C. R. Mecanique 347 (2019) 845–855

Contents lists available at ScienceDirect

Comptes Rendus Mecanique

www.sciencedirect.com

Data-Based Engineering Science and Technology / *Sciences et technologies de l'ingénierie basées sur les données*

Data-driven modeling and learning in science and engineering

Francisco J. Montáns [a,*], Francisco Chinesta [b], Rafael Gómez-Bombarelli [c], J. Nathan Kutz [d]

[a] *Escuela Técnica Superior de Ingenieros Aeronáuticos, Universidad Politécnica de Madrid, Plaza Cardenal Cisneros 3, 28045 Madrid, Spain*
[b] *ESI Group Chair @ PIMM, Arts et Métiers ParisTech, 151, boulevard de l'Hôpital, 75013 Paris, France*
[c] *Department of Materials Science and Engineering, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA*
[d] *Department of Applied Mathematics, University of Washington, Seattle, WA 98195, USA*

ARTICLE INFO

Article history:
Received 24 July 2019
Accepted after revision 10 August 2019
Available online 14 November 2019

Keywords:
Data-driven science
Data-driven modeling
Artificial intelligence
Machine learning
Data-science
Big data

ABSTRACT

In the past, data in which science and engineering is based, was scarce and frequently obtained by experiments proposed to verify a given hypothesis. Each experiment was able to yield only very limited data. Today, data is abundant and abundantly collected in each single experiment at a very small cost. Data-driven modeling and scientific discovery is a change of paradigm on how many problems, both in science and engineering, are addressed. Some scientific fields have been using artificial intelligence for some time due to the inherent difficulty in obtaining laws and equations to describe some phenomena. However, today data-driven approaches are also flooding fields like mechanics and materials science, where the traditional approach seemed to be highly satisfactory. In this paper we review the application of data-driven modeling and model learning procedures to different fields in science and engineering.

© 2019 Académie des sciences. Published by Elsevier Masson SAS. All rights reserved.

1. Introduction

The walk of humankind in life is a complex journey of learning through observation and experiencing of the world, from which we collect property data, sometimes easily quantifiable, sometimes more qualitative. Also, from observation, we relate events to that property data. It is from repetitive experiences from which we can determine some patterns that relate events to data, and events to events themselves. In the case of science discovery, these patterns and relations are formalized in laws and equations, the data are formalized in properties and variables, and the observations are formalized in event measurements, which may be actions or properties themselves. Laws and equations, typical in science, allow us to perform predictions and facilitate the transmission of the learning procedure in a very compact manner, with the minimum amount of information. However, the classical process of learning in science is a slow process which needs much observational experience, usually from expensive proposed experiments, as to discover the main variables involved and their influence on events for a probably huge amount of possible combinations, missing frequently unforeseen relevant variables. Furthermore, the classical scientific approach is hypotheses-driven and hence is biased by them. The scientific method was established

* Corresponding author.
E-mail addresses: Fco.Montans@upm.es (F.J. Montáns), Francisco.Chinesta@ensam.eu (F. Chinesta), rafagb@mit.edu (R. Gómez-Bombarelli), kutz@uw.edu (J.N. Kutz).

https://doi.org/10.1016/j.crme.2019.11.009
1631-0721/© 2019 Académie des sciences. Published by Elsevier Masson SAS. All rights reserved.

because of the natural biases and weaknesses of the human mind, including the natural bias humans have when seeking metaphysical explanations which are not based on real observations ([1] and [2], citing Francis Bacon in Novum Organum Scientiarum, AD 1620). However, the classical scientific method is still biased by the deductive thinking of the human mind. Data-driven procedures seek, if possible, an unbiased implicit approach to our learning experience based on raw data from real observations. These procedures have the additional advantage of testing correlations between different variables and observations, learning unforeseen patterns in nature and allowing us to discover new scientific laws or even more, performing predictions without the availability of such laws. We are living the era of data science, which impacts all aspects of life [3]. Data-driven procedures are giving rise to a new economy. Data-based science will also change our lives and how we do science. Data collection, data-mining [4] and data-visualization will also be of paramount importance in science discovery. Data-driven scientific discovery is considered the fourth paradigm [5], so founding agencies have begun to invest significantly in data-driven science (The Economist, Briefing 5/6/2017). For example, National Science Foundation of the USA granted in 2014 $31 million in awards to lay the groundwork in data-science (NSF News Release 14-132). The purpose of this brief review of the relatively novel field of data-driven modeling in science and engineering, is to give a scent of different approaches and applications in several scientific and engineering fields of data-driven modeling. Therefore, sampling some approaches and applications is our purpose; no intention is made to include all possible works, applications or procedures, but just to give a big picture of some paths followed in different fields.

2. Data-driven discovery in science

2.1. Computer power facilitates computer-based discovery

The ever-increasing rise of computational power in the past few decades has led to significant advances in statistical and machine learning techniques [6]. The collection of algorithms and data mining methods developed as a result have formed the core mathematical architecture of artificial intelligence (AI) agents, and although AI has a long history in scientific discovery [7], data-driven approaches in modern computers can now ingest and process algorithms at scale. This has been largely enabled by the plummeting costs of sensors, computational power, and data storage technologies. Indeed, such vast quantities of data afford us new opportunities for data-driven discovery, which, as mentioned, has been referred to as the 4th paradigm of science [5].

In most application fields in the engineering, physical and biological sciences, physical models are expressed as a set of governing constitutive relations, spatio-temporal relations, and/or dynamical systems. Data-driven discovery in these application areas are specifically constructed to discover constitutive relations and differential governing equations in a wide variety of fields, [8], conservation laws [9] or propagation of nonlinear waves [10], spatio-temporal dynamics [11], development of predictive procedures for molecular dynamics for nanoscale flow [12], and health monitoring [13]. In particular, data-driven techniques may be extremely important in complex areas of life sciences [14], allowing for unveiling unknown biological mechanisms [15]. Thus, there are increasingly funding initiatives to obtain new methods, software tools and training within the framework of Big Data analysis in health [13]. In the field of data-driven modeling in agro-environmental science, an overview may be found in [16] and therein references.

From the Schrödinger equation of quantum mechanics to Maxwell's equations for electromagnetic propagation, knowing the governing laws has allowed for transformative technological impact in society (e.g., smart phones, internet, lasers, satellites). And just as Newton built upon the work of Kepler and others, proposing the existence of gravity in order to derive $F = ma$ and explain Kepler's elliptic orbits, the discovery of a fundamental governing law is critical for technological development, enabling unprecedented engineering and scientific progress, such as sending a rocket to the moon.

2.2. Algorithms for data-driven modeling, discovery of laws and learning physical constraints

The recent and rapid increase in the availability of measurement data of physical systems has spurred the development of many data-driven methods for modeling and predicting dynamics. At the forefront of data-driven methods are deep neural networks (DNNs). DNNs not only achieve superior performance for tasks such as image classification, but they have also been shown to be effective for future state prediction of dynamical systems [10]. A key limitation of DNNs, and similar data-driven methods, is the lack of interpretability of the resulting model: they are focused on prediction and do not provide governing equations or clearly interpretable models in terms of the original variable set. An alternative data-driven approach uses symbolic regression to identify directly the structure of a nonlinear dynamical system from data [17]. This works remarkably well for discovering interpretable physical models, but symbolic regression is computationally expensive and can be difficult to scale to large problems. However, the discovery process can be reformulated in terms of sparse regression [8], providing a computationally tractable alternative, thus leveraging the power of symbolic regression with computational tractability. These contrasting techniques show the diversity of strategies that can be employed to extract meaningful physics from data. They also highlight the fact that machine learning and artificial intelligence algorithms may be capable of learning physics principles and constraints [9]. Using modern sparse regression architectures and neural networks, several critical tasks may be enacted from data alone: (i) the discovery of first principles models, (ii) the identification of physical constraints and conservation laws, and (iii) improved models using known physics. A diversity of architectures

allows one to also develop black box and gray box[1] modeling strategies for complex systems where physics is only partially known. Additionally, the architecture not only allows one to impose physics constraints, or bake-in physics, but it can also discover physical constraints that need to be baked-in, i.e. one can constrain learning and one can learn constraints [18]. Thus, not only may parsimonious and interpretable physical models be discovered as a direct result of such strategies, but critical insights such as conservation laws and physical constraints could also be discovered. These innovations have the potential to discover generalizable models which can be modified to handle multi-scale physics, noisy systems, and limited data.

