AIを活用したスマート農業
―その実現のために必要なことと将来展望―

岩田　洋佳*

令和２年４月21日，当会は一般社団法人農林水産奨励会との共催により，東京大学大学院農学生命科学研究科の岩田洋佳准教授をお迎えして，第１回農業懇話会を Web 会議の方式で開催しました。これは，新型コロナウイルス感染症の流行拡大により政府より出された緊急事態宣言に基づいた東京都の要請を受け，通常の講演会から持ち方を変更したものです。以下は講演と質疑応答の内容です。

講　演

１．はじめに


岩田　洋佳 氏

世界人口の増加傾向は続いており，近い将来食料が世界的大問題になるだろうといわれています。つい最近までは，2050年の人口が80億人を超え，その頃で頭打ちだろうと言われていましたが，2017年の推計では97億人を超え，2050年以降も増加し続けるという予測になっています。

また，地球環境も変動してきて温暖化の進展や各種の自然災害の頻発が問題となってきており，こういう中で農業を営んでいくことが容易ではなくなってきていると思います。

それに加え，国連食糧農業機関（FAO）が2011年に出している Livestock in food security（畜産と食糧の安全保障）の報告では，世の中に中流階級の人が増えているので肉の消費量が上がると予測しています。2010年比で2050年には1.5倍，場合によっては２倍以上になると考えています。そうすると，家畜の餌と人間が食べる穀物との競合も当然生じてきます。

それとコインの表裏だの関係になりますが，食料が不足傾向になると，農業がビジネスになるという考え方が出てきます。例えば，経済産業省（2016）の試算では，2030年にはバイオエコノミー（生物資源やバイオテクノロジー活用した経済活動）は180兆円（食料関係が36％）の巨大市場になると予想されています。

２．AIと機械学習・深層学習

最近よく耳にする AI とは Artificial Intelligence の略語で，「人工知能」と訳されます。一般社会では，この AI の中に機械学習や深層学習も含めて考える場合も多いようです。なお，「機械学習」は Machine Learning の訳語で略して ML，「深層学習」は Deep Learning の訳語で，略して DL とも

＊いわた　ひろよし　東京大学大学院　農学生命科学研究科　准教授，（株）Quantomics　技術アドバイザ

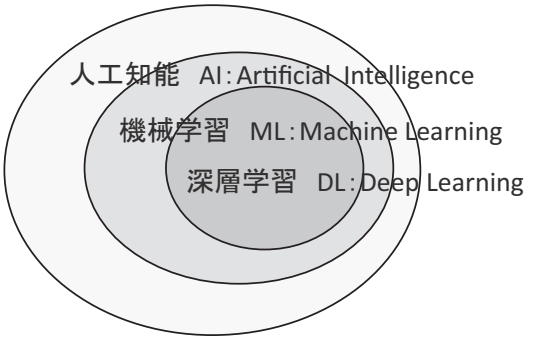

図１　人工知能・機械学習・深層学習の概念の包含関係

（Chollet（2017），巣籠 悠輔監訳（2018）より改変）

呼ばれています。

この深層学習が非常に注目されているところから，逆に AI という言葉が改めて注目を浴びているのです。つまり，AI が一番古い概念で，「AIって何？」という際には，押さえている範囲が最も大きいのが AI で，AI の中の一部が機械学習（ML），ML の中の一部が深層学習（DL）です（**図１**）。

AI と機械学習の違い

AI の一部が機械学習だと説明しましたので，それらの「違い」というのは変ですが，あえてどういうところが違うのか説明します。皆さんの中には1970年代，1980年代に「エキスパートシステム」という技術があったのを覚えている方もいると思います。これは，AI という言葉が一気に認知度を得た時代に期待されたシステムです。つまり，エキスパートシステムは，「専門家の頭の中にあるルールをコンピュータの中に記述し，次に，データを入れていくと，そのデータをルールによって処理して，『そのデータの場合にはこう対応しなさいね』と答えるシステム」です（**図２上**）。

このエキスパートシステムには非常に大きな期待が寄せられました。例えば，私は育種が専門なので，育種家が持っていて，頭の中で判断しているノウハウを全部コンピュータ上に記述します。あるいは，生育に障害を受けた作物を見て何の病害に罹っているか判断している植物病理の専門家の頭の中を AI で置き換えようとします。これによって，今まで専門家がやってきた仕事をコンピュータに置き換えられるのではないかという期待が起こったわけです。

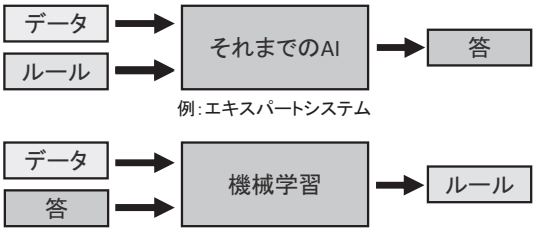

図２　人工知能と機械学習の違い

［Chollet（2017），巣籠 悠輔監訳（2018）より改変］

ただ，これが簡単ではなかったのは，ルールを記述することが実は非常に難しいことが分かってきたためです。そのため，エキスパートシステムには大きな期待がかけられて，それなりの成果も挙げたのですが，いったん下火になりました。

そこで次に登場したのが機械学習です。これは，「ルールを学習しよう」というものです。つまり，元のデータを入力すると答えを出していく「ルール」があって，元の「データ」をその「ルール」に入れたら「答え」が出力されてくるのではなくて，元の「データ」とその「答え」を入力して，その間に存在する「ルール」をあぶり出そうというのが機械学習です（**図２下**）。

機械学習の仕組み

図３は，機械学習の仕組みを非常に簡単に説明したものです。今，元のデータは x_1，x_2という２次元の平面上に散らばっている点の位置です。一方，答えは点の色です。実は，機械はまだ知らない状態なのですが，（例

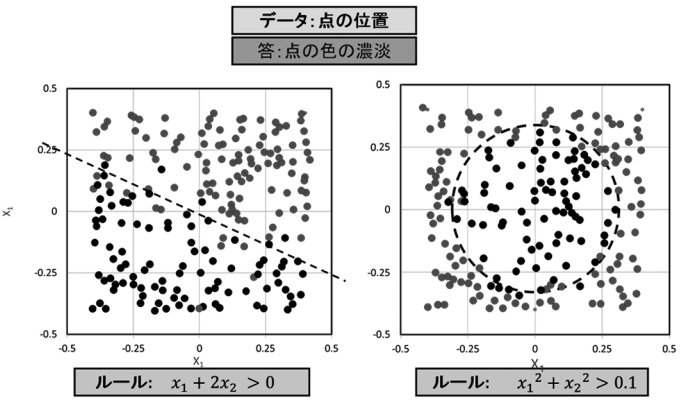

図３　「データ」と「答」から機械学習により「ルール」を導出

（左：例１，右：例２）

の役割です。

さらに，一時期一世を風靡したサポートベクターマシンという機械学習の仕組みを使うと，上述したようなデータを直線で二分するとか円形のような単純な形の枠の内と外に分けるような単純なルールから，渦巻きの模様のようなかなり複雑な（非線形の）境界線で二分するルール（**図４下**の例３）も抽出することができます。

機械学習では，データが入力されると，そのデータに対してある変換を行って，そのデータを少し違う次元の空間に持っていきます。例えば，横軸 x1と縦軸 x2という平面の中でここのデータの位置を考えるのではなく，例４（**図４上**）の場合だと，個々のデータを〖x_1〗^2+〖x_2〗^2と変換をしてやり，横軸を x_1，x_2，縦軸を〖x_1〗^2+〖x_2〗^2とした空間を考えると，前述した円で囲まれた部分が縦軸のある値より下，外側が上と非常に簡単に両者の範囲を分けることができます。

今回示した３例のうち，最初の例では，直線で切り分け，例２では円形の中と外に切り

１）では，x_1+ 2x_2>0 だと点の色は灰色になるというルールを裏で私が設け，そのルールに従って私が答えを作っています。そして，このデータと答えを機械学習します。そうすると，機械がルールを編み出します。実際には，機械が境界線を学習してどこかに境界線を引くというルールを導き出します。統計学に詳しい人に種明かしをすると，ここでは機械の中に「多項モデルのロジスティック回帰」という回帰分析の拡張法が仕組んであります。これも一つの機械学習の仕組みです。

上述の**図３左**（例１）では，あぶり出されたルールは一次不等式だと推定されましたが，**図３右**（例２）のようなデータと答えが与えられた場合には，黒丸のデータが分布している真ん中辺りを原点として，灰色の点が〖〖x_1〗^2+〖x_2〗^2>0.1〗_ であるという条件のルールだとも学習してきます。つまり，元データと答を与える（入力する）ことにより，その背後にあるルールをあぶり出すというのが機械学習

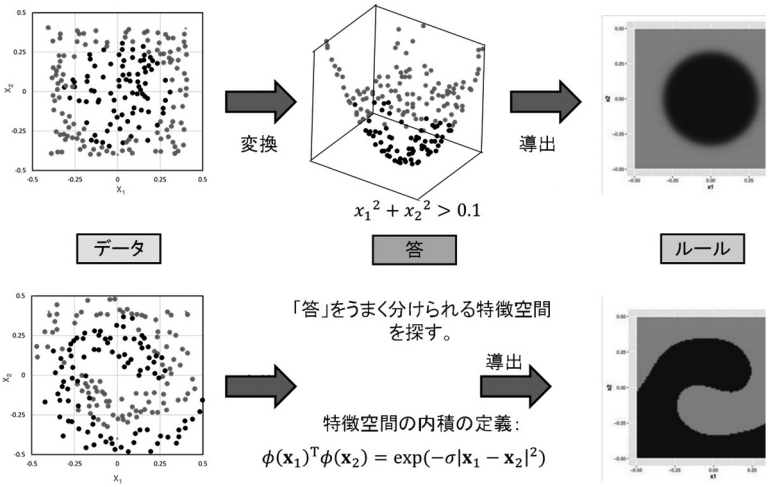

図４　次元を上げた特徴空間で「答」をうまく分けられる「ルール」を導出

［例３（下）：次元を上げて複雑な答えを判別，例４（上）：平面を３次元化した特徴空間］

分けるという非常に簡単なものでしたが，そうではない場合でも，何らかデータ変換をすると，そのデータを分けるルールがあぶり出され，それを元の空間に戻してやればルールが導き出せることになります。

