Statistics and Probability Letters 136 (2018) 63–67

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

The role of statistics in the era of big data: A computational scientist' perspective

Alfio Quarteroni *

CMCS, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
MOX, Politecnico di Milano, Milan, Italy

ARTICLE INFO

Article history:
Available online 21 February 2018

Keywords:
Numerical models
Computational methods
Data assimilation
Uncertainty quantification
Data analysis

ABSTRACT

In their modern implementation, computational models based on first principles from Physics can dramatically benefit from the recent explosion of Data Science. In fact, these two branches of applied mathematics can virtuously interplay, and at a large extent they already do.

1. Modeling Based Scientific Computing (MBSC)

Modeling Based Scientific Computing (MBSC) is a branch of Computational Science. It is built on mathematical models derived from first principles expressing the laws of nature and on accurate and efficient numerical algorithms for their approximation. It exploits scientific software designed for powerful computational platforms to solve the associated (often, large scale) algebraic problems, and validate the computed solution against reality. *Input data* (under the form of medical images, geometrical shapes, initial conditions, boundary conditions, forcing terms, model coefficients, etc.) are essential for model definition, while *data from measurements and observations* are crucial for model validation.

In a deterministic setting, MBSC has been tremendously successful in achieving truly predictive capabilities that led to substantial design improvements in automotive and aerospace industry, better exploitation strategies in oil industry, accurate weather forecast, the control and optimization of power plants for cleaner and less polluting emissions, to name just a few. See Rüde et al. (2016).

MBSC is traditionally used *in conjunction with theory and experiments*. However, it has also been employed in cases where mathematical theory is not yet available (for instance to simulate complex multi-physics problems, e.g. in computational medicine), or when experimental data are dangerous or impossible to achieve (like for nuclear tests, the reentry of space vehicles from the upper atmosphere, the simulation of extreme events such as earthquakes or volcanic eruptions, etc.).

As reported in the Preface of the Fourth Paradigm book dedicated to Jim Gray (Hey et al., 2009), the history of science has been historically developing along *four phases*: the first based upon empirical science and observations, the second upon theoretical science and mathematically-driven insights, the third upon computational science and simulation-driven insights, the fourth upon data-driven insights of modern scientific research.

MBSC has marked the third phase, whereas we have now entered the fourth, data driven, phase. This temporal partition, however, is not disjoint: it overlaps.

Virtually unlimited amount of scientific data are nowadays generated from multiple sources, such as Internet and digital networks, broad networks of sensors, large scale experiments or measures (from micro, such as high energy physics, to

* Correspondence to: CMCS, Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland.
E-mail address: alfio.quarteroni@polimi.it.

https://doi.org/10.1016/j.spl.2018.02.047
0167-7152/© 2018 Elsevier B.V. All rights reserved.

macro, as those generated by Earth observation satellites, or by mixing streaming and historical data). This volume of data, suitably combined with statistical models, can provide new investigation tools in those areas or circumstances where MBSC are *inapplicable* because first principles are lacking or inappropriate to model and simulate complex processes. Otherwise, they can be *combined* with MBSC for data assimilation and to quantify uncertainties in the results provided by models based on first principles. This is even more crucial when MBSC is applied in new areas including the social sciences, humanities, business, finance, and government policy, where uncertainty, hazard and randomness play a major role.

2. Data driven models

Dual to MBSC is *data driven scientific discovery* in the era of big data. According to this new paradigm, statistics-based models from data mining and machine (statistical) learning are triggered on large data sets to analyze and predict complex phenomena.

Data are without hypotheses about what they might show. "We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot" (Anderson, 2008).

Data management, data mining, visual analytics for visual communication of the results issuing from complex data analyses, statistics and deep learning, represent basic pillars of data science. Challenges are research reproducibility (data should be seen and saved, software code available, workflows replayed), reusability (how to use data on new workflows), creation of knowledge graphs (who uses the data and for doing what), and IP protection.

With the traditional MBSC paradigm, data are ancillary to the model. With the data driven paradigm, the dream is to allow statistical algorithms to unveil the laws and patterns governing complex data systems when first-principles cannot. This is a fascinating temptation. It exploits massive data sets and massive computational power running in model parallelism and data parallelism.

