Materials science with large-scale data and informatics: Unlocking new opportunities

Joanne Hill, Gregory Mulholland, Kristin Persson, Ram Seshadri, Chris Wolverton, and Bryce Meredig

Universal access to abundant scientific data, and the software to analyze the data at scale, could fundamentally transform the field of materials science. Today, the materials community faces serious challenges to bringing about this data-accelerated research paradigm, including diversity of research areas within materials, lack of data standards, and missing incentives for sharing, among others. Nonetheless, the landscape is rapidly changing in ways that should benefit the entire materials research enterprise. We provide an overview of the current state of the materials data and informatics landscape, highlighting a few selected efforts that make more data freely available and useful to materials researchers.

Introduction

Data-intensive science has been described as the "fourth paradigm" for scientific exploration, with the first three being experiments, theory, and simulation.[1] While the value of data-intensive research approaches are becoming more apparent, the field of materials science has not yet experienced the same widespread adoption of these methods (as has occurred in biosciences,[2] astronomy,[3] and particle physics[4]). Nonetheless, the potential impact of data-driven materials science is tremendous: Materials informatics could reduce the typical 10–20 year development and commercialization cycle[5] for new materials. We see plentiful opportunities to use data and data science to radically reduce this timeline and generally advance materials research and development (R&D) and manufacturing.

In this article, we discuss the current state of affairs with respect to data and data analytics in the materials community, with a particular emphasis on thorny challenges and promising initiatives that exist in the field. We conclude with a set of near-term recommendations for materials-data stakeholders. Our goal is to demystify data analytics and give readers from any subdiscipline within materials research enough information to understand how informatics techniques could apply to their own workflows.

Challenges surrounding data: The status quo in materials

There are five principal barriers to broader data sharing and large-scale meta-analysis within the field of materials science. This section enumerates and discusses the following barriers in depth: (1) opaque buzzwords in materials informatics, which prevent a typical materials scientist from readily seeing how data-driven methods could apply to their work; (2) idiosyncrasies in individual researchers' preferred data workflows; (3) a wide variety of stakeholders, who often have conflicting goals, hailing from corresponding diverse research areas; (4) limited availability of structured data and agreed-upon data standards; and (5) a lack of clear incentives to share data.

Proliferation of buzzwords

Like many areas of science, materials informatics is unfortunately hamstrung by the proliferation of buzzwords whose meanings are not clear to researchers in the broader materials community. To a first approximation, machine learning, data mining, and artificial intelligence are roughly interchangeable and refer to the use of algorithms to approximately model patterns in data. Materials informatics, in analogy to bioinformatics,

Joanne Hill, Citrine Informatics, USA; jo@citrine.io
Gregory Mulholland, Citrine Informatics, USA; greg@citrine.io
Kristin Persson, Lawrence Berkeley National Laboratory, USA; kapersson@lbl.gov
Ram Seshadri, University of California, Santa Barbara, USA; seshadri@mrl.ucsb.edu
Chris Wolverton, Northwestern University, USA; c-wolverton@northwestern.edu
Bryce Meredig, Citrine Informatics, USA; bryce@citrine.io
doi:10.1557/mrs.2016.93

refers to developing an understanding of materials using data and algorithms. Thus, machine learning is a key tool for researchers in the materials informatics domain.

While "big data" has become a fashionable term in the materials informatics and broader materials communities, the reality is that very few materials researchers outside of large user facilities and some specialized communities such as tomography and combinatorial chemistry generate data that meet the traditional "3V's" big-data definition of high volume, velocity, and variety.[6] Real-world examples of big data include YouTube streaming four billion hours of video per month (which is over six exabytes, or six billion gigabytes, at 1080p resolution), and Twitter receiving about half a million tweets per minute during the Brazil–Germany soccer match in the 2014 World Cup. To give a contrasting example from materials science, performing 100,000 density functional simulations to predict the electronic structures of many known crystalline materials—while scientifically very useful—is not the domain of big data. To store all of the meaningful scientific outputs of the simulations would not require more than a few terabytes of storage, which fits easily on a single hard drive (considered low volume); the simulations finish relatively slowly (i.e., low velocity—completed calculations arrive at the rate of a few results per hour, perhaps, and not thousands of results per second); and the output data are all completely uniform (i.e., no variety by definition). YouTube, Twitter, Google, and Facebook deal with big data; most materials researchers do not, although that fact takes nothing away from the unique data challenges facing the materials community.

We also wish to draw a clear distinction between computational materials science and materials informatics. The former generally refers to using physics-based tools, such as density functional theory (DFT), molecular dynamics, or phase-field simulations, to model behavior of materials. In contrast, models developed by materials informatics are data based, not physics based (i.e., there is no underlying governing equation, such as the Schrödinger equation), nor is informatics specific to "computational" researchers. For example, experimentalists performing tomographic or high-throughput x-ray diffraction studies generate tremendous quantities of data and may thus turn to algorithmic approaches to process and understand these data at scale. We would designate such activities as materials informatics.

Finally, many researchers, especially those who study structural metals, may be familiar with the subdiscipline of computational materials science called integrated computational materials engineering (ICME).[7] This framework involves connecting physics-based models at various length scales (e.g., atomistic simulation, dislocation modeling, thermodynamic modeling, continuum modeling) to predictively model alloy systems. Materials informatics can complement ICME in two ways: (1) by predictively supplying key materials property parameters for underlying ICME models, if those parameters are not known *a priori*; and (2) by integrating the outputs of an ICME workflow into higher-level machine learning-based models of materials behavior.

Idiosyncratic data workflows

Data workflows in materials vary considerably and depend on a number of factors, such as specific research focus, data-acquisition techniques, and individual researcher's personal preferences. While materials researchers have long been producing a wealth of knowledge relating to the processing–structure–properties–performance relationships of materials, this information is generated, analyzed, and disseminated in such a wide variety of ways that researchers face tremendous difficulties in reusing and repurposing others' data. We have found in the course of software usability interviews with materials researchers at universities and companies that virtually every individual makes a unique set of choices among characterization or simulation tools, data warehousing methods, data analysis methods, and data reporting avenues. This fragmentation among workflows makes centralization and standardization of materials data far more challenging; we visualize the situation in **Figure 1**.

Wide variety of stakeholders and research areas

Materials science is a broad and interdisciplinary field in which progress emerges from complex interactions between producers of data (i.e., researchers at universities, government labs, and industry), funding agencies, makers of equipment and software, and distributors of research results (often journal publishers). These stakeholders are all instrumental in the materials field, but often do not share aligned incentives when it comes to making materials data broadly available.

