データ科学とAI

河原　吉伸

1. はじめに

昨今では急速に社会への浸透が進むAI技術であるが，この流れは様々な分野における科学研究でも同様である．チューリング賞の受賞者であるJ. Gray博士が2009年に提唱した「データ集約型科学」は，経験科学から理論科学，そして計算機を用いたシミュレーション・ベースの科学に続く「第4の科学」として位置づけられ，今後の科学研究の標準的なスキームになろうとしている.[1] データ集約型科学自体は，データベースやワークフロー管理，可視化，クラウドコンピューティングなどを含む，データ収集からその分析，そして科学的知見へのフィードバックまでにいたる総合的な枠組みである．その中で，集約されたデータから現代科学のコンテキストに直接アクセスするためには，数理統計や機械学習，計算科学をはじめとした様々な数理的方法が本質的な役割をはたす．

タイトルにもある「データ科学」という言葉の定義は，今でもいくつかの異なる文脈で用いられているように見受けられる．しかし一般には，科学研究におけるなんらかの特定の目的の下，統計やデータ解析，機械学習などの様々な数理やモデリング手法，アルゴリズム，計算機システムを横断的に駆使して，目的に資する情報をデータから抽出する一連の方法やアプローチを扱う学際領域のことを，データ科学として捉えるという考え方が広く受け入れられている.[2,3] このようにデータ科学は‘統計’と‘計算’という側面が重要である一方，そもそも科学研究は人類の知識獲得の手段の一つである訳なので，データから得られる情報と‘知識’をどのように結びつけるのか，という視点も必要不可欠である.[4] 本稿では，筆者の理解する範囲で，AIを用いてこのデータ科学にどのような視点からアプローチすることが大事なのか，最近の話題とともに考えていきたい．

2. データ科学におけるAIの役割

様々な科学分野における研究はこれまで,「仮説と検証」，つまり対象となる現象の観察を通して，数理モデルを用いた表現による一般化とその解析・シミュレーションを繰り返すことで進展してきた．このアプローチは依然として科学研究の基本的な姿勢ではあるが，一方で，上述のように，飛躍的な計測技術・情報インフラの発展を背景にその考え方は変わりつつある．つまり，膨大な観測・計測データ，いわゆるビッグデータを扱う必要性から，従来通りの観察という姿勢ではコストが膨大になってしまったり，そもそも本質が何かを見極めることが困難になってしまうという問題に直面することになり，多くの分野で科学研究への新たなアプローチが求められるようになっている．

このような問題に対してデータ科学が果たす役割は様々であるが，特に次の二つの観点が重要であると筆者は考えている (図1も参照のこと)．ま

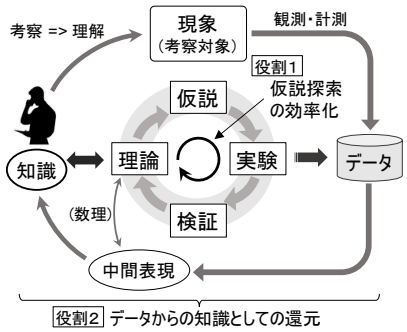

図 1　科学研究における仮説・検証とデータ科学

ず一つは,「1. 仮説探索の効率化」である．つまり，膨大なデータを元にした，仮説立案や検証のための情報の提示や自動化などを通して，仮説・検証ループの反復を削減してそのプロセスを加速する．そしてもう一つの重要な役割は「2. データからの知識としての還元」である．データから抽出された情報を，科学者が蓄積してきた知識と結びつけ，科学的知見としてフィードバックする機能である．

データ科学のこのいずれの役割においても，我々はなんらかの統計的手法の適用を考える訳であるが，その際に注意する必要があるのが，いわゆる「予測」と「推測」の違いである.[5,6] 残念ながら，専門家の間でも，往々にしてこれらの考え方を混同して手法が適用されていることが多々見受けられる．AI 的な文脈において「予測」という言葉が用いられる場合，必ずしもシミュレーション，つまり時間的に将来の対象の状態を推定することを意味する訳ではない．より一般に，モデル構築時においては未知だった新しいデータに対して，その対応する量を推定する問題をさす．例えば，手元にある画像データを用いて被写体を指定するモデルを作る場面において，モデル構築時に用いていなかった新しい画像の中に何が写っているかを推定する場合，‘予測’という言葉が用いられる．一方で推測というのは，データを生成しているメカニズムそのものを推定 (同定) する問題をさす．統計的検定は，この場合に一般的によく用いられるアプローチの一つである．

