データサイエンス再考

村　田　嘉　弘[1]
鈴　木　　斉

Abstract

Data Science is nowadays an important tool for researches, new product developments, creation of new businesses, and Artificial Intelligence. Many countries have found the power of Data Science, and are trying to educate students as Data Scientists. So Data Science has been introduced by journalism and introductory books. But the word 'Data Science' is often ambiguous. Its ambiguity gives rise to confusions among people. In this article, we define the notions of 'Data Science as a procedure' and 'Data Science as a field of study' to clarify the meaning of Data Science. And we report the near future of Data Science in connection with XAI (eXplainable AI) and Quantum Computer.

Keywords : Data Science as a procedure, Data Science as a field of study, Machine Learning, XAI, Quantum Computer

1．はじめに

「経済学」は人間の経済活動とその結果生じる経済現象や制度を対象とする学問，「経営学」は企業等の経営の実態と取り巻く諸環境を対象とする学

1　埼玉学園大学経済経営学部特任教授，長崎大学名誉教授

問,「物理学」は自然現象のうち物理的現象と呼ばれるものを対象とする学問。このように「○○学」というとその学問の研究対象が存在するのが通例であるが,「データサイエンス」という用語は「○○学」という用語と同列に語ることのできない部分がある。

「統計学」「計量経済学」「マーケティングリサーチ」のような学問としての市民権を得た用語と異なり，データ分析やコンピューターに関わる分野は一時的なバズワードと言ってもよいような用語が溢れて来た。KDD（Knowledge-Discovery in Databases），データマイニング，ビッグデータ，BI（Business Intelligence），SIS（Strategic Information System）等，マスコミを賑わし豊かな未来を描いた用語が次々と登場してきた。ではこれらの用語は一時的なもので消え去ったのかというと，実は「データサイエンス」という大きな流れの中に取り込まれている。そして「データサイエンス」なるものは意味する範囲がますます拡大し揺るぎない学問領域になりつつあるように見える。

本論文は「データサイエンス」について来歴と現況を俯瞰し，「データサイエンス」という用語の定義，および，向かう方向について論ずることを目標としている。少なくとも「『データサイエンス』という用語は人により意味が異なる」という現状からの脱却を試みることとする。そのため，本論文は以下のような構成からなる：

1．はじめに
2．データサイエンスの現状
　2-1．データサイエンスの歴史
　2-2．データサイエンス教育
3．データサイエンスの再定義
　3-1．プロセスとしてのデータサイエンス
　3-2．プロセスとしてのデータサイエンスの詳細

「データサイエンス」の重要性が指摘されながら，意味の混乱，理解の混乱が起きている現状を鑑み，現時点と当面の方向性を整理し無用な混乱を低減することに本論文が貢献できることを願う。

2．データサイエンスの現状

2-1．データサイエンスの歴史

情報系の用語が本来の意味とは異なって使われることが多々あるのと同じように，「データサイエンス」という用語も初出時には異なる意味で使われていたとされる（[野村]）。

文献の中に明確に登場するのはピーター・ナウアの著作（[Naur]）が最初であるとされ，確かに著作の中では

1.8. A Basic Principle of Data Science

という項目を立てており，第1章のSummaryの中で

> Data science is the science of dealing with data, once they have been established, while the relation of data to what they represent is delegated to other fields and sciences.

と述べている。まさに「データを扱う科学」をData Scienceとしており，

課題や多様な分析手法という意識はないように見える。

野村総合研究所の有賀友紀・大橋俊介（[野村]）によると，現在使用されているData Scienceという用語に最も近い定義を行ったのは統計学者の林知己夫（[Hayashi]）であるとされる。[Hayashi]の中で林は

> Data Science is not only a synthetic concept to unify statistics, data analysis and their related methods but also comprises its results. It includes three phases, design for data, collection of data, and analysis on data. Fundamental concepts and various methods based on it are discussed with a heuristic example.
>
> （和訳）
>
> データサイエンスは，統計，データ分析，およびそれらに関連する方法を統合するための総合的な概念であるだけでなく，その結果も含む概念である。それは，データの設計，データの収集，およびデータの分析の3つのフェーズを含む。基本的な概念とそれに基づくさまざまな方法は発見的な例により議論される。

と述べている。

現代の「データサイエンス」という用語はこの林の定義をベースにしており，手法や関連事項，対象領域が拡大しているにも関わらず，敢えて再定義をせず，そのままにして緩やかな理解のまま済ませているように見える。

一方，現在Big Techと呼ばれる米国のIT企業達がコンピューターを駆使して購買データや消費者のプロファイルを収集分析することで売り上げを伸ばしていることが人々に知られ，またDeep Learning技術による実用的なAIが登場するにつれ，データサイエンスの威力は世界に広く知られることとなった（[Foote]）。

データサイエンスとは何かについては漠とした理解のまま，しかし，その重要性はますます高まってきているというのがデータサイエンスの現状であると言える。

2-2．データサイエンス教育

データサイエンスが強力な力を持っていることが理解されると，各国でデータサイエンティスト育成が始まった。

米国に後れを取ったと日本が認識したのは2014年頃である。

日本学術会議 情報学委員会 E-サイエンス・データ中心科学分科会により『提言 ビッグデータ時代に対応する人材の育成』(2014年 9 月11日）が出されたが，「１ 作成の背景」（[学術]）に書かれた次の文章が日本の当時の状況をよく表している。

