Proceedings of the Twenty-Sixth RAMP Symposium
Hosei University, Tokyo, October 16-17, 2014

機械学習の三つのレベルとデータ駆動型科学

Three Levels of machine learning and data-driven science

岡田 真人[1*]

Masato Okada

概要　本講演では文部科学省科学研究費補助金新学術領域研究「スパースモデリングの深化と高次元データ駆動科学」(平成 25 年度〜29 年度) の紹介を行なう．このプロジェクトの目的は，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し，高次元データ駆動科学を創成することである．その目的を達成するために，以下の三つの戦略を用いる．(1) 今後 5 年で飛躍的発展が確実視される枠組みであるスパースモデリングに重点投資し，(2) 分野をまたぐモデルの構造的類似性を明確化することで，分野の壁を取り去り，知識伝播を飛躍的に加速する．(3) 実験家と理論家が有機的に協働することで，仮説の提案／検証ループを効率的に稼働させる体制の規範モデルを確立する．

このプロジェクトを遂行する上で，スパースモデリング，さらにそれを含む機械学習について，脳科学およびコンピュータビジョンに大きな影響を与えた David Marr の三つのレベルの視点が重要であり，これがデータ駆動科学にとって必要不可欠な基本概念であることを説明する．

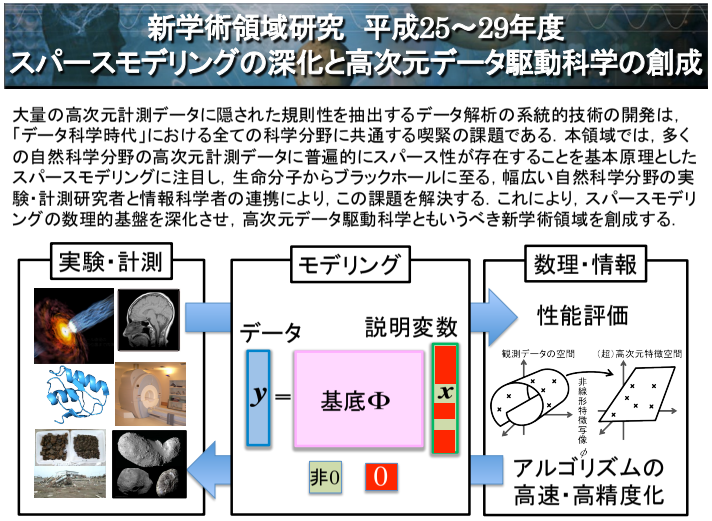
キーワード　スパースモデリング，データ駆動型科学，機械学習

1　東京大学大学院 新領域創成科学研究科 複雑理工学専攻, 〒277-8561 千葉県柏市柏の葉 5-1-5 東大基盤棟 701
Department of Complexity Science and Engineering, the University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa-shi, Chiba-ken 277-8561, Japan

∗ E-mail address: okada@k.u-tokyo.ac.jp

機械学習の三つのレベルと データ駆動科学

東京大学・大学院新領域創成科学研究科
岡田真人



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協働
   仮説の提案／検証ループを効率的に稼働させる体制

圧縮センシングによるブラックホール(BH)の直接撮像　A02-3:天文学班(本間)

おとめ座銀河団の
巨大楕円銀河　M87の銀河中心核

BHからの電波
の検出に成功

(*Science*, Online September 27 2012)

スパースモデリング
高次元データからの効率的情報抽出

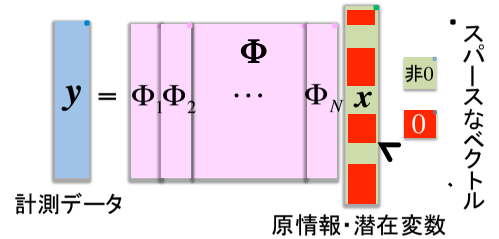
直感的な理解，仮説の提案，
潜在構造の推定が困難

多くの分野の共通課題
攻略すべきボトルネック

第一原理からの演繹が難し
い，生物学，地学に顕著

圧縮センシング
スパース性でサンプリング定理を超える

線形計測：ブラックホール，MRI，NMR・・・

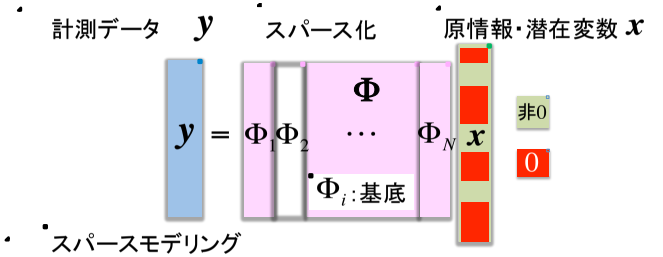


変数の個数が，式の数より多い ⇒ 解が求まらない

スパースなベクトル

変数の要素に0が多いベクトル．

圧縮センシング
スパース性でサンプリング定理を超える



潜在変数がスパース(0が多い)
0の場所を推定しながら，方程式を解く

$$E(x)=\left\|y-\sum_i \Phi_i x_i\right\|^2+\lambda\sum_i |x_i| \qquad \Phi_i：基底$$

データの再構成　スパースな変数

圧縮センシングによるMRIの高速撮像
A01-1:医学班(富樫)

