データ駆動科学の三つのレベル

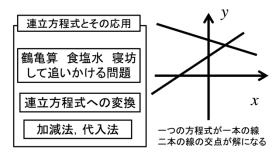
東京大学 大学院新領域創成科学研究科 複雑理工学専攻 教授 岡田真人

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。

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

連立方程式とデータ駆動科学

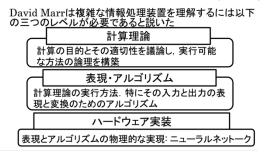


(五十嵐, 竹中, 永田, 岡田, 応用統計学, 2016)

離することができる。一番上のレベ 図 1 連立方程式とデータ駆動科学 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

David Marrの三つのレベル (1982)



David Marr Vision: A Computational Investigation into the Human Representation and Processing of Visual Information (1982)

図 2 David Marr の三つのレベル

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

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いたが、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://sparsemodeling.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

新学術領域研究 平成25~29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

領域代表岡田真人の個人的な狙い 世界を系統的に記述したい その方法論と枠組みを創りたい ヒトが世界を認識するとは?

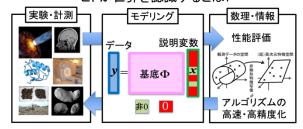


図 3 スパースモデリングの深化と高次元 データ駆動科学の創成

データ駆動科学の三つのレベル (2016)

計算理論(対象の科学, 計測科学)

データ解析の目的とその適切性を議論し、実行可能な方法の論理(方略)を構築

モデリング(統計学,理論物理学,数理科学)

計算理論のレベルの目的,適切さ,方略を元に,系をモデル化し,計算理論を数学的に表現する

【表現・アルゴリズム(統計学,機械学習,計算科学) モデリングの結果得られた計算問題を,実行するのための アルゴリズムを議論する.

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

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。

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、自然をあえて忘れて、自然をあえて忘れて、加減とでの記述をあえて忘れて、加減とでの記述をあえて立りが、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

領域の目的と戦略

目的:高次元データ駆動科学の創成

大量の高次元データから 仮説(モデル)を系統的に 導く方法論を「生物」,「地学」分野に確立し、それを実 践するための研究体制のコアを我が国に形成する.

3つの戦略

- 1. スパースモデリングに重点投資 今後5年で飛躍的発展が確実視される枠組み
- 2. 分野の壁を取り去り、知識伝播を飛躍的に加速分野をまたぐモデルの構造的類似性を明確化
- 3. 実験家と理論家との有機的恊働 仮説の提案/検証ループを効率的に稼働させる体制

図 5 領域の目標と戦略

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。

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、ベイズ理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。

このようなトレーニングを積むこ

 データ駆動科学の 三つのレベル
 連立方程式とその応用

 計算理論
 鶴亀算 食塩水 など

 表現・モデリング
 連立方程式への変換

 アルゴリズム
 加減法、代入法

連立方程式の応用と

(五十嵐, 竹中, 永田, 岡田, 応用統計学, 2016)

図 6 連立方程式の応用とデータ駆動科学

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。