Chinesta and Cueto *Advanced Modeling and Simulation in Engineering Sciences(2022)9:21*
https://doi.org/10.1186/s40323-022-00234-8

Advanced Modeling and Simulation in Engineering Sciences

REVIEW Open Access

Empowering engineering with data, machine learning and artificial intelligence: a short introductive review

Francisco Chinesta[1*] and Elias Cueto[2]

*Correspondence: Francisco.Chinesta@ensam.eu
[1]PIMM Lab & ESI Chair, Arts et Metiers Institute of Technology, Paris, France
[2]I3A, Universidad de Zaragoza, Zaragoza, Spain

Abstract

Simulation-based engineering has been a major protagonist of the technology of the last century. However, models based on well established physics fail sometimes to describe the observed reality. They often exhibit noticeable differences between physics-based model predictions and measurements. This difference is due to several reasons: practical (uncertainty and variability of the parameters involved in the models) and epistemic (the models themselves are in many cases a crude approximation of a rich reality). On the other side, approaching the reality from experimental data represents a valuable approach because of its generality. However, this approach embraces many difficulties: model and experimental variability; the need of a large number of measurements to accurately represent rich solutions (extremely nonlinear or fluctuating), the associate cost and technical difficulties to perform them; and finally, the difficulty to explain and certify, both constituting key aspects in most engineering applications. This work overviews some of the most remarkable progress in the field in recent years.

Keywords: Data-driven learning, Physics-informed learning, Physics-augmented learning, Machine learning, Artificial intelligence

Introduction

Engineering sciences have acquired a proved maturity in what concerns to modeling, simulation and experiments, the three so-called pillars of engineering. These have enabled an unprecedented technological development in almost every technology domain: space, transport and mobility, energy, machinery, civil and industrial infrastructures, smart industry, smart cities and nation, among others.

Existing models are a heritage of centuries of fruitful science. Once calibrated and validated, they exhibit both accuracy and robustness, even in the presence of uncertainty. Models developed under the impulsion of experimental techniques enable the access to the smallest scales, with increasing accuracy and efficiency, for observing, measuring and interacting with the deepest details of the surrounding reality.

Models also benefit from the unprecedented advances in applied mathematics and computer science. Thus, a fast and accurate solution of complex mathematical models was

possible, even when components, structures and systems exhibit extremely large uncertainty and complex couplings. Thus, the design (under multi-disciplinary optimality constrains), the diagnosis, the prognosis and predictive engineering are nowadays part of the everyday engineering practice—the engineer's mission!.

The twenty-first century started with revisited challenges: the dream or need of managing systems of increasing size and complexity, that involve the finest scales, exhibit uncertainty, variability and fluctuating behaviors. But new objectives also exist, such as the need of developing digital twins of the physical entities. These twins should be able to emulate their behavior and, consequently, enable an efficient dialog between the user or the controller and the digital twin, instead of making that dialog with the real entity.

A digital twin (DT) aims at representing the reality in a complete, concise, accurate and efficient way. It is more than a continuously calibrated physics-based model (something that is achieved through adequate data assimilation) and also much more than a simple transfer function, that relates specific inputs with specific outputs. Even if they are very precise and operate very fast, they fail to be general enough to address any query. DT comprise three main functionalities: (i) an accurate model, able to replicate or emulate the reality with the required level of fidelity; (ii) a digital platform for processing that model and making predictions from it at the required feedback rate; and (iii) data with multiple missions: calibrating physics-based models, learning data-driven-models, making diagnosis, validating predictions, ...

Advances in sensing for data acquisition, data transfer, data storage, data analytics, ... have facilitated and boosted the irruption of DT in almost all the scientific and technological domains [24,40,67,123]. Now, we must conciliate accuracy with velocity, make it fast and well—the engineer's dream!

In this context, we should address a first question:

- What are the limits or difficulties we are confronted to in the current engineering practice?

with a natural second one:

- which framework is needed, such that it is able to conciliate explainability and understanding (the foundations of knowledge), efficiency and pragmatism, needs and resources?

for finally

- identifying the available methodologies,

and

- illustrating their use in the domain of materials and structures, fluids and flows, processing and multi-physics coupling, complex systems and systems of systems.

Coming back to the first question, it can be noticed that several models exhibit a limited fidelity with respect to the reality that they are expected to represent. This is due to different reasons of practical and/or epistemic nature. In fact, reality seems to be much richer that our approach to it, our conceptualization of the physics. To enrich it, thus enhancing their predictability capabilities, further research works are needed. Some of

them will run for many years to reach the targeted accuracy. This is the way how science and technology advanced during the past centuries.

Nothing seems to be really new from the fundamental point of view at present. Maybe only the fact that our society, to which scientists and engineers belong, becomes more and more impatient, and queries are expected to be responded instantaneously. An insatiable impatience seems to be the only novelty with respect to the process of transforming observations into knowledge accomplished during centuries of successful science and technology.

From now on, three main challenging scenarios are addressed:

- Improving the efficiency of optimal design procedures by enabling fast, accurate and a complete exploration of the design space—that is, the space spanned by the design parameters—by improving existing techniques and proposing new advanced methodologies.
- Diagnosis works quite well based solely on data analysis. As soon as data, as described later, is classified so as to construct a sort of catalog or dictionary of faults, pattern recognition will be enough for performing online diagnosis. The prescription of corrective actions, needed for accurate prognosis, is also discussed later.
- Finally, models exist and are accurate enough. However, they are difficult to manipulate with the prescribed accuracy, under stringent real-time constraints. Very often, model order reduction (MOR) techniques do not suffice or are too complex to implement. To improve performance models thus are degraded (coarsened). In that case, the risk becomes that of making wrong predictions very fast. Efficient technologies are then needed to ensure fast an accurate predictions.

The transition between the twentieth and twenty-first centuries was accompanied of a hatching of technologies able to address the just referred challenges, conciliating efficiency and accuracy:

- First, the solution of state-of-the-art physics-based models has been accelerated so as to accomplish real-time response by using advanced MOR techniques. These techniques neither reduce nor modify the model itself, they simply reduce the complexity of its solution by employing more adapted approximations of the unknown fields [23].

  Model Order Reduction techniques express the solution of a given problem (usually governed by a PDE) by employing a reduced basis with strong physical or mathematical content. Sometimes these bases are extracted from some offline solutions of the problem at hand, as in the Proper Orthogonal Decomposition (POD) or the Reduced Basis (RB) methods. When operating within the reduced basis approach, the complexity of the solution scales with the size of this basis. It is, in general, much smaller than the size of the general-purpose approximation basis employed in the Finite Element Method (FEM), whose size scales with the number of nodes involved in the mesh that discretizes the domain in which the problem is defined. Even if the use of a reduced basis implies a certain loss of generality, it enables impressive computing-time savings while guaranteeing acceptable accuracy. This is true, in general, as soon as the problem solution continues to live in the space spanned by the reduced basis. The main drawbacks of these techniques are: (i) their limited generality when addressing situations far from the ones employed to construct the reduced basis; (ii) the diffi-

culties of addressing nonlinear problems, that require the use of advanced strategies; and (iii) their intrusive character with respect to their use in well-experienced and validated software.

For circumventing, or at least alleviating, the just referred computational issues, an appealing route consists of constructing the reduced basis at the same time that the problem is solved, as Proper Generalized Decompositions (PGD) do [20–22]. However, PGD is even more intrusive than POD and RB. Thus, non-intrusive PGD procedures have been proposed recently that construct the parametric solution of the problem from a number of high-fidelity solutions. These are obtained offline for different choices of the model parameters. Among these techniques we can mention the SSL-PGD, that considers hierarchical separated bases for interpolating the precomputed solutions [9], or its sparse counterpart [58,110].

Once the parametric solution of the problem at hand is available, it can be particularized online for any choice of the model parameters, enabling simulation, optimization, inverse analysis, uncertainty propagation, simulation-based control, ... all of them under stringent real-time constraints [23].

- With the democratization of technologies around sensing, metrology, communication, storage and computation, massive data-acquisition experienced an exponential increase in many new domains. In other domains like spectrography, tomography, thermography, among others, massive data has already been present for many years. Data were not the real revolution. The real revolution, as just argued, was their democratization and the colonization of many domains of science and technology. This happened first where models were inexistent, or where models existed but their predictions did not fulfill the desired expectations in terms of accuracy or speed of computation.
- Thus, many initiatives irrupted worldwide, pointing out the opportunities that the use of machine learning and artificial intelligence could represent [75]. In particular, in France the collective report entitled *AI for Humanity*, whose writing was leaded by the 2010 Fields Medal awardee Cedric Villani [127], was the starting point, followed with the four french 3AI initiatives: AI focusing on medicine, mobility, aeronautics and space, environment, city and nation, human and social sciences, ... Similar activities saw the light in all the countries around the world, almost simultaneously.

