FUTURE OF PRIVACY FORUM

Discussion Draft

SEPTEMBER 2024

Data Clean Rooms: A Taxonomy & Technical Primer

BY AARON MASSEY

Technologist and Senior Policy Analyst for Advertising Technologies and Platforms, FPF

Forward

Data clean rooms are an increasingly discussed tool in industry, utilized in advertising and marketing, health care, and academic research, as well as for regulatory compliance. Leading thinkers have already started to analyze the capabilities of data clean rooms, including the challenges that they could address,[1] as well as the challenges that their use may raise for organizations.[2] The Future of Privacy Forum would like to recognize and thank the IAB Tech Lab and the Center for Information Policy Leadership for their prior work and analysis of data clean rooms and many of the Privacy Enhancing Technologies (PETs) that make them possible.[3] Our goal has been to build on this excellent analysis while offering new ways to think about data clean room technologies.

Data clean rooms are not a monolith. Each different formulation of a data clean room and the unique combinations of PETs that the data clean room may use is likely to implicate questions of regulatory compliance. This Primer provides a taxonomy of data clean rooms based on governance mechanisms, technological protections, and risk mitigations. We then provide an introduction to common PETs utilized in data clean rooms, their use cases, and a description of different implementation options.

The below also serves as a basis for meaningful future legal analysis of common data clean room configurations. As explained in a report on data clean rooms by the IAB Tech Lab, it is important for every party involved in a data clean room to understand their obligations.[4] However, those obligations may evolve with time and pursuant to new enforcement actions. For instance, in 2023, IAB Tech Lab indicated that "it is the responsibility of the Data Contributor to ensure that their datasets have the required compliance with applicable privacy regulations." However, the FTC's 2024 Complaint against X-Mode outlined potentially applicable responsibilities for data recipients as well, including "implement[ing] certain safeguards at a reasonable cost" and "audit[ing] the process by which its suppliers obtain consent."[5]

Understanding, deploying, and documenting data protection measures beyond contracts is increasingly critical.[6] As a next step, the Future of Privacy Forum is kicking off a project to analyze the interplay between the data clean room and global privacy and data protection regulations. Data clean rooms are likely to affect organizations' compliance with laws that limit the sale or sharing of data, restrict the transfer of data across geographic jurisdictions, and provide for the implementation of privacy principles like data minimization and use limitations.

If you have feedback on this Primer or would like to connect with FPF on our future data clean rooms work, please send a message to adpractices@fpf.org with your contact information and specific area of interest.

Introduction

This paper provides a framework for policy practitioners and lawmakers to understand the complexity and nuances of many implementations of data clean rooms and how this complexity may impact a statutory compliance analysis for organizations using them.

Conceptually, data clean rooms address an old problem: how can knowledge be learned across two data sets when restrictions exist that prevent the direct combination of those data sets? Organizations are increasingly facing privacy and security challenges that restrict sharing of personal information, whether for use by researchers, business partners, marketers or other purposes. Clean rooms offer the potential to enable a more secure data analysis and collaboration environment while avoiding risks of improper sharing of personal information. We begin by defining clean rooms and exploring their historical development. We will discuss some of the use cases for clean rooms and the factors driving their adoption. Additionally, we will examine the cost implications of implementing clean room solutions, particularly in the context of existing cloud infrastructure. Next, we examine the different models of clean rooms, comparing and contrasting their key features and use cases. We will explore traditional contract-based clean rooms, as well as more widely used modern clean rooms such as aggregation filter-only, infiltration/exfiltration protective, identity-based, and purpose-specific clean rooms. Each model offers distinct advantages and is suited for different data sharing scenarios.

To best explain how clean rooms function, we will then explore the underlying technologies that power them. This includes private set intersection (PSI), secure multi-party computation (SMPC), fully homomorphic encryption (FHE), confidential computing, and differential privacy. We will discuss how these technologies work together to protect data privacy while enabling meaningful analysis.

Finally, we will raise relevant legal questions implicated by clean rooms, including whether data clean rooms fit into common regulatory structures, how the use of a clean room may impact an analysis related to an organization's sale or sharing of data, and the implications of clean rooms for regulation pertaining to cross-border data transfer. We also include an appendix that provides more information on government-issued guidance documents detailing responses to new privacy-enhancing technologies (PETs) and, for some, providing information on new and emerging regulatory frameworks.

A comprehensive understanding of clean room technologies and their potential applications is increasingly important for data protection and compliance professionals. By exploring the different models, underlying technologies, and legal considerations, we hope to provide the tools necessary to empower organizations to start to think through the relevant legal analysis that may impact their decision to rely on clean rooms for secure and collaborative data analysis. FPF plans to continue to engage on these important questions following the publication of this introductory report and invites feedback from those researching, implementing, or analyzing the use of clean room technologies in practical settings.

Part 1: What are Data Clean Rooms?

The concept of the data "clean room" has grown in popularity over the past several years. In short, a clean room helps an organization work with others to address some data access needs if that organization doesn't have the data necessary to do so on its own, while simultaneously respecting the interests of other organizations and regulatory limitations.

Etymology and History of Data Clean Rooms

The name "clean room" was originally inspired by silicon fabrication clean rooms, where electronic components were manufactured. When making a silicon wafer, like a CPU, even the smallest speck of dust can destroy the product. All personnel and equipment entering a silicon fabrication facility must ensure they are not contaminating the facility. To that end, personnel wear cleanroom suits, also called "bunny suits", which ensure no outside contaminant affects the delicate manufacturing inside the facility. Thus, "clean room" originally referred to a place where everything was clean, and the outside world was, for lack of a better word, dirty.[7]

Data clean rooms that use privacy enhancing technologies were established to provide robust privacy protections for both organizations seeking to protect their data and for data subjects protected by laws, regulations, or data governance practices, including in some cases mathematical guarantees. Although the technologies behind clean rooms are complex and often unintuitive, they can provide meaningful protection for data.[8] This paper focuses only on the methods and techniques used for protecting privacy while maintaining utility, which is the core goal of any data clean room.

What is a Clean Room?

Data clean rooms are not like the fabrication clean rooms from which they get their name. In fact, a more accurate name might be a "data mud room," in that data clean rooms are the place to take otherwise risky analysis to ensure that it doesn't implicate other data or systems.[9] A common definition is similar to the following: "A data clean room is a piece of software that allows two companies to match data without either party getting any direct access to the data of the other party."

This definition is limited in a couple of ways. The first is that data clean rooms have been around long before software was involved. The problem they solve is foundational: needing to do analysis on the collected first party data of two or more partners. Data had to be pooled and then restricted. In older data clean rooms, data was protected contractually, sometimes using a purpose-specific corporation just for the analysis. The second limitation in this definition of data clean rooms is that it is limited to two parties, whereas in practice data clean rooms operate with many more than just two parties.

A more accurate definition might be something like the following: "A data clean room is a collaboration environment where two or more companies or their partners can perform data analysis on collective data sets and choose what they reveal to one another." This definition may, however, be too general and high-level to provide adequate description. Because data clean rooms are built from a diverse array of privacy enhancing technologies and used in a wide variety of circumstances, an accurate definition is often not illuminating.

Perhaps the clearest definition of data clean rooms comes from their most common use case: advertising. Here's how Adweek defines data clean rooms: "a privacy-compliant technology that lets separate companies combine data sets—to find matches in duplicated audiences or identify audiences—without the need for data to be transferred or stored in a centralized location." This is, of course, particular to advertising, and it also narrows the definition down to the common technologies used in digital advertising. This more narrow definition may be more conceptually useful when seeking to understand common data clean rooms deployments, but it is important to recognize that data clean rooms are composite systems with a long tail of possible use cases rather than singular or purpose-specific technologies.

What problems can a data clean room address?

