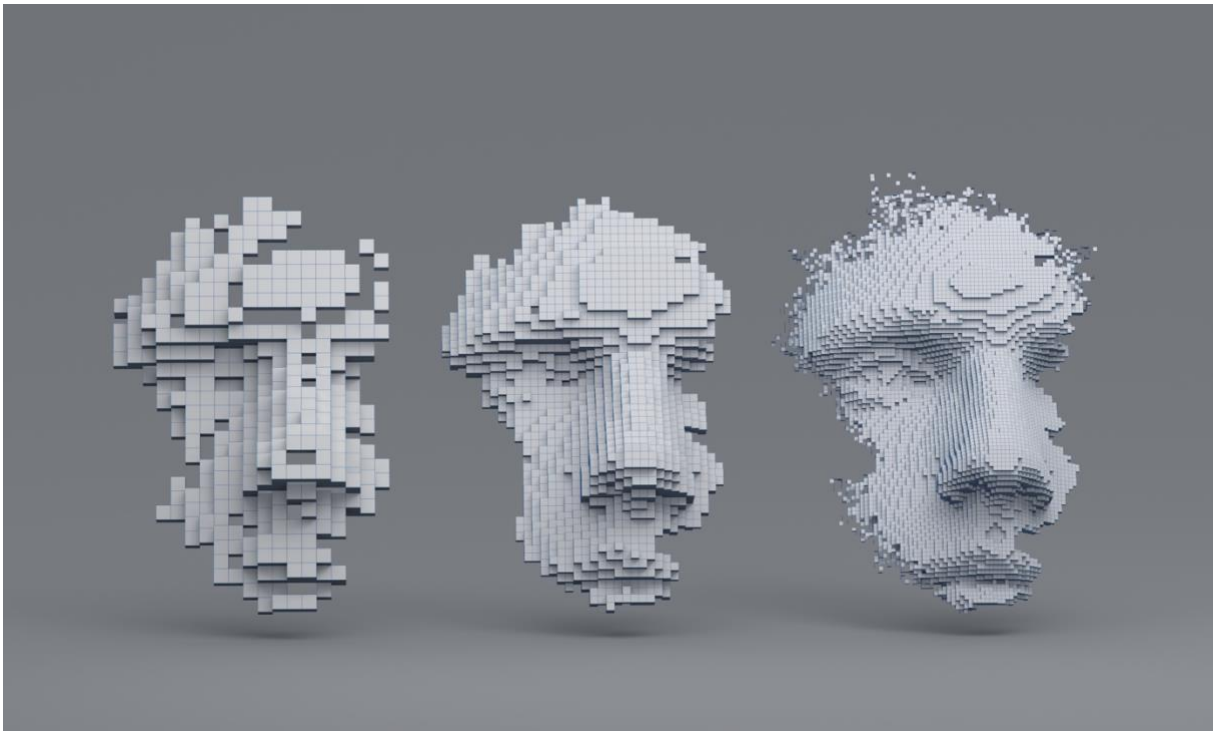
Palantir

Beyond Anonymisation

A comprehensive approach to handling personal data

palantir.com



Introduction

Anonymisation is often considered the panacea of data protection: it helps organisations derive useful insights from data while allowing them to avoid the legal and privacy risks associated with processing personal data... or so the theory goes.

However, as a standalone measure anonymisation can fall short. To cite a famous example, a graduate student was once able to use nominally anonymised hospital data from Massachusetts to send the state's Governor his own health records.

We should therefore be careful about the use of the word 'anonymised'. As the UK's Information Commissioner's Office (ICO) points out, it risks leaving organisations with the false sense of security that data is impossible to re-identify under any circumstances.

In this whitepaper we want to build on the UK's Understanding Patient Data's reconceptualisation of anonymisation as 'de-personalisation'. First, we explore a range of techniques that enable organisations to minimise the exposure of personal data. We'll argue that anonymisation redefined as such can reduce the risk associated with data processing but does not necessarily absolve organisations of their data protection responsibilities – nor need that be the primary goal. We'll then demonstrate how our customers can safely process personal data in Palantir Foundry by relying on a range of built-in privacy-protective means.

What is anonymisation?

Anonymisation traditionally refers to the process of removing identifiable information from datasets so that the people described in the data cannot reasonably be re-identified. Anonymisation is useful because it allows organisations to work with data derived from individuals without putting the privacy of those individuals at risk.

