Explainer: De-Identifying Data[1]

A key element of many data governance frameworks is managing privacy issues associated with data. Much data concerning human activity will include personal information - information about identifiable individuals. Such information is subject to data protection laws across numerous jurisdictions.

One often discussed mechanism for using data that includes personal information is to strip such data of elements that can potentially identify individuals - a process known as “de-identification”. “Personal information” includes both direct and indirect identifiers. Direct identifiers include information that can be used to identify an individual independently or in combination with other information (such as name, phone number, address, credit card number, or social insurance number, etc.). Indirect identifiers include information that can be used to identify an individual when combined with other available information such as background knowledge of the individual or other available datasets.

Data de-identification inevitably involves a trade-off between detail and privacy. More detailed datasets will include more information, including more personal information. Privacy protection, on the other hand, requires sharing less detail, which may prove to limit the usefulness of the de-identified dataset. However, privacy is more fundamentally about the right and ability to control information about oneself. In particular, individuals may not want their personal information to be used for purposes they don't approve of, even if information has been de-identified. Data protection rights extend even to de-identified personal information through other fair information principles such as purpose specification, accountability, and openness. It is important to remember that when the individual concerned sees data about themselves, they recognize it as their data, regardless of whether it can be linked back to them or to the group they belong to. They know that it has been collected, used, and disclosed.

The efficacy of de-identification strategies is highly contentious. One prominent researcher, noted computer scientist Cynthia Dwork, has said that “de-identified data isn’t”. There is no guarantee that de-identified data cannot be re-identified. The risk of re-identification is significant. Despite this, de-identification strategies are often deployed in data sharing initiatives.

When they are used, data de-identification strategies should address the following considerations:

1. Caution: Over-Estimating the Efficacy of De-Identification Strategies
2. Best Practices and Accepted Standards

[1] This explainer is informational only. Its purpose is to give context to the accompanying document. It summarizes and simplifies a body of writing in a field that is contentious and complex. It does not represent the views of the CCLA, which is on record as objecting to data de-identification as an ineffective means of safeguarding privacy, and preventing privacy harms and violations.

3. Standard Approaches
4. Risk of Harm
5. Re-Identification Risk
6. Applying De-Identification Techniques
7. Considerations for Specific Kinds of Data

1. **Caution: Over-Estimating the Efficacy of De-Identification Strategies** - As stated above, the risk of re-identification is significant. Security and privacy advocates are critical of approaches to evaluating the risk of re-identification; poor security measures applied to data sets; and the underestimation of adversary knowledge, motivation, and resources.

2. **Best Practices and Accepted Standards -** There are no universal or legislated protocols or standards for de-identifying data. Frameworks have been proposed, but not generally adopted, for publicly responsible "trusted user" vehicles that would minimize the risk of re-identification. Particular fields have developed standards and best practices to approaching de-identification exercises. For example, the *Pan-Canadian De-Identification Guidelines for Personal Health Information* (2007), developed by Dr. Khaled El Emam, a well-known proponent of de-identification, proposes a widely consulted means for sharing personal health information in Canada.

3. **Standard Approaches -** In the absence of standard approaches to de-identification of data in a specific domain, the Office of the Information and Privacy Commissioner of Ontario has produced *De-Identification Guidelines for Structured Data* in 2016. The document develops a multi-step process for de-identification by calculating the risk of re-identification for classified variables before introducing specific de-identification techniques. The document describes a standard approach to de-identifying data. It does not purport to be a panacea for privacy issues associated with data. It offers a systematic approach to de-identification tasks that, if replicated, will seek to improve organization data sharing and provide an opportunity to address privacy harms. Other similar documents include the American National Institute of Standards and Technology's *Guide to Protecting the Confidentiality of Personally Identifiable Information*, and the EU's Article 29 Working Party's *Opinion 05/2014 on Anonymization Techniques*.

4. **Risk of Harm -** A de-identification strategy should account for the risk to individuals of inadvertent disclosure or re-identification by any means. Given that the risk of re-identification is significant, an organization should seriously consider the wisdom of releasing de-identified information of a sensitive nature that would be highly valuable to an attacker and result in privacy violations or inflict harm on individuals if re-identified. Note that it can be difficult to distinguish between what is and is not "sensitive". Some kinds of information (*e.g.*, health, financial) are normally considered sensitive. But non-sensitive information can become sensitive when combined with other personal data.

5. **Re-Identification Risk -** The extent of de-identification required to protect personal privacy should be proportional to the risk of re-identification. Determination of risk should consider multiple factors, such as the sensitivity and detail of the information, the potential harm to individuals in the event of inappropriate use, the potential value to outside parties of re-identified data, and the likelihood of a re-identification attack.

6. **Applying De-Identification Techniques -** A range of de-identification techniques are available. These include masking direct identifiers, modifying the size of equivalent classes, pseudonymization, encrypting values, adding random noise (*e.g.*, "salting" or "peppering"), removing columns of identifiable variables, generalizing data over groups of records, differential privacy, and suppressing quasi-identifiers. Decision-makers must exercise discretion and care in selecting techniques appropriate to the data involved.

7. **Considerations for Specific Kinds of Data -** Specific kinds of data involve specific challenges for de-identification strategies. some non-exclusive examples:

    a. **Location Data** - There is no evidence to support the claim that location data can be effectively anonymized. See our accompanying document, Location Tracking Explainer.

    b. **High-Dimensional Data** - High-dimensional data features records that include many independent fields. Big data analytics applications thrive on high-dimensional data. Accordingly, high-dimensional data is now the norm, not the exception. De-identification techniques have had limited success with high-dimensional data given that there are so many options for cross-referencing re-identification attacks. As a result, until proven otherwise, it must be assumed that de-identification of this data is unlikely to be effective. This is an evolving area and new approaches and standards are emerging, caution and care are required.