Federal Register Notice 87 FR 35250, https://www.federalregister.gov/documents/2022/06/09/2022-12432/request-for-information-on-advancing-privacy-enhancing-technologies, June 9, 2022

Request for Information (RFI) on

Advancing Privacy Enhancing Technologies

Palantir

Palantir

ORGANIZATION INFORMATION

Organization Name: Palantir Technologies Inc. ("Palantir")
Organization Type: Industry
Organization Description: Palantir builds innovative software that helps organizations make sense of disparate data at scale. Palantir's software incorporates robust security, privacy, and data protection capabilities at every level of its platforms, allowing organizations to individually secure each piece of integrated data, set granular access controls, and maintain context-enriched audit logs of user activity. These features, baked in from the beginning of development, help ensure responsible, proportionate, and legitimate interactions with sensitive information. Organizations in national security, defense, law enforcement, health, and finance trust Palantir to safeguard, manage, and analyze their most important data assets.

We built privacy controls, data protection, and governance into our platforms from the start. Our implementation of privacy-enhancing technologies (PETs) across our platforms has helped our organization build trust with the institutions whose work relies on safekeeping information and protecting the data of their constituents.

INTRODUCTION

Data sharing and analytics is a critical component to solving many of the nation's most pressing problems: from critical emergency response (e.g., natural disasters, pandemics) and supply chain bottlenecks to carbon footprint tracking and clean energy development. While these projects vary in scope and impact, they all begin with the same fundamental question: is the relevant data available? If so, what is the quality of this data, in terms of completeness, accuracy, and timeliness? From that foundation, organizations may logically turn to questions about which PETs they might most effectively deploy to protect their most valuable assets. However, without a solid data foundation, novel PETs can be a distraction or, worse yet, fundamentally mislead, alter, or otherwise hide data quality issues and impede operational outcomes and decisions.

Data quality is a central concern because it is the foundation upon which all successful data and analytics projects are built. The application of PETs is also contingent upon data quality. A single bad dataset has the potential to compromise an entire data-driven initiative, triggering consumer distrust and rendering the effort moot. Worse, it could get into the hands of users and be misinterpreted or actioned upon.

Sometimes these lapses in data quality are simply a result of having the wrong source data (e.g., an incorrect mapping table), sometimes the data just represents a different perspective of correctness (e.g., different definitions of metrics between business groups), or sometimes it is due to mistakes in the data preparation (e.g., incorrect data cleaning logic).

When users are blind to assumptions or issues in their data, the decisions made upon that flawed data can propagate quickly throughout an organization, with successive levels of users left none the wiser. This is why data quality must be paired with robust data transparency: all users, with relevant roles and permissions, should have the full context necessary to appropriately use and trust their data.

Coupled with a strong foundation of quality data, governance and technology are equal enablers of effective PET implementation. Rigorous privacy protections coded into governance policy is socio-technical and must factor in the institutional workings as much as the pure data/technology considerations.

RESPONSES TO QUESTIONS

1. Specific research opportunities to advance PETs

Privacy-enhancing technologies (PETs) are best understood not as standalone tools, but as instruments within more complete systems. Integrated together with other PETs as well as other technical products or organizational governance procedures, PETs produce a contextually configurable, holistic data governance arsenal. Whilst innovative approaches such as homomorphic encryption may herald great promise as privacy 'cure-alls' in specific circumstances, this should not distract from the tremendous effects that can already be generated by combining a comprehensive data foundation with powerful and flexible data governance tools such as granular access controls and sophisticated data minimization regimes.

Accelerating the adoption of individual and novel PETs should therefore factor in three critical systems-level considerations:

- **Integration**: As with any software product, proper technical integration is key. Integration overhead, execution risk, and ongoing maintenance requirements can be minimized by relying on commercially available software wherever possible, rather than developing one-off integrations.
- **Governance**: Technology-driven governance is a powerful tool to promote best practices and the enforcement of PETs. For example, privacy-enhancing governance policies can and should be effectively applied, enforced, and monitored in integrated software platforms. Technology-driven governance can automate difficult or time-consuming actions, such as the correct application of policies and security controls to specific users, that might otherwise hamper the adoption of PETs.
- **Orchestration**: Software-based orchestration acts as a force multiplier for PET adoption. Instead of relying on a patchwork solution backed by complex, manual processes and policies, software can reduce total cost of ownership (TCO) and remediate the burdensome technical sophistication required by end users to implement or use PETs. In this way, commercially available software is a critical component in the adoption of PrivacyOps, which integrates teams within an automation-driven common framework that enables communication and collaboration for most important practices of privacy compliance.

