解説

データサイエンス・生成AIの紹介

株式会社アトラエ 上級データサイエンティスト　杉山　聡
データサイエンス VTuber アイシア＝ソリッド

1．はじめに

ここ数年，新しいデータの利活用の事例やAIの登場が相次ぎ，生活がどんどん変わってきています。2022年に登場したChatGPTを使ったことがあるでしょうか。宿題や勉強内容がわからないときにChatGPTに質問すれば，無制限にいつまででも，私たちに合わせた解説を提供してくれます。登場時は誤った返答も多かったですが，最近はかなり正確になりました。また，TikTokやYouTubeショートなどでは，私たちがなんとなく動画を見ているだけで，常に興味ある動画を提供し続けてくれます。

これらの仕組みを裏で支えているのが，データサイエンスや生成AIです。この記事では，データサイエンスや生成AIとは何か，どのような仕組みで動いているかを紹介します。

2．データサイエンスとは

データサイエンスとは，データの力を活用し，ビジネス課題を解決する活動で，この仕事をする人は，よくデータサイエンティストと呼ばれます。「ビジネス課題を解決」と言われてもピンとこないかもしれませんが，「このケーキの値段を変えずにもっと美味しくしたい」とか「このゲームをもっと面白くしたい」など，私たちの生活に直接関わるものが多いです。

最近では，スポーツの文脈でもデータサイエンスが行われることがあります。その文脈では，例えば野球なら「この投手の投球時の肘の位置をあと5 mm上げれば球速が上がるだろう」とか，バスケットボールなら「相手の攻勢が激しく失点が続いている時，誰を投入すべきか」などを検討します。このように，実現したい目標に対して，データの力を活用して実現を目指す活動がデータサイエンスです。

3．身近なデータサイエンスとプロレベルのデータサイエンス

実は，データサイエンスは大人だけのものではありません。高校生の皆さんであっても，すでにデータサイエンス的な活動は行っています。データサイエンスの本質は，データを用いて状況を理解し，判断し，行動を変えることです。例えば，模試の結果を見て，得点率や平均点との比較を通して分野別の得意・不得意を把握し，勉強時間の配分を変えたとしたら，これはもう立派なデータサイエンス的活動です。

このような日常でのデータ利活用の延長に，プロのデータサイエンスがあります。データサイエンティストには，4つのレベルがあります。まず一番身近なレベルが「見習いレベル」です。統計やプログラミングなどの知識を基に，今例に挙げたような判断を正しく行える人たちです。次のレベルが「独り立ちレベル」です。課題解決の本質を見抜き，大学以降で習うような専門的な分析手法を使いこなし，大規模なデータベースも難なく扱える人たちです。一般の会社の最上位レベルが「棟梁レベル」です。全社規模のプロジェクトを，部署間の対立などを乗り越えて実現する人や，最先端のアルゴリズムを使いこなして難しい分析課題を解決する人，大規模なデータベースや複雑なデータも扱える技術者などが該当します。最後に，「業界を代表するレベル」があり，複数の企

業を取りまとめたり国レベルのプロジェクトを動かしたりする人や，現在の分析手法では解析不能な問題に対して新しい手法を生み出して不可能を可能にする人，超大規模なデータ・AIを扱うシステムを安定的に動かせる人などが該当します。

4．データサイエンティストになるには

ここまでに紹介した仕事の中に，カッコいいと感じるものはあったでしょうか。もしあった場合は，データサイエンティストを目指して勉強してみても良いでしょう。データサイエンティストになる道筋には，大きく分けて２つあります。

１つめが，大学のデータサイエンス学部に進学することです。データサイエンス学部では，分析手法やプログラミングなどの基礎技術を学べるとともに，実際のビジネス課題の解決に挑戦できるプログラムを持つ大学も多いです。データサイエンス学部で幅広く学ぶことが１つの王道です。

２つめは，自分の興味あるテーマの学部に進学したあと，データサイエンティストを目指す方法です。最近は，各領域の専門家がデータ分析を学び，データサイエンティストになるケースが増えています。例えば，経済学の専門家がデータ分析を学び，経済データを分析して政策立案に活かしたり，医療の専門家がAIを学び，AI創薬やAI医療の分野で新しい会社を起こしたりしている事例がたくさんあります。興味があることを学びながらデータ分析の使い方も学び，双方の知識をかけ合わせて活用するのも良い道筋です。

5．生成AIとは

生成AIの話に入る前に，まずAIを紹介します。AIはArtificial Intelligenceの略で，日本語では人工知能と言います。AIの定義には様々な流儀がありますが，その最も広い定義が，「人間と同等かそれ以上の情報処理を実現するシステム」です。実は，AIは，実現するとAIと呼ばれなくなる宿命にあります。例えば，日本語かな変換・検索・乗換案内などは，それが実現する前まではAIと考えられていましたが，今では当たり前過ぎてわざわざAIと呼ばれません。画像生成AIがAIと呼んでもらえるのも向こう数年だけの話で，すぐに，「アプリを使ったらほしい画像を生成できるなんてのは当たり前」になり，AIとは呼んでもらえなくなるでしょう。