  Engineering benefited of that impulse, with different approaches to the employ of the data:

  1. Some applications were essentially or purely based on the employ of data. This is the case of applications making use on pattern recognition (based on data classification), widely employed in diagnosis.
  2. Dynamic Data Driven Application Systems (DDDAS) represent an intimate dialog between models and data, where data is used for keeping the models calibrated continuously, whereas the models serve to drive the data collection [27].
  3. In the third case, data is used with learning purposes, to be employed to make predictions or anticipate anomalous behaviors.

  In the third item listed above, three levels of learning, depending on the relative weight of physics and data, can be distinguished:

1. The first learning approach considers essentially data, that is assumed to be constrained by first principles and their associated variational formulations only. This is the most genuine *data-driven learning*.
2. In the second learning approach, physics (and more generally, all the existing knowledge) is assimilated during the learning process, giving rise to the so-called *physics-informed learning*, whose popularity is growing very rapidly.
3. In the third approach, more than solely informing the learning process, physics augments the learning. For example, calibrated physics-based models are enriched with a data-driven model (eventually physics-informed) for representing the deviation between the observed reality and the predictions obtained from the calibrated physics-based model itself. This third framework, a sort of transfer learning, is called here *physics-augmented learning*.

The two remaining main protagonists are data themself, as well as the techniques for quantifying the learned model confidence:

1. What data?; at which scale?; where and when? which metrics should we consider for quantifying, expressing and comparing them? how transform the data into knowledge, or how extract the last from the former?
2. How to verify and validate the learned models?; how to explain and certify? Engineering remains to be extremely dependent on the confidence to quantity risks, that must be mastered in the best manner.

As mentioned before repeatedly, because of fundamental or practical reasons the considered models (when they exist) do not allow to attain the required accuracy or rapidity in their predictions. Thus, techniques based on the use of data (solely, informed or augmented by the physics) are becoming appealing alternatives for replacing, enriching or augmenting the existing models. To empower engineering and, therefore, conciliate accuracy and velocity from the smart use of physics and data, advanced methodologies focusing on data, learning and verification and validation (V&V) must be used, adapted or proposed.

The next sections revisit the just referred three topics: (i) data; (ii) learning; and (iii) V&V.

Methodologies: data

Our staring point is Data, where the upper case "D" is voluntarily used to emphasize the impressive richness that this word embraces, as discussed in the current section.

Data and metrics

Data have a double nature: a qualitative essence and a quantitative extension. First, data exist with respect to a given target, that is, with respect to a given objective. In this sense, data become goal-oriented. Then, to quantify them, an appropriate metric is needed. We can visualize such a need by thinking on a ruler to measure the position of an object in the space, or its dimensions; a balance for measuring its weight; or a thermometer for measuring its temperature.

The situation becomes more complex when data cannot be directly represented in a vector space. For example, data representing a manufactured product, could consist of a

sort of identity card comprising the list of constituents, their percentages, the different processing parameters and even the name of the employee that produced it. All this information could be sorted into a list that will define the product's identity card. Two different products become represented by two distinct data. The important issue here is the metric to be used to compute the distance between both products. It seems obvious that calculating that distance between two products is much more controversial that computing the Euclidian distance between two points drawn on a paper.

Thus, data cannot exist without a metric for describing and manipulating them. When it is not defined a priori, this metric must be learned at the same time that the classification is carried out. It is the way that most of AI-based classification techniques proceed. These are considered later.

Data reduction and intrinsic dimensionality

With physics and its target both defined, an important issue concerns the features to be considered to infer the desired output. The optimal choice can be stated as follows: no more than the strictly needed, neither less than the required ones. This is easy to express but difficult to apply.

Removing useless features can be performed by using different techniques. Some of them are of a statistical nature (analysis of variance, ...), others are based on the dimensionality of the manifold in which the multidimensional data (assumed to be expressible in a vector space) is embedded. In the last case, we cite here linear (e.g. Principal Component Analysis (PCA) [79]) or nonlinear dimensionality reduction. Among them, Locally Linear Embedding (LLE) [107], kernel Principal Component Analysis (kPCA) [79], local Principal Component Analysis (ℓ-PCA), Multidimensional Scaling (MDS) [79], t-distributed Stochastic Neighbor Embedding (t-SNE) [87], among others.

The just referred strategies are also known as manifold learning techniques. All of them aim at removing linear and nonlinear correlations and then, approximating the intrinsic dimension of the data embedding manifold. There is not a universal technique for performing nonlinear dimensionality reduction and calculating the exact intrinsic dimensionality of data. All of them involve a series of hyper-parameters, assumptions, hypotheses and sometimes technical choices (e.g. the so-called kernel trick in the kPCA). All of them work well with data expressed in vector spaces. However, different variants or alternatives exist for addressing more complex data, involving categorial features, sometimes incomplete, enabling the discrimination between useful and useless features [31,32,56]. Autoencoders (AE), a particular form of Neural Network (NN) architecture, represent an alternative route to nonlinear dimensionality reduction [45,51,113].

Combined features

Sometimes, all the features retained for the explanation of a certain target reveal to be useful. However, they act in a combined manner. Imagine for a while, the velocity, density and viscosity of a fluid in a flow. These features can be measured (different devices exist for this). However, the flow is solely described from a feature that combines the former three: the Reynolds number. However, it can not be measured directly, but indirectly obtained from the three just referred measurable features.

Discovering and extracting these combined hyper-features could be of great relevance from the modeling viewpoint, and they are of major relevance when accessing to knowledge. However, its explicit identification is not an easy and direct task. Some techniques perform the task, but in a *black-box* sense, as it is the case when using the previously introduced autoenccoders.

Time series

Time series, representing the system response at equivalent conditions, usually differ if they are compared from their respective values at each time instant. That is, two time series, even when they describe the same system in similar conditions, never match perfectly. Thus, they differ even if they resemble in a certain metric that should be learned. For example, our electrocardiogram measured during two consecutive minutes will exhibit a resemblance, but certainly both of them are not identical, thus making a perfect match impossible. A small variation will create a misalignment needing for metrics less sensible to these effects. The same rationale applies when comparing two profiles of a rough surface, two images of a foam taken in two close locations, ... they exhibit a resemblance even if they do not match perfectly.

Thus, techniques aiming at aligning data were proposed. In the case of time-series, Dynamic Time Warping, DTW [94,115] has been successfully applied in many domains. The theory of optimal transport arose as a response to similar issues [126]. Another alternative consists of renouncing to *align* the data, and focussing on extracting the adequate, goal-oriented descriptors of these complex data, enabling comparison, clustering, classification and modeling (from non-linear regressions) [81].

Finally, data transformation can be performed to represent them in a more compact way in a more appropriate space. This is done, for instance, in the Fourier or wavelet transform, or the one based on persistence homology, the so-called Topological Data Analysis (TDA) revisited in the next section.

Data representation

As just discussed, sometimes features are associated with observable and measurable quantities. However, these features—very pertinent from the point of view of the technician—are much less pertinent from the point of view of the modeler. It is important to note that the more features are considered, the higher is the volume of data needed to accomplish learning tasks.

Sometimes, the complexity of a learned model depends on the chosen observables. For describing the solar system (aiming at modeling mechanistically its movement), one can make use of a metric for locating the planets with respect to the Earth (Ptolemy) or the sun (Kepler) at each time instant. Both are valid (no absolute frame of reference exists), but the former leads to a much more complex model than the later. Thus, the complexity of a model strongly and intimately depends on the description chosen for the data.

Certain models seem to be complex when described in the so-called physical space (usually the Euclidean space and time). However, with alternative descriptions, data become sparser and then more suitable for modeling purposes. Imagine for a while a periodic sinus function, whose description needs a certain number of data (dictated by the Shanon and Nyquist representation theorem). Its description becomes simpler and more compact as

soon as it is expressed in the Fourier space. Many representation spaces exist (wavelets, Fourier, DCT, ...) and their choice is part of the whole problem solution. This rationale represents the foundations of compressed sensing (the interested reader can refer to [60] and the references therein).

As previously mentioned, the complexity of the learned model depends on the data description. An adequate manipulation of the data could reduce the nonlinearity of the learned model (as discussed above in the Ptolemy versus Kepler confrontation). When looking for higher linearity, sometimes the dimensionality increases, as is the case of the kPCA. It does not need to make the mapping explicit because the only need when applying the PCA in the high-dimensional space is having a scalar product. Here, the Mercer theorem allows us to compute the scalar product in the intermediate space from the one calculated at the original space by means of the so-called kernel trick [79].

Other valuable transformation lies in the use of Topological Data Analysis (TDA), based on the persistence homology. It becomes an appealing alternative for describing data with large topology content [14,98]. This is the case of time series or images of microstructures (foams, polycrystals, composite materials, ...) [33–35,137]. TDA offers compact and concise metrics able to discriminate complex data from its intrinsic topology. Other possibility for addressing such a complex data consists of extracting some valuable statistical descriptors (statistical moments, pair-correlation, covariogram, ... [122]) on which applying usual learning strategies discussed later [137].

Last but not least, even if most data sets accept a valuable representation in the form of a list—this is very common in machine learning—there are more appropriate representations that take into account neighborhood, invariances, etc. Images are then easily convoluted and graphs decomposed in segments and vertex on which the learning procedures efficiently apply, as discussed later.

