Three levels of data-driven science

Yasuhiko Igarashi¹, Kenji Nagata¹, Tatsu Kuwatani², Toshiaki Omori³, Yoshinori Nakanishi-Ohno¹ and Masato Okada¹,⁴

¹Graduate School of Frontier Sciences, the University of Tokyo, 5-1-5, Kashiwanoha, Kashiwa, Chiba, 277-8561, Japan
²Department of Solid Earth Geochemistry, Japan Agency for Marine-Earth Science and Technology, 2-15 Natsushima-cho, Yokosuka 237-0061, Japan
³Department of Electrical and Electronic Engineering, Graduate School of Engineering, Kobe University, 1-1 Rokkodai-cho, Nada-ku, Kobe, Hyogo 657-8501, Japan
⁴ERATO, Okanoya Emotional Information Project, Japan Science Technology Agency, 2-1 Hirosawa, Wako, Saitama, 351-0198, Japan

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

Abstract. A research project, called “the Initiative for High-dimensional Data-Driven Science through Deepening of Sparse Modeling” is introduced. A concept, called the three levels of data-driven science, is proposed to untie the complicated relation between many fields and many methods. This concept claims that any problem of data analysis should be discussed at different three levels: computational theory, modeling, and representation/algorithm. Based on the concept, how to choose a suitable method among several candidates is discussed through our study on spectral deconvolution. In addition, how to find a universal problem across the disciplines is presented by explaining our proposed ES-SVM method. Moreover, it is illustrated that the hierarchical structure of data analysis should be visualized and shared. From these discussions, we believe that data-driven science is mother of science, namely, a scientific framework that drives many fields of science.

1. Introduction

Recent progress in the technology of experimentation and measurement makes it possible to obtain a huge amount of high-dimensional data. Needless to say, this trend has forced scientists to change their attitudes toward data. Effective use of high-dimensional data involves a strong framework and a flexible methodology to make the tight connection of information science to the original purpose of data analysis derived from various scientific disciplines. The aim of this project is to establish a novel framework for natural science, namely, data-driven science.

We noticed that sparse modeling (SpM) is a key technology of data-driven science. In the field of information science, SpM has recently received much attention. The basic notions of SpM are as follows. First, high-dimensional data are sparse and can be essentially expressed by a small number of variables. Second, the number of explanatory variables should be reduced without loss of accuracy. Finally, explanatory variables are selected objectively, and models of target phenomena are constructed automatically.

In this study, extracting latent structures from high-dimensional data by SpM is focused on. An effective framework of data-driven science in various fields of natural science is proposed, and some activities by the organizers of the conference, HD³-2015, Grant-in-Aid for Scientific Research on Innovative Areas, also known as Sparse Modeling [1], are described.
Figure 1. Three levels of data-driven science.

This paper consists of five sections. First, the concept of *three levels of data-driven science*, namely, a guiding principle of data-driven science, is proposed in section 2. This concept claims that any problem of data analysis should be discussed at different three levels: computational theory, modeling, and representation/algorithm. Section 3 explains how to choose a suitable representation/algorithm in the case where there are several algorithms for a certain computational theory of data analysis, giving the example of spectral deconvolution shared by various disciplines [2]. In section 4, we demonstrate that it is important to find an essential problem of data-driven approach that is latent in a variety of fields and develop a universal modeling to the problem, while the applications of our proposed “exhaustive-search method for support vector machines (ES-SVM)” to geoscience, psychology, and so forth are introduced. In sections 3 and 4, the initiative for data-driven science and the deepening of SpM, respectively, are described [1]. In section 5, a three-step approach to materials science is presented, and an example of applying the three levels of data-driven science to specific fields of natural science is described. Finally, the topics covered in the preceding sections are summarized, and the future prospects of data-driven science are discussed.

2. The three levels of data-driven science

Recently, data science was featured in one of the leading journals *Science*. In its special issue, it was reported that a common method for analyzing image data performs well in the fields of both astronomy and life science, though these fields completely differ in terms of target objects and spatial and temporal scales [5]. The versatility and broad scope of data-analysis methods inspire us to believe that a general principle exists behind a wide variety of data, and that universally applicable algorithms can be developed on the basis of this principle. It follows that one of the most important purposes of data-driven science is the pursuit of a fundamental principle. As described in section 3, on the other hand, the same set of spectroscopic data can be analyzed by several different methods. It is thus necessary to choose the most suitable method. That choice hinges on the question, gAccording to what criteria should the appropriateness of data-analysis methods to specific purposes be evaluated?h Moreover, the relation between target data and analysis methods is not one-to-one but many-to-many. Accordingly, the promotion of
data-driven science will untie this complex correspondence systematically.

