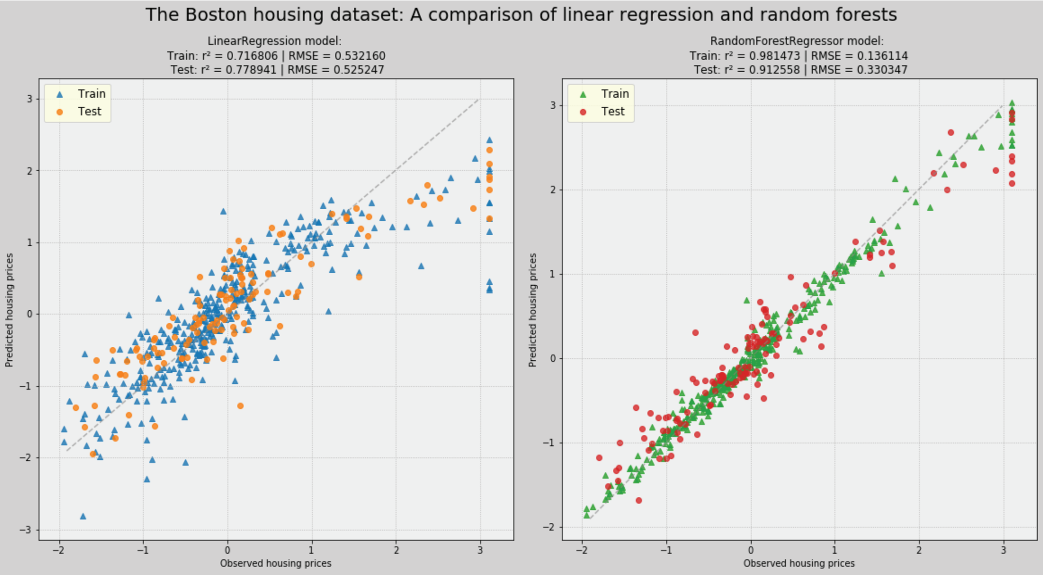
SPLIT DECISIONS: PRACTICAL MACHINE LEARNING FOR EMPIRICAL LEGAL SCHOLARSHIP

James Ming Chen[*]

2020 MICH. ST. L. REV. 1301



ABSTRACT

Multivariable regression may be the most prevalent and useful task in social science. Empirical legal studies rely heavily on the ordinary least squares method. Conventional regression methods have attained credibility in court, but by no means do they dictate legal outcomes. Using the iconic Boston housing study as a source of

[*] Professor of Law and Justin Smith Morrill Chair in Law, Michigan State University; Visiting Scholar, Faculty of Economics and Business, University of Zagreb (Ekonomski Fakultet, Sveučilište u Zagrebu); Executive Vice President and Chief Data Scientist, Silver Leaf Capital LLC. Portions of this Article draw upon James Ming Chen, *An Introduction to Machine Learning for Panel Data*, 27 INT'L ADVANCES ECON. RSCH. 1 (2021), a more technically comprehensive presentation of machine-learning methods for panel data.

price data, this Article introduces machine-learning regression methods. Although decision trees and forest ensembles lack the overt interpretability of linear regression, these methods reduce the opacity of black-box techniques by scoring the relative importance of dataset features. This Article will also address the theoretical tradeoff between bias and variance, as well as the importance of training, cross-validation, and reserving a holdout dataset for testing.

TABLE OF CONTENTS

INTRODUCTION

Empirical legal scholarship, like many other branches of the social sciences, relies almost entirely on generalized linear models. The ultimate questions of criminal justice—guilt versus innocence, death versus life—present questions of binary classification.[1] Other questions, such as market power inquiries in antitrust law, requiring the estimation of some parameter or variable on a continuous scale.[2] Even more tangibly, the law may question whether the final prices offered by automobile dealers to customers are fair.[3] For decades, courts and commentators have relied on multivariable regression to solve these types of quantitative problems.[4]

Perhaps no task is more prevalent, or more useful, in the social sciences than the prediction of a numerical value through its relationship with other variables. This is the domain of "regression," defined broadly as a family of "technique[s] for estimating a mathematical relationship between factors on the basis of numerical data."[5] Perhaps the most valuable contribution of "regression analysis is to organize and explain data that may [otherwise] appear to be random."[6] As is evident from the simplest binary model of logistic regression, classification can be cast in continuous rather than discrete

[1] *See* David C. Baldus, Catherine M. Grosso, George Woodworth & Richard Newell, *Racial Discrimination in the Administration of the Death Penalty: The Experience of the United States Armed Forces (1984-2005)*, 101 J. CRIM. L. & CRIMINOLOGY 1227, 1239 (2012); David C. Baldus, George Woodworth, Catherine M. Grosso & Aaron M. Christ, *Arbitrariness and Discrimination in the Administration of the Death Penalty: A Legal and Empirical Analysis of the Nebraska Experience (1973-1999)*, 81 NEB. L. REV. 486, 545 (2002).

[2] *See* Ira Horowitz, *Market Definition in Antitrust Analysis: A Regression-Based Approach*, 48 S. ECON. J. 1, 7 (1981).

[3] *See* Ian Ayres, *Fair Driving: Gender and Race Discrimination in Retail Car Negotiations*, 104 HARV. L. REV. 817, 836–41 (1991).

[4] *See* Franklin M. Fisher, *Multiple Regression in Legal Proceedings*, 80 COLUM. L. REV. 702, 702 (1980).

[5] *See* Michael O. Finkelstein, *Regression Models in Administrative Proceedings*, 86 HARV. L. REV. 1442, 1442 (1973).

[6] Bazemore v. Friday, 478 U.S. 385, 403 n.14 (1986) (citing Fisher, *supra* note 4, at 705–07).

terms.[7] In multinomial as well as binary classification, these tasks require an estimate of the probability that a particular observation will fall into a particular class.[8] Linear regression, logistic regression, and multinomial generalizations of logistic regression all belong to the broader class of generalized linear models.[9]

By far the most popular tool for regression is the multivariable generalization of the ordinary least squares method.[10] Linear regression is by far the most popular method for evaluating panel data in economics, and perhaps in the social sciences at large. Every spreadsheet and statistics package performs linear regression. Regression results are widely and readily understood. The scale and sign of coefficients, along with *p*-values and *t*-statistics, communicate valuable information among all scientists conversant with conventional statistics.

Despite these benefits, linear regression may not be the most accurate method for making predictions from panel data. Machine learning and artificial intelligence have dramatically expanded the range of tools that lawyers and legal scholars may exploit. Open-source software and a burgeoning coding community have introduced these methods to a broader audience.

These methods expose a cultural split within science. Multivariable regression by ordinary least squares represents the dominant statistical culture.[11] This dominant culture assumes that the data are generated by a specific stochastic data model. Machine learning represents the competing culture of algorithmic models.[12] The algorithmic culture suspends assumptions regarding the mechanism

[7] *See* J. S. Cramer, *The Early Origins of the Logit Model*, 35 STUD. HIST. & PHIL. BIOLOGICAL & BIOMEDICAL SCIS. 613, 614 (2004); Juliana Tolles & William J. Meurer, *Logistic Regression Relating Patient Characteristics to Outcomes*, 316 J. AM. MED. ASS'N 533, 533–34 (2016).

[8] *See* Strother H. Walker & David B. Duncan, *Estimation of the Probability of an Event as a Function of Several Independent Variables*, 54 BIOMETRIKA 167, 167–68 (1967). *Compare* DAVID W. HOSMER, JR., STANLEY LEMESHOW & RODNEY X. STURDIVANT, APPLIED LOGISTIC REGRESSION 1–34 (3d ed. 2013) (introducing logistic regression as a method for predicting dichotomous, yes/no answers), *with id.* at 35–48 (presenting multiple logistic regression analysis as a generalization of the basic, binary method).

[9] *See generally* ANNETTE J. DOBSON & ADRIAN G. BARNETT, AN INTRODUCTION TO GENERALIZED LINEAR MODELS 104–19, 149–96 (4th ed. 2018).

[10] *See* LEE EPSTEIN & ANDREW D. MARTIN, AN INTRODUCTION TO EMPIRICAL LEGAL RESEARCH 176–82 (2014).

[11] *See* Leo Breiman, *Statistical Modeling: The Two Cultures*, 16 STAT. SCI. 199, 206 (2001).

[12] *See id.*

by which data is generated and distributed. Consequently, the algorithmic culture enjoys a wider range of algorithms promising greater accuracy and perhaps deeper understanding of data at any scale.[13]

According to the "no free lunch" theorem of machine learning, we cannot know ahead of time which model is best suited to a particular set of data.[14] Consequently, the most practical approach lies in applying the widest feasible range of methods. *A priori* assumptions cannot supplant experimentation. Not altogether subtly, this Article urges legal scholars, economists, and other social scientists to draw liberally from methods on both sides of the boundary between the statistical and algorithmic cultures.

Linear regression and its closest kin hold a commanding advantage within law: At least in some contexts, courts accept these methods as reliable and authoritative. Though some authorities have been quick to limit or criticize conventional regression techniques, courts greet these methods with grudging respect and, more often than not, open acceptance. This Article therefore begins by reviewing Supreme Court cases that have directly discussed regression. History suggests that machine-learning alternatives to conventional regression techniques will percolate through administrative and judicial proceedings in settings where quantitative inference is paramount before eventually winning broader acceptance.

This Article then introduces the family of machine-learning methods based on decision trees. These methods, at their root, depend on decision trees to divide data, variable by variable. Statistically informed extensions, such as bagging and pasting, increase the explanatory reliability of decision trees. Ensembles of decision trees harness the Delphic wisdom of numerous miniature regressors. This Article applies these basic methods to the Boston housing study as an iconic instance of regression tasks of greatest interest to law.

Decision trees and ensembles can also perform classification tasks. These methods can supplement classification through binomial or multinomial logistic regression. The configuration and optimization of a tree- or forest-based classifier follow almost exactly the same steps as those required for regression. The machine-learning

[13] *See id.*

[14] David H. Wolpert, *The Lack of A Priori Distinctions Between Learning Algorithms*, 8 NEURAL COMPUTATION 1341, 1343 (1996) (internal quotation marks omitted).

techniques described in this Article may therefore be adapted to classification.

Machine learning is no panacea. Its uptake in law, as in other disciplines, requires a sober evaluation of advantages and pitfalls. Trees and forests lack the overt interpretability of linear regression. Machine-learning packages often compensate for the opacity of these "black-box" techniques by scoring the relative importance of dataset features. In light of these theoretical and practical considerations, this Article will also address the tradeoff between bias and variance, as well as the importance of training, cross-validation, and reserving a holdout dataset for testing.

I. CONVENTIONAL REGRESSION METHODS BEFORE THE SUPREME COURT: A PAGE OF HISTORY[15]

Legal authorities have embraced conventional regression methods for the past half-century.[16] Although many lawyers and judges remain uncomfortable interpreting (let alone constructing) a regression model, this basic scientific tool has won widespread if not unconditional acceptance throughout law. One strand of that history is worth revisiting. The Supreme Court's treatment of various forms of linear regression offers a possible preview of the law's attitude toward regression through machine learning. "The law," after all "embodies the story of a nation's development" across the ages, "and it cannot be dealt with as if it contained only the axioms and corollaries of a book of mathematics."[17]

A. The Broader Context: Law and Social Science

Legal acceptance of regression belongs to an older, broader tradition traceable to the brief filed by future Justice Louis Brandeis in the 1908 case of *Muller v. Oregon*.[18] That brief, devoted almost entirely to scientific studies on the effects of long working hours on women rather than legal precedent, is widely (though not universally)

[15] *Cf.* N.Y. Tr. Co. v. Eisner, 256 U.S. 345, 349 (1921) ("Upon this point a page of history is worth a volume of logic.").

[16] *See* Finkelstein, *supra* note 5, at 1442; Fisher, *supra* note 4, at 702.

[17] O. W. HOLMES, JR., THE COMMON LAW 5 (Mark DeWolfe Howe ed., 1963) (1881).

[18] 208 U.S. 412 (1908).

regarded as a milestone in the legal role of social science.[19] As a Supreme Court Justice, Brandeis's embrace of empiricism is perhaps most evident in his dissent in *New State Ice Co. v. Liebmann*.[20] In support of a state legislature's prerogative to regulate the ice industry, Justice Brandeis recited the prevailing scientific literature on food spoilage and refrigeration.[21]

Brown v. Board of Education explicitly invoked "modern authority" on "psychological knowledge" regarding racial differences in children's reactions to dolls decorated to look either white or black.[22] The Supreme Court invoked so-called doll studies as evidence of "a feeling of inferiority . . . that may affect [black children's] hearts and minds in a way unlikely ever to be undone."[23] *Brown* heralded the first and arguably still the most important triumph of social science in American law.[24] Psychologists and educational experts concede,

[19] *See, e.g.*, Marion E. Doro, *The Brandeis Brief*, 11 VAND. L. REV. 783, 792 (1958); Jonathan Yovel & Elizabeth Mertz, *The Role of Social Science in Legal Decisions*, *in* THE BLACKWELL COMPANION TO LAW AND SOCIETY 410, 414 (Austin Sarat ed., 2004). For thoughtful critiques of the mythology of the Brandeis brief, see David E. Bernstein, *Brandeis Brief Myths*, 15 GREEN BAG 9, 9–10 (2011); Noga Morag-Levine, *Facts, Formalism, and the Brandeis Brief: The Origins of a Myth*, 2013 U. ILL. L. REV. 59, *passim* (2013).

[20] 285 U.S. 262, 280–311 (1932) (Brandeis, J., dissenting).

[21] *See* Daniel A. Farber, *Reinventing Brandeis: Legal Pragmatism for the Twenty-First Century*, 1995 U. ILL. L. REV. 163, 175–76 (1995). *See generally* G. Alan Tarr, *Laboratories of Democracy? Brandeis, Federalism, and Scientific Management*, 31 PUBLIUS: J. FEDERALISM 37 (2001).

[22] 347 U.S. 483, 494, 495 & n.11 (1954); *accord* Kenneth B. Clark & Mamie P. Clark, *Emotional Factors in Racial Identification and Preference in Negro Children*, 19 J. NEGRO EDUC. 341, 344–50 (1950) (summarizing the research cited in *Brown*).

[23] *Brown*, 347 U.S. at 494.

[24] *See generally, e.g.*, Ludy T. Benjamin, Jr. & Ellen M. Crouse, *The American Psychological Association's Response to* Brown v. Board of Education*: The Case of Kenneth B. Clark*, 57 AM. PSYCH. 38 (2002); Stuart W. Cook, *Social Science and School Desegregation: Did We Mislead the Supreme Court?*, 5 PERSONALITY & SOC. PSYCH. BULL. 420 (1979); Harold B. Gerard, *School Desegregation: The Social Science Role*, 38 AM. PSYCH. 869 (1983); John P. Jackson Jr., *The Scientific Attack on* Brown v. Board of Education, *1954–1964*, 59 AM. PSYCH. 530 (2004); Janet Ward Schofield & Leslie R. M. Hausmann, *School Desegregation and Social Science Research*, 59 AM. PSYCH. 538 (2004) (suggesting that *Brown* may have precipitated a decline in research on school desegregation and its effects). *But see, e.g.*, William E. Doyle, *Can Social Science Data Be Used in Judicial Decisionmaking?*, 6 J.L. & EDUC. 13, 16 (1977) (attributing *Brown* to constitutional theory rather than scientific influence).

however, that "public policies promoted with the doll test . . . have not remedied the deep disparities of racial inequality in U.S. education."[25]

Finally, the Court's treatment of expert testimony foreshadows how the judiciary may eventually treat machine learning. As recently as the early 1990s, some federal courts observed the so-called *Frye* rule, which provided that "expert testimony . . . must be sufficiently established to have gained general acceptance in the particular field in which it belongs."[26] The *Frye* regime ended with the 1993 case of *Daubert v. Merrell Dow Pharmaceuticals, Inc.*[27] Rejecting *Frye*'s standard of general acceptance, *Daubert* relied on the express language of the Federal Rules of Evidence, which provide that "scientific, technical, or other specialized knowledge" may be admitted as expert testimony if such knowledge "will assist the trier of fact to understand the evidence or to determine a fact in issue."[28] Most of the commentary on *Daubert* characterizes this decision as shifting the criterion for determining scientific validity from acceptance within a scientific discipline to some sort of evaluation by admittedly nonexpert, generalist judges.[29]

B. Conventional Regression Models Before the Most Numerate Branch: The Supreme Court at the Bar of Statistics[30]

The Supreme Court has explicitly cited or discussed traditional methods of regression since the 1970s. The history of conventional regression models before the Court, brief in temporal and doctrinal terms, falls into three phases. The first phase spanned the decade after the Court's reinstatement of the death penalty after the 1972 *Furman*

[25] Gwen Bergner, *Black Children, White Preference:* Brown v. Board*, the Doll Tests, and the Politics of Self-Esteem*, 61 AM. Q. 299, 302 (2009); *accord* Kenneth B. Clark, *The Social Sciences and the Courts*, 17 SOC. POL'Y 33, 37 (1986).

[26] Frye v. United States, 293 F. 1013, 1014 (D.C. Cir. 1923).

[27] 509 U.S. 579, 587 (1993).

[28] FED. R. EVID 702(b); *accord Daubert*, 509 U.S. at 588.

[29] *See, e.g.*, Shana M. Solomon & Edward J. Hackett, *Setting Boundaries Between Science and Law: Lessons from* Daubert v. Merrell Dow Pharmaceuticals, Inc., 21 SCI., TECH., & HUM. VALUES 131 (1996); *cf.* Edward K. Cheng & Albert H. Yoon, *Does* Frye *or* Daubert *Matter? A Study of Scientific Admissibility Standards*, 91 VA. L. REV. 471, 503 (2005) (finding no statistically significant difference arising from state courts' decisions either to adhere to *Frye* or to adopt *Daubert*).

[30] *Cf.* ALEXANDER M. BICKEL, THE LEAST DANGEROUS BRANCH: THE SUPREME COURT AT THE BAR OF POLITICS (2d ed. 1986); Paul H. Edelman & Jim Chen, *The Most Dangerous Justice: The Supreme Court at the Bar of Mathematics*, 70 S. CAL. L. REV. 63 (1996).

v. Georgia decision had imposed a *de facto* nationwide moratorium on capital punishment.[31]

A second phase consisted of a 1986–1987 trilogy of cases involving racial bias in voting, employment, and (again) capital punishment. Judicial resistance to this class of statistical methods climaxed in the third and most controversial of these cases, *McCleskey v. Kemp*.[32] Cases since *McCleskey*, reflecting greater though inconsistent and perhaps grudging acceptance of regression methods, represent a third phase. Class certification under the Federal Rules of Civil Procedure has become the only legal setting outside criminal justice in which open judicial hostility to the evidentiary and normative claims of regression might hinder legal acceptance of generalized linear methods.

