Vanderbilt Law Review

Volume 73
Issue 3 *Issue 3*

Article 1

4-2020

Statistical Precedent: Allocating Judicial Attention

Ryan W. Copus

VANDERBILT LAW REVIEW

VOLUME 73 APRIL 2020 NUMBER 3

ARTICLES

Statistical Precedent: Allocating Judicial Attention

Ryan W. Copus[*]

Suffering from a well-covered "crisis of volume," the U.S. Courts of Appeals have patched together an ad hoc system of triage in an effort to provide cases with sufficient attention. For example, only some cases are assigned to central staff, analyzed by law clerks, orally argued, debated over by judges, or decided in published opinions. The courts have evaded overt disaster by increasing the number of active, senior, and visiting judges, but adding personnel poses its own demands on attention—judges must also pay attention to one another in order to coherently develop and apply the law. With too little time and too many voices, they have increasingly abandoned the effort to coordinate that uniform approach to judging: the courts now create traditional precedent in less than 10% of cases, some larger courts have

* Climenko Fellow and Lecturer on Law, Harvard Law School. As this Article is based on my doctoral dissertation, I cannot hope to acknowledge everyone who talked through the ideas, commented on drafts, or provided research assistance, and I apologize for omissions. I am grateful to my dissertation committee, Kevin Quinn (co-chair), Justin McCrary (co-chair), Sean Farhang, and Anne Joseph O'Connell for guidance; Jacob Gersen, Marin Levy, Rob MacCoun, and Holger Spamann for insights and encouragement; Ryan Hubert, Hannah Laqueur, and Julian Nyarko for painfully struggling through the analytic foundations; Climenko Fellows for their criticism and framing advice; and the research assistance of Jonathan Korn and Cullen O'Keefe. My thoughts on this project have evolved considerably, and some of the views presented here differ from those I discussed with the above-mentioned individuals.

stopped the practice of circulating opinion drafts to the full court, and en banc proceedings are initiated at a miniscule rate.

This Article explains and illustrates how courts can leverage advances in artificial intelligence to more fairly and effectively allocate attention. A machine-generated mapping of a court's historical decision patterns—what I term "statistical precedent"—can help a circuit court locate the district court, agency, staff attorney, law clerk, and panel decisions that are most incompatible with the court's collective jurisprudence. Statistical precedent can also aid the court in identifying areas of law that are most in need of development. With the ability to locate likely errors and opportunities for law development, the circuit courts could distribute attention so as to revitalize their contribution to the rule of law.

INTRODUCTION

The U.S. Courts of Appeals were once admired for their wealth of judicial attention and for their generosity in distributing it.[1] At least by legend, almost all cases were afforded what William Richman and William Reynolds have termed the "Learned Hand Treatment."[2] Guided by Judge Learned Hand's commandment that "[t]hou shalt not ration justice,"[3] a panel of three judges would read the briefs, hear oral argument, deliberate at length, and prepare multiple drafts of an opinion.[4] Once finished, the judges would publish their opinion, binding themselves and their colleagues in accordance with the common-law tradition. The final opinion would be circulated to and read by every judge in the circuit, providing nonpanel judges with an opportunity to provide feedback or evaluate a decision for en banc review. And on top of this extensive attention was a reasonable chance for yet more, as the Supreme Court reviewed approximately 3% of the circuit courts' decisions.[5] But darker days were ahead.

A caseload explosion greatly diminished the courts' reservoir of judicial attention.[6] Between 1960 and 2010, the courts' caseload

1. *See* WILLIAM M. RICHMAN & WILLIAM L. REYNOLDS, INJUSTICE ON APPEAL: THE UNITED STATES COURTS OF APPEALS IN CRISIS 3 (2013) (noting the "received and perhaps idealized tradition of the operation of the circuit courts").

2. *Id.* at 3.

3. Judge Learned Hand, Keynote Address at the Legal Aid Society's 75th Anniversary (1951), https://www.legalaidnyc.org/historical_event/thou-shalt-not-ration-justice/ [https://perma.cc/23EX-BX7G].

4. Judge Hand apparently "wouldn't even let a law clerk write a sentence, not one sentence." MARVIN SCHICK, LEARNED HAND'S COURT 107 n.92 (1970) (quoting Harold R. Medina, *The Decisional Process*, 20 B. BULL. N.Y. COUNTY LAW. ASS'N 94, 99 (1962)).

5. *See* COMM'N ON STRUCTURAL ALTS. FOR THE FED. COURT OF APPEALS, FINAL REPORT 12 tbl.2-1 (Dec. 18, 1998), https://www.law.berkeley.edu/wp-content/uploads/2019/03/Commission-on-Structural-Alternatives-for-the-Federal-Courts-of-Appeals-1998.pdf [https://perma.cc/2DQC-5H6J].

6. The caseload crisis and its ill effects have been the subject of a long line of studies and congressional commissions. For a detailed review, see RICHMAN & REYNOLDS, *supra* note 1, at 128–64.

increased by 1,436%.[7] The courts responded to this precipitous rise in workload with a series of moves to reduce the time and effort that judges spent on each case. They employed an army of staff attorneys to help decide cases and draft opinions, increased the number of law clerks from one to three or four per judge, and curtailed the availability of oral argument such that in 2017, it was provided in less than 20% of cases.[8] Deliberation among judges on a panel is, by most accounts, rare, and almost 90% of decisions are made in terse, unpublished, and nonprecedential opinions.[9] The courts now review a mere 0.19% of decisions en banc, down from 1.5% in 1964.[10] And the Supreme Court has similarly reduced its contribution, reviewing only 0.1% of circuit court decisions,[11] down from approximately 3% in 1950.[12]

The shortage of attention threatens to undermine the courts' ability to decide cases correctly and develop the law coherently. Without the time to carefully consider each case, circuit court judges—traditionally serving as the main source of error correction in the federal courts—will inevitably make more errors of their own. Research, for example, shows that reversal rates in civil appeals declined as more attention was funneled to address the influx of immigration appeals.[13] And, as already noted, the circuit courts have

7. *Id.* at 6.

8. *Table B-10: U.S. Courts of Appeals—Cases Terminated on the Merits after Oral Arguments or Submission on Briefs, by Circuit, During the 12-Month Period Ending September 30, 2017*, U.S. CTS., http://www.uscourts.gov/statistics/table/b-10/judicial-business/2017/09/30 (last visited Apr. 5, 2020) [https://perma.cc/4KJX-DBRH].

9. *Table B-12: U.S. Courts of Appeals—Type of Opinion or Order Filed in Cases Terminated on the Merits, by Circuit, During the 12-Month Period Ending September 30, 2017*, U.S. CTS., http://www.uscourts.gov/statistics/table/b-12/judicial-business/2017/09/30 (last visited Apr. 5, 2020) [https://perma.cc/A8B9-QSLB].

10. A. Lamar Alexander, Jr., Note, *En Banc Hearings in the Federal Courts of Appeals: Accommodating Institutional Responsibilities (Part I)*, 40 N.Y.U. L. REV. 563, 564 (1965).

11. Roy E. Hofer, *Supreme Court Reversal Rates: Evaluating the Federal Courts of Appeals*, 2 LANDSLIDE (Jan.–Feb. 2010), https://ipo.org/wp-content/uploads/2013/03/Supremecourt reversalrates.pdf [https://perma.cc/F7FY-HEY9].

12. The Federal Judicial Center reports that 3,064 decisions were terminated by the U.S. Courts of Appeals in 1950. *Caseloads: U.S. Courts of Appeals, 1892-2017*, FED. JUD. CTR., https://www.fjc.gov/history/courts/caseloads-us-courts-appeals-1892-2017 (last visited Apr. 5, 2020) [https://perma.cc/84AL-YS6N]. It also reports that the U.S. Supreme Court granted certiorari in 103 cases in 1950. *Caseloads: Supreme Court of the United States, Petitions for Certiorari, 1923-1969*, FED. JUD. CTR., https://www.fjc.gov/history/courts/caseloads-supreme-court-united-states-petitions-certiorari-1923-1969 (last visited Apr. 5, 2020) [https://perma.cc/FC87-WDB7].

13. *See* Bert I. Huang, *Lightened Scrutiny*, 124 HARV. L. REV. 1109 (2011) (demonstrating that civil reversal rates in the Second and Ninth Circuits fell in correlation with a heavier immigration workload).

dramatically reduced their contribution to the development of law.[14] In brief, the courts are struggling to perform their two main functions: error correction and law development.

The ostensibly obvious solution—more judges—creates its own drains on attention. Judging is a social, collective enterprise. In order to apply and develop a coherent system of law, judges need attend to not only their own cases, but also to *one another*. In the age of legal realism, we cannot rely on a mechanical jurisprudence to coordinate the consistent application and development of law.[15] And the proliferation of judges—250% since 1960[16]—increases both the difficulty and importance of judges paying attention to other judges. Each judge has her own judicial philosophy, set of heuristics, and idiosyncrasies. When small in number, judges can learn and adapt to other judges, fitting their own unique judicial style into the broader jurisprudence of their courts. But in larger courts, judges work with one another less frequently, are unable to keep abreast of precedent produced by their colleagues,[17] and lose touch with the norms that support a common sense of justice.[18] More extreme panels, ideological

14. District courts, facing a dearth of precedent from the circuit courts, have increasingly turned to themselves for legal guidance. By my count (using an automated citation counter), between 1993 and 2013, district courts almost tripled the rate at which they cite to other district court opinions in their published opinions. Perhaps unsurprisingly, the notion of a "district court split" has become much more common: the phrase returned only ten results in a LexisNexis search of district court opinions issued in 2004, but the same search of 2013 district court opinions returned seventy results.

15. *See* Chad M. Oldfather, *Error Correction*, 85 IND. L.J. 49, 76–79 (2010) (discussing how legal realism has led to an acceptance of indeterminacy and a more equitable form of review).

16. RICHMAN & REYNOLDS, *supra* note 1, at 6. Note that this does not account for increased reliance on senior and visiting judges.

17. *See The Case for Restructuring the Ninth Circuit: An Inevitable Response to an Unavoidable Problem*: *Hearing on Oversight of the Structure of the Federal Courts Before the Subcomm. on Oversight, Agency Action, Federal Rights and Federal Courts of the S. Comm. on the Judiciary*, 115th Cong. 9 (2018) [hereinafter O'Scannlain Statement] (written testimony of Diarmuid F. O'Scannlain, Circuit Judge, U.S. Court of Appeals for the Ninth Circuit) ("[E]ven our own judges have difficulty simply staying abreast of the circuit's ever-expanding caselaw."); *Bringing Justice Closer to the People: Examining Ideas for Restructuring the 9th Circuit: Hearing Before the Subcomm. on Courts, Intellectual Prop., and the Internet of the H. Comm. on the Judiciary*, 115th Cong. 4–5 (2017) [hereinafter Kleinfeld Statement] (written statement of Andrew J. Kleinfeld, Circuit Judge, U.S. Court of Appeals for the Ninth Circuit):

> Judges on the same court should read each other's decisions. We are so big that we cannot and do not. That has the practical effect that we do not know what judges on other panels are deciding. It is odd word usage to call a public body a "court," in the singular, if its judges do not ever sit together as one body, and do not even read each other's opinions. We may get the quotes right from other panels' decisions, but there is no way anyone can get a feel for our court, as all attorneys do for smaller courts.

18. *See, e.g.*, COMM'N ON STRUCTURAL ALTS. FOR THE FED. COURT OF APPEALS, *supra* note 5, at 29 ("[T]here is consensus among appellate judges throughout the country . . . that a court of appeals, being a court whose members must work collegially over time to develop a consistent and coherent body of law, functions more effectively with fewer judges . . . ."); Stephen L. Wasby,

or otherwise, are impaneled,[19] and the court becomes less capable of monitoring and correcting their excesses.[20] In the words of one judge on the U.S. Court of Appeals for the Ninth Circuit, larger courts struggle to form a "reckonable court."[21]

In summary, judging has become a more time-pressured and solipsistic exercise.[22] In order to thoughtfully and coherently apply and develop the law, courts must be careful in allocating their limited attention. As it stands, the courts are struggling to patch together an ad hoc triage system. Little is known about who makes triage decisions or how they are made, and practices differ considerably across circuits, but they are routinely a product of discretion and

Communication in the Ninth Circuit: A Concern for Collegiality, 11 U. PUGET SOUND L. REV. 73, 129–32 (1987) (reporting concerns of Ninth Circuit judges that the size of the court reduces collegiality); *see also* O'Scannlain Statement, *supra* note 17, at 8–9 ("[T]he sheer number of judges on our court often means that we work 'together' only nominally. . . . It should be no surprise that it becomes difficult to establish effective working relationships in discerning the law when we sit together so rarely."); *Rebooting the Ninth Circuit: Why Technology Cannot Solve Its Problems: Hearing Before the Subcomm. on Privacy, Tech. and the Law of the S. Comm. on the Judiciary*, 115th Cong. 12–13 (2017) (written statement of Richard C. Tallman, Circuit Judge, U.S. Court of Appeals for the Ninth Circuit):

> Collegiality is extremely important in our appellate system. The genius of the appellate process is founded upon the close collaboration of jurists who combine their independent judgment, informed by their personal experiences, and apply their collective wisdom to decide the issues presented by an appeal. Only by sitting together regularly can members of a court come to know one another and work most effectively in common pursuit of the right answer under the Rule of Law.

19. D.H. Kaye, *On a Mathematical Argument for Splitting the Ninth Circuit*, 48 JURIMETRICS J.L. SCI. & TECH. 329 (2008); *see also* Richard B. Saphire & Michael E. Solimine, *Diluting Justice on Appeal?: An Examination of the Use of District Court Judges Sitting by Designation on the United States Courts of Appeals*, 28 U. MICH. J.L. REFORM 351, 372–75 (1995) (discussing evidence that visiting district court judges increase aberrancy of decisions).

20. *See* O'Scannlain Statement, *supra* note 17, at 11 ("Our court regularly receives around 800 petitions for en banc review a year. . . . Identifying which of those 800 petitions merits further review is a labor-intensive task . . . . There are, alas, only so many hours in a day."); *Review of the Report by the Commission on Structural Alternatives for the Federal Courts of Appeals Regarding the Ninth Circuit and the Ninth Circuit Reorganization Act: Hearing Before the Subcomm. on Admin. Oversight and the Courts of the S. Comm. on the Judiciary*, 106th Cong. 84 (1999) (statement of Andrew J. Kleinfeld, Circuit Judge, U.S. Court of Appeals for the Ninth Circuit) ("When a circuit [g]rows to a size such that its judges cannot read and correct other panels' decisions, district judges and lawyers trying to figure out what the law is are compelled to say that it depends on who is on the panel.").

21. Kleinfeld Statement, *supra* note 17, at 7.

22. *See* Erwin N. Griswold, *The Federal Courts Today and Tomorrow: A Summary and Survey*, 38 S.C. L. REV 393, 405–06 (1987):

> [T]his sparse review promotes a lack of discipline among judges sitting on the courts of appeals. . . . What we have . . . is a collection of very able judges who work very hard, but essentially on an individual basis, without very much in the way of careful guidance, and far too little authoritative guidance . . . . The consequence is that the system of precedent on which the common law is based has lost much of its structure and influence. . . . In essence, what we now have is rapidly becoming a discretionary approach to justice.

proxy.[23] Staff attorneys are assigned to make initial decisions and draft opinions in pro se, immigration, social security, and "straightforward" appeals.[24] Oral argument is denied where the "decisional process would not be significantly aided by oral argument."[25] An opinion is supposed to be published if it "establishes, alters, modifies, clarifies, or explains a rule of law."[26] En banc is reserved for circumstances where it is "necessary to secure or maintain uniformity of the court's decisions" or if an appeal "involves a question of exceptional importance."[27] Many courts have effectively abandoned the effort to keep judges aware of their court's new precedent, jettisoning the practice of precirculating opinions to nonpanel colleagues.[28]

This Article argues that a system of *statistical precedent* can help the courts more fairly and effectively allocate attention, thereby promoting the courts' error-correcting and law-developing functions. Like traditional precedent, statistical precedent is the product of a court's historical decisions. But in contrast to traditional precedent, which is based on outcomes and reasoning in a handful of judge-identified "similar" cases, statistical precedent is based on finely tuned patterns automatically mined from large-scale datasets of previous decisions. In short, a statistical precedent is a precise, rigorous, and machine-generated answer to a critical question: How frequently has the court reversed cases *like this one*? By exploiting the statistical associations between circuit court decisions and case characteristics (e.g., case subject matter; lower court outcome; identity of the lower court judge; whether a case was decided by motion to dismiss, summary judgment, bench trial, or jury trial; text content of briefs; the presence of an amicus brief), we can use a court's past decisionmaking patterns to predict the probability that a court will reverse each lower court decision. And the information embedded in

23. *See* Marin K. Levy, *The Mechanics of Federal Appeals: Uniformity and Case Management in the Circuit Courts*, 61 DUKE L.J. 315 (2011).

24. *Id.* at 331, 346.

25. FED. R. APP. P. 34(a)(2). At least sometimes, staff attorneys have substantial influence in this decision. *See, e.g.*, 5TH CIR. R. 34.13(A):

> The judges of the court screen cases with assistance from the Staff Attorney. When the last brief is filed, a case is generally sent to the Staff Attorney for prescreening classification. If the Staff Attorney concludes that the case does not warrant oral argument . . . . [t]he clerk then routes the case to 1 of the court's judges.

26. 4TH CIR. R. 36(a).

27. FED. R. APP. P. 35(a).

28. *See* Levy, *supra* note 23, at 365 n.330 (offering the Second Circuit as an example of a court that "almost never precirculates opinions beyond the original panel").

such a prediction is invaluable to restoring the circuit courts' central role in the justice system.

Specifically, statistical precedent provides courts with three critical pieces of information. First, it allows courts to identify which of its decisions—whether made by a staff attorney, law clerk, judge, or panel—are most incompatible with the court's collective jurisprudence. Second, statistical precedent lets a court know which appeals would likely be correctly decided even with limited attention: cases with very high or low statistical precedent represent the "easy" cases that are almost always decided the same way. Third, it allows the court to identify the "hard" cases that provide the most promising opportunities to develop the law; a statistical precedent close to 50% indicates that the governing law is insufficient to generate a judicial consensus as to the proper outcome.