Quality and quantity

After proving that data can not be dissociated from the envisaged goal, and that their quantity is a relative concept that depends on many facts and considerations, to complete the picture the complexity of the learned model becomes strongly dependent on all the just referred choices.

Artificial intelligence (AI) is usually associated to big data. However, in engineering sciences; *smart* (or *useful*) is preferred to *big*. Devices for performing measurements with the required volume, accuracy and acquisition rate, are expensive. Performing measurements becomes very often technically complex (this is related to device placement, for instance), with many other difficulties related to data transfer (mainly in the case of remote sensing), data-storage, data-treatment, ... This will force us to have needing big enough computational infrastructures. It is also important to note that the optimal datum at the optimal location remains sometimes unattainable, because of technical issues or even because of security or safety regulations.

It is at that point that the data/knowledge couple becomes a very convenient option. Until now, we emphasized the fact that data serves to create knowledge, or for enhancing the existing one. Now, we are pointing out that existing knowledge allows data to become smarter. For that purpose, imagine for a while that we are interested in knowing the temperature at Paris on the first of January, 2022. Based on our accumulated experience

(knowledge or simply common sense) we can place a thermometer at the Sorbonne square and register the temperature early in the morning, at noon and at mid-night. Thus, one thermometer and 3 measurements are enough, whereas in absence on any experience or knowledge, one could be tempted to put hundreds of thermometers at each street and register the temperature each millisecond, to finally realize that most of this big data are simply useless. Thus, in engineering applications the choice is obvious: *smart* instead of *big*, even if in many cases (e.g. tomography, thermography, DLV, ...) the smartest data remains usually extremely big!

The hybrid framework (something which we call physics-augmented learning later on), looks for a data-driven enrichment of the physics-based model. The nonlinearity is in general much smaller (most of the nonlinearity is expected to be already explained by the physics-based model), so that it needs much less data. This is a real added-value of that hybrid framework.

Another way of reducing the amount of measurements consists of discovering the best locations and time instants where to perform the measurements. These techniques are grouped under the name of *active learning* [116]. In them, the existing or acquired knowledge is employed to drive the data acquisition procedure. Different techniques exist, based on the pre-existing knowledge (as in transfer learning), statistical sensing (with their foundation in the Bayes' theorem), some ones inspired from robotics (e.g., SLAM), stochastic learning or those based on information theory and the associated key concept of entropy (in the sense of the information theory). All of them aim at measuring the minimum amount and most relevant data.

When focusing on quality, the main aim is not having deterministic data by removing most of its noise, or pushing the noise to the lower scales, by significantly improving the measurement devices with the associated cost. Waiting for making it better, the data variability can be addressed by using standard or advanced filters, widely employed in data assimilation (e.g., Kalman and extended Kalman filters), or by taking into account the data variability within more adapted stochastic learning settings. The data noise is not the most formidable enemy, the most dangerous is bias. Outliers can be more easily identified and their impact limited by simply removing them or by using techniques more robust to their presence (e.g.,the L1 norm).

Grouping and classifying

Data is usually grouped, expecting that a datum belonging to a group shares some property with the the rest of the group's members. If European citizens are grouped by nationality, as soon as we identify a Spanish citizen we could assume that he speaks Spanish.

The concept is quite simple. However, organizing data in groups in a supervised or unsupervised manner is much more technical because of the need of using a metric to compare data or to evaluate data proximity.

Since data clustering (unsupervised) or classification (supervised) entail a learning process, in a certain sense, both techniques will be described and discussed in the next section.

Data augmentation and completion

When data is not abundant enough, data augmentation techniques can be applied. If some knowledge exists (reduced basis or data manifold) completing or augmenting data can be

performed easily. Interpolating data in complex nonlinear manifolds remains to be quite a technical issue [99]. Data can be augmented by generating extra data by using symmetry considerations or other kind of physics-based knowledge.

Another valuable black-box approach proceeds by combining a data generator and a data discriminator, with one trying to betray the other. This is the rationale behind the so-called Generative Adversarial Networks (GAN) [128], where the data generator allows augmenting the learning process.

Methodologies: learning

With data and all their just mentioned richness at hand, everything seems ready to learn models that relate the features with the target. In what follows, we address three learning modalities, depending on the relative weight of preexisting knowledge (physics) and data: (i) data-driven learning; (ii) physics-informed learning; and (iii) physics-augmented learning.

Data-driven learning

Data-driven learning is mainly based on the use of data. Different techniques are revisited in what follows.

Clustering and classification

Unsupervised clustering proceeds by grouping the data depending on their relative distance. For that purpose a metric must be available. When considering the k-means technique [88,89], and once the number of groups is predefined, the datum is grouped in such a manner that a datum belonging to a certain group, remains closer to the group's centre of gravity than to the center of gravity of any other group.

When the *ruler* does not exist (e.g., the before mentioned identity cards of manufactured products), the distribution of the data in groups must follow a certain criterion. For example, the class to which it belongs. However, to be useful, the border delimiting the different groups must be defined, like the borders of the different European counties. These borders are the ones separating data in the most robust way, in the sense of maximizing the distance from the data to the borders. With the purpose of obtaining the best classification, these separators can be linear or nonlinear, like the nonlinear borders of the European countries. Now that the map of Europe is defined, if one individual is found in Spain, one could assume (with a certain risk) that he or she speaks Spanish. This is reasonable from a probabilistic point of view, but there exists a non-zero probability that our individual is simply a tourist than does not speak a single word in Spanish! Obviously, more safe classification exists for inferring the spoken language.

Numerous techniques for data classification exist: Code to Vector (C2V) [2], Support Vector Machine (SVM) [26], Decision Trees (DT) [72] or its random forest counterpart [10], Neural Networks and Deep Neural Networks (DNN) [45], often convolutional (CNN) when addressing images [124], or Graph Neural Networks (GNN) [11] when applied on data structured on graphs.

For enhancing the classification performance, the so-called *boosting* procedures have been proposed and are nowadays widely and successfully employed [36,37]. Other techniques at mid-way between the supervised and unsupervised ones are proving their supe-

riority and then attracting more and more the interest of analysts, in particular the *semi-supervised* [140,141] and *self-supervised* techniques [78]. It is worth highlighting *reinforcement learning* procedures [66,97,119], that are becoming a major protagonist in AI.

When the target is quantitative and continuous, the resulting models consists of linear or nonlinear regressions addressed in the next sections.

Linear regression

If, for a while, we consider data (input and output) sorted in vectors, that for the sake of simplicity but without loss of generality, are assumed to be of the same size, the simplest model consists of simply looking for the linear application (a constant squared matrix) that, applied on the input data, results in the associated output data (or its inverse if one prefers). The rang of the resulting learned matrix depends on the intrinsic dimensionality of the manifold in which data is embedded. An incremental procedure for constructing it, inspired of the Dynamic Mode Decomposition (DMD, described later) was proposed in [105].

Nonlinear regressions

In the nonlinear case different possibilities exist. The simplest ones, based on polynomial approximations, become inefficient in two main situations: (i) when the polynomial degree increases, needed for describing for example non-polynomial nonlinearities; and (ii) when the number of features (model parameters) increases.

One possibility for addressing the multi-parametric case consists of using separated representations. These are at the heart of the so-called Proper Generalized Decomposition, PGD. That separated representation computes sequentially the approximation involving each parametric dimension (with the other dimensions frozen within an alternated directions fixed point algorithm). Thus, all the data is available to solve the problem in each parametric dimension, enabling the use of rich enough approximation bases. However, the problem considered globally, becomes under-determined, and in that case the number of solutions becomes undetermined (infinite). All them describe very accurately the data used in the regression construction (training data-set), but will provide very poor predictions (overfitting) outside the training dat-set (the so-called test data-set or at any other data point).

To avoid overfitting phenomena, different regularizations exist. Some consider adapted basis and their associated collocation points (for instance hierarchical orthogonal bases and their associated Gauss-Lobatto nodes in the case of the Sparse Subspace Learning (SSL) [9]). Others proceed by enriching the approximation sequentially while incrementing the polynomial degree [58]; or those making use of sparse regularizations. The so-called Sparse Identification of Nonlinear Dynamics (SINDy) regression [12] uses very rich approximation bases (by mixing polynomial with any other function expected contributing to the target) and then selecting the sparsest combination of the those functions for explaining the available data. Sparsity was combined with separated representations in our former works to conciliate multi-parametric settings with richness, small amounts of data, while circumventing overfitting [110]. In that paper, different regularizations (for instance Elastic-Net, Ridge, Lasso, ..., that give rise to the so-called rsPGD and s2PGD formulations) were compared. This is also the case for their combination with an anchored-ANOVA

formulation [121], employed for recovering the sensibility indicators that the analysis of variance provides, as well as to better address non-polynomial nonlinearities [73].