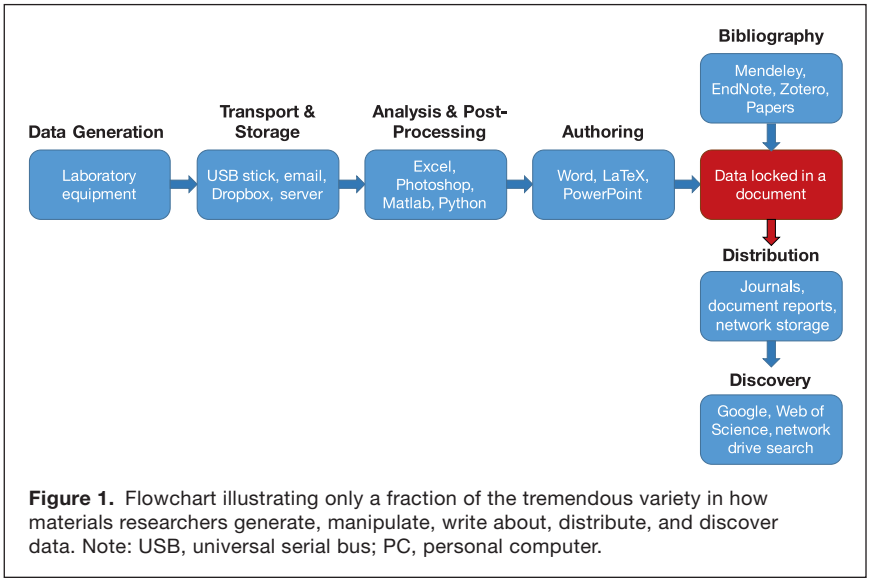


Figure 1. Flowchart illustrating only a fraction of the tremendous variety in how materials researchers generate, manipulate, write about, distribute, and discover data. Note: USB, universal serial bus; PC, personal computer.

For example, researchers may wish to keep particularly exciting results private for extended periods to avoid the risk of others publishing those results, and they even withhold negative results from publication altogether; industrial researchers generally withhold their most interesting results as trade secrets; characterization equipment-makers often design proprietary data formats in an effort to differentiate their tools through software; and publishers have the incentive to generate revenue from controlling access to materials publications and data, for example, by offering for-pay databases.[8,9] **Figure 2** lists some of the stakeholders in the materials-data landscape, broken down by category.

Data decentralization, limited access to structured data, and missing data standards

Decentralization

The substantial diversity among subdisciplines within materials science and engineering is often cited as a reason why a unified data infrastructure for materials research is impractical; instead, the community will be forced to adopt a federated system of smaller databases.[10] One can make the counterargument, however, that enabling cross-pollination among different areas of materials and creating a "one-stop-shop," comprehensive data clearinghouse is crucial to the advancement of materials research, and hence, we should focus on building such a system in spite of the inherent challenges of doing so. The National Institute of Standards and Technology (NIST) Materials Data Curation System[11] and Citrine Informatics' Citrination platform[12] are two such very broad materials-data infrastructures whose goal is to structure and store a wide variety of materials research data.

The current materials data landscape is a highly fragmented patchwork quilt of smaller databases, each customized to present information from a specific subdiscipline. We have created an extensive, yet inevitably incomplete, list of materials-data resources (see **Table I**).

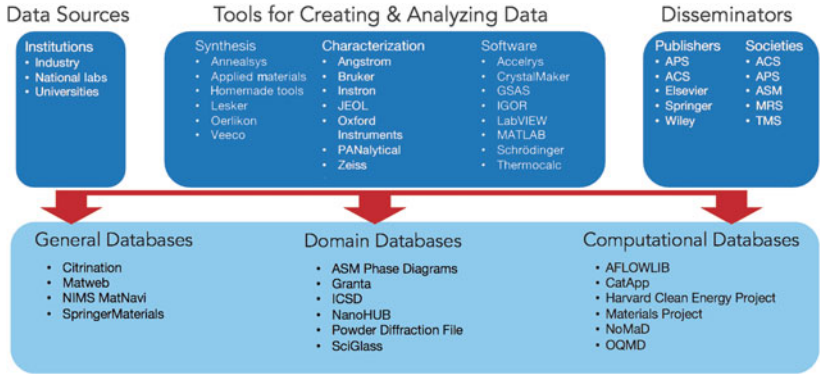


Figure 2. A partial list of the hundreds of organizations that all play important roles in the generation, distribution, and storage of materials data. Note: AFLOW, Automatic-FLOW for Materials Discovery; APS, American Physical Society; ACS, American Chemical Society; MRS, Materials Research Society; TMS, The Minerals, Metals and Materials Society; ASM, American Society for Metals; NIMS, National Institute for Materials Science; ICSD, Inorganic Crystal Structure Database; AFLOWLIB, Repository associated with the AFLOW project.

Limited access to structured data

The vast majority of materials-data resources available today are optimized for "low-throughput" human consumption (e.g., via a graphical interface). Modern data analytics techniques, however, rely on systematic computational access to very large stores of data through an application programming interface (API).[13,14] While systematic access to large data sets is widespread (e.g., the genomics community),[15] the current status quo in materials data is fundamentally incompatible with state-of-the-art methods of computationally extracting insights from data. In Table I, we note that the vast majority of data resources are not bulk downloadable, which essentially renders them unavailable to data analytics unless they are first scraped or extracted by other software methods. For example, the Inorganic Crystal Structure Database (ICSD)[16] is an invaluable, authoritative collection of crystallographic data on tens of thousands of materials; because it is not bulk downloadable, however, researchers face significant barriers to analyzing its contents in aggregate.

Missing data standards

Data standards are another key to revolutionizing the materials data landscape. Data decentralization in materials has led to a wide variety of choices in terms of data storage techniques. Most of the data resources in Table I employ idiosyncratic data formats under the hood, and the materials community has few widely adopted data standards (the Crystallographic Information File, or CIF, is a notable exception; it is the gold standard for representing crystal structure data).[17] There are general repositories, such as Dryad and Figshare, that store data from a large number of unrelated scientific fields.[18,19] Repositories such as these allow data to be uploaded in any format. While this broadens public access to raw scientific data, it does not necessarily facilitate reuse and analytics, as the information is often formatted in such a way that other researchers would have tremendous difficulty interpreting it.

The lack of data standards in materials greatly complicates the task of gaining useful insight from large-scale materials data. Flexible, uniform, computer-readable data standards should be established to enable data to be shared and systematically mined. Task forces and working groups have been convened to address this issue, but achieving broad agreement on data standards among diverse stakeholders has proven challenging. Citrine Informatics is working to nucleate grassroots support for a flexible JavaScript Object Notation (JSON)-based materials-data format[20] that provides a semistructured means to represent a wide range of materials data, but success with this initiative will depend strongly on uptake by the materials community at large. We provide a more detailed overview of some

Table I. A list of some notable materials-data resources.

