ARTICLES

THE DARK DATA QUANDARY

DANIEL J. GRIMM*

The digital universe remains a black box. Despite attaining high-technology capabilities like artificial intelligence and cognitive computing, "Big Data" analytics have failed to keep pace with surging data production. At the same time, the falling costs of cloud storage and distributed systems have made mass data storage cheaper and more accessible. These effects have produced a chasm between data that is stored and data that can be readily analyzed and understood. Enticed by the promise of extracting future value from rising data stockpiles, organizations now retain massive quantities of data that they cannot presently know or effectively manage. This rising sea of "dark data" now represents the vast majority of the digital universe.

Dark data presents a quandary for organizations and the judicial system. For organizations, the inability to know the contents of retained dark data produces invisible risk under a spreading patchwork of digital privacy and data governance laws, most notably in the medical and consumer protection areas. For courts increasingly confronted with Big Data-derived evidence, dark data may shield critical information from judicial view while embedding subjective influences within seemingly objective methods. To avoid obscuring organizational risk and producing erroneous outcomes in the courtroom, decision-makers must achieve a new awareness of dark data's presence and its ability to undermine Big Data's vaunted advantages.

---

* Adjunct Professor of Law, *Georgetown University Law Center*. Senior Trial Attorney, Division of Enforcement, U.S. Commodity Futures Trading Commission ("CFTC" or "Commission"). The Author conducted the research for and wrote this Article in his personal capacity and not in his official capacity as a CFTC employee. The analyses and conclusions expressed in this Article are those of the Author, and do not reflect the views of other Division of Enforcement employees, the CFTC staff, the Commission itself, or the United States government.

TABLE OF CONTENTS

INTRODUCTION

We are increasingly awash in data. So awash, in fact, that it no longer seems necessary to mention. The inevitability of a data-driven society powered by "Big Data"[1] analytics has long been a common mantra within the popular press, across industries, and among scholars.[2] The dawn of the Big Data era, fueled by a rate of data accumulation[3] that casts Moore's Law[4] for the quintupling of computing power in an antiquated glow, has bred optimism about its potential for seemingly infinite applications, including medicine, energy, financial markets, cybersecurity, and more recently, the law.[5]

---

1. "Big Data" is "the accumulation and analysis of unusually large datasets." Andrew Guthrie Ferguson, *Big Data and Predictive Reasonable Suspicion*, 163 U. PA. L. REV. 327, 352 (2015). In more specific terms, "Big Data" is a storage and analysis process designed to address the 'three Vs' of modern datasets: volume, variety, and velocity. *See id.* at 352–53; *see also* Andrew McAffee & Erik Brynjolfsson, *Big Data: The Management Revolution*, HARV. BUS. REV. (Oct. 2012), https://hbr.org/2012/10/big-data-the-management-revolution.

2. *See, e.g.*, Tom Breur, *Big Data and the Internet of Things*, 3 J. MARKETING ANALYTICS 1, 3 (2015) ("The second wave of Big Data growth, triggered by large-scale application of machine-to-machine traffic, is more like a tsunami than a wave. Unstoppable and irreversible."); *see also* Liran Einav & Jonathan Levin, *The Data Revolution and Economic Analysis*, 14 INNOVATION POL'Y & ECON. 1, 1 (2014) ("The media is full of reports about how big data will transform business, government, and other aspects of the economy."); Travis B. Murdoch & Allan S. Detsky, *The Inevitable Application of Big Data to Health Care*, 309 J. AM. MED. ASS'N 1351, 1352 (2013); Paul Ohm, *The Underwhelming Benefits of Big Data*, 161 U. PA. L. REV. ONLINE 339, 346 (2013) ("Big Data is coming, like it or not."); Steve Lohr, *Sizing up Big Data, Broadening Beyond the Internet*, N.Y. TIMES: BITS BLOG (June 19, 2013, 11:09 PM), http://www.cs.columbia.edu/igert/courses/E6898/Sizing_Up_Big_Data.pdf.

3. To put current rates of data accumulation into perspective, consider that "in 2016 we produced as much data as in the entire history of humankind through 2015." Dirk Helbing et al., *Will Democracy Survive Big Data and Artificial Intelligence?*, SCI. AM. (Feb. 25, 2017), https://www.scientificamerican.com/article/will-democracy-survive-big-data-and-artificial-intelligence.

4. For a detailed consideration of Moore's Law, see Robert R. Schaller, *Moore's Law: Past, Present, and Future*, IEEE SPECTRUM (1997), http://clifton.mech.northwestern.edu/~me381/papers/scalinglaw/moores-law.pdf.

5. *See, e.g.*, Caryn Devins et al., *The Law and Big Data*, 27 CORNELL J.L. & PUB. POL'Y 357, 362 (2017) (challenging "the widespread optimism about [Big Data's] potential uses in the legal system"); *see also* H.V. Jagadish et al., *Big Data and its Technical Challenges*, 57 COMM. ACM 86, 86 (2014) ("Big Data analysis now drives nearly every aspect of society, including mobile services, retail, manufacturing, financial services, life sciences, and physical sciences."); Michael Mattioli, *Disclosing Big Data*, 99 MINN. L. REV. 535, 540 (2014) ("Popular wisdom in technology circles holds that no avenue of human endeavor will not soon be touched and transformed by [Big Data]."); Neil M. Richards & Jonathan H. King, *Big Data Ethics*, 49 WAKE FOREST L. REV. 393, 393 (2014) (describing the "'Big Data' Revolution" as reaching "all kinds of human activities and

While there is no shortage of detractors,[6] Big Data advocates have promoted the belief that wealth buried within troves of data created by the Web, internet-enabled mobile devices, "smart" products, social media, and countless other sources can be readily mined and extracted.[7] The promise of hidden value has made mass data storage an organizational imperative. This digital hoarding is enabled by the falling costs of cloud systems and distributed computing solutions, which free data from the limitations of local storage.[8] All that is required to unlock the value of our amassed data, we are told, is the ability to identify obscured connections and patterns, which can be unearthed through advanced analytics that turn data into information,

---

decisions," including "medicine, education, voting, law enforcement, terrorism prevention, and cybersecurity").

6. For critiques of Big Data, see, for example, Devins et al., *supra* note 5; *see also* Kate Crawford & Jason Schultz, *Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms*, 55 B.C. L. REV. 93 (2014); Margaret Hu, *Big Data Blacklisting*, 67 FLA. L. REV. 1735 (2015); Ian Kerr & Jessica Earle, *Prediction, Preemption, Presumption: How Big Data Threatens Big Picture Privacy*, 66 STAN. L. REV. ONLINE 65 (2013); Jonas Lerman, *Big Data and Its Exclusions*, 66 STAN. L. REV. ONLINE 55 (2013); Neil M. Richards & Jonathan H. King, *Three Paradoxes of Big Data*, 66 STAN. L. REV. ONLINE 41 (2013); Nicholas P. Terry, *Protecting Patient Privacy in the Age of Big Data*, 81 UMKC L. REV. 385 (2012).

7. Many suggest we now live in an era of '"ubiquitous computing,'" in which the "world . . . is filled with 'intelligent' devices" that "make it possible to add a remarkable variety of intelligent functions to what were previously 'dumb' tools and appliances." PETER B. SEEL, DIGITAL UNIVERSE: THE GLOBAL TELECOMMUNICATION REVOLUTION 18–19 (2012); *see also* Steven M. Bellovin et al., *It's Too Complicated: How the Internet Upends* Katz, Smith, *and Electronic Surveillance Law*, 30 HARV. J.L. & TECH. 1, 9 (2016) ("Big Data collection and the ready availability of personal data—peoples' GPS locations, Facebook likes, etc.—are now pervasive, even ubiquitous sources of information . . . .").

8. *See, e.g.*, Tim Phillips, *How AI Can Untap the Dark Data Goldmine*, ECONOMIA (May 23, 2018, 10:26 AM), https://economia.icaew.com/features/may-2018/how-ai-can-untap-the-dark-data-goldmine (referring to the "plunging cost of cloud storage" as a basis for retaining data for potential future use); *see also* ERNST & YOUNG, BIG DATA: CHANGING THE WAY BUSINESSES COMPETE AND OPERATE 5 (2014) [hereinafter E&Y, Changing the Way], https://www.ey.com/Publication/vwLUAssets/EY_-_Big_data:_changing_the_way_businesses_operate/$FILE/EY-Insights-on-GRC-Big-data.pdf ("Cloud computing enables companies to use prebuilt big data solutions, or quickly build and deploy a powerful array of servers, without the substantial costs involved in owning physical hardware."). For a more involved discussion of cloud computing, see, for example, Primavera De Filippi & Miguel Said Vieira, *The Commodification of Information Commons: The Case of Cloud Computing*, 16 COLUM. SCI. & TECH. L. REV. 102 (2014).

and information into wealth.[9] Advances in artificial intelligence,[10] machine learning,[11] and cognitive computing[12] are increasingly viewed as the keys to unlocking this wealth.[13]

Yet, spinning bytes into gold is only the most obvious effect of the Big Data revolution. Big Data's growing prominence within our networked society reveals grander designs, including new modes of knowledge production steeped in perceptions of data-driven omnipotence and objectivity. Data scientists now derive conclusions "'born of the data' rather than 'born from theory,'" while proponents of a new empiricism have declared "the end of theory" altogether.[14]

Regardless of how the knowledge-production argument is ultimately resolved, there is little doubt that Big Data-driven decisions hold a

---

9. Such wealth is expected to be significant. *See, e.g.*, Angela Byers, *Big Data, Big Economic Impact?*, 10 I/S J.L. & POL'Y FOR INFO. SOC'Y 757, 764 (2015) ("Big data has the potential to create tens and perhaps hundreds of billions of value in many sectors . . . .").

10. Artificial intelligence has defied common definition. Matthew U. Scherer, *Regulating Artificial Intelligence Systems: Risks, Challenges, Competencies, and Strategies*, 29 HARV. J.L. & TECH. 353, 359 (2016) (contending that a practical definition of artificial intelligence is elusive because intelligence is often tied to human characteristics). For purposes of this Article, "artificial intelligence" refers to technologies that cause machines "to do tasks that would normally require human intelligence." Stefan van Duin & Naser Bakhshi, *Part 1: Artificial Intelligence Defined*, DELOITTE (Mar. 28, 2017), https://www.deloitte.com/fi/fi/pages/technology/articles/part1-artificial-intelligence-defined.html.

11. Machine learning is an artificial intelligence method which, "[b]roadly speaking, . . . involves computer algorithms that have the ability to 'learn' or improve in performance over time on some task." Harry Surden, *Machine Learning and Law*, 89 WASH. L. REV. 87, 88 (2014).

12. Cognitive computing systems are "a category of technologies that uses natural language processing and machine learning to enable people and machines to interact more naturally." Richard Cave, *How Cognitive Systems Will Make Personalized Learning a Reality*, IBM (May 4, 2016), https://www.ibm.com/blogs/watson/2016/05/cognitive-systems-will-make-personalized-learning-reality.

13. *See* Randy Bean, *How Big Data is Empowering AI and Machine Learning at Scale*, MIT SLOAN MGMT. REV. (May 8, 2017), https://sloanreview.mit.edu/article/how-big-data-is-empowering-ai-and-machine-learning-at-scale. Distributed ledger and blockchain technologies are likely to provide an additional avenue for extracting value from large datasets. *See, e.g.*, Jeremy Epstein, *When Blockchain Meets Big Data, the Payoff Will be Huge*, VENTUREBEAT (July 30, 2017, 12:10 PM), https://venturebeat.com/2017/07/30/when-blockchain-meets-big-data-the-payoff-will-be-huge (discussing the potential for blockchain technology to create valuable opportunities in advanced data analysis).

14. Rob Kitchin, *Big Data, New Epistemologies and Paradigm Shifts*, BIG DATA & SOC'Y 1, 3 (2014). Kitchin argues that the Big Data revolution has placed two groups in conflict: the new empiricists, who believe that data analytics can reveal truth without binding inquiry to hypotheses, and those who subscribe to a new brand of data-centric science that "seeks to hold to the tenets of the scientific method" by "generat[ing] hypotheses and insights 'born from the data' rather than 'born from the theory.'" *Id.* at 5–6.

certain appeal.[15] Decision-makers are drawn to Big Data-born decisions because they are believed to harness the power of high technology for enhanced precision and fact-inclusiveness, conjuring "the mythical omniscient actor from rational choice theory, accounting for all available information, probabilities of events, and potential costs and benefits."[16] Faith in all-knowing algorithms[17] has created a "more is better" narrative around the collection and processing of data, based on the assumption that more data supplies additional grist for the analytics mill, resulting in more accurate decisions than traditional human inquiry can produce.[18] Stated differently, in our networked Big Data society, "the smartest person in the room is the room."[19]

Big Data-driven decisions may also be viewed as qualitatively preferable to other decisions because data is assumed to be free from human bias.[20] Decisions resulting from Big Data analytics are often perceived to be inherently objective, as data-driven conclusions are assumed to be steeped in facts and evidence undisturbed by human interference.[21] Presumed objectivity has infused Big Data conclusions

---

15. *See, e.g.*, McAffee & Brynjolfsson, *supra* note 1 ("The evidence is clear: Data-driven decisions tend to be better decisions."); *see also* Matthew A. Waller & Stanley E. Fawcett, *Data Science, Predictive Analytics, and Big Data: A Revolution that Will Transform Supply Chain Design and Management*, 34 J. BUS. LOGISTICS 77, 77 (2013) ("Data are widely considered to be a driver of better decision making and improved profitability, and this perception has some data to back it up.").

16. Devins et al., *supra* note 5, at 362.

17. "The term 'algorithm' comes from computer science . . . and refers to an automatic rule that uses numerical inputs to produce some result . . . ." Angèle Christin et al., *Courts and Predictive Algorithms* 1 (2015), http://www.law.nyu.edu/sites/default/files/upload_documents/Angele%20Christin.pdf.

18. *See, e.g.*, Kalev Leetaru, *Does More Data Really Lead to Better Decision Making?*, FORBES (June 14, 2016, 9:13 PM), https://www.forbes.com/sites/kalevleetaru/2016/06/14/does-more-data-really-lead-to-better-decision-making ("There is a popular adage within many quarters of the 'big data' world that the more data you have, the more accurate your decision making will be.").

19. Mark Andrejevic, *The Big Data Divide*, 8 INT'L J. COMM. 1673, 1676 (2014) (quoting DAVID WEINBERGER, TOO BIG TO KNOW: RETHINKING KNOWLEDGE NOW THAT THE FACTS AREN'T THE FACTS, EXPERTS ARE EVERYWHERE, AND THE SMARTEST PERSON IN THE ROOM IS THE ROOM xiii (2011)).

20. *See, e.g.*, Allan G. King & Marko J. Mrkonich, *"Big Data" and the Risk of Employment Discrimination*, 68 OKLA. L. REV. 555, 555 (2016) ("Big Data utilizes methods that largely eliminate discretion, and unconscious bias, from the selection process.").

21. *See, e.g.*, *How is Big Data Analytics Transforming Corporate Decision-Making?*, ERNST & YOUNG (July 4, 2016), https://consulting.ey.com/how-is-big-data-analytics-transforming-corporate-decision-making ("Big data can transform how decision-makers view business problems and inform strategic decisions, allowing them to rely upon objective data.").

with a curious "prominence and status,"[22] rooted in the belief that data is "raw, objective, and neutral—the 'stuff of truth itself.'"[23]

The legal system is increasingly enticed by Big Data's promise of more fact-inclusive and objective decision-making. Big Data has established footholds across the legal landscape, not only in the emerging fields of data protection and cybersecurity, but also in traditional areas like criminal law.[24] Especially significant is the adaptation of Big Data in matters of evidence, which carries the potential to affect an endless array of cases.[25]

But for a society that is daily becoming more reliant on its data, we still know relatively little about "the oil of the digital era."[26] While data volumes grow exponentially, our ability to analyze and leverage the data we are creating has, perhaps surprisingly, lagged behind.[27] A key

---

22. Gernot Rieder & Judith Simon, *Datatrust: Or, the Political Quest for Numerical Evidence and the Epistemologies of Big Data*, BIG DATA & SOC'Y 1, 3 (2016) (quoting S. Leonelli, *What Difference Does Quantity Make? On the Epistemology of Big Data in Biology*, BIG DATA & SOC'Y 1, 2 (2014)).

23. *Id.* (quoting Lisa Gitelman & Virginia Jackson, *Introduction*, *in* "RAW DATA" IS AN OXYMORON 2 (Lisa Gitelman ed., 2013)).

24. *See* Devins et al., *supra* note 5, at 366 (describing how Big Data's predictive modeling already "has transformed areas of law ranging from financial regulation to pre-trial release and sentencing determinations in criminal cases").

25. *See id.*

26. *The World's Most Valuable Resource is No Longer Oil, but Data*, ECONOMIST (May 6, 2017), https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data.

27. *See, e.g.*, Leandro DalleMule & Thomas H. Davenport, *What's Your Data Strategy?*, HARV. BUS. REV. (May–June 2017), https://hbr.org/2017/05/whats-your-data-strategy? ("Cross-industry studies show that on average, less than half of an organization's structured data is actively used in making decisions—and less than 1% of its unstructured data is analyzed or used at all."); *see also* Justine Brown, *Data's Dark Side: What Can't Be Seen Can't Be Controlled*, BISCOM (July 14, 2016), https://www.biscom.com/datas-dark-side-cant-seen-cant-controlled (quoting Biscom CEO Bill Ho, who stated, "'While companies increasingly embrace data-driven strategies and decisions, the ability to analyze and use all the data is lagging behind the collection of data'"); John Gantz & David Reinsel, *The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East*, IDC VIEW 3 (2012), https://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf ("[W]hile the portion of the digital universe holding potential analytic value is growing, only a tiny fraction of territory has been explored."); Richard Harris, *More Data Will be Created in 2017 than the Previous 5,000 Years of Humanity*, APP DEVELOPER MAG. (Dec. 23, 2016), https://appdevelopermagazine.com/4773/2016/12/23/more-data-will-be-created-in-2017-than-the-previous-5,000-years-of-humanity (quoting Sencha CEO Art Landro as saying, "More data was created in the last two years than the previous 5,000 years of humanity . . . . Yet, recent research has found that less than 0.5 percent of that data is actually being analyzed for operational decision making"); Arif Mohamed, *Digital Transformation Makes Data Management Top Priority*, CIO (May 2, 2017, 5:54 PM), https://www.cio.com/article/3193717/it-industry/digital-

driver of the gulf between data accumulation and analytical insight is the phenomenon of "dark data." While no universal definition exists, early accounts of dark data in scientific literature describe it as information generated by failed experiments and not published or distributed, making it "nearly invisible" to the broader scientific community.[28] Descriptions of dark data within the digital world characterize it as data that is "hidden or undigested"[29] or "uncategorized, unmanaged, and unanalyzed."[30] While dark data is commonly unstructured[31]—often text-based, but not "analytics-ready"[32]—any data, in any form, can become dark.[33] Dark data is constantly being produced by organizations, the internet, personal mobile devices, and innumerable other sources.[34]

Far from representing a small slice of the digital universe,[35] dark data comprises the great majority of all existing data. By late 2017, it was

---

transformation-makes-data-management-top-priority.html ("Too many businesses lack a clear insight into the mountain of information they're sitting on.").

28. P. Bryan Heidorn, *Shedding Light on the Dark Data in the Long Tail of Science*, 57 LIBRARY TRENDS 280, 281 (2008); *see also* Adam R. Ferguson et al., *Big Data from Small Data: Data-Sharing in the 'Long Tail' of Neuroscience*, 17 NATURE NEUROSCIENCE 1442, 1443 (2014) (discussing the role of dark data in the neuroscience context).

29. Tracie Kambies et al., *Dark Analytics: Illuminating Opportunities Hidden Within Unstructured Data*, *in* TECH TRENDS 2017: THE KINETIC ENTERPRISE 21, 21 (Deloitte Univ. Press, 2017), https://www.deloitte.com/insights/us/en/focus/tech-trends/2017/dark-data-analyzing-unstructured-data.html.

30. ARMA Int'l, *Information at the Edge of Enterprise: Big Data . . . Dark Data . . . Your Data*, VIEWPOINTE (2013) [hereinafter Viewpointe White Paper], https://1pdf.net/big-datadark-datayour-data-viewpointe_58621c02e12e89e37fe76c10.

31. *See, e.g.*, Phillips, *supra* note 8 ("Dark data may often be unstructured . . . .").

