はしがき

データサイエンスを学ぶということはどのようなことなのでしょうか。Python や R といったプログラム言語を学び，それらのプログラムにあるパッケージを用いて分析することでしょうか。もちろん，プログラムやパッケージを用いて何らかのデータ分析を行う作業スキルを身に付けることは重要ですが，データサイエンスがどのような背景で誕生し，何を意味するのか，従来のデータ分析の手法とどのように異なるのかといったことを理解することも重要です。ただ，これまで出版された書籍では，どちらかというと分析手法に焦点を当てたものが多く，先に述べたデータサイエンス全体の理解を進めるものはあまりありませんでした。

また，近年の機械学習の手法の発達には著しいものがありますが，統計学と機械学習の両方を一冊で学べる書籍というものもほとんどありませんでした。ビッグ・データの時代になり，社会のさまざまなセクターでデータ分析が活用される現在では，データサイエンス全体の理解を進め，統計学と機械学習のエッセンスとそれを支える基礎的な知識を一冊にまとめた書籍が求められています。このことが本書を執筆する動機になっています。

本書は，大学でデータサイエンスを初めて学習する人向けの教科書として，また，大学や企業においてこれからデータ分析を行う人，データサイエンスを基礎から学びたい人向けに執筆しています。データ分析の初学者が，分析手法についての書籍や論文を読み進める際に必要となる，知識・考え方を理解できるような構成になっています。そのため，ベクトルやデータ空間といった，これまで入門的な統計学やデータ分析の書籍ではあまり扱われてこなかった内容や，必要に応じてネイピア数などの基礎的な用語についても説明しています。

本書は全 9 章で構成されていますが，第 1 章から第 3 章はデータサイエンスの定義やデータそのものについての理解を深めるための章です。例えば，データ分析の書籍には，データを空間に布置するという表現がときどき出てきますが，そのイメージが把握できるように，データと空間についても取り上げています。第 4 章から第 8 章は実際にデータを分析する際に必要な知識を体系

的にまとめたものです。最近の機械学習がデータサイエンスの領域に与えた影響を考慮し，機械学習の発展的な書籍，ならびに学術論文を円滑に読み進めることができるようまとめています。第 9 章はデータサイエンスの応用知識だけでなく限界についてもまとめています。データサイエンスは万能ではありません。限界を理解して活用することが求められます。また，データサイエンスの時代には，調査データを扱っていた時代とは異なる新しい問題が生じます。これらの問題に対しても目を向ける必要がありますが，データサイエンスの入門書で，このような問題を取り上げた書籍はあまり見かけないため，この最後の章で扱っています。

本書は，第 1 章から 1 つずつ章を読み進めるという読み方をしていただいても，第 1 章と第 9 章を読み，データサイエンスの概要を理解し，それから各章を読み進めていただいても構いません。また，自身の興味のままに好きな箇所から読み進めても問題ありません。本書はデータサイエンスの入門書ですので，読み終えた後に，自身の知識が足りない領域，さらに詳しく知りたい領域が理解できれば，どのような読み方をしていただいても問題ありません。また，本書を読み終えた後，どのような形でも良いので，実際のデータを分析してもらえれば，著者としてこれ以上の幸せはありません。

本書は，著者だけの力で書き上げたものではありません。早稲田大学データサイエンス研究所に所属する研究者や共同研究先の担当の方とともに議論したことが，本書の土台となっています。執筆した原稿については，横浜市立大学の越仲孝文先生，坂巻顕太郎先生に初稿を読んでいただき，さまざまな有意義な助言をいただきました。第 6 章の「データ分析の手法」と第 8 章の「データの分析事例」における具体的な分析や作図では，早稲田大学大学院創造理工学研究科博士後期課程の阪井優太氏，修士課程の良川太河氏にかなりの手助けをしていただきました。

また，横浜市立大学データサイエンス学部の有志の学生の皆さんには，原稿に対して読み手の視点で意見をもらいました。さらに，有斐閣の岡山義信氏には，本書の企画段階からさまざまなご尽力をいただき，著者の作業の進捗がはかばかしくない状況でも常に適切な助言と励ましのお言葉をかけていただきました。岡山氏の適切な助言，提案がなければ本書の出版は難しかったと思われ

ます。この場を借りて，ご協力をいただいた皆様には心から感謝を申し上げる次第です。

2022 年 11 月

上田　雅夫
後藤　正幸

■ウェブサポートページのご案内
以下のページで本書で紹介している分析例の再現方法など，情報を提供しております。ぜひご活用ください。

http://www.yuhikaku.co.jp/books/detail/9784641166110

目　次

第 1 章

データサイエンスとは

この章では，データサイエンス全体の理解を進めるため，データサイエンスの定義を明らかにし，その後で，データサイエンスの特徴，ならびに，その適用領域について説明します。同時に，なぜ，今データサイエンスが注目されているのか，データ環境の変化という点から解説をします。

1 データサイエンスとは

●── データサイエンスとは何か

次頁の図 1.1 は，Google Trends[1] を用いて「Data Science」というキーワードを，検索地域を「すべての国」，検索期間を 2010 年～2019 年に設定し，検索した結果です。データサイエンスという言葉が，2013 年から右肩上がりに上昇しており，データサイエンスという言葉の関心が，年を経るごとに高まっていることが理解できます。ただし，データサイエンスという言葉は，これだけ注目を集めていますが，その言葉の定義が曖昧であると指摘されており（Press 2013），データサイエンスが何であるか，明確に説明できる人は，残念ながらそれほど多くありません。そこで，データサイエンスを深く理解するために，まず，データサイエンスとは何かを明らかにする必要があります。

データサイエンスという言葉の定義には，Dhar（2013）の「データから一般化できる方法で知識を得るための学問」という定義があります。Wikipedia でデータサイエンスを調べてみると，「データサイエンスは構造化あるいは非構造化データから知識やインサイトを抽出するための科学的な手法，アルゴリ

1 Google で検索されたキーワードの時系列変化を可視化するサービス。Google Trends のサイトから利用可能（https://trends.google.co.jp/trends/?geo=JP）。

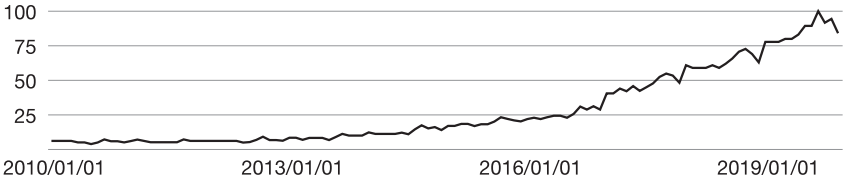

図 1.1　Google Trends の検索結果（“Data Science” すべての国）

ズムやシステムを用いる学際的な研究領域」（筆者訳）と定義しています。広辞苑では「統計学・計算機科学・情報科学などを応用し，各種のデータが持つ意味・法則性を探り出し，また，その手法を研究すること」と定義しています。

