PRIVACY AND SYNTHETIC DATASETS

Steven M. Bellovin, Preetam K. Dutta, and Nathan Reitinger
Columbia University, Department of Computer Science

ABSTRACT

Sharing is a virtue, instilled in us from childhood. Unfortunately, when it comes to big data—i.e., databases possessing the potential to usher in a whole new world of scientific progress—the legal landscape prefers a hoggish motif. The historic approach to the resulting database–privacy problem has been anonymization, a subtractive technique incurring not only poor privacy results, but also lackluster utility. In anonymization's stead, differential privacy arose; it provides better, near-perfect privacy, but is nonetheless subtractive in terms of utility. Today, another solution is leaning into the fore, synthetic data. Using the magic of machine learning, synthetic data offers a generative, additive approach—the creation of almost-but-not-quite replica data. In fact, as we recommend, synthetic data may be combined with differential privacy to achieve a best-of-both-worlds scenario. After unpacking the technical nuances of synthetic data, we analyze its legal implications, finding both over and under inclusive applications. Privacy statutes either overweigh or downplay the potential for synthetic data to leak secrets, inviting ambiguity. We conclude by finding that synthetic data is a valid, privacy-conscious alternative to raw data, but is not a cure-all for every situation. In the end, computer science progress must be met with proper policy in order to move the area of useful data dissemination forward.

Table of Contents

This is an unedited draft version.
Please cite to 22 STAN. TECH. L. REV ___ (2019).

INTRODUCTION

Synthetic data is a viable, next-step solution to the database-privacy problem: You are in a database;[1] sharing your secrets has the potential to unlock incredible breakthroughs across a vast number of disciplines;[2] but keeping your secrets private—while at the same time maintaining the usefulness of the data—is a nontrivial problem.[3] Enter: synthetic data, leveraging the power of machine learning to create an almost-but-not-quite replica of your data (as well as the data of others).

Historically, the way to share private information without betraying privacy was through anonymization,[4] stripping the data of *all* identifiers that could potentially uniquely identify an individual or group of individuals.[5] Anonymization, however, proved to be

---

[1] This happened not because you are famous, have over ten friends on Facebook, or even because you clicked "agree" to more Terms of Service contracts than you can count. This happened because you live in the 21st century. *See* ALAN F. WESTIN, PRIVACY AND FREEDOM 366-67 (1967) ("[T]he rapid computer development and usage throughout American society means that vast amounts of information about individuals and private groups in the nation are being placed in computer-usable form. More and more information is being gathered and used by corporations, associations, universities, public schools, and governmental agencies. And as 'life-long dossiers' and interchange of information grow steadily, the possibilities increase that agencies employing computers can accomplish heretofore impossible surveillance of individuals, businesses, and groups by putting together all the now-scattered pieces of data."); *see also* BRUCE SCHNEIER, DATA AND GOLIATH: THE HIDDEN BATTLES TO COLLECT YOUR DATA AND CONTROL YOUR WORLD 44 (2015) ("It's counterintuitive, but it takes less data to uniquely identify us than we think. Even though we're all pretty typical, we're nonetheless distinctive. It turns out that if you eliminate the top 100 movies everyone watches, our movie-watching habits are all pretty individual. This is also true for our book-reading habits, our Internet-shopping habits, our telephone habits, and our web-searching habits. We can be uniquely identified by our relationships."). Moreover, this database possesses the quantifiable parts of you—which may be more than you suspect. The database may include not only your name, where you live, where you work, who you know, and how to contact you, but likely a few other sensitive and interesting tidbits as well, such as how often you talk to your mother, where you like to go on Friday nights, or whether you are pregnant. *See infra* notes 37-40 and accompanying text; *see generally* JULIA ANGWIN, DRAGNET NATION (2014).

[2] *See* RAMEZ ELMASRI & SHAMKANT B. NAVATHE, FUNDAMENTALS OF DATABASE DESIGN 3 (7th ed. 2016) ("Databases and database systems are an essential component of life in modern society: most of us encounter several activities every day that involve some interaction with a database."); RAGHU RAMAKRISHNAN & JOHANNES GEHRKE, DATABASE MANAGEMENT SYSTEMS 4 (3rd ed. 2003) ("The amount of information available to us is literally exploding, and the value of data as an organizational asset is widely recognized."); Jane Yakowitz, *Tragedy of the Data Commons*, 25 HARVARD J.L. & TECH. 1, 4 (2011) ("[T]he social utility of the data commons is misunderstood and greatly undervalued by most privacy scholars.").

[3] *See* Matthew Fredrikson et al., *Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing*, *in* PROCEEDINGS 23RD USENIX SEC. SYMPOSIUM 19, 29 (2014) (showing that utility and privacy cannot be both achieved in the context of personalized warfarin dosing—"Put simply: our analysis indicates that in this setting where utility is paramount, the best known mechanisms for our application do not give an ε [i.e., the privacy loss parameter indicating the deviation between the original and modified data] for which state-of- the-art [differential privacy] mechanisms can be reasonably employed.")

[4] Anonymization, as used in this sentence, refers to the colloquial understanding of the term—which is more accurately defined as deidentification. *See infra* Section I.B (discussing the history of anonymization, starting with deidentification and leading to synthetic data).

[5] *See, e.g.,* Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization,* 57 UCLA L. REV. 1701, 1703 (2010) ("Imagine a database packed with sensitive information about many people. . . . Now imagine that the office that maintains this database needs to place it in long-term storage or disclose it to a third party without compromising the privacy of the people tracked. To

anything but a "silver bullet."[6] From the AOL search-query debacle to the Netflix Prize affair, it seemed trivial for just about anyone with some level of computer aptitude to "join" auxiliary information with a series of "perturbed" data points and unveil the very data that anonymization was designed to protect.[7]

The well-documented failures of anonymization prompted aggressive research on data sanitization, ranging from *k*-anonymity[8] in the late 1990s to today's highly acclaimed privacy mechanism, differential privacy.[9] But the basic tradeoff between utility and privacy—an inverse relationship to be sure—still remains.

The aim of this Article is to present the next, best step in sanitized data release, "synthetic" data.[10] In essence, take an original (and thus sensitive) dataset, use it to train[11] a machine learning enabled generative model,[12] and then use that model to produce

---

eliminate the privacy risk, the office will anonymize the data, consistent with contemporary, ubiquitous data-handling practices. First, it will delete personal identifiers like names and social security numbers. Second, it will modify other categories of information that act like identifiers in the particular context—the hospital will delete the names of next of kin, the school will excise student ID numbers, and the bank will obscure account numbers.").

[6] *Id.* at 1736; *see also* Arvind Narayanan & Edward W. Felton, *No Silver Bullet: De-Identification Still Doesn't Work* (July 9, 2014), http://randomwalker.info/ publications/no-silver-bullet-de-identification.pdf.

[7] *See* Arvind Narayanan & Vitaly Shmatikov, *Robust-De-Anonymization of Large Datasets (How to Break Anonymity of the Netflix Prize Dataset)*, 1, 2 (2008), https://arxiv.org/pdf/cs/0610105v1.pdf ("How much does the attacker need to know about a Netflix subscriber in order to identify her record in the [anonymized] dataset . . . very little."); Paul Ohm & Scot Pettet, *What if Everything Reveals Everything?*, *in* BIG DATA IS NOT A MONOLITH 46-47 (2016) (discussing what the authors believe to be a not-so-distant world where any piece of information reveals all pieces of information).

[8] *See generally* Pierangela Samarati & Latanya Sweeney, *Protecting Privacy when Disclosing Information: k-Anonymity and Its Enforcement through Generalization and Suppression* (1998), https://epic.org/privacy/reidentification/Samarati_Sweeney_paper.pdf; *see also* Subsubsection I.B.3.i.

[9] *See* Cynthia Dwork, *Differential Privacy*, *in* INT'L COLLOQUIUM ON AUTOMATA, LANGUAGES & PROGRAMMING 1, 1-2 (2006); see also Tore Dalenius, *Towards a Methodology for Statistical Disclosure Control*, *in* STATISTIK TIDSKRIFT 429 (1977); Michael Hilton, *Differential Privacy: A Historical Survey*, http://www.cs.uky.edu/~jzhang/CS689/PPDM-differential.pdf (last visited Aug. 7, 2018); *see also* Subsubsection I.B.3.ii.

[10] *See infra* Part II.

[11] When using machine learning, one first prepares a model of the likely input. This is done by feeding the program sample data, known as "training data." The program then "learns" its characteristics, and uses that knowledge to process subsequent input data. *See* HAL DAUMÉ, A COURSE IN MACHINE LEARNING 8-18 (2013) ("At a basic level, machine learning is about predicting the future based on the past. For instance, you might wish to predict how much a user Alice will like a movie that she hasn't seen, based on her ratings of movies that she has seen. This prediction could be based on many factors of the movies: their category (drama, documentary, etc.), the language, the director and actors, the production company, etc.").

[12] A generative model, per analogy, is like trying to identify a language someone is speaking by first learning many different languages and then matching one of those languages to the one being spoken. *See* Sargur N. Srihari, *Machine Learning: Generative and Discriminative Models*, 10 https://cedar.buffalo.edu/~srihari/CSE574/Discriminative-Generative.pdf (last visited Aug. 7, 2018). For a more rigorous description, see Andrew Ng, *CS229 Lecture Notes*, http://cs229.stanford.edu/notes/cs229-notes2.pdf (last visited Aug. 7, 2018) ("Consider a classification problem in which we want to learn to distinguish between elephants (y = 1) and dogs (y = 0), based on some features of an animal. Given a training set, an algorithm like logistic regression or the perceptron algorithm (basically) tries to find a straight line—that is, a decision boundary—that separates the elephants and dogs. Then, to classify a new animal as either

realistic, yet artificial data that nevertheless has the same statistical properties as the underlying, real data.[13] The end result may be compared to counterfeit money. Although the appropriately-sized and printed paper may appear genuine on first blush, a keen eye reveals its inauthenticity (e.g., perhaps the weight is ever-so-slightly lacking or the color-shifting ink is too monochromatic).[14] The goal of synthetic data is thus to create an as-realistic-as-possible dataset, one that not only maintains the nuances of the original data, but does so without endangering important pieces of personal information.[15]

But how do privacy-protecting statutes interpret this new method of data generation? If a trained model were to generate a synthetic dataset full of fictitious people it would plainly not offend strict interpretations of personally identifiable information—e.g., knowing the full extent of Mickey Mouse's medical history does not offend HIPAA because Mickey Mouse is not a real person. On the other hand, depending on how the machine learning model is trained and how broadly a statute is written, a synthetic dataset may "leak" (although the probability of such an event is remarkably small) just enough information to be considered offending—e.g., if the synthetic database's version of Mickey Mouse just happened to live at identifiably similar street address as a real person this may indeed run afoul of HIPAA.

To analyze synthetic data's legality, we first briefly discuss the database-privacy problem and outline a few privacy metrics that have populated the field post-anonymization-failure awareness. Next, in Part II, we present a case study on synthetic data using a real, practical dataset. Here, we look at the veracity of synthetic data and take a practical dive into its strengths and limitations. We then tie the two worlds together in Part III and assess synthetic data from a legal vantage, reviewing "vanilla" synthetic data (i.e., data generation without additional sanitization techniques) and differentially private synthetic data. Finally, we offer technical and legal recommendations for the community.

---

an elephant or a dog, it checks on which side of the decision boundary it falls, and makes its prediction accordingly. Here's a different approach. First, looking at elephants, we can build a model of what elephants look like. Then, looking at dogs, we can build a separate model of what dogs look like. Finally, to classify a new animal, we can match the new animal against the elephant model, and match it against the dog model, to see whether the new animal looks more like the elephants or more like the dogs we had seen in the training set. Algorithms that try to learn p(y|x) directly (such as logistic regression), or algorithms that try to learn mappings directly from the space of inputs X to the labels {0, 1}, (such as the perceptron algorithm) are called discriminative learning algorithms. [A]lgorithms that instead try to model p(x|y) (and p(y)). These algorithms . . . are called generative learning algorithms. For instance, if y indicates whether an example is a dog (0) or an elephant (1), then p(x|y = 0) models the distribution of dogs' features, and p(x|y = 1) models the distribution of elephants' features."); *see generally* Andrew Y. Ng & Michael I. Jordan, *On Discriminative vs. Generative Classifiers: A Comparison of Logistic Regression and Naive Bayes*, *in* NIPS (2002).

[13] To use an analogy, synthetic data is like replacing the pieces of a jigsaw puzzle to create a different picture; even though all the puzzle pieces *fit* together in the same way (i.e., each piece has similar, yet synthetic, attributes), the overall image has changed—importantly, and hopefully, the change is not discernable but nonetheless protects privacy.

[14] U.S. Currency Education Program, *Quick Reference Guide*, https://www.uscurrency.gov/sites/default/files/download-materials/en/Quick-Reference-Guide-5-10-20-50-100.pdf (last visited Aug. 1, 2018).

[15] To make certain, differential privacy precautions may be additionally added while creating the new data. *See infra* Section II.C. Thus, synthetic data does not challenge differential privacy, but is instead a more refined approach to protecting privacy with synthetic data.

In short, although no solution to the database-privacy problem is a "silver bullet,"[16] synthetic data is a promising next step, offering several advantages over historic methods of deidentification. Most importantly, synthetic data allows us to step away from the deidentification–reidentification arms race and focus on what really matters, useful data. That being said, because the method is relatively new, its meshing with legal statutes is both over and under inclusive—statutes thinking of "identification" in binary terms may accept the wholesale value of synthetic data, even though unique-enough data may nonetheless "leak" information; and statutes that consider identification broadly may prohibit synthetic data, even though risk of a leak, practically, is minimal.[17] Therefore, this Article recommends that the privacy community view synthetic data as yet another valid tool in the ever growing privacy tool belt; one that should be better accommodated by the law in terms of explicate permissions and limitations, but has the potential to offer great benefits when used properly.

I. THE DATABASE-PRIVACY PROBLEM

What is privacy? At its most general, privacy orbits around the right to be left alone, as it was originally contemplated by Warren and Brandeis, and later with Prosser.[18] From there, however, the concept has experienced its fair share of refactoring.[19]

A. Privacy: A Database Perspective

[16] *See* Narayanan & Felton, *supa* note 6 ("When faced with the challenge of fostering data science while preventing privacy risks, the urge to preserve the status quo is understandable. However, this is incompatible with the reality of re-identification science. If a 'best of both worlds solution exists, de-identification is certainly not that solution.").

[17] *See infra* Section III.B.

[18] *See* Samuel D. Warren & Louis D. Brandeis, *The Right to Privacy*, 4 HARVARD L. REV. 193, 195 (1890) (taking issue with how a new technology of the time—yellow journalism—permitted "what is whispered in the closet [to] be proclaimed from the house-tops"); William L. Prosser, *Privacy*, 48 CALIF. L. REV.383, 389 (1960) (using tort law to place emphasis on four different categories of invasions on a plaintiff's "right to be let alone" (quoting THOMAS M. COOLEY, A TREATISE ON THE LAW OF TORTS OR THE WRONGS WHICH ARISE INDEPENDENT OF CONTRACT 29 (2d ed. 1888)).

[19] *See generally* DANIEL J. SOLOVE, THE DIGITAL PERSON: TECHNOLOGY AND PRIVACY IN THE INFORMATION AGE 56-72 (2004). To be sure, privacy is simply the historical response to an age-old maxim: an irksome new technology unveiling a previously unidentified social norm (e.g., consider the "Doomsday Book," which permitted a Norman king to record his subject's property in an unmodifiable way—the change in technology, from "hearing the testimony of reliable local men to looking up a book kept by the Exchequer," permitted the king unprecedented powers of surveillance and ownership). *See* John Henry Clippinger, *Digital Innovation in Governance: New Rules for Sharing and Protecting Private Information*, *in* RULES FOR GROWTH: PROMOTING INNOVATION AND GROWTH THROUGH LEGAL REFORM 386-89 (2011) ("The term 'privacy' is derived from the Latin term, *privatus*, meaning separated from the rest...By separating out an individual's right for private information from that of a group, public, or government, the right of privacy forms the basis for a broad base of individual rights such as dignity, speech, worship, and happiness.") (citing DIALOGUS DE SCACCARIO: THE COURSE OF THE EXCHEQUER 64 (Charles Johnson ed. 1983); M.T. CLANCHY, FROM MEMORY TO WRITTEN ENGLISH 20 (3d ed. 2013)).

