Privacy and Synthetic Datasets

Steven M. Bellovin,[*] Preetam K. Dutta,[†] and Nathan Reitinger[‡]

22 STAN. TECH. L. REV. 1 (2019)

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 is either too greedy or too Laissez-Faire. Either all identifiers must be stripped from the data, rendering it useless, or one-step removed personally identifiable information may be shared freely, freely sharing secrets. In part, this is a result of the historic solution to database privacy, 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 the familiar ambiguity—privacy statutes either overweigh (i.e., inappropriately exclude data sharing) or downplay (i.e., inappropriately permit data sharing) the potential for synthetic data to leak secrets. We conclude by finding that synthetic data is a valid, privacy-conscious alternative to raw data, but not a

[*] Steven M. Bellovin is the Percy K. and Vida L.W. Hudson Professor of Computer Science at Columbia University, affiliate faculty at its law school, and a Visiting Scholar at the Center for Law and Information Policy at Fordham University School of Law.

[†] Preetam Dutta is a doctoral student at the Department of Computer Science at Columbia University.

[‡] Nathan Reitinger is an attorney and a master's student at the Department of Computer Science at Columbia University.

cure-all. In the end, computer science progress must be met with sound policy in order to move the area of useful data dissemination forward.

TABLE OF CONTENTS

INTRODUCTION

Synthetic data is a viable, next-step solution to the database-privacy problem: You[1] are in a database;[2] sharing your secrets and allowing data

1. Moreover, this database possesses the quantifiable facts about you, which may be more than you suspect. For example, 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. 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:

[T]he rapid pace of computer development and usage throughout American

scientists to analyze each and every aspect of your life has the potential to unlock incredible breakthroughs across a vast number of disciplines;[3] but keeping your secrets private—while at the same time maintaining the usefulness of the data—is a nontrivial problem.[4] 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,[5] stripping away *all* identifiers that could

---

> 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.

ALAN F. WESTIN, PRIVACY AND FREEDOM 366-67 (1967). *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.").

3. *See* RAMEZ ELMASRI & SHAMKANT B. NAVATHE, FUNDAMENTALS OF DATABASE SYSTEMS 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 3 (3d 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 HARV. J. L. & TECH. 1, 2 (2011) ("[I]n 2001, John J. Donohue and Steven D. Levitt presented shocking evidence that the decline in crime rates during the 1990s, which had defied explanation for many years, was caused in large measure by the introduction of legalized abortion a generation earlier. [This] stud[y] and many others have made invaluable contributions to public discourse and policy debates, and . . . would not have been possible without anonymized research data . . . .").

4. *See* Matthew Fredrikson et al., *Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing*, 2014 PROC. USENIX SECURITY SYMP. 17, 27 (showing that utility and privacy cannot be both achieved in the context of personalized warfarin dosing—even with the application of differential privacy).

5. Anonymization, as used in this sentence, refers to the colloquial understanding of the term—which is more accurately defined as deidentification. Briefly, it means removing names and other obviously identifying information, and perhaps replacing them with random values. *See infra* Part I.B (discussing the history of anonymization, starting with de-identification and moving to synthetic data).

potentially uniquely identify an individual or group of individuals.[6] Anonymization, however, proved to be anything but a "silver bullet."[7] From the AOL search-query debacle to the Netflix Prize affair, it seemed trivial with even novice computer aptitude to "join"[8] auxiliary information with a series of "perturbed"[9] data points and unveil the very data that anonymization was designed to protect.[10]

The well-documented failures of anonymization have prompted aggressive research on data sanitization, ranging from *k*-anonymity[11] in the late 1990s to today's highly acclaimed privacy mechanism, differential privacy.[12] But the basic tradeoff between utility and privacy—an inverse relationship—still remains.

---

6. Paul Ohm describes how this is possible:

> 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 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.

Paul Ohm, *Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization*, 57 UCLA L. REV. 1701, 1703 (2010) (emphasis in original).

7. *Id.* at 1736; *see also* Arvind Narayanan & Edward W. Felton, *No Silver Bullet: De-Identification Still Doesn't Work* (July 9, 2014), https://perma.cc/X6PZ-X9EP (archived Nov. 14, 2018).

8. In a general sense, the database "join" operation merges two tables on a common condition. For example, two records that have the same value for "social security number" can be merged. *See* ELMASRI & NAVATHE, *supra* note 3, at 107-09.

9. "Perturbed" here refers to the traditional, remove-name-and-zip-code styled sanitization techniques which often fail to exclude information which may be linked together to reidentify individuals. *See infra* Part I.B.1.

10. *See* Arvind Narayanan & Vitaly Shmatikov, *Robust De-Anonymization of Large Sparse Datasets*, 2008 IEEE SYMP. SECURITY & PRIVACY 111, 111-112 ("We demonstrate that an adversary who knows only a little bit about an individual subscriber can easily identify this subscriber's record in the dataset. Using the Internet Movie Database as the source of background knowledge, we successfully identified the Netflix records of known users, uncovering their apparent political preferences and other potentially sensitive information."); 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 future where any single piece of true information can allow someone to infer all other true information).

11. *See generally* Pierangela Samarati & Latanya Sweeney, *Protecting Privacy when Disclosing Information:* k*-Anonymity and Its Enforcement Through Generalization and Suppression* (1998), https://perma.cc/FF9E-FTGF (archived Nov. 7, 2018); *see also* Subpart I.B.3.i.

12. *See* Cynthia Dwork, *Differential Privacy*, 33 INT'L COLLOQUIUM AUTOMATA, LANGUAGES & PROGRAMMING 1, 1-2 (2006); *see also* Tore Dalenius, *Towards a Methodology*

The aim of this Article is to present a new, better alternative to sanitized data release, "synthetic" data.[13] In essence, take an original (and thus sensitive) dataset, use it to train[14] a machine-learning enabled generative model,[15] and then use that model to produce realistic, yet artificial data that nevertheless has the same statistical properties. Consider, for example, a database of salaries. A differentially private version of the database would

for Statistical Disclosure Control, 15 STATISTIK TIDSKRIFT 429 (1977); Michael Hilton, *Differential Privacy: A Historical Survey*, https://perma.cc/J3HT-DMWB (archived Nov. 7, 2018); *see also* Subpart I.B.3.ii.

13. *See infra* Part II.

14. 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 may be based on many factors: the category of movie (e.g., drama or documentary), the language, the director and actors, or the production company. *Id.*

15. 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://perma.cc/YEH8-H3K7 (archived Nov. 7, 2018). For a more rigorous description, see Andrew Ng, *CS229 Lecture Notes* 1, https://perma.cc/FSK6-73YZ (archived Nov. 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 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 $p(y \mid x)$ directly (such as logistic regression), or algorithms that try to learn mappings directly from the space of inputs χ to the labels $\{0, 1\}$, (such as the perceptron algorithm) are called discriminative learning algorithms. [A]lgorithms that instead try to model $p(x \mid y)$ (and $p(y)$) . . . [t]hese algorithms are called generative learning algorithms. For instance, if y indicates whether a[n] example is a dog (0) or an elephant (1), then $p(x \mid y = 0)$ models the distribution of dogs' features, and $p(x \mid y = 1)$ models the distribution of elephants' features.").

have, e.g., the same average salary but the individual entries would be different[16] than the underlying, real data.[17] The end result may be compared to counterfeit money. Although the appropriately-sized 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).[18] 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.[19]

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 an 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.,

16. The statistical properties of data are a function of the type of data, whether it is an image or a piece of text. In the general context, statistical properties of the training-set data are descriptive features that are kept close to the ground truth in the synthetic data. *See, e.g.*, William Li et al., *Using Algorithmic Attribution Techniques to Determine Authorship in Unsigned Judicial Opinions*, 16 STAN. TECH. L. REV. 503, 509-11, 525 (2013) (using the statistical properties of text—i.e., Justice Alito often uses the words "set out in" and "the decision of" while Justice Roberts often uses "the first place" and "without regard to"—to unveil authorship).

17. 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.

18. U.S. CURRENCY EDUC. PROGRAM, QUICK REFERENCE GUIDE (2017).

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

data generation without additional sanitization techniques) and differentially private synthetic data. Finally, we offer technical and legal recommendations for the legal community.

In short, although no solution to the database-privacy problem is a "silver bullet,"[20] 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, the method is relatively new and its meshing with legal statutes is both over- and under-inclusive: On the one hand, statutes thinking of "identification" in binary terms may accept the wholesale value of synthetic data, even though unique-enough data may nonetheless "leak" information; on the other, statutes that consider identification broadly may prohibit synthetic data, even though risk of a leak, practically, is minimal.[21] 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 explicit 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 is the right to be left alone, as it was originally contemplated by Samuel Warren and Louis Brandeis, and later by William Prosser.[22] From there, however, the concept has experienced its fair share of refactoring.[23]

---

20. *See* Narayanan & Felton, *supra* note 7, at 8 ("If a 'best of both worlds' solution exists, de-identification is certainly not that solution.").

21. *See infra* Part III.B.

22. *See* Samuel D. Warren & Louis D. Brandeis, *The Right to Privacy*, 4 HARV. L. REV. 193, 195 (1890) (taking issue with how a new innovation of the time—yellow journalism—permitted "what is whispered in the closet [to] be proclaimed from the housetops") (quotation omitted); 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)).

23. *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 ;. 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

A. Privacy: A Database Perspective

Congress's response to these watershed pieces of legal scholarship, along with the influential 1973 study,[24] was to enact a lattice of statutes targeting areas of highly sensitive data.[25] Though not the exclusive avenue for privacy protection, these statutes form a meshwork that, though protective of privacy, have impeded data sharing. Protected sectors range from health (HIPAA) to finance (FCRA), and often hinge the statutory shield on the definition of "personally identifiable information" (PII).[26] Put simply, if a fact (i.e., a datum[27] in the database) contains PII, then it is protected and cannot

---

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))).