Even if the just discussed techniques perform quite well with mild nonlinearities, when the nonlinearities become intense and strongly non-polynomial, without an a priori knowledge on its character, these techniques fail to perform correctly in the scarce data limit. In that case the use of NN, and more particularly Deep Neural Networks (associated with so-called deep learning) [45] are the most appealing alternatives for addressing intense and general nonlinear behaviors. The universal approximation theorems, introduced for approximating functions [96] and then extended for approximating functionals and operators, explain the gain of popularity of these techniques, see [18,19] respectively. The gain in efficiency results at the cost of becoming the predictions less explainable and the necessity of higher volumes of data for performing the learning.

Other than the standard DNN (having several internal layers of neurons of different widths, both being network hyper-parameters), when the data to be manipulated consists of images, curves or graphs, CNN or GNN introduced before are preferable.

Dynamical systems

In the context of dynamical systems, one aims at learning the application that allows computing the state variables at a certain time instant, from the knowledge of the state at the previous time step. Dynamic Mode Decomposition, DMD, computes a matrix (linear model) for that purpose, that is, a matrix (the model) able to updating (in time) the system state from the current state [111]. The best constant matrix enabling the representation of all the available data is computed accordingly in general in a least-squares sense. Some constraints can be added during the learning process in order to ensure the stability of the resulting time integrator (related to the spectral radius of the matrix that is being learned) [109].

This rationale can be extended for considering nonlinear models, by assuming locally linear representations [108], as already employed in model order reduction techniques involving reduced bases. Other more general framework concerns the use of the Koopman operator theory [13,132].

Finally, the use of NN is also a valuable route, where the so-called residual NN (rNN) are being successfully employed for integrating nonlinear dynamical systems [100]. Nonlinear Autoregressive Exogenous NN, NARX, allow taking into consideration longer memory effects [8].

Miscellaneous

To finish this section, we would like to mention two additional learning scenarios. The first is the one related to the incomplete observation of the system state. The learned model must take this fact into account.

In computational mechanics, it is usual to perform static or dynamic condensation, when trying to express the internal (slave) degrees of freedom dependent on the master ones (here, the ones that are accessible for observation). It is easy to prove that in the static and linear case, a condensed model, relating input and outputs in the region under scrutiny can be defined and learned. Actions applying in the hidden region are accessible from their effects on the observed region. The transient case is a bit more technical, but

under certain conditions such a condensed model continues to exist [106], and in the most general case a series of internal variables can be defined and learned. The use of the so-called recurrent-NN [55] generalizes the just referred rationale and even extend them to nonlinear settings. Echo state network (ESN) improves recurrent neural networks. ESN has a non-trainable sparse connected recurrent part (dynamic reservoir) in a hidden layer whose weights are generated randomly and remain unchanged during the training [135].

The second scenario concerns the case in which different types of data must be integrated into the learning process. Multi-Modal Learning represents a very valuable route for defining efficient learning frameworks. The so-called Boltzmann machines (inspired from statistical mechanics) perform well in those settings [117].

Physics-informed learning

The learning procedures just considered were based almost exclusively on data. However, as soon as some knowledge exists, one is tempted to assimilate it in the learning process. This is at the origin of *transfer learning* techniques [131]. Knowledge on physics can sometimes be introduced in the NN-based learning process through the adequate definition of the so-called *loss* functions, whose minimization allows computing the weights of the different neuron connections.

Moving forward, one could imagine a regression describing the unknown field, whose space and time evolution is governed by a partial differential equation, PDE, subjected to the corresponding initial and boundary conditions. Thus, the PDE residual can be included into the loss function, as well as the residual in the fulfillment of the initial and boundary conditions. The learning process can be efficiently performed because of the possibility of taking the derivatives of a NN-based regression by using automatic differentiation [6]. This rationale is at the one present in the so-called Physics Informed Neural Networks, PINN, [102], extended to the discovery of operators in [86]. A self-consistent formulation was proposed in [129].

A variant consists of enforcing thermodynamic consistency. In the reversible framework, a regression is performed to find the particular form of the Hamiltonian, whose gradient in phase space results in the time derivative of the state variables. Thus, from data reporting the time evolution of the state, the free energy and the conservation operator (Hamiltonian) are learned, leading to a symplectic integrator. In the most general irreversible case, the free energy and the entropy, as well as the conservation and dissipation operators, are all them learned by subjecting them to some thermodynamic consistency constraints. These techniques are known as Thermodynamic Informed Neural Networks, TINN, or Structure Preserving NN, SPNN, with a rich recent literature [7,43,46,49,50,80,90,139].

In some cases the learning problem is formulated from the differential form of the GENERIC model [43]. However, variational formulations are also available, as the one of Herglotz (contact geometry) [125] or the one making use of the Onsager variational formulation that involves the so-called *Rayleighian* [54].

Thus, the PINN operates by replacing the usual finite element-based functional approximation by a NN-based regression, which is very general, efficient and robust for describing strongly nonlinear functions, its main advantage. The price to be paid is the necessity of solving nonlinear optimization problems even when solving a linear PDE. Its ther-

modynamic variant allows learning potentials (the trickiest issue arises when modeling thermo-mechanical systems) while incorporating any constraint (symmetries, ...).

Physics-augmented learning and hybrid modeling

Hybrid models consist of two contributions. It is the result of a physics-based model and the data-driven model, which takes into account the deviation between the measured physical reality and the physics-based model prediction [3,24,92,108,109]. The main advantages of the augmented framework is double. First, the possibility of explaining the (usually) most important part of the resulting hybrid (or augmented) model, the one concerning the physics-based contribution. Second, with the deviation much less nonlinear that the observed reality itself (the physics-based model contains an important part of such nonlinearity), less data suffices for constructing the data-driven model.

The practical implementation of this simple rationale faces to three main challenges:

1. The physics-based model must be calibrated (by assimilating the collected data) and solved at a feedback rate compatible with the evolution of the physical system, even faster if we want to anticipate future events. Model Order Reduction techniques—revisited in the Introduction section—are main protagonists in enabling real-time physics. These technologies have nowadays acquired at a proved maturity, and some of them have been integrated into commercial simulation softwares and computational platforms. Trying to minimize their intrusiveness, different minimally invasive methodologies were proposed. In general, these techniques operate by defining first a Design of Experiments, DoE, then, by computing a high-fidelity solution at the different points of that DoE, for, finally, constructing the surrogate (also known as metamodel or response surface) by using an appropriate regression (whose choice depends on the amount of available data, that at its turn depends on the cost of each high-fidelity resolution).
2. The data-driven model describing the difference between the measure and the prediction given by the physics-based model just described, must be created and interrogated in real-time.
3. Data are essential, as already mentioned, for calibrating the physics-based model and for learning the data-driven one. Both procedures could ask for different kind of data, collected in different locations and times, and having different natures (for instance, times series, images, ...) Accurate and fast data-assimilation techniques are also compulsory, and robust enough for addressing the data variability as previously discussed.

Methodologies: verification and validation

Making a decision based on a catalog (or dictionary), after recognizing a pattern, needs and adequate data classification, the construction of a dictionary and its enrichment, and also techniques for performing efficient searches inside. This process is the main one in diagnosis. On the other hand, making a prediction of the state of a system (prognosis) needs to make use of a model (physics-based, data-driven or hybrid).

However, the major issue in both actions, diagnosis and prognosis, concerns the level or degree of confidence. In many domains of engineering, this confidence is much more than

a simple added value. In engineering, most components, structures and systems must be certified before of being employed to fulfill regulations.

In the case of physics-based models, centuries of science with well stablished and well experienced models and solution procedures, enabled the requested confidence to ensure the functioning and the risks of a design or decision. However, when models are learned from data, and solely from data, many questions come into play: (i) the data considered in the learning process were the adequate and with the adequate quantity and quality, thus enabling to extract all the richness present in the physical phenomenon under study?; (ii) Was the sampling vast enough for covering all the functioning states? This second question is motivated by the difficulty of data-driven models to extrapolate far from the data that served to create the model (interpolation is safer than extrapolation).

Other than the previous points concerning the sources, other concern the learning process itself: was the chosen learning technique the most adequate? For instance, using a linear regression for modeling a nonlinear behavior does not seem the best choice. Moreover, many regression techniques involve a number of hyper-parameters to be finely tuned, again from the collected data.

Depending on the confidence granted to the learned model, its predictions could be used in an automatic way (high degree of confidence) or as simple suggestions offered to the decision-maker (lower level of confidence).

In the context of the simulation-based engineering, one of its key branches concerns the verification and validation (V&V). The former quantifies the error between the actual solution and the one produced numerically, and the last quantifies the agreement between the model solution and the real system behavior itself. In what concerns the verification, different error indicators and estimators, a priori and a posteriori, have been proposed. The last are based on the computed solution, and the former on the model properties (the ones associated to the differential operators and the considered approximations and discretization). Other than estimating the error, confidence intervals can be also derived and even more, with certified bounds.

However, verification and validation remain much less developed in what concerns learned models. There is a large variety of available learning methods, most of them operating in a black-box mode, involving a number of hyper-parameters, that without any a priori knowledge, must be themselves finely tuned from the available data to maximize the predictive performances. Thus, there are several sources of inaccuracy, among them: (i) the data, and in particular its alignment with respect to the goal, quantity, quality, ... (ii) the accuracy of the considered learning technique; (iii) the partition of the available data into the two data sets, the training data-set and the test data-set; (iv) the position of the evaluation point with respect to the position of the training data-points.

