TRUATA.

In collaboration with

FUTURE OF PRIVACY FORUM

Mindful Data Mining

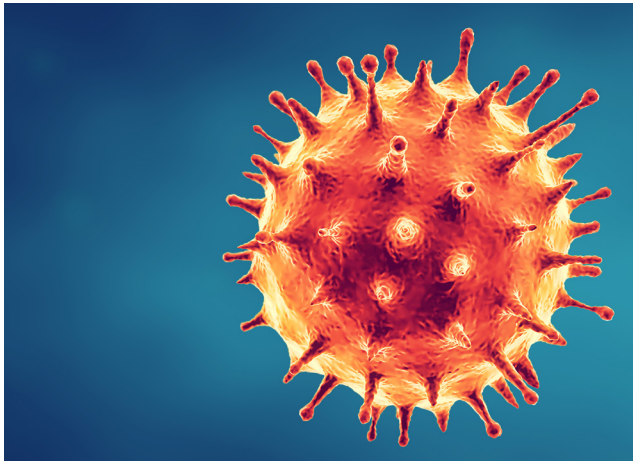
Responsible data analysis can deliver profound social good without eroding the data protection rights of individuals

TRUATA. FUTURE OF PRIVACY FORUM

The spread of the novel coronavirus (SARS-CoV-2) has changed the behaviour of hundreds of millions of people unlike any event in generations.

The lack of successful treatments or a vaccine has meant widespread lockdowns in an attempt to control what has quickly become a pandemic. While scientific research seeks medical solutions, governments, health authorities and the private sector are seeking to deploy data and technology to track, predict, and contain the spread of the virus.

Dr. Maurice Coyle
Chief Data Scientist at Truata.



Working with Data During a Crisis

In this article, we will discuss the ways in which thoughtful and responsible data analysis can deliver profound social good using data without eroding the data protection rights of individuals that have been painstakingly built up in recent decades. Data analysts need to understand what uses of data will be lawful, transparent and fair and need to be particularly mindful of potential endangerment of under-represented or otherwise vulnerable populations.

There is an understandable and necessary rush to produce data analyses to help government agencies, commercial companies and most of all citizens with information as to the spread of a virus such as SARS-CoV-2.

Data provides information to track the progress of the virus and demonstrates how containment measures are affecting the spread of the virus.

At the same time, those working with data must be aware of the potential risks of working on such urgent and important tasks. Inaccurate data may cause unnecessary stress or panic and place a further strain on public resources.

It may also give a false sense of comfort and result in people relaxing their responses to the virus, leading to more rapid spread.

Tips for Working with Data

The fundamentals of data science must be strictly observed to ensure accurate insights are shared, and well-informed decisions can be made as outlined below.

- **Testing assumptions.** Data without context can be misleading and all factors that may be affecting an analytical outcome must be considered, controlled for, and communicated. Choice of statistical tests, methodology, and metrics can all greatly affect the outcome and they are usually based on an assumption regarding the data (such as its distribution) that should be tested and verified [1,2].

- **Distribution analysis.** The basic properties of data distributions are incredibly important to understand prior to analysis. Whether data is uniform or multimodal, skewed or normal, long- or fat-tailed all affect how analysis should be performed. Statistics such as mean, median, or standard deviation, may be misleading when relied upon for multimodal distributions, for example [3].

- **Sampling approach.** Where sampling is required for practical reasons, the sampling approach can affect the analytic outputs by introducing bias. Analysts must ensure data samples are representative which requires knowledge of the population being sampled. The distribution analysis above is one component of this understanding [4].

- **Data cleansing and preparation.** In the early days of a virus such as SARS-CoV-2, data is by nature sparse and great care must be taken to cleanse and prepare the data prior to analysis to remove or mitigate sources of bias [5].

- **Integrity of analytical outputs.** Where possible, multiple sources of data should be used to corroborate any findings. Ideally, the data used to generate analytical insights should be made public so the results can be verified.

