Prediction is not everything, but everything is prediction

Leonardo Egidi

Dipartimento di Scienze Economiche, Aziendali, Matematiche e Statistiche 'Bruno de Finetti', Universita` degli Studi di Trieste, Trieste, Italy.

E-mail: legidi@units.it

Jonah Sol Gabry

Department of Statistics, Columbia University, New York, USA.

E-mail: jgabry@gmail.com

Abstract. Prediction is an unavoidable task for data scientists, and over the last few decades, statistics and machine learning have bec~~a~~ome the most popular 'prediction weapons' in many fields. However, prediction should always be associated with a measure of uncertainty,~~;~~ because from it, ~~only~~ we can only reconstruct and falsify the model/algorithm decisions. Ma~~-~~chine learning methods offer many point-~~ ~~predictions, but they rarely yield ~~some~~ a measure of uncertainty, whereas statistical models usually do a ~~bad~~ poor job of~~in~~ communicating predic~~-~~tive results. According to ~~the~~ Popper's falsification~~ism~~ theory, natural and physical sciences can be falsified on the grounds of wrong predictions~~:~~; ~~t~~ho~~ugh~~wever, ~~for social sciences~~ this is not always true in the social sciences. We move then to a weak instrumentalist philosophy: ~~p~~Predictive accuracy is not always constitutive of scientific success, especially in the social sciences.

Keywords: Prediction; Popper's falsificationism philosophy; Weak instrumentalism; Predictive accuracy; Machine learning

1. Introduction

As motivated by ~~the~~ falsificationism, ~~approach (~~Popper~~,~~ (1934), and many philosophers of science, prediction has a primary role in the progress of science; however, this is often ~~a~~ controversial ~~argument~~ —see Kuhn (1962) and Lakatos (1976) for some criticisms. Popper argues that theories, ~~in order~~ to be scientific, must be falsifiable on the grounds of their predictions: wrong predictions should perhaps pus~~h the~~he scientists to reject their theories or to re-formulate them~~,~~; conversely, exact predictions should corroborate a scientific theory. Popper's philosophy is instrumentalist in a strong sense (Hitchcock and Sober, 2004) when applied to physical and natural sciences: predictive accuracy is constitutive of scientific success, not only symptomatic of it, and prediction works as a confirmation theory tool for science.

Since the 1940s, with the growing availability of fast computers and the use of simulation routines, science expanded its boundaries and extended the existing frameworks in new di~~r~~me~~c~~ ~~t~~nsions; ~~think,~~ for instance, think of~~at~~ the Manhattan project in Los Alamos, when the problem of neutron diffusion in fissionable material allowed Stanislaw Ulam and Nicholas Metropolis to invent and develop Markov Chain Monte Carlo Methods through the ENIAC computer. In particular, the birth and ~~the~~ growth of probabilistic and statistical methods have made the 'debut of science in society' possible, whereas the growing ability of data and the development of sophisticated

Commented [.1]: This document has been edited for consistency with US English (AmE) conventions.

Please see any unspecified guidelines for further instructions, and feel free to ignore any changes or comments that conflict with your requirements.

Suggestion: Consider a consistent font throughout.

Commented [.2]: Suggestion: *incorrect*

Commented [.3]: Edited for structure/flow.

Commented [.4]: For clarity and grammatical construction, choose either:
Popper's falsificationism
or
Popper's falsification philosophy

Commented [.5]: Edited for structure/clarity.

Commented [.6]: Because they are included in your word count but no style guide was indicated, your citations have been edited for consistency and structure.

Please ensure that each citation is accurate and complete and present in your reference list.

Commented [.7]: Edited for concision.

Commented [.8]: Acronyms should be spelled out upon first use, followed by the acronym itself in parentheses. Subsequently, only the acronym needs to be used in the text.

computational tools starting from the 1950s and 1960s opened the door to~~ the~~ data science revolution; the 1990s transformed ~~then~~ data science into a global oracle, and data scientists g~~ot~~ained more ~~and more~~ credibility as the availability of ~~new~~ modern machinery~~ies~~ grew.

For many of us, data science and statistical methods are scientific~~, by means of~~with tools de~~-~~signed to formulate a theory (model) from some evidence (data) and generalize this hypothesis by induction. Over the last few decades, statistics and machine learning (ML) have bec~~a~~ome the most popu~~-~~lar 'prediction weapons' for both social and natural sciences, including frameworks such as weather's forecasting, presidential elections, planets' motions, global warming, ~~G~~gross d~~D~~omes~~-~~tic p~~P~~roduct, etc. However, there is often a clear separation between these two fields: statistics is usually seen as a discipline ~~whic~~that extracts information from~~ the~~ current data, whereas ML~~machine learning~~ is usually designed to predict new events. ~~Th~~Ho~~ugh~~wever, many times, the right weapons are embraced by the wrong people. The predictive power in statistics is a small, ~~n~~ elegant ~~small~~ gun~~,~~ with small bullets and good properties~~ and small bullets~~, whereas in ML~~machine learning~~ it is a bazooka~~,~~ with big bullets and devastating effectiveness~~ and big bullets~~. The statistician ~~perfectly~~ knows the gun's details and how it is~~to~~ used~~ it,~~; the machine learner is rarely aware of the bazooka's properties. Much literature ~~ab~~o~~ut~~n ML~~ma- chine learning~~ methods (Breiman et al., 2001) is based on their ability to successfully predict test set data, but (almost) nothing is said about the technical assumptions required to tune/build the algorithms; conversely, many statistical methods are claimed to be good upon the check of their residuals on the training data, but rarely on the ground of some forecasting abilities on holdout samples.

The main novelty of this paper is the *weak instrumentalist* position for prediction, under which predictive accuracy is constitutive of scientific success only when the underlying statistical methods are falsifiable and transparently designed to predict out-of-sample events. In other w~~ay sai~~ords, there are many contexts, especially in the social sciences, where falsification through the prediction's fallacy should be replaced by a more consistent idea of falsification: we believe this position may be beneficial for the so-called ~~"~~hard sciences~~"~~ as well. From one side, mathemat~~-~~ical and quantitative laws formulated by Galilei and Newton were physical and deterministic~~ laws~~, ~~by means of~~with which ~~future~~ particular future facts could have been predicted wi~~th~~ the absolute precision; from the other side, probabilistic and statistical laws designed to describe human be~~-~~haviours and social facts are stochastic laws,~~ by means of~~with which particular future ~~particular~~ events could be predicted with an intrinsic amount of uncertainty. Of course, as statisticians, we want to do our best ~~in~~to predict~~ing~~ future social events, but we cannot entirely evaluate a model's performance only on the grounds of its predictive accuracy~~ only~~. Using Popper's terminology, wrong social science~~s~~ predictions should not be the only tool to falsify a theory.

Prediction is an unavoidable task for scientists working with data, but it is not all ~~what~~ we need, especially when framed in social science frameworks; moreover, prediction should always be associated with a measure of variability, because from variability only we are able to reconstruct and falsify the model/algorithm decisions. ML~~achine learning~~ methods offer many point~~-~~ predictions, but they rarely yield ~~some~~ a measure of uncertainty, whereas statistical models, when predicting new items, usually do a bad job ~~in~~of communicating the results. Weak instru- mentalist philosophy should push ~~the~~ statisticians to embrac~~e more~~e the bazooka more when needed, and the machine learners to use a more precise gun when a bazooka is unnecessary.