Relying on a diversity of interoperable fundamentals rather than narrow, "magic bullet" technologies will be advantageous in large or complex processing environments which feature a range of heterogeneous data sources and workflows. Not all of these may be amenable to a single PET; instead, an interoperable, layered "fundamentals" approach is necessary. This approach is akin to best practices in related fields like cyber security, in which a defense-in-depth strategy adds intentional redundancy and strengthens protections against varied and sometimes unanticipated failure-modes.

Given the interconnected nature of PETs implementation and governance, future educational and training programs would benefit from the inclusion of privacy-focused governance topics. Such programs should focus on the human aspect of governance policy and implementation, as well as how technology can expedite and more consistently enforce privacy-enhancing tools, techniques, and best practices.

Like the current state of PETs, there are many ways to expand and advance research in support of PETs. These methodologies and approaches suffer from many of the same problems facing academic, medical, or other research: progress is slow, collaboration is limited, and the process is fraught with barriers. PETs research would benefit from software platforms that can enable secure, collaborative, and reproducible research, which would enhance outcomes including the development and improvement of existing PETs.

2. Specific technical aspects or limitations of PETs

PETs are only effective if the underlying data foundation is sound. Organizations need a comprehensive overview of their processing operations, including the ability to check their data for quality, accuracy, and

representativeness. In the absence of this, organizations will struggle to configure and apply PETs effectively and sustainably. This is true regardless of the specific platform architecture that a data governance team is responsible for overseeing, whether data is held in a single data lake, a federated system, or various siloed systems.

The use of PETs can be useful in connection with data protection requirements, e.g., granular access controls that enable controllers to limit processing of personal data or prompts that ask users to confirm the purpose of processing prior to accessing sensitive data. These approaches can fail when implemented either on a poorly-understood data foundation, or as standalone, "magic bullet" solutions that fail to bridge the gap from academic innovation to real world impact due to the factors described below:

- **PETs' technical sophistication**: The technical sophistication of PETs yields a number of challenges for organizations attempting to implement and enforce them across their enterprise:
  - Novel PETs are often difficult to implement and utilization without an underlying orchestration infrastructure.
  - Effective PETs governance may require semi-automation or augmentative controls over the processes and policies supported by some PETs.
- **Configuring solutions marketed as one-size-fits-all**: Such solutions often claim "out of the box" functionality that must actually be contextually situated and adapted to the particular processing and threat circumstances (e.g., the specifics of particular ontologies, data, use cases, attack vectors).
- **Lack of interoperability and the inadequacies of point solutions**: Operators are often left to navigate whether the specific PET interoperates with the other governance or oversight processes on which the system depends.
- **High set-up costs**: The set-up costs required to establish PETs may be prohibitive for certain organizations that would otherwise like to deploy PETs. The expertise to deploy PETs may also come at a high cost as the required skillsets are highly specialized and in limited supply.
- **Lack of scalability and prohibitive computational requirements**: Organizations may find that the additional costs accrued in the transition from controlled environment testing to real-world operations on large scale enterprise platforms are prohibitive.
- **Adaptability**: Organization leaders will need to assess whether their proposed PET solutions are applicable to anticipated future changes in processing operations on dynamic, changing environments. PETs that are highly context dependent and brittle should be avoided due to the up-front costs and minimal long-term viability and impact.
- **Lack of communicability and comprehensibility**: PETs should be interoperable and promote communication across the range of relevant stakeholders in the privacy-promoting space (e.g., operations, IT, legal). Effective PETs can be integrated within a large enterprise solution that can be tailored for the variety of relevant stakeholders and users to promote, rather than hinder, communication and comprehensibility.
- **Misguided understanding**: Resistance to PETs adoption may also stem from the belief that PETs can impede and introduce friction to the process of innovation. This view should be challenged: in our experience, this perception is misguided. In fact, the absence of PETs can undermine consumer and citizen trust in personal data processing in both the public and private sectors. The lack of trust and consensus can serve as a more significant impediment to innovation.