24. Warren & Brandeis, *supra* note 22; U.S. DEP'T HEALTH, EDUC. & WELFARE, RECORDS COMPUTERS AND THE RIGHTS OF CITIZENS (1973), https://perma.cc/NZN7-P4R7.

25. Notably, this approach differs from the one adopted by the United Kingdom, which has been called "expansionist" and protects data 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://perma.cc/6XMH-LSGP.

26. *See, e.g.*, Fair Credit Reporting Act of 1970, 15 U.S.C. §§ 1681; 1681a(c)–(d) (2017); Privacy Act of 1974, 5 U.S.C. § 552a(a)(2) (2017); FERPA, 42 U.S.C. § 1320g(a)(5)(a) (2017); Health Insurance Portability and Accountability Act (HIPAA) of 1996, 42 U.S.C. § 1320 (2017); Driver's Privacy Protection Act of 1994, 18 U.S.C. § 2721(a) (2017); Right to Financial Privacy Act of 1978, 12 U.S.C. §§ 3401–3422 (2017); Foreign Intelligence Surveillance Act of 1978, 15 U.S.C. §§ 1801–1811 (2017); Privacy Protection Act of 1980, 42 U.S.C. § 2000aa (2017); Cable Communications Policy Act of 1984, 47 U.S.C. § 551 (2017); Electronic Communications Privacy Act of 1986, 18 U.S.C. §§ 2510–2522 (2017); Computer Matching and Privacy Protection Act of 1988, 5 U.S.C. § 552a (2017); Telephone Consumer Protection Act of 1991, 47 U.S.C. § 227 (2017); Identity and Assumption Deterrence Act of 1998, 18 U.S.C. § 1028 (2017); Gramm-Leach-Bliley Act of 1999, 15 U.S.C. §§ 6801–6809 (2017); Uniting and Strengthening America by Providing Appropriate Tools Required to Intercept and Obstruct Terrorism Act of 2001 (USA Patriot Act), 107 Pub. L. No. 56, 115 Stat. 272 (2001) (codified as amended in scattered sections of the U.S. Code); CAN-SPAM Act of 2003; Video Voyeurism Prevention Act of 2004, 18 U.S.C § 1801 (2017); Video Privacy Protection Act of 1988, 18 U.S.C. § 2710 (2017). *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).

27. *See* MICHAEL J. HERNANDEZ, DATABASE DESIGN FOR MERE MORTALS: A HANDS-ON GUIDE TO RELATIONAL DATABASE DESIGN 43 (3d ed. 2013) ("The values you store in the database

be shared; if the fact does not contain PII, then it is not protected and may be shared freely.[28] The problem comes from delineating PII from non-PII.

In fact, because of the statutory mosaic in which PII has been iteratively defined, the term is nearly impossible to understand. 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.[29] 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[30] in any way, perfect utility is achieved.[31]

---

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.") (emphasis in original).

28. Under this framework, the question of privacy changes 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," although 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 re-identification, and the trustworthiness of the data aggregator. *See* Ohm, *supra* note 6, at 1727, 1765-68.

29. 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 pattern: 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, and therefore whether sharing is impermissible. The same is true for synthetic data, except the sharing would be done with synthetic data rather than original data.

30. Put simply, perturbation means modification—one of the simplest techniques being the addition of random values alongside original values. *See infra* Part I.B.1; *see also* Hillol Kargupta et al., *On the Privacy Preserving Properties of Random Data Perturbation Techniques*, 3 PROC. IEEE INT'L CONF. DATA MINING 99, 99 (2003).

31. *See* Ohm, *supra* note 6, 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* THEORY CRYPTOGRAPHY CONF. 363 (2005))).

B. Databases

A database is simply a collection of data. Be it physical[32] or digital, the "database" is more technically defined as the "organized collection of factual relations."[33]

It is likewise important to note that databases are not inherently threatening to privacy.[34] Indeed, the database is not a new concept borne from the Internet or computer. Before the proliferation of computerized information, data describing individuals manifested itself in physical locations.[35] And these physical locations were, for the most part, geographically disparate. To concatenate database 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).[36] With the depressed compilation of data, privacy rights were more easily protected using traditional, redacted-name-and-zip-code methods.

---

32. *See, e.g.*, *The Technium: One Dead Media*, KK (June 17, 2008), https://perma.cc/4MF2-VD6L ("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[, these cards] 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.").

33. *See* HERNANDEZ, *supra* note 27, 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 3, 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.") (emphasis in original).

34. *But see* SIMSON GARFINKEL, DATABASE NATION: THE DEATH OF PRIVACY IN THE 21ST CENTURY 5 (2001).

35. *See* SOLOVE, *supra* note 23, at 13 (noting how records were mostly kept by hand in various offices); Carolym Puckett, *The Story of the Social Security Number*, 69 SOC. SEC. BULLETIN 55, 56 (2009) (noting that there were 12 regional offices); *Comput. Privacy: Hearings Before the Subcomm. on Admin. Practice and Procedure of the S. Comm. on the Judiciary*, 19th Cong. 74 (1967) (statement of Arthur R. Miller, professor of law, University of Michigan) ("Privacy has been relatively easy to protect in the past for a number of reasons: (1) large quantities of information about individuals have not been available; (2) the available information generally has been decentralized and has remained uncollected and uncollated . . . .").

36. SOLOVE, *supra* note 23,at 14 ("Technology was a primary factor in the rise of information collection. The 1880 census [an early attempt at mass information collection] required almost 1,500 clerks to tally information tediously by hand—and it took seven

But all this changed once seemingly unlimited columns,[37] cheap storage, and centralized access became more ubiquitous.[38] As our society merges itself with the digital world, information is more easily amassed.[39] Not only that, but linking different kinds of databases is also practical—unlocking the potential for *en masse* learning.[40] From the social sciences to medicine to

years to complete.").

37. In database terms, this is known as a field (i.e., the vertical groupings in a Microsoft Excel document). *See* HERNANDEZ, *supra* note 27, 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.'") (emphasis in original).

38. *See* SOLOVE, *supra* note 23, at 14 ("As processing speeds accelerated and as memory ballooned, computers provided a vastly increased ability to collect, search, analyze, and transfer records.").

39. *See id.* at 15 ("Today, federal agencies and departments maintain almost 2,000 databases, 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.") (internal citations omitted).

40. *See* Tim Berners-Lee, *The Next Web*, TED 10:45-15:00 (Feb. 2009), https://perma.cc/4J2Y-D9YE (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://perma.cc/99JW-7Z65 (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).

modern day business operations, storing, analyzing, and reproducing information has become routine.[41] Some have even referred to this type of interaction (colloquially termed "big data") as the 21st century's microscope.[42] But our brave, big data world is not without drawbacks.[43]

Because the information collected concerns more and more private minutiae,[44] the valuable byproducts of the knowledge come at an increasing cost to privacy. 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 scandalous assumption?[45]—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 twenty-five products that are indicative

41. *See* ELMASRI & NAVATHE, *supra* note 3, 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.").

42. VIKTOR MAYER-SCHONBERGER & KENNETH CUKIER, BIG DATA: A REVOLUTION THAT WILL TRANSFORM HOW WE LIVE, WORK, AND THINK 18 (2014) ("Big data marks an important step in humankind's quest to quantify and understand the world. A preponderance of things that could never be measured, stored, analyzed, and shared before is becoming datafied. Harnessing vast quantities of data rather than a small portion, and privileging more data of less exactitude, opens the door to new ways of understanding. It leads society to abandon its time-honored preference for causality, and in many instances tap the benefits of correlation.").

43. *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 is 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, etc.), 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 (*e.g.* for example, free email accounts, social networks, etc.) 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).").

44. For example, consider biometric data. *See, e.g.*, JENNIFER LYNCH, EFF & IMMIGRATION POLICY CENTER, FROM FINGERPRINTS TO DNA: BIOMETRIC DATA COLLECTION IN U.S. IMMIGRANT COMMUNITIES & BEYOND 4 (2012).

45. Charles Duhigg, *How Companies Learn Your Secrets*, N.Y. TIMES, Feb. 16, 2012, at MM30 ("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?'").

of pregnancy. Indeed, the panoply of what Target knew was creepily extensive.[46] But Target, along with its usual compatriots like Facebook, Amazon, Netflix, and Alphabet,[47] had been collecting this kind of data for years, yielding increasingly intimate details of our lives as the technology improved.[48] And with these intimate details came humbling returns—during the years when Target implemented its targeted baby advertisements, the company's revenues increased from $44 billion to $67 billion.[49] But the real question for these companies is not how to monetize big data insights, but how to do it without the understandably negative optics.[50] Stated simply, how can this data be usefully applied *without* stepping on anyone's privacy toes? Historically, the answer has been anonymization.

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

Early on—and still making the rounds today[51]—the assumption was that if you stripped out enough identifying information from a dataset, the

---

46. *See* Nick Saint, *Eric Schmidt: Google's Policy Is to 'Get Right up to the Creepy Line and Not Cross It,"* BUS. INSIDER (Oct. 1, 2010, 2:44 PM), https://perma.cc/HCX7-SSJP; *see also* Duhigg, *supra* note 45 ("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.").

47. Many companies collect large amounts of data on individuals; some market this data to other companies. *See* FED. TRADE COMM'N, DATA BROKERS: A CALL FOR TRANSPARENCY AND ACCOUNTABILITY 11-18, 23-31 (2014).

