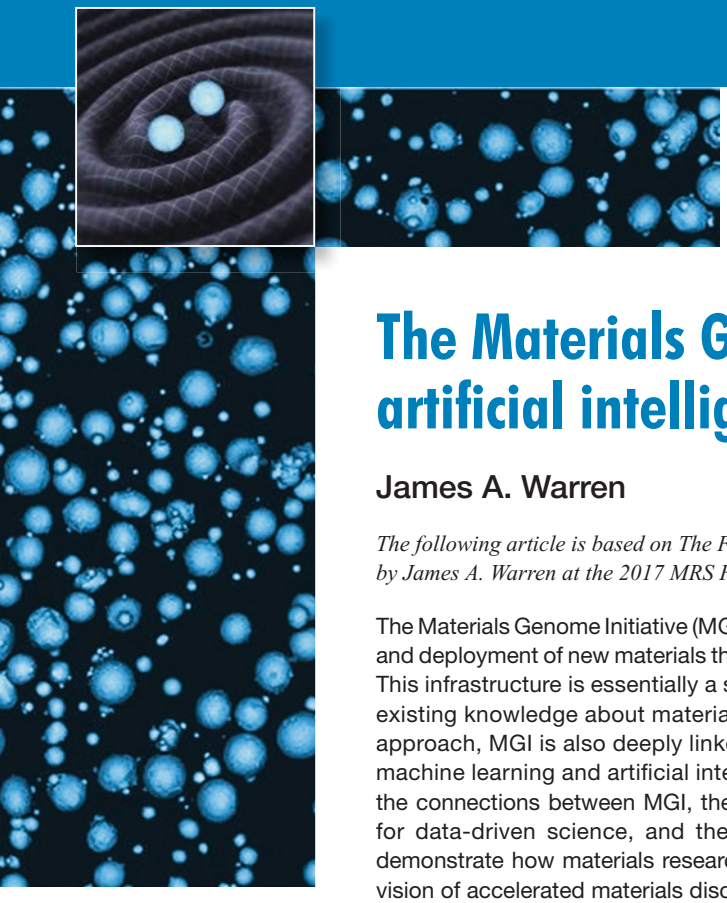


The Materials Genome Initiative and artificial intelligence

James A. Warren

The following article is based on The Fred Kavli Distinguished Lectureship in Materials Science given by James A. Warren at the 2017 MRS Fall Meeting Plenary Session in Boston, Mass.

The Materials Genome Initiative (MGI) seeks to accelerate the discovery, design, development, and deployment of new materials through the creation of a materials innovation infrastructure. This infrastructure is essentially a system for providing data and tools that encapsulate our existing knowledge about materials, and the means to create new knowledge. Given this approach, MGI is also deeply linked to the ongoing exponential growth in applications of machine learning and artificial intelligence (AI) to materials research. This article explores the connections between MGI, the consequent need for data publication, the implications for data-driven science, and the application of AI to materials design. Examples will demonstrate how materials research is transforming in remarkable ways, and that the MGI vision of accelerated materials discovery is within reach.

The Materials Genome Initiative

There is a profound link between the Materials Genome Initiative (MGI) and artificial intelligence (AI). To demonstrate this, we must first define what is meant by each. MGI is a federal "initiative," but what does that really mean? When the US White House announces an initiative (e.g., the Brain Initiative, the National Nanotechnology Initiative, or the Materials Genome Initiative), it's simply a statement by the US President that something is a government priority, and that relevant federal agencies should work together to address that issue. While each agency pursues their own mission, the initiative connects and harmonizes these efforts toward something greater. In the case of the MGI,[1] there has been active participation and coordination by the Department of Energy, the National Science Foundation, the National Institute of Standards and Technology (NIST), the Department of Defense (all the service branches), as well as NASA, National Nuclear Security Administration, the Food and Drug Administration, the National Institutes of Health, and the US Geological Survey.

The other term, AI, is used here in a less than rigorous manner to define a computational technique (e.g., machine learning, deep neural nets, convolutional neural nets) to discover correlations in high dimensional data sets. We will explore the connection after a tour through the MGI and the data infrastructure.

