ARTICLE

PREDICTIVE ANALYTICS AND LAW

Harry Surden[*]

Abstract

Many fields, including finance, medicine, and software engineering routinely rely upon predictive analytics in order to improve decision-making. Predictive analytics is the practice of using data, analyzed by algorithms, to inform uncertain decisions. The algorithms used often arise from the field of artificial intelligence. Recently, although still in its early stages, predictive analytics has also begun to emerge within the field of law. Lawyers, for example, have begun to rely upon analyzed data to better estimate litigation outcomes, damage awards, and even the success of individual motions. This article will examine predictive analytics as used in litigation, describing its methods, uses, state of the art, and limits. One central conclusion is that attorney predictions can be improved by relatively basic analytic interventions. For example, attorneys should aim to root certain litigation predictions in real-word data where it is available, rather than relying upon purely intuitive estimates.

Importantly, advanced predictive analytics remains as much art as science and as such, requires careful use and application. Consequently, the users of predictive analytic data in law must be able to properly interpret the results and place them in context in order to make better and more accurate legal predictions. Notably, predictive models are only as good as the data that goes into them, and many legal data sets may have skews or reflect selection bias because they are based upon publicly released data. For instance, the damage amounts reported in some public datasets may be misleading and unrepresentative of damage amounts generally, due to the fact that the majority of damages data goes unreported due to settlement under confidentiality agreements. A lawyer relying upon such predictive data must understand the various selection effects that may have gone into the data, and properly adjust their interpretations or risk making wrong or overly confident decisions. This article will also explore some of the ways in which attorneys engaged in predictive analytics must appropriately calibrate their analyses.

[*] Professor of Law, University of Colorado Law School, Affiliated Faculty, Stanford University CodeX Center for Legal Informatics.

INTRODUCTION

Professionals often operate in environments of uncertainty. Sometimes this uncertainty is about future events. [1] For example, a doctor may be indecisive about whether a surgery will improve a patient's condition, or an attorney might have difficulty predicting a client's outcome in pending litigation. Other times, there is uncertainty, not about the future, but about some aspect of the present. This is usually due to incomplete information. For example, even after a positive diagnostic test result, a doctor may not know if a patient actually has a particular disease. Interestingly, it may be true as a factual matter that the patient either does, or does not, have the disease at that moment. However, the doctor only has limited information about that underlying reality, as the test result might be a false positive. Likewise, a corporate counsel might only be able to estimate the likelihood that a firm owes particular contractual duties. Perhaps these duties are truly specified in a particular set of contractual clauses somewhere, but as a practical matter, they may be obscured within a voluminous trove of firm-wide contracts.

Predictive analytics is the practice of using data, analyzed by algorithms, to assess such uncertain contexts.[2] Many fields have incorporated such formal, data-oriented practices into their decision making and analysis. For example, credit card firms routinely use predictive analytics to identify fraudulent transactions in real-time. Within logistics, firms computationally analyze past data to help predict future demand or supply chain interruptions. Other domains ranging from finance, transportation, electronic commerce, software, entertainment to

[1] Probability is often informally associated with uncertainty about *the future*. But it is more accurate to associate probability with uncertainty wherever it occurs, future, past or present. It so happens that many of the things that are uncertain are those in the future (e.g., results in an upcoming election). Because of randomness and other unpredictable factors, we will not know the outcome of future events until after they happen. However, we can also be uncertain about things in the present moment (e.g. Is there treasure buried under this rock?) or even the past (e.g. Did Nixon deliberately erase the Watergate recordings? Did this defendant commit the crime?). Probability can thus express uncertainty about the future, present, or past.

[2] Jaquie Finn et al., Predictive Analytics for Healthcare (2020).

professional sports today routinely use predictive analytics to aid in decision making.

Such a data-driven approach is sometimes contrasted with less formal predictive processes.[3] Professionals, for instance, also produce estimates in environments of uncertainty but typically do so using combinations of trained judgment, analysis, intuition, common sense, domain knowledge, experience and other factors.[4] Historically, within the practice of law, such prediction based upon trained judgment has been the norm while the use of analytics comparatively rare. Recently, although still in early stages, predictive analytics has begun to emerge within legal practice.[5] Attorneys are starting to incorporate computationally analyzed data, rather than judgment alone, to better estimate uncertain aspects of law, such as litigation outcomes, damage awards, and even the probability of success of individual motions.[6]

This article explores the role of predictive analytics in law, describing its methods, uses, state of the art, and limits. Because predictive analytics has been able to improve decision making in fields outside of law, there are reasons to believe that it can improve analysis at least in *some* legal contexts.[7] However, there are a few points worth emphasizing at the outset. First, it is a common misperception that analytics involves replacing analytical data for trained judgment or substituting humans with computers. This is not what experience in other fields has shown. Rather,

[3] Richard E. Nisbett et al., The Use of Statistical Heuristics in Everyday Inductive Reasoning., 90 Psychological Review 339–363 (1983).

[4] See, e.g. Lauren Vogel, Gut Feelings a Strong Influence on Physician Decisions, 190 CMAJ E998–E999 (2018) (describing how physicians often make intuitive predictions), Gerd Gigerenzer, How I Got Started: Teaching Physicians and Judges Risk Literacy, 28 Appl. Cognit. Psychol. 612–614 (2014), Jaquie Finn et al., Predictive Analytics for Healthcare (2020) ("Historically, healthcare professionals have relied on patient self-reporting and their own judgment to understand and predict how a disease might progress.")

