ADVANCED SCIENCE Open Access
www.advancedscience.com

Data-Driven Materials Science: Status, Challenges, and Perspectives

*Lauri Himanen, Amber Geurts, Adam Stuart Foster, and Patrick Rinke**

Data-driven science is heralded as a new paradigm in materials science. In this field, data is the new resource, and knowledge is extracted from materials datasets that are too big or complex for traditional human reasoning—typically with the intent to discover new or improved materials or materials phenomena. Multiple factors, including the open science movement, national funding, and progress in information technology, have fueled its development. Such related tools as materials databases, machine learning, and high-throughput methods are now established as parts of the materials research toolset. However, there are a variety of challenges that impede progress in data-driven materials science: data veracity, integration of experimental and computational data, data longevity, standardization, and the gap between industrial interests and academic efforts. In this perspective article, the historical development and current state of data-driven materials science, building from the early evolution of open science to the rapid expansion of materials data infrastructures are discussed. Key successes and challenges so far are also reviewed, providing a perspective on the future development of the field.

1. Introduction

In this perspective article, we review the current state of data-driven materials science with a focus on materials data infrastructures. Data-driven invokes associations with big data, data management, open data and artificial intelligence (e.g., machine learning). The public debate of these terms is currently dominated by internet giants like Google, Amazon, and Facebook who also lead the technological development of data infrastructures, algorithms, and analysis tools. Compared to these e-commerce and social media developments, the field of data-driven materials science is still under construction. By way of analogy, it is nonetheless still instructive to imagine a Materials "Google"—the Materials Ultimate Search Engine (MUSE). In this article, we address what it takes to develop such a search tool for materials.

Materials science, the study of the characteristics and applications of materials is a well established discipline that combines chemistry, physics, and engineering research. Materials scientists frequently dream of designing new materials from scratch for use in society.[1] However, instead of finding new materials using the MUSE, they discover new materials through conventional experimental, theoretical, or computational research (see left panel of **Figure 1**). This pipeline through which new materials are discovered, designed, developed, manufactured, and deployed remains slow, costly, and highly inefficient: By the time a new material comes to market, the patent protection of the original invention is at the end of its tenure, and proprietary advantage is lost[2] (see also ref. [3]). By applying data science to materials research, we now have a way to accelerate the materials value chain from discovery to deployment.

Data science has developed out of the growing demand for open science combined with the meteoric rise of AI and machine learning. As these innovative technologies allow ever-larger datasets to be processed and hidden correlations to be unveiled, data-driven science is emerging as the fourth scientific paradigm[4,5] (cf. Figure 1) following the first three eras of experimentally, theoretically, and computationally propelled scientific discoveries. Often connected to the fourth industrial revolution[6] or the second machine age,[7] such data-driven approaches permeate science, business, politics, and even social life. Since materials innovation is a critical, well-recognized driver of economic development and societal

L. Himanen, Dr. A. Geurts, Prof. A. S. Foster, Prof. P. Rinke
Department of Applied Physics
Aalto University
P.O. Box 11100, 00076 Aalto, Espoo, Finland
E-mail: patrick.rinke@aalto.fi
Dr. A. Geurts
Department of Management Studies
Aalto University
P.O. Box 11100, 00076 Aalto, Espoo, Finland
Dr. A. Geurts
TNO, Netherlands Organization for Applied Scientific Research
Expertise Center for Strategy and Policy
Anna van Beurenplein 1, DA 2595 The Hague, Netherlands
Prof. A. S. Foster
Graduate School Materials Science in Mainz
Staudinger Weg 9, 55128 Mainz, Germany
Prof. A. S. Foster
WPI Nano Life Science Institute (WPI-NanoLSI)
Kanazawa University
Kakuma-machi, Kanazawa 920-1192, Japan
Prof. P. Rinke
Theoretical Chemistry and Catalysis Research Centre
Technische Universität München
Lichtenbergstr. 4, D-85747 Garching, Germany
The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/advs.201900808.

DOI: 10.1002/advs.201900808

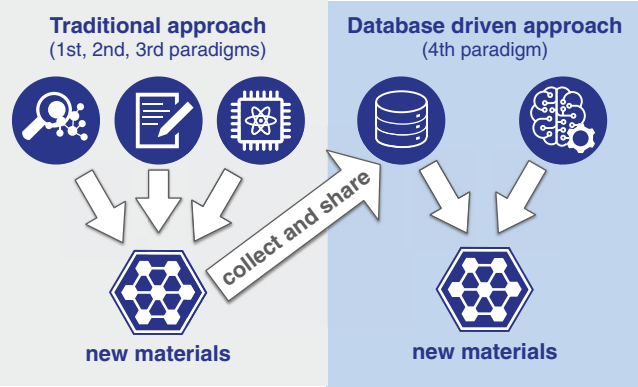

Figure 1. Materials discovery schematic. In the traditional approach, new materials are discovered by experimentation, theory, or computation (also referred to as 1st, 2nd, and 3rd paradigms and symbolized by the three icons at the top of the left panel). In the 4th paradigm of data-driven materials science, available data is gathered in data infrastructures, and machine learning approaches discover new materials.

progress, it is important that new trends, such as data science, are embraced if they have the potential to advance the field.

Data-driven materials science and materials informatics are umbrella terms for the scientific practice of systematically extracting knowledge from materials datasets. This practice differs from traditional scientific approaches in materials research by the volume of processed data and the more automated way information is extracted (cf. Figure 1), for example, through the use of machine learning (see refs. [5,8–23] for recent review articles on machine learning in materials science). In our MUSE analogy, this would be the search and find part. In addition to data processing and data analysis tools, data-driven materials science also requires physical infrastructures that host and preserve that data. These would be the data storage part of our MUSE example, which, as physical infrastructures, require dedicated community efforts and sustained investment to become and remain operational.

Stakeholders in academia, industry, governments, and the public attach different meanings and expectations to data-driven materials science. The actual material science is carried out in academia and research and development (R&D) departments in industry. Scientists at universities and companies not only produce materials data that could then be stored in data facilities, they are also the primary user group of materials data infrastructures. In the wake of digitalization, industry has a further interest in digitizing materials data and incorporating data-driven materials science into their value chain. Policy makers and governmental or private funding agencies may have an interest in promoting open science data and can stir scientific developments through policy and funding decisions. The general public benefits from materials science by quality-of-life enhancement through new products and technologies. They have an indirect interest in data-driven materials science as a means to accelerate innovations and follow developments in science and open data in the media. Together these stakeholders form an ecosystem of mutual benefit. The vitality of this ecosystem is crucial for the success and the longevity of data-driven materials science.

In this article, we embed our perspective in the emerging field of data-driven materials science in the context of the open science movement, which has shaped the philosophy and design of several materials science data infrastructures. We discuss how these infrastructures grew historically from simple databases into data centers that then progressed into materials discovery platforms, and we detail the current state of data infrastructure. A list of current challenges provides the gateway to the second part of this article, in which we delve deeper into data organization, acquisition, quality, and machine learning. We conclude with an industrial perspective that addresses the future and longevity of materials data infrastructures.

2. Open Science Movement

Many of the fundamental aspects of data-driven materials science are built upon the key elements of the Open Science movement. The European Commission outlines[24] Open Science as "...a new approach to the scientific process based on cooperative work and new ways of diffusing knowledge by using digital technologies and new collaborative tools." Here we reflect on those aspects of Open Science that are particularly relevant to the birth and future of data-driven materials science.

Openness in science was initially curtailed by the prestige wars between the patrons of early scientists and their associated, convoluted encryption schemes.[25] Once more professional scientific practice developed, scientists embraced the idea of accessibility of research as a cornerstone of progress, and this has been generally mandated by public policy. As early as 1710 in the UK, the Copyright Act endowed the ownership of copyright to authors rather than publishers, encouraging authors to deposit manuscripts into national libraries to make them publicly accessible. In addition to accessibility, public accountability and scientific reproducibility have remained powerful driving forces in the way science has been conducted and disseminated, and significant deviations from these norms are of great concern to the community.[26] More recently in the 1990s, the development of the internet transformed this debate as it became possible to make nearly all aspects of the scientific research process easily accessible, from preliminary data to final publications. While arguments over fair allocation of rewards for scientific achievement versus full and early research dissemination remain challenging,[27] and intellectual property management regularly introduces conflicts,[28] the era of Open Science[29,30] (or indeed Open Innovation[31]) is here to stay and contributes to scientific advancement overall.[32] Open-access journals and data, and open-source software have significant impacts on the Open Science movement.

2.1. The Rise of Open Access Publishing

Building on the foundations of the very first online journals, websites like arXiv (established in 1991) took the first steps in providing Open Access to scientific publications. As more content became available online, and the need for physical copies of journals in libraries rapidly diminished, many

expected a significant reduction in the cost of journal subscriptions. When this did not happen, it catalyzed the Open Access movement and other alternatives to conventional scientific publishing practices. At present, over 50% of newly published articles are Open Access, and conservative estimates place achievement of complete Open Access by 2040.[33,34] Current Open Access approaches tend to fall into two classes (or hybrids thereof[34]): gold, where the article is freely available at the point of publication; and green, where the authors can deposit the article in a public repository, for example, at their home institution. Some publishers require an embargo period before deposition in a public repository, but there is little evidence in terms of publisher income to support the existence of such embargoes.[27] Many funding agencies have embraced Open Access publishing as a way to improve public transparency and accountability, and these agencies have made it a condition for support—this includes all European Union funding for 2020 and beyond.[35] As such, Open Access is at the heart of the Open Science movement and certainly overlaps with one of the critical developments in data-driven science, Open Data.

2.2. Open Access Data

The initiative to make data Open Access can be traced to efforts to establish scientific global data centers in the 1950s,[36] largely as a way to store data long term and make it internationally accessible—all data was fully available for the cost of printing and delivery. Following this change, demands for scientific data sharing continued to rise, especially after the development of the internet and the tantalizing prospect of easy upload and download of data globally.

While many scientists were quick to embrace this, it took a decade for Open Data to appear as a clear objective and topic for scientific policy. In 2004, science ministers of most developed countries signed an agreement that all publicly funded archive data should be made available, with the guidelines for this following in 2007.[37] As is often the case, the scientific communities themselves were ahead of policy changes, and many bespoke scientific databases had already proliferated, providing data repositories in almost every field across the globe. There are now thousands of them, and finding useful ways to search for a relevant repository, let alone data within it, requires serious effort.[38]

Motivation to make this effort is increasing rapidly, with many journals and funding agencies demanding the availability of data tied to publication or grants. Contributing to many aspects of the Open Science initiative, the Public Library of Science[39] has pioneered this development, with a clear policy on data sharing for its publications and likely rejection if policies are not followed. Other major publishers have also been active, with at least the creation of specific Open Data journals,[40] policies,[41] and collaboration with Open Data initiatives.[42] Many funding agencies now insist on a data management plan with all submissions, and this plan must give a detailed account of how data will be stored, secured, and shared—with particular attention to the Open Science rules of the agency in question.

In an attempt to provide unifying guidelines for the widely varying groups interested in Open Data and to aid in data management development, the FAIR Data Principles were established.[43,44] These principles have been adopted by several major players in global data management (see **Table 1**). The ideas behind making data searchable, accessible, flexible, and reusable at the core of FAIR are also the concepts that make the power of data-driven science actually attainable.

2.3. Open-Source Software for Science

The development of open-source software entails the final element of the Open Science movement. Its development started in parallel with the earliest computing hardware efforts, with nearly all software freely available in the public domain as part of large academic and corporate collaborations. Since the relative cost of software compared to hardware has increased, this openness began to steadily decline until the early 1980s with the launch of the GNU project and the parallel explosion of Linux and the internet in the early 1990s. This provided a powerful platform and toolset for the collaborative development of software that could then be freely downloaded, culminating in the active open-source movement in 1997.[45] In particular, it suited the kind of focused, rapidly changing software that characterizes nearly all scientific applications.

In 2005, the creation (by Linus Torvalds) and rapid adoption of Git as a distributed revision control system, closely followed by hosting site GitHub, put the seal on the standard approach for open-source scientific software development that remains to this day. It became possible to manage updates to codes from a large development team, while providing a platform for feedback, bug notification, and feature requests from users. It is now possible to find Open Source software for nearly every aspect of a scientific project,[46] from electronic lab notebooks,[47] experimental toolsets,[48] and simulation packages,[49] to machine learning libraries[50] and online collaborative writing sites.[51] With freely accessible data, Open Access publications explaining the science behind it, and a wealth of open-source software to mine it, the way is clear for innovative data-driven science.

3. Materials Data Infrastructures

Having established the context for Open Science, we next review the emergence of materials data infrastructures that collect, host, and provide materials data to stakeholders. We first reflect on early digital materials infrastructures before discussing the current state.

3.1. Development of Materials Infrastructures

The increasing capabilities of first-principles methods—and the increasing capabilities of computational science in general—have accelerated materials researchers interest for new, computer-based pathways to materials discovery and design—better, faster, and cheaper than ever before. Perhaps

Table 1. List of current major materials data infrastructures. The entries are divided into non-commercial (top) and commercial (bottom). Note that some platforms are named after the leading research project and may host multiple services under different names. As contact person we listed the director(s) of each infrastructure, in such cases, where they were clearly identifiable. Data volume numbers reflect the state in April 2019.

Name	Website	Contact	Overview	Ref.
AFLOW	aflowlib.org	Stefano Curtarolo, Duke University	Computational data consisting of 2 118 033 material compounds and 281 698 389 calculated properties with focus on inorganic crystal structures. Incorporates multiple computational modules for automating high-throughput first principles calculations.	[83,91]
Computational Materials Repository	cmr.fysik.dtu.dk	Kristian Thygesen and Karsten Jacobsen, DTU	Computational datasets from a diverse set of applications. Data creation and analysis with the Atomic Simulation Environment (ASE).	[92–94]
Crystallography Open Database	crystallography.net		Open-access collection of crystal structures of organic, inorganic, metal–organic compounds and minerals, excluding biopolymers.	[95,96]
HTEM	htem.nrel.gov	Caleb Phillips and Andriy Zakutayev, NREL	Properties of thin films synthesized using combinatorial methods. Contains 57 597 thin film samples, across a wide range of materials (oxides, nitrides, sulfide, intermetallics).	[97,98]
Khazana	khazana.gatech.edu	Rampi Ramprasad, Georgia Institute of Technology	Platform to store structure and property data created by atomistic simulations, and tools to design materials by learning from the data. Tools include Polymer Genome and AGNI.	[99–101]
MARVEL NCCR	nccr-marvel.ch	Nicola Marzari, EPFL	Materials informatics platform for data-driven high-throughput quantum simulations. Data available at materialscloud.org, powered by the AiiDA-infrastructure.	[85]
Materials Data Facility (MDF)	materialsdatafacility.org	Ben Blaiszik and Ian Foster, University of Chicago	Data publication network for computational and experimental datasets. Data exploration through the Forge python package.	[102,103]
Materials Project	materialsproject.org	Kristin Persson, LBNL	Online platform for materials exploration containing data of 86 680 inorganic compounds, 21 954 molecules and 530 243 nanoporous materials. Develops various open-source software libraries, including pymatgen, custodian, FireWorks, and atomate.	[84,104]
MatNavi/NIMS	mits.nims.go.jp	Yibin Xu, NIMS	An integrated material database system comprising ten databases, four application systems and the NIMS Structural Datasheet Online.	[105]
NOMAD CoE	nomad-coe.eu	Matthias Scheffler, FHI/ Max Planck Society	Provides storage for full input and output files of all important computational materials science codes, with multiple big-data services built on top. Contains over 50 236 539 total energy calculations.	[106,107]
Organic Materials Database	omdb.mathub.io	Alexander Balatsky, Nordita	Open access electronic structure database for 3-dimensional organic crystals. Contains approximately 24 000 materials.	[108,109]
Open Quantum Materials Database	oqmd.org	Chris Wolverton, Northwestern University	Database of DFT-calculated thermodynamic and structural properties with focus on inorganic crystal structures. Contains 563 247 entries with support for full download and advanced usage through the qmpy python package.	[90,110]
Open Materials Database	openmaterialsdb.se	Rickard Armiento, Linköping University	Computational database primarily based on structures from the Crystallography Open Database. Data creation and analysis with High-Throughput Toolkit (httk).	[111,112]

Table 1. Continued.