3. Data-driven modeling in mechanical engineering and materials science

3.1. The change of paradigm in solid mechanics

Whereas data-driven (big-data) applications have been extensively used in many fields for more than a decade, this type of approach has attracted the attention only recently to researchers in the field of modeling of solids. One of the reasons for this is that traditionally, the mechanics of solids has followed a quite successful deterministic approach in which, with relatively little available experimental data, relatively meaningful predictions where obtained in general, complex situations. Furthermore, information about the behavior of a material has been traditionally passed to the community through the specification of few material parameters for a specific constitutive model. However, the data-driven change of paradigm has reached also the solid mechanics community. These are most probably the main reasons. (1) Currently the computational power is large, so the analysis of nonlinear solids is routinely being performed in industry. This has fostered interest in simulating more complex materials, which at the same time are generated and optimized by material scientists to achieve some desired properties. (2) The variety found in biological materials, including living materials, along with the difficulty of their characterization because of the different structure from specimen to specimen, from location to location, and because of aging and time, also prompted the search for modeling tools that are not based on a specific modeling structure or function, but that can represent a larger variety of materials, conceptually similar but with a wide span of possible behaviors. Within this family of approaches, are constitutive manifold approaches. (3) Currently there is a large amount of available material data for many types of materials, so there is a need for unstructured modeling that is capable for assimilating these data, possibly obtained from diverse types of testing or observations under different conditions. (4) There are currently successful model order reduction techniques which reduce the curse of dimensionality, allowing us to develop a modern version of the slide rule, precomputing off-line the problem for many possibilities and saving a reduced representation of them, so information can be easily passed over and used to rebuild specific solutions at a given time.

3.2. Constitutive manifolds and reduced representations for multiscale analysis

One of the main problems addressed currently by data-driven models is the multiscale analysis of heterogeneous materials, see [19] and the more recent works [20–25]. For the case of soils, multi-field, multi-scale poroplasticity data-driven modeling using recursive homogenizations and deep learning has been performed by Wang and Sun [26]. Nonlinear hyperelastic problems have been also addressed in [27] through constitutive maps. Here, one of the main purposes of the data-driven procedures is to virtually test microstructured solids, probably with very complex microstructures, and to develop a constitutive manifold, see Fig. 1. This manifold is computed off-line, often through reduced sampling and reduced representation (Fig. 1), for example hyper-reduction [28,29] and Proper Generalized Decompositions (PGD) [20,30,31]. PGD approximations have also been used combined with the LATIN approach for multiscale analysis [32]. In [33], PGD is efficiently considered for data-assimilation. Once an off-line representation is obtained, during analysis at the continuum level, a material behavior representation closest to the precomputed manifold is searched. The advantage is the general representation of material behavior and moderate online computational effort. To this regard, clustering techniques have been presented as a tool for avoiding the curse of dimensionality [34,35]. Clustering has been used for long time in data driven techniques, see, e.g., [36]. In [37,38], see also [39], numerically explicit spline potentials (NEXP) are proposed to represent the hyperelastic behavior of multiscale materials, and the coefficients are computed sampling the material in a strain space. The advantage is that an analytical differentiable function is available and constitutive tangents for Newton schemes are easily computed.

Within the framework of solid mechanics, there are other applications where data-driven approaches are increasingly pursued. For example, in [40], within the context of Integrated Computational Materials Science Engineering (ICME), reduced-order data-driven modeling is employed for assessing the high cycle fatigue performance of polycrystalline alpha-Ti structures. Fatigue problems are also analyzed in [41] and also in [24], where a data-driven approach is used to identify the small fatigue crack driving force. A review of data science in materials science is given in [42]. Diffusion in random heterogeneous media is studied in [43]. A review of data-driven techniques, classification of variables and material properties, and specially of machine learning in materials informatics, is given in [44]. In [45], model reduction techniques are

[1] In contrast to theoretical (fully determined from physics, no data needed) and black-box models (fully determined from data, no physical understanding needed), a "gray box" model is a model employing some physics or basic understanding, but which still needs to be calibrated from, or completed with, data.

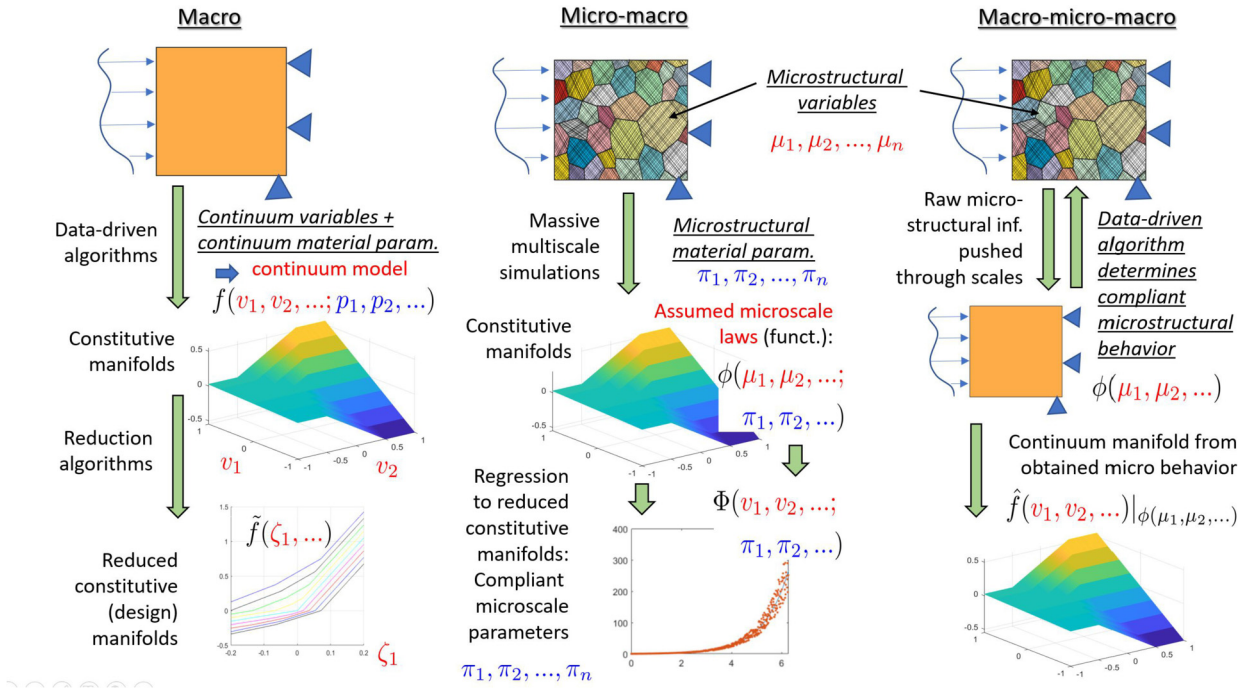

Fig. 1. Data-driven approaches in solid mechanics. Macro (left): variables and possible parameters are macroscopic; data-driven procedures determine the constitutive/design manifolds from observed macroscopic behavior. Micro-macro (center): a model for the microscopic behavior is used, with material parameters meaningful at the microscopic scale. Massive simulations are performed to develop either micro or macro constitutive manifolds as a function of parameters at the micro scale. Reduced representations may be used at any level. Macro-micro-macro (right): Minimal microscopic information is used (e.g., the material structure), assumed constitutive laws are avoided. The raw kinematic microstructural dependencies are pushed to the continuum scale, carrying the microstructural variables. Compliant macro-micro behavior is obtained at the continuum level, solving both behaviors at once.

also employed for viscoplastic analyses. Complex relations between variables of processes and structure relations in additive manufacturing are also obtained through a multiscale, multiphysics approach in [31]. A different application is given in [22], where the nonlinear anisotropic electrical response of graphene/polymer nanocomposites is studied employing computational homogenization based on neural networks. Neural networks have also been used in the characterization of materials with a flexible analytical model in the background (e.g., [46], [47]). This same group used genetic algorithms and data-based metamodels of friction laws (friction response surfaces) in the design of tires to optimize their wear performance [48], [49].

