






Data-driven glass/ceramic science research: Insights from the glass and ceramic and data science/informatics communities

Eileen De Guire¹  | Laura Bartolo² | Ross Brindle³ | Ram Devanathan⁴ | Elizabeth C. Dickey⁵  | Justin Fessler⁶ | Roger H. French⁷  | Ulrich Fotheringham⁸ | Martin Harmer⁹  | Edgar Lara-Curzio¹⁰ | Sarah Lichtner³ | Emmanuel Maillet¹¹ | John Mauro¹²  | Mark Mecklenborg¹ | Bryce Meredig¹³ | Krishna Rajan¹⁴ | Jeffrey Rickman⁹ | Susan Sinnott¹² | Charlie Spahr¹ | Changwon Suh³ | Adama Tandia¹⁵ | Logan Ward¹⁶ | Rick Weber¹⁷

¹The American Ceramic Society, Westerville, Ohio

²Northwestern-Argonne Institute for Science and Engineering, Northwestern University, Evanston, Illinois

³Nexight Group, Silver Spring, Maryland

⁴Energy and Environment Directorate, Pacific Northwest National Laboratory, Richland, Washington

⁵Department of Materials Science and Engineering, North Carolina State University, Raleigh, North Carolina

⁶IBM Watson, Arlington, Virginia

⁷Department of Materials Science and Engineering, Case Western Reserve University, Cleveland, Ohio

⁸SCHOTT AG, Mainz, Germany

⁹Department of Materials Science and Engineering, Lehigh University, Bethlehem, Pennsylvania

¹⁰Mechanical Properties and Mechanics Group, Oak Ridge National Laboratory, Oak Ridge, Tenn

¹¹Materials Science and Engineering, GE Global Research, Niskayuna, New York

¹²Department of Materials Science and Engineering, The Pennsylvania State University, University Park, Pennsylvania

¹³Citrine Informatics, Redwood City, California

¹⁴Department of Materials Design and Innovation, University at Buffalo, Buffalo, New York

¹⁵Corning Incorporated, Corning, New York

¹⁶Globus Labs, University of Chicago, Chicago, Illinois

¹⁷Materials Development Inc., Arlington Heights, Illinois

Correspondence

Eileen De Guire, The American Ceramic Society, 550 Polaris Pkwy, Ste 510 Westerville, OH 43082.
Email: edeguire@ceramics.org

Funding information

NIST AMTech Award No. 70NANB15H073.

Abstract

Data-driven science and technology have helped achieve meaningful technological advancements in areas such as materials/drug discovery and health care, but efforts to apply high-end data science algorithms to the areas of glass and ceramics are still limited. Many glass and ceramic researchers are interested in enhancing their work by using more data and data analytics to develop better functional materials more efficiently. Simultaneously, the data science community is looking for a way to access materials data resources to test and validate their advanced computational learning

Justin Fessler is now with Salesforce, Washington, DC.

Charlie Spahr retired from ACerS 12/31/18.

algorithms. To address this issue, The American Ceramic Society (ACerS) convened a Glass and Ceramic Data Science Workshop in February 2018, sponsored by the National Institute for Standards and Technology (NIST) Advanced Manufacturing Technologies (AMTech) program. The workshop brought together a select group of leaders in the data science, informatics, and glass and ceramics communities, ACerS, and Nexight Group to identify the greatest opportunities and mechanisms for facilitating increased collaboration and coordination between these communities. This article summarizes workshop discussions about the current challenges that limit interactions and collaboration between the glass and ceramic and data science communities, opportunities for a coordinated approach that leverages existing knowledge in both communities, and a clear path toward the enhanced use of data science technologies for functional glass and ceramic research and development.

KEYWORDS

glass, modeling/model, simulation

1 | INTRODUCTION

In 2015–2017, The American Ceramic Society (ACerS) led activities to plan for the launch of the Functional Glass Manufacturing Innovation Consortium (FGMIC) and the development of *Driving Functional Glass Manufacturing Innovation: A Technology Roadmap to 2025*.¹ Supported with funding from the National Institute of Standards and Technology (NIST) Advanced Manufacturing Technology (AMTech) program, the roadmap identifies challenges that currently constrain functional glass manufacturing and provides pathways for developing, improving, and implementing advanced technologies over the next 10 years that are critical to industry innovation. The FGMIC roadmap captured a wide swath of technologies needed for advanced manufacturing of functional glasses that fell into five broad categories: Characterizing materials structures and properties, Modeling and simulating performance, Optimizing manufacturing processes, Data infrastructure and best practices, and Workforce development and coordination.

The need for increased coordination between the glass community and the data science and informatics community emerged as a key theme throughout the roadmap that cut across all categories. Based on roadmap findings, key benefits to the application of data science to glass include:

1. **Tailored properties:** By generating structure-property data from characterizations coupled with modeling and existing glass databases, manufacturers will gain an increased understanding of the structural characteristics of glasses, enabling them to more accurately develop glasses with improved functionality, more consistent and reliable performance, and reduced costs.

2. **Development of new glass compositions:** Using well-established glass theory and simulation, coupled with data-driven glass informatics for information sifting or screening (ie, the driving concept of the Materials Genome Initiative [MGI]), can help narrow down the essentially infinite glass composition space.
3. **Improved process control:** The sophisticated combination of heterogeneous data from varied theoretical calculations, empirical modeling, and processing characterizations will help companies pinpoint opportunities for maximizing process efficiency, which can reduce product development time and minimize manufacturing costs.

Recognition of the convergence of data science with glass science during the FGMIC project brought into focus the timeliness of the data opportunity for glass science and manufacturing. Major funding agencies have established funding initiatives to address data science in many fields, including materials science. For example, searching the term “artificial intelligence” on www.grants.gov, the United States government repository for funding opportunities in May 2019 yielded 90 announcements from agencies including those encompassing the materials science community such as the National Science Foundation, the Department of Defense (Air Force Research Laboratory, the Office of Naval Research, DARPA, Army Research Laboratory), NASA, and the Department of Energy.

Additionally, ACerS has observed similar paradigms, opportunities, and challenges for data-driven discovery for ceramic science as evidenced by symposia at ACerS conferences on topics such as computational materials science and a “ceramic genome.” Staff discussions with members and ACerS leaders, as well as informal observation of topics appearing in ACerS peer-review journals, point to an increased

use of data science for ceramic materials science similar to the glass community. Indeed, since the workshop reported on here, similar workshops have taken place to address issues related to the convergence of ceramic science and data challenges. For example, in June 2018, Lehigh University faculty, N.B. Carlisle and J.M. Rickman, organized a 2-day workshop on the convergence of materials research and multi-sensory data science to develop a roadmap for understanding human interaction with large amounts of data for materials discovery.² The potential of data-driven, computational learning approaches for accelerating the discovery of cost-effective advanced materials has been gaining momentum through four major technical thrusts: Materials Informatics (MI) since the early 2000s, Integrated Computational Materials Engineering (ICME) since 2008, MGI since 2011, and High Performance Computing for Manufacturing (HPC4MFG) since 2015.^{3–6} Several national initiatives—including the National Nanotechnology Initiative (NNI)⁷ and the National Strategic Computing Initiative (NSCI)⁸—share the data-driven research and development (R&D) vision and have adopted strategic objectives that acknowledge the need to maintain an intimate connection with the key driving principles of the MGI. In addition, the Networking and Information Technology R&D (NITRD) Program released the *National Artificial Intelligence (AI) Strategic Plan* that outlined research priorities closely matching those identified in the *MGI Strategic Plan*.⁹

Several workshop events related to AI- or data-driven materials science include the DOE Advanced Manufacturing Office (AMO)'s 2017 “Workshop on AI Applied to Materials Discovery and Design,”¹⁰ the “2nd Novel Materials Discovery (NOMAD) Industry Workshop,”¹¹ “CECAM Big-Data driven Materials Science,”¹² and the “CECAM workshop on Machine Learning in Atomistic Simulations.”¹³ These activities, however, were not exclusively designed for a diverse spectrum of participants across data science, data-driven materials science (ie, bilingual experts in data/materials), and materials science (especially glass and ceramic science). As a professional society, ACerS wanted to identify opportunities where it could help these communities efficiently work together and advance the ability of both communities to harness the potential of big data in materials science, especially glasses and ceramics.

To address the need for collaboration between the data science and glass and ceramic communities, ACerS held a 1-day facilitated workshop in February 2018 in Silver Spring, Maryland. The workshop brought together a select group of more than 20 experts in the data science, informatics, and ceramic/glass communities from glass and ceramic manufacturing companies, universities, national laboratories, government agencies, and ACerS to discuss the greatest opportunities and mechanisms for facilitating increased

collaboration and coordination between these communities. Care was taken to evenly balance expertise in data science/informatics and ceramic/glass science to avoid “silos,” encourage diversity of thought, and uncover challenges and opportunities that tend to be invisible within discrete communities. Several participants' expertise qualified them as “dual citizen experts” in data science as well as materials science.

This article summarizes the results of this workshop, including:

1. A description of current challenges that limit interactions and collaboration between the glass and ceramic and data science communities,
2. Prioritized key challenges, gaps, and needs that cross-cut the glass and ceramic sciences and the data science communities,
3. Priority opportunities for collaborative activity that leverage the expertise of both the glass and ceramic sciences and the data science communities.

2 | WORKSHOP PROCESS

The ultimate objectives of the workshop were to achieve:

1. A mutual understanding of the crosscutting challenges and opportunities for collaboration (eg, conferences, training, content needs) across the glass and ceramics science and data science communities,
2. An understanding of the role of individual stakeholders from the glass and ceramic and data science communities in facilitating further activity and investment,
3. An overview of the value proposition and next steps for pursuing top-priority opportunities identified through the workshop process

To provide all workshop participants with a common baseline for discussion, the workshop began with brief presentations on the current state of data-driven materials research and development, data science and informatics, and data science applied to the field of glass and ceramics. These presentations, which were provided by Bryce Meredig, Roger H. French, and John C. Mauro, are summarized in the Current State-of-the-Art of Data-Driven Glass and Ceramic Science section of this article.