32. Amir Gandomi & Murtaza Haider, *Beyond the Hype: Big Data Concepts, Methods, and Analytics*, 35 INT'L J. INFO. MGMT. 137, 137 (2015).

33. *See* IRON MOUNTAIN, DARK DATA TASK FORCE REPORT: IDENTIFICATION AND REMEDIATION OF DARK DATA IN LAW FIRMS 5 (2015), www.ironmountain.com/resources/whitepapers/d/dark-data-task-force-report-identification-and-remediation-of-dark-data-in-law-firms ("Dark data is largely unstructured, such as real-time communications and documents, but can also be semi-structured, for example XML code, or structured, as in a database.").

34. *See, e.g.*, Charles Babcock, *IBM Cognitive Colloquium Spotlights Uncovering Dark Data*, INFO. WEEK (Oct. 14, 2015, 10:05 AM), http://www.informationweek.com/cloud/software-as-a-service/ibm-cognitive-colloquium-spotlights-uncovering-dark-data/d/d-id/1322647 (describing "frenetic pace" of dark data production that is set to "multiply" through the Internet of Things); *see also* Irfan Khan, *Dark Data Tells Many Tales*, ITWORLD (Sept. 11, 2012), https://www.itworld.com/article/2720835/it-management/dark-data-tells-many-tales.html ("Every enterprise accumulates dark data. Companies don't try to hoard this unanalyzed information, it just happens because it's created almost everywhere.").

35. The term "digital universe" appears to have been coined by John Gantz and David Reinsel, who describe it as "a measure of all the digital data created, replicated, and consumed in a single year." Gantz & Reinsel, *supra* note 27, at 1.

widely estimated that eighty percent of existing data was dark, with that figure expected to reach ninety-three percent or higher by 2020.[36] At the same time, an exploding "internet of things"—rapidly becoming the "internet of everything"[37]—could, if upper-end estimates are believed, integrate up to 212 billion data-collecting devices by 2020.[38]

As with current data volumes, the bulk of the coming data tidal wave will be dark, causing our visibility into the digital universe to continually lag its expansion.[39] This divide between data volumes and data analysis exposes a paradox at the heart of the information age: the "information-data dichotomy," in which mass data creation makes it more difficult, rather than easier, to identify relevant information.[40] The "main culprit [is] dark data."[41]

The gap between data storage technologies and advanced analytics that can extract information and insights from dark data is a quandary for decision-makers. The most basic problem is that dark data can make a wide array of legal risks effectively invisible, resulting in miscalculation and error.[42] For example, how can an organization comply with laws governing the treatment of personally identifiable information (PII) unless it knows what PII it possesses? How can an

---

36. Babcock, *supra* note 34 (quoting remarks by John Kelly of IBM); *see also* Sanjay Srivastava, *Shedding the Light on Dark Data*, GENPACT BLOG (Nov. 1, 2017), http://www.genpact.com/insight/blog/shedding-the-light-on-dark-data ("It's estimated that unstructured data accounts for over 80 percent of all business data. And given trends in data proliferation, it's projected to grow to nearly 95 percent by 2020.").

37. *See, e.g.*, Tom Bajarin, *The Next Big Thing for Tech: The Internet of Everything*, TIME (Jan. 13, 2014), http://time.com/539/the-next-big-thing-for-tech-the-internet-of-everything (describing the "Internet of Everything" as referring to the trend of "adding connectivity and intelligence to just about every device in order to give them special functions").

38. *The Internet of Things is Poised to Change Everything, Says IDC*, BUS. WIRE (Oct. 3, 2013, 8:24 AM), http://www.businesswire.com/news/home/20131003005687/en/Internet-Poised-Change-IDC; *see also* Helbing et al., *supra* note 3 ("Everything will become intelligent; soon we will not only have smart phones, but also smart homes, smart factories and smart cities.").

39. Mika Javanainen, *Shedding Light on Dark Data in the IoT Era*, ENTERPRISETECH (Sept. 21, 2016), https://www.enterprisetech.com/2016/09/21/shedding-light-dark-data-iot-era.

40. ROCKET SOFTWARE, DATA VIRTUALIZATION: SHINE THE LIGHT ON DARK DATA 3 (2017), http://www.rocketsoftware.com/sites/default/files/resource_files/gartner_rdv_newsletter_final_041917_0.pdf.

41. *Id.*

42. *See, e.g.*, Christopher Bouton, *Pharma's Next Big Discovery: AI*, PHARMA LETTER (Sept. 5, 2018), https://www.thepharmaletter.com/article/pharma-s-next-big-discovery-ai (writing, in the pharmaceutical context, that because "as much as 90% of the data within a given organization is dark or siloed . . . companies are making important decisions with incomplete information").

organization adequately protect sensitive data from malicious cyber agents—and meet its obligations under mounting cybersecurity legal regimes—if it lacks insight into the content and location of its retained data? The present inability of advanced analytics to interpret mountainous quantities of retained dark data virtually ensures that regulated data will evade compliance systems, posing risks to organizations and individuals.

Dark data can also complicate judicial decision-making, which now commonly draws on Big Data processes like predictive analytics[43] and pattern recognition.[44] Criminal law in particular is replete with examples of dark data's ability to sabotage justice. The history of wrongful convictions resulting from absent or withheld evidence—effectively dark data at trial—portends the life-altering harm that can result when exonerating data remains dark to judges and juries.[45] The risk of harm is amplified in the Big Data era, as erroneous, data-driven conclusions may now acquire a patina of omnipotence and objectivity that insulates them from serious challenge.[46]

The concern is not that Big Data's supposed fact-inclusive objectivity cannot aid legal decision-making, it is that we often forget that "objectivity is [also] compatible with error: [a]n objective interpretation is not necessarily a correct one."[47] Nor is objectivity even guaranteed, as Big Data processes often remain stubbornly mired in the subjective framing they were designed to replace.[48] These flaws are

---

43. Predictive analytics "is the use of data, statistical algorithms and machine learning techniques to identify the likelihood of future outcomes based on historical data." *Predictive Analytics: What it is and Why it Matters*, SAS, https://www.sas.com/en_us/insights/analytics/predictive-analytics.html (last visited Feb. 5, 2019).

44. "[P]attern recognition is a branch of machine learning that emphasizes the recognition of data patterns or data regularities in a given scenario." *Pattern Recognition*, TECHOPEDIA, https://www.techopedia.com/definition/8802/pattern-recognition-computer-science (last visited Feb. 5, 2019). Pattern recognition "can be either 'supervised,' where previously known patterns can be found in a given data, or 'unsupervised,' where entirely new patterns are discovered." *Id.*

45. *See, e.g.*, Brandon L. Garrett & Peter J. Neufeld, *Invalid Forensic Science Testimony and Wrongful Convictions*, 95 VA. L. REV. 1, 76 (2009) (discussing wrongful convictions resulting from improperly-withheld exculpatory evidence).

46. *See* Devins et al., *supra* note 5, at 359–62, 371–72 (describing the "illusion of objectivity" that shields Big Data methods despite subjectivity embedded within algorithms).

47. Owen M. Fiss, *Objectivity and Interpretation*, 34 STAN. L. REV. 739, 747–48 (1982) (describing the authority maintained by an objective, interpretive legal rule, even if the interpretation is incorrect).

48. *See* Kitchin, *supra* note 14, at 5 ("[J]ust as data are not generated free from theory, neither can they simply speak for themselves free from human bias or framing . . . . Making sense of data is always framed—data are examined through a particular lens that influences how they are interpreted.").

often difficult to detect and harder to remediate. The gravest consequence of the dark data quandary is that factual determinations and, ultimately, legal decisions with implications for life and liberty may rest on unknowingly incomplete or erroneous data, but will achieve augmented credibility through Big Data's presumed omnipotence and objectivity.[49]

Two caveats are necessary before proceeding further. First, dark data's cryptic nature does not excuse fatalism. The response to newfound awareness that unknown, potentially risk-laden data may reside within an organization cannot be to continue ignoring or mismanaging that data. Nor should the challenge of assessing Big Data-derived evidence give way to acceptance without consideration of whether dark data may have distorted results. Instead, awareness of the dark data quandary should spur efforts to question and address the limitations of Big Data technologies.

Second, this Article does not dispute the value that Big Data adds to a host of applications, and the Author agrees that Big Data can be critiqued while also acknowledging that it will produce "significant, new, life-enhancing . . . benefits."[50] Rather, this Article offers a moderate proposal: decision-makers should be aware of the constraints dark data can place on organizational risk management, and on the accuracy and objectivity of Big Data-driven conclusions. Most importantly, decision-makers must avoid assuming the validity or completeness of Big Data-driven conclusions without deeper examination.

This Article proceeds in three parts. Part I provides a background discussion of data types and structures, which is necessary to understand the terminology used throughout this Article. Part II presents the invisible risk problem, which occurs when organizations retain dark data without the present ability to effectively analyze and interpret it. Central to the invisible risk problem is what this Article calls the "storage imperative": the drive to collect more and more data, even dark data that lacks a present use, out of the belief that analytical tools will ultimately unlock hidden value within that data.

Part II also illustrates specific ways that dark data can turn organizational risk invisible by considering medical privacy under the Health Insurance Portability and Accountability Act of 1996 and

---

49. *See* Devins et al., *supra* note 5, at 359–62. Big Data "models cannot measure or predict all of the relevant variables that may influence the legal system," and "Big Data cannot decide questions of meaning, equity and justice—though it risks doing so under the guise of objectivity, evidence and science." *Id.* at 359, 362.

50. Ohm, *supra* note 2, at 339–40 ("'Big Data' has become nearly synonymous with 'data analysis,' and data analysis is a lynchpin of modern science. To argue against Big Data is to argue against science. That is not my brief.").

consumer protection under section 5 of the Federal Trade Commission Act. Part III addresses dark data's implications for the use of Big Data evidence in the courtroom. Part III argues that dark data challenges the Big Data narratives of omnipotence and objectivity, and suggests that judges should consider the limitations dark data may place upon Big Data-derived evidence.

The purpose of this Article is to shed light on dark data, which has to date escaped attention in legal scholarship. This Article neither advocates specific data management efforts, nor recommends how judges should treat particular types of Big Data-derived evidence. Instead, this Article is a call for heightened vigilance in decision-making, and it will have done its job if it assists decision-makers in asking the right questions.

I. BACKGROUND

A. Structure and Non-Structure

Further discussion is premature without a working definition of "data." The term "data" is context-dependent, but commonly refers to recorded facts that can become information,[51] or is used more directly as a synonym for information.[52] Data is often defined as digital,[53] though it need not be.[54] This Article uses "data" to mean measurements or observations that often become, but are not

51. *See, e.g.*, Ilkka Tuomi, *Data Is More than Knowledge: Implications of the Reversed Knowledge Hierarchy for Knowledge Management and Organizational Memory*, 16 J. MGMT. INFO. SYS. 103, 105 (1999) ("Data have commonly been seen as simple facts that can be structured to become information.").

52. *See, e.g.*, *Data*, CAMBRIDGE DICTIONARY, https://dictionary.cambridge.org/us/dictionary/english/data (last visited Feb. 5, 2019) (defining data as "information collected for use"); *see also Data*, MERRIAM-WEBSTER DICTIONARY, https://www.merriam-webster.com/dictionary/data (last visited Feb. 5, 2019) (defining data as "factual information (such as measurements or statistics) used as a basis for reasoning, discussion, or calculation").

53. *See, e.g.*, *Data*, OXFORD ENGLISH DICTIONARY, http://www.oed.com/view/Entry/296948?rskey=vmQdkM&result=1&isAdvanced=false#eid (last visited Feb. 5, 2019) (defining data in the computing context as "[q]uantities, characters, or symbols on which operations are performed by a computer, considered collectively. Also (in non-technical contexts): information in digital form").

54. *See supra* notes 52–53.

inevitably, information.[55] From this definition, "data" can be divided into two main types: structured and unstructured.[56]

1. Structured data

While comprising a relatively small sliver of the digital universe, structured data has long been the focus of most organizations' data management efforts.[57] Structured data is data that has been organized in relational databases[58] and is readable in structured query language (SQL).[59] Common examples of structured data internal to an organization include financial records, human resources records, and

55. *See* Ashby Monk et al., *Data Management in Institutional Investing: A New Budgetary Approach* 3–4 (2017), https://ssrn.com/abstract=3014911 (defining data as "any recorded measurement or observation about the world" and explaining that "data becomes information when it is given sufficient context to be useful for decision-making").

56. Some have argued that what is often called "unstructured" data is really "semi-structured," as most data contains metadata that provides certain information about the data. *See, e.g.*, Drew Robb, *Semi-Structured Data*, DATAMATION (July 3, 2017), http://www.datamation.com/big-data/semi-structured-data.html. Nonetheless, this Article, like most literature on the topic, divides data into structured and unstructured varieties. *See id.* (explaining, "for the sake of simplicity, data is loosely split into structured and unstructured categories").

57. *See, e.g.*, Bryan Lapidus, *Structured Data vs. Unstructured Data for FP&A and Treasury*, ASS'N FIN. PROFS. (Mar. 6, 2017), https://www.afponline.org/trends-topics/topics/articles/Details/structured-data-vs.-unstructured-data-for-fp-a-and-treasury ("While structured data is estimated to be about 20 percent of current data, it is the main source of information that we in finance use and create."); *see also* ARVIND SATHI, ENGAGING CUSTOMERS USING BIG DATA: HOW MARKETING ANALYTICS ARE TRANSFORMING BUSINESS 109 (2014) (explaining that customer data management solutions have, "over the past decades, . . . been focused primarily on intraorganization sources of traditional 'structured' data").

58. *See, e.g.*, Ahmed Abbasi et al., *Big Data Research in Information Systems: Toward an Inclusive Research Agenda*, 17 J. ASS'N INFO. SYS. i, iii (2016); *see also* Malavika Jayanand et al., *Big Data Computing Strategies*, *in* BIG DATA: CONCEPTS, METHODOLOGIES, TOOLS, AND APPLICATIONS 793, 796 (Info. Resources Mgmt. Assoc. ed., 2016) ("Data that resides in a fixed field within a record or file is called structured data."); Stephen Kaisler et al., *Big Data: Issues and Challenges Moving Forward*, IEEE 995, 999 (2012) (describing structured data as possessing "well-defined data definitions (often in tables) as stored in relational databases").

59. For a brief explanation of SQL, see *SQL Tutorial*, W3RESOURCE, https://www.w3resource.com/sql/tutorials.php (last visited Feb. 5, 2019) (explaining that SQL "manag[es] data in relational database management system[s]").

the organization's Web data.[60] Common external structured data includes credit data, real estate records, and mobile phone data.[61]

Relational databases storing structured data rely on predefined schemas that sort data into rows, columns, and, ultimately, tables.[62] Organizing data into rigid schemas allows a relational database user to run SQL commands across the database, including executing queries for particular data.[63] SQL also allows a user to "join" data from different tables for compilation and comparison.[64]

The value of structured data is that its "standardized pieces . . . are identifiable and accessible by both humans and computers."[65] Organized and analytics-ready, structured data offers an ease of use that eludes unstructured data.[66] For example, a relational database might arrange data by name, date, time, and email address, allowing each component of an email message to be easily identified, retrieved, and analyzed.[67]

2. Unstructured data

Unstructured data is effectively the opposite of structured data;[68] that is, unstructured data does not "fit neatly into traditional structured

60. *See, e.g.*, Ramesh Nair & Andy Narayanan, *Benefiting from Big Data: Leveraging Unstructured Data Capabilities for Competitive Advantage*, BOOZ & CO. 3 (2012), https://www.strategyand.pwc.com/media/file/Strategyand_Benefiting-from-Big-Data.pdf.

61. *Id.* It is worth noting that describing data as being "external" to an organization depends on the organization being discussed. Credit reporting information may be external to most organizations, but it is internal to the credit reporting bureaus.

62. *See, e.g.*, Marc L. Berger & Vitalii Doban, *Big Data, Advanced Analytics and the Future of Comparative Effectiveness Research*, 3 J. COMP. EFFECTIVENESS RES. 167, 172 (2014) (describing relational databases as organizing data into "predefined data structures"); *see also A Relational Database Overview*, ORACLE, https://docs.oracle.com/javase/tutorial/jdbc/overview/database.html (last visited Feb. 5, 2019) (describing tabular data organization within relational databases).

63. *See* ORACLE, *supra* note 62. For example, a SQL "select" command will select and display specified information within the database. *Id.*

64. *See id.*

65. *What is Structured Data?*, U.S. SEC. & EXCH. COMM'N, https://www.sec.gov/structureddata/what-is-structured-data (last visited Feb. 5, 2019).

66. Pete Johnson, *Big Data, Dark Data, Unstructured Data—What Does It All Mean?*, AI FOUNDARY (Nov. 9, 2016), https://www.aifoundry.com/newsroom/blog/big-data-dark-data-unstructured-data-what-does-it-all-mean (writing that structured data is "usually much easier to understand" than unstructured data).

67. *See* Ray Bernard, *Big Data and Privacy for Physical Security*, SECURITY INDUS. ASS'N (Nov. 14, 2017), https://www.securityindustry.org/2017/11/14/big-data-and-privacy-for-physical-security.

68. *See, e.g.*, *Unstructured Data*, IOTONE, https://www.iotone.com/term/unstructured-data/t695 (last visited Feb. 5, 2019) ("[U]nstructured data usually refers

formats or databases."[69] Filling out a much larger portion of the digital universe than structured data,[70] unstructured data cannot be easily adapted to relational databases.[71] Examples of human-generated unstructured data include free-form text, such as the content of email messages and documents,[72] as well as non-textual data like photo, audio, and video files.[73] Unstructured data often includes "exhaust data," or data created as an unplanned "byproduct" of other, deliberate activity, such as Web use.[74]

While ideal for navigating relational databases, SQL contains significant limitations for managing unstructured data.[75] As schemas must be established before populating a relational database, unstructured data is often held in "not only structured query language" (NoSQL) non-relational databases, which replace tabular organization with a "data first, schema later" approach.[76] Freedom from schema

---

to information that doesn't reside in a traditional row-column database . . . . It's the opposite of structured data—the data stored in fields in a database.").

69. BERNARD MARR, DATA STRATEGY: HOW TO PROFIT FROM A WORLD OF BIG DATA, ANALYTICS AND THE INTERNET OF THINGS 89 (2017).

70. *See, e.g.*, Christine Taylor, *Structured vs. Unstructured Data*, DATAMATION (Mar. 28, 2018), http://www.datamation.com/big-data/structured-vs-unstructured-data.html ("[T]here is simply much more unstructured data than structured. Unstructured data makes up 80% and more of enterprise data, and is growing at the rate of 55% to 65% per year."); *see also* E&Y, Changing the Way, *supra* note 8, at 12 ("By some estimates, more than 80% of the data within organizations is unstructured and unfit for traditional processing."); Joe Mullich, *Harnessing the Potential of Unstructured Data*, WALL ST. J. (Dec. 10, 2012), http://online.wsj.com/ad/article/datamanagement-harness; Nair & Narayanan, *supra* note 60, at 5.

71. *See, e.g.*, April Reeve, *Big Data and NoSQL: The Problem with Relational Databases*, DELL EMC (Sept. 7, 2012), https://infocus.dellemc.com/april_reeve/big-data-and-nosql-the-problem-with-relational-databases (describing limitations of relational databases in managing unstructured data).

72. Pierre Dorion, *What is Unstructured Data and How is it Different from Structured Data in the Enterprise?*, TECHTARGET, https://searchstorage.techtarget.com/feature/What-is-unstructured-data-and-how-is-it-different-from-structured-data-in-the-enterprise (last visited Feb. 5, 2019) (suggesting that although emails and documents are organized in a database, such as Microsoft Exchange, the body of the message is freeform text); *see also* Richard Stiennon, *Are Dark & Unstructured Data Putting Your Business at Risk?*, BLANCCO TECH. GRP. (Mar. 23, 2017), https://www.blancco.com/blog-dark-unstructured-data-business-risk ("Unstructured data often appears in the form of e-mails, memos, chats, white papers, marketing materials, images, presentations, and video files.").

73. *See* Gandomi & Haider, *supra* note 32, at 137; Taylor, *supra* note 70.

74. Terry, *supra* note 6, at 389–90; *see also* Reeve, *supra* note 71 (stating that "[r]elational databases . . . don't scale well to very large sizes" and "don't do unstructured data search very well . . . .").