[5] See, e.g. Daniel Martin Katz, Quantitative Legal Prediction--or--How I Learned to Stop Worrying and Start Preparing for the Data-Driven Future of the Legal Services Industry, 62 Emory L.J. 909 (2013).

[6] *See, e.g.,* Daniel Martin Katz, Michael J. Bommarito & Josh Blackman, A General Approach for Predicting the Behavior of the Supreme Court of the United States, 12 PLoS ONE e0174698 (2017), KEVIN D. ASHLEY, ARTIFICIAL INTELLIGENCE AND LEGAL ANALYTICS: NEW TOOLS FOR LAW PRACTICE IN THE DIGITAL AGE (6th printing ed. 2019)., Ed Walters, Data-Driven Law: Data Analytics and the New Legal Services (2019), 12.

[7] AJAY AGRAWAL, JOSHUA GANS & AVI GOLDFARB, PREDICTION MACHINES: THE SIMPLE ECONOMICS OF ARTIFICIAL INTELLIGENCE (2018).

predictive analytics is best thought of as a tool that can help *supplement, support, and improve* the analysis of trained human experts, such as attorneys.[8] The appropriate use of analytics can provide professionals with a more comprehensive set of information upon which to base complex, uncertain decisions. [9]

Moreover, although predictive analytics can improve analysis in some legal settings, this does not mean that it is necessarily useful in *all* legal contexts. Rather, this article will emphasize that predictive analytics often remains as much trained art as science, and as such requires skill to understand where and how it can be appropriately used in law. If employed without adequate understanding, or in an inappropriate setting, it can lead to inferior decisions. To illustrate, consider that predictive analytics requires information about the world to be translated to a well-structured form, such as numerical data, that a computer can easily process. This is not always straightforward, for in many legal contexts there is relevant information that is difficult to express as data but that attorneys understand to be crucial. For instance, a particular client might come across as particularly sympathetic (or unlikeable) in ways that may meaningfully affect legal outcomes. Such nuances, while germane to actual legal outcomes might be difficult to fully capture in data, and as such many not be sufficiently represented in computational models.

Similarly, there is a real risk that unsophisticated legal users of predictive analytics will make unjustified decisions if they do not properly understand their analytics model, its limits, and what the analysis is actually telling them. For instance, predictive models in law are often built upon the data that happens to be available, such as government databases of filed federal or state lawsuits. However, such litigation information may not be representative of legal disputes broadly at every stage. Rather this data may exhibit selection-bias effects since they reflect the small subset of legal

[8] *See, e.g.,* AJAY AGRAWAL, JOSHUA GANS & AVI GOLDFARB, PREDICTION MACHINES: THE SIMPLE ECONOMICS OF ARTIFICIAL INTELLIGENCE (2018).

[9] AJAY AGRAWAL, JOSHUA GANS & AVI GOLDFARB, PREDICTION MACHINES: THE SIMPLE ECONOMICS OF ARTIFICIAL INTELLIGENCE (2018).

disputes that have progressed all the way to litigation, compared to the wider universe of all informal and formal legal disagreements that never even reach an attorney, let alone a courtroom. Such selection effects may provide skewed results, if care is not taken. Lawyers using predictive analytics must therefore understand the various distortions that have gone into the data and computer mode and properly adjust their interpretations, or risk making wrong, or overly confident predictions.

In sum, even when available, legal predictive analytics do not, on their own, provide reliable answers to uncertain legal assessments. Rather, the users of such tools in law must have the skills to properly interpret and contextualize results in order to actually make better decisions. However, when properly used, predictive analytics has the promise to improve decision-making within for lawyers in certain contexts.

I. What Are Predictive Analytics?

"Predictive analytics" describes the process of making estimations about some unknown aspect of the world (whether in the future, past, or present) by incorporating results from computer systems that have analyzed relevant data. [10]

Broadly speaking, the concept of predictive analytics is often contrasted against less formal predictive methods. For instance, physicians have historically diagnosed patient conditions by incorporating information such as symptoms, diagnostic test information, domain knowledge about diseases and clinical research, patient medical history and applying trained judgment, intuition, and experience.[11] By contrast, a doctor using predictive analytics to diagnose a patient's disease status would still use all of the informal processes above but might add to that assessment more formal analyses of patient characteristics associated with

[10] AJAY AGRAWAL, JOSHUA GANS & AVI GOLDFARB, PREDICTION MACHINES: THE SIMPLE ECONOMICS OF ARTIFICIAL INTELLIGENCE (2018).

[11] DANIELE CHIFFI, CLINICAL REASONING: KNOWLEDGE, UNCERTAINTY, AND VALUES IN HEALTH CARE (2021), https://doi.org/10.1007/978-3-030-59094-9 (last visited Nov 4, 2021).

disease data from the broader population. Similarly, attorneys are thought to make predictive judgments for clients by applying trained judgment, intuition, domain knowledge of the law, reasoning, and experience to the facts of a particular situation.[12] An attorney using predictive analytics might, in addition to the factors above, incorporate data such as litigation success rates in a particular jurisdiction, or root damage award predictions in computational models that incorporate past damage award data.

A. Basic and Advanced Predictive Analytics

It is helpful to conceptually divide predictive analytics into basic and advanced predictive analytics. "Basic predictive analytics" can be thought of as simple summary statistics of relevant data, such as averages (e.g., the average litigation damage award in negligence automobile accident cases in a jurisdiction) and base-rate probabilities (e.g., the typical plaintiff win rate in such negligence lawsuits in a jurisdiction). By contrast, "advanced predictive analytics" involve computational models that incorporate modern machine-learning methods from artificial intelligence to examine data for predictive patterns. This article will examine both.

