機械学習と宇宙論

東京大学理学系研究科 吉田直紀・森脇可奈

アブストラクト
大型望遠鏡を用いる宇宙観測からはペタバイト級の画像データが生み出され，天文学者は巨大データを分析してはじめて宇宙や天体に関する情報を得ることができる．機械学習やAIを用いたデータ分析技術の開発は世界中ですすめられており，手法とスピードの差が観測プロジェクトの成否の鍵となることも多い．宇宙論や天文学分野での機械学習の応用と実観測データへのアプリケーションについて紹介する．

キーワード
深宇宙観測，機械学習，画像解析，シミュレーション

1. はじめに：データサイエンスとしての宇宙論

近年の地上大型望遠鏡や宇宙望遠鏡を用いた観測により，私たちの宇宙の成り立ちや天体の形成進化について多くのことが明らかになってきた．地球から 100 億光年も離れた場所にある星々の光をとらえ，太陽系外の惑星の姿を直接見ることさえできる。高性能の観測装置を用いて精彩な天体の画像を取得できるようになり，この 10 年ほどの間に天文学における観測データの量は爆発的に増加してきた．近い将来には一つの望遠鏡から生み出されるデータがエクサバイト級になると予想されている．例えば、我が国が米国ハワイ島に設置したすばる望遠鏡の巨大カメラ Hyper-Suprime Cam（HSC）の焦点面には 104 個の CCD チップが搭載され，1 つのスナップショットで 10 億ピクセルの画像が生成される．一晩の観測で，数百ギガバイトのデータが生みだされ，天文学者はその日のうちに相応の量のデータを分析することになる．また，10 カ国以上が参加する国際共同電波観測プロジェクト Square Kilometer Array による観測データ生成速度は 1 秒あたり 1 テラバイト近くに達し，6 時間の観測で数ペタバイトのデータが生成される．全データを物理的に保存し続ける実用的な技術は未だ存在せず，必要なデータのみを保存することが合理的だろう．しかし，保存するべきデータをどのようにして知り，どのように選択すればよいのかは現時点で誰にも分からない．こうした課題に対して，効率的で信頼性の高い機械学習または AI ベースの手法に対する期待は大きい．

膨大なデータを処理することさえできれば，宇宙観測によってさまざまな科学的成果が得られる．遠くの宇宙で発生する超新星爆発の様子や，全く新しいタイプの天体の発見，多数の銀河の位置を同定して行う宇宙の地図づくりなど，期待される成果は多岐にわたり，数千人の研究者が巨大データ分析に従事することになるだろう．本稿では，天文学や

宇宙論の研究で使われている機械学習や AI について解説する．実観測データへの応用例も紹介するが，個々の手法やアルゴリズムについての詳細は割愛し，どのようなプロセスを経て科学成果が得られるのかに焦点を絞る．以下では，機械学習や深層学習，生成系 AI などを含めた広い意味でのデータ分析手法を ML/AI 手法と総称する．

2. 機械学習の宇宙観測への応用

天文学や宇宙論では撮像画像から分光データ，天体の明るさの時系列データなどさまざまなタイプのデータを取り扱う．本節でははじめに天体画像の分析に関するアプリケーションを紹介しよう．

2.1 超新星の自動検出と分類

超新星とは質量の大きな星がその一生の最期に起こす巨大な爆発現象のことである．寿命をむかえた星が急に明るく輝き，可視光以外にも X 線やガンマ線などにより観測することもできる．肉眼で見られる超新星はめずらしく，何十年に一回程度の頻度になってしまうが，専門の大型望遠鏡を用いて超新星探しを行うと，最近では 1 年間に数千個以上も発見される．米国が南米チリに建設中のベラ-ルービン望遠鏡が稼働し始めると，年間数万個もの超新星が検出されるようになるだろう．そのような超新星探査の宇宙科学としての主目的は，Ia 型超新星とよばれる特別なタイプを発見，同定し，その天体までの距離を測定することで宇宙膨張の歴史を明らかにすることである．1990 年代後半には数えるほどの Ia 型超新星を用いて距離測定が行われ，現在の宇宙の加速度的膨張の発見につながった（2011 年のノーベル物理学賞）．