In Section 2, we revise the steps required to formulate a scientific theory and review the role of prediction for natural sciences from Galilei's law of falling bodies to Albert~~h~~ Einstein's general relativity~~ of Albert Einstein~~. ~~Moreover, w~~We also analyz~~s~~e the confirmation theory approach, both in natural and

Commented [.9]: Simple and concise – well done.

Commented [.10]: NB: In AmE, 'which' is used in non-restrictive phrases, and a comma is used before 'which'. Alternatively, 'that' is used for restrictive phrases, and no comma is used.

Commented [.11]: Edited for structure/flow.

Commented [.12]: Great imagery

Commented [.13]: Edited for structure/tone.

Commented [.14]: Most?

Commented [.15]: Edited for redundancy.

NB: *So-called* eliminates the need for scare quotes (quotation marks often used to alert readers that a term is used in a nonstandard/slang, ironic, or other special sense). Like any such device, scare quotes lose their force and irritate readers if overused. Consider removing these throughout.

Commented [.16]: Suggestion: *On one hand*

Commented [.17]: Edited for redundancy.

Commented [.18]: Suggestion: *On the other hand*

Commented [.19]: Suggestions:
incorrect
false

Commented [.20]: suggestion: *communicate results poorly*

Commented [.21]: NB: Try to avoid the use of transitional words/phrases (e.g., moreover, in addition, therefore, furthermore, in fact, thus) unless necessary. These should be used rarely in a text, and only when it is essential that the ideas between two different sentences be clearly linked. If the ideas in the two sentences can stand on their own, a transitional word/phrase is not needed.

This comment applies to the whole document.

Commented [.22]: Edited per AmE.

social sciences. In Section 3, we focus on predictions for statistical learning, ~~whereas~~ while a ~~the~~ weak instrumentalist philosophy is detailed in Section 4. Section 5 proposes an applied example for the football Russia World Cup 2018, ~~whereas~~ and Section 6 concludes.

2. Prediction for science or science for prediction?

2.1. I~~t i~~s prediction part of ~~the~~ science design?

The main stages required to formulate a scientific law are summarized by Russell (1931) as follows: (1) observation of some relevant facts~~;~~, (2) formulation of a hypothesis underlying and explaining the previously mentioned facts ~~above;~~, and (3) deduction of some consequences from this hypothesis. In his opinion, the modern scientific method ~~i~~was born with Galileo Galilei, father of the law of falling bodies, and with Johannes Kepler, who discovered the three laws of planetary motion:

> Scientific method, as we understand it, comes into the world full-fledged with Galileo (1564-1642), and, to a somewhat lesser degree, in his contemporary, Kepler (1571-1630). [...] They proceeded from observation of particular facts to the establishment of exact quantitative laws, by means of which future particular facts could be predicted.

Commented [.23]: Is this a quotation? If so, please include a citation to avoid plagiarism.

Then, the law of universal gravitation of Isaac Newton embodied the two previous theories, whereas the theory of the general relativity of Albert Einstein generalized ~~the~~ Newton's theory. Thus, in the last 500 years, physics—and, more generally, science—advanced by falsification and generalization of ~~the~~ previous theories, by providing new and more exciting theories to predict new natural facts, and highlighting the confirmation nature of prediction. In general, as Hitchcock and Sober (2004) argue, mathematical descriptions of the invariant behavio~~u~~r of a physical phenomenon are essentially predictive~~:~~; further experiments and observations can validate these theories.

However, ~~the~~prediction's link ~~of prediction~~ with the scientific laws is ~~in our opinion~~ more ambiguous than what people are usually inclined to think. The following questions arise: ~~i~~Is prediction a central step in science? Is prediction a relevant aim of science? A negative answer to the first question could be seen in disagreement with some *instrumentalist* scientists, who would claim that, from an instrumental perspective, predictive success is not merely *symptomatic* of scientific success~~, but it~~ — is also *constitutive* of scientific success (Hitchcock and Sober, 2004). A more sophisticated answer could be~~:~~ that prediction is not explicitly part of the formulation of a scientific hypothesis (1)–(3) *at the time the law is posed*, but it becomes relevant and relevant as science advances; the chain of events ~~whic~~that brought Newton to generalize the theories of Galilei and Kepler first, and Einstein to revisit the gravitational law of Newton then, was supposedly based on the fallacy of some predictions, and it gained sense only *ex-post*. The fact that the bodies in proximity to the earth surface were revealed by Newton to not fall exactly with a constant acceleration—the acceleration slightly rises as they get closer to the earth—did not make ~~the~~ Galilei's law of constant acceleration for falling bodies less scientific, or totally wrong from a scientific point of view. Scientific falsification detected by wrong predictions (Popper, 1934) is a powerful and exceptional tool, but ~~alo~~in~~g~~ this paper, we ~~feel to warn about~~ caution its abuse/misuse.

Commented [.24]: Edited for redundancy.

Anything not cited or referenced would be assumed to be your opinion.

Commented [.25]: Edited for structure/clarity.

~~Over the last decades, s~~Scientific prediction has recently~~s~~ bec~~a~~ome popular not only in the context of physics and natural science, but for the social sciences as well. Steps (1)–(3) above are widely used by social scientists and statisticians to build consistent theories about human and social behavio~~u~~rs:

~~r~~Recently,~~ it emerged clearly~~ the need to build a quantitative population's laws with the aim to mimic~~ the~~ physical nature's laws emerged. However, the role played by prediction in social sciences is even more obscure (Popper, 1944, 1945) and much more controversial than for the natural sciences, though data scientists ar~~e every~~e~~ry day~~ ~~mo~~incre~~-~~asi~~n~~n~~d more~~gly asked to build 'weapons of mass prediction' in ~~many~~ varying social contexts. Perhaps~~,~~ the actual outcome may be far away from the predictions: Trump's win in the US ~~P~~presidential ~~E~~election~~s~~, Brexit, and the Leicester's Premier League's win were very low-probability events, but they all occurred ~~during~~ in 2016. Can all of these rare events falsify the finest algorithms and models designed to not predict their occurrence? Our naive and tentative answer is no, they can~~no~~'t. We give more intuition~~s~~ about this in the next section.

2.2. *Prediction as a confirmation theory approach*

For Popper (~~Popper,~~ 1934), a theory is scientific only if it is falsifiable, where the falsification of a theory is meant to b~~e the~~e the possibility ~~t~~of compar~~e~~ing its predictions with the observed data. In his view, theories whose predictions conflict with any observed evidence must be rejected: predic~~-~~tion corroborates (or confirms) a theory when it survives an attempt at falsification; prediction delegitimizes a theory when it does not pass the falsification test.

The confirmation nature of prediction is crucial in the natural sciences, such as physics. In gen~~-~~eral, as Hitchcock and Sober (2004) argue, mathematical descriptions of the invariant behavio~~u~~r of a physical phenomenon—such as Newton's and Keplero's laws~~,~~ or Maxwell's equations—are essentially predictive~~:~~; further experiments and observations can validate these theories.

