A. Prakash, S. Sandfeld

Chances and Challenges in Fusing Data Science with Materials Science

Chancen und Herausforderungen bei der Verschmelzung von Datenwissenschaft und Werkstoffwissenschaft

The working group "3D Data Science" is headed by Prof. Dr. Stefan Sandfeld.

Received: May 25, 2018
Accepted: May 29, 2018

Eingegangen: 25. Mai 2018
Angenommen: 29. Mai 2018
Übersetzung: V. Müller

Abstract

Data science and informatics have emerged as the fourth paradigm of scientific research over the past decade. Although the impact of this new paradigm is very apparent in many scientific fields and has seen many success stories, the field of materials informatics – data science and informatics for materials science and engineering – is still in its infancy. Based on the availability of data, the field of materials science would be ideal for data analytics and informatics, particularly if such data is shared with the larger materials science community. In this work, we discuss the advantages of digitalization and data science, current challenges for experiments and simulations involving data manage-

Kurzfassung

Datenwissenschaft und -informatik sind im letzten Jahrzehnt als Viertes Paradigma der wissenschaftlichen Forschung in Erscheinung getreten. Obwohl die Auswirkungen dieses neuen Paradigmas in vielen wissenschaftlichen Gebieten deutlich zum Ausdruck kommen und bereits in vielen Fällen erfolgreich angewandt wurden, steckt die Werkstoffinformatik – Datenwissenschaft und -informatik im Bereich der Materialwissenschaft und Werkstofftechnik – noch immer in den Kinderschuhen. Auf Basis der Verfügbarkeit von Daten würde sich das Gebiet der Werkstoffwissenschaften ideal für Datenanalyse und -informatik eignen, vor allem wenn solche Daten von einer größeren Gemeinschaft von Werkstoffwissenschaftlern genutzt wird. Diese

Authors:

Aruna Prakash, Stefan Sandfeld Micromechanical Materials Modelling (MiMM), Institute of Mechanics and Fluid Dynamics, Technische Universität Bergakademie Freiberg (TUBAF), Lampadiusstr. 4, 09599 Freiberg, Germany;
e-mail: arun.prakash@imfd.tu-freiberg.de

ment, acquisition and sharing, and look at possible solutions.

Arbeit diskutiert die Vorteile der Digitalisierung und Datenwissenschaft, aktuelle Herausforderungen bei experimentellen Untersuchungen und Simulationen, die Datenmanagement, -erhebung und -nutzung beinhalten und stellt mögliche Lösungsansätze vor.

1. Introduction

Scientific exploration in the field of materials science and engineering (MSE) has traditionally evolved around the three main paradigms of experiments/empirical reasoning, theory/modeling and computation/simulation (cf. Fig. 1). In the past decade, data science and informatics (DSI) has evolved as the fourth paradigm [1], and has shown great potential for significantly accelerating materials development [2, 3]. DSI distinguishes itself from the computational paradigm, in that the latter involves solution methodologies to well formulated problems, whilst the former deals with pattern recognition and finding links between different sets of data [4]. The quality of the solution in the computational paradigm is strongly dependent on the quality, sophistication and predictive capability of the underlying model. By contrast, DSI is model-free; the quality of the solution is determined solely by the underlying data. This characteristic feature makes it particularly attractive for increased synergy between the three principal paradigms, particularly between experiments and computations, or even between methods of a particular paradigm. Some examples of such studies, which result in high data requirements, include:

- High throughput experiments and simulations, including tomography investigations [e. g. 5–10].

1. Einleitung

Wissenschaftliche Untersuchungen auf dem Gebiet der Materialwissenschaft und Werkstofftechnik (MSE, Materials Science and Engineering) haben sich traditionell um die drei Hauptparadigmen – Experimente/empirische Begründung, Theorie/Modellierung und Berechnung/Simulation (vgl. Bild 1) – herausgebildet. Im letzten Jahrzehnt sind Datenwissenschaft und -informatik (DSI, Data Science and Informatics) als Viertes Paradigma [1] in Erscheinung getreten und haben großes Potenzial für eine deutliche Beschleunigung der Entwicklung neuer Werkstoffe gezeigt [2, 3]. DSI unterscheidet sich vom rechnergestützten Paradigma insofern, dass letzteres Lösungsmethodiken zu gut formulierten Problemen beinhaltet, während sich DSI mit Mustererkennung und der Erkennung von Verknüpfungen zwischen verschiedenen Datensätzen beschäftigt [4]. Die Qualität des Lösungskonzepts, das auf dem rechnergestützten Paradigma beruht, hängt stark von der Qualität, Differenziertheit und Vorhersagbarkeit des zugrundeliegenden Modells ab. Im Unterschied dazu arbeiten Datenwissenschaft und -informatik modellfrei; die Qualität des Lösungsansatzes hängt allein von den zugrundeliegenden Daten ab. Diese charakteristische Eigenschaft ist besonders interessant für das Erreichen eines besseren Zusammenwirkens der drei Hauptparadigmen, v. a. von experimentellen Untersuchungen und Berechnungen oder sogar von Verfahren innerhalb eines bestimmten Paradigmas. Bespiele von Untersuchungen, die zu hohen Datenanforderungen führen, sind u. a.:

- Hoch-Durchsatz-Untersuchungen und -Simulationen, darunter tomografische Untersuchungen [z. B. 5–10].

- Experimentally informed large-scale atomistic simulations [e. g. 11–13].
- High throughput crystal plasticity simulations, including integrated computational materials engineering studies [e. g. 14, 15].
- Combined simulation strategies [16–20].
- Multiscale methods, including concurrent frameworks [21–23, 49].

DSI shows excellent potential for obtaining further and new insightful information from such studies [24–26]. The attractive feature of DSI is that it is independent of the technique used to acquire the data, and can hence be used by experimentalists, modelers and simulation scientists alike. In this regard, DSI acts more like a reservoir that is fed by the three other paradigms (see, Fig. 1). This opens up a multitude of opportunities to gain deeper and better insights by combining data from various sources, particularly experiments and simulations, thus improving our knowledge on material behavior. With the application of currently available DSI toolsets having already shown great success in many fields, and many being developed, the times ahead are indeed promising for materials scientists.

The availability of data is hence at the very heart of DSI. Buzzwords like Digitalization, Industry 4.0, Digital Twin, Big Data, etc. make their appearance in this context. The pervasive nature of such buzzwords notwithstanding, it is important to understand their impact and relevance from the point of view of an individual researcher. That is to say, to answer the simple question that materials scientists are perhaps

- Experimentell gestützte, großskalige atomistische Simulationen [z. B. 11–13].
- Hoch-Durchsatz-Simulationen zur Kristallplastizität, darunter auch computergestützte Studien auf dem Gebiet der Werkstoffentwicklung [z. B. 14, 15].
- Kombinierte Simulationsstrategien [16–20].
- Multiskalen-Methoden, darunter nebenläufige Frameworks [21–23, 49].

DSI zeigt außerordentlich großes Potenzial hinsichtlich der Gewinnung neuer und aufschlussreicher Informationen aus solchen Studien [24–26]. Eine interessante Eigenschaft ist, dass DSI unabhängig von dem Verfahren ist, mit dem die Daten erhoben wurden, und somit gleichermaßen von Experimentatoren, Modellierern und Simulationswissenschaftlern genutzt werden kann. In dieser Hinsicht verhält sich DSI eher wie ein Speicher, der durch die anderen drei Paradigmen gespeist wird (s. Bild 1). Somit eröffnen sich vielerlei neue Möglichkeiten, um tiefere und bessere Erkenntnisse zu erlangen, indem Daten aus diversen Quellen, besonders aus experimentellen Untersuchungen und Simulationen kombiniert werden, was somit dazu beiträgt, unser Verständnis für das Verhalten von Werkstoffen zu verbessern. Da die Anwendung von derzeitig verfügbaren DSI-Software-Werkzeugen bereits große Erfolge in vielen Bereichen verbuchen konnte und derzeit etliche neue Software-Werkzeuge entwickelt werden, sehen Werkstoffwissenschaftler einer vielversprechenden Zukunft entgegen.

