

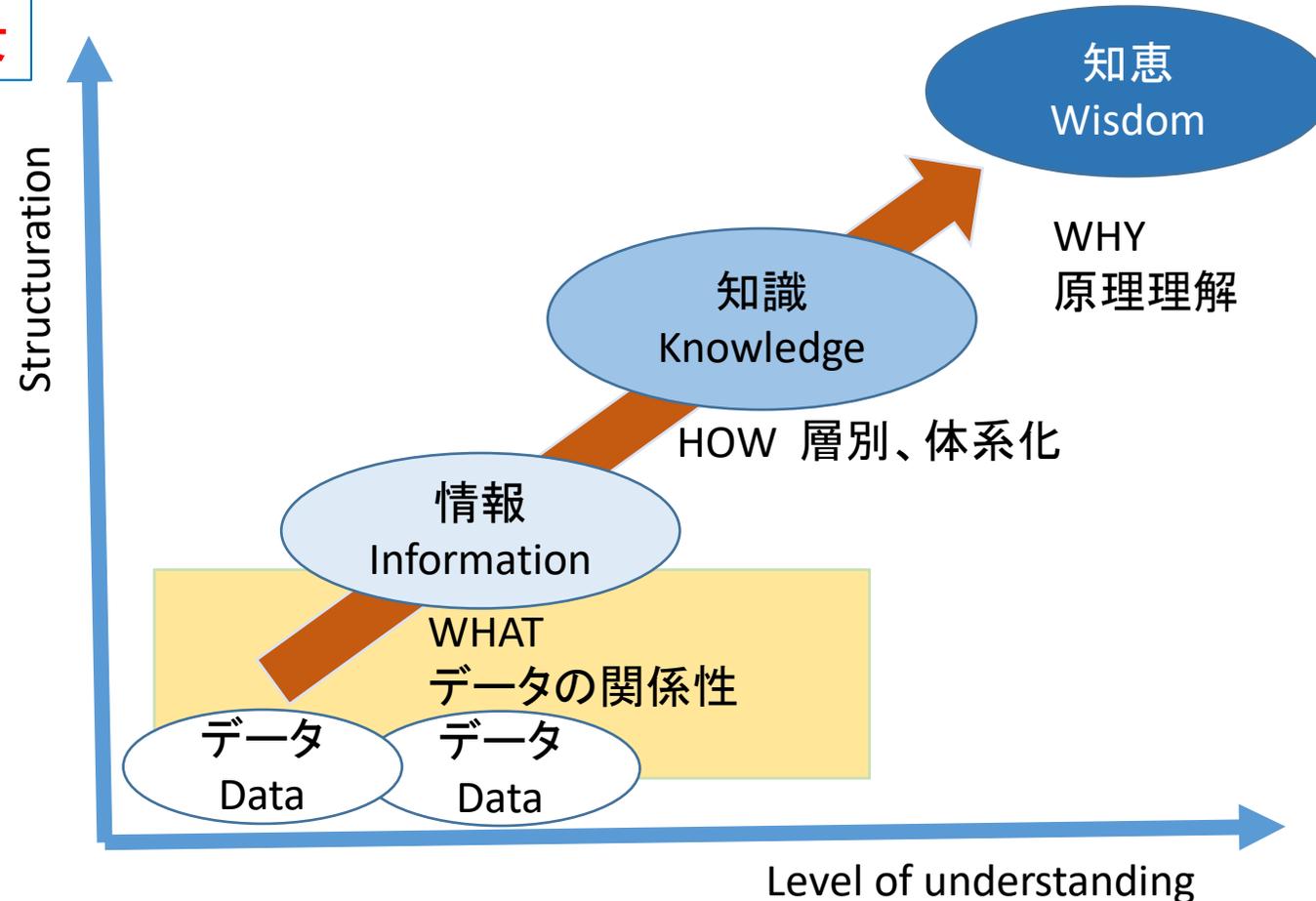
データ科学における予測と理解の両立を目指して
— 分かるとは何か？ —

寺倉清之
2018/05/21

Wisdom Hierarchy (DIKW)