The remainder of the workshop focused on collaborative brainstorming and prioritization in a process facilitated by Nexight Group. After brainstorming grand challenges and gaps in the implementation of data science for the data-driven glass and ceramics landscape, each participant voted on the top challenges they believed ACerS is

well-positioned to help address. The results of this voting exercise were discussed, leading to the identification of four priority areas of ACerS-led opportunities for improved glass and ceramics and data science community coordination, which have been summarized in this article as (a) Bridging Physics-based, Empirical, and Data-Driven Models; (b) Creating Collaborative Data Infrastructure; (c) Boosting Interdisciplinary Education; and (d) Enhancing Data Usability and Visualization. While these discussions focused on opportunities that could be led by ACerS, the opportunities identified reflect the needs at the intersection of the data/informatics and the ceramic/glass communities and could be implemented by another professional society or other coordinating body.

In breakout groups, the participants defined activities for each opportunity, as well as the associated activity value proposition, scope, format, metrics of success, and next steps. All participants had an opportunity to review and refine the content developed by the other three breakout groups. The challenges and opportunities identified during the Glass and Ceramic Data Science Workshop discussions serve as the basis for this article.

3 | CURRENT STATE-OF-THE-ART OF DATA-DRIVEN GLASS AND CERAMIC SCIENCE

Integrating today's emerging data science and technology capabilities within the scope of glass and ceramic science would enlarge the materials and manufacturing R&D network while allowing the data science/informatics community to benefit from access to materials and manufacturing data resources for testing and validating advanced AI algorithms.

The materials community has started to demonstrate how advanced AI technologies can be used to rapidly and successfully solve materials and manufacturing problems. In the community, there are currently efforts to assess how to take full advantage of high-end AI tools and algorithms with respect to the following three types of machine intelligence:

1. Perception intelligence to obtain insight into the strategy for energy materials discovery and optimal design by analyzing complex materials/manufacturing data¹⁴
2. Prediction intelligence to predict materials behaviors by analyzing data such as high-dimensional microstructural or topographical images of materials^{15,16}
3. Prescription intelligence to enable automated root-cause analyses or real-time (ie, in-situ or operando) data monitoring for enhanced manufacturing processes

Nevertheless, researchers in the ceramic and glass science community and the data science community still need a common language for working together, a common understanding of the opportunities and limitations of their disciplines, and the opportunity to interact with leading specialists in their fields. What compounds the problem even further is the volume of scientific literature published on an annual basis within and across disciplines, and the human inability to comprehensively consume all of this literature and understand which research and information is pertinent within the field, while continuing to conduct ongoing research—thus growing the need for more advanced AI within R&D.

Selected experts kicked off the workshop with presentations on data-driven materials R&D and the challenges and benefits of using massive amounts of data.

3.1 | AI in materials R&D

In his opening presentation, Bryce Meredig emphasized that AI algorithms should be used to plan future experiments (ie, process development), not to attempt to discover miracle materials in one shot. An example of AI-driven experimental design is Balachandran et al's work on high-temperature ferroelectrics, wherein the goal was to identify compositions with high Curie temperatures that also should crystallize in the perovskite phase.¹⁷

According to Meredig, AI is most credible to domain experts when it is interpretable, not a black box. AI must also be able to handle typical issues in materials design, such as small, sparse materials data; uncertainty; and high dimensionality of many experimental design spaces, while leveraging the laws of physics. He presented two examples to illustrate use of AI in the context of materials discovery (below and sidebar).

The first example is Hutchinson et al's approaches that demonstrate the power of transfer learning to overcome data scarcity, which is a common issue in materials data.¹⁸ Transfer learning uses information from one dataset to inform a model on another to continuously bridge sparse data while preserving the contextual differences in the measurement. In predicting experimental band gap energies with the assistance of density functional theory-based band gap energies, authors showed that transfer learning results with only 67 experimental band gap energies are more accurate than generic materials informatics with 337 experimental measurements. As a second example, Ling et al tackled the optimization of composition and processing for desired properties of materials by using sequential learning (also called optimal experimental design or active learning) to fit data-driven models that incorporate uncertainty information in the analysis of experimental data.¹⁹

Machine learning: A data-driven approach for discovering oxygen ion conductors

Bryce Meredig and Sossina M. Haile

Oxygen ion conductors represent one of the most widely studied and technologically useful classes of electrochemical materials, finding applications ranging from oxygen sensors to fuel cells to gas separators.^{20–23} The rich research literature on these materials makes them an attractive candidate for machine learning (ML)-driven discovery, as ML is most effective if large datasets are available to inform data-driven models.

While the research literature provides extensive experimental measurements on a wide variety of oxygen ion conductors, a typical publication represents an *unstructured* source of data—its natural language text is comprehensible to human readers, but not ML algorithms. Thus, as a prerequisite to applying ML to oxygen ion conductors, we created the largest (to the best of our knowledge) publicly accessible *structured* database of oxygen ion conductor properties by extracting materials data from hundreds of published research articles. The resulting database contains 3,039 Physical Information File²⁴ records on 2340 different materials, and is freely available on the Open Citrination²⁵ platform at <https://citrination.com/datasets/151085>.

Using the aforementioned database as a source of training data, we created ML models to predict the $\ln(\sigma_0)$ prefactor and activation energy for compounds of (in principle) arbitrary chemical composition; the predictive accuracy of these two models is illustrated in Figure 1. Of course, factors beyond composition such as microstructure also influence ionic conductivity measurements, and these effects cannot be fully captured in composition-only models. Nonetheless, a key advantage of ML models for materials screening is that they can evaluate candidates ~6 orders of magnitude faster than density functional theory calculations.²⁶ We employed our ML models to screen a large composition space and downselect to a short list of promising oxygen ion conductors, some of which have demonstrated compelling performance in subsequent laboratory investigations. Full details of this work will be described in a forthcoming publication.²⁸

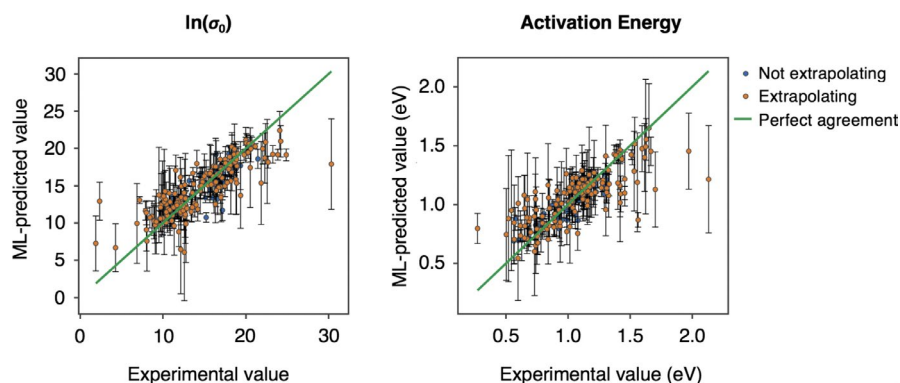


FIGURE 1 Predictive accuracy of ML models trained to estimate $\ln(\sigma_0)$ and activation energy for oxygen diffusion in oxide compounds. Both plots show results from three-fold cross-validation, and include unique uncertainty estimates (displayed as error bars) for every point.²⁷

3.2 | Data science: Informatics and analytics

While it is typical to deal with small data in materials science, some particular research areas require the consideration of big data analytics. Roger H. French shared recent data-driven R&D efforts for real-world big data generated from multiple heterogeneous photovoltaic (PV) test sites.²⁹ Handling and analyzing the massive amount of time-series data from these sites is critical for understanding PV system degradation and lifetime performance (ie, the performance loss rate).³⁰

Combined with a nonrelational data warehouse (wherein data storage is optimized for the specific requirements of the type of data being stored, rather than structured in a typical tabular schema), his research group uses Apache Hadoop³¹/Hbase/Spark³² architecture for distributed and high-performance (ie, petabyte and petaflop) computing (Dist/HPC) for three different types of modeling: predictive modeling to identify stress and response relationships;³³ network/inferential modeling for stress-mechanism-response relationships;³⁴ and machine learning modeling for image processing and

classification of PV cell degradation types from PV module electroluminescent (EL) images.³⁵

French highlighted the following key challenges in applying data analytics to engineering fields:

1. Open data science tools (ie, open source, R, Python, or Git and shared Git repositories such as Github, BitBucket, or GitLab)^{36–39} to reproduce research by code version control and collaboration,
2. In-place data analytics such as distributed R-analytics in Hadoop/Spark Hadoop Distributed File System (HDFS),
3. Data integration (eg, laboratory data such as spectral or image data with time-series data generated from PV test sites or accelerated weathering exposure experiments),
4. Graph-based network modeling of complex system interactions.

Data from deployed real-world materials systems are a new complement to laboratory experiments: Enabling degradation studies of PV modules and power plants under realistic conditions and lifetimes

Roger H. French, Alan J. Curran, Ahmad M. Karimi

Photovoltaic (PV) power plants, using crystalline silicon (c-Si) PV modules are products that come with 25-year product warranties and are exposed to harsh outdoor conditions with little maintenance in this time period. One of the grand challenges for PV has been achieving these long lifetimes while extending the lifetime of PV modules to 50 years. Achieving this requires an understanding of the multiple simultaneous and sequential degradation mechanisms that are active. Lifetime and degradation science⁴⁰ is an approach to these long lifetime, complex system, degradation pathway problems, that adds data science and big data analytics approaches to the more traditional, hypothesis-driven, laboratory-based research methods.

Using statistical and machine learning methods of data science has had good success in many areas such as materials design.⁴¹ For long lifetime materials science problems, we introduce an epidemiological approach, studying larger populations of real-world systems, spanning a wider range of exposure conditions (stressors), to enable data-driven modeling and statistically significant scientific findings, instead of just observational results.

This requires access to large volumes of data; in our case time series PV power plant data from 5000 PV inverters located at 787 PV power plant sites, distributed globally across 13 Köppen-Geiger (KG) climate zones^{41,42} with detailed weather, and insolation data and metadata (information describing the data) about the PV systems' performance. We also need to draw upon both distributed computing to analyze "petabyte scale" datasets, along with high performance computing for "petaflop" computations on the data.