> 情報通信技術の飛躍的進歩によって，多くの学術研究分野や社会において時々刻々ビッグデータが蓄積しつつある。ビッグデータには膨大な知識や潜在的価値が埋蔵されているため，その有効活用が今後の学術や産業発展の鍵となっており，激しい国際競争が始まっている。
>
> 当分科会ではビッグデータ活用のために今や喫緊の課題となっている，データ中心科学の確立と学術分野への普及・定着のための方策の検討を行ってきたが，その中で新しい科学のための新しい人材，すなわちデータサイエンティストの育成が特に重要であることを認識した。データサイエンティストの育成は，既に海外では急速に進められており，我が国においても直ちに着手しなければ，学術研究や産業界におけるビッグデータ活用において大きく立ち遅れる恐れがある。
>
> そこで，当分科会ではビッグデータ時代に対応した人材育成の在り方を中心に検討を行い，本提言を作成した。

この文章の中に出て来る「データ中心科学」が現在のデータサイエンス（正確には本論文で定義する「学問としてのデータサイエンス」）である。

この提言を受け，日本政府も文部科学省を通じてデータサイエンス教育に本格的に取り組み始めた。

平成28年(2016年)12月に，北海道大学・東京大学・滋賀大学・京都大学・大阪大学・九州大学の 6 大学が数理及びデータサイエンスに係る教育強化の

拠点校として選定された。そして，これらの大学を核として「データサイエンス教育強化拠点コンソーシアム」が組織化された。令和元年（2019年）には，国立大学20校が協力校として参加し，全国を 6 ブロック（北海道・東北，関東・首都圏，中部・東海，近畿，中国・四国，九州・沖縄））に分けて，データサイエンス教育の普及に取り組み始めた（［教育 1 ］）。

「データサイエンス教育強化拠点コンソーシアム」は日本のデータサイエンス教育の標準を作ろうとしているので，コンソーシアムが提供する資料（［教育 2 ］）からデータサイエンスに関する我が国の現在の認識を伺うことができる（図 1 ）。

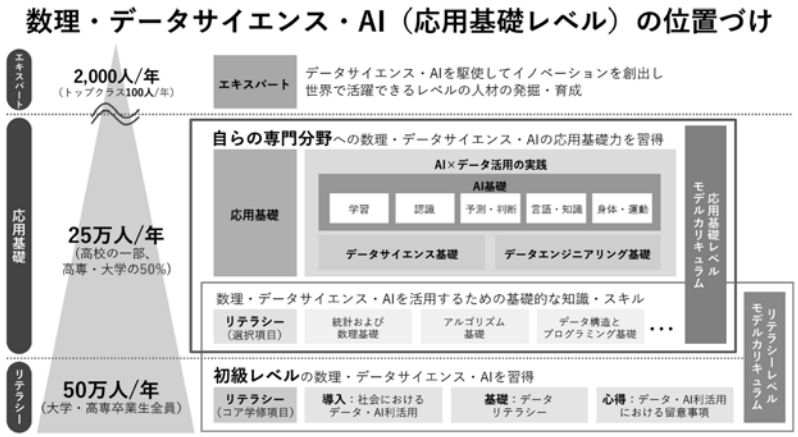

図 1

出所：［教育 2 ］の p 8

3．データサイエンスの再定義

2 章でデータサイエンスの現状を見てきたが，概ね，林知己夫による定義とその後の AI との関係等を踏まえて「データサイエンス」を理解していると言える。そこで，本論では，データ分析とその活用プロセスとしての「データサイエンス」（プロセスとしてのデータサイエンス）と，プロセスとして

のデータサイエンスにおける諸要素を研究対象とする「データサイエンス」（学問としてのデータサイエンス）に分けて論ずることにする。

3-1．プロセスとしてのデータサイエンス

「プロセスとしてのデータサイエンス」を最も簡単に定義すると，以下のようになる。

定義

「プロセスとしてのデータサイエンス」とは

- 世の中の様々な課題を（探索的にまたは実証的に）解決するために
- 課題の生じている現象・事象から
- それらの現象・事象を表すデータ群（データサイズは小規模から超大規模まで，種類は数値データ・文字データ（文章）・画像データ・音声データ等）を入手し
- そのデータを統計的手法や非統計的手法を用いて分析することで
- 現象・事象の中に潜む規則や法則を発見し
- それらを利用して，新たな価値（新商品・新サービス・未来予測・新たな政策等）を生み出す

一連の手順・手法のことである。ただし，分析および規則法則の発見がコンピューターによる機械学習という形態で行われる場合も含むものとする。

図示すると図2のようになる。

この定義から分かるように，「プロセスとしてのデータサイエンス」は従来の統計分析，データマイニング，BigData分析，マーケティングリサーチ，探索的分析，実証研究的アプローチ，生命科学におけるBioinformaticsなどをすべて含むものとなっている。

