ISSN 1674-1056
http://cpb.iphy.ac.cn
http://iopscience.iop.org/cpb
中国物理 B
Chinese Physics
中国物理 Chinese Physics B

The materials data ecosystem: Materials data science and its role in data-driven materials discovery

Hai-Qing Yin(尹海清), Xue Jiang(姜雪), Guo-Quan Liu(刘国权), Sharon Elder, Bin Xu(徐斌), Qing-Jun Zheng(郑清军), Xuan-Hui Qu(曲选辉)

Citation: Chin. Phys. B . 2018, 27(11): 118101. **doi:** 10.1088/1674-1056/27/118101
Journal homepage: http://cpb.iphy.ac.cn; http://iopscience.iop.org/cpb

What follows is a list of articles you may be interested in

Combinatorial synthesis and high-throughput characterization of copper-oxide superconductors

J Wu, A T Bollinger, X He, I Božović
Chin. Phys. B . 2018, 27(11): 118102. **doi:** 10.1088/1674-1056/27/118102

Recent advances in non-Pb-based group-IV chalcogenides for environmentally-friendly thermoelectric materials

Bing-Sheng Du(杜炳生), Ji-Kang Jian(简基康), Hai-Tao Liu(刘海涛), Jiao Liu(刘骄), Lei Qiu(邱磊)
Chin. Phys. B . 2018, 27(4): 048102. **doi:** 10.1088/1674-1056/27/4/048102

Germanene nanomeshes:Cooperative effects of degenerate perturbation and uniaxial strain on tuning bandgap

Yan Su(苏燕), Xinyu Fan(范新宇)
Chin. Phys. B . 2017, 26(10): 108101. **doi:** 10.1088/1674-1056/26/10/108101

Geometrically induced π-band splitting in graphene superlattices

Yanpei Wei(魏艳佩), Tiantian Jia(贾甜甜), Gang Chen(陈刚)
Chin. Phys. B . 2017, 26(2): 028103. **doi:** 10.1088/1674-1056/26/2/028103

Templated synthesis of highly ordered mesoporous cobalt ferrite and its microwave absorption properties

Li Guo-Min, Wang Lian-Cheng, Xu Yao
Chin. Phys. B . 2014, 23(8): 088105. **doi:** 10.1088/1674-1056/23/8/088105

TOPICAL REVIEW — Physics research in materials genome

The materials data ecosystem: Materials data science and its role in data-driven materials discovery*

Hai-Qing Yin(尹海清)[1,2,3,†], Xue Jiang(姜雪)[1,3], Guo-Quan Liu(刘国权)[1,3], Sharon Elder[4], Bin Xu(徐斌)[1], Qing-Jun Zheng(郑清军)[5], and Xuan-Hui Qu(曲选辉)[1,2,6]

[1] *Collaborative Innovation Center of Steel Technology, University of Science and Technology Beijing, Beijing* 100083, *China*
[2] *Beijing Laboratory of Metallic Materials and Processing for Modern Transportation, University of Science and Technology Beijing, Beijing* 100083, *China*
[3] *Beijing Key Laboratory of Materials Genome Engineering, University of Science and Technology Beijing, Beijing* 100083, *China*
[4] *Computer Science and Engineering, The Pennsylvania State University, University Park, PA* 16802, *USA*
[5] *Kennametal Inc.*, 1600 *Technology Way Latrobe, PA* 15650, *USA*
[6] *Institute of Advanced Materials and Technology, University of Science and Technology Beijing, Beijing* 100083, *China*

(Received 7 May 2018; revised manuscript received 20 August 2018; published online 19 October 2018)

Since its launch in 2011, the Materials Genome Initiative (MGI) has drawn the attention of researchers from academia, government, and industry worldwide. As one of the three tools of the MGI, the use of materials data, for the first time, has emerged as an extremely significant approach in materials discovery. Data science has been applied in different disciplines as an interdisciplinary field to extract knowledge from data. The concept of materials data science has been utilized to demonstrate its application in materials science. To explore its potential as an active research branch in the big data era, a three-tier system has been put forward to define the infrastructure for the classification, curation and knowledge extraction of materials data.

