# RE-IDENTIFICATION OF "ANONYMIZED DATA"

Boris Lubarsky*

CITE AS: 1 GEO. L. TECH. REV. 202 (2017)
https://perma.cc/86RR-JUFT



## I. INTRODUCTION

Today, almost everything about our lives is digitally recorded and stored somewhere. Every interaction with technology creates data about that user. Each credit card purchase, medical diagnosis, Google search, Facebook post, or Netflix preferences is another recorded data point about that individual user. Beyond that, every census report, home purchase, voter registration, medical history, and cell phone geolocation is recorded and stored. This data is then analyzed and used by the entities that collect it. Netflix analyzes user preferences to recommend movies; medical researchers study patient data to find new treatments and cures; and Google reviews search queries to improve its search results. That aggregated data is also sold and transmitted to third parties, such as analytics companies, marketing companies, or commercial data brokers.

Currently there are some legal protections that aim to prevent the disclosure or sale of personally identifiable information, such as name, Social Security numbers, and medical conditions when data is sold or transmitted.

* GLTR Staff Member; Georgetown Law, J.D. expected 2018; Georgetown University, B.S.F.S. 2011. 

However, if that data is scrubbed, of a small category of personally identifiable information it can be considered "anonymized" data.[1] There is no regulation of "anonymized" data: it can be sold to anyone and used for any purposes. The theory is that once the data has been scrubbed, it cannot be used to identify an individual person and is therefore safe for sale, analysis, and use.[2]

The proliferation of publicly available information online, combined with increasingly powerful computer hardware, has made it possible to re-identify "anonymized" data. This means scrubbed data can now be traced back to the individual user to whom it relates. Scrubbed data is commonly re-identified by combining two or more sets of data to find the same user in both. This combined information often reveals directly identifying information about an individual. Re-identification of anonymized data has grave privacy and policy implications as regulators, businesses, and consumers struggle to define privacy in the modern permanently-recorded age.

## II. LEVELS OF IDENTIFIABILITY

Personal data exists on a spectrum of identifiability. Think of a staircase. At the top is data that can directly identify an individual: a name, phone number, or Social Security number. These data are collectively called "direct identifiers."

The second step down is data that can be indirectly, yet unambiguously, linked to an individual. Only a very small amount of data is needed to uniquely identify an individual – 63% of the population can be uniquely identified by the combination of their gender, date of birth, and zip code alone.[3] These data are collectively called "indirect identifiers."

The third step down the staircase is data that can be ambiguously connected to multiple people – physical measurements, restaurant preferences, or individuals' favorite movies. The fourth step down is data that cannot be linked to any specific person – aggregated census data, or broad survey results. Finally, there is data that is not directly related to individuals at all: weather reports and geographic data.

In this way, information that is more difficult to relate to an individual is placed lower on the staircase. However, as data becomes more and more

---

[1] Paul Ohm, *Broken Promises: Responding to the Surprising Failure of Anonymization*, 57 UCLA L. REV. 1701, 1754 (2010).

[2] *Id*. at 1755.

[3] Philippe Golle, *Revisiting the Uniqueness of Simple Demographics in the US Population*, 5 ACM WORKSHOP ON PRIVACY ELEC. SOC'Y 77, 78 (2006) (using 2000 census data).

scrubbed of personal information, its usefulness for research and analytics directly decreases. As a result, privacy and utility are on opposite ends of this spectrum – maximum usefulness from the data at the top of the staircase and maximum privacy at the bottom of the staircase. As data gets more and more scrubbed – its usefulness for analysis decreases.

