データ科学における予測と理解の両立を目指して
― データ科学と物理 ―

寺倉清之
2019/6/5

予測と理解

ここで問題にする理解は、
　新事実の発見の後に、天才の閃きによる新しい理解、
というのではなく
　データから機械学習的に得る理解、
というような意味

津田先生

（第１回JST workshop （2013/2/11) では）

機械学習の目的は次の2つ

１．原因究明 （理解）

２．予測

それぞれの目的を徹底的に追及する場合は、2つは別の課題。

山本一成氏 （ポナンザの開発者）

『人工知能はどのようにして「名人」を超えたのか？』

第2章：黒魔術とディープラーニング　－科学からの卒業－

第1節：機械学習によってもたらされた「解釈性」と「性能」のトレードオフ

・・・

最9節：還元主義的な科学からの卒業

平成30年5月21日(月)　　13:00 ～ 17:40

JST東京本部別館1階ホール（東京都千代田区五番町7 K’s五番町)

12:30-　　Open

13:00-　浅野 哲夫　（JAIST)
　　開会挨拶

13:15-　寺倉 清之　(NIMS/JAIST)
　　趣旨説明

13:35-　溝口 理一郎　(JAIST/ISTC-CNR)
　　「分かる」を支援するAIとは？

14:20-　日高 昇平　（JAIST)
　　情報・学習・理解：機械学習から
　　機械理解の定式化に向けて

15:05-　★★★ ブレイク ★★★