Keywords: Materials Genome Initiative, materials data science, data classification, life-cycle curation

PACS: 81.05.Zx, 89.20.Ff **DOI:** 10.1088/1674-1056/27/11/118101

1. Introduction

As the practice of obtaining information and insight from data, data science has become a very familiar term to researchers from various disciplines.[1] This concept was first introduced in the 1960s and lasted for a few decades. In 1996, statistician CF Jeff Wu used the term again, describing the discipline as an extension of statistics. Nowadays, huge amounts of scientific data are produced by simulations, high-throughput scientific instruments, satellites, telescopes, and so on. The availability of big data is revolutionizing how research is conducted and has led to the emergence of a new paradigm in science based on data-intensive computing and analytics. Data science is defined as the fourth paradigm of data-intensive scientific discovery, alongside experimentation, theory and calculation.[2] Furthermore, the release of the Big Data R&D Initiative in 2012 has accelerated the development of data science.

Data science has been applied in diverse disciplines in recent years. An integrated data science pipeline is used to identify latent signals for QT-DDIs by using electrocardiogram data in electronic health records.[3] A data-driven approach was taken to simulate human mobility, provide spatial models, and predict the nationwide consequences of mass switching to electric vehicles.[4] To address the challenges of the prevailing development of big data, as well as the Data Science Journal, new journals including the Journal of Data Science and Analytics (JDSA)[5] and Data Science and Engineering (DSE)[6] were launched to stimulate scientific innovation and practice in data management and data-intensive applications.

In this study, we will introduce a data ecosystem for materials data, and aim to organize it into a coherent portrait of the scientific study of materials data, which are related to each other and to the materials science and engineering disciplines. The materials data ecosystem is comprised of data sources and data science. There are diverse kinds of materials data sources: publications, records from facilities and computation tools, third-party data, and so on, which are not covered in this text in detail. Materials data science, inherently cross-functional and at the very highest level of data study, is investigated here.

2. Data science in material science and engineering

In 1999, John R. Rodgers introduced the new concept of materials informatics and defined it as an effective data management tool for new materials discoveries. Somewhat later, Integrated Computational Materials Engineering

*Project supported by the National Key R&D Program of China (Grant No. 2016YFB0700503), the National High Technology Research and Development Program of China (Grant No. 2015AA03420), Beijing Municipal Science and Technology Project, China (Grant No. D161100002416001), the National Natural Science Foundation of China (Grant No. 51172018), and Kennametal Inc.

†Corresponding author. E-mail: hqyin@ustb.edu.cn

© 2018 Chinese Physical Society and IOP Publishing Ltd

(ICME, 2008) and the Materials Genome Initiative (MGI, 2011) attracted more attention worldwide on integrating computational capabilities, data management, and experimental techniques.[7] Although materials databases were built in many countries with universal access to abundant scientific data—and some have become fundamental to materials computation—materials data and materials informatics[8] received their first recognition when compared with computation and experimentation in materials innovation. The concept of materials data infrastructure was put forward based on the integration of ICME and materials informatics[9], however, the diversity of materials science has yet to be exhibited. Therefore, a system which enables one to virtually express real-world materials details as well as data mining needs to be built.

3. Infrastructure of materials data science

Materials data science is data science applied in materials science and engineering, aiming to take advantage of data to discover the science beneath the observed phenomena and production. It is a new, inherently cross-disciplinary approach. Currently, the market's ever-demanding requirements push people to need to understand the whole production chain of materials. Therefore, materials researchers strive to build a link between the five core points: chemical composition, microstructure, manufacturing, properties, and performance in service.