Name	URL	Category	Free/Non-Free
3D Materials Atlas	cosmicweb.mse.iastate.edu/wiki/display/home/Materials+Atlas+Home	3D Characterization	Free
AFLOWLIB	aflowlib.org	Computational	Free
AIST Research Information Databases	www.aist.go.jp/aist_e/list/database/riodb	General Materials Data	Free
American Mineralogist Crystal Structure Database	rruff.geo.arizona.edu/AMS/amcsd.phP	Minerals	Free
ASM Alloy Center Database	mio.asminternational.org/ac	Alloys	Non-Free
ASM Phase Diagrams	www1.asminternational.org/AsmEnterprise/APD	Thermodynamics	Non-Free
CALPHAD databases (e.g., Thermocalc SGTE)	www.thermocalc.com/products-services/databases/thermodynamic	Thermodynamics	Non-Free
Cambridge Crystallographic Data Centre	www.ccdc.cam.ac.uk/pages/Home.aspx	Crystallography	Non-Free
CatApp	suncat.stanford.edu/catapp	Catalysts	Free
Chemspider	www.chemspider.com	Chemical data	Free
CINDAS High-Performance Alloys Database	cindasdata.com/products/hpad	Alloys	Non-Free
Citrination	citrination.com	General Materials Data	Free
Computational Materials Repository	cmr.fysik.dtu.dk	Computational	Free
CRC Handbook	www.hbcpnetbase.com	General Materials Data	Non-Free
CrystMet	cds.dl.ac.uk/cgi-bin/news/disp?crystmet	Crystallography	Non-Free
Crystallography Open Database	http://www.crystallography.net	Crystallography	Free
DOE Hydrogen Storage Materials Database	www.hydrogenmaterialssearch.govtools.us	Hydrogen Storage	Free
Granta CES Selector	www.grantadesign.com/products/ces	General Materials Data	Non-Free
Handbook of Optical Constants of Solids, Palik	N/A	Hard-Copy Sources	Non-Free
Harvard Clean Energy Project	cepdb.molecularspace.org	Computational	Free
Inorganic Crystal Structure Database	cds.dl.ac.uk/cds/datasets/crys/icsd/llicsd.html	Crystallography	Non-Free
International Glass Database System	www.newglass.jp/interglad_n/gaiyo/info_e.html	Glass	Non-Free
Knovel	app.knovel.com/web/browse.v	General Materials Data	Non-Free
Matbase	www.matbase.com	General Materials Data	Free
MatDat	www.matdat.com	General Materials Data	Non-Free
Materials Project	www.materialsproject.org	Computational	Free
MatNavi (NIMS)	mits.nims.go.jp/index_en.html	General Materials Data	Free
MatWeb	www.matweb.com	General Materials Data	Free
Mindat	www.mindat.org	Minerals	Free
NanoHUB	nanohub.org	Nanomaterials	Free
Nanomaterials Registry	www.nanomaterialregistry.org	Nanomaterials	Free
NIST Materials Data Repository (DSpace)	materialsdata.nist.gov/dspace/xmlui	General Materials Data	Free
NIST Interatomic Potentials Repository	www.ctcms.nist.gov/potentials	Computational	Free
NIST Standard Reference Data	www.nist.gov/srd/dblistpcdatabases.cfm	General Materials Data	Non-Free
NIST Standard Reference Data	www.nist.gov/srd/onlinelist.cfm	General Materials Data	Free
NoMaD	nomad-repository.eu/cms	Computational	Free
Open Knowledge Database of Interatomic Models (Open KIM)	openkim.org	Computational	Free
Open Quantum Materials Database	oqmd.org	Computational	Free

Table I. A list of some notable materials-data resources.

Name	URL	Category	Free/Non-Free
Pauling File	paulingfile.com	General Materials Data	Non-Free
Pearson's Handbook: Crystallographic Data	N/A	Hard-Copy Sources	Non-Free
Powder Diffraction File (PDF)	www.icdd.com/products/index.htm	Crystallography	Non-Free
PubChem	pubchem.ncbi.nlm.nih.gov	Chemical data	Free
Reaxys	www.elsevier.com/solutions/reaxys	Chemical data	Non-Free
Scifinder/ChemAbstracts	scifinder.cas.org	Chemical data	Non-Free
SciGlass	www.sciglass.info	Glass	Non-Free
SpringerMaterials	materials.springer.com	General Materials Data	Non-Free
Metallurgical Thermochemistry, Kubaschewski	N/A	Hard-Copy Sources	Non-Free
TEDesignLab	www.tedesignlab.org	Thermoelectrics	Free
Total Materia	www.totalmateria.com	General Materials Data	Non-Free
UCSB-MRL thermoelectric database	www.mrl.ucsb.edu:8080/datamine/thermoelectric.jsp	Thermoelectrics	Free

Note: AFLOWLIB, Automatic-FLOW for Materials Discovery; AIST, National Institute of Advanced Industrial Science and Technology (Japan); ASM, American Society for Metals; CALPHAD, CALculation of PHAse Diagrams; SGTE, Scientific Group Thermodata Europe; CINDAS, Center for Information and Numerical Data Analysis and Synthesis; CRC, Chemical Rubber Company; DOE, US Department of Energy; CES, Cambridge Engineering Selector; NIMS, National Institute for Materials Science; NIST, National Institute of Standards and Technology; KIM, Knowledge Database of Interatomic Models; UCSB MRL, University of California, Santa Barbara Materials Research Laboratory.

existing standards in our recommendations for key next steps. Here, we simply note that organizations such as IEEE and W3C are potential hardware, software, and Internet-focused models for promulgating data standards in materials.

Lack of incentives

The typical materials researcher today experiences minimal incentive for sharing data. Other research communities, such as biological sciences and astronomy, are often used as exemplars of data-dissemination practices; however, these groups have unique sets of data sharing requirements, norms, and incentives that may not directly transfer to materials.[21] In the materials community specifically, it is not clear that making one's research data broadly available will lead to any of the following: (1) enhanced impact and more citations for one's work; (2) improved funding opportunities; or (3) improved chances at professional advancement and promotion. The National Science Foundation (NSF) and US Department of Energy (DOE) are two major funding agencies that now require a data management plan for funded research,[22,23] though it is not clear what the consequences might be for researchers who do not make a good-faith effort to deliver on these plans. Going forward, it will be vital for funding agencies and journal publishers to encourage data sharing by rewarding researchers who offer their data to the community or by prescribing data warehousing practices as does the National Institutes of Health and many biological sciences journals.

Not only does the materials community lack incentives for data sharing, it also lacks an obvious forcing function that necessitates a culture of structured data access and advanced analytics. In genomics, for example, the volumes of data produced even in routine laboratory experiments are so great that data-driven approaches are essential in the field. In contrast, many materials science and engineering researchers have been able to continue using traditional "small data" generation and analysis approaches, in spite of the potential advantages of harnessing large-scale data analytics techniques to inform research in many sub-fields.

The situation is gradually changing. We believe that researchers who adopt data standards and make their research widely available via data repositories will win in several important ways by: (1) getting ahead of competitors by integrating machine learning and data analytics in their workflows; (2) making their research more discoverable—on the Citrine Informatics' Citrination materials data platform, for example, a highly accessed data set from the Open Quantum Materials Database[24] can attract 50–100 views per week (see **Figure 3**), comparable engagement to a high-profile paper in a reputed journal two to three months after publication; and (3) saving time in finding and analyzing data by taking advantage of automation and software. McKinsey estimated that knowledge workers, which include materials scientists, spend about 20% of their time looking for information;[25] establishing a materials data infrastructure can help reduce this overhead time burden.