データサイエンスは文字通り，データ ＋ サイエンスの二語で構成されており，サイエンス ＝ 科学が何を示すのか理解できると，データサイエンスの理解が進みます。Oxford Learner’s Dictionary では，サイエンスとは「実験などで証明することができる事実をもとにした，自然や物質の構造や振る舞いについての知識」と定義しています。このサイエンスの定義を，データサイエンスに当てはめますと，「分析，調査，実験，観察で得られた事実（データ）をもとに，統計的構造の推定や予測，因果関係の評価などを通じて客観的に正しい意思決定や行動に結びつけるプロセスに関する知識」というように表現することができます。データから知識を導くには，何らかの手法で分析する必要があるため，先に述べた定義に分析に関する部分を追加すると，データサイエンスの定義は，「統計学や機械学習などの手法を用い，調査や実験などにより得られたデータの中に隠れた構造やデータ自体の動きを明らかにすることである」となるでしょう。

データは，何かの行動の結果得られるものです。調査のデータは，回答者が調査票の項目について回答した結果です。POS データ（Point of Sales Data：販売時点データ）は，人々の購買行動の結果であり，ソーシャル・メディアのデータは，人々が Facebook などのソーシャル・メディアにおける閲覧，投稿などの行動をした結果のデータです。単に，事実を確認するだけであれば，これらのデータを集計するだけで十分ですが，社会における人々の行動は複雑で，その複雑さを考慮すると何らかの分析を行う必要があります。POS データを用いて商品 A の販売実績を確認するのであれば，商品 A の日別の販売実

績を集計するだけで十分ですが，なぜ，商品 A が他の商品よりも売れたのかを理解するためには，何らかの分析が必要になります。表やグラフでも商品 A の売れ行きに影響を与える要因を理解することができますが，表やグラフでは限界があります。グラフで表現できるのは，3 軸までです。売上に 1 つを使うと，残りの軸は 2 つであり，要因は 2 つまでしか扱えません。また，グラフや表を用いて表現しても，それぞれの要因の純粋な定量的な影響度は不明です。そのために，統計学や機械学習の知識が必要となります。

ビッグ・データの時代になり，企業が扱えるデータの種類が多岐にわたるようになったことも，分析をするうえで，統計学や機械学習（含むデータマイニング）の手法に関する知識が，要求される一因となりました。例えば，ソーシャル・メディアのデータに含まれる，「フォローする」「フォローされる」という関係性のデータを用いて，影響力のある人を特定する場合，人数が少なければ手作業でも特定できますが，現在のソーシャル・メディアのデータの規模になると，作業を効率化させる何らかの方法を用いる必要があります。その目的のために，多様な方法が提案されてきました。

データの種類が増え，さまざまな分析手法が利用できる現在では，目的とデータに見合った手法の選択は，実務に活用できる情報をデータから得るうえで不可欠です。そのために，統計学や機械学習について学ぶ必要があります。例えば，クレジットカードのデータを分析し，顧客満足度を高めるには，会員の属性マスターと購買履歴データから，自社の売上に貢献している会員の特徴を理解する必要があります。クレジットカードなどの購買履歴データは，膨大な量になります。その膨大な量のデータから効率的に情報を得るためには，決定木という手法が適しています。また，E コマースの購買履歴データは，扱う商品の種類が多いため，疎のデータ（0 が多いデータ）になりやすく，そのようなデータから情報を得るため PLSA（Probabilistic Latent Semantic Analysis）のように次元圧縮の手法が発達しました。同じ目的に対し，さまざまな手法が提案されている現在では，目的とデータに応じた手法を選択することが，データサイエンティストに求められています。本書の第 8 章を参考に，どの分析手法で，どのようなことが理解できるかのイメージをつかむとよいでしょう。

加えて，データから何らかの情報を得るには，目的だけではなく，「データ」

にも注目する必要があります。本格的な分析の前に，集計やグラフによりデータの特徴を十分に理解したうえで（データに特徴を語ってもらったうえで），手法を選択することが多いでしょう。データを表やグラフで表すことを可視化（visualization）といいますが，こちらについても分析手法同様にいくつかの注意するべき点があります。データの内容を正しく読み手に伝えるには，その注意点を守る必要があります。

●── データマイニングとデータサイエンス

データサイエンスを考えるうえで，2000 年頃，ブームとなったデータマイニングについて理解すると，データサイエンスの特徴が，よく理解できるでしょう。データマイニングとは，大量の非集計のデータから，効率よく情報を収集する手法であり，古川・守口・阿部（2003）は，著書の中で「実務的には探索的非集計のデータ分析と予測の作業」と定義しています。データサイエンスとデータマイニングは，どちらも大量のデータから，実務に貢献する情報を得るという点で類似していますが，データマイニングがもてはやされた時代のデータと現在のデータでは，データ自体が大きく異なります。データマイニングの時代は，主に，顧客の購買履歴データを扱っていましたが，現在は，センサーやソーシャル・メディアを経由して得られるデータを扱うようになりました。購買履歴データでは，データは数値と文字（テキスト）として保存されており，数字や文字のデータを分析できれば十分でしたが，データサイエンスの時代になると，数字や文字に加え，音声，画像，位置情報といったデータも分析対象となり，現在では，目的に応じて幅広いデータを分析することが求められています。表 1.1 にあるように，購買履歴データとソーシャル・メディアのデータでは，データの種類，非構造化データ[2]の扱い，さらに，分析結果を活用する部署など，大きく異なり，データマイニングの時代と今とでは，データ分析に必要な技術ならびに活用領域が大きく異なることが理解できるでしょう。

2　ある統一したルールでデータベースに蓄積され，列と行でその内容を指定できるデータを構造化データと呼びます（第 2 章でもう一度言及します）。代表的な非構造化データとして Web のアクセス・ログがあります。

