Blueprint for GDPR

2nd Edition

Updated and Expanded

A blueprint for GDPR Data Safe Havens. Compliance controls alone do not unlock Big Data value. Big Data innovation requires embedding BigPrivacy-enabled GDPR Data Safe Havens that protect data when in use for compliant decentralised processing, repurposing, combination and sharing.

If you are not using the BigPrivacy-enabled Data Safe Havens – you are not maximising Big Data value under the GDPR and evolving data protection regulations.

January 2019

M Gary LaFever[1]
CEO & Co-Founder
Anonos Inc.

Anonos®

BigPrivacy® Unlocks Data

Blueprint for GDPR
2nd Edition
Updated and Expanded

(Rev 1.2)

PREFACE

This 2nd Edition of the Anonos BigPrivacy GDPR Blueprint provides updated information and examples of **"Data Safe Havens" which are *explicitly recognized combinations of GDPR legal & technical safeguards that maximise Big Data value by leveraging GDPR compliant pseudonymisation***:

GDPR Data Safe Havens

1. **Legacy Consent Data Transformation - Avoid Deletion: SaveYourData®**
2. **Legitimate Interest Big Data Processing**
3. **Legal Secondary Processing of Big Data**
4. **Data Minimisation; Data Protection by Design and By Default in Support of Big Data**
5. **Expanded Big Data Use and Sharing Opportunities**
6. **Compliant Cloud Processing for Big Data**

In the year since the publication of the 1st Edition of this BigPrivacy GDPR Blueprint in January 2018, Anonos has received recognition as a **Gartner Cool Vendor** for innovative technology, **Gartner, IDC** and **Forrester** have all highlighted BigPrivacy in numerous reports and blogs, and Anonos software has been certified under the EuroPrivacy certification scheme as complying with European GDPR requirements. The certification by EuroPrivacy (www.europrivacy.org), completed using the "Privacy Flag" certification scheme developed under a research project co-funded by the European Commission and Switzerland, highlights that Anonos SaveYourData software meets GDPR requirements for compliant pseudonymisation. Most recently, IDC has published a special report entitled **Anonos' SaveYourData – a EuroPrivacy Certified Solution – "Deep Freezes" Enterprises' Existing Personal Data Sets as They Plan Analytics Strategies** available at www.anonos.com/DoNotDeleteYourData.

KEY TAKE AWAYS

BigPrivacy® technology harmonizes two objectives which have previously been in opposition: it delivers data value through data use, sharing and combination while simultaneously enabling GDPR-compliant data protection. BigPrivacy enables GDPR-specific advantages to accrue to each of the three major internal stakeholders: Business, Technology, and Compliance.

- To achieve a sustainable competitive advantage and drive real value and revenue growth in today's increasingly regulated information economy, organisations must **be on the forefront of maximizing the value** of their data; yet, at the same time, they must **comply with new, more stringent requirements** under the GDPR and similar evolving data protection regulations. These global data protection regulations raise serious issues relevant to different stakeholder groups – both external and internal.

- For an organisation to be truly successful, it must **harmonize data protection and data use**. This can be accomplished by involving all stakeholders within the organisation. This blueprint is divided into three parts – covering *Business, Technology and Compliance* matters. Regardless of an individual's specific perspective, this harmonization is **the means for achieving and sustaining a data-centric competitive advantage.**

- Heightened GDPR requirements for consent are nearly impossible to satisfy with respect to "Big Data" (as defined in this blueprint) processing when iterative analyses, correlations and computations cannot be described with required specificity at the time of consent. In addition, consent cannot be made a condition for receiving a product or service – a data subject must be offered a genuine choice, or their consent is not freely given. If consent is not obtained in full compliance with GDPR requirements, "**the data subject's control becomes illusory and consent will be an invalid basis for processing, rendering the processing activity unlawful.**"[2]

- A data controller **must disclose the lawful basis for data use at the time of collecting personal data**, so it must decide in advance the applicable lawful basis and it cannot modify or "swap" the lawful basis in the course of processing. For example, a data controller "is not allowed to retrospectively utilise the legitimate interest basis in order to justify processing, where problems have been encountered with the validity of consent."[3]

- The GDPR has **no grandfather provision or other exemption allowing for the continued use of "Legacy Consent Data"** (as defined in this blueprint) collected using now legally non-compliant broad-based consent. "If a controller is unable to renew consent in a compliant way and is also unable – as a "one off" situation - to make the transition to GDPR compliance by basing data processing on a different lawful basis while ensuring that continued processing is fair and accounted for, **the processing activities must be stopped."**[4]

- To lawfully process iterative Big Data and to legally use Legacy Consent Data (as defined in this blueprint), GDPR-compliant legal bases are required. After May 25, 2018, **companies that continue to rely on broad-based consent will not be complying with GDPR requirements**. Failure to comply with GDPR obligations exposes parties, including co-data controller and data processor partners, to fines of up to the greater of 20 Million Euros or 4% of annual turnover (global gross revenue), plus additional significant obligations, liability and exposure.

- This brings us to **BigPrivacy**[5] – the patented state-of-the-art technology developed over the last seven years by Anonos, which can serve as **a blueprint for the commercial deployment of technology solutions** that enforce principles of **Functional Separation** to support **Legitimate Interest processing** which, as this blueprint explains, helps organisations to satisfy Data Safe Haven criteria under the GDPR and other evolving global data protection laws while protecting the fundamental rights of data subjects.

ContactUs@BigPrivacy.com

TABLE OF CONTENTS

INFOGRAPHIC*

Shortcomings of Existing Compliance Technology

Organisations need to operate "at the speed of digital business" while at the same time remaining compliant. However, current data protection technologies require organisations to choose whether to favour business at the expense of compliance **– or –** to favour compliance at the expense of business – leaving technology in the middle to reconcile this tension.

* The above matrix represents a subset of functionality considered most relevant by the author for supporting "Data Safe Havens" – explicitly recognized combinations of GDPR legal & technical safeguards that maximise Big Data value by leveraging GDPR compliant pseudonymisation – necessary to maximise the value of personal data use for data processing as described in this blueprint. This matrix is not intended as a complete or exhaustive analysis of the capabilities of the technologies, principles or approaches listed.

** GDPR Article 25 imposes a new mandate for Data Protection by Design and by Default[6] (DPbDD) that is much more than just privacy by design. **It is the most stringent implementation of privacy by design.** DPbDD requires that data protection be applied by design and by default at the earliest opportunity (e.g., by pseudonymising data in compliance with Article 4(5)) and requires that steps must be taken to make use of the data, as compared to the pre-GDPR default, where data is available for use by default and steps must be taken to protect it. DPbDD requires granular, context-sensitive control over data so that **only that data necessary at any given time, and only as required to support each authorized use, is made available.**

*** Support for non-consent legal basis is important because: (1) consent must be freely given and does not extend to the collection and use of personal data that is not strictly needed to provide a product or service, data sets therefore may contain both (a) data for which consent is a valid legal basis and (b) data for which consent is not a valid legal basis; (2) certain data uses do not support the level of specificity required for consent to serve as a valid legal basis; and (3) EU member state laws may prohibit the use of consent as a legal basis for certain personal data uses.

**** For example, de-identification requirements under US laws like the Health Insurance Portability and Accountability Act (HIPAA), the Children's Online Privacy Protection Act (COPPA), and the Gramm-Leach-Bliley Act (GLBA).

Summary Benefits of BigPrivacy Technology

- To maximise Big Data (as defined in this blueprint), **new technical and organisational measures like those enforced by BigPrivacy technology are necessary** to safeguard the privacy rights of data subjects in situations where consent is impractical, unavailable or unattainable.

- **BigPrivacy leverages GDPR-certified technology** to process information to be used for **actionable analytical insights and tangible business benefits in compliance with regulatory Data Safe Havens** (as defined in this blueprint) for processing personal data.

- **By technologically enforcing Functional Separation-based risk mitigation measures**, BigPrivacy technology helps to eliminate the (now) false dichotomy between (i) favouring business or meeting compliance requirements and (ii) between revenue and risk.