Integration of Data, Information to higher levels

本日の講演会の背景



科学の発展の初期：主に観測結果からの法則の探索・推論

ケプラーの法則、メンデレーエフの周期律表、Newton力学、
万有引力、Maxwell 方程式、熱力学、統計力学、量子力学

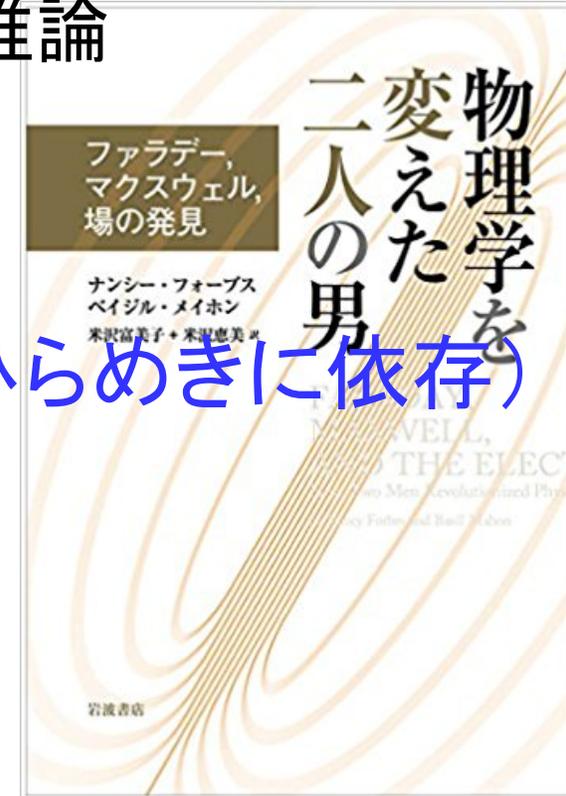
帰納的方法（+仮説的推論）の時代（天才のひらめきに依存）
abduction

原理や法則の確立

演繹的方法の時代（第3の科学は代表格）

情報技術、自動化技術の発展：**big data** 生成とその処理技術

新帰納的方法の時代（第4の科学、スマート化）



新法則(特に、因果律の根幹の法則)の発見

新奇物質発見

物質・材料開発 I

1957 BCS理論

その後、物性物理の終焉説が起こった。

P. W. Anderson の “More is different” (1972) による悲観論からの脱却

その後の新物質発見の例

1976 カーボンナノチューブ(遠藤守信)

1985 フラレン

1986 銅酸化物超伝導体

1991 カーボンナノチューブ(飯島澄男)

2005 トポロジカル絶縁体の理論提案

2006 鉄系超伝導体

2007 トポロジカル絶縁体の実験的存在確認

これらの新物質発見が
基礎的な物性発現の機構
の提案と実証という
活発な物質科学の活動
の展開に繋がっている。

これらは、経験、洞察、ひらめき、偶然、などなどによる新奇物質発見

物質・材料開発 II

科学技術の進歩に後押しされた物質開発

この場合は、目標とする物性や機能が明確なものが多い。

物質設計

物質設計は、本来逆問題なので実現が困難

- * Design principle によるスクリーニング
- * データ科学の活用 (bio-informatics, cheminformatics, materials informatics)

基本的に、これは機能最適化

発見における2つの基本的問題

1. 物質・材料開発

- ・物質・材料開発 II の機能最適化は、現在のデータ科学でもかなり対応できる。
- ・しかし、物質・材料開発 I の全く新奇な物質の開発は違う問題。

鉄系超伝導体の最初の発見と、 T_c がより高い関連物質の探索は別の問題。

細野、Parrinello,

2. 基本法則の発見

- ・Newton の運動方程式、万有引力の法則
- ・Maxwell の電磁気の方程式
- ・量子力学の基本方程式
- ・ . . .

物質・材料設計は逆問題

順問題

解析型

逆過程

逆問題

開拓型

演繹的

given

原因

結果

物質・材料

因果関係

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

帰納的

given

原因

結果

materials

相関関係

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。

回帰解析

$$\text{機能: } F = f(x_1, x_2, \dots) \quad (1)$$

$$\text{記述子: } x_1, x_2, \dots \quad (2)$$

目的の機能 F と記述子 (x_1, x_2, \dots) の間の関係式 (1) を導くだけでなく、記述子 (x_1, x_2, \dots) 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.

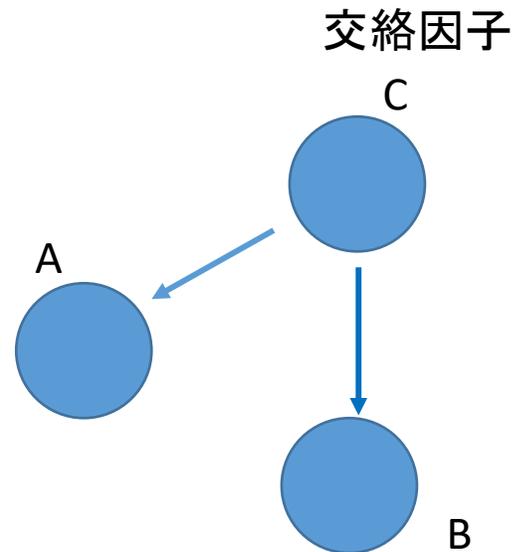


George Edward Pelham Box
(October 18, 1919 – March 28, 2013)

Prediction and understanding are generally different targets.