表 1.1　購買履歴データとソーシャル・メディア・データの違い

	購買履歴データ	ソーシャル・メディア・データ
概要	購買の記録	投稿・反応の記録（双方向の記録）
保存されるデータの種類	数字	文字，画像，動画，音声，反応の結果
非構造化データの有無	無	有
主に活用する部署	営業，マーケティング	営業，マーケティング，広報，宣伝

●―― データサイエンスの構成技術

データサイエンスに関し，「データサイエンス ＝ 分析技術」という誤解があります。データサイエンスは，Deep Learning などの分析手法のみを学習し，理解すれば，ビジネスなどの実務に役立つ情報が得られるというわけではありません。また，Python のように機械学習と親和性の高いプログラミング言語を習得することだけがデータサイエンスではありません。先に示したデータサイエンスの定義からも理解できるように，データサイエンスの目的は，データに埋もれた構造やデータの挙動を，データと活用の目的に見合った手法で明らかにすることです。そのためには，まず，分析するデータの特徴を理解することから始めます。

データの特徴を理解するうえで，平均値や分散などの基礎統計量の算出，表やグラフの作成などを行う必要がありますが，これらの作業をする前に，分析するデータには，どのような変数が含まれているのか，それらの変数はどのように収集されたのか（収集したのか），といった点を理解するところから始まります。ビジネスに利用するデータが，調査データが主流のときは，分析者は当該のデータについて，調査票の作成を通して分析する前から熟知していましたが，現在は，データウェアハウスに「蓄積された」データを分析することが多く，分析者はデータの内容を十分に理解する機会を経なくても，分析することができます。ただし，そのような態度で，手許にあるデータを分析することが，データから十分な情報を引き出せるという保証はありません。むしろ，デ

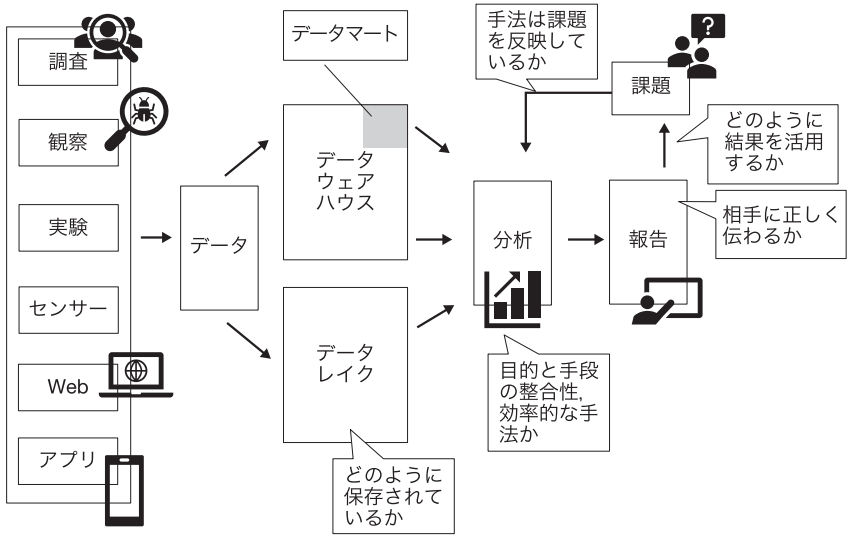

図 1.2　データサイエンスの構成要素

ータの内容を理解したうえで分析するほうが，分析を効率的に進めることができ（分析時の手戻りを少なくする），分析結果を読み取るうえでも，間違った解釈をすることなく進めることができるでしょう。

データを分析する目的は，何らかの課題を解決することであり，分析した後にもフォローする必要があります。例えば，分析した結果を組織で利用するために，関連部署へどのように働きかけるかも考えるべきです。図 1.2 にデータサイエンスを構成する各要素をまとめています[3]。図 1.2 にある通り，データサイエンスを構成する要素は，データの収集から分析結果の報告，活用までの道筋づくりまで含み，幅広い知識が求められます。

なお，データサイエンティストに求められる能力というと，分析力が真っ先に思い浮かぶかと思いますが，分析力だけでは十分ではありません。データを分析する目的は，先に述べたように，課題を解決することであり，分析結果を

3　データウェアハウス内のデータを分析する際，利用する部署によっては，分析メニューや必要な集計タイプが異なるため，あらかじめ，目的に応じてデータウェアハウス内のデータを小分けにしておくと効率的です。そのような小分けにしたデータを「データマート」と呼びます。

意思決定に用いることです。そのため，できるだけ素早く，分析結果を意思決定に活用できる形にすることが求められます。例えば，複数のカテゴリーの販売実績のデータを分析し，その結果をまとめる際は，出力された分析結果をコピーし貼り付ける作業を，プログラムを書き自動化するといったことが考えられます。分析に関する作業を行う際は，分析結果の加工といった，最終段階を見越して，合理的に作業を行うことができるように考えてから作業を行うべきでしょう。

●―― データサイエンスとデータ

POS データやソーシャル・メディアのデータが，現在のように企業の意思決定に用いられる以前は，調査データを主に用いて意思決定を行ってきました。ある目的に対し調査を設計・実施し，調査から得られたデータを分析していました。1980 年代の後半から情報システムの発達にあわせて，データウェアハウスに取引データ（代表的なものとして POS データ）を蓄積し，その蓄積されたデータを，意思決定に用いるようになりました。その後，企業のロイヤルティ・プログラム[4]の進展に伴い，POS データに個人が識別できる ID が付き，この ID 付き POS データが，小売業などの企業の意思決定に利用されるようになりました。この時代になると，データは分析の度に収集されるのではなく，データウェアハウスに蓄積されているデータから，分析の目的に応じて，切り出されるようになりました。

データウェアハウス内のデータは，あらかじめ決められた項目のみを保存するように設計されています。また，効率的に蓄積するため，データは正規化[5]されたテーブルとして保存されています。そのため，データウェアハウス内には，複数のテーブルが保存されています。実際のデータ分析では，複数のテーブルを結合し，分析用のデータを作成します。時には作成したデータを用いて，新しい変数を作成するといった作業を行います。分析用のデータを作成

4　顧客の利用実績に応じて，割引，優遇価格，ノベルティ（おまけ）などを提供するプログラムのこと。航空会社のマイレージ・プログラムがその代表です。

5　データの重複をなくして管理すること。商品の販売実績に商品名やサイズを記録し，商品マスターにも商品名やサイズを記録するのではなく，商品の販売実績には販売実績のみ，商品マスターには商品の情報のみを記録し管理することです。

