Introduction

Enrico Camporeale[*], *Simon Wing*[†], *Jay R. Johnson*[‡]

[*]Centrum Wiskunde & Informatica, Amsterdam, The Netherlands [†]Johns Hopkins University, Laurel, MD, United States [‡]Andrews University, Berrien Springs, MI, United States

A common goal of scientific disciplines is to understand the relationships between observable quantities and to construct models that encode such relationships. Eventually any model, and its supporting hypothesis, needs to be tested against observations—the celebrated Popper's falsifiability criterion (Popper, 1959). Hence, experiments, measurements, and observations—in one word *data*—have always played a pivotal role in science, at least since the time of Galileo's experiment dropping objects from the leaning tower of Pisa.

Yet, it is only in the last decade that libraries' bookshelves have started to pile up with books about the data revolution, big data, data science, and various modifications of these terms. While there is certainly a tendency both in science and publishing to re-brand old ideas and to inflate buzzwords, one cannot deny that the unprecedented large amount of collected data of any sort—be it customer buying preferences, health and genetic records, high energy particle collisions, supercomputer simulation results, or of course, space weather data—makes the time we are living in unique in history. The discipline that benefits the most from the explosion of the *data revolution* is certainly machine learning. This field is traditionally seen as a subset of artificial intelligence, although its boundaries and definition are somehow blurry. For the purposes of this book, we broadly refer to machine learning as the set of methods and algorithms that can be used for the following problems: (1) make predictions in time or space of a continuous quantity (regression); (2) assign a datum to a class within a prespecified set (classification); (3) assign a datum to a class within a set that is determined by the algorithm itself (clustering); (4) reduce the dimensionality of a dataset, by exposing relationships among variables; and (5) establish linear and nonlinear relationships and causalities among variables.

Machine learning is in its golden age today for the simple reason that methods, algorithms, and tools, studied and designed during the last two decades (and sometimes forgotten), have started to produce unexpectedly good results in the last 5 years, exploiting the historically unique combination of big data availability and cheap computing power.

The single methodology that has been popularized the most by nonspecialist media as the archetype of machine learning's groundbreaking promise is probably the massive multilayer neural network, which is often referred to as *deep learning* (LeCun et al., 2015). For instance, deep learning is the technology behind the recent successes in image and speech recognition (with the former recently achieving better-than-human accuracy; He et al., 2015) and the first computer ever defeating a world champion in the game of Go (Silver, 2016).

The popular media often focus on the technological applications of machine learning, which has propelled recent advances in many areas, such as self-driving cars, online fraud detection, personalized advertisement and recommendation, real-time translation, and many others (Bennett and Lanning, 2007; Sommer and Paxson, 2010; Guizzo, 2011). However, we believe that it makes sense to ask whether machine learning could even change the process of scientific discovery.

Looking specifically at physics, the process of developing a model often relies on some form of the well-known Occam's razor: the simplest model that can explain the data is preferred. As a consequence, an important characteristic of most physics models is that every step of the process that led to their development is completely intelligible by the human mind. Such models are referred to as *white-box* models, suggesting that each component (including the set of assumptions) is transparent. Despite its marvelous achievements, the human brain has a very limited ability to process data, especially in high dimensions. This might be trivially related to the fact that the basic way of understanding data is graphical, and it is hard to visualize more than three variables in a single plot. Hence, the relationships between observable quantities that are encoded in white-box physics models usually do not explore high dimensional spaces. This human limitation does not mean that such models are "simple"; on the contrary they can be quite complicated, sometimes requiring formidable numerical methods to produce results that can be compared against observations. Essentially, all first-principles physics models are white-box models.

Contrary to the modus operandi of the white boxes (one could perhaps say of the human mind), machine learning algorithms focus essentially on two characteristics: being accurate and being robust against new data (i.e., being able to generalize). Indeed, the guiding principle concerns the trade-off between complexity and accuracy to avoid overfitting (see Chapter 4).

Hence, in contrast to white-box models, machine learning methods are often referred to as *black-box*, signifying that the mathematical structure and the relationships between variables are so complicated that it is often not useful to try to understand them, as long as they deliver the expected results. For example, and referring again to deep learning, one can certainly unroll a neural network to the point of deriving a single closed formula that relates inputs and outputs. However, such a formula would generally be incomprehensible and completely useless from a science-based perspective, although some features may be related to physical processes.

We need to mention a third, in-between paradigm, obviously called *gray-box modeling* that has recently emerged. Whereas white-box models are accurate but computationally slow (often much slower than real time when it comes to forecasting), and black-box models are fast but very sensitive to noise and outliers, the idea of gray box is to employ reduced physics models, and to calibrate the assumptions or the free parameters of the models via machine learning techniques. Gray box is often used in engineering modeling, and it is gradually making its way into more fundamental physics. In particular, we believe that the skepticism that surrounds machine learning in certain physics communities will be eventually overcome by embracing gray-box models, which allow the use of prior physical information in a more transparent way.

