Chapter 1
Dimensions, Bits, and Wows in Accelerating Materials Discovery

Lav R. Varshney

Abstract In this book chapter, we discuss how the problem of accelerated materials discovery is related to other computational problems in artificial intelligence, such as computational creativity, concept learning, and invention, as well as to machine-aided discovery in other scientific domains. These connections lead, mathematically, to the emergence of three classes of algorithms that are inspired largely by the approximation-theoretic and machine learning problem of dimensionality reduction, by the information-theoretic problem of data compression, and by the psychology and mass communication problem of holding human attention. The possible utility of functionals including dimension, information [measured in bits], and Bayesian surprise [measured in wows], emerge as part of this description, in addition to measurement of quality in the domain.

1.1 Introduction

Finding new materials with targeted properties is of great importance to technological development in numerous fields including clean energy, national security, resilient infrastructure, and human welfare. Classical approaches to materials discovery rely mainly on trial-and-error, which requires numerous costly and time-intensive experiments. As such, there is growing interest in using techniques from the information sciences in accelerating the process of finding advanced materials such as new metal alloys or thermoelectric materials [1, 2]. Indeed the national *Materials Genome Initiative*—a large-scale collaboration to bring together new digital data, computational tools, and experimental tools—aims to quicken the design and deployment of advanced materials, cf. [3, 4]. In developing these computational tools, there is a

L. R. Varshney (✉)
Coordinated Science Laboratory and Department of Electrical
and Computer Engineering, University of Illinois at Urbana-Champaign,
Urbana 61801, USA
e-mail: varshney@illinois.edu

T. Lookman et al. (eds.), *Materials Discovery and Design*, Springer Series
in Materials Science 280, https://doi.org/10.1007/978-3-319-99465-9_1

desire not only for supercomputing hardware infrastructure [5], but also advanced algorithms.

In most materials discovery settings of current interest, however, the algorithmic challenge is formidable. Due to the interplay between (macro- and micro-) structural and chemical degrees of freedom, computational prediction is difficult and inaccurate. Nevertheless, recent research has demonstrated that emerging statistical inference and machine learning algorithms may aid in accelerating the materials discovery process [1].

The basic process is as follows. Regression algorithms are first used to learn the functional relationship between features and properties from a corpus of some extant characterized materials. Next, an unseen material is tested experimentally and those results are used to enhance the functional relationship model; this unseen material should be chosen as *best* in some sense. Proceeding iteratively, more unseen materials are designed, fabricated, and tested and the model is further refined until a material that satisfies desired properties is obtained. This process is similar to the active learning framework (also called adaptive experimental design) [6], but unlike active learning, here the training set is typically very small: only tens or hundreds of samples as compared to the unexplored space that is combinatorial (in terms of constituent components) and continuous-valued (in terms of their proportions). It should be noted that the ultimate goal is not to learn the functional relationship accurately, but to discover the optimal material with the fewest trials, since experimentation is very costly.

What should be the notion of *best* in iteratively investigating new materials with particular desired properties? This is a *constructive machine learning* problem, where the goal of learning is not to find a good model of data but instead to find one or more particular instances of the domain which are likely to exhibit desired properties. Perhaps the criterion in picking the next sample should be to learn about a useful dimension in the feature space to get a sense of the entire space of possibilities rather than restricting to a small-dimensional manifold [7]. By placing attention on a new dimension of the space, new insights for discovery may be possible [8]. Perhaps the criterion for picking the next sample should be to choose the most informative, as in *infotaxis* in machine learning and descriptions of animal curiosity/behavior [9–13]. Perhaps the goal in driving materials discovery should be to be as surprising as possible, rather than to be as informative as possible, an algorithmic strategy for accelerated discovery one might call *surprise-taxis*. (As we will see, the Bayesian surprise functional is essentially the derivative of Shannon's mutual information [14], and so this can be thought of as a second-order method, cf. [15].)

In investigating these possibilities, we will embed our discussion in the larger framework of data-driven scientific discovery [16, 17] where theory and computation interact to direct further exploration. The overarching aim is to develop a viable *research tool* that is of relevance to materials scientists in a variety of industries, and perhaps even to researchers in further domains like drug cocktail discovery. The general idea is to provide researchers with cognitive support to augment their own intelligence [18], just like other technologies including pencil-and-paper [19, 20] or internet-based tools [21, 22] often lead to greater quality and efficiency of human thought.

When we think about human intelligence, we think about the kinds of abilities that people have, such as memory, deductive reasoning, association, perception, abductive reasoning, inductive reasoning, and problem solving. With technological advancement over the past century, computing technologies have progressed to the stage where they too have many of these abilities. The pinnacle of human intelligence is often said to be creativity and discovery, ensconced in such activities as music composition, scientific research, or culinary recipe design. One might wonder, then, can computational support help people to create and discover novel artifacts and ideas?

