材料工学とデータ駆動科学

東京大学・大学院新領域創成科学研究科
複雑理工学専攻
岡田真人
okada@edu.k.u-tokyo.ac.jp

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
  - 情報統計力学、データ駆動科学 (2004/07 –)

村田先生と井上先生との関係

- 理化学研究所　脳科学総合研究センター　甘利チーム 副チームリーダー　(2001 - 04/06)
- 村田先生は、私の前任の副チームリーダー。早稲田大学へのご昇進で、私の副チームリーダー就任を強く勧めてくださった。
- 井上先生は、京大医学研究科大学院生、理研JRA(Junior Research Associate)で、和t氏は井上先生の学院論文の共同研究者。その後、私のポスドク研究員を1年間を務めていただき、東工大樺島研の助教から、早稲田大学へ

内容

- 本講演の目的
- データ駆動科学
- 材料/デバイスの機能発現の3ステップモデル
- 情報数理基盤のベイズ推論とスパースモデリング
- ベイズ推論を計測科学に適用したコンパクトな体系のベイズ計測
  - ベイズ計測三種の神器
- $y=ax+b$の線形回帰、スペクトル分解を述べ、さらに機能発現の3ステップモデルの例として大久保研との共同研究を紹介する。
- 材料工学の展望

内容

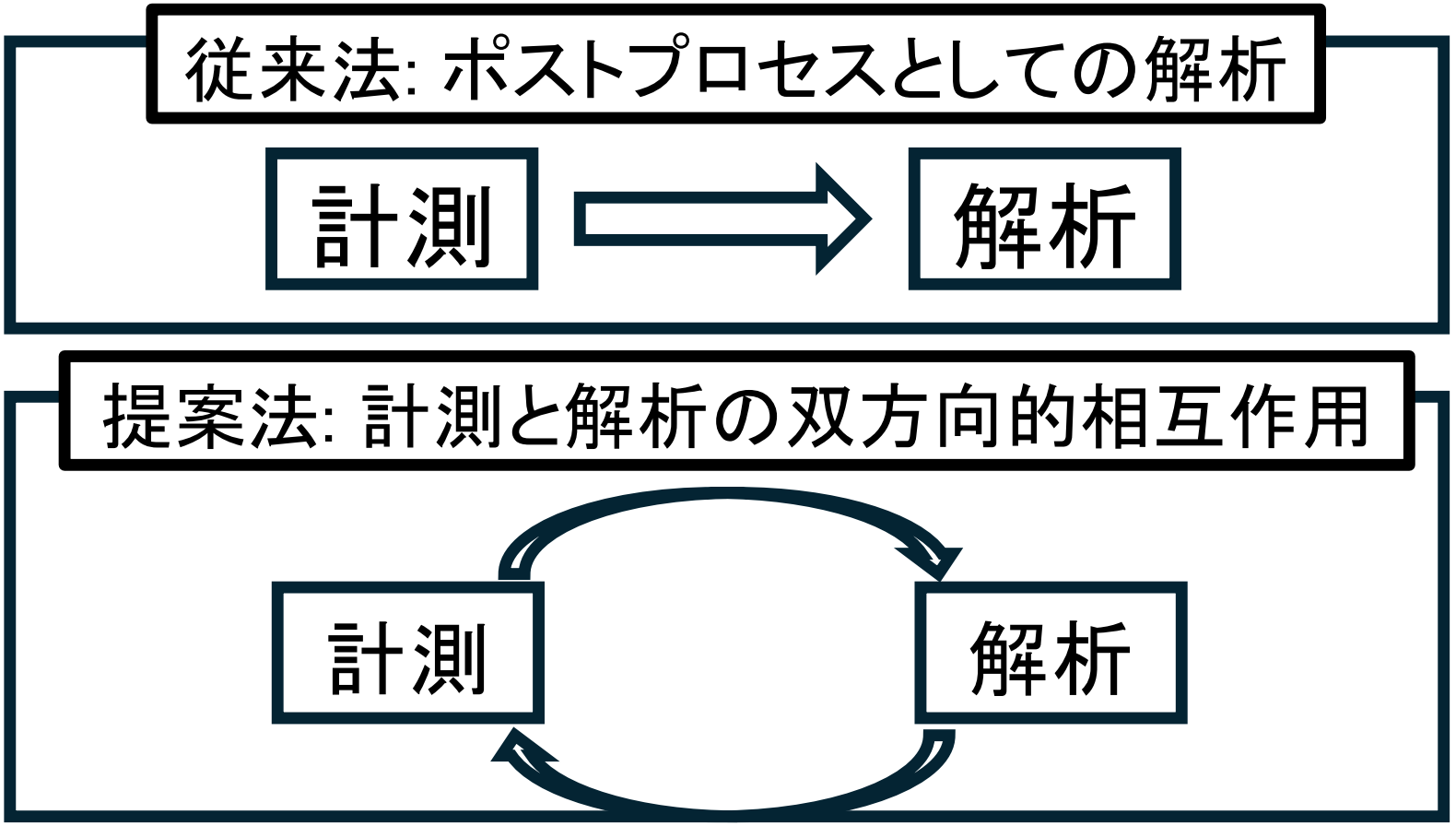
- 本講演の目的
- データ駆動科学
- 材料/デバイスの機能発現の3ステップモデル
- 情報数理基盤のベイズ推論とスパースモデリング
- ベイズ推論を計測科学に適用したコンパクトな体系のベイズ計測
  - ベイズ計測三種の神器
- $y=ax+b$の線形回帰、スペクトル分解を述べ、さらに機能発現の3ステップモデルの例として大久保研との共同研究を紹介する。
- 材料工学の展望

本講演の目的

- 実験/計測のポストプロセス/付帯事項として捉えられがちなデータ解析が、材料工学を推進する要。
- その学理として、データ駆動科学その情報数理基盤のベイズ計測とスパースモデリング
- データ駆動科学の三つのレベルと機能発現の3ステップモデルの紹介
- ベイズ計測の実例として、y=ax+bとスペクトル分解のベイズ計測の紹介
- 機能発現の3ステップモデルの具体例の大久保研との共同研究
- 材料工学の展望

ベイズ計測による
計測と解析の双方向相互作用

計測限界から実験計画へ

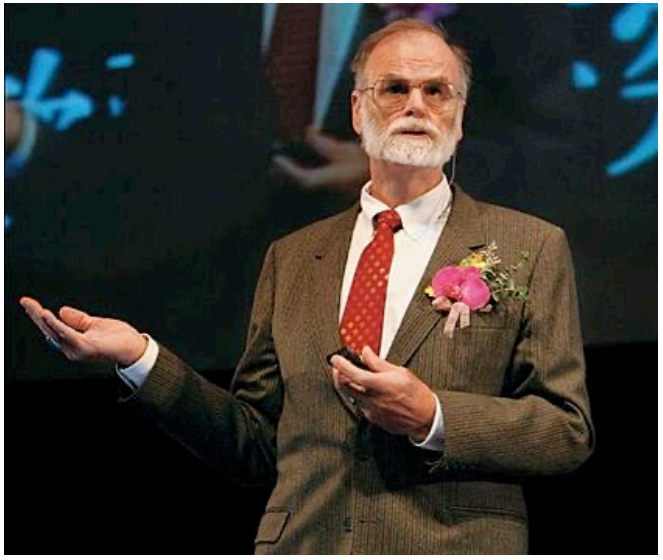


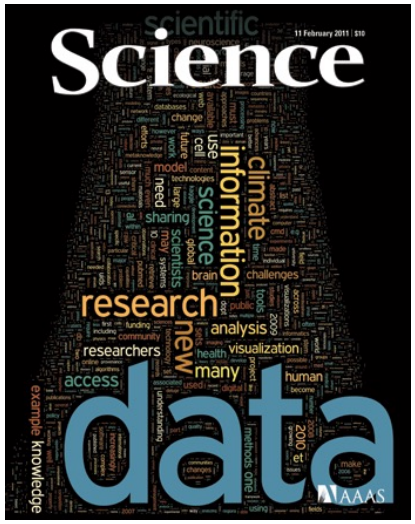
実験家もベイズ計測を使いこなす時代へ

内容

- 本講演の目的
- データ駆動科学
- 材料/デバイスの機能発現の3ステップモデル
- 情報数理基盤のベイズ推論とスパースモデリング
- ベイズ推論を計測科学に適用したコンパクトな体系のベイズ計測
  - ベイズ計測三種の神器
- $y=ax+b$の線形回帰、スペクトル分解を述べ、さらに機能発現の3ステップモデルの例として大久保研との共同研究を紹介する。
- 材料工学の展望

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





[*Science*, Feb. 2011]

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)

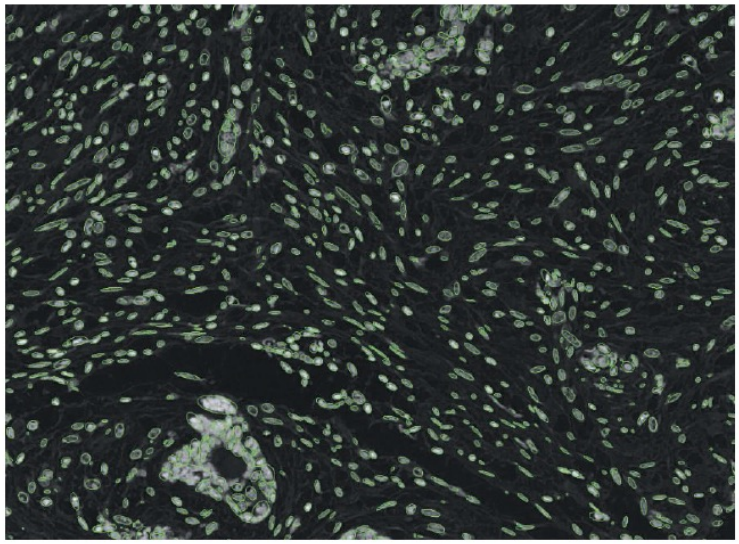
⇩

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

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



NEWS

Is There an Astronomer in the House?

[*Science*, Feb. 2011]

With biomedical researchers analyzing stars and astronomers tackling cancer, two unlikely collaborations creatively solve data problems

pathologists look for different biomarkers—specific proteins—in the patient's cancerous

と単純に喜んで良いのか?! ⇒ 必要なことは

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

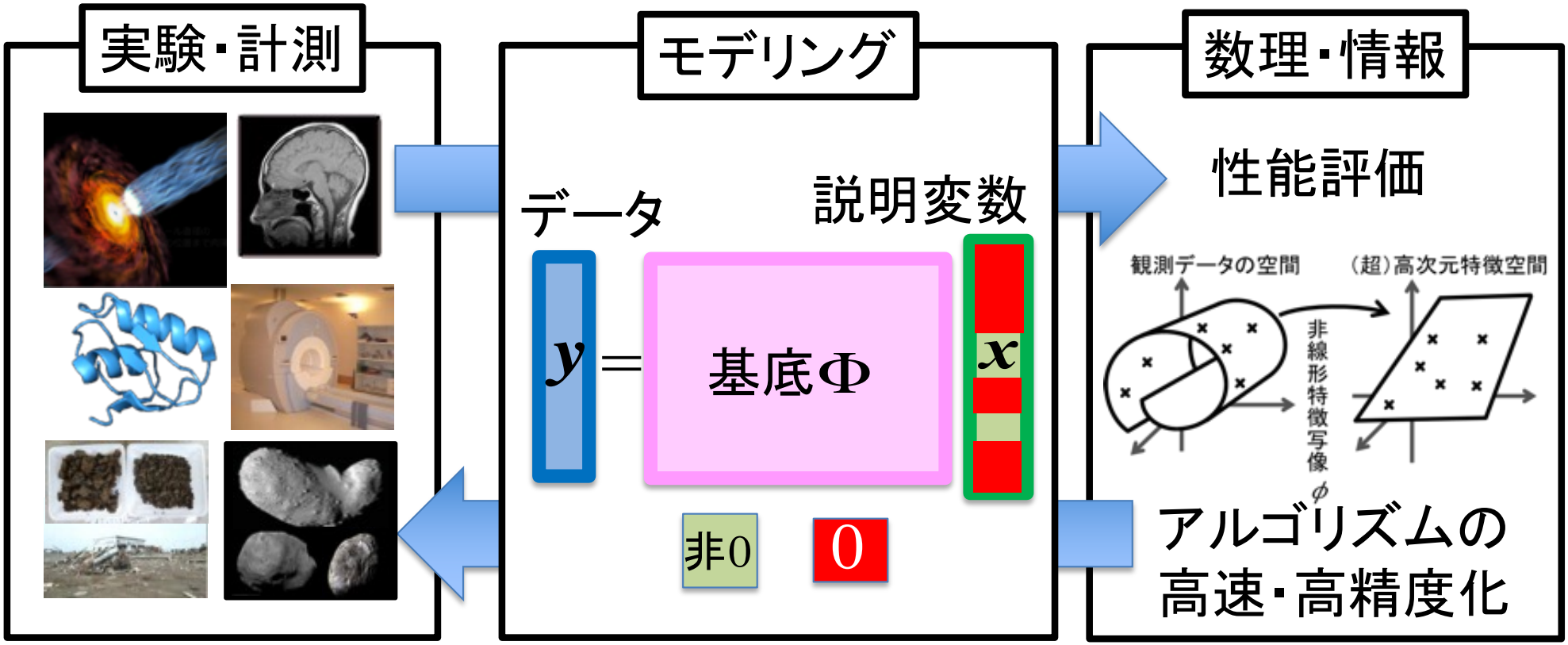
データ駆動科学とは

- 機械学習などの人工知能を使い，各学問分野の問題を解いていくというアプローチ
- 実験/計測/計算データの背後にある潜在的構造の抽出に関して，データが対象とする学問に依存しない普遍的な学問体系
- 同じアルゴリズムがスケールや対象を超えて，有用であることが多いという経験的事実を背景として，その理由を問い，背後にある普遍性から，データ解析自体を学問的対象とする枠組み．
- 全ての実験/計測のデータ解析をデータ駆動学の三つのレベルの鋳型に押し込んで考える

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

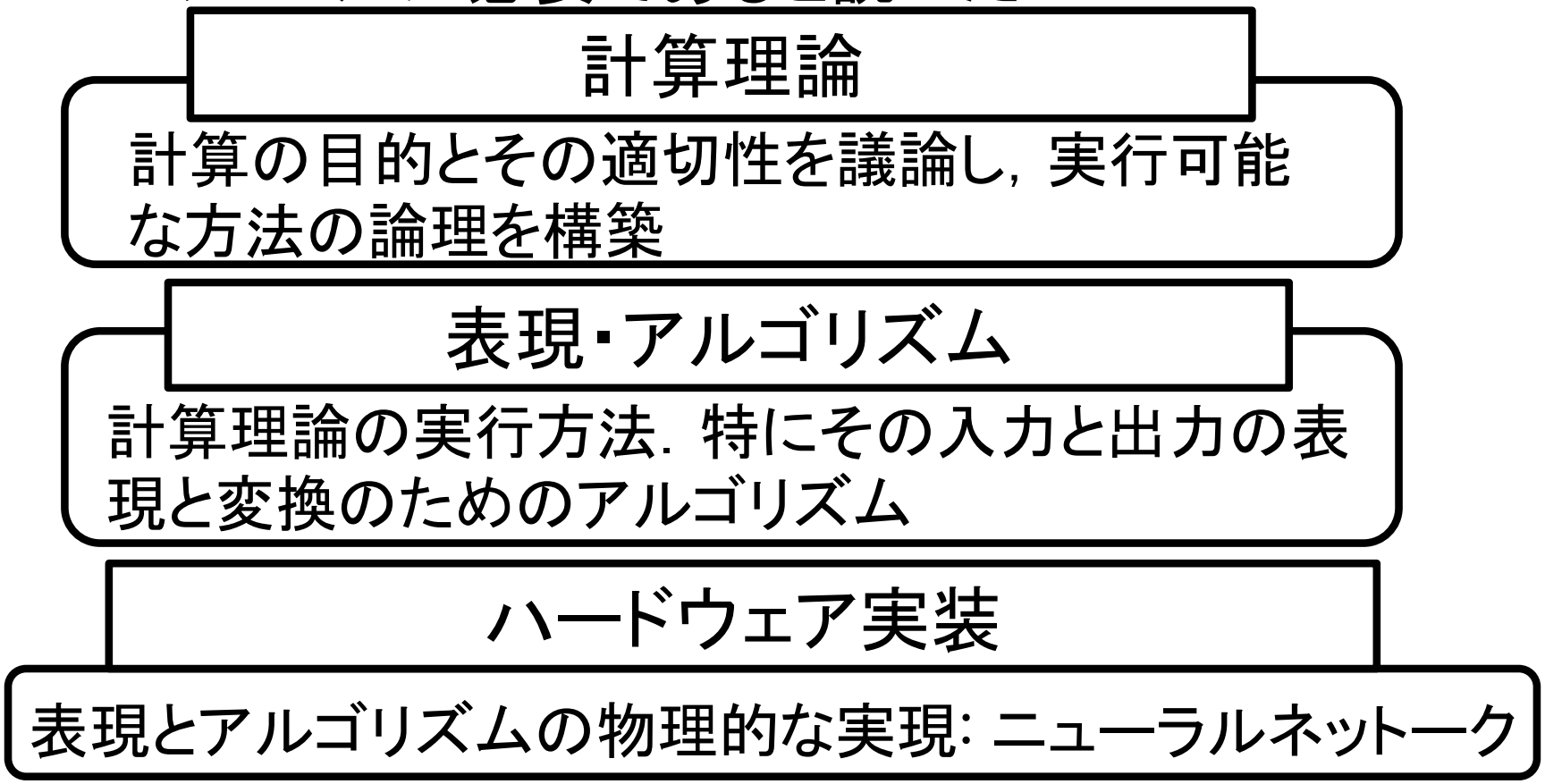
領域代表岡田真人の個人的な狙い

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?



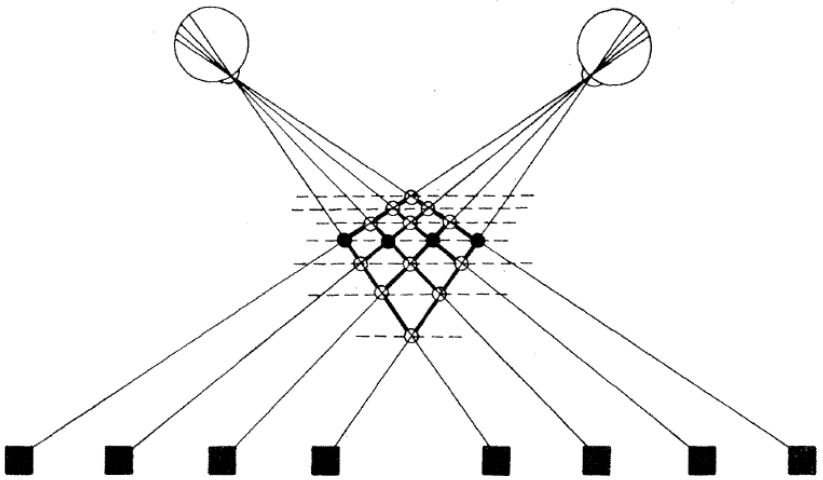
David Marrの三つのレベル (1982)

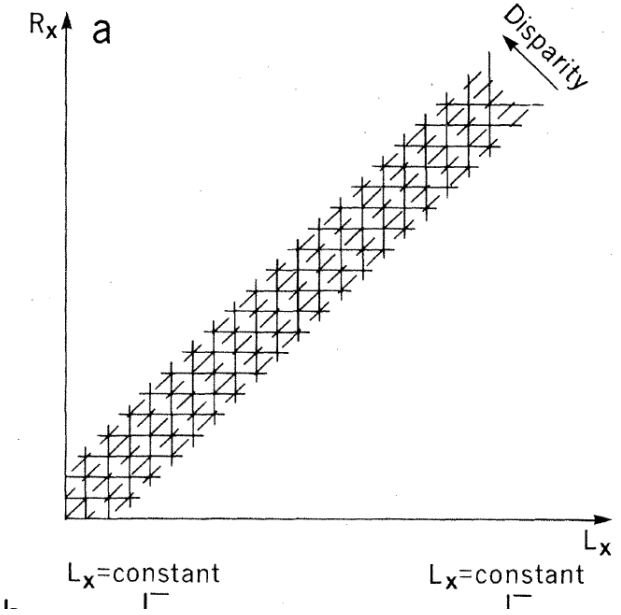
David Marrは複雑な情報処理装置を理解するには以下の三つのレベルが必要であると説いた



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

Marr Poggio: Cooperative Computation of Stereo Disparity *Science*, (1976)





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.

