The Great Regression. Machine Learning, Econometrics, and the Future of Quantitative Social Sciences

Julien Boelaert, Etienne Ollion

The Great Regression

Machine Learning, Econometrics, and the Future of Quantitative Social Sciences

Julien Boelaert : Post-doctorant iPOPs (SAGE, Strasbourg)
Etienne Ollion : Chercheur CNRS (SAGE, Strasbourg) et chercheur associé à l'université de Linköping

Summary
What can machine learning do for (social) scientific analysis, and what can it do to it? A contribution to the emerging debate on the role of machine learning for the social sciences, this article offers an introduction to this class of statistical techniques. It details its premises, logic, and the challenges it faces. This is done by comparing machine learning to more classical approaches to quantification – most notably parametric regression– both at a general level and in practice. The article is thus an intervention in the contentious debates about the role and possible consequences of adopting statistical learning in science. We claim that the revolution announced by many and feared by others will not happen any time soon, at least not in the terms that both proponents and critics of the technique have spelled out. The growing use of machine learning is not so much ushering in a radically new quantitative era as it is fostering an increased competition between the newly termed classic method and the learning approach. This, in turn, results in more uncertainty with respect to quantified results. Surprisingly enough, this may be good news for knowledge overall.

Résumé
Que peut faire l'apprentissage automatique (*machine learning*) pour les sciences sociales, et que peut-il lui faire ? Cet article propose une introduction à cette classe de méthodes statistiques. Il détaille ses prémisses, sa logique, et les défis qu'elle pose pour les sciences (sociales). Il le fait au moyen d'une comparaison avec d'autres approches quantitative plus conventionnelles, les régressions paramétriques en premier lieu, et ce tant au niveau général qu'en pratique. Au-delà de l'exercice méthodologique, l'article se propose de revenir sur débats houleux qui entourent le *learning*. Il revient pour se faire sur le rôle et les conséquences possibles de l'usage de l'apprentissage statistique. Il soutient que la révolution promise par beaucoup et crainte par d'autres ne se produira pas de sitôt, ou en tout cas pas dans les termes souvent évoqués. Le changement de paradigme évoqué de manière prophétique n'aura pas lieu. Plutôt, une concurrence accrue entre différentes formes de quantification du monde social va se mettre en place. Contre toute attente, cette incertitude croissante pourrait être de bon augure pour la connaissance en général.

Jorge Luis Borges once described a band of monkeys locked up in a room solely furnished with typewriters[1]. Given a few eternities, he argued, the primate's random strokes on the keyboards would produce not only meaningful sentences, but also world-class literature and scientific discoveries. “Everything”, the Argentine novelist wrote, “would be in its blind volumes. Everything: the detailed history of the future, Aeschylus' *The Egyptians*, the exact number of times that the waters of the Ganges have reflected the flight of a falcon, the secret and true name of Rome.” And, of course, the complete collected writings of Shakespeare. Facts and fiction alike would be present in this utopian “total library” (Borges, 1939).

This image of a battalion of immortal typist monkeys haphazardly trying out numerous combinations in an attempt to solve a given problem is probably how many view machine learning today. In the view of its critics, a statistical method that tries to replace sound mathematical reasoning by sheer brute computation is about as likely to yield interesting results as an army of apes randomly stroking keyboards[2]. Machine learning proponents would certainly disagree with this characterization. They would for instance rightfully argue that their algorithms do not work at random. These algorithms certainly exploit randomness, but they do so while following very specific rules that allows them, contrary to Borges’ monkeys, to learn from their mistakes. The fact remains, however, that machine learners do tend to embrace some form of radical inductivism. As their argument goes, a learning machine can, through clever use of intense computation and with virtually no pre-established conceptions about the data at hand, efficiently uncover its hidden structure, and indeed detect complex correlations that would take mere human statisticians ages to find.

These debates are nothing new, echoing classic oppositions about the primacy of data vs. models, or about induction vs. deduction. Almost twenty years ago, a provocative paper on the “two cultures” of statistics sparked an intense conversation within the statistical community. Written by machine learning figurehead Leo Breiman, the text made a powerful case for the “algorithmic” culture while trying to respond to the main criticisms leveled against it (see Breiman, 2001 and its rejoinders). While this debate is still raging, the context has changed: over time, the machine learning approach has led to such important successes that it can no longer be dismissed on purely theoretical grounds. In the recent years, learning (or artificial intelligence) algorithms have written important pages in the history of science and technology. Disease detection, face recognition, real-time translation in multiple languages: all rely on some machine learning algorithm, and all perform their task better, or at least faster, than any human. Today, the apes are indeed writing Shakespeare-like verses, and driving themselves around the streets of various cities.

These recent successes owe much to the rise in computational power and the growing availability of data, which have done much to increase the standing of machine learning. Moreover, they occur at a time of crisis for statistics, on two separate fronts. The more conventional statistical methods, with probabilistic hypotheses designed in the early twentieth century to deal with small random samples, are drowning in the rising tide of large, non-

1 The authors are grateful to Jean-Yves Bart, Marie Bergström, Arthur Charpentier, Sébastien Chauvin, Illias Garnier, Satu Helske, Estiatorio Margaritas, Jean-Philippe Touffut, the participants of the IAS lectures series (Linköping university), of the CREST (ENSAE) workshop, of the Berkeley sociology colloquium series, and of the University of Chicago Mixed Methods workshop. We are also grateful to the editorial board of the *Revue Française de Sociologie* for insightful comments. The data was accessed via Linköping University and paid for by several grants, including ERC grant agreement no 324233, Riksbankens Jubileumsfond (DNR M12-0301:1), and the Swedish Research Council (DNR 445-2013-7681 and DNR 340-2013-5460) and the Excellence Initiative of the University of Strasbourg.

2 In fact, when the experiment was effectively carried out in the 2000s, the results turned out to be quite disappointing. In thirty days, the six macaques locked up with a computer only produced five pages of text, mostly made of “s.” After a short while, the lead animal started bashing the keyboard with a rock, while others used it as a lavatory. See http://news.bbc.co.uk/2/hi/3013959.stm, retrieved on February 25th, 2018.

random and dirty datasets. Additionally, statisticians themselves are increasingly vocal in their fundamental critique of the many ill uses of the tools that have come to dominate large swaths of quantitative science - linear regressions and hypothesis testing (see Wasserstein and Lazar, 2016). The time may thus seem ripe for a rapid take-over of science by machine learning algorithms. And indeed, the idea that they will deeply transform numerous disciplines is now regularly evoked. Famed computer scientist Pedro Domingos, for one, recently wrote that machine learning "follows the same procedure [as classic statistics] of generating, testing, and discarding or refining hypotheses. But while a scientist may spend her whole life [doing so], machine learning can do it in a fraction of second" (Domingos, 2015, p. 13). Machine learning, he concluded, is "the scientific method on steroids". To him, it is thus "no surprise that it is revolutionizing science" (*ibid.*).

For machine learning to offer meaningful results in science, however, it will take more than just automation or computer power. While its achievements in certain areas are beyond dispute, and while scientists are increasingly discussing its relevance for their endeavors, certain central problems remain. The question of the respective roles of data and theory is one. Another is the lack of guarantees that comes with machine learning results: the very flexibility that makes the models so powerful makes it virtually impossible to mathematically demonstrate that their results are optimal. Yet another issue is even more problematic for scientific investigation: the merit of these algorithms is mainly judged in terms of predictive capacities, with little attention given to interpretation. Flexible models may be very accurate, but they are also much harder to interpret than a standard regression. In other words, they may be very good at telling whether a person will do something (buy a product, develop a disease, get married), but they have trouble telling us why. That marketing companies, internet giants or applied disciplines can make profitable use of techniques designed for prediction is understandable. What scientists who are more interested in the explanation of their chosen object can gain from them remains unclear.

A contribution to the ongoing debate on the role of machine learning for (social) science, this article offers an introduction to this class of statistical techniques. It starts by detailing its premises, its logic, and the challenges it currently faces. Because the field is much too diverse to warrant a uniform treatment, the rest of our argument is restricted to one key subclass of algorithms, called supervised learning. These algorithms are of particular interest for quantitative social science because they are aimed at generalizing one of its most dominant and ubiquitous tools, parametric regression[3]. The article systematically compares the two approaches, from a theoretical standpoint first, then on a large-scale dataset. Beyond sheer methodological clarification, the article argues that the statistical revolution announced by many is not likely to happen anytime soon, at least not in the terms spelled out by its proponents. Rather than ushering in a radically new scientific era in quantitative methods, the dissemination of machine learning is fostering an increased competition between the two approaches, one that brings about more uncertainty about quantified results. Surprisingly enough, we argue, this uncertainty may be good news for science overall.

1. What is machine learning?

"Machine learning", "artificial intelligence", "statistical learning", "data mining", "pattern recognition"… the diversity of overlapping denominations reflects both the diversity

3 In all the following, we will refer to *parametric regression* for the wide family of models comprising linear regression, its logistic and Poisson cousins, non-linear least squares and other methods that both rely on an explicit formulation of the model and are interpreted through statistical significance tests. They are parametric in the sense that their main objective is to identify the value of certain parameters, which are hypothesized to play a key role in the phenomenon under study.

of approaches and the diverging research goals of the various scientific communities that have contributed to the development of the field. As an intellectual endeavor, machine learning is probably best described as a crossover between disciplines, statistics and computer science in the first place. From the former, it retains the creation of mathematical models meant to capture the underlying structure of otherwise unmanageable datasets. From the latter, it borrows the skills to produce efficient algorithms for solving complex optimization problems.

Machine learning is currently the most popular form of artificial intelligence ("AI"), a term it is sometimes equated with nowadays. In the early 1950s, the progressive dissemination of digital computers reignited a century-old quest: producing a machine that could emulate human reasoning (Buchanan, 2005), or at least acquire and improve knowledge. A wealth of methods and principles developed over time, most of which can be said to fall into one of the two great families of AI, namely the top-down and the bottom-up approach (Copeland, 2000). The top-down approach is also known as "symbolic AI", because it consists in making computers manipulate symbolic representations of the world using logical rules, in an effort to mimic the way a human would juggle concepts and knowledge using her reason[4]. The bottom-up approach takes the opposite stance: artificial intelligence can be generated from interconnected networks of simple calculating units. This approach, which comprises much of what is called machine learning today, was largely influenced by a theory called "connectionism", in which animal intelligence is thought to emerge from the connections between its biological neurons.