In addressing this question, we will take inspiration from related problems including computational creativity, concept learning, and invention, as well as from machine-aided discovery in other scientific domains. Connections to related problems lead, mathematically, to the emergence of three classes of accelerated discovery algorithms that are inspired largely by the approximation-theoretic [23] and machine learning problem of dimensionality reduction [24], by the information-theoretic problem of data compression [25, 26], and by the psychology and mass communication problem of holding human attention. The possible utility of functionals including dimension, information [measured in bits], and Bayesian surprise [measured in wows], emerge as part of this description, in addition to measurement of quality in the domain. It should be noted that although demonstrated in other creative and scientific domains, accelerated materials discovery approaches based on these approximation-theoretic and information-theoretic functionals remain speculative.

1.2 Creativity and Discovery

Whether considering literary manuscripts, musical compositions, culinary recipes, or scientific ideas, the basic argument framing this chapter is that it is indeed possible for computers to create novel, high-quality ideas or artifacts, whether operating autonomously or semi-autonomously by engaging with people. As one typical example, consider a culinary computational creativity system that uses repositories of existing recipes, data on the chemistry of food, and data on human hedonic perception of flavor to create new recipes that have never been cooked before, but that are flavorful [27–29]. As another example, consider a machine science system that takes the scientific literature in genomics, generates hypotheses, and tests them automatically to create new scientific knowledge [30]. Some classical examples of computational creativity include AARON, which creates original artistic images that have been exhibited in galleries around the world [31], and BRUTUS, which tells stories [32]. Several new applications, theories, and trends are now emerging in the field of computational creativity [33–35].

Although several specific algorithmic techniques have been developed in the literature, the basic structure of many computational creativity algorithms proceed by first taking existing artifacts from the domain of interest and intelligently performing a variety of transformations and modifications to generate new ideas; the design space has combinatorial complexity [36]. Next, these generated possibilities are assessed to predict if people would find them compelling as creative artifacts and the best are chosen. Some algorithmic techniques combine the generative and selective steps into a single optimization procedure.

A standard definition of creativity emerging in the psychology literature [37] is that: *Creativity is the generation of an idea or artifact that is judged to be novel and also to be appropriate, useful, or valuable by a suitably knowledgeable social group.* A critical aspect of any creativity algorithm is therefore determining a meaningful characterization of what constitutes a good artifact in the two distinct dimensions of novelty and utility. Note that each domain—whether literature or culinary art—has its own specific metrics for quality. However, independent of domain, people like to be surprised and there may be abstract information-theoretic measures for surprise [14, 38–40].

Can this basic approach to computational creativity be applied to accelerating discovery through machine science [41]? Most pertinently, one might wonder whether novelty and surprise are essential to problems like accelerating materials discovery, or is utility the only consideration. The wow factor of newly creative things or newly discovered facts is important in regimes with an excess of potential creative artifacts or growing scientific literature, not only for ensuring novelty but also for capturing people's attention. More importantly, however, it is important for pushing discovery into wholly different parts of the creative space than other computational/algorithmic techniques can. Designing for surprise is of utmost importance.

For machine science in particular, the following analogy to the three layers of communication put forth by Warren Weaver [42] seems rather apt.

Level A (The technical problem)
Communication: How accurately can the symbols of communication be transmitted? *Machine Science*: How accurately does gathered data represent the state of nature?
Level B (The semantic problem)
Communication: How precisely do the transmitted symbols convey the desired meaning? *Machine Science*: How precisely does the measured data provide explanation into the nature of the world?
Level C (The effectiveness problem)
Communication: How effectively does the received meaning affect conduct in the desired way? *Machine Science*: How surprising are the insights that are learned?

A key element of machine science is therefore not just producing accurate and explanatory data, but insights that are surprising as compared to current scientific understanding.

In the remainder of the chapter, we introduce three basic approaches to discovery algorithms, based on dimensions, information, and surprise.

1.3 Discovering Dimensions

One of the central problems in unsupervised machine learning for understanding, visualization, and further processing has been *manifold learning* or *dimensionality reduction*. The basic idea is to assume that a given set of data points that have some underlying low-dimensional structure are embedded in a high-dimensional Euclidean space, and the goal is to recover that low-dimensional structure. Note that the low-dimensional structure can be much more general than a classical smooth manifold [43, 44]. Such machine learning-based approaches generalize, in some sense, classical harmonic analysis and approximation theory where a fixed representation, say a truncated representation in the Fourier basis, is used as a low-dimensional representation [23].

The most classical approach, principal components analysis (PCA) [45, 46], is a linear transformation of data defined so the first principal component has the largest possible variance, accounting for as much of the data variability as possible. The second principal component has the highest variance possible under the constraint that it is orthogonal to the first principal component, and so on. This linear transformation method, accomplished by computing an eigenbasis, also turns possibly correlated variables into values of linearly uncorrelated variables. It can be extended to work with missing data [47]. One of the distinguishing features of PCA is that the learned transformation can be applied directly to data that was not used to train the transformation, so-called *out-of-sample extension*.

There are several nonlinear dimensionality reduction algorithms that first construct a sparsely-connected graph representation of local affinity among data points and then embed these points into a low-dimensional space, trying to preserve as much of the original affinity as possible. Examples include locally linear embedding [48], multidimensional scaling methods that try to preserve global information such as Isomap [49], spectral embeddings such as Laplacian eigenmaps [50], and stochastic neighbor embedding [51]. Direct out-of-sample extension is not possible with these techniques, and so further techniques such as the Nyström approximation are needed [52].