Why conceptually divide predictive analytics into basic and advanced versions? The reason is that much of the attention from the media, business sector, and academic literature tends to focus upon the complex, state-of the art advanced predictive analytics systems built upon artificial intelligence methods and large datasets.[13] However, research shows that professionals can see improved decision-making simply by incorporating very basic data, such as averages or base-rates, into some predictive processes that

[12] PAUL BREST & LINDA HAMILTON KRIEGER, PROBLEM SOLVING, DECISION MAKING, AND PROFESSIONAL JUDGMENT: A GUIDE FOR LAWYERS AND POLICYMAKERS (1st ed ed. 2010), 16.
[13] *See, e.g.,* AJAY AGRAWAL, JOSHUA GANS & AVI GOLDFARB, PREDICTION MACHINES: THE SIMPLE ECONOMICS OF ARTIFICIAL INTELLIGENCE (2018).

are today based primarily on trained intuition.[14] While it is true, as shall be discussed, that advanced predictive systems built upon machine-learning can also provide impressive results in particular uncertain situations, that need not be the focus for most attorneys. Rather, significant improvements can be achieved by attorneys just by tethering their predictions to even rudimentary data, rather estimating solely on intuition, judgment, and personal experience alone.

1. Basic Predictive Analytics and Bayesian Reasoning

Studies have shown that professionals can make better predictions when they incorporate even basic analytical data, such as averages or base-line probabilities into their forecasting. Consider an example from medicine involving what is known as the base-rate fallacy. To understand this, it is first helpful to have a basic background on Bayesian reasoning.[15]

The concepts of probability and Bayesian reasoning are related. We live in a world where many things are uncertain and unpredictable. Probability refers to the various processes that we use for estimating things that are uncertain (e.g. uncertainty of prevailing in a hypothetical lawsuit). Some of these probability processes are mathematically formal while others are approximate and intuitive. Bayesian reasoning refers a particular aspect of any probability estimation process: once we have already some probability estimate, how should we *change* that original probability estimate – upwards ("It's now seems more likely" or downwards ("It now seems less likely") once we receive new, additional information that is relevant (e.g., "It seems more likely that we will prevail after learning new information that significantly strengthens our client's legal position.").

We can think of there being two variants of Bayesian reasoning: intuitive, informal Bayesian reasoning and mathematically formal

[14] *See e.g.,* William M. Grove & Paul E. Meehl, *Comparative Efficiency of Informal (subjective, Impressionistic) and Formal (mechanical, Algorithmic) Prediction Procedures: The Clinical–Statistical Controversy.*, 2 PSYCHOLOGY, PUBLIC POLICY, AND LAW 293–323 (1996).

[15] Bayes' reasoning, and Bayes' rule, is named after Thomas Bayes who discovered the mathematical relationship in the 18th Century. JAMES V. STONE, BAYES' RULE: A TUTORIAL INTRODUCTION TO BAYESIAN ANALYSIS.

Bayesian reasoning.[16] Humans are thought to mostly engage in informal, and approximate Bayesian reasoning most of the time, as opposed to the more mathematically formal and precise version. For that reason, let us focus first on intuitive, informal Bayesian reasoning to help us develop an intuition about the process. I will then explain the more formal mathematical approach.

a. Intuitive Bayesian Reasoning

Humans live in a world of uncertainty and are constantly estimating the probability of uncertain events using heuristics and approximations. Informal Bayesian reasoning simply refers to the process of intuitively changing our probability estimates up or down, based upon new relevant information that we learn. A familiar example from everyday life will help illustrate this point. Imagine that you are uncertain as to whether you think it will rain on a particular day, and you are trying to decide whether to bring an umbrella to work or not. For simplicity, assume that you do not have access to a weather forecast. You can probably make a rough guess at the outset based upon the month. Say that it is June, and you know from experience that it is fairly unlikely to rain on a typical day in June where you live, as compared to the rest of the year. However, there is additional information that you can use to improve your prediction – looking out the window. If you, see a bright sunny sky free of clouds, you might think that your original estimate was reasonable. If, by contrast, you see a dark sky full of clouds, you might change your initial, informal probability estimate of rain upwards. You know from experience that it tends to rain more on cloudy days. Thus, based upon that new information, your belief that it might rain today is probably going to be higher than before you had looked out the window. Perhaps you might opt to bring an umbrella to work.

This is an example of the intuitive type of Bayesian reasoning that people do all the time. In that instance, you began with some original estimate – a *prior* belief about how likely it was to rain on

[16] PAUL BREST & LINDA HAMILTON KRIEGER, PROBLEM SOLVING, DECISION MAKING, AND PROFESSIONAL JUDGMENT: A GUIDE FOR LAWYERS AND POLICYMAKERS (1st ed ed. 2010).

any particular day in June. That prior probability estimate – a low chance of rain - was an approximation based upon your rough impression of the average rate of rain - your experience that it tended to rain fairly relatively infrequently in June in your area. You then acquired some new, pertinent information – you looked out the window and saw dark, cloudy skies. Further, you knew from experience that dark clouds tend to be associated with rain and can be a reasonably good predictive signal of that uncertain event. You then intuitively reasoned that it was probably more likely to rain than you had originally thought, having observed the dark sky.

In other words, you revised your prior probability estimate upwards after incorporating new, relevant information associated with the uncertain thing you were trying to predict. This way of thinking is referred to as *intuitive* Bayesian thinking, because it roughly approximates the more formal mathematical Bayesian process (described below), in the sense that both involve updating earlier probability estimates for uncertain things (e.g. will it rain today) based upon learning new relevant information or evidence (e.g. dark clouds). However, although this approximate process is quite useful for everyday predicting, as shall be discussed, intuitive Bayesian reasoning sometimes departs from formal Bayesian reasoning in ways that can lead to inaccurate intuitive predictions.