するといったスキルは，調査データが主流であったときには，ほとんど必要とされていませんでしたが，現在では，データ分析に従事する人にとって備えていなければいけない必須スキルです[6]。加えて，データウェアハウスに蓄積されたデータを分析するうえで，調査データを主に分析していた時代には，顧みられることがなかった新しい注意点が生じました。1 つはデータの更新に関する点であり，もう 1 つは分析対象期間の設定という点です。

データウェアハウスに保存されたデータは，ある周期（日次や週次など）で更新されます。その際に，保存されていたデータの内容に関し更新が行われる場合があります[7]。例えば，ある商品の名称が変更され，その変更が過去にわたって更新されるため，その商品の過去の実績を分析する際に，以前の商品名称が残っておらず，その名前で分析できないということがあります。そのような問題を回避するためにも，データの更新情報については，分析前に確認する必要があります。後者については，データウェアハウスのデータは，同一の対象者（この場合は人ではない場合もあります）の時系列の記録ですが，適切な分析対象期間を設定しないと，正しい結果が得られないことがあります[8]。データウェアハウスに蓄積されているデータを分析するには，購買間隔などの消費の行動の特徴を理解して設定する必要があります。

ビッグ・データの時代になり，さまざまなデータが意思決定に利用できるようになると，従来のデータウェアハウスでは，対応できなくなり，新たにデータレイクという概念が提唱されました（Woods 2011）。これまでのデータウェアハウスでは，データ分析の目的にあわせて，収集するデータの構造を事前に決定し，その構造化されたデータを蓄積していました。この方式では，データの活用がデータを蓄積する動機となっており，データを効率的に利用できるという点で優れていますが（構造を決定したデータを蓄積しているため，分析を行いやすい），企業の意思決定に活用する，データの種類が増加した現在では，この方法では自社の課題に十分に対応できない場合があります。

データの種類が増え，更新頻度が早い現在では，データの利用目的を考え，

6　本書では，第 3 章の「データ空間の構成法」で説明します。
7　商品や顧客情報をまとめたマスターも更新され，これらの管理も必要です。
8　詳しくは，第 2 章で図 2.7 を用いて説明していますので，そちらを参照してください。

その目的に沿ってデータウェアハウスを設計していては時間が掛かり，検討している間に，重要なデータが蓄積できないという問題が生じる可能性があります。その問題に対応するためには，まずは必要と思われるデータを蓄積し，分析する際に，データを分析の目的に応じて，蓄積されているデータを加工し，分析するほうが効率的な場合もあります。そのために，データをまず保存しておく，容れ物が必要になり，システムに組み込まれるようになりました（この容れ物がデータレイクです）。

また，データウェアハウスに保存されているデータも，利用頻度が高いものもあれば低いデータもあります。頻度が高いデータは別に管理する，使用目的に応じて管理するなど，データウェアハウス全体を，細分化して管理することが効率的な場合があります。その際，管理しやすいように小型化した（もしくは一部の）データウェアハウスをデータマートと呼びます。現在は，大量のデータをさまざまな部署で利用し，意思決定に用いています。そのため，意思決定が迅速に行えるように，データを扱う仕組み自体も，時代にあわせて変更する必要があります。

●―― データサイエンスの特徴

データサイエンスは，データを分析し，意思決定に利用することを 1 つの目的にしていますが，従来の統計学やデータマイニングと大きく異なる点もあります。これまでは，データの背景にあるメカニズムを明らかにすることが主な目的でしたが，データサイエンスの時代では，敵対的生成ネットワーク（Generative Adversarial Networks：GANs[9]）に代表されるように，データを新たに作り出すことも目的の 1 つになっています。統計学，特に時系列分析では，1 から t までの期間のデータを用いて，モデルを作成し，$t+1$ 以降のデータを予測します。この $t+1$ 以降の予測値はある意味，モデルから生成されたデータではありますが，応用される領域の広さで考えますと，今の生成モデルほどではありません。

また，分析に用いるデータからも，データサイエンスの特徴を示すことがで

9　用意されたデータの特徴を学習し，新たなデータを作成する手法です。Goodfellow et al.（2014）によって提案されました。

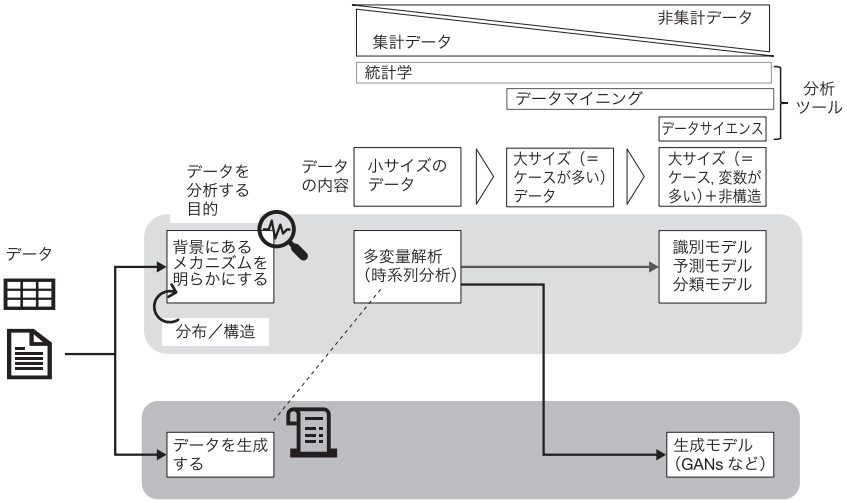

図 1.3　データサイエンスの特徴

きます。これまでのデータ分析では，分析のためのマシン（計算機，コンピュータ）の性能や，データ収集のコストなどのさまざまな制約があり，データのサイズは小さく，その小さなデータを分析するための手法が検討されましたが，先ほど説明したような情報技術の発達によって，データのサイズが大きくなり，同時にデータ自体も構造が定まらない非構造なデータ[10]が増えました。そのようなデータを分析するのに，統計学以外にもさまざまな手法が検討・提案され，図 1.3 で示すように，現在のデータサイエンスを形作っています。図 1.3 で重要な点は，現在のデータサイエンスが，統計学，データマイニングの上に成立しており，これらの知識がないと正しく活用できない（目的に応じた適切な結果を得ることができない）という点です[11]。