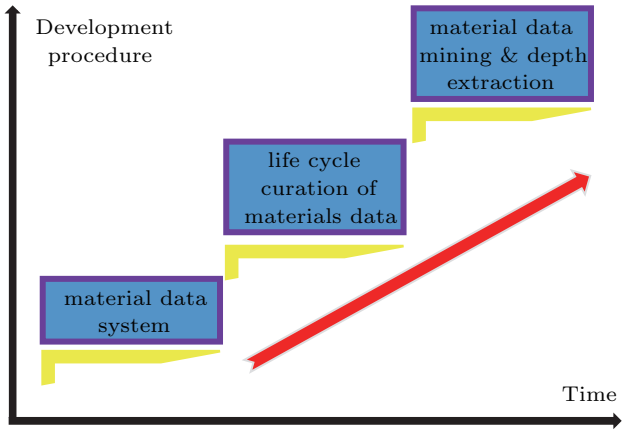


Fig. 1. (color online) Infrastructure of materials data science.

Knowledge engineering connects data with information, knowledge, and intelligence from a conceptual perspective. When combined with scientific disciplines, data are endowed with relevant meaning and the complicated correlations among data can be explained. To promote data science into a sub-field in materials science and engineering, a three-tier infrastructure of materials data science was introduced, as shown in Fig. 1, showing the main procedure of materials data development, that is, data production, management and application. The first and fundamental tier is the materials data system, the second tier is the life cycle curation of materials data, and the third and highest tier is materials data mining and deep extraction.

3.1. The scientific system of materials data

The generation of a single material datum represents a kind of attribute with the characteristics of a specific material, and may have the potential to be applied with limited scope according to the material's nature and its inherent correlation. As there are several methods of classification for materials, each of which has pros and cons, the classification of materials data is certainly the same. So, borrowed from the materials scientific data sharing network,[10] the materials data system consists entirely of 11 categories on the first level shown in Fig. 2, including fundamental materials, metal & alloys, ceramics, organic materials, composites, biomaterials, information materials, energy materials, natural materials, building materials, and road & transport materials data. With the materials data system, each single datum will be classified into a category, which will provide the following data curation steps and ultimately the data applications for the specific groups of materials. In the data system, the fundamental data differ from the other categories of certain materials, with the primary role being to provide overall information on single chemical elements. The crystal structure and properties such as the thermodynamic parameters of the binary, ternary, and other combinations of elements are also pivotal contents of the fundamental data category.

The categories can be further classified into more sublayers according to the classification system of each individual material. Furthermore, the system is also feasible as a basis for knowledge to construct a metadata system for the materials.

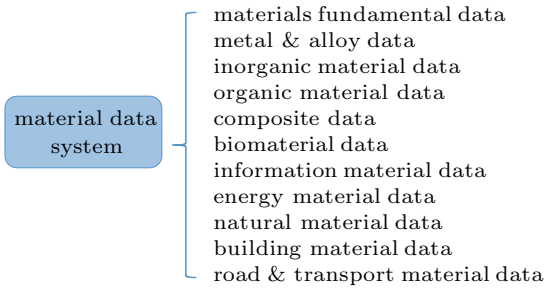


Fig. 2. (color online) The first level of the materials data system, borrowed from the materials scientific data sharing network. When data science is incorporated into materials science, the material theories as well as the data standards are integrated to ensure accurate data representation in the materials discipline.

3.2. Life cycle curation of materials data

Data, which are the product in the data era, exhibit common characteristics of real commodities and experience a similar development process. The life cycle of data involves production, storage, updates, management, publication, application, and finally deletion or long-term storage for re-use. The comparison of the life cycle process, shown in Fig. 3, indicates the similarity between scientific data and industrial products.