75. Berger & Doban, *supra* note 62, at 172.

76. *SQL vs. NoSQL Databases: What's the Difference?*, UPWORK, https://www.upwork.com/hiring/data/sql-vs-nosql-databases-whats-the-difference (last

allows many NoSQL databases to store enormous volumes of data, especially when integrated with cloud and distributed systems that maximize scalability.[77] Greater storage capacity flows logically from the nature of non-relational databases: by avoiding the need to initially organize data by schema, a database can store it in much greater volumes.[78] Depending on the volume involved, unstructured data may be stored in giant non-relational databases called "data lakes."[79] The ability to handle large volumes of schema-less data has made NoSQL databases key components of Big Data architecture, with major social media companies using them to power huge volumes of free-form text, picture, and video files.[80]

External unstructured data, such as customer social media posts, can be particularly valuable to organizations because they can reveal insights that may be otherwise unavailable. As one commentator remarked: "Your customers have plenty to say about you and the industry you serve, but they probably don't say it in the same language your database speaks."[81] Put another way, extracting value from external unstructured data requires work, which is increasingly performed by Big Data analytics that apply artificial intelligence and related high-technology methods to vast datasets.[82]

B. Darkness and Light

The concepts of structured and unstructured data provide a foundation for discussing dark data. At base, dark data is data that is

---

visited Feb. 5, 2019); *see also* SNOWFLAKE COMPUTING, INC., FAST, EFFICIENT PROCESSING OF SEMI-STRUCTURED DATA 2 (2015), https://www.snowflake.net/wp-content/uploads/2015/06/Snowflake_Semistructured_Data_WP_1_0_062015.pdf ("[S]tructured data requires a fixed schema defined in advance.").

77. DINO ESPOSITO & ANDREA SALTARELLO, MICROSOFT.NET: ARCHITECTING APPLICATIONS FOR THE ENTERPRISE 370 (2014).

78. *See* UPWORK, *supra* note 76.

79. Mandy Chessell et al., *Dive into Analytics and the Data Lake*, IBM, https://developer.ibm.com/tv/developers-and-data-lake-analytics (last updated on Feb. 9, 2018) ("A data lake is a storage repository that holds an enormous amount of raw [or refined] data in native format until it is accessed.").

80. Berger & Doban, *supra* note 62, at 172 ("No SQL databases power many of the largest websites that contain large amounts of unstructured data," including Facebook and Twitter).

81. Joe Hewitson, *What's Your Data Strategy Missing?*, IBM (Apr. 13, 2017), https://www.ibm.com/information-technology/whats-your-data-strategy-missing.

82. *See, e.g.*, Blair Hanley Frank, *IBM Declares AI the Key to Making Unstructured Data Useful*, VENTUREBEAT (July 11, 2017, 5:01 PM), https://venturebeat.com/2017/07/11/ibm-declares-ai-the-key-to-making-unstructured-data-useful.

unknown or unused,[83] usually because it has been collected but not analyzed.[84] Dark data is often described as "information assets [that] organizations collect, process and store during regular business activities, but generally fail to use for other purposes."[85] Equating dark data to "information assets" suggests that value is present within the unanalyzed data lying dormant in server logs and cloud accounts.

As the concept of "dark data" refers to whether data is known or visible rather than to a structural quality, dark data can be understood to occupy one end of a three-point "visibility spectrum."[86] The visibility spectrum connects structured, unstructured, and semi-structured data to, respectively, the qualities of "light," "dark," and "grey."[87] Structured data, arranged in relational databases with rigid schema, is the most visible, or "light," while unstructured data's defiance of analytics-readiness makes it the least visible, or "dark."[88] Semi-structured data is not fully structured, but it carries metadata that lends it partial, or "grey," visibility.[89]

The visibility spectrum is intuitive, as unstructured data is more likely to be dark, while structured data stored within a relational database will almost always be light.[90] Still, while structure plays an important role in determining visibility, "any data could become dark depending on the way the business . . . uses it."[91] The following subsections explore various ways that data can become or remain dark.

---

83. *See* Saurabh Sharma, *Shedding Light on Dark Data*, CIO (May 27, 2015, 7:56 AM), https://www.cio.com/article/2926089/data-analytics/shedding-light-on-dark-data.html.

84. *See, e.g.*, *Dark Data*, CONPERIO TECH. SOLUTIONS, https://conperio.com/dark-data (last visited Feb. 5, 2019) (remarking that "dark data has come especially to denote operational data that is left unanalyzed").

85. *IT Glossary*, GARTNER, http://www.gartner.com/it-glossary/dark-data (last visited Feb. 5, 2019).

86. The term "visibility spectrum" appears to have been developed in a white paper by Indus Valley Partners. *See* Tom Coughlin, *Analysis of Dark Data Provides Market Advantages*, FORBES (July 24, 2017, 10:20 PM), https://www.forbes.com/sites/tomcoughlin/2017/07/24/analysis-of-dark-data-provides-market-advantages.

87. *See id.*

88. *Id.*

89. *See id.*; *see also* Robb, *supra* note 56 (describing metadata within semi-structured data).

90. *See* IRON MOUNTAIN, *supra* note 33, at 5 ("Dark data is largely unstructured, such as real-time communications and documents . . . ."); *see also* Johnson, *supra* note 66 ("Because there are components to the document such as field names and descriptions, structured data tends to be less dark.").

91. Bob Laurent, *What is Dark Data?*, BETANEWS, https://betanews.com/2017/04/19/what-is-dark-data (last visited Feb. 5, 2019); *see also* Arvind Purushothaman, *Time Has Come for Enterprises to Mine Dark Data for Decision Making*, TECH OBSERVER (May 13, 2018, 12:03 AM), https://techobserver.in/article/ opinion/time-has-come-for-enterprise-to-mine-dark-data-for-decision-making-arvind-purushothaman-virtusa ("Dark data can be both structured and unstructured.").

1. Default data

Data creation has become a default consequence of nearly every activity in modern life.[92] Of the swelling quantity of dark data being amassed today, most is "not consciously collected," but is the byproduct of other, deliberate activity.[93] As mobile devices, social media applications, audit programs, and other networked entities perform their tasks, they automatically create data byproducts that may go unanalyzed and remain dark.[94] As knowingly creating structured data also creates extraneous and unknown data that is retained but not analyzed,[95] unmanaged volumes of dark data will likely continue to outstrip our ability to analyze and understand the data we are creating.[96]

2. Situational effects

Dark data may also result from accidental or unplanned situational effects. For example, an employee who, unknown to an organization, copies data onto a laptop or a flash drive and removes it from the organization has created dark data as far as the organization is concerned. While the data may be structured and light where it resides on the organization's servers, the unauthorized copy may be unknown to the organization and beyond its control, making it dark to the organization.[97]

---

92. Ganesh Moorthy, *Dark Data: The Two Sides of the Same Coin*, ANALYTICS, http://analytics-magazine.org/dark-data-two-sides-coin (last visited Feb. 5, 2019) (stating that "[d]ata generation is a default" in modern life); *see also* Anita L. Allen, *Protecting One's Own Privacy in a Big Data Economy*, 130 HARV. L. REV. F. 71, 71 (2016); Neil M. Richards, *The Dangers of Surveillance*, 126 HARV. L. REV. 1934, 1940 (2013) (explaining that the rise of the Internet of Things will "subject[] more and more previously unobservable activity to electronic measurement, observation, and control"); Babcock, *supra* note 34; Khan, *supra* note 34.

93. Bob Laurent, *What is Dark Data? Alteryx Shines a Light*, ITPROPORTAL (Apr. 19, 2017), http://www.itproportal.com/features/what-is-dark-data-alterxy-shines-a-light; *see also* Terry, *supra* note 6, at 389–90.

94. Babcock, *supra* note 34.

95. *See Digging into Dark Data Can Reap Benefits*, ANNALECT (Sept. 7, 2017), https://www.annalect.com/digging-into-dark-data-can-reap-benefits ("When every interaction, transaction, and engagement gets captured, brands must prioritize what gets immediately utilized and what gets pushed to the wayside . . . . This often means un- or semi-structured data . . . is left to hang out in the archive 'just in case.'").

96. *See* Babcock, *supra* note 34; *see also* Khan, *supra* note 34 ("Organizations simply generate far more data than they can currently exploit.").

97. *See* Sharon D. Nelson & John W. Simek, *Cybersecurity Basics*, 88 OKLA. B.J. 1549, 1553 (2017) (explaining that employees can create dark data by stealing data "or leav[ing] it on flash drives, their home devices, etc.").

Situational effects may occur on a larger scale, such as system failures or hacking events that result in the destruction, theft, or locking[98] of data. Curiously, while data destruction and data theft can both create dark data, they do so in opposite ways. The former erases previously-known data, rendering it dark unless it has been backed up, while the latter creates additional, unknown copies of data that an organization cannot control.

3. Organizational effects

Dark data is also created, often unknowingly, by organizational effects that turn otherwise usable light data into dark data.[99] A common problem especially among large organizations is the "siloing" of data into isolated or poorly-integrated repositories.[100] If Office A creates data relevant to Office B's mission, but does not share the data with Office B, the data is effectively dark as to Office B.[101] In this example, the data is at once both light and dark within the same organization, and could become wholly light if Offices A and B integrate their separate data silos.

4. Deliberate creation

Still more data is intentionally left dark. Organizations subject to resource constraints that require balancing data-management priorities may choose to focus their data management efforts on high-priority data while intentionally leaving other data dark. Resource management, budgeting, and organizational objectives can all impact decisions about which data must be presently analyzed, and which may be warehoused for later consideration.

---

98. An example of data locking is found in ransomware, which "allows wrongdoers to control, damage, and interrupt systems; deny access to data; and destroy or otherwise harm the data's integrity—all *without* actual acquisition of the data." James A. Scherer et al., *Ransomware—Practical and Legal Considerations for Confronting the New Economic Engine of the Dark Web*, 23 RICH. J.L. & TECH. 1, 22 (2017).

99. *See, e.g.*, McKinsey & Co., *The Age of Analytics: Competing in a Data-Driven World*, MCKINSEY GLOBAL INST. 3 (2016), https://www.mckinsey.com/~/media/mckinsey/business%20functions/mckinsey%20analytics/our%20insights/the%20age%20of%20analytics%20competing%20in%20a%20data%20driven%20world/mgi-the-age-of-analytics-full-report.ashx ("Many incumbents struggle with switching from legacy data systems to a more nimble and flexible architecture to store and harness big data.").

100. *See* Sharma, *supra* note 83 ("Because collected data sits in separate silos, it is often difficult to systematically bring it together to produce a clear, cohesive picture. This is especially true for companies with legacy IT systems . . . .").

101. *See* David Greenfield, *OSIsoft Shines a Light on Dark Data*, AUTOMATIONWORLD (May 1, 2018), https://www.automationworld.com/article/industry-type/all/osisoft-shines-light-dark-data (explaining that "dark data" can "refer to data that is being collected, but gathered in silos so that it is not aggregated with other data").

II. INVISIBLE RISK

The primary unintended consequence of hoarding dark data is the creation of invisible risk. Organizations presently lacking the means to analyze inflowing dark data have taken to storing it for later use, motivated by the view that advanced analytics will ultimately unlock hidden value within data repositories.[102] But for all its promise, dark data can also be a vector of significant risk: when an organization cannot readily identify or interpret its data, it similarly cannot identify and manage risks concealed within that data.

Data-born risk is atomized across the data lifecycle, as collection, use, disclosure, and destruction of data can all carry distinct legal risks from regulatory agencies and private plaintiffs.[103] For instance, if an organization cannot identify or analyze sensitive medical data residing within its retained dark data, it may be unable to comply with its obligations governing the protection, use, and disclosure of that data. The risks of this de facto invisibility are growing, as a patchwork of federal, state, and international data governance laws has steadily expanded over the last several decades.[104] Dark data can frustrate efforts to comply with these laws.

In addition to posing challenges under existing legal frameworks, dark data creates broader risks under emerging cybersecurity legal regimes. Organizational resource allocation and cybersecurity priorities can indirectly magnify the appeal of dark data to bad actors, including rogue employees and external cybercriminals.[105] Unlike light, structured data, dark data is not used for presently-defined

---

102. Alex Woodie, *The Growing Menace of Data Hoarding*, DATANAMI (June 13, 2016), https://www.datanami.com/2016/06/13/growing-menace-data-hoarding ("Instead of storing only data that has a proven business value, companies are now storing any piece of data that has a remote chance of providing value in the future.").

103. *See* Javier Salido & Doug Cavit, *Trustworthy Computing: A Guide to Data Governance for Privacy, Confidentiality, and Compliance*, MICROSOFT 1 (2010), http://mscorp.indsyntest.com/perspective/pdf/sec-Data_Governance_-_Moving_to_Cloud_Computing.pdf ("Organizations that want to move confidential data to the cloud should systematically identify incremental risks to data privacy and security in the information lifecycle . . . .").

104. *See, e.g.*, Ieuan Jolly, *Data Protection in the United States: Overview*, THOMSON REUTERS PRAC. LAW (July 1, 2017), https://content.next.westlaw.com/Document/I02064fbd1cb611e38578f7ccc38dcbee/View/FullText.html?contextData=(sc.Default)&transitionType=Default&firstPage=true&bhcp=1 (describing growth in state and federal data protection laws).

105. Johan Holder, *The True Cost of Unstructured 'Dark Data' in the GDPR Era*, COMPUTING (Aug. 31, 2017), https://www.computing.co.uk/ctg/opinion/3016440/the-true-cost-of-unstructured-dark-data-in-the-gdpr-era; *see also infra* note 235 and accompanying text.

purposes, which can make it a low priority for data protection and cybersecurity measures.[106] This neglect can make dark data an attractive target for unauthorized access and theft, exposing organizations to significant legal, financial, and reputational risk.[107]

A. The Storage Imperative

The core of the dark data quandary for organizations is the present technological divide between data storage systems and Big Data analytical tools. While Big Data analytics are constantly evolving, fueled by rapid progress in predictive methods and artificial intelligence, the capabilities of existing tools have yet to scale with an expanding digital universe.[108] In lamenting the inability of data analytics to keep pace with soaring data volumes, one commentator wrote: "[W]e are simply not equipped to deal with this constant deluge of data. Compounding this effect is the fact that most of this unanalyzed data is unstructured."[109]

But while the flood of data may defy large-scale analysis, it can, in relative terms, be stored.[110] As data volumes have risen, developments

106. *See* Holder, *supra* note 105; *see also infra* note 235 and accompanying text.

107. *See, e.g.*, Jeff John Roberts, *A Surprise in the Equifax Breach: Victims Likely to Get Paid*, FORTUNE (Oct. 10, 2017), http://fortune.com/2017/10/10/equifax-class-action (explaining that Equifax will likely have to pay more than $1 billion to consumers who were harmed by its 2017 data breach); *see also Implications and Consequences of a Data Breach on a Business*, CYPRESS DATA DEF. (2017), https://www.cypressdatadefense.com/security-assessments/why-security-testing-is-important/implications-and-consequences-of-a-data-breach-on-a-business (last visited Feb. 5, 2019) ("Failing to uphold proper information security standards . . . may result in a significant loss of revenue due to an increased negative sentiment from customers who were affected, and potential customers who choose to put their trust in another company.").

108. *See* Gantz & Reinsel, *supra* note 27.

109. Moorthy, *supra* note 92; *see also* Barclay Blair, *The Total Cost of Owning Unstructured Information: Decoding Information Governance, Big Data and eDiscovery*, NUIX WHITEPAPER 3 (2012), http://docplayer.net/1375017-White-paper-the-total-cost-of-owning-unstructured-information-about-the-author-decoding-information-governanc e-big-data-and-ediscovery.html ("Data volumes are growing, but *unstructured information* is growing faster than our ability to manage it."); Thomas H. Davenport et al., *How 'Big Data' is Different*, MIT SLOAN MGMT. REV. (July 30, 2012), https://sloanreview.mit.edu/article/how-big-data-is-different ("There is no question that organizations are swimming in an expanding sea of data that is either too voluminous or too unstructured to be managed and analyzed through traditional means.").

110. *See, e.g.*, E&Y, Changing the Way, *supra* note 8, at 5. One explanation for the growth in data storage is Kryder's Law, which posits that computing storage density is increasing in a manner similar to the growth of microprocessor capacity described by Moore's Law. *See* RUSSELL WALKER, FROM BIG DATA TO BIG PROFITS: SUCCESS WITH DATA AND ANALYTICS 10 (2015) ("Mark Kryder, the former CTO of Seagate, observed that data storage capacity has been increasing in time and that the cost of that storage has

in cloud computing and distributed systems[111] have displaced local, "on premises" data storage, opening up new and cheaper options for warehousing huge quantities of data.[112] Transitioning away from local data storage has also brought cost savings, making it "easy to store dark data and not think about it."[113] The consequence of these trends is that individuals and organizations "have all become hoarders," retaining "immense, previously unfathomable amounts of data simply because they *can*."[114]

Aided by rising storage capabilities and enticed by the Big Data narrative that nearly all data holds value waiting to be unlocked,[115] mass

---

been decreasing in time. These realities mean that producing and storing more data has been constantly easier over time.").

111. A noteworthy example is edge computing, which reduces latency and can offer real-time data analysis by performing operations at the "edge" of a network rather than by making contact with centralized cloud servers. Bob O'Donnell, *Edge Computing is Reshaping the Cloud*, WESTERN DIGITAL BLOG (June 13, 2018, 1:24 PM), https://blog.westerndigital.com/edge-computing-reshaping-cloud.

112. *See* Christine Hall, *Survey: On-Prem Data Centers Lowest Investment Priority for IT Shops*, DATACENTER KNOWLEDGE, (Aug. 22, 2017), https://www.datacenterknowledge.com/business/survey-prem-data-centers-lowest-investment-priority-it-shops ("[D]ata centers now have the lowest priority for new spending among a list of five categories," which a recent study "attributes to increasing reliance on cloud infrastructure, cloud storage, and [software as a service]."); *see also* E&Y, Changing the Way, *supra* note 8, at 5 ("Cloud computing enables companies to use prebuilt big data solutions . . . without the substantial costs involved in owning physical hardware.").

113. Scott Etkin, *Don't Be Spooked by Dark Data*, DATA INFORMED (Oct. 30, 2015, 5:30 AM), http://data-informed.com/dont-be-spooked-by-dark-data (quoting an interview with Peter Vescuso, Chief Marketing Officer of VoltDN) (on file with the American University Law Review).

114. Omer Tene & Jules Polonetsky, *A Theory of Creepy: Technology, Privacy and Shifting Social Norms*, 16 YALE J.L. & TECH. 59, 84 (2013) [hereinafter Tene & Polonetsky, *Social Norms*]; *see also* Joshua Klein, *Why Hoarding Your Data is Hurting Your Business*, INC. (July 10, 2017), https://www.inc.com/joshua-klein/why-hoarding-your-data-is-hurting-your-business.html (finding that organizations have become "compulsive data hoarders").

115. *See, e.g.*, Viktor Mayer-Schönberger & Yann Padova, *Regime Change?: Enabling Big Data Through Europe's New Data Protection Regulation*, 17 COLUM. SCI. & TECH. L. REV. 315, 319–20 (2016); *see also* Kambies et al., *supra* note 29, at 23 (stating that unstructured dark data that has been left untapped due to technical constraints may contain "valuable information on pricing, customer behavior, and competitors"). Kambies et al. also quote Greg Powers, Vice President of Technology at Halliburton, as saying the following about dark data: "[T]here's so much potential value buried in this darkness that I flip the frame and refer to it as 'bright data' that we have yet to tap." Kambies et al., *supra* note 29, at 27; *see also* Andrea Peterson, *Companies Have More Data than Ever. That's Risky*, WASH. POST (Jan. 7, 2015), https://www.washingtonpost.com/news/the-switch/wp/2015/01/07/companies-have-more-data-than-ever-thats-risky (explaining that technology companies that have not determined how to

data collection and storage has become an organizational imperative. Organizations are collecting and storing more data, most of it unstructured and dark, than at any other time in human history.[116] In an age where it has become theoretically possible to store all the world's data on DNA strands, we can expect technology to continually test the limits of data storage.[117] Many organizations already store nearly all data they encounter.[118]

Modern society is fueling the storage imperative by repricing data's value as an asset. While once valuable in relation to a known, discrete objective, data is now often assumed to possess high, if indeterminate, latent value.[119] In this estimation, data's value "can only be fully reaped as the data is . . . reused over and over again for different purposes," which may not be known at the time of collection.[120] The search for latent value "creates a very strong economic incentive" to collect data of unknown present value and "to keep the data for as long as possible."[121]

Organizations have responded to data's future value proposition and falling storage costs by hoarding huge quantities of data that they cannot presently interpret or understand.[122] As one commentator explained,

---

monetize collected data have often "decide[d] to hoard data under the assumption that it may be useful to them someday down the line, even if they haven't figured out how yet").