15:20-　田村 亮　　(NIMS/東大)
　　機械学習からのフィードバックによる
　　材料科学的理解の促進事例

16:05-　ダム ヒョウ チ　（JAIST)
　　データマイニングを活用した物質空間の
　　理解に向けて

16:50-　瀧川 一学　　（北大)
　　機械学習は真の発見に寄与できるのか？

17:35-　伊藤 聡　　(NIMS)
　　閉会挨拶








理解の2つの問題

・ある現象の原因の理解

因果関係の理解　　（物理での ）understandable

原因究明

・機械学習による予測が、どのようにしてなされたか、その仕組みや根拠を知る

AI のホワイトボックス化

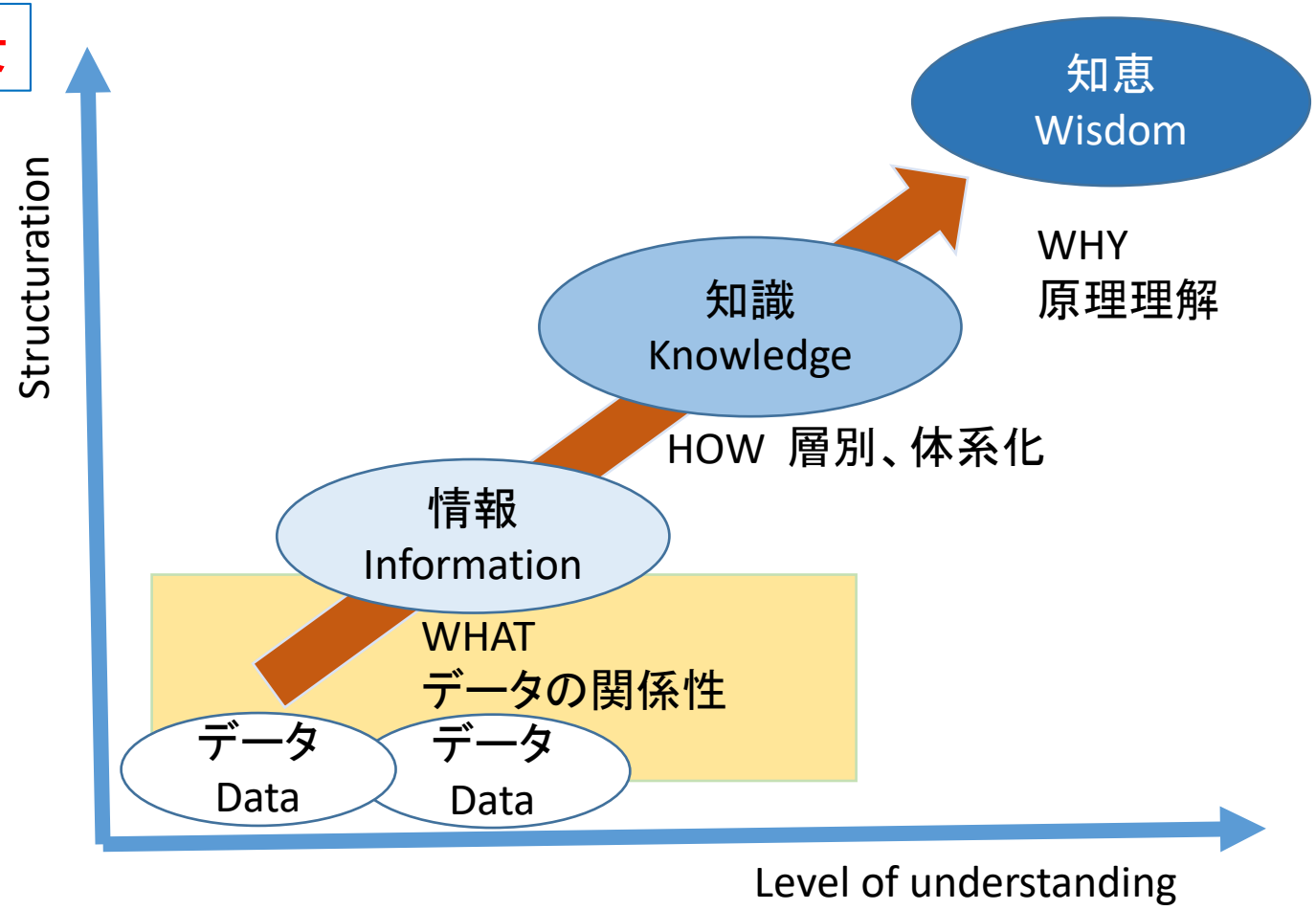
説明可能性　　（機械学習での）explainable

真鍋明氏作成

Wisdom Hierarchy (DIKW)

Integration of Data, Information to higher levels

本日の講演会の背景



J. Rowley: Journal of Information Science 33(2007) 163

内容

データ科学の意義

物質科学における解析型研究と開拓型研究

More is different

最近の研究から：“Reliable and explainable machine learning”