In 2011, the MGI was announced by the Obama administration[2] with a goal of decreasing by 50% the time to market of a new material by accelerating its discovery, design, development, and deployment—all at a lower cost. This initiative sought to achieve these goals through the creation of a materials innovation infrastructure (MII), as shown in **Figure 1.** The MII (comprised of computation, experiment, and digital data) is the substrate for achieving the desirable ends detailed in the outer ring of the graphic, such as clean energy, healthier people, and enhanced security.

The final (bottom right) quarter of the outer ring, next-generation workforce, is different, as it is both a means to MGI ends as well as its own desirable outcome. Much of MII will be open (or accessible to all who can pay), and thus accessible to the world, in the same manner as most scientific literature, but can only provide an advantage to those who can use it. The national competitive edge derives from having the sophisticated workforce that can take advantage of this infrastructure.

MGI does not just foster the discovery of new materials; it's more about developing sufficient knowledge to design materials, and then ensure their subsequent deployment by optimizing desired properties along the manufacturing continuum. The MII—computational tools, experimental tools, and digital data—are really just the elements of science and engineering. What's profound is not these elements,

James A. Warren, Materials Genome Program, National Institute of Standards and Technology, USA; james.warren@nist.gov
doi:10.1557/mrs.2018.122

© 2018 Materials Research Society
https://doi.org/10.1557/mrs.2018.122 Published online by Cambridge University Press

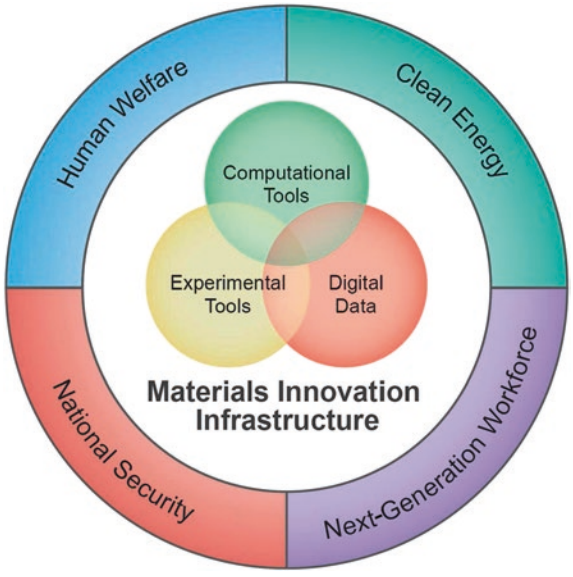


Figure 1. The Materials Genome Initiative: Venn diagram of the Materials Innovation Infrastructure.

but their intersections. The MGI seeks to integrate these elements as tightly as possible.

Thus, the MGI, in no small way, begins with the classic multiscale modeling paradigm that has been batted around for at least 20 or 30 years (depending on which community you're coming from). The scales move from the quantum level out to the nano-, micro-, and macrolevels. There is actually a lot of supposition that goes into such a picture. These models are imperfect, requiring numerous inputs to be useful. The hierarchy of models must be fed with information to actually have any chance of them working. This information includes data from experiments, as well as data from lower-level, trusted models. The key is deciding which information to use, and which models are best. Once you've made those choices, you then need to exercise this model chain, and hopefully create a designed material with targeted properties. Even more daunting, the researcher would really like to exercise the inverse problem, where the properties are specified and the knowledge system informs you how to make the material. Understanding that this is our ultimate goal, the MII can be framed as an integrated system to create and capture knowledge about materials, and make that knowledge useful to a broad audience of materials designers.

A classic example

Before discussing how MGI and AI have been linked, let's examine the pre-AI state of the art in MGI approaches. An example of that comes from Lass and co-workers[3] at NIST who collaborated with the US Mint on a new alloy for the 5-cent coin. The goal was to design a new alloy that had less nickel and copper than the current alloy (which is expensive), while preserving the look, corrosion properties, mechanical properties, and electrical conductivity. The NIST-Mint Team employed a "classic" materials-by-design approach, which has its roots in techniques pioneered by G. Olson.[4] One assembles an Olson diagram as shown in **Figure 2**, relating the processing, microstructure, and properties. Note that all of the science and engineering is embedded in the arrows of Figure 2. If there are sufficiently well-parameterized models, one can then work backward toward the materials processing and determine the optimum process to create a new material. For this study, the crucial quantitative exercise was creating composition-dependent models for electrical conductivity and color (yellowness), phase equilibria, and, of course, cost, in the quaternary system (Cu-Ni-Zn-Mn). This is an excellent example of integrated computational materials engineering (ICME),[5] a key concept in the founding MGI white paper,[2] and remains an enormously important modality for materials discovery and design.