Nonetheless, intuitive Bayesian reasoning is a useful predictive process commonly used by attorneys, doctors and other professionals. For instance, a doctor may be uncertain whether a patient has a particular disease upon identifying a few relevant symptoms. But that doctor may become more confident that the patient actually has that disease after receiving a positive result in a diagnostic test that is designed to detect that condition. In that case, the doctor is using intuitive Bayesian reasoning – she is revising her original probability estimate that "the patient might have the disease" upwards to "there is a pretty good chance the patient has the disease" upon receiving new, pertinent information – a positive test result.

Similarly, attorneys often assess the strength of a client case based upon an initial intake of information and a quick assessment under the law. But as attorneys learn new information about the situation, they often change their assessment. For example, in a negligence context involving an automobile accident, an attorney might be pessimistic about an injured client's chance of prevailing, given no indication of careless conduct by the other driver. Under the law, negligence would require a showing that the party that caused the injury – the other driver – was driving carelessly. However, the attorney might revise her assessment upwards after learning of data showing that the other party was texting while driving at the time of the accident. She might now think that her client has a stronger case – a higher probability of proving negligent conduct in a hypothetical lawsuit – than she originally thought before learning of these facts. In this instance, the attorney engaged in intuitive Bayesian reasoning. She made an original probability assessment by assessing her client's facts under the law, and then revised her probability estimate of prevailing upwards ("stronger case") from her prior estimate ("weaker case") after receiving additional, relevant information associated with a higher chance of success.

Of course, Bayesian reasoning tells us that learning new information associated with *lower* probabilities should similarly cause a professional to revise their estimate downwards, that the uncertain event is *less* likely. For example, if the doctor had instead received a negative diagnostic test result – information that is negatively associated with having the disease – she should now think that the chances of her patient actually having that disease are less likely than her prior, less-informed diagnostic assessment. Similarly, if the attorney had instead learned that it was *his client* who had been texting while driving, rather than the other driver, the attorney's original assessment of the case's strength might drop even further, considering how this new information might undermine his client's legal position in a future negligence case. In essence, Intuitive Bayesian reasoning involves using heuristics and approximations to make better guesses about the world as we gather new information that suggests things are more or less likely than we had originally thought.

b. Formal Bayesian Reasoning

Formal Bayesian reasoning has a similar spirit to the informal process just discussed: it involves changing probability estimates upwards or downwards, based upon learning new information that is associated with higher or lower probability. The difference is that formal Bayesian reasoning uses mathematical processes and actual calculations to produce these upwards or downwards probability shifts, rather than impressionistic beliefs. Formal Bayesian reasoning also requires actual quantities and numerical probability estimates in order to conduct assessments. I will first provide a brief (largely non-technical) overview formal Bayesian reasoning, and then explain how it relates to predictive analytics and improving the decision-making of professionals in certain uncertain contexts.

Formal Bayesian analysis has specific (and sometimes confusing) terminology, but it is worth describing those terms because the language is commonly used.[17] Let us revisit the rain example. In any prediction, we are trying to estimate the chances of some uncertain thing. The term that Bayesian analysis sometimes uses to refer to the thing that we are trying estimate is "the hypothesis." In this example, the hypothesis would be the probability it is going to rain today.

We began with an intuitive guess as to the probability of rain before we had looked out the window and gathered more information, and then we revised our guess after looking out the window and seeing dark skies. In Bayesian terminology, we call our previous estimate of the chance of rain, before we looked out the window, the "prior probability." This earlier estimate sometimes informally it is referred to as the "prior." Why is it called the "prior"? It is because we had an "old" probability estimate (e.g. "It's probably not going to rain because it's June") that we made prior to getting new relevant information. But now that we have additional information (e.g. dark skies), we need to update our prior, less-informed estimate to a current, more informed estimate that incorporates what we have learned (e.g., "It

[17] STONE, *supra* note 16.

seems more likely that it will rain, now that I see that there are dark clouds in the sky"). That new, more informed and hopefully more accurate estimate is referred to as the "posterior probability", because it is our best revised guess *after* (or posterior to) learning additional, relevant information.

The main difference between formal and intuitive Bayesian is that formal Bayesian analysis uses actual numerical probabilities and calculations. So, instead of non-numerical prior probability estimates, such as "it is unlikely to rain today because it is June" we have to assign an actual numerical probability, a number between 0 (0% or not possible) and 1 (100% or certain). A decimal such as .1 (or 10%) represents a 10% probability of rain in June, whereas a number such as .9 would represent a 90% chance.

Where did we get such a prior probability number from? In other words, how can we possibly come up with any numerical estimate for any uncertain thing initially? In Bayesian statistics, initial probability estimates are subjective, meaning that we can, in principle, assign any percent that we think is right given our intuition. However, just because Bayesian statistics allows us to assign any arbitrary probability to any uncertain event (e.g. 90% chance of rain), does not mean that arbitrary assignment is *accurate* or reflects the reality of the world. Rather, as I shall discuss, it is important to get the prior probability as accurate to reality as possible if we can, or we risk making very inaccurate initial and updated predictions, in light of new information.

Consider the task of trying to put an actual number on the probability of rain on a typical day in June in your area. One approach could be to assign an estimate based upon your experience and your loose impression of how often it rains (e.g. "I remember it raining about 20% of the time"). However, as much psychological research has shown, such heuristic approximations can be notoriously inaccurate as they are subject to multiple cognitive biases.