個別のデータ分析手法がそのままに止まらず，大きく「プロセスとしてのデータサイエンス」に包含されるようになったのには以下の理由がある：

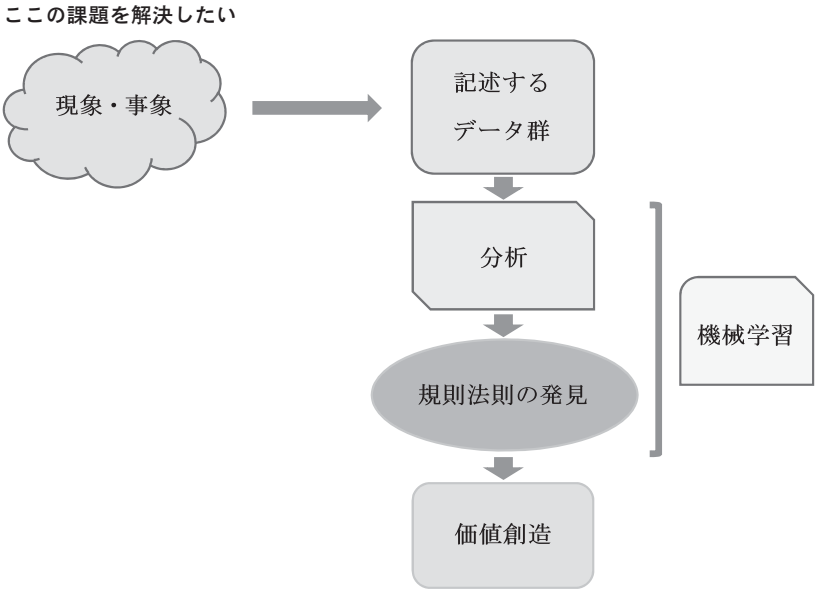

図 2

出所：筆者作成

(1) ITの発達により多種多様なデータを収集できるようになったこと。
インターネット上の各種サーバーが取得する膨大な情報，SNS等で人々が発信する文字情報・音声情報・画像情報，通信インフラ（モバイルネットワーク上の基地局等）の取得する膨大な定型データ，IoTにより各種機器から収集される膨大な情報などが存在し，手続きを踏めばこれらを入手できるようになったこと。

(2) コンピューターの性能が向上し，身近なPCでも高度な分析を行えるようになってきたこと。

(3) 分析ツールが整備され，安価に高度な分析が行えるようになってきたこと。

(4) ニューラルネットワークのようなデータを学習し規則法則を内在化させ

る手法も取り入れることで，分析および規則法則の発見の段階が一体化することも良しとした。この考え方が一連の過程の自動化や機械学習（Machine Learning）手法として発展してきた。

(5) データサイエンス分野の分析手法を機械学習（Machine Learning）手法としてAIにインプリメントすることが進んできたこと。逆に分析の自動化などAI機能を利用した分析が行われるようになったこと。

(6) (1)から(5)を総合することで，様々な分野で非常に強力な成果を生み出せるようになったこと。

従って，従来型のデータ分析においても，データの種類・データの量・分析手法・AIを含むITの活用法のどれかの要素が従来の形態を超えていれば，「プロセスとしてのデータサイエンス」であると言えることになる。

3-2．プロセスとしてのデータサイエンスの詳細

「プロセスとしてのデータサイエンス」の定義は3-1で述べたとおりである。ここではもう少し細部に踏み込むことにする。

(1)「世の中の様々な課題を（探索的にまたは実証的に）解決する」

課題が存在し解決したいと考えている人または組織をAとする。データ収集と分析を行う人または組織をBとする。AとBが異なれば，Bから見てAはクライアントである。もちろん，AとBが同一のことも多い。ここでAとBを区別するのは，AとBとの関係により，分析手法や価値創造が大きく異なる結果となるからである。AとBとの関係が影響を及ぼすときには，以下各段階で注意を述べる。

次にAが課題を解決したいと言っても，大きく分けて次のような2通りの場合がある：

(a) 課題の発生している現象・事象に関する知見が十分でなく，探索的な分析を行いたいとき。

(b) 課題の発生している現象・事象に関する知見がある程度あり，仮説に基づく検証・実証を行いたいとき。

(a)(b)は単独で実施されることもあれば，(a)の後に(b)，逆に(b)の結果(a)が必要になるということもある。

また，分析全体の実施方法は通常１回切りの場合がほとんどであるが，分析が手軽になってきたため，ソフトウェア開発のように以下の各種タイプの分析を行えるようになりつつあることにも注意したい：

- ウォーターフォール型（事前計画通りに分析を１回実行する）
- プロトタイプ型（ひな形となる分析を行った後，本格分析を行う）
- アジャイル型（短期間で実施できる本格分析を繰り返して改善を図る方法）

Aが何を目標としているかで以降の内容が大幅に変化するので，AとBが異なる場合は，最初の段階から十分な協議が必要である。

(2)「課題の生じている現象・事象」

統計分析で言えば，母集団をきちんと特定するということに相当する。推測統計学を学習するときに母集団のことを学ぶにも関わらず，現実のデータ分析では母集団があいまいのまま，または十分な検討をされないままということが多い。

課題を解決したいAが特定商品の売り上げを問題にしているとき，どの範囲を母集団とするかは非常に重要である。店舗はどの範囲を対象とするのか，期間はどの範囲にするのか，ターゲットとしては特定商品のみで良いのか，より広い範囲の商品をターゲットにすべきか。母集団の構造は地域や時間により変化するということを理解しておかないと間違った結果を導くことになったり，底の浅い分析を行うことになる。母集団を十分に吟味しておくことが間違いのない分析の第１歩である。

