De-identification White Paper
DRAFT Jan. 31, 2015

By Jules Polonetsky, Christopher Wolf, and Kelsey Finch

The current de-identification landscape is rife with uncertainty and risk for organizations, consumers, and regulators alike. While the policy debate continues to assume a binary framework where personal information is either identified or not, already a wide range of intermediary approaches appear in practice. Predicated on disagreements and misunderstanding about what data is or should be considered de-identified, the divide between how de-identification techniques are discussed and how they are deployed continues to grow. In order to find a path forward, we must change both how de-identification is discussed and how it fits within broader data protection and privacy debates.

This paper seeks to describe the current de-identification debate and explain the basis for the discord among policymakers and stakeholders. Next, we reframe the data spectrum, placing personal information that has been subjected to technical and/or administrative controls into three categories with different sets of controls and privacy protections applying to each set. Then, we examine the relevance of our approach to the existing U.S. and EU frameworks, both in theory and in practice. Finally, we describe critical measures for the success of de-identification and pseudonymization, so that important data uses can advance in a manner that considers both their benefits and risks.

Introduction

Since long before the computer age, consumers' personal information has been used to drive new product development and help organizations maintain their current services. Even in the pre-industrial age shopkeepers recorded their customers' transactions and purchasing habits, census takers collected individual demographic information by hand, and communities allocated resources on the basis of personal information. Modern computational power has magnified these activities, unleashing the era of Big Data where sophisticated analytics and large, detailed databases work together constantly to put personal data to new uses. Today, personal data are driving scientific and medical advances, more inclusive curricula, more efficient infrastructure, and a revolutionary wave of innovative technologies and services around the globe.

In numerous fields of inquiry, researchers are putting geolocation, health, traffic, education, environmental, census and mobile carrier data, among others, to new and unanticipated uses. Even when collected to provide or maintain services, personal data can support secondary analysis on a vast scale. The ability to track and analyze various data trends over time has led to advances in education, health, public services, business and technology to the benefit of both individuals and society. Appendix A to this document provides a number of detailed examples of such uses of data.[1]

[1] Appendix A forthcoming.

In today's world, researchers believe that additional advances will be achievable if they have access to increasingly extensive and detailed sets of information. Researchers depend on personal and quasi-personal data to improve equipment and analytics engines across every industry. Their impact can already be seen in new tools and techniques helping to make x-ray machines better, drugs safer and more effective, and transportation more secure. Much of this societally beneficial research also takes place in private hands. In the business world, just as in scientific communities, access to larger and more detailed data sets is seen as a door to better and safer products, more helpful customer service, and a more competitive information economy. However, while the scale and scope of the data available today bring new, unexpected benefits, they also introduce new and unexpected privacy risks.

In many of these areas, organizations apply de-identification or pseudonymization techniques and measures in order to minimize or eliminate privacy risks to individual data subjects. To de-identify data, those elements which were "personal" because they referred to a particular individual are eliminated. To pseudonymize data, on the other hand, organizations attempt to obscure the connections between individuals and their personal information without completely destroying those connections. As pseudonymous data is therefore still link*able* to an individual, they pose a slightly higher privacy risk, although still far lower than the risk of unaltered personal data. At the same time, the more that data is manipulated, the less useful and reliable it becomes.

De-identification and pseudonymization enable critical public and private research by allowing for the maintenance and use – and, in certain cases, the sharing and publication – of valuable information sets, even those based on sensitive personal information. However, widespread debate continues to take place over the ways and means, as well as fundamental feasibility, of de-identification. The abundance of Big Data is believed to undermine de-identification efforts through powerful new computing capabilities that can identify data previously considered to be non-identifiable. In many cases, de-identification relies on organizational commitments to keep data from public disclosure and subject to legal or administrative protections, but these promises are often viewed skeptically by critics.

Critics also point to well publicized examples of re-identification, such as the re-identification of individuals from the release of purportedly de-identified AOL Search data, a Massachusetts medical database, or Netflix recommendations. In each of these cases, "Even though administrators had removed any data fields they thought might uniquely identify individuals, researchers . . . unlocked identity by discovering pockets of surprising uniqueness remaining in the data." Given the sophistication of the data handlers in these cases and the repeated success of re-identification attacks, critics conclude that "de-identification fails to resist inference of sensitive information either in theory or in practice."

Defenders of de-identification argue that the attacks were on databases that were not credibly de-identified in any acceptable manner and should not be used to undermine respect for serious de-identification measures. In some cases the debate is over the role of utility of data and relevant risks. Can de-identification that takes into account relevant risks, but not all risks, and that seeks to ensure that the final data set has utility for the intended uses be considered acceptably de-identified? Must every data set (even if held internally) be considered to be a public data set due to potential breaches? Can we trust organizations when they commit to keep data internal or share it only with trusted partners?

If we do not find a way to resolve these questions, we all stand to lose. There are important policy roles for stakeholders working to advance the cutting edge of de-identification science, as well as for those seeking to facilitate the widespread adoption of pragmatic de-identification measures. Furthermore, we must not lose sight of the important role of pseudonymous data in these debates. If de-identification and pseudonymization render data unusable, or are held to impossible standards, we will be denied the socially beneficial activity arising from uses of that data, whether in private or public hands. The ability to distinguish between individuals underlies

many legitimate business and scientific activities in both the U.S. and EU. A black-and-white approach risks both over- and under-protecting personal information, or needlessly sacrificing the utility of data.

The De-Identification Landscape

The Debate

Academics, technologists, regulators, advocacy groups, and businesses have sought for years to establish common standards for the de-identification of personal data. Despite broad consensus around the need for and value of de-identification, the debate as to whether and when data can truly be said to be de-identified appears interminable. The axiom that 'data can be either useful or perfectly anonymous but never both' rings louder than ever.[2]

Rather than discuss *how* to de-identify personal information, the discussion has increasingly turned to *whether* personal information can be (or can be said to be) de-identified. Today, broader policy discussions about de-identification largely fall apart first when attempting to determine what level re-identification risk, if any, is acceptable for data to still be called 'de-identified' and second when considering whether or not organizational controls should be considered in that calculus.[3] Although technical research into new de-identification techniques remains ongoing, each time a new de-identification solution is proposed another new re-identification risk seemingly raises its head – and then the debate returns whether or not that data can still be properly considered de-identified. This cycle has even led to some researchers contending that all de-identification efforts may be futile.[4]

Policymakers have often sought to draw bright lines around de-identification by simply declaring that certain data can be shared based upon how it has been aggregated and/or its intended use, whether or not de-identification experts agree. In Colorado, the Public Utility Commission has adopted a "15/15" methodology to govern aggregated consumer data. In order to be considered sufficiently protected for public sharing, a data sample must contain more than 15 customers and no single customer's data may comprise more than 15 percent of the total.[5] On the other hand, a California utility commission took an approach that weighed the appropriateness of sharing based on who the recipient of the data would be and what purposes it would be used for. Even in the education world, regulators determine whether information has been reasonably de-identified based on whether others in the school community may have additional identifying information. Even HIPAA takes a step down this road, providing that healthcare organizations may deem their data "de-identified" under the Safe Harbor standard by simply removing 17 categories of identifiers from a data file – although it then leaves a catch-all 18th category, unwilling to entirely commit to it.[6] While having a clear methodological standard eases the compliance burden of de-identifying information, these methods are often criticized by statisticians.