そんな短命な宿命のAIですが，ちょうど2025年の今は，生成AIと呼ばれるAIが流行しています。生成AIとは，何かを生成するシステムです。対話AIであれば返答の言葉を生成し，画像生成AIなら画像を生成します。他にも最近では，動画生成AIや，３Dモデルの生成AIなど，多様な生成AIが登場しています。これらの生成AIには，現在の技術では，大きく２つの流儀があります。それが，自己回帰モデルと拡散モデルです。これらの仕組みについて，次章以降で説明します。

6．自己回帰モデルの生成AI

自己回帰モデルの生成AIは，ChatGPTなど，主にテキストを生成する生成AIに用いられています。特に言語系に用いられる場合は，大規模言語モデル（Large Language Models / LLM）とも呼ばれます。この大規模言語モデルがテキストを生成する仕組みは以下の通りです。まず，大規模言語モデルでは，文章をいきなり生成するのではなく，１単語ずつ順番に生成していきます[1]。インターネット上に膨大に存在する文章を基に，「今までに入力された文章を基にすると，次はどの単語が妥当か？」という問題を高い精度で解くことで，自然な対話を実現するのみならず，非常に幅広い知識を持ってどんな疑問にも答えてくれる能力を手にしました。

しかし，よく考えてみると，「次の単語を高精度に予測する」だけで，あれだけ知的な応対ができるのは非常に不思議ですよね。これについて，2019年に発表されたGPT−2の例を紹介します（初代ChatGPTがGPT−3.5と言われているので，

[1] 正確には，単語よりもう少し細かい「トークン」と呼ばれる単位で生成を行っています。

GPT-2はその1.5世代前のモデルです)。次の文章のXXXXに入る単語を予想してみてください。

「スーツケースにトロフィーが収まらなかったんだよね。XXXXが大きすぎたからね。」

この問題は，人間には簡単に解けるでしょう。トロフィーをスーツケースに収めようとしたものの，大きすぎて収まらなかったので，大きすぎたXXXXはトロフィーとわかります。ですが，これをコンピュータープログラムで判断するのは至難の業です。なぜなら，XXXXにスーツケースを入れてもトロフィーを入れても文法的には問題がないので，どちらを入れるべきか判断がつかないからです。これを正確に判断するためには，スーツケースが容器でトロフィーが中に収めるものであるとか，収めたいものが容器よりも大きいと入らないなど，現実世界の常識をコンピューターに教え込む必要があります。通常このような常識は膨大にあるので，コンピューターに全て正確に教え込むことは難しく，この問題はなかなか解けなかったのです。

ここでGPT-2の登場です。インターネット上の膨大な文章を学習したGPT-2では，XXXXに「スーツケース」が入る確率と，「トロフィー」が入る確率を計算することができます。そして，XXXXに「トロフィー」が入る確率のほうが高いと計算されます。つまり，GPT-2の研究で，インターネット上の膨大な文章を学習したモデルには，この手の常識が自然に組み込まれるという現象が発見されたのです。

これに加えて，2020年に発表された「スケール則」の論文が決め手でした。ものすごく大雑把にスケール則を説明すると，「２倍学習すれば２倍賢くなる」「10倍学習すれば10倍賢くなる」「この法則は学習量を何倍にしても成立すると思われる」という発見です。つまり，大量のお金を用意して，大量の計算機を買って，たくさん優秀な人を雇って，大量のデータを用意して，とにかくたくさん学習すれば，今までに見たこともないような賢いAIを作成可能だということです。こうして作られているのが，現代の大規模言語モデルなのです。スケール則は全くとどまるところを知らず，向こう数年はAIの性能はいくらでも上がり続けるでしょう。

７．拡散モデルの生成AI

拡散モデルの生成AIは，画像生成AIや動画生成AIなど，基本的には言語系以外の生成AIに用いられている技術です。ここでは，画像生成AIを例に拡散モデルの仕組みを紹介します。コンピューターでは，画像のデータは，ピクセルと呼ばれる小さいマス目の色を指定することで定義されています。例えば，100×100の画像であれば，合計10,000個のピクセルがあり，各ピクセルが何色かを記録することで，画像のデータが保存されています。色のデータは，光の３原色である赤・緑・青の色の強さで指定されているので，１ピクセルあたり３つの数値でデータが指定されます。そのため，100×100の画像であれば，100×100×３＝30,000個の数値が１つの画像に対応します。