(3)「現象・事象を表すデータ群(データサイズは小規模から超大規模まで,種類は数値データ・文字データ(文章)・画像データ・音声データ等)の入手」

母集団を特定した段階で,課題となる現象・事象を記述するデータを収集できるかがまず問題となる。個人情報保護法の成立以来,情報保護意識が高まったこともあり,分析に必要なデータを入手しにくくなってきた。そのため,公開情報や本来必要とするデータの代替データで分析を行うということも生じている。課題を解決したいAが独自データを持つ場合は,それをフルに活用すべきである。

また,データサイエンスというとBigDataを分析するものと思い込んでいる人々もいるが,小規模データの分析を行うデータサイエンスも存在する。例えば,日本国内の47都道府県の産業構造分析を行おうとすると,データ件数としては47件にしかならない。しかし,多数の属性により特徴づけられているため,データサイエンスによる分析が必要となる。

どのような種類のどのような範囲のデータを収集すべきかはAと分析者Bの協議により明確にする必要がある。最初の段階でAの目標が明確になっているはずなので,その目標に合う種類のデータを収集するのであるが,実はこれが大変難しい。分析を実行してみて初めてデータの不足・変数の不足が明らかになることが多い。

これらの事情により,データ収集・データ分析の前には「データ分析のデザイン」を実施するが普通である。そこでは,

- 目標達成のために収集すべきデータ全体の関係図を描き,
- 想定する分析方法を選択し,
- データ収集の方法・時期・期間を明らかにし,
- 質問票等独自調査が必要な場合は質問票の設計を行う。

特定母集団の分析を行うためには,その集団の外部環境のデータを収集しなくてはいけないことが多く,収集すべきデータ全体をできるだけ事前に描

き出すようにすることは必須である。この「データ分析のデザイン」はAとBの間で十分な時間をかけて実施すべきである。

(4)「統計的手法や非統計的手法を用いての分析」

現代のデータ分析の多くが統計分析を超えてデータサイエンスと呼ばれるようになった大きな理由の一つが非統計的手法による分析ができるようになったことにある。

これら非統計的手法は1990年代にKDD（Knowledge-Discovery in Databases）と呼ばれ，その後，データマイニングと呼ばれるようになった手法群である。コンピューターの計算能力の向上により，人手では不可能な大量で複雑な計算を実行できるようになったため，このような手法が誕生した。

ただし，統計的手法を用いるにせよ，非統計的手法を用いるにせよ，分析を行うには最低でも以下の手順を踏むことになる。

(A) データクレンジング（欠損値や外れ値等のあるデータの処理等）
(B) データ加工（どのような加工を行うかで規則法則が見えるか見えないかが大きく左右されるため，分析者の腕の見せ所となる。）
(C) データ分析（どのような手法を選択するか。手法によって規則法則が見えたり，見えなかったりする。）

また，Aがクライントであるとき，分析者Bは他者への説明がしやすいデータ分析手法を選ぶことになる。そのため，規則法則が目に見えにくい手法やなぜそうなるかの説明がしにくい手法は敬遠される。最終の価値創造はAの判断によるものであるため，Aが理解しにくい分析方法はやはり避けるべきものとなる。データ分析手法に関しては(7)と 3 - 4 で改めて述べる。

(5)「現象・事象の中に潜む規則や法則の発見」

統計的手法は標本や母集団全体の特徴を把握するためのものであると言える。層別を行った上で各層の特徴を調べるということも行われているが，一

般的には集団全体の特徴を把握し，必要ならば推測する。

一方，非統計的手法は標本や母集団全体の中の一部の集合にのみ当てはまる規則や法則を浮かび上がらせてくれる。

店舗の売り上げデータは，多様な消費者の購買行動の結果によるものである。売り上げデータ全体から分かる客単価や消費傾向も重要な指標とはなるが，どのような特性を持つ消費者がどのような購買行動を取っており，どのような消費者群がいるかを把握できる方が好ましい。マーケティングリサーチはそのような事実の解明を目指しており，データサイエンスの非統計的手法はこの点で非常に強力である。

(6)「分析結果を利用した新たな価値（新商品・新サービス・未来予測・新たな政策等）の創造」

「プロセスとしてのデータサイエンス」の目標は(1)の「世の中の様々な課題を（探索的にまたは実証的に）解決する」であった。そのために(2)から(5)の手順を踏んだのである。(5)で明らかになった「現象・事象の中に潜む規則や法則」をどのように活用するかが本来の目標である。

当然ながらこの最終段階は課題の存在する分野により，また課題の性質により様々な様相を見せる。

まず，課題を解決したいAが望むような規則や法則を発見できたかが問題である。分析者Bは往々にして分析そのものに拘り，それをどのように活用するかをなおざりにすることがある。そのようなことがないようにするため，プロセスの要所要所でAとBが協議し，何を明らかにすべきか常に意識を共有すべきである。

また，AとBが異なるのであれば，Aが理解しやすい手法で規則や法則を発見し，丁寧に説明する必要がある。

分析者Bは規則や法則の精度・信頼度を重視するが，精度・信頼度が低くてもAにとっては重要な発見であることもあり，その逆もあり得る。Aは課

題に関わる事柄を総合的に把握しており，分析に用いたデータ以外のことで，規則や法則の重要性を判断する。例えば，規則・法則の発見過程では用いなかった条件（コストや機会のこと等）も併せ考えて経営判断することは十分にあり得る。

また，多くの場合，一連のプロセス(1)〜(5)で分析は終わったと考えがちであるが，本当の探求はこれからであると気がつくこともある。むしろより深い視野・より広い知見を求めて次なるプロセスを開始すると考えた方が好ましい。(1)で述べたように，分析が手軽にできるようになってきた現在，

