

Turab Lookman
Stephan Eidenbenz
Frank Alexander
Cris Barnes *Editors*

Materials Discovery and Design

By Means of Data Science and Optimal
Learning



Springer

Springer Series in Materials Science

Volume 280

Series editors

Robert Hull, Troy, USA

Chennupati Jagadish, Canberra, Australia

Yoshiyuki Kawazoe, Sendai, Japan

Richard M. Osgood, New York, USA

Jürgen Parisi, Oldenburg, Germany

Udo W. Pohl, Berlin, Germany

Tae-Yeon Seong, Seoul, Republic of Korea (South Korea)

Shin-ichi Uchida, Tokyo, Japan

Zhiming M. Wang, Chengdu, China

The Springer Series in Materials Science covers the complete spectrum of materials physics, including fundamental principles, physical properties, materials theory and design. Recognizing the increasing importance of materials science in future device technologies, the book titles in this series reflect the state-of-the-art in understanding and controlling the structure and properties of all important classes of materials.

More information about this series at <http://www.springer.com/series/856>

Turab Lookman · Stephan Eidenbenz
Frank Alexander · Cris Barnes
Editors

Materials Discovery and Design

By Means of Data Science and Optimal
Learning

Editors

Turab Lookman
Theoretical Division
Los Alamos National Laboratory
Los Alamos, NM, USA

Stephan Eidenbenz
Los Alamos National Laboratory
Los Alamos, NM, USA

Frank Alexander
Brookhaven National Laboratory
Brookhaven, NY, USA

Cris Barnes
Los Alamos National Laboratory
Los Alamos, NM, USA

ISSN 0933-033X

ISSN 2196-2812 (electronic)

Springer Series in Materials Science

ISBN 978-3-319-99464-2

ISBN 978-3-319-99465-9 (eBook)

<https://doi.org/10.1007/978-3-319-99465-9>

Library of Congress Control Number: 2018952614

© Springer Nature Switzerland AG 2018

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

This book addresses aspects of data analysis and optimal learning as part of the co-design loop for future materials science innovation. The scientific process must cycle between theory and design of experiments and the conduct and analysis of them, in a loop that can be facilitated by more rapid execution. Computational and experimental facilities today generate vast amounts of data at an unprecedented rate. The role of visualization and inference and optimization methods, in distilling the data constrained by materials theory predictions, is key to achieving the desired goals of real-time analysis and control. The importance of this book lies in emphasizing that the full value of knowledge-driven discovery using data can only be realized by integrating statistical and information sciences with materials science, which itself is increasingly dependent on experimental data gathering efforts. This is especially the case as we enter a new era of big data in materials science with initiatives in exascale computation and with the planning and building of future coherent light source facilities such as the upgrade of the Linac Coherent Light Source at Stanford (LCLS-II), the European X-ray Free Electron Laser (XFEL), and Matter Radiation in Extremes (MaRIE), the signature concept facility from Los Alamos National Laboratory. These experimental facilities, as well as present synchrotron light sources being upgraded and used in novel ways, are expected to generate hundreds of terabytes to several petabytes of in situ spatially and temporally resolved data per sample. The questions that then arise include how we can learn from this data to accelerate the processing and analysis of reconstructed microstructure, rapidly map spatially resolved properties from high throughput data, devise diagnostics for pattern detection, and guide experiments toward desired information and create materials with targeted properties or controlled functionality.

The book is an outgrowth of a conference held in Santa Fe, May 16–18, 2016 on “Data Science and Optimal Learning for Materials Discovery and Design”. In addition, we invited a number of other authors active in these efforts, who did not participate in Santa Fe, to also contribute chapters. The authors are an interdisciplinary group of experts who include theorists surveying the open questions and future directions in the application of data science to materials problems, and experimentalists focusing on the challenges associated with obtaining, analyzing,