In the SDLE Research Center at Case Western Reserve University, we have ingested the 1- to 15-minute interval time-series power datasets from the PV power plant sites and their inverters under a data use agreement signed by 14 companies. We de-identify and anonymize their systems, followed by data ingestion into our Hadoop cluster, and merge them into one comprehensive dataset.²⁹ In this manner we can present any company's results in comparison to the rest (the "bench"), and we can compare performance to climate zone, PV module or PV inverter brands and models, and to PV cell types and module materials.

CRADLE (Common Research Analytics and Data Lifecycle Environment), our computational environment for data science and analytics (Figure 2) is based on the open source distributed computing platform Hadoop, with Hbase, a NoSQL (or nonrelational) database, and Spark (in-memory distributed processing), which is integrated into CWRU's high-performance computing cluster. It provides the ability to ingest, annotate, query, assemble, search, explore, and develop statistical and machine learning models using Petabyte scale volumes of real-world data. This architecture has been successfully employed for multiple DOE SETO,⁴³ ARPA-E,⁴⁴ ARO,⁴⁵ and industry funded research projects.

Quantifying the performance loss rate (PLR) of a PV power plant, which for a modern PV module is about 0.5%/year, is an analytical challenge because the output of the PV power plant varies continuously throughout the day from 0 kW at night to its maximum output at solar noon. To determine a small, annual change, from systems with such dynamic daily behavior, requires detailed data-driven modeling. The XbX method for PLR determination⁴⁶

determines the degradation rate (R_d) of the PV module once other system factors have been considered. The PV inverter time-series data are segmented into time segments of length X , either daily, weekly or monthly, and for each segment, a β predictive model is created to determine the dependence of the power output on the weather conditions. The next data-driven model, the ξ longitudinal temporal model, determines a fit for the predicted power over time and can either be a linear or a piecewise fitted change point model, and the yearly PLR is calculated from the slope, or slopes, of the model. The final inferential step is the γ model, which is used to compare the available metadata across many inverters and, using variable selection and rank-ordering, determines the most significant variables affecting the PLR.

Using industry data sources gives us access to large volumes of data, from many systems and PV module brands, over long time periods and diverse climate zones. For an example, spatio-temporal PLR result, consider 7- to 12-year long, 15-minute interval, time-series power from 353 PV inverters. Computing the PLR for one system typically runs for 2 hours on an HPC compute node, but with distributed and high-performance computing, all 353 systems can be analyzed in 4 hours. Using variable selection in the γ model, we found (Figure 3) that the first rank-ordered predictor of PLR for these systems is the PV module brand, demonstrating that the detailed materials choices in the c-Si PV cells and the encapsulation and packaging materials of the PV module play a critical role in lifetime performance.

These kinds of unbiased, data-driven analytics, now possible using data science methodologies combined with distributed and high-performance computing on tens and hundreds of terabytes of real-world data would have been unthinkable 10 years ago. These new capabilities represent a new front in our scientific studies of critically important and complex materials systems and a complementary tool to add to our traditional approaches to materials science problems.

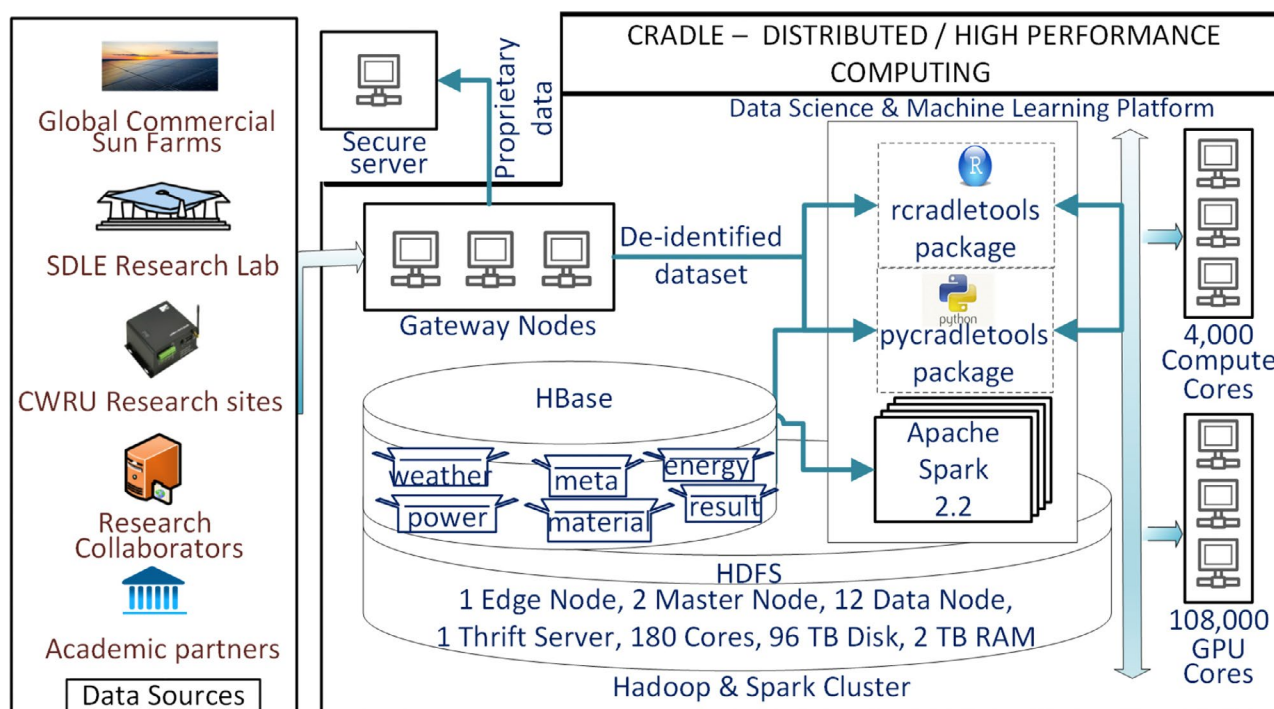


FIGURE 2 CRADLE is a 96 Tb, 180 core, distributed computing environment based on Hadoop/Hbase/Spark, embedded in CWRU's high performance computing cluster

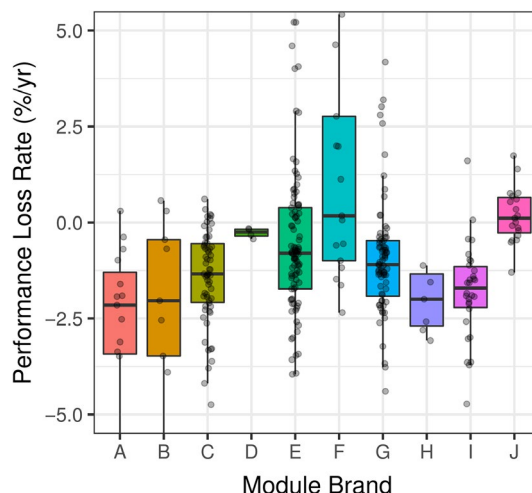


FIGURE 3 Performance Loss Rate distribution based on the PV module brand. Different brands show different performance, both in distribution and in the magnitude of the PLR

3.3 | Decoding the glass genome

Developing new functional glasses is a formidable challenge due to the extremely large search space of elements in the periodic table and continuously adjustable compositions due to the noncrystalline structure of glasses. In addition, balancing between product attributes (eg, ion exchange properties, damage resistance, or relaxation) and manufacturing attributes (eg, liquidus viscosity, forming temperature, or refractory compatibility) is key to successful glass design.

John C. Mauro reviewed his approach for decoding the glass genome (ie, glass structure-property relationships analogous to the materials genome) to reduce this search space, which has been successfully employed to provide guidance to experiments in the area of functional glasses.^{48,49} Recent efforts to decode the glass genome include developing new predictive capabilities with physics-based and machine learning models to optimize functionalities by adjusting composition subject to constraints.

According to Mauro, developing new data-oriented models to explore more diverse properties and chemistries of functional glasses provides many opportunities in applications ranging from energy and the environment to healthcare and information technologies. However, he also pointed out that there are challenges for data-driven glass R&D, including:

1. Maintaining data integrity, reporting, and security,
2. Creating and maintaining dynamically updated models,
3. Incorporating process data,
4. Bridging physical and empirical modeling approaches,
5. Leveraging complex data (eg, images),
6. Creating interatomic potentials from machine learning.

4 | CURRENT DATA SCIENCE COMMUNITY CHALLENGES THAT LIMIT INTERACTIONS WITH THE GLASS AND CERAMIC COMMUNITY

Workshop participants identified a variety of grand challenges to advancing data-driven glass and ceramic science and prioritized the most difficult barriers that crosscut both the data science and glass and ceramic communities, which are abridged in Tables 1 and 2. See supplementary material for tables (Tables S1-S10) summarizing each challenge and detailed sub-challenges.

Table 1 reports sub-challenges that received four or more votes from participants (participants could vote for up to five sub-challenges), or at least votes from 20% of participating experts from the data science and informatics community. See the supplementary material for complete tables of sub-challenges and votes.

4.1 | Data science community challenge #1: Lack of infrastructure and data sharing

The success of the Human Genome Project in Biotechnology has inspired the materials community to develop a “genomic” approach to developing new materials with programs such as MGI.⁵ Research programs based on MGI principles seek to use data and computational infrastructures to enhance the processes of discovery and development of lower cost functional materials with improved or tailored properties—discovering these materials more quickly than laboratory methods alone would achieve. To accelerate data-driven research processes, it is critical to have more accessible databases to make information searches easier.