Throughout all three phases, the constitutionality of capital punishment has persistently influenced the Supreme Court's reaction to regression models, their underlying methods, and the ultimate question of legality. At a slightly higher level of generality, the Court has quite often focused on statistical studies aiming to quantify the impact of racial bias on a legally consequential act. This tendency in the types of controversies that bring regression methods before the high court has narrowed the Justices' focus, and not necessarily in a way that promotes the integration of regression as a scientific technique into legal analysis.

Regression models in law, especially in administrative proceedings, fall into two broad categories. Some controversies place "primary concern" on "the value of the[] coefficients" of explanatory variables in a regression model.[33] Others direct "the focus of attention" toward "the computed value of the dependent variable."[34] Legal controversies seeking to clarify the impact of racial bias tend to place greater emphasis on the coefficients and statistical significance of independent variables than on fitted values of the target variable.[35] Allegations of racial discrimination in criminal justice, voting rights, and employment have dominated the Supreme Court cases that have

[31] *See* 408 U.S. 238, 238–40 (1972) (per curiam).

[32] *See* 481 U.S. 279, 279–80 (1987).

[33] Finkelstein, *supra* note 5, at 1445.

[34] *Id.*

[35] Bazemore v. Friday, 478 U.S. 385 (1986), is an important exception. *Bazemore* focused on racial differences in salaries predicted by a regression model, a classic instance in which a model's predictions were more valuable than its coefficients and *p*-values.

asked the Justices to evaluate regression studies and to decide the law's receptivity to this class of scientific methods.

C. Phase One: Matters of Life and Death

The Court's earliest opportunity to examine regression models originated in the 1968 Eighth Circuit case of *Maxwell v. Bishop*.[36] The future Justice Blackmun rejected a study showing highly racialized patterns in the imposition of the death penalty for rape in southern states between 1945 and 1965.[37] "Standing by themselves," then-Judge Blackmun reasoned, the "facts as to rape charges in Garland County [Arkansas]" and their statistical evaluation "disclose nothing from which conclusions of unconstitutionality . . . may appropriately be drawn."[38]

Affirming on grounds with no explicit connection to race, the Supreme Court had no occasion in *Maxwell* to review then-Judge Blackmun's treatment of the statistical evidence.[39] A quarter-century later, Justice Blackmun would renounce all efforts to reconcile capital punishment with the Constitution.[40] The full Court would eventually forbid the imposition of the death penalty for the rape of an adult.[41] Even more broadly, the Court would forswear capital punishment for any crime not involving treason or the death of a human victim.[42] Regression played no meaningful role in any of those cases, which rested solely on legal and moral reasoning.

The Court would soon have further opportunities to evaluate regression studies in the context of the death penalty. Invalidating all death sentencing schemes then prevailing, the Justices placed a moratorium on capital punishment in the 1972 case of *Furman v. Georgia*.[43] Justice Marshall's concurrence invoked scientific studies on the deterrent effects of the death penalty, albeit without directly discussing their underlying methodology.[44] In *Gregg v. Georgia*, one of five 1976 decisions that reconsidered and collectively restored the

[36] 398 F.2d 138 (8th Cir. 1968), *vacated*, 398 U.S. 262 (1970).

[37] *See* Samuel R. Gross, *David Baldus and the Legacy of* McCleskey v. Kemp, 97 IOWA L. REV. 1905, 1906–07 (2012).

[38] *Maxwell*, 398 F.2d at 148.

[39] *See* 398 U.S. at 263.

[40] *See* Callins v. Collins, 510 U.S. 1141, 1143–59 (1994) (Blackmun, J., dissenting).

[41] *See* Coker v. Georgia, 433 U.S. 584, 592, 597–600 (1977).

[42] *See* Kennedy v. Louisiana, 554 U.S. 407, 437 (2008).

[43] 408 U.S. 238 (1972).

[44] *See id.* at 238, 352–55 & nn. 124–25 (Marshall, J., concurring).

death penalty after *Furman*, the principal opinion of Justices Stewart, Powell, and Stevens described "[s]tatistical attempts to evaluate the worth of the death penalty as a deterrent to crimes by potential offenders" as "simply . . . inconclusive."[45] In the companion case of *Roberts v. Louisiana*, Justice White likewise cited the "inconclusive nature of statistical studies" on capital punishment.[46]

The task of closely examining regression methodology fell to Justice Marshall. His *Gregg* dissent may be characterized as the first Supreme Court opinion that paid close attention to the mechanics and legal significance of regression. Justice Marshall disputed the premises and the conclusion of a paper by Isaac Ehrlich, which had "found a negative correlation between changes in the homicide rate and changes in execution risk" and surmised that "each additional execution in the United States" between 1933 and 1967 "might have saved eight lives."[47]

Regression analysis of the death penalty's hypothesized deterrent effect was the first application of this method to influence a Supreme Court decision. Justice Marshall reviewed the literature responding to Ehrlich's study, much of it published in law reviews.[48] Justice Marshall's criticisms reflected a sophisticated understanding of multiple regression and its limitations—or at least of the literature responding to Ehrlich. Justice Marshall observed that Ehrlich may have erred in "compar[ing] execution and homicide rates on a nationwide, rather than a state-by-state, basis," in such a way that "[t]he aggregation of data from all States—including those that have

[45] Gregg v. Georgia, 428 U.S. 153, 184–85 (1976).

[46] Roberts v. Louisiana, 428 U.S. 325, 355–56 & n.7 (1976) (White, J., dissenting).

[47] *Gregg*, 428 U.S. at 234 (Marshall, J., dissenting). *See generally* Isaac Ehrlich, *The Deterrent Effect of Capital Punishment: A Question of Life and Death*, 65 AM. ECON. REV. 397 (1975).

[48] *See Gregg*, 428 U.S. at 235 n.8 (citing, *inter alia*, Peter Passell, *The Deterrent Effect of the Death Penalty: A Statistical Test*, 28 STAN. L. REV. 61 (1975); David C. Baldus & James W. Cole, *A Comparison of the Work of Thorsten Sellin and Isaac Ehrlich on the Deterrent Effect of Capital Punishment*, 85 YALE L.J. 170 (1975); William J. Bowers & Glenn L. Pierce, *The Illusion of Deterrence in Isaac Ehrlich's Research on Capital Punishment*, 85 YALE L.J. 187 (1975); Jon K. Peck, *The Deterrent Effect of Capital Punishment: Ehrlich and His Critics*, 85 YALE L.J. 359 (1976); Isaac Ehrlich, *Deterrence: Evidence and Inference*, 85 YALE L.J. 209 (1975); Isaac Ehrlich, *Rejoinder*, 85 YALE L.J. 368 (1976)). Peter Passell and John Taylor later published a further response to Ehrlich, *see* Peter Passell & John B. Taylor, *The Deterrent Effect of Capital Punishment: Another View*, 67 AM. ECON. REV. 445 (1977), which Justice Marshall cited in draft form. *See Gregg*, 428 U.S. at 235 n.8 (Marshall, J., dissenting).

abolished the death penalty—obscures the relationship between murder and execution rates."[49]

Justice Marshall also criticized "the quality of Ehrlich's data, his choice of explanatory variables, his failure to account for the interdependence of those variables, and his assumptions as to the mathematical form of the relationship between the homicide rate and the explanatory variables."[50] Finally, Justice Marshall cited "[t]he most compelling criticism of the Ehrlich study"—namely, the vulnerability of its conclusions "to the choice of the time period included in the regression analysis"—as grounds for reserving "severe doubts on the reliability of Ehrlich's tentative conclusions" and rejecting their usefulness "in assessing the deterrent impact of the death penalty."[51]

A majority in *Gregg* rejected Justice Marshall's approach. As then-Justice Rehnquist would later write in a different context, those comments in dissent were "just that: comments in a dissenting opinion."[52] Relying upon the Justices' collective failure to embrace the social science on the deterrent effect of the death penalty, the Court continued to confess its own "difficulties with sophisticated statistical methodology" in the contemporaneous antitrust case of *Illinois Brick Co. v. Illinois*.[53] *Illinois Brick* denied antitrust standing to an indirect, downstream purchaser of goods whose prices had alleged been raised by a price-fixing conspiracy among manufacturers.[54] In addition to *Furman*, *Gregg*, and *Roberts*'s failure to embrace either view of the deterrent effect of the death penalty, *Illinois Brick* recognized "the difficulties that have been encountered, even in informal adversary proceedings, with the statistical techniques used to estimate" elasticities of supply and demand.[55] The Court therefore deemed it "unrealistic to think that elasticity studies introduced by expert witnesses will resolve" controversies over the extent to which price increases attributable to an anticompetitive conspiracy have been passed on by the direct purchaser to its own customers.[56]

In the same 1976 term during which it decided *Illinois Brick*, the Court decided a trilogy of cases upholding the use of a different

[49] *Gregg*, 428 U.S. at 235 (Marshall, J., dissenting).

[50] *Id.* at 235 n.8.

[51] *Id.* at 235–36.

[52] U.S. R.R. Ret. Bd. v. Fritz, 449 U.S. 166, 176 n.10 (1980).

[53] 431 U.S. 720 (1977).

[54] *Id.* at 742.

[55] *Id.*

[56] *Id.*

statistical technique, the comparison of two sample proportions using the binomial distribution.[57] In the first of these cases, the Court permitted the use of binomial probability to detect unlawful discrimination in jury selection.[58] Confronted with allegations of gross underrepresentation of Mexican–Americans in Texas juries, the Court found that "difference[s] between the expected and observed number of Mexican–Americans" amounted to eleven or even twenty-nine standard deviations.[59]

Two other cases extended the application of binomial distributions to employment discrimination.[60] Comparing black teachers in the Hazelwood (Missouri) School District with the black representation in the overall workforce of St. Louis, the Court again found standard deviations far exceeding the "two or three" that would be sufficient to reject the null hypothesis of random, non-racially discriminatory hiring.[61] From origins traceable to the 1960s, the Court thus expanded the use of statistical disparities to detect racial bias beyond jury selection.[62] In that context, the Justices had "permitted a finding of constitutional violation even when the statistical pattern does not approach . . . extremes" so "stark" as to be evident to the untrained, naked judicial eye.[63]

These employment cases reaffirmed earlier acknowledgements that "[s]tatistical analyses have served and will continue to serve an important role" in redressing alleged discrimination.[64] The Court unequivocally declared that "[s]tatistics are equally competent in

[57] *See generally* Lawrence Brown & Xuefeng Li, *Confidence Intervals for Two Sample Binomial Distribution*, 130 J. STAT. PLANNING & INFERENCE 359 (2005).

[58] *See* Castaneda v. Partida, 430 U.S. 482, 496 & n.17 (1977).

[59] *Id.* at 496 n.17. *See generally* Michael O. Finkelstein, *The Application of Statistical Decision Theory to the Jury Discrimination Case*, 80 HARV. L. REV. 338, 353–56 (1966).

[60] *See* Hazelwood Sch. Dist. v. United States, 433 U.S. 299, 309–13 (1977); Int'l Brotherhood of Teamsters v. United States, 431 U.S. 324, 339–40 n.20 (1977).

[61] *Hazelwood*, 433 U.S. at 311 n.17.

[62] *See* Alexander v. Louisiana, 405 U.S. 625, 630 n.9 (1972); Turner v. Fouche, 396 U.S. 346, 359 (1970); Whitus v. Georgia, 385 U.S. 545, 552 & n.5 (1967).

[63] Village of Arlington Heights v. Metropolitan Hous. Dev. Corp., 429 U.S. 252, 266 n.13 (1977); *cf.* United States v. Lopez, 514 U.S. 549, 563 (1995) (questioning whether substantial effects on interstate commerce were "visible to the naked eye" in the absence of "Congressional findings" regarding gun possession in school zones).

[64] Mayor of Philadelphia v. Educ. Equal. League, 415 U.S. 605, 620 (1974); *accord Teamsters*, 431 U.S. at 339.

proving employment discrimination."[65] Although these cases did not directly involve regression methodologies, they did establish the Court's willingness to treat statistical evidence as rebuttable rather than "irrefutable."[66] The Court ultimately described statistics as a tool of "infinite variety" whose "usefulness depends on all of the surrounding facts and circumstances."[67]

Experience with statistical tests of significance regarding racial disparities may have softened the Court's skepticism. By the 1980s, the Court began to indicate greater willingness to entertain regression studies. In the context of employment discrimination, the Court acknowledged that "statistical technique[s] or other method[s]" could "quantify the effect of sex discrimination on . . . wage rates."[68] In criminal cases, the Justices realized, "sophisticated statistical sampling surveys and complex regression analyses" could likewise expose "racial bias."[69]

On multiple occasions during the early 1980s, individual Justices criticized their colleagues for failing to accord greater weight to the predictive value of regression. Justice White reached his own reckoning with regression and race in the 1980 voting rights case of *City of Mobile v. Bolden*.[70] His dissent explicitly recognized that "[r]egression analyses covering" municipal elections throughout the 1960s and 1970s had "confirmed the existence of severe bloc voting" along racial lines.[71]

In a case contesting the apportionment of anadromous fish between Idaho, Oregon, and Washington, Justice O'Connor disputed the special master's "undue skepticism" of the "linear least squares regression method."[72] Praising this method's "predictive value, if used properly," Justice O'Connor argued that "[c]ourts can rely on the same sort of calculations that agencies charged . . . with management of fisheries perform."[73]

[65] *Teamsters*, 431 U.S. at 339.

[66] *Id.* at 340.

[67] *Id.*; *accord Hazelwood*, 433 U.S. at 312.

[68] County of Washington v. Gunther, 452 U.S. 161, 181 (1981).

[69] Stephens v. Kemp, 469 U.S. 1043, 1051–52 (1984) (excusing "indigent, uneducated, incarcerated petitioners" from allegations of "inexcusable neglect for having failed" either "to perform or to underwrite" such studies).

[70] 446 U.S. 55 (1980).

[71] *Id.* at 98 (White, J., dissenting).

[72] Idaho *ex rel.* Evans v. Oregon, 462 U.S. 1017, 1038 n.8 (1983) (O'Connor, J., dissenting).

[73] *Id.*

Justices Stevens, joined by Justices Brennan and Marshall, objected to the Court's summary disposition of *Lehman v. Trout*.[74] The majority had remanded a class action suit for further "findings of fact, based on new evidence if necessary," on the "evidentiary value" of statistical evidence alleging sex discrimination.[75] The dissenters described and defended the "statistical evidence, expert testimony, and nonstatistical evidence covering specific instances of discrimination."[76] They validated the district court's reliance on "several variant regressions," including at least one model "using a logarithmic, rather than a linear, equation," which all "produced *statistically significant* results."[77] Justices Stevens described yet another "linear model which included dummy variables" for a range of educational and occupational predictors and sex as the "dependent variable."[78] The Court's "excessive reliance on summary dispositions," he alleged, gave short shrift "to the probative value of respondents' evidence after the most truncated of presentations."[79]

Justice Brennan, joined by Justice Marshall, sharply criticized the Court's denial of certiorari in the 1984 capital punishment case of *Stephens v. Kemp*.[80] Consistent with the suggestion that the Court's trilogy of 1977 cases endorsing the use of the binomial distribution to evaluate claims of bias had represented a legal breakthrough, Justice Brennan observed that comprehensive statistical studies of bias had barely begun to emerge during the late 1970s and early 1980s.[81] Indeed, the Fifth Circuit had held that "the then extant social science evidence" was "inadequate as a matter of law to raise a colorable claim of discrimination in Georgia's capital sentencing system."[82] This line of criticism drew its power from Justice Brennan's implicit assumption that the Court would entertain and credit "sophisticated statistical sampling surveys and complex regression analyses of racial

[74] 465 U.S 1056 (1984).

[75] *Id.* at 1056.

[76] *Id.* at 1057 (Stevens, J., dissenting).

[77] *Id.* at 1059 n.4.

[78] *Id.* at 1061 n.7.

[79] *Id.* at 1062.

[80] 469 U.S. 1043, 1043 (1984) (Brennan, J., dissenting from denial of certiorari).

[81] *Id.* at 1051 (citing Spencer v. Zant, 715 F.2d 1562, 1582 (11th Cir. 1983) for the proposition that that "the pertinent statistical studies . . . were available only through oral testimony" as "late as May 1982").

[82] *Id.* (citing Spinkellink v. Wainwright, 578 F.2d 582 (5th Cir. 1978), *cert. denied*, 440 U.S. 976 (1979); Smith v. Balkcom, 660 F.2d 573 (5th Cir. 1981), *cert. denied*, 459 U.S. 882 (1982)).

bias."[83] Three years later, in *McCleskey v. Kemp*, a majority of the Justices would disagree.[84]

D. Phase Two: Regression and Racial Reckoning—*Thornburg v. Gingles*, *Bazemore v. Friday*, and *McCleskey v. Kemp*

The decade-long progression beyond the Supreme Court's first, extensive engagement with regression analysis in *Gregg v. Georgia* and its companion cases on the restoration of the death penalty set the stage for three cases, all decided in 1986 and 1987, that still influence the Court's approach to this statistical tool. All three cases remain good law. Despite their common roots in racial justice, these cases combined to steer the Court's approach to regression in radically different, doctrinally contingent directions.

1. *Voting Rights:* Thornburg v. Gingles

The Court had surprised Congress and voting rights advocates by adopting the "intent test" of purposeful racial discrimination in the 1980 case of *City of Mobile v. Bolden*.[85] Congress amended section 2 of the Voting Rights Act in 1982 to restore the "results test" that the Court had articulated in the 1973 case of *White v. Regester*.[86] *Thornburg v. Gingles* made it "clear that a violation could be proved by showing discriminatory effect alone."[87]

The Court's treatment of regression in *Gingles* focused on a narrow methodological point. An expert witness for the voting rights plaintiffs had deployed "two complementary methods of analysis—extreme case analysis and bivariate ecological regression analysis—in order to determine whether blacks and whites in [North Carolina] differed in their voting behavior."[88]

The Court approvingly quoted the district court's conclusions as to the validity of the methods and the reliability of their outcomes. "[B]oth methods [were] standard in the literature for the analysis of racially polarized voting."[89] The Court observed that "the data reflected positive relationships and . . . the correlations did not happen

[83] *Id.* at 1052.

[84] 481 U.S. 279, 286, 312–13 (1987).

[85] 446 U.S. 55, 74 (1980).

[86] 412 U.S. 755, 765–66 (1973).

[87] 478 U.S. 30, 35, 73 (1986).

[88] *Id.* at 52–53 (footnote omitted).

