学融合とデータ駆動科学

東京大学・大学院新領域創成科学研究科
基盤科学系　複雑理工学専攻
岡田真人

日時:2025年4月16日(水) 16:50-18:35

概要

- 学融合を実践する肝は、融合する分野の共通基盤を明確にすることである
- 今日の学融合セミナーのテーマは「データ駆動」であるので、その共通基盤として、科学/工学を系統的に取り扱うことが可能なデーター駆動科学を紹介し、データー駆動科学を用いた学融合の試みを紹介する。

内容

- 自己紹介
- 学融合の目的と実践
  - データ駆動科学の三つのレベル
  - ベイズ的スペクトル分解を例にして
- 可解線形回帰モデル
- SPring-8全ビームラインベイズ化計画
- まとめと今後の展開
- 集中講義の紹介

自己紹介

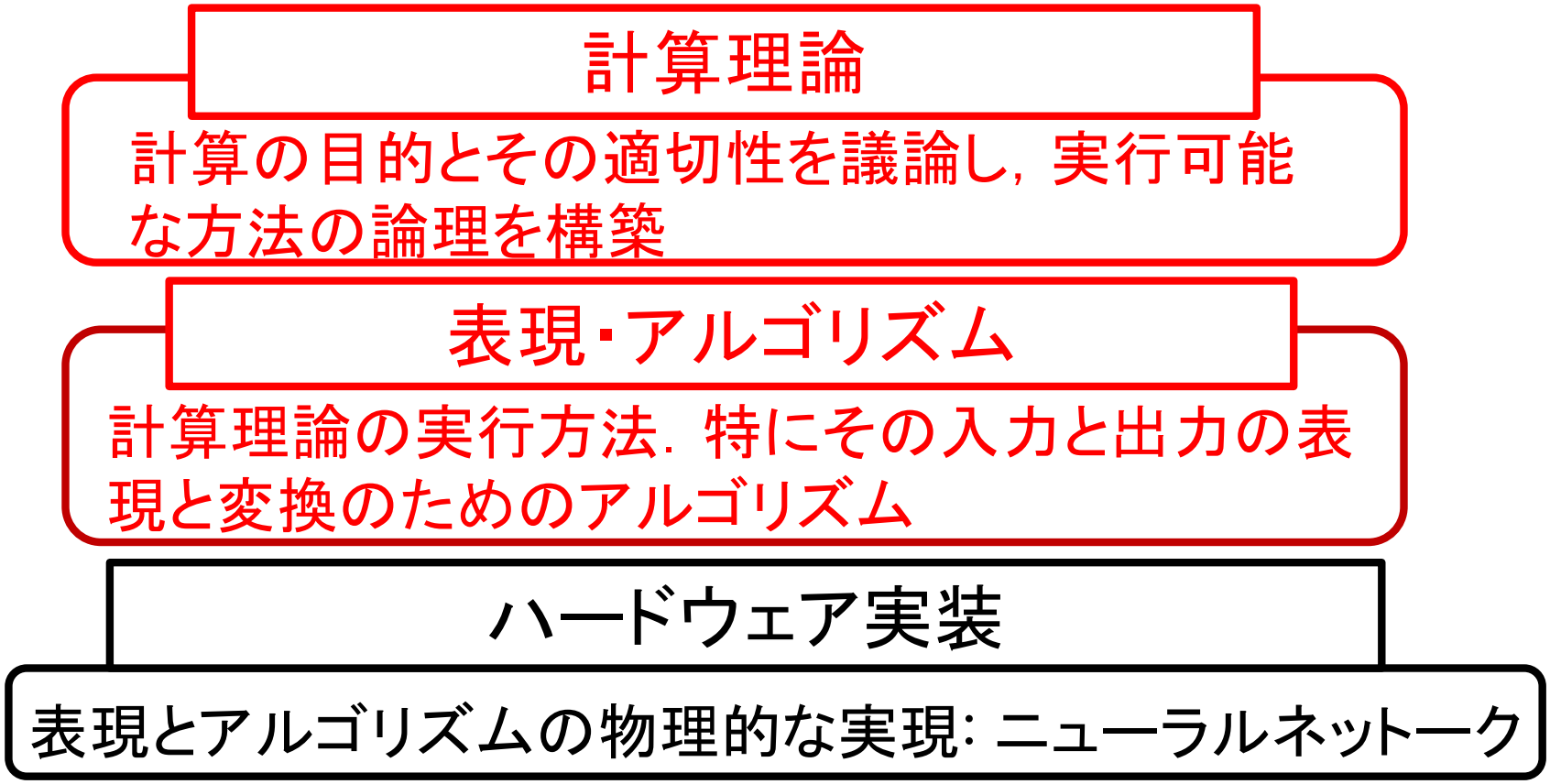
- 大阪市立大学理学部物理学科 (1981 - 1985)
  - アモルファスシリンコンの成長と構造解析
- 大阪大学大学院理学研究科(金森研) (1985 – 1987)
  - 希土類元素の光励起スペクトルの理論
- 三菱電機 (1987 - 1989)
  - 化合物半導体(半導体レーザー)の結晶成長
- 大阪大学大学院基礎工学研究科生物工学(福島研) (1989 - 1996)
  - 畳み込み深層ニューラルネット
  - 情報統計力学(ベイズ推論と統計力学の数理的等価性)
- JST ERATO 川人学習動態脳プロジェクト (1996 - 2001)
  - 計算論的神経科学
- 理化学研究所 脳科学総合研究センタ (甘利T) (2001 - 04/06)
  - ベイズ推論，機械学習，データ駆動型科学
- 東京大学・大学院新領域創成科学研究科 複雑理工学専攻
  - データ駆動科学 (2004/07 –)

内容

- 自己紹介
- 学融合の目的と実践
  - データ駆動科学の三つのレベル
  - ベイズ的スペクトル分解を例にして
- 可解線形回帰モデル
- SPring-8全ビームラインベイズ化計画
- まとめと今後の展開
- 集中講義の紹介

David Marrの三つのレベル (1982)

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



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

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)