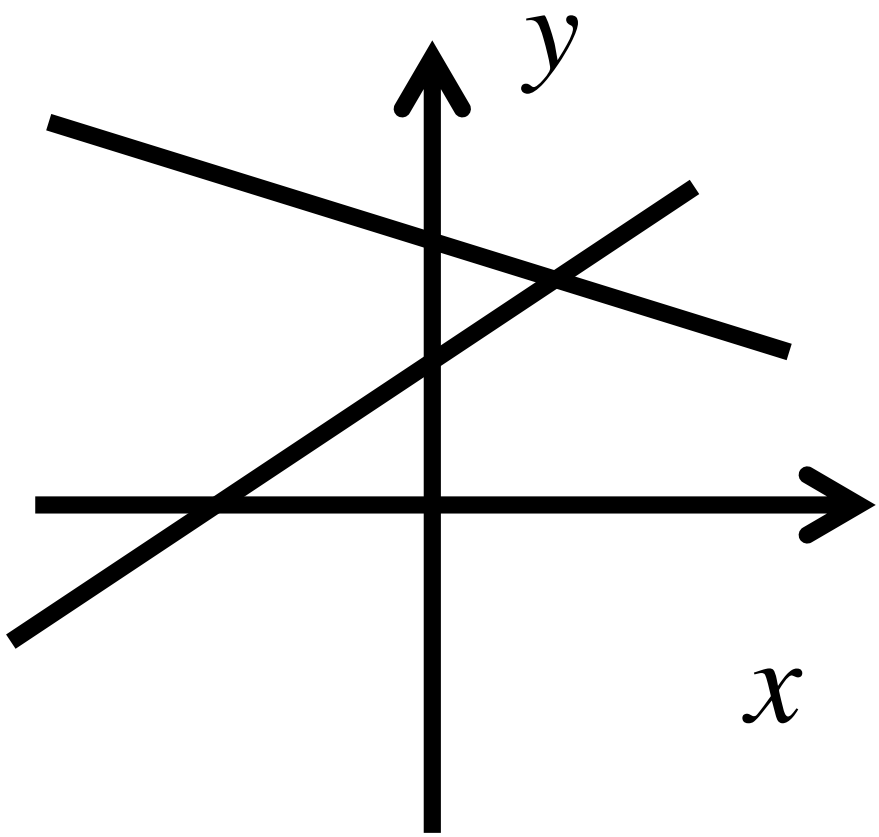
連立方程式とデータ駆動科学

連立方程式とその応用

鶴亀算　食塩水　寝坊
して追いかける問題

連立方程式への変換

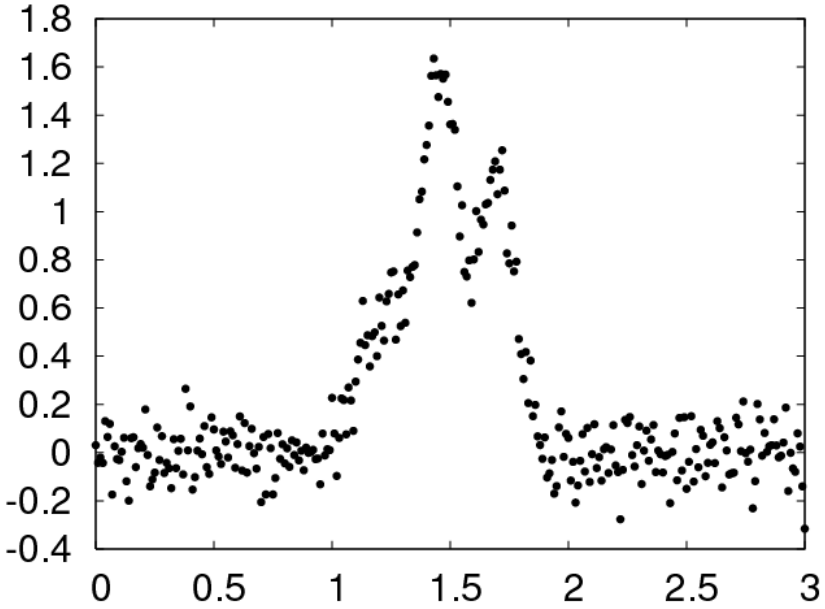
加減法，代入法

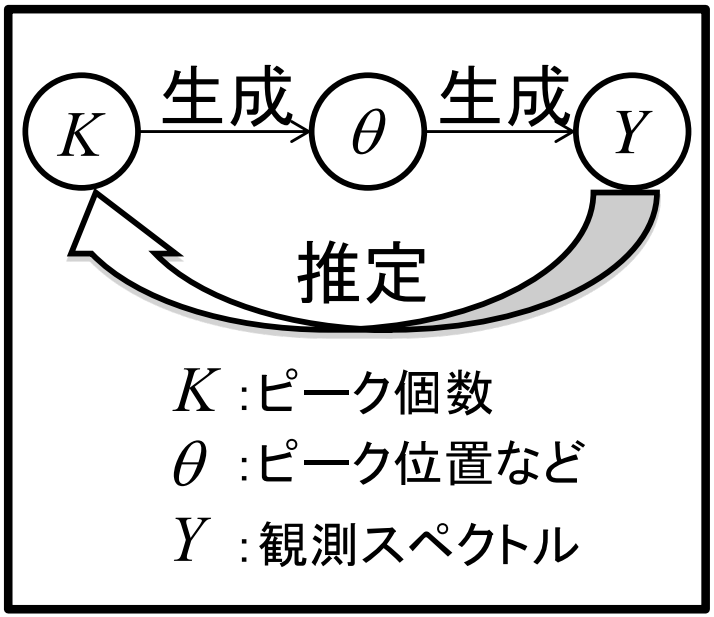


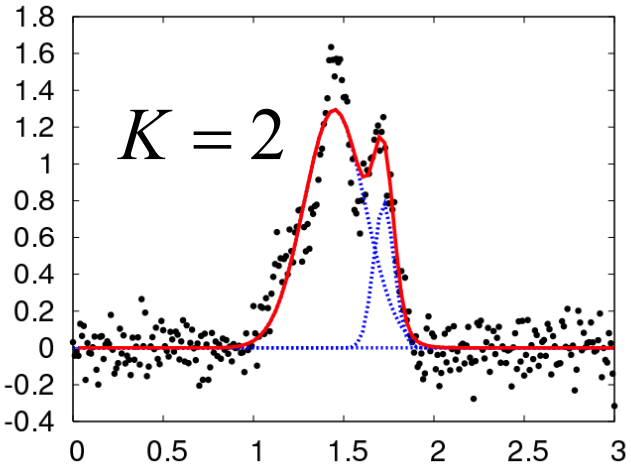
一つの方程式が一本の線
二本の線の交点が解になる

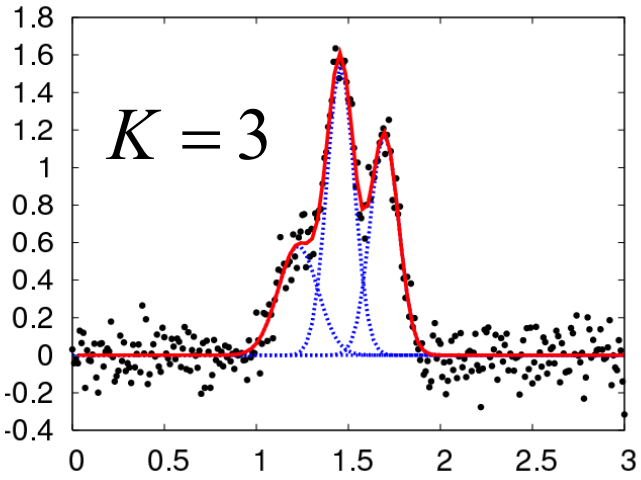
（五十嵐，竹中，永田，岡田，*応用統計学*，2016）

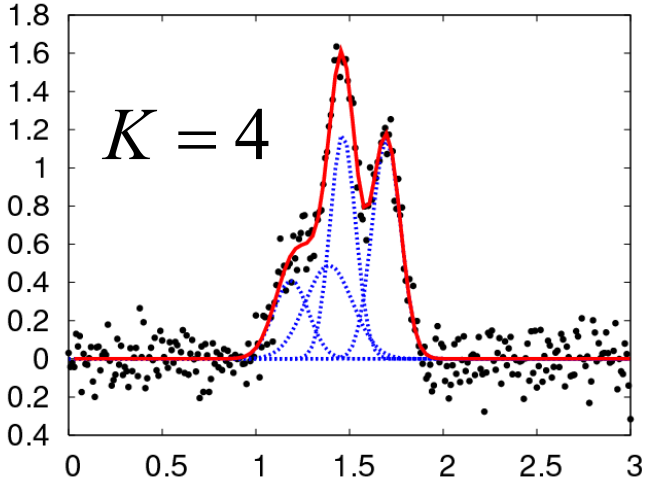
ベイス的スペクトル分解: Kをどう選ぶか











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する。
　誤差関数の最小化では、単峰性関数が多い方が誤差
　関数は小さい。そこで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K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するとモデリングする

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.

データ駆動科学の三つのレベルと計測関連企業の高収益化

過当競争化するプログラムではなく、顧客が担当すべき計算理論をコンサルし、高収益化を目指す

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.

計測関連企業は
優良ソリューションビジネスを目指せ

計算理論(対象の科学，計測科学)

計算理論の構築は、インタビューによる
顧客のニーズの掘りお越しに対応.

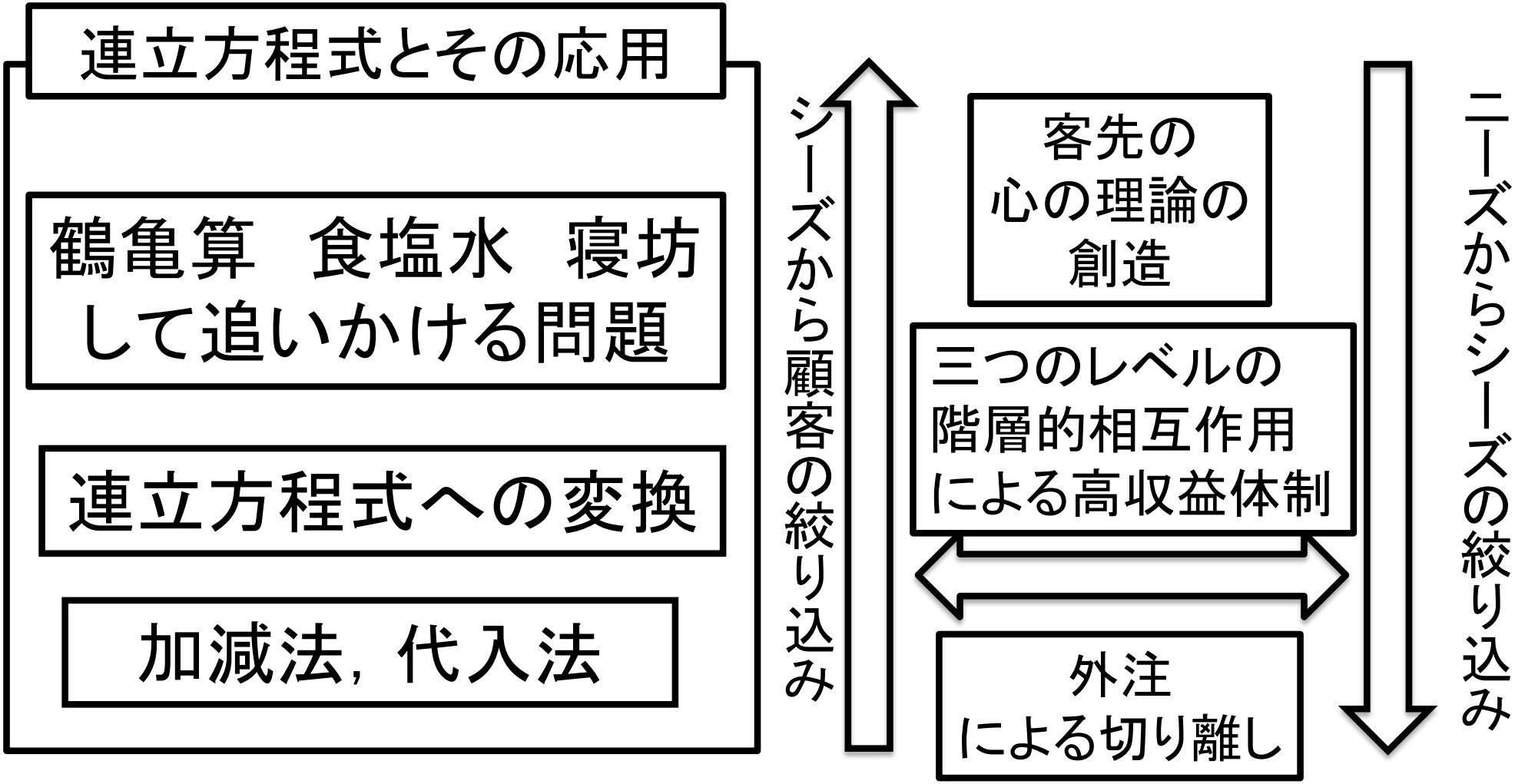
モデリング(統計学，理論物理学，数理科学)

顧客ニーズの数理モデル化による推理的絞り込み

表現・アルゴリズム(統計学，機械学習，計算科学)

顧客によらない表現とアルゴリズムによる
低コスト化と外注化

連立方程式からの
高収益体制化へのヒント



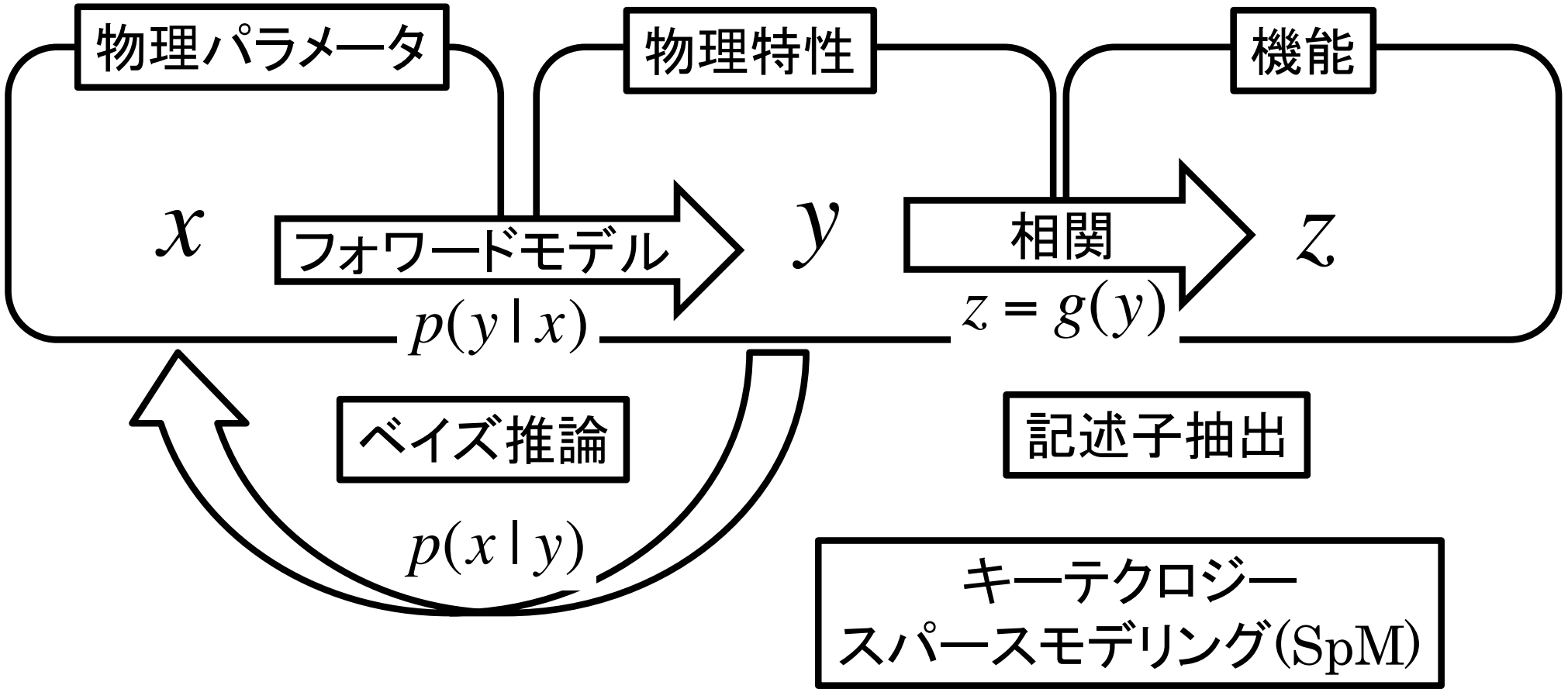
（五十嵐，竹中，永田，岡田，*応用統計学*，2016）

内容

- 本講演の目的
- データ駆動科学
- 材料/デバイスの機能発現の3ステップモデル
- 情報数理基盤のベイズ推論とスパースモデリング
- ベイズ推論を計測科学に適用したコンパクトな体系のベイズ計測

  –ベイズ計測三種の神器
- $y=ax+b$の線形回帰、スペクトル分解を述べ、さらに機能発現の3ステップモデルの例として大久保研との共同研究を紹介する。
- 材料工学の展望

機能発現の3ステップモデル



第15回NIMS フォーラム 2015年10月7日（水）
マテリアルズ・インフォマティクスとは何か-物質材料科学とデータ駆動科学-
https://mns.k.u-tokyo.ac.jp/pdf/2015nims.pdf

Igarashi, Nagata, Kuwatani, Omori, Nakanishi-Ohno, and Okada “Three levels of data-driven science” International meeting on High-dimensional Data-Driven Science (HD3-2015), *Journal of Physics: Conference Series*, 699 (2016) 012001(2016)

機能発現の3ステップモデル

- 一般に、利益に直結する機能は、数理モデル一気に記述できない。
- 物理プロセスが数理モデルで記述できている場合は、ベイズ推論ベイズ推論で、特徴量抽出
- スパースモデリング(SpM)で、機能から特徴量選択
- SpMの結果として得られた特徴量をコントロールすることで、所望の機能特性を得る。
- 材料工学を機能発現の3ステップモデルの鋳型に押し込める

内容

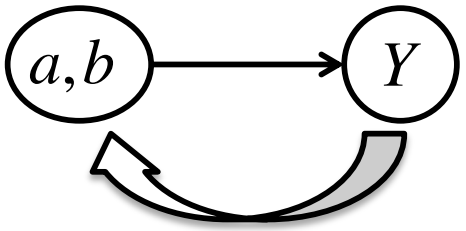
- 本講演の目的
- データ駆動科学
- 材料/デバイスの機能発現の3ステップモデル
- 情報数理基盤のベイズ推論とスパースモデリング
- ベイズ推論を計測科学に適用したコンパクトな体系のベイズ計測

  －ベイズ計測三種の神器
- $y=ax+b$の線形回帰、スペクトル分解を述べ、さらに機能発現の3ステップモデルの例として大久保研との共同研究を紹介する。
- 材料工学の展望

ベイズ推論
詳細は後ほど説明

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

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，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

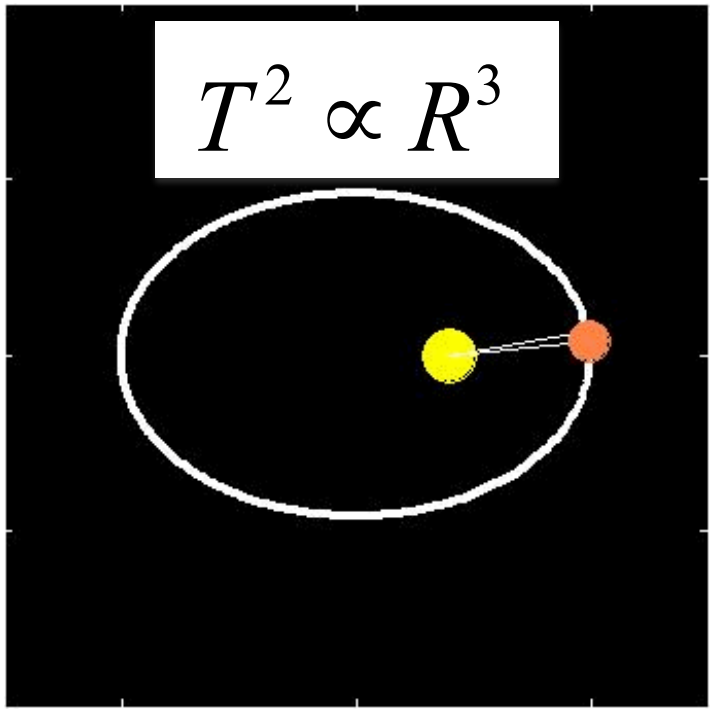
物理学とスパースモデリング(SpM)

- 古典力学や量子力学の前段階で、スパースモデリング(SpM)は活用されている歴史
- ニュートン力学に対するKeplerの法則
  - 公転周期Tと公転半径R
- 前期量子論
  - プランクの輻射の理論、アインシュタインの光量子仮説
- これらは全て、そのレベルを記述する数理モデルがない段階で、実験データから特徴量をヒトが決め、その特徴量を用いて、現象を定量的に記述する現象論である

Keplerの法則

ティコ・ブラーエの
天体観測データ

Keplerの法則



観測データからヒトが直感で特徴量TとRを抽出し
その定量的現象論を提案

線形回帰モデル

- 目的変数 : y
- 準備した特徴量 : $(x_1, x_2, \ldots, x_p)$

関係式 $y = g(x_1, x_2, \ldots, x_p)$ に対して，線形和による近似を考える

$$y = g(x_1, x_2, \ldots, x_p)$$
$$\approx w_1 x_1 + w_2 x_2 + \cdots + w_p x_p$$