●—— データサイエンスの適用領域

社会活動，特に，企業活動においてデータサイエンスが注目される理由に

10　非構造なデータについては第 2 章で説明します。

11　この点については第 4 章の「データ生成のメカニズム」で，統計学の確率分布と現在の生成モデルとの関係を例に取り上げて説明します。

は，次の 2 点があります。1 つは，先に述べたようにデータの種類が，飛躍的に増加した点です。もう 1 つは，企業にはデータを利用し，より効率のよい活動を行い，改善に結びつけたいという潜在的なニーズがある点です。これまでは，企業のさまざまな活動を測定するデータは，技術的もしくは費用的に収集することが難しかった，または，そのようなデータ自体がありませんでした。現在のように豊富なデータが身近にあれば，データを活用するための手法の検討や，データ活用を業務フローへ組み込もうとする動きは自然なことです。

例えば，広告代理店にとって，ラジオ，テレビ，新聞，雑誌，インターネットといったメディアに適した広告を配信することは，業務の根幹をなすことです。ただし，ラジオについては適切な広告配信を行うことはテレビほど容易ではありませんでした。その理由は，テレビの視聴率のように，継続して利用状況を記録したデータを得ることが容易ではなかったことが原因です（消費者調査では，視聴率のデータのように高頻度に行うことは，コスト面で難しいという現状があります）。しかしながら，現在では，インターネットラジオ「radiko（ラジコ）」の視聴履歴を用いて，テレビの視聴率のようにラジオの利用状況が定量的に理解できるので，ラジオの広告効果の測定を行い，広告配信の効率化を行っています[12]。

データサイエンスは，企業の広告以外の業務における効率化にも貢献しています。新卒採用は人事部の採用担当にとって，大変負担の大きい業務です。コニカミノルタ社では，新卒採用時の適性検査から社員を 6 つのセグメントに分類し，セグメント別に入社後の行動を追跡し，どのセグメントの人材を拡充するかの方向性を明らかにし，採用担当者の負担の軽減に努めています[13]。また，活動量のデータを用いてコールセンターの生産性を管理し，生産性の向上に貢献できるという報告もあります（渡邊ほか 2013）。図 1.4 にあるように，現在では，企業の活動の中心にデータサイエンスがあり，企業の各部署，企業の内部活動と外部活動のそれぞれについて，データ分析した結果を意思決定に利用することが可能です。

図 1.4 に挙げたデータサイエンスを活用する業務については，調査データを

12　『日経流通新聞』2018 年 11 月 23 日付。
13　『日経産業新聞』2018 年 10 月 22 日付。

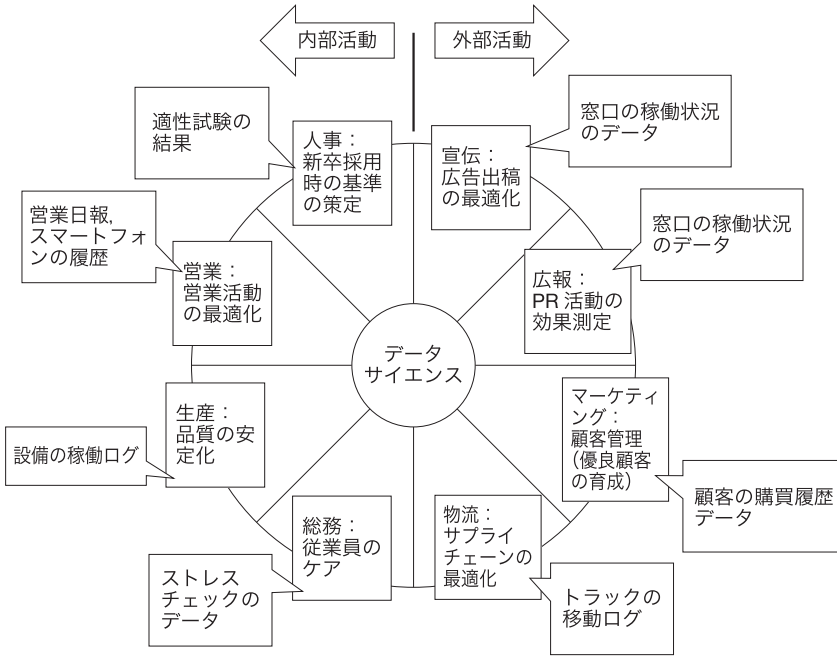

図 1.4　企業内の業務とデータサイエンスの利用

用いることも可能です。調査データは，調査項目を工夫することで，多様な内容のデータを収集することができますが，コストが掛かるために回数を制限せざるを得ないといった課題や，調査で収集できる内容は回答者の記憶に保存されている内容だけであるといった課題があります。さらに，調査では分析者が必要とする情報を収集することは容易なことではなく，調査票の設計の段階から注意する必要があります。そのため，マーケティング・リサーチのテキストでは，調査票の作成方法について少なくないページを割いています。

POS データが，マーケティングに活用されるようになった理由に，調査データに比べて品質やコスト面で優れているという側面があります。調査で消費者から過去に購入したものを聞き出すことはできますが，半年前の買い物について正確に答えられる人は，ほとんどいないでしょう。一方，POS データは，レジでスキャンされた結果がそのまま蓄積されるため，消費者の購買行動がデータとして正確に記録されています。また，POS データは小売店の日々の決

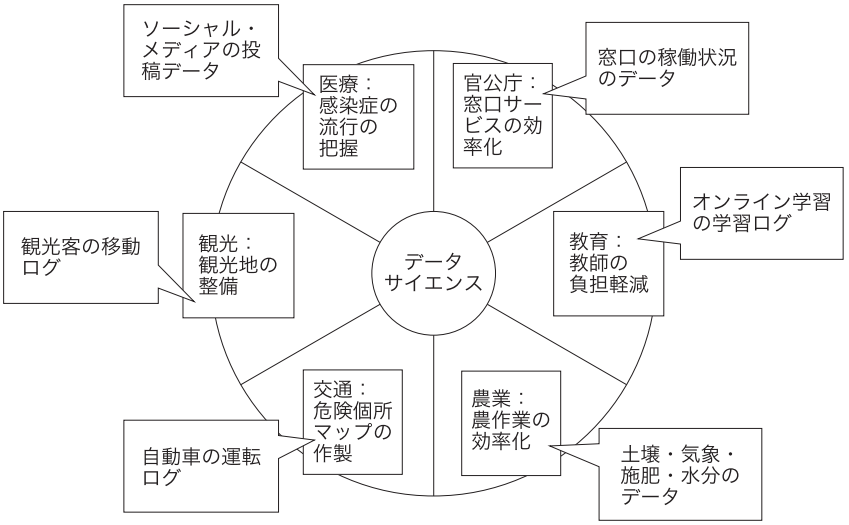

図 1.5　社会の各セクターとデータサイエンスのテーマ

済の結果として蓄積されるため，収集に関しては調査のようなコストが掛かりません。品質やコストの面で優れた代替データがあれば，今使用しているデータから代替データにシフトすることは当然でしょう。