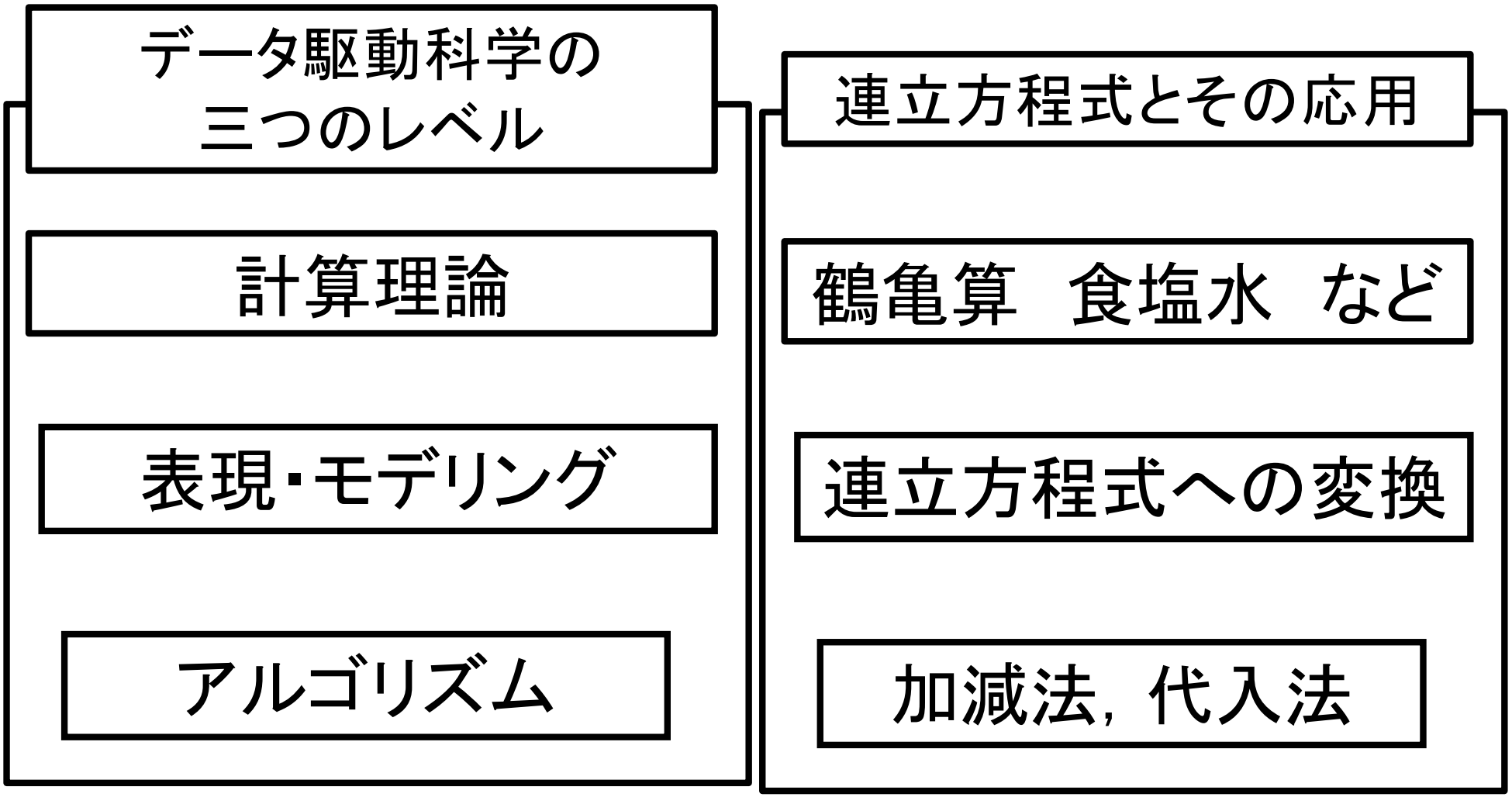
計算理論のレベルの目的，適切さ，方略を元に，系をモデル化し，計算理論を数学的に表現する

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.

連立方程式の応用と
データ駆動科学の三つのレベル



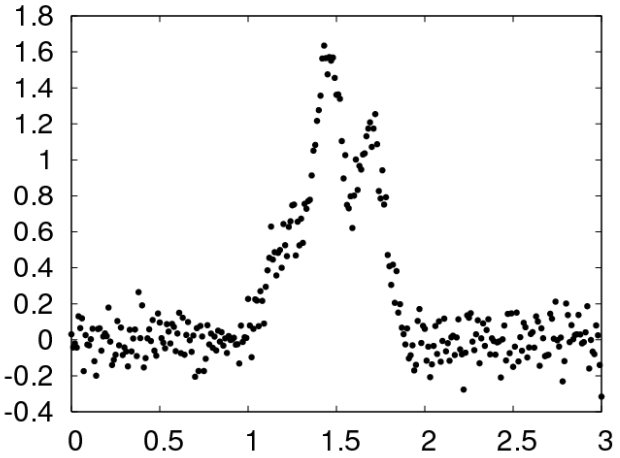
（五十嵐，竹中，永田，岡田，*応用統計学*，2016）

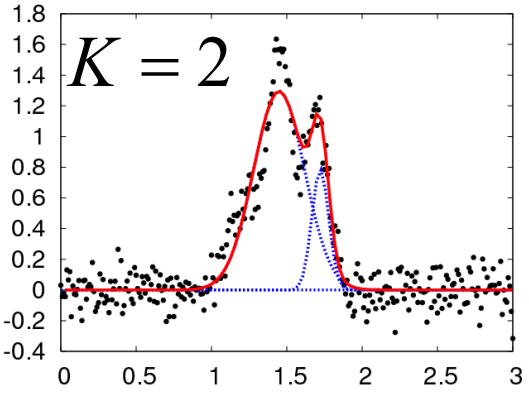
スペクトル分解

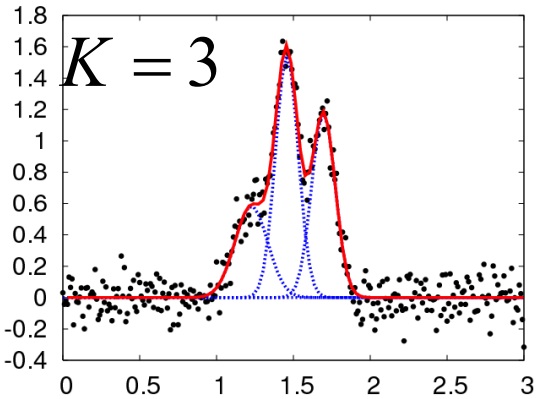
永田賢二，杉田誠司，岡田真人
東大新領域

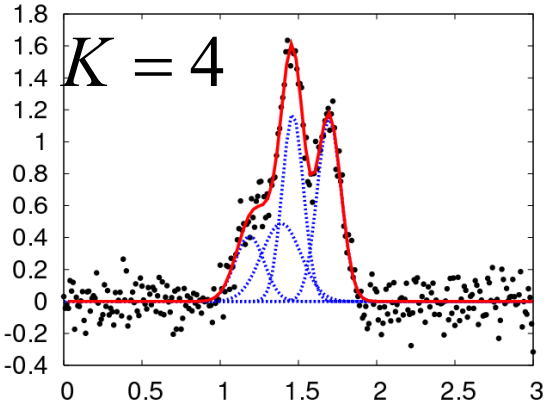
Kenji Nagata, Seiji Sugita and Masato Okada, "Bayesian spectral deconvolution with the exchange Monte Carlo method", *Neural Networks*, **28**, 82-89 (2012)

ベイズ的スペクトル分解









Nagata, Sugita and Okada, Bayesian spectral deconvolution with the exchange Monte Carlo method, *Neural Network*s 2012

スペクトル分解の三つのレベル (1/2)

スペクトル分解の計算理論

データ解析の目的: 多峰性スペクトルから背後にある離散電子のエネルギー準位を推定する

データ解析の適切さ: 多峰性スペクトルを単峰性関数の線形和で表し、その単峰性関数の個数を推定し、その単峰性関数の中心位置を電子のエネルギー準位とする
統計学の交差検証誤差やベイズ　的モデル選択で単峰性関数の数を決める。