ただし w_i は回帰係数

予測モデルの汎化性能

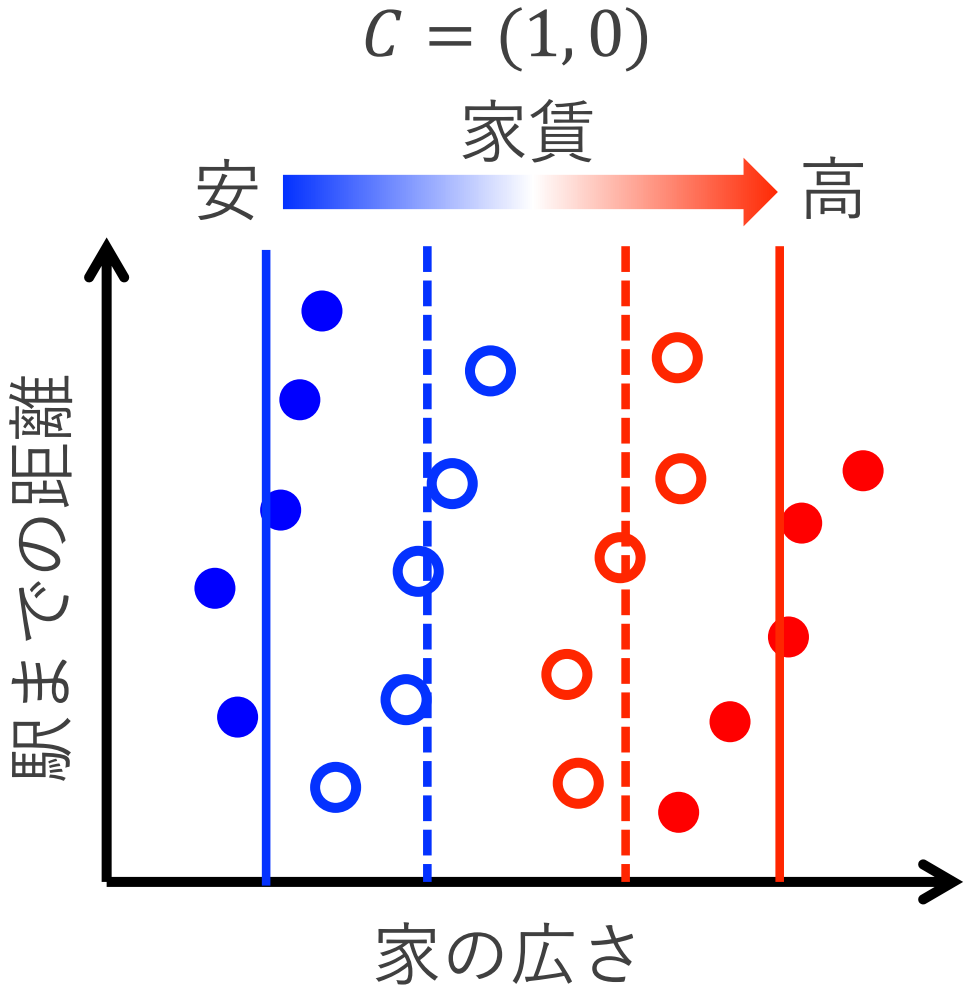
機械学習モデルに求められる性質

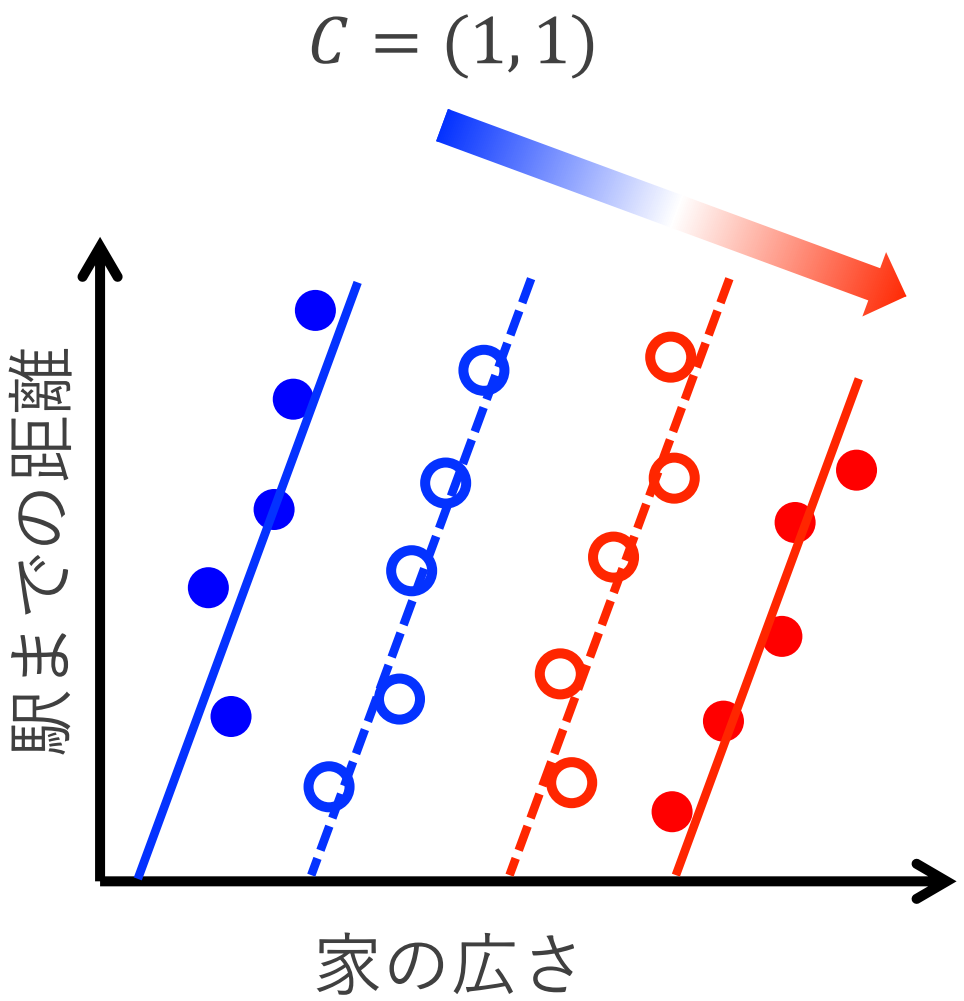
→ 未知データを上手く予測すること（汎化性能）

$$y \approx w_1 x_1 + w_2 x_2 + \cdots + w_p x_p$$

汎化性能を高めるためには，必要な特徴量を見極めることが重要となる

特徴量選択の効果①



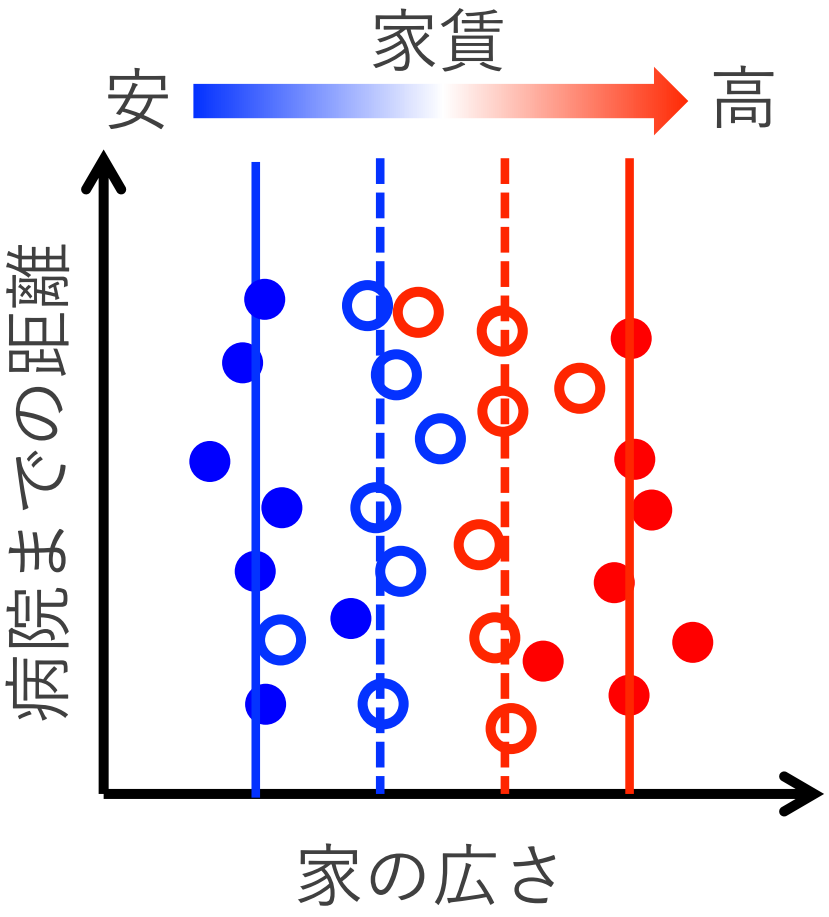


$Error(1, 0) > Error(1, 1)$

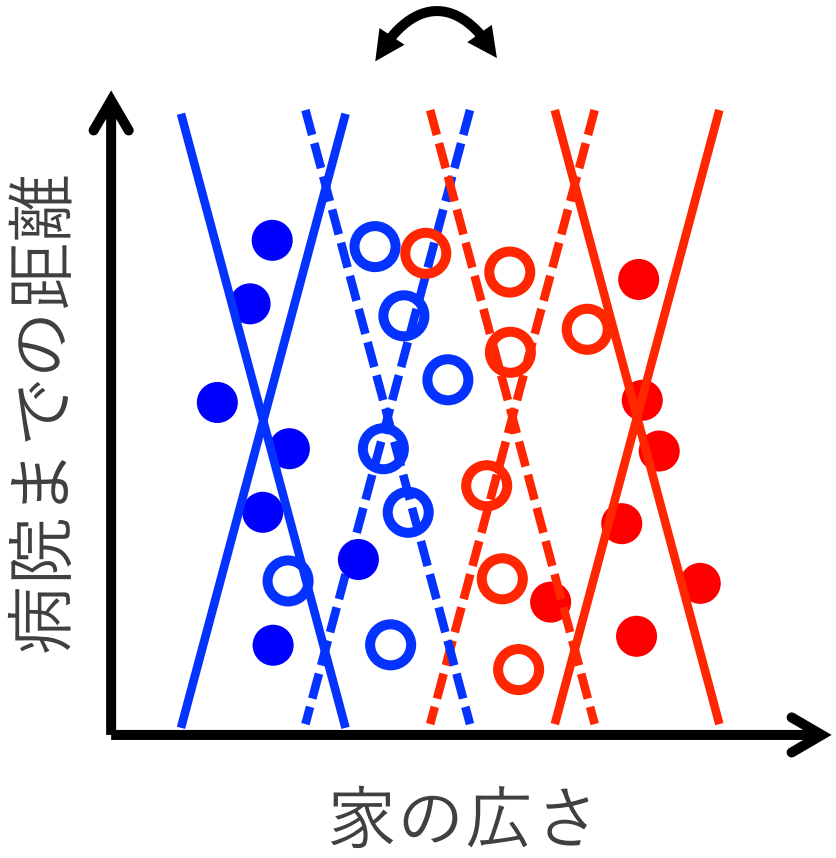
- 両方とも予測に必要

特徴量選択の効果②

家の広さのみ $C = (1, 0)$



家の広さと病院までの距離
$C = (1, 1)$



$Error(1, 0) < Error(1, 1)$

- 家の広さ　　：必要
- 病院までの距離：不要

インジケータベクトルによるモデルの指定

インジケータベクトル $\boldsymbol{c}$ によって部分モデルを定義

$$\boldsymbol{c} = (1, 1, 0, \ldots, 0, 1, 0, 1) \in \{0,1\}^p$$

- i 番目の要素が特徴量 i に対応
  - 1:モデルに含まれる
  - 0:モデルに含まれない
- インジケータベクトルは 2^p 状態を取る
  - その内1つは空のモデル

厳密な特徴量選択を行うにはどの手法であっても計算量が指数関数的に増加する(Cover and Van Campenhout, 1977)

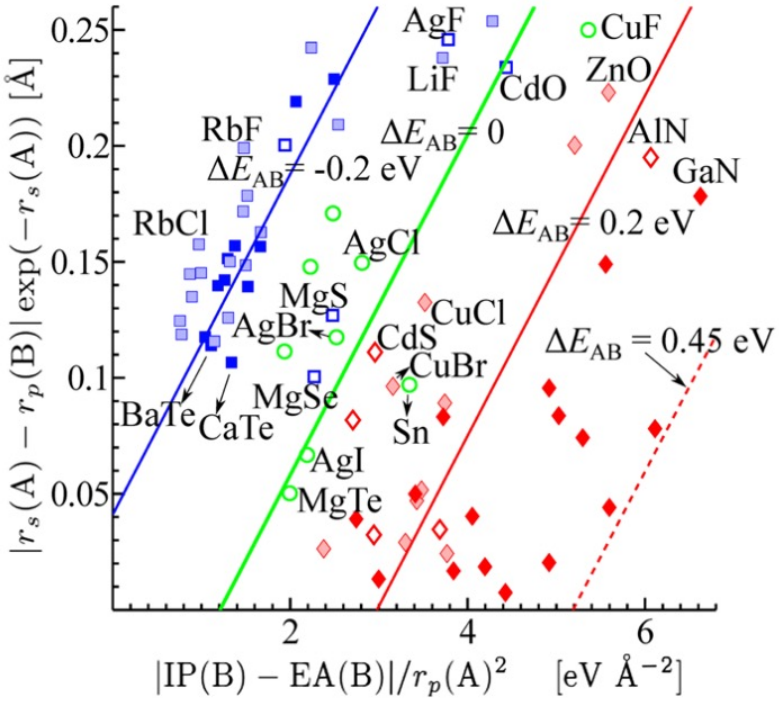
物理学における特徴量選択の重要性

- 物理的知見の抽出
  - 選ばれた特徴量と目的変数の関係性を考察することで，物理現象への知見が得られる
    - Ghiringhelli *et al.*, 2015, Igarashi *et al.*, 2018
- 特徴量の準備の省力化
  - 選ばれた特徴量のみ準備すればよいため，実験の計測時間や回数及び数値シミュレーションの回数を削減することが可能
    - Nakanishi-Ohno *et al.*, 2016

材料科学における特徴量選択の応用

1. 電池材料に関する化学反応の安定性予測
   Sodeyama *et al*, 2018

2. 半導体化合物の結晶構造予測
   Ghiringhelli *et al*., 2015(右図)

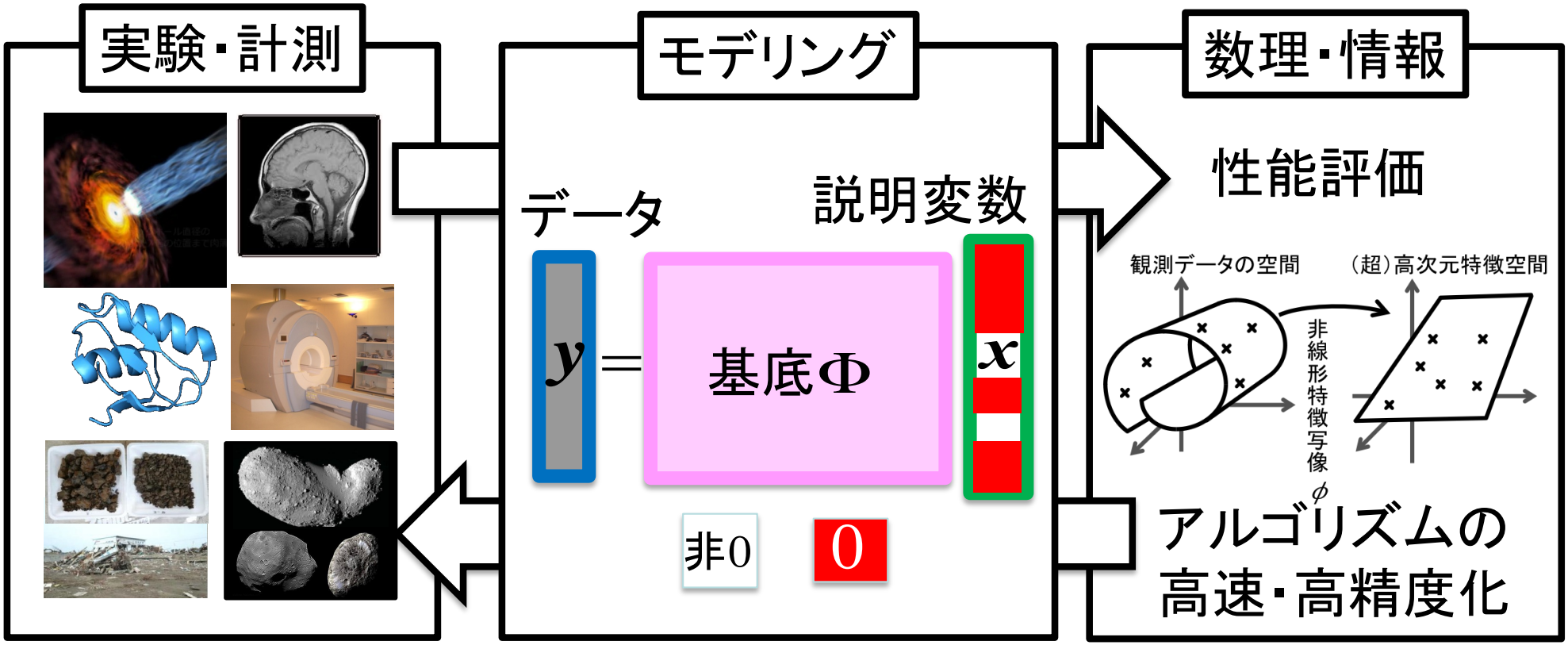


2つの特徴量からなる線形モデルによる結晶構造予測

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

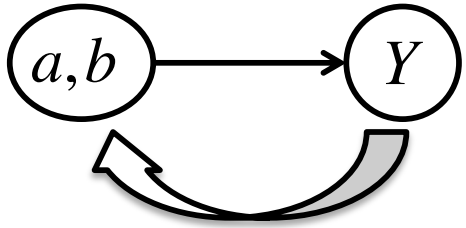
- 本講演の目的
- データ駆動科学
- 材料/デバイスの機能発現の3ステップモデル
- 情報数理基盤のベイズ推論とスパースモデリング
- ベイズ推論を計測科学に適用したコンパクトな体系のベイズ計測
  - ベイズ計測三種の神器
- $y=ax+b$の線形回帰、スペクトル分解を述べ、さらに機能発現の3ステップモデルの例として大久保研との共同研究を紹介する。
- 材料工学の展望

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

ベイズ推論 (2/2)

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

高3数Cで学ぶ確率積の公式による世界記述

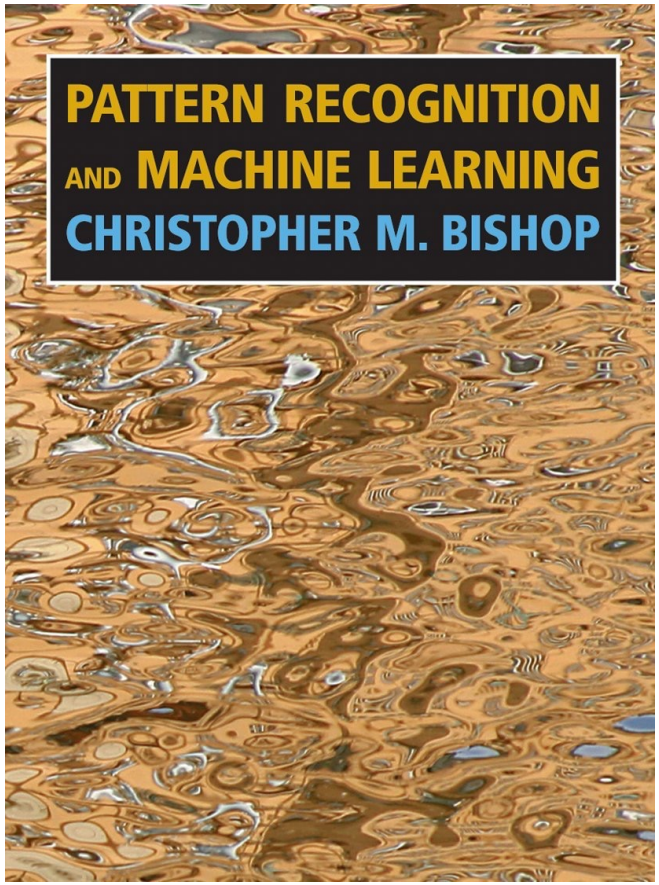
$p(Y,a,b)$は神の世界記述

$p(Y,|\ a,b)\ p(a,b)$はヒトの世界記述

$p(a,b|\ Y)\ p(Y)$はデータ解析

驚くべきことに因果律と推定がイコールで結べる

ベイズ推論とベイズ計測は違う



分厚い本を読む必要はない．分厚い部分のほとんどは近似アルゴリズムの説明

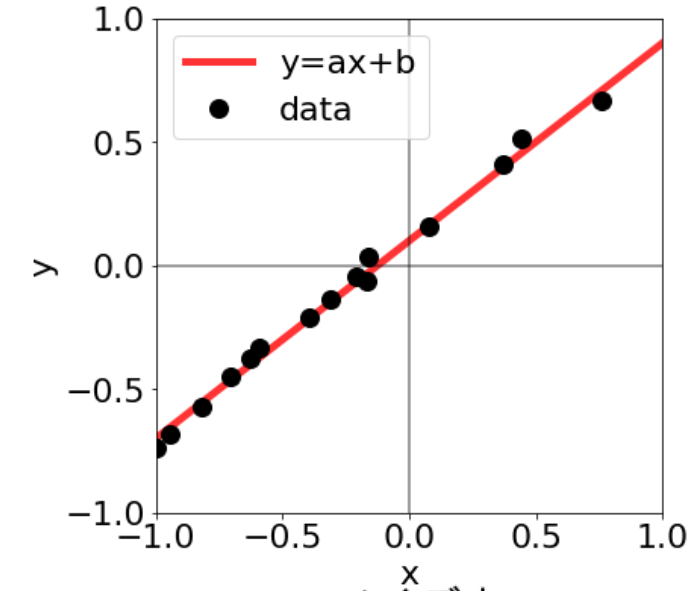
ベイズ計測三種の神器
表現とアルゴリズムが少数であることが重要

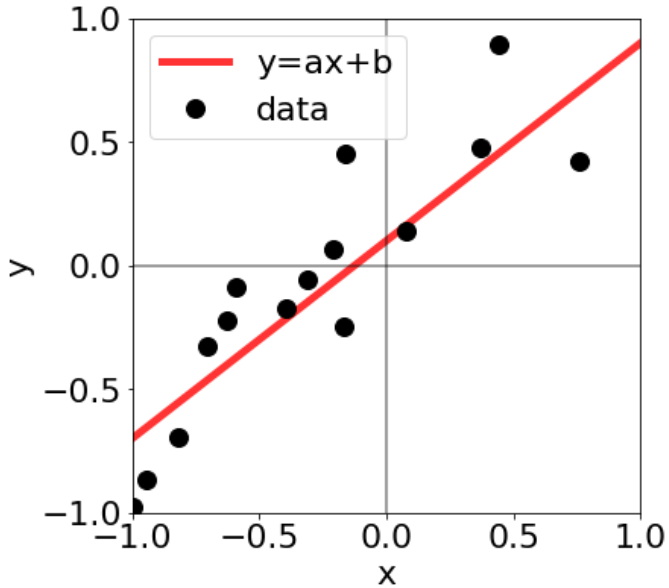
- パラメータの事後確率推定: 数理モデルのフリーパラメータを決める系統的枠組み
- ベイズ的モデル選択: 複数モデルをデータだけから選択する系統的枠組み
- ベイズ統合: 同一物質に対する複数の計測データを統合する系統的枠組み