[89] *Id.* at 53 n.20.

by chance."[90] Accordingly, the Court endorsed the district court's conclusion that "the correlation between the race of the voter and the voter's choice of certain candidates was statistically significant."[91]

With the support of the United States as *amicus curiae*, North Carolina challenged this statistical approach. The state contended that "bivariate statistical analyses which merely demonstrated a *correlation* between the race of the voter and the level of voter support for certain candidates" failed "as a matter of law" to "prove that race was the primary determinant of voters' choices."[92] Rather, the government appellants argued that "only multiple regression analysis, which can take account of other variables which might also explain voters' choices," could adequately handle the effect of factors such as "party affiliation, age, religion, income[,] incumbency, education, campaign expenditures," or even the "distance that a candidate lived from a particular precinct" and thereby "prove that race was the primary determinant of voter behavior."[93]

The Court rejected what it considered the equivalent of a "new intent test" demanding that "a *specific factor*—racial hostility—*determined* white voters' ballots" through a "demonstrat[ion] that other potentially relevant *causal factors*, such as socioeconomic characteristics and candidate expenditures, do not correlate better than racial animosity with white voting behavior."[94] In so concluding, the Court quoted the "prohibitive[]" cost and nearly insurmountable operational challenges of rendering these factors "as interval-level independent variables for use in a multiple regression equation."[95]

2. *Employment Discrimination*: Bazemore v. Friday

The contemporaneous 1986 case of *Bazemore v. Friday* squarely presented the question of whether "a regression analysis [may] be treated as probative evidence of discrimination where the analysis does not incorporate every conceivable relevant variable."[96] *Bazemore*

[90] *Id.* at 53 n.22.

[91] *Id.* at 53.

[92] *Id.* at 61.

[93] *Id.* at 61–62 (internal quotation marks omitted).

[94] *Id.* at 72.

[95] *Id.* at 73 (quoting Peyton McCrary, *Discriminatory Intent: The Continuing Relevance of "Purpose" Evidence in Vote-Dilution Lawsuits*, 28 HOW. L.J. 463, 492 (1985)).

[96] 478 U.S. 385, 393 n.4 (1986) (Brennan, J., concurring). *Bazemore* generated an unusual configuration of opinions. With respect to the use of regression analyses to establish racially discriminatory patterns and practices, the Court adopted

centered upon allegations that the North Carolina Agricultural Extension Service, before creating a unitary system in 1965, "maintained two separate, racially segregated branches and paid black employees less than white employees."[97]

Bazemore focused "heavily on multiple regression analyses designed to demonstrate that blacks were paid less than similarly situated whites."[98] The Court asked whether this sort of "expert statistical evidence" could be admitted to "establish by a preponderance of the evidence that racial discrimination was the company's standard operating procedure—the regular rather than the unusual practice."[99]

The *Bazemore* plaintiffs' "regressions used four independent variables—race, education, tenure, and job title"—to predict each employee's salary.[100] Similar regressions conducted by the extension service added sex and job performance; it is unclear whether an "experience" variable materially differed from the plaintiffs' "tenure" variable.[101] Statistically significant applications of these "regressions purported to demonstrate that in 1974 the average black employee earned $331 less per year than a white employee with the same job title, education, and tenure," and $395 less in 1975.[102]

The Fourth Circuit rejected these regression analyses for two reasons. First, the data included "salary figures which reflect the effect" of discrimination before the passage of the Civil Rights Act of 1964.[103] Second, the court of appeals criticized the plaintiffs' regression analysis for failing "to consider county-to-county differences in salary increases."[104] That omission, argued the lower court, fatally undermined the aspiration of an "appropriate regression

"the reasons stated in the concurring opinion of Justice Brennan," which all other Justices joined. *Id.* at 386 (per curiam). With respect to the North Carolina Agricultural Extension Service's obligations to desegregate 4-H and Extension Homemaker Clubs, the Court adopted Justice White's concurrence for five Justices. *See id.* at 387–88 (per curiam). In a separate partial dissent, Justice Brennan objected to the remedial treatment of the extension service's educational programs. *See id.* at 409 (Brennan, J., dissenting in part). *Bazemore*'s concurrences, taken in concert, effectively constituted the opinion of the Court.

[97] *Id.* at 394 (Brennan, J., concurring).

[98] *Id.* at 398.

[99] *Id.* at 397–98 (internal quotation marks omitted) (quoting Int'l Brotherhood of Teamsters v. United States, 431 U.S. 324, 336 (1977)).

[100] *Id.* at 398.

[101] *Id.*

[102] *Id.* at 399.

[103] *Id.*

[104] *Id.*

analysis of salary" to "include *all* measurable variables thought to have an effect on salary level."[105] In so reasoning, the Fourth Circuit attacked the use of regression in *Bazemore* on twin grounds of overinclusion and underinclusion, the legal analogues to Type I (*alpha*) and Type II (*beta*) errors in statistical hypothesis testing.[106]

The Supreme Court swiftly disposed of the lower court's objection to the inclusion of "salary disparities created prior to 1972 and perpetuated thereafter."[107] Relying on wholly legal reasoning rather than their understanding of statistical methodology, the Justices deemed this evidence relevant because "hold[ing] otherwise would have the effect of exempting" discrimination before the passage of Title VII.[108]

The Court's response to the alleged Type II error arising from the omission of possibly relevant variables was more comprehensive and more consequential in its impact on future judicial treatment of regression analyses. Despite conceding that "the omission of variables from a regression analysis may render the analysis less probative than it otherwise might be," the Court ruled that the categorical exclusion of an underinclusive regression from evidence as "plainly incorrect."[109]

This portion of *Bazemore* identified a common concern in statistical analysis: omitted variable bias.[110] A more sophisticated fixed-effects model could address concerns over county-specific differences by treating each North Carolina county as a statistically

[105] *Id.*

[106] *See* R.S. Radford, *Statistical Error and Legal Error: Type One and Type Two Errors and the Law*, 21 LOY. L.A. L. REV. 843, 851–55 (1988); Mario J. Rizzo & Frank S. Arnold, *An Economic Framework for Statutory Interpretation*, 50(4) L. & CONTEMP. PROBS. 165, 168–69 (Fall 1987). *See generally* Matthew D. Lieberman & William A. Cunningham, *Type I and Type II Error Concerns in fMRI Research: Re-Balancing the Scale*, 4 SOC. COGNITIVE & AFFECTIVE NEUROSCIENCE 423 (2009); R. Lilford & N. Johnson, *The Alpha and Beta Errors in Randomized Trials*, 322 NEW ENG. J. MED. 780 (1990); Saskia Litière, Ariel Alonso & Geert Molenberghs, *Type I and Type II Error Under Random-Effects Misspecification in Generalized Linear Mixed Models*, 63 BIOMETRICS 1038 (2007).

[107] *Bazemore*, 478 U.S. at 395 (Brennan, J., concurring).

[108] *Id.*

[109] *Id.* at 400.

[110] *See, e.g.*, Carlos Cinelli & Chad Hazlett, *Making Sense of Sensitivity: Extending Omitted Variable Bias*, 82 J. ROYAL STAT. SOC'Y 39 (2020); Kevin A. Clarke, *The Phantom Menace: Omitted Variable Bias in Econometric Research*, 22 CONFLICT MGMT. & PEACE SCI. 341 (2005): Kevin A. Clarke, *Return of the Phantom Menace: Omitted Variable Bias in Political Research*, 26 CONFLICT MGMT. & PEACE SCI. 46 (2009).

distinct entity.[111] Notably, Justice Marshall's *Gregg* dissent had objected to a similar failure to account for state-by-state differences in the deterrent effect of capital punishment. [112]

Crucially, the Court observed that the "failure to include variables will [normally] affect the analysis' probativeness, not its admissibility."[113] *Bazemore* did not involve the extreme case of a "regression[] so incomplete as to be inadmissible as irrelevant."[114] Instead, the Court recognized "that a regression analysis that includes [fewer] than 'all measurable variables' may serve to prove a plaintiff's case."[115] Whether such an incomplete regression analysis discharges the plaintiff's "burden . . . to prove discrimination by a preponderance of the evidence" ultimately hinges "on the factual context of each case in light of all the evidence" presented by both parties.[116]

Bazemore's treatment of variables that arguably should be included in a regression analysis forecloses the feckless and innumerate strategy of declaring that "many factors" affect a contested decision such as "an individual employee's salary."[117] At the very least, *Bazemore* compels a party contesting the validity of regression to make a genuine "attempt . . .— statistical or otherwise—to demonstrate" that a proper evaluation of the relevant factors would leave "no significant disparity" warranting legal attention or sanction.[118]

Bazemore was the relatively unusual case in which regression analysis placed greater emphasis on "the computed value of the dependent variable" than on "the value of the[] coefficients" of independent variables.[119] The specific allegations of racial

[111] *See, e.g.*, Jushan Bai, *Panel Data Models with Interactive Fixed Effects*, 77 ECONOMETRICA 1229 (2009); Larry V. Hedges, *Fixed Effects Models*, *in* THE HANDBOOK OF RESEARCH SYNTHESIS 285–300 (Harris Cooper & Larry V. Hedges, eds. 1994). *See generally* PAUL D. ALLISON, FIXED EFFECTS REGRESSION MODELS (2009). A different article in this issue makes extensive use of fixed effects models to isolate country-specific effects in an evaluation of tariff policies around the world. *See* James Ming Chen, Thomas Poufinas, Charalampos Agiropoulos & George Galanos, *Principles of Political Economy and the Taxation of Nations: Econometric and Machine-Learning Evaluation of Tariffs*, 2020 MICH. ST. L. REV. 1361 (2020).

[112] Gregg v. Georgia, 428 U.S. 153, 234–36 (1976) (Marshall, J., dissenting).

[113] *Bazemore*, 478 U.S. at 400 (Brennan, J., concurring).

[114] *Id.* at 400 n.10.

[115] *Id.* at 400.

[116] *Id.*; *see also* Tex. Dep't of Cmty. Affs. v. Burdine, 450 U.S. 248, 252 (1981).

[117] *Bazemore*, 478 U.S. at 403 n.14 (Brennan, J., concurring).

[118] *Id.*

[119] Finkelstein, *supra* note 5, at 1445.

discrimination in that case make it easier to understand the shift in emphasis. The state agricultural extension service had admittedly been segregated *de jure* before 1965. The dispute hinged on pay disparities by race, persisting past the creation of a single, unitary service through the union of formerly segregated branches. Regression as a descriptive or even prescriptive tool can quantify the difference in pay in tangible, as in the $331 or $395 reduction in annual salary borne by the black agricultural extension employees in *Bazemore*.

3. *Death Penalty:* McCleskey v. Kemp

A difference in the perception (if not the predictive accuracy) of regression analysis dominated *McCleskey v. Kemp*.[120] *McCleskey* is arguably the most important case in which the Supreme Court reviewed this class of statistical methods. Warren McCleskey, a black man, had been sentenced to death in Georgia for the murder of a white police officer during the course of an armed robbery.[121] McCleskey's Eighth and Fourteenth Amendment challenges to his sentence rested on "the Baldus study," which "examine[d] over 2,000 murder cases that occurred in Georgia during the 1970's."[122]

After "subject[ing] his data to an extensive analysis [and] taking account of 230 variables that could have explained" racial disparities in capital sentencing "on nonracial grounds," David Baldus concluded "that, even after taking account of 39 nonracial variables, defendants charged with killing white victims were 4.3 times as likely to receive a death sentence as defendants charged with killing blacks."[123] That model also concluded that "black defendants were 1.1 times as likely to receive a death sentence as other defendants."[124] The study concluded that "black defendants, such as McCleskey, who kill white victims have the greatest likelihood of receiving the death penalty."[125]

McCleskey distinguished capital sentencing, "and the relationship of statistics to that decision," from jury "venire-selection [and] Title VII cases" in which the Court had upheld resort to

[120] 481 U.S. 279 (1987).

[121] *See id.* at 283–85.

[122] *Id.* at 286. David Baldus and his coauthors published their work as David C. Baldus, Charles Pulaski, & George Woodworth, *Comparative Review of Death Sentences: An Empirical Study of the Georgia Experience*, 74 J. CRIM. L. & CRIMINOLOGY 661 (1983).

[123] *McCleskey*, 481 U.S. at 287.

[124] *Id.*

[125] *Id.*

statistical evaluation in general and regression analysis in particular.[126] Treating each capital jury as a unique deliberative body taking account "of innumerable factors that vary according to the characteristics of the individual defendant and the facts of the particular capital offense," the Court denied the validity "of an inference drawn from . . . general statistics to a specific decision in a trial and sentencing."[127] The Court criticized the absence of a "practical opportunity" for the state "to rebut the Baldus study" and McCleskey's statistical arguments.[128]

Demanding (but not finding) "exceptionally clear proof" as a precondition to overturning the "discretionary judgments" within the state's "criminal justice process," the Court held "that the Baldus study is clearly insufficient to support an inference that any of the decisionmakers in McCleskey's case acted with discriminatory purpose."[129] The Court rejected McCleskey's even broader suggestion "that the Baldus study proves that the State as a whole has acted with [the] discriminatory purpose" of "enact[ing] or maintain[ing] the death penalty . . . *because of* an anticipated racially discriminatory effect."[130]

The Court also rejected McCleskey's reliance on the Baldus study for the proposition that "the Georgia capital punishment system [was] arbitrary and capricious in *application*" insofar as "racial considerations" led the state to impose a sentence upon him that was disproportionately harsh relative to other murder cases.[131] Observing that "[e]ven Professor Baldus does not contend that his statistics *prove* that race enters into any capital sentencing decisions," the Court declared that "[s]tatistics at most may show only a likelihood that a particular factor entered into some decisions."[132] The Court ultimately declined "to accept the likelihood allegedly shown by the Baldus study as the constitutional measure of an unacceptable risk of racial prejudice influencing capital sentencing decisions."[133]

"At most," the Court concluded, "the Baldus study indicates a discrepancy that appears to correlate with race."[134] That quantum of

[126] *Id.* at 294.

[127] *Id.*

[128] *Id.* at 296.

[129] *Id.* at 297.

[130] *Id.* at 297–98.

[131] *Id.* at 308.

[132] *Id.*

[133] *Id.* at 309.

[134] *Id.* at 312.

statistical evidence, the Court held, did "not demonstrate a constitutionally significant risk of racial bias affecting the Georgia capital sentencing process."[135] This conclusion rested heavily upon David Baldus's confession, consistent with scientific norms, that statistics "obviously . . . cannot say . . . to a moral certainty" that race dictated the outcome of McCleskey's trial or sentencing.[136] If anything, the creativity and reach of social science restrained the Court from adopting a rationale allowing constitutional challenges, "at least in theory," to "be based upon any arbitrary variable, such as the defendant's facial characteristics, or the physical attractiveness of the defendant or the victim," upon the delivery of "some statistical study" finding "influen[ce] in jury decisionmaking."[137]

Justice Blackmun's dissent in *McCleskey* emphasized the inconsistencies between the majority's treatment of capital sentencing and other legal settings, such as jury selection and employment, "in which [the Court] long has accepted statistical evidence and has provided an easily applicable framework for review."[138] He rejected the Court's assertion that those settings differed materially and in kind rather than degree from capital sentencing because there allegedly "are fewer [relevant] variables" and because the applicable "statistics relate to fewer entities."[139]

Justice Brennan's dissent even more forcefully defended the Baldus study. "McCleskey's statistics have particular force," he wrote, "because most of them are the product of sophisticated multiple-regression."[140] "[D]esigned precisely to identify patterns in the aggregate," multivariable regression "is particularly well suited to identify the influence of impermissible [legal] considerations," even where judges are unable "to reconstitute with certainty any individual decision that goes to make up [an unlawful] pattern."[141]

Justice Brennan recited the central lesson of *Bazemore v. Friday*.[142] A "multiple-regression analysis need not include every conceivable variable," he observed, "as long as it includes those variables that account for the major factors that are likely to influence

[135] *Id.* at 313.

[136] *Id.* at 308 n.29.

[137] *Id.* at 317–18 (footnotes omitted).

[138] *Id.* at 350 (Blackmun, J., dissenting).

[139] *Id.* at 362.

[140] *Id.* at 327 (Brennan, J., dissenting).

[141] *Id.*

[142] 478 U.S. 385 (1986).

decisions."[143] Justice Brennan concluded that Baldus's "statistical evidence . . . thus relentlessly documents the risk that McCleskey's sentence was influenced by racial considerations."[144] In a bitter but memorable Parthian volley, Justice Brennan described the *McCleskey* majority's fear of "widespread challenges to all aspects of criminal sentencing" as motivated by "a fear of too much justice."[145]

D. Phase Three: A Decision Tree Grows in Washington—The Supreme Court's Regression Jurisprudence Comes of Age

Supreme Court cases since *McCleskey* have never fully bridged the racially charged divide exposed by that decision. By the same token, regression survived *McCleskey*'s rejection of the Baldus study. In certain settings, and perhaps only among certain Justices, this basic social science tool retains a critical kernel of legal acceptance.

1. *Racial Disparities Reconsidered*

As the most dangerous of the Justices' encounters with regression analysis, *McCleskey v. Kemp* may have represented the high-water mark of the Supreme Court at the bar of statistics.[146] That case pitted some of "the best empirical studies on criminal sentencing ever conducted" against lower court judges who openly "revealed contempt" for "rinky-dink regressions that accounted for only a few variables" and "prove[d] nothing other than . . . the adage that anything may be proved by statistics."[147]

McCleskey's impact may be muted. To be sure, much of the academic commentary has lamented *McCleskey*'s barrier to judicial acceptance of statistical evidence of discriminatory purpose.[148] Lower

[143] *McCleskey*, 481 U.S. at 327–28 (Brennan, J., dissenting).

[144] *Id.* at 328.

[145] *Id.* at 339.

[146] *See* Edelman & Chen, *supra* note 30.

[147] Randall L. Kennedy, McCleskey v. Kemp*: Race, Capital Punishment, and the Supreme Court*, 101 HARV. L. REV. 1388, 1400 & n.45 (1988) (quoting McCleskey v. Kemp, No. C87-1517A, at 12 (N.D. Ga. Dec. 23, 1987)).

[148] *See, e.g.*, Henry Louis Gates, *Statistical Stigmata*, 11 CARDOZO L. REV. 1275, 1282–83 (1990); Sharad Goel, Maya Perelman, Ravi Shroff & David Alan Sklansky, *Combatting Police Discrimination in the Age of Big Data*, 20 NEW CRIM. L. REV. 181, 199–200 (2017); Aziz Z. Huq, *The Consequences of Disparate Policing: Evaluating Stop and Frisk as a Modality of Urban Policing*, 101 MINN. L. REV. 2397, 2453–54 (2017); Michael Selmi, *Statistical Inequality and Intentional (Not Implicit) Discrimination*, 79 LAW & CONTEMP. PROBS., no. 3, 2016, at 199, 213 n.74.

courts, however, appear to have seized upon Justice Powell's observation that McCleskey "relie[d] *solely* on the Baldus study."[149] Courts routinely admit statistical evidence resembling the Baldus study in these circumstances:

(1) where statistics are paired with circumstantial or other nonstatistical evidence,[150]

(2) where statistical evidence of discrimination focuses on a single prosecutor or law enforcement officer,[151] or

(3) cases not involving capital punishment or prosecutorial discretion.[152]

Moreover, the Court's subsequent jurisprudence on racially selective prosecution suggests—albeit in a backhanded way—that the judiciary might not blind itself entirely to statistical evidence of racial bias in criminal justice. Unexpected, indirect support for regression emerged in *United States v. Armstrong*.[153] This 1996 case pegged the "requirements for a selective-prosecution claim" to "ordinary equal protection standards" wholly in harmony with *McCleskey*.[154] Specifically, the "claimant must demonstrate that the federal prosecutorial policy 'had a discriminatory effect and that it was motivated by a discriminatory purpose'" by showing "that similarly situated individuals of a different race were not prosecuted."[155]

Although *Armstrong* did not directly address statistical methodology, it at least implicitly rebutted any suggestion that *McCleskey* forecloses the consideration of statistical evidence in cases alleging purposeful discrimination. *Armstrong* rejected "the presumption that people of all races commit all types of crimes," a baseline wholly at odds with "the premise that any type of crime is the exclusive province of any particular racial or ethnic group."[156] This

[149] McCleskey v. Kemp, 481 U.S. 279, 293 (1987) (emphasis added); *accord* Reva B. Siegel, *Blind Justice: Why the Court Refused to Accept Statistical Evidence of Discriminatory Purpose in* McCleskey v. Kemp*—and Some Pathways for Change*, 112 NW. U. L. REV. 1269, 1288 n.113 (2018).