スペクトル分解のモデリング

多峰性スペクトルを単峰性関数の線形和に観測ノイズが付加されれて生成されるとモデリングする

スペクトル分解の定式化

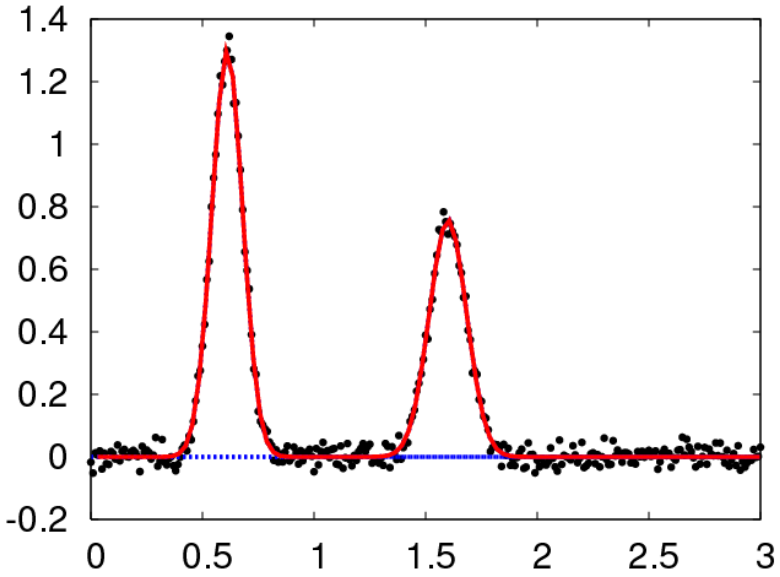
ガウス関数(基底関数)の足し合わせにより，スペクトルデータを近似

観測データ：$D = \{x_i, y_i\}_{i=1}^{n}$

x_i：入力　y_i：出力

$$f(x;\theta) = \sum_{k=1}^{K} a_k \exp\left(-\frac{b_k\left(x-\mu_k\right)^2}{2}\right)$$

$$\theta = \{a_k, b_k, \mu_k\} \quad k = 1, \cdots, K$$



二乗誤差を最小にするようにパラメータをフィット（最小二乗法）

$$E(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - f\left(x_i;\theta\right)\right)^2$$

スペクトル分解の三つのレベル (2/2)

スペクトル分解の表現・アルゴリズム

多峰性スペクトルを単峰性関数の線形和に観測ノイズが付加されれて生成するとモデリングし、ベイズ推論を適用することで、K個の単峰性関数の大きさ、位置、幅。大きさの事後確率を求める。各Kに対して、ベイズ的自由エネルギーを求め、ベイズ的自由エネルギーを最小にするKを求める。そのK個の単峰性関数の位置を、電子のエネルギー準位とする。

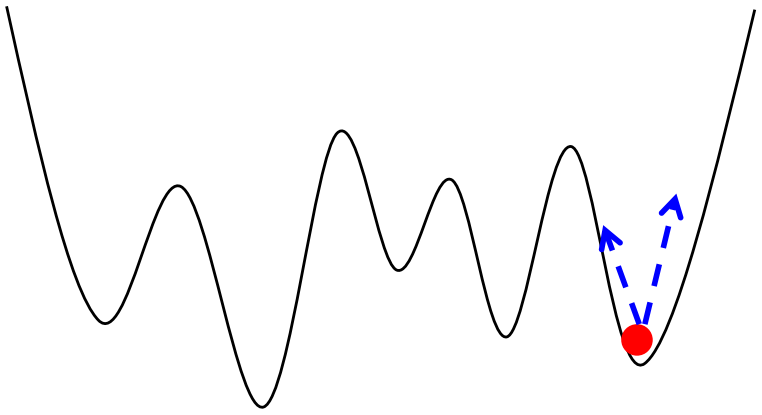
Igarashi, Nagata, Kuwatani, Omori, Nakanishi-Ohno and M. Okada, “Three Levels of Data-Driven Science”, *Journal of Physics: Conference Series*, 699, 012001, 2016.

Nagata, Sugita and M. Okada, “Bayesian spectral deconvolution with the exchange Monte Carlo method”, *Neural Networks*, 28, 82-89 2012.

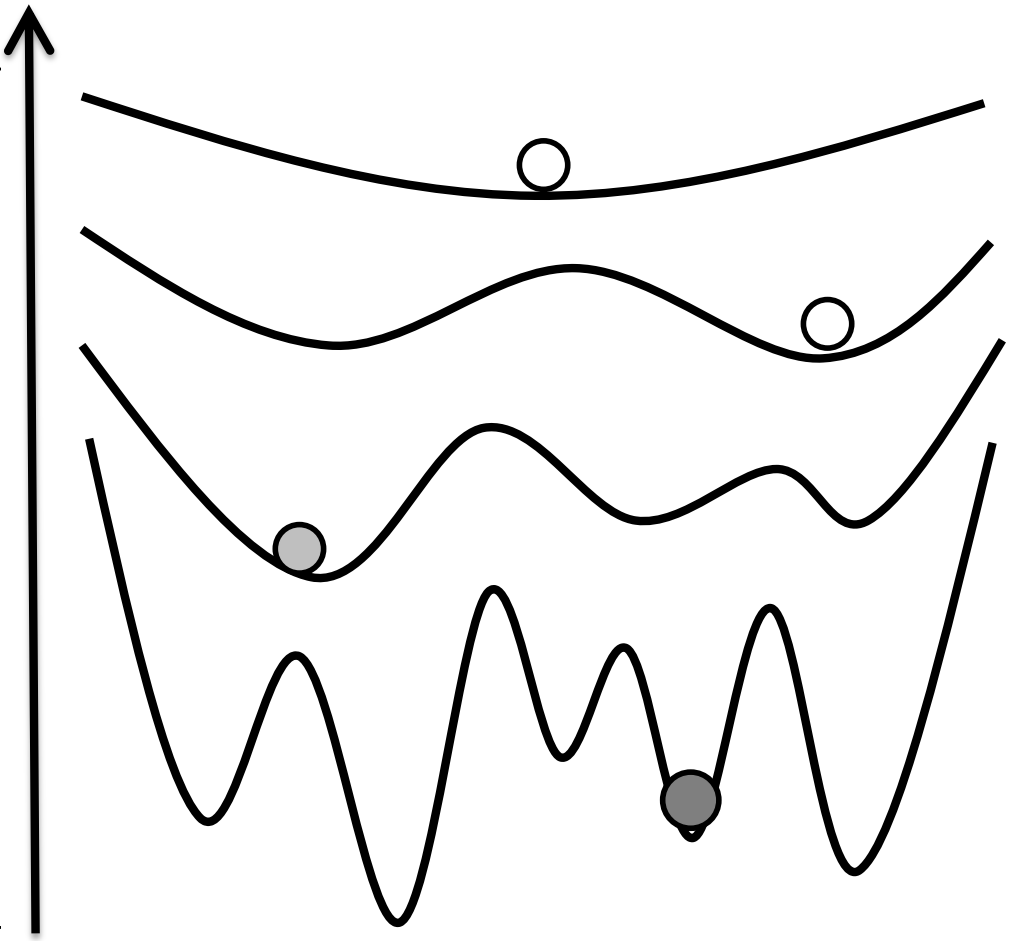
モンテカルロ法の適用
レプリカ交換モンテカルロ法

メトロポリス法

$$p_\beta(\theta) \propto \exp\left(-\frac{n}{\sigma^2}\beta E(\theta)\right) p(\theta)$$