3.3. Continuum approach and representation functions

As mentioned, data-driven approaches in solid mechanics entail usually a large number of finite element computations to determine the behavior of the material for a representative amount of combinations and loading domain, typically tens [50] or hundreds of thousands [38] of analyses. Then, it is important to select a proper reduced sampling and representation approach [21]. Somehow in this line is the cubic spline representation for hyperelastic materials [51,52], which has been extended to anisotropic materials [53,54] and damage [55]. This type of procedure has also been employed in large strain nonequilibrium (strain-level dependent) viscoelasticity [56,57]. Data-driven approaches are also used in solid mechanics for different purposes. For example, to be able to determine the behavior of a material from observed deformation patterns of structures [58–60]. The idea is to fully substitute the material laws by constitutive manifolds [61,62] preserving only the conservation laws. A data-driven formulation fully consistent with thermodynamics was proposed in [33,63]. The initial small strain deterministic approach in [58] has also been extended to noisy data and dynamics [64], where Shannon entropy-maximizing schemes, frequently used in image processing [65] are employed. Similar problems are addressed in the animation industry, where realistic simulations of the deformations of cloth for realistic visual perception are also pursued [66]. A hybrid approach combining first order models with data-based enrichment was addressed in [67].

3.4. Bio-inspired models and data-driven modeling in biomechanics

In the field of biomechanics, and specially in characterizing soft tissues, cells and their behavior, data-driven approaches look promising, because a deep knowledge that may bring traditional laws, and even relations between variables, is lacking. For example, in [68] data-driven reduced-order models from atomistic simulations are employed to develop a microtubule model for cells. The interest in data-driven approaches for biomechanics is highlighted in [69]. Within the context of biological systems, a polynomial order heuristic algorithm is developed in [70] with the purpose of inferring the governing

behavior of dynamical systems. A review of data-driven modeling of biological processes, at different scales and from different perspectives, is given in [71]. Biological systems also inspire data-driven approaches as the so-called artificial immune systems (see, e.g., [72–74]). This is an adaptative computational system which replicates some properties of the immune system as error tolerance, redundancy and diversity of systems, distribution of tasks, dynamic learning, system adaptation and, specially, self-monitoring. This technique has been used, for example, by [75,76] for damage detection in composites and in Structural Health Monitoring (SHM). In SHM, the combination of physics-based assessment and of data-driven procedures in damage identification can be found in [77]. Previous data-driven approaches, as stochastic subspace identification, were also used for finite element model updating in damage evolution in [78].

3.5. Physics- and structure-based data-driven approaches

The difficulties in considering all aspects of engineering modeling and science with data-driven procedures, and the interest in taking advantage of the learnings from the classical approach, is currently motivating a mixed approach in which data-driven modeling is guided by some physical insight [79]. The purpose of developing mixed approaches is to improve the reliability of the obtained relations through fundamental principles, like conservation laws (e.g., energy conservation and maximum entropy).

The usual approaches, explained in the previous paragraphs, are either macro or micro-macro approaches, see Fig. 1. In the former, all the procedure is performed and the continuum scale. In micro-macro (structure-based) approaches, constitutive relations are defined at the microscale as a function of material parameters. In a macro-micro-macro approach (Fig. 1), the microscale only defines some kinematic relations between micromechanical variables. These relations are pushed to the continuum scale relating them to macroscopic ones. At the continuum level, a data-driven method is applied determining the behavior at both scales [80].

4. Data-driven procedures in other engineering fields

Apart from the already mentioned engineering applications, data-driven procedures are being used in a large variety of other engineering fields. In this section we just sample representative applications in different topics.

4.1. Industrial processing

Data-driven techniques have been used for some years in industrial processing both for monitoring and for predicting. A review of these techniques may be found in [81]. Data-driven techniques combined with physically-based models are also present in virtual, digital and hybrid twins as reported in [82]. Data-driven modeling of the production processes in the automotive industry can be found in [83]. In the production of biofuels, created by microorganisms, where there is a need to engineer the microorganism's metabolism, the optimization of the host and the pathways as to maximize the production of the fuel is performed by data-driven approaches in [84]. Soft sensors are computer-based virtual sensors giving information about a process. They are used, for example, in new automobiles to give remaining fuel reading, avoiding oscillations of the gauge. An early review of data-driven soft sensors in the chemical production industry are reviewed in [85]. To improve the duration of products and to avoid problems due to stochastic variations in batch processes, a subspace-aided data-driven approach is proposed in [86] and applied to fed-batch penicillin production. Physics-based data-driven modeling is being proposed in production engineering and in control, e.g., to control heating/cooling systems in buildings [87].

4.2. Gas and oil industry

Of course, the oil and gas industry, as well as earth scientists and engineers, have also benefit from data-driven procedures in all aspects of modeling soil behavior, exploration and production, including seismic analysis, reservoir characterization, management and production. The book [88] gives a summary of data-driven techniques used in the field. Data mining, pattern recognition and machine learning algorithms are used in [89] for full reservoir modeling of shale assets for hydrocarbon production. Mixed physics-based, data-driven approaches are also starting to be proposed in this field, see, e.g., [90–92]. In the case of earthquake engineering, Song et al. [93] give data-driven computer codes for accurately predicting the performance of buildings from raw databases.

4.3. Fluid and particle dynamics

Data-driven procedures are being increasingly used in fluid dynamics, especially to improve accuracy of simulations in turbulence when using Reynolds Averaged Navier Stokes (RANS) modeling. Duraisamy [94] developed a data-driven approach to model turbulent and translational flows. From their data-driven procedure, applying inverse problems, they inferred the functional form of deficiencies in known closure models and then improved them using machine learning to obtain accurate predictions; see also [95]. Data-driven approaches (convolutional networks) are also used in [96] to accelerate Navier-Stokes simulations. Data-driven modeling to improve the prediction of the Reynolds stress anisotropy in the shear layer of jet-in-crossflow simulations is also used [97]. In [98], a review of data-driven techniques to model turbulence, with emphasis in

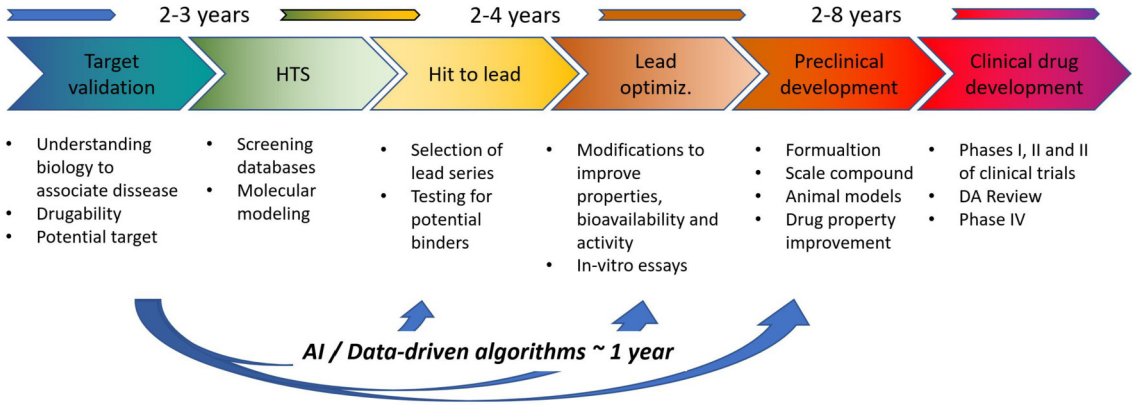

Fig. 2. The steps to drug-discovery. Artificial intelligence and data-driven procedures using multiple datasets/databases may substantially accelerate the first steps, in which molecules are selected, combined, improved and refined, saving important funds.

reducing uncertainties in RANS models is given. Data-driven dimension reduction procedures applied to dynamical systems, both for modal decompositions and for transfer functions, are studied in [99] among others. Reduced order representations based on dynamic mode decomposition methods [100] combined with proper orthogonal decompositions are also employed in naval hull shape optimization for improved drag and lift properties [101]. Three different methods to learn slosh dynamics and advancing the solution in time have also been studied in [102], namely Proper Orthogonal Decomposition, Locally Linear Embedding and Topological Data Analysis. These methods provided solutions faster than real time in a laptop computer.