New thinking around materials data

Having outlined the very real barriers facing widespread adoption of data-driven materials science, we now take a more

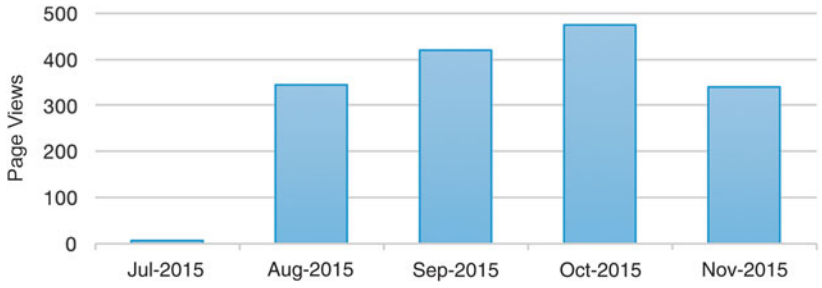


Figure 3. Access statistics (page views/month) for the Open Quantum Materials Database (OQMD) paper after it appeared on the Citrination platform. The spike in August 2015 resulted from Google indexing the OQMD data set on Citrination.

encouraging tone: the data status quo in materials science is changing for the better. Specifically, more data and research outputs are becoming available as open-access content; funding agencies are acknowledging meta-analyses of data as key to progress in materials research; several notable projects have emerged to unite software, data, and web infrastructure for the benefit of the materials community; and industrial stakeholders are beginning to recognize that a pressing need exists for open manufacturing data.

Open-access movement

The open-access (OA) paradigm, in which readers are able to view and (sometimes) repurpose published research at no cost, is gaining traction as key stakeholders jump on board. More publishers are adopting OA models, and increasing numbers of papers are appearing under a creative commons license, which makes content and data freely available. The Nature Publishing Group, for example, launched the journal *Scientific Data* in 2014, which is OA and dedicated specifically to redistributing important scientific data sets. Unfortunately, among the journal's 60+ recommended repositories, fewer than five are dedicated to materials and chemical data, illustrating just how badly materials is lagging other disciplines in terms of data warehousing. By contrast, the biological sciences have over 40 approved repositories across subdomains ranging from omics to taxonomy.

Governments around the world are intensifying their efforts to make research data more widely available. In the United Kingdom, the Royal Society has outlined recommendations that address the issues surrounding data sharing within the scientific community. They believe research data should be accessible, intelligible, assessable, and usable.[26] The United Kingdom and Ireland have also committed to improve copyright law to facilitate text and data mining (TDM), as they recognize it as an important technique for extracting insight from existing data.[27] In June 2014, the UK parliament passed a law that allows TDM of copyrighted materials for noncommercial purposes, as long as sufficient reference is made to the original work. The US White House's Office of Science and Technology Policy in 2013 directed national research agencies to prepare to make federally funded research outputs publicly accessible, and since 2012, the European Commission has been pushing for broader access to government-funded research. Thus, many agencies directly involved in financing the research enterprise are advancing initiatives to encourage wider dissemination of scientific data.

Data as a critical materials R&D enabler

Data-intensive approaches are proving to be valuable for materials discovery, reducing the time needed to search for new materials with desirable properties by shortlisting promising candidates. Recently, stakeholders have shown an interest in promoting activities that encourage the use of modern data-centric approaches to solve materials problems.[28,29] One notable example is the US Materials Genome Initiative (MGI), launched in 2011, to accelerate materials development and commercialization.[30–32] Specifically, the MGI aims to halve the time and money needed to shepherd novel materials from the laboratory to widespread commercial deployment. In a similar vein, in 2015, the US Air Force Research Laboratory, NIST, and NSF launched a Materials Science and Engineering Data Challenge to encourage the use of publicly available data to discover or model new material properties.[29] The purpose of this challenge is to demonstrate that researchers can extract entirely novel insights from already-published materials data sets; the challenge submission period ended in March 2016, and the winners will present their results at the Materials Science and Technology 2016 Conference in Salt Lake City.

Using data to scale from the laboratory to manufacturing

The goal of accelerating materials development and deployment, as expressed by the US MGI, does not end with fundamental materials discovery. Reliably manufacturing those materials at scale is frequently an even greater challenge, and both industry and government see opportunities for data to accelerate materials scale-up to manufacturing. Numerous efforts to address this challenge have emerged, from the Advanced Manufacturing Plan 2.0 report by the President's Council on Science and Technology[33] to the National Network for Manufacturing Innovation (NNMI) funded by several US government agencies.[34] While manufacturing broadly has always been a data-intensive endeavor, given that data-historian software has been ubiquitous in the manufacturing environment to log process data for at least two decades, modern analytics can now crunch these data to identify more complex relationships between environmental conditions, processing parameters, product quality, materials wear and lifetime characteristics, and many other metrics. Optimizing these parameters promises to yield more product, at lower cost, using less energy.

While manufacturing faces some of the same challenges as the materials R&D community, it also suffers from a unique and severe constraint: lack of publicly available data. Proprietary data exist in abundance, but many producers of materials carefully guard their manufacturing data as trade secrets; so while the nascent open-access movement grows, the sharing of manufacturing data lags substantially. To combat this, consortia are forming around precompetitive research in efficient, smart manufacturing. Three examples of these are the Smart Manufacturing Leadership Coalition at the University of Texas, the Digital Manufacturing node of the NNMI system, and the recent call for proposals for a Manufacturing Innovation Institute on Smart Manufacturing.[35–37] Critically, each of these includes substantial industrial involvement and sponsorship, ensuring that the tools and methods developed within them are relevant to real-world manufacturing challenges.

Case studies: Demonstrating the potential of materials data and analytics

Materials Project

The Materials Project[38,39] was instituted at Lawrence Berkeley National Laboratory in 2011 with the goal to create an open, collaborative, and data-rich ecosystem for accelerated materials design. The Project uses high-performance computing within a sophisticated integrated infrastructure comprising an open-source python-based analysis library, pymatgen,[40] a document-based schema-less database, and automated open-source workflow software, Fireworks,[41,42] to determine structural, thermodynamic, electronic, and mechanical properties of over 65,000 inorganic compounds by means of high-throughput *ab initio* calculations. More compounds and properties (e.g., elastic tensors, band structures, dielectric tensors, x-ray diffraction, piezoelectric constants, etc.)[43,44] are being added on a daily basis. The Materials Project, and related data-driven *ab initio* screening efforts, have led to a number of advances in energy materials discovery.[45]