48. SCHNEIER, *supra* note 2, at 4 ("The bargain you make, again and again, with various companies is surveillance in exchange for free service.").

49. *See* Duhigg, *supra* note 45.

50. *Id.* As an executive for Target pronounced: "'With the pregnancy products . . . we learned that some women react badly . . . . [T]hen we started mixing in all these ads for things we knew pregnant women would never buy, so the baby ads looked random. We'd put an ad for a lawn mower next to diapers. We'd put a coupon for wineglasses next to infant clothes. That way, it looked like all the products were chosen by chance.'" *Id.*

51. For example, FERPA explicitly allows data release if the information is deidentified, meaning all personally identifiable information has been removed; the Gramm-Leach-Bliley Act has been interpreted by the FTC to *not* protect "aggregate information or blind data [not containing] personal identifiers such as account numbers, names, or

data could be shared freely.[52] Though the approach is colloquially referred to as anonymization,[53] it is more accurately described as deidentification: sterilization via subtraction.[54] Under HIPPA's "safe harbor" provision,[55] medical data is freely shareable if all seventeen identifiers have been removed. Detailed and explicitly defined, HIPAA assumes that information

addresses"; and both the Cable Act and VPPA's definition of PII have been interpreted by the courts to *not* cover anonymized identifiers. *See* FERPA, 20 U.S.C. § 1232g(a)(5)(a) (2017) (listing facts constituting PII—and conversely, which facts are not PII and are therefore free to share); Gramm-Leach-Bliley Act of 1999, 15 U.S.C. §§ 6801–09 (2017); 16 CFR § 313.3(o)(2)(ii) (2018). In *Pruitt*, the court ruled that hexadecimal codes identifying customers and their purchases were not PII because the digits were not addresses or names. *See* Pruitt v. Comcast Cable Holdings, 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."). And in *In re Hulu*, the court found that "a unique anonymized ID alone is not PII." *In re* Hulu Privacy Litig., No. C 11-03764 LB, 2014 WL 1724344, at *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") (emphasis in original); *see also* 34 C.F.R. § 99.31(b) (2018) (discussing how the deidentification must reasonably ensure that a student's identity is not "personally identifiable"); 16 C.F.R. § 313.3(o)(2)(ii)(B); Benjamin Charkow, Note, *The Control Over the De-Identification of Data*, 21 CARDOZO ARTS & ENT. L.J. 195, 196-97 (2003) (noting "[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") (internal citation omitted).

52. *See* Ohm, *supra* note 6, at 1707-11 (recounting the staunch supporters of de-identification—spanning from industry to academia to administration).

53. Meaning "without a name or nameless" from the Greek *άνωνυμία*. *See* Zoltán Alexin, *Does Fair Anonymization Exist?*, 28 INT'L REV. 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 re-identification to consider the resulting dataset private).

54. *See* Ira S. Rubinstein & Woodrow Hartzog, *Anonymization and Risk*, 91 WASH. L. REV. 703, 710 (2016) (defining de-identification 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), https://perma.cc/N9TW-K86K)). 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* 14 PROC. INT'L CONF. KNOWLEDGE DISCOVERY & DATA MINING 70, 70, 78 (2008) (considering "trivial sanitization" to be the removal of all quasi-identifiers or sensitive attributes)).

55. *See generally* Health Insurance Portability and Accountability Act (HIPAA) of 1996, 42 U.S.C. § 1320 (2017); 45 CFR § 164.514. The identifiers are 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.

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?[56]

In actuality, identifying individuals using seemingly non-unique identifiers is far easier than a data sanitizer might hope.[57]

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,"[58] it is difficult to correctly guess which 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.[59] The result is an inescapable tradeoff: more representative data versus more privacy.

Full scale DNA sequencing—anonymized via de-identification—was publicly released in the 1990s as part of the Human Genome Project.[60] However, in 2004, researchers demonstrated that it was possible to link an individual's de-identified genomic data with publicly available single nucleotide

---

56. HIPAA does have a provision for "any other unique identifier not listed." 45 CFR § 164.514(b)(2). However, as Part I.B.2 demonstrates, even trivial data points may be linked with an identity. *See infra* Subsection I.B.2. Before the Netflix Prize reidentifications, it would have defied logic to consider liking non-blockbuster movies to be a unique identifier. *See* Narayanan & Shmatikov, *supra* note 10, at 122 fig.9 (2008) (showing how liking less popular movies makes deidentification more likely).

57. *See, e.g.*, Brian Hayes, *Uniquely Me!*, AM. SCI., https://perma.cc/UD6A-XM9C (archived Oct. 26, 2018).

58. 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 is necessary to achieve the desired anonymity. *See, e.g.*, Nathan Reitinger, Comment, *CAD's Parallel to Technical Drawings: Copyright in the Fabricated World*, 97 J. PAT. TRADEMARK OFF. SOC'Y 111, 113-14 (2015) (explaining how subtractive manufacturing starts with a large block of material and gradually whittles it away to form a desired object, while additive manufacturing gradually builds an object from the ground up); Samuel H. Huang et al., *Additive Manufacturing and Its Societal Impact: A Literature Review*, 67 J. ADVANCED MANUFACTURING TECH. 1191, 1191 (2013).

59. *See* Brian Parkinson et al., *The Digitally Extended Self: A Lexicological Analysis of Personal Data*, 44 J. INFO. SCI. 552, 552-53 (2017) (noting "the classification of data based on degrees of identifiability may fluctuate and become indeterminate.").

60. This was the result of policies adopted by several organizations, including the National Human Genome Research Institute, the Department of Energy, and the International Human Genome Sequencing Consortium. With a focus on open records, these policies generally recommended depositing data and resources into the public domain. *See Reaffirmation and Extension of NHGRI Rapid Data Release Policies: Large-Scale Sequencing and Other Community Resource Projects*, NAT'L HUMAN GENOME RESEARCH INST. (Feb. 2003), https://perma.cc/HT2Q-VDT3.

polymorphism data.[61] NIH reacted to the privacy woes by restricting access to individual-level genomic data on a permission-only basis.[62] But then, in 2008, researchers again showed that individuals could be identified in trace-amount, high-density genetic mixtures.[63] NIH clamped down harder on the weak link, restricting any access to genome-wide association studies.[64] Most recently, in 2013, researchers demonstrated yet again that it was possible to match NIH's snippet tandem repeats with consumer-focused, publicly available, genetic genealogy information, which permitted an individual's surname to be identified.[65] In response, NIH held its tune and moved age information from a public to non-public database.[66]

---

61. Zhen Lin et al., *Genomic Research and Human Subject Privacy*, 305 SCIENCE 183, 183 (2004) ("If 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 individual. In such a case, the rest of the genotypic, phenotypic, and other information linked to that individual in public records would be available.").

62. NIH required approval from a Data Access Committee to gain access to individual genomic data in the Database of Genotypes and Phenotypes. *See* Stacey Pereira et al., *Open Access Data Sharing in Genomic Research*, 5 GENES 739, 740 (2014).

63. 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, 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 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.").

64. *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), https://perma.cc/C8YU-S496 ("The US National Institutes of Health (NIH), the Broad Institute in Cambridge, Massachusetts, and the Wellcome Trust in London all decided to restrict access to data from genome-wide association (GWA) studies—which contain collections of thousands of people's DNA—after research suggested that it is possible to identify an individual from their genetic fingerprint even when their DNA is mixed together with that of many other people.").

65. Melissa Gymrek et al., *Identifying Personal Genomes by Surname Inference*, 339 SCIENCE 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.").

66. *See* Pereira et al., *supra* note 62, at 740 ("NIH worked with the [National Institute of General Medical Sciences] to move age information, which was previously publicly accessible, into the controlled-access part of the database."). *But see* Khaled El Emam at al., *A Systematic Review of Re-Identification Attacks on Health Data*, 6 PLOS ONE 1, 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.").

This vignette demonstrates not only that de-identification requires precise definitions of "unique identifiers," but also that de-identification suffers from an aging problem. When genomic data was originally released, consumer-opt-in genetic studies were not popular. To be sure, 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 de-identification–re-identification arms race.

3. The Confusing: Post-Anonymization-Failure Awareness

Realizing that anonymization promises much more than it manages to deliver, a number of alternative approaches have been suggested. Setting aside the purely legal solutions (often focusing on reframing PII),[67] one of the most popular paths in this terrain is to use computer science metrics to replace the historic means of deidentification.

The linchpin in each of these methods is to start with creating a metric for defining privacy because, as with all connotative definitions, using non-technical, non-mathematical descriptions invites ambiguity.[68] 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.

67. Various methods for addressing PII's failures have been proposed. *See* Schwartz & Solove, *supra* note 29, at 1894 (arguing for a division between "identified" and "identifiable" information and applying protection mechanisms based on the risk each category engenders); Rubinstein & Hartzog, *supra* note 54, at 706 ("[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 et al., *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.").

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

i. k-Anonymity

k-anonymity maintains privacy by *guaranteeing* that for every record in a database there are some number "*k*" of indistinguishable copies.[69] 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.[70] The suppression method follows a denotative definition and replaces a pivotal piece of identifying information in the original database with a meaningless placeholder.[71] 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 opposed to a single value (i.e., "Class Year" between one and three).[72]

---

69. *See* Samarati & Sweeney, *supra* note 11, 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.").

70. Roberto J. Bayardo & Rakesh Agrawal, *Data Privacy Through Optimal* k*-Anonymization*, 21 PROC. INT'L CONF. DATA ENGINEERING 217, 217 (2005) ("Suppression is the process of deleting cell values or entire tuples . . . . Generalization involves replacing specific values [such as a phone number] with more general ones [such as the area code alone.]").