A different application especially well-suited for data-driven procedures is the simulation of crowds. The behavior of crowds and individuals within a crowd has been frequently modeled as fluids and particles within it, but it is recognized that actual crowds follow complex laws because they react to external and internal stimuli. The works [103,104] develop an agent-based data-driven simulation procedure to simulate crowds in computer graphics simulations. The movements of the crowds have been obtained from aerial recordings, from which the behavior of individuals obtained from those of surrounding ones was devised. A similar work, but aimed at emergency plans, is presented in [105].

4.4. Bioengineering

Bioengineering is another field in which data-driven algorithms may find extraordinary applications. For example, physiological variables as blood glucose and hormones such as insulin, cortisol, epinephrine, glucagon, and their dynamic interrelations determine the metabolic conditions in patients with diabetes. Data-driven models for diagnosis, glucose prediction in insulin-dependent patients and treatment management may be found, for example, in the book edited by Marmarelis and Mitsis [106]. In [107], an inverse data-driven regression procedure is developed to compute the cardiac electric diffusivity from electrocardiogram signals. Also, in medical imaging, [108] employ non-parametric density estimation and edge confidence maps in the segmentation of brain images obtained from magnetic resonance imaging. Of course, data-driven tools as an aid to medical diagnosis and decisions have been used for more than two decades [109]. Examples of applications are [110] for breast cancer and [111] for coronary heart disease. A review of Big Data applications in biomedical research may be found in [112].

5. Data-driven chemistry and drug discovery

5.1. The relevance of data-driven approaches in chemistry

The discovery of new chemical compounds such as small molecule drugs, and the assignment of new application labels to existing ones are very complex processes. Usually, the lack of deterministic approaches to predict performance from structure and the complexity of carrying out discrete optimization over chemical graphs result in very costly trial-and-error tests to arrive at a product with the desired performance. The procedure of drug discovery typically follows different stages (see Fig. 2): (1) target validation, (2) primary and secondary assay development (high-throughput screening), (3) hit to lead compound, (4) lead optimization, (5) preclinical drug development and (6) clinical drug development [113].

The availability of curated labeled and unlabeled data in the chemical sciences is very large compared with some other physical sciences. Since its origins many decades ago, the Chemical Abstracts Service has compiled a list of over 144 million known substances, and about 67 million protein and DNA sequences (www.cas.org). More than 15,000 substances are added each day. The size of potential chemical space, however, is overwhelmingly larger than our ability to explore it by hand. Different estimates for the number of chemically accessible molecules range from 1,030 to 10,100. Therefore, the number of possible combinations to explore is very large. The speed at which data is being generated is higher than the speed at which

we can analyze them. In this regard, data-driven techniques are important, and they are increasingly being used in guessing or narrowing the search for compounds with given desired characteristics [114]. This is especially important because even though the number of possible targets has been increasing, the actual number of new drug launches is decreasing, whereas the costs associated with their development is increasing steadily [115].

5.2. Data-driven procedures in drug discovery

Computational modeling in drug discovery has been used for some time in industry. Because of the highly competitive landscape and the large economic incentives, drug discovery is the strongest driver in the development of cheminformatics and data-driven tools in chemistry, including data integration [116,117]. An early analysis of the integration of data and knowledge in drug discovery is given in the review paper of Searls [118]. As mentioned by [119], data-driven algorithms must be used to identify compounds with a minimum number of liabilities in lead-like and drug-like hits, since hit lists have more than 1,000 compounds if drug-like hits or leads are not discarded. High-throughput virtual screening approaches have a long history in drug discovery, where predictive machine learning tools are used to prioritize compounds and identify leads in internal databases with millions of compounds (Fig. 2). An interactive data-driven visual analytics technique, named ConTour, was developed by Partl et al. [120] to enable the analysis of compounds based on multi-correlated datasets. A review of data-reduction techniques in drug discovery using principal component analysis, Bayesian analysis, hierarchical clustering, similarity analysis and projections, can be found in [121] and therein references. One of the most appealing recent developments in computer-driven drug discovery is the combination of supervised machine learning models to predict performance with unsupervised models to generate novel promising compounds in a fully automatic manner. Combining the large number of unlabeled chemicals that somehow characterize the nature of accessible chemical space, and the abundant activity labels from high-through combinatorial chemistry, automatic chemical design is closer to practice. In this line, Gómez-Bombarelli et al. [122] have developed an automatic procedure based on continuous vector representations of molecules and the use of neural networks to perform inverse design of molecules. A large number of related works continue to explore the application of deep neural networks to this task, including syntax and grammar based generative models that can write chemical graphs [123–125], deep-reinforcement learning tools [126,127].

6. Data quality and stochastic processes

In any data-driven model or process, data quality is very important, since erroneous or biased data may produce erroneous models and erroneous decisions. A recent dramatic example may be the crash of two B-737-Max airplanes. Whereas the crashes are still under investigation, preliminary findings note that erroneous data from a single angle-of-attack sensor may have produced the anti-stall software to tilt the airplane down, an issue central to the fatal accidents [128]. It is apparent that data redundancy and data time-series analyses may, in most occasions, enhance data quality to yield better models and model predictions. According to ISO 9001:2015 standard, the definition and assessment of data quality depends on the context and use of data, see also [129,130]. There are several reviews on data-quality research, data curation (correction, reorganization, maintenance, integration and preparation of data), data-quality definitions and attributes in different fields, see for example [131,132], and [133] for linked data. The recent book [134] reviews many data quality aspects in different fields.

Within the context of model learning in science and engineering, quantitative data quality, as obtained from experiments, seems central to the quality of models and predictions themselves. In this regard, some characteristics (quality dimensions) are very important, as accuracy, completeness, consistency and credibility [135,129]. There are algorithms for data-quality analysis and curation. For time series, algorithms can detect and correct issues like data shifts, diverging patterns, unphysical patterns, mismatched records or trends, noise patterns, etc, see for example [136,137] and therein references. In some fields like Prognostic and Health Management of systems, data clustering plays a key role in differentiating multiple system conditions. Algorithms are reviewed in [138] regarding clustering differentiation and quality enhancement; see also therein references.

The quality in description of a process is often related to the understanding of its stochastic nature or that of the variables and observations involved, so the application of data-driven methods to stochastic phenomena and stochastic processes is also a current area of research. For example, Hou et al. [139] study the approximation capability of deep generative networks in capturing the posterior distribution in Bayesian inverse problems through learning a transport map. Raissi et al. introduce in [140] the concept of parametric Gaussian processes in order to encode massive amounts of data in a small number of data points. A remarkable feature is that their parametric Gaussian processes quantify the uncertainty of the predictions associated with the process imperfections. Surrogate models facilitate simpler and faster ways of inquiring approximate solutions in the space of design variables. For the case of stochastic, high-dimensional and variable-fidelity-source systems, Yang and Perdikaris [141] present a deep learning probabilistic procedure to construct these predictive data-driven surrogates, based on input-output pairs, with quantified uncertainty. Other works are those of Soize and Ghanem [142] which propose a procedure for generating realizations of a random vector in an unknown subset of the Euclidean space which is consistent with observational data of the vector, and that of Soize and Farhat [143], which presents a fast predictor-corrector approach for computing the vector-valued hyperparameter for a novel nonparametric probabilistic method with the purpose of quantifying the uncertainties of the model-form in nonlinear computational mechanics; see also therein references.

7. Conclusions

The 21st Century is considered the Century of Big Data. The change of century has brought a historic change in our society. Computers, Internet and the new digital devices are producing a large amount of data. Current computational power and cloud computing also allow for an unforeseen number of simulations. However, instead of being overwhelmed by such amount of raw information, we are learning how to take advantage of the new paradigm. Software companies have taught us that much benefit and key information may be obtained from data analytics. Then, we are learning new ways of doing things, among them, science and engineering.