In Europe, meanwhile, a generalized de-identification standard is rooted within the omnibus Data Protection Directive and further clarified by each national data protection authority. The guidance given by the Article 29 Working Party on anonymization has focused primarily on the role of technical de-identification measures, seeking minimal residual privacy risk before determining that data may be considered "anonymous." While the Working Party cogently presents the technical issues and privacy risks inherent in de-identification, its characterizations of acceptable re-identification risk have been understood by some as requiring near-zero risk,

[2] Paul Ohm, at 1704 http://uclalawreview.org/pdf/57-6-3.pdf

[3] Paul Ohm, Broken Promises of Privacy, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1450006; Cavoukian & Castro, Setting the Record Straight: De-identification Does Work, http://www2.itif.org/2014-big-data-deidentification.pdf; Narayanan & Felten, No Silver Bullet: De-identification still doesn't work, http://randomwalker.info/publications/no-silver-bullet-de-identification.pdf

[4] Is De-identification Sufficient to Protect Health Privacy in Research?, Mark A. Rothstein

[5] http://www.cpuc.ca.gov/NR/rdonlyres/8B005D2C-9698-4F16-BB2B-D07E707DA676/0/EnergyDataCenterFinal.pdf

[6] "…(R) Any other unique identifying number, characteristic, or code…" http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/guidance.html#safeharborguidance

an infeasible standard.[7] EU regulators, too, tend to concentrate on analyzing and applying the highest technical standards available to personal data in isolation, rather than evaluating de-identification measures in light of the practical difficulty of actually re-identifying any particular individual. U.S. regulators emphasize whether an individual's unique identity can be reasonably attributed to them as the locus of de-identification determinations. EU regulators, in contrast, increasingly focus on whether and when an individual can be singled out or treated differently than another individual on the basis of certain information.[8] If so, that information is deemed personal.

The Framework

Part of the force driving policy discussions into these definitional dead-ends is the inherent value of being able to call data de-identified. As one researcher has put it, anonymization is "ubiquitous, trusted and rewarded by law."[9] In the current legal frameworks of both the U.S. and the EU, "personal data" in the EU or "PII" in the U.S. operates as a legal trigger; as soon as data becomes personally identifiable, the full panoply of legal obligations and restrictions applies to it.[10] Accordingly, organizations around the world have structured their internal and external privacy policies around variations of "PII" data – and its converse, de-identified data – locking themselves further into a binary regime.

As Eloise Gratton has previously described, a "literal interpretation of the definition of *personal information* and of the term 'identifiable' has in many instances either an over-inclusive outcome, an under-inclusive outcome, or may trigger uncertainty as to which kind of information is in fact "identifiable."[11] While the definition of "personal information" is traditionally intended to be broad, when *any* data may "trigger a system in which organizations and industry players will incur additional costs for complying with [data protection laws]," unintended results arise.[12] In order to provide required privacy disclosures and gather consent for the collection of identifi*able* information, for example, organizations may be paradoxically forced to actually identify those individuals. Similarly, "it may be difficult for an organization collecting new types of data to grant access if this data has not even been processed."[13]

Moreover, "it is not always clear at what point a piece of data can be said to be *identifying* an individual"[14] and the legal uncertainty arising from this state of affairs has proven problematic. If neither organizations nor regulators can say whether particular data points are personal, compliance with data protection laws will continue to be suboptimal. There remains significant debate across and within multiple jurisdictions about how identifiability should even be measured, including "whether illegal means that may be used to identify an individual should be considered"; what kinds of costs and resources should be used by an organization to determine if certain data can 'identify' an individual and is therefore covered under the definition"; and "whether information should be evaluated alone or in correlation with other information available when attempting to determine if this information is 'identifiable.'"[15] In addition to shifting legal standards for identifiability, organizations must grapple with changing technologies and information-sharing practices that may increase the likelihood of being able to link data to an identified individual. [16]

---

[7] See, e.g., Khaled and Cecilia, http://idpl.oxfordjournals.org/content/early/2014/12/12/idpl.ipu033.full.pdf?keytype=ref&ijkey=K8xdZaj1rw3EzDx

[8] http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf (The opinion analyzes the robustness of three de-identification techniques on the basis of (i) is it still possible to single out an individual, (ii) is it still possible to link records relating to an individual, and (iii) can information be inferred concerning an individual?) p3.

[9] http://ico.org.uk/about_us/research/~/media/documents/anonymisation_seminar/ohm_slideshow.ashx

[10] Solove & Schwartz, PII 2.0 (and BNA follow-up); Gratton, *If Personal Information Is Privacy's Gatekeeper*

[11] Gratton 115 art.

[12] Gratton at 119

[13] Gratton at 119.

[14] Gratton at 124, siting Bercic & George.