Congress's response to these watershed pieces of legal scholarship, along with the influential 1973 study,[20] was to enact a meshwork of statutes targeting areas of highly sensitive data.[21] Though not the exclusive avenue for privacy protection, these statutes form a meshwork particularly ensnarling for data sharing. Protected sectors span from health (HIPAA) to finance (FCRA), and often hinge the statutory shield on the definition of "personally identifiable information" (PII).[22] Put simply, if a fact (i.e., a datum[23] in the database) is PII, then it is protected and cannot be shared; if the fact is not PII, then it is not protected and may be shared freely.[24] The problem, of course, comes from delineating PII from non-PII.

---

[20] Warren & Brandeis, *supra* note 18. U.S. Dep't Health, Edu. & Welfare, Records Computers and the Rights of Citizens (1973), https://www.justice.gov/opcl/docs/rec-com-rights.pdf.

[21] Notably, this approach differs from the one adopted by the United Kingdom, which has been called "expansionist" and protects information that "may" lead to personal information. *See generally* Paul M. Schwartz & Daniel J. Solove, *Reconciling Personal Information in the United States and European Union*, G.W. L. FACULTY PUBLICATIONS & OTHER WORKS 1, 10 (2013), https://scholarship.law.gwu.edu/cgi/viewcontent.cgi?article=2103&context=faculty_publications.

[22] *See, e.g.,* Fair Credit Reporting Act of 1970, 15 U.S.C. §§ 1681; 1681a(c)-(d) (2006); Privacy Act of 1974, 5 U.S.C. § 552a(a)(2) (2006); FERPA, 42 U.S.C. § 1320g(a)(5)(a) (2006); Health Insurance Portability and Accountability Act (HIPAA) of 1996, 42 U.S.C. § 1320 (2006); Driver's Privacy Protection Act of 1994, 18 U.S.C. § 2721(a) (2006); Bank Secrecy Act of 1970, Pub. L. No. 91-508; Right to Financial Privacy Act of 1978, 12 U.S.C. §§ 3401–3422; Foreign Intelligence Surveillance Act of 1978, 15 U.S.C. §§ 1801–1811; Privacy Protection Act of 1980, 42 U.S.C. § 2000aa; Cable Communications Policy Act of 1984, 47 U.S.C. § 551; Electronic Communications Privacy Act of 1986, 18 U.S.C. §§ 2510–2522; Computer Matching and Privacy Protection Act of 1988, 5 U.S.C. § 552a; Employee Polygraph Protection Act of 1988, 29 U.S.C. §§ 2001–2009; Telephone Consumer Protection Act of 1991, 47 U.S.C. § 227; Communications Assistance for Law Enforcement Act of 1994, Pub. L. No. 103-414; Personal Responsibility and Work Opportunity Reconciliation Act of 1996, Pub. L. No. 104-193; Identity and Assumption Deterrence Act of 1998, 18 U.S.C. § 1028; Gramm-Leach-Bliley Act of 1999, 15 U.S.C. §§ 6801–6809; USA Patriot Act of 2001; CAN-SPAM Act of 2003; Video Voyeurism Prevention Act of 2004, 18 U.S.C § 1801; Video Privacy Protection Act of 1988, 18 U.S.C. § 2710 (2006). *See generally* Omer Tene, *Privacy Law's Midlife Crisis: A Critical Assessment of the Second Wave of Global Privacy Laws*, 74 OHIO. ST. L.J. 1217, 1225 (2013) (discussing how the United States responded to privacy protection by grouping categories of particularly sensitive information and creating specific rules to regulate those categories). Note that one impetus for the Privacy Act was the Watergate scandal; *see* STANLEY I. KUTLER, THE WARS OF WATERGATE: THE LAST CRISIS OF RICHARD NIXON 589 (1990) ("In his 1974 State of the Union message, Nixon warned that technology had encroached on the right of personal privacy. . . . Congress readily responded, but a committee report grasped the irony inherent in its efforts when it credited the 'additional impetus' from the 'recent revelations connected with Watergate-related investigations, indictments, trials, and convictions.'"); *see also* COMM. ON GOV'T OPERATIONS, 93D CONG., LEGIS HISTORY OF THE PRIVACY ACT OF 1974: S. 3418 (Pub. L. 93-579) 8 (J. Comm. Print 1976) (background).

[23] *See* MICHAEL J. HERNANDEZ, DATABASE DESIGN FOR MERE MORTALS: A HANDS-ON GUIDE TO RELATIONAL DATABASE DESIGN 43 (3rd ed. 2013) ("The values you store in the database are data. Data is static in the sense that it remains in the same state until you modify it by some manual or automated process").

[24] Under this framework, the question of privacy turns from "does this data point invade someone's privacy" to "does this data point fit within the statute's definition of what should be protected." According to Professor Ohm, this is part of the problem with PII in general: The question should not be does this data fit (because the factual data could *always* "fit" with the right inference or SQL "inner join"—though it would not be traditionally protected because without the inner joint it doesn't fit), but rather, does this data pose a high-risk to privacy. Professor Ohm outlined several factors to help answer his question: sanitation technique reliability, public release of the data, quantity limitations, industry motives for research and reidentification, and the trustworthiness of the data aggregator. *See* Ohm, *supra* note 5.

In fact, because of the statutory mosaic PII is iteratively defined in, the term is only ascertainable in a general sense. Professors Schwartz and Solove have therefore categorized it into three different buckets: (1) PII as a tautology, where the statutory definition of PII swallows any data that relates to the individual; (2) public versus non-public PII, where the statute shields only "non-public" information, though non-public is not defined; and (3) explicit PII specifications, where only those statutorily defined facts (e.g., both first and last name) are protected.[25] On a wide lens, the limitations on data sharing may be thought of through these categories.

With that general legal framework in place, we can now more easily look at the problem at hand; specifically, how information stored in databases creates a tradeoff between privacy and utility. To be sure, if no data is shared, perfect privacy is achieved; if the database is not perturbed in any way, perfect utility is achieved.[26]

B. Databases

Databases were not borne from the computer or the Internet. Indeed, a database is just that, the collection of data. Be it physical[27] or digital, the "database" is more technically defined as the "organized collection of factual relations."[28]

It is likewise important to note that databases are not inherently threatening to privacy. Before the proliferation of computerization democratized information, data describing individuals manifested itself in physical locations.[29] And these physical locations were, for the most part, geographically disparate. To concatenate database

[25] Paul M. Schwartz & Daniel J. Solove, *The PII Problem: Privacy and a New Concept of Personally Identifiable Information*, 86 N.Y.U. L. REV. 1814, 1829-35 (2011). The prototypical legal case follows this tune: corporation A is sharing a user's data with corporation B. The user files suit and the court must determine whether the data is in fact PII (the sharing is impermissible) or not (the sharing is permissible). The same is true for synthetic data, except the sharing would be done with synthetic data rather than original data.

[26] *See* Ohm, *supra* note 5, at 1752-55 ("[P]erfect privacy can be achieved by publishing nothing at all—but this has no utility; perfect utility can be obtained by publishing the data exactly as received from the respondents, but this offers no privacy.") (quoting Shuchi Chawla et al., *Toward Privacy in Public Databases*, *in* 2 THEORY CRYPTOGRAPHY CONF. 363 (2005)).

[27] *See, e.g.*, The Technium: One Dead Media, KK (June 17, 2008), https://kk.org/thetechnium/one-dead-media/ ("Edge-notched cards were invented in 1896. These are index cards with holes on their edges, which can be selectively slotted to indicate traits or categories, or in our language today, to act as a field. Before the advent of computers were one of the few ways you could sort large databases for more than one term at once. In computer science terms, you could do a "logical OR" operation. This ability of the system to sort and link prompted Douglas Engelbart in 1962 to suggest these cards could [implement] part of the Memex vision of hypertext.").

[28] *See* HERNANDEZ, *supra* note 23 at 4 ("[A] database is an organized collection of data used for the purpose of modeling some type of organization or organizational process. It really doesn't matter whether you're using paper or a computer application program to collect and store the data. As long as you're gathering data in some organized manner for a specific purpose, you've got a database."); RAMAKRISHNAN & GEHRKE, *supra* note 2, at 4 ("A database is a collection of data, typically describing the activities of one or more related organizations. For example, a university database might contain information about the following: Entities such as students, faculty, courses, and classrooms[;] and [r]elationships between entities, such as students' enrollment in courses, faculty teaching courses, and the use of rooms for courses.").

[29] *See* SOLOVE, *supra* note 19, at 13 (noting how records were mostly kept by hand in various state offices).

information required a laborious effort, the kind of effort deterring not only collection itself, but also the linkage of relations (i.e., vehicle records were not easily combined with credit card records, though both were located in governmental databases).[30] With compilation of data depressed, privacy rights were more easily protected using traditional, redact-my-name-and-zip-code methods.

But all this changed once seemingly unlimited columns,[31] cheap storage, and centralized access became more ubiquitous.[32] As our society merges itself with the digital world, information is more easily amassed.[33] Not only that, but linking different kinds of databases is also practical—unlocking the potential for *en masse* learning.[34] From the social sciences to medicine to modern day business operations, storing, analyzing, and reproducing information has become a second nature.[35] Some have even compared this type of interaction (colloquially termed "big data") as the 21st century's microscope. But our brave, big data world is not without drawback.[36]

---

[30] *Id.* at 14 ("Technology was a primary factor in the rise of information collection. The 1880 census [one of the first attempts at mass information collection] required almost 1500 clerks to tally information tediously by hand—and it took seven years to complete.").

[31] In database terms, this is known as a field (i.e., the vertical groupings in a Microsoft Excel document). *See* HERNANDEZ, *supra* note 23 at 52 ("A field (known as an attribute in relational database theory) is the smallest structure in the database and it represents a characteristic of the subject of the table to which it belongs. Fields are the structures that actually store data. The data in these fields can then be retrieved and presented as information in almost any configuration that you can imagine. . . . Every field in a properly designed database contains one and only one value, and its name will identify the type of value it holds. This makes entering data into a field very intuitive. If you see fields with names such as FIRSTNAME, LASTNAME, CITY, STATE, and ZIPCODE, you know exactly what type of values go into each field. You'll also find it very easy to sort the data by state or look for everyone whose last name is 'Hernandez.'").

[32] *See* SOLOVE, *supra* note 19, at 14.

[33] *See id.* at 15 ("Today, federal agencies and departments maintain almost 2,000 databases, 18 including records pertaining to immigration, bankruptcy, licensing, welfare, and countless other matters. In a recent effort to track down parents who fail to pay child support, the federal government has created a vast database consisting of information about all people who obtain a new job anywhere in the nation. The database contains their SSNs, addresses, and wages.").

[34] *See* Tim Berners-Lee, *The Next Web*, TED 10:45-15:00 (Feb. 2009), https://www.ted.com/talks/tim_berners_lee_on_the_next_web#t-678345 (urging listeners to rally around the slogan "Raw Data Now" to usher in a new generation of innovations in science, medicine, and technology); Tim Berners-Lee, *The Year Open Data Went Worldwide*, TED (Feb. 2010), https://www.ted.com/talks/tim_berners_lee_the_year_open_data_went_worldwide#t-301141 (listing a few of the ways open data has changed the world, among them the real-time mapping of Haiti after the 2010 earthquake, allowing users to see the location of refugee camps, damaged buildings, and hospitals).

[35] *See* ELMASRI & NAVATHE, *supra* note 2, at 3 (For example, if we go to the bank to deposit or withdraw funds, if we make a hotel or airline reservation, if we access a computerized library catalog to search for a bibliographic item, or if we purchase something online—such as a book, toy, or computer—chances are that our activities will involve someone or some computer program accessing a database. Even purchasing items at a supermarket often automatically updates the database that holds the inventory of grocery items.").

[36] *See* Jordi Soria-Comas & Josep Domingo-Ferrer, *Big Data Privacy: Challenges to Privacy Principles and Models*, 1 DATA SCI. ENGINEERING 21, 21-22 (2016) ("The potential risk to privacy is one of the greatest downsides of big data. It should be taken into account that big data are all about gathering as many data as possible to extract knowledge from them (possibly in some innovative ways). Moreover, more than often, these data are not consciously supplied by the data subject (typically a consumer, citizen), but they are generated as a by-product of some transaction (e.g., browsing or purchasing items in an online store), or they are obtained by the service provider in return for some free service (for example, free email accounts,

Because the information collected concerns more and more private minutiae,[37] the valuable byproducts of the knowledge come at a higher and higher privacy cost. For example, back in 2012 a father became irate when his high school-aged daughter began receiving coupons from Target for maternity clothing and nursery furniture. Shocked by Target's gall—how could a corporation make such a solicitous assumption?[38]—the father angrily demanded Target stop the harassment.

In reality, Target had accurately predicted the girl's third trimester date based on an algorithm it developed by crawling its massive customer database and identifying approximately 25 products that are indicative of pregnancy. Indeed, the panoply of what Target knew was "creepily" extensive.[39] But the larger, more important picture was also evident: industry, and the government, had been collecting this kind of data for years, only yielding more and more intimate details of our lives as the technology improves.[40] But herein lies the problem, how can this data be usefully applied *without* stepping on anyone's privacy toes? At the time, the answer was anonymization.

1. The (Assumedly) Good: Privacy via "Anonymization"

Early on—and still making the rounds today[41]—the assumption was that if you stripped out enough identifying information from a dataset, the data could be shared

---

social networks) or as a natural requirement for some service (e.g., a GPS navigation system needs knowledge about the position of an individual to supply her with information on nearby traffic conditions)."

[37] For example, consider biometric data. *See, e.g.*, EFF, Immigration Policy Center, Jennifer Lynch, *From Finger Prints to DNA: Biometric Data Collection in the U.S. Immigrant Communities and Beyond* (2012), https://www.eff.org/document/fingerprints-dna-biometric-data-collection-us-immigrant-communities-and-beyond.

[38] Charles Duhigg, N.Y. TIMES, MM30 (Feb. 19, 2012) ("Andrew Pole had just started working as a statistician for Target in 2002, when two colleagues from the marketing department stopped by his desk to ask an odd question: 'If we wanted to figure out if a customer is pregnant, even if she didn't want us to know, can you do that?'").

[39] *See* Nick Saint, *Eric Schmidt: Google's Policy is to 'Get Right up to the Creepy Line and Not Cross it*,'" BUSINESS INSIDER (Oct. 1, 2010, 2:44 PM), https://www.businessinsider.com/eric-schmidt-googles-policy-is-to-get-right-up-to-the-creepy-line-and-not-cross-it-2010-10; *see also* Duhigg, *supra* note 38 ("For decades, Target has collected vast amounts of data on every person who regularly walks into one of its stores. Whenever possible, Target assigns each shopper a unique code—known internally as the Guest ID number—that keeps tabs on everything they buy. 'If you use a credit card or a coupon, or fill out a survey, or mail in a refund, or call the customer help line, or open an e-mail we've sent you or visit our Web site, we'll record it and link it to your Guest ID[.] We want to know everything we can.' Also linked to your Guest ID is demographic information like your age, whether you are married and have kids, which part of town you live in, how long it takes you to drive to the store, your estimated salary, whether you've moved recently, what credit cards you carry in your wallet and what Web sites you visit. Target can buy data about your ethnicity, job history, the magazines you read, if you've ever declared bankruptcy or got divorced, the year you bought (or lost) your house, where you went to college, what kinds of topics you talk about online, whether you prefer certain brands of coffee, paper towels, cereal or applesauce, your political leanings, reading habits, charitable giving and the number of cars you own.").

[40] Schneier, *supra* note 1, at 4 ("The bargain you make, again and again, with various companies is surveillance in exchange for free service.").