116. Anna Berge, *Adequacy in Documentation*, *in* LANGUAGE DOCUMENTATION: PRACTICES AND VALUES 51, 64 (Lenore A. Grenoble & N. Louanna Furbee eds., 2010) ("[D]ocumentation is greatly helped by new advances in technology, which allow us not only to document more but also to store and make accessible more data than ever before. What we expect out of documentation efforts is far greater than at any previous time . . . ."); Evelyn Kotler, *Stop Neglecting Your Dark Data*, ITPROPORTAL (Jan. 25, 2016), https://www.itproportal.com/2016/01/25/stop-neglecting-your-dark-data ("[W]e are amassing and storing more data than ever before."); *see also* Harris, *supra* note 27 (quoting Sencha CEO Art Landro).

117. Mike McRae, *Microsoft Plans on Storing Its Data on DNA in The Next 3 Years*, SCI. ALERT (May 27, 2017), https://www.sciencealert.com/microsoft-could-be-storing-data-on-dna-within-the-next-three-years (describing growing potential of DNA as a data storage medium).

118. Guy Betar, *Shining a Light on Dark Data*, CIO (Nov. 4, 2015, 11:17 AM), https://www.cio.com.au/article/588167/shining-light-dark-data (explaining that "[w]ith the expansion of digital storage capacity, and a corresponding reduction in the cost of such storage," organizations have developed a tendency to "store almost all data that [they] create[] or collect[]"); *see also* Stiennon, *supra* note 72 (writing that "[m]ost organizations" keep unstructured data from a variety of sources "without a plan for disposing of it").

119. Mayer-Schönberger & Padova, *supra* note 115, at 319–20.

120. *Id.*

121. *Id.* at 320.

122. *See* DalleMule & Davenport, *supra* note 27 ("[L]ess than half of an organization's structured data is actively used in making decisions—and less than 1% of its unstructured data is analyzed or used at all.").

"the plunging cost of cloud storage means you don't need [to] throw away your dark data unless you are sure it is too inconsistent or incomplete to be useful in the future."[123] Another asserted, "'[d]ark data has virtually limitless value.'"[124] This view that nearly all data is or will become valuable is the engine driving the Big Data machine: stockpiling data becomes an economical means of securing future wealth, which will be captured when analytical tools catch up to the data lakes.[125]

In contrast to data storage technologies, advanced analytical tools that can accurately interpret large quantities of dark data are costly and nascent, often drawing from emerging fields like artificial intelligence, machine learning, and cognitive computing. Acquiring and effectively implementing systems that use these technologies is neither easy nor cheap.[126] Moreover, organizations that do implement bleeding-edge Big Data systems may learn that artificial intelligence algorithms do not possess the mythical qualities often attributed to them. As one technology executive explained, "many of today's [artificial intelligence] algorithms are static," requiring reprogramming to account for "new sensors, new users, and new data streams."[127] Extracting insight from a rising sea of dark data is not automatic and demands "a new evolution of [artificial intelligence] that can adapt to a rapidly changing world."[128]

Large-scale data storage is relatively cheap, is becoming cheaper, and is subject to limited (and falling) technical barriers.[129] In contrast,

---

123. Phillips, *supra* note 8.

124. *See* Brown, *supra* note 27 (quoting Brad Anderson, Vice President of Big Data Informatics at Liaison Technologies); *see also* Kambies et al., *supra* note 29, at 22 (claiming that harvesting dark data "could prove to be something akin to a lottery jackpot"); Bob Laurent, *What Awaits Discovery Within 'Dark Data'?*, IDG CONNECT (Apr. 13, 2017), http://www.idgconnect.com/blog-abstract/25968/what-awaits-discovery-dark (claiming that dark data "is the spoil heap into which it's possible we've been throwing pearls . . . .").

125. *See* Peterson, *supra* note 115 and accompanying text (explaining that many companies "hoard data" assuming it will be useful in the future).

126. *See* Martin Heller, *10 Signs You're Ready for AI—But Might Not Succeed*, CIO (Aug. 29, 2017, 3:00 AM), https://www.cio.com/article/3219710/artificial-intelligence/10-signs-your-it-organizaiton-is-ready-for-artificial-intelligence.html (explaining the initial costs of hiring data analysts and scientists to help make artificial intelligence systems useful to those who invest in them).

127. Mike Montiero, *Healthcare's Dark Data Problem Needs a Super Human Solution*, MEDCITY NEWS (July 5, 2017, 1:21 AM), https://medcitynews.com/2017/07/healthcares-dark-data-problem-super-human-solution.

128. *Id.*

129. *See, e.g.*, E&Y, Changing the Way, *supra* note 8, at 5 (highlighting advances in cloud computing).

advanced analytics are expensive, are derived from emerging high technology fields, and may offer speculative returns on investment at distant points in the future.[130] The wearable technology industry is an example of this delayed gratification, as wearable devices are harvesting "massive" amounts of users' biomedical data, even though useful medical insights may be years away.[131] The delay between data collection and insight has not dented the wearable medical devices industry, which is expected to be worth more than $48 billion by 2023.[132]

Warehousing dark data in the hope of unlocking future value is not risk-free, as dark data may conceal very real present dangers. The following sections illustrate dark data's invisible risk problem by examining the medical privacy and consumer protection contexts.

B. Example 1: Medical Privacy

1. HIPAA background

Federal statutory protections for medical privacy in the United States revolve around the Health Insurance Portability and Accountability Act of 1996 (HIPAA),[133] which created a regulatory framework for the use and disclosure of protected health information (PHI).[134] The U.S. Department of Health and Human Services (HHS) defines PHI as "individually identifiable health information."[135] HHS regulates the use of PHI by HIPAA-covered entities, which include health care providers, health plans, and health care clearinghouses, as well as "business associates" of those entities.[136] HIPAA's entity-based construction has

130. *See* Heller, *supra* note 126 (describing the high initial overhead associated with implementing artificial intelligence systems).

131. Tom Foremski, *Dark Side of Wearables: Tsunami of Useless Medical Big Data*, ZDNET (Dec. 9, 2016, 12:26 PM), http://www.zdnet.com/article/dark-side-of-wearables-tsunami-of-useless-medical-big-data.

132. *Wearable Devices: Useful Medical Insights or Just More Data?*, SCIENCEDAILY (Aug. 2, 2018), https://www.sciencedaily.com/releases/2018/08/180802115622.htm (summarizing content originally published in FRONTIERS IN PHYSIOLOGY).

133. Pub. L. No. 104-191, 110 Stat. 1936 (1996).

134. 45 C.F.R. pts.160 and 164 (2018).

135. 45 C.F.R. § 160.103. Individually identifiable health information can include "demographic information collected from an individual" that "[i]s created or received by a health care provider, health plan, employer, or heath care clearinghouse," and "identifies the individual" or creates "a reasonable basis to believe" that the information "can be used to identify the individual." *Id.*

136. HIPAA applies to covered entities and their business associates, as defined by 45 C.F.R. § 160.103. *Covered Entities and Business Associates*, OFF. FOR CIV. RTS., DEP'T OF HEALTH & HUM. SERVS., https://www.hhs.gov/hipaa/for-professionals/covered-entities (last visited Feb. 5, 2019). Under HIPAA, a "business associate" of a covered

been criticized as its "original sin": "[t]he data protection model is structured around a group of identified health-care data custodians rather than around health-care data."[137]

The HIPAA framework for protecting PHI has meandered through numerous administrative rulemakings,[138] ultimately resulting in the Privacy Rule, the Security Rule, and the Breach Notification Rule, all of which HHS enforces through its Office for Civil Rights (OCR).[139] The foundational Privacy Rule[140] "strikes a balance" between prohibiting the unauthorized use or disclosure of PHI without impeding the use of medical data to advance public health.[141] This balance-striking effort resulted in the Privacy Rule's permissive treatment of the disclosure of "de-identified" PHI, which reflects an intent to reduce individual privacy risks while allowing PHI to be put to productive medical use.[142] Two de-identification methods comply with the Privacy Rule: the "Expert Determination" method, in which a qualified expert concludes that a disclosure of PHI poses minimal risk of identifying an individual, and the "Safe Harbor" method, which

---

entity is any person or entity who, on behalf of a covered entity, creates or handles PHI or performs services involving PHI. § 160.103.

137. Nicolas P. Terry, *Regulatory Disruption and Arbitrage in Health-Care Data Protection*, 17 YALE J. HEALTH POL'Y L. & ETHICS 143, 164 (2017).

138. For a detailed discussion of the evolution of HIPAA rules, see generally Stacey A. Tovino, *The HIPAA Privacy Rule and the EU GDPR: Illustrative Comparisons*, 47 SETON HALL L. REV. 973 (2017).

139. *About Us*, OFF. FOR CIV. RTS., DEP'T OF HEALTH & HUM. SERVS., https://www.hhs.gov/ocr/about-us (last visited Feb. 5, 2019); *Enforcement Process*, OFF. FOR CIV. RTS., DEP'T OF HEALTH & HUM. SERVS., https://www.hhs.gov/hipaa/for-professionals/compliance-enforcement/enforcement-process (last visited Feb. 5, 2019).

140. 45 C.F.R. pts. 160, 164(A), 164(E).

141. OFF. FOR CIV. RTS., DEP'T OF HEALTH & HUM. SERVS., OCR PRIVACY BRIEF: SUMMARY OF THE HIPAA PRIVACY RULE (2003), https://www.hhs.gov/sites/default/files/privacysummary.pdf.

142. *See* §§ 164.502(d)(2), 164.514(a)–(b); *see also De-Identification and its Rationale*, DEP'T OF HEALTH & HUM. SERVS., https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html#rationale (last visited Feb. 5, 2019). The Privacy Rule provides for two methods of de-identification: (1) the "Expert Determination" method, in which a qualified expert determines that the risk of identifying the individual is "very small," and (2) the "Safe Harbor" method, which requires the removal of 18 specific identifiers. § 164.514(b)–(c); *see also* DEP'T OF HEALTH AND HUM. SERVS., GUIDANCE REGARDING METHODS FOR DE-IDENTIFICATION OF PROTECTED HEALTH INFORMATION IN ACCORDANCE WITH THE HEALTH INSURANCE PORTABILITY AND ACCOUNTABILITY ACT (HIPAA) PRIVACY RULE 5–6 (2012) [hereinafter HHS, DE-IDENTIFICATION GUIDANCE], https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification (explaining methods for de-identification of PHI).

requires eighteen specified individual identifiers to be removed from PHI prior to disclosure.[143]

Like the Privacy Rule, HHS describes the Security Rule[144] as a tool of balance.[145] As a reaction to the digitization of medical records, the Security Rule was designed to protect PHI while allowing covered entities "to adopt new technologies to improve the quality and efficiency of patient care."[146] The Security Rule applies only to PHI in electronic format, or "e-PHI," which is a subset of the information covered by the Privacy Rule.[147] Among other requirements, the Security Rule mandates that covered entities "[p]rotect against any reasonably anticipated threats . . . to the security or integrity" of e-PHI,[148] thus establishing what amounts to a federal cybersecurity mandate in the healthcare space.

The HIPAA framework was strengthened in 2009, when the Health Information Technology for Economic and Clinical Health Act (HITECH) became law.[149] HITECH enhanced HIPAA penalties, made business associates directly liable for Privacy and Security Rule violations, and charged HHS with establishing what would become the Breach Notification Rule, which requires covered entities to promptly notify affected individuals of PHI breaches.[150]

---

143. The Safe Harbor method's eighteen individual identifiers are listed in 45 C.F.R. § 164.514(b)(2)(i)(A)–(R) and include telephone numbers, Social Security numbers, medical record numbers, Internet Protocol (IP) addresses, and biometric identifiers such as fingerprints.

144. 45 C.F.R. pts. 160, 164(A), 164(C).

145. *See* OFF. OF CIV. RTS., DEP'T OF HEALTH & HUM. SERVS., OCR PRIVACY BRIEF: SUMMARY OF THE HIPAA SECURITY RULE (2013) [hereinafter OCR SECURITY RULE SUMMARY], https://www.hhs.gov/hipaa/for-professionals/security/laws-regulations.

146. *Id.*

147. *Id.*

148. § 164.306(a)(2).

149. Pub. L. No. 111-5, 123 Stat. 115, 226–79 (2009). For a detailed discussion of the specific ways that HITECH strengthened HIPAA, see Melissa M. Goldstein & William F. Pewen, *The HIPAA Omnibus Rule: Implications for Public Health Policy and Practice*, 128 PUB. HEALTH REP. 554, 555 (2013).

150. 45 C.F.R. pts. 160, 164(A), 164(D) (implementing the Breach Notification Rule); *see also* OFF. FOR CIV. RTS., DEP'T OF HEALTH & HUM. SERVS., GUIDANCE TO RENDER UNSECURED PROTECTED HEALTH INFORMATION UNUSABLE, UNREADABLE, OR INDECIPHERABLE TO UNAUTHORIZED INDIVIDUALS, https://www.hhs.gov/hipaa/for-professionals/breach-notification/guidance (last visited Feb. 5, 2019) (specifying notification regarding PHI that has not been "rendered unusable, unreadable, or indecipherable" to unauthorized persons, including through acceptable encryption methods).

2. Dark data and the Privacy Rule

The HIPAA framework is a set of legal responses to technology's changing effects on medical privacy, beginning with the migration of medical records from filing cabinets to databases[151] and continuing through contemporary preoccupations with cybersecurity and data breaches.[152] In the Big Data era, HIPAA-covered entities must contend with the reality that data collection and storage technologies are outstripping analytics, causing databases to swell with dark data that cannot be readily analyzed or safeguarded.[153] The contents of this medical dark data may contain regulated PHI that must be used, disclosed, and protected within the confines of the HIPAA framework. The inability to identify and protect PHI residing within troves of dark data creates immediate, invisible risks to covered entities.[154]

HHS has alluded to the problem of dark data concealing PHI. In guidance addressing PHI de-identification under the Privacy Rule, HHS acknowledges that "[m]edical records are comprised of a wide range of structured and unstructured (also known as 'free text') documents."[155] By challenging the "unstated assumption" that sensitive data "only lives in structured formats,"[156] HHS warns that "PHI may exist in different types of data in a multitude of forms and formats," and that while PHI "may reside in highly structured database tables, such as billing records," it may also be present "in a wide range of documents with less structure

---

151. *See* Peter A. Winn, *Confidentiality in Cyberspace: The HIPAA Privacy Rules and the Common Law*, 33 RUTGERS L.J. 617, 617 (2002) (describing how the HIPAA Privacy Rule resulted from "a lack of confidence in the ability of traditional common law doctrines to protect personal health information" when "vast health information networks" replaced paper records).

152. *See, e.g.*, Michael H. Bauscher & Kortni M. Hadley, *Guidance on Cybersecurity: The HIPAA Breach Notification Rule*, CARTER LEDYARD & MILBURN LLP (July 14, 2017), http://www.clm.com/docs/8006343_3.pdf (discussing the HIPAA Security and Breach Notification Rules within the cybersecurity context).

153. *See* Montiero, *supra* note 127 (describing limitations on the ability to harness insights from medical dark data); *see also* Woodie, *supra* note 102 (describing the immense volume of data being retained by businesses despite minimal, if any, long term value); *supra* note 131 and accompanying text (explaining the lag period between collecting data and the ability to harness technologies that will make the data useful).

154. *See* Jay Savaiano, *Bring Healthcare's Dark Data to Light*, HEALTHCARE IT NEWS (Jan. 30, 2013, 12:00 AM), http://www.healthcareitnews.com/news/bring-healthcares-dark-data-light ("[O]ne of the biggest threats to compliance and security [in the healthcare industry] is 'dark data,' which is unaccounted for by IT departments that have no insight or centralized control over how it's being created, stored or used.").

155. HHS, DE-IDENTIFICATION GUIDANCE, *supra* note 142, at 29.

156. Andy Green, *Personally Identifiable Information Hides in Dark Data*, VARONIS (Apr. 30, 2013), https://blogvaronis2.wpengine.com/personally-identifiable-information-hides-in-dark-data.

and written in natural language, such as discharge summaries, progress notes, and laboratory test interpretations."[157]

The presence of unstructured—or dark—data does not relax Privacy Rule compliance obligations, as "[t]he de-identification standard makes no distinction between data entered into standardized fields and information entered as free text (i.e., structured and unstructured text) . . . ."[158] Irrespective of whether PHI is structured or unstructured, a covered entity seeking protection under the Safe Harbor de-identification method must remove each individual identifier "regardless of its location in a record if it is recognizable as an identifier."[159]

While the meaning of "recognizable" within the dark data context can be debated, the above passages create a bright-line standard for de-identification under the Privacy Rule.[160] By making no legal distinction between structured and unstructured data as it relates to the Privacy Rule identifiers, HHS suggests that OCR could bring an enforcement action based on improperly-disclosed PHI even where the disclosing party is unaware that PHI is buried within disclosed dark data.[161] Thus, any HIPAA-compliant disclosure under the Safe Harbor method requires locating, analyzing, and de-identifying any dark data that contains a Privacy Rule identifier.[162]

3. Dark data and the Security Rule

HIPAA's Security Rule creates a healthcare cybersecurity structure by "operationaliz[ing] the protections contained in the Privacy Rule" through safeguards intended to protect e-PHI from loss and misuse.[163] The Security Rule is built around three categories of safeguards—administrative, physical, and technical—that covered entities must

157. HHS, DE-IDENTIFICATION GUIDANCE, *supra* note 142, at 29.

158. *Id.*

159. *Id.*

160. *See* Green, *supra* note 156 (arguing that HHS does not make a distinction between structured and unstructured or dark data with respect to Privacy Rule compliance).

161. OCR regularly brings enforcement actions against covered entities for improper disclosure of PHI. *See, e.g.*, DEP'T OF HEALTH & HUM. SERVS., 21ST CENTURY ONCOLOGY RESOLUTION AGREEMENT AND CORRECTIVE ACTION PLAN ¶¶ I.2.A., II.6, https://www.hhs.gov/sites/default/files/21co-ra_cap.pdf (last visited Feb. 5, 2019) ($2.3 million settlement followed improper disclosure of PHI); DEP'T OF HEALTH & HUM. SERVS., CARDIONET RESOLUTION AGREEMENT ¶¶ I.2.C., II.1 (Apr. 3, 2017), https://www.hhs.gov/sites/default/files/cardionet-ra-cap.pdf ($2.5 million settlement resulted from allowing an unauthorized individual to access PHI).

162. HHS, DE-IDENTIFICATION GUIDANCE, *supra* note 142, at 29.

163. OCR SECURITY RULE SUMMARY, *supra* note 145. As noted above, the Security Rule only applies to electronic PHI.

address in their handling of e-PHI.[164] While the Security Rule establishes several general requirements for covered entities,[165] the safeguards adopt a "[f]lexibility of approach" that recognizes organizational differences.[166] To account for distinctions among covered entities, the safeguards provide "Required" and "Addressable" standards, the latter of which include a reasonableness assessment that accounts for organizational difference.[167]

a. Risk assessments

A critical "Required" administrative safeguard under the HIPAA Security Rule is the obligation of covered entities to "[c]onduct an accurate and thorough assessment of the potential risks and vulnerabilities to the confidentiality, integrity, and availability" of e-PHI under their control.[168] Conducting a risk assessment "is the first step" in complying with Security Rule obligations to "implement reasonable and appropriate security measures to protect against reasonably anticipated threats or hazards" to e-PHI.[169] As the starting point from which other Security Rule measures grow, conducting an adequate risk assessment is a "foundational" requirement for covered entities.[170]

In its Risk Analysis Guidance, OCR emphasizes that "an organization's risk analysis should take into account all of its e-PHI, regardless of [its source, location, or] the particular electronic

164. 45 C.F.R. §§ 164.306, 164.310, 164.312 (2017).

165. Compliance with the Security Rule requires covered entities to satisfy four general requirements: (1) take measures to protect "the confidentiality, integrity, and availability of" e-PHI; (2) identify and guard against "reasonably anticipated threats" to the security and integrity of e-PHI; (3) protect against "reasonably anticipated [impermissible] uses or disclosures"; and (4) ensure their workforces comply with Security Rule measures. § 164.306(a).

166. § 164.306(b).