To develop a guiding principle of data-driven science, it is useful to cast a spotlight on the three levels proposed by a famous neuroscientist David Marr. In his book *Vision*, he proposed the three levels at which any machine carrying out an information-processing task must be understood, and since the publication of that book, the concept of the three levels has considerably influenced neuroscience [4]. Also in computer science, there was a paradigm shift called “object orientation” and, since then, it has been argued that the hierarchy of problems should be recognized and utilized. Through the scientific activities of Sparse Modeling, we have found an importance of hierarchy for data-driven science as the object orientation in computer science, and have become convinced that the three levels shown in figure 1 give a novel insight into data-driven science.

The top (first) level, *computational theory*, in figure 1 comprises the goal and strategy for information processing as well as the appropriateness of data analysis. This level is the linchpin connecting the specific fields of natural science and data-driven science. For example, materials informatics aims to predict physical properties (such as electric conductivity, permittivity and magnetic permeability) from data obtained by large-scale computation regarding electronic states of materials. It is impossible to set a proper objective of data analysis without a background in materials science, on the basis of which the meaning of these physical properties and the mechanism of material functions are discussed in terms of physics and chemistry. This specific example shows that constructing computational theory requires a deep understanding of a target field of natural science in addition to knowledge about techniques for data analysis.

The significance of strategy setting at the top level (computational theory) is explained as follows. In general, data do not always include a sufficient amount of information to achieve the objective of data analysis. Let us consider the computational theory of stereopsis by binocular disparity. The relevant data are two images formed on retinas by visual inputs into both eyes. Actually, it is impossible to calculate the depth of what is seen from these two images only, because the correspondence between the two retinal images is still ambiguous. Marr and Poggio developed a computational theory for depth perception by assuming that the correspondence between retinal images is spatially smooth [6]. The assumption of smoothness is a strategy at the top level (computational theory). In addition, the appropriateness of the strategy is ensured by the fact that the spatial distributions of physical quantities usually vary smoothly. The above example shows that the objective, strategy, and appropriateness of data analysis are based on the knowledge of the target field of natural science.

The bottom level in figure 1 consists of two concepts: one is the choice of mathematical representation for the goal and strategy of computational theory, and the other is the algorithm to be used to execute the representation. This layer corresponds to machine-learning techniques and data-analysis methods. Between the computational-theory level and the representation and algorithm level, a modeling level, which enables smooth translation of computational theory into representation and algorithm, is necessary to construct an information-processing system called *data-driven science* (figure 1). The modeling level to formulate what is common to various purposes of data analysis plays a leading role in connecting various fields of natural science and information science. Thus, the modeling level forms the basis of data-driven science.

In this study, the concept of three novel levels, which consist of computation theory, modeling, and representation/algorithm, namely, the three levels of data-driven science was set as a guiding principle for data-driven science.

As stated above, the discovery of interdisciplinary structural similarity regarding data analysis is necessary to untie the complex relation between scientific data and analysis methods, and it is beyond conventional academic areas dealing with only a specific field. When researchers from different disciplines discuss the issue of data analysis, they often miss which level it oriented from, only to lose sight of the route to the solution. The breakthrough regarding this situation
involves removing the barriers between different disciplines, conveying common knowledge to different fields, and revealing interdisciplinary structural similarity. Hence, we insist that it is necessary to keep the three levels of data-driven science in mind constantly.

3. Spectral deconvolution
Spectral deconvolution is a framework for deconvoluting multimodal spectral data, such as that shown in figure 2(a), into the linear sum of unimodal basis functions such as Gaussian functions[2]. In this section, we consider spectral deconvolution, and describe how to apply the three levels of data-driven science (described in section 2) to a specific data-analysis problem.

Firstly, the computational theory of spectral deconvolution is described. In spectroscopy, each unimodal basis function is contained in a complex spectrum, such as that shown in figure 2(a), and corresponds to the electron energy levels of the target material. Therefore, the computational theory of spectral deconvolution is to determine the unknown electronic state of a measuring material. As the strategy of spectral deconvolution, the multimodal spectral data, such as that shown in figure 2(a), is deconvoluted into the sum of unimodal basis functions, and the position of each unimodal basis function, which corresponds to the energy state, is estimated. In the following, two representations, Bayesian estimation in figure 2(b) and L1VM (which stands for $l_1$ regularization vector machine and is one kind of SpM [7]), and their algorithms for this computational theory and its strategy, are described. Moreover, the appropriateness of each representation is evaluated by mathematically investigating the advantages and disadvantages of the representations and algorithms.