A well-known historical example of predictive confirmation in chemistry dates back to the middle of the 19th century—see Maher (1988) for a detailed version of the example. At that time, more than 60 chemical elements were known, and new ones continu~~ing~~ed to be discov-ered. Some prominent chemists attempted to determine their atomic weights, densit~~y~~ies, and other properties~~,~~ by collecting ~~many~~ experimental observations. In 1871, the Russian chemist Dmitri Mendeleev noticed that arranging the elements by their atomic weights, valences, and other chemical properties tended to show~~ a~~ periodical recurrences. He found some gaps in the pattern~~,~~ and ~~he~~ argued that these missing values corresponded to some existing elements ~~whic~~that had not yet been discovered.~~:~~ ~~H~~He named three of these elements (eka-alumin~~i~~um, eka-boron, and eka- silicon) and gave~~, by giving~~ ~~some~~ detailed descriptions of their properties. Despite the skepticism of the scientific community, ~~the~~ French Paul-Emile Lecoq de Boisbaudran in 1874, ~~the~~ Swedish Lars Fredrik Nilson in 1878, and ~~the~~ German Clemens Winkler in 1886 discovered three elements ~~whic~~that corresponded to descriptions of eka-alumin~~i~~um, eka-boron, and exa-silicon~~, respectively~~: these three elements are ~~better known~~ now respectively known as gallium, scandium, and germanium. ~~The predic- tive ability of~~ Mendeleev's predictive ability was remarkable—the Royal Society awarded him the Davy Medal in 1882—~~,~~ and the newly discovered elements ~~well~~ represented pieces of evidence ~~whic~~that confirmed the theory.

Predictive confirmation is still ambiguous in the social sciences. As argued by Popper (1944, 1945) and Sarewitz and Pielke Jr. (1999), the social sciences have long tried to emulate physical sci~~-~~ences in developing invariant mathematical laws of human behavio~~u~~r and interaction to predict economics quantities, elections, policies, etc.; many scholars agreed ~~about the fact~~ that a social theory should be judged on its power to predict (Friedman, 1953).

However, we believe that social science predictions require more ~~and more~~ motivation~~s~~ to validat~~e~~ the underlying theori~~y~~es. In the 2016 United States ~~P~~presidential ~~e~~Election, ~~the~~ Republican Donald Trump defeated~~ the~~ Democrat Hillary Clinton by winning the ~~e~~Electoral ~~c~~College (304 vs

Commented [.26]: Edited for concision/tone.

Commented [.27]: Solid question

Commented [.28]: Suggestion: This could be cut, and your meaning would remain intact.

Commented [.29]: NB: Avoid contractions in formal/academic writing.

Commented [.30]: This is unclear (intuition is a sense that cannot be 'given' or 'provided') – do you mean *we give more detail on this . . .*?

Commented [.31]: Suggestion:

see Maher (1988).
see Maher (1988) for details/an example.

Commented [.32]: Suggestion: *particular chemical properties*

Commented [.33]: NB: AmE prefers the serial (Oxford) comma.

Commented [.34]: Edited for structure/flow/concision.

Commented [.35]: Edited for concision.

Commented [.36]: Suggestion:

Donald Trump (Republican) defeated Hillary Clinton (Democrat)

Commented [.37]: Whether you place a period/full stop after *vs* will depend on your style guide.

If you do make changes to this selection, please make the same changes throughout the document to remain consistent.

227), but gaining a lower voter~~s'~~ percentage (46.1% vs 48.2%). According to various online poll aggregators, ~~Hillary~~ Clinton was given a 65% or 80% or 90% chance of winning the electoral college. As Gelman (2016b) argues:

> These probabilities were high because Clinton had been leading in the polls for months; the probabilities were not 100% because it was recognized that the final polls might be off by quite a bit from the actual election outcome. Small differences in how the polls were averaged corresponded to large apparent differences in win probabilities; hence we argued that the forecasts that were appearing, were not so different as they seemed based on those reported odds. The final summary is that the polls were off by about 2% (or maybe 3%, depending on which poll averaging you're using), which, again, is a real error of moderate size that happened to be highly consequential given the distribution of the votes in the states this year.

In November 2016, many modelers, inclu~~de~~ding Nate Silver, the founder of the well-known FiveThir~~-~~tyEight blog (https://fivethirtyeight.com), failed to predict ~~the~~ Trump's~~s'~~ win. However, it is naive to conclude that th~~o~~ese models failed because their underlying mechanism was wrong; rather, political science predictions cannot ~~entirely~~ act as theory's only confirmation tools~~, due to~~ for many reasons ~~attributed~~, for instance, ~~to~~ nonresponse and voter~~s'~~ turnout, as explained by Gel~~-~~man (2016a):

> Yes, the probability statements are not invalidated by the occurrence of a low-probability event. But we can learn from these low-probability outcomes. In the polling example, yes an error of 2% is within what one might expect from nonsampling error in national poll aggregates, but the point is that nonsampling error has a reason: it's not just random. In this case it seems to have arisen from a combination of differential nonresponse, unexpected changes in turnout, and some sloppy modeling choices. It makes sense to try to understand this, not to just say that random things happen and leave it at that.

3. The role of prediction in statistical learning: what we ~~usually~~ do, what we ~~n~~do not do, what we should do

3.1. From the observed to the observable

As statisticians, we often deal with a double task: first, creating a sound mathematical model to accommodate the data and retrieve useful inferential conclusions from parameters' estimates—in this section, we make no distinction between classical and Bayesian inference; second, using this model to make predictions. From a practical point of view, inference and prediction should act sequentially and appear as "~~"~~two sides of the same coin"~~"~~. ~~And both of them should~~ while contribut~~e~~ing to the ~~to the~~ statistical workflow by coherently accounting for ~~the~~ intrinsic model uncertainty.

However, the widespread feeling is that statistics has always been ~~thought as~~considered the *science of inference*, or *science of estimates*, and inference is ~~most of the time~~often considered ~~as~~ the dom~~-~~inan~~t~~ing side and seen as separate from prediction (adding reference?). Inference creates an underlying mathematical model of the data-generating process (Bzdok et al., 2018)~~,~~; its main task is to formulate a theory that adequately captures an unknown mechanism connecting po~~-~~tentially influential predictors with a response variable; the inferential laws should be as general

Leo: I slightly re-vised this part, I guess now this first paragraph could be ok and could better link the second paragraph

Jonah: I guess you agree with this viewpoint. Maybe adding a

Commented [.38]: Please verify the accuracy of all direct quotations prior to submission.

This comment applies to the whole document.

Commented [.39]: Edited per AmE.

Commented [.40]: Edited for concision/tone.

Commented [.41]: Edited for word choice/flow.

Commented [.42]: Please provide a citation here.

reference?

as possible, ideally valid for the population of interest, and not symptomatic of the observed data (it is out of the scope of this paper to review the distinct inferential approaches). Prediction instead moves from the observed to the unobserved (though observable in the future), being the action designed to forecast future events without requiring a full understanding of the under~~- der~~lying data-generation process. Predictive actions can be daily shared and accepted from the human common sense: each person is in fact more or less confident with weather's predictions or with presidential election predictions, but rarely that person is aware of the underlying statis~~-~~tical model required to produce that forecast, unless he is a statistician/data scientist. In such a

Leo: Second paragraph: slightly re- vised

~~-~~view, inference seems ~~hard~~ difficult and obscure, and prediction is easy and transparent to the people.

Commented [.43]: Suggestion: *simple*

Formatted: Indent: Left: 1.21", Right: 0.99", Space Before: 5.8 pt, Line spacing: Multiple 1.02 li

This is a paradoxical argument, since inference is often associated wi~~to~~th the action of *explaining* a given problem, and its results should be relevant and available to the majority of the popula~~-~~tion. However, statisticians look like those magicians who resemble their decks of cards by fictitiously adding/removing ~~some~~ extra cards~~,~~ that are not available to the audience. To continue with the metaphor, model parameters behave like these fake cards, which are incredibly relevant to build the trick (aka statistical theories), but do not exist in ~~the~~ real life; in brief, parameters are

Commented [.44]: Suggestion: This could be cut, and your meaning would remain intact.