Die Verfügbarkeit von Daten bildet somit das Herzstück von DSI. Schlagwörter wie Digitalisierung, Industrie 4.0, Digitaler Zwilling, Big Data, usw. treten in diesem Zusammenhang auf. Ungeachtet der Allgegenwärtigkeit solcher Schlagwörter, ist es wichtig, deren Bedeutung und Relevanz aus der Sicht eines einzelnen Forschers zu verstehen. D. h. die einfache Frage zu beantworten, die möglicherweise (und zu Recht) von Werkstoff-

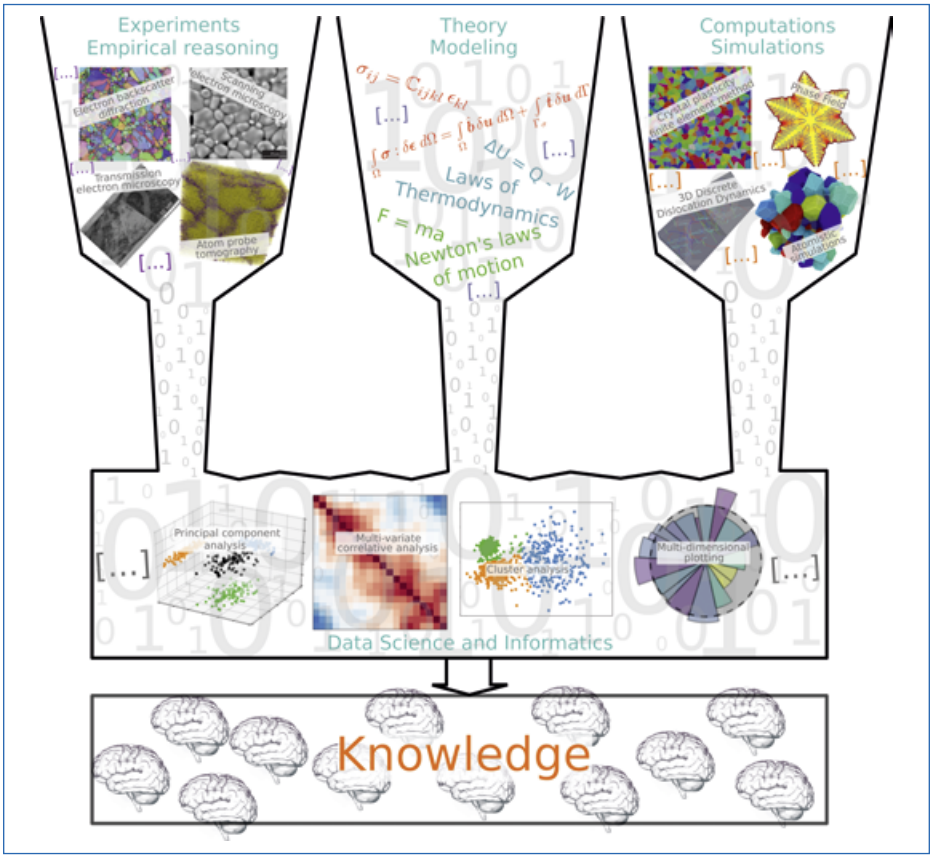

Fig 1: *A schematic representation of the four paradigms of scientific exploration. Traditionally science has evolved around the three paradigms of Experiments, Theory and Computations. With the advent of Data Science and Informatics, a new field has opened up, which allows, for instance, data from electron back-scatter diffraction and crystal plasticity finite element simulations, or from transmission electron microscopy and 3D dislocation dynamics simulations to be integrated and used for further analysis like, e. g. principal component analysis or cluster analysis, leading to improved knowledge on material behavior. Micoscopy Images (top left) courtesy of M. Motylenko und C. Wüstefeld ((HR)TEM), Stefan Martin (EBSD), TU Freiberg. Scanning electron microscopy image from Ref. [50]. Atom probe data courtesy of P. Felfer (FAU) and D. Gianola (UCSB). DDD image courtesy of D. Weygand (KIT).*

Bild 1: *Schematische Darstellung der Vier Paradigmen wissenschaftlicher Untersuchungen. Traditionell hat sich die Wissenschaft um die drei Paradigmen – Experimente, Theorie und Berechnungen – entwickelt. Der Einzug der Datenwissenschaft und -informatik eröffnet ein neues Gebiet, sodass z. B. Daten aus der Elektronenrückstreubeugung, aus kristallplastischen Finite-Elemente-Simulationen, aus Transmissionselektronenmikroskopie und 3D-Versetzungsdynamik-Simulationen für weitere Analysen integriert und genutzt werden können, z. B. für Hauptkomponenten- oder Clusteranalysen, was zu einem besseren Verständnis hinsichtlich des Werkstoffverhaltens führt. Mikroskopische Aufnahmen (oben links) mit freundlicher Genehmigung von M. Motylenko und C. Wüstefeld ((HR)TEM), Stefan Martin (EBSD), TU Freiberg. Bild zur diskreten Versetzungsdynamik mit freundlicher Genehmigung von D. Weygand (KIT).*

(and rightly so) asking: "What is digitalization all about? What is in it for me? Where are the current challenges? and where do I find information/tools/training for improved data handling in my own research and that of my group?" Although, perhaps seemingly selfish at the surface, these questions denote the central challenge in motivating individual researchers and small research groups towards the process of digitalization, particularly since researchers have only limited time and resources at their disposal.

The aim of the current paper is to discuss the impact and advantages of digitalization and DSI, current challenges involved in data acquisition, as well as management and sharing, in the context of characterization methods and simulations in MSE. We propose possible solution strategies and discuss some of them based on a few examples. This paper is essentially directed towards the individual researcher and small research groups – both experiments and simulation oriented, keeping the questions mentioned previously in focus. Our goal is that by reading this paper, researchers not only find answers to their questions, but are also motivated to share their data/methodology/software etc., in order to increase our knowledge about materials. We note that much of what is in this article is a result of discussions in the DGM Arbeitskreis on 3D Data Science, and as a result, is very much tuned to the German/European context. However, the discussion and the message in the current article is, in general, relevant to the global materials science community. For more details on some of the points discussed below, the reader is referred to the Strategy Paper of the DGM [27].

wissenschaftlern gestellt wird: „Um was geht es bei der Digitalisierung? Welcher Nutzen ergibt sich daraus für mich? Worin liegen die aktuellen Herausforderungen? Und wo gibt es Informationen/Tools/Fortbildungen zur besseren Datenverarbeitung für meine eigenen Forschungsprojekte und die meiner Arbeitsgruppe?" Obwohl diese Fragen oberflächlich betrachtet vielleicht egoistisch klingen mögen, symbolisieren sie eine der wesentlichen Herausforderungen, nämlich die Motivierung einzelner Forscher und kleiner Forschungsgruppen, den Schritt hin zur Digitalisierung zu wagen, besonders vor dem Hintergrund, dass die Zeit und die Ressourcen, die ihnen zur Verfügung stehen, begrenzt sind.

Ziel der vorliegenden Arbeit ist es, die Bedeutung und Vorteile der Digitalisierung und von DSI, aktuelle Herausforderungen bei der Datenerhebung, sowie deren Management und Nutzung im Kontext von Charakterisierungsmethoden und Simulationen auf dem Gebiet der Materialwissenschaft und Werkstofftechnik zu diskutieren. Es werden mögliche Lösungsansätze vorgestellt, von denen einige anhand von Beispielen diskutiert werden. Diese Arbeit richtet sich speziell an einzelne Forscher und kleine Forschungsgruppen – sowohl versuchs- als auch simulationsorientiert – und konzentriert sich dabei auf die bereits erwähnten Fragen. Ziel ist, dass das Lesen dieser Arbeit nicht nur Antworten auf die Fragen der Forscher liefert, sondern die Forscher auch motiviert, ihre Daten/Methodik/Software, usw. zu teilen, um unser Wissen über Werkstoffe zu vertiefen. Es wird angemerkt, dass ein großer Teil der vorliegenden Arbeit auf den Ergebnissen des DGM-Arbeitskreises „3D Data Science" beruht und somit deutlich auf den deutschen/europäischen Kontext abgestimmt ist. Die Diskussion und Botschaft der vorliegenden Arbeit ist allgemein dennoch von Bedeutung für die internationale Gemeinschaft der Werkstoffwissenschaftler. Für weitere Details zu einigen in dieser Arbeit diskutierten Punkte wird auf das DGM-Strategiepapier [27] verwiesen.

2. Why Digitalization and Digital Transformation?

Digitization, i. e., the conversion of analog/physical information into a digital representation, is where it all began. Digitalization is perhaps the logical next step, where businesses and processes are advanced into the digital era by leveraging digitization and digital technologies for handling and analyzing the digital form of data, and turning data into knowledge. Digital transformation is the corresponding process of changing a specific field or community. While the academic research sector of MSE is, in some respect, still at the beginning of this transformation, the commercial sector has indeed progressed further: Industry 4.0 has embraced digitalization through increased automation and data exchange, particularly for smart manufacturing. Altogether, digitalization has already led to ground-breaking innovation in many fields of engineering, notably in e-mobility, telecommunication and the energy sectors, and presents a unique opportunity for advanced materials development.

The main question that digitalization in materials science shall help to answer is the following: What is required – be it experimentally or computationally – to fully characterize and understand the behavior of Material X? One might be tempted to answer that we require an *as-complete-as-possible* listing of data on the desired material. This includes (i) data, ranging from electronic properties, through atomic positions to descriptions of microstructural features on different length scales, (ii) stress strain curves, effective material properties, etc., and (iii) a detailed description on how the said data

2. Warum Digitalisierung und digitaler Wandel?

Es begann alles mit der digitalen Umwandlung, d. h. der Umwandlung analoger/physikalischer Informationen in eine digitale Darstellung (engl. “digitization”). Digitalisierung (engl. “Digitalization”) ist wohl der nächste logische Schritt, durch den Unternehmen und Prozesse in das digitale Zeitalter geführt werden. Dies geschieht durch die Zunutzemachung digitaler Umwandlung und digitaler Technologien bei der Verarbeitung und Analyse von Daten, die in digitaler Form vorliegen, und durch das Umwandeln von Daten in Wissen. Der Digitale Wandel beschreibt den entsprechenden Umwandlungsprozess innerhalb eines spezifischen Gebiets oder einer Gemeinschaft. Während der akademische Forschungssektor MSE in mancherlei Hinsicht immer noch am Anfang dieses Wandels steht, ist der Handelssektor in der Tat schon weiter vorangekommen: Industrie 4.0 nimmt die Digitalisierung bereitwillig an und macht sie sich mittels stärkerer Automatisierung und Datenaustausch zu Nutze, besonders im Bereich intelligenter Fertigungsverfahren. Insgesamt hat die Digitalisierung bereits bahnbrechende Innovationen in vielen ingenieurwissenschaftlichen Bereichen hervorgebracht, vor allem in der E-Mobilität, Telekommunikation und im Energiebereich und stellt auch eine einzigartige Chance bei der Entwicklung moderner Werkstoffe dar.