and learning from data from large-scale user facilities, such as the Advanced Photon Source (APS) and LCLS. We have organized the chapters so that we start with a broad and fascinating perspective from **Lav Varshney** who discusses the relationship between accelerated materials discovery and problems in artificial intelligence, such as computational creativity, concept learning, and invention, as well as machine learning in other scientific domains. He shows how the connections lead to a number of common metrics including “dimension”, information as measured in “bits” and Bayesian surprise, an entropy-related measure measured in “wows”. With the thought-provoking title “Is Automated Materials Design and Discovery Possible?”, **Mike McKerns** suggests that the tools traditionally used for finding materials with desired properties, which often make linear or quadratic approximations to handle the large dimensionality associated with the data, can be limiting as global optimization requires dealing with a highly nonlinear problem. He discusses the merits of the method of “Optimal Uncertainty Quantification” and the software tool *Mystic* as a possible route to handle such shortcomings. The importance of the choice and influence of material descriptors or features on the outcome of machine learning is the focus of the chapter by **Prasanna Balachandran et al.** They consider a number of materials data sets with different sets of features to independently track which of the sets finds most rapidly the compound with the largest target property. They emphasize that a relatively poor machine-learned model with large error but one that contains key features can be more efficient in accelerating the search process than a low-error model that lacks such features.

The bridge to the analysis of experimental data is provided by **Alisa Paterson et al.** who discuss the least squares and Bayesian inference approaches and show how they can be applied to X-ray diffraction data to study structure refinement. By considering single peak and full diffraction pattern fitting, they make the case that Bayesian inference provides a better model and generally affords the ability to escape from local minima and provide quantifiable uncertainties. They employ Markov Chain Monte Carlo algorithms to sample the distribution of parameters to construct the posterior probability distributions. The development of methods for extracting experimentally accessible spatially dependent information on structure and function from probes such as scanning transmission and scanning probe microscopies is the theme of the chapter by **Maxim Ziatdinov et al.** They emphasize the need to cross-correlate information from different experimental channels in physically and statistically meaningful ways and illustrate the use of machine learning and multivariate analysis to allow automated and accurate extraction and mapping of structural and functional material descriptors from experimental datasets. They consider a number of case studies, including strongly correlated materials.

The chapter by **Brian Patterson et al.** provides an excellent overview of the challenges associated with non-destructive 3D imaging and is a segue into the next three chapters also focused on imaging from incoherent and coherent light sources. This work features 3D data under dynamic time dependence at what is currently the most rapid strain rates available with present light sources. The chapter discusses issues and needs in the processing of large datasets of many terabytes in a matter of

days from in situ experiments, and the developments required for automated reconstruction, filtering, segmentation, visualization, and animation, in addition to acquiring appropriate metrics and statistics characterizing the morphologies. **Reeju Pokharel** describes the technique and analysis tools associated with High Energy Diffraction Microscopy (HEDM) for characterizing polycrystalline microstructure under thermomechanical conditions. HEDM captures 3D views in a bulk sample at sub-grain resolution of about one micron. However, reconstruction from the diffraction signals is a computationally very intensive task. One of the challenges here is to develop tools based on machine learning and optimization to accelerate the reconstruction of images and decrease the time to analyze and use results to guide future experiments. The HEDM data can be utilized within a physics-based finite-element model of microstructure.

The final two chapters relate to aspects of light sources, in particular, advances in coherent diffraction imaging and the outstanding issues in the tuning and control of particle accelerators. In particular, **Edwin Fohtung et al.** discuss the recovery of the phase information from coherent diffraction data using iterative feedback algorithms to reconstruct the image of an object. They review recent developments including Bragg Coherent Diffraction Imaging (BCDI) for oxide nanostructures, as well as the big data challenges in BCDI. Finally, **Alex Sheinker** closes the loop by discussing the major challenges faced by future coherent light sources, such as fourth-generation Free Electron Lasers (FELs), in achieving extremely tight constraints on beam quality and in quickly tuning between various experimental setups under control. He emphasizes the need for feedback to achieve this control and outlines an extremum seeking method for automatic tuning and optimization.

The chapters in this book span aspects of optimal learning, from using information theoretic-based methods in the analysis of experimental data, to adaptive control and optimization applied to the accelerators that serve as light sources. Hence, the book is aimed at an interdisciplinary audience, with the subjects integrating aspects of statistics and mathematics, materials science, and computer science. It will be of timely appeal to those interested in learning about this emerging field. We are grateful to all the authors for their articles as well as their support of the editorial process.

