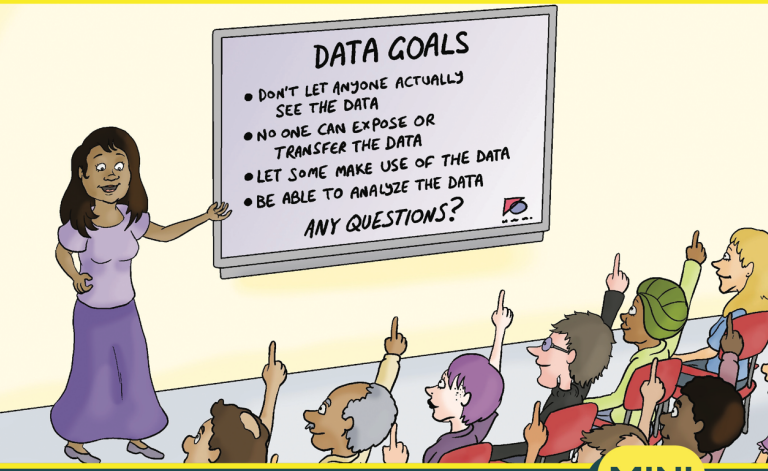
ConversationalGeek

Conversational Privacy Enhancing Technologies & Big Data

Mike Cobb (CISSP-ISSAP)



MINI Edition

Learn about:

- How privacy enhancing technologies (PETs) have evolved to revolutionize data sharing
- How to get started with PETs and incorporate them into you IT processes

Sponsored by

Duality

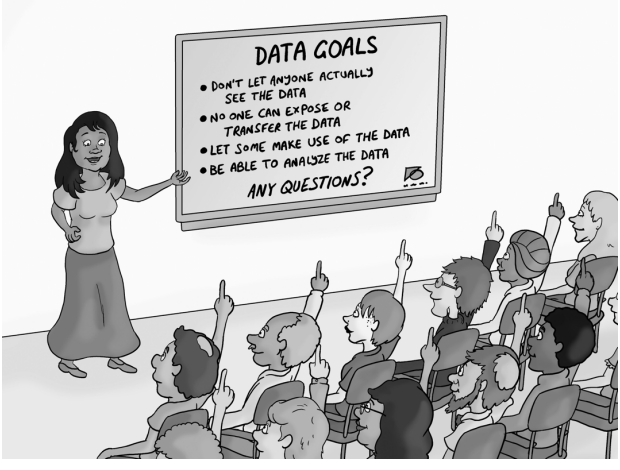
Sponsored by Duality

Duality is the leader in privacy preserving data collaboration empowering organizations worldwide to maximize the value of their data without compromising on privacy or regulatory compliance. Founded and led by world-renowned cryptographers and data scientists, Duality operationalizes privacy enhancing technologies (PETs) to accelerate data insights by enabling analysis and AI on encrypted data, while preserving data privacy, compliance and protecting valuable IP. A Gartner Cool Vendor, Duality was recently named a Tech Pioneer 2021 by the World Economic Forum (WEF) and listed on Fast Company's 2020 Most Innovative Companies.

For more information, visit dualitytech.com and follow us on LinkedIn and Twitter.

Conversational Privacy Enhancing Technologies & Big Data (Mini Edition)

by Mike Cobb



ConversationalGeek®

Conversational Privacy Enhancing Technologies & Big Data (Mini Edition)

Published by Conversational Geek® Inc.

www.ConversationalGeek.com

Trademarks

Warning and Disclaimer

Additional Information

For general information on our other products and services, or how to create a custom Conversational Geek book for your business or organization, please visit our website at www.ConversationalGeek.com.

Publisher Acknowledgments

All of the folks responsible for the creation of this book:

Author:	Mike Cobb
Project/Copy Editor:	Pete Roythorne
Content Reviewer(s):	Lucy Temprano
	Marcella Arthur

The “Conversational” Method

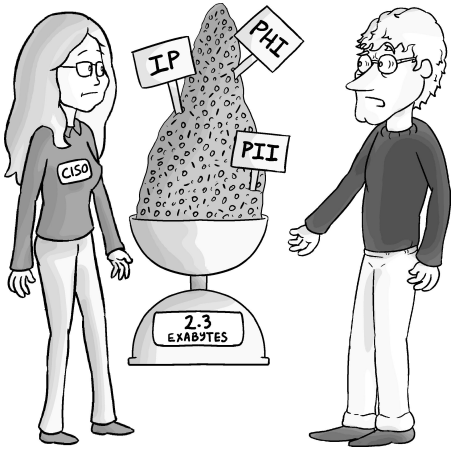
We have two objectives when we create a “Conversational” book. First, to make sure it’s written in a conversational tone so that it’s fun and easy to read. Second, to make sure you, the reader, can immediately take what you read and include it into your own conversations (personal or business-focused) with confidence.