Name	Website	Contact	Overview	Ref.
SUNCAT	suncat.stanford.edu	Thomas Francisco Jaramillo, SLAC/ Stanford University	Materials informatics center for atomic-scale design of catalysts. Online tools and computational results for 112 157 surface reactions and barriers available at catalysis-hub.org.	[89,113]
Citrine Informatics	citrine.io	Bryce Meredig and Greg Mulholland	A materials informatics platform combining data infrastructure and AI. Open database and analytics platform for material and chemical information available at the Citrination platform: citrination.com.	[114,115]
Exabyte.io	exabyte.io	Timur Bazhirov	Cloud-based modelling platform for materials informatics.	[116,117]
Granta Design	grantadesign.com	Mike Ashby and David Cebon	R&D organization offering data, tools and expertise for materials design.	[118]
Materials Design	materialsdesign.com	Clive M. Freeman, Erich Wimmer and Stephen J. Mumby	Software products and services for chemical, metallurgical, electronic, polymeric, and materials science research applications.	[119–121]
Materials Platform for Data Science	mpds.io	Evgeny Blokhin	Online edition of the PAULING FILE with focus on curated experimental data for inorganic materials.	[122,123]
MaterialsZone	materials.zone	Assaf Anderson and Barak Sela	Provides a notebook-based materials informatics environment together with experimental data.	[124]
SpringerMaterials	materials.springer.com	Michael Klinge	Curated data covering multiple material classes, property types, and applications. A set of advanced functionalities for visualizing and analyzing data provided through SpringerMaterials Interactive.	[125]

one of the first attempts to use materials information in a different and more efficient way was the development of the Calculation of Phase Diagrams (CALPHAD, 1970s) method and database, in which multiple calculations of phase diagrams were put in a centralized database to speed up the design and development of new alloys.[52] In the 1990s, the increasing capability to collect, store and analyze "big data" led researchers to explore the potential of data-science in scientific research (for more information, see ref. [4]). With these innovative ideas up in the air, material scientists at the Massachusetts Institute of Technology (MIT) developed tools to predict the properties of materials from datasets.[53] Around the same time, researchers at the Technical University of Denmark demonstrated the potential of evolutionary algorithms in finding materials with specific properties,[54] or to use high-throughput screening for candidate materials with key parameters to narrow down the number of required experiments.[55–57] The researchers at MIT even envisioned how with such computational tools a "virtual materials laboratory" could be build, in which new materials are designed and tested based on computer calculations.[53] These ideas eventually led to the launch of a curated database that is now called the Materials Project.[58,59] This Open Access (see Section 2.2) database would use high-throughput computing to uncover the properties of all known inorganic materials and enable future researchers to find appropriate materials through interactive exploration and data mining.[59,60]

As big data and data science became increasingly fashionable, the US government announced the launch of the Materials Genome Initiative (MGI) in 2011.[61] This initiative emphasized the usefulness of data informatics for materials discovery and design. As similar efforts were launched around the world promoting the availability and accessibility of digital data in science, a trend was set and a new paradigm of materials science emerged:[5] data-driven materials science. Set to reduce time and investment needed to support the typical 10–20 year research-development-commercialization cycle for new materials, more and more Open Access materials data initiatives opened worldwide, as illustrated by **Figures 2** and **3**.

Most of the early materials data initiatives started as databases that hosted data and offered search functionality with the idea to encourage materials scientists to share their data with a larger community. The launch of the Materials Genome Initiative became a defining moment in data-driven materials science (see Figure 3) as databases evolved into data centers that offered rudimentary materials and data analysis services. The emerging interest around data mining and AI made materials scientists increasingly eager to use such algorithms in their research. As a result, the focus of most centers transitioned to developing workflows that would enable scientists to search, mine, and query the databases. This marks another turning point in the history of data-driven materials science, with infrastructures becoming materials discovery platforms (see Figure 3), whose self-declared mission is to facilitate the discovery of novel materials.

The distinction between databases, data centers and materials discovery platforms introduced in the previous paragraph is based on the loose definitions given in the paragraph. The terminology reflects our impression of the evolution of

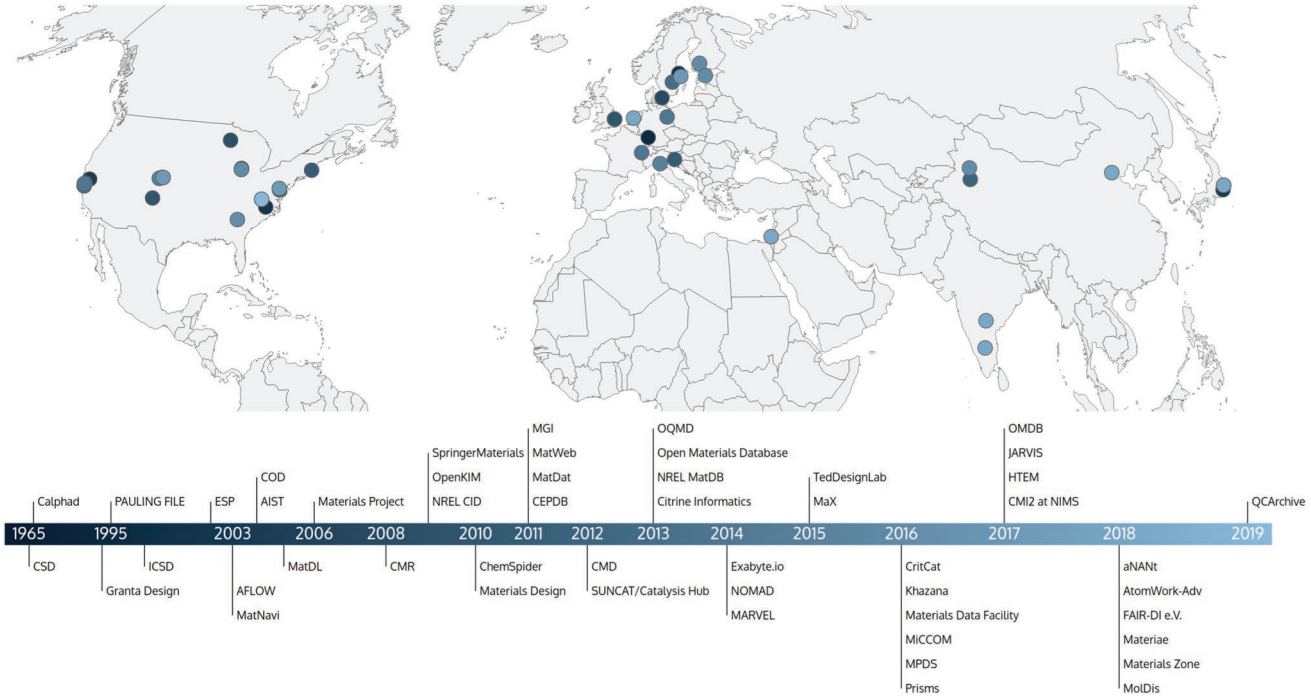

Figure 2. Timeline and geographic distribution of materials data infrastructures and companies. The colour of the dots represents the time of establishment. The map shows that historically more centers have emerged in the U.S. and Europe, with Asia catching up over time. In addition, the U.S. has a higher renewal rate than Europe, as can be seen in the larger number of lighter colored dots. CSD: Cambridge Structural Database, ICSD: Inorganic Crystal Structure Database, ESP: Electronic Structure Project, AFLOW: Automatic-Flow for Materials Discovery, AIST: National Institute of Advanced Industrial Science and Technology Databases, COD: Crystallography Open Database, MatDL: Materials Digital Library, CMR: Computational Materials Repository, NREL CID: NREL Center for Inverse Design, CEPDB: The Clean Energy Project Database, MGI: Materials Genome Initiative, CMD: Computational Materials Network, OQMD: Open Quantum Materials Database, NOMAD: Novel Materials Discovery Laboratory, MaX: Materials Design at the Exascale, MICCOM: Midwest Integrated Center for Computational Materials, MPDS: Materials Platform for Data Science, CMI2: Center for Materials Research by Information Integration, HTEM: High Throughput Experimental Materials Database, JARVIS: Joint Automated Repository for Various Integrated Simulations, OMDB: Organic Materials Database, QCArchive: The Quantum Chemistry Archive.

materials data infrastructures and provides a simple classification scheme to distinguish different infrastructure types. For the remainder of the article, we will use materials data infrastructure as the most general and encompassing term to refer to either of the three types.

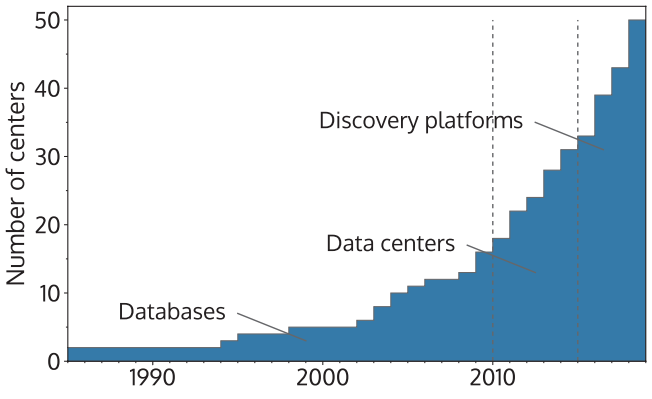

Figure 3. Number of materials informatics projects and infrastructures as function of time (see Figure 2 and Table 1 for details on individual projects and infrastructures). We divide the time axis into three periods that reflect the evolution of the data infrastructures (see text for details).

The spillover effect from data science to materials science is currently boosting the emerging field of data driven materials science or materials informatics.[5] The computational possibilities of machines to analyze and detect patterns in data has created a new feedback loop in the relationship between hypothesis and experiment, which facilitates the next step to mix human trial-and-error experimental and computational research with "artificial intuition" (or to use data mining tools to approach human-like intuition to suggest candidate materials that are further refined via computational and experimental research).

Big data and data science are also prevalent in other scientific fields. In chemistry, databases emerged earlier than in materials science,[62–64] as exemplified by Chemical Abstracts Service (CAS), the principal chemical database provider[65,66] whose first database was created in 1965.[66,67] Carefully produced and curated datasets were essential for developments in quantum chemistry.[68–71] In particle physics, the CERN Open Data Portal[72] offers more than 1 petabyte of open data for research conducted at their facilities. In biology, a variety of databases and metadatabases store biological information, for example, ConsensusPathDB[73] for human protein–protein, genetic, metabolic, signaling, gene regulatory, and drug–target interactions; the protein data bank[74] that houses 3D structural data of large biological molecules; and the International Nucleotide

Sequence Database Collaboration[75,76] that collects and disseminates DNA and RNA sequences. In this perspective article, our focus is on materials science, but it is clear that the "4th scientific paradigm"[4] is emerging in other fields as well.

3.2. Overview of Current Materials Infrastructures

Figure 3 depicts a clear rise of active materials infrastructures, many of which have developed into very mature and stable services used in everyday research processes.[77–80] Table 1 shows a summary of the most prominent materials discovery platforms in existence today. As these platforms have matured, the range of different services they provide has grown (for another perspective on the components of materials data infrastructures see ref. [21]), and **Table 2** shows their features. Data infrastructures that have emerged at the time of submission of this article, such as, e.g., QCArchive[81] have not yet been included in Tabels 1 and 2.

Perhaps the most important service that a data platform has to offer is an efficient distribution channel for the data stored within. Often the data is accessible through a webpage that its clients can access online. This has the lowest adoption barrier since no additional software is needed and the data can be explored visually through a browser. Examples of such services include the Novel Materials Discovery Laboratory (NOMAD) Encyclopedia,[82] AFLOWlib,[83] the Materials Project,[84] and the Materials Cloud.[85] A browser-based method is, however, rarely useful for materials informatics applications, which require automated access to large volumes of data. To facilitate access to large data volumes, it is typical to offer an application programming interface (API) to users to enable automatic data crawling. This is often done by defining a Representational State Transfer (REST) or GraphQL interface to the data.[86–89] These interfaces allow automated access through programmable queries. Another way, as adopted by OQMD,[90] for example, is to offer an offline version of the database as a direct download to users. Offline access provides the most flexibility and performance but typically requires knowledge on how to interact with the underlying database with Structured Query Language (SQL) or object-relational-mapping (ORM). That said, a full download is not practical for large data volumes.

As the amounts of data produced by materials science increases, a practical concern over long-term storage of this data is emerging. There is also increasing pressure from funding agencies and other institutions to ensure the correct

Table 2. Services provided by the selected materials data infrastructures. Open Access: provides partial or full free access to data. Computational data: contains data originating from software simulations. Experimental data: contains data originating from experiments. Data upload: allows upload of own data, with the possibility of issuing Digital Object Identifiers (DOIs). Workflow management tools: provide or collaborate in the development of open-source software tools for workflow management. Web API: data can be accessed remotely with automated scripts. Data analysis tools: provide online or offline data analysis tools, including machine learning.

	Open access	Comp. data	Exp. data	Data upload (DOIs)	Workflow management tools	Web API	Data analysis tools
AFLOW	✓	✓			✓	✓	✓
Computational Materials Repository	✓	✓			✓		✓
Crystallography Open Database	✓	✓	✓	✓			
HTEM	✓		✓	✓		✓	✓
Khazana	✓	✓	✓				✓
MARVEL NCCR	✓	✓		✓	✓		✓
Materials Data Facility (MDF)	✓	✓	✓	✓ (DOI)[a)]		✓	
Materials Project	✓	✓			✓	✓	✓
MatNavi/NIMS	✓	✓	✓				✓
NOMAD CoE	✓	✓		✓ (DOI)		✓	✓
Organic Materials Database	✓	✓					✓
Open Quantum Materials Database	✓	✓					✓
Open Materials Database	✓	✓		✓	✓	✓	✓
SUNCAT	✓	✓				✓	✓
Citrine Informatics	✓[b)]	✓	✓	✓		✓	✓
Exabyte.io						✓	✓
Granta Design		✓	✓				✓
Materials Design		✓	✓				✓
Materials Platform for Data Science	✓[c)]	✓	✓			✓	✓
Materials Zone			✓				✓
Springer Materials			✓				✓

[a)]Upload requires access to private/institutional storage space; [b)]Open access to a subset of data; [c)]Open access to limited set of materials properties.

and safe long-term storage of data. To answer this demand, some data infrastructures now provide data storage services for materials data. Currently, Springer-Nature lists two recommended data repositories for materials science:[126] the NOMAD Repository[107] and the Materials Cloud.[127] Both of these free services are for computational materials data, accept uploads from any source, and guarantee data storage for at least 10 years after data deposition. Often the data volumes in experimental studies, especially in imaging, far outnumber computational efforts. For instance, electron microscopes can easily generate tens of gigabytes of data in a day of operation.[128] Because of this higher volume, it is much more challenging to organize central and free data storage for experimental data. Instead the storage space is provided by the host university or laboratory, as in the case of the Materials Data Facility,[102] which is a collaboration between US universities and research centers. In addition to the materials-science-specific storage solution, there are also free solutions to store generic scientific data, such as Zenodo,[129] Dryad,[130] Figshare,[131] and Dataverse.[132]

The online analytics tools[85,117,133] provided by data infrastructures are fairly modern additions that have emerged from the rise and popularity of interactive browser content and notebook-based environments, such as the Jupyter notebook.[134] These online tools range from simple tutorials to realistic materials property prediction and materials discovery through machine learning. They can be used without local hardware or software resources and have therefore become an important channel for dissemination and learning. Some platforms also participate in the development of Open Source software (see Section 2.3) libraries for performing offline data analysis on materials data.[135–138] Such libraries have high reuse value for scientists working with materials data and, through Open Source distribution and contribution mechanisms, can remain in active use beyond the lifetime of individual projects.

The value of materials data has also been recognized by materials informatics companies. We have included a selection of these companies in Tables 1 and 2. A major selling point of these companies is the access they provide to privately owned, highly curated materials property data that is inherently valuable in R&D. In sufficiently large quantities, this kind of materials data can help firms immensely in selecting optimal materials for products, without having to spend additional expenses on building their own research infrastructure. Another recently emerging business model revolves around selling access to software environments with a Software as a Service (SaaS) model. In this model, companies offer on-demand access to preconfigured cloud-based environments for materials informatics. Such services can be valuable for companies and research laboratories because they can be used according to current demand, do not require large one-off investments in hardware, and do not require specialized skills in software configuration and system management.

3.3. Current Challenges

Having reviewed the current state of data infrastructures in materials science, we now return to the MUSE analogy. The previous two sections illustrated that despite enormous process in data-driven materials science, several challenges need to be overcome before a powerful materials search engine and discovery tool takes shape. The challenges are depicted in **Figure 4** and are raised here briefly before being discussed in detail in corresponding sections.