Datafication

MGI seeks to improve our ability to design and deploy new materials. To accomplish this, we need better data and models. Thus, MGI evolves with the recognition that improvements in computational power, coupled with the disruptive potential of the Internet, can further accelerate materials design.

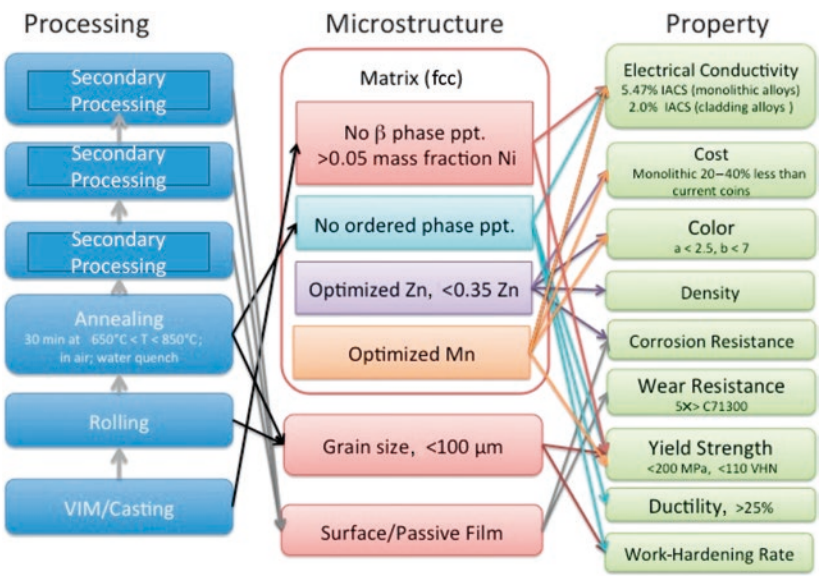


Figure 2. An example of an Olson diagram.[3] In this case, Lass and co-workers, using best-practice integrated computational materials engineering principles, designed a new quaternary (Cu-Ni-Zn-Mn) alloy that preserves the properties of the existing US 5-cent coin while lowering costs. Note: VIM, vacuum induction melting; fcc, face-centered cubic; ppt., precipitation; IACS, International Annealed Copper Standard; VHN, Vickers hardness.

https://doi.org/10.1557/mrs.2018.122 Published online by Cambridge University Press

Understanding data as the raw material, we now ask, how do we feed this infrastructure with the data it needs to make better predictions? One approach is to perform a large number of computations and then mine that information, which is the established operation of a number of density functional theory (DFT) database projects.

The MGI was partially inspired, both in approach and in name, by the Materials Genome Project, now renamed The Materials Project to avoid confusion.[6] The Materials Project has a rather simple idea at its core—to do as many DFT calculations as possible, over a range of material types, and then mine those results in a database and perform analytics to discover materials with sought-after properties. The database was viewed as a "materials genome," and there the metaphor for MGI was born.* There are a number of other efforts using DFT databasing, including Automatic Flow for Materials Discovery (Duke University), the Open Quantum Materials Database (Northwestern University), The Novel Materials Discovery (NOMAD) Laboratory (EU), and the Automated Interactive Infrastructure and Database (École Polytechnique Fédérale de Lausanne). Regardless of the specific DFT method used, these databases demonstrate that a sufficient quantity of well-curated data is a doorway to materials discovery.

Beyond computing using trusted models, another approach to getting the required data is via high-throughput experimental methods (which we will touch upon at the end of this article). First, we consider how we might get data from "everywhere." This process requires that one discover the needed data and then mine it. Thus, to be successful, one must be able to mine relevant literature and mine the (currently scant) published data. Optimally, data publication would be more common, and viewed as a community best practice. Thus, lurking in the background of the MGI is a necessary rethinking of how we publish materials research. The existing publication paradigm is an accident of history. If you were designing the scientific publication system today instead of letting it evolve over more than 500 years, it a might be a little bit different, optimized toward better scientific outcomes.