・ウォーターフォール型（事前計画通りに分析を１回実行する）
・プロトタイプ型（ひな形となる分析を行った後，本格分析を行う）
・アジャイル型（短期間で実施できる本格分析を繰り返して改善を図る方法）

のうちの「プロトタイプ型」「アジャイル型」の実現可能性が高くなってきている。

(7)「分析および規則法則の発見がコンピューターによる機械学習という形態で行われる場合も含む」

データサイエンスにおける各種分析手法が機械学習機能としてなぜAIにインプリメントされているのかの理由については「3-5．AIとの関係」で述べることにし，ここでは，「プロセスとしてのデータサイエンス」においてAIの機械学習機能を利用する際の注意点を述べる。

データサイエンスにおける「分析」「規則法則の発見」の過程をAIに行わせるときに問題となるのが，規則法則の理解可能性である。課題解決を望むAと分析者Bが異なるときは特に顕著であり，最終的な「価値創造」を行うためには，発見できた規則法則がどのような内容であるのかをAが理解できることが求められる。規則法則がブラックボックス化していては，最終的な経営判断を下すのにリスクが高すぎる。そこで，現在提唱されているのが，

XAI（eXplainable AI：説明可能なAI）という概念である（[大坪]）。Aの理解可能性とは分析者B側からすると説明可能性であり，その様な分析をしてくれるAIが求められている。そのためXAIは今まさに発展途上であり，多くの新しい手法が開発されている。

3-3．学問としてのデータサイエンス

研究者がデータサイエンスという用語を用いるときには，学問としてのデータサイエンスを意味することがある。

定義

「学問としてのデータサイエンス」とは人間やコンピューターが行う「プロセスとしてのデータサイエンス」そのものとそれを取り巻く環境，分析等の結果生じる現象を研究対象とする学問である。従って，「プロセスとしてのデータサイエンス」の

- 適用領域，適用上の条件や課題
- データ収集手法，データ管理技術
- データクレンジング手法，データ加工技術，加工手法
- データ分析手法の改善・開発・評価
- 規則法則の抽出方法
- クライアントへの説明方法
- 一連の分析手続きの見直し
- 分析手法等の他への応用
- 必要とする機器・コンピューター

また

- データサイエンス教育
- データサイエンスの可能性および影響

等が研究テーマとなる。

この段階になりようやく，「データサイエンス」を「経済学」「経営学」等

と同様な用語として認識できる。「経済学」が人間の経済活動とその結果生じる経済現象や制度を対象とする学問，「経営学」が企業等の経営の実態と取り巻く諸環境を対象とする学問であったように，「学問としてのデータサイエンス」は人間やコンピューターが行う「プロセスとしてのデータサイエンス」とそれを取り巻く環境，分析等の結果生じる現象を対象とする学問であると言える。

3-4．分析手法

3-2(4)で述べたデータ分析手法を整理し概観しておく。

まず，手法全体を図示すると図3のようになる。

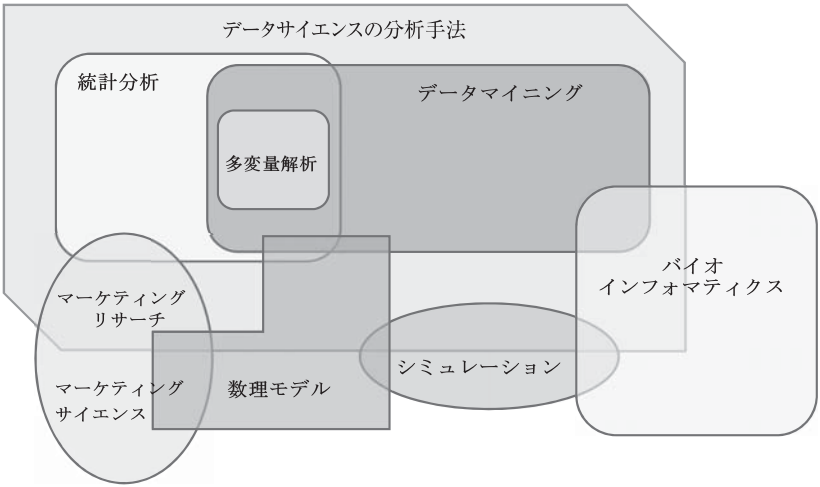

図3

出所：筆者作成

ここで統計的手法とデータマイニング（KDD）について主な手法を概観しておく。

・統計的手法

通常の統計学で学ぶ記述統計学，推測統計学

多変量解析

数理科学分野・生物学分野・心理学分野・経済学分野等の個別分野で用いる統計学的手法

・データマイニング手法

アソシエーション分析

決定木分析

集団学習

SVM（Support Vector Machine）

ニューラルネットワーク各種（SOM[2],CNN[3],RNN[4],GAN[5]も含む）

強化学習

ネットワーク分析

テキストマイニング各種

XAI（eXplainable AI） 各種

3-5．AIとの関係

人工知能（AI）をどのように実現するかという長い道のりにおいてはいくつもの課題があったが，推論のさせ方や知識の表現形式に関する研究は1980年代までに一応の成果を得ており，その後大きな課題となったのが，自律的な知識獲得をどのように実現するかであった（[小高]）。