According to a survey conducted by the Materials Research Society (MRS) in conjunction with the Minerals,

TABLE 1 Data science community challenges that limit interaction with glass and ceramics community

Data community challenges	Sub-challenges	Workshop participant votes
<i>Challenge #1</i> Lack of infrastructure and data sharing	No methods/techniques for collecting data from journals	14
	Role of journal publishers is unclear	
	Data repositories of commercial publishers require journals to be open to text mining/natural language processing (NLP)	
	Difficult to find a host for a large, precompetitive database	6
	Lack of robust NLP methods to extract unstructured data	4
<i>Challenge #2</i> Lack of Appropriate Computational Approaches to Glass and Ceramic Materials Questions	Difficult to select/match the right model to fit the data	4
<i>Challenge #3</i> Insufficient Cultural and Educational Support	Lack of “codes” to represent correlations; problem might be intractable with conventional math descriptions	4
	Not enough encouragement of diversity of all types: cultural and gender, but could extend to cross-disciplinary training	6
	Learning curve is large for classically trained materials scientists to make professional transition to materials data scientists	4
<i>Challenge #4</i> Limited Data Accessibility and Usability	User interfaces are underdeveloped	9
	Lack of common terminology and nomenclature (ontology)	6
<i>Challenge #5</i> Lack of Data Reliability and Quality Control	Uncertainty quantification (UQ), and uncertainty propagation through analyses are highly dependent on the material type, composition, model, and experimental method used to synthesize/characterize; this complexity is difficult to navigate. UQ methods analyze input uncertainties and their propagation through models and processes to understand uncertainties in the output and sensitivity to variations in the input.	5

Metals & Materials Society (TMS) in 2013, engineers and scientists often use materials databases for experimental design, experimental data interpretation and modeling, model validation, input for calculations, materials selection, and/or engineering design.⁴⁹ For example, data mining of the Inorganic Crystal Structure Database (ICSD) coupled with electronic structure calculations allows the computational identification of unknown materials such as two-dimensional materials,^{50,51} transparent conducting oxides,⁵² or ABX family of compounds such as half-Heusler compounds (NaMgX where X = P, Sb, As) for the application from thermoelectric materials to topological insulators.⁵³ Additionally, some

large databases related to electronic structures have been released^{54,55} and other materials simulation results.^{56,57}

At present, the glass and ceramic community still must work with a limited number of databases. Glass researchers, for example, rely on the Glass Property Information System (SciGlass) or International Glass Database System (INTERGLAD)^{58,59} while high demand exists to use physical and chemical properties databases along with other compositions, properties, and structures to develop advanced materials. However, only a handful of publicly accessible ceramic property databases exist and are housed at national labs, universities, and professional societies.⁶⁰ Functional ceramic oxides, for instance, are

Glass and ceramics community challenges	Sub-challenges	Workshop participant votes
<i>Challenge #1</i> Difficulty Bridging Physics-Based, Empirical, and Data-Driven Models	Lack of benchmarking models with standards	9
	Difficulty bridging physics-based and empirical models; Development of UQ along with physics-based models is difficult and not generally done.	7
	Interatomic potential development for glass and ceramic systems does not take advantage of machine learning	7
<i>Challenge #2</i> Need for Workforce Training and Education	Lack of interdisciplinary education (materials science and data science together)	13
	Lack of formal training in statistics	4
<i>Challenge #3</i> Inadequate Characterization Approaches	Experimental techniques are not “designed for statistical inference” or customized for AI applications	8
<i>Challenge #4</i> Lack of Data Infrastructure	Lack of data infrastructure, which is a key prerequisite for broad application of AI in materials	9
<i>Challenge #5</i> Limited Industry Communications and Interaction	No clear role of professional societies or publishers (eg, ACerS journals) in collecting and encouraging data sharing	6

TABLE 2 Glass and ceramics community challenges that limit interaction with data science community

an extremely important and diverse class of materials for high-end energy applications ranging from transparent conducting oxides to oxide-based photocatalytic materials.

A prototypical database (ie, materials property data bank) of functional glass and ceramic materials should be constructed in an accelerated way for data-driven approaches. The desirable database will contain engineering information derived from experimentation, theory, and existing legacy data. In addition, it is critical to have information such as a wide range of synthesis recipes to develop new glass and ceramic materials for various applications. Published literature is the most valuable source of relevant information for materials behaviors. A large volume of data such as text, sequential data, and metadata (ie, data about data) is often buried in the growing volume of literature. A fundamental limitation of the data available through literature is that information is not in a structured format. Thus, there is a critical need to develop more efficient ways to populate data from literature or web.

The success of genomics in biology originates from the use of high-end natural language processing (NLP) to capture unstructured genomic sequential data in a computer-readable form to guide experimentation. Inspired by recent work on applying text extraction and machine learning from scientific literature for materials synthesis,⁶¹ the glass and ceramic community must

work with the data science community to utilize NLP for rapid information retrieval and develop a database for functional materials.^{62,63} (Note that NLP focuses on deriving meaning from information retrieved from texts while text mining focuses on extracting data and linkage analysis of information.)

The goal of the data science and glass and ceramic communities should be to rapidly populate the glass and ceramic database through the information retrieval process aided by NLP. Using NLP from various texts, the communities will map unstructured text into a structured form. The objectives will be: (a) database construction—population of the database incorporating text-retrieved information using NLP; (b) database updating with automatic annotation and manual curation; and (c) accelerated materials discovery, including process enhancement through data analytics and other computational learning methodologies for compositions, stoichiometries, and synthesis controls of target materials.

The data science and glass and ceramic communities could start by collecting glass- and ceramic-related information from existing databases, such as SciGlass and ICSD for rules related to crystal composition and structure formation. ICSD offers scientists evaluated crystal structure data for more than 210 000 technologically relevant ceramic materials such as superconductors, fireproof materials, and intermetallic compounds. The

communities could simultaneously collect materials behaviors from the literature, performing database population with an implementation of NLP, while also expanding data population from texts to images, tables, and captions. The communities need to continue this process for figure-text combination (ie, image classifications of graphical, diagram, bar/line chart, spectra, microscopy, captions, and tables) through the final stage of the project and develop annotation (attaching information to materials sequencing), manually curate data, and perform tests with collaborators. It is critical to closely collaborate with research areas such as NLP, databases, and information systems in the data science community. During the process, federal collaborators such as NIST should be considered. The combined community needs to adopt best practices for archival data and metadata formats, which will be further discussed in 4.4.

4.2 | Data science community challenge #2: Lack of appropriate computational approaches to glass and ceramic materials questions

As noted above, data analytics algorithms should be designed to fit the glass and ceramic community's needs, but many glass and ceramic researchers have difficulty selecting and matching the right models to fit their data. An absolute construction of an optimal model that fits the data with best values is not readily available, as the selection of the best input variables for the best models remains an open problem in the data science community. Accordingly, there is always a trade-off between the accuracy of computation and the available inputs. Moreover, the selection of best input variables or the development of new computational codes requires a clear understanding of the physical problem, which is a big challenge to data scientists in general.⁵² The challenge in developing a multi-scale model becomes harder when there is a gap between empirical, physics-based, and data-driven models. To overcome the challenge, according to Curtarolo et al.,⁶⁴ the materials science community often uses "descriptors—empirical quantities or combination of materials properties (ie, a materials gene) ranging from microscopic parameters to macroscopic parameters—as input variables to construct computational models." Once the best descriptors that depict and capture the physical behaviors of glasses/ceramics are identified or developed by the glass and ceramic community, the data science community can collaborate closely with the materials community for better model development via iterative coupling of computational code/model development, validation, and interpretation.

4.3 | Data science community challenge #3: Insufficient cultural and educational support

Today, data technology is centered on solving convergent or transdisciplinary problems (eg, autonomous vehicles or the Internet of Things) in a more synergistic R&D setting under the various terms of "networks," "ecosystems," and "communities."

While data science is becoming more pervasive in science and engineering, there is a critical need for data scientists who can easily cross over boundaries between data science and engineering/science disciplines where the data is generated and utilized. There is a complementary need for materials scientists to integrate the "T-shaped" skills of data science into their education and abilities, where the vertical stem of the "T" refers to depth of knowledge in materials, and the horizontal top refers to cross-disciplinary skills that enable collaboration with data science professionals.⁶⁵ Thus, applying data science knowledge to real science/engineering problems as a data scientist still presents challenges due to the interdisciplinary nature of research and development. To address this issue, workshop attendees emphasized the need for cultural and educational diversity of the data science field, which could be achieved by demonstrating the power of data-driven approaches to other disciplines. Leveraging the concept of "teaching and learning from data," the data science community could develop new courses that address contemporary data science issues and increase awareness of computational methods for data-driven materials design, leading to skilled glass and ceramic researchers able to apply data science knowledge.⁶⁶

Because the goals and scope of activities to tackle cultural and educational issues should be determined through demand from both the glass and ceramic sciences and the data science communities, it will be crucial to gain and shape academic, industry, and federal agency perspectives on emerging data-driven technologies. This is critical for aligning the glass and ceramic community needs with valuable data science concepts and resources and to eventually establish a cohesive strategy for broadening the R&D scope of both communities.

4.4 | Data science community challenge #4: Limited data accessibility and usability

In general, technological advances in science and engineering are increasingly dependent on the ability of researchers to effectively manage and use data.⁶⁷ Desired data management frameworks for data-driven glass and ceramic R&D should encompass all activities related to the generation, release, sharing, and reuse of data and to improve the reproducibility of science.^{68,69} To maximize the value of data frameworks and enable researchers to fully leverage complex data, the glass and ceramic community must combine well-planned, well-executed data management strategies; well-organized data; and well-designed infrastructure for data sharing.^{70,71} Unfortunately, the glass and ceramic research field has long been lacking a uniform approach to data usability, despite the efforts of the broad science and engineering communities and government agencies to promote data capture and archiving for research projects with respect to the four data principles—findability, accessibility, interoperability, and reusability (FAIR)⁷²—and to require data management plans.^{73,74}

One exception is the Crystallographic Information File standards that are promoted by the International Union for

Crystallography with the purpose of developing an ontology for machine-readable crystallographic information (see <https://www.iucr.org/resources/cif> and references therein). Another example is the ACerS-NIST Phase Equilibria Diagrams database.