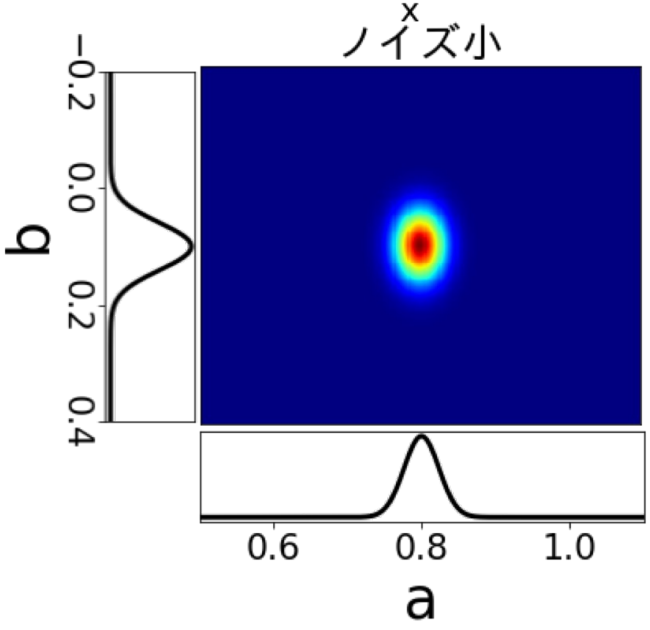
内容

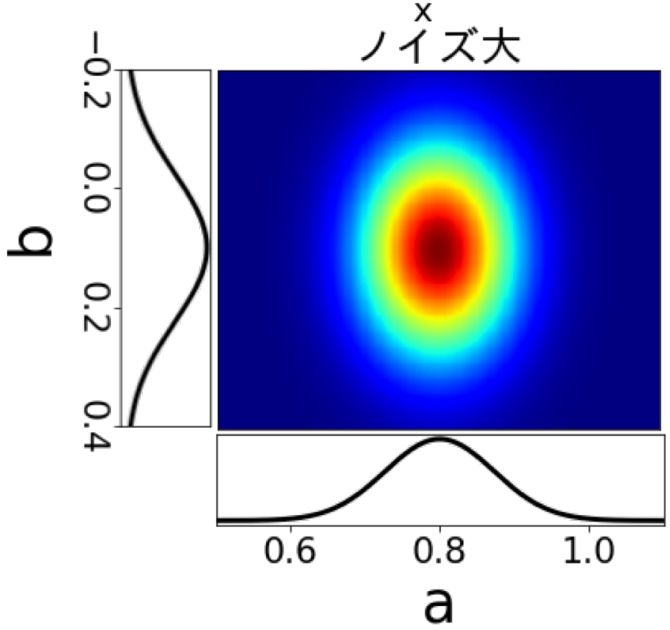
- 本講演の目的
- データ駆動科学
- 材料/デバイスの機能発現の3ステップモデル
- 情報数理基盤のベイズ推論とスパースモデリング
- ベイズ推論を計測科学に適用したコンパクトな体系のベイズ計測
  - ベイズ計測三種の神器
- $y=ax+b$の線形回帰、スペクトル分解を述べ、さらに機能発現の3ステップモデルの例として大久保研との共同研究を紹介する。
- 材料工学の展望

直線回帰$y=ax+b$のベイズ計測



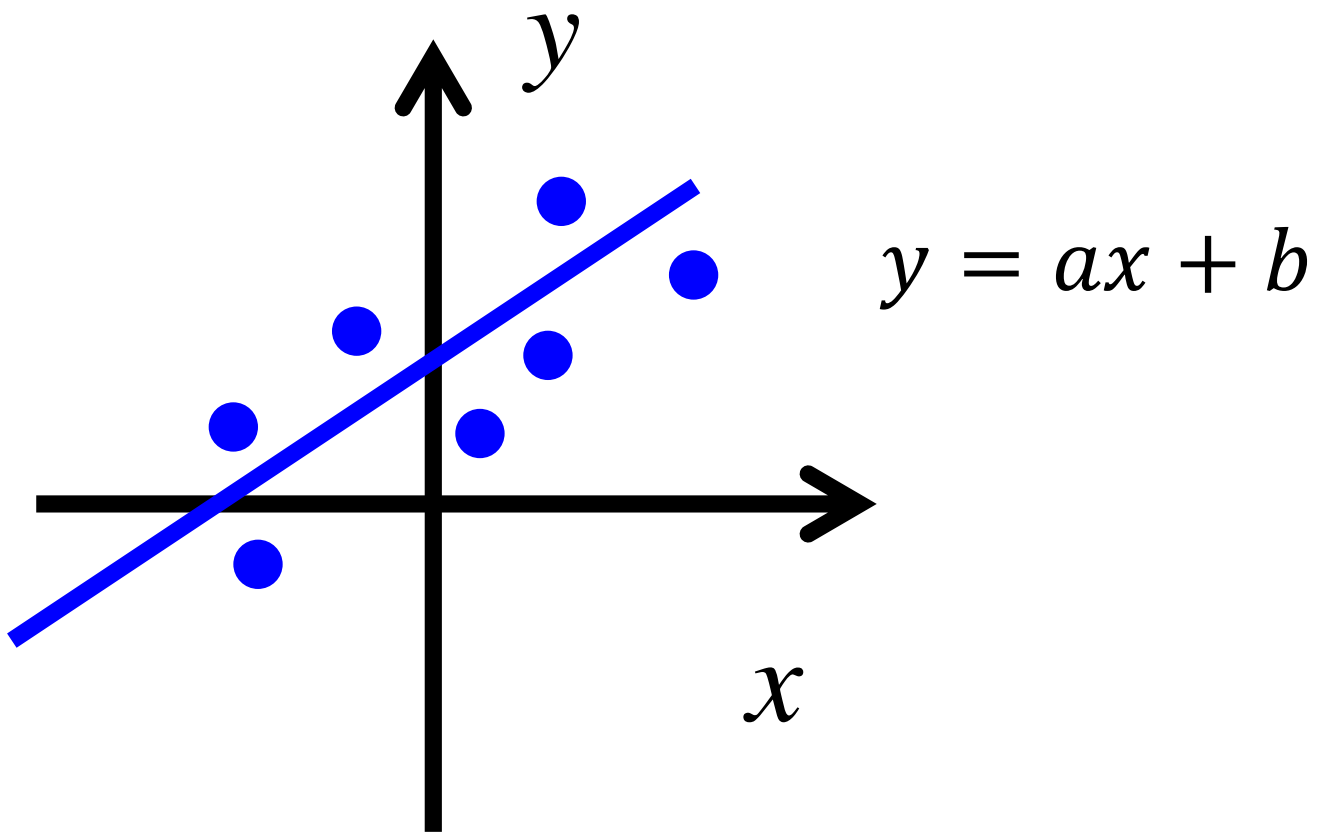






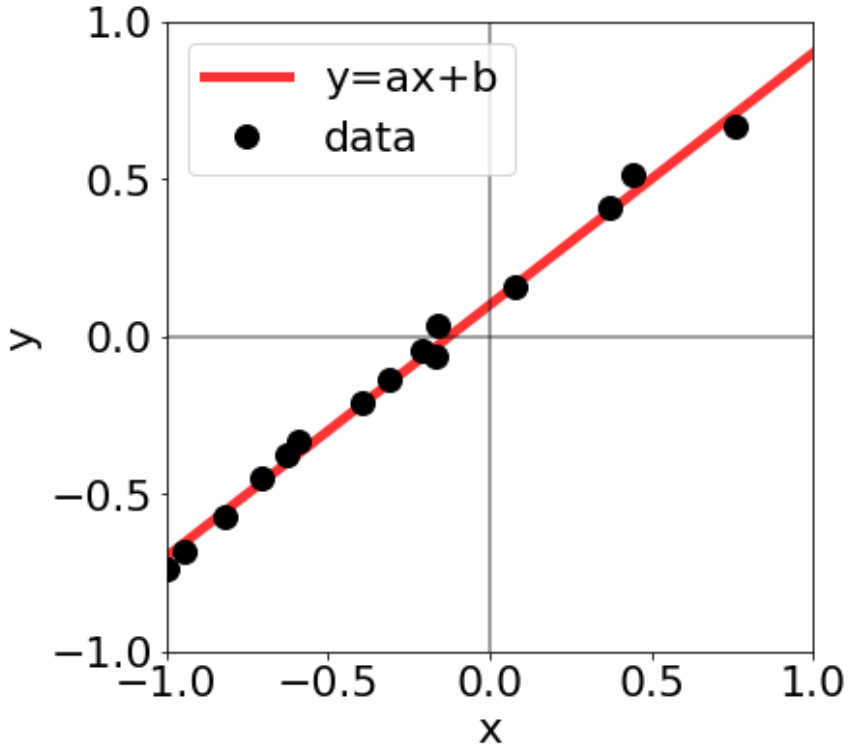
ベイズ計測の利点
$y=ax+b$の取り扱いを通じて

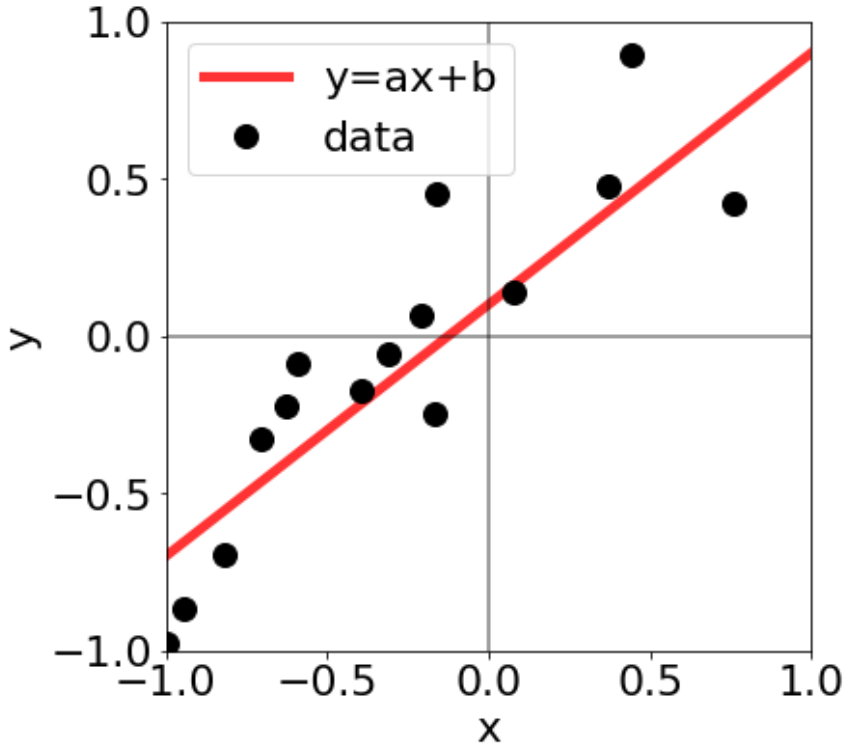
現状でも用いられている最も簡単な例



傾き a ： 系の線形応答、バネ定数、電気伝導度、誘電率

ベイズ計測の利点
$y=ax+b$の取り扱いを通じて





この二つの推定精度の違いを数学的に表現したい

準備として従来手法の最小二乗法

$y=ax+b$の最小二乗法

$$E(a,b)=\frac{1}{N}\sum_{i=1}^{N}(y_i-(ax_i+b))^2$$

二乗誤差$E(a,b)$を最小にするようにパラメータをフィット（最小二乗法）

$\overline{x}=\frac{1}{N}\sum_{i=1}^{N}x_i=0$ とする場合

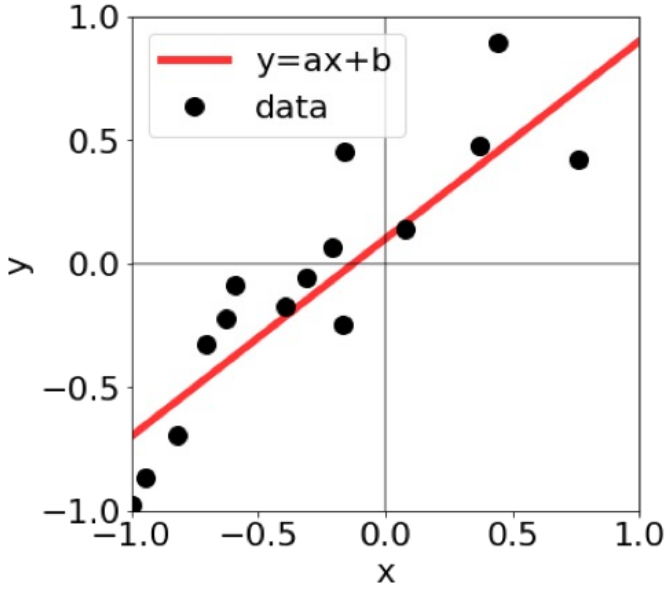
$$E(a,b)=\frac{1}{2}\left(\overline{x^2}\left(a-\frac{\overline{xy}}{\overline{x^2}}\right)^2+(b-\bar{y})^2-\frac{\overline{xy}^2}{\overline{x^2}}-\bar{y}^2+\overline{y^2}\right)$$

$$a_0=\frac{\overline{xy}}{\overline{x^2}}\quad b_0=\bar{y}$$

$$=\mathcal{E}_a(a)+\mathcal{E}_b(b)+E(a_0,b_0)\geqq E(a_0,b_0)$$

平均：$\bar{y}=\frac{1}{N}\sum_{i=1}^{N}y_i$，分散：$\overline{x^2}=\frac{1}{N}\sum_{i=1}^{N}x_i^2$

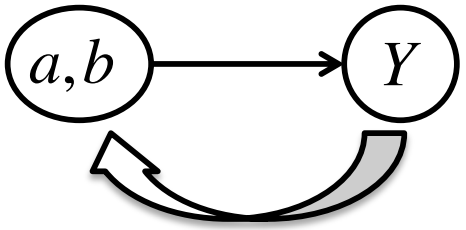
$$\overline{xy}=\frac{1}{N}\sum_{i=1}^{N}x_iy_i$$



ベイズの定理による
神器1: パラメータの事後確率推定 (1/4)

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

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，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

神器1: パラメータの事後確率推定 (2/4)

$$y_i = ax_i + b + n_i$$

$$p(n_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{n_i^2}{2\sigma^2}\right)$$

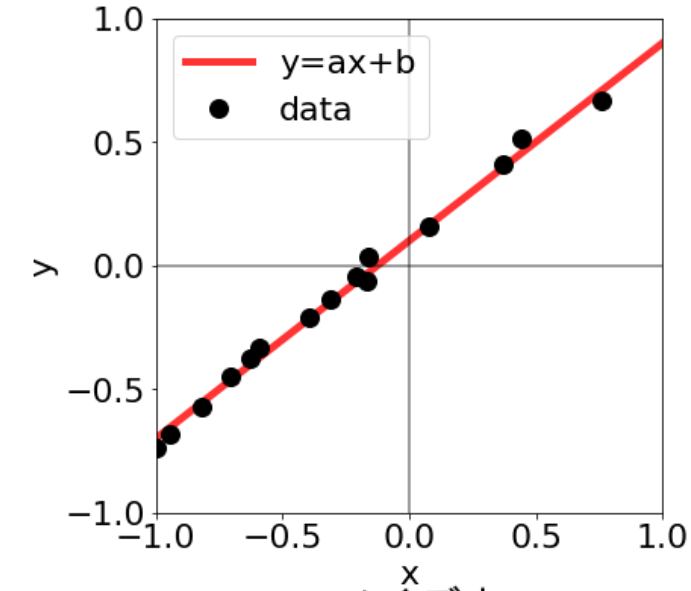
$$p(n_i) = p(y_i|a,b) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - (ax_i + b))^2}{2\sigma^2}\right)$$

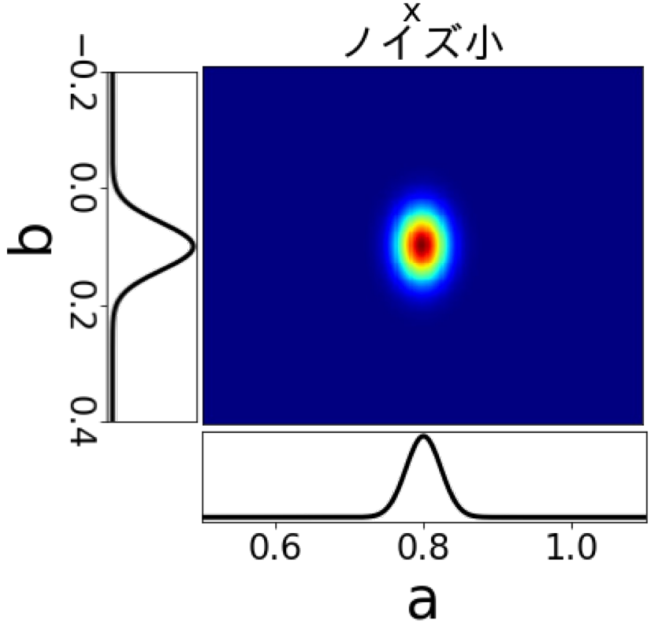
$$\begin{aligned} p(Y|a,b) &= \prod_{i=1}^{N} p(y_i|a,b) \\ &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^N \exp\left(-\frac{\sum_{i=1}^{N}(y_i - (ax_i + b))^2}{2\sigma^2}\right) \\ &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^N \exp\left(-\frac{N}{\sigma^2}E(a,b)\right) \end{aligned}$$

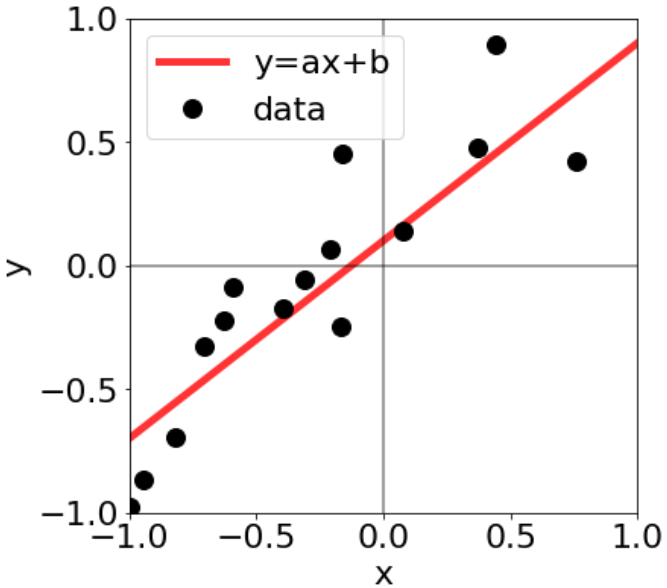
神器1: パラメータの事後確率推定 (3/4)

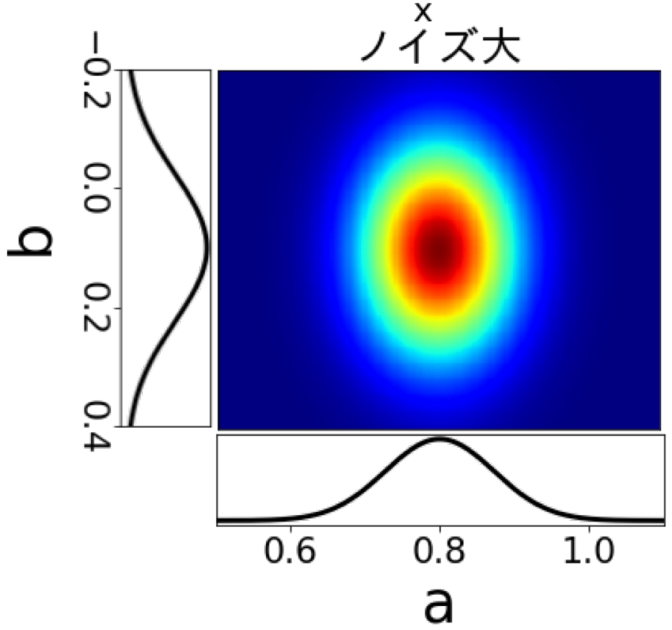
$$
\begin{aligned}
p(a,b|Y) &= \frac{p(Y|a,b)p(a,b)}{p(Y)} \propto p(Y|a,b) \\
&= \exp\left\{-\frac{N}{\sigma^2}\Big(\mathcal{E}_a(a)+\mathcal{E}_b(b)+E(a_0,b_0)\Big)\right\} \\
&\propto \exp\left\{-\frac{N}{\sigma^2}\Big(\mathcal{E}_a(a)+\mathcal{E}_b(b)\Big)\right\} \\
&= \exp\left\{-\frac{N\overline{x^2}}{2\sigma^2}(a-a_0)^2+\frac{N}{2\sigma^2}(b-b_0)^2\right\}
\end{aligned}
$$

神器1: パラメータの事後確率推定 (4/4)