Another approach that supports direct out-of-sample extension is dimensionality reduction using an autoencoder. An autoencoder is a feedforward neural network that is trained to approximate the identity function, such that it maps a vector of values to itself. When used for dimensionality reduction, a hidden layer in the network is constrained to contain only a small number of neurons and so the network must learn to encode the vector into a small number of dimensions and then decode it back. Consequently, the first part of the network maps from high to low-dimensional space, and the second maps in the reverse manner.

With this background on dimensionality reduction, we can now present an accelerated discovery algorithm that essentially pursues dimensions in order to prioritize investigation of data. This Discovery through Eigenbasis Modeling of Uninteresting Data (DEMUD) algorithm, due to Wagstaff et al. [7], is essentially based on PCA and is meant not just to prioritize data for investigation but also provide domain-specific explanations for why a given item is potentially interesting. The reader will notice the fact that novel discovery algorithms could be developed using other dimensionality reduction techniques that can be updated and with direct out-of-sample extension in place of PCA, for example using autoencoders.

The basic idea of DEMUD is to use a notion of uninterestingness to judge what to select next. Data that has already been seen, data that is not of interest due to its category, or prior knowledge of uninterestingness are all used to iteratively model what should be ignored in selecting a new item of high interest. The specific technique used is to first compute a low-dimensional eigenbasis of uninteresting items using a singular value decomposition $U\Sigma V^T$ of the original dataset X and retaining the top k singular vectors (ranked by magnitude of the corresponding singular value). Data items are then ranked according to the reconstruction error in representing in this basis: items with largest error are said to have the most potential to be novel, as they are largely in an unmodeled dimension of the space. In order to initialize, we use the whole dataset, but then proceed iteratively in building up the eigenbasis. Specifically, the DEMUD algorithm takes the following three inputs: $X \in \mathbb{R}^{n \times d}$ as the input data, $X_U = \emptyset$ as the initial set of uninteresting items, and k as the number of principal components to be used in X_U. Then it proceeds as follows.

Algorithm 1 DEMUD [7]

```
Let $U = SVD(X, k)$ be the initial model of $X_U$ and let $\mu$ be the mean of the data
**while** discovery is to continue and $X \neq \emptyset$ **do**
    Compute reconstructions $\hat{x} = UU^T(x - \mu) + \mu$ for all $x \in X$
    Compute error in reconstructions $R(x) = \|x - \hat{x}\|_2 = \|x - (UU^T(x - \mu) + \mu)\|_2$ for all $x \in X$
    Choose $x' = \text{argmax}_{x \in X} R(x)$ to investigate next
    Remove this data item from the data set and add it to the model, i.e. $X = X \backslash \{x'\}$ and $X_U = X_U \cup \{x'\}$.
    Update $U$ and $\mu$ by using the incremental SVD algorithm [53] with inputs $(U, x', k)$.
**end while**
```

The ordering of data to investigate that emerges from the DEMUD algorithm is meant to quickly identify rare items of scientific value, maintain diversity in its selections, and also provide explanations (in terms of dimensions/subspaces to explore) to aid in human understanding. The algorithm has been demonstrated using hyperspectral data for exploring rare minerals in planetary science [7].

1.4 Infotaxis

Having discussed how the pursuit of novel dimensions in the space of data may accelerate scientific discovery, we now discuss how pursuit of information may do likewise. In Shannon information theory, the *mutual information* functional emerges from the noisy channel coding theorem in characterizing the limits of reliable communication in the presence of noise [54] and from the rate-distortion theorem in characterizing the limits of data compression [55]. In particular, the notion of information rate (e.g. measured in bits) emerges as a universal interface for communication systems. For two continuous-valued random variables, $X \in \mathscr{X}$ and $Y \in \mathscr{Y}$ with corresponding joint density $f_{XY}(x, y)$ and marginals $f_X(x)$ and $f_Y(y)$, the mutual information is given as

$$I(X; Y) = \int_{\mathscr{Y}} \int_{\mathscr{X}} f_{XY}(x, y) \log \frac{f_{XY}(x, y)}{f_X(x) f_Y(y)} dx dy.$$

If the base of the logarithm is chosen as 2, then the units of mutual information are bits. The mutual information can also be expressed as the difference between an unconditional entropy and a conditional one.

There are several methods for estimating mutual information from data, ranging from plug-in estimators for discrete-valued data to much more involved minimax estimators [56] and ensemble methods [57]. For continuous-valued data, there are a variety of geometric and statistical techniques that can also be used [58, 59].

Mutual information is often used to measure informativeness even outside the communication settings where the theorems are proven, since it is a useful measure of mutual dependence that indicates how much knowing one variable reduces uncertainty about the other. Indeed, there is an axiomatic derivation of the mutual information measure, where it is shown that it is the unique (up to choice of logarithm base) function that satisfies certain properties such as continuity, strong additivity, and an increasing-in-alphabet-size property. In fact, there are several derivations with differing small sets of axioms [60].

