APL MATERIALS **4**, 053208 (2016)

Perspective: Materials informatics and big data: Realization of the "fourth paradigm" of science in materials science

Ankit Agrawal[a] and Alok Choudhary
Department of Electrical Engineering and Computer Science, Northwestern University, Evanston, Illinois 60208, USA

(Received 13 January 2016; accepted 3 April 2016; published online 15 April 2016)

Our ability to collect "big data" has greatly surpassed our capability to analyze it, underscoring the emergence of the fourth paradigm of science, which is data-driven discovery. The need for data informatics is also emphasized by the Materials Genome Initiative (MGI), further boosting the emerging field of materials informatics. In this article, we look at how data-driven techniques are playing a big role in deciphering processing-structure-property-performance relationships in materials, with illustrative examples of both forward models (property prediction) and inverse models (materials discovery). Such analytics can significantly reduce time-to-insight and accelerate cost-effective materials discovery, which is the goal of MGI. [http://dx.doi.org/10.1063/1.4946894]

INTRODUCTION

The field of materials science relies on experiments and simulation-based models to understand the "physics" of different materials in order to better understand their characteristics and discover new materials with improved properties for use in society at all levels. Lately, the "big data" generated by such experiments and simulations has offered unprecedented opportunities for application of data-driven techniques in this field, thereby opening up new avenues for accelerated materials discovery and design. The need for such data analytics has also been emphasized by the Materials Genome Initiative (MGI),[1] which envisions the discovery, development, manufacturing, and deployment of advanced materials twice as fast and at a fraction of the cost.

Four paradigms of science

In fact, these developments in the field of materials science are along the lines of how science and technology overall have evolved over the centuries. For thousands of years, science was purely empirical, which here corresponds to metallurgical observations over the "ages" (stone, bronze, iron, steel). Then came the paradigm of theoretical models and generalizations a few centuries ago, characterized by the formulation of various "laws" in the form of mathematical equations; in materials science, the laws of thermodynamics are a good example. But for many scientific problems, the theoretical models became too complex with time, and an analytical solution was no longer feasible. With the advent of computers a few decades ago, a third paradigm of computational science became very popular. This has allowed simulations of complex real-world phenomena based on the theoretical models of the second paradigm, and excellent examples of this in materials science are the density functional theory (DFT) and molecular dynamics (MD) simulations. These paradigms of science have contributed in turn towards advancing the previous paradigms, and today these are popular as the branches of theory, experiment, and computation in almost all scientific domains. The amount of data being generated by these experiments and simulations has given rise to the fourth

[a]ankitag@eecs.northwestern.edu. URL: http://eecs.northwestern.edu/~ankitag.

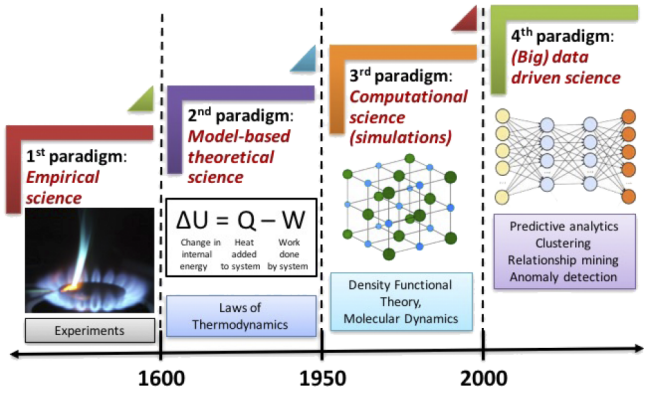

FIG. 1. The four paradigms of science: empirical, theoretical, computational, and data-driven.

paradigm of science[2] over the last few years, which is (big) data driven science, and it unifies the first three paradigms of theory, experiment, and computation/simulation. It is increasingly becoming popular in the field of materials science as well and has, in fact, led to the emergence of the new field of materials informatics. Figure 1 shows these four paradigms of science.

Big data

Before going further, it would be useful to expand on the concept of big data, and what it means specifically in the context of materials science. The "bigness" (amount) of data is certainly the primary feature and challenge, but several other characteristics can make the collection, storage, retrieval, analysis, and visualization of such data even more challenging. For example, the data may be from heterogeneous sources, may be of different types, may have unknown dependencies and inconsistencies within it, may have parts that are missing or not reliable, may be generated at a rate that could be much greater than what traditional systems can handle, may have privacy issues, and so on. These issues can be summarized by the various Vs associated with big data—volume, velocity, variety, variability, veracity, value, and visualization. Of these, the first three (volume, velocity, and variety) are specific to big data and others are features of any data, including big data. Further, each application domain can also introduce its own nuances to the process of big data management and analytics. It is worth noting that in many areas of materials science, until recently, there was more of a no data than a big data problem, in the sense that open, accessible data have been rather limited; however, recent MGI-supported efforts[3,4] and other similar efforts around the world are promoting the availability and accessibility of digital data in materials science. These efforts include combining experimental and simulation data into a searchable materials data infrastructure and encouraging researchers to make their data available to the community. Thanks to such efforts, it is fair to say that the sheer complexity and variety in materials science data becoming available nowadays requires the development of new big data approaches in materials informatics. For example, there are numerous kinds of experimental and simulation-based materials property data (e.g., physical, chemical, electronic, thermodynamic, mechanical, structural), engineering/processing data (e.g., heat treatment), image data (e.g., electron backscatter diffraction), spatio-temporal data (e.g., tomography, structure evolution), unstructured textual data (materials science literature), and so on. Of course, several of these types of data are often coupled together. Several excellent review articles on big data in materials science are available in the literature.[5–7]

