

An iceberg floats in a blue ocean under a blue sky with white clouds. The visible tip of the iceberg is small and jagged, while the submerged portion is much larger and more complex. The water is a deep blue, and the sky is a lighter blue. The overall scene is a metaphor for hidden risks in data.

TRUATA.

Beneath the surface:

Understanding hidden privacy risks in your data

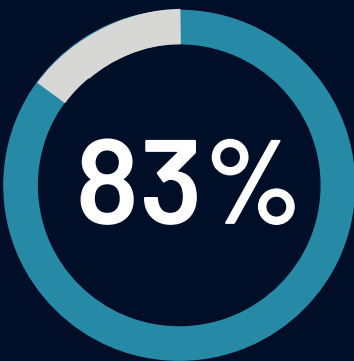
Value creation in a data-driven economy

Extracting value from data — securely and ethically — has fast become a business imperative. Those who want to survive and thrive in the next decade are getting to grips with both the revenue opportunities and the risks associated with data value chains.

Already, organizations are redesigning and remapping operations to embed privacy-by-design at the core of enterprise data strategies. And, as they vie for trust and integrity in a privacy-conscious world, business leaders are turning to privacy-enhancing technologies (PETs) that can keep their commercial data machines well-oiled without burning bridges with consumers or sparking hefty sanctions from regulatory authorities.

Dr. Maurice Coyle
Chief Data Scientist
Trūata

Yangcheng Huang
Director -
Software Engineering
Trūata



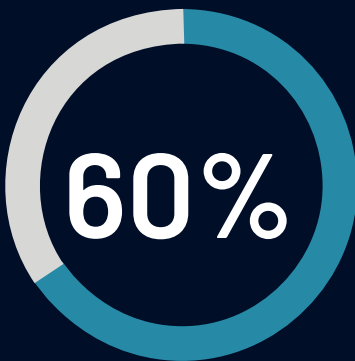
Privacy legislation worldwide has been well received with 83 percent seeing a positive impact¹

1.8x

Organizations continue to invest in privacy and estimate the return on investment on average at nearly 1.8 times the spending¹

\$8bn

Through 2022, privacy-driven spending on compliance tooling will rise to \$8 billion worldwide²



By 2025, 60% of large organizations will use one or more privacy-enhancing computation techniques in analytics, business intelligence or cloud computing³

↑ 13%

The average privacy budget has increased by 13% from 2021 to 2022 with smaller organizations (50-249 employees) increasing budgets from \$1.1 million to \$1.5 million.¹

7x

Innovations in data privacy software and privacy-enhancing technologies (PETs) are set to grow by seven times in 2022⁴

¹ Cisco ² Gartner ³ Gartner ⁴ G2

The power (and pitfalls) of emerging PETs

Emerging privacy-enhancing technologies are being hailed for their ability to help organizations minimize data protection risks (often referred to as privacy risks – the term used throughout this paper) without de-valuing data, unlock data-driven innovation and enhance digital trust in the marketplace.

However, PETs are not a silver bullet for protecting personal data. Despite the heavy investments and rapid adoption in privacy-enhancing technologies, many solutions are still leaving businesses unnecessarily exposed to risk, which may negate the cost of investing in such technologies in the first place. There are an increasing number of software and solution providers in the market that claim to remove privacy risks from data; however, more often than not, only unique personal identifiers (such as email addresses, credit card numbers etc.) are being addressed and dealt with. These “direct identifiers” or “personally-identifiable information (PII)” only make up a small proportion of the overall re-identification risk as we shall see below. Due to the sheer size and depth of information contained in today’s data assets, the majority of risks are indirect risks created by “quasi-identifiers” – and most technologies fail to address these hidden dangers.

If all sources of re-identification risk are not identified, both direct and indirect, they cannot be effectively managed. And measuring risks in datasets of today’s size requires sophisticated automation and quantified privacy technology, blended with the right expertise, to deliver a scalable solution that doesn’t compromise on data privacy or data utility.

As privacy-enhancing technologies mature and become more prevalent in the market, businesses looking towards third-party solutions will need to take the time to really understand the strengths and weaknesses of the technologies that they are matching to their business needs in order to avoid exposing or devaluing their data.

Getting to grips with the terminology:

PII versus personal data

When talking about personal data and data protection, the terms can vary according to the jurisdiction. **Personally Identifiable Information (PII)** is a term often used outside of Europe, particularly in North America, whereas **personal data** is the term generally used in Europe due to prevailing European laws, most recently the GDPR. In a nutshell, personal data under the GDPR is a broader term that includes PII and non-PII. So, while an IP address or a device ID is likely to be considered non-PII in America, it is usually considered personal data under the EU’s GDPR.

What is PII?

Personally identifiable information (PII) has been defined as:

“Information which can be used to **distinguish or trace an individual’s identity**, such as their name, social security number, biometric records, etc. **alone, or when combined with other personal or identifying information** which is **linked or linkable to a specific individual**, such as date and place of birth, mother’s maiden name, etc.”⁵

What is personal data?

Personal data has been defined under the GDPR as:

“Personal data” shall mean **any information relating to an identified or identifiable natural person** (‘Data Subject’); an **identifiable person** is one who can be identified, **directly or indirectly**, in particular by reference to an **identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity** of that natural person.”⁶

5. [US Department of Commerce: Office of Privacy & Open Government](#)

6. <https://gdpr.eu/eu-gdpr-personal-data/>

Why do quasi-identifiers matter?

