

Review

Information Fusion for Multi-Source Material Data: Progress and Challenges

Jingren Zhou ¹, Xin Hong ¹ and Peiquan Jin ^{1,2,*}

¹ School of Computer Science and Technology, University of Science and Technology of China, Hefei 230027, China

² Key Laboratory of Electromagnetic Space Information, Chinese Academy of Sciences, Hefei 230027, China

* Correspondence: jpq@ustc.edu.cn

Received: 16 July 2019; Accepted: 19 August 2019; Published: 22 August 2019



Abstract: The development of material science in the manufacturing industry has resulted in a huge amount of material data, which are often from different sources and vary in data format and semantics. The integration and fusion of material data can offer a unified framework for material data representation, processing, storage and mining, which can further help to accomplish many tasks, including material data disambiguation, material feature extraction, material-manufacturing parameters setting, and material knowledge extraction. On the other side, the rapid advance of information technologies like artificial intelligence and big data, brings new opportunities for material data fusion. To the best of our knowledge, the community is currently lacking a comprehensive review of the state-of-the-art techniques on material data fusion. This review first analyzes the special properties of material data and discusses the motivations of multi-source material data fusion. Then, we particularly focus on the recent achievements of multi-source material data fusion. This review has a few unique features compared to previous studies. First, we present a systematic categorization and comparison framework for material data fusion according to the processing flow of material data. Second, we discuss the applications and impact of recent hot technologies in material data fusion, including artificial intelligence algorithms and big data technologies. Finally, we present some open problems and future research directions for multi-source material data fusion.

Keywords: information fusion; multi-source material data; knowledge discovery

1. Introduction

Multi-source material data integration [1] and information fusion is one of the most important techniques that support the Materials Genome Initiative. It is also the foundation for achieving material computing with distinctive features, integration of experiment data and industrial data, and knowledge discovery from material data. The former U.S. president Barack Obama clearly pointed out that the general purpose of the Materials Genome Initiative was to double the rate of discovering, developing, making and using advanced materials. Multi-source data fusion can help us to discover new materials, as well as to advance the development of materials. The multi-source property of material data means that material data usually come from different application domains. Multi-source material data are often heterogeneous, meaning that material data from different sources vary in data format, data description, data quality and data processing methods. Materials are substances that people can use to make useful parts, devices and things. The development of materials plays a key role in the progress of human society. Material science involves the nature and applications of substances in every scientific and engineering field. Material science is an interdisciplinary science that studies the interrelation of material processing technologies, material microstructures and material macro performance.

Many kinds of theories are involved in material science, including solid-state physics, material chemistry, mechanical engineering, biology, electrical engineering and computer science. It is a widely accepted idea that integrating material science with other disciplines may lead to new findings or materials. For example, when combined with electronic engineering, material science can derive electronic material; when combined with mechanics, material science can derive structural material. Regarding the integration of material science with computer science, it has been a hot trend to discover new rules from heterogeneous multi-source material data, as multi-source material data contain rich information about material structure, material molecular formula, chemical reactions between materials, etc.

However, multi-source material data fusion is not a trial task. First of all, multi-source heterogeneous data usually have different formats, including structural data, semi-structural data and unstructured data. To realize multi-source heterogeneous data integration and fusion, a unified representation framework needs to be designed first [2]. Material data differ from other kinds of data in many aspects. First, material data are messy with complex correlations. Second, the semantic operations among material data sources are complex. Third, material data have high and inconsistent dimensions, meaning that there are not only one-dimension data, but also two-dimension data, three-dimension data, or even high-dimension data. Data fusion first needs to integrate the multi-source heterogeneous data into a unified data processing system. Another challenge lies in the accuracy of material ontology mapping. One kind of material can be represented in different languages, and its scientific name can change with time. In addition, acronyms and synonyms are common in material science. Note that it is already a big challenge in computer science to correctly analyze data with different encodings and formats. Thus, how to resolve conflicts that exist in multi-source material data is a big problem.

To the best of our knowledge, the community is currently lacking a comprehensive review of the state-of-the-art techniques on multi-source material data fusion. Although there are a few related reviews in this area, most of them are focused upon specific application domains. Moreover, new information technologies, like artificial intelligence algorithms [3] and big data technologies [4], are not reflected in previous reviews. Compared to previous works, this review presents a systematic analysis on both the algorithms and intrinsic features of material data fusion. Basically, the unique features of our study are two-fold. First, we present a systematic categorization and comparison framework for material data fusion according to the processing flow of material data. Second, we discuss the applications and impact of recent hot technologies in material data fusion, including artificial intelligence algorithms and big data technologies. In addition, we discuss the open problems in material data fusion, and present some future research directions for this area.

The remainder of this paper is organized as follows: Section 2 introduces the background of multi-source material data fusion. Section 3 discusses the current progress of multi-source material data fusion. Section 4 presents the future research directions of multi-source material data fusion, and finally in Section 5 we conclude the paper.

2. Background

In this section, we present the basic concepts in material data and multi-source data fusion, and particularly discuss the motivations of multi-source material data fusion.

Material data are produced in a wide range of industrial and scientific environments. They are difficult to be described by a unified model, due to the complex features of material data. Material data fusion is closely related with data gathering and integration. The most fundamental challenge of multi-source heterogeneous data integration and fusion is that there are big challenges in every small process of processing material data, and every process cannot be processed separately. Currently, Oracle, Facebook, Google and some other big IT companies already provide technologies for data format conversion and data integration. However, they cannot support the fusion of material data, as material data involve many kinds of data, which are hard to be processed within one framework. Table 1 shows a classification of material data according to material data structure and format.

Table 1. Classification of material data according to data format and structural degree.

Type	Data Structure	Data Format
T1	unstructured	Web pages, Word, PDF and other document data
T2	structured	Excel spreadsheets data
T3	Semi-structured	XML (Extensible Markup Language), JSON (JavaScript Object Notation), RDF (Resource Description Framework) and other formats
T4	unstructured	Key-value data, graph data and other NoSQL data
T5	unstructured	Images, videos and other file data
T6	structured	Data that can be retrieved from SQL databases with interfaces

Basically, the main characteristics of material data can be summarized as follows:

- **Multi-Source.** A great amount of material data usually comes from different science institutions, material factories and manufacturing enterprises that make all kinds of materials.
- **Heterogeneity.** The heterogeneity of material data means that different kinds of material data are usually represented in different structures, e.g., relational tables, semi-structural formats, and unstructured texts.
- **Unknowingness.** For the time being, material science still has a bottleneck in determining what features can affect material performance. Material data are corresponding to many attributes; many materials still have unknown features.
- **Irretrievability.** Many materials that are important for manufacturing, the aerospace industry and transport undertakings are usually kept confidential. Material data generated from manufacturing and scientific research need to be transformed into formatted data, which is very time consuming.
- **Complexity.** There is a great variety of hidden correlations among material data, and it is hard to recognize data correlations accurately.
- **Multi-Scale.** Material data can be described in many scales: Electronic scale, atom-molecular scale, micro-scale, mesoscale and macroscale.

Multi-source heterogeneous data fusion aims to combine material knowledge to reinforce the value of existing knowledge, and to extract the most valuable knowledge for material discovering, so it is necessary for people from material science and people from computer science to work together, which increases the difficulty of data fusion. The data conflicts in multi-source data are still a difficult problem to solve. The current mainstream method is to calculate the weight of data sources in terms of credibility, and then calculate the weighted average, but the details of the calculation method (e.g., the credibility of databases) are completely determined under specific situations, so it is very difficult to find a unified set of conflict resolution methods for material data fusion. Table 1 summarizes the major material data types and data formats.

2.1. Material Data Collection and Data Sharing

Materials can be roughly divided into composite materials and simple materials. There are many ways to describe materials, such as using a molecular formula [5,6]. For composite materials, simple representations like a molecular formula are not enough, and new ways need to be devised, e.g., using the combination ratio of various materials in a composite material. Further, as the conditions required for the various chemical reactions of materials are critical, how to describe the composition, ingredient, manufacturing process and properties of materials is very important. The highly specialized data requires manual and standardized definitions by experts in relevant fields, which is time consuming and costly.