例３（**図４下**）の渦巻きは「特徴空間における内積の定義」という複雑なことをやっています（例１や例２の学習方法を一般化して拡張しています）。ここでは，その内容を理解していただく必要はありませんが，下式のような変換をします：

〖"ϕ(" x_1)〗^T "ϕ(" x_2)=exp(−σ〖|x_1−x_2 |〗^2)

実はこの内積の定義ができると，いろいろな統計手法がその特徴空間の中で使えるようになります。この式の中のＴの次元は無限にとれるので，一旦非常に高い次元に持っていって，そこでシンプルにデータの区分けができるようになると，そこから元の２次元データに数式を戻してくることにより，渦巻き型のような複雑なルールも導き出せます。

機械学習と深層学習の違い

先ほど**図１**で説明したように，深層学習は機械学習の一種ですが，どこで区別しているのでしょうか。簡単に言うと，「深層」(Deep)という言葉がそれを表しています。

深層学習の概念を**図５**に示しましたが，先ほどの機械学習における元データと答えと似た話になります。深層学習では，入力というところから元のデータが入力されます。また，「真値」というのが機械学習で言う答えにあたります。そして，その両者の間に整合性が合うようにさせた予測値が計算されます。大事なところは，データを変換するところです。

データ変換自体は，機械学習のところでもありましたが，変換するところが多層になっていて，何度もデータ変換を行いながら最後の予測値に結び付けるのが深層学習の特徴的なところです。先ほどのサポートベクターマシンは，こういうデータ変換の層が１カ所あるいは２カ所程度しかないシャローラーニング（Shallow leaning, 浅層学習）と呼ばれているのに対し，深層学習では，ここの部分が分厚いのが特徴です。この分厚いデータ変換層では，情報の「蒸留」と新たな組み立てをやっています。

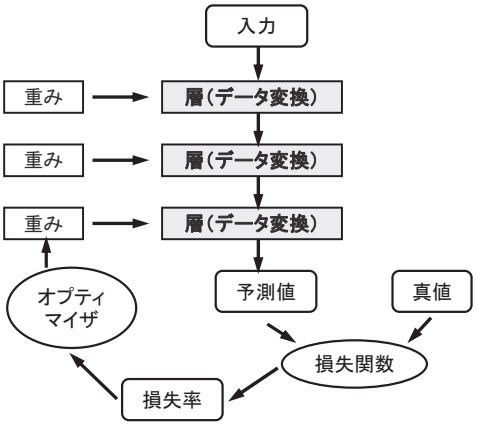

図５　深層学習におけるデータの蒸留方法

［Chollet(2017)，巣籠 悠輔監訳(2018)より改変］

機械学習ではデータ前処理が必要

一旦，機械学習の場合を振り返ってみます。例えばイネの籾の形を見てそのイネがどの品種なのかを予測したいなら，まず特徴を抽出するため，画像解析により，籾の長さ，幅，面積，周囲長，面積/周囲長比，籾の色合い等を定量化して抽出し，その情報をさらに，「特徴エンジニアリング」と呼ばれる方法で要約します。

特徴エンジニアリングにはいろいろな方法がありますが，複数あるデータの種類を主成分分析という方法で要約・圧縮して減らす方法があります。また，先ほどのサポートベクターマシンでは元の複雑な形のデータをカーネル法という方法で次元を上げて単純な形に変換して計算し，そこで得られた特徴を元の

次元に戻すという方法で要約を行います。このような方法でまず画像などの元データの前処理を行ってから，次のステップである「学習」に使います。これが機械学習です。

深層学習ではデータ前処理が不要

ところが，深層学習では，画像データを学習する際，前処理なしでいきなり画像データを入力しても直ちに学習ができます。

つまり，人間が画像を見た時に頭の中で事前に画像の中からいろいろな特徴を抽出して要約する画像解析の操作を深層学習では，その多層の階層の中でよきにはからってやってくれ，データを蒸留して自動的に組み立ててくれるというところが特徴的なところです。ここがいわゆる機械学習の中でもシャローラーニングと言われるものとの大きな違う点です。

深層学習の高性能化，大衆化

このような画像の要約ができるようになると，例えば，複数の動物の画像からゾウを見つけてくることも可能になります。そこでは，最初の層でゾウのいろいろな部分の形が抽出され，それを何らかの形で組み立て，最終的には特徴的な耳や腰の形が抽出され，「これって，ゾウだよね」というように分かるようところまで学習するのが深層学習の特徴的なところです。

画像を分類する能力のコンテストが2010年から行われています。これは画像が与えられて，何が写っているかをリストアップし，対象としているものが入っていれば正解とするコンテストです。深層学習が登場したのは2010年で，最初はエラーが多かったのですが，それ以降は深層学習の手法の改良が進み，2015年には人間の能力を機械が超えてきています。その後もさらに改良が進み，2017年には，元々エラー頻度が0.28もあったのが0.023と10分の１の誤差になって高性能化が進んできています。

また，私が最近驚いたのは，DeepL という無料でも使える翻訳ソフトの精度が高いことです。私も，よく学生の草稿論文を添削しますが，その草稿よりも DeepL が書いた英文を直す方がずっと楽なのではというぐらい優れた能力を有しています。

こんな状況なので，深層学習に人が飛びつかないわけがないので，インターネット書店で Deep Learning と検索すると，相当数の関係書籍が出てきます。Deep Learning と英語で書いてあって，ちょっと小難しそうに見えるものから漫画の混じるものまであって，大衆化が進んでいます。

GPU の発達が大衆化を加速

深層学習が一気に大衆化した理由を少し説明します。コンピュータゲームでは，最近はリアルな画像なのかバーチャルなのか分からないぐらい画像が非常にきれいになっているのをご存じかもしれません。それは，GPU（Graphical Processing Unit）という元々はグラフィック処理専用のプロセッサをうまく使うことによって，今までお話ししてきた深層学習の計算が高速度でできるようになったからです。

近年，この GPU が非常に安価になり，また TensorFlow や PyTorch，Keras のような使いやすいパッケージソフトが出てきて，市販の参考書を見て４，５日それらのソフトを触った程度の学生でも，すぐに深層学習のプログラムを書いて試すことができる状況になっています。

例えば，「Deep Learning で上司の顔を認識して画像を隠す」というおそらく2016年が最終更新の非常に有名なホームページがあります。どこまで本当なのか分からないのですが，ある会社員が上司の顔をウェブカメラで

撮って学習して，その上司が歩いて近づいてくると知らせてくれて，仕事と関係の無い画面も隠してくれるというシステムを深層学習によって作ったという記事です。まさに大衆化が進んでいることが分かる典型的な例です。

AIの夏と冬

実は深層学習が出てきたおかげで，現在は「AIの夏」といわれています。夏とわざわざ言うのは，「AIの冬」があったからです。例えば，先に言及したエキスパートシステムは，それが悪かった訳ではないのですが，それで可能になると思われた期待感が高すぎたのと人間の頭の中にあるルールを明文化することが難しかったため，その後，一度「AIの冬」がありました。期待していろいろ投資をしたのに，結局，あまり使えなかったのです。

現在は，また新しく「夏が来た」といわれています。しかし，今後また「第３の冬」[注1)]が来るという恐れも十分あります。それは，深層学習に対する大きな期待感から，逆に期待を裏切られた感が出てこないかと心配されているからです。

そうならないようにするには，使う側がAIをどう使えばいいかを理解することが重要です。

３．AIを用いた育種の高速化

ここでは，作物育種の過程でAIの技術である機械学習がどう使えるのかというお話しをします。

まず，作物の育種を行うというのは，親同士を交配させて，その後代（子孫）から良い系統を選ぶことですが，実はそんなに簡単ではありません。もちろん，交配して新しい後代を作るのにも時間がかかりますが，その特性を評価して，中から良いものを選ぶのにも時間がかかります。それは栽培試験に時間が必要だからです。

例えば年１回栽培できる作物だと栽培試験による評価は年１回しかできません。果樹のように果実の評価を行えるまでに何年間も待つ必要のある作物もあります。私はスギのゲノムプロジェクトに参画したことがありますが，スギの場合は評価できるまでに20年，30年待たなければならないと言われていました。このように，育種の場合，栽培試験による評価の結果を待つというのが大きなネックになります。

機械学習を育種に組み込む

既に説明したように，機械学習の仕組みにより，それぞれの評価候補となる植物体のDNAのデータを「元データ」，栽培試験で得られるそれぞれの植物体の形質についての評価項目のデータを「答え」として作った仕組みを示したのが**図６**です。植物体のDNAデータというのは，DNAに書き込まれているはずの遺伝的な性能の情報ですので，このDNA情報と植物体の性能を突き合わせる。つまり，DNAの情報をデータとして与えて，それからそれに対応する植物体の性能を答えとして与えて，その間にある「ルール」を機械学習させます（実際の機械学習では，量的遺伝学で使われる統計手法を用いることがほとんどです）。