To fully understand the function of a data clean room, it is useful to focus on the problems that data clean rooms solve. If a company needs to transfer data with a partner to answer a question accurately, but they both have a business interest in protecting that data, then what they need is a data clean room of some kind. Similarly, if an organization that has a lot of data that carries significant privacy interests for the data subjects, particularly if that organization is subject to privacy laws that restrict what data is allowed to be used in a calculation or disclosed to third parties, then they may need a data clean room to be able to support key business activities.

Currently, the most common problem data clean rooms are being used to solve is advertising analytics. The trend both in technology development and in recent regulation has been to bar or deter promiscuous data sharing, as has become the norm in certain markets, and toward more private first-party data sets. This makes advertising analytics more challenging, and data clean rooms are a commonly used approach to address this challenge. Two or more parties can leverage data clean rooms to get the analytics answers they need while either minimizing or eliminating their data sharing, depending on the configuration of the data clean room.

Data clean rooms cannot address all privacy risks. If the privacy concern is the ethics of the business practice, it may not be addressed by mathematically guaranteeing that the data sharing is limited, de-identified, or aggregated. Concerns about uses of data involve issues that often go far beyond narrow conceptions of data protection and involve competition, ethics, fairness, equity, and broad societal concerns. Although many of the most significant concerns about harms caused by widespread sharing of data, data breaches, revealing of embarrassing information, data sharing with law enforcement and such can be significantly minimized with data clean rooms, the broader social concerns also need to be addressed.[10] However, it is significant that risks involving data protection specific concerns can be significantly minimized and broader issues may be easier to tackle. In order to do so, the particular configuration and use of the clean room remains critical. As such, a thorough understanding of the relevant techniques and how well they serve these goals is essential for policymakers and practitioners.

How are data clean rooms being used?

Data clean rooms are used in a wide variety of industries, and this trend is growing as the use of software continues to proliferate and data analytics become more and more central to more and more business models. Technical and regulatory trends are also pushing data clean rooms into the spotlight as a critical technology for generating shared insight while limiting and controlling data sharing. Any industry that deals in broad data analysis may have some role for data clean rooms.

ADVERTISING AND MARKETING

The most common place where data clean rooms are utilized is in advertising and marketing. Advertisers often have customer data from a previous campaign, while retailers have first party data such as purchase records. Advertisers and retailers can match their data with data from a data broker or data from an advertiser (or both) to determine the success of an ad campaign. They can also want to determine who they should be targeting for their next ad campaign. Revenue sharing arrangements can be determined based on which party contributed the most to making sales while protecting individual privacy using a data clean room. Finally, they can use the data to conduct market research for new products and services or to determine what products are popular and what sorts of innovations might prove successful. But advertising and marketing are not the only areas where you would need a data clean room.

HEALTH CARE

Other areas where data clean rooms are currently being used include health care. Health researchers conducting research on the effectiveness of a treatment regimen or medication need access to comprehensive records about how the patient did without compromising patient privacy. This research must also be done without violating intellectual property agreements or data collection restrictions for the companies involved in creating the treatment.

ACADEMIC RESEARCH

Academic research is another sector where data clean rooms are currently used. Third parties providing sensitive data to researchers, such as those conducting research into the effect of social media on mental health, will want to have assurances that the data is used appropriately. For instance, the entity providing the information often wants to ensure that access is minimized to only those authorized to do this work and only for the duration of the study. Where relevant, data clean rooms may also be designed to limit researchers to examining aggregate results for cohorts of people rather than on individuals.

GOVERNMENT USE

Governments use data clean rooms for budgetary planning, municipality planning, and public health research. To these ends, governments collect detailed information about corporate and individual taxes, the people who are moving into or out of an area, the activities people do regularly, traffic data, family size, and other raw sensitive information. When governments work with third party contractors and service providers, they are often restricted by law in what data they are able to transfer to those entities. Different government authorities at the city, county, state, and federal level are also commonly restricted by law in their data sharing practices. Data clean rooms are an important solution for these circumstances, enabling more productive and efficient collaborations with government data.

REGULATORY COMPLIANCE

Finally, data clean rooms are useful even within a single organization that has two or more data sets collected with different data minimization restrictions or use prohibitions. Each data set can be treated as if it originated from a separate data partner, and a data clean room can be set up to enable whatever collaboration is possible given the restrictions on both of the original sets of data. Note that this use of a data clean room isn't about access to the underlying data because the data in both data sets is controlled by the same entity. In this case, the purpose of the data clean room is to ensure the results of any analysis respect the restrictions each independent data set must comply with.

Part 2: Clean Room Taxonomy: Four Central Models for Data Clean Rooms

The breadth and variety of data clean rooms does not lend itself easily to mental models, but can be broadly conceptualized into four central models for data clean rooms that generally represent all approaches taken by industry as a whole. These models are not cleanly separated, though each model can be characterized by its fundamental approach to the privacy and utility trade-off available with data clean room technologies.

Model Type	Governed By	Technological Protections	Availability	Best Case Risk
Contracts	Contract alone	None	Rare, primarily because of the lack of protection	High risk even in the best case scenario because of the lack of technical verification.
Contract+ (Input / Output Filters)	Contract backed by Input / Output Filters	Statistical data release tools like Differential Privacy	Common. This has become the minimally viable approach in most industries.	Moderate risk with better protection for data classification harms than for use prohibitions.
Identity Matching	Contract backed by both Input / Output filters and use restrictions limited to the identity scheme.	Data protections determined by a combination of both statistical data release tools and use restrictions determined by standard identity technologies.	Common. This is the standard approach to data clean rooms in many industries.	Low risk when using PETs for both identifying data to transfer and for the identity matching.
Custom Configurations	Contract backed by filters, data release algorithms, and mathematical guarantees of limited analytical inferences.	Depends on the purpose of the clean room, but the ideal case is able to ensure strong use prohibitions in addition to standard privacy protections through custom SMPC or FHE implementations.	Rare. Most standard business data flows do not need this level of protection, but many of the scenarios that require it would not be possible without it.	Low risk in the best case, but fully understanding the problem space and building out a solution is not a trivial task. The biggest risk may simply be implementation complexity.

Model One: Contracts

As explained in section one, data clean rooms are not new. Organizations have needed to transfer data to one another for decades, and this sharing has always been something that required some protections. Before the development of more sophisticated technology, restrictions were often provided through contractual terms. These contracts sought to ensure that both or all parties to the collaboration understood what requirements they were operating under and were satisfied. Thus, the first model is the least technology intensive solution to data clean rooms.

Typical data collectives using contracts set up a contract that allowed for everyone to pool data into one giant database, conduct any analysis, and then delete whatever access they had to the pooled data. Contracts can specify what analysis may be done, what each party can learn from the analysis, and when the sharing will end. However, contracts are limited in terms of their ability to protect data from other parties to the collaboration or against insider threats, which does limit the utility of this model. Even if privacy regulations do not restrict data sharing, first-party data is often a critical asset that organizations seek to protect. Participating organizations may not have robust technical reassurance or verification regarding use of the data that violates the contract. That said, these collaborations still happen both because they are easy to set up and because of their low overhead costs.

Although the sharing in this model is enforced only through contract, that does not mean technology is not important in the successful completion of a data clean room based on contracts. Parties involved must still take standard precautions to protect data from data breaches and may also have to consider purpose limitations and data minimization. Access control mechanisms, along with auditing requirements, should also be implemented. Finally, planning for incident response may also be required. Technology and technical expertise is critical for each of these, although it is more similar to the precautions to be taken when setting up any data storage and does not contain unique aspects related to data clean room technologies.

Contract-only clean rooms between two direct parties are rare in our modern data protection environment, and many discussions of clean rooms simply omit them. However, they are still important to note for three reasons. First, all other approaches to clean rooms still rely on a contractual analysis to ensure the technical details are correctly and fully understood. The use of PETs does not eliminate the need to perform a diligent contractual analysis. Second, organizations that hold two or more data sets collected under different terms of use may actually be operating in practice under this model. Spending time and money on a clean room that provides technical protections for an analysis is less attractive when all the data is held by a single organization and competitive data sharing concerns do not apply. Third, contracts are often the method of compliance when a neutral third party provides a matching service to facilitate matching of data sets for further analysis or activity. For example, a third party mailing house may facilitate one party sending direct mail to the customers of another party, without providing information back to either party and not retaining the data sets.

