Big Data Discovery, Privacy, and the Application of Differential Privacy Mechanisms[♦]

James A. Sherer, Jenny Le, and Amie Taal[*]

Introduction

The age of information overload presents important implications for data privacy that are increasingly evident in eDiscovery and internal investigation practices, where the associated data collections now evidence characteristics of traditional big data,[1] including related size and complexity hurdles that challenge traditional data processing models. Big data analytics are progressively used in eDiscovery and internal investigations to manage cost and efficiency and return analysis simply unavailable in years past. But big data also presents challenges in terms of absolute volume and additional elements of velocity, variety, and variability. Each of these elements in turn increases the potential amount of personally identifiable information ("PII") within a dataset, and when operating in concert, can magnify data privacy concerns.

These concerns will not upend the current practice and the increasing trend of using big data for eDiscovery and investigative processes, as a number of "class actions and other lawsuits typically invoke statistical sampling"[2] and other matters use "statistical and qualitative analysis, in conjunction with explanatory and predictive models"[3] as core components to practice. The use of these data sets and attendant analytics will instead continue to grow, and as analytics become more sophisticated and eDiscovery and investigative datasets (the "Collections") become bigger and richer, there is an increasing danger that unsecure big data analytics will unwittingly—or intentionally—unveil PII. This paper introduces legal and information governance practitioners to a new breed of algorithmic techniques and evaluates whether the application of these techniques is sufficient to mitigate the danger of PII disclosure. A resounding "maybe" is the conclusion, and even that depends upon practitioner decision-making on the front end of the process. As this paper explains, there is no mathematical *lapis philosophorum*[4] waiting in the wings, but there are tools and practices that can help mitigate these new concerns.

As mentioned above, Collections, once indexed and refined, form databases of unstructured (and, less frequently but still importantly, structured[5]) data sources, where admittedly arbitrary definitions of "big cases"

♦ Submitted on May 14, 2105 as a Refereed Paper for the ICAIL 2015 Workshop on Using Machine Learning and Other Advanced Techniques to Address Legal Problems in E-Discovery and Information Governance ("DESI VI Workshop") and accepted for publication in the July, 2015 issue of *The Computer & Internet Lawyer Journal*.

[*] James A. Sherer is Counsel in the New York office of Baker Hostetler LLP; Jenny Le is a Vice President at Evolve Discovery, and Amie Taal is a Vice President at Deutsche Bank. The views expressed herein are solely those of the authors, should not be attributed to their places of employment, colleagues, or clients, and do not constitute solicitation or the provision of legal advice. Special thanks to Kevin Wallace for his assistance with this Paper.

[1] Kate Crawford and Jason Schultz, *Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms*, 55 B.C.L. Rev. 93, 96 (2014), http://lawdigitalcommons.bc.edu/bclr/vol55/iss1/4 (Big Data is "a generalized, imprecise term that refers to the use of large data sets in data science and predictive analytics" that "may include personal data generated from a variety of sources"), *citing* Omer Tene and Jules Polonetsky, *Big Data for All: Privacy and User Control in the Age of Analytics,* 11 Nw. J. Tech. & Intell. Prop. 239, 240 (2013), http://scholarlycommons.law.northwestern.edu/njtip/vol11/iss5/1.

[2] David J. Walton, *How lawyers and law firms operate in a Big Data world*, INSIDECOUNSEL (Apr. 11, 2014), http://www.insidecounsel.com/2014/04/11/how-lawyers-and-law-firms-operate-in-a-big-data-wo.