[150] *See, e.g.*, Chavez v. Ill. State Police, 251 F.3d 612, 647–48 (7th Cir. 2001); Floyd v. City of N.Y., 959 F. Supp. 2d 540, 603 (S.D.N.Y. 2013).

[151] *See, e.g.*, Belmontes v. Brown, 414 F.3d 1094, 1127 (9th Cir. 2005); Jefferson v. Terry, 490 F. Supp. 2d 1261, 1340 (N.D. Ga. 2007).

[152] *See, e.g.*, Mehta v. Vill. of Bolingbrook, 196 F. Supp. 3d 855, 863–64 (N.D. Ill. 2016); Smith v. City of Chi., 143 F. Supp. 3d 741, 754–56 (N.D. Ill. 2015); *Floyd*, 959 F. Supp. 2d at 562. *See generally* Siegel, *supra* note 149, at 1288 & nn.113–15 (reviewing all three of these types of cases).

[153] 517 U.S. 456 (1996).

[154] *Id.* at 465 (quoting United States v. Wayte, 470 U.S. 598, 608 (1985)).

[155] *Id.*

[156] *Id.* at 469 (internal quotations omitted).

proposed analytical starting point was "contradicted by the most recent statistics" showing that "[m]ore than 90% of the persons sentenced in 1994 for crack cocaine trafficking were black, 93.4% of convicted LSD dealers were white, and 91% of those convicted for pornography or prostitution were white."[157] "Presumptions at war with presumably reliable statistics," *Armstrong* concluded, "have no proper place in the analysis of this issue."[158]

Despite erecting another formidable barrier to private individuals alleging racial bias in law enforcement, *Armstrong* did rehabilitate the idea that statistical analysis may reveal racial disparities in criminal justice. In its willingness to accord dispositive legal weight to racial differences, *Armstrong* is closer in spirit to *Thornburg v. Gingles* and *Bazemore v. Friday* than *McCleskey v. Kemp*. Indeed, the basic racial statistics invoked in Armstrong are suitable for analysis through the bivariate approaches upheld in *Gingles*.

2. *Class Certification*

In the years since *McCleskey*, the Supreme Court has questioned the value of regression analysis in one other major body of legal doctrine: certification of a class under Federal Rule of Civil Procedure 23. This skein of cases began in 2011 with *Wal–Mart Stores, Inc. v. Dukes*.[159] *Dukes* involved "one of the most expansive class actions ever," a "class comprising about one and a half million plaintiffs, current and former female employees of . . . Wal–Mart who allege[d] that the discretion exercised by their local supervisors over pay and promotion matters" discriminated against them as women.[160]

These Wal–Mart employees sought certification under Rule 23(b)(2), which permits certification when "the party opposing the class has acted or refused to act on grounds that apply generally to the class, so that final injunctive relief or corresponding declaratory relief is appropriate respecting the class as a whole."[161] The employees also sought to satisfy Rule 23(a)'s requirement of "questions of law or fact common to the class."[162] In support of both claims, the employees submitted "statistical evidence about pay and promotion disparities

[157] *Id.* (internal citations omitted).

[158] *Id.* at 469–70.

[159] 564 U.S. 338 (2011).

[160] *Id.* at 342.

[161] FED. R. CIV. P. 23(b)(2); *see Dukes*, 564 U.S. at 345–46.

[162] FED. R. CIV. P. 23(a)(2).

between men and women" as well as "the testimony of a sociologist, Dr. William Bielby, who conducted a 'social framework analysis' of Wal–Mart's 'culture' and personnel practices" and found the company "'vulnerable' to gender discrimination."[163]

The two bodies of expert testimony met distinct but ultimately negative fates. The Court pounced upon Bielby's concession "that he could not calculate whether 0.5 percent or 95 percent of the employment decisions at Wal–Mart might be determined by stereotyped thinking."[164] The Court questioned whether the sociological "framework analysis" might even have qualified as admissible expert testimony under Federal Rule of Evidence 702 and *Daubert*.[165] The sociologist's inability to quantify the impact of stereotyped thinking on Wal–Mart's employment practices allowed the Court to "safely disregard what he has to say."[166] Bielby's social framework analysis "provide[d] no verifiable method for measuring and testing any of the variables that were crucial to his conclusions."[167]

The *Dukes* plaintiffs' statistical evidence, by contrast, "consist[ed] primarily of regression analyses performed by" a statistician and a labor economist.[168] The statistician's analysis of regional and national data "compar[ed] the number of women promoted into management positions with the percentage of women in the available pool of hourly workers."[169] It found "statistically significant disparities between men and women at Wal–Mart" and could explain those disparities "only by gender discrimination."[170] The labor economist's regression analysis "concluded that Wal–Mart 'promotes a lower percentage of women than its competitors.'"[171]

The Court rejected this proffer on grounds reminiscent of *McCleskey*. Justice Scalia, writing for the majority, refused to infer "uniform, store-by-store disparity," or even disparity at the district level, from findings made at the regional and national levels.[172] Even

[163] *Dukes*, 564 U.S. at 346.

[164] *Id.* at 354.

[165] *Id.*; *see* FED. R. EVID. 702; Daubert v. Merrell Dow Pharms., Inc., 509 U.S. 579 (1993).

[166] *Dukes*, 564 U.S. at 354–55.

[167] John Monahan, Laurens Walker & Gregory Mitchell, *Contextual Evidence of Gender Discrimination: The Ascendance of "Social Frameworks*,*"* 94 VA. L. REV. 1715, 1747 (2008); *accord Dukes*, 564 U.S. at 354 n.8.

[168] *Dukes*, 564 U.S. at 356.

[169] *Id.*

[170] *Id.*

[171] *Id.*

[172] *Id.* at 357.

more fundamentally, he wrote, even a finding of "a pay or promotion pattern" across "*all* of Wal–Mart's 3,400 stores . . . would still not demonstrate" the necessary "commonality of issue."[173] In light of the virtual certainty that individual store managers "will claim to have been applying some sex-neutral, performance-based criteria," the mere demonstration "that Wal–Mart's policy of discretion has produced an overall sex-based disparity [did] not suffice."[174]

To no avail, Justice Ginsburg defended the plaintiffs' "regression analyses" as "sufficient to raise an 'inference of discrimination.'"[175] She decried the "majority's contention to the contrary" as a misguided and "arcane disagreement about statistical method—which the District Court resolved in the plaintiffs' favor."[176]

Controversy over class certification returned two terms later, in the 2013 antitrust case of *Comcast Corp. v. Behrend*.[177] The same configuration of Justices rejected another regression analysis offered in support of a class certification. Cable television subscribers alleged that Comcast concentrated its holdings of cable systems in the Philadelphia area by swapping its holdings elsewhere with other cable holding companies.[178] These practices collectively constituted "'clustering,' a strategy of concentrating operations within a particular region."[179]

The plaintiffs described themselves as "subscribers in the Philadelphia cluster" who had been harmed by Comcast's elimination of competition and maintenance of "prices for cable services above competitive levels."[180] They sought certification under Federal Rule of Civil Procedure 23(b)(3).[181] The Court described that rule as a "demanding," even "adventuresome innovation" for situations not "clearly call[ing] for" class-action treatment."[182]

Justice Scalia's majority opinion concluded that the *Behrend* plaintiffs had failed to prove that "questions of law or fact common to class members predominate over any questions affecting only

[173] *Id.*

[174] *Id.*

[175] *Id.* at 372 (Ginsburg, J., dissenting).

[176] *Id.* at 372 n.5.

[177] 569 U.S. 27 (2013).

[178] *See id.* at 29–30.

[179] *Id.* at 29.

[180] *Id.* at 30.

[181] FED. R. CIV. P. 23(b)(3).

[182] *Behrend*, 569 U.S. at 34 (quoting Wal–Mart Stores, Inc. v. Dukes, 564 U.S. 338, 362 (2011) (quoting Amchem Prods., Inc. v. Windsor, 521 U.S. 591, 614–15, 623–24 (1997))).

individual members."[183] The cable subscribers "relied solely on . . . a regression model comparing actual cable prices in . . . Philadelphia . . . with hypothetical prices that would have prevailed but for [Comcast's] allegedly anticompetitive activities."[184] This study sharply distinguished between the two primary tasks of regression analysis: prediction and correlative inference. Although the model was able to "calculate[] damages of $875,576,662 for the entire class," it "did not isolate damages resulting from any one theory of antitrust impact."[185]

The majority thought it fatal that the model "calculated damages resulting from 'the alleged anticompetitive conduct as a whole'" without "attribut[ing] damages to any one particular theory of anticompetitive impact."[186] Justice Scalia conceded that "[t]his methodology might have been sound, and might have produced commonality of damages, if all four . . . alleged [market] distortions [had] remained" available as theories of antitrust liability.[187]

But the district court had foreclosed three of the hypotheses as a matter of law. The "model's inability to bridge the differences between supra-competitive prices in general and supra-competitive prices attributable to" the only remaining plausible theory of antitrust liability prevented the certification of this plaintiff class under Rule 23(b)(3).[188]

Justice Ginsburg again dissented. She decried the Court's willingness to "consider fact-based matters, namely what this econometric multiple-regression model is about, what it proves, and how it does so"—all matters better consigned to the "two lower courts' related factual findings to the contrary."[189] In her view, the model had achieved its more modest goal of showing the mere fact that "Comcast's conduct brought about higher prices," without "purport[ing] to show precisely *how*" those higher prices came about.[190] She argued that the Court should have dismissed the writ of certiorari as improvidently granted and given the lower courts freedom to conduct their own "underlying considerations [of] detailed,

[183] FED. R. CIV. P. 23(b)(3); *accord Behrend*, 569 U.S. at 34.

[184] *Behrend*, 569 U.S. at 31–32.

[185] *Id.* at 32.

[186] *Id.* at 36–37.

[187] *Id.* at 37.

[188] *Id.* at 38.

[189] *Id.* at 46 (Ginsburg, J., dissenting).

[190] *Id.* at 48.

technical, and fact-based" disputes over antitrust injury and the regression models designed to measure it.[191]

Unlike *Bazemore v. Friday*, *Behrend* delivered a concrete prediction of antitrust damages in a setting where a majority of the Justices had emphasized the other contribution of regression to legal decisionmaking: the attribution of results to independent predictors through the size and scale of coefficients.[192] The Court implicitly demands a sort of evaluation similar to sensitivity analysis and Lagrangian multipliers in linear programming and other branches of operations research.[193] Sensitivity analysis describes the impact of marginal changes in the coefficients of an objective function and on the "left-" and "right-hand-side" of operational constraints.[194] Sensitivity analysis can also inform regression models.[195] Federal courts are familiar with this tool.[196]

3. *Sense and Sensitivity*

The vast majority of other Supreme Court cases after *McCleskey* suggest that most Justices have reached a stable though occasionally grudging *détente* with the probative claims of regression analysis. In voting rights cases, Some Justices evidently regret the Court's

[191] *Id.*

[192] *See* Finkelstein, *supra* note 5, at 1445.

[193] *See, e.g.*, Ashok D. Belegundu, *Lagrangian Approach to Design Sensitivity Analysis*, 111 J. ENG'G MECHS. 680 (1985); Emanuele Borgonovo & Elmar Plischke, *Sensitivity Analysis: A Review of Recent Advances*, 248 EUR. J. OPERATIONAL RSCH. 869 (2016); Frank H. Clarke, *A New Approach to Lagrange Multipliers*, 1 MATHEMATICS OPERATIONS RSCH. 165 (1976); Anthony V. Fiacco, *Sensitivity Analysis for Nonlinear Programming Using Penalty Methods*, 10 MATHEMATICAL PROGRAMMING 287 (1976). *See generally* EMANUELE BORGONOVO, SENSITIVITY ANALYSIS: AN INTRODUCTION FOR THE MANAGEMENT SCIENTIST (Springer 2017).

[194] *See, e.g.*, Alireza Ghaffari Hadigheh, Oleksandr Romanko & T. Terlaky, *Sensitivity Analysis in Convex Quadratic Optimization: Simultaneous Perturbation of the Objective and Right-Hand-Side Vectors*, 2 ALGORITHMIC OPERATIONS RSCH. 94 (2007).

[195] *See, e.g.*, H. Christopher Frey & S.R. Patil, *Identification and Review of Sensitivity Analysis Methods*, 22 RISK ANALYSIS 553 (2002); Roy E. Welsch, *Regression Sensitivity Analysis and Bounded-Influence Estimation*, *in* EVALUATION OF ECONOMETRIC MODELS 153–67 (Jan Kmenta & James B. Ramsey eds., 1980). *See generally* SAMPRIT CHATTERJEE & ALI S. HADI, SENSITIVITY ANALYSIS IN LINEAR REGRESSION (John Wiley & Sons 2009).

[196] *See, e.g.*, *In re* Lamictal Direct Purchaser Antitrust Litig., 957 F.3d 184, 193 n.4 (3d Cir. 2020); Tadros v. Celladon Corp., 738 F. App'x 448, 448 (9th Cir. 2018); WMI Holdings Corp. v. United States, 891 F.3d 1016, 1027 (Fed. Cir. 2018).

endorsement in *Thornburg v. Gingles* of "extreme case analysis and bivariate ecological regression analysis."[197] These simpler methods of bivariate regression. more succinctly summarized as "bivariate regression analysis," "measure[] merely the correlation between race and candidate preference" without "directly control[ling] for other factors."[198] Bivariate regression analysis, as Justice Thomas has disapprovingly noted, has "become the norm for determining cohesion in vote dilution cases."[199] In a similar spirit, Chief Justice Roberts has lamented his colleagues' rejection of "the District Court's parsing of the statistical evidence" and the "typical[]" presentation of "regression analyses of past voting records" as evidence of "an effective Latino opportunity district," notwithstanding what he considered spurious evidence of noncompactness.[200]

More generally, Justices who have questioned the validity of regression analyses in settings such as class certification have not hesitated to question the inadequacy of statistical models brought before the Court, relative to hypothetical or even idealized models incorporating more variables and more data. Skepticism toward the utility of regression, stern enough to block its application in death penalty and class certification disputes as a matter of law, can evidently yield to an opportunistic embrace of the method's predictive and explanatory potential. A sudden appreciation for methodological thoroughness can enable Justices to emphasize and exploit Type II errors of underinclusion in statistical models. The faithless translation of statistical methods betrays a latent instrumental purpose to subjugate the mathematics of regression to ideological goals in law.[201] If this be cynicism, make the most of it.[202]

Dissenting from denial of certiorari in a government contracting case involving race-conscious preferences, Justice Scalia argued that

[197] Thornburg v. Gingles, 478 U.S. 30, 52–53 & n.20 (1986).

[198] Holder v. Hall, 512 U.S. 874, 904 n.13 (1994) (Thomas, J., concurring in the judgment).

[199] *Id.*

[200] League of United Latin Am. Citizens v. Perry, 548 U.S. 399, 500 (2006) (Roberts, C.J., concurring in part, concurring in the judgment in part, and dissenting in part).

[201] *See* Lawrence Lessig, *Fidelity in Translation*, 71 TEX. L. REV. 1165, 1171–73 (1993).

[202] *See* JON KUKLA, PATRICK HENRY: CHAMPION OF LIBERTY 71 (2017) ("If this be treason, make the most of it."). For iconic sources treating translation as an inherently treasonous enterprise, see Gregory Rabassa, *If This Be Treason: Translation and Its Possibilities*, 44 AM. SCHOLAR 29 (1975); GREGORY RABASSA, IF THIS BE TREASON: TRANSLATION AND ITS DYSCONTENTS (2005).

"the government should have been required to produce a regression analysis controlling for [nonracial] factors if it wished to rely on statistical disparities."[203] He dismissed claims to "the relationship between minority ownership and size-and-experience in the Denver construction industry" because the "disparity studies" that the city had conducted "did not address those variables."[204]

Alongside Justice Scalia, author of the class certification cases that were so skeptical of regression analyses, Justice Thomas has expressed grave misgivings over this method. An unusual variation on the theme of *McCleskey v. Kemp* arose in the 2015 case of *Glossip v. Gross*.[205] *Glossip* involved the legal significance of regression analysis of the factors most likely to influence the imposition of capital punishment. Dissenting from the Court's affirmance of a death sentence,[206] Justice Breyer invoked empirical studies.[207] In Justice Thomas's words, those studies showed "that the primary explanation . . . for the gap between . . . egregiousness scores" assigned to a crime "and the actual sentences was not the race or sex of the offender or victim, but the locality in which the crime was committed."[208]

Justice Thomas condemned "these studies [as] inherently unreliable because they purport to control for egregiousness by quantifying moral depravity in a process that is itself arbitrary, not to mention dehumanizing."[209] Justice Thomas took exception to one study that "assigned 'depravity points' to identify the 'worst of the worst' murderers."[210] "We owe victims more than this sort of pseudoscientific assessment of their lives," he concluded.[211]

[203] Concrete Works of Colo., Inc. v. City & Cnty. of Denver, 540 U.S. 1027, 1032 (2003) (Scalia, J., dissenting from denial of certiorari to 321 F.3d 950 (10th Cir. 2003)).

[204] *Id.*

[205] *Compare* Glossip v. Gross, 576 U.S. 863 (2015), *with* McCleskey v. Kemp, 481 U.S. 279 (1987).

[206] *See Glossip*, 576 U.S. at 918–19 (Breyer, J., dissenting).

[207] *See* John J. Donohue III, *An Empirical Evaluation of the Connecticut Death Penalty System Since 1973: Are There Unlawful Racial, Gender, and Geographic Disparities?*, 11 J. EMPIRICAL LEGAL STUD. 637, 640 (2014); Robert J. Smith, *The Geography of the Death Penalty and Its Ramifications*, 92 B.U. L. REV. 227, 231–32 (2012).

[208] *Glossip*, 576 U.S. at 903 (Thomas, J., concurring).

[209] *Id.*

[210] *Id.* (discussing David McCord, *Lightning Still Strikes: Evidence from the Popular Press That Death Sentencing Continues to Be Unconstitutionally Arbitrary More Than Three Decades After* Furman, 71 BROOK. L. REV. 797, 833–34 (2005)).