Data infrastructure

Pursuing these notions, the US NSF (National Science Foundation) funded a study[7] of the state of the materials data innovation infrastructure, and what it will take to get it to a level where the goals of MGI can be achieved. While there has been a great deal of focus on "open data," and the social and technical benefits of open data, these benefits do not directly accrue to the person generating and sharing the data. That is, there is an incentive problem. One can argue that there are immediate benefits to research in better data management, but only at the cost of a change in the research workflow. Research reproducibility and continuity are enhanced, while data loss is minimized. However, it is most likely that when the tools become sufficiently good, researchers will be curating and sharing their data by "accident." Data will be placed on a "platform" because it has a tool that enables better science to be done, and data curation and sharing emerge as a positive side effect.

In pursuit of this infrastructure, inspiration has come in particular from the astronomy and astrophysics community, which has been enormously successful in developing a data infrastructure. Essentially, all major telescopes are hooked into the virtual astronomical observatory[8] and if you use the infrastructure to do a search on any point in the sky, you can get all of the known data—radio frequencies, optical frequencies, x-ray frequencies—about that point in the sky for a specified solid angle. Can this infrastructure model be copied for materials? The problem is significantly more complicated for the field of materials than astronomy, with processing and synthesis variables, and a host of different measurement types (not to mention commercial, proprietary, and export-restricted data). Yet, this is a matter of scope. The task is big, but not impossible. To begin, the community is focusing on a small number of computational and experimental techniques. One area of considerable promise is data from photon and neutron sources. Beamline data are expensive to obtain, and growing exponentially, as detectors' data rates increase following a Moore's Law-type curve. Beamlines are the "big science" of our time and that's largely driven by materials research.

To meet the needs of the community to better curate materials data, NIST is developing open-source software that is highly configurable, so that it can accommodate the rich diversity of materials data. This tool, the Materials Data Curation System[9] is used in tandem with a "schema" that describes the specifics of the data. A library of these schemas is also maintained so that one need not start from scratch when developing a data capture/creation protocol. If the work to be curated is not in the library of schemas, then NIST can provide expertise to quickly develop a new schema.

One of the first problems facing researchers is finding the relevant existing data and expertise so that research can proceed from the existing state of the art. To address this, NIST has created software called the NIST Materials Resource Registry (NMRR).[10] The NMRR allows discovery of materials databases, software, institutions, and application programming interfaces (APIs) relevant to materials research. In addition, the NMRR software federates diverse sources (i.e., geographically disparate institutions can populate their own instances of the NMRR, and each instance harvests the metadata from the other instances, yielding a global search infrastructure). While the initial and primary use of the NMRR is information discovery, the ability to use this search to discover APIs will ultimately drive development of the NMRR, converting it into a tool that allows a researcher to go directly from discovery to analysis.

Footnote

*In 2002, independently of the Materials Genome Project (MGI), Zi-Kui Liu trademarked this name. He kindly agreed to allow the MGI effort to use the name.

https://doi.org/10.1557/mrs.2018.122 Published online by Cambridge University Press

Data fusion

The infrastructures previously described are but a few of the numerous efforts under way, of which many are in the mode of researchers seeking solutions, and most are likely to fail. Success in the long run is inevitable, as the promise of the data infrastructure is the promise of better, faster research. Scientists will upload their data to a service so they can use tools on that service to do new science. If other researchers have done the same, the fusion of these data sets can drive further discovery and higher quality results. An example of such a mode of research has been developed at the Materials Data Facility (MDF). The MDF is based at the Center for Hierarchical Materials Design (CHiMaD), a NIST Center of Excellence, and a collaboration of Northwestern University, The University of Chicago, and Argonne National Laboratory. The MDF has access to a wide range of materials data resources and has done "deep indexing" on these resources to allow data fusion. L. Ward, a researcher at the MDF, has developed a nice example (**Figure 3**), where three DFT results, one from the University of Connecticut (UConn)[11] and two data sets hosted in a NIST repository are combined.[12] The figure shows the mean absolute error (MAE) of machine learning models trained on these sets. Each model was trained to predict the force acting on an aluminum atom, and its accuracy was measured against an original test set (Figure 3a) and training set (Figure 3b) used by Botu et al. in their 2017 paper.[11] Ward and his collaborators first only used the original data from Botu et al. and were able to replicate the accuracies of that work, but then added, successively, two data sets from the NIST Materials Data Repository. By adding the NIST data into the training set, they increased the accuracy of the model by 30%. While this is not surprising, our ability to do such a data fusion is nearly impossible without the right infrastructure. This is what all research should be like, where new research data are added into the sum of existing knowledge, thereby multiplying the knowledge in the field.