"The big data era has created a new scientific paradigm: collect data first, ask questions later. When the universe of scientific hypotheses that are being examined simultaneously is not taken into account, inferences are likely to be false. The consequence is that follow up studies are likely not to be able to reproduce earlier reported findings or discoveries" (Candes, 2017).

The deep learning revolution, a modern reincarnation of artificial neural networks according to some, aims at finding common representations across domains by replacing piles of codes with data and learning. "It is a powerful class of machine learning models, a collection of simple, trainable mathematical functions that are compatible with many variants of machine learning" (Dean, 2016).

Speech and image recognition, object recognition and detection, machine translation, language modeling, are domains where deep learning is showing fantastic achievements. There is no (yet) evidence, however, that deep learning can be equally successful for simulating physics based processes such as, say, complex flow fields featuring multiple scales interactions. And, of course, many others.

More in general, there seems to be a need to fill the gap between data providers, data scientists and computational scientists to collaborate for improving the predictive capabilities of computational models at large; a room for cooperation, rather than competition, between data science and computational science.

3. The interplay

Models based on first principles can extract from large scientific data sets valuable insights that can go far beyond what can be recovered by black-box statistical modeling alone. Aided by the availability of large data sets, domains such as biology, medicine, and even social sciences, are increasingly becoming quantitative sciences (King, 2014).

Computational scientists' belief is that "Models based on first principles are essential components of systems that extract valuable insights from massive scientific data, insights that tend to go far beyond what can be recovered by black-box statistical modeling alone" (Rüde et al., 2016).

Model-based scientific computing and data science are indeed strongly interconnected in the modern way to design numerical simulation processes, as schematically represented with the help of the synthetic diagram of Fig. 1.

In this diagram, *real world* means any real life problem in any possible field. For the sake of exposition, I will refer to one domain that is attracting increasing interest for numerical simulation, that of computational medicine, and, more specifically, to the mathematical model and numerical simulation of the behavior of the human heart. The dream is that one day a virtual version of a human heart may help medical doctors diagnose heart disease and determine the best treatment for a specific patient, without the need for unnecessary invasive clinical practices.

First principles are called into play for the set up of the individual core cardiac models, represented by: the electrophysiology (the process that drives the rhythm of the heart), the passive and active mechanics of the cardiac muscle (that determines contraction and dilation of the myocardium), the microscopic force generation in sarcomeres (the basic contractile units of the cardio-myocytes), the blood flow in the heart chambers (two ventricles and two atria), and the dynamics of the four (tricuspid, pulmonary, mitral and aortic) valves. The coupling of these models through suitable transmission conditions that express dynamic and kinematic interactions yield the global mathematical model (Quarteroni et al., 2017a, b).

Data are essential to "close up" this system of differential equations. They express: the shape of the specific heart at hand (these geometrical data are extracted from medical images); initial conditions on velocity and pressure of blood in