そうすると，今度はそのルールを用いて，新たな植物体からはDNAのデータだけを取り出しておいて，そのルールを当てはめると，そこから答えにあたる，本来栽培試験により初めて判明するはずの各植物体の形質を導き出す（推定する）ことができます。

注１）AIには既に70年代に１回目の冬があった。これは，当時のコンピュータの推論・探索能力が期待ほどではなく，成果があまり出なかったためである。1995年ごろからAIは２回目の冬の時代を迎えた。

こういう手順を取ると，かなり効率的に答えを導くことができて，例えば育種の場合には一番時間がかかっている栽培試験による形質評価のところをショートカットできるようになります。

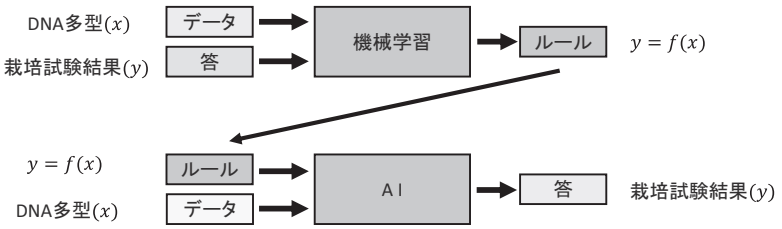

図６　AI を用いた育種の高速化の仕組み

ソバのゲノミック育種の例

普通ソバのゲノミック選抜（多数の DNA マーカー情報から特性を予測する選抜法）に機械学習を導入しようと，トヨタ自動車，筑波大学，京都大学と５年間のプロジェクトを行いました。具体的な選抜のスキームでは，年に１回は圃場で栽培ができるので，まず，圃場で栽培している植物体から DNA を採取してデータとなる DNA 多型の情報[注2)]と，開花までの日数，花数，千粒重などの答えとなる収量関係形質の情報を収集して，これらのデータと答えから機械学習によりルールを得ます（**図６上**）。

得られたルールを用いて圃場で栽培している世代の集団から選抜した植物体で採種を行い，その種子（GS1世代）を用いて温室で２作目を栽培します。温室なので，屋外で栽培しないと分からない答えは得られませんが，先ほど得ておいたルールに基づいて選抜をします（**図６下**）。

この仕組みを使うと，栽培試験をせずに（答えを試験栽培で実際に求めなくても）選抜が出来るので，時間短縮ができるというところが重要です。そして，年に２回の選抜を３年間行い，GS6世代まで世代が進めたところでの個体当たりの種子数は，元の集団（母集団）

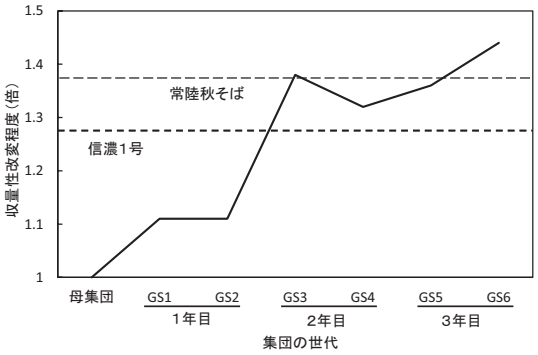

図７　ソバの母集団を基準とした各ゲノミック選抜集団（GS）の収量性改変程度

［Yabe, Iwata et al.（2018）より］

の1.44倍にまで改良できました（**図７**）。

イネの世代短縮育種に適用

実は，温室のような人工環境を使えば，育種材料の世代を１年にかなり進められます。イネでは年４世代が可能なので，例えば，両親の交配によって得た子供（後代）の集団の内，一部（兄弟に相当）を圃場で栽培し，各植物体から DNA 多型のデータと答えに相当する諸形質のデータを得て，それらから機械学習でルールを求めることができます（**図６上**）。

一方，温室で１年に４世代回して栽培している残りの兄弟植物にこのルールを当てはめると（DNA 多型のデータで選抜），年４回の選抜により品種改良を進めていくことができます（**図６下**）。

注２）複数の植物体間で DNA 配列が異なる部分を「多型」といい，多型の存在する部位が栽培特性の差異を起因する遺伝子の存在部位の候補となる。

機械学習で超優良系統を選抜

このように機械学習を導入するとどのぐらい育種の効率が上がるか評価してみます。実際に育種を行う場合，花の色のように遺伝だけではっきり性質が決まる特性は少数派で，収量性や草丈のような量的な性質の多くは，育ち方（環境）によって決まる部分と遺伝によって決まる部分があります。全体の内，遺伝的な要因で決まる割合を遺伝率といいますが，例えば，収量性に関わるような形質ではおおむね30～40％が遺伝によって決まるとされており，残りは生育環境のばらつきの影響で決まります。そうすると，栽培試験によってこの1,000系統のトップを選抜するのは実際にはかなり難しいことになります。

図８は，遺伝率が0.2と先ほどよりもさら厳しい状態を想定して，各世代100個体の内からトップのものを選抜する循環選抜の効率をシミュレーションした結果を示しています。すると，10世代ぐらい経過した時点，つまり３年目に入る頃から，その選抜系統の能力は，図中に破線で示されたレベル，すなわち従来の方法で1,000個体中のトップを選んできた系統の能力を超えることが出来る結果となりました。したがって，機械学習を用いて非常に高速に品種改良できることが分かります。

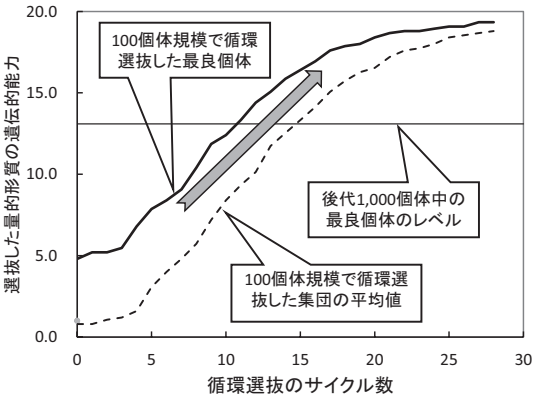

図８　シミュレーションで予測したゲノミック選抜の効率

（２品種交配の後代で，量的遺伝子座48個，遺伝率＝0.2，100個体規模の循環選抜を想定）

ナシの品種改良の例

果樹の育種を行う場合，実際に交配して後代を作るのは大変な時間と労力が必要です。しかし，それをコンピュータの中で仮想的に行って，どんな形質を持った子どもが出てくるのかを予測することが出来れば，育種家にとって重要な情報になります。

例えば，ナシの品種が「幸水」を含めて80ほどあり，「幸水」よりも早生で果実のサイズも大きいものを育種しようとする場合，従来ですと非常に多くの品種同士の組合せで交配を行い，それらの子孫を果実が結実するまで栽培し早晩性や果実の大きさを計測する必要があり，膨大な手間と時間がかかりました。

しかし，機械学習により，各品種の DNA 多型のデータから早晩性や果実の大きさを推定できるルールが導き出されておれば，コンピュータシミュレーションにより任意の品種間の組合せによる子孫の性質を推定できます（**図６**）。

実際の機械学習の方法は，メンデル遺伝の法則を使っているだけなのでそんなに難しい理屈ではありません。例えば２種類の遺伝的な性質が親から子に同時に遺伝する場合，２つの性質は「遺伝的に連鎖している」と言います。また，連鎖状態にあった２つの遺伝子が組換えによって分離し，親から子に同時に遺伝しなくなる現象も遺伝子の組換えとして知られており，その現象はある確率で起こります。そして，その確率は実験で求めることが出来ます。そうして親の有する遺伝子の情報と組換え確率の情報があれば，子どもがどのような DNA を持つようになるかという答えはコンピュータ上で極めて容易に求められます。

実際の操作では，ナシ80品種の植物体から

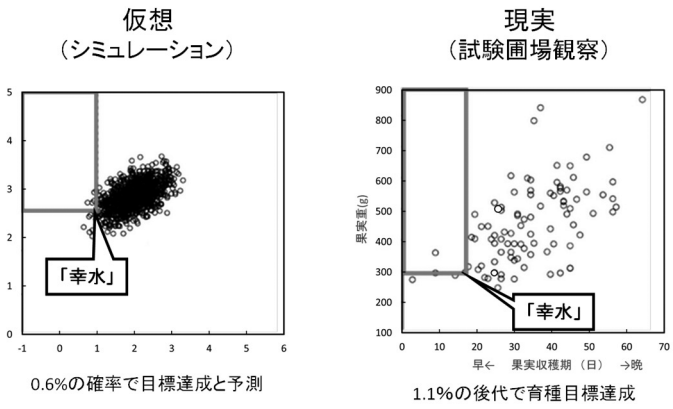

図９　シミュレーションによるナシの交配後代の能力予測（左）と実際の交配結果（右）

（「幸水」よりも早生で果実の大きいものが育種目標，Iwata et al., 2013）

DNA のデータを取り，各品種の特性を答えとして，両者からルールを求めておきます（**図６上**）。次に，親の交配で出てくる子供（後代）の DNA データをコンピュータのシミュレーションで求めておいて，上記のルールに当てはめると，実際には交配をしていないのに，ある親同士を交配したときにどんな特性の子供が生まれるかを予測できます。

