Materials Data Science: Current Status and Future Outlook

Surya R. Kalidindi[1] and Marc De Graef[2]

[1]School of Mechanical Engineering and School of Computational Science and Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332

[2]Department of Materials Science and Engineering, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213-3890; email: degraef@cmu.edu

Annu. Rev. Mater. Res. 2015. 45:171–93

The *Annual Review of Materials Research* is online at matsci.annualreviews.org

This article's doi:
10.1146/annurev-matsci-070214-020844

Keywords

materials database, materials data management, materials data analytics, materials e-collaboration platform, process-structure-property linkage

Abstract

The field of materials science and engineering is on the cusp of a digital data revolution. After reviewing the nature of data science and Big Data, we discuss the features of materials data that distinguish them from data in other fields. We introduce the concept of process-structure-property (PSP) linkages and illustrate how the determination of PSPs is one of the main objectives of materials data science. Then we review a selection of materials databases, as well as important aspects of materials data management, such as storage hardware, archiving strategies, and data access strategies. We introduce the emerging field of materials data analytics, which focuses on data-driven approaches to extract and curate materials knowledge from available data sets. The critical need for materials e-collaboration platforms is highlighted, and we conclude the article with a number of suggestions regarding the near-term future of the materials data science field.

1. INTRODUCTION

1.1. What Is Data Science?

Data science, initially considered to be an extension of the statistical sciences, is now clearly established as an independent discipline (1, 2). Some researchers have made bold claims on the increasingly important role that data science is expected to play in current and future scientific endeavors (3, 4). In this regard, it is important to distinguish data science, which focuses on extracting high-value information or knowledge from available data, from computational science, which addresses solution methodologies to rigorously formulated problems. As a simple example, consider a sophisticated physics-based multiscale materials modeling tool that simulates structure-property relationships. Computational science deals with the challenges involved in solving the governing field equations under specified materials constitutive laws and imposed boundary and initial conditions. Data science addresses the extraction of the embedded low-dimensional linkages between the various inputs and outputs involved in the numerical simulation. In this pursuit, data science leverages and assimilates well-established concepts and results from statistics, applied mathematics, computational science, computer science, information science, digital signal processing, and systems theory. As one might expect, data science aims to be a cross-cutting discipline that can be applied to a broad range of application domains. Indeed, data science has already enjoyed many remarkable successes in disparate application domains, including recommendation systems (5), personal informatics (6), drug discovery (7), decision systems (8), and health care (9).

At its core, data science is composed of two primary components. The first component can be broadly identified as data management (DM) and includes robust and reliable storage, aggregation, archival, retrieval, and sharing protocols. This first step is critical and necessary, as pursuing data science without easy access to reliable data would be impossible. It is important to embrace a very broad definition of data in this context. Indeed, large collections of music files, or a company's inventory, can be considered to be data. When one does experiments in materials science, data can include not just the files produced by the instruments used, but also various other pertinent details about the experiment (e.g., environmental conditions, details of how the test specimens were prepared, instrument settings, and so on). It is customary in the field of data science to refer to salient data about data as metadata. Metadata might include key information such as the file type or format, file size, and time of creation, and its central purpose is to enhance the utility of the data by providing important information on the context and content, thereby allowing these data to be discovered in appropriate searches by potential users. In this regard, metadata might even include keywords, categories, and other relevant metrics that describe the essential content of the data. Carefully designed metadata and metadata databases are crucial to ensuring and improving data longevity and usefulness.

The second task in data science centers around data analytics (DA) and is aimed at mining the embedded high-value information via noise or background filtering, data fusion, uncertainty analyses, statistical analyses, dimensionality reduction, pattern recognition, regression analyses, machine learning, and statistical learning. It is important to understand and recognize that these techniques need to be somewhat customized for different application domains, especially to be able to take advantage of the legacy knowledge already curated historically by the experts in the specific domain. Indeed, much of the practical utility and impact of data science arises from this second task.

Increasingly, in recent years, there has been a strong recognition for the critical need of a third component of data science: e-collaboration science (e-CS). This component deals with the online tools designed specifically to seed and nurture cross-disciplinary research collaborations between

application domain experts and data scientists. Through such collaborations, one aims to achieve (*a*) a synergistic integration of specific disciplinary workflows with emerging data science tools and (*b*) the concomitant enhancement of the overall research productivity. e-Collaborations also have the potential to dramatically accelerate the processes of multidisciplinary research team formation with a desired and otherwise unachievable combination of expertise. As a simple illustration of the importance of e-collaboration tools, consider the impact of email on the research productivity of individual researchers in all domains of science and technology. One can make similar observations regarding the prevalent use of Dropbox (10) and GitHub (11) by researchers worldwide. HUBzero (12) and nanoHUB (13) serve as early examples of successful new online scientific communities that share data and tools. These examples attest to the positive impact that e-collaboration platforms can have in enhancing and accelerating the cross-disciplinary collaborations needed to integrate emerging data science protocols into existing disciplinary workflows.

Given that a significant fraction of academic research is funded by federal or state tax dollars, community access to data is also becoming increasingly important. Data that were originally acquired as part of a research project may well be transitioned to become part of an educational effort. In 2005, the National Science Board of the National Science Foundation published a report (14) on "Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century." Long-term data preservation and mechanisms for making data available to diverse groups (e.g., researchers, students, and data scientists) form the focus of this report. In this regard, it is reiterated that simply providing access to data would not be adequate; to facilitate optimal use of the available data, the user often needs critical metadata (including, for example, when, where, and how the data were created; what machines were used under what conditions; and what data processing has been conducted prior to releasing the data). The e-collaboration platforms described above can enhance such forms of data sharing by providing the relevant context, discussions, and annotations of the data in ways that add tremendous value to the end user.

1.2. What Is Big Data?

The phrase Big Data has recently spread across all fields of science and engineering and has become somewhat of a catchall name covering many different situations. Big Data can refer to very large, structured data sets that require dedicated and known algorithms for the extraction of typically well-defined parameters, for instance, particle tracking in a Navier-Stokes fluid turbulence simulation (15), elementary particle trajectories from collider data (16), and stellar spectral data from the Hubble Legacy Archive (17). Big Data can also refer to unstructured data collections for which sets of data analysis routines are used to parse potential connections and trends between seemingly unrelated data items, to predict future behavior of returning customers, and so on. Before we take a closer look at data science in the materials field (Section 2), a brief survey of Big Data in other fields can provide some useful insights into and context for the role of Big Data in the materials community. In each case, we highlight DM, DA, and e-CS efforts as potentially relevant examples.

Some of the largest data sets can be found in the elementary particle physics community, clustered at a number of experimental locations, such as the Conseil Européen pour la Recherche Nucléaire (CERN) in Geneva (18), the Fermi National Accelerator Laboratory (Fermilab) near Chicago (19), and the Relativistic Heavy Ion Collider at the Brookhaven National Laboratory on Long Island (20). CERN's Large Hadron Collider (LHC) produces a sustained data stream of approximately 300 MB/s (21), amounting to approximately 15 PB per year. This data stream represents only a fraction of the total data generated by all the collider instruments; decisions on which data to keep are generally made in real time (DA). Each of these labs provides vast digital

storage space (DM): The total reported (archival) storage capacity amounts to more than 200 PB at CERN and more than 500 PB at Fermilab. Efficient worldwide access to these data has required the creation of dedicated file storage and retrieval systems (DM and e-CS), such as dCache (22), jointly created and maintained by the Deutsches Elektronen-Synchrotron (DESY, Hamburg, Germany) (23) and Fermilab, and the Worldwide LHC Computing Grid (24), the world's largest computing grid. Recently, CERN started to make its data available to the public (e-CS) via its Open Data Portal (16), which provides access to both data and data analysis algorithms (DA) as well as to educational resources.

In the biological and medical communities, the National Center for Biotechnology Information (25) provides a central portal to an impressive array of databases (DM) covering all length scales from nucleotides, genes, and proteins to cells and the genomes of more than 1,000 organisms. Genetic sequences provide an example of Big Data that is simultaneously structured (specific well-posed queries regarding base-pair sequences can be executed using well-known algorithms) and unstructured (it may not always be clear a priori which questions need to be asked or even how to ask them). Data-mining approaches (DA) can then provide novel insights and discoveries (26). The (micro)biological field has many aspects in common with the materials field, in particular the presence of multiple length scales. Once a protein or enzyme structure has been solved to within a given atomic positional accuracy, then that molecule can be regarded as a known building block that can partake in chemical processes; this is equivalent to knowledge of a material's crystal structure. There is little uncertainty associated with this knowledge. Only when these building blocks are combined into more complex configurations (e.g., a cell in biology or a material's microstructure) do variability and uncertainty become important; each transition to a larger length scale brings with it an increase in variability, which eventually gives rise to a rich data landscape that must be analyzed by both structured and general data-mining approaches (DA).