The history is partly technical, and both traditions experienced cyclical episodes of great expectations, followed by years of relative disinterest. Starting in the 1950s, researchers garnered sufficient results to attract interest in AI, both from the scientific community and from a wider audience. This initial golden era came to an end in the late 1960s, when several unsolved issues, such as the impossibility to carry out certain simple tasks or prohibitive computational requirements, started to be considered as unsurpassable. As often exaggerated promises had not been kept, both funding and interest plummeted abruptly, leading to what is known today as the first "AI winter." For over a decade, AI and machine learning were regarded by many as past science fiction (Crevier 1993, p. 203). In the 1980s, a new generation of artificial neural networks called "multilayer perceptrons" overcame these technical difficulties and, although they were quickly overshadowed by more efficient techniques, initiated a revival of the machine learning approach. Neural networks came back again in the 2000s, when so-called deep learning models suddenly gave much improved results when faced with much larger databases; they are currently one of the most fashionable methods in the whole field. Over the years, the progressive refinements of techniques, the rise in computational power and the increase in available data all concurred to help removing bottlenecks that had previously halted the development of AI.

The history of the field is also partly commercial. In the 1950s, artificial intelligence received large subsidies from private companies and public administrations, which allowed blue-sky research. Likewise, the current phase of intense development owes much to the massive investments made by numerous companies, including technology giants like Facebook, Google, Baidu and Microsoft, that dedicate large chunks of their Research & Development budgets to continually improving their standing in this highly competitive domain. Because they have significantly higher resources than universities, and can offer compensations that far exceed any academic salary, they attract numerous qualified researchers, PhD students and seasoned academics alike. For over a decade now, they have

4 Expert systems were a very successful example of symbolic AI, whose many commercial applications in the 1980s included medical diagnosis, chemical analysis and credit authorization.

been a transforming force in the field, sometimes steering research toward their own goals, and blurring the line between academia and industry.

What does machine learning do?

Machine learning is classically defined as "the field of study that gives computers the ability to learn without being explicitly programmed," the art of programming computers to "optimize a performance criterion using example or past data" (Alpaydin, 2010). Most definitions thus stress the role of empirics in the elaboration of the model, and intensive computation. Beyond this point however, there is little agreement about what machine learning is, and even less about what it should be.

One possible response is to consider what the algorithms do. Work in the field of machine learning is divided into several overarching tasks: numerical optimization, function approximation, visualization of multidimensional data, problem solving... Goals and tasks are thus varied, and not all of them are relevant for social science. A few of them are, though, insofar as they pursue the goals of classic statistical techniques with renewed means. One particularly relevant distinction is that between supervised and unsupervised learning. In *supervised learning*, the goal is to predict the values of an outcome variable *y*, based on the values of a set of predictor variables *x*. To achieve a good result, the algorithm should find good approximations to the unknown relationship between the predictors and the outcome variable. This is, of course, a standard task in statistics, called regression in the case of a continuous output (*e.g.* when trying to predict an individual's income based on age, gender, level of education, *etc*), or classification in the case of a categorical output (*e.g.* when predicting success or failure at an exam, based on students' social backgrounds). Expanding on the standard method of linear regression, which has been known since the nineteenth century, many different algorithms tackle this problem from very different angles: feedforward neural networks, support vector machines, decision trees and their extensions such as random forests, are but a few popular examples.

In *unsupervised learning*, there are no *y* values to predict, and instead the focus is on the detection of regularities in a set of *x* variables[5]. This family of methods can be further divided into two subtasks, clustering and dimensionality reduction. Clustering is the grouping of observations into coherent classes, whereas dimensionality reduction consists in mapping a multidimensional dataset in more manageable form, typically two-dimensional plots. Some classic unsupervised algorithms are already part of the standard toolset of quantitative social science: hierarchical clustering and k-means on the clustering side, factorial analysis (correspondence analysis such as it was developed by Benzecri and popularized by Bourdieu) for dimensionality reduction. Over the years, many techniques have been developed to generalize them and overcome some of their limitations: self-organizing maps are one example, dating from the 1980s, blending the clustering and the dimensionality reduction approaches; t-SNE and deep autoencoders are more recent developments, highly publicized in the machine learning literature.

Comparing machine learning to more classic statistical techniques can be difficult, because the frontier between the two is not always clear-cut. For one thing, the main tasks they are asked to perform are largely the same. To add to the confusion, linear regression and PCA are often the first methods taught in machine learning classes. Some strong differences do nonetheless set apart the common econometric use of parametric models from the machine learning approach to prediction. The following section focuses on the differences between these two strands of supervised learning.

5 Whereas in prediction tasks the outcome variable plays the role of a supervising variable, where only the changes in *x* that are relevant to those in *y* must be accounted for, in unsupervised learning all variables are treated equally, as the observations are not "labelled" by an *y* value; hence the name *unsupervised*.

2. Four Salient Differences

Mathematical foundations versus empirical efficacy

The first major difference pertains to the theoretical foundations of the two families of methods. The parametric modeling approach has solid foundations in mathematical statistics and probability theory. This can be readily observed in its methodological articles, which are usually centered around mathematical theorems, and thick with algebraic or probabilistic formulae. Indeed, the overall quality of a parametric method is mainly evaluated on the basis of its theoretical properties. The simple linear regression model, for instance, owes much of its popularity to the fact that its estimator is known to be unbiased, consistent and efficient[6] if its hypotheses are fulfilled.

This stands in sharp contrast with the bulk of the supervised machine learning literature. Surely mathematics are not altogether absent from its papers and manuals, but they are mostly mobilized to explain what the models do, rather than to prove their desirable properties. More generally, its ties to mathematical statistics, and probability theory in particular, are much looser. Indeed, successful machine learning models owe their popularity to their empirical efficacy rather than to their theoretical properties. To put it bluntly, a good supervised learning model is one that gives good predictions on a wide range of empirical problems. Significantly, some of the most widely used algorithms are poorly understood from a mathematical standpoint. Random forests offer a prime example of this surprising fact: while they have been shown to be a highly effective and easy-to-use predictive model in many different settings, the question of just why they work so well remains open, largely due to the difficulty of expressing the workings of the complete algorithm into tractable mathematical formulae (Biau and Scornet, 2016).

This major difference is further refracted into various practical aspects of both types of methods. One striking difference is that parametric models typically have exact, unique, optimal and tractable solutions, while most machine learning methods have to rely on approximate optimization. Given a specified model and a dataset, a parametric regression's parameters are computed using an algebraic formula, which is bound to yield the best possible solution - given the initial hypotheses. Most supervised machine learning models, on the other hand, are much too complex to have a known solution that can be expressed in a mathematical formula. Instead, the search for good parameters - the "training" phase - is algorithmic: it follows a procedure that typically consists in a series of steps that are repeated many times, in order to progressively approach a satisficing solution. The solutions found by machine learning algorithms are thus approximate, and their optimality is rarely guaranteed: there is always a possibility that a better solution exists but wasn't found by the training procedure[7]. In addition to this, the fact that many training procedures are stochastic (*ie* make use of random number generation in their search for a good model) has an unsettling consequence: two consecutive runs of the same algorithm on the same dataset may yield different solutions.

Perhaps the aspect in which supervised machine learning suffers most from its comparative lack of mathematical foundations is in the interpretation of its results, and

6 Following the Gauss-Markov theorem.

7 The fact that most machine learning results cannot be guaranteed to be optimal has one important consequence in practice, in the case of negative results. If a given learning procedure fails to find a good relationship between the predictors and the outcome variable, the failure may either be due to the fact that no meaningful relationship exists in the data, or to a failure of the training algorithm to find a good solution to its intricate optimization problem. Negative results (such as the absence of a relationship) are thus very hard to establish using supervised machine learning tools.

inference[8]. The known theoretical properties of a parametric regression ensure that, once the model is trained, we know not only the value of the estimated parameter (*e.g.* the marginal effect of education level on personal income), but also its theoretical distribution (*e.g.* a normal distribution centered around the true value of the parameter). This makes it possible to evaluate the quality of the estimation, and the extent of its uncertainty. These properties can then be harnessed, usually in the form of inferential tests, to answer the question "is there a statistically significant wage gap between genders, when taking into account the level of education and social background of the studied individuals?" Results from machine learning procedures, on the other hand, rarely grant such definite inference: not much is known of their quality (apart from measures of goodness of fit), so that they can usually not answer such precise questions, or that they give a measure of uncertainty to their answer.

Explicit modeling versus flexible forms

In the supervised setting, both parametric modeling and machine learning share a common goal: they try to model an outcome variable y as a function of predictor variables, x_1, x_2, x_3 so that y is approximated by $y = f(1, x_2, x_3)$. Different specifications of the transformation f are the main distinction between different kinds of models. In the parametric setting, it is almost entirely specified *ex ante*: the model, all the transformations of the variables and their potential interactions must be explicitly provided by the researcher. Fitting these models to the data then consists in tuning a few free parameters so that the predictions are as close as possible to the observed true values of the outcome variable. The model can take a very simple form, such as the familiar "naive" linear model:

$$y = a + b\,x_1 + c\,x_2 + d\,x_3.$$

Alternatively, it can take a much more sophisticated form, and include any number of interactions between variables, or nonlinear (*e.g.* polynomial) transformations of the predictors or parameters.

The specific mathematical form of the model should be guided by a detailed theory about the phenomenon under study (or, in technical terms, the "data-generating process"), and the predictor variables should be chosen on strong empirical or theoretical grounds. In economics for instance, models of individual or collective behavior are supposed to reflect particular specifications of the actors' utility functions. Moreover, it should be entirely defined *ex ante*: the progressive addition or transformation of variables to optimize a criterion (such as p-values) is strongly advised against, a rule that is as strict as it is frequently transgressed. In common statistical practice, decisions as to which variables, transformations or interactions to include or exclude from a regression model are often based on preliminary results (such as descriptive plots or bivariate significance tests), or made by fitting many regression models sequentially. However common it might be, this practice is strictly prohibited by statistical theory, at least if further statistical inference is to be carried out, such as statistical significance tests. The reason for this interdiction is that iterative model construction tends to artificially increase statistical significance. Had the data been even slightly different, the many modeling decisions taken prior to computing the final model might have led to different choices. This uncertainty, however, is not taken into account in the reported standard errors and p-values of statistical software. In other words, a routine and seemingly harmless procedure such as the iterative adjustment of the model frequently leads to artificially improved p-values (a practice sometimes called "p-value hacking").

In contrast, most supervised machine learning procedures start from the opposite stance: they treat the relationship between variables as fully unknown, and try to train a

8 Here the word inference is understood in the general meaning of what can be said about the results, their stability and how they confirm or contradict a given assertion about the data-generating process, see (Efron and Hastie, 2016, p. 3-8).

complex, flexible model to fit the data. Their starting point is that the reality expressed in the data is too complex to be known in advance, and that the model should be built from the data rather than specified *ex ante*. While they do feature a model (a set of mathematical formulae that link the predictors to the outcome), the key aspect of their model is *flexibility*, the capacity to automatically adapt to the data at hand. This is not to say that there are no choices to be made by the researcher when running machine learning algorithms. They are usually governed by two sets of parameters: one that fixes a level of complexity for the model (*i.e.* how far from linearity the model is allowed to go), and one that governs the way in which the algorithm searches for an optimal solution[9].

