Harvard Data Science Review • Issue 1.1, Summer 2019

A Balanced Perspective on Prediction and Inference for Data Science in Industry

Nathan E. Sanders[1]

[1]Communicating Science (ComSciCon), Harvard University, Cambridge, Massachusetts, United States of America

The MIT Press

Published on: Sep 18, 2019

DOI: https://doi.org/10.1162/99608f92.644ef4a4

ABSTRACT

The strategic role of data science teams in industry is fundamentally to help businesses to make smarter decisions. This includes decisions on minuscule scales, such as what fraction of a cent to bid on an ad placement displayed in a web browser, whose importance is only manifest when scaled by orders of magnitude through machine automation. But it also extends to singular, monumental decisions made by businesses, such as how to position a new entrant within a competitive market. In both regimes, the potential impact of data science is only realized when both humans and machine actors are learning from data and when data scientists communicate effectively to decision makers throughout the business. I examine this dynamic through the instructive lens of the duality between inference and prediction. I define these concepts, which have varied use across many fields, in practical terms for the industrial data scientist. Through a series of descriptions, illustrations, contrasting concepts, and examples from the entertainment industry (box office prediction and advertising attribution), I offer perspectives on how the concepts of inference and prediction manifest in the business setting. From a balanced perspective, prediction and inference are integral components of the process by which models are compared to data. However, through a textual analysis of research abstracts from the literature, I demonstrate that an imbalanced, prediction-oriented perspective prevails in industry and has likewise become increasingly dominant among quantitative academic disciplines. I argue that, despite these trends, data scientists in industry must not overlook the valuable, generalizable insights that can be extracted through statistical inference. I conclude by exploring the implications of this strategic choice for how data science teams are integrated in businesses.

Keywords: industry, entertainment, communication, inference, bibliometrics

1. Introduction

The explosive uptake of data science in industry can be attributed to the enormous innovation enabled by pooling developments in quantitative methods across disparate disciplines, and the additional potential emergent at their intersection. Interdisciplinary research in all fields unlocks tantalizing possibilities, but data science is unique in that it brings together two domains–statistics and computation–that are integral to essentially all fields of science, engineering, the digital humanities, and related fields in academia as well as industry (Blei & Smyth, 2017). For many practitioners, the excitement of reading new work or participating in conferences in data science is driven by the opportunity to encounter a diversity of ideas; to learn from the hard-won example of methods that have incubated within varied fields.

But as marvelous as advancements in machine learning and other data science methodologies and technologies may be, they do not create value for business on their own. Value is created when people have the insight to apply these techniques to new problems, to extend their capabilities beyond what was originally contemplated, or to use them as tools that support people in making good decisions and taking appropriate actions.

In "An Executive's Guide to Machine Learning," Pyle and San José defined three stages to the application of machine learning, data science, and artificial intelligence in the business world. They call these stages "description," "prediction," and "prescription" (2015). This framework has been adopted widely in the business community.[1] They branded the "description stage" as "Machine Learning 1.0," the collection of data in databases to facilitate online processing and question answering. They defined the "prediction" stage, which they denoted as the current state of the art, to mean using models to predict future outcomes. Reflecting the present "urgency" they associated with businesses' adoption of this capability, they used the term "prediction" or related conjugations 10 times in their nine-page article.

It is understandable that a contemporary observer would form the perspective that prediction has been the principle preoccupation of data science. For example, the popular online platform Kaggle[2] has engaged hundreds of thousands of users, some veterans and some first-time modelers, to participate in data science competitions since 2010. Kaggle has become a highly influential and constructive entry point into the practice of data science and experience on the platform is frequently cited by job seekers and recruiters as a key way to build credentials for the data science job market.[3] Kaggle always frames its competitions as prediction challenges: the purpose of the actions of data scientists in Kaggle competitions is defined to be the improvement of predictive performance metrics. There is rich discussion on the platform of how users can improve the scores of their models, but relatively little discussion of what can be learned about the systems they are modeling from their models' development and application.

Finally, Pyle and San José anticipate a third stage, "prescription," that involves human learning from and interpretation of models to explain why outcomes occur the way they do, which they present as the aspirational future of machine learning. But to scientists, "prescription" is a highly recognizable modality that may be broadly translated to the statistical term "inference." Referring to inference explicitly, Pyle and San José urged practitioners to move beyond "classical statistical techniques [that] were developed between the 18th and early 20th centuries for much smaller data sets than the ones we now have at our disposal" (p. 47). They could have looked back even farther in time: it is not an exaggeration to say that the origins of this kind of "prescriptive" rational reasoning from data facilitated by conceptual and mathematical models can be traced across 4,000 years of the history of science (Franklin, 2015).

Applications of inferential reasoning to modern technologies and problems already motivates much of modern science. To name a few examples: generations of advancement in causal inference enables measurement of the causal effects of salient interventions from uncontrollable observational datasets (Imbens & Rubin, 2015; Pearl, 2014), new algorithms for Bayesian inference enable expectations to be computed over high dimensional models that capture the behavior of complex probabilistic systems (e.g., Betancourt, Byrne, Livingstone, & Girolami, 2017), and the field of interpretable machine learning has generated elegant mechanisms to explain so-called “black box” models in comprehensible terms (e.g., Doshi-Velez & Kim, 2017; Guidotti et al., 2018). All these methods are already in use across virtually every field of science in one form or another. In agreement with Pyle and San José, it is certainly valuable for business executives to move beyond the strategic goal of prediction and to recognize the opportunity for data science to enhance our understanding of data and the systems that generate it.

In this article, I offer an accessible introduction to the duality between inference and prediction in data science intended for practitioners in industry (§2). Using evidence from a textual analysis of research abstracts from technical preprints, I show that there is a growing imbalance such that prediction is increasingly dominant in the marketplace of ideas for data science (§3) and draw connections to the circumstance in industry described above. I then illustrate the mutual dependence and complimentary importance of inference and prediction using examples from the entertainment industry, focusing on the box office projection and advertising attribution tasks (§4). Finally, I examine some implications of these trends for how organizations conceive of and communicate about data science (§5).

2. The Duality of Inference and Prediction

The terms inference and prediction are used widely and not entirely consistently across the connected domains of data science, from theoretical statistics to computer science to medicine to entertainment and beyond, and in everyday parlance. These variations in conceptualization, terminology, and even mathematical notation make it challenging to communicate clearly about high level concepts to an audience as diverse as industrial data scientists. I will attempt to tackle this challenge here by appealing to descriptions, examples, illustrations, and clarifying contrasts I have found useful in discussions with colleagues.