Die Kernfrage, die mit Hilfe der Digitalisierung im Bereich der Werkstoffwissenschaften beantwortet werden soll, ist die folgende: Was ist nötig – sei es experimentell oder rechnergestützt – um das Verhalten des Werkstoffs X vollständig zu charakterisieren und zu verstehen? Man könnte sich zu der Antwort verleiten lassen, dass eine Auflistung von Daten zum gewünschten Werkstoff, die so vollständig wie möglich ist, erforderlich ist. Darunter fallen (i) Daten zu elektronischen Eigenschaften, zu Atompositionen bis hin zu Beschreibungen von Gefügeeigenschaften auf verschiedenen Längenskalen, (ii) Spannungs-Dehnungs-Kurven,

was obtained. While this might seem the utopian goal of digital transformation from the point of view of a computer scientist, utilizing our existing knowledge and experience as materials scientists will help us intelligently choose only the data which is really needed, thus reducing redundancies. Nonetheless, the materials scientists of the future would be unburdened from mundane tasks of cataloging and managing data, since these would be automated, allowing one to fully delve into developing and implementing improved methods for data acquisition and analysis, together with developing advanced algorithms for handling and analysis, resulting in better interpretation of material behavior and advanced theories.

3. What is in it for me as a Materials Researcher?

A key factor that drives our everyday research is the availability of data, through which we derive important insights into material behavior. Generating such data, either from experiments or simulations, requires the knowledge of methodologies, equipment and tools. For instance, generating EBSD data requires the user to not only have an understanding of diffraction, but also know how the specimens under consideration were prepared, how the specimen is oriented during measurement, the equipment used, the lighting conditions and indexing rate used, etc., among other important information about the actual procedure itself. In the world of atomistic simulations, the scientist needs to know exact details on the interatomic potential used, the code used for simulations, the details on the numerical toolbox (e. g. coupling constants, time increment, etc.), steps for generating the at-

effektive Materialeigenschaften, etc. und (iii) eine detaillierte Beschreibung, wie die jeweiligen Daten gesammelt wurden. Auch wenn dies für einen Informatiker das utopische Ziel des digitalen Wandels zu sein scheint, hilft uns Werkstoffwissenschaftlern das bereits vorhandene Wissen und unsere Erfahrung bei der intelligenten Auswahl von genau den Daten, die wirklich ausschlaggebend sind, wodurch Redundanzen verringert werden. Trotz alledem würde der Werkstoffwissenschaftler der Zukunft von der stumpfsinnigen Aufgabe befreit sein, Daten zu katalogisieren und zu verwalten – dies geschieht automatisiert – und könnte sich damit vollständig der Entwicklung und dem Einsatz verbesserter Verfahren für die Datenerhebung und -analyse widmen sowie der Entwicklung erweiterter Algorithmen für Verarbeitung und Analyse, woraus sich eine bessere Interpretation des Werkstoffverhaltens und weiterentwickelte Theorien ergeben.

3. Welcher Nutzen ergibt sich daraus für mich als Werkstoffwissenschaftler?

Ein Schlüsselfaktor, der die tagtägliche Forschung vorantreibt, ist die Verfügbarkeit von Daten, aus denen sich wichtige Einblicke in das Werkstoffverhalten ableiten lassen. Das Generieren von Daten, entweder auf Grundlage von Experimenten oder Simulationen, erfordert die Kenntnis von Methoden, Gerätschaften und Tools. Z.B. ist es für die Erhebung von EBSD-Daten erforderlich, dass der Nutzer nicht nur das Phänomen der Beugung versteht, sondern neben weiteren wichtigen Informationen bzgl. der eigentlichen Durchführung auch weiß, wie die entsprechenden Proben vorbereitet wurden, wie die Probe während der Messung orientiert sein muss, welche Ausrüstung verwendet wird, welche Lichtverhältnisse und Indizierungsrate erforderlich sind, etc. In der Welt atomistischer Simulationen benötigt der Wissenschaftler genaue Angaben zum verwendeten interatomaren Potenzial, zum verwendeten Code

omistic structure etc. Such key information or meta data is crucial for reproducing the results.

Within a digitalized work flow, such meta data, together with information about the data processing, would be part of the publication itself, making reproduction of scientific results much easier or possible in the first place. Digitalization can hence result in research data becoming more accessible and usable [28, 29].

Furthermore, the knowledge and availability of various tool sets would ease the learning curve of researchers. As mentioned earlier, DSI also opens up further possibilities in terms of new methods and techniques to analyze material behavior; for the contemporary materials scientist this is of particular interest, since these methods are invariant to the process used – experiments or computations – to acquire the dataset, allowing one to obtain deeper insights into the processing-structure-property-performance relationship for a material.

4. Current Challenges

The path of digital transformation in MSE requires the acquisition, management, analysis and dissemination of data. In particular, we note that the acquisition and subsequent processing of experimental data needs to be performed digitally, and efforts must be taken to increase the synergy between experiments and simulations. A number of challenges exist on this path towards digitalization (see, e. g., [25, 26]). These challenges are generally listed as

des Simulationsprogramms, zu den digitalen Werkzeugen (z. B. Kopplungskonstanten, Zeitschrittweite, etc.) und zu den einzelnen Schritten bei der Erzeugung atomistischer Strukturen, etc. Solche Schlüsselinformationen oder Metadaten sind entscheidend für das Reproduzieren von Ergebnissen.

Beim digitalen Workflow wären solche Metadaten sowie Informationen zur Datenverarbeitung Teil der Publikation selbst, sodass ein Reproduzieren wissenschaftlicher Ergebnisse deutlich vereinfacht bzw. erst ermöglicht wird. Die Digitalisierung kann somit dazu führen, dass Forschungsdaten besser zugänglich und nutzbar werden [28, 29].

Darüber hinaus würde die Kenntnis und Verfügbarkeit verschiedenster Software-Werkzeuge die Lernkurve von Forschern abflachen. Wie bereits zuvor erwähnt, werden durch DSI auch weitere Möglichkeiten im Hinblick auf neue Verfahren und Techniken zur Analyse des Werkstoffverhaltens eröffnet; für den Werkstoffwissenschaftler von heute ist dies von besonderem Interesse, da diese Methoden für den Prozess, der zur Gewinnung der Datensätze – ob durch experimentelle Untersuchung oder Berechnung – angewandt wird, gleich sind, wodurch ein tieferes Verständnis des Zusammenhangs zwischen Bearbeitung, Gefüge, Eigenschaften und Leistungsfähigkeit eines Werkstoffs ermöglicht wird.

4. Aktuelle Herausforderungen

Um den Digitalen Wandel im Bereich MSE einzuleiten, ist die Erhebung, das Management, die Analyse und Veröffentlichung von Daten erforderlich. Besonders sei hier angemerkt, dass die Erhebung und anschließende Verarbeitung von aus Experimenten gewonnenen Daten digital erfolgen muss und Anstrengungen unternommen werden müssen, um ein besseres Zusammenwirken von Experimenten und Simulationen zu erreichen. Es gibt einige Herausforderungen, die im Hinblick

The Four Vs – Volume, Variety, Veracity and Velocity. Although they are generally discussed vis-à-vis Big Data, these challenges are present for any data itself.

auf den Digitalen Wandel zu bewältigen sind (s. z. B. [25, 26]). Diese Herausforderungen werden gemeinhin als die „Vier Vs“ bezeichnet – Volume (Menge), Varietey (Vielfältigkeit), Veracity (Vertrauenswürdigkeit) und Velocity (Geschwindigkeit). Obwohl diese normalerweise vor dem Hintergrund von Big Data diskutiert werden, gelten diese Herausforderungen für jede Art von Daten.

Volume: In recent years, improved imaging technologies and high throughput experiments, together with increased computing power and the availability of high performance computing resources has resulted in large volumes of data from both experiments and simulations that are amenable to Big Data approaches. The size of data is in the range from a few terabytes to few hundreds of terabytes, or even petabytes. This is very much in contrast to the situation just a couple of decades ago, where materials science and engineering suffered from a lack of data, rather than from big data. The challenge of “Big”ness of this data is not only in handling such large volumes of data which require significantly improved infrastructure, including storage space and network bandwidth, but also in the availability of tool sets to analyze such data. Financing such infrastructure can easily be beyond the capability of a typical research group; long term storage of data is incompatible with short-term project based funding of most funding agencies. Universities, on the other hand, point out to the project-specific nature of the generated data and expect researchers to obtain third party funding for the same. Even if data is made available via third party resources, analyzing them requires the knowledge of sophisticated methods and tools. The fields of life sciences, astronomy, or particle physics are often quoted as leading examples of DSI; in addition to developing their tools and software they have been able to achieve this success by imparting training. To date,

Volume: In den vergangenen Jahren wurden durch verbesserte bildgebende Verfahren, Hoch-Durchsatz-Untersuchungen sowie durch verbesserte Rechenleistung und die Verfügbarkeit von Hochleistungs-Datenverarbeitungsressourcen große Datenmengen aus Experimenten und Simulationen generiert, welche sich für Big-Data-Ansätze eignen. Die Größe der Daten liegt im Bereich von wenigen hundert Terabyte bis hin zu Petabytes. Dies steht im deutlichen Gegensatz zur Situation noch vor einigen Jahrzehnten, als sich die Materialwissenschaft und Werkstofftechnik nicht etwa großen Datenmengen, sondern einem Mangel an Daten gegenüber sahen. Die Herausforderung in Bezug auf die Menge dieser Daten liegt nicht nur in der Verarbeitung solch großer Datenmengen, die eine deutlich verbesserte Infrastruktur, darunter Speicherplatz und Netzwerkbandbreite, erfordern, sondern auch in der Verfügbarkeit von Tools zur Analyse solcher Daten. Die Finanzierung einer solchen Infrastruktur kann leicht die Möglichkeiten einer typischen Forschungsgruppe übersteigen; die langfristige Archivierung von Daten lässt sich nicht mit der kurzfristigen projektbasierten Förderung vieler Trägereinrichtungen vereinbaren. Andererseits weisen Universitäten auf den projektspezifischen Charakter der erzeugten Daten hin und erwarten von Forschern, dass diese dafür Fördermittel von Dritten beschaffen. Selbst wenn Daten durch Drittmittel bereitgestellt werden, erfordert deren Analyse das Wissen um ausgefeilte Methoden und Tools. Die Bereiche Biowissenschaften, Astronomie oder Teilchenphysik werden oft als führende Beispiele für DSI genannt; neben

efficient and effective working with large datasets is still not the common practice in the MSE community.