More and more high-profile data breaches caused by quasi-identifiers have been hitting the headlines in recent years, showing just [how damaging it can be to assume that data has been effectively de-identified](#) or anonymized by simply mitigating direct identifiers within a dataset.

From [companies nonchalantly asserting that their data is de-identified](#), non-personal or even anonymous because it does not contain a 'name' column or other PII, to privacy oversights by the likes of [Netflix](#) and [AOL](#) showing how easy it can be to trace information back to specific individuals using a small number of attributes, the failure to address quasi-identifiers has left the identity of millions of people exposed. And these incidents are becoming more prevalent as awareness grows around indirect risks, with many organizations and tech giants coming under fire for their failure to properly protect consumers privacy and coming under attack from those who know how to exploit the vulnerabilities.

One of the more recent examples circles around [Facebook targeting](#), where a new research paper highlighted that identifying an individual becomes a simple game of probability when you have multiple 'clues' – or multiple indirect data points – that can be put together to single out a unique individual from its **2.8 billion users**.

“The results from our model reveal that the 4 rarest interests or 22 random interests from the interests set FB [Facebook] assigns to a user make them unique on FB with a 90% probability”

Quasi-identifiers, therefore, pose significant risks to businesses and to people; they also present a major challenge for emerging privacy technologies that are designed to bring about balance between the need for data analytics and the need to protect the rights of individuals in a digital-first world.

Quasi-identifiers are typically types of information that are extremely important when analyzing a dataset to derive insights that can help make informed business decisions, so simply removing them can considerably impact data utility and stunt data-driven innovation. In order to preserve maximum data utility, the risk of re-identification must be reduced to an acceptable level that also protects individuals' privacy.

What is an indirect identifier / quasi-identifier?

Indirect identifiers are types of information that cannot identify an individual on their own but could identify an individual when used in combination with other information in a dataset. For example: country of birth, race, religion, sexual orientation, employment information, health-related data, financial information, information about behavioral events such as the time, location and amount of a retail transaction.

Indirect identifiers are also referred to as **quasi-identifiers**. When certain quasi-identifiers are combined together, they can uniquely identify an individual – much like a fingerprint. For such reasons, most data protection regulations outline that, in order to have a measurement of risk, it is critical to understand the likelihood of re-identification based on quasi-identifiers.



Inferences may still be drawn from the dataset, especially if attributes are correlated or have strong logical relationships.

[ARTICLE 29 Data Protection Working Party](#)

A study found that **99.98%** of individuals could be correctly re-identified in any dataset using 15 demographic attributes⁷

99.98%

Successful re-identification

A 3-month study of the credit card records for 1.1 million people shows that just 4 spatiotemporal points are enough to uniquely re-identify **90%** of individuals⁸