“Geek in the Mirror” Boxes

We infuse humor and insight into our books through both cartoons and light banter from the author. When you see one of these boxes it’s the author stepping outside the dialog to speak directly to you. It might be an anecdote; it might be a personal experience.

Within these boxes I can share just about anything on the subject at hand. Read ’em!

Understanding the Privacy Enhancing Technologies (PETs) Landscape



"So you want me to make all that data encrypted, sharable, and useful?"

Data marts and data warehouses, Big Data, data lakes, and even data swamps! Wow, that sounds like a lot of data. It is, and we're all collecting it in

greater variety and in ever greater volumes. Those that are turning all that data into knowledge are quickly becoming the leaders of their industry or field. In fact, how smart you are with your data is dictating how successful your organization is, and it has for some time. When the world started to go digital, those companies that invested heavily in databases to collect and store data were the first to be able to offer faster and smarter services – think Amazon, Uber, and Airbnb.

These vast quantities of data led to the development of different types of storage systems like NoSQL and Redis to handle both structured and unstructured data so it could be properly organized and quickly accessed. During the 1990s and 2000s the pioneers of mass digital data use began to analyze and mine this data to gain insights into customers and processes to further improve their businesses. Today's big data business intelligence systems transform raw data into easy-to-digest actionable insights and have proved to be essential in predicting trends in customer behavior, eliminating inefficiencies, and identifying areas of future growth. For example, having invested heavily

in big data analytics, Netflix is estimated to have saved $1 billion per year on customer retention due to improved personalization[1]. Can you see a trend here? Companies that make the most of their data perform extremely well.

In 2017, the Economist declared that data had replaced oil as the world's most valuable resource[2]. Unlike oil though, the supply of data is endless, and it can be used multiple times in multiple ways to gain new insights.

In recent years, data analysis has advanced to machine learning and artificial intelligence to eke

---

[1] www.statista.com/topics/842/netflix, and www.forbes.com/sites/enriquedans/2018/05/27/how-analytics-has-given-netflix-the-edge-over-hollywood

[2] www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data

out even deeper actionable insights in order to stay ahead of the competition. And the flow of data just keeps growing. By 2025, it's estimated that 463 exabytes of data will be created each day globally – the equivalent of nearly 213 million DVDs per day[3] – and as data analysis technologies improve, more of this data can be turned into useful information, increasing opportunities for industries, governments, and your organization.

Exciting Opportunities! Is there a Catch?

The collection and consumption of all this data, much of it personally identifiable information (PII), has quite rightly given rise to a wall of data protection and privacy related legislation. Every organization that collects or uses personal data has to contend with both horizontal and vertical data privacy legislation. Horizontal legislation applies across all industries and sectors: the EU's General Data Protection Regulation (GDPR), the California

---

[3] www.weforum.org/agenda/2019/04/how-much-data-is-generated-each-day-cf4bddf29f/

Consumer Privacy Act (CCPA), and California Privacy Rights Act (CPRA) are obvious examples of how governments and states are trying to regulate how the data of millions of citizens is managed to ensure organizations have a legal basis to process and retain it.

Vertical privacy legislation covers a specific industry, so if you're operating in the healthcare industry, for example, you have to comply with the Health Insurance Portability and Accountability Act (HIPAA), while the Gramm-Leach-Bliley Act (GLBA) applies to anyone in the financial industry. Fines for non-compliance with data protection laws can be extremely high: Amazon has already received a GDPR fine of $877 million[4], and more countries like Canada, Australia, India, and China are updating their privacy laws and the level of possible fines.

Clouds of Unusable Data

These often complex and onerous legislative environments have put most companies in a

[4] www.datarainbow.eu/2145-2/

position where they cannot utilize all of their data and remain compliant. So, without strong anonymity and privacy guarantees, medical, financial, military, and other high-security domain research isn't possible. This means around 60% to 73% of data sitting in enterprise data centers goes unused for analytics[5].