Variety: Variety refers to the heterogeneous nature of data, that is primarily due to the wide variety of used methods and techniques. The reason is two-fold: Many phenomena in materials are inherently multiscale and therefore require descriptions on many, possibly interlinked length and time scales. An example is plasticity, which is governed by dislocations, which in turn consist of displaced atoms. On the other end of the length scale, e. g., the macroscopic hardening behavior is an emergent property of the phenomena on smaller scales. The second reason for the heterogeneous nature of data is the fact that many materials science problems are of interdisciplinary nature and may require the expertise of physicists, chemists, biologists and engineers alike, resulting in the usage of a wide variety of equipment and tools. The field of tribology with its roots in nanoscience, chemistry and engineering is a particularly instructive example for this. Heterogeneity of datasets may not be completely avoidable, particularly when using commercial software with proprietary file formats. But even open source simulation codes suffer from this problem. This makes it difficult to combine datasets from different sources and perform analysis, or even use some data as input for other methods as, e. g., needed in experimentally informed simulations or in multiscale approaches.

der Entwicklung ihrer eigenen Tools und Software konnten sie diesen Erfolg durch entsprechende Fortbildungen verbuchen. Bislang ist effizientes und effektives Arbeiten mit großen Datenmengen immer noch nicht die allgemeine Praxis innerhalb der MSE-Gemeinschaft.

Variety: Variety bezieht sich auf den heterogenen Charakter von Daten, welcher sich hauptsächlich durch die Vielzahl der angewandten Verfahren und Techniken ergibt. Der zweifache Grund liegt hierin: Viele Phänomene, die bei Werkstoffen beobachtet werden, sind grundsätzlich skalenübergreifend und erfordern somit Beschreibungen auf vielen, möglicherweise verknüpften Längen- und Zeitskalen. Ein Beispiel ist die Plastizität, die auf Versetzungen beruhen, welche wiederum aus versetzten Atomen bestehen. Am anderen Ende der Längenskala liegt bspw. das makroskopische Verfestigungsverhalten als auftretende Eigenschaft, bestimmt durch Phänomene auf kleineren Skalen. Der zweite Grund für den heterogenen Charakter von Daten ist die Tatsache, dass viele Probleme innerhalb der Werkstoffwissenschaften interdisziplinärer Natur sind und gegebenenfalls die Expertise sowohl von Physikern, Chemikern, Biologen als auch Ingenieuren erfordern, was die Nutzung einer Vielzahl von Geräten und Tools zur Folge hat. Das Feld der Tribologie, das seine Wurzeln in der Nanowissenschaft, Chemie und Ingenieurwissenschaft hat, ist hierfür ein besonders aufschlussreiches Beispiel. Die Heterogenität von Datensätzen kann nicht vollständig vermieden werden, vor allem, wenn kommerzielle Software mit firmeneigenen Dateiformaten genutzt wird. Aber sogar Open-Source-Codes zu numerischen Simulationen weisen dieses Problem auf. Dies erschwert die Kombination von Datensätzen aus verschiedenen Quellen und die Durchführung von Analysen oder sogar die Nutzung bestimmter Daten als Input für andere Verfahren, wie es bspw. bei Simulationen, die auf experimentellen Untersuchungen beruhen, oder bei Multiskalen-Ansätzen erforderlich ist.

Veracity: Veracity refers to the "trueness" or correctness of data, which can be guaranteed only if every dataset is accompanied by a detailed documentation of not only the data itself, but also on the work flow used to generate or acquire the data. Furthermore, details on checks performed to ensure the quality of the dataset is also desirable. Such practices are not yet standardized across the MSE community. A simple illustration of such a workflow, that involves generation of atom probe tomography informed atomistic samples (cf. Ref. [11]) is shown in Fig. 2. This workflow involves open-source software (Blender [48], NanoSCULPT [44]) as well as custom-built scripts and programs in multiple programming languages viz. Python, R and Fortran. Rarely is the work flow documented in terms of the exact sequence of tools/scripts/software used, and even rarely in terms of the exact version, making the replication of data almost impossible. A related problem is that of availability – even in cases where the work flow is documented, reviewing a dataset (e. g., in a journal peer review process) may fail due to lack of software licenses, and as a result, the evaluation process can only submit a plausibility of the data. The lack of uniform standards is also visible in many further aspects: for instance, it is unclear if, when and where raw data ought to be stored, since such storage is rarely regarded as necessary. Accepted practice is to draw scientific insights from the raw data and store only the meta data information. Furthermore, raw data is usually stored locally on the machine where the experiment was performed or the simulation was run and rarely backed-up. The raw data is often discarded shortly after the findings are published as an article in a journal. Evaluating the published dataset, however, would require the original raw data, which is either unavailable, or requires significant effort to be reproduced. These problems

Veracity: Veracity bezieht sich auf die "Wahrheit" oder Richtigkeit von Daten, die nur dann garantiert werden kann, wenn zu jedem Datensatz eine genaue Dokumentation vorliegt, die sich nicht nur auf die Daten selbst bezieht, sondern auch den Workflow beschreibt, durch den die Daten erzeugt und gewonnen wurden. Außerdem sind Angaben zu den Kontrollen, die zur Sicherstellung der Qualität des Datensatzes durchgeführt wurden, wünschenswert. Solche Praktiken sind innerhalb der MSE-Gemeinschaft noch nicht vereinheitlicht. Eine einfache Darstellung eines solchen Workflows, der die Erzeugung von atomistischen Proben mit Hilfe tomographischer Atomsonden (vgl. [11]) beinhaltet, ist in Bild 2 zu sehen. Dieser Workflow beinhaltet Open-Source-Software (Blender [48], NanoSCULPT [44]) sowie kundenspezifische Scripts und Programme in mehreren Programmiersprachen, nämlich Python, R und Fortran. Selten wird der Workflow im Hinblick auf die genaue Sequenz der angewandten Tools/Scripts/Software dokumentiert, sogar selten im Hinblick auf die genaue Version, was das Reproduzieren von Daten fast unmöglich macht. Ein damit zusammenhängendes Problem ist das der Verfügbarkeit – sogar in Fällen, in denen der Arbeitsablauf dokumentiert ist, kann die Überprüfung eines Datensatzes (z. B. durch ein Peer-Review-Verfahren für Veröffentlichungen in Fachzeitschriften) aufgrund fehlender Software-Lizenzen fehlschlagen und so kann nur eine Bewertung hinsichtlich der Plausibilität der Daten erfolgen. Das Fehlen einheitlicher Standards spiegelt sich auch in vielen weiteren Gesichtspunkten wider: zum Beispiel ist unklar, ob, wann und wo Primärdaten gespeichert werden sollten, da deren Speicherung selten als notwendig erachtet wird. Eine gängige Praxis ist, dass wissenschaftliche Erkenntnisse aus Primärdaten gewonnen werden und nur die Metainformationen gespeichert werden. Außerdem werden Primärdaten normalerweise nur lokal auf dem Gerät abgespeichert, mit dem das Experiment durchgeführt wurde oder auf dem die Simulation ausgeführt wurde und das selten unter Sicherung der Daten. Primärdaten

