# COMPETING ALGORITHMS FOR LAW:
## SENTENCING, ADMISSIONS, AND EMPLOYMENT

*Saul Levmore & Frank Fagan*[*]

Abstract

*Algorithms have found their way into courtrooms, college admission committees, and human resource departments. While defendants and other disappointed parties have challenged the use of algorithms on the basis of due process or similar objections, it should be expected that they will also challenge their accuracy, and attempt to present algorithms of their own in order to contest the decisions of judges and other authorities. The problem with this approach is that people who can transparently see why they have been algorithmically denied rights or resources can manipulate an algorithm by retrofitting data. Demands for full algorithmic transparency by policymakers and legal scholars are therefore misguided. To overcome algorithmic manipulation, we present the novel solution of algorithmic competition. This approach, versions of which have been deployed in finance, works well in law. We show how the state, a university, or an employer should set aside untested data in a lock-box. Parties to a decision then develop their respective algorithms and compete. The algorithm that performs best with the lock-box data wins. While this approach presents several complications which the Article discusses in detail, it is superior to full disclosure of data and algorithmic transparency.*

***



[*] Levmore is the William B. Graham Distinguished Service Professor of Law at the University of Chicago Law School; Fagan is an Associate Professor of Law, EDHEC Business School, France. We benefited from discussions with colleagues at a University of Chicago Law School workshop, and with Concetta Balestra Fagan and Eliot Levmore.

## I. INTRODUCTION

When the past is thought to predict the future, it is unsurprising that machine learning, with access to large data sets, wins prediction contests when competing against an individual, including a judge. Just as computers predict next week's weather better than any human working alone, at least one study shows that machine learning can make better decisions than can judges when deciding whether or not to grant bail.[1] Courts have, therefore, started accepting machine learning when contemplating prison sentences as well as bail releases or denials.[2] In both cases, failure is judged largely on the basis of recidivism, and partly on the basis of a failure to appear for court hearings or to abide by requirements issued by parole officers.[3] The use of data to decide an individual's fate raises constitutional and other legal questions, but given courts' acceptance of data thus far, this Article sets these constitutional and ethical questions aside and examines the use of data with the new tools now available to litigators and courts. In any event, and as we will see, there

[1] *See* John Kleinberg, et. al., *Human Decisions and Machine Predictions*, 133 Q. J. Econ. 237, 237-38 (2018) (developing policy simulations that reduce crime by 24.7% with no change in jailing rates or that reduce jailing rates by 41.9% with no change in crime rates). For related work, see Himabindu Lakkaraju & Cynthia Rudin, *Learning Cost-Effective and Interpretable Treatment Regimes*, 54 Proceedings Machine L. Res. 166, 166 (2017); Jongbin Jung, *Simple Rules for Complex Decisions*, Stanford Univ. Working Paper 1, 1 (2017).

[2] *See* Brandon Garret & John Monahan, *Judging the Use of Risk Assessment in Sentencing*, Judicature (forthcoming 2019) (documenting an increased use in statistical risk assessment tools in sentencing and, notably, the endorsement of those tools by drafters of the 2017 revision to the Model Penal Code and the FIRST STEP Act, a sweeping federal sentencing reform enacted in 2018); John Logan Koepke & David G. Robinson, *Danger Ahead: Risk Assessment and the Future of Bail Reform*, 93 WASH. L. REV. 1725, 1729 (2018) (documenting the introduction of statistical risk assessment tools for pretrial release decisions in 14 states since 2012).

[3] *See, e.g.*, 18 U.S.C. § 3142 (b) (providing for pretrial release of the accused on the basis of flight and crime risk).

are many other applications of machine learning in law, and some steer clear of constitutional objections. It should also be noted that the persistent objections take different form when data are used by courts as opposed to legislatures. Some of this Article's suggestions are probably best directed at legislatures, rather than at judges ruling on single cases with litigants who may not have the time or resources to engage in algorithmic arguments. Nevertheless, a single judge making a binary decision is a good place to start, in part because there already exist such cases, as discussed presently in Part I.A. A careful look at the application of machine learning to a legal decision about the length of a person's prison term, reveals a number of things that can help us understand the future of machines and algorithms in law.[4] While this Article offers a set of suggestions for improving the use of machine learning in judicial decision-making, it is also applicable to legislatures that might be amenable to significant changes in several areas of law.

### *A. Algorithmic Sentencing in Criminal Cases*

We begin with, and then dwell on, the relatively recent and well known Wisconsin Supreme Court decision in *State v. Loomis*.[5] The state had accused Eric Loomis of participating in a drive-by shooting, and he eventually pled guilty to two lesser charges. In preparation for sentencing by a lower court, a Wisconsin Department of Corrections officer produced a presentencing investigation report that included an assessment of risk based on an outside firm's algorithm. The firm had used data, given to it by the state, about many previous recidivist (and non-recidivist) wrongdoers after their release. The

[4] We use the term "machine learning" for a tool that finds hidden connections, reaches conclusions, and often improves on its own after humans have given it goals and data. Machine learning is best understood as a subset of artificial intelligence, which is a general term that includes things that humans have done (or still do) which can be carried out by a machine. We use both terms specifically in a way that indicates a transfer of investigation of facts and decision-making away from humans. *See* Frank Fagan & Saul Levmore, *The Impact of Artificial Intelligence on Rules, Standards, and Judicial Discretion*, 93 S. Cal. L. Rev. n.1 (forthcoming 2019).

[5] 881 N.W.2d 749 (Wis. 2016). For a discussion on the dangers of bias in risk assessment tools such as COMPAS, the tool used in *Loomis*, Julia Angwin et al., *Machine Bias*, ProPublica (May 23, 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing; *but see* Anthony W. Flores et al., False Positives, False Negatives, and False Analyses: A Rejoinder to "Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And it's Biased Against Blacks.", 80 Fed. Prob.. 38 (2016) (disputing the Angwin article and suggesting there might be less racial bias than previously thought); *see generally* Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, Cass R Sunstein, *Discrimination in the Age of Algorithms*, 10 Journal of Legal Analysis113-174 (2018); *see also* Bernard E. Harcourt, *Risk As a Proxy for Race*, University of Chicago Public Law & Legal Theory Working Paper No. 323 (2010) (advocating that courts should not attempt to predict recidivists because of the serious dangers of racial inequity.)

algorithm, "COMPAS," and the method by which it was produced, was disclosed neither to either court nor to the defendant. The state supreme court held that a trial court's use of an algorithmic risk assessment in sentencing does not violate the defendant's due process. Loomis lost.[6]

The algorithm was formed by inspecting a large data set, as well as the defendant's criminal history (along with data provided by the defendant in an interview) as a means of estimating the risk of recidivism.[7] Such algorithms do not (yet) perform a cost-benefit analysis, which might take account of the cost of incarceration and the cost of various potential and predicted crimes, but a judge is free to do so, armed with the algorithm's forecast. Defendant did not like the outcome of the lower court's evaluation (six years in prison), and objected to the court's reliance on an algorithm with unrevealed details. For example, if the algorithm included race or a factor highly correlated with race, the defendant might have had serious grounds for objection.[8] The algorithm almost surely took sex and age into account, and the defendant insisted on a right to individualized justice rather than an aggregative (or for that matter to an individualized) approach that took sex into account. The outside firm that produced the prediction offered by the state to the lower

[6] Other due process claims have been brought against risk assessment tools in Iowa, but have been rejected for procedural reasons. *State v. Guise*, 921 N.W.2d 26, 31 (2018); *State v. Gordon*, 921 N.W.2d 19 (Iowa 2018). The core issue, whether or not risk assessment tools are permissible if they indirectly use race, has not been resolved. One pro-se litigiant has brought an equal protection claim against COMPAS in Wisconsin, and the court held that *Loomis* does not restrict a litigant from seeking relief under an equal protection suit. At the time this Article is being written, a trial date has not been set. *Henderson v. Stensberg*, No. 18-CV-555-JDP, 2020 WL 1320820 (W.D. Wis. Mar. 20, 2020). It seems likely that courts will eventually exclude tools that intentionally look for inputs that are correlated with race, but it seems inevitable that tools, such as employment status, will be permitted even though they obviously are correlated with race.

[7] For a comprehensive description of COMPAS' inner workings, *see* Tim Brennan, William Dieterich, Beate Ehret, *Evaluating the Predictive Validity of the COMPAS Risk and Needs Assessment System*, 36 Criminal Justice and Behavior 21, 22-25 (2009).

[8] If a risk assessment tool did explicitly include race, a defendant could make a conventional equal protection claim. In a different context, the University of Texas's "Personal Achievement Index," which explicitly considered an applicant's race, passed strict scrutiny when it showed "concrete and precise goals" in relation to educational diversity. *Fisher v. Univ. of Tex. at Austin*, 136 S. Ct. 2198 (2016). *See* Aziz Huq, *Racial Equity in Algorithmic Criminal Justice*, 68 Duke L.J. 1043, 1084, 1098 ("[T]he weight of precedential evidence (as well as common sense) suggests that the mere fact that a decision-maker can *observe* the race of subjects does not mean that resulting action is therefore invalid [on equal protection grounds]). That COMPAS does not rely on race as an explicit training variable "reflect[s] corporate risk aversion, not an effort at legal compliance." *Id* at 1097.

court, claimed that the algorithm as well as details about its inputs and methods were trade secrets.[9]

In siding with the state on the important issues, the supreme court emphasized that there was no evidence that the lower court specifically took sex into account and, in any event, the algorithmic result was just one of several factors used by the lower court in reaching its sentencing decision. Future courts were admonished to use multiple factors and individualized assessments as well as the added benefit of these mysterious algorithms.[10]

Let us imagine that Wisconsin outsourced its consideration of the available data to firm W, rather than to firms X and Y, which might also have been eager for the work. It is immediately obvious that if Wisconsin had actually employed all three companies, and then preferred the prediction generated by W, because it suggested the harshest sentence for Loomis, defendant would have an excellent claim.[11] Algorithm shopping makes for bad statistics; indeed, the state would need to reveal the results produced by X and Y, just as it cannot secretly shop for its favorite psychiatric exam or lie detector test.[12] That defendants can do so is another matter, and most are constrained

---

[9] Again, many variables considered by judges (and also by algorithms) are likely to be correlated with gender or race. For instance, given prevailing levels of segregation, home address and zip code can surely serve as correlated proxies. *See* Anupam Datta et. at., "*Proxy Discrimination in Data-Driven Systems: Theory and Experiments with Machine Learnt Programs"* arXiv, https://arxiv.org/abs/1707.08120v1. Federal sentencing clearly prohibits the use of "race, sex, national origin, religion, creed, and socio-economic status." Federal Sentencing Guidelines § 5H1.10. Proxies can be expected to be prohibited in future Guidelines and other methods of sentencing to the extent that their correlation to clearly prohibited variables is strong and easily identifiable. Thus, the Sentencing Commission exclude factors like age and drug abuse, even though "empirical research has shown that … [those factors predict] recidivism." 2018 U.S.S.G. chapter 4. As we shall see, in the world of competing algorithms, the state, employer, university, or other entitiy sets the parameters of the competition, which can entail requirements to exclude certain variables identified by the competition's sponsor, including correlated proxies. Eventually courts will need to decide how to accept or exclude variables that are correlated with race. See Joan Petersilia, Susan Turner, *Guideline-Based Justice: Prediction and Racial Minorities*, 9 Crime & Just. 151, 173-75 (1987) (expressing skepticism that predictive algorithms can elimite racial disparities in sentencing since eliminating racially correlated variables decreses predictive accuracy.)

[10] *Loomis*, 881 N.W.2d. at 769; *see also* State v. Jones, No. 2015AP2211-CRNM, 2016 WL 8650489, at 4 (Wis. Ct. App. Nov. 29, 2016) (applying the *Loomis* requirement that judges must consider "many factors" in addition to algorithimic risk assessments while sentencing to avoid due process violations).