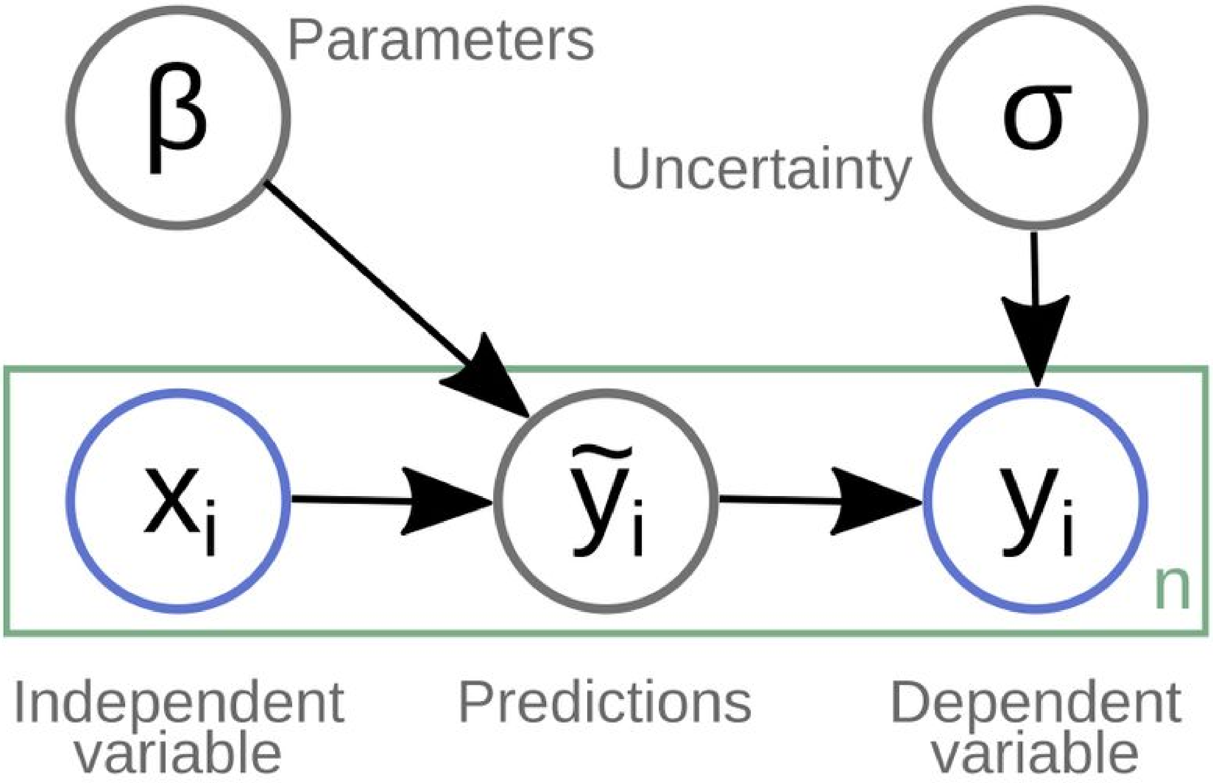

Figure 1. A graphical diagram (see, e.g., Lauritzen, 1996) of a simple supervised machine learning model. The observed variables are outlined in blue and unobserved variables of the model in gray; the green plate represents the dimensionality of the data, n.

2.1. Definitions and perspectives

I define the terms inference and prediction in practical terms as follows:

- **Predictions:** The outputs emitted by a model of a data generating process in response to a specific configuration of inputs.

- **Inferences**: The information learned about the data generating process through the systematic comparison of predictions from the model to observed data from the data generating process.

To elaborate, consider the straightforward case of a linear regression model. With respect to the concepts of inference and prediction, this example is generally representative of predictive model-based inference and supervised machine learning. For a more general, non model-dependent discussion of inference, see e.g. Betancourt (2015).

Figure 1 shows the model’s essential components, enumerated as follows:

- A set of independent variables, x_i, that are observed and provided to the model as input data or “predictors.”

- A dependent variable, y_i, that is also observed and provided to the model as training examples of output data.
- Predictions, $\tilde{y}_i$, synthetic output data that are generated by the model and intended to match y_i as well as possible.
- A set of inferred parameters, β, that serve to transform the inputs into the predictions.
- An uncertainty measure, σ, that characterizes the magnitude of typical errors in predictions.

The variables x_i and y_i are "observables," while β and σ are model parameters representing features of the model that cannot be directly observed. The prediction $\tilde{y}_i$ represents the best fit of the model to the observed variable y_i.

A sample application from the entertainment industry, which will be detailed in §4.1, is the modeling of box office outcomes for new theatrical film releases. In box office projection models, the independent variables are typically characteristics of films such as their cast composition and genre classification, among many others. The dependent variable of interest may be the total box office gross of a film. In the linear regression case, the β parameters directly represent the effects of each independent variable on the dependent variable, such as the additional dollars in gross attributed to the selection of a genre favored by audiences. The uncertainty measure indicates the amount of variance expected between the revenue predicted by the model and its actual outcome.

Both inference and prediction are truly integral to the functioning of the model and both have effects that cannot be ignored in the practical application of the model. The parameters can only be learned by comparison of the predictions to dependent variable observations through the model training process. If the model's predictions do not match reasonably well the observed outcomes of the data generating process, inferences about that process will be unreliable. Likewise, the predictions themselves are generated directly from the combination of the parameters and the independent variables. If the operation of the model through its parameters cannot be explained and understood, there will be little basis to build confidence in the model's predictive ability for future realities or in new domains.

However, I have observed that many analysts and organizations choose to invest their time and attention primarily on one function or the other, and perceive of the relative importance of inference and prediction as imbalanced. In particular, as discussed in §1, a notion that prevails in many circles of industry is that prediction is the primary concern of data science. Figure 2 contrasts the two perspectives on the example model. Panel a) illustrates the "balanced perspective," where the parameter inferences and predictive outputs are viewed with equal interest. Panel b) highlights the "prediction-oriented" perspective, where the predictive outputs of the model carry outsized interest compared to the other essential elements of the model.

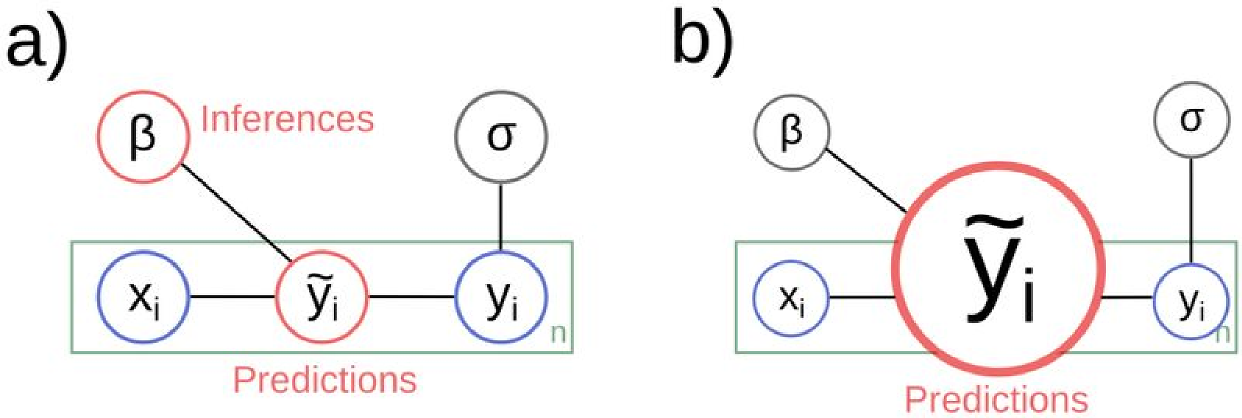

Figure 2. Illustration of perspectives on the sample model of Figure 1: a) a balanced perspective equally valuing inference and prediction and b) a prediction-oriented viewpoint.

Figures 1 and 2 serve to illustrate that prediction and inference are two distinct goals of the modeling process which both offer value to organizations and are inextricably connected to each other in the modeling process, but can be viewed in different ways. Both perspectives are valid in different contexts and it is important for analysts and organizations to consider and recognize the appropriate orientation for a particular data science project.

In some operational applications in industry, the predictive outputs of a model will be integrated directly into an automated system and the values of the inferred parameters and other model behaviors will never be inspected; this exemplifies the prediction-oriented perspective. For example, the developer of an online streaming platform implementing a collaborative filtering algorithm (see e.g., Ricci, Rokach, & Shapira, 2015) may deploy the predictive outputs of their model to provide recommendations. The recommendations are obtained by fitting users' time spent viewing video on the platform, perhaps without concern for the parameters of this model or the drivers of the users' behavior. In this "black box" modeling regime, the parameters are simply a means to an end; nuisances that can be entrusted to a well engineered automated learning framework and overlooked thereafter. (That said, there are numerous inferential insights that can be extracted about consumer content preferences, and about the content itself, from collaborative filtering algorithms, e.g. Tintarev and Masthoff, 2015.)

Guidotti et al. (2018) provide a useful criteria for when truly black box predictive models are appropriate and, therefore, when inference, explanation, or interpretability are unnecessary: "an explanation could be not required if there are no decisions that have to be made on the outcome of the prediction" (p. 93:5). Of course, in many contexts in science and industry broadly, making decisions on the basis of data is the primary underlying motivation for applying data science.

At the other extreme, the predictive outputs of a model may be used solely as a means for model fitting in order to produce inferences and may be scarcely commented on; an "inference-oriented perspective." For example, an astronomer may make careful measurements of the brightness of a supernova explosion for the purpose of inferring the physical parameters of a progenitor star through the comparison of the brightness observations to models motivated by astrophysical theory. In this context, future predictions of the observables are uninteresting in and of themselves. Identifying the physical parameters of the stars is the goal of the study, though these parameters cannot be measured directly; even if it were possible to place a star on a balance to measure its mass, the observation of the supernova itself follows the conflagration of the star. The brightness measurements are a means to an end; incidental observables that serve the purpose of constraining the values of physical stellar parameters through model training validated by predictive performance against those observables. (This example particularly informs the perspective I bring to the industrial context, e.g., Sanders, Betancourt, & Soderberg, 2015; Sanders, Soderberg, et al. 2015)

As I will explore further below, there is evidence that the prediction-oriented perspective is increasingly dominant in many fields and I argue that there would be benefit to more frequent use of the balanced viewpoint.