それをコンピュータ上でシミュレーションを行ったのが**図９**（**左**）です。横軸が収穫期，縦軸が果実サイズです。そうすると，「幸水」よりも早生で大きな果実を付ける子孫（太枠の中）がわずかですが存在するはずと予測されました。そして，候補となった品種の組合せの子孫を実際に試験圃場で栽培して評価を行ったところ，早生で果実の大きな系統が得られました（**図９右**）。つまり，これは，機械学習で得られたルールを使うと，今までできなかったことが可能になったという例です。

ソルガム育種の例

バイオエネルギー作物としても使われているソルガムという飼料穀物では，世界中のあらゆる不良環境に耐性を持った品種の開発が期待されます。そこで同じように，作物体から DNA 多型のデータを取り，一方，世界中の不良環境がある土地における栽培実験により得られた特性を答えとすれば，機械学習でその答えが出てくるルールが分かります（**図６上**）。ルールが分かれば，不良環境でわざわざ選抜しなくても，日本の中で DNA のデータだけで選抜を行い（**図６下**），不良環境にフィードバックできるという枠組みを考えました。

私たちはこの枠組みにより，東大の堤伸浩教授と科学技術振興機構の CREST プログラムにより，塩害耐性が非常に強いソルガムを他の作物が全然育たない塩害地で栽培してバイオマスを作るというプロジェクトを行いました。そこでは，メキシコの塩害地で栽培試験して得られたソルガム系統の栽培特性を答えにして，日本で取ったソルガムの DNA 多型のデータからルールを抽出しました（**図６上**）。ルールが導き出されると，沖縄の温室で育てた苗から取った DNA のデータだけで選抜を行い，それを交配する操作を繰り返しました。そうすると，メキシコのための品種を日本で効率的に育種することが可能になりました。

機械学習応用のためのポイント

以上の例から，機械学習とは，データと答えからルールを抽出し，そのルールを使って新しいデータから答えを早く導き出す，つまり，ルールの発見と利用という枠組みであることを理解して下さい。

４．AI 利用の注意点

訓練データの重要性

AI を使った農業あるいはスマート農業をうまく発展させるには，農業のどの場面で AI の枠組みをうまく使っていくかというところに知恵を絞ることが重要になります。中

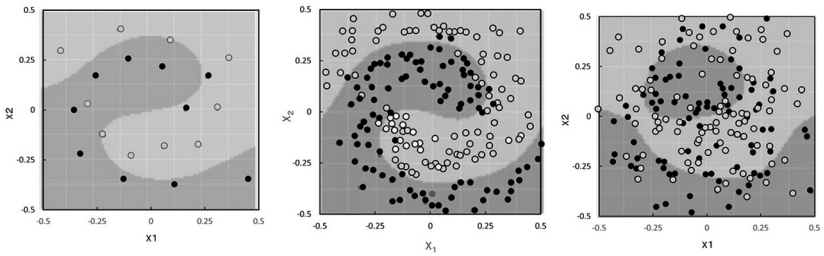

図10　訓練データの質と量が導出されるルールに及ぼす影響

でも注意すべき点は，訓練データを作るところです。**図10**（中央）は**図４**と同じ渦巻きのデータです。このようなデータと答えのセットを「訓練データ」と言います。こう称するのは，機械学習において学習しようとしている機械を訓練するためのデータだからです。当然，機械を訓練するデータの量や質は導出されるルールの質に影響してきます。

例えば，**図10**（**左**）のようにデータ数が少なければ，ルールとして認識される渦巻きの形も曖昧になり，一方，データの質が悪くなれば，境界がうまく見つけられず，**図10右**のように渦巻きであることを認識できなくなります。

これと同じことはスマート農業を行う場面でも重要で，いかに質の良い，あるいは量の多いデータを訓練データとして得るかということが重要になります。先ほどのゲノムのDNA データを基にしたルール作成とそれを用いた選抜という話にしても，そのネックになるのは訓練データ作成の部分です。訓練データの中には当然，答えの要素が入っていますが，この答えというのは本来栽培試験により出てくるものなのでむやみに省略できないからです。

訓練データ取得にひと苦労

そもそも，私がなぜ機械学習の利用を思いついたかというと，栽培試験の結果を得るのに非常に時間がかかって全体のボトルネックになっていたからです。代わりに，データから答えが導かれるルールを先に求めておけば，新しいデータをそのルールに当てはめるだけで答えが導き出せるというスキームを作りたかった訳です。ところが，このルールを得るための栽培試験データ自体が訓練データとして必要なので，このデータもたくさん取らなければいけないのです。

結局，訓練するためにはデータを収集しなければいけません。そのためには，当然今までの栽培試験のやり方を効率的にすることも考えなければなりません。これらの問題意識から，例えば，2017年に「Nature」誌に出た記事では，ロボットを使った生体の計測（フェノタイピング[注3)]）や育種のデータ管理システムの話題が紹介されています。当然，歴史的に蓄積されてきたデータを紙からどう起こすかという話もありますが，ここではまず高効率フェノタイピングの話題を紹介します。

高効率フェノタイピング

写真１（**右上**，**口絵参照**）は私たちが使っているフェノタイピングのシステムです。マルチを敷いた上の栽培植物の列をまたいで計測器が移動していますが，**写真１**（**左上**，**口絵参照**）を見ると，計測器の上部両端のカメラでマルチの境目を検知しながら進んでいます。マルチを完全にまっすぐ敷くのはなかなか難しいので，マルチを見ながらまっすぐ進んでいくようにしています。また，この計測器に付いたアームを必要に応じて動かしなが

注３）植物体の遺伝情報ではなく草丈や開花日などの表現型を測定すること

写真１　高精度フェノタイピングの実際

右上：マルチを跨いだ自動走行計測システム，左上：計測システムの拡大図，
左下：上空で撮影するドローン，右下：測定データはバーコード管理
（岩田らの科技振興機構 CREST プロジェクトによる）

ら写真を撮ることもできます。

写真１（左下，口絵参照）では，上空にマルチスペクトルカメラを付けたドローンが飛んでおり，これで作物を撮影しながら圃場の作物の計測をしています。写真右下は収穫した作物の計測をバーコードで管理しているところです。このバーコードリーダーは全部タブレットにつながっていて，草丈の計測だろうがなんであろうが全部バーコード化してあるので，スーパーのレジ係のようにバーコードでデータを読んでいけばいいという仕組みです。

ドローンの利用

写真２（口絵参照）は先ほどのドローンで撮った動画のスナップショット（**左**）とその動画から作った３次元**図**（**右**）です。**右図**はリアルに見えますが，ドローンを飛ばして動画や多数の写真を撮り，そのデータから畑の３次元構造を再構築した仮想画像です。バイオエネルギー用のソルガムは大型で，本来計測が大変ですが，このシステムを使うと作物が毎日畑でどうなるのかを非破壊で計測することができます。

あらゆる農業生産現場でこういうことを行うのはすぐには難しいかもしれませんが，先ほどの「データ」と「答え」を準備するところには，こういう訓練データを準備していくことが必要になってきます。

データファーム

上記のような考えをさらに端的に強調したのが，東京大学の平藤雅之特任教授と郭威助教が推進している「データファーム」という考え方です。ファームなので農場ですが，そこは作物を収穫するのではなくて，データを収穫するファームです。つまり，このデータファームを各地に作って機械学習のために使う訓練データを収穫し，そこから前述したルールを抽出しようというものです。そのための特殊な機械や特殊なセッティングで栽培試験をやる必要があるので，それを実現し，そこで訓練データを取ろうという考え方です。こうやって，何か工夫しないと訓練データをうまく集めるのは難しいだろうというのが，私の見解です。

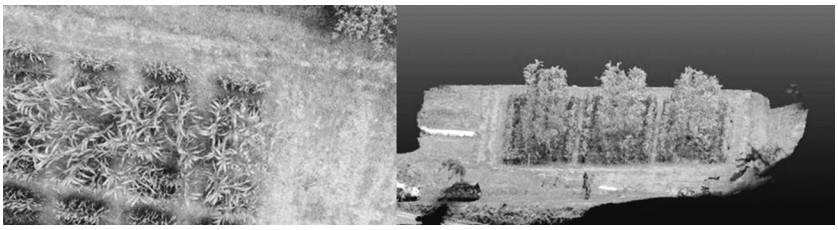

写真２　ドローンのリモートセンシング解析による経時的な作物生長の非破壊計測

（左：ドローンによる動画撮影のスナップショット，
右：動画の画像解析により作物の３次元構造を推定）

データのデジタル化

普段からやっている仕事の中にうまくデータが蓄積される仕組みを組み込んでいくことも重要で

す。最近の統計を見ると，日本種苗協会で登録されている野菜品種は9,056品種あります。ダイコンだけでも600品種を超える品種が登録されています。日本にはたくさんの種苗会社があり，野菜を扱っているところだけでも約100社あります。わが国の種苗産業の持つポテンシャルは非常に高いのです。ただし，いろいろな会社の方とお話をしていると，どの会社もデータのデジタル化があまり進んでいないように思います。

例えば，今でも育種材料の形質の計測は主に手で行い，データは野帳にて記帳し，それを表計算ソフトに打ち込んでいます。一方，ゲノムデータを使った解析を行っている種苗会社もたくさんありますが，そのデータはその部署で独自に取られています。さらに，栽培圃場に環境センサーを設置しているところでもそのデータは独自に取られており，データ処理がバラバラでなされているのが現状です。

このように，手で取って野帳に書いたデータをまた写したりすると，誤りが生じやすいのはもちろん，データの管理が不十分だったり担当者がいなくなったりすると，せっかく取ったデータが死蔵されたり継続性が潰えたりしてしまうということもあり得ます。また，ゲノムデータと環境データを一緒に解析しようにも，改めてデータの間の紐づけをしないと，当然，データ間の関連をモデル化もできないということになります。

