Data Privacy – A Risk Assessment
Report to the CEO of Newco Industries.

Owen R. Evans

Data Governance 5300 – Final Project

5.5.2021

Data Privacy - A Risk Assessment
Report to the CEO of Newco Industries

1. Introduction and Overview

The following analytical report establishes a set of clear guidelines and affords a comprehensive risk assessment pertaining to the release of our potentially sensitive data to either private or public parties. The release of any data set that is not effectively de-identified to remove personally identifiable information (PII) may subject Newco Industries to potential liability issues that may result in severe fines and ultimately damage our hard-fought brand. The following report thus aims to accomplish the following…

- Provide an overview of the ethical and regulatory issues associated with the release of sensitive datasets.
- Discuss the potential threats to data privacy.
- Determine how and why we should de-identify sensitive data.
- Perform a comprehensive data disclosure risk assessment that will ultimately test the robustness of our de-identification efforts.

2. Ethics & Regulations – Balancing Data Utility and Privacy

The impetus behind the following technical guidance is to provide a responsible framework of policies and methodologies that allow Newco Industries to maximize the utility of our data stores while still sufficiently protecting data privacy. Article 17 of the International Covenant on Civil and Political Rights (ICCPR) has deemed the right to privacy an inherent human right that affords a measure of protection from unlawful and arbitrary intrusions of "privacy, family, home or correspondence." From an ethical perspective, Newco Industries is obligated to ensure protection against the release of sensitive information that may cause undue financial, physical, psychological and/or reputational harm to any of our customers.

Beyond the ethics pertaining to data privacy, Newco Industries would also be subject to certain laws and regulations governing the protection and disclosure of data. In the United States, where Newco conducts most of its business activities, there is unfortunately no unified Federal law governing data privacy. Instead, there are a number of sector-specific laws governing the release of personal information within certain industries. For instance, the Health Insurance Portability and Accountability Act (**HIPAA**) protects personal health information, and the Family Educational Rights and Privacy Act (**FERPA**) protects student education records. In the absence of any defining Federal privacy legislation, many states (in particular California) have drafted more stringent data privacy laws. As a US based business, the chaotic conglomerate of sector specific Federal laws and stringent state acts, might make it difficult to draft for Newco Industries to draft and execute an effective and compliant data privacy policy. The challenge is rooted in desire to

ethically extract as much utility and value from our data stores, without running afoul of current and/or impending privacy legislation.

It is imperative that we draft a data privacy policy largely in line with the current legislative trends. The rise of big data over the last 5 years and the absence of a cohesive Federal data privacy act, there has been an unprecedented amount of wide-sweeping data privacy legislation enacted by various states. As it turns out, many of these state regulations are principally based upon the European-based General Data Protection Regulation (GDPR). The GDPR has been in effect in the EU since 2018 and governs data protection and privacy with stringent requirements and punitive fines for non-compliance. Academics have stated that the GDPR is the most "consequential regulatory development in information policy in a generation", that may serve as the future model for federal data privacy regulation in the United States [i]. For this reason, it is recommended that Newco Industries base their data privacy / security policy specifically to ensure compliance with the GDPR. We should anticipate the fact that as we grow our business beyond the US border, we will be subject to various data privacy regulations across the globe. By basing our data privacy policy on the guidelines brought forth in the GDPR, we will ensure compliance with the most stringent global regulation, anticipate any future changes in US federal law and allow us to conduct business on a global scale.

According to the GDPR, and business that collects or uses personal information, including special category or sensitive information, may process such information only in certain circumstances. This type of data is highly restricted in scope, requires consumer consent and additional security safeguards. In order to be exempt from the GDPR and to allow release of data to third parties, organizations are required to ensure that the data does not contain "any information relating to an identified or identifiable natural person." The GDPR further defines an "identifiable natural person as a person who can directly or indirectly be identified with the help of an identifier such as a name, location, an ID number or an online identifier, as well as physical, physiological, economic, cultural or social features of that natural person". What this means for a data privacy policy based upon the tenets of the GDPR, is that the data stores of Newco Industries must be anonymized, pseudo-anonymized or de-identified prior to use and/or release. These processes, and how to implement, are defined further below and are meant to ensure a high level of protection over personal information in the event of an inadvertent or deliberate data breach.