71. *Id.*

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

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 illustrates the tradeoff between utility and privacy—the table is less useful now because each individual row is less unique. There has been a loss of utility for the gain of privacy.

ii. Differential Privacy

Differential privacy is a popular and robust method[73] that rose to prominence following the famous shortcomings demonstrated in the Netflix Prize affair.[74] While there are many forms of the general technique, its primary

73. 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://perma.cc/GE7G-EV6V (archived Oct. 26, 2018) ("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.") (emphasis in original).

74. *See generally* Narayanan & Shmatikov, *supra* note 10, at 118-23; Ohm, *supra* note 6, 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. 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,

goal is to maximize the accuracy of queries from a database while limiting or minimizing the potential for privacy leakage.[75] 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 modified from Professor Dwork and Roth's recent work.[76]

Imagine a scenario in which someone asks you the question: "Do you like ice cream?"[77] This question has a binary, yes or no, answer. However, it could be modified with the aid of a coin toss.[78] Prior to answering, a coin is tossed, and if a head is the result, you answer the question with the truth. Otherwise, you will give a "random" answer (which in this case is another coin toss with a predefined "yes" if heads and "no" if not).[79]

While it is possible to deduce the probability of people who like ice cream, the individuals answering this question now have "deniability."[80] In other words, 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 essence of differential privacy: Because of the introduction of randomness (i.e., a person's veracity depends on a coin toss) which produces deniability, you may now say "I may or may not be 'in' the database."[81]

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 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.'") (internal citations omitted).

75. Although not the first to introduce differential privacy, Cynthia Dwork's survey is commonly cited. *See* Cynthia Dwork, *Differential Privacy: A Survey of Results*, 5TH INT'L CONF. THEORY & APPLICATIONS OF MODELS OF COMPUTATION 1, 2-3 (2008).

76. Cynthia Dwork & Aaron Roth, *The Algorithmic Foundations of Differential Privacy*, 9 FOUND. & TRENDS THEORETICAL COMP. SCI. 211 (2013).

77. *Id.* at 238-39.

78. *Id.*

79. *Id.*

80. *Id.* at 225-26 (discussing how "'[p]rivacy' comes from the plausible deniability of any outcome").

81. Those responsible for the database's secrets may now 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." Cynthia Dwork, *The Promise of Differential Privacy: A Tutorial on Algorithmic Techniques*, 52 IEEE FOUND. COMP. SCI. 1, 1 (2011); *see also* Cynthia Dwork, *A Firm Foundation for Private*

Differential privacy has many strengths, but as with all methods, it is not a panacea.[82] For example, if enough identical queries are asked, the power of deniability is diluted.[83] Eventually, repeat queries may be able to take all answers together and disambiguate falsity from truth.[84] Additionally, if the query being asked requires high specificity, then it is more difficult to permit deniability.[85] For example, if a query asks for the minimum GPA in a group of students, it will be hard to tell a lie while also providing 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 the utmost importance.[86] One study focusing on Warfarin dosing found that privacy was only sufficiently protected if differential privacy was used, but that differential privacy destroyed utility.[87] "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."[88] Stated another way, if absolute utility is needed, even *de minimis* sanitization has an adverse effect.

Data Analysis, 54 COMM. ACM 86, 91 (2011) ("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.").

82. For one critique of the metric, see Bambauer et al., *supra* note 67, at 754 ("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://perma.cc/2U98-D798 (critiquing Fool's Gold for misreading differential privacy's value).

83. *See generally* Bill Howe, *Weakness of Differential Privacy*, COURSERA (last accessed Aug. 1, 2018), https://perma.cc/2GPN-FWAB (finding that differential privacy is best suited for low-sensitivity areas and has problems with repeat queries).

84. *See* Matthew Green, *What Is Differential Privacy*, CRYPTOGRAPHY ENGINEERING (June 15, 2016), https://perma.cc/B37U-GP43 ("But there's a big caveat here. Namely, while the amount of 'information leakage' from a single query can be bounded by a small value, this value is not zero. Each time you query the database on some function, the total 'leakage' increases—and can never go down. Over time, as you make more queries, this leakage can start to add up.").

85. *Id.*

86. Matthew Fredrikson et al., *supra* note 4, at 19 (finding utility and privacy mutually exclusive in regard to warfarin dosing studies); *id.* at 29 ("[F]or ε values that protect genomic privacy, which is the central privacy concern in our application, the risk of negative patient outcomes increases beyond acceptable levels.").

87. 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) was too high, then utility was gained but privacy was lost; however, if ε was too low, then privacy was gained but utility was lost, resulting in adverse patient outcomes. *See id.* at 26-27, 29.

88. *Id.* at 29.

In summary, 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 a heavily redacted document which perfectly maintains privacy by liberally striking all remotely identifying phrases (i.e., deidentification). Unfortunately, the document is also 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, without allowing you to know whether the FBI answers truthfully). But not in all cases. We are still in a negative-sum game, less data for more privacy, and the threat remains 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, legally, real data cannot be shared, but, practically, deidentified data lacks sufficient utility. In those moments, having a synthetic dataset may present a best-of-both-worlds solution—shareable, yet similar-to-original data.

Before illustrating the veracity and limitations of synthetic data using a case study, we will first outline the core concepts underlying synthetic data. Here, we start with an oft-cited yet poorly-understood term: machine learning.

A. In Brief: Machine Learning

When Ada Lovelace sat down to ponder one of the world's first computer programs, the automatic calculation of Bernoulli numbers, she did so in painstaking detail.[89] She considered each and every step in a laborious, mathematical fashion.[90] As anyone with experience in programming knows,

89. *See generally* Stephen Wolfram, *Untangling the Tale of Ada Lovelace*, WIRED (Dec. 22, 2015), https://perma.cc/H6U3-K9HW.

90. 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

this is exactly what the art of programming requires.[91] 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?[92] This is exactly what machine learning does, often relying on a neural network[93] at its core.[94]

While neural networks are not new, the concept has garnered tremendous attention recently given the multitude of problems that are now tractable as a result of improvements in computer hardware and cheaper prices for that hardware.[95] In the past, a necessarily large neural network (i.e., large enough to produce worthwhile results) could not be trained without rooms of computers and rows of graphics cards. Today, thanks in part to the gaming culture's endless hunger for higher-performance graphics cards, the computations may be done at home. Because the neural network is a central component of machine learning, which is used in the creation of synthetic data, our understanding of synthetic data will start there.

(Dec. 28, 2017), https://perma.cc/Z47B-B7FP.

91. *See* MARKO PETKOVŠEK ET AL., A=B, at 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* 1 DONALD E. KNUTH, THE ART OF COMPUTER PROGRAMMING (2d. ed. 1973).

92. *See* Jeremy Howard, *The Wonderful and Terrifying Implications of Computers that Can Learn*, TED (Dec. 2014), https://perma.cc/J42E-GBYC (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).

93. Neural networks, as well as other techniques such as regression, may be considered a subset of machine learning, which is itself a subset of artificial intelligence.

94. Research in the neural network domain dates back decades, at least as far back as Walter Pitts's 1942 article on the subject. *See* Walter Pitts, *Some Observations on the Simple Neuron Circuit*, 4 BULLETIN MATHEMATICAL BIOPHYSICS 121, 121 (1942) (explaining "[a] new point of view in the theory of neuron networks").

95. Ophir Tanz describes this history:

> "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 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)."

Ophir Tanz, *How Video Game Tech Makes Neural Networks Possible*, TECHCRUNCH (Oct. 27, 2017), https://perma.cc/D9RE-397W.

The easiest way to understand a neural network is to first see how one behaves after its training is complete, as we will describe.[96] From there, we will work backwards and explain how training occurs. For purposes of explanation, we will use a convolutional neural network (CNN). These networks are often used for image classification and provide the easiest means to begin to understand an otherwise difficult-to-illustrate concept.[97]

1. The Neural Network

The first place to start is with a mathematical representation of some goal; here, our goal will be to correctly identify hand-drawn digits captured in a digital image. Digital images like those you see on your computer screen are made up of pixels; each pixel is one specific color, and that color results from mixing a specific amount of red, green, and blue—known to your computer as a combination of numbers.[98] Assume we start with the digital image of a hand-drawn one.

---

96. For further introduction, see Jeremy Howard, *Lesson 3: Deep Learning 2018*, YOUTUBE (Dec. 30, 2017), https://perma.cc/EK6G-LD39; *see also* Otavio Good, *A Visual and Intuitive Understanding of Deep Learning*, YOUTUBE (Nov. 5, 2017), https://perma.cc/3JJD-G2CF.

97. Although we later describe the production of synthetic data from a text standpoint, images may be synthetically replicated as well. *See* John T. Guibas et al., *Synthetic Medical Images from Dual Generative Adversarial Networks*, *in* 31 PROC. NEURAL INFO. PROCESSING SYSTEMS 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.").

98. David Eck explains the concept with even greater clarity and comprehensiveness: "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." DAVID J. ECK, INTRODUCTION TO COMPUTER GRAPHICS 11 (version 1.2, 2018). Continuing:

> 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.

Id. at 19 (emphasis omitted); *see also* Victor Powell, *Image Kernels* (Jan. 29, 2015), https://perma.cc/LLY3-MJ79.

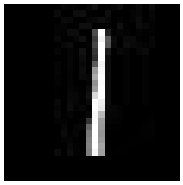


Figure 1[99]

We can then assign 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.[100] 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."[101]

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 2[102]

99. Yann LeCun et al., *The MNIST Database of Handwritten Digits*, https://perma.cc/UN3B-AVZH (archived Oct. 15, 2018). Note that reproducing this image requires writing a basic computer program that interacts with data provided.