超新星は数多くの銀河の画像の中の局所的な明るさの変動として現れるため，大量の画像の差分をとることで検出できる(図 1)．しかし一晩に数百個もの検出となれば大掛かりな画像処理システムが必要となる．異なる日（時）に観測して得られた画像データの差分をとっても，後に残ったもののすべてが天体に対応するわけではない．天候や大気の状態も含めて，観測条件は刻一刻と変わるものであり，単純な画像差分では考慮しきれなかった明るい部分のわずかな差や，宇宙線が CCD にヒットすることによるエラー，さらには人工衛星や小惑星のような移動天体も差分画像には残っている．もちろん後者は移動天体を探す研究者にとっては貴重な検出例となるのだが，超新星を探索する我々にとっては，それらの例は「非検出」として扱わなくてはならない．近年の ML/AI にとっては画像分類は得意分野でもあり，十分な数の例を用いて学習すれば，（差分）画像の中の超新星を迅速に見つけることができる．我々の研究チームはすばる HSC を用いた超新星サーベイを行い，ML/AI によってさまざまなタイプの超新星を発見した．2016 年 11 月から 2017 年 4 月までの半年の間に，六分儀座にある COSMOS 領域と呼ばれる部分を 52 回繰り返し観測した結果，7 万を超える変動天体候補を検出し，その中から 2 000 個ほどの超新星候補を発見した．一晩の観測で 50 個以上の新たな超新星を発見できた計算になる．特に

地球からの距離が 70 億光年以上も離れたところにある，重要な超新星を 100 個以上も検出しており，この数はこれまでのサーベイ観測による蓄積をはるかに超えるものである．