MACHINE LEARNING AND SPACE WEATHER

Space weather is the study of the effect of the Sun's variability on Earth, on the complex electromagnetic system surrounding it, on our technological assets, and eventually on human life. It will be more clearly introduced in Chapter 1, along with its societal and economic importance.

This book presents state-of-the-art applications of machine learning to the space weather problem. Artificial intelligence has been applied to space weather at least since the 1990s. In particular, several attempts have been made to use neural networks and linear filters for predicting geomagnetic indices and radiation belt electrons (Baker, 1990; Valdivia et al., 1996; Sutcliffe, 1997; Lundstedt, 1997, 2005; Boberg et al., 2000; Vassiliadis, 2000; Gleisner and Lundstedt, 2001; Li, 2001; Vandegriff, 2005; Wing et al., 2005). Neural networks have also been used to classify space boundaries and ionospheric high frequency radar returns (Newell et al., 1991; Wing et al., 2003), and total electron content (Tulunay et al., 2006; Habarulema et al., 2007). A feature that makes space weather very remarkable and perfectly posed for machine learning research is that the huge amount of data is usually collected with taxpayer money and is therefore publicly available. Moreover, the released datasets are often of very high quality and require only a small amount of preprocessing. Even data that have not been conceived for operational space weather forecasting offer an enormous amount of information to understand processes and develop models. Chapter 2 will dwell considerably on the nature and type of available data.

In parallel to the above-mentioned machine learning renaissance, a new wave of methods and results have been produced in the last few years, which is the rationale for collecting some of the most promising works in this volume.

The machine learning applications to space weather and space physics can generally be divided into the following categories:

- *Automatic event identification*: Space weather data is typically imbalanced, with many hours of observations covering uninteresting/quiet times, and only a small percentage of data of useful events. The identification of events is still often carried out manually, following time-consuming and nonreproducible criteria. As an example, techniques such as convolutional neural networks can help in automatically identifying interesting regions like solar active regions, coronal holes, coronal mass ejections, and magnetic reconnection events, as well as to select features.
- *Knowledge discovery*: Methods used to study causality and relationships within highly dimensional data, and to cluster similar events, with the aim of deepening our physical understanding. Information theory and unsupervised classification algorithms fall into this category.
- *Forecasting*: Machine learning techniques capable of dealing with large class imbalances and/or significant data gaps to forecast important space weather events from a combination of solar images, solar wind, and geospace in situ data.
- *Modeling*: This is somewhat different from forecasting and involves a higher level approach where the focus is on discovering the underlying physical and long-term behavior of the system. Historically, this approach tends to develop from reduced descriptions based on first principles, but the methods of machine learning can in theory also be used to discover the nonlinear map that describes the system evolution.

We will certainly see increasing applications of machine learning in space physics and space weather, falling in one of these categories. Yet, we also believe it is still an open question whether the amount and the kind of data at our disposal today is sufficient to train accurate models.

SCOPE AND STRUCTURE OF THE BOOK

The aim of this book is to bridge the existing gap between space physicists and machine learning practitioners. On one hand, standard machine learning techniques and off-the-shelf available software are not immediately useful to a large part of the space physics community that is not familiar with the jargon and the potential use of such methods; on the other hand, the data science community is eager to apply new techniques to challenging and unsolved problems with a clear technological impact, such as space weather.

The first part of the book is intended to provide some context to the latter community which might not be familiar with space weather forecasting. Chapter 1 summarizes the *Societal and Economic Importance of Space Weather*, while Chapter 2 describes the *Data Availability and Forecast Products for Space Weather.*

The second part offers a short, high-level overview of the three main topics that will be discussed throughout the book: *Information Theory* (Chapter 3), *Regression* (Chapter 4), and *Classification* (Chapter 5). Obviously, we refer the reader to more specific textbooks for in-depth explanation of these concepts.

The last part is devoted to applications covering a broad range of subdomains.

Chapter 6, *Untangling the Solar Wind Drivers of Radiation Belt: An Information Theoretical Approach*, is concerned with an application of information theory to study the classical problem of discerning different solar wind input parameters and quantifying their different roles in driving the radiation belt electrons.

Chapter 7, *Emergence of Dynamical Complexity in the Earth's Magnetosphere*, tackles the Earth's magnetosphere complexity from the standpoint of system science, studying classical concepts such as scale-invariance, self-similarity, and multifractality in the context of the analysis of time series of geomagnetic data.