The limiting conditions of PETs adoption outlined above may be mitigated with an approach to innovation that focuses not just on the narrow purpose impacts of specific products in standalone, controlled environments, but also aims to address the full ecosystem and lifecycle of data management in complex real-world systems. In this more holistic setting, privacy risks may be better addressed through a combination of several inter-related and reinforcing technological safeguards including but not limited to:

- **Access Permissions**: Ensuring that users only have access to precise subsets of data necessary for their responsibilities.

- **Action Permissions**: Restricting permissions to conduct potentially sensitive actions, such as importing, exporting, transferring, or combining data to those users who absolutely need to do so.
- **Marking Data**: Persistently tagging sensitive datasets to clearly indicate their sensitivity, and to restrict actions such as joining them with datasets bearing other markings that may be risky in combination.
- **Obfuscation by Default**: Making data encrypted and unreadable by default. Users must enter an acceptable justification in order to decrypt necessary subsets of the data.
- **Auditing**: Empowering oversight bodies to check and verify compliance with data governance policies around de-identified data, and that no spurious, malicious, or risky actions are undertaken.
- **"Inferring" Sensitive Data**: Running background checks to infer sensitive data across the system, automatically flagging and locking down sensitive data uploaded accidentally or de-identified insufficiently.
- **Testing and Validation**: Providing the ability to do validations and "battle-test" anonymized data before it is shared more widely within the system or exported for external use.
- **Data Lineage**: Leveraging lineage tracking to understand how data is flowing within the system: which users have access to what level of identifiable data, and for what purposes at different stages.

3. Specific sectors, applications, or types of analysis that would particularly benefit from the adoption of PETs

Broadly, PETs can benefit a broad range of sectors, applications, and analyses by promoting consumer confidence and public trust. The most effective communications regarding relatively complex and niche areas like emerging PETs must directly and simply address people's basic concerns about the use of their data.

This means understanding what people's concerns are, explaining how a technology addresses these concerns, providing an approachable overview of how the tech works, what the outcomes are, and why these outcomes are desirable. This discussion of outcomes is integral to the broader adoption of PETs because it demonstrates their value beyond the privacy community. Namely, PETs are critically important because they lead to protections in contexts that carry the most risk or harms, particularly in the realm of public health and medical research.

Given the potential complexity of this subject, concise, easily understood, easy-to-consume communications are important: illustrations and video demonstrations can help in this regard. These should be posted somewhere prominent and easily available, and where relevant, should be displayed or shared by projects leveraging these PETs. This must be part of an ongoing cycle—trust in any one technology is contingent on a broader sense that the developer behind the technology is trustworthy, which requires demonstrating a continuous investment in thoughtfulness and a responsible approach to technology. These communications will be particularly effective if they are developed or released in collaboration with an already highly trusted individual or institution.

For example, scientific and health/medical researchers have demonstrated the value of leveraging a centralized research platform to manage data and its use in a secure and traceable environment. At the National Institutes of Health (NIH), data scientists, clinicians, and researchers use commercial software for organization-wide data, aggregation, harmonization, sharing, and knowledge management; data quality improvements; the use of imaging and genomics data for patient care; informatics workflows; AI/ML model management and performance monitoring; and analysis of high-throughput screening data. Organizations benefit from using commercial software because it is:

- **Quickly Deployable**: Commercial software platforms often require minimal configuration out-of-the-box, allowing rapid deployment of the platform to meet NIH needs in days or weeks instead of months or years.

- **Highly Scalable**: Commercial software is typically backed by the latest in dynamic scaling technologies, allowing it to scale in lockstep with user, compute or data size—preventing latency issues and workflow restrictions.
- **Easily Maintainable and Highly Secure**: Under the software as a service (SaaS) continuous delivery model, cloud-based software instances receive frequent updates and security patches. This software is often highly (re)configurable to reflect changing needs—and doesn't require large-scale reinvestment for its sustainability.

The benefit of leveraging commercial software to implement and enforce PETs is demonstrated by NIH's National COVID Cohort Collaborative (N3C). At the beginning of the COVID-19 pandemic, the National Center for Advancing Translational Sciences (NCATS) and the Center for Data to Health (CD2H) team required the ability to securely and rapidly integrate, harmonize, and make available clinical data relating to the COVID-19 outbreak in the United States. With a diverse and complex clinical data landscape, NCATS required a flexible data infrastructure that could integrate previously siloed data sources (e.g., Electronic Health Records (EHR)) as well as open source data (e.g., Social Determinants of Health data, US demographic data, etc.). Prior to the creation of N3C, technical barriers to data ingestion, harmonization, and secure sharing capabilities slowed critical collaboration. Furthermore, many entities had access to partial views of the current situation from their own limited EHR; no single entity had access to a comprehensive view of COVID-19 patient data needed for scientific research.