- **BigPrivacy's award-winning technology** creates standardised **Variant Twin®** data assets that can be sourced, curated, combined and shared in a trusted, predictable and legally compliant manner. The ability to standardise and scale data asset utilisation transforms data cost centres into revenue centres by extending the value of data – both internally within, and externally outside, an organisation – **to support a whole new ecosystem of compliant Big Data value by enabling**:

    - Aggregation of data across jurisdictions and between different legal entities.
    - Analytics to be processed on protected decentralised data.
    - Cloud-based processing of Big Data analytics, AI, ML and DX.
    - Legal repurposing, combining and sharing of data.

ContactUs@BigPrivacy.com

FOREWORDS

DATA PROTECTION MEGATRENDS

Martin Abrams, Executive Director and Chief Strategist
The Information Accountability Foundation (IAF)
http://informationaccountability.org

There are two data protection megatrends going on today. The first is the breaking wave of transformational data processing laws, regulations, and guidance evolving around the globe, epitomized by the GDPR. The second is the evolution of a data trust deficit into a full-fledged legitimacy conundrum. Yet people expect all the value of a highly observational world. How do global organisations reconcile the growing importance of data analytics, artificial intelligence, and machine learning with the increasingly complex and multi-jurisdictional regulations on lawful data use? And furthermore, how do they maintain trust that is based on both value and protection? The Information Accountability Foundation believes accountability-based information policy management—being a trusted data steward—is the key element of the answer.

The GDPR requires accountability specifically. It requires organisations to have policies, and the processes to put those policies into effect. Those processes rest on new technologies that are demonstrable to assure conditions set by policy actually are actionable. The GDPR introduces these new controls in the form of technical and organisational measures necessary to support data protection by design and by default. Comprehensive data protection impact assessments that balance the interests of all stakeholders is part of organisational controls. Pseudonymisation, as newly defined under the GDPR, is another methodology that enables fine-grained, risk-managed, use case-specific controls necessary to support data protection by design and by default, particularly the fundamental data protection law principle of data minimisation. Data protection by design and by default embodies the goal of making technology controls that support appropriate uses.

A central core of data protection accountability and ethics is the will and ability to demonstrate that you can, in fact, keep your promises. Technologies that enforce data protection by design and by default show data subjects that in addition to coming up with new ways to derive value from data, organisations are pursuing equally innovative technical approaches to protecting data privacy—an especially sensitive and topical issue given the epidemic of data security breaches around the globe.

Vibrant and growing areas of economic activity—the "trust economy," life sciences research, personalized medicine/education, the Internet of Things, personalization of goods and services—are based on individuals trusting that their data is private, protected, and used only for appropriate purposes that bring them and society maximum value. This trust cannot be maintained using outdated approaches to data protection. We must embrace new approaches like data protection by design and by default to earn and maintain trust and more effectively serve businesses, researchers, healthcare providers, and anyone who relies on the integrity of data.

Traditional approaches to data processing often involve the use of static identifiers that enable the ability to infer—or single out or link to—a data subject because static identifiers, when used across multiple data sets, enable the overlay of the data sets so data that is not identifiable by itself, when combined with other overlapping data, leads to re-identification of a data subject. Conversely, data protection by design and by default can leverage dynamically changing identifiers to probabilistically prevent the ability to infer identifying information pertaining to a data subject across multiple data sets or data combinations—all in a manner that is capable of supporting mathematic analysis, audit, and enforcement.

New technologies are being introduced to implement data protection by design and by default. Anonos' first-of-its-kind patented BigPrivacy technology is one example that supports "proportional" use of data in a manner that is responsive to the variety and complexity of different potential uses of data. Specifically, BigPrivacy can reveal different levels and types of information to the same and/or different parties at different times, for different purposes, at different places—and with respect to each, only as necessary for each proposed use of data. By ensuring that only the minimum information necessary for each appropriate purpose is processed by "dialling-up" or "dialling-down" the linkability (or identifiability) of data, BigPrivacy helps to support accountable, ethical, fair, and legal data use.

- Martin Abrams
  Executive Director and Chief Strategist
  The Information Accountability Foundation (IAF)

ETHICAL TOOLS FOR CONTROLLING DISCLOSURE

Jules Polonetsky, Chief Executive Officer
The Future of Privacy Forum (FPF)
https://fpf.org/

Writing in The New Yorker on December 19, 2016 about the work of sociologist Beryl Bellman, Malcolm Gladwell, said ""A secret isn't invalidated by its disclosure, it's defined by its disclosure. What makes a secret a secret is simply the operating instructions that accompany its movement from one person to the next."

Today's world is awash in secrets captured and disclosed by data-driven products and services. With all the personal information collected by wearables, smart homes, social media, smart cars, and innumerable other data-centric offerings, few companies are truly promising individuals privacy. Rather they are committing to responsible use of the data and controlled disclosure. The massive volume, variety and velocity of data created and captured by the ever-increasing numbers of data-driven offerings highlights the need for technical tools that enable those personal information commitments.

Traditionally, de-identification has been a primary method for enabling access to and use of data while protecting individuals' privacy. De-identification has even sometimes been viewed as a "silver bullet" enabling organisations to reap the benefits of data processing while avoiding operational risks and legal requirements. However, scientists have repeatedly demonstrated that purportedly de-identified data sets can be vulnerable to re-identification attacks thereby casting doubt on the extent to which de-identification is a credible method for using and deriving value from data while protecting privacy. Compounding the uncertainty is the fact that re-identification risks only increase as computing technologies become ever faster and the data-centric products and services generate increasingly more data for linkage and analysis.

Thus, weak or unproven promises of de-identification are no longer acceptable to regulators around the world. Proven techniques and processes like the Anonos BigPrivacy technology are the minimum bar called for to support unlocking the value of data while respecting the rights of individuals. While no "silver bullet," if implemented correctly dynamically implemented de-identification can provide the technical operating instructions for both effective legal compliance and an operating system for respecting the secrets shared by individuals.

- Jules Polonetsky
  Chief Executive Officer
  The Future of Privacy Forum (FPF)

CO-CREATION PRIVACY, SECURITY AND SHARING CONCERNS

Richard Bradbury, Vice President & UKI Managing Director
Hitachi Vantara
https://www.hitachivantara.com

At Hitachi, we believe many of the data-driven products and services that will shape our future have yet to be discovered. Yet, it is clear that traditional innovation cycles are not producing the results needed for today's ever-more competitive, connected and convergent environment. Increasingly, organisations are discovering that the most fertile ground for developing, managing and monetizing information to develop data-driven products, services and experiences are at the intersection between companies, their customers and a host of other players in the innovation ecosystem. This requires a new approach called "co-creation," which is the process of innovating with partners in order to create new value for business stakeholders, for customers and for society at large.

Technology such as the Internet of Things (IoT) plays a key role in driving and facilitating these new partnerships, enabling more integration and the availability of vast amounts of data to multiple stakeholders. Hitachi is one of the few companies in the world with decades of experience in developing both operational technology and information technology, the core building blocks of IoT. Hitachi recently commissioned a study to explore the extent to which co-creation is being adopted by companies across industry sectors and the benefits of a more collaborative co-creation approach to innovation. The research surveyed over 500 senior executives and directors at multi-billion-dollar revenue companies across a range of sectors in Europe to understand the state of co-creation and its impact on Social Innovation.

Our research reveals that some of the greatest barriers to co-creation are concerns related to privacy and data security and the lack of a culture that encourages sharing and collaboration around ideas. Data Protection by Design and by Default, as newly defined and required under the EU General Data Protection Regulation (GDPR), is a critical new approach to helping resolve these issues to unlock the full value of data.

The leading innovators of the future will master a much more agile and fluid approach to innovation, where value is co-created with customers, partners, academic institutions and other segments of society. Hitachi believes Data Protection by Design and by Default, as enabled by BigPrivacy technology to enable technical enforcement of policies, can be instrumental in breaking apart the silos that can plague companies, allowing entire departments down to individual employees to connect, acquire and share knowledge with one another in a privacy-respectful yet information-rich manner.