Commented [.45]: Suggestions: *are* *operate*

Jonah: Do you like the magician metaphor? Edit as you wish

~~-just some~~ fictitious and technical devices used to explain and approximate ~~the~~ complexity. Rather, only prediction links the observed with the observable and is accessible to the people: it is never a matter of parameters' interpretation~~,~~; it only requires a check of the discrepancy between

Formatted: Indent: First line: 0.21", Space Before: 3.45 pt

Commented [.46]: Edited for structure/concision.

~~-~~observed and future events, and can doubtless be done by anyone.

Formatted: Indent: Left: 1.21", First line: 0.21", Right: 0.99", Space Before: 3.45 pt, Line spacing: Multiple 1.02 li

When framed in a predictive task, many technical questions arise. First ~~of all~~, should we use all the data to build a reasonable/useful model~~,~~ or take only a portion of the sample to accomm~~o~~- date the model (the training set), using the remaining values ~~to~~ for valida~~te~~tion and testing ~~the model~~ (the validation and test sets)? Is an overfitting model suited enough for predictive purposes in out-of- sample scenarios? Should we trust more a model ~~whic~~that ~~w~~accurate~~l~~ly accommodates the current data or a model/algorithm with a~~n~~ high predictive accuracy for future predictions? These and many oth~~-~~ ers apparently naive questions pushed many scholars to debate about the supposed supremacy of

Leo: Third paragraph: revised

Commented [.47]: Edited for tone/structure.

Jonah: Do you like this?

~~-~~prediction over accommodation (Maher, 1988; Hitchcock and Sober, 2004; Worrall, 2014). Ac~~-~~cording to his own point of view, the statistician should ask himself whether

Formatted: Indent: First line: 0.21", Space Before: 3.5 pt

~~-~~he wants models that are true—or approximately true—or predictively accurate.

Formatted: Indent: Left: 1.21", First line: 0.21", Right: 0.99", Space Before: 3.5 pt, Line spacing: Multiple 1.02 li

~~In our opinion t~~There is not a clear domain of one approach over the other: inference and prediction are not enemies, but could be strong allies to reveal the truth. Moreover, even more importantly, we strongly believe that (almost) everything in statistics is predictive, or, at least, may be read ~~unde~~from a predictive point o~~f~~v view. Even though many statisticians seek to mask their theories/models only by claiming the relevance of their estimation/inferential process, they are rarely aware of the predictive essence underlying their procedures. For illustration~~s~~ purposes, consider a logistic/probit regression model for diagnosing diabetes testing positive probabilities using biochemical variables (such as glucose, insulin, mass, etc.) as predictors. A summary for these kinds of models is usually documented by the adoption of odds-ratios, con- fidence/predictive intervals for the parameters, *p*-values and other estimation-driven measures. However, the hidden task of such a model is ~~intrinisecally~~intrinsically predictive: rather than ~~overly~~ hyperfocus- ing on the numerical impact of the classical/Bayesian estimates, the focus here should be on the general probability of being tested positive to the diabetes, by considering existing or new values for glucose, insulin, etc~~.~~. ~~As a further example, c~~Consider the usual randomized clinical trials set to assess some drugs' efficacy as another example: these studies are rarely conducted with the idea of predicting a useful and healthy behavior in the population (the hidden and final aim, according to us)~~, rather~~.

Leo: Fourth paragraph: revised

Commented [.48]: Edited for concision.

Commented [.49]: Edited for word choice/tone.

Commented [.50]: Edited for structure/tone.

Rather, they are built to find and display a statistically significant effect, according to the statistical significance of some parameters. This estimation obsession makes statistics obscure, whereas speaking in predictive terms, openly accessible to the population, would make statistics more transparent. This is the reason why we provocatively claim that *prediction is not everything~~,~~ but (almost) everything, in statistics, could be described in predictive terms.*

Jonah: This paragraph is new, and tries to collect our recent ideas. Edit as you wish

Commented [.51]: Please verify all variables/equations/formulae prior to submission.

3.2. *Generalization performance of a statistical model*

Leo: I revised a bit this section and I change its title. I guess now is more clean

Assessing the generalization performance of a learning method related to the predictive performance on a test set is a central task in modern data~~-~~ science. In this section, we review some well-known approaches and highlight their main merits and weaknesses.

Suppose we use a set of observations $\tau = (x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ to fit (or train) the sta~~-~~tistical model $Y = f(X) + \varepsilon$, where the single x_i is the observed value of the predictor/covariate X_i, the single y_i is the observed value of the dependent/response variable Y_i, f is an unknown mathematical function of X, and ε is the random error. Fitting a model means producing an estimate $\hat{f}$ for the unknown f: in the univariate linear regression case $f(X) = \beta_0 + \beta_1 X$.~~,~~ T~~t~~his is translated in estimating the parameters β_0, β_1 by finding $\hat{\beta}_0, \hat{\beta}_1$. Let (x_{n+1}, y_{n+1}) be a new observation not used to train the model, we then would like to take the expectation across all such new values and define the *test mean square error (MSE)*, or *generalization error*:

$$\text{Test MSE} \equiv \mathrm{E}\left[(y_{n+1} - \hat{f}(x_{n+1}))^2 | \tau\right], \tag{1}$$

where $\hat{f}(x_{n+1})$ is the prediction for y_{n+1} produced by $\hat{f}$, the expectation is taken across all new unseen predictor-response pairs (x_{n+1}, y_{n+1}) and the training set τ is cons~~d~~idered to be fixed. Unfortunately, it is commonly unfeasible to calculate the test MSE~~,~~ because we are often in a situation in wh~~ere~~ich we do not have any test data available. As suggested by Hastie et al. (2009), it does not seem possible to estimate the conditional error in (1) appropriately~~,~~ given only the information in the same training set τ. For such reasons, we may introduce a related quantity, the *expected test MSE*, or *expected generalization error*, which is usually estimated by cross~~-~~-validation (Hastie et al., 2009) and related statistical methods:

$$\text{exp t}\cancel{\text{T}}\text{est MSE} \equiv \mathrm{E}[\text{Test MSE}] = \mathrm{Var}(\hat{f}(x_{n+1})) + \mathrm{Bias}^2(\hat{f}(x_{n+1})) + \mathrm{Var}(\varepsilon), \tag{2}$$

Commented [.52]: This has been revised for the sake of consistency.

where the expectation is taken across many training sets, $\mathrm{Var}(\hat{f}(x_{n+1}))$ is the variance for $\hat{f}(x_{n+1})$, $\mathrm{Bias}^2(\hat{f}(x_{n+1}))$ is the squared bias, $\mathrm{Var}(\varepsilon)$ is the error variance. This final term, ~~is~~ known as the irreducible error, is the minimum lower bound for the test MSE. Since we only ever have access to the training data points (including the randomness associated with the ε values) we cannot~~'t~~ ever hope to get a "more accurate" fit than what the variance of the residuals offer.

The predictive goal is to select the model where the expected test MSE is lowest: the statisticians should try to minimize ~~the~~ exp t~~T~~est MSE by choosing the model that simultaneously has low variance and low bias. However, this is a relevant and difficult challenge, much debated within the statistical community: more complex models, with lower bias, tend to overfit the data~~,~~ by yielding poor predictive results and then higher variance; conversely, too simple mod- els tend to not fit the data adequately and have higher bias. Statistical procedures often incur in the *bias–variance trade-off* (James et al., 2013)~~,~~; the challenge is to find a compromise by controlling for both ~~the~~ bias and ~~the~~ variance.