While statistical precedent may have the capacity to transform the administration of justice, I offer a set of limited reforms for the more immediate future: (1) when a panel decision deviates widely from statistical precedent, the court should flag it and circulate it to nonpanel judges so that they have an opportunity to offer feedback and consider it for en banc review; (2) if such an outlier decision is made in an unpublished opinion, it should also be added to a public "high-risk" list so as to discourage abuse of this particularly controversial form of justice; (3) courts should stop using proxies (e.g., "pro se" as a proxy for "easy affirmance") when deciding which cases should be assigned to staff attorneys and instead use statistical precedent to identify the consensus affirmances and reversals that are most appropriate for assignment to central staff; and (4) judges should default to publishing opinions when statistical precedent is close to 50%.

This Article proceeds in five parts. Part I introduces the basic process of mapping a court's statistical precedent. In a nontechnical manner, I explain how machine learning can be leveraged to craft an individually tailored precedent for each case. I also compare statistical precedent to the traditional rule of precedent, discussing its relative strengths and weaknesses. Of particular importance is that traditional precedent becomes less effective as caseloads and court sizes grow, while statistical precedent increases in accuracy with the size of datasets. In brief, statistical precedent is a system of precedent suited for the modern world.

Part II reviews Marin Levy's "resource allocation framework" for assessing the allocation of judicial attention.[29] Levy argues that error correction tends to be maximized by conserving judicial attention when a case would likely be decided correctly without it.[30] Law development, too, suffers little if these cases receive minimal judicial attention, as the types of cases that can be decided correctly without judicial attention are unlikely to involve legal issues that need clarification.[31] I argue that Levy's framework is limited by its implicitly formalist treatment of error. "Error" is deeply contested, and the fact that judges have conflicting notions of error is a defining feature of adjudication.[32] What if some panels would assign error to a case and others would not? The framework also obscures the fact that judicial attention can be allocated to correct a circuit court's own errors and that it can occur in a multistage process. In short, it is not *cases* that need attention, but *decisions*—lower court and agency decisions, yes, but also staff attorney, law clerk, and panel decisions. The circuit courts do not employ a "Two-Track system."[33] It is a sprawling, multilevel system of review.

Part III presents an expanded conceptual framework. I introduce the concepts *degree of error* and *degree of instability*. The degree of error is the extent to which a decision departs from a court's collective judgment. For example, if a panel reverses a case that only 10% of possible panels would reverse, the panel's decision has a 90% degree of error, and the court's error-correcting function would be promoted by focusing the court's attention on such an outlier. A case's degree of instability is the extent to which possible panels would disagree as to its correct outcome. Instability is maximized where half of panels would reverse a case and half would affirm. I argue that a case with high instability is an opportunity to develop law: if the governing law cannot generate consensus among judges, it is also likely failing to provide society the ability to plan and organize its affairs.

Part IV empirically demonstrates that statistical precedent can usefully estimate each decision's degree of error and instability. I use

29. Marin K. Levy, *Judicial Attention as a Scarce Resource: A Preliminary Defense of How Judges Allocate Time Across Cases in the Federal Courts of Appeals*, 81 GEO. WASH. L. REV. 401, 422 (2013).

30. *See id.* at 414–20 (describing the judicial response to increasingly scarce resources).

31. *See id.* at 430–33.

32. For example, between 1995 and 2013, at least 40% of civil cases in the Ninth Circuit could have been decided differently if they had been assigned to one panel rather than another. Ryan Copus & Ryan Hübert, Detecting Inconsistency in Governance 5 (July 26, 2018) (unpublished manuscript), https://ssrn.com/abstract=2812914 [https://perma.cc/2M2X-VZXA].

33. RICHMAN & REYNOLDS, *supra* note 1, at xii.

a dataset of Ninth Circuit civil decisions made between 1996 and 2010 to build a model of the Ninth Circuit's statistical precedent.[34] I then use the model to estimate the degree of error and instability for each district court decision reviewed by the circuit court in 2011 and 2012 and validate the estimates by testing them against traditional indicators of error and law development. A circuit court decision with a high estimated degree of error is significantly more likely to have a dissenting opinion, negative subsequent appellate history, and negative analysis in future opinions. Furthermore, opinions disposing of cases with a high estimated degree of instability are published more often and cited more frequently. In summary, as judged by judges, statistical precedent can accurately identify erroneous decisions and opportunities for developing law.

Part V discusses the details of actually adopting a system of statistical precedent. In addition to elaborating on the set of reforms introduced above, I consider some of the core concerns with algorithm-aided justice. These concerns, I argue, are largely evaded by using algorithms to allocate attention rather than to automate or recommend decisions on the merits. I also explain how courts can obtain the most useful summary of their statistical precedent and overcome concerns that the coders of statistical precedent might embed their own normative preferences and biases. The Article concludes with a brief discussion of political feasibility.

I. AN INTRODUCTION TO STATISTICAL PRECEDENT

Both the basic process of mapping a court's statistical precedent and its ability to help with everyday issues of judicial administration are intuitively accessible. Below, I describe the general idea of statistical precedent and, in order to help build a basic understanding, compare it to the traditional rule of precedent.

A. Mapping a Court's Statistical Precedent

Imagine that a circuit court judge is worried about her court's distribution of judicial attention. Though she generally trusts the considered judgment of her circuit court colleagues, she worries that judicial judgment, including hers, is too often ill-considered: judges

34. The dataset does not include administrative agency or habeas corpus cases. The variables include the nature of suit, whether the plaintiff has legal representation, identity of the district court and judge, the district court's ABA rating, the outcome at the district court, the number of parties, and the number of major law firms. For more details on the dataset, see Copus & Hübert, *supra* note 32.

provide only cursory review of staff attorney decisions, spend little time thinking about those cases that they decide with unpublished opinions, rely heavily on law clerk bench memos, and are too often driven by ideological preconceptions and heuristics. They must sometimes err in their error correction. She is particularly concerned about improving in two domains. First, she is worried that she and her colleagues are not adequately reviewing recommendations by staff attorneys.[35] Although she cannot possibly provide each of those decisions with a comprehensive assessment, can she somehow flag likely errors and make sure that she at least provides those cases with her focused attention? Second, she is concerned that she should be more active in monitoring panel decisions. While her circuit is large enough that judges have stopped precirculating their opinions to the full court, she would like to be aware of decisions that are particularly unusual.

Perhaps her court's historical decisions can provide insight. If the court has generally reversed a particular type of case, maybe a staff attorney's or panel's decision to affirm that type of case is a good candidate for her focused attention. She tests the idea out with one of her recent cases. She begins simply: she looks up the reversal rate for civil cases with a pro se plaintiff filed in the last ten years. Panels have reversed 16% of such cases, significantly higher than she would have thought. That's useful information—she should probably be paying more attention to recommendations to affirm those cases. But the search seems too broad: she wants to know more about *this particular type* of case. She zeroes in: civil cases where there was a pro se plaintiff, a corporate defendant, federal question jurisdiction, decided on summary judgment, plaintiff prevailed, nature of suit is contract, decided by Judge Smith of the U.S. District Court for the Northern District of California on the report and recommendation of a Magistrate Judge Johnson, and the district court opinion was

35. Former Ninth Circuit Judge Alex Kozinski nicely explains the concern:

> [T]he circuit shares approximately 70 staff attorneys, who process roughly 40 percent of the cases in which we issue a merits ruling. When I say process, I mean that they read the briefs, review the record, research the law, and prepare a proposed disposition, which they then present to a panel of three judges during a practice we call "oral screening"—oral, because the judges don't see the briefs in advance, and because they generally rely on the staff attorney's oral description of the case in deciding whether to sign on to the proposed disposition. After you decide a few dozen such cases on a screening calendar, your eyes glaze over, your mind wanders, and the urge to say O.K. to whatever is put in front of you becomes almost irresistible.

Alex Kozinski, *The Appearance of Propriety*, LEGAL AFF. (Jan.–Feb. 2005), http://www.legalaffairs.org/issues/January-February-2005/argument_kozinski_janfeb05.msp [https://perma.cc/NS45-FQM3].

published. No results. That was obviously too specific, so she deletes the search parameters for nature of suit and magistrate judge. Six cases match the less restrictive search, and the circuit court reversed four of them (66%). She suspects that the sample size is too small to trust. Published district court opinions are uncommon in pro se cases: maybe she can get a bigger sample by eliminating publication as a search parameter. She gets 231 results with twenty-one reversals (about 10%). But is that the best estimate? Is opinion publication an important correlate of reversal that this estimate ignores? Should she try additional searches?

The judge is struggling to find the search query that optimizes the "bias-variance tradeoff."[36] Figure 1 helps to convey the concept. She wants a low bias, low variance estimate, as represented by the target in the upper left corner. Unfortunately, lower bias generally means higher variance, and lower variance generally means higher bias. Why? An unbiased estimate of a case's chances of being reversed uses all information about that case—it aims for the center of the target. But by using all of the information, the number of comparable cases (i.e., cases with the same characteristics) dwindles, and any estimate based on such a small number of cases is likely to be unreliable—our dart player is aiming for the center, but she has a shaky (high variance) hand. By ignoring some characteristics about the case of interest, say by leaving the fact that the plaintiff prevailed out of the search query, we increase the number of cases we are basing an estimate on, but we move the aim away from the center of the target, towards cases where the plaintiff did not prevail. The dart player's hand is steadier, but it is no longer aiming at the center.

36. *See* Scott Fortmann-Roe, *Understanding the Bias-Variance Tradeoff*, SCOTT FORTMANN-ROE (June 2012), http://scott.fortmann-roe.com/docs/BiasVariance.html [https://perma.cc/DH6Y-EH8T].

FIGURE 1[37]

Variance can be a serious issue because it increases exponentially as more characteristics are added to the search. In a world with extensive electronic records, the list of available characteristics can be almost endless, so this "curse of dimensionality" can be a troublesome problem.[38] For example, even with only ten dichotomous variables (e.g., decided by summary judgment or not, plaintiff prevailed or not, district court opinion published or not), there are 2^{10} (or 1,024) different types of possible cases. Even with a moderately sized dataset of ten thousand, we'd expect only ten of each type of case. With such small sample sizes, estimates would have extremely high variance.

Fortunately, we do not have to choose between adding a characteristic to the search inquiry or simply ignoring it. With techniques like multiple regression, we can *partially* add characteristics to the "search inquiry" (the quotes are now necessary because the *partial* addition of characteristics involves mathematical operations that are more sophisticated than a simple search inquiry, and we would be more accurate to now call it a statistical model). Rather than observing the reversal rate for the rare contracts case where the pro se plaintiff prevailed on summary judgment on the report and recommendation of a magistrate judge, we could instead start with the much more common civil case where the pro se plaintiff prevailed on summary judgment (contracts or not, on the report and

37. *Id.* at fig.1.

38. The phrase "curse of dimensionality," widely used in statistical conversations, was coined in RICHARD BELLMAN, DYNAMIC PROGRAMMING ix (2003).

recommendation of a magistrate judge or not). Worried that we have disregarded an important predictor of reversal (i.e., worried that we have taken on too much bias in the effort to reduce variance), we could try different methods of incorporating the case subject matter or magistrate judge as predictors. We might, for example, see how reports and recommendations of magistrate judges are associated with reversal rates for all cases and add that to our baseline estimate. Or perhaps we suspect that the association is unique for pro se civil cases, so we instead check how magistrate reports are associated with reversal for that subgroup of cases.

The problem is now even starker: With all of the choices about which variables to add, which to add partially, and how to add them partially, how can we possibly figure out the "search query"—the statistical model—with the best mix of bias and variance? In other words, how do we find the dart player with the optimal combination of aim and steadiness?

Machine learning provides a solution, effectively automating the process of creating statistical models and testing them for optimal accuracy. With a supply of predictor variables (e.g., nature of suit, prevailing party) and an outcome variable (e.g., reversal), we can let a machine train itself to identify which combinations of predictor variables are most helpful in predicting the outcome. Algorithms can learn from and adapt to the data, iteratively building models on subsets of data and testing themselves against different subsets to construct a predictive model. This is how the judge can find the best statistical answer to the question, "How often has the court reversed a case *like this one*?"[39] It is how we can best identify a case's statistical precedent.

Imagine, then, that the judge has access to a model built with machine learning algorithms and the universe of the court's decisions over the last five years. She enters all of the information for her case: in the last five years, her court has reversed 85% of similar cases. The staff attorney's recommendation to affirm now looks suspicious, and maybe she should take a closer look at the briefs. But she is still struggling to understand the meaning—and value—of statistical precedent. An analogy to traditional precedent can help build more intuition.

39. It is important to understand that "like this one" will generally not be cognizable—it is unlikely to refer to a set of cases with the same set of characteristics (e.g., contract cases where the plaintiff won in the Northern District of California on a motion for summary judgment), because machine learning will draw on information from other categories of cases to generate more accurate predictions.

B. Traditional Versus Statistical Precedent

The rule of traditional precedent, by which cases are resolved to conform to past decisions and, in turn, generate law to govern future cases, differs from statistical precedent in a number of obvious ways: traditional precedent is communicated in natural language, while statistical precedent is communicated in mathematical language; traditional precedent guides decisions on the merits, while I am arguing that statistical precedent should merely guide the focusing of attention on cases;[40] and traditional precedent is, at least ideally, based on legally relevant factors, while statistical precedent utilizes both legally relevant and irrelevant factors in summarizing historical decisions. Despite these differences, statistical precedent largely serves the same ends as traditional precedent. And given heavy caseloads and large courts, I will argue that it can serve those ends more effectively while simultaneously restoring the waning power of traditional precedent.

One standard justification for the traditional rule of precedent is that past decisions and reasoning embody a collective wisdom that an individual or small group of judges is unlikely to surpass. In this vein, Adrian Vermeule identifies four major theories: informational, evolutionary, traditional, and deliberative.[41] He succinctly states the core of each theory:

> [T]he aggregate judgment of many might employ dispersed information better than the judgment of one; the judgments of many heads, over time, might weed out bad policies or institutions through an evolutionary process . . . tradition might embody the contributions of many minds; finally, deliberation and argument among the many might contribute diverse perspectives, resulting in better policies or institutions than any one could devise.[42]

Whatever the merit of each individual theory, it is surely the case that previous decisions capture a valuable resource of collective wisdom.

Statistical precedent also captures collective wisdom. It efficiently summarizes how a court has decided similar cases, allowing the court to identify and focus attention on decisions that most depart from its collective wisdom. Of course, unlike traditional precedent, it does not capture the *reasoning* of previous decisions—only the

40. While statistical precedent could theoretically provide guidance on the merits, there are serious problems with employing it in such a manner. *See infra* Section V.C (addressing various concerns with using statistical precedent).

41. *See* Adrian Vermeule, *Many-Minds Arguments in Legal Theory*, 1 J. LEGAL ANALYSIS 1, 4 (2009).

42. *Id.*

outcome of that reasoning. But that relative deficiency comes with extraordinary benefits. Statistical precedent can be communicated in one single, objective figure, thereby evading a key limitation of traditional precedent: the fact that different judges can interpret and apply the same precedent in different ways.[43] By supplementing traditional precedent with statistical precedent, courts could add an objective indicator of a decision's deviation from judges' collective judgment.

The rule of traditional precedent is also justified by an appeal to the value of legal certainty. As pithily expressed by Justice Brandeis, "[I]n most matters it is more important that the applicable rule of law be settled than that it be settled right."[44] Thus, traditional precedent may be valuable even where collective wisdom is unwise. So too with statistical precedent, although in a less robust manner. Traditional precedent can be read and interpreted by businesses, organizations, and individuals as they try to plan their affairs and predict the outcomes of hypothetical or actual litigation. Statistical precedent is less valuable to noncourt actors. The problem is in the mismatch between the cases that form the basis of statistical precedent—those cases that make it into the appellate court—and the much larger set of cases that people want guidance on. Statistical precedent can accurately model the former but not the latter. Combined with the fact that it does not include the reasons for an outcome, statistical precedent would thus likely be of little direct use to potential litigants in evaluating the merits of their case. Nonetheless, insofar as statistical precedent helps the courts attend to and correct the decisions that depart furthest from its collective practices (including the practice of deferring to traditional precedent), it can promote consistent and predictable decisionmaking.

Traditional precedent also aids in assuring that like cases are treated alike. Similarly, statistical precedent, which allows a court to locate and funnel attention to the cases that have *not* been treated like similar cases, can help make sure that a court abides by the fundamental tenant of equality.[45]

43. *See* Karl N. Llewellyn, *Remarks on the Theory of Appellate Decision and the Rules or Canons About How Statutes Are to Be Construed*, 3 VAND. L. REV. 395, 396 (1950) ("[S]ince there is always more than one available correct answer [to a disputed issue of law], the court always has to select.").

44. Burnet v. Coronado Oil & Gas Co., 285 U.S. 393, 406 (1932), *overruled in part by* Helvering v. Mountain Producers Corp., 303 U.S. 376 (1938).

45. For the moment, I am minimizing a complication to this claim. Traditional precedent promises to assure equal treatment for cases whose facts are similar in *legally relevant* ways. The conception of "similarity" implied by statistical precedent is, at least as a technical matter, agnostic to the distinction between legally relevant and irrelevant facts. Statistical precedent

Finally, traditional precedent helps the court preserve resources. Justice Cardozo, for example, justified the rule of precedent on the grounds that "the labor of judges would be increased almost to the breaking point if every past decision could be reopened in every case, and one could not lay one's own course of bricks on the secure foundation of the courses laid by others who had gone before him."[46] One could say something similar of judicial attention and statistical precedent in the modern courts: without the ability to rely on the information embedded in datasets of past decisions, judges could not hope to find the current decisions that most need their attention.

Of course, the extent to which traditional precedent actually serves the above goals is a contentious issue. Advocates of the indeterminacy thesis doubt that precedent can meaningfully constrain decisions.[47] Political scientists have produced an essentially uncountable number of studies purporting to show the dominant influence of political ideology on judicial decisionmaking.[48] And there is no shortage of objections to those critiques of precedent.[49]

Whatever success traditional precedent has had in allowing judges to coordinate across time and cases to promote collective wisdom, legal certainty, equality, and efficiency, it is struggling under modern conditions. The simple evidence of that fact is that courts have all but stopped using it: as noted above, less than 10% of decisions now establish precedent.[50] The most obvious reason for the retreat from

merely summarizes the collective decisions of a court. Thus, if a court has been responsive to legally irrelevant case facts, statistical precedent will also tend to be responsive to those facts. In short, if a court's shared conception of error is faulty, statistical precedent will reflect that fault. In Section III.A, I argue that attention must be distributed according to some conception of error, and that whatever its faults, a court's collective conception of error is our best option. Furthermore, in Section V.C, I explain that because I am arguing that statistical precedent should only be used to allocate attention—not to automate or recommend decisions—this largely mitigates the concern that algorithms would cement historical faults into the justice system.