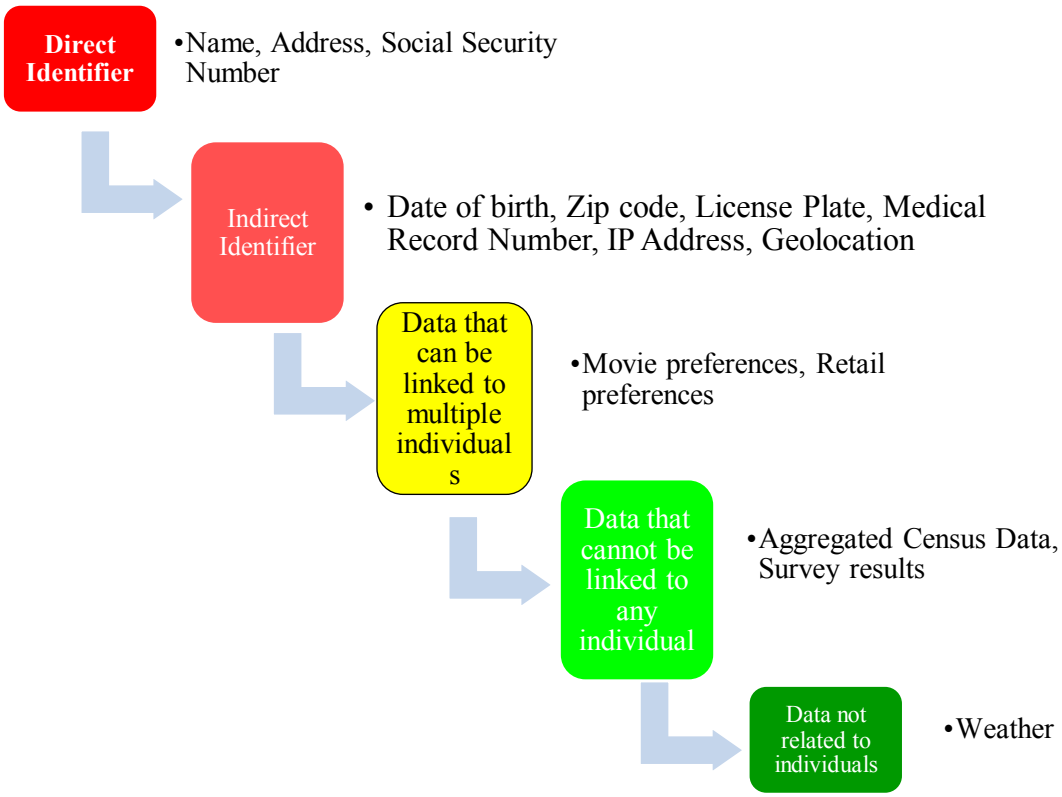


**Figure 1. Examples of Personal Information**

## III. THE FOUR TYPES OF DATA SCRUBBING

The process of scrubbing, or removing, identifying information from data is done by a collection of approaches, tools, and algorithms.[4] There are four main categories of techniques used to scrub data of identifying information – (1) removing data; (2) replacing data with pseudonyms; (3) adding statistical "noise"; and (4) aggregation.

The first two techniques are used mainly for direct identifiers and the latter two are used for indirect identifiers. While direct identifiers can be easily removed or replaced, indirect identifiers can be difficult to remove as they might be important for later analysis or use. For example, a dataset for medical research would be nearly worthless if data scrubbing removed the symptoms or diagnosis.

### *A. Deletion or Redaction*

The first method is to entirely remove or redact any data that directly identifies a person – names, Social Security numbers, etc. This can often be accomplished through automation. In a database with structured data – in a chart that identifies its variable – an entire column of identifying information can be deleted easily and automatically. In the table below, the first column, "Name," can be removed without compromising the usefulness of the data for future research.

| **Name** | **DOB** | **Zip Code** | **Gender** | **Race** | **Diagnosis** |
|---|---|---|---|---|---|
| Adam Smith | 1/1/1970 | 20002 | M | Caucasian | Congestive Heart Failure |
| Betty Davis | 2/2/1980 | 20001 | F | African American | Pneumonia |
| Carlos Hernandez | 3/3/1990 | 20007 | M | Hispanic | Addison's Disease |

**Table 1: Sample Structure Database, Raw Data**

[4] Simson Garfinkel, *De-Identification of Personal Information*, 8053 NAT'L INST. OF STANDARDS & TECH. INTERNAL REP. 1, 6 (2015), http://dx.doi.org/10.6028/NIST.IR.8053 [https://perma.cc/7X86-FBFG].