Cognitive biases are systematic errors or deviations in judgment or assessment that humans tend to make, due the tendency of our brain's analytical systems to rely upon mental shortcuts or heuristics. Consider just one cognitive bias (among many others) that can make intuitive probability estimates quite inaccurate: the availability heuristic. The availability heuristic refers to the tendency of the human mind to make estimates based upon information that is comparative easy to recall or that happen come to mind. The problem is that information that happen to arise in memory may not be statically representative of the underlying phenomenon that we are aiming to estimate. So, for example, in subjectively assigning a probability of rain 20% we might base this upon our general impression and memory of the past of it raining about 1 out of 5 days the previous June.

The major problem is that our memories and impression may not accurately reflect the underlying numerical reality as to how many days it actually rained. Due to bounded memory and cognition, we probably don't accurately remember the actual weather on every day from a year past. Moreover, we might selectively remember certain rainy days (or sunny) days because they left a bigger impression due an unusual rainfall or something else memorable. The important point is that largely intuitive estimates can sometimes be extremely inaccurate due to the availability bias, and multiple cognitive biases inherent in human reasoning.

Fortunately, there is a very sensible choice of prior probability that can get us surprisingly far in many circumstances where it is available – and that is the average or "base-rate" probability, based upon data. What is the average probability (or base rate) of some event? It is the typical rate at which we expect that the to occur over many possible times. Since the goal of probability is to make the *best* estimate about uncertain events that we can under the circumstances, we can often do better than intuitive, cognitive bias-prone probability estimates, simply by looking at basic data from the past. So, for example, if it is available, we might gather data for weather in June in our area over the past five years, and see the fraction (or proportion) of days that it was sunny, out of the total number of days. Say, for instance, we examined data of 100

June days from past years and saw that it rained on only 10 of those 100 days. We would now have an estimate, based on data, of the base-rate probability of rain in June - 10% (10/100) probability of rain on an average day. Base-rate estimates, or average probabilities, are simply fractions. In this case, we take the number of days it did rain (10) and divide it by the number of days it could have rained (100), and we come up with a simple proportional estimate (1/10 or 10% or .1). To be clear, in many instances of prediction, such base-rate calculations may not available. Sometimes the event that we are trying to estimate may be unique so that past data would not be available. In such a case, a purely subjective estimate might be the best we can do. But in other cases (as in many instances in law) relevant base-rate data *is* available and it is simply not being used.

Such a base-rate estimate – even though the calculation is extremely simple (it is merely a fraction) can be surprisingly powerful because it allows us to have a reasonably solid handle on the prior probability since it is rooted in actual data. To be clear, looking at sample data from the past provides an *estimate of average probability*. It is not *the* probability in any objective sense. But in many cases such an average is the best that we can do, and it is often much better than the subjective, impressionistic probability assessments that people intuitively make, and that are subject to cognitive biases. So, one important point is that predictions can often be made much more accurate simply by moving from away a purely subjective probability estimates to those tethered to some reasonably representative data when such data is available. So, in the case of the rain example (and many other estimates) a reasonable point to start for a basic first-cut estimate is with a base-rate, or average probability – a fraction rooted in past data.

The important point for this discussion is that simply by having an base-rate arising from data, rather than purely from intuition, one can dramatically improve one's predictions. However, let me briefly mention the other aspects of formal Bayesian analysis. We mentioned that Bayesian probability involves revising our probability estimates in light of new, informative data. How do we actually calculate our more accurate, new probabilities based upon additional information?

Depending upon the data that's available, we can sometimes use the formulas of conditional probability. Without getting into the details, consider that with conditional probability, we are no longer asking the question "What is the average chance of rain in June?" Rather we have new, informative information – dark skies, so we ask the question, "Given that we have observed dark skies, what is the average chance of rain in June?" Again, the math amounts to little more than basics fractions within our 100 day dataset of past weather for June. To estimate conditional probability, we would examine the number of days with dark skies. Say that there are 15 dark sky days (out of 100 total days). Then we can focus in on only those dark sky days, and figure out the fraction of those 15 days that it rained. Imagine that it rained on 9 out of 15 of those dark days, which is 9/15 or .60 or 60%. In this example, we have determined the predictive relationship of dark skies and rain: Of the days when there have been dark skies in the past, 60% of them have resulted in rain. So, our best updated guess now, having seen dark skies, is that there will be a 60% chance of rain, revised up from our original prior estimate of 10% before we had looked out the window and gained that additional information.

However, sometimes we have access information about the relationship between the thing we are trying to predict (e.g. rain), and the associated predictive signal (e.g. dark skies), except in a less useful form. For instance, imagine that we had access instead to the converse data from the prior paragraph: the proportion of rainy days in June that had dark skies. In other words, given the number of rainy days in June, what proportion of those rain days had dark skies. Observe that this is a different conditional probability than we discussed previously. Previously we had the following informationL given that we had days with dark skies, what proportion of those resulted in rain. Here we have the converse: given that we had days with rain, what fraction of those rainy days had dark skies. Interestingly, using the mathematic relationship known as Bayes' rule, we can sometimes still make the correct calculations as to how we might update our predictions for rain given after having observed a cloudy sky.