- Richard Bradbury
  Vice President & UKI Managing Director
  Hitachi Vantara

ContactUs@BigPrivacy.com

Anonos®

Blueprint for GDPR 2nd Edition

Updated and Expanded

(Rev 1.2)

I. INTRODUCTION

Today, the greatest asset for many organisations is not on their balance sheet. Rather, it is the information that drives all aspects of data-centric innovation and value creation, including the future of iterative analytics, artificial intelligence ("AI"), machine learning ("ML") and digital transformation ("DX"). (In this blueprint, we collectively refer to the combination of these three activities as "Big Data.")

Gartner predicts that by 2020, more than 40% of enterprise revenue will come from digital business.[7] Similarly, IDC forecasts that by 2020, 50% of the Global 2000 will see a majority of their business coming from their ability to create digitally-enhanced products, services and experiences.[8] **This is the age of *infonomics, "the theory, study and discipline of assigning economic significance to information." Infonomics "provides the framework for businesses to measure, manage and monetize information as a real asset."***[9]

What many organisations do not realize is that these information assets are under attack from a threat which, if not addressed, may destroy much of their information's value – or simply prevent it from ever being realized. The threat is the growing tension between, on the one hand, the value of ever-increasing data linkages, correlations and new discoveries made possible via Big Data and, on the other hand, the risks these inherently bring, leading to new requirements for lawful processing of personal data contained within Big Data. This tension is evident in recently enacted (and pending) laws that significantly increase requirements for securing legally enforceable consent from individuals ("data subjects") for Big Data processing. Of these, none has greater significance than the European Union General Data Protection Regulation (GDPR), which went into full effect on May 25, 2018, following a two-year implementation period after the 2016 passage of the Pan-European law. The GDPR makes it clear that organisations must protect individual "personal data"[10] or face the risk of injunctions terminating the processing of illegal data and fines up to 20 million Euros or 4% of consolidated global gross revenues, *whichever is greater*. This is neither conjecture nor a prediction; rather, this significantly increased risk exposure is a reality.

To achieve a sustainable competitive advantage and to drive real value and revenue growth in today's increasingly regulated information economy, **organisations must be on the forefront of maximizing the value of their Big Data assets; yet, at the same time, they must comply with new, more stringent requirements under the GDPR and similar evolving data protection regulations. These global data protection regulations raise serious issues relevant to different stakeholder groups – both external and internal.**

External Stakeholder Groups

External stakeholders include data protection authorities, vertical industry regulators, outside auditors, stockholders, individual data subjects, and boards of directors. A high-level overview of some of the perspectives of these different external stakeholder groups follows:

- **Data Protection Authorities** – Under the GDPR, data protection authorities (DPAs) are the regulators with authority to issue injunctions to stop or suspend illegal data processing and impose fines equal to the greater of 20 million Euros or 4% of consolidated global gross revenues. While DPAs may or may not be too busy expanding staff, etc.[11] in the near term to aggressively enforce GDPR requirements, as indicated below, they are not the only external parties of interest.

- **Vertical Industry Regulators** – Vertical industry regulators (e.g., banking, insurance, telecommunications, healthcare, etc.) will likely require companies under their jurisdiction to confirm that they are in compliance with the GDPR given the magnitude of potential financial exposure. For example, central banks (e.g., the European Central Bank (ECB), which administers monetary policy of the euro area, and the Bank of England (BoE), the central bank of the United Kingdom) will likely require banks under their jurisdiction to confirm compliance with the GDPR due to the material adverse impact on the banks if they are found to be noncompliant.

- **Outside Auditing Firms** – External auditors retained by organisations to conduct financial audits have the obligation to ensure that issued reports accurately represent the financial viability of the organisations. If such auditing firms do not adequately represent the GDPR preparedness of a client in an audit, they may be liable for failing to adequately audit financial statements.[12] These auditors include the "Big 4" – Deloitte, KPMG, Ernst & Young and PwC – as well as the many other accounting firms who perform audits and issue annual reports on companies.

- **Stockholders** – Material adverse financial results may arise from an organisation not complying with the GDPR and other evolving regulations because of the magnitude of impact on operations from losing access to data, regulatory fines, etc. Inaccurate, misleading or incomplete information about a company's efforts to comply with data protection requirements in publicly filed documents may be relied upon by investors in making investment decisions to their detriment, thereby giving rise to potential claims by stockholders, including under class actions.

- **Individual Data Subjects** – the GDPR for the first time authorizes quasi-class-action law suits by representatives of individual data subjects (e.g., EU non-governmental organisations and advocacy groups) and clarifies that recovery is possible for non-monetary losses like damage to reputation, emotional distress, pain and suffering, etc. – items generally not recoverable under other jurisdictions' laws (e.g., in the US) for data protection, privacy or security claims. Individuals are also concerned about damages from identify theft, privacy and security breaches. In fact, **some commentators predict that liability to data subjects will exceed fines under the GDPR.**[13]

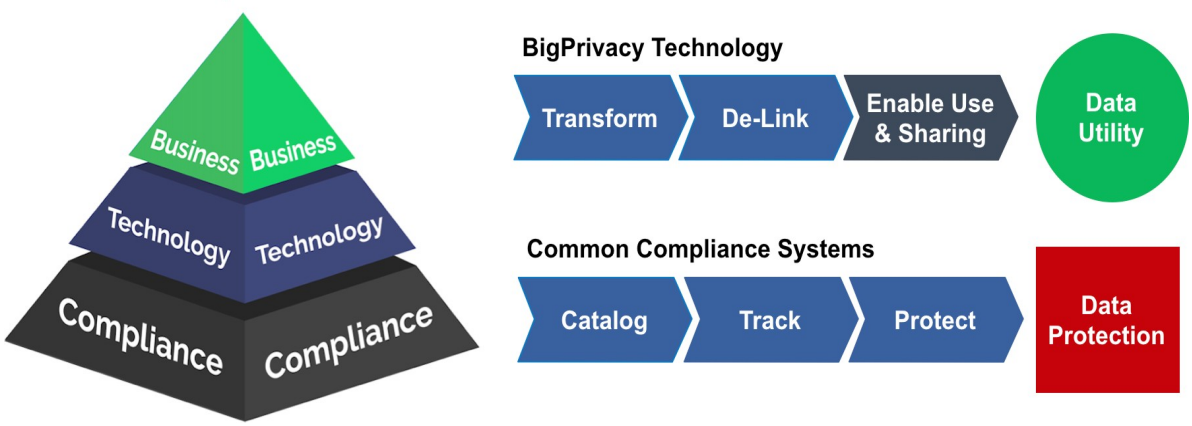
- **Boards of Directors** – There is a global trend toward directors being held personally liable for wrongdoings resulting from legislative changes and increased enforcement activity. A *Financier Worldwide* article notes, "The risk here is also exacerbated by the fact that the General Data Protection Regulation (GDPR) will impose much more stringent burdens on companies that process European Economic Area (EEA) citizens' data from May 2018. If significant penalties are imposed on companies under the GDPR – like the maximum penalties of 4 percent of an organisation's worldwide turnover – shareholders, regulators and others may look to the board to ascertain what went wrong."[14] In the past, CEOs and other C-level executives have had to resign following significant data breaches, for example.