Of particular interest here is the pursuit of information as a method of discovery, in an algorithm that is called infotaxis [9–13]. The infotaxis algorithm was first explicitly discussed in [9] who described it as a model for animal foraging behavior. The basic insight of the algorithm is that it is a principled way to essentially encode exploration-exploitation trade-offs in search/discovery within an uncertain environment, and therefore has strong connections to reinforcement learning. There is a given but unknown (to the algorithm) probability distribution for the location of the source being searched for and the rate of information acquisition is also the rate of entropy reduction. The basic issue in discovering the source is that the underlying probability distribution is not known to the algorithm but must be estimated from available data. Accumulation of information allows a tighter estimate of the source distribution. As such, the searcher must choose either to move to the most likely source location or to pause and gather more information to make a better estimate of the source. Infotaxis allows a balancing of these two concerns by choosing to move (or stay still) in the direction that maximizes the expected reduction in entropy.

As noted, this algorithmic idea has been used to explain a variety of human/animal curiosity behaviors and also been used in several engineering settings.

1.5 Pursuit of Bayesian Surprise

Rather than moving within a space to maximize expected gain of information (maximize expected reduction of entropy), would it ever make sense to consider maximizing surprise instead. In the common use of the term, pursuit of surprise seems to indicate a kind of curiosity that would be beneficial for accelerating discovery, but is there a formal view of surprise as there is for information? How can we compute whether something is likely to be perceived as surprising?

A particularly interesting definition is based on a psychological and information-theoretic measure termed *Bayesian surprise*, due originally to Itti and Baldi [38, 40]. The surprise of each location on a feature map is computed by comparing beliefs about what is likely to be in that location before and after seeing the information. Indeed, novel and surprising stimuli spontaneously attract attention [61].

An artifact that is surprising is novel, has a wow factor, and changes the observer's world view. This can be quantified by considering a prior probability distribution of existing ideas or artifacts and the change in that distribution after the new artifact is observed, i.e. the posterior probability distribution. The difference between these distributions reflects how much the observer's world view has changed. It is important to note that surprise and saliency depend heavily on the observer's existing world view, and thus the same artifact may be novel to one observer and not novel to another. That is why Bayesian surprise is measured as a change in the observer's specific prior probability distribution of known artifacts.

Mathematically, the cognitively-inspired Bayesian surprise measure is defined as follows. Let $\mathscr{M}$ be the set of artifacts known to the observer, with each artifact in this repository being $M \in \mathscr{M}$. Furthermore, a new artifact that is observed is denoted D. The probability of an existing artifact is denoted $p(M)$, the conditional probability of the new artifact given the existing artifacts is $p(D|M)$, and via Bayes' theorem the conditional probability of the existing artifacts given the new artifact is $p(M|D)$. The Bayesian surprise is defined as the following relative entropy (Kullback-Leibler divergence):

$$s = D(p(M|D)||p(M)) = \int_{\mathscr{M}} p(M|D) \log \frac{p(M|D)}{p(M)} dM$$

One might wonder if Bayesian surprise, $s(D)$, has anything to do with measures of information such as Shannon's mutual information given in the previous section. In fact, if there is a definable distribution on new artifacts $q(D)$, the expected value of Bayesian surprise is the Shannon mutual information.

$$E[s(D)] = \int q(D) D(p(M|D)||p(M)) dD = \int \int_{\mathscr{M}} p(M, D) \log \frac{p(M|D)}{p(M)} dM dD,$$

which by definition is the Shannon mutual information $I(M; D)$. The fact that the average of the Bayesian surprise equals the mutual information points to the notion that surprise is essentially the derivative of information.

Let us define the weak derivative, which arises in the weak-* topology [62], as follows.

Definition Let $\mathscr{A}$ be a vector space, and f a real-valued functional defined on domain $\Omega \subset \mathscr{A}$, where Ω is a convex set. Fix an $a_0 \in \Omega$ and let $\theta \in [0, 1]$. If there exists a map $f'_{a_0} : \Omega \to \mathbb{R}$ such that

$$f'_{a_0}(a) = \lim_{\theta \downarrow 0} \frac{f[(1-\theta)a_0 + \theta a] - f(a_0)}{\theta}$$

for all $a \in \Omega$, then f is said to be *weakly differentiable in* Ω *at* a_0 and f'_{a_0} is the *weak derivative in* Ω *at* a_0.

If f is weakly differentiable in Ω at a_0 for all a_0 in Ω, then f is said to be *weakly differentiable*.

The precise relationship can be formalized as follows. For a fixed reference distribution $F_0 = q(D)$, the weak derivative of mutual information is:

$$I'_{F_0}(F) = \lim_{\theta \downarrow 0} \frac{(I((1-\theta)F_0 + \theta F_0) - I(F_0))}{\theta} = \int s(x)q(x)dx - I(F_0)$$

Indeed, even the Shannon capacity C of communication over a stochastic kernel $p(M|D)$ can be expressed in terms of the Bayesian surprise [63]:

$$C = \max_{q(D)} I(M; D) = \min_{p(M)} \max_{d \in \mathscr{M}} s(d),$$

therefore all communicated signals should be equally surprising when trying to maximize information rate of communication.