Model Two: Contracts Plus: Input and Output Filtering

The second model for data clean rooms, which may be called “contracts plus”, is similar to the first one but with added technical measures. That is, data clean rooms in this approach also have a contract, and parties to the contract must also be aware of and prepared for the standard data security and privacy practices required for their jurisdiction. Additionally, participating organizations still pool all of their data into a database, and the data is still available in a raw format where participants with access to the database can perform the analysis they need. The most significant difference is that in this model, technical measures protect both inputs and outputs. These technical protections generally fall into one of three categories:

1. A restriction on data input: Limiting either the amount of data that can be added or when data can be added to the collaboration.

2. A restriction on queries: Limiting either the attributes that can be queried (i.e., no queries over quasi-identifiers), limiting the total number of queries, or limiting the frequency of queries.

3. A restriction on result sets returned: Limiting any results returned to aggregations rather than result sets with individual records or limiting result sets to those that have been filtered with some statistical privacy protections such as differential privacy or k-anonymity.

The integration of technical measures can serve a number of purposes. For instance, technical measures that moderate the number of times you can run the same query over the data can be used to prevent revelation of details that could be inferred based on changes that may have happened to the underlying data between the queries. Because of the integration of these technical restrictions, this model is more complex than the first model, but it remains broadly similar in the sense that the analysis is still conducted over essentially raw data.

Collaborators in a contract plus model are not required to use any of the technologies that you may see in more protective environments. For example, these collaborations may not use differential privacy, fully homomorphic encryption, or secure multi-party computation. In fact, many technical solutions are essentially just a formally defined interface from one storage solution to another. For example, an organization that stored all their data in an on-prem structured relational database may wish to collaborate with an organization that primarily uses a cloud-based document store. If they wish to collaborate, then they would need some interface for this collaboration.

The interface does not have to copy or store data from either partner, but it must be able to pull from both data sets to answer queries, such as those that motivated the collaboration in the first place. For example, if a business wants to conduct a retargeting-type advertising campaign on an online platform without revealing all of their customers' contact information, then the interface must be able to determine which customers also have accounts on the platform. One contract+ approach to this problem would be to limit customer information provided to those customers who have affirmatively stated that they have an account on the platform in question.

A contract plus approach can be technically challenging, particularly as the number of partners or the number of different data storage approaches increases. Some of the more sophisticated solutions have dedicated more resources to ensuring interoperability and security than they have to advanced privacy protection mechanisms.

The reason for this is that many technical solutions on the market are primarily offering a collaboration interface from one data warehouse provider to another. For example, if one organization stores their data in a graph database and they want to partner with an organization that stores their data in a document store, then they will necessarily have to have an interface built to perform analysis over the collected data. This can be technically challenging, and some of the more sophisticated solutions have dedicated far more resources to ensure interoperability than they have to advanced privacy protection mechanisms.

Model Three: Identity Matching with PETs

Model three is all about identity matching, and there are two competing general approaches to doing this. Regardless of approach, model three data clean rooms build on both previously discussed models of clean rooms; they leverage contracts and they may use technical measures to restrict inputs, queries, and results. Identity matching clean rooms are more narrowly focused in terms of the use cases for which this model is an ideal fit. In fact, this model is, of the four models, the closest to an advertising-oriented solution in the data clean room space. Identity matching clean rooms are designed to match between two or more data sets based purely on identity, whether matching one-to-one or one-to-many or whether matching deterministically or probabilistically. If the purpose of the collaboration is to target advertising or measure campaign activation rates, then the only results required take the form of matched identity records across

the parties involved. Identity matching clean rooms are the most widely deployed commercial data clean room today. Additionally, deployments are increasing in part because of industry transitions away from widespread universal identifiers like the third-party cookie.

Two distinctly different approaches have formed to address this need. The first is an identity-oriented solution in which clean room vendors use or develop deterministic, probabilistic, or hybrid identifiers and then perform matches in clean rooms with participating organizations that also use these identifiers. The second approach is attribute-oriented solution: without a strong universal identifier, identity matching clean rooms must join first-party data sets by matching, either deterministically or probabilistically, other attributes of the records in the dataset.

As with much to-do with data clean rooms, the dichotomy between identity-oriented solutions and attribute-oriented solutions is a somewhat blurry line. Many identity-oriented solutions can be derived from attributes, so retailers who have been collecting first-party data for decades can begin using identity matching clean rooms either by performing an attribute-oriented match or by transitioning their identity solution and then using a clean room designed around it. Many opportunities exist here, in part because retailers who have collected large amounts of personal data in a first-party context may now be better positioned to compete with third-party data ecosystems that must account for additional privacy compliance considerations when sharing data with third parties. The third-party cookie made it cheaper and easier to match identity data with a wide variety of third parties. The ongoing industry transition away from third-party cookies and towards other approaches for digital advertising has driven many retailers to start their own advertising business based on their own first-party data.

Identity matching clean rooms can satisfy several different organizational goals. One goal would be to allow advertisers and publishers to use their first-party data, along with data purchased from other organizations, to define an addressable audience and activation criteria for that audience. This is an extremely common goal in advertising, and why identity matching clean rooms are the most common model for clean rooms in use today.[11] Another reason identity matching clean rooms are used is for organizations to learn more about their own customers or members, either for market research or for data enrichment. For example, an organization may be able to discern a new demographic that appears particularly interested in their products, which they may then be able to build an advertising campaign to target. Organizations able to identify new potential customers by matching against another organization's first party data may find similar value in this method of clean room analysis.

Finally, organizations may seek to match on identity for straightforward attribution and measurement of advertising campaigns. Campaign optimization or measuring Return on Ad Spend ("ROAS") both depend on being able to connect views of advertisements to purchases, but purchase data may only be available to first-parties not directly connected with the advertisement view. Of course, first-party retailers do not want to provide unvarnished access to their data. An identity-focused data clean room is a natural solution.

Model Four: Custom Clean Room Configurations

Although each of the first three models of data clean rooms built on the capabilities of the previous model, the fourth model is best thought of not as the capstone of this progression so much as a boutique collaboration, a unique opportunity that may not have a clear market beyond the participants involved.

Given this, it is unsurprising that custom data clean rooms may be built for many distinct reasons. For example, custom clean rooms may be the result of large legacy databases that are not compatible with the identity-oriented solutions offered by identity matching data clean rooms. When identity-oriented solutions are not possible, analytics based on other data attributes may still be desirable.

Many clean room providers operate as a service that requires some customization precisely to fill in the gaps where other models fail to address organizational needs. This is not dissimilar to data warehouse providers,

who often send staff on-site to their customers for weeks at a time to set up large retail infrastructure and assist with a transition from an older system.

A custom clean room approach may become more common as artificial intelligence becomes more embedded in the world of data analytics. For example, a large retailer with enough data on browsing history and purchasing patterns may be able to use a feature vector with a confidence interval as a means of categorizing users for analytics without relying on more traditional identity solutions.[12] Using a confidence interval allows analytics to match data between two people who are both likely to make a purchase even if they don't have an exact match for the predicted likelihood in their feature vectors. This approach enables more accurate analytics and insights when organizations seek to answer common marketing questions, like how many people in my dataset would be likely to purchase Chicago sports apparel, without data exchange and without relying on identifiers or learning anything about the other datasets in the clean room.

Custom clean rooms may also be used when the participants want more technical protections than would be available in identity-focused data clean rooms. Consider a healthcare research scenario where the actual name or marketing identifiers of the individuals involved is less important than the data involved in their medical treatment. The goal is to produce research that is useful for the medical community while not putting individual patients at risk by exposing their participation in the study or revealing the details of their diagnosis. This is not always an easy trade-off and often involves some quasi-identifier (i.e., a data point that partially identifies a person) available in the data set that would need to be protected through aggregation or a more formal privacy protection technique like differential privacy. Medical research that relies on perfectly accurate individual data points may require full access to all the data through solicitation of consent for use and sharing of the data. For areas where custom configurations of clean rooms for medical research provide protection without compromising research value, customizations are not likely to evolve into a new common configuration that is generally applicable to medical research and will continue to require customization with data partners for each new collaboration.