[3] ERNST &YOUNG, *Forensic data analytics - Globally integrated compliance review, litigation support and investigative services*, http://www.ey.com/Publication/vwLUAssets/ey-forensic-technology-and-discovery-services/$FILE/ey-forensic-technology-and-discovery-services.pdf

[4] *See also panacea*, *nostrum*, or *catholicon.*

[5] Conrad Jacoby, Jim Vint & Michael Simon, *Databases Lie! Successfully Managing Structured Data, The Oft-Overlooked ESI*, 19 RICH. J.L. & TECH 9 (2013), http://jolt.richmond.edu/v19i3/article9.pdf.

may "start at three million" documents and range to "hundreds of millions" for just one client.[6] We considered these Collections "big data" by virtue of their size and complexity, but there are additional wrinkles to these data sets. By their very operation, Collections are overbroad,[7] and in addition to the breadth of Collections generally, certain types of early case assessment Collections also involve an investigation of the data before litigation even begins. This may implicate "data privacy issues when [the practitioner] is basically on an early fishing expedition."[8] These data privacy concerns apply to both the original collectors of the data as well as the parties to whom the Collections are provided.

Concerns associated with "protecting privacy and confidentiality in computer network data collections" have been discussed in other disciplines,[9] but these discussions are generally focused on only one type of data analysis. In practices applied to Collections, however, data analytics may be applied in any (or all) of the following stages: (1) collection; (2) processing; (3) cleansing; (4) integration and analysis; (5) refinement of original intent purposes, (6) review and revision; and (7) production.[10] Impacts on data privacy exist at each stage, including the disclosure or unmasking of PII. Further issues can arise through the sometimes *ad hoc* nature of the Collection process, such as storing all the "crown jewels" in one place, leaving these enriched data sources less protected than they would be otherwise segregated within the organization. These additional issues are beyond the scope of this paper. Instead, we focus on the implications of big data analytics on Collections that may impact privacy. And because one particular method—the use of Differential Privacy (or "DifP") techniques—was posited as the industry "gold standard" for data privacy,[11] we chose to examine whether the application of DifP could save practitioners from themselves when confronting and utilizing these types of big data collections.

Data Privacy as a Factor of Collection Analysis

There is no debate that these big data Collections are being used—and are increasingly being used—within the practice of law. Some note that present-day practitioners must "understand…when to marshal big data analytics to build a case" and that "data automatically generated by social media applications and mobile devices constitute a potential treasure trove of evidence."[12] Others state that the use of technology assisted review "uses algorithms in much the same way that Amazon can offer you selections based on what you've bought in the past."[13] The promise is that further integrations of technology and data, such as data extraction, may lead practitioners "beyond simple pattern-matching" and into providing "the ability to make inferences based on a set of rules."[14]

---

[6] David J. Parnell, *John Tredennick And Mark Noel Of Catalyst, On Technology Assisted Review*, FORBES (Feb. 18, 2015), http://www.forbes.com/sites/davidparnell/2015/02/18/john-tredennick-mark-noel-catalyst-technology-assisted-review/.

[7] Jack Halprin, *The Legal Hold Action Plan - Best Practices for Meeting the Preservation Obligation*, ABA QUICKCOUNSEL (May 3, 2011), http://www.acc.com/legalresources/quickcounsel/Preservation-Obligation-QC.cfm.

[8] Sheryl Nance-Nash, *Predictive coding and emerging e-discovery tools*, Corporate Secretary (Aug. 14, 2013), http://www.corporatesecretary.com/articles/ediscovery-and-records-management/12507/predictive-coding-and-emerging-e-discovery-tools/.

[9] Daniel Kifer & Ashwin Machanavajjhala, *Pufferfish: A Framework for Mathematical Privacy Definitions*, ACM Transactions on Database Systems TODS, 39(1), 2014, http://www.cse.psu.edu/~dkifer/papers/pufferfishjournal.pdf.

[10] Heiko Müller & Johann-Christoph Freytag, *Problems, Methods, and Challenges in Comprehensive Data Cleansing*, TECHNICAL REPORT HUB-IB-164, Humboldt-Universität zu Berlin, Institut für Informatik (2003), http://www.informatik.uni-jena.de/dbis/lehre/ss2005/sem_dwh/lit/MuFr03.pdf.

