COLLECTION ANXIETY

Jane Bambauer†

I. THE COLLECTION PARADOX..................................................... 198
II. IS COLLECTION COUNTERPRODUCTIVE? .................................. 199
III. IS COLLECTION TOO COSTLY? .................................................. 202
IV. LIMITATIONS................................................................. 204

DNA is not the only path to exoneration.1 Some of the wrongfully accused exculpate themselves using video recordings, transaction records, and other digital information.2 Given the gluttonous data collection practices of our government,3 these exoneration stories should be much more frequent than they are. Digital Innocence by Joshua A.T. Fairfield and Erik Luna puts forward an elegantly simple proposal: criminal defendants should have the same opportunities to access digital records that law enforcement officers and prosecutors do.4 Hear, hear!

Digital Innocence so persuasively demonstrates the exoneration potential of consumer transaction data that the article creates a dilemma for the authors: as the case for defensive use of consumer data is strengthened, so, too, is the case for large-scale collection of data in the first place. In fact, the innocence potential for consumer data arguably justifies not only the collection and retention of data by private third parties but the far more forbidding prospect of increased government data collection. After all, the state is in the best position to collect and permanently store data to ensure that exonerating information is preserved. Better yet, law enforcement agencies could use the data to exclude the suspect before an arrest or charge is brought in the first place.

These implications are unacceptable to Fairfield and Luna. The authors carefully constrain their arguments in favor of defensive access

† Associate Professor of Law, University of Arizona James E. Rogers College of Law. B.S., Yale College; J.D., Yale Law School.

1 And it never was. Between 1989 and 2012, only 37 percent of exonerations involved DNA evidence (though of course the proportion was much higher for sexual assault exonerations). SAMUEL GROSS & MICHAEL SHAFFER, NATIONAL REGISTRY OF EXONERATIONS, EXONERATIONS IN THE UNITED STATES, 1989–2012, at 22 (2012).


4 Fairfield & Luna, supra note 2.
to information already collected and retained. They ask, “If the powerful are going to spy on us and collect data about our lives, shouldn’t we know what information is being gathered and have access to the resulting databases?” The emphasis is on “if.” They also come out strongly against the secret and pervasive data collection practices of the NSA and other arms of government, calling them “anathema to a liberal, open democracy.”

Fairfield and Luna are wise to put these limits on their argument if for no other reason than that the exoneration potential of already collected data can be taken seriously without the distraction of the collection question. But Fairfield and Luna do not appear to limit their argument for purely practical or instrumental reasons. Their strong declarations against collection show they would be loath to contribute anything that could be used to justify more invasions of our privacy.

We have seen this before. The battle between the academy’s love for exoneration and hatred for collection has been raging in the context of DNA databases for several years. Digital information is just another venue for the long fight between privacy and exoneration. Generally speaking, the academy splits its allegiances. Privacy is more important at the collection stage, and exoneration is paramount thereafter. Attention is rarely given to the pesky inconsistency between the anti-collection position and the subsequent vindication of those very same collections through exoneration stories.

Of course, not every superficial inconsistency proves to be so on closer inspection. Perhaps exoneration stories make the best of a bad situation. Perhaps the few hundred reversed convictions do not, and cannot, justify large-scale collection of personal information. Two powerful arguments can be made in favor of the “pro-exoneration but anti-collection” instinct that Fairfield and Luna possess. The first is that innocence and noncollection do not actually compete: although already collected information can and should be used to exonerate, if we were to move to a system of increased collection and use of consumer data, we could very well wind up with more rather than fewer wrongful convictions.

The second explanation for the “pro-exoneration, anti-collection” viewpoint is that even if increased collection of data could marginally

---

5 Id. at 986.
6 Id.
7 Id. (“To be clear, the following [discussion of the use of Big Data to prevent wrongful convictions] is not an apologia for data gathering in service of national security or commercial interests.”).
8 See, e.g., Erin Murphy, Relative Doubt, 109 Mich. L. Rev. 291, 313 (2010) (discussing the pros and cons of familial DNA searching); see also longer discussion infra Part I.
9 See Fairfield & Luna, supra note 2, at 989–91, for further discussion of the current academic literature.
improve the rate of exoneration, or marginally decrease the rate of wrongful conviction, these marginal improvements cannot justify the tremendous costs to privacy.