2.2. Material Data Representation

The current common method is to define the format and specification of the data in advance using XSD (XML Schema Definition) [7]. Tabular forms can well describe the reaction and change of various steel materials under the corresponding conditions, and even record the experimental data

accumulated by large steel mills. However, the data of the reaction process is often described using functional graphs. This type of data is difficult to describe in a tabular form.

Standardization, storage and a semantic expression of material data are also important for describing the composition and process conditions of materials, and then an analysis of quantitative structure and performance can help optimize the material making process and improve performance. How to ensure that the description of data is always consistent and accurate in semantics is an important challenge that needs to be solved in the process of ontology alignment during the integrating process.

2.3. Material Data Integration and Storage

Material data from multiple data sources are likely inconsistent and have different data granularities and patterns. Therefore, material data integration is facing some difficulties. The following are some questions that need answering in material data integration:

- How to establish a unified data schema and data model for material data?
- Does heterogeneous data require multiple model representations?
- How to solve the conflicts in multi-source data?
- How to accurately and effectively find the correct data?
- How to semantically identify ontology in material data?

Material data storage is also a key issue in material data integration and fusion. Traditional, relational database systems cannot meet the requirements well when storing complex material data. They are especially inefficient for storing functional graphs and description documents. Relational database systems are used to store structural data in terms of the relational model, but material data include complex document data, recording reaction processes, performance test results, interaction between materials, etc. Material data also record how material properties vary with various conditions with functional graphs, as well as non-text data, such as material shape images, material production process images and molecular structure images. All of these data elements are difficult to be stored in relational databases.

2.4. Material Data Fusion

Data fusion can be generally accomplished in three layers, namely data layer, feature layer, and decision layer. One problem that often arises in the data layer fusion is the data conflicts in multi-source heterogeneous data. These conflicts are difficult to be accurately resolved by regular computer programming methods. At the same time, multi-source material data are often with multiple formats. Most material data do not have a uniform representation standard. Therefore, it is necessary to unify the data representation of material data before data fusion. In addition, when dealing with multi-modal material data, there will be more troubles, e.g., chart data, text data and image data are difficult to be integrated perfectly. Material data includes, not only static data such as text and XML, but also dynamic data such as functional graphs that describe how a variable is constantly changing in a certain process. Another problem comes from the fusion of dynamic data, such as the fusion of functional curves from multiple data sources. Since the function represents the relationship between the dependent variable and the independent variable over a time period, the curve contains the numerical pairs corresponding to the infinite number of time points. Therefore, it is very difficult to judge which function curve can most accurately describe the process. In addition, the data describing the 3D model of the molecular composition [8] of the material is also difficult to be fused, and the multimodal data contained in the multi-source heterogeneous material data is also difficult to integrate. In addition to the above problems, feature extraction of material data is also a difficult task. Meanwhile, when judging the results of the fusion at the decision-making level, certain expert support is needed due to the high professionalism of material science.

2.5. Discovering Material Knowledge

Material data fusion is not just the integration of multi-source heterogeneous data into the same database, but more importantly, the information contained in each data source is combined to

obtain knowledge that can be used for material production and R&D (Research and Development). Knowledge discovery often requires methods such as knowledge reasoning and machine learning. Due to the characteristics of material data, discovering material knowledge from multi-source material data faces the following problems:

- One important feature of material data is the small sample characteristic of material data. As we all know, the reason why machine learning has achieved some remarkable achievements in recent years is the development of big data in recent years, there is a large amount of data to improve the accuracy of machine learning. Therefore, applying small sample material data to machine learning to obtain high-quality material knowledge is a knotty issue.
- In order to get new material knowledge, machine learning and knowledge reasoning are needed to realize data mining between data, while traditional methods only use data mining methods. How to use the graph-based method to do the relationship reasoning between material entities to improve the efficiency and accuracy of reasoning is also a big challenge.
- Material data requires a lot of domain knowledge and rules for knowledge derivation. Combing the method of machine learning with the artificial power of experts is an open challenge.

3. Progress of Information Fusion for Multi-Source Material Data

In this section, we introduce the technical progress in multi-source material data fusion. In order to accelerate the research and development of materials, a number of methods for big data analysis are required. Based on this demand, scientists have built high-throughput computing platforms, such as AFLOW (Automatic Flow) [9] in the US, which is a high-throughput computational software for calculating properties. MatCloud [10] is the first high-throughput material integrated computing and data management platform developed by the Materials Genetics Laboratory of the Chinese Academy of Sciences. This platform integrates material data, and performs high-throughput simulation calculations to generate massive data. After that, it uses material informatics to find the relationship between the structural information contained for prediction. In China, the Institute of Materials and Genomics Engineering of Shanghai University is actively researching the fusion of material data with the cooperation of other schools of computer science and technology, in the process of researching the genome of materials for applications such as simulation computing and machine learning. Figure 1 shows the general process of multi-source heterogeneous data fusion.

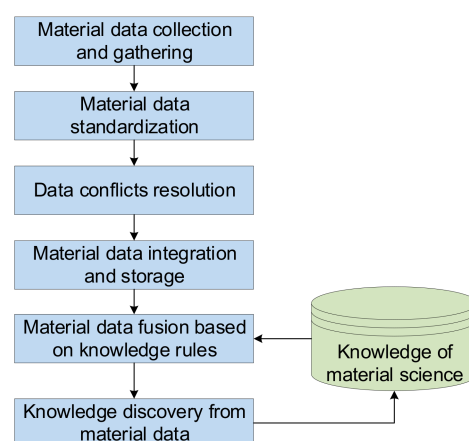


Figure 1. General process of multi-source material data fusion.

3.1. Data Collection and Processing of Material Data

According to the general process of material data fusions shown in Figure 1, we first need to define the various data items for materials. Take the aluminum matrix composite material as an example; when designing the material, it is necessary to design the particle shape, size, orientation and the distribution

statistics of the composite material. The material preparation needs to consider the diameter, height and weight of the material. It needs to be accurate to describe the interfacial tensile strength of aluminum matrix composites when describing the material characterization. The material processing stage needs to consider the maximum diameter, thickness and maximum rolling thickness of the forging. The LIMS (Laboratory Information Management System) [11,12] technology can be used for collecting material data automatically, analyzing data and material data management in many fields.

Next, we need to run simulation software to get simulation data. At present, there are some simulation frameworks [13] for material properties, which can ensure the normal operations of the workflow. Material data are produced in a variety of ways, such as mechanical properties—tensile tests measure elastic modulus, yield strength, tensile strength, elongation after fracture, breaking strength, maximum force non-proportional elongation, maximum force total elongation rate and the total elongation of the fracture. Considering the multi-source heterogeneity of material data, data fusion techniques based on majority voting strategies [14,15] are widely used, and source weight can be computed to make the fusion result closer to the data from more reliable sources [16].

In addition, machine-learning models including SVM (Support Vector Machines), nearest neighbors and other machine learning methods, can also be used to discover the relationship between material data. The application of thermal barrier coating materials in aerospace applications is very important. The factors affecting the performance of thermal barrier coatings are currently unclear, and the ways to optimize the process of improving performance still remains unrevealed. Thus, it is necessary to quantify the intrinsic relationship between all factors and performance, and to use machine learning to obtain the optimal solution [17], so that the bonding strength of the thermal barrier coating on the turbine guide vanes of domestic aviation engines can be increased. When analyzing the high-precision image data of ceramic complex microstructures, an image is segmented based on complex network and deep learning. The main steps are to establish the relationship between the surface image layer structure, feature extraction, texture analysis, judgment of microstructure and performance. With the ability to simultaneously describe multivariate microstructure factors and complex process parameters (such as the melting factor model), we can indirectly establish the quantitative relationship between process and microstructure, and provide scientific guidance for coating performance improvement. At the same time, first-principles as well as finite element calculations [18] are used to optimize the composition and structure of the material. Some examples are: Establishing multiple regression equations, performing a significance test, removing insignificant variables and quantifying the magnitude of each influencing factor. Figure 2 shows how material data are collected and stored into databases.