[11] *See* Daubert v. Merrell Dow Pharmaceuticals, Inc., 509 U.S. 579, 590 (1993) (holding that scientific evidence must be based upon scientific principles and methods in order to establish "a standard of evidentiary reliability").

[12] For several examples in which prosecutors sought biased experts to support the state's case, and were eventually discovered leading to reversals, see *passim* Paul C. Gianelli, *The*

by cost and the government's ability to show the jury that a witness regularly produces results that favor defendants. In any event, we return to the matter of multiple and competing algorithms below, as the title of this Article suggests.

Imagine further that Wisconsin offers W a huge amount of data about past defendants, including features known before sentencing, as well as behavior after they are released. The data might include persons from all 50 states in order to build up the data set.[13] One state will rarely have access to all the information from other states but, over time, professionals like W can build up large data sets. The manner of data recordation is likely to differ among states, and this can amount to an omitted variable but, generally speaking, the bigger the data set the easier and more accurate is W's work. Data from multiple states is likely superior to data from Wisconsin alone, even though it is possible that Wisconsinites are somehow different from released prisoners in other states.

### *B. Imperfect Algorithms*

It is easy to see that W's statistical technique is made imperfect, or at least more difficult, by the fact that the data are incomplete and tarnished because there is not a random selection of offenders. W can only study the background and behavior of those who were paroled, and make guesses about how those who remain incarcerated might have acted had they been released. To see this difficulty, imagine that judges are biased against accused people who are very tall, perhaps because judges are intimidated by persons much larger than themselves. If tall persons are imprisoned until they are of an age when they are virtually incapable of committing serious crimes, then we do not know whether Loomis's height matters.

Even if the bias against tall people is less severe, or is limited to some judges, it can easily reduce the number of relevant observations in W's data set, and make it impossible to assess properly the statistical likelihood that Loomis will be an early recidivist.[14] Loomis might be very tall, and the data might have shown that such a tall subject has rarely or never been a recidivist.

---

*Abuse of Scientific Evidence in Criminal Cases: The Need for Independent Crime Laboratories*, 4 Va. J. Soc. Pol'y & L. 439 (1997).

[13] For example, initial development of COMPAS, the risk assessment tool used in *Loomis*, consisted of 30,000 survey responses given between January 2004 and November 2005 by inmates, probationers, and parolees from multiple jurisdictions. *See* Koepke & Robinson, *supra* note 2 at 1758. Another pretrial risk assessment tool, developed by the Laura and John Arnold Foundation, drew on 750,000 cases from 300 jurisdictions. *Id.*

[14] *See infra* Part IV.

However, machine learning along with W's ingenuity is now stacked against Loomis because non-recidivists who look like him are in prison because of judicial bias; they are not in W's algorithm in a way that does justice for tall people. This is the case even if the state provides data about height, and W has the good sense to feed this data into the machine learning process. It is possible that the "machine" will see that few recidivists were tall, but inasmuch as few (or no) tall persons were released in the first place, there is not much to learn from the available data.

This is the price we pay for not having a large pool of randomly selected subjects released after a short period of incarceration, and for not allowing law to over-rule some judges and randomly assign short sentences.[15] It is also a problem that arises because states are unlikely to gather and record all available data; someone decides what to record and how to record it. Of course, defendant is free to point this out to a judge and, as we will argue, one of the great advantages of encouraging judges to use common sense, and not to rely entirely on the prediction produced by companies like W, is that it offers a defendant the opportunity to point to a characteristic of his that was excluded from the information available during the algorithm's development.[16] Even without this safety valve, data are of course useful, and machine learning can produce superior predictions. The safety valve, if available, makes this more likely—though it requires defense attorneys to be sophisticated about statistical methods, and to have the resources to evaluate the available data or to acquire additional data.[17]

Among other advances, in Part II we suggest some solutions to the problem of strategically designed and yet competitively successful algorithms. The idea is in part that when parties—such as a criminal defendant seeking a

---

[15] *See* Michael Abramowicz, Ian Ayres & Yair Listokin, *Randomizing Law*, 159 U. PENN. L. REV. 929, 964-74 (2011) (asserting that law should randomly assign people and firms to different legal rules to assess their efficacy).

[16] We develop this point in the context of selecting between rules and standards in Fagan & Levmore, *supra* note 4. When legal environments are subject to rapid change, machine learning is disadvantaged. Combinations of human judges and machines that apply standards or broad rules are superior to machines acting alone that apply narrow rules. The same is true when legal environments are inconsistent across space and the patterns identified by algorithms in one place are substantially different in another. *See* Frank Fagan, *Standardized Data Collection: Legal Requirements, Guidelines, or Competition?* 2 GNLU J. L. & Econ. (forthcoming 2019).

[17] The need has long been noted, but may be growing. *See* David H. Kaye, *Statistics for Lawyers and Law for Statistics*, 89 Mich. L. Rev. 1520, 1520 (1991) (noting the need); Janice Arellano, *Statistics and Law Practice*, ABA Practice Points (Dec. 23, 2015), https://www.americanbar.org/groups/litigation/committees/minority-trial-lawyer/practice/2015/statistics-and-law-practice/ (noting the urgency).

shorter sentence—offer competing algorithms, there is a second-mover advantage. One solution we offer involves opening the competition to the world. As we will see, it is possible to have a fair competition among algorithms, where "fair" refers both to the battle for statistical virtuosity as well as fairness to the defendant.

Before proceeding to more complex problems and solutions, it is useful to ask what we mean by "superior" predictions. How do we know if machine learning has done a good job? How does Wisconsin know which purveyor of algorithmic results to select, and whether it is worth paying for its services? Is there such a thing as a prosecution-friendly source of these algorithms, and therefore a provider to which defendants should object? In the world of Finance, where a great deal of progress with artificial intelligence can be observed, success is much easier to identify. A straightforward approach is to see whether more profit is earned with a given algorithmic approach—a term that refers to machine learning, usually accomplished with some insight about what data to offer the machine, and often what theory to investigate. Money offers an obvious tool with which to measure performance. A more sophisticated approach in the world of finance, would look for risk-adjusted returns, rather than apparent profit. The analyst should evaluate the product of the modern statistical technique with the overall performance of a market or particular investment (like Treasury bonds), adjusting for risk. It must be apparent that identifying superiority in the world of finance, or for that matter in medicine where years of life might make for a decent measure of success, is easier than in the world of law. Not all crimes committed by released prisoners inflict equal costs, and there is no easy way to compare a crime committed soon after release from one committed three years later.[18]

## II. COMPETING WITH THE STATE'S ALGORITHMS

### *A. Overfitting and Dividing Data*

Even if we can agree on measures of performance, it is important to see that statistical methods of the sort used in *Loomis*, as well as in finance and medicine, always face the danger of overfitting. Any programmer can look at a large set of data drawn from observable past experience, and develop some algorithm that appears to match the available information. Data might be fit not with a straight line (an assumption of linearity) or even a power function,

[18] To some extent, law does attach values to different crimes. It provides sentencing guidelines that can be used to say how much worse an armed robbery is than a mere breaking and entering. If one is satisfied with these measures, then the argument in the text becomes more straight-forward.

but with something full of zigs and zags, conventionally fitted to meandering curves. If this algorithm is developed from data, or even limited data, about the performance of stocks in 2010-2015, for example (though we will return to criminal defendants), it will be the case that someone who magically had this information and produced an algorithm at the start of 2010 could have made a fortune "predicting" the next five years and investing accordingly. If the goal is to earn profit from predictions about 2016-2017, a decent and now widely accepted approach, or solution to this problem of over-fit,[19] is to develop an algorithm based on 2010-2013 data, and then see how it performs on the "withheld" data available from 2014-2015.[20] This is significantly better than testing the algorithm on 2013 data, because that data has already influenced the prediction algorithm that was based on data from 2010 through 2013. Lest all this sounds easy, it should be noted that, in practice, successful pure machine learning based on 2010-2013 data, and then tested on the withheld data from 2014-2015, rarely performs much better than what can be produced by talented competitors when it comes to 2016-2017, even though the competitors did worse when tested on the withheld 2014-2015 data.[21] The financial world is just not the same after several years, and of course other investors have also changed their behavior based on their understanding of the available data and changed circumstances. Nevertheless, someone seeking to predict 2016-2017 prices, might as well use the algorithm that did best on the withheld 2014-2015 data, but it is unfortunate to see that while the machine-learning strategy of withholding data corrects for overfitting, it

[19] The problem of overfitting is easy to see with an example. Suppose we want to predict whether the roll of a pair of dice is more likely than not to come up twelve when thrown. We might build an algorithm and collect data from a number of rolls with various dice, noting their color, weight, the time the roll was thrown, whether the experimenter crossed fingers, and the temperature within the room. If the dice are fair, the correct answer produced by the algorithm is "no." If it happens that the dice are thrown thousands of times and that on five of these rolls with an orange pair of dice, weighing 7.2 grams, thrown at 3:15 P.M. in a 64-degree room, with the experimenter crossing her fingers, the dice come up with two sixes, then the algorithm may incorrectly construct a path that predicts "yes" in that case. The algorithm has overfit the data, and fallen prey to the fact that *some* combination of observations is likely, though hardly expected to be seen again in a set aside set of observations. *See* Stuart Russel & Peter Norvig, *Artificial Intelligence: A Modern Approach* 705 (Pearson 3d ed 2010).

[20] *See generally*, George E. P. Box et Al., *Time Series Analysis: Forecasting and Control* (5th ed. 2015) (using differencing to handle non-stationary data).

[21] This has led to the introduction of "Long Short-Term Memory" algorithms which seek to better preserve persistent features of earlier data. *See* Sima Siami Namin & Akbar Siami Namin, *Forecasting Economic and Financial Time Series: ARIMA vs. LSTM*, ARXIV.ORG (Mar. 16, 2018), https://arxiv.org/abs/1803.06386 at 8 (describing how LSTM uses memory gates to filter earlier data that is useful for prediction).

is not a perfect solution. A decent theory that guides "supervised" or assisted machine learning is often a superior approach.[22]

If the discussion to this point seems pessimistic, we should keep in mind the goal of machine learning in both law and financial investing. Machine learning that is relatively unsupervised,[23] with state-of-the-art dividing and withholding of data, surely outperforms an individual forecaster or charlatan investment adviser, but the real winner is apt to be more supervised machine learning, and often with so much supervision that it can hardly be called machine learning. The 2010-2013 data is likely to be, and is sensibly, investigated with data that is manipulated by a theory provided by an intelligent human. Intelligently drafted starting points, and selection of characteristics, are likely to reduce overfitting while also making better use of the available data.

In the world of finance, for instance, one who has a theory about how recent elections or new climate change statistics affect stock performance, and therefore a theory about what to look at in constructing an algorithm to test the two sets of data, with one withheld, is likely to do far better when it comes to predicting the future or a third set of data. In most settings there is a limit to the data that can be offered (even) to the most capable computers. With or without limits, an algorithm can easily overfit unless it is equipped with a theory or some human judgment. In the financial world, for example, a statistician can find that Apple stock goes up in price when the Yankees won by 3 runs on the previous day. An acolyte of the new tool might say, with hind-sight, that machine learning has uncovered the fact that big victories raise the spirits of stock exchange investors, especially in New York, who then buy Apple because of their new enthusiasm. It will be seen as an example of data revealing a new theory, which is then of predictive value, if future victories by the team are followed by increases in the price of Apple stock. Statistically speaking, this result might also be confirmed when tested on

[22] A decent theory reduces the number of candidate hypotheses, often exponentially, that must be tested with data. *See* Russel & Norvig, *supra* note 19 at 697 ("There is a tradeoff between the [empirical representation…] of a hypothesis space and the complexity of finding a good hypothesis within that space.") *See also* Pedro Domingos, *The Master Algorithm* 73 (Basic Books 2015) (noting that when simply one variable is added to a dataset, the number of candidate hypotheses for explaining that data expands exponentially).