Internal Stakeholder Groups

For an organisation to be truly successful, it must harmonize data protection and data use by fully involving and harmonizing different stakeholder groups within the organisation. No matter an executive's or individual's area of responsibility, they will need buy-in from other stakeholder groups to achieve demonstrable harmonization of data protection and data innovation. This is why we have divided this blueprint into three parts, accordingly covering Business, Technology and Compliance matters. Regardless of an individual's specific perspective, this harmonization is the means for achieving and sustaining an *infonomics* competitive advantage. But even this is not enough. Rather, it is a strong foundation. Having established that, **the imperative for all organisations is to understand how to manage their Big Data assets in the post-GDPR data stewardship era.**

More Than Compliance Is Required

Compliance solutions on the market today were not designed to address new requirements now that broad-based data subject consent fails to legally support Big Data processing under the GDPR and other evolving data protection laws. They were designed to limit or prevent situations that expose an organisation to potential penalty, liability and third-party claims but presume that a valid legal basis to process Big Data already exists. To maximise Big Data value, new technical and organisational measures like those uniquely supported by BigPrivacy GDPR-certified, award-winning technology are necessary to safeguard the privacy rights of data subjects in Big Data processing situations where consent is impractical, unavailable or unattainable. **The blueprint for enforcing Functional Separation-based risk mitigation measures to support valid Legitimate Interest processing of Big Data is BigPrivacy technology.**



II. BUSINESS BENEFITS OF BIGPRIVACY

BigPrivacy's GDPR-certified, award-winning technology creates standardised **Variant Twin®** data assets that can be sourced, curated, combined and shared in a trusted, predictable and legally compliant manner. The ability to standardise and scale data asset utilisation transforms data cost centres into revenue centres by extending the value of data – both internally within, and externally outside, an organisation – **to support a whole new ecosystem of compliant Big Data value by enabling**:

- **Aggregation of data across jurisdictions and between different legal entities.**
- **Analytics to be processed on protected decentralised data.**
- **Cloud-based processing of Big Data analytics, AI, ML and DX.**
- **Legal repurposing, combining and sharing of data.**

The vehicle for maximizing the value of Big Data assets lies in what we refer to in this blueprint as **Data Safe Havens** – that is, explicitly recognized combinations of GDPR legal & technical safeguards that maximise Big Data value by leveraging GDPR compliant pseudonymisation (see Section IV below for further explanation of Data Safe Havens). If you can satisfy these Data Safe Haven requirements, then you can indeed maximise the value of personal data use for Big Data processing. ***The implications here are hard to overstate.*** Compliance with Data Safe Haven requirements means enabling your Big Data to be processed for uses in the future that cannot be described with specificity today.

The GDPR is the beginning, not the end. Other global regulations are following the lead of the GDPR and are imposing similar restrictions while specifying the same or very similar Data Safe Haven criteria.

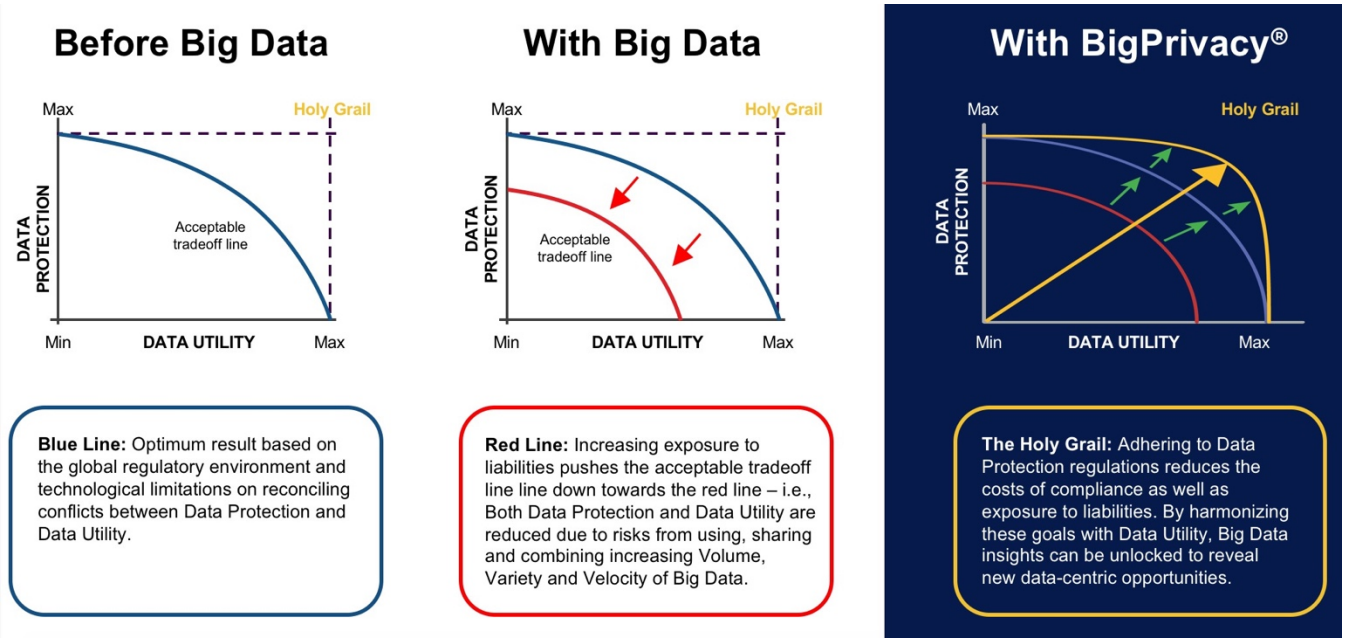
EU personal data that was lawfully processed for years – even for decades – may expose your organisation to significant legal liability starting May 25, 2018 under the GDPR.

The question it is not whether desired data processing activities are technically possible, but rather whether desired processing activities are lawfully permissible.

Concerns regarding the lawful use of personal data for Big Data processing extend beyond newer laws like the GDPR. Further complications are that laws, which predate modern technologies, are being applied to *new* uses of data. For example, in 2015, the US Federal Trade Commission ("FTC") issued a report on the Internet of Things ("IoT").[15] In this report, the FTC stated that to protect the privacy of IoT data, one should delete the data. This was so absurd on its face that two commissioners refused to endorse the report.[16]

This brings us to BigPrivacy, the patented technology developed over the last seven years by Anonos, which serves as a blueprint for the commercial deployment of technology solutions that enforce principles of Functional Separation which, as this blueprint explains, help organisations to satisfy Data Safe Haven criteria under laws like the GDPR while protecting the fundamental rights of data subjects.

What BigPrivacy technology does is to reduce risk from using data (i.e., "de-risk" data) anywhere in the data flow – from the point of collection to the data lake itself – to create data that are non-identifying, dynamically de-identified derivative versions of original data[17] that we refer to as **Variant Twin®** data. Variant Twins enable organisations to *lawfully* process, combine and share Big Data to maximise its value, by helping to justify the processing of personal data under GDPR Article 6(1)(f) **Legitimate Interest** grounds, and also helping to comply with the data minimisation principle and Data Protection by Design and by Default obligations. Because the GDPR does not focus on technology for its own sake but rather on encouraging the use of technical and organisational measures to help protect the fundamental privacy rights of data subjects, **BigPrivacy Variant Twin data provides precisely those types of measures (pseudonymisation and data minimisation, among others), all within the broader Data Safe Haven contexts.**



Practically speaking, the GDPR frequently recommends *pseudonymisation* of personal data. While *anonymisation* of information means (in the EU) irrevocably severing all links between data and the data subject, *pseudonymisation* (as newly defined under the GDPR) means maintaining those links (suitably accessible by encrypted keys and the like) in the hands of authorized parties only and requiring access to secured keys to see the underlying data or to reveal linkages to underlying data. For example, imagine that someone with an incurable medical condition is having their health data used to further investigational drug discovery. If an effective drug is discovered, GDPR-defined pseudonymisation enables the person to be contacted, treated and cured, whereas anonymisation makes it theoretically impossible to find that person again.

In addition to being a specifically enumerated means of helping to achieve Data Protection by Design and by Default, pseudonymisation is cited more than ten additional times in the GDPR[18] as an exemplary safeguard to help harmonize **"the protection of fundamental rights and freedoms of natural persons in respect of processing activities" while enabling the free flow of data to advance legitimate business objectives.**

A long-standing tenet of EU data protection law, embodied in the GDPR, is the concept of *data minimisation*. This includes disclosing the smallest amount of data necessary to the smallest number of people needing it, and being able to disclose different data to different people, all in accordance with those persons' actual minimum authorized data use needs and requirements. Data minimisation principles are at the heart of Data Protection by Design and by Default requirements.

The principle of Functional Separation involves using technical and organisational safeguards to separate information value from identity to enable the discovery of trends and correlations independent from applying the insights gained to the data subjects concerned. Under the GDPR, Functional Separation is embodied within the definitional requirements for GDPR compliant pseudonymisation that the information value of data is separated from identity and that additional secured information is required to relink information value to identity only under authorized conditions. The principal of Functional Separation exists under other evolving data protection laws using different terms - e.g., "De-Identification" under the California Consumer Protection Act and the proposed Indian Data Privacy Law and "Anonymization" under the Brazil Data Protection Law.