2.2. Conceptual parallels

The duality between inference and prediction as defined in this section parallels, but is distinct from, other well known conceptual dualities that confront data scientists. Here I examine a few related concepts to clarify the distinctions between them.

First, Breiman (2001) identified a conflict of “culture” in statistical modeling, identifying a “data modeling culture” that operates on the assumption that there exists a parameterizable model that can explain the data generating process and an “algorithmic modeling culture” that assumes that “Nature forms the outputs y from the inputs x by means of a black box with complex and unknown interior” (p. 209). Breiman asserted that 98% of all statisticians at the time of his writing belonged to the data modeling culture, while the algorithmic modeling culture was already dominating in other fields. He advocated for the use of algorithmic models by exploring the Occam dilemma: “Accuracy generally requires more complex prediction methods.” Breiman’s debate between modeling cultures, or model types, is not the duality I examine here. Instead, the duality explored in this section corresponds to Breiman’s two “goals” for analyzing data (in his § 1) of “prediction” (similar to my definition above) and “information” (similar to my definition for inference above), rather than the two “approaches” of data and algorithmic modeling. Both goals can be pursued via either approach. The simultaneous pursuit of Breiman’s two goals would be analogous to the balanced perspective advocated in this article. Furthermore, information extraction (Breiman’s term) or inference (mine) from a model need not be confined to parameter estimation, as in the example above. Other methods for analysts to extract information from and interpret the modeling process are discussed in §5.1 and elsewhere in this article.

Second, I distinguish my definition of inference from the narrow domain of frequentist hypothesis testing. In some domains, particularly psychology, statistical inference has historically been synonymous with hypothesis testing (see e.g. Killeen 2005, Krueger 2001, Schmidt 1996). In my formulation, hypothesis testing would be one approach among a broad class of methodologies for learning about the data generating process that also includes Bayesian methods, techniques for interpreting deep learning models, and others discussed elsewhere in this article. My definition of "inference" is more similar to the concept of "scientific inference" discussed by e.g. Hubbard, Haig, and Parsa (2019): "discovery of replicable and empirically generalizable findings" (p. 91). In an exploration of the purpose of hypothesis testing, *p*-values, and significance levels, Billheimer (2019) advocate for "Predictive inference" summarized as follows: "Rather than infer the value of a parameter that can never be observed, our inferential focus should be the prediction of future observable quantities" (p. 291). Billheimer's recommendation, building on work by de Finetti (1937), Geisser (1993), and others, is that testable predictions of future observable values should be the currency for evaluating the performance of a model and identifying the reliability of inferences about parameters. This is compatible with the balanced perspective advocated in this section and similar to the concept of "correspondence to observable reality" articulated as a virtue of statistical practice by Gelman and Hennig (2017).

Next, I consider the familiar distinction between correlation and causation. The rich literature on causal inference (see e.g. Imbens and Rubin, 2015; Pearl, 2014; Wang & Blei, 2018) carefully defines the meaning of the causal effect of an intervention assigned to a unit and establishes multiple frameworks and a variety of empirical methods for measuring causal effects from observational and experimental data. In both business and research settings, constraints on the ability to control assignment mechanisms and other system factors often limit the extent to which causal effects can be isolated and measured. As a result, it is often necessary for data scientists to, for example, analyze descriptive correlations within datasets or to model data with known (and unknown) confounding variables that may not be fully observed. For analyses focused on the goal of prediction, data scientists must recognize how these limitations affect the generalizability of their models. A predictive model that learns a correlation between a particular predictor and a dependent variable of interest may perform poorly on out of sample cases where an unobserved confounding predictor or an additional cause has changed. For analysis focused on the goal of inference, it is critical to understand the limitations of a dataset or study design for identifying causation to avoid over-interpreting inferences.

Finally, inferences under any particular model (however simple or complex) are subject to the assumption that the model accurately describes the data generating process. Amrhein, Trafimow, and Greenland (2019) suggests treating inferential statistics as "unstable local descriptions of relations between models and the obtained data" (p. 262). Analysts should fit a variety of models, systematically compare their performance, and generalize when possible using continuous model expansion to help mitigate the effects of this localization (Betancourt, 2015; Draper, 1995; Lavine, 2019). Ultimately these considerations provide an explication of how inference is a useful procedure for analysts and organizations to learn from data. The iterative process of designing models, applying them to data, checking their predictive performance, and interpreting the models' parameters and behavior all promotes understanding of the data generating system being modeled; see e.g. Gelman et al. (2013) and Betancourt (2018) for Bayesian formulations of a principled process for model development.

3. The Rise of Prediction in the Literature

As indicated in §1, the imbalanced prediction-oriented perspective on data science has become dominant in forums ranging from the executive suites in many industries to the online community of data science learners and practitioners. This aligns with trends in the academic research community, who themselves have increasingly focused on the predictive aspect of modeling.

In this section, I demonstrate this trend by textual analysis of academic preprints. While the linguistic expression of the broad concepts of inference and prediction explored in §2 eludes strict and comprehensive quantification, a simple analysis of word frequency signals the relative rates at which researchers deploy the inferential and predictive frames.

Table 1. Synonymous terms for salient root words used in Figures 3–4.

Root Word	Synonyms
Predict	predict, predictability, predicted, predicting, prediction, predictions, predictive, predictor, predictors, predicts
Infer	infer, inference, inferences, inferencing, inferential, inferred, inferring, infers

To supply data for this analysis, I construct a textual database of research abstracts queried from the scholarly pre-print server the arXiv[4] using their API.[5] The arXiv is divided into domain specific categories and each category is further divided into subcategories such as astro-ph.GA (“Astrophysics of Galaxies”) and cs.NE (“Computer Science: Neural and Evolutionary Computing”). The date of first publication and rate of publication varies greatly across subcategories, with some having received thousands of submissions per year for decades and others being effectively quiescent. I gathered all abstracts published from 2005-01-01 to 2018-12-31 across the 141 subcategories of the following arXiv categories, totaling more than 1.3 million abstracts: astro-ph (Astrophysics), cond-mat (Condensed Matter), cs (Computer Science), econ (Economics), eess (Electrical Engineering and Systems Science), math (Mathematics), nlin (Nonlinear Sciences), physics (Physics), q-bio (Quantitative Biology), q-fin (Quantitative Finance), and stat (Statistics). Not included are several additional categories in the physics domain such as nucl-th (Nuclear Theory) and quant-ph (Quantum Physics), which are anticipated to display similar trends as the primary physics category.

With this data, we can examine the relative frequency and time evolution of the use of “infer,” “predict,” and related terms in arXiv categories and sub- categories over time. I adopt the following notation: The total count of abstracts from a subcategory, s, in a given year, y, is denoted $N_{s,y}$. The concepts, c, of inference and prediction are identified by the words “infer,” “predict,” and other semantically synonymous terms listed in Table 1. The presence of these terms is counted within abstracts from each subcategory and aggregated by year as $\mathtt{N}(\mathtt{c})_{\mathrm{s,y}}$. If abstracts include words from both the “infer” and “predict” lists of Table 1, they are counted in both categories. The growth of abstract submissions to a subcategory between years a and b is calculated simply as $\mathtt{G}(\mathtt{N})_{\mathrm{s,a,b}} = \mathtt{N}_{\mathrm{s,a}}/\mathtt{N}_{\mathrm{s,b}}$. The rate of use of a term α in a subcategory in a given year a is expressed as $\mathtt{U}_{\mathrm{s,\alpha,a}} = \mathtt{N}(\alpha)_{\mathrm{s,a}}/\mathtt{N}_{\mathrm{s,a}}$. The growth rate of the usage of the same term between years a and b in a subcategory is calculated as $\mathtt{G}(\mathtt{U})_{\mathrm{s,\alpha,a,b}} = \mathtt{U}_{\mathrm{s,\alpha,a}}/\mathtt{U}_{\mathrm{s,\alpha,b}}$. The relative usage growth rate between two terms α and β between years a and b in a subcategory is calculated as $\mathrm{R}_{\mathrm{s,\alpha,\beta,a,b}} = \mathtt{G}(\mathtt{U})_{\mathrm{s,\alpha,a,b}}/\mathtt{G}(\mathtt{U})_{\mathrm{s,\beta,a,b}}$.