The response to all these questions seems compulsory for gaining confidence on the AI outcomes. To move beyond the usual performance indicators, based on the difference between the predictions and the collected data in both data-sets, the training and the test, the physics-informed and physics-augmented frameworks offer new possibilities of enhancing confidence.

Applications

This section revisits the use of data-driven techniques on different domains: materials and structures, fluids and flows, processes and couplings, and finally complex systems. For a more extensive review the interested reader can refer to [29] and the numerous references therein.

Data-driven materials and structures

In the context of materials, technologies able to access the finest scales of materials to perform observations and measurements, combined with the technologies for assimilating or learning from collected data, have experienced remarkable progresses. They enable to bridge the different scales in the description of materials. For that purpose, an efficient dialog between data and models, the last being based on the physics, on the data or having a hybrid nature, has been the key progress [95].

The trickiest issue was and continues to be data themself: what features to consider, how to represent them, how to represent the different chemical elements, the atomic structure and bonds, the macromolecules conformation with their topology and the crystallographic structure, the dislocations and other localized defects, ... how to assimilate that data into the models and how to construct data-driven models from them, among many other questions without a definitive (unique and general) response [1,4,42,103,112,130,134].

At the mesoscopic scale, far from the atoms and the finest description of the structure of the matter, but still far from the scale of the part, one is interested in describing the effective behavior of a representative volume of the considered solid material, and again different approaches are possible and are being widely considered by the scientific community:

1. The first is almost based on the collected data, complemented with trusted first principles (and their associated variational formulations). The collected data is expected to describe solely the material phenomenological behavior, without the need of assuming any template or further assumptions, as considered in the seminal work of Ortiz [69], with many others that followed [15,30,48,70,71,118].
2. A second approach considers the collected data to lie on a manifold embedded into the higher-dimensional behavior space. The manifold dimensionality depends on the complexity of the behavior (linear or nonlinear, reversible or history dependent, ...) As soon as the collected data allows to infer the manifold structure, then first principles are solved with the data-driven behavior manifold [57,59,62,76].
3. A third approach is much more aligned with the physics-informed rationale. The collected data are used with a number of quite general rules representing, in a quite general form, the material description, for constructing the so-called *constitutive manifold* that intimately embraces data and existing knowledge [74].
4. Other physics-informed approaches, within a thermodynamical setting, proposed a regression of the free energy, from which the behavior result by simple differentiation, and the regression is then tuned with respect to the collected data, while enforcing during the construction as many conditions as constraints dictated by the existing knowledge (symmetries, objectivity, among others) [76,77,91,138]. Other physics-informed approaches were proposed in more complex (dissipative) settings [43,49,50].

5. Finally, within the augmented rationale (or hybrid paradigm) the real behavior can be assumed represented by a first order one (calibrated at best from the available data) complemented by an enrichment (or correction) filling the gap between the collected data and the predictions obtained from the chosen and calibrated model [44], with again some constraints applying during the data-driven enrichment model construction (e.g., convexity of the yield function [61]).

Other works address multi-scale problems and the use of the so-called Deep Material Network [38,39,83,84], micro and macrostructural analysis [82], learning constitutive equations from indirect observations or the plasticity modeling [41,47,52,53,93,133].

Despite of the important progresses recently accomplished, many challenges remain open and are attracting a lot of interest within the scientific community, among them, the ones concerning:

- The description and evolution of the so-called internal variables, able to condensate at the current time all the effects of the past material history. Some proposals exist, however, [50].
- The material description at the microscopic scale to take into account the subjacent physics, that can be impacted by a bad (or too poor) choice of the descriptors and model features. This point is essential to address the inverse problem of finding the best atomic or microscopic structure for attaining the optimal macroscopic properties, at the origin of the so-called *materials by design*.
- At the scale of the structure, a wide topic concern the Structural Health Monitoring, SHM. The loss of performance can be motivated by some amount of local damage, expected to be identified from the analysis of experimental data (diagnosis). However, to give more quantitative predictions on present or future consequences or actions, a model seems a valuable option. The diagnosis and prognosis [101,120] must be accompanied of an effective sensing.

  For that purpose, physics-augmented learning (the hybrid modeling approach) seems to be particularly well adapted. We could assume that the real structure can be expressed from its undamaged counterpart (assumed well modeled) complemented by a correction that removes from the undamaged model the mechanical performances at the location were damage occurs. To locate and quantify that searched correction, the (local and global) structure equilibrium, as well as the collected data, suffice for calculating the data-driven physics-informed model correction.
- In the case of very large structures, the model can not retain all the details. A resolution level able to represent all the structural details will be numerically untreatable, and by coarsening its representation (as usually carried out in practice), the effects of the details are lost. The Grail consists of enriching the model, without increasing its size or resolution, for accurately representing the collected data. In that case, as the model should be enriched everywhere within the hybrid modeling approach, the correction will become too rich with respect to the usually scarce, available data. When the structure is subjected to loads living in a certain reduced subspace, the structural problem can be formulated in a reduced space, and there, the enrichment becomes local and few data suffice.

Data-driven fluids and flows

Concerning fluids, we find first the complex rheology associated to the so-called Non-Newtonian fluids, whose behavior, in general nonlinear, also depends on a series of conformational coordinates (also known as configurational), whose number and time evolution have to be modeled. Since these fluids are composed of entangled macro-molecules, or consist of concentrated particle suspensions, the correlation of these fine descriptions with the resulting mesoscopic properties, very much resembles the just addressed discussion about the data-driven description of solid materials.

Fluids, even those with the simplest rheology, linear (Newtonian) fluids, face other difficulties, such as incompressibility. This is also found in some classes of solid materials and meta-materials, and the nonlinear advective term at the origin of turbulence. The main consequences are:

- The need to employ appropriate discretization schemes with respect to the advective terms and with respect to the mixed formulations (to address the fluid incompressibility or the non-Newtonian rheology).
- The need to describe turbulence throughout all the scales at which its effects manifest.
- The model change: Stokes model (linear and elliptic) at low Reynolds numbers, whose only difficulty comes from the incompressibility constraint. The Navier–Stokes, NS, model at higher Reynolds numbers, where the advective effects must be addressed, and then, after a certain critical Reynolds number, turbulence comes into play. Finally, the NS model degenerates into the compressible Euler model where discontinuities (shocks) can appear, needing for adequate numerical schemes for capturing and resolving them.
- The strongly nonlinear couplings in presence of phase change, chemical reactions, combustion, ...

All these difficulties entail extremely fine discretizations (in space and time), challenging the most powerful computing platforms. It is for that reason that machine learning techniques are attracting the more and more interest, without the aim of being exhaustive:

- To model the complex rheology from the use of any machine learning procedure (data-driven, physics-informed or physics-augmented), operating more at the fluid scale than at the one of the flow.
- To discover or tune discretization schemes with the optimal properties (stability and accuracy).
- To model and describe turbulence. In that sense, a hybrid approach could assume a first order model and enrich it (from a data-driven correction) to better represent the experimental findings.
- To enrich coarser representations within the hybrid approach for conciliating accuracy and effectiveness.
- To construct (internal or external) flow (aerodynamic or hydrodynamic) surrogates with respect to a number of features (geometry, inflow velocity, ...), and then, including them in the optimization loop or in any application needing real-time flow evaluations.

Some valuable references covering the topics previously discussed are [5,64,65,85,104, 114,142].

Data-driven processes

Processes transform matter in structures, properties into performance. Processes involve solids and fluids, structures (e.g., stamping), flows (e.g. injection or extrusion) and all the physics with the associated couplings. Thus, a process becomes a multi-parametric transfer function that groups all the parameters characterizing the incoming materiel with all the ones that are characteristic of the process itself.

Here, one usually looks for a function that expresses the final, targeted performance as a function of a number of features. The goal is discovering the best material/process couple associated with the optimal performances, enabling *performance by design*. For that purpose, the different data clustering and classification, and the different data-driven (informed or not) regressions are being widely employed.

In general, most modeling approaches remain too coarse-grained and lack of generality. Relating the oven temperature to the time evolution of the temperature of a thermally treated part can not be generalized to a part with different size or geometry, for example. To enhance generality, the data features must be enriched or/and the learning process informed or augmented.

To the just referred difficulties, we should add the ones coming from the multi-physics coupling: electromagnetic forming, induction or micro-waves based processes, thermal treatments (heating or tempering), chemical (e.g. reactive resin transfer moulding), mechanical (vibrations, ultrasounds, shoot penning, ...) or the ones coming from the fact of considering multi-physics performances (thermal, acoustic, damping, ...) Mastering all the connections is crucial for inverting the design arrow, enabling materials and performance by design.

There is an exponential increase in the number of publications reporting that modeling route, in many technology domains: machining and drilling [17,25,68], additive manufacturing [143], reactive extrusion [16,63], induction hardening [28], chemical reactions [136], among many others.