とは言え，30,000個の数値をでたらめに用意しても，意味ある画像にはなりません。例えば本物の顔写真のデータなら，その画像の大部分のピクセルはその人の肌の色になるはずですし，隣接するピクセルは似た色になるなど，データには一定の規則が現れます。完全にランダムに30,000個の数値を用意して作った偽物の画像の場合，このような規則がある普通の画像は生じないのです。

ここで，拡散モデルの登場です。拡散モデルでは，インターネット上にある大量の画像を用いて学習することで，画像データが「本物っぽい度合い」を計算することができます。そして，これを活用すると，様々な画像を生成できるようになるのです。実際には，以下のプロセスで画像を生成します。まず，30,000個の数値を完全にランダムに決めて，偽物の画像を作ります。次に，拡散モデルを用いて，画像の各ピクセルの色をどのように修正したら「本物っぽい度合い」が高まるかを計算します。この計算結果を用いて，各ピクセル

の色を少し修正して，ちょっとだけ本物に近づけます。この作業を何回も繰り返すことで，最後には本物のような画像が生成できるのです[2]。

実は，この修正の方向性は自由に調整できます。ですので，「美味しいドーナツを生成したい」と指定すれば，「本物の美味しいドーナツっぽい度合い」を高めるように各ピクセルの色を修正し，美味しいドーナツの画像を生成できるのです。

8．10年くらい先の未来予測

最後に，データサイエンスと生成AIが発展した未来について，確実度の高い予測をいくつか紹介します。10年先と言えば，今の高校生が25–28歳の頃です。仕事に慣れてきて，人によっては大きな仕事をはじめる頃です。大きく分けて，宇宙・ロボティクス・X情報学の３つを紹介します。

まず，宇宙開発が大きく進展していると思われます。SpaceXがロケットの再利用を実用化させました。巨大なブースターが難なく着陸し，再利用される様子はもう見飽きた人もいるでしょう。飽きたということは，常識化したということで，AIがAIと呼ばれなくなる現象に似ています。10年ではまだ無理だと思われますが，2040年くらいには宇宙旅行は大金持ちなら行けるくらいまで安くなり，2060年くらいには大金持ちが月旅行くらいなら行けるようになるかもしれません。ここまで発展するかはわかりませんが，向こう10年は，宇宙開発は伸び続けます[3]。

ロボットも圧倒的に発展することが確定している分野の１つです。高校生の方々は，AIなんかよりロボットを勉強したほうが良いかもしれません。ロボットが発展する理由は大きく分けて２つ，AI開発と，工場の自動化です。AIは，インターネット上の膨大なデータを学習してここまで賢くなりましたが，りんごを見たこともなければ，雨に濡れて寒いと思ったこともありません。実体験の不足が，今のAIの致命的な弱点だと言われています。ですが，そんなもの，ロボットにAIを乗せて野に解き放てば解消される問題ですよね。というわけで，AIのさらなる発展のためにも，ロボット開発はどんどん進んでいきます。また，工場の自動化のために，人型ロボットが確実に発展します。工場は，人間が働く前提で設計されています。そのため，新規のロボットを個別の工場別に作るのではなく，汎用人型ロボットを作ってすべての工場に解き放ったほうが，一気に自動化が進みます。つまり，売れるのです。その市場規模は膨大なので，汎用ヒト型ロボットの開発も確実に進みます。

最後にX情報学です。英語ではXInformaticsと言い，Xの研究領域でデータ・AIを活用して研究を進める分野です。このXには生物や材料などが入ります。例えば生物情報学なら，各個人の体のデータを解析して，個人別に最適な薬を調合することができるようになります。例えば材料情報学では，AIを用いて新素材の候補物質を発見することができます。この技術が発展すれば，今まで治らなかった病気が治り，というかそもそも病気になる前に薬が処方されて病気にならなくなり，電池が全然切れないスマホがとても軽い材料で作られ，通信速度も今の100倍！ なんてことが可能になります。

総じて，どの領域でも，データの力をAIで解き放ち，今日の不可能が明日の可能になり，僕たちが想像するSF的未来がどんどん実現されていきます。この時代は，生きているだけで新技術が出てくるので，とても楽しい時代となるでしょう。また，その時代を前に進める人になることもできます。当然，誰もが技術者になるわけではないですが，技術者になりたい人にとっては，人類史上最も面白い時代だと言って良いでしょう。

---

[2] この説明は，拡散モデルをエネルギーベースモデルと解釈し，スコアを用いたフローでノイズ除去するプロセスを表しています。他には，拡散モデルはノイズ除去のモデルだという説明も有名です。

[3] 本題ではないので詳しくは書きませんが，宇宙は国防と直結する上，資源獲得の夢があるので，確実に発展し続けます。