These formalisms are all well and good, but it is also important to have operational meaning for Bayesian surprise to go alongside. In fact, there are several kinds of operational meanings that have been established in a variety of fields.

- In defining Bayesian surprise, Itti and Baldi also performed several psychology experiments that demonstrated its connection to attraction of human attention across different spatiotemporal scales, modalities, and levels of abstraction [39, 40]. As a typical example of a such an experiment, human subjects were tasked with looking at a video of a soccer game while being measured using eye-tracking. The Bayesian surprise for the video was also computed. The places where the Bayesian surprise was large was also where the human subjects were looking. These classes of experiments have been further studied by several other research groups in psychology, e.g. [64–67].
- Bayesian surprise has not just been observed at a behavioral level, but also at a neurobiological level [68–70], where various brain processes concerned with attention have been related to Bayesian surprise.

- In the engineering of computational creativity systems, it has empirically been found that Bayesian surprise is a useful optimization criterion for ideas or artifacts to be rated as highly creative [27–29, 71]. Likewise in marketing [72], Bayesian surprise has been found to be an effective criterion for designing promotion campaigns [73].
- In the Bayesian model comparison literature, Bayesian surprise is also called *complexity* [74] and in thermodynamic formulations of Bayesian inference [75], an increase in Bayesian surprise is necessarily associated with a decrease in free-energy due to a reduction in prediction error. It should ne noted, however, that Bayes-optimal inference schemes do not optimize for Bayesian surprise in itself [74].
- In information theory, Bayesian surprise is sometimes called the *marginal information density* [76]. When communicating in information overload regimes, it is necessary for messages to not only provide information but also to attract attention in the first place. In many communication settings, the flood of messages is not only immense but also monotonously similar. Some have argued that "it would be far more effective to send one very unusual message than a thousand typical ones" [77]. The Bayesian surprise therefore arises in information-theoretic studies of optimal communication systems. One example is in highly-asynchronous communication, where the receiver must monitor the channel for long stretches of time before a transmitted signal appears [78]. Moreover, we have shown that Bayesian surprise is the natural cost function for communication just like log-loss [79] is the natural fidelity criterion for compression [14] (as follows from KKT conditions [80]). One can further note that there is a basic tradeoff between messages being informative and being surprising [14].

Given that Bayesian surprise has operational significance in a variety of psychology, neurobiology, statistics, creativity, and communication settings, as well as formal derivative relationships to mutual information, one might wonder if an accelerated discovery algorithm that aims to maximize Bayesian surprise might be effective. In particular, could surprise-taxis be a kind of second-order version of infotaxis? This direction may be promising since recent algorithms in accelerated materials discovery [81] imitate the human discovery process, e.g. by using an adaptive scheme based on Support Vector Regression (SVR) and Efficient Global Optimization (EGO) [82] and demonstrating on a certain family of alloys, M_2AX phases [83].

In developing a surprise-taxis algorithm for materials discovery, however, one may need to explicitly take notions of quality into account, rather than just pure novelty concerns, since there may be large parts of the discovery space that have low-quality possibilities: a Lagrangian balance between differing objectives of surprise and quality.

1.6 Conclusion

Although mathematically distinct, various problems in machine learning and artificial intelligence such as computational creativity, concept learning [84], invention, and accelerated discovery are all quite closely related philosophically. In this chapter, we have suggested that there may be value in bringing algorithmic ideas from these other related problems into accelerated materials discovery, especially the conceptual ideas of using dimensions, information, and surprise as key metrics for algorithmic pursuit.

It is an open question whether any of these ideas will be effective, as they have been in their original domains that include exploring minerals on distant planets [7], modeling the exploratory behavior of organisms such as moths and worms [9, 11], and creating novel and flavorful culinary recipes [27–29]. The data and informatics resources that are emerging in materials science, however, provide a wonderful opportunity to test this algorithmic hypothesis.

Acknowledgements Discussions with Daewon Seo, Turab Lookman, and Prasanna V. Balachandran are appreciated. Further encouragement from Turab Lookman in preparing this book chapter, despite the preliminary status of the work itself, is acknowledged.