Data-driven complex systems of systems

Complex systems of systems represent one of the most challenging scenarios. The system size, entanglements, the variability and uncertainty propagating far from its source, the presence of emergent behaviors, chaotic dynamics, etc., make it difficult to proceed with either fully data-driven techniques (not enough data) or physics-based model (too deterministic and unable to cover large systems while keeping the right degree of resolution).

The use of physics-informed and physics-augmented learning

procedures combined with some physics-based model of components or system parts, and fully data-driven model of the other parts, where no model or knowledge exist, seems a valuable option for succeeding their modeling.

This framework is expected contributing to enhanced smarts grids, smart cities and nations, smart industry (including the economic ecosystem), mobility networks, ... that will constitute without any doubt the next technological revolution.

Conclusions

In this short review we have revisited the main methodologies available to acquire knowledge from data. They make use of data almost exclusively, while others incorporate physics and knowledge in different ways, by informing or by augmenting the learning process.

The three main protagonists have been revisited: data with its amazing richness, machine learning procedures, and the ones enabling to gain confidence on data-driven designs and decisions.

In the second part of the paper, we have revisited four major application domains, by referring to some existing works, and highlighting some remaining major challenges.

As Winston Churchill once said in another, very different context: *Now this is not the end. It is not even the beginning of the end. But it is, perhaps, the end of the beginning.*

Acknowledgements
Author acknowledges the support of the ESI Group through its research chair at Arts et Metiers Institute of Technology and the University of Zaragoza.

Author contributions
All the authors participated equally in the paper content. Both authors read and approved the final manuscript.

Availability of data and materials
Interested reader can contact authors.

Declarations

Competing interests
The authors declare that they have no competing interests.

Received: 14 February 2022 Accepted: 25 September 2022
Published online: 27 October 2022

References

1. Agrawal A, Choudhary A. Perspective: materials informatics and big data: realization of the fourth paradigm of science in materials science. APL Mater. 2016;4:053208.
2. Argerich C, Ibanez R, Barasinski A, Chinesta F. Code2vect: an efficient heterogenous data classifier and nonlinear regression technique. C R Mecanique. 2019;347:754–61.
3. Argerich C, Carazo A, Sainges O, Petiot E, Barasinski A, Piana M, Ratier L, Chinesta F. Empowering design based on hybrid twin: application to acoustic resonators. Designs. 2020;4:44.
4. Bartok AP, Kondor R, Csanyi G. On representing chemical environments. Phys Rev B. 2013;87: 184115.
5. Bar-Sinai Y, Hoyer S, Hickey J, Brenner MP. Learning data-driven discretizations for partial differential equations. Proc Natl Acad Sci. 2019;116:15344–9.
6. Baydin AG, Pearlmutter BA, Radul AA, Siskind JM. Automatic differentiation in machine learning: a survey. J Mach Learn Res. 2018;18:1–43.
7. Bertalan T, Dietrich F, Mezic I, Kevrekidis IG. On learning Hamiltonian systems from data. Chaos. 2019;29: 121107.
8. Billings SA. Nonlinear system identification: NARMAX methods in the time, frequency and spatio-temporal domains. Hoboken: Wiley; 2013.
9. Borzacchiello D, Aguado JV, Chinesta F. Non-intrusive sparse subspace learning for parametrized problems. Arch Comput Methods Eng. 2019;26:303–26.
10. Breiman L. Random forests. Mach Learn. 2001;45:5–32.
11. Bronstein M, Bruna J, Cohen T, Velickovic P. Geometric deep learning, grids, groups, graphs, geodesics and gauges. arXiv:2104.13478.
12. Brunton S, Proctor JL, Kutz N. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. PNAS. 2016;113(15):3932–7.
13. Brunton SL, Brunton BW, Proctor JL, Kaiser E, Kutz JN. Chaos as an intermittently forced linear system. Nat Commun. 2017. https://doi.org/10.1038/s41467-017-00030-8.
14. Carlsson GG. Topology and data. Bull Am Math Soc. 2009;46(2):255–308.
15. Carrara P, De Lorenzis L, Stainier L, Ortiz M. Data-driven fracture. Comput Methods Appl Mech Eng. 2020;372: 113390.
16. Casteran F, Delage K, Cassagnau Ph, Ibanez R, Aregerich C, Chinesta F. Application of machine learning tools for the improvement of reactive extrusion simulation. Macromol Mater Eng. 2020. https://doi.org/10.1002/mame.202000375.
17. Charalampous P. Prediction of cutting forces in milling using machine learning algorithms and finite element analysis. J Mater Eng Perform. 2021;30:2002–13.
18. Chen T, Chen H. Approximations of continuous functionals by neural networks with application to dynamic systems. IEEE Trans Neural Netw. 1993;4(6):910–8.

19. Chen T, Chen H. Universal approximation to nonlinear operators by neural networks with arbitrary activation functions and its application to dynamical systems. IEEE Trans Neural Netw. 1995;6(4):911–7.
20. Chinesta F, Ladeveze P, Cueto E. A short review in model order reduction based on Proper Generalized Decomposition. Arch Comput Methods Eng. 2011;18:395–404.
21. Chinesta F, Leygue A, Bordeu F, Aguado JV, Cueto E, Gonzalez D, Alfaro I, Ammar A, Huerta A. Parametric PGD based computational vademecum for efficient design, optimization and control. Arch Comput Methods Eng. 2013;20(1):31–59.
22. Chinesta F, Keunings R, Leygue A. The proper generalized decomposition for advanced numerical simulations. A primer. Springerbriefs. Berlin: Springer; 2014.
23. Chinesta F, Huerta A, Rozza G, Willcox K. Model order reduction. In: Stein E, de Borst R, Hughes T, editors. The encyclopedia of computational mechanics. 2nd ed. Hoboken: Wiley; 2015.
24. Chinesta F, Cueto E, Abisset-Chavanne E, Duval JL, El Khaldi F. Virtual, digital and hybrid twins: a new paradigm in data-based engineering and engineered data. Arch Comput Methods Eng. 2020;27:105–34.
25. Chupakhin S, Kashaev N, Klusemann B, Huber N. Artificial neural network for correction of effects of plasticity in equibiaxial residual stress profiles measured by hole drilling. J Strain Anal. 2017;52(3):137–51.
26. Cristianini N, Shawe-Taylor J. An introduction to support vector machines and other kernel-based learning methods. New York: Cambridge University Press; 2000.
27. Darema F. Engineering/scientific and commercial applications: differences, similarities, and future evolution. In: Proceedings of the Second Hellenic European Conference on mathematics and informatics. HERMIS. 1994;1:367–74.
28. Derouiche K, Garois S, Champaney V, Daoud M, Traidi K, Chinesta F. Data-driven modelling for multi-physics parametrized problems—application to induction hardening process. Metals. 2021;11(5):738.
29. Dimiduk DM, Holm EA, Niezgoda SR. Perspectives on the impact of machine learning, deep learning, and artificial intelligence on materials, processes and structures engineering. Integr Mater Manuf Innov. 2018;7:157–72.
30. Eggersmann R, Kirchdoerfer R, Reese S, Stainier L, Ortiz M. Model-free data-driven inelasticity. Comput Methods Appl Mech Eng. 2019;350:81–99.
31. Escofier B. Traitement Simultane de Variables Quantitatives et Qualitatives en Analyse Factorielle. Les Cahiers de Analyse des Donnees. 1979;4(2):137–46.
32. Escofier B, Pages J. Analyses Factorielles Simples et Multiples. Dunod. 2008.
33. Frahi T, Yun M, Argerich C, Falco A, Chinesta F. Tape surfaces characterization with persistence images. AIMS Mater Sci. 2020;7(4):364–80.
34. Frahi T, Chinesta F, Falco A, Badias A, Cueto E, Choi HY, Han M, Duval JL. Empowering advanced driver-assistance systems from topological data analysis. Mathematics. 2021;9:634.
35. Frahi T, Falco A, Vinh Mau B, Duval JL, Chinesta F. Empowering advanced parametric modes clustering from topological data analysis. Appl Sci. 2021;11:6554.
36. Freund Y, Schapire R. A decision-theoretic generalization of on-line learning and an application to boosting. J Comput Syst Sci. 1997;55(1):119–39.
37. Friedman J. Greedy function approximation: a gradient boosting machine. Ann Stat. 2000;29:1189–232.
38. Gajek S, Schneider M, Bohlke T. An FE-DMN method for the multiscale analysis of short fiber reinforced plastic components. Comput Methods Appl Mech Eng. 2021;384: 113952.
39. Gajek S, Schneider M, Bohlke T. On the micromechanics of deep material networks. J Mech Phys Solids. 2020;142: 103984.
40. Ghanem R, Soize C, Mehrez L, Aitharaju V. Probabilistic learning and updating of a digital twin for composite material systems. IJNME. 2020. https://doi.org/10.1002/nme.6430.
41. Ghavamian F, Simone A. Accelerating multiscale finite element simulations of history-dependent materials using a recurrent neural network. Comput Methods Appl Mech Eng. 2019;357(1):112594.
42. Gomez-Bombarelli R, Wei JN, Duvenaud D, Hernandez-Lobato JM, Sanchez-Lengeling B, Sheberla D, Aguilera-Iparraguirre J, Hirzel TD, Adams RP, Aspuru-Guzik A. Automatic chemical design using a data-driven continuous representation of molecules. ACS Cent Sci. 2018;4(2):268–76.
43. Gonzalez D, Chinesta F, Cueto E. Thermodynamically consistent data-driven computational mechanics. Contin Mech Thermodyn. 2019;31:239–53.
44. Gonzalez D, Chinesta F, Cueto E. Learning corrections for hyper-elastic models from data. Front Mater Sect Comput Mater Sci. 2019;6:14.
45. Goodfellow I, Bengio Y, Courville A. Deep learning. Cambridge: MIT Press; 2016.
46. Greydanus S, Dzamba M, Yosinski J. Hamiltonian neural networks. 2019. arXiv:1906.01563v3.
47. Hartmaier A. Data-oriented constitutive modeling of plasticity in metals. Materials. 2020;13(7):1600.
48. He Q, Chen J. A physics-constrained data-driven approach based on locally convex reconstruction for noisy database. Comput Methods Appl Mech Eng. 2020;363: 112791.
49. Hernandeza Q, Badias A, Gonzalez D, Chinesta F, Cueto E. Deep learning of thermodynamics-aware reduced-order models from data. J Comput Phys. 2021;426:109950.
50. Hernadez Q, Gonzalez D, Chinesta F, Cueto E. Learning non-Markovian physics from data. J Comput Phys. 2021;428:109982.
51. Hinton GE, Zemel RS. Autoencoders, minimum description length and Helmholtz free energy. In: Advances in neural information processing systems 6 (NISP 1993). Morgan-Kaufmann. 1994:3–10.
52. Huang DZ, Xu K, Farhat C, Darve E. Learning constitutive relations from indirect observations using deep neural networks. J Comput Phys. 2020;416:109491.
53. Huang D, Fuhg JN, Weibenfels C, Wriggers P. A machine learning based plasticity model using proper orthogonal decomposition. Comput Methods Appl Mech Eng. 2020;365:113008.
54. Huangy S, Hey Z, Chem B, Reina C. Variational Onsager Neural Networks (VONNs): a thermodynamics-based variational learning strategy for non-equilibrium PDEs. arXiv:2112.09085.
55. Hughes TW, Williamson IAD, Minkov M, Fan S. Wave physics as an analog recurrent neural network. Sci Adv. 2019;5(12):eaay6946.