[41] For example, FERPA explicitly allows data release if the information is deidentified, which is defined as the removal of all personally identifiable information; the Gramm-Leach-Bliley Act has been interpreted by the FTC to *not* protect non-PII data, which is defined as "[i]nformation that does not identify a consumer, such as aggregate information or blind data [not containing] personal identifiers such as account

freely.[42] Though the approach is colloquially referred to as anonymization (i.e., "without a name or nameless" from the Greek *ἀνωνυμία*[43]), it is more accurately described as deidentification: sterilization via subtraction.[44] A good example comes from arguably the bellwether of deidentification, HIPAA.[45] Under its "safe harbor" provision, medical data is freely shareable if seventeen identifiers have been removed.[46] Detailed and explicitly defined, HIPAA assumes that information lacking these identifiers is of no privacy concern: How could Jane Doe's privacy be affected if no one knows her name, address, or social security number?[47]

---

numbers, names, or addresses"; and both the Cable Act and VPPA's definition of PII have been interpreted by the courts to *not* cover anonymized identifiers—in *Pruitt*, hexadecimal codes identifying customers and their purchases were not viewed as PII because, in the court's eyes, the digits were simply not addresses or names,[41] and in *In re Hulu*, the court found that "a unique anonymized ID alone is not PII." *See* 34 C.F.R. § 99.31(b) (discussing how the deidentification must reasonably ensure that a student's identity is not "personally identifiable"); 16 C.F.R. § 313.3(0)(2)(ii)(B), https://www.gpo.gov/fdsys/pkg/CFR-2010-title16-vol1/pdf/CFR-2010-title16-vol1-sec313-3.pdf; *see also* Benjamin Charkow, Note, *The Control Over the De-Identification of Data*, 21 CARDOZO ARTS & ENT. L.J. 195, 196-97 (2003) (noting how "[c]ongressional statutes and related administrative agency regulations typically exclude information from protection once the information has been modified in such a way that the data subject can no longer be identified" and arguing that "no privacy interest is retained in de-identified information"). Pruitt v. Comcast Cable Holdings, LLC, 100 F. App'x 713, 715 (10th Cir. 2004) ("[T]he converter box code—without more—provides nothing but a series of numbers…. Without the information in the billing or management system one cannot connect the unit address with a specific customer; without the billing information, even Comcast would be unable to identify which individual household was associated with the raw data in the converter box."). In re Hulu Privacy Litig., No. C 11-03764 LB, 2014 WL 1724344 *10-11 (N.D. Cal. Apr. 28, 2014) (using Pruitt as a standard and finding that "an anonymous, unique ID without more does not constitute PII.").

[42] *See* Ohm, *supra* note 5, at 1707-11 (recounting the staunch supporters of deidentification—spanning from industry to academia to administration).

[43] *See* Zoltán Alexin, *Does Fair Anonymization Exist?* 28 INT'L REVIEW L. COMPUTERS & TECH. 21, 21 (2014) (finding that HIPAA's safe harbor permits anonymization, meaning that stripping out the seventeen named identifiers permits a small enough chance of reidentification to consider the resulting dataset private).

[44] *See* Rubinstein & Hartzog, *supra* note 8, at 710 (defining deidentification as "the process by which data custodians remove the association between identifying data and the data subject") (citing Simson L. Garfinkel, *De-Identification of Personal Information*, NISTIR 8053 (2015), http://nvlpubs.nist.gov/nistpubs/ir/2015/NIST.IR.8053.pdf). Sanitization is also a term used to describe a similar process. *See* Justin Brickell & Vitaly Shmatikov, *The Cost of Privacy: Destruction of Data-Mining Utility in Anonymized Data Publishing*, *in* PROC. 14TH ACMKDD INT'L CONF. KNOWLEDGE DISCOVERY & DATA MINING 70, 70 (2008) (considering "trivial sanitization" to be the removal of all quasi-identifiers or sensitive attributes).

[45] *See generally* Health Insurance Portability and Accountability Act (HIPAA) of 1996, 42 U.S.C. § 1320 (2006); 45 CFR § 164.514.

[46] To be sure, an expert may also certify the data's anonymity. Among the seventeen identifiers, an eighteenth catch-all provision is included. In total, these include: name, geographic subdivision smaller than a state (including and address and part of the zip code), anything including a date, telephone number, vehicle identifiers like license plate number, fax number, serial number, email address, URLs, social security number, IP address, medical record numbers, biometric identifiers, insurance number, facial images, account numbers, professional license number, and any other unique identifier not listed. 45 CFR § 164.514(b)(2)(i).

[47] To be sure, HIPAA does have a provision for "any other unique identifier not listed." However, as Section I.B.2 demonstrates, what deidentification misses is the fact that even trivial data points may be linked with an identity—e.g., before the Netflix Prize reidentifications, it would be defy logic to consider that liking non-blockbuster movies is a unique identifier. *See* Arvind Narayanan & Vitaly Shmatikov, *Robust-De-Anonymization of Large Sparse Datasets*, *in* PROC. IEEE SYMPOSIUM SECURITY & PRIVACY 1, 12 Fig. 9 (2008) (showing how liking less popular movies makes deidentification more likely).

Unfortunately, just because a database looks anonymous, does not mean it is. As a long line of academics have stressed, identifying individuals using seemingly non-unique identifiers is a much easier task than a data sanitizer might hope.[48]

2. The Bad: Reidentification Awareness

Because the core of deidentification is the removal of unique identifiers, a premium is necessarily placed on precisely defining what constitutes a unique identifier. Indeed, by relying on "subtractive uniqueness"[49] it is very difficult to correctly guess what attributes should be removed (i.e., one man's trash is another man's treasure) while maintaining the necessary idiosyncrasies for the data to remain useful.[50] The result is an inescapable tradeoff, more representative data versus more privacy.

Consider that full scale DNA sequencing—anonymized via deidentification—was publicly released in the 1990s as part of the Human Genome Project.[51] But, in 2004, researchers demonstrated how it was possible to link an individual's deidentified genomic data with publicly available single nucleotide polymorphism data.[52] NIH reacted to the privacy woes by restricting access to individual-level genomic data on a permission-only basis.[53] But then, in 2008, researchers again showed that individuals could be identified in trace-amount, high-density genetic mixtures.[54] NIH clamped down harder on the weak

---

[48] *See, e.g.*, Brian Hayes, *Uniquely Me!*, AM. SCI., https://www.americanscientist.org/article/uniquely-me.

[49] Similar to the way subtractive manufacturing produces a desired object by removing material until the object is created, deidentification takes an original description and removes as much of it as necessary to achieve the desired anonymity. *See* Samuel H. Huang, Peng Liu, Abhiram Mokasdar & Liang Hou, *Additive Manufacturing and Its Societal Impact: A Literature Review*, 67 J. ADV. MANUFACTURING TECH. 1191, 1191 (2013) (discussing how traditional manufacturing techniques remove material from an object in order to whittle down a final product).

[50] *See* Brian Parkinson, David E. Millard, Kieron O'Hara & Richard Giordano, *The Digitally Extended Self: A Lexicological Analysis of Personal Data*, 1 J. INFO. SCI. 1, (2017) (noting how "the classification of data based on degrees of identifiability may fluctuate and become indeterminate").

[51] This was the result of policies adopted by several organizations, including the National Human Genome Research Institute, the Department of Energy, and International Human Genome Sequencing Consortium. With a focus on open records, these policies generally recommended depositing data and resource into the public domain. *See* Reaffirmation and Extension of NHGRI Rapid Data Release Policies: Large-scale Sequencing and Other Community Resource Projects, NHGRI (Feb. 2003), https://www.genome.gov/10506537/.

[52] Zhen Lin, Art. B. Owen, Russ B. Altman, *Genomic Research and Human Subject Privacy*, 305 SCIENCE 183, 183 (2004) ("[I]f someone has access to individual genetic data and performs matches to public [single nucleotide polymorphism (SNP)] data a small set of SNPs could lead to successful matching and identification of the individuals. In such a case, the rest of the genotypic, phenotypic, and other information linked to that individual in public records would be available.").

[53] NIH did this, via the Database of Genotypes and Phenotypes. Institutional approval is needed prior to accessing individual genomic data. *See* Stacey Pereira, Richard A. Gibbs & Amy L. McGuire, *Open Access Data Sharing in Genomic Research*, 5 GENES 739 (2014)

[54] This is the common "security via aggregation" theory, which was applied to batch-samples containing many participants' data. *See* Nils Homer et al., Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays, 4 PLOS GENETICS 1, 1, 9 (2008) ("Considering privacy issues with genetic data, it is now clear that further research is needed to determine how to best share data while fully masking identity of individual participants. However, since sharing only summary data does not completely mask identity, greater emphasis is needed for providing

link, restricting any access to genome-wide association studies.[55] Finally, and most recently in 2013, researchers demonstrated yet again how it was possible to match NIH's snippet tandem repeats with consumer-focused, publicly available, genetic genealogy information, permitting an individual's surname to be identified.[56] In response, NIH held its tune and moved age information from a public to non-public database.[57]

What this vignette demonstrates is not only that deidentification requires precise definitions of "unique identifiers," but also that deidentification suffers from an aging problem. When genomic data was originally released, consumer-opt-in genetic studies were not popular. It is difficult enough to pin down exactly what data identifies individuals, but it is even more difficult to accurately predict what potential auxiliary information could be available in the future—i.e., the deidentification–reidentification arms race.

3. The Confusing: Post-Anonymization-Failure Awareness

Realizing that anonymization talks more than it walks,[58] a number of alternative approaches have been suggested. Setting aside the purely legal solutions (often focusing on reframing PII),[59] one of the most popular paths in this terrain is to use computer science metrics to replace the historic means of deidentification.

The lynchpin in each of these methods must be explicitly stated: The first place to start is with creating a metric for defining privacy, because, as with all connotative

---

mechanisms to confidentially share and combine individual genotype data across studies, allowing for more robust meta-analysis such as for gene-environment and gene-gene interactions.").

[55] See Natasha Gilbert, Researchers Criticize Genetic Data Restrictions: Fears Over Privacy Breaches are Premature and Will Impede Research, Experts Say, NATURE NEWS (Sept. 4, 2008), http://www.nature.com/news/2008/080904/full/news.2008.1083.html ("Although scientists will still be able to access data eventually, in practice getting all the relevant bureaucratic approvals can take months and this can be a huge deterrent to researchers and a restriction on scientific progress.").

[56] Melissa Gymrek et al., *Identifying Personal Genomes by Surname Inference*, 339 SCIENCE 321, 321, 324 (2013) ("This study shows that data release, even of a few markers, from one person can spread through deep genealogical ties and lead to the identification of another person who might have no acquaintance with the person who released his genetic data.").

[57] *See generally* Pereira, Gibbs & McGuire, *supra* note 53, at 739-40 (using patient survey results to show that a majority would prefer open access genomic data release if given the option). But see Khaled El Emam, Elizabeth Jonker, Luk Arbuckle & Bradley Malin, *A Systematic Review of Re-Identification Attacks on Health Data*, 6 PLOS ONE 1, (2011) ("The current evidence shows a high re-identification rate but is dominated by small-scale studies on data that was not de-identified according to existing standards. This evidence is insufficient to draw conclusions about the efficacy of de-identification methods.").

[58] Specifically, anonymization promises much more than it manages to deliver.

[59] Solving PII's failures is not without disagreement. *See* Schwartz & Solove, *supra* note 25 at 1894 (2011) (arguing for a division between "identified" and "identifiable" information and applying protection mechanisms based on the risk each category engenders); Ira S. Rubinstein & Woodrow Hartzog, *Anonymization and Risk*, 91 WASH. L. REV. 703, 706 (2016) ("[T]he best way to move data release policy past the alleged failures of anonymization is to focus on the process of minimizing risk, not preventing harm."). *Compare* Andrew Chin & Anne Klinefelter, *Differential Privacy as a Response to the Reidentification Threat: The Facebook Advertiser Case Study*, 90 N.C. L. REV. 1417, 1423 (2012) (arguing that differential privacy could be a workable standard to replace traditional anonymization techniques) *with* Jane Bambauer, Krishnamurty Muralidhar & Rathindra Sarathy, *Fool's Gold: An Illustrated Critique of Differential Privacy*, 16 VAND. J. ENT. & TECH. L. 701, 754 (2014) ("In its strictest form, differential privacy is a farce. In its most relaxed form, it is no different, and no better, than other methods.")

definitions, using non-technical, non-mathematical descriptions invites ambiguity.[60] Indeed, it is because of these metrics that computer scientists are able to provide quantifiable guarantees; specifically, a measure of *how much* privacy is protected not only in the typical case, but also in the face of an "attacker" attempting to obtain secrets.

i. k-*Anonymity*

k-anonymity maintains privacy by *guaranteeing* that for every record in a database there are some number "*k*" of indistinguishable copies.[61] Stated otherwise, no single row in the table is unique because it cannot be distinguished from at least *k* others. The fundamental guiding principle of *k*-anonymity is that it tries to map at least *k* entities to what is considered identifying information in a database.

To better understand how this sanitization technique works, consider the following table, which pairs an individual with a computing "task" (i.e., accessing a file via "touch," creating a file via "create," or removing a file via "delete").

Name	Class Year	Phone Number	Task
Bill	1	123-345-6789	Touch
Alice	1	234-345-4567	Touch
Becky	2	345-456-5678	Create
Bob	2	456-567-6789	Delete

In an attempt to anonymize the table above, we can use a combination of two common techniques, both leading to *k*-anonymity: suppression and generalization.[62] The suppression method follows a denotative definition and replaces a pivotal piece of identifying information in the original database with a meaningless placeholder.[63] In our example, we will remove "name" and "phone number" and insert a "#" as a symbolic placeholder. The other technique, generalization, employs a broadening approach to add uncertainty, aggregating rows (i.e., "Class Year" of one) to create a range of values as

[60] Felix T. Wu, *Defining Privacy and Utility in Data Sets*, 84 U. COLO. L. REV. 1117, 1121, 1125-26 (2013).

[61] *See* Samarati & Sweeney, *supra* note 8 at 1 ("[W]e address the problem of releasing person-specific data while, at the same time, safeguarding the anonymity of the individuals to whom the data refer.").

[62] Roberto J. Bayardo & Rakesh Agrawal, *Data Privacy Through Optimal* k-*Anonymization*, *in* INT'L CONFERENCE DATA ENGINEERING 1, 1 (2005) ("Suppression is the process of deleting cell values or entire tuples. Generalization involves replacing specific values such as a phone number with a more general one, such as the area code alone.").

[63] *Id.*

opposed to a single value (i.e., "Class Year" between one and three).[64] Applying these two techniques to our simple dataset results in the following:

Name	Class Year	Phone Number	Task
#	$1 \leq \text{Year} \leq 3$	#	Touch
#	$1 \leq \text{Year} \leq 3$	#	Touch
#	$2 \leq \text{Year} \leq 4$	#	Create
#	$2 \leq \text{Year} \leq 4$	#	Delete

The newly suppressed and generalized dataset now has a *k* value of two for Class Year since there are two records for any class (i.e., two rows have a Class Year from one to three and two rows have a Class Year from two to four). Importantly, though the example is oversimplified, it nonetheless illustrates of the clear tradeoff in utility for privacy—the table is less useful now that each individual row is less unique. There has been a loss of utility for the gain of privacy.

ii. Differential Privacy

Another popular and robust method is differential privacy.[65] Differential privacy rose to prominence following the famous shortcomings of the Netflix Prize affair.[66] While

---

[64] *Id.*; *see also* Sheng Zhong, et al., *Privacy-Enhancing* k-*Anonymization of Consumer Data*, *in* PRINCIPLES OF DATABASE SYSTEMS 139, 139-40 (2005).