A series of web applications provide users with the capability to perform advanced searches and useful analyses (e.g., phase diagrams, reaction-energy computations, band-structure decomposition, novel structure prediction, Pourbaix diagrams).[38,40,46] The calculated results and analysis tools are freely disseminated to the public via a searchable online web application, and the data are easily accessed and downloaded through the first implemented Materials Application Programming Interface (Materials API).[14] A high-level interface to the Materials API has been built into the pymatgen analysis library that provides a powerful way for users to programmatically query and analyze large quantities of materials information. While most of the available data are computed and produced in-house, the Project recently launched MPComplete, which allows Project users to submit desired structures to be simulated within DFT and MPContribs,[47,48] a software framework within which users may upload external materials data—either computed or measured—and develop apps within the Project's infrastructure. Today, the Project has more than 18,000 registered users and attracts 300+ distinct users every day to the site, spanning industry, academia, and government.[38]

The Open Quantum Materials Database

The Open Quantum Materials Database (OQMD)[24,49] is a high-throughput database currently consisting of ~400,000 DFT total energy calculations of compounds from the ICSD and decorations of commonly occurring crystal structures. OQMD is open (without restrictions) and is online.[50,51] Users can (1) search for materials by composition, (2) create phase diagrams (T = 0K), (3) determine ground-state compositions, (4) determine whether equilibrium (any two-phase tie line) exists between any two phases, (5) visualize crystal structures, or (6) download the entire database for their own use. The OQMD has been used to perform high-throughput computational screening of many types of materials, such as structural metal alloys,[52] Li battery materials,[53] and high-efficiency nanostructured thermoelectrics.[54] Much of the software and tooling surrounding the OQMD is open source and available for anyone to use and build upon.

Expert-led database building from literature

The practice and effectiveness of aggregation of experimental data is exemplified in two widely used databases of crystal structures. The ICSD, hosted by the Fachinformationszentrum (FIZ) Karlsruhe in Germany contains over 180,000 entries on the crystal structures of minerals, metals, and other extended solid–inorganic compounds.[16] The older Cambridge Crystallography Data Centre in the United Kingdom compiles and distributes the Cambridge Structural Database (CSD), a repository of experimentally determined organic and metal–organic crystal structures that currently exceeds 800,000 entries.[55] Both of these databases owe their success to some combination of early and widespread adoption, encouragement from journal publishers, and the clarity and utility of the .cif crystal structure format.[17,56] A third example of a careful and useful compilation of structural data is the Protein DataBank.[57] In materials science, the recent proliferation of computationally generated databases of materials structures and some of their computed properties have all been rooted in the ICSD.

Similar searchable online databases of materials properties, particularly those related to functional materials, are not as readily available. For example, no repository exists even for something as simple as the magnetic or ferroic ordering temperatures of inorganic compounds. Perhaps the resource that comes closest to what is required is the Landolt–Börnstein handbook series, which dates to the late 19th century and is now available electronically at the for-pay SpringerMaterials database.

Going forward, it is clear that the impetus for the creation of such databases must be associated with journal-mandated

requirements for the deposition of relevant property data, appropriately curated and formatted, in precisely the same manner that is mandated for crystal structure information. Recognizing the need and utility of such databases, there have been recent attempts to physically mine the literature to better understand the landscape of thermoelectric materials[58] and lithium- and lithium-ion-battery materials (**Figure 4**).[59] The process involves gathering appropriate publications, deciding the key data in the publications, and then employing a combination of students and postdoctoral fellows to perform data extractions. Quantitative experimental and simulation results reported in publications must be physically entered into a text file, frequently through the process of manually digitizing plots in the publications by using freeware tools such as DataThief. At this stage, metadata such as the unit-cell volume of the compound being measured, the elemental abundance of the constituents, or the preparation method are also entered. Finally, the text file is read into web-based visualization suites using software such as HighCharts, which is freely available for use to academic, not-for-profit entities.

The data, once available for visualization, can be highly useful for further prediction,[60] including through machine learning.[61] The insight that can be gained simply by looking at the data, appropriately plotted, cannot be overstated. As an example, the previous exercise as applied to thermoelectric materials quickly illustrates the large regions of parameter space where searching for new high-performing materials would be futile.

Citrination materials-data analytics platform

Citrination[12] is a materials data platform that extracts new insights from large-scale materials data. The platform ingests messy materials data sources, such as papers, patents, or existing databases, and extracts clean, structured facts from these files (e.g., $T_{m,H_2O} = 0°C$, where T_{m,H_2O} is the melting temperature of water). The resulting data can be used to train machine learning models of materials behavior, which Citrine deploys as web apps to accelerate R&D, manufacturing, and sales efforts in the materials industry. Citrination also represents one of the world's largest collections of completely free and open materials data, usable by any researcher worldwide, with over three million records and counting.[12,62,63]

Two forms of data exist within Citrination: (1) structured data, which contain clearly defined and formatted data points, and (2) unstructured data, which include images, PDFs, and documents in other formats. The data represent a collection of information from a wide variety of sources, both experimental and computational, and have been added either by the company or by the site's users themselves. The platform supports a wide variety of materials metadata, using an underlying hierarchical data standard. This makes it easy for data to be understood, evaluated, and cited. The structured data is searchable and can be programmatically accessed using the site's API. Thus, Citrination makes it easier for researchers to access and analyze materials data at a scale that has not been previously possible.

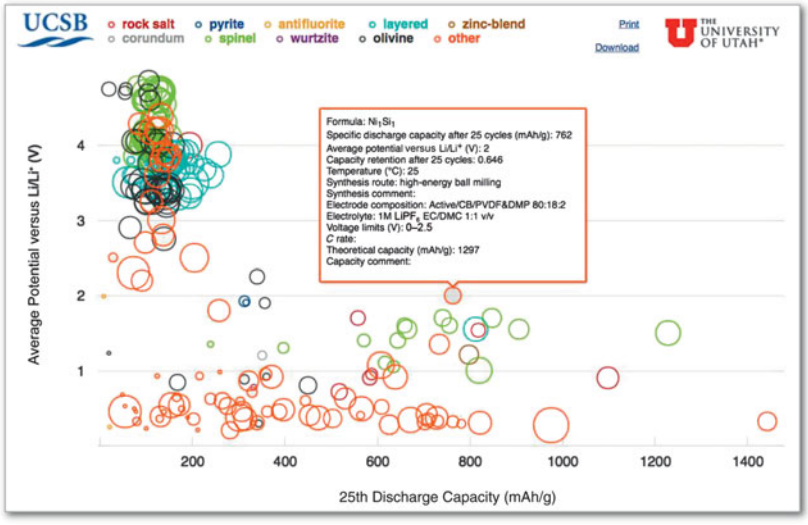


Figure 4. Screenshot of a battery data-mining resource,[72] displaying the average potential of the electrode as a function of the calculated discharge capacity after the 25th cycle. Such visualizations make it easier for materials scientists to identify important patterns in very large data sets. The symbol size is proportional to the percentage capacity that is retained after 25 cycles (larger is better), and the color indicates the crystal structure type of the active electrode material. Hovering the mouse over data points allows metadata to be read, and clicking on the points takes the reader to the original literature.[73] For thermoelectrics, see Reference 71. Note: CB, carbon black; PVDF&DMP, poly(vinylidene fluoride) & dimethyl phthalate; EC, ethylene carbonate; DMC, dimethyl carbonate; *C*, charge rate.