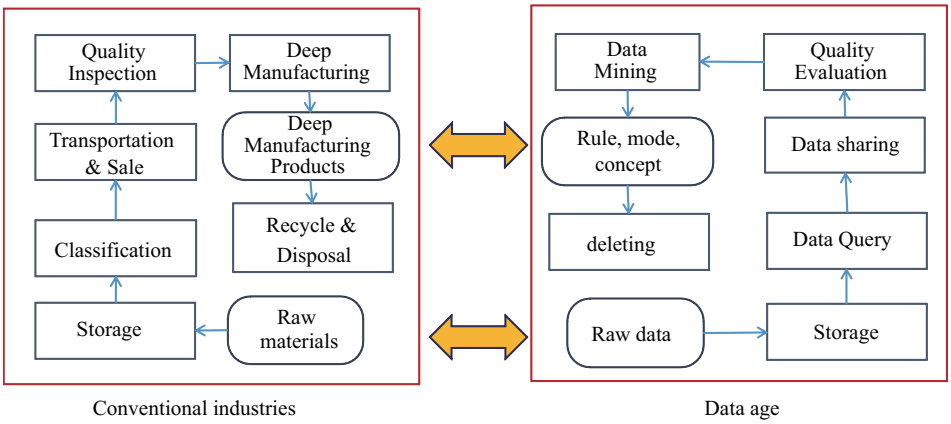


Fig. 3. (color online) Analogy of product processes in the data era and in the real economy.

3.2.1. Description system of materials data

The description of scientific data is associated with storage and presentation activities. Based on the materials data system, as well as data sources from computation, experimentation, characterization and industry, we are drafting a standard for materials scientific data description, whereby the materials data are divided into three groups, that is, the experimental data, the computational data, and the production data. The experimental data are further divided into two sub-groups, one is the experimental data of bulk materials and the other is the data on specific subjects such as coatings and corrosion.

For different groups and sub-groups, the attributes being collected in the databases are different due to their inherent generation and usage, which requires comprehensive information, covering everything from the quantum to the macro-scale in academic activities, as described in ICME. As the MGI describes, the goals of time-halving and cost-halving will be fulfilled when endeavors focus on an innovative re-use of the data. To meet the requirement for innovative material discovery, it is mandatory to add a detailed description of the material production process for experiments, and prerequisites for computational data, which were occasionally omitted in the past. In the past, due to the limited approaches in materials research, most of the information about data generation processes was omitted, and the re-use of data was restricted solely to questions of materials' performance.

To emphasize the whole chain of materials production and optimization, the integrity of each item of data is especially significant. The key attributes for the three groups of materials data are listed in Table 1. It should be noted that data from both the intermediate stages and the non-optimized samples are collected due to their potential for materials design[11] and optimization.

Table 1. Attributes of the computational, experimental, and production data in the infrastructure of the MGI and data mining.

Data type		Attributes in general (besides the data source and data collector)
Experimental data	data on bulk materials	brand/material name chemical compositions raw material information properties data, testing conditions, equipment & institution unstructured/descriptive file (eg. SEM, image) processing parameters
	data on specific research	basic material information (brand, material name, chemical compositions, properties data, supplier) properties data, testing conditions, equipment & institution unstructured/descriptive file (eg. SEM, image) processing parameters
Computational data		material name/structure software information model/algorithm and its source chemical compositions properties files/data constraints of computation (prerequisites, initial conditions, boundaries)
Industrial data		brand/material name properties data description of product company information

3.2.2. Storage of materials data

The database, an organized collection of data, has been the typical way to define, store, update, and administer data whereby the data are accessible to query and retrieve. The inorganic crystal structure database (ICSD), the Pauling file, databases for thermodynamic computation, and so on, are those that are specifically used and associated with the software of first-principles calculation, thermodynamics, and properties simulations. Others are mostly about the properties obtained from past research and industrial activities.

The database is well-developed for the curation of raw materials data, while data warehouses have appeared in recent years for data mining and to store specific-topic and integrated data from one or more disparate sources. Besides databases and data warehouses, cloud storage provides a brand new choice. Currently, cloud computing has been applied to provide Paas and Iaas services equipped with both hardware and software in some supercomputing centers and companies.