However, simply redacting or removing data is not foolproof. For example, medical records contain large amounts of unstructured text such as transcriptions of conversations and hand written notations.[5] Direct identifiers might not be clearly marked, and important medical information may be mistaken for personal information and deleted accidentally.[6] If a data set is released with insufficient de-identification, the missed direct or indirect identifiers can be used to re-identify the individual involved.[7]

### *B. Pseudonyms*

| Name | DOB | Zip Code | Gender | Race | Diagnosis |
|---|---|---|---|---|---|
| NAME1 | 1/1/1970 | 20002 | M | Caucasian | Congestive Heart Failure |
| NAME2 | 2/2/1980 | 20001 | F | African American | Pneumonia |
| NAME3 | 3/3/1990 | 20007 | M | Hispanic | Addison's Disease |

**Table 2: Structured Database with pseudonym replacing names.**

The second approach is a process called pseudonymization replacing data with pseudonyms that are either randomly generated or determined by an algorithm. Pseudonymization preserves the usefulness of the data but replaces the identifying information. This approach shares weaknesses with data deletion – direct identifiers can be difficult to identify and replace, and indirect identifiers are inadvertently left in the dataset.

Pseudonymization comes with its own additional weaknesses. If the pseudonyms are not assigned randomly but by a predetermined algorithm, the data can be re-identified. For example, in 2014 the New York City Taxi and Limousine Commission released a dataset of all taxi trips taken in New York City that year.[8] Before releasing the data the Taxi and Limousine Commission attempted to scrub it of identifying information, specifically they

[5] *Id*. at 30.
[6] *Id*.
[7] *Id*.
[8] Anthony Tockar, *Riding with the Stars: Passenger Privacy in the NYC Taxicab Dataset*, NEUSTAR RESEARCH (Sept. 15, 2014), http://research.neustar.biz/author/atockar/ [https://perma.cc/5LZG-YZM8].

pseudonymized the taxi cab medallion numbers and driver's license numbers. Bloggers were, however, able to discover the algorithm used to alter the medallion numbers and then reverse the pseudonymization.[9]

Pseudonyms also cease to be effective if the same unique pseudonym is continually used throughout a dataset, in multiple datasets, or for a long period. Any of these increase the likelihood an individual will by identified by non-direct identifiers associated with that unique pseudonym.[10] Some pseudonymization is meant to be reversible and a "key" is kept to reverse the process. This adds an extra level of security for medical records, for example, but still allows full access to the patient's identifying information. However, as long as a key is retained or the algorithm can be reverse engineered or discovered, pseudonymization can be easily reversed and the data re-identified.

### *C. Statistical Noise*

The third approach to scrubbing datasets of indirect personally identifying information is to introduce statistical "noise." Like pixelating someone's face on TV or in an image, statistical noise allows the viewer to "see" the data, but uses static to obscure the identity of the individuals involved. Static can be introduced in a number of ways such that identifying a specific individual becomes difficult. These include:

- Generalization: Specific values can be reported as a range. For instance, a patient's age can be reported as 70-80 instead of giving a full birthdate.[11]
- Perturbation: Specific values can be randomly adjusted for all patients in a dataset. For example, systematically adding or subtracting the same number of days from when a patient was admitted for care.[12]
- Swapping: Data be exchanged between individual records within a dataset.

As with all scrubbing techniques, the more direct or indirect identifiers are removed, or obscured with static, the less useful the data is for research and analytics.

---

[9] *Id.*; Garfinkel, *supra* note 4, at 17.
[10] Garfinkel, *supra* note 4, at 17.
[11] *Id*. at 20.
[12] *Id*.