This Essay takes on both of these arguments and finds them lacking.\textsuperscript{10} The “pro-exoneration but anti-collection” instincts that dominate the legal academy exaggerate the risks that come with large collections of consumer data.\textsuperscript{11} They also ignore the significant benefits that come from data collection.\textsuperscript{12} These benefits include new constraints on police discretion, the reduction of systemic bias in law enforcement, and, of course, vindicating the rights of the innocent.\textsuperscript{13}

This Essay makes the uneasy and unpopular case for increased collection and retention of consumer data—by the government, no less. To be clear, this is a thought experiment, and leaves many crucial questions for future research. I personally do not hold strong convictions that the government should engage in massive data collection and retention. But at the same time, for reasons sketched out in this Essay, the strong convictions against government data collection felt by the overwhelming majority of legal scholars are equally premature, and do not hold up well against careful analysis.

The Essay proceeds in four parts. Part I describes the paradox that Fairfield and Luna embrace by recognizing the salutary effects of consumer data without endorsing the collection of that very same data. Parts II and III describe and challenge the strongest arguments in favor of this conceptual splitting. Part II argues that on balance, increased collection of consumer data is likely to increase (rather than decrease) the accuracy of arrests, prosecutions, and convictions. Although it would be convenient if collection made criminal investigation outcomes worse, so that privacy and innocence could work together, this is wishful thinking. Privacy and accuracy really are at odds.

Part III responds to the argument that privacy is a more valuable social interest than marginal improvements in accuracy. It argues that privacy should not automatically trump the competing interests in government data collection because the potential benefits from data collection go beyond vindication of the falsely accused. Data-driven law enforcement can also limit police discretion and unearth systemic bias. However, all of the rosy potential for data described in Parts II and III is contingent on the implementation of fair data collection and use practices. So Part IV concludes by acknowledging some of the assumptions and important limitations that cabin the Essay’s arguments
to the realm of the theoretical.

I

THE COLLECTION PARADOX

*Digital Innocence* walks a fine line between data optimism and pessimism. On one hand, the authors proclaim their fidelity to the privacy movement, and criticize the government and private parties for collecting as much information as they do. On the other hand, the authors celebrate the exoneration potential of this same data. This puts Fairfield and Luna in a precarious position, but it is not unique to them. The utility of DNA databases for Innocence Project work has created similar tension between privacy and exoneration.

The propriety of DNA collection has left forensic experts in disorder. Because DNA databases add obvious value for accuracy and exoneration, scholars have been tripping over themselves to try to explain why DNA should not be collected in bulk. Elizabeth Joh, Erin Murphy, and Andrea Roth have each criticized the collection of DNA samples from arrestees. Joh objects that arrestee DNA databases create incentives for police to increase arrests and for legislatures to create new crimes so that the police can expand their databank. Murphy argues that familial DNA matching (where a partial match leads police to suspect a relative of somebody in the DNA database) permits law enforcement to proceed on the basis of “guilt by association.” Roth argues that minorities will be disproportionately represented in an arrestee DNA database, which will exacerbate existing disparities in criminal enforcement rates. Each of these problems could be avoided if the police had access to a population-wide DNA database—a solution reluctantly endorsed by Roth but dutifully shunned by the Supreme

14 See Fairfield & Luna, supra note 2, at 986.
15 See id. at 1075–76.
16 For example, Jason Kreag, writing about post-conviction access to the federal DNA database, explicitly puts the collection debate aside and focuses on access to databases that already exist. Jason Kreag, *Letting Innocence Suffer: The Need for Defense Access to the Law Enforcement DNA Database*, 36 CARDOZO L. REV. (forthcoming 2015).
20 Id. at 297. See also D.H. Kaye & Michael E. Smith, *DNA Identification Databases: Legality, Legitimacy, and the Case for Population-Wide Coverage*, 203 WIS. L. REV. 413 (2005). Erin Murphy also recognizes that this is the logical endpoint if we have confidence in DNA matching methods. Erin Murphy, *The Art in the Science of DNA: A Layperson’s Guide to the Subjectivity Inherent in Forensic DNA Typing*, 58 EMORY L. J. 489, 511 (2008) (“Either our confidence in such forensic techniques should be so great that we should not bristle at universal application, or we should recognize that the fallibility and subjectivity of such methods spark
Court.21 Justice Scalia dubbed the idea a “genetic panopticon.”22 Given the Fourth Amendment’s laxness on government collections of third-party data,23 the Court will have to grapple with the pros and cons of the “digital panopticon” soon enough.