3.1. Bayesian spectral deconvolution
Bayesian spectral deconvolution [2] is explained as follows. The following equation is the mathematical representation for the strategy of the spectral deconvolution,

$$G(x; \theta, K) = \sum_{k=1}^{K} a_k \exp \left(-\frac{(x - \mu_k)^2}{2\sigma_k^2}\right),$$ (1)
where the number of Gaussian functions \((K)\) shows the number of energy levels for the measured material. Parameter set \(\theta\) is defined as \(\theta = \{a_k, \mu_k, \sigma_k\}_{k=1}^K\). The mathematical representation for the computational theory of the spectral deconvolution is to appropriately determine \(K\) and \(\theta\) from data set \(D = \{x_i, y_i\}_{i=1}^N\) with \(N\) data points, which compose the multimodal spectral data shown in figure 2(a).

Parameter set \(\theta\) is often optimized by minimizing the following mean-square error function:

\[
E(\theta, K) = \frac{1}{2N} \sum_{i=1}^N (y_i - G(x_i; \theta, K))^2. \tag{2}
\]

This process is called the method of least squares. It can be used to determine not only \(\theta\) but also \(K\) as follows. Although \(E(\theta, K)\) can be reduced to almost zero by increasing \(K\), the obtained curve \(G(x; \theta, K)\) is an excellent fit to not only signal but also noise, so it gives a very poor representation of the measured material. Bayesian estimation can optimize not only \(\theta\) but also \(K\) by stochastically formulating an observation model and by tracking back causal laws with Bayes’ theorem.

To be more precise, the causal laws, which produce observation data \(y\), are stochastically formulated. First, it is assumed that the number of energy levels for the measured material \((K)\) is generated according to probability \(p(K)\), and \(\theta\) is obtained from conditional probability \(p(\theta|K)\). For simplicity , \(p(K)\) and \(p(\theta|K)\) are set as constants. The observed data \((y)\) are assumed to be the sum of a signal following equation (1) and a Gaussian noise \((\epsilon)\) with zero mean and variance 1; that is, \(y = G(x; \theta, K) + \epsilon\). Given \(K\) and \(\theta\), \(y\) is generated from the conditional probability of input \(x\) as

\[
p(y|x, \theta, K) = \frac{1}{\sqrt{2\pi}} \exp \left(- \frac{(y - G(x; \theta, K))^2}{2} \right). \tag{3}
\]

Under the assumption that each data, \((x_i, y_i)\), is independently obtained, the conditional probability of obtaining set \(D\) of \(N\) data given \(\theta\) can be written as \(p(D|\theta, K) = \prod_{i=1}^N p(y_i|x_i, \theta, K) \propto \exp(-NE(\theta, K))\). Then, applying Bayes’ theorem and marginalization, it is possible to obtain the posterior probability of the parameter set, \(p(\theta|D, K)\), given data set \(D\) and \(K\) as

\[
p(\theta|D, K) = \frac{p(D, \theta, K)}{p(D, K)} = \frac{p(D|\theta, K)p(\theta|K)p(K)}{p(D)} \propto \exp(-NE(\theta, K)), \tag{4}
\]

where it is assumed that \(p(K)\) and \(p(\theta|K)\) are constant. From equation (4), \(\theta\) that maximizes \(p(\theta|D, K)\) is found to be equal to \(\theta\) obtained by the least-square method.

Posterior probability \(p(K|D)\) of \(K\) given the data \(D\) is given by

\[
p(K|D) = \frac{p(D, K)}{p(D)} = \frac{\int d\theta p(D, \theta, K)}{p(D)} \propto \int d\theta \exp(-NE(\theta, K)). \tag{5}
\]

Bayesian estimation uses the value of \(K\) that maximizes \(p(K|D)\) on the basis of the maximum a posteriori estimation. The right term of equation (5) corresponds to the partition function in statistical mechanics. Hence, free energy \(F(K)\) is defined by using the partition function as follows:

\[
F(K) = - \log \int d\theta \exp(-NE(\theta, K)). \tag{6}
\]
According to the monotonicity of logarithms, the value of $K$ that maximizes $p(K|D)$ is equal to the one that minimizes $F(K)$.