The cloud computing platform will definitely be utilized by more materials data researchers once the privacy and intellectual property issues in materials communities have been settled.

Therefore, databases, data warehouses, and cloud storage are three alternative candidates for materials researchers to optimize data storage of their own data resources.

3.2.3. Science-data sharing & publication

In recent years, the contradiction between data sharing and proprietary interest protection has become more obvious and is turning into a global problem to tackle. In academia, which is one of the most significant sources of scientific data, the owners of data are reluctant to share, or may even reject sharing, any supplemental information, as well as the final outcomes, which are essential to understanding or reproducing the data, due to the protection of intellectual property and potential commercial value, although government funding agencies and scientific journals require one to do so. In materials science and the like, there are some serious issues related to national security on some specific topics and it is prohibited to share data. Therefore, difficulty in discriminating where to draw boundaries in this context impedes communication on data. Tim Austin[12] considered data citation as a possible solution.

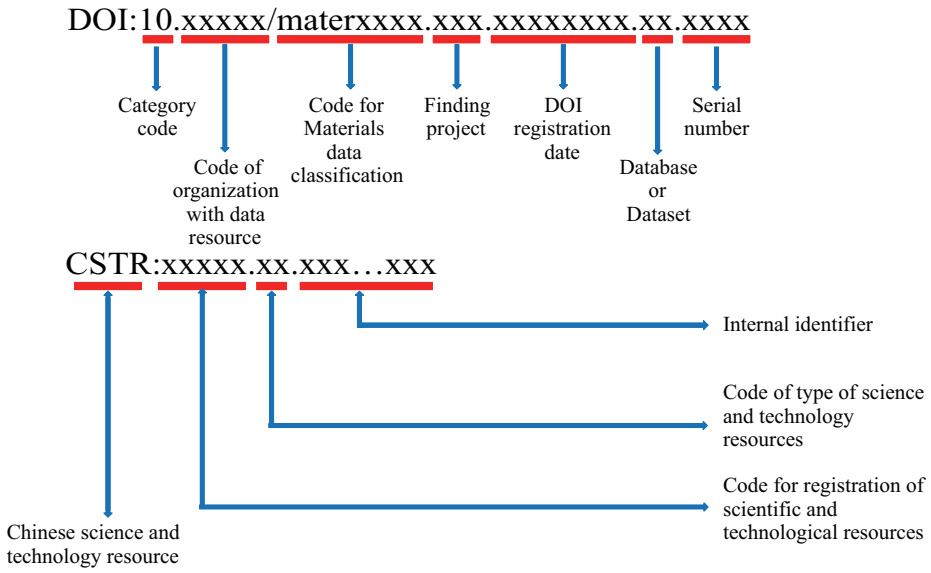


Fig. 4. (color online) The DOI and CSTR systems for identifying materials data.

The combination of two widely-available methods relating to papers, which are identification and citation, may be one feasible solution. The use of digital object identification (DOI) has been a common method worldwide for protection of the intellectual property of papers in publication in recent years,[13] by which the papers are uniquely labelled with a series of numbers and letters once the registration of DOI is done. Similarly, DOI for scientific data began for geographic data in China a few years ago[14] and now a formal and detailed format for registration and citation has been established. Later, a DOI system for materials data was founded in China, based on work on the National Materials Scientific Data Sharing Network, shown in Fig. 4. There are two parts to the system, coupled with each other to express the registered data/data set uniquely. For both systems, the data resource organization, usually a university or institution, states where the data come from. The code of materials data classification, where mater is the abbreviation for the term material, consists of two levels of materials data classification in Subsection 3.1, with the first level shown in Fig. 2. The science and technology resource identification (CSTR) system covers all the digital objects including data.