Data-driven procedures focus on data and try to extract variables and relations directly from raw data, giving frequently more accurate responses without the use of classical analytical laws and equations. However, many open questions remain, and in some occasions, drawbacks have been found as the lack of fulfillment of some physical principles. Then, new physics-based data-driven procedures are getting in.

Data-driven procedures are also used to predict what science will be done in science (a field known as Science of Science [144]), which may be relevant to funding agencies and to researchers looking for long term funding [145]. However, we are at the start of a new epoque in which data science has shown many remarkable success cases, so data-driven algorithms are not needed for predicting that funding agencies will increasingly invest more and more funds in data-driven science.

Acknowledgements

FJM acknowledges support from Agencia Estatal de Investigación of Spain, grant PGC-2018-097257-B-C32. JNK acknowledges support from the Air Force Office of Scientific Research (AFOSR) grant FA9550-17-1-0329.

References

[1] Wikipedia team (Addressed 11/22/2018): https://en.wikipedia.org/wiki/Novum_Organum.
[2] F. Mazzocchi, Could big data be the end of the theory in science? A few remarks on the epistemology of data-driven science, EMBO Rep. 16 (10) (2015) 1250–1255.
[3] L. Cao, Data science: a comprehensive overview, ACM Comput. Surv. 50 (3) (2017) 43.
[4] U. Fayyad, G. Piatetsky-Saphiro, P. Smyth, From data mining to knowledge discovery in databases, AI Mag. 17 (3) (1996) 37–54.
[5] T. Hey, S. Tansley, K.M. Tolle, The Fourth Paradigm: Data-Intensive Scientific Discovery, Vol. 1, Microsoft Research, Redmond, WA, 2009.
[6] C.M. Bishop, Pattern Recognition and Machine Learning, Information Science and Statistics, Springer-Verlag New York Inc., Secaucus, NJ, USA, 2006.
[7] P. Langley, J.M. Zytkow, Data-driven approaches to empirical discovery, Artif. Intell. 40 (1989) 283–312.
[8] S.L. Brunton, J.L. Proctor, J.N. Kutz, Discovering governing equations from data by sparse identification of nonlinear dynamical systems, Proc. Natl. Acad. Sci. (2016), 201517384.
[9] J.-C. Loiseau, S.L. Brunton, Constrained sparse Galerkin regression, J. Fluid Mech. 838 (2018) 42–67.
[10] M. Raissi, P. Perdikaris, G.E. Karniadakis, Physics informed deep learning (Part II): data-driven discovery of nonlinear partial differential equations, arXiv:1711.10566, 2017.
[11] S.H. Rudy, S.L. Brunton, J.L. Proctor, J.N. Kutz, Data-driven discovery of partial differential equations, Sci. Adv. 3 (4) (2017) e1602614.
[12] P. Angelikopoulos, C. Papadimitriou, P. Loumoutsakos, Data driven, predictive molecular dynamics for nanoscale flow simulations under uncertainty, J. Phys. Chem. B 117 (47) (2013) 14808–14816.
[13] P.E. Bourne, V. Bonazzi, M. Dunn, E.D. Green, M. Guyer, G. Komatsoulis, J. Larkin, B. Russell, The NIH big data to knowledge (BD2K) initiative, J. Amer. Med. Inform. Assoc. 22 (6) (2015) 1114.
[14] N. Merchant, E. Lyons, S. Goff, M. Vaughn, D. Ware, D. Mickos, P. Antin, The iPlant collaborative: cyberintraestructure for enabling data to discovery for the life sciences, PLoS Biol. 14 (1) (2016) e1002342.
[15] A. Gaudinier, S.M. Brady, Mapping transcriptional networks in plants: data-driven discovery of novel biological mechanisms, Annu. Rev. Plant Biol. 67 (2016) 575–594.
[16] R. Lokers, R. Knapen, S. Janssen, Y. van Randen, J. Jansen, Analysis of Big Data technologies for use in agro-environmental science, Environ. Model. Softw. 84 (2016) 494–504.
[17] M. Schmidt, H. Lipson, Distilling free-form natural laws from experimental data, Science 324 (5923) (2009) 81–85.
[18] K. Kaiser, J.N. Kutz, S.L. Brunton, Discovering conservation laws from data for control, in: 57th IEEE Conference on Decision and Control, Miami Beach, FL, Dec 17–19, 2018, 2018, pp. 6415–6421.
[19] B.A. Le, J. Yvonnet, Q.C. He, Computational homogenization of nonlinear elastic materials using neural networks, Int. J. Numer. Methods Eng. 104 (2015) 1061–1084.
[20] F. El Halabi, D. González, J.A. Sanz-Herrera, M. Doblaré, A PGD-based multiscale formulation for non-linear solid mechanics under small deformations, Comput. Methods Appl. Mech. Eng. 305 (2016) 806–826.
[21] F. Fritzen, O. Kunc, Two-stage data-driven homogenization for nonlinear solids using a reduced order model, Eur. J. Mech. A, Solids 69 (2018) 201–220.
[22] X. Lu, D.G. Giovanis, J. Yvonnet, V. Papadopoulus, F. Detrez, J. Bai, A data-driven computational homogenization method based on neural networks for the nonlinear anisotropic electrical response of graphene/polymer nanocomposites, Comput. Mech. 64 (2019) 307–321.
[23] N.H. Paulson, M.W. Priddy, D.L. McDowell, S.R. Kalidindi, Data-driven reduced-order models for rank-ordering the high cycle fatigue performance of polycrystalline microstructures, Mater. Des. 154 (2018) 170–183, https://doi.org/10.1016/j.matdes.2018.05.009.
[24] A. Rovinelli, M.D. Sangid, H. Proudhon, W. Ludwig, Using machine learning and a data-driven approach to identify the small fatigue crack driving force in polycrystalline materials, Comput. Mech. 4 (2018) 35, https://doi.org/10.1038/s411524-018-0094-7.
[25] W. Yan, S. Lin, O.L. Kafka, Y. Lian, C. Yu, Z. Liu, J. Yan, S. Wolff, H. Wu, E. Ndip-Agbor, M. Mozaffar, K. Ehmann, J. Cao, G.J. Wagner, W.K. Liu, Data-driven multi-scale multi-physics models to derive process-structure-property relationships for additive manufacturing, Comput. Mech. 61 (2018) 521–541.
[26] K. Wang, W. Sun, A multiscale multi-permeability poroplasticity model linked by recursive homogenizations and deep learning, Comput. Methods Appl. Mech. Eng. 334 (2018) 337–380.
[27] I. Temizer, T.I. Zohdi, A numerical method for homogenization in non-linear elasticity, Comput. Mech. 40 (2007) 281–298.
[28] D. Ryckelynck, A priori hyperreduction method: an adaptive approach, J. Comput. Phys. 202 (2005) 346–366.
[29] D. Ryckelynck, Hyper-reduction of mechanical models involving internal variables, Int. J. Numer. Methods Eng. 77 (1) (2009) 75–89.