100. *See* Howard, *supra* note 96, at 50:33 (using a spreadsheet to map a larger grid); *see also* Jeremy Howard, *deeplearning1*, GITHUB (Dec. 31, 2016), https://perma.cc/7T6P-HLJL (containing the spreadsheet used in the exercise).

101. *See* Howard, *supra* note 96.

102. Here we see a 5x5 grid as a mathematical representation of the digital image of a one; the contours of the one are outlined by those points where the values are 1.0 instead of 0.0. For a larger example of this representation, see Jean-Carlos Paredes, *Understanding Neural Networks Using Excel*, MEDIUM: TOWARDS DATA SCI. (Nov. 18, 2017), https://perma.cc/HF35-SRTE.

The neural network would then run the grid through a series of "convolutions."[103] A convolution is simply a filter[104] or filters applied to the numbers that make up the grid.[105] Here is an example filter which happens to highlight the vertical lines in the grid:

Filter		
1	0	-1
1	0	-1
1	0	-1

Figure 3[106]

To apply the filter, we first slide the filter on top of the input grid, covering a 3x3 set of cells at a time, or nine of the twenty-five cells in the input grid. We then multiply each cell in the input against each cell in the filter and find the sum (e.g., the top-right 3x3 grid would be: 1 * 1 + 0 * 0 + 0 * - 1. We sum all nine results together, and record them as the result of a single convolution. If the sum is negative, we can use a zero in its place.[107] We shift the

103. A convolution pulls out the features of an image (e.g., detecting edges, sharpening, or blurring). *See* Ujjwal Karn, *An Intuitive Explanation of Convolutional Neural Networks*, DATA SCI. BLOG (Aug. 11, 2018), https://perma.cc/3VLE-UXFG ("The primary purpose of [c]onvolution . . . is to extract features from the input image. Convolution preserves the spatial relationship between pixels by learning image features using small squares of input data."); *see also* Alex Krizhevsky et al., *ImageNet Classification with Deep Convolutional Neural Networks*, *in* 60 COMM. ACM 84, 87 fig.2 (2017) (pictorially representing each of the convolutions).

104. A filter is a sequence of mathematical operations on the grid. Adit Deshpande explains a filter this way:

> Now, the best way to explain a conv[olutional] 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).

Adit Deshpande, *A Beginner's Guide to Understanding Convolutional Neural Networks*, https://perma.cc/4X7B-ENCH (archived Oct. 15, 2018) (emphasis omitted).

105. *See* Howard, *supra* note 92.

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

107. 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 96; *see also* Yee Whye Teh & Geoffrey E. Hinton, *Rate-Coded Restricted Boltzmann Machines for Face Recognition*, *in* 13 PROC. NEURAL INFO. PROCESSING SYSTEMS 872, 872 (2000).

3x3 filter one column to the right and start again. Here, it turns out there are nine possible locations for the filter, and so the output of this filter (i.e., one layer in the neural network) has nine cells.[108] 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 4

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 5

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.[109]

---

108. A filter may be applied to the top-left corner and moved one pixel at a time to the right until the end of that row is reached. The filter may then be moved back to the first column but shifted one pixel down, and then shifted again pixel by pixel to the right until the end of that row is reached. This continues until the end of the input is reached. *See, e.g.*, Jonathan Hui, *Convolutional Neural Networks (CNN) Tutorial*, JONATHAN HUI BLOG (Mar. 16, 2017), https://perma.cc/TD5B-43MT.

109. These may focus on the outer edges of the drawing, the horizontal edges, or practically anything imaginable. Soham Chatterjee explains why this is valuable:

> 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 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.

Soham Chatterjee, *Different Kinds of Convolutional Filters*, SAAMA (Dec. 20, 2017), https://perma.cc/6JB7-RWQ3; *see also* Deshpande, *supra* note 104.

These convolutions, in combination with other layers,[110] 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 2x2 pooling layer, where we halve 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.[111] 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 6

The end result of the fully connected layer is the number 4.5 (i.e., the far-right column would be $3 * .5 + 3 * .5 + 3 * .5 = 4.5$). This is the result of a single convolution layer plus a single fully connected layer. If this is the

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

111. Howard, *supra* note 96.

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.[112]

In practice, this sequence of adding filters plus a fully-connected layer to produce a single number would occur several times. With each number, we start with the original "input layer," apply a series of filters, and end up with a single number at the output layer.[113] The process teases out distinguishing characteristics of the hand-drawn images of numbers—e.g., the number one is traditionally a single vertical line surrounded by empty space.

This grid of numbers together (here we would have ten of them) will form the basis of our prediction as to which digit (i.e., 0-9) the original image represents.[114] The prediction is often obtained by using the softmax function, a method used to calculate a probability distribution.[115] Softmax starts by finding the 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, the softmax function removes any negative numbers (via exponentiation) and distinguishes more likely predictions from less likely ones.[116] The result will be a set of probabilities—one for each digit—adding up to 1 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.[117] We are now done with the pre-trained CNN and have our prediction in hand: in this case, the model will predict that the image is undoubtedly of the digit "1."

As for what trains a model like this in the first place, it suffices to say this occurs through a process of gradually improving the filters and weights

---

112. *Id.*

113. *Id.*

114. *Id.*

115. 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 by the sum. *See* John S. Bridle, *Probabilistic Interpretation of Feedforward Classification Network Outputs, with Relationships to Statistical Pattern Recognition*, *in* 68 NEUROCOMPUTING 227, 231-32 (1989).

116. *See* Ji Yang, *ReLU and Softmax Activation Functions*, GITHUB (Feb. 11, 2017), https://perma.cc/4VZH-DCTV ("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.").

117. *See, e.g.*, Li Wan et al., *Regularization of Neural Networks Using DropConnect*, 30 PROC. INT'L CONF. MACHINE LEARNING 1058, 1063 (2013) (using a new method and finding the result to be a 0.21% error rate—"We note that our approach surpasses the state-of-the-art result of 0.23% (Ciresan et al., 2012), achieving a 0.21% error rate, without the use of elastic distortions (as used by (Ciresan et al., 2012)).").

discussed above.[118] Starting out, values are randomly assigned to the filters and weights (i.e., the parameters). Then, typically through a process known as stochastic gradient descent,[119] the values assigned to 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).[120] Stated otherwise, the weights and filters are adjusted to find the optimal combination of numbers to achieve some desired result such as the accurate identification of hand-drawn digits.

But this raises 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 determines how we design the neural network: context. The CNN we used above to identify hand-drawn characters can walk through layers of convolutions in an independent fashion, without one layer's result affecting the operation of the others.[121] For example, an edge-detecting filter's ability to detect edges does not depend on what filters were applied before it. No layer depends on the others. A neural network required to understand sentences, however, would need to be designed differently because speech relies heavily on context and on order. And context comes from memory—something a CNN lacks, but an RNN boasts.

A recurrent neural network (RNN) uses the same layer-by-layer approach as the CNN. The main difference for an RNN is that another input is added after the first filter is applied.[122] For example, 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.[123] That is, instead of simply adding more layers of convolutions on top, we merge the equivalent of another picture to the mix. This

118. For an overview, see *How Neural Networks Are Trained*, MACHINE LEARNING FOR ARTISTS, https://perma.cc/Q2X8-S9S7 (archived Oct. 15, 2018).

119. *See generally* Sebastian Ruder, An Overview of Gradient Descent Optimization Algorithms 2 (June 15, 2017) (manuscript), https://perma.cc/H8B2-3UYY.

120. *See* Vitaly Bushaev, *How Do We 'Train' Neural Networks*, MEDIUM: TOWARDS DATA SCI. (Nov. 15, 2017), https://perma.cc/88AX-ZZ2C; Howard, *supra* note 96.

121. 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 roles is independent of the other.

122. Howard, *supra* note 92.

123. *Id.*

process allows the network to gain memory. The end result, like the final 4.5 we produced in the CNN example, depends not only on a single input entered in the beginning, but also on the intermediate input injected within the hidden layers.

This brings us to the last concept in our nutshell: Generative Adversarial Networks (GANs).[124] The sine qua non of a GAN, a recent invention,[125] is its ability to generate *similar* data.[126] 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 subparts.

3. Generative Adversarial Network

The easiest way to think of a GAN is through the example of the production of counterfeit money.[127] 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[128]), a GAN uses two models pitted against each other in an iterative loop.[129] 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.[130]

---

124. *See generally* Ian J. Goodfellow et al., *Generative Adversarial Nets*, *in* 27 PROC. NEURAL INFO. PROCESSING SYSTEMS 2672, 2672 (2014); *see also* Martin Arjovsky et al., *Wasserstein GAN* (Dec. 6, 2017), https://perma.cc/R4JU-KTJE.

125. *See* Goodfellow, *supra* note 124 (seminal paper on GANs published in 2014). In fact, GANs are one of the reasons synthetic data has recently received attention. *See e.g.*, JAKUB LANGR & VLADIMIR BOK, GANS IN ACTION 2 (2018) ("[W]hile artificial intelligence and machine learning have historically been excellent at teaching computers to discern ever more intricate patterns in data and master ever more complex gameplays, they have been poor at teaching them to generate new data—something that we, humans, do every day as we converse with one another, come up with innovative solutions to problems, and express our creativity through art. This all changed in 2014 when Ian Goodfellow invented [GANs]. . . . GANs have achieved remarkable results that have long been considered virtually impossible for artificial systems, such as the ability to generate fake images in real-world-like quality, turn a scribble into a photograph-like image, or turn a video footage of a horse into a running zebra—all without the need for vast troves of painstakingly-labeled training data. As such, [GANs are] hailed by industry experts as one of the most important innovations in deep learning."); Christopher Bowles et al., *GAN Augmentation: Augmenting Training Data Using Generative Adversarial Networks* (Jan. 8, 2019), https://perma.cc/K9SH-73L2 (discussing the use of GANs to increase the availability of medical training data).