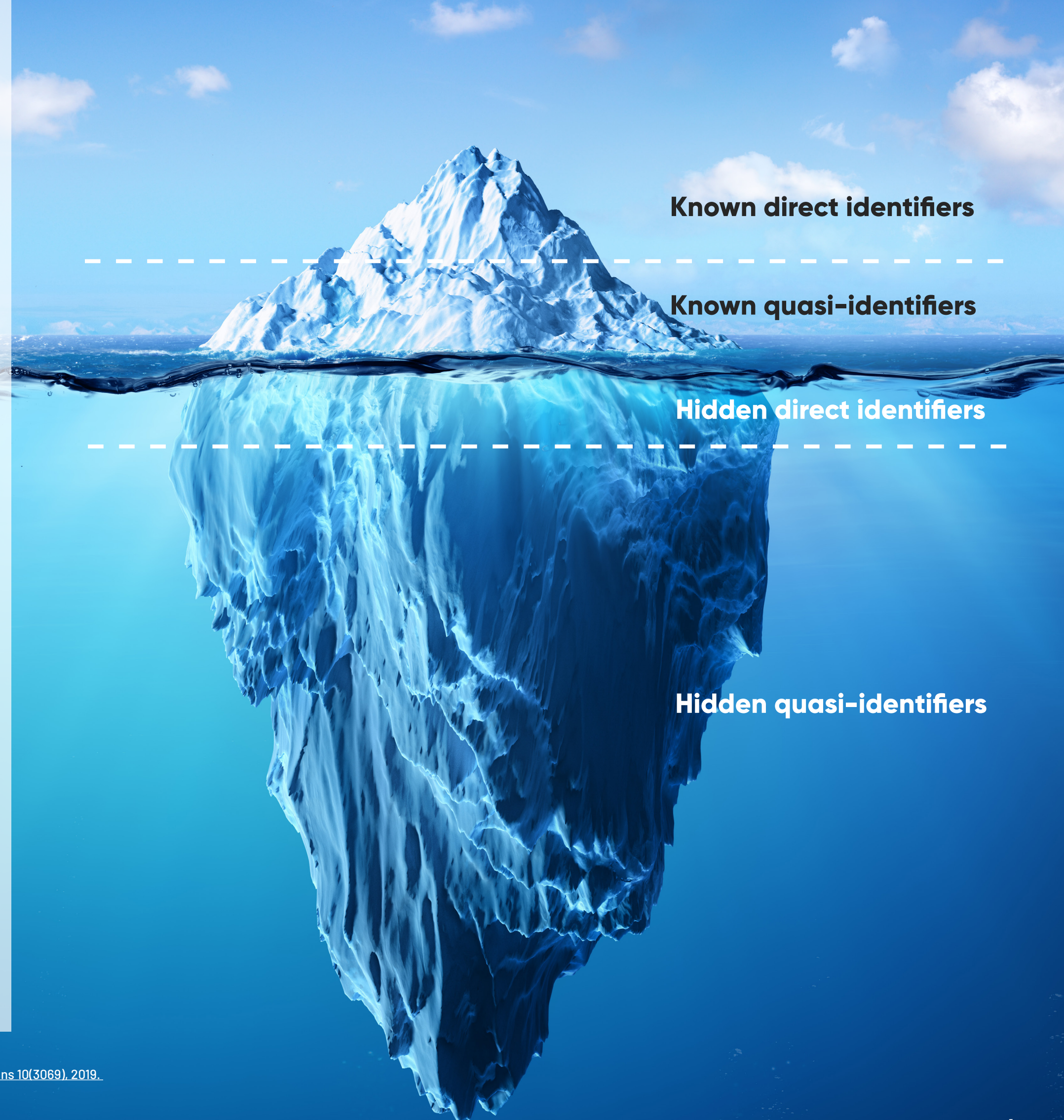
90%

Successful re-identification

Based on only 3 parameters (gender, date of birth and a 5-digit ZIP), a study found that it was possible to uniquely identify **87%** of the entire US population⁹

87%

Successful re-identification



⁷ Estimating the success of re-identifications in incomplete datasets using generative models, Rocher et al. Nature Communications 10(3069), 2019.

⁸ Unique in the shopping mall: On the reidentifiability of credit card metadata, De Montjoye et al. Science 347(6221), 2015.

⁹ Simple Demographics Often Identify People Uniquely, Latanya Sweeney, Carnegie Mellon University, Data Privacy, 2000.

The risks of disclosing personal data

When datasets contain personal information that can identify an individual, there can be conflict between the commercial goals of data use and the privacy protection of individuals. De-identification attempts to resolve this conflict by removing or altering some of the data attributes that would identify an individual, while allowing useful information and detail to remain.

When executed effectively, de-identification ensures that data cannot be matched to the person it is associated with. As mentioned, much of the hype around the de-identification of data—and many of the privacy-enhancing technologies that offer de-identification—pivots around the detection and mitigation of **direct identifiers or PII**, when the greater risks lie in the indirect (or hidden) quasi-identifiers. Similar risks should also be flagged in the growing synthetic data market, whereby many solution providers claim to have developed technology that can reproduce ‘fake’ or ‘dummy data’ by observing real statistical distributions. However, if these datasets are statistically similar to the original then they will almost certainly have reproduced risky quasi-identifiers.

The direct risks that are known, and can be seen, can be managed effectively, but **all** datasets have hidden sources of risk that outnumber these direct risks – even synthetic and pseudonymized datasets. We will examine the extent to which this occurs below.

If objective quantification of re-identification risks is not performed before and after privacy-enhancing technologies are used, it is simply not possible to know if the desired outcome has been achieved and, therefore, whether a dataset can safely be used for a particular purpose.

Figure 1 shows a dataset that does not contain any direct identifiers and contains an ‘age’ field that has been rounded to the nearest decade. Although this dataset could be considered well-protected, and some may even claim it is “anonymized”, it nevertheless contains re-identification risks. For example, all *age-sex-country* combinations are unique, as are all *age-sex* pairs. If another dataset containing direct identifiers existed, or some other source of information about the individuals within the dataset, then they could be re-identified, and their sensitive health information exposed.

	Age	Sex	Country	Diagnosis
1	30	M	USA	Heart disease
2	40	M	GBR	Kidney damage
3	20	F	GBR	Lung cancer
4	20	M	GBR	Heart disease
5	40	F	FRA	Asthma

Figure 1: A dataset that does not contain any direct identifiers may still allow for re-identification of individuals

Managing risk in the algorithmic age

Most companies understand the importance of managing big data risks. However, understanding **what they need to do** to protect data privacy and understanding **how to do it at scale** is where the gap in knowledge can leave organizations – and their data – vulnerable.

This business conundrum has led to a rapid rise in the adoption of privacy-enhancing technologies to assist with critical data challenges. But while there are many software solutions in the market that can identify and mitigate direct identifiers in datasets, these technologies still fail to address the indirect identifiers which makeup the biggest proportion of the risk.

Even with relatively few columns in a dataset, the number of possible quasi-identifiers is beyond the human capacity to detect and manage with any degree of confidence. When handling a dataset with just 50 columns, there are more possible combinations of columns than there are stars in the Milky Way galaxy. So, with the sheer volumes of data that businesses are managing today, it's impossible to expect humans to wade through that data and accurately detect all the risks.

In the era of big data analytics, businesses need a solution that enables them to measure and mitigate both the direct and the indirect privacy risks in datasets at scale. Not only this, but to move with the agility needed in a data-driven economy, businesses now require software that brings sophisticated automation to the risk management process while providing the flexibility to dynamically transform data to meet specific thresholds based on the context of the data use. Software that provides objective, consistent ways to measure and manage risk ensures confidence in data decision-making and use.