Part 3: Technologies Behind Data Clean Rooms

Many modern data clean rooms are complex systems built upon a foundation of diverse privacy-enhancing technologies (PETs). These PETs serve as essential building blocks, each with its own strengths, weaknesses, and tradeoffs. Adoption of more technically complex clean room implementations is being driven both by regulatory developments and the need for technical protections of business interests. This Part examines five core PETs commonly used in data clean rooms: private set intersection (PSI), secure multi-party computation (SMPC), homomorphic encryption, confidential computing, and differential privacy.

Modern data clean rooms leverage a combination of PETs to achieve their privacy-preserving goals. PSI, SMPC, homomorphic encryption, and confidential computing provide the computational foundation, while differential privacy and other attribute protection techniques contribute to input privacy and output utility. Understanding these technologies is crucial for effectively evaluating, implementing, or utilizing data clean rooms.

Technology	Definition	Issue Addressed	Implementations
Private Set Intersection (PSI)	A secure multiparty computation designed to identify the data all parties share in common without revealing additional details to the parties involved.	Matching records without sharing complete data sets.	The cryptographic primitives underlying implementations of PSI take typically one of three common approaches: cryptographic hashing, oblivious transfer, or secure multi-party computation
Secure Multi-party Computation (SMPC)	A family of algorithmic techniques for performing secure calculations over data provided by two or more parties and returning results with mathematical guarantees restricting the information transferred.	Provides a more general solution than PSI in that it can perform calculations on the attributes of records without sharing data sets.	Many implementations exist, falling into two primary groups: garbled circuit implementations for two party computation and multi-party protocols

Technology	Definition	Issue Addressed	Implementations
Homomorphic Encryption (HE)	A form of encryption that allows for computation on encrypted data without requiring the data to be decrypted.	Provides strong protection for scenarios where data must not be decrypted in any way on any system.	Four basic implementations exist: partially homomorphic, somewhat homomorphic, leveled fully homomorphic, and fully homomorphic encryption
Confidential Computing	A hardware-based technique for protecting data in use through cryptographically guaranteed data isolation.	Provides a computationally efficient approach for protecting data in use, but requires attestation of data and analytical techniques shared.	Hardware support available from most major chip manufacturers, with essentially two levels of protection: full virtual machine protection and per-process isolation.
Differential Privacy	A family of algorithms designed to protect statistical data releases by adding precisely calculated noise that provides protection for individual data points while maintaining aggregate statistics for the set as a whole.	Protects individual data points in data sets that must be transferred in an unencrypted form to be useful.	Many implementations exist, protecting a range of data release scenarios and addressing both individual data points as well as groups or subgroups within a larger data set. The family of algorithms used in differential privacy can be thought of in two categories: global techniques and local techniques.

Private Set Intersection

Private Set Intersection (PSI) is a cryptographic technique essential to modern data clean rooms. Its core function is to identify common elements between two or more datasets without revealing the entire contents of either set, particularly sensitive information, to the parties involved. PSI addresses a conceptually simple problem: how two parties can learn what data they have in common without revealing all of their data to one another. This is not only possible, but it can be done with several methods and potentially many partners. It also remains an area of active research in the computing community.

Traditionally, discovering common data elements involved exchanging complete datasets, a process that can create significant privacy risks. PSI offers a solution by identifying matching elements using cryptographic primitives that make it impossible for an involved party to learn any information about the non-matching records that may be held by each other party.[13] This is particularly valuable in scenarios where data sensitivity necessitates restricted access, which is becoming a common regulatory restriction.

The cryptographic primitives underlying implementations of PSI is typically one of three common approaches: cryptographic hashing, oblivious transfer, or secure multi-party computation. Cryptographic hashing works in scenarios where the attributes to be matched can be identified in advance and a perfect match is all that is required. For instance, a lumberyard seeking to advertise on a social media platform can use a PSI based on cryptographic hashing of customer email addresses to identify customers who have registered for the social media platform using that same email address without disclosing their entire customer base or the social media platform revealing all of its user details. This is the simplest form of PSI as it only involves two datasets with identical matching requirements, and comes with some limitations. For instance, if an email address in one database includes an extraneous character (like a period at the end), and the email address in the other database does not, it would not be flagged as a match. These drawbacks can be more significant in other use cases with more varied data, and many real-world applications do not neatly conform to such idealized conditions. Datasets often vary in size, schema, and data quality. For example, people may use different email addresses in different contexts even when using other identifiers, like a physical address, that match. These complexities require more sophisticated PSI algorithms that can handle imperfect data and provide probabilistic matches.

The second common technique underlying many PSI implementations is oblivious transfer, which refers to a set of private information retrieval techniques where one party sends multiple pieces of information to one or more other parties and remains unaware which one of these pieces were actually used by the recipients involved in the computation.[14]

Oblivious transfer is also widely used as part of several implementations of secure multi-party computation (SMPC), which is the third common technique underlying many PSI implementations. Although SMPC can be used for arbitrary computations and is commonly used in other data clean room operations, it can also be used for PSI applications that require a more sophisticated matching algorithm than simply using cryptographic hashes or oblivious transfer on its own.

Despite advancements, PSI remains an area of active research and development. Its implementation often necessitates custom solutions tailored to specific use cases. The absence of a general implementation framework underscores the need for careful data preparation and algorithm selection when applying PSI in practical scenarios. PSI has limitations and it is essential to recognize them in advance. The accuracy of matches can be influenced by data quality and algorithm parameters. Additionally, as the number of parties involved increases, the complexity of PSI calculations grows, potentially impacting performance and in some cases affecting privacy guarantees or thresholds. Implementation options may also have implications for privacy, security, and regulatory compliance, which is why they should be evaluated by privacy technologists and legal professionals.

Even with these limitations, PSI is a cornerstone of many data clean rooms, enabling secure and privacy-preserving data collaboration. Its versatility and adaptability to diverse data environments make it a valuable asset for organizations seeking to extract matches from common data without compromising other sensitive information.

Secure Multi-party Computation

Secure multi-party computation (SMPC) is a cryptographic technique that enables multiple parties to collaboratively compute an output to a function over their combined inputs without revealing individual contributions to any party involved. Emerging in the late 1970s, SMPC has become a critical subfield of modern cryptography. Unlike traditional cryptographic methods that focus on securing data in transit or at rest, SMPC ensures the confidentiality and integrity of data throughout the computation process. This implies that no party, including those involved in the computation, can access the raw data provided by others. SMPC has seen significant advancements in the last two decades, with both generic and specialized approaches emerging. While generic protocols offer flexibility, specialized protocols often provide superior efficiency. Notably, SMPC has been widely applied in auction and voting scenarios, and its potential in secure machine learning is increasingly explored.

The versatility of SMPC is underscored by its ability to compute any arbitrary function, which is why it can be used as part of private set intersection but also for more complex tasks. This characteristic makes it a valuable tool for data clean rooms, where complex analyses often extend beyond simple data intersections. For instance, calculating aggregate statistics, such as average salaries by gender across multiple Fortune 500 companies, can be achieved through SMPC without compromising the privacy of individual salary data. Of course, the calculation could be anything. An average is a relatively simple calculation, but SMPC can handle arbitrarily complex calculations. This capability is crucial in an era where data collaboration is essential but mitigating privacy risk is paramount.