データサイエンスは企業だけではなく，官公庁，医療，公共サービス（交通など），文化，学校，農業，観光などの社会の幅広いセクターにおいても活用されています。例えば，本田技研工業のインターナビは自動車の走行中のデータをインターネット経由ですべて収集しています。収集されたデータの中には，急ブレーキを踏んだ箇所についてのデータがあり，急ブレーキの多い箇所と現場の状況を照らしあわせることで，より安全な道路環境へと改善することができます[14]。交通事故を減らすためには，道路の状況を改善することは有効な手段の 1 つですが，道路の現状について，常に最新の状況に更新されるデータがなかったため，危険な場所を明らかにし改善することは困難でした。しかしながら，急ブレーキの情報をカーナビ経由で収集することができるようになったため，そのデータを用いて道路状況の改善をすることができるようになり

14　『日経コンストラクション』2013 年 8 月号。

ました。同時に，データサイエンスが公共サービスにも活用可能という理解が広がりました。このように，収集されるデータの種類が多くなればなるほど，社会のさまざまな領域，セクターでデータを利用することができるようになります。

社会の各セクターとデータサイエンスが活用できるテーマ例をまとめると図 1.5 のようになります。この図 1.5 から，データサイエンスが日々の生活に関するセクターの中心にあり，さまざまな場面で活用可能であることが理解できるでしょう。

●―― データサイエンスに期待されること

組織（企業，公的団体など）の業務は，以下の 3 つに大きく分けることができます。

- より大きなアウトプットを得るための業務（効果を生み出す業務）
- アウトプットに至る過程を改善する業務（効率を最大化する業務）
- リスクに対応する業務

この 3 つの中でデータサイエンスを活用できる業務は，主に後者の 2 つです。

効果を生み出す業務とは，0 を 1 にする業務であり，例えば，既存の市場を一変するような創造的な（時には破壊的な）商品・サービスの開発です。このような業務については，担当者の気づきやひらめきが必要とされるだけではなく，新しい商品・サービスを作り出すという動機が要求されます。そのため，データサイエンスにとって苦手な領域です。過去の結果であるデータからは，過去の延長のような結果が得られやすいという面があります。スマートフォンがまだ世にないときに，携帯電話の利用ログのデータを分析しても，現在のスマートフォンのような携帯電話が開発できたとは考え難いでしょう。フィーチャーフォンのデータからは，恐らく，フィーチャーフォンの機能を改良するなど，フィーチャーフォンの延長線上の製品しかできないからです。

効率を最大化する業務は，1 を 100 にする業務であり，例えば，工場の生産

性の向上などがこれにあたり，データサイエンスを適用しやすい領域です。データサイエンスが，効率性を上げる業務に適用しやすい理由は，現在のデータ環境によるところが大きいでしょう。インターネットが普及し，さまざまな活動がネットを経由してデータとして収集できるようになりました。ネット経由で収集されたデータは，企業や公共団体といった組織の活動の結果であり，活動の巧拙が結果としてデータに保存されます。そのため，収集されたデータを分析することで，活動の巧拙の原因（活動の課題）を発見し，それをもとに現状の改善活動を行うことができます。

企業の活動において，将来の予測を立てることは容易ではなく，常にリスクが伴います。未知のリスクに対応するには，過去の事例が参考になります。そのため，過去の事例が多ければ多いほどリスクの低下を図ることができます。また，蓄積する過去の事例については，その質も重要です。蓄積する事例は，情報の欠損が少なく，高い品質を保持していることが望ましいでしょう。人間は経験した過去の事例を，記憶として蓄積していますが，1 人の人間が蓄積できる経験の量には限りがあります。また，人間は過去の事実を，時間の経過とともに忘却してしまうため（記憶をそのままの形で維持することも難しい），事例の質という面でも問題があります。事例の量および質について考慮すると，蓄積されたデータを用い，将来のリスクに備えることが望ましいでしょう。

データから有用な情報を得ることは，分析を通じて過去の事例を，誰でも利用できる結果に変換することでもあります。データサイエンスにより，これまで，個人や組織の一部で所有していた知識を，定量的な分析を経て，組織内の全員に共有することも可能となるでしょう。知識の量・質が企業の競争力や公共団体が提供するサービスの質を左右する現在において，データサイエンスに注目が集まるのは，ある意味自然なことでしょう。

データを分析した結果を提示したときに，関係者から「そんなことはすでに知っている」と言われることがあります。データ分析が，人々の経験を分析した結果として，誰もが理解しやすい形に変換していることを考えると，この反応は，ある意味当然の反応でしょう。むしろ，注意する必要があるのは，自分の経験と異なる結果，新しい発見があったときです。そのようなときは，分析の過程でどこか手違いが生じていないかなど，結果を解釈するにあたり，慎重

に進めるべきでしょう。

2　本書の構成

この節では，本書の理解を進めるため各章の構成および概要について説明します。本書では，データサイエンスの総合的な理解を目的に，全部で 9 章の構成にしています。各章のタイトルは以下の通りです。

第 1 章　データサイエンスとは
第 2 章　データ収集のための基礎知識
第 3 章　データ空間の構成法
第 4 章　データ生成のメカニズム
第 5 章　データの可視化手法
第 6 章　データ分析の手法
第 7 章　データ活用のフレームワーク
第 8 章　データの分析事例
第 9 章　データ分析上の注意点と応用知識

各章の関係を図でまとめると図 1.6 の通りです。

●―― 各章の概要

本書を通して，データサイエンスを十分に理解してもらうには，最初から読み進めてもらっても，関心のある章から読み進めてもらっても大丈夫です。関心のある章から読んでみたいという人のために，ここでは，先に挙げた 9 つの章について，それぞれの章の概要を説明します。

第 1 章では「データサイエンス」が何を意味しているのかを定義し，本書におけるデータサイエンスの位置づけを明らかにしました。定義を明らかにした後，第 2 章～第 4 章で，データについて説明を行います。第 2 章～第 4 章は，データサイエンスの基礎として位置づけられ，データの定義からデータがどのように生成されるのかを説明し，データサイエンスを支えるデータについての