Key next steps

Nucleation around data standards

The amount of data in the materials community, as in many other areas of science and human endeavor, is increasing exponentially, making data management an urgent priority. To enable seamless data sharing and increase the usability of published data, data standards are required. Historically, efforts to create standards for materials data storage have focused on XML schemas. Over a decade ago, NIST developed MatML for storing materials data.[64] Other examples of XML schema, developed specifically for materials data storage, are MatDB and NMC-MatDB.[65] However, none of these has achieved wide adoption in the field.[66] We hypothesize that the greatest barrier to adoption to any proposed new data standard is that users do not see the value in adopting a standard that is not already widespread. "Seeding" a new data standard with a large quantity of useful materials data could help mitigate this problem.

JSON[67] has emerged alongside XML as a preferred file format for hierarchical data formatting, and JSON is now used for asynchronous

browser/server communication among other applications. The JSON format is similar to XML in many ways and provides good flexibility in terms of the scope of data that can be structured in this format. Most modern computer programming languages provide native parsing and generation of JSON files. This makes it a good candidate to be used to store materials data of varying types and for various purposes. The Materials Information File (MIF) is an open JSON-based file format, created by Citrine Informatics, to store diverse materials data ranging from standard entropy curves to x-ray diffraction patterns to DFT simulation outputs.[20] A number of key objects have already been defined to encompass common materials concepts (e.g., materials, measurements, phases, phase diagrams), and new objects can be readily created as needed.[20] The MIF format can thus evolve to accommodate materials data from every subdiscipline of materials. With this potential flexibility and ability to store various types of data, Citrine's goal is to build broad support for the MIF as a data standard within the materials community, and has been tackling the data standards problem by generating millions of MIF records and making them publicly available.

Data consolidation

Given the inherently diverse nature of materials data, consolidation is a major challenge. At present, most data repositories focus on a specific subset of materials data, and while this allows them to specialize, it means that it is often difficult to extract value across numerous data resources. In addition, different repositories structure data idiosyncratically, and the ease of access is highly variable; a single unified infrastructure would greatly streamline data analysis.

Data-intensive visualization

Data visualization is a key research activity that conveys information efficiently. Visualizations assist with highlighting patterns within data and identifying useful and important trends. Some classic examples of materials-related visualizations are Ashby plots, relating density and modulus across classes of materials, and Pettifor maps[68] of intermetallic crystal structures. We believe that improved systematic access to materials data will dramatically enhance the community's ability to generate such information-rich visualizations. **Figure 5** illustrates how a Python script can be used to programmatically generate an Ashby plot using public data from the Citrination platform.

Better software

Generally, materials informatics is only accessible to those who have deep experience in computer programming and data science. This is because the most currently available informatics tools rely on some degree of programming ability to analyze and manipulate data. As there are many materials scientists without such a background, it is vital that materials informatics are democratized in order to allow widespread access to the benefits of large-scale materials data analysis. For this to become a reality, software must be developed that will be intuitive and easy to use for materials experts who do not also possess training in computer science or data science. Such a goal requires the development of sophisticated user interfaces that expose the power of materials data without miring the user in jargon, arcane tuning parameters, or unfamiliar syntax. Such tools are just now emerging for widespread consumption by the materials community; examples include Citrine's thermoelectric materials recommendation engine,[61,69] Materials Project's Pourbaix diagram generator,[46] OQMD's grand canonical linear programming-based[70] phase stability evaluator, and the University of California, Santa Barbara Materials Research Laboratory and The University of Utah's thermoelectric data visualizer.[58]

Summary

This article discusses the challenges and opportunities associated with data-intensive materials research. With respect to

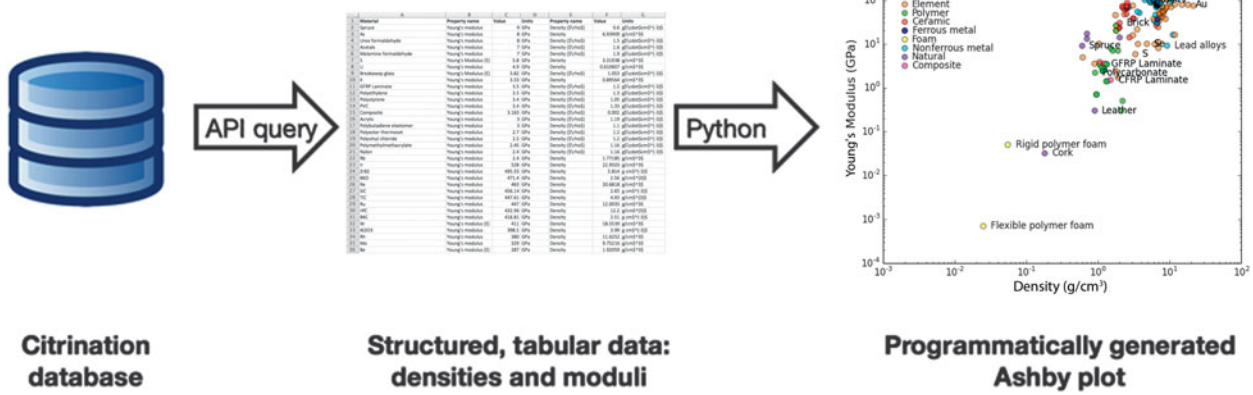


Figure 5. An Ashby plot auto-generated using structured data accessed via the Citrination API. Programmatic data access allows for the construction of key materials data visualizations, such as Ashby plots, without tens of hours of manual data entry. Note: AISI, American Iron and Steel Institute; GFRP, glass-fiber-reinforced plastic; CFRP, carbon-fiber-reinforced plastic.

its integration of large-scale data analysis, materials science lags behind other scientific disciplines; however, the situation is rapidly changing for the better. In particular, funding agencies, journal publishers, industry, government labs, and university researchers are aligning to make materials research data more accessible and useful to the community. We highlighted four specific efforts within the materials research community—Materials Project,[39] Open Quantum Materials Database,[50] expert database curation at the University of California, Santa Barbara and The University of Utah,[71] and the Citrination platform—[12] all of which involve aggregating, analyzing, or visualizing large quantities of materials research data at no cost to users. We expect these and related efforts to gather momentum as materials research continues to benefit from broader access to large data sets.

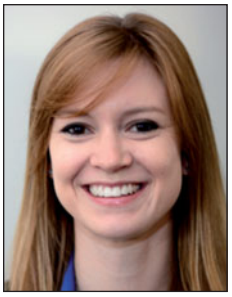
Acknowledgment

K.P. was supported by the Materials Project (Grant # EDCBEE), supported by the US Department of Energy Office of Science, Office of Basic Energy Sciences Department under Contract No. DE-AC02-05CH11231. R.S. thanks the National Science Foundation for support of this research through NSF-DMR 1121053 (MRSEC). C.W. gratefully acknowledges funding support from DOC NIST award 70NANB14H012 (CHiMaD).