[211] *Id.* at 904.

Justice Thomas more directly attacked regression methodology in the 2019 case of *Flowers v. Mississippi*.[212] Allegedly discriminatory peremptory challenges to jurors took center stage. Justice Thomas criticized the majority for simplistically assuming "that the only relevant difference between [struck] black jurors . . . and seated white jurors is their race."[213] He insisted that a proper regression analysis should "control[] for other potentially relevant variables" before it can "even begin to provide probative evidence of discrimination."[214] "Indeed," he concluded, "it is difficult to conceive of a statistical study that could possibly control for all of the relevant variables in this context, including tone of voice, facial expressions, and other relevant information."[215] This criticism echoed a criticism voiced by Richard Posner: "a statistical study that fails to correct for salient explanatory variables, or even to make the most elementary comparisons, has no value as causal explanation."[216]

That both *Glossip* and *Flowers* involved criminal justice should not be particularly surprising. The Court as a whole has shown extreme discomfort over reliance on statistical evidence in death penalty cases and other criminal controversies. For every skeptic such as Justice Thomas in *Glossip*, Justices such as Justice Sotomayor invoke multivariable regressions showing a statistically significant correlation between election years and judicial decisions to impose the death penalty by overriding a lesser penalty imposed by a jury.[217] The arc of death penalty jurisprudence from *Furman* to the *Gregg* quintet and *McCleskey* has not perceptibly bent: Capital sentencing rates, measured nationally or inside Texas as a particularly revelatory test state, expose the same sort of arbitrary fluctuation that the Court condemned half a century ago in *Furman*.[218]

Flowers does expose a curious failure by the Supreme Court to embrace the use of the binomial distribution in jury selection cases. Fisher's exact test of binomial probability offers a deterministic way

[212] 139 S. Ct. 2228 (2019).

[213] *Id.* at 2261 (Thomas, J., dissenting).

[214] *Id.* at 2262.

[215] *Id.*

[216] People Who Care v. Rockford Bd. of Educ. Sch. Dist. No. 205, 111 F.3d 528, 537 (7th Cir. 1997); *accord Flowers*, 139 S. Ct. at 2262 (Thomas, J., dissenting).

[217] *See* Woodward v. Alabama, 134 S. Ct. 405, 408–09 (2013) (Sotomayor, J., dissenting from denial of certiorari).

[218] *See* Scott Phillips & Alena Simon, *Is the Modern American Death Penalty a Fatal Lottery? Texas as a Conservative Test*, 3 LAWS 85, 94–96 (2014).

to evaluate whether binary outcomes are the product of chance.[219] The technique is not alien to American law; federal launch safety standards prescribe tests of binomial significance for the evaluation of debris risk.[220] The specter of bias along lines of race and sex haunts peremptory challenges to jurors and has vexed the law for decades.[221]

A formidable body of scholarship, however, has not yet persuaded the Supreme Court to formally prescribe the binomial theorem as a tool for evaluating this class of claims.[222] In its initial expression of "interest[]" in calculations of binomial probability, the Court considered that analysis "unnecessary to [its] disposition" of a jury selection case.[223] Though no longer exotic, this tool has fallen short of providing firm guidance for judicial decisions. Statistics and regression, "while not controlling upon the courts by reason of their authority, do constitute a body of experience and informed judgment to which courts and litigants may properly resort for guidance."[224] Like informal expressions of agency guidance in administrative law, these tools wield the "power to persuade," despite "lacking power to control."[225]

Despite occasional expressions of dissatisfaction and skepticism by some Justices, most contemporary Supreme Court cases reveal general comfort with regression analysis. By the same token, the

[219] *See* R.A. Fisher, *On the Interpretation of χ^2 from Contingency Tables, and the Calculation of P*, 85 J. ROYAL STAT. SOC'Y 87, 93–94 (1922). *See generally* DAVID SALSBURG, THE LADY TASTING TEA: HOW STATISTICS REVOLUTIONIZED SCIENCE IN THE TWENTIETH CENTURY (2002).

[220] *See* 14 C.F.R. §§ A417.25(b), C417.9(b) (2021).

[221] *See, e.g.*, *Flowers*, 139 S. Ct. 2228; Foster v. Chatman, 136 S. Ct. 1737 (2016); Snyder v. Louisiana, 552 U.S. 472 (2008); Miller-El v. Dretke, 545 U.S. 231 (2005); Johnson v. California, 545 U.S. 162 (2005); J.E.B. v. Alabama *ex rel.* T.B., 511 U.S. 127 (1994); Georgia v. McCollum, 505 U.S. 42 (1992); Edmonson v. Leesville Concrete Co., 500 U.S. 614 (1991); Holland v. Illinois, 493 U.S. 474 (1990); Batson v. Kentucky, 476 U.S. 79 (1986).

[222] *See, e.g.*, Bruce E. Barrett, *Detecting Bias in Jury Selection*, 61 AM. STATISTICIAN 296 (2007); Roger Allan Ford, *Modeling the Effects of Peremptory Challenges on Jury Selection and Jury Verdicts*, 17 GEO. MASON L. REV. 377 (2010); Joseph L. Gastwirth, *Case Comment: Statistical Tests for the Analysis of Data on Peremptory Challenges: Clarifying the Standard of Proof Needed to Establish a* Prima Facie *Case of Discrimination in* Johnson v. California, 4 L., PROBABILITY & RISK 179 (2005); Joseph L. Gastwirth, *Statistical Testing of Peremptory Challenge Data for Possible Discrimination: Application to* Foster v. Chatman, 69 VAND. L. REV. EN BANC 51 (2016).

[223] Whitus v. Georgia, 385 U.S. 545, 552 n.2 (1967); *accord* Alexander v. Louisiana, 405 U.S. 625, 630 n.9 (1972).

[224] Skidmore v. Swift & Co., 323 U.S. 134, 140 (1944).

[225] *Id.*; *accord* United States v. Mead Corp., 533 U.S. 218, 235 (2001).

Justices have not uniformly and enthusiastically encouraged the tool's uptake. This universal tool of statistical evaluation in the social and natural sciences has flourished despite the high court's history of curbing its legal applicability.

If anything, individual Justices often express frustration with their colleagues' failure to embrace the method more enthusiastically. Justice Stevens has urged the Court to welcome "a sophisticated regression analysis" of "the efficacy [and] effects of . . . handgun ban[s]," in light of the Court's lack of "the technical capacity and the localized expertise to assess" such measures' "wisdom, need, and propriety."[226] In a legal setting far removed from the Second Amendment, Chief Justice Roberts has urged securities law litigants to "introduce evidence of the *existence* of price impact" in publicly regulated capital markets through "'event studies'—regression analyses that seek to show that the market price of . . . stock tends to respond to pertinent publicly reported events."[227]

"There is no reason to belabor this line of analysis."[228] Though some observers take solace in the tendency of the judiciary (and especially the Supreme Court) to rely more heavily on empiricism and data-based decisionmaking, the law has had at best a turbulent relationship with social science in general and regression analysis in particular.[229] The Justices "are not statisticians," and at most "the views of experts (or their absence) might help [the Court] understand (though not control [its] determination" of legal questions.[230] The

[226] *See* McDonald v. City of Chicago, 561 U.S. 742, 903 & n.46 (2010) (Stevens, J., dissenting) (quoting Griswold v. Connecticut, 381 U.S. 479, 482 (1965)).

[227] *See* Halliburton Co. v. Erica P. John Fund, Inc., 573 U.S. 258, 280 (2014). *See generally* John J. Binder, *The Event Study Methodology Since 1969*, 11 REV. QUANTITATIVE FIN. & ACCT. 111 (1998); Louis Ederington, Wei Guan & Lisa (Zongfei) Yang, *Bond Market Event Study Methods*, 58 J. BANKING & FIN. 281 (2015); Ana Paula Serra, *Event Study Tests: A Brief Survey*, 2 REVISTA ELECTRÓNICA DE GESTÃO ORGANIZACIONAL 248 (2004).

[228] Craig v. Boren, 429 U.S. 190, 204 (1976).

[229] *See, e.g.*, Paul S. Appelbaum, *The Empirical Jurisprudence of the United States Supreme Court*, 13 AM. J.L. & MED. 335, 335–36 (1987); Rachael N. Pine, *Speculation and Reality: The Role of Facts in Judicial Protection of Fundamental Rights*, 136 U. PA. L. REV. 655, 657 (1988); Donald N. Bersoff & David J. Glass, *The Not-So* Weisman: *The Supreme Court's Continuing Misuse of Social Science Research*, 2 U. CHI. L. SCH. ROUNDTABLE 279, 279 (1995); Jeffrey M. Shaman, *Constitutional Fact: The Perception of Reality by the Supreme Court*, 35 U. FLA. L. REV. 236, 236–38 (1983).

[230] Zuni Pub. Sch. Dist. No. 89 v. Dep't of Educ., 550 U.S. 81, 100 (2007). *See generally* Joseph L. Gastwirth, *A 60 Million Dollar Statistical Issue Arising in the Interpretation and Calculation of a Measure of Relative Disparity:* Zuni Public

Court's "methodology [for] analyzing scientific information" is neither "sophisticated [n]or consistent," but rather "result-oriented," reactive, and haphazard.[231] Judicial use of science, unsurprisingly, resembles the use of precedent, the raw material from which legal reasoning is crafted.[232] The decline of law as an autonomous discipline, monotonic and steep, might yet be asymptotic.[233]

II. A PROLOGUE TO PRACTICAL MACHINE LEARNING: GATHERING AND PREPARING DATA

Having outlined one branch of the legal history of conventional regression methods, this Article now turns to a concrete demonstration of practical machine learning for law and legal scholarship. This prologue will discuss the Boston housing dataset and its preparation.

A. The Boston Housing Dataset

The data used in this demonstration of machine-learning methods comes from David Harrison and Daniel Rubinfeld's effort to predict housing prices in Boston's 506 census tracts in 1978.[234] The Boston housing dataset has become the social science equivalent of *Drosophilia melanogaster*, the common fruit fly, in biology.[235] This dataset has become a popular teaching tool for machine learning and other predictive methods.[236] The scikit-learn machine learning library for Python includes the Boston housing dataset.

School District 89 v. U.S. Department of Education, 5 L., PROBABILITY & RISK 33 (2006); Joseph L. Gastwirth, *The U.S. Supreme Court Finds a Statute's Description of a Simple Statistical Measure of Relative Disparity 'Ambiguous' Allowing the Secretary of Education to Interpret the Formula:* Zuni Public School District 89 v. U.S. Department of Education II, 7 L., PROBABILITY & RISK 225 (2008).

[231] Dean M. Hashimoto, *Science as Mythology in Constitutional Law*, 76 OR. L. REV. 111, 114 (1997).

[232] *See* John Monahan & Laurens Walker, *Social Authority: Obtaining, Evaluating, and Establishing Social Science in Law*, 134 U. PA. L. REV. 477, 477–78 (1986).

[233] *See* Richard A. Posner, *The Decline of Law as an Autonomous Discipline: 1962–1987*, 100 HARV. L. REV. 761, 761–63 (1987).

[234] *See* David Harrison, Jr. & Daniel L. Rubinfeld, *Hedonic Housing Prices and the Demand for Clean Air*, 5 J. ENV'T ECON. & MGMT. 81, 96–98 (1978).

[235] *See generally* Stanley Fields & Mark Johnston, *Whither Model Organism Research?*, 307 SCIENCE 1885 (2005) (discussing biological research on model organisms such as yeasts, nematodes, and fruit flies).

[236] *See* DAVID A. BELSLEY, EDWIN KUH & ROY E. WELSCH, REGRESSION DIAGNOSTICS: IDENTIFYING INFLUENTIAL DATA AND SOURCES OF COLLINEARITY 244–

The proper gathering and preparation of a dataset such as the Boston housing dataset enable the application of nearly all machine-learning models alongside conventional linear regression. Panel data, once rendered in the two-dimensional format most compatible with Excel and advanced statistical software such as R, Stata, or SPSS, can be exported in comma-separated value (CSV) format. The Pandas package for Python can import data in CSV format and put it immediately to work with minimal preprocessing in every machine learning model evaluated or even mentioned in this Article.[237]

Although the specific prices in the Boston housing dataset are woefully out of date, the need to evaluate residential real estate prices remains quite germane to a wide range of legal issues. Mortgage fraud is litigated with regularity.[238] So are other aspects of the subprime mortgage crisis widely blamed as the trigger of the Great Recession of 2008–2009.[239] Homebuyers and -sellers have alleged price-fixing conspiracies and other anticompetitive conduct among real estate agents.[240] Accurate modeling of housing prices would advance the proper resolution of these disputes.

61 (1980); THOMAS W. MILLER, MARKETING DATA SCIENCE: MODELING TECHNIQUES IN PREDICTIVE ANALYTICS WITH R AND PYTHON § 6.4 (2015); J.R. Quinlan, *Combining Instance-Based and Model-Based Learning*, *in* PROCEEDINGS OF THE TENTH INTERNATIONAL CONFERENCE OF MACHINE LEARNING 236–43 (Morgan Kaufmann Publishers, Inc., 1993).

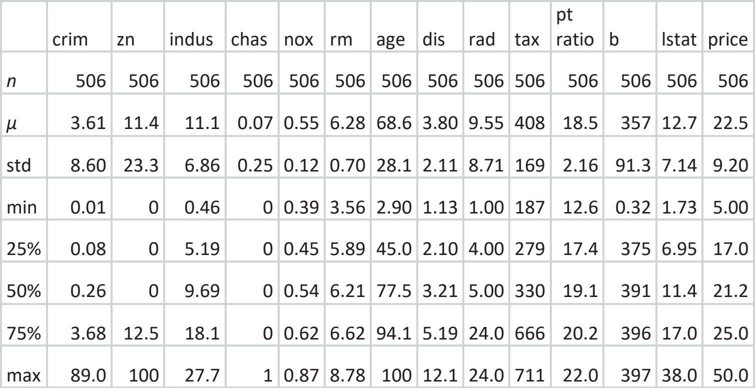
[237] *See generally* WES MCKINNEY, PYTHON FOR DATA ANALYSIS: DATA WRANGLING WITH PANDAS, NUMPY, AND IPYTHON (2d ed. 2017).

[238] *See, e.g.*, United States v. Phillips, 731 F.3d 649 (7th Cir. 2013) (en banc); United States v. Beecroft, 825 F.3d 991 (9th Cir. 2016); Matthew A. Edwards, *Punishing Hope? Materiality and Immateriality in Federal Mortgage Fraud Cases Under 18 U.S.C. § 1014*, 22 U. PA. J. BUS. L. 492, 513–21 (2020).

[239] *See, e.g.*, *In re* Lehman Bros. Mortg.-Backed Secs. Litig., 650 F.3d 167 (2d Cir. 2011); Makor Issues & Rts., Ltd. v. Tellabs, Inc., 513 F.3d 702 (7th Cir. 2008).

[240] *See* Moehrl v. Nat'l Ass'n of Realtors, No. 19-CV-01610, 2020 WL 5878016, at *3 (N.D. Ill. Oct. 2, 2020).

This is a numerical summary of the Boston housing dataset:



	crim	zn	indus	chas	nox	rm	age	dis	rad	tax	pt ratio	b	lstat	price
n	506	506	506	506	506	506	506	506	506	506	506	506	506	506
μ	3.61	11.4	11.1	0.07	0.55	6.28	68.6	3.80	9.55	408	18.5	357	12.7	22.5
std	8.60	23.3	6.86	0.25	0.12	0.70	28.1	2.11	8.71	169	2.16	91.3	7.14	9.20
min	0.01	0	0.46	0	0.39	3.56	2.90	1.13	1.00	187	12.6	0.32	1.73	5.00
25%	0.08	0	5.19	0	0.45	5.89	45.0	2.10	4.00	279	17.4	375	6.95	17.0
50%	0.26	0	9.69	0	0.54	6.21	77.5	3.21	5.00	330	19.1	391	11.4	21.2
75%	3.68	12.5	18.1	0	0.62	6.62	94.1	5.19	24.0	666	20.2	396	17.0	25.0
max	89.0	100	27.7	1	0.87	8.78	100	12.1	24.0	711	22.0	397	38.0	50.0

Table 1.

Kernel density estimation is a generalization of the histogram.[241] Kernel density estimation offers (among other things) a way to visualize the shape of stochastically distributed data.[242] These kernel density estimates provide comprehensive exploratory analysis of all variables in the Boston housing dataset:

[241] *See* George R. Terrell & David W. Scott, *Variable Kernel Density Estimation*, 20 ANNALS STAT. 1236, 1239–40, 1242 (1992).

[242] *See* Mats Rudemo, *Empirical Choice of Histograms and Kernel Density Estimators*, 9 SCANDINAVIAN J. STAT. 65, 65–78 (1982).