1990年代は丁度インターネットの整備普及が始まった時期であり，インターネットから情報を収集し，そこから「何らかの方法」により知識を獲得するという発想に辿り着くのは自然な流れであった。このコンピューターの

2　SOM（Self-Organizing Map　自己組織化マップ）

3　CNN（Convolutional Neural Network　畳み込みニューラルネットワーク）

4　RNN（Recurrent Neural Network　再帰型ニューラルネットワーク）

5　GAN（Generative Adversarial Network　敵対的生成ネットワーク）

自律的な知識獲得を「機械学習」と呼ぶ。そして,「何らかの方法」の部分として最適であったのがデータサイエンスの各種手法であった。

人工知能（AI）分野で研究されていたニューラルネットワークはデータサイエンスにおいて実用化されており，データサイエンス側のその他多くの手法と共に人工知能（AI）の「機械学習」という機能に還元された。

現代では，ソフトウェアが機械学習機能を備えていればAIであるとまで緩く解釈されるようになっているが，本来は，機械学習機能はAI実現のための一要素である。

機械学習とは上述の経緯で登場したものであるため，データサイエンスの各種手法が機械学習機能において利用されるのは当然であり，逆にAIという名称でデータサイエンスの各種手法を使用させるプログラム群をデータサイエンスで利用することも当然な流れである。

データサイエンスの各種手法はAIに当然のごとく利用されるが，AIのブラックボックス化が進む中で，AIが導き出す規則法則の理解可能性・説明可能性が重視されるようになった。決定木分析のような古典的な分析手法は精度は今一つでも，結果の分かりやすさ（理解可能性・説明可能性）は抜群である。そのため，分析精度も高く説明可能性も高いXAI（eXplainable AI）の研究開発が現在精力的に進められている（[Gun]［大坪］）。

4．データサイエンスの今後

「2-2．データサイエンス教育」で述べたように，データサイエンス教育が国の主要政策の一つであるため，教育界では「データサイエンス」「データサイエンティスト」という用語がトレンドになっている。しかし，ビジネスの世界では，データサイエンティストがAIの機能を開発したり，AIを使って分析をするのは当たり前であり，「AI」という用語の方が主流となってきている。「統計学」「統計分析」が目新しくない用語として各学問領域に定着しているのと同様に，今後「データサイエンス」という用語が正確な意味で

定着していくことが好ましい。

用語の動向がどうなるかを離れ，「プロセスとしてのデータサイエンス」「学問としてのデータサイエンス」そのものがどうなるかについて，現在の動向から推測すると，データサイエンスにおいて，今後次のような発展拡大があるのは間違いないと思われる。

- 適用領域の拡大
- 新手法の開発
- 自動化
- 量子コンピューターの利用

4-1．適用領域の拡大

ビジネス界ではAIによるデータサイエンスが当たり前になってきているとは書いたものの，ビジネスにデータサイエンス・AIが大いに役立つという認識が広がっているということであり，自社内で「プロセスとしてのデータサイエンス」を実際に活用しているかというと，先進的な大企業とそうでない企業，特に中小企業との格差は広まる一方のようである。その意味では，社会全体としてのデータサイエンスの活用はまだまだこれからであると言える。

現在どのような事例があるか国内に絞って見てみる。以下で事例の引用先とする株式会社NTTデータ数理システムは1982年にCADと科学技術計算を専門とする企業として設立され，日本におけるデータサイエンスのリーディングカンパニーとして日本におけるデータサイエンス普及と高度化に大いに貢献してきた会社である。2012年にはNTTデータグループに入り，NTTデータ本体におけるデータサイエンティスト育成の中核ともなっている企業である（[数理]）。

この数理システム社のホームページにある活用事例が日本でのデータサイエンスの活用状況の一つの目安となる。分野ごとの事例紹介が「マーケティ

ング」「Web解析」「レコメンデーション」「施設配置・立地戦略」「物流ロジスティックス」「スマートグリッド・エネルギー」「ネットワーク」「金融」「コールセンター」「スケジューリング」「半導体・LSI・MEMS」「教育」と区分されていることから，社会の広い範囲でデータサイエンスが活用されていることが分かる。

この中の「マーケティング」の例を一覧表示してみる（次ページ表１）。

数理システム社と共同で分析やシステムの開発を行った企業の例であるため，企業独自の例は含まれていないが，企業がそれぞれの課題を解決するためにプロセスとしてのデータサイエンスを活用している様が伺える。

4 - 2．新手法の開発

3 - 2 (7)および3 - 5で述べたように，XAI（eXplainable AI）の手法の研究が精力的に進められている（[Gun]［大坪]）。ここでは代表的な手法としてLIMEとSHAPについて概説する。

LIME（Local Interpretable Model-Agnostic Explanations）は2016年にMarco T. Ribeiro, Sameer Singh, Carlos Guestrin（[Ribeiro]）が提唱した手法であり，“局所的に解釈可能な モデルに依存しない説明”の意味である。その名の通り，特徴量からなるデータ空間Dにおいて定義されたモデルfがあり，$a \in D$におけるfの予測値$f(a)$が分かるとき，$a \in D$の近傍Uでfを近似する解釈しやすい近似モデル（またはモデルの集合）ξ_aを獲得して代替して説明する手法である。