[23] *See* Russel & Norvig, *supra* note 19 at 694-95 (explaining that unsupervised learning involves algorithmic observation of patterns in data even though no explicit instructions are given by the observer). We say *relatively* unsupervised because at some early point a decision must be made about the data that is fed to the process. This choice of data implicitly reflects some intuitions or theorizing about what is relevant and what might yield useful insights, and especially surprising ones.

withheld data, which is of course limited; the Yankees win by three runs just a few times each season, and Apple has been publicly traded for just two decades. Baseball scores, like the number of letters in players' names, might wisely be excluded from the data offered to the machine, and this might make room for better predictions. The trick is to exclude things that bring about merely chance correlations, while leaving room for machine learning to help us find true connections that were not previously imagined.[24]

This early excursion into machine learning may give the reader clues about its use in courts. Still, another word of caution is in order. Machine learning, whether assisted or not, has a great advantage in using a large data set or, as we have seen, two such sets. Thousands of stocks and bonds, traded over many hours a day and more than 250 days a year, yield a very large data set, even after the Yankees' results are excluded. In Wisconsin, the statistician has a tougher time. The number of prisoners released after a given number of years of incarceration, and then the number of recidivists committing serious crimes, is relatively small, even before other variables are considered. When it comes to assessing variables and thus making predictions that a judge alone would not have intuited, the number of variables is often too small to produce (predictively) significant results. Someone like Loomis may benefit from an algorithm that considered the category of a convicted person who was over 30 years of age, had a job to which he could return, and a long-term association with a romantic partner and church. These variables are not like three-run baseball victories, as it is easier to come up with theories about why they might help predict behavior after release from prison. But there are probably too few matches in the data, and this is surely the case if we throw in characteristics that might cause a released person to avoid re-incarceration, like someone who loves Fresca or another legal drink not available in prison. Only unsupervised learning would discover such a predictor[25]—and it is this ability that seems to attract law professors to artificial intelligence—but it is most unlikely to do so with any reliability when it comes to legal questions rather than stock exchanges, because the data set is insufficiently large. Statisticians like data sets with ten million observations to divide and study,[26]

---

[24] This is, for example, the intuition behind regularization techniques, such as "Lasso," or "least absolute shrinkage and selection operator." Variables with small, non-zero coefficients can be understood as correlations brought about by chance. Lasso, and other regularization techniques, shrink those coefficients to zero in order to amplify the algorithm's ability to identify true connections. There is, of course, some cost along with the benefit. *See* Robert Tibshirani, *Regression Shrinkage and Selection via the Lasso*, 58 J. Royal Stat. Soc. Series (B) Methodological 267, 267 (1996) (developing the Lasso technique).

[25] Repeated human investigation could discover such a predictor though it surely would suffer from overfit.

[26] See Ian Goodfellow et Al., *Deep Learning 2* (2016) ("As of 2016, a rough rule of

and such numbers will simply be unavailable in any legal application that comes to mind.

Returning now to the evaluation of an algorithm, we observed that the metric in law is more difficult to state than it is in finance; medicine probably falls in between. The existing and well-crafted literature on bail determinations asked whether the machine outperformed a judge, or an average judge, and it found that the machine wins,[27] though perhaps the judge could have been better equipped if offered the information regularly offered to the machine.[28] Still, it is likely that an algorithm would yield fewer recidivists, holding the percentage of released persons constant, but perhaps this is not a good measure of success, as some recidivists might commit more horrific crimes.[29] Three judges working quickly might by majority vote do better than one slow judge or one algorithm. We do not claim to have a solution to this problem of accurate measurement and fair competition, and so we simply proceed with the hope, or intuition, that Wisconsin can assess W's performance.

In the best of all worlds, as we will see, a state like Wisconsin might have selected W, over algorithms produced by competing firms, X and Y, because the state may have tested all these firms' algorithms and, taking cost into account, found W to be superior (however that was measured). As we have seen from our digression to the neater world of finance, the sophisticated way to do this would be for Wisconsin to divide available data, and give one portion to the applicants, and then test their algorithms on the withheld data.[30]

---

thumb is that a supervised deep learning algorithm will generally achieve acceptable performance with around 5,000 labeled examples *per category* and will match or exceed human performance when trained with a dataset containing at least 10 million labeled examples.") (emphasis added).

[27] *See* Kleinberg, et al., *supra* note 1.

[28] None of the experiments or simulations matched the algorithm against a judge equipped with past data on previous bail decisions, though the algorithm was given that information. On occasion a judge probably recognizes an accused person from an earlier case assigned to the same judge, but overall, judges would surely perform better if provided with a spreadsheet or table containing the characteristics of recidivists.

[29] While a bail algorithm can be specifically trained to predict the risk of violent crimes, the algorithm will identify fewer recidivists only if violent crime and recidivism are negatively correlated. *See id.* at 272-75. Our point is that the algorithm may reduce recidivism by a smaller amount (when compared to human judges working alone) in exchange for reducing violent crimes by a greater amount, and that the social value of the two quantities is difficult to compare.

[30] Even so, data scientists can invalidate their results by inadvertently "peeking" at the withheld data. Suppose the algorithm is finely adjusted to several configurations, which are then individually tested on the withheld data. If the analyst selects the configuration on the basis of the withheld data error rate, then information about the withheld data has inadvertently leaked into the testing algorithm. The best approach, in order to obtain an

This offers protection against overfitting.[31] If X can see all the available data, X can construct an algorithm that does very well—when not limited to a simple function but rather able to zig and zag. But this algorithm will do poorly with new data or actual cases.[32] This is true even for large data sets. Thus, it might happen by chance that a great majority of the recidivists had surnames containing five letters or more. X submits an algorithm that draws on this observation, that perhaps only an artificial intelligence notices, by sifting through thousands of observations. It is now most unlikely that this characteristic helps X to do well when its algorithm is tested against the withheld data—unless there is something "true" (perhaps because of a correlation with some other characteristic) about lengthy surnames. In any event, if W was selected because its algorithm (which presumably performed best for W, and better than the judges', on the proffered data and then also) performed best on the withheld data, Wisconsin can in good faith tell the court that W's prediction is at least as good as anything the court had previously used, and also that the definition of success or "goodness" is reasonable. Both are being evaluated with something like the number of released persons who committed a felony within five years, though over time more sophisticated measurement tools can be developed. Comparisons and the search for superiority must always involve some degree of guesswork because we cannot observe the inaccurate and unfortunate long sentences assigned by W and judges.[33] It is tempting to say that if Wisconsin has insufficient data to withhold for testing, it could simply test W, X, and Y on *future* data; it could reserve judgment while waiting for the behavior of released persons in the next year or two. This data is obviously and reliably unavailable to the state, as well as to W, X, and Y, and thus forms a nice set for testing. One problem with this approach is that one or two years will not provide enough data. Another is that it is unlikely that any judge or legislature will like the idea of telling prisoners that their terms of incarceration are unknown at present, but will be decided and revealed long in the future based on the behavior of other persons in what are likely to be somewhat different environments.

---

independent evaluation of the algorithm, is to lock away the test data until learning is completely finished. *See* Russel & Norvig, *supra* note 19 at 708-09.

[31] Note that this approach does not *eliminate* overfitting, dividing data simply helps the analyst identify instances where overfitting has occurred. *See id.* Algorithm development is similar to scientific falsification. The analyst cannot objectively "prove" that the algorithm does not suffer from overfit, but the algorithm can be understood as properly specified, or validated, until otherwise demonstrated. At its core, an algorithm is a hypothesis.

[32] *See id.* at 696 (noting the obvious, that models containing high-degree polynomials (what this Article calls zigs and zags) are more likely to overfit the data).

[33] We return to this important point in Part IV.

### *B. Algorithms for the Defense*

What if a clever defendant used the machine-learning plot against the plotter, hoisting it by its own petard? Imagine that Loomis says: "You relied on this invisible algorithm, W1, to categorize me as someone relatively likely to commit another crime after a short prison sentence. Give me all the data that was used to justify the outside expert, W's, prediction, and allow me to develop an algorithm that not only does better than a mere judge, but that also outperforms W1, which you used for sentencing me to six years in prison. If my superior algorithm, D1, suggests a shorter prison sentence for me, then we should throw out W1. We should either reward me by setting me free or, at the very least, reduce my sentence to that suggested by my superior algorithm, D1." This is an attractive argument, and something like it might also have been made about the way that success has been measured, but we have promised to set that problem aside. Readers might like to pause here and think about defendant, D's, suggestion. It is that the state's use of W1 might be biased against someone like D. This is especially so because no one gets to see W's algorithm, as it is being withheld as a trade secret.[34] D wants the opportunity to develop an algorithm that does a better job than W1, and of course that suggests less prison time for D. To do this, D says he needs all the data that went into the formation of W1.

In order to evaluate this idea, it is important to think about the procedures available to defendants, as well as the best practices for developing predictive algorithms. Defendant can be accidently or willfully strategic. Just as defendant can shop among lie detector services and psychiatric evaluations, defendant, if well funded by an organization or other sources, can shop around for the most favorable algorithm—without the prosecutor knowing about this shopping spree and revealing it to the judge. This strategy works if the defendant is given access to the withheld data. The problem is exacerbated by the fact that the defendant argues (perhaps successfully) that he should be allowed to look at the withheld data in a way that W could not, assuming best practices were followed and data was withheld from W, before it was used to test W's algorithm.[35] Defendant can look at his own characteristics, like height, and try those out in constructing a favorable algorithm that is superior to W1 with respect to all the available data, and that then also benefits defendant. Defendant can try many variations, because he has the advantage

[34] Recall that this was a significant objection advanced by Loomis, who asserted that he was denied information considered by the court at sentencing. 881 N.W.2d at 786.

[35] This is also true if W was chosen over X and Y, as it might be if the state had the funds and inclination to compare statisticians during the first round of developing its algorithm.

of seeing all the data—if his "clever" suggestion to the court is accepted. With enough testing, defendant is likely to find some characteristic that he shares with non-recidivists—after testing on the withheld as well as the delivered data. Ironically, D may deploy variables that are yet more linked to race and other features that courts and legislatures will have sought to exclude from decision-making processes that influence sentencing. Courts may give D more leeway in this regard than it will give the state.

One obvious response by the government is, therefore, to explain to the court that defendant's advantage is too great if he is given access to both sets of data since defendant will be able to easily retrofit irregular characteristics to the government's algorithm. A defendant, like D, should also be required to produce an algorithm that defeats W1 when tested on withheld data; defendant should not "see" the withheld data. If defendant is given only the first set of data, he could still try to develop an algorithm that does reasonably well on this data set, and that also favors the defendant. Perhaps this is not as hard as it looks; defendant can probably replicate W1, and then add to it some characteristic that is so unique to defendant that it is unlikely to cause this algorithm, D2, to do any worse than W1. All defendant needs to add is some variable that is not found in recidivists in the data set offered to W and now in defendant's hands. Defendant might also have more leeway when it comes to variables that are highly correlated with race, to take the most important example, and defendant will use such variables only when they are to defendant's advantage.

The court might be impressed if defendant can do just as well as the outside expert. Moreover, it is possible that a court would allow defendant to defeat W1 by introducing variables that help D but that are not predictively significant.[36] This is not as far-fetched as it sounds. Imagine that D is very

[36] Technically, machine learning does not calculate statistical significance when choosing the relevant features and characteristics to include in an algorithm. Instead, model selection is the product of optimization and regularization. We could use the term "statistical significance" here for readers unfamiliar with the differences between machine learning and statistics, since that term tracks the basic intuition. Thus, if D strategically chooses a rare characteristic like home address of 4 letters, then D might defeat W, because W could never use this characteristic in its algorithm as it has either been optimized or regularized. Optimization generally involves excluding variables that provide no additional analytical power, that risk overfit, and that tax computational resources. If W has never seen D's rare characteristic, then it surely would not have been included in its algorithm. If W had seen it just a few times, it very likely would have excluded it in order to optimize model performance. See note 37 *infra*. The other likely and related possibility is that W had "regularized" its algorithm. If so, it would have removed variables that appeared irrelevant and that would likely have little impact on its model. On regularization, see *supra* note 24. In addition, specific variables can be scored in terms of "importance." These measures

old or had a significant knee injury while incarcerated. D will say: "Look, it is implausible that I will be a recidivist, as I am already 63.2 years old with a severely injured knee, and it is also impossible for me to show that age or injury is predictively significant in terms of your fear of my future criminal activity. You have no other person with these characteristics in any subset of your data." At a minimum this hypothetical shows that machine learning alone, with predictively significant data, cannot carry the day for the government.