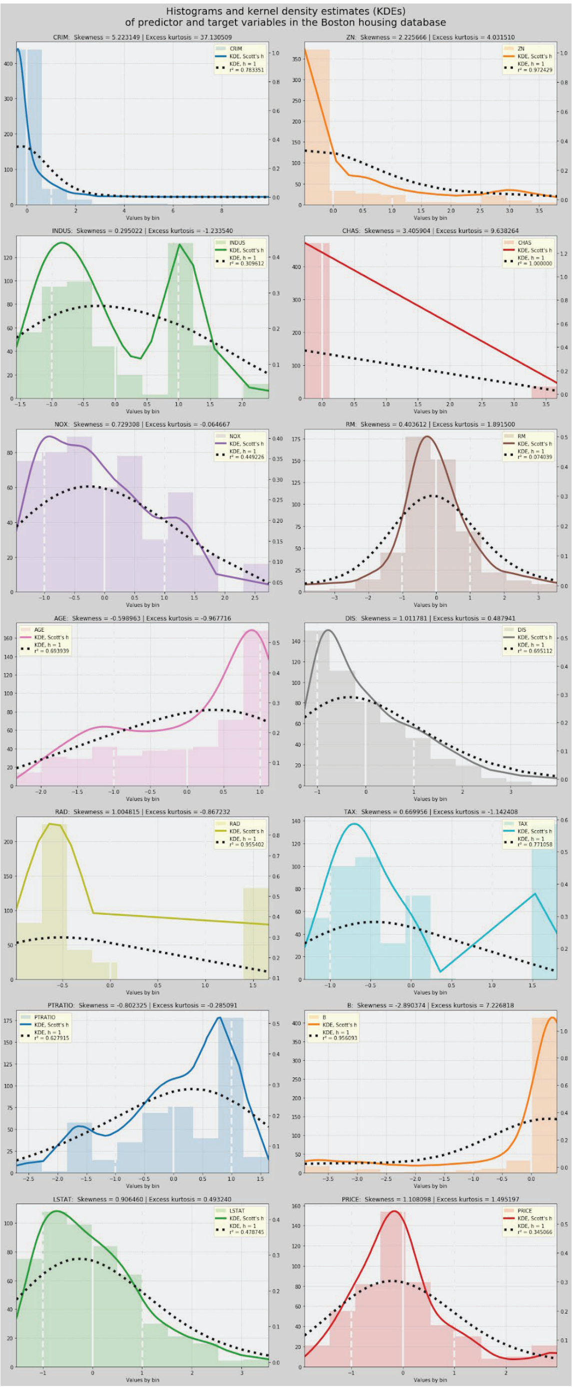

Histograms and kernel density estimates (KDEs)
of predictor and target variables in the Boston housing database
CRIM: Skewness = 5.223149 | Excess kurtosis = 37.130509
CRIM
KDE, Scott's h
KDE, h = 1
r^2 = 0.783351
0
100
200
300
400
0
2
4
6
8
0.0
0.2
0.4
0.6
0.8
1.0
Values by bin
ZN: Skewness = 2.225666 | Excess kurtosis = 4.031510
ZN
KDE, Scott's h
KDE, h = 1
r^2 = 0.972429
0
50
100
150
200
250
300
350
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
0.0
0.2
0.4
0.6
0.8
1.0
Values by bin
INDUS: Skewness = 0.295022 | Excess kurtosis = -1.233540
INDUS
KDE, Scott's h
KDE, h = 1
r^2 = 0.309612
0
20
40
60
80
100
120
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
0.1
0.2
0.3
0.4
Values by bin
CHAS: Skewness = 3.405904 | Excess kurtosis = 9.638264
CHAS
KDE, Scott's h
KDE, h = 1
r^2 = 1.000000
0
100
200
300
400
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
0.0
0.2
0.4
0.6
0.8
1.0
1.2
Values by bin
NOX: Skewness = 0.729308 | Excess kurtosis = -0.064667
NOX
KDE, Scott's h
KDE, h = 1
r^2 = 0.449226
0
20
40
60
80
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
Values by bin
RM: Skewness = 0.403612 | Excess kurtosis = 1.891500
RM
KDE, Scott's h
KDE, h = 1
r^2 = 0.074039
0
25
50
75
100
125
150
175
−3
−2
−1
0
1
2
3
0.0
0.1
0.2
0.3
0.4
0.5
Values by bin
AGE: Skewness = -0.598963 | Excess kurtosis = -0.967716
AGE
KDE, Scott's h
KDE, h = 1
r^2 = 0.693939
0
20
40
60
80
100
120
140
160
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
0.1
0.2
0.3
0.4
0.5
Values by bin
DIS: Skewness = 1.011781 | Excess kurtosis = 0.487941
DIS
KDE, Scott's h
KDE, h = 1
r^2 = 0.695112
0
20
40
60
80
100
120
140
−1
0
1
2
3
0.0
0.1
0.2
0.3
0.4
0.5
Values by bin
RAD: Skewness = 1.004815 | Excess kurtosis = -0.867232
RAD
KDE, Scott's h
KDE, h = 1
r^2 = 0.955402
0
50
100
150
200
−0.5
0.0
0.5
1.0
1.5
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Values by bin
TAX: Skewness = 0.669956 | Excess kurtosis = -1.142408
TAX
KDE, Scott's h
KDE, h = 1
r^2 = 0.771058
0
20
40
60
80
100
120
140
−1.0
−0.5
0.0
0.5
1.0
1.5
0.2
0.3
0.4
0.5
0.6
Values by bin
PTRATIO: Skewness = -0.802325 | Excess kurtosis = -0.285091
PTRATIO
KDE, Scott's h
KDE, h = 1
r^2 = 0.627915
0
25
50
75
100
125
150
175
−2.5
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
0.1
0.2
0.3
0.4
0.5
Values by bin
B: Skewness = -2.890374 | Excess kurtosis = 7.226818
B
KDE, Scott's h
KDE, h = 1
r^2 = 0.956093
0
50
100
150
200
250
300
350
400
−3.5
−3.0
−2.5
−2.0
−1.5
−1.0
−0.5
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Values by bin
LSTAT: Skewness = 0.906460 | Excess kurtosis = 0.493240
LSTAT
KDE, Scott's h
KDE, h = 1
r^2 = 0.478745
0
20
40
60
80
100
−1
0
1
2
3
0.0
0.1
0.2
0.3
0.4
Values by bin
PRICE: Skewness = 1.108098 | Excess kurtosis = 1.495197
PRICE
KDE, Scott's h
KDE, h = 1
r^2 = 0.345066
0
20
40
60
80
100
120
140
160
−1
0
1
2
0.1
0.2
0.3
0.4
0.5
Values by bin

Figure 1.

The final subplot shows the target variable, PRICE, as a nearly normal distribution, except a small subset of high-priced houses. This seemingly modest deviation from Gaussian normality will pose difficulties for linear regression—and an opportunity for machine-learning alternatives to demonstrate superior predictive accuracy.

B. Data Preparation

The supervised machine-learning methods applied to this dataset required the splitting of data into randomized subsets for training and testing. This practice, rarely followed in conventional econometrics, ensures that machine learning methods do not merely memorize labels or values associated with data to be predicted.[243] Holding out 25%—an admittedly arbitrary but frequently observed ratio—of the dataset for testing helps ensure the generalizability of any supervised machine learning model to data not seen during training.[244]

Accordingly, preparation of the data began with a 75% to 25% randomized division of observations between a training set and a holdout test set. Setting a determined seed for the pseudo-random number generator ensured reproducible results.

Many machine-learning algorithms perform more accurately when data is scaled.[245] I applied standard scaling to training data. Standard scaling ensures that machine learning evaluates and reports all results in terms of Gaussian *z*-scores, or multiples of a dependent or independent variable's standard deviation from its mean. Critically, scaling must begin exclusively on training data. Test data must be scaled according to the distribution of values in the training data. This separation of training and test data keeps data leakage from contaminating all predictive tests.[246]

C. Linear Regression of the Boston Housing Dataset

Conventional linear regression permits the expression and interpretation of the model in closed form, with coefficients and *p*-values familiar to all social scientists:

[243] *See* ANDREAS C. MÜLLER & SARAH GUIDO, INTRODUCTION TO MACHINE LEARNING WITH PYTHON: A GUIDE FOR DATA SCIENTISTS 17–18 (2017).

[244] *See id.*

[245] *See id.* at 134–42.

[246] *See id.* at 138–40.

Closed-form expression of the linear model (based on a train/test split)
Statistical significance — p < 0.001: ***; 0.01: **; 0.05: *; 0.10, +:

PRICE =

Intercept	0.000000
CRIM	-0.120264 **
ZN	0.150448 ***
INDUS	0.029518
CHA	0.074704 *
NOX	-0.280434 ***
RM	0.221709 ***
AGE	0.021906
DIS	-0.352755 ***
RAD	0.299396 ***
TAX	-0.202809 *
PTRATIO	-0.239119 ***
B	0.063051 +
LSTAT	-0.452595 ***

These are beta coefficients, expressed in Gaussian terms of according to the standard scaling of the data. Tests of statistical significance are based on a two-tailed *t*-test for Pearson's correlation coefficient for each of the independent variables relative to PRICE, the dependent variable.[247] Accuracy as measured by r^2 for test set predictions is quite respectable for traditional linear regression—approximately 0.779:

Training set score:	0.716806
Test set score:	0.778941

Predictive accuracy aside, perhaps the greatest value of linear regression lies in its ease of interpretation. The original Boston housing study sought to quantify the impact of nitrogen oxide levels (as a proxy for all forms of pollution) on residential real estate prices. The three stars accompanying the NOX variable indicate statistical significance in this respect: There was no more than a 0.1% probability that the relationship of this variable to price could have arisen solely

[247] For a worked example illustrating beta coefficients and the calculation of statistical significance based on *p*-values in the context of multivariable regression, see LARRY HATCHER, ADVANCED STATISTICS IN RESEARCH: READING, UNDERSTANDING, AND WRITING UP DATA ANALYSIS RESULTS 262–67 (2013).

by chance. The negative sign accompanying this variable's coefficient indicates a negative correlation between NO_x levels and housing price by census tract.

Finally, the standard scaling of all variables enables us to gauge the absolute impact of each variable on the predicted price. The absolute values of these beta coefficients imply that air pollution (as measured by NOX) has more than half the impact of LSTAT, which in turn is the most influential determinant of housing prices. The LSTAT variable measures the proportion of the population in each tract that could be (and was) characterized as having low socioeconomic status.

III. Machine-Learning Methods and Some Preliminary Results

We turn at last to machine learning. This part introduces the fancifully named dendrological class of methods based on decision trees and forest ensembles.

A. Decision Trees and Forest Ensembles

The classification and regression tree (CART) algorithm is the basis for a dazzling constellation of machine learning methods.[248] The resulting decision trees and forests (stochastically assembled ensembles of decision trees) often outperform linear regression. They are not limited to linear relationships. All decision tree-based algorithms are robust in the presence of outliers. These algorithms are also quite forgiving of misspecified models. The inclusion of weakly predictive or even wholly nonpredictive variables generally does not weaken a decision tree or tree-based ensemble.

Bifurcating the data according to values for each independent variable generates a decision tree predicting the average price per house in each of Boston's 506 census tracts. This basic machine-learning model immediately improves r^2 relative to the OLS baseline by nearly 0.100.

[248] *See generally* Leo Breiman, Jerome H. Friedman, Richard A. Olshen & Charles J. Stone, Classification and Regression Trees (1984); Wei-Yin Loh, *Classification and Regression Tree Methods*, *in* Encyclopedia of Statistics in Quality and Reliability 315 (Fabrizio Ruggeri, Ron S. Kenett & Frederick W. Faltin eds., 2008).

```
Training set score:     0.920483
Test set score:         0.876399
```

One weakness of decision trees and ensembles based upon them is that they are not amenable to evaluation according to *p*-values and conventional tests of statistical significance. One of the fiercest debates in machine learning involves the tension between less accurate but more readily interpreted "white box" models and more accurate but heuristically opaque "black box" models.[249] Methodological diversity within machine learning, however, offers solutions along a more refined spectrum of "gray" solutions offering different mixtures of accuracy and interpretability.[250] In practice, different applications will call for blends of white box models, black box models, and expert judgment.[251] The balance between interpretive clarity and sufficiency of data lies at the heart of the "credibility revolution" in empirical legal studies.[252]

Decision trees and ensemble methods based upon them do quantify the contribution of each predictive variable. All tree-based methods in scikit-learn report "feature importances," a vector of values whose sum is 1 and whose individual values correspond to each regressor's contribution to the model's predictions.[253] Specifically, feature importances in scikit-learn "is a weighted average, where each node's weight" in a decision tree or across all trees in a forest "is equal to the number of training samples that are associated with it."[254]

This doughnut plot reveals that the percentage of residents with lower socioeconomic status and the average number of rooms per

[249] *See* Cynthia Rudin, *Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead*, 1 Nature Mach. Intel. 206, 206–08 (2019).

[250] *See* Emmanuel Pintelas, Ioannis E. Livieris & Panagiotis Pintelas, *A Grey-Box Ensemble Model Exploiting Black-Box Accuracy and White-Box Intrinsic Interpretability*, Algorithms, Jan. 5, 2020.

[251] *See* Octavio Loyola-González, *Black-Box Vs. White-Box: Understanding Their Advantages and Weaknesses from a Practical Point of View*, 7 IEEE Access 154096, 154096–154113 (2019).

[252] *See* Ryan Copus, Ryan Hübert & Hannah Laqueur, *Big Data, Machine Learning, and the Credibility Revolution in Empirical Legal Studies*, *in* Law as Data: Computation, Text & the Future of Legal Analysis 21, 21 (Michael A. Livermore & Daniel N. Rockmore eds., 2019).

[253] *See* Aurélien Géron, Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems 198–99 (2d ed. 2019).

[254] *Id.* at 198.

house (a proxy for size) account for more than 80% of the predictive power of a basic decision tree.

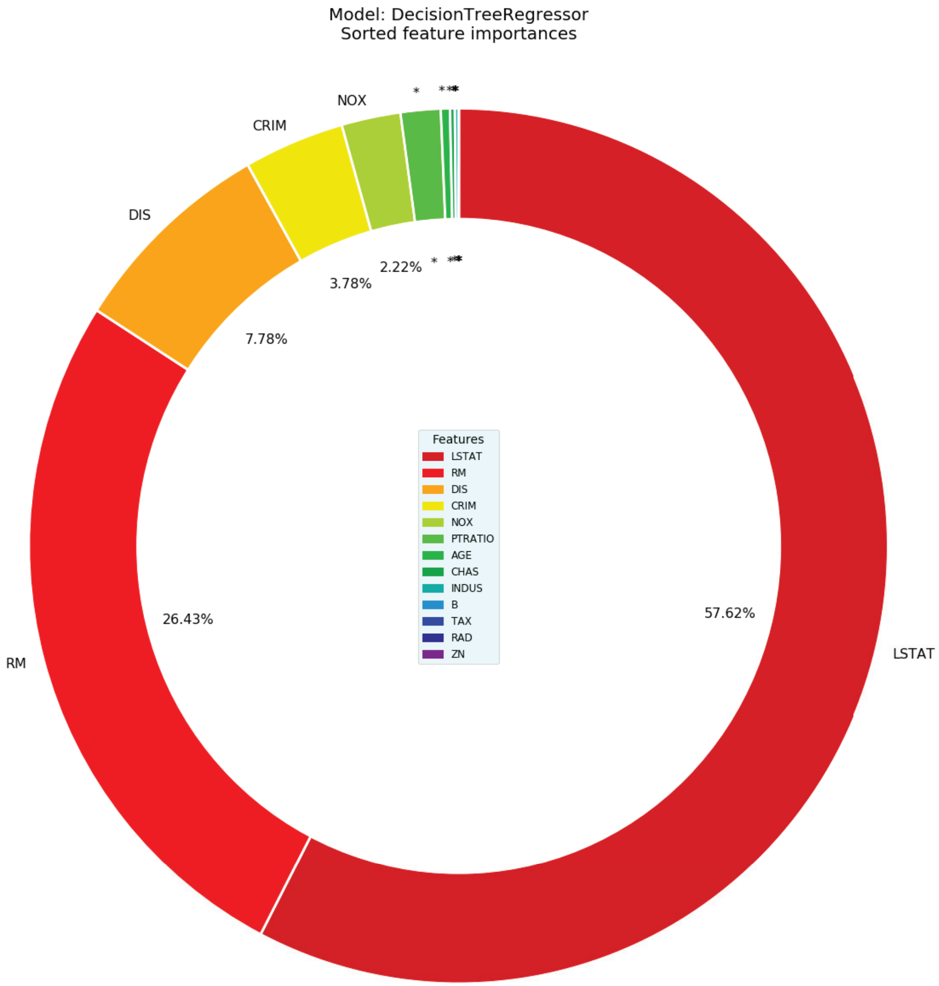


Figure 2.

The simplest way to diversify the results of a decision tree algorithm is to sample random subsets (either with or without replacement) of the training set. Bagging, short for *bootstrap* *aggregating*, takes samples with replacement.[255] Pasting takes samples without replacement.[256] Even in an infinite bagging process, 1/*e* of any dataset (approximately 0.368) will escape sampling:[257]

$$\lim_{x \to \infty} \left(1 - \frac{1}{x}\right)^{x} = \frac{1}{e}$$

As a result, the "out-of-bag" subset of training instances *not* chosen in bagging provides an additional validation set by which to evaluate the effectiveness and generalizability of the decision tree on previously unseen data.[258]

Bagging improves both the training and test performance of the decision tree algorithm on the Boston housing dataset. The loss of accuracy in the out-of-bag score, relative to the test score, does counsel some caution in the interpretation of these results.

	Decision tree	Bagging
Training set score:	0.920483	0.941659
Test set score:	0.876399	0.900210
Out-of-bag score:		0.854082

Ensemble and boosting methods based on decision trees include random forests, extremely randomized trees (extra trees), adaptive boosting (AdaBoost), gradient boosting, and extreme gradient boosting (XGBoost). All of these methods, plus stochastic variants of the gradient boosting algorithms, use exactly the same syntax within the scikit-learn application programming interface (API) for Python.

In anticipation of applying these more advanced methods, this Article will explain the bias-variance tradeoff. This property of all machine-learning models explains why collections of individually weaker predictors might improve accuracy.

[255] *See* Leo Breiman, *Bagging Predictors*, 24 MACH. LEARNING 123, 123–24 (1996).

[256] *See* Leo Breiman, *Pasting Small Votes for Classification in Large Databases and On-Line*, 36 MACH. LEARNING 85, 85–86 (1999).

[257] *See* GÉRON, *supra* note 253, at 195 & n.6.

[258] *See, e.g.*, Tom Bylander, *Estimating Generalization Error on Two-Class Datasets Using Out-of-Bag Estimates*, 48 MACH. LEARNING 287, 288 (2002); Gyeongcheol Cho, Kwanghee Jung & Heungsun Hwang, *Out-of-Bag Prediction Error: A Cross Validation Index for Generalized Structured Component Analysis*, 54 MULTIVARIATE BEHAV. RSCH. 505, 506–08 (2019).

B. Hyperparameter Testing and the Bias-Variance Tradeoff

Mastery of the *bias-variance* tradeoff is essential to the proper use of machine learning. This dilemma arises from an intrinsic property of all supervised machine learning models: Greater inaccuracy, or bias, in the estimates of the parameters of a model can reduce the variance among parameter estimates across samples.[259] The impossibility of perfectly reconciling the tension between bias and variance hampers efforts to apply supervised machine learning more generally to data on which such algorithms have not trained.[260]

Roughly speaking, *bias* refers to a method's overall accuracy, particularly in training. Excessive bias results in a model that *underfits* its data. As practitioners of generalized linear methods in economics may have observed with polynomial models, however, highly accurate models do not provide reliable results unless they generalize well to new, unseen data.[261] High-variance models tend to *overfit* training data. *Variance* therefore affects the generalizability and consistency of results with new data.

This image illustrates the bias-variance tradeoff as a quest to minimize prediction error.[262] At optimal complexity, a model strikes the best attainable balance between underfitting and overfitting training data.

[259] *See* Ron Kohavi & David H. Wolpert, *Bias Plus Variance Decomposition for Zero-One Loss Functions*, *in* MACHINE LEARNING: PROCEEDINGS OF THE THIRTEENTH INTERNATIONAL CONFERENCE (ICML '96) 275 (1996).

[260] *See* Stuart Geman, Élie Bienenstock & René Doursat, *Neural Networks and the Bias/Variance Dilemma*, 4 NEURAL COMPUTATION 1, 9–10 (1992).

[261] *See, e.g.*, Douglas M. Hawkins, *The Problem of Overfitting*, 44 J. CHEM. INFO. & COMPUT. SCIS. 1, 1–12 (2004); *cf.* Yaohao Peng & Mateus Hiro Nagata, *An Empirical Overview of Nonlinearity and Overfitting in Machine Learning Using COVID-19 Data*, CHAOS, SOLITONS & FRACTALS, June 30, 2020, at 1–16.