Publication of scientific data is regarded as one of the means of data sharing, as well as of evaluating the contributions of data collectors. The DOI/CSTR provide the information of the data as the metadata for database management and information querying and retrieval.[15] Sungbum Park *et al.* produced an IS success model for evaluating the application of the DOI system, and indicated that both data content including the features and information quality are significant factors to influence organizational benefits by means of perceived usefulness and user satisfaction.[16] Accordingly, the DOI of materials data should be implemented at the point when the data are collected and integrated into the databases due to its high correlation to the application of databases. Unlike in the field of human health data which is creating a global coalition of data resources,[17] currently, an internationally accessible materials data infrastructure hardly appears, however, DOI is paving the way towards this goal.

3.2.4. Data transfer in cross-scale modeling and simulation

Cross-scale modeling is an interesting topic following the exploitation of multiscale modeling.[18] Smart manufacturing of materials requires cross-scale modeling, simulation and control, by taking advantage of a combination of information and materials knowledge, where data are the fundamental elements and data transfer across scales is crucial. So the characteristics of big data in the smart manufacturing of materials are high dimension and complicated correlation rather than high volume.

Krishna Rajan pointed out that there currently lacks a unified way to explore patterns of behavior across correlative databases.[19] To bridge the gaps between the databases and cross-scale research activities, it is essential to understand the input and output for each scale, that is, the relevant attributes as prerequisites and boundary conditions for computation/experiment, and the results in a data format. A knowledge-based understanding of the exploitation process of powder metallurgy materials is shown in Fig. 5, where all the input parameters and the output are listed for scales of both the computation phase and the fabrication phase. As you can see, the output does not fit the input exactly on the subsequent scale, which indicates that a comprehensive understanding of the entire chain of material design and production will be the only solution. An interface is required to bridge the gaps in input and output between two scales which have a close relationship with each other when a cross-scale computation is expected.