[15] Gratton at 126, 128, 134

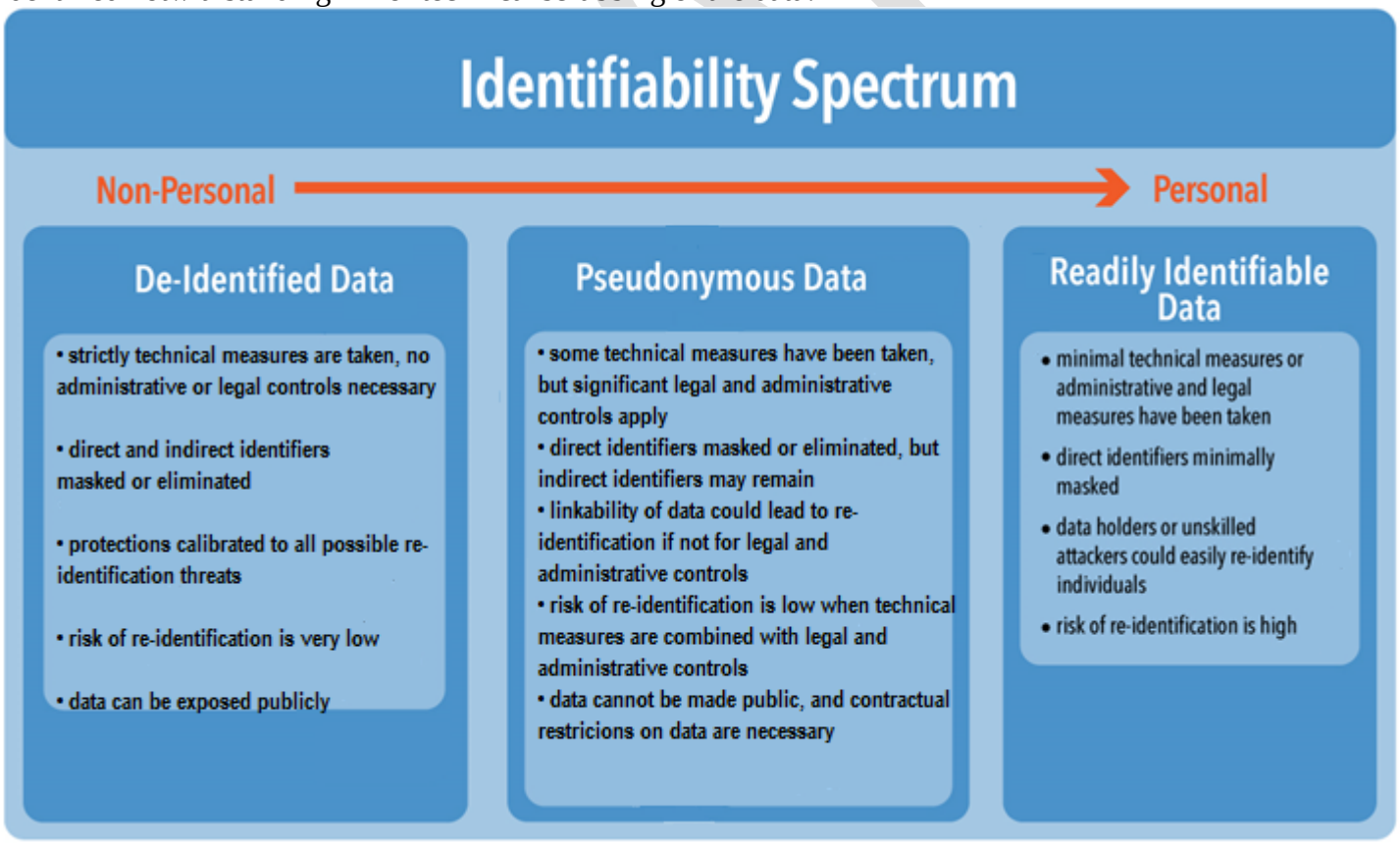
[16] Solove & Schwartz 11-23-12 at 4

Leading scholars have suggested that reforming the current framework to encompass a wider spectrum of data, a so-called "PII 2.0" continuum that instead categorizes data as "identified, identifiable, or non-identifiable."[17] Depending on the location of data on this spectrum, different legal requirements would apply. Despite wide agreement that this movement away from a binary definition of PII was an accurate reflection of practical reality, policymakers have yet to respond by reassessing current law. While this paper later seeks to demonstrate that, in practice, government and industry standards have started to reflect this sliding scale of identifiability, first it seeks to further develop the stages of this identifiability spectrum.

A New De-Identification Taxonomy

While the rhetorical debate surrounding de-identification rages, organizations continue to employ a range of techniques to de-identify, obscure, and share their data. As these methods offer widely varying levels of protection and obscurity, depending on the context of their use, they have too often become square pegs forced into the round, all-or-nothing holes of the current PII framework. In order to help advance the debate of when data should be properly termed "de-identified," we have examined this wider range of practices and reclassified data on a spectrum according to its state of identification. A more nuanced understanding of how organizations are protecting their data on the ground will help the entire de-identification community better assess and respond to privacy risks.

The following chart offers the de-identification debate a new taxonomy, categorizing data along a spectrum of identifiability. As described in detail immediately below, data that has been technically manipulated is classified as "de-identified" if it has been rendered non-linkable through the use of technical controls, "pseudonymous" if it is linkable but protected by legal and administrative controls, and "readily identifiable" if it can be easily re-identified notwithstanding minor technical scrubbing of the data.



[17] Id. But see infra Part II.

De-Identified Data

At the furthest end of the identifiability spectrum, where data pose the least privacy risk and are considered least personal, are de-identified data. These are data with such low, near-zero privacy risk that they can be shared freely and publicly. In order to create data suitable for public release, statisticians and data scientists seek to apply sophisticated statistical, encryption and other mathematical processes to data sets in order to achieve permanent, impenetrable de-identification.[18] Both direct and indirect or quasi-identifiers are modified in order to protect against future re-identification attacks, and legal and administrative controls are not available. Because these data are to be released publicly, technical protections are the data's last and only line of defense against re-identification attempts. While we leave for others the debate as to whether it is ever appropriate to describe data as "anonymous," it is our opinion for now that the only subset of de-identified data that should be described as "anonymous" is this one, where appropriately applied technical measures render data permanently unlinkable.

De-identification requires an inclusive view of privacy and re-identification risk. Accordingly, these de-identified data are intended to withstand any potential re-identification threats, from world-class external attackers to malicious insiders with existing knowledge of the individuals. When evaluating the effectiveness of a particular de-identification effort, researchers determine whether attackers can identify information about individuals in a certain data set with *any* certainty. If even minimal re-identification is possible, the data is generally not considered to have been acceptably de-identified.[19] This inquiry is enormously useful in moving the science of de-identification forward.

However, in many cases achieving the near-zero re-identification risk sought by technical proponents renders data unusable for the purposes it was gathered.[20] Furthermore, limited guidance is available to organizations to determine what level of re-identification risk is appropriate given the value of a particular data set. Because the risks are different, applying the strictest available technical tools has costs if applied in the same way to ordinary business purposes and data used for research on sensitive health topics. An overly strict standard will decrease utility and increase expenses. Critics of some proposed uses of such data may not be concerned about such negative impact, but similar analysis could leave important databases needed for research inaccessible to scientists.[21] Inflexible confidentiality or consent rules often intended to limit commercial tracking, such as in the proposed European Data Protection Regulation, may inadvertently undermine the utility of cancer and other disease registries, and the critical health research they facilitate.[22]

A number of technical methodologies hold great promise to both protect data against re-identification and maintain their utility, but these are not yet broadly feasible. Primary among these is differential privacy, which "ensures that the removal or addition of a single database item does not (substantially) affect the outcome of any analysis."[23] Despite its potential, differential privacy mechanisms do not prevent all sensitive disclosures, require substantial infrastructure investments and sophisticated users, and do not provide the granular data that organizations may sometimes require.[24] It has thus have proven difficult for many organizations to employ and has not been widely adopted in practice.[25] However, other mechanisms to enable highly perturbed data continue to be deployed in new contexts, as organizations struggle to maximize both privacy and utility. One recent

[18] Khaled; Sweeney; Wu; Narayanan & Felten; Dwork
[19] Narayanan & Felten
[20] Gellman note FN 10 (Ohm, citing Brickell & Shmatikov)
[21] Barth-Jones; Khaled (p 140 "Distortions to the data that produce results that do not make sense erode the trust of the data analysts in the data and act as barriers to the acceptability of the techniques used to protect the privacy of the data.").
[22] http://www.ejcancer.com/article/S0959-8049(13)00845-9/abstract?cc=y
[23] http://research.microsoft.com/pubs/74339/dwork_tamc.pdf, http://research.microsoft.com/en-us/projects/databaseprivacy/dwork.pdf
[24] See new Scientific American article on differential privacy
[25] http://research.neustar.biz/2014/09/08/differential-privacy-the-basics/

example is Google's application of Randomized Aggregatable Privacy-Preserving Ordinal Response (RAPPOR), a technology that enables anonymously collecting statistics from end-user, client-side software by locally applying privacy-protective randomized response.[26]

Pseudonymous Data

Moving along the identifiability spectrum we find pseudonymous data: data that is neither fully identified nor fully personal, but that has protected against re-identification by both technical and administrative controls. This data may also be described as obscured, as it has been subjected to some technical modification or masking, albeit without the same technical rigor as de-identified data. Importantly, a combination of technical and administrative protections must achieve a sufficiently *low* risk of re-identification. In the EU, this data is considered personal, but may weigh in favor of data processing during the legitimate interests balancing test. In the U.S., if it is protected by sufficient administrative and legal controls (presuming reasonable technical measures exist), it is considered not personal.