Processing-structure-property-performance (PSPP) relationships

So what can materials informatics and big data do for a real-world materials science application? It is well-known that the key of almost everything in materials science relates to PSPP relationships,[8] which are far from being well-understood. Figure 2 shows these PSPP relationships, where the deductive science relationships of cause and effect flow from left to right, and the

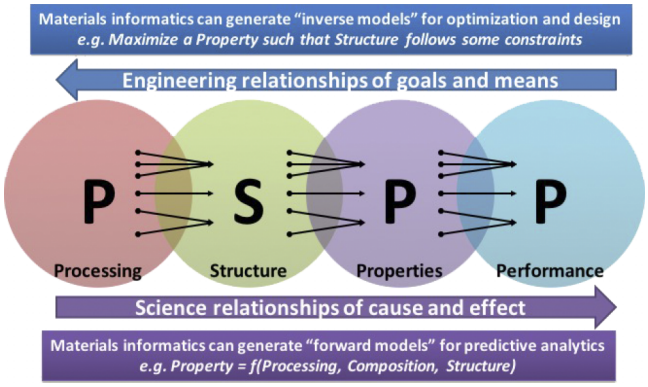

FIG. 2. The processing-structure-property-performance relationships of materials science and engineering, and how materials informatics approaches can help decipher these relationships via forward and inverse models.

inductive engineering relationships of goals and means flow from right to left. Further, it is important to note that each relationship from left to right is many-to-one, and consequently the ones from right to left are one-to-many. Thus, many processing routes can potentially result in the same structure of the material, and the same property of a material could be potentially achieved by multiple structures. Each experimental observation or simulation can be thought of as a data point for a forward model (e.g., a measurement or calculation of a property given the processing, composition, and structure parameters). A database of such data points can be used with a materials informatics approach such as predictive analytics to build data-driven forward models that can run in a very small fraction of the time it takes for doing the experiment or simulation. This acceleration of forward models can not only help to guide future simulations and experiments, but also make it possible to realize the inverse models, which are much more challenging and critical for materials discovery and design. The construction of inverse models is typically formulated as an optimization problem wherein a property or performance metric of interest is intended to be maximized or minimized, subject to the various constraints on the representation of the material, which is typically in the form of a composition- and/or structure-based function. The optimization process usually involves multiple invocations of the forward model, and thus having a fast forward model is extremely valuable. Further, since these inverse relationships are one-to-many, a good inverse model should be able to identify multiple optimal solutions (if they exist), so as to have the flexibility to select the material structure that can be attained in the easiest and most cost-effective manner.

In the rest of this article, we shall first discuss a generic workflow for conducting materials informatics, and then illustrate it with examples of some recent advances in this field in terms of development and application of big data analytics approaches for building both forward and inverse models for PSPP relationships. In particular, we will take the example of steel fatigue prediction using a set of experimental data (forward models),[9] predicting the stability of a compound using DFT simulation data (forward models) and subsequent discovery of stable ternary compounds (inverse models),[10] and structure-property optimization of a magnetoelastic material (inverse models).[11]

KNOWLEDGE DISCOVERY WORKFLOW FOR MATERIALS INFORMATICS

Figure 3 depicts a typical end-to-end workflow for materials informatics. A variety of raw materials data as discussed earlier could be stored in heterogenous materials databases in potentially different data formats. Given a task at hand (say developing property prediction models), the first step is to understand the data format and representation, and do any necessary preprocessing to ensure the quality of the data before any modeling, and remove or appropriately deal with noise, outliers, missing values, duplicate data instances, etc. Usually the instances and/or attributes responsible for these are removed if they are easily identifiable and sufficient data are available, but optimally utilizing incomplete data remains an active area of research. Examples of data preprocessing steps include discretization, sampling, normalization, attribute-type conversion, feature