167. § 164.306(d)(1). The implementation specifications for "Required" standards must be performed. Implementation specifications under "Addressable" standards trigger an assessment of whether the specification is "reasonable and appropriate" for the entity "when analyzed with reference to the likely contribution to protecting [e-PHI]." § 164.306(d)(3)(i). Covered entities must enact the specification if it is reasonable and appropriate; if it is not, the entity must document the reasons why, and must implement an "equivalent alternative" if that is itself reasonable and appropriate. § 164.306(d)(3)(ii)(B)(2).

168. § 164.308(a)(1)(ii)(A).

169. OFF. FOR CIV. RTS., DEP'T OF HEALTH & HUM. SERVS., GUIDANCE ON RISK ANALYSIS REQUIREMENTS UNDER THE HIPAA SECURITY RULE 1–2 (2010), https://www.hhs.gov/sites/default/files/ocr/privacy/hipaa/administrative/securityrule/rafinalguidancepdf.pdf.

170. *Id.* at 1.

medium in which it is created, received, maintained or transmitted."[171] Stated more directly, "[a]n organization must identify where [its] e-PHI is stored, received, maintained or transmitted."[172] The failure to search "all IT equipment, applications, and data systems utilizing e-PHI" can risk costly enforcement actions.[173]

Dark data can significantly complicate Security Rule risk assessments. As the nature of dark data can prevent organizations from knowing that it exists or understanding its content, e-PHI concealed within dark data may be exceedingly difficult to effectively assess and safeguard.[174] Size and complexity may blind large organizations to e-PHI within their dark data, as dispersed and siloed data repositories can make data management difficult.[175] Small organizations may face different challenges, including resource constraints that complicate the task of identifying dark data and determining whether it may hide e-PHI.[176]

By casting a cloud of uncertainty over the completeness and accuracy of risk assessments, dark data can frustrate organizations' ability to satisfy their Security Rule obligations. Unknowingly incomplete risk assessments can produce cascading vulnerabilities, as the inability to

---

171. *Id.* at 5.

172. *Id.*

173. *See, e.g.*, DEP'T OF HEALTH & HUM. SERVS., NEW YORK AND PRESBYTERIAN HOSPITAL ("NYP") RESOLUTION AGREEMENT ¶¶ I.2.b., II.6, https://www.hhs.gov/sites/default/files/ocr/privacy/hipaa/enforcement/examples/ny-and-presbyterian-hospital-settlement-agreement.pdf (last visited Feb. 5, 2019) (NYP agreed to pay $3.3 million for e-PHI disclosures during data breach and failure to perform a thorough risk assessment). OCR entered into a similar Resolution Agreement with the Trustees of Columbia University in the City of New York ("Columbia") following the same data breach. DEP'T OF HEALTH & HUM. SERVS., COLUMBIA RESOLUTION AGREEMENT ¶ II.6, https://www.hhs.gov/sites/default/files/ocr/privacy/hipaa/enforcement/examples/columbia-university-resolution-agreement.pdf (last visited Feb. 5, 2019) (resulting in $1.5 million resolution amount).

174. *See* Ken Spinner, *Is Dark Data Putting Your Organization at Risk?*, DATA CTR. KNOWLEDGE (May 21, 2018), http://www.datacenterknowledge.com/industry-perspectives/dark-data-putting-your-organization-risk ("If your dark data holds, for example, a Word document with employee PII . . . your organization may be violating regulations such as . . . HIPAA . . . . Unfortunately, many companies don't know this data is even on their network and [they] fail to secure it.").

175. *See IT Leaders Fear Data Fragmentation is Putting Businesses at Risk*, MIMECAST, https://www.mimecast.com/resources/press-releases/dates/2013/6/it-leaders-fear-data-fragmentation-is-putting-businesses-at-risk (last visited Feb. 5, 2019) ("IT managers believe that fragmentation of corporate data across their IT infrastructure and an emerging 'Shadow IT' network of user devices or consumer cloud services outside their control, is putting their organizations at risk and driving up costs.").

176. *See* Christina Donnelly & Geoff Simmons, *Small Businesses Need Big Data, Too*, HARV. BUS. REV. (Dec. 5, 2013), https://hbr.org/2013/12/small-businesses-need-big-data-too (discussing resource constraints that can hamper small organizations' use of Big Data).

locate or identify e-PHI may defeat security measures resulting from the assessment. Just as damaging is the possibility of a false confidence effect, as misplaced faith in a faulty risk assessment may cause overreliance on security efforts that are prone to unseen deficiencies.[177] The combination of invisible risk and over-confidence in the effectiveness of security measures that address other, known risks may create blind spots in which a covered entity's "information security team may not even realize that they have [e-PHI] in their organization until it gets breached."[178]

b. Other safeguards

In addition to the challenges it poses for risk assessments, dark data can vex organizational compliance with other HIPAA Security Rule safeguards.[179] The technical safeguard of access control[180] is a clear example. Well-functioning access controls limit the availability of e-PHI to persons or software programs that have been given access rights[181] and, when authentication standards are in place,[182] can demonstrate that they are in fact the persons or programs to whom access has been granted. Failure to establish access controls over e-PHI can expose covered entities to data breaches and resulting OCR enforcement actions.[183]

---

177. Jonathan Litchman, *The False Promise of HIPAA for Healthcare Cybersecurity*, HEALTH IT SECURITY (Mar. 8, 2016), https://healthitsecurity.com/news/the-false-promise-of-hipaa-for-healthcare-cybersecurity.

178. *OCR Cyber-Awareness Monthly Update*, OFF. FOR CIV. RTS., DEP'T OF HEALTH & HUM. SERVS. (Mar. 3, 2016), https://www.hhs.gov/sites/default/files/hipaa-cyber-awareness-monthly-issue2.pdf.

179. While the Security Rule's technical safeguards are most relevant to this Article, the connection between physical safeguards and dark data cannot be ignored. *See, e.g.*, 45 C.F.R. § 164.310(d)(1) (2018) (providing for policies and procedures that "govern the receipt and removal of hardware and electronic media that contain [e-PHI] into and out of a facility, and the movement of these items within the facility"). Establishing security protocols over electronic devices can help prevent employees from creating dark data by removing e-PHI from an organization without the organization's knowledge. *See* Nelson & Simek, *supra* note 97, at 1553 (describing how employees can create dark data by transferring data to portable devices).

180. *See* § 164.312(a)(1).

181. *Do the Security Rule Requirements for Access Control, Such as Automatic Logoff, Apply to Employees Who Telecommute or Have Home-Based Offices if the Employees Have Access to Electronic PHI (e-PHI)?*, OFF. FOR CIV. RTS., DEP'T OF HEALTH & HUM. SERVS., https://www.hhs.gov/hipaa/for-professionals/faq/2004/do-the-security-rule-require mennts-for-access-control-apply-to-employees-that-telecommute (last visited Feb. 5, 2019).

182. *See* § 164.312(d) (detailing authentication standards).

183. *See, e.g.*, DEP'T OF HEALTH & HUM. SERVS., SOUTH BROWARD HOSPITAL DISTRICT D/B/A MEMORIAL HEALTHCARE SYSTEM (MHS) RESOLUTION AGREEMENT ¶¶ I.2.A, II.6 (Feb. 14, 2017), https://www.hhs.gov/sites/default/files/memorial-ra-cap.pdf (MHS

Dark data can create roadblocks to implementing effective access controls over e-PHI. The challenge of locating dark data and determining its contents may preclude a covered entity from effectively situating dark data within its data security framework. For example, if a particular server holds dark data of unknown content, an organization may inadvertently fail to place adequate access control restrictions on the server, risking Security Rule violations if e-PHI is present.[184] The difficulty in cataloguing dark data and incorporating it into data security processes can leave sensitive data unprotected from impermissible access.

Another example of dark data's mischief involves the Security Rule's encryption provision.[185] According to OCR, encryption can aid the access control standard[186] by preventing data from being accessed unless valid decryption keys are presented.[187] Encryption can also advance the Security Rule's transmission security standard, which requires covered entities to implement measures to guard against unauthorized access to e-PHI transmitted over electronic networks.[188] While encryption is an "Addressable" implementation standard and covered entities have "flexibility to determine when, with whom, and what method of encryption to use," OCR has taken the position that covered entities "must encrypt" transmissions where a "significant" risk of unauthorized access or interception exists.[189]

The problem is that encryption standards are rarely applied to the universe of an organization's data,[190] and risk assessments may miss the presence of

---

agreed to pay $5.5 million for failing to manage access controls, which contributed to a former employee accessing the PHI of 80,000 persons); *see also* DEP'T OF HEALTH & HUM. SERVS., REGENTS OF THE UNIVERSITY OF CALIFORNIA RESOLUTION AGREEMENT ¶ I.2.B.ii (July 6, 2011), https://www.hhs.gov/sites/default/files/ocr/privacy/hipaa/enforcement/examples/uclahsracap.pdf (describing access control failures that resulted in improper viewing of e-PHI).

184. *See supra* note 183 (providing examples where liability was imposed for inadequate access control restrictions).

185. Encryption is an "Addressable" implementation specification under the access control standard. § 164.312(a)(2)(iv).

186. § 164.312(a)(1).

187. OFF. FOR CIV. RTS., DEP'T OF HEALTH & HUM. SERVS., SECURITY STANDARDS: TECHNICAL SAFEGUARDS, 2 HIPAA Security Series 1, 6–7 (2007) [hereinafter OCR SECURITY SERIES], https://www.hhs.gov/sites/default/files/ocr/privacy/hipaa/administrative/securityrule/techsafeguards.pdf.

188. § 164.312(e)(1), (2)(ii).

189. OCR SECURITY SERIES, *supra* note 187, at 12.

190. Nathan Cranford, *IBM Debuts Universal Encryption Mainframe to Combat Hackers*, RCR WIRELESS NEWS (July 18, 2017), https://www.rcrwireless.com/20170718/business/ibm-debuts-universal-encryption-mainframe-tag27 (noting the large amount of computational power required to encrypt and decrypt data).

dark data altogether, causing it to evade security measures.[191] As dark data is unknown and unanalyzed, it is more likely to be omitted from limited encryption protocols,[192] leaving holes in e-PHI protection efforts.

C. Example 2: Consumer Protection

The Federal Trade Commission (FTC) has drawn on its vast consumer protection jurisdiction to reach across industries and establish an outsized enforcement role in the privacy space.[193] Having used its consumer protection mandate to become America's "lead privacy law enforcer,"[194] the FTC has in recent years increasingly recognized that privacy protection is tightly intertwined with data security,[195] both of which were historically fractured by spider webs of industry-specific legal controls.[196] Heightened understanding of the extensive financial damage that data breaches can inflict on

191. *See supra* Section II.B.3.a.

192. Matthew Davis, *Dark Data is a Risk and an Opportunity for Small Businesses*, FUTURE HOSTING (Feb. 23, 2017), https://www.futurehosting.com/blog/dark-data-is-a-risk-and-an-opportunity-for-small-businesses ("Dark data is usually not encrypted or subject to the same protection as data that is known to be sensitive . . . .").

193. *See* Daniel J. Solove & Woodrow Hartzog, *The FTC and the New Common Law of Privacy*, 114 COLUM. L. REV. 583, 588 (2014) [hereinafter Solove & Hartzog, *FTC Privacy*].

194. Edith Ramirez, Former Commissioner, FTC, Keynote Address at the Georgetown Law Center 2011 Computer, Freedom and Privacy Conference: Learning From History: Mobile and the Future of Privacy 6 (June 14, 2011) [hereinafter Ramirez, Georgetown Address], https://www.ftc.gov/public-statements/2011/06/learning-history-mobile-and-future-privacy.

195. *See* Julie Brill, Former Commissioner, FTC, Remarks Before the International Conference of Data Protection and Privacy Commissioners: Big Data and Consumer Trust: Progress and Continuing Challenges 3 (Oct. 15, 2014), https://www.ftc.gov/system/files/documents/public_statements/592771/141015brillicdppc.pdf ("Data security is also an FTC priority because . . . there is no privacy without appropriate data security.").

196. Jolly, *supra* note 104 (discussing the "patchwork system of federal and state laws and regulations that can sometimes overlap, dovetail and contradict one another" in the data protection sphere).

consumers[197] has deepened the FTC's interest in data security as a consumer protection issue.[198]

The FTC has applied its consumer protection powers to the privacy and data protection spheres by addressing behavior that might otherwise escape the reach of sectoral legal constraints.[199] As "the law of privacy and data security is so fragmented, so magma-like in its nature, the FTC has had an unusually influential role" in these areas by "embracing certain standards and norms that have achieved a decent level of consensus."[200] Through expansive consumer protection enforcement powers covering "nearly every industry,"[201] the FTC has been able to steadily bring privacy and, later, data security within the ambit of its principal consumer protection enforcement tool, section 5 of the Federal Trade Commission Act (FTC Act).[202]

---

197. Hacking and other unauthorized data access can cause significant harms, including identity theft as well as blackmail, stalking, and physical harm in extreme cases. Vincent R. Johnson, *Cybersecurity, Identity Theft, and the Limits of Tort Liability*, 57 S.C. L. REV. 255, 256–57 (2005); *see also* James C. Cooper, *Separation Anxiety*, 21 VA. J.L. & TECH. 1, 10 (2017) ("[E]ven if identity theft does not result in direct financial losses, the time and hassle of reestablishing one's identity is harmful. Then there are subjective harms, which include any direct psychic or embarrassment costs . . . ."). In addition to its individual harms, identity theft takes a massive economic toll, with one estimate suggesting that American consumers lost $17 billion to the crime in 2017. Jessica Rich, *Beyond Facebook: It's High Time for Stronger Privacy Laws*, WIRED (Apr. 8, 2018, 8:00 AM), https://www.wired.com/story/beyond-facebook-its-high-time-for-stronger-privacy-laws.

198. *See* Julie Brill, Former Commissioner, FTC, Keynote Address at the Center for Strategic and International Studies Workshop on Stepping into the Fray: The Role of Independent Agencies in Cybersecurity: On the Front Lines: The FTC's Role in Data Security 2 (Sept. 17, 2014) [hereinafter Brill, CSIS Speech], https://www.ftc.gov/system/files/documents/public_statements/582841/140917csisspeech.pdf ("Data security is one of our top consumer protection priorities. In our enforcement actions and policy initiatives, we focus on the harms that consumers may suffer when companies fail to keep information secure.").

199. Woodrow Hartzog & Daniel J. Solove, *The Scope and Potential of FTC Data Protection*, 83 GEO. WASH. L. REV. 2230, 2293 (2015) [hereinafter Hartzog & Solove, *FTC Data Protection*].

200. *Id.*

201. *Id.* at 2236. Hartzog and Solove identify the following industries as falling within the FTC's consumer protection jurisdiction: "automotive, financial, health, retail, online services, hospitality, entertainment, manufacturing, data processing, food and beverage, transportation, and many more." *Id.*

202. 15 U.S.C. §§ 41–58 (2012). *See* Maureen K. Ohlhausen, Former Commissioner, FTC, Remarks before the Federal Communications Bar Association: Painting the Privacy Landscape: Informational Injury in FTC Privacy and Data Security Cases 2 (Sept. 19, 2017), https://www.ftc.gov/system/files/documents/public_statements/1255113/privacy_speech_mkohlhausen.pdf ("Our primary privacy and data security tool is enforcement under our section 5 authority to protect consumers from deceptive or unfair acts or practices."); *see also* Brill, CSIS Speech, *supra* note 198, at 3 ("The main

Section 5 broadly prohibits "unfair or deceptive acts or practices in or affecting commerce"[203] and grants the FTC enforcement authority over such conduct[204] where a U.S. nexus is present.[205] While the FTC also enforces certain sectoral consumer protection laws,[206] it is the evolution of section 5 enforcement that has driven the FTC's progressively deeper forays into the privacy and data protection spaces.[207] As technology has advanced, so too has the FTC applied its section 5 enforcement powers to new arenas, including mobile devices and applications,[208] cloud security,[209] the Internet of Things,[210] encryption,[211] and phishing.[212]

---

legal authority that the FTC uses in the data security space is section 5 of the FTC Act, which gives us the ability to stop unfair or deceptive acts or practices.").

203. 15 U.S.C. § 45(a)(1).

204. § 45(a)(4)(B) (granting the FTC the power to seek "[a]ll remedies" available to it in response to unfair or deceptive acts or practices); § 45(b) (providing for FTC administrative proceedings in response to unfair or deceptive acts or practices); § 45(m) (authorizing the FTC to commence civil enforcement actions in response to unfair or deceptive acts or practices).

205. The FTC is limited to pursuing unfair or deceptive acts or practices that cause or are likely to cause reasonably foreseeable injury within the United States, or which involve material conduct occurring within the United States. § 45(a)(4)(A).

206. An example is the Standards for Safeguarding Customer Information Rule ("Safeguards Rule"). 16 C.F.R. pt. 314 (2018). The Safeguards Rule, which governs the treatment of customer information at defined financial institutions, was implemented pursuant to sections 501 and 505(b)(2) of the Gramm-Leach-Bliley Act. 15 U.S.C. §§ 6801(b), 6805(b)(2) (2012).

207. Hartzog & Solove, *FTC Data Protection*, *supra* note 199, at 2293; *see also* Ohlhausen, *supra* note 202, at 1–2 ("[T]he FTC is the primary U.S. enforcer of commercial privacy and data security obligations," and has "brought more than 500 privacy and data security related cases.").

208. *See, e.g.*, Gen. Workings Inc., FTC Docket No. C-4573, 2016 WL 2894073 (Apr. 18, 2016) (applying section 5 enforcement power to mobile applications that have access to sensitive information); *see also* Snapchat, Inc., FTC Docket No. C-4501, 2014 WL 7495798 (Dec. 23, 2014); Credit Karma, Inc., FTC Docket No. C-4480, 2014 WL 4252397 (Aug. 13, 2014); Fandango, LLC, FTC Docket No. C-4481, 2014 WL 4252396 (Aug. 13, 2014).

209. *See, e.g.*, Complaint at *2, ASUSTeK Comput., Inc., FTC Docket No. C-4587, 2016 WL 4128217 (July 18, 2016) (involving a cloud feature on ASUS devices marketed as private and secure that was in fact vulnerable to attackers).

210. *See, e.g.*, TRENDnet, Inc., FTC Docket No. C-4426, 2014 WL 556262 (Jan. 16, 2014) (involving networked devices including security cameras and routers).

211. *See, e.g.*, Superior Mortg. Corp., 140 F.T.C. 926, 928 (Dec. 14, 2005), 2005 WL 6241024 (alleging that defendants failed to safeguard website interfaces, including through strong password policies and data encryption).

212. *See, e.g.*, Complaint for Permanent Injunction and Other Equitable Relief, FTC v. Hill, No. H-03-5537 (S.D. Tex. Dec. 3, 2003) (alleging unfair and deceptive practices in an email phishing scam).

Many leading names in the technology sector have now been on the receiving end of section 5 enforcement actions.[213]

1. Unfairness and deception

Facts supporting FTC claims under section 5's unfairness and deception prongs "frequently overlap," and it is not always possible to "completely disentangle the two theories."[214] Each nonetheless contains nuance worth exploring. As "[t]he FTC's data security enforcement actions initially focused on deception," so too will this Article.[215]

Section 5 deception theory is rooted in a 1983 FTC Policy Statement that identified three elements "undergird[ing] all deception cases."[216] To establish that an act is deceptive under section 5, the FTC must show: (1) "a representation, omission or practice that is likely to mislead the consumer;" (2) that the act in question is deceptive when considered "from the perspective of a consumer acting reasonably in the circumstances;" and (3) materiality, in that the representation, omission, or practice "is likely to affect the consumer's conduct or decision with regard to a product or service."[217]

As "[m]ost deception involves written or oral misrepresentations, or omissions of material information,"[218] FTC consumer privacy actions have traditionally invoked "a deception theory of broken promises," in which a company violates representations it voluntarily made in its privacy policies.[219] Subsequent FTC enforcement began to include "a broader conception of deception . . . that did not rely only on explicit promises made."[220] Data security cases are among the new "[g]eneral [d]eception" actions that abandon traditional "broken promises" theory to reach acts unrelated to the breach of written policies, such as deceptively inducing consumers to download spyware.[221]

---

213. *See, e.g.*, Complaint for Permanent Injunction and Other Equitable Relief, FTC v. Uber Techs., Inc., No. 3:17-cv-00261 (N.D. Cal. Jan. 19, 2017); Oracle Corp., FTC Docket No. C-4571, 2016 WL 1360808 (Mar. 28, 2016); Complaint for Permanent Injunction and Other Equitable Relief, FTC v. Amazon.com, Inc., No. 2:14-cv-01038 (W.D. Wash. July 10, 2014); Facebook, Inc., FTC Docket No. C-4365, 2012 WL 3518628 (July 27, 2012); Google Inc., 152 F.T.C. 435 (Oct. 13, 2011), 2011 WL 11798458.