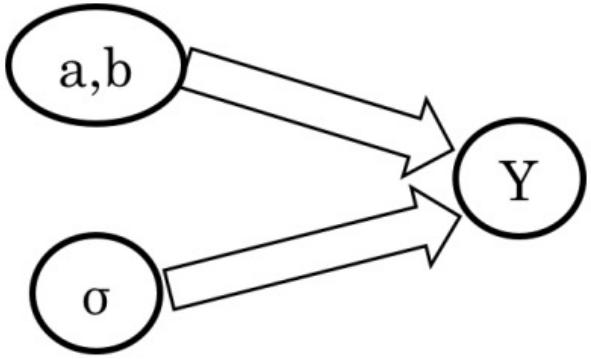








神器1: パラメータの事後確率推定
ノイズ分散推定

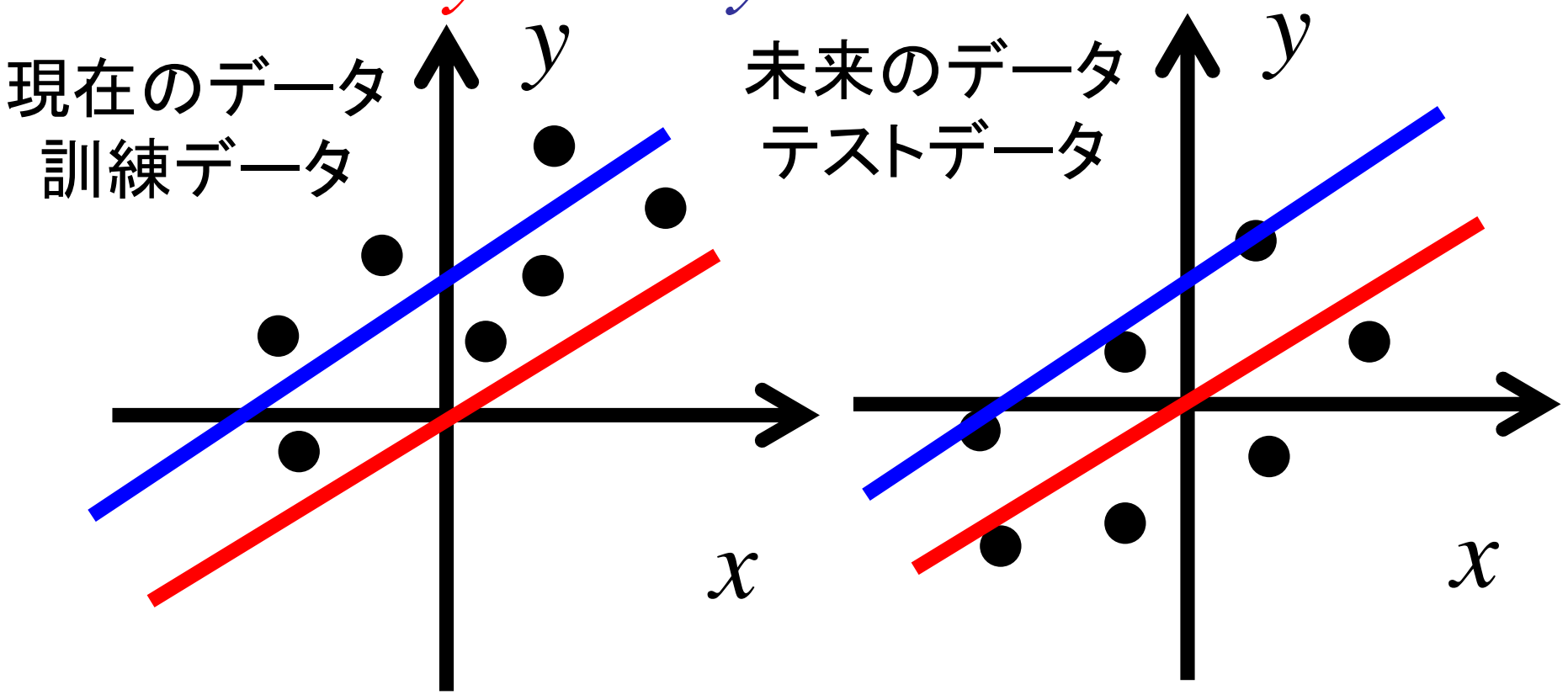


$$
\begin{aligned}
p(\sigma^2|Y) &\propto \int dadb\, p(Y|a,b,\sigma^2) \\
&= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \int dadb \exp\left\{-\frac{N}{\sigma^2}E(a,b)\right\} \\
&= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N \left\{\exp\left(-\frac{N}{\sigma^2}E(a_0,b_0)\right) + \int da \exp\left(-\frac{N\overline{x^2}}{2\sigma^2}(a-a_0)^2\right) + \int db \exp\left(-\frac{N}{2\sigma^2}(b-b_0)^2\right)\right\} \\
&= (2\pi\sigma^2)^{-\frac{N-2}{2}} (N^2\overline{x^2})^{\frac{1}{2}} \exp\left(-\frac{N}{\sigma^2}E(a_0,b_0)\right) \qquad (25)
\end{aligned}
$$

$$
\sigma^2 = \frac{NE(a_0,b_0)}{N-2} = \frac{1}{N-2}\sum_{i=1}^{N}\{y_i - (a_0x_i + b_0)\}^2
$$

問題意識　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

神器2: ベイズ的モデル選択

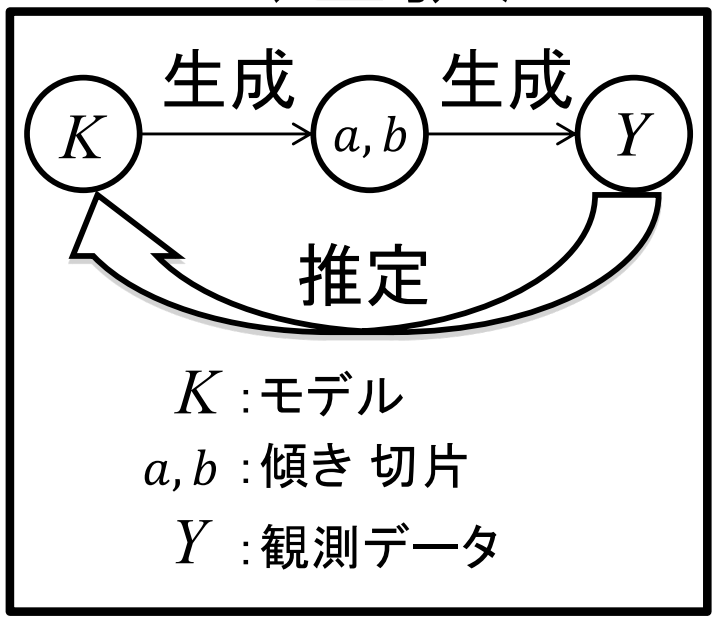
1. 欲しいのは　$p(K \mid Y)$

2. θがないぞ

3. $p(K,\theta,Y)$ の存在を仮定

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

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



4. 無駄な自由度の系統的消去: 周辺化，分配関数

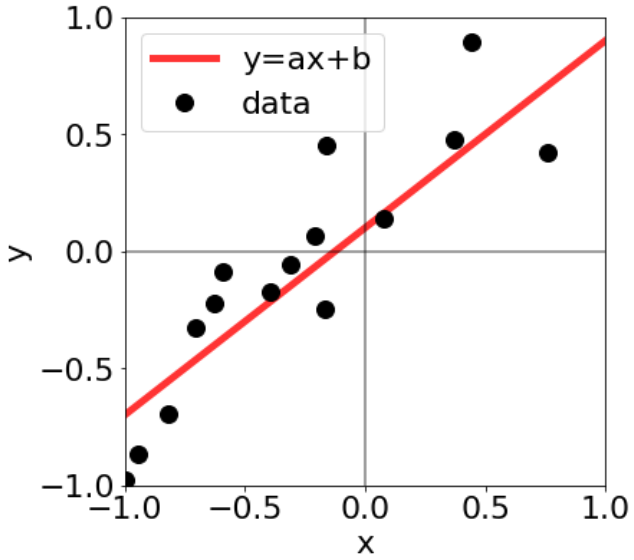
$$p(K,Y) = \int p(K,\theta,Y)\,d\theta$$

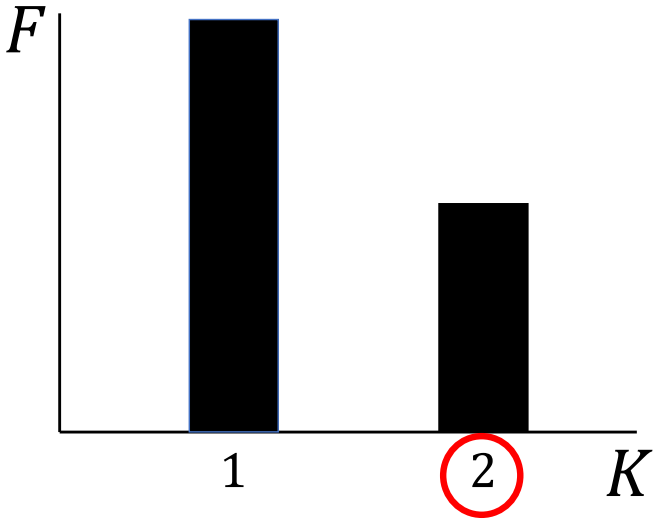
$$p(K \mid Y) = \frac{p(Y \mid K)p(K)}{p(Y)} \propto p(K)\int \exp\left(-nE(\theta)\right)p(\theta)\,d\theta$$

$$F(K) = -\log \int \exp\left(-nE(\theta)\right)p(\theta)\,d\theta$$

自由エネルギーを最小にするモデル K を求める.

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

まとめ: ベイズ計測三種の神器 $y=ax+b$の解析取り扱いを通じて

- 従来の最小二乗法
  - 1.物理パラメータの点推定
- ベイズ計測

1. 物理パラメータの確率分布推定
2. データからのベイズ的モデル選択
3. ベイズ統合: 今回は説明を省略
   - 今回は説明しない

内容

- 本講演の目的
- データ駆動科学
- 材料/デバイスの機能発現の3ステップモデル
- 情報数理基盤のベイズ推論とスパースモデリング
- ベイズ推論を計測科学に適用したコンパクトな体系のベイズ計測
  - ベイズ計測三種の神器
- $y=ax+b$の線形回帰、スペクトル分解を述べ、さらに機能発現の3ステップモデルの例として大久保研との共同研究を紹介する。
- 材料工学の展望

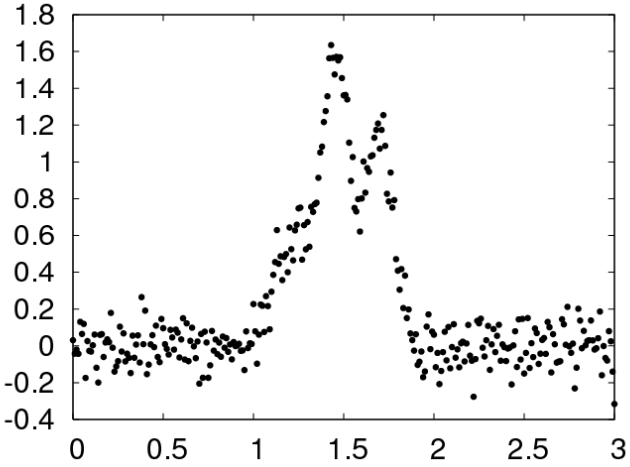
ベイズ推論とベイズ計測は違う

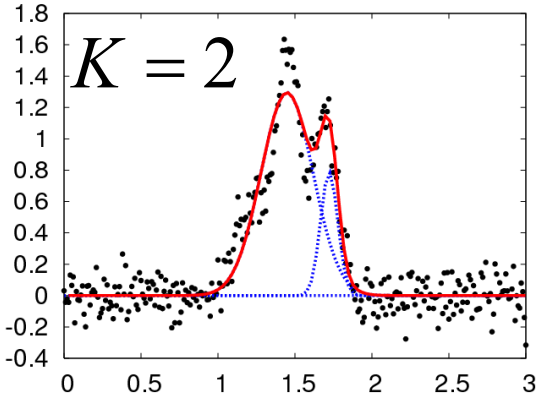
分厚い本を読む必要はない．分厚い部分のほとんどは近似アルゴリズムの説明

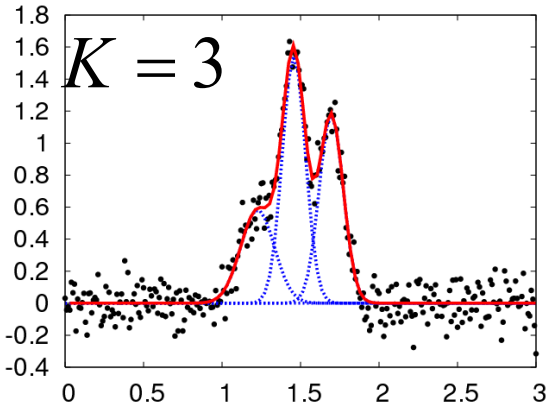
ベイズ計測の習得法

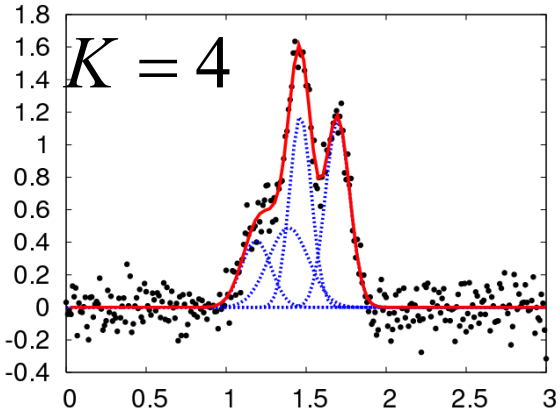
1. $y=ax+b$への解析的計算の適用
2. レプリカ交換モンテカルロ法の導入
   $y=ax+b$への数値計算の適用
   ベイズ的スペクトル分解
3. 各課題に取り組む

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

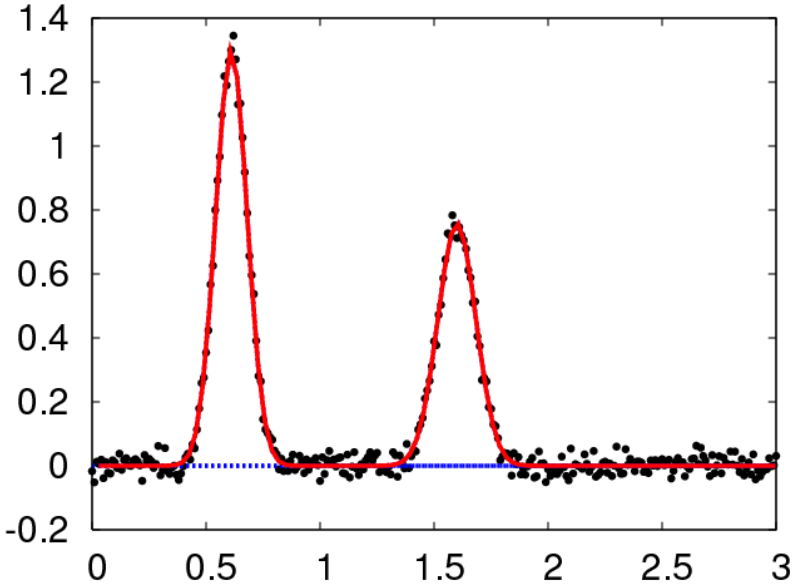
ガウス関数(基底関数)の足し合わせにより，スペクトルデータを近似

観測データ：$D = \{x_i, y_i\}_{i=1}^{n}$

x_i：入力　y_i：出力

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

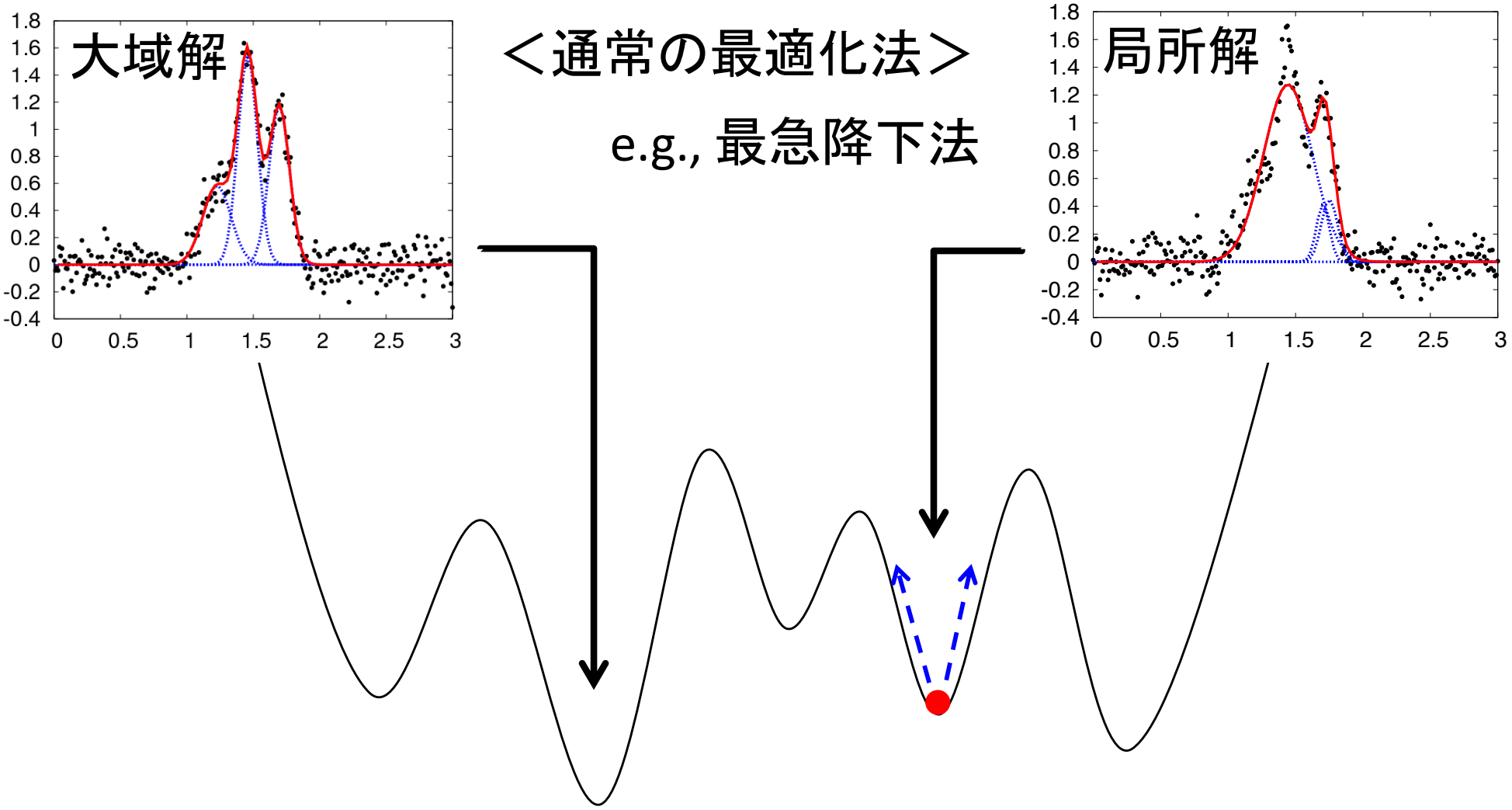
$$\theta = \{a_k, b_k, \mu_k\} \quad k = 1, \cdots, K$$



