2025 年 6 月 1 日(日)

# データ駆動科学の三つのレベル

東京大学　大学院新領域創成科学研究科　複雑理工学専攻
教授　岡田真人

この解説では、データ解析に階層性が存在することを指摘し、それに基づきデータ駆動科学の三つのレベルを提唱する[1]。

その階層性の具体例は、中学二年生で学ぶ連立方程式とその応用問題である。図 1 に示すように、鶴亀算や食塩水の問題などの一見全く違った問題が、連立方程式の枠組みで統一的に議論できることは、中学二年生で学ぶ。

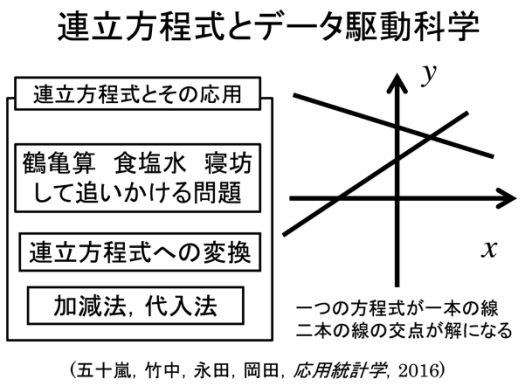



図 1 連立方程式とデータ駆動科学

これらの実世界の複雑な問題は、図 1 のように、三つのレベル(階層)に分離することができる。一番上のレベルは問題を自然言語で記述するレベルである。次に、その自然言語の記述を連立方程式を用いて数理モデル化するモデリングのレベルである。最後は、連立方程式を加減法や代入法で解くアルゴリズムのレベルである[2]。

このように鶴亀算や食塩水の問題などのデータ解析の問題は、すべて上述の三つのレベルで記述されるという立場がある。これを明確に述べたのが脳神経科学者であり人工知能研究者の David Marr (1945-1980) である。David Marr は彼の遺作の Vision[3]で、複雑な情報処理機械を理解するためには、図 2 に示す David Marr の提唱する David Marr の三つのレベルによるアプローチが必要であると述べた。一番上のレベルは情報処理の目的を自然言語で記述する計算理論レベルである。次に、その自然言語の記述を数理モデル化し、その数理的な問題を解くことを考える表現とアルゴリズムのレ

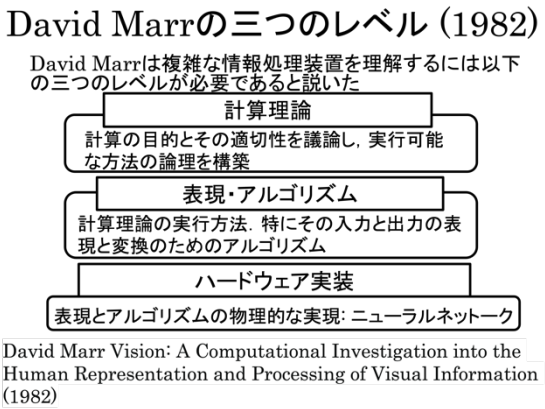



図 2 David Marr の三つのレベル

ベルである。最後は、アルゴリズムのハードウエア実装を取り扱うハードウェア実装のレベルである。

前述の連立方程式の例では、表現とアルゴリズムを別のレベルに割り当てたが、David Marr は脳の機能の理解することを主眼にしていたので、脳における機能のハードウェア実装を意識するために、数理的な枠組みある表現とアルゴリズムを同じレベルに置き、それを表現とアルゴリズムのレベルとした。

我々が図 3 に示す文部科学省科学研究費補助金「新学術領域研究」平成 25 年度～29 年度スパースモデリングの深化と高次元データ駆動科学の創成（略称 疎性モデリング Initiative for High Dimensional Date Driven Science through Deepening Sparse Modeling, http://sparse-modeling.jp/)で提案したデータ駆動科学の三つのレベルである(図 4)。我々は、心理学/行動科学や工学を含む自然科学全般でのデータ解析は、David Marr の三つのレベルを参照した、図 4 のデータ駆動科学の三つのレベルで階層的に行う必要があることを述べた。我々は疎性モデリングのヒアリングにおいて、図 5 に示すように、疎性モデリングの目標は日本と世界に先立ちデータ駆動科学のコアを形成するとしていた。そのためのフレームワークが図 4 のデータ駆動科学の三つのレベルである。図 6 に示すように、データ駆動科学の三つのレベルは前述の連立方程式の三つのレベルと完全に対応している。