The flexibility of their models is what allows machine learning procedures to bear the promise of *universal approximation*: if a relationship does exist in the data, then the flexible model is, in theory at least, capable of finding and reproducing it. This allows them to do away with many constraints inherent to parametric modeling, the most obvious being monotonicity. In the simplest form of linear regression (a linear combination of the variables without further transformations or interactions, what we have called a naive model), the effect of each predictor on the outcome is the same for the whole population. While mathematically appealing, this hypothesis is regularly contradicted in empirical analyses. For instance, young women from working-class families and/or from an ethnic minority have long outperformed the males of their social groups at school in France, a situation that was reversed at the other end of the social spectrum (Baudelot and Establet, 1992). Or again, using GSS data from the 1990s, an article showed that sexual orientation used to play a different role respective to earnings, as lesbian women earned 20–30% more than similar heterosexual women (Carpenter, 2006, p. 258) - a fact the author attributed to the more limited (sexual) division of labor in the latter couples. In both cases, the situation may have changed over time, but the idea remains true: that the effect of predictor variables is context-dependent and overall non-linear is the norm rather than the exception in the social sciences.

Many of these limitations can of course be overcome in the parametric setting. One can introduce non-monotonous effects by making the model more complex, *e.g.* by including a squared age variable, or interactions terms between gender and social background. However, all such refinements have to be specified (by hand) by the researcher, as parametric models are not designed to detect anything by themselves. Hence the temptation of trying to build a parametric model step by step[10]. But as stated above, this poses serious problems for further interpretation, especially in terms of p-values. Machine learning solves this problem in a somewhat radical way: p-values are virtually absent from the whole literature.

Validation criteria

These remarks beg the question of the assessment of model quality. Here, too, machine learning and parametric modeling differ. In the latter, quality is mainly assessed through p-values. While goodness-of-fit measures, such as the R^2, AIC or BIC, are sometimes reported, they are mainly used for model comparison, and rarely interpreted for themselves.

9 The sensitivity to these two types of parameters varies widely among different types of machine learning models. Some, such as random forests, need practically no tuning, and can be successfully be run with default parameters in most situations. Others, such as feedforward neural networks, allow (and often require) intensive fine-tuning of both sets of parameters.

10 Some approaches, such as the popular stepwise regression, even automate this trial and error process. This has sometimes been called data-mining, often disparagingly coming from statisticians (Chatfield, 1995 notes that the practice, especially when combined with interpretation of p-values, has been called “logically unsound and practically misleading”, or even “a quiet scandal”). When used carefully and without hypothesis testing, it may be regarded as a primitive and crude form of machine learning (it is typically addressed at the very start of the discipline’s manuals).

P-values, by difference, are central, as they are used to test specific hypotheses. Conventionally, a p-value smaller than 0.05 for some parameter of interest in a regression model is interpreted as proof of the existence of a significant effect. P-values and statistical significance testing have a long history, going back to Neyman and Pearson in the early twentieth century, and have been central to the development of many scientific disciplines as we know them today. Their importance in current research cannot be overstated: in most quantitative research, p-values decide whether a study makes it to print or not.

One of the most disorienting aspects of supervised machine learning for the newcomer is thus that it is devoid of any such metric. P-values and statistical tests are almost altogether absent. Instead, its universal criterion of model quality is prediction. The logic is the following: if the model is able to predict well, then it has probably succeeded in capturing the patterns that link the predictors to the outcome variable. In other words, the ultimate measure of quality is the ability of the model to accurately guess the outcome value. While reminiscent of the common R^2 of linear regression, the measures in machine learning differ in one key aspect: they are computed by *cross-validation*, *i.e.*, on observations that were not used to build the model. This is done in order to prevent overfitting: if the same data are used to train the model and evaluate its quality, there is a high risk that the model will stick too close to that specific dataset, and not generalize well to other data[11].

The simplest way to implement cross-validation is to divide the dataset into two subgroups: the "train set" (typically comprising two thirds of the population's observations) is fed to the algorithm during the training phase, in order to fit the model. The "test set" (the remaining third of the observations) is held out during training, and only used to evaluate the model's quality: after training, the predictors of the test set are fed to the trained model, which yields predicted values that are then compared to the actual values of the outcome variable in the test set. This train/test approach is just one possibility among many, but the principle is clear: the quality of a model is assessed by its power of *generalization*, its ability to correctly predict observations from other samples of the same population.

Analysis of the results

Finally, the type of results and interpretations also vary between the two approaches. With parametric regressions, the results mostly consist in interpreting the estimated parameters and their standard errors. Four main types of interpretation can be drawn from such models: the statistical significance of an effect ("controlling for other predictors, gender has a significant effect on wages, at the 5% level"), the sign of this effect ("women earn less than similar men"), its magnitude ("women earn a% less than similar men"), or a confidence interval for this magnitude ("the gender gap is between b% and c% with probability 95%"). Parametric models are generally sufficiently simple for their parameters to have an interpretable meaning, one that is shaped by the prior modeling choices. Being parsimonious in the way they link the predictors to the outcome, they provide simple answers to specific research questions – either a parameter for a given variable, or a binary answer to the question of significance. All these interpretations have direct implications in terms of understanding of the phenomenon under study, and can be immediately used outside the regression model they were based on. Furthermore, the very questions they answer are worked into the modeling choices, so that one typically tailors a parametric model to extract such knowledge from the data at hand.

11 A model that overfits is akin to a student who, in order to prepare for a math exam, would learn the all of the numerical answers given in class, without understanding the reasoning. This student would succeed in the unlikely case the test perfectly replicates the examples seen in class, but fail should the teacher change any number in the exercise. Learning the answers is not the same as being able to generalize them to other contexts.

The situation is quite different in supervised learning models, which usually do not allow direct interpretation of their parameters. This is a notorious downside to their greater flexibility. Rather than parameters, the main result of most learning algorithms is, once again, their prediction, a type of information social scientists are not used to working with[12]. Once the model has been trained, it can output a predicted value of the outcome variable for any given observation. This type of information can certainly be relevant in engineering problems (such as assigning a category to an image), or for commercial uses. Likewise, it can reap – and indeed already has reaped – significant results in applied domains such as medicine, where some algorithms routinely detect certain pathologies better and faster than professionals. But what can more fundamental disciplines do with prediction, when their main questions have less to do with diagnosis and more with explanation? Since most academic disciplines are more concerned with answering "why" rather than "whether" someone will "do" something (be it success at an exam, consumption of cultural goods, or the spread of a disease), this feature of machine learning has long been described as a major impediment to its acclimatization in science.

While they are certainly problematic, these traits are also well known by proponents of machine learning. Some even argue that prediction may not be such a bad tool overall, and that correct interpretation and good prediction actually go hand in hand (Breiman, 2001). Their reasoning is the following: a model that achieves better prediction is bound to extract more information from the data, and to capture the phenomenon under study more precisely. Surely, the argument goes, there is more to be learned from an income model that accounts for 90% of income variance than from one that accounts for 60%. Furthermore, several tools have been developed over the years to "open the black box" and extract meaningful interpretations from the otherwise obscure models. Most are model-specific, but a few are generalizable to all predictive models (including parametric models). Some offer insights into the global interpretability of the model and demonstrate which variables play the most important role; others explore the role of one variable across the dataset. These measures are detailed and illustrated in the following section, which compares parametric regression and supervised machine learning on an empirical case.

3. Differences in Practice: Determinants of salary in Sweden

How do supervised learning methods and parametric regression compare in practice? The following pages present a comparative analysis of the two approaches on a classic question of social sciences, the determinants of individual wages, on a single dataset. To do so, we consider the usual moments of quantitative analysis: model building, assessment of model quality, and interpretation of the results.

The analysis is carried out on the Swedish labor force for 2012. The rationale for choosing this country is that Sweden practices population registration (*folkbokföring*). As such, it keeps detailed information about all of the persons living on its soil, and has done so since the seventeenth century at least[13]. The continued collection of varied information over decades, along with the merging of various administrative registers (census, taxes, land register, educational attainments), makes the Swedish register a well-known trove of reliable, fine-grained mass data, which can be accessed through *Statistika centralbyrån* (SCB), the country's statistical service. For each individual living in Sweden, one can thus access data

12 It should be stressed that the term "prediction", in this context, does not mean predicting the future. A supervised model acts like a function: when given a set of input values, it produces an output value, which is called a prediction because it represents what the model "thinks" the outcome variable should be. It can thus apply to any observation.

13 A prerogative of the Swedish church until the official separation of church and state, population registration is now conducted by the tax office.

on a wide range of items including civil status, composition of the household, identity of spouse and children if applicable, as well as place and company of employment, eligibility to taxes, days of sick leave taken in the previous 12 months, etc. All these are updated and stored yearly. Other information, such as results from tests taken during conscription or real-estate transactions, can also be retrieved. The register thus offers both large-scale and quality data, a combination that is quite rare nowadays in the social sciences.

With these sources, we analyzed income in Sweden in 2012. A classic question in the quantitative social scientific literature since Jacob Mincer's famous equation (1974), its determinants are overall well-known. For instance, certain regularities like the gender wage gap and the role of age have consistently been established in the literature, both in Sweden and abroad. This approach may lack the excitement of novelty, but it also offers precious landmarks when investigating a new method. For the sake of the analysis, not all items were used – only nine predictor variables were retained (see insert). The rationale for selecting only a few variables is that the use of many predictor variables (whose role is at best unclear) is likely to unduly favor the machine learning algorithms. The output variable for all models is the logarithm of the yearly gross wage[14], as this transformation helps linear models cope with the usually non-Gaussian distribution of wages. We furthermore restricted the study to the employed population aged 18 to 65, with a gross salary exceeding 50,000 crowns a year (about 5,000 euros, less than a third of the full-time observed minimum wage), to remove part-time workers. Again, this was done in order not to unduly unfavor the linear model, which is notoriously sensitive to outliers and heterogeneous samples. This left us with 4,109,447 observations.

We fit three models to these data. The first one is the simplest form of linear regression, which takes as predictors the nine aforementioned variables, without any further transformation or interaction effect. We call this model "naive" because it does not require any domain-specific knowledge. The second model builds on this model by adding interaction effects and variable transformations. It does so using a "lasso," a popular variant of the linear model in the machine learning context. Finally, the third model uses a fully flexible machine learning algorithm: random forests.