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

“some models are useful” の例

1. Kepler, Galileo, Newton

Kepler の師、Tycho Bráhe による 16 年間に亙る天体運航の観測データを遺産として受け取り、そこから Kepler の3法則を導いた。

Kepler が得たのは、基本的に相関関係。それで天体運航を理解したと言えるか？
本当に理解したのは、因果関係の根本まで行きついた Newton に至ってからか？

物理学とは何だろうか 上 朝永振一郎（岩波書店）

2. Faraday, Maxwell

- Faraday は自ら行った実験を注意深く観察し、深い洞察を通して、電磁誘導、電場、磁場などの場の概念まで得た。
- Faraday の結果を基に、Maxwell は電磁場の基本方程式を完成させた。

「電力と磁力の性質についてのファラデーの構想に、数学的な解釈がクラーク・マックスウェルによって与えられた今、ファラデーの同時代人には漠然としてはいはっきりしなかった彼の言葉の裏に隠されたものが、如何に正確で精度が高かったのかが分かる。最高の数学的分析が必要な多くの一般的定理を、数式を只の一つも用いず、秩序だった推論と本能的な一種の直観によって、ファラデーは見出した。これは、本当に驚異的なことだ。」 ヘルマン・フォン・ヘルムホルツ (1881)

物理学を変えた二人の男」 Nancy Forbes, Basil Mahon 原著、
米沢富美子、米沢恵美 訳（岩波書店）

補足－1

講演会の趣旨説明では、Kepler が得たものは相関関係だった、と簡単に一言で話してしまっただが、朝永先生の本を、小さい字の註まで注意深く読み直して、Kepler の解析が如何に大変なものだったかをよく理解できた。

Tycho Bráhe のデータと、Kepler の頃のドメイン知識を使って、現在の通常の機械学習の手法で、Kepler が得た相関関係を再現できるか、と問われたら、きっとできない、と答える。

それを理解してもらうために、朝永先生の本から関連の部分の要約を以下に記す。

1. まず、Kepler は火星の運行の解析をします。その際の、Kepler の解析が非常に複雑で困難であったのは、Tycho の観測データは、地球から見た太陽の方向、地球から見た火星の方向のデータだけであって、距離の情報は全くなかったから。
2. Kepler の最初の作業仮説は、火星の運動は等速円運動であるが、太陽の位置は円の中心からズレている、としたプトレマイオス流のモデルであった。
3. Kepler は数年かけてその解析を進めてほぼ完成した最後の段階で、角度にして僅か 8 分の Tycho の観測データとの誤差があることが分かった。
4. この僅かな誤差が有意の相違だと判断して、Kepler は最初の作業仮説を捨てて、新しく正しい軌道の形を求めた。いろいろと試した挙句に、長円型の軌道がうまく行くことを見つけた。

補足－2

機械学習で、Kepler の相関関係を再現できそうにない、という理由は以下のようなこと。

通常の機械学習で上の作業を代行した場合に予想されることは、観測データに伴う誤差を正規分布として扱い、Kepler の見つけた 8 分の誤差は 360 度の上にはばら撒かれると、誤差の割合は僅かに

$$8/(360 \times 60) = 0.00037$$

に過ぎないので、きっと最初の作業仮説を捨てられない、という結末が予想される。

その上でなお、Kepler は理解に至ってはいない、という理由は、Kepler が正しく面積速度一定を結論していたとは言え、角運動量の概念、ましてや角運動量の保存、力学的エネルギーの概念、エネルギー保存などは全く知らなかった、ということに因る。

- ・全く新奇な物質の発見
- ・因果律の根本の発見、

はどのようなスキームで整理されるか？

データ科学は、存在するデータ間の内挿（最近は外挿も）が得意。

外挿はできても、画期的な新物質発見は無理？

既存のデータに基づいた探索は、結局は最適化か？

物質・材料開発 I（新奇物質発見）への対応は？

直交軸を探せ（細野）

仮説的推論 (abduction) の方法論的開発

1. 演繹法

前提1: 人間であれば死ぬ。(大前提)

前提2: ソクラテスは人間である。(小前提)

結論: ソクラテスは死ぬ。(3段論法の帰結)

必然的な推論

2. 帰納法

前提1: ソクラテスは人間である。(小前提)

前提2: ソクラテスは死ぬ。(3段論法の帰結)

結論: 人間であれば死ぬ。(大前提)