Although this word frequency analysis is a blunt measurement instrument, its results readily suggest some interesting changes over time.

Figure 3 illustrates these results for some popular arXiv subcategories. The fast growing stat.ML (“Statistics: Machine Learning”) subcategory, which has grown by $\sim 8\times$ in annual submissions since 2012, is illustrative of the topical evolution of many subcategories. In stat.ML, inference dominated the discussion during the early period of the subcategory in the late 2000’s. But the subcategory has trended quickly towards prediction over the past half-decade, with the use of prediction synonyms rising by ~25% since 2015 and inference synonyms falling by ~35% over the same period. Similarly, the cs.CV (“Computer Science: Computer Vision and Pattern Recognition”) subcategory has grown by $> 10\times$ since 2012. Prior to that time, the use of inferential terms was somewhat more common than prediction terms. Since that time, the use of both terminologies has grown in this field. But the use of predictive terms has grown at $\sim 2.5\times$ the rate of the growth of inferential terms and now dominates over them by more than $\sim 2\times$.

One of the more volatile cases is cs.AI (“Computer Science: Artificial Intelligence”). Before 2010, usage of predictive terms was stable and inferential terms increasing. Between 2010 and 2013, this field saw a dramatic rise in inference-related terms ($> 2\times$ over 2 years) while predictive term usage continued at similar rates. Since 2013, prediction terms have grown by $> 2\times$ while inferential terms have fallen by half. The abstract submission rate has roughly tripled over this time period.

A final interesting direct contrast are the subcategories stat.AP (“Statistics: Applications”) and stat.TH (“Statistics Theory”). stat.TH had even usage of inferential and predictive language in 2012. While the fraction of articles mentioning prediction has stayed constant since that time and total article submissions have risen a modest ~60%, the discussion of inference has grown ~50%, a trend noticeably contrary to the fast-growing subcategories discussed previously. Much the opposite, stat.AP has seen a ~70% growth in prediction discussion dating back to 2012, with a flat long term trend in the use of “infer” and synonyms. These figures suggest that discussion of statistical inference is increasingly concentrated in certain subcategories like stat.TH and moving away from subcategories where it used to be more prevalent, like stat.ML, cs.CV, cs.AI, and stat.AP.

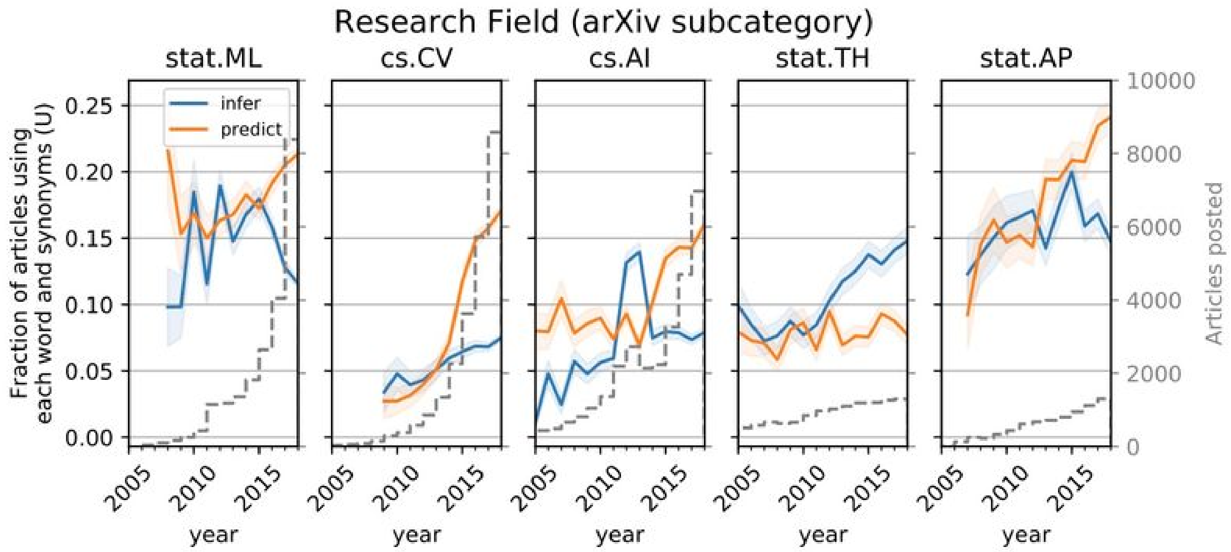

Figure 3. Average rate of usage (U) of "infer," "predict," and synonymous terms (as defined in Table 1) within abstracts of different arXiv subcategories over time (colored lines) and standard error (colored regions). The total volume of abstracts (N) in each subcategory is also shown (gray dashed line).

The faster growing research disciplines tend to be the same ones that have increasingly focused on prediction in recent years. Figure 4 illustrates this finding at the subcategory level. Each timepoint is calculated using a rolling boxcar mean with a 3-year trailing window, i.e. an estimate of the rate at each time point from the average of sequential aggregations of the prior 3-year period. Subcategory data points with fewer than 100 articles in a year or extremely low (< 1%) usage of either term set are excluded. Evaluated as a simple correlation at the subcategory level weighted by 2018 post volume, the trend between article submission growth and the relative usage growth for "prediction" over "inference" is fairly strong (weighted correlation coefficient $\rho = 0.59$).

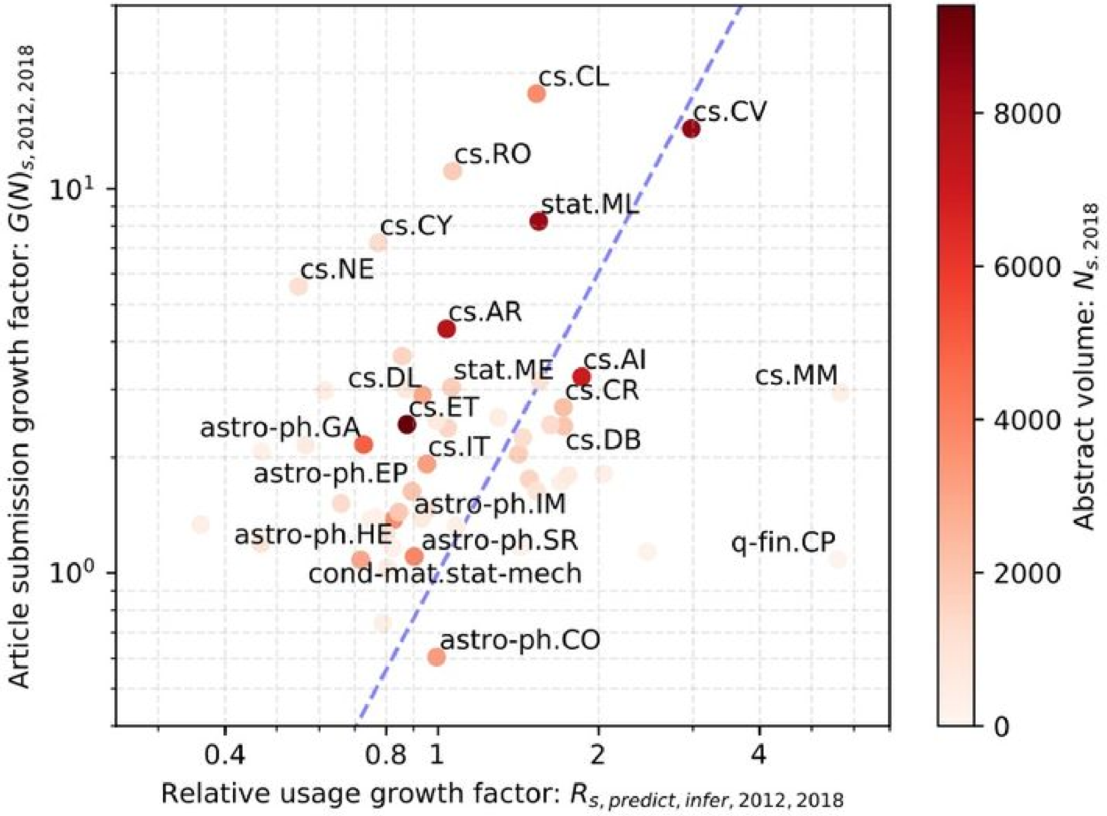

Figure 4. Comparison between article submission growth and the growth in the relative usage of inference and prediction terms for arXiv subcategories. The growth is calculated over the period 2012 to 2018 using a rolling boxcar mean. The shading denotes the total submission volume as of 2018. The blue line shows a least squares linear trend weighted by 2018 post volume. For clarity, only the largest (by 2018 post volume) third of the points and those with extreme growth rates are labeled.