Data that are *not* intended for public disclosure do not necessarily need to be subjected to the same stringent technical measures as data in the previous category to ensure re-identification risks remain low. And because these data will not be released to the entire world, technical measures are not the sole line of defense against re-identification: instead, legal and administrative controls can be used to minimize any residual privacy risk. Therefore, organizations may be able to use less data-destructive technical measures, supplemented by stricter administrative and legal controls, to achieve a sufficiently low risk of re-identification.

Pseudonymization spans a wide range of practices, and requires legal and organizational protections (such as contractual promises not to re-identify data or segregating personal and non-personal data) in order to be relied upon. It may include data that is: only temporarily de-identified; masked by weaker technical applications;[27] protected against some credible threats but not others; or that is maintained in a manner that is obscured, but could be easily identified by the holder of the data, if legal or policy commitments are ignored. The critical distinction is that while direct identifiers are removed or manipulated just as they are for de-identified data, pseudonymization does not technically mask indirect identifiers, or relies on administrative measures (which may be bypassed) to de-link indirect identifiers from data subjects. Thus, data is not obviously or easily linked to an individual, but does remain link*able*. Because of this, pseudonymous data should not be released publicly, and additional contractual safeguards may be needed before such data can be shared.

Accordingly, risk assessments for pseudonymous data may be conducted more pragmatically, focusing not on the every possible attack vector but instead on those that are truly feasible or likely to be available to an attacker. This more narrowly risk-based assessment[28] thus considers re-identification risks "in the particular circumstances involved, having regard to such factors as the motives and capacity of the organization or individual to re-identify the information."[29] As well as the practical risk of an attack, other considerations may include the sensitivity of the data and the risk of harm if an attack is successful. By focusing their de-identification efforts on realistic threat models, organizations can more easily employ and assess the technical mechanisms available to them to in order to find the best fit for their particular purposes.

Furthermore, pseudonymous protections can be calibrated on a case-by-case basis to protect against a wide range of likely re-identification threats, both internal and external. Administrative and legal controls are particularly effective against unskilled or opportunistic re-identification, such as by peeping employees or vendors, with contractual agreements providing both deterrents to and punishments for re-identification. Even

[26] https://static.googleusercontent.com/media/research.google.com/en/us/pubs/archive/42852.pdf
[27] Such as constraining names, adding noise, character scrambling, character masking, truncation, or encoding. Khaled 164-66.
[28] Khaled book
[29] Cavoukian & Khaled June 2014

inadvertent re-identification – such as when a researcher accidentally recognizes a family member in a data set – can be protected against through administrative measures, such as automating data processing so that no human interacts with it; utilizing non-persistent data; or shifting data processing to another country, where the risk of accidental recognition is significantly decreased.

By working with a risk-based approach, organizations can be more precise in calibrating the balance between protecting privacy and preserving utility with their data sets. As a result, this has proved particularly fertile ground for the development of new de-identification techniques. One of the most well- known tools for practical technical de-identification is the Privacy Analytics Risk Assessment Tool (PARAT) developed by Dr. Khaled El Emam, which "de-identifies information in a manner that simultaneously minimizes both the risk of re-identification and the degree of distortion to the original database."[30] Another innovative technical approach to de-identification is the Anonos "Dynamic Data Obscurity" method, which "dynamically segments de-identifiers to data stream elements at various stages" causing data to undergo a physical transformation so data is no longer identifiable to third parties without assistance from the user to which data pertains while still preserving full access to all underlying data.[31]

Many organizations use the term 'de-identified' or 'anonymous' for this type of data because it does require credible steps to hide individuals' identities be taken. Pseudonymization techniques might preserve the utility of data better than de-identification techniques, for example, and still render quite low re-identification risk. In addition, explicitly describing this data as personal might limit an organization's ability to use and share it, even for non-commercial or legitimate purposes. There are any number of uses for data that do require the information to be linkable and to include a significant number of historic details, within both the corporate and scientific spheres. For instance, key-coded data, in which personal data "have been stripped of direct identifiers and replaced by a key to avoid unwanted or unintended re-identification" by anyone without the key, only a limited remains indirectly identifiable even when backed by robust administrative safeguards securing and limiting access to the key.[32] The ability to link research data to individual data subjects may be necessary during clinical trials, for example, to enable treatment if a researcher discovers that follow-up medical attention is required.[33] Key-coded data is used extensively in a range of sectors where limited re-identification may become necessary or desirable under special circumstances, including pharmaceutical research, scientific and historical research, marketing analysis, and online and mobile services.[34]

Readily Identifiable Information

Finally, the furthest end of the spectrum contains data that has been only superficially manipulated or released in a manner where it could be readily and easily linked to an individual. Even if such data does not explicitly identify an individual, or some minimal administrative or legal controls are present, if the risk of re-identification is *high* then data should not be considered pseudonymous. Data that are personal in some context (e.g., nine unique digits comprising a social security number) or that are have been linked to an identifier only temporarily may fall within this category, as they may be marginally protected when taken out of context but remain vulnerable to any minimally intensive re-identification effort. In most cases, this data should be considered explicit PII, subject to the full complement of data protection laws.

[30] Cavoukian & Khaled June 2014 (p 13), also http://www.privacyanalytics.ca

[31] Gary LaFever, IAPP Privacy Perspectives article (Oct. 20, 2014) available at https://privacyassociation.org/news/a/what-anonymization-and-the-tsa-have-in-common/ and comments to FTC and Mauritius DPC officials available at http://www.anonos.com/anonos-enabling-bigdata/

[32] http://www.epag-thinktank.eu/docs/whitepapers/EPAG_Whitepaper_Key_Coded_Data.pdf

[33] http://www.cov.com/files/Publication/26174ea1-6641-457f-990c-f874b10f7670/Presentation/PublicationAttachment/350459db-4a73-4ac4-b59a-fa8f354473ee/oid64167.pdf

[34] http://www.epag-thinktank.eu/docs/whitepapers/EPAG_Whitepaper_Key_Coded_Data.pdf (find substitute source)