126. *See* Goodfellow, *supra* note 124.

127. *Id.* at 2672.

128. This is unlike the CNN or RNN, both of which typically only use one model. *See supra* Subparts II.A.1-2.

129. *Id.*

130. Notably, there are no constraints on the specific types of models used in the

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 repeatedly, assuming theoretically ideal conditions, an equilibrium is reached in which the discriminator is unable to distinguish between real and fake data.[131] This is why GANs are becoming the go-to for synthetic data generation[132]—they have the ability to generate similar data (e.g., deepfakes[133]) with better results than seen before.

Concluding our brief discussion of machine learning, it is important to remember the core[134] technology at hand: the neural network, the layers of convolutions discussed in the first example.[135] Importantly, regarding synthetic data, these networks may also be paired with differential privacy.[136] The addition of differential privacy to neural networks drew interest as early as 2016;[137] however, it was not until 2018 that researchers realized the potential implications and advantages of applying the technique to GANs.[138] Without delving into the technical details, privacy is added by implementing noise into the training data (i.e., creating the filters and weights). With this understanding at hand we may now move onto our case

---

GAN.

131. 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." *What Is the Nash Equilibrium and Why Does It Matter?*, ECONOMIST (Sept. 7, 2016), https://perma.cc/WCX9-KQJN; *see also* Tim Salimans et al., *Improved Techniques for Training GANs*, 30 PROC. NEURAL INFO. PROCESSING SYSTEMS 2234 (2016).

132. *See infra* note 148 and accompanying text.

133. *See* James Vincent, *All of These Faces Are Fake Celebrities Spawned by AI*, THEVERGE (Oct. 30, 2017, 7:05 AM), https://perma.cc/795N-9L87 ("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.").

134. And for the deft reader, 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* (Oct. 2018), https://perma.cc/L8AZ-ZCFF; *see also* Howard *supra* note 92.

135. *See supra* Part II.A.1.

136. *See generally* Nissim et al., *supra* note 73.

137. *See* Martín Abadi et al., *Deep Learning with Differential Privacy*, 2016 PROC. CONF. COMPUTER & COMMS. SECURITY 308, 308 (applying differential privacy techniques to machine learning language modeling).

138. *See* Liyang Xie et al., *POSTER: A Unified Framework of Differentially Private Synthetic Data Release with Generative Adversarial* Network, 2017 PROC. CONF. COMPUTER & COMMS. SECURITY 2547.

study demonstrating the feasibility and utility of synthetic data generated through a GAN.

B. Case Study: Generating and Evaluating Synthetic Data

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).[139] 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.[140] Paradoxically, despite interest in insider threat detection, it is an area devoid of publicly available data since the data is expensive to attain and privacy-invasive by nature. This is because the data necessary to detect an insider threat is very fine-grained,[141] and its collection causes 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.[142]

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.[143] The original data was acquired by having each cadet install software on their machine collecting information

---

139. *See* Ewen Macaskill & Gabriel Dance, *NSA Files: Decoded*, GUARDIAN (Nov. 1, 2013), https://perma.cc/2YW7-ZN5Z.

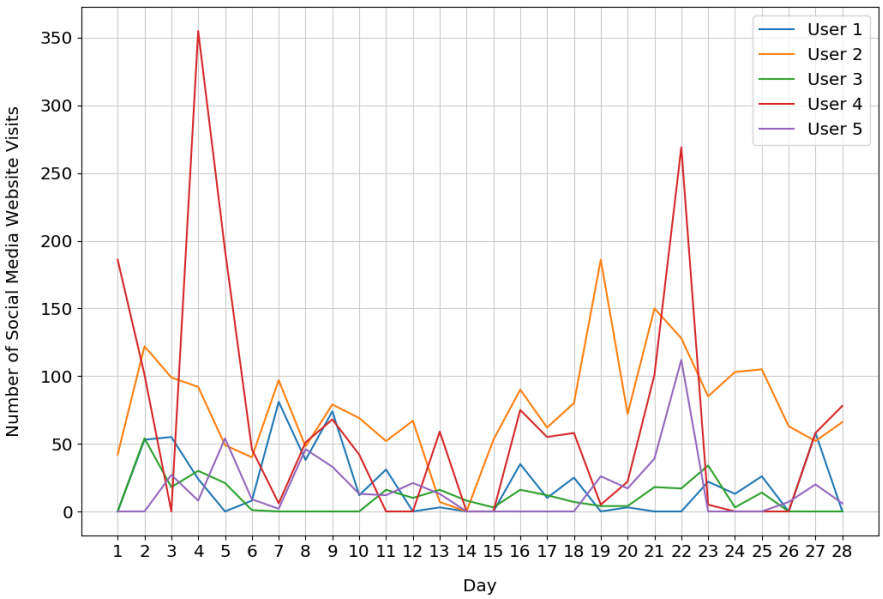
140. *See generally* INSIDER ATTACK AND CYBER SECURITY: BEYOND THE ATTACKER (Salvatore J. Stolfo et al. eds. 2008).

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

142. *See* Preetam Dutta et al., *Simulated User Bots: Real Time Testing of Insider Threat Detection Systems*, 2018 IEEE SYMP. SECURITY & PRIVACY WORKSHOP 228, 228. The data from the West Point cadets was gathered under an IRB-approved protocol.

143. *See id.* at 232. 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/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." *Id.* 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.

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.[144]



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, and detail), an RNN is the best neural network architecture for the job.[145] 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).[146] This type of RNN leverages not only an

144. 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.

145. *See supra* Part II.A.2 (explaining how RNNs are typically used for text-based generation).

146. *See generally* Sepp Hochreiter & Jürgen Schmidhuber, *Long Short-Term Memory*, 9 NEURAL COMPUTATION 1735, 1735 (1997); *see also* Christopher Olah, *Understanding LSTM Networks*, COLAH'S BLOG (Aug. 27, 2015), https://perma.cc/EE26-C6Y6

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.

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).[147] This metric looks at the frequency of word-use in a document. After grouping, we could then assess the similarities or differences between the raw and synthetic data.

("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.").

147. *See* Gerard Salton & Christopher Buckley, *Term-Weighting Approaches in Automatic Text Retrieval*, *in* 24 INFO. PROCESSING & MGMT. 513, 513 (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,* 1998 PROC. IEEE INT'L CONF. ON ACOUSTICS, SPEECH & SIGNAL PROCESSING 645, 647-48; *see also* Charu C. Aggarwal & Philip S. Yu, *A Condensation Approach to Privacy Preserving Data Mining*, ADVANCED DATABASE TECH. 183, 183 (2004).

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.[148] What is more, even beyond our case study, similar research concurs in our results: Synthetic data is a valid alternative to original data.[149]

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.[150] 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."[151] 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.[152] In summary, the "usefulness" of synthetic data has been validated by not only our work, but also the work of others.

However, this is not to say that synthetic data is the "silver bullet" data scientists and privacy activists have been searching for.[153] As the deidentification saying goes, just because the dataset appears anonymous does not

148. *See* Dutta et al., *supra* note 142.

149. *See generally* Neha Patki et al., *The Synthetic Data Vault*, 2016 IEEE INT'L CONF. DATA SCI. & ADVANCED ANALYTICS 399, 400-10 (demonstrating a technique—the synthetic data vault—used to create synthetic data from five publicly available datasets).

150. *See id.* at 409 ("[For] 7 out of 15 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. MACHINE LEARNING HEALTHCARE 286, 296 (2017) ("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.").

151. *See* Patki et al., *supra* note 149, at 409.

152. *See* Noseong Park et al., *Data Synthesis Based on Generative Adversarial Networks*, *in* PROC. 11TH VLDB ENDOWMENT 1071, 1074 (2018).

153. Chong Huang et al., *Context-Aware Generative Adversarial Privacy*, 19 ENTROPY 656, 656 (2017) (implementing a "context-aware privacy framework"); Brett K. Beaulieu-Jones et al., *Privacy-Preserving Generative Deep Neural Networks Supporting Clinical Data Sharing* (July 5, 2017), https://perma.cc/X65F-36VE (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); Vincent Bindschaedler et al., *Plausible Deniability for Privacy-Preserving Data Synthesis*, *in* PROC. 10TH VLDB ENDOWMENT 481, 481 (2017) ("[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); Matt Fredrikson et al., *Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures*, *in* PROC. 22ND CONF. COMPUTER & COMMS. SECURITY

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

1322, 1322 (2015) ("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."); Aleksei Triastcyn & Boi Faltings, *Generating Artificial Data for Private Deep Learning* (June 7, 2018), https://perma.cc/UWR9-XQXB ("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.").

1. Too Individualized

First off, one inherent characteristic of synthetic datasets is that they may leak information.[154] In computer science parlance, this is referred to as overfitting a model, which may result in particular data being "leaked."[155] 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.[156] Consequently, that fact will be reflected in the model, and may show up in some synthetic records.[157] 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 excessively.

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

154. The term leak here means to permit discovery of facts that were assumedly hidden by the process of synthetic data generation. *See* Samuel Yeom et al., *Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting*, 31 IEEE COMP. SECURITY FOUNDS. SYS. 268, 281 (2018) (finding privacy leaks in machine learning models as owed to overfitting—and even other, more subtle features of the models); *see also* Tyler Hunt et al., *Chiron: Privacy-Preserving Machine Learning as a Service* (Mar. 15, 2018), https://perma.cc/QZ8S-UHZY.