To enable scientific research of COVID-19 patient data, NCATS and CD2H partnered with Palantir to configure its commercially available software solution to support the N3C data enclave. N3C provides a secure, national resource of EHR and related data that can serve as a foundation and approach for future multi-stakeholder, multi-system medical research efforts. N3C is highly secure and governable, national research data infrastructure. N3C provides an expanding data and research asset comprised of harmonized data from 5.5 million COVID-positive patients and 815 million visits across 74 different provider sites. The research enclave enables team science among a community of more than 2,600 researchers from 280 institutions; a collaborative approach resulting in the publication of more than 37 peer-reviewed research projects in the world's leading medical journals.

As an open and interoperable data infrastructure, N3C enables researchers to overcome the complexities of COVID-19 data heterogeneity and sensitivity. Through open APIs and out-of-the-box connectors, N3C allows researchers to share and leverage lab and other health data for COVID-19-related research in open and interoperable formats. By using open source languages, researchers can also publish their work from N3C to GitHub. N3C impact includes:

- **Collaborative Science and Research**. To date, there are approximately 400 research projects in N3C run by more than 2,600 researchers. This includes projects to identify the efficacy of repurposed drugs for COVID-19, investigate how social determinants of health (SDoH) affect real-world outcomes for COVID-19 patients, and characterize the pharmacoepidemiology of COVID-19. These projects and many others are made uniquely possible by N3C.
- **Privacy Preserving Record Linkage (PPRL)**. The commercial software underlying N3C can integrate directly with PPRL PETs to securely link de-identified data across data sources. For example, N3C has used PPRL to link clinical data from EHR records with additional information such as images (TCIA), viral variants information (NCBI), mortality data, and Medicare claims. This provides numerous benefits: Analysis of x-ray images provides deeper insight into the impact of the disease on the lungs, variant data enables understanding of the clinical differences caused by different variants, mortality data promotes an accurate picture of patient outcomes, and claims data provides a complete picture of the medicines a patient is being prescribed. This enables a richer understanding of the disease while preserving the privacy of patients.
- **N3C External Dashboard**. With open APIs, N3C populates an external dashboard with data suitable for public consumption and displays that data in an accessible web portal. In this way, N3C fosters data transparency with the public to promote understanding and spur innovation and

scientific discovery.

- **Interoperability with Observational Health Data Sciences and Informatics (OHDSI) Ecosystem**. With JSON definitions that are used by OHDSI, researchers can input concept sets and cohort definitions that they defined in the ATLAS tool into N3C and then in turn export those created in N3C to other OHDSI environments.

4. Specific regulations or authorities that could be used, modified, or introduced to advance PETs

PETs can be useful tools for engendering trust in data privacy considerations required to support collaborative consortia and multi-party deployments/partnerships. However, additional technical and governance frameworks should be established to ensure risk averse institutions are confident in the efficacy of PETs and the underlying infrastructure for collaborative or other data-sharing environments. Fundamental features should include:

- **Purpose-based Access Controls**: As discussed in more detail in response to Question 6, purpose-based access controls play a critical role in fundamental trust in PETs as well as environments where PETs should be applied to protect data and promote participation.
- **Collaboration functions, to include branching and version control**: A shared environment should include clear provenance and traceability, including the ability to branch data and code as well as maintain clear version control.
- **Data integration and compounding knowledge**: Participating organizations know that they will benefit from large-scale, PET-protected data and access to compounding knowledge.
- **User friendly environment**: To ensure the broadest possible participation and utilization of PETs within a shared environment, PETs should be applicable in an intuitive manner. Any shared capability should include user-friendly security, compliance, and audit features for platform administrators.

Additionally, clarifications to regulatory ambiguities would greatly benefit the PETs community. For example, OSTP/NITRD could facilitate agency-specific or broader initiatives to define the preferred, context-specific technical means and operating procedures for the deletion of sensitive data[1]. Information blocking poses another opportunity for meaningful clarification.