AI and machine learning

Finally, we return to the question, "What does AI have to do with MGI?" Since 2012, data-driven materials science has been the much sought-after prize of a working MGI infrastructure. But this idea was not widely discussed as it seemed a far-off goal. Now, with the basics of the infrastructure taking shape, and the ongoing revolution in AI, much has changed.

There has been an absolute explosion in the amount of work conducted in this area, and the MGI has had, and will continue to have, an important role in enabling AI approaches to materials design. A casual literature search will demonstrate that "materials informatics" and "machine learning" are now hot topics in materials research. Where is the link to MGI?

As previously stated, MGI seeks to build systems that can create and capture knowledge about materials, and then make both the system and knowledge useful to a broad audience of materials designers. To do that, we must construct models that encapsulate large amounts of information about nature in a compact form. We cannot analyze the physical world without models. A recent example is the Laser Interferometer Gravitational-Wave Observatory (LIGO). In 2016, LIGO measured the collision of two black holes 1.3 billion light years from us.[13] This required measurements of strain of approximately 10^{-21}, which seems positively absurd to a materials scientist. When a gravitational wave passes through the detector, it yields a specific pattern of length changes. To make sense of the signal, that pattern must be compared with a supercomputer simulation of the collision of two black holes 1.3 billion years away. That is, Einstein's theory of gravity must be solved in this terrifically complex case, and that, in turn, enables the measurement (and further confirms the theory). So far, Einstein's equations have held up nicely.

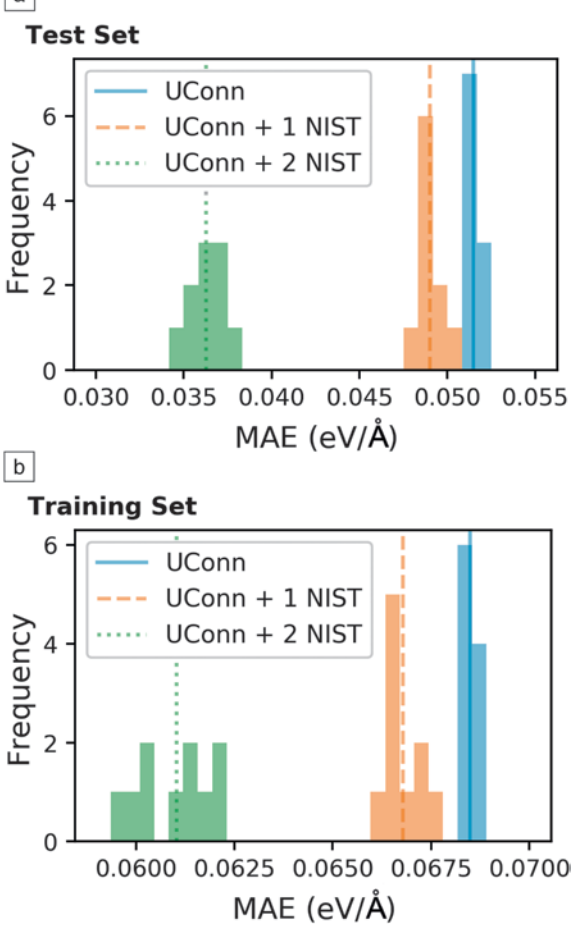


Figure 3. Mean absolute error (MAE) of machine learning models trained on three different training sets. Each model was trained to predict the force acting on an aluminum atom using the methods of Botu et al.,[11] and its accuracy was measured against the original (a) test set and (b) training set. This chart shows the histogram of the error for 10 repetitions using only the original data (blue), and the original data combined with one or two data sets from the National Institute of Standards and Technology (NIST) Materials Data Repository (orange, green).[12] Vertical lines show the mean of 10 repetitions.

