*Abstract for the **Data Science & Social Research** Conference*
University of Napoli Federico II , February 17-19, 2016

Predictive versus Generative Modelling: a Challenge for (Social) Sciences

G. Saporta
CEDRIC, Conservatoire National des Arts et Métiers, Paris
gilbert.saporta@cnam.fr

Key Words: black-box models, data science, inference, machine learning

We carry on the reflexion initiated in [1]. Donoho [2], reinterpreting [3], notes that most developments in academic statistics are oriented towards inference: ie fitting, estimating and validating generative models, preferably parcimonious, rather than for prediction where "*unfortunately, accuracy and simplicity (interpretability) are in conflict*". Classical inference corresponds to the role of statistics as an auxiliary of sciences. However in most sciences, a good model should also provides accurate predictions, which becomes the sole criterium in decision sciences like pattern recognition, customer behaviour, etc. The most efficient predictive models are rather black-box algorithms like random forests or deep learning.
Some consider that "*statistics is the least important part of data science*" [4] while others claim that even science is obsolete [5]! Meanwhile, renowned scientists [6] are calling practitioners of social sciences to study Machine Learning . The use of black-box models fitted for massive data is probably the main challenge for social sciences due to their lack of interpretability. Getting better predictions, thanks to a better understanding of the real world, needs to combine statistics and machine learning with causal inference.

References:

[1] G.Saporta (2008). Models for Understanding versus Models for Prediction, In P.Brito, ed., Compstat Proceedings, Physica Verlag, 315-322

[2] D.Donoho (2015). 50 years of Data Science, Tukey Centennial workshop, https://dl.dropboxusercontent.com/u/23421017/50YearsDataScience.pdf

[3] L.Breiman (2001). Statistical Modeling: The Two Cultures, Statistical Science, 16, 3, 199–231

[4] A.Gelman (2013). http://andrewgelman.com/2013/11/14/statistics-least-important-part-data-science/

[5] C.Anderson (2008). The End of Theory: The Data Deluge Makes the Scientific Method Obsolete, http://www.wired.com/2008/06/pb-theory/

[6] H.Varian (2014). Big Data: New Tricks for Econometrics, Journal of Economic Perspectives, 28, 2, 3–28