There are at least two distinct reasons that society might wish to limit the collection of information despite its potential to exonerate the innocent.

The first is a counter-productivity argument.24 If the government is permitted to vacuum up all of our personal data indiscriminately, we will wind up with more false accusations and wrongful convictions. With a world of historical transaction data, the police will always be able to find somebody to accuse for any given crime, and some portion of those accused will be factually innocent. Under this theory, innocence and anti-collection are not at odds. Even if data occasionally improves the accuracy for somebody falsely accused, it will just as frequently wrongly wrap some innocent person into an investigation.

The second is a fundamental rights argument.25 Even if data collection can reduce the rate of wrongful arrests and convictions, the incurable damage done to privacy, public trust, and civil liberties are too great to seriously entertain unlimited government collection of personal information.

The remainder of this Essay will uncover and explore the assumptions driving these two defenses of the anti-collection, pro-exoneration position. Neither defense is persuasive. On the other hand, the benefits that could come from extensive data collection are contingent on optimistic assumptions about the existence of transparency and accountability mechanisms, which are not currently in place.26 Because we are still in the early days of data analytics and information collection technologies, the Essay’s exploration will be abstract, the conclusions tentative and qualified.

II
IS COLLECTION COUNTERPRODUCTIVE?

The first argument against data collection posits that collecting and
hoarding data will actually increase wrongful arrests and convictions. It will attract police attention to evidence that appears suspicious but is actually innocent. That is, mass data collection will increase the number of innocent people who are arrested, charged, or convicted.27 If this is true, the argument is unassailable: even on a pure cost-benefit calculus, the required invasions of privacy would fail to serve their purpose of avoiding false arrests and convictions.

Some scholars have made this counter-productivity argument against forensic evidence. They have debunked the presumed reliability of bite-mark analysis28 and fingerprinting.29 Even DNA matching has caught some criticism.30 Carole McCartney, a British criminal law scholar and expert in forensic evidence, has raised important questions about the quality control of forensics (including DNA typing). She makes ominous claims about its future: “[W]hilst forensic science may aid in exculpating innocent suspects, it can also convict the innocent. Forensic science may yet come to usurp ‘fairness’ and render obsolete the presumption of innocence.”31

Danielle Citron and Frank Pasquale raise similar concerns about the data-ravenous credit scoring industry.32 Citron and Pasquale accuse the industry, despite its access to rich and finely grained data, of scoring people in arbitrary and discriminatory ways.33 Police access to digital information could have a similar counterintuitive relationship with

27 This could occur one of two ways. The first, explored here, is that the data could decrease the average accuracy of investigations. The second is that even if accuracy is improved, the volume of digital information could cause an increase in the number of investigations. If this occurs, it is possible to have a more accurate system in terms of hit rate while nevertheless causing more false arrests in absolute numbers. To economize on space here, I will not address this second possibility. But I have written about it elsewhere and have identified solutions. Jane Bambauer, Hassle, 113 MICH. L. REV. (forthcoming, 2014).


30 Murphy, supra note 8, at 309, 315; Murphy, supra note 20, at 490–92; Murphy, supra note 28, at 754–56. Murphy takes great pains to make clear that DNA matching is much more objective and scientific than other forms of forensic science but that even properly performed DNA typing requires discretion and maintains a risk of error. Murphy, supra note 20, at 490–92. But she does, in the end, suggest that the “specter of erroneous attribution . . . should loom largely in the minds of policymakers as they approve expansive forms of DNA databasing and searching.” Id. at 511.

31 CAROLE MCCARTNEY, FORENSIC IDENTIFICATION AND CRIMINAL JUSTICE 67 (2012). McCartney also worries that reliance on forensic evidence can undermine some of the criminal procedure protections we have long relied on to preserve values other than accuracy and efficiency. Id. at 64, 66. These are explored in Part III.


33 Id. However, the authors do not compare the fairness of credit outcomes based on today’s data-rich credit scoring to the outcomes one would expect to see without consumer data—using only income, wages, or other basic application information. Without this baseline comparison, the damning conclusions Citron and Pasquale draw are premature.
accuracy in the criminal justice system.