このように，種苗産業は，いろいろな品種が持っている遺伝的ポテンシャルに関するデータを引き出すには最も良い場所なのですが，実は所有しているデータをうまくビッグデータ化することも，その恩恵にあずかることも，今はかなり難しい状況にあると考えています。

データ駆動型育種プラットフォーム

このような問題意識から，第２期 SIP プロジェクト[注4)]の中で，小さな企業でも，自社で得られたデータを蓄積しつつ，すぐに利用できるシステムを作るプロジェクトを行っているところです。

具体的には，社内あるいは社外（クラウドとして）にデータセンターを設け，そこに各種の API（各種のソフトウエアの機能を組み込み関数のように容易に共有できる仕組み，Application Program Interface）があって，データの入出力や視覚化，さらに，作物ゲノムのデータや環境のデータを紐づけながら作物の特性のデータに入力していくことができるようにします。そうすると，機械学習のところで説明したようにルール化ができるので，それを基に育種の意思決定をすることができます。

例えば，交配組合せを決める場合には，ある親品種同士なら，どのくらいの数の後代（子ども）を作れば育種目標に到達できそうかという意思決定にも利用できます。また，農業生産現場でもルールが得られれば，そのルールを基にした意思決定ができるようになります。ただし，それをやるには，その前提として，必要なデータをいかにうまく集めるかが重要になります。それには，前述したデータファームのような特殊な農場を使って収集するという考えもありますが，普段から行っている育種事業の仕事の中にうまくデータが集積されてくるシステムを作っておく方が近道です。

世界には，米国のコーネル大学やフランスの INRA（国立農学研究所）など，われわれと同じようなことを考えているグループもいます。例えば，コーネル大学で作っている

注４）平成30年から内閣府が実施している「戦略的イノベーションプログラム」

BreedBase システムは育種のデータ管理システムで，その中に各種の API が用意されています。われわれもコーネル大学や INRA と協力しながら，国内20社近くと一緒にこのシステムの開発を進めています。

ビッグデータのインパクト

このように，データが自動的にたまるシステムを作ることは極めて重要です。例えば，農研機構と私たち東京大学はリンゴの659系統（品種未登録）と既往185品種のゲノムデータと栽培特性のデータを入力してルール抽出用の訓練データを一緒に作っています。この段階ではまだ特定の形質の選抜を目的にしている訳ではないのですが，例えば，酸味など果実の特性をコントロールしている遺伝子がゲノム中のある位置にありそうだということが分かってきます。結局，大きなデータは大きなパワーになります。

農業の例ではありませんが，2013年の「Science」誌には，13万人のデータを使用して，人間の最終学歴とゲノムに関連が見られる何かの遺伝子の存在が示唆されるという論文が載って話題になりました。現在ではそれに113万人のデータが集められているので，ゲノム染色体上での遺伝子の位置がよりはっきり分かり，実はそれが言語理解に関わる遺伝子らしいということが分かっているようです。

結局，莫大なデータで解析できるようになると，個々の遺伝子で説明できる部分はそれほど大きくなくても，ある特性を説明できてしまう遺伝子が例えば３つ程度特定できてしまうというのが驚きです。

なお，上記の解析で研究者が使った13万人という数は非常に多いように聞こえますが，育種の現場で育種家が見ている材料，あるいは生産者が見ている植物の数はものすごい数です。そういうデータがうまく集まってきてそれが機械の訓練に使えれば，とんでもなく新しい情報として別の解析ができるようになるかもしれません。

ビッグデータを得るための工夫

実は，農業分野でビッグデータなるものはまだ世の中にはありません。ビッグデータを使ってスマート農業をという話はよく出てきますが，今はそれをどう収集するかということをよく考えなければいません。

そこで意識しなければいけないのは，前述したデータファームのような特殊な農場を作るという方法がありますし，あるいは，これまでの作業の流れの中にうまくデータが蓄積されるシステムを作っていくというのも一つの方法です。さらに，過去のデータをいかに掘り出すかということも重要です。農業というのは年に１回しかできないことが多いので，圃場試験を行ったとしても今後の20年間で20年分のデータしか取れません。しかし，過去のデータに遡れば，もっとたくさんのデータが集まるかもしれません。

データのラベリング

ここまで，機械学習のためのデータ準備が大変だと説明しましたが，実は機械に学習させるのも大変です。

写真３（口絵参照）は小麦の穂を撮影しているところです。ここでは右の写真の中に多数ある穂をバウンディングボックスという仮想上の四角い箱形で囲んで，機械が小麦の穂を認識できるように学習させています。この作業はアノテーション（正答のラベル付け，ラベリング）と呼ばれています。深層学習を行う際の「訓練データ」はこのようにして作成しています。実はこの作業は大変なので，こういう作業自身が巨万の富を生むビジネスになるとまでいわれています。

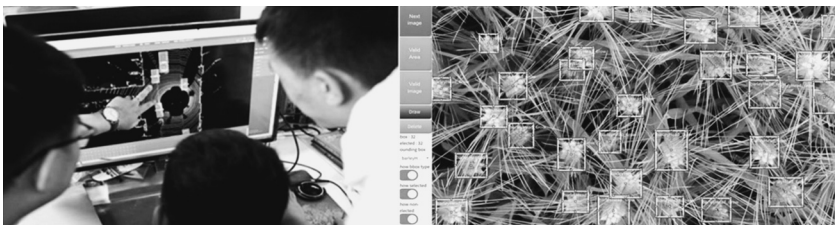

写真３　機械学習用の訓練データ作成作業

〔左：専門家が画像のアノテーションを行う，
右：小麦の群落画像から穂の部分だけを仮想のバウンディングボックスで囲む，
（東京大，生態調和農学機構，郭威氏提供）〕

ラベリング作業の改善

図11は**図10**と同じ設定ですが，**図11**（**左上**）では○と●は既にラベリングされているデータ，灰色の点は未実施で答えがまだ分からないデータです。言い換えると，前２者は既に栽培試験がされているデータで，後者はまだなされていないデータに相当します。従って，栽培試験で新しい答えを追加しながら新しいルールに更新していくわけです。

図11（**右上**）は，**図11**（**左上**）のまだ答えの量が不十分な状況でルールを推定した場合です。ここでは，ルールを二値化しておらず，中間にはグレーゾーンを設けています。この状態（学習データ数40）のデータと答えから得られたルールを用いて，まだ答えの分からないデータの答えを予測すると，グレーゾーン内ではある程度の誤答（×）が出て，正解の確率（正解率）は66％にとどまっています。

そこで，さらにデータと答えの数を増やしてルール（すなわちモデル）を改良する必要が出てきます。つまり，これまで２年間栽培試験をしてきたものに，さらに３年目の栽培試験を行う必要が出てきます。その時に，どこを選んで試験をするかを考える必要が出てきます。ラベリングの話で例えると，どのデータに正答のラベルを付けるかが重要になります。

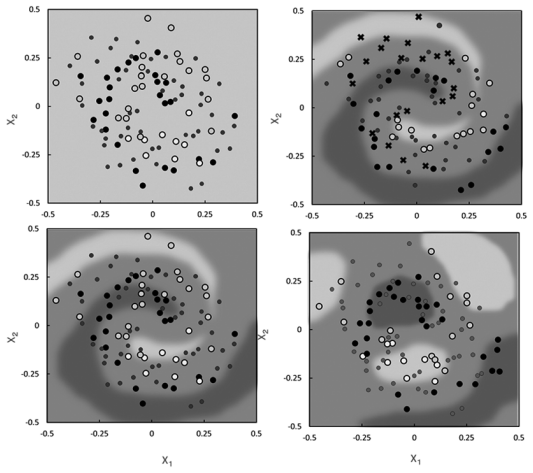

図11　機械学習における訓練データのラベリングによる「ルール」改善方法

（左上：一部のデータをラベリング，右上：左の訓練データでの正解率=66％，左下：グレーゾーンを新たにラベリングして追加したルールで正解率78％に向上，右下：無作為にラベリングを追加すると正解率が低下。
ラベリングにより○と●と判定，×：誤判定，・：未ラベリング，地の色の濃淡は推定された判別ルール）

機械にアクティブ・ラーニング

新たに栽培試験を行うべき場所は，当然，間違え易い，あまり自信がないグレーゾーンを確認する方がいいでしょう。つまり，40個あった学習データに，グレーゾーンのデータ10個を加えて学習させる（専門家がそういう目で見てラベリングする）と，**図11**（**左下**）の例では正解率が78％まで上がっています。このとき，新たな10個を無作為に選んでしまうと，あまり正解率が上がらない場合（**図11右下**）も出てきます。

つまり，機械に１回学習をさせ，機械の正解率が不十分と判断した場合は，その時点で不明確な部分（グレーゾーン）だけを専門家がラベルする。そして，その専門家がラベルしたものを使って機械は利口になっていくということを繰り返します。こういう学習のさせ方を「アクティブ・ラーニング（能動学

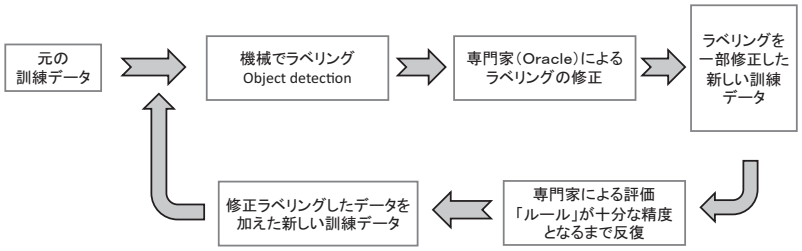

図12　専門家による“手助け”を効率的に利用する機械のアクティブ・ラーニング

［Ghosal et al.（2019）の図を改変］

習）[注5)]」と言います。

さて，機械の能動学習では，機械が自発的にあるデータの答えが知りたい場合にだけ人間が正答のラベリング，すなわちアノテーションをつけていきます。こうするのは，人間がアノテーションするのは実はコストが高いということでもあります。