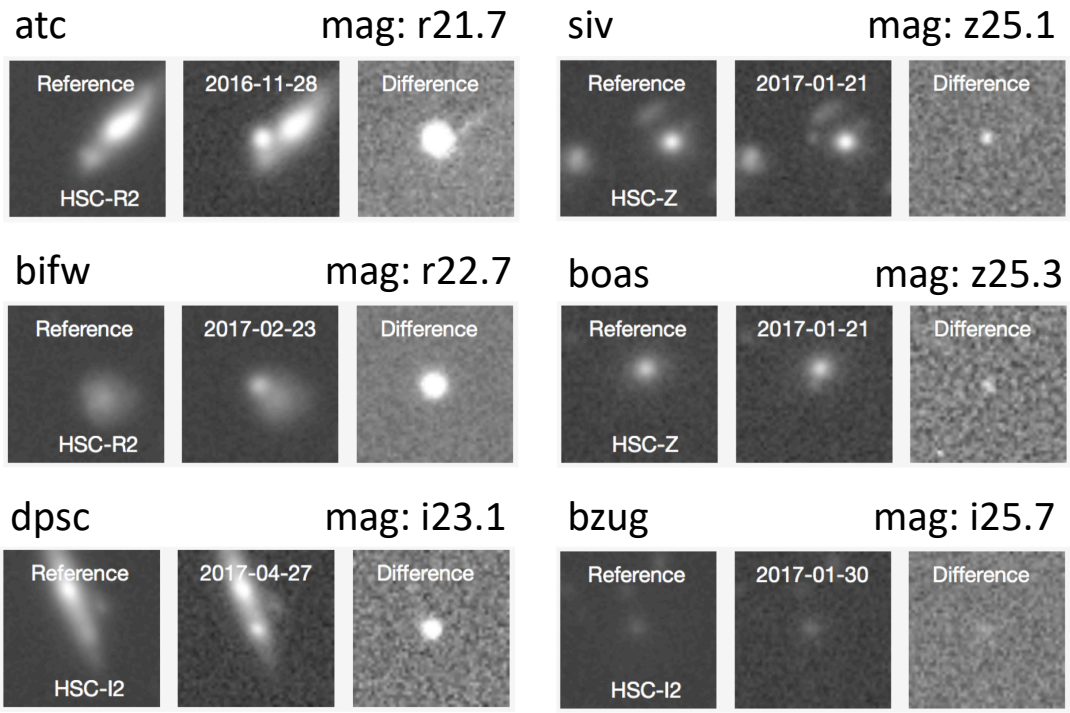

図 1 すばる HSC サーベイで発見された超新星の例．それぞれ左から参照用画像，観測された画像，これらの差分画像になっており，左上に超新星のアルファベット名，右上には等級(明るさ）が示されている．等級の値が大きいほど暗い天体である．

集光力で勝る大望遠鏡を用いる利点は遠くの暗い超新星まで検出できることである．図 1 では 26 等級，すなわち肉眼でみえる星の 1 億分の 1 の明るさの天体（の変化）まで検出している．

超新星はその爆発メカニズムやもとの星のタイプによって様々な型があり，それぞれ明るさの変化パターンや変化の時間に特徴がみられる．本来は型の分類は分光スペクトルを用いて行われるが，光を波長ごとに分けて測定する分光観測は一般に時間がかかる．そこで私たちの研究グループは，明るさの変動と色の情報だけから超新星の型を自動判定する分類器を開発した．

図 2 に，実際にすばる望遠鏡による観測データ分析に用いられた，畳み込みニューラルネットワークの構造を示す．観測画像データ処理は図の左側から順に行われる．はじめに超新星を検出し，明るさの変動を求めるため，差分解析を行う．異なるフィルターを通した多色の観測画像(図 2 では Ｙｚｉｒｇ の 5 つの観測バンド)を用いることで超新星の特徴を正確に捉えることができる．数万個もの画像例を用いて十分なトレーニングを行なった後は，ネットワークに観測画像を与えると即座に超新星を分類できるようになる．この判

別器が優れている点は，ある観測夜での超新星の画像が各フィルターに対し 1 枚でもあれば，90 パーセント以上の精度で Ia 型を抽出できることである．

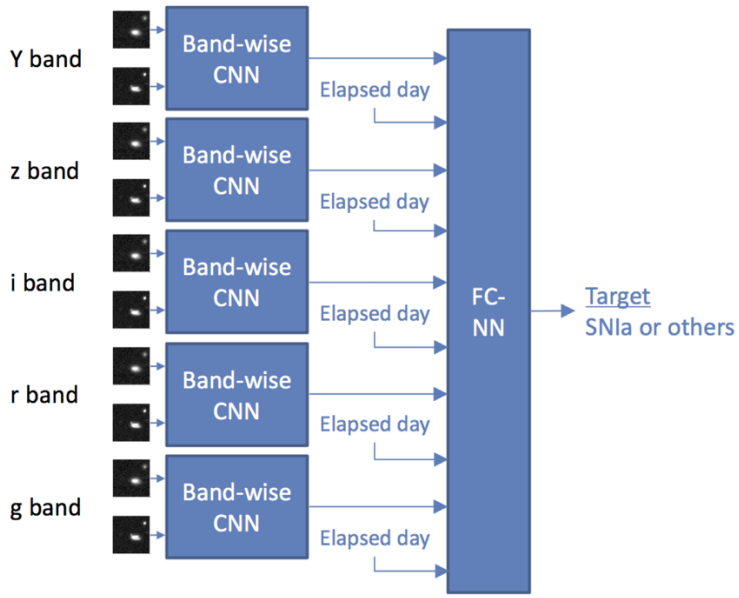

図 2 超新星分類に用いられた学習器の構造．左側から入力する 5 色の画像に対して画像の畳み込みなどの操作を順に行なって明るさの変動や特徴を検出し，一番右側の全結合層で超新星タイプを判定、出力するようにトレーニングされる．