The COVID-19 pandemic has demonstrated the importance of processing sensitive personal data securely and at scale. Public and private organisations have an obligation to keep their citizens and employees safe without compromising the protection of their personal data; and in order to leverage public health data to make informed decisions, organisations need to have a range of technical tools at their disposal. However, users at these organisations should only have access to the data they need to achieve a legitimate processing purpose, as defined under applicable laws, regulations, and institutional or organisational policies. ICO considers anonymisation an important tool for balancing innovation with individuals' privacy, particularly in the use of open data for the public good.

The anonymisation approach an organisation takes depends on the purpose of the processing and the risk associated with the processing. As a rule of thumb, the more sensitive the data and the greater the harm associated with re-identification, the more robust an organisation's approach to anonymisation needs to be. At the same time, anonymisation can result in the loss of data utility, potentially rendering analyses performed on anonymised data useless.

What are different ways to achieve anonymisation?

To understand how data can be anonymised in practice, we highlight some key approaches to deidentifying data below, using the notional example of a 37-year-old Indian man named Yash Patel living in a six-person household in postcode W1D 37X. (Note that age, gender, address, and ethnicity have been chosen at random, and for illustration purposes only.)

PSEUDONYMISATION

Pseudonymisation is the process of replacing or removing any directly identifying information from a dataset. However, the data can still be reidentified in combination with other data. For instance, let's assume Yash was diagnosed with COVID-19 in March 2020. An example of pseudonymisation would be to replace Yash's personally identifying information with an

artificial identifier, or pseudonym: “AB1236C was diagnosed with COVID-19 in March 2020.” Pseudonymising data can be a useful approach to anonymisation when an organisation does not need to know the identity of someone within a dataset, but needs to analyse someone’s data across datasets. For instance, public health officials may be interested to know from a different data source that AB1236C suffers from high blood pressure. Note that if adequate information can be obtained by linking records from different data sources pertaining to AB1236C, it might be possible to infer that the records belong to Yash. Since the risk associated with re-identification remains high, pseudonymised data is often considered separately from anonymised data. Under the EU General Data Protection Regulation (GDPR), pseudonymised data is considered personal data.

GENERALISATION / AGGREGATION

Generalisation refers to the process of converting uniquely identifying properties into broader categories. For instance, a record revealing data of a 37-year-old Indian man who lives in postcode W1D 37X can be generalised to belong to a man aged between 35-38 who lives in Westminster. Aggregation is a special case of generalisation where generalised entries are combined to represent groups instead of individuals in the anonymised dataset. The extent to which data can and should be generalised again depends on the purpose of processing and the sensitivity of the data.

OBFUSCATION

Obfuscation refers to the process of obscuring or concealing the meaning of data. Masking is the most common form of obfuscation where data is replaced with realistic-looking but inauthentic data. Pseudonymisation can be achieved through masking as well, where the pseudonym becomes the masked value. An obfuscated version of Yash's record could be: *******, who lives in a six-person household, in postcode W1D *** was diagnosed with COVID-19 in March 2020. Encryption is an example of a more complex obfuscation technique where the identifiable data point is replaced with an encrypted value, which can be reversed through decryption. On the other hand, hashing is an example of permanent

obfuscation; determining the original value from its appropriately hashed version is not possible.

SYNTHETIC DATA

Synthetic data refers to data that is generated to represent some basic characteristics about the people in the original dataset. For instance, public health officials may not need to know that Yash Patel was diagnosed with COVID-19 in March 2020, but they may be interested to know how many Indian men aged between 35-38 were diagnosed with COVID-19 in postcode W1D 37X in March 2020. All they need is identically shaped data (i.e., the data format for the original and synthetically generated anonymised dataset is the same) which at least somewhat accurately captures the frequencies of COVID-19 for relevant demographic categories across different postcodes. In other words, synthetic data is representative data. Which characteristics need to be retained in synthetic data depends on the purpose of processing. Organisations need to be careful that the characteristics intended to keep the data representative do not make the data uniquely identifiable, e.g., when Yash Patel is the only Indian man aged 37 diagnosed with COVID-19 in postcode W1D 37X, and the synthetic version of this data accurately represents frequencies at the postcode and age level.

k-ANONYMISATION

k-anonymisation refers to a statistical process to transform a dataset such that for each identifying property within the dataset, there needs to be at least k-1 records with the same properties. It can be obtained by using a combination of techniques described above, like masking and generalisation. For example, to have the record of a 37-year-old Indian man living in a six-person household in postcode W1D 37X be part of a 4-anonymised dataset, there need to be at least three other records which share the same identifying attributes. This ensures that if someone tries to find Yash Patel in this dataset based on these identifiers, they will end up with four different records without knowing which data entry corresponds to Yash. If this is not true in the original dataset, then we can generalise age values and/or mask the postcode until the 4-anonymised property becomes an invariant of the dataset.