3. Data De-Identification – Why and How?

As directed by the data privacy tenets in the GDPR, we must ensure that any personal information is removed from the data set prior to use/release. In essence, the use or release of any data set that is effectively de-identified (and has no personally identifiable information) cannot fundamentally violate the privacy of individuals. While this may seem simple in principle, the process of de-identification may be somewhat complex. The complexity arises in the definition of a de-identified dataset. Full anonymization of a data set is a subset of de-

identification that is considered highly secure as it involves complete removal of all direct and indirect personal identifiers. By definition, full anonymization of a data set will involve obfuscation of all useful attributes, rendering the utility of the data set null and restricting innovation. A fully randomized data set for instance will have little to no value.

To address this concern and to foster innovation, the GDPR allows for pseudo-anonymization. This process involves "the processing of personal data in such a way that the data can no longer be attributed to a specific data subject without the use of additional information". Pseudo-anonymization requires the substitution of direct identifiers with artificial values, or pseudonyms, and the allowance of some quasi-identifiers provided that the risk of re-identification is acceptable. This process will allow organizations to maintain compliance with the GDPR while improving the utility of the data set. It provides a sufficient answer to striking a balance between data privacy and data utility.

The increased data utility of pseudo-anonymization does come at a cost, however. In general, the more useful a dataset is (i.e. lower level of deidentification), the higher the risk of re-identification. The inclusion of quasi-identifiers or key variables may inadvertently render a dataset vulnerable to inference attacks, where an intruder can deduce personally sensitive information by comparing the pseudo-anonymized data set with public data sets. The classic example of re-identification via an inference attack occurred in 2016 where the potential release of then Massachusetts Governor Weld's personal information via linkage of a deidentified health data set with voter registration records [[ii]]. This work, carried out by L. Sweeney, indicated that there is sufficient information in the public domain to conclusively identify 87% of the population with only three identifiers – zip code, date of birth and gender. It is clear that data owners and processors need to exercise extreme due diligence prior to the data release of pseudo-anonymized data sets. It is imperative that we follow established Best Practices and exercise effective de-identification via proper masking procedures, followed by a quantitative risk assessment of potential re-identification. Any de-identified data set that exceeds our established metrics for data disclosure risk cannot be released and will need to be further processed to ensure a high level of data privacy.

4. Data Review

The data set in question (Merrimack Survey Results) contains vital and useful demographic, financial and educational data concerning our customers. It has been effectively pseudo-anonymized by replacing names with a randomly assigned identification codes. For the purposes of this risk assessment, we are primarily interested in releasing education & job category information of 100 random customers as a function of age and gender (Table 1). The following risk assessment will provide a quantitative measure of the probability of re-identification for this specific data release. This exercise will determine whether such a data release will pose a threat to data privacy for four separate threat scenarios.

Table 1. First 10 records of the Parsed Merrimack Customer Dataset.

CustomerID	Gender	Age	EducationYears	JobCategory
3964-QJWTRG-NPN	1	20	15	1
0648-AIPJSP-UVM	0	22	17	2
5195-TLUDJE-HVO	1	67	14	2
4459-VLPQUH-3OL	0	23	16	2
8158-SMTQFB-CNO	0	26	16	2
9662-FUSYIM-1IV	0	64	17	3
7432-QKQFJJ-K72	1	52	14	1
8959-RZWRHU-ST8	1	44	16	1
9124-DZALHM-S6I	1	66	12	1
3512-MUWBGY-52X	0	47	11	6

5. Threat Modeling

The risk of re-identification via release of the data set in question will be modeled according to the following four threat scenarios:

- **T1:** The adversary is internal, the access to data is legitimate, and the attack is deliberate. This situation occurs when there is an active attempt by the data owner/process to re-identify individuals within the data set. The probability of this type of an attack depends upon the controls and measures put forth in our data security plan and the motives and skills of the intruder. Because Newco Industries has recently implemented a comprehensive data security plan, the probability of T1 is very low. **Pr(Attempt,T1) is 0.1 (or 10%).**