As OSTP/NITRD is aware, information blocking is a practice by a health IT developer, health information network, health information exchange, or health care provider that is "likely to interfere with access, exchange, or use of electronic health information (EHI)," except in exceptions as identified by HHS. "EHI," however, is defined as "electronic protected health information (ePHI) to the extent that it would be included in a designated record set, regardless of whether the group of records are used or maintained by or for a covered entity." To qualify as PHI, such information must:

1. Identify (or reasonably could be used to identify) an individual;
2. Relate to past, present, or future physical or mental health conditions of an individual, the provision of health care to an individual, or payment for care, and;
3. Be maintained or transmitted in any form of media.

This definition presents a challenge for the use of data as it relates to some PETs, since PETs may be used to fundamentally change the nature of data to remove its PHI characteristics. As such, some PETs could render information blocking rules inapplicable. In a PET environment, the obligation for entities to share EHI could evaporate, reversing the important progress the ONC Cures Act Final Rule has (and will continue) to propel. We recommend expanding EHR Conditions of Certification (CoC) to adapt to the implementation of PETs, in whatever technical capacity OSTP/NITRD sees fit within the current

[1] An explanation of our approach to "Designing for Deletion" is below: https://blog.palantir.com/designing-for-deletion-palantir-explained-6-adfe25fda810

regulatory landscape for information sharing/blocking, which does not adequately account for the adoption and use of PETs.

In addition, the future of data analytics and PETs would benefit from expansion of information blocking regulations to cover sharing information that is not EHI. For example, it is unclear how an organization would train a machine learning model using federated learning on data accessible only through Fast Healthcare Interoperability Resources (FHIR)/US Core Application Programming Interfaces (APIs). API standards specified in 45 CFR § 170.215 were simply not designed with PETs in mind. Sharing broader data sets, even when not entirely necessary given the use case, can vastly increase the impact of analysis, and the power of that analysis to improve programmatic and organizational effectiveness.

5. Specific laws that could be used, modified, or introduced to advance PETs

Due to the high up-front cost of PET implementation, the public's tenuous perception of organizations' ability to protect privacy, and the limits of industry self-regulating, the legislative branch would be well-advised to enact policies to incentivize the implementation of PETs across key industries. Broadly, this could include:

- Establishing funding for PET prototyping and subsidies for the implementation of PETs in the next LHHS appropriations bill.
    - OSTP/NITRD could consider a budgetary authority request to fund an incentive program that encourages PET adoption. Such an incentive would not only speed the adoption of PETs, but it would advance many principles of data exchange outlined elsewhere in the U.S. Government's data policy work, for example in the health context with the Department of Health and Human Services (HHS) Trusted Exchange Framework and Common Agreement (TEFCA):
        - **Safe Harbors**: The adoption of PETs by entities liable for inappropriate information disclosure and/or use should be incentivized through the expansion of existing safe harbor regulations (such as HIPAA Safe Harbor) and the establishment of new safe harbor regulations to cover the use of PETs.
        - **Standardization**: Ensures that sensitive data will be subject to the same, and highest, level of protection that does not vary based on the sophistication of the agency establishing the PET.
        - **Openness and Transparency**: Offers clear standards to the community as to the types of security measures adopted within the PET.
        - **Cooperation and Non-Discrimination**: Establishes standard rules around access to PETs, ensuring organizations of all sizes can benefit and removing the barrier of establishing (or independently assessing) the strength of an independent PET.
        - **Privacy, Security, and Safety**: Improves access to anonymized data for analytical/machine learning purposes to occur within those secure data enclaves.
        - **Equity**: Levels the playing field by enabling organizations of various sizes and technical skillsets to securely access and analyze the data, which can also lead to important health equity data insights and improvements.
        - **Public Health**: Accelerates the opportunities for data analysis to address health disparities and public health issues.
- The inclusion of PET implementation as a condition of maintaining certification for health information technology under 42 U.S.C. § 300jj–11(c)(5). This could be included, for example, as part of a new iteration of the Cures Act (21st Century Cures Act/H.R. 34).