[65] Andrew Chin & Anne Klinefelter, *Differential Privacy as a Response to the Reidentification Threat: The Facebook Advertiser Case Study*, 90 N.C. L. REV. 1417, 1423 (2012) (arguing that differential privacy could be a workable standard to replace traditional anonymization techniques); Kobbi Nissim, et al., *Differential Privacy: A Primer for a Non-technical Audience*, VAND. J. ENT & TECH. LAW (forthcoming), https://privacytools.seas.harvard.edu/files/privacytools/files/pedagogical-document-dp_new.pdf ("Intuitively, a computation protects the privacy of individuals in the data if its output does not reveal any information that is specific to any individual data subject. Differential privacy formalizes this intuition as a mathematical definition. Just as we can show that an integer is even by demonstrating that it is divisible by two, we can show that a computation is differentially private by proving it meets the constraints of the definition of differential privacy. In turn, if a computation can be proven to be differentially private, we can rest assured that using the computation will not unduly reveal information specific to a data subject.").

[66] *See* Narayanan & Shmatikov, *supra* note 7; Ohm, *supra* note 5, at 1720-22 ("On October 2, 2006, about two months after the AOL debacle, Netflix, the 'world's largest online movie rental service,' publicly released one hundred million records revealing how nearly a half-million of its users had rated movies from December 1999 to December 2005.90 In each record, Netflix disclosed the movie rated, the rating assigned (from one to five stars), and the date of the rating. Like AOL and GIC, Netflix first anonymized the records, removing identifying information like usernames, but assigning a unique user identifier to preserve rating-to-rating continuity . . . To improve its recommendations, Netflix released the hundred million records to launch what it called the 'Netflix Prize,' a prize that took almost three years to claim. The first team that used

there are many forms of the general technique, its primary goal is to maximize the accuracy of queries from a database while limiting or minimizing the potential of privacy leakage.[67] Theoretical computer scientists are fond of the method due to its strict mathematical formulations and provable guarantees. For our purposes, a high-level understanding may be attained through a simple example, one which is modified from Professor Dwork's recent work.[68]

Imagine a scenario in which someone asks you the question: "Do you like ice-cream?"[69] This answer has a binary, yes or no, answer. However, it could be modified with the aid of a coin toss.[70] Prior to answering, a coin is tossed, and if a head is the result, the person answering the question tells the truth. If not heads, the person will give a "random" answer (which in this case is another coin toss with a predefined "yes" if heads and "no" if not).[71]

Notably, though it is possible to deduce the probability of people who like ice-cream, the individuals answering this question now have "deniability."[72] Indeed, although combining some basic facts about the independence of events may produce a probability distribution, the individuals are now permitted to say "I may or may not have answered truthfully." And this is the gravitas of differential privacy: Because of the introduction of randomness (i.e., a person's veracity depends on a coin toss) which produces deniability, the individuals may now say "I may or may not be 'in' the database."[73]

To be sure, differential privacy has many strengths; but as with all methods, it is not a panacea.[74] For example, if enough identical queries are asked, the power of deniability is diluted.[75] Eventually, repeat queries may be able to take all answers together and

---

the data to significantly improve on Netflix's recommendation algorithm would win one million dollars. . . . Two weeks after the data release, researchers from the University of Texas, Arvind Narayanan and Professor Vitaly Shmatikov, announced that 'an attacker who knows only a little bit about an individual subscriber can easily identify this subscriber's record if it is present in the [Netflix Prize] dataset, or, at the very least, identify a small set of records which include the subscriber's record.'").

[67] Although not the first to introduce differential privacy, a commonly cited survey. *See* Cynthai Dwork, *Differential Privacy: A Survey of Results*, in Int'l Conference Theory & Applications of Models of Computation 1, 2-3 (2008).

[68] Cynthia Dwork & Aaron Roth, *The Algorithmic Foundations of Differential Privacy*, 9 FOUNDATIONS & TRENDS THEORETICAL COMP. SCI. 211 (2013).

[69] *Id.* at 29-30.

[70] *Id.*

[71] *Id.*

[72] *Id.* at 15-16 (discussing how "'[p]rivacy' comes from the plausible deniability of any outcome").

[73] Allowing those responsible or protecting the database's secrets to say: "You will not be affected, adversely or otherwise, by allowing your data to be used in any study or analysis, no matter what other studies, data sets, or information sources, are available." *Id.* at 211 ("Differential privacy will ensure that the ability of an adversary to inflict harm (or good, for that matter)—of any sort, to any set of people—should be essentially the same, independent of whether any individual opts in to, or opts out of, the dataset.").

[74] For one critique of the metric, see Bambauer, Muralidhar & Sarathy, *supra* note 59 at 754 (2014) ("In its strictest form, differential privacy is a farce. In its most relaxed form, it is no different, and no better, than other methods."). *But see* Frank McSherry, *Differential Privacy for Dummies*, GITHUB (Jan. 4, 2017), https://github.com/frankmcsherry/blog/blob/master/posts/2016-02-03.md (critiquing Fool's Gold for misreading differential privacy's value).

[75] *See generally* Bill Howe, *Weakness of Differential Privacy*, COURSERA (last accessed Aug. 1, 2018), https://www.coursera.org/lecture/data-results/weaknesses-of-differential-privacy-50Y9k (finding that differential privacy is best in low-sensitivity areas and has problems with repeat queries).

disambiguate falsity from truth. Additionally, if the query being asked requires high specificity, then it is more difficult to permit deniability.[76] For example, if a query asks about the minimum GPA in a group of students, it will be hard to tell a lie while at the same time provide a useful answer, because there is only one student with the lowest GPA.

In fact, some studies suggest that utility and privacy are mutually exclusive attributes if utility of the data is of upmost importance.[77] One particular study focusing on Warfarin dosing found that privacy was only sufficiently protected if differential privacy was used—but differential privacy was found to destroy utility.[78] "We show that differential privacy substantially interferes with the main purpose of these models in personalized medicine: for ε values [i.e., a measure of "how" protective of privacy the database is] that protect genomic privacy . . . the risk of negative patient outcomes increases beyond acceptable levels."[79] Stated otherwise, if absolute utility is needed, even *de minimis* sanitization has an adverse effect.

In a summary with a little more legal gusto, the solutions to the database-privacy problem thus far may be likened to requesting a contentious document from the FBI pursuant to a Freedom of Information Request. The FBI may return an incredibly redacted document which perfectly maintains privacy by liberally striking out all phrases that were not completely benign (i.e., deidentification). Unfortunately, the document is useless; it is impossible to plunder the document's gems when all that can be seen are black highlights. Using updated methods of sanitization will help the FBI be more responsive—*k*-anonymity (i.e., replacing names, dates, and locations with symbols or grouping important facts together) or differential privacy (i.e., allowing you to ask the FBI specific questions, but not knowing whether the FBI is answering truthfully)—but not in all cases. We are still in a negative-sum game, less data for more privacy, and there is still the threat that joining auxiliary information with existing data could unveil secrets. This brings us to yet another solution posed by the computer science literature, synthetic data.

II. SYNTHETIC DATA

Synthetic data may be thought of as "fake" data created from "real" data. The beauty of it stems from its grounding in real data and real distributions, which make it almost indistinguishable from the original data. Its impetus, in this context, comes from the fact that there are many times when the best data that may be shared, real data, is legally protected and cannot be shared; conversely, the practical data that can be shared,

---

[76] *Id.*

[77] Matthew Fredrikson et al., *Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing*, *in* PROCEEDINGS 23RD USENIX SECURITY SYMPOSIUM 19, 29 (2014) (showing that utility and privacy cannot be both achieved in the context of personalized warfarin dosing—"Put simply: our analysis indicates that in this setting where utility is paramount, the best known mechanisms for our application do not give an ε [i.e., the privacy loss parameter indicating the deviation between the original and modified data] for which state-of- the-art mechanisms can be reasonably employed.").

[78] The study did use a new method to assess utility, in which differentially private results and non-sanitized results were used to suggest warfarin dosing amounts. In the end, if the privacy loss parameter (ε) were too high, then utility was gained but privacy was lost; however, if ε were too low, then privacy was gained but utility was lost, resulting in adverse patient outcomes. *See id.* at 26-27, 29.

[79] *Id.* at 29.

deidentified data, is not as useful as it should be. It is in those moments that having a synthetic dataset may present a best-of-both-worlds solution.

Prior to engaging with our case study illustrating the veracity and limitations of synthetic data, it will be helpful to have a nutshell outline of the core concepts underlying synthetic data. Here, we start with the oft-cited, yet poorly-understood term, machine learning.

A. In Brief: Machine Learning

When Ada Lovelace sat down to ponder about one of the world's first computer programs, the automatic calculation of Bernoulli numbers,[80] she did so in painstaking detail.[81] Each and every step considered in a laborious, mathematical fashion.[82] Yet, as anyone with experience in programming knows, this is exactly what the art of programming requires.[83] But what if the computer could learn to calculate the Bernoulli numbers on its own; what if by showing the computer specific data the computer could interpret the data in a useful fashion and, after many iterations, replicate desired behavior?[84] This is exactly what machine learning does, often relying on a neural network at its core.[85]

Neural networks have garnered tremendous attention given the multitude of problems that are now tractable as a result of improvements in computer hardware and cheaper prices for that hardware.[86] The easiest way to understand a neural network is to

---

[80] Used to "express the value of the zeta function at integer even values." Pascal Sabah & Xavier Gourdon, *Introduction on Bernoulli's Numbers* 1, 1 (2002), http://math.ucr.edu/~res/math153/s12/bernoulli-numbers.pdf.

[81] *See generally* L.F. Menabrea, *Sketch of the Analytical Engine Invented by Charles Babbage*, 3 SCI. MEMOIRS 666 (1843), https://johnrhudson.me.uk/computing/Menabrea_Sketch.pdf.

[82] For an image of the resulting program, see Gene Kogan, From Deep Learning Down: An Excavation of Mathematics Reveals the Continuity of Our Knowledge, MEDIUM (Dec. 28, 2017), https://medium.com/@genekogan/from-deep-learning-down-85e527a5fe7b.

[83] *See* MARKO PETKOVŠEK, HERBERT S. WILF & DORON ZEILBERGER, A=B vii (1997) ("Science is what we understand well enough to explain to a computer. Art is everything else we do. During the past several years an important part of mathematics has been transformed from an Art to a Science: No longer do we need to get a brilliant insight in order to evaluate sums of binomial coefficients, and many similar formulas that arise frequently in practice; we can now follow a mechanical procedure and discover the answers quite systematically."); *see generally* DONALD E. KNUTH, THE ART OF COMPUTER PROGRAMMING (vols. 1-4 1968).

[84] *See* Jeremy Howard, *The Wonderful and Terrifying Implications of Computers that can Learn*, TED (Dec. 2014), https://www.ted.com/talks/jeremy_howard_the_wonderful_and_terrifying_implications_of_computers_that_can_learn (describing how Arthur Samuel wanted to write a program that could play—and beat—him at checkers, eventually coming upon the idea that the computer program should "learn" to play checkers by playing against itself).

[85] Research in the neural network domain dates back several decades and can be attributed as early as Walter Pitts's work published in 1942. *See* Walter Pitts, *Some Observations on the Simple Neuron Circuit*, 4 BULLETIN MATHEMATICAL BIOPHYSICS 121 (1942).

[86] *See* Ophir Tanz, *How Video Game Tech Makes Neural Networks Possible*, TECHCHRUNCH (Oct. 27, 2017), https://techcrunch.com/2017/10/27/how-video-game-tech-makes-neural-networks-possible/ ("When id Software's John Carmack released Doom in 1993, he had no inkling that his gory first-person shooter--one of the first to feature a 3D environment, and easily the most popular at that time--would help spark a revolution in how machines process information. Six years later, Nvidia released the GeForce 256, the first graphical processing unit (GPU) built specifically to produce 3D graphics for the burgeoning game

see how one behaves after the training process is complete—i.e., using a pre-trained model.[87] From there, we will work backwards and explain on a broad level how training occurs.

For the specific type of network, the archetypical educational example, a convolutional neural network (CNN), will be best. These networks are often used for image classification and provide the easiest means to better understand an otherwise difficult to illustrate concept.[88]

1. The Neural Network

The first place to start is with a mathematical representation of some goal; here, the goal is to correctly identify hand-drawn digits. Images consist of a series of pixels, and pixels consist of a series of numbers (red, green, and blue) which together make a specific color, forming all of the images you see on your computer screen.[89] Assume we start with the hand-drawn image of a one.

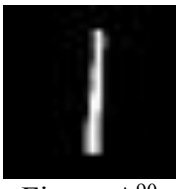


Figure A[90]

industry. In the 17 years since, GPUs have become not merely a staple of high-end gaming, which was the original primary reason for their development, but a driving force behind major advances in artificial intelligence (AI).").

[87] *See* Jeremy Howard, *Lesson 3: Deep Learning 2018*, YOUTUBE (Dec. 30, 2017), https://www.youtube.com/watch?v=9C06ZPF8Uuc; Otavio Good, *A Visual and Intuitive Understanding of Deep Learning*, YOUTUBE (Nov. 5, 2017).

[88] Although we later describe the production of synthetic data from a text standpoint, images may be synthetically replicated as well. *See* John T. Guibas, Tejpal S. Virdi & Peter S. Li, *Synthetic Medical Images from Dual Generative Adversarial Networks*, *in* NEURAL INFO. PROCESSING SYS. 1, 1-2 (2017) ("We propose a novel pipeline for generating synthetic medical images, allowing for the production of a public and extensive dataset, free from privacy concerns.").

[89] *See* DAVID J. ECK, INTRODUCTION TO COMPUTER GRAPHICS 11-19 (2018) ("A digital image is made up of rows and columns of pixels. A pixel in such an image can be specified by saying which column and which row contains it. . . . The colors on a computer screen are produced as combinations of red, green, and blue light. Different colors are produced by varying the intensity of each type of light. A color can be specified by three numbers giving the intensity of red, green, and blue in the color. Intensity can be specified as a number in the range zero, for minimum intensity, to one, for maximum intensity. This method of specifying color is called the RGB color model, where RGB stands for Red/Green/Blue. For example, in the RGB color model, the number triple (1, 0.5, 0.5) represents the color obtained by setting red to full intensity, while green and blue are set to half intensity. The red, green, and blue values for a color are called the color components of that color in the RGB color model."); *see also* Victor Powell, *Image Kernels: Explained Visually*, http://setosa.io/ev/image-kernels/ (last visited Aug. 1, 2018).

[90] Yann LeCun, Corinna Cortex & Christopher J.C. Burges, *The MNIST Database*, http://yann.lecun.com/exdb/mnist/ (last visited Aug. 1, 2018).

We can then give pixels in the drawing a weight between 0.0 and 1.0, depending on the pixel-value at each location. The closer to white, the higher the number; the closer to black, the lower the number.

Imagine the simplified case where we map our image to a five-by-five grid.[91] The result would look something like the following, where we can see a black area (i.e., "0.0") with a line of white down the center (i.e., "1.0"). This starting grid is known as our "input."[92]

Input				
0.0	0.0	1.0	0.0	0.0
0.0	0.0	1.0	0.0	0.0
0.0	0.0	1.0	0.0	0.0
0.0	0.0	1.0	0.0	0.0
0.0	0.0	1.0	0.0	0.0

Figure B[93]

The neural network would then run the grid through a series of "convolutions."[94] A convolution is simply a filter[95] or filters applied to the numbers that make up the grid.[96] Here is an example filter which happens to highlight the vertical lines in the grid:

Filter

[91] *See* Howard, *supra* note 87 (using an excel spreadsheet to map a larger grid); *see also* Jeremy Howard, *deeplearning1*, GITHUB (Dec. 31, 2016), https://github.com/fastai/courses/tree/master/deeplearning1/excel.

[92] *See* Howard, *supra* note 87.

[93] Input, a four-by-four grid with "activated" points being those where the pixel values move from 0 to 1. For a larger version of this, see Jean-Carlos Paredes, Un*derstanding Neural Networks Using Excel*, MEDIUM: TOWARDS DATA SCIENCE (Nov. 19, 2017), https://towardsdatascience.com/understanding-convolutions-using-excel-886ca0a964b7.