The US National Oceanic and Atmospheric Administration (NOAA) (27) manages several data centers that are focused on oceanographic, climatic, and geophysical data. NOAA collects (DM), processes (DA), and distributes (e-CS) satellite data and a variety of derived products, as well as the scientific algorithms needed to extract interpretable information from the vast numbers of data downloaded from near earth orbits each day. Data can be accessed in real time (e-CS) by using the NOAA View Data Exploration Tool (28). For more detailed data analysis, NOAA provides the National Operational Model Archive and Distribution System (29), which gives the scientist user direct access to historical archived data related to global weather patterns, climate, geophysics, and so on. Data are made available via dedicated data servers (30) that employ the OPeNDAP framework (31) to make local data accessible to remote locations (DM and e-CS). The framework also allows for limited remote data analysis (DA) of data sets on the server.

NASA's Earth Observing System Data and Information System (32) is a large, distributed data center program for managing and archiving earth-related data from satellites, aircraft, and field measurements (DM). Each of the 12 data centers has specific responsibilities for production, archiving, and distribution (e-CS) of earth science data products regarding sea ice, atmospheric convection, vegetation indices, clouds, soil moisture, etc. Each of the member data centers hosts multiple open-access data sets; for instance, the National Snow and Ice Data Center (33) makes more than 370 data sets available to the public.

These examples provide only a snapshot of the bewildering number of open-access scientific databases available worldwide and make it abundantly clear that the materials community still has a long way to go in terms of DM, DA, and e-CS. Materials data have not yet reached the Big Data designation, although there are signs (34) that we are on the verge of a materials data revolution. Although the materials data sets of today exhibit tremendous variety, they do not yet exhibit either a sufficiently large volume or high velocity (see Section 2.2) to be designated as Big

Data. However, recent advances in both computational and experimental capabilities are about to usher materials data into the emerging Big Data era. The Integrated Computational Materials Engineering initiative (35) and the Materials Genome Initiative (36) are steps in the right direction, but there is a clear need for a practical and robust approach to large-scale materials data storage and dissemination, which will require a substantial financial investment in data infrastructure and dedicated personnel.

2. MATERIALS DATA SCIENCE

2.1. Distinguishing Features of Materials Data

Before we take a deep dive into the world of materials data science, it will be instructive to understand and appreciate the salient features of materials data that demand a certain degree of customization of the currently available data science tools. Virtually all considerations in materials science and engineering revolve around three main quantities: (*a*) synthesis and/or processing routes, (*b*) hierarchical materials internal structure, and (*c*) properties or performance characteristics. Of these three quantities, it is relatively easy to represent the important details of quantities *a* and *c* as digital data. In contrast, most of the information related to quantity *b* exists in the form of images produced by various materials characterization tools.

The conversion of images to digital data is a fairly complicated task. Although certain other application domains (e.g., astronomy, biomedicine, security) also capture their raw data as images, the type and nature of the image analytics needed in materials science are quite different from those of these other application domains. **Figure 1** shows a collage of materials structure images taken at different length scales through different techniques, illustrating the broad variety of image data and hinting at the inherent complexities in subsequent analyses.

Several aspects of the images used by the materials community distinguish the field from other image-rich fields:

1. For materials images acquired at a selected length scale, it is typically not known a priori which of the image features are responsible for a material's performance characteristics. This situation is very different from that of other fields, such as astronomy, in which the possible image features (stars, nebulae, galaxies, etc.) are known beforehand and the role of each image feature is often well understood. Materials specialists often make subjective (intuitive) guesses on what they consider to be the most important local features in a given image. Therefore, the same set of images can lead to different interpretations by different experts.
2. The diversity of image features is matched by the diversity of techniques and protocols employed by materials researchers. Materials science is a relatively young field of study that is fast evolving; the past decade has seen an explosion of characterization techniques used to quantify a material's internal structure at different length scales (37). There do not yet exist broadly adopted standards either in the protocols employed in characterizing the material or in reporting the results produced in such studies. Consequently, the current practices lead to inconsistent documentation and reporting of both data and metadata, as well as to issues in the long-term preservation of the data.
3. Each of the techniques employed in studying a hierarchical materials structure, for instance, 2D surface scans or 3D tomographic reconstructions of the sample interior, is usually aimed at only a specific structural feature and thus produces incomplete information on a material's structure. Furthermore, images of the same surface obtained using optical or electron microscopes, X-rays, and Raman spectroscopes often reveal very different information about

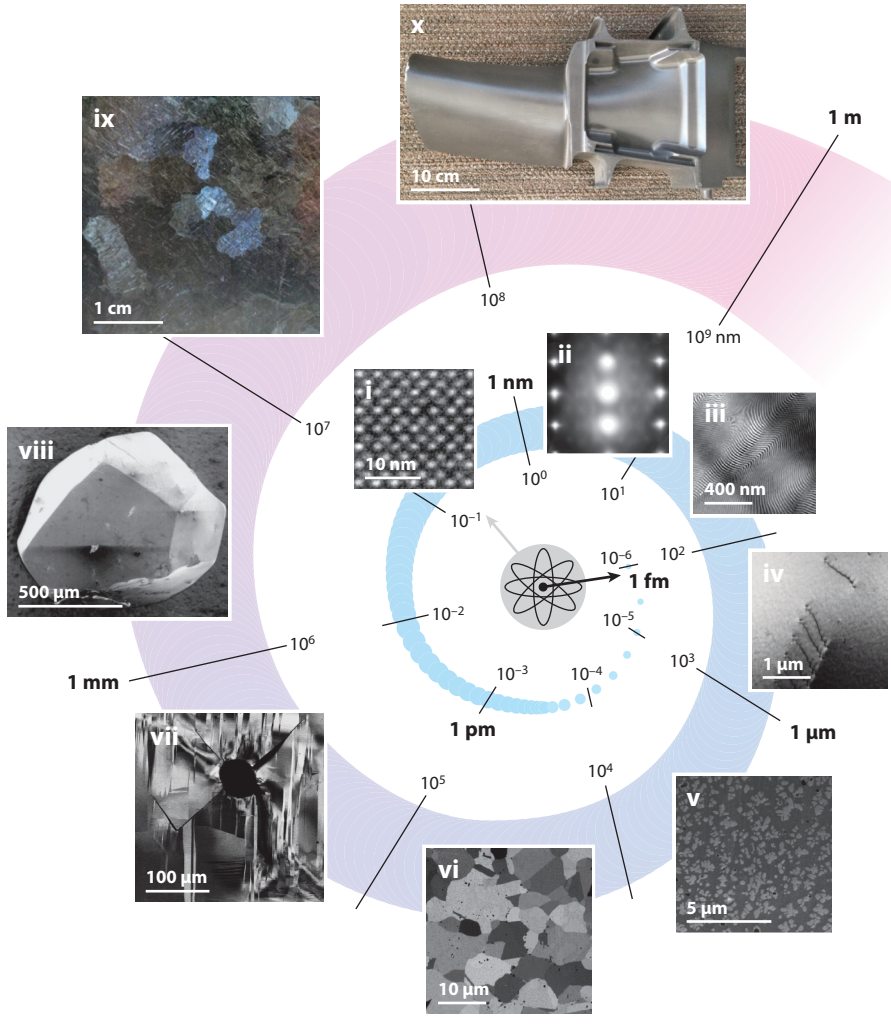

Figure 1

Collage of materials images spanning a range of length scales from atomistic to macroscopic, starting at the center and moving clockwise in increasing microstructural length scale: (*i*) atomic-resolution image of $BaTiO_3$ along [001]; (*ii*) diffuse electron scattering in [112] zone axis orientation, indicating the presence of short-range order in Cu–15% Al; (*iii*) lamellar poly(styrene-*b*-isoprene) block copolymer microstructure (courtesy of M. Bockstaller); (*iv*) dislocation array in Cu–15% Al; (*v*) dendritic γ' precipitates in a Rene-88DT superalloy; (*vi*) polycrystalline grain microstructure in an IN100 superalloy (courtesy of M. Uchic); (*vii*) type II twins and magnetic contrast in a multiferroic Ni_2 MnGa alloy; (*viii*) a centimeter-size single extracted grain of Ni_2 MnGa; (*ix*) centimeter-sized grains in architectural titanium; and (*x*) a single-crystal superalloy turbine blade (courtesy of T. Pollock).

the same underlying materials structure. Consequently, one needs to judiciously fuse these different pieces of information, a process that is often impeded by differences in the accuracy and resolution limits of the different tools, by the inability to scan the exact same regions by using asynchronous data acquisition, and by the fact that the geometrical characteristics of 2D and 3D sampling grids can be very different for each observation technique.

Although major advances are being made to address the impediments listed above, for the time being one has to factor the above considerations in developing suitable data science tools for the materials community. In the following subsections, we review recent developments in the emerging field of materials data science.

2.2. Main Objectives of Materials Data Science

Materials with superior combinations of properties play a critical enabling role in the development of advanced technologies. Therefore, a grand challenge for the materials community lies in the acceleration of the rate at which new materials can be designed, manufactured, and deployed. These strategic objectives have been clearly articulated in numerous high-profile national documents (e.g., References 35 and 36). These documents express a consensus opinion that one of the key ingredients to achieving the desired acceleration lies in the development and broad utilization of suitable data science tools (encompassing DM, DA, and e-collaboration platforms) by the materials community.