A Path Forward: Expanding the Framework

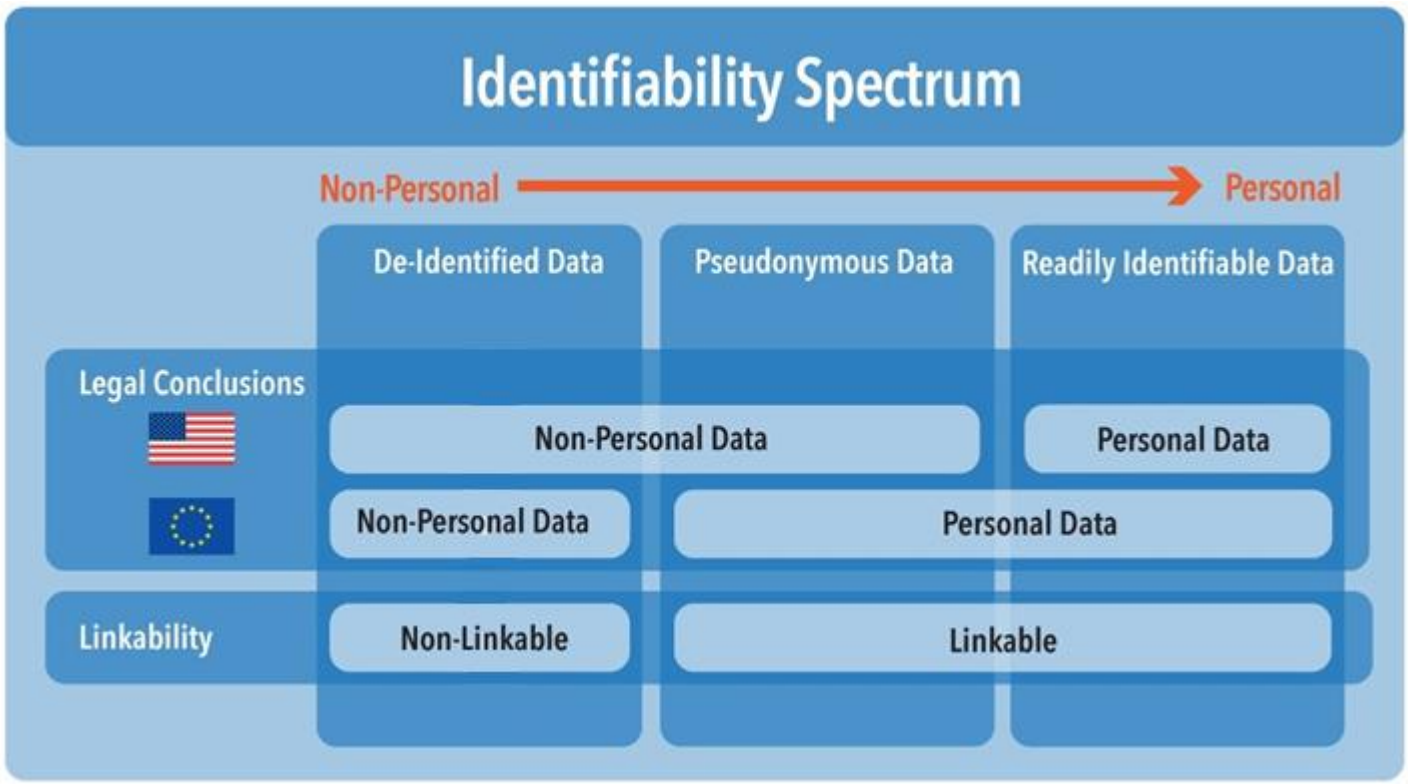
The current binary policy view has led to an impasse where researchers and organizations make increasing uses of that data in formats they consider acceptably and pragmatically de-identified while policymakers seek to ensure that the same data falls within the scope of data protection laws. And yet, as Paul Ohm writes, "[n]o matter how effectively regulators follow the latest re-identification research, folding newly identified data fields into new laws and regulations, researchers will always find more data field types they have not yet covered. The list of potential PII will never stop growing until it includes everything."[35] In order to help bridge the gap, we propose a new framework that will more granularly categorize data along the spectrum of identifiability, allowing organizations to more accurately reflect their efforts to de-identify, pseudonymize, or otherwise protect data and ensuring that policymakers on either side of the de-identification debate no longer simply speak past each other.

As we described above, we believe that there already exists a wider range of obscured or "de-identified" data than is conceived of by the current PII framework. Rather than distort these practices – and the threat and utility models underlying them – by forcing them into either "identified" or "de-identified" pigeonholes, we suggest introducing more flexibility to the PII framework by recognizing a spectrum of de-identification practices.

Accordingly, we propose that "de-identified" terminology be reserved for data that cannot be linked to a particular individual because of the elimination of direct and indirect identifiers through comprehensive technical controls. This data would be considered *non-personal* and exempt from data protection regulations. It would be freely shareable. Next, "pseudonymous" data, in which direct but not indirect identifiers are removed and administrative or technical measures reduce re-identification risk to "low," would be considered personal in some circumstances and non-personal in others. This data could never be made public, and, as described below, would often be subject to some, but not all, data protection requirements. Finally, data in which direct but not indirect identifiers are only nominally masked, administrative or legal controls are negligible or non-existent, and where the risk of re-identification is high would be considered "readily identifiable" and considered *personal information*, subject to all relevant data protection laws.

Mapping the expanded data taxonomy to the PII framework produces the following chart, which describes each data category in terms of its 'linkability' and legal status under U.S. and EU law:

[35] Broken Promises of Privacy, 57 UCLA L. Rev. 1701, 1742 (2010).



The inclusion of pseudonymous data in this spectrum presents our most significant break with the current PII framework. This data is a conundrum in that it often does mask the data subject's identity, but not always reliably enough to establish a guarantee of re-identification against a determined technical attacker. With additional legal and administrative controls, however, the actual risk of identification can be minimal. Subjecting pseudonymous data to the full set of Fair Information Practice Principles, however, could perversely incentivize organizations to maintain the data in a *more* identified manner. For example, an organization that would otherwise maintain data in a pseudonymous state (thus decreasing privacy risk) might be required to re-identify the individual in order to comply with the individual's statutorily mandated access right to the data. In many other cases, including critical medical and social research projects, explicit consent requirements or a restriction on sharing would completely eliminate the utility of the data.

Our framework addresses this dilemma by proposing that pseudonymous data be subject to some, but not all, data protection requirements, in proportion with the risk of re-identification and the nature of the data. Depending on the totality of the circumstances, pseudonymous data could be subject to a range of increasingly intensive legal elements, such as consent, data minimization principles, sharing restrictions, use limitations, or other important privacy protections. Within this framework, data that is more personally identifying or that faces greater re-identification risks (such as location records) would be subject to certain set of privacy-protective standards (such as technically rigorous pseudonymization, or opt-in consent). Data that is less identifiable or that is less likely to be attacked, however, would be held to correspondently less intensive standards (such as opt-out consent, or increased reliance on administrative controls). An organization's evaluation of the risks and benefits of processing data in particular ways may include an obligation to err on the side of privacy over utility as the risk of attack and the risk of harm increase.

Pseudonymous data naturally encompasses a diverse range of data, with correspondingly diverse threat models to be protected against and potential uses to support. Organizations must therefore be careful in assessing the default privacy protections applicable to a particular data set and the fairness of a particular use. In order to ensure that the claimed benefits of a particular use are appropriately weighed against its potential privacy risks,

we have previously proposed that organizations engage in Data Benefit Analysis (DBA).[36] A complement to the traditional Privacy Impact Assessment (PIA), the DBA assesses such variables as the "nature of the benefit, the identity of the beneficiary and the likelihood of success" and feeds the results into existing PIAs in order to craft "a balanced, comprehensive view of big data risks and rewards." This process recognizes that, in some circumstances, a small amount of privacy risk may be worth accepting if the ultimate result will lead to much larger benefits. A DBA allows organizations to rationally measure the potential benefits to consumers and society at large that will arise out of using personal data in a particular way, in order to then determine whether those benefits outweigh the privacy risks also arising from it.[37]

It is important to note that re-identification risk is not determined solely on the type of controls utilized, or the type of data to be protected, but rather by a combination of case-specific factors and threats. There may be circumstances in which non-sensitive data is adequately protected by pseudonymization, just as there may be circumstances when more comprehensive de-identification measures are required. However, by making it more clear externally by what criteria data is considered de-identified,[38] organizations will be able to debate the sufficiency of their measures on fair ground. Rather than engage in a false debate about whether data is capable of being re-identified under any circumstances, discussions can then turn to whether the data has been appropriately de-identified based on the risk of re-identification, the safeguards in place, and the potential benefits of preserving a particular amount of utility in the data.