While information blocking regulations help unlock access to electronic health information (EHI), PETs can enable access to health data without the disclosure of protected health information (PHI)/personally identifiable information (PII), thereby enhancing patient privacy protections. As set forth in 45 C.F.R. Part 171, the Information Blocking Final Rule prohibits Actors from undertaking any practice likely to

interfere with, prevent, or materially discourage access to, exchange of, or use of EHI. Under these regulations:

- There is no actual obligation to share data once it has been de-identified (i.e., once it is no longer EHI).
- There is an obligation to share the data when it is EHI (i.e., when it is still identifiable) as and when appropriate (per relevant statutes and regulations).
- The API standards specified in 45 CFR § 170.215 were not designed, nor are they being implemented, with PETs in mind. For example, it is unclear how organizations could train an ML model using federated learning on data accessible via FHIR/US Core APIs. Instead, organizations are incentivized to share EHI (which often incurs liability for covered entities (CEs)) even when unnecessary given the applicability and availability of PETs.

Given the potential of PETs to mitigate concerns related to EHI, PHI, and PII, PETs would benefit from changes to the existing body of legislation, regulation, and policy by:

- Expanding the EHR Chain of Custody (CoC) to cover PETs implementation.
- Expanding existing Information Blocking regulations to the sharing of non-EHI information. The scope of non-EHI information will vary depending on the specific PET and should be established clearly in regulations and guidance.

6. Specific mechanisms, not covered above, that could be used, modified, or introduced to advance PETs

While much effort and attention has recently been directed towards de-centralized tools for addressing privacy risks, it is important to acknowledge that, for both technical and institutional reasons, this class of PETs has fundamental limitations. There is and will remain a critical role for centralized information and analytics environments for the foreseeable future. Centralized software suites, therefore, can be a powerful tool to increase the efficacy of data sharing and analytics to benefit individuals and society. Software can be used to effectively implement policy, enable partnerships, or create collaborative research workspaces. For example, OSTP/NITRD may recommend reliance on templatized data sharing agreements outlining risk mitigation and mutual responsibilities. Risk mitigation and mutual responsibilities can be applied to the mechanism of data sharing itself through the Governance and technical infrastructure by leveraging a central data sharing platform to track provenance and use. Trust in these programs (e.g., a data sharing agreement or partnership between public and private organizations), and the tools backing them, requires capabilities to enforce security measures and provide provenance. One such mechanism is the configuration and implementation of purpose-based access controls. Controlling data, and tracking its use across a platform, is a powerful mechanism to build trust and enable better outcomes.

This creates an acute challenge for data governance teams. Tracking who has access to what information and why, across thousands of datasets and thousands of users, quickly becomes exceptionally complex.

This challenge grows exponentially with organizational scale. As the number of access requests grows, so too does the number of potential failures. Auditing decisions to grant access is difficult if the access requests were made by telephone, email, or even in person. Data governance teams often rely on more technical colleagues to grant access to the data itself, and this can make it hard for the data governance team to check whether their decision has been appropriately enforced.

To mitigate these challenges as data scale and use grows, organizations require a close integration between data governance process and access control system. Purpose-based access controls aim to:

- Introduce structure and clarity to data access decisions.
- Capture missing context and make it available to the people who need it.
- Build intuitive tooling for non-technical data governance teams to enforce rules.

Instead of applying for access to an individual data set, a potential user applies for access to a purpose. The purpose is set by data governance teams to contain data specifically scoped to help the user meet their goal—no more, no less. Every user must apply to a purpose, and they only have access to the data that's been assigned to that purpose.

Data governance teams must record a rationale for their decision at the same time they grant a user access to data. Likewise, data owners must record a rationale when they approve the use of a data set for a purpose. Recording these justifications prompts both sides to continually consider the necessity and proportionality of their decisions. The output of that assessment can be captured in commercial software, making it available to data governance teams for review.

At any point, an auditor can understand not just who has access to what data, but also why they were given access—with all the context that went into that decision.

7. Risks related to PETs adoption

Perhaps the greatest risk to PETs adoption is inevitable disappointment and distrust that is bound to arise when PETs are touted as silver-bullet solutions to privacy challenges, but fail to deliver fully or as promised. Part of this risk is a consequence of privacy literature and technology landscape littered with artifacts of a seemingly binary categorization of data: anonymized and non-anonymized (or raw) data. The underlying presumption being that anonymized data, by being stripped of designated identifiers, has been sufficiently cleansed so as to wholly eliminate the risk of direct (personal) attribution. The problem with this binary notion is that data can almost never be fully anonymized (there will always be some residual re-attribution risk in the face of sufficiently motivated and well-resourced adversary) and the degree to which one pursues complete anonymization often will come at the price of diminished utility of the data.