データ科学の意義

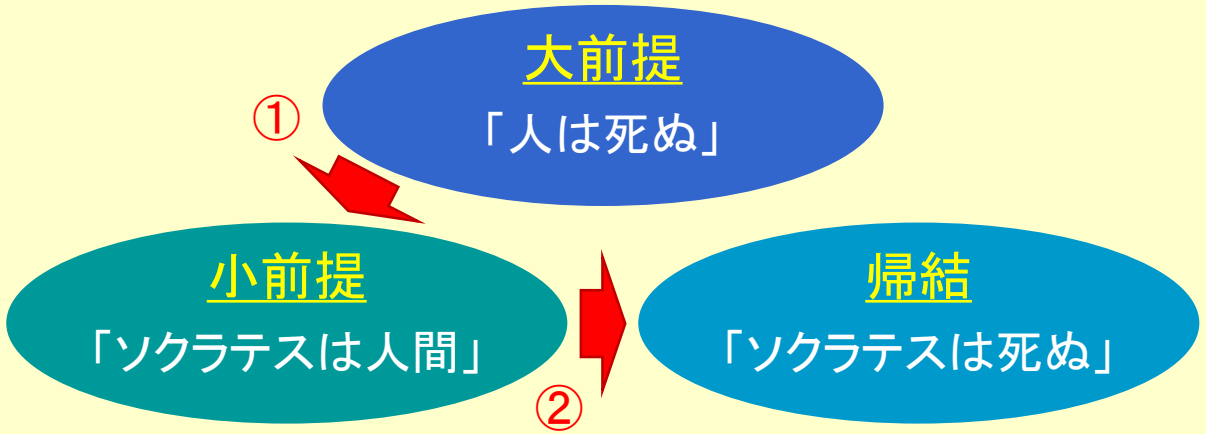
開発の効率化だけではない

自然科学の４本柱と推論の４要素

データ科学が加わって、自然科学の基盤が整った。

推論の3形式

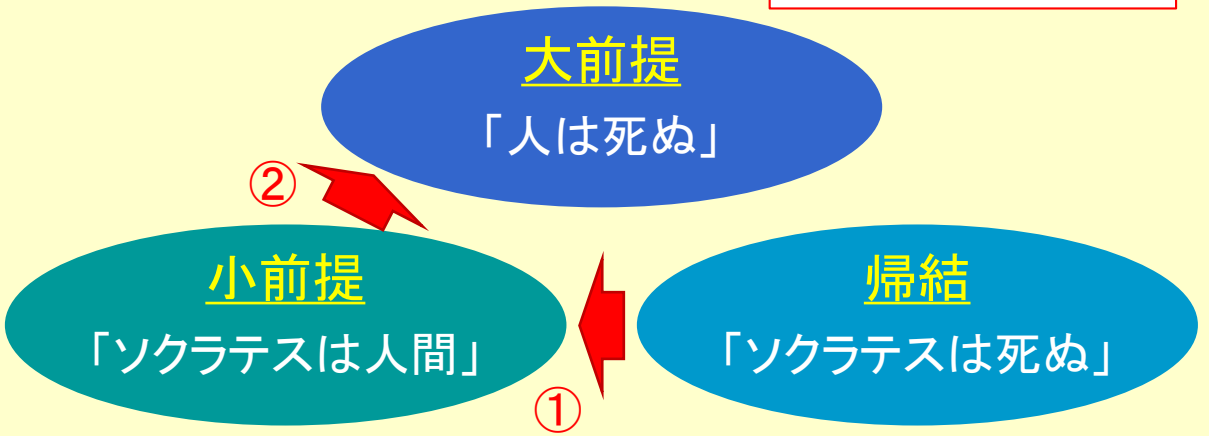
演繹法：必然的な推論



例えば、計算科学

帰納法：蓋然的な推論

間違った大前提を導く可能性あり

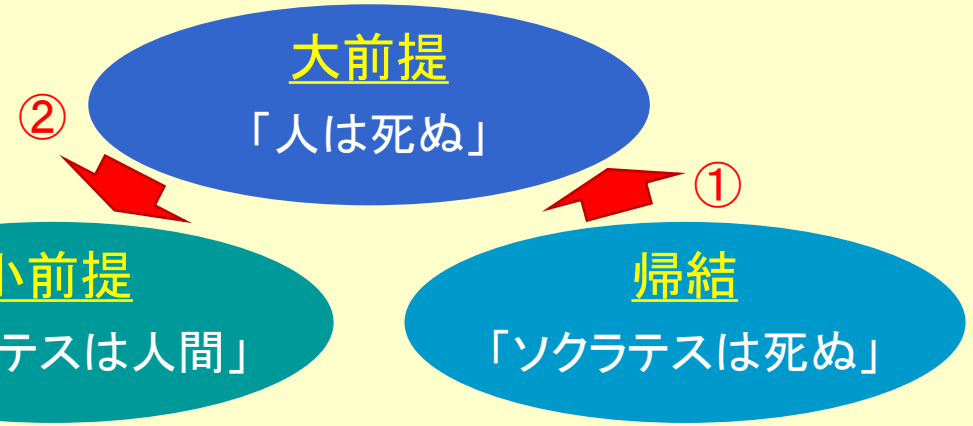


データ科学

（仮説推論）：蓋然的な推論

アブダクション

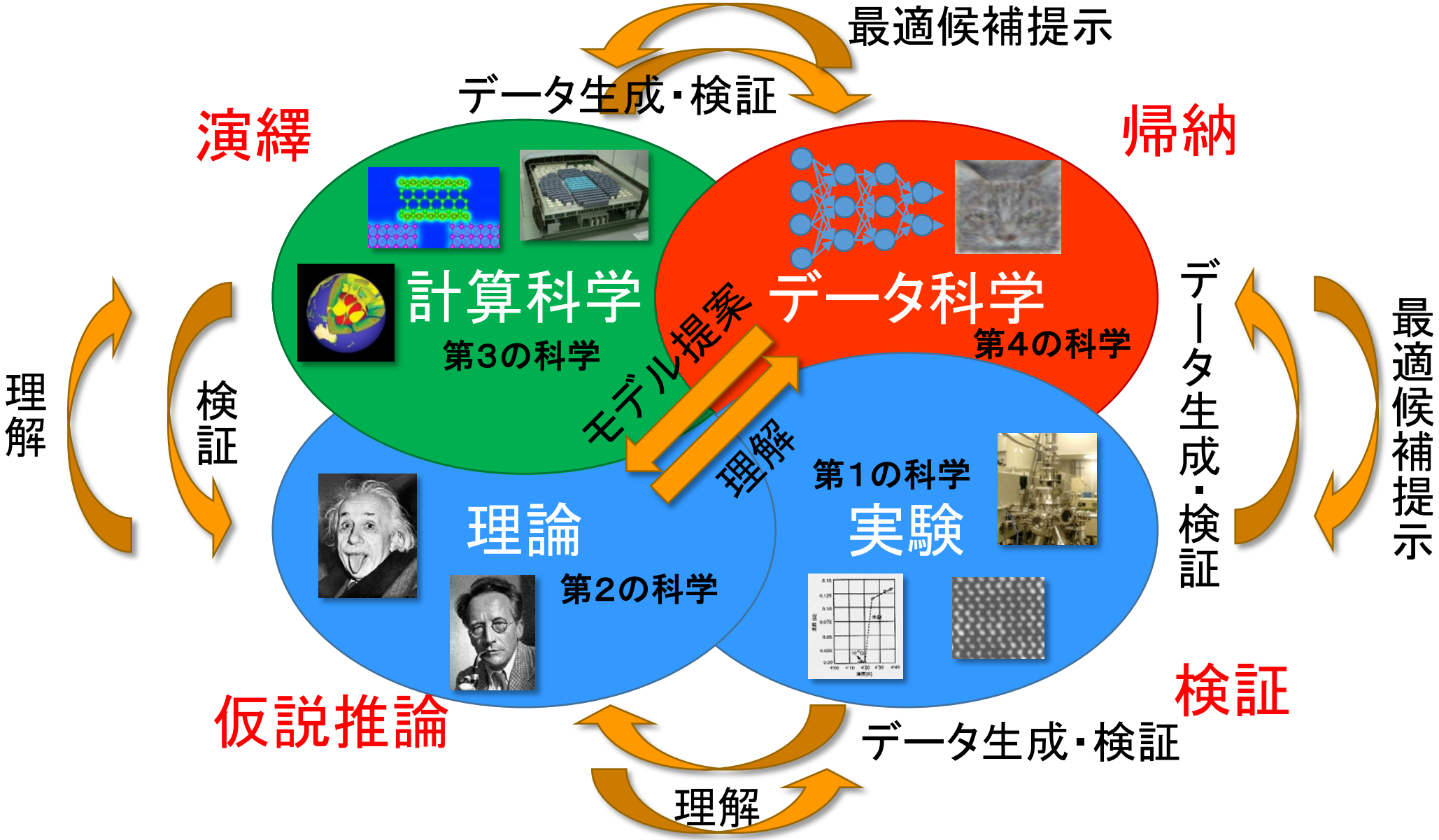
間違った小前提を導く可能性あり



理論

C. S. Peirce
(1839 – 1914)

（推論の3形式＋検証）と 自然科学の4本柱 との関係



物質科学における解析型研究と開拓型研究

解析から予測へ、予測から設計へ

自然科学、例えば物理、では、多くの研究が解析型研究。

観測された現象を支配する基本原理（法則）を見出すことが非常に重視される。

法則が得られたら、それに基づいて（因果関係により）種々の予測が可能になる。

一方で、新現象の発見、新物質の発見が重要であり、物質設計は物質科学の重要課題。
（私は）これを開拓型研究と呼んでいる。

物質・材料設計は逆問題

順問題 ⇔ 逆過程 ⇔ 逆問題

解析型

演繹的

given

原因：物質・材料 → 因果関係 → 結果：物性、機能

物質・材料研究では、強力な演繹的研究手法がある。:
実験、理論、計算科学

開拓型

帰納的

given

原因：materials ← 相関関係 ← 結果：物性、機能

データマイニング（機械学習）は強力な帰納的研究手法。

寺倉清之：MATERIAL STAGE 15巻10号 p.1 (2016)