In summary, it is of the utmost importance that scientific principles are adhered to when analysing data related to a crisis such as the novel coronavirus. Care must be taken to ensure bias and noise do not influence outputs; scientific methods must be used, and above all, transparent, verifiable communication of results is critical.

Sensitive Data Types

Often, the most powerful types of data are also the most sensitive. Health records and location data are two types of data that would be incredibly useful for tracking the spread of the coronavirus and researching ways to find a cure.

They are also the most sensitive from a privacy point of view, and the strictest levels of privacy safeguards must be applied to these kinds of data, even during a global pandemic.

Once data has been made available for analysis, it cannot easily be recalled, so measures must be taken to protect the data and data subjects they related to before any release or access is provided. Such actions may include:

- **Minimisation;** ensuring that only the data strictly required for analysis are made available for analysis means singling out and linkability risks are minimised.

- **Aggregation;** where possible, data should be aggregated into a form that is useful for a particular kind of analysis so that event- or data subject-level detail is not available. In many use cases related to using location data, aggregate forms of the data are sufficient.

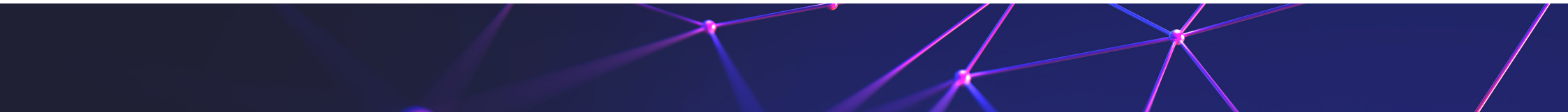
- **Access rather than sharing;** Providing tools and mechanisms for analysis to occur over a dataset without providing direct access to the data add protection against misuse of the data and enable the data to be recalled or deleted when it has served its purpose.

Maintaining Standards of Privacy and Data Protection in a Time of Crisis

Neither the fundamental human right to privacy nor adherence to principles of data protection disappear during a crisis, so when implementing any data analytics project the usual standards should be applied.

The pandemic will pass, and when it does any personal data that has been made available without adequate safeguards will represent a privacy risk and a potential liability under data protection regulations.

Here we discuss several aspects of privacy enhancing techniques that we have used to great effect within Trūata to ensure the highest levels of privacy are achieved while delivering the desired level of analytic utility.



Quantifying privacy

When working with personal data or preparing data for use by analysts, the principles of privacy by design and privacy by default should apply. Objective measurement of the re-identification risk within a dataset to quantify the level of personal data it contains is essential. This goes beyond simply identifying data such as email addresses, phone numbers, real names, etc, that could directly identify an individual.

There are many types of risk within a dataset, and certain fields can combine to form a "fingerprint" or "quasi-identifier" that can uniquely identify an event or individual. Where personal data exists, attempts should be made to aggregate or otherwise transform the data so that it retains important analytical characteristics without exposing individuals.

At all times, the data owner should ensure that the data to be used is minimised to the greatest extent possible.

Robust and objective risk quantification is required both before and after any anonymisation or aggregation techniques have been applied, to measure the change in overall risk profile and residual risk of re-identification.

Synthetic data

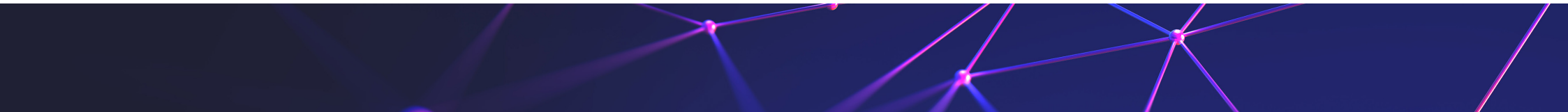
In certain cases, a synthesised version of the dataset may be used to develop business logic or train models. However, the final analysis must usually be performed on the original data. This is because synthesised data that retains correlations between columns usually reproduces many of the quasi-identifiers in the original data and thus retains much of the re-identification risk. To produce a privacy-safe synthetic dataset, the correlations between high-risk quasi-identifiers should be perturbed or removed. Releasing such datasets and providing tools or processes to enable business logic to be executed or models to be trained on the original data mean that the conflicting needs to preserve privacy while generating accurate analytic results are both met.