Students of machine learning might suggest that the state can anticipate the D2 strategy by underfitting—in this case by offering less data or a loose model in the first step to both W and then to D. This will make it more difficult for D to insert a characteristic that survives competition with W1 when applied to the withheld data.[37] Put differently and more cynically, it suggests that W1 should be created so that it just barely defeats the human judge. One problem with this response is that courts may be unimpressed with a marginal improvement over human judges. Another is that D may still defeat or match W1 by offering a model that values an unusual characteristic. Finally, there may be political, statistical, or financial reasons to find W through a competition also involving X, Y, and other potential vendors; this competition will have produced an algorithm that outperforms the human judge but also offers D enough information to follow the strategy described here.

It is plausible that the government would be disappointed because a court would allow D's (first clever) argument to prevail, in part because it might think that to compensate for defendant's inability to see the details of W's process, defendant should have access to the data first withheld from W. In that case, a clever approach for the government is to divide the data in *three*. The first set is offered to W (and perhaps to X and Y),[38] and is then tested

---

generally report the mean decrease in accuracy when predicting training data when a given variable is excluded from the algorithm. See Gareth James et Al., *An Introduction to Statistical Learning with Applications in R* 330 (2013). Thus, if the mean accuracy of W with the home address letter-length variable is 50.71% and remains unchanged with that variable's inclusion, it likely would be excluded on the basis of unimportance.

[37] Ideally, W would be optimized with a technique that begins with the smallest, simplest model, which increases in size until it begins to overfit. This is the basic approach of algorithmic model selection. *See* Russel & Norvig, *supra* note 19 at 709. However, one can imagine the reverse approach. Begin with a large, complicated model, and decrease its size until it is sufficient for use, even if other, more accurate models are available. Note the importance of algorithmic competition for encouraging accuracy when parties construct algorithms strategically. Various methods of this kind might generate winners and losers in any competition among algorithms, an idea we advance in subpart D below.

[38] Again, this may apply to all the bidders if W competed with X and Y.

against the withheld second set, as before.[39] The defendant, D, is then free to see these two sets and construct its algorithm that will easily defeat W2.[40] But now W2 is compared to D2 by comparing their relative success on the third set of data, which has been withheld from both W and D. Again, success must be defined, but we have assumed that superiority is somehow successfully measured against the results produced by a human judge. This division-in-three seems like a good approach—and is consistent with requisite testing and validation for the inclusion of variables and the selection of model structure; without set-aside data, testing and validation is impossible. A hypothesis' strength is measured by subjecting it to repeated testing, not by developing it and then immediately accepting it as true. In the world of algorithms—which are in essence, complicated hypotheses that predict outcomes—data must be set aside or awaited, in order for testing to occur.[41] This is probably the case for all counterintuitive conclusions arising out of empirical work; in the absence of reliable priors, testing on previously unrevealed data is ideal. Note, however, that division-in-three, though a novel means of combatting over-fitting, still favors D, because this defendant has had access to more data and is thus likely to outperform W2 when both are tested on the third data set. On the other hand, it is barely possible that D will do worse because the variable it added, as something unique to it, might to D's surprise be found in several recidivists in the third set of data.[42]

---

[39] Note that these sets are a bit smaller as there has been a division in three rather than two, and so there is a greater chance of mis-specifying the model, and perhaps generating overfit if too few observations lead to a rigid algorithm. If the data sets are sufficently large, however, then the third part can actually help avoid overfit. Now, there is a training set, a second training set which can be used to "correct" for overfit by introducing yet more training data that was initially hidden from the algorithm, and a third, final testing set. Stephen Marsland *An Algorithmic Perspective* 20 (2nd ed. 2015).

[40] We refer to the algorithm that is based on the first set, but that has been tested on the second. Note that the text sets aside the question of whether D may only use statistically significant variables along with arguments about personal characteristics that appeal to the judge's common sense.

[41] Note the analytical similarity to "leave-some-out" or "k-fold cross-validation," in which the analyst divides the data into *k* number of groups, sets one group of the data aside, trains the algorithm with the remaining groups, and then tests the newly trained algorithm on the set-aside data. The process is repeated for each division *k* of the data. *Id.*; Russel & Norvig, *supra* note 19 at 708-709 (3d ed. 2010); Brett Lantz, *Machine Learning With R* 319 (2013). However, this process is carried out generating several testing data sets from the original in-sample data. Our suggestion is that, when there is sufficient data, algorithmic competitions should use out-of-sample data multiple times.

[42] Put differently, D has the advantage if victory is determined not by comparing R-squared results (which would require D to find statistically significant variables), but by success with the out-of-sample data.

The division-in-three strategy is vulnerable to the objection that soon other defendants will appear and they will either have access to all the data or we will need to divide the data into four and five and so forth, until the sets are so small as to be useless. In Section II.E below, we take up such questions of how to proceed following the first competition among algorithms. But here it is useful to note an interesting though perhaps minor complexity, as well as an objection the government might raise to the idea of competing algorithms: Defendant can "cheat." To see this, we can turn back to the more straightforward procedure in which the data is divided in two, and imagine that the court does not give the defendant the withheld, second set of, data. Knowing that there will be a test on the withheld data, D wishes he could have access to it in order to defeat W1 by successfully overfitting an attribute unique to D that results in a favorable prediction. Imagine that because of budgetary restrictions and in order to avoid another omitted variable, the data set is drawn entirely from Wisconsin.[43] Now the defendant can look at the first set of data, once offered to W and now offered to D, and then look for news reports of recidivists. D can then adjust its algorithm to "predict" recidivists based on the advantage of knowing and investigating characteristics of these known recidivists. Of course, D will exclude any characteristics that also point to himself. This strategy will work even if the identities of recidivists in the first set of data are not revealed, because by looking at a group of known recidivists, defendant can be virtually certain that some will be found in the withheld data. A cynic might say that W has probably done this as well, but professional ethics[44] might restrain W and, in any event, under some questioning this strategy can be discovered because there cannot possibly be a good trade-secret claim with regard to this information.

It is time to recap the most likely scenario. The state employs an expert W, and W produces an algorithm from a large set of data given to it by the state. W's goal is to do better than a judge. W's algorithm, W1, is then tested against withheld data containing cases that were already decided by judges along with several years of outcomes with respect to prison sentences of various length. W's algorithm is found to be superior to the typical judge's, and the algorithm is then offered to a court and used along with other

[43] Koepke & Robinson, *supra* note 2, document several examples. One involved Florida's Pretrial Risk Assessment, developed with 1,757 cases from January and March 2011. Another focused on Ohio's Pretrial Assessment Tool, developed with "over 1,800" cases from September 2006 to October 2007. Finally there is an example drawn from, Virginia's Pretrial Risk Assessment Instrument, developed with 1,971 cases from July 1, 1998 to June 30, 1999 and later revised with data from 2005. *Id. at* 1758.

[44] *See* Code of Conduct for Professional Data Scientists, Rule 5d, Oxford-Munich Code of Conduct (2018) http://www.code-of-ethics.org/code-of-conduct/.

information to sentence a defendant, D, like Loomis. D now argues that he can construct an algorithm superior to W1, and D's aim is to find such an algorithm that is also favorable to himself. In one scenario, the state will be required to give D all the available data, including that withheld from W in the first step. D is then likely to beat W1 because D has more information than did W, and D can retrofit, in a manner of speaking, in a way that now benefits D. It is not hard to imagine that a court will equip D with both sets of data. We suspect that most readers (as well as the authors found here) were at first attracted to the "clever argument" advanced by the defendant some seven paragraphs above, and so there is good reason to expect judges to do so as well. Finally, even if courts reject this argument, there are advantages that D might exploit simply by looking for self-serving characteristics of real recidivists who are reported and described in the news, and are likely to be found in withheld data.[45] This advantage of D is most plausible if D is able to include variables that are not predictively significant.[46]

Finally, it must be noted that law is not like finance or medicine—fields where data scientists have devoted much attention. Data scientists are likely to respond to much of the discussion here by insisting that retrofitting must simply not be allowed. These scientists regularly divide data to test hypotheses, but the excluded data must be invisible to the designer of an algorithm. Here, on the other hand, we have put forward clever or seductive arguments that defendants might make, and then suggested ways to offset the strategic algorithms, or simply hypotheses, that defendants and other second movers might advance. Why not simply ban retrofitting or withhold all data from defendants—or disappointed college applicants, as discussed presently (and as a first step in extending our analysis to areas far removed from criminal sentencing), and other parties that do not like the results produced by data science? The answer lies in the very nature of the adversarial system. It is based on each side developing arguments that defeat or out-maneuver opponents. It is unrealistic to think that courts will deny defendants the opportunity to compete with the state's algorithmic decision-making—or even its old-fashioned hypotheses backed up by several observations. The adversarial method might be thought of as constitutionally required or as a kind of competition that is the American way; either way it is presently

[45] This is not to say that self-serving characteristics could be inserted into the algorithm without a plausible theory underlying their predictive signifigance. *See infra* Section II.D.

[46] By "not predictively significant" we mean that such variables lack predictive power for W, but are, nonetheless, unique to D. *See supra* note 36. It is apparent that much depends on this last assumption, as discussed presently. D's task is otherwise difficult inasmuch as W has presumably identified the correct model. D may do best by taking his variable that is not significant and offering it to the court, as encouraged by *Loomis*, as something representing individualized justice, that can be added to the logarithmic input.

embedded in our legal system and not likely to be pushed aside to conform to the habits of present-day data scientists.

Even so, data scientists should be comfortable with our proposal since algorithmic competition can broaden data collection effort and sharpen predictive models.[47] Retrofitting is not a problem if the state insists on data division and disallows statistically or predictively unimportant variables. If D evaluates the available data and puts forward a feature unique to D, it will either be discarded as insignificant when tested against the withheld data, or shown to be useful for predicting crime. D's efforts, even if characterized by intentional retrofitting, are either disqualified or harnessed by W. Retrofitting is simply not possible when data is divided and set aside for demanding tests.[48] If, on the other hand, the state permits D to submit predictively insignificant variables, then, as we discuss presently, they should at least be accompanied by plausible theories.

### *C. Over-ruling Prediction with Causality or New Information*

A very old D, or one who receives a significant knee injury while incarcerated, has the better of the argument when pitted against W. W's algorithm may predict that D will recidivate on the basis of his social and criminal factors, but it will fail to account for the defendant's physical disadvantages, now emphasized by D, and perhaps unknown to W. D will now simply fall into the small set of defendants that W misclassified. Machine-learning enthusiasts might suggest that the state can under-fit, as we suggested above, or perhaps W can simply add an additional variable that accounts for D's impediment. Over time, sufficiently large recidivism data on defendants like D will eliminate the classification error. If either of these arguments appears doubtful to a court that is considering the use of W (or X or Y), it is because the misclassification error can be immediately addressed with the input of a human judge. The judge does not require a predictive model fed with sufficiently large data in order to discern that D, who is incapacitated, will not recidivate. Put differently, the human judge, unlike state-of-art machine learning, is equipped with causal reasoning.[49] In short,

---

[47] *See infra* Section II.E.

[48] There remains the possibility that D's variable might be erroneously identified as predictively important when tested against the withheld data, but not because D has retrofitted the algorithm. Indeed, W would have made the error on its own had W1 discovered the erroneous variable first. The error is a straightforward instance of overfit.