SMPC has spawned a diverse array of protocols and implementations. Broadly categorized, these approaches fall into two primary groups: garbled circuit implementations for two party computation, sometimes abbreviated 2PC, and multi-party protocols. Garbled circuit implementations, which only work when there are only two parties, are constructed by the sender and evaluated by the receiver in a manner that reveals only the output of the calculation, not the circuit's internals or the original data provided as an input.[15] Once evaluated by the recipient party, the recipient repeats the process in reverse to ensure both parties are able to share the full output of the computation.[16]

Multi-party computation (MPC) protocols, in contrast to garbled circuits, employ alternative cryptographic primitives.[17] MPC methods have received a great deal of attention in the last two decades as a result of its use in creating cryptocurrency wallets and related technologies seeking to perform proof of work calculations. Although several MPC methods are commonly used, they generally provide privacy benefits comparable to those used in the 2PC example above and differ in that they are more complex and less computationally efficient. This complexity and inefficiency may prove prohibitive depending on the business use case.[18] Given the complexity and nuances of SMPC, regardless of the number of parties involved in the computation, consulting with cryptography experts is advisable when selecting the appropriate protocol and implementation approach.

Homomorphic Encryption

Homomorphic encryption is a cryptographic technique that enables computations to be performed directly on encrypted data without requiring decryption. Unlike many other privacy-enhancing technologies, it operates in a centralized manner, eliminating the need for multi-party protocols, such as those used in PSI or SMPC. Homomorphic encryption is particularly valuable in scenarios where both data and queries are sensitive, such as cloud-based database searches.[19]

A quintessential example involves a cloud-based search engine that allows users to conduct encrypted searches against an encrypted dataset and returns an encrypted result set. The cloud-based host would be mathematically prevented from discerning the search terms, retrieved data, or results, as all computations occur on encrypted information.

Identifying potential drug interactions is a practical use case where homomorphic encryption demonstrates its value. If two pharmaceutical companies are interested in determining drug interactions without revealing sensitive patient information or proprietary pharmaceutical developments, they can use homomorphic encryption to conduct the potentially complex calculations necessary to return the desired information. This would not be possible with more limited techniques.

Today, homomorphic encryption encompasses four primary implementation categories: partially homomorphic, somewhat homomorphic, leveled fully homomorphic, and fully homomorphic encryption.[20] Fully homomorphic encryption represents the most robust form of homomorphic encryption, supporting arbitrary computations without limitations. However, despite its potential, fully homomorphic encryption presents substantial computational challenges. The process of encryption, decryption, and calculation of results is computationally intensive, hindering scalability and maintainability.

Modifications to computations can be difficult and may require significant re-engineering regardless of which category of homomorphic encryption is used, limiting practical flexibility. Recall that homomorphic encryption is best used in scenarios where both the data and the queries are sensitive. Although these scenarios exist, the more common practical scenario is that the data are sensitive and the queries are not. Consequently, for many use cases where homomorphic encryption might be considered, alternative approaches like PSI or SMPC often offer more efficient and practical solutions if their limitations can be accommodated.

Confidential Computing

Confidential computing represents a distinct approach to privacy-enhancing technologies, relying on isolation rather than encryption.[21] It does this using two key technologies: trusted execution environments (TEEs) and attestation services. TEEs are separate and secure areas within a central processing unit ("CPU") that isolate data processing from the rest of a computer system. Attestation services provide cryptographic verification that the TEE is operating according to a set of rules provided when the confidential computation is initiated. By employing TEEs and attestation services, confidential computing can facilitate the creation of a secure enclave within system memory for any arbitrary computation. This isolated environment protects data as it is in use from unauthorized access by even privileged users on the host system. While initial implementations of confidential computing were limited to virtual machine-level isolation, advancements now permit isolation at the process and even AI model levels. This evolving landscape offers promising opportunities for future applications.

Confidential computing is an alternative approach to techniques like SMPC and homomorphic encryption. While homomorphic encryption can also protect data in use, confidential computing eliminates the computational overhead associated with performing operations on encrypted data. Confidential computing allows for standard software packages, libraries, and programming techniques within the secure enclave, facilitating the use of existing software tools and frameworks and lowering the cost of development, improving computational efficiency, and making strong privacy protection accessible and practical for many organizations. While not offering the same cryptographic guarantees as SMPC or homomorphic encryption, confidential computing excels in computational performance and flexibility.

Confidential computing has already seen adoption in a variety of industries, including healthcare, banking, and online advertising. For example, large research hospitals may coordinate analysis of randomized controlled trials in confidential computing environments to protect both patient records and potentially proprietary data analysis. Financial institutions may perform joint fraud detection operations in confidential computing environments to identify fraudulent transaction patterns without exposing the underlying data. Advertisers may conduct real time bidding auctions for online advertising to protect both individual privacy and advertiser auction history. More recently, organizations in many domains may use confidential computing to protect either the training or use of neural network-based artificial intelligence models.

Confidential computing, like other PETs, is not without its challenges and trade offs. The establishment of a TEE and the use of an attestation service to verify the integrity of the environment and its contents requires technical expertise. Although confidential computing may make it easier to achieve data privacy goals, the technology alone does not eliminate the need to understand relevant laws, regulations, and guidance for data processing. Additionally, careful consideration must be given to the design of computational outputs to prevent unintended information leakage. PSI, SMPC, and homomorphic encryption require extensive upfront design to specify exactly what elements of the computation must be protected, which makes it harder for engineers to inadvertently produce a system that leaks sensitive information. However, confidential computing isolates the entire computation without requiring a formal design to protect specific elements of the computation. As a result, it is possible for the output of a data analysis performed in a confidential computing context to reveal sensitive information.[22]

Differential Privacy

Differential privacy is a family of mathematical algorithms designed to safeguard individual privacy while enabling the receipt of valuable statistical information derived from datasets. A core principle of differential privacy involves introducing calculated noise into data to obscure specific details while preserving overall data patterns. By adding noise to individual data points, differential privacy ensures that the release of aggregate statistics does not compromise the privacy of any particular individual. This approach makes it mathematically infeasible to infer precise information about specific records from the released data. Differential privacy, though not directly useful in supporting a joint calculation among several parties like the previous technologies, plays an essential role in many data clean rooms. It is widely used in statistical data releases where precise aggregate values are important, like advertising.

The family of algorithms used in differential privacy can be thought of in two categories: global techniques and local techniques. In both cases, the goal is to produce a group analysis of common data that is then to be released to everyone in the group, but differences in the trust model dictate fundamentally different approaches. These categories are defined by where and when statistical noise is added to individual data points. In a global differentially private approach, a trusted curator has access to all the raw individual data points and is able to add noise to individual data points all at once just prior to releasing the data. For traditional clean room applications where organizations are interested in working with their raw first party data, global differential privacy is an extremely valuable tool. If all collaborators can agree to trust a single curator, then they can take advantage of a different set of algorithms than an environment where each party is responsible for adding statistical noise to their own data prior to combining them for analysis.

In a local differentially private approach, there is no trusted curator and statistical noise is added to each data point individually as a part of the algorithm that collects the data for the data set. Local differential privacy provides more protection to individuals, but the resulting data set is noisier and requires more data collection to achieve the same level of aggregate utility in statistical calculations. Organizations seeking to take advantage of differential privacy must know how to select the particular approach that is appropriate for their situation.

The application of global differential privacy requires careful consideration of several factors. The specific method for introducing noise, commonly defined by a Laplacian curve, is crucial and should be conducted by people with expertise and training in statistical data analysis or privacy enhancing technologies. Additionally, ensuring that the underlying dataset has changed enough over time is essential to maintain privacy guarantees. Techniques like subsampling or incorporating new data can help achieve this or decrease the time between statistical data releases.

Despite its strengths, differential privacy is not without limitations. Challenges include the potential for errors in noise calculation, susceptibility to side-channel attacks, and the impact of floating-point arithmetic.[23] Nevertheless, differential privacy has emerged as a leading method for protecting privacy in data publishing. Numerous organizations, including the US Census Bureau, Google, Apple, Facebook, and Microsoft, have adopted differential privacy for various applications, demonstrating its practical utility and effectiveness.