The relation between article submission growth, $G(N)$, and increasing predictive focus, R, is also evident at the category level (Table 2). Among the categories studied here, the three fastest expanding categories are also the three with the highest trend towards predictive language usage: cs, stat and q-fin. The stable-volume astro-ph, nlin, and cond-matter categories, in contrast, are trending towards more inferential language usage.

These results show that scholarly articles in the fastest growing quantitative research fields, as proxied by abstracts posted to the arXiv, have increasingly focused on prediction since about 2012. It is reasonable to conclude that this shift in focus has contributed to the prevailing perspective in industry discussed in §1, where data science emerged as a prominent area of investment over roughly the same time period. Due to the inter-dependent nature of techniques developed in one domain and studied or applied in the other, the rise of prediction-focused literature in the academic literature may also have been driven in part by demands originating in industrial practice.

Table 2. Summary of the arXiv growth rate statistics displayed in Figure 4 aggregated by category. The growth is calculated over the period 2012 to 2018 using a rolling boxcar mean; see the text for details and definition of terms.

	Relative ‘predict’ usage growth: R	Article submission growth: G(N)	2018 article volume: N (1000’s)
nlin	0.45	1.01	1.81
astro-ph	0.83	1.15	19.97
cond-mat	0.96	1.18	23.57
math	0.98	1.34	44.69
q-bio	1.22	1.43	3.02
physics	0.98	1.61	19.97
q-fin	1.58	1.66	1.30
cs	1.30	2.98	71.19
stat	1.46	3.52	13.43

4. Applications from Entertainment

In my application domain, entertainment, I have found the duality of prediction and inference to be a useful consideration when developing strategy for a variety of different business challenges. Our group develops and deploys methods to model, understand, and influence consumer behavior and market systems using techniques including natural language processing (Gundogdu, Sanghvi, & Harrigian, 2018; Harrigian, 2018; Ning, Qu, Cai, & Sanders, 2018), Bayesian inference (Lei, Sanders, & Dawson, 2017; Sanders & Lei, 2018), image recognition, multi-modal deep learning, matrix factorization, and more. Below, I will examine the problems of box office projection and advertising attribution as instructive examples of this duality.

4.1. Box Office Projection

The task of “box office projection” is to model the consumer market that generates revenue via ticket sales for the theatrical exhibitions of a film in one or more territories or worldwide. The most common approach to the task is to construct averages over the historical revenue performance of comparable films identified heuristically based on similarity of film content or production metadata. Model based (regression) approaches are also frequently applied, with independent variables including production characteristics (such as the production budget of a film), talent characteristics (such as the “starpower” of an actor or director as measured from past box office gross or awards), the marketing support behind a film (such as the advertising expenditure and features describing the ad campaign strategy), measures of audience response (such as digital trailer views or volume of social media conversation), and more. For at least the past three decades, a wealth of literature on this task has been produced by the academic community (see e.g. Asur & Huberman, 2010; Basuroy, Chatterjee, & Ravid, 2003; De Vany & Walls, 1999; Elberse, 2007; Lee, Park, Kim, & Choi, 2018; Wang & Blei, 2018), and many industry groups, including film producers and distributors as well as independent vendors, have invested in proprietary data collection and models for this task.

Consider how the perspectives of §2 apply to this task. From the predictive perspective, the goal of box office projection is to predict the revenue generated by the theatrical release. This has value to help studios anticipate the financial outcome of a film, model the expected financial risk and return of their release portfolio, or analyze the strength of their expected competition on a release weekend. From the inferential perspective, the goal of box office projection is to understand the structure and dynamics of the theatrical market. This enables studios to articulate the properties of their film and the marketplace that generates risk for a release and to reason about how to alter production, marketing, and other factors under their control to optimize the return from each product.

Both of these sets of outcomes are of significant interest to studios. One modeling perspective’s set of outcomes is not inherently better than the other, but they are different from each other. Yet the predictive orientation has been most prominent in public interest and discussion.[6] Near theatrical release (within a few weeks of a film’s debut), predictions of box office models are routinely reported by the industry press.[7] In this near-release regime, progress has been made in engineering and integrating digital signals from social and search platforms (Liu, Ding, Chen, Chen, & Guo, 2016; Panaligan & Chen, 2013; Shruti, Roy, & Zeng, 2014). Moreover, online prediction market communities offer non-model based mechanisms for anticipating performance (Elberse & Anand, 2006; Karniouchina, 2011; Pennock, Lawrence, Giles, & Nielsen, 2001). Despite these advancements, variance in box office projections near the time of release is notoriously high (see e.g. Walls, 2005). Earlier in the production lifecycle, typically years before the film’s release, is the critical “greenlighting” stage, when a studio decides whether or not to invest in a film concept. The variance of possible outcomes during that stage is much higher still. Fundamental production and marketing variables may not have been set at that point and the future state of the market is much more difficult to foresee. Predictive modeling during greenlighting is therefore less common (but see e.g. Eliashberg, Hui, & Zhang, 2007; Ghiassi, Lio, & Moon, 2015; Lash & Zhao, 2016).

Given all this context, there is much to refer inference as a high leverage goal of box office projection. Inference allows studios to learn generalizable strategies for production that can be relied upon even in regimes where the absolute predictive outputs of the same model have high variance and limited utility for financial applications. Predictive modeling is widespread near theatrical release, but at this stage of the film lifecycle most production decisions have already been executed. The actual predicted dollar value for the gross output by a box office projection model near release is not highly actionable. The most important outcomes from this modeling, from the studio perspective, is the opportunity to adjust marketing and distribution strategy based on inferences about how predicted gross depends on factors such as audience awareness within different territories and demographics. In the greenlighting phase, predictive precision is highly degraded as described above, but inferences about variation in box office performance by production characteristics such as actor caliber, positioning (the genre framing of the film emphasized to audiences), and sensitivity to audience reception can be highly impactful for product development and release planning. Across all time periods, an understanding of uncertainty–both in the predicted outcome and its relationships with the independent variables–is critical given the high variance inherent to the market and the portfolio management and risk mitigation goals of studios. While it need not be so, analysis of uncertainty is often absent from prediction-oriented modeling approaches for box office projection, as in many of the examples cited above.

4.2. Advertising Attribution

In advertising, the multi-channel attribution modeling task (see e.g. Abhishek, Fader, & Hosanagar, 2012; Berman, 2018; Dalessandro, Perlich, Stitelman, & Provost, 2012; Gupta & Zeithaml, 2006; Yadagiri, Saini, & Sinha, 2015) is to allocate the value of a consumer conversion (a behavior such as a product purchase or website visit) across the individual “impressions” that causally contributed to that outcome. Impressions are defined as advertisement exposures on different channels, such as television and online social media, or “organic” interactions with a brand such as word of mouth. This modeling enables measurement of the effectiveness of each channel, or “platform,” on influencing consumer behavior.

However, rigorous classical causal attribution modeling is not possible in the practical context of most advertising campaigns. It is prevented by incomplete individual-level data on consumer exposure across key online and offline platforms, a lack of consumer conversion data (particularly for offline behaviors), a lack of integration between exposure and conversion datasets when they are available, and an inability to randomize exposure at the individual level. In particular, in the U.S. film industry, the vast majority of tickets are purchased at the brick and mortar box office, and hence not associated with the consumer’s identity by digital tracking; there is little or no ability for studios to capture individual ad exposure logs for many major advertising channels, including broadcast and cable television and online social media. In practice, researchers generally need to accept data that are missing by platform (introducing substantial systematic errors associated with non-attributed platforms), data that are missing by person at random (introducing substantial sampling error depending on the number of observations achieved), and/or data that are missing by person not at random (introducing systematic errors based on demographic, platform usage, or other factors that explain the missingness). It is common, for example, to only apply attribution models to a small subset of available marketing channels where data are more readily available or to a “panel” of consumers that have opted in to more detailed tracking, which may have small sample size and may not be representative of the general population.