[11] Jane Bambauer, Krishnamurty Muralidhar & Rathindra Sarathy, *Fool's Gold - An Illustrated Critique of Differential Privacy*, Vanderbilt JET, Vol. 16 No. 4 (2014) at 703, http://www.jetlaw.org/wp-content/uploads/2014/06/Bambauer_Final.pdf.

[12] Walton, *supra* note 2.

[13] Nance-Nash, *supra* note 8.

[14] *Id.*

These Collections also give rise to new data privacy concerns. More benign considerations of PII may center on the ability to distinguish, identify, trace, or link information about an individual,[15] and to subsequently "nudge" them or influence their behavior.[16] However, the types of legal-based analysis that focus on the "types of data that are commonly used for authenticating an individual, as opposed to those that violate privacy, that is, reveal some sensitive information about an individual"[17] may be more concerning. The loss of PII may include identity theft; embarrassment; blackmail; a loss of public trust; legal liability; and/or remediation costs.[18] More extrapolated examples include perpetuating discriminatory practices; individualized uses of health information; and predictive policing.[19] These may also extend to the problems of "aggregate information" which, while it "gets less attention than the problem of protecting individual records…is most relevant to business data where aggregates reflect different kinds of business secrets."[20] Finally, all of these considerations are underscored by the real issue of whether or not the data relied upon is valid. Issues related to "data accuracy and integrity"[21] permeate and inform each and every data release concern.

In short, the unintentional dissemination of PII and related and extended other secrets—through the operation of a Collection—can cause real problems. And that this may occur is unsurprising given present-day enriched data sets as well as the technology assisted review tools that "can quickly analyze millions of documents for subtle patterns"[22] and provide insights never before available to practitioners. However, most general *legal* discussions do not incorporate a consideration of how applied technology within the practice of law impacts privacy[23]—and specifically, how the increasing data sizes associated with data collections, including those within the scope of eDiscovery or similar investigations, need to address the implications of removing PII or the related ability to piece PII together. Other times, the data privacy concerns might center only on the location of the data and the manner in which it is collected. Still another issue associated with legal collections of data is that even "simpler" approaches, where the release of only "aggregate" information, those statements about large groups of people, seems like a facially workable measure. However, "even this approach is susceptible to breaches of privacy,"[24] and these collections may be manipulated to operate in a diametrically opposed way.

[15] Erika McCallister, Tim Grance & Karen Scarfone, *Guide to Protecting the Confidentiality of Personally Identifiable Information (PII)*, NIST SPECIAL PUBLICATION 800-122, 2-1 (2010), http://csrc.nist.gov/publications/nistpubs/800-122/sp800-122.pdf.
[16] David Stewart, *Protecting Privacy in the Age of Big Data and Analytics*, WALL STREET RISK & COMPLIANCE JOURNAL (Nov. 3, 2014), http://deloitte.wsj.com/riskandcompliance/2014/11/03/protecting-privacy-in-the-age-of-big-data-and-analytics/.
[17] Arvind Narayanan & Vitaly Shmatikov, *Myths and Fallacies of Personally Identifiable Information*, VIEWPOINTS, ACM, Vol. 53 No. 6 (2010), https://www.cs.utexas.edu/~shmat/shmat_cacm10.pdf.
[18] McCallister et al., *supra* note 15.
[19] Crawford, *supra* note 1.
[20] Kifer & Machanavajjhala, *supra* note 9.
[21] President's Council of Advisors on Science and Technology, *Report to the President, Big Data and Privacy - A Technological Perspective* (2014) at xii, https://www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_big_data_and_privacy_-_may_2014.pdf.
[22] Parnell, *supra* note 6.
[23] Infotech Europe, *Paradigm shifts - random thoughts on predictive coding, data privacy, IBM, neuroscience and other stuff as we close out the year* (Dec. 23, 2012), http://www.infotecheurope.com/paradigm-shifts-random-thoughts-on-predictive-coding-data-privacy-ibm-neuroscience-and-other-stuff-as-we-close-out-the-year/. ("There is never a technical/scientific speaker to discuss how technology has simply eroded our privacy … some of it willingly").
[24] Erica Klarreich, *Privacy By the Numbers - A New Approach to Safeguarding Data*, SCIENTIFIC AMERICAN (Dec. 31, 2012), http://www.scientificamerican.com/article/privacy-by-the-numbers-a-new-approach-to-safeguarding-data/.