Leo: Maybe this part can be dropped/reformulated. Is the expected test MSE relevant here? how is connected with the topics?

Formatted: Font: 8 pt

When building a model for real-life applications to extract information from the data, it is good practice to keep in mind this bias~~-~~–variance trade-off. Nevertheless, it is often problematic to assess the performance of a statistical model by looking directly at the elements in Equa- tion (2). For this~~such a~~ reason, many statistical methods such as cross-validation, bootstrap, Akaike Information Criterion (AIC)~~IC~~, and BIC have been proposed to provide reasonable estimates of the expected generalization error and~~ then~~ help model~~l~~ers ~~in~~ choos~~e~~ing ~~among~~ candidate models. However, not all of these meth~~-~~ods are (always) effective in estimating the prediction error on out-of-sample datasets: AIC and Deviance Information Criterion (DIC)~~IC~~ suffer from the conditioning on a point estimate, by estimating the performance of the plugin predictive density, as claimed by Gelman et al. (2014), whereas cross-validation is ap~~-~~pealing but can be computationally expensive and also is not always well defined in dependent data ~~data~~ settings.

Jonah: This last conclusion could be revised

3.3. *Information criteria*

In our practice, prediction should not be assimilated to "~~‘~~take a rabbit out of a hat~~’~~," but to look~~ing~~ at its inherent uncertainty. Splitting the predictions' uncertainty in variance and squared bias has been proved to be useful from a theoretical point of view~~,~~; however, it can appear bogus and artificial when framed in practical data analysis: how long does it take to control~~ling~~ and lower~~ing~~ the bias and the variance of a learning method? How much should the statistician stretch his model to avoid problematic bias~~-~~–variance tradeoffs?

Jonah: What do you think? I've edited slightly this paragraph

In ~~much previous~~the literature ~~about~~ on predictive accuracy, as for the AIC~~kaike Information Crite- rion (AIC)~~ (Akaike, 1973), there is no~~t any~~ role played by the model's uncertainty, since the measure of the model's accuracy is evaluated conditionally on parameters' point~~s~~ estimates, the maximum likelihood point estimate. Even the DIC~~eviance Information Criterion (DIC)~~ (Spiegelhalter et al., 2002), for many years a milestone for Bayesian model comparisons, is conditioned on a plugin~~-in~~ estimate, the posterior mean, with the number of parameters of AIC replaced by a measure of effective number of parameters.

Leo: Slightly revised

Rather, if we are framed in a Bayesian context, we intend the unobserved and future values $\tilde{y}$ to come from the posterior predictive distribution, denoted here by $p(\tilde{y}|y)$, which incorporates the intrinsic uncertainty propagating from the parameters—summarized by the posterior distribution—to the observable future values. Recent proposals, such as the Watanabe-Akaike Information Criteria (WAIC) (Watanabe, 2010) and Leave-One-Out cross~~—~~validation Information Criteria (LOOIC) (Vehtari et al., 2017), go in the direction of data granularity~~,~~ by defining~~tion of~~ the expected log~~-~~–pointwise predictive density for a new dataset (ELPPD). These approaches require the computation of the log-pointwise predictive density $p(\tilde{y}_i\, y)$ for each new observable value $\tilde{y}_i$ and have the desirable property of averaging over the posterior distribution.

Jonah: You are the expert here, dom you like it?

Although all the predictive information criteria may fail in some practical situations, LOOIC and WAIC offer the possibility ~~to~~of provid~~e~~ing a measure of predictive accuracy based on ~~the~~ sin- gle data points~~,~~ in a computationally efficient way (both the methods are implemented in the loo R package (Vehtari et al., 2019)). Despite not conclusive for the predictive accuracy of a statistical model, these techniques allow in many situations to compare distinct models by ~~the~~ acknowledg~~ement~~ing~~ of~~ an intrinsic uncertainty propagating from the parameters to the observable future values: in such a viewpoint, *observable values*, and not *parameters*, are really relevant. A transparent predictive tool should encompass data, parameters, and future data~~ all together~~, not focusing on parameters estimates/plugin~~-in~~ predictive densities alone~~:~~; in such a way, the fal~~-~~sification of a single piece makes the joint model falsifiable. In Section 4, we make this point even more

Commented [.53]: Edited for structure/flow.

Commented [.54]: This is only used once (here). Consider reworking or deleting.

Formatted: Font: Not Italic

Formatted: Font: Not Italic

Commented [.55]: Suggestion: *germane*

Commented [.56]: Edited for structure/clarity.

Formatted: Font: Not Italic

~~even more~~ clear.

Jonah: What do you think?

Formatted: Font: 8 pt

3.4. The two cultures

Leo: I am not sure this is the right place for this section...maybe moving at the beginning of section 3?

Formatted: Font: 8 pt

Formatted Table

As brilliantly argued by Breiman et al. (2001), there are two cultures in the use of statistical modeling to reach conclusions from data: a stochastic data model consisting of predictors, parameters and random noise to explain the response variable *y* is adopted by the data modeling culture; a function of the predictors to predict the response variable *y* is assumed by the algorithmic modeling culture, also named ~~machine learning (~~ML~~)~~ culture. The two approaches strongly differ in ~~strongly differ in~~ their validation: goodness-of-fit tests vs. predictive accuracy on out-of-sample data. It is evident that the data modeling culture—linear regression, generalized linear models, Cox model, etc.—is aimed at extracting some information about how nature is associating the response variable to the dependent variable, whereas the algorithmic culture—decision and classification trees, neural nets—is more oriented to predicting future values of the response variable given the values of the predictors.

The historical appeal of the ML field dates back the mid-1980s, when neural nets and decision trees became incredibly popular (Breiman et al., 1984) in areas where parametric data models were not applicable, such as speech ~~recognition~~, image ~~recognition~~, and handwriting recog~~-~~nition~~,~~ and prediction in financial markets. In analy~~z~~sing real data from these fields, the only criterion to evaluate these algorithms was predictive accuracy: this is translated in finding an
algorithm $f(x)$ able to be a good predictor for *y* for future values of *x*, the so~~-~~called *test set*. To alleviate the degree of overfitting and the lack of predictive robustness in decision trees~~,~~ in the mid-1990s, some data scientists argued that by aggregating many trees and perturbing the train~~-~~ing set, using bagging (Breiman, 1996), boosting (Freund et al., 1996), or random forests (Ho, 1995)~~,~~ dramatically increased the predictive accuracy of the trees~~,~~ by decreasing the variance.

Data scientists are used to training their procedures on the *training set*, which is chosen at the beginning ~~in many possible ways~~. A common strategy is to select the first half of a dataset to train the algorithm, and the second half to test it; another strategy consists of selecting only a percentage—say, ~~the~~ 75% of the dataset—and us~~e~~ing the remaining ~~25%~~percentage to test the algorithm. However, a small change in the dataset can cause a large change in the final predictions, and some adjustments are often required to increase the algorithm's robustness. This "~~"~~training shak~~-~~ing"~~"~~ is popular despite controvers~~ialy,~~ and ~~gives~~ allows us ~~the possibility~~ to discuss ~~about~~ the eventual supremacy of the prediction ~~supremacy of the prediction~~ over model construction and interpretation.

Commented [.57]: Edited for structure/flow.

Leo: Slightly edited and shortened

In what follows, w~~e feel to~~ relax these boundaries and merge the two cultures~~,~~ by referring to this fusion as the field of statistical learning. The list of considerations contained in the next section is valid for ~~both~~ algorithmic and stochastic modelers aimed at predictive purposes.

