LUNAR VENTURES

Privacy Enhancing Technologies

Part 1: *The privacy infrastructure of tomorrow is being built today.*

This paper is part of the *Lunar Ventures: Insight Series*

Authors:
Alberto Cresto
Jaivi Jayakumar

Disclaimer

This document (the “Document”) has been prepared by Berlin Innovation Ventures GmbH (“Lunar Ventures”). Lunar Ventures is registered in Germany at the Local Court of Charlottenburg with registration number HRB 190056 B. Lunar Ventures is an appointed representative of Berlin Innovation Ventures 1 GmbH & Co. KG (“Lunar I”) which is authorized and regulated by the German Federal Financial Supervisory Authority.

No undertaking, warranty or other assurance is given, and none should be implied, as to, and no reliance should be placed on, the accuracy, completeness or fairness of the information or opinions contained in this Document. The information contained in the Document is not subject to completion, alteration and verification nor should it be assumed that the information in the Document will be updated. The information contained in the Document has not been verified by Lunar Ventures, Lunar I or any of its associates or affiliates. The Document should not be considered a recommendation by Lunar Ventures, Lunar I or any of its directors, officers, employees, agents or advisers in connection with any purchase of or subscription for securities. Recipients should not construe the contents of this Document as legal, tax, regulatory, financial, or accounting advice and are urged to consult with their own advisers in relation to such matters. The information contained in the Document has been prepared purely for informational purposes. In all cases persons should conduct their own investigation and analysis of the data in the Document. The information contained in the Document has not been approved by the Federal Financial Supervisory Authority. This Document does not constitute, or form part of, any offer of, or invitation to apply for, securities nor shall it, or the fact of its distribution, form the basis of or be relied upon in connection with any contract or commitment to acquire any securities. Any forecasts, opinions, estimates, and projections contained in the Document constitute the judgement of Lunar Ventures and are provided for illustrative purposes only. Such forecasts, opinions, estimates, and projections involve known and unknown risks, uncertainties and other factors which may cause the actual results, performance or achievements to be materially different from any future results, performance or achievements expressed or implied by such forecasts, opinions, estimates and projections. Accordingly, no warrant (express or implied) is or will be made or given in relation to, and (except in the case of wilful fraud) no responsibility or liability is or will be accepted by Lunar Ventures, Lunar I or any of its directors, officers, employees, agents or advisers in respect of, such forecasts, opinions, estimates and projections or their achievement or reasonableness. Recipients of the Document must determine for themselves the reliance (if any) that they should place on such forecasts, opinions, estimates and projections. Information contained in the Document may not be distributed, published or reproduced in whole or in part or disclosed to any other person. The distribution of any document provided at or in connection with the Document in jurisdictions other than Germany may be restricted by law and therefore persons into whose possession any such documents may come should inform themselves about and observe any such restrictions.

About this white paper

About Lunar Ventures

Lunar Ventures is an early seed venture fund based in Berlin. We invest anywhere in Europe in technical teams building moonshot infrastructure software companies. Get in touch with us at hello@lunarventures.eu

About the authors:

Alberto Cresto, Principal
VC investor with 20+ deeptech investments executed across Europe and North America. Alberto has advised startups and public companies on M&A initiatives, minority investments and new business development in high tech sectors since 2015.

Jaivi Jayakumar, Analyst
While completing advanced studies at ESCP Business School, Jaivignesh has worked in roles closely associated with IT, manufacturing, product development, consulting, project management, and venture capital.

A special thanks to our Chief Scientist Elad and to our CTO Mick for the priceless insights and continued support in preparing this white paper.

Table of contents

Executive Summary

At Lunar we believe many new exciting technologies powered by breakthroughs in computer science are on the verge of making it into the real world and become building blocks of a new generation of tech infrastructure. This is why we started Lunar Ventures after all. *Privacy Enhancing Technologies* (PETs) are one of the most exciting spaces for us as technologists and as investors. The purpose of this document is to provide to a non-technical audience a primer into this space by shedding some light into what these technologies are, how they work, what they are best suited for and which companies are playing a role in this nascent ecosystem.

Crucially PETs change the way we interact with data. The biggest paradox in data-driven business models has always been the inability to unlock new opportunities while ensuring confidentiality, integrity, and security at the same time. The data that we (individuals and businesses) generate daily has massive potential to improve our experiences, which invariably come at the cost of our privacy. Acting on the longstanding trade-off between generating insights out of data and maintaining its confidentiality, PETs enable us to run analyses while respecting the secrecy of the underlying data. This has some important implications.

Firstly it pushes companies and individuals to more confidently share their data, under the guarantee that it will remain confidential — increasing the amount of data overall available to improve existing products and services, and develop new ones.

Secondly it allows internal departments and different companies to collaborate confidentially on sensitive datasets and analyses, while protecting the secrecy of both the data and the IP behind the analytics methods — breaking many of the data silos erected over time by regulation and competitive concerns.

While doing so, we believe PETs will generate an amount of value in the trillion dollar range, and completely re-shape for the better the way individuals and organizations produce, share, consume and interact with data.

The early use-cases of PETs range from banking to healthcare, through IT infrastructure and insurance, and to advertising — with more constantly popping up as the interest in these technologies rises rapidly.

Different PETs achieve different goals, have distinct pros and cons, and their adoption comes with different types of costs (often in terms of lost efficiency and ease of use): we will review these both at a general level, and will then further deep dive into each technique.

Finally, to make things more tangible, we will offer a peek into the nascent PET market and into what startups and corporates are doing in the space.

This paper focuses on the current state of PETs and will be followed with another document "*Collaborative Computing: Making data sharing cheaper, faster and easier with partnership-enhancing technologies*" discussing the outlook and future development of the PET market.

Lunar PET Landscape

LUNAR VENTURES

ENABLING INFRASTRUCTURE

SOFTWARE-BASED

PII De-identification

Data Anonymization & DP

DATAGUISE · aircloak · LEAPYEAR · HACERA · nymiz · PRIVITAR · ANONOS · MMUTA · KIPROTECT · TUMULT LABS

Synthetic Data

hazy · gretel · YDATA · Statice · MOSTLY·AI · Facteus · SYNTHO · TRUATA. · SYNTHESIZED · TONIC

Private Computation

Software based

Tr'pleBl'nd · SHIELD · Adap · Partisia · apheris AI · inpher · Duality · 華控清交 HUA KONG TSINGJIAO · PRIVECOMMS · ZAMA · DATAFLEETS · SHIELDIO · ENVEIL · Oblivious AI · nth party · SHOJI · PlatON · appliedblockchain · prenode · ARPA · fetch.ai · XAIN. · qedit · LIGERO · sherpa.ai · Ntropy network. · MagicCube · PRIVATE STORAGE · SCALEOUT · S20.AI

Hybrid (software & hardware)

cosmian · sharemind · OASIS LABS · enigma · secretarium

HARDWARE-ENABLED

Hardware based

Fortanix · EDGELESS SYSTEMS · decentriq · SCONE · MADANA · CYSEC · anjuna · GRAPHENE · TrustKernel · evervault

Providers of Confidential Cloud

Google Cloud · Microsoft Azure · Alibaba Cloud · IBM Cloud · aws

PET Hardware

IBM · AMD · RISC-V · arm · intel

SPECIFIC USE-CASES

Identity

PEER MOUNTAIN · nuggets · KRIPTAN · NuID · civic · sovrin identity for all · sedicii

User Data

Glimpse Protocol · Canopy · INSIGHTS NETWORK · MIMIRIUM · INFOSUM

Key Management

UNBOUND · KMS CHAIN · KEYLESS · SEPIOR · SPHERITY · MIRACL

Medical

InfoShield · Kenya.ai · Lynx.MD · OWKIN

Cryptocurrency & Ledgers

Enduser Facing

ZenGo · Zcash · MONERO · curv · Zcoin · MW · ZEAN

Ledger Infrastructure

BOLT LABS · NuCypher · Qredo · LOOPRING · MATTER LABS · CYBERIAN · SHARDX · THRESHOLD · ATOMRIGS LAB · keyTango · CYBAVO · AZTEC · STARKWARE · Fireblocks

Did we forget someone or do you have any feedback?

Feel free to reach out to

Alberto Cresto

Lunar PET Landscape Cheatsheet (1/2)

	PRIVATE COMPUTATION (CRYPTOGRAPHY)		PIIs DE-IDENTIFICATION	
	SOFTWARE BASED	**HARDWARE BASED (TEEs MIDDLEWARE)**	**SYNTHETIC DATA**	**DIFFERENTIAL PRIVACY**
What is the value proposition of these companies?	• Allow mutually-distrusting parties to confidentially collaborate on data analysis or to analyze data while keeping it secret		• Remove Personally Identifiable Information from data analyses • Create a "surrogate" data set that has the statistical/overall characteristics of the original, but does not contain sensitive information	• Perform aggregate queries on a dataset, while ensuring the queries' results do not expose any PIIs
Techniques	• **HE**: ensures that both the data and the result of the analysis remains secret, removing the need to trust the location where the analysis takes place. • **SMPC**: enables multiple mutually-distrusting parties to collaborate on a joint analysis on confidential data, preventing any participant from learning anything about the inputs provided by the other parties. • **ZKP**: allows data provided by one party to remain secret while being verified by another party.	• **Trusted Execution Environments**: a secure partition of a larger chip/SoC that secures the execution environment of the analysis by isolating it completely from the rest of the machine processes.	• **Synthetic Data**: multiple data generation techniques which create an artificial data set mimicking the properties and correlations of an original, confidential dataset	• **Differential Privacy**: its main goal is to protect the privacy of any individual providing his information to a database that is used for aggregate analysis.
What enabled this?	• Advances in computer Science and cryptography	• Advances in hardware hardening and virtualization	• Various advances (depending on the technique used)	• Advances in machine learning and statistics
How do they look?	• A cryptographic protocol telling participants what computations to perform, what information to encrypt and how, and where to send it	• Software component that orchestrates a hardware enclave (often provisioned by a cloud provider)	• Software generating novel datasets to be used in place of the original	• Software sitting between the data analyst and the sensitive dataset, offering an anonymized view by adding noise to the "true" results of the queries
What computational efficiency/ performance losses do we incur?	• Large efficiency hit (often 1000x-10,000x or more) but rapidly improving with R&D progress	• Moderate performance hits (depending on the type of computation, CPU or GPU architectures, etc)	• No performance hit	• No performance hit

Lunar PET Landscape Cheatsheet (2/2)

	PRIVATE COMPUTATION (CRYPTOGRAPHY)		PIIs DE-IDENTIFICATION	
	SOFTWARE BASED	**HARDWARE BASED (TEEs MIDDLEWARE)**	**SYNTHETIC DATA**	**DIFFERENTIAL PRIVACY**
Impact on accuracy or validity of results	• Perfect accuracy/validity		• Validity of results depends on the quality of the software and the expertise of the data scientist in charge of producing the synthetic data	• Validity and accuracy are high if correctly used. However, it is often not possible to apply a full analysis due to restrictions like a "privacy budget"
Difficulty of integration and workflow impact	• Often hard to integrate into existing systems, requires assistance of security experts	• Requires the adaptation and deployment of applications to a new environment • As of 2021, there are no technological barriers to integrating this	• Does not require extensive integration since it generates a new data set that can be used instead of the original one • May create a convoluted workflow, where the data scientist creates hypotheses on the synthetic data, then verified by a 3rd party on the original data	• Important impact on workflow, as the DP software often ends being the UX of the data analyst / the interface to the data being analyzed • Made people quite unhappy — this system does not jibe well with the way companies work with data
Required expertise from the developer or data scientist	• Requires deep and nuanced understanding of security and cryptography (except in some uniquely easy-to-integrate cases, e.g. Zama)	• Does not require particular expertise, assuming reasonably strong middleware (e.g. Anjuna)	• The data scientist who anonymizes the data needs expertise to create synthetic data with characteristics consistent to the original dataset, and/or that properly hides sensitive information • The data scientist who consumes the anonymized data does not need particular expertise	• Does not require particular expertise from the data scientist
Commercial maturity	• Early	• Medium	• Mature	• Medium
Confidentiality assurance	• Confidentiality is mathematically guaranteed: as long as integration was done well (which is challenging), privacy is entirely preserved	• Confidentiality is based on hardware trustworthiness • Proven to be vulnerable to multiple types of side-channel attacks • With these caveats in mind, TEEs are secure as long as integration was done well	• Confidentiality is based on the skills of the data scientists and the technique used • Lack of standardization and best practices prevents predictable assurances of confidentiality	• Confidentiality is guaranteed — as long as the system parameters were chosen well • However, there is an inherent tradeoff between: the desired level of privacy, the number of queries that the system can perform and their accuracy.

Introduction

Data has grown to be one of our most valuable assets — and researchers, entrepreneurs and companies have made immense efforts over time to keep it secure and private.

Data is at risk:

- when **at rest** in our computers' and clouds' storage;
- when **in transit** over our telecommunication networks;
- and when **in use**, analyzed by computers to generate insights.

Originally data was rarely shared outside the premises of an organization. Until the advent of the internet, most attention to privacy was devoted to keeping data encrypted while **at rest**, in case an intruder managed to get access to the secured location of the data — driving adoption of technologies such as *symmetric key encryption*. With symmetric key encryption data is made unreadable (encrypted ciphertext) and again fungible (decrypted plaintext) using the same encryption key (typically a password) which is to be shared with all parties needing access to the data. The first use of such encryption dates back nearly 4,000 years ago, when the ancients would apply simple but effective techniques like shifting the alphabet by a few positions in order to write a secret message — and became increasingly sophisticated in modern times (think of the Enigma Machine used in WW2, which also marked the dawn of modern computation).

As the internet evolved in an interconnected and dynamic network, and as it became the norm to exchange data at scale, the focus of encryption shifted to protecting data **in transit**. However symmetric encryption was simply not up for the job: it would have been completely nonsensical to send to the receiver the very secret password needed to decrypt the message — over the very same unsecure communication channel which we were trying to secure in the first place! Researchers put their heads down, and a new generation of technologies like TLS (Transport Layer Security), powered by *asymmetric key encryption*, made even the most sensitive data secure while transferred between different parties. The novelty of asymmetric key encryption was to use a different key for encryption and a different one for decryption: the sender could encrypt a message with the receiver's *public key* (known to everyone) — and this message would only be decryptable by the receiver's *private key* (a personal password).

Our data is secure when at rest and in transit. But we are still forced to trust any counterparty handling our data for analysis.

This finally removed the need to exchange passwords among parties needing to collaborate on data, historically the achilles heel of symmetric encryption.

Encrypted data communications made it possible, for example, to securely use credit cards online, unlocking the world of ecommerce and giving rise to giant companies like Amazon (remember 2002, when we were still afraid to buy those concert tickets online?). It thus opened the door for secure communication between *two trusted, non-competing and non-adversarial parties* — creating an opportunity for Fintech, SaaS, remote work (working and living during a pandemic would look very different without the backbone of TLS) and basically anything running in the cloud.

If we follow the state of the art in security today, our data would be secure at all times it's intended: when stored on our premises, or on the premises of another party we had decided to transfer it to, and while in transit to our desired trusted recipients. Nonetheless there's always an assumption on the trustworthiness of the parties we share data with: we trust them to keep it confidential, or use it only for the specific purposes that were agreed upon.