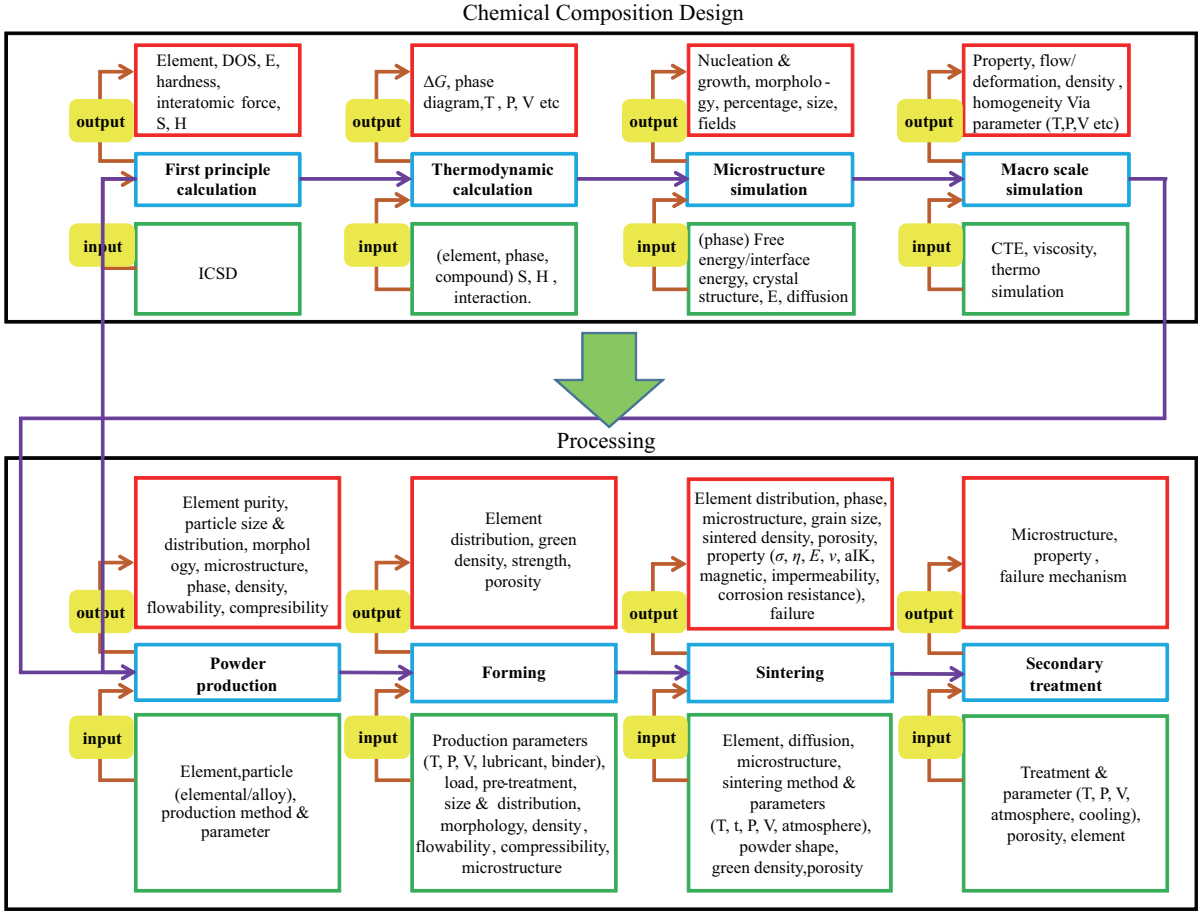


Fig. 5. (color online) Computation at different scales of modeling (chemical composition design) and simulation (processing optimization) with the input parameters and output, showing the feasibility of data transfer among scales.

3.3. Data-driven materials research

Data science is regarded as the fourth paradigm of data-intensive scientific discovery, alongside experimentation, theory, and calculation.[2] In materials science and engineering, with the emergence of materials informatics, integrated computational materials engineering and the materials genome, materials data are going beyond collection and integration and entering a new stage of application for the exploration and discovery of new or alternative materials, which is the core of data-driven materials research and a further step ahead in materials design. Materials informatics is becoming a methodology for data mining and machine learning in materials science.[20]

According to its functionality, the research of data mining in materials science is divided into two categories, one for the creation of new materials based mainly on first-principles calculations, and the other for the improvement of properties by optimizing composition and processing.

In the past few years, the breakthrough of combining the MGI and data mining has emerged swiftly. The discovery of brand-new functional materials candidates, especially for clean energy storage, has been frequently reported in the journals Science and Nature and the term materials code appeared in a cover article of the journal Nature.[21] High throughput first-principles calculations (HTCs) make it possible to obtain massive volumes of data, leading to the most abundant data resource for data mining and machine learning.[22] The Materials Project,[23] Automatic Flow for Materials Discovery (AFLOWLib), Open Quantum Materials Database (OQMD), Novel Materials Discovery (NoMaD) repository, CatApp Database, and Computational Materials Repository (CMR)[24] are the newly-established *ab initio* databases, where millions of data are integrated. By using the methods of principal component analysis, regression, neural networks, and Bayesian algorithms, materials with tailored properties have been discovered, such as $Ti_{50.0}Ni_{46.7}Cu_{0.8}Fe_{2.3}Pd_{0.2}$.[25,26]

The integration of materials design and processing optimization[27] boosts research into solving the problems for the full work flow.[28,29] In this case, data range from the calculated elements to the detailed parameters in fabrication, and data mining extends the ideas of ICME to extract semantic connections, which are central to solving tough problems of integration, cleaning, and analysis among the attributes in experimentation and large-scale production.[10,30]

4. Prospects and challenges

Materials data play a vital role in materials research. Industrial applications of materials data will be a positive stimulus for the systematic establishment and implementation of materials data science on research as well as education. Smart manufacturing aims to take advantage of advanced information and manufacturing technologies to enable flexibility in physical processes, therefore industria 4.0 enables one to apply the data to the whole work flow and the opportunity to push materials data science forward into a knowledge engineering system to realize artificial intelligence (AI) in materials innovation and production.