Jonah: what do you think?

4. Predictive instrumentalism and how to make predictive models transparent and falsifiable

4.1. Weak instrumentalism philosophy

Jonah: I relaxed the distinction between ML and statistics, and talk more generally in terms of statistical learning

As statisticians and (data) scientists, demanded to build models for social and physical sciences, our efforts should be addressed to produce good, transparent and well posed algorithms/models, and make them falsifiable upon a strong check (Gelman and Shalizi, 2013).

However, lately the need for powerful prediction weapons emerged, especially in the machine learning field, and the goodness of a modeling procedure is often associated with its

Table 1. Weak instrumentalism summary

General science	
p1	Predictive accuracy is not always constitutive of scientific success
p2	Scientific falsification on the ground of wrong predictions is sometimes misleading, especially in social sciences (Trump’s election, Leicester win, Brexit)
p3	Supposedly valid scientific theories should exist before the future data have been revealed
p4	Prediction is not explicitly part of the formulation of a scientific hypothesis at the time the law is posed, but it becomes relevant and relevant as science advances
Statistics	
p5	Take care of variability in the statistical predictions
p6	If necessary, go beyond the distinction between inference and prediction, and consider a joint model for data, parameters and future data (falsificationist Bayes)
p7	Rather than reasoning in terms of variance and bias, reason more in terms of predictive information criteria and posterior predictive distribution
Machine Learning	
p8	‘Shaking the training set’ to improve predictive accuracy is an obscure step
p9	Avoid to tune the algorithm with the only task to improve predictive accuracy
p10	To be falsifiable, ML techniques need to be transparently posed

predictive ability on out-of-sam~~o~~ple scenarios. As a consequence, only good predictive models are retained, whereas the others, even when sophisticated and well built, are discarded; predictive accuracy became the only discrimination’s tool to decide between good and bad statistical models/algorithms. We refer to this philosophical position as *strong instrumentalism*, for which the predictive accuracy carried out by the algorithms is constitutive—and not only symptomatic—of ~~the~~ broader scientific success. People strictly adhering to this philosophical perspective are usually inclined toward~~s~~ some “munging” procedures, such as “shaking the train- ing set”, or “over-tuning” some tuning parameters to ensure lower variance and higher accuracy; for the most of the time, these data scientists seem apparently ready to do “whatever it takes” to improve over the previous methods.

It is worth ~~to~~ stressing that evaluating a model/algorithm in light of its ability to predict future data is not shameful at all; conversely, it turned out to be beneficial in many areas, for instance where a parametric stochastic model failed to be really generative and useful. However, even if predictions of future data were good tools to falsify a posed theory, ~~many times~~ some strong instrumentalist techniques — lack ~~of~~ a general and valid theoretical framework. As an illustrative example, the number of predictors at each split of a random forest is a tuning parameter fixed at $\sqrt{p}$ in most cases, but in practice, the best values for these parameters will depend on the problem; if the method (the theory) is tuned and selected on the grounds of its predictive accuracy, the underlying theory to be falsified is bogus, and not posed in a transparent way.

As mentioned above, our skepticism regards the recently dominant role of prediction in falsifying our models~~,~~; for such a reason, we ~~would claim to~~ adhere to the *weak instrumentalism* position: in brief, predictions and predictive accuracy are a central task of science, but only sometimes do they ~~are~~ constitut~~iv~~e ~~of~~ scientific success.

Leo: Slight revision and cut

Leo: Slight revised

Commented [.58]: Suggestion: *considers*

Commented [.59]: Edited for concision/tone.

4.2. The falsificationist Bayesianism framework: going beyond inference and prediction

Gelman and Shalizi (2013) argue that a key part of Bayesian data analysis regards~~ the~~ model checking through posterior predictive checks. In such a view, the prior is seen as a testable part of the Bayesian model and is open to falsification~~:~~; from such intuition, Gelman and Hennig (2017) name this framework *falsificationist Bayesianism*.

As stated by Gelman et al. (2013), the process of Bayesian data analysis can be idealized by dividing it into the following three steps:

(a) Setting up a full probability model——a joint probability distribution—for all observable and unobservable quantities in a problem. The model should be consistent with knowledge about the underlying scientific problem and the data collection process.

(b) Conditioning on observed data: calculating and interpreting the appropriate posterior distribution~~,~~ (i.e., the conditional probability distribution of the unobserved quantities of ulti~~-~~mate interest)~~,~~ given the observed data.

(c) Evaluating the fit of the model and the implications of the resulting posterior distribution: ~~h~~How well does the model fit the data, are the substantive conclusions reasonable, and how sensitive are the results to the modeling assumptions in step (a)? In response, one can alter or expand the model and repeat the three steps.

In the above paradigm, predictions are never mentioned. ~~But~~However, this does not mean that predictions are not relevant in the Bayesian paradigm. Denoted by $\tilde{y}$ the unobserved vector of future values, we may derive the posterior predictive distribution as

$$p(\tilde{y}|y) = \int p(\tilde{y}|\vartheta)p(\vartheta|y)d\vartheta, \tag{3}$$

where $p(\vartheta|y)$ is the posterior distribution for ϑ, whereas $p(\tilde{y}|\vartheta)$ is the likelihood function for future observable values. Equation (3) ~~m~~ca~~y~~n be resamp~~b~~led in the following way:

$$p(\tilde{y}|y) = \frac{p(\tilde{y}, y)}{p(y)} = \frac{1}{p(y)}\int p(\tilde{y}, y, \vartheta)d\vartheta. \tag{4}$$

From Equation (4), we immediately notice that whenever we are interested in predictions, we need to consider a joint model $p(\tilde{y}, y, \vartheta)$ for both the observed data y and the unobserved quantities $\tilde{y}, \vartheta$. This joint model incorporates both the likelihood and the prior, being $p(\tilde{y}, y, \vartheta) = p(\tilde{y}|\vartheta)p(y|\vartheta)p(\vartheta)$. Thus, the joint model for the predictions, the data, and the parameters is transparently posed~~,~~ and open to falsification when the observable $\tilde{y}$ becomes known.

To summarize the above discussion, we collect in Table 1 the main points ~~whic~~that follow from the weak instrumentalist philosophy. We divided them into three categories: the first one collects general considerations about the role of prediction in modern science and data science, whereas the second and the third one propose some tips for statisticians and machine learners, respectively.

5. Applied example: ~~football~~ Russia World Cup 2018

Commented [.60]: Edited per AmE.

Commented [.61]: Edited for word choice/tone.

Commented [.62]: Edited for clarity. Please check.

Jonah: I moved here the table, what do you think?

In this section, we review and motivate a simple example of football prediction in light of the weak instrumentalist philosophy proposed in the previous section and summarized in Sec- tion ??. In particular, we put in evidence the influence of the training set for future predictions by revealing some paradoxical considerations in ML results from a small-sample case. We con- sider here the dataset containing the results of all ~~the~~ 64 tournament's matches (48 of the group stages, and 16 of the knockout stage) for the FIFA World Cup 2018 hosted in Russia and won by France.

Commented [.63]: Please check.