Until very recently analyzing and extracting insights from data always required access in plain-text which means we could not extract and share the value of data while keeping the data secret.

Unfortunately, the assumption of trust has proven to be consistently incorrect, as counterparty organizations have used data for activities that were not previously agreed upon, or unilaterally changed terms of service. To counter these questionable practices, increasingly fuelled by powerful corporate and private interests, regulators have put many restrictions on how we handle data, with whom it can be shared, and set specific rights for its owners.

But no matter how much regulation you throw at it, handing over data for analysis to a counterparty always imposes some risks and requires some level of trust — regardless of how secure it is at rest or in transit. Taking this risk is a deliberate decision we make considering the trade-off between maximising the value of data and protecting the privacy and the interests of its owners, within an increasingly stringent legal framework.

This tradeoff exists for one simple reason: until very recently data could not be kept confidential while **in use**. Analyzing and extracting insights from data had always required access in plain-text which means *we could not share the value of data while keeping the data secret*. This requirement has wildly limited the way we collaborate on data and prevented us from exploiting its full potential. Computer scientists have long fantasized about a theoretical "Blind Turing Machine", a computer completely oblivious to the tasks it is executing and the data it is processing. With the rise of cloud computing and the advent of outsourced computation as the standard paradigm for data analytics, the opportunity cost grew even bigger, pushing research efforts to focus on developing methods to process and understand data, while preserving its privacy. The leading cryptographers and academic researchers in the world have been spending decades on this topic by

now, and many of the recent Turing and Godel awards have been awarded for advancements in this field. After decades in academia, this research is finally reaching an inflection point and gaining grounds in commercial applications, giving birth to the brave new world of *Privacy Enhancing Technologies* (PET).

PETs enable potentially adversarial parties to collaborate on sensitive data without needing to rely on mutual trust

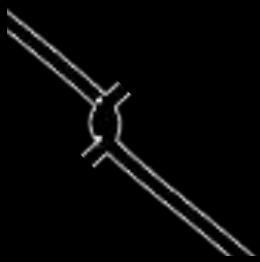
PETs already promise to revolutionize the way we generate, transact, exchange and consume data. It is our deep belief that it will soon unlock trillions in economic value — besides hopefully a better world from a social perspective: democratizing encryption carries the promise of granting individuals the same privacy that was until now the luxury of nation states. This space, while still lagging behind in the hype cycle, does succeed in attracting interest from major investors and corporates. A new generation of startups is rapidly flourishing across the globe (with more than $850M raised so far!) pushed by technology entrepreneurs, and backed by a handful of VCs familiar with the science and engineering of PETs. On the other hand, the largest tech companies (and even some industrial early adopters) are embracing PETs for use in their own products and repackaging that experience to build PET building blocks that will pose the basis of a new generation of enabling infrastructure.

This document is meant to be a primer to PETs. We hope to help our readers better understand what problems they solve; which benefits and costs come with their adoption; which techniques and tools are normally used; which early applications are emerging; and which companies and startups are pushing the boundaries of this nascent space.

Section 1: Why PET?

What are PETs?

PETs are a set of cryptographic techniques and protocol, architectural designs, data workflows, and systems of hardware and software that enable adversarial parties to collaborate on sensitive data without needing to rely on mutual trust.

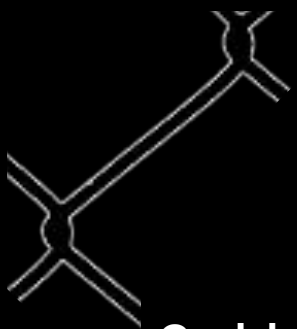


What problems do PET solve?

Handling sensitive data and sharing it with third parties impose large liabilities which have prevented us from exploiting the full potential of the data ecosystem. PETs enable adversarial parties to collaborate on sensitive data without needing to rely on mutual trust.

What are the opportunities?

Like previous waves of cryptography, PETs adoption could unlock a trillion dollar opportunity by helping us extract more value from existing data, driving the creation of even more of it and enabling a new generation of services and use-cases to flourish.



Goal, benefits and cost of adoption.

PETs' main goal is to increase the level of confidentiality when multiple parties collaborate together on sensitive data. PETs drastically improve the trade off between generating value from data and respecting its secrecy, but can incur meaningful accuracy and efficiency losses.

Section 1: Why PETs?

PETs hold the promise to revolutionize the way we generate, transact, exchange and consume data — enabling us to share its value, while keeping it secret.

What problems PETs solve?

Handling sensitive data and sharing it with third parties impose large liabilities which have prevented us from exploiting the full potential of the data ecosystem. PETs enable adversarial parties to collaborate on sensitive data without needing to rely on mutual trust.

PETs are a set of cryptographic techniques, architectural and protocol designs, data science workflows, and systems of hardware and software **that enable potentially adversarial parties to collaborate on sensitive data without needing to rely on mutual trust.** Fundamentally, PETs eliminate or reduce the amount of information leaked in data analysis — be that information related to the data being used, the result of the analysis or the parties involved in the process. In order to understand what problems PETs solve, let's look at an imaginary use-case together:

Silvia Roche, a brilliant health-professional-turned-founder, is working on a novel digital health startup ("STDHelp") that diagnoses sexually transmitted diseases (STDs) by utilizing computer vision to analyze patients' pictures of their private parts. STDHelp is a phone application that lets you upload pictures and get a nearly instant diagnosis. The service is appealing because it's cheap, instantaneous, and complies with prevalent taboos around these topics — for many it's far less uncomfortable consulting a machine than a person on these matters. When a user uploads a picture to STDHelp, this gets analyzed in order to detect STDs, and the user is provided with a painless diagnosys. STDHelp's strategy is to quickly expand the scope of detected issues.
The service works using a computer vision model which is offered as-a-service to STDHelp by a machine learning company ("CV-ML.ai"). Silvia just needs to tailor the CV model to her specific needs by finding and using a few thousand images of STD infections to train for her specific use case.

The healthcare example of STDHelp is a small and simple, but realistic example that we can use to explain the main forms of PET available today, while providing tangible use cases. While being a great idea, Silvia's plan faces a few challenges. Firstly, Silvia will have a very hard time getting access to the images needed to train the computer vision model. Data coming from medical records is highly private and fragmented: scattered across dozens of clinics, each requiring the consent of all of its hundreds of patients — and alternatives are scarce. Without data Silvia could not

go far. Secondly, STDHelp's strategy requires handling and sharing very sensitive information about its users with its own service providers (CV-ML and others):

- identity and demographic information,
- whether they are affected by an STD,
- receiving and storing their private pictures,
- etc.

It's hard to imagine users feeling comfortable using such a solution. Finally, handling that data would be both a regulatory nightmare and potentially a large liability.

PETs are a set of cryptographic methods, architectural designs, data science workflows, and systems of hardware and software that enable adversarial parties to collaborate on sensitive data without needing to rely on mutual trust.

Silvia is well aware of these problems and knows that PETs can help. In an example case like this, PETs could mitigate the challenges by preventing any party (internal or external to the process) from learning sensitive information about any other party. With PETs Silvia could train CV-ML's computer vision model on the sensitive patient pictures held by clinics — without ever actually accessing, seeing or moving from their secure premises those private pictures — and develop a very accurate detector for her medtech application. When an STDHelp user uploads a picture for diagnosys, he could be sure that neither STDHelp nor CV-ML (or anyone else) will ever be able to learn about his identity or physically see the picture (these are examples of "*Input Privacy*"). In parallel, PET could assure users that even if an eavesdropper with malicious intent (for example blackmailing the user!) manages to get their hands on the diagnosys, this would not leak any information about the picture that was analyzed, the identity of the user, the diagnosys, and more (these are examples of "*Output Privacy*"[1]). PETs don't only benefit end users. In a case like this, PET could ensure CV-ML that its proprietary computer vision model would remain secret even when used by STD Help so that its IP could never be leaked or stolen.

STDHelp uses PETs to solve its challenges in securing Personally Identifiable Information (PII), intellectual property and trade secrets.

Zooming out from this singular example case, we can see that PETs are going to revolutionize healthcare (and almost any other sector) over the next 10-20 years. This would unlock advances in precision medicine, allowing vast scale machine learning projects on data which today is siloed inside hospitals; and favoring a general step change in digitizing medicine — likely creating hundreds of billions of dollars in value along the way

[1] This is not a full categorization of the types of privacy. In fact, workflows where PETs are deployed are often very complex and very different from one another, with many players being involved. When setting up systems with so many parties, confidentiality becomes harder and involves consideration on who is allowed to know what objects, what kind of leakage is catastrophic and which is considered acceptable, and so on. These questions are usually dealt with on a case by case basis. Some of these can be dealt with systematically, using "data access" or "data governance" policies which are usually orthogonal to PETs and outside the scope of this document.

— along with incalculable improvements to human lives.

A large majority of digital businesses, and many traditional businesses, increasingly face such hurdles due to challenging regulations, corporate policies, data breaches and competitive concerns. It is safe to assume these challenges cause an aggregate of billions of dollars in losses and missed opportunities. In general, handling of sensitive data is a big liability for all parties involved, and the businesses that solve these challenges will be well positioned to benefit from a once-in-a-lifetime opportunity.

Generalizing from our mock application example, some of the uses of PETs today can be summarized as follows:

- **Performing analysis on encrypted data** — removing the need to trust an analytics provider.
- **Enabling multiple entities to confidentially collaborate in data analysis** — by privately providing their respective input to a joint analysis and sharing its output.
- **Analyzing data directly where it is stored** — removing the need to physically move and collect data, and enabling only specific analyses to take place directly at the secure premises of the data owner.
- **Anonymizing sensitive data** — before sharing it with other internal departments or to third party organizations.
- **Using synthetic data** — replacing the original data to completely remove the need to handle sensitive data, while being able to run the same analyses.
- **Protecting the intellectual property, trade secrets and know-how** — while it is being used by third parties which may be tempted to steal it.
- **Creating a "safe haven" for sensitive data and applications** — ensuring that these are firmly partitioned from the rest of the system (phone, laptop, etc).
- **Ensuring results of aggregated analyses do not reveal any information about individual inputs.**
- **Proving a quality or a statement** — without needing to provide any supporting evidence.

Thanks to the "horizontal" nature of their capabilities, PETs hold the power to completely transform how we interact with data, across virtually any industry (see an overview of early applications further in the document).
PETs will soon get adopted by every company that works with, or wants to work with more sensitive information; that has valuable IP and know-how that is held back by competitive concerns; or that wishes to use outsourced (i.e. cloud) analytics services on data it does not want or cannot share. All these will prove as an edge in competitiveness, compliance and innovation across all sectors.
Understanding the economic value behind this secular transition is hard. Nonetheless, in the next section, we offer a simple attempt at sizing at least its order of magnitude.

What is the opportunity?

Like previous waves of cryptography, PETs adoption could unlock a trillion dollar opportunity by helping us extract more value from existing data, driving the creation of even further data, and enabling a new generation of services and use-cases to flourish.

Privacy Enhancing Technologies have relatively recently become a focus of the private sector and with it, a new market opportunity is growing. Venture capitalists have invested more than $850m into the space as the transition from academia to commercial applications has already begun. The economic opportunity of PETs is present but still challenging to quantify as industry adoption is so early in its lifecycle. Further, in the case of infrastructural technologies (like fibre optic networking, internet browsers, smart phones and 5G), the bulk of the value is not created directly as a well defined market size but rather indirectly by unlocking opportunities in other industries. As an example, the entire market referred to as 'The Future of Work' — that is growing exponentially since 2020 — is dependent on advancement in a range of underlying infrastructural technologies.

Trying to evaluate a market size in the case of PETs might just be putting the cart before the horse.

In predicting adoption of infrastructural technologies, founders and investors alike search for the 'killer use case' as a first vector to drive then mainstream adoption across other verticals. Blockchain and VR are examples of infrastructural technologies looking for mainstream adoption through 'killer use cases'. Since we see the potential to unlock value as a precursor to market size, this analysis focuses on the potential value that could be unlocked by PETs.

Privacy Enhancing Technologies have recently become a focus of the private sector and with it, a new market opportunity is growing.

Our estimate of the potential value unlocked by privacy enhancing technologies is $1.1T to $2.9T (USD).

We define this value as contribution to global GDP through additional digital and data-driven commercial activity over 20 years to 2040. Two methodologies are presented to support our estimate:

1. comparison to economic value created by historical asymmetric encryption adoption
2. estimating incremental data usage that is possible thanks to PETs' unique capabilities.

We present two approaches because the range of industries impacted is so broad that introducing a second methodology offers a different perspective to verify the underlying impact. To support the argument, we make two main assumptions:

1. Like before the advent of asymmetric encryption, there exists information that is not being shared or fully exploited due to various concerns (competitive, regulatory, etc.)[2]
2. A substantial amount of that data will be unlocked by PETs, and will result in similar economic growth and the rise of new sectors.

Under these two assumptions, we hypothesize that PETs will play a key role in the next two decades of the digital economy, creating new sectors, unlocking dramatic value in the trillion dollar-range to global annual GDP.

Comparison to Asymmetric Encryption

The purpose of PETs is to increase the confidentiality of digital operations. Their economic impact can be estimated by looking at the historical impact of technologies that have increased confidentiality in the past. One such example is asymmetric encryption — the underlying technology used in TLS (transport layer security) protocols that secure communication over the internet.

The introduction of asymmetric encryption in the 90s and its expanded adoption in the 2000s enabled the development of new internet use cases beyond web surfing. You may recall the time when common wisdom suggested that using credit cards online was too "dangerous". Indeed before the introduction of TLS, credit card information would be delivered in plain-text and was open to eavesdropping and man-in-the-middle attacks, leaving early adopters exposed to the risk of financial fraud. Concert tickets were an early product to move online, but it was only a few brave souls (or overly committed Spice Girls fans) that would take the risk. The adoption and evolution of online transactions from credit card purchases to online bank statements, to online payments to mobile-first banking was a decades long transition. The 'dot-com' bubble attempts to migrate to 'the net' consumer purchasing like books (Amazon), pet supplies (pets.com) or used goods (eBay) predate mainstream adoption of online purchasing. The mainstream adoption and consumer trust in any ecommerce company is wholly reliant on the use of asymmetric encryption. As every retailer is now forced to have an online presence during the 2020-2021 pandemic and every consumer must be comfortable purchasing online, it is clear that asymmetric encryption is a key infrastructural technology unlocking value through the use case.

The introduction of asymmetric encryption secured data in transit, enabling businesses and consumers to securely share data, relying on mutual trust, enabling the development of new internet use cases beyond web surfing.

[2] It is further arguable that what is preventing us from gaining more value out of data today is not the lack of more computing power or the ability to move data around in a more expeditious and efficient manner (like it was the case two decades ago). The bottleneck is rather our inability to exploit more and better the ever increasing amount of data, and do so catching up with more stringent regulatory frameworks — addressing rising public and private concerns on how our data is being gathered, stored and used.