図12は前述の郭威氏のグループの論文（Ghosal et al., 2019）からの引用ですが，①機械（object director）に取りあえずラベル付けをさせる，②ラベルを付けたものの中から人間が「これ間違っているよ」と修正する，③修正された結果を機械がもう１回学習するというループを回ります。何か出来のあまり良くない子どもの面倒を親が見ているような感じです。つまり，「まずちょっとやってみなさい」と言って，正答だったところはもう教えなくてもいい訳で，間違えたところを丁寧に教えてあげるというイメージです。深層学習はまだ意外と手がかかります。なお，「人が機械に丁寧に教えてあげる」という部分をこの分野では「オラクル（正解を教えてる者，神託者）と称する人間がバウンディングボックスを付ける」と表現します。

注５）この用語は，教員からの押し付けがましい授業ではなく，生徒自身に考えてもらう授業のことにも使うが，今回の例とは関係がない。

ラーニングをさらに効率化

まず，コムギの穂がたくさん写っている写真を機械が見て「穂かどうか自信がない」としたものをバウンディングボックスにより四角で囲んで指定したいのですが，これも結構手間がかかります。そこで，人間が機械に丁寧には教えないで，穂があるところをクリックする程度の結構いい加減な教え方をします。そうすると，別の「機械学習する機械」が「こういうふうにバウンディングボックスを作るつもりだったのでは？」と出力してくるので，それを人間が「正解である」と教えるか，あるいは，ちょっと間違っていれば，バウンディングボックスの位置を修正するという感じの処理をします。つまり，人間がまず完全には正解とはいえない形で機械に教え，それを機械が類推してちゃんとした正解を出し，それでも間違っていれば，それを人間が修正して答え合わせを行っていくことを繰り返していきます。

先ほどの例でいうと，子どもが自信なさそうにしていると少しだけヒントを与える，しかし，ヒントを基にしても間違えるようなときにはきちんと教え直すというような感じの方法もあります。こういう方法をうまく作っていかないとラベル付けも大変です。

つまり，ラベリングというのは極めて大変な作業で，深層学習の華やかな世界の向こう側には地道にラベリングしている世界があるのです。

AI は外挿が苦手

AI がラベリング以外にもう一つ苦手にしているものとして，外挿[注6)]というものがあります。

例えば，カンキツのゲノム情報から果実の大きさを予測する研究について，訓練データとして，カンキツ類のゲノム情報を「データ」，果実の大きさが「答え」として，それから答えを予測する「ルール」の推定を行っていた際，訓練データの中にブンタンやハッサクのように極端に大きいカンキツが含まれていなかった場合にも本当に大きいものを予測できるのかという疑問が生じました。これは，学術用語ではまさに外挿と言います。しかし，育種というのはこれまでに無かった高収量の品種を作っていくのが仕事ですから，基本的に外挿を行っていくことが必要です。したがって，外挿が不得意では困ります。

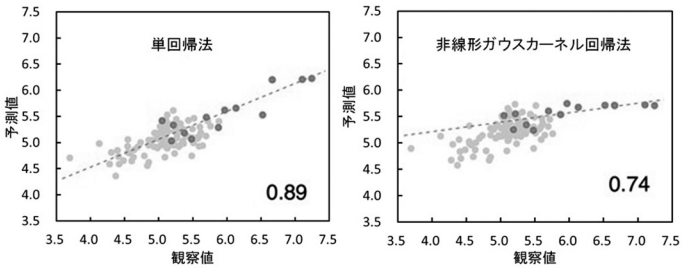

図13　非常に果実の大きい品種を除いた訓練データから異なる回帰法で導出した「ルール」によるゲノミック選抜の予測値

〔予測値：●，数値は相関係数，両軸とも対数変換値。
（Minamikawa et al.,（2017）の図を改変）〕

そこで，私たちのグループの南川舞氏が2017年に発表した論文の中で，２つの手法でカンキツの果実の大きさの機械学習を行って外挿の能力を見てみました。つまり，大きな樹種の答えは伏せておいて，あまり大きくないカンキツのデータばかりで訓練したルールを使って答え合わせをしました。

そうすると，**図４**の渦巻き図形も解いてきた非線形ガウスカーネル回帰法という優秀な機械学習法でも，元のデータに含まれた果実サイズを大きく超える予測はできていません（**図13**）。一方，単純な単回帰のモデルでは外挿も比較的得意にしていて，けた外れに大きいカンキツ類もそれなりに大きいものだと予測できました。

機械学習（深層学習）でも外挿を苦手にしない手法もあるのですが，多くの方法が外挿を苦手にするのは基本的にデータが与えられたとき，それに対応する答えをいかに効率的に導くかというルールになっているからです。そのため，非常に性能の良い非線形ガウスカーネル回帰法であっても，「今まで見てきた中で一番大きなもの」までは予測できてもそれを超えたものの予測はなかなかできないわけです。

外挿能力の克服法

外挿能力の欠点を克服する方法もあります。農業関係からは少し離れますが，地下資源探査で使用する方法を例として出します。**図14**で地の色の濃いところは有用鉱物や石油の埋蔵量の多いところ，薄いところは掘っても何もないところですが，現時点ではだれも

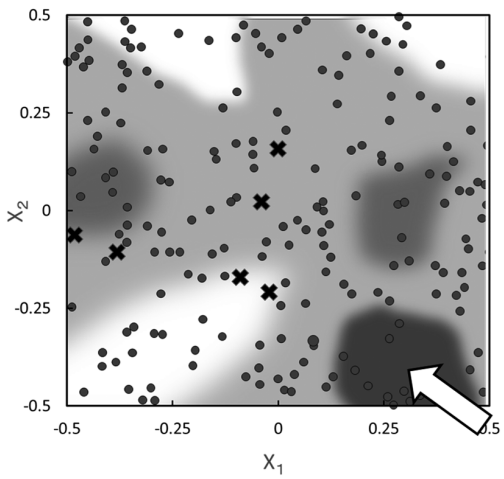

図14　鉱物埋蔵量が最も多いところを探査する方法

〔●：掘削予定地，×：試掘地点，地の色：濃（多）～淡（少），矢印：埋蔵量が最も多い地点〕

注６）既知の数値データを元にそのデータの範囲外の数値を求めること）

このルールは分かっていません。矢印で示した所が一番鉱物量が多いところなので，ここを見つけたいわけです。

試験掘削予定の全部地点（●）を一気に調査できないので，最初は６つの×の地点の調査だけを行ったデータだけがあります。そこから何とかして多量埋蔵地点にたどり着きたいわけです。これは，まさに外挿の世界です。既に得ているデータを元に，比較的高い含量だった調査地点の周りを探査するだけでは，本当の高含量埋蔵地点まではたどり着けません。これが外挿の難しさです。

ベイズ最適化法で外挿に対処

そこで，ベイズ最適化という方法を使ってみました。これも先に導き出しておいたルールを出発点にはするのですが，使い方が違います。つまり，ルールを基にして予測された予測値そのものではなくて，ある程度不確かさを伴った予測します。

通常だと訓練データによりルールを導き出して，そこに新しいデータを入力しますが，そのデータが訓練データを用意する際に用いたデータとよく似た値だと，新しいデータに対応した答えはその訓練データで得た答えの近傍内でしか得られません。従来の方法では仮にその答えの値が65だとすると，前後に例えば５の比較的狭い幅の不確かさをつけて65±5の範囲を探査すべきだと予測していました。

しかし，訓練データに使用したデータの範囲をはずれたデータを新しいデータとして入力する場合には，不確かさが大きくなるものと考え，ルールから導かれた答えがたとえ50と低くても50±25として不確かさの幅を広げてやります。推定する値の上下の幅が標準偏差だとみなすと**図15**のようになります。そうすると，訓練データで用いた範囲内で導出したルールにより推定される答えが現状品種の

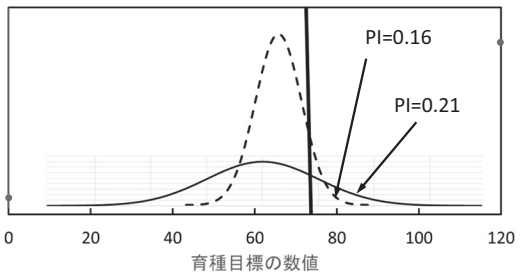

図15　予測値の不確実性も考慮に入れた目標の効率的な探索

〔縦棒：現状の最良の特性値；予測値の不確実性も考慮に入れると，不確実性が大きい場合のほうが（実線）が小さい場合（破線）よりも改良される確率（PI：現状の最良値より良い値をもつ確率）が高くなる場合がある。〕

特定の数値である70点を超える確率は，従来の推定方法では16％ですが，新しい推定法では21％となり，後者の方が不確かであるがゆえに優先的に探索すべきだという答えになります。

この話を農業の現場で応用すべきなのは，いろいろな条件で栽培しながら最適な栽培方法を見つけていくのがかなり難しいからです。栽培試験そのものが年１回しかできないかもしれないし，１個のデータを得るのに非常に時間を要するような場合もあります。そういう場面ではここで紹介した探索方法は非常に有効です。