- **T2:** The adversary is internal, the access to data is legitimate, and the attack is inadvertent. A recipient of a non-public dataset may inadvertently recognize a record in the dataset based upon previous knowledge. The observer is able to perform a reidentification based upon the observation of a quasi-identifier with a substantially high uniqueness. The probability of an inadvertent acquaintance attack is ultimately dependent on the prevalence of the variable in question and the average number of acquaintances (m = 150). For the data set in question, we would utilize the prevalence of a variable at the extremes of our population. In this case, it would be appropriate to utilize education years >22, which would represent those with doctorate degrees (an extreme value for this variable, 1% assumed). The probability of an acquaintance attack is thus given by Pr(acquaintance) = 1-(1-(% of Ph.D's)^m = 1-(1-0.01)^{150} or 0.05. **Pr(Acquaintance Attack, T2) is 0.779 or 78%.**

- **T3:** The adversary is external, there is a data breach, and the nature of the attack is deliberate. This attack is your classic external extruder attack. The HMISS analytics report from 2012 places infers that the probability of an external data breach is 27%. **Pr(Breach Attack, T3) is 0.27 or 27%.**

- **T4:** The adversary is external, there is legitimate access to the data files and the nature of the attack is deliberate. An external adversary with legitimate access infers an attack from a third-party vendor or contractor. A recent research report by the Ponemon Institute has indicated that 59% of organizations in the US had suffered from a third-party data breach in 2017. **Pr(Breach Attack, T4) is 0.59 or 59%.**

6. Data Masking Procedures

The probability of an attack or data breach is shown above and can be partially mitigated with sound data security practices and strict oversight over third party vendors and contractors. However, in the unlikely event of a data breach it is imperative that we process data in a fashion that ensures that the probability re-identification is kept to below a critical threshold. This will ensure that we meet our ethical, financial and legal obligations to ensure privacy of sensitive information. As indicated above, we must ensure that all direct identifiers are removed, encrypted or randomized. Leaving direct identifiers in the dataset will ensure that the probability of re-identification upon a data breach is 100%. Beyond just direct identifiers, quasi-identifiers can be further processed to ensure greater data privacy. This comes at the expense of utility, but it may be necessary to reduce the risk of re-identification for some data sets via inference or record linkage attacks. The masking of data can occur through a variety of methods – randomization, encryption, scrambling, substitution. For the purposes of this risk assessment report, we will focus mainly on the randomization of both sex and age. Randomization can be facilitated by most software programs by generating a random number column and sorting the selected variable according to this randomized number column. Randomization removes all linkages to pseudonyms or identifiers and will thus significantly reduce the risk of re-identification. It is important to note that randomization still imparts value to data, but only in the aggregate sense. Its main purpose is to de-identify the selected data on the record-level. Shown in is the first 10 records of the dataset in question after randomization of age and gender according to established practices.

Table 2. First 10 entries of a processed dataset with randomized Gender and Age variables.

Randomized on Gender and Age				
CustomerID	Gender	Age	EducationYears	JobCategory
3964-QJWTRG-NPN	1	71	15	1
0648-AIPJSP-UVM	0	19	17	2
5195-TLUDJE-HVO	0	43	14	2
4459-VLPQUH-3OL	0	71	16	2
8158-SMTQFB-CNO	0	24	16	2
9662-FUSYIM-1IV	0	68	17	3
7432-QKQFJJ-K72	1	39	14	1
8959-RZWRHU-ST8	0	25	16	1
9124-DZALHM-S6I	1	53	12	1
3512-MUWBGY-52X	0	63	11	6

7. Disclosure Risk Assessment

In order to test whether the de-identified data is sufficiently processed to ensure a minimum level of data privacy, it is recommended to conduct a comprehensive and quantitative risk analysis of re-identification. A framework for this type of analysis exists in the Healthcare Industry and is based upon the careful work of both Kniola[iii] and El Emam[iv]. Both researchers consider the determination of risk an exercise in determining the joint probability of reidentification and a successful breach/attempt. For data loss to occur, both of these events must occur simultaneously and thus a determination of joint probability is appropriate. According to the general product rule of probability, a joint probability can be calculated as the product of the conditional probability and marginal probability via…

$$P(A,B) = P(A|B) * P(B)$$

T1: Pr(Re-Id, Attempt) = Pr(Re-Id | Attempt) * Pr (Attempt)
T2: Pr(Re-Id, Acquaintance) = Pr(Re-Id | Acquaintance Attempt) * Pr (Acquaintance Attempt)
T3: Pr(Re-Id, Breach) = Pr (Re-Id | Breach) * Pr (Breach)
T4: Pr(Re-Id, External Breach) = Pr (Re-Id | External Breach) * Pr (External Breach)