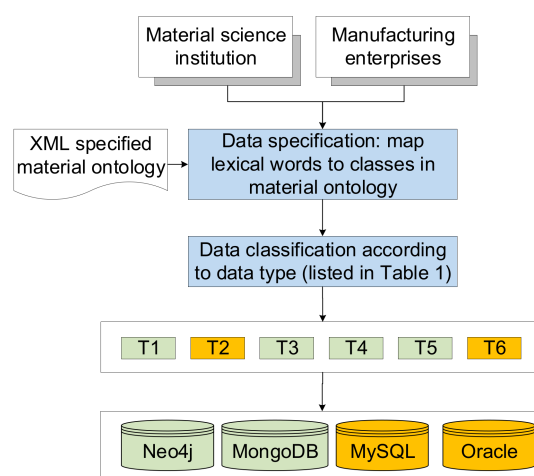


Figure 2. Material data collection and storage process.

3.2. Unified Representation of Material Data

The unified representation of material data is important to the integration of computational data, experimental data, test data, characterization data and service data in material science. Taking aluminum-based composite as an example, the whole process of processing, warehousing, database establishment and database storage is very important, and there are existing techniques of building an aluminum-based composite database platform that can be used continuously [19]. The machine-learning model of material data was applied to the qualitative and quantitative modeling of the full chain data of high-performance aluminum matrix composites. The service data of the aluminum matrix composite is added to the database, along with the simulation data. Descriptions of heterogeneous material data often use the predictive nonlinear multiscale model theory [2]. That is, the material data is normalized by constructing a multi-scale, multi-level data model. At present, the conceptual model is mainly used to describe the physical properties and the processing of material data [20]. Multi-source heterogeneous data is described with uniform tokens and views, all of which is described by a unified conceptual model.

In order to cope with the complex characteristics of material data, several existing technologies in NoSQL databases [21] can be adopted to build a database platform for material data. These technologies include document databases, key-value databases, column databases, and graph databases. The document-oriented databases represented by MongoDB and CouchDB are mainly designed to store, acquire and manage the document-based data. The minimum unit of data storage in document databases is the document. The document attributes stored in the same table can be different. The data can be stored in various formats such as JSON and XML. The key-value storage databases represented by Redis and BerkeleyDB provide efficient storage for key-value pairs. They can efficiently retrieve values for a given key. The column databases represented by HBase store data by column family. The column databases offer the convenience of storing structured and semi-structured data, facilitating data compression, and reducing I/O (Input and Output) costs for queries toward certain columns. The graph databases represented by Neo4J and FlockDB are convenient to store graph-based relations among material data. The graph data structure is also helpful to construct knowledge graphs over integrated material data. In addition to the above NoSQL database technologies, in real applications it is also possible to use a variety of databases to store material data, resulting in a hybrid database that can store different types of material data.

3.3. Integration of Big Material Data

The differences between data integration and data fusion are hard to describe. In fact, some researchers acknowledge that these two concepts are almost the same under most circumstances. Nevertheless, data integration is also regarded as the primary phase of data fusion by many researchers. In this paper, we recognize data integration as a procedure of integrating material data sources, data disambiguation, and data storage in a uniform platform. There are some existing integration techniques, and some typical techniques include multi-level data integration, full-vectorization data integration, and knowledge-graph-based data integration. The existing truth-finding techniques used in data integration are analyzed and discussed in the literature [22]. Due to the continuity of material data, certain data ranges and heterogeneity, it is preliminarily judged that CRH (Conflict Resolution on Heterogeneous data) [23] may be a suitable framework for material data. Other researchers proposed to perform material data integration based upon semantic recognition [24,25] and ontology technologies [26], such as lexical semantic similarity calculation, ontology alignment, semantics recognition with models like CRF (Conditional Random Field), and rule-based ontology matching. Data integration with semantic-index-enabled knowledge bases was also proposed in the literature [27]. Material ontology can be indexed semantically with the help of knowledge bases. Knowledge base generation consists of manual work [28] and algorithms like entity matching [29,30] and extracting semantic relations [31,32].

3.4. Algorithms for Material Data Fusion

Multi-source data fusion is a process of cognizing, synthesizing and judging a variety of data. The data involved in the fusion process often have properties such as multi-source, heterogeneity and incompleteness. Generally, data fusion methods can be categorized as probability-based methods, evidence reasoning methods and knowledge-based methods [33–35], but another point-of-view is to divide data fusion into data-level fusion, feature-level fusion and decision-level fusion [36,37], as shown in Figure 3. The data-level fusion focuses on data collection and preprocessing, through which all data sources are transformed, cleaned and integrated into a fused data store. The feature-level fusion mainly consists of two processes, namely feature extraction and feature fusion. The feature extraction process aims to extract features from original data items or sampled data items. The extracted features are inputted into the feature fusion process to form fused features. For example, physical features can be fused with chemical features to produce fused features for materials. The decision-level fusion is regarded as a multi-model classification, where different classifiers are examined on features as well as data items to get the optimal model for decision making, e.g., predicting the performance of a future material.

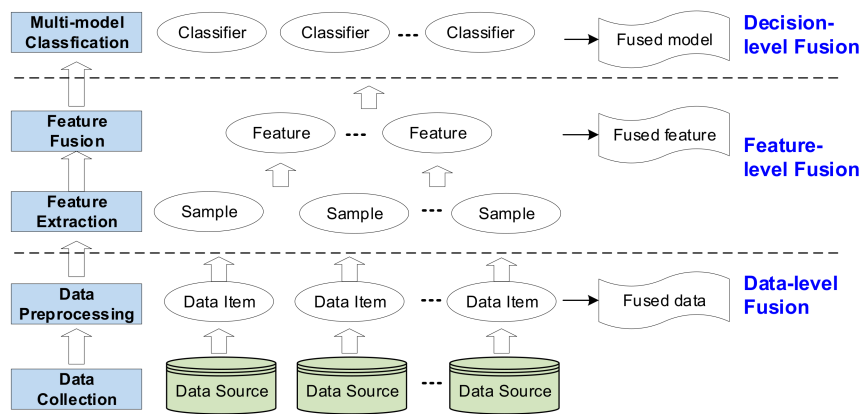


Figure 3. The fusion process, including three levels.

3.4.1. Data-Level Fusion

Data-level fusion is the lowest level of data fusion, which aims to directly process the acquired raw data and resolve possible conflicts. For resolving data conflicts in data fusion, there are many conflict resolution and data merging methods based on different ideas, which can be classified as follows:

1. Manual methods, e.g., using crowdsourcing to solve conflicts that are easy to be judged by humans [38]. The basic goal of crowdsourcing algorithms is to select a set of tasks to ask crowds, in order to improve the confidence of data under limited budgets, which is an optimization problem. The crowdsourcing fusion system architecture is shown in Figure 4.

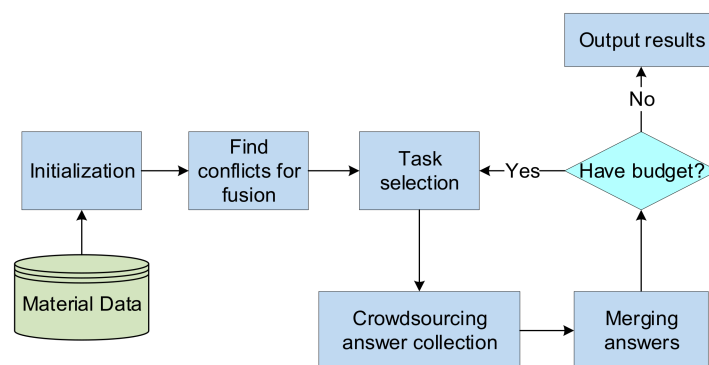


Figure 4. The working process of a crowdsourcing fusion system.