### *D. Aggregation*

The fourth scrubbing approach, aggregation, is closely related to statistical noise. Instead of releasing raw data, the dataset is aggregated and only a summary statistic or sub-set is released. For example, a dataset might only provide the total number of patients treated, rather than each patient's individual record. However, if only a small subsample is released, the probability of re-identification increases.[13]

| No. Patients | No. Female |
|---|---|
| 3 | 1 |

**Table 3: Aggregated Patient Data**

In an aggregated dataset, an individual's direct or indirect identifiers are withheld from publication. However, the summary data must be based on a broad enough range of data to not lead to the identification of a specific individual. For instance, in the above example, only one female patient visited the hospital. She would be easier to re-identify than if the data included thirty women who had spent time at the hospital.

As with all four scrubbing techniques, the more direct or indirect data that is removed about an individual, the less useful the data becomes.[14] Data utility and individual privacy are on opposite ends of a spectrum. The more the data is scrubbed the less useful it is. The more indirect, or direct, variables that are in a dataset the more useful it is for analysis but at the cost of individual privacy. As a result there is an incentive for data re-identification – the more specific a data set is the more useful it is for research, marketing, and nefarious purposes. If you re-identify "anonymized" data you have much greater information about a specifically identified person while being outside the current regulatory framework of reporting and data security laws.

## IV. RE-IDENTIFICATION

Data re-identification occurs when personally identifying information is discoverable in scrubbed or so called "anonymized" data. When a scrubbed data set is re-identified, either direct or indirect identifiers become known and

[13] *Id*.
[14] Ohm, *supra* note 1, at 1754.

the individual can be identified. Direct identifiers reveal the real identity of the person involved, while the indirect identifiers will often provide more information about the person's preferences and habits. Scrubbed data can be re-identified through three methods: insufficient de-identification, pseudonym reversal, or combing datasets. These techniques are not mutually exclusive; all three can be used in tandem to re-identify scrubbed data.

### *A. Insufficient De-Identification*

Insufficient de-identification occurs when a direct or indirect identifier inadvertently remains in a data set that is made available to the public. Both structured and unstructured data are prone to re-identification, as inadvertently leaving direct or indirect identifiers can lead to the discovery of a person's identity. Structured data are those that organize the information into tables with identified values. Tables 1-3 above are structured data, as the column containing the name, date, zip code etc. are clearly identified. Unstructured data is basically everything else—it is usually plain text and can be much more variable. Internet searches, doctors' notes, and voice commands are all unstructured data.

#### 1. Insufficiently De-Identified Structured Data

Insufficiently de-identified structured data can occur when indirect identifiers are left in a data set, either inadvertently or for utility purposes. In the mid-1990's, Massachusetts purchased health insurance for state employees and subsequently released records summarizing every state employee's hospital visits.[15] Then-governor of Massachusetts William Weld, assured the public that the data had been properly scrubbed.[16] The fields containing explicit identifiers such as name, address, and Social Security numbers were removed, however, the record still contained almost a hundred unscrubbed attributes per patient that were unscrubbed.[17] Latanya Sweeney, then a graduate student, obtained the data and used the Governor's zip code, birthday, and gender to identify his medical history, diagnosis, and prescriptions.[18]

---

[15] *Id*. at 1719.
[16] Henry T. Greely, *The Uneasy Ethical and Legal Underpinnings of Large-Scale Genomic Biobanks*, 8 ANN. REV. GENOMICS & HUM. GENETICS 343, 352 (2007).
[17] Latanya Sweeney, *Uniqueness of Simple Demographics in the U.S. Population* (Laboratory for Int'l Data Privacy, Working Paper LIDAP-WP4, 2000).
[18] *Id*.