https://doi.org/10.1557/mrs.2018.122 Published online by Cambridge University Press

What about the more typical situation when we have little in the way of a good quantitative model? For example, biology has such fantastically complex systems that it often seems beyond hope that physical models will ever be found, or if found, then they may not be soluble. There are many materials systems which are also so complex that it may not be practical to use physics-based models to quantify materials behavior. Hence, we can expect to learn a great deal from the approaches taken by biologists. Just what do we need to do and how do we find the middle ground where we can apply data-driven methods alongside traditional physics-based models?

If one reflects on the situation, it is clear that AI is simply a method for creating models. You have some descriptors (your inputs), and predictions (your outputs). AI often (sometimes purposefully) feels like a black box, depending on the approach, but it boils down to a method linking these inputs and outputs. AI might also be viewed as a method for efficient coarse graining, allowing for linking of a short-length-scale phenomenon with more macroscopic properties, without necessarily reckoning with intermediate scales.

Similarly, if one thinks about the role of physics, it's really just an approach to generate quantitative models of reality. What separates traditional physics from AI is the presumption of causality. AI is inherently about correlations. However, this is where those of us working in the physical sciences have an edge over fields that lack physics-based models. We can exploit the underlying ground truths (causes) to validate and enhance our AI approaches. It is the fusion of these two methods that really makes AI of such potential interest for materials discovery and design.

AI and materials characterization

Having begun the challenging task of enabling the discovery and interoperation of materials data, we are now beginning to see the payoff in applications of AI to materials research. Currently, success is often achieved via significant custom coding and intimate understanding of the data, as opposed to a more automatic integration that the data infrastructure promises. Five years ago, the application of AI to materials design was viewed as a far off, if not an implausible, vision. This is no longer the case.

DeCost and co-workers[14] have applied machine-learning techniques to the problem of classifying alloy powders, that are used as feedstock for additive manufacturing. The ability of such algorithms to detect patterns (such as human faces) is now central to many useful tools in our daily lives. Taking micrographs of powders as a starting point, defining a number of descriptors (various short length scale descriptions of the local geometry), the researchers used methods coming from the field of computer vision to detect these feature descriptors and correlate specific alloy patterns with specific feature sets. This type of likeness is analogous to how people analyze microstructure, but such algorithms can find correlations that humans would be unlikely to identify (and our descriptors will often be coarser). Sometimes the correlations are spurious, but sufficient algorithm training and cross-validation improves the reliability of these methods. In this case, the researchers were able to accurately identify the type of metal powder 95% of the time, which is better than a human trained to do the same thing. As characterization of structure is at the heart of materials science, AI promises (and has already proven) to be a powerful aid in discovery.

AI, materials discovery, and autonomous science

A final example is the work of A. Gilad Kusne and co-workers at NIST, as well as collaborators at the SLAC National Accelerator Laboratory and the University of Maryland.[15] The work uses AI methods to rapidly measure experimental ternary phase diagrams. A wafer is covered, via sputtering, with the three components of interest in a "Gibbs Triangle" of material, where each corner of the triangle is a pure element, the edges are the binary alloys, and the center is equal percentages of each component. The samples are prepared in a way that attempts to ensure they are in the equilibrium crystal structure (for a specified temperature) for each composition. The goal then is to find the phase changes in this "library" of compositions to identify where the crystal structure changes. We can do this using AI.

A bright x-ray beam is rastered (this was done at SLAC) over the sample, and AI is employed to detect when the pattern changes sufficiently. It should be noted that this is image recognition again, albeit this time of a diffraction pattern. We don't actually need the crystal structures, just when the pattern changes, which must be the phase boundary. This was successful—as one can compare against the known phase diagram (determined using more arduous methods), the approach can be validated, and subsequently applied to as yet unexplored systems. This is classic combinatorial science (or high throughput), a method that has gained new life, because machine learning allows for rapid analysis of the library.