超新星の探索を含む，天体の明るさや現象の時間変化に着目する研究は時間領域天文学とよばれ、新たな分野として近年急速に発展しており，日本でも革新的な観測が行われている．Tomo-e Gozen (以下 Tomo-e)は 84 個の CMOS センサーで構成され，木曽天文台のシュミット望遠鏡に搭載して毎秒 2 フレームという頻度で夜空の動画を撮ることができる．Tomo-e は秒単位あるいは分単位という短い時間で変動する天体を捉えることで最短時間領域を開拓できると期待されている．人間が普通に行う観測では，同じ箇所を毎晩探索するというように変動検出の最小単位は 1 日となってしまう．したがって歴史的にも明るさを数日から数十日で変える変光星や超新星しか知られていなかった．しかし最新の望遠鏡による探索により，明るさも変動の時間も随分異なるものが多数見つかるようになった．秒単位で明るさを変える天体など想像もつかないかもしれないが，実際には X 線や γ 線を 1 秒ほどの間に放出して輝く天体（γ 線バースト）が多数見つかっており，また波長の長い電波域では，１ミリ秒程度で明るくなる高速電波バーストも発見されている．それらの特異な天体がどのように明るくなり，どこから高エネルギーの電磁波を放出するのかなどその物理的メカニズムは謎である．極めて短い時間の間に激しく変化する天体は宇宙にはたくさん存在するのだろう．短時間領域にはまだ探索しつくされていない広大な世界

がひろがっているかもしれない．Tomo-e は時間領域天文学の最先端を切り開くと期待されているが，そのデータ分析は前人未到の分野である．Tomo-e は一晩の観測でおよそ 30 テラバイトの動画データを生み出す．Youtube の動画 300 万個分ものデータになる．この中から，ランダムな場所で一瞬だけ光る天体を検出するのは至難の作業である．さいわい、他分野の研究や一般社会で実装されているような ML/AI による異常検知が役立ちそうだ．工場ラインの故障検知からクレジットカードの不正使用まで，異常検知は社会での応用も多い課題である．宇宙観測では，全体に丸く対称的であったり，淡くぼんやりと輝いているなど，画像としては単純なものが膨大に存在する中での微小な変化(異常)を確実に検出しなくてはならないため，他の用途とは大きく異なる側面もある。そのような特殊な目的で開発された ML/AI アプリケーションが他の様々な問題にも適用できることを期待している．

2.2 宇宙の地図づくり

宇宙には数多くの銀河が存在するが，それらは広大な空間の中でばらばらに存在するわけではなく，大きな塊やフィラメント状の特徴的な構造を形作る．従来は，数百万個もの銀河を一つ一つ時間をかけて詳細に分光観測することで，差し渡しが 100 億光年にもおよぶ領域の銀河の地図をつくっていた．通常，完了までに数年から十年ほどかかる作業である．最近になって，銀河の分布をさぐるための強力な手段が提案され，実際にいくつかの観測が開始された．本節では「輝線強度マッピング」とよばれるこの最新の観測とそのデータ分析法について解説しよう．図 4 に，ある銀河の分光スペクトルの一例を示す．特定の波長で強度の大きい輝線がいくつか見られる．これらは銀河内の原子やイオンから放出されるものであり，それぞれ固有の波長が定まっているが，実際の観測では銀河が遠ざかる速度のために本来よりも長い波長の光として観測される．遠ざかっていく救急車の音がドップラー効果で低く聞こえるのと同じように，我々から遠ざかっていく遠方の銀河からの光は波長が伸びるのだ．このため，銀河からの輝線を検出し波長のずれを測ることで，その奥行き方向の位置（距離）を特定し，銀河の三次元分布を得ることができる．