The central objective of materials data science is to facilitate an efficient mining of large materials data sets, resulting in the extraction and identification of high-value materials knowledge. The core materials knowledge of high value—not only to the materials community but also to the product design and manufacturing communities—can be represented as process-structure-property (PSP) linkages. Kalidindi (38) has proposed a systematic hierarchical set of objectives for materials data science, reproduced here as **Figure 2**. At the first level of data analysis, the main objectives are (*a*) to extract trends on the evolution of selected features during a given manufacturing route and (*b*) to study how these details affect effective properties/performance characteristics of interest for the material. Although such correlations do not provide comprehensive information on the PSP linkages of interest, they usually provide valuable guidelines for the materials development community. One might characterize these higher-value descriptions of PSP linkages as information (see **Figure 2**).

At the next level, one aims to extract more rigorous, reliable, and complete PSP linkages from the available data; this information can be characterized as materials knowledge. The

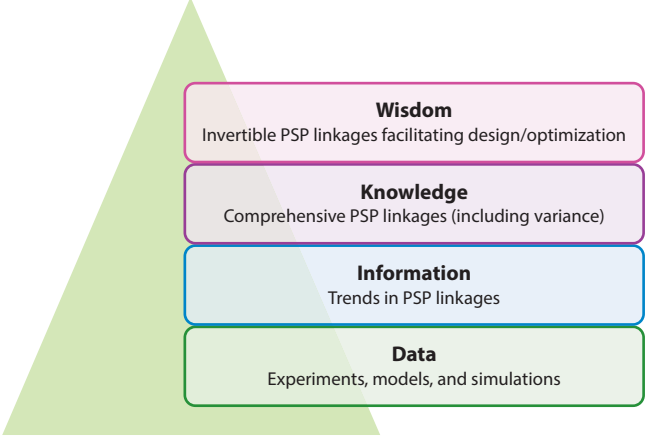

Figure 2

Hierarchical scheme for the main objectives of materials data science. Abbreviation: PSP linkages, process-structure-property linkages.

comprehensive PSP linkages available at this stage should be sufficiently reliable to allow for a rigorous quantification of the inherent uncertainty; the available PSP linkages may be successfully employed to simulate manufacturing processes of interest and to predict the performance of the final product. However, the main focus in the data transformations at the knowledge level continues to be in the forward direction (process → structure → properties). At the final stage of data transformation, effort should be focused on establishing invertible PSP linkages that allow for customized process and materials design for targeted applications (i.e., address inverse problems). This highest level of understanding of PSP linkages can be characterized as wisdom.

Deeper reflection on the objectives stated above, combined with the need to consider complex materials physics at multiple hierarchical length scales, makes one realize the critical need for collecting, storing, and distributing large quantities of materials data. The number of data needed to accomplish these ambitious goals is large enough to justify the application of the Big Data infrastructure and toolsets. However, at the current time, the materials data science is very much at a nascent stage, and a number of additional developments are needed to propel the field forward. Generally speaking, Big Data is qualified by the five Vs—volume, velocity, variety, veracity, and value—although sometimes variability and visualization are also included (39). Let us briefly review where we currently stand in materials data science with respect to the five basic concepts, beginning with the objective Vs: volume, velocity, and variety.

2.2.1. Volume. The quantity (volume) of experimental data generated by a single materials research group is typically relatively small relative to many other fields. Electron microscopes, for instance, can easily generate tens of gigabytes of data in a day of operation; in high-energy diffraction microscopy experiments at a synchrotron beamline (40), far-field observations tend to generate several hundreds of gigabytes per experiment, whereas near-field observations are far more data intensive, at several terabytes per experiment. Advances in detector design could push this volume significantly higher in the near future. X-ray tomography data acquisition with high-speed, high-resolution cameras (2,000 × 1,500 pixel 16-bit images at a frame rate of 70 Hz, resulting in approximately 25 GB of data per minute) easily generates multiterabyte data sets, and a typical multiday beamline run routinely produces 20 TB of data. Interestingly, in contrast with collider operations at CERN and Fermilab, which provide centralized high-speed data storage and archiving facilities, user facilities such as the Advanced Photon Source (41) typically provide only limited centralized storage facilities; instead, data are usually stored on user-provided media and are transported by car between the beamline and the user's home institution. The number of such large data sets available from various research groups is currently relatively small, and transferring data from one location to another is not straightforward because of a lack of appropriate infrastructure. Long-term data protection is often achieved by simply storing the drives on a shelf instead of mounting them in RAID (redundant array of inexpensive/independent disks)-like configurations, because the probability of disk failure in a RAID system is too high when large numbers of disks are used (see Section 2.4.1).

If one were to somehow aggregate, in one place, all available experimental materials data for a single given materials system, the total data volume would likely be relatively small. However, the number of simulation data that could be generated could be very significant. The current practices in the materials field typically do not include uploading and sharing of simulation data sets. Indeed, it would actually be a good idea for materials simulation experts to openly share their computational data sets in ways that would allow data scientists to mine these data sets for PSP linkages. A concerted effort in this direction may very well provide new opportunities for the application of Big Data concepts to materials data sets but will require an infrastructure to facilitate data uploads and downloads.

2.2.2. Velocity. Velocity refers to the speed with which data are generated, stored, and processed. With the possible exception of high-speed data collection at beamlines mentioned in the previous subsection, the rate at which most experimental materials data sets are currently being acquired typically does not qualify this activity as Big Data. A modern scanning electron microscope running an electron backscatter diffraction experiment may be capable of acquiring and indexing more than 1,000 patterns per second, but this operation mode does not usually involve storage of all these patterns on disk; the data generation rate drops significantly when all the data are stored. For computational data sets, the velocity is in principle limited only by the number of available processors and by how fast data can be written to disk. Commercial organizations and online search engines are generally much further advanced in the area of real-time data processing and thus operate at a significantly higher velocity than any current materials data efforts.

2.2.3. Variety. As mentioned above and illustrated in **Figure 1**, materials data sets tend to exhibit tremendous variety. By this metric, we can likely qualify the materials data science challenge as a Big Data challenge. It is important to remind ourselves that the tremendous variety in materials data sets arises from the need to employ different techniques to obtain different pieces of information about the hierarchical structure of a material. Merging or fusing these varied data into coherent data sets across multiple length scales will likely pose a significant challenge.

2.2.4. Veracity and value. The remaining two Vs of Big Data are usually considered to be the subjective Vs. Materials data sets exhibit substantial uncertainty because of the use of nonstandard and continuously evolving protocols. However, the typical practice in the field does not require data generators to quantify and report the uncertainty associated with their data sets. Similarly, we do not currently have widely adopted protocols to evaluate the value of the materials data sets. As McNulty (39) states: "The Value [of Big Data] lies in rigorous analysis of accurate data, and the information and insights this provides."

A December 2014 search of the Compendex and Inspec databases for the period 2000–2014 turned up only two English-language journal articles that had the terms "materials" and "Big Data" in their title (34, 42); a search using only "Big Data" as the keyword produces 879 matches for the same time period, the majority of them dealing with data analysis, cloud computing, data mining, and data handling as the main focus areas. These search results, along with the discussion in the preceding subsections, confirm that we are in the nascent stage of materials data science. A number of ongoing efforts are aimed at addressing the challenges described above; we highlight some of these efforts next.

2.3. Materials Databases

Several recent vision-setting documents for the materials community (35, 36) emphasize the critical need for open-access materials databases that capture, curate, and archive the critical information needed to facilitate accelerated development of new/improved materials and their deployment in emerging technologies. Although these reports do not provide a clear road map, they do emphasize the anticipated benefits from the availability of these new databases. One of the immediate benefits is that these databases have the potential to minimize the unintended repetition of effort from multiple groups of researchers, possibly over multiple generations. As noted above, core materials knowledge that needs to be aggregated, curated, validated, and archived is often expressed in the form of PSP linkages, i.e., high-value information that needs to be communicated to the manufacturing and product design value chain. The main challenge in designing and building the desired databases stems from the fact that most materials of interest exhibit rich internal structures that span multiple length scales. Therefore, to rigorously catalog a materials system, one needs to

adopt an extensible structure quantification framework that is broadly applicable to all materials systems of interest; in other words, the framework should be able to quantify the rich heterogeneity at a multitude of length scales in different materials systems. Because of the tremendous difficulty of including the hierarchical materials structure in knowledge databases, most efforts at building such databases either have focused on a single length scale by restricting attention to a limited number of structure measures or have completely ignored the structure information.

When materials databases are focused on PSP linkages at a single length scale, they need only to address a limited amount of variety in the data sets. Consequently, it is possible to design and build valuable databases that focus on physical properties of specific chemical elements or compounds that do not depend on microstructure. One approach for the creation of materials databases is described in detail in Reference 43. Examples of such databases include Citrine Informatics (44) for physical properties of nearly 30,000 chemical compounds, the Clean Energy Project database (45) for electronic properties of organic compounds used in plastic solar cells, The Materials Project (46) at MIT/LBNL and the Automatic-FLOW for Materials Discovery project (47) at Duke University for large-scale data from electronic structure computations of compounds, CALPHAD (48) for computationally derived thermodynamic properties of various thermodynamic phases, and the Open Quantum Materials Database (49) at Northwestern University for density functional theory–calculated thermodynamic and structural properties of nearly 300,000 compounds. The reader is also referred to the large number of reference data sets and databases maintained by the National Institute of Standards and Technology (NIST) (50) for various aspects of materials structures and properties; these data sets and databases are usually focused on specific aspects at selected length or structure scales. NIST also maintains a repository for interatomic potentials (51).