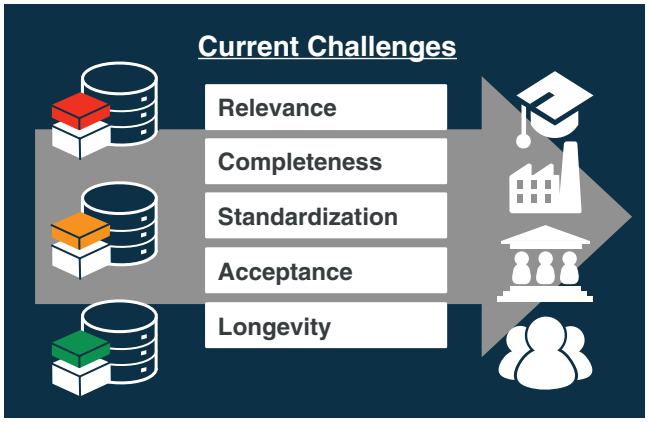

Figure 4. Challenges faced by materials data infrastructures (on the left) on the way to increase the adoption by stakeholders from academia, industry, governments and the public (depicted on the right).

3.3.1. Relevance and Adoption

Materials data infrastructures must provide relevant data and information to be adopted by stakeholders, be it scientific communities, industries, governments, or the public. Relevance is determined by data volume, data type, and data quality, and entails data completeness, and data homogeneity. Different communities will have different specifications of these terms, which makes it challenging to develop general and interdisciplinary infrastructures that can be adopted. Relevance and adoption are therefore closely related to the subsequent challenges of completeness, standardization, and acceptance.

Relevance also includes tools that operate on the data and help users to classify, analyze, and correlate data. Machine learning has gained the most prominence in this regard and is reviewed in Section 7. Since machine learning is always data hungry, it makes sense to integrate machine learning applications directly into materials data infrastructures. Challenges to such one-stop-shop solutions, which would increase the acceptance of materials infrastructures, include the wide variety of available machine learning approaches and data diversity. For data to be informative for machine learning algorithms, its features and properties need to be relevant and complete.

3.3.2. Completeness

Completeness is "the quality of being whole or perfect and having nothing missing." While ideal completeness is hard to attain in practice, data infrastructures today suffer from a real, severe completeness problem: they contain mostly computational and almost no experimental data (cf. Table 2). This state of affairs is rooted in the historic development of data-driven materials science presented in Section 3.1. Since computational

data comes in a digital format, computational scientists were early adopters of database platforms. Moreover computational data is currently more homogeneous and easier to curate than experimental data.[139] Facilitating a seamless comparison between computational and experimental data is, however, an important step toward validating theoretical predictions and in driving materials discovery and development efforts:[139] materials that are identified as promising still require further evaluation, selection, and experimentation. Building synergies among computational and experimental databases thus remains an important challenge for the future of data-driven materials science, which we address in Sections 4 and 5.

3.3.3. Standardization

Some form of standardization is essential in the widespread adoption of a new paradigm or technology.[140,141] Stakeholders can only participate in the development of a technology if they speak a common language. The language analog in data-driven materials science is metadata. Metadata provides relations (the grammar) between data items (the words). Developing standardized metadata for materials science that is informative, exhaustive, and adaptable is an outstanding challenge. We address the first steps toward creating a materials ontology in Section 4. Such an ontology, or classification system, would be the foundation for materials science metadata and the evolution of different materials science dialects into a common language.

3.3.4. Acceptance and Ecosystems

Materials data infrastructures will only be useful if they are accepted as a useful tool by various stakeholders. Apart from being relevant and complete, data infrastructures have to be user friendly to be widely adopted. User friendliness includes easy upload and download of data. Easy data upload also facilitates completeness since it reduces hurdles to data sharing. Widespread acceptance furthermore requires trust in the stored data, and this can only be achieved through data curation. Data curation is the management and quality control of data throughout its lifecycle, from creation and initial storage to the time when it is archived for posterity or becomes obsolete and is deleted. We address the challenges pertaining to data creation and curation in Sections 5 and 6.

Infrastructure acceptance is different for different stakeholders. Current infrastructures are predominantly built and used by scientists in academia, as detailed in the previous section. Industry interest and participation has not been systematically studied, and it is dependent mostly on anecdotal evidence. Some materials companies leverage the value of reference databases (e.g., IBM[142] and ASM International[143]), while others contract the services of intermediaries (cf. Table I and Table II). Apart from this, industry seems to still be exploring the opportunities and potential benefits of materials informatics[144] without full engagement with academia.

The disconnect between academic and corporate R&D in many fields makes industry involvement more difficult in this specific case. A hurdle to widespread industry adoption is the materials gap—the fact that industry requires other data than what is currently stored in the available materials data platforms. Ecosystems that facilitate the interaction between academic, corporate, governmental, and public stakeholders are a potential solution that we discuss further in Section 9.

3.3.5. Longevity and Diffusion

With increasing awareness for open and data-driven science, national, and international funding for the development of Open Science (see Section 2) is rising, and new materials data platforms are emerging in their wake. However, longevity and diffusion of innovations and new technologies are rarely considered by funding agencies, and long-term financial support for sustained operation is not guaranteed. The initial wave of digital materials infrastructures were built predominantly by materials scientists whose main focus lies in basic science. The long-term maintenance and usability of infrastructure is often only a secondary priority for most scientists. As a result, these digital infrastructures are in danger of becoming digital ruins of the expansion of Open Science. We discuss potential solutions in Section 9.

Next we explore central topics and applications around data-driven materials science to provide insight into these challenges and into the successes of data-driven materials science.

4. Materials Ontology

We begin our more detailed review sections with the relevance, completeness, and standardization challenges. One of the first decisions in the planning of a materials data infrastructure is which types of materials will be relevant to its intended user base. The most complete representation of these materials will then have to be stored in the database of the infrastructure. The storage requires standardization and the development of metadata formats. Storing only the raw materials data without any metadata would be futile because raw data is neither searchable nor suitable for machine learning.

The development of a metadata framework requires materials classification schemes from which metadata entries and relations can be derived. Crude classification schemes group materials by their functional properties (electronic, optical, mechanical), topological characteristics (bulk, surface, nanotube, polymer, see **Figure 5**), or by material type (ceramics, metals, glasses, polymers, or composites). More sophisticated classification schemes are clearly needed to facilitate data-driven materials science. In addition, the origin of the raw data needs to be encoded in the metadata. For real samples, these would be the synthesis and processing conditions and the history of the sample since creation. For virtual samples, the generating computer code and the computational settings need to be known. This clearly demands the classification and organization of materials data in a materials ontology.

Ontology is originally a field of philosophy defined as the study of properties, events, processes, and relations of existence.[146] In computer science, the term ontology has been coopted to more specifically mean a formal collection of entities, relationships between those entities, and inference rules that

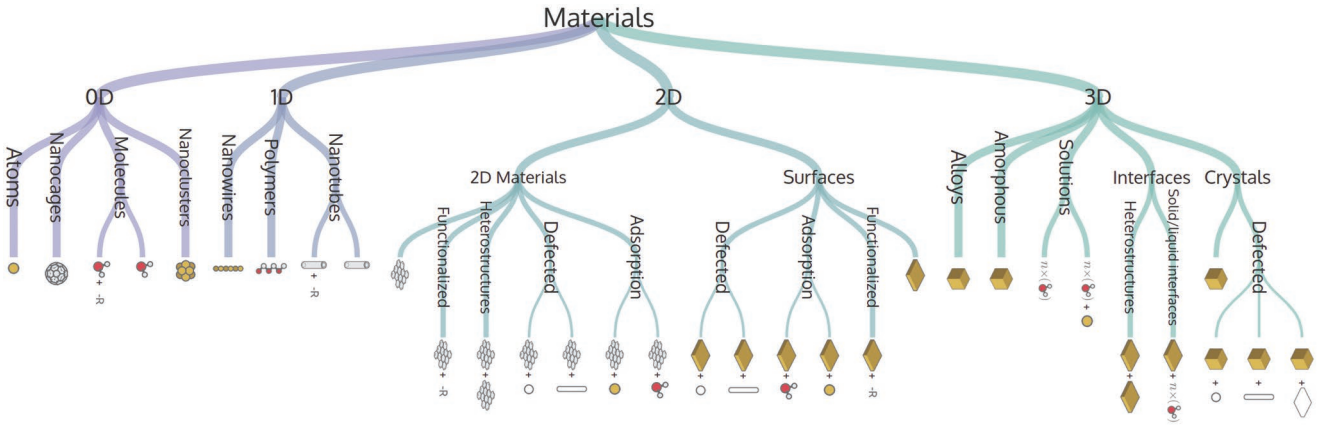

Figure 5. Example of an ontological hierarchy for the structural characterization of materials: a materials tree of life. Adapted under the terms of the Creative Commons Attribution 4.0 International License.[145] Copyright 2018, the Authors, Published by Springer Nature.

are shared by a community. In materials science, a materials ontology would be a classification scheme for materials, their properties, units, and limits, and their interrelations. Defining an ontology is conceptually important for the purpose of establishing a standard that can be shared by different people working with the same data, and it is a practical necessity in database design. The ontology concept is also closely linked to the ability to search data: the ontology defines the available search terms and facilitates semantic reasoning, which then facilitates complex searches.

Creating the necessary machinery for ontologies in materials science is a tremendous task. An ontology structure has to be developed, suitable formats and standards for encoding meaning have to be defined, and wide-spread adoption of the ontology has to be ensured. For example, a substantial part of information available on the internet today consists of human written text. To interpret the information contained within this text requires human reasoning or sophisticated natural language processing software. But there is a complementary standard by the World Wide Web Consortium called Semantic Web that defines an explicit, machine-readable format (Resource Description Framework, RDF) to organize information on the web. It provides an ontology language (Web Ontology Language, OWL) to describe ontologies for sharing concepts across content creators.[147] If this semantic web standard were embraced by the web community, it would significantly boost information sharing across the internet and unleash the power of automated semantic reasoning by artificial intelligence.[148] As of now, this powerful idea remains largely unrealized.

Similarly in materials science, a standardized ontology that ensures a complete representation of materials has not yet emerged. Currently various ontologies and less-than-formal standards compete. NOMAD Meta-info,[86,149] ESCDF,[86] and OpenKIM[150] are the first attempts to categorize computational results in atomistic materials science. PLINIUS[151] is used in the field of ceramics, ONTORULE[152] in the steel industry, SLACKS[153] for laminated composites, and PIF,[154] Ashino,[155] EMMO,[156] MatOnto,[157] Premap,[158] and MatOWL[159] represent general materials science data. Although the development of these materials ontologies has accelerated, they are not nearly as mature as in other fields, for example, the biosciences.[160] Especially for industrial purposes, these publicly available ontologies are typically insufficient, forcing companies to create their own internal, domain-specific ontologies.[156]

The lack of standardization aggravates data sharing. Computational science, for example, still relies on file-based data exchanges between different codes. Such file-based data exchange requires interfacing software, significant human resources, and expertise on how the data is structured. Moreover, incompatible standards lead to conversion errors and data loss. These interoperability problems could be solved by a common ontology and a standardized representation of knowledge within this ontology. Fitting existing and novel data into such an ontological framework would still be a tedious and error-prone task for humans. Existing tools and techniques could, however, be used to simplify and automate this process. Data curation services[161] help in organizing and annotating data, natural language processing can be used to mine data from scientific literature,[162,163] and automated structural classification helps in categorizing the contents.[145,164,165]

The ontologies themselves could also be constructed semi-automatically by observing the nomenclature and the relation of concepts used in the literature.[166,167] If widely adopted in materials infrastructures, this standardization would enable the vision of a powerful search platform for materials science. Once defined and filled, AI solutions would benefit from it. One could envision virtual AI agents helping scientists to answer complex questions related to material performance and synthesis by analyzing materials databases and scientific literature. Such AI agents would not only aid basic research, they would also help businesses that could then more effectively leverage existing scientific knowledge in their R&D.

Recently there have been promising efforts in trying to unify the nomenclature and standards in materials science by the European Materials Modelling Council,[156] the Research Data Alliance (RDA),[168] and by a collaboration between NOMAD scientists and the Centre Européen de Calcul Atomique et Moléculaire (CECAM).[86] A concrete example of such collaboration is the Open Databases Integration for Materials Design (OPTiMaDe) consortium.[169] OPTiMaDe is building a common interface for accessing data from multiple materials platforms. The diversity of subfields and stakeholders in materials science

might make it impossible to define one universal materials ontology. We, however, recommend that unifying ontologies whenever possible and disseminating these efforts to the materials science community are key steps in making the most out of the rich body of materials data created with the modern experimental and computational methods discussed next.

In summary, standardization facilitates data sharing. A materials ontology is a classification scheme for materials that enables standardization. Ontologies also ensure completeness of materials data since everything that falls outside of an ontology by definition indicates a lack of completeness in the ontology. Attempts at constructing materials ontologies are underway. However, more needs to be done to ensure that relevant materials and relevant materials properties are incorporated into existing materials infrastructures. Otherwise our MUSE will return irrelevant information when queried.

5. Data Creation

We now stay with the challenges of relevance and completeness and address how enough relevant data can be generated to feed a materials infrastructure. Once again, we encounter standardization but this time in the context of standards for generating data. We briefly review techniques and recent improvements in data creation methodology—so-called high-throughput methods—that are enabling experimental and computational scientists to efficiently create data for data-hungry repositories and applications.

For experimental materials data, the introduction and refinement of deposition and analysis methods has had perhaps the greatest impact on data creation efficiency. In 1965, the first composition gradients could be achieved in thin-film material codeposition,[170] offering a more efficient replacement for the one-by-one creation and study of materials. Since then, multiple improved materials synthesis and characterization techniques have been introduced.[171–175] They have enabled the rapid generation of composition–structure–property relationships.[176–180] State-of-the-art deposition techniques, such as combinatorial laser-molecular beam epitaxy (CLMBE)[173] introduced around the year 2000, can be used to create temperature and composition gradients across the sample and provide control of the deposition in three dimensions.

These new methods facilitate a finer and more complete sampling of structural phases and chemical compositions in a single experiment. They efficiently create materials libraries—experimental samples with one or more composition or phase gradients. Each library represents part of a well-defined materials space. Measurements from such materials libraries are now made accessible through Open Access online services, such as the High Throughput Experimental Materials (HTEM) database.[97]

In contrast, and perhaps surprisingly, the high-throughput creation of computational materials data has only become common practice in the 21st century.[90,181,182] Thanks to Moore's law and massively parallel computing architectures, available computational power has increased steadily, and computational data creation has quickly taken advantage of this power, even surpassing experimental efforts. For example, the Open Quantum Materials Database (OQMD) performs virtual high-throughput materials synthesis by decorating known crystal structure prototypes with new elements. It has now grown from the initial set of roughly 30,000 experimentally known crystal structures from Inorganic Crystal Structure Database (ICSD) to more than 560,000 computationally predicted materials.[90,110] High-throughput workflows have now matured and are increasingly applied to screen also complex properties such as coupling and reorganization energies in organic crystals.[183,184]

In the creation of such massive datasets, it is increasingly important to adhere to computational standards. This standardization has been pioneered by the Materials Project and the Automatic-Flow for Materials Discovery (AFLOW)-consortium (see Table 1), with comprehensive specifications for the methodological details, such as k-point grid densities, cutoff energies, pseudopotentials, and convergence criteria, related to density functional theory (DFT) calculations.[185–187] This standardization ensures that data can be made cross-compatible within a database or even across databases.

Relatively recent additions to computational materials science are workflow management tools like FireWorks,[188] atomate,[189] AiiDa,[190] and AFLOWπ.[191] These tools enable researchers to build automated and robust workflows for creating consistent datasets. Workflows connect different computational steps and checks into a single computational graph. The computational steps generate data, and checks aid with automated recovery from errors that might occur in a computational step, for instance due to incorrect computational settings or hardware failures. An example of a workflow graph from FireWorks is given in **Figure 6**.

In summary, materials data needs to be created in sufficient volumes for materials infrastructures to be relevant and complete enough. To fulfill this need, high-throughput experimental and computational methods have emerged. The level of automation and efficiency provided by these methods ensures that the bandwidth at which materials data can be created should not be an issue for the MUSE of the future.

6. Data Quality

From data generation, we move on to data quality. As already alluded to in the previous section, the quality of data is related to standardization in the generation of data and considerably affects the acceptance of data infrastructures.