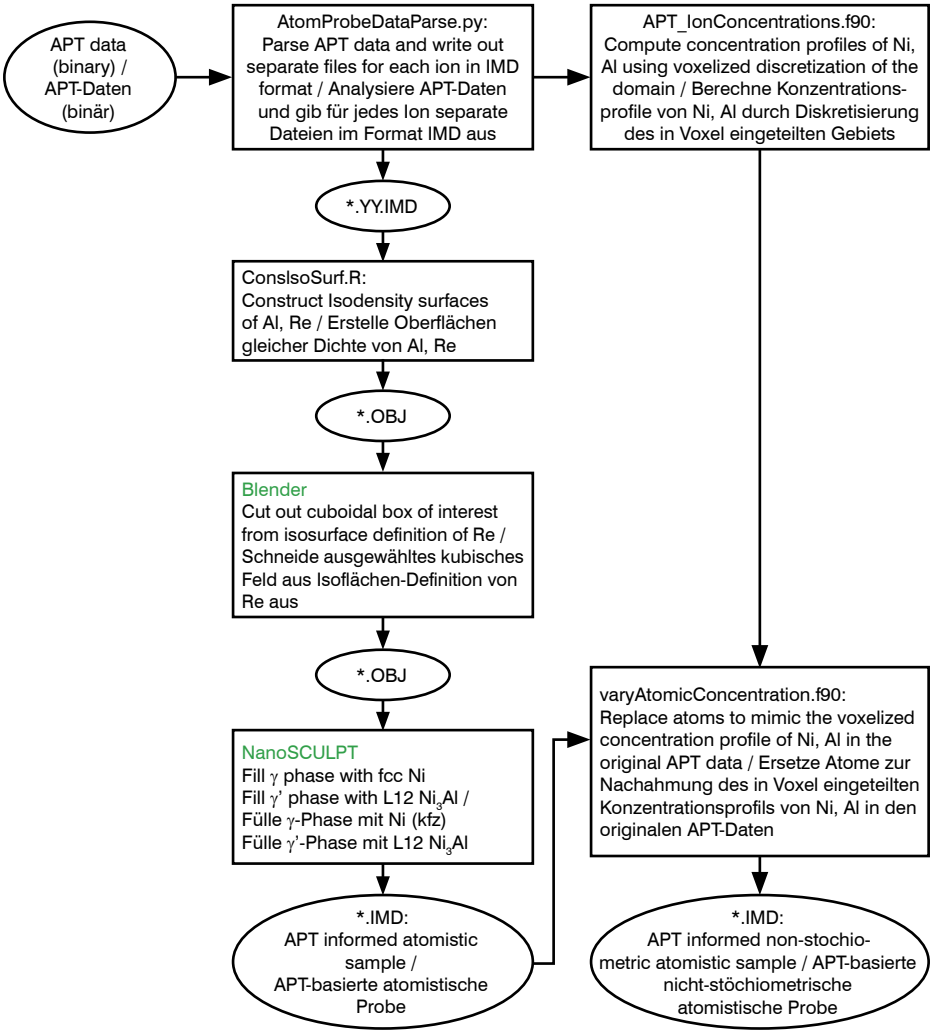

Fig 2: *A workflow diagram illustrating the generation of atom probe tomography (APT) informed stoichiometric and non-stoichiometric atomistic samples in Ref. [11]. Files are indicated in ellipses, and the processing method in a rectangular box. Only the original APT data obtained from experiments is in a binary file format. The workflow involves open source software (denoted in green) and custom built scripts/programs in different programming languages – Python (*.py), R (*.R), Fortran (*.f90).*

Bild 2: *Workflow-Diagramm zur Erzeugung von stöchiometrischen und nicht-stöchiometrischen atomistischen Proben mit Hilfe tomographischer Atomsonden (APT, Atom Probe Tomography) aus Ref. [11]. Dateien werden durch Ellipsen angegeben, die Bearbeitungsverfahren werden durch Rechtecke angezeigt. Nur die ursprünglichen APT-Daten aus experimentellen Untersuchungen liegen im Binärdateiformat vor. Der Workflow beinhaltet Open-Source-Software (grün markiert) und kundenspezifische Scripts/Programme in verschiedenen Programmiersprachen – Python (*.py), R (*.R), Fortran (*.f90).*

are particularly aggravated when a researcher leaves the group.

Velocity: Velocity refers to the rate with which data is produced. The bottleneck is, in particular, visible in real time data analysis, e. g., from a digital camera. While handling large data streams is clearly a formidable task, in many current MSE applications this mainly reduces to handling large amounts of data. Nonetheless, it is well possible that the bandwidth, or the connection and linkage of data streams from, e. g., different microscopy methods will require tailored strategies as well.

Other challenges: Besides the aforementioned four Vs, there is at least one additional, major obstacle for digital transformation and DSI in the MSE community: The lack of a strong **data sharing culture**. A consequence of current established scientific practices is that only the scientific insights, and not the acquired data itself, is deemed relevant for a publishable study. The career path of the scientist is influenced by the number and quality of publications, together with associated statistics such as citations and h-index. Since data itself cannot be published or cited, there appears to be little incentive for research groups to make their data available. Researchers appear to be even more cautious about sharing the data with the wider community due to lack of clear guidelines that clarify how such data may be further used, and who is responsible for missing data and/or misinterpretation and misrepresentation of shared data. It is also unclear if a mere citation to the original publication present-

werden oft kurz nach der Veröffentlichung der Ergebnisse in einem Artikel oder einer Fachzeitschrift verworfen. Eine Bewertung des veröffentlichten Datensatzes würde allerdings die originalen Primärdaten erfordern, die entweder nicht verfügbar sind oder die nur mit großem Arbeitsaufwand reproduziert werden können. Diese Probleme werden besonders dann verstärkt, wenn ein Forscher die Gruppe verlässt.

Velocity: Velocity bezieht sich auf die Geschwindigkeit, mit der Daten erzeugt werden. Engpässe werden vor allem deutlich bei Echtzeit-Datenanalysen, z. B. bei Digitalkameras. Die Handhabung großer Datenströme stellt zweifelsohne eine schwierige Aufgabe dar. Bei vielen Anwendungen im Bereich MSE geht es meist nur um die Handhabung großer Datenmengen. Dennoch ist es gut möglich, dass die Bandbreite oder Verbindung und Verknüpfung von Datenströmen bspw. aus verschiedenen Mikroskopieverfahren ebenfalls maßgeschneiderte Strategien erfordern.

Weitere Herausforderungen: Neben den bereits erwähnten „Vier Vs" gibt es mindestens ein weiteres großes Hindernis für den Digitalen Wandel und DSI innerhalb der MSE-Gemeinschaft: Das Fehlen einer ausgeprägten **Kultur des Datenaustauschs**. Eine Folge der aktuell etablierten wissenschaftlichen Praktiken liegt darin, dass lediglich die wissenschaftlichen Erkenntnisse und nicht die gewonnenen Daten selbst als relevant für eine publizierbare Studie angesehen werden. Die berufliche Karriere von Wissenschaftlern wird unter anderem durch die Anzahl und Qualität ihrer Veröffentlichungen sowie dazugehörigen Statistiken, z. B. Quellangaben und dem h-Index, beeinflusst. Da Daten selbst nicht veröffentlicht oder angegeben werden können, scheint es für Forschungsgruppen wenig Anreiz zu geben, ihre Daten zur Verfügung zu stellen. Forscher scheinen beim Datenaustausch mit einer größeren wissenschaftlichen Gemeinschaft sogar noch vorsichtiger zu sein, da es keine eindeutigen Regeln gibt, die definieren, wie solche Daten weiter genutzt werden können und wer ver-

ing the data would suffice (as dictated by current practice), or if the primary authors need to be acknowledged as co-authors in the new publication, and furthermore, how such data is to be cited in secondary publications. For instance, usage of datasets from material databases results in citations for the database, and not the original work to which the data is attributed.

5. Some Steps towards Digitalization

Digitalization has different facets and hence requires a number of different actions and aspects to be taken into account. In the following, we present a choice of what we consider the most important ones.

5.1 Data Formats: Standardization and Interfaces

To tackle the problem of variety – a bottleneck for efficient data exchange between methods and groups – there are two fundamentally different approaches: (i) develop a unifying standard for file formats and/or interfaces that is applicable and well-accepted within a specific field of work (e.g., raw data in X-ray tomography or finite element analysis); (ii) accept the heterogeneous nature of data, and develop interfaces and converters – with sufficient documentation – to ensure compatibility between different file and/or data formats. We believe that in order to gain widespread acceptance, a healthy mix of the two strategies is required. Indeed, the two strategies can be seen as complementary and synergetic [30]. The NOMAD project [31] is one such example; it is a code-independent

antwortlich für fehlende Daten und/oder Fehlinterpretationen und falsche Darstellungen ausgetauschter Daten ist. Ebenfalls ist nicht klar, ob eine einfache Nennung der ursprünglichen Publikation, in der die Daten vorgestellt wurden, ausreichen würde (entsprechend der aktuellen Praxis) oder ob die Hauptautoren als Co-Autoren der neuen Veröffentlichung genannt werden müssen und wie diese Daten außerdem in Zweitveröffentlichungen angegeben werden sollen. Zum Beispiel wird bei der Nutzung von Datensätzen aus Werkstoffdatenbanken die entsprechende Datenbank angegeben und nicht die ursprüngliche Arbeit, der die Daten zugeschrieben werden.

5. Einige Schritte hin zur Digitalisierung

Die Digitalisierung hat verschiedene Facetten und setzt somit voraus, dass verschiedene Maßnahmen und Gesichtspunkte berücksichtigt werden. Im Folgenden wird eine Auswahl der von uns als am wichtigsten erachteten Punkte vorgestellt:

5.1 Dateiformate: Standardisierung und Schnittstellen

Um das Problem der Vielfältigkeit anzugehen – ein Hindernis im Hinblick auf effizienten Datenaustausch zwischen Verfahren und Gruppen – ergeben sich zwei gänzlich verschiedene Ansätze: (i) die Entwicklung eines vereinheitlichenden Standards für Dateiformate und/oder Schnittstellen, der in einem bestimmten Tätigkeitsfeld anwendbar und allgemein anerkannt ist (z.B. Primärdaten bei der Röntgentomographie oder Finite-Elemente-Methode); (ii) die Akzeptanz, dass Daten einen heterogenen Charakter haben und die Entwicklung von Schnittstellen und Konvertern – mit ausreichender Dokumentation – um die Kompatibilität zwischen verschiedenen Datei- oder Datenformaten sicherzustellen. Wir glauben, dass eine gesunde Mischung beider Strategien erforderlich ist, um breite Akzeptanz zu erlangen. Beide Strategien können in der Tat als komple-

database which stores data from electron structure calculations from a wide variety of codes, and converts them to a format that allows for analytics to be directly performed on the database. An example on a different context is the HDF5 format [32] which provides a standard for storing possibly compressed data and which at the same time provides an XML-like interface for storing meta data along with the data such that the data structure itself can act as documentation. HDF5 is platform independent and works on Windows, Linux or Mac. Furthermore, wrappers for the most important programming languages including Python and Matlab exist. In Fig. 3 we show a mock dataset containing force-displacement data along with an image and meta data. Creating this file requires one line of Matlab code for each attribute (meta data entry) or two lines of code for any dataset.