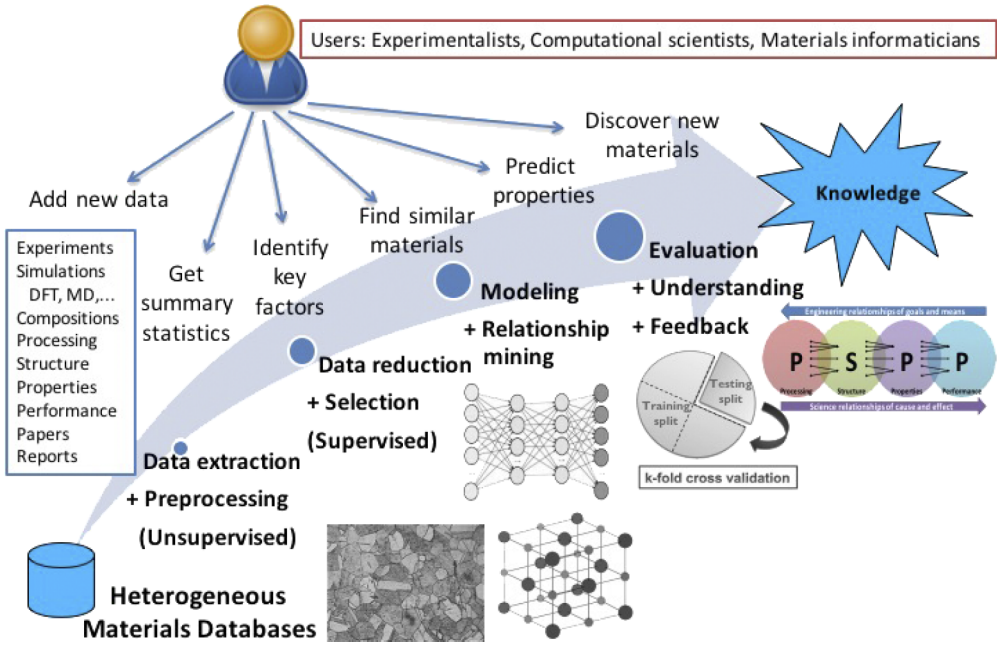

FIG. 3. The knowledge discovery workflow for materials informatics. The overall goal is to mine heterogenous materials databases and extract actionable PSPP linkages to enable data-driven materials discovery and design.

extraction, feature selection, etc. Such data preprocessing can either be supervised or unsupervised, based on whether the process depends on the target attributes (here the property of the material to be predicted), and are thus usually considered separate stages in the workflow.

Once appropriate data preprocessing has been performed and the data are ready for modeling, one can employ supervised data mining techniques for predictive modeling. Caution needs to be exercised here to appropriately split the data into training and testing sets (or use cross validation), else the model may be prone to overfitting and show over-optimistic accuracy. If the target attribute is numeric (e.g., fatigue strength, formation energy) regression techniques can be used for predictive modeling and if it is categorical (e.g., whether a compound is metallic or not), classification techniques can be used. Some techniques are capable of doing both classification and regression. There also exist several ensemble learning techniques that combine the results from base learners in different ways, and have shown to improve accuracy and robustness of the model in some cases. Table I lists some of the popular predictive modeling techniques. Apart from predictive modeling, one can also use other data mining techniques such as clustering and relationship mining depending on the goal of the project, for instance, to find group similar materials or discovering hidden patterns and associations in the data.

Proper evaluation of data-driven models is crucial. A data-driven model can, in principle, "memorize" every single instance of the dataset and thus result in 100% accuracy on the same data, but will most likely not be able to work well on unseen data. For this reason, advanced data-driven techniques that usually result in black-box models need to be evaluated on data that the model has not seen while training. A simple way to do this is to build the model only on part of the data, and use the remaining for evaluation. This can be generalized to k-fold cross validation, where the dataset is randomly split into k parts. $k-1$ parts are used to build the model and the remaining one part is used for testing, and the process is repeated k times with different test splits. Cross-validation is a standard evaluation setting to eliminate any chances of over-fitting. Of course, k-fold cross-validation necessitates building k models, which may take a long time on large datasets. It is also important to note that cross-validation is supposed to be of the entire workflow and not just of the predictive model. Hence, any supervised data preprocessing should also be considered along with predictive modeling while performing cross-validation in order to get unbiased estimates of accuracy. Quantitative assessments of the models predictive accuracy

TABLE I. Popular predictive modeling algorithms.

Modeling technique	Capability	Brief description
Naive Bayes[12]	Classification	A probabilistic classifier based on Bayes theorem
Bayesian network[13]	Classification	A graphical model that encodes probabilistic conditional relationships among variables
Logistic regression[14]	Classification	Fits data to a sigmoidal S-shaped logistic curve
Linear regression[15]	Regression	A linear least-squares fit of the data w.r.t. input features
Nearest-neighbor[16]	Both	Uses the most similar instance in the training data for making predictions
Artificial neural networks[17,18]	Both	Uses hidden layer(s) of neurons to connect inputs and outputs, edge weights learnt using back propagation
Support vector machines[19]	Both	Based on the structural risk minimization, constructs hyperplanes multidimensional feature space
Decision table[20]	Both	Constructs rules involving different combinations of attributes
Decision stump[21]	Both	A weak tree-based machine learning model consisting of a single-level decision tree
J48 (C4.5) decision tree[22]	Classification	A decision tree model that identifies the splitting attribute based on information gain/gini impurity
Alternating decision tree[23]	Classification	Tree consists of alternating prediction nodes and decision nodes, an instance traverses all applicable paths
Logistic model tree[24,25]	Classification	A classification tree with logistic regression functions at the leaves
M5 model tree[26,27]	Regression	A tree with linear regression function at the leaves
Random tree	Both	Considers a randomly chosen subset of attributes
Reduced error pruning tree[21]	Both	Builds a tree using information gain/variance and prunes it using reduced-error pruning to avoid over-fitting
AdaBoost[28]	Ensembling	Boosting can significantly reduce error rate of a weak learning algorithm
Bagging[29]	Ensembling	Builds multiple models on bootstrapped training data subsets to improve model stability by reducing variance
Random subspace[30]	Ensembling	Constructs multiple trees systematically by pseudo-randomly selecting subsets of features
Random forest[31]	Ensembling	An ensemble of multiple random trees
Rotation forest[32]	Ensembling	Generates model ensembles based on feature extraction followed by axis rotations

can be done with several classification/regression performance metrics such as accuracy, precision, recall/sensitivity, specificity, area under the receiver operating characteristics (ROC) curve, coefficient of correlation (R), explained variance (R^2), Mean Absolute Error (MAE), Root Mean Squared Error ($RMSE$), Standard Deviation of Error (SDE), etc.