Technology Takeaways

This exploration of privacy-enhancing technologies (PETs) commonly employed in data clean room environments reveals a set of complex technology options, each designed to solve some part of overall problems addressed by data clean rooms. It is imperative to recognize that data clean rooms are not singular technologies themselves, but rather composite systems reliant on a diverse, and sometimes overlapping, array of PETs.

A salient characteristic of these PETs is that many of them are better thought of as a family of approaches and that they are more defined by the problem they solve than by the particular approaches taken. Private

set intersection, secure multi-party computation, confidential computing, homomorphic encryption, and differential privacy each provide for multiple approaches and implementations to achieve their goals. This diversity necessitates a comprehensive understanding of the available options when designing and implementing data clean room solutions. Consequently, individuals seeking to leverage data clean rooms must possess a deep familiarity with both the overarching PETs and their specific variants to make informed decisions regarding technology selection and implementation.

Another critical characteristic common to each of the technologies outlined is that they do not obviate the need for serious engagement with traditional data privacy considerations. Policy analysis, data privacy, and information security cannot be "solved" simply by using a particular technology. The remainder of this paper details existing legal requirements related to data clean rooms and provides an overview of the challenges practitioners will face when adopting them along with questions they should consider when designing solutions.

Part 4: Challenges, Questions, and Existing Legal Requirements

Data clean rooms raise privacy considerations both related to and beyond existing legal requirements. Organizations interested in building, maintaining, or participating in data collaborations using clean room technologies must consider the tradeoffs involved carefully. This Part introduces important challenges, questions, and existing legal requirements that must be considered when evaluating data clean rooms. A comprehensive evaluation of data clean rooms necessitates a thorough examination of existing regulatory guidance, both on the technologies themselves as well as on issues that organizations may want clean rooms to address.

Do data clean rooms impact organizational compliance with legal restrictions on the sale or sharing of data?

Sales, sharing, and transfers of data are commonly limited in regulatory approaches to data protection. Such limitations are seen in a variety of regulations, from the EU's General Data Protection Regulation (GDPR) to the California Consumer Privacy Act (CCPA) to name only two. Under the CCPA, the scope of a "sale" or "share" of data both implicate information transferred to a third-party.[24]

Given that a clean room is often established to limit a third party's ability to access data, there is an argument that clean rooms may provide an operational way to allow for some third-party analysis without running afoul of legal requirements. However, a recent enforcement action by the California Attorney General has introduced new complexity to this analysis and hints that enforcers may decide that many implementations of clean rooms do not comply with statutory requirements.[25]

Key questions raised under GDPR will include who is the controller of the processing taking place in a data clean room, meaning who is going to be legally responsible for the processing of personal data taking place there? What is the legal responsibility arrangement among the organizations that pool the data in the clean room? For instance, are they joint controllers? To what extent is there processing of personal data in the clean room? Are there instances where the data is anonymous, or will it always be considered personal data?

Whether the use of an in-house, first-party data clean room by third-parties, either as paying customers or as part of a co-operative agreement, would be considered a sale, share, or transfer of data under any specific regulation is case specific, and the specific structure of the clean room and technologies utilized is likely to be central to that analysis. For instance, a clean room configured to accept queries that only return aggregate answers protected by PETs and only for a limited set of purposes is arguably not transferring or providing unrestricted access to any underlying data to a third party.

Do data clean rooms help with compliance with data localization requirements or cross-border transfer restrictions?

A growing number of jurisdictions have some restrictions on the transfer of personal data across borders. Several countries have implemented "data localization" laws that require that personal data from within that jurisdiction is only processed locally.[26] With the GDPR, the EU has taken an approach in limiting transfer of personal data to jurisdictions that do not offer essentially equivalent protections.[27] In addition, a 2024 U.S. Executive Order seeks to limit transfer of personal information in certain circumstances to those deemed "countries of concern."[28]

Organizations in different global jurisdictions that want to collaborate in a data clean room must consider whether this use constitutes a cross-border data transfer within any of these relevant frameworks.

For instance, under the GDPR, data may generally only be transferred to a country that has received a decision of adequacy related to legal protections from the European Commission, such as the adequacy decision issued for the Data Privacy Framework between the U.S. and the EU[29], or, alternatively, in a circumstance where there are "appropriate safeguards."[30] In exceptional cases, derogations like individual consent can also be relied on.[31] The "appropriate safeguards" are limitatively listed in the GDPR and may include "standard contractual clauses" (SCCs), which now must also be accompanied by a transfer impact assessment (TIA) and consideration of supplementary measures to ensure sufficient protection of personal data from disproportionate or unwarranted government access.[32]

Any use of a clean room, including its specific technical underpinnings, is likely to be taken into consideration in both a TIA and the assessment of "supplementary measures" as well as the requirements of programs like the Data Privacy Framework. Although a complete analysis depends on configuration and implementation details of the particular data clean room, the protections provided by data clean room environments may make cross-border data transfers viable even in environments that do not have an adequacy decision.

Do data clean rooms offer technical solutions to incorporating principles of data minimization or use limitation?

Many data privacy regulations define and mandate the incorporation of guiding principles of data protection that include data minimization and use limitation.[33] These efforts include the GDPR, and several US state laws including California.[34] While, as with other legal questions, the particular PETs adopted in a clean room are integral to the analyses around respect for data minimization and use limitations, unfortunately, the particular PETs used in a data clean room may have distinct, and sometimes contradictory effects on each of these questions.

A consideration of a clean room implementing different PETs illustrates this tension. Differential Privacy provides protections that minimize the amount of raw individual data made available to partners in a data clean room collaboration, but those partners can use the aggregate data that is provided with few restrictions. There are no technical limits to the types of questions that could be asked outside of the statistical noise provided to protect individual data points. On the other hand, SMPC and Homomorphic Encryption both require more understanding of the particular use cases to be enabled by the computation. Only the analysis enabled by the particular algorithm implemented are possible with those technologies.

As with many questions about data clean rooms, the details matter. However, in general, the technologies examined in this paper can be thought of as follows:

1. **Private Set Intersection:** Allows for the adoption of strong data minimization policies, but once the intersection has been identified, any analysis can be conducted on that data.
2. **Secure Multi-party Computation:** Able to provide for defined uses and prohibit all others, but any data can be provided as input to the computation.

3. **Homomorphic Encryption:** Allows somewhat for the delineation of limited uses while also protecting the secrecy of the input data. For example, a host of a system using homomorphic encryption will not know what search was performed on a database or what the search returned to the user, and the user can enter any search terms and receive all applicable results for their search. All in all, Homomorphic Encryption provides better protection for use limitations than for data minimization requirements.
4. **Confidential Computing:** Provides for isolation of the data while it is being used. This technology does not require or provide hard technical limitations for either data minimization or use limitation requirements. Confidential computing offers a flexibility not available in other PETs, but the tradeoff for this flexibility is that data minimization and use prohibition requirements must be accounted for through other means.
5. **Differential Privacy:** Provides strong data minimization protection for individual records, but does not limit or restrict the analysis that can be performed in the aggregate.

What are other policy considerations organizations should consider in implementing clean rooms?

Data clean rooms offer a secure solution for organizations to conduct data analysis while prioritizing privacy. By leveraging PETs, clean rooms enable collaborative data analysis while protecting sensitive information. While clean rooms today are too slow and limited for many use cases, for instance to be used as a replacement for the third-party cookie ecosystem, they have great potential to be utilized in a variety of contexts in the future as the technology further develops.

While data clean rooms offer several benefits, it's important to acknowledge their limitations:

- Data clean rooms may prove to be too technically complex for a lay user to easily understand, meaning the "creepy" factor associated with data-driven insights may erode trust even when the analysis that produces those insights is protected by a data clean room.
- The technical requirements for data clean rooms are more onerous than traditional open-web methods based on third-party data access.
- Data clean rooms don't significantly alter existing challenges related to ensuring that informed consent is received by all parties for the processing of their data.

To address these challenges, organizations must prioritize clear communication about the technologies used in data collaborations with regulatory authorities.