Trūata has focused its research and product development on bringing to market a next-generation, frictionless solution that empowers businesses to conduct automated, in-house de-identification that addresses both the direct and indirect risks in datasets at scale. To-date, we believe there is no software on the market like Trūata Calibrate that can as successfully and intelligently traverse this space to not only detect but also mitigate these risks.

Even with relatively few columns in a dataset, the number of possible quasi-identifiers is beyond the human capacity to detect and manage with any degree of confidence.

Beneath the surface: uncovering your biggest risks

In order to highlight just how much risk is retained in datasets when quasi-identifiers are ignored, we ran a risk analysis on data assets of varying size using **Trūata Calibrate**.

The six datasets used for analysis varied in size from 14 to 75 columns, with an average of 45 columns.



From the results of the six datasets that underwent analysis using **Trūata Calibrate** shown in Figure 2, it is clear to see that direct identifiers / PII fields make up a low percentage of the data analyzed. This ranges from 6% to 21%, with **the average dataset containing just 12% direct identifiers / PII fields**.

Singling out, linkability and inference risks still exist in the remaining 88% of columns, highlighting that the majority of the risks are overlooked if only direct identifiers are considered during a risk analysis.

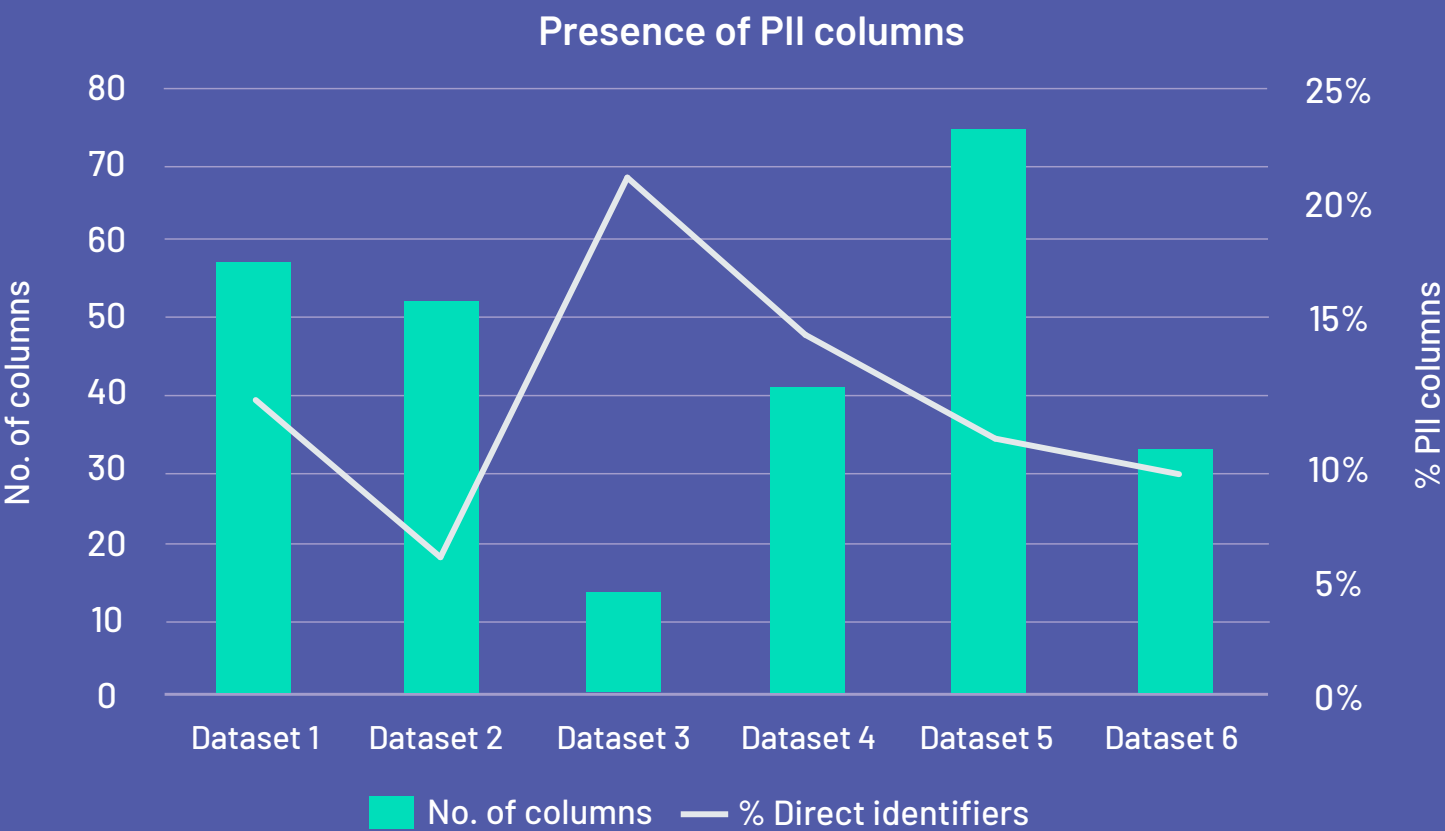


Figure 2 – Statistics that highlight the % of direct identifiers in different datasets

Anatomy of a quasi-identifier

In order to understand whether the re-identification risks present in our example datasets arise from PII fields or not, we will examine how often PII fields appear in the riskiest quasi-identifiers as found by Trūata Calibrate.