According to workshop participants, effective user interfaces and licenses^{75,76} must ensure the maximum value of open data,⁷⁷ data accessibility, usability, and reproducibility.⁷⁸ The user interfaces should consider how data will be treated during all R&D phases of a project and should define what happens with the data after the project concludes.^{79–82} In addition, a well-designed data management plan provides better data quality control; facilitates data access, reuse, and validation; and leads to enhanced R&D efficiency by saving R&D time and cost.⁸¹ Both the data science and glass and ceramic communities need to address these data accessibility and usability challenges to achieve the end goals of data flexibility, community consensus, and realized R&D opportunities through effective data management.

To further improve data-access and data-use practices/agreements, workshop attendees recognized that professional societies such as ACerS need to precisely assess the extent to which the science and engineering community has adopted data management initiatives.

4.5 | Data science community challenge #5: Lack of data reliability and quality control

Ensuring data quality for long-term reliability has been a long-standing challenge in both the data science and glass and ceramic communities (Table 2). The issue of data quality is particularly important, not only for enhanced data-driven models but also for fundamental research interests in measurement and characterization science. Indeed, it is one of the recurring themes of several recent roadmap/workshop reports. For instance, ASM International's workshop report on Materials Data Analytics sponsored by NIST addressed the challenge from a data science point of view. According to this report, identifying sources of uncertainty from data, models, and experiments, as well as tracking uncertainty propagation, is complicated but is also critical to advancing materials data analytics.⁸³ The ACerS technology roadmap report for functional glass described data reliability challenges from a viewpoint of glass science in particular: "the glass industry currently works with very limited reference standards and a small set of known standard compositions for instrumentation calibration and chemistries for multicomponent glasses."¹

5 | GLASS AND CERAMIC COMMUNITY CHALLENGES THAT LIMIT INTERACTIONS WITH THE DATA SCIENCE COMMUNITY

Table 2 reports sub-challenges that received four or more votes from participants (participants could vote for up to five

sub-challenges), or about at least votes from 20% of votes participating experts from the perspective of the glass and ceramic science community. See the supplementary material for complete tables of sub-challenges and votes.

5.1 | Glass and ceramic community challenge #1: Difficulty bridging physics-based, empirical, and data-driven models

As discussed above, the objective of decoding the glass genome is to enable the predictive design of new glassy materials. However, designing functional materials, in particular glasses or ceramics for specific applications, is a complex process due to the need to determine optimal combinations of material chemistry, processing routes, and synthesis parameters that meet specific requirements such as mechanical properties.⁴⁷ Current efforts to address this issue in the glass and ceramic community include the use of a combined approach of physics-based and empirical modeling with data-driven techniques, including statistical learning and machine learning algorithms, to capture and simulate information about glass structure, topology, and/or chemical bonding. However, a lack of benchmarking models with standards is still an open problem, which is a corollary of the fundamental difficulty bridging physics-based models to empirical or data-driven models.

To address this challenge, the materials science community has been focused on interatomic potential development. For example, as a part of its response to MGI, NIST launched the Interatomic Potential Repository which provides a source for interatomic potentials (or force fields) and evaluation tools to enable researchers to perform simulations and modeling.⁸⁴ The repository is intended to include broader classes of materials such as various metals, semiconductors, oxides, and carbon-containing systems. The glass and ceramic community needs to develop empirical, transferable interatomic potentials for simulations of glass and ceramics at the atomic level, which are in good agreement with experimental results. Here, a good transferability means the ability to capture a wide range of composition, processing history, and coordination states of atoms.^{85,86}

5.2 | Glass and ceramic community challenge #2: Need for workforce training and education

The main purpose of science and engineering is to enhance the quality of human life, and effective education for each level of the workforce—from K-12 students to incumbent professionals—is an indispensable component of realizing this mission. Here, the main characteristic of the desired workforce for the glass and ceramic community is the ability to independently solve current glass and ceramic materials challenges through innovative data science approaches. To fulfill this increased demand for skilled

interdisciplinary experts in glass and ceramic and data sciences, it is critical to develop effective curriculum and teaching strategies.

Applying the knowledge of data scientists to materials education requires well-developed teaching tools and computational methodologies that are designed to significantly boost student motivation to learn various concepts of glass and ceramic science. Good examples include: scientific data representation tools that aim to remove bottlenecks in visualization of complex glass and ceramic data to uncover hidden relationships, patterns, and outliers; data-oriented pedagogical tools utilizing statistics and data analytics; and tools to enhance the teaching and learning of terminologies commonly used in both data and glass and ceramic sciences.

An increased understanding of structure-process-property relationships will be a key enabler of successful R&D in the glass and ceramic community. Therefore, it is also crucial to establish an integrated computational framework based on materials data analytics for teaching (micro) structure-process-property relationships of glasses and ceramics. There are broad activities for education of materials science, including the well-known teaching web page MATTER,⁸⁷ CES EduPack,⁸⁸ and materials education symposia,⁸⁹ but pedagogical tools and approaches being developed by the glass and ceramic science and data science communities should be uniquely designed to focus on real-world complex glass and ceramic science problems.

Developing data, programming, and statistical literacy

Elizabeth Dickey

The increasing abundance of data in almost every aspect of glass/ceramics R&D and manufacturing is increasing the demand for students and postgraduates with highly developed skills in data science, programming and statistics, with the ability to apply these skills to the design of new materials and materials processes and to identify statistical correlations with property and performance metrics.

Several universities in the United States are exploring new approaches to graduate education under the auspices of the National Science Foundation Research Traineeship Program,⁹⁰ which challenges the academic community to develop new graduate training paradigms in “Data-Enabled Science and Engineering” and “Harnessing the Data Revolution.” Emerging Ph.D. training programs such as Texas A&M’s *Data-Enabled Discovery and Design of Energy Materials*⁹¹ and North Carolina State University’s *Data-Enabled Science and Engineering of Atomic Structure* (Figure 4), are highly interdisciplinary programs that often link curricula and research training activities in materials science and engineering with statistics, applied mathematics, industrial engineering, and/or computer science. These programs also teach strong communication and interaction skills, which are necessary for working in highly interdisciplinary environments.^{92,93} In addition, specialized master degree programs are emerging, such as Case Western Reserve University’s accelerated M.S. in materials science and engineering with data science and their undergraduate concentration in materials data science and minor in applied data science,⁹⁴ which in addition to traditional materials science courses, incorporates data science modeling, prediction, and statistical and machine learning, along with a capstone research project on applied data science.

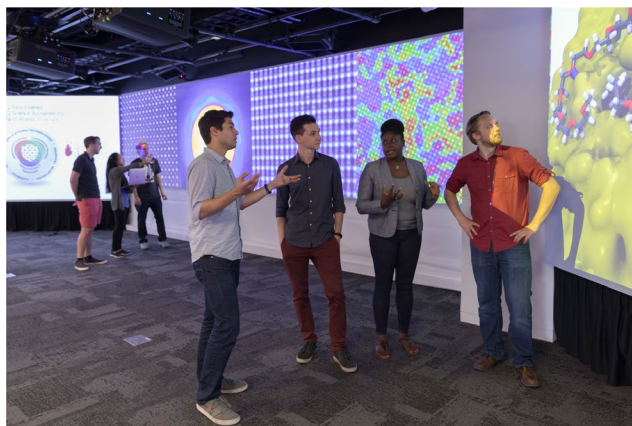


FIGURE 4 Graduate students in the NSF Research Traineeship on the Science and Engineering of Atomic Structure at NC State University, discuss challenges in the statistical analysis of experimental and computational data and opportunities to incorporate machine learning algorithms. Credit: NCSU

5.3 | Glass and ceramic community challenge #3: Inadequate data fusion approaches

While huge amounts of microstructure-related glass and ceramic materials data have been generated through recent advancements in imaging techniques, there is still a need to understand structural mechanisms as a function of materials synthesis or processing routes, regardless of physical length scales ranging from the atomic level to the mesoscale. However, there is no single characterization/metrology tool to scan all detection ranges. In addition, it is not practical to use various in-line characterization tools due to the complexity of glass and ceramic processes.

One of the promising approaches to solve these challenges is to design experimental techniques for AI applications or statistical inference-based information interpretation. For example, a wide range of AI algorithms can be used to merge various characterization data—in particular, to fuse image data at different scales—for full insight into physical phenomena of complex ceramic or glass materials. According to Rajan, the usefulness and evaluation of materials descriptors has been a fundamental question in the field of statistical inference.⁹⁵ The concept of statistical inference will be a value-added benefit to existing state-of-the-art experimental approaches by quantitatively identifying information that influences structure-process-property relationships in glass and ceramic science.

5.4 | Glass and ceramic community challenge #4: Lack of data infrastructure

A well-functioning, collaborative R&D environment for emerging technologies such as data science is an essential asset to the data-driven glass and ceramic community. Data infrastructure for the glass and ceramic community is urgently needed to facilitate not only sharing glass and ceramic data, but also exchange of computational codes, algorithms, and ideas through structured communication and engagement with internal/external experts. It is essential to have a clear and concise strategy for cross-functional planning and implementation to address the current challenges of establishing a collaborative data infrastructure.

5.5 | Glass and ceramic community challenge #5: Limited industry communications and interaction

Over the last 50 years, there have been a significant number of technology movements, which are still driving the technology revolution. During the workshop, attendees explored the technology flow, connections, and open/sharing of data across technology movements. They concluded that the glass and ceramic community needs clear communication

and defined mechanisms for interaction with the data science community for more active application of data-driven approaches. To this end, both communities should identify and create flexible, community-driven practices to encourage the adoption of universal FAIR data principles.⁵⁵ Such efforts via open dialogue, such as through professionally facilitated meetings, could have two major impacts:

1. **Promoting glass and ceramic community consensus:** Apart from sharing data and promoting public engagement between all stakeholders, data exchange practices and data policies created through community consensus would be major community-wide discussion topics, under the collaborative environment discussed in Section 4.44.1.
2. **Facilitating data-related (eg, data availability, data standardization, or data ownership) education:** Community consensus on a collaborative environment for open data and data sharing/exchange will help educators teach data-related issues through a series of data-oriented education programs aimed at developing a workforce skilled in data-driven glass and ceramic science.