Besides the above, rather technical aspects, there is also an aspect of "variety" that is directly related to the underlying physics: when combining or comparing different experimental and/or simulation approaches, each of them often concentär und synergetisch angesehen werden [30]. Das NOMAD-Projekt [31] ist ein solches Beispiel; es handelt sich um eine Code-unabhängige Datenbank, in der Daten aus Berechnungen zur Elektronenstruktur ausgehend von einer Vielzahl von Codes gespeichert und in ein Format umgewandelt werden, welches Analysen direkt in der Datenbank ermöglicht. Ein Beispiel aus einem anderen Kontext ist das Datenformat HDF5 [32], das ein Standard für die Speicherung von eventuell komprimierten Daten ist und das gleichzeitig eine XML-ähnliche Schnittstelle zur Speicherung von Daten und Metadaten bietet, sodass die Datenstruktur selbst als Dokumentation dienen kann. HDF5 ist plattformunabhängig und läuft auf Windows, Linux oder Mac. Außerdem gibt es Wrapper für die wichtigsten Programmiersprachen, darunter auch Python und Matlab. Bild 3 zeigt einen Pseudo-Datensatz, der Kraft-Weg-Daten zusammen mit einem Bild und Metadaten enthält. Das Erstellen dieser Datei erfordert eine Zeile Matlab-Code für jedes Attribut (Metadaten-Eintrag) oder zwei Codezeilen für jeden Datensatz.

Neben den oben genannten, recht technischen Gesichtspunkten gibt es außerdem den Aspekt der Vielfältigkeit („Variety"), der in direktem Zusammenhang mit der zugrundeliegenden Physik steht: bei der Kombination oder dem Vergleich verschiedener experimenteller

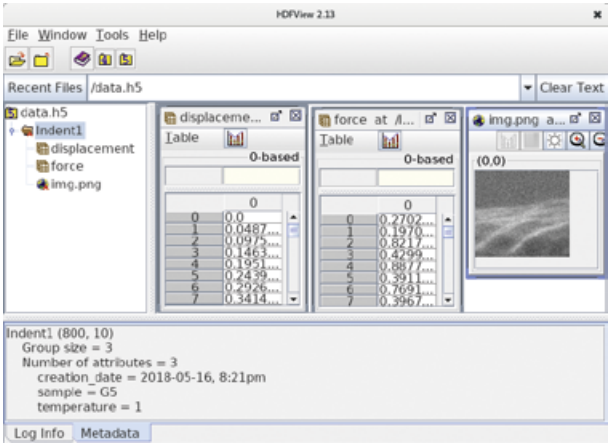

Fig 3: *Screenshot of the content of a HDF5 file, which is able to store heterogeneous data, including even images. Additionally, meta data, like e. g., creation date, sample ID, temperature etc. can be stored as well, and is shown for the group "indent1" at the bottom of the window.*

Bild 3: *Screenshot des Inhalts einer HDF5-Datei, die heterogene Daten speichern kann, u. a. auch Bilder. Zudem können Metadaten wie z. B. Erstellungsdatum, Proben-ID, Temperatur, etc. gespeichert werden, zu sehen unten im Fenster für die Gruppe "indent1".*

trates on different phenomenon or length/time scale. In these cases, physics-based conversions of different data types have to be sought [33, 34], which may involve averaging or filtering of data, e. g., when going from high-resolution data to coarse grained descriptions. Knowledge of the underlying materials scientific problem is, in such cases, very important for deciding which details need to be included.

5.2 Integrated Workflow Tools

In order to ensure faithful replication of every step along the data processing chain, it is necessary to use integrated workflow tools, that can create an automated protocol of the individual steps and software used in the chain. Such a tool is essentially a platform with a working environment that has access to tools and software in use by the researcher. As a result, a multitude of open source, commercial and self-developed software and scripts can be used, and the workflow is automatically registered in the platform. The automation removes the onus from the user, who may sometimes forget to document the workflow if done manually.

Such a procedure can be successfully implemented as pipelines as is to be found in tools like the synthetic microstructure generation tool DREAM3D [35], or via processing graphs as is the case of LabView [36]. Other tools like Taverna [37], or Drake [38], are more generic, and allow for the overall workflow to be encapsulated, and furthermore, backed up via a cloud storage.

5.3 Training on Best Practices for Software Development

Contemporary scientific software, particularly free and open-source software displays a wide spectrum of documentation

und/oder Simulationsansätze konzentriert sich jeder Ansatz oft auf verschiedene Phänomene oder Längen-/Zeitskalen. In diesen Fällen muss eine auf der Physik basierende Umwandlung verschiedener Datentypen angestrebt werden [33, 34], was mit einer Mittelung oder Filterung von Daten verbunden sein kann, z. B. wenn von hochauflösenden Daten in grobkörnige Beschreibungen übergegangen wird. In solchen Fällen ist die Kenntnis des zugrundeliegenden werkstoffwissenschaftlichen Problems für die Entscheidung, welche Angaben einzubeziehen sind, überaus wichtig.

5.2 Integrierte Workflow-Tools

Um eine originalgetreue Reproduktion eines jeden Schritts der Datenverarbeitungskette sicherzustellen, ist die Nutzung integrierter Workflow-Tools erforderlich, die ein automatisches Protokoll der einzelnen Schritte und der in der Kette verwendeten Software erstellen können. Ein solches Tool ist im Grunde eine Plattform mit einer Arbeitsumgebung, die Zugriff auf Tools und Software hat, die vom Forscher verwendet werden. Folglich können eine Vielzahl an Open-Source-, kommerzieller oder selbstentwickelter Software und Scripts verwendet werden und der Workflow wird automatisch auf der Plattform erfasst. Dank der Automatisierung entfällt diese Pflicht für den Nutzer, der bei der manuellen Erfassung bisweilen vergessen könnte, den Workflow zu dokumentieren.

Ein solches Verfahren kann erfolgreich mit Hilfe von Pipelines implementiert werden, wie beim DREAM3D-Tool zur Erzeugung synthetischer Gefüge [35] oder durch Verlaufsgraphen wie bei LabView [36]. Andere Tools wie Taverna [37] oder Drake [38] sind allgemeiner und ermöglichen die Kapselung des gesamten Workflows und zudem die Sicherung mittels cloud-basierter Speicherung.

5.3 Weiterbildung in bewährten Vorgehensweisen bei der Software-Entwicklung

Aktuelle wissenschaftliche Software, besonders kostenlose oder Open-Source-Software zeigt ein breites Spektrum an Dokumentation

and adherence to standards. One may broadly classify such software on three different tiers:

- Tier 1 contains software that are distributed by dedicated groups/institutions, and is well maintained with regular patches, bug fixes and updates of patches. Examples of such software include LAMMPS (Atomistic simulations), DREAM3D (Synthetic microstructure generation) [35], Deal.II (finite element toolbox) [40]. Such packages usually contain extensive documentation, and have a support system either via forums or support groups.
- Tier 2 contains software that is usually maintained by individuals and small research groups, that are made available to the wider community, albeit with varying levels of documentation. Support is usually provided by the development team itself, and rarely has a support group. Examples include ParaDiS [41], MicroMEGAS [42], DAMASK [43], NanoSCULPT [44], FE2AT [45], etc.
- Tier 3 contains task-specific scripts and tools, like parsers and job schedulers, or even software written for larger functions and purposes, including simulations and analysis. These are usually developed by individual scientists, who in many cases have little or no training in formal requirement analysis, software development and management, and quality assurance. Such tools, for instance, are rarely archived properly with version numbers, which is extremely important for replicating a dataset at a later stage.

To ensure consistent standards, but easy software development, where the quality procedures do not overwhelm a researcher, we recommend the following quality

und Einhaltung von Standards. Solche Softwares lassen sich grob in 3 verschiedene Ebenen einteilen:

- Ebene 1 beinhaltet Software, die von engagierten Gruppen/Institutionen vertrieben wird und durch regelmäßige Patches, Bugfixes und Patch-Updates zuverlässig gepflegt wird. Beispiele hierfür sind u. a. LAMMPS (Atomistische Simulationen), DREAM3D (Erzeugung synthetischer Gefüge) [35] und Deal.II (Finite-Elemente-Toolbox) [40]. Solche Pakete beinhalten für gewöhnlich eine umfangreiche Dokumentation und System-Support entweder über Foren oder Support-Gruppen.
- Ebene 2 beinhaltet Software, die für gewöhnlich von Einzelpersonen oder kleinen Forschungsgruppen gepflegt und einer der größeren Gemeinschaft zur Verfügung gestellt wird, wenn auch mit unterschiedlichen Dokumentationsstufen. Der Support erfolgt normalerweise durch das Entwicklerteam selbst, in seltenen Fällen gibt es Support-Gruppen. Beispiele sind u. a. ParaDiS [41], MicroMEGAS [42], DAMASK [43], NanoSCULPT [44], FE2AT [45], etc.
- Ebene 3 beinhaltet aufgabenspezifische Scripts und Tools wie Parser und Scheduler oder sogar Software, die für umfangreichere Funktionen und Zwecke programmiert wurde, darunter Simulationen und Analysen. Diese werden für gewöhnlich von einzelnen Wissenschaftlern entwickelt, die oftmals wenig oder gar keine praktische Ausbildung im Hinblick auf formelle Bedarfsanalyse, Software-Entwicklung und -management sowie Qualitätskontrolle haben. Z.B. werden solche Tools selten korrekt unter Angabe der Version archiviert, was überaus wichtig für die Reproduktion eines Datensatzes zu einem späteren Zeitpunkt ist.