これらの予測と推測の違いは，上記のデータ科学の 2 つの役割を考える上でも重要である．つまり，仮説探索の効率化においては，予測が十分に機能し仮説立案に使えるのであれば，データの背後にあるメカニズムを知る必要は必ずしもない場合も多い．一方で，データからの抽出情報を知識と結びつけるためには，その統計的な評価や背後のメカニズムとの関係の議論は必須となる．つまり，推測の観点がより重要となる．従って，これらの違いを理解した上で，それぞれの目的のために作られたモデルや手法を適切に使い分けることが，データ科学へ取り組む際には必要不可欠となる．

3. データにより駆動される科学

科学研究におけるデータ科学の重要な役割の一つが，仮説探索の効率化であることは先述の通りである．つまり，データ科学は，科学研究における仮説・検証のループを加速する一つの礎として機能する.「データ駆動科学」という言葉も最近ではよく耳にするが，これは正に，データ科学のこの側面をうまく捉えた表現であるとも言える．

ベイズ最適化

この役割の重要な成功例としては，材料科学分野への適用があげられる.[7] 例えば，所望の材料特性などを得るための実験計画における，機械学習モデルを用いたベイズ最適化 (またはブラックボックス最適化) は大きな注目を集める枠組みである．

ベイズ最適化の基本的な考え方は次のようである.[*1] まず，調整し得る変数 $\boldsymbol{x}$ (例えば材料の混合具合や加工温度など) から，最大化したい材料特性への関数 f を，これまで得られているデータから推定される代理モデル $\hat{f}$ として保持しておく．一般にベイズ最適化の文脈では，代理モデル $\hat{f}$ としてはガウス過程回帰がよく用いられる.[*2] なお，代理モデル $\hat{f}$ は，あくまでこれまでのデータから推察される f の‘代理’なので，これまであまり探索できていない $\boldsymbol{x}$ の領域では不確実性が伴うことに注意する．関数 f を最大化する $\boldsymbol{x}$ となり得る

*1)　より詳しい内容については解説[8] などを参照されたい．
*2)　ガイス過程回帰については書籍[9] などが参考になる．

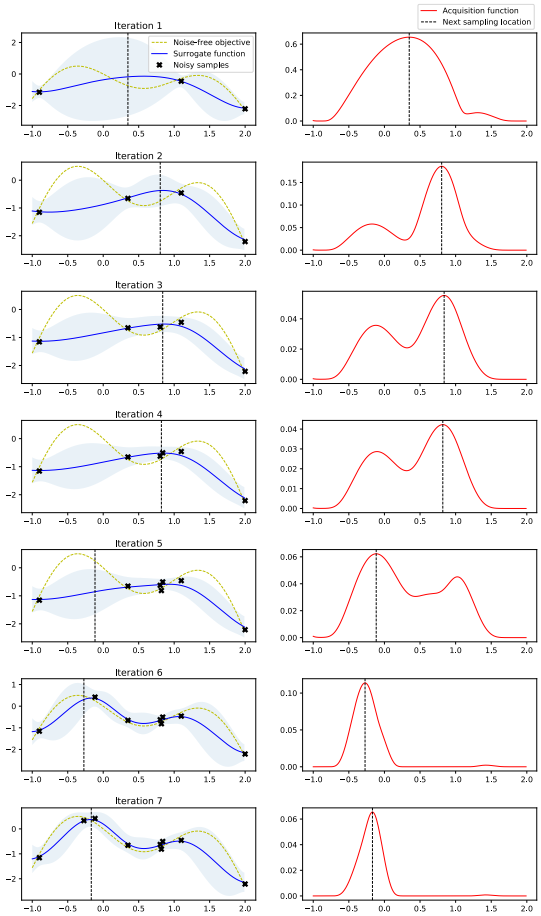

図 2　代理モデルとしてガウス過程回帰を用いた場合のベイズ最適化の数値例.