46. BENJAMIN CARDOZO, THE NATURE OF THE JUDICIAL PROCESS 149 (1st ed. 1921).

47. For an extended discussion of the indeterminacy thesis, see Lawrence B. Solum, *On the Indeterminacy Crisis: Critiquing Critical Dogma*, 54 U. CHI. L. REV. 462 (1987).

48. For an overview of research into extra-legal influences on judging, see Jeffrey J. Rachlinski & Andrew J. Wistrich, *Judging the Judiciary by the Numbers: Empirical Research on Judges*, 13 ANN. REV. L. & SOC. SCI. 203 (2017).

49. *See, e.g.*, Harry T. Edwards & Michael A. Livermore, *Pitfalls of Empirical Studies that Attempt to Understand the Factors Affecting Appellate Decisionmaking*, 58 DUKE L.J. 1895 (2009).

50. *See Table B-12: U.S. Courts of Appeals—Type of Opinion or Order Filed in Cases Terminated on the Merits, by Circuit, During the 12-Month Period Ending September 30, 2017*, *supra* note 9 (revealing that nearly 90% of decisions are now made in nonprecedential opinions). Although courts have dramatically reduced the production of formal legal precedent, there are arguments that nonpublished opinions create a body of informal precedent. Lauren Robel, *The Practice of Precedent:* Anastasoff*, Noncitation Rules, and the Meaning of Precedent in an Interpretive Community*, 35 IND. L. REV. 399, 401 (2002).

precedent is that judges simply do not have the time to carefully construct opinions fit for publication: traditional precedent is a victim of the courts' need to triage.[51] But there are many other likely reasons that traditional precedent is becoming so rare. A high rate of opinion publication might "add[ ] to the clutter, and sometimes confusion, of our multitudinous array of published decisions."[52] The increase in the number of judges may also make precedent collectively less intelligible: if court-developed law is Ronald Dworkin's chain novel,[53] then when written in a large court, it is a novel written by a cacophonous collection of authors.[54] Moreover, many of the cases that make up the modern courts' dockets—such as those reviewing social security or immigration decisions—may involve the type of bulk, fact-intensive areas of law that are particularly resistant to the constraining forces of precedent.[55]

Statistical precedent, in contrast, thrives under modern conditions. Because there are more judges contributing to a court's body of decisions, it can draw on a more diverse collection of viewpoints that strengthen the collective wisdom embedded in a court's decisions. And because there are more decisions, statistical precedent can more accurately track that wisdom: a larger dataset allows machine learning to dig deeper into the statistical connections between case variables and case outcomes. Statistical precedent is also robust to the fact-intensive areas of law that may make traditional

51. *See* Alex Kozinski, *In Opposition to Proposed Federal Rule of Appellate Procedure 32.1*, 51 FED. LAW 36, 38 (2004) ("[T]he process of anticipating how the language of the disposition will be read by future litigants and courts, and how small variations in wording might be imbued with meanings never intended—takes exponentially more time and must be reserved, given our caseload, to the cases we designate for publication.").

52. Boyce F. Martin, Jr., *In Defense of Unpublished Opinions*, 60 OHIO ST. L.J. 177, 197 (1999).

53. *See* RONALD DWORKIN, LAW'S EMPIRE 228–38 (1986) (comparing the interpretive processes of law and literature through the invented genre of "chain novel").

54. *See Unpublished Judicial Opinions: Hearing Before the Subcomm. on Courts, the Internet, and Intellectual Prop. of the H. Comm. on the Judiciary*, 107th Cong. 58 (2002) (statement of Alex Kozinski, Circuit Judge, U.S. Court of Appeals for the Ninth Circuit) ("We want to speak clearly through . . . published opinions. And given that we have over two dozen judges doing the speaking, plus 10 senior judges, plus visiting judges, you can actually get quite a cacophony going . . . .").

55. *See* Carolyn Shapiro, *The Limits of the Olympian Court: Common Law Judging Versus Error Correction in the Supreme Court*, 63 WASH. & LEE L. REV. 271, 293 (2006):

> The larger the body of law and the more fact-intensive the inquiry, identifying all or even most relevant factually analogous cases becomes difficult or impossible. With a mass of precedent from which to choose, judges may well "decid[e] intuitively . . . what is the right result and then scour[ ] legal texts for the [precedent] that will justify the intuition."

(alterations in original) (quoting John Braithwaite, *Rules and Principles: A Theory of Legal Certainty*, 27 AUSTL. J. LEGAL PHIL. 47, 63 n.61 (2002)).

precedent less effective. Because statistical precedent merely summarizes outcomes rather than the reasoning process by which those outcomes are justified, it is as much at home with standards as it is with rules.

Importantly, statistical precedent can also help strengthen traditional precedent. If time pressures and the deleterious effects of a cluttered jurisprudence keep judges from producing precedent at a high rate, they need to make sure that their law-developing efforts are spent wisely. And statistical precedent can alert courts to the cases whose outcomes are most unpredictable and, thus, likely most underdetermined by existing law.[56]

Although I hope the comparison to traditional precedent is usefully intuitive, a conceptual framework is needed to fully understand statistical precedent and how it can be crafted to best serve the administration of justice. Part II reviews the existing "resource allocation framework" for assessing the distribution of judicial attention, and Part III expands that framework so that we can better understand what statistical precedent can offer courts.

II. THE RESOURCE ALLOCATION FRAMEWORK

How should courts allocate their limited attention? Marin Levy has provided the most ambitious answer to that question, and her answer provides the basis for mine.[57] In this Part, I summarize her application of the resource allocation framework to the issue of judicial attention. But I also argue that the framework is limited by an overly formalistic treatment of "error" and that we should move from a case-based to decision-based framework.

A. The Basics

Most scholarly literature on the triaging of judicial attention has criticized what William Richman and William Reynolds termed a "Two-Track" system of justice: powerful litigants can expect their arguments to be heard, considered, and resolved by Article III judges, while the claims of powerless litigants will be resolved on the briefs by a staff attorney, getting only cursory review by actual judges.[58] These

56. *See infra* Section III.B (discussing how relative degrees of instability can signal areas where the law needs further development).

57. Levy, *supra* note 29.

58. RICHMAN & REYNOLDS, *supra* note 1, at xii.

"separate and unequal" tracks of justice are undoubtedly troubling.[59] And for many scholars, the answer is to end the need for triage: either increase the number of judges[60] or limit the number of cases[61] so the court can provide full judicial attention to each case.

But Levy asks us to be "realists."[62] Congress is unlikely to either radically increase the courts' supply of judicial attention or to decrease the demand for judicial attention. Scholars, therefore, need to start addressing whether and how courts can do better with their limited resources. In brief, how can the court use its main input—judicial attention—to maximize its two main outputs—error correction and law development?[63]

With respect to a court's error-correcting function, Levy proposes that courts conserve judicial attention when a case is likely to be decided correctly without it.[64] She proposes two categories of cases that would be likely to satisfy this criteria: "(1) those that are most likely to be reviewed effectively through a nonargument review process and (2) those that are least likely to have errors upon arrival at the appellate courts."[65]

In the first category, she proposes, are those cases that raise issues that the court repeatedly confronts.[66] As courts (including staff attorneys) become more familiar with the complexities of an issue, it should be easier for them to identify errors, and there should thus be little need for judicial attention.[67] As an example, she offers appeals from the Board of Immigration Appeals ("BIA") denying an asylum application.[68] Some circuits decide hundreds of such appeals annually, and most of the appeals involve the same issue: "[W]hether an adverse credibility finding by the BIA is supported by substantial evidence."[69]

For the second category of cases—those that are least likely to have errors—she proposes two promising subcategories.[70] First,

59. David C. Vladeck & Mitu Gulati, *Judicial Triage: Reflections on the Debate over Unpublished Opinions*, 62 WASH. & LEE L. REV. 1667, 1668 (2005).

60. *See* RICHMAN & REYNOLDS, *supra* note 1, at xiii.

61. *See* JUDICIAL CONFERENCE OF THE U.S., LONG RANGE PLAN FOR THE FEDERAL COURTS 19–20 (Dec. 1995), https://www.uscourts.gov/sites/default/files/federalcourtslongrangeplan_0.pdf [https://perma.cc/SXH5-L5KD].

62. Levy, *supra* note 29, at 401.

63. *See id*. at 424–25.

64. *See id*. at 431.

65. *Id*.

66. *See id.*

67. *Id.*

68. *See id*. at 431–32.

69. *Id*. at 432.

70. *See id*.

district courts are unlikely to have made errors in deciding frivolous appeals, such as those from tax protestors.[71] Second, cases that have already undergone a "meaningful layer of review" should be less likely to arrive in the appellate courts with error.[72] For example, she theorizes that Social Security cases should rarely contain a material error because they have already been reviewed by an administrative law judge, the Social Security Administration Appeals Council ("SSAAC"), and a district court before arriving in the appellate court.[73]

And which appeals are least likely to be important for advancing the court's law-development goals? Levy argues that they are largely the same cases that need the least error-correcting attention: frivolous appeals and those that repeatedly involve the same core issues (e.g., asylum applications), which are unlikely to need clarification of the law.[74]

Taking stock of the courts' current practices, Levy provides a preliminary defense: they seem to be placing the right cases in the low-attention track.[75] But she also stresses the tentative nature of her defense.[76] Perhaps the courts are depriving the wrong cases of attention. How could they know? She recommends that courts randomly select some of the appeals that are currently receiving limited judicial attention and provide them with more attention (e.g., assign them to chambers or track them for oral argument).[77] If the publication and reversal rates of those randomly selected cases turned out to be significantly higher than the cases that were not selected, it would provide evidence that the court's triage system was malfunctioning.[78]

In summary, Levy argues that judges should allocate more attention to cases that are likely to be erroneously decided without it. While her framework serves as the conceptual foundation for my paper, it leaves two core issues underdeveloped: What does it mean for decisions to be in "error," and how can courts actually find them?

71. *Id.*
72. *Id.* at 433.
73. *Id.*
74. *See id.*
75. *See id.* at 435.
76. *See id.* at 439.
77. *See id.* at 441.
78. *Id.* at 441–42.

B. The Limitations

"Error" is contested: judges disagree with other judges. Sometimes this disagreement is readily apparent, such as when a judge issues a dissenting opinion or a court reverses a panel decision in an en banc proceeding. But empirical research shows that judicial disagreement is far more prevalent than dissents or en banc decisions would suggest. Because cases are randomly assigned to panels, researchers can show that some types of panels systematically reach different outcomes than other types of panels. And the rate of disagreement can be striking. Cass Sunstein and coauthors found, for example, that a panel of three judges all appointed by a Democratic president is 86% more likely to decide in favor of the plaintiff in a gay rights case than a panel of all Republican appointees, 49% more likely to decide in favor of an affirmative action plan, and 46% more likely to decide for the plaintiff in a sex discrimination case.[79] And such high rates of inconsistency are not limited to politically salient issues: at least 40% of all civil cases in the Ninth Circuit could be decided differently based on panel assignment.[80]

The fact that judges disagree over whether a decision is in error poses challenges to Levy's framework. By whose conception of error should courts allocate attention? If she means that courts should limit their judicial attention when *all* of its panels would agree, the framework is a poor match for the scope of the problem. For example, it would provide little guidance for identifying which 10% of opinions should be published unless—implausibly—panels were in complete consensus in 90% of cases. And which 20% of cases should be tracked for oral argument? The courts resources are so constrained that they do not just need to know which cases they can safely pay less attention to—they also need to know which cases are especially in need of attention.

The first step in making the framework viable is choosing between two plausible options: courts could either allocate attention according to each panel's conception of error or according to some collective, court conception of error. Theoretically, panel-centric and court-centric conceptions could lead to vastly different allocations of judicial attention. Consider, for example, the assignment of cases to staff attorneys. Perhaps there are some lower court decisions that 10% of panels would reverse and that 90% of panels would affirm.

79. *See* CASS R. SUNSTEIN ET AL., ARE JUDGES POLITICAL? AN EMPIRICAL ANALYSIS OF THE FEDERAL JUDICIARY 20 tbl.2-1 (2007).

80. Copus & Hübert, *supra* note 32, at 18. The results exclude habeas and agency review.

Assuming that staff attorneys generally share the views of the collective court, staff attorneys would likely draft an opinion to affirm the case.[81] Under a panel-centric allocation scheme, these cases should be assigned to judicial chambers for 10% of panels and to staff attorneys for 90% of panels. But under a court-centric scheme, the cases would be assigned to staff attorneys regardless of which panel was assigned to the case.

The distinction might seem to be of mere theoretical interest, but it points to a second limitation: conceptual analysis can only provide a rough approximation of the decisions that need judicial attention. For example, Levy argues that cases that have already been meaningfully reviewed for error before they reach the appellate court, such as Social Security appeals, are good candidates for less attention.[82] This again belies an overly formalist view of error—might the circuit court's conception of error differ from the administrative law judge's, the SSAAC's, and the district court judge's conception of error? And even if most social security appeals do not need judicial attention, might there be some that do?

The rough results of conceptual analysis also obscure the possibility for allocating attention in a multistage process. For example, perhaps, as Levy argues, it is true that asylum appeals are likely to be decided correctly by staff attorneys because they repeatedly raise the same legal issues.[83] Nonetheless, staff attorneys might still make mistakes in some cases, and conceptual analysis does little to help us identify those mistakes. With more precise estimates of the "correct" decision, judges could allocate their attention to reviewing the most at-risk staff attorney opinions.

Levy's implicit assumption that judges are in consensus as to error further masks the fact that judicial attention also needs to be allocated to the work of other judges in order to correct the courts' own

81. *Cf.* Richard Posner, *Will the Federal Courts of Appeals Survive Until 1984? An Essay on Delegation and Specialization of the Judicial Function*, 56 S. CAL. L. REV. 761, 775 (1983):

> Because the staff attorney is not selected by the individual judge, he owes his loyalty to the court as a whole (perhaps too indistinct an entity to command much loyalty), rather than to the individual judge to whom he is from time to time assigned. There can be no assurance that the staff attorney will share the outlook and values of that judge, and he will not have a chance to acquire that outlook and those values, or at least understand them sympathetically, by working intimately with the same judge over a period of months or years. For these reasons the staff attorney will ordinarily be less able to function effectively as a judge's alter ego . . . .

82. *See* Levy, *supra* note 29, at 433.

83. *See id.* at 431–32.

errors and aid the development of a coherent and lucid body of law.[84] Most obviously, a circuit court can allocate attention by evaluating a case for en banc review, and, perhaps, actually taking the case en banc. But other forms of judicial attention might also help correct a circuit court error. For example, some courts make drafts of opinions available to the entire court, providing an opportunity for off-panel judges to provide feedback.[85] How should a court allocate these forms of attention? The next Part enriches the resource allocation framework to remedy these limitations and make room for understanding the role that statistical precedent can play.

III. EXPANDING THE FRAMEWORK: ERROR AND INSTABILITY

I adopt the basics of Marin Levy's resource allocation framework, but I set out more explicit targets for advancing the court's core functions. I argue that a court can generally promote its error-correcting function by focusing attention on decisions with a high *degree of error*, and that a court can generally promote its law-developing function by focusing on decisions with a high *degree of instability*.

A. The Degree of Error

It may seem awkward to speak of decisions having "degrees" of error, but I do not think it should. Assessing error can be a difficult task. In deciding whether a decision should be reversed, a judge might consider precedent, statutes, legislative history, policy, values, and

84. The U.S. Courts of Appeals have played a critical role in maintaining quality, predictability, and consistency of decisionmaking in the relatively decentralized and high-volume federal district courts. But as the circuit courts have themselves transformed into behemoth systems of adjudication, a question presents itself: Who will correct the circuit courts' errors? The Supreme Court confesses to have relinquished the job. *See, e.g.*, Stephen G. Breyer, *Reflections on the Role of Appellate Courts: A View from the Supreme Court*, 8 J. APP. PRAC. & PROCESS 91, 92 (2006) (noting that the Supreme Court "is not a court of error correction"). But by strategically allocating judicial attention, the courts of appeals could serve as robust correctors of their own errors.

85. *See, e.g.*, Marsha S. Berzon, *Dissent, "Dissentals," and Decision Making*, 100 CALIF. L. REV. 1479, 1490 (2012) ("Not infrequently, an off-panel judge will circulate a memorandum to the panel identifying what that judge views as an error in the panel's opinion and suggesting revisions."). But judges might also have to choose which panel opinions to even look at. *Id.* at 1490 n.49:

> In some circuits, draft opinions are circulated to the entire court before they are published. *See, e.g.*, 7TH CIR. R. 40(e). The Ninth Circuit does not adhere to this practice because of our size. We do, however, precirculate summaries of opinions, and we are quite receptive to altering opinions after publication based on feedback from our colleagues.

how the decision might fit with future decisions of the court—and a judge might also consider whether or how much each of these should even be considered. We should thus not naively treat error as dichotomous: any assessment that a decision is in error is, or at least should be, implicitly accompanied by a degree of certainty with respect to that assessment. And we are rarely, if ever, completely certain that a decision should or should not be reversed. If prompted to think about our assessment that a particular case should be reversed, we may, for example, admit to lingering doubts about the scope of our legal research, the strength of the policy analysis, the propriety of consulting legislative history, or even the wisdom of our motivating values. Thus, after reflection, we might think that a decision is *more or less* in error. Though our legal systems may often operate on ones and zeroes, we should be honest and humble enough to admit that our assessments do not. To speak more honestly, it is better to say that the court's error-correcting function is about making sure decisions that are *more* in error are reversed.

While individuals can make their own degreed assessments of error, how should a court, as a collective entity that must allocate attention to correcting errors, make assessments? My argument is that it should try to aggregate the assessments of its panels. More specifically, I define a lower court or agency decision's *degree of error* as the percentage of all possible panel combinations that would reverse a decision if they were to carefully evaluate it.[86] Importantly, any decision by a circuit court (e.g., by a staff attorney, law clerk, or panel) also has a degree of error. If the circuit court decision is to affirm, its degree of error is the same as the lower court or agency decision. If the decision is to reverse, the degree of error is the opposite. For example, if a staff attorney's opinion recommends reversing a lower court decision that has a 20% degree of error, the staff attorney opinion has an 80% degree of error.