56. Husson F, Josse J. missMDA: Handling missing values with/in multivariate data analysis (principal component methods). R package version 1.10. 2016. https://CRAN.R-project.org/package=missMDA.
57. Ibanez R, Borzacchiello D, Aguado JV, Abisset-Chavanne E, Cueto E, Ladeveze P, Chinesta F. Data-driven non-linear elasticity. Constitutive manifold construction and problem discretization. Comput Mech. 2017;60(5):813–26.
58. Ibanez R, Abisset-Chavanne E, Ammar A, Gonzalez D, Cueto E, Huerta A, Duval JL, Chinesta F. A multi-dimensional data-driven sparse identification technique: the sparse Proper Generalized Decomposition. Complexity. 2018. Article ID 5608286.
59. Ibanez R, Abisset-Chavanne E, Aguado JV, Gonzalez D, Cueto E, Chinesta F. A manifold learning approach to data-driven computational elasticity and inelasticity. Arch Comput Methods Eng. 2018;25(1):47–57.
60. Ibanez R, Abisset-Chavanne E, Cueto E, Ammar A, Duval JL, Chinesta F. Some applications of compressed sensing in computational mechanics. Model order reduction, manifold learning, data-driven applications and nonlinear dimensionality reduction. Comput Mech. 2019;64:1259–71.
61. Ibanez R, Abisset-Chavanne E, Gonzalez D, Duval JL, Cueto E, Chinesta F. Hybrid constitutive modeling: data-driven learning of corrections to plasticity models. Int J Mater Form. 2019;12:717–25.
62. Ibanez R, Gilormini P, Cueto E, Chinesta F. Numerical experiments on unsupervised manifold learning applied to mechanical modeling of materials and structures. CRAS Mecanique. 2020;348(10–11):937–58.
63. Ibanez R, Casteran F, Argerich C, Ghnatios C, Hascoet N, Ammar A, Cassagnau P, Chinesta F. Data-driven modeling of reactive extrusion. Fluids. 2020;5(2):94.
64. Jiang C, Vinuesa R, Chen R, Mi J, Laima S, Li H. An interpretable framework of data-driven turbulence modeling using deep neural networks. Phys Fluids. 2021;33: 055133.
65. Jin X, Cai S, Li H, Karniadakis GE. NSFnets (Navier-Stokes flow nets): physics-informed neural networks for the incompressible Navier-Stokes equations. J Comput Phys. 2021;426: 109951.
66. Kaelbling LP, Littman ML, Moore AW. Reinforcement learning: a survey. Artif Intell Res. 1996;4:237–85.
67. Kapteyn MG, Willcox KE. From physics-based models to predictive digital twins via interpretable machine learning. 2020. arXiv:2004.11356v3.
68. Kim D, Kim TJY, Wang X, Kim M, Quan Y, Woo OhJ, Min S, Kim H, Bhandari B, Yang I, Ahn S. Smart machining process using machine learning: a review and perspective on machining industry. Int J Precis Eng Manuf-Green Technol. 2018;5(4):555–68.
69. Kirchdoerfer T, Ortiz M. Data-driven computational mechanics. Comput Methods Appl Mech Eng. 2016;304:81–101.
70. Kirchdoerfer T, Ortiz M. Data driven computing with noisy material data sets. Comput Methods Appl Mech Eng. 2017;326:622–41.
71. Kirchdoerfer T, Ortiz M. Data-driven computing in dynamics. Int J Numer Methods Eng. 2018;113(11):1697–710.
72. Kirkwood CW. Decision Tree primer. 2002. http://creativecommons.org/licenses/by-nc/3.0/.
73. Kubicek M, Minisci E, Cisternino M. High dimensional sensitivity analysis using surrogate modeling and high dimensional model representation. Int J Uncertain Quantif. 2015;5(5):393–414.
74. Ladeveze P, Neron D, Gerbaud P-W. Data-driven computation for history-dependent materials. C R Mecanique. 2019;347(11):831–44.
75. Lam R, Horesh L, Avron H, Willcox KE. Should you derive, or let the data drive? An optimization framework for hybrid first-principles data-driven modeling. 2017. arXiv:1711.04374v1.
76. Latorre M, Montans FJ. What-you-prescribe-is-what-you-get orthotropic hyperelasticity. Comput Mech. 2014;53(6):1279–98.
77. Latorre M, Montans FJ. Experimental data reduction for hyperelasticity. Comput Struct. 2020;232: 105919.
78. LeCun Y. Self supervised learning. https://www.youtube.com/watch?v=SaJL4SLfrcY.
79. Lee JA, Verleysen M. Nonlinear dimensionality reduction. New York: Springer; 2007.
80. Lee K, Trask NA, Stinis P. Machine learning structure preserving brackets for forecasting irreversible processes. 2021. arXiv:2106.12619v1.
81. Lhermitte S, Verbesselt J, Verstraeten W, Coppin P. A comparison of time series similarity measures for classification and change detection of ecosystem dynamics. Remote Sens Environ. 2011;115:3129–52.
82. Liu Z, Fleming M, Liu WK. Microstructural material database for self-consistent clustering analysis of elastoplastic strain softening materials. Comput Methods Appl Mech Eng. 2018;330:547–77.
83. Liu Z, Wu CT, Koishi M. A deep material network for multiscale topology learning and accelerated nonlinear modeling of heterogeneous materials. Comput Methods Appl Mech Eng. 2019;345:1138–68.
84. Liu Z, Wu CT. Exploring the 3D architectures of deep material network in data-driven multiscale mechanics. J Mech Phys Solids. 2019;127:20–46.
85. Loiseau J, Noack B, Brunton S. Sparse reduced-order modelling: sensor-based dynamics to full-state estimation. J Fluid Mech. 2018;844:459–90.
86. Lu L, Jin P, Karniadakis GE. DeepONet: learning nonlinear operators for identifying differential equations based on the universal approximation theorem of operators. 2020. arXiv:1910.03193v3.
87. Maaten L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res. 2008;9:2579–605.
88. MacKay D. Chapter 20—An example inference task: clustering. information theory, inference and learning algorithms. Cambridge: Cambridge University Press. 2003:84–292.
89. MacQueen J. Some methods for classification and analysis of multivariate observations. In: Proceedings of 5th Berkeley symposium on mathematical statistics and probability. University of California Press. 1967:281–97.
90. Masi F, Stefanou I, Vannucci P, Maffi-Berthier V. Thermodynamics-based Artificial Neural Networks for constitutive modeling. 2020. arXiv:2005.12183v1.
91. Minano M, Montans FJ. WYPiWYG damage mechanics for soft materials: a data-driven approach. Arch Comput Methods Eng. 2018;25:165–93.
92. Moya B, Badias A, Alfaro I, Chinesta F, Cueto E. Digital twins that learn and correct themselves. Int J Numer Methods Eng. 2022. https://doi.org/10.1002/nme.6535.
93. Mozaffar M, Bostanabad R, Chen W, Ehmann K, Cao J, Bessa MA. Deep learning predicts path-dependent plasticity. PNAS. 2019;116(52):26414–20.