点を探索する際には，基本的には代理モデル $\hat{f}$ で関数値が大きくなる点を候補とする一方，この不確実性も考慮して探索点を選択する．この戦略を「探索と活用」とも言う．なお図 2 は，変数が 1 次元の場合の，ガウス過程回帰を用いたベイズ最適化の数値例である．

ベイズ最適化を用いた実験計画は，材料科学分野に限らず，種々の科学分野における第一原理計算のパラメータ探索や，創薬における分子構造の探索など，多くの領域で適用されている．

代理モデル

材料科学分野におけるベイズ最適化の適用では，仮説 (この場合，材料特性の組合せや実験パラメータなど) から材料特性への関数の機械学習モデルによる表現と，探索と活用に基づく仮説空間の効率的な探索，の 2 つの要素がうまく組み合わせ用いられている．このケースでは前者として，ガウス過程回帰が用いられることが多いというのは前述のとおりである．この理由は，ガウス過程回帰が予測の信頼度を出力できるため，不確実性を定量化でき探索と活用に利用できるためであった．これに限らず，データ科学の様々な場面において，対象の性質をモデリングするためにどのような代理モデルを用いるか，という選択がデータ科学の成功を分ける重要なファクターとなり得る．

代理モデルを得るシンプルでかつ有用な手段の一つは，入出力関係を表すと予想される関数 (記述子) のセット $\boldsymbol{\phi}(\boldsymbol{x}) := [\phi_1(\boldsymbol{x}), \ldots, \phi_m(\boldsymbol{x})]^\mathsf{T}$ ($\boldsymbol{x}$ は予測を与えるための入力にあたる変数) を事前に用意しておき，データからそれらの中で重要と判断されるものを重み付けして選択するという方法である．例えばこの際，統計や機械学習で議論されるスパース推定がよく用いられる．スパース推定とは，モデル中のパラメータのうち，その多くが 0 となるように推定するための一連の方法をさす.*3) 言い換えると，予測に重要となる記述子の部分空間を抽出する手続きであるとも言える．例えば，線形モデル $\hat{y} = \boldsymbol{w}^\mathsf{T}\boldsymbol{\phi}(\boldsymbol{x})$ (各記述子への重み付け $\boldsymbol{w}$ により目的変数 y の予測 $\hat{y}$ を与える) の場合には，ベクトル $\boldsymbol{w}$ をスパース推定することで，重みが非ゼロとなる記述子を用いて目的変数にあたる量を予測するモデルを獲得する．

スパース推定を実現するための方法はいくつか知られているが，最もよく用いられるものの一つにスパース正則化がある．正則化は，一般にモデルを推定する際に，各モデル f を用いた予測に対する損失を表す関数 $\ell(f)$ へ，別の基準で与えられたペナルティ項 $\Omega(f)$ (正則化項と呼ばれる) を同時に与えて最適化する枠組みをさす．

$$\min_{f \in \mathcal{F}} \ \ell(f) + \lambda \cdot \Omega(f)$$

ただし，$\mathcal{F}$ は推定する関数の範囲を表す集合として

*3)　より詳しい内容については書籍[10] などを参照されたい．

用いている．また λ は両項のバランスを調整するパラメータである．スパース正則化は上記において，正則化項として ℓ_1 正則化などのスパース正則化関数を用いる．例えば線形モデル $f(\boldsymbol{x}) = \boldsymbol{w}^\mathsf{T}\boldsymbol{\phi}(\boldsymbol{x})$ の場合は，与えられたデータ $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$ を用いて，二乗誤差 $\ell(f) = (1/N)\sum_{i=1}^N (y_i - \boldsymbol{w}^\mathsf{T}\boldsymbol{\phi}(\boldsymbol{x}_i))^2$，$\ell_1$ 正則化 $\Omega(f) = \|\boldsymbol{w}\|_1 (:= \sum_{i=1}^m |w_i|)$ として，上式の基準に従い重み $\boldsymbol{w}$ を最適化する．