Data access control vs data sharing

To enable analysts to actually work with a particular dataset, it is useful to provide tools that enable the analysis to be performed but without sharing the data or without providing direct access to the underlying, row-level data.

This layer of abstraction separates the analyst from the actual data while enabling the analysis to proceed. In the age of Big Data, researchers have access to huge amounts of publicly available data and computation resources at very low prices.

For high-dimensional datasets, this means that linking with another dataset or inferring sensitive information about individuals is very hard to prevent [6]. Thus, preventing access to raw data is likely to be critical to ensure re-identification of individuals cannot occur.



Trusted 3rd Party

Quite often, the full value of data is realised by combining data from different sources. In these cases, it is useful to ensure that the integrity and privacy of the different datasets are preserved by using a trusted 3rd Party to manage and combine the different datasets.

This ensures purpose limitation is observed, along with other principles of privacy by design in enabling powerful analysis. Legal trusts, like Truata, are well placed to act as a trusted third party where there are multiple parties, as trusts can act independently and objectively and can be designed to prevent the use of data in any manner other than strictly for anonymising the data, carrying out analytics and providing insights back to the multiple parties involved.

In summary, data owners need to be mindful of the inherent risks in providing any access to data for analysis. Maintaining existing hard-won privacy and data protection standards is critical to ensure there is no trade-off between individuals' rights and social good. Using privacy-enhancing technologies to quantify privacy risks in a dataset, to separate analysts from the underlying data, and to generate safe datasets means societal good may be delivered without compromising individual privacy.

Conclusions

The value that data can bring during a time of crisis such as the COVID-19 pandemic is compelling. Data is of critical importance to prevent and contain the spread of the virus, and to protect vulnerable citizens from its continued spread. However, it is of equal importance that the fundamental human rights of those same citizens whose data is being analysed are not traded to achieve this societal good.

We have outlined ways in which data scientists should conduct their analyses in a time of crisis, observing fundamental principles of data analysis to ensure timely, appropriate communication of insights derived from data.

We have also outlined how the core principles of data protection can be observed even in a time of crisis to ensure the decades of work that have gone into achieving high standards of data protection and privacy are not eroded.

Find out more at truata.com

References

1. Garson, G.D. Testing Statistical Assumptions (Statistical Associates Publishing, 2012).
http://www.statisticalassociates.com/assumptions.pdf

2. Godsey, B. Check your assumptions about your data (2017).
https://towardsdatascience.com/check-your-assumptions-about-your-data-20be250c143

3. Barber, M. Data science concepts you need to know! Part 1 (2018).
https://towardsdatascience.com/introduction-to-statistics-e9d72d818745.

4. National Research Council . Frontiers in Massive Data Analysis, Chapter 8 (2013).
https://www.nap.edu/read/18374/chapter/10

5. Krishnamurthy, P. Understanding Data Bias.
https://towardsdatascience.com/survey-d4f168791e57

6. Rocher et al. Estimating the success of re-identifications in incomplete datasets using generative models. Nat Commun 10, 3069 (2019).
https://www.nature.com/articles/s41467-019-10933-3

Truata Headquarters: Silverstone House, 1st Floor, Ballymoss Road, Sandyford. Dublin D18 A7K7. Ireland | +353 1 566 8468 | truata.com | ©2021 Truata Limited ("Trūata") | All rights reserved. Trūata's marks and logos and all other proprietary identifiers used by Trūata herein ("Trūata Marks") are all trademarks and/or trade names of Trūata and/or its affiliates. Any other products, services, or company names referenced herein may be trademarks of Trūata or of other companies; if trademarks of other companies, there is no endorsement or affiliation, expressed or implied, claimed by Trūata. ARTICLE_DATAMINING_2020_004