6 | COLLABORATIVE OPPORTUNITIES BETWEEN THE GLASS AND CERAMIC COMMUNITY AND THE DATA SCIENCE COMMUNITY

After identifying and prioritizing challenges, the workshop participants achieved consensus on the top four opportunity areas for collaboration between the glass and ceramic community and the data science community. Focus on these opportunities could help stimulate a culture of data and code sharing, support the creation of accessible data and algorithm repositories, and facilitate the development of community-developed standards needed for a robust data infrastructure.

Working in small groups, the participants identified near-term action plans for each opportunity area, outlining key activities, value propositions, scope, format, success metrics, and next steps. The action plan matrices are outlined below.

6.1 | Opportunity #1: Bridging physics-based, empirical, and data-driven models

One of the collaborative opportunities is to bridge physics-based, empirical, and data-driven models. However, interatomic potential development for glass and ceramic materials modeling does not take advantage of machine learning. Moreover, there is a need to design experimental

TABLE 3 Collaborative Opportunity #1: Bridging Physics-based, Empirical, and Data-driven Models

Activity	Value Proposition	Scope	Format	Success Metric	Next Steps
Topical meeting	Build a new technical community and assemble skill set	Data science and ceramic/glass around these technical areas	Co-locate with an established meeting	Community comes back, and the event becomes self-perpetuating	Find champions and organizers
• Standalone data/glass meeting			• Recruit top invited speakers		
• May be specific to narrow technical topics					
Short courses or webinars	Help build excitement and competencies	Pitch content at appropriate levels for glass/ceramics data scientists	Online or in-person classes	Many people sign up for these courses	Topics and features
• Data scientists to learn glass and ceramic science	Use data analytics in education of members			At least 10 participants to start; more than 20 would be success	
• Glass and ceramic scientists to learn data science					
Influence funders (agencies, companies, other partners)	Increase funding across the board to advance the article/paper and its outcomes	Encompass points raised in this meeting	Report on this workshop with diverse authorship	Increased journal articles	Recruit co-authors
			White paper on data science and ceramics/glass	More abstracts at meetings—enough for a session (6–8)	

TABLE 4 Collaborative Opportunity #2: Creating Collaboratory Infrastructure

Activity	Value Proposition	Scope	Format	Success Metric	Next Steps
Highlight existing efforts	Guidance for members' data management policies	All glass/ ceramics-related data, regardless of journal	Website	Adoption	Create task force for developing recommendations
Create domain database editors				Large number of visitors to site	Engage publications (communications and editors)
Guidelines for publishing certain types of data for publishers					
Develop glass-related metadata requirements					
Make NIST/ACerS phase diagrams more accessible	Selling database revenue for ACerS	Phase diagrams and SciGlass/INTERGLAD update	Access-controlled online data platform	Comprehensive and usable data	Identify SciGlass owner
	Data for users			Number of users	Contact Japanese Ceramics Society
	Enable new science and enhance value of published data				Discuss next steps with NIST for phase diagrams
Special issue for data-driven ceramic science	Encourage good practices	Glass and ceramics-related journals	Special mark on articles	Views of papers	Launch pilot programs
Special mark for articles with good practice	Be a leader, increase visibility		Special issue	Change in number of "good" papers over time	

TABLE 5 Collaborative Opportunity #3: Boosting Interdisciplinary Education

Activity	Value Proposition	Scope	Format	Success Metric	Next Steps
Materials Scientist of Tomorrow Data science skillset definition	Workforce development, increased salary, career flexibility	Formal and continuing education standards Tie to industry needs	Report or publication <ul style="list-style-type: none">Highlight role models	Number of downloads, citations, and programs adopting these guidelines	Identify actors (universities, national labs, industry) Communicate with other professional societies
Research on how to address data science education Promote exchange of best practices	Pedagogical framework	Faculty, student development Network for expertise	Data science page on ACerS website Recommendations for online courses Tutorials for educators Educational modules, easy-to-use examples Massive Open Online Course (MOOC)	Number of students Diversity	Identify instructors
Materials data science symposium	Member engagement Networking Convergence Skill development	Implement within existing conferences Target more international meetings	Symposium at existing conferences	10,000 (±) people at symposium Happens by MS&T 2019 or sooner	Identify a dynamic, TED-talk style spokesperson to champion this space at MS&T Identify the right existing meetings to target Communicate to organizers of meetings to start coordinating
Encourage diversity	Data science is a natural fit to enhance diversity—it's accessible (cost-wise) to many more universities than expensive capital equipment (democratizes cutting-edge research capabilities)				Find experts to do "show-and-tell" demos for students

techniques for inference and AI applications. Table 3 outlines an opportunity matrix for addressing these issues.

6.2 | Opportunity #2: Creating collaborative data infrastructure

One of the recurring themes in the workshop was to create data infrastructure, which is a key prerequisite for broad application of data science in ceramic/glass R&D as well as facilitation of data, codes, and information sharing with enhanced data standardization and ontology. Table 4 is an opportunity matrix for creating collaborative data infrastructure.

6.3 | Opportunity #3: Boosting interdisciplinary education

The importance of preparing the 21st-century scientists who understand both ceramic and glass science and data science cannot be overemphasized; there is currently a critical need for skilled positions focused on the use of state-of-the-art data technologies for advancing materials R&D. Table 5 is an opportunity matrix that describes suggested approaches to developing collaborative mechanisms across the glass and ceramic and data science communities to prepare the next-generation workforce.

University at Buffalo Department of Materials Design and Innovation: Weaving data science through a new materials science curriculum

Krishna Rajan

In 2015, the University at Buffalo (UB) established the Department of Materials Design and Innovation (MDI) that *has* introduced a completely novel curriculum by bringing a data perspective in all its core subjects for the study of materials theory, characterization, synthesis, processing and computational and simulation techniques (Figure 5). This bold decision to create an entirely new department, instead of modifying and/or merging existing curricular programs, was motivated by the belief that a truly new paradigm based on data-driven materials science requires rethinking how we teach and learn materials science. MDI uses new cross-disciplinary methodologies and novel pedagogical approaches to promote the convergence of physics, computer science, mathematics, chemistry, engineering disciplines, and materials science. This was made possible through the hiring of a new cadre of faculty, every one of whom has convergent skillsets in *both* data science and domain applications. MDI's curriculum uses the mathematical rigor of data science to explore and navigate the core concepts that define structure-property-processing relationships in all classes of materials, and informatics as the "connective tissue" between experiments and computation to uncover patterns and relationships in an accelerated manner that would otherwise not be easily seen or be left undiscovered. MDI's pedagogical approach heightens *awareness* of data in all facets of materials science and engineering, develops statistics, machine learning and mathematics skills, and shows how to judiciously use the tools of data analytics with the theory of materials behavior. MDI has succeeded in attracting students from diverse backgrounds who, through their formal coursework and research projects, are learning to communicate and collaborate across disciplines. Some of the students come with physical/life sciences and engineering backgrounds, interested in adding informatics training to their repertoire, while students with computer science and mathematics backgrounds find in MDI opportunities to enter the materials science domain.



FIGURE 5 University at Buffalo's new Department of Materials Design and Innovation brings a data perspective to all materials science core courses including theory, characterization, synthesis, processing, and computational methods. Professor and department chair, Krishna Rajan, instructs students in the new curriculum. Credit: Rajan, U. Buffalo

TABLE 6 Collaborative Opportunity #4: Enhancing Data Usability and Visualization

Activity	Value Proposition	Scope	Format	Success Metric	Next Steps
User interface design challenge or visualization contest	Bringing data science and ceramic science students together	Build an intuitive interface <ul style="list-style-type: none">• Instrument• Remote operation• Density functional theory (DFT) calculation	Student chapter competition (like Mug Drop)	Number of participants	Design the rules and format
	• Education		Data visualization challenges on how to communicate multi-dimensional data in single visual	Ability to make design tool useful to other disciplines, diversify users	Enlist organizers
	• Chapter competition				Find sponsors
Symposia on visualization and interfaces	Accelerated discovery				
	Recruitment	Build a simulator (eg, virtual reality activities with goggles)	3D experiential learning roller coaster ride through energy landscape	Number of visitors and downloads	Develop content, with member contributions
	Education		Real and/or online (download)		Find sponsors
Virtual reality hub in exhibit area	Research	Symposium on visualization and user interface design for materials	Lectures and demonstration	Attendance	Identify organizers and speakers
	Education				
	Networking				
3D experiential learning journey-through-glass rollercoaster ride across energy landscape					
Data visualization tool with journal publications	Education	Facilitate upload of videos	Publicize tool	Number of downloads	Journals should encourage upload of visualization/ video as supplemental information
	Accelerate discovery	Develop new tools	Encourage authors to upload videos		Develop visualization tools for various formats

6.4 | Opportunity #4: Enhancing data usability and visualization

Enhanced user interfaces for increased data accessibility and usability are beneficial to both glass and ceramic and data science communities. The glass and ceramic community would enlarge the nationwide data-driven materials technology network while allowing the data science community to access materials-related data resources for testing and validating advanced high-end computational algorithms. Suggested routes for success stories in both communities will show how advanced data technologies can be used to rapidly and successfully solve glass and ceramic research problems (Table 6).

7 | KEY CONCLUSIONS

Several key conclusions can be drawn from the discussions at the Glass and Ceramics Data Science Workshop, including:

1. Demand for new materials with extraordinary functionality creates an urgent need for use of rapid discovery and refinement techniques. Data-driven methods, encompassing tools such as machine learning, artificial intelligence, and big data informatics, offer a new paradigm for researchers to discover new glasses and ceramics efficiently and effectively. Development of close collaborations between data science and materials science research communities is an essential element of realizing the objective.
2. Leaders in the glass and ceramic research community already are embracing data science/informatics tools to study complex materials, processes, and applications. Key collaborative challenges to solve include bridging physics-based, empirical, and data-driven models; creating collaborative data infrastructure; boosting interdisciplinary education; and enhancing data usability and visualization. Solutions to these challenges will require fluency in data science as well as materials science.
3. Glass and ceramic researchers will need near-expert proficiency in data science and informatics, and vice-versa, making education in data science methods is increasingly important for glass and ceramic researchers. Universities and ACerS are best-positioned to respond to this need with input from industrial customers.
4. The moment is opportune for ACerS with key members to lead efforts to solve data challenges, create cross-disciplinary education and networking opportunities, and raise the level of awareness of the opportunities that advances in data science offer for advancing the state-of-the-art for glass and ceramics.
5. While the workshop's goal was to provide direction to ACerS, these findings will serve as useful guideposts for any organization building a data science and informatics strategy for its constituents.