仮に記述子のセットが事前にうまく必要十分に用意できていたとすると，スパースに (少数のもので) 表現された代理モデルは，これ自体仮説の考察に有益な情報となり仮説・検証ループの加速が期待できる．重要と判断された記述子を解析したり，より良い記述子を導出する手がかりとできるためである．しかし一方，そもそも未知の現象に対して有効な記述子を準備することが容易ではない場合も多く，また手元にあるデータを等しく表現できる記述子の集合は単一とは限らないなど，推定自体の困難さも内在するので注意が必要である．これに限らず，どのように現象を表すモデルを準備するか，という問題はデータ科学で常に注意すべき事項としての認識が必要である．

4. データ抽出情報と科学的知識のリンク

従来的な科学的方法を加速するデータ科学のもう一つの重要な役割は，知識へとつながる情報をデータから抽出する機能である．この際，抽出された情報がどのように科学的知識と結びつけられるのか，という問題を考えることは重要な視点の一つである．

科学における知識は「方程式」として記述されるのが一般的である．特に科学研究においては，自然科学・社会科学によらず動的な現象を扱うことが多いが，これらを表現するためには，常／偏微分方程式を用いるのが一般的である．本稿では，このような考察対象のなんらかの特性 (項目間の関係や，時間的挙動など) を表す方程式を「数理モデル」と呼ぶことにする．昨今では，(時系列) データから直接的に微分方程式を発見しようという試みなども報告が見られる．例えば，ワシントン大の Kutz 教授らのグループは，先にも言及したスパース推定を用いて，事前に用意した関数のセットの中から方程式中の項として時系列データを表すものを選択することで，微分方程式を推定しようというアプローチを報告している.[11] また，同様に事前に準備した関数のクラスや数理的仮定などを元に，探索を繰り返して物理法則の発見を再現する興味深い研究も報告されている[12]．しかしながら，データから直接方程式を推定しようという問題設定は，これらの例の場合のように，その現象を必要十分に表す方程式の項や仮定をうまく事前に絞り込むことができない限り，様々な技術的な難しさが内在している．例えば，仮に得られた方程式が時系列データを精度よく予測できたとしても，それは必ずしもその現象 (データ生成のメカニズム) を表すものと近くなるとは限らない.[13] 同定性の問題 (同じデータを表すモデルが複数存在し得る) や，最適化等の計算による誤差なども綿密に分析する必要もある．上記の報告のように，最近の機械学習手法を用いることで，直接方程式を発見するアプローチは特定の側面ではうまくいっているようには見えるが，今後も十分な検証が必要であろうというのが筆者の考えである．

それでは，直接方程式を発見するアプローチ以外に，どのようにしてデータから科学的知見につながるような情報を抽出する方法が考えれば良いのか．この問いは極めて深遠で様々な観点から考えることが必要であるとは思うが，ここでは 2 つの主要な考え方に着目してみたい．

4.1 中間表現

この問題を考える際にまず大事な視点の一つとして，科学的知識 (数理モデル) とデータをつなぐ中間表現を適切に選択するということがあげられる．ここまで何度か取り上げたような，適切な記述子のセットを事前に用意しておく，というのもこの最も単純なアプローチとして捉えられる．なぜなら，記述子がデータからうまく選択することができれば，その記述子を解釈することで科学的な知見へとフィードバックできる可能性がある．た

だし先述のように，そもそも未知の現象に対して有効な記述子のクラスを準備することは容易ではない場合も多く，また様々な推定自体の困難さも内在するので注意が必要である．

あるいは最近では，量子論で状態を表す量として用いられる波動関数をニューラルネットにより表し，その推定を通して多体問題を解く研究も報告されている．[14] 波動関数は量子多体問題の状態を表す十分な情報を含んでいるため，それが推定されることでこの問題における様々な物理学的知識と直接リンクする．また例えば，Google Brain の Greydanus 博士らは，物理的な現象を表すモデルの推定において，物理分野で一般的に系の記述に用いられるハミルトニアンを表すニューラルネットワークを用いる枠組みを提案している．[15] これらの例のように，各分野での先見知識を元に，一般的に用いられる適切な表現に対して推定問題が設定できるのであれば，それはデータから得られたモデルが直接科学的知識とつながる有効なデータ科学のアプローチとなり得る．