[94] *See* Alex Krizhevsky, Ilya Sutskever & Geoffrey E. Hinton, *ImageNet Classification with Deep Convolutional Neural Networks*, *in* 25 ADVANCES NEURAL INFO. PROCESSING SYS. 1, 2-3 (2012), http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf.

[95] A filter is a sequence of mathematical operations on the grid. *See, e.g.*, Adit Deshpande, *A Beginner's Guide to Understanding Convolutional Neural Networks, https://adeshpande3.github.io/A-Beginner%27s-Guide-To-Understanding-Convolutional-Neural-Networks/* (last visited Aug. 1, 2018) ("Now, the best way to explain a [convolutional] layer is to imagine a flashlight that is shining over the top left of the image. Let's say that the light this flashlight shines covers a 5 x 5 area. And now, let's imagine this flashlight sliding across all the areas of the input image. In machine learning terms, this flashlight is called a filter (or sometimes referred to as a neuron or a kernel) and the region that it is shining over is called the receptive field. Now this filter is also an array of numbers (the numbers are called weights or parameters). . . . As the filter is sliding, or convolving, around the input image, it is multiplying the values in the filter with the original pixel values of the image (aka computing element[-]wise multiplications).").

[96] *See* Howard, *supra* note 87.

1	0	-1
1	0	-1
1	0	-1

Figure C[97]

To apply the filter, we multiply each cell in the input against each cell in the filter and then find the sum (i.e., the top-right 3x3 grid would be: 1*1 + 0*0 + 0*-1 for the top row plus the second and third row in the same fashion).[98] If the sum is negative, we can use a zero in its place.[99] This is what the grid looks like on the third iteration, this three-by-three slice becoming the sum of each multiplication (i.e., 3).

Filter Applied to Top-Right Corner				
0.0	0.0	1.0 * 1	0.0 * 0	0.0 * -1
0.0	0.0	1.0 * 1	0.0 * 0	0.0 * -1
0.0	0.0	1.0 * 1	0.0 * 0	0.0 * -1
0.0	0.0	1.0	0.0	0.0
0.0	0.0	1.0	0.0	0.0

Figure D

The end result of this convolution (i.e., a single layer in our convolutional neural network) is the following smaller grid:

Result of Single-Layer Convolution		
0	0	3
0	0	3
0	0	3

Figure E

A CNN may have any number of convolutions, each used in a stacking fashion to highlight some aspect of the pixels that make up the drawing.[100] These convolutions, in

[97] Filters may also be called kernels. Using a pre-trained model, these filters are pre-determined.

[98] The filter is applied to the top-left corner and moved one pixel at a time until the end is reached. It is then moved one pixel down and the row is repeated, until the end of the input is reached. *See* Jonathan Hui, *Convolutional Neural Networks (CNN) Tutorial*, JONATHAN HI BLOG (Mar. 16, 2017) https://jhui.github.io/2017/03/16/CNN-Convolutional-neural-network/

[99] This is known as the rectified linear unit, simply the maximum between zero and the result of the convolution (i.e., negative numbers are replaced with zero). *See* Howard, *supra* note 87; *see also* Yee Whye The & Geoffrey E. Hinton, *Rate-Coded Restricted Boltzmann Machines for Face Recognition*, *in* ADVANCES NEURAL INFO. PROCESSING SYS. (2000), https://papers.nips.cc/paper/1886-rate-coded-restricted-boltzmann-machines-for-face-recognition.pdf.

[100] These may focus on the outer edges of the drawing, the horizontal edges, or practically anything imaginable. *See* Soham Chatterjee, *Different Kinds of Convolutional Filters*, SAAMA (Dec. 20, 2017), https://www.saama.com/blog/different-kinds-convolutional-filters/ ("By being able to learn the values of different filters, CNNs can find more meaning from images that humans and human designed filters might not be able to find. More often than not, we see the filters in a convolutional layer learn to detect abstract

combination with other layers,[101] make up the architecture of the model (i.e., stacked layers form the "deep" part of deep learning). For example, the next layer may be a pooling layer, for instance a two-by-two pooling layer, where we half the dimension of the grid by taking the maximum number out of each four-cell block. Another common method would be a fully-connected layer, in which we find the matrix product of the grid by multiplying the full grid with a layer of pre-determined weights.[102] Here is our example with a fully-connected layer:

Result of Single-Layer Convolution		
0	0	3
0	0	3
0	0	3

Hypothetical Weights		
.2	0	.5
.3	.1	.5
.1	0	.5

Fully Connected Layer		
0 * .2	0 * 0	3 * .5
0 * .3	0 * .1	3 * .5
0 * .1	0 * 0	3 * .5

Result of Fully Connected Layer
4.5

Figure F

The end result of the fully-connected layer is the lone number 4.5 (i.e., the far-right column would be 3 * .5 + 3 * .5 + 3 * .5 = 4.5). If this is the last step in our architecture, then this would be known as the "output," with each layer in-between input-to-output known as a "hidden" layer.[103] In practice, a fully-connected layer would be applied several times (using various weights each time), resulting in a series of individual numbers.[104] And those numbers form the basis for our prediction of which digit the original image represents.

In the case of digits 0-9, we would have ten lone numbers in the end, as owed to our iterations of a fully-connected layer.[105] Then, moving from a single number to probabilities, a common function to use would be softmax.[106] Softmax starts by finding the

concepts, like the boundary of a face or the shoulders of a person. By stacking layers of convolutions on top of each other, we can get more abstract and in-depth information from a CNN."); Deshpande, *supra* note 95.

[101] Another layer may be a fully-connected layer where each digit in the grid is multiplied by a pre-specified weight.

[102] Howard, *supra* note 87.

[103] *Id.*

[104] *Id.*

[105] *Id.*

[106] Softmax will first use the exponential function of each number (i.e., e to the power of each number—$e^{4.5}$) and then find the sum of each of those numbers and divide the result of the exponential function

exponential function of each of those numbers (e.g., $e^{4.5} = 90.01$), finds the sum of the result, and then divides each number by the sum. Essentially, we are removing any negative numbers (via exponentiation) and distinguishing more likely predictions from less likely ones.[107] The result will be a set of probabilities adding up to one with (hopefully) the correct number having the highest probability. If we are using a robust model, we may anticipate an accuracy of nearly 100 percent.[108] We are now done with the pre-trained CNN and have our prediction in hand.

Turning to training the model, it suffices to say this is the process of gradually improving the filters and weights discussed above.[109] Starting out, these parameters are randomly assigned. Then, typically through a process known as stochastic gradient descent,[110] the parameters are optimized to minimize a particular loss function (e.g., during each training loop we check the model's predictions against known outcomes, the difference between the two being the loss).[111] Stated otherwise, the weights and filters are adjusted to find the optimal combination of numbers to achieve a desired result, such as the identification of distinguishing features in hand-drawn digits.

But this begs another question relevant to synthetic data. How would this process work if we were using text instead of pixels?

2. Recurrent Neural Network

Switching gears from images to text, one attribute remains dispositive: context. A CNN used to identify hand-drawn characters walks through layers of convolutions in an independent fashion.[112] For example, even after our final, fully-connected layer, we could have applied another layer of convolutions followed by another fully-connected layer. Each layer is not dependent on the other. Sentences, on the other hand, are different because speech relies heavily on context, heavily on order. And context comes from memory—something a CNN lacks. However, this does not mean the two have nothing in common. In fact, just the opposite.

---

by the sum. *See* John S. Bridle, *Probabilistic Interpretation of Feedforward Classification Network Outputs, with Relationships to Statistical Pattern Recognition*, 68 NEUROCOMPUTING 227, 231-232 (1989).

[107] *See* Kulbear, *ReLU and Softmax Activation Functions*, GITHUB (Feb. 11, 2017), https://github.com/Kulbear/deep-learning-nano-foundation/wiki/ReLU-and-Softmax-Activation-Functions (last visited Aug. 1, 2018) ("The softmax function squashes the outputs of each unit to be between 0 and 1, just like a sigmoid function. But it also divides each output such that the total sum of the outputs is equal to 1 . . . . The output of the softmax function is equivalent to a categorical probability distribution, it tells you the probability that any of the classes are true.").

[108] *See* Li Wan et al., *Regularization of Neural Networks Using DropConnect*, 28 PROC. 30TH INT'L CONFERENCE MACHINE LEARNING, 1, 6 (2013) (resulting in a .21% error rate).

[109] For an overview, see *How Neural Networks are Trained*, MACHINE LEARNING FOR ARTISTS https://ml4a.github.io/ml4a/how_neural_networks_are_trained/ (last visited Aug. 1, 2018).

[110] *See generally* Sebastian Ruder, *An Overview of Gradient Descent Optimization Algorithms*, 1, 2 (2017), https://arxiv.org/pdf/1609.04747.pdf (last visited Aug. 1, 2018).

[111] *See* Vitaly Bushaev, *How do We 'Train' Neural Networks*, MEDIUM: TOWARDS DATA SCIENCE (Nov. 15, 2017), https://towardsdatascience.com/how-do-we-train-neural-networks-edd985562b73 (last visited Aug. 1, 2018); Howard, *supra* note 87.

[112] Independence implies that one element does not affect another; for example, the chance of rolling a six on a dice two consecutive times is not impacted by whether or not you roll a six the first time. Each of the rolls is independent of the other.

A recurrent neural network (RNN) uses the same layer-by-layer approach as the CNN (i.e., deep learning).[113] Here, the prediction may be the next word in a sentence or the next character in a word, but the iterative stepping remains the same. We start with an input, use a filter, and receive a result.[114] In more mathematical terms, the input is run through a filter which produces activations in a narrowing fashion until a function like softmax is used to create predictions.[115]

The main difference for an RNN is that new input is added after the first filter.[116] For instance, if we were using a character-based model, the first input would be the first character of a word (e.g., the character "t"), followed by the application of a filter, followed by the next character (e.g., the character "h") added through something like a matrix multiplication.[117] That is, instead of simply adding more layers of convolutions, we add the equivalent of another picture to the mix. This process allows the network to gain memory. The end-result, like the final 4.5 we produced in the CNN example, is dependent not only on a single input entered in the beginning, but also the intermediate input injected during the hidden layers.

This, finally, brings us to the last concept in our nutshell: Generative Adversarial Networks (GANs).[118] The beauty of a GAN—a recent invention[119]—is its ability to generate *similar* data.[120] The newness of GANs, however, should not be mistaken for novelty; GANs are built upon the exact same foundations we have seen in the previous sections.

3. Generative Adversarial Network

The easiest way to think of a GAN is through the production of counterfeit money.[121] A counterfeiter (i.e., the generator) attempts to produce the most realistic-looking fake money, while a detective (i.e., the discriminator) seeks to spot the fraudulent activity. In this way (i.e., using a generator *and* discriminator[122]), a GAN uses two models pitted against each other in an iterative loop.[123] Notably, GANs may rely on either type of neural network, a CNN or an RNN, for a foundation. The important feature is rather in the interplay between the two roles.[124]

---

[113] *See* Jeremy Howard, *Lesson 6: Deep Learning 2018*, YOUTUBE (Dec. 30, 2017); *see also* Hiromi Suenaga, *Deep Learning 2: Part 1 Lesson 6*, MEDIUM (Jan. 10, 2018).

[114] Howard, *supra* note 113.

[115] *Id.*

[116] *Id.*

[117] *Id.*

[118] *See generally* Ian J. Goodfellow, et al., *Generative adversarial Nets*, 27 Advances Neural Information Processing Systems (2014); *see also* Martin Arjovsky, Soumith Chintala & Léon Bottou, *Wasserstein GAN*, https://arxiv.org/pdf/1701.07875.pdf (last visited Aug. 1, 2018).

[119] In fact, GANs are one of the reasons synthetic data has recently received attention. *See* Goodfellow, *supra* note 118.

[120] *Id.*

[121] *Id.* at 1.

[122] Unlike the CNN or RNN which typically only use one model. *See supra* subsections II.A.1-2.

[123] *Id.* at 1-2.

[124] Notably, there are no constraints on the specific types of models used in the GAN.

Specifically, the generator's measurement of success depends on the detective's ability to correctly identify falsity, and vice-versa. If the game is played over and over, assuming theoretically ideal conditions, an equilibrium is reached in which the discriminator is unable to distinguish between real and fake data.[125] This is why GANs are becoming the go-to for synthetic data generation.[126] The ability to generate similar data (i.e., deep fakes[127]) with better results than seen before.

Taking a step back, it is important to remember the core[128] technology at hand: the neural network.[129] These are the layers of convolutions discussed in the first example.[130] Additionally, these networks may be paired with differential privacy.[131] The addition of differential privacy to neural networks drew interest as early as 2016.[132] However, it was not until 2018 that researchers realized the potential implications and advantages of applying the technique to GANs.[133] Without delving into the technical details, privacy is added by implementing noise into the training procedure we saw before. The proof and demonstration of this concept is beyond the scope of this article. With that understanding at hand we may now move onto our case study demonstrating the feasibility and utility of synthetic data generated through a GAN.

B. Case Study: Generating and Evaluating Synthetic Data

---

[125] The end result here is an equilibrium between the generator and discriminator, known as Nash equilibrium: "In a Nash equilibrium, every person in a group makes the best decision for herself, based on what she thinks the others will do. And no-one can do better by changing strategy: every member of the group is doing as well as they possibly can." The Economist, https://www.economist.com/the-economist-explains/2016/09/06/what-is-the-nash-equilibrium-and-why-does-it-matter (Sept. 7, 2016); *see also* Tim Salimans, Ian Goodfellow, Wojciech Zaremba & Vicki Cheung, *Improved Techniques for Training GANs*, *in* NIPS 2 (2016).

[126] *See infra* note 148 and accompanying text.

[127] *See* James Vincent, *All of These Faces Are Fake Celebrities Spawned by AI*, THEVERGE (Oct. 30, 2017, 7:05 AM), https://www.theverge.com/2017/10/30/16569402/ai-generate-fake-faces-celebs-nvidia-gan ("By working together, these two networks can produce some startlingly good fakes. And not just faces either — everyday objects and landscapes can also be created. The generator networks produce[] the images, the discriminator checks them, and then the generator improves its output accordingly. Essentially, the system is teaching itself.").

[128] And the core of the core is this, the universal approximation theorem: *any* real-world problem which is able to be mathematically mapped as a continuous function can be solved with nearly-perfect accuracy by using a neural network. And in more mathematical terms, "neural networks with a single hidden layer can be used to approximate any continuous function to any desired precision." Michael Nielsen, *Neural Networks and Deep Learning*, http://neuralnetworksanddeeplearning.com/chap4.html (last visited Aug. 1, 2018); *see also* Howard *supra* note 129.

[129] Jeremy Howard, *Lesson 1: Deep Learning 2018*, YOUTUBE (Dec. 30, 2017), https://www.youtube.com/watch?v=IPBSB1HLNLo.

[130] *See supra* subsection II.A.1.

[131] *See generally* Nissim et al., *supra* note 65.

[132] *See* Martín Abadi et al., *Deep Learning with Differential Privacy*, https://arxiv.org/pdf/1607.00133.pdf (last visited Aug. 1, 2018) (applying differential privacy techniques to machine learning language modeling).

[133] *See* Liyang Xie et al., *Differentially Private Generative Adversarial Network*, https://arxiv.org/pdf/1802.06739.pdf (last visited Aug. 7, 2018).

The first step in producing synthetic data is to acquire an original, raw dataset, which is often difficult. Consider the area of insider threat detection. An insider threat is an individual or group of individuals who betray the trust of the organization and expose information about the organization to others for motives often misaligned with those of the company. The area commanded international spotlight when a Booz Allen Hamilton contractor, Edward Snowden, shared classified documents from the National Security Agency (NSA).[134] Mr. Snowden's leak not only spurred widespread concern, as dealings of the NSA became available to the public, but also caused research to explode on how to thwart insider threats.[135] Paradoxically, despite interest in insider threat detection, it is an area devoid of data. This is because the data necessary to detect an insider threat is very fine grained,[136] causing privacy concerns. With this in mind, we opted to use a previously-attained, private dataset maintained by Columbia University's Intrusion Detection Lab, the West Point dataset.[137]

1. Database Selection and Synthetic Data Generation

The West Point dataset tracks the computer interactions of 63 West Point cadets over a one-month period.[138] The original data was acquired by having each cadet install software on their machine collecting information on *all* aspects of use (i.e., editing documents, viewing webpages, opening files, and any other activity occurring on the computer). This resulted in a wide variety of comparable relationships. For example, the number of website visits per user per day (*see* figure below) or the time spent online versus time spent writing documents.[139]

[134] *See* Ewen Macaskill & Gabriel Dance, *NSA Files: Decoded*, THEGUARDIAN (Nov. 1, 2013), https://www.theguardian.com/world/interactive/2013/nov/01/snowden-nsa-files-surveillance-revelations-decoded#section/1.