[262] *See* Frank J. W. M. Dankers, Alberto Traverso, Leonard Wee & Sander M. J. van Kuijk, *Prediction Modeling Methodology*, *in* FUNDAMENTALS OF CLINICAL DATA SCIENCE 101, 107 fig.8.3 (Pieter Kubben, Michel Dumontier & Andre Dekker eds., 2019).

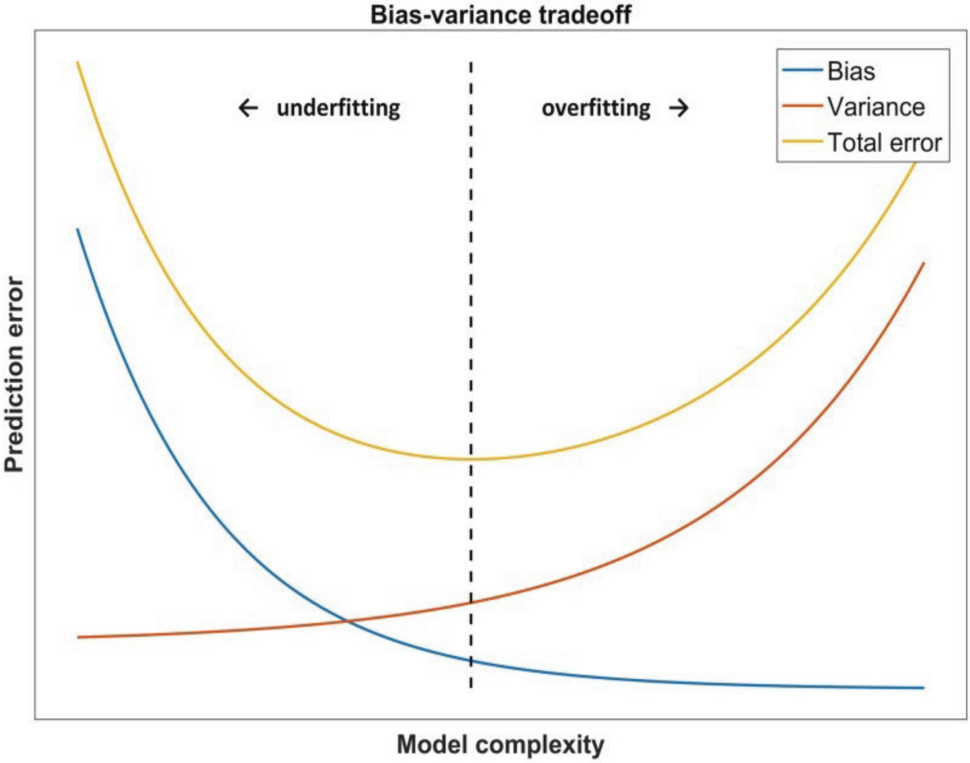


Figure 3.

Most machine-learning models offer a wide, sometimes daunting, list of adjustable hyperparameters. If these settings are not properly tuned, a model may fall far short of its predictive potential. Ways to explore a conceptually hyperparameter space include grid search and random search.[263]

Sufficiently large datasets provide the luxury of a three-way split between training, validation, and test subsets. For example, the celebrated MNIST dataset of handwritten digits (which made a vital contribution to the early development of optical character recognition) is divided into 60,000 training observations, 10,000 validation observations, and 10,000 test observations.[264] An intermediate holdout subset of validation data would enable us to strike the optimal balance between bias and variance before we apply a model with ideally tuned

[263] *See* MÜLLER & GUIDO, *supra* note 243, at 267–82.

[264] *See* Ernst Kussul & Tatiana Baidyk, *Improved Method of Handwritten Digit Recognition Tested on MNIST Database*, 22 IMAGE & VISION COMPUTING 971, 971–81 (2004).

hyperparameters to the final holdout subset of data designated as the "test set."

Figure 4.[265]

With 506 observations, the Boston housing dataset is relatively small. One way to try different hyperparameters without leaking final holdout data into training is *k-folds cross-validation*.[266] We can split the training data into *k* even smaller subsets and use each of those "folds" as a synthetic validation set.

IV. ADVANCED MACHINE-LEARNING RESULTS

Midway through life's journey, I have by no means abandoned all hope of understanding the love that moves the sun and other stars. Divine! For the moment, though, let us enter a sylvan clearing where we will surely glimpse random forests and extra trees.

[265] *See* Sample Images from MNIST Test Dataset (Illustration), *in* Josef Steppan, *File:MnistExamples.png*, WIKIMEDIA COMMONS, https://commons.wikimedia.org/wiki/File:MnistExamples.png [https://perma.cc/QYQ6-VUU2] (last visited Mar. 15, 2021).

[266] *See* MÜLLER & GUIDO, *supra* note 243, at 258–59.

A. Random Forests

Among ensemble and boosting methods based on aggregations of decision trees, random forests are perhaps the simplest.[267] They require the tuning of only two hyperparameters: the maximum number of features that a randomized tree may contain, plus the maximum depth of each tree (or the number of splits we will allow within each tree). Randomizing the threshold for each predictor yields an even more stochastic algorithm called *extremely random trees*, or extra trees.[268]

The admitted tedium of hyperparameter tuning gives way to the beauty of visualizations depicting the search for the ideal bias-variance balance:

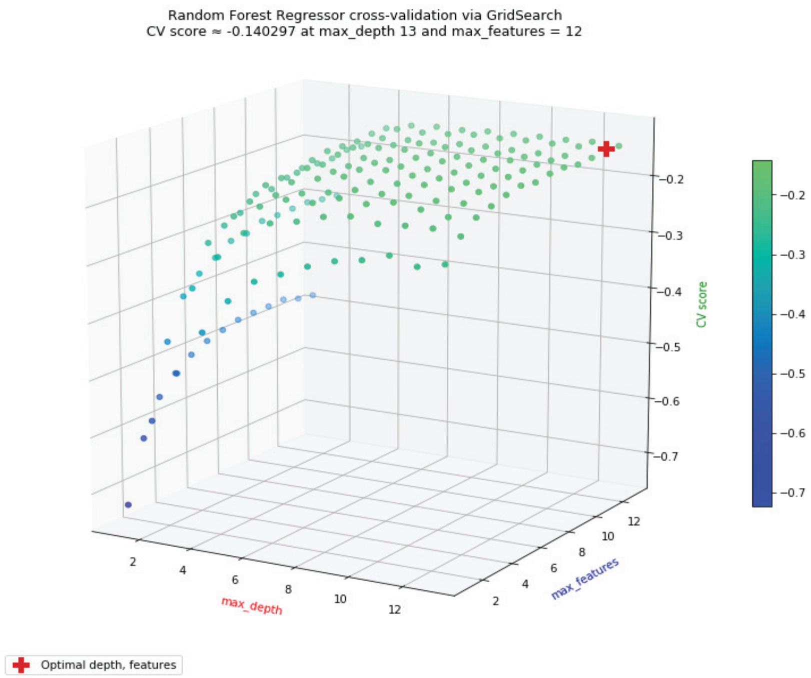


Figure 5.

[267] *See* Leo Breiman, *Random Forests*, 45 MACH. LEARNING 5, 5–6, 10–11 (2001); Tin Kam Ho, *Random Decision Forests*, *in* 1 PROCEEDINGS OF THE THIRD INTERNATIONAL CONFERENCE ON DOCUMENT ANALYSIS AND RECOGNITION 278, 278–82 (1995).

[268] *See* Pierre Geurts, Damien Ernst & Louis Wehenkel, *Extremely Randomized Trees*, 63 MACH. LEARNING 3, 5–7 (2006).

The ideal hyperparameter settings yield a random forest model that should generalize well to unseen data. The optimal random forest model chosen by 5-fold cross-validation improves r^2 by roughly 0.133:

Training set score:	0.981473
Test set score:	0.912558

Once again, feature importances report the relative weight of each regressor within the Boston housing dataset:

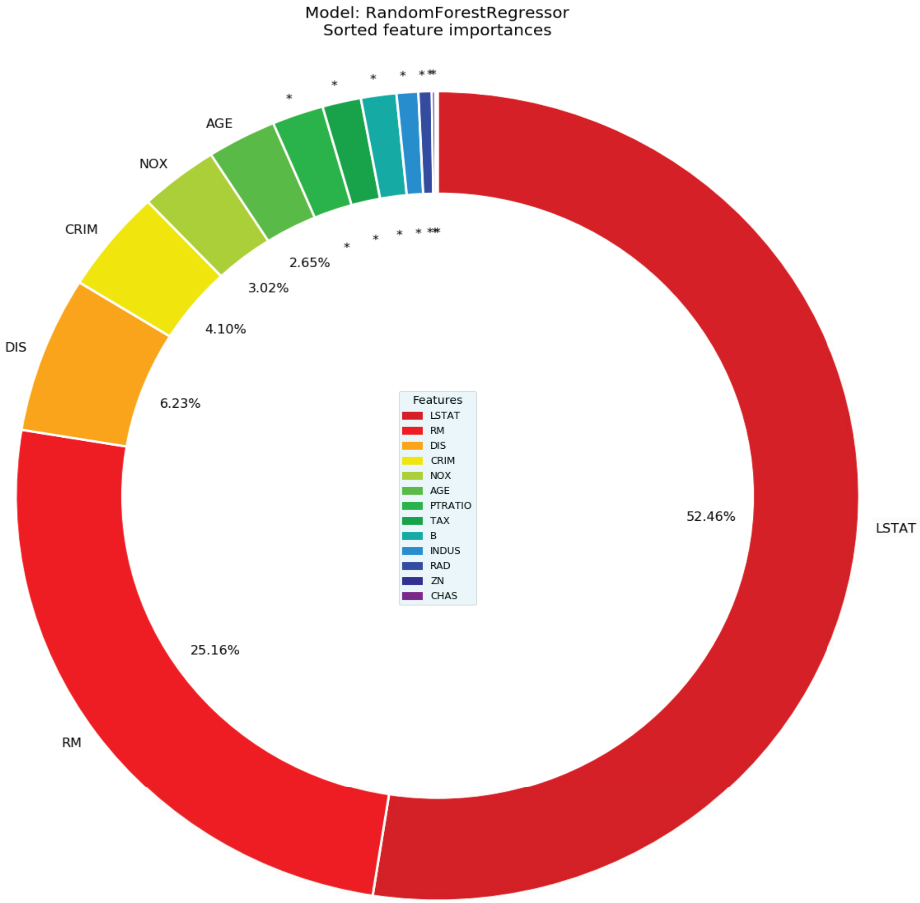


Figure 6.

Two consequences of the random forest method are immediately apparent. First, its accuracy is much higher than that of the baseline linear regression model. A scatterplot dramatically highlights the improvement from 0.779 test set r^2 to 0.912:

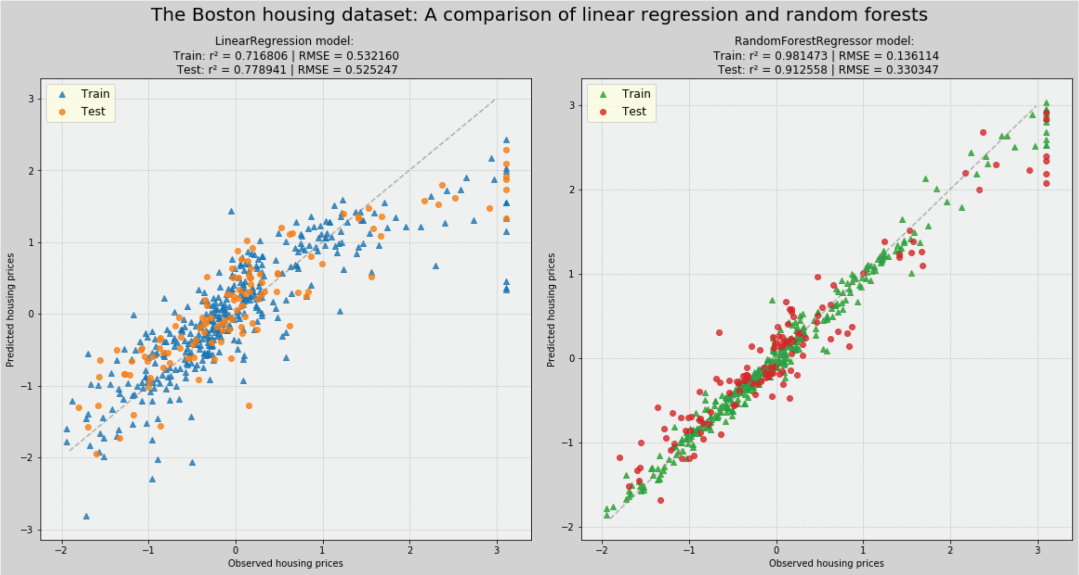


Figure 7.

The superior performance of random forests is most pronounced among the highest-valued observations in the Boston housing dataset. Machine learning outperforms multivariable linear regression in predicting prices in Boston's most expensive census tracts. Data that might otherwise be discarded as "outliers" becomes much more tractable in machine learning.

The traditional approach to statistics has devised all sorts of devices for managing possibly spurious outliers. Trimming crudely discards all data beyond points that a researcher regards as too extreme.[269] The slightly less destructive process of "winsorizing" clips outliers at an arbitrarily low or high level and assigns the corresponding minimum or maximum value in place of possibly

[269] *See* Brenton R. Clarke, *Empirical Evidence for Adaptive Confidence Intervals and Identification of Outliers Using Methods of Trimming*, 36 AUSTL. J. STAT. 45, 48–51 (1994); Edward J. Lusk, Michael Halperin & Frank Heilig, *A Note on Power Differentials in Data Preparation Between Trimming and Winsorizing*, 1 BUS. MGMT. DYNAMICS 23, 24–25 (2011).

extreme values.[270] For its part, the law occasionally expresses a conscious "purpose to exclude statistical outliers."[271]

By contrast, the intrinsic robustness of tree- and forest-based methods of machine learning counsels the retention of all data. The emergence of machine learning has revealed "the unreasonable effectiveness of data."[272] Given sufficient data, very different algorithms attain almost identical results on complex problems such as natural language disambiguation.[273] Performative convergence in spite of differences in algorithmic complexity suggests the primacy of data over theoretical elaboration and experimental design. "[I]nvariably, simple models and a lot of data trump more elaborate models based on less data."[274]

A key corollary of the unreasonable effectiveness of data is a systematic preference for retaining data as observed, with neither trimming nor winsorizing, in all forms of machine learning. The proponents of the unreasonable-effectiveness hypothesis have responded directly to scientists and scholars "who are worried about the curse of dimensionality and overfitting of models to data": "all the experimental evidence . . . suggests that throwing away rare events is almost always a bad idea," because the phenomena of greatest interest to practitioners of machine learning "consist[] of individually rare but collectively frequent events."[275]

A second consequence flows from the feature importances generated by the random forest model. LSTAT and RM, two variables quantifying (respectively) the proportion of persons of lower socioeconomic status and the average number of rooms per house in each census tract, jointly account for more than three-quarters of the predictions generated by the optimal random forest model. Air pollution, as measured by nitrogen oxides as a convenient proxy,

[270] *See* W. J. Dixon, *Simplified Estimation from Censored Normal Samples*, 31 ANNALS MATHEMATICAL STAT. 385, 388–89 (1960); Cecil Hastings, Jr., Frederick Mosteller, John W. Tukey & Charles P. Winsor, *Low Moments for Small Samples: A Comparative Study of Order Statistics*, 18 ANNALS MATHEMATICAL STAT. 413, 413–14 (1947); John W. Tukey, *The Future of Data Analysis*, 33 ANNALS MATHEMATICAL STAT. 1, 17–19 (1962).

[271] Zuni Pub. Sch. Dist. No. 89 v. Dep't of Educ., 550 U.S. 81, 91 (2007).

[272] *See* Alon Halevy, Peter Norvig & Fernando Pereira, *The Unreasonable Effectiveness of Data*, 24 IEEE INTELLIGENT SYS. 8, 8–12 (2009).

[273] *See generally* Michele Banko & Eric Brill, *Scaling to Very, Very Large Corpora for Natural Language Disambiguation*, *in* PROCEEDINGS OF THE 39TH ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS 26 (2001).

[274] Halevy, Norvig & Pereira, *supra* note 272, at 9.

[275] *Id.*

lagged far behind. The probability that a prediction would hinge on this variable scarcely reached 3%.

The contrast between feature importances and coefficients in the linear model counsels some skepticism toward what now appears to be the illusory clarity of ordinary least squares regression. The linear model supported the original Boston housing study's hypothesis that residential real estate prices reflect the negative impact of air pollution. The random forest model's feature importances diminish the weight that we might otherwise ascribe to this factor.[276]

Most human experts would probably agree that average home size and the "character" of a neighborhood, as a very thinly disguised euphemism for the presence of poor people, have a far greater impact on real estate prices. To the extent that air pollution does affect housing prices, its impact may reflect "environmental racism," or the tendency with which pollution is directed toward nonwhite

[276] As demonstrated in Section II.C of this Article, linear regression based on Gaussian-scaled data generates beta coefficients. *See* HATCHER, *supra* note 247, at 262–67. *Compare* Thomas B. Newman & Warren S. Browner, *In Defense of Standardized Regression Coefficients*, 2 EPIDEMIOLOGY 383 (1991), *with* Michael H. Criqui, *On the Use of Standardized Regression Coefficients*, 2 EPIDEMIOLOGY 393 (1991), *and* Sander Greenland, James J. Schlesselman & Michael H. Criqui, *The Fallacy of Employing Standardized Regression Coefficients and Correlations as Measures of Effect*, 125 AM. J. EPIDEMIOLOGY 349 (1987). Beta coefficients and their corresponding *p*-values can be converted into a vector of "emulated" feature importances resembling the true feature importances generated by dendrological machine-learning models such as random forests. *See* James Ming Chen, *Interpreting Linear Beta Coefficients Alongside Feature Importances in Machine Learning*, 49 ATL. ECON. J. (forthcoming 2021).

Formally, the emulated feature importance of a linear regression variable may be expressed as:

$$f_v = \frac{|\beta_v|(1-p_v)^\gamma}{\sum_{j=1}^{m} |\beta_j| (1-p_j)^\gamma}$$

where β indicates a beta coefficient, p indicates its p-value, the subscript v identifies a specific predictive variable, j is an indexing variable, and m indicates the number of predictive variables. The superscript γ indicates the possibility that the vector of 1 minus the p-values for each independent variable can be raised to an arbitrary power. Since $1 - p_j$ is strictly nonnegative, $\gamma \in \Re$.

My own work adheres to simple polynomial values for γ. A value of $\gamma = 2$ instead of 1 amplifies the "penalty" placed on p in a way that corresponds to the ridge (or ℓ_2) and lasso (or ℓ_1) norms in penalized linear regression. *See, e.g.*, Minjung Kyung, Jeff Gill, Malay Ghosh & George Casella, *Penalized Regression, Standard Errors, and Bayesian Lassos*, 5 BAYESIAN ANALYSIS 369 (2010); Mohammed El Anbari & Abdallah Mkhadri, *Penalized Regression Combining the L_1 Norm and a Correlation Based Penalty*, 76 SANKHYA B 82 (2014); Art B. Owen, *A Robust Hybrid of Lasso and Ridge Regression*, 443(7) CONTEMP. MATHEMATICS 59 (2007).

inhabitants.[277] Machine learning thus sharpens inferences drawn from more traditional methods of predictive inference and from subjective human judgment.