モード分解

また一つのアプローチとして，考察対象が複雑なものであったとしても，我々が理解できる単純なものに分解して解析しようという方法は，データ科学においてもよく用いられる．例えば，統計解析でもよく用いられてきた主成分分析や，時系列データへのフーリエ解析などは，現在でもデータ科学の多くの場面でも有効な手段として活用されている．さらに，それらの古典的な方法を，カーネル法やニューラルネットなどの機械学習的原理によって一般化 (非線形化) した方法も，多くの場面で成功裏に用いられている．

この方向性に関連して昨今注目されているものに，動的な現象から得られるデータの解析に対する「動的モード分解」と呼ばれる方法がある．[16, 17] この方法は，当初は流体分野において提案されたデータ解析手法で，例えば流体現象の計測データのような時空間的広がりをもつデータに対しては，特徴的な周期性を共有する時空間的な共振パターンを抽出することができる．図 3 は，2 次元の円

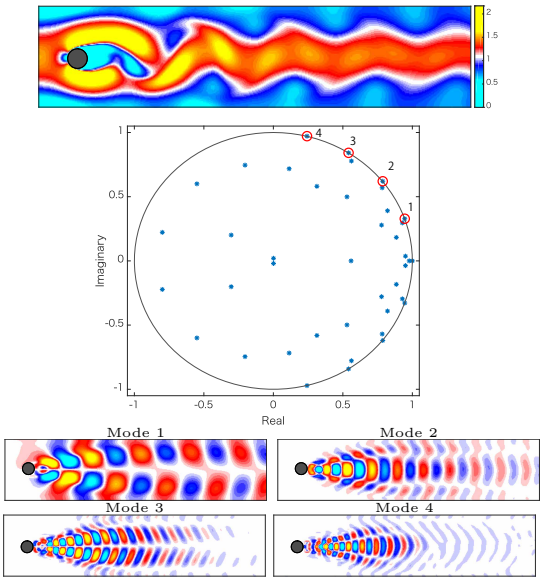

図 3 円柱周りの 2 次元流体のシミュレーションデータに対する動的モード分解の数値例．(上) スナップショットの例，(中) 推定された固有値，(下) 代表的な推定された動的モード (丸印の固有値に各々対応)．

柱まわりの流れ (レイノルズ数 100) に対して，動的モード分解を適用した数値例である．動的モード分解で推定された固有値は複素数でその位相が時間的な周期性を示し，またこれに対応する動的モードと呼ばれる空間パターンが各々得られる．円柱周りの流れはカルマン渦と呼ばれる組織構造が見られることが知られているが，これに近い構造がデータから抽出されている様子が分かる．最近では，動的モード分解は流体分野に限らず，様々な動的現象を扱う科学領域への適用が報告されている．

動的モード分解の重要な点の一つは，推定量と，データを生成していると想定されるメカニズム (力学系) との関係が数学的に議論できることにある．(非線形) 力学系の作用素論的解析，特に，クープマン作用素と呼ばれる線形作用素のスペクトルとの関係が知られている．[18] クープマン作用素，またはその随伴であるペロン・フロベニウス作用素は，数学，または応用数学分野でも盛んに研究されており，力学系の様々な数学的原理や，制御理

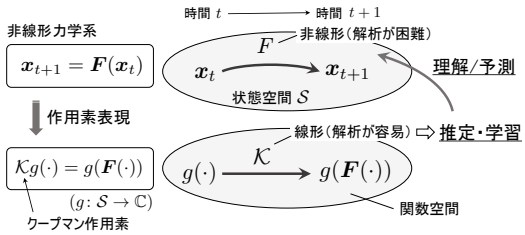

図 4　動的現象のデータ解析に対するクープマン作用素を用いたアプローチの概念図.