2. Truth finding methods. In the literature [39], the authors ranked multiple material data sources by considering the relationships between them to discover truth in material science databases. CRH [23] is an amenable framework for material data fusion, which resolves conflicts from different sources for each entry. It provides more accurate results compared with the voting-based approach. The proposed objective function in the CRH framework characterizes the overall difference between unknown truths and input data, while modeling source reliability as unknown source weights. The CRH framework can improve the performance of existing approaches due to its ability of tightly coupling various data types in the conflict resolution and source reliability estimation process. Running time on both single and Hadoop cluster machines demonstrates its efficiency, which has great potential for material data fusion due to the large amount of material data.

3. Mathematic statistic-based methods, e.g., MRF (Markov Random Field). In the literature [40], measurements from one or multiple sensors need to be combined in a consistent manner in order to build a grid map, which is a grid-mapping problem. GMRFs (Gaussian Markov Random Fields) can be used to learn a prior map by using the conditional independence property between spatial locations to obtain a representation of the state with a sparse information matrix. GMRFs are MRFs where all variables ξ_1, \dots, ξ_n are jointly Gaussians. The Markov property is depicted as: For $i \neq j$, ξ_i and ξ_j are independent, conditional on the other variables. The continuously indexed GMRF is used to build the prior map $p(\xi|X) \sim N(\eta, Q^{-1})$. Considering n observations y_i at locations $x_i, i = 1, \dots, n$, the model is shown in Equation (1).

$$y_i = \xi_i + \varepsilon_i = f(x_i) + \varepsilon_i, \xi \sim N(m(X), Q^{-1}(X, X)) \tag{1}$$

where m is the mean function, the function f models the underlying process and ε_i is the i^{th} noise term. The noise is considered independent, distributed as $N(0, \sigma_{\varepsilon,i}^2)$. Having obtained the spatially correlated prior map $p(\xi|X) \sim N(\eta, Q^{-1})$, we model the new independent measurements z as Gaussian distributed, which is shown in Equation (2).

$$p(z|\xi, x) = N(\mu_z, Q_z^{-1}) \tag{2}$$

3.4.2. Feature-Level Fusion

Feature-level fusion aims to perform data fusion based on feature information like categories, characteristics, primitives (edges or textures), or state vectors of the underlying output. Feature-level fusion comprehensively analyzes and processes feature information. Feature-level fusion algorithms can be mainly divided into three categories:

1. Algorithms based upon probability-based methods, e.g., the Bayesian decision theory. The Bayesian decision theory is a classic pattern recognition method. The known model space Ω includes C model classes, denoted as $\Omega = (\omega_1, \dots, \omega_C)$, and the unknown samples are composed of N -dimensional real numbers, denoted as $x = [x_1, x_2, \dots, x_N]^N$. According to the Bayesian decision theory with the minimum error rate, if the sample is within class j , then this class is the model class with the largest posterior probability under the condition of the known sample x . This decision process is expressed as $x \rightarrow \omega_j$, such that ω_j can be expressed by Equation (3).

$$\omega_j = \arg \max_{k=1, \dots, C} P(\omega_k|x) \tag{3}$$

here, $P(\omega_k|x)$ represents the posterior probability of the k^{th} class, $k \in [1, C]$. The classical method for combining continuous conditional measures is to use the Bayes Rule, which (roughly) involves multiplying the measures together, and then normalizing via an integral operation. In the literature [41], the authors combined a class of data fusion rules to get the combination of two (or more) probability distributions in the presence of unknown correlations with the Bayesian data fusion for improving fusion performance.

2. Algorithms based on the Sparse Representation Theory, which is mostly used for image fusion [42–44]. The Sparse Representation Theory was proposed for signal processing. Its basic idea is to use the prior knowledge of data sparseness to find a linear representation of the target signal from as few atoms as possible in an over-complete dictionary. The main steps in most sparse-representation-based image fusion methods include the following procedures:

- Segment the source images into some overlapping patches, and rewrite each of these patches as a vector.
- Perform sparse representation on the source image patches using pre-defined or learned dictionaries.
- Combine the sparse representations by some fusion rules.
- Reconstruct the fused images from their sparse representations.

3. Feature fusion algorithms based upon the Deep Learning Theory [45,46]. In this kind of algorithm, fusion features are obtained by fusing features obtained from multiple neural networks. From the feed-forward neural network such as the perceptron, multi-layer perceptron to other feedback neural networks such as RBM (Restricted Boltzmann Machine), DBM (Deep Boltzmann Machine), and CNN (Convolutional Neural Network), deep learning has been demonstrated to be highly efficient for feature learning [45]. Deep learning can be used for feature learning in image fusion and object detection [47]. In machine learning, multi-modal data collected by various sensors are often used for fusion, while dynamic fusion of multi-modal data is often performed by the vectorization deep learning method [48,49], and the model training method is used to achieve satisfactory results. When processing multi-modal data, it is necessary to use techniques such as image segmentation and the vector extraction of image sequences [50]. The defect of neural networks is its large computation complexity.

3.4.3. Decision-Level Fusion

Decision-level fusion aims to construct appropriate expert systems and knowledge bases, and to develop fusion rules suitable for decision-making. The fusion result is to provide decision basis and decision support for command and control decisions, to analyze the site situation, support and assist decision-making, the evaluation of decisions, evaluation of system performance, and so on.

The advantage is that different types of data can be fused, the calculation amount is small, the fault tolerance ability and the anti-interference ability are strong, but the disadvantages are also obvious, e.g., the loss of data information will bring the precision down. Decision-level fusion often uses the DS (Dempster-Shafer) evidence theory, the fuzzy theory, statistical decision-making, production rules and neural networks to make the final comprehensive decision.

1. DS Evidence Theory [51]. The DS Evidence Theory is a typical and widely applicable decision making-method. The DS Evidence Theory [52,53] is favored for its ability of dealing with uncertainty, integration of measurement information, and reasonable theoretical derivation. In the DS model, the frame of discernment (FoD) is denoted by Θ , which is a set of N hypotheses that are mutually exclusive and exhaustive, i.e., $\Theta = \{H_1, H_2, \dots, H_N\} = \{H_i | i = 1, 2, \dots, N\}$. The basic probability assignment (BPA) is a mass function $m: 2^\Theta \rightarrow [0, 1]$, which satisfies Equation (4).

$$\sum_{A \subseteq \Theta} m(A) = 1 \quad (4)$$

The belief function (Bel) is defined by Equation (5).

$$Bel(A) = \sum_{B \subseteq A} m(B) \quad (5)$$

and the plausibility function (Pl) is defined by Equation (6).

$$Pl(A) = \sum_{B \cap A \neq \emptyset} m(B) \quad (6)$$

Finally, the DS combination rule can be represented by Equation (7).

$$m(A) = [m_1 \oplus m_2](A) = \frac{1}{1-k} \sum_{A_i \cap B_j} m_1(A_i) \cdot m_2(B_j) \quad (7)$$

where \oplus represents the orthogonal sum operator, and k is the global conflict factor demonstrating the conflict degree between m_1 and m_2 .

2. Production rules [54,55]: Production rules are a common means of knowledge representation, which represents the causality in the form of “IF-THEN”. This form of rules reflects the behavioral characteristics of humans solving a class of problems, which can be solved by applying these rules cyclically. Material science can give us many scientific rules, which can be expressed in the form of “IF-THEN”. Production rules can increase the automation degree of material data fusion.