[49] *See* Ryan Copus et al., *Big Data, Machine Learning, and the Credibility Revolution in Legal Studies* in Michael A. Livermore & Daniel N. Rockmore, *Law as Data: Computation, Text, and the Future of Legal Analytics* 56 (2019) (describing differences between machine learning and predictive inferences on the one hand, versus statistics and

one solution to the problem presented by D's particularity is to allow D to seek human help in overruling the winning algorithm. Presumably, a judge can also overrule the algorithm in the government's favor. We suspect that this is the path law will take in the near future.[50]

Another solution is to allow D to take advantage of his personal knowledge and ask the court to require W to develop a new algorithm, W3, that incorporates D's new information. This is not simply a matter of retrofitting. Instead, D might say something like: "Look, now that I see my own situation, it occurs to me that there might be a sufficient number of other people who are old or who suffered a knee injury, and now I see that this variable should have been included in your original search in the data for the algorithm that best predicts recidivism. You may or may not have collected such data, but the burden should be on you to include this data or explain why it is unavailable. Why don't you explore the data again, with information about knees (or kinds of friends made in prison or church attendance), and see if you can find a W3 that defeats your own W2 or W1." The government might resist this argument unless there is a requirement that the new variable meet a certain level of significance; in other words, D cannot simply point to an unusual characteristic he shares with non-recidivists, but D must name a variable that proves to be significant (and available) in the large pool of data available to W.[51]

Finally, a third possibility is that the government is allowed to (or simply be permitted as it likes) to look at D's individual characteristics and then on its own suggest new variables in order to create a W4 that will make D worse off than before. This is analogous to the familiar claim by teachers that if a student asks for an exam to be re-graded, the student should be aware that the grade might be lowered because of the re-examination of the "data" that the student believed could only raise the assigned grade.

There is much to be said in favor of the second response described here but, as previously suggested, we think that this sort of competition among algorithms is unlikely in the near future. Experienced data scientists will

---

causal inferences on the other, within the context of legal research); *see also* Judea Pearl, *Theoretical Impediments to Machine Learning With Seven Sparks from the Causal Revolution*, arXiv: 1801.04016v1 (2018) (noting that machine learning remains unequipped with the tools of causal reasoning, but that advances in graphical and structural models could make counterfactuals computationally tractable, and that causal reasoning with machines could therefore be on the horizon).

[50] *See* Fagan & Levmore, *supra* note 4 (describing this intuition in detail).

[51] Thus, D might ask the court to consider a model of causal, as opposed to predictive, inference.

object to any attempt to improve with hindsight on the original winning algorithm, while lawyers and most citizens will surely prefer a version of the adversarial system with a judge deciding when to overrule the winning algorithm submitted by the government.

The important idea here is that all algorithms, and especially those designed to be predictive, have some problems. There is the problem of defining success, and then that of taking changed circumstances into account. In general, statisticians compare how much of the variation observed in the outcome variable (recidivism) is explained by the independent variables (social and criminal histories, gang tattoo, and so on). The more a model can accurately explain variation in something like observed recidivism outcomes, the "better."[52] Most of us, and certainly judges, are drawn to variables that are accompanied by obvious and intuitive causal relationships such as smoking with lung cancer. Even if other personal factors and independent variables diverge, smoking can explain a great deal of the variance in lung cancer outcomes. Other factors, such as a 20-year career as a professional coal miner, may be less obvious because they can only explain a small amount of the variance in cancer outcomes over a population that represents many professions. Nevertheless, the coal miner variable may be significant and carry a large coefficient. Comparing causal models with and without the coal miner variable can be accomplished with various diagnostic tests of measurement error.[53] Comparison strategies are enhanced with division of data and out-of-sample testing. In the end, statistics is both an art and a

[52] As mentioned above, models that generate predictive inferences can rely on optimization, regularization, and measures of importance to evaluate the usefulness of predictive variables. *See supra* note 30. In contrast, students of statistics may recall that the amount of variation in causal outcomes explained by the inputs is referred to as R-squared. *See* Jeffrey M. Woolridge, *Introductory Econometrics: A Modern Approach* 200-01 (5th ed. 2013) (providing the definition and noting that inferences are difficult to make in the presence of low R-squared values since most of the variation is the result of unknown factors). However, greater R-squared does not necessarily mean a "better" model. When data units are transformed, say for seasonal adjustments, deflating, logging, or differencing, comparisons of R-squared are inaccurate. In particular, time series data tend to inflate values of R-squared. *See* Robert Nau, *What's a Good Value for R-Squared*, Statistical Forecasting: Notes on Regression and Times Series Data (accessed Oct. 28, 2019), https://people.duke.edu/~rnau/rsquared.htm. For this reason, data should be divided and out-of-sample model performance should be compared.

[53] For a good example of how health statisticians might approach the cancer-coal miner hypothetical and employ rigorous diagnostics, *see* Louis Anthony Cox, Jr., *Quantifying and Reducing Uncertainty About Causality in Improving Public Health and Safety*, 1490, 1492-93, Handbook of Uncertainty Quantification (2017). Similar methods might be used by a judge or D to check the causal claims produced by W1, or W's competitiors.

science, and no single method is perfect or best for all situations.[54] Both causal and predictive models are apt to benefit from competition.

### *D. Simple and Dramatic Reforms with Competing Algorithms*

We have come far enough to offer various solutions to the problems described thus far, before moving away from criminal sentences to other applications. Our suggestions overcome the several advantages enjoyed by D, if D is well advised by statisticians and is facing a sympathetic or intuitively-minded judge. First, law should not admit evidence from D when D uses an algorithm that defeats the state purely on the basis of the state's data; it should reject what we have called defendant's "clever argument." However, if the state wants to submit algorithmically created evidence, it should be required to conduct a competition. One obvious idea for extracting the benefit of competitive algorithms is for the state to spend resources on W, X, and Y, and then use the algorithm that wins when applied to the withheld data. The greater the number of competitors, the less likely will it be that D can defeat the winner without retrofitting from pieces of the withheld data. D can use information about himself, but D must defeat W, X, and Y, all of whom will be judged on their performance with the withheld data.

A more interesting and more democratic approach is not to hire W in the first place, but to offer the first set of data to the public, and for the state to offer a reward to the person or entity that produces the best algorithm when tested against withheld data. This idea is borrowed from *Kaggle*, a website the reader might wish to examine.[55] It is used in the real world of competing algorithms. Instead of committing resources to W, and forcing defendants to find resources in order to create their own algorithms, the state would offer this money as a reward to the public winner. D is of course free to join in this competition, and to insert characteristics that are friendly to D, but these insertions are quite unlikely to help. D might even advertise his own characteristics, with the hope that some participant in the competition, perhaps part of a law school clinical program, will be inclined to favor the

[54] Robert Nau, *What's the Bottom Line? How to Compare Models*, Statistical Forecasting: Notes on Regression and Times Series Data (accessed Oct. 28, 2019), https://people.duke.edu/~rnau/compare.htm (noting there is no absolute criterion for "good" values of statistical diagnostic tests).

[55] Interested readers should visit https://www.kaggle.com/, and in particular its list of competitions at https://www.kaggle.com/competitions. For example, at the time this Article is being written, there are more than 15 open competitions, 7 with monetary prizes, while 344 have been completed in the past. The leading competition in terms of prize size is the 2019 Data Science Bowl, in which the objective is to "uncover factors to help measure how young children learn."

impoverished under-dog and try out these characteristics in the hope of winning. An interesting twist on this idea is to allow D not only to offer information about himself to the public, but also to offer his own prize for an algorithm that does best when applied to the data withheld by the state. D, and any sympathetic supporters, hope that the competition suggested here will benefit D. D has some advantage here, because the state is not looking for a winner that suggests the longest prison sentence. Still, even if D is well funded, it is unlikely that more competitors or multiple competitions reach very different results as far as D is concerned. D (and we) might argue that these competitions are in the spirit of the beyond-a-reasonable-doubt standard.

We do not claim that our competing-algorithm idea is perfect. Again, competitors can look for reports in the news or other public records in order to gain advantage by over-fitting, and find recidivists that might be in the withheld data. It is possible that this severe form of over-fitting can be offset by offering less data in the first step, and withholding more. Put differently, some intentional under-fitting might offset the strategic over-fitting. Our claim is that this method is superior to that followed in present-day Wisconsin, or by judges working without the benefit of machine learning.

Indeed, the major alternative to our suggestion is not really an alternative, but something that can be added to it. It is also something that the Wisconsin Supreme Court (no doubt) unknowingly invited when it said that the state could use an algorithm, like W1, so long as it was not the only factor considered by the lower court judge in the sentencing decision.[56] The idea is that any variable D wants to add to the previously chosen algorithm must come with some theory that appeals to the judge. If D adds the number of symbols in his address to improve on the performance of W1, the judge can reject D2 because it is almost surely an example of over-fitting, and what we might think of as retrofitting. But if D is able to provide an attractive theory—and even one that D constructs after forging an algorithm based on all the available data (and then designed to be friendly to characteristics that D knows of himself)—then the court can accept D's clever argument, because it is backed up by a plausible theory. For example, D might introduce an algorithm that includes as a variable defendant's attendance at church services, or care for his children. These variables would benefit D. A judge might find that this variable—however over-fitted and retrofitted—makes sense because the judge finds it plausible that such a characteristic might prevent recidivism and the defendant's fear of re-incarceration. Our intuition

[56] *See* 881 N.W.2d 749 at 769.

is to prefer public competition (including D) for a winning algorithm, as the title of this Article suggests. On the other hand, *Loomis* itself suggests the last approach. We recognize, however, that in the near future courts might be disinclined to engage in such dramatic use of the new tool of machine learning. If so, they are more likely to accept the first solution, that the state should simply show that it conducted some competition among algorithm makers. Even this would be a significant step forward.

### *E. Post-Competition Practices*

Realistically, current practices as well as the ideas advanced in this Article must confront the question of what to do when new defendants come on the scene. If D, or a competition open to the public, has succeeded in defeating the state's algorithm, or simply constituted the method of producing the winning algorithm, then a new defendant, whom we might call 2D, can try to defeat the prevailing algorithm. Imagine the case where the state produces an algorithm and then D defeats it on data withheld from both D and the state, but presumably with an entry that benefits D. D's algorithm is now the one that 2D must defeat, for there is no reason to continue to accept the state's algorithm, as it is now second-best. Indeed, the state itself should use D's algorithm until it is defeated either by an updated algorithm solicited by the state, or by an improved algorithm brought in by 2D or some later defendant, 3D.

It might seem at first that ongoing improved algorithms are unlikely, but this ignores the fact that 2D might have the advantage of access to new data that come into play between D and 2D's time. If the state gathers this data, most of it can be withheld from 2D, though again 2D will have some information available from newspaper reports and perhaps other public sources or other defense attorneys. More realistically, 2D has the advantage of changed circumstances and can argue that sentencing should be evaluated on the basis of the most recent data since, after all, up-to-date data should more accurately predict recidivism. The world has changed since the earlier, winning algorithm came into being, and this benefits 2D as well as competitors in any new public competition. For example, the state may have installed ankle bracelets to keep track of released prisoners, and this may have changed the likelihood and distribution of recidivism. It may have installed cameras or simply improved its policing strategies. The new winning algorithm is now likely, in circumstances where recidivism is responsive to new advances in technology (and changing environments generally), to weight observations in favor of recently acquired data. The state might argue that set-aside data in the lock box continues to serve as valid testing ground, and it will be difficult

for 2D to defeat the previous winner on this proving ground, but it is hard to see, as a matter of due process or as a matter of data science, why the old withheld data should be used forever. If the state is in control of the competition among algorithms, law will need to develop some reasonable rule about how often new competitions ought to be funded. This is not a simple matter. If data between 2020 and 2021 is favorable to D, he will prefer that the entire data-set be based on 2020-2021 data. However, there will be features of the older data that continue to help predict D's recidivism. For example, variables like height and weight may predict outcomes between 2010 through 2021 with stable accuracy, but other features like D's unique family history may be sensitive to social change, and predict with great accuracy only when applied to new data.