Chapter 8, *Application of NARMAX to Space Weather*, reviews the several uses of the methodology based on Nonlinear AutoRegressive Moving Average with eXogenous inputs models to space weather, focusing on geomagnetic indices and radiation belt electrons.

Chapter 9, *Probabilistic Forecasting of Geomagnetic Indices Using Gaussian Process Models*, presents an application of Gaussian process (GP) regression with a particular emphasis on model selection and design choice. GP can be understood in the context of Bayesian inference, and it is a particularly promising tool for space weather prediction, for its natural ability to provide probabilistic forecasts.

Chapter 10, *Prediction of MeV Electron Fluxes With Autoregressive Models*, focuses on relativistic electrons in the radiation belts and on relevant forecasting verification techniques for autoregressive models. The approach employed in this chapter represents a nice example of a gray-box modeling discussed earlier.

Chapter 11, *Artificial Neural Network for Magnetospheric Conditions*, discusses an application of feed-forward neural networks to the problems of electron density estimation in the radiation belt and the specification of waves and flux properties.

Chapter 12, *Reconstruction of Plasma Electron Density From Satellite Measurement Via Artificial Neural Networks*, is also concerned with the study of radiation belt electron density via neural networks, although using a completely different approach to derive input features, and emphasizing model selection and verification.

Chapter 13, *Classification of Magnetospheric Particle Distribution Using NN*, tackles an unsupervised multicategory classification problem: clustering particle distribution in pitch-angle from Van Allen Probes data. The machine learning method chosen for this task is a class of neural networks called self-organizing map.

Chapter 14, *Machine Learning for Flare Forecasting*, discusses the recent progresses in solar flare forecasting, comparing several types of machine learning algorithms, and some relevant computing aspects.

Chapter 15, *Coronal Holes Detection Using Supervised Classification*, presents results on the problem of coronal holes detection, comparing different techniques including support vector machine and decision trees. The chapter has a useful hands-on approach, with a direct link to MATLAB software available on the author's website.

Finally, Chapter 16, *Solar Wind Classification Via the K-Means Clustering*, presents an unsupervised clustering technique to divide the solar wind in different types, based on their characteristics measured by instruments on the Advanced Composition Explorer.

In conclusion, we believe that this book provides an up-to-date portrait of some state-of-the-art applications of machine learning to space weather. However, some important works have inevitably been left out. In particular, we would like to mention the recent progress in the prediction of solar flares and coronal mass ejections using Solar Dynamic Observatory data via support vector machine and automatic feature extraction (Bobra and Couvidat, 2015; Muranushi, 2015; Bobra and Ilonidis, 2016; Jonas et al., 2017); the use of data assimilation (Koller et al., 2007; Shprits et al., 2007; Arge et al., 2010; Innocenti et al., 2011; Godinez et al., 2016; Lang et al., 2017); and uncertainty quantification and ensemble techniques (Schunk, 2014; Guerra et al., 2015; Knipp, 2016; Camporeale, 2016).

ACKNOWLEDGMENTS

The authors would like to thank several colleagues that have helped in reviewing the chapters: George Balasis, Shaun Bloomfield, Monica Bobra, Joe Borovsky, Jacob Bortnik, Algo Carè, Michele Cash, Veronika Cheplygina, Xiangning Chu, Gregory Cunningham, Rob Decker, Veronique Delouille, Mariusz Flasinski, Simone Garatti, Manolis Georgoulis, Larisza Krista, Naoto Nishizuka, Juan Valdivia, Shawn Young, and Chao Yue.

The authors would like to thank Mathworks for providing complimentary MATLAB licenses to some of the authors.

This work was partially supported by the NWO-VIDI grant 639.072.716 and NASA grants (NNX15AJ01G, NNX16AR10G, and NNX16AQ87G).

References

Arge, C.N., et al., 2010. Air force data assimilative photospheric flux transport (ADAPT) model. In: AIP Conference Proceedings, vol. 1216, 1.

Baker, D.N., 1990. Linear prediction filter analysis of relativistic electron properties at 6.6 RE. J. Geophys. Res. Space Phys. 95 (A9), 15133–15140.

Bennett, J., Lanning, S., 2007. The netflix prize. In: Proceedings of KDD Cup and Workshop.

Boberg, F., Wintoft, P., Lundstedt, H., 2000. Real time Kp predictions from solar wind data using neural networks. Phys. Chem. Earth Part C 25 (4), 275–280.

Bobra, M.G., Couvidat, S., 2015. Solar flare prediction using SDO/HMI vector magnetic field data with a machine-learning algorithm. Astrophys. J. 798 (2), 135.