This is an empirical argument and one that is not likely to be winning. In order to maintain the belief that collecting more data would make things worse, it is not enough to point to the potential for error. (Every criminal investigation system will have some.) Instead, scholars who wish to show that a new investigation tool underperforms the old ones must be prepared to demonstrate not just error but comparatively worse error. More precisely, to reject a new investigation method outright, the error from the combination of old and new methods must be worse than the old methods alone.

Once we begin to catalog the old methods of proving criminal conduct, the counter-productivity argument seems implausible. Traditional prosecutions are based on testimony—either the victim’s, an eyewitness’s, or the accused’s own confession. All of these forms of evidence are known to be unreliable.34 Thus, when scholars criticize the lack of scientific rigor in new tools for criminal investigation and prosecution,35 they overlook the fact that the entire enterprise of American criminal justice lacks scientific rigor.

Pushing police away from digital information forces investigations to maintain heavy reliance on informants, eyewitness testimony, and confessions. With information as rich as geolocation, nearly complete purchasing histories, and Google Glass video recordings, an assumption that increased use of data would make things worse requires some mental gymnastics.

The counter-productivity argument would have simplified the ethical and philosophical questions raised by mass data collection if it had been descriptively correct. But because it is not, we must move on

---

34 Gross & Shaffer, supra note 1, at 40 (finding that mistaken eyewitness identifications and false confessions are contributing factors to wrongful convictions). Studies of eyewitness identifications in police station lineups find that about one-third of the eyewitnesses who identify anybody at all choose a filler. See State v. Henderson, 27 A.3d 872, 885–89 (N.J. 2011) (describing empirical studies that tested the reliability of eyewitness identifications). The International Association of Chiefs of Police has said about eyewitness identification: “[O]f all investigative procedures employed by police in criminal cases, probably none is less reliable than the eyewitness identification. Erroneous identifications create more injustice and cause more suffering to innocent persons than perhaps any other aspect of police work.” Id. at 885–86 (quoting INT’L ASS’N OF CHIEFS OF POLICE, TRAINING KEY NO. 600, EYEWITNESS IDENTIFICATION 5 (2006)).

35 See, e.g., Mnookin et al., supra note 29; Paul C. Giannelli, The “Science” of Wrongful Convictions, 18 CRIM. JUST. 55, 55 (2003). Some scholars worry that, even if an empirical approach to criminal identification is more accurate than the usual forms of proof, such evidence will be presumed infallible. See, e.g., Laurence H. Tribe, Trial By Mathematics: Precision and Ritual in the Legal Process, 84 HARV. L. REV. 1329, 1334 (1971) (stating that mathematical arguments in trials may be given more weight than they deserve because advanced math is “at once impenetrable by the layman and impressive to him”). But this contradicts other theories that jurors are more convinced by a good-faith eyewitness in court identification than anything else. See Watkins v. Sowders, 449 U.S. 341, 352 (1981) (Brennan, J., dissenting).
to the second, trickier defense of the anti-collection instinct.

III

Is Collection Too Costly?

The next defense of the anti-collection position accepts that data collection may reduce wrongful convictions and asserts that the interests in privacy are nevertheless more valuable. This is the dark, less-talked-about companion to our commitment to let “ninety-nine guilty go free”—we are willing to lock up the occasional innocent in order to promote the civil liberties of the rest. Framed this way, the commitment to privacy comes at a price, but the trade-offs are not necessarily unreasonable. After all, all systems have error. Using all possible means to minimize the chance of false conviction can create intolerable costs.

We have never seriously considered driving down false convictions to the lowest possible rate. To give one silly but illustrative example, even before the advent of computing, the state could have minimized the risk of false convictions by allowing (or requiring) all homes to be forcibly searched for evidence that might exculpate a criminal suspect. But this would violate the Fourth Amendment and the values that undergird it. The Constitution has determined for us that privacy can frustrate some innocence-serving searches in order to ensure that Americans can count on having some minimally adequate amount of privacy and security. These privacy commitments predate the digital revolution, and they will persist.

Most of the criminal procedure and privacy scholars who examine the issue of government data collection have correctly identified the perils that Big Data has brought to criminal investigation. They have described the dramatic changes in the quantity of information that can be collected, the variety of previously unconnected information that can be aggregated, and the ease with which these large data troves can be searched for wide-ranging purposes (both valid and not).