Los Alamos, NM, USA
Los Alamos, NM, USA
Los Alamos, NM, USA
Brookhaven, NY, USA

Turab Lookman
Stephan Eidenbenz
Cris Barnes
Frank Alexander

Contents

1	Dimensions, Bits, and Wows in Accelerating Materials	
	Discovery	1
	Lav R. Varshney	
1.1	Introduction	1
1.2	Creativity and Discovery	3
1.3	Discovering Dimensions	5
1.4	Infotaxis	6
1.5	Pursuit of Bayesian Surprise	8
1.6	Conclusion	11
	References	11
2	Is Automated Materials Design and Discovery Possible?	15
	Michael McKerns	
2.1	Model Determination in Materials Science	16
2.1.1	The Status Quo	16
2.1.2	The Goal	16
2.2	Identification of the Research and Issues	17
2.2.1	Reducing the Degrees of Freedom in Model Determination	17
2.2.2	OUQ and mystic	19
2.3	Introduction to Uncertainty Quantification	21
2.3.1	The UQ Problem	21
2.4	Generalizations and Comparisons	24
2.4.1	Prediction, Extrapolation, Verification and Validation	24
2.4.2	Comparisons with Other UQ Methods	25
2.5	Optimal Uncertainty Quantification	27
2.5.1	First Description	28

2.6	The Optimal UQ Problem	31
2.6.1	From Theory to Computation	31
2.7	Optimal Design	36
2.7.1	The Optimal UQ Loop	36
2.8	Model-Form Uncertainty	40
2.8.1	Optimal UQ and Model Error	40
2.8.2	Game-Theoretic Formulation and Model Error	41
2.9	Design and Decision-Making Under Uncertainty	42
2.9.1	Optimal UQ for Vulnerability Identification	42
2.9.2	Data Collection for Design Optimization	43
2.10	A Software Framework for Optimization and UQ in Reduced Search Space	44
2.10.1	Optimization and UQ	44
2.10.2	A Highly-Configurable Optimization Framework	45
2.10.3	Reduction of Search Space	46
2.10.4	New Massively-Parallel Optimization Algorithms	49
2.10.5	Probability and Uncertainty Toolkit	50
2.11	Scalability	53
2.11.1	Scalability Through Asynchronous Parallel Computing	53
	References	54
3	Importance of Feature Selection in Machine Learning and Adaptive Design for Materials	59
	Prasanna V. Balachandran, Dezhen Xue, James Theiler, John Hogden, James E. Gubernatis and Turab Lookman	
3.1	Introduction	60
3.2	Computational Details	62
3.2.1	Density Functional Theory	62
3.2.2	Machine Learning	63
3.2.3	Design	63
3.3	Results	64
3.4	Discussion	73
3.5	Summary	76
	References	77
4	Bayesian Approaches to Uncertainty Quantification and Structure Refinement from X-Ray Diffraction	81
	Alisa R. Paterson, Brian J. Reich, Ralph C. Smith, Alyson G. Wilson and Jacob L. Jones	
4.1	Introduction	81
4.2	Classical Methods of Structure Refinement	83
4.2.1	Classical Single Peak Fitting	83
4.2.2	The Rietveld Method	84

4.2.3	Frequentist Inference and Its Limitations	86
4.3	Bayesian Inference	87
4.3.1	Sampling Algorithms	89
4.4	Application of Bayesian Inference to Single Peak Fitting: A Case Study in Ferroelectric Materials	90
4.4.1	Methods	92
4.4.2	Prediction Intervals	93
4.5	Application of Bayesian Inference to Full Pattern Crystallographic Structure Refinement: A Case Study	94
4.5.1	Data Collection and the Rietveld Analysis	95
4.5.2	Importance of Modelling the Variance and Correlation of Residuals	96
4.5.3	Bayesian Analysis of the NIST Silicon Standard	97
4.5.4	Comparison of the Structure Refinement Approaches	97
4.5.5	Programs	99
4.6	Conclusion	100
	References	101
5	Deep Data Analytics in Structural and Functional Imaging of Nanoscale Materials	103
	Maxim Ziatdinov, Artem Maksov and Sergei V. Kalinin	
5.1	Introduction	104
5.2	Case Study 1. Interplay Between Different Structural Order Parameters in Molecular Self-assembly	106
5.2.1	Model System and Problem Overview	106
5.2.2	How to Find Positions of All Molecules in the Image?	107
5.2.3	Identifying Molecular Structural Degrees of Freedom via Computer Vision	108
5.2.4	Application to Real Experimental Data: From Imaging to Physics and Chemistry	112
5.3	Case Study 2. Role of Lattice Strain in Formation of Electron Scattering Patterns in Graphene	115
5.3.1	Model System and Problem Overview	115
5.3.2	How to Extract Structural and Electronic Degrees of Freedom Directly from an Image?	116
5.3.3	Direct Data Mining of Structure and Electronic Degrees of Freedom in Graphene	117
5.4	Case Study 3. Correlative Analysis in Multi-mode Imaging of Strongly Correlated Electron Systems	121
5.4.1	Model System and Problem Overview	121