A study showed that 63% of the population can be uniquely identified by the simple combination of their gender, date of birth, and zip code available from census data.[19] As a result of Sweeney's study on re-identification, Congress passed the Health Insurance Portability and Accountability Act (HIPAA) in 1996. HIPAA regulates personally identifiable information in medical records.[20] Its "Safe Harbor" provision specifically established that only the first three digits of a zip code could be reported in scrubbed data.[21]

#### 2. Insufficiently De-Identified Unstructured Data

In 2006 AOL released 20 million search queries for 650,000 users, from three months of data.[22] AOL attempted to scrub the data of any direct or indirect identifiers: it deleted direct identifiers such as usernames and IP addresses. To preserve the data's utility, AOL replaced that information with unique identifying numbers through pseudonymization.[23] Because each user had a unique number, each user's search results could be viewed as a group. Soon after the release, two New York Times reporters were able to track down a sixty-two year-old widow in Georgia by analyzing her AOL searches.[24]

### *B. Pseudonym Reversal*

Pseudonyms are only an effective scrubbing mechanism if they cannot be reversed. There are several ways pseudonymization can be defeated. Some pseudonyms are designed to be reversible and a "key" is kept to reverse the process, however, this precludes their security function. Secondly, the longer the same pseudonym is used for a specific individual, the less secure and easier it is to re-identify that individual. Thirdly, if the method used to assign pseudonyms is discovered or becomes know the data can be re-identified.

#### 1. New York City Taxi and Limousine Data

---

[19] Golle, *supra* note 3, at 78.
[20] 45 C.F.R. § 164.514
[21] 45 C.F.R. § 164.514(b)(2) (2013) (stating that the safe harbor also identified seventeen other specific types of data that would have to be removed before the statute would apply).
[22] Ohm, *supra* note 1, at 1717.
[23] *Id*.
[24] *See* Michael Barbaro & Tom Zeller, Jr., *A Face Is Exposed for AOL Searcher No. 4417749*, N.Y. TIMES (Aug. 9, 2006), http://www.nytimes.com/2006/08/09/technology/09aol.html [https://perma.cc/4297-VE9X].

After the New York Taxi dataset of all taxi rides in 2014 was de-identified by reverse identifying the medallion pseudonym, a data scientist intern at Neustar discovered he could find pictures taken of celebrities entering or leaving taxicabs with the medallion number in the picture.[25] He used indirectly identifying information—medallion number, time, and date—to locate specific rides in the dataset released by the New York City Taxi and Limousine commission. From those 3 indirect identifiers, the intern then used the dataset to identify pick up location, drop of location, amount paid, and even amount tipped.[26]

### *C. Combining or Linking Datasets*

The most powerful tool for re-identifying scrubbed data is combing two datasets that contain the same individual(s) in both sets. Dr. Sweeney was able to re-identify Governor Weld's supposedly "anonymized" set of medical data by linking two databases together. She purchased the voter rolls from Cambridge, where Weld resided, then combined those rolls with the hospital data. Six people in Cambridge shared Weld's birthday, of those, half were men and only one lived in Weld's zip code.[27] In this way she circumvented the scrubbing procedures and re-identified the "anonymized" data.

When two or more anonymized datasets are linked together, they can then be used to unlock other anonymized datasets. Once one piece of data is linked to a person's real identity, that data can then be used to destroy the anonymity of any virtual identity with which that data is associated. The ability to link even supposedly innocuous data exposes people to potential harm because of this.[28]

#### 1. Netflix Prize Data

In 2006, Netflix publicly released one-hundred million records revealing hundreds of thousands of user ratings from 1999 through 2005 and offered a million dollar prize for the first team to significantly improve Netflix's movie recommendation algorithm. [29] Although the data contained no direct

[25] Tockar, *supra* note 9.
[26] *Id*.
[27] Sweeney, *supra* note 18.
[28] Ohm, *supra* note 1, at 1746.
[29] *Id*. at 1720; *The Netflix Prize Rules*, NETFLIX, http://www.netflixprize.com/rules (last visited June 12, 2010) [https://perma.cc/RA6L-LB8B].