マダガスカルに適するイネ品種の探索

地球規模課題対応国際科学技術協力プログラム（SATREPS）[注7)]で国際農研センター（JIRCAS）の M. Wissuwa 博士らと行っている研究では，マダガスカルに適したイネ系統の選定を行いました。この際にベイズ最適化法を適用しました。

アフリカ南東部のマダガスカルではイネが主要な農作物で，米の一人当たり消費量は日

注７）科学技術振興機構（JST），日本医療研究開発機構（AMED），国際協力機構（JICA）が共同実施している開発途上国の研究者が共同で行う研究プログラム

本人の倍です。しかし単収は極めて低く，日本の2/3ぐらいです。これはリン欠乏で鉄過剰である上，農家は貧しく肥料も農薬も撒けません。したがって，このような条件下でも多収となる品種が欲しいというニーズがあります。一方，フィリピンにある国際イネ研究所（IRRI）では，イネ3,000系統の遺伝資源の全DNA配列を既に解読しています。

そこで，私たちは，それらの系統の一部をマダガスカルで栽培試験し，データとして各系統のDNA配列を，マダガスカルでの栽培試験の結果を答えとして，両者を突き合わせルールをまず導きました。それに基づいて，IRRIの保有する3,000系統のうちのどれを次にマダガスカルで試験栽培すべきか判断しようというのがプロジェクトの計画です。

もちろん，3,000系統全部をマダガスカルで栽培試験をすればいいのですが，それらを一気に栽培試験することは困難です。通常の試験規模が1年100系統だとすると，3,000系統なら30年もかかってしまいます。そこで，もう既に読んであるDNAの配列をうまく使うことを考えました。しかし，この設計でもやはり外挿の問題が出てきます。例えば，最初に選んできた100系統の範囲内で良いものを予測できるというのでは全く役に立たちません。その範囲を超えたものを見つけたい訳です。

このように，不確かさを考慮しながら探索していかないとなかなか目的に行き着けません。ですから，全く同じルールに基づいてはいるのですが，ひと工夫すると外挿となる場合でも有効な探索もできるようになり，遺伝資源も早く良いものが得られるようになりました。

5．農業現場へのAI適用の将来展望

AIの「眼」としての利用

農業とAIを考える上で非常に重要になるのが，「眼」の存在です。農業の規模を単に拡大すると，それがそのまま収益の増加につながらないのはよくある話です。平成23年度の農業経営統計調査では，規模を拡大しても経営耕地面積が10ha当たりを越えると面積当たりあるいは固定資産当たりの所得は下がっています。これは規模を拡大すると栽培管理等に目が届かなくなり，管理の質が低下してしまうからです。このため，農地全体に目が届くようにする必要があります。

5年前に農林水産省の「革新的技術創造促進事業（異分野融合共同研究）」プロジェクトで担当した課題では，ドローンを使って多数の圃場の見回りを行い，得たデータを解析用のクラウドメモリー上に蓄積し，それを解析して現場にフィードバックする方法を検討しました。

当時からドローンの性能自体はそんなに変わっていませんが，取れてくるデータは**図16**（**口絵参照**）のようなもので，ドローンを飛ばして広い範囲の水田を撮影します。元の画像は斜めからなので，真上から見ているように修正し，それを3次元処理，表面温度測定，植物の生育状況の測定[注8)]等の解析をするとかなり高次な情報が取れます。

例えば，時系列画像があると，イネの倒伏予測も可能です。**図17**で，9月27日のイネ収穫前の画像を地表高モデル（DSM, Digital Surface Model）で3次元処理した画像では倒伏と推定された部分があり，実際の写真でも確認されました。そこで，7月28日のDSMデータをみると，倒伏と推定された部分はその時点でやや盛り上がっており，過繁茂だった可能性があり（背の高い雑草が生えていた可能性もある），そういう状態が2カ

注8）各地点の波長から葉緑素含量を推定するNDVI（Normalized Difference Vegetation Index，正規化植生指数）やさらに改良したENDVIがよく利用される。

図16　ドローンで撮影した圃場の画像加工例

〔左：画像190枚からの３次元図，中：左の図を真上から見たものに加工，右：推定された標高（疑似カラーで表されている）〕

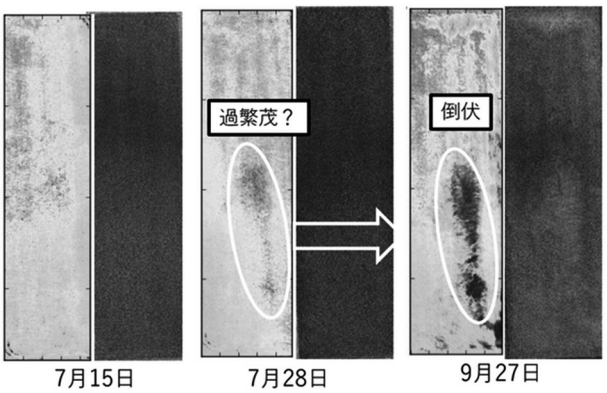

図17　ドローンで時系列に撮影した圃場の画像からの倒伏予測

〔各撮影日の図の左側：地標高図（橙色ほど高い），右側：幾何補正画像（真の色に近い）〕

月後の倒伏につながった可能性があるわけです。

これを活用して，２カ月前の画像から機械学習でここは倒れそうだと予測を行って，台風が来て被害が激しくなる前に刈り取ってしまおうという意思決定も出来そうです。実際，そういうモデルを作ったところ，７月のデータで倒伏を予測できることが分かりました。

つまり，ドローンの「眼」で水稲圃場の植生表面の高さを経時的に推定していたから，こういうモデリングができるようになったのです。今後，こういうデータを蓄積すれば，もっと違う発見もできて，いろいろなルールが抽出できると思います。

AIの「眼」に対する期待

ドローンの「眼」に対する期待が非常に高まっています。例えば，昨年イスラエルで行われた農業のベンチャーが集まる会議に出てきました。住友商事はイスラエルのTARANIS社というリモートセンシングでフィールドモニタリングをするサービスを提供する会社に対して大きな投資をしようとしています。また，ドイツのバイエルクロップサイエンス社も提携してアルゼンチンで事業展開するようです。

イスラエルのSeeTree社はリモートセンシングのサービスを提供するスタートアップ企業で，私も実際に会社を訪問してCEOと話をしました。その会社はターゲットをカンキツに絞って大きな資金調達がされているようでした。ドローンを使った「眼」は極めて重要な農業上の手段になると思います。

ドローンの「眼」利用の懸案

ただ，「眼」があればいいというものではありません。人間は，眼の背後には脳があるので，即座に情報処理して「これは危ないから逃げろ！」という判断ができます。しかし，現在のドローンには「眼」しか付いていないので，観測した後の情報処理に時間がかかってしまいます。ドローンで圃場を測定中に病害や生理障害が出ていて，すぐにでも何かした方がいいという場合でも適切なアクションを起こせない場合があります。

何がそれを阻んでいるかというと，データが巨大なので生産者圃場からのデータ収集が大変だからです。今は通信規格が5Gになってきつつあるので状況が変わってくるのかもしれませんが，それでも大変です。実際にわれわれが2015年頃にプロジェクトを行っていた時には，解析用のクラウドメモリーの部分がネックになり，当初の最終目標まで行き着

けませんでした。

先日，あるサイトに，「農業に5G，それ要る？　スマート農業『ぼったくり投資案件』３選」とありました。笑い事ではありません。「ドローンは開発途上　空撮データ解析が遅い例も」とあって，夏撮影したデータの解析に時間がかかってその年の稲作には解析結果が使えなかった例を紹介し，だから本当にデータを有効利用するには即時性も必要と言うことでした。

私たちが前述のプロジェクトを行った時のボトルネックもそこで，解析に時間がかかることと取得データの送信方法が懸案でした。現地試験は福島県須賀川市の（有）西部農場で行っていたのですが，社長さんから「インターネットでやるより郵便の方が速いのでは？」と言われてしまい，「そうですね」とメモリを郵送で送ろうという話まで出たぐらいです。

この問題に対する解決方法の一つにとして，ドローンそのものに「脳みそ」を積むことも考えています。幸い，小型の深層学習装置があるので，それをドローンに搭載するとドローンが飛びながら情報解析を行い，着陸後に重要な情報だけをクラウドメモリーに転送する，あるいは本当に即時性を必要とする場合には飛んでいる状態でドローンから生産者に何か「注意喚起」を出して，即時のアクションができるようにするのが良いと思います。このように「眼」があればいいというわけではなくて，「眼」を支えるシステムを作ることが極めて重要です。

農業におけるトランスオミクスの利用

農業のデータを取っていく場合，相手にしているのは植物や動物です。植物や動物は生き物なので，外側の見た目だけを測っているだけでは分からないことがたくさんあります。そういう場合には，現在，「○○オミクス」と呼ばれる網羅的情報がたくさんあります。オミクス[注9)]と言われているものは大抵の場合，計測効率が非常にいいものなので，きちんと測っておくことが重要です。そうしておくと，例えば作物の根圏に生息している微生物がどうなっていると作物がどのように育つかというルールも見えてきます。

そのようになれば，例えば栽培管理を最適化するために，単に散水するときに作物がどれだけ生長したのかを見るだけではなくて，そのときの植物の内部状態，あるいはその植物の根を取り囲んでいる根圏微生物がどう変わったのかという情報を基にしてモデルを作ることができるでしょう。そうすれば，もっと作物側に寄り添った精度の良いモデルになるのではないかと考えています。

