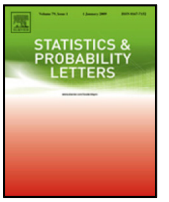
ARTICLE IN PRESS

Statistics and Probability Letters xx (xxxx) xxx–xxx

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 xxxx

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@epfl.ch.

https://doi.org/10.1016/j.spl.2018.02.047
0167-7152/

Please cite this article in press as: Quarteroni A., The role of statistics in the era of big data: A computational scientist' perspective. Statistics and Probability Letters (2018), https://doi.org/10.1016/j.spl.2018.02.047.

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

Please cite this article in press as: Quarteroni A., The role of statistics in the era of big data: A computational scientist' perspective. Statistics and Probability Letters (2018), https://doi.org/10.1016/j.spl.2018.02.047.