[30] F. Chinesta, A. Leygue, F. Bordeu, J.V. Aguado, E. Cueto, D. Gonzalez, I. Alfaro, Parametric PGD based computational vademecum for efficient design, optimization and control, Arch. Comput. Methods Eng. 20 (1) (2013) 31–59.
[31] D. Neron, P. Ladeveze, Proper generalized decomposition for multiscale and multiphysics problems, Arch. Comput. Methods Eng. 17 (2010) 351–372.
[32] M. Cremonesi, P.-A. Neron, D. Guidault, P. Ladeveze, A PGD-based homogenization technique for the resolution of nonlinear multiscale problems, Comput. Methods Appl. Mech. Eng. 267 (2013) 275–292.
[33] D. González, A. Badias, I. Alfaro, F. Chinesta, E. Cueto, Model order reduction for real-time data assimilation through extended Kalman filters, Comput. Methods Appl. Mech. Eng. 326 (2017) 679–693.
[34] M.A. Bessa, R. Bostanabad, Z. Liu, A. Hu, D.W. Apley, C. Brinson, W. Chen, W.K. Liu, A framework for data-driven analysis of materials under uncertainly: countering the curse of dimensionality, Comput. Methods Appl. Mech. Eng. 320 (2017) 633–667.
[35] S. Tang, L. Zhang, W.K. Liu, From virtual clustering analysis to self-consistent clustering analysis: a mathematical study, Comput. Mech. 62 (2018) 1143–1460.
[36] P. Ma, C.I. Castillo-Davis, W. Zhong, J.S. Liu, A data-driven clustering method for time course gene expression data, Nucleic Acids Res. 34 (2006) 1261–1269.
[37] J. Yvonnet, D. Gonzalez, Q.-C. He, Numerically explicit potentials for the homogenization of nonlinear elastic homogeneous materials, Comput. Methods Appl. Mech. Eng. 198 (2009) 2723–2737.
[38] J. Yvonnet, E. Monteiro, Q.-C. He, Computational homogenization method and reduced database model for hyperelastic heterogeneous structures, Int. J. Multiscale Comput. Eng. 11 (2013) 201–225.
[39] L. Xia, P. Breitkopf, Multiscale structural topology optimization with an approximate constitutive model for local material microstructure, Comput. Methods Appl. Mech. Eng. 104 (2015) 1061–1084.
[40] N.H. Paulson, M.W. Priddy, D.L. McDowell, S.R. Kalidindi, Data-driven reduced-order models for rank-ordering the high cycle fatigue performance of polycrystalline microstructures, Mater. Des. 154 (2018) 170–183, https://doi.org/10.1016/j.matdes.2018.05.009.
[41] D. Ryckelynck, Hyper-reduction framework for model calibration in plasticity-induced fatigue, Adv. Model. Simul. Eng. Sci. 3 (15) (2016).
[42] S.R. Kalidindi, M. De Graef, Materials data science: current status and future outlook, Annu. Rev. Mater. Res. 45 (2015) 171–193.
[43] B. Ganapathysubramanian, N. Zabaras, Modeling diffusion in random heterogeneous media: data-driven models, stochastic collocation and the variational multiscale method, J. Comput. Phys. 226 (2007) 326–353.
[44] R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi, C. Kim, Machine learning in materials informatics: recent applications and prospects, Comput. Mat. 3 (2017) 54, https://doi.org/10.1038/s41524-017-0056-5.
[45] N. Relun, D. Neron, P.A. Boucard, A model reduction technique based on the PGD for elastic-viscoplastic computational analysis, Comput. Mech. 51 (2013) 83–92.
[46] C. Zopf, M. Kaliske, Numerical characterization of uncured elastomers by a neural network based approach, Comput. Struct. 182 (2017) 504–525.
[47] I. Kopal, I. Labaj, M. Harnicarova, J. Valicek, D. Hruby, Prediction of the tensile response of carbon black filled rubber blends by artificial neural network, Polymers 10 (6) (2018) 644.
[48] A. Serafinska, N. Hassoun, M. Kaliske, Numerical optimization of wear performance. Utilizing a metamodel based friction law, Comput. Struct. 165 (2016) 10–23.
[49] W. Graf, M. Gütz, F. Leichsenring, M. Kaliske, Computational intelligence for efficient numerical design of structures with uncertain parameters, in: 2015 IEEE Symposium Series on Computational Intelligence, 2015, pp. 1824–1831.
[50] S. Bhattacherjee, K. Matous, A nonlinear manifold-based reduced order model for multiscale analysis of heterogeneous hyperelastic materials, J. Comput. Phys. 313 (2016) 635–653.
[51] T. Sussman, K.J. Bathe, A model of incompressible isotropic hyperelastic material behavior using spline interpolations of tension-compression data, Commun. Numer. Methods Eng. 25 (1) (2009) 53–63.
[52] J. Crespo, M. Latorre, F.J. Montáns, WYPIWYG hyperelasticity for isotropic, compressible materials, Comput. Mech. 59 (1) (2017) 73–92.
[53] J. Crespo, F.J. Montáns, Function-refresh algorithms for determining the stored energy density of nonlinear elastic orthotropic materials directly from experimental data, Int. J. Non-Linear Mech. 107 (2018) 16–33.
[54] J. Crespo, F.J. Montáns, General solution procedures to compute the stored energy density of conservative solids directly from experimental data, Int. J. Eng. Sci. 141 (2019) 16–34.
[55] M. Miñano, F.J. Montáns, WYPiWYG damage mechanics for soft materials: a data-driven approach, Arch. Comput. Methods Eng. 25 (2018) 165–193.
[56] M. Latorre, F.J. Montáns, Fully anisotropic finite strain viscoelasticity based on a reverse multiplicative decomposition and logarithmic strains, Comput. Struct. 163 (2016) 56–70.
[57] M. Latorre, F.J. Montáns, Strain-level dependent nonequilibrium anisotropic viscoelasticity: application to the abdominal muscle, J. Biomech. Eng. 139 (10) (2017) 101007.
[58] T. Kichdoerfer, M. Ortiz, Data-driven computational mechanics, Comput. Methods Appl. Mech. Eng. 304 (2016) 81–101.
[59] A. Leygue, M. Coret, J. Rethore, L. Stainier, E. Verron, Data driven constitutive identification, 2017, HAL ID: hal-01452492v2.
[60] L.T.K. Nguyen, M-A. Keip, A data-driven approach to nonlinear elasticity, Comput. Struct. 194 (2018) 97–115.
[61] R. Ibañez, D. Borzacchiello, J.V. Aguado, E. Abisset-Chavane, E. Cueto, P. Ladeveze, F. Chinesta, Data-driven non-linear elasticity: constitutive manifold construction and problem discretization, Comput. Mech. 60 (2017) 813–826.
[62] R. Ibañez, E. Abisset-Chavanne, J.V. Aguado, D. Gonzalez, E. Cueto, F. Chinesta, A manifold-based methodological approach to data-driven computational elasticity and inelasticity, Arch. Comput. Methods Eng. 25 (1) (2018) 47–57.
[63] D. González, F. Chinesta, E. Cueto, Thermodynamically consistent data-driven computational mechanics, Contin. Mech. Thermodyn. 31 (1) (2019) 239–253, https://doi.org/10.1007/s00161-018-0677-z.
[64] T. Kirchdoerfer, M. Ortiz, Data driven computing with noisy material data sets, Comput. Methods Appl. Mech. Eng. 326 (2017) 622–641.
[65] S.F. Gull, J. Skilling, Maximum entropy method in image processing, IEE Proc. F, Commun. Radar Signal Process. 131 (1984) 646–659.
[66] H. Wang, J.F. O'Brien, R. Ramammoorthi, Data-driven elastic models for cloth: modeling and measurement, ACM Trans. Graph. 30 (4) (2011) 71.
[67] R. Ibañez, E. Abisset-Chavanne, D. Gonzalez, J.L. Duval, E. Cueto, F. Chinesta, Hybrid constitutive modeling: data-driven learning of corrections to plasticity models, Int. J. Mater. Form. 12 (4) (2019) 717–725, https://doi.org/10.1007/s12289-018-1448-x.
[68] Y. Feng, S. Mitran, Data-driven reduced-order model of microtubule mechanics, Cytoskeleton 75 (2018) 45–60.
[69] J. Ku, J.L. Hicks, T. Hastie, J. Leskivec, C. Re, S.L. Delp, The mobilize center: an NIH big data to knowledge center to advance human movement research and improve mobility, J. Amer. Med. Inform. Assoc. 22 (6) (2015) 1120–1125.
[70] E. Cai, P. Ranjan, P. Pan, M. Wuebbens, D. Marculescu, Efficient data-driven model learning for dynamical systems, in: Proceedings of the Intelligent Systems for Molecular Biology Meeting, ISMB 2016, Orlando, July 8-12, 2016, 2016.
[71] J. Hasenauer, N. Jagiella, S. Hross, F.J. Theis, Data-driven modeling of biological multi-scale processes, J. Coupled Syst. Multiscale Dyn. 3 (2) (2015) 101–121, https://doi.org/10.1166/jcsmd.2015.1069.
[72] B. Chen, C. Zang, Artificial immune pattern recognition for structure damage classification, Comput. Struct. 87 (2013) 1394–1407.
[73] S.A. Hofmeyr, S. Forrest, Architecture for an artificial immune system, Evol. Comput. 8 (4) (2000) 443–473.
[74] J.E. Hunt, D.E. Cooke, Learning using an artificial immune system, J. Netw. Comput. Appl. 19 (2) (1996) 189–212.