It is easy to imagine that D and his competitor, in this example the state, will argue over the rate of environmental change. The state will be more likely to assert that change is slow so as to preserve the value of its set aside data. D will certainly be advantaged if he successfully argues that the lock-box data is no longer representative of the current environment, from which he can distinguish himself.[57] Predictive analysis is at a disadvantage in fast-moving environments where the past provides little insight about the future. In those instances, law does better with human-machine combinations.[58] A sympathetic judge who evaluates D's assertions can demand that the competition use new data. If new data is unavailable, the judge can overrule prediction on the basis of D's arguments as we have discussed above in Section II.C. Of course, D's case can be enhanced if he presents a causal model indicating that his unique family history is associated with non-recidivism.

Apart from human-machine combinations that rely on the human judge, there are other possibilities worth exploring, especially if environmental change is relatively slow. Let us assert, for example, that although 2D is able to give greater weight to recent data in constructing a new algorithm, the state's set-

[57] For instance, California Penal Code § 3041.5(b)(3) provides that inmates found unsuitable for parole by the Californian Board of Parole Hearings receive a subsequent hearing either three, five, seven, ten, or fifteen years later, but inmates may request new advanced hearings on the basis of a "change in circumstances or new information." Research has shown that of all the factors the Board considers when making a parole decision, it gives the greatest weight to psychiatric risk assessments. Hannah Laqueur & Anna Venancio, *A Computational Analysis of California Parole Suitability Hearings* in Michael A. Livermore & Daniel N. Rockmore (Eds.), Law as Data 207-08 (2019). Those evaluations are valid for five years, but inmates must be evaluated anew if a petition to advance parole is granted. *Id.* at 202.

[58] *See* Fagan & Levmore, *supra* note 4 at 6.

aside testing data need not be updated until there is a large amount of new data. Every new defendant cannot demand a newly created set for testing, but once there are many observations it will become reasonable to require the state to update the set aside data, inasmuch as circumstances will really have changed–and at some point there is enough new data to justify a new set.

Once time (and successive algorithms) is brought into the analysis, there is the question of whether D can ask that his prison sentence be recalculated based on the new information. The claim is a version of the familiar question of when law, and especially criminal and constitutional law, ought to be retroactive. It also raises the question of whether, in the quest for clean, unseen, divided data, law might simply say that algorithms will be tested against *future* data. The first suggestion is more easily set aside; it is like the student who wants an exam regraded but hesitates because the professor might find that the exam grade should in fact be lowered. If the possibilities are asymmetrical, there is the danger of too many requests. Similarly, if prison sentences are updated, then they ought to be done so for all defendants, and courts will never want to do this in a way that surprises incarcerated persons with longer sentences. If so, then it is statistically incorrect to update for the few who desire it.

But what about the harder question and more attractive idea of using future data instead of dividing past data, and hoping that the set-aside batch is unavailable to the algorithm makers? As a practical matter, it is hard to imagine courts telling the accused that their sentences are presently unknown. Perhaps, at the time of trials, sentences could be adjusted upward, to the high end of the sentencing guideline scale, and then some years later reduced according to an algorithm based on future data that becomes available in the period following the earlier trial and the development of the winning algorithm.[59] Another possibility is to divide the existing data in three, test the first defendant's algorithm on the withheld data, while binding defendant's lawyer to secrecy. No information can be revealed to other lawyers or parties, so that all future defendants' algorithms can be tested on the same withheld data. This sort of nondisclosure requirement might require statutory assistance, and might also require a rule that the same lawyer (and hired expert) cannot be involved in a future case with similar characteristics. The problems may seem unmanageable, but note that if future defendants appear

[59] Unsurprisingly, there is already evidence that judges use machine generated predictions to reduce sentences more often than they do to increase them. See Megan T. Stevenson & Jennifer L. Doleac, *Algorithmic Risk Assessment in the Hands of Humans*, (November 18, 2019). Available at SSRN: https://ssrn.com/abstract=3489440 or http://dx.doi.org/10.2139/ssrn.3489440.

after some time passes, the problem is mitigated quite naturally, as new data will be available for testing.

Finally, assuming there is not enough data to divide it into very many parts, there remains the idea of using randomly selected data from the past, though we have already argued that this data has already been included in the set that produced the earlier algorithm, and is therefore tainted in an important way. Perhaps data used by defendants can be excluded from the sample, much the same way populations are sampled without replacement, but the complications are real and potentially open the door for unwanted biases and strategic behavior.

It is plain that the right choice among these options depends on the amount of available data, the expected number of future defendants,[60] and other factors. We do not claim to know the right answer, as our goal is to draw attention to the strengths as well as the problems associated with algorithms in legal settings. If the emphasis here is on the problem of–and solutions to–overfitting, then it is probably useful to begin with the idea of dividing data in three. The leading alternative is to hope that courts have the patience to use future data as the "set aside" data, along with the idea of handing out larger prison sentences with the expectation that many will later be reduced. As explained, this potential reduction responds to the possibility that because of changed circumstances, new data (previously unavailable and thus hidden at the time of the initial sentencing) will bring about new algorithms.

## III. COMPETING ALGORITHMS FOR PRIVATE DECISIONS WITH LIMITED DATA

### *A. University Admissions and Predictive Variables*

We now turn our attention to questions of algorithmic allocation of scarce resources and what to do when algorithms are improved with statistically unimpressive variables. We recognize that the data sets involved here are relatively small, and what we call algorithms might be little more than hypotheses that are conventionally tested. Still, there is a great deal in common among these examples; our title refers to the idea of competing algorithms, but the larger subject is how to deal with retrofitting, whether one

[60] We should note that there is something to be said for limiting the observations to recidivism by "similar" defendants. It is unlikely that those incarcerated for drunk driving, for example, will has the same proclivity for recidivism of any kind as will those incarcerated for armed robbery. But this feature is presumably one of the characteristics found in the winning algorithms. If it is significant, then it further reduces the size of the relevant data pool.

has a very large number of observations or fewer observations with which to test or advance an hypothesis. Some of the examples discussed presently can be described in terms of sorting, but the high stakes that they often present, also lend themselves to competition and to nuances that are now familiar. Consider examples like a university setting aside a fixed number of seats for its incoming class; a law firm or consultancy that can only promote so many of its associates to partner; a business firm that can only hire a limited number of employees or appoint a handful of directors to its board; and a foundation that distributes a limited number of grants or awards. In each of these cases, the evaluator wishes to make the best possible match between applicants and scarce resources, and it can rely on a combination of human judgment and algorithmic input to do so. What has been done by humans is again likely to improve with some input from, or delegation to, machines. By now it should be plain that a human and a machine are apt to do better than either working alone.[61] In all these cases a human must decide on the performance goals, including the desired mix of winners across several dimensions. An algorithm might be excellent at determining the weight that ought to be given to SAT scores, but it needs to be told what the SAT scores are trying to predict.

It is easy to picture a university that has developed an algorithm for admitting students. The algorithm might adjust as it goes along in order to diversify a class, and it might learn from the performance of students admitted in the past. As its classes are diversified it will have access to more data about performance. But now imagine that an eager applicant wants to offer a competing algorithm, without knowing the details of the target university's own algorithm. In turn, an applicant (or a group of applicants who feel under-represented) could be expected (with the assistance of professional admissions counselors no doubt) to produce a competing algorithm that favors the personal characteristics of these applicants while simultaneously meeting some threshold qualifications sought by evaluators, normally at the instruction of deans or other university officials. This, after all, is the approach applicants currently take; they craft personal statements, experiences, and interview strategies in order to appeal to the needs of the university. Persuasion is both a science and an art. The deployment of algorithms for allocating scarce resources can be understood as an attempt to suppress persuasive artistry and elevate bureaucratic reasoning in decisions but, as we have seen, algorithmic manipulation can be accomplished by carefully retrofitting data.[62]

[61] *See* Fagan & Levmore, *supra* note 4 (describing this intuition in detail).
[62] *Supra* Section II.

The problem, or perhaps it is best called a reality, is presented even in the simplest cases where the university has quite straightforward goals that are revealed to all, much as sentencing decisions might be based on predictions about recidivism and the seriousness of later crimes. Imagine that the university simply wants to admit students who will perform well in university courses and be admitted to the most selective graduate programs. It develops a straightforward admissions algorithm to predict academic success, and it may or may not differentiate itself from its competitors who also seek high performers. The winning algorithm, developed by dividing sufficiently big data as discussed earlier, gives weight to high school grades depending on the prior performance of applicants from given high schools, and also uses ACT and SAT scores. It might also take account of performance in spelling bees and math contests. It is easy to see that an applicant, who observes the university's admissions and rejections, can reason backwards and come close to guessing the university's algorithm. Now the applicant offers a competing algorithm that makes use of some feature of her own that is also present in observable successful students at the targeted university. For example, the applicant may improve on the reigning algorithm by adding in membership on a high school debate team, and this retrofitted algorithm favors our clever applicant. Note that the variable added by the applicant's algorithm is more appealing (to humans but not necessarily to machines) than is something like the number of letters in the applicant's mother's first name or a parent's employment status; it is easy to believe a story about how the debate team (or parent's) experience is predictive of academic success. As before, the applicant's cleverness is easy to over-rate. Had the data been divided into three, so that the applicant could not see it all before constructing her algorithm, it is far less likely that she could develop a superior algorithm. But the more important point here is that there is room for competition among algorithms. If the university is able to state its goals, then competition among algorithms holds great promise.

Put differently, it is common for universities, and certainly law schools, to be offered information by testing companies, For example, a law school learns how much weight to assign to college grades and the LSAT in order to best predict performance in that law school's first-year courses. The best weights can vary among undergraduate schools. It is apparent tht if the law school admits very few students from one undergraduate institution, the performance of these students in the first year offers less information than if the sample were larger. Similarly, the law school might like to know whether it should admit more physics majors from that undergraduate institution, but as it seeks to add variables with few observations, its statistical conclusions become shakier. This tendency suggests the importance of thinking about the

difference, if any, between retrofitting where there is a large pool of data and where there are but few observations. How many observations are needed before it is sensible for a law school to say "With this LSAT score and an excellent recommendation from Professor Eze, we ought to admit the applicant, because that undergraduate professor has sent us excellent students in the past?" The question is in some ways identical to, but in other ways easily distinguished from, the ability of a defendant in Wisconsin to offer a competing algorithm, and for this reason we leave it for another day.[63]

In any event, one reason for focusing on the case of university admissions, is that it is likely to be one where the "court" (in this case the university) is unable or unwilling to state its goals. A university wants much more than high grades on its exams. It wants some diversity, better sports teams, future donors, and leadership within the student body.[64] Every university would like to produce future American presidents and Nobel Prize winners. The more it is open about how it weights these goals (if that is even possible), the easier it will be for applicants to retrofit. An applicant can point to a particular life experience that she shares with some successful students or alumni. Again, a great deal depends on whether we demand that features be predictively significant. Note that the retrofitting is not necessarily a problem. In our earlier case of prison sentences, it was easier to state the goal in terms of recidivism. In the case of university admissions, retrofitting may lead to a different group of admitted students, but it will be very hard to say that this group was inferior to that which would have been admitted by the university's earlier algorithm. The more complicated the goal, the harder it is to fault retrofitting. On the other hand, the more the goal is defined by a weighted set of multiple factors, the harder it might be for the retrofitter to overcome the defensive, and statistically powerful, case for setting aside data and then requiring independent testing and validation of the retrofitter's algorithm.

---

[63] For now, it may be helpful for the reader to first distinguish between predictive versus causal models. Predictive models generally require 5,000 labeled examples *per category* of an outcome, as well as all of the predictive variables that accompany those outcomes, in order to provide acceptable levels of accuracy. *See* Goodfellow, et al., *supra* note 26. Causal models may provide insights with as few as 100 observations, but may not hold up well when sample sizes are increased, significance thresholds are reduced, or when the effect of the variable in question is small. *See, e.g.*, Jill E. Fisch, Jonah B. Gelbach & Jonathan Klick, *The Logic and Limits of Event Studies in Securities Fraud Litigation*, 96 Texas L. Rev. 552, 618 (2018) (describing the circumstances of statistical power within the context of securities fraud litigation). For instance, if the causal effect of Professor Eze's recommendation letter is observable in 6 out of 100 students, and only increases student performance in contracts, then the variable is likely irrelevant, even if initially found statistically significant because of the impact of the contracts grade on the students' first year grade point averages.

