2014年).....	127
12.3 データ分析の導入(2015年).....	129
12.4 データ収集の対策と機械学習による来客数予測(2016年).....	129
12.5 予測精度向上に向けたAIによる通行人の画像解析の導入 (2017年).....	130
12.6 業務の省力化に向けたDX化(2019年).....	131
12.7 経営への効果.....	132
12.8 データにもとづく冷静な経営判断の事例.....	132
12.9 サービス産業活性化への挑戦.....	133
12.10 おわりに.....	134
<b>第13章 DX時代の歩き方</b> .....	137
13.1 DX時代に求められるスキル.....	137
13.2 今後、社会や組織で求められる人材像.....	138
引用・参考文献.....	141
索引.....	145

## 【コラム】

統計の用語について .....	58
制約条件の理論 .....	90
QRコードによる電子決済を勧めたがるワケ .....	100
AIは人間の仕事を奪ってしまうのか？ .....	121

## ★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 : 

パスワード : 

## ★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。

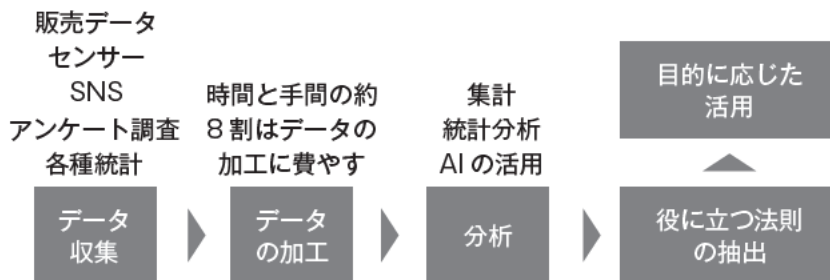


図 2.1 データサイエンスのワークフロー



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のでしょうか。「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。

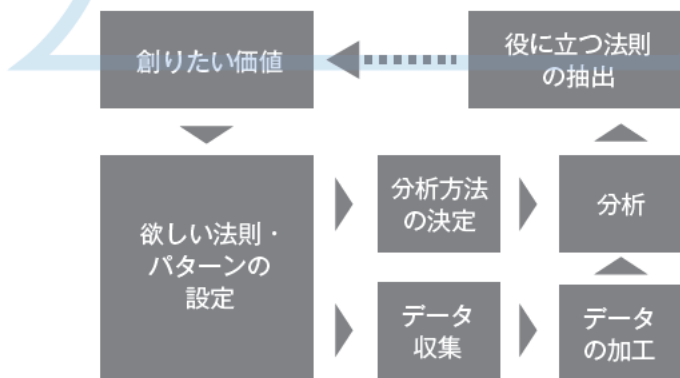


図 2.2 データサイエンスのデザインの概念

## 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のでしょうか？ 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はなく、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

- 【英数字】**
- 100% 積み上げ棒グラフ 33  
 AI 103  
 API 95  
   —エコノミー 96  
 CFI 81  
 CSV 形式 98  
 DX 6, 137  
   —時代 5, 6, 137  
   —人材 6  
 GFI 81  
 GPS 101  
 HTML 94  
   —タグ 95  
 HTTP/HTTPS 95  
 IC タグ 117  
 Internet of Things 92  
 IoT 92  
 Key Performance Indicator 132  
 KPI 132  
 NLC 120  
 $\rho$  値 54, 71  
 QR コード 101  
 RMSEA 81  
 Robotic Process Automation 129  
 RPA 129  
 SEM 80  
   —の適合度指標とその基準 81  
 Shift\_JIS 97  
 Structural Equation Modeling 80  
 Theory of Constraints 90  
 TOC 90
- $t$  検定 55  
 Unicode 97  
 Web API 95  
 Web スクレイピング 95
- 【あ 行】**
- 新しい価値 9  
 アナログ 91  
 アンケート調査 93  
 意思決定 27, 61  
 因果関係 79  
   —の構造化 80  
 インサイト 3  
 因子 26  
 印象操作 44, 49  
 インターネット 92  
 円グラフ 32  
 エンコーディングルール 97  
 オープンデータ 93  
 音声認識 113  
 オンラインアンケート 93
- 【か 行】**
- 回帰 107  
   —分析 70  
 カイ二乗検定 54  
 科学的アプローチ 7  
 学習器 106  
 隠れ層 110  
 可視化 30  
 風が吹けば桶屋が儲かる 79  
 仮説 7