This conception of error has a number of appealing normative features. The first is epistemic. Tying it to collective judicial judgment leverages the wisdom of a wise crowd.[87] Federal circuit court judges

86. An even better definition would aggregate each individual panel's degreed assessment of error. But the ultimate goal will be to estimate the hypothetical decisions, and panels do not provide their degreed assessments in the real world—they either disturb a lower court decision or affirm it. I thus settle for the current definition, which, I will argue, can plausibly be estimated.

87. There are multiple ways to understand the epistemic benefit. For one, if we assume that each panel is better than a coin flip at correctly deciding cases, Condorcet's Jury Theorem shows that the probability of getting the correct answer increases with the number of votes. Here, I invoke the polling model of the theorem, as described by Paul H. Edelman. Paul H. Edelman, *On Legal Interpretations of the Condorcet Jury Theorem*, 31 J. LEGAL STUD. 327, 333 (2002).

are among the most respected jurists in the nation and, as designated experts in law, their collective assessment merits substantial epistemic deference.[88]

A second benefit of tying the degree of error to the judgment of circuit court judges is democratic legitimacy.[89] Federal circuit court judges, of course, are nominated by the president and confirmed by the senate. Thus, even where we strongly disagree with the collective judgment of circuit court judges in some areas of law, our disagreement does not have the same stamp of institutional legitimacy. Efforts to promote a court's error-correcting function should focus on judicial conceptions of error. We may strive to change a court's judgment, either through arguments directed at its existing members or through efforts to have judges appointed whose assessments of error better match our own, but we should hesitate to undermine (or prevent improvements in) a court's error-correcting function merely because we disagree with its conception of error.

To the extent that one is unpersuaded by the epistemic or legitimacy benefits of a court-centric definition of error, its potential for reducing inconsistency in decisionmaking may warrant deference. Inconsistent decisionmaking, whether due to the idiosyncrasies of different panels' judgments[90] or simple panel oversights,[91] undermines

88. I do not want to shy away from the claim that collective judicial judgment is, on average, superior to an individual panel's judgment. While I discuss other benefits of focusing judicial attention on the decisions that the highest percentage of panels would reverse, statistical precedent loses much of its appeal if one is not convinced that a high level of judicial support for reversal is a good indication that a decision should be reversed. Curiously, there seems to be a tendency to mentally foreground those imagined situations where our own judgments are in the minority. For example, we readily imagine those situations where most judges would decide against (or in favor) of a plaintiff in an employment discrimination suit, but where we would bravely and wisely rule in favor of (or against) the plaintiff. Why should our brave and wise decision be subject to extra judicial scrutiny!? But for most of us, most of the time (and for most actual judges, most of the time), the much more realistic concern is that a miscarriage of justice escapes notice.

89. *See* Michael B. Abramowicz, *En Banc Revisited*, 100 COLUM. L. REV. 1600, 1602 (2000):

> Just as we structure legislatures around majoritarian principles, so too, I will argue, should we seek to ensure that when a panel reaches a decision, it is the decision that a majority of all judges on the courts of appeals would reach if given adequate time to consider the issue. A decision is thus "correct" if it is the hypothetical majoritarian one.

(footnote omitted).

90. *See, e.g.*, Kleinfeld Statement, *supra* note 17, at 7:

> No district judge and no lawyer can, by reading even a few hundred of our decisions, predict what our court will do in the next case. Even if the decisions could be read, there are over 3,000 combinations of judges who may wind up on panels, so the exercise would not be worth the time. At best, the bar can predict that we will restate our clear holdings as controlling law, though different panels may apply the same holdings to similar facts in different ways. The disparateness will naturally be higher in unpublished dispositions.

the ability of lower courts, litigants, businesses, and individuals to predict and comply with legal requirements. Focusing judicial attention according to a collective conception of error—even if that conception is flawed—can help a court reduce inconsistency and promote predictability.

Dissemination of decisions' degrees of error could also increase a court's total supply of judicial attention by increasing the expected benefit of extra judicial effort. Consider, for example, a judge who wishes to be more active in the court's error-correcting role. She has thought about trying to review more of the staff attorney drafts as well as some of her colleagues' drafts.[92] But the universe of options is overwhelming, and she figures that she would be wasting her time if she simply selected opinions to review at random. She might thus choose to proceed as normal, quickly checking the staff-attorney drafts that are her official responsibility and paying attention only to her assigned panel's cases. But if she had access to cases' degrees of error and could thus readily identify a subset of more troublesome decisions, she might decide to invest the additional effort.

While I propose that a court should *generally* focus judicial attention on decisions with higher degrees of error if the court's goal is to promote error correction, a higher degree of error may not always be a good target for judicial attention. Cases with an extremely high degree of error may not—at least immediately—warrant judicial attention. For example, while a court could allocate attention to make sure that a lower court decision with a 95% degree of error is reversed, staff attorneys would also likely reverse such an "easy" case. It could thus make sense to assign the case to a staff attorney. But the general rule that judicial attention should be allocated to decisions with high degrees of error would immediately come back into play if the drafted opinion unexpectedly recommended affirming the lower court decision:

91. *See, e.g.*, Huang, *supra* note 13, at 1130–37 (providing evidence that the Second and Ninth Circuits reduced their reversal rates in civil cases once they were overwhelmed by immigration appeals).

92. *See* Berzon, *supra* note 85, at 1490 ("Not infrequently, an off-panel judge will circulate a memorandum to the panel identifying what that judge views as an error in the panel's opinion and suggesting revisions."). At least without the aid of statistical precedent, many courts can do little to review panel opinions. *Id.* at 1490 n.49. Commentators have stressed the importance of cross-panel sharing and feedback. Robert A. Leflar recommends that opinions

> be circulated to all the judges on the entire court under an arrangement by which other judges may within a specified short time report their objections to it with requests that the original panel reconsider its position. The panel would not be bound to do so, but could. This arrangement at least would give all the judges some opportunity for input into the original panel's ultimately authoritative precedential decision.

Robert A. Leflar, *The Multi-Judge Decisional Process*, 42 MD. L. REV. 722, 729–30 (1983).

the panel should prioritize and intensely review a staff attorney decision with such a high degree of error.

The degree of error may also be an imperfect target for judicial attention insofar as it does not track the importance of an error. For example, the lower court's decision in a *small* contractual dispute may have a 70% degree of error while another district court's decision in a *large* contractual dispute may have only a 60% degree of error. One could reasonably believe that a court's error-correcting function is better served by targeting judicial attention at the large contractual dispute even though it has a lower degree of error. Or perhaps a district court decision has a lower degree of error than another, but the former decision made multiple errors while the latter made only one. One might again reasonably believe that it is more important to correct the decision with more errors. But there is little reason to think that the possible disconnects between the degree of error and its importance would be systematic (i.e., degree and importance of error are unlikely to be negatively correlated), so degree of error would still, on average, provide a good target for judicial attention if the court's goal is error correction.[93]

B. The Degree of Instability

Like error, the judicial development of law is the subject of contentious debates. Should courts develop law with small, incremental steps, or should their judicial opinions provide broad guidance in an effort to resolve issues beyond those that are immediately presented by the case under consideration? Should they incorporate policy analysis into the law? Should they promote standards or rules? We can skip these questions, as there is little need

93. Note that I make no claim about *how high* a decision's degree of error must be to warrant the court allocating attention to it in order to correct the error. For example, attentiveness to decisions with less than a 50% degree of error may promote error correction despite the fact that most panels would not believe the decision should be reversed. In fact, it is even possible that there is *no lower court decision* that a majority of panels would assess as in error. There may nonetheless be reasons that the court should continue to reverse cases with a higher degree of error. For one, the possibility of reversal could be important for incentivizing district court judges and agencies to do better: even if most panels would not reverse a particular decision, it is possible that the district judge or agency could have decided the case such that an even larger percentage of panels would favor affirming. For example, instead of resolving the case on a motion to dismiss, the district court judge could have permitted discovery and resolved the case on a motion for summary judgment instead, perhaps satisfying even more panels' notions of justice. *See* Jonah B. Gelbach & David Marcus, *Rethinking Judicial Review of High Volume Agency Adjudication*, 96 TEX. L. REV. 1097, 1101 (2018) (arguing that judicial review of agency decisions can help agencies identify and fix systematic problems). Regardless, the issue is largely academic. The ultimate goal will be to estimate each decision's degree of error, and estimates will be not be sufficiently precise to support a debate on such a fine-tuned issue.

to weigh in on *how* the law should be developed in order to suggest *where* it should be developed. And on this issue, courts already provide guidance in their rules on opinion publication. If, according to those rules, judges should develop law where the resolution of a case "establishes, alters, modifies, clarifies, or explains a rule of law,"[94] then it stands to reason that a court can promote its law-developing function by directing judicial attention toward cases where the relevant law is most in need of being established, altered, modified, clarified, or explained.

I propose that judges should seek to make these adjustments to law in cases where judicial assessments of error are most likely to conflict. If the current state of the law is insufficient to generate consensus among judges as to the correct outcome in a case, it is a good indication that the law needs development: lower court judges, litigants, businesses, and individuals are also likely to be confused as to what the law requires of them. More specifically, I propose that a court can promote its law-developing function by focusing attention on the cases with a high *degree of instability*, defined as the percentage of all possible panel combinations whose decisions would conflict with the majority of hypothetical panel decisions.[95] Thus, the degree of instability is maximized at 50%, where 50% of panels would reverse and 50% of panels would affirm a case. In contrast, if 80% of hypothetical panels would decide a case in a given way, the degree of instability is only 20%.

Of course, a higher degree of instability may not always represent a better opportunity for developing law. First, the degree of instability is not necessarily related to its importance. For example, even if judges widely disagree over the outcome of a case, the case's fact patterns may be so far removed from any that are likely to occur in the future that development of the law would have little practical effect. Or a case may be of such public importance that it should be decided in a published opinion regardless of its contribution to law.[96] But there is little reason to believe that the degree and importance of instability systematically conflict, so instability should still, on average, provide a good target for the development of law.

94. 4TH CIR. R. 36(a).

95. The term "instability" may seem like an odd choice over more natural words like "disagreement," "dissension," or "conflict." But terms like "disagreement" suggest that panels openly reach conflicting decisions. "Instability" is meant to stress the extent to which assessments may fluctuate with different panels, whether panels are aware of that fact or not.

96. Indeed, circuit publication rules generally include a provision addressing public importance. *See, e.g.*, 9TH CIR. R. 36-2(d) (requiring publication if the disposition involves "a legal or factual issue of unique interest or substantial public importance").

Second, instability may be more or less remediable. For example, in some cases, curing instability might require addressing so much complexity that the attempt to develop law would likely muddy the waters. Or there may be wisdom in allowing legal issues to percolate and develop in the district courts rather than jumping at the first chance to resolve an unfamiliar issue,[97] especially if the panel lacks expertise in the area.[98] Alternatively, issues in an unstable case may be so intensely contested that any attempt to provide clarity would fail to build judicial consensus, engendering more or less veiled defiance of precedent instead. And, unlike the importance of instability, there are a priori reasons to think that remediability and degree of instability *are* negatively correlated: complexity, unfamiliarity, and intensity of disagreement may all be *causes* of existing instability.[99] Nonetheless, whether judges choose to resolve instability or not, they should at least be aware of it.[100] Knowledge of its existence and the presumed toll it takes on those who must plan affairs in law's shifting shadow should inspire judges to more deeply reflect and communicate in an earnest search for a way forward.[101]

97. *See, e.g.*, Frederick Schauer, *Do Cases Make Bad Law?*, 73 U. CHI. L. REV. 883, 915 (2006) (discussing the possibility of "delaying the very process of rulemaking until enough cases arose such that the rulemaking body could have the benefit of having seen multiple examples of some larger problem").

98. *See* Edward K. Cheng, *The Myth of the Generalist Judge*, 61 STAN. L. REV. 519, 548–50 (2008) (discussing benefits of subject-matter specialization in opinion authorship).

99. It is not difficult to imagine ways that complexity, unfamiliarity, and intensity of disagreement could each be a cause of instability. For example, in factually complex areas of law, where unique fact patterns can be difficult to account for ex ante, decisionmaking may be resistant to the constraints of rule-based precedent. *See generally* Louis Kaplow, *Rules Versus Standards: An Economic Analysis*, 42 DUKE L. J. 557 (1992). Thus, high instability may exist *because* factual complexity prevents the development of precedent that could reliably constrain judges. Unfamiliarity may cause instability because ideology may fill gaps where data is sparse, and allowing district courts to assess and develop arguments could help promote consensus at the circuit level. Finally, intensity of disagreement may have caused judges to avoid clarifying law in the past such that those intense disagreements remain unresolved.

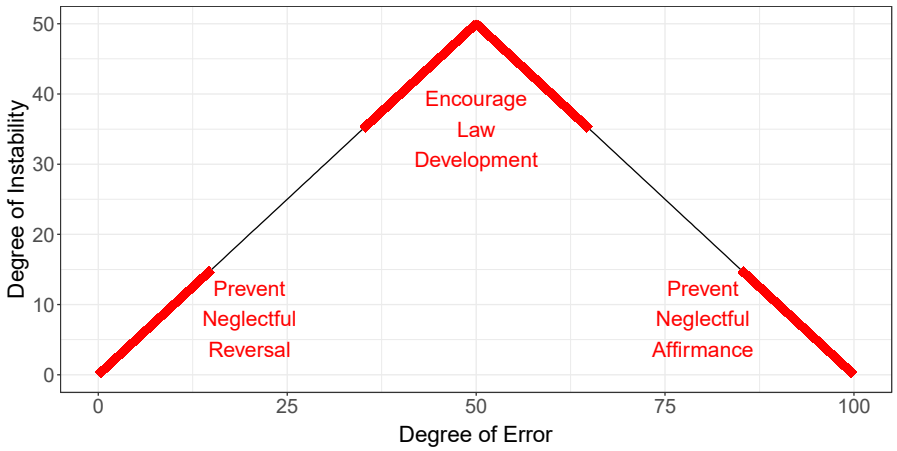
100. Consider, for example, possible ways forward in the face of high complexity. If factual complexity is the cause of high instability in a set of cases, it may indicate that the court should establish a different level of deference. If reversal is little more than a coin flip, it is not clear that circuit court review is accomplishing much. Reducing or increasing deference could help bring judges to the same page.

101. *See, e.g.*, Patricia M. Wald, *The Problem with the Courts: Black-Robed Bureaucracy, or Collegiality Under Challenge?*, 42 MD. L. REV. 766, 785 (1983):

> As it is now, except with each panel, judges learn of each others' views only through circulated written opinions which, in the court's pressured work environment, often gain more dust than readership. It might make sense for the judges of the court to meet occasionally to discuss areas of law in the circuit that may need clarification, or have been left a bit murky. The purpose of sharing views on such topics would not be to establish a fixed agenda for action and definitely not to decide abstract issues. Rather, its purpose would be to make us more sensitive to our colleagues' interests

Note that instability is a simple function of error, as illustrated in Figure 2. At low degrees of error, the values increase together, but they diverge from one another at high degrees of error. We can think of decisions with either extremely high or low error as the "easy" decisions—essentially all panels agree on what the correct outcome is.[102] Decisions with extremely low degrees of error are easily identifiable affirmances, and decisions with extremely high error are easily identifiable reversals. Judges are unlikely to find good opportunities for developing law at either end: if precedent is meant to build predictability in legal systems, there is little room for creating that predictability where it already exists.[103] But, of course, this means that the general rules I have suggested—that courts focus attention on decisions with a high degree of error to promote error correction and focus attention on decisions with a high degree of instability to promote law development—can be in tension with one another.

FIGURE 2: DEGREE OF ERROR AND INSTABILITY



Though the tension between the court's error-correcting and law-developing functions is real, it can be alleviated by a multistage

and views, and perhaps to establish a general aura of agreement on our responsibility, as a court, to identify and to elucidate particular subjects.

102. This Article does not engage with the indeterminacy thesis. An "easy" case, used in the sense here, may still be legally indeterminate according to some conceptions of the law. For example, judges might generally agree on the outcome of a case due to widely shared policy views—views that might not count as "legal" sources according to some philosophies.

103. *See* Frederick Schauer, *Precedent*, 39 STAN. L. REV. 571, 597–98 (1987) ("The most commonly offered of the substantive reasons for choosing strong over weak precedential constraint is the principle of predictability.").

approach to allocating attention.[104] Judges need not initially allocate their attention to reviewing lower court decisions that have an extremely high degree of error—they can conserve their attention by letting staff attorneys or law clerks make a first effort, waiting to allocate their attention and review the decisions that do not reverse those cases. Of course, the review of those high-error staff attorney or law clerk drafts for error would not tend to promote law development, but we should expect such errors to be rare.

With the conceptual framework in place, we are now in position to understand the value of statistical precedent. It is a way to estimate each case's degree of error and instability. Although necessarily based on datasets of historical decisions, it can be used to solve a problem of prediction: How would the court's current judges, as a collective, apply existing law to resolve each case? As I show in the next Part, statistical precedent is surprisingly effective at doing so.

IV. AN ILLUSTRATION: THE NINTH CIRCUIT'S STATISTICAL PRECEDENT

In order to demonstrate the ability of statistical precedent to usefully predict the degree of error and instability of future decisions, this Part uses the Ninth Circuit's statistical precedent between 1996 and 2010 (the training set, with 16,357 observations) in order to estimate the degree of error and instability for each lower court and panel decision in 2011 and 2012 (the test set, with 1,890 observations). I then show that the panel decisions with higher degrees of error were indeed more likely to be accompanied by traditional indicators of error: they were more frequently accompanied by dissents, had more subsequent negative appellate history, and were more frequently subject to negative analysis in future opinions. Higher instability estimates were also associated with law development: cases with higher instability were more frequently published, and published decisions with higher instability were cited more frequently.

104. Here, I focus on possible divergences between the decisions in error and the those that provide good opportunities for developing law. A court's error-correcting and law-developing goals could be in tension even where the decisions completely overlap (e.g., where there are no cases with greater than 50% degree of error), simply because time spent developing law is time not correcting errors (and vice versa). We could, for example, imagine a court purely dedicated to quickly correcting as many errors as possible and never authoring precedential opinions. Alternatively, we could imagine a court dedicated to writing comprehensive and high-quality opinions in only the most important cases. I do not address the relative importance of error correction as opposed to law development, as judges undoubtedly have more informed views than I do about the issue. My argument is that whatever split judges choose, knowledge regarding cases' error and instability would be instrumental.