The details of the Bayes' rule calculation are somewhat complex and not relevant for our purposes. However, there *are* a few important points to draw out of the Bayes' rule discussion with respect to making good predictions. Bayes' rule is simply a slightly more complex way of updating our prior probabilities upwards or downwards based upon new information when we have information (e.g. probability of dark skies given rain) that is slightly less useful than the information that we really want (e.g. the probability of rain given dark skies). So, it must be transformed mathematically. However, the first point is that in many cases, the data that is available for us to make updated predictions comes in the less wieldy form that requires Bayes' rule. The second is that accurate calculations under Bayes' rule often depend *heavily* on getting an accurate base-rate. If we get the base-rate wrong our Bayes' rule calculations for our updated probability in light of new information, can be wildly inaccurate. This will be illustrated below

2. Improving Decision-Making With Basic Analytics

As the prior discussion showed, predictions based upon intuition can be skewed due to cognitive biases. By contrast, the use of simple fundamental analytics – averages and base-rates - can help improve predictive analysis. Let us first look at a well-known example of the base-rate fallacy in medicine, and then see how that same fallacy applies to law, and how analytics data can help in such a circumstance.

The base-rate fallacy occurs when we are making a prediction, but we do not accurately represent or properly take into account the base-rate, or average probability of the thing that we are estimating, when we are updating our probabilities in light of new information. Consider a doctor who is trying to determine the probability that her patient has Disease A after her patient has received a positive test diagnostic result for that disease. It is well known that diagnostic tests are not always accurate – they sometimes produce false positive or false negative results. Imagine that the doctor knows some data about the accuracy of the test – the sensitivity: if the test is given to who actually has Disease A, 90% of the time it will return a positive result (a true positive), and 10% of the time it will return a negative result (a false negative). A doctor, relying upon intuition, might be tempted upon seeing the positive result and considering the 90% sensitivity rate information, to conclude that there is a 90% probability that the patient now has the disease, given that she received a positive test result. But this is incorrect. What the doctor wants to know is the probability the patient has the disease given that there has been a positive test result. But the doctor actually has information about the converse in her test information: she has the probability that this diagnostic test will a positive test result, given that the patient actually has the disease.

This example is similar to the rain example discussed. The conditional probability data that the doctor wished she had was the following: given that we have a true positive test result, what is the probability that the patient has the disease. In reality, the doctor has access only the converse data: given that the patient has the disease, what is the probability of getting a positive test result.[18]

As discussed earlier, even with such unwieldly converse data, the doctor can still correctly calculate the probability that the patient has the disease given a positive test result. It is just that the doctor must have an accurate base-rate calculation in order to do the calculation properly. In medicine, the base-rate is called the prevalence, and it refers to the percentage of the population who has the disease. We can think of it as the average probability of a random person in the population having the disease. If 3 out of 10 people have a particular disease, then the base-rate (or prevalence) is 30%. Other diseases are rarer, and can have rates such as 1% (1 out of 100 or .01) or even .01% (1 out of 10,000).

[18] The reason is that this converse data is often easier to collect. A scientist can find 100 people that are already known to be sick with the disease, administer the test, and then see the proportion of those known sick individuals who test positive. Then the sensitivity or accuracy of the test can be calibrated

Applying Bayes' rule in this way to rare diseases – those with a low base-rate or prevalence – can lead to surprising and counter-intuitive results. For example, let's imagine that Disease A has a base-rate (or prevalence) of 1% or .01 (1 out of 100 people in the population have it). And we are using a test with a 90% sensitivity rate (90 out of 100 people who are truly sick who take the test, actually test positive). Using Bayes' rule we come to a counter-intuitive result: even with a test with a 90% sensitivity rate, the probability that the patient actually has the disease, given a positive test result, is only about 9%. This might be quite shocking to doctor unfamiliar with statistics, as she might expect the probability to be around 90% (which is the quoted test sensitivity), rather than the actual 9% result. The key observation is that the very low base-rate of 1% (only 1 in 100 people actually have the disease) is "anchoring" the probability down, even after updating to include the positive test result. If we were to randomly test people, 99% would not have the disease. So even with a highly sensitive test, we happen to be testing many people who do not actually have the disease. Most of the results are thus false positives, due to the sheer number of people in the population without the disease.

There are many ways that the above probability calculation could have gone wrong. First, the doctor might not have had an accurate handle on the base-rate, and this inaccurate base rate would have greatly distorted the prediction about the patient's disease status. Imagine that instead of relying upon actual prevalence data, the doctor simply estimated the base-rate using intuition and experience, and arrived at a 20% prevalence rate (rather than the actual 1%). In that case, using the Bayes' rule the updated probability would have been wildly inaccurate: the doctor would have calculated that her patient had 70% chance of having the disease (given her inaccurate base-rate estimate), whereas the patient's true probability was only 9%. Second, it is confusing to understand whether one has the predictive data that we want (the probability of disease given a positive test result, the probability of rain given a dark sky), or the more unwieldy converse that requires

Bayes' rule (the probability of positive test result given illness, the probability of dark skies given rain). Finally, the calculation itself - applying Bayes' Rule (or even basic conditional probability) - is mathematically complicated and prone to calculation error.

The larger point is that many of these diagnostic inaccuracies in can be resolved by simply using basic predictive analytics. Research has shown that doctors whose use such analytical systems in contexts like this make better predictions. The medical analytics systems can have access to actual prevalence or base rate data about disease, thereby preventing inaccurate intuitive assessments of probability.[19] Moreover, the systems themselves do the correct calculations automatically, obviating the need to understand the different types of data, or the mathematical formulas. Thus, the predictive capabilities of the medical professional, with the right patient and population data aided by computation, and appropriately contextualized, can lead to improved predictive and diagnosis accuracy compared to an informal assement with little contextual data.