[64] *See supra* note 8 and accompanying text.

It may seem apparent that our applicant would be eager to see the university's algorithm and all of its data in order to successfully retrofit her candidacy. It is also tempting to make the argument that the candidate *should* be able to access the algorithm and data in order to legitimately plead her case, especially if the university only permits predictively useful variables. After all, if the applicant observes that a number of students who had worked at a car wash during high school performed extremely well at the university, then this observation could benefit the applicant and strengthen the university's algorithm at once. But this approach would be a mistake. Data must be set aside as a precondition for validation. Once again, testing an algorithm on withheld data permits the analyst to identify and discard unimportant variables and retain important ones. This critical step in the algorithm-building process reduces the likelihood of overfit and lends credibility to the final predictive model.[65] It is impossible to credibly validate any variable as predictively important without testing. Withholding data from the applicant facilitates testing, and permits her to legitimately assert that the car-wash variable matters. An applicant who is knowledgeable in data science, and confident in the car-wash variable, would demand that data be withheld so that she could convincingly validate that variable. If not, her reliance on car-wash experience is merely an unsubstantiated hypothesis made attractive for a moment thgrough retrofitting and perhaps some claim about how working at a car wash teaches things that prove useful in the eyes of a university administrator.

The argument here has become complex and open to a variety of objections—but also to improvement. As a practical matter, universities might respond to the advantages and problems associated with competing algorithms and data division by announcing that it will admit 60% of its class on the basis of academic performance in various courses and standardized exam results, and that these admissions decisions will be made by the winning algorithm. Interest groups and other observers might battle over the 60% number, but it is plausible that this will simply cause other universities to be secret about the percent of admissions decisions delegated to winning algorithms. The balance of the admitted class will be admitted by humans who can be expected to experiment with a variety of goals.

It is apparent that more thought should be given to the value of algorithmic features that do not meet the usual requirements for predictive importance, or its causal analog with which most readers, and certainly legal academics, are familiar: statistical significance. In some cases, statistical insignificance is

[65] *See supra* note 41 and accompanying text.

not troubling. Think of a patron who has one bad meal in Restaurant A on Tuesday, and then one delicious experience in Restaurant B on Wednesday. The patron is then asked to choose the best location for a birthday dinner to be held the next month. It is perfectly rational for the patron to choose B, even though he has just one observation to call upon. If an algorithm were developed to predict good restaurant meals, the single observation would be of little use if statistical significance were required. On the other hand, most people intuit a Bayesian approach to the problem of choosing among restaurants.[66] They might well choose Restaurant C over A, because no information is more promising than the single negative experience in A. The observer has no prior, let us say, and then the single positive experience in B updates the observer. It seems silly to say that the statistical insignificance means that there is no reason to expect A to be inferior to B or C. Note that if the birthday dinner is to be held on a Wednesday, it would be almost laughable for the observer to say that both C and A are fine choices because the data suggests that restaurants are disappointing on Tuesdays. A clever retrofitter, such as the owner of A who is eager for more business, might insist that the day of the week is the key variable, but common sense or priors push the observer to value the named restaurant much more than the day of the week.[67]

The restaurant example, and the importance of priors returns us to the idea developed in Part II, that some theory can go a long way in accepting or rejecting retrofitted algorithms, and especially so with limited data and no opportunity to divide data. It is not mysterious that we value the statistically "insignificant" observation of one bad meal, while rejecting the retrofitted algorithm that includes information about the number of letters in the applicant's mother's last name.

### *B. Employment Decisions*

[66] Jame Joyce, *Bayes' Theorem*, The Stanford Encyclopedia of Philosophy (Spring 2019 Edition) archived at https://plato.stanford.edu/archives/spr2019/entries/bayes-theorem/ ("A hypothesis is confirmed by any body of data that its truth renders probable").

[67] Note that the singular addition of Tuesdays or Wednesdays can be chosen from many features such as time of day, whether it is sunny, the color of the restaurant floor, and other variables with no apparent connection to food quality. The inclusion of just one of many "unapparent" variables by the owner of A should be troubling as it suggests a coincidental relationship (overfit) as well as strategic variable selection to reach a desired outcome (retrofit). With a data set containing millions of observations, however, unanticipated relationships can be revealed, but then inspected for plausibility and, finally but crucially, tested on withheld data.

At first blush, employment decisions seem like a poor area for competing algorithms, because the data set for a given employer or even for a given job description in an industry is likely to be even smaller than that available to most universities, and thus of relatively limited use. The advantage of data science, and machine learning in particular, is its ability to find connections across large data sets, along with the ability to ignore or even disprove conventional stereotypes. A human is far more likely to rely on predictively insignificant variables; these may come with theories, but the theories are developed ex post. The human might say something like "the best two CEOs I have observed during my career had law degrees so we should hire a CEO with a law degree." Another human, or board member, might say "Let us be careful about how we define best. If we look at the rate of return on assets and correct for risk—and also compare the results with those earned by other firms in our industry—we get a better measure than just looking at the increase in our stock's price." Neither of these approaches requires data science, and it is arguable that the best CEO is someone who will work well alongside the company's existing employees in the state in which it conducts business. Data science is nearly useless here because there are few observations about CEO performance alongside many plausible factors. We do not have many observations, and competing algorithms are unlikely to improve decision-making.

Every student who searches for a good teacher for her particular learning style, and every faculty member looking for a new dean or colleague, is aware of this problem. A decent theory with an observation or two is likely to be more useful than a conventional data scientist or victor in a public competition among algorithm makers. Another way to make this point is to see that dividing data is normally of little help in making employment decisions. We would ridicule a committee member who said: "We agree on who have been our two most successful faculty members over the past five years, and I note that both went to Columbia, both were 30 years old when hired, and both vacationed last summer in France. I will vote for candidates who possess these characteristics, and there is no point in interviewing anyone who does not." This is pure retrofitting, even though it is possible to construct theories about the value of these inputs. The usual means of testing the retrofit is unavailable. The same is true for most large organizations. The fact that diversity or teamwork is important adds to the skepticism about using data that an algorithm maker might discover. Moreover, it would be laughable if an applicant observed that the organization favored Columbia graduates, and then said at an interview: "I can discern your hiring pattern, but actually the last three Nobel-Prize-winning faculty members you had were all born on September the 14th; no other faculty members were born on

that date; and I too was born on the 14th of September." The statistical observation would hardly be helped with some theory about why September babies were especially talented.

It is tempting to say that these examples show that predictive significance is important. Perhaps it is important for finding a lemon in a haystack, as when excluding the bad restaurant discussed in the previous section, since there may be many observations to support this finding, but not so useful for finding a treasure, or needle, in the haystack, since there will likely be very few observations. The characteristics of two faculty members or two CEOs do not provide much useful information. On the other hand, perhaps the right question is not whether hiring can be done well with statistical methods and competing algorithms, but rather whether the algorithmic approach is superior to the familiar one that convinces some hiring committee chair-persons that they are good at identifying talent. Humans tend to use home-made algorithms, or simply hypotheses, and these are not tested for predictive significance.

Despite all this skepticism, there is room for competing algorithms in the employment market. For example, an employer might say that it values the number of patents awarded, or a university might say that it values citation counts. It becomes apparent that it hires applicants who have succeeded along these metrics while they are on the job market. A competing algorithm might now be structured by looking at the performance of all success stories across the country. If every Nobel Prize winner started out at the University of Chicago and had been hired at Chicago after graduate school at Berkeley, there might be a case for Columbia's interviewing only those applicants who are presently at Chicago and who were also educated at Berkeley. Further retrofitting by applicants is unlikely to yield anything useful. But the important point is that data science is now barely appropriate. The data set has grown; it might be divisible in two and even in three; and the objection based on predictive importance is no more powerful here than it is with respect to the conventional human-directed hiring process. A strange way to say this last point is that algorithms that make room for predictively unimportant and irregular features should be rejected in favor of a Bayesian approach that depends on priors—and which rely on theories that are not themselves developed after the evidence is in. Retrofitting is a problem for Bayesian humans as it is for algorithmic decision-making.

## IV. SYNTHETIC ALGORITHMS

### *A. Inferring Counterfactuals*

Much of the enthusiasm for introducing algorithms in law is based upon the false belief that algorithmic decisions are completely data-driven. As we have seen, W can only study the behavior of those who were set free, and then make educated guesses about how those who remained incarcerated might have acted had they been released. If W ignores these retained persons, W's data set is severely biased, or simply uninformed; we would like to account for past errors and successes, and some of the errors involved retaining prisoners who would have imposed no costs if released. W's assessment of the incarcerated group is especially important because judges choose continued incarceration for a reason. The incarcerated group is hardly random.[68] Algorithmic performance depends upon how accurately W can infer what would have happened to the incarcerated defendants had they been paroled or granted bail. Clearly, this algorithmic decision is based upon severely limited data. A complete solution is stubbornly imperfect. A larger data set is not the answer because the type of data required by W does not exist. In short, W must infer a counterfactual. This inference is based upon observable data, and gaps are filled with theory and assumptions to produce what we might call a synthetic, as opposed to a data-driven, algorithm.[69]

The challenge of such synthetic algorithms is quite different from that of predicting the future when circumstances have changed. In the case of stocks, we do not expect an algorithm to do particularly well in predicting the price of a given stock one month from today, because there are many omitted variables.[70] Some of these are inevitable simply because of the passage of time. Similarly, the behavior of people released from prison in 2021 may be affected by many factors that were unobserved, and even non-existent, in 2019. Machine learning's ability to handle many variables at once is its strength, but omitted variables inevitably limit algorithmic performance.[71] In contrast, the challenge with respect to predicting behavior after release from

[68] This means that omitted variable bias poses a significant challenge since unobserved characteristics of jailed defendants will very likely be correlated with the fact that they have been jailed.

[69] Statistical methods that use theory and assumptions to label data are sometimes called synthetic methods. Throughout this Section, we have these methods in mind.

[70] Compare stock prediction to the frequently cited (and puffed up) examples of machine learning's triumphs in games of Jeopardy, Chess, and Go. Machines triumph in those games because the rules are fixed, variables are tractable, and the prediction environment is stable. Deep Blue, Stockfish, and AlphaGo, would undoubtedly perform less successfully if the rules of Trivia, Chess, and Go changed every few years.

[71] See Leslie Valiant, *Probably Approximately Correct* 61-62 (2013) (explaining that learning cannot occur when the context of a generalization is changing).

prison comes largely from all the unobserved counterfactuals; an entire category of data is missing. The same is true in medicine because we are normally unwilling to randomize in certain ways; we will not deny medical treatment or food types to subjects, and certainly not to randomly selected subjects. In both settings, we are at the mercy of our ability to theorize about the hypothetical outcomes of some class of persistently unobservable events or the discovery of natural experiments.

Notice how many problems are solved if the task is to develop an identification, or "pattern matching," algorithm, as it is sometimes called. For example, many humans are good at identifying plants and even other humans. We see faces and recognize people we have seen before, or have seen on television or in photographs. Few of us would be any good at these tasks if there were millions of subjects to identify, because we can keep just hundreds or several thousand in mind. A machine can obviously keep more samples in mind. Computer scientists have worked hard on pattern-matching algorithms. It is easy to see the application to police work. Intuitively, it seems that machine learning has an easier task here. Every time it correctly identifies a face or body, and every time it misidentifies one, it improves its knowledge base. Moreover, the challenge of changing circumstances can be met by updating pictures in the data bank, and this improves facial-recognition by both humans and machines. The algorithm improves with each success and failure. In contrast, the judge (and the statistician working in criminal law) does not get to see what would have happened to people who remain incarcerated. Whether the incarcerated defendant, if he had been released, would have flown the jurisdiction or committed a crime remains unknown.