But scholars have failed to fully appreciate how the searchability of data—the use of filters and responsive algorithms—can fundamentally alter the basic assumptions about both privacy and law enforcement. To use another silly but illustrative example, suppose it were possible to automatically scan the contents of every American’s home in order to find the house


38 This point is not lost on Fairfield and Luna, though. In building the case for Digital Innocence, the authors recognize the importance of refined searches in order to be confident that defendants will be able to find what they need. Fairfield & Luna, supra note 2, at 1072. These limited searches would likely restrict defendants’ access to records that may be relevant to their case.
concealing a murder weapon without discovering any other personal and intimate details. This sort of magical searching process could help protect the wrongfully accused without interfering with privacy and autonomy to the extent that physical searches of every home would.

Because of searchability, Big Data has potential that goes well beyond strengthening the efficiency of law enforcement (a good thing but not necessarily a paramount interest). Big Data can also constrain discretion, decrease systemic bias, protect the innocent, and improve government accountability. Put simply, Big Data can promote equitable law enforcement.39

If done properly, data-driven policing will decrease reliance on the observations, discretion, and assumptions of individual police officers. By searching databases for patterns that suggest a high probability of criminal conduct (either by searching for facts related to a specific criminal incident40 or by using an algorithm that reliably detects criminal behavior), police can reduce their reliance on traditional, but flawed, investigation methods.41

Pattern-based data mining is likely to improve the distributional fairness of the criminal justice system through two mechanisms. First, for any given crime, data mining provides an option for building suspicion without resorting to vague factors like “high crime area[s]”42 and “furtive movements,”43 which tend to concentrate in poor and minority neighborhoods. Second, pattern-based data mining improves the detection of different sorts of crimes—crimes like fraud, insider trading, and identity theft—which brings the attention of law enforcement to criminals who have greater social status and wealth.44

These distributional benefits are separate and distinct from the interests in protecting the innocent—the interest built up in Digital Innocence but then undervalued in the authors’ denouncement of data collection. The potential benefits for wrongly accused individuals is

---

39 I explore each of these in much more depth in other work. See Bambauer, supra note 27 (discussing police discretion to search and ways to constrain that discretion); Jane Bambauer, Defending the Dog, 91 OR. L. REV. 1203 (2013) (describing how technologies, both old and new, can reduce reliance on police discretion).
40 Christopher Slobogin gives special consideration to investigation procedures that move from the facts of a crime toward the identity of a suspect instead of the other way around. He calls these “event-driven” investigations. CHRISTOPHER SLOBOGIN, PRIVACY AT RISK: THE NEW GOVERNMENT SURVEILLANCE AND THE FOURTH AMENDMENT 186 (2007).
41 Tal Zarsky has argued that pattern-based data mining has the potential to radically reduce law enforcement bias and inequities if it is done right. Tal Z. Zarsky, Governmental Data Mining and its Alternatives, 116 PENN. ST. L. REV. 285, 289–90, 311–12 (2011); Tal Z. Zarsky, Automated Prediction: Perception, Law, and Policy, 9 COMM’NS. OF THE ACM 33, 35 (2012).
43 See, e.g., People v. Woods, 64 N.Y.2d 736, 737 (N.Y. 1984); Bambauer, supra note 27.
greater than even Fairfield and Luna let on. Data collection could protect wrongly suspected individuals at every point in the criminal justice process. Accurate data could undermine the probable cause that the police would otherwise have. This would protect the innocent person from search and arrest, and certainly from wrongful conviction. And of course, as Fairfield and Luna demonstrate, if criminal charges are brought against the wrong person, that person would have all the motivation needed to search for exculpating data. But in order to exculpate, the data first must exist.

The protection of the innocent and the equitable treatment of all are some of the ends that privacy is meant to achieve. So, when a conception of privacy works against those things—obstructing exoneration and frustrating evenhanded enforcement—it is all the harder to defend its moral supremacy. To be clear, neither of these interests (reduced discretion or protection of the innocent) requires law enforcement to have unbounded access to personal data or to perform searches on any target it chooses without justification. But they do require the collection of large amounts of data that comprehensively describe the population.