元画像　　スパースモデリング　　従来法

データ獲得時間を短縮
⇒予防医療に革新，患者負担の軽減

MRIが実用化されて30年
装置の改良でデータ獲得時間5分の1

データ科学の勃興(Jim Gray, 1944-2012)
Fourth paradigm

第1の時代： 経験科学
(数千年前～，アリストテレス)

↓

第2の時代： 理論科学
(数百年前～，ライプニッツ)

↓

第3の時代： 計算科学
(数十年前～，フォン・ノイマン)

↓

第4の時代：データ科学

[*Science*, Feb. 2011]

天文学における高次元データ解析手法が，全く対象とスケールの異なる生命科学でも有効に働く

[*Science*, Feb. 2011]

多様な視点の導入による革新的展開
普遍的な視点による分野を越えたアナロジー／普遍性への探究心
普遍的な原理にもとづく新しい解析法の発展

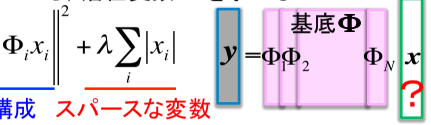
計測から潜在構造推定へ
基底選択から基底学習へ

基底選択

データ $\boldsymbol{y}$ と基底関数 Φ から，潜在変数 $\boldsymbol{x}$ を求める

$$E(\boldsymbol{x}) = \left\| \boldsymbol{y} - \sum_i \Phi_i x_i \right\|^2 + \lambda \sum_i |x_i|$$

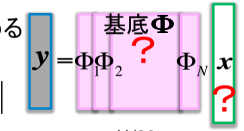
データの再構成　スパースな変数



基底学習

変数 $\boldsymbol{x}$ だけでなく，基底 Φ も同時に求める

$$E(\Phi, \boldsymbol{x}) = \left\| \boldsymbol{y} - \sum_i \Phi_i x_i \right\|^2 + \lambda \sum_i |x_i|$$



11/39

基底学習： データから潜在構造を知る
B01-2:スパースモデリング班(岡田)

$$E(\Phi, \boldsymbol{x}) = \left\| \boldsymbol{y} - \sum_i \Phi_i x_i \right\|^2 + \lambda \sum_i |x_i|$$

幅広い生物・地学分野の喫緊のテーマ
各分野のフラッグシップを選定

- A01-1: 医学班(富樫・京大)
  新たな診断・治療の実現
- A02-1: 地球科学班(駒井・東北大)
  津波防災対策への提言
- A01-2: 生命科学班(木川・理研)
  タンパク科学の質的変化
- A02-2: 惑星科学班(宮本・東大)
  次世代探査戦略の創出
- A01-3: 脳科学班(谷藤・理研)
  モノを見分ける脳のしくみ
- A02-3: 天文学班(本間・国立天文台)
  ブラックホールの直接撮像

スパースモデリングの有用性が確実視できる題材を選定
⇒　これらを起爆剤に公募研究・周辺分野に成果を波及

ケプラーの法則

$$T^2 \propto R^3$$

大きな半径(R)を持つ惑星ほど
ゆっくり(T)太陽の周りを回る

古典力学を規範として

$$T^2 \propto R^3$$

ティコ・ブラーエ
(1546-1601)

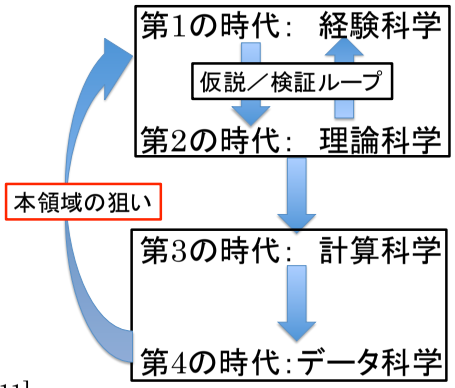
ヨハネス・ケプラー
(1571-1630)

多量のデータ ⟶ 少数の説明変数 ⟶ 現象論
スパース性

科学の出発点は全てデータ駆動科学

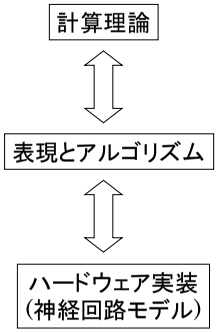
- 科学:データを少数の説明変数で記述する.
- 天体観測の結果を分析し，思索を重ねることで得られた古典力学はこの模範例である.

データ科学の勃興（Jim Gray）



[*Science*, Feb. 2011]

計算論的神経科学
Computational Neuroscience



David Marr (1945-1980)　　Vision (1982)

機械学習の三つのレベル

David Marrは複雑な情報処理装置を理解するには以下の三つのレベルが必要であると説いた

計算理論
情報処理(データ解析)の目標，方略，適切さ

表現とアルゴリズム(狭義機械学習)
計算理論の表現(ベイズ推論)と，そのアルゴリズム

ハードウェア実装
アルゴリズムがどのように物理的に実現されるか

応用分野の存在が生命線の機械学習においては
応用分野(物質材料科学，地球惑星科学など)の知見に基づく計算理論の構築が必須であり，それをどのようにスパースモデリングやベイズ推論に基づき表現するかが重要

方略 ⟷ 表現

2014年12月14日～17日
公開シンポジウム

- チュートリアル　2014年12月14日(日)
- 成果報告会　12月15日(月)～17日(水)
- 会場：東工大すずかけ台キャンパス　大学会館多目的ホール