We therefore recommend in discussions of PETs a terminological movement away from the use of "anonymization" (and even slightly more nuanced concepts like "pseudonymization") and instead suggest a focus on "de-identification" as a concept that captures the spectral rather than binary nature of privacy risks. Focusing on de-identification enables a more clear-eyed view of a range of techniques and spectrum of attendant impacts in reducing (but not necessarily eliminating) re-identification risks. In this framing, the following class of tools and techniques might be better conceptualized and characterized as methods for minimizing re-identification risk to varying degrees and with corresponding benefits and drawbacks:

- **Generalization**: Reducing the granularity of information (e.g., converting Date of Birth to Age or Age Range).
- **Aggregation**: Grouping data about individuals together and continuing analysis at the aggregate level.
- **Obfuscation**: Hiding or disguising identifying data to unauthorized parties, perhaps by masking or encryption.
- **Dynamic Minimization**: Showing only parts of the data depending on the needs or role of the user.
- **Synthetic Data**: Producing artificial data that replicates important underlying trends in the original data.

8. Existing best practices that are helpful for PETs adoption

Certain questions can be applied to guide the adoption of relevant, PETs implementation best practices. These questions help organizations establish PETs beyond a multitude of various de-identification techniques, most of which are subject to a tradeoff between optimizing data utility and minimizing re-identification risk. To make data less identifiable, organizations should consider:

- **How sensitive is the data?** There are many ways data could be sensitive: it could contain information on protected characteristics such as health, gender, or ethnicity; or it may be in some

other sense intimate, personal, or confidential. A related question to ask is "What would the potential harm to these individuals be if this information were re-identified?"

- **How easy is it to re-identify the data?** To answer this question, consider how unique the individual data point is, i.e., to how many individuals it could apply. The fewer people to whom it can apply, the higher the risk of re-identification.
- **What happens if it's joined with other data?** Consider the other data in your system, now and in the future. Could that data, joined with the de-identified data, result in re-identification? How likely is such a join to occur (in the system or if the data is published elsewhere)? What protections are in place to guard against it?

Organizations must also consider their internal accountability, oversight, and governance structures, and how these structural artifacts could either increase or decrease the likelihood of re-identification. Namely, organizations should consider impacts across four categories:

- **Users**: How many users will have access to this data? How will this change over time? Risk grows with each user who gains access. This is particularly important if users might be motivated to try re-identifying the data, perhaps to learn about public figures or people they know.
- **Permissions**: How much data can users access? What other data can they access (i.e., outside of the platform where they access the de-identified data), and could this be combined with the de-identified data? Do these users have permissions that would allow them to import, export, or transfer the data in unanticipated ways?
- **Policies**: Are there clear data governance policies in place, and how well does the average user understand them? Does the platform enforce these policies? Can data governance teams monitor and measure compliance?
- **Metadata**: Are datasets within the platform clearly labelled and described, so that data governance and operational users can quickly understand their sensitivity, intended use, and the applicable policy protections?

9. Existing barriers, not covered above, to PETs adoption

The greatest challenge to the deployment of PETs comes when attempting to bridge the gap from developing purely academic innovations to deploying these innovations at scale in complex, changing, expansive real world processing platforms. Additional hurdles to implementing PETs at this step include:

- **Technical Skillset Shortages**: Some of the most critical and fundamental PETs are underutilized because they require unique technical skills, ensuring that those individuals tasked with implementing PETs and best practices are almost always under-resourced. One such example is the practice of audit log analysis: even with tools and automations to support the analysis, this can be a difficult and time consuming task. Specific challenges include inconsistent log formats, decentralized logs, and an expert knowledge requirement to perform change log analysis. Broadly, this gap in PET literacy or familiarity demonstrates how so often the individuals tasked with implementing and enforcing PETs are over-worked or under-resourced.
- **Corollary Staffing Shortages**: The implementation and enforcement of PETs often falls to oversight/governance teams. PETs introduce a paradox whereby many of the most effective PETs work best when they augment human review/oversight, but without such oversight teams in the first place, it is hard to implement PET-driven protections. This demonstrates the socio-technical nature of data governance and privacy, where both technical and human controls are required to uphold data governance effectively and in context.