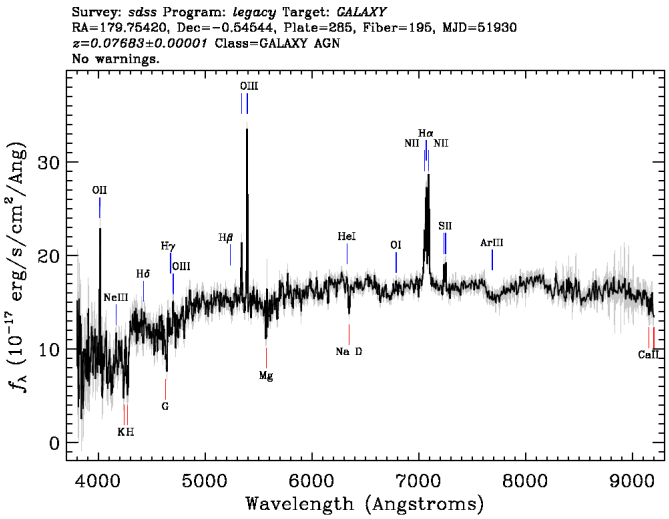

図 3 スローンデジタルスカイサーベイによって得られた銀河の分光スペクトルの一例．水素イオンや酸素イオンなど，いくつかの同定された輝線がラベルづけされている．http://skyserver.sdss.org/dr16/en/tools/chart/navi.aspx より．

最近では個々の銀河の輝線を検出するかわりに，天球面上の広い領域を一括して分光観測し，膨大な数の銀河からの輝線シグナルを一気に観測することが可能になっている．こうした観測は輝線強度マッピング観測と呼ばれ，現在さまざまな波長帯での観測が計画されている．広い領域を効率的に探査できるという大きなメリットがある一方で，種々の天体からの光が同じ波長帯で重なってしまい，視線方向（奥行き）の情報を得るのは難しい．この研究上の課題に対しては畳み込みニューラルネットワークを用いたシグナル抽出の手法が提案されている．畳み込みにより入力画像（観測データ）にフィルターをかけることで画像中の特徴的な構造を選択的に取り出す操作である．このような「シグナル分離器」が適切に動作するためには，膨大な事前学習を行わなくてはならない．私たちの研究グループは敵対的生成アルゴリズムを採用し，ノイズが混入する観測データから銀河分布を再現することに成功した．この学習方法では，シグナル分離器(生成器)の他にもう一つ別の判別器を用意する．図 6 に示すように，判別器はシグナル分離によって生成されたデータと真の大規模構造データを分別する．より具体的には，判別器は入力画像に対して一つの数字を返すような「関数」としてトレーニングされ，真のデータに対しては 1，生成された（偽の）データに対しては 0 を返すように学習する．一方シグナル分離器は，判別器を出来るだけ騙せるように（生成した画像が判別器に本物であると判断されるように）学習する．このようにして二つのネットワークを敵対的に学習させることで，高い精度でシグナルを抽出できるようになる．図 6 に示す「偽のデータ」は，実際に学習済みのシグナル分離器を模擬観測データに適用した場合の出力である．真のデータにおいて見られる銀河の特徴的な分布がよく再現できていることがわかる．

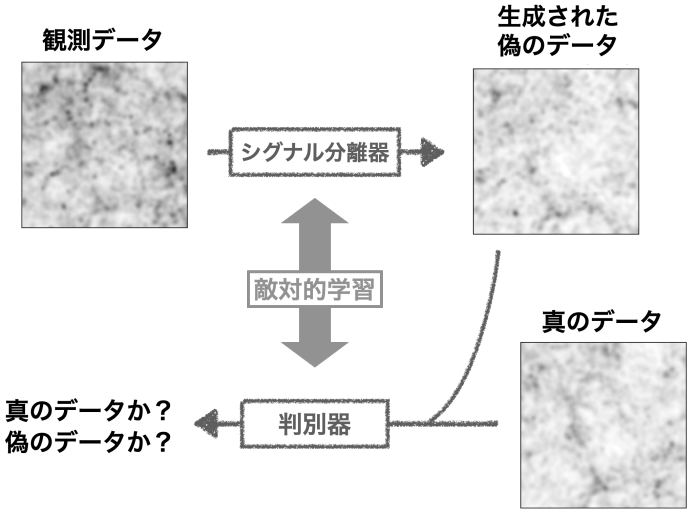

図 4. 敵対的生成アルゴリズムの概念図．生成器は観測データ画像からシグナルだけを取り出し，人工的な（偽の）データを作り出す．判別器は生成された画像の確らしさを判別し，この二つが「協働」することで互いの能力が向上していく．