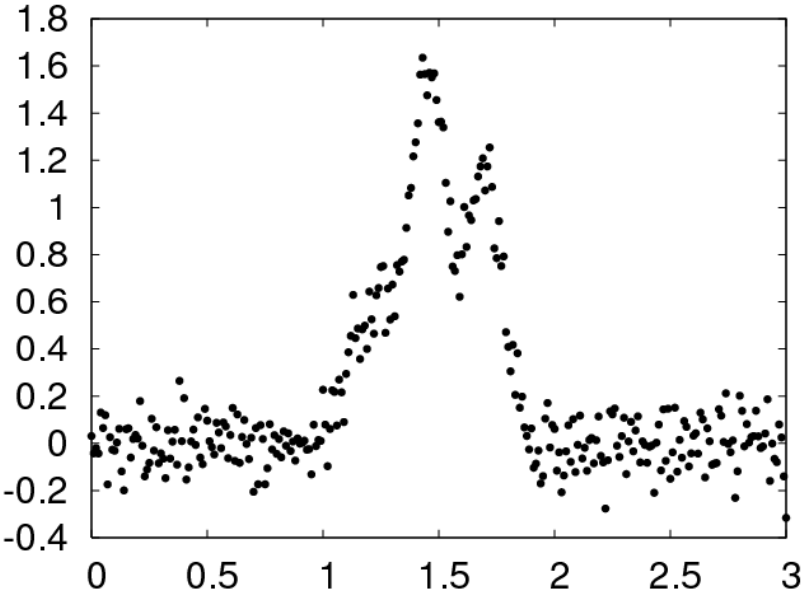


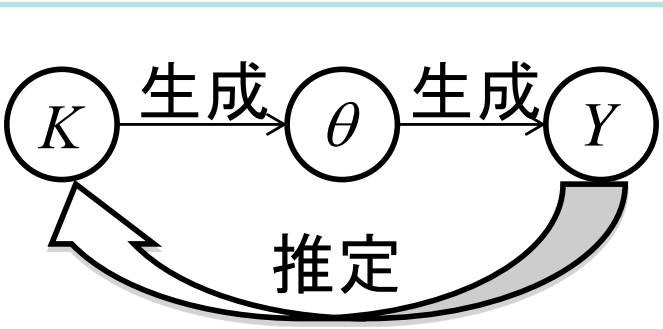
レプリカ交換モンテカルロ法



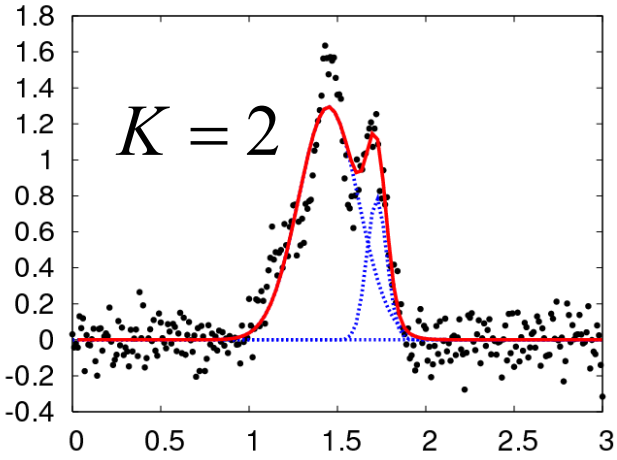
K. Hukushima, K. Nemoto, *J. Phys. Soc. Jpn.* **65** (1996).

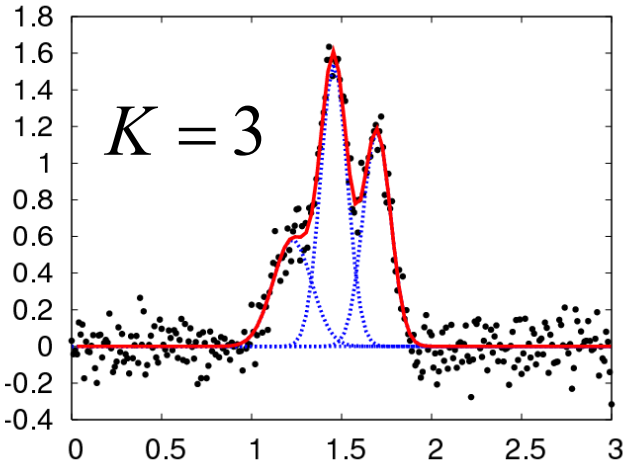
モデル選択: Kをどう選ぶか

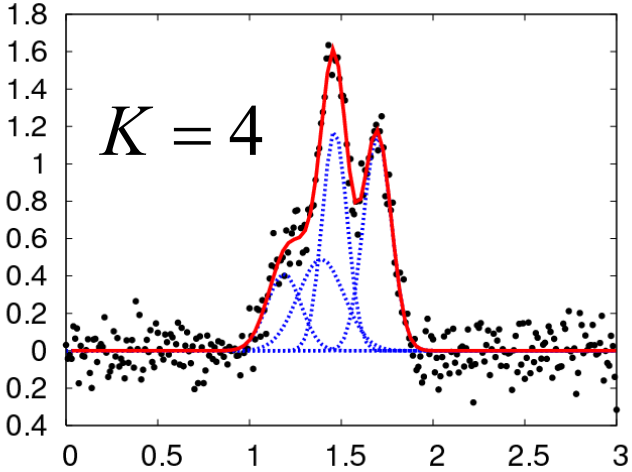




K :ピーク個数
θ :ピーク位置など
Y :観測スペクトル







Nagata, Sugita and Okada, Bayesian spectral deconvolution with the exchange Monte Carlo method, *Neural Networks* 2012

モデル選択: 自由エネルギーの導入

1. 欲しいのは　$p(K|Y)$

2. θがないぞ

3. $p(K,\theta,Y)$ の存在を仮定

$$p(K,\theta,Y) = p(Y|\theta,K)p(K)$$

$$p(Y|\theta,K) = \prod_{i=1}^{n} p(y_i|\theta) \propto \exp(-nE(\theta))$$

K:ピーク個数
θ:ピーク位置など
Y:観測スペクトル

4. 無駄な自由度の系統的消去:周辺化，分配関数

$$p(K,Y) = \int p(K,\theta,Y)d\theta$$

$$p(K|Y) = \frac{p(Y|K)p(K)}{p(Y)} \propto p(K)\int \exp(-nE(\theta))p(\theta)d\theta$$

$$F(K) = -\log\int \exp(-nE(\theta))p(\theta)d\theta = \boxed{E-TS}$$

自由エネルギーを最小にする個数　Kを求める.

スペクトル分解のハードウェア実装

Table 1. Execution times [seconds] of different algorithms and computing implementations.

Algorithm \ Computing	CPU	GPU
MCMC method	10371.5 sec	1326.9 sec
SVI metohod	260.8 sec	7.2 sec