The California Consumer Protection Act (CCPA) enforces Functional Separation by what is defined as protected "Personal Information" under the Act. Personal Information includes "information that identifies, relates to, describes, is capable of being associated with, or could reasonably be linked, directly or indirectly, with a particular consumer or household." The CCPA extensive list of Personal Information includes "static" and even "probabilistic" tokens (replacement identifiers) used to replace personal information that "more probable than not" can be used to identify a consumer or device. While restrictions under the CCPA do not apply to "de-identified data," traditional approaches to de-identification do not satisfy the heightened requirements for de-identification under the CCPA. CCPA de-identification requirements are not satisfied using "static" and "probabilistic" tokens (replacement identifiers) because they fail to adequately separate information value from identity to prevent unauthorized re-identification of data subjects via the so-called "Mosaic Effect," discussed later.

Anonos' first-of-its-kind patented BigPrivacy technology helps to safeguard the rights and interests of data subjects by technically enforcing data minimisation, Data Protection by Design and by Default and Functional Separation principles leveraging certified GDPR compliant pseudonymisation. **This approach reconciles the growing importance of Big Data processing with increasingly complex and multi-jurisdictional restrictions on lawful data use.** Equally important, and elaborated on in Section III of this blueprint, BigPrivacy uniquely uses dynamic, rather than static pseudonymous identifiers, thereby reducing re-identification of personal information via the Mosaic Effect.

For these reasons, BigPrivacy uniquely enables data controllers[19] and data processors[20] to unlock the full value of Big Data and maximise the value of information in the global data-driven economy while also respecting the rights of individuals. ***BigPrivacy enables organisations to move beyond mere compliance and highlight a commitment to ethical, legal, and regulatory compliant operations that benefits customers, brands, reputations, and bottom lines.***

III. TECHNOLOGY BENEFITS OF BIGPRIVACY

Evolving Global Regulatory Restrictions

The graphic to the left depicts the awkward position that technology teams find themselves in within the context of the GDPR. They are "sandwiched" between conflicting goals and objectives of their business and compliance colleagues. On the one hand, business colleagues desire technology that enables access, use, sharing and combining of Big Data needed to maximise benefits, value and revenues for their organisation. But at the same time, compliance colleagues are looking for technology that enables compliance with evolving global concerns related to:

- **Misuse:** Risk from data misuse, abuse, unauthorized disclosure and loss, which can limit access to and use of desired data; and
- **Liability:** Suitable technical and organisational measures can help to reduce impacts from the expanded exterritorial reach of the GDPR, dramatically increased levels of liability, and (effectively) joint and several obligations among data controllers and processors up and down the data service "stack."

Limitations of Security-Only and Privacy-Only Compliance Solutions

Organisations need to operate "at the speed of *digital* business" while at the same time remaining compliant. However, current approaches require organisations to choose whether to favour business at the expense of compliance or to favour compliance at the expense of business – leaving technology in the middle to reconcile this tension. Technology developed to support compliance prior to the GDPR fell into one of two broad categories: security-only or privacy-only technology:

- **Security-Only:** These "point solution" tools focus on preventing unauthorized use or loss of data. Security-only technologies can help to comply with certain GDPR security requirements, but are not architected, and fail, to satisfy new requirements necessary under the GDPR for Data Safe Haven use of non-identifying Big Data.

  Security tools such as encryption, hashing, static or stateless tokenization, data masking and related approaches help to protect against the unauthorized identification of data subjects using data that directly reveals the identity of a data subject within a single data source (i.e., static data use). However, those tools do nothing to protect against unauthorized re-identification of data subjects by correlating data attributes that exist in multiple data sources (i.e., dynamic data use) to reveal identity via linkage attacks (the "Mosaic Effect"[21]).

- **Privacy-Only:** Prior to the GDPR, privacy was protected primarily by implementing siloed technologies (e.g. preference/consent management tools), publishing privacy and data use commitments (e.g. privacy statements), and using written contracts, "click-through" agreements and Terms of Service (ToS) that set forth what organisations are authorized to do, or not do, with data. However, for non-technical, non-preventive, contract-based measures to remain effective, controllers require resources and access to monitor compliance by their counterparties to contractual commitments. Such monitoring is typically unavailable or impractical to implement. Contract-based measures had also placed the risk from inadequate data protection on data subjects, due to limited recourse in practice against data controllers and processors for privacy violations.

  Technologies developed to safeguard privacy rights either work on a binary access/no access basis (e.g., data masking) or on an aggregated basis to support generalized statistics/analysis. In today's changing regulatory landscape, these technologies either fail to comply with new GDPR standards for modern digital processing or cannot support business needs for increased access to personal data under controlled conditions. As examples, a technology that enforces "enclave" protection by limiting access to data will not support large-scale production use by definition and technology that aggregates or anonymizes data removes the risk of unauthorized re-identification at the cost of constraining the utility and value of the data because it eliminates linkages.

Data Safe Haven Protection & Enablement – Data Protection by Design and by Default

BigPrivacy reconciles conflicts between data protection and utility to support Data Safe Haven data protection and enablement of personal data. BigPrivacy does much more than simply manage and govern data – it leverages emerging technologies to process information to be used for actionable analytical insights and tangible business benefits in compliance with regulatory Data Safe Havens for processing data (i.e., Variant Twin data). In the context of the GDPR, BigPrivacy helps to reduce costs of GDPR compliance and exposure to liabilities while increasing the value of Variant Twin data as an asset with the controlled ability to relink to identifying versions of data with proper legal basis (informed consent or other) – all in a GDPR-compliant risk-based manner. By technologically enforcing risk mitigation measures required for Data Protection by Design and by Default, BigPrivacy helps avoid having to make tough choices between favouring business or compliance resulting in increased risk or lost revenue respectively.

Security-only and privacy-only solutions are coarse-grained, top-down solutions, rather than fine-grained, bottom-up solutions. Conversely, BigPrivacy leverages patented, risk-based dynamic de-identification capabilities to programmatically support wide ranges of granularity. BigPrivacy's fine-grained, bottom-up approach enables the disclosure of only a part of a data subject's personal data, all the way down to a single data element such as individual direct identifiers (e.g., name) and indirect identifiers (e.g., birthdate or address); it also enables different parts of the data subject's personal data to be disclosed to different people in different places over different time periods for different purposes. BigPrivacy enables organisations to continue to conduct business and process data without compromising data subjects' privacy.

BigPrivacy reconciles data protection and utility via continuous, adaptive and risk-based fine-grained, bottom-up controls to help achieve data minimisation and measures required for Data Protection by Design and by Default as noted below.

BigPrivacy Applies Data Element Level Controls

CONTINUOUS

Dynamic De-Identification

- Content (entropy/obscurity inserted at data element level)
- Context (technically enforced use case-specific policies)

Artificial Intelligence

- Fine-grained generalization and suppression enforced via programmatic privacy actions

Variant Twins

- Non-identifying digital twins

Cross Device/Application

ADAPTIVE

Inputs

- Structured data
- Unstructured data
- Mesh (disparate sources)

Outputs

- Privileged access management (PAM) controls based on credentials of:
  - Person
  - Place
  - Purpose
  - Other

RISK-BASED

Dynamic De-Identification = *Next-generation analytics* with "next-gen" improvements over:

- Tokenization: Defeats mosaic effect from use of static tokens; enables granular controlled relinking.
- Masking: Reversible, data versus presentation layer, obscured information value is not just hidden.
- Multi-Party Computation (MPC): Granular control over levels of disclosure and relinking.
- Differential Privacy: Benefits of differential privacy with controlled re-linkability for 100% context retention; programmatic k-anonymity & l-diversity risk-based adjustments.

BigPrivacy leverages **Continuous**, **Adaptive** and **Risk-Based** controls to enforce Data Protection by Design and by Default by leveraging three critical mechanisms down to the individual data element level, rather than only at the individual user level. These three mechanisms are **Pseudonymisation**, **Functional Separation** and **Data Minimisation**.