155. *See generally* Michael Veale, Reuben Binns & Lilian Edwards, *Algorithms that Remember: Model Inversion Attacks and Data Protection Law*, 376 PHIL. TRANSACTIONS ROYAL SOC'Y 1, 3- 6 (2018) ("It has been demonstrated that machine learning models are vulnerable to a range of cybersecurity attacks that cause breaches of confidentiality. Confidentiality attacks leak information to entities other than those whom designers intended to view it.").

156. *See* Nicolas Papernot & Ian Goodfellow, *Privacy and Machine Learning: Two Unexpected Allies?*, CLEVERHANS-BLOG (Apr. 29, 2018), https://perma.cc/W93V-6VR8 ("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.").

157. 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* DELTA JUNCTION DATING, https://perma.cc/X2VE-NHBA (archived Oct. 26, 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.

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.[158]

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 trying to avoid.[159] 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.[160]

A better solution here is to use differential privacy in combination with synthetic data generation.[161] Though a relatively new technique, the results are very promising.[162] Utility is sustained through data generation and privacy is obtained—up to a certain threshold—via the robust guarantees of differential privacy.[163] However, setting that threshold[164] 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.[165]

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

158. *But see* Florian Tramèr et al., *Stealing Machine Learning Models via Prediction APIs*, *in* 25TH USENIX SECURITY SYMP. 601, 601 (2016) (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).

159. *See* Ohm, *supra* note 6, at 1716-31 (describing the failures of anonymization).

160. Nicholas Carlini et al., The Secret Sharer: Measuring Unintended Neural Network Memorization and Extracting Secrets 11-12 (Feb. 22, 2018) (unpublished manuscript), https://perma.cc/JD9Z-5DPQ.

161. Cynthia Dwork & Vitaly Feldman, *Privacy-Preserving Prediction*, 75 PROC. MACHINE LEARNING RESEARCH 1, 1 (2018); Carlini et al., *supra* note 160, 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).

162. H. Brendan McMahan et al., *Learning Differentially Private Recurrent Language Models*, 6TH INT'L CONF. ON LEARNING REPRESENTATIONS, May 2018, at 1-2.

163. *See* Abadi et al., *supra* note 137 (applying differential privacy techniques to machine learning language modeling).

164. *See supra* note 87 and accompanying text (referring to the epsilon parameter).

165. *See supra* Subpart I.B.3.ii.

data to force leakage. These attacks are known generally as adversarial machine learning.[166] 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.[167] Along those same lines, other research[168] goes even further to suggest that if the attacker has full access to the model's code,[169] 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

166. *See* Ling Huang et al., *Adversarial Machine Learning*, 4 PROC. WORKSHOP SECURITY & ARTIFICIAL INTELLIGENCE 43, 43 (2011); Ivan Evtimov et al., *Robust Physical-World Attacks on Machine Learning Models*, COMPUTER VISION & PATTERN RECOGNITION, June 2018, at 1.; *see also* Alexey Kurakin et al., *Adversarial Machine Learning at Scale*, 5TH INT'L CONF. ON LEARNING REPRESENTATIONS, Apr. 2017, at 1; Gamaleldin F. Elsayed et al., *Adversarial Examples that Fool Both Computer Vision and Time-Limited Humans*, 32 PROC. NEURAL INFO. PROC. SYS. 1, 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, http://perma.cc/AR3J-UGHF (archived Oct. 26, 2018).

167. *See* Fredrikson et al., *supra* note 4, at 17 ("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 et al., *Machine Learning Models that Remember Too Much*, 2017 PROC. CONF. ON COMPUTER & COMMS. SECURITY 587, 600 (2017) ("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.").

168. *See* Fredrikson et al., *supra* note 4. 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 et al., *Membership Inference Attacks Against Machine Learning Models*, 2017 IEEE SYMP. SECURITY & PRIVACY 3, 3; Yunhui Long et al., Understanding Membership Inferences on Well-Generalized Learning Models 1 (Feb. 13, 2018) (unpublished manuscript), https://perma.cc/GAU3-AV62 ("[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.").

169. 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. *See* Shokri et al., *supra* note 168, at 3.

access only,[170] the attacker could learn whether a data point was "in" the data set used to train the model.[171]

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.[172] In fact, out of many possible solutions to the model-access problem, differential privacy has been noted as the only solution to sufficiently protect privacy while maintaining utility.[173] In this sense, differential privacy provides both a way to escape data leakage and adversarial machine learning.[174]

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.[175] The conundrum, though improved, is not completely solved by synthetic data.[176]

---

170. 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.*

171. *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.*

172. *See* Long et al., *supra* note 168, at 13 ("Differential privacy is a prominent way to formalize privacy against membership inference."); Bindschaedler et al., *supra* note 153 (using a form of differential privacy plus synthetic data).

173. Long et al., *supra* note 168. *But see* Dwork & Feldman, *supra* note 161.

174. Carlini, *supra* note 160, 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.").

175. 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* PRINCIPLES DATABASE SYSTEMS 202, 202-03 (2003).

176. *See, e.g.*, Briland Hitaj et al., *Deep Models Under the GAN: Information Leakage from Collaborative Deep Learning*, 2017 PROC. CONF. COMPUTER & COMMS. SECURITY 603,

III. SYNTHETIC DATA'S LEGALITY

Turning to the legal world, one 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 response. 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[177] the produced data may be deemed "vanilla" synthetic data. The generation process is as bare-bones as possible. Data in, data out. Unfortunately, as Part II.C demonstrates, this could result in data leakage: secrets in, secrets out.

Per data leakage, pairing vanilla synthetic data with privacy statutes 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 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, 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.

606-07; Jamie Hayes et al., LOGAN: Evaluating Privacy Leakage of Generative Models Using Generative Adversarial Network 1-2 (Aug. 21, 2018) (unpublished manuscript), https://perma.cc/S9UF-VQVN. On the other hand, see Stefanie Koperniak, *Artificial Data Give the Same Results as Real Data—Without Compromising Privacy*, MIT NEWS (Mar. 3, 2017), https://perma.cc/GWC5-YSTG.

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

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. Moreover, considering evenly distributed training data, even if the model reproduced a particular datum that turned out to be real, this would not automatically mean an individual could be deidentified.[178] Assuming an adversary learns zip code 10004 within the database is real (e.g., assume the row in the database has the following fields: <name>, <zip code>, <HIV status>) this 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.[179] 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. Yet, consider that even complex computer science methods of extracting secrets from vanilla synthetic data (i.e., attempts to identify "leaks"[180] in the dataset) have been shown to be hit or miss.[181] Researchers using sophisticated methods to extract secrets in a vanilla synthetic dataset were only successful[182] three times in seven—even when the likelihood that a secret was in the synthetic dataset was over four thousand times more likely than a random word.[183]

178. To visualize this, see Figure 4 in Triastcyn and Faltings's work, *supra* note 153. The image displays real versus fake pictures of numbers, illustrating how many numbers look similar, but are not.

179. 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 graver consequences.

180. *See supra* Part II.C.1 (describing leaky data).

181. Carlini, *supra* note 160, at 10, tbl.5.

182. The researchers did not exhaustively search for a secret but 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.* (using social security number and credit card number).

183. *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.

For another perspective, look at it under the lens of the *Sander* case, where professor Richard Sander litigated over access to the State Bar of California's bar admission database.[184] On remand from the California Supreme Court, the California Superior Court assessed whether any method of de-identification would sufficiently protect students' privacy rights under FERPA while allowing Professor Sander to use the database.[185] Of four proposed methods relying on deidentification, the court found only one sufficiently protected privacy by using k-anonymity—though it destroyed utility.[186] Importantly, Professor Sweeney pegged the most-privacy-preserving method proposed as having a 31% chance of reidentification.[187] And to the court, this far exceeded the acceptable privacy risk.[188] The court prohibited the data from being disclosed.[189]

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)—making disclosing sensitive bar data unwarranted in one case or unwise in the other.

Likewise, vanilla synthetic data makes no guarantee that a dataset is 100 percent free of *all* real identifiers. Where, as here in *Sander*; sensitive facts are on the line, urging a court to permit data release will be a tough sell. In other words, when data disclosure is governed by a stringent statute,

---

184. *See* Sander v. State Bar of California, No. CPF-08-508880 (Nov. 7, 2016) (on file with author); *see generally* Latanya Sweeney et al., *Saying It's Anonymous Doesn't Make It So: Re-Identifications of "Anonymized" Law School Data*, J. TECH. SCI., Nov. 2018.

185. Sander, No. CPF-08-508880 at *19.

186. *Id.* at *4. The other proposals included various levels of *k*-anonymity, removal of traditional identifiers like name and zip code, and the use of a secure data enclave. *Id.* at *17-21.

187. *Id.* at *17-21; *see also* Sweeney et al., *supra* note 184, 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.

188. *See* Sweeney et al., *supra* note 184, at 9 ("[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%. Sander, No. CPF-08-508880 at *19.

189. Sander, No. CPF-08-508880 at *21-22.

any chance of identification—no matter how low—may prohibit the release of a dataset into the wild.[190]

2. Under-Inclusive Privacy

Insensitivity to chance identification is also possible. With statutes 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, the generated synthetic data would present a "theoretical" rather than concrete chance of identification: any expected data leakage is unlikely to enable the identification of an individual. Here, however, the chance is likely low enough for a court to permit disclosure. Consider the unifying themes in *Pruitt*, *In re Hulu*, *Eichenberger*, and *In re Nickelodeon*.