These are just a few of the many anonymisation techniques organisations can use. Most use a variant or combination of the techniques described above, and accordingly provide different deidentification guarantees. These approaches are not mutually exclusive. To obtain a k-anonymised dataset, one could use both generalisation and obfuscation, for instance. It's also important to note that some of them highlight the procedures involved in the anonymising process, while others emphasise the end state i.e., the mathematical guarantees that the anonymised dataset must satisfy.

Why is anonymisation hard to do?

The biggest challenge associated with anonymisation is striking the right balance between retaining the utility of the data while protecting the privacy of the individuals whom those data represent. In practice, it can be difficult to determine what level of deidentification is appropriate, balancing the sensitivity of the data with the purpose of processing.

THE PROBLEM WITH LOW-LEVEL AGGREGATES

Even when consolidated for aggregation, individual datasets may still represent only small groups of individuals. These are considered low-limit aggregate groups, and they pose a risk of reidentifying individuals who belong to these groups. In certain cases, mere membership in a low-limit aggregate group can have repercussions for individuals – performance evaluations, for example. On the other hand, suppressing the presence of a low-limit aggregate can deteriorate data utility. The probability of ending up with low-limit aggregates in anonymised data increases as we increase the granularity at which we aggregate. For example, a dataset is likely to have fewer individuals in postcode W1D 37X who are aged between 35 and 38 years, who live in a six-person household, and who were diagnosed with COVID-19 in March 2020 as compared to the number of individuals who match all those attributes but with the location generalised from a specific postcode to a government district; in this case generalise from W1D 37X to Westminster in Central London. Lower granularity for location information in the aggregated dataset reduces the risk of low-limit aggregates.

REIDENTIFICATION THROUGH DATA LINKING

An anonymised dataset that might seem safe for wider sharing can still be susceptible to re-identification – even if it's been verified to not reveal individual information. Partial information from the anonymised data, linked with insights from other data sources (e.g., datasets already in the public domain) can be used to reverse engineer aspects of the anonymisation process. To cite a famous example, researchers were able to re-identify a subset of users in the Netflix Prize dataset, which contains anonymous movie ratings of Netflix subscribers, by obtaining additional information from the Internet Movie Database (IMDb).

LOSS OF UTILITY AND ACCURACY

Techniques like data suppression or generalisation are not always feasible for anonymisation purposes. They can lead to negative consequences like loss of utility or misleading results from analysis performed on top of anonymised data. To illustrate, let's consider an extension of the example above, this time considering three individuals living in postcode W1D 37X and aged between 35-38 years, who live in a six-person household, and who were diagnosed with COVID-19 in March 2020. In order to remove the low-limit aggregate, we could suppress the number 3 in this data, or alternatively we could remove this entry from the dataset entirely (another form of suppression). For any macro analysis on the anonymised dataset, these three individuals will not be represented in the aggregated statistics. The accuracy concerns become more severe as more low-limit aggregates are suppressed in the anonymisation process. The trade-off between effective anonymisation and retaining utility for various kinds of data analysis becomes quite evident in this scenario.

CHALLENGE OF LIVE DATA STREAMS

Anonymisation is usually implemented as a one-time static operation, but non-trivial to perform on live data streams while still maintaining deidentification guarantees. In order to efficiently process data streams, which are continuous and usually unbounded, each record should be processed as it comes in. Most anonymisation techniques are only effective on

static cuts of data where retroactive updates to any single data entry are not a concern.
In light of these challenges, it's clear that there is no silver bullet for obtaining perfectly, or rather appropriately, anonymised datasets. More generally, the efficiency and effectiveness of anonymisation workflows depend on a variety of factors, including access to relevant data assets; clarity around and awareness of the purpose of anonymisation; and additional controls to protect sensitive versions of data.

How, then, does one design a big data platform that goes beyond providing the ability to write one-off anonymisation pipelines to building an ecosystem for responsible data processing as a whole? Where a group of researchers can supplement their publication with a deidentified version of the underlying data allowing the research community to verify their results? Where data engineers are able to build and maintain applications to monitor COVID-19 trends without getting access to individual patient details?