Shown above are the relevant equations to determine the joint probability of re-identification and a data breach/attempt. The product of these equations affords an overall risk or probability of data re-identification for each threat situation. The last term in each probability statement is a marginal probability that estimates the risk of a data breach alone. These values have been determined through an extensive literature search and are listed explicitly in Section 5 of this analytic report. The first term is a conditional probability that seeks to determine the probability of re-identification given that a successful data breach attempt has occurred. It is a measure of how "anonymous" the data is and whether our data masking processes were truly successful.

To determine the conditional probability of reidentification, Pr (Re-Id | Breach), one will need to garner a sense for the distinguishability of records contained within the data set. The more unique a data record is, the higher the probability of re-identification. To determine the conditional probability of re-identification one will thus need to calculate the inverse of the size of the equivalence class to which the record belongs. The more records in an equivalence class, the harder it is to distinguish it. An equivalence class is a set of records that share a common set of quasi-identifiers. Shown in Table 3 are the equivalence classes and sizes for the first 50 records of the parsed Merrimack Survey Database. Depending on the intent of the risk assessment procedure, one can utilize the maximum or average value of Pr(Re-ID | attempt) in order to determine the overall risk of disclosure. For our purposes, we will utilize the average probability **calculated as 0.51.**

Using the equations in Section 7 and the pre-determined marginal probabilities of a data breach or attempt, one can then calculate the overall joint probability of reidentification resulting from a data breach/attempt.

T1: Pr(Re-Id, Attempt) = Pr(Re-Id | Attempt) * Pr (Attempt)
T1: Pr(Re-Id, Attempt) = 0.51 * 0.10 = 0.051 or **5.1%**

T2: Pr(Re-Id, Acquaintance) = Pr(Re-Id | Acquaintance Attempt) * Pr (Acquaintance Attempt)
T2: Pr(Re-Id, Acquaintance) = 0.51 * 0.78 = 0.398 or **39.8%**

T3: Pr(Re-Id, Breach) = Pr (Re-Id | Breach) * Pr (Breach)
T3: Pr(Re-Id, Breach) = 0.51 * 0.27 = 0.138 or **13.8%**

T4: Pr(Re-Id, External Breach) = Pr (Re-Id | External Breach) * Pr (External Breach)
T3: Pr(Re-Id, External Breach) = 0.51 * 0.59 = 0.30 or **30%**

Interestingly, lowest probability of data disclosure is embodied in threat scenario T1 – internal, deliberate. The low probability of this scenario is brought forth via the implementation of a robust data security plan with adequate controls and employee training. It is important to establish a threshold of risk necessary to proceed with data release/sharing. Based upon these results would one have high confidence that the data set was sufficiently de-identified to limit the risk of re-identification? For non-health related datasets, it may be permissible to have a maximum overall risk of re-identification of 30%. The risks determined for this exercise are considerably higher this threshold and further processing, masking or generalization will be necessary to modify the size of equivalence classes.

Table 3. Determination of Equivalence Class Size for Parsed Merrimack Data Set. Gender and Age variables are randomized and are thus not included in the determination of equivalence class.