In only two of the six examined datasets, PII fields were found among the riskiest quasi-identifiers, where they constitute 50% and 64% of quasi-identifiers. Across all datasets, 6,176 (equating to 85%) of the riskiest quasi-identifiers do not contain any PII fields. This highlights the fact that if only PII fields are considered and treated as part of a risk management exercise, the vast majority of risky quasi-identifiers will be left untouched.

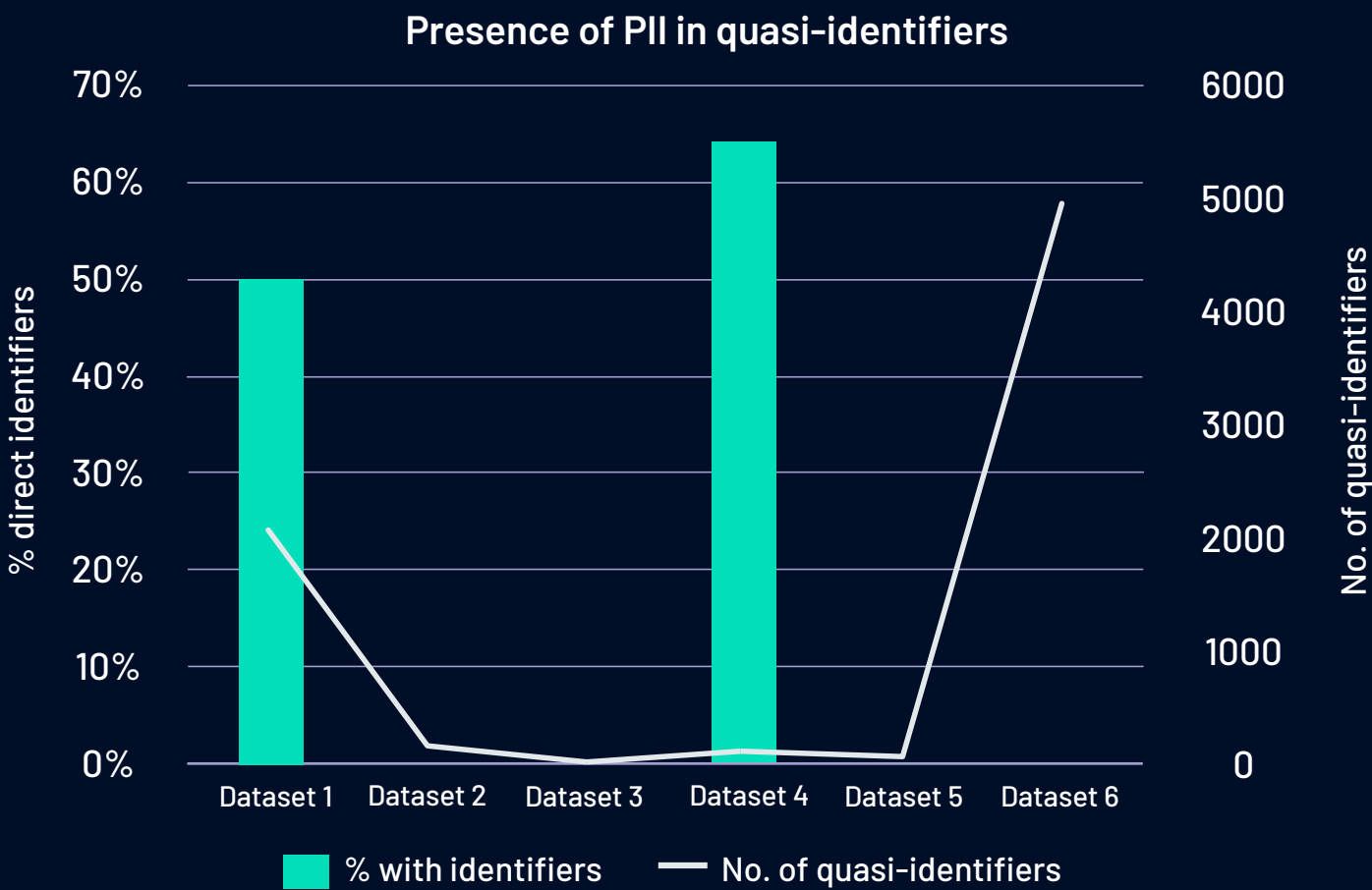
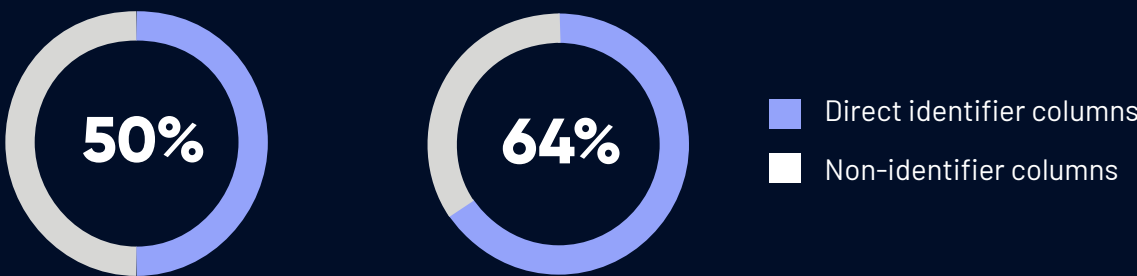