The Center for Hierarchical Materials Design (CHiMaD) at Northwestern University brings together state-of-the-art computational methods, curated materials databases, and novel integrated experimental methods under NIST sponsorship and in collaboration with the University of Chicago and Argonne National Laboratory. The center is developing publicly accessible databases of CALPHAD protodata, thermodynamic and structural properties of polymers, and materials Big Data repositories. The CALPHAD protodata effort focuses on archiving, for the first time, the thermodynamic properties of elements, compounds, and alloys that underlie phase diagrams. The NanoMine database, created using the NIST Materials Data Creator interface, will contain data on the properties of polymer-nanoparticle mixtures that will be used in polymer-matrix composites. The Materials Data Facility will enable researchers to upload and share the large data sets generated by 3D, 4D, and spectroscopic materials characterization approaches as well as density functional theory calculations. CHiMaD will also create the data-mining tools to extract knowledge from the databases.

Several materials databases focus on macroscale physical properties of various commercially available materials without a rigorous description of the associated materials' microstructures; they serve a valuable purpose for the product design community to facilitate deployment of advanced materials available today in emerging technologies. However, more often than not, the materials available today will likely need some additional tweaking in their manufacturing process so that they can be optimized for a selected application. To undertake such a tweaking in a cost-effective manner, it is essential to document and track the evolution of a material's hierarchical structure throughout the various unit manufacturing processes employed. Nevertheless, the property-centered databases described above serve an important role in materials development efforts. Examples of such databases include MatWeb (52), with properties for more than 105,000 compounds, and Granta CES Selector (53). The NIMS Materials Database (54) is perhaps unique in this regard because it actually includes the processing history employed in the manufacture of the materials; this processing history implicitly includes information on a material's microstructure.

In building and deploying material databases for macroscale properties of hierarchical materials, it would be beneficial to learn from the successes and failures of prior efforts. In a recent summary, Freiman et al. (55) discuss the lessons learned from building and maintaining the National Materials Property Data Network (56), which was operated commercially until 1995. Although this network contained numerous databases on many different materials, it proved too costly to maintain because of (*a*) a lack of a single common standard for information on the broad range of materials included in the database, (*b*) enhancements/modifications to testing procedures used for evaluating properties of interest that sometimes invalidated previous measurements, and (*c*) the placement of onus on a single entity as opposed to joint development by multiple entities that possessed the broad range of requisite sophisticated skill sets. Notwithstanding these shortcomings, a major deficiency of these historical databases aimed at capturing the macroscale properties of hierarchical materials is that they do not include the important details of a material's microstructure, which includes all important features of the hierarchical structure of the material. The microstructure-property connections are many-to-one, and composition alone is inadequate to correlate with the many macroscale physical properties of most advanced materials systems. Indeed, the spatial placement of local states in the internal structure (i.e., morphological attributes) at various constituent hierarchical length scales plays an influential role in determining the overall properties of the material (57–59).

It is becoming increasingly clear that materials databases will have to include microstructure information (38, 60); only the microstructure space can serve as the common higher-dimensional space in which the much-needed knowledge of PSP correlations can be reliably stored and communicated. Any attempt to represent this core knowledge in lower-dimensional spaces (e.g., processing-property correlations or composition-property correlations) should be expected to potentially result in substantial loss of fidelity. The recognition of the important role of microstructure has led to the design and deployment of novel microstructure-centered materials databases and tools. Examples of such resources include DREAM.3D (61), the Materials Atlas (62), and the Computational Materials Data Network (63). These newer-generation microstructure-centered platforms offer exciting avenues for accelerated materials development efforts. DREAM.3D, in particular, keeps track of the workflow, i.e., the individual steps that a user carried out in the process of analyzing a data set, in addition to providing a standardized file format for data storage (64).

Professional materials societies have taken on an active role in the area of materials databases and DM. For instance, in 2002, the Material Data Management Consortium (65), a group of materials-oriented organizations in the aerospace, defense, and energy sectors, was initiated by ASM International, the Life Prediction Branch of NASA Glenn Research Center, and Granta Design. The consortium provides each member with tools for data mining and data analysis as well as ensuring integrity and traceability. The Minerals, Metals and Materials Society (TMS) (66) actively supports the Materials Atlas (62), which provides a repository for 3D experimental and simulation data, as well as software tools and tutorials for using the data. TMS also houses the Materials Cyberinfrastructure Portal (67), which serves as an online access point for critical tools and resources that can accelerate materials innovation processes.

2.4. Materials Data Management

DM refers to all practical aspects related to the handling of data—e.g., hardware, file systems and formats, data redundancy, archiving, unique file naming, and data transmission—as well as the issue of who stores and curates the data. Each of these aspects poses unique problems when data sets become very large or when very large numbers of files are involved. Providing remote access

to very large data sets also raises issues because download times can become prohibitively long, even on high-speed Internet connections.

The question of who becomes the curator of materials data is an important one and is as yet unanswered for the materials community. One could make a strong case for the involvement of government laboratories, e.g., the US Department of Commerce's NIST. The instruments capable of generating large quantities of experimental data are typically located at national laboratories, which suggests that materials data centers should be established at such locations. Furthermore, there is a need for a community-wide data-archiving strategy; national laboratories appear to be the logical entities for creating and maintaining such a strategy.

2.4.1. Data storage hardware and file systems. The practical aspects of Big Data storage are very different from those faced by individual academic laboratories; in a typical materials characterization laboratory, a judiciously sized RAID system (68) may well be sufficient to provide storage capacity for day-to-day and archival operations. In a RAID system, redundancy is achieved by means of parity-based error correction methods, which means that the data are spread over multiple disks; when a single disk fails, the missing data can be rebuilt from the information on the other disks. For very large data sets, however, one quickly reaches a point at which the failure probability of individual drives leads to the high likelihood that a second drive will fail while the system is running the lengthy rebuild to recover from the first failure.

Hard disk failures are not as uncommon as one might think. Manufacturers publish data on drive reliability in two different forms: (*a*) mean time to failure (MTTF), typically approximately one million hours for modern drives, and (*b*) annualized failure rate (AFR) (the percentage of drives expected to fail per year). Research into disk failure data (69) shows that drive failures are significantly more common than the MTTF and AFR numbers would lead one to believe. Regarding RAID systems, the analysis indicates that failure of a second disk during a rebuild from a previous failure is nearly four times more likely than an isolated failure. For this reason, large data centers opt to duplicate all data three times to provide redundancy, instead of working with potentially unreliable RAID systems. Google File System, which is a highly distributed file system on relatively cheap hardware and which can manage hundreds of terabytes of storage across thousands of disks on thousands of clients, is one highly visible example of this data redundancy strategy (70). Other relevant file systems (71) include ZFS (zettabyte file system), developed at Sun Microsystems (72), and Hadoop (73). ZFS stores files in so-called zpools, which are not constrained to specific hardware devices and can grow to very large size by adding individual devices (similar to adding RAM to a PC). Hadoop is an Apache open-source architecture for storage, transformation, and analysis of very large data sets. Hadoop consists of the Hadoop Distributed File System along with a programming model (MapReduce) for parallel computations over large data sets.

2.4.2. Data-archiving strategies and file formats. Although there is no real technical limit to the number of files in a single directory, in practice most operating systems get bogged down quickly when more than a few thousand files are present in a single folder. This scenario can easily arise in typical materials characterization experiments when data (e.g., electron backscatter diffraction patterns) are stored in individual, small image files (74). Working with nested subfolders, each with a relatively small number of files, eventually becomes problematic as well, due to file name length limitations; when the file name includes the entire directory path, excessive nesting of subfolders can easily exceed the file system limit. The solution to this problem is to store data in large files, which have an internal structure that resembles that of an entire file system but that does not suffer from file system limitations.

To be useful for large and varied data sets, a file format must (*a*) accommodate metadata, (*b*) accept arbitrary multidimensional data structures, (*c*) allow for a hierarchical data structure,

(*d*) have read and write routines that are relatively easy to use, and (*e*) (ideally) be open source (75). The file formats that satisfy one or more of these requirements are HDF (Hierarchical Data Format) (76), netCDF (Network Common Data Form) (for storage of array-oriented scientific data) (77), PDB (Protein Data Bank) (used predominantly in the biological community) (78), FITS (Flexible Image Transport System) (used predominantly in astronomy) (79), and DICOM (Digital Imaging and Communications in Medicine) (80). Among these formats, HDF stands out because it is widely supported on desktop and high-performance computing platforms, is accessible from most programming languages, is scalable (with no size limit) and searchable (allows for random access), is open source, is well documented, and is continuously updated by the HDF group (76). HDF is already used in a number of materials-related environments, notably DREAM.3D (61, 64), and there is an ongoing dialogue with materials characterization instrument vendors to incorporate HDF-formatted files as an output option into a variety of data acquisition programs.

The US Library of Congress, which houses several petabytes of digital data, also runs the National Digital Information Infrastructure and Preservation Program (81), which maintains a series of tools for digital data preservation, archiving, and annotation of a variety of data types, from books and other textual material to audio and video data. The Community Owned Digital Preservation Tool Registry (82) provides links to nearly 400 different tools related to backup, DM, digital repositories, file formats, metadata, validation, version control, and many other objectives, and the materials community would likely benefit from the already available archival tools.