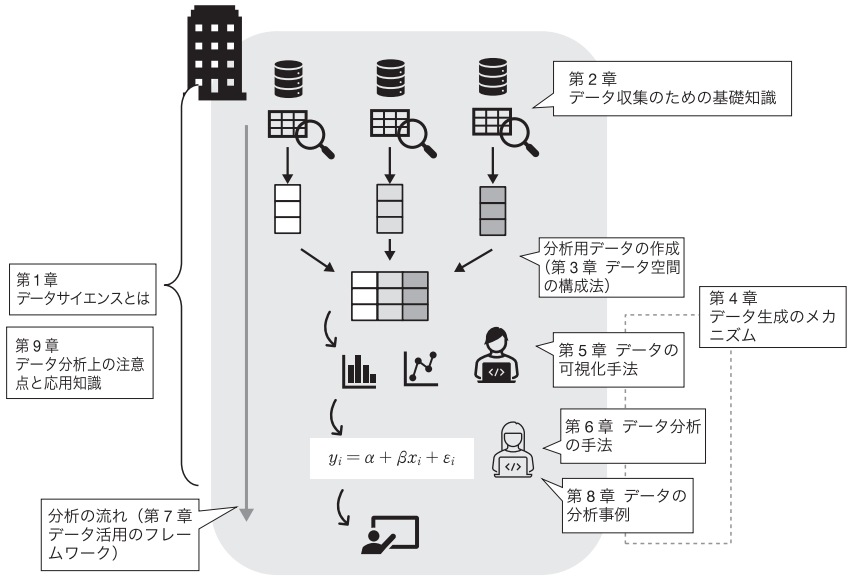

図 1.6　各章の関係

理解を深めます。データサイエンスはデータを支配する何らかの構造，ルールを明らかにするものであるため，データ自体について理解することは不可欠です。

図 1.7 では，機械学習のシステムが（図中の真ん中の黒い四角）が実社会において極めて小さい部分であり，関連する他のシステムのほうが大きく，無視できないことを示しています（Sculley et al. 2015）。例えば，機械学習の新しいアルゴリズムとデータの収集では，新しいアルゴリズムのほうが注目されている面が多々ありますが，実際には，データ収集もそれ以上に重要です。このことは，図 1.7 の「Data Collection」が大きく表示されていることからも理解できます。また，現在のデータ分析では，データウェアハウス上に保存されているさまざまなデータをつなぎ合わせ，分析用のデータを作成することが多く，そのプロセス自体の妥当性の評価は，欠かせません。分析した結果はデータの質に依存するわけですから，「Data Verification」が大きな面積を占めるのも当然です。

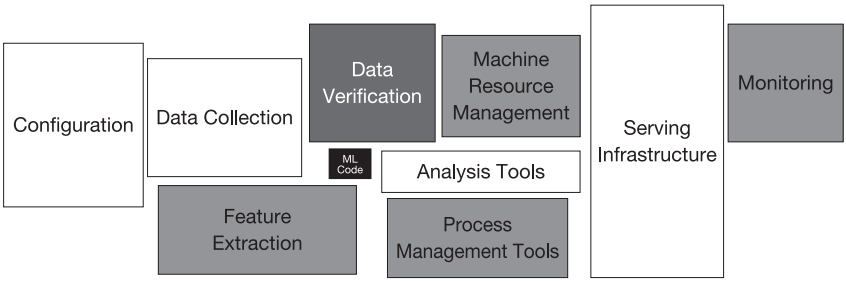

図 1.7　機械学習のシステムと関連するシステム（Sculley et al. 2015）

機械学習は，データを分析して，社会に貢献する何らかの示唆を導き出す道具ですが，分析した結果から意味のある示唆を得るには，意味のあるデータを分析する必要があります。また，よい結果を得るためには，データを収集する際に，いくつか注意するべき点があり，それらについては，第 2 章・第 3 章で説明します。また，収集したデータを分析する前には，データが重複していないか，異常値が含まれていないかの確認を行い，重複しているデータは削除し，異常値と思われるデータについては，分析に含めるか否かを慎重に検討する必要があります。これらの作業はデータクリーニング（もしくはデータクレンジング）と呼ばれ，第 3 章において説明を行います。第 4 章は分析対象のデータがどのようなメカニズムで生成されるのかについて説明します。また，データが生成される背景には何らかの確率分布を考えて分析しますが，この章では分析に用いる代表的な確率分布についてまとめています。

第 5 章～第 8 章は，データサイエンスの活用を踏まえた，分析手法の説明を中心とした応用のパートになります。データを分析するには，まず，そのデータの特徴を，表やグラフから確認します。その際，表やグラフの特徴，利用する理由を理解しておく必要があります。第 5 章において，それらの説明を行います。データサイエンスは，科学の一領域ですが，自然科学と異なり，データの中にある構造や法則を発見することだけがその目的ではなく，データを分析して得られた結果を用いて，社会に働きかけることも目的の一部となります。社会の課題を解決する一助として，統計学や機械学習の手法を用いてデータの分析を行いますが，これらの領域において提案されている手法は膨大にあり，

これらの全体像を第 6 章で示します。データ分析を円滑に行うには，膨大な手法の中から手許にあるデータと当該の目的に見合った手法を選択できるかが，結果に大きな影響を与えます。そのために第 7 章において分析の目的と手法の関係について事例を交えて説明を行い，その理解を深めます。また，データサイエンスは目的を意識することが重要ですが，それは，得られた分析結果をどのように活用するかを考えてのことです。そのため，第 8 章において，読者の皆さんも利用可能なオープンソース・ライブラリ上で公開されているデータを分析した事例をもとに，どのように活用するのかといった点についても説明します。また，分析自体の進め方については第 7 章で説明します。

第 9 章においてはデータ分析上の注意点と応用知識についてまとめます。データサイエンスは統計学をベースに，高度に発展した情報技術や機械学習などの分析技術を融合して発展した研究領域です。その現状や将来性に関して説明します。加えて，AI の技術・適用できる範囲を説明しながら，データサイエンスの適用領域について説明します。人工知能（Artificial Intelligence：AI）の限界について述べられた書籍，*Artificial Unintelligence: How Computers Misunderstand the World*（Broussard 2018）や『AI にできること，できないこと』（藤本・柴原 2019）などが刊行されていますが，これは，AI を活用するには，ＡＩの限界を理解する必要があるからです。同じことは，データサイエンスにもあてはまります。ともすれば，データサイエンスを用いれば，組織の課題すべてが解決できると思われがちですが，データサイエンスの利用には，注意すべき点があり限界があります。データサイエンスにできることと，できないことを理解して初めて，データサイエンスによる課題解決を図ることが可能でしょう。