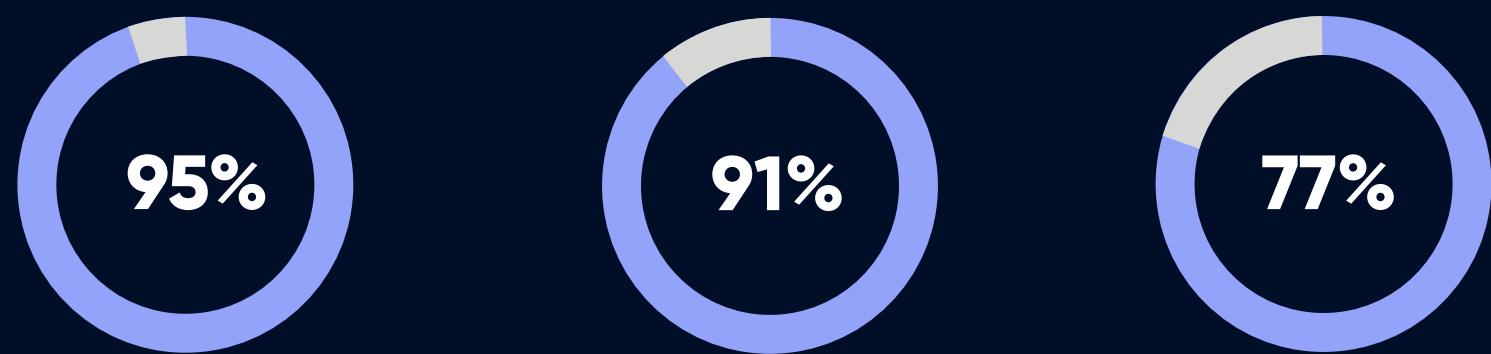
Figure 3 – Statistics that outline the risk content of quasi-identifiers

How risky are quasi-identifiers?

When running each dataset through Trūata Calibrate, it quickly becomes clear that hidden risks make up the vast majority of risk in datasets – no matter what the size of the dataset. Next we will examine how much re-identification risk these datasets actually contain.

As mentioned above, when the fields that make up the riskiest quasi-identifiers are examined, PII fields in the riskiest quasi-identifiers are only found in two datasets.

In these datasets, the PII fields were in the minority, with 91% and 95% of the riskiest fields being non-PII fields. Furthermore, up to 77% of non-PII fields appear in risky quasi-identifiers, highlighting that it is not just a few non-PII fields where the majority of risk resides.



As shown in Figure 4, five out of the six tables had an average risk score above 90% across its quasi-identifiers, with three of these having an average risk score of 99%. The dataset with the lowest average risk score contained 61% risk.