All three models are presented in sequence. In order to be able to apply machine learning validation standards, the studied population was randomly divided into three samples prior to any treatment: the training sample (70% of the studied population, 2,876,613 observations) was used to fit the models[15], the validation sample (15%) was used to fine-tune the models' meta-parameters, and the remaining 15% was used only once at the very end, as a test sample to estimate the generalization error of each final model.

DESCRIPTION OF VARIABLES

The predictor variables retained in all models are the following:
Age: A numerical variable, indicating the person's age in 2012.
Gender: A categorical variable with two levels (man / woman).
Occupation: A categorical variable with seven ordered levels (Managers, Skilled employees,

14 The logarithm is the most common transformation for the econometric study of wage equations. It helps linear models cope with the usually non-Gaussian distribution of wages, so that the partial effect of each variable is relative rather than absolute (e.g. an extra year of schooling might be associated with, say, a 5% increase in wages, rather than an increase of 3,000 Swedish crowns).

15 While it is unusual to split the data into training and validation samples in the context of parametric regression, it is worth noting that in this specific case where the data describe the complete population, fitting a linear regression on a random sample makes it conform better to the probabilistic hypotheses of inference (p-values, for instance, lose their theoretical justification when computed on a complete population).

Intermediary occupations, Office and service activities (low qualifications), Low-skilled workers, Military, and Other occupations).
Children between 0 to 10: A numerical variable indicating the number of children under the age of 10 in 2012.
Region: A categorical variable with 21 unique levels indicating the region of residence.
Family Type: A categorical variable with 4 levels indicating the type of relationship (single, couple) along with the presence of children in the household.
Educational Level: A categorical variable with 5 levels indicating the highest educational achievement (including an "Unknown education" level).
Unemployment: A continuous variable indicating the number of days of paid unemployment in the previous year.
Citizenship: A categorical variable with 4 levels indicating the individual's citizenship (Swedish, Nordic countries, European Union, Others).

Three models, one dataset and one question

We start with a simple *linear regression model*, in which each predictor variable has a linear association with the output variable (logarithmic gross wage). This "naïve" model is fitted through ordinary least squares on the training sample. The R^2 is 0.36, meaning that the model bridges 36% of the error gap between a null model, which would predict the average wage for all observations, and a perfect prediction. This corresponds to a (root mean squared) prediction error of 0.44 which means that, on average, the predicted gross wage lies somewhere between 64% and 155% of the true observed gross wage[16].

The estimated coefficients and associated statistics are reported in table 1. As could be expected given the large number of observations, nearly all coefficients have p-values very close to zero, making this standard interpretation tool virtually useless. This is a well-known limitation of significance testing: large numbers of observations push standard deviations (and thus p-values) toward 0. Conceived at a time when data was scarce, statistical tests do not fare well in the era of big data[17]. On the other hand, the estimated parameters are readily interpretable: the coefficient for Gender-Woman (-0.26) means that, according to this model, women earn around 23% less[18] than comparable men, all other things being equal. Similarly, each additional year of age is found to be associated with a 1% increase in gross wage, *ceteris paribus*.

The simplicity of these association effects reflects the simplicity of the chosen model: a 1% increase in wage for each additional year of age is an interesting and remarkably parsimonious result, but it is very probably an over-simplification, as wages are known to rise faster in the early stages of a professional career, and the temporal dynamics of wages are most probably different for men and women. It is thus tempting to try and specify a better model, by adding, removing and combining terms that might take into account such non-linear and interaction effects. This is what we do for the second model.

16 Because our models predict logarithmic wages, the models' errors have to be turned into percent of change to be interpretable in terms of nominal wages. A root mean squared error of A thus becomes a (100 x exp(+/- A)) % error in nominal wages.

17 (Saporta 2006, p. xxxii) gives a telling example of this fact. When measuring the linear correlation between two continuous variables with one million observation, the absolute correlation need only be 0.002 to be significant at the 5% level. Such a small correlation is, of course, of no interest for interpretation.

18 Because gross wage enters our regression function in logarithmic form, an estimated coefficient b for predictor X means that a unit increase in X is, on average, and all other variables held equal, associated with a 100 x (exp(b) - 1) % variation of gross wage.

	Estimate	Std. Error	Marginal effect on gross wage (%)	p-value
(Intercept)	7.534	0.003429		$< 10^{-15}$
Age	0.0111	0.000023	1.1	$< 10^{-15}$
Gender: Man	*ref*			
Gender: Woman	-0.2572	0.000571	-22.7	$< 10^{-15}$
Family: Couple with no children	*ref*			
Family: Couple with child(ren)	0.0649	0.000897	6.7	$< 10^{-15}$
Family: Single parent	0.0445	0.001143	4.6	$< 10^{-15}$
Family: Single & others	0.0631	0.000901	6.5	$< 10^{-15}$
Children 0-10	-0.0010	0.000401	-0.1	0.011523
Education: Below high school	*ref*			
Education: High school degree	0.0982	0.000922	10.3	$< 10^{-15}$
Education: Higher ed - 2 years or less	0.0723	0.001337	7.5	$< 10^{-15}$
Education: Higher ed - over 2 years	0.1185	0.001074	12.6	$< 10^{-15}$
Education: Unknown	0.0170	0.003675	1.7	0.000003
Occupation: Intermediary Occupations	*ref*			
Occupation: Managers	0.3388	0.001203	40.3	$< 10^{-15}$
Occupation: Military	0.1042	0.004306	11.0	$< 10^{-15}$
Occupation: Office & Service (LowQ)	-0.2648	0.000837	-23.3	$< 10^{-15}$
Occupation: Others	-0.5658	0.001669	-43.2	$< 10^{-15}$
Occupation: Skilled Employees	0.1172	0.000864	12.4	$< 10^{-15}$
Occupation: Workers	-0.2086	0.000883	-18.8	$< 10^{-15}$
Days of unemployment	-0.0031	0.000009	-0.3	$< 10^{-15}$
Citizenship: EU, Except Nordic	*ref*			
Citizenship: Nordic countries	0.0435	0.002997	4.4	$< 10^{-15}$
Citizenship: Other countries	-0.1169	0.002734	-11.0	$< 10^{-15}$
Citizenship: Sweden	0.0388	0.001936	4.0	$< 10^{-15}$
Multiple R-squared: 0.3586, Adjusted R-squared: 0.3586 F-statistic: 4.022e+04 on 40 and 2876572 DF, p-value: < 2.2e-16				

Table 1: Estimation results for the "naive" linear regression on Swedish register data. *Dependent variable: logarithmic gross wage. Reference case: man, in a couple with no children, below high school education level, Intermediary occupation, EU (except Nordic) citizenship.*

For better readability, 21 coefficients for County of residence were omitted from the table (of which 15 showed p-values that were virtually 0).

***

The second model, which we shall call *sophisticated linear*, builds on the naive one by adding a squared age term and an interaction effect between age and gender. This allows a non-monotonic association between age and wages, and specific age effects for men and women. Because we arrived at this specific model formulation by a process of trial and error, trying multiple specifications and choosing the best we could find, we used the lasso procedure instead of standard linear regression[19]. This lasso is still a linear model, but its optimization follows a machine learning approach (see insert), circumventing the problems that trial-and-error poses to standard regression and hypothesis testing. We thus chose the model's specification based on its prediction quality on the validation set. Once this was done[20], the final generalization error was evaluated on the test set, yielding a R^2 of 0.40.

The lasso: parametric models without p-values

The lasso is a popular procedure that was invented in the 1990s, and may be seen as the machine learning way of performing linear/parametric regression. In a sense, it is still a parametric model, because the form of the relationship between predictors and outcome is specified by hand in the exact same way it would be for a linear model. However, it uses a different optimization algorithm than standard linear models, because it has a different cost function to minimize (see Hastie et al, 2008, p. 68).

Instead of just finding the parameters that minimize the sum of squared errors, the lasso also seeks to limit the complexity of the fitted model, by forcing the parameters toward small absolute values. This is done to overcome some well-known limitations of linear regression (possible overlearning, overinterpretation of p-values…) while at the same time retaining its ease of interpretation. The intuition is that, for a given model specification, higher parameters mean more complex models: smaller parameters imply that a change in some predictor is associated with a small change in the outcome, making the overall model more careful.

The complexity of the model is measured by the sum of its absolute parameters, and is controlled by an additional parameter λ. For zero λ the lasso has the same solution as an unconstrained linear model, while for a very large λ all coefficients are forced toward zero. Between these two extremes lies an optimal value, the λ that yields the best generalization error. This optimal value is found, as typically for a machine learning procedure, by cross-validation on the training sample.

Just like a standard parametric model, a lasso yields estimated parameter values that can be immediately interpreted. However, these parameters are not associated with standard errors and p-values; instead, the constrained cost function ensures that "useless" variables have a parameter exactly equal to 0, so that they are left out of the model.

The lasso is thus a useful cross-over between standard parametric modeling and machine learning. It has been extensively used in many empirical settings over the last two decades, and been given many extensions. Additionally, and quite uniquely for a machine learning model, it

19 The attentive reader might note that, moving from our first to our second model, we make two changes at once: the model's parametric specification and the optimization algorithm, thus making comparison more difficult. We would argue, however, that if anything the lasso is supposed to perform better than ordinary least squares in terms of generalization error, so that our second model is actually giving parametric specifications their best chance.

20 All lasso computations were executed using the *glmnet* package for R (Friedman, Hastie, Tibshirani, 2010), using the default parameters and 10-fold cross-validation.

can be fully justified in terms of Bayesian statistics.

The estimated coefficients are reported in table 2, and while their value may be interpreted in roughly the same way as for the simple linear model, interpretation is made harder by the fact that all predictors were standardized prior to fit, and by the higher complexity of the model (the effect of age, for instance, can no longer be summarized by a single number, as it enters the predictive model through a linear term, a quadratic term, and an interaction term with one of the genders). Note the absence of standard errors and p-values next to the coefficients, as such statistics cannot be computed using the lasso procedure. Instead, the lasso has forced some coefficients to zero (not reproduced in the table): according to the lasso procedure, and given the specific model we chose, these variables are better left out in order to achieve good generalization. The non-zero coefficients, on the other hand, are deemed useful for prediction; this is notably the case for the squared age term, the gender-age interaction, and the number of children. Finally, because all variables were standardized, the absolute value of their parameters can serve as a measure of the relative contribution of each variable to the fit. The model can thus be said to be dominated by the age, gender and occupation variables.