Predictions from attribution models for individual consumer behavior, or indeed bulk predictive performance measures for attribution models, should therefore not be taken at face value. They will depend sensitively on the aforementioned systematic sources of error, and hence they may not generalize well to real world scenarios. For example, an attribution model incorporating the effect of web display and television ads may not be a reliable predictor of the actual purchase behavior of a consumer who is also influenced by social media ads, not to mention word of mouth and other organic channels.

Nonetheless, the output of attribution models can provide a critical input to other important models in the marketing domain. Measurements of platform effectiveness can be integrated with or provide comparisons for media mix models (e.g. Kannan et al., 2017), which identify the optimal distribution of a media budget across available advertising platforms, and models for bid optimization, which identify the appropriate value of an individual advertising impression (e.g. Edelman, Ostrovsky, & Schwarz, 2007). In this way, attribution models can inform decisions made by advertisers about aspects of campaigns they directly control, although the dependent variable (individual consumer product purchasing choices) and unobservable variables (platform effectiveness measures) of attribution models themselves are not directly controllable. The accuracy of the platform effectiveness measurements from the attribution model may be independently validated by the predictive performance of these dependent models.

One may view attribution modeling as inherently a problem of statistical inference: the intent is to measure an unobservable parameter (platform effectiveness). Indeed, Ji, Wang, and Zhang (2016) and Lei, Sanders, and Dawson (2017) explicitly formulate attribution modeling as a Bayesian inference task.

However, as in all supervised learning tasks, inferences from attribution models must be calibrated on the basis of their predictive performance on observed outcomes (§2). Because platform effectiveness is an unobservable parameter, there is no ground truth to directly validate its inferences, similar to the stellar physical parameters inferred from supernova observations discussed in §2. Therefore, Ji et al., (2016) and Lei et al. (2017) both assess inferences from models based on their predictive performance on consumer behavioral data such as the AUC, F1-score, and pointwise predictive density. While, as in the box office projection case, the variance of these individual predictions may be high, a rigorous inference procedure will assess the uncertainty of inferences on quantities such as platform effectiveness measurements, characterize their dependency on other model parameters and assumptions, and test their sensitivity to model mis-specification related to issues like platform coverage. In this way, advertisers can extract meaningful and reliable information about advertising channels despite limitations in predictive precision.

4.3. Industry Generalizations

Both the examples in this section illustrate applications where neither a predictive- or inference-oriented perspective by itself is adequate to extract all the available value from data and modeling investments made by businesses. The balanced perspective, able to extract information and insights from the modeling process while also using predictive measures to study the reliability and boundaries of those inferences, should be preferred.

The examples in this section also showcase the role of inference and prediction in different regimes of decision power. In some circumstances, companies or other actors will have direct control over an independent variable in a model, therefore providing indirect decision power over the outcome from a system (modeled as the dependent variable). An example would be the casting decisions in film production, contributing to box office performance. In this domain, inferences about the role of the independent variable in the system are directly actionable as they can provide decision support for choices made about that independent variable. In another regime, the actor may have much more tenuous decision power over the dependent variable (or even none at all). Examples would include models to predict macroeconomic trends or attribution models applied to measure the latent effectiveness of media platforms. In this regime, inferences from models of systems lacking decision power can inform choices made in related contexts. For example, inferences about the role of housing start rates in predicting macroeconomic outcomes can support the use of housing starts as a leading indicator in making investment or product release decisions, and inferences about platform effectiveness are actionable because they inform media mix models used to make decisions about media spending on different platforms. Model design processes for data science in industry should assess the actionability, e.g. the decision support role, of both inferential and predictive aspects of models.

5. Discussion

In the previous sections, I have described prediction as the output from models for data and inference as a mechanism by which people can learn from the comparison of models to data (§2). I have observed that there is a prevailing focus on prediction for data science in industry (§1) and shown that there is a growing focus on prediction in the research literature (§3), but have offered examples of the importance of a balanced perspective incorporating both inference and prediction in applied work in the entertainment industry (§4).

In this section, I will assess some of the possible implications of the trend towards prediction-oriented discourse among data scientists. In particular, because the success of data scientists in communicating about the workings and results of models is critical to the ability of an organization to learn from its data, I will explore the role that communication plays in the actual work of and outcomes from data science.

5.1. Implications of the Trend Towards Prediction

Arguably, the distinctions drawn in §2 are purely semantic and modelers working under either the predictive or balanced perspective can choose to address the full range of modeling challenges ascribed to inference and prediction without regard for these terms. Even if so, semantic distinctions can have real consequences within organizations.

The terms we as researchers use and emphasize in educating students, communicating to the public, developing strategy with and conveying results to executives, and talking to investors also impact the paths we follow in doing data science. Data scientists and other business stakeholders are collaborators in the definition of any modeling task. When data scientists present and justify their modeling work purely in terms of predictive performance, it tends to be evaluated on that basis. When a studio embarks on box office "projection," it implicitly frames the market modeling problem around prediction outputs rather than inference and information extraction, regardless of the actual use cases for the model. If the industry nomenclature referred to box office "contributor," "driver," or "attribution" analysis, the inferential goals of this modeling task may be more prominent and perhaps the literature cited in §4.1 would more frequently discuss causal inference and analysis of uncertainty. This co-dependence emphasizes the imperative for data scientists to communicate thoughtfully, clearly, and effectively within organizations to ensure that their modeling work is aligned to business objectives.

While the rate analysis of term usage from the quantitative academic literature from §3 is not a causal temporal analysis, it is not difficult to imagine that fields of application that have historically focused on statistical inference may be slower to recognize and adopt the benefits of new techniques driven by the predictive modeling community, such as advances in deep learning, and vice versa. Companies can mitigate the effects of this siloing by constructing data science teams with diverse representation of prior methodological and applied experience, encouraging interaction across functional and disciplinary divisions, promoting external collaborations and participation in conferences, and setting an expectation for reading sources and journals from a variety of disciplines and trans-disciplinary sources.

Within the research community, the recent investment in predictive methodologies suggests an opportunity to further capitalize on inferential techniques that compliment these advancements. Exciting new work on visualizing and interpreting the activations of deep neural networks (e.g. Li, Liu, Chen, & Rudin, 2017; Olah et al., 2018; Yosinski, Clune, Nguyen, Fuchs, & Lipson, 2015), new mechanisms for understanding the workings of machine learning models (e.g., Doshi- Velez & Kim, 2017; Guidotti et al., 2018; Samek, Wiegand, & Müller, 2017), and approaches for probabilistic deep learning (e.g. Gal & Ghahramani, 2016; Lake, Salakhutdinov, & Tenenbaum, 2015) are just a few exemplars of the opportunity that exists at the intersection of predictive and inferential approaches.

5.2. Data Science as a Language for Research

Communicating about any complex technical subject is challenging. Yet communication in industrial data science is further compounded by the diversity of audiences that may be stakeholders for any given model. Within a company, there may be other data scientists focused on similar problems, data scientists working with entirely different modalities of data, software engineers, creative experts (such as marketers or product designers), operations managers, executive decision makers, and more that all need to interpret and act upon the results of the same model. While there is certainly nothing new about the need for disparate actors across an organization to be coordinated with each other, the mutual agreement that they should coordinate on the basis of models of data, and data science generally, is perhaps the fundamental consequence of the “big data revolution” (McAfee, Brynjolfsson, Davenport, Patil, & Barton, 2012).

In this way, data science itself can be viewed as a new common basis for communication, inquiry, and understanding in both science and industry. For example, within business, data-driven strategy development may be formulated as a decision theoretic response to statistical inferences. Operational execution in the age of automation already routinely takes the form of a predictive modeling task. In this sense, data science may manifest a new common language shared by researchers across sectors and domains.