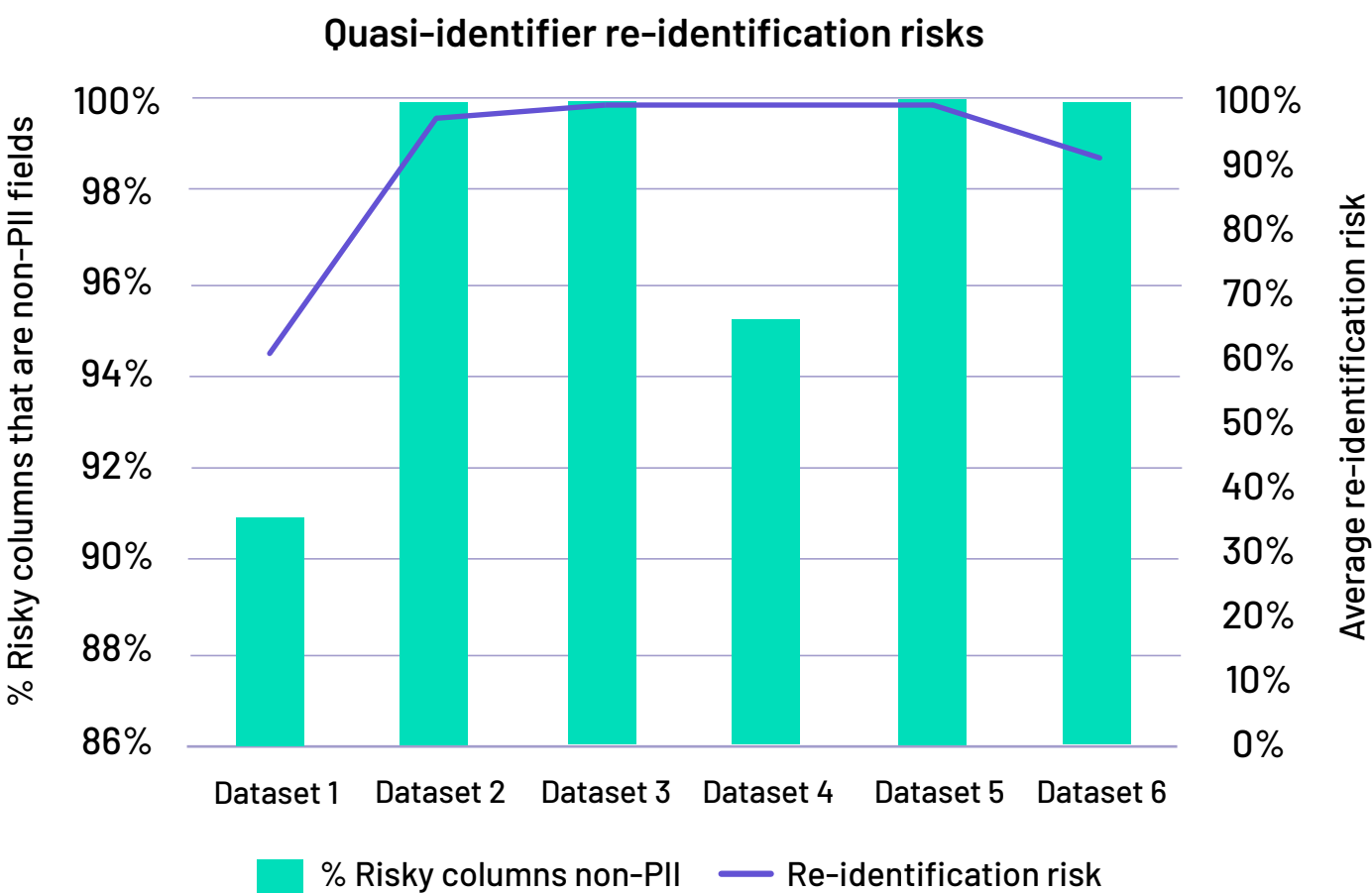


Figure 4: Re-identification risk measurements across all datasets

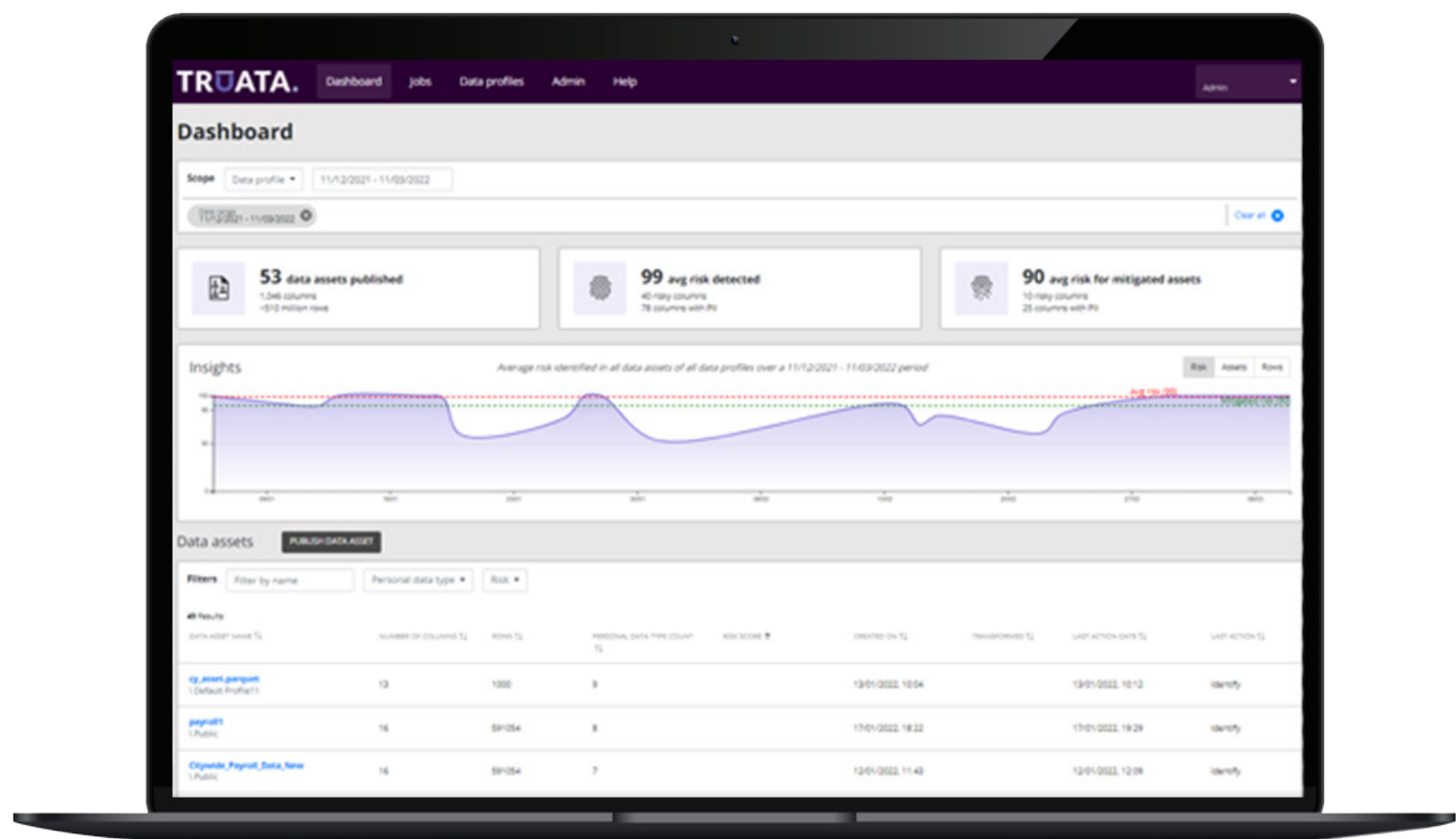
The analysis and results presented above highlight the issue that exists in the market today. If a business is only addressing the sources of direct identifiers or PII fields, then it is leaving itself vulnerable. The vast majority of high-risk fields are not PII; they are made up of hidden risks in the form of quasi-identifiers.