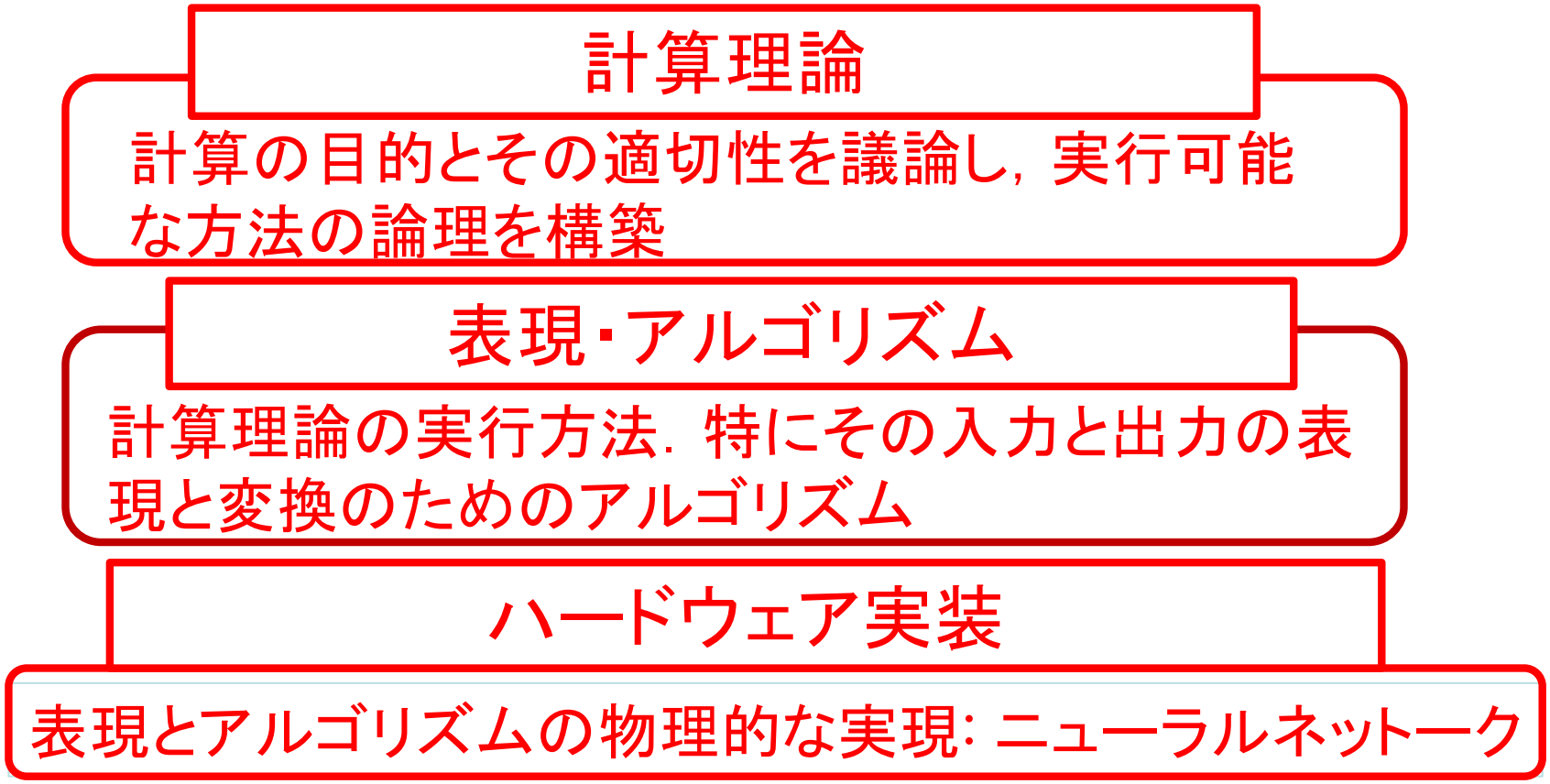
260.9/7.2=36.1 times faster

Rapid, Comprehensive Search of Crystalline Phases from X-ray Diffraction in Seconds via GPU-Accelerated Sparse Inference

Murakami, Nagata, Matsushita and Demura[1]
National Institute for Materials Science: NIMS

David Marrの三つのレベル (1982)

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



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

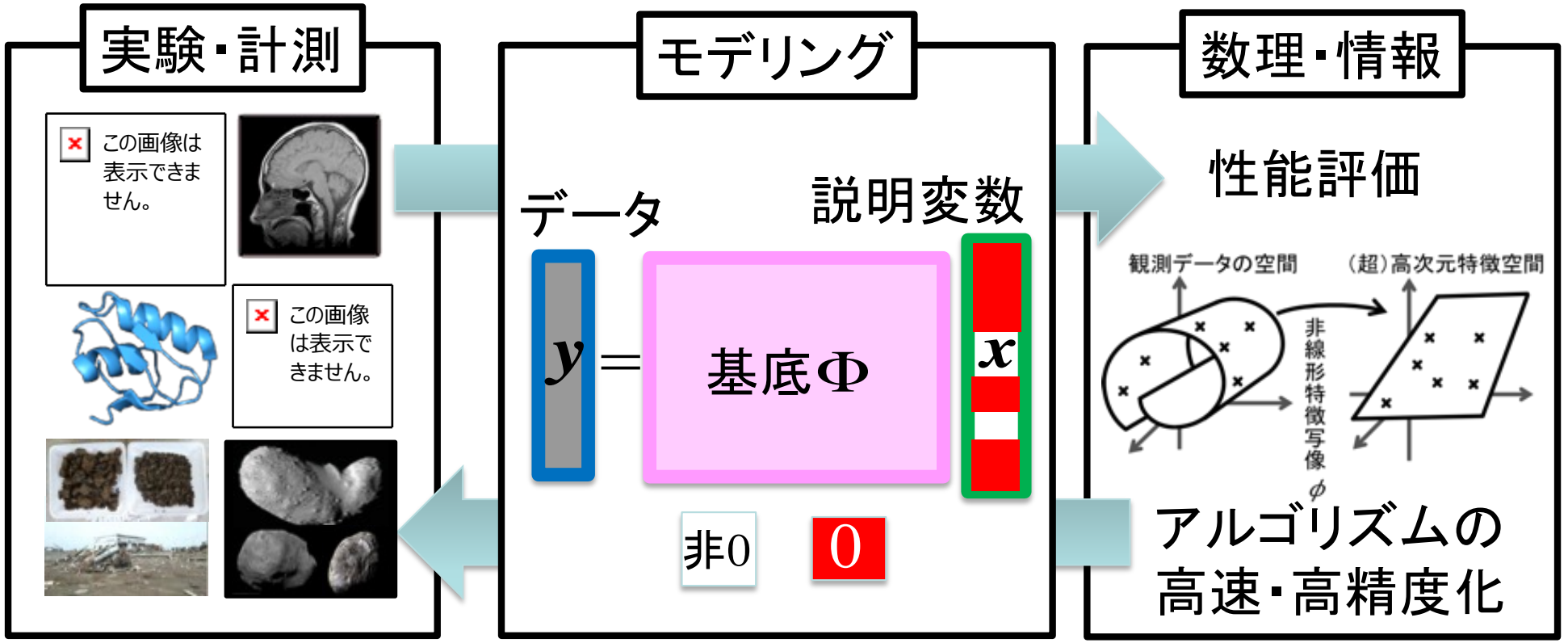
異分野共創でどのような分野のブレークスルーを期待できるか?

- 計算理論
  - データの生成元の分野、計測科学
    - その分野の限界の突破と計測科学の刷新
- モデリング/アルゴリズム
  - 物理学、数理科学、統計学
    - モデリング手法の深化
  - 深層ネットの数理(特異統計学)と計測科学
- ハードウェア実装
  - 計算機科学: 新規アルゴリズムの新たな実装

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?



内容

- 自己紹介
- 学融合の目的と実践
  - データ駆動科学の三つのレベル
  - ベイズ的スペクトル分解を例にして
- 可解線形回帰モデル
- SPring-8全ビームラインベイズ化計画
- まとめと今後の展開
- 集中講義の紹介

線形回帰の可解ベイズ計測

片上舜[A]，柏村周平[A]，永田賢二[c]，水牧仁一朗[c]，
岡田真人[A]

A 東大新領域，B NIMS，C 熊大

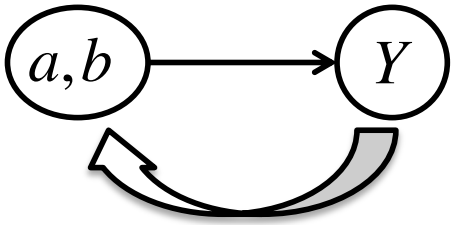
Shun Katakami, Shuhei Kashiwamura, Kenji Nagata, Mizumaki Masaichiro and Masato Okada, "Mesoscopic Bayesian Inference by Solvable Models", https://arxiv.org/abs/2406.02869

ベイズ計測とは?