The resulting knowledge from this workflow can be represented in the form of invertible PSSP relationships, thereby facilitating materials discovery and design. This workflow can be used and leveraged at different stages by various stakeholders such as experimentalists, computational scientists, and materials informaticians. For example, one can add new data as they are generated, get summary statistics, identify key factors influencing a material property, find materials similar to a given material in the database, develop forward predictive models, use it to predict properties of new materials, and finally develop inverse models to discover materials with a desired property or set of properties.

A lot of research has sprung up over the last decade utilizing data mining on materials science data,[9–11,33–41] and all of them can be said to use some flavor of the above-described materials informatics workflow. It is also worth noting that this workflow is essentially a materials science adaptation of existing similar workflows of data-driven analytics in other domains, as most of the advanced techniques for big data management and informatics come from the field of computer science and more specifically high-performance data mining,[42–50] via applications in many different

domains like business and marketing,[51–53] healthcare,[54–60] climate science,[61–63] bioinformatics,[64–68] and social media analytics,[69–71] among many others.

ILLUSTRATIVE EXAMPLES OF MATERIALS INFORMATICS

Steel fatigue prediction

Agrawal *et al.*[9] used data from the Japan National Institute of Material Science (NIMS) MatNavi database[72] to make predictive models for fatigue strength of steel. Accurate prediction of fatigue strength of steels is important for several advanced technology applications due to the extremely high cost and time of fatigue testing and potentially disastrous consequences of fatigue failures. In fact, fatigue is known to account for more than 90% of all mechanical failures of structural components.[73] The NIMS data included composition and processing attributes of 371 carbon and low-alloy steels, 48 carburizing steels, and 18 spring steels. The materials informatics approach used consisted of a series of steps that included data preprocessing for consistency using domain knowledge, ranking-based feature selection, predictive modeling, and model evaluation using leave-one-out cross-validation (a special case of cross-validation where $k = N$, the number of instances in the dataset; it is generally used with small datasets for a more robust evaluation) with respect to various metrics for prediction accuracy. Twelve regression-based predictive modeling techniques were evaluated, and many of them were able to achieve a high predictive accuracy, with R^2 values ~0.98, and error rate <4%, outperforming the only prior study on fatigue strength prediction reported R^2 values of <0.94. In particular, neural networks, decision trees, and multivariate polynomial regression were found to achieve a high R^2 value of >0.97. It is well known in the field of predictive data analytics that after a certain point it becomes more and more difficult to increase prediction accuracy. In other words, an improvement from 0.94 to 0.97 should be considered more significant than an improvement from 0.64 to 0.67. It was also observed from the scatter plots that the three grades of steels were well separated in most cases, and different techniques tended to perform better in different regions, which was also reflected in the distribution of errors. For example, multivariate polynomial regression gave the best R^2 of 0.9801 but also gave very poor predictions for some carbon and low alloy steels. It is encouraging to see good accuracy despite the limited amount of available experimental data; however, it was also concluded that the methods resulting in bimodal distribution of errors or the ones with significant peaks in higher error regions need to be used with caution even though their reported R^2 may be high. Nonetheless, this work successfully exhibited the use of data science methodologies to build forward models on experimental data connecting processing and composition directly with performance, in the context of PSPP relationships.

Stable compound discovery

In another work, Meredig, Agrawal *et al.*[10] employed a similar predictive analytics framework on simulation data from quantum mechanical DFT calculations. Performing a DFT simulation on a material requires its composition and crystal structure as input. Traditionally, there has been a lot of effort in the field of computational materials science on crystal structure prediction (CSP), which takes the composition as input and performs a global optimization of all possible unit cells to find the one with lowest formation energy. Thus, both CSP and DFT are computationally expensive, and this work used a database of existing DFT calculations to build forward models predicting the property of a material (in this case formation energy) using a variety of attributes all of which could be derived solely based on the materials stoichiometric composition. Of course, the predictive models are trained on DFT data, which implicitly use structure information, but once the model is built, it can subsequently be used to predict the formation energy of new materials without crystal structure as input. It was found that the developed models were excellent at predicting DFT formation energies to which they were not fit, with R^2 score greater than 0.9, and MAE well within DFT's typical agreement with experiment. Thus, these forward models could predict formation energy without any structural input, about six orders of magnitude faster than DFT, and without