If so, part of our common work as practitioners in the field of data science is to create and standardize the very terms of discussion for research and decision making within businesses going forward. The choices we make as a community today in how to describe our purpose and our work may reverberate for decades by framing discussions not only in classrooms and journals, but also in laboratories, offices, and board rooms. Consider the phrase “big data.” It has the virtue of signaling part of what is exciting about this new field–the ability to manipulate, apply analytical methods to, and extract value from data on a scale once unimaginable–but has the vice of neglecting other aspects important to data science. It devalues the significance and complexity of work done on not-“big” datasets, which comprises much of the cutting edge work in academia and industry; it fails to invoke the coequal role of modeling in data science; and it ignores the critically important issue of data quality.

In general, there is room for data scientists to identify alternative language that communicates their meaning more clearly and directly to diverse audiences. Sometimes this can be accomplished by a translation, e.g. replacing a specific term like “singular value decomposition” with a generic term like “recommender system.” Other times, it may require deliberate explication of a confusing or misunderstood concept. For example, with respect to the communication of uncertainty in sensitive areas of public interest, Morton, Rabinovich, Marshall, and Bretschneider (2011) recommended emphasizing the actionable potential of the upside of the risk profile of climate change, and Manski (2015) comment on the risks of not reporting uncertainty in the publication of economic data by government agencies. Both provide recommendations for framing estimates of uncertainty as specific and useful assessments of the variability or risk associated with a system that should have productive consequences on decisions made in response to an analysis. The optimal approach to communicating any important topic with the potential for ambiguity will vary by subject and audience and deserves the thought and consideration of the data scientist as a domain expert.

Simplicity and concision are virtues in technical communication, but they should be used to elegantly explain difficult concepts and not to obscure or avoid them. When it comes to topics like uncertainty, models that may seem like black boxes, and inference about unobservable parameters, data scientists should strive to communicate more about modeling within organizations, not less. Likewise, Manski (2018) expressed the hope that "concealment of uncertainty is a modifiable social norm" addressable by increased awareness. It is incumbent upon the data scientist to identify the most effective ways to provide context for and to explain why their model acts the way it does, why they believe its implications, how they've made relevant methodological choices and assumptions, and what caveats remain in their implementation or analysis. Not every facet of a model needs to be belabored, but the ones that are most important to the data scientist will generally have significance for the audience as well. Communication should be viewed as an integral part of an outcome from a balanced model development process.

Acknowledgments

I thank the following individuals for insightful feedback on this work: the anonymous reviewers and editors, Karim Kassam, Xiao-Li Meng, Allen Riddell, and Arjun Sanghvi. I thank my colleagues at WarnerMedia Applied Analytics, WarnerMedia, and the Communicating Science Conference series (ComSciCon) for countless enlightening conversations that led to the perspective described here.

Disclosure Statement

Nathan Sanders has no financial or non-financial disclosures to share for this article.

References

Abhishek, V., Fader, P., & Hosanagar, K. (2012). Media exposure through the funnel: A model of multi-stage attribution. *SSRN*. https://doi.org/10.2139/ssrn.2158421

Amrhein, V., Trafimow, D., & Greenland, S. (2019). Inferential statistics as descriptive statistics: There is no replication crisis if we don't expect replication. *The American Statistician*, *73*(Suppl. 1), 262–270. https://doi-org.ezp-prod1.hul.harvard.edu/10.1080/00031305.2018.1543137

Asur, S. & Huberman, B. A. (2010). Predicting the future with social media. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology* (Vol. 1, pp. 492–499). IEEE. https://doi.org/10.1109/wi-iat.2010.63

Basuroy, S., Chatterjee, S., & Ravid, S. A. (2003). How critical are critical reviews? The box office effects of film critics, star power, and budgets. *Journal of Marketing*, *67*(4), 103–117. https://doi.org/10.1509/jmkg.67.4.103.18692

Berman, R. (2018). Beyond the last touch: Attribution in online advertising. *Marketing Science, 37*(5), 685–853. https://doi.org/10.1287/mksc.2018.1104

Betancourt, M. (2015). A unified treatment of predictive model comparison. *arXiv*. https://doi.org/10.48550/arXiv.1506.02273

Betancourt, M. (2018). *Towards a principled bayesian workflow (pystan)*. GitHub. Retrieved from https://github.com/betanalpha/jupyter_case_studies/blob/master/principled_bayesian_workflow/principled_bayesian_workflow.ipynb

Betancourt, M., Byrne, S., Livingstone, S., Girolami, M., et al. (2017). The geometric foundations of hamiltonian monte carlo. *Bernoulli, 23*(4A), 2257–2298. https://doi.org/10.3150/16-bej810

Billheimer, D. (2019). Predictive inference and scientific reproducibility. *The American Statistician*, *73*(Suppl. 1), 291–295. https://doi.org/10.1080/00031305.2018.1518270

Blei, D. M. & Smyth, P. (2017). Science and data science. *Proceedings of the National Academy of Sciences*, *114*(33), 8689–8692. https://doi.org/10.1073/pnas.1702076114

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, *16*(3), 199–231. https://doi.org/10.1214/ss/1009213726

Dalessandro, B., Perlich, C., Stitelman, O., & Provost, F. (2012). Causally motivated attribution for online advertising. In *Proceedings of the sixth international workshop on data mining for online advertising and internet economy* (p. 7). ACM. https://doi.org/10.1145/2351356.2351363

De Vany, A. & Walls, W. D. (1999). Uncertainty in the movie industry: Does star power reduce the terror of the box office? *Journal of Cultural Economics*, *23*(4), 285–318. https://doi.org/10.1023/A:1007608125988

Doshi-Velez, F. & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv*. https://doi.org/10.48550/arXiv.1702.08608

Draper, D. (1995). Assessment and propagation of model uncertainty. *Journal of the Royal Statistical Society: Series B (Methodological)*, *57*(1), 45–70. https://doi.org/10.1111/j.2517-6161.1995.tb02015.x

Edelman, B., Ostrovsky, M., & Schwarz, M. (2007). Internet advertising and the generalized second-price auction: Selling billions of dollars worth of keywords. *American Economic Review*, *97*(1), 242–259. https://doi.org/10.1257/000282807780323523

Elberse, A. (2007). The power of stars: Do star actors drive the success of movies? *Journal of Marketing*, *71*(4), 102–120. https://doi.org/10.1509/jmkg.71.4.102

Elberse, A. & Anand, B. (2006). Advertising and expectations: The effectiveness of pre-release advertising for motion pictures. *Harvard Business School*.

Eliashberg, J., Hui, S. K., & Zhang, Z. J. (2007). From story line to box office: A new approach for green-lighting movie scripts. *Management Science*, *53*(6), 881–893. https://doi.org/10.1287/mnsc.1060.0668

Finetti, B. d. (1937). La prevision: Ses lois logiques, ses sources subjectives. anuales de l'institut henripoineare, vol. 7, 1964. [Foresight: Its logical laws, its subjective sources. Studies in subjective probability.] New York: Wiley, 93–158.

Franklin, J. (2015). *The science of conjecture: Evidence and probability before pascal*. JHU Press.

Gal, Y. & Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International conference on machine learning* (pp. 1050–1059).

Geisser, S. (1993). Predictive inference. CRC Press.

Gelman, A. & Hennig, C. (2017). Beyond subjective and objective in statistics. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *180*(4), 967–1033. https://doi.org/10.1111/rssa.12276

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

Ghiassi, M., Lio, D., & Moon, B. (2015). Pre-production forecasting of movie revenues with a dynamic artificial neural network. *Expert Systems with Applications*, *42*(6), 3176–3193. https://doi.org/10.1016/j.eswa.2014.11.022

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018a). A survey of methods for explaining black box models. *ACM Computer Survey*, *51*(5), 1–42. https://doi-org.ezp-prod1.hul.harvard.edu/10.1145/3236009

Gundogdu, A. S., Sanghvi, A., & Harrigian, K. (2018). Recognizing Film Entities in Podcasts. *arXiv.* https://doi.org/10.48550/arXiv.1809.08711

Gupta, S. & Zeithaml, V. (2006). Customer metrics and their impact on financial performance. *Marketing Science*, *25*(6), 718–739. https://doi.org/10.1287/mksc.1060.0221

Harrigian, K. (2018). Geocoding without geotags: A text-based approach for Reddit. *arXiv*. https://doi.org/10.48550/arXiv.1810.03067

Hubbard, R., D. Haig, B., & A. Parsa, R. (2019). The limited role of formal statistical inference in scientific inference. *The American Statistician*, *73*, 91–98. https://doi.org/10.1080/00031305.2018.1464947