Variable	**Coefficient**		**Variable**	**Coefficient**
(Intercept)	7.9555		Occupation: Intermediary Occupations	*ref*
Age	0.8198		Occupation: Managers	0.0807
Age (squared)	-0.7107		Occupation: Military	0.0072
			Occupation: Office & Service (LowQ)	-0.1094
Gender: Man	*ref*		Occupation: Others	-0.0792
Gender: Woman	-0.1655		Occupation: Skilled Employees	0.0469
Age x Gender:Man	0.0415		Occupation: Workers	-0.0851
Family: Couple with no children	*ref*		Days of unemployment	-0.0884
Family: Couple with child(ren)	-0.0081			
Family: Single parent	-0.0089		Education: Below high school	*ref*
Family: Single & others	0.0030		Education: High school degree	0.0376
			Education: Higher ed - 2 years or less	0.0111
Children 0-10	-0.0245		Education: Higher ed - over 2 years	0.0409
			Education: Unknown	-0.0006
Citizenship: EU, Except Nordic	*ref*			
Citizenship: Nordic countries	0.0073			
Citizenship: Other countries	-0.0157			
Citizenship: Sweden	0.0159			

Table 2: Estimation results for the “sophisticated” linear regression, estimated by lasso on Swedish register data.

Dependent variable: logarithmic gross wage. All variables (dummy transformations and interactions included) were scaled to unit variance prior to estimation. Reference case: man, in a couple with no children, below high school education, intermediary occupation, and EU (except Nordic) citizenship[21].

We thus achieved a slightly better generalization result by using a more sophisticated linear model than the first one. This is, of course, encouraging, but still begs the question: is this really the best possible model? Despite our best efforts to build a good and meaningful model, and our (limited) knowledge of the subject, it might very well be that we missed some important effects or interactions. We therefore turn to a fully flexible model to try and extract the most information from the data.

***

The random forest algorithm was introduced in 2001, and has been acclaimed ever since for its adaptiveness and ease of use (see Hastie *et al.*, 2008, p. 587-603). Without going into the technicalities of the algorithm, it can be said that it uses multiple instances of an earlier supervised learning procedure called *classification and regression trees* (CART[22]) in what is called an *ensemble*. The idea of an ensemble is, in a sense, democratic: combine the opinions of many voters, and you get a single powerful decision. The random forest applies this principle by averaging the predictions of many low-bias, high-variance and de-correlated trees[23]. A typical forest is a collection of hundreds of trees, each of which is slightly perturbed in a process involving random numbers, hence the name. Its individual trees are relatively poor predictors, but as an ensemble it is a powerful low-bias, low-variance model. The resulting complexity of the random forest also makes it a typical black box.

Simply giving the nine predictor variables and the outcome variable to a random forest implementation[24] yields an R^2 of 0.45 on the test sample. This is a notably better prediction score than our best effort with linear models (see figure 1 for diagnostics plots of the three trained models). This was, furthermore, achieved effortlessly since the model construction effort was carried out by the computer, without the researcher doing a thing. The algorithm has thus captured a more accurate picture of how wages vary along with social characteristics in Sweden. However, immediate interpretation was lost in the process. Indeed, there is no reading of coefficients in random forests, as the combination of variables it produces is an intricate entanglement of tree branches. In fact, should the full mathematical equation of the trained model be written down, it would fill many pages (this is thus never done). Interpretation, instead, has to rely on prediction.

21 For improved readability, 21 coefficients for county of residence were omitted from the table (all of which except Stockholm showed very small absolute coefficients, three of which were 0).

22 Classification and regression trees are simple algorithms for supervised learning (see Hastie *et al.*, 2008, p. 305-312). They work with recursive binary partitioning: first find the explanatory variable that best separates the output variable into well-separated groups, split the population in two, then repeat the operation on each of these resulting sub-populations, then on the four subsequent sub-populations, *etc*. The result has a tree form that makes it easy to interpret.

23 The individual trees of a random forest are said to have low bias in the sense that, on average, they give a prediction that is close to the desired target value; they have high variance because a small change in the data can result in a large change in the fitted tree; they are de-correlated in the sense that they are forced, by various random perturbations, to fit the data in many different ways.

24 For all random forest computations we used the very efficient *ranger* package (version 0.8.0) for R (Wright and Ziegler, 2017), with 500 trees and 4 variables per split.

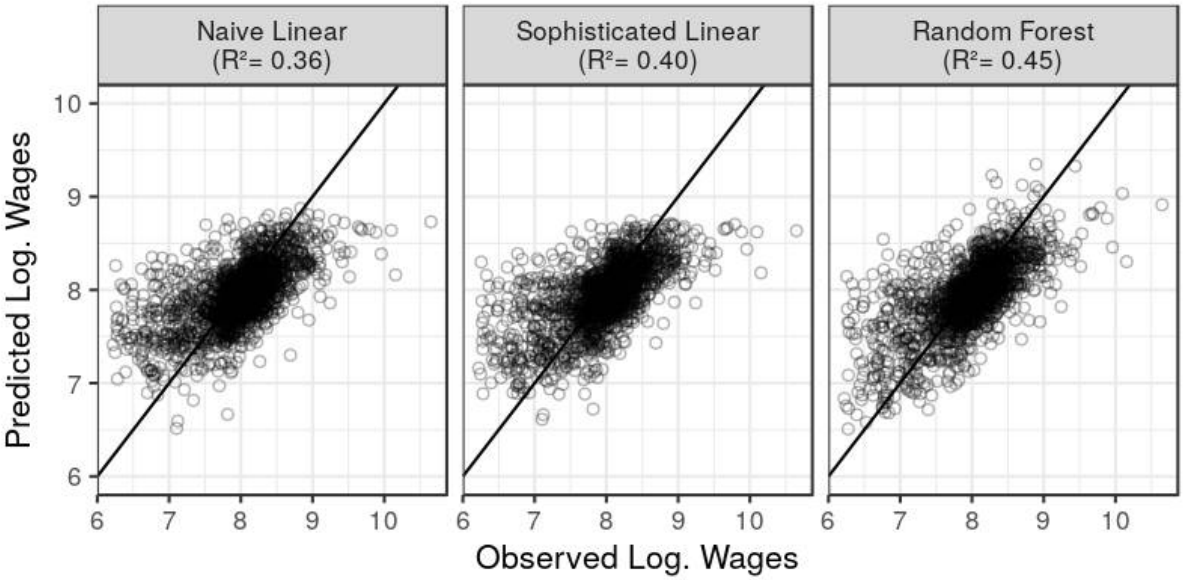

Figure 1: Diagnostics plots, observed vs. predicted wages.
Sample of 1,000 test observations, from the Swedish wages data. The horizontal axis represents observed logarithmic wages, the vertical their predicted counterparts, for three different models. The diagonal represents perfect generalization, where each out-of-sample observation is correctly predicted by the model.

Model interpretation

The first key of interpretation of supervised models, especially in the machine learning context, is their predictive power. In terms of generalization error, the random forest is clearly the best of the three: its R^2 of 0.45 means that it bridges 45% of the gap between the null model and perfect prediction (compared to 36% and 40% for the naive and sophisticated linear models). Its root mean squared error on the test set is 0.4063. In terms of nominal gross wages, this means that most random forest predictions lie between 67% and 150% of the true observed wages (65%-153% for the sophisticated linear, 64%-155% for the naive linear model). This is certainly a step forward, although admittedly still far from perfect[25]. In exactly what way does the random forest surpass the linear models? The diagnostics from figure 1 give us a first indication, showing that the flexible model does a better job at capturing variations at the ends of the wage spectrum: its predictions are better for the nearly 40% of the population whose log-wages are under 7.5 or above 9, and its predicted wages are less narrowly distributed than is the case for both linear models. All in all, it is thus a better prediction machine, in the sense that it gives a more accurate answer to the question "how much does someone with this set of social properties earn?"

Prediction alone, of course, is not very useful to a social scientist, and the black box aspect of the random forest model, with individual parameters that are meaningless for direct interpretation, can be unsettling. Fortunately, a variety of tools have been designed to extract interpretable information from predictive models. The next pages present two of the most

25 A word on perfect prediction might be in order here. While it is, in theory, the ultimate goal of any predictive model, it can of course never be attained in practice, at least in social science. For one thing, individual and social behavior are always partly random, so that there always remains a portion of incompressible "noise" in the data. Furthermore, even complex models trained on millions of cases cannot compensate for low-quality predictor variables: measurement errors, dubious proxies and omitted variables are very common in social scientific research, and are most probably the real bottleneck that prevents good prediction.

common, which share the benefit of being applicable regardless of the type of model used: permutation importance and partial dependence.

Permutation importance: which variables matter for prediction?

As its name suggests, the *permutation importance* measure aims at assessing the relative (predictive) importance of the variables at hand, using permutation. It exploits the fact that once the model has been trained, it works as a prediction machine that can be fed any predictor data. Its principle is simple: using a subset of the population under study (preferably the test set), the values of a given variable are randomly permuted (rearranged in a random order), leaving all other variables intact. When these perturbed data are passed through the model, the prediction quality is degraded (when compared to prediction on the unperturbed data). This degradation in prediction quality is interpreted as an importance measure for the current variable: a variable that makes the model much worse when it is permuted is important, while a variable that can be permuted without affecting prediction quality is not important. The procedure is repeated for each variable in sequence, and may be repeated several times for each variable to improve robustness. The whole process thus offers a standardized, interpretable measure of variable importance that can be compared across predictors (both continuous or categorical), does not depend on the type of learning model used, and demands only a single trained model[26].

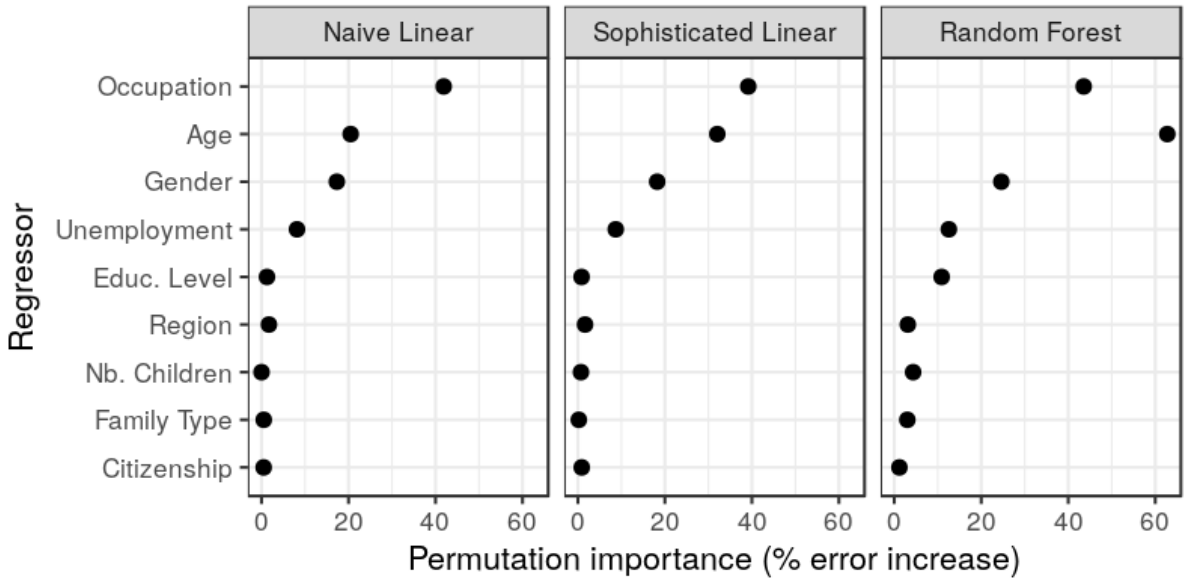

Figure 2: Variable importance estimation for the three models (effect of random permutation of each variable on prediction accuracy).
Occupation is an important predictor for all models (prediction error increases by around 40% when this predictor is shuffled); Age is interpreted as not very important in the Naive

26 Prediction importance could also be measured by training multiple models, and measuring the degradation in prediction quality that results from removing each variable; this, however, is not practical, as it demands training a large number of different models, a number that grows exponentially with the number of regressors.