ここまでは主に二次元画像を念頭に解説してきたが，同様の手法は三次元データに拡張することができる．この場合，スペクトル輝線の固有波長という物理的な情報をネットワークに与えることでさらに高い精度でのシグナル分離が可能であると分かった．図 5 に輝線強度マッピング観測で得られる三次元データに敵対的生成ネットワークを適用した場合の結果を示す．観測データ（上段）はノイズが支配的で，中段に示したような銀河からの輝線シグナルは埋もれてしまっている．十分にトレーニングされた敵対的生成ネットワークは、観測データから即座にシグナルを抽出し(下段)，結果は実際の分布を非常によく再現している．この学習プロセスでは次節でも紹介する宇宙の構造形成の数値シミュレーションを用いて 50000 個もの 3 次元分布とノイズ場を用意し，生成器と判別器をトレーニングした．

ML/AI の多次元データへの応用は爆発的にすすんでいる． 3 次元データの分析は衛星から地球を観測するといったリモートセンシングなどでも需要があり，今後は環境問題等の社会的課題にも適用できるような手法の開発を目指したい．

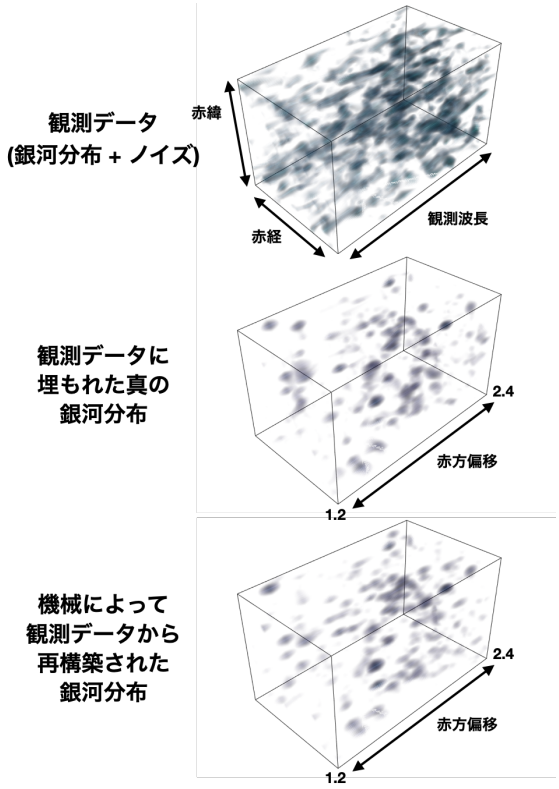

図 5 輝線強度マッピングを想定した 3 次元の模擬観測データに敵対的生成ネットワークによるシグナル抽出を試みた。種々のノイズが混入する観測データ(上)から真の銀河分布(中央)を再構築することができる（下の結果).