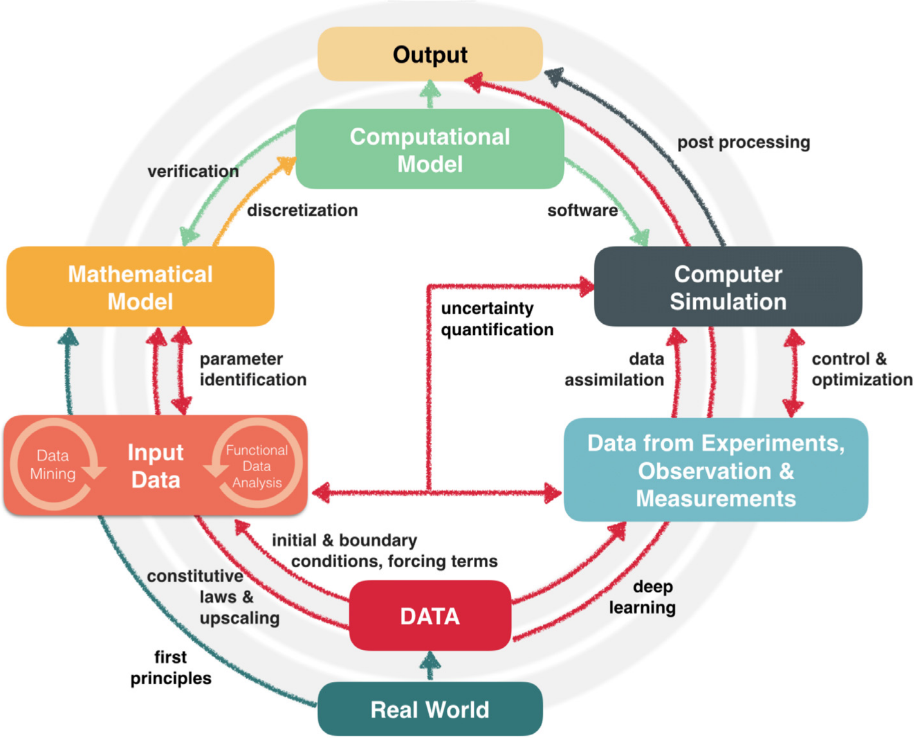

Fig. 1. The interplay between data science and model-based scientific computing.

the cavities, displacements of the myocardium, etc., that are necessary to represent the state of the dynamical system at the initial time of the simulation; the boundary conditions describing the state of the system at its external boundaries; the body forces acting on the system (e.g. the pre-stress regime characterizing the myocardium); the coefficients that characterize the patient specific tissue properties as well as the distribution of the fibers and collagen sheets composing the myocardium.

MRI (magnetic resonance imaging) and CT (computer tomography) play a major role for acquisition of cardiac images. Often, due to heart motion, temporally resolved acquisitions are performed, providing up to 20–30 frames per heartbeat. The main interest is in the left ventricle, due to its vital importance and its pronounced thickness, ranging between 6 and 16 mm. An initial class of ventricle segmentation methods makes use of little (or even no) a priori information. Usually in these methods the endocardium is first segmented by means of thresholding.

Data mining and functional data analysis (Ramsay and Silverman, 1997) are essential tools to "model" these complex data patterns, to filter the intrinsic noise (e.g. by regularization functional involving elliptic differential operators, (Azzimonti et al., 2014) and extract the most meaningful components (or patterns).

Specific automatic methods that have been developed for cardiac image segmentation are based on strong prior information concerning the shape of the ventricles, which is included in the segmentation algorithm by means of statistical models (Frangi et al., 2001). These strategies are suited to cardiac segmentation because variability in heart shape among patients, as opposed to arteries, is very small in normal conditions. Statistical model-based segmentation strategies rely on identifying an average shape of available geometries forming a training set, and modeling the variability within the latter. This is usually done by means of principal component analysis of positions and, if needed, displacements, allowing computation of the eigenvalues and eigenvectors of the covariance matrix related to the training set. These strategies allow automatic segmentation without user intervention, at the expense of needing a training set. For example, deformable models have been extended to this framework by adding a term to the functional to be minimized that penalizes the distance to a reference model (e.g. the mean shape of the training set). Another very common statistical model-based strategy is atlas-guided segmentation. Given an atlas, that is, an integrated image from multiple segmentations, a registration procedure is performed based on mapping the coordinates of the image under investigation to those of the atlas. This transformation is then applied to the atlas obtaining the final segmentation. The registration process could be based on non-rigid transformations that account for elastic deformations.

Cardiac image segmentation represents a noticeable instance of mutual interaction of statistical methods with model based scientific computation. See Petitjean and Dacher (2011) for a review.

To correctly model electrical propagation and mechanical contraction in the myocardium, fiber orientation has to be accurately described. Indeed, the conduction velocity of the action potential propagation assumes different values along the fibers than in the tangential direction. Moreover, the stretching ability of the myocardium is facilitated along the fiber direction.

Diffusion-tensor-MRI is an MRI technology able to identify fiber orientation, but it is not yet used every day in clinical practice, and it is difficult to apply because of heart movement. An alternative strategy is based on computational generation of the fiber orientation to provide a plausible configuration, for example by means of the solution of a Poisson equation, or by using the unscented Kalman filter. See Sermesant (2016) and references therein, Nagler et al. (2017).