Let (y_n^H, y_n^A) denote the observed number of goals scored by the home and ~~the~~ away team in the n-th game, respectively. A general bivariate Poisson model allowing for goals' correlation (Karlis and Ntzoufras, 2003) is the following:

$$
\begin{aligned}
Y_n^H, Y_n^A \mid \lambda_{1n}, \lambda_{2n}, \lambda_{3n} &\sim \text{BivPoisson}(\lambda_{1n}, \lambda_{2n}, \lambda_{3n}) \\
\log(\lambda_{1n}) &= \vartheta + \text{att}_{h_n} + \text{def}_{a_n} + \frac{\gamma}{2} w_n \\
\log(\lambda_{2n}) &= \vartheta + \text{att}_{a_n} + \text{def}_{h_n} - \frac{\gamma}{2} w_n \\
\log(\lambda_{3n}) &= \theta_0,
\end{aligned}
\tag{5}
$$

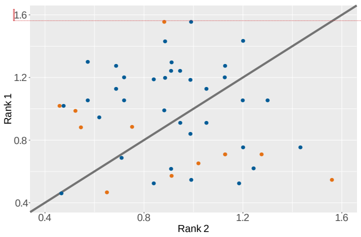
where the case $\lambda_{3n} = 0$ is reduce~~s~~d to the double Poisson model (Baio and Blangiardo, 2010). $\lambda_{1n}, \lambda_{2n}$ represent the scoring rates for the home and the away team, respectively, where: ϑ is the common baseline parameter; the parameters att_T and def_T represent the attack and the defen~~c~~se abilities, respectively, for each team T, $T = 1, \dots, N_T$; the nested indexes $h_n, a_n = 1, \dots, N_T$ denote the home and the away team playing in the n-th game, respectively; the only predictor is $w_n = (\text{rank}_{h_n} - \text{rank}_{a_n})$, the difference of the FIFA World Rankings (https://www.fifa. com/fifa-world-ranking/)—expressed in FIFA ranking points divided by 10^3—between the home and the away team in the n-th game, multiplied by a parameter $\gamma/2$. This last term tries to correct for the well-known phenomenon of *draw inflation* (Karlis and Ntzoufras, 2003), favo~~u~~ring ~~the~~ draw occurrence when teams are close in terms of their FIFA rankings. The value of the FIFA ranking difference w included in the models was considered on June 7th, only a ~~bunch~~ few ~~of~~ days before the tournament ~~takes~~ took place. In a Bayesian framework, attack and defen~~c~~se parameters are usually assigned some noninformative prior distributions (Baio and Blangiardo, 2010) and imposed a sum-to-zero constraint to achieve identifiability.

Formatted: Space Before: 3.7 pt, Line spacing: Multiple 0.89 li

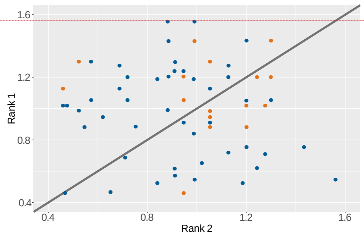
We decided to train our statistical models/ML techniques on distinct portions of matches from the group stage, where teams are more heterogeneous in terms of their FIFA rankings and actual strengths. To assess predictive performance between statistical models and ML al- gorithms in predicting football outcomes, we compare the double Poisson and the bivariate Poisson model, fitted by rstan package (Stan Development Team, 2018), with five ML proce- dures: Random Forest, Classification and Regression Trees (CART), Bagged CART, Multivari- ate Adaptive Regression Splines (MARS), and Neural Network, according to their standard use as provided by the caret package (Kuhn, 2019). The three different prediction scenarios are as:

A *Train* 75% of randomly selected group stage matches
Test Remaining 25% group stage matches

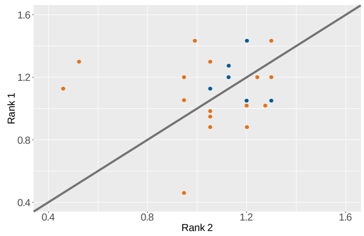
B *Train* Group stage matches
Test Knockout stage



(a) Scenario A



(b) Scenario B



(c) Scenario C

Figure 1. For each prediction scenario~~s~~, the values of the FIFA rankings for each match are shown in blue ~~colour~~ for the training set and in orange ~~colour~~ for the test set.

Commented [.64]: Static images are unable to be edited with this service. Please ensure that all information in each image is accurate and complete according to your guidelines.

This comment applies to the whole document.

Commented [.65]: Edited for redundancy.

Commented [.66]: Edited for redundancy.

C *Train* Group stage matches for which bo~~th~~ the teams have a ~~Fifa~~ FIFA ranking greater than 1
Test Knockout stage.

Figure 1 displays for each scenario the values for the FIFA rankings for the training set matches (blue points) and the test set matches (orange points), along with the line Rank 1 = Rank 2, implying that the ranking difference is $w = 0$. In Scenario A, the test set matches are randomly selected from the group stage, and they do not show any particular pattern around ~~the~~ line $w = 0$. In Scenarios~~scenarios~~ B and C, test set matches belong to the knockout stage, where the teams are expected to be stronger and closer to each other in terms of their rankings. In fact, the majority of the orange points (13 out of 16) ~~is~~are displayed toward~~towards~~ the bottom right corner—higher rankings—and closer to the line $w = 0$—closer strengths. Scenario B uses more and more data to predict test set results—all ~~the~~ 48 group stage matches—whereas Scenario C only six matches.

Commented [.67]: Edited per AmE.

Formatted: Space Before: 0 pt, Line spacing: Multiple 0.85 li

Commented [.68]: Suggestion: Consider parentheses here to avoid confusion with too many consecutive em dashes.

Figure 2 depicts the posterior predictive distribution (p5 and p7) of the number of goals scored by France and Croatia during the final from the bivariate Poisson model. Darker regions are associated with higher probabilities, whereas the red squa~~d~~re ~~is in~~ correspond~~ence~~s with the observed result, 4-2. From this plot, one could be tempted to conclude that the bivariate Poisson model completely failed to predict the match; however, the global probability of France winning

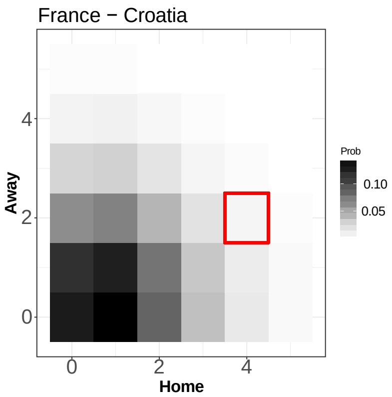


Figure 2. Posterior prediction distribution of the goals for the final France~~-~~–Croatia

within the 90 minutes—obtained summing the single probabilities over the lower triangle of the plot—is about 42%, against the 29% chance of winning for Croatia (p1 and p2). F~~r~~or this plot, only we can acknowledge the intrinsic variability in our model predictions (p5).

To ~~h~~ga~~ve~~in a glimpse ~~abou~~into statistical and ML procedures' predictive performance, Table 2 shows the accuracy in the predictions for the seven methods and the three scenarios. Assuming that higher predictive accuracies should not entirely suggest the best scientific methods (p1), we analyz~~s~~e the performance of the methods by focusing on pro and cons. As suggested by Figure 1a, Scenario A is the~~ most~~ nois~~y~~iest in terms of rankings' differences, ~~be~~wi~~ng~~th its test set constituted by matches randomly chosen from the group stage, without any kind of pattern. As~~ it~~ is intuitive, ML techniques (Random Forest and Neural Nets)~~,~~ perform better, since they 'shake' the training set (p8) in such a way as to retrieve the highest predictive accuracy. The ML performances dramatically decrease in Scenarios B and C, where learning from the training set should be focused on predicting the knockout stage. ML algorithms learn less and in a very random way, but it is not clear why (p10). As already argued, the choice of the training and the test set can dramatically change the predictive performance of the ML algorithms, which over~~-~~perform statistical models only when considering a portion of the group stage to predict the remaining group stage matches. Should ~~mayb~~we ~~w~~perhaps conclude that statistical models are better scientific tools to predict the World Cup? Not at all (p1), but we can learn from this example to improve over the next World Cups (p4).