214. FTC v. Wyndham Worldwide Corp., 799 F.3d 236, 245 (3d Cir. 2015).

215. Brill, CSIS Speech, *supra* note 198, at 3.

216. James C. Miller III, *FTC Policy Statement on Deception*, FED. TRADE COMM'N (Oct. 14, 1983), https://www.ftc.gov/system/files/documents/public_statements/410531/831014deceptionstmt.pdf.

217. *Id.*

218. *Id.*

219. Solove & Hartzog, *FTC Privacy*, *supra* note 193, at 628–29.

220. Hartzog & Solove, *FTC Data Protection*, *supra* note 199, at 2235–36.

221. Solove & Hartzog, *FTC Privacy*, *supra* note 193, at 630–31.

Like deception cases, section 5 unfairness actions have evolved alongside the FTC's expanding role in the data security space. Modern understandings of unfairness under section 5 arise from a 1980 FTC Policy Statement that defined unfair acts or practices as those that cause unjustified consumer injury that is not outweighed by offsetting consumer or competitive benefits.[222] In 1994, Congress amended the FTC Act to incorporate the 1980 definition, resulting in a statutory prohibition on "unfair" acts or practices that are "likely to cause substantial injury to consumers which is not reasonably avoidable by consumers themselves and not outweighed by countervailing benefits to consumers or to competition."[223]

In 2005, the FTC brought its first standalone unfairness case in the data protection and cybersecurity contexts, alleging unfair acts or practices by a wholesaler that experienced a breach of customer credit card data.[224] In the years after, spurred by the absence of a comprehensive federal cybersecurity regime, "the FTC, left alone to police the vast number of data practices not covered by specific internet privacy legislation, has increasingly begun to apply the 'unfairness' prong of section 5 to data security cases."[225] The FTC now pursues unfairness cases against companies that have "hewed to their privacy policies but nevertheless failed . . . to implement adequately robust cybersecurity measures."[226] At least one federal appeals court has validated the FTC's position that the failure to maintain reasonable and appropriate cybersecurity measures can constitute unfair practices under section 5.[227]

2. Dark Data and Section 5

Vacuuming up and warehousing dark data can produce data privacy and cybersecurity harms to consumers. In alluding to the dark data problem, former FTC Chairwoman Edith Ramirez remarked,

---

222. Michael Pertschuk et al., *FTC Policy Statement on Unfairness*, FED. TRADE COMM'N (Dec. 17, 1980), https://www.ftc.gov/public-statements/1980/12/ftc-policy-statement-unfairness.

223. 15 U.S.C. § 45(n) (2012); *see also* Brill, CSIS Speech, *supra* note 198, at 3–4.

224. BJ's Wholesale Club, Inc., 140 F.T.C. 465, 468 (Sept. 20, 2005).

225. Stuart L. Pardau & Blake Edwards, *The FTC, the Unfairness Doctrine, and Privacy by Design: New Legal Frontiers in Cybersecurity*, 12 J. BUS. & TECH. L. 227, 239–40 (2017).

226. *Id.* at 229.

227. FTC v. Wyndham Worldwide Corp., 799 F.3d 236, 249 (3d Cir. 2015) (affirming denial of motion to dismiss section 5 unfairness claim based on a data breach that exposed personal data associated with hundreds of thousands of consumers). Ongoing court battles are being waged over the extent to which the FTC can use section 5 to bring enforcement actions against "intangible" harms resulting from data breaches. *See* LabMD, Inc. v. FTC, 678 F. App'x 816, 820–21 (11th Cir. 2016).

"[c]ompanies often tell the FTC that they cannot innovate unless they are broadly permitted to collect information about consumers, on the theory that they may one day identify a new use for it."[228] Ramirez added that this approach to consumer data—what this Article calls the storage imperative[229]—"is fundamentally at odds with privacy protection."[230]

As to data protection and cybersecurity, Ramirez explained that "the FTC often sees data retained long past its usefulness to the company that collected it. Although it has no continuing use to the company, it is highly attractive to a hacker."[231] Recognizing the risks of unused, accumulated data, the FTC has admonished businesses not to "collect or keep data you don't need."[232] This warning reverberates mightily in the Big Data era.

Retained dark data poses unique cybersecurity risks because its position at the periphery of data management systems often leaves it relatively unprotected and vulnerable to breach.[233] Former FTC Commissioner Julie Brill has expressed concern "that the vast collection of data about consumers can unintentionally . . . include sensitive information, and . . . the necessary heightened protections are not being provided."[234] This is especially the case with dark data. As cybersecurity measures tighten controls over structured data, cybercriminals have turned their attention to unstructured and dark data, which can be rich in value while lightly guarded.[235] Cybersecurity

---

228. Ramirez, Georgetown Address, *supra* note 194, at 7.

229. *See supra* Section II.A.

230. Ramirez, Georgetown Address, *supra* note 194, at 7.

231. *Id.*

232. *App Developers: Start with Security*, FED. TRADE COMM'N, https://www.ftc.gov/tips-advice/business-center/guidance/app-developers-start-security (last visited Feb. 5, 2019).

233. Davis, *supra* note 192.

234. Julie Brill, Former Commissioner, FTC, Remarks at Fordham University School of Law: Big Data, Big Issues 2 (Mar. 2, 2012), https://www.ftc.gov/public-statements/2012/03/big-data-big-issues.

235. *See* Jory Heckman, *Do Agencies Need an 'Awakening' About What Their Data Is Worth?*, FED. NEWS NETWORK (Oct. 26, 2018, 8:29 AM), https://federalnewsnetwork.com/big-data/2018/10/do-agencies-need-an-awakening-about-what-their-data-is-worth (quoting Donna Ray, Executive Director of the U.S. Homeland Security Department, who stated, '"Dark data is becoming the number-one target for most . . . cyber threats, because people realize it's out there and it's a treasure trove of information.'"); *see also* Michelle Drolet, *Protect Your Unstructured Data with User Behavior Analytics*, CSO (Mar. 21, 2017, 9:01 PM), https://www.csoonline.com/article/3182910/security/protect-your-unstructured-data-with-user-behavior-analytics ("Carelessly handled unstructured data is an easy target, and it can prove very valuable for hackers. Since unstructured data may not be monitored, attacks and successful exfiltrations often go unnoticed for long periods."); Edward Goings, *Shining a Light on Dark Data: Securing Information Across the Enterprise*, CIO (Jan. 12, 2016, 7:18 AM), https://www.cio.com/article/3016799/data-breach/shining-a-light-on-dark-data-securing

resources are finite, and organizations will naturally steer protections to known high-value data, leaving more porous controls for unstructured and dark data of unknown value and sensitivity.[236]

Dark data's opacity, in combination with the storage imperative, can place organizational behavior in direct tension with FTC warnings to achieve visibility into "everywhere sensitive data might be stored."[237] Already, the FTC has brought numerous section 5 enforcement actions following breaches of sensitive data that was stored without a present business need.[238] As these enforcement actions show, dark data is too often under-protected, inviting cyber intrusions and the parade of harms they bring.[239]

Constant data creation and capture also produces an environment favorable to inadvertent data collection, which can result in organizations unknowingly acquiring and mishandling sensitive dark data. Inadvertent data collection is not necessarily accidental collection. Rather, inadvertent data collection can occur when an organization intends to capture data—often, a lot of data—but does so in a way that erroneously includes data that the organization does not

---

-information-across-the-enterprise (explaining that because dark data often resides in under-protected locations, "[a]n attacker that can find . . . a poorly defended server will compromise it and reap the same rewards as if they had compromised the primary location of that data"); Holder, *supra* note 105 (discussing how hackers are turning their attention to unstructured data as organizations "lock down sensitive information in structured systems"); Juliette Rizkallah, *The Big (Unstructured) Data Problem*, FORBES (June 5, 2017, 7:00 AM), https://www.forbes.com/sites/forbestechcouncil/2017/06/05/the-big-unstructured-data-problem (referring to unstructured data as a "new attack vector" for hackers as businesses focus on protecting structured data); Spinner, *supra* note 174 (asserting that more unsecured dark data means that hackers have more opportunities to access valuable corporate information).

236. *See supra* note 235 (characterizing unstructured data as vulnerable to hackers because it is difficult to manage and is often under-protected); Davis, *supra* note 192 (explaining that dark data is less protected than structured data).

237. FED. TRADE COMM'N, PROTECTING PERSONAL INFORMATION: A GUIDE FOR BUSINESS 2 (2016), https://www.ftc.gov/system/files/documents/plain-language/pdf-0136_proteting-personal-information.pdf.

238. *See, e.g.*, Accretive Health, Inc., FTC Docket No. C-4432, 2014 WL 726603, ¶¶ 6(c)–(d) (Feb. 4, 2014) (alleging that defendant failed "to ensure that employees removed information from their computers for which they no longer had a business need," and used "consumers' personal information in training sessions with employees and fail[ed] to ensure that the information was [subsequently] removed . . . ."); *see also* Ceridian Corp., 151 F.T.C. 514 ¶ 8 (June 8, 2011), 2011 WL 3568986 (alleging that the defendant "created unnecessary risks to personal information by storing it indefinitely on its network without a business need . . . ."); DSW Inc., 141 F.T.C. 117 ¶ 7 (Mar. 7, 2006), 2006 WL 6679055 (alleging that the defendant "created unnecessary risks to . . . information by storing it in multiple files when it no longer had a business need to keep the information . . . .").

239. *See supra* note 235; *see also* Davis, *supra* note 192.

intend to, or actively attempts to avoid, collecting.[240] Technical failures often account for inadvertent data collection, as seen in two FTC enforcement actions involving automated collection technologies.[241] This Article will focus on the *Upromise* enforcement action,[242] which demonstrates the legal liability that can befall organizations that inadvertently collect sensitive dark data.

3. Upromise, Inc.

In 2012, the FTC charged Upromise, Inc., a Massachusetts company offering college savings account rebates to members who purchase goods and services from Upromise partner merchants.[243] To direct consumers to its partner merchants, Upromise provided an internet browser toolbar that highlighted its partners in online search results.[244] The Upromise toolbar also included a "Targeting Tool"—a modified version of the toolbar designed to direct personalized advertisements to users based on their Web browsing data.[245]

The Targeting Tool caused the Upromise toolbar to "collect extensive information about consumers' online activities," including websites visited, hyperlinks clicked, search terms, usernames, and passwords.[246] The Targeting Tool collected data in the background as users navigated the internet, leaving consumers with no practical way to determine the scope of data collection.[247]

The FTC alleged that the Upromise Privacy Statement represented to consumers that the toolbar "might 'infrequently' collect some personal information," but that "a filter, termed a 'proprietary rules engine,' would 'remove any personally identifiable information.'"[248] The Privacy Statement also represented that Upromise would take "'every commercially viable effort . . . to purge [Upromise's] databases of any personally identifiable information.'"[249] The FTC alleged that Upromise did not adhere to these representations, as the Targeting

240. *See* Brad Stone, *Google Says It Collected Private Data by Mistake*, N.Y. TIMES (May 14, 2010), https://www.nytimes.com/2010/05/15/business/15google.html.

241. *See* Compete, Inc., 155 F.T.C. 264 (Feb. 20, 2013), 2013 WL 8364898; *see also* Upromise, Inc., FTC Docket No. C-4351, 2012 WL 1225058 (Mar. 27, 2012).

242. Upromise, Inc., FTC Docket No. C-4351, 2012 WL 1225058.

243. Complaint ¶ 3, Upromise, Inc., FTC Docket No. C-4351, 2012 WL 1225058 (Mar. 27, 2012).

244. *Id.* ¶ 4.

245. *Id.* ¶¶ 5–7.

246. *Id.* ¶ 7.

247. *Id.*

248. *Id.* ¶ 8.

249. *Id.*

Tool caused Upromise to gather, retain, and transmit sensitive consumer information, including as vulnerable clear text.[250]

At this point, the FTC's case against Upromise reads like a traditional section 5 action based on the alleged breach of self-imposed Privacy Statement representations (i.e. deception) and unsafe handing of consumer data (i.e. unfairness). Indeed, the FTC ultimately charged Upromise with three counts of section 5 deception based on representations made to consumers, as well as a fourth unfairness count resulting from Upromise's "failure to employ reasonable and appropriate measures to protect consumer information."[251] What sets *Upromise* apart from other section 5 cases is that the action turns, in significant part, on the failure of automated technical controls that Upromise implemented for the purpose of narrowing the scope of data collection. As the FTC alleged:

> [A]lthough a filter was used to instruct the Targeting Tool to avoid certain data, the filter was too narrow and improperly structured. For example, although the filter was intended to prevent the collection of financial account personal identification numbers and would have prevented collection of that data if a website used the field name "PIN," the filter would not have prevented such collection if a website used field names such as "personal ID" or "security code."[252]

This passage suggests that Upromise attempted to avoid collecting sensitive data, but was foiled by data filter definitions that were too narrow to serve their intended purpose, placing Upromise unknowingly in breach of its Privacy Statement.[253] The sensitive data was likely dark as to Upromise, which seemingly was unaware that the browser toolbar was collecting data that Upromise's filters were designed to exclude. While the *Upromise* action appears to address structured data that slipped through overly-narrow filter definitions and was thus dark to Upromise, the risk of unwittingly obtaining sensitive data is even higher when unstructured or truly dark data is collected, as it may lack formal organization readable to automated filters.[254]

---

250. *Id.* ¶ 10.
251. *Id.* ¶¶ 15–20.
252. *Id.* ¶ 9.
253. *See id.* ¶ 15.
254. *See* SECUROSIS, L.L.C., UNDERSTANDING AND SELECTING A DATA LOSS PREVENTION SOLUTION 8, https://securosis.com/assets/library/reports/DLP-Whitepaper.pdf (last visited Feb. 5, 2019) (arguing that rule-based data analysis is ideal "for detecting easily identified pieces of structured data like credit card numbers, social security numbers, and healthcare codes/records," but it "[o]ffers very little protection for unstructured content like sensitive intellectual property").

While Upromise initially resolved the matter without paying a civil monetary penalty—likely because its collection of sensitive data was seemingly inadvertent—it nonetheless agreed to perform significant remedial measures under a consent order carrying a twenty-year term.[255] The lesson of *Upromise* is that well-intentioned automated controls designed to limit data collection may not protect against section 5 liability when those controls fail and result in the collection of sensitive data that is dark to the collecting organization.

As the FTC alleged in the similar *Compete, Inc.*[256] action, organizations can reduce their section 5 liability if they "assess and address the risk that . . . data collection software [will] collect sensitive consumer information that it [is] not authorized to collect."[257] Implementing measures to identify the contents of collected dark data and to ensure that automated controls governing data collection are functioning properly will go far in addressing this risk.

D. Emerging Legal Regimes

While HIPAA and the FTC Act represent significant federal data privacy and cybersecurity efforts, neither fully captures the rising prominence of these areas.[258] Developments within the financial

---

255. Decision and Order, Upromise, Inc., FTC Docket No. C-4351 (Mar. 22, 2012), https://www.ftc.gov/sites/default/files/documents/cases/2012/04/120403upromisedo.pdf. On March 16, 2017, the United States, acting on behalf of the FTC, filed a new complaint against Upromise, alleging that it breached the terms of the 2012 consent order. Complaint for Civil Penalty, Injunction, and Other Relief, United States v. Upromise, Inc., No. 17-10442 (D. Mass. Mar. 16, 2017), https://www.ftc.gov/system/files/documents/cases/upromise_complaint_stamped.pdf. Upromise agreed to undertake additional remedial measures and pay a $500,000 civil penalty to resolve the 2017 action. Stipulated Order for Permanent Injunction and Civil Penalty Judgment, United States v. Upromise, Inc., No. 17-10442 (Mar. 23, 2017), https://www.ftc.gov/system/files/documents/cases/upromise_order_-_3-23-17.pdf.

256. Complaint ¶¶ 15, 17(c), Compete, Inc., FTC Docket No. C-4384 (Feb. 20, 2013), 2013 WL 8364898 (charging section 5 violations following inadvertent collection of sensitive consumer data where automated filters failed).

257. *Id.* ¶ 17(c).

258. Even taking the FTC alone, section 5 represents only one data protection tool in the agency's enforcement arsenal. In addition to the FTC Act, the FTC has jurisdiction over sectoral data protection laws that may be implicated by dark data, such as the Gramm-Leach-Bliley Act (GLBA), which covers defined financial institutions. Pub. L. No. 106-102, 113 Stat. 1338 (1999) (codified in scattered sections of 12 U.S.C. (2000) and 15 U.S.C. (2000)). *See, e.g.*, Complaint ¶¶ 8, 16–17, 22, Premier Cap. Lending, Inc., FTC File No. 072-3004, 2008 WL 4892987 (Nov. 6, 2008) (following data breach, alleging that defendant violated the GLBA by providing a mortgage company with credit data access without assessing its cyber controls or requiring an inventory of the sensitive data it retained).

industry offer a timely example of the deepening emphasis placed on cybersecurity in particular. To mitigate the risk of debilitating cyberattacks that lock or destroy financial data, a group of U.S. financial institutions has launched a "doomsday project" dubbed Sheltered Harbor, which uses encrypted data vaults to prevent cyberattacks from spreading financial panic.[259] Sheltered Harbor has been compared "to seed banks, the Arctic vaults where governments keep basic material for agriculture, to be accessed in case [of] a nuclear attack."[260] The analogy's dire tone underscores the outsized importance cybersecurity has obtained in recent years.[261]

The public sector is also increasingly focused on cyber risks. Heightened concern with systemic risks to global markets[262] and a seemingly-endless parade of data breaches[263] has raised cybersecurity's profile at all levels of

259. Telis Demos, *Banks Build Line of Defense for Doomsday Cyberattack: The Sheltered Harbor Project is Meant to Ensure that Every U.S. Bank Has a Protected, Unalterable Backup that Can be Used to Serve Customers in Case of a Major Hack*, WALL ST. J. (Dec. 3, 2017), https://www.wsj.com/articles/banks-build-line-of-defense-for-doomsday-cyberattack-1512302401. For more information about Sheltered Harbor, see *Sheltered Harbor Fact Sheet*, FS-ISAC (Oct. 2018), https://shelteredharbor.org/images/ShelteredHarbor/Documents/Sheltered-Harbor-Fact-Sheet-2018-10-24.pdf.

260. Demos, *supra* note 259 (paraphrasing remarks by Sheltered Harbor CEO Steven Silberstein).

261. *See* OFF. OF FIN. RES., CYBERSECURITY AND FINANCIAL STABILITY: RISKS AND RESILIENCE 1, 6–7 (2017), https://www.financialresearch.gov/viewpoint-papers/files/OFRvp_17-01_Cybersecurity.pdf (describing cybersecurity incidents as "a key threat to financial stability" and discussing Sheltered Harbor as one response).

262. *SEC Chairman Clayton Issues Statement on Cybersecurity*, U.S. SEC. AND EXCH. COMM'N (Sept. 20, 2017), https://www.sec.gov/news/press-release/2017-170 (quoting SEC Chairman Jay Clayton as stating, '"Cybersecurity is critical to the operations of our markets and the risks are, . . . in many cases, systemic.'").

263. *See* Taylor Armerding, *The 18 Biggest Data Breaches of the 21st Century*, CSO (Dec. 20, 2018, 5:01 AM), https://www.csoonline.com/article/2130877/data-breach/the-biggest-data-breaches-of-the-21st-century (identifying major cyber breaches).

government, resulting in sharpened federal attention[264] and the introduction of more than 240 state cybersecurity bills and resolutions in 2017.[265]

In addition, increased awareness of cybersecurity's implications for digital privacy—laid bare by the 2017 Equifax consumer data breach[266]—and louder clamoring for a "right to be left alone" in a world of ubiquitous data collection, have produced a political environment primed for new privacy laws. A framework for what may come is rising across the Atlantic, in the form of the European Union's newly-implemented General Data Protection Regulation (GDPR).[267]

The GDPR establishes a number of data-protection and data-privacy measures that reach U.S. data processors[268] and controllers[269] that conduct business in the European Union (EU), target EU customers, or retain EU customer data.[270] In addition to its broad cybersecurity mandates that include data encryption and post-incident restoration,[271]

---

264. The financial institution and defense contracting sectors have recently seen significant federal cybersecurity efforts. In 2016, federal financial regulators released proposed rules for enhanced cybersecurity standards for certain entities with total consolidated assets of $50 billion or more. Enhanced Cyber Risk Management Standards, 81 Fed. Reg. 74,315 (Oct. 26, 2016) (to be codified at 12 C.F.R. pt. 30, 12 C.F.R. Chap. II, 12 C.F.R. pt. 364). Similarly, the Department of Defense finalized a rule establishing cybersecurity standards and requiring defense contractors to report network penetrations. Defense Federal Acquisition Regulation Supplement: Network Penetration Reporting and Contracting for Cloud Services, 81 Fed. Reg. 72,986 (Oct. 21, 2016) (to be codified at 48 C.F.R. pts. 202, 204, 212, 239, and 252).