Asymmetric encryption secured data in transit, enabling businesses and consumers to securely *share data, relying on mutual trust*. This made sectors like ecommerce, ebanking & fintech, and pretty much any other cloud-based business rise and thrive[3]. PETs, a new wave of cryptography, are now pushing the data economy one step further, enabling businesses and consumers to securely *share only the value of data, without having to rely on mutual trust.*

The United Nations Conference On Trade And Development (UNCTAD) estimates the digital economy accounted for up to 15.5% of the world GDP or a total $11.5T[4] in 2019. It is arguable that asymmetric encryption was a key enabler for at least ecommerce ($2.4T contribution to GDP) and cloud and e-services ($470B to GDP)[5]. Encryption is not singularly responsible for unlocking this $2.9T value creation, and many other technologies and trends (broadband communication, internet penetration, etc.) were also vital to the success of these sectors. However cryptography is still a key driver for their existence and responsible for a material part of this growth. To put things into perspective this impact was achieved over roughly a 30-year period, starting around 1990. In comparison, we estimate that PETs are already roughly around 10 years into this “30 year” period.

Now that we have estimated that asymmetric encryption was a key component in creating upwards of $2.9T of annual GDP, we argue that PETs are likely to have a similar economic impact over the next 20 years.

PETs, a new wave of cryptography, are now pushing the data economy one step further, enabling businesses and consumers to securely share only the value of data, without having to rely on mutual trust.

Estimating data value unlocked

An alternative analysis can be based on the value of data in the digital economy. It has long been a cliche to describe ‘data as the new oil’: the metaphor makes the salient point that data has created new business models in adtech, increased productivity with machine learning models and improved industrial operations through smart sensors. Data is not universally good (for example, it can reinforce biases through recommendation engines), but for better or for worse data is immensely valuable, and clearly has the potential to create economic value. Privacy enhancing technologies expand the use cases of data by making it more readily available to more parties. Consequently, PETs will unlock value by expanding the use of data beyond its current set of use cases.

McKinsey estimates that only 1% of the world’s data is being used for analysis.

[3] See here for further reading on cryptography adoption and importance in internet sectors in the early 2000s

[4] Digital Economy Report 2019, UNCTAD

[5] 2020 data from Statista

McKinsey estimated that only 1% of the world's data is being used for analysis[6]. It is reasonable to assume that a part of the value of the digital economy is dependent on data being analyzed, and the insight from McKinsey further reinforces our previous assumption that there exists information that has value and that is not being used today.

From these premises, we make two additional claims:

1. At least 10% of the $11.5T value of the digital economy is based on analyzing that 1% of data being used; and that
2. PETs thanks to their revolutionary capabilities will allow us to utilize an additional 1% of the data available over the next 20 years (which is likely a conversative estimate).

If the above appears reasonable, then we can estimate that PETs could unlock more than $1.1T[7] in annual global GDP in a 20 year timeframe.

PETs will enable to utilize data we are not currently using for analysis, and expand the universe of use cases for data we are already using.

Now that we have discussed which problems PETs solve, and what kind of impact one could expect from these technologies, we are ready to dive deeper into why companies adopt them, what benefit they get, and what costs come with PETs' adoption.

[6] McKinsey, 2015

[7] The above analyses attempt to quantify the value that will be created by PETs. They do not quantify the revenue potential or the market size of the PET vendors themselves. It is difficult to estimate the market size of PET vendors today and we leave this question unanswered.

Goals, benefits and costs of adopting PET

PETs' main goal is to increase the level of confidentiality when multiple parties collaborate together on sensitive data. PETs drastically improve the trade off between generating value from data and respecting its secrecy, but can incur meaningful accuracy and efficiency losses.

In this section we address what benefits and costs are related to PETs' adoption and deployment. Indeed PETs impose non obvious business and technical trade-offs which should be carefully considered vis-à-vis the specific goals of a given organization.

PETs' main goal is to increase the level of confidentiality when multiple parties collaborate together on sensitive data. The technical use cases that are possible by applying one or several PETs create benefits much like asymmetric encryption did in the 90s. Similarly, PETs are expected to unlock immense value much like asymmetric encryption did in the 2000s/2010s, by enabling new opportunities like social networks, eCommerce and all the other cloud enabled business we enjoy using today.

PETs' main goal is to increase the level of confidentiality when multiple parties collaborate together on sensitive data.

Benefits of PET adoption

The following is a brief and non-exhaustive list of benefits from employing PETs, though we expect entrepreneurs will soon expand this list far beyond our limited VC imagination:

- **Increased collaboration:** In our example earlier in this paper, PETs helped ML-CV and STDHelp collaborate in a confidential way on an innovative product. ML-CV is assured that its machine learning IP remains secret, and STDHelp is assured that the privacy of its users is protected. In general, PETs can help internal departments or different organizations collaborate with their respective data and IP, while being assured that no confidential information is being disclosed in the process.
- **Improved trust in outsourcing data analyses:** PETs can assure companies that their sensitive data are kept confidential when analyzed by third parties (e.g. from any cloud service).
- **Regulatory compliance:** By applying PETs, STDHelp avoids onerous regulation (e.g. GDPR) and reduces the inherent liability related to handling very sensitive user pictures.
- **Competitive differentiation:** by adopting PETs STDHelp builds a strong value proposition centered on privacy — and stands out among competing services. PETs can help companies reposition vis-a-vis rising concerns and gain more adoption from privacy-aware users.
- **Faster product testing and development:** increased access to data needed for product development and testing shortens times to build and bring data-enabled products to market.

- **More data and more services:** by winning users' confidence STDHelp will gain further adoption, consequently gaining access to more data, and thus will be able to develop new services. The confidential use of data pushes users and companies to share more information previously considered too sensitive (you can think of DNA, browsing history, supply chain information, etc.) — allowing service providers to develop applications & services previously unviable due to regulation or lack of data.
- **Reduced cyber security risk:** With PETs data is less exposed as it can be secured when at rest, in transit and in use. The financial impact of cyber incidents is dramatically reduced by eliminating the exposure of sensitive information.
- **Protection of trade secrets:** PETs assure ML-CV that its proprietary algorithms for STD detection can be deployed outside of their premises or private cloud without any risk of disclosing their IP.

As we will see when reviewing the techniques, different PETs offer different privacy protections and benefits, and satisfy different goals. Building a PET-powered application to address a specific use-case implies understanding the nuances related to each specific PET technique, and how to best combine them into a single comprehensive system.

Cost of adoption

Now that we have a clearer picture of what one can get out of PETs, it is time to look into their downsides. Indeed, the adoption of PETs offers great advantages to organizations and individuals alike, but it does not come without friction and costs.
To do so we first need to understand the concepts of data utility and utility costs. We have already mentioned that one of the key purposes of PETs is to improve the longstanding balancing act between being able to analyze data and respecting its confidentiality. There are many things we can do with data that create utility[8]. The more things we can do with data (types of analyses we can run on it, wealth of insights we can generate from it, etc.), the higher its utility.

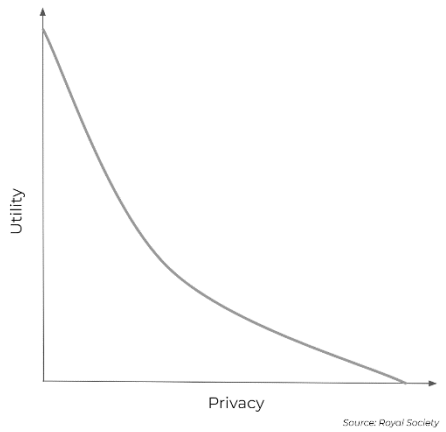
Traditionally more confidentiality meant less utility: the more we want to keep our data secret, the less it is fungible and useful, and vice versa.

Thus, data has its maximum utility when it is in the clear, in plain text: we can deep dive into it, analyse it in all sorts of ways available to us, and we can transform it as we see fit to allow even more useful analyses. In order words: plain text data is fully fungible. However maximum utility comes at the cost of zero confidentiality: plain text data reveals all of its secrets to anyone receiving it or handling it.

[8] Utility can be defined as the amount of value we can extract out of data and how efficient (computing cost, latency, communication costs, etc) that process is: the higher the efficiency, the higher the value. The maximum utility is benchmarked to the analyses we can do and the insights we can generate when the data is in the clear.

Conversely, data can have maximum confidentiality when it is encrypted. Plain text is transformed into ciphertext — a string of indecipherable values — at the cost of its utility: encrypted data is typically not (or not fully) fungible, and thus without (or having reduced) utility.

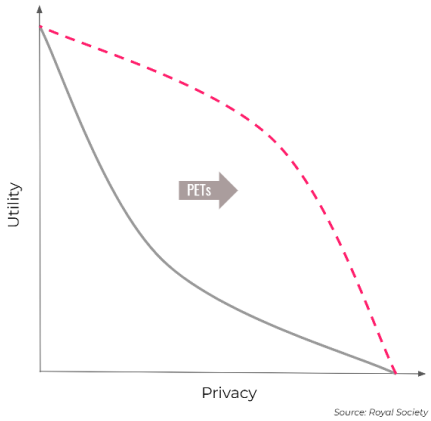
Interpolating between these two points — (*max. utility; min. confidentiality*) vs. (*min. utility; max. confidentiality* — we can plot a curve which defines the relationship between utility and confidentiality.



As we can easily see, an increase in confidentiality is associated with a reduction in utility: *the more we want to keep our data secret, the less it is fungible and useful*, and vice versa.

PETs manipulate this curve by shifting it upwards and to the right, into a much more favourable balance between confidentiality and utility.

PETs do not exempt us entirely from the trade-off between confidentiality and utility. PETs make the data analysis process less efficient, reducing utility



However, PETs do not exempt us entirely from the trade-off between confidentiality and utility. The use of PET often diminishes the level of utility relative to plain-text data by making the data analysis process less efficient. These decreases in efficiency are called “utility costs”. The nature and intensity of such costs varies significantly based on the specific PET technique being considered and the architectural choices being made. Two systems using the same PET techniques may show different profiles, based on how the system is conceived, developed, and deployed.

Typically, PETs may incur one or more of the following utility costs:

- **Accuracy & efficacy:** some PET techniques produce significant changes to the sensitive data being analyzed, reducing the accuracy of the analysis being performed. Some others limit the types of analysis we can perform on data — and our ability to produce a full spectrum of insights.
- **Computational overhead:** very often the adoption of PETs comes at the cost of extra computational power. Simple computations may become

very large when adapted for specific PET workflows — at times even totally preventing the economic or practical viability of the use-case being explored.

- **Communication overhead:** some PETs leverage distributed analysis, thus requiring intense communication among the parties involved. Others create a large expansion of the data, putting a strain on the communication bandwidth of the system.
- **Latency:** the time delay between the start of an analysis and its conclusion constitutes its latency, often as a result of the communication and computational overheads of PET. This latency, when compared to an analogue non-PET analysis, may suffer dramatic regressions. In other words, analysis under PET may take significantly longer times to conclude.
- **Integration & engineering complexity:** the adoption of PETs almost always implies building complex systems and architectures and integrating them into existing workflows. These tasks require expert resources and often result in engineering costs which would not be incurred otherwise.

When building PET-powered applications, organizations should perform a thorough analysis of the benefits and costs related to addressing a specific use-case. Even when the benefits of adopting PET are immense, the cost related to their adoption could still be economically unsustainable.

Now that we have a general understanding of the benefits and costs associated with adopting PETs, let's proceed with an introduction to some of the techniques we are most excited about here at Lunar.

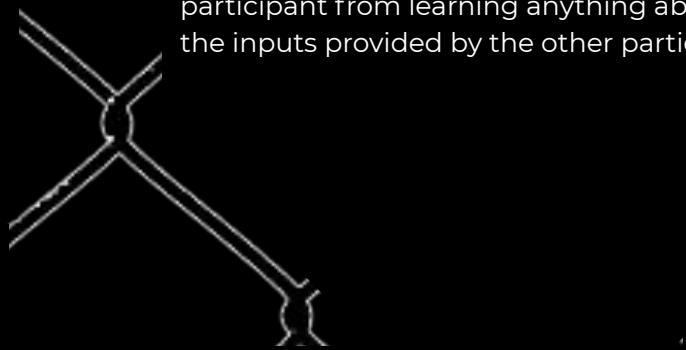
Section 2: PET techniques

Homomorphic Encryption

Homomorphic encryption cryptographically ensures that both the data and the result of the analysis remains secret, removing the need to trust the location where the analysis takes place.

Secure Multi-Party Computation

Secure Multi-Party Computation enables multiple mutually-distrusting parties to collaborate on a joint analysis on confidential data, preventing any participant from learning anything about the inputs provided by the other parties.



Differential Privacy

Differential privacy's main goal is to protect the privacy of any individual providing his information to a database that is used for aggregate analysis.

Zero-knowledge Proofs

Zero-knowledge-proofs allow data provided by one party to remain secret while being verified by another party. It acts as an auditing system which allows the underlying information not to be disclosed in full to the auditor.

Synthetic Data Generation

Synthetic data is an artificial data set that mimics the properties and correlations of an original, confidential dataset

Federated Machine Learning

Federated learning ships machine learning models to the locations where the data is stored to perform the training locally, decoupling the training process from the need to access, share and store all the data in a centralized location.

Trusted Execution Environments

A Trusted Execution Environment is a secure partition of a larger chip/SoC that secures the execution environment of the analysis by isolating it completely from the rest of the machine processes.

Section 2: PET techniques

PETs are a set of cryptographic techniques, architectural and protocol designs, data science workflows, and systems of hardware and software that enable adversarial parties to collaborate on sensitive data without needing to rely on mutual trust.

In this section we offer a high-level overview of some prominent PET approaches and techniques. For each technique we present a very brief technical and scientific overview, outline some scenarios where the technique is useful, address its limitations and shortcoming, and present some opportunities for innovators in the space.

It's important to note some of the below are proper cryptographic techniques (or inventions), some are more general approaches (e.g. a family of techniques), and some are very general goals which may be achieved through any set of techniques or mix and match thereof.

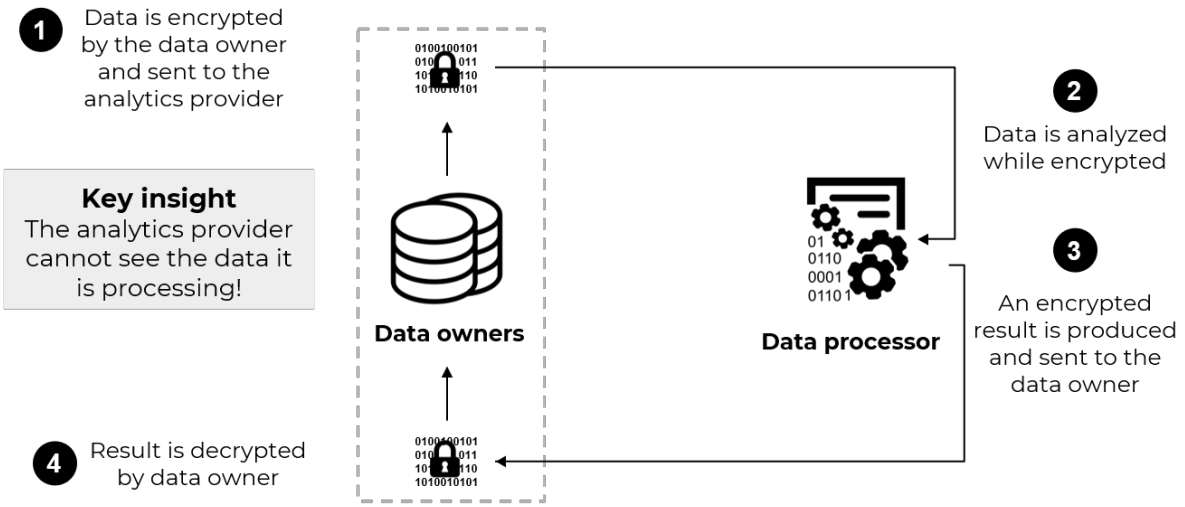
Since we are a fund that specializes in computer science, we decided to place particular focus on the techniques that have cryptographic roots and are primarily software-based[9]. Specifically, we look at Homomorphic Encryption, Secure Multiparty Computation, Zero-Knowledge Proofs, Differential Privacy, Synthetic Data, Federated Learning and Trusted Execution Environments.