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）

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

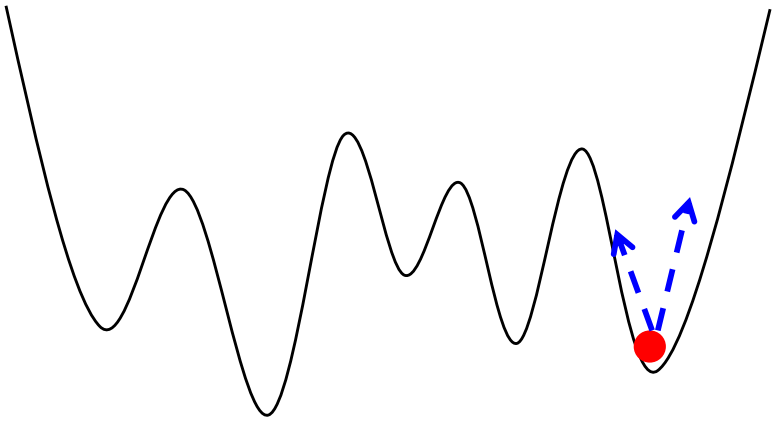
スペクトル分解従来法: 最急降下法
誤差関数は局所解を持つ



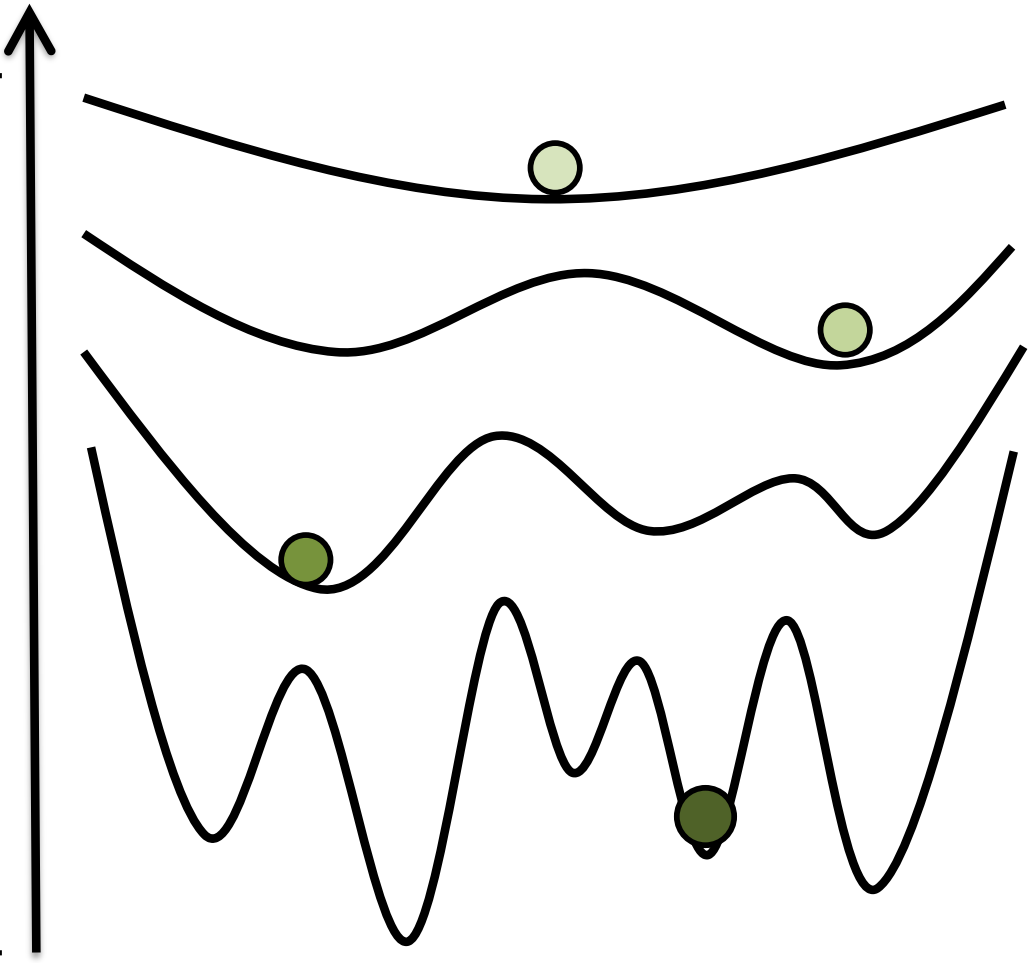
モンテカルロ法の適用
レプリカ交換モンテカルロ法

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: 自由エネルギーの導入

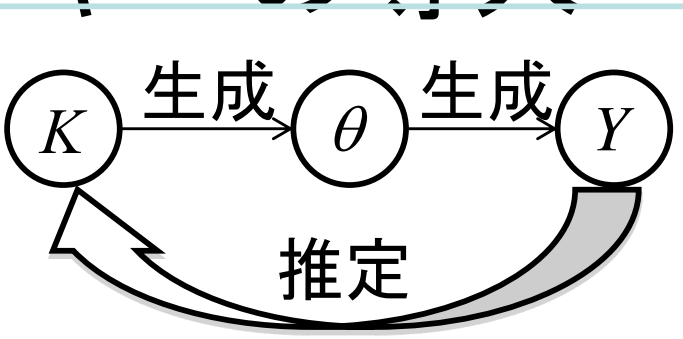
1. 欲しいのは　$p(K|Y)$

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

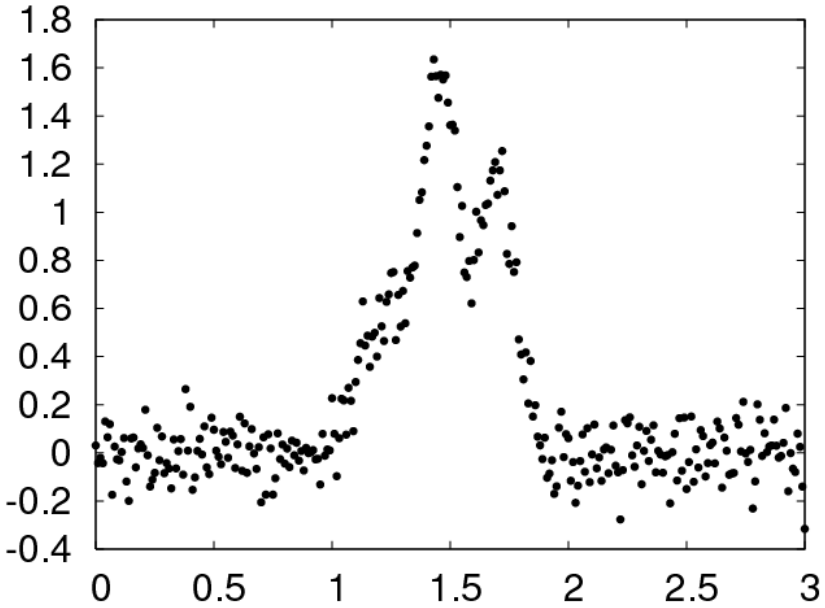
$$p(K,Y) = \int p(K,\theta,Y)d\theta$$

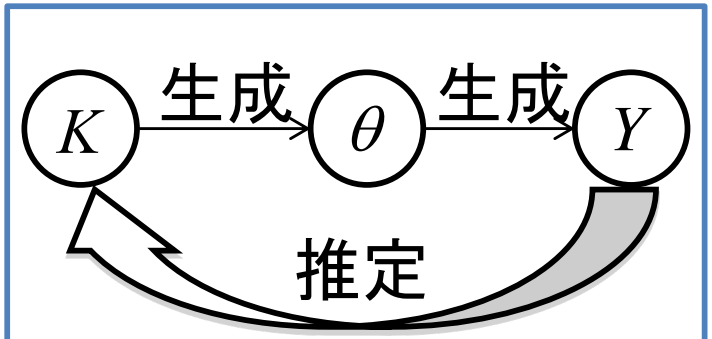
$$p(K|Y) = \frac{p(Y|K)p(K)}{p(Y)} \propto p(K)\int \exp(-nE(\theta))p(\theta)d\theta$$

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

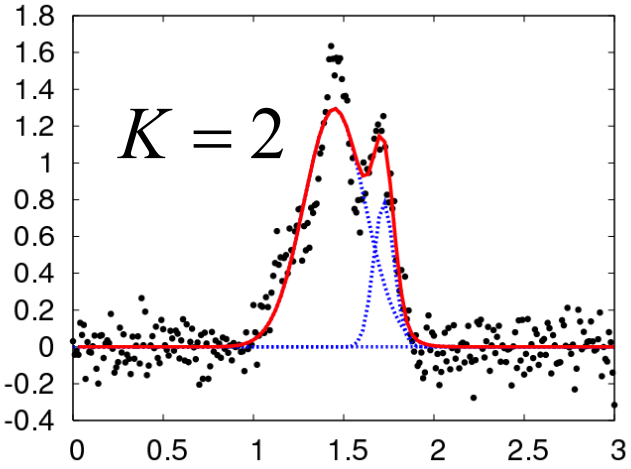
自由エネルギーを最小にする個数　Kを求める.

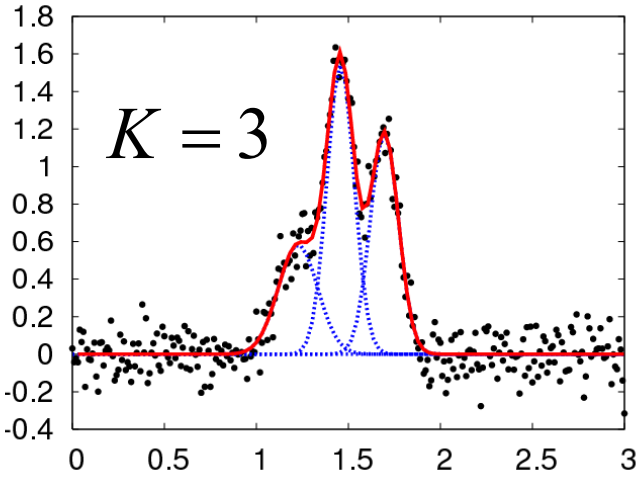
モデル選択: Kをどう選ぶか

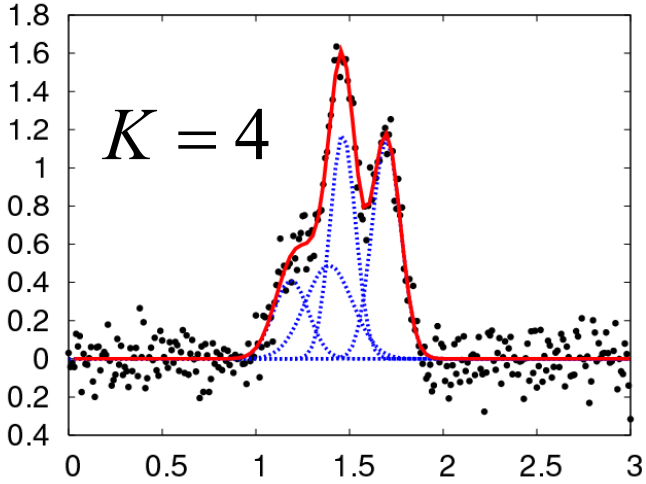




K :ピーク個数
θ :ピーク位置など
Y :観測スペクトル



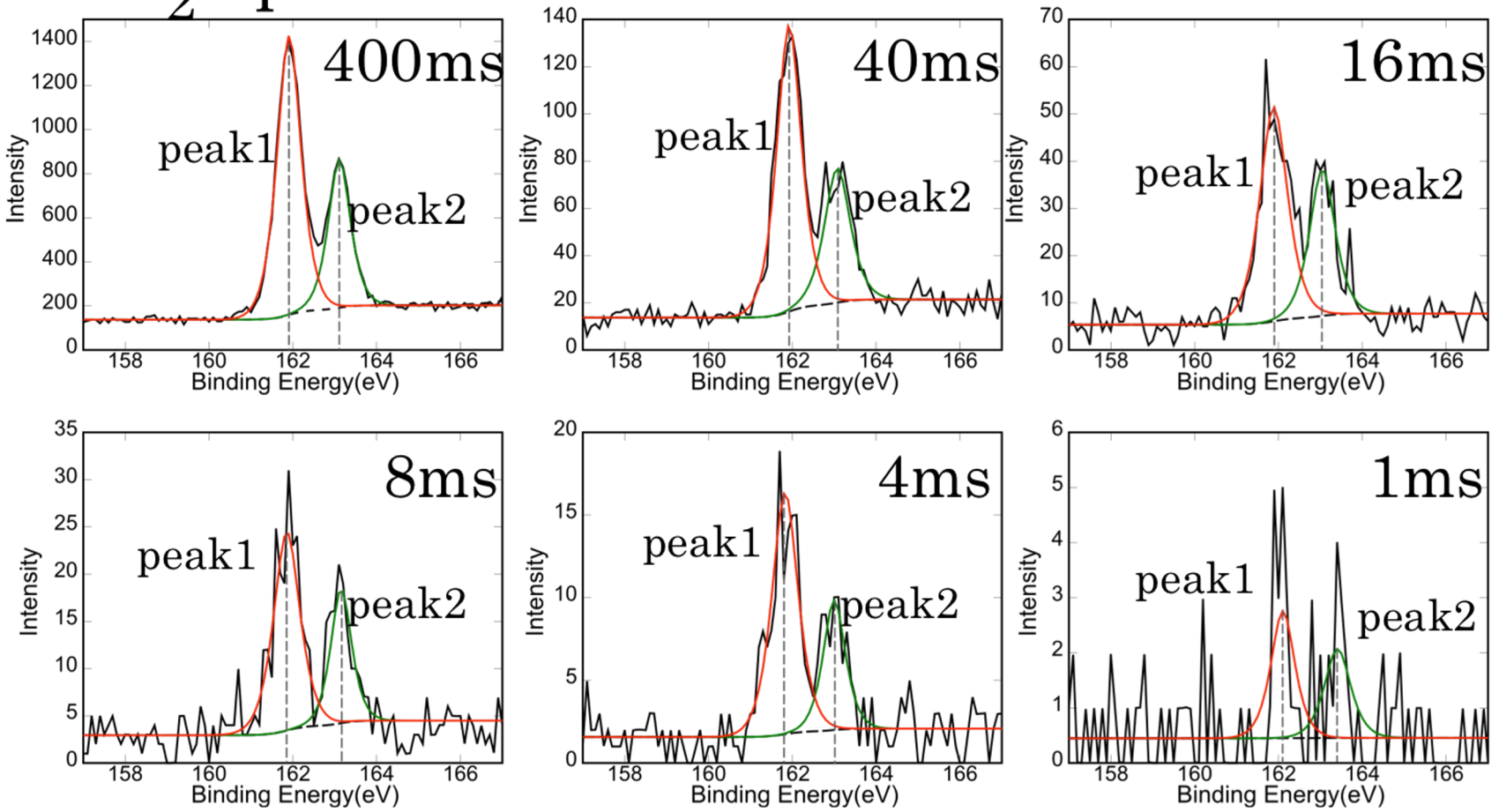




Nagata, Sugita and Okada, Bayesian spectral deconvolution with the exchange Monte Carlo method, *Neural Networks* 2012

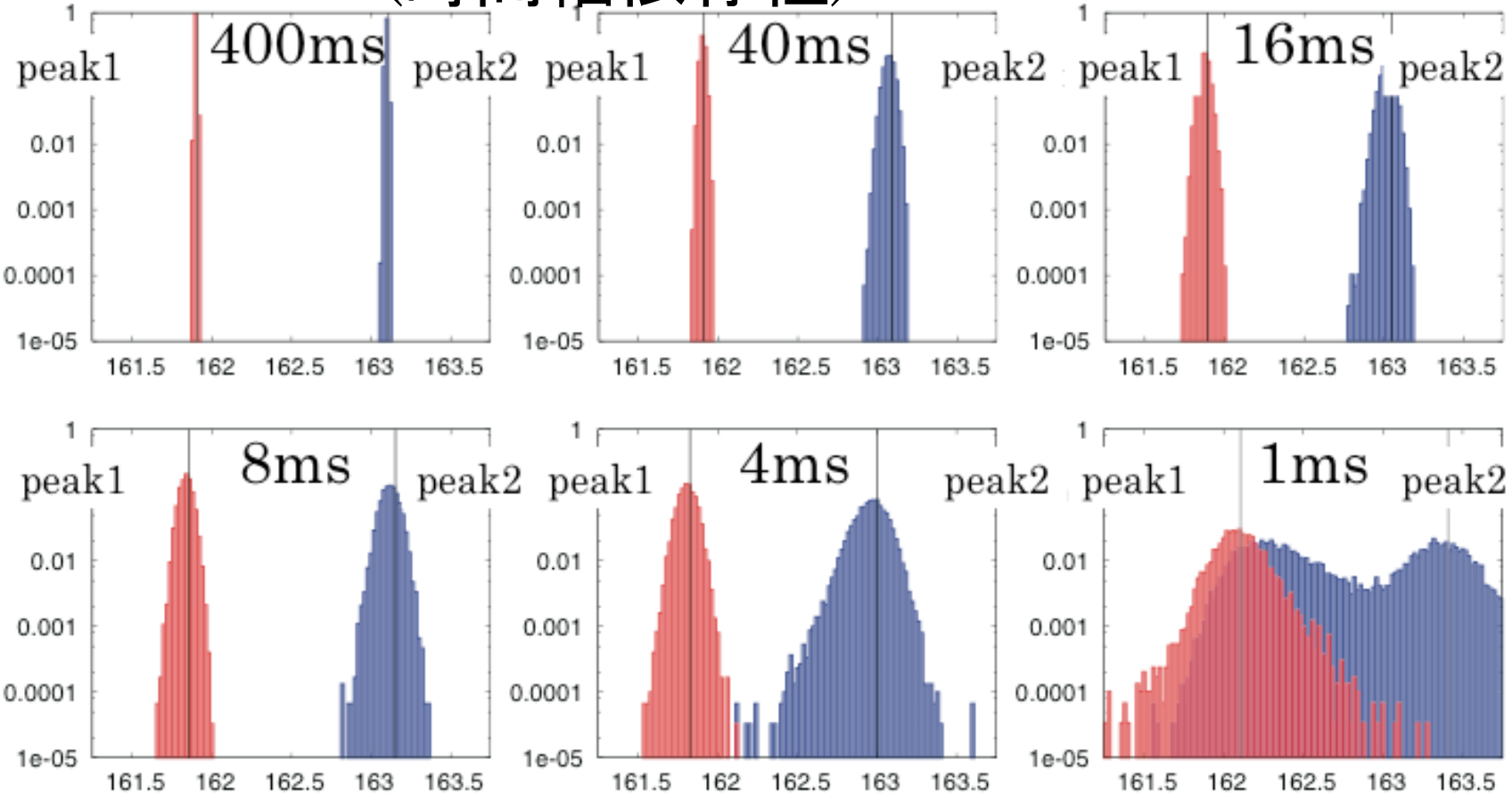
計測限界の理論的取り扱い (4/9)
(Nagata *et al.* 2019)

MoS_2 2p （時間幅依存性）



計測限界の理論的取り扱い (6/9)
(Nagata *et al.* 2019)

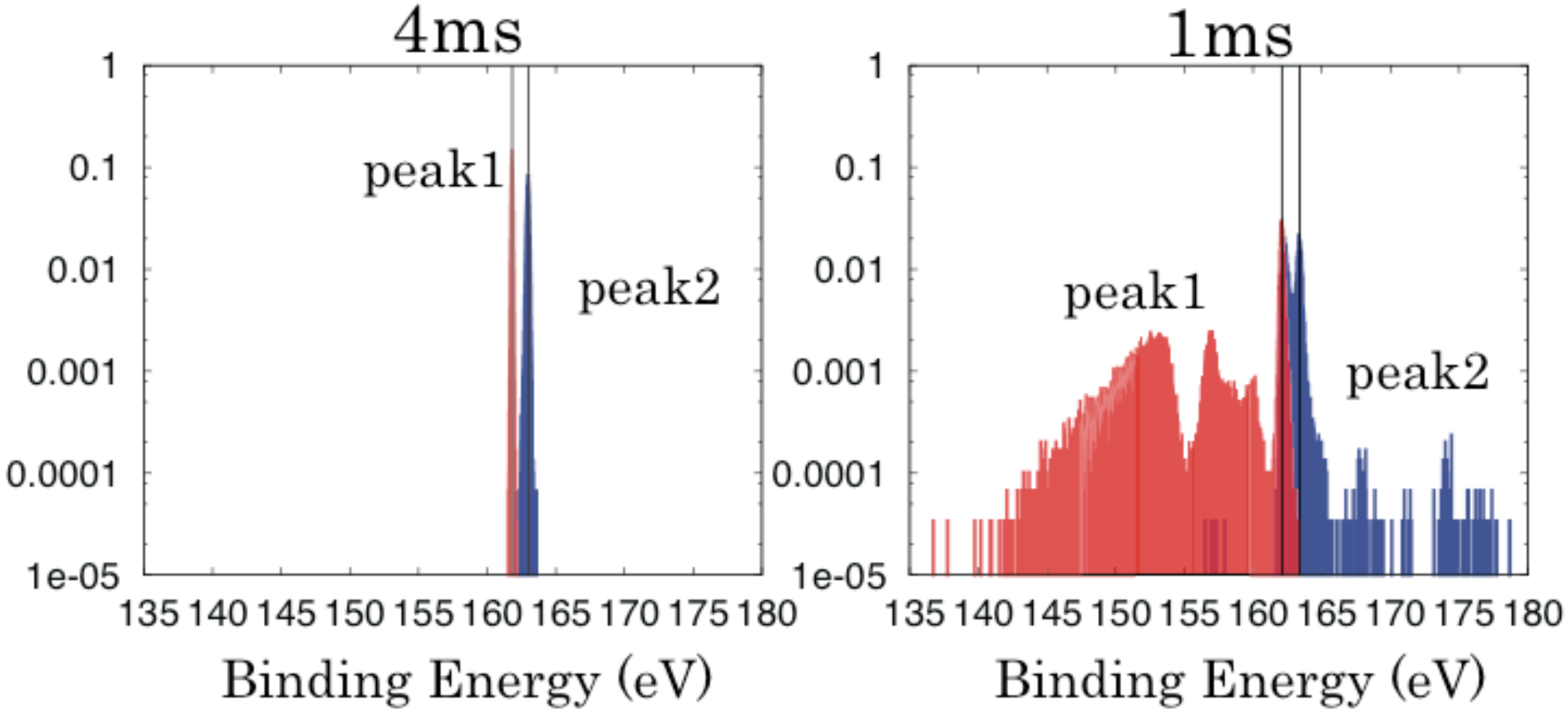
ベイズ計測: ベイズ推論によって,
ピーク位置のベイズ事後確率を計算
(時間幅依存性)



計測限界の理論的取り扱い (7/9)
(Nagata *et al.* 2019)

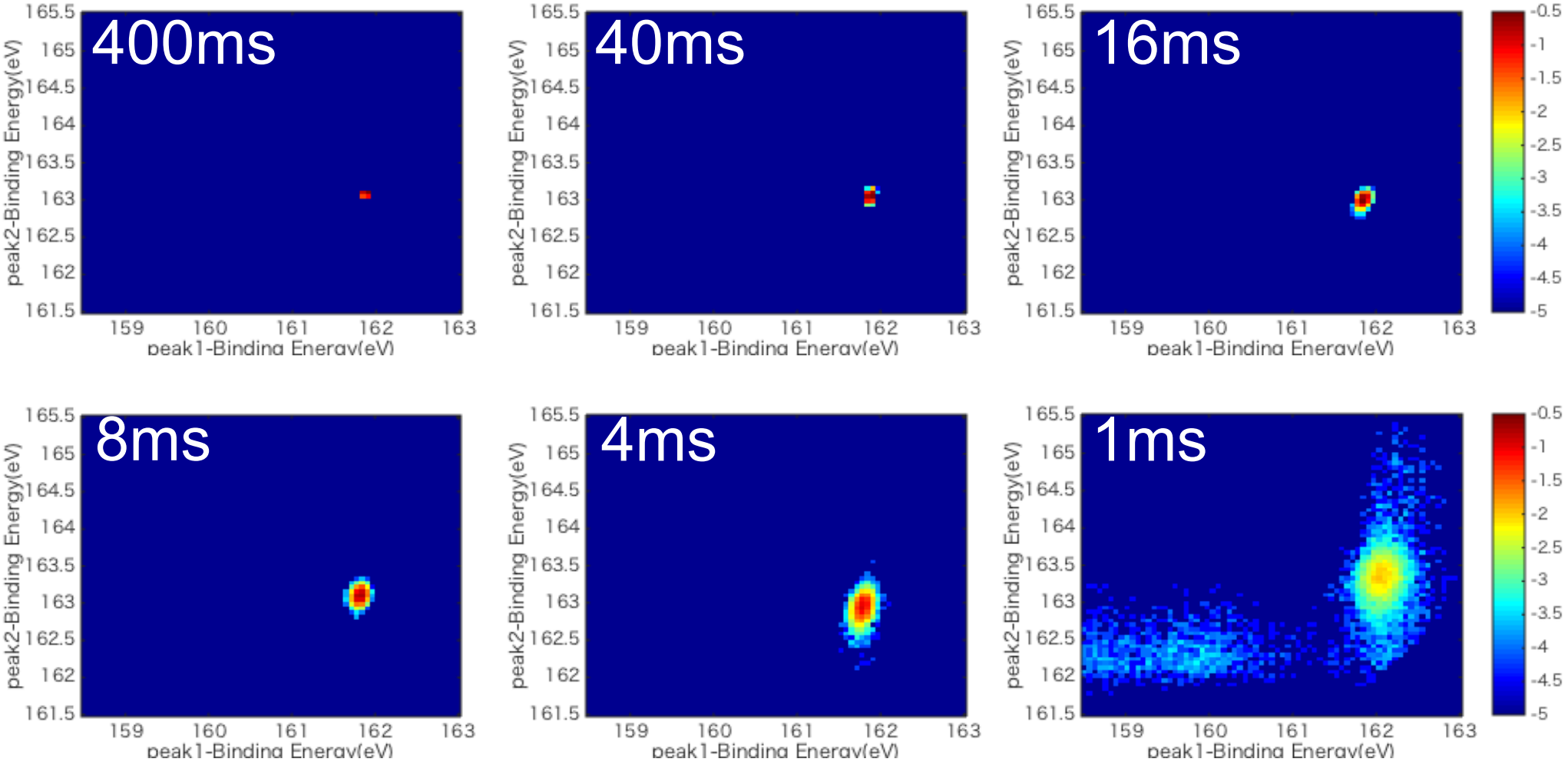
ベイズ計測: ベイズ推論によって,
ピーク位置のベイズ事後確率を計算

(時間幅依存性)