For the mechanical problem involving the muscle region, data that are commonly available include the stresses exerted by the blood on the endocardium of the left ventricle and the endocardial and/or epicardial vessel wall displacements. Stresses are usually obtained from measurements of aortic pressure. Vessel wall displacements can be obtained from dynamic MRI or CT images, yielding 20–30 frames per heartbeat, providing the position of the endocardium and epicardium at multiple times. After suitable post-processing, these techniques can provide an estimate of the vessel wall displacement (and thus velocity) by comparing two consecutive frames. The endocardial vessel wall velocity, thanks to a continuity argument, could also be interpreted as the blood velocity at the interface with the endocardium.

Another measurement that can be easily provided by means of Doppler echocardiographic methods or PC-MR is the flow rate at the mitral and aortic valve orifices. With PC-MR technology, measurement of blood velocity is possible in principle at any point of the ventricles and atria chambers. These data can be used to provide boundary conditions for the core cardiac models.

From a modeling perspective, parameters involved in the cardiac mechanical model depend on the chosen constitutive law. Appropriate up-scaling techniques can however be designed in order for improving the constitutive law in macroscopic models by suitably accounting for data measurements available at the microscopic scale.

Several techniques are available for parameter estimation. Point estimates, relying on either variational or sequential methods, provide optimal least-squares estimates by minimizing a cost functional accounting for the misfit between measured data and state observations. Other techniques yield confidence regions or, more generally speaking, the possibility of characterizing the probability distribution of the unknown parameters provided they are described in terms of random variables; this is the goal of statistical inversion theory relying for example on Bayesian inference.

Since different inputs may have produced the observed outcome, instead of finding the most likely input configuration resulting in the observation performed, one can incorporate all possible information about the unknown inputs that are available prior to the measurement.

In the specific context of cardiovascular modeling, parameter estimation is necessary for model calibration/personalization, for the purpose of diagnosis or treatment. Indeed, parameters that are not directly measurable (e.g. for tissue conductivity or elasticity moduli for arterial vessels) are tuned in such a way that the outcome of the numerical model is able to reproduce patient-specific data. This process can rely on data assimilation. This rather generic term encompasses a wide range of techniques exploited to merge measurements and images into the mathematical model in order to improve numerical simulations. Variational data assimilation methods, or ensemble and unscented Kalman filters in case of dynamical systems, are two common techniques that have proven to be very efficient in cardiovascular models (and, much earlier, in numerical weather forecast).

Although in MBSC outcomes are computed from inputs via a deterministic process, input data are often contaminated by experimental noise, or cannot be fully ascertained. Due to *uncertainty*, computational simulations have to be performed for a set of different parameter configurations. How confidently can results be predicted based on large-scale simulations if the models and data comprise inherent uncertainties? *Sensitivity analysis* has the aim to assess the robustness of the output with respect to variations of uncertain input, and can be regarded as a *forward uncertainty quantification* problem. Investigating the propagation of input uncertainties via computed outputs means evaluating suitable statistics of the outputs (such as expected values, moments, confidence bands), which are functions of the parameters affected by uncertainty. On the other hand, the solution of *optimal control* and *inverse identification* problems also depends on the experimental noise affecting observations and measurements used during the identification process, or the set-up of a desired target. Evaluating uncertainties in this case, providing suitable confidence intervals for the estimated quantities (not simply point estimates) and characterizing the statistical distribution of the unknown parameters, all represent *inverse uncertainty quantification* problems. In this second case, quantifying uncertainties is even more important because an inverse problem is intrinsically ill-posed. New methods have recently been developed that build on statistical techniques such as *Monte Carlo methods*, *Bayesian inference*, and *Markov decision processes*. All these approaches require the repeated, iterative solution of the state system, the ensemble of partial differential equations that model the complex cardiac dynamics based on first principles.

4. Conclusions

Data science and computational science share a common mathematical core. They are both rooted in solid foundations of algorithms, linear and nonlinear algebra, statistics, computer science, and deep knowledge of specific domains. This common core is already exploited in the modern paradigm of MBSC, as it emerges from the scheme of Fig. 1. Statistical models operating on complex data sets are, in general, more efficient when they can exploit an insightful knowledge of the "physical process" that has generated the data, thanks to existing accurate and well-calibrated computational models. Educational programs that will shape the new generations of computational and data scientists should be inspired by this synergetic interplay (Baker et al., 2010).