Homomorphic Encryption

Homomorphic encryption cryptographically ensures that both the data and the result of the analysis remains secret, removing the need to trust the party performing the analysis or the location where it takes place.

Homomorphic Encryption (HE) is a property of certain encryption systems which allows computation to be performed directly on encrypted data without needing to decrypt it, and producing an encrypted result. Thus, HE systems protect both the input and output data, at rest and during the computation. HE is ideal when outsourcing computation on sensitive data to unsecure environments or to untrusted parties; HE cryptographically ensures that both the data and the result of the analysis remains secret.

[9] The universe of privacy-enhancing techniques is actually much broader and includes pure design choices like enforcing stringent best practices in user data policies, access controls and others. We purposely ignore that part of the spectrum, as we feel it is well covered elsewhere and fairly orthogonal to the "trustless" spirit of computer-science-based techniques.



Partial Homomorphic Encryption schemes, which allow some simple restricted forms of data analysis, have been known since the 1970's. In 2009, Craig Gentry described the first Fully Homomorphic Encryption (FHE) scheme — which in principle allows *any* data analysis to be performed on the encrypted data. However, Gentry's method was not optimized for real-world usage, and is too slow for real world applications due to the massive computation overheads. In Gentry's method, the Data Processor can perform its computation directly on the encrypted data, but that computation might take a billion times more CPU resources than performing it directly on the unencrypted data. A large amount of research work has been dedicated since 2009 to invent simpler and faster schemes. For example, in 2013 IBM Research released HElib, a library which improves the performance of FHE by several orders of magnitude over Gentry's method[10]. Today, there are multiple open source homomorphic encryption libraries[11] available and suitable for different applications. One approach we find particularly promising is that of our portfolio company Zama, creator of the Concrete HE library which greatly reduces the overheads incurred when applying neural networks directly on the encrypted data..

Utility costs & limitations

HE is one of the most powerful PETs, but it also creates dramatic overheads in computation, latency, and communication. Most HE methods require any computation to be re-written as a circuit of logical gates — a process that in many scenarios can turn a small program into a gigantic computation. The larger the circuit — the larger the CPU cost of HE. For some types of computations, a transformation to small circuits can be automatically achieved (e.g. this is the case for Zama, which automatically translates neural networks into circuits); in other types of computations, achieving realistic running times requires a specialist to

[10] Gentry's method is considered a "first-generation" method while today's methods are considered "fourth-generation methods": for a detailed a timeline see e.g. HE on wikipedia

[11] *A Review of Homomorphic Encryption Libraries for Secure Computation*

perform this transformation. The additional CPU cost incurs higher energy consumption, cloud costs, and latency.

Historically, performance slowdowns have prevented real-world adoption. But with increased interest from industry, we are witnessing a "Moore's Law for HE": an onslaught of advancements in both science and engineering are quickly improving efficiency of HE methods.

In HE systems, the encrypted data is much larger than its unencrypted version (the "plaintext"). This blow-up in data size increases the load on communication bandwidth, which consequently also increases latency.

In some applications, HE can be challenging to integrate with existing systems without requiring substantial changes in established workflows and data pipelines. This often results in high engineering costs. Additionally, HE is a low level cryptographic primitive and in some applications, building secure protocols based on HE may require the expertise of an expert cryptographer.

As HE is making its way into real world use-cases, some concerns have been raised around the lack of verifiability of computation: since outsourced computation is the leading use-case for HE, the data owner's ability to verify that the computation was executed correctly will likely be a critical and non-trivial request. However, efficient mechanisms for verifying the integrity of the performed computation are quite easy to add to HE libraries, and this shouldn't be a concern in the medium-to-long term.

Opportunity

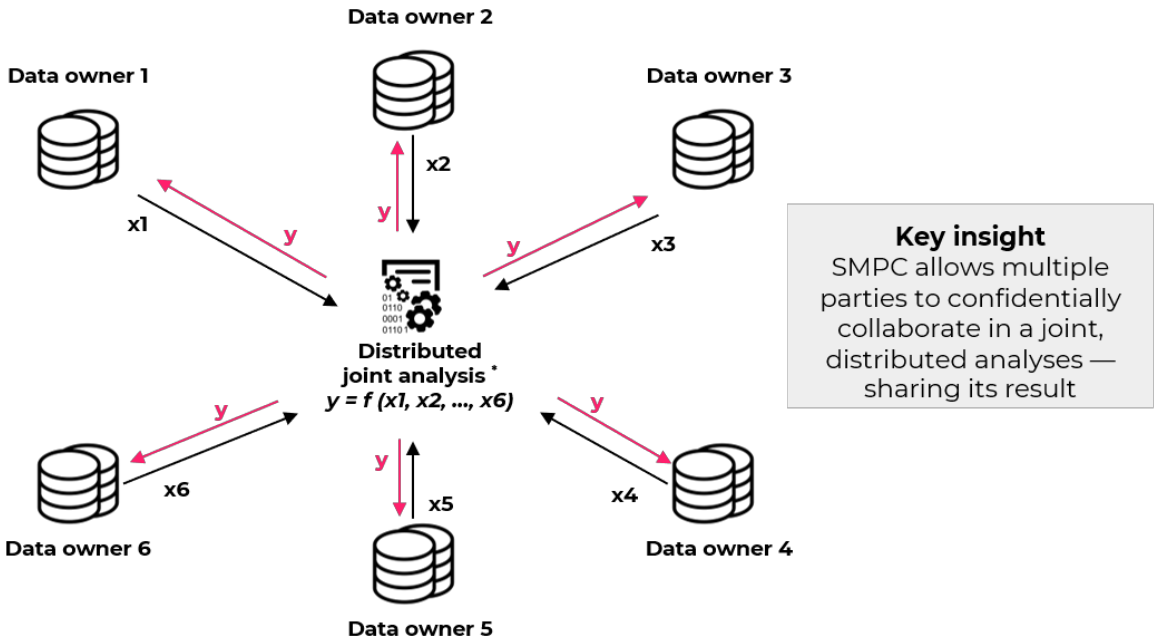
From a theoretical perspective, HE is fully proven to work securely and correctly. But applied solutions are still very early in their industry adoption, and are often limited to small applications that perform simple analytics. More complex solutions are only now being productized. Most of these are building the bottom-of-the-stack infrastructure that will make HE viable for many applications over the medium term.

Historically, performance slowdowns created by HE have prevented its real-world adoption. But with increased interest from industry, we are witnessing a "Moore's Law for HE": an onslaught of advancements in both science and engineering are quickly improving efficiency of HE methods. As a result, HE is becoming fertile ground for industrial PoCs, pushed by a new generation of startups. We expect that strong HE development tools will emerge in the coming years, improving access of non-experts to this technology.

One unique potential application of HE is "Private Machine Learning": allowing to run machine learning models on highly-confidential data. We expect to see substantial level activity and innovation happening in this domain.

Secure Multi-Party Computation

Secure Multi-Party Computation enables multiple mutually-distrusting parties to collaborate on a joint analysis on confidential data, preventing any participant from learning anything about the inputs provided by the other parties.



Key insight
SMPC allows multiple parties to confidentially collaborate in a joint, distributed analyses — sharing its result

* this is a visual simplification. The analysis does not run in a central location. SMPC enables distributed analyses. The computation occurs in a distributed fashion at each of the participating nodes, coordinated by the SMPC protocol.

Secure Multi-Party Computation (SMPC), is a family of cryptographic methods with the goal of enabling multiple mutually-distrusting parties to perform a joint computation, while keeping their respective inputs secret. One simple example is an auction: during an auction the winner is selected by performing the computation: "what is the highest bid?". This computation is normally done by an auctioneer, with full visibility on the bids of each participant. SMPC systems however allow performing such computation even in cases where no party trusts another (not even the auctioneer!) to know the bids — yet a winning bid still needs to be decided (see more below). SMPC prevents any participant from learning anything about the inputs (e.g. bids) provided by the other parties. To perform the joint computation, the participating parties follow a *communication protocol*: a set of instructions and inter-communications that, overall, implement a distributed program. Ultimately, SMPC allows parties to collaborate without the need for a trusted authority: were a trusted authority to exist (i.e. an arbitrator or judge or an auctioneer), participants could send their confidential inputs to this trusted party, which could compute the answer and broadcast it back to the participants. SMPC achieves the exact

same behavior, but replaces the trusted third party by a cryptographic protocol.

Small-scale SMPC implementations have been used for some time now, e.g. in order to enhance the security of key management systems.

SMPC was first formally introduced as secure two-party computation (2PC) in 1982 by Andrew Yao in the context of "Yao's Millionaires' Problem". SMPC protocols typically have high requirements in both CPU, bandwidth and latency. This caused potential industry applications to stall for many years. But in recent years, SMPC's commercial availability has been increasing. A famous early application of SMPC is the 2008 application to Danish Sugar Beet Auctions. SMPC is often applied to spectrum auctions — indeed, auctions are a natural setting where multiple parties (the bidders and askers) need to compute a joint result (the market equilibrium) while the inputs themselves (the bids and asks) are often commercially (or otherwise) confidential. Auctions are also high-value processes which are not sensitive to latency — a setting that lends itself to SMPC's strengths and weaknesses. As SMPC increases in efficiency, we expect the sphere of viable use cases to grow dramatically.

Utility costs & limitations

The most challenging aspect of bringing SMPC to market is system integration. When running an SMPC protocol, all participants learn the result of the computation. Yet, the result might carry some information on the underlying confidential data. If the parties repeatedly apply the SMPC protocol, more and more data will "leak". Quantifying and controlling such leakage is best done by a cryptography expert. SMPC protocols are also challenging to integrate without creating other potential vulnerabilities that may be exploited by a sufficiently-motivated attacker to learn confidential information. Overall, while SMPC is a provably-secure technique in theory, it requires a strong security expert or cryptographer in order to integrate properly and securely. Due to these challenges, it is difficult to "productize" SMPC. Horizontal companies that focus on SMPC have traditionally been consultancies, rather than product startups. But there are early signs that this might be changing. The difficulty of establishing the practical security of deployed SMPC is also a potential blocker to adoption: techniques that are so prone to bad integration often arouse justified suspicion by developers and users alike.

While the theory behind SMPC is relatively mature, actual commercial products are still very early in development, mainly held back by integration challenges.

SMPC also creates substantial overheads on computation and network communications, often increasing such resource requirements by several orders of magnitude. Furthermore, since SMPC is a distributed software running on a set of interconnected nodes, it heavily taxes the underlying network and on the participants' CPU resources. Large scale deployments which are sensitive to

latency may incur high engineering expenses.

Opportunity

Like other cryptographic privacy-preserving technologies, SMPC began being integrated into real world applications towards the end of the 2010s. Applications into mainstream market segments are still in their early phases, with stakeholders building up confidence that SMPC can deliver on its claims. While the theory behind SMPC is relatively mature, actual commercial products are still very early in development. Simpler use-cases where computations are mostly local and with fewer interaction among parties — like distributed voting, private bidding and auctions, and key management — have reached an early product stage in the market. While ease of programming and integration is a critical factor to potential end users, not much has been done to develop the required tooling, limiting SMPC to bespoke applications addressing one or a few specific use-cases. SMPC is difficult to configure correctly and currently requires highly customised client and server software for deployment. Companies and startups are now shifting their attention to enabling more general purpose and configurable systems to support multiple use-cases with a common layer of enabling infrastructure.

In the medium-to-long term, a wealth of high-value applications exists. Indeed, our guess is that SMPC will be a vital component in a large universe of data collaboration applications which are going to create immeasurable business value: much of the trillions of dollars of value yet to be created by PETs is likely to stem from such data-collaboration applications powered by SMPC.

Zero-Knowledge Proofs

Zero-knowledge-proofs allow data provided by one party to remain secret while being verified by another party. It acts as an auditing system which allows the underlying information not to be disclosed in full to the auditor.

Zero Knowledge Proofs (ZKP) are a family of cryptographic methods whose goal is to allow data provided by one party (the prover) to be verified by another party (the verifier), without revealing that data in the process. A simple example would be proving I have a positive credit score without providing any of my financial information (e.g. for privacy reasons). A sound application of ZKP ensures that if (and only if) the proof is correct, it will be accepted by the verifier and that the proof itself is the only information shared with the verifier. The verifier is able to rely on the proof because the system is built in a way that the probability of the proofer being able to cheat is statistically negligible[12]. ZKP obviously provides Input Privacy, as the proof does not reveal any information about the private input used. ZKP also provides a guarantee of output

[12] See the Alibaba Cave story for a demonstration of this mechanism.

correctness (guarantee that the claim of the prover is true if the verification is successful).
ZKP can act as an auditing system whenever the underlying information is private and it should not be disclosed in full. Early use-cases span from KYC, credit scoring, asset or data custody to digital advertising, where ZKP can prove that a target viewer matches the criteria of a campaign without revealing any sensitive information about her. Another similar area which is seeing increased adoption is around identification and authentication, where ZK-powered tools can allow users to authenticate while keeping their identity secret.
The advent of blockchain has pushed the adoption of an increasing number of applications that leverage ZKP. Maybe the most notable application of ZKP is the cryptocurrency Zcash, which, using zero-knowledge Succinct Non-interactive Argument of Knowledge (zk-SNARK), has dramatically enhanced the level of privacy by keeping secret transactions' and users' data.

Zero knowledge proofs were introduced in the work of Goldwasser, Micali and Rockoff in the 1980s at MIT. Since then academia has developed different types of schemes depending on the types of statements supported, the level of interaction needed between parties, and how the overheads of the process are split among the parties. For example, some zero-knowledge systems require that the prover and verifier interact during the verification of the proof while others enable asynchronous generation and verification of the proof.

Utility costs & limitations

The utility cost of using ZKP varies widely based on the type of system and the characteristics of its specific design (level of interaction among parties, size and complexity of the proof, etc), which may reflect in computational overhead, latency and strain on communication bandwidth. For example, the efficiency of ZKP is often dictated by the length of the proof and the computation complexity of the prover and the verifier.
The design of the system may also have meaningful implications on who bears those costs and how they are spread across prover and verifier.
Compared to HE and SMPC, ZKP is best suited to address only specific types of use-cases that can be reconducted under the general need of confidentiality while verifying information.

ZKP has seen meaningful level of adoption in the blockchain space — a trend expected to replicate across other industries.