3. Fuzzy theory [56,57]: Fuzzy theory refers to the theory that uses the basic concept of fuzzy sets or continuous degree of membership functions. Fuzzy sets provide a form for dealing with less rigorous information. Let Z be the set of elements, and z denotes a class of elements of Z , i.e., $Z = \{z\}$. The fuzzy subset A in Z is represented by the membership function $\mu_A(z)$, which is between $[0,1]$. In the literature [58], the AHP (Analytic Hierarchy Process) approach [59] was introduced to determine the priorities of different monitoring models, after which a fuzzy decision fusion strategy was developed for process monitoring applications. The overall priority index PRT_k for each model has been obtained based on monitoring a specific fault in the process by a monitoring statistic, e.g., squared prediction error (SPE). Then, based on the determined posterior probabilities, a fuzzy discriminant matrix corresponding to SPE can be formulated for fusing the probabilities generated through different monitoring models. A fuzzy decision system could be considered to handle the material data-fusion process, which can satisfy the purpose of online process monitoring.

3.4.4. AI Algorithms for Material Data Fusion

AI algorithms have already been studied in material knowledge discovering, e.g., predicting new materials by using massive sample data from experiments. Stefano et al. [60] proposed a factor-based framework that incorporated a-priori knowledge as constraints to solve the pattern decomposition problem. They also proposed a new pattern decomposition algorithm that aimed at solving a sequence of (mixed integer) quadratic programs. Kiapour et al. [61] devised an attribute-based approach to recognition in X-ray scattering images, and demonstrated applications to image annotation and retrieval, which is helpful to the design of automatic analysis techniques for probing the physical structure of materials. Le Bras et al. [62] provided the first publicly available dataset for the phase-map identification problem, along with a parameterized synthetic data generator to assess the quality of proposed approaches, as well as tools for data visualization and solution evaluation. They fulfilled the goal of stimulating the development of new computational techniques for the analysis of material data. Jha et al. [63] introduced a novel deep regression network with individual residual learning that placed shortcut connections after each layer, so that each layer learned the residual mapping between its output and input. They showed that the proposed approach was able to provide significantly better prediction performance than state-of-the-art machine learning approaches.

Actually, AI algorithms are the main-stream techniques, not only in material knowledge discovery, but also in other areas. However, this review focuses on material data fusion, and we have to clarify that material data fusion is the foundation of material knowledge discovery, because it determines the quality of material data that are processed by material knowledge discovery. On the other hand,

many AI algorithms, such as classification models and deep learning models, can be considered in feature-level fusion and decision-level fusion.

3.4.5. Big Data Algorithms for Material Data Fusion

Material data is regarded as a kind of big data, because it has the key properties of big data, known as the 4V (volume, variety, velocity and veracity) property. Big data techniques are applicable in material science. For example, data organization and indexing techniques in big data can be used to organize material data. Guzun et al. [64] implemented the BSI (Bit-Sliced Indices) arithmetic over Hadoop MapReduce for the top- k (preference) query in big data analytics. Their experiments showed that the BSI approach was much faster than previous methods. Gowanlock et al. [65] proposed a heterogeneous sorting algorithm that utilized both multi-core CPUs and many-core GPUs. The sorting performance was better than the parallel reference implementation when sorting data that exceeds the memory capacity of GPUs. Gowanlock et al. also demonstrated that the heterogeneous sorting algorithm achieved high efficiency in large datasets. Liu et al. [66] presented the first deep learning solution toward a scientific image indexing problem, using a collection of over 300K microscopic images. The results showed that their proposal was 54% better than a dictionary lookup method, which is now popularly used in material data science. Qu et al. [67] proposed a high-throughput infrastructure for the automated calculation of molecular properties with a focus on battery electrolytes, aiming to ultimately address all chemical components present in the electrolyte, as well as the interactions between them. Their framework was used to perform complex computation on very large data sets and is proven to be efficient and scalable for computing complex properties.

4. Future Research Challenges

In recent years, due to the increasing applications of high-throughput experiments and material characterization methods, and the explosive growth of material data, the data fusion of multi-source heterogeneous materials based on material genomes has become an important direction of the future material genome. In this section, we present some future research directions toward multi-source material data fusion.

4.1. Optimizing Material Data Fusion Techniques

There are many sources of material data such as experiment, calculation, production and literature data. In addition, many data formats of material data exist, including structured, semi-structured and unstructured data, along with various combinations of columns. The current data fusion technologies not only lose a lot of useful information, but also cannot combine the domain knowledge to eliminate erroneous data when solving data conflicts. Therefore, a comprehensive consideration of data freshness, data sources and related expertise in data fusion will be a major research direction in the future.

Basically, the optimization of material data fusion technologies is supposed to focus on the following aspects:

1. Optimization based upon deep learning models. How to choose the appropriate machine-learning model and choose the appropriate parameters are essential for the fusion result. Deep learning primarily uses neural networks, which include supervised neural networks and no/semi-supervised learning. Supervised neural networks include deep neural networks, recurrent neural networks and convolutional neural networks. Unsupervised pre-training networks include Boltzmann machines, restricted Boltzmann machines, deep belief networks, generative adversarial networks and self-encoders. The choosing of a model not only affects the validity of the results, but also the cost of training.

2. Optimizing the algorithms used in data fusion. This includes algorithms such as weighted average, Kalman filter [68], Bayesian estimation, statistical decision, evidence theory, fuzzy theory, neural network and production rules. These techniques do not have a best algorithm that is obviously better than the others on solving conflicts between multiple data sources, so the combining results of

various algorithms can be considered in the future [69]. In addition, the accuracy of the data source corresponds to the weight of the data source in data fusion, which has a huge impact on the result of the fusion algorithm. In the future, it is necessary to optimize the estimation of the accuracy of the data source, and if necessary, the semi-manual method can be used.

3. Optimization from the perspective of knowledge discovery and knowledge use. In the whole process of data fusion, it is necessary to process and fuse the data using the knowledge of the material field and the generated rules. After the fusion results are obtained, knowledge extraction can be performed by means of machine learning and knowledge graph. In the future, we should consider adding the new knowledge we have learned to the original knowledge set, so as to get a positive feedback, and gradually improve the performance and effect of data fusion.

4.2. Processing Data with Unknown Types

In the process of data collection, it is very likely to encounter data of new formats and new types, which is an open-set recognition problem [70].

If the semantic meaning of the data is not well recognized, the automated data collection process cannot be completed. The future data processing direction will be mainly semantic ontology recognition technology based on ontology theory. At the same time, the machine learning method based on the deep learning framework is used to identify new types of data, and the results of learning are evaluated with certain manual judgment to give feedback. It is also possible to integrate various sorted material data to form a knowledge graph [71], which is convenient for knowledge reasoning and the discovery of new knowledge.

4.3. Protection of Data Property in Data Fusion

Today, block chain [72] is the main solution for multi-party sharing of files and data. We need to ensure that data ownership is always the data provider, so we also need to pay attention to the privacy protection of data fusion. By the data fusion of multi-source material data and knowledge discovery using machine learning, it is highly likely that the original data cannot be included in the results, but data protection for the data provider is still required. Therefore, in the future, it is necessary to record the flow changes of the data throughout the fusion process in detail.

When the final result is obtained, the data can be traced to obtain the source of the data. Using techniques such as block chain and Ethereum [73] to solve the problem of smart contract and technology sharing [74], data providers can confidently contribute their important data.

4.4. Verifying the Effects of Machine Learning

When we learn the best process parameters and other data items through machine learning, we cannot immediately judge the effectiveness of machine learning, because we can only admit that machine learning has effectively discovered useful knowledge of material data after applying the acquired knowledge to the production of materials and really improving the manufacture of materials. However, the important issue is that this process is time consuming and labor intensive, so it will make the whole process of data fusion and knowledge discovery most inefficient. After we use machine learning to obtain material knowledge, how to conduct experiments to judge the validity of knowledge and give effective feedback on machine learning is a research difficulty. Therefore, the main research direction in the future is how to combine the corresponding domain knowledge, so that the entire artificial intelligence work process can be feed backed by the generated knowledge to improve the structure of the neural network. Material data encounters many inconveniences in actual use. For example, when querying thermal barrier ceramic coating data, if you use relational database storage, it only works out on the situation where the relationship is relatively simple, but many materials have complicated structures, so the graph database storage is used, which has stronger semantic expression capabilities, and the knowledge graph provides a feasible way to organize material entities into entity networks. Many experiments in the materials industry are often very time consuming and expensive.