identifiers, within weeks of the data's release, two researchers were able to re-identify a subset of specific people by cross-referencing the Netflix data with IMDB.com ratings. Using just six ratings of obscure movies, the researchers re-identified individuals 84% of the time (if they were in both datasets).[30] Including an approximate time of the rating was made allowed identification 99% of the time. Although this only worked to re-identify Netflix users that also had IMDB accounts, the Netflix information could be cross-referenced with social media movie preference found on online dating apps and Facebook for similar results.[31]

## V. CONCLUSION

The current regulatory framework is predicated on the supposition that data that has been scrubbed of direct identifiers is "anonymized" and can be readily sold and disseminated without regulation because, in theory, it cannot be traced back to the individual involved. However, today's techniques of re-identification effectively nullify scrubbing and compromise privacy. The examples of Governor Weld, Netflix, AOL, and NYC taxi illustrate how any data scrubbed of direct personal identifiers can still be readily re-identified if it is combined with another set that also contains data about the same individuals.

Once a dataset is released to the public it can never be strengthened, only 'weakened' by future information that may be released that could lead to that information being re-identified.[32] Re-identification can also be achieved by anyone from government entities, to data brokers, to blackmailers, and is nearly impossible to trace. Once a comprehensive database of previously "anonymized" data is created, it can readily be de-identified. One data broker, InfoUSA, alone claims to have data on 235 million US consumers and uses 29 billion records from over 100 sources to update its database of raw data every year.[33]

The re-identification of anonymized data presents serious policy and privacy implications. For example, this information can be used to bypass password recovery mechanisms for email and bank accounts. Consider that Sarah Palin's email was famously hacked in 2008 when someone guessed her

---

[30] Arvind Narayanan & Vitaly Shmatikov, *Robust De-Anonymization of Large Sparse Datasets*, PROC. 2008 IEEE SYMP. ON SEC. & PRIVACY 111, 121 (2008).
[31] *Id*.
[32] Ohm, *supra* note 1, at 1717.
[33] *Data Quality*, INFOUSA, https://www.infousa.com/data-quality/ (last visited Mar. 29, 2017) [https://perma.cc/ZRZ7-CXT8].

password recovery question that she met her husband at "Wasilla high."[34] Re-identification of scrubbed data can lead to the publication of sensitive or embarrassing information from a person's past that they may not want their employer, spouse, or community to discover. Medical history, sexual preferences and proclivities, reproductive choices, or even details of one's conception can be discovered. Today, that sensitive or embarrassing piece of information more than likely resides on an "anonymized" dataset up for sale. Without regulation of re-identified anonymized data, employers, neighbors, and blackmailers have an unprecedented window into an individual's most private information.

Consider the smartphone. You can unlock it with facial recognition; give it a command or ask it directions; make a credit card payment; and confirm it by swiping your fingerprint. Every one of those interactions is a recorded data point about that individual user: facial print, voice print, finger print, credit card information, and geolocation – on one device. All that information is stored; sent to third parties; reviewed and analyzed; and, after a brief scrubbing, sold on the commercial market.

There is no comprehensive data privacy law in America – it is regulated on an ad-hoc, sector-by-sector basis. None of this patchwork of laws and regulations covers "anonymized" data. There is no duty to report if data has been re-identified. There is no private cause of action for an individual seeking redress for re-identified data, and no external way to verify if a private entity has privately de-identified "anonymized" data exists. The theory that data scrubbed of personally identifying information cannot be re-identified has time and again been shown to no longer hold true. Our current ad-hoc approach is antiquated and inadequate for addressing the new technological challenges re-identified data presents.

[34] *See* Kim Zetter, *Palin E-Mail Hacker Says It Was Easy*, WIRED (July 18, 2008, 10:05 AM), https://www.wired.com/2008/09/palin-e-mail-ha/ [https://perma.cc/HH54-QVWJ].