Opportunity

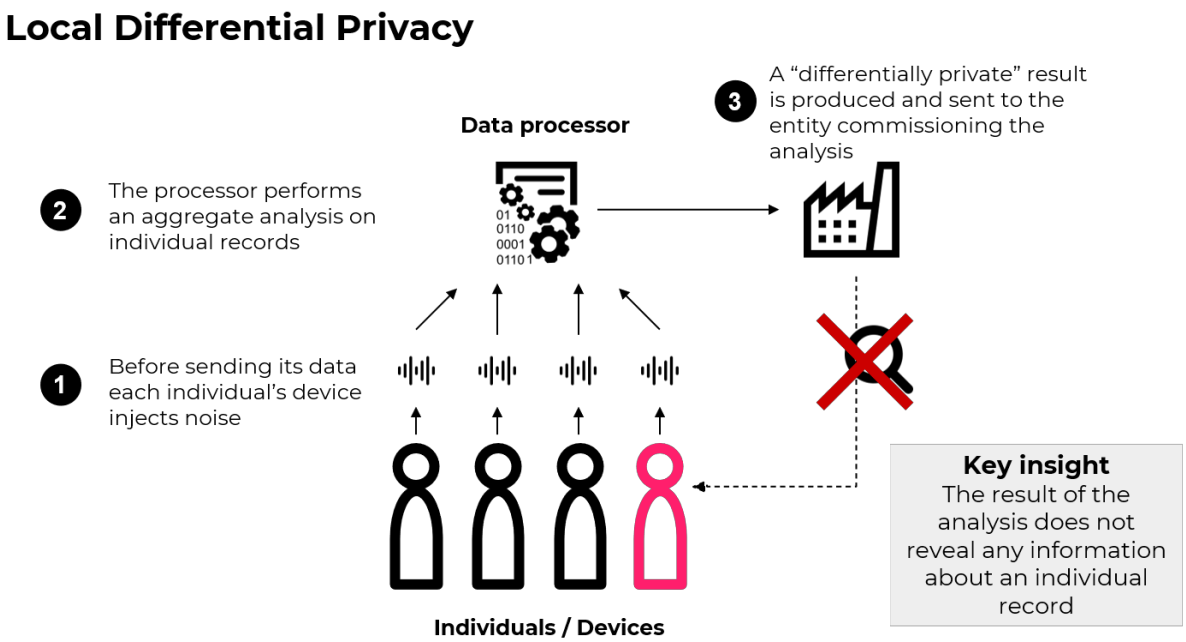
Overall the ZKP space is still very early in its development, with a lot of innovation happening in academia or through technical papers of independent researchers. When it comes to transforming ZKP science into technology, the space still suffers from a lack of standardization: efforts in this direction started only mid-2018, but ZKP is still fragmented across a multitude of different languages and implementations which hamper compatibility.

ZKP adopters are expected to be experts in this field. Successfully employing ZKP requires being able to appreciate the small nuances between different systems and implementations, which can result in meaningful differences in outcome when applied to real use-cases.

Differential Privacy

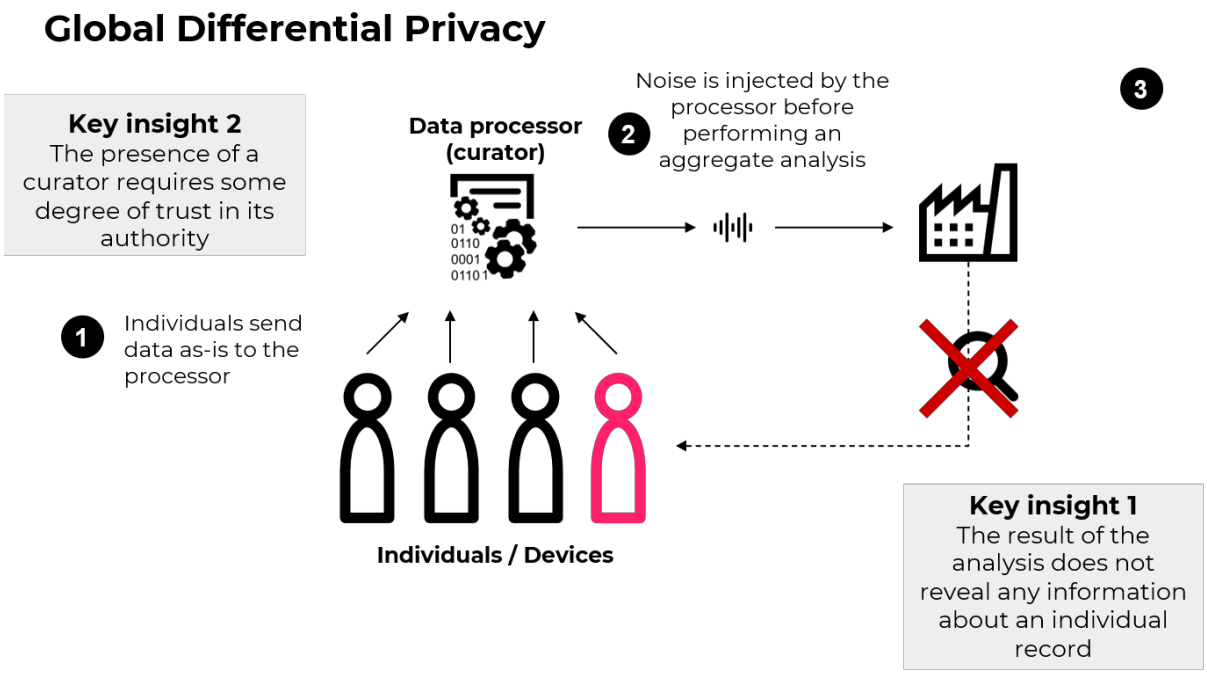
Differential privacy's main goal is to protect the privacy of any individual providing their information to a database that is used for aggregate analysis.

The main goal of Differential privacy (DP) is to protect the privacy of any particular record of data that's used as one of the inputs to statistical analysis. Thus DP deals with Input Privacy. A simple example is my personal medical record that is being stored within a hospital's database alongside many other patients' records, and used for aggregated statistics. Differential privacy quantifies and limits the amount of information that could be revealed about myself (or any other individual record) from the output of the statistical analysis.



DP is an intrinsic property of the analysis process: it requires an aggregate analysis to provide comparable statistical output even when removing an individual input from the database (regardless of the input being removed and regardless of the database being used). In practice DP is achieved by adding a certain level of noise (or obfuscation) into the analysis. Noise can be applied by individual data owners before committing their data

to the dataset (Local DP) or by a curator[13] at a later time (Global DP). Ideal use-cases are large statistical analyses (like Apple or Samsung gathering usage statistics from their users, or analyses performed by national statistics organizations).



Differential privacy is just now starting to be used in real use-cases. A healthy number of new companies have begun developing new products, leveraging differentially private statistics to address specific niches. Recently, the focus has shifted to providing configurable infrastructure that would make DP available for a larger and expanding set of applications.

Differential privacy has its roots in statistical organizations, which have long collected data under a promise that it would only be used for aggregated statistics and could not be traced back to a specific individual. With the explosion of the amount of data, the amount of work needed to keep that confidentiality promise led to a need for an increase in research. In the early 2000s it became clear that keeping confidentiality of private data requires adding some amount of noise in the dataset (the *Fundamental Law of Information Recovery).* Shortly after the amount of noise was quantified by Cynthia Dwork who formulated that *"overly accurate answers to too many questions will destroy privacy in a spectacular way"*.

Utility costs & limitation

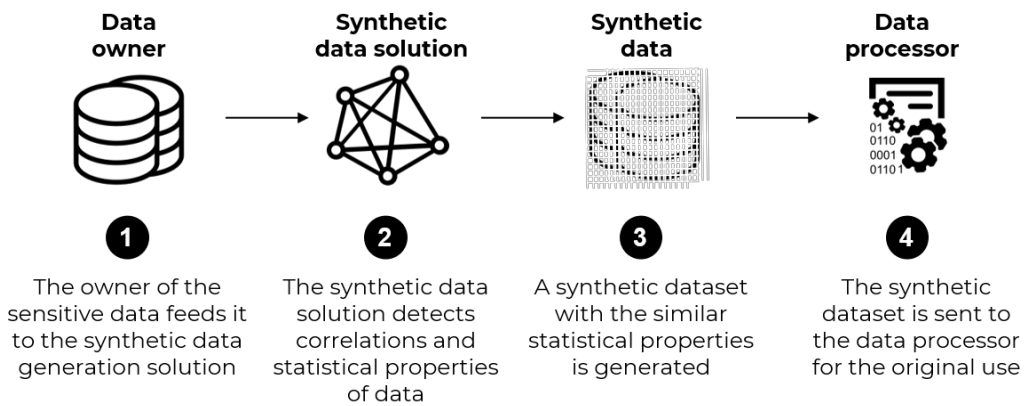
Differential privacy would typically incur only a moderate computational overhead when compared to non-private alternatives. From a utility cost perspective however, employing

[13] It is worth noting the presence of a curator implies some degree of trust in a central authority

differential privacy would often require balancing a three-pronged trade-off between the desired level of privacy; the number of queries that can be allowed to be performed on a given data set (often called the 'privacy budget'); and the desired accuracy of these queries. Designing a DP system for commercial use often requires fixing one of these three variables based on business requirement. As an example a company may need to require a very stringent level of privacy. Given this business imperative "anchoring" the privacy level, the company is then left to decide between the accuracy of the queries and the number of queries the system could support — before starting to run the risk of revealing information about individual records. Thus, the higher the number of queries the system can support, the lower their accuracy must be in order to maintain the desired level of privacy. Indeed, one important limitation of DP is that given an infinite amount of queries, any information could be revealed on the underlying data. Key implications are that either the querier needs to be a trusted entity; or in cases where third, untrusted parties are able to pose queries, the data owner must establish "privacy budgets" that limit the global amount of queries that can ever be run. This limitation poses non-trivial challenges when the number of queries is not known a-priori or simply having a limit is completely impractical for business purposes.

Synthetic Data Generation

Synthetic data is an artificially generated data set that mimics the properties and correlations of an original, confidential dataset.



Synthetic data generation, as the name implies, is a set of methods aimed at producing artificially generated data that mimics real world data. Thus, it is another form of Input Privacy. Indeed, an alternative approach to handling sensitive data, is to artificially create a synthetic dataset. Given that this dataset has the same mathematical and statistical properties of the real world sensitive data — when analyzed it would deliver similar results to those of real

world data. While not explicitly referenced by data privacy regulations, properly created synthetic data is not subject to regulatory restrictions as it can guarantee full anonymization. This is achieved by recreating a completely new (fully synthetic) dataset that does not contain any of the original data or a new dataset where only a subset of the data is replaced with artificial data (partially synthetic).

At its core generating synthetic data implies understanding correlations between the fields of a dataset and recreating a new one that reflects the same correlations. Properties or correlations in the original data set can be selectively included or excluded during generation of synthetic data to offer varying degrees of anonymization. However, to be effective the new data set should maintain the original structure so that traditional analysis tools can be used without any changes. Several techniques can be used to create synthetic data, but machine learning is playing a pivotal role in advancing the field. Generative models are becoming the most common way to produce synthetic data, and use deep learning to autonomously learn the statistical distribution of a dataset, and generate a new, nearly identical one statistically.
The same deep learning techniques are used on video data to produce 'deep fake' videos, but the underlying principles can be applied to other data types. Indeed, the scope of applications for synthetic data goes much beyond PETs. As an example, Neurolabs uses synthetic data to rapidly train computer vision algorithms on demand and at scale. In some cases, generated synthetic data can be more time and cost effective than collecting real world data. Generally speaking when the parameters and the generation environments are set, creating additional synthetic data becomes both fast and cheap.

Synthetic data can be used through the entire data lifecycle - integration, centralization, processing and publishing. For example, the use of artificial data can ease concerns about sharing data across internal departments or across different organizations. It can help circumvent regulations controlling the length of time PII can be stored. By mitigating privacy concerns, synthetic data accelerates the testing and development of machine learning models and eases the burden of publishing insights that would otherwise include sensitive data. More generally, synthetic data accelerates the development of any data-enabled product or service.

At its core generating synthetic data implies understanding correlations between the fields of a dataset and recreating a new one that reflects the same correlations.

Early efforts to generate synthetic audio data can be traced back to the 1930s, and got a boost in the 1970s with software synthesizers. However, attempts to use synthetic data for privacy-preserving statistics were introduced in the early 1990s for a national census. The space advanced rapidly with the advent of Generative Adversarial Networks (GANs) and synthetic data started being used for model training for e.g. self-driving cars, retail applications, simulations, etc.

Utility costs & limitations

Generating synthetic data still requires acquiring an original dataset to model, which in some cases may be difficult. The methods used for generating synthetic datasets (e.g. GANs) may also have some specific limitations, resulting in synthetic data that does not perfectly reflect the original and reduces the accuracy of analyses. Even if the accuracy loss from the limitation of a given technique is acceptable, the quality of the synthetic data generated in this way is still largely dependent on the overall quality of the model used. In fact the use of synthetic data transfers some of the risk associated with the sensitive data to the underlying model used for the generation. Analyses from synthetic data must often be backtested against analyses from the original data to validate accuracy.
In contrast, synthetic data could lead to a leak of information. It is not a trivial task to understand and quantify the risk of data leakage through a reverse engineered model.

Synthetic data can be used through the entire data lifecycle - integration, centralization, processing and publishing.

Finally, generative models are often not able to detect industry specific constraints relating to datasets. The creation of effective artificial data sets requires a domain expert to set appropriate boundaries and limits to be applied to the synthetic dataset. Granting access to a domain expert effectively transfers the risk to this individual. Granting access to the model and its parameters should not be taken lightly.

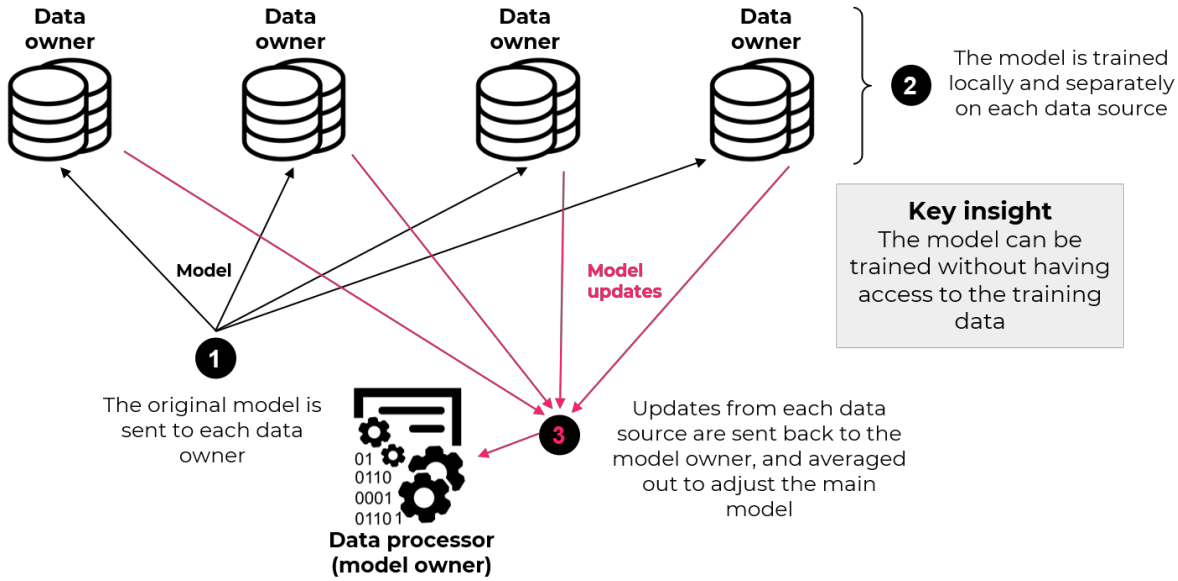
Opportunity

Use of synthetic data for PET purposes started in the 1990s and commercial offerings pushed by a new generation of startups blossomed in the last few years. The ecosystem of tools needed for at-scale use of synthetic data is still emerging. One of the most recent advances is the release of an open source toolkit by MIT in October 2020 (“the Synthetic Data Vault”), aiming to become a one-stop shop for the use of artificial data for privacy-preserving purposes. There’s still a notable lack of standardization and multiple techniques (from simple linear regressions, to more complex decision trees and generative models[14]) are being used in an ad-hoc way to generate synthetic data for specific use-cases. Well documented best practices and approaches are still to emerge.