ACKNOWLEDGMENTS

This workshop and the resulting report were supported by NIST AMTech Award No. 70NANB15H073, Jean-Louis Staudenmann, program officer. E.C. Dickey acknowledges the support of the National Science Foundation under Grant No. DGE-1633587. R. French acknowledges the support of DOE-EERE SETO award DE-EE-000 8172. B. Meredig and S.M. Haile acknowledge the support of ARPA-E under contract DE-AR0000707. Their project is a collaboration between Citrine Informatics and the research group of Sossina M. Haile at Northwestern University.

AUTHOR CONTRIBUTIONS

Sarah Lichtner, Changwon Suh, and Ross Brindle of Nexight Group designed and facilitated the workshop with direction from ACerS (Eileen De Guire and Charlie Spahr). Sarah Lichtner, Changwon Suh, and Eileen De Guire contributed to writing the manuscript. All authors contributed to the identification of challenges and opportunities at the intersection of glass and ceramic science and data science that laid the foundation for this work.

ORCID

Eileen De Guire  <https://orcid.org/0000-0002-1602-5056>

Elizabeth C. Dickey  <https://orcid.org/0000-0003-4005-7872>

Roger H. French  <https://orcid.org/0000-0002-6162-0532>

Martin Harmer  <https://orcid.org/0000-0003-1389-3306>

John Mauro  <https://orcid.org/0000-0002-4319-3530>

REFERENCES

1. The American Ceramic Society. Functional glass manufacturing innovation consortium. <https://ceramics.org/professional-resources/functional-glass-manufacturing-innovation-consortium>. Accessed May 26, 2019.
2. Carlisle NB, Rickman JW. Workshop on the convergence of materials research and multi-sensory data science roadmap. Lehigh University; 2018 [Unpublished].
3. Rajan K. Materials informatics. *Mater Today*. 2005;8(10):38–45.
4. National Research Council. Committee on Integrated Computational Materials Engineering, National Materials Advisory Board, Division on Engineering and Physical Sciences. Integrated computational materials engineering: A transformational discipline for improved competitiveness and national security. Washington, DC: National Academies Press. 2008;ISBN:9780309178211.
5. Materials genome initiative. <https://www.mgi.gov>. Accessed November 25, 2018.
6. High Performance Computing for Manufacturing. <https://hpc4mfg.llnl.gov>. Accessed November 25, 2018.

7. Kalil T, Bruce A. cochairs. National Science and Technology Council Committee on Technology. National nanotechnology initiative strategic plan. http://www.nano.gov/sites/default/files/2016_nni_strategic_plan_public_comment_draft.pdf. Accessed November 25, 2018.
8. Holdren JP, Donovan S. cochairs. The National Strategic Computing Initiative Executive Council. National strategic computing initiative strategic plan. https://www.whitehouse.gov/sites/whitehouse.gov/files/images/NSCI%20Strategic%20Plan_20160721.pdf.pdf. Accessed November 25, 2018.
9. Felten E, Garriss M. cochairs. National Science and Technology Council, Networking and Information Technology Research and Development Subcommittee. The national artificial intelligence research and development strategic plan. [cited Nov. 25, 2018]. https://www.nitrd.gov/PUBS/national_ai_rd_strategic_plan.pdf. Accessed November 25, 2018.
10. Office of Energy Efficiency and Renewable Energy. Workshop on artificial intelligence applied to materials discovery and design. <https://energy.gov/eere/amo/events/workshop-artificial-intelligence-applied-materials-discovery-and-design>. Accessed November 3, 2018.
11. Centre Européen de Calcul Atomique et Moléculaire. École Polytechnique Fédérale de Lausanne. 2nd NOMAD (Novel Materials Discovery) Industry Workshop. <https://www.cecami.org/workshop-1377.html>. Accessed November 25, 2018.
12. Centre Européen de Calcul Atomique et Moléculaire. École Polytechnique Fédérale de Lausanne. Big-data driven materials science. <https://www.cecami.org/workshop-1437.html>. Accessed November 25, 2018.
13. Centre Européen de Calcul Atomique et Moléculaire. École Polytechnique Fédérale de Lausanne. Machine learning in atomistic simulations. <https://www.cecami.org/workshop-746.html>. Accessed November 25, 2018.
14. Le TC, Winkler DA. Discovery and optimization of materials using evolutionary approaches. *Chem Rev*. 2016;116:6107–32.
15. DeCost BL, Holm EA. A computer vision approach for automated analysis and classification of microstructural image data. *Comput Mater Sci*. 2015;110:126–33.
16. DeCost BL, Holm EA. Exploring the microstructure manifold: Image texture representations applied to ultrahigh carbon steel microstructure. *Acta Mater*. 2017;133:30–40.
17. Balachandran PV, Kowalski B, Sehirlioglu A, Lookman T. Experimental search for high-temperature ferroelectric perovskites guided by two-step machine learning. *Nat Commun*. 2018;9:1668.
18. Hutchinson ML, Antono E, Gibbons BM, Paradiso S, Ling J, Meredig B. Overcoming data scarcity with transfer learning. 2017, arXiv preprint arXiv:1711.05099. Presented at the 31st Conference on Neural Information Processing Systems (NIPS 2017). Long Beach, CA; 2017.
19. Ling J, Hutchinson ML, Antono E, Paradiso S, Meredig B. High-dimensional materials and process optimization using data-driven experimental design with well-calibrated uncertainty estimates. *Integr Mater Manuf Innov*. 2017;6(3):207–17.
20. Steele B. Oxygen ion conductors and their technological applications. *Solid State Ionics*. 1992;13(2):17–28.
21. Jasinski P, Suzuki T, Anderson HU. Nanocrystalline undoped ceria oxygen sensor. *Sens Actuators B Chem*. 2003;95(1–3):73–7.
22. Haile SM. Fuel cell materials and components. *Acta Mater*. 2003;51(19):5981–6000.
23. Kharton VV, Yaremchenko AA, Kovalevsky AV, Viskup AP, Naumovic EN, Kerko PF. Perovskite-type oxides for high-temperature oxygen separation membranes. *J Membrane Sci*. 1999;163(2):307–17.
24. Michel K, Meredig B. Beyond bulk single crystals: a data format for all materials structure-property-processing relationships. *MRS Bull*. 2016;41(08):617–23.
25. O'Mara J, Meredig B, Michel K. Materials data infrastructure: a case study of the Citrination platform to examine data import, storage, and access. *JOM*. 2016;68(8):2031–4.
26. Meredig B, Agrawal A, Kirklin S, Saal JE, Doak JW, Thompson A, et al. Combinatorial screening for new materials in unconstrained composition space with machine learning. *Phys Rev B*. 2014;89(9):094104.
27. Ling J, Hutchinson M, Antono E, Paradiso S, Meredig B. High-dimensional materials and process optimization using data-driven experimental design with well-calibrated uncertainty estimates. *IMMI*. 2017;6(3):207–17.
28. Huang R, Davenport TC, Meredig B, Haile SM. Private communication.
29. Hu Y, Gunapati VY, Zhao P, Gordon D, Wheeler NR, Hossain MA, et al. A nonrelational data warehouse for the analysis of field and laboratory data from multiple heterogeneous photovoltaic test sites. *IEEE J Photovoltaics*. 2017;7(1):230–6.
30. French RH, Podgornik R, Peshek TJ, Bruckman LS, Xu Y, Wheeler NR, et al. Degradation science: mesoscopic evolution and temporal analytics of photovoltaic energy materials. *Curr Opin Solid St M*. 2015;19:212–26.
31. Apache Hadoop. Apache Software Foundation; 2018. <http://hadoop.apache.org>. Accessed July 13, 2018.
32. Apache Software Foundation. Apache Spark™—Unified analytics engine for big data; 2018. <http://spark.apache.org>. Accessed July 13, 2018.
33. Klinke AG, Gok A, Ifeanyi SI, Bruckman LS. A non-destructive method for crack quantification in photovoltaic backsheets under accelerated and real-world exposures. *Polym Degrad Stab*. 2018;153:244–54.
34. Huang W-H, Wheeler N, Klinke A, Xu Y, Du W, Gok A, et al. netSEM: Network structural equation modelling; 2018. <https://CRAN.R-project.org/package=netSEM>. Accessed June 14, 2018.
35. Jahn U, Herz M, Köntges M, Parlevliet D, Paggi M, Tsanakas I, et al. Review on IR and EL imaging for PV field applications. IEA-PVPS Task 13; 2018. <http://www.iea-pvps.org/index.php?xml:id=480>. Accessed June 10, 2018.
36. GitHub. A development platform, San Francisco, CA. <http://www.github.com>. Accessed July 29, 2018.
37. Atlassian Bitbucket. Developer of collaboration and productivity software. San Francisco, CA: Atlassian Corporation Plc. <http://www.bitbucket.org>. Accessed July 29, 2018.
38. Chansler R, Kuang H, Radia S, Shvachko K, Srinivas S The Hadoop distributed file system; 2018. <http://www.aosabook.org/en/hdfs.html>. Accessed July 29, 2018.
39. GitLab. Open source single application for DevOps lifecycle.. San Francisco, CA: GitLab Inc. <http://www.gitlab.com>. Accessed July 29, 2018.
40. Balachandran PV, Kowalski B, Sehirlioglu A, Lookman T. Experimental search for high-temperature ferroelectric perovskites guided by two-step machine learning. *Nat Commun*. 2018;9:1668.