ベイズ推論

$$p(Y,a,b) = p(Y \mid a,b)\,p(a,b) = p(a,b \mid Y)\,p(Y)$$

生成(因果律)



＜ベイズの定理＞

$$p(a,b \mid Y) = \frac{p(Y \mid a,b)\,p(a,b)}{p(Y)} \propto \exp\left(-nE(a,b)\right)p(a,b)$$

$p(a,b \mid Y)$:事後確率。データが与えられたもとでの物理パラメータの確率.

$p(a,b)$:事前確率。あらかじめ設定しておく必要がある。これまで蓄積されてきた科学的知見

ベイズ計測三種の神器

1.物理パラメータの事後確率分布定

2. モデル選択

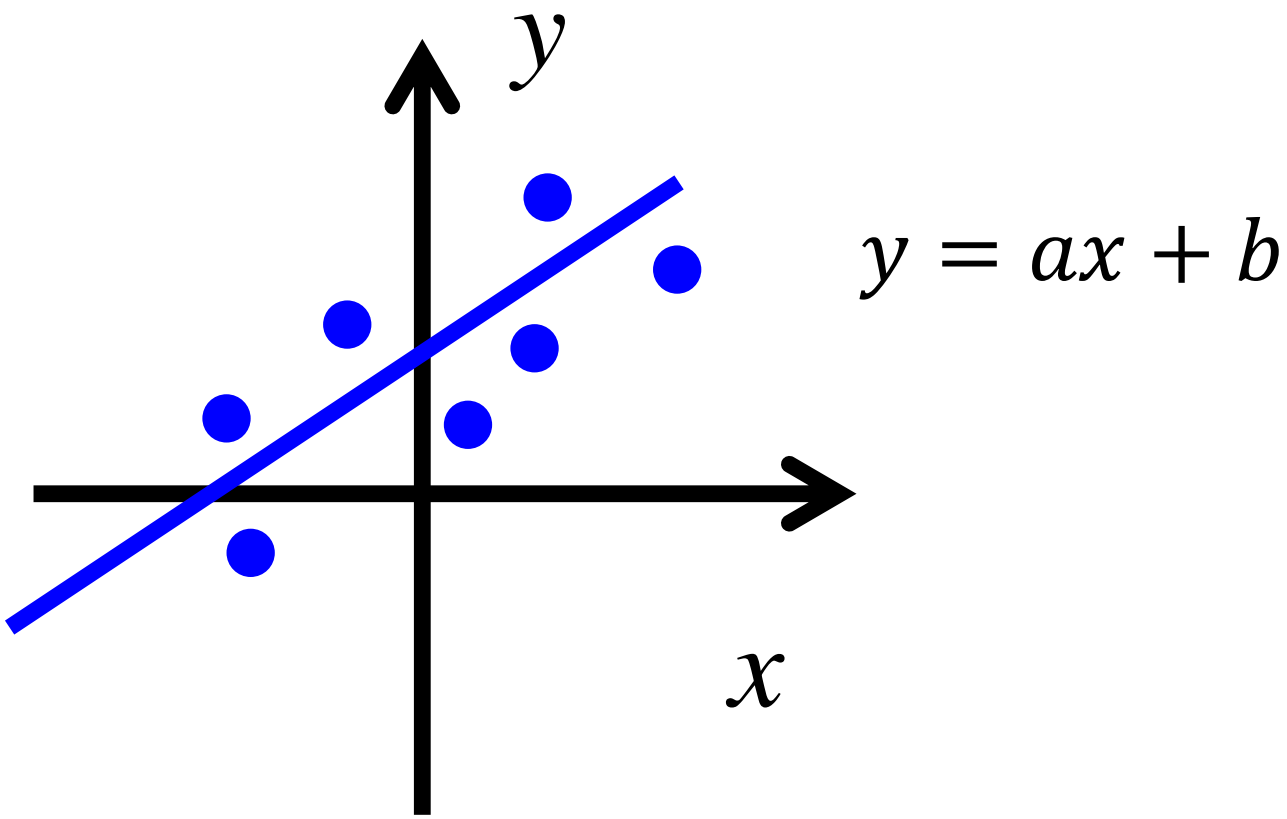
3. データ統合

ベイズ計測の三種の神器

- 神器1: パラメータの事後確率推定
  - 計測限界
- 神器2: ベイズ的モデル選択
  - データを説明する複数のモデル候補から、データのみで、一つのモデルを選択
- 神器3: ベイズ統合
  - 一つに物質に対する複数計測(マルチーモーダル計測)からパラメータを一組み決定