[14] Comparative Evaluation of Synthetic Data Generation Methods

Federated Machine Learning

Federated learning ships machine learning models to the locations where the data is stored to perform the training locally, decoupling the training process from the need to access, share and store all the data in a centralized location.



Standard machine learning approaches require access (often from multiple sources) and consolidation of training data in a single location. The downside of this architecture is that the data scientist sometimes needs to be granted lawful access to each data source individually - a daunting process, particularly for sensitive data. As an example, consider the challenge of getting consent from thousands of patients registered across hundreds of hospitals to use their personal data to develop a machine learning model for early disease detection.

Federated learning (FL), in contrast, is a group of architectural approaches with the goal of decentralizing machine learning. While there could be various ways to achieve this goal, one prominent approach towards FL would typically ship the machine learning models to the locations (nodes) where the data is stored to perform the training locally. Under this approach the models are trained locally and independently on each distributed dataset producing individual model updates which are sent back to the central server. There they are aggregated into a single consolidated global model through an “averaging process”.

Other alternative approaches to achieve FL would include, as an example, utilizing SMPC to distribute and parallelize the training of a *single* neural network model over multiple data

sources (as opposed to training several independent models and averaging them out as in the previous example). For the sake of simplicity, throughout the rest of the document we will refer mainly to the approach based on "model averaging", which was first presented in 2017 by Google AI[15] and first employed in Google's Android Keyboard (Gboard): when Gboard shows a typing suggestion on your phone, the phone locally stores information about the current context and whether you chose that suggestion. The on-device history is then analyzed locally one the phone. Then a model update is sent to Google's server, where it is averaged out with all the other users' model updates, in order to improve the next iteration of Gboard's prediction model.

This technique decouples the model training process from the need to access, share and store all the data in a centralized location, allowing machine learning algorithms to learn from multiple, decentralized data sets.

By removing the need to share the data sources, FL enables multiple entities to collaborate in the development of models and enables the machine learning provider to remain largely disconnected from the sensitive datasets. FL can also be used in conjunction with other PET techniques. For example SMPC and HE can be used to ensure that the IP behind the ML models sent to the multiple data sources remains secret and inaccessible to the data owners. Thus DP can be applied to the individual datasets to ensure the aggregated model is differentially private.

Enhancing the privacy of the data used for machine learning is just one of the potential goals of FL, and may not be the key driver for adoption. Indeed, FL offers other important advantages, like making the overall training process more scalable in case of numerous datasets: assuming nodes have their own compute and storage resources, FL can have some "auto-scaling" properties where the capacity of the system is linearly extended by the additions of a new node. For the purposes of this document, we decided to focus more on its privacy enhancing properties.

Utility costs & Limitations

Training on multiple datasets poses non-trivial challenges due to potential differences in data collection and storage methods from different nodes. These differences can reduce the accuracy of the system and, when using FL as a PET, are often hard to resolve since data cannot be accessed, 'cleaned' or 'standardized' directly by the data scientist before analysis. Thus, in some cases, FL systems should potentially account for a confidential and remote data preparation process, which can be a non-trivial task. Working in such a federated system, a data scientist would not enjoy a very rapid process of exploring the data and iterating on the models being built.

[15] *Federated Learning: Collaborative Machine Learning without Centralized Training Data"*

FL systems require an efficient and secure communication network that becomes more complex as additional nodes are added. FL systems must accommodate variability and availability in the storage, computational, and communication capabilities of the nodes; and keep working when potentially some of them are unavailable.
Similar to SMPC, dishonest behaviours from network participants are a risk. Federated learning can thus be subject to two forms of "poisoning":

- data poisoning refers to a malicious attempt to tamper with one or more data sources
- model poisoning refers to tampering with the machine learning model used by one or more nodes.

Finally, transmitting model updates in place of raw data enhances but does not fully guarantee confidentiality. A model update sent back by a given node contains *some* information about the node's training data[16]. For this reason, FL is often used in combination with other PET techniques to boost the confidentiality of the overall system.

These cases require a nuanced understanding of the scientific and practical trade-offs between accuracy, efficiency and costs imposed by combining multiple PETs.

Opportunities

FL has become an exciting field of experimentation for corporates and startups alike, thanks to the release of tools like TensorFlow Federated. Current initiatives are aimed at easing some of the recurrent challenges when deploying secure federated learning. These include network challenges like reducing the size of the models shipped to data locations, reducing latency, improving network throughput and making FL systems resilient to intermittent network or node availability. Another substantial area of work is how to build more secure FL systems by integrating other PETs. The state of the art for these types of integrations is in its very early days. Many startups are trying to tackle how to best balance the trade-offs imposed by each technique and explore for which use-cases they are a best fit.

[16] *Quantification of the Leakage in Federated Learning*

Trusted Execution Environment

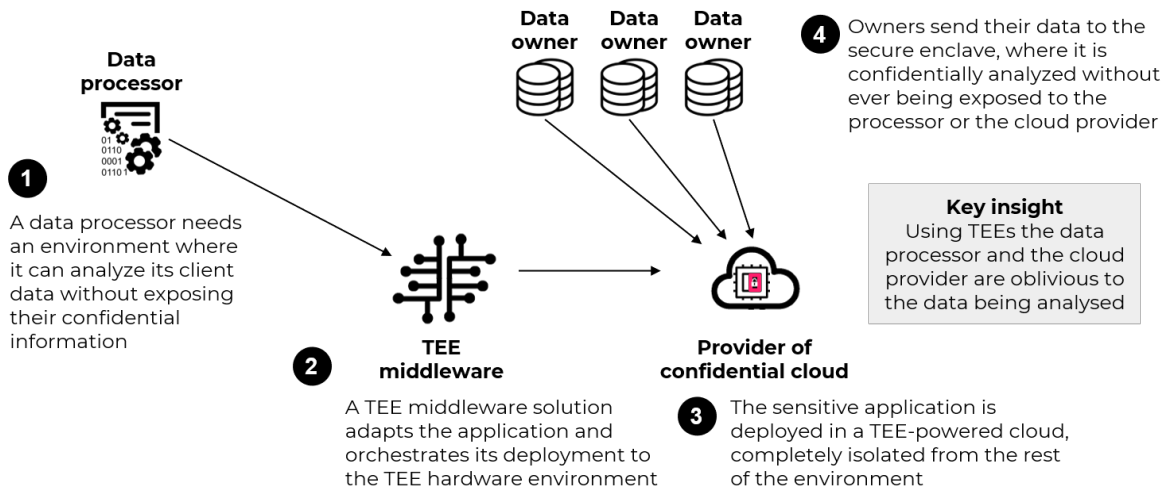
A Trusted Execution Environment is a secure partition of a larger chip/SoC that secures the execution environment of the analysis by isolating it completely from the rest of the machine processes.

A Trusted Execution Environment (TEE) is a secure partition of a larger chip/SoC that is able to support any type of computation. Unlike other approaches covered previously in this report, this is a *hardware first* approach that relies on implementation on chip. Rather than attempting to secure inputs cryptographically, TEEs secure the execution environment of the computation (the "enclave") by isolating it completely from the rest of the machine processes, *including from the operating system*.

A TEE is a secure area of a main processor that safeguards code and data loaded inside of it

Key insight
TEEs completely isolate sensitive data and apps from the rest of the device

Trusted execution environment in the cloud



Other processes relying on the enclave receive an attestation of its integrity and an attestation that the code is being executed correctly. By keeping data encrypted while at rest, and decrypting it only when inside the secure enclave for use, TEEs provides Input Privacy. Depending on the computations being executed, the attestation of integrity of the enclave and the attestation of correct code execution can make TEEs a form of Output Privacy.

Originally developed for mobile applications handling sensitive information, TEEs are making their way into consumer electronics, industrial devices, drones and more — pushed by

the increasing need for these devices to access third-party applications and the internet. TEE hosted computing is also becoming increasingly available among cloud providers.
Compared to other cryptographically-enabled techniques, TEEs scale better and more efficiently. Indeed as the size of the input data increases different TEEs can be linked together to analyze multiple data sources in an integrated fashion. TEEs are flexible enough to be used in light weight mobile applications for payments and identity; password and key management; data rich use-cases like relational databases; enforcing digital rights management in streaming applications (Netflix was an early adopter in 2011); and even server-scale cloud applications.

A Trusted Execution Environment (TEE) is a secure partition of a larger chip/SoC that is able to support any type of computation.

Computing environments for secure and private execution have existed since the 1970s, when Secure Elements (tamper resistant platform for hosting code and sensitive data) found their first application in smart cards and later became mainstream with payment cards. Secure elements are not a fully fledged computing environment but rather purpose-built hardware to host sensitive data (like keys) and very specific applications of limited but critical scope. Trusted Execution Environments overcome the computational and memory constraints of Secure Elements and are able to support any type of computation.

In 2008, the Open Mobile Terminal Platform first defined TEE as a set of hardware and software components able to support applications while meeting one of two security levels:

- resistance against software-only attacks or
- resistance to both software and hardware attacks.

Originally developed for mobile applications handling sensitive information, TEEs are making their way into consumer electronics, industrial devices, drones and more

Since then TEEs have become an increasingly important component of the mobile ecosystem, and are now expanding to cloud applications for enterprise use. In 2020, Global Platform became the main body in pushing forward standardization of the technology.

Utility costs & limitations

The computational costs of using TEEs is substantial (ranging from 20% increases for small applications to 700-800% in extreme cases) but arguably lower compared to software-only techniques like HE and SMPC. The use of TEEs does not carry any loss of information nor limit the types of computations capable of being performed and consequently offers a very good level of performance relative to analysis in the clear.

Adopting TEEs requires the use of special hardware that needs to be integrated into the device during manufacturing. While these hardware modules are

present in most mobile devices, their presence cannot be assumed for other types of devices. Engineering resources are often needed to update software applications to match the specific software frameworks of TEEs, with SDKs offered by hardware OEMs and third party vendors. In most cases the use of TEEs demands some degree of collaboration between TEE providers and application developers to make sure that a specific level of security is met.

Despite offering meaningful improvements to the state of security in mobile devices and even cloud applications, from a theoretical perspective TEEs do not provide a mathematical or cryptographic guarantee of confidentiality. TEEs have been subject to some side channel attacks[17] (like Replay, TOCTOU, and Foreshadow) exploiting memory caches, speculative execution, and more — prompting security experts to argue over the real degree of isolation and security offered by TEEs.

Despite offering improvements to the security of mobile devices and cloud applications, from a theoretical perspective TEEs do not provide a mathematical or cryptographic guarantee of confidentiality.

Opportunity

TEEs are commercially available and widely deployed in mobile devices. Their adoption in other devices, architectures, and use-cases is expanding. Intel, ARM, AMD, IBM and RISC-V are key players in the TEE space with hardware and software offerings that have reached solid commercial status. TEEs underwent intense standardization efforts which vastly helped their compatibility with an array of applications. However, more standardization is needed since different hardware and software implementations still exist (e.g. the process-based model by Intel's SGX and the VM-based model by AMD). Applications running in a TEE must be developed specifically for different hardware technologies, leading to cumbersome and inefficient development. These shortcomings are proving to be fertile ground for innovative startups offering a better development experience with a wealth of software development kits that streamline the adaptation of applications for specific platforms. Several startups are focusing on building infrastructures that efficiently deploy secure applications at scale. This is done through middleware layers mediating cloud providers and applications providers who are looking for more secure execution.

[17] *A Survey of Published Attacks on Intel SGX*

Section 3: the emerging PET market

The PET Market

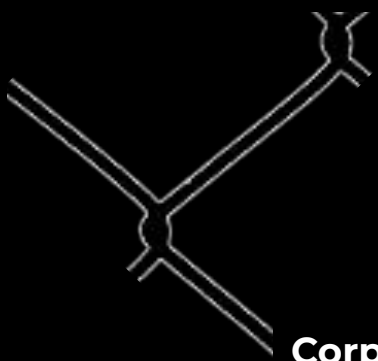
Though in its very early stages and still behind in the hype cycle, the PET market is seeing a strong inflection point. A new generation of startups is flourishing, while PET is increasingly becoming part of the tech giants' strategic agenda.

Startup Landscape

The PET startup landscape is expanding rapidly and currently skewed towards building enabling infrastructure. Further verticalization still requires scientific and engineering advancements in the underlying technologies.

Early PET applications

The scope and breadth of PET applications is staggering and reflects the paradigm shifting nature of the underlying technologies.



Corporate Initiatives

Tech corporates have been early adopters for their internal tools and developer evangelists for their PET open source frameworks. The focus will soon shift to building PET-powered commercial applications.

Section 3: The Emerging PET Market

Though still in its very early stages and behind in the hype cycle, the PET market is seeing a strong inflection point. A new generation of startups is flourishing, while PET is increasingly part of the tech giants' strategic agenda.

Now that we have a general understanding of what PETs are and what they are good for, we are ready to take a closer look at how this nascent market is shaping up. PETs are rapidly making it into the agenda of founders and giant corporates alike with a plethora of initiatives starting over the last few years. In this section we will cover three topics: we discuss the PET startup landscape and its most notable actors; we provide a quick overview of the most interesting PET-powered use-cases across industries and we provide a review of recent corporate initiatives in the space.

Startup Landscape

The PET startup landscape is expanding rapidly and currently skewed towards building enabling infrastructures. Further verticalization still requires scientific and engineering advancements in the underlying technologies.

At Lunar Ventures, PET is one of the technology trends that we are most excited about. Since launching the fund, we have worked to get to know as many startups in this field as possible. We deeply believe that PET will be a key building block of tomorrow's technology infrastructure, and that winning companies will see outsized adoption in the market. Venture capitalists have poured more than $850m into the PET space, demonstrating it is a fertile and promising ground for startups. Indeed, there has been an explosion in the number of new companies popping up with PET-powered applications or with solutions tackling key bottlenecks in the development, management and deployment of confidential technologies.