41. Rubel F, Brugger K, Haslinger K, Auer I. The climate of the European Alps: shift of very high resolution Köppen-Geiger climate zones 1800–2100. *Meteorol Z.* 2016;26(2):115–25.
42. Bryant C, Wheeler NR, Rubel F, French RH. Kgc: Koeppen-Geiger Climatic Zones. 2017. <https://cran.r-project.org/web/packages/kgc/index.html>. Accessed November 20, 2017.
43. Peshek TJ, Fada JS, Hu Y, Xu Y, Elsaeti MA, Schnabel E, et al. Insights into metastability of photovoltaic materials at the mesoscale through massive I-V analytics. *J Vac Sci Technol B.* 2016;34:050801.
44. Pickering EM, Hossain MA, French RH, Abramson AR. Building electricity consumption: data analytics of building operations with classical time series decomposition and case based subsetting. *Energy Build.* 2018;177:184–96.
45. Verma AK, French RH, Carter J. Physics-informed network models: a data science approach to metal design. *Integr Mater Manuf Innov.* 2017;6:279–87.
46. Curran AJ, Hu Y, Haddadian R, Meakin D, Peshek TJ, French RH. Determining the power change rate of 373 plant inverter's time-series data across multiple climate zones, using a month-by-month data science analysis. *IEEE PVSC-44*, June 25–30, 2017. <http://www.ieee-pvsc.org/PVSC44/>. Accessed July 17, 2019.
47. Mauro JC. Decoding the glass genome. *Curr Opin Solid St M.* 2018;22(2):58–64.
48. Mauro JC, Tandia A, Vargheese KD, Mauro YZ, Smedskjaer MM. Accelerating the design of functional glasses through modeling. *Chem Mater.* 2016;28(12):4267–77.
49. Ward CH, Warren JE, Hanisch RJ. Making materials science and engineering data more valuable research products. *Integr Mater Manuf Innov.* 2014;3:22.
50. Inoshita T, Jeong S, Hamada N, Hosono H. Exploration for two-dimensional electrides via database screening and ab initio calculation. *Phys Rev X.* 2014;4(3):031023.
51. Lebègue S, Björkman T, Klintonberg M, Nieminen RM, Eriksson O. Two-dimensional materials from data filtering and ab initio calculations. *Phys Rev X.* 2013;3(3):031002.
52. Hautier G, Miglio A, Ceder G, Rignanese G-M, Gonze X. Identification and design principles of low hole effective mass p-type transparent conducting oxides. *Nat Commun.* 2013;4(1):2292.
53. Zakutayev A, Zhang X, Nagaraja A, Yu L, Lany S, Mason TO, et al. Theoretical prediction and experimental realization of new stable inorganic materials using the inverse design approach. *J Am Chem Soc.* 2013;135(27):10048–54.
54. Saal JE, Kirklin S, Aykol M, Meredig B, Wolverton C. Materials design and discovery with high-throughput density functional theory: the open quantum materials database (OQMD). *JOM.* 2013;65(11):1501–9.
55. AFLOW: automatic—FLOW for materials discovery. [cited Nov. 25, 2018]. www.aflowlib.org
56. Jain A*, Ong SP*, Hautier G, Chen W, Richards WD, Dacek S et al. (*=equal contributions). The materials project: a materials genome approach to accelerating materials innovation. *APL Mater.* 2013;1(1):011002.
57. Jain A. Materials project. The materials project. 2018. <https://materialsproject.org/>. Accessed July 13, 2018.
58. AKos GmbH. Sciglass-glass property information system. <http://www.akosgmbh.de/sciglass/sciglass.htm>. Accessed November 25, 2018.
59. International glass database system: INTERGLAD Ver. 7. http://www.newglass.jp/interglad_n/gaiyo/info_e.html. Accessed November 25, 2018.
60. Freiman S, Rumble J. Current availability of ceramic property data and future opportunities. *Am Ceram Soc Bull.* 2013;92(3):34–9.
61. Kim E, Huang K, Saunders A, McCallum A, Ceder G, Olivetti E. Materials synthesis insights from scientific literature via text extraction and machine learning. *Chem Mater.* 2017;29:9436–44.
62. Cohen KB, Hunter LE. Text mining for translational bioinformatics. *PLOS Comput Biol.* 2013;9(4):e1003044.
63. Kajikawa Y, Abe K, Noda S. Filling the gap between researchers studying different materials and different methods: A proposal for structured keywords. *J Info Sci.* 2006;32(6):511–24.
64. Curtarolo S, Hart G, Nardelli MB, Mingo N, Sanvito S, Levy O. The high-throughput highway to computational materials design. *Nat Mat.* 2013;12:191–201.
65. Hughes D, French RH. Crafting a minor to produce T-shaped graduates. 2016. http://tsummit.org/files/T-Summit_Speaker_Abstracts-2016.pdf. Accessed November 25, 2018.
66. Business Higher Education Forum. Creating a minor in applied data science. 2016. <http://www.bhef.com/publications/creating-minor-applied-data-science>. Accessed August 16, 2016.
67. Wadia C, Stebbins M. It's time to open materials science data. White House: The White House, Office of Science and Technology Policy. 2015. <https://www.whitehouse.gov/blog/2015/02/06/its-time-open-materials-science-data>. Accessed April 27, 2015.
68. Announcement: reducing our irreproducibility. *Nature.* 2013;496:398–398.
69. Peng RD. Reproducible research in computational science. *Science.* 2011;334:1226–7.
70. Marburger III JH. Harnessing the power of digital data for science and society: Report of the interagency working group on digital data to the Committee on Science of the National Science and Technology Council. 2009. <https://catalog.data.gov/dataset/harnessing-the-power-of-digital-data-for-science-and-society-report-of-the-interagency-wor>. Accessed May 30, 2019.
71. Allison J, Cowles B, DeLoach J, Pollock T, Spanos G. Integrated computational materials engineering (ICME): Implementing ICME in the aerospace, automotive, and maritime industries. Warrendale: The Minerals, Metals & Materials Society; 2013. http://d3em.tamu.edu/wp-content/uploads/2016/04/Report-TMS_icme_study_2013PDF.pdf. Accessed July 17, 2019.
72. DOE Public Access Plan. Department of Energy, (n.d.). <http://energy.gov/downloads/doe-public-access-plan>. Accessed November 27, 2016.
73. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. Comment: the FAIR guiding principles for scientific data management and stewardship. *Sci Data.* 2016;3:160018.
74. National Science Foundation. Dissemination and sharing of research results. <https://www.nsf.gov/bfa/dias/policy/dmp.jsp>. Accessed November 25, 2018.
75. Choose a License (n.d.). Choose an open source license. <http://choosealicense.com/>. Accessed April 8, 2016.
76. Open Data Commons. Open Data Commons Open Database License (ODbL). 2018 [cited April 8, 2016]. <http://opendatacommons.org/licenses/odbl/>. Accessed April 8, 2016.
77. Hendler J, Holm J, Musialek C, Thomas G. US government linked open data: semantic.data.gov. *IEEE Intell Syst.* 2012;27(3):25–31.

78. Lowndes J, Best BD, Scarborough C, Afflerbach JC, Frazier MR, O'Hara CC, et al. Our path to better science in less time using open data science tools. *Nature Ecol Evol*. 2017;1:0160.
79. DataONE. Earth, data observation network for Earth. <https://www.dataone.org>. Accessed July 17, 2019.
80. Michener WK. Ten simple rules for creating a good data management plan. *PLoS Comput Biol*. 2015;11(10):1–9.
81. Chaussabel B, Ueno H, Banchereau J, Quinn C. Data management: it starts at the bench. *Nat Immunol*. 2009;10(12):1225–7.
82. Kowalczyk ST. Before the repository: Defining the preservation threats to research data in the lab. JCDL '15 Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital LibrariesKnoxville; 2015.
83. ASM International. Materials data analytics: a path-finding workshop results. 2015. <https://www.asminternational.org/documents/10192/25925847/ASM+MDA+Workshop+Report+Final.pdf/0e29644e-a439-4928-a07a-8718817a46e4>. Accessed November 25, 2018.
84. Becker CA, Tavazza F, Trautt ZT, de Macedo R. Considerations for choosing and using force fields and interatomic potentials in materials science and engineering. *Curr Opin Solid St M*. 2013;17(6):277–83.
85. Wondraczek L, Mauro JC. Advancing glasses through fundamental research. *J Eur Ceram Soc*. 2009;29:1227–34.
86. Wang M, Krishnan N, Wang B, Smedskjaer MM, Mauro JC, Bauchy M. A new transferable interatomic potential for molecular dynamics simulations of borosilicate glasses. *J Non Cryst Solids*. 2018;498:294–304.
87. UK Centre for Materials Education. The Higher Education Academy. <http://www.materials.ac.uk/elearning/matter/>. Accessed November 25, 2018.
88. Granta Materials Intelligence. CES 2018 EduPack. <http://www.grantadesign.com/education/>. Accessed November 25, 2018.
89. Granta Design. Materials education symposia.. <http://www.materialseducation.com/>. Accessed November 25, 2018.
90. Dickey EC, Greer A. Big data meets materials science: Training the future generation. *Am Ceram Soc Bull*. 2017;96(6):40–4.
91. Chang CN, Semma B, Pardo ML, Fowler D. Data-enabled discovery and design of energy materials (D3EM): Structure of an interdisciplinary materials design graduate program. *MRS Adv*. 2017;2(31–32):1693–8.
92. Committee on Facilitating Interdisciplinary Research, National Academy of Sciences, National Academy of Engineering, and Institute of Medicine. Facilitating interdisciplinary research. The National Academies Press; 2005. <http://www.nap.edu/catalog/11153.html>. Accessed July 17, 2019.
93. Strober M. Interdisciplinary conversations: challenging habits of thought. Palo Alto, CA: Stanford University Press; 2010.
94. Business Higher Education Forum. Creating a minor in applied data science. 2016. <http://www.bhef.com/publications/creating-minor-applied-data-science>. Accessed August 16, 2016.
95. Rajan K. Materials informatics: The materials “gene” and Big Data. *Annu Rev Mater Res*. 2015;45:153–69.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: De Guire E, Bartolo L, Brindle R, et al. Data-driven glass/ceramic science research: Insights from the glass and ceramic and data science/informatics communities. *J Am Ceram Soc*. 2019;102:6385–6406. <https://doi.org/10.1111/jace.16677>