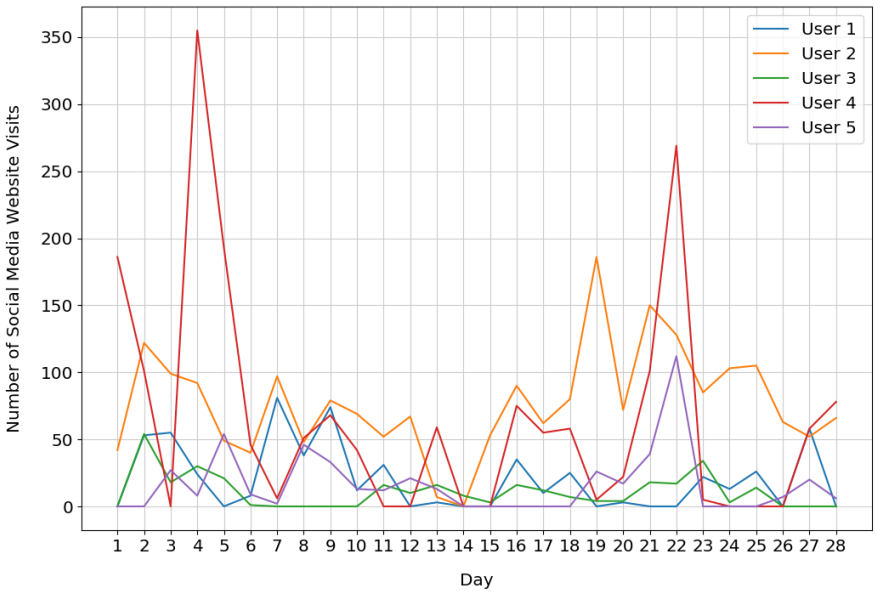
[135] *See, e.g.*, INSIDER ATTACK AND CYBER SECURITY: BEYOND THE ATTACKER (Salvatore J. Stolfo, Steven M. Bellovin, Shlomo Hershkop & Angelos D. Keromytis eds., 2008).

[136] Detecting insider threats is only attained from thorough, intimate access to a user's interactions with their computer.

[137] *See* Preetam Dutta, Gabriel Ryan, Aleksander Zieba & Salvatore Stolfo, *Simulated User Bots: Real Time Testing of Insider Threat Detection Systems*, *in* WORKSHOP ON RESEARCH FOR INSIDER (2018). The data from the West Point cadets was gathered under an IRB-approved protocol.

[138] *See id.* The earliest installations of this software occurred on January 15, 2015, and the latest installations were on February 13, 2015. Each user had a participant and device Windows System ID and Unique ID number. The cadets had up to three extraction dates for the data from their machines, ranging from February 10, 2015, for the first pull to March 12, 2015, for the last data collection. Notably, the data collected did suffer from periods of technical difficulties due to the data collection software agent. However, despite this fact, the data still provides a valuable resource and a wealth of information regarding normal user behavior.

[139] Although the users are from a homogenous population with similar roles, the users have diverse usage habits. Data heterogeneity is an important characteristic to consider when modeling and analyzing data since it is intuitively and pragmatically impossible to differentiate individuals if they all resemble one another.



Notably, it is easy to see how users may be differentiated by their actions—some often use social media (i.e., users four and two) and others do not (i.e., users three and five).

The next step is to select the type of neural network to be used. Given the type of data contained in the West Point dataset (e.g., the columns of the dataset contain user ID, timestamp, action, detail), a RNN is the best neural network architecture.[140] Additionally, it would be most ideal for the RNN to take into account various prior actions when making predictions. For this reason, we used a specific type of RNN known as Long Short-Term Memory (LSTM).[141] This type of RNN leverages not only an RNN's ability to maintain some form of memory, but also the ability to remember important events over varying time periods. We primed the LSTM with the following inputs: previous event, previous time step, previous day-of-week, and previous hour-of-day. Additionally, as illustrated above, we used the GAN technique and pitted a generator against a discriminator. The generator was able to produce a predicted next event and predicted next time step while the discriminator was able to check these predictions for accuracy. This process formed the basis for our synthetic data.

---

[140] *See supra* Subsection II.A.2 (explaining how RNNs are typically used for text-based generation).

[141] *See generally* Sepp Hochreiter & Jürgen Schmidhuber, *Long Short-Term Memory*, 9 NEURAL COMPUTATION 1 (1997); *see also* Christopher Olah, *Understanding LSTM Networks*, COLAH'S BLOG (Aug. 27, 2015), http://colah.github.io/posts/2015-08-Understanding-LSTMs/ ("Sometimes, we only need to look at recent information to perform the present task. For example, consider a language model trying to predict the next word based on the previous ones. If we are trying to predict the last word in 'the clouds are in the [next word],' we don't need any further context—it's pretty obvious the next word is going to be sky. In such cases, where the gap between the relevant information and the place that it's needed is small, RNNs can learn to use the past information. But there are also cases where we need more context. Consider trying to predict the last word in the text 'I grew up in France . . . I speak fluent [next word].' Recent information suggests that the next word is probably the name of a language, but if we want to narrow down which language, we need the context of France, from further back. It's entirely possible for the gap between the relevant information and the point where it is needed to become very large. Unfortunately, as that gap grows, RNNs become unable to learn to connect the information. . . . Long Short Term Memory networks—usually just called 'LSTMs'– are a special kind of RNN, capable of learning long-term dependencies.").

2. Evaluation of Synthetic Data

To assess the efficacy of our generated data, we clustered both the raw data (i.e., the cadets' computer interactions) and synthetic data around similar actions—i.e., intuitively, the trail of actions left by users naturally groups around commonalities like frequency of social media use. To accomplish this task, we used term frequency inverse document frequency (TF-IDF).[142] This metric looks at the frequency of word-use in a document. After grouping, we could then assess the similarities or difference between the raw and synthetic data.

As expected, when checking the clustered synthetic groups against the clustered raw groups we found little to no variance. In other words, our synthetic data, for all but privacy infractions, was the same.[143] What is more, even beyond our case study, similar research concurs in our results: Synthetic data is a valid alternative to original data.[144]

A budding body of research has found that when comparing analysis using original data to analysis using synthetic data, for the most part, the results are indistinguishable, even by domain experts.[145] In fact, researchers have gone so far as to conclude that "scientists can be as productive with synthesized data as they can with control data."[146] Moreover, other publications suggest that in the face of reidentification (i.e., the thorn in the side of deidentification), synthetic datasets leave *no* room for leakage.[147] In summary, the "usefulness" of synthetic data has been validated by not only our work, but also the work of others.

---

[142] *See* Gerard Salton & Christopher Buckley, *Term-weighting approaches in automatic text retrieval*, *in* INFORMATION PROCESSING & MANAGEMENT 24, 5 (1988). This process initially looks at the frequency with which words occur in a document (i.e., the text-based actions assigned to each user). The problem is that insignificant words often occur with high frequency (such as "the" or "a"). *Id.* TF-IDF therefore pivots to ascribe a higher weight to less-common words and a lower weight to more-common words. *Id.* In the end, the combination of these two components yields a useful metric for grouping users by similar actions. For specifics, we used a Gaussian Mixture Model that minimized the Bayesian Information Criterion. *See* Scott Chen & P.S. Gopalakrishnan, *Clustering via the Bayesian information criterion with applications in speech recognition, in* ACOUSTICS, SPEECH & SIGNAL PROCESSING (1998); *see also* Charu C. Aggarwal & Philip S. Yu, *A Condensation Approach to Privacy Preserving Data Mining*, in PROC. 9TH INT'L CONFERENCE ON EXTENDING DATABASE TECH.,183-99 (2004).

[143] *See* Dutta, Ryan, Zieba & Stolfo, *supra* note 137.

[144] *See generally* Neha Patki, Roy Wedge & Kalyan Veeramachaneni, *The Synthetic Data Vault*, *in* INTERNATIONAL CONFERENCE ON DATA SCIENCE AND ADVANCED ANALYTICS 399, 400-10 (2016) (demonstrating a technique—the synthetic data vault—used to create synthetic data form five publicly available datasets).

[145] *See* Patki, Wedge & Veeramachaneni, *supra* note 144 ("in 7 out of 17 comparisons, we found no significant difference between the accuracy of features developed on the control dataset vs. those developed on some version of the synthesized data…"); Edward Choi, et al., *Generating Multi-Label Discrete Patient Records Using Generative Adversarial Networks*, 68 PROC. MACH. LEARNING HEALTHCARE, 1, 10 (2018), https://arxiv.org/pdf/1703.06490.pdf ("The findings suggest that medGAN's synthetic data are generally indistinguishable to a human doctor except for several outliers. In those cases, the fake records identified by the doctor either lacked appropriate medication codes, or had both male-related codes (*e.g.* prostate cancer) and female-related codes (*e.g.* menopausal disorders) in the same record.").

[146] *See* Patki, Wedge & Veeramachaneni, *supra* note 144.

[147] *See* Noseong Park, et al., *Data Synthesis Based on Generative Adversarial Networks*, *in* PVLDB, 1071, 1074 (2018), http://www.vldb.org/pvldb/vol11/p1071-park.pdf.

However, this is not to say that synthetic data is the "silver bullet" data scientists and privacy activists have been searching for.[148] As the deidentification saying goes, just because the dataset appears anonymous does not mean it is. For synthetic data, this means that without adding privacy-preserving features like differential privacy, there still remains risk of data leakage.

C. Risk of Data Leakage: Limitations of Synthetic Data

Synthetic data alone is not the end-game for database privacy: it too has limitations. These include the uniqueness of data used to train the machine learning model, the ability of an attacker to use adversarial machine learning techniques, and the type of questions being asked of the dataset. Moreover, as discussed in Part III, the ceiling on each of these

[148] Matt Fredrikson, Somesh Jha & Thomas Ristenpart, *Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures*, in CCS (2015), https://www.cs.cmu.edu/~mfredrik/papers/fjr2015ccs.pdf ("Computing systems increasingly incorporate machine learning (ML) algorithms in order to provide predictions of lifestyle choices, medical diagnoses, facial recognition, and more. The need for easy 'push-button' ML has even prompted a number of companies to build ML-as-a-service cloud systems . . . . The features used by these models, and queried via APIs to make predictions, often represent sensitive information. In facial recognition, the features are the individual pixels of a picture of a person's face. In lifestyle surveys, features may contain sensitive information, such as the sexual habits of respondents. In the context of these services, a clear threat is that providers might be poor stewards of sensitive data, allowing training data or query logs to fall prey to insider attacks or exposure via system compromises. [We] introduce new attacks that infer sensitive features used as inputs to decision tree models, as well as attacks that recover images from API access to facial recognition services. . . . One example from our facial recognition attacks is depicted in Figure 1: an attacker can produce a recognizable image of a person, given only API access to a facial recognition system and the name of the person whose face is recognized by it."); Vincent Bindschaedler, Reza Shokri & Carl A. Gunter, *Plausible Deniability for Privacy-Preserving Data Synthesis (Extended Version)*, *in* CRYPTOGRAPHY & SECURITY (2017), https://arxiv.org/pdf/1708.07975.pdf ("[T]he major open problem is how to generate synthetic full data records with *provable* privacy, that experimentally can achieve acceptable utility in various statistical analytics and machine learning settings. In this paper, we fill this major gap in data privacy by proposing a generic theoretical framework for generating synthetic data in a privacy-preserving manner.") (emphasis in original); Chong Huang et al., *Context-Aware Generative Adversarial Privacy*, in ENTROPY (2017), https://web.stanford.edu/~kairouzp/gap_entropy.pdf (implementing a "context-aware privacy framework"); Brett K. Beaulieu-Jones, Zhiwei Steven Wu, Chris Williams & Casey S. Greene, *Privacy-Preserving Generative Deep Neural Networks Supporting Clinical Data Sharing*, BIORXIV (2017), https://www.biorxiv.org/content/biorxiv/early/2017/07/05/159756.full.pdf (acknowledging the failure to non-sanitized synthetic data to hold up against even simple privacy attacks—and therefore incorporating differential privacy into their machine learning models); Aleksei Triastcyn & Boi Faltings, *Generating Artificial Data for Private Deep Learning*, https://arxiv.org/pdf/1803.03148.pdf (last visited Aug. 1, 2018) ("Following recent advancements in deep learning, more and more people and companies get interested in putting their data in use and employ [machine learning models] to generate a wide range of benefits that span financial, social, medical, security, and other aspects. At the same time, however, such models are able to capture a fine level of detail in training data, potentially compromising privacy of individuals whose features sharply differ from others. Recent research . . . suggests that even without access to internal model parameters, by using hill climbing on output probabilities of a neural network, it is possible to recover (up to a certain degree) individual examples (e.g. faces) from a training set. The latter result is especially disturbing knowing that deep learning models are becoming an integral part of our lives, making its way to phones, smart watches, cars, and appliances. And since these models are often trained on customers' data, such training set recovery techniques endanger privacy even without access to the manufacturer's servers where these models are being trained.").

limitations hinges on the particular law being applied. As fodder for that later legal analysis, each of these limitations are discussed in turn.

1. Too Individualized

First off, one inherent characteristic of synthetic datasets is that they may "leak" information.[149] In computer science parlance, this is referred to as overfitting a model, which may result in particular data being "leaked." Consider the graph of social media use above, showing the outlier count of over 350 visits to social media websites by user four. A machine learning model must take this into account.[150] Consequently, that fact will be reflected in the model, and may show up in some synthetic records.[151] Under an absolute definition of privacy—no leakage whatsoever, in any reasonable amount of time—this latter result is unacceptable, since only that one person used social media that much.

We thus have a dilemma: even with a reasonable distribution of input records (i.e., one that does not exhibit habitual cases such as only one party performing a particular action with sufficient frequency to sway the model) there may be *at least some* risk that some quantity of the original data could be leaked. Moreover, bounding that leakage by quantifying "how hard" it is to reverse the model to find a leak is an open-ended problem.[152]

Ideally, a technical solution could be developed. Although one solution might be to use a synthesizing algorithm to replace the actual cadet's anomalous behavior with a different one, this is simply anonymization, the very technology whose failures we are

---

[149] *See* Samuel Yeom, Irene Giacomelli, Matt Fredrikson & Somesh Jha, *Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting*, *in* COMP. SEC. FOUNDATIONS SYMPOSIUM 1, 22-23 (2018), https://arxiv.org/pdf/1709.01604.pdf (finding privacy leaks in machine learning models as owed to overfitting—and even other, more subtle features of the models); Tyler Hunt et al., *Chiron: Privacy-Preserving Machine Learning as a Service*, https://arxiv.org/pdf/1803.05961.pdf (last visited Aug. 1, 2018).

[150] *See* Nicolas Papernot & Ian Goodfellow, *Privacy and Machine Learning: Two Unexpected Allies?*, CLEVERHANS-BLOG (Aprl. 29, 2018), http://www.cleverhans.io/privacy/2018/04/29/privacy-and-machine-learning.html ("Machine learning algorithms work by studying a lot of data and updating their parameters to encode the relationships in that data. Ideally, we would like the parameters of these machine learning models to encode general patterns ('patients who smoke are more likely to have heart disease') rather than facts about specific training examples ('Jane Smith has heart disease'). Unfortunately, machine learning algorithms do not learn to ignore these specifics by default. If we want to use machine learning to solve an important task, like making a cancer diagnosis model, then when we publish that machine learning model (for example, by making an open source cancer diagnosis model for doctors all over the world to use) we might also inadvertently reveal information about the training set. A malicious attacker might be able to inspect the published model and learn private information about Jane Smith.").