3. Improving Attorney Decision Making with Basic Analytics

Within law, attorneys commit similar errors in probabilistic assessment. Thus, access to even basic predictive analytics data is likely to improve predictions in some contexts. First, consider the analogous base-rate problem in law. Lawyers are frequently asked to perform predictions for clients, such as the probability of success of a hypothetical case in a particular jurisdiction under the client's specific set of facts. Often estimates are made using the informal methods previously mentioned. However, as the previous discuss has made clear, informal estimates in which probabilities generated from intuitive processes strongly diverge from the underlying true probabilities can lead to highly inaccurate predictions.

For instance, consider an attorney who informally estimates, based upon his experience and judgment, that the

[19] Amos Cahan & James J Cimino, *A Learning Health Care System Using Computer-Aided Diagnosis*, 19 J. MED. INTERNET RES. e54 (2017).

probability of a typical plaintiff prevailing in a negligence dispute in his jurisdiction seems to be about 50%. As the previous discussion indicated, such informal predictions are subject to a variety of cognitive biases, including availability and selection effects. Imagine that the true-base rate – the average rate at which plaintiffs actually prevail in the jurisdiction is closer to 15%. At the outset, any of the attorney's informal predictions are going to be significantly inaccurate, as they are based in an unrealistic base-rate probability. Moreover, as the attorney revises her probability judgments upwards or downwards, based upon new information that is favorable or unfavorable to her client, her updated probabilities will be even more inaccurate, as these new predictions will be strongly distorted by the original, incorrect base-rate.

As in the medical context however, a relatively simple method of improving such legal predictions is to simply anchor initial estimates in real-world data. This is the type of basic predictive analytics that we see emerging in the realm of litigation. Firms such as Lex Machina, LexisNexis, Clio, and Bloomberg, have begun to provide easy access to basic predictive analytics data. Such basic statics include information about base-rates – average win rates for plaintiffs in particular types of lawsuits in certain jurisdictions, average damage awards, motion success rates, etc.

Attorneys can thus make more informed, and likely more accurate predictions simply by making relatively minor changes to their workflow that incorporate simple statistical summary data that is easy for non-technically trained audiences to understand. For example, the attorney above could use basic predictive analytics to determine a reasonable estimate, based upon past data, of the actual probability of a typical plaintiff prevailing in a negligence dispute in his jurisdiction. Such an estimate, since it is based upon actual data is likely to result prior probability that is closer to the true average rate that that the attorney is trying to estimate. Since lawyers also commit probability errors that are analogous to the medical base-rate errors described previously, even this relatively minor intervention of basing initial probability estimates on actual data is likely to improve later probability estimates.

B. Advanced Predictive Analytics in Law

1. Predictive Analytics, Data, and Machine Learning

Although perhaps somewhat less commonly used today by attorneys than the basic analytics systems just described, more complex predictive analytics systems are also emerging within law. Such advanced systems offer more sophisticated predictive capabilities, but also require more sophistication to build, use, and interpret.

Most advanced legal analytics systems are built using machine-learning. Machine-learning refers to a family of technological approaches arising out the field of artificial intelligence. The major characteristic of a machine-learning method is the ability to detect patterns in data. Common machine learning techniques that the reader might have encountered include logistic regression, support vector machines, naïve bayes, and neural-network based approaches that fall under the category of deep-learning.

The principal way that a machine-learning system is built is by providing a learning algorithm with example data. Such machine-learning algorithms are designed to detect patterns in data, and those detected patterns can later be used for prediction. For example, consider a high-level (and oversimplified) description of a machine-learning system that could be used to estimate the probability of prevailing in litigation. Such an advanced predictive system could be built by providing a machine learning algorithm with example data representing past cases (or legal disputes), and associated outcomes from those cases, such as whether the case settled or went to trial, costs, and damage amounts.

A crucial step in this process involves translating relevant real-world information about actual cases into structured data that a computer can process. Thus, at some point, one has to create a computer readable dataset of past cases. To accomplish this, those with domain expertise in will have to decide upon a set of features (or characteristics) of cases that they believe are likely to be useful in making predictions about legal outcomes. For instance, in a negligence automobile injury case, one predictive feature might be how careless the other driver was, or another might be whether that driver had an elevated blood alcohol level. Those domain experts will have to determine a suitable translation process from the relevant characteristics of the cases to numerical data.

Next, the individual cases must be "coded" or translated from subjective case characteristics (e.g. How careless was the defendant?) to numerical data that can be processed by a computer (e.g. carelessness rating on a scale from 0 – 5). Typically, each case would be coded along multiple dimensions that represent different possible case characteristics – (e.g. age of lead plaintiff and defendant). Sometimes a dataset will have hundreds or thousands or even more features, usually represented numerically. Datasets are typically represented in row-column format, with each row representing an individual example (in this instance a row would represent an individual case), and the columns representing the individual features of the case that distinguish them from one another (e.g. Column 1 Plaintiff name, column 2 Plaintiff age, etc). Associated with each case, typically in columns to the right of predictive features, would be data about how the case resolved and other information that might be useful to predict (e.g., Did the case settle before the trial? Did it go to trial? Did the plaintiff prevail at trial? What were the damages)

Having converted real-world case information into a numerical data set, the next step would be typically to train a machine-learning model by providing it with these examples. Machine-learning algorithms are designed to analyze multiple examples and find patterns within that example data that are the most predictive. For instance, since each case in the prior case dataset has a series of numerical features that distinguish the case, and a case outcome – so-called labeled data (e.g. Did the plaintiff go to trial and prevail?), the machine-learning algorithm would be able to "learn" those case features that tend be most predictive of, say, a plaintiff prevailing. After analyzing a sufficiently large number of labeled examples, the machine-learning algorithm can then encode the pattern that it detected in a machine-learning model. That model now contains a numerical representation of the case characteristics that machine-learning algorithm determined are the most effective at predicting case outcomes.