Once computational or experimental data sets are placed in an appropriate file format, including all relevant metadata, the next step is to place the file in a repository and provide it with, ideally, a unique and persistent identifier by which it can be accessed remotely. There is already a worldwide system in place for the creation of such identifiers: the Digital Object Identifier (DOI, ISO 26324) (83), managed by the International DOI Foundation (IDF) (84). This system is widely used in the publication and entertainment world, and more than 100 million unique identifiers have been assigned through this system. The IDF coordinates Registration Agencies that provide the actual DOI services and registrations. The DOI of a document remains fixed over the lifetime of the document, whereas the document's location may change; the metadata associated with the DOI contain, among other items, the current URL for the document. Under the hood, the DOI system is built on top of the Handle System (85), developed by the Corporation for National Research Initiatives (86), which provides infrastructure for unique and persistent identifiers of digital objects.

A coherent and community-wide approach to materials data science should, at the very least, include a materials-oriented Registration Agency that assigns DOIs to data sets and maintains the DOI metadata. In combination with large-scale data repositories, such a materials registration agency could provide and manage a data-archiving strategy for the materials community. This agency could, for instance, incorporate all the metadata for all materials data sets into the DOI metadata while the actual data sets would remain on different servers. This setup would make searching for particular data types easy because all the metadata would be consolidated by a single agency.

2.4.3. Data access strategies. Collecting data files in a dedicated repository with unique and persistent document identifiers will be useful only if those files can then be accessed remotely whenever and wherever they are needed. Among the many available network file transfer protocols, GridFTP (87), originally created through a collaboration between the University of Chicago and the Argonne National Laboratory, is perhaps the most commonly used approach for large-volume data transfer between sites. Based on the standard file transfer protocol (FTP), GridFTP provides reliability and security to the server-client connection, including authentication, encryption, automatic fault recovery for unattended file transfers, and scripting capabilities. The

standard protocol is maintained by the Internet Engineering Task Force (88), and Globus Online (89) provides a popular open-source toolkit for building data grids.

Although individual data files can be very large, the user often does not need the entire file but needs only a small subset of the data. Selective access to data subsets requires careful planning of the overall data layout within the file, which can be accomplished by means of multiresolution data-gridding approaches (90). An interesting example of such an approach can be found in the fluid dynamics community and consists of a 27-TB data set of 1,024 time samples on a $1{,}024^3$ spatial grid (91). The data set is publicly available, and a variety of grid-based functions, such as differentiation and interpolation, can be executed rapidly from remote user locations via dedicated web-based data-processing functions; no portion of the data is ever transferred to the remote location. This rapid availability is accomplished by the use of a special data-ordering method that guarantees that the data needed for numerical differentiation (e.g., a 3D neighborhood of voxels) are nearly always located close together in the data file so that the cache miss rate is minimized. The data ordering is known as Z-ordering or Morton ordering, based on the Peano-Hilbert space-filling curve, and indexes the 3D data by interleaving the bit representations of the individual coordinate indices into a single index; if two points are close together in 3D, then they should also be close together in the data file. Such a data-ordering approach could also be useful for 3D and 4D materials data in cases in which one needs localized microstructural information and requires the dedicated implementation of both the data format and the web-based interface for accessing the data. This implementation, in turn, would require close collaborations with the computer science community.

2.5. Materials Data Analytics

As described in **Figure 2**, the central goal of materials DA is to extract high-value information from available compilations of materials data sets. This high-value information is generally expressed in the form of PSP linkages to facilitate easy insertion into the product design and manufacturing endeavors. The central challenge in arriving at these high-value PSP linkages stems from the fact that a material's internal structure spans a multitude of length scales and potentially timescales and exhibits rich details (including a variety of disorders and defects; see **Figure 1**) that essentially control a material's properties of interest. Therefore, any efforts to extract PSP linkages have to start with a rigorous quantification of the hierarchical microstructure. Herein lies the main challenge. If one were to employ highly simplified measures of the internal structure, then the extracted PSP linkages likely would not be of sufficient accuracy and reliability for guiding advanced materials development. In contrast, if one were to include a very large number of structure measures by using a comprehensive framework that accounted rigorously for the many details of a material's internal structure, the extracted PSP linkages would likely be unwieldy and impractical for use in development efforts, which would essentially demand inverse solutions. The injection of data-driven methods might be most beneficial in precisely this area. In data-driven approaches, one employs algorithms that objectively identify the salient features in the assembled data set. One of the important consequences of the data-driven approaches is that the assessment of the features is likely to change as more data become available. In other words, the definitions of salient materials structures in data-driven approaches are dynamic, at least until a sufficiently large data set that uniformly samples the entire domain of interest is assembled (the reader is reminded that the complete space of all theoretically possible materials structures is an unimaginably large space).

The conventional approaches used in the materials development field have treated a material's microstructure largely as images and have resorted to an ad hoc or intuitive selection of certain salient image features as the measures of the material's microstructure. For example,

microstructures in structural metals are typically quantified by the overall chemical composition, by the volume fractions of the different phases, and by the average grain (crystal) size. Furthermore, the conventional protocols utilized in this field have largely required a substantial amount of manual input [e.g., placing lines on a micrograph to estimate the average grain size (92, 93)]. Only in recent years have these protocols been modified to include the conversion of microstructure images to digital signals (mostly grayscale images from optical and scanning electron microscopes) that allow for application of modern image analysis tools (94). One of the main consequences of the ad hoc approaches utilized thus far is that only very simple quantitative correlations between a material's structure, its associated properties, and their evolution during readily available manufacturing unit processes have been established and utilized successfully. For example, the most commonly used structure-property linkages are those established by the Hall-Petch rules (95, 96), which suggest that the macroscale yield strength of a metal is inversely proportional to the square root of the average grain size. Likewise, at the lower length scale, the slip resistance (defined in a region smaller than the individual crystals in a polycrystal) is directly proportional to the square root of the local averaged dislocation density. A number of simple laws capture the influence of the overall chemical composition on the yield strength of the alloy [e.g., solid-solution strengthening (97)] and the effect of precipitate type and size on the yield strength of the material [e.g., precipitation hardening (98)]. These correlations employ only the simplest microstructure measures, i.e., averaged values. To extract and utilize quantitative correlations between a material's microstructure and defect-sensitive properties such as fatigue strength or toughness, it is imperative to develop new protocols that employ more advanced measures of the material's microstructure.

The emerging interdisciplinary field of materials informatics, which is still very much in its infancy, focuses mainly on data-driven approaches designed to extract and curate materials knowledge from available materials data sets. The customization and adoption of modern information technology tools, such as image-based search engines, data mining, machine learning, and crowdsourcing, can potentially identify completely new avenues for successfully addressing some of the most difficult challenges described above. Much of the initial focus in this emerging field has been on materials discovery through combinatorial chemistry and variations of crystal structures at a single length/structure scale (99–104).

More recently, the focus in materials DA has shifted to rigorous analyses of microstructure images. In almost all experimentally measured microstructure images, one of the first steps in the analysis involves some form of segmentation. Segmentation, defined as partitioning the observation space into disjoint regions or classes, is a very broad research area, and there are almost as many segmentation methods as there are practitioners. It would be unrealistic to expect to find a segmentation algorithm that would work perfectly on all possible microstructures. Although the human mind is very good at segmenting complex scenery, trying to reproduce this on a computer is "fearsomely complex," in the words of Serra (105). Nevertheless, there have been significant recent advances, in particular in the area of segmentation of grain boundary networks (106) and Bayesian segmentation methods for multiphase microstructures (107); a general-purpose open-source segmentation tool, the EM/MPM Workbench, is now available for microstructure images for which the intensity histogram can be approximated as a superposition of Gaussians (108, 109).

In many situations, the segmentation step is typically followed by various forms of feature identification. This research field is, once again, very broad, and the number of image features or feature vectors proposed in the literature is rather large. There is at present no reliable technique to automatically decide which feature vectors will be needed to accurately describe and quantify the essential content of an image. Feature vectors were recently used on a number of different microstructure types, including moment invariants for microstructures with precipitates (110, 111) and the Minkowski functionals, and related Betti numbers, for the mechanisms of grain growth

(112, 113). For an in-depth overview of statistical microstructure descriptors, we refer the reader to Ohser & Mücklich (114).

In recent years, a versatile and extensible framework has been created for the representation of a material's internal structure that promises to be applicable to a broad range of materials systems. A material's structure at any selected length scale is captured as a digital signal (115) denoted as m_s^h, which denotes the probability that a specified spatial bin (or voxel) indexed by s is physically occupied by a potential local state indexed by h. Because the values of m are bounded between zero and one (in many cases, a simple binary scale suffices), this approach produces a generalized representation for a broad range of materials systems at different length/structure scales. The local state can include any information needed to define a material's properties at the length scale of the spatial voxels and can include chemical composition, phase identifiers, orientations (e.g., in crystalline phases), and defect densities (e.g., dislocation density). In addition to transforming a material's structure into a versatile digital signal, this approach inherently treats the material's structure as a stochastic process because of the probabilistic interpretation of the variable m. The digital signal representation of structure offers many advantages, including fast computation of spatial correlations (38, 116–118); automated identification of salient structure features in large data sets (119); extraction of representative volume elements from an ensemble of data sets (120–122); reconstructions of structures from measured statistics (117, 123–125); building of real-time, searchable structure databases (126, 127); and mining of high-fidelity, multiscale structure-performance-structure evolution correlations from physics-based models (128–132).