The PET space can be very intimidating at first glance. It is a tangled web of very complex disciplines like cryptography, computer science and hard core engineering — in an early stage of development. Each technique has its own pros and cons, key use-cases and limitations. It is not rare that multiple techniques are used in combination, with the aim to finetune a killer solution for a specific problem or use case. In order to help founders and investors approaching the space, we include an overview of all the companies that we are aware of that are making use of PETs. We don't assume this is comprehensive and are happy to expand this list based on suggestions from our readers. (If you are aware of a company in this space which we didn't include please reach out!)

Lunar PET Landscape

LUNAR VENTURES

ENABLING INFRASTRUCTURE

SOFTWARE-BASED

PII De-identification

Data Anonymization & DP

DATAGUISE · aircloak · LEAPYEAR · HACERA · nymiz · PRIVITAR · ANONOS · MMUTA · KIPROTECT · TUMULT LABS

Synthetic Data

hazy · gretel · YDATA · Statice · MOSTLY·AI · Facteus · SYNTHO · TRUATA. · SYNTHESIZED · TONIC

Private Computation

Software based

Tr'pleBl'nd · SHIELD · Adap · Partisia · apheris AI · inpher · Duality · 華控清交 HUA KONG TSINGJIAO · PRIVECOMMS · ZAMA · DATAFLEETS · SHIELDIO · ENVEIL · Oblivious AI · nth party · SHOJI · PlatON · appliedblockchain · prenode · ARPA · fetch.ai · XAIN. · qedit · LIGERO · sherpa.ai · Ntropy network. · MagicCube · PRIVATE STORAGE · SCALEOUT · S20.AI

Hybrid (software & hardware)

cosmian · sharemind · OASIS LABS · enigma · secretarium

HARDWARE-ENABLED

Hardware based

Fortanix · EDGELESS SYSTEMS · decentriq · SCONE · MADANA · CYSEC · anjuna · GRAPHENE · TrustKernel · evervault

Providers of Confidential Cloud

Google Cloud · Microsoft Azure · Alibaba Cloud · IBM Cloud · aws

PET Hardware

IBM · AMD · RISC-V · arm · intel

SPECIFIC USE-CASES

Identity

PEER MOUNTAIN · nuggets · KRIPTAN · NuID · civic · sovrin identity for all · sedicii

User Data

Glimpse Protocol · Canopy · INSIGHTS NETWORK · MIMIRIUM · INFOSUM

Key Management

UNBOUND · KMS CHAIN · KEYLESS · SEPIOR · SPHERITY · MIRACL

Medical

InfoShield · Xenya.ai · Lynx.MD · OWKIN

Cryptocurrency & Ledgers

Enduser Facing

ZenGo · Zcash · MONERO · curv · ZCOIN

Ledger Infrastructure

BOLT LABS · NuCypher · Qredo · LOOPRING · MATTER LABS · CYBERIAN · SHARDX · THRESHOLD · ATOMRIGS LAB · keyTango · CYBAVO · AZTEC · STARKWARE · Fireblocks

Did we forget someone or do you have any feedback?

Feel free to reach out to

Alberto Cresto

In preparing this market map we aimed to find an optimal trade-off between presenting something simple yet effective and to provide visual dimensions to categorize companies and strategies into buckets.
The more fundamental distinctions are:

The implementation that is being used:

1. Companies that rely on "Software PETs" (originating from computer science, statistics or cryptography)
2. Companies using PETs that are "Hardware Based" (relying at least partly on hardware).

The scope of the problems that companies are trying to solve:

1. Companies leveraging PETs to tackle **specific use-cases.** These can be addressing specific industries (e.g. healthcare) or more industry-agnostic use-cases (for example, identity management, key management, blockchain scalability, etc).
2. Companies building "**enabling infrastructure**" to streamline how PETs are developed, deployed and set up — while remaining fairly agnostic to the underlying use-cases.

When considering "hardware-based" vs "software based" PETs, it is important to note that companies leveraging hardware-based PETs use a very different technology stack compared to their software-based peers, and are part of a more mature ecosystem that is largely driven by tech corporates from the cloud ("Confidential Cloud") and semiconductor space. Additionally, the decision to adopt software PETs or hardware PETs has several implications across technical, business and commercial considerations.

There are a few important caveats to note in the "specific use-cases" vs "enabling infrastructure" distinction. Firstly, this framework is a categorization tool rather than a perfect science. PET is a nascent technology and distinguishing between these two groups can be at times hard. We used our best judgement in grouping companies but the market map should still be used with a grain salt. You will notice the landscape is heavily skewed towards **enabling infrastructures.** There are a number of reasons for this. Firstly, most of the underlying techniques left academia recently and have not yet become well scalable and understood technologies. There's plenty still to be built at a fundamental level to make these technologies easier (or sometimes even viable) to use and deploy. To some extent we expect any companies in this space to be contributing to enabling infrastructure. We believe there are advancements in the underlying technology that need to happen before further and clearer "verticalization" can occur and we expect this market map to show much neater distinctions in a few years.

PET will be a key building block of tomorrow's technology infrastructure and winning companies will see outsized adoption in the market

Secondly, a company may begin by focusing on the horizontal technology improvement but may ultimately

mature into a verticalized solution. Many of the companies we categorized as enabling infrastructure position themselves as addressing a multitude of use-cases. This is likely due to the fairly "horizontal nature" of PETs and the fact that the need for confidentiality applies equally to a large number of sectors. Indeed, we expect the decision to go vertical vs horizontal to depend on a number of factors, including the background of the founding team and the nature of their network, the perceived ease of the go-to-market strategy, traction with lighthouse customers, etc.
It is also possible that some early stage start-ups have not yet found an evident product market fit, so they are not yet declaring a specific vertical and continue experimenting with customers across sectors. We are in the very early days of PETs.

Consequently, we limited the "**specific use-cases**" bucket to companies that have a very clear positioning in the market, and whose scope is very focused on solving a specific problem. However, we expect a number of the horizontal companies to be in the process of verticalizing themselves around their first lighthouse customers and assume that some verticalized companies may expand into other verticals over time and may become horizontal technology providers.

Since most of the groupings we have identified in the "specific use-case" category are well covered elsewhere (for example, "identity", "key management", and "cryptocurrency and ledgers" were among the very first adopters of PETs) we preferred to focus the rest of our analysis on companies building enabling infrastructures.

Venture capitalists have poured more than $850m into the PET space, demonstrating it is a fertile and promising ground for startups

We are very excited at the idea that PETs are on the verge of becoming pervasive across sectors. We feel that most of the considerations we make discussing enabling infrastructure are to an extent relevant to the PET space as a whole. We summarized our perspective in the table below.

Lunar PET Landscape Cheatsheet (1/2)

	PRIVATE COMPUTATION (CRYPTOGRAPHY)		PIIs DE-IDENTIFICATION	
	SOFTWARE BASED	**HARDWARE BASED (TEEs MIDDLEWARE)**	**SYNTHETIC DATA**	**DIFFERENTIAL PRIVACY**
What is the value proposition of these companies?	• Allow mutually-distrusting parties to confidentially collaborate on data analysis or to analyze data while keeping it secret		• Remove Personally Identifiable Information from data analyses • Create a "surrogate" data set that has the statistical/overall characteristics of the original, but does not contain sensitive information	• Perform aggregate queries on a dataset, while ensuring the queries' results do not expose any PIIs
Techniques	• **HE**: ensures that both the data and the result of the analysis remains secret, removing the need to trust the location where the analysis takes place. • **SMPC**: enables multiple mutually-distrusting parties to collaborate on a joint analysis on confidential data, preventing any participant from learning anything about the inputs provided by the other parties. • **ZKP**: allows data provided by one party to remain secret while being verified by another party.	• **Trusted Execution Environments**: a secure partition of a larger chip/SoC that secures the execution environment of the analysis by isolating it completely from the rest of the machine processes.	• **Synthetic Data**: multiple data generation techniques which create an artificial data set mimicking the properties and correlations of an original, confidential dataset	• **Differential Privacy**: its main goal is to protect the privacy of any individual providing his information to a database that is used for aggregate analysis.
What enabled this?	• Advances in computer Science and cryptography	• Advances in hardware hardening and virtualization	• Various advances (depending on the technique used)	• Advances in machine learning and statistics
How do they look?	• A cryptographic protocol telling participants what computations to perform, what information to encrypt and how, and where to send it	• Software component that orchestrates a hardware enclave (often provisioned by a cloud provider)	• Software generating novel datasets to be used in place of the original	• Software sitting between the data analyst and the sensitive dataset, offering an anonymized view by adding noise to the "true" results of the queries
What computational efficiency/ performance losses do we incur?	• Large efficiency hit (often 1000x-10,000x or more) but rapidly improving with R&D progress	• Moderate performance hits (depending on the type of computation, CPU or GPU architectures, etc)	• No performance hit	• No performance hit

Lunar PET Landscape Cheatsheet (2/2)

	PRIVATE COMPUTATION (CRYPTOGRAPHY)		PIIs DE-IDENTIFICATION	
	SOFTWARE BASED	**HARDWARE BASED (TEEs MIDDLEWARE)**	**SYNTHETIC DATA**	**DIFFERENTIAL PRIVACY**
Impact on accuracy or validity of results	• Perfect accuracy/validity		• Validity of results depends on the quality of the software and the expertise of the data scientist in charge of producing the synthetic data	• Validity and accuracy are high if correctly used. However, it is often not possible to apply a full analysis due to restrictions like a "privacy budget"
Difficulty of integration and workflow impact	• Often hard to integrate into existing systems, requires assistance of security experts	• Requires the adaptation and deployment of applications to a new environment • As of 2021, there are no technological barriers to integrating this	• Does not require extensive integration since it generates a new data set that can be used instead of the original one • May create a convoluted workflow, where the data scientist creates hypotheses on the synthetic data, then verified by a 3rd party on the original data	• Important impact on workflow, as the DP software often ends being the UX of the data analyst / the interface to the data being analyzed • Made people quite unhappy — this system does not jibe well with the way companies work with data
Required expertise from the developer or data scientist	• Requires deep and nuanced understanding of security and cryptography (except in some uniquely easy-to-integrate cases, e.g. Zama)	• Does not require particular expertise, assuming reasonably strong middleware (e.g. Anjuna)	• The data scientist who anonymizes the data needs expertise to create synthetic data with characteristics consistent to the original dataset, and/or that properly hides sensitive information • The data scientist who consumes the anonymized data does not need particular expertise	• Does not require particular expertise from the data scientist
Commercial maturity	• Early	• Medium	• Mature	• Medium
Confidentiality assurance	• Confidentiality is mathematically guaranteed: as long as integration was done well (which is challenging), privacy is entirely preserved	• Confidentiality is based on hardware trustworthiness • Proven to be vulnerable to multiple types of side-channel attacks • With these caveats in mind, TEEs are secure as long as integration was done well	• Confidentiality is based on the skills of the data scientists and the technique used • Lack of standardization and best practices prevents predictable assurances of confidentiality	• Confidentiality is guaranteed — as long as the system parameters were chosen well • However, there is an inherent tradeoff between: the desired level of privacy, the number of queries that the system can perform and their accuracy.

There are a few noteworthy start-ups that have been in the space for a few years and have drawn the mind space (and wallet space) of reputable VCs and angel investors:

Immuta

Immuta offers a no-code platform for data governance, control and policy creation with some anonymization features based on differential privacy. Founded in 2014, they have since raised $70m from Sequoia Seed Investments, Intel Capital, DFJ Growth and many others.

Fortanix

Fortanix has built a platform to manage and orchestrate the deployment of applications on Intel SGX's TEE. The solution can be used to enable runtime encryption of keys, data, and applications. Founded in 2016, they have raised $28m from Foundation Capital, Intel Capital, Neotribe Ventures and Inspovation Ventures.

Enveil

Enveil, leveraging HE, enables to securely search, cross-match, and derive insights from third-party data sources without revealing the contents of the search or analytic — or compromising the security or ownership of the underlying data. The platform is composed of a client application and a server application to be deployed where the data resides. Founded in 2016, they have raised $13m from IN-Q-TEL, MasterCard, Capital One and others.

Zcash

Zcash is a crypto currency which leverages ZKP to allow anonymous transactions: transactions are verified without revealing the sender, receiver or transaction amount. Selective disclosure features within Zcash allow a user to share some transaction details, for purposes of compliance or audit. Founded in 2015 and backed by Fred Ehrsam, Winklevoss Capital, Boost VC and others. They now have a market capitalization of $1.3B (Jan 2021).

Zama

Zama enables companies to process their customer's data in encrypted form. Zama's technology is based on a breakthrough homomorphic encryption scheme, which for the first time makes homomorphic deep learning practical without limitations on the type of networks used and with acceptable latency. Founded in 2019, Zama has raised a pre-seed round from Lunar Ventures, Charlie Songhurst and Brent Hoberman.

Owkin

By leveraging FL, Owkin enables parallel running of machine learning models on clinical data kept within the secure premises of multiple hospitals. Founded in 2016, they have raised €68m from Brent Hoberman, Eight Roads Ventures, Google Ventures and others.

Some Early PET Applications across Sectors

The scope and breadth of PET applications is staggering and reflects the "enabling infrastructure" nature of its technologies.

PETs are complex technologies originating from hard-core research done by computer scientists and cryptographers — and they can appear a bit abstract. We thought our audience could benefit from a 360° overview of how and where PETs are used in real life. In this section we look at how these technologies are used across a variety of sectors. Each sector's use-cases are reviewed in a brief and non-comprehensive way, as we preferred to keep a bird-eye view on the space and its emerging opportunities rather than zooming too much into a single use-case. In particular we don't give a full walkthrough of how to deploy one of these technologies inside a given sector. On the internet there's a number of more detailed use-cases.

The interested reader can look at:

- this blogpost for a thorough presentation of how to use SMPC on multiple healthcare data sources to identify and prevent heart failure;
- this blogpost presenting a zero-knowledge re-adaptation of the the "Millionaire's Problem";
- this whitepaper for a more detailed overview of Homomorphic Encryption applications;
- this presentation showing in detail how to build a Federated Learning system and how to apply it to a healthcare use-case.

For those looking for a more succinct overview instead, let's proceed with a quick scan of use-case across industries.

Banking

Financial institutions are exploring the use of SMPC and HE to improve credit scoring frameworks. Credit scoring frameworks require access to data from multiple, independent lending institutions like banks, credit card companies and merchants that offer credit to their customers. PET techniques allow data to be shared and analyses to be performed across different lenders without fully exposing the underlying sensitive data. Financial institutions can use ZKP to combine KYC and customer registration programs into a common onboarding utility, reducing duplicate efforts across partner institutions and limiting data sharing to the minimum required by regulations.

Consumer Technology

Infringements of customers' privacy are commonplace, particularly in digital advertisements and in the personalization of digital services. ZKP could be used to locally match ads with relevant consumers removing the need to push all of our sensitive data to dubious middlemen; and HE could be used to qualify users without exposing personal information. Similarly, PETs can be used to improve customer marketing models by connecting internal user data from different geographies while staying

compliant with differing privacy regulations like Europe's GDPR or California's CCPA. A few large technology companies are already major users of DP to gain insight into aggregate actions of its users while preserving their individual privacy. News publishers and other companies aiming to personalize their services could do so leveraging DP-powered personalized recommendation engines. Phone manufacturers are long-standing users of secure enclaves to prevent tampering on both a hardware and software level: these technologies are used to protect 3rd-party apps that process sensitive information (identification, e-commerce payments and more).