計測限界の理論的取り扱い (8/9)
(Nagata *et al.* 2019)

MoS_2 2p



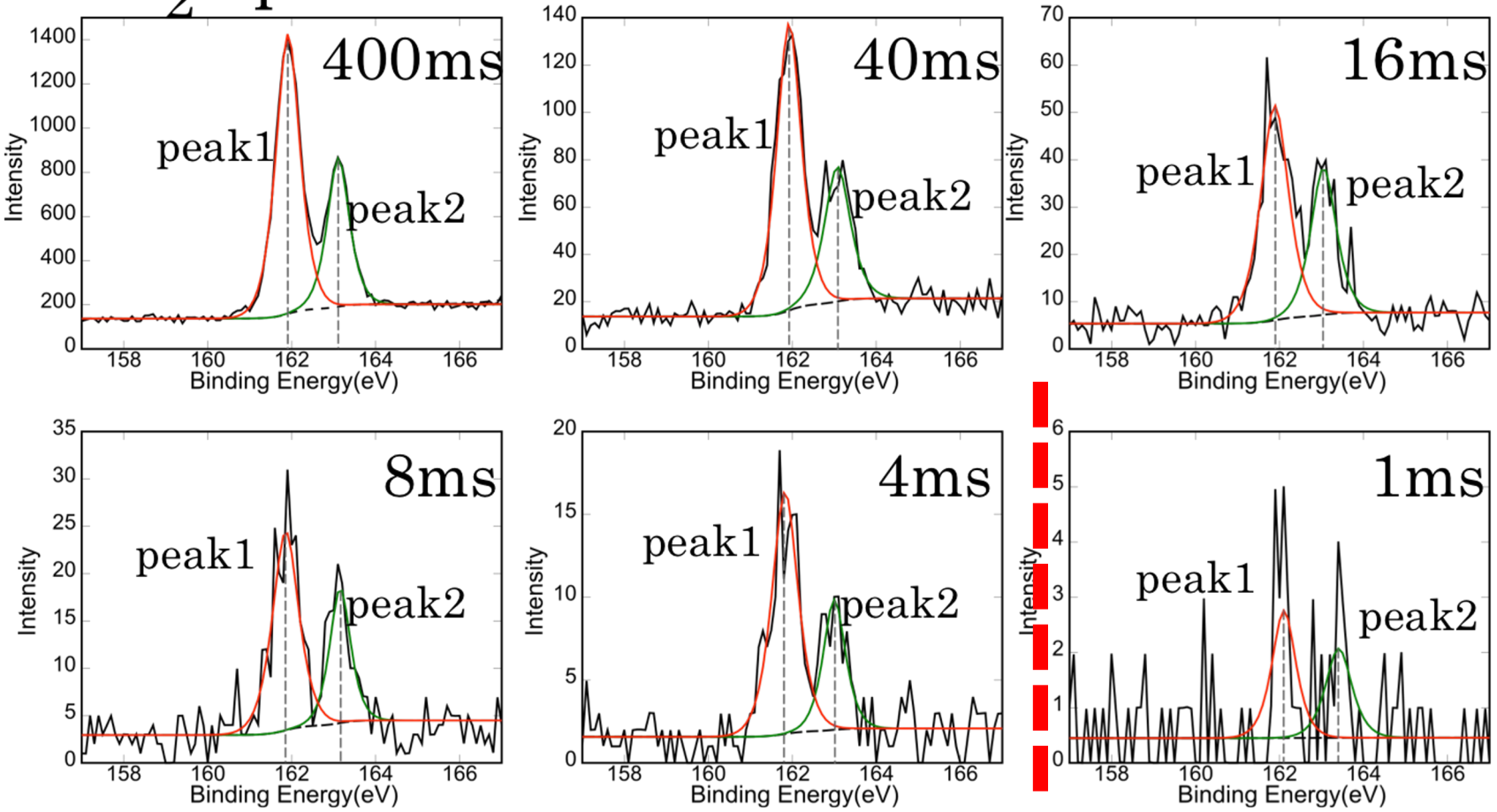
ここまで

計測限界の理論的取り扱い (9/9)
(Nagata *et al.* 2019)

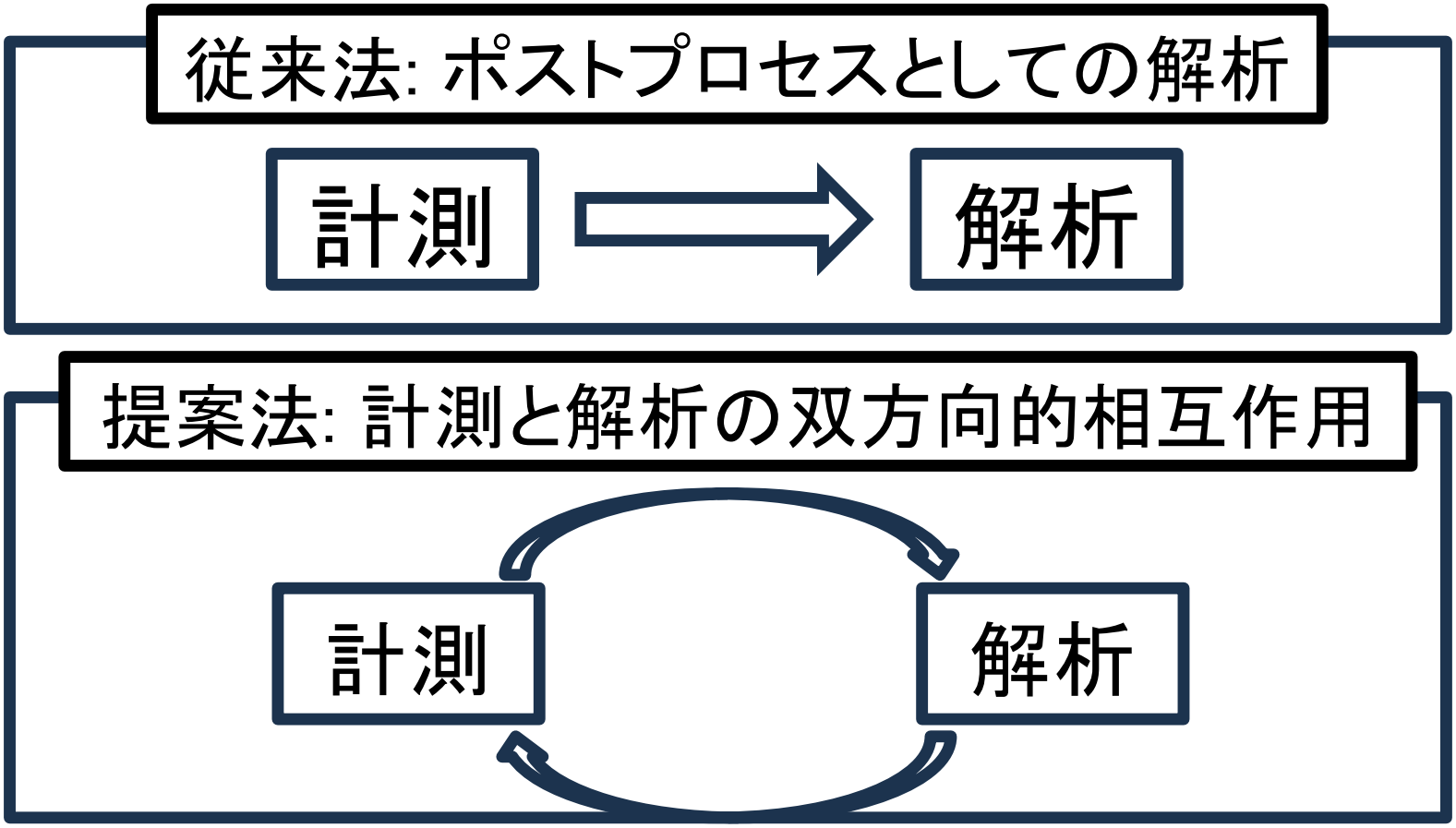
ベイズ計測: ベイズ推論によって，ピーク位置のベイズ事後確率を計算

戦略目標: 計測限界を定量的に評価できる枠組みの提案

MoS_2 2p （時間幅依存性）



Take Home Messege
ベイズ計測のインパクト

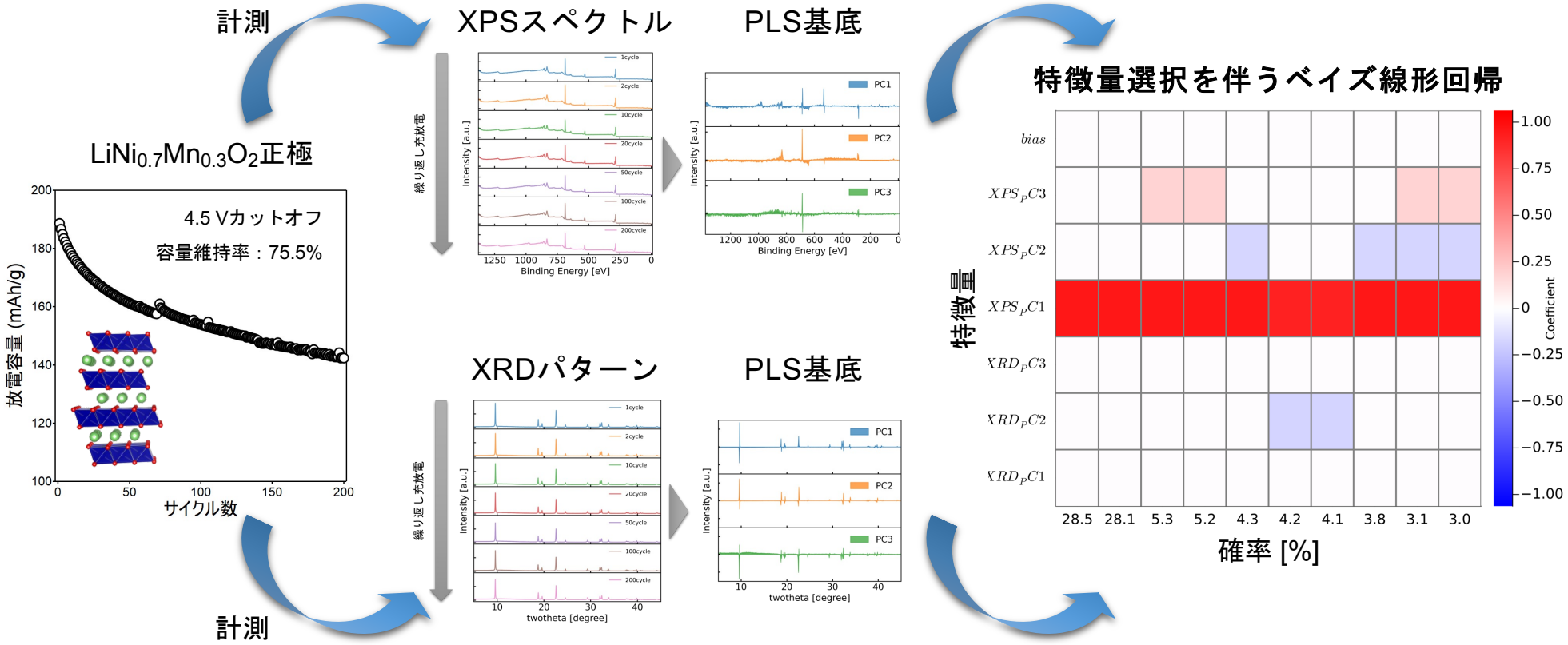


放射光科学のパラダイムシフト
キャリアップの千載一遇のチャンス

内容

- 本講演の目的
- データ駆動科学
- 材料/デバイスの機能発現の3ステップモデル
- 情報数理基盤のベイズ推論とスパースモデリング
- ベイズ推論を計測科学に適用したコンパクトな体系のベイズ計測
  - ベイズ計測三種の神器
- $y=ax+b$の線形回帰、スペクトル分解を述べ、さらに機能発現の3ステップモデルの例として大久保研との共同研究を紹介する。
- 材料工学の展望

リチウムイオン電池の正極材料に対する解析事例

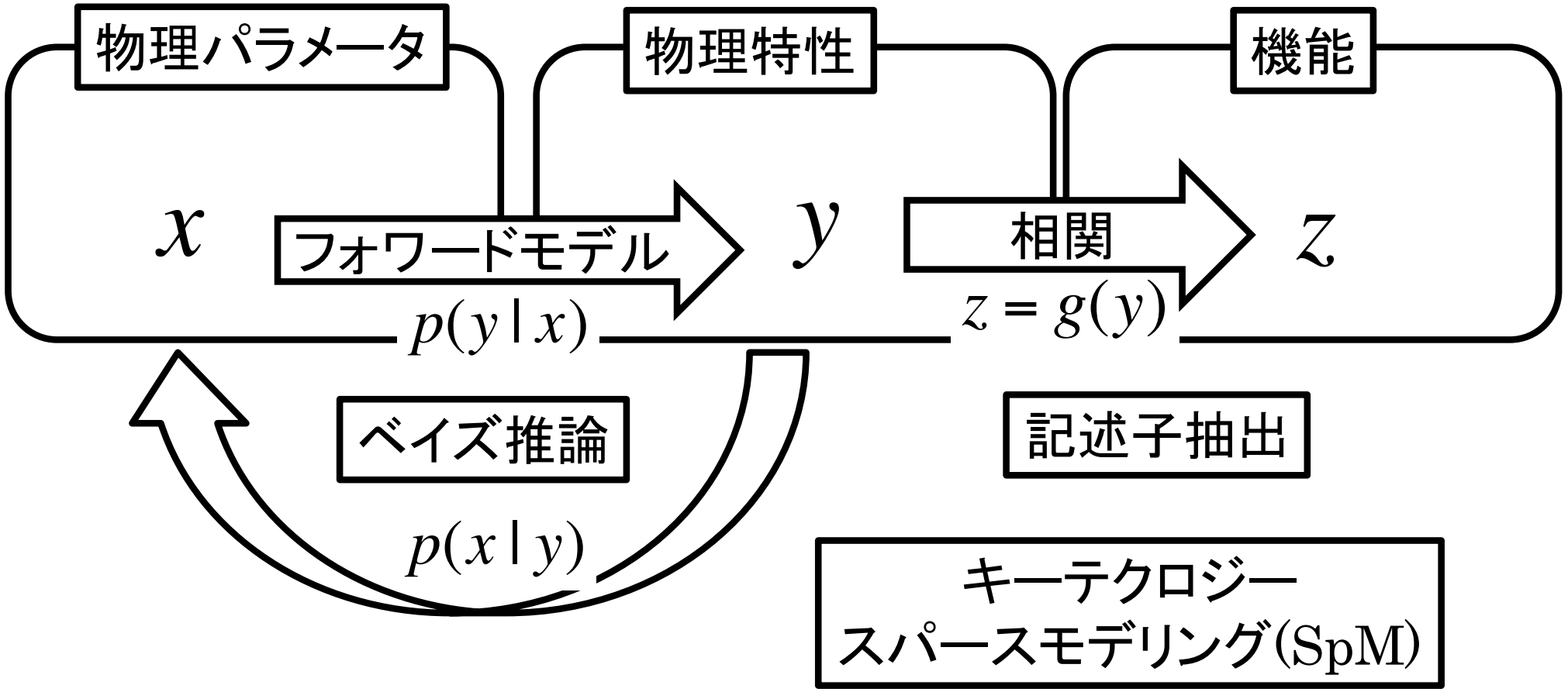


✓ サイクル劣化に対する説明性：XRD（バルク状態）< XPS（表面状態）

PLS(部分的最小二乗回帰 Partial Least Squares Regression

1. 入力ベクトル(XPSやXRD)に主成分分析
2. 目的変数が主成分の線形関数
3. XPSとXRDの主成分を全て用いて、その後は全状態探索型スパースモデリングで、どの主成分を用いるかを決める
4. 結果として、XRDの主成分のみが残った
5. 機能発現の3ステップモデルのロールモデル的研究

機能発現の3ステップモデル



第15回NIMS フォーラム 2015年10月7日（水）
マテリアルズ・インフォマティクスとは何か-物質材料科学とデータ駆動科学-
https://mns.k.u-tokyo.ac.jp/pdf/2015nims.pdf

Igarashi, Nagata, Kuwatani, Omori, Nakanishi-Ohno, and Okada “Three levels of data-driven science” International meeting on High-dimensional Data-Driven Science (HD3-2015), *Journal of Physics: Conference Series*, 699 (2016) 012001(2016)

内容

- 本講演の目的
- データ駆動科学
- 材料/デバイスの機能発現の3ステップモデル
- 情報数理基盤のベイズ推論とスパースモデリング
- ベイズ推論を計測科学に適用したコンパクトな体系のベイズ計測
  - ベイズ計測三種の神器
- $y=ax+b$の線形回帰、スペクトル分解を述べ、さらに機能発現の3ステップモデルの例として大久保研との共同研究を紹介する。
- 材料工学の展望

ベイズ計測の適用例

東京大学 岡田研究室

- 事後分布推定 and/or モデル選択
  1. スペクトル分解
  2. X線光電子放出スペクトル(XPS)
  3. X線吸収スペクトル(XAS)
  4. メスバウアー分光
  5. X線小角散乱スペクトル
  6. NMR
  7. 中性子非弾性散乱スペクトル
  8. 比熱
  9. 帯磁率
- ベイズ統合
  1. XPSとXAS
  2. 比熱と帯磁率