[75] M. Anaya, D.A. Tibaduiza-Burgos, F. Pozo, A bioinspired methodology based on an artificial immune system for damage detection in structural health monitoring, Shock Vib. 2015 (2014) 648097.
[76] M. Anaya, D.A. Tibaduiza, F. Pozo, Detection and classification of structural changes using artificial immune systems and fuzzy clustering, Int. J. Bio-Inspir. Comput. 9 (1) (2017) 35–52.
[77] F.J.M. Guerra dos Santos Cavadas, Structural Health Monitoring of Bridges: Physics-Based Assessment and Data-Driven Damage Identification, Ph. D. thesis, Universidade do Porto, 2016.
[78] A.S. Kompalka, S. Reese, O.T. Bruhns, Experimental investigation of damage evolution by data-driven stochastic subspace identification and iterative finite element model updating, Arch. Appl. Mech. 77 (8) (2007) 559–573.
[79] A. Karpatne, G. Alturi, J.H. Faghmous, M. Steinbach, A. Banerjee, A. Ganguly, S. Shekhar, N. Samatova, V. Kumar, Theory-guided data science: a new paradigm for scientific discovery from data, IEEE Trans. Knowl. Data Eng. 29 (2017) 2318–2331.
[80] V.J. Amores, J.M. Benitez, F.J. Montáns, Data-driven, structure-based hyperelastic manifolds: a macro-micro-macro approach, arXiv:1903.11545 [cond-mat.mtrl-sci].
[81] S. Yin, S.X. Ding, X. Xie, H. Luo, A review on basic data-driven approaches for industrial process monitoring, IEEE Trans. Ind. Electron. 61 (11) (2014) 6418–6428.
[82] F. Chinesta, E. Cueto, E. Abisset, J.L. Duval, F. El Khaldi, Virtual, digital and hybrid twins. A new paradigm in data-based engineering and engineered data, Arch. Comput. Methods Eng. (2019), https://doi.org/10.1007/s11831-018-9301-4, in press.
[83] J. Wang, Q. Chang, G. Xiao, N. Wang, S. Li, Data driven production modeling and simulation of complex automobile general assembly plant, Comput. Ind. 62 (7) (2011) 765–775.
[84] P.P. Peralta-Yahya, F. Zhang, S.B. Del Cardayre, J.D. Keasling, Microbial engineering for the production of advanced biofuels, Nature 488 (7411) (2012) 320.
[85] P. Kadlec, B. Gabrys, S. Strandt, Data-driven soft sensors in the process industry, Comput. Chem. Eng. 33 (4) (2009) 795–814.
[86] S. Yin, S.X. Ding, A.H.A. Sari, H. Hao, Data-driven monitoring for stochastic systems and its application on batch process, Int. J. Syst. Sci. 44 (7) (2013) 1366–1376.
[87] S.A. Vaghefi, M.A. Jafari, J. Zhu, J. Brouwer, Y. Lu, A hybrid physics-based and data driven approach to optimal control of building cooling/heating systems, IEEE Trans. Autom. Sci. Eng. 13 (2) (2014) 600–610.
[88] K.R. Holdaway, Harness Oil and Gas Big Data with Analytics: Optimize Exploration and Production with Data-Driven Models, Wiley, New Jersey, 2014.
[89] S. Esmaili, S.D. Mohaghegh, Full field reservoir modeling of shale assets using advanced data-driven analytics, Geosci. Front. 7 (1) (2016) 11–20.
[90] Y. Zhang, J. He, C. Yang, J. Xie, R. Fitzmorris, X-H. Wen, A physics-based data-driven model for history matching, prediction, and characterization of unconventional reservoirs, Soc. Pet. Eng. J. 23 (4) (2018) SPE-191126-PA.
[91] Z. Guo, A.C. Reynolds, H. Zhao, A physics-based data-driven model for history matching, prediction, and characterization of waterflooding performance, Soc. Pet. Eng. J. (2018), https://doi.org/10.2118/182660-PA.
[92] T. Kaneko, R. Wada, M. Ozaki, T. Inoue, Combining physics-based and data-driven models for estimation of WOB during ultra-deep ocean drilling, in: ASME 2018 37th International Conference on Ocean, Offshore and Arctic Engineering, Madrid, Spain, 2018, Paper OMAE2018-78229, V008T11A007.
[93] I. Song, I.H. Cho, R.W. Wong, An advanced statistical approach to data-driven earthquake engineering, J. Earthq. Eng. (2018), https://doi.org/10.1080/13632469.2018.1461713, in press.
[94] K. Duraisamy, Z.J. Zhang, A.P. Singh, New approaches in turbulence and transition modeling using data-driven techniques, in: 53rd AIAA Aerospace Sciences Meeting, AIAA SciTech Forum AIAA2015-1284, 2015, pp. 1–14.
[95] E.J. Parish, K. Duraisamy, A paradigm for data-driven predictive modeling using field inversion and machine learning, J. Comput. Phys. 305 (2016) 758–774.
[96] J. Tompson, K. Schlachter, P. Sprechmann, K. Perlin, Accelerating fluid simulation with convolutional networks, arXiv:1607.03597, 2017.
[97] J. Ling, A. Ruiz, G. Lacaze, J. Oefelein, Uncertainty and data-driven model advances for a jet-in-crossflow, J. Turbomach. 139 (2) (2016) 021008.
[98] K. Duraisamy, G. Laccarino, H. Xiao, Turbulence modeling in the age of data, Annu. Rev. Fluid Mech. 51 (2019) 357–377, https://doi.org/10.1146/annurev-fluid-010518-040547.
[99] S. Klus, F. Nüske, P. Koltai, H. Wu, I. Kevrekidis, C. Schötte, F. Noé, Data-driven model reduction and transfer operator approximation, J. Nonlinear Sci. 28 (3) (2018) 985–1010.
[100] V. Beltran, S. Le Clainche, J.M. Vega, Temporal extrapolation of quasi-periodic solutions via DMD-like methods, in: 2018 Fluid Dynamics Conference, AIAA Aviation Forum (AIAA 2018-3092), 2018.
[101] N. Demo, M. Tezzele, G. Gustin, G. Lavini, G. Rozza, Shape optimization by means of proper orthogonal decomposition and dynamic mode decomposition, arXiv:1803.07368v2, 2018.
[102] B. Moya, D. González, I. Alfaro, F. Chinesta, E. Cueto, Learning slosh dynamics by means of data, Comput. Mech. 64 (2019) 511–523.
[103] K.H. Lee, M.G. Choi, Q. Hong, J. Lee, Group behavior from video: a data-driven approach to crowd simulation, in: SCA '07: Proceedings of the 2007 ACM SIGGRAPH/Eurographics Symposium on Computer Animation, 2007, pp. 109–118.
[104] P. Charalambous, Y. Chrysanthou, The PAG crowd: a graph based approach for efficient data-driven crowd simulation, Comput. Graph. Forum 33 (8) (2014) 95–108.
[105] J. Zhang, H. Liu, Y. Li, X. Qin, S. Wang, Video-driven group behavior simulation based on social comparison theory, Physica A Statist. Mech. Appl. 512 (2018) 620–634.
[106] V. Marmarelis, G. Mitsis, Data-Driven Modeling for Diabetes, Lecture Notes in Bioengineering, Springer-Verlag, Berlin, 2014.
[107] O. Zettinig, T. Mansi, D. Neumann, B. Georgescu, S. Rapaka, P. Seegerer, E. Kayvanpour, F. Sedaghat-Hamedani, A. Amr, J. Haas, H. Steen, Data-driven estimation of cardiac electrical diffusivity from 12-lead ECG signals, Med. Image Anal. 18 (8) (2014) 1361–1376.
[108] J.R. Jimenez-Alanaiz, V. Medina-Banuelos, O. Yanez-Suarez, Data-driven brain MRI segmentation supported on edge confidence and a priori tissue information, IEEE Trans. Med. Imaging 25 (1) (2006) 74–83.
[109] J.P. Kassirer, Diagnostic reasoning, Ann. Intern. Med. 110 (11) (1989) 893–900.
[110] X.H. Wang, B. Zheng, W.F. Good, J.L. King, Y.H. Chang, Computer-assisted diagnosis of breast cancer using a data-driven Bayesian belief network, Int. J. Med. Inform. 54 (2) (1999) 115–126.
[111] Y. Xing, W. Wang, Z. Zhao, Combination data mining methods with new medical data to predicting outcome of coronary heart disease, in: International Conference on Convergence Information Technology, Gyeongju, South Korea, 21-23 Nov 2007, IEEE, 2007.
[112] J. Luo, M. Wu, D. Gopukumar, Y. Zhao, Big data application in biomedical research and health care: a literature review, Biomed. Inform. Insights 8 (2016) 1–10.
[113] K.H. Bleicher, H.J. Böhm, K. Müller, A.I. Alanine, Hit and lead generation: beyond high-throughput screening, Nat. Rev. Drug Discov. 2 (2003) 369–378.
[114] Y. Jing, Y. Bian, Z. Hu, L. Wang, X-Q.S. Xie, Deep learning for drug design: an artificial intelligence paradigm for drug discovery in the big data era, AAPS J. 20 (2018) 58.
[115] J. Dews, Strategic trends in the drug industry, Drug Discov. Today 8 (2003) 411–420.
[116] N. Kumar, B.S. Hendriks, K.A. Janes, D. de Graaf, D.A. Lauffenburger, Applying computational modeling to drug discovery and development, Drug Discov. Today 11 (17–18) (2006) 806–811.