Beyond Anonymisation

From our perspective, anonymisation should be considered as only one tool in the data protection toolkit. Organisations need not be afraid from processing personal data as long as they have technical and organisational means at their disposal that enable them to process that data lawfully and appropriately. With Palantir Foundry we are proud to have built a data platform that combines effective data tooling with a range of privacy controls. This more comprehensive approach enables our customers to achieve their processing purpose while staying on top of the risks, helping them determine not only the best approach to data minimisation, but also strengthening their overall data protection posture.
The following constitutes an overview of Foundry's core data protection features and how they can be leveraged to facilitate anonymisation in the broader context of processing sensitive data:

DATA INTEGRATION

Data integration lies at the heart of Palantir Foundry's power to enable organizations to process data both effectively and responsibly. For instance, an organisation tasked with better understanding the spread of COVID-19 is able to combine sensitive patient records with demographics data in a single interface. Having all relevant data in one place helps with implementing effective anonymisation workflows. As the first part of this white paper demonstrates, a one-size-fits-all anonymisation solution does not generally work. A robust and flexible depersonalisation process must account for all relevant data assets and corresponding metadata. This includes the granularity of information contained within the dataset to be anonymised, the subset of identifiable information that is required to achieve a legitimate processing purpose, specific rows or columns in the data which require additional protections, and an overview of additional data sources that contain information which could lead to undesirable reidentification.

ACCESS CONTROLS

Granular access controls at the project, datasource, dataset, and even sub-dataset level ensure that access to sensitive data sources is always limited to that which is necessary to meet legitimate processing needs. Users can work on the same dataset with different levels of permissions. For instance, a limited number of users with privileged permissions working with directly identifiable data, while the majority of users only has access to only de-personalised data downstream.

PURPOSE LIMITATION

Foundry enables users to structure data assets tied to specific purposes. All subsequent processing of data relies on purpose-based access controls. Instead of getting access to individual datasets, users get access to purposes. Purposes make it clear why a user might have access to personal data in Foundry, and when this access is no longer necessary. Users can also be reminded of the conditions of compliant and appropriate processing by means of popups configurable to appear before sensitive data operations. For instance, a

user can be prompted to enter a mandatory justification, subject to verification and approval, before marking an anonymised dataset safe for wider sharing.

TESTING AND VALIDATION

Analytical tools in Foundry provide the ability to do validations and battle test anonymised data before it is shared more widely within the system or exported for external purposes. This can help ensure, for example, that a k-anonymised dataset maintains the required statistical invariants before the data is published. The presence of trends that must be retained in the anonymised version of the data can be verified by performing the same analysis on the original dataset and its anonymised version, and then comparing the results side by side in Foundry. Additionally, these comparisons can be used to configure health monitors to flag if reidentification risk in anonymised data increases beyond a fixed threshold. This could happen due to ingestion of new data assets, which can increase the possibility of reidentification via record linking, or due to faulty updates in the anonymisation logic.

DATA LINEAGE

In order to keep track of which users have access to which level of identifiable data for what purposes, Foundry offers powerful lineage tracking and data provenance tools to get a comprehensive overview of how data is flowing within the system. Administrators can switch between different views to see details such as the source of raw data, who is using what version of data for different use cases, and the level of identifiability of those versions.

DATA INFERENCE

For particularly sensitive types of personally identifiable information (PII) such as personal health information (PHI), Foundry provides tools able to run background checks to infer PII/PHI across all data in the system, automatically flagging and preventing access to sensitive data which may have been uploaded accidentally or deidentified insufficiently. For example, if an otherwise anonymised version of a dataset still contains a postcode in a free-

text field, which may not have been scrubbed appropriately during the anonymisation process, compliance and data protection users can receive notifications to review this data in Foundry directly before approving and making the data available to users downstream.

NEED-BASED DECRYPTION

Occasionally, if a user does need access to directly identifying information in otherwise de-identified datasets, Foundry offers a flexible encryption and decryption service. Users will be prompted to enter a justification prior to decrypting the data, effectively striking a balance between protecting sensitive data and enabling users to do their job. The list of justifications can be reviewed later by relevant oversight bodies within or outside the organisation for verification and auditing purposes.

Protecting Privacy and Civil Liberties (PCL) at Palantir

Since our founding, we've believed that an information system becomes a liability when it lacks robust, built-in functionalities that facilitate processing data in accordance with legal and ethical obligations. This is why we have set data protection among our highest priorities in the development and deployment of Palantir software; and it is why government organisations, private companies, and major philanthropies trust our products to safeguard their most important data and analytical assets, including in the context of COVID-19.

To ensure our products can meet the legal and ethical requirements in practice, Palantir has invested in an in-house privacy and civil liberties (PCL) engineering team. This interdisciplinary team of software engineers, data scientists, philosophers, designers, and policy experts is available to meet with customers to help them develop and implement privacy-protective workflows in our platforms. Find out more about Palantir's approach to privacy and civil liberties engineering at https://www.palantir.com/pcl/.