5.4.2	How to Obtain Physically Meaningful Endmembers from Hyperspectral Tunneling Conductance Data?	122
5.5	Overall Conclusion and Outlook	126
	References	127
6	Data Challenges of In Situ X-Ray Tomography for Materials Discovery and Characterization	129
	Brian M. Patterson, Nikolaus L. Cordes, Kevin Henderson, Xianghui Xiao and Nikhilesh Chawla	
6.1	Introduction	130
6.2	In Situ Techniques	133
6.3	Experimental Rates	136
6.4	Experimental and Image Acquisition	141
6.5	Reconstruction	145
6.6	Visualization	146
6.7	Segmentation	148
6.8	Modeling	151
6.9	In Situ Data	152
6.10	Analyze and Advanced Processing	153
6.11	Conclusions	156
	References	158
7	Overview of High-Energy X-Ray Diffraction Microscopy (HEDM) for Mesoscale Material Characterization in Three-Dimensions	167
	Reeju Pokharel	
7.1	Introduction	167
7.1.1	The Mesoscale	168
7.1.2	Imaging Techniques	169
7.2	Brief Background on Scattering Physics	171
7.2.1	Scattering by an Atom	172
7.2.2	Crystallographic Planes	174
7.2.3	Diffraction by a Small Crystal	175
7.2.4	Electron Density	177
7.3	High-Energy X-Ray Diffraction Microscopy (HEDM)	178
7.3.1	Experimental Setup	178
7.3.2	Data Analysis	179
7.4	Microstructure Representation	181
7.5	Example Applications	183
7.5.1	Tracking Plastic Deformation in Polycrystalline Copper Using nf-HEDM	183
7.5.2	Combined nf- and ff-HEDM for Tracking Inter- granular Stress in Titanium Alloy	186
7.5.3	Tracking Lattice Rotation Change in Interstitial-Free (IF) Steel Using HEDM	187

7.5.4	Grain-Scale Residual Strain (Stress) Determination in Ti-7Al Using HEDM	189
7.5.5	In-Situ ff-HEDM Characterization of Stress-Induced Phase Transformation in Nickel-Titanium Shape Memory Alloys (SMA)	190
7.5.6	HEDM Application to Nuclear Fuels	191
7.5.7	Utilizing HEDM to Characterize Additively Manufactured 316L Stainless Steel	192
7.6	Conclusions and Perspectives	194
7.6.1	Establishing Processing-Structure- Property- Performance Relationships	196
	References	198
8	Bragg Coherent Diffraction Imaging Techniques at 3rd and 4th Generation Light Sources	203
	Edwin Fohtung, Dmitry Karpov and Tilo Baumbach	
8.1	Introduction	204
8.2	BCDI Methods at Light Sources	211
8.3	Big Data Challenges in BCDI	212
8.4	Conclusions	214
	References	214
9	Automatic Tuning and Control for Advanced Light Sources	217
	Alexander Scheinker	
9.1	Introduction	218
9.1.1	Beam Dynamics	220
9.1.2	RF Acceleration	222
9.1.3	Bunch Compression	223
9.1.4	RF Systems	224
9.1.5	Need for Feedback Control	226
9.1.6	Standart Proportional Integral (PI) Control for RF Cavity	227
9.2	Advanced Control and Tuning Topics	232
9.3	Introduction to Extremum Seeking Control	233
9.3.1	Physical Motivation	234
9.3.2	General ES Scheme	236
9.3.3	ES for RF Beam Loading Compensation	238
9.3.4	ES for Magnet Tuning	240
9.3.5	ES for Electron Bunch Longitudinal Phase Space Prediction	242
9.3.6	ES for Phase Space Tuning	246
9.4	Conclusions	249
	References	249
	Index	253

Contributors

Prasanna V. Balachandran Los Alamos National Laboratory, Los Alamos, NM, USA; Department of Materials Science and Engineering, Department of Mechanical and Aerospace Engineering, University of Virginia, Charlottesville, VA, USA

Tilo Baumbach Institute for Photon Science and Synchrotron Radiation, Karlsruhe Institute of Technology, Eggenstein-Leopoldshafen, Germany

Nikhilesh Chawla 4D Materials Science Center, Arizona State University, Tempe, AZ, USA