References

1. K.M. Tolle, D.S.W. Tansley, A.J.G. Hey, *Proc. IEEE* **99**, 1334 (2011).
2. B. Ramsundar, S. Kearnes, P. Riley, D. Webster, D. Konerding, V. Pande, *Machine Learning* (2015), available at http://arxiv.org/abs/1502.02072.
3. M. Banerji, O. Lahav, C.J. Lintott, F.B. Abdalla, K. Schawinski, S.P. Bamford, D. Andreescu, P. Murray, M.J. Raddick, A. Slosar, A. Szalay, D. Thomas, J. Vandenberg, *Mon. Not. R. Astron. Soc.* **406**, 342 (2010).
4. The CMS Collaboration, *Nat. Phys.* **10**, 557 (2014).
5. A. White, *MRS Bull.* **37**, 715 (2012).
6. P.C. Zikopoulos, C. Eaton, D. deRoos, T. Deutsch, G. Lapis, *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data* (McGraw Hill, San Francisco, 2011).
7. J. Allison, D. Backman, L. Christodoulou, *JOM* **58**, 25 (2006).
8. Scifinder, https://scifinder.cas.org.
9. http://materials.springer.com/welcome.
10. T.N. Bhat, L.M. Bartolo, U.R. Kattner, C.E. Campbell, J.T. Elliott, *JOM* **67**, 1866 (2015).
11. NIST Materials Genome Initiative, "Materials Data Curation System," available at https://mgi.nist.gov/materials-data-curation-system.
12. Citrine Informatics, Citrination, available at https://citrination.com.
13. R.H. Taylor, F. Rose, C. Toher, O. Levy, K. Yang, M.B. Nardelli, S. Curtarolo, *Comput. Mater. Sci.* **93**, 178 (2014).
14. S.P. Ong, S. Cholia, A. Jain, M. Brafman, D. Gunter, G. Ceder, K.A. Persson, *Comput. Mater. Sci.* **97**, 209 (2015).
15. National Center for Biotechnology Information (NCBI) Genome, available at http://www.ncbi.nlm.nih.gov/genome.
16. A. Belkly, M. Helderman, V.L. Karen, P. Ulkch, *Acta Crystallogr. B* **58**, 364 (2002).
17. S.R. Hall, F.H. Allen, I.D. Brown, *Acta Crystallogr. A Found. Crystallogr.* **47**, 655 (1991).
18. Dryad Digital Repository, http://datadryad.org.
19. Figshare, http://figshare.com.
20. Citrine Informatics, "MIF Schema," available at http://citrineinformatics.github.io/mif-documentation.
21. C.L. Borgman, *J. Am. Soc. Inf. Sci. Technol.* **63**, 1059 (2012).
22. National Science Foundation Directorate for Engineering, "NSF ENG Data Management Plan Requirements," available at https://www.nsf.gov/eng/general/dmp.jsp.
23. US Department of Energy, "Statement on Digital Data Management," available at http://science.energy.gov/funding-opportunities/digital-data-management.
24. J.E. Saal, S. Kirklin, M. Aykol, B. Meredig, C. Wolverton, *JOM* **65**, 1501 (2013).
25. M. Chui, J. Manyika, J. Bughin, R. Dobbs, C. Roxburgh, H. Sarrazin, G. Sands, M. Westergren, "The Social Economy: Unlocking Value and Productivity through Social Technologies" (McKinsey Global Institute, 2012).
26. *Science as an Open Enterprise* (The Royal Society, 2012).
27. *Standardization in the Area of Innovation and Technological Development, Notably in the Field of Text and Data Mining: Report from the Expert Group* (European Commission, 2014).
28. J.P. Holdren, *Memorandum for the Heads of Executive Departments and Agencies: Increasing Access to the Results of Federally Funded Scientific Research* (Executive Office of the President: Office of Science and Technology Policy, February 22, 2013).
29. Air Force Research Lab in partnership with National Institute of Standards and Technology and the National Science Foundation, "Materials Science and Engineering Data Challenge," available at https://www.challenge.gov/challenge/materials-science-and-engineering-data-challenge.
30. T. Kalil, C. Wadia, "Materials Genome Initiative for Global Competitiveness" (2011), available at https://www.whitehouse.gov/sites/default/files/microsites/ostp/materials_genome_initiative-final.pdf.
31. L. Kaufman, J. Ågren, *Scr. Mater.* **70**, 3 (2014).
32. C.M. Simon, J. Kim, D.A. Gomez-Gualdron, J.S. Camp, Y.G. Chung, R.L. Martin, R. Mercado, M.W. Deem, D. Gunter, M. Haranczyk, D.S. Sholl, R.Q. Snurr, B. Smit, *Energy Environ. Sci.*, **8**, 1190 (2015).
33. President's Council of Advisors on Science and Technology, "Report to the President, Accelerating US Advanced Manufacturing" (2014), available at https://www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/amp20_report_final.pdf.
34. https://www.manufacturing.gov/nnmi.
35. The University of Texas at Austin Cockrell School of Engineering: McKetta Department of Chemical Engineering, "Smart Manufacturing Leadership Coalition Wins DOE Funding," available at http://www.che.utexas.edu/2013/04/05/smart-manufacturing-leadership-coalition-wins-doe-funding. News release, April 5, 2013.
36. The White House, "President Obama Announces Two New Public-Private Manufacturing Innovation Institutes and Launches the First of Four New Manufacturing Innovation Institute Competitions," available at https://www.whitehouse.gov/the-press-office/2014/02/25/president-obama-announces-two-new-public-private-manufacturing-innovatio. News release, February 25, 2014.
37. DOE EERE, "Manufacturing Innovation Institute for Smart Manufacturing: advanced Sensors, Controls, Platforms, and Modeling for Manufacturing (15AD)."
38. 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, *APL Mater.* **1**, 011002 (2013).
39. http://www.materialsproject.org.
40. S.P. Ong, W.D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V.L. Chevrier, K.A. Persson, G. Ceder, *Comput. Mater. Sci.* **68**, 314 (2013).
41. A. Jain, G. Hautier, C.J. Moore, S.P. Ong, C.C. Fischer, T. Mueller, K.A. Persson, G. Ceder, *Comput. Mater. Sci.* **50**, 2295 (2011).
42. A. Jain, S.P. Ong, W. Chen, B. Medasani, X. Qu, M. Kocher, M. Brafman, G. Petretto, G.-M. Rignanese, G. Hautier, D. Gunter, K.A. Persson, *Concurr. Comput.* **27**, 5037 (2015), doi:10.1002/cpe.3505.
43. M. de Jong, W. Chen, T. Angsten, A. Jain, R. Notestine, A. Gamst, M. Sluiter, C.K. Ande, S. van der Zwaag, J.J. Plata, C. Toher, S. Curtarolo, G. Ceder, K.A. Persson, M. Asta, *Sci. Data* **2**, 150009 (2015).
44. M. de Jong, W. Chen, H. Geerlings, M. Asta, K.A. Persson, *Sci. Data* **2**, 150053 (2015).
45. A. Jain, A.Y. Shin, K.A. Persson, *Nat. Rev. Mater.* **1**, 15004 (2016).
46. K.A. Persson, B. Waldwick, P. Lazic, G. Ceder, *Phys. Rev. B Condens. Matter* **85**, 1 (2012).
47. P. Huck, D. Gunter, S. Cholia, D. Winston, A.T. N'Diaye, K. Persson, *Concurr. Comput.* 1–12 (2015), doi:10.1002/cpe.3698.
48. P. Huck, A. Jain, D. Gunter, D. Winston, K. Persson, presented at the *2015 IEEE 11th International Conference on e-Science*, Munich, Germany, August 31–September 4, 2015, available at http://dx.doi.org/10.1109/eScience.2015.75.
49. S. Kirklin, J.E. Saal, B. Meredig, A. Thompson, J.W. Doak, M. Aykol, S. Rühl, C. Wolverton, *NPJ Comput. Mater.* **1**, 15010 (2015), doi:10.1038/npjcompumats.2015.10.
50. http://oqmd.org.
51. Twitter @TheOQMD.
52. S. Kirklin, J.E. Saal, V.I. Hegde, C. Wolverton, *Acta Mater.* **102**, 125 (2016).
53. S. Kirklin, B. Meredig, C. Wolverton, *Adv. Energy Mater.* **3**, 252 (2013).
54. L.-D. Zhao, S. Hao, S.-H. Lo, C.-I. Wu, X. Zhou, Y. Lee, H. Li, K. Biswas, T.P. Hogan, C. Uher, C. Wolverton, V.P. Dravid, M.G. Kanatzidis, *J. Am. Chem. Soc.* **135**, 7364 (2013).
55. F.H. Allen, *Acta Crystallogr. B* **58**, 380 (2002).
56. I.D. Brown, B. McMahon, *Acta Crystallogr. B* **58**, 317 (2002).
57. H.M. Berman, *Nucleic Acids Res.* **28**, 235 (2000).