A. Pseudonymisation

Digital Rights Management ("DRM") techniques are already widely used by many companies to limit the numbers of copies individuals can make of, or otherwise how they can access, music, movies and other digital content. BigPrivacy uniquely employs DRM-like principles, but "stands DRM on its head" in a manner that Anonos refers to as Privacy Rights Management® or PRM®.[22] Specifically, BigPrivacy's fine-grained controls enable the selective use and sharing of "protected data" (defined for the purposes of this blueprint as personally identifiable data protected by regulation, e.g., personal data under the GDPR; and other data with respect to which selective access is required) with improved multi-stakeholder engagement, all without exposure to unnecessary privacy, security and degradation-of-value risks.

Current approaches to data protection employ outdated "static" approaches to pseudonymisation. As a result, supposedly pseudonymised data can be traced back to a data subject because the persistent or "static" pseudonymous tokens used for a given data element do not change. Searching for a random, pseudonymised string which repeats itself within or across databases can provide a malicious actor or interloper with enough information to unmask the identity of a data subject. Two well-known examples of such unauthorized re-identification are the AOL[23] and Netflix[24] search examples.

Now consider an alternative: what happens if the same data element is each time replaced with a different pseudonymised token, where each different token bears absolutely no algorithmic relationship to any others. The same malicious actor or interloper can no longer determine that these "dynamic" pseudonymous tokens belong or relate to the same data subject, let alone uncover a data subject's name or other identifying information. This is the approach BigPrivacy takes, replacing persistent, unchanging pseudonymous tokens used in current approaches to data protection with patented, dynamically-pseudonymised tokens (referred to herein as "Dynamic De-Identifiers") for each use and/or for each type of use. This substantially mitigates otherwise present risks of linkability and re-identification. By way of illustration, let us explore the differences between *static pseudonymisation* and BigPrivacy *dynamic pseudonymisation* more closely.

Static Tokenisation

Consider the following simple example. A database contains a postal code value of "20500," so static tokenisation creates a token value of, e.g., "6%3a8"; every time postal code 20500 is found, the exact same token ("6%3a8") replaces it. Due to advances in technology and threat-actor sophistication, such persistent (static) tokens can be readily linked back to individuals via the "Mosaic Effect" without requiring any access to additional information to reveal that token's value. An example of the "Mosaic Effect" is where three seemingly "anonymous" data sets that use persistent (static) pseudonyms – each composed of the zip code, birthdate and gender of US citizens – were combined to identify up to 87% of the population of the United States by name.[25]

BigPrivacy Dynamic Pseudonymisation

In contrast, Anonos BigPrivacy combats the Mosaic Effect by using Dynamic De-Identifiers as described above. BigPrivacy by default replaces structured and unstructured direct identifiers (such as name) and indirect identifiers (such as birthdate or address) with randomly generated dynamic pseudonymous tokens called "Replacement Dynamic De-Identifiers." Also, by default BigPrivacy replaces every data element with a different randomly generated Replacement Dynamic De-Identifier. If certain data fields are determined by a data controller as not requiring protection, the data controller can selectively "opt-out" those data elements from replacement. The data controller manages a highly secure master look-up database which contains information necessary to, under technically controlled conditions, relink all connections between direct and indirect (structured and unstructured) identifiers and Replacement Dynamic De-Identifiers.

BigPrivacy, however, does more than this. It not only changes pseudonymous tokens over time, but it can further limit their use to specific contexts involving people, place, purpose or other attributes. In fact, a paper funded in part by the EU Horizon 2020 Framework Programme for Research and Innovation,[26] which evaluated techniques for anonymizing, pseudonymising and de-identifying personal data under the GDPR, highlighted BigPrivacy's patented technique for "changing pseudonyms over time for each use or each type of use as a way to mitigate linkability." Further, rather than opposing or excluding privacy-enhancing techniques such as k-anonymity, l-diversity and t-closeness, the EU Horizon 2020 Framework Programme for Research and Innovation paper highlights that BigPrivacy technology supports their deployment to "provide stronger protection when the data controller observes that re-identification risks increase."[27]

In these cases, Anonos BigPrivacy uses dynamic pseudonymisation to introduce uncertainty (entropy) down to the data element level. From there, the data controller can selectively reveal only the original data (cleartext), which a given user is authorized to see down to a unique data element level, at a specific time for an authorized use. Dynamic versus static pseudonymisation is essential because the GDPR defines "pseudonymisation" as requiring that personal data cannot be attributed to a specific individual without the use of additional information that is subject to protective technical and organisational measures. With BigPrivacy, dynamic pseudonymisation is subject to fine-grained, selectively controlled and technically enforced policies and procedures under the purview of the data controller to enable Data Protection by Design and by Default.

B. Functional Separation

There is also a middle ground between disclosure and non-disclosure, through pre-processing or "data binning."[28] For example, someone who is age 34 is also between the ages of 30 and 40; someone whose blood pressure is 120/70 also belongs to the group of people whose blood pressure is lower than 130/80, etc. That is, instead of a specific data value contained in a given data element, one can refer to a range of data values within which the specific data value fits. With BigPrivacy, this is accomplished by using a different type of Dynamic De-Identifier, called an "Association Dynamic De-Identifier." This feature is in addition to enforcing selective access to protected data down to the data element level via dynamic replacement using Replacement Dynamic De-Identifiers.

Representing a specific value as being contained within a range of values has been referred to by some regulators as *Functional Separation*,[29] i.e., separating the information value of the data from means by which the individuals to whom that data pertains can be identified. Data controllers can therefore use BigPrivacy-enabled Association Dynamic De-Identifiers to "bin" (or categorize) data attributes that fall within the same class, set, group, etc., before being distributed or viewed for greater control over the dissemination of personal information. (*Further discussion about functional separation is included in Section IV below in the context of Safe Haven #3.*)

How can the information value of data be preserved even when re-identification is prevented in most cases to the end user? Because many uses of data simply require ranges in order to perform the statistical analyses being undertaken. Other uses of data may concern information in the aggregate, but not information that can theoretically be tied back to a particular individual. In such cases, data controllers can authorize processing of data that is less identifying (linkable) than original information.

In other words, Functional Separation earlier in the process enables data to be processed further for authorized purposes without enabling that data to be linked to the individual – unless expressly authorized by an individual with their specific and unambiguous consent or otherwise permitted under another recognized legal basis. Only under controlled, authorized conditions can Association Dynamic De-Identifiers be relinked to the original data.

We have referred above to BigPrivacy's ability to dynamically pseudonymise data at a very granular, fine-grained level. Functional Separation adds to this capability as compared to most traditional "access controls." BigPrivacy-enabled Functional Separation helps protect companies against insider threats, which is still a common security risk.[30] The technology can ensure that employees can only access, use and edit the minimum personal data necessary to do their jobs – and no more.

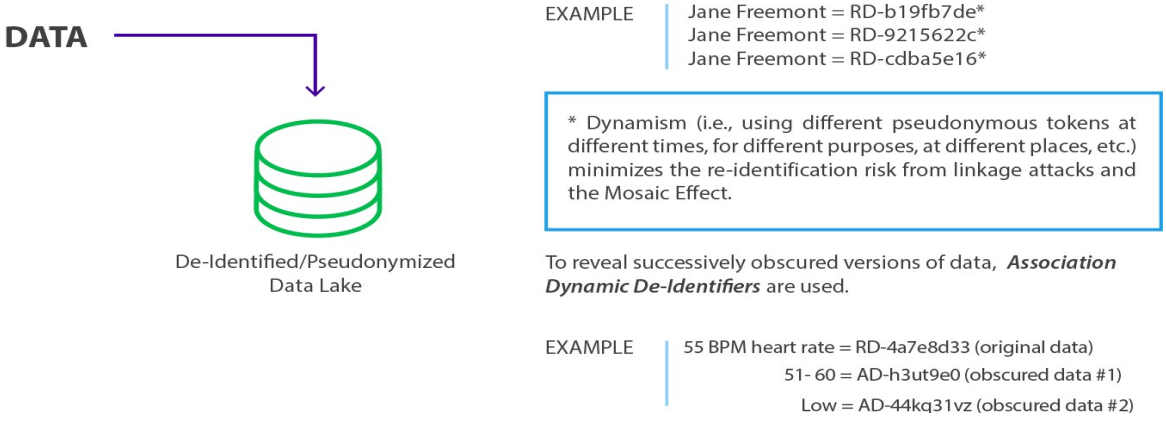
None of these advantages achieved through finer granularity prevent BigPrivacy technology from being used to reproduce or transmit up to 100% of the original value and utility of data; the difference is that BigPrivacy allows information access to the extent necessary to support each authorized use, which is essential to meeting GDPR Data Safe Haven requirements. BigPrivacy thus controls "identifying" (Replacement Dynamic De-Identifier) and "associating" (Association Dynamic De-Identifier) data elements so that data uses are possible only for those properly authorized through contextually and semantically relevant and secure keys. Any time additional properly- authorized data uses arise, BigPrivacy retains all the original data's value and utility, accessible and usable only under technologically enforced conditions by authorized parties.