If only the results of real experiments were used to verify the effectiveness of machine learning, the cost of money and time will be greatly increased. Therefore, the design and implementation of simulation experiments with results close to real experiments will be a very important research direction. Moreover, how to improve the effectiveness and accuracy of simulation experiments is an important research subject. Existing simulations can simulate the process of material synthesis, but it is impossible to verify the working effect of the material, and there is a long way to approach the real process of manufacturing and using materials. Therefore, if the reference value of simulation experiments to data verification can be increased in the future, the whole process of data fusion and knowledge reasoning will be accelerated. The simulation experiment includes not only the simulation program running on the computer, but also directly verifying the material performance with a professional material experiment facility [75].

5. Conclusions

With the increasing applications of big data and artificial intelligence in other industries, computer technologies such as machine learning and big data processing play an important role in industries. Material industries need to obtain knowledge from a large amount of data, and the fusion of multi-source heterogeneous material data becomes an important prerequisite for further data mining and knowledge discovery. New computer technologies such as machine learning [76], data cubes [77], and knowledge graphs [78] can be used to help the development of new materials and improvements in existing material-making processes, has had a significant impact on the advancement of the materials industry. Some applications as phase mapper [79] help scientists filter worthless material structures in the research of new materials, which has greatly accelerated the research and development of new materials. This paper is elaborated on the significance and difficulties of the integration and fusion of multi-source heterogeneous materials data. Based on the characteristics of material data, we present a comprehensive review on the existing related technologies and the challenges in multi-source heterogeneous material data fusion. Finally, some future research directions of the fusion and integration of multi-source heterogeneous materials data are discussed, through which we expect to provide valuable references for advancing the researches in this field.

Author Contributions: J.Z., conceptualization, methodology, investigation, and writing—original draft preparation, X.H., methodology and investigation; P.J., writing—review and editing, supervision, project administration, and funding acquisition.

Funding: This research was funded by the National Key Research and Development Program of China (grant number: 2018YFB0704400 and 2018YFB0704404) and the National Science Foundation of China (grant number: 61672479). The APC was funded by the National Key Research and Development Program of China (grant number: 2018YFB0704400 and 2018YFB0704404).

Acknowledgments: We would like to thank the editors and anonymous reviewers for their suggestions and comments to improve the quality of the paper.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Stonebraker, M.; Ilyas, I. Data integration: The current status and the way forward. *IEEE Data Eng. Bull.* **2018**, *41*, 3–9.
2. Matouš, K.; Geers, M.; Kouznetsova, V.; Gillman, A. A review of predictive nonlinear theories for multiscale modeling of heterogeneous materials. *J. Comput. Phys.* **2017**, *330*, 192–220. [[CrossRef](#)]
3. Wu, Z.; Wan, S.; Jin, P.; Yue, L. Discriminative feature learning with constraints of category and temporal for action recognition. In Proceedings of the 8th International Conference on Images and Graphics (ICIG), Tianjin, China, 13–16 August 2015; pp. 173–184.

4. Lyu, M.; Jin, P.; Zhang, Z.; Wan, S.; Yue, L. STEM: A simulation-based testbed for electromagnetic big data management. In Proceedings of the 30th International Conference on Software Engineering and Knowledge Engineering (SEKE), Redwood City, CA, USA, 1–3 July 2018; pp. 229–230.
5. Seko, A.; Togo, A.; Tanaka, I. Descriptors for machine learning of materials data. In *Nanoinformatics*; Tanaka, I., Ed.; Springer: Singapore, 2018; pp. 3–23.
6. Cano, G.; Garcia-Rodriguez, J.; Garcia-Garcia, A.; Sanchez, H.; Benediktsson, J.; Thapa, A.; Barr, A. Automatic selection of molecular descriptors using random forest: Application to drug discovery. *Expert Syst. Appl.* **2017**, *72*, 151–159. [[CrossRef](#)]
7. Jounaidi, A.; Bahaj, M. Designing and implementing XML schema inside OWL ontology. In Proceedings of the 2017 International Conference on Wireless Networks and Mobile Communications (WINCOM), Rabat, Morocco, 1–4 November 2017; pp. 1–7.
8. Punjani, A.; Brubaker, M.; Fleet, D. Building proteins in a day: Efficient 3D molecular structure estimation with electron cryomicroscopy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 706–718. [[CrossRef](#)] [[PubMed](#)]
9. Calderon, C.; Plata, J.; Toher, C.; Oses, C.; Levy, O.; Fornari, M.; Natan, A.; Mehl, M.; Hart, G.; Nardelli, M.; et al. The AFLOW standard for high-throughput materials science calculations. *Comput. Mater. Sci.* **2015**, *108*, 233–238. [[CrossRef](#)]
10. Yang, X.; Wang, Z.; Zhao, X.; Song, J.; Zhang, M.; Liu, H. MatCloud: A high-throughput computational infrastructure for integrated management of materials simulation, data and resources. *Comput. Mater. Sci.* **2018**, *146*, 319–333. [[CrossRef](#)]
11. Grand, A.; Geda, E.; Mignone, A.; Bertotti, A.; Fiori, A. One tool to find them all: A case of data integration and querying in a distributed LIMS platform. *Database* **2019**, *2019*, baz004. [[CrossRef](#)] [[PubMed](#)]
12. Blazek, P.; Kuca, K.; Krejcar, O. Concept of a module for physical security of material secured by LIMS. In Proceedings of the International Conference on Bioinformatics and Biomedical Engineering, Granada, Spain, 8–10 May 2018; pp. 352–363.
13. Wang, J.; Gao, F.; Vazquez-Poletti, J.; Li, J. High performance computing for advanced modeling and simulation of materials. *Comput. Phys. Commun.* **2017**, *211*, 1. [[CrossRef](#)]
14. Li, X.; Duan, Z.; Hanebeck, U.D. Performance ranking of multiple nonlinear filters using ranking vector and voting fusion. In Proceedings of the 2017 20th International Conference on Information Fusion (Fusion), Xi'an, China, 10–13 July 2017; pp. 1–8.
15. Singh, M.; Baruah, R.; Nair, S. A voting-based sensor fusion approach for human presence detection. In *Intelligent Human Computer Interaction*; Springer: Cham, Switzerland, 2016; pp. 195–206.
16. Yang, Y.; Bai, Q.; Liu, Q. Dynamic source weight computation for truth inference over data streams. In Proceedings of the 18th International Conference on Autonomous Agents and Multi Agent Systems, Montreal QC, Canada, 13–17 May 2019; pp. 277–285.
17. Raccuglia, P.; Elbert, K.; Adler, P.; Falk, C.; Wenny, M.; Mollo, A.; Zeller, M.; Friedler, S.; Schrier, J.; Norquist, A. Machine-learning-assisted materials discovery using failed experiments. *Nature* **2016**, *533*, 73. [[CrossRef](#)] [[PubMed](#)]
18. Li, Y.; Liu, G.R. A novel node-based smoothed finite element method with linear strain fields for static, free and forced vibration analyses of solids. *Appl. Math. Comput.* **2019**, *352*, 30–58. [[CrossRef](#)]
19. See, L.; Perger, C.; Duerauer, M.; Fritz, S.; Bechtel, B.; Ching, J.; Alexander, P.; Mills, G.; Foley, M.; O'Connor, M.; et al. Developing a community-based worldwide urban morphology and materials database (WUDAPT) using remote sensing and crowdsourcing for improved urban climate modelling. In Proceedings of the 2015 Joint Urban Remote Sensing Event (JURSE), Lausanne, Switzerland, 30 March–1 April 2015; pp. 1–4.
20. Xia, L.; Raghavan, B.; Breitkopf, P. Towards surrogate modeling of material microstructures through the processing variables. *Appl. Math. Comput.* **2017**, *294*, 157–168. [[CrossRef](#)]
21. Bhogal, J.; Choksi, I. Handling big data using NoSQL. In Proceedings of the 2015 IEEE 29th International Conference on Advanced Information Networking and Applications Workshops, Gwangju, Korea, 24–27 March 2015; pp. 393–398.
22. Li, Y.; Gao, J.; Meng, C.; Li, Q.; Su, L.; Zhao, B.; Fan, W.; Han, J. A survey on truth discovery. *ACM SIGKDD Explor. Newsl.* **2016**, *17*, 1–16. [[CrossRef](#)]
23. Li, Q.; Li, Y.; Gao, J.; Zhao, B.; Fan, W.; Han, J. Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation. In Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data, Snowbird, UT, USA, 22–27 June 2014; pp. 1187–1198.