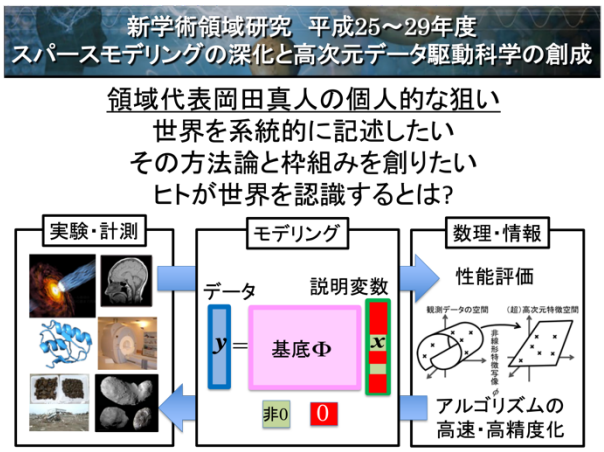
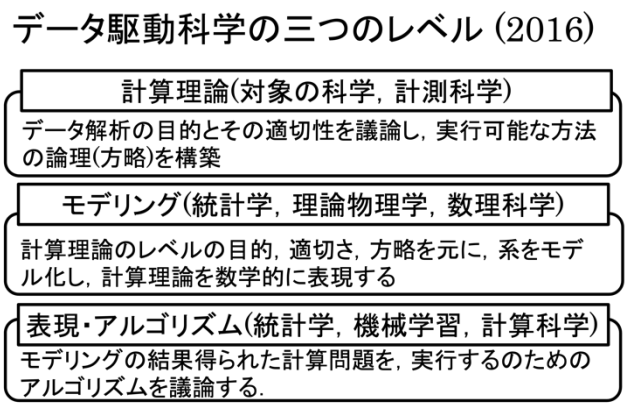



図 3 スパースモデリングの深化と高次元データ駆動科学の創成




Igarashi, Nagata, Kuwatani, Omori, Nakanishi-Ohno and M. Okada, “Three Levels of Data-Driven Science”, *Journal of Physics: Conference Series*, 699, 012001, 2016.

図 4 データ駆動科学の三つのレベル

この図 6 の対応から、データ駆動科学の習得の戦略が見えてくる。我々が連立方程式の応用問題を解く際には、最初に鶴亀算や食塩水の問題に取り掛からない。まずは、連立方程式を定義し、それを解くためのアルゴリズムである、加減法や代入法を習得する。その後に、そこで得られた知見をもとに、鶴亀算

や食塩水の問題の文章を解析し、問題文が必ず二つの部分から構成されていることに気づく。そして、それら二つの部分が、それぞれの方程式で数理的に表現できることに気づき、そこで、自然言語での記述をあえて忘れて、加減法や代入法のアルゴリズムを用いて問題を解く。ここから、心理学/行動科学や工学を含む自然科学全般でのデータ解析でも、連立方程式のような何らかの数理的な表現が存在し、それを解くためのアルゴリズムが存在するはずである。これが、ベイズ推論とスパースモデリングである。

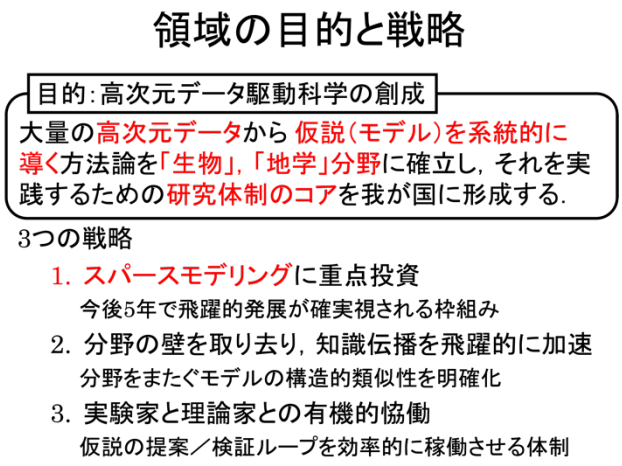



図 5 領域の目標と戦略

つまり、まずデータ解析を学ぶには、自分の取り扱う分野のことを一旦忘れて、ベイズ推論とスパースモデリングを学ぶ。そして、そこで得た知見から、データ解析の計算理論を見つめ直して、ベイズ推論とスパースモデリングで数理的に定式化し、解くのである。単刀直入にいうと、問題をデータ駆動科学の三つのレベルの鋳型に押し込んで取り扱うのである。

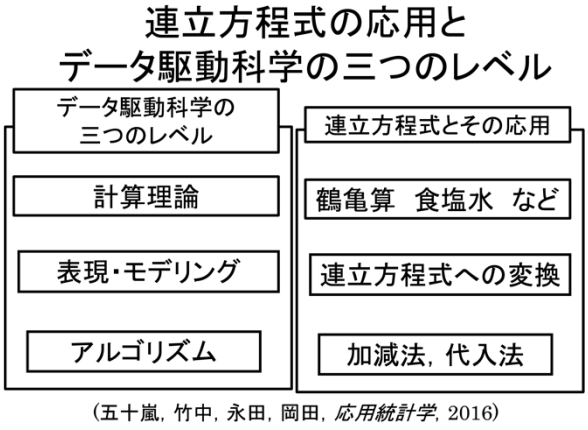



図 6 連立方程式の応用とデータ駆動科学

このようなトレーニングを積むことで、分野によらない普遍的なデータ解析のスキルを身につけることできる。つまり、データ駆動科学の三つのレベルは、ともすれば属人的な能力とみなされていた、多くの分野を普遍的位に取り扱う能力を、誰もが習得できる新しいパラダイムなのである。