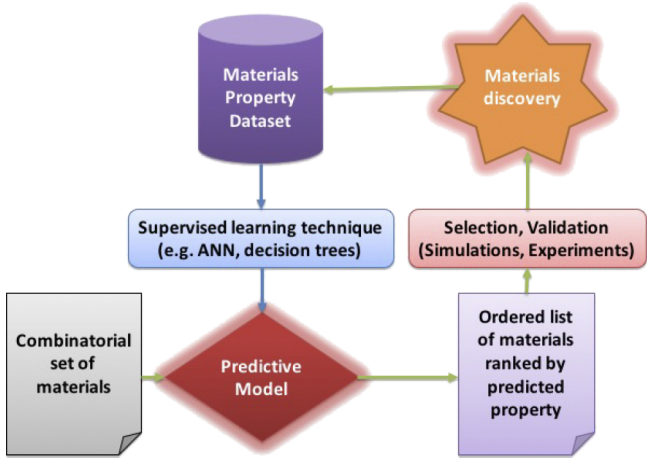

FIG. 4. A simple realization of the inverse models for PSPP relationships. The forward predictive model built using a supervised learning technique on a labeled materials dataset can be used to scan a combinatorial set of materials and thus convert this set to a ranked list, ordered by the predicted property. This can be followed by one or more screening steps to select and validate the predictions using simulation and/or experiments, thereby enabling data-driven materials discovery, which can in turn be fed back into the materials dataset to derive improved models, and so on. Blue arrows denote the forward model construction process, and green arrows denote the materials discovery process via inverse models.

sacrificing the accuracy of DFT compared to experiment. These models were subsequently used to conduct "virtual combinatorial chemistry" by scanning almost the entire unexplored ternary space of compounds of the form $A_xB_yC_z$, to identify compounds that are predicted to be stable. This is tantamount to the inverse models described earlier in this article. Here A, B, C were selected from a list of 83 technologically relevant elements, and x, y, z were obtained using an enumeration procedure taking into account the statistically most common compositions in the Inorganic Crystal Structure Database (ICSD) and basic charge balance conditions. A total of approximately 1.6×10^6 ternary compositions were scanned to get predictions of formation energy in a matter of minutes, in contrast to tens of thousands of years that would be required for DFT simulations to do the same. Many interesting insights were obtained as a result of this study, which identified about 4500 ternary compositions as predictions for new stable compounds. Of these, 9 were systematically selected for a full DFT crystal structure test, and 8 of them were explicitly confirmed to be stable. Interestingly, if the 4500 predictions reported in this study are experimentally confirmed, it would represent an increase in the total number of known stable ternary compounds by more than 10%. This realization of inverse models is depicted in Figure 4.

Data-driven microstructure optimization

The last example that we would like to discuss is the recent work on microstructure optimization of a magnetoelastic Fe-Ga alloy (Galfenol) microstructure for enhanced elastic, plastic, and magnetostrictive properties.[11] When a magnetic field is applied to this alloy, the boundaries between the magnetic domains shift and rotate, and this leads to a change in its dimensions. This behavior is called magnetostriction, and has many applications in microscale sensors, actuators, and energy harvesting devices. Desirable properties for such a material include low young's modulus, high magnetostrictive strain, and high yield strength. Theoretical forward models for computing these properties for a given microstructure are available, but inverse models to obtain microstructures with the desired properties are very challenging. Note that the approach used in the previous example—stable compound discovery where the forward model could scan the entire ternary composition space to realize the inverse model—would be prohibitive here, since the microstructure space is too large. In this example, the microstructure was represented by an orientation distribution function (ODF) with 76 independent nodes leading to a 76 dimensional inverse problem (each value representing positive volume density of a crystal orientation). For a conservative estimate of the

number of possible combinations, even if we assume just two possible values for each dimension, it would result in the order of 2^{76} combinations, which would take more than 2×10^9 years to enumerate, even assuming it takes just 1 μs to compute a property with forward models. Thus, high dimensionality of microstructure space along with other challenges such as multi-objective design requirements and non-uniqueness of solutions makes even the traditional search-based optimization methods incompetent in terms of both searching efficiency and result optimality. In this work, a machine learning approach to address these challenges was developed, consisting of random data generation, feature selection, and classification algorithms. The key idea was to prune the search space as much as possible by search path refinement (ranking the ODF dimensions) and search region reduction (identifying promising regions within the range of each ODF dimension) so as to try to reach the optimal solution faster. More details of this approach are available elsewhere.[74] It was found that this data-driven approach for structure-property optimization could not only identify more optimal microstructures satisfying multiple linear and non-linear properties much faster than traditional optimization methods, but also discover multiple optimal solutions for some properties that were unknown for this problem before this study.

CONCLUSION

To summarize and conclude, (big) data driven analytics, which is the fourth paradigm of science has given rise to the emergence and popularity of materials informatics, and it is of central importance in realizing the vision of the materials genome initiative. We discussed a generic workflow for materials informatics along with three recent illustrative applications where data-driven analytics has been successfully used to learn invertible PSPP relationships. Currently, the field of materials informatics is still pretty much in its nascent stage, much like what bioinformatics was 20 years ago. Interdisciplinary collaborations bringing together expertise from materials science and computer science, and creating a workforce equipped with such interdisciplinary skills, are vital to optimally harness the wealth of opportunities becoming available in this arena, and enable timely discovery and deployment of advanced materials for the benefit of mankind.

ACKNOWLEDGMENTS

The authors gratefully acknowledge support from AFOSR Award No. FA9550-12-1-0458, NIST Award No. 70NANB14H012, and DARPA Award No. N66001-15-C-4036.