2.6. Materials e-Collaboration Platforms

The practical realization of the data transformations described in **Figure 2** is feasible only with the accumulation of very large libraries of materials data that capture the relevant multiscale spatiotemporal information on materials internal structures covering a very broad range of materials classes. Given the amount of information needed, it is impossible for any single research group or single organization to assemble all this information. Furthermore, even if one were to somehow assemble these libraries, the expertise needed to efficiently mine them and curate the core materials knowledge needed to advance the accelerated development of new and improved materials would lie well outside the traditional skill sets of the materials practitioners and experts; this is actually the central focus of emerging e-collaboration platforms. It is therefore imperative to develop online collaboration platforms that are specifically designed to facilitate intimate collaborations between team members distributed geographically, organizationally, and on the basis of expertise.

Integrated workflows are foundational to the successful realization of the vision articulated above. These workflows would efficiently integrate high-value information and knowledge gleaned from different sources. Data are the essential currency of all transactions in such workflows. These integrated workflows are emerging at this time (e.g., Reference 133). The materials development community is poised to undertake the massive effort needed to design and validate such workflows in a systematic manner. This effort can be helped in significant ways if it is undertaken within a supporting Big Data cyberinfrastructure that will allow us to digitally capture the numerous trials undertaken in the development and validation of workflows (in both successful and failed trials). This database of workflows can then provide guidance for the design of new and improved workflows in the future, thereby allowing us to learn from successes and failures of past attempts in a highly objective manner.

Examples of such collaboration platforms are emerging and are becoming increasingly popular with the scientific research community. However, they currently tend to focus largely on the individual subtasks involved in the overall research enterprise. Prime examples include Google

Docs (134), Authorea (135), and ShareLaTex (136) for collaborative writing; Mendeley (137) and ResearchGate (138) for collaborative annotation and sharing of research documents; GitHub (11) and Sourceforge (139) for collaborative software development; Plotly (140) for collaborative data analysis and visualization; and Google+ (141) and LinkedIn (142) for e-teaming and networking.

Within the materials community, there have been several attempts to create e-collaboration sites, notably the Materials Microcharacterization Collaboratory at Oak Ridge National Laboratory (143) and the related TelePresence Microscopy Collaboratory at Argonne National Laboratory (144), which provide limited remote access to user instruments; neither of these sites has gained widespread visibility beyond the materials characterization community. TMS and NIST recently initiated the MGI (Materials Genome Initiative) Digital Data Community (145), a public discussion forum housed at LinkedIn, to help develop the materials innovation infrastructure that supports the MGI. Compared with the sites described in the previous paragraph, all these materials-centered e-collaboration efforts have been relatively small scale and low visibility. The Department of Energy currently supports the PRISMS Center (146) at the University of Michigan, which is developing the Materials Commons project. This project intends to create an evolving information repository as well as a virtual collaboration platform for the materials community and focuses on establishing an integrated multiscale modeling framework and the development of advanced open-source computational methods.

Whereas the above-mentioned collaboration platforms support specific subtasks involved in a collaborative research environment, there have also been efforts aimed at integrating one or more of these functionalities into comprehensive research collaboration platforms that allow for the development of emergent open-access research communities. Given the central role that computer codes play in any kind of data transformation, it is only natural that the collaboration platforms pay major attention to central functionalities related to code versioning, sharing, and curation. In this regard, several successful collaboration platforms have been built and deployed using HUBzero (12) software. One of the best-known examples is nanoHUB (13), which is focused on sharing and providing access to nanoscale science and engineering simulation tools. These platforms address the many challenges of code sharing by providing access to executable versions of highly sophisticated software on centralized computational resources (such as cloud computing resources, campus clusters, and national high-performance computing facilities). As another alternative, a number of recent efforts have focused on building collaboration communities (147) by using GitHub. These efforts have focused more on the transparency of the codes (through code versioning) and have allowed users to utilize their own independent computational resources for executing the codes. Approaches centered around open-source code sharing present tremendous flexibility and agility to the individual researcher in pursuing a wide range of new research avenues, many of which may not have been envisioned by the original code producer. GitHub-based research collaboration platforms also allow one to integrate other web services as needed (e.g., for sharing and annotating of data and results, discussions, and tracking of workflows). One such example is the recently launched maTIN environment at Georgia Tech (148), which involves collaborative research and education cutting across materials, manufacturing, and data sciences.

3. CONCLUSIONS

The emerging concepts described in this review build on advances in information sciences and DA and their application to materials data sets. The following low-hanging-fruit opportunities have been identified:

- The abundance of data generated by modern materials characterization equipment, as well as numerous sophisticated physics-based multiscale models, provides the incentive for the

development and implementation of data-driven protocols for objective decision support at various stages of materials development activities. Currently, many of these decisions are made in an ad hoc manner on the basis of empirical knowledge and the instincts of the experts involved in these workflows. This is one of the main factors leading to the expensive late-stage iterations that dramatically increase the overall cost and time of materials development efforts. Data-driven protocols may mitigate the inherent risk to a large extent, not only by making decisions more objective, but also by recording failures and successes such that the knowledge gained from these failures and successes is transferable to other endeavors.

- The availability of data and the use of data-driven protocols allow us to objectively quantify the uncertainty associated with the information and knowledge used in making decisions in materials development workflows. All knowledge or data generated by experiments or models are inherently associated with some level of uncertainty and are often incomplete. Rigorous use of statistics and probability theories offers a practical approach for addressing these challenges. DA and data-driven approaches are essential for taking advantage of these advanced tools.
- Success in materials development efforts is often predicated on the availability and engagement of cross-disciplinary expertise that covers both multimodal measurements and multiphysics simulations in a broad range of materials phenomena. Because this expertise is often highly localized in terms of specialization and is often distributed in terms of organizations and/or geography, there exist several hurdles to establishing highly productive collaborations. Modern data science and cyberinfrastructure provide the critically needed tools to mediate and accelerate such collaborations.
- Standardization and automation of workflows are highly desired for achieving process scalability. The current workflows used in most materials development efforts are highly customized and not scalable, which is a major contributing factor to the high cost of these efforts. Digital workflow recording is an important step toward standardization and automation. With digital capture of the workflows, it would become possible to identify best practices and implement them to achieve the desired acceleration of materials development at an affordable cost.

The materials community does not have the skill set needed to fully enable the materials data science revolution. We will need to work with scientists in other fields (e.g., computer science, data mining, statistics, and physics) to generate the necessary cyberinfrastructure. Given the relatively small size of the materials community compared with that of other engineering and science disciplines, convincing our colleagues in other fields that this endeavor is a worthwhile investment of time and resources may not be straightforward. However, the strategic importance of materials development for national security and the long-term economic health of the nation should provide sufficient justifications for funding agencies to make available the resources needed to establish a nationwide materials cyberinfrastructure.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

The authors acknowledge valuable input from Phil Withers, Euan Wielewski, Peter Voorhees, Michael Bockstaller, Charles Meneveau, Brian Puchala, John Allison, Michael Uchic, and Tresa

Pollock. Financial support from the Air Force Office of Scientific Research, MURI contract FA9550-12-1-0458, is gratefully acknowledged.