Acknowledgments

I would like to thank my colleagues A. Manzoni and C. Vergara for many stimulating discussions on this topic.

References

Anderson, C., 2008. The end of theory: the data deluge makes the scientific method obsolete. Wired, https://www.wired.com/2008/06/pb-theory/ (access: 17.05.17).

Azzimonti, L., Nobile, F., Sangalli, L.M., Secchi, P., 2014. Mixed finite elements for spatial regression with PDE penalization. SIAM/ASA J. Uncertain. Quant. 2014.

Baker, S., et al., 2010. 2016. Data-enabled science in the mathematical and physical sciences, National Science Foundation, http://www.nsf.gov/mps/dms/documents/DataEnabledScience.pdf.

Candes, E., 2017. Around the reproducibility of scientific research in the big data era: what statistics can offer? J.L. Lions Colloquium, UPMC, Paris, 17 March 2017 https://www.ljll.math.upmc.fr/contenu/article/lecons-j-l-lions-2017-17-03-2017-14h00-colloquium-e-candes.

Dean, J., 2016. Large-Scale Deep Learning for Intelligent Computer Systems, WSDM2016 (http://www.wsdm-conference.org/2016/slides/WSDM2016-Jeff-Dean.pdf).

Frangi, A.F., Niessen, W.J., Viergever, M.A., 2001. Three-dimensional modeling for functional analysis of cardiac images: A review. IEEE Trans. Med. Imaging 20 (1).

Hey, T., Tansley, S., Tolle, K., 2009. Jim gray on eScience: a Transformed scientific method. In: Hey, T., Tansley, S., Tolle, K. (Eds.), The Fourth Paradigm, Data Intensive Scientific Discovery. Microsoft Research, Redmond, Washington, pp. xvii–xxxi.

King, G., 2014. Restructuring the social sciences: Reflections from harvard's institute for quantitative social science. PS: Political Sci. Polit. 47, 165–172.

Nagler, A., Bertoglio, C., Stoeck, C.T., Kozerke, S., Wall, W.A., 2017. Maximum likelihood estimation of cardiac fiber bundle orientation from arbitrarily spaced diffusion weighted images. Med. Image Anal. 39, 56–77.

Petitjean, C., Dacher, J.-N., 2011. A review of segmentation methods in short axis cardiac MR images. Med. Image Anal. 15, 169–184.

Quarteroni, A., Lassila, T., Rossi, S., Ruiz-Baier, R., 2017a. Integrated heart-coupling multiscale and multiphysics models for the simulation of the cardiac function. Comput. Methods Appl. Mech. Engrg. 314, 345–407.

Quarteroni, A., Manzoni, A., Vergara, C., 2017b. The cardiovascular system: mathematical modelling, numerical algorithms and clinical applications. Acta Numer. 25, 365–590.

Ramsay, J.O., Silverman, B.W., 1997. Functional Data Analysis. Springer.

Rüde, U., Willcox, K., Curfman McInnes, L., De Sterck, H., Biros, G., Bungartz, H., Corones, J., Cramer, E., Crowley, J., Ghattas, O., Gunzburger, M., Hanke, M., Harrison, R., Heroux, M., Hesthaven, J., Jimack, P., Johnson, C., Jordan, K.E., Keyes, D.E., Krause, R., Kumar, V., Mayer, S., Meza, J., Mørken, K.M., Oden, J.T., Petzold, L., Raghavan, P., Shontz, S.M., Trefethen, A., Turner, P., Voevodin, V., Wohlmuth, B., Woodward, C.S., 2016. Research and education in computational science and engineering. SIAM Rev. September 2016 arXiv:1610.02608, [cs.CE], submited for publication see http://wiki.siam.org/siag-cse.

Sermesant, M., 2016. When Cardiac Biophysics Meets Groupwise Statistics: Complementary Modelling Approaches for Patient-Specific Medicine. Signal and Image Processing. Université de Nice - Sophia Antipolis.