近似モデルの集合ξ_aは，線形モデルの候補集合Gの中から

$$\xi_a = \arg\min_{g \in G} (L(f, g, \pi_a) + \Omega(g))$$

で得る。ただし，

$$L(f, g, \pi_a) = \sum_{z,\, z' \in U} \pi_a(z)(f(z) - g(z'))^2 \qquad \Omega(g)\text{は}g\text{の正則化項}$$

企業名	分析やシステムの概要
WOWOWコミュニケーションズ	解約理由分析による離反率抑制
監査法人トーマツデロイトアナリティクス	テキストマイニング・確率的潜在意味解析・ベイジアンネットワークを組み合わせての口コミ分析（宿泊予約サイト口コミ分析，観光地の口コミ分析）
オージス総研	行動観察とベイジアンネットワークでの生活者心理のモデリング
株式会社アサツーディ・ケイ	エージェント・ベースド・モデルによる消費者行動の表現
ソネット・メディア・ネットワークス株式会社	効果要因分析に基づくRTB（リアルタイムビッティング）
ビデオリサーチ	ビデオリサーチ社の独自調査データと顧客企業側の大規模データによるデータ統合ソリューション
ぐるなび	テキストマイニングによる ①アクセス上位店分析 ②ブランドイメージ比較 ③キャッチコピー分析
三菱食品	テキストマイニングによるコンビニデザートに対する生活者の意見分析（ブランド評価）
システム・ロケーション株式会社	過去の販売実績データに基づく自動車の残価自動計算システム
東都生活協同組合	併売分析，キャンペーンマネジメント
アイエックス・ナレッジ株式会社	アンケートの自由回答のテキストマイニング
横河電機	テキストマイニングによる3C　（Customer, Competitor, Company）分析

表1

出所：［数理］活用事例より筆者作成

$$\pi_a(z)=\exp\left(-\frac{D(a,z)^2}{\sigma^2}\right) \qquad D(a,z)\text{は}a,z\text{の距離}$$

とする（[Ribeiro]［大坪］）。元のモデルfに制約がないため適用範囲が広いが，$a\in D$の近傍Uが自動生成されるため，分析のたびにξ_aが変わってくるという難点がある。

SHAP（SHapley Additive exPlanations）は2017年にScott M. Lundberg, Su-In Lee（[Lund]）が提唱した手法である。協力ゲーム理論に登場するシャープレイ値の考え方を用いて，特徴量の予測への寄与度を計算する。これにより，どの特徴量がどの程度効いて予測結果が算出されているかを把握できることになる。SHAPには元となるモデルに依存しないKernelSHAP，決定木モデル用のTreeSHAP，DeepLearningモデル用のDeepSHAPなど様々な派生手法が存在する（[大坪]）。

4-3．自動化

プロセスにおけるデータサイエンスの「分析」「規則法則の発見」の部分はAIによる「機械学習」でも行われるようになったと述べたが，「記述するデータ群」の抽出の段階まで含め，自動化する動きが始まっている。

ツールを利用して人手で一連のプロセスを実行すると，専門家でもだいたい2か月程度の時間を要する。データクレンジング，分析手法(分析モデル)の選定，規則法則発見までの試行錯誤があるため，どうしても時間がかかってしまう。

しかし，データサイエンスの威力を認識し始めた企業ほど新たな分析に対する需要は高く，次々と分析課題が発生する。人力による数か月の分析を待ってはおれないという状況が生まれつつある。

このような中で，NECが米国シリコンバレーに設立したdotData Inc.が発売する「dotData」は「自動化により，あらゆる規模の企業でAI駆動な

企業風土を作り上げ，データドリブンDXの推進を支援する」（[Dot]）という機能を持つソフトウェアとして急速に注目を浴びている。

課題があると，関連すると思うデータベース（データウェアハウスやデータマート）を分析担当者がdotDataに指示するだけで，一連のプロセスが始まり，数日で結果が表示される。

このような自動化はますます進むものと思われる。

近未来には，課題そのものの発見，必要とするデータベースの列挙から始まり，データ収集・クレンジング・データ加工・分析・規則法則の発見・活用提言までをアシストする自動化ソフト（AI）が登場してもおかしくはないと思われる。

4-4．量子コンピューターの利用

19世紀末から20世紀初頭にかけて，数学の世界では集合論の建設に端を発し，数学そのものの基礎を研究するための論理学および数学基礎論の構築が進んだ。この流れは20世紀中葉のフランスの数学者集団ブルバキによる数学体系の再構築にまで及んだ。真偽2値の記号論理学の完成は，シャノンによるデジタル回路の考案，2進法を基礎とするコンピューターの構築，数字や文字・記号を2進数に対応させることで世界をデジタル化する発想へとつながった。このようにして，現代のコンピューターはほとんどのものが2進法を根本において構成されている。

2進法は，真と偽，1と0，ある・なし，高い・低いの間の変換を容易にし，現代のコンピューターが今のように隆盛を極めた根幹をなすとも言える。

一方，20世紀初頭から始まった量子力学の発見と建設は現象世界に対する認識を一変させることとなった。粒子性と波動性は極微の世界に限らず成り立つ物理法則であるが，極微の世界になるほど2重性が顕著となる。その性質は物質においても力（相互作用）においても見られる。2重性は，観測に

より粒子の状態が一意に定るまでは粒子が複数の状態の重ね合わせとして存在することも意味する。また，量子エンタングルメント（量子もつれ）状態にある 2 粒子間での量子情報のやり取り（量子テレポテーション）などの発見が量子力学をより一層豊かな量子情報論へと発展させた（[石坂]）。