Linear model, important in the Sophisticated Linear model, and very important in the Random Forest.

Figure 2 shows the results of permutation importance for the three trained models. For both linear models, the importance measure yields results that are in stark contrast with the common interpretation of its coefficients and p-values. The first difference is formal: while significance tests give a binary (yes-no) answer, permutation importance yields a score and an ordering. More importantly, whereas nearly all regression parameters were deemed significantly different from 0 (at the 0.1% level) and showed sizeable marginal effects, from a permutation point of view only Occupation, Age, Gender (and to some extent the number of days of unemployment in the past year) play an important part in prediction[27]. In comparing both sets of results, however, one should keep in mind that the two importance measures do not answer quite the same question. Consider, for instance, a dataset in which a predictor x_1 has a definite but very small effect on outcome y, independent from the other regressors. While a (well-defined) parametric model should find the effect to be significantly different from zero, x_1 would score low on permutation importance. Depending on what one means by "importance", one of several possible measures can thus be favored.

For the sophisticated linear (lasso) model, the permutation importance measure can be confronted to the estimated coefficients, again with contrasting results: while many variables have non-zero estimated parameters (such as the education and county of residence dummies), again only three or four variables can be deemed important for prediction, based on the permutation measure.

Comparing these models on permutation results alone, we find that all three agree on the main lines: Occupation, Age and Gender are always the most important variables, followed by numbers of days of unemployment. Additionally, the random forest draws attention to Education by giving it a moderate importance level. The order of the variables, however, varies from one model to another: the importance of Age increases as the complexity of the model increases (around 20% increased error for the naive linear model, 35% for the sophisticated linear, over 60% for the random forest). We may infer from this that wages and age are associated in a complex way, or at least one that is hard to accurately formulate into an explicit equation.

Such diverging results raise the question of which of the three importance measures is the most accurate, the answer to which is debatable. From a machine learning perspective, the fact that the random forest yields better (cross-validated) prediction makes it more trustworthy. However, the lack of inference on the permutation importance values (in the sense of a measure of uncertainty of the computed values) makes it somewhat dubious from a statistical perspective. This flaw applies more generally to most interpretation results of supervised machine learning, although in recent years some effort has been made in that direction: see (Efron and Hastie, 2016, chap. 20).

The *partial dependence* measure is another method that harnesses prediction power to gain insight into a model's representation of the data. In the naive linear model, the marginal effect of a given variable can be directly computed from its estimated parameter, and interpreted in the common *ceteris paribus* way. The sophisticated linear model has slightly more complex marginal effects due to the interaction terms, and these effects are allowed to

27 Of course, p-values by themselves are not sufficient to assess whether a variable does really play an important role in the model, as reminded by the American Statistical Association's statement. Further interpretation should rely on the size of the effects, as measured by the estimated parameters (and always keeping in mind that these strongly depend on the chosen model form). For instance, the marginal difference between men and women on gross wage is equivalent to a 20-year difference in terms of age, so that the effect of gender can be said to be more important.

vary between different sub-populations, as pre-specified in the model's construction. When using flexible models such as the random forest, the effects of a given predictor are allowed to vary freely according to the predictor's value, in a non-monotonic and non pre-specified way. Contextual interpretations can then be made about the local effects of variables (age, for instance, is not specified *ex ante* to play a monotonic or quadratic role, or to be independent from the other regressors). *Partial dependence* is a popular method to extract such information from black box models. As it cannot be summarized by a single digit, it is typically represented on bivariate "partial dependence plots".

Partial dependence: isolating the effect of a single variable in a complex model

The intuition for partial dependence is that, if one is to study the association of a single variable with the outcome, the effects of all other predictors can be neutralized by *averaging them out*. Just as permutation importance, partial dependence combines predictive power and perturbations of the original dataset to produce interpretable measures. The partial dependence of the outcome variable on a given predictor X_1 is computed in the following way: for each value x that X_1 takes in the dataset (and preferably in its test-subsample), a new dataset is created in which all values of X_1 are replaced by this single value x. If X_1 is gender, for instance, two synthetic datasets are created, each one with as many observations (N) as the base dataset: in the first table all observations are assumed to be women (while all other variables are left untouched), in the other all are set to be men. Both these synthetic datasets are fed, in turn, to the predictive model, yielding N predicted wages for synthetic women, and another N for synthetic men. Finally, we compute the average prediction for the synthetic men and for the synthetic women, and these two average predictions make up the partial dependence plot[28]. In the case of age, 47 different synthetic datasets of size N are created: one where all individuals are set to be 18 years old, one for 19, and so on until 65 years old.

28 As in the case of permutation importance, it is interesting to note that, while the partial dependence measure is used to answer a question similar to that of the *ceteris paribus* effects of variables in the linear model, it does not measure the exact same thing. While the estimated parameters of the logistic regression measure the marginal effect of each variable, *independently* from the other predictors, the partial dependence measures the *average* effect of a given predictor, for all observed values of the other predictors. In the former case, the other predictors are "controlled" by the specific form of the model, while in the latter they are "averaged out" (see Hastie, Tibshirani and Friedman, 2008, pp. 369-370 for a treatment of this difference).

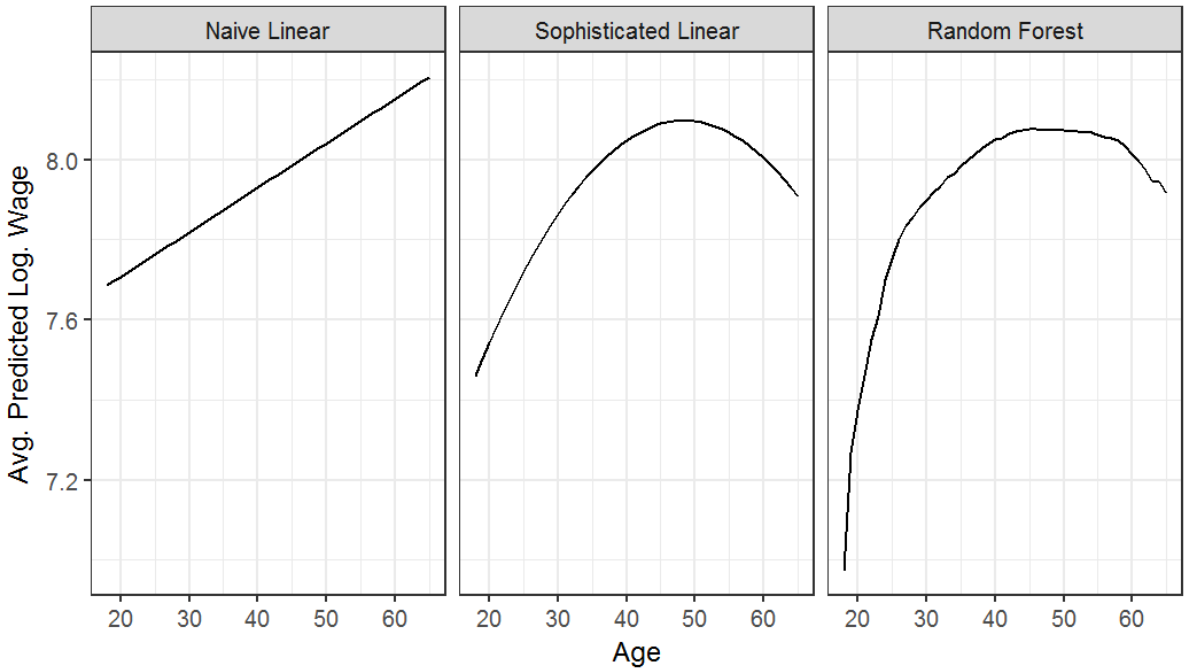

Figure 3: Partial dependence plot for age, computed from three predictive models
The horizontal axis is age. The vertical axis is the predicted log wage, obtained by averaging out all predictors other than age. This is interpreted as the specific contribution of age to the variations of wages, as understood by each model.

Figure 3 shows the partial dependence plots for the age regressor, for all three fitted models. The shape of the curves is a telling visual indicator of the way the regressor is used in the model. The naive model's curve is by design straight, while the sophisticated linear model's curve describes a section of a parabola: predicted wages rise steadily from 18 years on, peaking just before 50, and slowly decreasing afterwards. In these models, the shape of partial dependence is determined by the initial modeling choices: the naive linear specification can only yield a linear curve, the sophisticated linear is allowed to take a quadratic shape. The random forest, on the other hand, is free to let each regressor take an arbitrary form in the predictive model. As noted before, it predicts wages on a broader spectrum than the two other models; on figure 3 this is visible from the fact that it predicts much lower average wages for the youngest in the sample (18-25 years old). Its partial dependence curve is reminiscent of the quadratic shape of the sophisticated linear, but its irregular form is actually closer to piecewise linear: fast rising wages for the young, then a slower increase from 25 to 40, followed by a plateau between 40 and 60 years old, and a steady decrease afterwards. Such an irregular form would be near-impossible to attain with parametric models[29], the plateau shape being especially difficult to build using polynomial transformations. It is also probably a more accurate description of the association between age and wage in Sweden, judging by the lower generalization error of the random forest.

It must be noted that the same partial dependence method can be extended to the estimation of combined effects of two or more predictors, although the required amount of computation grows exponentially with the number of variables. The search for interesting combined effects of predictors, the modeling of which is theoretically one of the strong points

29 This is at least true for models were age is kept as a continuous variable. Piecewise linear functions are easy to construct if the turning points are already known (in this case, ages 25, 40 and 60), but the whole point of the random forest (and similar flexible models) is that it was able to detect these thresholds by itself.

of machine learning, thus largely falls to the researcher, who must try out different small subsets of predictor variables and visualize them on 2D or 3D plots. Another weakness of the partial dependence measure arises from the way it tries to force counterfactual prediction: when estimating the partial dependence on occupation, for instance, the model will be asked to predict synthetic observations that have no close equivalent in the training dataset (for instance women with a high education level doing manual construction labor), which can yield misleading results. Finally, as in the case of permutation importance, faith in the interpretation of partial dependence rests entirely on the credit given to the predictive power of the model, rather than on statistical theory.