コラム①　データサイエンスと日本

1960 年代に，現代のデータサイエンスに通じる 2 つの書籍が日本で出版されました。1 つが，林知己夫（1964）『市場調査の計画と実際』（村山孝喜との共著）であり，もう 1 つが，田口玄一（1962）『実験計画法』です。林先生はカテゴリカルなデータに注目し，その分析手法である数量化理論を提案され，田口先生は製品の品質の向上という課題を解決する，独自の手法を提案されました。どちらも現在の

データサイエンスがめざしている内容です。

林，田口両先生が 50 年以上前に提案された手法は，現在の大学でも授業で取り上げられ，また，企業などの現場において活用されており，時代が経った今でもその内容は色あせてはいません。データサイエンスというと，Google などのイメージにより，海外のほうが進んでいる印象を持つかもしれませんが，1960 年代に，現代にも通じる研究成果が報告されている日本は，世界から決して遅れをとっているわけではありません。また，データサイエンスを扱った書籍として，共立出版からデータサイエンス・シリーズ（全 12 巻）が 2001 年から出版されており，林・田口先生の後もデータから科学的な手法で情報を得ることに対し関心があったことが理解できます（林先生は，その後 2001 年に『データの科学』という本も上梓されています）。

むしろ，このような土壌があったからこそ，データサイエンスに関連する研究者や実務家が増え，今の時代を形成していると考えられます。ただし，データサイエンスをこれからさらに発展させるために重要なことがあります。岩崎（2019）はその著書で，「データサイエンス ＝（統計学 ＋ 情報科学）× 社会展開」としており，理論研究と応用研究のバランスの重要性を指摘しています。データから得られた示唆を，どのように社会に展開するかを考えることも重要ですが，社会展開だけを考えるデータサイエンスは意味がありません。なぜなら，社会の課題を解決するには，複雑な課題について取り組まざるを得ず，難しい判断が求められます。その際は，基礎に立ち返り，原理原則を理解したうえで分析を行う必要があります。基礎のない応用はありません。基礎がしっかりして初めて社会の課題に取り組めるため，基礎的な研究や理論研究も，データサイエンスでは取り組むべきでしょう。

課　　題

① データを用いて，社会（ビジネスを含む）の課題を解決した事例を挙げましょう。
② ①で挙げた事例について，なぜ，データを用いて課題が解決できたかを考えましょう。
③ 身近にあるデータを 5 つ挙げてください。それら 5 つのデータを用いて，どのような課題が解決できるかを考えましょう。
④ ③で挙げた課題に対し，データ以外に必要なものがないかを考えてみましょう。

参考文献

岩崎学（2019）『事例で学ぶ！　あたらしいデータサイエンスの教科書』翔泳社。

藤本浩司・柴原一友（2019）『AI にできること，できないこと――ビジネス社会を生きていくための 4 つの力』日本評論社。

古川一郎・守口剛・阿部誠（2003）『マーケティング・サイエンス入門――市場対応の科学的マネジメント』有斐閣。

渡邊純一郎・藤田真理奈・矢野和男・金坂秀雄・長谷川智之（2013）「コールセンタにおける職場の活発度が 生産性に与える影響の定量評価」『情報処理学会論文誌』54（4）：1470–1479。

Broussard, M.（2018）*Artificial Unintelligence: How Computers Misunderstand the World*, MIT press.

Dhar, V.（2013）‘‘Data science and prediction,’’ *Communications of the ACM*, 56（12）：64–73.

Goodfellow, I., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio（2014）‘‘Generative Adversarial Nets,’’ *Advances in Neural Information Processing Systems*, 27.

Press, G.（2013）‘‘Data Science:　What’s The Half-Life of A Buzzword?’’ https://www.forbes.com/sites/gilpress/2013/08/19/data-science-whats-the-half-life-of-a-buzzword/#39294227bfd3

Sculley, D., G. Holt, D. Golovin, E. Davydov, T. Phillips, D. Ebner, V. Chaudhary, M. Young, J-F. Crespo, and D. Dennison（2015）‘‘Hidden Technical Debt in Machine Learning Systems,’’ *NIPS*’15: *Proceedings of the* 28*th International Conference on Neural Information Processing Systems*, Volume 2: 2503–2511.

Woods, D.（2011）‘‘Big Data Requires a Big, New Architecture.’’ https://www.forbes.com/sites/ciocentral/2011/07/21/big-data-requires-a-big-new-architecture/

著者紹介

上田 雅夫（うえだ まさお）
現職：横浜市立大学データサイエンス学部教授
主著：『マーケティング・リサーチ入門』（共著）有斐閣，2018年。『マーケティング・エンジニアリング入門』（共著）有斐閣，2017年。

後藤 正幸（ごとう まさゆき）
現職：早稲田大学理工学術院教授
主著：『ビジネス統計――統計基礎とエクセル分析』（共著）オデッセイコミュニケーションズ社，2015年。『入門 パターン認識と機械学習』（共著）コロナ社，2014年。『確率統計学（IT Text）』（共著）オーム社，2010年。

データサイエンス入門
――データ取得・可視化・分析の全体像がわかる

Introduction to Data Science

2022年12月25日 初版第1刷発行

著　者　上田雅夫，後藤正幸
発行者　江草貞治
発行所　株式会社有斐閣
〒101-0051 東京都千代田区神田神保町2-17
http://www.yuhikaku.co.jp/
装　丁　麒麟三隻館
印　刷　大日本法令印刷株式会社
製　本　大口製本印刷株式会社
装丁印刷　株式会社亨有堂印刷所

落丁・乱丁本はお取替えいたします。定価はカバーに表示してあります。
©2022, Masao Ueda, Masayuki Goto
Printed in Japan. ISBN 978-4-641-16611-0

本書のコピー，スキャン，デジタル化等の無断複製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の利用でも著作権法違反です。

JCOPY 本書の無断複写（コピー）は，著作権法上での例外を除き，禁じられています。複写される場合は，そのつど事前に，（一社）出版者著作権管理機構（電話03-5244-5088，FAX03-5244-5089，e-mail：info@jcopy.or.jp）の許諾を得てください。