Big data is often characterized in terms of four Vs: volume, velocity, variety, and veracity. Each V poses a challenge, although the volume challenge could, in principle, be solved by more storage space, and the velocity challenge of faster data generation could be addressed by faster computer processing and accelerated measurement and fabrication techniques. Increased variety is more a challenge for standardization and ontology integration, yet it is also a benefit for machine learning and materials discovery algorithms. Veracity, however, is the most problematic because it is a softer measure of how to quantify the degree of trust in data and how to improve its trustworthiness.

Data veracity has two aspects: bias and variance. In an experiment or a calculation, the bias of the result is quantified as the offset of the average result from the ground truth,

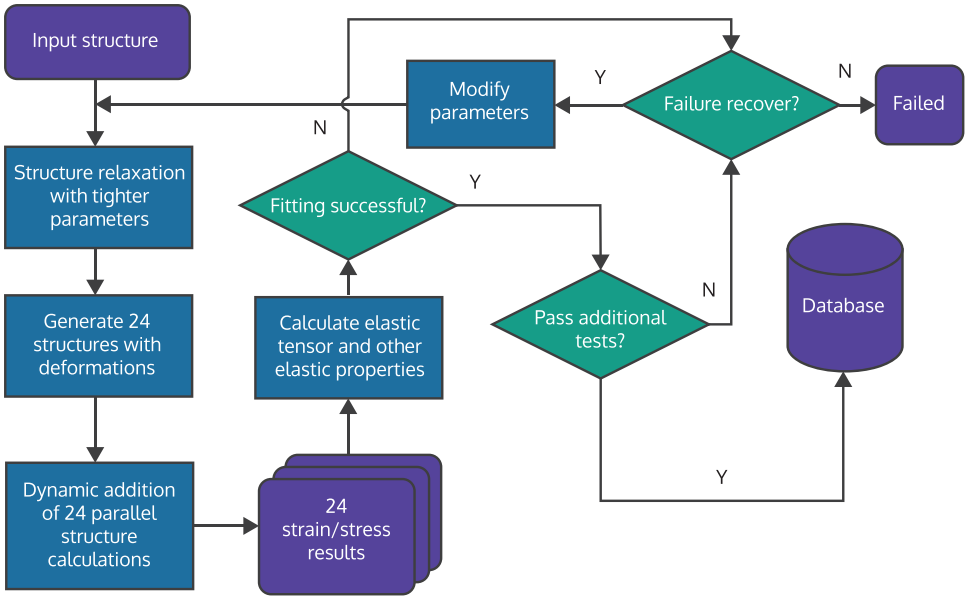

Figure 6. A computational workflow used in creating a dataset of elastic tensors with the FireWorks workflow manager. The indigo boxes correspond to inputs or results, lighter blue boxes correspond to actions, and green diamonds correspond to decisions. Adapted with permission.[188] Copyright 2019, Wiley.

whereas variance is quantified by its probabilistic definition as the spread of the results over identical runs. Note that the variance and bias discussed here originate from approximations in theoretical models or experimental uncertainties, not from stochastic processes in the experiment or calculations, for example in molecular dynamics simulations. We also disambiguate the use of bias and variance here from the same terminology commonly used in machine learning.

For users of data infrastructures, it is important to know the quality of available datasets. However both data bias and its veracity may be hard to quantify.

In the computational realm, variance can be caused by different computational environments or differences in implementation but it is typically negligible even between different software implementations.[192] Computational variance is also often one or several orders of magnitude lower than the corresponding experimental variance.[193] The veracity of computational data is thus dominated by bias—offset from the experimental ground truth. The estimation of this bias depends on access to experimental data or comparison to results from a higher-level theory.[194] The bias also depends on the types of chemical elements in the materials. Some computational approaches or approximations may break down for specific groups of elements, such as dispersion-governed compounds, magnetic materials, strongly correlated materials, and relativistic effects in heavy elements, leading to much larger errors for these groups of materials.[193]

While computational scientists have full control over their simulations, experimental scientists often face errors that are beyond the control of their experimental setup. Bias in measurements can be due to incomplete knowledge of a sample's content and history, as well as interactions between the sample and the environment. Variance can be caused by material imperfections and contaminants, and experimental uncertainties introduced by the equipment. As such, it is typically difficult to discriminate between bias and variance in experimental errors. If there are no comparable experimental facilities, this can also make it hard to assess the data quality. In some cases, quality-controlled commercial equipment and widely accepted standard procedures are available when performing measurements. However, there is often a need to use custom-built equipment or to measure materials for which the standard procedures are not applicable, making it harder to reproduce and validate results. This difficulty of controlling more elaborate experimental setups means that reliable experimental data exists for simple systems, such as small molecules[195] or elemental crystals,[193] but for more complex systems and measurements, the data quality may be harder to determine.

The combination of high bias in computational results and the difficulty of controlling errors in experiments makes the overall estimation of data veracity in materials science particularly hard. One example is given by the formation energies of crystals for which computational and experimental values notably differ. This discrepancy is caused by both systematic errors in the computational methodology and experimental uncertainties.[196] Due to the species-specific nature of the computational error and the vastness of compositional and structural space, it is impossible to make an exhaustive brute-force comparison between experiment and computation. However intelligent error extrapolation schemes are being developed. In one such scheme, the computational error of nonconverged energies for crystals with two different chemical species—compared to fully converged energies—can be estimated by using a linear combination of errors from solids with only one chemical species.[197] Such schemes could also be extended to estimate the error between experimental and computational data. This will require systematic data collection from both experiment and computation but it may prove to be fertile for practical error estimation.

In summary, data quality is important to ensure standardization of data and to increase acceptance of data infrastructures, but it is challenging to quantify. Two indicators of quality are bias and variance. Systematic data collection and new extrapolation schemes would facilitate bias and variance assessments in the future.

7. Data Analytics

To increase the adoption of data infrastructures, developers are adding tools and apps to their data platforms that operate directly on the data in the infrastructure's database. These tools add value to the data and enhance its *relevance*. Many tools now include machine learning. Here we briefly review the main types of machine learning and illustrate how they could add value to material infrastructures.

7.1. Introduction to Machine Learning

Machine learning is the scientific study of how to construct computer programs that automatically improve with experience.[198] More specifically, machine learning algorithms use statistical models and optimization algorithms to reveal patterns in training data to make predictions or decisions without being explicitly programmed to perform a certain task. The advantage over human learning is that computers can often handle much larger and higher dimensional data, and suitable approximations can be automatically found by monitoring how well the models generalize to unseen data.

Many of the statistical methods in machine learning have been around for decades. For example, Marvin Minsky built the first hardware implementation of a neural network[199] in 1951. Our current AI boom has been facilitated by the rapid hardware development for information storage and processing, the conscious effort of data gathering and curation, as well as increased developments of machine learning methods and libraries by private companies, the public sector and academia, driven by the potential that machine learning can unlock from previously untapped data resources. Today machine learning is a key ingredient of materials informatics, as showcased by various reviews on the topic.[5,8–10,13,14,17,18,22,23]

Machine learning can be divided into different subfields that are characterized by the available data. Supervised learning is the most mature and powerful of these subfields and is used in the majority of machine learning studies in the physical sciences.[17] Supervised learning applies in situations where a machine learning model is trained on input–output pairs from a real process to produce optimal outputs for unseen inputs. Typical applications are predictions of physical properties (like formation energies[200–202] or molecular properties[203–207]) given the input features of a material or process (e.g., geometry, physical properties, external conditions).

In unsupervised learning, only input data is given to a model but no output. The machine is then tasked with a learning objective, for example to find rankings or patterns for this input. Unsupervised learning is often used to preprocess input data, such as dimensionality reduction by principal component analysis (PCA),[208,209] or aiding the analysis of complex output data like visualization of high-dimensional data with T-distributed stochastic neighbor embedding (T-SNE)[97,210] or sketchmap.[211]

Finally reinforcement learning is a rapidly emerging field with promising applications in tasks that require machine creativity. In reinforcement learning, a model is given a task of choosing a set of actions to optimize a long-term goal. As such, it differs from supervised learning because no correct input–output pairs are presented for individual actions, but the training is a mixture of exploration and exploitation guided by a long-term reward.[212] This mode of learning can be useful in the exploration of compound and material spaces like exploration of grain-boundary structures with evolutionary algorithms[213] and the search for new molecules with objective-reinforced generative adversarial networks.[214]

The knowledge contained in a machine learning model is encoded in the parameters of the model, and it is, in principle, tractable. However the number of parameters can reach into the millions, which makes these models quite opaque to human interpretation. This is different from the scientific approach thus far, which has relied on deriving and discovering physical laws that are encoded in humanly readable equations. For commercial applications, the transparency of the models is not as important as their performance, but for advancing scientific understanding and wider acceptance, better human interpretability would be beneficial. Recent examples of approaches that analyze machine learning models to reveal their mechanisms include the analysis of input feature importance,[200] explicit formulation of the input in algebraic form,[215,216] and analysis of convolutional neural network filters.[217]

7.2. Specifics of Machine Learning in Materials Science

Currently the applications of machine learning in materials science are rich and diverse, ranging from catalyst design,[19,80] exploring the mechanisms of high-temperature superconductivity,[218,219] to predicting excitation spectra.[206] Building such applications can generally be broken down to four key steps: data acquisition, feature engineering, model building, and analysis. These steps are illustrated in **Figure 7**. They are, however, interdependent and often multiple iterations of each step are required to create a successful machine learning system. Specialized software frameworks[220–222] have been developed to aid the set-up and build and management of machine learning models.

While machine learning generally requires data, the amount of data depends on the specific problem. **Figure 8** illustrates the trade-off between the available data volume and the complexity of the underlying process for different machine learning approaches. Problems in the top-left corner are not suitable for machine learning due to the low amount of available data. The further to the right a problem sits, the more suitable it becomes for machine learning. In practice, it is often difficult to place machine learning methods and new problems in this diagram. Thus rapid prototyping of the problem is frequently a key to successful machine learning. Since we have control over only one of these parameters—the amount of data—the importance of open data access, materials databases, efficient

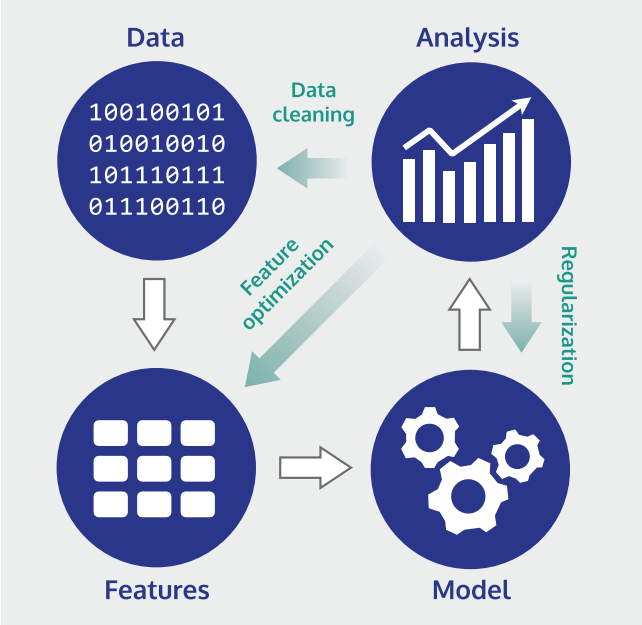

Figure 7. Key steps in building a machine learning model. The white arrows indicate the flow of data, green arrows indicate actions that can be identified and performed after analysis to improve the performance of the model.

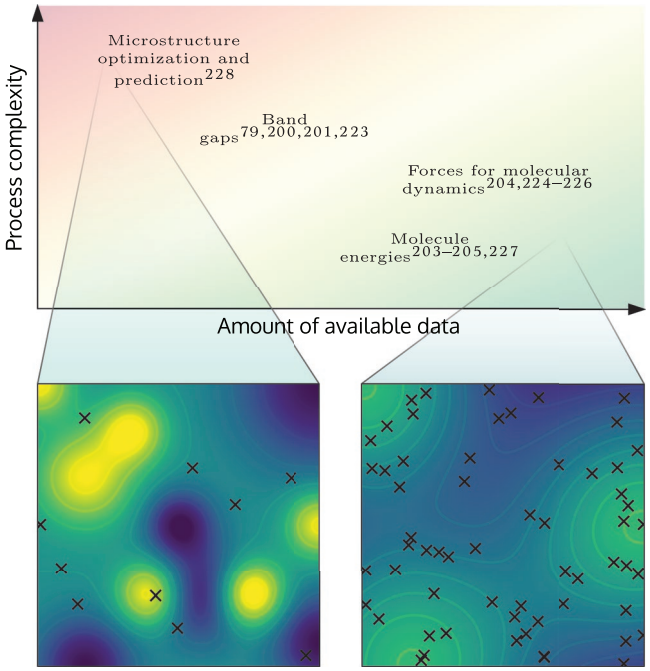

Figure 8. The machine learning domain in terms of data volume and the complexity of the physical process, with selected examples placed in this domain. The complexity of a physical process here means the complex, nonlinear structures present in the data. Two opposing learning scenarios, a hard and an easy one, are illustrated in the lower panel. In these two cases, the underlying physical process is represented by a colored contour map, and the sampling of this process is represented by black crosses.[223–228]

data creation, and data veracity is paramount for the success of machine learning.

Machine learning models expect input data in alpha-numerical form, typically as an array of letters or more often as numbers. Raw data, however, is usually unsuitable for machine learning. The first task in building a machine learning model is therefore to extract informative features from the raw data (cf. Figure 7). Feature engineering refers to the act of introducing domain expertise to the learning model by affecting which features are used. It can be beneficial to apply problem-specific feature transformations, called feature extraction,[212] which exploit known symmetries in the input for example, making it easier for the model to learn a unique mapping. This can be especially important for input features that encode atomic geometries. Physical properties exhibit symmetries with respect to translation, rotation, and index permutation in the Cartesian coordinates representing a geometry. Using a transformation that makes the input invariant with respect to these symmetries will help the learning process by creating a unique mapping from an atomic geometry to its properties. Such structural descriptors have been successfully applied in the prediction of molecular and crystalline properties,[203,205,229] and there their development has exploded in recent years.[202,203,205,229–238] To facilitate easier navigation through descriptor choices, application-neutral software libraries for descriptors are being developed.[239,240] In contrast to human-driven feature engineering, the optimal features can also be discovered more systematically by learning them directly from the data with feature learning. In the simplest form this can be achieved by methods like principal component analysis (PCA),[208] which is based on the variance of the input features. In the other extreme, the features may be formed by an encoder as a nonlinear latent space within an autoencoder neural network.[241] Analyzing the features used by the machine learning model in making a decision forms the basis of understanding and verifying the correctness of the model. Integrating such analysis into the workflow of building machine learning models is still lacking in many cases, hindering their acceptance and interpretability.

After feature selection (cf. Figure 7), the machine learning algorithm must be chosen (see different machine learning types discussed at the beginning of this section). Each algorithm has its own application domain, and there is currently no algorithm that is optimal for all problems.[10] This conundrum is also known as the "no free lunch theorem."[242] Some common choices include feed-forward neural networks,[243] decision trees,[244] kernel ridge regression,[245] support vector regression,[245] and Gaussian processes.[246] These approaches are common in computer science and are available in generic software packages that help select the best model for a task.[50,247–254] Apart from such generic approaches and packages, machine learning models are often customized to materials science. One example is the creation of custom neural network architectures[201,204,206,255,256] that have been designed specifically for atomistic geometries, reducing the need for feature engineering.

One practical concern in model selection is the amount of data that is available for training the model. Methods like kernel ridge regression require an inversion of a matrix whose size is proportional to the number of training samples. This

restricts their usability for large datasets because the time taken by a brute force inversion scales as (n^3) with the dataset size n. Other models like neural networks can handle larger datasets since they can be trained by using small batches of the training data, and their performance can be monitored during training. At the other end of the spectrum we find powerful tools for small datasets such as, regression with Gaussian processes and Bayesian optimization,[18,257] the extraction of effective materials descriptors with subgroup discovery[258] or compressed sensing as done in, e.g., the least absolute shrinkage and selection operator (LASSO) or the sure independence screening and sparsifying operator (SISSO).[215,216] Also some forms of input are better suited for certain models. For example, images exhibit a high degree of correlation between adjacent input points. Models that exploit such correlations, like convolutional neural networks,[259] may then be the best choice. In other cases, the input features have no apparent correlations or have completely different numerical ranges, and decision trees may exhibit the best performance.[244]

The final step in the machine learning workflow is the performance assessment (cf. Figure 7). This analysis guides all other aspects of learning - from excluding corrupt learning samples to optimizing the features and the model itself. The analysis step is general for all application areas of machine learning. For a more in-depth introduction, we refer the reader to existing literature.[198,212] The goal of this step is to ensure that the model generalizes well to unseen data. Two common problems are over- and underfitting. In over-fitting, the model becomes too specific. It reproduces the training data very well but performs poorly on new data. In underfitting, the model learns rules that are too general and averages through training and through new data. The balancing act between over- and underfitting is called the bias-variance trade-off, and it is typically controlled with cross-validation and careful dataset design. The whole dataset is usually split into a training set, from which a further validation set for hyperparameter optimization can be split off, and a test set. Model performance is then evaluated on the test set. The dataset contents and the exact way the data is split into training and test sets can affect the reported performance and in some cases lead to unrealistic results. Often the sampling of training examples is not very even in the input space, as the samples can exhibit high levels of clustering—for example, the dataset may comprise of multiple clustered material types that have very similar properties. In such cases the model is able to interpolate very well even if it has only been trained on one representative of each cluster, but its performance will start to deteriorate for unseen material types, which are hard to leave out of the training set with purely random selection. Due to this effect randomly split training and test sets can offer unrealistic performance metrics and alternative cross-validation strategies like leave-one-cluster-out cross-validation (LOCO CV)[260] offer more realistic performance metrics.