Using Bayesian estimation for spectral deconvolution has two difficulties. One is obtaining the correct $\theta$ for minimizing error function $E(\theta, K)$ (because of the local minima of the error function). The other is how to calculate the integral for $\theta$ in equations (5) and (6).

To overcome these difficulties, the exchange Monte Carlo (EMC) method [2], an algorithm of the Markov-chain Monte Carlo method, is used. It can efficiently sample from a probability distribution with complex local minima such as the spin glass system [8, 9]. Moreover, EMC has another advantage that it can calculate free energy $F(K)$ in equation (6) by using the sampling results[2].

3.2. Spectral deconvolution by L1VM

Spectral deconvolution by L1VM is explained as follows. The following function is a mathematical representation for the strategy of the spectral deconvolution using Gaussian functions:

$$G(x; \{a_i\}_{i=1}^N, \sigma) = \sum_{i=1}^N a_i \exp \left( -\frac{(x - x_i)^2}{2\sigma^2} \right),$$ (7)

where $a_i$ is the magnitude of the $i$th Gaussian function, and it is assumed that the variance of each Gaussian function is constant ($\sigma$) to reduce the number of estimated parameters.

It is assumed that a data set is much larger than the number of energy levels of the measured material ($K$), as shown in figure 2(a). A subset of the training samples is then efficiently selected by using sparsity-promoting priors, and the multimodal spectral data are deconvoluted into the sum of a few unimodal basis functions. It is considered that the position of each unimodal basis function corresponds to the energy state.

As denoted above, $l_1$ regularization is used, and parameters $\{a_i\}_{i=1}^N$ are optimized by minimizing the following $l_1$ regularized function:

$$E(\{a_i\}_{i=1}^N, \lambda) = \frac{1}{2N} \sum_{j=1}^N (y_j - G(x_j; \theta_2))^2 + \lambda \sum_{i=1}^N |a_i|. (8)$$

Here, to avoid over-fitting and optimize the parameters of L1VM, namely, $\lambda$ and $\sigma$, 10-fold cross-validation is used, and the one-standard-error rule, namely, the simplest (most regularized) model is taken, and parameters whose error is within one standard error of the minimal cross validation error are chosen[10], is used.

In the case of spectral deconvolution by L1VM, $K$ is obtained as a by-product. Although the results can lead to inappropriate spectral deconvolution, the evaluation function in equation (8) gives no local solution, and a global minimum solution can be obtained by using a polynomial time algorithm of number of data points $N$.

3.3. Comparison between results by Bayesian estimation and L1VM

An appropriate representation of the spectral deconvolution is determined by comparing the results obtained by the two representations, Bayesian estimation and L1VM, as explained above. Applying Bayesian estimation and L1VM to the data set (figure 2(a)), which are generated data and composed of the sum of three Gaussian functions, obtained the results as shown in figure 3.

Figure 3(a) shows that the Bayesian estimation selects $K = 3$ as plausible result by minimizing free energy $F(K)$. This result is considered a suitable outcome since the generated data are composed of three Gaussian functions. On the other hand, as shown in figure 3(b), the number of peaks abstracted from L1VM is 19, which is far from the true number of peaks,
Figure 3. Results of spectral deconvolution by Bayesian estimation (a) and L1VM (b), respectively. The parameters of L1VM are bandwidth $\sigma = 0.11$ and $\lambda = 0.026$. The black dots represent the multimodal data shown in figure 2(a). The solid line indicates the estimated functions, and the impulses show the estimated height of spectrum $a_k$.

Table 1. Comparison between computational costs, advantage and disadvantage by Bayesian estimation and L1VM

| Algorithm    | Cores | Time (mins) | Advantage            | Disadvantage                      |
|--------------|-------|-------------|----------------------|-----------------------------------|
| Bayes est.   | 10    | 157.4       | Effective representation | High computational cost           |
| L1VM         | 10    | 21.3        | $O(N^3)$ computational cost | Biased estimation result          |

$K = 3$. This is because L1VM uses a restricted model in which the variance of each Gaussian function is constant $\sigma$ and the position of the $i$th peak of the spectrum is fixed data $x_i$. The peak positions obtained by L1VM are seemingly divided into three regions, indicated by three peaks. Moreover, peak-fitting results obtained by L1VM are almost equal to those obtained by Bayesian estimation, as shown in figures 3(a) and (b).