Because of the importance to BigPrivacy of Replacement Dynamic De-Identifiers and Association Dynamic De-Identifiers, let us examine them more closely. Values of each and every selected identifying data element are replaced and/or augmented based on the desired type of dereferencing.[31]

1. **Identity Dereferencing** – when a Dynamic De-Identifier is used to replace a data element by pointing, via a key, to the value of the replaced data element, the Dynamic De-Identifier is called a *Replacement Dynamic De-Identifier*.

2. **Association Dereferencing** – when a Dynamic De-Identifier is used to obscure an underlying data element value by mapping that value into a broader range or by using some other association of/ correlation with the replaced data element, all so the value of the information conveyed in no way communicates any identifying information about the individual to whom the underlying data element value relates, the Dynamic De-Identifier is called an *Association Dynamic De-Identifier*.

Data elements are replaced with ***Replacement Dynamic De-Identifiers*** that are randomly assigned (not algorithmically derived) for maximum security. Access to a master look-up database is required to access keys that reveal original source (identifying) values of data elements and/or obscured versions of data elements and/or associations among data elements.



The above figure highlights BigPrivacy technology's Functional Separation capabilities. First, each instance of the data element "Jane Freemont" is replaced with a different randomly assigned Replacement Dynamic De-Identifier (i.e., RD-b19fb7de, RD-9215622c and RD-cdba5e16) that is then disclosed, so that, without access to keys that reveal the original values and/or correlations among these data elements, no relationships and no identifying information will be disclosed. Secondly, different levels of identifiability/obscurity are provided to different persons using the data as follows:

- **Scenario A:** one person may be provided with the key associated with Replacement Dynamic De-Identifier RD- 4a7e8d33 to reveal the data element value "55 BPM" (original data), representing a heart rate of fifty-five beats per minute.

- **Scenario B:** a second person is provided with the key associated with Association Dynamic De-Identifier AD-h3ut9e0 to reveal the obscured data element value of "51-60" (obscured data #1), representing a range of fifty-one to sixty beats per minute.

- **Scenario C:** a third person is provided with the key corresponding to the Association Dynamic De-Identifier AD-44kq31vz to reveal the obscured data element value of "Low" (obscured data #2), representing a heart rate that has been classified as low.

Replacement Dynamic De-Identifiers, as well as Association Dynamic De-Identifiers, change dynamically and are temporally unique[32] whether used for a different analysis or purpose by the same or a different party, where such parties may be different individuals within the same organisation. This dynamism mitigates linkability as highlighted in the EU Horizon 2020 Framework Programme for Research and Innovation funded paper cited above.

BigPrivacy thus provides and implements localized, technology-enforced policies for controlling the sharing of protected data in a dynamically de-identified format. A limited number of authorized parties within each data controller manage access to keys used to grant fine-grained access to "perturbed"[33] or original versions of data elements under technically controlled, risk-based, use case sensitive conditions.

C. Data Minimization

Each episode of Star Trek,[34] a popular science fiction television series involving space travel, begins with a speech by the captain of the (fictitious) starship *Enterprise* that closes with the phrase "…to boldly go where no man has gone before." The mission of Anonos, since inception, has been to boldly go where no man has gone before, however, in the context of enabling privacy-preserving data use and sharing, rather than space travel.

Imagine that two (or more) groups within the same organisation, or two (or more) separate organisations, desire to share, combine and process data in a privacy-preserving manner. In both scenarios, each party desires to learn the results of a coordinated analysis without revealing private data. This is exactly what BigPrivacy technology makes possible. Organisations (or groups) can use BigPrivacy to run algorithms against the union of private data, all without allowing any party to view the other parties' private information. This process, sometimes referred to as "Multi-Party Computation" or "MPC"[35] is accomplished by technically minimizing and harmonizing the data exchanged by the parties.

By technologically enabling the use of the minimum level of linkable (identifiable) data necessary for each processing to protect personal data on a per-authorized-use basis, BigPrivacy technologically enforces data minimisation for many applications, without requiring changes to existing systems. This is different from BigPrivacy's capability to accomplish Functional Separation of data as described above (e.g., by converting a specific value into a range of values in which it fits). In this instance, BigPrivacy restricts delivery to only those specific values that are genuinely needed for an authorized use – and no others. This is one of the essential aspects of data minimisation.

Data minimisation can be even further constrained. For example, BigPrivacy can automatically check requests to reveal data to ensure compliance with privacy-enhancing techniques such as k-anonymity[36] and l-diversity[37] levels established by a data controller.

An example of BigPrivacy-enabled Multi-Party Computation follows: For each party willing to share data, BigPrivacy ingests source data in its existing format and transforms the data into a harmonized, dynamically de-identified/pseudonymised format. As the data is de-identified, BigPrivacy converts it into a harmonized format so it can be processed among disparate parties. More technically, Anonos BigPrivacy enforces privacy-preserving functional interoperability without requiring syntactic interoperability,[38] semantic interoperability,[39] the disclosure of identifiable (linkable) private data or a reliance (solely) on non-technically enforced data-sharing arrangements.

A hypothetical example of BigPrivacy-enabled Multi-Party Computation is depicted below. In this example, the dark blue rectangles represent private data owned by Party A, and the dark green rectangles represent private data owned by Party B. Party A and Party B agree in advance on a Variant Twin schema that enables each of Party A and Party B to convert their identifying "1 to 1" deterministic private data into a granular but non-identifying Variant Twin version of their data (represented by the gold rectangles). They accomplish this by grouping their data into sets of records indistinguishable from each other with respect to certain identifiers (e.g., direct identifiers like name and/or indirect identifiers like birthdate).

Each such record set, referred to as a cohort or "equivalence class" (i.e., a class of individuals within a set that are in a given equivalence relation) consists of a specified minimum number of individuals that are indistinguishable one from another, thereby representing a re-identification risk level determined to be acceptable by the data controller for a desired intended data use. In this manner, one division within a financial institution (Party A) could exchange a Variant Twin version of their private data with another division (Party B), enabling the two divisions to run algorithms against the union of the private data without divulging identifying information.

After analysis and computation is completed, each party can map the results of the analysis back to original identifying data associated with applicable equivalence class Variant Twin information retained by their specific implementation of BigPrivacy technology.

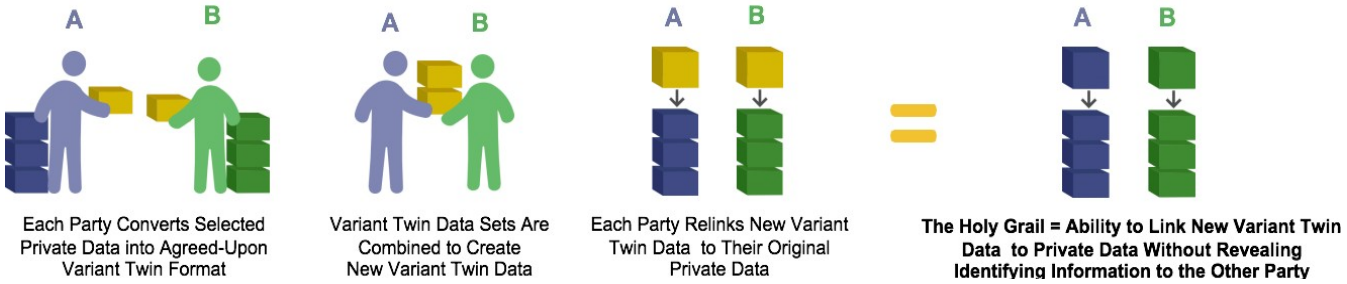
Gartner
Cool Vendor
2018

Gartner awarded Anonos Cool Vendor status for innovative technology *because of the ability of BigPrivacy to create nonidentifying versions of personalized data –**Variant Twins** – that enable compliant Big Data analytics, AI, ML & DX. Gartner highlights the benefits of BigPrivacy-enabled Multi-Party Computing (MPC) to maintain the confidentiality of Personal Data for compliant data sharing and collaboration, profiling and Next Best Action analysis.*

See https://www.anonos.com/coolvendor

Lawful Sharing of Non-Identifying Variant Twin Data (Structured and Unstructured) between Divisions or Organizations Creates New Data Value

1. Value of Combined Variant Twin Data
2. Value to Party A of Party A original private data enhanced by combined Variant Twin Data
3. Value to Party B of Party B original private data enhanced by combined Variant Twin Data



When data are undergoing processing, it is the required results which determine how little information needs to be disclosed in the first place. These constraints can include the minimum level of identifiability, obscurity or precision necessary for participating organisations to achieve results arising out of the union of their private data in a privacy-preserving manner.