Nikolaus L. Cordes Materials Science and Technology Division, Engineered Materials Group, Los Alamos National Laboratory, Los Alamos, NM, USA

Edwin Fohtung Department of Physics, New Mexico State University, Las Cruces, NM, USA; Los Alamos National Laboratory, Los Alamos, NM, USA

James E. Gubernatis Los Alamos National Laboratory, Los Alamos, NM, USA

Kevin Henderson Materials Science and Technology Division, Engineered Materials Group, Los Alamos National Laboratory, Los Alamos, NM, USA

John Hogden Los Alamos National Laboratory, Los Alamos, NM, USA

Jacob L. Jones Department of Materials Science and Engineering, North Carolina State University, Raleigh, NC, USA

Sergei V. Kalinin Oak Ridge National Laboratory, Institute for Functional Imaging of Materials, Oak Ridge, TN, USA; Oak Ridge National Laboratory, Center for Nanophase Materials Sciences, Oak Ridge, TN, USA

Dmitry Karpov Department of Physics, New Mexico State University, Las Cruces, NM, USA; Physical-Technical Institute, National Research Tomsk Polytechnic University, Tomsk, Russia

Turab Lookman Los Alamos National Laboratory, Los Alamos, NM, USA

Artem Maksov Oak Ridge National Laboratory, Institute for Functional Imaging of Materials, Oak Ridge, TN, USA; Oak Ridge National Laboratory, Center for Nanophase Materials Sciences, Oak Ridge, TN, USA; Bredesen Center for Interdisciplinary Research, University of Tennessee, Knoxville, TN, USA

Michael McKerns The Uncertainty Quantification Foundation, Wilmington, DE, USA

Alisa R. Paterson Department of Materials Science and Engineering, North Carolina State University, Raleigh, NC, USA

Brian M. Patterson Materials Science and Technology Division, Engineered Materials Group, Los Alamos National Laboratory, Los Alamos, NM, USA

Reeju Pokharel Los Alamos National Laboratory, Los Alamos, NM, USA

Brian J. Reich Department of Statistics, North Carolina State University, Raleigh, NC, USA

Alexander Scheinker Los Alamos National Laboratory, Los Alamos, NM, USA

Ralph C. Smith Department of Mathematics, North Carolina State University, Raleigh, NC, USA

James Theiler Los Alamos National Laboratory, Los Alamos, NM, USA

Lav R. Varshney Coordinated Science Laboratory and Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, USA

Alyson G. Wilson Department of Statistics, North Carolina State University, Raleigh, NC, USA

Xianghui Xiao X-ray Photons Sciences, Argonne National Laboratory, Argonne, IL, USA

Dezhen Xue State Key Laboratory for Mechanical Behavior of Materials, Xi'an Jiaotong University, Xi'an, China

Maxim Ziatdinov Oak Ridge National Laboratory, Institute for Functional Imaging of Materials, Oak Ridge, TN, USA; Oak Ridge National Laboratory, Center for Nanophase Materials Sciences, Oak Ridge, TN, USA

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

© Springer Nature Switzerland AG 2018

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]

- 1: Let $U = SVD(X, k)$ be the initial model of X_U and let μ be the mean of the data
 - 2: **while** discovery is to continue and $X \neq \emptyset$ **do**
 - 3: Compute reconstructions $\hat{x} = UU^T(x - \mu) + \mu$ for all $x \in X$
 - 4: Compute error in reconstructions $R(x) = \|x - \hat{x}\|_2 = \|x - (UU^T(x - \mu) + \mu)\|_2$ for all $x \in X$
 - 5: Choose $x' = \operatorname{argmax}_{x \in X} R(x)$ to investigate next
 - 6: Remove this data item from the data set and add it to the model, i.e. $X = X \setminus \{x'\}$ and $X_U = X_U \cup \{x'\}$.
 - 7: Update U and μ by using the incremental SVD algorithm [53] with inputs (U, x', k) .
 - 8: **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, $X \in \mathcal{X}$ and $Y \in \mathcal{Y}$ with corresponding joint density $f_{XY}(x, y)$ and marginals $f_X(x)$ and $f_Y(y)$, the mutual information is given as

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

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 $\mathcal{M}$ be the set of artifacts known to the observer, with each artifact in this repository being $M \in \mathcal{M}$. 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):

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

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.

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

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

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

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]:

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

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 be 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-taxi 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-taxi 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](https://arxiv.org/abs/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