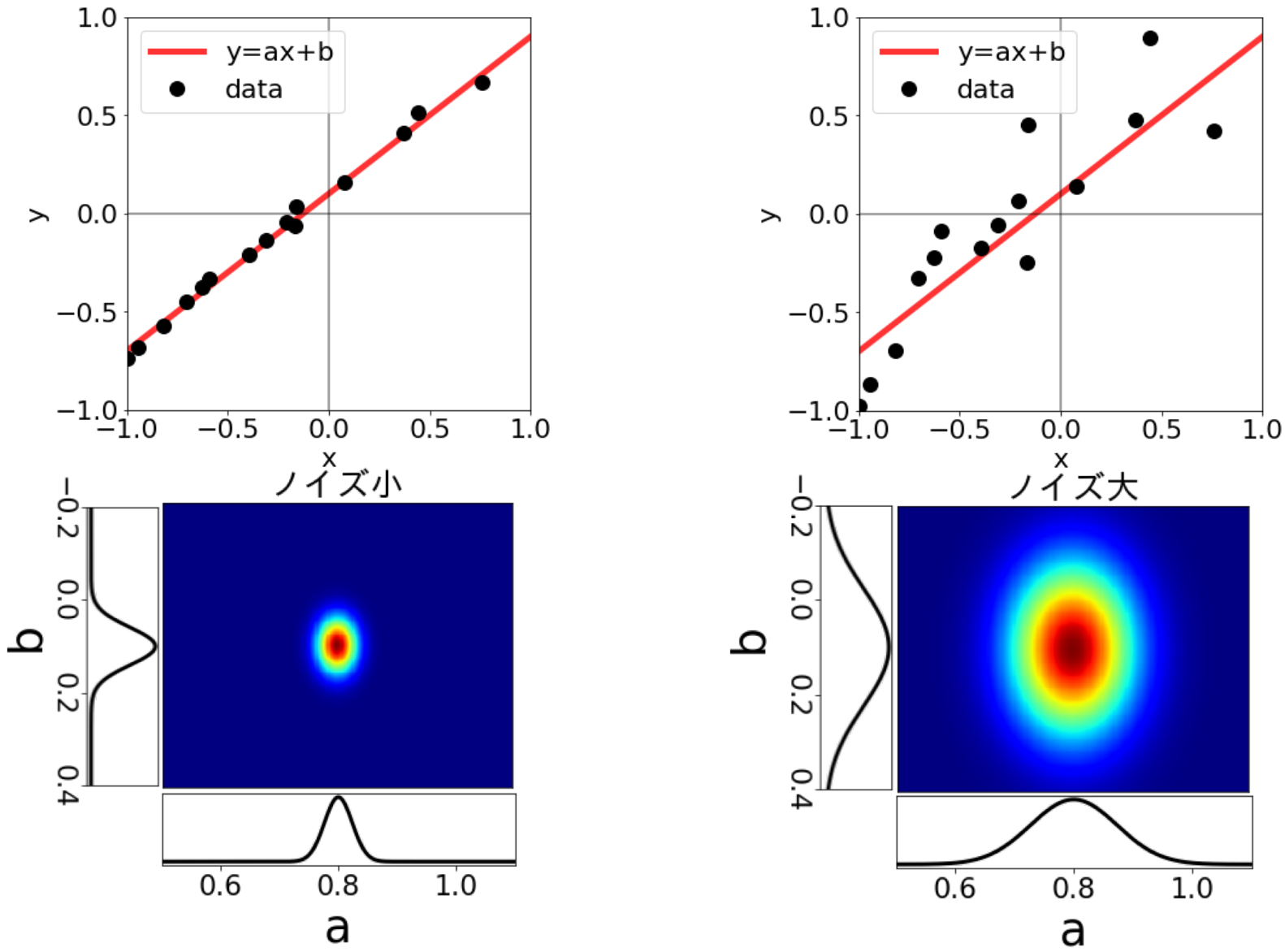
ベイズ計測の利点
$y=ax+b$の取り扱いを通じて

現状でも用いられている最も簡単な例



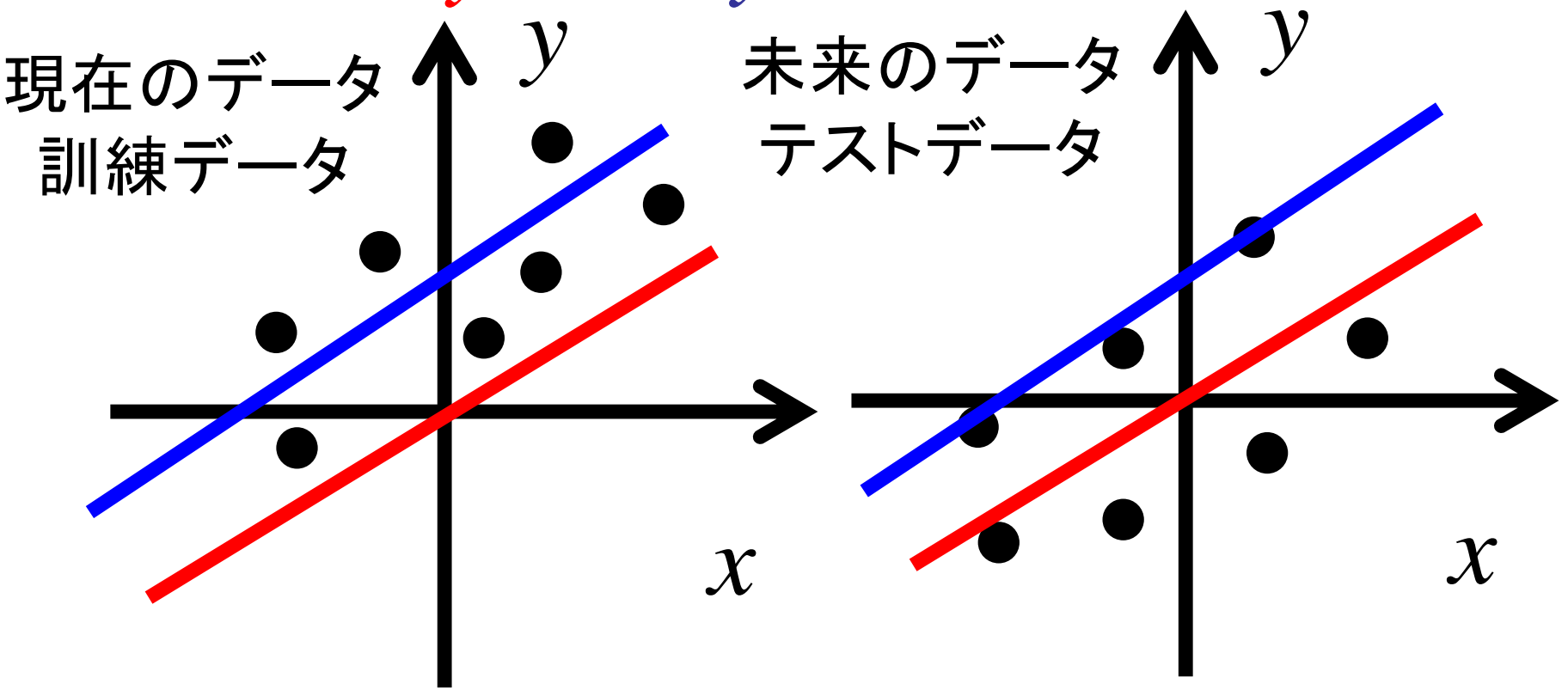
傾き a ： 系の線形応答、バネ定数、電気伝導度、誘電率

神器1: パラメータの事後確率推定 (4/4)



神器2: ベイズ的モデル選択

$y=ax$か$y=ax+b$か?