B. Beyond Basic Ensemble Methods: Boosting, Support Vector Machines, and Neural Networks

Once split and scaled for decision trees and random forests, data can undergo other machine-learning methods. Machine learning's "no free lunch" theorem counsels exploration of all available methods. No single method can be expected to provide the optimal solution for all datasets. At the same time, the "unreasonable effectiveness of data" suggests that these methods should converge toward similar solutions—as long as there are enough observations and as long as features have been properly selected and engineered. These principles point in opposite directions: Though machine-learning practitioners should try all methods, properly curated data should report similar results without regard to that choice among methods.

The tension between these theorems of machine learning gives way to a simple practical consideration: Properly preprocessed panel data can be fed, with no further modifications, for evaluation by the full range of regression methods in scikit-learn. Decision trees, forests, boosting methods, support vector machines, and a multilayer perceptron as the simplest neural network architecture are all available. Consequently, this Article will glance briefly at some of the machine-learning methods beyond basic ensembles.

Boosting represents a special class of ensembles that combine weak learners into a strong learner.[278] Each step in the sequential training of predictors strives to correct mistakes made by its predecessor.[279] The AdaBoost algorithm relies upon decision stumps, or decision trees truncated after a single split.[280] After each training instance, AdaBoost updates weights for each predictor.[281] Sequential

[277] *See generally, e.g.*, Robert D. Bullard, *Environmental Justice in the 21st Century: Race Still Matters*, 49 PHYLON 151 (2001); Rachel D. Godsil, *Remedying Environmental Racism*, 90 MICH. L. REV. 394 (1991); Ryan Holifield, *Defining Environmental Justice and Environmental Racism*, 22 URB. GEOGRAPHY 78 (2001).

[278] *See* Harris Drucker & Corinna Cortes, *Boosting Decision Trees*, 8 ADVANCES NEURAL INFO. PROCESSING SYS. 470, 472 (1996).

[279] *See* GÉRON, *supra* note 253, at 199.

[280] *See* Yoav Freund & Robert E. Schapire, *A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting*, 55 J. COMPUT. & SYS. SCIS. 119, 125–26 (1997).

[281] *See id.* at 126.

learning makes it difficult to implement AdaBoost through parallel computing and to scale it to larger datasets.[282]

The *gradient boosting* algorithm also adds predictors sequentially to an ensemble. Rather than adjusting the weights for each instance, as AdaBoost does, gradient boosting fits each new predictor to the previous predictor's residual errors.[283] Hyperparameters in gradient boosting control the ensemble's learning rate as well as the depth and growth of decision trees within the ensemble.[284] XGBoost, or Extreme Gradient Boosting, addresses the limits on speed and scalability that have plagued other boosting algorithms.[285] Training on random subsamples yields *stochastic* gradient boosting, which trades higher bias for lower variance and faster training.[286]

Support vector machines and neural networks represent two very different approaches to machine learning. Support vector machines fall into two categories, each named for the parameter by which it can be optimized (either epsilon or nu).[287] Better suited for small to medium-sized datasets, support vector machines are versatile enough to handle tasks such as classification, error and fraud detection, and even clustering, a form of unsupervised learning beyond the reach of

[282] *See* GÉRON, *supra* note 253, at 201.

[283] *See* Leo Breiman, *Arcing Classifiers*, 26 ANNALS STAT. 801, 809–10, 822–23 (1998); Jerome H. Friedman, *Greedy Function Approximation: A Gradient Boosting Machine*, 29 ANNALS STAT. 1189, 1192–94 (2001).

[284] *See* GÉRON, *supra* note 253, at 204.

[285] *See* Tianqi Chen & Carlos Guestrin, *XGBoost: A Scalable Tree Boosting System*, *in* PROCEEDINGS OF THE 22ND ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING 785, 789–91 (2016); Yingrui Zhou, Taiyong Li, Jiayi Shi & Zijie Qian, *A CEEMDAN and XGBOOST-Based Approach to Forecast Crude Oil Prices*, 2019 COMPLEXITY 4392785, at 4–5 (2019).

[286] *See* Jerome H. Friedman, *Stochastic Gradient Boosting*, 38 COMPUTATIONAL STAT. & DATA ANALYSIS 367, 369–70 (2002).

[287] *Compare* VLADIMIR VAPNIK, THE NATURE OF STATISTICAL LEARNING THEORY § 5.6, at 138–46 (2d ed., Springer 2000) (epsilon-optimized support vector machines), *with* Bernhard Schölkopf, Alex J. Smola, Robert C. Williamson & Peter L. Bartlett, *New Support Vector Algorithms*, 12 NEURAL COMPUTATION 1207, 1210–15 (2000) (nu-optimized). For contextual explanations of the difference between epsilon-optimized and nu-optimized support vector machines, see Jakub Langhammer & Julius Česák, *Applicability of a Nu-Support Vector Regression Model for the Completion of Missing Data in Hydrological Time Series*, 8(12) WATER 560, at 6 (2016); Fan Zhang, Chirag Deb, Siew Eang Lee, Junjing Yang & Kwok Wei Shah, *Time Series Forecasting for Building Energy Consumption Using Weighted Support Vector Regression with Differential Evolution Optimization Technique*, 126 ENERGY & BLDGS. 94, 95–97 (2016).

most other supervised methods.[288] Support vector machines are readily adapted to regression, typically a less computationally demanding task.[289]

Neural networks supply the muscle behind ambitious applications of computer vision, natural language processing, reinforcement learning, and robotics. Proponents of deep learning routinely predict that this form of artificial intelligence will be able to achieve all tasks entrusted to it.[290] By comparison, regression of economic panel data is a straightforward application of neural networks.[291]

The following six-way plot of the training and test set predictions for the Boston housing dataset shows the conventional linear model alongside five vastly superior machine-learning alternatives: bagging, random forests, XGBoost, a support vector machine, and a multilayer perceptron (or dense neural network). Both support vector regression and deep learning outperformed the random forest model, and linear regression by an even larger margin. In the aggregate, machine learning raised accuracy as measured by r^2 from 0.78 for linear regression to a range between 0.91 and 0.94.

[288] *See* Asa Ben-Hur, David Horn, Hava T. Siegelmann & Vladimir Vapnik, *Support Vector Clustering*, 2 J. MACH. LEARNING RSCH. 125, 125–26, 135 (2001).

[289] *See* Harris Drucker, Chris J.C. Burges, Linda Kaufman, Alex Smola & Vladimir Vapnik, *Support Vector Regression Machines*, 9 ADVANCES NEURAL INFO. PROCESSING SYS. 155, 155–58 (1997).

[290] *See* Karen Hao, *AI Pioneer Geoff Hinton: "Deep Learning Is Going to Be Able to Do Everything,"* 123 MIT TECH. REV. (Nov. 3, 2020), https://www.technologyreview.com/2020/11/03/1011616/ai-godfather-geoffrey-hinton-deep-learning-will-do-everything/ [https://perma.cc/K7AA-KUZX].

[291] *See* Fionn Murtagh, *Multilayer Perceptrons for Classification and Regression*, 2 NEUROCOMPUTING 183, 190, 192–94 (1991).

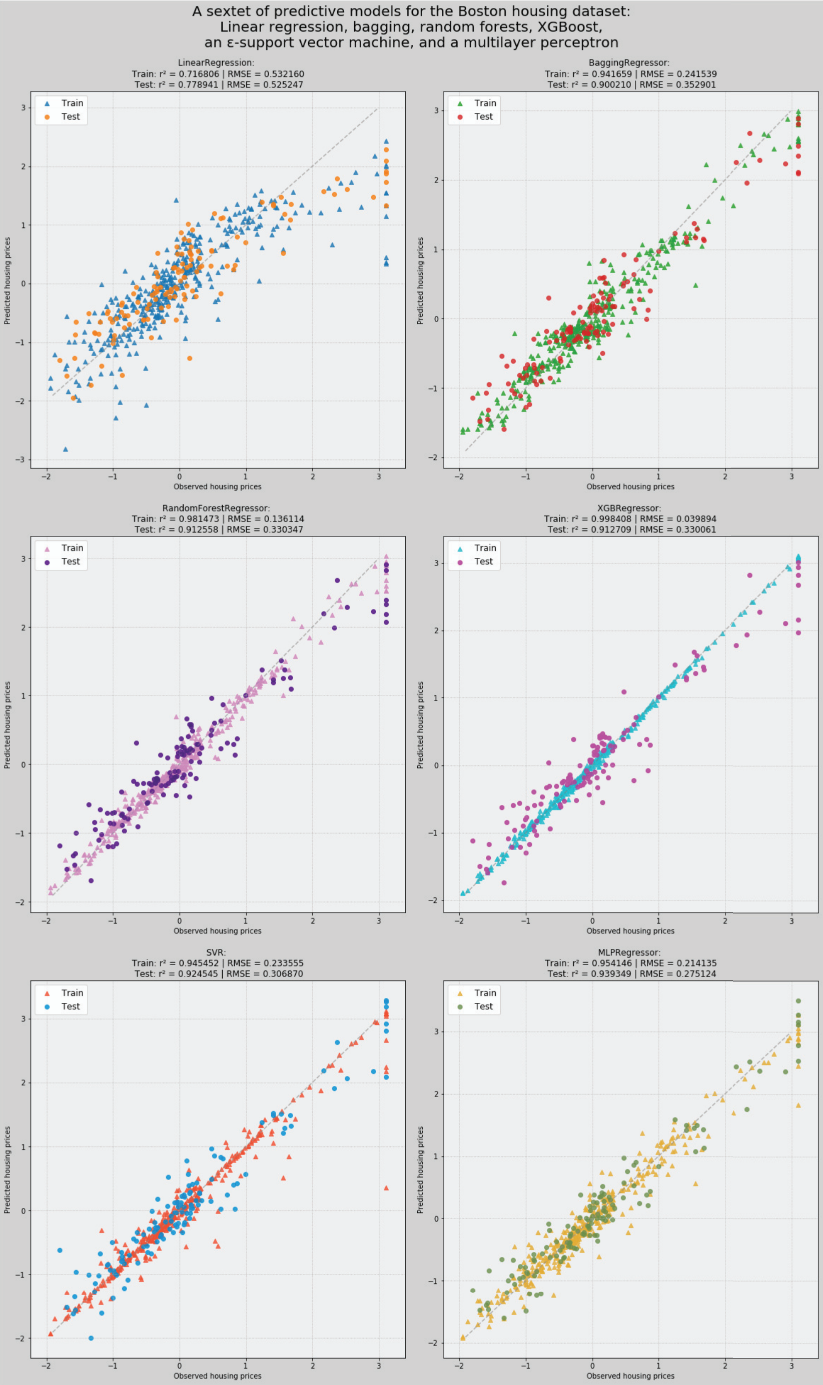

A sextet of predictive models for the Boston housing dataset:
Linear regression, bagging, random forests, XGBoost,
an ε-support vector machine, and a multilayer perceptron
LinearRegression:
Train: r² = 0.716806 | RMSE = 0.532160
Test: r² = 0.778941 | RMSE = 0.525247
BaggingRegressor:
Train: r² = 0.941659 | RMSE = 0.241539
Test: r² = 0.900210 | RMSE = 0.352901
RandomForestRegressor:
Train: r² = 0.981473 | RMSE = 0.136114
Test: r² = 0.912558 | RMSE = 0.330347
XGBRegressor:
Train: r² = 0.998408 | RMSE = 0.039894
Test: r² = 0.912709 | RMSE = 0.330061
SVR:
Train: r² = 0.945452 | RMSE = 0.233555
Test: r² = 0.924545 | RMSE = 0.306870
MLPRegressor:
Train: r² = 0.954146 | RMSE = 0.214135
Test: r² = 0.939349 | RMSE = 0.275124
Train
Test
Predicted housing prices
Observed housing prices
−3
−2
−1
0
1
2
3

Figure 8.

CONCLUSION: MACHINE LEARNING IN LAW AND LEGAL SCHOLARSHIP

This brief foray barely approaches the full potential for machine learning in law and legal scholarship. Enhancement of regression tasks through machine learning represents merely the first step toward a comprehensive approach to legal prediction.[292] A predictive approach to law search combines quantitative insights into search behavior with the science of information retrieval.[293] The precedential nature of law lends itself to evaluation through graph databases and network theory.[294] The very material of law—the words of statutes, rules, opinions, and orders—must yield its secrets to natural language processing, corpus linguistics, and computational linguistics.[295]

This Article's modest contribution to the basic toolkit of empirical legal studies urges law "to move beyond case studies, rhetoric, and conventional statistical methods" and toward a full embrace of "the empirical and technological methods" needed to evaluate law as a branch of data science.[296] The pragmatic roots of modern law have warned us that "a lawyer who has not studied economics and sociology is very apt to become a public enemy."[297] The twentieth century had not yet dawned when Oliver Wendell Holmes foresaw a legal future in which the master "of statistics and . . . economics" would dominate the master of "black-letter"

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

[293] *See* Faraz Dadgostari, Mauricio Guim, Peter A. Beling, Michael A. Livermore & Daniel N. Rockmore, *Modeling Law Search as Prediction*, 29 A.I. & L. (forthcoming 2021).

[294] *See generally* Greg Leibon, Michael Livermore, Reed Harder, Allen Riddell & Dan Rockmore, *Bending the Law: Geometric Tools for Quantifying Influence in the Multinetwork of Legal Opinions*, 26 A.I. & L. 145 (2018); J.B. Ruhl & Daniel Martin Katz, *Measuring, Monitoring, and Managing Legal Complexity*, 101 IOWA L. REV. 191 (2015).

[295] *See generally In re* Adoption of Baby E.Z., 266 P.3d 702 (Utah 2011); LAW AS DATA: COMPUTATION, TEXT & THE FUTURE OF LEGAL ANALYSIS (Michael A. Livermore & Daniel N. Rockmore eds., 2019); Stephen C. Mouritsen, The *Dictionary Is Not a Fortress: Definitional Fallacies and a Corpus-Based Approach to Plain Meaning*, 2010 BYU L. REV. 1915.

[296] Ruhl & Katz, *supra* note 294, at 244.

[297] STEPHEN W. BASKERVILLE, OF LAWS AND LIMITATIONS: AN INTELLECTUAL PORTRAIT OF LOUIS DEMBITZ BRANDEIS 229 (1994); *accord* Farber, *supra* note 21, at 175.

doctrine.[298] That prophecy is being fulfilled, though slowly and spasmodically.

"It is unrealistic to expect either members of the judiciary or state officials to be well versed in the rigors of experimental or statistical technique."[299] After decades of experience with conventional regression methods and models, the Supreme Court has managed "merely [to] illustrate[] that proving broad sociological propositions by statistics is a dubious business, and one that inevitably is in tension with the normative philosophy that underlies" the law.[300] Methods as novel and unfamiliar to legal decisionmakers as machine learning and artificial intelligence should not expect immediate acceptance in courts. Their accuracy impels their use in legal settings placing a premium on predictive power. Their opacity constrains their use as self-contained substitutes for conventional regression. Machine learning should be regarded as a complement to rather than a substitute for ordinary least squares and the broader family of linear methods.

This Article harbors no illusions that courts (or even administrative agencies) will embrace machine-learning methods for regression immediately and wholeheartedly. Courts have rejected the suggestion that their proceedings "should . . . be converted into a graduate seminar on economic forecasting."[301] It is far more likely that legal authorities, even when they begin considering machine-learning models, will express reservations as to their validity and reliability. We should also expect courts to calibrate their enthusiasm or skepticism toward machine learning according to the impact of these novel methods on the merits of individual cases and, even more so, on the ideological commitments held by judges. The history of the Supreme Court's treatment of conventional regression methods teaches as much.

Even in economics, machine-learning methods are unlikely to displace their conventional counterparts. Their lack of interpretive clarity ensures that closed-form expressions of a linear or polynomial regression model will almost always accompany a model based on an ensemble of decision trees, a support vector machine, or a neural network.

[298] Oliver Wendell Holmes, *The Path of the Law*, 110 HARV. L. REV. 991, 1001 (1997).

[299] Craig v. Boren, 429 U.S. 190, 204 (1976).

[300] *Id.*

[301] Doca v. Marina Mercante Nicaraguense, S.A., 634 F.2d 30, 39 (2d Cir. 1980), *cert. denied*, 451 U.S. 971 (1981); *accord* Jones & Laughlin Steel Corp. v. Pfeifer, 462 U.S. 523, 548 (1983).

In scientific as in doctrinal development, however, "[t]he law is not indifferent to considerations of degree."[302] Since the 1970s, the judiciary has slowly and fitfully accommodated some of the most commonplace and useful tools of social science. Because machine-learning regression methods are likely to complement rather than replace conventional methods, the legal acceptance of these methods will likely track the slow, fitful, but inexorable diffusion of methods such as ordinary least squares, classification through logistic regression, and event studies of financial markets. That process of diffusion, one might imagine, will resemble that of the conduction of heat through a solid medium.[303] Or, alternatively, the transmission of legal information will follow the physics of decay and relaxation[304]

It is not too far-fetched to imagine that a discipline that once spoke strictly of founders, framers, and *The Federalist Papers* might come to speak with equal fluency of Foucault and Fourier. Intellectual progress, no less than the movement of particles and waves, honors its own rhythms without regard to human desires or concerns. "A law of acceleration, definite and constant as any law of mechanics, cannot be supposed to relax its energy to suit the convenience of" human laws.[305] In time, waves reinforcing the links between conventional methods and their machine-learning counterparts "will have radiated so far that their undulatory motion, if discernible at all, will be too faint or obscure, too broken by cross-currents, to be [distinguished] by the law."[306]

[302] A.L.A. Schechter Poultry Corp. v. United States, 295 U.S. 495, 554 (1935) (Cardozo, J., concurring).

[303] *See generally, e.g.*, J. Unsworth & F. J. Duarte, *Heat Diffusion in a Solid Sphere and Fourier Theory: An Elementary Practical Example*, 47 AM. J. PHYSICS 981 (1979).

[304] *See generally, e.g.*, Alexander Lukichev, *Physical Meaning of the Stretched Exponential Kohlrausch Function*, 383 PHYSICS LETTERS A 2983 (2019); Graham Williams & David C. Watts, *Non-Symmetrical Dielectric Relaxation Behaviour Arising from a Simple Empirical Decay Function*, 66 TRANSACTIONS FARADAY SOC'Y 80–85 (1970).

[305] HENRY ADAMS, THE EDUCATION OF HENRY ADAMS: AN AUTOBIOGRAPHY 493 (Edmund Morris intro., 1996).

[306] Carter v. Carter Coal Co., 298 U.S. 238, 327 (1936) (Cardozo, J., dissenting in part and concurring in the result).