The results of Bayesian estimation and L1VM are compared in table 1. Using only the data set and stochastic formulation, Bayesian estimation can estimate not only $\theta$ but also $K$, and the method can be considered an effective representation for the spectral deconvolution. However, Bayesian estimation suffers from the disadvantages that error function $E(\theta, K)$ in equation (2) has local minima and high computational cost is incurred in deriving free energy $F(K)$. Actually, as shown in table 1, Bayesian estimation takes about 7 times longer to obtain the results than L1VM. Thus, it cannot easily handle a large-scale data set such as 2-dimensional spectral data and multimodal data with more than a hundred peaks. On the other hand, L1VM has the advantage of a polynomial computational cost, namely, $O(N^3)$, where the number of data sets is $N$, and is thus applicable to such large-scale data. At the same time, since L1VM uses the restricted model, the obtained parameters, such as the number of peak functions, could be different from those given by a true function. Due to this disadvantage, post-processing, for example, clustering neighboring peak functions, must be performed as shown in figure 3(b).

Multimodal spectral data can be analyzed by several different representations. It is necessary to appropriately recognize the advantages and disadvantages of each representation and keep the three levels of data-driven science in mind constantly, as shown in figure 1(b). Computational theory and strategy must therefore be taken into account when choosing the most suitable
representation and algorithm.

4. ES-SVM: Exhaustive Search with Support Vector Machine

The deepening of SpM encouraged by data-driven science is illustrated from the viewpoint of the problem of feature selection solved by the framework of supervised learning [3]. The purpose of feature selection is to select a significant combination of features among all given features and variables. In light of the three levels of data-driven science, the problem of feature selection is derived from the level of computational theory where it is requested to extract essential variables to describe data, and should be discussed at the level of representation and algorithm. Cover and Van Campenhout demonstrated that the only algorithm that can find the optimal subset of feature variables is the exhaustive-search method, which explores all of the solution space at the expense of a computational cost of $O(2^N)$ with dimension of data $N$ [11].

To avoid computational complexity, fast approximate algorithms, such as least absolute shrinkage and selection operator (LASSO), whose computational cost is $O(N^3)$ [12], have been proposed. The recent spread of SpM in various fields was triggered by the mathematical equivalence between the results of $l_1$-norm minimization and those of $l_0$-norm minimization under a certain condition. However, we cannot be too careful of the equivalence condition in applying SpM to raw data. For example, the case where SpM is utilized for extreme measurement in natural science is considered as follows. The amount of data obtained by MRI and black-hole imaging does not always satisfy the condition for success of approximate algorithms.

The task of latent-structure extraction is a representative application of feature selection and is very complicated. For example, in electrophysiology, a field of brain science, selecting neurons performing pattern recognition is one task done by SpM. In such a case, several sets of neurons can have a function of pattern recognition in the first place. In short, at the level of computational theory, it is assumed that the problem of feature selection has multiple solutions. LASSO is designed on the assumption that the problem has a unique solution. Consequently, when LASSO is applied to such a system with multiple solutions, it is unlikely to find the proper solutions. In a more serious case where no neuron is involved in the task of pattern recognition, LASSO is urged into a phantom set of neurons. Nagata et al. proposed ES-SVM (exhaustive search with support vector machine) to address these kind of problems with multiple or no solutions properly.

When a monkey was performing a pattern-recognition task requiring the identification of facial images, Eifuku et al. observed the neural activities of 23 neurons in the anterior inferior temporal (AIT) cortex of the monkey [13]. The neural activities are electrophysiologically measured by conducting a single-unit recording. In this study, the face images of 4 different identities viewed from 7 different angles were presented. Then, view-invariant neurons working on the pattern-recognition task with SpM were selected [3]. The neural activities were used to solve the binary-classification problem for face identification, and tried any combination of face images of 4 different identities, that is, $4C_2 = 6$ combinations. Two approximate algorithms of SpM, LASSO and SLR (sparse logistic regression) [14], were used for solving the feature-selection problem. To select the parameters of LASSO and SLR, LOOCV (leave-one-out cross validation) [7] was used. In LOOCV, 14 samples of 23 neurons were divided into 13 training data and 1 test data for calculating the generalization performance of unknown data. This training and validation operation was iterated with different partitioning 14 times to reduce variability. The parameters whose generalization error is minimal were then optimized. It was found that using LASSO and SLR makes it possible to select almost the same neurons working on 5 of the 6 binary classification problems irrespective of training data. On the other hand, LASSO can select no neuron working on the last problem of the 6 binary classification problem and SLR selected neurons for the last problem, which unsteadily vary as training data.