The dataset is from the universe of Ninth Circuit docket sheets for civil cases between 1996 and 2012.[105] A colleague and I wrote a computer script to extract key information from each docket sheet and linked it with the Federal Judicial Center's biographical directory of judges. Variables included the case subject matter, the prevailing party at the district court, the identity of the district court judge, the ABA ratings of the district court judge, the number of parties, the presence of repeat players (e.g., parties who frequently litigate appeals), whether there was federal question or diversity jurisdiction, the district court magistrate judge, whether a party was pro se, details about a party's legal representation (e.g., city attorney, LLP, LLC, Department of Justice), and the votes of each judge on a panel.

With full access to the Public Access to Court Electronic Records database, we could radically expand the collection of variables in order to improve the accuracy of statistical precedent. We could include, for example: whether the case was decided pursuant to a motion for summary judgment, a motion to dismiss, or a trial; computerized grades of litigant briefs; district-level statistical precedent; summaries of a district court opinion's citation network; and information about the standard of review. But my aim here is only to show that even with a more limited set of variables, a model of statistical precedent can locate errors and law-development opportunities. Later in this Article, I explain how courts can move beyond the proof of concept and obtain a transparently constructed, implementation-quality model of a court's statistical precedent.[106]

A. Modeling the Ninth Circuit's Statistical Precedent

Because this is only a proof of concept, I keep the technical details to a minimum. In short, I used the R "SuperLearner" package to build an initial model of the court's statistical precedent between 1996 and 2010.[107] The algorithm searches over multiple models, iteratively building each model on a subset of the training set and evaluating each model's predictions on a different subset to select a

105. Habeas cases and cases reviewing agency decisions are excluded. At the time of writing this Article, the process of extracting variables from the docket sheets for these cases is not yet finished. For a more complete description of the dataset, see Copus & Hübert, *supra* note 32.

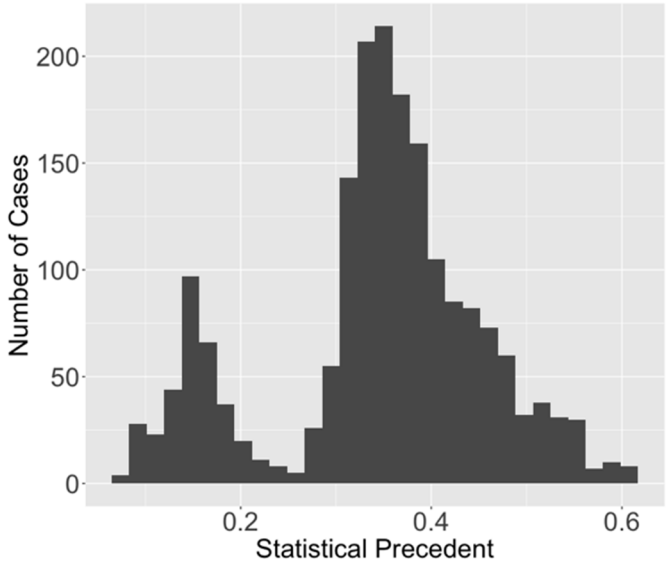
106. *See infra* Section V.A.

107. There is abundant literature on the SuperLearner package. For a particularly gentle introduction, see Daniel Gremmell, *Ensemble Learning in R with SuperLearner*, DATACAMP (Feb. 20, 2018), https://www.datacamp.com/community/tutorials/ensemble-r-machine-learning [https://perma.cc/7HWD-TUVB].

model that has the best mix of bias and variance for predicting reversal in new datasets (e.g., future cases).[108]

This initial model of statistical precedent is well calibrated for predicting the court's decisionmaking in 2011 and 2012. Figure 3 shows the distribution of statistical precedent, or the predicted proportion of hypothetical panels that would reverse each lower court decision. Cases are clustered around estimates of 30% to 40%, although cases have estimates as low as 7% and as high as 61%. On average, the estimates are accurate: regression results in the test set indicate that a 1% increase in estimated degree of error is associated with a 1.06% increase in the reversal rate.[109]

FIGURE 3:
INITIAL MODEL OF STATISTICAL PRECEDENT (2011–2012 TEST SET)



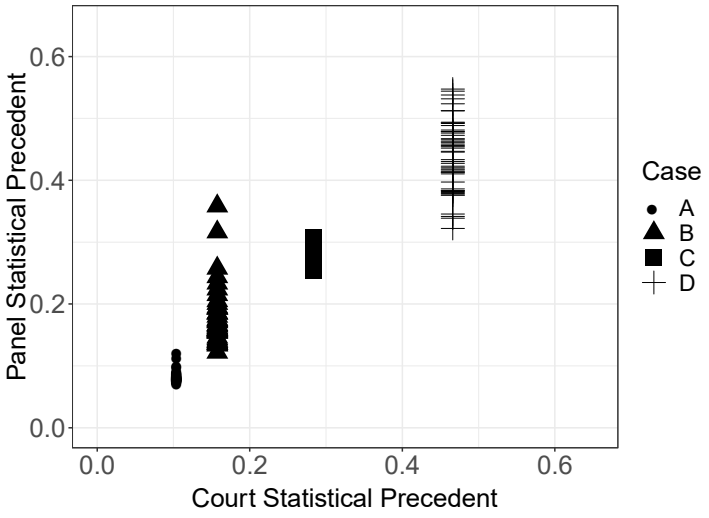
But we can improve on this initial model of statistical precedent by incorporating predictions about how different panels would decide each case. Figure 4 can help build intuition. It displays the relationship between the initial statistical precedent for four cases

108. SuperLearner, like most machine learning techniques, uses a process of cross validation to test the accuracy of candidate models and choose a weighted combination of multiple models with the optimal mix of bias and variance. *See id*. I include a LASSO regression, Random Forest, and Gradient Boosting Machine as candidate models. The Gradient Boosting Machine generated the lowest cross-validated mean squared error and received all of the weight.

109. Standard error = 0.09%. P-value = 0.000.

(*A*, *B*, *C*, and *D*) and predictions for how likely fifty different panels are to reverse each case. The spread of the panel predictions provides new information. Consider, for example, case *B* and case *C*. The spread of panel predictions for case *B* is high, suggesting that panels would frequently reach conflicting outcomes in case *B*. The spread of panel predictions for case *C* is less pronounced. This is unexpected—a case with an estimated error of about 28% (case *C*) should be more unstable than a case with an estimated error of about 16% (case *B*). The spread of the panel predictions suggests that the statistical precedent should be closer to 50% for case *B* and further from 50% for case *C*.

FIGURE 4:
PANEL-SPECIFIC STATISTICAL PRECEDENT IN FOUR CASES[110]



The technical details of estimating each panel's probability of reversing each case and then using the spread of those predictions to adjust the initial model of statistical precedent are moderately complex, and I have placed them in the Appendix.[111] The important takeaway is that I estimated how one thousand different panels would have decided each case in the test set, and the variability of those estimates can provide extra information about each case's degree of instability and, hence, degree of error. But is there any reason to

110. For illustrative purposes, I have slightly magnified the variation in the panel statistical precedent.

111. *See infra* Appendix, Part B.

believe that the estimates of one thousand different panels are informative? While we cannot observe whether the vast majority of those estimates track reality in any meaningful way, we can observe whether they help predict decisions of the panels that are *actually assigned* to decide a case.

Figure 5 displays the distribution of the difference between the assigned panel's predicted probability of reversing a case and the preliminary estimate of the collective court's degree of error. The average of these panel-court deviations is zero, which one would expect given that "the court" is ultimately an aggregation of its panels. But there are also substantial deviations, and Table 2 shows that those deviations are indeed predictive of actual reversal: controlling for the court's estimated degree of error, a 1% increase in panel-court deviation results in approximately a 1% increase in the reversal rate. This provides confidence that the estimates for the one thousand panels contain useful information.

FIGURE 5: DISTRIBUTION OF DIFFERENCES BETWEEN PANEL AND COURT STATISTICAL PRECEDENT (2011–2012 TEST SET)

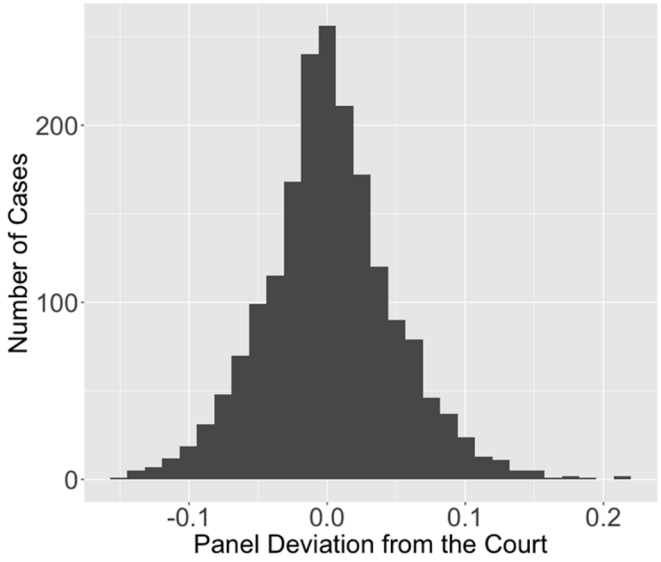


TABLE 1: REGRESSING REVERSAL ON PANEL DEVIATIONS FROM STATISTICAL PRECEDENT (2011–2012 TEST SET)

	Beta Estimate	*Standard Error*
Initial Statistical Precedent	1.04%***	0.09%
Panel Deviation	1.22%***	0.21%

*0.05, **0.01, ***0.001 statistical significance.

The panel-specific predictions also prove useful in updating the initial model of statistical precedent. Table 2 displays the results of a regression testing the predictive capacity of the adjustments. Controlling for the initial statistical precedent, the adjustments are associated with a 1.6% increase in reversal rate.

TABLE 2: REGRESSING REVERSAL ON ADJUSTMENTS TO STATISTICAL PRECEDENT (2011–2012 TEST SET)

	Beta Estimate	*Standard Error*
Initial Statistical Precedent	1.06%***	0.09%
Adjustments	1.64%***	0.46%

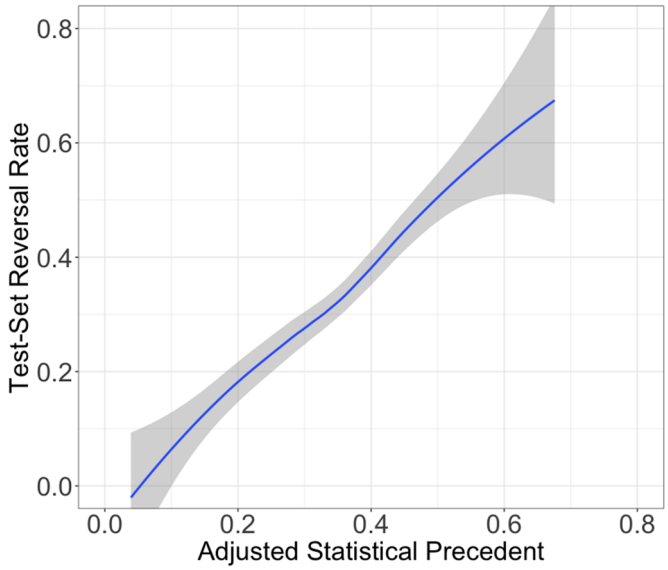
*0.05, **0.01, ***0.001 statistical significance.

Figure 6 shows that the adjusted model of statistical precedent matches the 2011–2012 court's actual reversal rates, providing the first piece of evidence that statistical precedent can accurately represent the current court's collective judgment. But it is not enough.[112] To show that statistical precedent can help the court, we need to show that decisions that deviate from statistical precedent are indeed incompatible with a court's jurisprudence, and we need to show

112. In technical terms, it is evidence that the predictive model is well calibrated, but it does little to show the model's discriminatory power. Even if *X%* of cases with an estimated *X%* error are reversed, those cases may have true error degrees that are far from *X%*. In the worst-case scenario, *X%* of those cases have true error of 100% (all panels would consistently reverse those cases) and *100 - X%* of those cases have true error of 0% (all panels would consistently affirm those cases). The estimates would be most useful to the court if they were accurate and could thus cleanly identify the decisions that more panels would disagree with. Insofar as the estimates are inaccurate, they would be less useful to the court: some decisions with lower estimated degrees of error would actually be more in error than decisions with higher estimated error.

that the cases with the most unstable statistical precedent are in fact promising opportunities for law development.[113]

FIGURE 6: ADJUSTED STATISTICAL PRECEDENT AND REVERSAL (2011–2012 TEST SET)[114]



B. Testing Statistical Precedent

To investigate whether statistical precedent accurately captures a court's collective wisdom, I test five hypotheses regarding the relationship between statistical precedent and traditional indicators of error and law development:[115]

113. Researchers are increasingly suspicious of traditional measures of discrimination, such as Area Under the Curve ("AUC"), for assessing the ultimate value of a predictive model. While useful in assessing the performance of one model against another, they do little to shed light on a model's value in real-world applications. *See, e.g.*, Jon Kleinberg et al., *Human Decisions and Machine Predictions*, 293 Q.J. ECON 237, 253 (2018) ("Measures such as [AUC], though, do not tell us whether the algorithm's predictions can improve on decision quality."). Nonetheless, some readers may be interested to know that the model of statistical precedent has an AUC of approximately 0.70. Note, though, that measures of discrimination are particularly uninformative in this application. The target of the prediction exercise—the degree of error—is not a dichotomous variable, although we can only assess it by reference to dichotomous outcomes. Thus, even a perfectly accurate model would have an AUC of less than 1 insofar as panels are inconsistent.

114. Figure 6 is a cross-validated, locally estimated scatterplot smoothing regression with 95% confidence intervals.

115. I note one limitation that applies to all except the hypothesis regarding opinion publication. As a practical matter, the hypotheses can be tested on only published opinions. Once

Hypothesis 1: Circuit court decisions with higher estimated degrees of error should be more frequently accompanied by dissents. In other words, as estimates indicate that more panels would disagree with the assigned panel's decision, a judge should be more likely to dissent.

Hypothesis 2: Circuit court decisions with higher estimated degrees of error should be more likely to have negative subsequent appellate history as measured by Shepard's citation services.

Hypothesis 3: Circuit court decisions with higher estimated degrees of error should be more likely to have negative analysis in future court opinions as measured by Shepard's citation services.

Hypothesis 4: Circuit court decisions in cases with higher estimated degrees of instability should be cited more frequently.[116]

Hypothesis 5: Circuit court decisions in cases with higher estimated degrees of instability should be published more frequently. We should expect that judges are already more likely to publish an opinion when the governing law is most underdeveloped. Thus, if unstable statistical precedent tracks underdeveloped law, judges should more frequently publish opinions when statistical precedent is unstable.

Table 3 displays the results of five regressions. The first three regressions test the relationship between the estimated degree of error and (1) dissent, (2) subsequent negative appellate history of the same

opinions are designated as unpublished, they are effectively ignored by courts, so the traditional indicators of error and law development do not show up for them. In fact, before 2007, Ninth Circuit appellate rules forbade lawyers from even citing unpublished opinions. *See* Sarah E. Ricks, *A Modest Proposal for Regulating Unpublished, Non-Precedential Federal Appellate Opinions While Courts and Litigants Adapt to Federal Rule of Appellate Procedure 32.1*, 9 J. APP. PRAC. & PROCESS 17, 20 (2007) (noting the prohibition on federal appellate courts restricting the citation of nonprecedential decisions after January 1, 2007 under Federal Rule of Appellate Procedure 32.1). While lawyers are now permitted to cite unpublished opinions, it is rare for other opinions to cite them, judges to dissent from them, or courts to review them en banc. The restriction of the analysis to published opinions poses a concern because they are not drawn from a random sample of cases.

116. I use Google Scholar citation counts. Note that there may be an ambiguous relationship between the value of precedent and citations. On the one hand, precedent in an unstable area of law may yield more citations as judges lean on it to guide future decisions in a still unstable (though perhaps more stable) area of law. On the other hand, precedent may decrease litigation of the issues it addresses by providing clarity to potential litigants, and precedent in unstable areas of law may be more likely to be superseded by newer precedent.

case,[117] and (3) negative analysis in future opinions.[118] The fourth and fifth regressions test the relationship between the estimated degree of instability and (4) citation-percentile ranking[119] and (5) opinion publication. As hypothesized, increases in error estimates are strongly associated with dissents, subsequent negative appellate history, and negative analysis in future opinions; increases in instability estimates are strongly associated with citations and opinion publication.

TABLE 3: REGRESSING TRADITIONAL INDICATORS ON ERROR ESTIMATES (2011–2012 TEST SET)

Degree	*Dissent*	*Negative Appellate History*	*Negative Analysis Percentile*	*Citation Percentile*	*Publication*
Error	0.3%**	0.4%**	0.3%**		
	(0.1%)	(0.1%)	(0.1%)		
Instability				0.6%**	1.34%***
				(0.2%)	(0.1%)

*0.05, **0.01, ***0.001 statistical significance. Robust standard errors in parentheses.

The results are strong evidence that statistical precedent can locate erroneous decisions and opportunities for developing law, allowing the court to revitalize the administration of justice. For example, consider dissents. For each additional estimated degree of error, the dissent rate increases by 0.3%. For the decisions with a degree of error above 70%, there is a remarkably high dissent rate of 29%. What might this mean for the court? Dissents can serve as a signal to the court that the decision should be considered for en banc review,[120] but they are an unreliable signal: all three of the panel members could have views that depart from the court's collective

117. Because subsequent negative appellate history could cause sharp reductions in citations, I do not control for the number of citations when testing the relationship between error and subsequent negative appellate history. Regardless, controlling for citations does not substantially affect the estimates.

118. As indicated by Shepard's signals. I exclude cases with subsequent negative appellate history so as not to allow appellate history to drive results in both regressions.

119. I use citation percentiles rather than pure citation counts due to the fact that citation counts are so widely distributed. Citation counts roughly follow a power-law distribution. *See* David G. Post & Michael B. Eisen, *How Long is the Coastline of the Law? Thoughts on the Fractal Nature of Legal Systems*, 29 J. LEGAL STUD. 545, 570 (2000). Using the log of citations rather than percentile rankings does not substantially change the results.