4. The advent of statistical learning?

In the light of the previous developments, it should be clear that machine learning offers an alternative approach to quantification. Will it replace more classic methods altogether, starting with parametric regression? The time may seem ripe for the computational approach to take over quantitative social science. Vast amounts of funding are being poured into the field. Just like big data in the last decade, “machine learning” has become the fashionable keyword to include in an abstract when applying for a grant or a prestigious conference. Similarly, the hype surrounding the field, whose results regularly make the headlines of international media, attracts masses of scholars and students. More importantly perhaps, these successes come at a time of heightened defiance towards the dominant method of parametric regression and its longtime companion, statistical significance testing. P-values, in particular, are suspected to lie at the heart of the crisis of replicability that has been documented in many areas of science, and their use and abuse has lately been condemned by the largest professional association in mathematical statistics (Wasserstein and Lazar, 2016). However, the current results offered by statistical learning, along with lessons from the history of science, should warn against unfettered optimism: the oft-evoked machine learning revolution, advocated by some and feared by others, may have to wait.

Three challenges for machine learning in science

One reason has to do with the type of results produced by these techniques. As evoked before, prediction is not a tool many scientists are used to working with. Surely, various tools have been developed to extract information from black-box models and to offer interpretation possibilities akin to those of parametric regression. However, the large variety of both machine learning models and interpretation tools, along with the lack of solid statistical inference, means that no golden standard for interpretation has appeared as of yet. This, combined with the fact that learning methods are not commonly taught in social science departments, makes publishing machine learning results much costlier than sticking to standard parametric modeling. Furthermore, the lack of guarantees as to the optimality of any given trained machine learning model, and the fact that their high flexibility could make them easier to manipulate toward a preconceived result, might make their results even more suspicious than those of parametric models.

In addition to the types of information offered by learning algorithms, the results themselves are often not as spectacular as could be hoped for. There are, of course, undeniable successes in certain areas. Automatic translation, self-driving cars, or, closer to science, the detection of rare diseases all bear witness to the efficiency of the approach - and the list is regularly growing longer. But even on its own criterion of choice, (cross-validated) predictive power, machine learning is regularly outperformed by more classic methods on social science datasets.

Here lies another challenge for machine learning: the size of the datasets. While parametric regression does not usually improve significantly beyond a few hundreds of observations, machine learning algorithms can require a very large number of training cases (observations) to live up to their promise of universal approximation. There is no rule as to how many observations are enough, as this is entirely dependent on the intrinsic complexity of the relationships to be modeled. And of course, this complexity remains largely unknown in advance[30].

Figure 4 illustrates how the three different models presented above were affected by size: all three were trained multiple times on random samples on the Swedish wages dataset, with sample sizes varying from 300 to 300,000 observations[31]. The linear and random forest models are the same as in section 3. XGBoost is another flexible learning model; it refers to "Extreme Gradient Boosting", a recent variation on the popular gradient boosting algorithm (Hastie et al, *op. cit.*, chapter 10). The linear model's prediction quality increases at first, when data is scarce, but reaches a plateau after 3,000 observations: its rigid specification prevents it from extracting more information from additional data. The random forest's prediction power, on the other hand, shows a steady increase as the dataset expands, with a curve hinting at the fact that it might do better still on a larger dataset (as indeed it does, to wit its results on the full near-three-million training set in section 3). The XGBoost curve is even more interesting: while this powerful model performs slightly better than the random forest after the 30,000 observations mark, it does consistently worse than even the linear model on datasets with under 10,000 observations. Unfortunately, 10,000 observations is already a large dataset for social sciences.

30 It should be noted that machine learning procedures are in fact capable of surpassing parametric models on small datasets, as witnessed by the many examples of only a hundred observations that can be found in machine learning textbooks. In the authors' experience, however, this is not the rule on social science datasets.

31 In order to assess this effect of size, we used 7 sample sizes, from 300 to 300,000 observations. Each model was trained on 20 random samples of each size, for increased robustness. As we used three different models, this adds up to a total of 7 x 20 x 3 = 420 trained models.

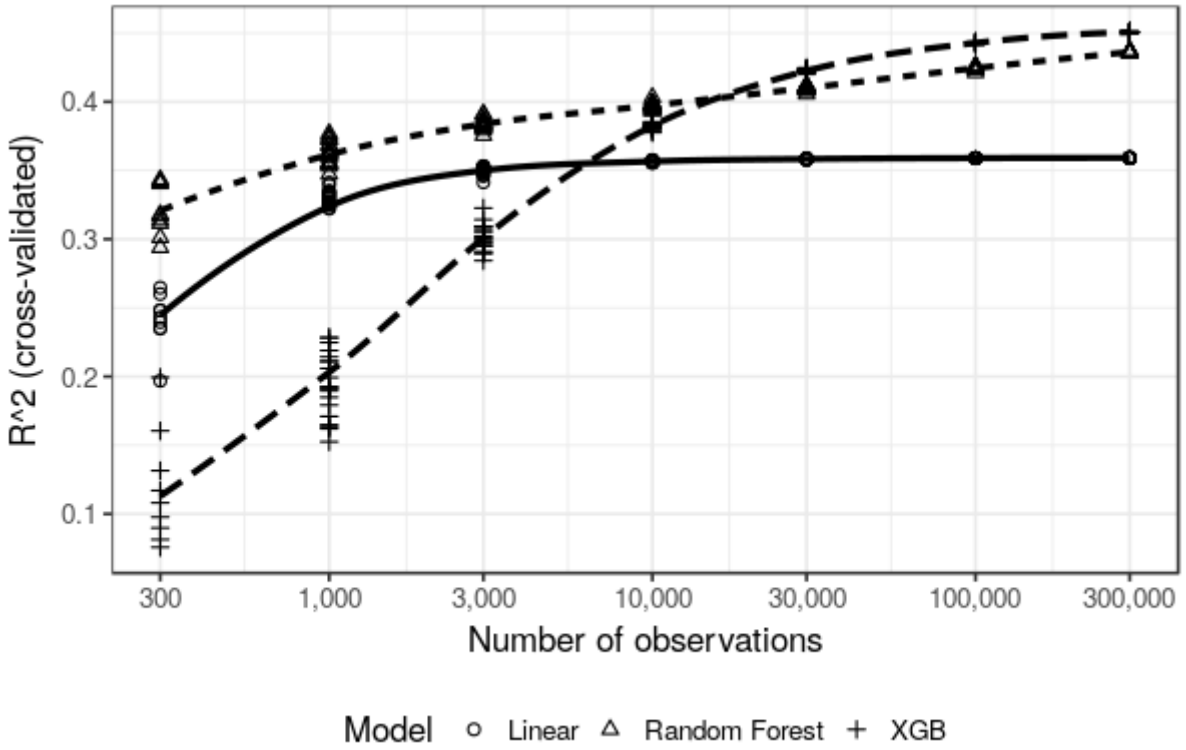

Figure 4: R^2 for three different predictive models, on random subsamples of varying sizes taken from the Swedish wages dataset. Here the R^2 is said to be cross-validated because it is computed on an independent test sample. A value of 0 means that the prediction is no better than the null model (predicting only the average wage for all observations), 1 means perfect prediction.

There is an important lesson to be gained from this: when faced with datasets that are too small (even with tens of thousands of observations in the case of complex patterns), machine learning procedures may not fulfil their promise of universal approximation. They may not even do better than more classic approaches. In fact, it is worth noting that while doing research for this article, the authors have tried out supervised learning algorithms on a good dozen datasets with varied properties and size. More often than not, standard parametric regression performed at least as well as advanced learning algorithms. This issue could be progressively solved by the ever-increasing mass of available data. The growing use of computers, the dissemination of sensors capturing varied aspects of everyday life, or the ease with which one can now format and merge databases, have led to a sharp increase in the size and number of datasets at hand. This might however not be sufficient. The majority of the newly and massively available data is, in fact, quite limited in information. As the former head of the U.S. Census bureau once put it, the current rise in data is not primarily due to an increase in what he called “design data” - data collected with a research question in mind. Instead, it consists in a new abundance of what he termed “organic data,” information collected for other purposes (the functioning of an administration, of a service, or for business matters). And while the latter may be converted into relevant information, this is in no way guaranteed (Grove, 2011). In the social sciences, big data all too often remain poor data.

There is yet another reason why machine learning may not replace standard methods altogether. The history of science reminds us that this is not the first time that (social) sciences have discussed the relative merits of a more inductivist approach over a theory-driven one. The cyclical debates about the primacy of empirics bear witness to this. This is of

course the case in sociology - one need only think of the periodic resurgences of such claims, for instance under the guise of “serendipity” or of “grounded theory”. This is also the case in economics, where empiricism is sometimes denounced as a form of “measurement without theory” (Koopmans, 1947), but sometimes vindicated too. Likewise, the stern defense of flexible models coming from proponents of machine learning echoes the oppositions between proponents of modeling and supporters of rich descriptions in many disciplines. The lessons of the past must be heeded here: the lack of resolution of these century-old disputes has less to do with the absence of an appropriate method, and more with the fact that by choosing to focus on one aspect, one necessarily loses sight of the other. There is no easy way out of established antinomies.

History also suggests that, in science at least, revolutions are less frequent than selective appropriation. In his book *Chaos of Discipline* (Abbott, 2001), Abbott defines this as a process of “fractalization,” whereby key insights from one approach are integrated within the other, while the two traditions keep on existing on their own. Such a process is surely ongoing between machine learning and parametric regression. In a recent example of what the data avalanche could do to econometrics, Hal Varian (2014) mentioned a few elements that could be fruitfully imported into this discipline from statistical learning, most notably tools for variable selection (the choice of which predictors to include in a regression model). Coming from econometrics too, (Charpentier *et al.*, *forthcoming*) make a case for a broader use of machine learning, and in particular for ascertaining the correct specification of the model. The article also evokes the use of prediction and the question of causal inference, two aspects that were largely tackled by Susan Athey, who in recent years has consistently looked for ways to improve the classic econometric approach in the light of the practices of statistical learning (Athey, 2017). Conversely, causal modeling, an endeavor that used to be the hallmark of econometrics, is gaining currency in the machine learning literature, following the influential work of Judea Pearl (2009).