This may be better than the situation in 2012 when just 0.5% of the 2.8 zettabytes of global data was used for analysis[6], but for C-suite executives and decision-makers involved with processing, analyzing, and monetizing data it's frustrating to be missing chances to gain a competitive edge or convert their data into valuable knowledge. For example, the World Economic Forum estimates that hospitals produce 50 petabytes of data per year, yet 97% is never used[7]. That's like having an

---

[5] www.forrester.com/blogs/hadoop-is-datas-darling-for-a-reason/

[6] www.theguardian.com/news/datablog/2012/dec/19/big-data-study-digital-universe-global-volume

[7] www.weforum.org/agenda/2019/12/four-ways-data-is-improving-healthcare

Olympic size swimming pool but only being able to put your toes in the water. It's a massive lost opportunity in terms of knowledge and insights that could not only improve decision-making and deliver better customer experiences but also change the world of healthcare and public services.

Is there a Solution?

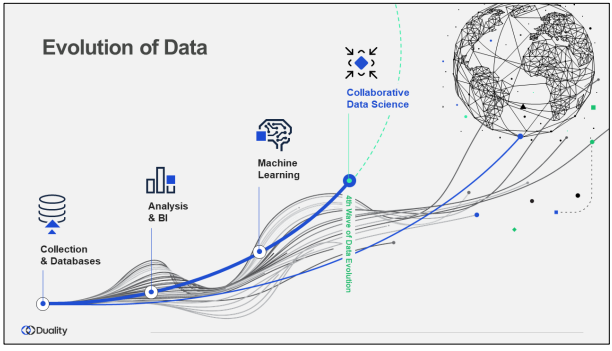
While strong privacy laws are essential, if you do not find a way to make decisions based on all of your data, you will not be able to compete against those that do, and when it comes to domains like health, the entire world loses out. For example, a series of COVID-19 research papers have been retracted as the health records underlying the studies cannot be shared for independent validation purposes[8].

[8] www.nature.com/articles/d41586-020-01695-w

When it comes to extracting value from data, good is no longer good enough. Data resources that are not fully utilized will lead to lost opportunities and lost business.

The lengthy process of establishing bespoke data sharing contracts stifles the ability of researchers to combine and research data from healthcare systems and the pharmaceutical industry which could lead to valuable insights; insights which could accelerate and optimize drug formulations and treatments, radically improving quality of care. Imagine the possible breakthroughs in cancer care if medical professionals could securely search and analyze decentralized medical data across organizational and regulatory boundaries.

Thankfully the evolution of data is entering a new phase and there are now ways for you to unleash the full commercial, scientific, and social potential of data while staying compliant.



Source: Revolutionizing the Data Sharing Landscape: Data Sharing in the Privacy Age

Privacy enhancing and privacy-preserving technologies (PETs) enable the analysis of highly sensitive customer information and intellectual property without sacrificing privacy or protection. They utilize a range of advanced cryptography and statistical techniques to allow data to be analyzed and shared while remaining “encrypted in-use”. PETs also support decentralized data analysis, so data doesn’t have to leave its jurisdiction of origin or even the data center in which it is currently stored. How great is that?

Tell me more about PETs

There are three commonly used PETs that you need to know about, not necessarily in-depth knowledge of the algorithms and methods they use, but how they can free up your data. They are homomorphic encryption, multiparty computation, and federated learning.

Homomorphic Encryption – enables computational operations on encrypted data so it remains secure and private even in untrusted environments. The results of any analysis remain encrypted and only the data owner can decrypt and view them. Slow performance for computationally heavy analysis has been an issue, so it is often used in conjunction with another PET, multiparty computation.

Multiparty Computation – leverages a cryptographic concept called additive secret sharing – a type of homomorphic encryption that allows multiple parties to analyze their combined data by distributing computations amongst them. It can work across untrusted third parties and jurisdictions as there is no need to expose or transfer data outside its local security controls. A secret share is a

fragment of incomplete data about the initial secret value from which it is derived, so affords no useful information on its own. However, when each party locally sums their secret shares to calculate a partial result and these partial results are recombined a valid result can be calculated without any tradeoff between data usability and data privacy.

Federated Learning – to gain sound intelligence a machine learning model needs real and relevant data. This data is often held on mobile devices, IoT sensors, and other edge devices and is protected by privacy laws. Federated learning enables multiple sets of decentralized, on-device data to train a smarter central model. A low-impact training model onboard a device sends no raw data to the central model, only the results needed for a specific computation.