Conclusion

This paper provided an overview of data clean rooms, including their historical development, use cases, and technological underpinnings. It provides practitioners with a taxonomy for thinking about clean rooms, breaking down options into four models and discussing their suitability for various data analysis scenarios. Finally, this paper introduces a series of critically important legal questions concerning how the implementation and use of clean rooms may address various regulatory requirements. The answers to these questions likely turn on technical details related to a particular model of clean room or the PETs that it utilizes. It will be important going forward for organizations to consider practical applications where clean room technology has been or will be implemented to provide greater insight into situations where clean rooms may or may not address specific needs.

Appendix: Government-Issued Guidance

Some governments have produced guidance on individual PETs, such as differential privacy or secure multi-party computation, but standards or frameworks for evaluating entire data clean room solutions do not yet exist. The United States, the United Kingdom, the European Union, Canada, and Singapore have each produced an evaluation of individual PETs with recommendations for data privacy practitioners. All of this guidance focuses either on PETs in general or on individual PETs, such as differential privacy or secure multi-party computation, rather than complete data clean room solutions found on the marketplace. This appendix provides an overview of the most relevant guidance available.

In the United States, the White House Office of Science and Technology Policy (OSTP) explored the responsible use of PETs.[35] Their National Strategy to Advance Privacy Preserving Data Sharing and Analytics cites several barriers to adoption and use of PETs, including limited awareness and understanding of the individual PETs, the lack of guidance surrounding the use of PETs, and the need to do further investigation on exactly how PETs are viewed by individuals in the context of data privacy decisions.

In the United Kingdom, the Information Commissioner's Office (ICO) developed a risk-benefit analysis tool exploring the use of several PETs.[36] Many of the technologies explored are commonly used in data clean rooms, although none of the individual PETs in the analysis fully comprise a data clean room environment on their own. The challenge for data privacy practitioners is to deconstruct data analytics solutions marketed as data clean rooms to understand whether they use the individual PETs outlined.

In the European Union, the European Union Agency for Cybersecurity (ENISA) has referenced PETs in a report in January of 2022 on data protection, highlighting differential privacy, homomorphic encryption, and secure multi-party computation.[37] ENISA's recommendations outline the circumstances where the Agency determined adopting these PETs was beneficial, but do not address whether or how combinations of PETs, as in a data clean room, preserves or extends these benefits.

The Infocom Media Development Authority (IMDA) in Singapore launched a PET Sandbox to support companies interested in exploring the application of PETs.[38] The idea is to invite companies to participate along with the IMDA in evaluating technologies so that the lessons learned from the evaluation can be shared, focusing on the context of regulatory compliance. A year after the launch of this program, Google partnered with IMDA in order to offer guidance to industry participants regarding the use of Google Privacy Sandbox to conduct advertising campaigns without requiring third-party cookies.[39] IMDA's goal was to better understand how the technology worked as a means of developing a better approach to regulating them. Although a similar approach could be taken with data clean room technologies, it has not yet been done.

While these efforts each provide valuable insights, a more comprehensive regulatory framework would mitigate the unique challenges faced by data privacy professionals seeking to examine data clean rooms. This framework should encompass not only the underlying PETs but also the overall architecture and governance of complete data clean room systems.

Endnotes

1 See, e.g., IAB Tech Lab, Data Clean Rooms Guidance and Recommended Practices (July 2023), *available at* https://iabtechlab.com/datacleanrooms/.

2 See, e.g., https://dualitytech.com/blog/data-clean-rooms-advantages-and-disadvantages/.

3 See, e.g., https://iabtechlab.com/datacleanrooms/ and also https://www.informationpolicycentre.com/cipl-blog/cipl-releases-paper-on-privacy-enhancing-and-privacy-preserving-technologies-understanding-the-role-of-pets-and-ppts-in-the-digital-age

4 *Supra* note 1.

5 See the FTC's Complaint against X-Mode here: https://www.ftc.gov/system/files/ftc_gov/pdf/X-ModeSocialComplaint.pdf

6 Organizations may no longer be able to rely exclusively on contractual language because, as the FTC indicated in X-Mode, "such language is insufficient to protect consumers from substantial injury." See *supra* note 4 at page 8.

7 Given the term "clean room" it may be natural for folks to associate data clean rooms with "privacy washing," or the act of putting limited or superficial protections on top of a practice that carries heavy privacy risk without addressing the risk itself. Privacy washing is an emerging and potentially serious concern, but it's not fair to blindly associate it with all data clean rooms.

8 A "clean room" should not be confused with the concept of "data cleaning"– the process of getting data into the proper format for analysis, including eliminating invalid or incomplete data. Data collected without a plan or organized effort is likely to require cleaning, which may consist of removing duplicate data or ensuring all dates or currencies are formatted correctly. This process remains exceedingly difficult to automate and typically involves some manual effort. Statistical data cleaning is not the purpose of a data clean room. Both involve preparing the data for analysis, but the methods and purposes of that preparation are not similar. The purpose of a data clean room is to enable data analysis or sharing without revealing restricted information, whereas the purpose of statistical data cleaning is to ensure the most accurate calculation possible with the data available.

9 Mud rooms allow anyone entering a house to take off muddy shoes in a designated room, so as to prevent mud from spreading to the rest of the house.

10 For more, see FPF's risk-utility framework in our draft publication "Advertising in the Age of Data Protection," available online: https://fpf.org/blog/examining-novel-advertising-solutions-a-proposed-risk-utility-framework/ The goal of this framework is to help organizations understand the inherent tradeoffs presented by the modern digital advertising landscape and the technology it is built upon. Rather than assigning value to the utility, risk, or social impact, the goal of the framework is to comprehensively identify the relevant factors for decision making.

11 Barthelemy, Lea. 2023. "Data Clean Rooms Are Now Essential for Audience Insights, Measurement, and Data Activation According to IAB State of Data Report." IAB. January 24, 2023. https://www.iab.com/news/state-of-data-2023-report-data-clean-rooms/

12 Feature vectors are just the bits of data that a machine learning algorithm uses to make a prediction about something, in this example shopping habits. "Feature (Machine Learning)." 2024. In Wikipedia. https://en.wikipedia.org/wiki/Feature_(machine_learning)#Feature_vectors

13 A cryptographic primitive is a low-level, well-understood mathematical building block used by cryptographers to build complete cryptographic protocols. For example, a one-way hash function may be adapted as part of a complete technique for identifying the intersection in a private set intersection. For more information, see https://en.wikipedia.org/wiki/Cryptographic_primitive

14 The complete mathematical details of this technique are beyond the scope of this paper. That said, we mention oblivious transfer here because it allows for a computationally efficient implementation of a probabilistic match in PSI.

15 Garbled circuits represent a foundational branch of research in cryptography that remains relevant today. They conceptualize the computation as an electrical circuit, obfuscating its structure to protect the privacy of inputs encoded by the circuit.

16 It may now be clear that oblivious transfer protocols mentioned in the previous subsection are based on this foundational principle for calculating results without sharing data. Oblivious transfer often serves as a critical component in modern SMPC implementations, allowing the receiver to select specific outputs without revealing their choice of which outputs were selected to the sender.

17 Shamir Secret Sharing and Additive Secret Sharing are notable examples of techniques used in these protocols.

18 Mitigating the cost and inefficiency of MPC implementations is an active area of research that may alter this analysis in only a few years. From an implementation standpoint, two primary strategies emerge. The first involves custom-building a solution tailored to specific computational requirements and using only the cryptographic primitives that remain relevant when the problem is fully and precisely specified. While this approach optimizes efficiency and privacy, it is resource-intensive and time-consuming. Alternatively, generic "compiled" frameworks offer a more rapid development path. By providing a higher-level abstraction, such as a programming language or domain-specific language, these frameworks enable developers to focus on problem definition rather than low-level cryptographic details. However, they may sacrifice some performance and customization flexibility compared to custom implementations. Of course, the benefit of this approach is that it reduces development costs and makes experimentation easier.