回帰解析

機能: $F = f(x_1, x_2, \cdots)$ (1)

記述子: $x_1, x_2, \cdots$ (2)

目的の機能Fと記述子$(x_1, x_2, \cdots)$の間の関係式(1)を導くだけでなく、記述子$(x_1, x_2, \cdots)$を決める。

記述子：①機能 F を制御する物理量であり、
②その機能に関して、対象の間の類似性を測るのに適している。

データからの回帰で得られる式（1）は、相関関係と呼ばれる。
相関関係は必ずしも因果関係ではない。 ➡ 予測の検証が必要

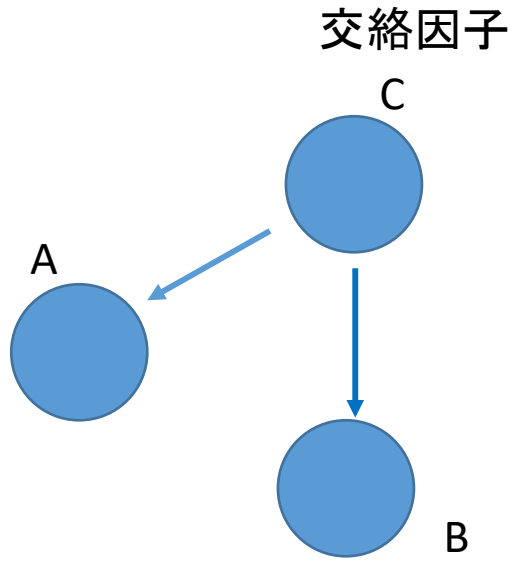
All models are wrong; some models are useful.

George Edward Pelham Box
(October 18, 1919 – March 28, 2013)

因果推論

交絡因子

相関関係と因果関係を混乱させる一つの重要な要素



AとBの間には、直接の関係はないが、
AもBもCの影響を受けるために、
見かけ上、AとBに関係があるように見える。

1. 交絡因子の問題の対応法が種々議論されているが、
   オントロジー解析で明確になるのでは？
2. 層別解析：　機械学習での「クラスター解析＋回帰」と関連

予測と理解（原因究明）

単純世界と複雑世界

予測と理解

ここで問題にする理解は、
　新事実の発見の後に、天才の閃きによる新しい理解、
というのではなく
　データから機械学習的に得る理解、
というような意味

津田先生

（第１回JST workshop （2013/2/11) では）

機械学習の目的は次の2つ

１．原因究明 （理解）

２．予測

それぞれの目的を徹底的に追及する場合は、2つは別の課題。

山本一成氏 （ポナンザの開発者）

『人工知能はどのようにして「名人」を超えたのか？』

第2章：黒魔術とディープラーニング　－科学からの卒業－

第1節：機械学習によってもたらされた「解釈性」と「性能」のトレードオフ

・・・

最9節：還元主義的な科学からの卒業

Vapnik: Einstein said “when the solution is simple, God is answering”. That is, if a law is simple we can find it. He also said “when the number of factors coming into play is too large, scientific methods in most cases fail.” In machine learning we deal with a large number of factors. So the question is what is the real world? Is it simple or complex? Machine learning shows that there are examples of complex worlds. We should approach complex world from a completely different position than simple worlds. For example, in a complex world one should give up explain-ability (the main goal in classical science) to gain a better predict-ability.

http://www.learningtheory.org/learning-has-just-started-an-interview-with-prof-vladimir-vapnik/