Privacy preserving aggregation using zero-sum masks (random values that cancel each other out when aggregated) encrypts individual results so the central model doesn't need to decrypt each message to compute the combined decrypted result. For additional security, differential privacy can be used to add noise to obscure atypical data

that could potentially expose an individual device and remove the risk of inference attacks reidentifying individuals from aggregate-level results when they are decrypted. While some devices are training the model, others can be used to test the quality of that training. Federated Learning can also be used to encrypt and combine sensitive data from multiple sources for analysis at a datacenter with the results sent back to the inquiring party.

All the data remains encrypted during analysis with only the inquiring party having access to the decrypted results, ensuring the data is fully protected the entire time. The efficient combination of PETs addresses the many privacy risks surrounding sensitive information while allowing access to high-quality data: a key requirement for discovery and innovation.

Other PETs that ensure data remains anonymous and private are synthetic data generation, generative adversarial networks, and zero-knowledge proofs. By combining these different technologies, you can solve the problem of analyzing vast quantities of confidential data for any

manner of projects, such as, improving user experiences, training self-driving cars with aggregated real-world driver behavior, or improving patient diagnostics and treatment plans; the applications are endless.

PETs enable collaborative rather than personalized learning so your business can learn from everyone without learning about anyone.

Use Cases

In the healthcare industry, PETs can remove the problem of working with and studying incomplete data sets which can produce ambiguous results. Real world evidence (RWE) studies can be conducted by aggregating or linking data from multiple data owners, clinics, hospitals, gyms, and so on without any risk of re-identification. As RWE studies are more complete, they can more accurately detect the effectiveness of a drug or treatment, or better predict how best to tackle an

emerging pandemic. A real opportunity to revolutionize the world of medical research and the world we live in.

The United Nations estimates that around $800 billion to $2 trillion are laundered every year with only 1% of illegal transactions being detected and acted upon[9]. A big obstacle to better detection and prosecution is that most leading banks only see 10%-25% of their customers' or prospective clients' full financial picture[10].

A U.S. bank that acts as a correspondent bank for example will have little knowledge of the nature of a transaction between a company in Spain transferring dollars to a company in Colombia as it's the Spanish and Colombian banks that are responsible for monitoring their customer's behavior and they cannot easily share information

---

[9] www.unodc.org/unodc/en/money-laundering/overview.html

[10] www.forbes.com/sites/ronshevlin/2021/11/15/americans-shadow-financial-lives-why-banks-dont-know-jack-or-jill/?sh=549b4480fe34

about them. This makes it really difficult to spot the multi-faceted signs of money laundering. It can take a minimum of six months to complete privacy compliance requirements before data can be cross-checked with another institution's data, creating big gaps in Anti-Money Laundering safeguards. This led to an eye-watering $2.7 billion in AML fines in 2021[11][11]. Using PETs to share data effectively yet securely between themselves, banks can greatly reduce investigation times and obtain a 360 view of a client's transaction history so faster and better decisions can be made to reduce risks and stop illegal activities.

As you can see, healthcare and finance are obvious beneficiaries of PETs, but any industry can greatly benefit from the deeper data analysis that they allow, particularly when it comes to improving customer experiences. For example, being able to tap into the millions of GPS data points collected from mobile devices would allow public transport operators to determine the habits of people using

---

[11] www.forbes.com/sites/forbestechcouncil/2022/03/24/lessons-from-the-seven-largest-aml-bank-fines-in-2021

public transport so operators could optimize bus schedules to match the needs of travelers – increasing transportation efficiency, reducing costs, while improving the service. A win-win for everyone. In fact, PETs have the potential to unlock untold value from data that until now has been off-limits. Have you thought how you could use PETs to take your organization to the next level yet?

Companies and even countries will get left behind if they don't prioritize securely sharing and analyzing their data, and PETs are the only way to do it without sacrificing privacy or competitive value.

Getting Started with PETs

Having to incorporate PETs into your IT processes is not a case of if but when. There's a danger however of your key stakeholders not appreciating the risks of delaying harnessing the power of PETs. Some may not even be aware that the big data and

privacy conundrum can be solved. This means you need to get all concerned in your organization around the same table so these technologies can be explained, questions answered, and doubts replaced with confidence that data research can be collaborative and not mutually exclusive.