Um sicherzustellen, dass einheitliche Standards existieren und die Software-Entwicklung einfach gehalten wird, um einen Forscher nicht durch Qualitätsverfahren zu überfordern,

levels for version control and documentation:

- Level 0: A bare minimum level of version control; every script/program where a piece of code changes the functionality is stored as a different file, for instance, with a "_v#" indicating the version number appended to the name (e.g. my_analysis_script_v07.m); On execution, the program shall provide the user with all necessary information for the current run; Documentation of functionality is provided in the code itself; User may be expected to read the source code in case of missing information.

- Level 1: Version control via a repository – source code is stored in a central repository and managed via a versioning system (git, mercurial, svn etc.). If compilation is required, an appropriate "Makefile" for build tools such as CMake or Make, along with corresponding instructions is also part of the repository. Detailed documentation is provided; hence source code screening is unnecessary.

- Level 2: Automated build server without/with packager – A central development environment is established to automate component and functionality testing. Documentation is web-based and can be independent of the build server; automated tools extract documentation from the code. The packager makes the process easier for the user by packing the entire program into a single file; "installation" can be performed by a single click. Documentation is self-contained in the program itself. Systems/Provide such as GitHuB [46] or GitLab [47] come with a number of tools all of which can be connected to the central repository and are additionempfehlen wird die folgenden Qualitätsstufen in Bezug auf Versionsverwaltung und Dokumentation:

- Stufe 0: Absolutes Mindestmaß an Versionsverwaltung; jedes Script/Programm, bei dem ein Stück Code die Funktionalität ändert, wird als eine andere Datei abgespeichert, z. B. mit einem „_v#", das die Version angibt und an den Namen angehängt wird (z. B. my_analysis_script_v07.m); Bei der Ausführung stellt das Programm dem Nutzer alle notwendigen Informationen für den aktuellen Durchlauf zur Verfügung; Die Dokumentation der Funktionalität ist im Code selbst enthalten; Vom Nutzer wird ggf. erwartet, im Fall von fehlenden Informationen den Quellcode zu lesen.

- Stufe 1: Versionsverwaltung durch ein Repository – der Quellcode wird in ein zentrales Repository gespeichert und durch ein Versionsverwaltungssystem (git, mercurial, svn etc.) verwaltet. Falls eine Kompilierung erforderlich ist, ist ein entsprechendes „Makefile" für Build-Tools wie CMake oder Make zusammen mit entsprechenden Anweisungen ebenfalls Teil des Repository. Eine detaillierte Dokumentation ist gegeben; somit ist eine Überprüfung des Quellcodes nicht erforderlich.

- Stufe 2: Automatisierte Server für fortlaufende Integration, auch Buildserver genannt, mit/ohne Packager – Eine zentrale Entwicklungsumgebung wird zur Automatisierung von Komponenten- und Funktionstests geschaffen. Die Dokumentation ist internetbasiert und kann unabhängig vom Buildserver sein; Automatisierte Tools extrahieren die Dokumentation aus dem Code. Der Packager erleichtert den Prozess für den Benutzer, indem das gesamte Programm in eine einzige Datei gepackt wird; Die „Installation" kann durch einen einzigen Klick ausgeführt werden. Die Dokumentation ist eigenständig und unabhängig im Programm selbst. Systeme/Dienste wie GitHuB [46] oder GitLab [47] sind erhältlich mit einer Reihe von

ally able to perform unit, functionality and integration tests.

The implementation of such standards, particularly levels 0 and 1, requires minimal effort, but often fails due to lack of time or motivation. Training on the available tools would help make the development workflow easier. Therefore, it would be extremely useful to include a minimal course program in materials science curricula that provides information and training on available tools and best practices, and usefulness of the same.

6. Conclusion

Digitalization in materials science and engineering shows great promise for bringing experiment and simulation through a data-related approach closer together. At the same time, with every step towards a digitalization the amount of available data increases. While data handling is certainly a challenge, we hope that the sheer number of new possibilities that are opening up is tempting enough for a larger number of researchers and research groups to invest the initial time and effort in order to realize the full potential of these new approaches. Additionally, we have to include some of these aspects in the regular academic education of students in MSE too. Last but not least, this process also requires the availability of new funding possibilities since large data handling and Big Data strategies require storage, hardware and additional manpower. Nonetheless, digitalization is a process that has already begun, and implementing small measures even in scattered groups will contribute to a successful progression.

Tools, die alle mit dem zentralen Repository verbunden werden können und zusätzlich in der Lage sind, Komponenten-, Funktions- und Integrationstests durchzuführen.

Die Einführung solcher Standards, besonders der Stufen 0 und 1, erfordert ein Minimum an Aufwand, scheitert jedoch oft an fehlender Zeit oder Motivation. Fortbildungen zu verfügbaren Tools würden dazu beitragen, den Entwicklungs-Workflow zu erleichtern. Aus diesem Grund wäre es außerordentlich nützlich, ein Mindestmaß an Lehrveranstaltungen in die Lehrpläne der Werkstoffwissenschaften aufzunehmen, in denen Informationen und Übungen zu den verfügbaren Tools sowie bewährten Vorgehensweisen angeboten werden und deren Nützlichkeit vermittelt wird.

6. Schlussfolgerung

Digitalisierung im Bereich der Materialwissenschaft und Werkstofftechnik ist ein vielversprechendes Unterfangen, durch das experimentelle Untersuchungen und Simulationen einander durch einen datenbasierten Ansatz näher gebracht werden. Gleichzeitig wird mit jedem Schritt hin zur Digitalisierung die Menge der verfügbaren Daten größer. Während die Datenverarbeitung sicherlich eine Herausforderung darstellt, hoffen wir, dass allein die Zahl der sich eröffnenden neuen Möglichkeiten für immer mehr Forscher und Forschungsgruppen Anreiz sind, die anfängliche Zeit und Mühe zu investieren, um das Potenzial dieser neuen Ansätze voll auszuschöpfen. Zusätzlich dazu müssen einige dieser Gesichtspunkte auch in die reguläre akademische Ausbildung von Studierenden im Bereich MSE einfließen. Zu guter Letzt verlangt dieser Prozess auch die Verfügbarkeit neuer Fördermöglichkeiten, da große Datenmengen und Big-Data-Strategien Speichermöglichkeiten, Hardware und zusätzliche Arbeitskraft erfordern. Dennoch ist die Digitalisierung ein Prozess, der längst begonnen hat und die Umsetzung kleiner Maßnahmen selbst in vereinzelten Forschungsgruppen wird zu einer erfolgreichen Weiterentwicklung beitragen.

Acknowledgements

The authors acknowledge funding from the European Research Council Starting Grant, "A Multiscale Dislocation Language for Data-Driven Materials Science," ERC Grant Agreement No. 759419 MuDiLingo. We thank Mykhaylo Motylenko Christina Wüstefeld, and Stefan Martin (Institute of Materials Science, TU Bergakademie Freiberg) for providing us with microscopy images, Peter Felfer (FAU) and Dan Gianola (UCSB) for providing atom probe tomogrpahy data, and Daniel Weygand for providing us with the discrete dislocation dynamics image in Fig. 1. The authors would also like to thank Dominik Steinberger, for the images for multivariate analysis, principal component analysis and cluster analysis, in Fig. 1.

Danksagungen

Die Autoren bedanken sich für die Förderung durch den „Starting Grant" des Europäischen Forschungsrats (ERC) „A Multiscale Dislocation Language for Data-Driven Materials Science", ERC Grant Agreement No. 759419 MuDiLingo. Wir danken Mykhaylo Motylenko, Christina Wüstefeld und Stefan Martin (Institut für Werkstoffwissenschaft, TU Bergakademie Freiberg) für die zur Verfügung gestellten mikroskopischen Aufnahmen und Daniel Weygand für die Aufnahme zur diskreten Versetzungsdynamik in Bild 1. Die Autoren danken außerdem Dominik Steinberger für die Aufnahmen zu Multivariaten Analysemethode, Hauptkomponentenanalyse und Clusteranalyse in Bild 1.