[1] Materials Genome Initiative for Global Competitiveness, OSTP, June 2011.
[2] T. Hey, S. Tansley, and K. Tolle, *The Fourth Paradigm: Data-Intensive Scientific Discovery* (Microsoft Research, 2009).
[3] Materials Genome Initiative Strategic Plan, National Science and Technology Council Committee on Technology Subcommittee on the Materials Genome Initiative, June 2014.
[4] C. H. Ward, J. A. Warren, and R. J. Hanisch, "Making materials science and engineering data more valuable research products," Integr. Mater. Manuf. Innovation **3**, 1–17 (2014).
[5] A. A. White, "Big data are shaping the future of materials science," MRS Bull. **38**, 594–595 (2013).
[6] S. R. Kalidindi and M. D. Graef, "Materials data science: Current status and future outlook," Annu. Rev. Mater. Res. **45**, 171–193 (2015).
[7] K. Rajan, "Materials informatics: The materials 'gene' and big data," Annu. Rev. Mater. Res. **45**, 153–169 (2015).
[8] G. B. Olson, "Computational design of hierarchically structured materials," Science **277**, 1237–1242 (1997).
[9] A. Agrawal, P. D. Deshpande, A. Cecen, G. P. Basavarsu, A. N. Choudhary, and S. R. Kalidindi, "Exploration of data science techniques to predict fatigue strength of steel from composition and processing parameters," Integr. Mater. Manuf. Innovation **3**, 1–19 (2014).
[10] B. Meredig, A. Agrawal, S. Kirklin, J. E. Saal, J. W. Doak, A. Thompson, K. Zhang, A. Choudhary, and C. Wolverton, "Combinatorial screening for new materials in unconstrained composition space with machine learning," Phys. Rev. B **89**, 1–7 (2014).
[11] R. Liu, A. Kumar, Z. Chen, A. Agrawal, V. Sundararaghavan, and A. Choudhary, "A predictive machine learning approach for microstructure optimization and materials design," Sci. Rep. **5**, 11551 (2015).
[12] H. George, "John and Pat Langley. Estimating continuous distributions in Bayesian classifiers," in *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence* (Morgan Kaufmann Publishers Inc., 1995), pp. 338–345.
[13] R. R. Bouckaert, "Naive Bayes classifiers that perform well with continuous variables," in *AI 2004: Advances in Artificial Intelligence* (Springer, 2004), pp. 1089–1094.
[14] D. Hosmer and S. Lemeshow, *Applied Logistic Regression* (John Wiley and Sons, Inc., 1989).