恐らく，ビッグデータというのは単に圃場でドローンを飛ばし，あるいはトラクターにカメラを付けて観測できるデータだけではなく，作物内部の生理的な情報も重要になると思います。今後，それらを解析しモデリングをすることは非常に重要になると思います。

農業と AI 教育

有名な言葉ですが，Harvard Business Review 誌に書かれた「データサイエンティストというのは，21世紀のもっともセクシーな職業だ」が一時期流行しました。私は，**図18**にも示しましたが，農業関係の技術者の中で，データサイエンスをやる人（アグリバイオ・データサイエンティストとしました）がもっと増えていくべきだろうと考えています。最初に食料問題のお話をしましたが，さらに，新型コロナウイルス感染症パンデミックもあり，国内でももっと食料を作れるようにしておく必要があるという議論も出てくる

注９）生物個体の生体内に存在する遺伝子やタンパク質等の網羅的情報のこと

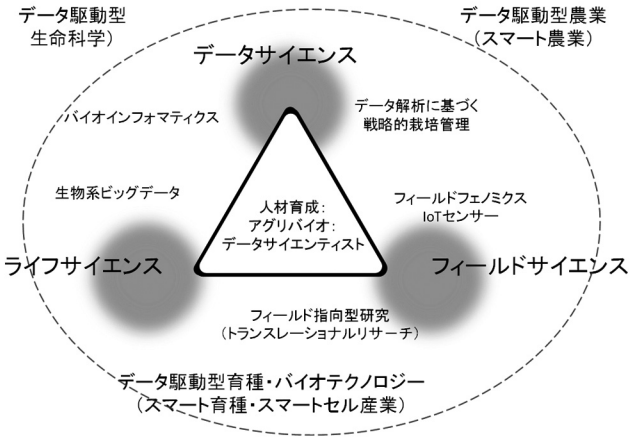

図18　今後人材育成をしたい農業関係の技術者の専門領域

（東京大学農学・生命科学科　大森良弘氏原図を改変）

と思います。そうなると，いかに効率的に農作物を作るかが重要で，そこではデータサイエンスが重要になります。

ただ，こういう人材は私の大学でもなかなかうまく育てられていません。なぜ難しいのかというと，**図18**に示した３分野を全部兼ね備えた人はなかなかいないのです。データサイエンス，ライフサイエンス，フィールドサイエンスはそれぞれ融合すべきですが，教員の場合を見ても大体はどこか一つの分野に所属しています。しかし，私はこの３分野をある程度理解できる学生を育てたいと思っています。

農業のためにAIを生かすということを考えるに当たり，最後の最後は若い有能な学生をどう育て，農業とAIの両方を理解できる人材をどう育成していくかというところに行き着きます。大学のなさなければならない役割は極めて大きいと考えています。

６．最後に

今は「AIの夏」と言われています。夏とわざわざ言われているのは，「AIの冬」が来るかもしれないということです。でも，そうならないようにするには，やはりそれを使う人たちがよく理解して使うことが重要です。

現状では，AIに使えるビッグデータがまだないので，それを取得するシステムをどう作るか，あるいは最近はドローンが「眼」になると言われていますが，その「眼」の背後のシステムもきちんと作る必要があります。さらに，何と言ってもデータ科学だと思うので，その手法をきちんと使える人材を育てることも重要です。

私は，AI研究について科学技術研究プロジェクトのあり方についても言いたいことがあります。短期的，近視眼的なプロジェクトというのは早期に何かを解決するには重要なのですが，これをきちんと成し遂げようと思うと，教育という視点も含めて長期的で遠視眼的な計画が大事で，そうしないと，本当にAIの冬が来てしまうかもしれないと危惧しています。

今日お話しした研究成果の多くは科学技術振興機構（JST），農林水産省からの予算によるものです。ただし，農林水産省の予算については，最近どうしても早期解決型のプロジェクトが多くなっており，なぜか初年度以降自動的に予算が削減されていくようなプロジェクトもあります。少し腰を落ち着けられるような骨太で遠視眼的な計画を立ててもらって，少し夢に取り組むようなプロジェクトを動かしてもらいたいと思ったりすることもあります。

質疑応答

良質の訓練データを現場でどう取得

「機械学習」という武器が加わっても，そのための「訓練データ」を作るにはやはり栽培試験データ等の量と質が決定的に重要とのお

話でした。育種の場合，それを行うのは専門家ですので，そのような点には比較的対処しやすいと思います。しかし，現在，農業現場で想定されているスマート農業では，農家の日常において機械作業等の営農を行いながら作物の生育状況や土壌条件のデータを収集し，それに基づいて機械が管理するということを想定していると思います。この場合，良質の「訓練データ」を大量に収集するには具体的にどのような点に注意することが必要でしょうか。

岩田：営農を行いながら機械学習の訓練データを収集するためには，計測データ間の紐付けを行い，それらの間の関連がわかるようにすることかと思います。

様々な要因を互いに紐付けるためには，例えば，田畑を多数のマス目（グリッド）に分割し，それぞれの分割において多面的にデータを収集します。これにより多面的なデータをマス目単位で紐付けでき，例えば施肥や耕うんのレベルが成長や生育状況に及ぼす影響，また，収量や品質に及ぼす影響などを機械に学習させることができます。ただし，収穫や品質など一枚の田畑を多数のマス目に分けてデータをとるのが難しい場合も少なくないと思います。その場合は，マス目単位でとられたデータと一枚の田畑単位でとられたデータの階層的な紐付けを解析のときに利用できるようにしておくことが重要かと思います。

いずれにしても，データ間の紐付けができるようにデータ収集をしておくことが互いの関連を解析するには必須となります。したがって，収集されるデータが互いにどのように紐付けができるのかを意識しながら，データ収集システムを設計することが重要だと思います。

交配親を超越した特性の育種に有効か

ナシの育種の事例では，モデルで予想された予測よりも現実の集団を用いた方が交配親の特性を上回っているように見えます（**図9**）。このように「美味しいものを食べるには挑戦が必要」といわれるように，かえって大きく異なる形質を有する親同士の組合せの方が予想を上回る特異な結果を得られる可能性が高くならないでしょうか。

岩田：予測値よりも観察値のほうが変異の幅が大きくなるのは，環境誤差による影響もあります。予測値のばらつきは遺伝分散によるものですが，観察値のばらつきでは遺伝分散に環境誤差分散が加わります。これにより，観察値は予測値よりも大きなばらつきを示します。

異なる特性をもつ親同士の組合せのほうが良いかどうかは，場合によります。もし，異なる特性をもつ組合せがいつも望ましいのであれば，例えば在来種と近代品種を交雑するのが良いかもしれません。しかし，実際には様々な形質が分離してしまい，近代品種が創り上げてきた遺伝子のセットが失われてしまう可能性も高くなります。こうした可能性についても事前に評価しておくという意味でも，モデルに基づくシミュレーションを行うことは重要です。

なお，果樹の場合は厳密には親同士の違いではなく，親の親間の違い，つまり子供の世代からみると，祖父母間の違いが子供の世代で分離します。

果樹育種への見通しについて

主要果樹のカンキツ，リンゴ，ブドウ，モモ，カキについては，ナシと同様，既にDNA多型のデータが蓄積されているようですが，実際の育種にAIが利用可能な状況になっているのでしょうか。

岩田：「DNA多型のデータ」が，どのような

データなのかに依存します。ゲノムをもとにした選抜を行うには，多数の品種・系統のDNA多型のデータが必要となります。そうしたデータがなければ，DNA多型と改良対象形質の表現型の間の関係を機械に学習させることができません。例えば，代表的な数品種についてDNA多型データがあるだけでは，DNA多型と表現型間の関係の学習は難しいです。最も望ましい状況は，育種を進めながら，育種集団のDNA多型データが日常的に収集され毎年データが蓄積していく状況です。こうした状況を作り出すことができれば，ゲノムとAIを用いた育種が汎用的に利用可能となります。また，こうした状況を作り出すことができれば，AIの能力は一般にデータが蓄積すればするほど向上するため，育種の効率も年々向上していくと期待されます。

機械学習の欠点・注意点

機械学習は膨大なデータからの壮大な帰納法，あるいは回帰分析のように理解しました。その問題点として「外挿」が不得意とのことでした。同様に，機械学習では因果関係の推論ができていないので，それが問題になる場合はあるのではないでしょうか。

岩田：おっしゃるように，回帰と同様に，機械学習も因果関係をモデル化しているわけではありません。言い換えると，予測精度が高いモデルができたとしても，入力が原因で出力が結果という関係があるとは限りません。その点については，モデルを利用するときに十分注意する必要があります。例えば，Aという原因がBとCに独立して影響を及ぼしているとき，Bを変化させてもCにそれに対応した変化が起こるとは限りません。一方，Aの変化はBとCに変化をもたらすので，Bの変化から間接的にCの変化を予測することは可能です。このように，優れた予測モデルが正しい因果関係を表すとは限らないので，使っているモデルの性質を念頭において利用する必要があるかと思います。

なお，予測のためのモデルと解釈のためのモデルは，多くの場合異なります。私の研究分野では，例えばゲノムに基づく選抜には前者のモデル，ゲノムに基づく遺伝子検出には後者のモデルが用いられます。いずれもDNA多型と表現型変異間の関連をモデル化したものですが，モデルを用いる目的が異なっており，そのために適したモデルが異なります。前者は予測精度の高さが最も重要であり，したがって機械学習，量的遺伝学の分野のさまざまなモデル化手法を用いて，予測精度の高いモデルが構築されます。後者は検出精度の高さが重要なので，例えば，多重検定による偽陽性を抑制するような注意深いモデルが用いられます。