To investigate the difficulties concerning selecting neurons working on the pattern-recognition
task, Nagata et al. proposed ES-SVM [3]. Let us formulate ES-SVM concretely. An indicator variable is defined as \( c \), which represents a combination of the selected neurons working on the pattern-recognition task and has the same dimensions as the number of neurons, that is, 23-dimension vectors. The \( i \)th variable \( c_i \) takes 0 or 1 and indicates that \( c_i = 1 \) if the \( i \)th neuron is contained in the combination, and \( c_i = 0 \) if it is not. By using the combination of features represented by the indicator variable \( c \), the average of LOOCV error, which is defined as prediction error \( CVE(c) \), is calculated. The prediction performance is evaluated by calculating \( CVE(c) \) for all combinations of neurons except \( c = (0, 0, 0, \cdots, 0) \), whose number is \( 2^{23} - 1 = 8388607 \). Figure 4 shows histograms of \( CVE(c) \). Figure 4(a) shows the result of one of the 6 binary classification problems on which almost the same neurons working are selected by the LASSO and SLR, irrespective of training data. Figure 4(b) represents the histogram of \( CVE(c) \), where no neuron should be selected by LASSO, and the selected neurons of SLR unsteadily vary as training data in LOOCV. Comparing figures 4(a) and (b) reveals that there is a certain connection between the shape of the histograms and the solution stability of LASSO and SLR. Accordingly, a method for comparing the obtained histograms with the histogram given by random guessing [3] is proposed in the following. More specifically, it is assumed that given data are randomly labeled and the probability that each of the data is correctly labeled was 0.5. We take a prediction of binary classification 14 times in LOOCV for all combinations, the averaged prediction error of all combinations are given by the following binomial distribution shown by the solid line in figures 4(a) and (b).

\[
p(k) = \frac{14!}{k!(14-k)!} (0.5)^k (1-0.5)^{14-k},
\]

where \( k \) is the number of misclassifications. Figure 4(b) shows that the histogram and the binomial distribution do not differ by much. This means that the neural activities data do not include information about view-invariant binary classification.

The above results are summarized as follows. As pointed out by Cover and Van Campenhout, no non-exhaustive sequential feature-selection procedure can be guaranteed to produce the optimal subset in the feature-selection problem[11]. Based on this knowledge, ES-SVM[3], in which a support vector machine (SVM) is used as the linear discriminant for the binary classification, was proposed. ES-SVM calculates the prediction error \( CVE(c) \) for all \( 2^N - 1 \) combinations of \( N \) dimensional features and makes a histogram of \( CVE(c) \). The histogram shows whether the given data contain the desired information or not. ES-SVM does not cause the problem of multiple solutions since it checks all candidates of feature combination without approximation. Moreover, the histogram of \( CVE(c) \) makes it possible to evaluate both existing and future approximate algorithms. Thus, ES-SVM can be considered a gmeta-algorithmh for judging the performance of approximate algorithms. Although the main disadvantage of ES-SVM is an exponential explosion of computational cost, recently, ES-SVM has been able to

![Figure 4. The histograms of \( CVE(c) \).](image-url)
calculate low-dimensional data, such as 23-dimension data of neural activities, in a realistic calculation time. ES-SVM is thus considered better than approximate algorithms, which do not assure accuracy. Since ES-SVM is an algorithm for solving the feature-selection problem, which has broad applicability, it has already been applied to geochemical discrimination of tsunami deposits and selection of NIRS channels [15, 16]. It is thus paving the way for a new trend in various academic disciplines. Moreover, the ES (exhaustive search) framework can easily be extended to regression problems. Actually, we have recently noticed a potential application to regression problems in astronomy [17].

To overcome the above-mentioned problem of computational cost, a statistical mechanical method, which consists of the exchange Monte Carlo (EMC) method and the method of multiple histograms [3], is proposed here. The basic idea behind this method is that $CVE(c)$ can be considered an energy when $c$ is seen as a binary spin. In this system, unlike a general spin system, $CVE(c)$ is not given the analytical function of $c$. This structure is similar to approximate Bayesian computation (ABC), which is of current interest. This statistical mechanical method is thus called gAES-SVM (approximate exhaustive search with support vector machine), and its effectiveness was demonstrated by estimating a histogram of $CVE(c)$ while incurring a computational cost of about $1/100$ of that of ES-SVM [3]. The histogram makes it possible to estimate the number of states which have, for example, $CVE(c) = 0$ and to verify whether the given data include information for binary classification even when the data size is large and computational cost exponentially explodes. Moreover, similar to ES-SVM, AES-SVM is a meta-algorithm for judging the performance of approximate algorithms.
5. Discussion and Perspective