References

1. T. Lookman, F.J. Alexander, K. Rajan (eds.), *Information Science for Materials Discovery and Design* (Springer, New York, 2016)
2. T.D. Sparks, M.W. Gaultois, A. Oliynyk, J. Brgoch, B. Meredig, Data mining our way to the next generation of thermoelectrics. Scripta Materialia **111**, 10–15 (2016)
3. A. Jain, S.P. Ong, G. Hautier, W. Chen, W.D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, K.A. Persson, The materials project: a materials genome approach to accelerating materials innovation. APL Mater. **1**(1), 011002 (2013)
4. M.L. Green, C.L. Choi, J.R. Hattrick-Simpers, A.M. Joshi, I. Takeuchi, S.C. Barron, E. Campo, T. Chiang, S. Empedocles, J.M. Gregoire, A.G. Kusne, J. Martin, A. Mehta, K. Persson, Z. Trautt, J. Van Duren, A. Zakutayev, Fulfilling the promise of the materials genome initiative with high-throughput experimental methodologies. Appl. Phys. Rev. **4**(1), 011105 (2017)
5. S. Curtarolo, G.L.W. Hart, M.B. Nardelli, N. Mingo, S. Sanvito, O. Levy, The high-throughput highway to computational materials design. Nat. Mater. **12**(3), 191–201 (2013)
6. B. Settles, Active learning literature survey. University of Wisconsin–Madison, Computer Sciences Technical Report 1648, 2009
7. K.L. Wagstaff, N.L. Lanza, D.R. Thompson, T.G. Dietterich, M.S. Gilmore, Guiding scientific discovery with explanations using DEMUD, in *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*, July 2013, pp. 905–911
8. J. Schwartzstein, Selective attention and learning. J. Eur. Econ. Assoc. **12**(6), 1423–1452 (2014)
9. M. Vergassola, E. Villermaux, B.I. Shraiman, 'Infotaxis' as a strategy for searching without gradients. Nature **445**(7126), 406–409 (2007)
10. J.L. Williams, J.W. Fisher III, A.S. Willsky, Approximate dynamic programming for communication-constrained sensor network management. IEEE Trans. Signal Process. **55**(8), 4300–4311 (2007)
11. A.J. Calhoun, S.H. Chalasani, T.O. Sharpee, Maximally informative foraging by *Caenorhabditis elegans*. eLife **3**, e04220 (2014)

12. R. Aggarwal, M.J. Demkowicz, Y.M. Marzouk, Information-driven experimental design in materials science, in *Information Science for Materials Discovery and Design*, ed. by T. Lookman, F.J. Alexander, K. Rajan (Springer, New York, 2016), pp. 13–44
13. K.J. Friston, M. Lin, C.D. Frith, G. Pezzulo, Active inference, curiosity and insight. Neural Comput. **29**(10), 2633–2683 (2017)
14. L.R. Varshney, To surprise and inform, in *Proceedings of the 2013 IEEE International Symposium on Information Theory*, July 2013, pp. 3145–3149
15. N. Agarwal, B. Bullins, E. Hazan, Second-order stochastic optimization for machine learning in linear time. J. Mach. Learn. Res. **18**(116), 1–40 (2017)
16. A. Karpatne, G. Atluri, J.H. Faghmous, M. Steinbach, A. Banerjee, A. Ganguly, S. Shekhar, N. Samatova, V. Kumar, Theory-guided data science: a new paradigm for scientific discovery from data. IEEE Trans. Knowl. Data Eng. **29**(10), 2318–2331 (2017)
17. V. Pankratius, J. Li, M. Gowanlock, D.M. Blair, C. Rude, T. Herring, F. Lind, P.J. Erickson, C. Lonsdale, Computer-aided discovery: toward scientific insight generation with machine support. IEEE Intell. Syst. **31**(4), 3–10 (2016)
18. B.F. Jones, The burden of knowledge and the 'death of the renaissance man': Is innovation getting harder? Rev. Econ. Stud. **76**(1), 283–317 (2009)
19. R. Netz, *The Shaping of Deduction in Greek Mathematics: A Study in Cognitive History* (Cambridge University Press, Cambridge, 1999)
20. L.R. Varshney, Toward a comparative cognitive history: Archimedes and D.H.J. Polymath, in *Proceedings of the Collective Intelligence Conference 2012*, Apr 2012
21. W.W. Ding, S.G. Levin, P.E. Stephan, A.E. Winkler, The impact of information technology on academic scientists' productivity and collaboration patterns. Manag. Sci. **56**(9), 1439–1461 (2010)
22. L.R. Varshney, The Google effect in doctoral theses. Scientometrics **92**(3), 785–793 (2012)
23. G.G. Lorentz, M. Golitschek, Y. Makovoz, *Constructive Approximation: Advanced Problems* (Springer, Berlin, 2011)
24. J.A. Lee, M. Verleysen, *Nonlinear Dimensionality Reduction* (Springer, New York, 2007)
25. T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression* (Prentice-Hall, Englewood Cliffs, NJ, 1971)
26. D.L. Donoho, M. Vetterli, R.A. DeVore, I. Daubechies, Data compression and harmonic analysis. IEEE Trans. Inf. Theory **44**(6), 2435–2476 (1998)
27. L.R. Varshney, F. Pinel, K.R. Varshney, D. Bhattacharjya, A. Schörgendorfer, Y.-M. Chee, A big data approach to computational creativity (2013). arXiv:1311.1213v1 [cs.CY]
28. F. Pinel, L.R. Varshney, Computational creativity for culinary recipes, in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI 2014)*, Apr 2014, pp. 439–442
29. F. Pinel, L.R. Varshney, D. Bhattacharjya, A culinary computational creativity system, in *Computational Creativity Research: Towards Creative Machines*, ed. by T.R. Besold, M. Schorlemmer, A. Smaill (Springer, 2015), pp. 327–346
30. R.D. King, J. Rowland, S.G. Oliver, M. Young, W. Aubrey, E. Byrne, M. Liakata, M. Markham, P. Pir, L.N. Soldatova, A. Sparkes, K.E. Whelan, A. Clare, The automation of science. Science **324**(5923), 85–89 (2009)
31. H. Cohen, The further exploits of AARON, painter, in *Constructions of the Mind: Artificial Intelligence and the Humanities*, ser. Stanford Humanities Review, vol. 4, no. 2, ed. by S. Franchi, G. Güzeldere (1995), pp. 141–160
32. S. Bringsjord, D.A. Ferrucci, *Artificial Intelligence and Literary Creativity: Inside the Mind of BRUTUS, a Storytelling Machine* (Lawrence Erlbaum Associates, Mahwah, NJ, 2000)
33. M.A. Boden, *The Creative Mind: Myths and Mechanisms*, 2nd edn. (Routledge, London, 2004)
34. A. Cardoso, T. Veale, G.A. Wiggins, Converging on the divergent: the history (and future) of the international joint workshops in computational creativity. A. I. Mag. **30**(3), 15–22 (2009)
35. M.A. Boden, Foreword, in *Computational Creativity Research: Towards Creative Machines*, ed. by T.R. Besold, M. Schorlemmer, A. Smaill (Springer, 2015), pp. v–xiii