Materials data science, as a form of data science, is an interdisciplinary field which combines materials science with computer science and math, as well as physics and chemistry. Collaboration is urgently needed to move towards meeting core requirements and goals, one of which is to achieve the integration of materials theory and knowledge with the algorithms and methods of data mining and machine learning.

References

[1] Vasant D 2013 *Data Sci. Prediction Commun. ACM* **56** 64
[2] Hey T, Tansley S and Tolle K 2009 *The fourth paradigm: data-intensive scientific discovery* [M] (Washington: Microsoft Corporation) pp. 109–130
[3] Lorberbaum T, Sampson K J, Woosley R L, Kass R S and Tatonetti N P 2016 *Drug Saf.* **39** 433
[4] Janssens D, Giannotti F, Nanni M, Pedreschi D and Rinzivillo S 2012 *Künstl Intell.* **26** 275
[5] Cao L B 2016 *Int. J. Data Sci. Anal.* **1** 1
[6] Bertino E 2016 *Data Sci. Eng.* **1** 1
[7] https://www.whitehouse.gov/blog/2016/08/01/materials-genomeinitiative-first-five-years
[8] Hill J, Mulholl, G, Persson K and Seshadri R 2016 *MRS Bull.* **41** 399
[9] Wong T T 2016 *JOM* **68** 2029
[10] Shi C X 1994 *Materials Lexicon* (New York: Chemical Industry Press)
[11] Raccuglia P, Elbert K C, Adler P D F, Falk C, Wenny M B, Mollo A, Zeller M, Friedler S A, Schrier J and Norquist A J 2016 *Nature* **533** 73
[12] Austin T 2016 *Mater. Discovery* **3** 1
[13] Gorraiz J, Melero-Fuentes D, Gumpenberger C and Valderrama-Zurián J C 2016 *J. Informetrics* **10** 98
[14] Liu C 2014 *Acta Geographica Sin.* **69** 3
[15] Khedmatgozar H R and Alipour-Hafezi M 2017 *Int. J. Inf. Management* **37** 162
[16] Park S B, Zo H J, Ciganek A P and Lim G G 2011 *Electron. Commerce Res. Appl.* **10** 626
[17] Anderson W P 2017 *Nature* **543** 179
[18] Yan W T, Ge W J, Smith J, Lin S, Kafka O L, Lin F and Liu W K 2016 *Acta Mater.* **115** 403
[19] Rajan K 2013 *Informatics For Mater. Sci. Eng.* **9** 21
[20] Kalidindi S R 2015 *Hierarchical Materials Informatics: Novel Analytics for Materials Data* (New York: Elsevier)
[21] Nosengo N 2016 *Nature* **533** 22
[22] Jain A, Persson K and Ceder G 2016 *APL Mater.* **4** 053102
[23] Jain A, Ong S P, Hautier G, Chen W, Richards W D, Dacek S, Cholia S, Gunter D, Skinner D, Ceder G and Persson K 2013 *APL Mater.* **1** 1
[24] Thygesen K S and Jacobsen K W 2016 *Science* **354** 180
[25] Xue D Z, Xue D Q, Yuan R H and Zhou Y M 2017 *Science* **125** 532
[26] Xue D Z, Balachandran P V, Hogden J, Theiler J, Xue D Q and Lookman T 2016 *Nat. Commun.* **7** 11241
[27] Singh S, Bhadeshia H, MacKay D and Carey H M 1998 *Iron-mak Steelmak* **25** 355
[28] Agrawal A, Deshpande P D, Cecen A, Basavarsu G, Choudhary A and Kalidindi S 2014 *Integrating Mater. Manufacturing Innovation* **3** 8
[29] Jeong J H, Ryu S K, Park S J, Shin H C and Yu J H 2015 *Comput. Mater. Sci.* **100** 21
[30] Miller R J 2015 *Proceedings* 30*th British International Conference on Databases*, BICOD 2015 Edinburgh, UK, July 6–8, 2015