58. M.W. Gaultois, T.D. Sparks, C.K.H. Borg, R. Seshadri, W.D. Bonificio, D.R. Clarke, *Chem. Mater.* **25**, 2911 (2013).
59. T. Sparks, L. Ghadbeigi, J.K. Harada, B. Lettiere, *Energy Environ. Sci.* **8**, 1640 (2015).
60. M.W. Gaultois, T.D. Sparks, *Appl. Phys. Lett.* **104**, 113906 (2014).
61. T.D. Sparks, M.W. Gaultois, A. Oliynyk, J. Brgoch, B. Meredig, *Scr. Mater.* **111**, 10 (2015).
62. https://www.whitehouse.gov/blog/2015/02/06/its-time-open-materials-science-data.
63. https://www.whitehouse.gov/blog/2015/07/20/unleashing-digital-data-accelerate-materials-discovery-development-and-deployment-0.
64. C. Sturrock, E. Begley, J. Kaufman, "MatML-Materials Markup Language Workshop Report" (National Institute of Standards and Technology, 2001).
65. T.S.P. Austin, H.H. Over, *Data Sci. J.* **11**, ASMD11 (2012).
66. X. Zhang, C. Zhao, X. Wang, *Comput. Ind.* **73**, 8 (2015).
67. https://tools.ietf.org/html/rfc7159.
68. D.G. Pettifor, *Solid State Commun.* **51**, 31 (1984).
69. M.W. Gaultois, A.O. Oliynyk, A. Mar, T.D. Sparks, G.J. Mulholland, B. Meredig, "A Recommendation Engine for Suggesting Unexpected Thermoelectric Chemistries," 1–7 (2015). [cond-mat.mtrl-sci] 5 Jan 2016; arXiv:1502.07635.
70. A.R. Akbarzadeh, V. Ozoliņš, C. Wolverton, *Adv. Mater.* **19**, 3233 (2007).
71. www.mrl.ucsb.edu:8080/datamine/thermoelectric.jsp.
72. L. Ghadbeigi, J.K. Harada, B.R. Lettiere, T.D. Sparks, *Energy Environ. Sci.* **8**, 1640 (2015).
73. http://tomcat.eng.utah.edu/sparks/battery.jsp. □



Joanne Hill studied chemistry and materials science at the University of Cape Town, South Africa. She now works as a data engineer at Citrine Informatics, focusing on automated data extraction and ingestion with the goal of making materials data more accessible and useful to everyone. Hill can be reached by email at jo@citrine.io.

Gregory Mulholland is the Chief Operating Officer at Citrine Informatics, the data platform for the physical world. He works with partners along the materials value chain to use state-of-the-art data science techniques to accelerate advanced materials discovery and deployment. He earned an MBA degree from Stanford University's Graduate School of Business, MPhil degree in materials science from the University of Cambridge, and a BS degree in electrical and computer engineering from North Carolina State University. Mulholland can be reached by email at greg@citrine.io.

Kristin Persson studies the physics and chemistry of materials using atomistic computational methods and high-performance computing technology, particularly for clean-energy production and storage applications. She is director of the Materials Project at Lawrence Berkeley National Laboratory, a multi-institution, multinational effort to compute the properties of all inorganic materials and provide the data and associated analysis algorithms to researchers free of charge. This project has been used to design novel photocatalysts, multivalent battery electrode materials, Li-ion battery electrode materials, and electrolytes for beyond-Li energy storage solutions. Persson can be reached by email at kapersson@lbl.gov.

Ram Seshadri is a professor in the Department of Chemistry and Biochemistry at the University of California, Santa Barbara (UCSB), and director of the UCSB Materials Research Laboratory: an NSF MRSEC. His group is active in the area of functional materials for energy conversion and storage, in addition to researching fundamental aspects of electronic, magnetic, and polar materials. He is also interested in materials education and outreach. Seshadri can be reached by email at seshadri@mrl.ucsb.edu.

Christopher Wolverton is a professor of materials science and engineering at Northwestern University. Before joining the faculty, he worked at the Research and Innovation Center at Ford Motor Company, where he was group leader for the Hydrogen Storage and Nanoscale Modeling Group. His research interests include computational studies of a variety of energy-efficient and environmentally friendly materials via first-principles atomistic calculations, high-throughput and data mining tools to accelerate materials discovery, and "multi-scale" methodologies for linking atomistic and microstructural scales. Wolverton can be reached by email at c-wolverton@northwestern.edu.

Bryce Meredig is a co-founder and the Chief Executive Officer of Citrine Informatics. His goal is to build software that enables every materials researcher to harness the world's entire corpus of materials data to accelerate their work. He earned a PhD degree in materials science from Northwestern University, an MBA degree from Stanford University, and a BAS degree in materials science and German from Stanford University. In addition to his role at Citrine, he is a consulting assistant professor of materials science at Stanford University. Meredig can be reached by email at bryce@citrine.io.