120. Deborah Beim et al., *Signaling and Counter-Signaling in the Judicial Hierarchy: An Empirical Analysis of En Banc Review*, 60 AM. J. POL. SCI. 490, 490 (2016). In my dataset, only three of the 415 published decisions without a dissent were reviewed en banc, while twelve of the eighty-three decisions with a dissent were reviewed en banc.

judgment, or one of the panel judges could disagree with the other panel members but be too pressed for time to author a dissenting opinion. In such cases, there will be no dissent to call the outlier decision to the court's attention. But why should a panel decision evade the broader court's scrutiny simply because all of its members happen to be ideologically aligned or because one member happens to be too busy to bother with a dissent? There was once a good answer: we do not know how to do any better. Statistical precedent changes that. Courts now have access to technology that can locate the presumptive injustices that deserve their attention, regardless of whether a dissent happens to have accompanied that injustice. And this is just one example of many. In the next Part, I discuss how courts could obtain, implement, and monitor a high-powered system of statistical precedent that could broadly improve the administration of justice.

V. ADOPTING STATISTICAL PRECEDENT

This Part addresses some of the subtler choices and challenges involved in adopting a system of statistical precedent. First, I explain how a court could select a high-quality model of its statistical precedent. I then propose four simple—and, I think, uncontroversial—reforms that could help introduce courts to the uses of statistical precedent. Finally, I address three commonly expressed concerns about the use of algorithms in the justice system: litigant gaming, embedded biases, and malfunctioning algorithms.

A. Selecting the Model of Statistical Precedent

The model of the Ninth Circuit's statistical precedent presented above was meant only to show that it can successfully locate errors and opportunities to develop the law. I am not suggesting that the Ninth Circuit begin using my model. It is undoubtedly possible to create models that are significantly more accurate in estimating degrees of error and instability. More data, more variables, better algorithms, extra weight to more recent years, less weight to the decisions of judges no longer on the court, incorporation of data from other circuits—there are many ways to improve and tailor statistical precedent so as to generate more accurate estimates of error and instability.

But if I am not offering a model, which model should the court use? A major issue is neutrality: modelers may, intentionally or not, embed their own preferences within their models. For example, a

modeler could choose to implement a regression model to generate predictions, thus having to manually select which variables to include, how to interact them, and what functional form to give any continuous variables. Even a modeler that is aiming to create the most predictive model might find herself unintentionally favoring models that reflect her individual conception of error rather than the court's conception.

A machine learning approach provides a considerable safeguard. For example, in building a model of the Ninth Circuit's statistical precedent, I used an assortment of algorithms, each of which include an automated process for selecting the most predictive variables and making them interact. I then let the data decide which of the models was most predictive, using the process of cross validation.[121] Nonetheless, if I had disliked the selected model (e.g., perhaps the model generated high estimates of error for decisions that I personally believed were correctly decided), I could have simply removed the model or added additional models in the hope that the process of cross validation would select a new model that, though perhaps less accurate, would better match my ideological preferences. Though it is much more difficult to ideologically tailor a model when using machine learning methods, it is possible.

Both neutrality and accuracy could best be assured by decentralizing the construction of models and selecting the model that performs best according to a prespecified, publicly communicated, and standardized criterion. Fortunately, the framework for such a process is already in place. Corporate and government institutions alike can now access high-quality predictive models that are tailored to their organization's specific objectives by sponsoring open competitions on the Kaggle website. Recently acquired by Google, Kaggle has run competitions for hundreds of organizations, including Microsoft, the National Football League, Expedia, and Home Depot.[122] Government organizations have also jumped in. For example, the U.S. Transportation Security Administration recently offered a $500,000 first-place prize for the creation of an algorithm to predict potential threats in airport security screenings.[123] The U.S. Courts of Appeals

121. For an accessible introduction to cross validation, see Jason Brownlee, *A Gentle Introduction to K-Fold Cross-Validation*, MACHINE LEARNING MASTERY (Aug. 8, 2019), https://machinelearningmastery.com/k-fold-cross-validation/ [https://perma.cc/3NFA-Y9QL].

122. For current competitions, see *Competitions*, KAGGLE, https://www.kaggle.com/competitions (last visited Apr. 5, 2020) [https://perma.cc/ZP4P-64E2].

123. *Passenger Screening Algorithm Challenge*, KAGGLE, https://www.kaggle.com/c/passenger-screening-algorithm-challenge (last visited Apr. 5, 2020) [https://perma.cc/6ZBA-GQSW]:

> As part of their Apex Screening at Speed Program, DHS has identified high false alarm rates as creating significant bottlenecks at the airport checkpoints. Whenever

should pursue a similar strategy to select a model of statistical precedent.[124]

I recommend that the courts hold annual competitions to predict the upcoming year's decisions. The court could then select the most predictive model for use in allocating judicial attention in the subsequent year. The process would ensure the statistical precedent remains current, allow modelers to build on the strengths of previous models, and ensure public transparency.

B. Proposals for Reform

Statistical precedent may one day allow radical changes to courts' operating procedures. For example, we could imagine a more finely tiered and gradual system of appellate review. There is little reason that all cases should be decided by a panel of three judges. The easy cases with extremely high or low error could, as an initial matter, be assigned to a single judge. If that judge's decision were made as expected, it could serve as the final decision. If the decision were unexpected (e.g., the judge affirmed a case with a high degree of error or reversed a case with a low degree of error), the case could be expedited for review by an additional judge. For moderately difficult cases with greater instability, the court might assign the traditional three-judge panel. And for the hard cases with instability estimates

> TSA's sensors and algorithms predict a potential threat, TSA staff needs to engage in a secondary, manual screening process that slows everything down. And as the number of travelers increase every year and new threats develop, their prediction algorithms need to continually improve to meet the increased demand.
>
> Currently, TSA purchases updated algorithms exclusively from the manufacturers of the scanning equipment used. These algorithms are proprietary, expensive, and often released in long cycles. In this competition, TSA is stepping outside their established procurement process and is challenging the broader data science community to help improve the accuracy of their threat prediction algorithms. Using a dataset of images collected on the latest generation of scanners, participants are challenged to identify the presence of simulated threats under a variety of object types, clothing types, and body types. Even a modest decrease in false alarms will help TSA significantly improve the passenger experience while maintaining high levels of security.

124. A frequent choice in Kaggle competitions that focus on datasets with dichotomous outcomes is the AUC, and it would be a strong option as a criterion for a court's model selection. One of the core advantages of AUC over other common metrics (e.g., the correct classification rate or F_1 score) is that it does not depend on a choice of threshold. This is particularly important in the context of statistical precedent because the ultimate goal is not to partition cases—the goal is to estimate the degree of error, which is not actually a zero or one. For clarity, consider the possibility that there are in fact no decisions that a majority of panels would reverse. We would still wish to know which cases have a higher degree of error. But with a measure like the correct classification rate, a perfectly accurate model would perform no better than a useless model that simply estimated a 0% degree of error for every decision.

close to 50%, courts could assign larger panels, effectively providing a mechanism for courts to resolve their internal inconsistencies and clarify law. But any such radical restructuring should be postponed until we have a better understanding of how statistical precedent operates. At least initially, it is probably the case that the information should simply be made available to judges, letting them explore the varied uses for the estimates. Below, I offer four moderate reforms that judges should consider implementing in the near future.

1. Flag Panel Decisions that Depart Widely from Statistical Precedent

There are, especially in large courts, simply too many opinions for courts to meaningfully review their own opinions. The Ninth Circuit, for example, produces on average two new published opinions per work day.[125] But "[t]he full court must, in order to prevent different panels from deciding cases inconsistently and thus greatly reducing the certainty of legal obligation, maintain a credible threat to rehear a case en banc if the panel deviates from the law of the circuit."[126]

Simply notifying all judges of how far each opinion deviates from statistical precedent would allow judges to at least meaningfully review the set of decisions that are most incompatible with their court's jurisprudence. Panels, unable to hide in the mass of opinions, would have more reason to try to decide cases in accordance with the court's collective conception of justice. Of course, many panels are undoubtedly trying to fit their decisions into the broader law and simply failing in that effort.[127] Thus, even in the absence of a credible threat to have the case reviewed en banc, such judges would be happy to receive feedback from other judges before their opinions are finalized.[128]

125. *See*, *e.g.*, Kleinfeld Statement, *supra* note 17, at 5:

> If we ignore the unpublished decisions (as most of us are forced to do, allowing for much error to go uncorrected in them), there were still 557 published dispositions, each with precedential force. Keeping up would require us to read around three per day, manageable if one is not on calendar, but generating a pile of about 15 plus the new ones that come in on Monday after a week on calendar. At that point, the opinions can only be glanced at to see if they affect pending cases or resolve matters in which the judge happens to have a particularly strong interest.

126. RICHARD A. POSNER, THE FEDERAL COURTS: CHALLENGE AND REFORM 133 (1996).

127. *See, e.g.*, O'Scannlain Statement, *supra* note 17, at 10 ("[I]t seems increasingly common for three judge panels to make sua sponte en banc requests for review of their own decisions, because they uncover directly conflicting Ninth Circuit precedent on a dispositive issue.").

128. *See, e.g.*, Berzon, *supra* note 85, at 1490 n.49 ("[W]e are quite receptive to altering opinions after publication based on feedback from our colleagues.").

2. Add Flagged Unpublished Decisions to a Public High-Risk List

Unpublished opinions are so numerous that they would probably overwhelm a simple flagging system—judges would still have too little attention to meaningfully review each other's problematic, nonprecedential decisions. Something more is needed if courts are to meaningfully attend to unpublished opinions.

And there are good reasons to believe that attention is needed. Justices Blackmun, Stevens, and Thomas have each independently charged circuit courts with abusing nonpublication,[129] and Judge Patricia Wald, former chief judge of the D.C. Circuit, wrote that nonpublication allows for "deviousness and abuse."[130] Empirical research lends support to those claims.[131] An Eighth Circuit panel went so far as to declare unpublished opinions unconstitutional.[132] Judge Wald succinctly summarizes the vast body of literature criticizing unpublished opinions:

> [I]t is argued that unpublished opinions: result in less carefully prepared or soundly reasoned opinions; reduce judicial accountability; increase the risk of nonuniformity; allow difficult issues to be swept under the carpet; and result in a body of "secret law" practically inaccessible to many lawyers. Furthermore, there is no uniformly enforced or practiced guidelines for making the publication decision; hence judges exercise considerable discretion in deciding when an opinion should be published, *i.e.*, when an opinion will become law.[133]

129. Adam Liptak, *Courts Write Decisions that Elude Long View*, N.Y. TIMES (Feb. 2, 2015), https://www.nytimes.com/2015/02/03/us/justice-clarence-thomas-court-decisions-that-set-no-precedent.html [https://perma.cc/3T9V-Y66M].

130. Patricia M. Wald, *The Rhetoric of Results and the Results of Rhetoric: Judicial Writings*, 62 U. CHI. L. REV. 1371, 1374 (1995).

131. *See* David S. Law, *Strategic Judicial Lawmaking: Ideology, Publication, and Asylum Law in the Ninth Circuit*, 73 U. CIN. L. REV. 817, 820 (2004) (finding "that there exists, for some judges, a significant relationship between how the judge votes on the merits of the case, and whether the case is published").

132. Anastasoff v. United States, 223 F.3d 898, 900 (8th Cir. 2000), *vacated as moot on reh'g en banc*, 235 F.3d 1054, 1056 (8th Cir. 2000).

133. Nat'l Classification Comm. v. United States, 765 F.2d 164, 173 n.2 (D.C. Cir. 1985) (separate statement of Wald, J.) (citing Wald, *supra* note 101, at 781–84). Boyce F. Martin, the former chief judge of the Sixth Circuit, has provided a similar list of criticisms. Martin, *supra* note 52, at 180:

- loss of precedent, that unpublished opinions are, in fact precedent but cannot be used as such;
- sloppy decisions, that judges are careless when they know they are writing an unpublished opinion;
- lack of uniformity, that panels cannot follow other panels when they are unaware of other panels' unpublished opinions;
- difficulty of higher court review, that the Supreme Court is far less likely to review an unpublished opinion than it is to review a published opinion;
- unfairness to litigants, that litigants deserve published opinions;

Despite the daunting size of the literature, I am aware of no effort to address the problems posed by unpublished opinions—except for arguments that the courts stop using them. But there is absolutely no sign that courts will reverse course on the use of unpublished opinions. The publication rate has continuously decreased, and there is no indication that judges are willing to relinquish the convenience of issuing unpublished opinions.[134]

The courts could implement a "high-risk list" for unpublished opinions, modeled after the "six-month list" that Congress instituted to reduce delays in the federal district courts. By law, the number of every judge's motions that have been pending for more than six months are made public.[135] While there is debate about whether the six-month list has been successful, its soft law approach is a promising way to balance the need for a flexible judiciary with accountability for unelected, life-tenured judges.[136] To rein in the inappropriate use of unpublished decisions, courts could publish lists of unpublished decisions that dispose of cases in ways that their colleagues would most likely disagree with. While such a high-risk list does not address all commentator concerns, it could help increase judicial accountability, decrease the risk of nonuniformity, and make it harder to sweep difficult issues "under the carpet." Maintaining a public list could shame judges in order to limit the abuse of unpublished opinions—they might be more attentive to cases, less likely to try and hide an outcome that is unsupported by the law, less likely to avoid difficult legal issues,[137] and less likely to agree to withdraw a dissent in exchange for nonpublication.[138]

Of course, there could be occasions when an unpublished opinion is justified (e.g., statistical precedent is inaccurate), and we

- less judicial accountability, that the unpublished opinion, particularly the per curiam, allows the judge to hide outside the public glare;
- less predictability, that any opinion provides a roadmap of the law and a sense of the direction in which the law is developing.

134. *See, e.g.*, Patrick Schiltz, *Much Ado About Nothing: Explaining the* Sturm Und Drang *over the Citation of Unpublished Opinions*, 62 WASH. & LEE L. REV. 1429, 1484 (2005) ("Judges have told me that being forced to treat their unpublished opinions as binding precedent would create chaos and that it would take decades to repair the damage.").

135. 28 U.S.C. § 476(a)(1) (2012).

136. *See* Miguel de Figueiredo et al., Against Judicial Accountability: Evidence from the Six Month List 6 (Feb. 20, 2018) (unpublished manuscript), https://www.ssrn.com/abstract=2989777 [https://perma.cc/Z2RZ-2T8Z].

137. *See* Wald, *supra* note 130, at 1374 ("I have seen judges purposely compromise on an unpublished decision incorporating an agreed-upon result in order to avoid a time-consuming public debate about what law controls.").

138. *See id.* ("I have even seen wily would-be dissenters go along with a result they do not like so long as it is not elevated to a precedent.").

would and should not expect the list to remain empty. But unusually frequent use of unpublished opinions with high error estimates should be disconcerting. To illustrate what the list might look like, Table 4 displays a 2010–2011 high-risk list for the Ninth Circuit, showing the number of unpublished decisions issued by each judge that deviated from statistical precedent by more than 70%. Judges are not uniformly using unpublished opinions. Judges William Fletcher, Ronald Gould, and Sidney Thomas lead the list; each were members of panels that issued more than ten high-risk unpublished opinions. Other active judges—who review the same cases on average—have much lower counts. It would be too much to claim that the former judges are strategically abusing nonpublication, but might they be underestimating just how "easy" their decisions are?[139]

139. *See* Daniel Kahneman et al., *Noise: How to Overcome the High, Hidden Cost of Inconsistent Decision Making*, HARV. BUS. REV., Oct. 2016, at 38, 43 ("Experienced professionals tend to have high confidence in the accuracy of their own judgments, and they also have high regard for their colleagues' intelligence. This combination inevitably leads to an overestimation of agreement.").

TABLE 4: HIGH-RISK UNPUBLISHED OPINIONS
(NINTH CIRCUIT 2010–2011)

William Fletcher	14	William Canby*	6
Ronald Gould	12	Procter Hug*	5
Sidney Thomas	11	John Noonan*	5
Richard Paez	10	Susan Graber	5
Atsushi Tashima*	9	Raymond Fisher	5
Richard Clifton	9	Joseph Farris*	5
Norman Smith	9	Michael Hawkins*	4
Betty Fletcher*	9	Ferdinand Fernandez*	4
Harry Pregerson	8	Edward Leavy*	3
Kim Wardlaw	8	John Wallace*	3
Johnnie Rawlinson	8	Carlos Bea	3
Jay Bybee	8	Andrew Kleinfeld*	2
Diarmuid O'Scannlain	7	Alex Kozinski	2
Mary Schroeder*	7	Arthur Alarcon*	2
Stephen Trott*	7	Marsha Berzon	2
Barry Silverman	7	Richard Tallman	2
Sandra Ikuta	7	Consuelo Callahan	2
Alfred Goodwin*	7	Robert Beezer*	1
Stephen Reinhardt	6	Dorothy Nelson*	1
M. McKeown	6	Thomas Reavley*^	1
Milan Smith	6	Pamela Rymer	1

*Senior Status, ^Visiting Judge

At the very least, both the court and the public should be aware of the wide variation in the use of unpublished opinions. The list would draw attention to a problem that is otherwise all too easy to ignore—inconsistency. If judges can differ so dramatically in their use of unpublished decisions, it is an indication that the criteria for publication are too vague to generate shared practices. Those criteria are the subject of the next proposal.

3. Include Unstable Statistical Precedent as a Criterion for Publication

With an overall publication rate of less than 10%,[140] it is critical that courts publish opinions where the law is most in need of development. The existing publication criteria do little to aid judges in making that determination. The criteria across circuits "amount to little more than saying that an opinion should not be published unless it is likely to have value as precedent."[141] Not only are they essentially tautological, but they are also wildly overbroad—almost every decision has some precedential value, but the courts simply cannot publish an opinion in every case. The criteria are flexible in an additional way: they say more about the actual opinion written than they do the underlying issues. The main criterion for publication is if an *opinion* "establishes, alters, modifies, clarifies, or explains a rule of law."[142] One judge might author an opinion describing the result as some mechanical application of existing law, while another might justify the result as some modification, alteration, or establishment of law.

There is no particular reason to believe that judges are good at identifying the cases that are best for creating precedential value, and they likely miss many opportunities for valuable law development.[143] The fact that judges display such wide variation in publication practices means that at least some of them are passing on the best opportunities.[144] As I have shown above, statistical precedent can locate those opportunities. By including unstable statistical precedent as a criterion, courts can help judges identify the cases that are most undetermined by existing law.