Privacy Considerations Applied in Legal Practice

Legal practitioners have recognized that data privacy is among the considerations at play in eDiscovery, where "professionals are [expected to be] familiar with local data transfer and privacy rules, state secrecy laws and other local requirements affecting international forensic data analytics."[25] There is even some scholarship on how US litigators can navigate US discovery rules in the face of European Union and other foreign data privacy statutes.[26] These considerations may "focus solely on the types of data that are commonly used for authenticating an individual, as opposed to those that violate privacy, that is, reveal some sensitive information about an individual."[27] This discussion is focused on the latter.

The discovery process is not a bar to the production of information that may contain privacy information; indeed, in litigation, courts routinely, by order, "require production, where necessary, of records that reflect medical treatment, *sometimes* with the identities of the actors redacted."[28] This extends as well to instances where "a health care provider may disclose protected medical information in response to a discovery request"[29] as long as "reasonable efforts have been made by such party to secure a qualified protective order."[30] The same holds true for the application of the Gramm-Leach-Bliley Act.[31]

There are some efforts put forth towards recognizing that data privacy comes in different categories or "confidentiality impact levels."[32] Traditional methods to ensure privacy include, tokenization, redactions[33] of the data,[34] or other types of de-identification measures that remove enough data such that "the remaining information does not identify an individual and there is no reasonable basis to believe that the information can be used to identify an individual."[35] These considerations may also include the more traditional means of obtaining protective orders[36] and/or filing under seal.[37] Despite the operation of these mechanisms, this may leave the issues associated with Big Data in the aggregate untouched, and some commentators have noted that the "versatility and power of re-identification algorithms imply that terms such as "personally identifiable" and "quasi-identifier" simply have no technical meaning" and, while "some attributes may be uniquely identifying on their own, *any attribute can be identifying in combination with others*."[38]

[25] ERNST &YOUNG, *supra* note 3.
[26] David W. Ichel, Peter J. Kahn & Theodore Edelman, *Current Approaches Taken in U.S. Litigation to Comply with Potentially Conflicting U.S. Discovery Obligations and EU and Other Foreign Data Privacy Statutes*, THE DUKE CONFERENCE (Nov. 2012), https://law.duke.edu/sites/default/files/images/centers/judicialstudies/Current%20Approaches.pdf.
[27] Narayanan & Shmatikov, *supra* note 17.
[28] *Metzger v. Am. Fidelity Assurance Co.*, No. CIV-05-1387-M, 2007 WL 3274921, at *1 (W.D. Okla. Oct. 23, 2007) (emphasis added).
[29] Stephen D. Feldman, *Practical Aspects of Privacy and Confidentiality in Litigation*, at 2, http://www.elliswinters.com/files/cle_manuscript.pdf.
[30] *Barnes v. Glennon*, No. 9:05-CV-0153, 2006 WL 2811821, at *5 n.6 (N.D.N.Y. Sep. 28, 2006).
[31] *Marks v. Global Mortgage Group Inc.*, 218 F.R.D. 492, 495-97 (S.D.W.V. 2003) (holding that the term "judicial process" includes civil discovery requests).
[32] McCallister et al., *supra* note 15.
[33] As currently required under F.R.C.P. 5.2(a), which require the redaction of social-security numbers, taxpayer-identification numbers, and financial-account numbers (except for last four digits); birth dates (except for the year of the individual's birth); and names of a minor (except for the minor's initials). F.R.C.P. 5.2 maintains limited exceptions associated with filings under seal.
[34] Feldman, *supra* note 29.
[35] McCallister et al., *supra* note 15.
[36] Ichel et al. *supra* note 26, at 12.
[37] *See* Feldman, *supra* note 29, at 6-15.
[38] Narayanan & Shmatikov, *supra* note 17 (emphasis original).