Relevance to the FTC Framing of PII

As Big Data and increasingly interconnected technologies continue to strain existing privacy norms, legislators and regulators have already begun exploring how to re-draw the lines between personal and non-personal. In the U.S., the FTC has acknowledged the broad consensus that "the traditional distinction between PII and non-PII has blurred and that it is appropriate to more comprehensively examine data to determine the data's privacy implications."[39] In order to address such concerns while still imposing some practical limits, the agency crafted a new PII standard, considering data personal when they are "reasonably linkable" to a particular consumer or device. At the same time, the FTC described three steps organizations can take to minimize such linkability, and thus their liability. Accordingly, the FTC considers data to be *not* "reasonably linkable," or de-identified, if an organization 1) takes reasonable measures to ensure that the data is de-identified, 2) commits publicly to maintaining and using the data in a de-identified fashioned, and 3) contractually prohibits downstream recipients of the data from attempting to re-identify it.

While the FTC's definition nominally still creates a linkable/non-linkable binary, it nevertheless captures many of the same factors embodied in our proposed categorization. Rather than holding all anonymous data to the highest technical standard, the FTC's approach recognizes the importance of administrative and legal protections when used in combination with reasonable, technical de-identification measures. The FTC's approach also acknowledges the significance of contextual factors in determining re-identification risk, noting that "what qualifies as a reasonable level of justified confidence depends on the particular circumstances, including the available methods and technologies. In addition, the nature of the data at issue and the purposes for which it will be used are also relevant."[40] Again, this recognition that different types of data may be more

[36] http://www.futureofprivacy.org/wp-content/uploads/FPF_DataBenefitAnalysis_FINAL.pdf

[37] As the Article 29 Working Party noted, "the very nature of the right to the protection of personal data and the right to privacy . . . are considered relative, or qualified, human rights. These types of rights must always be interpreted in context. Subject to appropriate safeguards, they can be balanced against the rights of others." Opinion 06/2014 on the notion of legitimate interests of the data controller under Article 7 of Directive 95/46/EC, *available at* http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp217_en.pdf.

[38] E.g., X set of sensitive data is being considered de-identified because it has been subjected to stringent perturbations and some administrative restrictions, while Y set of data is being considered de-identified because it is non-sensitive, direct and quasi-identifiers have been masked, and substantial legal and administrative controls exist on its use and sharing.

[39] Era of Rapid Change (2012)

[40] FTC Era of Rapid Change at 21

sensitive, or more at risk, or otherwise merit different levels of protection – particularly in the liminal zone between "reasonably linkable" and "non-linkable" – directly parallels our taxonomic approach.

One aspect of the FTC's definition of "reasonably linkable" requiring further discussion is the inclusion of data that identify devices rather than identifiers. The agency's current guidance simply indicates that data is personal to the extent it identifies "an individual *or device*," without further elaboration as to why or when any given device is assumed to belong to an identifiable individual.[41] Certainly many devices are used by only one individual, who may be readily identifiable, but not all. Some devices may be shared equally between two people, or ten, or more. In today's Internet of Things environment, many devices maybe part of an ecosystem and have no connection to any individual. Identifiers, including those that identify devices rather than users, come in a wide variety of formats and features; whether any particular device can be reasonably traced to an individual requires a case-by-case assessment, not a blanket assumption that all devices as personal. For example, some identifiers have look-up databases in the hands of the organization holding the data, while others have public look-up databases, creating two different levels of re-identification risk. Another identifier might be easily cleared by its users, although others could be hard-coded, or only clearable by resetting the entire device. And yet another identifier might only be used locally, while others are shared globally. All of these factors need to be accounted for in assessing the risk of re-identification

Despite the FTC's broad definition of personal data as incorporating reasonably identifiable devices, their enforcement actions reflect a more nuanced recognition that some identifiers create more privacy risk than others. For example, in its consent decree with Myspace, the agency emphasized that privacy concerns arose from the sharing of quasi-identifiers when there existed a publicly available look-up directory, which could be used to actually re-identify an individual, rather than from the transmission of the pseudonymous device identifier itself. This more nuanced and risk-based application of privacy rules to quasi-identifiers is also reflected in some U.S. courtrooms. For instance, a California court examined Hulu's unique User IDs "very practically" and determined a unique quasi-identifier, without more, is not PII under the VPPA. When that court examined Hulu's use of the Facebook "Like" button, which transmitted cookies containing unique Facebook IDs, however, it did find a violation of the VPPA, as "the link between the user and the video was more obvious." Given the wide variation that can exist in identifiers, treating all "device identifiers" the same under the letter of the law does not provide a useful framework for organizations, nor does it take into account the particularized risk of identification arising from each identifier.

While we agree that persistent or universal pseudonyms should generally be subject to more robust set of privacy protections than truly un-linkable data, we do not agree that they should be considered *per se* linkable, or fully personal. Instead, identifiers – whether they arise from a device or not – that are pseudonymous, subject to administrative controls and appropriate consumer protections, could be considered de-identified data in many cases. The more globally unique an identifier is, or the more clearly individual it is, or the more parties that can access it, the more private it should be treated. Identifiers that are only readable by one organization, or that are controllable by a consumer, or that can be shared between consumers, on the other hand, could reasonably be considered less private and subject to less stringent protections.[42]

Relevance to the EU Framing of PII

In Europe, too, the traditional definition of PII may soon see an official shift. The current European Data Protection Directive regulates information relating to a natural person who is "identified or identifiable,"[43] taking an inclusive approach in extending data protections to its people and their quasi-identifiers. Currently, only data that has been irreversibly de-identified and protected against "all the means likely reasonably to be used" by either the data controller or a third party can be considered de-identified.[44] However, European

[41] Id, emphasis added.
[42] See also https://www.cdt.org/files/pdfs/CDT-Pseudonymous-Data-DPR.pdf
[43] http://ec.europa.eu/justice/policies/privacy/docs/wpdocs/2007/wp136_en.pdf
[44] http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf

regulators, courts, and advisory groups have struggled with the concept of "identifiable" data for years.[45] Over the course of three years, for example, five French courts and the CNIL took contradictory positions on whether or not IP addresses were personal information.[46] European courts have not yet reached a consensus on how to determine if a particular type of data was personal.[47] Despite these struggles, it remains possible that pseudonymous data could appear in the final Data Protection Regulation in some form. Given the importance of pseudonymized data to a wide range of societally and individually useful purposes, it will be important for the final GDPR to include a definition of pseudonymous data and appropriate measures of protections and permissions when data is pseudonymized.