3.3 エミュレーション技術の開発

宇宙論の研究で数値シミュレーションは重要な役割を果たしてきた．そもそも天体や宇宙そのものといった対象はテーブルトップ実験が困難なものであるため，観測結果との直接比較や理論モデルの検証のためには数値実験（シミュレーション）が有効である．また，宇宙の構造や天体の形成については次のような特殊な事情によって，数値シミュレーションが大きな力を発揮している．一言でいえば，初期条件が観測的に知られているのである．宇宙初期には物質と光が入り混じった高温のスープのような状態であったが，その中にはわずかな物質密度の揺らぎが存在し，その揺らぎを種として宇宙の構造は形成されたと考えられている．これまでの様々な観測結果から，宇宙の進化に関する標準的な理論モデルが確立されており，基礎物理過程にしたがって，宇宙が誕生した頃の物質分布(初期条件)を統計的に正確に再現することを可能にしている．したがって，宇宙論研究のフォワードモデリングにおける残りの作業は，適切な物理過程を厳密に取り入れ，観測可能な領域内での天体や構造の形成と進化を克明に再現することである．

現代宇宙論ではいくつかのパラメータによって理論モデルが指定される．パラメータと

いっても物質密度や現在の空間膨張率のようにきちんと定義された物理量であり，それらの値が異なると宇宙の進化が変わる．これらの「宇宙論パラメータ」とよばれる基本的な量を正確に測り，正しい宇宙モデルを特定することが観測的宇宙論の大きな目的である．最重要の基本パラメータは 6 つあり，様々な観測からそれらの値を推定する際にはベイズ統計やマルコフ鎖モンテカルロなど，現代的な統計的推論の手法が採用される．残念ながら宇宙の平均の物質密度といった基本量でさえも直接的に測定する手段はなく，例えば物質の空間的分布に関する統計量，すなわち観測データの要約統計量が使用される．適切な要約統計量を用いることでデータベクトルの量を大幅に削減しつつ高精度の統計的推論が可能となる．このための強力な手法が本節で紹介するエミュレーションである．計算コストの高い数値シミュレーションを，より安価で正確な統計モデルに置き換えて推論プロセスを遂行する．

理論モデルの予言と観測データの比較は要約統計量ベースで行うため，正確な統計解析結果を得るには，宇宙の構造形成の数値シミュレーションをすべてのパラメータセットに対して行うことが望ましい．しかし計算速度および計算資源を考えるとこれは現実的ではない．富岳コンピュータを使っても 3 日かかるような大規模シミュレーションを何百回も行うわけにはいかないからだ．主要な宇宙論パラメータに限っても 6 次元パラメータ空間上でサンプル点を密に置き，各点で多数のシミュレーションを実行しなくてはならないため，典型的な次元の呪いに立ち向かうことになる．

私たちの研究チームは，ラテン超格子を用いてパラメータサンプリングを最適化し，ガウス過程を用いて任意のパラメータ点で統計量を計算するツール「Dark Emulator」を開発した．エミュレータの概念は，要約統計量の補間や外挿を利用する従来の方法とは根本的に異なる．エミュレータは回帰問題を扱う統計モデルであり，精緻なシミュレーションによって得られた入力と出力の関係を学習する．最終的に測定したい物理量のベイズ推論に基づいて出力を生成し，多くの場合，パラメトリックな手法を採用しない．これは，関数形とその係数を手動で決定しなければならない関数フィッティングなどとは対照的であり，エミュレータは複雑な問題に対してもデータ駆動的に較正することができる．実観測データの分析結果にしたがってサンプリングを密にするなど，自律的な構成にすることも可能である．図 6 に Dark Emulator の出力精度を示す.