[15] E. Weher, "Edwards, Allen, L.: An introduction to linear regression and correlation. (A series of books in psychology.) W. H. Freeman and Comp., San Francisco 1976. 213 S., Tafelanh., s 7.00," Biom. J. **19**, 83–84 (1977).
[16] D. W. Aha and D. Kibler, "Instance-based learning algorithms," Mach. Learn. **6**, 37–66 (1991).
[17] C. Bishop, *Neural Networks for Pattern Recognition* (Oxford University Press, 1995).
[18] L. Fausett, *Fundamentals of Neural Networks* (Prentice Hall, New York, 1994).
[19] V. N. Vapnik, *The Nature of Statistical Learning Theory* (Springer, 1995).
[20] R. Kohavi, "The power of decision tables," in *Proceedings of the 8th European Conference on Machine Learning, ECML '95* (Springer-Verlag, London, UK, 1995), pp. 174–189.
[21] I. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques* (Morgan Kaufmann Publication, 2005).
[22] J. Quinlan, *C4. 5: Programs for Machine Learning* (Morgan Kaufmann, 1993).
[23] Y. Freund and L. Mason, "The alternating decision tree learning algorithm," in *Proceeding of the Sixteenth International Conference on Machine Learning* (Citeseer, 1999), pp. 124–133.
[24] N. Landwehr, M. Hall, and E. Frank, "Logistic model trees," Mach. Learn. **59**, 161–205 (2005).
[25] M. Sumner, E. Frank, and M. Hall, "Speeding up logistic model tree induction," in *Knowledge Discovery in Databases: PKDD 2005* (Springer, 2005), pp. 675–683.
[26] Y. Wang and I. Witten, "Induction of model trees for predicting continuous classes," in *Proceedings of European Conference on Machine Learning Poster Papers, Prague, Czech Republic* (Springer, 1997), pp. 128–137.
[27] J. R. Quinlan, *Learning with Continuous Classes* (World Scientific, 1992), pp. 343–348.
[28] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," in Proceedings of the 13th International Conference on Machine Learning 96, 148-156 (1996).
[29] L. Breiman, "Bagging predictors," Mach. Learn. **24**, 123–140 (1996).
[30] T. Ho, "The random subspace method for constructing decision forests," IEEE Trans. Pattern Anal. Mach. Intell. **20**, 832–844 (1998).
[31] L. Breiman, "Random forests," Mach. Learn. **45**, 5–32 (2001).
[32] J. Rodriguez, L. Kuncheva, and C. Alonso, "Rotation forest: A new classifier ensemble method," IEEE Trans. Pattern Anal. Mach. Intell. **28**, 1619–1630 (2006).
[33] S. Curtarolo, D. Morgan, K. Persson, J. Rodgers, and G. Ceder, "Predicting crystal structures with data mining of quantum calculations," Phys. Rev. Lett. **91**, 135503 (2003).
[34] C. C. Fischer, K. J. Tibbetts, D. Morgan, and G. Ceder, "Predicting crystal structure by merging data mining with quantum mechanics," Nat. Mater. **5**, 641–646 (2006).
[35] G. Hautier, C. C. Fischer, A. Jain, T. Mueller, and G. Ceder, "Finding natures missing ternary oxide compounds using machine learning and density functional theory," Chem. Mater. **22**, 3762–3767 (2010).
[36] K. Gopalakrishnan, A. Agrawal, H. Ceylan, S. Kim, and A. Choudhary, "Knowledge discovery and data mining in pavement inverse analysis," Transport **28**, 1–10 (2013).
[37] P. Deshpande, B. P. Gautham, A. Cecen, S. Kalidindi, A. Agrawal, and A. Choudhary, "Application of statistical and machine learning techniques for correlating properties to composition and manufacturing processes of steels," in *2nd World Congress on Integrated Computational Materials Engineering* (John Wiley & Sons, Inc., 2013), pp. 155–160.
[38] A. G. Kusne, T. Gao, A. Mehta, L. Ke, M. C. Nguyen, K.-M. Ho, V. Antropov, C.-Z. Wang, M. J. Kramer, C. Long *et al.*, "On-the-fly machine-learning for high-throughput experiments: Search for rare-earth-free permanent magnets," Sci. Rep. **4**, 6367 (2014).
[39] R. Liu, Y. C. Yabansu, A. Agrawal, S. R. Kalidindi, and A. N. Choudhary, "Machine learning approaches for elastic localization linkages in high-contrast composite materials," Integr. Mater. Manuf. Innovation **4**, 1–17 (2015).
[40] P. V. Balachandran, J. Theiler, J. M. Rondinelli, and T. Lookman, "Materials prediction via classification learning," Sci. Rep. **5**, 13285 (2015).
[41] F. Faber, A. Lindmaa, O. A. von Lilienfeld, and R. Armiento, "Crystal structure representations for machine learning models of formation energies," Int. J. Quantum Chem. **115**, 1094–1101 (2015).
[42] Y. Bengio, "Learning deep architectures for ai," Found. Trends® Mach. Learn. **2**, 1–127 (2009).
[43] W. Fan and A. Bifet, "Mining big data: Current status, and forecast to the future," ACM SIGKDD Explor. Newsl. **14**, 1–5 (2013).
[44] A. Agrawal, M. Patwary, W. Hendrix, W.-k. Liao, and A. Choudhary, "Big Data and High Performance Computing," in *Cloud Computing and Big Data*, edited by L. Grandinetti, Advances in Parallel Computing Vol. 23 (IOS Press, 2013), pp. 192–211.
[45] M. Patwary, D. Palsetia, A. Agrawal, W.-k. Liao, F. Manne, and A. Choudhary, "Scalable parallel optics data clustering using graph algorithmic techniques," in *Proceedings of 25th International Conference on High Performance Computing, Networking, Storage and Analysis (Supercomputing, SC'13)* (ACM, 2013), pp. 1–12.
[46] Z. Chen, S. W. Son, W. Hendrix, A. Agrawal, W.-k. Liao, and A. Choudhary, "Numarck: Machine learning algorithm for resiliency and checkpointing," in *Proceedings of 26th International Conference on High Performance Computing, Networking, Storage and Analysis (Supercomputing, SC'14)* (ACM, 2014), pp. 733–744.
[47] Y. Low, D. Bickson, J. Gonzalez, C. Guestrin, A. Kyrola, and J. M. Hellerstein, "Distributed graphlab: A framework for machine learning and data mining in the cloud," Proc. VLDB Endowment **5**, 716–727 (2012).
[48] Y. Xie, D. Palsetia, G. Trajcevski, A. Agrawal, and A. Choudhary, "Silverback: Scalable association mining for temporal data in columnar probabilistic databases," in *Proceedings of 30th IEEE International Conference on Data Engineering (ICDE), Industrial and Applications Track* (IEEE, 2014), pp. 1072–1083.
[49] S. Jha, J. Qiu, A. Luckow, P. Mantha, and G. Fox, "A tale of two data-intensive paradigms: Applications, abstractions, and architectures," in *Big Data (BigData Congress), 2014 IEEE International Congress on* (IEEE, 2014), pp. 645–652.
[50] Y. Xie, P. Daga, Y. Cheng, K. Zhang, A. Agrawal, and A. Choudhary, "Reducing infrequent-token perplexity via variational corpora," in *Proceedings of the 53rd Annual Meeting of the Association of Computational Linguistics (ACL) and the 7th International Joint Conference on Natural Language Processing* (ACL Anthology, 2015), pp. 609–615, available at https://aclweb.org/anthology/P/P15/P15-2101.pdf.