Bobra, M.G., Ilonidis, S., 2016. Predicting coronal mass ejections using machine learning methods. Astrophys. J. 821 (2), 127.

Camporeale, E., 2016. On the propagation of uncertainties in radiation belt simulations. Space Weather 14 (11), 982–992.

Gleisner, H., Lundstedt, H., 2001. A neural network-based local model for prediction of geomagnetic disturbances. J. Geophys. Res. Space Phys. 106 (A5), 8425–8433.

Godinez, H.C., et al., 2016. Ring current pressure estimation with RAM-SCB using data assimilation and Van Allen Probe flux data. Geophys. Res. Lett. 43 (23), 11948–11956.

Guerra, J.A., Pulkkinen, A., Uritsky, V.M., 2015. Ensemble forecasting of major solar flares: first results. Space Weather 13 (10), 626–642.

Guizzo, E., 2011. How Google's self-driving car works. In: IEEE Spectrum, vol. 18.

Habarulema, J.B., McKinnell, L.A., Cilliers, P.J., 2007. Prediction of global positioning system total electron content using neural networks over South Africa. J. Atmos. Sol. Terr. Phys. 69 (15), 1842–1850.

He, K., et al., 2015. Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE International Conference on Computer Vision.

Innocenti, M.E., et al., 2011. Improved forecasts of solar wind parameters using the Kalman filter. Space Weather 9, S10005.

Jonas, E., et al., 2017. Flare prediction using photospheric and coronal image data. ArXiv preprint arXiv:1708.01323.

Knipp, D.J., 2016. Advances in space weather ensemble forecasting. Space Weather 14 (2), 52–53.

Koller, J., et al., 2007. Identifying the radiation belt source region by data assimilation. J. Geophys. Res. Space Phys. 112, A06244.

Lang, M., et al., 2017. Data assimilation in the solar wind: challenges and first results. Space Weather 15, 1490–1510.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436–444.

Li, X., 2001. Quantitative prediction of radiation belt electrons at geostationary orbit based on solar wind measurements. Geophys. Res. Lett. 28 (9), 1887–1890.

Lundstedt, H., 1997. AI techniques in geomagnetic storm forecasting. In: Tsurutani, B.T., Gonzalez, W.D., Kamide, Y., Arballo, J.K. (Eds.), Magnetic Storms. American Geophysical Union, Washington, D.C.

Lundstedt, H., 2005. Progress in space weather predictions and applications. Adv. Space Res. 36 (12), 2516–2523.

Muranushi, T., 2015. UFCORIN: a fully automated predictor of solar flares in GOES X-ray flux. Space Weather 13 (11), 778–796.

Newell, P.T., Wing, S., Meng, C.I., Sigillito, V., 1991. The auroral oval position, structure and intensity of precipitation from 1984 onwards: an automated on-line data base. J. Geophys. Res. 96, 5877–5882.

Popper, K., 1959. The Logic of Scientific Discovery. Routledge, London.

Schunk, R.W., 2014. Ensemble modeling with data assimilation models: a new strategy for space weather specifications, forecasts, and science. Space Weather 12 (3), 123–126.

Shprits, Y., et al., 2007. Reanalysis of relativistic radiation belt electron fluxes using CRRES satellite data, a radial diffusion model, and a Kalman filter. J. Geophys. Res. Space Phys. 112, A12.

Silver, D., 2016. Mastering the game of go with deep neural networks and tree search. Nature 529 (7587), 484–489.

Sommer, R., Paxson, V., 2010. Outside the closed world: on using machine learning for network intrusion detection. In: 2010 IEEE Symposium on Security and Privacy (SP).

Sutcliffe, P.R., 1997. Substorm onset identification using neural networks and Pi2 pulsations. Ann. Geophys. 15 (10), 1257–1264.

Tulunay, E., et al., 2006. Forecasting total electron content maps by neural network technique. Radio Sci. 41 (4), RS4016.

Valdivia, J.A., Sharma, A.S., Papadopoulos, K., 1996. Prediction of magnetic storms by nonlinear models. Geophys. Res. Lett. 23 (21), 2899–2902.

Vandegriff, J., 2005. Forecasting space weather: predicting interplanetary shocks using neural networks. Adv. Space Res. 36 (12), 2323–2327.

Vassiliadis, D., 2000. System identification, modeling, and prediction for space weather environments. IEEE Trans. Plasma Sci. 28 (6), 1944–1955.

Wing, S., et al., 2003. Neural networks for automated classification of ionospheric irregularities from HF radar backscattered signals. Radio Sci. 38 (4), 1–8.

Wing, S., et al., 2005. Kp forecast models. J. Geophys. Res. Space Phys. 110, A4.