In failing to address such risks, businesses are leaving themselves open to privacy breaches and at risk to the resulting fines, sanctions and reputational damage.

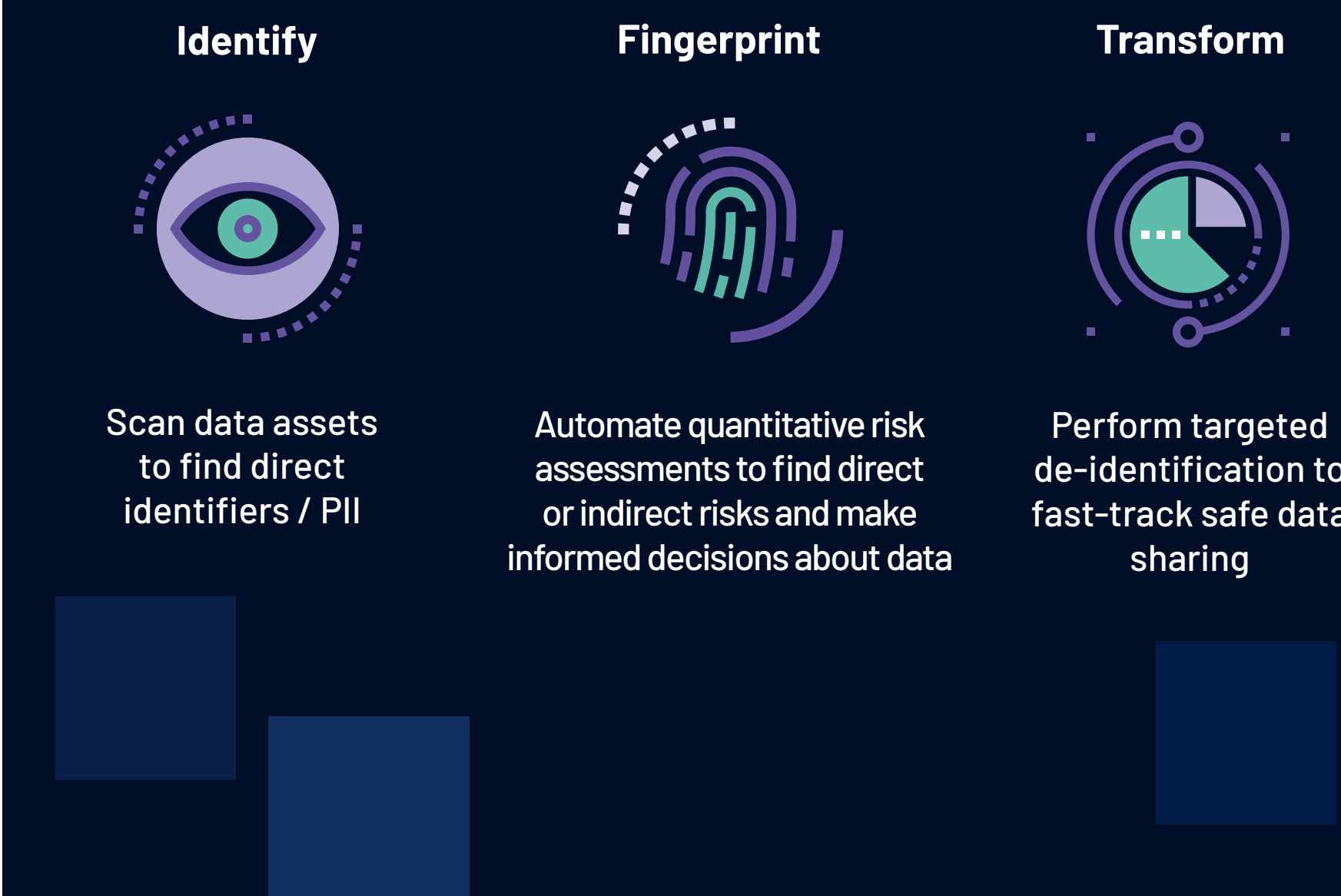
Bringing new levels of automation and sophistication to your risk analysis

Trūata Calibrate’s patented* Fingerprint technology makes automated quasi-identifier discovery possible, and this is what sets our software apart from other market players.


Powered by intelligent automation, Trūata Calibrate facilitates fast and effective risk measurement and mitigation via a centralized dashboard and easy integration with other solutions via flexible APIs. The platform provides a smart, standardized solution for managing privacy risks and ensures that data can be effectively transformed for safe use right across your business ecosystem.





* Patent pending



Get in touch

 Arrange a free demo session today: [book a demo](#)

 Speak to our team: info@truata.com

 Follow Trūata for more: 