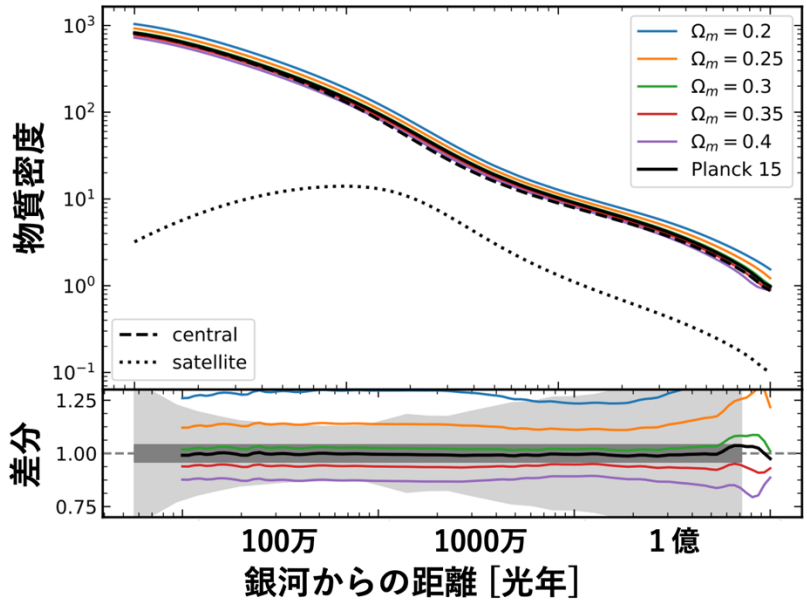

図 6 エミュレータによる物質分布の表現とその精度．パラメータΩで指定される宇宙モデルに対して，銀河まわりの物質密度分布を正確に出力する．図の下部には，数値シミュレーションの結果との残差も示し，残差（エミュレータの精度）はモデル間の相違にくらべて有意に小さくなっていることがわかる．

2019 年に公開した Dark Emulator は現在までに一万五千回以上ダウンロードされており，専門の宇宙論研究者に使われることも多くなってきた．実際に観測データの分析に使う研究者は世界にも数百人もいるか，というレベルなので，15000 回というのは驚きの数字である．簡単に調査した結果によれば，宇宙や物理に関係なく，ML/AI の研究開発に従事している人が様々な目的で使っているようである．エミュレーション技術を社会の様々なシミュレーションに応用するための礎とできるかもしれない．

4. 今後の宇宙観測とサイエンスする AI

世界中で次々と最新の望遠鏡が稼働し，広範囲の波長帯で宇宙を観測する計画がすすんでいる．電磁波以外にも重力波や高エネルギーニュートリノといった新しい情報源を複合的に使うマルチメッセンジャー解析も宇宙物理学の新たな潮流となっている．ところが次世代の観測によって提供されるデータの量は現在の技術や処理能力では太刀打ちできない．本稿で繰り返したように，科学成果にむすびつけるためには膨大な量のデータを迅速に処理し，統計解析を行うための効率的な ML/AI アプリを開発することが急務である．ML/AI の研究自体は急速に発展しており，科学研究だけでなく社会の実際的な問題に適用する新しいアプローチが文字通り毎日提案されている．一方で，ML/AI に固有の問題も懸念されており，偽の相関や非物理的な生成物など，従来の科学研究の手法とは相容れない

側面もある．将来の観測から真に重要な科学的成果を得るには，分析結果が物理学や統計学の観点から合理的に理解されなければならない．そのために説明可能な AI など，より信頼性の高い技術が求められている．

本稿では ML/AI を活用した宇宙観測データ分析を例として，いわば AI for Physics について紹介したが，逆のアプローチである Physics for AI も今後重要になると考えている．宇宙観測や基礎科学のデータはビッグデータとして十分な量で，測定誤差などもよくコントロールされており，データとして質のよいものが多い．物理法則や基礎理論はフォワードモデリングを信頼性のあるものにしており，多くの現象に対して直接数値シミュレーションの結果を ground truth として用いることができる．ML/AI 研究の play ground として基礎科学が果たす役割は大きいと思われる．

文字通り「宇宙から降ってくる」観測データには，宇宙の起源や成り立ち，さらには宇宙に起こる不思議な現象に関する情報が眠っている．歴史的には 16 世紀のティコ-ブラーエの観測ノートの分析から，やがて万有引力の法則の発見につながったことがよく知られている．現代の巨大観測データから AI が新たな現象やより根源的な法則を発見する日が来るかもしれない．日々開発される新しい手法の中から基礎科学の研究に適したものを選択し，信頼できるデータ分析を行っていくためには，データ科学や計算機科学の専門家との協力や，分野横断的な視野を持った研究が一層重要となる．