[117] H.P. Fisher, S. Heyse, From targets to leads: the importance of advanced data analysis for decision support in drug discovery, Curr. Opin. Drug Discov. Devel. 8 (3) (2005) 334–346.
[118] D.B. Searls, Data integration: challenges for drug discovery, Nat. Rev. Drug Discov. 4 (2005) 45–58.
[119] T. Wunberg, M. Hendrix, A. Hillisch, M. Lobell, H. Meier, C. Schmeck, H. Wild, B. Hinzen, Improving the hit-to-lead process: data-driven assessment of drug-like and lead-like screening hits, Drug Discov. Today 11 (2006) 175–180.
[120] C. Partl, A. Lex, M. Streit, H. Strobelt, A.-M. Wassermann, H. Pfister, D. Schmalstieg, ConTour: data-driven exploration of multi-relational datasets for drug discovery, IEEE Trans. Vis. Comput. Graph. 20 (12) (2014) 1883–1892.
[121] T.J. Howe, G. Mahieu, P. Marichal, T. Tabruyn, P. Vugts, Data reduction and representation in drug discovery, Drug Discov. Today 12 (1–2) (2007) 45–53.
[122] R. Gómez-Bombarelli, J.N. Wei, D. Duvenaud, J.M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T.D. Hirzel, R.P. Adams, A. Aspuru-Guzik, Automatic chemical design using a data-driven continuous representation of molecules, ACS Cent. Sci. 4 (2) (2018) 268–276.
[123] M.J. Kusner, B. Paige, J.M. Hernández-Lobato, Grammar variational autoencoder, arXiv:1703.01925 [stat], 2017.
[124] W. Jin, R. Barzilay, T. Jaakkola, Junction tree variational autoencoder for molecular graph generation, arXiv:1802.04364, 2018.
[125] H. Dai, Y. Tian, B. Dai, S. Skiena, L. Song, Syntax-directed variational autoencoder for structured data, arXiv:1802.08786, 2018.
[126] K. Popova, O. Isayev, A. Tropsha, Deep reinforcement learning for de novo drug design, Sci. Adv. 4 (7) (2018) eaap7885.
[127] G.L. Guimaraes, B. Sánchez-Lengeling, P.L.C. Farias, A. Aspuru-Guzik, C. Outeiral, P.L.C. Farias, A. Aspuru-Guzik, Objective-reinforced generative adversarial networks (ORGAN) for sequence generation models, arXiv:1705.10843, 2017.
[128] J. Nicas, J. Creswell, Boeing's 737 Max: 1960s Design, 1990s Computing Power and Paper Manuals, The New York Times, NY Edition, 9 Apr 2019, A1, 2019.
[129] J. Bicevskis, Z. Bicevska, G. Karnitis, Executable data quality models, Proc. Comput. Sci. 104 (2017) 138–145.
[130] M.S. Marev, E. Compatangelo, W. Vasconcelos, Towards a Context-Dependent Numerical Data Quality Evaluation Framework, Technical report, 2018, arXiv:1810.09399.
[131] R.Y. Wang, V.C. Storey, C.P. Firth, A framework for analysis of data quality research, IEEE Trans. Knowl. Data Eng. 7 (4) (1995) 623–639.
[132] J. Liu, J. Li, W. Li, J. Wu, Rethinking big data: a review on the data quality and usage issues, ISPRS J. Photogramm. Remote Sens. 115 (2016) 134–142.
[133] A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehman, S. Auer, Quality assessment for linked data: a survey, Semant. Web 7 (1) (2019) 63–93.
[134] C. Batini, M. Scannapieco, Data and Information Quality. Dimensions, Principles and Techniques, Springer Nature Switzerland AG, 2018.
[135] I. Kirchen, D. Schütz, J. Folmer, B. Vogel-Heuser, Metrics for the evaluation of data quality of signal data in industrial processes, in: IEEE 15th International Conference on Industrial Informatics, INDIN, 24-26 July 2017, 2017.
[136] G. Pastorello, D. Agarwal, T. Samak, C. Poindexter, B. Faybishenko, D. Gunter, R. Hollowgrass, D. Papale, C. Trotta, A. Ribeca, E. Canfora, Observational data patterns for time series data quality assessment, in: Proceedings of the IEEE 10th International Conference on eScience, vol. 1, IEEE Computer Society, 2014, pp. 271–277.
[137] R. Gitzel, Data quality in time series data. An experience report, in: Proceedings of CBI 2016 Industrial Track, Paris, France, 31 Aug 2016, 2016, pp. 41–49.
[138] Y. Chen, F. Zhu, J. Lee, Data quality evaluation and improvement for prognostic modeling using visual assessment based data partitioning method, Comput. Ind. 64 (2013) 214–225.
[139] T.Y. Hou, K.C. Lam, P. Zhang, S. Zhang, Solving Bayesian inverse problems from the perspective of deep generative networks, Comput. Mech. 64 (2019) 395–408.
[140] M. Raissi, H. Babaee, G.E. Karniadakis, Parametric Gaussian process regression for big data, Comput. Mech. 64 (2019) 409–416.
[141] Y. Yang, P. Perdikaris, Conditional deep surrogate models for stochastic, high dimensional, and multifidelity systems, Comput. Mech. 64 (2019) 417–434.
[142] C. Soize, R. Ghanem, Data-driven probability concentration and sampling on manifold, J. Comput. Phys. 321 (2016) 242–258.
[143] C. Soize, C. Farhat, Probabilistic learning form modeling and quantifying model-form uncertainties in nonlinear computational mechanics, Int. J. Numer. Methods Eng. 117 (2019) 819–843.
[144] S. Fortunato, C.T. Bergstrom, K. Börner, J.A. Evans, D. Helbing, S. Milojevic, A.M. Petersen, F. Radicchi, R. Sinatra, B. Uzzi, A. Vespignani, L. Waltman, D. Wang, A-L. Barabasi, Science of science, Science 359 (6379) (2018) eaao0185, https://doi.org/10.1126/science.aao0185.
[145] A. Clauset, D.B. Larremere, R. Sinatra, Data-driven predictions in the science of science, Science 355 (6324) (2017) 477–480.