[51] G. Linden, B. Smith, and J. York, "Amazon.com recommendations: Item-to-item collaborative filtering," IEEE Internet Comput. **7**, 76–80 (2003).
[52] Y. Zhou, D. Wilkinson, R. Schreiber, and R. Pan, "Large-scale parallel collaborative filtering for the netflix prize," in *Algorithmic Aspects in Information and Management* (Springer, 2008), pp. 337–348.
[53] Y. Xie, D. Honbo, A. Choudhary, K. Zhang, Y. Cheng, and A. Agrawal, "Voxsup: A social engagement framework," in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD) (Demo Paper)* (ACM, 2012), pp. 1556–1559.
[54] H. C. Koh, G. Tan *et al.*, "Data mining applications in healthcare," J. Healthcare Inf. Manage. **19**, 64–72 (2005), available at http://www.ncbi.nlm.nih.gov/pubmed/15869215.
[55] A. Agrawal, S. Misra, R. Narayanan, L. Polepeddi, and A. Choudhary, "Lung cancer survival prediction using ensemble data mining on seer data," Sci. Program. **20**, 29–42 (2012).
[56] J. S. Mathias, A. Agrawal, J. Feinglass, A. J. Cooper, D. W. Baker, and A. Choudhary, "Development of a 5 year life expectancy index in older adults using predictive mining of electronic health record data," J. Am. Med. Inf. Assoc. **20**, e118–e124 (2013).
[57] K. Lee, A. Agrawal, and A. Choudhary, "Real-time disease surveillance using twitter data: Demonstration on flu and cancer," in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '13* (ACM, New York, NY, USA, 2013), pp. 1474–1477.
[58] L. Liu, J. Tang, Y. Cheng, A. Agrawal, W.-k. Liao, and A. Choudhary, "Mining diabetes complication and treatment patterns for clinical decision support," in *Proceedings of 22th ACM International Conference on Information and Knowledge Management (CIKM 2013), San Francisco, USA* (ACM, 2013), pp. 279–288.
[59] K. Lee, A. Agrawal, and A. Choudhary, "Mining social media streams to improve public health allergy surveillance," in *Proceedings of IEEE/ACM International Conference on Social Networks Analysis and Mining (ASONAM)* (IEEE, 2015), pp. 815–822.
[60] C. K. Reddy and C. C. Aggarwal, *Healthcare Data Analytics* (CRC Press, 2015), Vol. 36.
[61] A. R. Ganguly, E. Kodra, A. Agrawal, A. Banerjee, S. Boriah, S. Chatterjee, S. Chatterjee, A. Choudhary, D. Das, J. Faghmous, P. Ganguli, S. Ghosh, K. Hayhoe, C. Hays, W. Hendrix, Q. Fu, J. Kawale, D. Kumar, V. Kumar, W.-k. Liao, S. Liess, R. Mawalagedara, V. Mithal, R. Oglesby, K. Salvi, P. K. Snyder, K. Steinhaeuser, D. Wang, and D. Wuebbles, "Toward enhanced understanding and projections of climate extremes using physics-guided data mining techniques," Nonlinear Processes Geophys. **21**, 777–795 (2014).
[62] C. Jin, Q. Fu, H. Wang, W. Hendrix, Z. Chen, A. Agrawal, A. Banerjee, and A. Choudhary, "Running map inference on million node graphical models: A high performance computing perspective," in *Proceedings of the 15th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid)* (IEEE, 2015), pp. 565–575.
[63] V. Lakshmanan, E. Gilleland, A. McGovern, and M. Tingley, *Machine Learning and Data Mining Approaches to Climate Science* (Springer, 2015).
[64] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped BLAST and PSI-BLAST: A new generation of protein database search programs," Nucl. Acids Res. **25**, 3389–3402 (1997).
[65] A. Agrawal and X. Huang, "PSIBLAST_PairwiseStatSig: Reordering PSI-BLAST hits using pairwise statistical significance," Bioinformatics **25**, 1082–1083 (2009).
[66] A. Agrawal and X. Huang, "Pairwise statistical significance of local sequence alignment using sequence-specific and position-specific substitution matrices," IEEE/ACM Trans. Comput. Biol. Bioinf. **8**, 194–205 (2011).
[67] S. Misra, A. Agrawal, W.-k. Liao, and A. Choudhary, "Anatomy of a hash-based long read sequence mapping algorithm for next generation DNA sequencing," Bioinformatics **27**, 189–195 (2011).
[68] A. ODriscoll, J. Daugelaite, and R. D. Sleator, "big data, hadoop and cloud computing in genomics," J. Biomed. Inf. **46**, 774–781 (2013).
[69] Y. Xie, Z. Chen, K. Zhang, Y. Cheng, D. K. Honbo, A. Agrawal, and A. Choudhary, "Muses: A multilingual sentiment elicitation system for social media data," IEEE Intell. Syst. **29**, 34–42 (2013).
[70] Y. Cheng, A. Agrawal, H. Liu, and A. Choudhary, "Social role identification via dual uncertainty minimization regularization," in *Proceedings of International Conference on Data Mining (ICDM)* (IEEE, 2014), pp. 767–772.
[71] R. Zafarani, M. A. Abbasi, and H. Liu, *Social Media Mining: An Introduction* (Cambridge University Press, 2014).
[72] See http://smds.nims.go.jp/fatigue/index_en.html for National Institute of Materials Science, accessed on Jan 12, 2016.
[73] G. E. Dieter, *Mechanical Metallurgy*, 3rd ed. (Mc Graw-Hill Book Company, 1986).
[74] R. Liu, A. Agrawal, Z. Chen, W. keng Liao, and A. Choudhary, "Pruned search: A machine learning based meta-heuristic approach for constrained continuous optimization," in *Proceedings of 8th IEEE International Conference on Contemporary Computing (IC3)* (IEEE, 2015), pp. 13–18.