All the key elements for successfully applying machine learning in materials science are in place, as illustrated also by the applications showcased in the next section. However, several challenges prevail. For example, selecting the optimal combination of data, features, machine learning models and analysis tools can be a formidable task, especially because the field is advancing so rapidly and practices become outdated quickly. Careful curation and standardization of both data and machine learning models can to some degree mitigate the problem, but not enough benchmark sets have been established in the community. Also, available data volumes are often still too small to apply machine learning tools that have been successful in other domains, e.g., commerce or social media.

Another challenge is the exchange of pre-trained models. Projects such as OpenML[261] and DLHub[262] are first examples for model-and-data sharing platforms that enable transfer learning, but more could be done. Metric assessment is a further challenge. The reported performance for machine learning models is an important selection criterion for adopting certain models or features. However, performance metrics are not yet standardized. Standardized datasets help, but more attention should be devoted to the selection of test and training sets to obtain more realistic error bars.

We have already discussed the challenge of interpretability. As the exact way input data informs the machine learning model is often blindly guided by the model optimization and hidden behind internal parameters, better methods for interpreting the decisions made by machine learning models are required. Although the natural sciences rarely have to worry about ethical consequences—unlike the social sciences that are now adopting AI into their decision making[263] —a critical evaluation of the decision mechanisms is important for understanding the shortcomings of machine learning models and to advance scientific understanding.

In summary, machine learning is a powerful concept for data analysis and materials informatics. Machine learning is a field undergoing very active development, and a plethora of suitable machine learning methods has been applied to materials science. Increasingly such machine learning tools are incorporated directly into data infrastructures. In our MUSE analogy, they will provide meaningful answers to our "searches."

8. Applications

Staying with relevance *and* adoption, we now briefly present areas in which data-driven materials science has been applied successfully. Success stories are important for the development of any field as they inspire trust and commitment in stakeholders. We have identified three major research objectives for which we think data-driven approaches have the largest impact on materials science: materials discovery, understanding materials phenomena, *and advancing materials modelling*. We review these three areas briefly and present relevant studies.

Materials discovery is a complex problem in which a list of target specifications are given and the optimal material is sought. Such discovery is often performed by using a forward solution—simply calculating the key properties for a pool of candidate materials to identify the best ones for further in-depth analysis, characterization, and verification. By using existing or dynamically generated materials data, scientists can build heuristic models (often through machine learning) to dramatically speed up the identification of best candidate materials for experimental synthesis. Although experimental synthesizability is often a bottleneck, there are multiple studies that have verified that this approach has the

power to accelerate the discovery of novel or improved materials. The examples that have already resulted in experimentally synthesized novel compounds include new molecules for efficient organic light emitting diodes (OLEDs),[264] polymer dielectrics for electrostatic energy storage,[265] novel gallide Heusler structures,[266] NiTi-based shape memory alloys with small thermal dissipation,[267] lead-free piezoelectrics[268] and metallic glasses,[269] and high-entropy alloys[270] for structural applications requiring hardness and corrosion resistance. Furthermore, multiple novel materials identified by this virtual screening are awaiting experimental validation, including photovoltaics,[271,272] photoelectrochemical water splitting materials,[273,274] topological insulators,[275,276] and novel binary or ternary crystal structures.[277,278]

If the desired candidate material is not in the pool of materials that are being screened with the forward solution, then it cannot be found in this way. In such cases, the problem has to be inverted, that is, a mapping from the target property space to materials space has to be found. For solid clusters, simple inverse relations have recently been established between X-ray absorption (XAS) spectroscopy[279–281] and coordination shells of atoms. For molecules, neural-network-based auto encoders and decoders[282] were combined with a grammar-based variational autoencoder[283] to map from the discrete molecular space into a continuous latent space (in which optimizations can be performed) and back again. Even with such sophisticated models, it is not easy to generate valid, synthesizable molecules and materials with the wanted properties, and inverse predictions remain difficult in practice.

Historically new materials were often discovered by first understanding fundamental materials phenomena and then applying them to look for other materials that might fit the same physical laws. Machine learning based predictions conceal these physical laws and do not provide the same understanding of materials-property relations. By changing the objective to understanding the underlying materials phenomena (asking why instead of what), we can hope to use reductionistic approaches to transcribe data into laws and equations that are more typically associated with scientific progress. In vast and high-dimensional materials data landscapes, for which human-intuition is ill-suited, this discovery of materials phenomena can be aided by a data-driven approach. The fundamental mathematical formulation of physical laws, such as conservation laws and differential equations, can be automatically deduced from data.[284–286] Methods based on compressed sensing[215,216] provide a systematic way of identifying the algebraic form of the descriptors that capture the underlying mechanisms behind material properties, providing a more natural basis for human interpretation. These methods have been successful in identifying physically meaningful descriptors that control the stability of perovskites[287] and monolayer metal oxides coatings.[288] Another idea is to map high-dimensional data into more easily human analyzable two- or 2D maps with unsupervised learning. This idea of "materials cartography" has been used to identify common features for high-temperature superconductor materials,[218] to group molecules into intuitive maps that can reveal key structure–property relations,[211] find phase transitions in complex systems,[289] or establish the key descriptors in the catalytic properties of metal surfaces.[290]

The final application area is related to advancing materials modelling with automated construction of surrogate models directly from data. These surrogate models can replace the laborious fitting of semiempirical models, and if trained with highly accurate data are able to reproduce complex chemical phenomena with very low computational cost by sacrificing some of the accuracy. Such AI-based modelling tools are able to assist even in very challenging tasks, as demonstrated by IBM RXN[291]—a free online tool that uses a machine translation inspired architecture to predict the product of chemical reaction from the structural formulas of the reactants. Another example is the creation of classical force fields from DFT training data.[204,224–226,292]

The promise of these methods is to achieve near DFT-level accuracy in the physical and chemical description of a system but at the much lower computational cost that is closer to classical force fields. These machine learned force fields enable studies of systems and mechanisms that have so far been out of the realm of computational studies, such as identifying the growth mechanism of amorphous carbon coatings[293] under deposition and the composition and activity of nanoclusters in aqueous solutions.[294] The implementations have now also made their way into established molecular dynamics software, where they can in some cases be used as a plug-and-play replacement for traditional force fields.[295–298]

Going one step up in the theoretical ladder, energies of very accurate, yet computationally very expensive coupled cluster theory calculations have been learned.[22,299–301] Another example is the training of exchange-correlation functionals or direct potential-to-density mappings from density or wave-function based methods.[302,303] Such efforts are an exciting step in expanding existing theoretical knowledge to new time and size regimes in materials modelling.

In summary, machine learning and data-driven approaches are being applied in materials science. The wider the range of successful applications, the higher the acceptance of materials data infrastructures will be. New applications will, in turn, challenge established machine learning methods, which will have to be further developed to address these challenges. This creates a feedback loop with developments in computer and data science and ensures that our MUSE continues to learn.

9. Stakeholder Relations

In previous sections, we have addressed the acceptance of materials data infrastructures from a technological viewpoint. We now reflect on how materials data infrastructures are currently received by different stakeholders. Strong support from stakeholders is needed to guarantee the diffusion of innovations to a wider pool of stakeholders and to ensure the longevity of data infrastructures.

Materials scientists in academia are actively pushing the frontiers of materials informatics to advance and accelerate materials design and discovery. However emerging fields require the interaction of various actors and stakeholders from different communities (academia, government, industry, and the public,

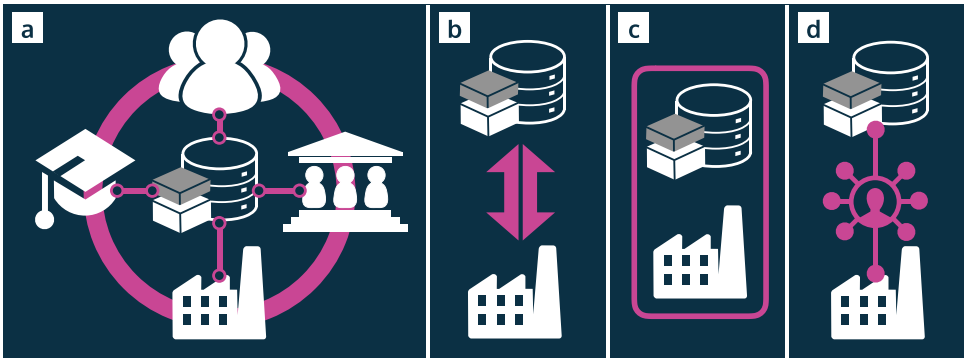

Figure 9. a) Schematic of an ecosystem in data-driven materials science with materials data platforms at the center. In this ecosystem, different stakeholders from universities, the public, industry, and government facilitate the development of a technology. b–d) Possible relationships between data platforms and industry, discussed in the text.

cf. **Figure 9**a) to generate understanding of the field and negotiate community boundaries.[304] A recent socioeconomic study investigated the emerging field of data-driven materials science. It identified that the field is scattered and largely lacking a supportive ecosystem with nonacademic stakeholders.[305,306]

The socioeconomic study identified visionary scientists at academic institutions who have pursued their idiosyncratic research objectives using trending topics in public discussions and governmental funding, including Open Science, Big Data, and AI. At the same time, government and funding agencies have provided strategic research openings focused on propelling the field of data-driven science in general. However, the focus on materials science applications or on finding and developing industry applications has been limited by the lack of targeted funding opportunities. With the exception of the US, which released substantial government funding to advance the field of data-driven materials science via the Materials Genome Initiative,[61] most government-sponsored funding schemes have been more general. In the European Union, two successful projects have been funded (MaX[307] and NOMAD CoE[106]). However both centers were facilitated by the European Union's Horizon 2020 high-performance computing grants rather than funding schemes focused on data-driven materials science. Until now, no call tailored to the exploration and exploitation of data-driven materials science has been issued by the EU, although this may change in the new framework program. Nevertheless national differences exist. For example in Switzerland, the importance of data-driven materials science has been acknowledged by the support for the MARVEL National Center of Competence in Research (NCCR).[85] In 2018 in Finland, a Future Makers funding call was opened specifically focused on the initiation of "high-level, ambitious strategic research openings that combine internationally top-level science and industrial impact ... to build long-term sustainable renewal and competitiveness of the Finnish technology industry based on ... data-based materials science".[308] Despite the call, none of the applications from materials science were funded.

The socioeconomic study further concludes that industry has remained seemingly reluctant to invest in data-driven scientific applications in materials research and development. This reluctance could be driven by government hesitation to fund data-driven materials science, but it could also be the result of the material industries' desire to work with proprietary databases, the generally long timeframe for the development of new materials (10–20 years), and/or the limited number of manufacturing employees with informatics backgrounds.[144] Industries do, however, capitalize on the creation and appropriation of (new) knowledge,[309] and new materials could advance such fields as health, energy, aerospace, automotive, semiconductor, and consumer goods:[144] materials informatics provides unprecedented opportunities for an industry to better use the existing vast "storehouses of information"[310] that firms possess to propel materials innovation at greater rates and lower costs. Despite the potential gains, uptake by industrial partners is challenging. Although materials companies generate enormous quantities of R&D data, this data is often undocumented and intangible. Before proprietary databases could be created, the archival data of companies needs to be structured, connected, and updated. Ignoring the identified challenges—acceptance (easy data upload and download, data curation, materials gap), standardization, and longevity—slows down industry adoption of data-driven materials science even further. In addition, firms face significant challenges to become or transform into data-driven organizations as they require different skills, knowledge, and resources.[311]

To capitalize on their data, three business models can be applied or are considered for application.[305,306] First, firms can consult and collaborate directly with established data platforms (see Figure 9b), for example those discussed in Figures 2 and 3. Traditionally such collaborations take the form of strategic interfirm alliances that influence companies' potential for knowledge creation.[312] Propelled by the drive for Open Science from policy players, Open Innovation is another potential form of collaboration in which firms open their internal innovation processes by purposefully allowing knowledge to freely circulate among all actors to accelerate internal innovation.[31] As a result, such data platforms offer data and services that can be used by academia and industries alike.

In the second model, if a firm does not wish to collaborate, it can consider building a proprietary digital infrastructure by dedicating a large, one-off investment in hardware and by acquiring software and specialized skills in software configuration and system management[313] (Figure 9c). This integration requires attracting computer scientists or materials scientists with extensive coding capabilities.

A final business model is developed around new intermediaries that position themselves between data platforms and industry (Figure 9d). Examples of such new intermediaries are materials informatics companies—often spin-off companies from academic efforts—that sell access to privately owned, highly curated materials property data that can be linked to data repositories within the firm and used in R&D processes. Collaboration with start-ups and companies (e.g., Citrine Informatics, Exabyte.io, Materials Design, and Granta Design[115,116,118,121]) can help firms develop new skills, capabilities, and knowledge.[312,314]

Each of these data platform–industry relationships have implications for the larger ecosystem of data-driven materials science. That is, industry and academia often hold contradicting interests. From the academic perspective, commercial business opportunities stimulate private data ownership and proprietary databases (e.g., Figure 9c), which can be detrimental for scientific progress since valuable data stays locked in the private domain. Furthermore, industries outside materials science are quickly recruiting academic employees with new coding and machine learning expertise in the field; this raises concerns over a possible "brain drain" from universities. Nevertheless the establishment and institutionalization of data science and machine learning for materials science within educational institutions could result in an increase in research funding, students, and industry interest or collaboration.[305]

At the same time, industries' need for these newly-skilled employees raises fear of "job loss" among established scientists within R&D departments in those firms. However, materials that are identified as promising through the materials informatics paradigm still require further evaluation, selection, experimentation, certification, and manufacturing. Building synergies among computational and experimental researchers therefore remains a key enabler toward reaping the benefits of data-driven materials science in firms.[305]

In summary, the field of data-driven materials science is still in its infancy, with an emerging ecosystem, ongoing community boundary negotiations, limited governmental funding, and an as yet disinterested industry. To establish data-driven materials science as a new paradigm in materials research, joint ecosystem efforts between research, industry, and public and governmental organizations are necessary.

10. Take-Home Messages

In this review article, we provide an overview of the current state of data-driven materials science. From a historical perspective, the field has matured greatly, but we identify key challenges—relevance, completeness, standardization, acceptance, and longevity—that still need to be resolved to create the MUSE. Better standardization of materials data through a materials ontology would immensely help sharing, integrating, and employing AI-powered analysis of materials data. Creating feedback mechanisms between experimental and computational data for error estimation provides a way toward solving the veracity problem in materials data. The use of machine learning is transformational for research, but it requires conscious efforts in the curation and standardization of both data and machine learning models, and techniques to make the models more interpretable. The synergy between academic developments and industrial interest remains a major challenge, but it is key to creating a sustainable ecosystem for materials data and expertise. Despite these challenges, there has been a dramatic rise in data-driven materials science using the full spectrum of this new paradigm. And doubtless we have only seen a glimpse of this data-driven revolution.

Acknowledgements

The authors thank Nina Granqvist, Kunal Ghosh, Ben Alldritt, Antti M. Rousi, Milica Todorović, Sven Bossuyt, Miguel Caro, David Gao, Matthias Scheffler, Bryce Meredig, and Heidi Henrickson for insightful discussions and a careful reading of our manuscript. Computing resources from the Aalto Science-IT project and CSC IT Center for Science, Finland, are gratefully acknowledged. This project had received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 676580 with The Novel Materials Discovery (NOMAD) Laboratory, a European Center of Excellence, and from the Jenny and Antti Wihuri Foundation. This work was furthermore supported by the Academy of Finland through its Centres of Excellence Programme 2015–2017 under project number 284621, as well as projects 305632, 311012, and 314862. A.S.F. was supported by the World Premier International Research Center Initiative (WPI), MEXT, Japan. This article is part of the Advanced Science 5th anniversary interdisciplinary article series, in which the journal's executive advisory board members highlight top research in their fields.