More Is Different

多数の集団は個とは異質である
：　金森先生による和訳

Broken symmetry and the nature of the hierarchical structure of science

Science 177, 393 (1972)　P. W. Anderson

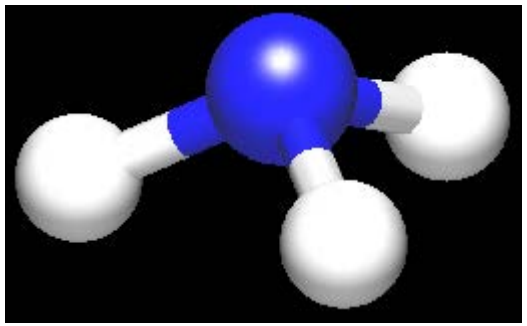
X	Y
Solid state or many-body physics	Elementary particle physics
Chemistry	Many-body physics
Molecular biology	Chemistry
:	:
Psychology	Physiology
Social science	Psychology

上記のハイアラーキにおいて、”X” は単に ”Y” の応用ということを意味するのではない。それぞれの段階に新しい法則、概念、一般化が必用になる。一つ前の段階 ”Y” でと同じように、”X” においてもインスピレーションや創造性が求められる。

More is different の例

空間の対称性と対称性の破れ

アンモニア分子 NH_3



N が上にあるのと下にあるのは
トンネル現象で等確率

固体の強誘電体

プラスの電気分極がトンネル現象でマイナスの
電気分極になることはない。Broken symmetry

力学特性の記述

原子スケール

原子変位に対する、原子に掛る力
ビリアル定理によるストレス

マクロスケール

連続体を使った弾性論

More is different の要約

1. 複雑な系になってもなお、その複雑系を支配する基本法則があるはず。

2. 要素が多数集まった系を支配する法則は、要素の振舞を支配する元の法則とは異なっており、そのことを創発特性という。

3. 社会まで含めた自然界は全て、“要素”と“要素の集合体”、という層が順に重なってできた階層構造を成しており、要素の集合体の振舞は創発特性として理解される。

この概念を、データ科学に活かす！

“More is different” の話

樺島祥介 [1]
東京工業大学大学院総合理工学研究科

5　おわりに

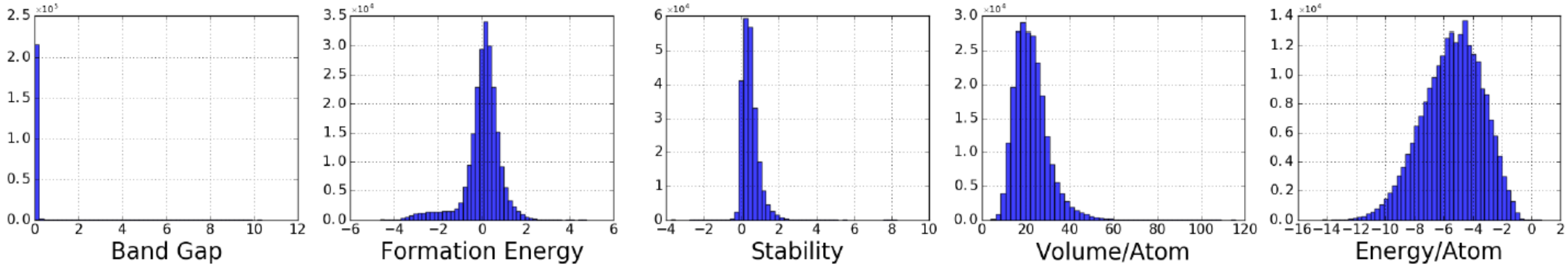
モノの科学とコトの科学の大きな違いは，自然という客観的かつ絶対的な存在に寄り掛かることが出来るか否か，という点にある．この違いが現時点での両分野における考えの進め方の違いに大きな影響を与えている．モノの世界の観察から得られる，量が増えた極限は漆黒の闇ではなく質の異なる明るい世界がきっと開けている，という安心感は自然が常に寄り添ってくれるという恵まれた状況になければ生じにくいのかもしれない．しかしながら一方で，どんなに美しい絵も，素敵な音の調べも，心ふるわせる一篇の詩も，もとをただしてしまうと単なる点やノイズ，字や声音に過ぎなくなる．小さな構成要素が沢山集まることで要素レベルでは予想もつかなかった性質や機能が集合体に宿る．この命題はコトの科学にも当てはまるのではないだろうか．これまでのところまだいくつかの事例しかないのだが More is different という「コトの見方」は情報科学の諸問題にももっと役立つ気がしてならない．