ところで，現代のコンピューターを構成する CPU，メモリ，光通信のどれも量子力学により性能や特性を調整している。微細構造物になるほど量子力学的効果が現れるのであるが，2 進数の性質を保つために，量子力学的効果を抑え込むための多大な努力が費やされて来た。

このような中で，微細素子が 2 進数のどちらかの状態に定まらない量子状態のまま計算を実行できないかという発想が生まれ，量子計算・量子回路・量子アルゴリズム（[Grover] [Shor]）の研究，量子コンピューター実現への研究開発が行われてきた。

現在，量子コンピューターは現実のものとなり，量子ゲート型の量子コンピューターである IBM 社の Q などの実機が稼働するようになった。

現在稼働中の量子コンピューターは NISQ（Noisy Intermediate-Scale Quantum）と呼ばれる小規模のものであるため，本格的な誤り耐性量子コンピューターの開発競争が凌ぎを削っている（[嶋田]）。

NISQ 量子コンピューターである Q など，量子コンピューターはまだ発展途上にあるとは言え，特定用途における量子コンピューターの性能に対する期待は大きく，国家レベルでの競争が行われている。

期待される用途の一つとして，データサイエンス分野での応用がある。

現在稼働中の Q は実機をオンラインで利用可能であるため，データサイエンスの可能性の探求と量子コンピューターの可能性の探求の双方の意味で，量子コンピューターを利用したデータサイエンスは研究開発や教育が今後ますます進む分野であると思われる。

今の所，量子コンピューターは CPU に相当する QPU だけが開発された状態であるため，パソコンのように使いこなすことはできず，現時点で Q を

利用するには，

- プログラミング言語Python
- 数学的知識（複素数体C上の線形代数，テンソル代数，フーリエ変換など）
- 論理回路に関する知識
- 量子力学の一部と量子情報論の一部

などが必要となる（[湊]）。そのため，初学者にはハードルが高い。しかし，教育環境の整備・指導法の整備，更には量子コンピューターメーカー側のプログラム開発環境整備により，少しずつではあっても環境の改善が進むものと思われる。

参考文献

［Dot］dotData社ホームページ（https://jp.dotdata.com/our-company/）

［Foote］Keith D. Foote,‘A Brief History of Data Science’, In Dataversity, 2016.（https://www.dataversity.net/brief-history-data-science/）

［学術］日本学術会議 情報学委員会 E-サイエンス・データ中心科学分科会 「提言 ビッグデータ時代に対応する人材の育成」（2014年 9 月11日）

［Grover］L.K.Grover,‘A fast quantum mechanical algorithm for database search’, In The 28th Annual ACM Symposium on the Theory of Computing（STOC ‘96）, pp.212-219, 1996.

［Gun］David Gunning,‘Explainable Artificial Intelligence（XAI）’, DARPA/I20.

［Hayashi］Chikio Hayashi,‘What is Data Science? Fundamental Concepts and a Heuristic Example’, In C. Hayashi, K. Yajima, HH. Bock, N. Ohsumi, Y. Tanaka, Y. Baba（eds）“Data Science, Classification, and Related Methods. Studies in Classification, Data Analysis, and Knowledge Organization”, Springer, Tokyo, 1998.

［石坂］石坂智・小川朋宏・河内亮周・木村元・林正人『量子情報科学入門』共立出版，2012年

［教育 1 ］数理・データサイエンス教育強化拠点コンソーシアム ホームページ（http://www.mi.u-tokyo.ac.jp/consortium/overview.html）

［教育 2 ］数理・データサイエンス教育強化拠点コンソーシアム「数理・データサイエンス・AI（応用基礎レベル）モデルカリキュラム～AI×データ活用の実践～」，2021年 3 月

［Lund］Scott M. Lundberg, Su-In Lee, ‘A Unified Approach to Interpreting Model Predictions’, arXiv:1705.07874v2［cs.AI］25 Nov 2017.

［湊］湊雄一郎・比嘉恵一朗・永井隆太郎・加藤拓己『IBM Quantumで学ぶ量子コンピュータ』秀和システム，2021年

［Naur］Peter Naur, “Concise Survey of Computer Methods”, Petrocelli Books, 1974.

［野村］野村総合研究所 有賀友紀・大橋俊介『RとPythonで学ぶ実践的データサイエンス＆機械学習』技術評論社，2019年

［小高］小高知宏『人工知能入門』共立出版，2015年

［大坪］大坪直樹・中江俊博・深沢祐太・豊岡祥・坂元哲平・佐藤誠・五十嵐健太・市原大暉・堀内新吾『XAI（説明可能なAI）そのとき人工知能はどう考えたのか？』リックテレコム，2021年

［Ribeiro］Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin, ‘“Why Should I Trust You?” Explaining the Predictions of Any Classifier’, arXiv:1602.04938v3［cs.LG］9 August 2016.

［Shor］P.W.Shor, ‘Polynomial-Time Algorithms for Prime Factorization and Discrete Logarithms on a Quantum Computer’, SIAM J. Comput., Vol.26, No.5, pp.1484-1509, 1997.

［嶋田］嶋田義皓『量子コンピューティング　基本アルゴリズムから量子機械学習まで』情報処理学会出版委員会監修，オーム社，2020年

［数理］株式会社NTTデータ数理システム ホームページ
（https://www.msi.co.jp/index.html）