Conflict of Interest

The authors declare no conflict of interest.

Keywords

artificial intelligence, databases, data science, machine learning, materials, materials science, open innovation, open science

Received: April 8, 2019
Revised: June 20, 2019
Published online: September 1, 2019

[1] G. Ceder, *Science* **1998**, *280*, 1099.
[2] T. W. Eagar, *Technol. Rev.* **1995**, *98*, 42.
[3] M. Boren, V. Chan, C. Musso, In *McKinsey on Chemicals*, Vol. 4. McKinsey & Company Industry Publications, **2012**.
[4] T. Hey, S. Tansley, K. Tolle, *The Fourth Paradigm: Data-Intensive Scientific Discovery*, Microsoft Research, Redmond, WA, USA **2009**.
[5] A. Agrawal, A. Choudhary, *APL Mater.* **2016**, *4*, 053208.
[6] K. Schwab, The Fourth Industrial Revolution, https://www.foreignaffairs.com/articles/2015-12-12/fourth-industrial-revolution (accessed: August 2019).
[7] E. Brynjolfsson, A. McAfee, *Second Machine Age. Work, Progress, and Prosperity in a Time of Brilliant Technologies*, W. W. Norton & Company, New York City, NY, USA **2014**.
[8] K. Rajan, *Mater. Today* **2005**, *8*, 38.
[9] K. Rajan, *Annu. Rev. Mater. Res.* **2015**, *45*, 153.
[10] Y. Liu, T. Zhao, W. Ju, S. Shi, *J. Materiomics* **2017**, *3*, 159.

[11] L. Zdeborová, *Nat. Phys.* **2017**, *13*, 420.
[12] M. Rupp, *Int. J. Quantum Chem.* **2015**, *115*, 1058.
[13] R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi, C. Kim, *npj Comput. Mater.* **2017**, *3*, 54.
[14] L. Ward, C. Wolverton, *Curr. Opin. Solid State Mater. Sci.* **2017**, *21*, 167.
[15] T. Mueller, A. G. Kusne, R. Ramprasad, *Machine Learning in Materials Science*. John Wiley & Sons, Hoboken, NJ, USA **2016**, Ch. 4, pp. 186–273.
[16] A. Zunger, *Nat. Rev. Chem.* **2018**, *2*, 0121.
[17] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, A. Walsh, *Nature* **2018**, *559*, 547.
[18] J. E. Gubernatis, T. Lookman, *Phys. Rev. Mater.* **2018**, *2*, 120301.
[19] B. R. Goldsmith, J. Esterhuizen, J. X. Liu, C. J. Bartel, C. Sutton, *AIChE J.* **2018**, *64*, 2311.
[20] K. Takahashi, Y. Tanaka, *Dalton Trans.* **2016**, *45*, 10497.
[21] The Minerals Metals & Materials Society (TMS), *Building a Materials Data Infrastructure: Opening New Pathways to Discovery and Innovation in Science and Engineering*. TMS, Pittsburgh, PA **2017**.
[22] A. P. Bartók, S. De, C. Poelking, N. Bernstein, J. R. Kermode, G. Csányi, M. Ceriotti, *Sci. Adv.* **2017**, *3*, 12.
[23] G. H. Gu, J. Noh, I. Kim, Y. Jung, *J. Mater. Chem. A* **2019**, *7*, 17096.
[24] K. Walsh, editor, *Open Innovation, Open Science, Open to the World—A Vision for Europe*, European Commission, Luxembourg City, Luxembourg **2016**.
[25] J. J. Regazzi, *Scholarly Communications: A History from Content as King to Content as Kingmaker*, Rowman & Littlefield, Lanham **2015**.
[26] D. Fanelli, *Proc. Natl. Acad. Sci. USA* **2018**, *115*, 2628.
[27] J. P. Tennant, F. Waldner, D. C. Jacques, P. Masuzzo, L. B. Collister, C. H. J. Hartgerink, *F1000Research* **2016**, *5*, 632.
[28] J. M. Esanu, P. F. Uhlir, in *The Role of Scientific and Technical Data and Information in The Public Domain: Proc. of A Symp*. The National Academies Press, Washington, D.C. **2003**.
[29] Open Science, https://openscience.com (accessed: August 2019).
[30] FOSTER, https://www.fosteropenscience.eu (accessed: August 2019).
[31] H. W. Chesborough, *Open Innovation: The New Imperative for Creating and Profiting from Technology*. Harvard Business Press, Brighton, MA, USA **2006**.
[32] E. C. McKiernan, P. E. Bourne, C. T. Brown, S. Buck, A. Kenall, J. Lin, D. McDougall, B. A. Nosek, K. Ram, C. K. Soderberg, J. R. Spies, K. Thaney, A. Updegrove, K. H. Woo, T. Yarkoni, *eLife* **2016**, *5*, 372.
[33] When will everything be Open Access?, https://blog.impactstory.org/oa-by-when/ (accessed: August 2019).
[34] H. Piwowar, J. Priem, V. Larivière, J. P. Alperin, L. Matthias, B. Norlander, A. Farley, J. West, S. Haustein, *PeerJ* **2018**, *6*, e4375.
[35] M. Schiltz, *PLOS Biol.* **2018**, *16*, 1.
[36] ICSU World Data System, https://www.icsu-wds.org (accessed: August 2019).
[37] OECD, *OECD Principles and Guidelines for Access to Research Data from Public Funding*, OECD Publishing, Paris, France **2007**.
[38] re3data.org, https://www.re3data.org (accessed: August 2019).
[39] PLOS, https://www.plos.org (accessed: August 2019).
[40] Scientific Data, https://www.nature.com/sdata/ (accessed: August 2019).
[41] Sharing Research Data for Journal Authors, https://www.elsevier.com/authors/author-services/research-data (accessed: August 2019).
[42] CODATA, The Committee on Data for Science and Technology, https://www.codata.org (accessed: August 2019).
[43] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. G. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. C. t Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, et al., *Sci. Data* **2016**, *3*, 160018.
[44] C. Draxl, M. Scheffler, *MRS Bull.* **2018**, *43*, 9.
[45] C. M. Kelty, *Two Bits: The Cultural Significance of Free Software*, Duke University Press Books, Durham, NC, USA **2008**.
[46] The OpenScience Project, https://openscience.org (accessed: August 2019).
[47] R. Kwok, *Nature* **2018**, *560*, 269.
[48] I. Horcas, R. Fernández, J. M. Gómez-Rodríguez, J. Colchero, J. Gómez-Herrero, A. M. Baro, *Rev. Sci. Instrum.* **2007**, *78*, 013705.
[49] S. Plimpton, *J. Comput. Phys.* **1995**, *117*, 1.
[50] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, *J. Mach. Learn. Res.* **2011**, *12*, 2825.
[51] Overleaf, https://www.overleaf.com (accessed: August 2019).
[52] L. Kaufman, J. Agren, *Scr. Mater.* **2014**, *70*, 3.
[53] G. Ceder, M. K. Aydinol, A. F. Kohan, *Comput. Mater. Sci.* **1997**, *8*, 161.
[54] G. H. Jóhannesson, T. Bligaard, A. V. Ruban, H. L. Skriver, K. W. Jacobsen, J. Nørskov, *Phys. Rev. Lett.* **2002**, *88*, 255506.
[55] J. Greeley, J. K. Nørskov, *Surf. Sci.* **2007**, *601*, 1590.
[56] J. Hummelshøj, F. Abild-Pedersen, F. Studt, T. Bligaard, J. K. Nørskov, *Angew. Chem.***2012**, *124*, 278; *Angew. Chem., Int. Ed.* **2012**, *51*, 272.
[57] J. Greeley, T. F. Jaramillo, J. Bonde, I. Chorkendorff, J. K. Nørskov, *Nat. Mater.* **2006**, *5*, 909.
[58] G. Ceder, D. Morgan, C. Fischer, K. Tibbetts, S. Curtarolo, *MRS Bull.* **2006**, *31*, 981.
[59] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, K. a. Persson, *APL Mater.* **2013**, *1*, 011002.
[60] A. Jain, G. Hautier, C. J. Moore, S. P. Ong, C. C. Fischer, T. Mueller, K. A. Persson, G. Ceder, *Comput. Mater. Sci.* **2011**, *50*, 2295.
[61] Materials Genome Initiative, https://www.mgi.gov/ (accessed: August 2019).
[62] M. Hann, R. Green, *Curr. Opin. Chem. Biol.* **1999**, *3*, 379.
[63] J. Noordik, *Cheminformatics Developments: History, Reviews and Current Research*, IOS Press, Clifton, VA, USA **2004**.
[64] P. Willett, *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2011**, *1*, 46.
[65] D. B. Baker, J. W. Horiszny, W. V. Metanomski, *J. Chem. Inf. Model.* **1980**, *20*, 193.
[66] D. W. Weisgerber, *J. Assoc. Inf. Sci. Technol.* **1997**, *48*, 349.
[67] D. P. Leiter, H. L. Morgan, R. E. Stobaugh, *J. Chem. Doc.* **1965**, *5*, 238.
[68] L. A. Curtiss, K. Raghavachari, G. W. Trucks, J. A. Pople, *J. Chem. Phys.* **1991**, *94*, 7221.
[69] L. A. Curtiss, K. Raghavachari, P. C. Redfern, J. A. Pople, *J. Chem. Phys.* **1997**, *106*, 1063.
[70] L. Goerigk, S. Grimme, *J. Chem. Theory Comput.* **2011**, *7*, 291.
[71] R. A. Mata, M. A. Suhm, *Angew. Chem.* **2017**, *129*, 11155; *Angew. Chem., Int. Ed.* **2017**, *56*, 11011.
[72] CERN Open Data Portal, https://opendata.cern.ch (accessed: August 2019).
[73] C. Wierling, H. Lehrach, R. Herwig, A. Kamburov, *Nucleic Acids Res.* **2008**, *37*, D623.
[74] H. Berman, K. Henrick, H. Nakamura, *Nat. Struct. Biol.* **2003**, *10*, 980.
[75] I. Karsch-Mizrachi, T. Takagi, G. Cochrane, *Nucleic Acids Res.* **2011**, *40*, D33.

[76] S. Brunak, A. Danchin, M. Hattori, H. Nakamura, K. Shinozaki, T. Matise, D. Preuss, *Science* **2002**, *298*, 1333.
[77] P. Sarker, T. Harrington, C. Toher, C. Oses, M. Samiee, J.-P. Maria, D. W. Brenner, K. S. Vecchio, S. Curtarolo, *Nat. Commun.* **2018**, *9*, 4980.
[78] T. Xie, J. C. Grossman, *Phys. Rev. Lett.* **2018**, *120*, 145301.
[79] G. Pilania, A. Mannodi-Kanakkithodi, B. P. Uberuaga, R. Ramprasad, J. E. Gubernatis, T. Lookman, *Sci. Rep.* **2016**, *6*, 19375.
[80] M. O. J. Jäger, E. V. Morooka, F. F. Canova, L. Himanen, A. S. Foster, *npj Comput. Mater.* **2018**, *4*, 37.
[81] QCArchive, https://qcarchive.molssi.org (accessed: August 2019).
[82] NOMAD Encyclopedia, https://encyclopedia.nomad-coe.eu (accessed: August 2019).
[83] AFlow—Automatic FLOW for Materials Discovery, https://aflowlib.org (accessed: August 2019).
[84] Materials Project, https://materialsproject.org (accessed: August 2019).
[85] Materials Cloud, https://www.materialscloud.org (accessed: August 2019).
[86] L. M. Ghiringhelli, C. Carbogno, S. Levchenko, F. Mohamed, G. Huhs, M. Lüders, M. Oliveira, M. Scheffler, *npj Comput. Mater.* **2017**, *3*, 46.
[87] R. H. Taylor, F. Rose, C. Toher, O. Levy, K. Yang, M. B. Nardelli, S. Curtarolo, *Comput. Mater. Sci.* **2014**, *93*, 178.
[88] S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson, G. Ceder, *Comput. Mater. Sci.* **2013**, *68*, 314.
[89] Catalysis Hub, https://www.catalysis-hub.org/ (accessed: August 2019).
[90] J. E. Saal, S. Kirklin, M. Aykol, B. Meredig, C. Wolverton, *JOM* **2013**, *65*, 1501.
[91] S. Curtarolo, W. Setyawan, S. Wang, J. Xue, K. Yang, R. Taylor, L. J. Nelson, G. Hart, S. Sanvito, M. Buongiorno Nardelli, N. Mingo, O. Levy, *Comput. Mater. Sci.* **2012**, *58*, 227.
[92] I. E. Castelli, D. D. Landis, K. S. Thygesen, S. Dahl, I. Chorkendorff, T. F. Jaramillo, K. W. Jacobsen, *Energy Environ. Sci.* **2012**, *5*, 9034.
[93] Computational Materials Repository, https://cmr.fysik.dtu.dk (accessed: August 2019).
[94] D. D. Landis, J. S. Hummelshøj, S. Nestorov, J. Greeley, M. Dułak, T. Bligaard, J. K. Nørskov, K. W. Jacobsen, *Comput. Sci. Eng.* **2012**, *14*, 51.
[95] S. Gražulis, A. Daškevič, A. Merkys, D. Chateigner, L. Lutterotti, M. Quirós, N. R. Serebryanaya, P. Moeck, R. T. Downs, A. Le Bail, *Nucleic Acids Res.* **2011**, *40*, 420.
[96] Crystallography Open Database, https://crystallography.net (accessed: August 2019).
[97] A. Zakutayev, N. Wunder, M. Schwarting, J. D. Perkins, R. White, K. Munch, W. Tumas, C. Phillips, *Sci. Data* **2018**, *5*, 1.
[98] HTEM, https://htem.nrel.gov (accessed: August 2019).
[99] Khazana: A Computational Materials Knowledgebase, https://khazana.gatech.edu/ (accessed: August 2019).
[100] C. Kim, A. Chandrasekaran, T. D. Huan, D. Das, R. Ramprasad, *J. Phys. Chem. C* **2018**, *122*, 17575.
[101] V. Botu, R. Batra, J. Chapman, R. Ramprasad, *J. Phys. Chem. C* **2017**, *121*, 511.
[102] B. Blaiszik, K. Chard, J. Pruyne, R. Ananthakrishnan, S. Tuecke, I. Foster, *JOM* **2016**, *68*, 2045.
[103] The Materials Data Facility (MDF), https://materialsdatafacility.org (accessed: August 2019).
[104] A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, K. A. Persson, *APL Mater.* **2013**, *1*, 011002.
[105] NIMS Materials Database (MatNavi), https://mits.nims.go.jp/index_en.html (accessed: August 2019).
[106] The Novel Materials Discovery (NOMAD) Laboratory, https://nomad-coe.eu/ (accessed: August 2019).
[107] NOMAD Repository, https://repository.nomad-coe.eu (accessed: August 2019).
[108] S. S. Borysov, R. M. Geilhufe, A. V. Balatsky, *PLOS ONE* **2017**, *12*, 1.
[109] Organic Materials Database, https://omdb.mathub.io (accessed: August 2019).
[110] OQMD, https://oqmd.org (accessed: August 2019).
[111] Open materials database, https://openmaterialsdb.se/ (accessed: August 2019).
[112] The high-throughput toolkit (httk), https://httk.openmaterialsdb.se (accessed: August 2019).
[113] SUNCAT, https://suncat.stanford.edu/ (accessed: August 2019).
[114] J. O'Mara, B. Meredig, K. Michel, *JOM* **2016**, *68*, 2031.
[115] Citrine Informatics, https://citrine.io/ (accessed: August 2019).
[116] Exabyte.io, https://exabyte.io/ (accessed: August 2019).
[117] T. Bazhirov, M. Mohammadi, K. Ding, S. Barabash, presented at APS March Meeting Abstracts, **2017**, C1.007.
[118] Granta Design, https://www.grantadesign.com/ (accessed: August 2019).
[119] X. Rozanska, J. J. P. Stewart, P. Ungerer, B. Leblanc, C. Freeman, P. Saxe, E. Wimmer, *J. Chem. Eng. Data* **2014**, *59*, 3136.
[120] X. Rozanska, P. Ungerer, B. Leblanc, P. Saxe, E. Wimmer, *Oil Gas Sci. Technol.* **2015**, *70*, 405.
[121] Materials Design Inc, https://www.materialsdesign.com/ (accessed: August 2019).
[122] E. Blokhin, P. Villars, in *Handbook of Materials Modeling: Methods: Theory and Modeling* (Eds: W. Andreoni, S. Yip), Springer International Publishing, Dordrecht, Netherlands **2018**, pp. 1–26.
[123] Materials Platform for Data Science, https://mpds.io (accessed: August 2019).
[124] MaterialsZone, https://www.materials.zone (accessed: August 2019).
[125] SpringerMaterials, https://materials.springer.com (accessed: August 2019).
[126] Springer Nature Recommended Repositories, https://www.springernature.com/de/authors/research-data-policy/repositories/12327124 (accessed: August 2019).
[127] Materials Cloud, https://www.materialscloud.org (accessed: August 2019).
[128] S. R. Kalidindi, M. De Graef, *Annu. Rev. Mater. Res.* **2015**, *45*, 171.
[129] Zenodo, https://zenodo.org/ (accessed: August 2019).
[130] Dryad, https://datadryad.org (accessed: August 2019).
[131] Figshare, https://figshare.com (accessed: August 2019).
[132] The Dataverse Project, https://dataverse.org (accessed: August 2019).
[133] NOMAD Big-Data Analytics, https://analytics-toolkit.nomad-coe.eu/home/ (accessed: August 2019).
[134] T. Kluyver, B. Ragan-Kelley, F. Pérez, B. Granger, M. Bussonnier, J. Frederic, K. Kelley, J. Hamrick, J. Grout, S. Corlay, P. Ivanov, D. Avila, S. Abdalla, C. Willing (Eds: F. Loizides, B. Schmidt), *Positioning and Power in Academic Publishing: Players, Agents and Agendas*. IOS Press, Clifton, VA, USA **2016**, 87–90.
[135] A. H. Larsen, J. J. Mortensen, J. Blomqvist, I. E. Castelli, R. Christensen, M. Dułak, J. Friis, M. N. Groves, B. Hammer, C. Hargus, E. D. Hermes, P. C. Jennings, P. B. Jensen, J. Kermode, J. R. Kitchin, E. L. Kolsbjerg, J. Kubal, K. Kaasbjerg, S. Lysgaard, J. B. Maronsson, T. Maxson, T. Olsen, L. Pastewka, A. Peterson, C. Rostgaard, J. Schiøtz, O. Schütt, M. Strange, K. S. Thygesen, T. Vegge, et al., *J. Phys. Condens. Matter* **2017**, *29*, 273002.
[136] S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson, G. Ceder, *Comput. Mater. Sci.* **2013**, *68*, 314.
[137] qmpy, https://oqmd.org/static/docs/index.html (accessed: August 2019).