Technology Assisted Review and Privacy-Preserving Data Analysis in eDiscovery

While there has been significant discussion around the application of technology assisted review in eDiscovery,[39] much less attention has been paid to proposed methods by which modern technologies and big data analytics can preserve—or obliterate—privacy in the context of eDiscovery. Even less focus has been directed towards the growing concern of linkage attacks[40] where the processed datasets may still expose PII even if the manner by which it does so is not immediately evident to the entity collecting the data or even in instances where data collections have been otherwise sanitized.[41] This may be a concern for those large datasets used in legal, regulatory, or other investigative collections.

Accepting that these concerns may not have percolated into legal practitioners' considerations, within other areas of study, the "problem of privacy-preserving data analysis has a long history spanning multiple disciplines."[42] These techniques include "generalizing the data" or making it less precise, in some cases by grouping continuous values; "suppressing the data" by deleting entire records or certain parts of records; "introducing noise into the data" by adding small variations into selected data; "swapping the data" where the administrator exchanges certain data fields of one record with the same data fields of another similar record (e.g., swapping the ZIP codes of two records); and "replacing data with the average value" or replacing a selected value of data with the average value for the entire group of data.[43]

Despite the application of these techniques among privacy practitioners, there is still skepticism that these generalized techniques of "de-identifying" records with sensitive individual data "by removing or modifying PII" are nothing more than a whitewash of legitimate and unaddressed privacy concerns, and are "increasingly meaningless as the amount and variety of publicly available information about individuals grows exponentially."[44] There is also the issue of data integrity and the degree of reliance practitioners may affix to these data sets once these techniques have been utilized. This is especially true with respect to big data, where by "combining the use of these data sets with predictive analytics, Big Data can dramatically increase the amount of related data that may be considered private"[45] and the "process can predict highly intimate information, even if none of the individual pieces of data could be defined as PII."[46] DifP is another widely-used big data privacy preservation method. It is a method enabling analysts to extract useful answers from databases containing personal information while offering strong individual privacy protections.[47]

In order to address these concerns in the context of eDiscovery, we considered first whether it was possible to define a mathematically rigorous definition of privacy[48] and, in doing so, we also considered the use of AI and mathematical algorithms to automate data privacy information culling.[49,50,51] This investigation, which

[39] Parnell, *supra* note 6.
[40] Cynthia Dwork & Aaron Roth, *The Algorithmic Foundations of Differential Privacy*, FOUNDATIONS AND TRENDS® IN THEORETICAL COMPUTER SCIENCE, Vol. 9, Nos. 3-4 (2014) 211-407, at 218 (7), http://www.cis.upenn.edu/~aaroth/Papers/privacybook.pdf.
[41] Daniel Kifer, *Attacks on Privacy and deFinetti's Theorem*, SIGMOD '09 PROCEEDINGS OF THE 2009 ACM SIGMOD INTERNATIONAL CONFERENCE ON MANAGEMENT OF DATA, 127-138, at 127, http://www.cse.psu.edu/~dkifer/papers/definetti.pdf.
[42] See Dwork & Roth, *supra* note 40.
[43] McCallister et al., *supra* note 15.
[44] Narayanan & Shmatikov, *supra* note 17.
[45] Crawford, *supra* note 1, at 98.
[46] Crawford, *supra* note 1, at 101.
[47] Javier Salido, *Differential privacy for everyone*, White Paper, MICROSOFT CORPORATION (2012), http://www.microsoft.com/en-us/download/details.aspx?id=35409.
[48] Kifer & Machanavajjhala, supra note 9.
[49] Frank McSherry & Kunal Talwar, *Mechanism Design via Differential Privacy*, FOCS '07 PROCEEDINGS OF THE 48TH ANNUAL IEEE SYMPOSIUM ON FOUNDATIONS OF COMPUTER SCIENCE, 94-103, http://www.msr-waypoint.net/pubs/65075/mdviadp.pdf.

evaluated the methods by which the protection of PII in large datasets has been addressed in other disciplines, led us to consider the application of DifP.