Pseudonymous data of the type we describe above, while protected by both technical and administrative controls, would still be considered identifi*able* and thus personal for the purposes of EU law. However, in keeping with our proposed framework, pseudonymous data could be a key factor to consider in assessing how EU data protection obligations apply. For example, we would propose that pseudonymous data could carry a rebuttable presumption that data processing is legitimate and that pseudonymization could be deemed a compatible use with regards to the purpose limitation principle. As existing EU data protection laws set the bar to de-identification higher than many countries, adding a "middle ground" to EU data protection laws would encourage not only more useful research, it would encourage the adoption of more reliable (albeit not technically perfect) technical and administrative controls.

Other rules to incentivize the creation and use of pseudonymous data have already been proposed both by EU member delegations and interested policy organizations.[48] As we have suggested previously, we believe that "pseudonymization should excuse controllers from certain obligations under the GDPR, such as obtaining explicit data subject consent or providing rights of access and rectification."[49] Others have similarly urged that "a general requirement that consent be 'explicit' is reasonable, but that for some categories of data, the 'legitimate interest' justification paired with a robust right to refuse processing is appropriate."[50] Previous drafts of the GDPR text have also suggested that controllers could be rewarded for utilizing pseudonymous data, as such processing could be presumed not to significantly affect the interests, rights or freedoms of the data subject.[51] One of the most repeated recommendations has been that "for unauthenticated pseudonymous data sets, it also be reasonable to excuse data controllers from obligations such as access rights and data portability."[52]

Furthermore, the legitimate interest analysis under Article 7 of the existing Data Protection Directive harmonizes with the risk analysis underlying both the U.S. approach to the use of pseudonymous data and our proposed framework above. The legitimate interests balancing test represents a fundamental recognition that privacy interests and data utility should be weighed together in certain circumstances.[53] The Article 29 Working Party has made clear that the application of appropriate measures "could, in some situations, help 'tip the balance'" in favor of the data controller's legitimate interests. Pseudonymization and personal data that are "less directly and less readily identifiable" are specifically mentioned as one such "less risky form[] of personal data processing," wherein the general likelihood of "data subjects' interests or fundamental rights and freedoms being interfered with is reduced."[54]

[45] Gratton p 126
[46] Id p 126
[47] Id 127, conflicting case law
[48] http://www.futureofprivacy.org/wp-content/uploads/FINAL-Future-of-Privacy-Forum-White-Paper-on-De-Id-January-201311.pdf, CDT paper, leaked 2013 draft
[49] http://www.futureofprivacy.org/wp-content/uploads/FINAL-Future-of-Privacy-Forum-White-Paper-on-De-Id-January-201311.pdf
[50] CDT paper
[51] http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//TEXT+TA+P7-TA-2014-0212+0+DOC+XML+V0//EN
[52] CDT, but also draft *id* and Jules-Omer paper.
[53] WP29 on legitimate interest processing
[54] Id. At 43

In order to further incentivize the use of pseudonymous data in Europe, we would urge policymakers to extend the legitimate interests test to not only personal and pseudonymous Article 7 data, but also to appropriately safeguarded and pseudonymized special categories of data under Article 8. Under this approach, if data of any sensitivity could be shown to have been subjected to credible pseudonymization, with sufficiently low risk of re-identification, then it would be permitted to undergo the balancing test. This is not at all to say that all uses of pseudonymized sensitive data would automatically be deemed legitimate, simply that there would be an opportunity to assess and balance "the legitimate interests of the controller, or any third parties to whom the data are disclosed, against the fundamental rights of the data subject."[55] The processing of pseudonymized health data for healthcare research or healthcare device maintenance, for example, may result in a different balance of interests than the processing of that same data for marketing purposes. That some uses of pseudonymized sensitive data may not pass the legitimate interest test does not mean that other, more compelling uses for that data should be thrown out like the proverbial baby in the bathwater.

It will be essential for the EU to have a pseudonymous category that allows the use of such data, or many standard and essential data uses will exist in an area of uncertainty. As we explain in the next section, organizations operating in Europe already operate under such uncertainty, finding their uses and protections for pseudonymous data subject to the views of individual DPAs.

PII 2.0 in Practice

While proponents of PII 2.0 models have urged policymakers to officially expand the PII framework, these efforts have yet to be reflected in the laws. Re-categorizing data may increase the precision of the de-identification debate, but without a framework that addresses the middle ground (i.e., pseudonymous data), de-identification in practice will remain strained. We argue, however, that the gap in real world application of these concepts is less than it appears on paper.

Outside the de-identification debate, we are already beginning to see an implicit recognition that certain states of data warrant different protections. As this approach gains broader recognition, we hope that it will provide a model for organizations and regulators to begin explicitly recognizing intermediate data states and assigning them tailored protections, further clarifying the divide between intermediate, identified and de-identified data. For these intermediate data sets, it must be clear that privacy restrictions do apply. In some cases consent may be required, or retention policies or commitments not to use data in certain discriminatory ways, or even notice or certain limited access. While regulators have been opaque in their application of such distinctions so far, the underlying logic to various recent decisions reflects this approach.

Indeed this has often been the case in the application of EU law, where regulators describe data sets such as web logs as personal, but then recognized the need for certain protections but not others. For example, German data protection authorities in Hamburg passed a resolution in 2009 that made the analysis of user behavior, based on the personal linkage of these data by using their full IP address, only permissible with the user's deliberate and explicit consent.[56] Most web analytics services, which gather such information as a matter of course, did not have practices in place to gather such consent, violating the new law. Rather than oust the service entirely, the DPA instead entered into a binding resolution with Google in 2011 implementing certain – but not all – of the law's protection measures. These included allowing users to opt-out, allowing website operators to request that IP addresses collected be 'anonymized' (by deleting the last digits) and requiring data processing agreements between Google and website operators using its Analytics. Website operators were also required to inform users about the use of Analytics in their privacy policies, including notice of the opt-out, and to delete data collected using previous, non-compliant analytics profiles.

[55] Id. At 3.
[56] http://www.iitr.us/publications/20-hamburg-data-protection-authority-data-protection-conforming-use-of-google-analytics.html

In Canada, the risk-based model is likewise in sync with this argument. Applying certain privacy protections to certain sets of data is perhaps most obvious in the increased obligations applied to sensitive data, as compared to non-sensitive data. For example, in January 2014, the Office of the Privacy Commissioner entered into a settlement with Google regarding retargeting of advertisements based on an individual's health-related searches. Canadian privacy law generally considers information collected for the purpose of online behavioral advertising to be personal information, and requires only implied consent from the consumer (an opt-out); however, sensitive information is treated differently, and requires express consent (an opt-in). In response to the complaint, Google agreed to increase its oversight of advertisers' remarketing campaigns and use of sensitive data.

Similarly, in the U.S. self-regulatory models for behavioral advertising and the Mobile Location Analytics (MLA) Code developed by the Future of Privacy Forum already bind a number of organizations to this approach. The National Advertising Initiative (NAI) Code of Conduct, for example, sets obligations for notice, choice, opt-out, and non-discrimination on data sets that are defined as "non-personal" – that is, neither anonymous nor personal – by their codes.[57] Sensitive data also require opt-in consent when used for interest-based advertising. The DAA Self-Regulatory Principles also set protections for pseudonymous identifiers, determining that "data is not considered PII under the Principles if the data is not used in an identifiable manner."[58] Here, an IP address is *not* PII when collected in isolation (and thus does not require consent or transparency when used for online behavioral advertising), but it *is* PII subject to the full set of Principles when it is "in fact linked to an individual in its collection and use."[59] The MLA Code, on the other hand, requires its organizations to provide in-store notice, to hash mobile device ID MAC addresses and to set discrimination and retention limits around a non-personal but not de-identified set of "de-personalized" data.