24. Wu, C.; Huang, C.; Chen, H. Expression recognition using semantic information and local texture features. *Multimed. Tools Appl.* **2018**, *77*, 11575–11588. [[CrossRef](#)]
25. Song, Y.; Cui, Y.; Han, H.; Shan, S.; Chen, X. Scene text detection via deep semantic feature fusion and attention-based refinement. In Proceedings of the 24th International Conference on Pattern Recognition (ICPR), Beijing, China, 20–24 August 2018; pp. 3747–3752.
26. Pai, F.; Yang, L.; Chung, Y. Multi-layer ontology based information fusion for situation awareness. *Appl. Intell.* **2017**, *46*, 285–307. [[CrossRef](#)]
27. Tekli, J.; Chbeir, R.; Traina, A.; Traina, C.; Yetongnon, K.; Ibanez, C.; Assad, M.; Kallas, C. Full-fledged semantic indexing and querying model designed for seamless integration in legacy RDBMS. *Data Knowl. Eng.* **2018**, *117*, 133–173. [[CrossRef](#)]
28. Jossé, G.; Schmid, K.; Züfle, A.; Skoumas, G.; Schubert, M.; Renz, M.; Pfoser, D.; Nascimento, M. Knowledge extraction from crowdsourced data for the enrichment of road networks. *Geoinformatica* **2017**, *21*, 763–795. [[CrossRef](#)]
29. Martins, P.; Marinho, Z.; Martins, A. Joint learning of named entity recognition and entity linking. In Proceedings of the 2011 International Conference on Computer Science and Network Technology, Harbin, China, 24–26 December 2019; p. 190.
30. Chen, Z.; Chen, Q.; Fan, F.; Wang, Y.; Wang, Z.; Nafa, Y.; Li, Z.; Liu, H. Enabling quality control for entity resolution: A human and machine cooperation framework. In Proceedings of the 2018 IEEE 34th International Conference on Data Engineering (ICDE), Paris, France, 16–19 April 2018; pp. 1156–1167.
31. Zhao, J.; Jin, P.; Liu, Y. Business relations in the web: Semantics and a case study. *J. Softw.* **2010**, *5*, 826–833. [[CrossRef](#)]
32. Jin, P.; Yang, J.; Zhao, J.; Liu, Y. A structural approach to extracting Chinese position relations from web pages. *J. Web Eng.* **2013**, *12*, 363–382.
33. Ding, W.; Jing, X.; Yan, Z.; Yang, L. A survey on data fusion in internet of things: Towards secure and privacy-preserving fusion. *Inf. Fusion* **2019**, *51*, 129–144. [[CrossRef](#)]
34. Wang, P.; Yang, L.T.; Li, J.; Chen, J.; Hu, S. Data fusion in cyber-physical-social systems: State-of-the-art and perspectives. *Inf. Fusion* **2019**, *51*, 42–57. [[CrossRef](#)]
35. Pires, I.; Garcia, N.; Pombo, N.; Francisco, F. From data acquisition to data fusion: A comprehensive review and a roadmap for the identification of activities of daily living using mobile devices. *Sensors* **2016**, *16*, 184. [[CrossRef](#)] [[PubMed](#)]
36. Taheri, S.; Toygar, Ö. Multi-stage age estimation using two level fusions of handcrafted and learned features on facial images. *IET Biom.* **2018**, *8*, 124–133. [[CrossRef](#)]
37. Ayed, S.B.; Trichili, H.; Alimi, A.M. Data fusion architectures: A survey and comparison. In Proceedings of the 15th International Conference on Intelligent Systems Design and Applications (ISDA), Marrakech, Morocco, 14–16 December 2015; pp. 277–282.
38. Chen, Y.; Chen, L.; Zhang, C. Crowdfusion: A crowdsourced approach on data fusion refinement. In Proceedings of the 2017 IEEE 33rd International Conference on Data Engineering (ICDE), San Diego, CA, USA, 19–22 April 2017; pp. 127–130.
39. Bélisle, E.; Huang, Z.; Gheribi, A. Truth discovery in material science databases. In Proceedings of the Australasian Database Conference, Melbourne, Australia, 4–7 June 2015; pp. 269–280.
40. Sun, L.; Vidal-Calleja, T.; Miro, J.V. Gaussian Markov random fields for fusion in information form. In Proceedings of the 2016 IEEE International Conference on Robotics and Automation (ICRA), Stockholm, Sweden, 16–21 May 2016; pp. 1840–1845.
41. Taylor, C.; Bishop, A. Homogeneous functionals and Bayesian data fusion with unknown correlation. *Inf. Fusion* **2019**, *45*, 179–189. [[CrossRef](#)]
42. Xia, Y.; Wan, S.; Jin, P.; Yue, L. Multi-scale local spatial binary patterns for content-based image retrieval. In Proceedings of the 9th International Conference on Active Media Technology (AMT), Maebashi, Japan, 29–31 October 2013; pp. 423–432.
43. Zhang, Q.; Liu, Y.; Blum, R.; Han, J.; Tao, D. Sparse representation based multi-sensor image fusion for multi-focus and multi-modality images: A review. *Inf. Fusion* **2018**, *40*, 57–75. [[CrossRef](#)]
44. Tian, Q.; Wan, S.A.; Jin, P.; Xu, J.; Zou, C.; Li, X. A novel feature fusion with self-adaptive weight method based on deep learning for image classification. In Proceedings of the 19th Pacific-Rim Conference on Multimedia (PC), Hefei, China, 21–22 September 2018; pp. 426–436.