[138] NOMAD CoE, nomad-lab, https://gitlab.mpcdf.mpg.de/nomad-lab (accessed: August 2019).
[139] K. Alberi, M. B. Nardelli, A. Zakutayev, L. Mitas, S. Curtarolo, A. Jain, M. Fornari, N. Marzari, I. Takeuchi, M. L. Green, M. Kanatzidis, M. F. Toney, S. Butenko, B. Meredig, S. Lany, U. Kattner, A. Davydov, E. S. Toberer, V. Stevanovic, A. Walsh, N.-G. Park, A. Aspuru-Guzik, D. P. Tabor, J. Nelson, J. Murphy, A. Setlur, J. Gregoire, H. Li, R. Xiao, A. Ludwig, et al., *J. Phys. D* **2018**, *52*, 013001.
[140] R. Garud, S. Jain, A. Kumaraswamy, *Acad. Manag. J.* **2002**, *45*, 196.
[141] N. Brunsson, A. Rasche, D. Seidl, *Organ. Stud.* **2012**, *33*, 613.
[142] R. C. Johnson, IBM Launches Accelerated Discovery Lab, https://www.eetimes.com/document.asp?doc_id=1319758 (accessed: August 2019).
[143] ASM International online databases, https://www.asminternational.org/materials-resources/online-databases (accessed: August 2019).
[144] B. Meredig, *Curr. Opin. Solid State Mater. Sci.* **2016**, *21*, 159.
[145] L. Himanen, P. Rinke, A. S. Foster, *npj Comput. Mater.* **2018**, *4*, 52.
[146] B. Smith, *Blackwell Guide to the Philosophy of Computing and Information* (Ed: L. Floridi). Blackwell Publishers, Cambridge, MA, USA **2003**, pp. 155–166.
[147] L. Feigenbaum, I. Herman, T. Hongsermeier, E. Neumann, S. Stephens, *Sci. Am.* **2007**, *297*, 90.
[148] T. Berners-Lee, J. Hendler, O. Lassila, *Sci. Am.* **2001**, *284*, 34.
[149] NOMAD Meta Info, https://metainfo.nomad-coe.eu/nomadmetainfo_public/archive.html (accessed: August 2019).
[150] E. B. Tadmor, R. S. Elliott, J. P. Sethna, R. E. Miller, C. A. Becker, *JOM* **2011**, *63*, 17.
[151] P. E. Vet, P.-h. Speel, N. Mars, in *Proc. of 11th European Conference on Artificial Intelligence (ECAI'94)* (Ed: A. G. Cohn), John Wiley & Sons, New York, NY, USA **1994**, pp. 187–205.
[152] C. de Sainte Marie, M. Iglesias Escudero, P. Rosina, In S. Rudolph, C. Gutierrez, editors, *Web Reasoning and Rule Systems*. Springer Berlin Heidelberg, Berlin, Heidelberg, **2011**, pp. 24–29.
[153] V. Premkumar, S. Krishnamurty, J. C. Wileden, I. R. Grosse, *Adv. Eng. Inform.* **2014**, *28*, 91.
[154] K. Michel, B. Meredig, *MRS Bull.* **2016**, *41*, 8, 617.
[155] T. Ashiron, *Data Sci. J.* **2010**, *9*, 54.
[156] European Materials Modelling Council, Report on Workshop on Interoperability in Materials Modelling, **2017**.
[157] K. Cheung, J. Drennan, J. Hunter, in *AAAI Spring Symposium: Semantic Scientific Knowledge Integration* (Eds: D. L. McGuinness, P. Fox, B. Boyan), The AAAI Press, Menlo Park, CA, USA **2008**, pp. 9–14.
[158] M. Bhat, S. Shah, P. Das, P. Kumar, N. Kulkarni, S. S. Ghaisas, S. S. Reddy, in *ICoRD'13* (Eds: A. Chakrabarti, R. V. Prakash), Springer India, India **2013**, pp. 1315–1329.
[159] X. Zhang, C. Hu, H. Li, *Data Sci. J.* **2009**, *8*, 1.
[160] X. Zhang, C. Zhao, X. Wang, *Comput. Ind.* **2015**, *73*, 8.
[161] Qresp: Curation and Exploration of Reproducible Scientific Papers, https://qresp.org (accessed: August 2019).
[162] M. C. Swain, J. M. Cole, *J. Chem. Inf. Model* **2016**, *56*, 1894.
[163] E. Kim, K. Huang, A. Saunders, A. McCallum, G. Ceder, E. Olivetti, *Chem. Mater.* **2017**, *29*, 9436.
[164] Y. Hinuma, A. Togo, H. Hayashi, I. Tanaka, e-prints, arXiv:1506.01455, **2015**.
[165] D. Hicks, C. Oses, E. Gossett, G. Gomez, R. Taylor, C. Toher, M. J. Mehl, O. Levy, S. Curtarolo, *Acta Crystallogr. A* **2018**, *74*.
[166] A. Maedche, S. Staab, *IEEE Intell. Syst.* **2001**, *16*, 72.
[167] A. Copestake, P. Corbett, P. Murray-Rust, C. Rupp, A. Siddharthan, S. Teufel, B. Waldron, in *Proc. of the UK e-Science Programme All Hands Meeting 2006 (AHM2006)* (Ed: S. J. Cox), National e-Science Centre, Edinburgh, Scotland **2006**.
[168] International Materials Resource Registries WG, https://www.rd-alliance.org/groups/working-group-international-materials-resource-registries.html (accessed: August 2019).
[169] OPTiMaDe, https://www.optimade.org (accessed: August 2019).
[170] K. Kennedy, T. Stefansky, G. Davy, V. F. Zackay, E. R. Parker, *J. Appl. Phys.* **1965**, *36*, 3808.
[171] J. J. Hanak, *J. Mater. Sci.* **1970**, *5*, 964.
[172] X. D. Xiang, X. Sun, G. Briceno, Y. Lou, K.-A. Wang, H. Chang, W. G. Wallace-Freedman, S.-W. Chen, P. G. Schultz, *Science* **1995**, *268*, 1738.
[173] H. Koinuma, M. Kawasaki, T. Itoh, A. Ohtomo, M. Murakami, Z. Jin, Y. Matsumoto, *Physica C* **2000**, *335*, 245.
[174] T. Fukumura, M. Ohtani, M. Kawasaki, Y. Okimoto, T. Kageyama, T. Koida, T. Hasegawa, Y. Tokura, H. Koinuma, *Appl. Phys. Lett.* **2000**, *77*, 3426.
[175] M. Kneiß, P. Storm, G. Benndorf, M. Grundmann, H. von Wenckstern, *ACS Comb. Sci.* **2018**, *20*, 643.
[176] H. Koinuma, I. Takeuchi, *Nat. Mater.* **2004**, *3*, 429.
[177] K. Rajan, *Annu. Rev. Mater. Res.* **2008**, *38*, 299.
[178] R. Potyrailo, K. Rajan, K. Stoewe, I. Takeuchi, B. Chisholm, H. Lam, *ACS Comb. Sci.* **2011**, *13*, 579.
[179] T. Chikyo, *Sci. Technol. Adv. Mater.* **2011**, *12*, 050301.
[180] H. von Wenckstern, D. Splith, A. Werner, S. Müller, M. Lorenz, M. Grundmann, *ACS Comb. Sci.* **2015**, *17*, 710.
[181] S. Curtarolo, W. Setyawan, G. L. Hart, M. Jahnatek, R. V. Chepulskii, R. H. Taylor, S. Wang, J. Xue, K. Yang, O. Levy, M. J. Mehl, H. T. Stokes, D. O. Demchenko, D. Morgan, *Comput. Mater. Sci.* **2012**, *58*, 218.
[182] S. Curtarolo, G. L. W. Hart, M. B. Nardelli, N. Mingo, S. Sanvito, O. Levy, *Nat. Mater.* **2013**, *12*, 191.
[183] C. Schober, K. Reuter, H. Oberhofer, *J. Phys. Chem. Lett.* **2016**, *7*, 3973.
[184] C. Kunkel, C. Schober, J. T. Margraf, K. Reuter, H. Oberhofer, *Chem. Mater.* **2019**, *31*, 969.
[185] A. Jain, G. Hautier, C. J. Moore, S. P. Ong, C. C. Fischer, T. Mueller, K. A. Persson, G. Ceder, *Comput. Mater. Sci.* **2011**, *50*, 2295.
[186] C. E. Calderon, J. J. Plata, C. Toher, C. Oses, O. Levy, M. Fornari, A. Natan, M. J. Mehl, G. Hart, M. B. Nardelli, S. Curtarolo, *Comput. Mater. Sci.* **2015**, *108*, 233.
[187] W. Setyawan, S. Curtarolo, *Comput. Mater. Sci.* **2010**, *49*, 299.
[188] A. Jain, S. P. Ong, W. Chen, B. Medasani, X. Qu, M. Kocher, M. Brafman, G. Petretto, G.-M. Rignanese, G. Hautier, D. Gunter, K. A. Persson, *Concur. Comp. Pract. E* **2015**, *27*, 5037.
[189] K. Mathew, J. H. Montoya, A. Faghaninia, S. Dwarakanath, M. Aykol, H. Tang, I. heng Chu, T. Smidt, B. Bocklund, M. Horton, J. Dagdelen, B. Wood, Z.-K. Liu, J. Neaton, S. P. Ong, K. Persson, A. Jain, *Comput. Mater. Sci.* **2017**, *139*, 140.
[190] G. Pizzi, A. Cepellotti, R. Sabatini, N. Marzari, B. Kozinsky, *Comput. Mater. Sci.* **2016**, *111*, 218.
[191] A. R. Supka, T. E. Lyons, L. Liyanage, P. D'Amico, R. A. R. A. Orabi, S. Mahatara, P. Gopal, C. Toher, D. Ceresoli, A. Calzolari, S. Curtarolo, M. B. Nardelli, M. Fornari, *Comput. Mater. Sci.* **2017**, *136*, 76.
[192] K. Lejaeghere, G. Bihlmayer, T. Björkman, P. Blaha, S. Blügel, V. Blum, D. Caliste, I. E. Castelli, S. J. Clark, A. Dal Corso, S. de Gironcoli, T. Deutsch, J. K. Dewhurst, I. Di Marco, C. Draxl, M. Dułak, O. Eriksson, J. A. Flores-Livas, K. F. Garrity, L. Genovese, P. Giannozzi, M. Giantomassi, S. Goedecker, X. Gonze, O. Grånäs, E. K. U. Gross, A. Gulans, F. Gygi, D. R. Hamann, P. J. Hasnip, et al., *Science* **2016**, *351*, 6280.