LITERATURE CITED

1. Cleveland WS. 2001. Data science: an action plan for expanding the technical areas of the field of statistics. *ISI Rev.* 69:21–26
2. Dhar V. 2013. Data science and prediction. *Commun. ACM* 56:64–73
3. Hey T, Tansley S, Tolle K, eds. 2009. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Redmond, WA: Microsoft Res.
4. Anderson C. 2008. The end of theory: The data deluge makes the scientific method obsolete. *Wired Mag.* 16:16.07
5. Linden G, Smith G, York J. 2003. Amazon.com recommendations: item-to-item collaborative filtering. *Internet Comput. IEEE* 7:76–80
6. Li I, Dey A, Forlizzi J. 2010. A stage-based model of personal informatics systems. In *Proc. SIGCHI Conference on Human Factors in Computing Systems*, pp. 557–66. New York: ACM
7. Hohman M, Gregory K, Chibale K, Smith P, Ekins S, Bunin B. 2009. Novel web-based tools combining chemistry informatics, biology and social networks for drug discovery. *Drug Discov. Today* 14:261–70
8. Tien JM. 2003. Toward a decision informatics paradigm: a real-time, information-based approach to decision making. *IEEE Trans. Syst. Man Cybern. C* 33:102–13
9. Wan TT. 2006. Healthcare informatics research: from data to evidence-based management. *J. Med. Syst.* 30:3–7
10. Dropbox, Inc. 2014. **http://www.dropbox.com**
11. GitHub. 2014. **https://github.com**
12. HUBzero. 2014. **https://hubzero.org**
13. nanoHUB. 2014. **https://nanohub.org**
14. National Science Board. 2005. *Long-lived digital data collections: enabling research and education in the 21st century*. Rep. NSB-05-40, National Science Board. **http://www.nsf.gov/pubs/2005/nsb0540**
15. Yu H, Kanov K, Perlman E, Graham J, Frederix E, et al. 2012. Studying Lagrangian dynamics of turbulence using in-demand fluid particle tracking in a public turbulence database. *J. Turbul.* 13:1–29
16. CERN Open Data Portal. 2014. **http://opendata.cern.ch**
17. Hubble Legacy Archive. 2014. **http://hla.stsci.edu**
18. CERN. 2014. **http://home.web.cern.ch/about**
19. Fermi National Accelerator Laboratory. 2014. **http://www.fnal.gov**
20. Relativistic Heavy Ion Collider. 2014. **http://www.bnl.gov/rhic**
21. Fuhrmann P. 2014. *dCache, the overview*. White Pap., dCache. **http://www.dcache.org/manuals/dcache-whitepaper-light.pdf**
22. dCache. 2014. **http://www.dcache.org**
23. DESY (Deutsches Elektronen-Synchrotron). 2014. **http://www.desy.de**
24. CERN: The Worldwide LHC Computing Grid. 2014. **http://home.web.cern.ch/about/computing/worldwide-lhc-computing-grid**
25. National Center for Biotechnology Information. 2014. **http://www.ncbi.nlm.nih.gov**
26. McDonald E, Brown C. 2014. *Working with Big Data in bioinformatics*. **http://www.aosabook.org/en/posa/working-with-big-data-in-bioinformatics.html**
27. NOAA (National Oceanic and Atmospheric Administration). 2014. **http://www.nesdis.noaa.gov**
28. NOAA View Data Exploration Tool. 2014. **http://www.nnvl.noaa.gov/view**
29. NOAA: National Operational Model Archive and Distribution System. 2014. **http://nomads.ncdc.noaa.gov/data.php**
30. GrADS Data Server. 2014. **http://grads.iges.org/grads/gds/index.html**
31. OPeNDAP. 2014. **http://opendap.org**
32. Earth Observing System Data and Information System. 2014. **https://earthdata.nasa.gov/about-eosdis**
33. National Snow and Ice Data Center. 2014. **http://nsidc.org/daac/data-sets.html**

34. White AA. 2013. Big data are shaping the future of materials science. *MRS Bull.* 38:594–95
35. Committee on Integrated Computational Materials Engineering, National Research Council. 2008. *Integrated Computational Materials Engineering: A Transformational Discipline for Improved Competitiveness and National Security*. Washington, DC: The National Academies Press. **http://www.nap.edu/catalog/12199/integrated-computational-materials-engineering-a-transformational-discipline-for-improved-competitiveness**
36. National Science and Technology Council, Executive Office of the President. 2011. *Materials genome initiative for global competitiveness*. **http://www.whitehouse.gov/sites/default/files/microsites/ostp/materials_genome_initiative-final.pdf**
37. Van Tendeloo G, Van Dyck D, Pennycook SE. 2012. *Handbook of Nanoscopy*. Weinheim, Ger.: Wiley-VCH
38. Kalidindi SR. 2015. Data science and cyberinfrastructure: critical enablers for accelerated development of hierarchical materials. *Int. Mater. Rev.* 60:150–68
39. McNulty E. 2014. *Understanding Big Data: the seven V's*. **http://dataconomy.com/seven-vs-big-data/**
40. Lienert U, Li SF, Hefferan CM, Lind J, Suter RM, et al. 2011. High-energy diffraction microscopy at the Advanced Photon Source. *JOM* 63:70–77
41. Advanced Photon Source, Argonne National Laboratory. 2014. **https://www1.aps.anl.gov**
42. Auciello O. 2013. The materials research community studies magnitude of Big Data. *MRS Bull.* 38:766–67
43. Mies D. 2002. Managing materials data. In *Handbook of Materials Selection*, ed. M Kutz, Chapter 17. New York: John Wiley & Sons
44. Citrine Informatics. 2014. **http://www.citrination.com**
45. Clean Energy Project. 2014. **http://cleanenergy.molecularspace.org**
46. The Materials Project. 2014. **http://www.materialsproject.org**
47. Automatic-FLOW for Materials Discovery. 2014. **http://www.aflowlib.org**
48. CALPHAD (Computer Coupling of Phase Diagrams and Thermochemistry). 2014. **http://www.calphad.org**
49. Open Quantum Materials Database. 2014. **http://oqmd.org**
50. NIST (National Institute of Standards and Technology) Data Gateway. 2014. **http://srdata.nist.gov/gateway/gateway?dblist=1**
51. NIST Material Measurement Laboratory. 2014. **http://www.ctcms.nist.gov/potentials/**
52. MatWeb. 2014. **http://www.matweb.com/**
53. Granta. 2014. **http://www.grantadesign.com/products/ces/**
54. MatNavi (NIMS Materials Database). 2014. **http://mits.nims.go.jp/index_en.html**
55. Freiman S, Madsen L, Rumble J. 2011. A perspective on materials databases. *Am. Ceram. Soc. Bull.* 90:28–32
56. Kaufman J. 1986. The National Materials Property Data Network, Inc.—the technical challenges and the plan. *Mater. Prop. Data* 1:159–63
57. Adams BL, Kalidindi SR, Fullwood DT. 2012. *Microstructure Sensitive Design for Performance Optimization*. Oxford, UK: Butterworth-Heinemann
58. Milton GW. 2001. *The Theory of Composites*. Cambridge, UK: Cambridge Univ. Press
59. Torquato S. 2002. *Random Hetereogeneous Materials*. New York: Springer-Verlag
60. Panchal JH, Kalidindi SR, McDowell DL. 2013. Key computational modeling issues in integrated computational materials engineering. *J. Comput. Aided Des.* 45:4–25
61. DREAM.3D. 2014. **http://dream3d.bluequartz.net**
62. Materials Atlas. 2014. **https://cosmicweb.mse.iastate.edu/wiki/display/home/materials+atlas+home**
63. Computational Materials Data Network. 2014. **http://www.asminternational.org/web/cmdnetwork/about**
64. Groeber MA, Jackson MA. 2014. DREAM.3D: a digital representation environment for the analysis of microstructure in 3D. *Integr. Mater. Manuf. Innov.* 3:5
65. Material Data Management Consortium. 2014. **http://www.mdmc.net**

66. TMS (The Minerals, Metals and Materials Society). 2014. **http://www.tms.org/**
67. The Materials Cyberinfrastructure Portal. 2014. **http://www.tms.org/cyberportal/**
68. Patterson DA, Gibson G, Katz RH. 1988. A case for redundant arrays of inexpensive disks (RAID). In *Proc. 1988 ACM SIGMOD International Conference on Management of Data*, pp. 109–16. Chicago: ACM
69. Schroeder B, Gibson G. 2007. Disk failures in the real world: What does an MTTF of 1,000,000 hours mean to you? In *Proc. 5th USENIX Conference on File and Storage Technologies* (*FAST'07*). San Jose, CA: USENIX
70. Ghemawat S, Gobioff H, Leung ST. 2003. The Google File System. In *Proc. 19th ACM Symposium on Operating System Principles*. Bolton Landing, NY: ACM
71. Healey CG. 2014. *CSC541: advanced data structures*. Course Notes, Dep. Comput. Sci., NC State Univ. **http://www.csc.ncsu.edu/faculty/healey/csc541/notes/file_sys.pdf**
72. Rodeh O, Teperman A. 2003. zFS—a scalable distributed file system using object disks. In *Proc. 20th IEEE Conference on Mass Storage Systems and Technology*, pp. 207–18. San Diego, CA: IEEE
73. Shvachko K, Hairong K, Radia S, Chansler R. 2010. The Hadoop distributed file system. In *Proc. 26th IEEE Symposium on Mass Storage Systems and Technologies*, pp. 1–10. Incline Village, NY: IEEE
74. Jackson M, Groeber M, Uchic M, Rowenhorst D, De Graef M. 2014. h5ebsd: an archival data format for electron back-scatter diffraction data sets. *Integr. Mater. Manuf. Innov.* 3:4
75. Jackson M, Simmons J, De Graef M. 2010. MXA: a customizable HDF5-based data format for multi-dimensional data sets. *Model. Simul. Mater. Sci. Eng.* 18:065008
76. The HDF Group. 2014. **http://www.hdfgroup.org/**
77. NetCDF. 2014. **http://www.unidata.ucar.edu/software/netcdf/index.html**
78. PDB (Protein Data Bank). 2014. **http://www.pdb.org/pdb/home/home.do**
79. FITS Support Office (NASA/Goddard Space Flight Center). 2014. **http://fits.gsfc.nasa.gov/**
80. DICOM. 2014. **http://medical.nema.org/**
81. National Digital Information Infrastructure and Preservation Program. 2014. **http://www.digitalpreservation.gov**
82. Community Owned Digital Preservation Tool Registry. 2014. **http://coptr.digipres.org/main_page**
83. ISO (International Organization for Standardization) 26234:2012. 2014. **http://www.iso.org/iso/catalogue_detail.htm?csnumber=43506**
84. DOI. 2014. **http://www.doi.org/**
85. Handle.Net. 2014. **http://www.handle.net/index.html**
86. Corporation for National Research Initiatives. 2014. **http://www.cnri.reston.va.us**
87. Globus Online GridFTP. 2014. **http://toolkit.globus.org/toolkit/docs/latest-stable/gridftp/**
88. Internet Engineering Task Force. 2014. **https://www.ietf.org**
89. Globus Online. 2014. **https://www.globus.org**
90. Kumar S, Edwards J, Bremer PT, Knoll A, Christensen C, et al. 2014. Efficient I/O and storage of adaptive-resolution data. In *Proc. International Conference for High Performance Computing, Networking, Storage and Analysis*, pp. 413–23. Piscataway, NJ: IEEE
91. Li Y, Perlman E, Wan M, Yang Y, Meneveau C, et al. 2008. A public turbulence database cluster and applications to study Lagrangian evolution of velocity increments in turbulence. *J. Turbul.* 9:1–29
92. ASTM Int. 2013. *Standard test methods for determining average grain size*. ASTM E112, ASTM Int.
93. ASTM Int. 2008. *Standard test methods for characterizing duplex grain sizes*. ASTM E1181, ASTM Int.
94. Russ J. 1992. *The Image Processing Handbook*. Boca Raton, FL: CRC
95. Hall E. 1951. The deformation and ageing of mild steel. III. Discussion of result. *Proc. Phys. Soc. B* 64:747–53
96. Petch N. 1953. Cleavage strength of polycrystals. *Iron Steel Inst. J.* 174:25–28
97. Argon A. 2008. *Strengthening Mechanisms in Crystal Plasticity*. Oxford, UK: Oxford Univ. Press
98. Reed-Hill R, Abbaschian R. 1994. *Physical Metallurgy Principles*. Boston: PWS. 3rd ed.
99. Rajan K. 2005. Materials informatics. *Mater. Today* 8:38–45
100. Gorse D, Lahana R. 2000. Functional diversity of compound libraries. *Curr. Opin. Chem. Biol.* 4:287–94
101. Curtarolo S, Morgan D, Persson K, Rodgers J, Ceder G. 2003. Predicting crystal structures with data mining of quantum calculations. *Phys. Rev. Lett.* 91:135503