IV LIMITATIONS

In the end, the assumptions that lead scholars like Joshua Fairfield and Erik Luna to fervently resist data collection may be wrong. Worse, they may be wrong and popular. By recognizing that the act of data collection is yoked to its potential for exoneration, Digital Innocence forces readers to think about an uncomfortable choice. We must select between two mutually exclusive worlds: one where the government has data that can make the criminal justice system better, and one where the government does not have data that it might abuse. For any category of personal data, the government can have it or the government cannot have it. A choice must be made between these two worlds.

The noncollection option has the advantage of certainty. The police will not be able to exploit information and run unlimited searches on dissidents, journalists, and personal enemies if they have no consumer data. But just as certainly, the government will not be able to make accuracy, efficiency, and fairness improvements using that data.

45 See Fairfield & Luna, supra note 2.

46 These concerns are not theoretical. Occasionally the government has accessed phone records of reporters, jeopardizing confidential sources and suggesting that the government is engaging in selective prosecution of journalists who have written critical articles. Mark Sherman, Government Obtains Wide AP Phone Records in Probe, ASSOC. PRESS., May 13, 2013. However, these instances seem to be rare and run against federal internal investigation policies. Brad A. Greenberg, The Federal Media Shield Folly, 91 WASH. U. L. REV. 457, 450 (2013).
Viewed this way, the collection option looks pretty attractive, especially if it is combined with precautions to restrict overuse and abuse of the data. However, this optimistic vision for mass data collection is hypothetical. Law enforcement agencies are just beginning to use data-driven methods. This Essay’s arguments contain some implicit assumptions that make them work.

Mainly, the benefits from large-scale data collection will not clearly outweigh the drawbacks so long as law enforcement can use personal data indiscriminately. The public will understandably withhold its trust if the police can use unfettered discretion to pick a target and access a long stream of data about him. So far, the Fourth Amendment precedent and various industry-specific privacy laws have done nothing to cut off the free-for-all for target-driven investigations. Indeed, the Supreme Court has crafted the Fourth Amendment to leave considerable latitude for police discretion. So, while data collection can decrease discretion and improve the equitable distribution of law enforcement, these benefits are speculative without firm legal commitments to reduce discretionary suspect-driven searches.

Moreover, the benefits of data collection that I trumpet in this Essay cannot be harnessed without transparency. Jack Balkin has warned that a government hooked on data (what he calls the National Surveillance State) is destined to become both a data glutton and a data miser. Like gluttons, the government will collect everything it can; and like misers, it will keep the data and its operations secret. Even if the gluttony is not as bad as we think it is, a lack of transparency will throw all the benefits into doubt. Transparency is a precondition for efficacy and accountability. Efficacy ensures that the promise of increased accuracy is not illusory, and accountability ensures that the anticipated decreases in discretion are real.

The form of transparency will have to balance the good kind of secrecy (preventing criminals from knowing how crime is detected) from the bad kind (preventing citizens from learning their government’s mistakes). But whatever its form, effective transparency

47 Moreover, law enforcement is not especially focused on limiting its discretion. To the contrary, law enforcement increasingly uses the services of Palantir and other information aggregators that allow police to comb through years’ worth of information after providing a name. LAPD’s Data Mining Program Has CIA Roots, CNN (May 25, 2014), http://www.cnn.com/video/data/2.0/video/tech/2014/05/25/cot-la-license-plates.cnn.html.
48 See Balkin, supra note 3, at 23.
49 Joh, supra note 17, at 283 (discussing Whren v. United States, 517 U.S. 806 (1996), and Atwater v. Lago Vista, 532 U.S. 318 (2001)).
50 Balkin, supra note 3, at 17.
51 Id.
52 Designing a good set of transparency rules is no easy task. See Mark Fenster, The Transparency Fix: Advocating Legal Rights and Their Alternatives in the Pursuit of a Visible State, 73
will have to incorporate the defensive access rights that Fairfield and Luna promote in *Digital Innocence*.

Regardless of its wisdom, government data collection is likely to accelerate. Defensive access to government and third-party data protects society in two ways. First, access to data advances the due process rights of innocent defendants. Second, even when a defendant accessing data is not factually innocent, such access provides a checking function on the government by uncovering ignored leads or unexplained differences in how the police enforces crime. The lessons from *Digital Innocence* will help bend the inevitable expansion of data-driven policing towards justice.