One of the future developments in data-driven science is the data-driven approach to numerical data obtained by large-scale simulations such as first-principles calculations in condensed matter physics. SpM was utilized to study the dynamics of anharmonic phonons and to extract an effective model of an alloy [18]. In these studies, numerical data play the role of measurement data shown in figure 2(b). It is important to project measurement data and numerical data onto the same effective model. Accordingly, we propose a framework based on effective models obtained by the data-driven approach shown in figure 5(a), to unify numerical data from large-scale simulations and experimental data from the most-advanced measuring equipment seamlessly. This framework was applied to the study of NiGa\textsubscript{2}S\textsubscript{4}, which is a triangular-lattice antiferromagnet, and is considered to be a typical system described by Anderson’s theory of resonating valence bonds (RVB). Recently, Takenaka et al. proposed a Bayesian framework of model selection in which an effective model described by a classical spin system is automatically extracted from Hartree-Fock calculation of electronic states [19]. In experimental studies of neutron scattering, on the other hand, the dispersion relation of a spin wave is related to the magnitude of exchange interactions between classical spins in effective models. In short, in the study of NiGa\textsubscript{2}S\textsubscript{4}, it is impossible to directly compare the numerical results of electronic states and the experimental results of neutron scattering; instead, the concept of effective models makes it possible to relate both experimental and numerical results, as shown in figure 5(a). In the framework of figure 5(a), the key to seamless integration of large-scale simulations and advanced experiments is Bayesian sensing and Bayesian estimation of effective models.

In terms of the characteristics of the specific science and its inherent universality which goes beyond the border of scientific fields, we will describe a concrete way to incorporate the data-driven approach into the specific fields of science. The following hierarchical structures exist in many fields of natural sciences: hierarchies from a genome via a protein, a cell to an individual in life science, hierarchies from behavior of a neuron to psychology and behavior science in brain science, and hierarchies of various scales of time and space in geosciences. These scientific fields have similarities: For all scientific fields, explanation ability has already reached a high level for many phenomena within each layer due to the significant technological progress of measurement instruments and computational equipment. On the other hand, systematic ways to connect among different layers for these scientific fields have not yet been developed. Data-driven science and its fundamental concept, namely, a three-level structure, can be a promising start point which connects among different layers. The procedure is as follows: First, each scientific field is described hierarchically according to the characteristics of the field. The characteristics of the field are represented in the obtained description. Next, data-driven science is used as the universal framework that connects adjacent layers.

The interpretation of the three levels of data-driven science is well illustrated in the field of materials science. It is useful to divide the development process of structural materials such as iron into three steps, as shown in the hierarchical expression in figure 5(b). The structural materials engineering aims to find the optimal values of process parameters in production for making and designing good materials in terms of durability such as strength and heat resistance. The process parameters and the desired properties are put at the first and third steps, respectively, in the hierarchical expression, as shown in figure 5(b). In the development of structural materials, it is important to parameterize the structure of material tissues [20]. Concretely, we enumerate tissue parameters such as grain size, grain shape, lath width, and grain boundary from images of metallographic structure obtained by electron microscopy. Then, we search for the optimal set of these structure parameters to predict the desired properties of materials, such as strength and heat resistance. The structure and tissue parameters are not always causally related to the material properties, but just a correlation between them can perform well in improving the function predictability. This framework has the same
mathematical structure that appears in a brain machine interface, where the intention of users is read out from electroencephalography [21], and in disaster management, where the height of a tsunami is estimated from the data obtained by sensors of DONET (Dense Oceanfloor Network system for Earthquakes and Tsunamis) [22]. We emphasize that the tissue and structure parameters enumerated by intuition do not always form a necessary and sufficient condition for explaining material functions. Therefore, we discuss a framework of supervised learning in which all the structure and tissue parameters that have been suggested so far are taken as inputs, and the desired functions are taken as outputs. In such a case, the technology of SpM will lead to the appropriate set of structure parameters.