102. Ceder G. 1998. Predicting properties from scratch. *Science* 280:1099–100
103. Breneman C, Brinson L, Schadler L, Natarajan B, Krein M, et al. 2013. Stalking the materials genome: a data-driven approach to the virtual design of nanostructured polymers. *Adv. Funct. Mater.* 23:5746–52
104. Cebon D, Ashby M. 2006. Engineering materials informatics. *MRS Bull.* 31:1004–12
105. Serra J. 2006. A lattice approach to image segmentation. *J. Math. Imag. Vis.* 24:83–130
106. Waggoner J, Simmons J, De Graef M, Wang S. 2013. Multi-structure propagation incorporating homeomorphism for materials image segmentation. *IEEE Trans. Image Process.* 22:5282–93
107. Comer M, Bouman CA, De Graef M, Simmons JP. 2011. Bayesian methods for image segmentation. *JOM* 63:55–57
108. EM/MPM Workbench. 2014. **http://www.bluequartz.net/?page_id=97**
109. Simmons J, Chuang P, Comer M, De Graef M, Uchic M, Spowart J. 2009. Application and further development of advanced image processing algorithms for automated analysis of serial section image data. *Model. Simul. Mater. Sci. Eng.* 17:025002
110. MacSleyne J, Simmons J, De Graef M. 2008. On the use of 2-D moment invariants for the automated classification of particle shapes. *Acta Mater.* 56:427–37
111. MacSleyne J, Simmons J, De Graef M. 2008. On the use of moment invariants for the automated analysis of 3-D particle shapes. *Model. Simul. Mater. Sci. Eng.* 16:045008
112. Hütler M, Rutledge G, Armstrong R. 2005. Crystal shapes and crystallization in continuum modeling. *Phys. Fluids* 17:014107
113. MacPherson R, Srolovitz D. 2007. The von Neumann relation generalized to coarsening of three-dimensional microstructures. *Nature* 446:1053–55
114. Ohser J, Mücklich F. 2000. *Statistical Analysis of Microstructures in Materials Science*. West Sussex, UK: John Wiley & Sons
115. Adams BL, Gao X, Kalidindi SR. 2005. Finite approximations to the second-order properties closure in single phase polycrystals. *Acta Mater.* 53:3563–77
116. Niezgoda SR, Fullwood DT, Kalidindi SR. 2008. Delineation of the space of 2-point correlations in a composite material system. *Acta Mater.* 56:5285–92
117. Fullwood DT, Niezgoda SR, Kalidindi SR. 2008. Microstructure reconstructions from 2-point statistics using phase-recovery algorithms. *Acta Mater.* 56:942–48
118. Fullwood DT, Niezgoda SR, Adams BL, Kalidindi SR. 2010. Microstructure sensitive design for performance optimization. *Prog. Mater. Sci.* 55:477–562
119. Niezgoda SR, Kalidindi SR. 2009. Applications of the phase-coded generalized Hough transform to feature detection, analysis, and segmentation of digital microstructures. *Comput. Mater. Contin.* 14:79–97
120. Niezgoda SR, Turner DM, Fullwood DT, Kalidindi SR. 2010. Optimized structure based representative volume element sets reflecting the ensemble-averaged 2-point statistics. *Acta Mater.* 58:4432–45
121. Wargo EA, Hanna AC, Çeçen A, Kalidindi SR, Kumbur EC. 2012. Selection of representative volume elements for pore-scale analysis of transport in fuel cell materials. *J. Power Sources* 197:168–79
122. Qidwai SM, Turner DM, Niezgoda SR, Lewis AC, Geltmacher AB, et al. 2012. Estimating response of polycrystalline materials using sets of weighted statistical volume elements (WSVEs). *Acta Mater.* 60:5284–99
123. Fullwood DM, Kalidindi SR, Niezgoda SR, Fast A, Hampson N. 2008. Gradient-based microstructure reconstructions from distributions using fast Fourier transforms. *Mater. Sci. Eng. A* 494:68–72
124. Bochenek B, Pyrz R. 2004. Reconstruction of random microstructures: a stochastic optimization problem. *Comput. Mater. Sci.* 31:93–112
125. Roberts AP. 1997. Statistical reconstruction of three-dimensional porous media from two-dimensional images. *Phys. Rev. E* 56:3203–12
126. Niezgoda SR, Kanjarla AK, Kalidindi SR. 2013. Novel microstructure quantification framework for databasing, visualization, and analysis of microstructure data. *Integr. Mater. Manuf. Innov.* 2:3
127. Kalidindi SR, Niezgoda SR, Salem AA. 2011. Microstructure informatics using higher-order statistics and efficient data-mining protocols. *JOM* 63:34–41
128. Kalidindi SR. 2012. Computationally-efficient fully-coupled multi-scale modeling of materials phenomena using calibrated localization linkages. *ISRN Mater. Sci.* 2012:305692

129. Fast T, Niezgoda SR, Kalidindi SR. 2011. A new framework for computationally efficient structure-structure evolution linkages to facilitate high-fidelity scale bridging in multi-scale materials models. *Acta Mater.* 59:699–707
130. Fast T, Kalidindi SR. 2011. Formulation and calibration of higher-order elastic localization relationships using the MKS approach. *Acta Mater.* 59:4595–605
131. Landi G, Niezgoda SR, Kalidindi SR. 2010. Multi-scale modeling of elastic response of three-dimensional voxel-based microstructure datasets using novel DFT-based knowledge systems. *Acta Mater.* 58:2716–25
132. Yabansu YC, Patel DK, Kalidindi SR. 2014. Calibrated localization relationships for elastic response of polycrystalline aggregates. *Acta Mater.* 81:151–60
133. Salem AA, Shaffer JB, Satko DP, Semiatin SL, Kalidindi SR. 2014. Workflows for integrating mesoscale heterogeneities in materials structure with process simulation of titanium alloys. *Integrating Mater. Manuf. Innov.* 3:24
134. Google Docs. 2014. **https://docs.google.com/**
135. Authorea. 2014. **https://www.authorea.com/**
136. ShareLaTeX. 2014. **https://www.sharelatex.com/**
137. Mendeley. 2014. **http://www.mendeley.com/**
138. ResearchGate. 2014. **http://www.researchgate.net/**
139. Sourceforge. 2014. **http://sourceforge.net/**
140. Plotly. 2014. **https://plot.ly/**
141. Google+. 2014. **https://plus.google.com/**
142. LinkedIn. 2014. **https://www.linkedin.com/**
143. Materials Microcharacterization Collaboratory. 2014. **http://web.ornl.gov/sci/doe2k/MICSReview/99/**
144. TelePresence Microscopy Collaboratory. 2014. **http://tpm.amc.anl.gov**
145. MGI (Materials Genome Initiative) Digital Data Community. 2014. **https://www.linkedin.com/groups/mgi-digital-data-community-7459917**
146. The PRISMS Center: Materials Commons. 2014. **http://prisms.engin.umich.edu/#/prisms**
147. Dabbish L, Stuart C, Tsay J, Herbsleb J. 2012. Social coding in GitHub: transparency and collaboration in an open software repository. In *Proc. ACM 2012 Conference on Computer Supported Cooperative Work*, pp. 1277–86. New York: ACM
148. maTIN. 2015. **http://materials.gatech.edu/matin**