論における安定性や非線形物理における縮約理論などの応用数学における様々な概念とも関係がある．動的モード分解はそのスペクトルを有限データから推定する数値計算手法の一つとして位置付けられるが，作用素論的方法に基づくデータ解析の重要性は，一般に非線形で数理的に扱うことが困難な系の時間発展を直接扱うのではなく，その動力学的特性へ作用素を用いた線形表現を介してアプローチする点にある (図 4 も参照)．なお，作用素論的方法に基づく動的現象のデータ解析に関連する導入的事項や技術的詳細については，著者による最近の解説もあるので参考にされたい.[19]

4.2 知識に寄り添う推定

中間表現をうまく選択することで，科学的知識と結びつけることが可能な情報がデータから抽出できる枠組みが得られる．しかしその際に忘れていけない重要な点は，それらがデータから ‘正確’ に推定できる設定を作ることである．ここで ‘正確’ にというのは，必ずしも精度が高いということを意味する訳ではない．先にも述べたように，データの背後にあるメカニズムに近づく推定を行うことが必要になる．例えば先の作用素表現は，データの背後にある複雑なメカニズムに原理的に対応が可能な，線形領域で解析を試みるアプローチである．そのため，推定における非線形性や局所最適性などの困難さを回避し得る場合も多い．このように，中間表現を考察する際には，その推定に伴う種々の困難さについても配慮することで有用な知見獲得へつながるアプローチが得られる．

また，一つの有効なアプローチとしては，推定されるモデルが先見知識の表現に近くなるように，なんらかの方法で推定プロセスをコントロールするという考え方があげられる．例えばテクニカルには，先見知識に近くなるように正則化項を設計してモデルを推定する方法などが有効な手段として用いられることも多い.[20] 先述のスパース正則化において，変数に関するより構造的な情報を取り込んだ正則化項を設計して機械学習を適用するアプローチは，従来からバイオインフォマティクス分野などで主に成功裏に適用されてきた.[*4] 多くのデータ科学の場面においては，併せて取り入れるべき有効な考え方の一つであろう．

5. 最後に

本稿では，機械学習をはじめとした AI の発展を背景に，データ科学により科学研究への向き合い方が変わりつつある中で，どのような考え方でこれらにアプローチしていけば良いのか，著者なりの見方でその一端について紹介してきた．またここまででも見てきたように，データ科学は，単に AI 関連技術の科学分野への応用という側面だけではなく，AI 分野における原理や方法に関する研究や，しいては AI 分野の研究者の考え方にも大きく影響しているように感じられる．著者自身も，機械学習分野の主要な国際会議である NIPS (Neural Information Processing Systems) に 2016 年に参加した際，素粒子物理学で著名な K. Cranmer 教授の当分野におけるデータ科学に関するキーノート講演を聞き胸が熱くなったのを鮮明に覚えている．一方で最近では，2018 年に当会議で最優秀論文賞となった Neural ODE [22] や，本稿でも紹介したハミルトニアン・ニューラルネット [15] などのように，どちらかといえば物理科学など周辺の科学分野で発展してきた考え方が，AI の革新へとつながる例も多く見られる．これらは，データ科学への注目から，AI へ関心を持ち参入する周辺分野の研究者が増えてきているためでもあるようである．

AI をコア技術としたデータ科学は，従来から

---

*4)　著者による簡単な解説[21] もあるので参考にされたい.

データの種類に依存しない汎用的な方法論を議論してきた機械学習がその主要な技術的基礎となっている．そのため，データ科学の多分野への拡がりは，これまで分野ごとに発展してきた技術や知識をつなぐ，分野横断型のアプローチとしてのデータ科学の側面を引き出す可能性もひめている．一方で，このような状況は，AIが今後，様々な科学分野の研究者が素養の一つとして身につける必要性をますます高めることも想像に難くない．いずれにしても，冒頭でも述べたように，データから科学的コンテキストに直接アクセスするデータ科学の深化・展開において，数学・数理科学がはたす役割は今後ますます重要になってくるであろう．

なお，第4節で紹介した作用素論的方法に基づく動的現象のデータ解析に関する理論的・技術的課題の解決や，さらなる発展的内容や応用を議論するプロジェクトとして，JST CREST「数学・数理科学と情報科学の連携・融合による情報活用基盤の創出と社会課題解決に向けた展開」領域において，2019年10月から，著者を代表として「作用素論的データ解析に基づく複雑ダイナミクス計算基盤の創出」(課題名) が開始されている．本課題における研究経過にも是非注目してほしい．*5)