熊本大学 赤井研究室

- 事後分布推定 and/or モデル選択
  1. フォトルミネッセンススペクトル

熊本大学 水牧研究室

- 事後分布推定 and/or モデル選択
  1. XRD

ベイズ計測の枠組みは様々な計測データに適用できる

（シミュレーションはノートPCで実行できる場合が多い）

SPring-8全ビームラインベイズ化計画

赤色BLが共用BL(JASRI担当):
計26本、今年度中に半分のBLで達成

全BL本数:62本

情報と放射光研究者のマッチング

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

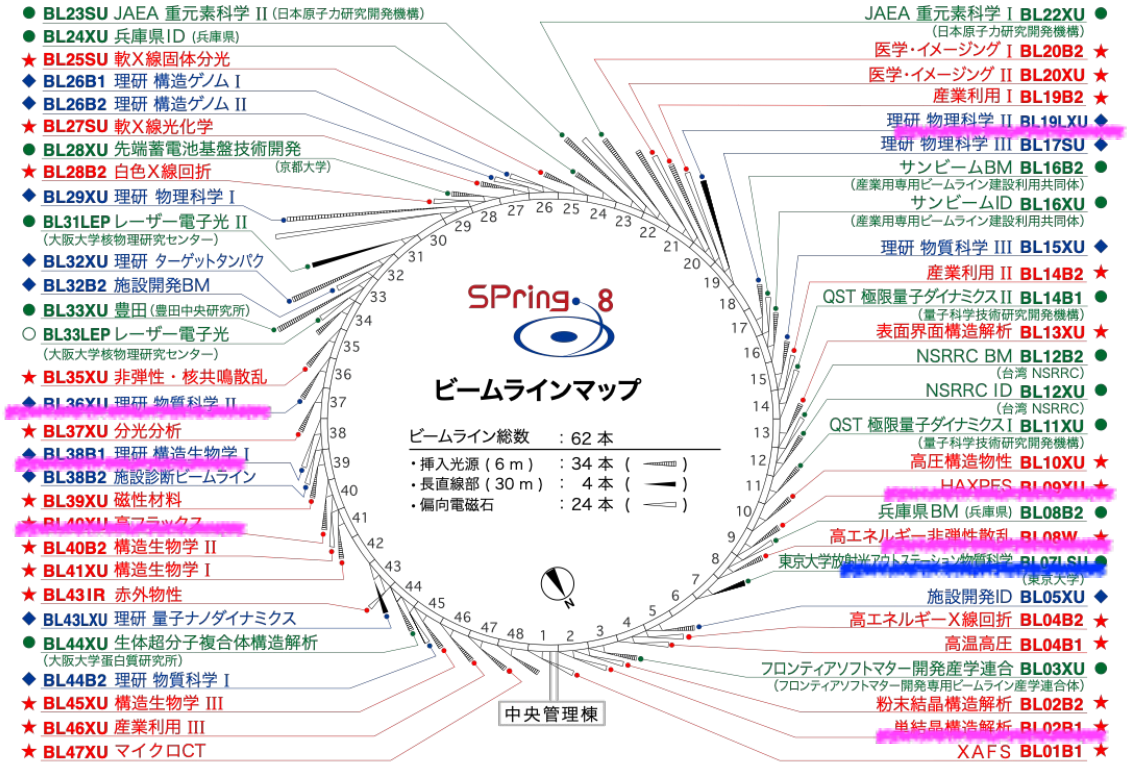
放射光ユーザーへの展開

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

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

SPring-8全ビームラインベイズ化計画

敬称略



赤色BLが共用BL(JASRI担当): 計26本

今年(2024)年度中に14BL/26の
ベイズ化が完了

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

情報と放射光研究者のマッチング

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

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

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年度中に14BL/26のベイズ化完了

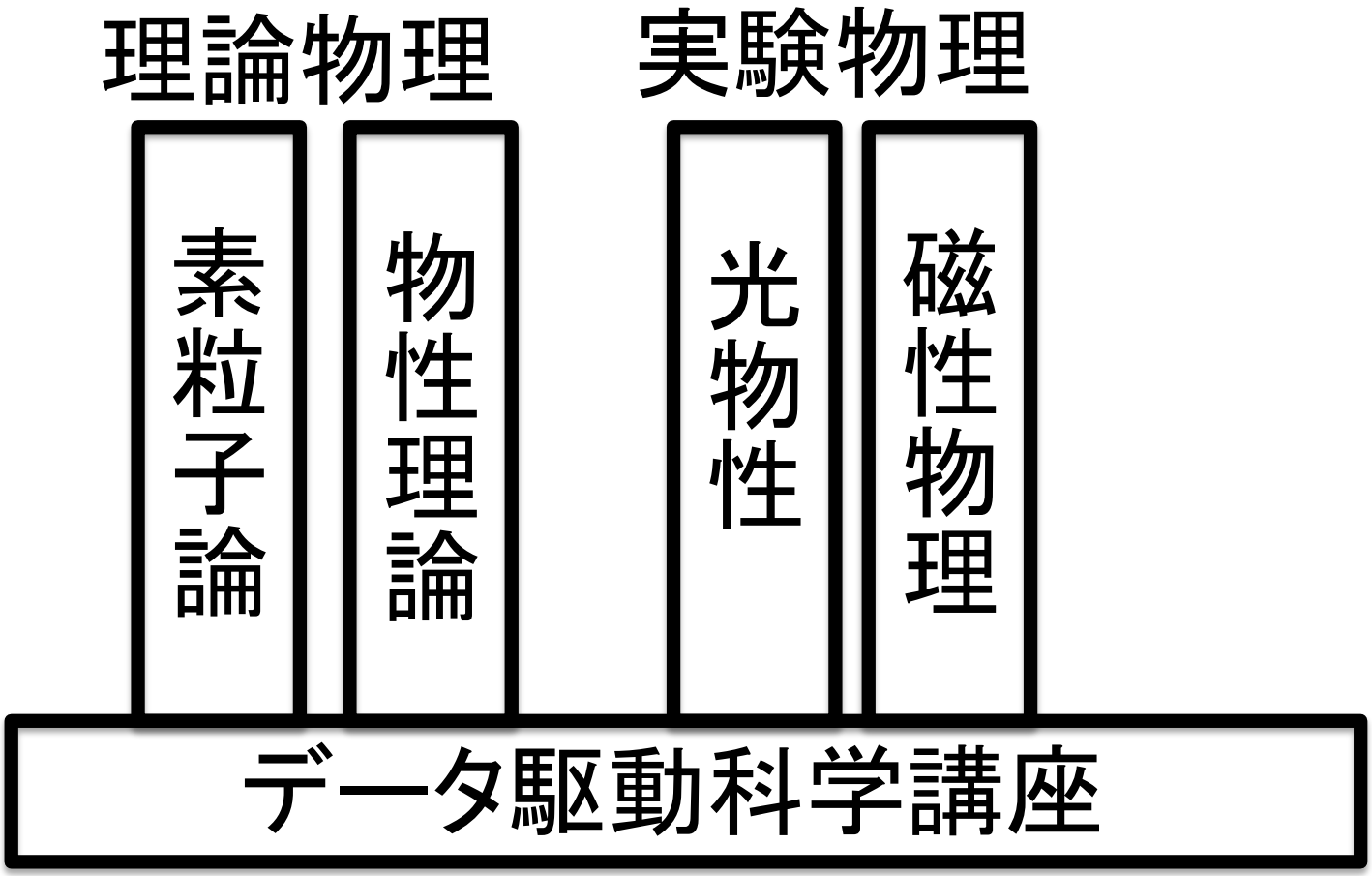
SPring-8全ビームライン
ベイズ化計画の波及効果

- 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日シンポジウム
- あいちSR: 2023年10月30日シンポジウム
- 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日シンポジウム
  －大盛況: ベイズ計測の国際展開
- 2024年9月6日SPring-8シンポジウム2024
- 佐賀LS: 2024年10月16日シンポジウム

データ駆動科学とと民間企業
サイバーフィジカルシステムの観点から

- 実験計測系などを具体的に特定して、そこから得られるデータの背後にある構造や知見を、現代的な機械学習アルゴリズムを用いて抽出する。
- 対象はセンサー系と情報処理系が強くカップルしたサイバーフィジカルシステム(CPS)と考えられる。
- 民間企業では、機械学習のアルゴリズム開発をする事はまれであり、顧客や開発のニーズに応じたCPSを開発することが多い。
- つまりデータ駆動科学はCPSの観点からも、民間企業就職に有利な学問分野である。

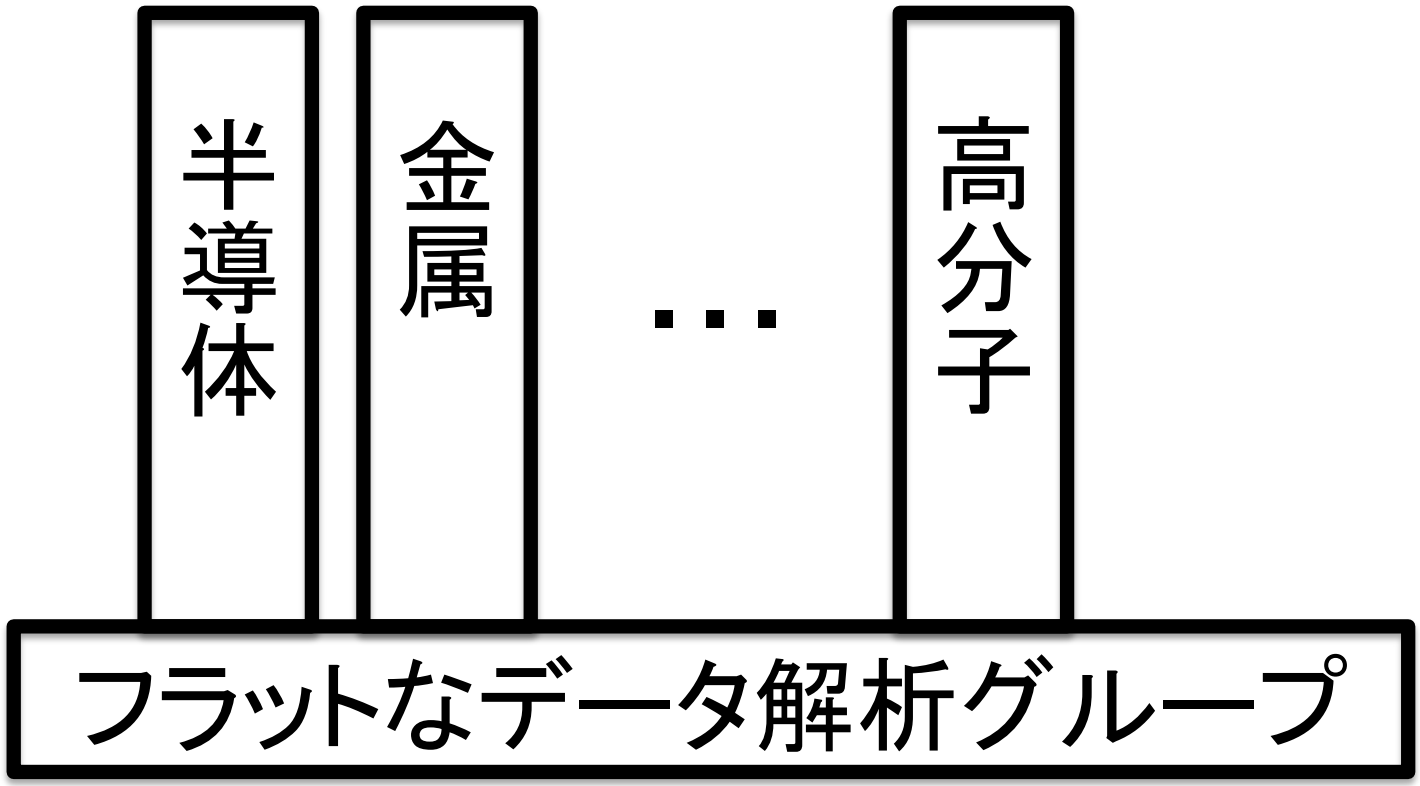
ベイズ計測とキャリアパス (1/3)
アカデミア編
物理学科/各学科に
データ駆動科学一講座導入



ベイズ計測とキャリアパス (2/3)

民間企業編

R&D組織のフラット化と人材の流動化



ある縦組織がリストラされても、フラットなデータ解析グループに沿って、他の縦組織にソフトランディングでき、人材の流動化が加速される。

ベイズ計測とキャリアパス (3/3)
放射光科学関連

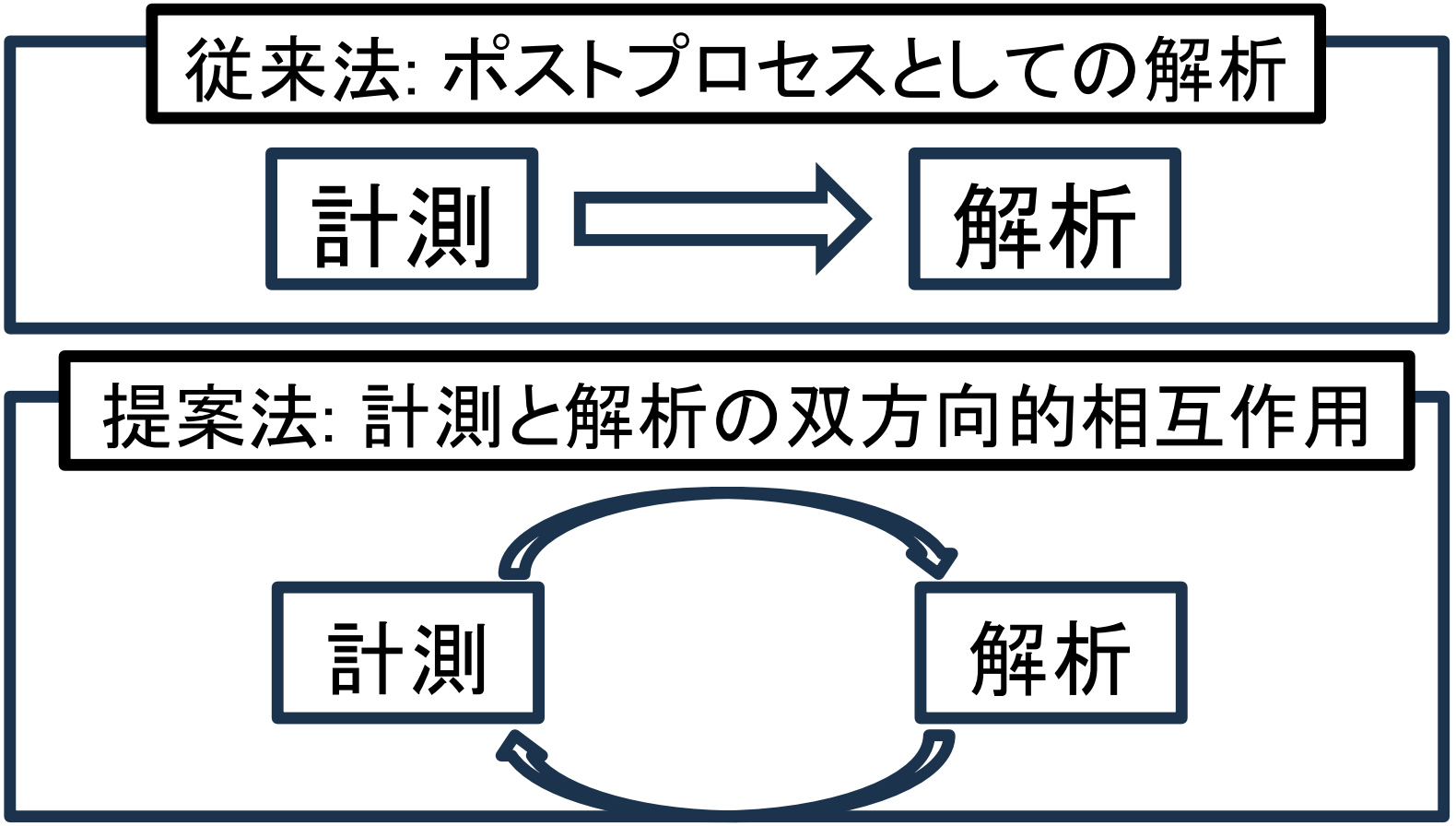
- SPring-8全ンビームラインベイズ化計画により、アカデミアだけでなく民間企業でもベイズ計測を取り入れる動きが急速に進んでいる。
- SPring-8/JASRIのような従来の計測サービス部門シーズと民間企業のニーズの乖離
- そこを埋めるために、新たな市場が形成

  －a.s.ist (東大岡田研の学生が起業)

  https://www.a-s-ist.com/
- ベイズ計測で、自分のポジションは、自ら創る.

ベイズ計測による
計測と解析の双方向相互作用

計測限界から実験計画へ



実験家もベイズ計測を使いこなす時代へ

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)

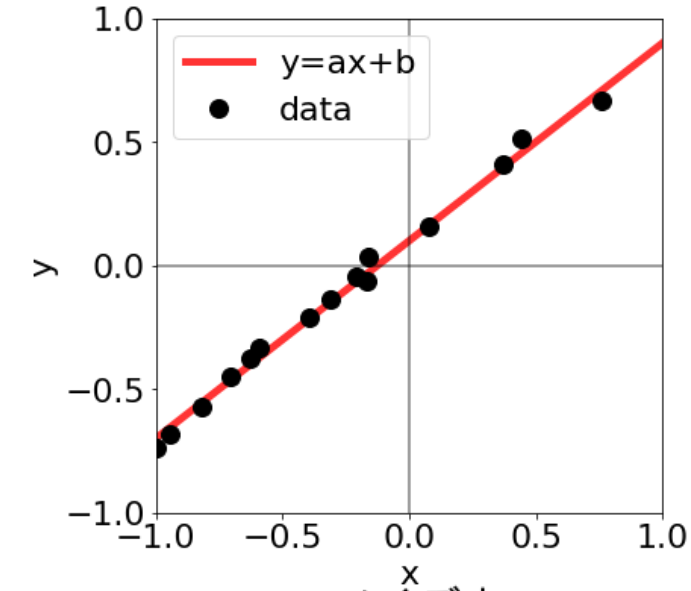
計算理論のレベルの目的，適切さ，方略を元に，系をモデル化し，計算理論を数学的に表現する

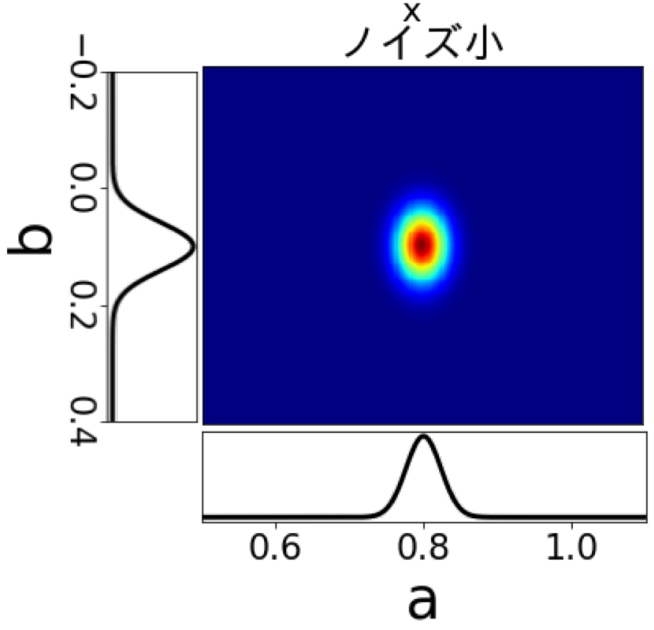
表現・アルゴリズム(統計学，機械学習，計算科学)

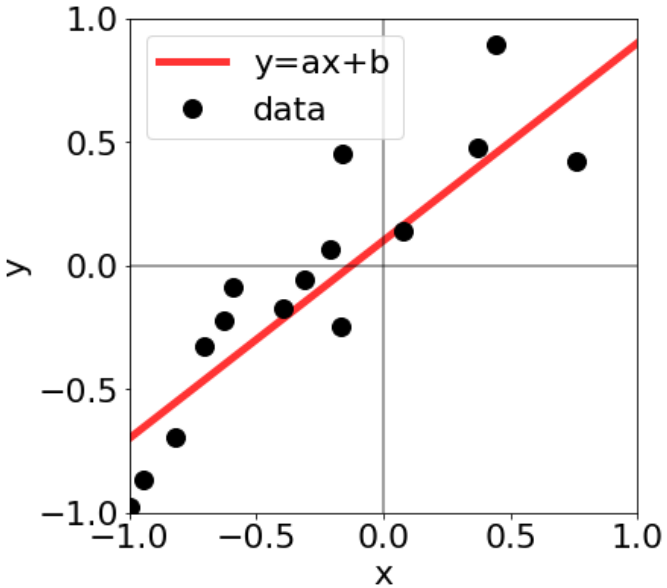
モデリングの結果得られた計算問題を，実行するのためのアルゴリズムを議論する.

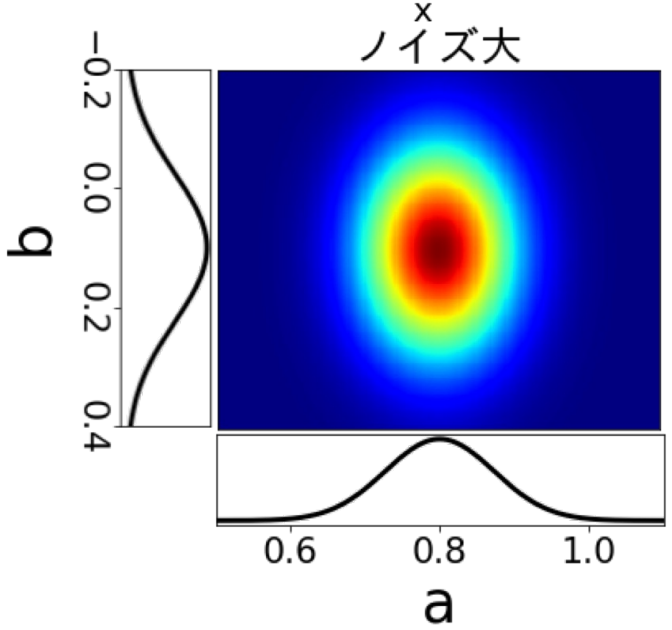
Igarashi, Nagata, Kuwatani, Omori, Nakanishi-Ohno and M. Okada, “Three Levels of Data-Driven Science”, *Journal of Physics: Conference Series*, 699, 012001, 2016.

神器1: パラメータの事後確率推定 (4/4)

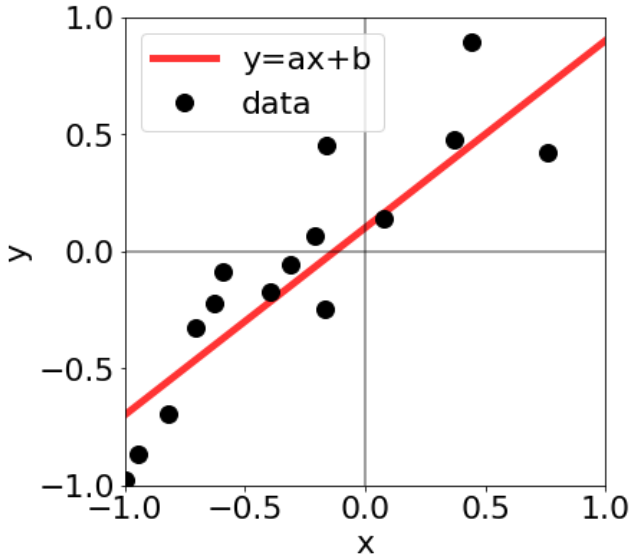


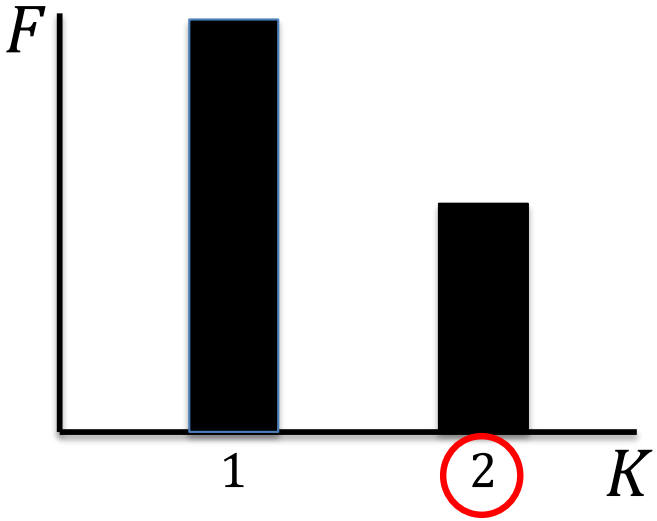






モデル選択: 自由エネルギー差





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

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

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

データのみからモデルを選択できる

まとめ

- 本講演の目的
- データ駆動科学
- 材料/デバイスの機能発現の3ステップモデル
- 情報数理基盤のベイズ推論とスパースモデリング
- ベイズ推論を計測科学に適用したコンパクトな体系のベイズ計測
  - ベイズ計測三種の神器
- $y=ax+b$の線形回帰、スペクトル分解を述べ、さらに機能発現の3ステップモデルの例として大久保研との共同研究を紹介する。
- 材料工学の展望