94. Muller M. Information retrieval for music and motion. Berlin: Springer; 2007.
95. Neggers J, Allix O, Hild F, Roux S. Big Data in experimental mechanics and model order reduction: today challenges and tomorrow opportunities. Arch Comput Methods Eng. 2018;25(1):143–64.
96. Nielsen M. Neural networks and deep learning. 2019. http://neuralnetworksanddeeplearning.com/chap4.html.
97. van Otterlo M, Wiering M. Reinforcement learning and Markov decision processes. In: Wiering M, van Otterlo M, editors. Reinforcement learning adaptation, learning, and optimization, vol. 12. Berlin: Springer; 2012.
98. Oudot SY. Persistence theory: from quiver representation to data analysis, American Mathematical Society. Mathematical surveys and monographs. 2010;209:2010.
99. Oulghelou M, Allery C. Parametric reduced order models based on a Riemannian Barycentric Interpolation. Int J Numer Methods Eng. 2021;122:6623–40.
100. Qin T, Wu K, Xiu D. Data driven governing equations approximation using deep neural networks. J Comput Phys. 2019;395(15):620–35.
101. Quaranta G, Lopez E, Abisset-Chavanne E, Duval JL, Huerta A, Chinesta F. Structural health monitoring by combining machine learning and dimensionality reduction techniques. Rev Int de Metodos Numericos en Calculo y Diseno en Ingenieria. 2019;35(1).
102. Raissi M, Perdikaris P, Karniadakis GE. Physics-informed neural networks: a deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. J Comput Phys. 2019;378:686–707.
103. Ramprasad R, Batra R, Pilania G, Mannodi-Kanakkithodi A, Kim C. Machine learning in materials informatics: recent applications and prospects. NPJ Comput Mater. 2017;3:54.
104. Ranade R, Hill C, Pathak J. DiscretizationNet: a machine-learning based solver for Navier-Stokes equations using finite volume discretization. Comput Methods Appl Mech Eng. 2021;378: 113722.
105. Reille A, Hascoet N, Ghnatios C, Ammar A, Cueto E, Duval JL, Chinesta F, Keunings R. Incremental dynamic mode decomposition: a reduced-model learner operating at the low-data limit. C R Mecanique. 2019;347:780–92.
106. Reille A, Champaney V, Daim F, Tourbier Y, Hascoet N, Gonzalez D, Cueto E, Duval JL, Chinesta F. Learning data-driven reduced elastic and inelastic models of spot-welded patches. Mech Ind. 2021;22:32.
107. Roweis T, Saul LK. Nonlinear dimensionality reduction by Locally Linear Embedding. Science. 2000;290:2323–6.
108. Sancarlos A, Cameron M, Abel A, Cueto E, Duval JL, Chinesta F. From ROM of electrochemistry to AI-based battery digital and hybrid twin. Arch Comput Methods Eng. 2021;28:979–1015.
109. Sancarlos A, Le Peuvedic JM, Groulier J, Duval JL, Cueto E, Chinesta F. Learning stable reduced-order models for hybrid twins A. Sancarlos, M. Cameron. Data Centric Eng. 2021;2:e10.
110. Sancarlos A, Champaney V, Duval JL, Cueto E, Chinesta F. PGD-based advanced nonlinear multiparametric regressions for constructing metamodels at the scarce-data limit. arXiv:2103.05358.
111. Schmid PJ. Dynamic mode decomposition of numerical and experimental data. J Fluid Mech. 2010;656:528.
112. Schmidt J, Marques MRG, Botti S, Marques MAL. Recent advances and applications of machine learning in solid-state materials science. NPJ Comput Mater. 2019;5:83.
113. Schmidhuber J. Deep learning in neural networks: an overview. Neural Netw. 2015;61:85–117.
114. Schwander L, Ray D, Hesthaven JS. Controlling oscillations in spectral methods by local artificial viscosity governed by neural networks. J Comput Phys. 2021;431: 110144.
115. Senin P. Dynamic time warping algorithm review. Technical report. 2008.
116. Settles B. Active learning literature survey. Computer Sciences Technical Report 1648. University of Wisconsin-Madison. 2009.
117. Srivastava N, Salakhutdinov R. Multimodal learning with deep Boltzmann machines. J Mach Learn Res. 2014;15:2949–80.
118. Stainier L, Leygue A, Ortiz M. Model-free data-driven methods in mechanics: material data identification and solvers. 2019. arXiv:1903.07983v2.
119. Sutton RS, Barto AG. Reinforced learning. An introduction. Cambridge: The MIT Press; 2018.
120. Taddei T, Penn JD, Yano M, Patera AT. Simulation-based classification: a model-order-reduction approach for structural health monitoring. Arch Comput Methods Eng. 2018;25(1):23–45.
121. Tang K, Congedo PM, Abgrall R. Sensitivity analysis using anchored ANOVA expansion and high order moments computation. [Research Report] RR-8531. 2014.
122. Torquato S. Statistical description of microstructures. Annu Rev Mater Res. 2002;32:77–111.
123. Tuegel EJ, Ingraffea AR, Eason TG, Spottswood SM. Reengineering aircraft structural life prediction using a digital twin. Int J Aerosp Eng. 2011;2011: 154798.
124. Venkatesan R, Li B. Convolutional neural networks in visual computing: a concise guide. Boca Raton: CRC Press; 2017.
125. Vermeeren M, Bravetti A, Seri M. Contact variational integrators. J Phys A Math Theor. 2019;52:445206.
126. Villani C. Optimal transport. Old and new. Berlin: Springer; 2006.
127. Villani C, et al. AI for humanity. 2018. https://www.aiforhumanity.fr/.
128. Wang K, Gou C, Duan Y, Lin Y, Zheng X, Wang FY. Generative adversarial networks: introduction and outlook. IEEE J Autom Sin. 2017;4(4):588–98.
129. Wang Z, Guet C. Self-consistent learning of neural dynamical systems from noisy time series. IEEE Trans Emerg Top Comput Intell. 2022. https://doi.org/10.1109/TETCI.2022.3146332.
130. Warren J. The materials genome initiative and artificial intelligence. MRS Bull. 2018;43(6):452–7.
131. Weiss K, Khoshgoftaar TM, Wang DD. A survey of transfer learning. J Big Data. 2016;3:9.
132. Williams MO, Kevrekidis G, Rowley CW. A data-driven approximation of the Koopman operator: extending dynamic mode decomposition. J Nonlinear Sci. 2015;25(6):1307–46.
133. Wu L, Nguyen VD, Kilingar NG, Noels L. A recurrent neural network-accelerated multi-scale model for elasto-plastic heterogeneous materials subjected to random cyclic and non-proportional loading paths. Comput Methods Appl Mech Eng. 2020;369: 113234.
134. Wu S, Kondo Y, Kakimoto M, Yang B, Yamada H, Kuwajima I, Lambard G, Hongo K, Xu Y, Shiomi J, Schick C, Morikawa J, Yoshida R. Machine-learning-assisted discovery of polymers with high thermal conductivity using a molecular design algorithm. NPJ Comput Mater. 2019;5:1–11.

135. Xue F, Li Q, Li X. The combination of circle topology and leaky integrator neurons remarkably improves the performance of echo state network on time series prediction. PLoS ONE. 2017;12(7):e0181816.
136. Yang W, Peng L, Zhu Y, Hong L. When machine learning meets multiscale modeling in chemical reactions. J Chem Phys. 2020;153:094117.
137. Yun M, Argerich C, Cueto E, Duval JL, Chinesta F. Nonlinear regression operating on microstructures described from Topological Data Analysis for the real-time prediction of effective properties. Materials. 2020;13(10):2335.
138. Zhang X, Garikipati K. Machine learning materials physics: Multi-resolution neural networks learn the free energy and nonlinear elastic response of evolving microstructures. Comput Methods Appl Mech Eng. 2020;372:113362.
139. Zhang Z, Shin Y, Karniadakis GE. GFINNs: GENERIC formalism informed neural networks for deterministic and stochastic dynamical systems. 2021. arXiv:2109.00092v1.
140. Zhu X. Semi-supervised learning. University of Wisconsin-Madison.
141. Zhu X. Semi-supervised learning literature survey. Madison: University of Wisconsin; 2008.
142. Zhuang J, Kochkov D, Bar-Sinai Y, Brenner MP. Learned discretizations for passive scalar advection in a two-dimensional turbulent flow. Phys Rev Fluids. 2021;6:064605.
143. Zohdi TI. Dynamic thermomechanical modeling and simulation of the design of rapid free-form 3D printing processes with evolutionary machine learning. Comput Methods Appl Mech Eng. 2018;331:343–62.

Publisher's Note