$y=ax+b$: 訓練誤差小

訓練誤差

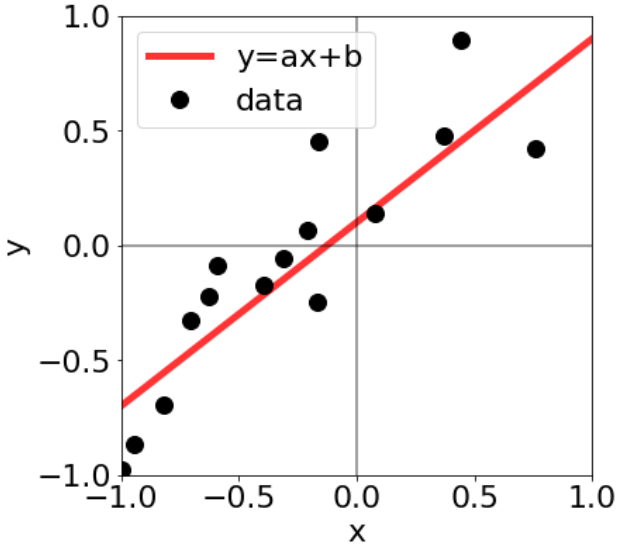
ノイズに過学習

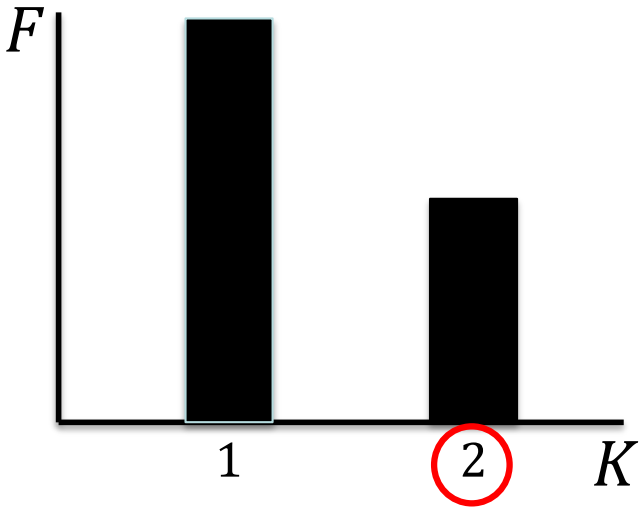
$y=ax$: 訓練誤差小

汎化誤差

モデル選択できる理由:汎化誤差は観測ノイズに依存する

モデル選択: 自由エネルギー差





- $K = 1 : y = ax$
- $K = 2 : y = ax + b$

$$F(K=1) = N\left\{\frac{1}{\sigma^2}E(a_0) + \frac{\log N}{2N}\right\}$$

$$F(K=2) = N\left\{\frac{1}{\sigma^2}E(a_0, b_0) + \frac{\log N}{N}\right\}$$

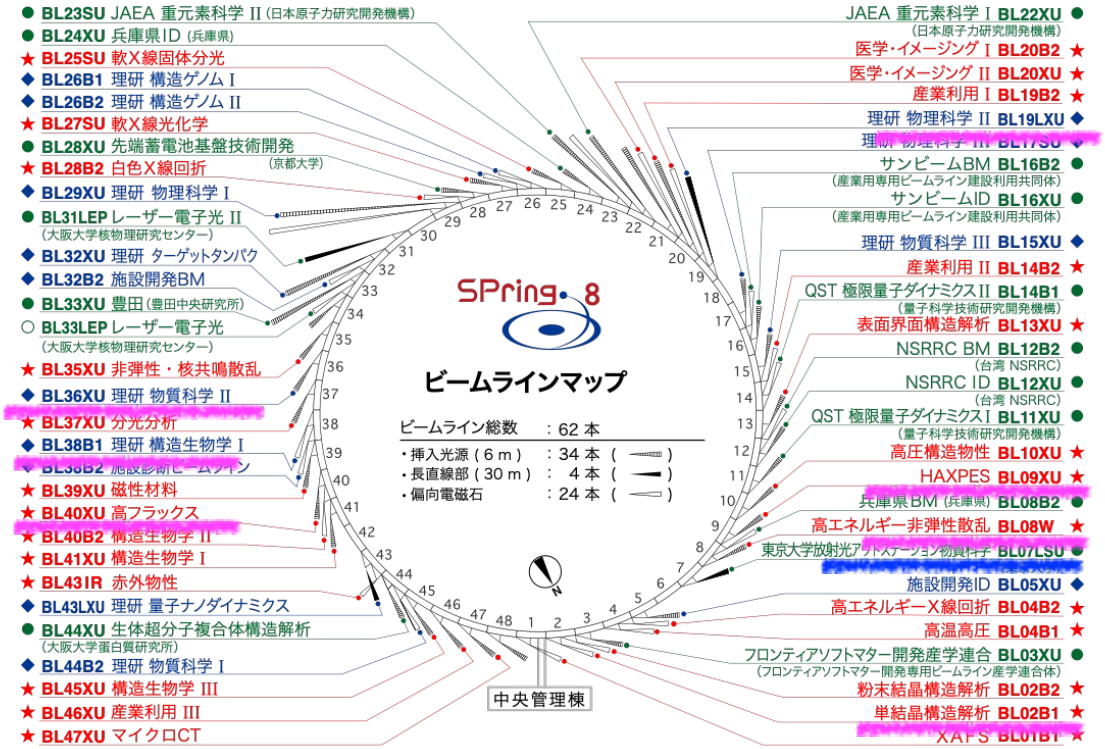
データのみからモデルを選択できる

内容

- 自己紹介
- 学融合の目的と実践
  - データ駆動科学の三つのレベル
  - ベイズ的スペクトル分解を例にして
- 可解線形回帰モデル
- SPring-8全ビームラインベイズ化計画
- まとめと今後の展開
- 集中講義の紹介

SPring-8全ビームラインベイズ化計画

敬称略



情報と放射光研究者のマッチング例

メスバウアー

BL35XU　岡田研学生+筒井

小角散乱

BL08B2　岡田研学生+桑本

BL19B2

XAS測定

BL37XU　岡田研学生+水牧

BL39XU

放射光ユーザーへの展開

時分割XRD

BL02B2　横山優一+河口彰吾、沙織

BL10XU　ユーザー：公立大、東工大

赤色BLが共用BL(JASRI担当): 計26本

2024年中に14BL/26のベイズ化が完了
2025年に全BLベイズ化完了
2023年度理事長賞受賞の波及効果により、
SPring-8全体のミッションとなり、
ベイズ化実績によりBLが評価される体制へ

年度	2021	2022	2023
導入	2	8	14
全BL	26	26	26

SPring-8

- アメリカのAdvanced Photon Source (APS), ヨーロッパのEuropean Synchrotron Radiation Facility(ESRF) と合わせて，世界3大放射光施設.
- 理研はSPring-8を「データ創出基盤」であると言っている．年間延べ1万人が利用.
- APSやESRFにおいてベイズ計測は導入されていない.
- 放射光におけるベイズ計測に関しては日本が最先端である.

SPring-8全ビームラインベイズ化計画

- 通常では系統的手法がない、モデル選択とデータ統合をベイズ計測で系統的に取り扱う
- フラッグシップ戦略: ベイズ計測をSPring-8に導入し、身近(近くにくるな症候群)な計測と他の大型計測施設への起爆剤とする.
- 2023年度JASRI理事長賞JASRIデータ駆動科学グループ横山優一氏受賞を契機に、全BLにベイズ計測利用の加速へ
- 2024年中に14BL/26のベイズ化完了

SPring-8全ビームライン ベイズ化計画の波及効果

- フラッグシップ戦略もあり、追従施設が続出
- SPring-8/JASRI: 2023年3月7日
- あいちSR: 2023年10月30日
- 日本放射光学会 若手研究会: 2024年9月2日
- 台湾(NSRRC): 2024年9月4日
- 佐賀LS: 2024年10月16日
- 広大HiSOR: 2024年11月18日
- PF: 2025年2月6日
- 広大HiSOR: 2025年3月5日

内容

- 自己紹介
- 学融合の目的と実践
  - データ駆動科学の三つのレベル
  - ベイズ的スペクトル分解を例にして
- 可解線形回帰モデル
- SPring-8全ビームラインベイズ化計画
- まとめと今後の展開
- 集中講義の紹介

まとめと今後の展開

- 学融合を実践する肝は、融合する分野の共通基盤を明確にすることである。
- 今日の学融合セミナーのテーマは「データ駆動」であるので、その共通基盤として、科学/工学を系統的に取り扱うことが可能なデータ駆動科学を紹介し、データ駆動科学を用いた学融合の試みを紹介データ駆動科学を用いた学融合の紹介した。
- 学融合を表面的に指向するのではなく、そのためには基盤が必須であることを認識

内容

- 自己紹介
- 学融合の目的と実践
  - データ駆動科学の三つのレベル
  - ベイズ的スペクトル分解を例にして
- 可解線形回帰モデル
- SPring-8全ビームラインベイズ化計画
- まとめと今後の展開
- 集中講義の紹介

HD3: 高次元次元データ駆動科学教育プログラム

- 「推論する」「測定する」「計算する」ことについて学融合的なカリキュラムで視野の広い人材を育てる。
- プログラム修了証書がもらえる。
  - 短期集中講義と基盤系各専攻の通常講義
  - 短期集中講義4単位以上の履修
  - 通常講義を含め合計6単位
  - 新領域創成科学研究科長よりプログラム修了書
- ホームページ
  http://sasakilab.k.u-tokyo.ac.jp/HD3/