References / Literatur

[1] Hey, T.; Tansley, S.; Tolle, K. (Eds.): Microsoft Research, Redmond, Washington, 2009

[2] Sumpter, B. G.; Vasudevan, R. K.; Potok, T.; Kalinin, S. V.: NPJ Computational Materials (2015), 15008

[3] Ramprasad, R.; Batra, R.; Pilania, G.; Mannodi-Kanakkithodi, A.; Kim, C.: NPJ Computational Materials (2017) 3, 54
DOI: 10.1038/s41524-017-0056-5

[4] Kalidindi, S. R.; De Graef, M.: Annual Reviews in Materials Research (2015) 45, 171–193
DOI: 10.1146/annurev-matsci-070214-020844

[5] Friedrich, H.; de Jongh, P. E.; Verkleij, A. J.; de Jong, K. P.: Chemical Reviews (2009) 109, 1613–1629
DOI: 10.1021/cr800434t

[6] Belianinov, A.; Vasudevan, R.; Strelcov, E.; Steed, C.; Yang, S. M.; Tselev, A.; Jesse, S.; Biegalski, M.; Shipman, G.; Symons, C.; Borisevich, A.; Archibald, R.; Kalinin, S.: Advanced Structural and Chemical Imaging (2015) 1, 6
DOI: 10.1186/s40679-015-0006-6

[7] Fernandez, J.-J.: Current Opinion in Solid State and Materials Science (2013) 17, 93–106
DOI: 10.1016/j.cossms.2013.03.002

[8] Curtarolo, S.; Hart, G. L. W.; Nardelli, M. B.; Mingo, N.; Sanvito, S.; Levy, O.: Nature Materials (2013), 12, 191–201
DOI: 10.1038/nmat3568

[9] Greeley, J.; Jaramillo, T. F.; Bonde, J.; Chorkendorff, I.; Norskov, J. K.: Nature Materials (2006), 5, 909–913
DOI: 10.1038/nmat1752

[10] Kelly, T. F.; Larson, D. J.: Annual Review of Materials Research (2012), 42, 1–31
DOI: 10.1146/annurev-matsci-070511-155007

[11] Prakash, A.; Guénolé, J.; Wang, J.; Müller, J.; Spiecker, E.; Mills, M. J.; Povstugar, I.; Choi, P.; Raabe, D.; Bitzek, E.: Acta Materialia (2015) 92, 33–45
DOI: 10.1016/j.actamat.2015.03.050

[12] Prakash, A.; Hummel, M.; Schmauder, S.; Bitzek, E.: MethodsX (2016) 3, 219–230
DOI: 10.1016/j.mex.2016.03.002

[13] Prakash, A.; Bitzek, E.: Materials (2017) 10, 88
DOI: 10.3390/ma10010088

[14] Tasan, C. C.; Diehl, M.; Yan, D.; Bechtold, M.; Roters, F.; Schemann, L.; Zheng, C.; Peranio, N.; Ponge, D.; Koyama, M.; Tsuzaki, K.; Raabe, D.: Annual Review of Materials Research (2015) 45, 391–431
DOI: 10.1146/annurev-matsci-070214-021103

[15] Lim, H.; Carroll, J. D.; Battaile, C. C.; Buchheit, T. E.; Boyce, B. L.; Weinberger, C. R.: International Journal of Plasticity (2014) 60, 1–18
DOI: 10.1016/j.ijplas.2014.05.004
[16] Möller, J. J.; Prakash, A.; Bitzek, E.: Modelling and Simulation in Materials Science and Engineering (2013) 21, 055011
DOI: 10.1088/0965-0393/21/5/055011
[17] Diehl, M.; Groeber, M.; Haase, C.; Molodov, D. A.; Roters, F.; Raabe, D.: Journal of Materials (2017) 69, 848–855
[18] Prakash, A.; Weygand, D.; Bitzek, E.: International Journal of Plasticity (2017) 97, 107–125
DOI: 10.1016/j.ijplas.2017.05.011
[19] Sandfeld, S.; Po, G.: Modelling and Simulation in Materials Science and Engineering (2015) 23, 085003
DOI: 10.1088/0965-0393/23/8/085003
[20] Steinberger, D.; Gatti, R.; Sandfeld, S.: Journal of Materials (2016) 68, 2065–2072
[21] McDowell, D. L.: International Journal of Plasticity (2010) 26, 1280–1309
DOI: 10.1016/j.ijplas.2010.02.008
[22] Dewald, M.; Curtin, W. A.: Modelling and Simulation in Materials Science and Engineering (2011) 19, 055002
DOI: 10.1088/0965-0393/19/5/055002
[23] Prakash, A.; Nöhring, W.; Lebensohn, R. A.; Höppel, H. W.; Bitzek, E.: Materials Science and Engineering A (2015) 631, 104–119
DOI: 10.1016/j.msea.2015.02.005
[24] Rajan, K.: Materials Today (2005) 8, 38–4
DOI: 10.1016/S1369-7021(05)71123-8
[25] Rajan, K.: Annual Review in Materials Research (2015) 45, 153–169
DOI: 10.1146/annurev-matsci-070214-021132
[26] Agarwal, A.; Choudhary, A.: APL Materials (2016) 4, 053208
DOI: 10.1063/1.4946894
[27] Sandfeld, S.; Dahmen, T.; Fischer, F. O. R.; Eberl, C.; Klein, S.; Selzer, M.; Nestler, B.; Möller, J.; Mücklich, F.; Engstler, M.; Diebels, S.; Tschuncky, R.; Prakash, A.; Steinberger, D.; Kübel, C.; Herman, H.-G.; Schubotz, R.: Strategiepaper – Digitale Transformation in der Materialwissenschaft und Werkstofftechnik. Available at https://www.dgm.de/medien/print-medien/strategiepapier-digitale-transformation/; Accessed on 28 May 2018
[28] Pfeif, E. A.; Kroenlein, K.: APL Materials (2016) 4, 053203
DOI: 10.1063/1.4942634
[29] Hill, J.; Mulholland, G.; Persson, K.; Seshadri, R.; Wolverton, C.; Meredig, B.: MRS Bulletin (2016) 41, 399–409
DOI: 10.1557/mrs.2016.93
[30] Ghiringhelli, L. M.; Carbogno, C.; Levchenko, S.; Mohamed, F.; Huhs, G.; Lüders, M.; Oliveira, M.; Scheffler, M.: NPJ Computational Materials (2017) 3, 46
DOI: 10.1038/s41524-017-0048-5
[31] The Novel Materials Discovery (NOMAD) Laboratory: European Center of Excellence. URL: https://www.nomad-coe.eu/; accessed 30 May 2018
[32] The HDF5 Group; URL: https://support.hdfgroup.org/, accessed 27 May 2018
[33] Gunkelmann, N.; Alhafez, I.; Steinberger, D.; Urbassek, H.; Sandfeld, S.: Computational Materials Science (2017) 135, 181–188
DOI: 10.1016/j.commatsci.2017.04.008
[34] Kositski, R.; Steinberger, D.; Sandfeld, S.; Mordehai, D.: Computational Materials Science (2018) 149, 125–133
DOI: 10.1016/j.commatsci.2018.02.058
[35] Blue Quartz Software. URL: https://dream3d.bluequartz.net; accessed 30 May 2018
[36] National Instruments; Laboratory Virtual Instrument Engineering Workbench (LabVIEW). URL: https://www.ni.com/labview; accessed 30 May 2018
[37] Taverna Workbench; URL: https://taverna.incubator.apache.org/; accessed 30 May 2018
[38] Drake Workflow Management Tool; Factual.com. Available at https://github.com/Factual/drake; accessed 30 May 2018
[39] LAMMPS: Large-scale Atomic/Molecular Massively Parallel Simulator. Available at http://lammps.sandia.gov; accessed 30 May 2018
[40] deal.II: An open source finite element library. Available at http://www.dealii.org; accessed 30 May 2018
[41] ParaDiS: Parallel Dislocation Simulator. Available at http://paradis.stanford.edu/site/about; accessed 30 May 2018
[42] MicroMegas: Open source program for dislocation dynamics simulations. Available at http://zig.onera.fr/mm_home_page/; accessed 30 May 2018
[43] DAMASK: The Düsseldorf Advanced Material Simulation Kit. Available at https://damask.mpie.de; accessed 30 May 2018
[44] NanoSCULPT: A tool/methodology to generate complex and realistic structures for atomistic simulations. Available at https://bitbucket.org/arunpksh/nanosculpt/; accessed 30 May 2018

[45] FE2AT: Finite Element informed Atomistic Simulations. Available at https://bitbucket.org/arunpksh/fe2at; accessed 30 May 2018
[46] GitHub. URL: https://github.com; accessed 30 May 2018
[47] GitLab. URL: https://about.gitlab.com; accessed 30 May 2018
[48] Blender: Open Source 3D Creation Suite. Available at https://www.blender.org; accessed 30 May 2018
[49] He, W. J.; Zhang, S. H.; Prakash, A.; Helm, D.: Computational Materials Science (2014) 82, 466–476 DOI: 10.1016/j.commatsci.2013.10.023
[50] Bueno, P. R.; Varela, J. A.: Materials Research (2006) 9, 293–300. Licensed under *Creative Commons Attribution License* DOI: 10.1590/S1516-14392006000300009

Bibliography

DOI 10.3139/147.110539
Pract. Metallogr. 55 (2018) 8; page 493–514

ISSN 0032–678X

Aruna Prakash

obtained his PhD from the Karlsruhe Institute of Technology. He then worked as PostDoc at Fraunhofer IWM and senior scientist at the Friedrich-Alexander-Universität Erlangen-Nürnberg. In 2018, he joined the Chair of Micromechanical Materials Modelling at the TU Bergakademie Freiberg as senior scientist.

Stefan Sandfeld

obtained his PhD from The University of Edinburgh. After a postdoctoral stay at the Karlsruhe Institute of Technology, he joined the Friedrich-Alexander-Universität Erlangen-Nürnberg as senior scientist. Since 2017, he is Professor of Micromechanical Materials Modelling at the TU Bergakademie Freiberg.