# Prediction is not everything, but everything is prediction

Leonardo Egidi
*Dipartimento di Scienze Economiche, Aziendali, Matematiche e Statistiche 'Bruno de Finetti', Università 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**.

## 1. Introduction

## 2. Prediction for science or science for prediction?

### *2.1. It is prediction part of the science design?*

Bertrand Russell's scheme, Cartesio

### *2.2. Prediction as a confirmation theory approach*

Popper, Kuhn, Mayo

For Popper (Popper, 2005), a theory is scientific only if is falsifiable, where the falsification of a theory is meant to be the the possibility to compare its predictions with the observed data. In his view, theories whose predictions conflict with any observed evidence must be rejected: prediction 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 natural sciences, such as physics. In general, as Hitchcock and Sober (2004) argue, mathematical descriptions of the invariant behaviour 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 continuing to be discovered. Some prominent chemists attempted to determine their atomic weights, density 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 recurrence. He found some gaps in the pattern, and he argued that these missing values corresponded to some existing elements which had not yet been discovered: he named three of these elements eka-aluminium, eka-boron, and eka-silicon, by giving some detailed description 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 which corresponded to descriptions of eka-aluminium, eka-boron, and exa-silicon, respectively: these three elements are better known now as gallium, scandium and germanium. The predictive ability of Mendeleev was remarkable—the Royal Society awarded him the Davy Medal in 1882—, and the new discovered elements well represent pieces of evidence which confirmed the theory.

Predictive confirmation is still ambiguous in social sciences. As argued by Popper, previsione and Sarewitz and Pielke Jr (1999), social sciences have long tried to emulate physical sciences in developing invariant mathematical laws of human behaviour 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 motivations to validate the underlying theory. In the 2016 United States Presidential Election the Republican Donald Trump defeated the Democrat Hillary Clinton by winning the Electoral College (304 vs 227), but gaining lower voters' 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 youre 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, included Nate Silver, the founder of the well-known FiveThirtyEight blog (`https://fivethirtyeight.com`), failed to predict the Trumps' win. However, it is naive to conclude that those models failed because their underlying mechanism was wrong; rather, the political science predictions cannot entirely act as theory's confirmation tools, due to many reasons attributed, for instance, to nonresponse and voters' turnout, as explained by Gelman (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: its 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 statistics

Statistics has always been thought as the *science of inference*, or *science of estimates*, and inference is always seen as separate from prediction. Inference is based on an underlying mathematical model for the data-generating process (Bzdok et al., 2018), its main task is to describe an unknown mechanism working through generalization: the inferential laws should in fact be as broad 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 moves from the observed to the unobserved, being the action designed to forecast future events without requiring a full understanding of the data- generation process. Each person is more or less confident with the weather's predictions or with presidential election predictions, but rarely that person is aware of the underlying statistical model required to produce that forecast, unless he is a statistician/data scientist. In such a view, inference seems hard and obscure, and prediction easy and close to the people. This is often a paradoxical argument, since the inference is often associated to the *explanation* of the problem, and should be relevant and available to the majority of the population

As statisticians, we are often faced with a double task. First, we must create a sound mathematical model to accomodate the data and retrieve useful inferences for our parameters—there is not distinction here between classical and Bayesian statistics, they are both designed to draw inferential conclusions, either in form of point estimates/ confidence intervals or in terms of posterior quantiles/credibility intervals. Then, we should use this model to make predictions, but this is rarely accounted by the statisticians in a transparent way.

For illustration purposes only, we consider the classical linear regression model, where $y_n$ denotes the response variable, $X$ is the $n \times p$ predictor matrix, and $\alpha, \beta_1, \beta_2, \ldots, \beta_p$ are the $p+1$ parameters—the intercept $\alpha$ and the $p$ regression parameters, we have

$$y_n = \alpha + \sum_{k=1}^{p} \beta_j x_{nk} + \varepsilon_n,\ n = 1, 2, \ldots, N, \tag{1}$$

with $\varepsilon_n \sim \mathcal{N}(0, \sigma_\varepsilon^2)$. Once the model has been estimated and we have retrieved some parameters' estimates $\hat{\alpha}, \hat{\beta}_1, \beta_2, \ldots, \hat{\beta}_p$, we could use them to provide a point forecast $\tilde{y}_{n+1}$ for the unobserved $y_{n+1}$ associated to the predictor vector $\tilde{x}_{n+1k}$. To account for uncertainty, rather than using a point forecast, we could also compute the prediction interval for $\tilde{y}_{n+1}$.

Once the value $y_{n+1}$ is known, the statistician can validate his prediction and check its plausibility. This *ex post* validation is not available when fitting the model, and as such should not represent the plausibility of the model fit itself, since the model construction did not require $y_{n+1}$. In this misalignment between the fitting of the model without the $n+1$-th unit and the forecast validation of $y_{n+1}$ there is space for an infinite debate about the scientific role of prediction.

### *3.1. Overfitting and data accomodation*

Even a simple linear regression case poses many challenges: should the statistician use all the $N$ data to build a reasonable/useful model, or could he take only a portion of the sample to accomodate the model (the training set) and use the remaining values to validate the model (the test set)? This apparently naive question pushed many scholars to debate about the presumed

supremacy of prediction over accomodation (Maher, 1988; Hitchcock and Sober, 2004; Worrall, 2014). According to this competition, the statistician should ask himself whether he wants models that are true—or approximately true—or predictively accurate.

It is well-known that more complex models tend to yield poor predictive results

### *3.2. The falsificationist Bayesianism framework*

Gelman's scheme of Bayesian inference. Draw connection with Bertrand Russell

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 ultimate interest, given the observed data.

(c) Evaluating the fit of the model and the implications of the resulting posterior distribution: 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 1? In response, one can alter or expand the model and repeat the three steps.

In the above paradigm, predictions are never mentioned. But 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

$$p(\tilde{y}|y) = \int p(\tilde{y}|\theta)p(\theta|y)d\theta, \tag{2}$$

where $p(\theta|y)$ is the posterior distribution for $\theta$, whereas $p(\tilde{y}|\theta)$ is the likelihood function for future observable values. In the linear regression case (1), the posterior predictive distribution for the future observation $\tilde{y}_{n+1}$ is given by:

$$p(\tilde{y}_{n+1}|y) = \int \mathcal{N}(\alpha + \sum_{k=1}^{p} \beta_p \tilde{x}_{n+1k}, \sigma_\varepsilon^2) p(\alpha, \beta_1, \beta_2, \ldots, \beta_p|y) d\alpha d\beta_1 d\beta_2 \ldots d\beta_p.$$

Equation (2) may be resambled in the following way:

$$p(\tilde{y}|y) = \frac{p(\tilde{y}, y)}{p(y)} = \frac{1}{p(y)} \int p(\tilde{y}, y, \theta) d\theta. \tag{3}$$

From Equation (3) we immediately notice that whenever we are interesting in predictions, we need to consider a joint model $p(\tilde{y}, y, \theta)$ for both the observed data $y$ and the unobserved quantities $\tilde{y}, \theta$. This joint model incorporates bot the likelihood and the prior, being $p(\tilde{y}, y, \theta) =$

$p(\tilde{y}|\theta)p(y|\theta)p(\theta)$. 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.

### *3.3. Communication duties*

## 4. The role of prediction in machine learning

Breiman, Bzdok, Popper (we could argue that ML procedures are not falsifiable!)

## 5. Going beyond inference and prediction: a tentative unifying approach

## 6. Applied examples

## 7. Conclusions

## References


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

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

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.

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.

Popper, K. (2005). *The logic of scientific discovery*. Routledge.

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

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