[193] K. Lejaeghere, V. Van Speybroeck, G. Van Oost, S. Cottenier, *Crit. Rev. Solid State* **2014**, *39*, 1.
[194] I. Y. Zhang, A. J. Logsdail, X. Ren, S. V. Levchenko, L. Ghiringhelli, M. Scheffler, *New J. Phys.* **2019**, *21*, 013025.
[195] L. A. Curtiss, K. Raghavachari, P. C. Redfern, J. A. Pople, *J. Chem. Phys.* **2000**, *112*, 7374.
[196] S. Kirklin, J. E. Saal, B. Meredig, A. Thompson, J. W. Doak, M. Aykol, S. Rühl, C. Wolverton, *npj Comput. Mater.* **2015**, *1*, 15010.
[197] Error estimates from high-accuracy electronic structure reference calculations, online notebook, https://analytics-toolkit.nomad-coe.eu/notebook-edit/data/shared/tutorialsNew/errorbars/errorbars_html.bkr (accessed: August 2019).
[198] T. Mitchell, *Machine Learning*, McGraw-Hill, New York **1997**.
[199] S. Russell, *Artificial Intelligence: A Modern Approach*, Prentice Hall, Upper Saddle River, NJ **2010**.
[200] O. Isayev, C. Oses, C. Toher, E. Gossett, S. Curtarolo, A. Tropsha, *Nat. Commun.* **2017**, *8*, 15679.
[201] T. Xie, J. C. Grossman, *Phys. Rev. Lett.* **2018**, *120*, 145301.
[202] L. Ward, R. Liu, A. Krishna, V. I. Hegde, A. Agrawal, A. Choudhary, C. Wolverton, *Phys. Rev. B* **2017**, *96*, 1.
[203] M. Rupp, A. Tkatchenko, K.-R. Müller, O. A. von Lilienfeld, *Phys. Rev. Lett.* **2012**, *108*, 058301.
[204] K. T. Schütt, H. E. Sauceda, P. J. Kindermans, A. Tkatchenko, K. R. Müller, *J. Chem. Phys.* **2018**, *148*, 24.
[205] F. A. Faber, A. S. Christensen, B. Huang, O. A. Von Lilienfeld, *J. Chem. Phys.* **2018**, *148*, 24.
[206] K. Ghosh, A. Stuke, M. Todorović, P. B. Jørgensen, M. N. Schmidt, A. Vehtari, P. Rinke, *Adv. Sci.* **2019**, *6*, 1801367.
[207] A. Stuke, M. Todorović, M. Rupp, C. Kunkel, K. Ghosh, L. Himanen, P. Rinke, *J. Chem. Phys.* **2019**, *150*, 204121.
[208] I. T. Jolliffe, *Principal Component Analysis*, Springer, New York, NY **2002**.
[209] K. Rajan, C. Suh, P. F. Mendez, *Stat. Anal. Data Min.* **2009**, *1*, 361.
[210] L. van der Maaten, *J. Mach. Learn. Res.* **2008**, *9*, 2579.
[211] M. Ceriotti, G. A. Tribello, M. Parrinello, *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 13023.
[212] C. Bishop, *Pattern Recognition and Machine Learning*, Springer, New York **2006**.
[213] A. Samanta, B. Li, R. E. Rudd, T. Frolov, *Nat. Commun.* **2018**, *9*.
[214] G. Lima Guimaraes, B. Sanchez-Lengeling, C. Outeiral, P. L. Cunha Farias, A. Aspuru-Guzik, *e-prints, arXiv:1705.10843*, **2017**.
[215] L. M. Ghiringhelli, J. Vybiral, S. V. Levchenko, C. Draxl, M. Scheffler, *Phys. Rev. Lett.* **2015**, *114*, 105503.
[216] R. Ouyang, S. Curtarolo, E. Ahmetcik, M. Scheffler, L. M. Ghiringhelli, *Phys. Rev. Mater.* **2018**, *2*, 083802.
[217] A. Ziletti, D. Kumar, M. Scheffler, L. M. Ghiringhelli, *Nat. Commun.* **2018**, *9*, 1.
[218] O. Isayev, D. Fourches, E. N. Muratov, C. Oses, K. Rasch, A. Tropsha, S. Curtarolo, *Chem. Mater.* **2015**, *27*, 735.
[219] V. Stanev, C. Oses, I. Takeuchi, *npj Comput. Mater.* **2018**, *4*, 29.
[220] L. Ward, A. Dunn, A. Faghaninia, N. Zimmermann, S. Bajaj, Q. Wang, J. Montoya, J. Chen, K. Bystrom, M. Dylla, K. Chard, M. Asta, K. Persson, G. Snyder, I. Foster, A. Jain, *Comput. Mater. Sci.* **2018**, *152*, 60.
[221] M. Haghighatlari, J. Hachmann, ChemML—A Machine Learning and Informatics Program Suite for Chemical and Materials Data Mining, https://hachmannlab.github.io/chemml.
[222] DeepChem, https://deepchem.io (accessed: August 2019).
[223] Y. Zhuo, A. M. Tehrani, J. Brgoch, *J. Phys. Chem. Lett.* **2018**, *9*, 1668.
[224] J. S. Smith, O. Isayev, A. E. Roitberg, *Chem. Sci.* **2017**, *8*, 3192.
[225] A. P. Bartók, M. C. Payne, R. Kondor, G. Csányi, *Phys. Rev. Lett.* **2010**, *104*, 1.
[226] Z. Li, J. R. Kermode, A. De Vita, *Phys. Rev. Lett.* **2015**, *114*, 096405.
[227] S. De, A. P. Bartók, G. Csányi, M. Ceriotti, *Phys. Chem. Chem. Phys.* **2016**, *18*, 13754.
[228] R. Liu, A. Kumar, Z. Chen, A. Agrawal, V. Sundararaghavan, A. Choudhary, *Sci. Rep.* **2015**, *5*, 1.
[229] O. Isayev, C. Oses, C. Toher, E. Gossett, S. Curtarolo, A. Tropsha, *Nat. Commun.* **2017**, *8*, 15679.
[230] G. Montavon, K. Hansen, S. Fazli, M. Rupp, F. Biegler, A. Ziehe, A. Tkatchenko, A. V. Lilienfeld, K.-R. Müller, in *Advances in Neural Information Processing Systems*, Vol. 25 (Eds: F. Pereira, C. J. C. Burges, L. Bottou, K. Q. Weinberger), Curran Associates, Red Hook, NY, USA **2012**, pp. 440–448.
[231] K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O. A. Von Lilienfeld, K. R. Müller, A. Tkatchenko, *J. Phys. Chem. Lett.* **2015**, *6*, 2326.
[232] F. Faber, A. Lindmaa, O. A. v. Lilienfeld, R. Armiento, *Int. J. Quantum Chem.* **2015**, *115*, 1094.
[233] A. P. Bartók, R. Kondor, G. Csányi, *Phys. Rev. B* **2013**, *87*, 184115.
[234] H. Huo, M. Rupp, *e-prints, , arXiv:1704.06439*, **2017**.
[235] J. Behler, *J. Chem. Phys.* **2011**, *134*, 074106.
[236] M. Gastegger, L. Schwiedrzik, M. Bittermann, F. Berzsenyi, P. Marquetand, *J. Chem. Phys.* **2018**, *148*, 24.
[237] F. A. Faber, A. Lindmaa, O. A. Von Lilienfeld, R. Armiento, *Phys. Rev. Lett.* **2016**, *117*, 2.
[238] W. Pronobis, A. Tkatchenko, K. R. Müller, *J. Chem. Theory Comput.* **2018**.
[239] L. Himanen, M. O. J. Jäger, E. V. Morooka, F. Federici Canova, Y. S. Ranawat, D. Z. Gao, P. Rinke, A. S. Foster, e-prints, arXiv:1904.08875, **2019**.
[240] A. S. Christensen, F. A. Faber, B. Huang, L. A. Bratholm, A. Tkatchenko, K.-R. Muller, O. A. von Lilienfeld, QML: A Python Toolkit for Quantum Machine Learning, https://github.com/qmlcode/qml (accessed: August 2019).
[241] I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning*, MIT Press, Cambridge, MA, USA **2016**.
[242] D. H. Wolpert, W. G. Macready, *IEEE Trans. Evol. Comput.* **1997**, *1*, 67.
[243] J. Schmidhuber, *Neural Netw.* **2015**, *61*, 85.
[244] L. Rokach, O. Maimon, *Data Mining With Decision Trees: Theory and Applications*, World Scientific Publishing Co., Inc., River Edge, NJ, USA, 2nd edition **2014**.
[245] J. Shawe-Taylor, N. Cristianini, *Kernel Methods for Pattern Analysis*, Cambridge University Press, New York, NY, USA **2004**.
[246] C. Rasmussen, *Gaussian Processes for Machine Learning*, MIT Press, Cambridge, MA **2006**.
[247] Chollet, François and others, Keras, https://keras.io (accessed: August 2019).
[248] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, X. Zheng TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, https://www.tensorflow.org.
[249] T. Chen, C. Guestrin, in, *Proc. of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD'16. ACM, New York, NY, USA, **2016**, pp. 785–794.
[250] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell, in *Proc. of the 22Nd ACM Int. Conf. on Multimedia* (Eds: K. A. Hua, Y. Rui, R. Steinmetz, A. Hanjalic, A. Natsev, W. Zhu), MM '14. ACM, New York, NY, USA **2014**.
[251] Theano Development Team, *e-prints, abs/1605.02688*, **2016**.

[252] PyTorch, https://pytorch.org (accessed: August 2019).
[253] GPy: A Gaussian process framework in python, https://github.com/SheffieldML/GPy (accessed: August 2019).
[254] Eclipse Deeplearning4j Development Team, Deeplearning4j: Open-source distributed deep learning for the jvm, apache software foundation license 2.0, https://deeplearning4j.org.
[255] F. Arbabzadah, S. Chmiela, K. R. Müller, A. Tkatchenko, *Nat. Commun.* **2017**, *8*, 6.
[256] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, G. E. Dahl, in *Proc. of the 34th Int. Conf. on Machine Learning* (Eds: D. Precup, Y. W. Teh), PMLR, Sydney, Australia **2017**, p. 1263.
[257] M. Todorović, M. U. Gutmann, J. Corander, P. Rinke, *npj Comput. Mater.* **2019**, *5*, 35.
[258] B. R. Goldsmith, M. Boley, J. Vreeken, M. Scheffler, L. M. Ghiringhelli, *New J. Phys.* **2017**, *19*, 013031.
[259] Y. LeCun, Y. Bengio, In M. A. Arbib, editor, *Convolutional Networks for Images, Speech, and Time Series*. MIT Press, Cambridge, MA, USA, **1998**, pp. 255–258.
[260] B. Meredig, E. Antono, C. Church, M. Hutchinson, J. Ling, S. Paradiso, B. Blaiszik, I. Foster, B. Gibbons, J. Hattrick-Simpers, A. Mehta, L. Ward, *Mol. Syst. Des. Eng.* **2018**, *3*, 819.
[261] J. Vanschoren, J. N. van Rijn, B. Bischl, L. Torgo, *SIGKDD Explor.* **2013**, *15*, 49.
[262] R. Chard, Z. Li, K. Chard, L. Ward, Y. Babuji, A. Woodard, S. Tuecke, B. Blaiszik, M. J. Franklin, I. Foster, *e-prints, arXiv:1811.11213*, **2018**.
[263] G. Irving, A. Askell, *Ai safety needs social scientists* **2019**, https://doi.org/10.23915/distill.00014 (accessed: August 2019).
[264] R. Gómez-Bombarelli, J. Aguilera-Iparraguirre, T. D. Hirzel, D. Duvenaud, D. Maclaurin, M. A. Blood-Forsythe, H. S. Chae, M. Einzinger, D. G. Ha, T. Wu, G. Markopoulos, S. Jeon, H. Kang, H. Miyazaki, M. Numata, S. Kim, W. Huang, S. I. Hong, M. Baldo, R. P. Adams, A. Aspuru-Guzik, *Nat. Mater.* **2016**, *15*, 1120.
[265] A. Mannodi-Kanakkithodi, G. M. Treich, T. D. Huan, R. Ma, M. Tefferi, Y. Cao, G. A. Sotzing, R. Ramprasad, *Adv. Mater.* **2016**, *28*, 6277.
[266] A. O. Oliynyk, E. Antono, T. D. Sparks, L. Ghadbeigi, M. W. Gaultois, B. Meredig, A. Mar, *Chem. Mater.* **2016**, *28*, 7324.
[267] D. Xue, P. V. Balachandran, J. Hogden, J. Theiler, D. Xue, T. Lookman, *Nat. Commun.* **2016**, *7*, 1.
[268] D. Xue, P. V. Balachandran, R. Yuan, T. Hu, X. Qian, E. R. Dougherty, T. Lookman, *Proc. Natl. Acad. Sci. USA* **2016**, *113*, 47, 13301.
[269] F. Ren, L. Ward, T. Williams, K. J. Laws, C. Wolverton, J. Hattrick-Simpers, A. Mehta, *Sci. Adv.* **2018**, *4*, 4.
[270] C. Wen, Y. Zhang, C. Wang, D. Xue, Y. Bai, S. Antonov, L. Dai, T. Lookman, Y. Su, *Acta Mater.* **2019**, *170*, 109.
[271] L. Yu, A. Zunger, *Phys. Rev. Lett.* **2012**, *108*, 068701.
[272] D. J. Baquião, G. M. Dalpian, *Comput. Mater. Sci.* **2019**, *158*, 382.
[273] I. E. Castelli, F. Hüser, M. Pandey, H. Li, K. S. Thygesen, B. Seger, A. Jain, K. A. Persson, G. Ceder, K. W. Jacobsen, *Adv. Energy Mater.* **2015**, *5*, 1400915.
[274] I. E. Castelli, T. Olsen, S. Datta, D. D. Landis, S. Dahl, K. S. Thygesen, K. W. Jacobsen, *Energy Environ. Sci.* **2012**, *5*, 5814.
[275] K. Yang, W. Setyawan, S. Wang, M. Buongiorno Nardelli, S. Curtarolo, *Nat. Mater.* **2012**, *11*, 614.
[276] G. Cao, H. Liu, X.-Q. Chen, Y. Sun, J. Liang, R. Yu, Z. Zhang, *Sci. Bull.* **2017**, *62*, 1649.
[277] G. Hautier, C. C. Fischer, A. Jain, T. Mueller, G. Ceder, *Chem. Mater.* **2010**, *22*, 3762.
[278] B. Meredig, A. Agrawal, S. Kirklin, J. E. Saal, J. W. Doak, A. Thompson, K. Zhang, A. Choudhary, C. Wolverton, *Phys. Rev. B* **2014**, *094104*, 1.
[279] J. Timoshenko, D. Lu, Y. Lin, A. I. Frenkel, *J. Phys. Chem. Lett.* **2017**, *8*, 5091.
[280] J. Timoshenko, A. Anspoks, A. Cintins, A. Kuzmin, J. Purans, A. I. Frenkel, *Phys. Rev. Lett.* **2018**, *120*, 225502.
[281] C. Zheng, K. Mathew, C. Chen, Y. Chen, H. Tang, A. Dozier, J. J. Kas, F. D. Vila, J. J. Rehr, L. F. J. Piper, K. A. Persson, S. P. Ong, *npj Comput. Mater.* **2018**, *4*, 12.
[282] R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams, A. Aspuru-Guzik, *ACS Cent. Sci.* **2018**, *4*, 268.
[283] M. J. Kusner, B. Paige, J. M. Hernández-Lobato, in *Proc. of the 34th Int. Conf. on Machine Learning* (Eds: D. Precup, Y. W. Teh), PMLR, Sydney, Australia, **2017**, pp. 1945–1954.
[284] M. Schmidt, H. Lipson, *Science* **2009**, *324*, 81.
[285] S. H. Rudy, S. L. Brunton, J. L. Proctor, J. N. Kutz, *Sci. Adv.* **2017**, *3*, 4.
[286] M. Raissi, P. Perdikaris, G. Karniadakis, *J. Comput. Phys.* **2019**, *378*, 686.
[287] C. J. Bartel, C. Sutton, B. R. Goldsmith, R. Ouyang, C. B. Musgrave, L. M. Ghiringhelli, M. Scheffler, *Sci. Adv.* **2019**, *5*, 2.
[288] A. S. M. Jonayat, A. C. T. van Duin, M. J. Janik, *ACS Appl. Energy Mater.* **2018**, *1*, 6217.
[289] L. Wang, *Phys. Rev. B* **2016**, *94*, 195105.
[290] A. J. Chowdhury, W. Yang, E. Walker, O. Mamun, A. Heyden, G. A. Terejanu, *J. Phys. Chem. C* **2018**, *122*, 28142.
[291] P. Schwaller, T. Laino, T. Gaudin, P. Bolgar, C. Bekas, A. A. Lee, *e-prints, arXiv:1811.02633*, **2018**.
[292] T. D. Huan, R. Batra, J. Chapman, S. Krishnan, L. Chen, R. Ramprasad, *npj Comput. Mater.* **2017**, *3*, 37.
[293] M. A. Caro, V. L. Deringer, J. Koskinen, T. Laurila, G. Csányi, *Phys. Rev. Lett.* **2018**, *120*, 166101.
[294] N. Artrith, A. M. Kolpak, *Nano Lett.* **2014**, *14*, 2670.
[295] V. Botu, R. Batra, J. Chapman, R. Ramprasad, *J. Phys. Chem. C* **2017**, *121*, 511.
[296] A. Singraber, J. Behler, C. Dellago, *J. Chem. Theory Comput.* **2019**, *15*, 1827.
[297] H. Wang, L. Zhang, J. Han, W. E, *Comput. Phys. Commun* **2018**, *228*, 178.
[298] A. P. Bartók, J. Kermode, N. Bernstein, G. Csányi, *Phys. Rev. X* **2018**, *8*, 041048.
[299] J. T. Margraf, K. Reuter, *J. Phys. Chem. A* **2018**, *122*, 6343.
[300] J. S. Smith, B. T. Nebgen, R. Zubatyuk, N. Lubbers, C. Devereux, K. Barros, S. Tretiak, O. Isayev, A. Roitberg, ChemRxiv, Preprint, 2018.
[301] D. M. Wilkins, A. Grisafi, Y. Yang, K. U. Lao, R. A. DiStasio, M. Ceriotti, *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 3401.
[302] J. C. Snyder, M. Rupp, K. Hansen, K. R. Müller, K. Burke, *Phys. Rev. Lett.* **2012**, *108*, 1.
[303] F. Brockherde, L. Vogt, L. Li, M. E. Tuckerman, K. Burke, K. R. Müller, *Nat. Commun.* **2017**, *8*, 1.
[304] N. Granqvist, J. Laurila, *Organ. Stud.* **2011**, *32*, 253.
[305] A. Geurts, Ethnographic field notes from Database-Driven Materials Science & Technology study.
[306] A. Geurts, N. Granqvist, P. Rinke, manuscript in preparation.
[307] MaX - Materials design at the Exascale, https://www.max-centre.eu/ (accessed: August 2019).
[308] Future makers funding program 2018, https://techfinland100.fi/future-makers-funding-program-2018/ (accessed: August 2019).
[309] R. Katila, G. Ahuja, *Acad. Manage. J.* **2002**, *45*, 1183.
[310] R. Garud, P. Nayyar, *Strategic Manage. J.* **1994**, *15*, 365.
[311] A. Geurts, *Technol. Forecast. Soc. Change* **2018**, *128*, 311.
[312] M. A. Schilling, *Organ. Sci.* **2015**, *26*, 668.
[313] D. Teece, *Res. Policy* **1986**, *15*, 285.
[314] A. Gawer, M. A. Cusumano, *Platform Leadership How Intel, Microsoft, and Cisco Drive Industry Innovation*, Harvard Business Press, Brighton, MA, USA **2002**.