参考文献

1) T. Hey, S. Tansley, & K. Tolle (eds.) (2009): *The Fourth Paradigm: Data-Intensive Scientific Discovery*, Microsoft Research.
2) V. Dhar (2013): Data science & prediction, *Commun. of the ACM*, **56**(12): 64–73.
3) L. Cao (2017): Data science: A comprehensive overview, *ACM Comput. Sur.*, **50**(3): 43 (2017).
4) D. Blei, & P. Smyth (2017): Science and data science, *PNAS*, **114**(33): 8689–8692.
5) L. Breiman (2001): Statistical Modeling: The Two Cultures, *Statistical Science*, **16**(3): 199-231.
6) G. James, D. Witten, T. Hastie, & R.Tibshirani (2013): *An Introduction to Machine Learning with Applications in R*, Springer.
7) R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi, & C. Kim (2018): Machine learning in materials informatics: Recent applications and prospects, *npj Comput. Mater.*, **3**: 54.
8) E. Brochu, V.M. Cora, & N. de Freitas (2010): A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning, arXiv:1012.2599.
9) 持橋大地, 大羽成征 (2019): ガウス過程と機械学習, 講談社サイエンティフィック (機械学習プロフェッショナルシリーズ).
10) 冨岡亮太 (2015): スパース性に基づく機械学習, 講談社サイエンティフィック (機械学習プロフェッショナルシリーズ).
11) B.L. Brunton, J.L. Proctor, & J.N. Kutz (2016): Discovering governing equations from data by sparse identification of nonlinear dynamical systems, *PNAS*, **113**(15): 3932–3937.
12) S.M. Udrescu, & M. Tegmark (2020): AI Feynman: A physics-inspired method for symbolic regression, *Sci. Adv.*, **6**(16): 2631.
13) C. Perretti, S. Munch, & G. Sugihara (2013): Model-free forecasting outperforms the correct mechanistic model for simulated and experimental data, *PNAS*, **110**(13): 5253–5257.
14) G. Carleo, & M. Troyer (2017): Solving the quantum many-body problem with artificial neural networks, *Science*, **355**(6325): 602–606.
15) S. Greydanus, M. Dzamba, & J. Yosinski (2019): Hamiltonian Neural Networks, in *Adv. in Neural Info. Proc. Sys. 32*, pp.15379–15389.
16) C.W. Rowley, I. Mezić, S. Bagheri, P. Schlatter, & D.S. Henningson (2009): Spectral analysis of nonlinear flows, *J. Fluid Mechanics*, **641**: 115–127.
17) J.N. Kutz, S.L. Brunton, B.W. Brunton, & J.L. Proctor (2017): Dynamic Mode Decomposition: Data-Driven Modeling of Complex Systems, SIAM.
18) K.K. Chen, J.H. Tu, & C.W. Rowley (2012): Variants of Dynamic Mode Decomposition: Boundary Condition, Koopman, and Fourier Analyses, *J. of Nonlinear Sci.*, **22**: 887–915.
19) 河原吉伸: 非線形力学系の作用素論的データ解析と動的モード分解: 正定値カーネルを用いた定式化とその拡張を中心として, 数理解析研究所講究録 (*in press*).
20) A. Karpatne et al. (2017): Theory-Guided Data Science: A New Paradigm for Scientific Discovery from Data, *IEEE Trans. on Knowl. Data Eng.*, **29**(10) 2318–2331.
21) 河原吉伸 (2016): 造的スパース推定とその最適化, 電子情報通信学会誌 (特集「スパースモ デリングの発展 ー原理から応用までー」), **99**(5): 386–391.
22) R. T.Q. Chen, Y. Rubanova, J. Bettencourt, & D. Duvenaud (2018): Neural Ordinary Differential Equations, in *Adv. in Neural Info. Proc. Sys. 31*, pp.6571–6583.

(かわはら よしのぶ，九州大学／理化学研究所)

*5) http://jp.comput-dynamics.org/