4. Use Statistical Precedent to Assign Cases to Staff Attorneys

The courts' treatment of pro se appeals is a high-profile issue at the moment. Judge Richard Posner's retirement, driven in part by what he regarded as his court's neglect of pro se appellants, and his

140. *See supra* note 50 and accompanying text.

141. POSNER, *supra* note 126, at 165.

142. 4TH CIR. R. 36(a); *see* 9TH CIR. R. 36-2(a) (designating a written disposition as an opinion if it "[e]stablishes, alters, modifies or clarifies a rule of federal law").

143. *See generally* Donald R. Songer et al., *Nonpublication in the Eleventh Circuit: An Empirical Analysis*, 16 FLA. ST. U. L. REV. 963, 984 (1989) (concluding that the criteria for publication provide little guidance and are inconsistently applied).

144. My own analysis of Ninth Circuit opinions shows that different panels can disagree over the decision to publish in more than 30% of civil cases.

recent book have vaulted the issue into the public spotlight.[145] Courts routinely assign pro se appeals to staff attorneys for an initial decision—decisions that judges are supposed to review before signing off on them.[146] While judges are obviously not bound by a staff attorney's initial decision, Posner claims that judges tend to "rubber stamp" them,[147] and the overwhelming recommendation by staff attorneys in the Seventh Circuit is to affirm (83%).[148] Courts presumably assign pro se appeals to staff attorneys because they collectively have a low degree of error—most panels would not assign error to the lower court. In the resource allocation framework, there is thus little need to dedicate judicial attention to those cases. But do all pro se appeals have low degrees of error? Should perhaps some of them be assigned to chambers, provided legal representation, and tracked for oral argument? While overworked judges and staff attorneys may be able to identify some of those cases, error estimates from statistical precedent could likely aid that effort.

Figure 7 shows the distribution of error estimates for pro se appeals and for all other appeals in the Ninth Circuit. According to these estimates, while "pro se" is not an awful proxy for low merit, it is far from perfect. Approximately 10% of civil pro se appeals have error estimates that place them in the range of other civil cases. Thirty-eight percent of pro se appeals with error estimates higher than 20% were reversed, but what percentage of the affirmances would have switched to reversals had the court allocated more judicial attention to them? We do not know, but I see little justification for assigning those cases to staff attorneys along with all of the other pro se appeals.[149] Furthermore, as I discuss in Section V.D below, the error estimates for pro se cases may be understated relative to other cases if statistical precedent reflects the court's historical deprivation of attention to them. It is thus particularly important that courts at least provide more attention to those cases that depart widely from statistical precedent.

145. *See* RICHARD A. POSNER, REFORMING THE FEDERAL JUDICIARY: MY FORMER COURT NEEDS TO OVERHAUL ITS STAFF ATTORNEY PROGRAM AND BEGIN TELEVISING ITS ORAL ARGUMENTS 135–44 (2017) (describing most judges and staff attorneys as unsympathetic towards pro se litigants).

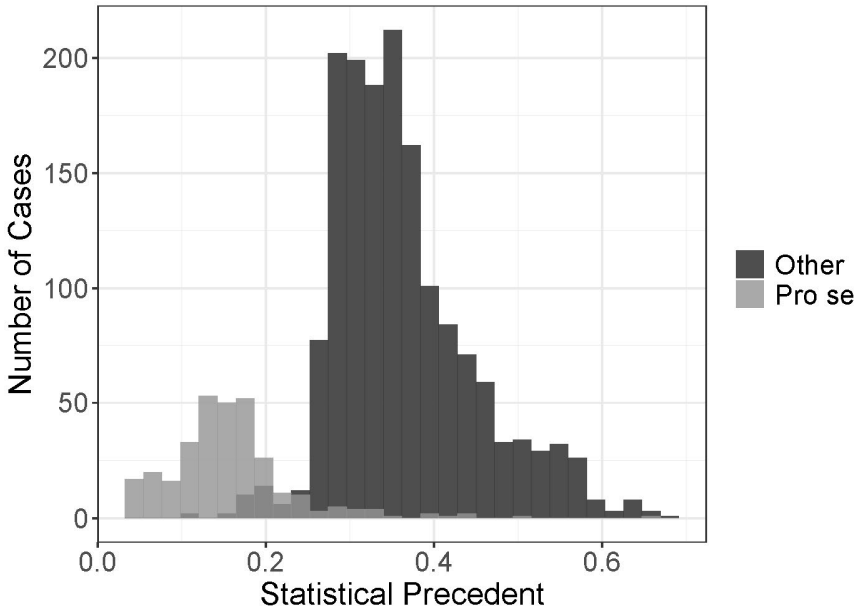
146. *See generally* Levy, *supra* note 23, at 380–81 (detailing the docketing practices of five circuit courts).

147. POSNER, *supra* note 145, at 6.

148. *Id.* at 9.

149. Note that I do not know whether some of these pro se appeals were ultimately assigned to judicial chambers. But this points to another problem with the current triage system—it is not a transparent system.

FIGURE 7: STATISTICAL PRECEDENT FOR PRO SE APPEALS



C. Some Concerns

I acknowledge that the combination of artificial intelligence and law makes many people uneasy. But data analytics is changing industries, and the courts risk committing widespread injustice by continuing to abstain. Finance, banking, sales, medicine, elections, and sports have all been deeply impacted by the adoption of predictive technology and improvements in this technology. Our justice systems have not kept pace, and it should be alarming that sports franchises are so much more advanced. What happens in our courts matters, and we should be doing better. Why aren't we?

There are three misconceptions that I think help explain why predictive technology has been so slow to catch on in courts. First, in many of our justice systems, there is no obvious way to measure a case's merit without simply deciding it: we use courts as our scales of justice. Parts of the criminal justice system stand out as exceptions: recidivism is a critical and measurable outcome, predictions of which can proxy for the merits of an individual's case and inform decisionmaking.[150] This likely explains why some parts of the criminal

150. *See, e.g.*, Kleinberg et al., *supra* note 113, at 243 ("Recidivism, which is one relevant input to sentencing someone who has been found guilty, can be predicted.").

justice system have been uniquely accepting of analytics. In contrast, there is seemingly nothing we can predict that would inform, for example, a panel's decision in an employment discrimination or free speech case. But to the extent that we trust the judgment of our judges and value consistency in decisionmaking, *their judgments* are a valuable target of prediction. Much like historical recidivism patterns can help parole boards find individuals that should be paroled, historical reversal patterns can help courts find—and pay attention to—decisions that should be reversed or need legal clarification.

The second misconception is that an algorithm must include and properly process legally relevant variables to be useful. Most of us believe that law is a critically important factor in judicial decisionmaking. Thus, the thinking goes, an algorithm that cannot process legally relevant variables in a legally relevant manner cannot accurately assess legal merit. And if it can, is that not evidence of radical legal realism and a threat to our belief in the rule of law? Those concerns are understandable but ultimately misplaced. Even where law is the dominant factor in decisions, an algorithm only needs access to variables that are *statistically associated* with law to generate accurate predictions. And the mechanisms by which cases are selected into the circuit courts make such correlations readily plausible.[151] For example, the beliefs of attorneys and district court judges are presumably important determinants of which cases are ultimately appealed. If those judgments are responsive to legal merits, then variables for attorneys and district judges (or variables that correlate with those variables) can be important predictors of legal merit. Thus, while legally relevant variables could very well increase the accuracy of predictions, they are not a prerequisite. In fact, legally relevant variables may be particularly unimportant predictors where law matters most: if the development of precedent matters, then the legal relevance of legally relevant variables will shift over time, making them poor predictors of future decisionmaking. Moreover, there is robust empirical evidence that models of judicial decisionmaking can be accurate without access to law. For example, statistical predictions based on a simple set of six general case

151. *See* George L. Priest & Benjamin Klein, *The Selection of Disputes for Litigation*, 13 J. LEGAL STUD. 1, 4 (1984) (presenting a model of the litigation process driven by economic factors, such as "the expected costs to parties of favorable or adverse decisions, the information that parties possess about the likelihood of success at trial, and the direct costs of litigation and settlement").

characteristics outperformed legal specialists in predicting Supreme Court decisions.[152]

Third, the combination of algorithms and justice inspires dystopian visions of machines deciding our fates, violating our notions of due process in the name of a more efficient system. But as demonstrated in this article, algorithmically aided justice need not be about coarse efficiency, and it need not raise due process issues. Statistical precedent is not about saving money or time—it is about ensuring just decisions and producing precedent that can help society navigate the complexities of law. Nor is statistical precedent about machines deciding our fates. It is not even about machines *recommending* outcomes to judges. It is about guiding the much less deliberate and less informed choice of where to even focus attention—a choice that is generally not itself the product of very focused attention.

In fact, as I argue in this Section, statistical precedent—as a tool to guide attention rather than recommend or automate merit decisions—largely evades the standard objections to algorithm-assisted decisionmaking.

1. Litigant Gaming

Algorithmic decisionmaking can suffer from Campbell's Law: "The more any quantitative social indicator is used for social decision-making, the more subject it will be to corruption pressures and the more apt it will be to distort and corrupt the social processes it is intended to monitor."[153] Distortion pressures are one of the core reasons that this Article does not propose that error estimates be used to recommend, much less automate, the ultimate outcome of an appeal. For example, imagine that the selected algorithm used information about law firms to make predictions, and that appeals by litigants represented by a Vault 100 law firm tended to have high

152. Theodore W. Ruger et al., *The Supreme Court Forecasting Project: Legal and Political Science Approaches to Predicting Supreme Court Decisionmaking*, 104 COLUM. L. REV. 1150, 1151–52 (2004):

> In advance of the oral argument date, we obtained predicted outcomes using two methods—one a statistical model that forecasts outcomes based on six general case characteristics, and the other a set of independent predictions from a large group of legal specialists, each making particularized assessments of one or more cases. . . . [T]he machine did significantly better at predicting outcomes than did the experts. While the experts correctly forecast outcomes in 59.1% of cases, the machine got a full 75% right.

153. Donald T. Campbell, *Assessing the Impact of Planned Social Change*, 2 EVALUATION & PROGRAM PLAN. 67, 85 (1979) (emphasis removed).

estimated degrees of error. While it is possible that such representation is causally responsible for the higher degrees of error, it could also be the case that it is a mere correlation. If the latter, and circuit courts were deciding actual case outcomes by reference to error estimates, litigants might alter their behavior and hire Vault 100 firms to artificially boost their chances of winning.

Moving from the substantive outcome to the distribution of attention greatly diminishes the risk of litigant gaming. When a weak (strong) appeal has an artificially inflated (deflated) degree of error, there is a worry that using an algorithm to help decide the outcome will cause the court to incorrectly reverse (affirm) the appeal. Efforts to inflate or deflate error estimates in order to manipulate judicial attention are likely to be less attractive. The battle for judicial attention is simply less consequential: unless the courts' shortage of judicial attention is much worse than we think, an appeal still has a reasonable probability of being decided correctly even with limited judicial attention.

Regardless of the payoff from successful gaming, it would generally be a costly, complex, and unpredictable endeavor. Many of the variables included in predictive models would likely be drawn from district court and agency litigation. And given the expense and the importance of prevailing at initial stages, litigants would hesitate to sacrifice optimal litigation strategies merely in anticipation of avoiding judicial attention at the circuit court should they happen to mistakenly prevail. Even understanding how to manipulate an algorithm would be difficult for litigants: the so called "black box" of machine learning, often derided for its opaqueness,[154] provides a safeguard against manipulation. Those parties sophisticated enough to successfully manipulate statistical precedent are likely to receive extensive attention regardless. Finally, attempts at manipulation may be undone (or even backfire) as the litigation process continues to unfold. Because variables frequently interact in machine learning models, attempts to manipulate them would often have to account for future events.[155]

154. *See, e.g.*, Cynthia Rudin, *Algorithms and Justice: Scrapping the 'Black Box,'* CRIME REP. (Jan. 26, 2018), https://thecrimereport.org/2018/01/26/algorithms-and-justice-scrapping-the-black-box/ [https://perma.cc/FXL4-PU8Y] (criticizing proprietary "black box" tools for their potential to produce flawed calculations, which if unnoticed, could become the basis of a court decision).

155. For example, suppose that a court's algorithm uses a variable that can be manipulated at a relatively low cost, such as the prevalence of citations in each party's briefs. The ultimate effect of one litigant's citation count on the estimates may depend on the other litigant's citation count.

In summary, the minimal payoff for manipulating judicial attention, coupled with its cost, complexity, and uncertainty, is likely to make litigant gaming rare and inconsequential. Nonetheless, it is admittedly difficult to predict the extent of litigant gaming a priori. The opportunities for gaming ultimately depend on the models selected by a court, the ease with which impactful variables can be manipulated, the ability of litigants to manipulate those variables covertly so as to not draw unwanted attention, and how judges and courts choose to implement the models. It may even be possible, though I think unlikely, that individual judges' assessments of error would be influenced by exposure to an error estimate. It is an issue that courts and scholars would need to assess upon selection of models and continue to monitor as the models are implemented.

If litigant gaming does indeed pose a problem, the courts could employ a number of strategies to combat it. Most simply, the courts could limit the variables that modeling teams are permitted to use to those that are costly to manipulate (e.g., variables derived from district court and agency litigation). While this would come at the cost of predictive accuracy, the protection against gaming could warrant it. A more sophisticated approach could involve selecting not just the best performing model, but the subset of models that meet some specified level of predictive performance.[156] Insofar as the models produce similarly accurate predictions through a diverse set of variables and mechanisms, averaging the predictions from the subset of models would reduce the opportunities for gaming while minimally impacting predictive power.

2. Status Quo Bias

Because statistical precedent relies on datasets of historical decisions, there may be a concern that it would tie the courts to outdated conceptions of error. While such a status quo bias might be a concern if algorithms were being used to guide merit decisions, it is difficult to imagine how targeting judicial attention could substantially impair a court's ability to develop new conceptions of error. Furthermore, if models are selected in accordance with the procedure I outline above, older decisions would only be used insofar as they help predict decisions in the year before a model is

156. For more discussion of this point, see Hannah Laqueur & Ryan Copus, Synthetic Crowdsourcing: A Machine-Learning Approach to Inconsistency in Adjudication 26 (Dec. 6, 2017), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2694326 [https://perma.cc/GSZ8-K6R9].

implemented, which should alleviate concerns that the models are tying courts to outdated conceptions of error.

More worrisome is the possibility that statistical precedent helps to cement historical judicial failures to identify decisions that both past and present judges would—if they paid more attention—agree are in error. The degrees of error and instability are defined by reference to hypothetical panel decisions that are made after *close* evaluation. But some actual panel decisions are likely inattentive mistakes—had the panel more carefully considered the case, it would have made a different decision.[157] For example, some decisions may be a result of the fact that a panel gave only cursory review to a staff attorney's or law clerk's draft of an opinion. Thus, any dataset used to build a model will include decisions that are the result of either close evaluation or a panel's inattentive mistake. Moreover, there is generally no way to identify which decisions were a product of a mistake, and these mistakes can have implications for the accuracy of error and instability estimates.

Under certain conditions, these inattentive panel mistakes can result in systematically deflated error estimates for particular sets of cases.[158] Pro se appeals may be in particular danger. If many judges have effectively given up on searching for meritorious pro se appeals,[159] estimates will tend to understate the error of those appeals, helping to continue the deprivation of judicial attention (although note that Section V.B.4 showed that some pro se appeals can have high estimated error).

There are at least three ways to address the possibility that courts have systematically deprived certain cases of attention. First, courts could provide separate treatment for pro se appeals (or other types of cases that may be systematically affirmed due to inattentive mistake) so as to avoid ignoring them when their deflated error estimates are lower than other, less meritorious cases. In brief, courts could compare error estimates for pro se appeals against other pro se appeals, setting a separate, pro-se-specific threshold. One might think of this as an affirmative action program for pro se appellants, meant to rectify historical deficits of attention. Second, as noted above, the court should continuously conduct tests on an algorithm's performance

157. *See infra* Appendix, Part A for a more detailed discussion of inattentive panel mistakes and their effects on error and instability estimates.

158. *See infra* Appendix, Part A.

159. *See, e.g.*, POSNER, *supra* note 145, at 135–36 (noting that most judges were uninterested in pro se appeals).

by providing extensive judicial attention to randomly selected cases. This could help courts identify cases that need extra attention.

The third solution relies on law professors and the bar donating their judgment to the courts. Statistical precedent relies on datasets of decisions, but there is no reason—technical, constitutional, or statutory—that those decisions must be actual judicial decisions. We could rely on other sources of collective wisdom to help build the dataset. For example, the Federal Judicial Center, in conjunction with the circuit courts, could establish blue-ribbon committees of lawyers and professors to carefully review briefs from current cases, conduct legal research, and donate their recommendations to a dataset. Insofar as the decisions conflict with the decisions of central staff, they would provide an immediate alert that a case needs more attention. But these decision donations would also yield benefits in the future. First, the machine learning algorithms used to build statistical precedent perform better with larger datasets, and a blue-ribbon committee could provide that additional data without further taxing judicial resources. Second, the availability of multiple decisions on the same case is a particularly valuable source of statistical information.[160] Third, the committee could focus on correcting for systematic blind spots the courts may have. By focusing on pro se cases, for example, the blue-ribbon committee could effectively embed more attention for pro se appellants into the system.[161]

Concerns about status quo bias also highlight the need to provide a baseline level of attention for all cases. Courts cannot, either under the guidance of statistical precedent or under the guidance of rough preconceptions (e.g., "pro se" as a proxy for "low merit"), safely deprive cases of a minimal level of attention. A baseline level of attention helps prevent major injustices, allows statistical precedent to update with changes in the merits of appeals, and promotes procedural fairness.[162]

160. Using actual judicial decisions to estimate instability is difficult, requiring complex computations. *See infra* Appendix, Part B. Multiple decisions on the same case, even if they are not from judges, provide directly observable evidence of a case's degree of instability.

161. Levy suggests that courts experiment with cases that have historically received low levels of judicial attention by providing them with more judicial attention so that we can estimate the extent of outcome-determinative attention shortages. Levy, *supra* note 29, at 441–42, I support her proposal, but I think it could be usefully supplemented with the work of a blue-ribbon committee.

162. Although I have not focused on procedural justice in this Article, it fits nicely with a system of statistical precedent, as a baseline level of attention is critical for maintaining and improving its accuracy.