A Consideration of Differential Privacy

DifP has been alternatively presented as the mechanism by which society may give "researchers access to vast repositories of personal data while meeting a high standard for privacy protection"[52] or as "a computationally rich class of algorithms" that satisfies a "robust, meaningful, and mathematically rigorous definition of privacy."[53] In application, it attempts to do "two important things at once…First, it defines a measure of privacy, or rather, a measure of disclosure—the opposite of privacy. And second, it allows data producers to set the bounds of how much disclosure they will allow"[54] in a given set of database queries. DifP in action, we discovered, is an attempt to address "the paradox of learning nothing about an individual while learning useful information about a population"[55] which may have implications in litigation that relies upon statistical analysis, such as "pattern and practice" employment discrimination class action cases.[56]

DifP operates according to a basic framework where the DifP algorithm employed operates to mask the value of any specific record within the data. When employed, and if the records are independent, changes to any specific record within the data will not materially impact the effect of the DifP algorithm's output as applied to a query, even if the viewer has access to both the output of the algorithm and the values of the rest of the records. However, if records are not independent, the viewer may determine the value of certain records given (again) the output of the algorithm and the values of the rest of the records.

The literature presented a number of additional considerations associated with the operation of DifP, including: (1) data cannot be fully anonymized and remain useful (this was reiterated time and time again); (2) the re-identification of anonymized records or *linkage attacks* are not the only risks;[57] (3) queries over large sets are not protective; (4) query auditing is problematic; (5) summary statistics are not "safe;" (6) there is an inherent danger in "ordinary facts;" and (7) not all datasets are "typical."[58] In contrast to the basic framework above, these additional considerations do not embody one particular algorithm that is a DifP operator; instead, DifP functions as "a mathematical guarantee that can be satisfied by an algorithm that releases statistical information on a data set. Many different algorithms satisfy the definition."[59]

One such DifP application we considered is "Pufferfish," a Framework for Mathematical Privacy Definitions.[60] We thought this framework might suffer in application to legal, regulatory, or investigatory collections because it requires a set of potential secrets, or "an explicit specification of what [the administrator] would like to protect;"[61] however, in application, each "secret" may, in fact, be the value of a

---

[50] Dwork & Roth, *supra* note 40.
[51] Bambauer et al., *supra* note 11.
[52] Klarreich, *supra* note 24.
[53] Dwork & Roth, *supra* note 40, at 222 (11).
[54] Bambauer et al., *supra* note 11, at 712.
[55] Dwork & Roth, *supra* note 40, at 216 (5).
[56] Allan G. King, *"Gross Statistical Disparities" as Evidence of a Pattern and Practice of Discrimination – Statistical versus Legal Significance*, 22 THE LABOR LAWYER 271 (2007), http://www.americanbar.org/content/dam/aba/publishing/lel_flash/LL_king.authcheckdam.pdf.
[57] Kifer, *supra* note 41.
[58] Dwork& Roth, *supra* note 40, at 217-220 (6-10).
[59] Cynthia Dwork, Frank McSherry, Kobbi Nissim & Adam Smith, *Differential Privacy – A Primer for the Perplexed*, Joint UNECE/Eurostat work session on statistical data confidentiality, WP. 26 (2011), http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2011/26_Dwork-Smith.pdf.
[60] Kifer & Machanavajjhala, *supra* note 9.
[61] *Id.*

given record, and Pufferfish operates to protect the value of every record. We considered an additional challenge to be (within this application and within a set of expectations centered on protections of privacy in the way people expect when records are independent) the authors' note that: "privacy definitions that can provide privacy guarantees without making any assumptions provide little utility beyond the default approach of releasing nothing at all."[62] This set of expectations is among the limitations of DifP application within the context of eDiscovery and related investigatory activity. Instead of viewing data points and records as singular instances brought together for statistical analysis, eDiscovery practices aim to measure "richness" of relatedness or relevance within the dataset, where interrelatedness is at least a by-product of a well-selected review set.[63]