The courts in these cases, focusing on the CCPA and VPPA, held that "anonymized" identifiers (i.e., user IDs, device serial numbers, or hexadecimal codes) could be publicly shared without violating consumers' PII. In *Yershov*, facing a relatively straightforward issue, the court found the combination of geolocation, device identification, and content viewed to be PII and therefore protected.[191] However, in *Pruitt* the court faced a more nebulous situation, debating whether converter box codes used by a company to map customers with business information were also PII.[192] Because the codes were simply a series of digits, and the mapping from codes to customers was not publicly shared, the court found these codes to be non-PII.[193]

---

190. 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 *Sander*, this is not a sure bet, and may result in clashing expert opinions.

191. Yershov v. Gannett Satellite Info. Network, Inc., 820 F.3d 482, 485 (1st Cir. 2016) (disclosing the following information: "[E]ach time Yershov watched a video clip on the App, Gannett disclosed to Adobe the title of the viewed video, Yershov's unique Android ID, and the GPS coordinates of Yershov's device at the time the video was viewed. Using this information, Adobe was able to identify Yershov and link the videos he had viewed to his individualized profile maintained by Adobe.").

192. *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.").

193. *Id.* at 715 ("To receive [digital cable service], subscribers must have a special

Building on this holding, *In re Hulu*[194] and *Eichenberger*[195] drew the line on the mere *theoretical* possibility of data linkage when sharing information.[196] These courts held that "randomly"[197] generated user IDs like the converter box codes in *Pruitt* were useless without a master table mapping the codes to real identifiers. Curtly stated, if reidentification of data is hard or time-consuming, that means the data is not PII. As 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

converter box installed and attached to their telephone line. The converter boxes, manufactured by Motorola, transmit and store (1) pay-per-view purchase information, (2) system diagnostic information and (3) settop bugging information. Each converter box contains a code displayed in hexadecimal format indicating the date of a pay-per-view purchase and a source identifier for the pay-per-view channel. The converter box stores a maximum of sixty-four purchases. When total purchases exceed that number, the newest purchase information overwrites the oldest purchase. The converter box also contains a code (again displayed in hexadecimal format) signifying the total number of purchases and payments generated through that particular box. Individual subscriber information is not contained within the converter box, but an identifying number known as a 'unit address' allows Comcast to match the subscriber's purchases to its billing system. The billing system contains the name and address of the household member responsible for payment.").

194. *In re* Hulu Privacy Litig., No. C 11-03764 LB, 2014 WL 1724344, at *10-11 (N.D. Cal. Apr. 28, 2014).

195. Eichenberger v. ESPN, Inc., 876 F.3d 979, 985 (9th Cir. 2017). *But see In re* Vizio, Inc., 238 F. Supp. 3d 1204, 1212 (C.D. Cal. 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").

196. *In re* Hulu Privacy Litig., 2014 WL 1724344, at *11 ("In sum, the statute, the legislative history, and the case law do not require a name, instead require the identification of a specific person tied to a specific transaction, and support the conclusion that a unique anonymized ID alone is not PII but context could render it not anonymous and the equivalent of the identification of a specific person."); Eichenberger, 876 F.3d 985-86 ("Plaintiff alleges that Defendant disclosed to Adobe: (1) his Roku device serial number and (2) the names of the videos that he watched. As Plaintiff concedes, that information *cannot* identify an individual unless it is combined with other data in Adobe's possession—data that ESPN never disclosed and apparently never even possessed. Indeed, according to Plaintiff, Adobe can identify individuals only because it uses a complex 'Visitor Stitching technique' to link an individual's Roku device number with other identifying information derived from 'an enormous amount of information' collected 'from a variety of sources.' We conclude that an ordinary person could not use the information that Defendant allegedly disclosed to identify an individual.") (emphasis in original).

197. *But see* Jason M. Rubin, *Can a Computer Generate a Truly Random Number?*, ASK AN ENGINEER (Nov. 1, 2011), https://perma.cc/V9QU-K5F7 ("'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.'").

> 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.[198]

Finally, pushing the line the farthest, *In re Nickelodeon Consumer Privacy Litigation*[199] found that networking cookies and an IP address were not PII—making *Yershov*'s understanding come full circle: "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."[200] According to this line of cases at least, non-detective work's boundary is the combination of geolocation, device identification, and content viewing history.

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*,[201] and even if relatively simple techniques could be used to track users across time with the released "non-identifiers," as seen in *In re Nickelodeon*.[202] However, the problem with permitting the sharing of these datasets containing only *theoretical* risk of identification is legion. This line of reasoning is the same one abhorred by the cavalcade of academics criticizing the historical means of anonymization.[203] Indeed, that the database "join"

198. *See Eichenberger*, 876 F.3d at 985.

199. 827 F.3d 262 (3d Cir. 2016).

200. *Id.* at 289 (citing Yershov v. Gannett Satellite Info. Network, Inc., 820 F.3d 482, 486 (1st Cir. 2016)).

201. 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. *See supra* note 74.

202. *In re Nickelodeon*, 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.").

203. 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

operation would not work on synthetic data does not mean the method is free of all threats.

As outlined in Part 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.[204] 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.[205] 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, prohibiting data release in the face of a low chance of identification. On the other hand, we may overestimate synthetic data's protection, and we can all agree that identifying individuals in a medical dataset, for instance, should be avoided. Congress should strike a new balance between over- and under-inclusive protection in light of new understandings of privacy-preserving techniques and the chance of identification. However, in the short term, a technical solution—differential privacy—should be pushed.

B. Differentially Private Synthetic Data

Ultimately, data leakage and the threat of techniques like adversarial machine learning result in the same dilemma identified in Part II.C: Even with a reasonable distribution of input records, there exists a *theoretical* 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 (and the optimal means of applying differential

same time ensured because "theoretic" identification is not identification at all? And the same is true for vanilla synthetic data.

204. *See* Long et al., *supra* note 168 and accompanying text.

205. *See supra* notes 176-84 and accompanying text.

privacy to synthetic data is not yet settled[206]), differential privacy nonetheless provides a better way of assuring privacy given 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 theoretical exercise, meaning the model resists even sophisticated attempts to reveal identities. In the *Sanders* case, synthetic data plus 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).[207] 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 are not silver bullets. 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.

No privacy preserving technique will completely solve the database-privacy problem. Indeed, if utility is of paramount concern, neither synthetic data nor differential privacy (nor even 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, 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.

206. *See supra* Part II.C.
207. *See* Carlini, *supra* note 160, at 12.

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 appreciate the full benefits or risks of synthetic data.[208] And third, synthetic data—if constructed properly—may solve Professor Ohm's failure of anonymization.[209] We briefly summarize the first two points in concrete recommendations.

First, we recommend that data managers and researchers use synthetic datasets instead of anonymized ones when appropriate. The chief reason for this recommendation is to avoid the arms race between deidentification and reidentification. True, secure anonymization via deidentification may be possible, albeit difficult;[210] however, the availability of secondary sources of information unknown to the sanitizer of the real data makes it a risky bet.[211] With synthetic datasets, we largely escape that trap.

Second, we recommend privacy statutes be amended to enable the use of synthetic datasets. Most of today's statutes are absolute: they bar disclosure of PII. While the actual metrics may be statistical—i.e., the HIPAA rules effectively use *k*-anonymity[212]—the goal is the same. No information may be disclosed about identifiable individuals.

Synthetic datasets are different. These datasets protect privacy through the addition of statistically similar information, rather than through the stripping away of unique identifiers. This, in turn, invites statutory ambiguity: the resulting datasets may leak information, but these leaks may or may not be enough to bar disclosure (resulting in over and under inclusive privacy coverage).

---

208. *See supra* Part II.A.

209. *See* Ohm, *supra* note 6.

210. *See* Dataverse, *HarvardX-MITx Person-Course Academic Year 2013 De-Identified Dataset, Version 2.0*, https://perma.cc/886F-TG67 (archived Oct. 26, 2018).

211. This was the case in the AOL and Netflix data dumps: linking the original data with other sources was sufficient to re-identify some of the records. *See supra* note 74 and accompanying text.

212. 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* Subpart II.B.3.i.

Our recommendation is therefore to face the ambiguity head on. New or amended statutes should accommodate synthetic data,[213] accepting the possibility of *measurably*[214] small privacy leakage in exchange for perhaps mathematically provable protection against reidentification.[215] 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 same tradeoff.[216]

CONCLUSION

Databases play a key role in scientific progression. Yet, the very fuel of databases, open data, is more like shale oil given our current privacy laws. Since the early days of reidentification, a Catch-22 has emerged: We have the ability to gather and process enormous amounts of information, but we are forced to act cautiously because of the ambiguity found in our legal statutes. We must either act too greedily by thoroughly stripping all identifiers from the data before sharing it, thereby making it useless, or too *Laissez-Faire* by being permitted to disclose the fact that "a local high school teacher rented a particular obscure French film," even if there was only one high school teacher in the area who spoke French.

Synthetic data offers progress. Though not a silver bullet, the method allows us to put an end to the deidentification–reidentification arms race

213. 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(a)(2)(A) (1984) ("[T]he term 'personally identifiable information' does not include any record of aggregate data which does not identify particular persons"—likewise, a new statute could include a statement that differentially private synthetic data is not PII).

214. 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.

215. 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*, 2014 IEEE SYMP. SECURITY & PRIVACY 359, 359.

216. 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.*, Maureen Ohlhausen, Acting Chair, Remarks at the Painting the Privacy Landscape: Informational Injury in FTC Privacy and Data Security Cases (Sept. 19, 2017).

and focus on what matters: useful, private data. To this extent, we recommend the privacy community accept synthetic data as a valid, next step to the database privacy problem.