[151] Though social media visits are relatively benign, consider the frequent visitation to an incredibly specific website such as Delta Junction Dating, a dating website geared toward the roughly 850-person town of Delta Junction, Alaska. *See* https://www.connectingsingles.com/dating_4_35641_0_1/delta_junction_dating.htm (last visited Aug. 1, 2018). If this were part of the analysis it may work its way into the synthetic dataset. And in some sense, this is privacy leakage: real data has appeared. In another sense, it is not; someone receiving the data could not easily tell which records are real and which are synthetic. However, if a single record appears frequently in the generated data, it would be likely assumed to be a reflection of the actual input data.

[152] *But see* Florian Tramèr et al., *Stealing Machine Learning Models via Prediction APIs*, *in* USENIX SECURITY (2016), https://arxiv.org/pdf/1609.02943.pdf (engaging in something similar to reversal by showing the successful duplication of machine learning model functionality given only pre-trained models in query-based interactions).

trying to avoid.[153] Other techniques falling into this category—i.e., regularization methods like weight-decay or drop out (i.e., discarding certain pieces of potentially-sensitive data during training)—are equally ill suited.[154]

A better solution here is to use differential privacy in combination with synthetic data generation.[155] Though a relatively new technique, the results are very promising.[156] Utility is sustained though data generation and privacy is obtained—up to a certain threshold—via the robust guarantees of differential privacy.[157] However, setting that threshold[158] will be key in achieving a balance between privacy and utility. These two pillars remain a tradeoff. For this reason, while adding differential privacy to synthetic data generation does help in the data leakage sense, it does not offer a silver bullet.[159]

2. Adversarial Machine Learning

A second limitation to synthetic data concerns situations where an attacker attempts to exert influence over the process of generating synthetic data to force leakage. These attacks are known generally as adversarial machine learning.[160] Notably, these attacks require more than the mere possession of synthetic data. Rather, the ability to have access to the model used to generate synthetic data (e.g., the particular convolutions and weights used in the CNN example given above) is a prerequisite.

Consider a pre-trained image recognition model similar to the one demonstrated above for the classification of digits, but aimed at faces. Recent research demonstrates that if the attacker has access to this model, and a little auxiliary information such as a person's name, the faces of those used to train the model could be uncovered.[161] Along those same

---

[153] *See* Ohm *supra* note 5.

[154] Nicholas Carlini et al., *The Secret Sharer: Measuring Unintended Neural Network Memorization and Extracting Secrets* 1, 11-12, https://arxiv.org/pdf/1802.08232.pdf (last visited Aug. 1, 2018).

[155] Cynthia Dwork & Vitaly Feldman, *Privacy-Preserving Prediction*, *in* CONFERENCE LEARNING THEORY (2018), https://arxiv.org/pdf/1803.10266.pdf; Carlini, *supra* note 154 at 13 (finding not only that neural networks memorize and generate secrets even when secrets are alarmingly rare, but that the use of differential privacy in combination with training neural network works better than any other sanitization technique).

[156] H. Brendan McMahan, Daniel Ramage, Kunal Talwar, Li Zhang, *Learning Differentially Private Recurrent Language Models*, *in* ICLR 2018, https://arxiv.org/pdf/1710.06963.pdf.

[157] *See* Abadi et al., *supra* note 132 (applying differential privacy techniques to machine learning language modeling).

[158] *See infra* notes 78-73 and accompanying text (referring to the epsilon parameter).

[159] *See supra* section I.B.3.ii

[160] *See* Ling Huang et al., *Adversarial Machine Learning*, *in* AI SEC. 1, 1 (2011); *see also* Alexey Kurakin, Ian J. Goodfellow & Samy Bengio, *Adversarial Machine Learning at Scale*, *in* ICLR 1, 1 (2017), https://arxiv.org/pdf/1611.01236.pdf; Gamaleldin F. Elsayed et al., *Adversarial Examples that Fool both Computer Vision and Time-Limited Humans*, https://arxiv.org/pdf/1802.08195.pdf (last visited Aug. 1, 2018) (tricking image recognition software into thinking a panda was a type of monkey); Ian J. Goodfellow, Jonathon Shlens & Christian Szegedy, *Explaining and Harnessing Adversarial Examples*, *in* ICLR 1, 3 (2015); Julia Evans, *How to Trick a Neural Network Into Thinking a Panda Is a Vulture*, CODEWORDS, https://codewords.recurse.com/issues/five/why-do-neural-networks-think-a-panda-is-a-vulture; s*ee* Ivan Evtimov et al., *Robust Physical-World Attacks on Machine Learning Models*, https://arxiv.org/pdf/1707.08945.pdf (last visited Aug. 1, 2018).

[161] *See* Fredrikson, Jha & Ristenpart, *supra note* 148; Matthew Fredrikson, Eric Lantz & Somesh Jha, *Privacy in Pharmacogenetics: An End-to-End Study of Personalized Warfarin Dosing*, *in* USENIX 17,

lines, other research[162] goes even further to suggest that if the attacker has full access to the model,[163] then up to 70% of the original data used to train the model could be uncovered. Not only that, but even with limited input-output access only,[164] the attacker could learn whether a data point was "in" the data used to train the model.[165]

The point here is that while synthetic data itself may escape the reidentification woes, not all aspects of its use are invulnerable. In particular, sharing machine learning models used for training on sensitive data should not be taken lightly. Yet, even from this perspective computer science literature points to differential privacy.[166] In fact, out of many possible solutions to this model-access problem, differential privacy has been noted as only one to sufficiently protect privacy while maintaining utility.[167] In this sense, differential privacy provides both a way to escape data leakage and adversarial machine learning.[168]

---

17 (2014), https://www.usenix.org/system/files/conference/usenixsecurity14/sec14-paper-fredrikson-privacy.pdf ("Performing an in-depth case study on privacy in personalized warfarin dosing, we show that suggested models carry privacy risks, in particular because attackers can perform what we call model inversion: an attacker, given the model and some demographic information about a patient, can predict the patient's genetic markers"); *see also* Congzheng Song, Rhomas Ristenpart & Vitaly Shmatikov, *Machine Learning Models that Remember too Much*, in CONFERENCE COMPUTER & COMMUNICATIONS SECURITY (2017), https://arxiv.org/pdf/1709.07886.pdf ("ML cannot be applied blindly to sensitive data, especially if the model-training code is provided by another party. Data holders cannot afford to be ignorant of the inner workings of ML systems if they intend to make the resulting models available to other users, directly or indirectly. Whenever they use somebody else's ML system or employ ML as a service (even if the service promises not to observe the operation of its algorithms), they should demand to see the code and understand what it is doing.")

[162] *See* Fredrikson, Jha & Ristenpart, *supra* note 161. It is important to note that the authors of this work make several assumptions about the models, which may not necessarily be feasible or realistic. However, the work does highlight the possibility of potentially dangerous leaks. But see Reza Shokri, Marco Stronati, Congzheng Song & Vitaly Shmatikov, *Membership Inference Attacks Against Machine Learning Models*, https://arxiv.org/pdf/1610.05820.pdf (last visited Aug. 1, 2018); Yinhui Long et al., *Understanding Membership Inferences on Well-Generalized Learning Models*, https://arxiv.org/pdf/1802.04889.pdf (last visited Aug. 1, 2018) ("[These] type[s] of attacks can have a significant privacy implication such as re-identifying a cancer patient whose data is used to train a classification model.").

[163] In this sense, we are referring to white-box access. White-box attacks allow the attacker to see "how" the sausage is made—i.e., how the machine learning code is built. The attacker does not have access to the data used in training, but is able to supply training data and see what comes out. *Id.*

[164] This is a black-box attack. The attacker is *only* able to give known input and observe output. The attacker may not see the code manipulating the input or alter that code in any way. *See id.*

[165] *See id.* This is known as the membership inference attack; "[G]iven a data record and black-box access to a model, determine if the record was in the model's training dataset." *Id.*

[166] *See* Long et al., *supra* note 162 ("Differential privacy is a prominent way to formalize privacy against membership inference."); Bindschaedler, Shokri & Gunter, *supra* note 148 (using a form of differential privacy plus synthetic data).

[167] Long et al., *supra* note 162. *But see* Dwork & Feldman, *supra* note 155.

[168] Carlini, *supra* note 154 at 13 ("Only by developing and training a differentially-private model are we able to train models with high utility while protecting against the extraction of secrets in both theory and practice.").

3. Non-Universality

Finally, as with all other methods, synthetic data even with differential privacy is not a cure-all. Indeed, the hard-limit reality of data sanitization is that there will always be some situations when the demands of individuality will not be satisfied by any privacy-preserving technique, no matter how finely tuned. For example, suppose the intended use is a particular statistical query: what percentage of records satisfy some property? If the result must be highly accurate and almost no sanitization is used, then an untrustworthy data custodian may be able to reconstruct the original data with 99% accuracy; conversely, if the results must be private, then even minimal amounts of noise may derail the needed accuracy.[169] The conundrum, though improved, is not completely solved by synthetic data.[170]

III. SYNTHETIC DATA'S LEGALITY

Turning to the legal world, the question remains: is synthetic data legal; does synthetic data protect privacy at least as much as a to-be-applied statute would mandate? Though the answer may appear straightforward—yes, fake data is not real—the nuances of data leakage and the mosaic used to define privacy require a more detailed approach. We therefore group the analysis into two categories: (1) "vanilla" synthetic data; and (2) differentially private synthetic data.

A. Vanilla Synthetic Data

When a generative model is trained without applying any form of data sanitization during or after training[171] the produced data may be deemed "vanilla" synthetic data. The generation process is as bare bones as possible. Data in, data out. Unfortunately, as Section II.C demonstrates, this could result in data leakage: secrets in, secrets out.

Per data leakage, pairing vanilla synthetic data with privacy statues results in both over and under inclusive statutes. Statutes thinking of PII in absolute terms (i.e., no privacy loss is permitted no matter how small the chance of leakage) may not permit synthetic datasets to be shared, even though the likelihood of identifying an individual is low. Conversely, statutes using a less stringent approach may underestimate the risk where more caution is needed. To illustrate each of these points, consider a large training dataset with

---

[169] More precisely, this applies if there are *n* records and the maximum error in a query must be much less than $\sqrt{n}$. *See* Irit Dinur & Kobbi Nissim, *Revealing Information While Preserving Privacy*, *in* SYMPOSIUM ON PRINCIPLES OF DATABASE SYSTEMS (2004).

[170] *See, e.g.*, Briland Hitaj, Giuseppe Ateniese & Fernando Perez-Cruz, *Deep Models Under the GAN: Information Leakage from Collaborative Deep Learning*, *in* CCS 603, 606-07 (2017); Jamie Hayes, Luca Melis, George Danezis & Emiliano De Cristofaro, *LOGAN: Evaluating Privacy Leakage of Generative Models Using Generative Adversarial Network*, 1, 1-2, https://arxiv.org/pdf/1705.07663.pdf (last visited Aug. 1, 2018). On the other hand, see Stefanie Koperniak, *Artificial Data Give the Same Results as Real Data—Without Compromising Privacy*, MIT NEWS (Mar. 3, 2017), http://news.mit.edu/2017/artificial-data-give-same-results-as-real-data-0303.

[171] This also assumes sanitization techniques are not used after the fact, such as applying differential privacy when querying the database.

few outliers. This would give the generative model its best chance of hiding secrets found in the original data.

1. Over Inclusive Privacy

Under one of the strictest privacy standards, HIPAA, privacy is assumed if a database is stripped of all seventeen identifiers believed to uniquely describe individuals, such as name, geographic indicators like zip code, and anything else that could reasonably be considered a "unique identifier." If the dataset lacks these identifiers, then it may be shared freely.

To be sure, synthetic data would *most likely* not contain any of the "real" identifiers found in the training data—all of the unique identifiers outlined by HIPAA would be replaced with machine-generated counterparts.[172] Moreover, considering evenly distributed training data, even assuming the model reproduced a particular datum that turned out to be real does not automatically mean an individual has been identified.[173] That an adversary learns zip code 10004 within the database is real does not mean any of the information related to the zip code is real or that the zip code provides any clues to uncovering the identity of an individual.[174] Synthetic data may not be "joined" with auxiliary information in the same sense as a deidentified dataset—the matchings would pair on fake data.

True enough, some sense of privacy has been lost with the hypothetical zip code leakage, but is this enough to prohibit sharing outright? HIPPA is clear; the dataset must lack all identifiers (or be verifiably secure according to an expert). Seemingly, then, the case is closed and the data may not be shared in the lacking-identifier sense. Yet, consider that even sophisticated computer science methods of extracting secrets from vanilla synthetic data (i.e., attempts to identify "leaks"[175] in the dataset) has been shown to be hit or miss.[176] When researchers used sophisticated methods to extract secrets in a vanilla

[172] In some sense, synthetic data may be compared to virtual child pornography (VCP). Synthetic data, on the surface, like VCP, appears indistinguishable from its original counterpart. From this perspective, then, it might violate privacy standards, just as congress originally prohibited VCP. *See* Child Pornography Prevention Act of 1996, 18 U.S.C. § 2256 (8). However, upon closer inspection, we find that the database has been artificially populated. Technically speaking, no individual's data is "in" the synthetic database; the database consists of specifically perturbed data collectively matching the original. Likewise, Ashcroft v. Free Speech Coalition found that differences in production were dispositive—one, a byproduct of actual harm, and the other, the photographic embodiment of an idea. *See* Ashcroft v. Free Speech Coalition, 535 U.S. 234, 239-41 (2002) ("For instance, the literal terms of the statute embrace a Renaissance painting depicting a scene from classical mythology, a 'picture' that 'appears to be, of a minor engaging in sexually explicit conduct.' The statute also prohibits Hollywood movies, filmed without any child actors, if a jury believes an actor 'appears to be' a minor engaging in 'actual or simulated . . . sexual intercourse.'") (citing 18 U.S.C. § 2256(2)).

[173] To visualize this, see Figure 4 in Triastcyn and Faltings work, *supra* note 148. The image displays real versus fake pictures of numbers, illustrating how many numbers look similar, but are not.

[174] On the opposite end of the spectrum, if an adversary learns the first and last name of someone in the database is real, that obviously has more grave consequences.

[175] *See supra* section II.C.1 (describing leaky data).

[176] Carlini, *supra* note 154 at 10, Table 5.

synthetic dataset, they were only successful[177] three out of seven times—even when the likelihood that a secret was in the synthetic dataset (i.e., the likelihood of a leak) was over four thousand times more likely than a random word.[178]

For another perspective, look at it under the lens of the *Sanders* case, where professor Richard Sander litigated over access to the State Bar of California's bar admission database.[179] After remand by the California Supreme Court, the California Superior Court was tasked with assessing if any method of de-identification would sufficiently protect students' privacy (FERPA) while allowing Professor Sander to use the database effectively.[180] Out of four proposed methods relying heavily on deidentification, the court found only one sufficiently protected privacy, *k*-anonymity—though it destroyed utility.[181] The proposals included various levels of *k*-anonymity, removal of traditional identifiers like name and zip code, and the use of a secure data enclave.[182]

Importantly, professor Sweeny's input on the matter pegged the most-privacy-preserving method proposed as having a 31% chance of reidentification.[183] And to the court, this far exceeded the acceptable privacy risk.[184] The data was prohibited from disclosure.[185]

Though non-precedential, the conclusion is clear: anonymization either skirted utility (i.e., k-anonymity) or privacy (i.e., historical anonymization techniques like removal of zip code and name)—neither of which were worth the disclosure of sensitive bar data.

---

[177] To be sure, the researchers did not exhaustively search for a secret and merely ran extraction algorithms for one hour. *See id.* at 10. Additionally, the "secrets" researchers were searching for were specific phrases (e.g., a passcode or credit card number) and not data which is relatively benign as a standalone data point—like discovering a zip code which *may* be real, but which is not linked to any real information directly. *Id.* at 10 (using social security number and credit card number).