This trained-machine learning model can now be used to make predictions on new cases. When a new, never-before-seen case comes along, information about that new case can be converted into data that matches the earlier case. Then that data about the new case can be provided to the trained-machine learning model which will can output a probabilistic prediction for that case, based upon patterns gleaned from past case data (e.g. In this new case, a 60% probability of success is predicted).

It is not evident how common such advanced predictive analytics are in law today, as much of the internal development is kept secret, but there is evidence that some law firms, as well as financial firms such as hedge funds, are using such advanced predictive analytical systems to make legal predictions. However, it is not clear how widely used these systems are, nor is it clear how much better the predictions of these systems are, compared to the baseline estimates offered by basic analytical systems.

One reason that advanced analytics systems have been slower to arise in law, as compared to other areas, probably has to do, in part at least, with the relative inaccessibility of legal data for computational analysis. In science, engineering, business, and other domains that routinely use predictive analytics, there are often widely available sources of data to analyze in the public domain. Often these widely available engineering datasets are produced by universities or governments and are available for others to build upon at will. By contrast, within law, much potential legal data remains inaccessible for analysis for all practical purposes. First, the vast majority of legal data remains private and confidential. For example, if one were to desire to create a system capable of making accurate predictions in the realm of contracts, one would need a source of contract data to analyze that was representative of contract data broadly. However, most commercial contracts are kept confidential and are not available for public analysis. Additionally, much of the data about law remains in human readable text. By contrast, most machine-learning analytical systems operate best when dealing with highly structured, primarily numerical data. While certain natural language processing technologies have become moderately capable at automatically analyzing legal documents, for the most part, highly accurate data analysis will require a human translation step from natural language to structured data. Thus, in addition to being generally inaccessible due to secrecy, the data that is available tends to exist in forms that are less amenable to computation and analysis.

2. Decisions About the Future Based Upon Past Data

The core of all predictive analytics practices is thus making decisions, at least in part, based upon data. The central logic of predictive analytics systems is that *past* data can be helpful in predicting present or future qualities that are uncertain. Implicit in this concept is that the uncertain thing that we are trying to predict (is similar, in a meaningful way, to the past data). If the past data is not representative of future data, or the new example that we are trying to predict is significantly different from the past, predictions can go awry.

There are thus several caveats to consider in both advanced and basic predictive analytics systems when used in law. As the prior discussion indicated, analytic systems are built by computationally analyzing past data. To be the most useful and robust in future prediction, the data upon which such systems are built should be representative. However, in many cases, there are likely to be skews or distortions in the data, primarily due to selection bias effects. In many cases, these distortions in the training data can produce in accurate or distorted predictive results.

For example, consider the basic analytical data that is provided by firms such as Lex Machina. As described earlier, such base-rate data and average data is extremely useful, as it can provide estimates that are rooted in actual cases. However, many of these analytics systems draw their information from public sources, such as the Federal PACER litigation data system, or state-level case databases. Such data will reflect selection biases, due to the fact that they represent legal disputes that have proceeded all the way to litigation. This, of course, represents only a small fraction of legal disputes broadly. The vast majority of disputes about law occur informally and never make it to an attorney. Many of those that do reach an attorney may never make it past the counseling or advice stage. Of those that do, many of those are resolved by attorneys through negotiation or cease and desist actions. Many of those that make it beyond that stage are resolved by formal settlement prior to litigation. Finally, a small subset of the wider universe of disputes actually results in formal litigation. Moreover, of the disputes that are commenced, the vast majority end before trial, either through settlement or summary judgment. Thus, the data available in electronic lawsuit databases suffer from a variety of selection bias effects that do not necessarily make them representative of all legal disputes generally. This is not necessarily problematic, as long as these biases and distortions are properly accounted for by those who use this data. However, most attorneys are not trained in data analysis. So there is a real risk that attorneys draw unwarranted conclusions from the data.

Similar selection bias problems can affect even the advanced predictive analytics systems described earlier. Consider a large law firm that builds a machine-learning predictive analytics system based upon its own internal data of past cases and legal disputes. This too is likely to reflect biases that, if not properly accounted for, can lead lawyers astray. For one, machine-learning systems often (but not exclusively) operate best when they have large numbers of examples, in the tens of thousands or millions. By contrast, a firm may have a comparatively small training set of only thousands of cases, and there might be enough differences among the examples cases to make prediction not as robust as possible. Additionally, the sample cases will exhibit the bias of those cases that tend to go that firm (e.g. perhaps due to attorney relationships). That is not necessarily problematic for predicting future cases, as long as that selection bias is represented in the new cases as well, but it still may produce unwarranted predictive skews. Again, such issues are not catastrophic, as long as they are properly accounted for by users who contextualize them.

Additionally, some cases are unique or have special features that are difficult to capture in data. For example, imagine a new case that is meaningfully different from past cases, in ways that are not reflected in the data. Perhaps the case involves an accident that received an extreme amount of publicity. Because of the unique characteristics of the new case are not fully accounted for in the data and the predictive analytics model, any analytics predictions for that case may be misleading in light of these difference. Once again, the key is for the attorneys to properly and appropriately contextualize and interpret the results in light of the limitations of the data, and the unique characteristics of their case.

Conclusion

This article has survey some the basic and advance predictive analytics systems used in law today. A central conclusion is that improvements in legal predictive capabilities can be achieved by simple interventions.