Kusne et al. are now taking AI to the "next level," after recognizing that rastering the beam over the entire sample is inefficient. Bayesian-style approaches are being employed to decide which next measurement maximizes an increase in knowledge (this follows the groundbreaking work of Lookman and colleagues[16]). The beam now goes from sampling the whole space to hopping about, and nails down the phase boundaries in an efficient way.

This can remarkably go even one step further. Once we realize that we can increase knowledge by measuring in the "best" places, we should also put AI in charge of making samples. We can synthesize the sample that AI indicates as giving the greatest likelihood of increasing knowledge, then measure that sample and repeat. Additionally, because AI is tireless, one can use more components and allow for large variations in processing techniques. This is autonomous experimentation, and it is the future of materials science.

Conclusion

This article discussed how the goals of MGI—particularly the central importance of model construction to the practice of

https://doi.org/10.1557/mrs.2018.122 Published online by Cambridge University Press

science and engineering and the development of the materials data infrastructure—have both prepared the way for AI and materials, and provided the intellectual framework to accelerate the discovery, design, development, and deployment of materials. MGI has helped to frame both technological and policy issues around opportunities to accelerate materials research, and will continue to do so until these ideas are so ingrained that the concepts will be invisible. Finally, it seems certain that when AI-powered MGI approaches are coupled to the ongoing democratization of manufacturing methods, the consequences for society and the economy will be profound.[17]

Acknowledgments

Thanks to E. Lass, B. DeCost, L. Holm, L. Ward, A. Gilad Kusne, J. Hattrick-Simpers, and my colleagues at NIST and the CHiMaD.

References

1. https://mgi.nist.gov.
2. https://www.mgi.gov.
3. E. Lass, M. Stoudt, C.E. Campbell (forthcoming).
4. G. Olson, *Science* **277** (5330), 1237 (1997).
5. National Research Council, "Integrated Computational Materials Engineering: A Transformational Discipline for Improved Competitiveness and National Security" (National Academies Press, Washington, DC, 2008).
6. https://materialsproject.org.
7. The Minerals, Metals & Materials Society (TMS), "Building a Materials Data Infrastructure: Opening New Pathways to Discovery and Innovation in Science and Engineering" (TMS, Pittsburgh, 2017).
8. C.H. Ward, J.A. Warren, R.J. Hanisch, *Integr. Mater. Manuf. Innov.* **3** (22), (2014).
9. https://mgi.nist.gov/materials-data-curation-system.
10. https://mgi.nist.gov/materials-resource-registry.
11. V. Botu, R. Batra, J. Chapman, R. Ramprasad, *J. Phys. Chem. C* **121** (1), 511 (2017).
12. https://materialsdata.nist.gov.
13. B.P. Abbott et al., *Phys. Rev. Lett.* **116**, 061102 (2016).
14. B.L. DeCost, H. Jain, A.D. Rollett, E.A. Holm, *JOM* **69**, 456 (2016).
15. J.R. Hattrick-Simpers, J.M. Gregoire, A.G. Kusne, *APL Mater.* **4**, 053211 (2016).
16. D. Xue, P.V. Balachandran, J. Hogden, J. Theiler, D. Xue, T. Lookman, *Nat. Commun.* **7**, 11241 (2016).
17. N. Gershenfeld, A. Gershenfeld, J. Cutcher-Gershenfeld, *Designing Reality* (Basic Books, New York, 2017). □



James A. Warren is the director of the Materials Genome Program at the National Institute of Standards and Technology (NIST). He received his AB degree in physics from Dartmouth College and his PhD degree in theoretical physics from the University of California, Santa Barbara. He has been at NIST since 1992, first as a National Research Council postdoctoral researcher, (1995) then as co-founder and director (2001) of the NIST Center for Theoretical and Computational Materials Science. From 2005 to 2013, he was the leader of the Thermodynamics and Kinetics Group. His research focuses on computational models of solidification, grain structures, creep, and wetting. In 2010, he was an Ad Hoc Committee member within the National Science and Technology Council (NSTC) that crafted the Materials Genome Initiative (MGI). He has served as the executive secretary of the NSTC MGI Subcommittee since 2012. Warren can be reached by email at james.warren@nist.gov.



https://doi.org/10.1557/mrs.2018.122 Published online by Cambridge University Press