3. Malfunction

Like any technology, algorithms can simply malfunction. Perhaps data is stored in some new way such that the algorithm misprocesses it and generates poor predictions. Or maybe there is some major change in the world that causes a widespread disconnect between historical and present statistical associations.[163]

We can and should try to be careful to make sure that algorithms do not malfunction, but we also have to prepare for the possibility that they might, especially if they are deeply embedded into a core institution like the federal judiciary. And the most important thing we can do is make sure that there is a system in place for detecting any serious malfunction. Fortunately, statistical precedent provides immediate and constant feedback: because it is ultimately a prediction of a how a case will be treated once it is provided more attention, the courts would be able to continuously test those predictions.

But it is also theoretically possible that statistical precedent could malfunction in a subset of cases that it is not recommending receive additional attention. Again, a baseline level of attention would alert the court to any major malfunctions with respect to those cases. But courts could provide additional safeguards. They might, for example, randomly provide extensive attention to a small sample of cases to make sure that the results are consistent with the algorithmic predictions. Such randomized checks could also be more targeted, with the court giving greater weight to sets of cases for which the accuracy of statistical precedent is a particular concern (e.g., pro se appeals).[164]

CONCLUSION

Is statistical precedent a politically feasible approach to mitigating the effect of burdensome caseloads and large courts on the quality of appellate justice, or is it a technocratic fantasy? If the experience of other industries is a good indication, the use of algorithms in adjudication will be met with heavy skepticism. The struggle for acceptance is most famously documented in American

163. I have struggled without success to come up with some plausible example. But as unlikely as such a world-changing event might be—at least such that the distribution of judicial attention would remain on the list of things we care about—it still seems worth consideration.

164. This is a simple extension of Levy's proposal that the courts conduct randomized tests of judicial attention. *See* Levy, *supra* note 29, at 441–442 ("[Courts] could randomly select a percentage of cases that normally receive nonargument track treatment and instead give them full judicial treatment . . . . One could then examine the outcomes in those cases and compare them to the outcomes in the rest of the argument cases.").

baseball,[165] but the basic story is transsubstantive and now effectively a trope: decisionmakers overestimate their own judgments and underestimate analytics—until a competitor embraces analytics. If you want to win an election, a championship, a stock market trade, or even your fantasy football league, you embrace analytics. Courts face no such competitive pressure from one another.

The success of analytics in other industries may help mitigate doubts about its value in the courts, but there is a path forward for statistical precedent even if it is met with widespread skepticism. Most importantly, statistical precedent could begin working without Congressional action, centralized adoption by circuits, or even substantial judicial interest—litigant desire for judicial attention could drive the adoption of statistical precedent. If estimates existed, parties might want to include them in their briefs (e.g., to avoid assignment to staff attorneys), petitions for rehearing en banc, and requests for oral argument. If so, legal research services like Westlaw or Lexis—organizations that are already collecting massive amounts of data on litigation in federal courts—might have the incentive to generate estimates of error and instability for their clients. Thus, even if unwilling to lead the way, the courts might be eased into statistical precedent.[166] But courts should not wait. Only they have full access to the data, can implement court-wide procedures for allocating attention, and can ensure public transparency.[167]

Although I focus on the U.S. Courts of Appeals in this Article, statistical precedent could be implemented in a wide variety of adjudication systems that are struggling with burdensome caseloads and untethered judges. One could imagine statistical precedent in state intermediate appellate courts, the Social Security Administration's Office of Disability Adjudication and Review,

165. Baseball's "Moneyball" story is well known because of Michael Lewis's best-selling book, *see* MICHAEL LEWIS, MONEYBALL: THE ART OF WINNING AN UNFAIR GAME (2003), which inspired a movie starring Brad Pitt that was nominated for six Academy Awards.

166. It is also encouraging that the courts already have experience with a rudimentary form of statistical precedent—the Federal Sentencing Guidelines. The sentencing guidelines were generated and adopted with a philosophy similar to the one that statistical precedent rests on—leveraging the collective, historical sentencing patterns to build a model for future sentencing. The ultimate success of the guidelines may have been hindered by flaws in both design and implementation. Perhaps most importantly, the guidelines are almost certainly subject to litigant (prosecutor) gaming. *See, e.g.*, Frank O. Bowman, III, *The Failure of the Federal Sentencing Guidelines: A Structural Analysis*, 105 COLUM. L. REV. 1315, 1336 (2005) (discussing the transfer of discretion from judges to prosecutors). But the very fact of their existence and continued role in the federal judicial system is a promising sign for the political fortunes of statistical precedent.

167. I presume that Westlaw or LexisNexis, as profit-driven companies, would not make their models of statistical precedent available to the public.

immigration courts, parole boards, the Patent and Trademark Office—the list is almost endless. I hope this Article can provide a guide for other adjudication systems as well.

APPENDIX

A. Inattentive Panel Mistakes

The degree of error and instability are defined by reference to hypothetical panel decisions that are made after *close* evaluation. But some actual panel decisions are likely inattentive mistakes—had the panel more carefully considered the case, it would have made a different decision. For example, some decisions may be a result of the fact that a panel gave only cursory review to a staff attorney's or law clerk's draft of an opinion. Thus, any dataset used to build a model will include decisions that are the result of either close evaluation or a panel's inattentive mistake. Moreover, there is generally no way to identify which decisions were a product of a mistake, and these mistakes can have implications for the accuracy of error and instability estimates.

Of course, insofar as mistakes are infrequent, they have no effect on the accuracy of error or instability estimates: if they are rare in the data, they will not substantially affect the estimates of error. But if mistakes are common, their effect on the estimates depends on the type of mistakes that panels make and how they are distributed across cases. For illustrative purposes, let us assume that panels frequently make mistakes.

Consider three types of mistakes that panels could make.[168] First, for some sets of cases, panels might have something close to a constant probability of making a mistake: panels that would reverse a case after careful evaluation might mistakenly affirm with a 20% probability, and panels that would affirm might mistakenly reverse with a 20% probability. Such mistake patterns might be common, for example, in those cases where a panel relies heavily on the recommendations of law clerks. Second, panels might be much more likely to mistakenly affirm some sets of cases than they are to mistakenly reverse them. Decisions on cases assigned to staff attorneys might have such mistake patterns if judges generally provide cursory attention to staff attorney recommendations and would only scrutinize recommendations to reverse. Third, some sets of cases might be more prone to mistake as instability increases: if panels are split between themselves on whether a case should be

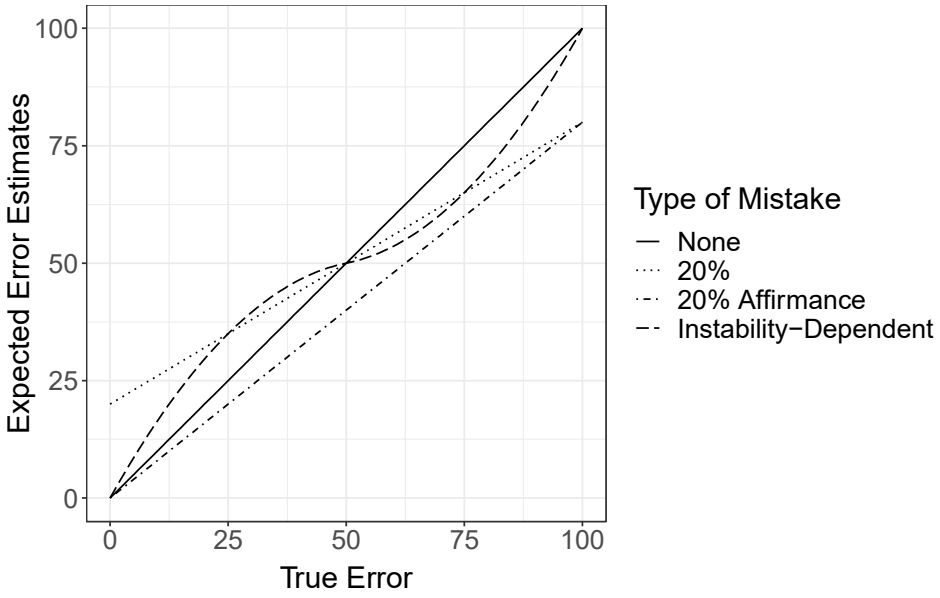
168. These mistake types are not exhaustive of all possible types of mistakes. For example, I leave out what I think is the unlikely possibility that panels are much more prone to mistakenly reversing some sets of cases than mistakenly affirming them. I believe the three types of mistakes I discuss capture the core of what we should be worried about.

reversed, an individual panel may also wrestle with the correct decision, stopping before it arrives at the decision that more reflection would have yielded.[169]

Appendix Figure 1 displays the expected effects from the three different types of mistakes on estimated error as well as expected estimates where there are no mistakes. Note that if the types of mistakes made are constant across all cases, there is little reason for concern that error estimates would be an unreliable guide: for each type of mistake, we would expect estimates to increase with the true degrees of error. But there are concerns if different types of mistakes are made in different sets of cases, reflected by the gaps between lines. For example, consider four different sets of cases, each with true error of 75%. If one set of cases is not subject to mistakes, we would expect those cases to each have an error estimate of 75%. But expected estimates for sets of cases subject to mistakes would be different: an expected 65% error estimate for cases subject to either a general 20% or instability-dependent probability of mistake and an expected 60% error estimate for cases subject to a 20% probability of an affirmance mistake. Thus, although all sets of cases have a true error of 75%, a court using the estimates to focus judicial attention for the purposes of error correction would prioritize some over others.

169. In Appendix Figure 1, I assume that panel mistakes occur at 80% of instability. While I suspect that mistakes are not so frequent, the assumption aids visualization.

APPENDIX FIGURE 1: THE EXPECTED EFFECT OF PANEL MISTAKES ON ESTIMATES OF DISTRICT COURT ERROR



The first thing to note is that even a high rate of inattentive panel mistakes (e.g., 20%) does not cause dramatic departures from the world where there are not mistakes. Second, there are two ways to look at the issue of inattentive panel mistakes. On the one hand, an abundance of panel mistakes would cause error estimates to be less accurate. On the other hand, an abundance of panel mistakes would mean that courts are more in need of statistical precedent. And even where mistakes are plentiful, error estimates can still direct attention—albeit less precisely—toward the decisions that need it.

Nonetheless, the sets of decisions that are routinely and mistakenly affirmed are of particular concern. Their systematically deflated error estimates can prevent courts from rectifying past deficiencies in attention. The problem may be especially pronounced with pro se appeals and other sets of appeals that are routinely assigned to staff attorneys.

Inattentive panel mistakes also affect estimates of instability. As with error estimates, expected instability estimates still tend to increase with true instability. But there is an exception if mistakes are systematically more likely to be affirmances. Because such mistakes would cause systematic understatements of error, and because instability first increases with error and then begins to

decrease after error reaches 50%, the effects are dependent both on which side of the 50% threshold a case is and whether the mistakes cause the case to cross the threshold.

B. Estimating Instability and Adjusting Error Estimates

Estimating instability requires predicting whether each panel would reverse each case, and this poses an acute challenge: data is critical to accurate estimates, but because the same three judges sit together so infrequently, data on any one panel is extremely limited. My approach consists of five main steps. First, I use each individual judge's decisions in the primary training set to build models of each individual judge.[170] Because most individual judges have made a large number of decisions, a machine learning algorithm can make progress in modeling an individual judge's decisions. For each judge, I also build a model on a randomly selected control group of other judges' votes.[171] With both a control and treatment model for each judge, it is possible to estimate each judge's case-specific "voting deviation," or how the judge's probability of voting to reverse each case differs from her colleagues' probabilities. The remaining problem is that we are interested in estimating each panel's probability of reversing each case, but we have little idea how three judges' different voting deviations aggregate to form an ultimate panel decision (previous research has clearly demonstrated that judges do not vote sincerely—their votes are affected by the other members of the panel[172]). The second step addresses this problem. The idea is to estimate the panel reversal probability by building a predictive model with the secondary training set, which can now have panel member voting deviations as variables. The difficulty is that the secondary training set is small—we just used the larger, primary training set to estimate the voting

170. I group together judges who have made fewer than one hundred decisions (mostly judges sitting by designation).

171. The reason for this complexity is that machine learning is data hungry—more data allows for more sophisticated and accurate models. Thus, simply comparing predictions generated from models of different judges who have decided a different number of cases may inflate estimates of interpanel conflict. Consider, for example, two panels that would decide all cases in an identical manner. One panel has decided only fifty cases, while another has decided one hundred. For the first panel, the best an algorithm might be able to do is predict the overall mean—there may not be enough data for the algorithm to take on more variance to reduce bias. Say, then, that the predictions for the first panel are 30% for all cases. For the second panel, there is enough data for a slightly more complex model. For simplicity, assume the model generates two different predictions: 10% for pro se appeals and 50% for represented appeals. Without accounting for sample size, we would consistently estimate a 20% disagreement rate between the panels even though they would actually decide all cases identically.

172. SUNSTEIN ET AL., *supra* note 79, at 20–21 tbl.2-1.

deviations—so we cannot rely on it for the task of both aggregating voting deviations and for predicting reversal more generally. My solution is to use the unitary court model (developed in order to obtain preliminary estimates of error and instability) to generate probability of reversal estimates that can be used as variables in the secondary training set. The secondary training set, though relatively small, now has access to condensed information on the voting tendencies of both individual judges and the collective court, and we can efficiently estimate the aggregate results. In the fourth step, then, I use the judge models (built with the primary training set) to generate predictions for each judge's voting deviations for each case in the test set. The test set now includes, *as variables,* predictions from both the unitary court model and for each judge's predicted effect on the outcome of a case. In the fifth step, I then use the model that was built on the secondary training set (to efficiently aggregate voting idiosyncrasies and collective court patterns) in order to estimate the probability that one thousand randomly selected panel combinations would reverse each case in the test set.[173]

Appendix Figure 2 displays the basic adjustment process. The independent estimates can first serve as a check on the preliminary estimates. Here, the relationship is encouraging but far from ideal. The desired pattern is beginning to take shape: independent estimates of instability increase with error at first and then level off and even begin decreasing after 50% error. But the estimates could clearly be more consistent—one witness is reporting a 6'2" suspect while the other is reporting a 5'10" suspect.

The next three panels of Appendix Figure 2 illustrate the adjustment process. The second panel displays independent estimates of instability after they are rescaled to match the scale of the preliminary estimates of instability.[174] The third panel shows the relationship between the rescaled estimates and the error probabilities after they have been adjusted to better match the rescaled estimates of instability.[175] The fourth panel displays the final

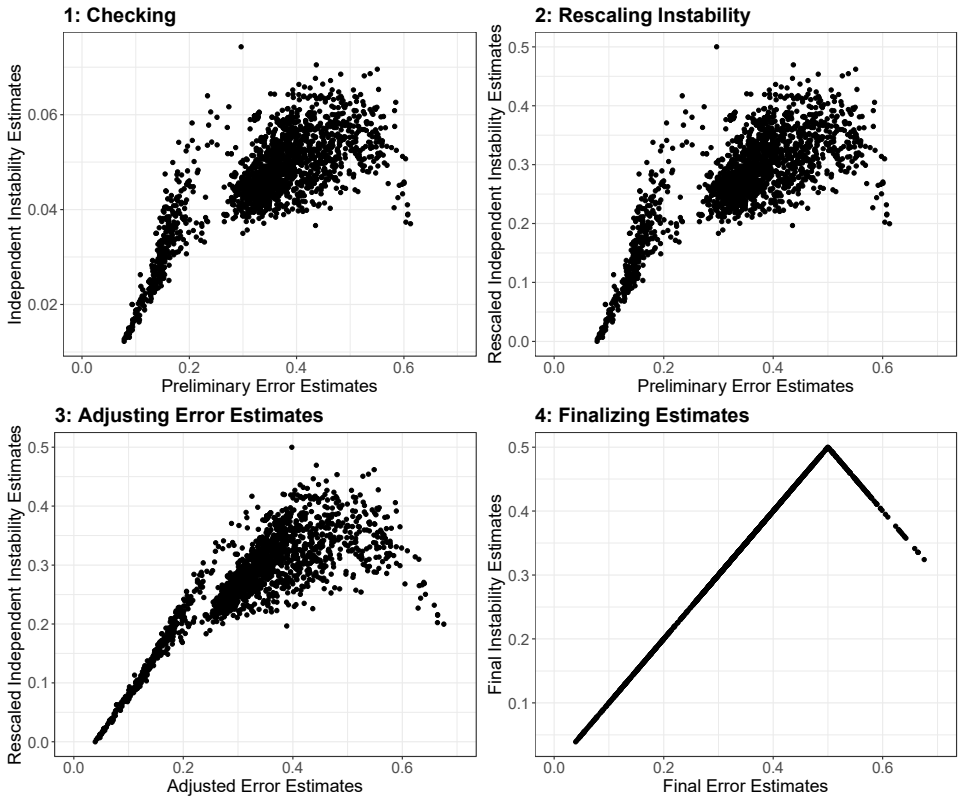
173. For computational convenience, I estimate reversal probabilities for one thousand randomly selected panels rather than all possible panel combinations.

174. I also make use of outside information in hopes of improving the rescale. First, instability cannot be higher than 50%. Second, it is very likely that there are a number of cases that all panels would agree should be reversed. I thus rescale from 0% to 50%.

175. In accordance with the discussion in Section II.C, I increase the weight given to the probabilities of error that are implied by the independent estimates of instability as the preliminary estimates of error move further from the 50% error threshold. For each percentage point from the threshold, I give the independent estimates 3% more weight, with a max weight of 50% for the independent estimates. A case with a 49% or 51% preliminary error probability is thus adjusted 3% toward the independent estimate, and a case with a 48% or 52% preliminary

stage of estimation, where the final estimates of error are matched with final estimates of instability. Appendix Figure 3 displays the distributions of those adjustments to error estimates.

APPENDIX FIGURE 2: TRIANGULATING FINAL ESTIMATES OF ERROR AND INSTABILITY



error probability is adjusted 6% toward the independent estimate. The choices of 3% increases and a cap of 50% weight were selected as a matter of judgment. Optimal weightings and caps should be the subject of future research. Here, my intent was to show that the estimation process could be effective even without knowledge of the decisions in 2012 and 2013.

APPENDIX FIGURE 3: DISTRIBUTION OF ERROR ADJUSTMENTS (TEST SET)