[178] *Id.* at 9-10. Training on the Enron corpus and hunting for secrets—looking for "real" social security numbers or credit card numbers—secrets were successfully extracted, in most cases, when the likelihood of a "secret" showing up was over four thousand times more likely than a random phrase.

[179] *See* Sanders v. State Bar of California, No. CPF-08-508880 (Nov. 7, 2016) (on file with author); *see generally* Latanya Sweeney, Michael von Lowenfeldt & Melissa Perry, *Saying it's Anonymous Doesn't Make it so: Re-Identifications of "Anonymized" Law School Data*, J. OF TECH. SCI. (Nov. 7, 2017), http://techscience.org/a/2017110702.

[180] *Sanders*, No. CPF-08-508880 at *19.

[181] *Id.* at *4.

[182] *Id.* at *17-21.

[183] *Id.*; *see also* Sweeney, Lowenfeldt & Perry, *supra* note 179 at 74-78. Stepping back, however, one of the reasons for this high risk of reidentification was that each proposal for sanitization revolved around starting with unique data and removing uniqueness bit by bit. The database's core was built on unique identifiers. Conversely, with synthetic data, the database's core would be built on fake, machine-generated data. Data leakage only concerns the possibility that one of the data points is real and is enough to tip off identification, presenting a much lower, theoretic risk to privacy.

[184] *See* Sweeney, Lowenfeldt & Perry, *supra* note 179 at 7 ("[T]he Court found that the percentage of unique records that exist after application of three of the four protocols is significantly higher than other acceptable norms. In particular, minority groups are more vulnerable to re-identification than their white counterparts. The Court also found considerable risk in "attribute disclosures," that is, inferences that can be drawn about applicants by virtue of their membership of a particular group."). To be sure, although HIPAA did not apply, the court used HIPAA as a benchmark in assessing risk. The petitioners in the case argued that HIPAA would accept a .22% risk of reidentification while the respondents argued for .02 to .04%. *Sanders*, No. CPF-08-508880 at *19.

[185] *Sanders*, No. CPF-08-508880 at *21-22.

Likewise, vanilla synthetic data makes no guarantee that a dataset is 100 percent free of *all* real identifiers. Getting away from that fact and urging a court to permit data release would be a difficult sale if, as in Sanders, sensitive facts are on the line. In this way, a stringent statute plus the uncertainty of chance identification—even if low—may prohibit any proposed release of a dataset into the wild.[186]

2. Under Inclusive Privacy

Insensitivity to chance identification is also possible. With statues like CCPA or VPPA, statutorily protected identifiers relate to a specific, unique piece or pieces of information. A database must lack these identifiers to be considered shareable, even if the pieces do not fall into the traditional category of name or zip code. And again, looking at evenly distributed training data, that there may be data leakage does not automatically mean an individual has been identified. In this sense, the generated data presents a "theoretical" rather than concrete chance of identification. And here, that chance may not be high enough for a court to bar disclosure. Consider the unifying themes in *Pruitt*, *In re Hulu*, *Eichenberger*, and *In re Nickelodeon*.

These cases, focusing on the CCPA and VPPA, permitted "anonymized" identifiers (i.e., user IDs, device serial numbers, or hexadecimal codes) to be publicly shared without violating consumers' PII. In distinguishing an earlier case, *Yershov*, where the court considered the combination of geolocation, device identification, and content viewed to be PII, the courts in *In re Hulu* and *Eichenberger* tossed aside the mere *theoretic* possibility of data linkage when sharing user IDs.[187] Building on *Pruitt*,[188] the courts held that "randomly"[189] generated user IDs were useless, and therefore not PII, without a master table connecting IDs to real identifiers like name and zip code. Likewise, *In re Nickelodeon* found that networking cookies and an IP address were not PII—pointing to the threshold

[186] To be sure, this is not to say the dataset should be released; only that under the right circumstances, the risk of identification may be overprotected and a less-stringent protection policy might better coax the wheels of research. Additionally, if the data concerned HIPAA, the provision regarding expert satisfaction could be used. 45 CFR § 164.514(b)(1). However, as seen in *Sanders*, this is not a sure bet, and may result in clashing expert opinions.

[187] Eichenberger v. ESPN, Inc., 876 F.3d 979, 985, 2017 U.S. App. LEXIS 24168, *13, 46 Media L. Rep. 1039, 2017 WL 5762817 (9th Cir. Wash. November 29, 2017). *But see* In re Vizio, Inc., 238 F. Supp. 3d 1204, 1224, 2017 U.S. Dist. LEXIS 60780, *48 (C.D. Cal. March 2, 2017) (finding the following assortment of collections to satisfy PII: "up to 100 billion content 'viewing data points' along with detailed information about a consumer's digital identity, such as consumers' IP addresses, zip codes, MAC addresses, product model numbers, hardware and software versions, chipset IDs, region and language settings, as well as similar information about other devices connected to the same network"

[188] *See* Pruitt v. Comcast Cable Holdings, LLC, 100 F. App'x 713, 716 (10th Cir. 2004) ("Without the information in the billing or management system one cannot connect the unit address with a specific customer; without the billing information, even Comcast would be unable to identify which individual household was associated with the raw data in the converter box.").

[189] Jason M. Rubin, *Can a Computer Generate a Truly Random Number?*, ASK AN ENGINEER (Nov. 1, 2011), https://engineering.mit.edu/engage/ask-an-engineer/can-a-computer-generate-a-truly-random-number/ ("'One thing that traditional computer systems aren't good at is coin flipping,' says Steve Ward, Professor of Computer Science and Engineering at MIT's Computer Science and Artificial Intelligence Laboratory. 'They're deterministic, which means that if you ask the same question you'll get the same answer every time.'").

foreshadowed by *Pruitt*: "There is certainly a point at which the linkage of information to identity becomes too uncertain, or too dependent on too much yet-to-be-done, or unforeseen detective work."[190] Perhaps it was best stated in *Eichenberger*:

> The manager of a video rental store in Los Angeles understood that if he or she disclosed the name and address of a customer—along with a list of the videos that the customer had viewed—the recipient of that information could identify the customer. By contrast, it was clear that, if the disclosure were that "a local high school teacher" had rented a particular movie, the manager would not have violated the statute. That was so even if one recipient of the information happened to be a resourceful private investigator who could, with great effort, figure out which of the hundreds of teachers had rented the video.[191]

While it was obvious to the courts that an address or geolocation may be PII, the introduction of randomness when paired with identifiers has not found favorable protection, even if some portion of potentially sensitive information like viewing history is tied to the "anonymous codes," as seen in *Pruitt*,[192] and even if relatively simply techniques could be used to track users across time with the released "non-identifiers," as seen in *In re Nickelodeon*.[193]

The problem with permitting the sharing of these datasets containing only *theoretic* risk of identification is legion. This line of reasoning is the same one abhorred by the cavalcade of academics criticizing historical means of anonymization.[194] Indeed, that the database "join" operation would not work on synthetic data does not mean the method is free of all threats.

As outlined in Section II.C, adversarial machine learning may be incredibly successful in uncovering secrets found in the training data, in some cases revealing up to 70% of the real, underlying records. Likewise, membership inference, another successful attack, would allow an attacker to glean sensitive information about the training data; specifically, whether the record attempting to be matched was used to train the model. Either way, synthetic data does not insulate privacy completely.

In summary, synthetic data's newness acts like a double-edged sword. On the one hand, the statutory lines drawn around privacy could result in over inclusive protection if a high-bar statute is applied. Rightfully, identification of individuals in a medical dataset should be avoided, but prohibiting data release in the face of a low chance of identification is not an ideal outcome. On the other hand, neither is it beneficial to overestimate synthetic data's protection. That reversal of generated data only exists in the "theoretic" sense does

---

[190] In re Nickelodeon Consumer Privacy Litig., 827 F.3d 262, 289, 2016 U.S. App. LEXIS 11700, *62 (3d Cir. N.J. June 27, 2016) (citing Yershov v. Gannett Satellite Info. Network, Inc., 820 F.3d 482, 486, 2016 U.S. App. LEXIS 7791, *9 (1st Cir. Mass. April 29, 2016)).

[191] *See Eichenberger*, 876 F.3d at 985.

[192] Although knowing User001's viewing history sounds benign, consider what would happen if User001 watched a particularly rare TV show in which 90% of its watchers come from one geographic location. Similar to the AOL search query reidentifications, anonymity in name alone may not be true anonymity if the user left bread crumbs in each of their recorded actions.

[193] *In re Nickelodeon Consumer Privacy Litig.*, 827 F.3d at 283 ("To an average person, an IP address or a digital code in a cookie file would likely be of little help in trying to identify an actual person.").

[194] It is like releasing stackable Lego blocks one at a time, but throwing caution to the wind because hard work alone could not possibly muster the energy to build a tower. A Lego block's very nature is aggregation. How can our "privacy" be dead while at the same time ensured because "theoretic" identification is not identification at all? And the same is true for vanilla synthetic data.

not mean the synthetic data is omnipotent. A more appropriate balancing of chance identification is needed; in other words, some manner of congressional response should take place. However, in the short term, a technical solution—differential privacy—should be pushed more heavily.

B. Differentially Private Synthetic Data

Ultimately, data leakage and the threat of techniques like adversarial machine learning result in the same dilemma identified in Section II.C: even with a reasonable distribution of input records there exists a *theoretic* possibility that original data may be leaked. Moreover, because privacy statutes do not speak to "fake" data, a door is left open for better or worse. The chance of identification may be inappropriately heightened or dampened depending on the statute at hand, the techniques used to train the model, and the ability to quantify the risk of identification. This uncertainty is problematic, and could lead to consequences paralleling the Netflix prize affair or the AOL search query debacle. Fortunately, a way forward has been identified, differential privacy.

Differential privacy's robust guarantees calm not only the fear of data leakage, but also the risks of adversarial machine learning. Although the technique is relatively new, even in the computer science literature (i.e., the optimal means of applying differential privacy to synthetic data is not yet settled[195]), differential privacy nonetheless provides a way of assuring privacy in the face of chance identification. Consider each of the examples discussed above.

When looking at HIPAA, the use of differentially private synthetic data would turn a "hit or miss" identification into a purely theoretical exercise, meaning the model resists even sophisticated attempts to reveal identities. In the *Sanders* case, synthetic data plus the differential privacy would likely give the court comfort in a "guarantee" of privacy post-release of the bar admission database. The court would have assurance that individuals could not be identified (i.e., a "secret" would not be any more likely "in" the database than any other datum) and professor Sander would have assurance that the data remains useful (i.e., research on these methods suggest only a 10% drop in accuracy).[196] And in the VPPA cases, even more intimate details could be shared with less risk. In *Yershov*, the court likely would have swayed toward permissible sharing if it knew that individuals had an incredibly low chance of identification.

In summary, using differential privacy in combination with synthetic data solves many of the problems owed to the limitations of the data generation process. However, we would be remiss if we did not make it absolutely clear that synthetic data and even differentially private synthetic data is not a silver bullet. Yes, differentially private synthetic data takes the chance of identification to a much safer level than vanilla synthetic data, but this does not mean it escapes all flaws entirely.

On the whole, no privacy preserving technique will completely solve the database-privacy problem. Indeed, if utility is of paramount concern, neither synthetic data or differential privacy or the combination of the two will resolve the conflict. Although synthetic data aids the database-privacy problem by using additive techniques rather that subtractive ones, and presents a statistically nearly-identical replica of the original data,

[195] *See supra* Section II.C.
[196] *See* Carlini, *supra* note 154 at 12.

this does not change the fact that the original data has been reshaped. The most ideal data to use in any analysis will always be original data. But when that option is not available, synthetic data plus differential privacy offers a great compromise.

IV. RECOMMENDATIONS

From the above analysis, three things are clear. First, synthetic datasets are better than traditional anonymization techniques (i.e., deidentification). Second, current laws are inadequate; while synthetic data may be permissible under some circumstances, the statutes do not cover the full benefits or risks of synthetic data.[197] And third, synthetic data—if constructed properly—may solve Professor Ohm's failure of anonymization.[198]

Firstly, the most important reason to use synthetic datasets instead of anonymized ones is that they avoid the arms race between deidentification and reidentification. True, secure anonymization via deidentification may be possible, albeit difficult;[199] however, the availability of secondary sources of information unknown to the sanitizer of the real data makes it a risky bet.[200] With synthetic datasets, we escape that trap entirely.

Second, most of today's privacy statutes are absolute: they bar disclosure of PII. While the actual metrics may be statistical—the HIPAA rules effectively use *k*-anonymity[201]—the goal is the same. No information may be disclosed about identifiable individuals. Synthetic datasets are different. They protect privacy through the addition of statistically similar information, rather than through the stripping away of unique identifiers. This, in turn, invites ambiguity: the resulting datasets leak information, and these leaks may or may not be enough to bar disclosure—resulting in over and under inclusive privacy coverage.

On the legal front, the solution to leakage is to face the ambiguity head on. New or amended statutes should accommodate synthetic data,[202] accepting the possibility of *measurably* small privacy leakage in exchange for perhaps mathematically provable protection against reidentification.[203] The exact amount of leakage is, of course, context-dependent; there is no reason that each U.S. sector-specific privacy statute should have the

---

[197] *See supra* Section II.A.

[198] *See* Ohm, *supra* note 5.

[199] *See* Dataverse, *HarvardX-MITx Person-Course Academic Year 2013 De-Identified dataset, version 2.0* (last accessed Sept. 8, 2017), https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/26147.

[200] This was the case in the AOL and Netflix data dumps: linking the original data with other sources was sufficient to reidentify some of the records. *See supra* note 66 and accompanying text.

[201] The purpose of stripping out 17 identifiers is to generalize the records, similar to how *k*-anonymity seeks to replace individuality with groupings. *See supra* Subsubsection II.B.3.i

[202] Notably, because of the many different definitions of privacy and PII, it is most likely that amendments must be made to key statutes. For example, it might be possible for HIPAA to explicitly embrace synthetic data under its expert determination statute, while it might require a statutory statement for statutes such as CCPA. *See* Cable Communications Policy Act of 1984, 47 U.S.C. § 551 (stating that "aggregate" data is not PII could be accompanied by a statement that differentially private synthetic data is not PII).

[203] Such tradeoffs have been used very successfully to obtain dramatic improvements in the performance of encrypted search algorithms. *See, e.g.*, Vasilis Pappas et al., *Blind Seer: A Scalable Private DBMS*, *in* IEEE SYMPOSIUM ON SECURITY & PRIVACY (May 2014).

same tradeoff.[204] Additionally, it is likely possible, though not yet available, to provide provable bounds on information leakage, with a stated error bound on queries. This opens up an interesting possibility: that the law provides a safe harbor for organizations that use this technique and are thus willing to incur less-than-perfect answers. There are still challenges, notably the possibility that software bugs may result in inadvertent leakage despite the guarantees of the algorithm; that said, it is worth exploring.

CONCLUSION

Databases play a key role in scientific progression. Yet, the very fuel of databases, open data, acts like a mire when paired with the meshwork of our privacy laws. Since the early days of reidentification a catch 22 has emerged: though we have the ability to gather and process incredibly enormous amounts of information, the legal morasses surrounding sharing this information acts prohibitively.

Synthetic data offers progress. Though not a silver bullet, the method allows us to put an end to the deidentification–reidentification arms race and focus on what matters, useful data. To this extent, we recommend that the privacy community accept synthetic data as a valid, next step to the database-privacy problem.

[204] There is a large debate in the privacy community about what constitutes "harm." To some, harm occurs only from direct financial loss or compromise of medical information; to others, disclosure of any private information is *a priori* harmful. *See, e.g.*, *Painting the Privacy Landscape: Informational Injury in FTC Privacy and Data Security Cases*, remarks by Acting Chair Maureen Ohlhausen, September 19, 2017, https://www.ftc.gov/system/files/documents/public_statements/1255113/privacy_speech_mkohlhausen.pdf.