Imbens, G. W. & Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Ji, W., Wang, X., & Zhang, D. (2016). A probabilistic multi-touch attribution model for online advertising. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management* (pp. 1373–1382). ACM. https://doi.org/10.1145/2983323.2983787

Kannan, P. et al. (2017). Digital marketing: A frame- work, review and research agenda. *International Journal of Research in Marketing*, *34*(1), 22–45. https://doi.org/10.1016/j.ijresmar.2016.11.006

Karniouchina, E. V. (2011). Are virtual markets efficient predictors of new product success? The case of the Hollywood stock exchange. *Journal of Product Innovation Management*, *28*(4), 470–484. https://doi.org/10.1111/j.1540-5885.2011.00820.x

Killeen, P. R. (2005). An alternative to null-hypothesis significance tests. *Psychological Science*, *16*(5), 345–353. https://doi.org/10.1111/j.0956-7976.2005.01538.x

Krueger, J. (2001). Null hypothesis significance testing: On the survival of a flawed method. *American Psychologist*, *56*(1), 16-26. https://doi.org/10.1037//0003-066x.56.1.16

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, *350*(6266), 1332–1338. https://doi.org/10.1126/science.aab3050

Lash, M. T. & Zhao, K. (2016). Early predictions of movie success: The who, what, and when of profitability. *Journal of Management Information Systems*, *33*(3), 874–903. https://doi.org/10.1080/07421222.2016.1243969

Lauritzen, S. L. (1996). *Graphical models*. Clarendon Press.

Lavine, M. (2019). Frequentist, bayes, or other? *The American Statistician*, *73*(Suppl. 1), 312–318. https://doi.org/10.1080/00031305.2018.1459317

Lee, K., Park, J., Kim, I., & Choi, Y. (2018). Predicting movie success with machine learning techniques: Ways to improve accuracy. *Information Systems Frontiers*, *20*(3), 577–588. https://doi.org/10.1007/s10796-016-9689-z

Lei, V., Sanders, N., & Dawson, A. (2017). Advertising attribution modeling in the movie industry. In *Online materials for StanCon 2017* (Vol. 2017). StanCon. Retrieved from http://mc-stan.org/events/stancon2017-notebooks/stancon2017-lei-sanders-dawson-ad-attribution.html

Li, O., Liu, H., Chen, C., & Rudin, C. (2017). Deep learning for case-based reasoning through prototype's: A neural network that explains its predictions. *arXiv*. https://doi.org/10.48550/arXiv.1710.04806

Liu, T., Ding, X., Chen, Y., Chen, H., & Guo, M. (2016). Predicting movie box-office revenues by exploiting large-scale social media content. *Multimedia Tools and Applications*, *75*(3), 1509–1528. https://doi.org/10.1007/s11042-014-2270-1

Manski, C. F. (2015). Communicating uncertainty in official economic statistics: An appraisal fifty years after Morgenstern. *Journal of Economic Literature*, *53*(3), 631–53. https://doi.org/10.1257/jel.53.3.631

Manski, C. F. (2018). Communicating uncertainty in policy analysis. *Proceedings of the National Academy of Sciences*, *116*(16) 7634–7641. https://doi.org/10.1073/pnas.1722389115

McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D., & Barton, D. (2012). Big data: The management revolution. *Harvard Business Review*, *90*(10), 60–68. https://hbr.org/2012/10/big-data-the-management-revolution

Morton, T. A., Rabinovich, A., Marshall, D., & Bretschneider, P. (2011). The future that may (or may not) come: How framing changes responses to uncertainty in climate change communications. *Global Environmental Change*, *21*(1), 103–109. https://doi.org/10.1016/j.gloenvcha.2010.09.013

Ning, S., Qu, X., Cai, V., & Sanders, N. (2018). Clust-LDA: Joint model for text mining and author group inference. *arXiv*. https://doi.org/10.48550/arXiv.1810.02717

Olah, C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K., & Mordvintsev, A. (2018). The building blocks of interpretability. *Distill*. https://distill.pub/2018/building-blocks

Panaligan, R. & Chen, A. (2013). Quantifying movie magic with google search. *Google Whitepaper—Industry Perspectives+ User Insights*.

Pearl, J. (2014). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Elsevier.

Pennock, D. M., Lawrence, S., Giles, C. L., & Nielsen, F. A. (2001). The real power of artificial markets. *Science*, *291*(5506), 987–988. https://doi.org/10.1126/science.291.5506.987

Pyle, D. & San José, C. (2015). An executive's guide to machine learning. *Mckinsey Quarterly*, (3), 44–53. https://www.mckinsey.com/industries/technology-media-and-telecommunications/our-insights/an-executives-guide-to-machine-learning

Ricci, F., Rokach, L., & Shapira, B. (2015). Recommender systems: Introduction and challenges. In F. Ricci, L. Rokach, & B. Shapira (Eds.), *Recommender systems handbook* (pp. 1–34). Springer. https://doi.org/10.1007/978-1-4899-7637-6_1

Samek, W., Wiegand, T., & Müller, K.-R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv*. https://doi.org/10.48550/arXiv.1708.08296

Sanders, N. & Lei, V. (2018). The role of prior information in inference on the annualized rates of mass shootings in the united states. *Statistics and Public Policy*, *5*(1), 1–8. https://doi.org/10.1080/2330443x.2018.1448733

Sanders, N. E., Betancourt, M., & Soderberg, A. M. (2015). Unsupervised transient light curve analysis via hierarchical Bayesian inference. *The Astrophysical Journal*, *800*(1), 36. https://doi.org/10.1088/0004-637x/800/1/36

Sanders, N. E., Soderberg, A. M., Gezari, S., Betancourt, M., Chornock, R., Berger, E., ... Waters, C. (2015). Toward characterization of the type IIP supernova progenitor population: A statistical sample of light curves from Pan-STARRS1. *The Astrophysical Journal*, *799*(2), 208. https://doi.org/10.1088/0004-637x/799/2/208

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods*, *1*(2), 115–129. https://doi.org/10.1037//1082-989x.1.2.115

Shruti, Roy, S. D., & Zeng, W. (2014). Influence of social media on performance of movies. In *2014 IEEE International Conference on Multimedia and Expo Workshops*. IEEE. https://doi.org/10.1109/icmew.2014.6890664

Tintarev, N. & Masthoff, J. (2015). Explaining recommendations: Design and evaluation. In F. Ricci, L. Rokach, & B. Shapira (Eds.), *Recommender systems handbook* (pp. 353–382). Boston, MA: Springer US. https://doi-org.ezp-prod1.hul.harvard.edu/10.1007/978-1-4899-7637-6_10

Walls, W. D. (2005). Modeling movie success when 'nobody knows anything': Conditional stable- distribution analysis of film returns. *Journal of Cultural Economics*, *29*(3), 177–190. https://doi.org/10.1007/s10824-005-1156-5

Wang, Y. & Blei, D. M. (2018). The blessings of multiple causes. *arXiv.* https://doi.org/10.48550/arXiv.1805.06826

Yadagiri, M. M., Saini, S. K., & Sinha, R. (2015). A non-parametric approach to the multi-channel attribution problem. In *Lecture Notes in Computer Science: Vol. 9418. Web Information Systems Engineering – WISE 2015* (pp. 338–352). Springer. https://doi.org/10.1007/978-3-319-26190-4_23

Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., & Lipson, H. (2015). Understanding neural networks through deep visualization. *arXiv*. https://doi.org/10.48550/arXiv.1506.06579

Footnotes

1. A Google query at the time of this writing returns about 300,000 results: https://www.google.com/searchei=_Xv7W7CzJsm8ae64vfgC&q=mckinsey+machine+learning+describe+pr-

edict+prescribe, retrieved April 28, 2019 ↩

2. https://www.kaggle.com/ ↩

3. See e.g. https://towardsdatascience.com/ how-to-build-a-data-science-portfolio-5f566517c79c and references therein. ↩

4. https://arxiv.org ↩

5. https://arxiv.org/help/api/user-manual ↩

6. See, e.g., https://www.theringer. com/movies/2018/8/2/17641822/ box-office-reporting-mojo-the-numbers-marvel-star-wars for a discussion ↩

7. See, e.g., https://www.boxofficemojo.com/news/?id= 4455&p=.htm or https://pro.boxoffice.com/category/ boxoffice-forecasts/ ↩