The legal team, once they understand how PETs work, should be able to give quicker approval to collaborative ventures as your organization will remain compliant. It is crucial that this is all accomplished before you realize you're falling behind the competition.

An important question you have to answer is which PET specialist vendor to choose and how their platform can be integrated into your existing systems. Any provider will be happy to present how their solution works, but the most important feature of any privacy preserving collaboration platform has to be that it makes compliance a non-issue: data is guaranteed to never be exposed. This means it has to work within your governance frameworks and be able to prove that data is always being handled securely.

Do ask to see real-life case studies of any vendor's solution. Although you may be an early adopter you don't want to be a guinea pig. Look for examples of scalability if you need to analyze large data sets rapidly and ask for an explanation of the data science and cryptography used. Your chosen solution must enable your organization to maximize the value of its data, for whatever purpose, while working together with partners and even your competitors.

Don't be one of those companies that deploy PETs purely to meet its privacy compliance obligations – it is so much more powerful than that.

It's understandable that your CEO and CISO may be reluctant to risk a privacy violation by using what may be perceived as untried and untested technology, but PETs are becoming recognized security-enabling and privacy tools. For example, the European Data Protection Board, which oversees the enforcement of GDPR, and the

European Union Agency for Cybersecurity has published technical guidance supporting multi-party computation as a valid privacy-preserving safeguard[12] while the proposed H.R. 847, Promoting Digital Privacy Technologies Act would require the U.S. National Science Foundation to support research on privacy enhancing technologies[13]. This is probably one of the reasons why Gartner recently predicted that 60% of enterprises would have already used PETs, or be in the process of implementing them, by 2025[14].

---

[12] www.enisa.europa.eu

[13] www.govinfo.gov/app/details/BILLS-117hr847rh

[14] Gartner, *Top Trends in Privacy Enhancing Computation 20*22

The Big Takeaways

Data is now the world's most valuable resource. If it is fully harnessed it can save lives, improve products and services, and increase efficiency and profits. However, a lot of potentially valuable data is protected by stringent privacy legislation to prevent it from being abused or misused. These laws have greatly limited the pool of data that can be used to find better ways of doing things, but they no longer need to be a barrier to innovation.

Although they have been around for a while, privacy enhancing technologies have now matured to the point where they can be applied to real-life problems. In Gartner's Top Strategic Technology Trends for 2022 report[15] privacy enhancing computation was highlighted as a key trend enabling secure data processing. They are also being promoted by government institutions around the world as a way for those with legitimate access

---

[15] https://www.gartner.com/en/information-technology/trends/top-strategic-technology-trends-for-spain-2022-emea-pd

to copious amounts of sensitive and personal data to share and analyze it safely and securely.

There are a growing number of PET-related products and services coming onto the market, some from startups and some from mature companies who are trying to shift into this growing market. To successfully harness the potential of PETs it's critical that you choose a vendor who has experience and expertise in this field and whose platform is enterprise-ready and able to utilize multiple PETs in order to solve your specific and unique business problems.

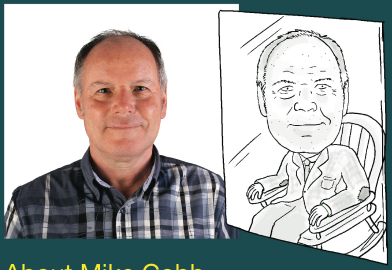
PETs are already being used in a variety of contexts, and the advantages that early adopters gain will soon become apparent as their improved products, services, and experiences begin to dominate. This is clearly a technology your company needs to understand and embrace as it's the next stage in the evolution of data – unleashing the power of data. Data collaboration is the future for most enterprises looking to get or stay ahead of the competition and PETs are the only effective and efficient way you can achieve it.

Which PET is Best?

Duality

dualitytech.com

Data is now the world's most valuable resource. However, a lot of potentially valuable data is protected by stringent privacy legislation, which is limiting the pool of data available to find better ways of doing things. This book looks at how privacy enhancing technologies (PETs) can enable data to be used for the greater good.



About Mike Cobb

Michael Cobb is a renowned security author with over 25 years of experience in the IT industry. He co-authored the book IIS Security and has written countless technical articles for leading IT publications. He has also been a Microsoft Certified Database Manager and registered consultant with the CESG Listed Advisor Scheme (CLAS).