36. M. Guzdial, M.O. Riedl, Combinatorial creativity for procedural content generation via machine learning, in *Proceedings of the AAAI 2018 Workshop on Knowledge Extraction in Games*, Feb 2018 (to appear)
37. R.K. Sawyer, *Explaining Creativity: The Science of Human Innovation* (Oxford University Press, Oxford, 2012)
38. L. Itti, P. Baldi, Bayesian surprise attracts human attention, in *Advances in Neural Information Processing Systems 18*, ed. by Y. Weiss, B. Schölkopf, J. Platt (MIT Press, Cambridge, MA, 2006), pp. 547–554
39. L. Itti, P. Baldi, Bayesian surprise attracts human attention. Vis. Res. **49**(10), 1295–1306 (2009)
40. P. Baldi, L. Itti, Of bits and wows: a Bayesian theory of surprise with applications to attention. Neural Netw. **23**(5), 649–666 (2010)
41. J. Evans, A. Rzhetsky, Machine science. Science **329**(5990), 399–400 (2010)
42. C.E. Shannon, W. Weaver, *The Mathematical Theory of Communication* (University of Illinois Press, Urbana, 1949)
43. N. Verma, S. Kpotufe, S. Dasgupta, Which spatial partition trees are adaptive to intrinsic dimension?, in *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence (UAI '09)*, June 2009, pp. 565–574
44. M. Tepper, A.M. Sengupta, D.B. Chklovskii, Clustering is semidefinitely not that hard: nonnegative SDP for manifold disentangling (2018). arXiv:1706.06028v3 [cs.LG]
45. K. Pearson, On lines and planes of closest fit to systems of points in space. Lond. Edinb. Dublin Philos. Mag. J. Sci. **2**(11), 559–572 (1901)
46. H. Hotelling, Analysis of a complex of statistical variables into principal components. J. Educ. Psychol. **24**(6), 417–441 (1933)
47. S. Bailey, Principal component analysis with noisy and/or missing data. Publ. Astron. Soc. Pac. **124**(919), 1015–1023 (2012)
48. S.T. Roweis, L.K. Saul, Nonlinear dimensionality reduction by locally linear embedding. Science **290**(5500), 2323–2326 (2000)
49. J.B. Tenenbaum, V. de Silva, J.C. Langford, A global geometric framework for nonlinear dimensionality reduction. Science **290**(5500), 2319–2323 (2000)
50. M. Belkin, P. Niyogi, Laplacian eigenmaps for dimensionality reduction and data representation. Neural Comput. **15**(6), 1373–1396 (2003)
51. L. van der Maaten, G. Hinton, Visualizing data using t-SNE. J. Mach. Learn. Res. **9**, 2579–2605 (2008)
52. Y. Bengio, J.-F. Paiement, P. Vincent, O. Delalleau, N.L. Roux, M. Ouimet, Out-of-sample extensions for LLE, Isomap, MDS, eigenmaps, and spectral clustering, in *Advances in Neural Information Processing Systems 16*, ed. by S. Thrun, L.K. Saul, B. Sch (2003)
53. J. Lim, D.A. Ross, R. Lin, M.-H. Yang, Incremental learning for visual tracking, in *Advances in Neural Information Processing Systems 17*, ed. by L.K. Saul, Y. Weiss, L. Bottou (MIT Press, 2005), pp. 793–800
54. C.E. Shannon, A mathematical theory of communication. Bell Syst. Tech. J. **27**, 379–423, 623–656 (1948)
55. C.E. Shannon, Coding theorems for a discrete source with a fidelity criterion. IRE Natl. Conv. Rec. (Part 4), 142–163 (1959)
56. J. Jiao, K. Venkat, Y. Han, T. Weissman, Minimax estimation of functionals of discrete distributions. IEEE Trans. Inf. Theory **61**(5), 2835–2885 (2015)
57. K.R. Moon, A.O. Hero, III, Multivariate f-divergence estimation with confidence, in *Advances in Neural Information Processing Systems 27*, ed. by Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, K.Q. Weinberger (MIT Press, 2014), pp. 2420–2428
58. A.O. Hero III, B. Ma, O.J.J. Michel, J. Gorman, Applications of entropic spanning graphs. IEEE Signal Process. Mag. **19**(5), 85–95 (2002)
59. Q. Wang, S.R. Kulkarni, S. Verdú, Universal estimation of information measures for analog sources. Found. Trends Commun. Inf. Theory **5**(3), 265–353 (2009)
60. J. Aczél, Z. Daróczy, *On Measures of Information and Their Characterization* (Academic Press, New York, 1975)