265. *Cybersecurity Legislation 2017*, NAT'L CONF. OF STATE LEGISLATURES (Dec. 29, 2017), http://www.ncsl.org/research/telecommunications-and-information-technology/cybersecurity-legislation-2017.

266. The 2017 Equifax breach resulted in the release of sensitive data concerning 143 million American consumers. Seena Gressin, *The Equifax Data Breach: What to Do*, FED. TRADE COMM'N (Sept. 8, 2017), https://www.consumer.ftc.gov/blog/2017/09/equifax-data-breach-what-do.

267. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation), 2016 O.J. (L119) 1 [hereinafter GDPR].

268. A "processor" under the GDPR "means a natural or legal person, public authority, agency or other body which processes personal data on behalf of the controller." *Id.* art. 4(8), 2016 O.J. (L119) at 33.

269. A "controller" under the GDPR is, in relevant part, "the natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data." *Id.* art. 4(7), 2016 O.J. (L119) at 33.

270. ALLEN & OVERY, PREPARING FOR THE GENERAL DATA PROTECTION REGULATION 5 (2018), http://www.allenovery.com/SiteCollectionDocuments/Radical%20changes%20to%20European%20data%20protection%20legislation.pdf; *see also* Martin James, *7 Steps to GDPR for US Companies*, INFO. WEEK (July 4, 2017, 7:00 AM), https://www.informationweek.com/strategic-cio/security-and-risk-strategy/7-steps-to-gdpr-for-us-companies/a/d-id/1329235?.

271. *See* GDPR, *supra* note 267, art. 32, 2016 O.J. (L119) at 51–52.

the GDPR greatly expands the rights of "data subjects"[272] by codifying a right to be forgotten, which allows people to demand erasure of their personal data in certain circumstances.[273] The GDPR also provides data subjects with the right to restrict certain processing of their personal data,[274] and to demand that data processors provide them with their data in "a structured, commonly used . . . format," facilitating "data portability."[275] Data processors must also implement default measures that limit the processing of personal data.[276] Breach notification requirements are also established.[277] Penalties for violating the GDPR are significant, and can reach the greater of twenty million Euros or up to four percent of a company's "worldwide annual turnover."[278]

Dark data creates severe risks under the GDPR.[279] The GDPR reflects the assumption that data processors know a great deal about the data they collect. It also demands that they be sufficiently nimble to quickly erase, limit the use of, and produce data in readable form.[280] Under the GDPR, "[t]he invisible file" now "could potentially cost an organization millions if not managed properly or removed appropriately on request."[281] While most early GDPR compliance efforts have focused on structured data, the law is not limited "to data that might be in a structured format, it applies to *all* data."[282] As a result, the GDPR—and any subsequent legal regimes that adopt its broad digital privacy and

---

272. A "data subject" under the GDPR is "an identified or identifiable natural person." *Id.* art. 4(1), 2016 O.J. (L119) at 33.

273. *Id.* art. 17, 2016 O.J. (L119) at 43–44.

274. *Id.* art. 18, 2016 O.J. (L119) at 44–45.

275. *Id.* art. 20, 2016 O.J. (L119) at 45.

276. *Id.* art. 25(2), 2016 O.J. (L119) at 48.

277. *Id.* art. 33, 2016 O.J. (L119) at 52.

278. *Id.* art. 83(5), 2016 O.J. (L119) at 82–83.

279. *See, e.g.*, Sam Jefferies, *The Three Types of Data Putting Law Firms at Risk*, LEGAL FUTURES (July 9, 2018, 12:00 AM), https://www.legalfutures.co.uk/blog/the-three-types-of-data-putting-law-firms-at-risk ("Dark data is a serious threat to GDPR compliance . . . . Failing to provide all the information because the documents were undiscoverable can lead to costly disputes, drawn-out negotiations, and financial penalties."); *see also* Mike Pannell, *GDPR Keeps Us All Awake at Night—It's High Time to Get Our Sleep Sorted*, PUBLICTECHNOLOGY (May 2, 2018), https://publictechnology.net/articles/partner_article/bt/gdpr-keeps-us-all-awake-night-%E2%80%93-it%E2%80%99s-high-time-get-our-sleep-sorted (explaining that many organizations seeking to comply with the GDPR "are focusing their attention on their obvious data, but neglecting the less well controlled information," and that "[u]nstructured and [d]ark [d]ata can equally contain personal data" subject to the GDPR).

280. Holder, *supra* note 105 (noting under the GDPR, organizations "need to fully understand the data [they] hold, or at least be able to quickly produce such data if requested").

281. *Id.*

282. *Id.*

data security tenets—poses a monumental challenge to organizations with caches of dark data.

III. DECISION DISTORTION

Thus far, this Article has described dark data as a source of invisible risk for organizations. This Section will address a different problem: dark data's ability to quietly distort the promised completeness, accuracy, and objectivity of Big Data-driven evidence used in judicial proceedings.

The slow-grinding nature of legal change has often cast the U.S. court system as "a crude and belated tool" in the face of rapid technological development.[283] It may therefore come as a surprise that the legal system has provided fertile ground for Big Data methods.[284] Much of law's initial embrace of Big Data technologies is born of necessity, as the massive volumes of data that have come to define modern litigation demand technological solutions that far exceed prior methods of discovery management.[285] The advent of giant datasets has sparked a more than $10 billion e-discovery industry that uses Big Data analytics to parse and make sense of financial data, emails, and other digital material.[286] Reaching the discovery stage in big league litigation can require using numerous computers to analyze terabytes of data.[287]

In criminal law, Big Data's influence is spreading from data-driven, predictive policing in America's streets[288] to matters of fairness and freedom in its courtrooms. Judges now commonly use predictive

283. Tene & Polonetsky, *Social Norms*, *supra* note 114, at 73.

284. John O. McGinnis & Russell G. Pearce, *The Great Disruption: How Machine Intelligence Will Transform the Role of Lawyers in the Delivery of Legal Services*, 82 FORDHAM L. REV. 3041, 3052 (2014) ("Predictive analytics is now coming to law. Indeed, law, with its massive amounts of data from case law, briefs, and other documents, is conducive to machine data mining that is the foundation of this new predictive science.").

285. *See* Kenneth J. Withers, *Electronically Stored Information: The December 2006 Amendments to the Federal Rules of Civil Procedure*, 4 NW. J. TECH. & INTELL. PROP. 171, 173–74, 176 (2006) (discussing litigation challenges of electronically-stored information).

286. *$17.3 Billion eDiscovery Market—Global Forecast to 2023*, PR NEWSWIRE (June 19, 2018, 1:15 PM), https://www.prnewswire.com/news-releases/17-3-billion-ediscovery-market—global-forecast-to-2023–300668568.

287. Mike DeCesaris, *Using Big Data in Gathering Expert Testimony*, LAW360 (July 1, 2015, 12:12 PM), https://www.cornerstone.com/Publications/Articles/Using-Big-Data-in-Gathering-Expert-Testimony.

288. *See* Elizabeth E. Joh, *The New Surveillance Discretion: Automated Suspicion, Big Data, and Policing*, 10 HARV. L. & POL'Y REV. 15, 16–19 (2016) (discussing the impacts of new technologies on police decisions regarding suspect identification and monitoring); *see also* Ferguson, *supra* note 1, at 331 (noting the influence of big data on law enforcement's development of reasonable suspicion).

analytics to set bail and make pretrial release decisions.[289] Big Data risk assessment models are widely used in sentencing, where data concerning criminal history and offense characteristics drive conclusions about recidivism risks.[290] Recently in *State v. Loomis*,[291] the Wisconsin Supreme Court upheld the use of an algorithmic risk assessment tool in criminal sentencing, rejecting the defendant's due process challenge to facing six years imprisonment based on proprietary computer code that was disclosed neither to the defendant nor the Court.[292] Wisconsin is not alone in using data-driven risk assessments in sentencing; several states even require them.[293]

While the use of Big Data analytics to inform legal decision-making is not inherently problematic, much can go wrong. In particular, the widening divide between stockpiled dark data and analytical tools that can make sense of it ensures that the contents of many datasets remain largely opaque to database operators. While datasets that are narrow by design can help isolate relevant data, improve query hits, and reduce search costs, concern is warranted when conclusions are derived from databases rife with dark and possibly legally-relevant data that is invisible or inaccessible to the operator. And if dark data can confound technologists armed with advanced analytical tools, how can judges and juries assess—let alone accurately assess—the reliability and quality of Big Data-derived digital evidence? The temptation to replace independent assessment with proxies based on assumed technical prowess may be difficult to resist.

This is all the more troubling because conclusions are not treated equally. Conclusions born of Big Data are infused with a credibility steeped in assumed objectivity and omnipotence that quintessentially "human" processes cannot match.[294] This credibility can make incomplete or erroneous conclusions easier to accept and harder to question.

---

289. Devins et al., *supra* note 5, at 396–97; Christin et al., *supra* note 17, at 1–3.

290. Devins et al., *supra* note 5, at 396–97.

291. 881 N.W.2d 749 (Wis. 2016).

292. *Id.* at 753; *Wisconsin Supreme Court Requires Warning Before Use of Algorithmic Risk Assessments in Sentencing* —State v. Loomis*, 881 N.W.2d 749 (Wis. 2016)*, 130 HARV. L. REV. 1530, 1530–33 (2017) [hereinafter *HLR on Loomis*] (describing the Court's advisement that an algorithmic risk assessment cannot be the sole basis for sentencing).

293. *HLR on Loomis*, *supra* note 292, at 1536 (identifying state statutes that provide for algorithmic risk assessments in criminal sentencing).

294. *See* Devins et al., *supra* note 5, at 362 (likening Big Data to the "mythical omniscient actor from rational choice theory"); *see also* Ric Simmons, *Quantifying Criminal Procedure: How to Unlock the Potential of Big Data in Our Criminal Justice System*, 2016 MICH. ST. L. REV. 947, 950 (2016) (claiming that in the criminal justice context, Big Data offers "the promise of increased fairness and greater objectivity . . . ."); Paula

By exposing the relative smallness of large datasets and anchoring advanced analytics to the unshakable risk of "missing the needle," dark data lays bare a basic problem with Big Data: purportedly fact-inclusive, all-seeing conclusions can be incomplete or wrong. Worse yet, they may result from subjective choices about database framing and construction, contradicting avowed freedom from human bias. The choice of what raw data to feed an algorithm—and what data to leave dark and unanalyzed—is a decision that will ultimately affect the algorithm's conclusions, regardless of how objective its methods may otherwise be.

A. Big Data's Legal Appeal

Certain aspects of Big Data are naturally appealing to lawyers and judges. As some scholars have remarked, "[l]egal tradition prizes consistency, stability, and uniformity in legal rules," and "Big Data promises . . . a scientific and evidence-based approach to law."[295] This approach resonates with the law's technocratic influences. For instance, behavioral law and economics—which uses "evidence-based, nudge-related and objective approaches" to improve legal decision-making—melds easily with data-driven rationality.[296] Or, as Richard Posner has argued, improving accuracy and limiting randomness in criminal proceedings enhances law's influence by making deterrence more effective.[297] More broadly, common law outcomes revolve around stare decisis, which transcends any particular case to advance "the broader societal interests in evenhanded, consistent, and predictable application of legal rules."[298]

Big Data is also appealing to lawyers and judges because it is perceived to be "indifferent" and "can be comprehensive in scope

---

Dantas, *The Future of Justice is Watson*, IBM BIG DATA & ANALYTICS HUB (Oct. 19, 2015), http://www.ibmbigdatahub.com/blog/future-justice-watson (claiming that while Watson, an IBM cognitive computing system, "does make errors, . . . its errors are random. Watson is not capable of making systematic errors based on political ideology, gender, upbringing or any number of factors that can creep into legal decisions made by humans").

295. Devins et al., *supra* note 5, at 358.

296. *Id.* at 363. For more on behavioral law and economics, see Christine Jolls, Cass R. Sunstein & Richard Thaler, *A Behavioral Approach to Law and Economics*, 50 STAN. L. REV. 1471 (1998).

297. Richard A. Posner, *An Economic Approach to the Law of Evidence*, 51 STAN. L. REV. 1477, 1483 (1999) ("The more accurate the process of determining guilt is, the less random punishment will be, and so the greater will be the law's deterrent effect.").

298. Thomas v. Wash. Gas Light Co., 448 U.S. 261, 272 (1980).

where lawyers are limited in experience."[299] By cleansing decision-making of subjectivity, Big Data promises "reality, unfiltered."[300] To take an example from the trial context, replacing subjective eyewitness testimony with the objectivity of data analytics would seem to obviate the trial as a stage upon which to present "different views of reality in a manner designed to produce a functional set of conclusions about what happened."[301] With Big Data analytics, differing accounts of what happened are no longer meaningful; instead, the data reveals what *actually did* happen. Rare is the decision-maker who is not enticed by the promise of "results with greater truth, objectivity, and accuracy."[302]

But for all of Big Data's appeal to legal deciders, data-driven outcomes may still prompt anxiety when applied by a judicial system charged with promoting intangible values like fairness and justice. A basic response to fears of overly-technocratic outcomes might be that prosecutors and judges "do not blindly follow the results provided by algorithms," but will consider them in light of their independent knowledge and experience.[303] Indeed, as the Wisconsin Supreme Court explained in *Loomis*, the trial court considered the algorithmic risk assessment at issue "along with other supporting factors" that justified the sentence handed down.[304] The algorithm was not outcome-determinative.[305]

Yet, while discretion and independent judgment remain, cognitive and behavioral research reveals that it is "psychologically difficult and rare to 'override' the recommendations provided by an algorithm," and in fact "judges and prosecutors are likely to follow the predictions provided by risk-assessment tools."[306] Like other people, judges and prosecutors may sense that algorithmic conclusions "generally seem[] more reliable, scientific, and legitimate than other sources of

299. Dru Stevenson & Nicholas J. Wagoner, *Bargaining in the Shadow of Big Data*, 67 FLA. L. REV. 1337, 1346 (2015).

300. Julie E. Cohen, *What Privacy is f or*, 126 HARV. L. REV. 1904, 1921 (2013).

301. Charles Nesson, *The Evidence or the Event? On Judicial Proof and the Acceptability of Verdicts*, 98 HARV. L. REV. 1357, 1389 (1985).

302. Crawford & Schultz, *supra* note 6, at 96.

303. Christin et al., *supra* note 17, at 7.

304. State v. Loomis, 881 N.W.2d 749, 765 (Wis. 2016).

305. *Cf.* Simmons, *supra* note 294, at 954 (discussing a scenario where "the results from big data's predictive algorithms could be outcome determinative, meaning that a police officer or a judge would only consider the algorithm's output and ignore all other evidence").

306. Christin et al., *supra* note 17, at 7.

information, including one's feelings about an offender."[307] As will be seen, the natural inclination to follow the algorithm underscores the need for heightened caution when weighing Big Data-derived evidence.

B. Gatekeepers and Fact Finders

A central problem for courts confronted with Big Data evidence is that it may appear as inscrutable as it is appealing. Detailed understandings of scientific and technical methods are not the purview of judges. As Justice Breyer quipped, "[a] judge is not a scientist, and a courtroom is not a scientific laboratory."[308] Even so, scientific and technical evidence is often the currency of truth in complex cases. Recognizing as much, the Supreme Court in *Daubert v. Merrell Dow Pharmaceuticals, Inc.*[309] determined that Federal Rule of Evidence 702 "confides to the judge some gatekeeping responsibility in deciding questions of the admissibility of proffered expert testimony."[310] *Daubert* calls on federal judges to assess at the outset "the scientific validity—and thus the evidentiary relevance and reliability—of the principles that underlie [the] proposed submission."[311] Evidence that fails this test is not presented to fact finders.[312]

Performing the *Daubert* gatekeeping function has always been difficult.[313] The use of Big Data methods in the courtroom makes the task significantly more so. As an empirical practice, Big Data may take

---

307. *Id.*; s*ee also* Ben Dickson, *Artificial Intelligence Has a Bias Problem, and It's Our Fault*, PC MAG. (June 14, 2018, 8 PM), https://uk.pcmag.com/features/96336/artificial-intelligence-has-a-bias-problem-and-its-our-fault ("Under the illusion that AI is cold, mathematical calculation devoid of prejudice or bias, humans may tend to trust algorithmic judgment without questioning it."); Deirdre K. Mulligan, Remarks at the FTC Fintech Forum: Artificial Intelligence and Blockchain 4 (Mar. 9, 2017), https://www.ftc.gov/system/files/documents/public_events/1051963/ftc_fintech_forum_ai_and_blockchain_-_deirdre_k_mulligan_transcript.pdf ("[W]e know from research that when a machine tells us something, people are far less likely to question, to ask about its priors, or its limits, or its background assumptions. And therefore it is taken as kind of ground truth.").

308. Hon. Stephen Breyer, *Introduction* to FED. JUDICIAL CTR., REFERENCE MANUAL ON SCI. EVID. 4 (3d ed. 2011), https://www.fjc.gov/sites/default/files/2015/SciMan3D01.pdf.

309. 509 U.S. 579 (1993).

310. *Id.* at 600 (Rehnquist, C.J., concurring in part and dissenting in part).

311. *Id.* at 593–95 (setting forth factors a court may consider in evaluating the admissibility of scientific evidence: (1) whether the method used to identify the evidence is based on a testable hypothesis; (2) whether the method is peer-reviewed and published; (3) the method's rate of error; (4) the existence and maintenance of standards controlling the method's operation; and (5) whether the method is generally accepted within the scientific community).

312. *See id.* at 591–92.

313. Breyer, *supra* note 308, at 4.

a "scientific approach,"[314] but it often eschews the scientific method.[315] By allowing patterns and connections drawn from large datasets to stand alone, Big Data often dispenses with the process of asking questions, forming hypotheses, and testing results.[316] As Verizon explains, "[w]e don't necessarily need to know how to ask a question or which data items we need to query; instead, we rely on algorithms that find answers in very large data stores."[317] And those answers, by themselves, are viewed as sufficient—correlation can now stand on its own, regardless of causation.[318] In other words, "[w]ho knows why people do what they do? The point is they do it, and we can track and measure it with unprecedented fidelity."[319]

Of course, this correlation-only approach is at odds with the judicial system's treatment of evidence. Can judges adequately assess the reliability of evidence derived from an algorithm that has supplied a *what* (correlation) with no corresponding *why* (causation)? How can a fact finder ultimately weigh such evidence? More generally, is such evidence even relevant when assessing a legal charge steeped in causation? After all, "[d]efendants are incarcerated, and indeed put to death, because their actions 'caused' a particular consequence."[320]

While a technical process need not "exactly mirror the fundamental precepts of the so-called harder sciences" to be reliable under *Daubert*, experts must nonetheless be able "to test the underlying hypotheses and review the standards controlling the technique's operation in an attempt to reproduce the results originally generated."[321] Without a hypothesis that can be tested and reproduced, any assessment of reliability must demand, at the very minimum, a detailed understanding of how an algorithm identified a given correlation.

---

314. Devins et al., *supra* note 5, at 358.

315. *See* Chris Anderson, *The End of Theory: The Data Deluge Makes the Scientific Method Obsolete*, WIRED (June 23, 2008, 12:00 PM), https://www.wired.com/2008/06/pb-theory (advocating replacing the traditional scientific method with correlative relationships extracted from large datasets).

316. *See id.*; *see also* Mattioli, *supra* note 5, at 541 ("[B]ig data draws insights from records gathered automatically and indiscriminately a priori.").

317. VERIZON, HOW TO THRIVE ON THE FRONTIERS OF DATA 3 (2014), http://www.verizonenterprise.com/resources/whitepaper/wp_thriving-frontiers-of-data_en_xg.pdf.

318. Anderson, *supra* note 315; Kenneth Neil Cukier & Viktor Mayer-Schoenberger, *The Rise of Big Data*, FOREIGN AFF. (2013), https://www.foreignaffairs.com/articles/2013-04-03/rise-big-data ("Big data helps answer what, not why, and often that's good enough.").