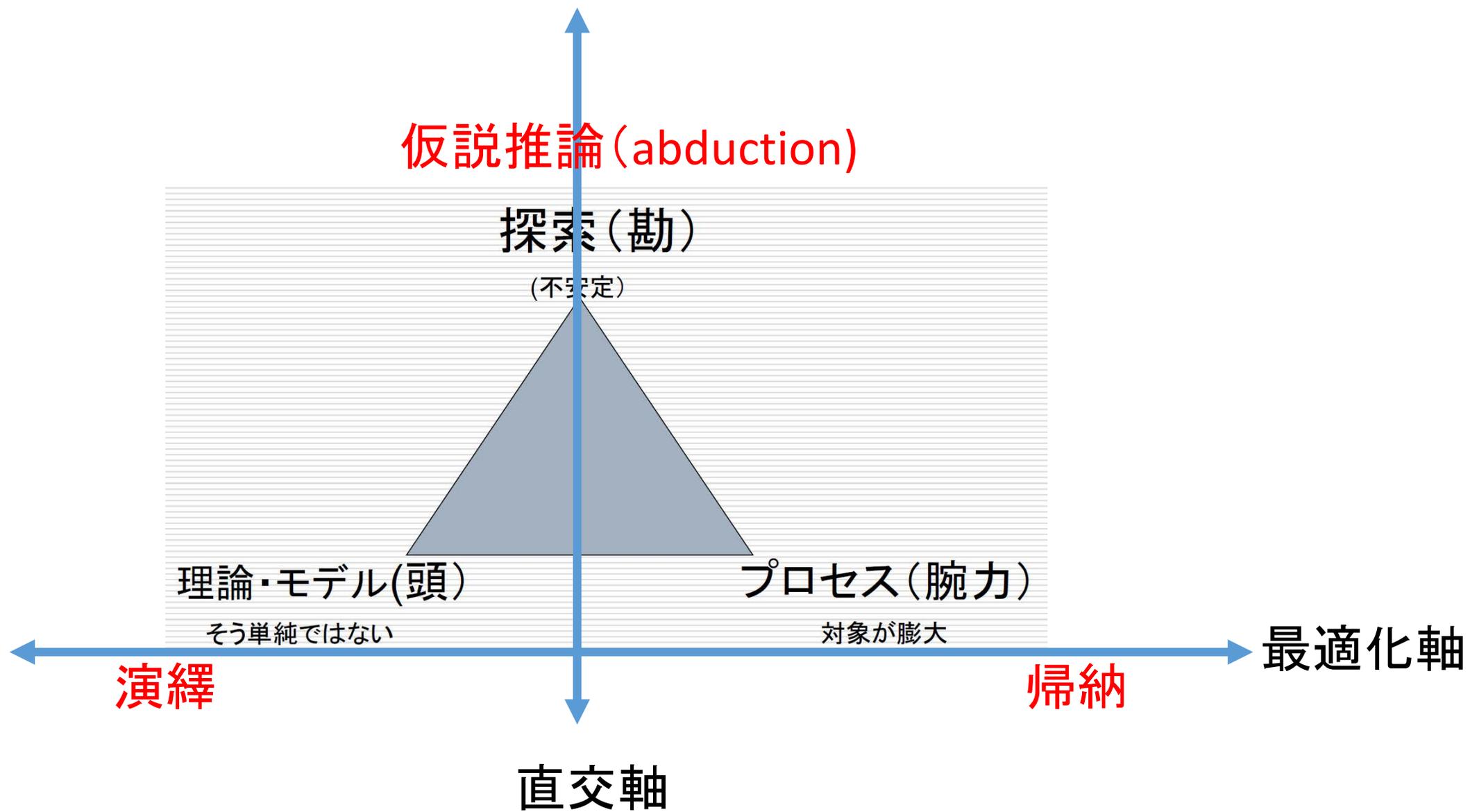
蓋然的な推論

3. アブダクション (仮説的推論)

前提1: ソクラテスは死ぬ。(3段論法の帰結)

前提2: 人間であれば死ぬ。(大前提)

結論: ソクラテスは人間である。(小前提)



予測可能性と説明可能性

単純世界と複雑世界

予測と理解

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。

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

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/>

ここでの新しい問題点

1. 説明可能性 vs. 予測可能性
2. complex world vs. simple world

More Is Different

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”の応用ということの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。

情報の縮約

フレーム問題

人工知能の急速な進歩

情報の縮約（自己符合化器、畳み込みニューラルネット）
沢山の情報から本質的要因を抽出す。情報を知識にする基本。

人工知能が比較的得意とする作業



Google の猫

フレーム問題

あるタスクを実行するのに、関係ある知識だけを取り出して使う。

人工知能が不得意とする作業



飽きずに総当たりをしてくれるというありがたい側面は、
一方では不要なことに時間と労力を使うことになる。

理解する、あるいは分かる、ということには、種々の側面がある。

理解する、あるいは分かるは、知的好奇心として重要、
という一般的な意義もあるが、
科学的新発見の基盤、
とも思われる。