45. Zhang, L.; Xie, Y.; Xidao, L.; Zhang, X. Multi-source heterogeneous data fusion. In Proceedings of the 2018 International Conference on Artificial Intelligence and Big Data, Chengdu, China, 26–28 May 2018; pp. 47–51.
46. Zhao, X.; Dai, X.; Jin, P.; Zhang, H.; Yang, C.; Li, B. OperaMiner: extracting character relations from opera scripts using deep neural networks. In Proceedings of the 24th International Conference on Database Systems for Advanced Applications (DASFAA), Part III. Chiang Mai, Thailand, 22–25 April 2019; pp. 542–546.
47. Yan, L.; Wan, S.; Jin, P.; Zou, C. Airplane fine-grained classification in remote sensing images via transferred CNN-based models. In Proceedings of the 5th International Conference on Geo-Spatial Knowledge and Intelligence (GSKI), Chiang Mai, Thailand, 8–10 December 2017; pp. 318–326.
48. Du, L.; Wang, Y.; Song, G.; Lu, Z.; Wang, J. Dynamic network embedding: An extended approach for skip-gram based network embedding. In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence Main Track (IJCAI-18), Stockholm, Sweden, 13–19 July 2018; pp. 2086–2092.
49. Wang, S.; Zhang, J.; Zong, C. Learning multimodal word representation via dynamic fusion methods. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; pp. 5973–5980.
50. Moreira, D.; Avila, S.; Perez, M.; Moraes, D.; Testoni, V.; Valle, E.; Goldenstein, S. Multimodal data fusion for sensitive scene localization. *Inf. Fusion* **2019**, *45*, 307–323. [[CrossRef](#)]
51. Ye, F.; Chen, J.; Li, Y.; Kang, J. Decision-making algorithm for multisensor fusion based on grey relation and DS evidence theory. *J. Sensors* **2016**, *2016*, 3954573:1–3954573:11. [[CrossRef](#)]
52. Deng, Y. Generalized evidence theory. *Appl. Intell.* **2015**, *43*, 530–543. [[CrossRef](#)]
53. Wu, G.; Li, L.; Li, L.; Wu, X. Web news extraction via tag path feature fusion using ds theory. *J. Comput. Sci. Technol.* **2016**, *31*, 661–672. [[CrossRef](#)]
54. Ali, M.; Ali, R.; Khan, W.A.; Han, S.; Bang, J.; Hur, T.; Kim, D.; Lee, S.; Kang, B. A data-driven knowledge acquisition system: An end-to-end knowledge engineering process for generating production rules. *IEEE Access* **2018**, *6*, 15587–15607. [[CrossRef](#)]
55. Obregon, J.; Kim, A.; Jung, J. RuleCOSI: Combination and simplification of production rules from boosted decision trees for imbalanced classification. *Expert Syst. Appl.* **2019**, *126*, 64–82. [[CrossRef](#)]
56. Alcantud, J.; Giarlotta, A. Necessary and possible hesitant fuzzy sets: A novel model for group decision making. *Inf. Fusion* **2019**, *46*, 63–76. [[CrossRef](#)]
57. Farhadinia, B.; Herrera-Viedma, E. Multiple criteria group decision making method based on extended hesitant fuzzy sets with unknown weight information. *Appl. Soft Comput.* **2019**, *78*, 310–323. [[CrossRef](#)]
58. Ge, Z.; Liu, Y. Analytic hierarchy process based fuzzy decision fusion system for model prioritization and process monitoring application. *IEEE Trans. Ind. Inform.* **2018**, *15*, 357–365. [[CrossRef](#)]
59. Ahmed, F.; Kilic, K. Fuzzy analytic hierarchy process: A performance analysis of various algorithms. *Fuzzy Sets Syst.* **2019**, *362*, 110–128. [[CrossRef](#)]
60. Ermon, S.; Le Bras, R.; Suram, S.; Gregoire, J.; Gomes, C.; Selman, B.; Van Dover, R. Pattern decomposition with complex combinatorial constraints: Application to materials discovery. In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, Austin, TX, USA, 25–30 January 2015; pp. 636–643.
61. Kiapour, M.; Yager, K.; Berg, A.; Berg, T. Materials discovery: Fine-grained classification of X-ray scattering images. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision, Steamboat Springs, CO, USA, 24–26 March 2014; pp. 933–940.
62. Le Bras, R.; Bernstein, R.; Gregoire, J.; Suram, S.; Gomes, C.; Selman, B.; Van Dover, R. Challenges in materials discovery—synthetic generator and real datasets. In Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, Québec City, QC, Canada, 27–31 July 2014; pp. 438–443.
63. Jha, D.; Ward, L.; Yang, Z.; Wolverton, C.; Foster, I.; Liao, W.; Choudhary, A.; Agrawal, A. IRNet: A general purpose deep residual regression framework for materials discovery. In Proceedings of the 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Anchorage, AK, USA, 4–8 August 2019; pp. 2385–2393.
64. Guzun, G.; Tosado, J.E.; Canahuate, G. Scalable preference queries for high-dimensional data using map-reduce. In Proceedings of the 2015 IEEE International Conference on Big Data (Big Data), Santa Clara, CA, USA, 29 October–1 November 2015; pp. 2243–2252.
65. Gowanlock, M.; Karsin, B. Sorting large datasets with heterogeneous CPU/GPU architectures. In Proceedings of the 2018 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW), Vancouver, BC, Canada, 21–25 May 2018; pp. 560–569.

66. Liu, R.; Agrawal, A.; Liao, W.; Choudhary, A.; De Graef, M. Materials discovery: Understanding polycrystals from large-scale electron patterns. In Proceedings of the 2016 IEEE International Conference on Big Data (Big Data), Washington, DC, USA, 5–8 December 2016; pp. 2261–2269.
67. Qu, X.; Jain, A.; Rajput, N.; Cheng, L.; Zhang, Y.; Ong, S.; Brafman, M.; Maginn, E.; Curtiss, L.; Persson, K. The Electrolyte Genome project: A big data approach in battery materials discovery. *Comput. Mater. Sci.* **2015**, *103*, 56–67. [[CrossRef](#)]
68. Wu, B.; Huang, T.; Jin, Y.; Pan, J.; Song, K. Fusion of high-dynamic and low-drift sensors using Kalman filters. *Sensors* **2019**, *19*, 186. [[CrossRef](#)] [[PubMed](#)]
69. Dong, W.; Kam, M. Integration of multiple adaptive algorithms for parallel decision fusion. In Proceedings of the 2016 Annual Conference on Information Science and Systems (CISS), Princeton, NJ, USA, 16–18 March 2016; pp. 355–359.
70. Neira, M.; Júnior, P.; Rocha, A.; Torres, R. Data-fusion techniques for open-set recognition problems. *IEEE Access* **2018**, *6*, 21242–21265. [[CrossRef](#)]
71. Zhang, X.; Liu, X.; Li, X.; Pan, D. MMKG: An approach to generate metallic materials knowledge graph based on DBpedia and Wikipedia. *Comput. Phys. Commun.* **2017**, *211*, 98–112. [[CrossRef](#)]
72. Ferreira, M.; Rodrigues, S.; Reis, C.; Maximiano, M. Blockchain: A tale of two applications. *Appl. Sci.* **2018**, *8*, 1506. [[CrossRef](#)]
73. Grishchenko, I.; Maffei, M.; Schneidewind, C. Foundations and tools for the static analysis of ethereum smart contracts. In Proceedings of the International Conference on Computer Aided Verification, Oxford, UK, 14–17 July 2018; pp. 51–78.
74. Alsayed Kassem, J.; Sayeed, S.; Marco-Gisbert, H.; Pervez, Z.; Dahal, K. DNS-IdM: A blockchain identity management system to secure personal data sharing in a network. *Appl. Sci.* **2019**, *9*, 2953. [[CrossRef](#)]
75. Gabardi, M.; Chiaradia, D.; Leonardis, D.; Solazzi, M.; Frisoli, A. A high performance thermal control for simulation of different materials in a fingertip haptic device. In Proceedings of the International Conference on Human Haptic Sensing and Touch Enabled Computer Applications, Pisa, Italy, 13–16 June 2018; pp. 313–325.
76. Ahneman, D.; Estrada, J.; Lin, S.; Dreher, S.; Doyle, A. Predicting reaction performance in C–N cross-coupling using machine learning. *Science* **2018**, *360*, 186–190. [[CrossRef](#)] [[PubMed](#)]
77. Xie, X.; Hao, X.; Pedersen, T.; Jin, P.; Chen, J. OLAP over probabilistic data cubes I: Aggregating, materializing, and querying. In Proceedings of the 32nd IEEE International Conference on Data Engineering (ICDE), Helsinki, Finland, 16–20 May 2016; pp. 799–810.
78. Zhao, M.; Wang, H.; Guo, J.; Liu, D.; Xie, C.; Liu, Q.; Cheng, Z. Construction of an industrial knowledge graph for unstructured Chinese text learning. *Appl. Sci.* **2019**, *9*, 2720.
79. Bai, J.; Xue, Y.; Bjorck, J.; Le Bras, R.; Rappazzo, B.; Bernstein, R.; Suram, S.; Van Dover, R.; Gregoire, J.M.; Gomes, C.P. Phase mapper: Accelerating materials discovery with AI. *AI Mag.* **2018**, *39*, 15–26. [[CrossRef](#)]