### *B. Judicial Faith in Synthetic Algorithms*

Most machine learning applications resolve these unknowns, or "unlabeled outcomes," with probability scores or Bayesian procedures that essentially assume that the jailed defendant's propensity to flee or recidivate is well matched by the propensity of a released defendant with similar characteristics.[72] Thus, if W observes that a released defendant, B1, with characteristics Q and R, flees the jurisdiction soon after release, then W will impute the flight label to retained defendant C, who shares the same characteristics, perhaps with some adjustment for factors that are shared or not shared with other released defendants, B2 and B3, whose behaviors have

[72] *See* Kleinberg et al., *supra* note 1 at 244 (noting that recent work in computer science on bail algorithms acknowledges the problem of unobserved counterfactuals and that all of the methods for addressing it rely on "a 'selection on the observables' assumption to impute outcomes [to jailed defendants]").

also been observed. The problem with this approach, of course, is that judges or algorithms may have selected B1, B2, and B3 for early release without noting various characteristics. Suppose W thoroughly examines criminal histories and age, but fails to record Bs' early childhood experiences or church attendance. W may have thought that these histories could not possibly be relevant or the Bs may not recollect facts from their own childhood.[73] Inasmuch as C's early childhood strongly resembles those of B1, B2, and B3, W's limited information may be unfortunate, though it is often not particularly problematic. The unobserved variables do not, after all, change C's propensity to misbehave after release, but it could well be that the failure to note the similar childhood experiences of C and the Bs leads to a false prediction about C's suitability for early release.

1. Algorithms without theory

Recent work has attempted to demonstrate how algorithms like W1 can improve upon judicial decisions to grant release by predicting the flight and recidivism risk of defendants that the judge has already decided to release.[74] Suppose W demonstrates that the riskiest 1% of defendants, such as those who have committed violent crimes, recidivate at a rate of 60%. Suppose further that judges, in the aggregate, would on their own release these defendants at a rate of 45%. By retaining those whom the algorithm predicts to be high risk, but that a lenient judge would otherwise release, overall accuracy can be improved. While it is possible that high-risk defendants may possess unobserved characteristics, the data on released high-risk defendants can show whether those that the algorithm predicts to be high-risk actually recidivate. While this approach may partially side-step the challenge of inferring counterfactuals, it does nothing to address the reality of changed circumstances.[75] The environment of the riskiest 1% of defendants in 2020 may change, perhaps because new medicines become available or technical courses are offered to at-risk offenders.

This approach also fails to address the judge's (or society's) appetite for risk, but that failure is manageable. Incarceration is costly but so is recidivism. One approach is to structure the algorithm so that it matches the risk level (or

[73] Childhood amnesia is documented and likely. *See* Mark L. Howe, *Memory Development* in Lynn S. Liben & Ulrich Müller (eds.) *Handbook of Child Psychology and Development Science*, Vol. 2 Cognitive Processes 217 (7th ed. 2015) (noting that even memories which do occur within the first 5 to 10 postnatal years tend to be poorly integrated and less durable).

[74] Kleinberg et al., *supra* note 1 at 261-269.

[75] It also cannot observe the behavior of the defendants jailed by the most lenient judges.

expected net cost) that has been produced by judges in the recent past. A successful algorithm will do this and incarcerate fewer people, thus reducing social costs. An alternative is to match the level of incarceration that has been accepted in the past, but to show that the outcome of algorithmic decision-making that is now used, reduces serious recidivism. Note, in passing, that it is easy to imagine that some judges are better than others, and that the W1 algorithm should be used in place of some judges but not others.

2. From facial recognition to trademark confusion

It is important (for judges and all of us) to be skeptical of synthetic approaches that require questionable assumptions. Empirically minded legal academics are often impressed with large data sets when it is the quality of hypotheses that matters most. Facial recognition software is so successful because it operates in a setting where data alone serves the needs of machine learning. Synthetic algorithms, on the other hand, require good theory and assumptions. Virtually every legal application will require synthetic algorithms because unobserved counterfactuals are involved, and thus some theory is required, as discussed earlier.[76]

Consider, for example, an application to the law regarding trademark confusion. Imagine that the British firm that owns the Holiday Inn brand hotels wants to show that a new set of hotels, bearing the brand name of Holiday Hotels, infringes on its Holiday Inn trademark. The claim is that the name confuses customers who may have a bad experience at a Holiday Hotel and therefore downgrade their view of Holiday Inns. Plaintiff must show that consumers are in act confused; they do in fact see a Holiday Hotel and think it is related to Holiday Inn. Courts have adopted a multifactor test to assess this sort of confusion. Most jurisdictions evaluate the "similarity of the marks" and permit parties to support their claims with survey evidence of confusion.[77] Similarity is typically established by experts and resolved by fact-finders, but imagine that P, the owner of Holiday Inn, develops an algorithm, similar to facial recognition, that precisely measures the number and placement of pixels, as well as their color, in order to provide evidence of trademark similarity. Suppose further that P's algorithm returns a similarity score. It should be obvious that if D, the owner of Holiday Hotel, were to develop an algorithm based upon the number, placement, and color

[76] *See supra* subpart II.A. See also Fagan & Levmore, *supra* note 4 at subpart III.C.2, noting that law often allocates mutually exclusive rights simultaneously and therefore creates winners and losers. Allocation, like imprisonment, creates unobserved counterfactuals.

[77] *See, e.g.*, Polaroid Corp. v. Polaroid Electronics Corp., 287 F.2d 492, 495 (2d Cir. 1961) (Friendly, J.); AMF, Inc. v. Sleekcraft Boats, 599 F.2d 341 (9th Cir. 1979).

of pixels, D's result would approximate P's. The similarity score is objectively derived from the underlying data, and is identical in approach to the pattern-matching and facial recognition algorithms. It generally requires no additional theory or assumptions to assess the fact, or likelihood, of confusion.[78]

But P and D will also argue about the other, established "confusion factors," including the evaluation of survey evidence. Surveys generally consist of litigants simply showing the two marks to a sample of consumers and asking them if they would be confused. Clever, and well paid, marketing experts might show several pictures and then ask which brand is more likely to offer a swimming pool. If consumers cannot recall that it was all the Holiday Inns they were shown, but only one in eight of the Holiday Hotels, then they are confused. P will prevail on this "evidence of confusion" factor, if the judge or jury is impressed with P's survey evidence.

Note the similarity of the trademark problem to that of assessing post-release criminality. Judges (and algorithms) can observe the characteristics of products and their trademarks as they can of incarcerated defendants, and they can then predict confusion and recidivism on the basis of observed characteristics. However, they cannot easily label a trademark as confusing or a defendant as a recidivist unless they permit free circulation of trademarks and defendants, and then observe outcomes. It is conceivable that enough customers can be found who have actually passed both Holiday Inns and Holiday Hotels, but even then, these are unlikely to be typical or randomly drawn customers. As a practical matter, the only solution is to assign hypothetical and unobserved labels to the incarcerated defendants and some likely hotel customers.

Law's faith in the accuracy of a synthetic algorithm about bail or shortened prison sentences should be predicated upon the credibility of the assumptions and theory required to assign those labels. Survey evidence in trademark confusion cases can be understood the same way. The credibility of the survey depends on how well the responses of the sampled consumers replicates the actual confusion that would have occurred throughout the population, or potential customers, had the trademark been permitted to circulate. We ought to prefer the survey evidence based on subjects who more closely resemble likely future customers. Courts should certainly prefer

[78] Note that P and D must evaluate the same trademark under identical conditions to reach this result. Perhaps P may insist on evaluating the trademark at night, in dim light. Even data-driven algorithmic decisions depend upon human data selection. *See supra* note 4.

larger over smaller sets of subjects, though it might be quite difficult to discover that P or D hired multiple marketing experts, and now simply brings to court the one whose result puts its claim in the best light. Law is a long way from requiring pre-specification of experimental methods.

Basic algorithmic tasks which are purely data-driven, such as pattern-matching and facial recognition, can be embraced to the extent that the future is likely to resemble the past, and the data represents the true population. Competitions for superior accuracy are then likely to be resolved on the basis of data volume, enhanced by the suitable division and withholding of data, as discussed earlier.[79] But when law is faced with unobserved counterfactuals, the problem, and its solution, is more complicated. In such cases, analytically sound hypotheses matter more than large data sets. Intuitions and serious work about financial markets offer an easy way to see this point. We have suggested that competition among algorithms, or even among marketing studies, offers a way for courts to find their way toward improved decisions that benefit from data and empirical methods. Still, when there is "missing" data because of unavailable counterfactuals or other reasons, synthetic algorithms come into play. Law now faces the difficult problem of seemingly attractive but misleading arguments, and we have suggested that competition among algorithms, and even competitions outsourced to the public and the growing number of curious and remarkably skilled data scientists eager to construct winning algorithms, may be the thing of the future.

It is apparent that law presents a tough challenge for machine learning. Most legal questions present small data sets, changed circumstances, and unobserved outcomes. Legal problems could not be more different than problems of facial recognition. Law needs prediction rather than identification algorithms. It may be an easy thing for machines to do better than judges, but the important task is likely to be to convince judges that they can get the most out of machine learning by sponsoring competitions among algorithms.[80] Judges and other lawmakers may not like the idea that law can be outsourced, but we have shown that competing algorithms offer an attractive strategy for bringing data analysis into legal decision-making.

---

[79] *See supra* subpart II.A.

[80] *Cf.* Nina Grgic-Hlaca, *Human Decision Making with Machine Assistance: An Experiment on Bailing and Jailing*, MPI Collective Goods Discussion Paper (2019), https://ssrn.com/abstract=3465622 (conducting an experiment and finding that giving machine advice on recidivism to lay-judges has only a small effect, and is biased in the direction of predicting no future criminal behavior).

## V. CONCLUSION

Machine learning has, inevitably, found its way into law and it is likely to expand its reach. This Article has exposed some of the problems that this new method—developed in settings with much larger data sets than are normally found in law, and without the complexity that unobserved outcomes present in law—presents for lawmakers. Beginning with the competition between prosecutors and criminal defendants, we have developed several novel solutions that should find their place in a variety of legal areas, beginning with criminal sentencing influenced by the likelihood that released persons will become recidivists.

First, there should be some limits on the ability of the second mover (normally the sentenced offender) to introduce an algorithm that is more favorable to him than that introduced by the state. The key insight here is that the second mover, especially if well funded, can tilt the algorithm in a way that favors his own characteristics. He is able to retrofit data. He should, however, be allowed to explain to a judge why these individual characteristics matter, but it is poor statistics to allow an algorithm that was able to be retrofitted. A more sophisticated innovation is to allow the second mover to participate at the very outset in a competition among algorithms without knowing the content and strategy of other algorithms. Without this knowledge it is unlikely that retrofitting will be much help. Finally, and most interesting, the state (or a party or court involved in a civil case or in constructing tax or environmental policy) might discover the best algorithm, and certainly one that performs better than human judges who will still be able to specify how success is to be measured, by encouraging a public competition. The trick here is to withhold data and then test competing algorithms on withheld data. Funds that the state now uses to present empirical evidence could be used to reward winning algorithms, whether produced by a defendant, a law school clinic, or, more likely, by a mere enthusiast who has developed talent in the new area of data science. An obvious advantage of this approach is that it levels the playing field for impoverished defendants. Some of the energy now directed to finance and other profitable fields could be used to improve law.

This Article has also addressed the problem of unobserved data, and the need for synthetic algorithms to take account of counterfactuals. In criminal law, as in medicine, it is impossible to study perfect control groups, and yet there is a need to imagine what unreleased persons would have done if released, just as there is a need to estimate what would have happened if some suffering individuals had not received a given treatment. Here, too, we have suggested

some solutions but, in the long run, encouraging competition among algorithms is likely to hold great promise.

***