CustomerID	Gender	Age	EducationYears	JobCategory	Equivalence Class	Size	Pr(Re-ID | Attempt)
7164-HJCLUK-29N	0	29	8	1	A	1	1.00
8241-PWPONH-62O	1	55	8	2	B	3	0.33
6592-QBSENI-6Y4	1	46	8	2	B	3	0.33
2041-WAUPOH-84L	0	73	8	2	B	3	0.33
1811-FUILIP-L3S	0	68	9	5	C	1	1.00
8795-FYOXCT-P09	1	36	10	1	D	3	0.33
2654-JRLXAI-HC7	0	33	10	1	D	3	0.33
4724-UVMKNV-Z1I	1	18	10	1	D	3	0.33
1743-WAHPKT-DNW	1	49	10	2	E	1	1.00
8268-VLELYK-IJK	1	73	10	3	F	1	1.00
8514-TCRHWV-DPM	0	66	10	5	G	1	1.00
6203-JLXPIJ-YQZ	0	37	10	6	H	2	0.50
4048-TDAQGT-9Y6	1	69	10	6	H	2	0.50
2228-KOLOPU-FY3	0	51	11	1	I	4	0.25
6441-FJUWZQ-7G8	1	49	11	1	I	4	0.25
0926-OZQUQX-GXA	0	74	11	1	I	4	0.25
4115-RQKQTM-RXB	0	65	11	1	I	4	0.25
5538-DRGHIE-PF0	1	46	11	2	J	4	0.25
8170-YHWTSQ-825	1	29	11	2	J	4	0.25
7217-UECHSF-PCR	1	49	11	2	J	4	0.25
1150-ESKSKG-FRM	0	58	11	2	J	4	0.25
2866-TTOTKL-TA7	0	42	11	3	K	2	0.50
8888-ERPHEF-H3V	0	21	11	3	K	2	0.50
1114-UELZXT-QT7	1	54	11	4	L	1	1.00
3512-MUWBGY-52X	0	63	11	6	M	1	1.00
9124-DZALHM-S6I	1	53	12	1	N	1	1.00
7634-AVNEXZ-7AG	1	70	12	1	N	1	1.00
9809-NBCXAS-1RB	0	22	12	2	O	2	0.50
7245-IUATZR-VLV	1	29	12	2	O	2	0.50
4226-TEHKPJ-GKG	0	71	12	3	P	2	0.50
5472-GUGMBS-ZVK	1	72	12	3	P	2	0.50
8846-JJEGFU-I69	0	26	13	1	Q	1	1.00
5052-NKTQMH-BKA	0	68	13	2	R	1	1.00
4225-PZZDIY-IBH	0	66	13	6	S	2	0.50
1441-JNZRPW-8KQ	1	22	13	6	S	2	0.50
7432-QKQFJJ-K72	1	39	14	1	T	4	0.25
9714-LCYJJB-SCS	1	42	14	1	T	4	0.25
5007-HXEVVL-NW5	0	40	14	1	T	4	0.25
0309-KOSUPM-CSV	1	23	14	1	T	4	0.25
5195-TLUDJE-HVO	0	43	14	2	U	2	0.50
2687-GRSYRT-YRW	1	72	14	2	U	2	0.50
2041-PNMGHX-TXJ	0	28	14	3	V	2	0.50
4973-FHPBZQ-AAA	0	18	14	3	V	2	0.50
0649-TBFJFL-QU4	1	25	14	6	W	2	0.50
0712-WQXYVV-HUP	0	27	14	6	W	2	0.50
3964-QJWTRG-NPN	1	71	15	1	X	4	0.25
4974-FUBHDF-Z7L	1	36	15	1	X	4	0.25
8101-YOBRYX-REF	0	22	15	1	X	4	0.25
7441-RFDNVF-7SP	1	28	15	1	X	4	0.25
6246-WBWHHS-GFB	0	32	15	2	Y	4	0.25

8. Summary and Conclusions

The preceding provides a comprehensive framework to ensure sufficient data privacy of data sets to limit the overall risk of re-identification. It is anchored by a probabilistic quantitative risk assessment that affords a decisive measure of the effectiveness of any de-identification process. The implementation of this framework will ensure that Newco Industries remain compliant with the strict regulations of the GDPR, meet our ethical obligations to our customers and reduces the risk of reputational damage due to a data breach.

[i] Hoofnagle, Chris; van der Sloot, Bart; Borgesius, Frederik Zuiderveen (10 February 2019). "The European Union general data protection regulation: what it is and what it means". *Information & Communications Technology Law*. **28**: 65–98. doi:10.1080/13600834.2019.1573501.

[ii] L. Sweeney. k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 557-570.

[iii] Kniola, L. "*Calculating the Risk of Re-Identification of Patient Level Data Using a Quantitative Approach*", **2016,** PhUSE Paper DH09.

[iv] Khaled El Emam. "*Concepts and Methods for De-identifying Clinical Trial Data1*", **2015,** Appendix B in Sharing Clinical Trial Data: Maximizing Benefits, Minimizing Risk, National Academies Press.