Cryptocurrency

Despite wide-spread beliefs, Bitcoin and other traditional cryptocurrency are only pseudonymous, not anonymous since all transactions are exposed on a public ledger. Anyone, like the SEC or law enforcement agencies or a merchant paid from a particular address, that could match a blockchain address to an individual could see her entire transaction history. Blockchain-powered currencies like Zcash can utilize ZKPs and other PETs to protect user information rather than storing data in the clear. Digital wallets and other authentication solutions could leverage ZKPs and SMPC to enhance privacy of users accessing their services by obfuscating real world identities.

Healthcare

Synthetic data could be used to bootstrap and speed-up the research and development efforts of new services without access to patient data. Governments could utilize DP to enhance the privacy of contact tracing applications for pandemics: these algorithms allow governments to uncover statistical information about the population while ensuring that the underlying information can never be identified through reverse engineering. Healthcare providers can leverage HE to establish genetic data exchanges that protect the underlying genomic data against cyber threats or to create novel machine learning algorithms that learn without seeing the underlying medical records. SMPC and FL could be used to coordinate and run large scale analyses in parallel data held by multiple hospitals without requiring lengthy permissions to transfer data off-premises or centralize that private data.

Industrial

The industrial sector has been a laggard in technology adoption but with the rise of Industry 4.0, it has started to produce large amounts of data. Unfortunately, most of it is still locked in silos due to competitive concerns. OEMs could use FL to enable predictive maintenance of IoT machinery for customers in the manufacturing, oil refining or automotive industries while respecting their respective need for confidentiality. Companies with large supply chains like Boeing or Airbus could also use synthetic data, HE and FL for the joint and collaborative development of novel products with other companies like engine manufacturers, Rolls Royce and Pratt and Whitney. Defense contractors and armed forces could use SMPC and HE to monitor the performance of their fleets without exposing sensitive information such as usage and location. Satellite companies could utilize SMPC and TEEs to prevent outer-space accidents by predicting the occurrence of satellite collisions without exposing the underlying locations.

Analytics and Cloud services

The range of cloud services that could be improved by adopting PETs is very large, and we include only a few examples. Generally speaking these services run in outsourced computing environments (i.e. cloud) where customers send their data. ML providers could use HE to convert traditional machine learning models into end-to-end encrypted equivalents, maximizing security and confidentiality of both their IP and their customers' data. Similarly TEEs and HE could enable a fully private cloud, where the customer data is inaccessible to everyone even while in use.

Insurance

PETs have the power to increase access to sensitive data that could be used to more efficiently price policies, resulting in lower prices and better services. Insurance providers can utilize FL, DP, and ZKP to detect insurance fraud through collaborative data sharing across their peers, brokers and reinsurers while respecting the privacy of the individual customers. Similarly, they could leverage SMPC and FL to enhance their underwriting models, by gathering important information about their customers while fully respecting their confidentiality.

Security

Response to cyber threats have long lacked wide-spread collaboration. Cyber security providers could utilize SMPC, HE, TEEs, and ZKP to engage in collaborative detection of cyber threats and mitigate the spread of untargeted attacks like ransomware by sharing information. These companies can perform search operations across multiple parties and networks to detect the occurrence of suspicious IT patterns whilst protecting the privacy of the concerned entities. At a more general level, PETs wildly improve the security postures of corporates and governments favouring less centralized data repositories, end-to-end encryption and reducing the need to access, store and secure sensitive datasets.

Trading

SMPC, HE, and ZKPs allow traders to privately compute on their clients' assets held within and outside their institutions without exposing the underlying information, enhancing their margin calculation models. More, traders can utilize SMPC and HE to execute trades without revealing their positions. These algorithms allow clients to find trading partners without exposing their orders before execution, hedging their risks to unfavourable pricing and market inefficiencies.

Corporate Initiatives

Tech corporates have been early adopters for their internal tools and developer evangelists for their PET open source frameworks. The focus will soon shift to building PET-powered commercial applications.

While startups and academia are playing a key role in moving the sector forward, tech giants — with their data enabled business models — are obvious early contributors and adopters. For them, PETs can improve access to more data and also consist in an opportunity to reposition their own value proposition around privacy and confidentiality. Indeed, they have already started building early product offerings built on both in-house development and strategic acquisitions. With Gartner predicting that half of all large companies will adopt privacy-enhancing computation by 2025, let's look at what's already happening in the market.

Apple

Apple has put security and privacy of its users at the center of its value proposition. It is not surprising that it was an early adopter in using DP to enhance the experience of its users while balancing between insight-generation and privacy-protection. Apple uses this framework across multiple use cases such as, from optimization of Safari's memory usage, discovering most recurrent energy consumption patterns, to identifying local slang words and trending emojis. Apple also developed and uses proprietary "Secure Enclave" processors to host TEEs on its iPhones, iPads, Macs, TVs, Watches, and HomePods. These processors are isolated from the main processor to provide military-grade security in handling cryptographic keys as well as running sensitive applications. Apple also intensively researches FL in an attempt to bring it under its portfolio of privacy-enhancing techniques. FL is currently used in Siri to enhance speaker recognition models through localized data.

Google

Google has been one of the first large companies trialing and adopting PETs techniques, thanks to the breadth of in-house products and services available for experimentation. Google has implemented Federated Learning (of which it was an early evangelist already in 2017) in Gboard — the default keyboard on most Android smartphones. Gboard continuously learns from contextual interactions with its auto-complete features and processes the on-device history to optimize its query-suggestion model.

Despite growing use of the technology in its own products, Google's B2B offering is still at a very early stage and focused on developer tools. Google has released "TensorFlow Federated" and "TensorFlow Privacy" as open-source frameworks to test their models using FL / DP techniques and launched "Private Join and Compute", an open-source framework using SMPC / HE to enable the extraction of valuable insights. The most advanced tools are related to TEE:

- "Asylo", an open-source framework abstracting the back-end complexities related to the development and management of applications;
- "Confidential Computing", a confidential VM solution that enables real-time encryption of data during computation;
- "Trusty" is a secure operating system that employs TEEs on Android, allowing developers to deploy sensitive applications with ease.

Facebook

Under continuous criticism for privacy violations, Facebook is actively experimenting with PETs. Facebook assisted independent academic research leveraging DP to help academia analyse aggregated groups without identifying the underlying users. The company has open-sourced algorithms for private set intersection using SMPC and HE. Facebook has also released the "CrypTen" and "Opacus" open-source frameworks for developers to train machine learning models using SMPC and DP. The company has also recently released a paper outlining a novel approach for a more fair allocation of resources federate learning.

Microsoft

Microsoft Research actively engages in state-of-the-art research on PETs and was one of the first cloud providers to offer the technology in their product suite. "Azure Confidential Computing"is a cloud-based VM solution for developers to use TEEs. They have attracted partnerships with start-ups like Signal, Fireblocks, Anjuna, and Fortanix. Microsoft has also released "SEAL", a set of open-source HE libraries and "WhiteNoise", a Python package to apply DP techniques to ML models on Azure.

IBM

IBM is a leader in the TEE space, offering "IBM Secure Execution" and "IBM Secure Service Container". RedHat launched "Enarx", an open-source project that simplifies use of TEEs by easing the need to entirely rewrite applications. Earlier this year, IBM released an HE Toolkit aimed at MacOS, iOS, and Linux developers. They also released "IBM Federated Learning" and "IBM Differential Privacy Library" - Python frameworks that empower developers and researchers alike to experiment with PETs. In the academic realm, IBM has partnered with Columbia University to jointly manage a center that researches PETs.

Tech Consulting Firms

EY has released "Ops Chain", an open-source protocol to use ZKP for secure and private transactions on public blockchains. Deloitte has partnered with Qedit to leverage ZKP on "Eduscrypt", its in-house blockchain platform that lets its clients track and validate employee qualifications on a notarized blockchain without sharing sensitive information. Accenture built a demo version of a dark pool to enable privacy-preserving trade-matching using SMPC.

Bosch

A notable example of early adopters from the industrial sector, Bosch has started "Trustworthy Computing" an initiative to incorporate PETs such as SMPC and HE in their solutions. Bosch Research also collaborates with startups such as Edgeless Systems and Airclock within the "Open Bosch" program to ensure data sovereignty for their partners

Conclusion and Further Readings

Our goal with this document was to help our readers understand the various privacy enhancing technologies, the problems they solve and the massive opportunities they hold.

PETs are a new wave of cryptography that acts on the longstanding trade-off between keeping information secure and keeping information fungible and useful. Thus, PETs enable businesses to extract more value out of their data while respecting the privacy of stakeholders involved. Similar to previous cryptographic technologies, we expect PETs to play a pivotal role in enabling new business models and creating new industries. We believe PETs will make a portion of existing but currently unused data available for analysis, and to incentivize the generation of more data — with a sizable impact on the overall data economy.

PETs are a promising field because of their unique capabilities — but also as their infancy makes this a blue ocean for founders and investors with the right skills to capitalize on it. The wide ranging potential benefits also come with material monetary, efficiency and accuracy costs. Both costs and benefits of PET adoption should be closely considered when exploring their use in real world applications.

The gamut of techniques and approaches that play a role in the PET space is wide. Thus we preferred to focus on a few main techniques, namely Homomorphic Encryption, Secure Multi-Party Computation, Zero-Knowledge Proofs, Differential Privacy, Synthetic Data Generation, Federated Machine Learning, and Trusted Execution Environments. For these techniques we presented a brief summary of their scientific roots and inner workings, what uses they are more suited for, what are their limitations and what opportunities are available to push the field forward.

The document also looked at the state of the PET market. Thanks to their horizontal capabilities PETs have several early use-cases across a large set of industries. Both startups and established corporates are playing an active role in the space. We included a detailed landscape analysis of all the startups we came across that are active in the PET ecosystem. We see a widespread focus in the market on building enabling infrastructure (compared to building focused applications addressed towards a specific use-case or industry). We believe this reflects both the horizontal nature of PETs capabilities and the early state of the space, where much still needs to be built to make further use-cases economically profitable and technically viable. Finally we reviewed PET-related initiatives from a number of large corporates. While most of the efforts are centered around increasing the privacy of their respective cloud offerings with hardware PETs, big tech has also released a notable number of open source tools and frameworks to streamline use and adoption of software-based PETs.

Admittedly, there's much more to be said and discussed about PETs. For example we did not cover their expected adoption curve, where we see the most exciting opportunities, the drivers and constraints that will shape the growth, and the overall outlook of the space. We really love talking about this topic and, in fact, we are not done yet!

This document will be followed by a paper that looks at the future of PETs. The paper titled: "Collaborative Computing: Making data sharing cheaper, faster and easier with partnership-enhancing technologies" explores the trends shaping the development of PETs. It predicts collaborative computing to be the largest new technology market to develop in the 2020s. By 2030, data marketplaces enabled by PETs, in which individuals, corporates, machines and Governments trade data securely, will be the second largest ICT market after the Cloud.

Resources

- Protecting privacy in practice
- The Challenge of Training Artificial Intelligence in the Age of Privacy | OpenMind
- Federated Learning: Collaborative Machine Learning without Centralized Training Data
- How Does Ocean Compute-to-Data Relate to Other Privacy-Preserving Approaches?
- A Beginners Guide to Federated Learning | by Dr. Santanu Bhattacharya | HackerNoon.com
- Tools for Organizations in a Rapidly Evolving Data Privacy Landscape
- Privacy-Preserving Machine Learning 2019: A Year in Review
- Privacy-preserving data analysis
- Historical insight into the development of Mobile TEEs
- Trusted execution environment - Wikipedia
- Protecting privacy in practice
- UN Handbook on Privacy-Preserving Computation Techniques
- What are the differences between HSM and SE?
- Trusted Execution Environment (TEE) 101: A Primer
- Foreshadow, SGX & the Failure of Trusted Execution
- Current Trusted Execution Environment landscape
- Can I Use Intel's SGX for Secure Computation in the Cloud Yet?
- Federated Learning: Collaborative Machine Learning without Centralized Training Data
- Asylo
- Titan M makes Pixel 3 our most secure phone yet (excluded)
- Google Cloud Confidential Computing
- Learning Statistics with Privacy, aided by the Flip of a Coin (excluded)
- Google Online Security Blog: Helping organizations do more without collecting more data
- Enabling developers and organizations to use differential privacy
- Introducing TensorFlow Privacy: Learning with Differential Privacy for Training Data
- Introducing TensorFlow Federated — The TensorFlow Blog
- Android's Trusty TEE
- Differential Privacy Overview
- Protection Against Reconstruction and Its Applications in Private Federated Learning
- Apple's Secure Enclave
- https://www.technologyreview.com/2019/12/11/131629/apple-ai-personalizes-siri-federated-learning/
- Demystifying the Secure Enclave Processor
- IBM Secure Execution for Linux
- IBM Releases Fully Homomorphic Encryption Toolkit for MacOS and iOS
- https://enarx.dev/
- Columbia University and IBM Establish New Center to Accelerate Innovation in Blockchain and Data Transparency
- z Systems Secure Service Container User's Guide
- Private matching for compute enabling compute on private set intersections
- New privacy-protected Facebook data for independent research on social media's impact on democracy
- Facebook - Introducing Opacus: A high-speed library for training PyTorch models with differential privacy
- CrypTen: A new research tool for secure machine learning with PyTorch
- EY releases zero-knowledge proof blockchain transaction technology to the public domain to advance blockchain privacy standards
- Deloitte Leverages Zero-Knowledge Proof On New Eduscrypt Platform
- Powering larger insights while preserving privacy
- Differential privacy in machine learning (preview) - Azure Machine Learning
- Azure Confidential Computing – Protect Data-In-Use
- Microsoft SEAL: Fast and Easy-to-Use Homomorphic Encryption Library
- Trustworthy computing – data sovereignty while connected
- https://riscv.org/
- Intel® Software Guard Extensions (Intel® SGX)
- Trusted Execution Technology (Intel® TXT) Overview
- Arm TrustZone technology
- AMD PRO Security
- Federated Learning: Challenges, Methods, and Future Directions – Machine Learning Blog | ML@CMU | Carnegie Mellon University
- Federated Learning: Collaborative Machine Learning without Centralized Training Data
- What is Federated Learning?. The field of machine learning is... | by ODSC - Open Data Science
- Introduction to Federated Learning and Challenges | by Kelvin
- What Is Federated Learning?
- Zero-Knowledge Proofs, Explained
- Zero Knowledge Proof: how to maintain privacy in a data-based world
- Zero Knowledge Proofs: An illustrated primer – A Few Thoughts on Cryptographic Engineering
- Introduction to privacy-preserving synthetic data | by Elise Devaux | Statice
- How generating synthetic data can protect organizations
- Up Next What Is Synthetic Data?
- Synthetic data
- 10 Use Cases for Privacy-Preserving Synthetic Data
- Privacy and Synthetic Datasets | Stanford Law School
- The real promise of synthetic data
- Global vs Local Differential Privacy | by shaistha fathima

LUNAR VENTURES