註
モノの科学：自然科学
コトの科学：情報科学

“Reliable and explainable machine learning methods for accelerated material discovery”
B. Kailkhura, B. Gallagher, S. Kim, A. Hiszpanski and T. Y-J. Han, arXiv:1901.02717v2

現在の MI の問題点

・MI では、殆どの場合データが少ないか、歪みが大きいか、という深刻な状況。
例として、OQMD (Open Quantum Materials Database) も歪みが大きい。



・更に、MI での狙いはしばしば、知られていなかった性能の物質の探索。
➔ 歪んだデータにおける、データ点が少ない領域の探索
実質的には外挿になる。予測精度の低下、誤った理解に導く可能性が高い。

・予測と理解の相克のため、より良い理解を得るために、予測性能を犠牲にする。
➔ explainable だが、misleading な予測

論文の内容の概略

1. Underrepresented and distributionally skewed data から学習する場合、訓練、検定、不確かさの定量化における従来の方法の欠点を明らかにする。

2. 上記のような困難な状況下においても、説明可能で信頼度が高い機械学習の方法を提案する。その方法は次のような要素から成っている：

   - データが少ない領域の探索に耐えられるような機械学習の訓練法

   - 標準的な機械学習の方法が使えるように、訓練データを操作するサンプリングアルゴリズム　(sub sampling)

   - 狙いからのバイアスをより良く把握するために、信頼のおける評価計量と不確かさの定量化

具体的な提案内容

・それぞれの目的変数の値に従って、データをいくつかのサブクラスに分割

・それぞれのサブクラスに対して、単純なモデルを訓練
　この分割は、説明可能性の向上をめざすもの
　単純なモデルの使用により、予測性能は低下

・予測性能の向上のために、転移学習を活用

・MLシステムの解釈性向上のために、手法の枠組みに「論拠生成器（rationale generator component) を追加。この生成要素の目的は、
1) 予測に対する説明：prototypes を与えて、物質科学者の知識による判断を助ける。 Decision-level
2) 回帰モデルの説明：物質のサブクラス毎に重要な記述子を与えて、サブクラスに関する説明を与える。 model-level

・機械学習による予測の評価計量と信頼度スコアの提案
・具体的課題への適用による、新しい仕組みの有効性のデモンストレーション

model-level, decision-level の説明可能性

Decision Level の説明

機械学習がその特定の decision/prediction をした理由を説明する。これまでは
知られていない化合物の予測については、知られている類似化合物を提示する。
類似度の定量化には、記述子が heterogeneous なので、Gowerの計量を用いる。
更に、大きい回帰の問題を、多クラス分類して単純な回帰問題にすることにより、
予測に至る論理的な過程も示すことができる。

Model Level の説明

記述子のうちどれが予測において重要な役割を果したか、重要な記述子がどのように
組み合わされたか、を知れば物質科学者は自動 ML システムを理解し、信頼する。
ここで提案する pipeline (regression + classification) により、単一回帰モデルの場合より
fine grained な記述子の重要性の情報を与える。
特に、記述子の重要性を2段階で示す : 1) 異なった種類の化合物を区別する
ために重要な記述子と 2) 同種類の化合物に対する回帰における重要な記述子。

複雑なシステムにおける、予測と理解の相克

Vapnikは「複雑な現実世界においては、よりよい予測をしようとすれば、理解を諦めなければならない」と主張した。この文言を真とすると、その対偶もまた真であり、それは「理解をしようとすれば、よりよい予測を諦めなければならない」ということになる。

More is different

1. 複雑な系になってもなお、その複雑系を支配する基本法則があるはず。

2. 要素が多数集まった系を支配する法則は、要素の振舞を支配する元の法則とは異なっており、そのことを創発特性という。

3. 社会まで含めた自然界は全て、“要素”と“要素の集合体”、という層が順に重なってできた階層構造を成しており、要素の集合体の振舞は創発特性として理解される。

この概念を、データ科学に活かす！