19 While the concept of homomorphic encryption was introduced in the 1970s, significant advancements towards performing arbitrary calculations using homomorphic encryption were not achieved until lattice-based methods emerged around 2009.

20 Homomorphic encryption relies on a similar mathematical foundation to garbled circuits in that the cryptographic primitives used reduce the problem to something similar to solving an electronic circuit. The four primary categories of homomorphic encryption are roughly defined by how many different types of gates are available to represent the circuit. For example, partially homomorphic encryption consists entirely of one gate, either addition or multiplication. Describing these differences in terms of the complexity of real world data privacy concerns is non-trivial, but the same general tradeoff seen in SMPC implementations is also true for homomorphic encryption: a more complex circuit allows for more functionality with privacy protection at the cost of computational complexity and efficiency.

21 FPF has a more detailed analysis of the policy implications of Confidential Computing available here: https://fpf.org/resource/confidential-computing-and-privacy-policy-implications-of-trusted-execution-environments/

22 Recall that one of the reasons we included discussion of contract-only clean rooms in Part 2 of this paper was that the contractual analysis identifying specific privacy risks and outlining a detailed approach to mitigating them is necessary for all models of data clean rooms. To varying degrees, PSI, SMPC, and Homomorphic Encryption all require some of this analysis as part of their technical implementation. Confidential Computing has no such hard technical requirement, even if the policy requirement remains.

23 A detailed analysis of the mathematical limitations of differential privacy is available in Dwork and Roth's The Algorithmic Foundations of Differential Privacy. https://www.cis.upenn.edu/~aaroth/privacybook.html

24 https://www.oag.ca.gov/privacy/ccpa#sectiona

25 In 2024, the California Attorney General announced a settlement with DoorDash, a food delivery platform. The complaint described DoorDash as participating in a cooperative marketing arrangement without providing appropriate disclosure or offering appropriate opt out mechanisms. DoorDash's participation involved data transfers to the cooperative that the Attorney General deemed "sales." To be clear, DoorDash was not using a data clean room, but the Complaint's language was broad enough to raise questions about whether the use of a data clean room would have made a material difference. Specifically, it determined that a key concern preventing remediation was that "consumer personal information and inferences about DoorDash's customers had already been sold downstream." A data clean room could potentially eliminate downstream sales of consumer personal information, but the purpose of a data clean room is to preserve the ability to sell inferences based on that information. Because both are mentioned in the complaint, the counterfactual analysis of a similar scenario that included a data clean room is unclear.

See FPF's summary of this here: https://fpf.org/wp-content/uploads/2024/06/Enforcement-Report-FINAL.pdf ; See also California Attorney General's press release announcing this enforcement action: https://oag.ca.gov/news/press-releases/attorney-general-bonta-announces-settlement-doordash-investigation-finds-company And see also the full complaint detailing the downstream sales: https://oag.ca.gov/system/files/attachments/press-docs/DoorDash%20Complaint.pdf

26 These countries include Australia, Canada, China, India, Russia and others. Some countries limit the scope of data localization laws to apply to a particular industry or domain, such as health data. For an overview of this trend, see: https://scholarlycommons.law.emory.edu/elj/vol64/iss3/2/ ; For a detailed example of data localization in China, see: https://fpf.org/blog/new-fpf-report-demystifying-data-localization-in-china-a-practical-guide/

27 In particular, the European Union (EU) requires processing of personal information of EU persons take place in jurisdictions that meet or exceed EU data protection standards. Several approaches exist to discern whether transfers comply with this standard, including Adequacy Decisions: https://commission.europa.eu/law/law-topic/data-protection/international-dimension-data-protection/adequacy-decisions_en ; Standard Contractual Clauses: https://commission.europa.eu/law/law-topic/data-protection/international-dimension-data-protection/standard-contractual-clauses-scc_en ; and Binding Corporate Rules: https://commission.europa.eu/law/law-topic/data-protection/international-dimension-data-protection/binding-corporate-rules-bcr_en

28 In February 2024, President Biden issued an executive order designed to protect Americans' sensitive personal information, including financial, health, location, and other personally identifiable information from transfer to "countries of concern." The executive order details implications for data brokers, companies building artificial intelligence technologies, and other companies that depend on data transfers. https://www.whitehouse.gov/briefing-room/presidential-actions/2024/02/28/executive-order-on-preventing-access-to-americans-bulk-sensitive-personal-data-and-united-states-government-related-data-by-countries-of-concern/

29 The Data Privacy Framework requires organizations to self-certify the presence of certain protections, including data integrity and accountability for any data transferred to third parties. https://www.dataprivacyframework.gov/key-requirements.

30 Without an adequacy decision from the European Commission that confirms the requirements are met or exceeded by the jurisdiction to which the data is transferred, the GDPR authorizes cross-border transfers of personal data when the data exporter "has provided appropriate safeguards, and on the condition that enforceable data subject rights and effective legal remedies for data subjects are available." See Art. 46 of the GDPR: https://gdpr-info.eu/art-46-gdpr/ ; See also International Data Transfers, European Data Protection Board, https://www.edpb.europa.eu/sme-data-protection-guide/international-data-transfers_en

31 Article 49 GDPR.

32 Most organizations rely on standard contractual clauses (SCCs), or model contracts that serve to raise the level of data protection to GDPR-like standards: https://www.skadden.com/insights/publications/2022/03/privacy-cybersecurity-update However, the Court of Justice of the European Union determined in 2020 that SCCs are not alone sufficient and that organizations that rely on SCCs to transfer data to a non-adequate jurisdiction must conduct a transfer impact assessment (TIA) and consider "supplementary measures" to ensure sufficient data protection, depending on the laws of the importer country. See the EU Commission's Q & A for the new standard contractual clauses developed in response to the Schrems II judgement: https://commission.europa.eu/law/law-topic/data-protection/international-dimension-data-protection/new-standard-contractual-clauses-questions-and-answers-overview_en ; And see also FPF's comparison of the EU's new standard contractual clauses with two other frameworks for international data transfers. https://fpf.org/blog/fpf-report-not-so-standard-clauses-an-examination-of-three-regional-contractual-frameworks-for-international-data-transfers/

33 https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/data-protection-principles/a-guide-to-the-data-protection-principles/the-principles/. "Purpose limitation" indicates that data is "collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes; further processing for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes shall not be considered to be incompatible with the initial purposes." "Data minimization" requires that data processing is adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed.

34 GDPR: https://gdpr-info.eu/art-5-gdpr/ ; UK's ICO: https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/data-protection-principles/a-guide-to-the-data-protection-principles/ ; California's CCPA Section 1798.100(c) reads: "A business' collection, use, retention, and sharing of a consumer's personal information shall be reasonably necessary and proportionate to achieve the purposes for which the personal information was collected or processed, or for another disclosed purpose that is compatible with the context in which the personal information was collected, and not further processed in a manner that is incompatible with those purposes." For more: https://leginfo.legislature.ca.gov/faces/codes_displayText.xhtml?division=3.&part=4.&lawCode=CIV&title=1.81.5
Note that Section 7002(a) of the CCPA regulations clarify further that "a business's collection, use, retention, and/or sharing of a consumer's personal information shall be reasonably necessary and proportionate to achieve: (1) The purpose(s) for which the personal information was collected or processed, which shall comply with the requirements set forth in subsection (b); or (2) Another disclosed purpose that is compatible with the context in which the personal information was collected, which shall comply with the requirements set forth in subsection (c)."

35 https://www.whitehouse.gov/wp-content/uploads/2023/03/National-Strategy-to-Advance-Privacy-Preserving-Data-Sharing-and-Analytics.pdf

36 https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/data-sharing/privacy-enhancing-technologies/

37 https://www.enisa.europa.eu/publications/data-protection-engineering/@@download/fullReport

38 https://www.imda.gov.sg/how-we-can-help/data-innovation/privacy-enhancing-technology-sandboxes

39 https://rsvp.withgoogle.com/events/privacysandbox-sg