319. Anderson, *supra* note 315.

320. King & Mrkonich, *supra* note 20, at 563.

321. Elcock v. Kmart Corp., 233 F.3d 734, 747 (3d Cir. 2000).

This is no easy task. Big Data algorithms powered by artificial intelligence—particularly of the neural network and deep learning varieties—can possess the unique ability to "program[] themselves . . . in ways we cannot understand."[322] Rather than operating within the bounds of predefined and transparent rules, biology-inspired neural networks may learn organically, producing decision processes that defy human understanding.[323] This opacity confounds not only those lacking deep technical knowledge, as even the creators of some algorithms "cannot fully explain their behavior."[324]

Consequently, it may not always be possible to determine the precise methodology that an artificial intelligence system has used to arrive at a conclusion, let alone whether the resulting evidence is sufficiently reliable to reach the jury under *Daubert*. Assessing reliability becomes more difficult still in the case of artificial intelligence algorithms created by non-parties, which often shield the inner workings of their creations behind intellectual property and trade secret laws.[325] Subsequently, if evidence generated by these proprietary algorithms is submitted to a jury, critical voices may charge that such evidence "is not provided to educate . . . it is offered as a conclusion to be deferred to by the fact finder."[326]

Gatekeepers and fact finders have good reason to be vigilant in considering evidence derived from Big Data. Taken together, Big Data's natural appeal to legal decision-makers and the frequent

---

322. Will Knight, *The Dark Secret at the Heart of AI*, MIT TECH. REV. (Apr. 11, 2017), https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai.

323. *Id.*; *see also* Heidi Vogt, *Artificial Intelligence Rules More of Your Life. Who Rules AI?*, WALL ST. J. (Mar. 13, 2018), https://www.wsj.com/articles/artificial-intelligence-rules-more-of-your-life-who-rules-ai-1520933401 (discussing neural networks and other artificial intelligence systems that derive conclusions through methods that may defy human understanding).

324. Knight, *supra* note 322.

325. *See* Richards & King, *supra* note 6, at 42 ("[W]hile big data pervasively collects all manner of private information, the operations of big data itself are almost entirely shrouded in legal and commercial secrecy."); *see also* Omer Tene & Jules Polonetsky, *Judged by the Tin Man: Individual Rights in the Age of Big Data*, 11 J. TELECOMM. & HIGH TECH. L. 351, 361 (2013) [hereinafter Tene & Polonetsky, *Tin Man*] ("[T]he machine is covered by an opaque veil of secrecy, which is backed by corporate claims of trade secrecy and intellectual property."); Eric Van Buskirk & Vincent T. Liu, *Digital Evidence: Challenging the Presumption of Reliability*, 1 J. DIGITAL FORENSIC PRAC. 19, 23 (2006) (explaining that proprietary source code is often unavailable for inspection in legal proceedings).

326. Ronald J. Allen, *Fiddling While Rome Burns: The Story of the Federal Rules and Experts* 2 (Nw. Pub. L. Res. Paper No. 17-29, 2017), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3080628.

inscrutability of its methods produce a significant risk of masking underlying errors and biases. Contrary to assertions of omnipotent visibility and objectivity through data-centric analysis freed from human bias, Big Data conclusions can be both wrong and subjective. As the following pages address, dark data can be a significant driver of Big Data error, and a key indicator of hidden subjectivity.

In particular, the presence of dark data can challenge claims of Big Data omnipotence by creating digital blind spots. Supposed all-encompassing conclusions may unknowingly omit relevant data, distorting analysis and producing limited or inaccurate legal outcomes. Even worse, dark data can mask unseen biases concealed within Big Data methods. As dark data can result not only from technological limitations but also from conscious choices about database construction and analysis, its presence may reveal subjective framing decisions that can steer outcomes in undisclosed directions.

C. The "N=All" Myth

Central to Big Data's appeal is the belief that it produces knowledge from an "N=All" position, in which analytical tools draw conclusions from all or nearly all relevant data points.[327] Sampling becomes obsolete if N=All, as algorithms can generate results that account for "the entire background population."[328] Dispensing with sampling is assumed to remove both sampling error and subjective bias, allowing data to speak for itself, resulting in more accurate conclusions.[329]

The rising presence of dark data throws cold water on the N=All assumption. Large datasets often contain a relatively small sliver of structured data that common Big Data tools can digest.[330] Big Data literature, while painting a picture of omnipotence and fact-inclusiveness,[331] often glosses over the presence of unstructured and dark data when describing datasets and analytical methods. As Amir Gandomi and Murtaza Haider explain, "[t]he popular discourse on big data, . . . focuses on predictive analytics and structured data. It ignores the largest component of big data, which is unstructured . . . ."[332]

---

327. Tim Harford, *Big Data: Are We Making a Big Mistake?*, FIN. TIMES (Mar. 28, 2014), https://www.ft.com/content/21a6e7d8-b479-11e3-a09a-00144feabdc0.

328. *Id.*; *see also* Cukier & Mayer-Schoenberger, *supra* note 318 (explaining that Big Data allows users to "collect and use a lot of data rather than settle for small amounts or samples").

329. Harford, *supra* note 327.

330. *See* Babcock, *supra* note 34.

331. *See* Devins et al., *supra* note 5, at 359–62, 371–72.

332. Gandomi & Haider, *supra* note 32, at 137.

Ignoring dark data can undermine results and lead to errors in decision-making. To bring dark data into the analytical fold, Stanford University's DeepDive system organizes unstructured dark data into SQL databases that can be queried and analyzed with common Big Data tools.[333] DeepDive is motivated by the awareness that "[d]ark data often holds information that is not available in any other format," and that the failure to consider dark data can influence results.[334] As "success in analytical tasks is often limited by the data available, the information embedded in dark data can be massively valuable."[335]

As the DeepDive programmers realize, leaving information within dark data invisible can create profoundly negative consequences for decision-making. Long before the Big Data era, the criminal justice system has grappled with the most tragic of errors: the conviction of innocent persons for crimes they did not commit.[336] Wrongful convictions have been revealed through DNA evidence of innocence—essentially, dark data at trial—long after criminal defendants have been sentenced and incarcerated.[337]

Erroneous forensic evidence has been responsible for numerous wrongful convictions,[338] "demonstrat[ing] the potential danger of giving undue weight to evidence and testimony derived from imperfect testing and analysis."[339] Mirroring claims of Big Data's omnipotence, assertions that flawed forensic science processes "employ methodologies that have perfect accuracy and produce no errors . . . hampered efforts to evaluate the usefulness of the forensic science disciplines," contributing to wrongful convictions.[340]

---

333. Christopher De Sa et al., *DeepDive: Declarative Knowledge Base Construction*, 45 SIGMOD Rec. 60, 60, 64 (2016), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5361060/pdf/nihms826683.pdf; Ce Zhang et al., *Extracting Databases from Dark Data with DeepDive*, SIGMOD (2016), https://cs.stanford.edu/people/chrismre/papers/modiv923-zhangA.pdf.

334. Zhang et al., *supra* note 333.

335. *Id.*

336. *See* Garrett & Neufeld, *supra* note 45, at 76 (identifying wrongful conviction cases, including those resulting from the failure to disclose exculpatory evidence).

337. *See id.* at 76 n.244 (explaining that evidence "did not surface until after post-conviction DNA testing, post-exoneration investigations, or civil suits").

338. Michael J. Saks, *Scientific Evidence and the Ethical Obligations of Attorneys*, 49 CLEV. ST. L. REV. 421, 424 tbl. 2 (2001).

339. Mark A. Godsey & Marie Alou, *She Blinded Me with Science: Wrongful Convictions and the "Reverse CSI-Effect,"* 17 TEX. WESLEYAN L. REV. 481, 491 (2011) (quoting NAT'L ACAD. OF SCI., NAT'L RES. COUNCIL OF THE NAT'L ACADS., STRENGTHENING FORENSIC SCIENCE IN THE UNITED STATES: A PATH FORWARD 4 (2009) [hereinafter NAT'L ACAD. OF SCI.], https://www.ncjrs.gov/pdffiles1/nij/grants/228091.pdf.

340. NAT'L ACAD. OF SCI., *supra* note 339, at 47.

The wrongful conviction cases of the past are a warning for the Big Data era. Unless care is taken to ensure that dark data likely to contain relevant evidence is identified, analyzed, and presented to fact finders, the unjust results of the past will reoccur in digital form. Effort must be made to ensure that claims of Big Data's supposed fact-inclusiveness do not produce similar error-insulating effects in the digital era.

The criminal law context offers endless scenarios where dark data may conceal exonerating digital evidence. In their call for a new conception of "digital innocence," Joshua A.T. Fairfield and Erik Luna argue that extensive data storage and vast databases, while serving prosecutors, also "guarantee the existence of exonerating evidence, stored somewhere, proving the innocence of suspects and defendants."[341] For example, geolocated and time-stamped social media data might support a defendant's alibi.[342] "Smart home" entry and exit data might prove that a defendant spent insufficient time within a residence to have committed an alleged crime.[343] Web traffic data may show that an illegal insider stock tip did not originate from a defendant's computer.[344]

Of course, timing is everything. While insights may exist within vast databases, it may not be possible to presently extract them.[345] This is especially true where exonerating insights are buried within dark data, which can elude the reach of existing analytical tools.[346] Troublingly, while dark data may render significant portions of large datasets effectively invisible, the assumed completeness and accuracy of Big Data-derived conclusions may nonetheless remain unquestioned.[347] The inherent opacity of dark data and the high appeal of Big Data methods may mask digital blind spots, causing incomplete and erroneous decisions to escape serious challenge.[348]

D. Subtle Subjectivity

In addition to producing decision errors, dark data can reveal cracks in Big Data's vaunted objectivity. Analytical tools are often assumed to

---

341. Joshua A.T. Fairfield & Erik Luna, *Digital Innocence*, 99 CORNELL L. REV. 981, 986 (2014).

342. *Id.* at 1072.

343. *See id.* at 991 (describing the potential for defense counsel to cross reference time and geolocation information).

344. *See id.* at 1001–02 (highlighting the increased capability of comprehensive browser-history tracking).

345. *See id.* at 1072 ("[D]iscovery may have to wait until databases become big enough or connected enough for meaningful analysis.").

346. *See supra* Section II.A.

347. *See supra* Section III.A.

348. *See* Devins et al., *supra* note 5, at 358, 365–67 (describing Big Data's appeal to legal decision-makers).

be disinterested distillers, mining vast datasets for naturally occurring connections and patterns.[349] Big Data processes are believed to be neutral extractors of insight, rather than creators driven by subjective impressions of what should be found.[350] This supposed objectivity is believed to enhance decisional accuracy by removing bias and allowing conclusions to be drawn from the data alone.[351]

Big Data's celebrated objectivity is vulnerable to the critique that it is little more than an illusion. While "[d]ata crunching may appear to be an exact science," in actuality it is "laden with subjective input from researchers who decide which data to analyze, questions to examine, and purposes to pursue."[352] Seemingly objective processes may conceal "the hidden assumptions of the programmers and policymakers . . . about which scientific theories are valid, what data should be considered, and what level of error is acceptable."[353]

Dark data provides additional ammunition for critiques of Big Data objectivity. The presence of dark data, while virtually inevitable,[354] may reveal subjective machinations behind seemingly objective Big Data methods. Dark data is often the product of technological limitations, such as the case of an organization that lacks the analytical tools necessary to interpret the dark data it has collected.[355] In other situations, however, dark data is brought to light or is left functionally invisible because of deliberate, subjective choices.[356]

In addressing institutional effects that can produce dark data, one commenter described data management practices as "almost completely subjective" because they rely on personalized schemas tailored to individuals' particular needs.[357] As this observation reveals, choosing to lighten particular dark data, and deciding how to organize dark data that has been lightened, requires subjective determinations about relevance, categorization, and labeling.[358] These determinations

---

349. *See supra* note 21 and accompanying text.

350. *See* Cohen, *supra* note 300, at 1921.

351. *See* Crawford & Schultz, *supra* note 6, at 96.

352. Tene & Polonetsky, *Tin Man*, *supra* note 325, at 353.

353. Andrea Roth, *Trial by Machine*, 104 GEO. L.J. 1245, 1250 (2016).

354. *See supra* notes 34 and 92 and accompanying text.

355. *See* Kambies et al., *supra* note 29, at 21 (describing technical hurdles for dark data analysis).

356. *See supra* Section I.B.4.

357. Alex Woodie, *Beware the Dangers of Dark Data*, DATANAMI (Aug. 18, 2015), https://www.datanami.com/2015/08/18/beware-the-dangers-of-dark-data.

358. *See* Alon Halevy et al., *The Unreasonable Effectiveness of Data*, 24 IEEE INTELLIGENT SYS. 8, 11 (2009) (explaining that Web tables "represent how different people organize

are not value-free and may bear on the ultimate conclusions drawn from datasets. For example, a database operator could decide to code certain data as "unrelated," causing it to be excluded from search queries. Conversely, categorizing the same data as "related" or "important" could have the effect of directing search queries to that data, affecting results.

While any effort to organize data will reflect the organizer's preferences, the effects of subjectivity in dark data management can run deeper. Resource constraints and organizational priorities collide with the costs of and technical barriers to lightening dark data, requiring initial choices about *which* dark data should be brought to light at all, and which should be left unanalyzed and effectively invisible. Indeed, "[c]lassification is an important step in separating the dark data worthy of illumination from the redundant, obsolete and trivial data."[359]

Classification thus reflects judgments,[360] which may or may not be benign. The difficulty is that a determination not to illuminate particular dark data can be used to exclude certain data from analysis altogether, thus affecting and potentially skewing results. Choosing to leave certain data in the dark can embed subjectivity more deeply than in the case of categorizing light data, which, while also organized according to the categorizer's interpretations, is nonetheless likely to remain accessible for independent review and analysis. Subsequent review is far more difficult in the case of unanalyzed dark data, which may simply remain in the void.[361]

Artificial intelligence systems that operate free from active human direction do not avoid subjectivity concerns.[362] Artificial intelligence algorithms are often tasked with "executing the instructions of human programmers based on data or material inputted by human operators."[363] As humans determine the input data that artificial intelligence systems ingest, subjectivity is built-in and inevitable, even

---

data—the choices they make for which columns to include and the names given to the columns").

359. Viewpointe White Paper, *supra* note 30, at 7.

360. *See* Halevy et al., *supra* note 358, at 11 (describing how a single expression can have multiple different meanings).

361. *See* Heidorn, *supra* note 28, at 281 (giving an example of scientific dark data as "exist[ing] only in the bottom left-hand desk drawer of scientists on . . . media that is quickly aging and soon will be unreadable").

362. *See* WILL HURD & ROBIN KELLY, COMM. ON OVERSIGHT AND GOV. REFORM, SUBCOMM. ON INFO. TECH., RISE OF THE MACHINES: ARTIFICIAL INTELLIGENCE AND ITS GROWING IMPACT ON U.S. POLICY 11 (2018), https://www.hsdl.org/?abstract&did=816362.

363. Roth, *supra* note 353, at 1270.

if active human interference ends once an algorithm is deployed.[364] Even if an artificial intelligence system could be designed to entirely ignore bias, the initial framing of input data could still skew results toward a desired outcome by omitting data likely to support or disprove a particular conclusion.[365] As "systems are designed to capture certain kinds of data," the decision to leave other data dark will affect an algorithm's conclusions by narrowing the background dataset from which patterns can be drawn, regardless of how objective the pattern-drawing process may be.[366]

E. The Need for Judicial Scrutiny

Technological limitations surrounding dark data analysis cannot justify avoiding the issue of dark data in the courtroom. First, the judicial system and its participants must recognize that dark data is a built-in constraint that precludes Big Data-derived conclusions from deserving the gloss of fact-inclusive omnipotence they often receive.[367] Acknowledging the limitations that dark data can place on the completeness of Big Data-derived evidence will reduce the likelihood that erroneous analysis is vested with the false confidence that contributed to past wrongful conviction cases.[368]

Second, courts can seek to ensure, to the extent possible in light of technological limitations, that dark data likely to contain relevant evidence has been identified, disclosed to defendants in criminal cases,[369] and presented to fact finders. Declining to scrutinize dataset composition and analytical methods risks hardwiring errors into the evidence process, jeopardizing the veracity of resulting legal conclusions. Courts and

---

364. *See, e.g.*, Dickson, *supra* note 307 (explaining that deep learning algorithms "can inherit covert or overt biases" from input data); *see also* Omer Tene & Jules Polonetsky, *Taming the Golem: Challenges of Ethical Algorithmic Decision-Making*, 19 N.C. J.L. & TECH. 125, 130 (2017) ("[E]ven without active human editorial intervention, no algorithm is fully immune from the human values of its creators . . . . Algorithms codify human choices about how decisions should be made."); Kitchin, *supra* note 14, at 5 ("Even if [a Big Data] process is automated, the algorithms used to process the data are imbued with particular values and contextualized within a particular scientific approach."); Mattioli, *supra* note 5, at 546 ("Data is often deeply infused with the subjective judgments of those who collect and organize it.").

365. Kitchin, *supra* note 14, at 5.

366. *Id.*

367. *See* Devins et al., *supra* note 5, at 401–02, 405–07 (listing Big Data limitations).

368. *See* NAT'L ACAD. OF SCI., *supra* note 339, at 4 (discussing imprecise or exaggerated expert testimony contributing to the admission of erroneous or misleading evidence).

369. *See* Brady v. Maryland, 373 U.S. 83, 87 (1963) (holding that the prosecution must disclose evidence that is material to guilt or punishment).

litigants must rigorously question Big Data-derived conclusions to ensure that relevant and exculpatory evidence has not been swept away.

Third, where objectivity is concerned, the point is not that subjectivity in data collection and analysis is inherently nefarious, but that Big Data methods are often falsely characterized as neutral when they are not.[370] False objectivity can lend subjective conclusions unwarranted deference in the courtroom.[371] To combat this distortion, courts must shed light on the collection and construction of datasets used to produce courtroom evidence, including by pressing litigants to identify data that has been left dark, as well as the reasons for that decision.

Finally, artificial intelligence systems that derive conclusions from seemingly inscrutable methods[372] should not be permitted to evade judicial inquiry in matters of evidence. In such situations, judges can exercise their *Daubert* function to demand that algorithms producing courtroom evidence be "'inspectable'" and '"able to explain [their] output.'"[373] While the adversary process will naturally promote this level of inquiry in many instances, other cases, particularly those involving indigent or pro se defendants, will call for an increased judicial role to ensure the reliability of Big Data evidence.[374]

CONCLUSION

The capability gulf between Big Data analytics and data storage technologies has produced a vast accumulation of retained dark data. The impulse to collect and store is understandable: there is little doubt that dark data can act as digital camouflage, concealing information of great value. And there is the possibility that artificial intelligence, blockchain, or other emerging technologies will ultimately out-engineer the dark data problem altogether, allowing hidden insights to be readily and economically extracted.[375]

---

370. *See* Ferguson, *supra* note 1, at 402 ("[L]ike other quantitative systems used for decisionmaking, big data-based predictive policing will appear to be objective and fair when it may in fact reflect subjective factors and structural inequalities.").

371. *See* Rieder & Simon, *supra* note 22, at 3 (describing the narrative that data is characterized by "trust, truth, and objectivity," contributing to the perceived neutrality of Big Data).

372. *See* Knight, *supra* note 322.

373. *See* HURD & KELLY, *supra* note 362, at 11 (quoting testimony from Charles Isbell).

374. *See* Section III.B. (discussing courts' gatekeeping role under *Daubert*).

375. While technology may eventually "solve" the dark data problem as a matter of analytics, such solutions are likely to raise important privacy and social concerns that will demand serious consideration.

But until such time, organizations must devote newfound attention to the invisible risks that may lie buried within their dark data. Managing the risks of dark data requires an initial awareness that beneath its gloss of omnipotence, Big Data technology currently peers into a very narrow slice of the digital universe. Failing to address dark data's digital blind spots not only risks liability under an expanding array of legal regimes, but can also result in false confidence being placed in incomplete or erroneous conclusions, inviting error.

Courts, too, must carefully examine Big Data-derived evidence to prevent high technology from cloaking omissions, errors, and subjective influences that can compromise the accuracy and fairness of judicial determinations. As Big Data assumes a greater role in American courtrooms, judges must ardently exercise their gatekeeping function to ensure that Big Data-derived evidence does not produce unjust outcomes for life and liberty.