This constraint joins the cautions provided by current literature, and at least one author argues that "differential privacy's strict and inflexible promises force a data producer to select from two choices: either to obliterate the data's utility or give up on the type of privacy that differential privacy promises."[64] In contrast, however, other academics note that DifP is still under development and also note that, in the application of Pufferfish and for the application of DifP generally, at this time its use requires an expert "to make assumptions explicit" such that the "domain expert needs to specify the potential secrets and discriminative pairs"[65] or determine which DifP algorithm may provide meaningful utility.

Not only does DifP require knowledge about the secrets it wishes to protect, its very operation lends itself *against* the "needle in a haystack" approach required in much (but not all) of eDiscovery and related investigatory activity.[66] In point of fact, "[f]or a query system to satisfy differential privacy, the system *must add noise* that ensures it only returns results such that the disclosure for everybody stays within certain predetermined bounds."[67] This highlights the difficulty of applying considerations of DifP within the databases that are geared towards the discovery of specific facts: "[a]ll database query systems serve the purpose of providing reasonably accurate information. Research results are the *raison d'être* for the query system in the first place. Inaccurate responses can be useless."[68]

We followed up our research by reaching out within the academic community associated with the algorithms that would support the application of DifP to determine if our considered application was valid. There were significant critiques levied against the application of DifP, with one author responding that current research recognizes that the use of "as is" outputs from DifP algorithms is not the best strategy, and additional statistical processing is needed to improve results.[69] Others voiced stronger concerns, stating that, while DifP was championed "as a practical solution to the competing interests in research and confidentiality" and poised for adoption as "the gold standard for data privacy," such adoption "would be a disastrous mistake."[70] But that could not be the end of the story; while DifP would clearly not solve all issues associated with unintentional disclosures of PII, perhaps it could appropriately address considerations of those class actions

[62] *Id.*
[63] Herbert L. Roitblat, *Measurement in eDiscovery*, White Paper, ORCATEC LLC (2013) at 2, http://www.theolp.org/Resources/Documents/Measurement%20in%20eDiscovery%20-%20Herb%20Roitblat.pdf.
[64] Bambauer et al., *supra* note 11, at 730.
[65] Kifer & Machanavajjhala, supra note 9.
[66] Discovery Subcommittee Advisory Committee on Civil Rules, *Notes of September 13, 2011 Conference Call*, at 3, http://www.uscourts.gov/uscourts/RulesAndPolicies/rules/Agenda%20Books/Civil/CV2011-11.pdf. ("The reality with litigation is that a lot of what is produced is not used in discovery, much less at trial. Only a small percentage of the information proves to be important. It is really discovery that we are doing, and once we discover something important we go from there. We have to look through the haystack to find the needle.").
[67] Bambauer et al., *supra* note 11, at 713.
[68] *Id.*, at 720.
[69] Michael Hay, Vibhor Rastogi, Gerome Miklau & Dan Suciu, *Boosting the Accuracy of Differentially-Private Histograms Through Consistency*, PROCEEDINGS OF THE VLDB ENDOWMENT, Vol. 3, No. 1 (2010), http://www.vldb.org/pvldb/vldb2010/papers/R91.pdf.
[70] Bambauer et al., *supra* note 11, at 701.

and other lawsuits that typically invoke statistical sampling noted above. Here, the literature cautioned users of DifP algorithms that the tool at hand must fit the purpose for which it is utilized; that is, "the limitations of a particular differentially private algorithm don't necessarily apply to all differentially private algorithms."[71] This is further evidence that DifP—and the algorithms that implement it as a concept—may be applicable, and even warranted, but appropriate implementation is necessary to confirm that it is operating as intended, and that there are no practitioner misunderstandings about what DifP does *not* do as part of its protections of PII within Collections.