—の立案 22  
 画像認識 113  
 価値の創造 18  
 間接効果 86  
 関連 61  
 機械学習 23, 25, 103, 130  
 棄却する 54  
 疑似相関 66  
 帰無仮説 53  
 強化学習 108, 113  
 教師あり学習 106, 112  
 教師なし学習 107  
 クラスタリング 107  
 グローバルデータ 73  
 クロス集計 34  
 群 56  
 結果 79  
 —系 81  
 原因 79  
 —系 81  
 検定 54  
 構造方程式モデリング 80  
 交絡因子 66  
 誤差 81  
 個人情報 96  
 —保護法 93  
 個票データ 93

## 【さ 行】

最小二乗法 70  
 採択する 54  
 錯覚 48  
 散布図 39  
 サンプルング 52  
 サンプル 52, 58

—サイズ 59  
 —数 59  
 時系列データ 38  
 次元圧縮 108  
 システムズ・アプローチ 90  
 自然言語処理 113  
 自然言語分類 120  
 重回帰分析 72, 77, 129  
 出力層 110  
 需要予測 117  
 省人化 117  
 情報 2, 91  
 —の記録 91  
 —の伝達 91  
 植物工場 118  
 人工知能 103  
 数値データ 38  
 スマートフォン 92  
 正解ラベル 106  
 正規化 99  
 正の相関 62  
 制約条件の理論 90  
 センサー 97  
 全体最適化 90  
 相関係数 62  
 相関分析 62, 129  
 総合効果 86  
 ソリューション思考 138

## 【た 行】

対数 75  
 —をとる 75  
 対立仮説 53  
 多重共線性 72  
 多重比較 56

単純集計 31  
弾性値 77  
直接効果 86  
著作権 95  
創りたい価値 4, 11, 14, 18  
ディープラーニング 104, 110, 113  
データ 2, 29, 92  
——エンジニアリングスキル 5  
——活用スキル 5  
——駆動型経営 125  
——形式 30  
——収集 11  
——の加工 12  
——の収集・蓄積 92  
——の整理と加工 100  
——の二次利用 93  
——の発信 92  
データクレンジング 97  
データサイエンス 1  
——スキル 5  
——のアプローチ 7  
データサイエンティスト 6  
データマイニング 21  
テーブル 98  
適合度指標 80  
デジタル 91  
動画配信 115  
統計 92  
——解析 22  
——データ 92  
——的仮説検定 22, 51, 52  
特微量 105  
度数データ 31  
特化型 AI 103

## 【な行】

ニューラルネットワーク 110  
入力層 110  
ネットリサーチ 93

## 【は行】

パス 81  
——係数 80, 82  
バックキャストイング 19, 135  
罰則 108  
パラメータ 59  
判断 61  
汎用型 AI 103  
ビジネスインテリジェンスツール  
134  
ヒストグラム 74  
ビッグデータ 22, 103  
標準誤差 71  
標本 52, 58  
フィールド 98  
フォアキャストイング 19  
負の相関 62, 65  
部分最適化 90  
分析 9, 12, 30  
——方針 15  
——方法の決定 12  
分類 107  
偏回帰係数 71, 72, 77  
変数 105  
偏相関係数 66  
棒グラフ 32, 33  
報酬 108  
欲しい法則 11, 14, 18  
母集団 52, 58  
母数 59

**【ま 行】**  
マーケティング 101  
マイクロデータ 93  
無線技術 92, 96  
メインキー 98  
文字コード 97  
モデル 80  
——の確からしさ 80  
問題解決 88

**【や 行】**  
役に立つ法則の整理・抽出 13

有意確率 53  
有意水準 54

**【ら 行】**  
立体グラフ 49  
利用ポリシー 95  
レーダーチャート 40  
レコード 98

**【わ 行】**  
割合データ 33



## 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 東京都渋谷区千駄ヶ谷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

© Yoshiomi Otsuka 2023

ISBN 978-4-8171-9769-6

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

本書の全部または一部を無断でコピー、スキャン、デジタル化などの複製をすることは、著作権法上での例外を除き禁じられています。本書を代行業者等の第三者に依頼してスキャンやデジタル化することは、たとえ個人や家庭内での利用でも著作権法違反です。