Next, we focus on the connection between the first and second steps in figure 5(b). The objective of computational theory is to optimize the process parameters in order to generate the above-obtained structure parameters. In this case, it is not important to estimate process parameters that completely restore the observed electron microscopic images, but it is necessary to estimate the process parameters that restore the appropriate structure parameters, that is, the statistical parameters of the observed electron microscopic images. This task requires the framework of approximate Bayesian computation (ABC) that has recently been developed. ES-SVM proposed in section 4 can be regarded as one example of ABC. Because forward models, which are incorporated into the Bayesian estimation, are generally phenomenological equations such as the phase-field model, it is necessary to systematically extract an effective model from a first-principles model.

This paper reviews the present situation of the project called gInitiative for High-dimensional Data-Driven Science through Deepening of Sparse Modeling.h We proposed three levels, which can be regarded as the fundamental concept of data-driven science, and introduced a spectral deconvolution as a specific example. Then, we showed that the creation of data-driven science made it possible to propose the ES-SVM method, which results in deepening of sparse modeling. Finally, we proposed a framework that solves scientific problems in various fields in order to practice data-driven science. In this framework, hierarchical structures of a scientific field are visualized as shown in figure 5(b), and adjacent hierarchical structures are linked by data-driven approaches. From the above discussion, we believe that data-driven science is mother science, namely, a scientific framework that will drive many fields of science.

Acknowledgments
We thank S. Tokuda, H. Takenaka and H. Sakamoto for useful comments and discussion to this paper. This work was supported by a Grant-in-Aid for Scientific Research on Innovative Areas (No. 25120005,25120009,25120010) from the Ministry of Education, Culture, Sports, Science and Technology of Japan, and funding Exploratory Research for Advanced Technology, Okanoya Emotional Information Project from the Japan Science and Technology Agency.

References
[1] Grant-in-Aid for Scientific Research on Innovative Areas Initiative for High-dimensional Data-Driven Science through Deepening of Sparse Modeling, http://sparse-modeling.jp/
[2] Nagata K, Sugita S, and Okada M 2012 Neural Netw. 28 82–89
[3] Nagata K, Kitazono J, Nakajima S, Eifuku S, Tamura R and Okada M 2015 IPSJ Transactions on Mathematical Modeling and Its Applications 8 23–30
[4] Marr D 1982 Vision (Cambridge MA: MIT press)
[5] Reed S 2011 Science 331 696–697
[6] Murry K P 2012 Machine learning: a probabilistic perspective (Cambridge MA: MIT press)
[7] Geyer C J 1991 in Comput. Sci. Stat. Proc. 23rd Symp. Inter. ed E M Keramidas (Fairfax Station) 156–163
[8] Mukoshiba K and Nemoto K 1996 J. Phys. Soc. Jpn 65 980–988
[9] Hastie T, Tibshirani R and Friedman J 2001 The Elements of Statistical Learning (Berlin: Springer) 216
[10] Cover T M and Canpenhout J M 1977 IEEE Trans. Syst., Man, Cybern., Syst. 7 657–661
[12] Tibshirani R 1996 J. R. Statist. Soc. B 58 267–288
[13] Eifuku S, Souza, W C, Tamura R, Nishijo H and Ono T 2004 J. Nuueorphysiol. 91 358–371
[14] Yamashita O, Sato M A, Yoshioka T, Tong F and Kamitani Y 2008 Neuroimage 42 1414–1429
[15] Kuwatani T, Nagata K, Okada M, Watanabe T, Ogawa Y, Komai T and Tsuchiya N 2014 Scientific Reports 4 7077
[16] Ichikawa H, Kitazono J, Nagata K, Manda A, Shimamura K, Sakuta R, Okada M, Yamaguchi M K and Kakigi R 2014 Front. Hum. Neurosci. 8 480
[17] Uemura M, Kawabata K S, Ikeda S, Maeda K, Wu H-Y, Watanabe K, Takahashi S and Fujishiro I 2015 Proc. Int. Meeting of HD3-2015 22
[18] Nelson L J, Hart G L W, Zhou F, and Ozolina V 2013 Phys. Rev. B 87 035125
[19] Takenaka H, Nagata K, Mizokawa T and Okada M 2014 J. Phys. Soc. Jpn. 83 124706
[20] Simmons J P, Chuang P, Comer M, Spowart J E, Uchic M D and Greaf M De 2009 Model. Simul. Mater. Sci. Eng. 17 025002
[21] Yokota Y, Igarashi Y, Okada M and Naruse Y 2015 Proc. IEEE Eng. Med. Biol. Soc. in press
[22] Igarashi Y, Baba T, Sato K, Hori T and Okada M 2015 to be submitted