61. D. Kahneman, *Attention and Effort* (Prentice-Hall, Englewood Cliffs, NJ, 1973)
62. D.G. Luenberger, *Optimization by Vector Space Methods* (Wiley, New York, 1969)
63. I. Csiszár, J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, 3rd edn. (Akadémiai Kiadó, Budapest, 1997)
64. E. Hasanbelliu, K. Kampa, J.C. Principe, J.T. Cobb, Online learning using a Bayesian surprise metric, in *Proceedings of the 2012 International Joint Conference on Neural Networks (IJCNN)*, June 2012
65. B. Schauerte, R. Stiefelhagen, "Wow!" Bayesian surprise for salient acoustic event detection, in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2013)*, May 2013, pp. 6402–6406
66. K. Takahashi, K. Watanabe, Persisting effect of prior experience of change blindness. Perception **37**(2), 324–327 (2008)
67. T.N. Mundhenk, W. Einhuser, L. Itti, Automatic computation of an image's statistical surprise predicts performance of human observers on a natural image detection task. Vis. Res. **49**(13), 1620–1637 (2009)
68. D. Ostwald, B. Spitzer, M. Guggenmos, T.T. Schmidt, S.J. Kiebel, F. Blankenburg, Evidence for neural encoding of Bayesian surprise in human somatosensation. NeuroImage **62**(1), 177–188 (2012)
69. T. Sharpee, N.C. Rust, W. Bialek, Analyzing neural responses to natural signals: maximally informative dimensions. Neural Comput. **16**(2), 223–250 (2004)
70. G. Horstmann, The surprise-attention link: a review. Ann. New York Acad. Sci. **1339**, 106–115 (2015)
71. C. França, L.F.W. Goes, Á. Amorim, R. Rocha, A. Ribeiro da Silva, Regent-dependent creativity: a domain independent metric for the assessment of creative artifacts, in *Proceedings of the International Conference on Computational Creativity (ICCC 2016)*, June 2016, pp. 68–75
72. J.P.L. Schoormans, H.S.J. Robben, The effect of new package design on product attention, categorization and evaluation. J. Econ. Psychol. **18**(2–3), 271–287 (1997)
73. W. Sun, P. Murali, A. Sheopuri, Y.-M. Chee, Designing promotions: consumers' surprise and perception of discounts. IBM J. Res. Dev. **58**(5/6), 2:1–2:10 (2014)
74. H. Feldman, K.J. Friston, Attention, uncertainty, and free-energy. Front. Hum. Neurosci. **4**, 215 (2010)
75. K. Friston, The free-energy principle: a rough guide to the brain? Trends Cogn. Sci. **13**(7), 293–301 (2009)
76. J.G. Smith, The information capacity of amplitude- and variance-constrained scalar Gaussian channels. Inf. Control **18**(3), 203–219 (1971)
77. T.H. Davenport, J.C. Beck, *The Attention Economy: Understanding the New Currency of Business* (Harvard Business School Press, Boston, 2001)
78. V. Chandar, A. Tchamkerten, D. Tse, Asynchronous capacity per unit cost. IEEE Trans. Inf. Theory **59**(3), 1213–1226 (2013)
79. T.A. Courtade, T. Weissman, Multiterminal source coding under logarithmic loss. IEEE Trans. Inf. Theory **60**(1), 740–761 (2014)
80. M. Gastpar, B. Rimoldi, M. Vetterli, To code, or not to code: lossy source-channel communication revisited. IEEE Trans. Inf. Theory **49**(5), 1147–1158 (2003)
81. P.V. Balachandra, D. Xue, J. Theiler, J. Hogden, T. Lookman, Adaptive strategies for materials design using uncertainties. Sci. Rep. **6**, 19660 (2016)
82. D.R. Jones, M. Schonlau, W.J. Welch, Efficient global optimization of expensive black-box functions. J. Glob. Optim. **13**(4), 455–492 (1998)
83. M.F. Cover, O. Warschkow, M.M.M. Bilek, D.R. McKenzie, A comprehensive survey of M_2AX phase elastic properties. J. Phys.: Condens. Matter **21**(30), 305403 (2009)
84. H. Yu and L.R. Varshney, Towards deep interpretability (MUS-ROVER II): learning hierarchical representations of tonal music, in *Proceedings of the 6th International Conference on Learning Representations (ICLR)*, Apr 2017