Conclusion

DifP is not a stand-alone solution for the privacy considerations that attend eDiscovery and investigation big data sets. However, its consideration will absolutely add value, if carefully and appropriately applied, in those instances where issues associated with class actions and other lawsuits that typically invoke statistical sampling arise. In fact, DifP may be uniquely suited for application to just those instances, where the amount of data that would be at issue is so immense, by virtue of a putative class, to provide real insights into data without the time and effort required to first (attempt to) sanitize the data prior to the analysis. These efforts might benefit from the "auditor" type approach, where an administrator would "audit the sequence of queries and responses, with the goal of interdicting any response if, in light of the history, answering the current query would compromise privacy."[72] While some attacks related to the operational challenges of certain DifP algorithms sometimes center on unlimited queries against databases, this type of Collection would have very different purposes. Interrogatories and document requests are by their very nature limited; queries run against Collections, even those hosted in opposing or third-party available data rooms, could be limited in much the same way. In that instance, the auditor could examine (a) the suggested DifP algorithm(s); (b) the manner in which such DifP algorithm was utilized; and (c) the output prior to production or utilization by an opposing or third party.

As far as addressing PII concerns more generally, there seem to be no absolute technological solutions represented in the literature at this time, even though academics note that DifP is among the best available tools, and that "this line of investigation—in differential privacy and in no other approach to private data analysis" allows researchers "to maintain a quantitative measure of the cumulative privacy loss suffered by an individual in a given database."[73] However, a data-centric approach and use of analytics can pre-identify potential privacy issues beyond rule-based analysis and to (1) reduce collection/processing of non-relevant private information (and narrowing the scope of the discovery period[74]); and (2) better identify the private information or potentially private information in existing in datasets. These components of privacy by design[75,76] also underpin good data practices generally, where as a general rule, organizations "should minimize the use, collection, and retention of PII to what is strictly necessary to accomplish their business purpose and mission."[77]

The recognition that technology changes over time is also ensconced in advice regarding the application of big data generally, where "[p]olicies and regulation…should not embed particular technological solutions, but rather should be stated in terms of intended outcomes [to] avoid falling behind the technology."[78] This was a consistent refrain, with commentators noting that with Big Data concerns, "a flexible model based more on

[71] Dwork et al., *supra* note 59, at 5.
[72] Dwork & Roth, *supra* note 40, at 219 (8).
[73] Dwork et al., *supra* note 59, at 6.
[74] Ichel et al. *supra* note 26, at 12.
[75] Ann Cavoukian, David Stewart & Beth Dewitt, *Have it all - Protecting privacy in the age of analytics*, White Paper, DELOITTE (2014), http://www2.deloitte.com/content/dam/Deloitte/ca/Documents/Analytics/ca-en-analytics-ipc-big-data.pdf.
[76] Stewart, *supra* note 16.
[77] McCallister et al., *supra* note 15.
[78] President's Council of Advisors, *supra* note 21.

values and less on specific procedures will be more likely to endure over time."[79] Therefore, practitioners should consider DifP in applicable instances (of which there may be more in the future) but should start with the basic underpinnings of privacy by design at the inception of these types of projects and seek to limit the types of data collected.

Further, when designing privacy by design measures, practitioners should also actively consider incorporating considerations of protective orders, redaction, tokenization, anonymization, data "swaps" and suppression techniques, and other creative measures to shield PII. While these solutions seem to fall well short of perfection, in the event of a disclosure of PII, courts and regulatory bodies may look to the behavior of the practitioners and whatever demonstrable actions those practitioners took. In short, perfection should not be the enemy of good preventative measures, and an admonition to do one's best is the best takeaway the current literature provides.

[79] Crawford, *supra* note 1, at 118.