If one examines their behavior, rather than their rhetoric, then, it appears that policymakers have long accepted a more sliding-scale understanding of PII, considering some intermediate state data personal but not requiring all of the elements of the law be applied to it. Similarly, even as organizations continue to claim that intermediate state data is de-identified or non-personal, they apply significant privacy requirements to it, including notice, choice, anti-discrimination provisions, retention limits, and more.

Critical Factors for a Successful De-Identification Framework

In order for structural changes to the de-identification and PII framework to be meaningful, the de-identification debate needs to be grounded in more nuanced terminology and de-identification practices need to be more transparent.

Transparency

Much of the current de-identification debate has been dedicated to strawmen, with both sides talking past one another about what is or is not de-identification in what has become a zero-sum discussion. In order to advance de-identification policy – and earn consumers' and regulators' trust – organizations need to be more transparent about what data they maintain, how it is used, how it is protected and what threats it is protected against. Critics and concerned consumers will not be satisfied with vague promises of "anonymity."

If data is claimed to be anonymous or de-identified, organizations should make clear by what standard they have made that determination, to aid others in understanding both the possible utility of the data and the possible threats to it. While security and trade secret rationales may prevent organizations from disclosing the exact details of their technical safeguards or administrative and contractual requirements, organizations could still describe the types of protections they have instituted. Within our above framework, an organization that

[57] (although many participating organizations claim anonymity for such data)
[58] https://www.aboutads.info/resource/download/seven-principles-07-01-09.pdf
[59] Id.

describes its data as "pseudonymous" should be able to inform consumers if or when, for example, it key-codes their personal information, trains its employees on privacy and security, or relies on contractual agreements to prevent onward sharing of data. Another organization, describing its data as "de-identified" and suitable for public use, could describe to consumers in plain language that it has utilized highly technical measures to remove or perturb both direct and indirect identifiers so that they are no longer linkable to the data, but that no further administrative controls have been utilized.

While organizations should be transparent about the method of de-identification and administrative controls they have in place, those must also be backed up by legal force. Increased transparency, in the form of public statements or representations about if and how data has been de-identified, also creates accountability. In the U.S., public promises regarding privacy are enforceable by the Federal Trade Commission under Section 5 of the FTC Act, while in Europe national Data Protection Authorities can and will continue to investigate and enforce such notices under the DPD and the proposed GDPR. Without increased transparency, industry standards and best practice guidance will not develop and debates about what de-identification is will stop us from realizing what it could be.

In order to build a framework in which de-identification encompasses both the technical frontier of de-identification science and more pragmatic technical measures, administrative controls and commitments not to re-identify data must be strictly enforced. And in order for pseudonymous data to be accepted, there must be a shared understanding of what administrative or legal measures are and should be buttressing data which are linkable, but not linked, to individuals, so that they can be utilized productively without undue re-identification risk. To help accomplish these goals, it is important that measurable standards be adopted so that de-identification and pseudonymization practices can be assessed and meaningful certifications can be published.

Terminology

As we discussed before, there are a range of reasons why the current terminology has been overused, and why organizations have stretched "anonymous" to cover a wider range of practices. We propose that, by a consistent terminology to describe how data is de-identified and protected, organizations can be more transparent about what they are doing; researchers can more accurately evaluate real-world re-identification risks; and regulators can better tailor their activities and guidance to strike the correct balance between protecting privacy and preserving the utility of data. Not only must data be categorized consistently, those categorizations must be able to accommodate new discoveries and advances by data scientists and re-identification specialists. In keeping with this approach, we propose recognizing a spectrum of de-identified data; pseudonymous data; and readily identifiable data.

Before we can turn to the debates of real importance in de-identification around the efficacy of de-identification data sets that have had rigorous de-identification methodologies applied, we must first cut through the confusion surrounding it. Both sides of the debate should work together to increase transparency around what efforts organizations can and do undertake to protect data and what protections are really being utilized. This need to find common terminology will not be an isolated event, accomplished once and then set aside. What we mean by technical de-identification and pseudonymization will necessarily change over time, as the techniques that are capable of de-identifying, or even re-identifying data, will continue to evolve. The terminology should stay the same, but the specifics of how such techniques are applied will change with technological advances. Consistency and clarity between stakeholders as to what they mean when they say 'de-identification' will be critical in ensuring that de-identification policy can continue to progress.

Next Steps

It is clear that we need a fundamental shift in public de-identification policy debates, both in terminology and approach. Rather than continue to debate whether data is or is not personal or de-identified, discussion and

policies must move towards a more nuanced approach to data that embraces the reality of varying risks of re-identification as well as acknowledging the risks and benefits of different uses of data. Data can exist in different states, subject to different threat models and better suited to different sets of protections or subject to different risks.[60] By re-examining terms like "de-identified" and "pseudonymous," and using them only in very clear and defensible circumstances, we believe that appropriate protections could be found to preserve both utility and privacy along the whole spectrum of personal information.

Once we have left the binary identified/de-identified model, we can begin deciding what the rules should be for pseudonymous data and building tools to support them. Data that is intended to for public release, for example, requires the strongest technical de-identification measures, as it "provides the sole line of defense protecting individual privacy."[61] Depending on what the data is to be used for, this may mean applying differential privacy tools to add noise to the dataset; scrubbing or aggregating certain fields; or hashing, salting or key-coding inputs and imposing additional administrative safeguards to buttress those techniques.[62] For data intended only for internal use, or more limited sharing, less invasive de-identification techniques may suffice when buttressed by administrative and legal controls, with those controls becoming more comprehensive as the risk of re-identification rises. Standard controls may include robust data security and use policies; access limits; employee training; data segregation guidelines; data deletion policies; individual access and correction rights; contractual limits on third parties' access, use, and sharing of data; penalties for contractual breaches; or auditing rights on service providers or business associates. A full examination of such controls is beyond the scope of this paper, but for additional details a technical paper is forthcoming.

Conclusion

Currently, a legacy legal structure is straightjacketing policy in this area by insisting on a binary identified/de-identified categorization of data and all-or-nothing privacy protections. This binary categorization has found its way into some legal models, including proposals for the next generation of privacy and regulatory oversight under the draft European General Data Protection Regulation.[63] In its place, we need to develop new models recognizing a spectrum of data states with varying restrictions and protections based on the actual utility and threat risks to that data. To do this, we must first recognize the full spectrum of data, from identified to de-identified and everything in between, as it already exists in practice. Only once we have reframed the debate can we develop legislative models that reflect the relevant choices and protections that should attach when data is less than personal.

[60] Swire, Practical Obscurity and Practical De-Identification: A Typology of Levels of De-Identification (forthcoming)
[61] Lagos & Polonetsky, Public vs. Nonpublic Data
[62] Omer & Jules, Seeing the whole spectrum paper
[63] Should the proposed GDPR adopt a pure binary requirement, it risks getting ahead of this important debate.