Knowledge and uncertainty

Hybridization does not mean that no change is under way. In the words of Abbott, “a fractal distinction produces both change and stability” (2001, p. 21). To capture this change, one must switch focal lenses and consider not what social sciences can do with machine learning, but rather what the rise of this set of techniques can do to social sciences. When it comes to their methods, many disciplines have been through deep introspection. Initiated in the first decade of the twenty-first century with a series of proven frauds and blatant invention of data, a movement for more transparency in scientific research emerged. It was reinforced by several failed attempts at replicating various studies. One of the most famous may be the Reproducibility project, spearheaded by psychologist Brian Nosek. Looking at 100 recently published studies in the field, he and his colleagues were able to reproduce only a third of the results proclaimed by the papers (Open Science Collaboration, 2015). Although the authors themselves nuance this result by stating that several reasons may account for such a low figure, the overall picture is grim. One reason is well known, and has to do with the organization of scientific production; the reduction of research budgets and the increased conditioning of grants on publications have led to more publication bias: publishing only conclusive results - and being in constant dire need to publish some - tends to generate false conclusions.

This is nonetheless not the only, and maybe not the even the main cause for this crisis that looms over most contemporary applied quantitative research. In a remarkable discussion of the issues plaguing contemporary quantitative research, John Ioannidis explained that the reason why “Most Published Research Findings Are False” had to do with the ubiquity of statistical significance testing (2005). And in fact, since its introduction in social sciences

after the 1950s, hypothesis testing, usually by way of parametric regression, has become omnipresent. With variations between disciplines and countries, they have largely reshaped the way researchers deal with numbers in the social sciences. This omnipresence, though, came at a cost. Speaking about U.S. sociology, Abbott famously contended that the adoption of the "generalized linear model" gave way to a cognitive change: increasingly, researchers started to think about the world as if it had the properties of the model, thus giving way to a generalized linear vision of reality (1988). The quasi universal adoption of this one tool certainly forced many to format their research so as to be able to implement a regression - even when the method was not warranted. Having become the main token of scientificity, parametric regression narrowed the scope of possible quantitative research.

Here may lie the most lasting impact of machine learning for social science, once the considerable hype surrounding it settles down. The rapid dissemination of statistical learning in many disciplines is indeed favoring the emergence of an alternative approach to quantification. By offering results that are sometimes, but certainly not always, in line with standard regression; by advancing different validation criteria; and most importantly by promoting another way to work with quantitative data. As the fundamental critique of parametric models and statistical testing gains currency, and as a plausible alternative emerges, parametric regression is being rooted out of its quasi monopolistic position. The aforementioned limitations of the machine learning approach, however, make it somewhat incommensurable, in the terms famously proposed by Kuhn. Its results are generally of a different nature than those of regression, and this prevents it from fully filling the opening space[32].

Paradoxically, the greater diversity of statistical techniques and criteria is thus making scientists more ignorant about their object of study: on the one hand, the statistical certainties of old are now known to be ill-founded, on the other hand the newer methods do not even pretend to give any definite answers regarding scientific hypotheses. All in all, this forces scientists to adopt a humbler position towards their data. In this sense, greater uncertainty may not be such a problem. Controlled ignorance might indeed be a scientific blessing in this time of patented replicability crisis. In a more recent endeavor, Nosek assembled three dozen teams of researchers to assess the magnitude of implicit bias in quantitative research. Each starting with the same question and the same database, the teams reached very different conclusions (Silberzahn et al., 2017). The article drawn from this experience subsequently made a powerful case for collective scrutiny. According to the authors, crowdsourcing of quantitative methods is necessary to debunk potential biases. Though obviously commendable, this option is not always feasible, as it would require too many people's time and energy - especially if it is to be done at every stage of research. But the multiplication of quantitative standards can offer such a critical eye. The presence of an alternative approach to quantification could become a way to ascertain our results. By forcing us to multiply standpoints, it would trigger an unsettling but eventually productive confrontation. After all, it has long been known that monoculture yields decreasing returns after a while[33].

32 If anything, Bayesian statistics might be the best candidate for this office, as it provides a rigorous theoretical framework on which to build more traditional and interpretable models - albeit again at the cost of increased complexity and loss of familiarity.

33 In other places or disciplines, the statistical hegemon may be different. The problem of methodological monoculture extends well beyond parametric regression.

Conclusion

Clifford Geertz once wrote that the immense success experienced by certain ideas is as justified as it is problematic. Speaking about concepts such as that of "culture" in the early twentieth century, he wrote that these ideas "resolve so many fundamental problems that they seem also to promise that they will resolve all fundamental problems, clarify all obscure issues." However, the anthropologist quickly noted, this achievement often becomes an issue in itself, as it can quickly transform into solipsism. "The sudden vogue of a *grande idée*," he added, is always at risk of "crowding out almost everything else for a while" (Geertz, 1973, p. 310). The same goes for methods. In quantitative research, few techniques have been as central as parametric regression (in its various forms of linear regression, logistic, Poisson, nonlinear least squares, *etc.*). Born itself at the turn of the twentieth century, it has produced momentous changes in numerous scientific areas. It also became so ubiquitous that its use has not only become rigidified – it may have become unproductive.

Judging from its resounding successes of the early twenty-first century, supervised machine learning may seem slated to be the next statistical *grande idée*. Emerging at a time of crisis for conventional statistics in general, and the common econometric use of parametric models in particular, this renewed approach to quantification may have sparked a new hope. By freeing itself from constraining and easily-violated probabilistic hypotheses, by being able to harness the ever-growing masses of data and computation power, and to automatically detect intricate patterns the researcher could not even have formulated, it bears the promise of making regression great again. While all this is true, the present article argues that it is very unlikely that machine learning will become the next methodological hegemon, at least in social science. This argument is not founded on the methods' alleged abandonment of theory: they are in no way a tool that magically extracts knowledge from raw data, nor can they be fruitfully used without the guidance of a precise research question and theoretical framework. The lack of solid understanding of their results, along with their need for large datasets to perform well, are much more serious obstacles to its prophesied rise to prominence. What machine learning does seem likely to do, however, is force scholars to denaturalize their methods. By widening the scope of possible uses of quantification, and exposing the weak points of yesterday's certainties, it is favoring a humbler attitude toward statistical results.

There are, no doubt, some unnerving aspects to machine learning. Not the least is the arrogance of some of its users, who buy into (or try to sell) the rhetoric of an always upcoming revolution. Against such naiveté, Geertz again had some cautionary words: "after we have become familiar with the new idea [...] it no longer has the grandiose, all-promising scope, the infinite volatility of apparent application it once had." Past that point, "our attention shifts to isolating [what the idea is good for], to disentangling ourselves from a lot of pseudoscience to which, in the first flush of celebrity, it has also given rise" (Geertz, 1973, p. 311).

Reference list

Abbott A., 1988, "Transcending General Linear Reality", *Sociological Theory*, 6(2), p. 169-186.

Abbott A., 2001, *Chaos of Disciplines*. The University of Chicago Press

Alpaydin E., 2010, *Introduction to Machine Learning*, MIT Press.

Athey S., 2017, "Beyond Prediction: Using Big Data for Policy Problems", *Science*, 335(6324), p. 483-485.

Baudelot C., Establet R., 1992, *Allez les filles. Une révolution silencieuse*. Le Seuil.

Biau G., Scornet E., 2016, "A Random Forest Guided Tour", *TEST*, 25, p. 197-227.

Borges J. L., 2000 [1939], "The Total Library", in *Non-Fiction 1922–1986*, Penguin Press, London, p. 214–216.

Breiman L., 2001, “Statistical modeling: The two cultures (with comments and a rejoinder by the author)”, *Statistical Science*, 16(3), p. 199–231.
Buchanan B., 2005, “A (very) Brief History of Artificial Intelligence”, *AI Magazine*, Winter 2005, p. 53-60.
Carpenter C. 2006, “Self-Reported Sexual Orientation and Earnings: Evidence from California,” *Industrial and Labor Relation Review*, 58(2), pp. 258-273.
Charpentier A., Flachaire E., Jakubowicz J., Ly A., 2017, “Econométrie et Machine Learning,” *working paper*.
Chatfield C., 1995, “Model Uncertainty, Data Mining and Statistical Inference”, *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 158(3), p. 419-466.
Copeland B. J., 2000, “What is Artificial Intelligence?”, *Alan Turing net archive*, http://www.alanturing.net/turing_archive/pages/Reference%20Articles/What%20is%20AI.html, accessed February 2018.
Crevier D., 1993, *AI: The Tumultuous Search for Artificial Intelligence*, Basic Books.
Domingos P., 2015, *The Master Algorithm. How the Quest for the Ultimate Machine Learning Will Remake Our World*, Basic Books.
Efron B., Hastie T., 2016, *Computer Age Statistical Inference*, Cambridge University Press.
Friedman J., Hastie T., Tibshirani R., 2010, “Regularization Paths for Generalized Linear Models via Coordinate Descent”, *Journal of Statistical Software*, 33(1), p. 1-22.
Geertz, C., 1973, “Thick Description. Toward an Interpretive Theory of Culture,” in *The Interpretation of Cultures*, Basic Books.
Grove R., 2011, “Three Eras of Survey Research,” *Public Opinion Quarterly*, 75(5), pp. 861-871.
Hastie T., Tibshirani R., Friedman J., 2009, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Second Edition, Springer Series in Statistics, New York, Springer.
Ioannidis J. P. A., 2005, “Why Most Published Research Findings are False”, *Plos Medicine*, 2(8).
Mincer J., 1974, *Schooling, Experience, and Earnings*, National Bureau of Economic Research.
Ollion É., Boelaert J., 2015, “Au-delà des big data. Les sciences sociales face à la multiplication des données numériques”, *Sociologie*, 6(3), p. 295-310.
Open Science Collaboration, “Estimating the reproducibility of psychological science. A large-scale assessment suggests that experimental reproducibility in psychology leaves a lot to be desired.", *Science*, 28 Aug 2015, Vol. 349, Issue 6251.
Pearl J., 2009, “Causal inference in statistics: An overview”, *Statistics Surveys*, 3, p. 96-146.
Saporta G., 2006, *Probabilité, analyse de données et statistiques*, Paris, Technip, 2d edition.
Silberzahn R., Uhlmann E. L., Martin D. P., Anselmi P., Aust F., Awtrey E. C., Bahník Š., *et al.* 2017. “Many Analysts, One Dataset: Making Transparent How Variations in Analytical Choices Affect Results”. *PsyArXiv*. September 21 (www.psyarxiv.com/qkwst, accessed January 19th 2018).
Varian H., 2014, “Big Data: New Tricks for Econometrics,” *Journal of Economic Perspectives*, 28(2), p. 3-28.
Wasserstein R.L., Lazar N.A., 2016, “The ASA's Statement on p-Values: Context, Process, and Purpose,” *The American Statistician*, 70(2), p. 129-133.
Wright M. N., Ziegler A., 2017, “ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R”, *Journal of Statistical Software*, 77(1), p. 1-17.