By concluding, from this simple case~~-~~ study, we cannot openly falsify our statistical/ML tech- niques on the ground of future predictions. However, Poisson models seem to be less sensitive to the training set structure, and then falsifiable in a broader sense.

Table 2. Prediction accuracy for the selected methods~~,~~ according to three prediction scenarios.

Train	75% group	100% group	rank > 1
Test	25% group	knockout	knockout
Random forest	0.67	0.25	0.44
Bagged CART	0.67	0.31	0.37
CART	0.58	0.31	0.19
MARS	0.58	0.38	0.49
NN	0.67	0.25	0.44
Double Pois.	0.58	0.50	0.56
Biv. Pois.	0.58	0.56	0.56

6. Discussion

Prediction is central in the progress of science and has bec~~a~~ome even more relevant in statistics and data science, as the availability of new computational tools has bec~~a~~ome common to accommodate data and predict new events. The entire field of science changed ~~a lot~~drastically over the last decades, new disciplines entered ~~in~~ the scientific gotha, and the social sciences became a new frontier~~s~~ where predictive accuracy was ~~strongly~~ required.

Natural and physical sciences progressed ~~by means of~~with Popper's falsificationism ~~philosophy~~, ~~whose one of the~~a main consequence~~s~~ of which is ~~the~~ strong predictivism: scientific theories should be fal~~-~~sified in light of wrong predictions. ~~Th~~Ho~~ugh~~wever, social sciences are not falsifiable in the same way: some social events—~~P~~presidential elections, football results, and polic~~ies~~y effects—are not perfectly predictabl~~e due~~e ~~t~~for many reasons, such as data origins and unpredictable human behavio~~u~~rs. In this paper, we relax the assumptions behind strong instrumentalism, and we provide a bunch of points—see Table 1—to frame statistical and ML~~machine learning~~ techniques within ~~the~~ a weak in~~-~~strumentalist philosophy~~, whose~~ in which the main proposals regard algorithm~~'s~~ transparency and variability in ~~the~~ predictions.

As statisticians ~~demanded~~ required to build good models to accommodate complex data, we ~~feel to~~must warn statisticians~~s~~ and data science users about the role of prediction. Predictive accuracy is not always constitutive of scientific success: ~~p~~Prediction is not everything, but it is vit~~d~~al, and it is our responsibility to choose ~~between~~ the gun or the bazooka.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *In: Petrov, B.N., Csaki, F. (eds.) Proceedings of the Second International Symposium on Information Theory, pp. 267–281. Akademiai Kiado, Budapest (1973). Reprinted in: Breakthroughs in Statistics*, pp. 610–624. Springer, New York (1992).

Baio, G. and M. Blangiardo (2010). Bayesian hierarchical model for the prediction of football results. *Journal of Applied Statistics 37*(2), 253–264.

Breiman, L. (1996). Bagging predictors. *Machine learning 24*(2), 123–140.

Breiman, L. et al. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science 16*(3), 199–231.

Commented [.69]: Edied for word choice/tone.

Commented [.70]: This word shows no Google or Google Scholar results, save for a town in Germany. Please check.

Commented [.71]: Edited for redundancy.

Alternate suggestion: *necessary*

Commented [.72]: Edited for structure/clarity/flow.

Commented [.73]: Edited per AmE.

Commented [.74]: Suggestion:
several
numerous

Commented [.75]: Suggestion: *(see Table 1)*

Commented [.76]: Edited for structure/flow.

Commented [.77]: Edited for word choice/clarity.

Commented [.78]: Edited for tone.

Commented [.79]: Edited for structure/flow/tone.

Commented [.80]: Your references were not edited, per your request.

Breiman, L., J. Friedman, R. Olshen, and C. Stone (1984). *Classification and regression trees*. Wadsworth.

Bzdok, D., N. Altman, and M. Krzywinski (2018). Points of significance: statistics versus machine learning.

Freund, Y., R. E. Schapire, et al. (1996). Experiments with a new boosting algorithm. In *icml*, Volume 96, pp. 148–156. Citeseer.

Friedman, M. (1953). *Essays in positive economics*. University of Chicago Press.

Gelman, A. (2016a). Election surprise, and three ways of thinking about probability.

Gelman, A. (2016b). Explanations for that shocking 2% shift.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin (2013). *Bayesian data analysis*. Chapman and Hall/CRC.

Gelman, A. and C. Hennig (2017). Beyond subjective and objective in statistics. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 180*(4), 967–1033.

Gelman, A., J. Hwang, and A. Vehtari (2014). Understanding predictive information criteria for bayesian models. *Statistics and computing 24*(6), 997–1016.

Gelman, A. and C. R. Shalizi (2013). Philosophy and the practice of bayesian statistics. *British Journal of Mathematical and Statistical Psychology 66*(1), 8–38.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

Hitchcock, C. and E. Sober (2004). Prediction versus accommodation and the risk of overfitting. *The British journal for the philosophy of science 55*(1), 1–34.

Ho, T. K. (1995). Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, Volume 1, pp. 278–282. IEEE.

James, G., D. Witten, T. Hastie, and R. Tibshirani (2013). *An introduction to statistical learning*, Volume 112. Springer.

Karlis, D. and I. Ntzoufras (2003). Analysis of sports data by using bivariate poisson models. *Journal of the Royal Statistical Society: Series D (The Statistician) 52*(3), 381–393.

Kuhn, M. (2019). *caret: Classification and Regression Training*. R package version 6.0-84.

Kuhn, T. S. (1962). The structure of scientific revolutions. *Chicago and London*.

Lakatos, I. (1976). Falsification and the methodology of scientific research programmes. In *Can theories be refuted?*, pp. 205–259. Springer.

Maher, P. (1988). Prediction, accommodation, and the logic of discovery. In *PSA: Proceedings of the Biennial meeting of the philosophy of science association*, Volume 1988, pp. 273–285. Philosophy of Science Association.

Commented [.81]: Edit: *Bayesian*

Commented [.82]: Edit: *Bayesian*

Commented [.83]: Edit: *Poisson*

Commented [.84]: Lowercase

Popper, K. (1934). *The logic of scientific discovery*. Routledge.

Popper, K. (1944). The poverty of historicism, ii. a criticism of historicist methods. *Economica 11*(43), 119–137.

Popper, K. (1945). The poverty of historicism, iii. *Economica 12*(46), 69–89.

Russell, B. (1931). *The scientific outlook*. Routledge.

Sarewitz, D. and R. Pielke Jr (1999). Prediction in science and policy. *Technology in Society 21*(2), 121–133.

Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. Van Der Linde (2002). Bayesian measures of model complexity and fit. *Journal of the royal statistical society: Series b (statistical methodology) 64*(4), 583–639.

Stan Development Team (2018). RStan: the R interface to Stan. R package version 2.18.2.

Vehtari, A., J. Gabry, Y. Yao, and A. Gelman (2019). loo: Efficient leave-one-out cross-validation and waic for bayesian models. R package version 2.1.0.

Vehtari, A., A. Gelman, and J. Gabry (2017). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and computing 27*(5), 1413–1432.

Watanabe, S. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research 11*(Dec), 3571–3594.

Worrall, J. (2014). Prediction and accommodation revisited. *Studies in History and Philosophy of Science Part A 45*, 54–61.