Each organisation retains sole control over its own protected data; therefore, each organisation retains the sole ability to relink these results at an identifying level under technically controlled, authorized conditions. In this manner, BigPrivacy increases the accuracy of processing as outlined in the data scientist expert determination attached to this blueprint as Appendix 1.

BigPrivacy thus provides very specific benefits over other privacy-preserving techniques. For example, consider differential privacy and homomorphic encryption:

- **Differential Privacy** – Differential Privacy imposes trade-offs between data utility and information leakage. With Differential Privacy, "the more you protect individual privacy, the less accurately you can compute aggregate statistics about the collection. There's no free lunch."[40]

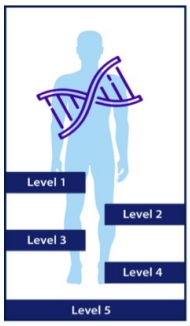
  When using Differential Privacy, high-level aggregated global statistics are made available, but detailed, accurate data cannot be achieved.[41] BigPrivacy does not suffer from this trade-off; rather, it decreases the need to distort, delete or otherwise impair the quality or efficiency of data when processing, all of which are limitations of Differential Privacy, while simultaneously supporting the ability of a data controller to relink to identifying versions of data under technically controlled, authorized conditions.

- **Homomorphic Encryption** – Homomorphic encryption allows one to compare, search or otherwise process data that remains encrypted (i.e., without ever decrypting it). Many consider it to be the "holy grail" for research because of its potential to enable comparison of separate, fully encrypted, sensitive datasets without revealing underlying sensitive data.[42] However, homomorphic encryption does not support sophisticated analyses[43] like detailed-level data mining because it does not support the mathematical operations necessary to perform this level of processing.[44]

BigPrivacy supports data mining at a detailed level, plus re-linkability to identifying data under controlled conditions. In contrast to homomorphic encryption, which involves limited variables (or predicates) and is expensive to compute, BigPrivacy supports easy-to-compute context- and privacy-preserving derived variables (or predicates) specifically designed to support desired authorized processing.

In genomic research, BigPrivacy technology enables organisations to share, combine and analyse data at predetermined non-identifying phenotypal (e.g., disease state) and genotypal (DNA) levels to enable information-rich but privacy-preserving data processing.

BigPrivacy enforces lawful use, sharing, combining and controlled linking of data using Variant Twins. *If new authorized data uses arise, all original data value and utility are retained to support them under technologically enforced conditions.*



Level 1: Pathways bearing mutations and subjects in binary cohort

Level 2: Level 1 + Genes bearing mutations and detailed disease classification

Level 3: Level 2 + Specific gene variants and disease class scores

Level 4: Level 3 + Hapmap haplotype results and full disease history

Level 5: Level 4 + Full SNP data and full patient record

Genomics data is one of the most privacy-sensitive forms of information. Yet, BigPrivacy enables organizations to share, combine and analyze data at predetermined non-identifying phenotypal (e.g., disease state) and genotypal (DNA) levels to enable information-rich but privacy-preserving "further processing." For example, in the figure above, phenotypal/genotypal data at Level 2 or 3 (and possibly Level 4) may be made available for information-rich but privacy-preserving "further processing."

Organizations can use, share, combine and control linking of data without revealing protected data to enable information-rich but privacy-preserving processing by limiting disclosure using Variant Twins.

Without revealing protected data, organisations can similarly share, compare and analyse non-identifying health-related data. This BigPrivacy dynamic de-identification and pseudonymisation enables information-rich but privacy-preserving processing by limiting disclosure to predetermined classes, sets, groups, etc.

Data controllers can find significant value in the capability to comply with legal obligations to perform such processing while enhancing the value of identifying protected data under their control by relinking the results of privacy-preserving processing. This is made much simpler by leveraging BigPrivacy-enabled ***pseudonymisation****,* ***functional separation*** *and* ***data minimisation****.*

IV. COMPLIANCE BENEFITS OF BIGPRIVACY – DATA SAFE HAVENS[45]

Many organisations believe that the GDPR requires "pure play" anonymisation, where the linkability and full context of data is lost in the process, in order to lawfully process Big Data. Anonos has spent six years exploring, in depth, the GDPR and predecessor EU data protection laws to understand "Data Safe Havens" – **explicitly recognized combinations of GDPR legal & technical safeguards that maximise Big Data value by leveraging GDPR compliant pseudonymisation** – that do not require the loss of linkability and full context of data.

Anonos' patented BigPrivacy technology helps to support the below enumerated GDPR Data Safe Havens to allow organisations to comply with the GDPR while retaining far greater leverage from Big Data analytics, AI, ML and DX. This is what differentiates Anonos BigPrivacy and delivers substantial value to organisations capitalizing on this advantage.

SAFE HAVEN #1: LEGACY CONSENT DATA TRANSFORMATION - AVOID DELETION: SAVEYOURDATA®

BigPrivacy SaveYourData software can be used by data controllers to exercise their "one off" opportunity to transform data collected using (now) non-compliant broad-based consent ("Legacy Consent Data") to a state that supports Legitimate Interest processing as an alternate (non-consent) legal basis **to avoid:**

i. **Having to delete valuable data;**

ii. **The risk of injunctions ordering immediate suspension of data processing; and**

iii. **Exposure to significant fines.**

Many organisations historically relied on general broad-based consent as their lawful basis for processing EU personal data, but such Legacy Consent Data is no longer legal to possess (in *either* encrypted or unencrypted format) or process under the GDPR. The GDPR has no "grandfather provision" or "exemption" that allows for ongoing storage or use of (now) illegal Legacy Consent Data, thereby exposing organisations to injunctions ordering the immediate suspension of data processing and substantial fines for failure to delete illegal data.[46] The GDPR-certified pseudonymisation capabilities of BigPrivacy SaveYourData software transform data to help support Legitimate Interest as an alternate (non-consent) legal basis.[47]

In late October 2018, IDC published the report "**Anonos' SaveYourData - a EuroPrivacy Certified Solution - "Deep Freezes" Enterprises' Existing Personal Data Sets as They Plan Analytics Strategies**"[48] as a Public Service Announcement so that data-driven organisations can avoid the catastrophic result of deleting valuable information that is crucial for Big Data insights and analytics. By Implementing Anonos SaveYourData software, organisations can transform Legacy Consent Data that would otherwise be deleted.

The following language from the IDC report summarizes the situation:

> *The European Data Protection Board (EDPB) endorsed the Article 29 Working Party requirements for consent under Regulation 2016/679 (WP259 rev.01) that specify the requirements for and limitations of using consent as a legal basis for processing EU personal data under the GDPR. [The] Consent Requirements acknowledge that the GDPR changed the definition of consent and that all data — collected before and after the GDPR — that fail to meet new strict GDPR consent requirements for specificity and unambiguity are no longer legal to "process" — which term under GDPR Article 4(2) includes mere storage of data. The GDPR has no "grandfather provision" or exemptions allowing for continued use of data collected using (now) illegal non-compliant consent. Storing or processing this data exposes organisations to regulator injunctions blocking access and use of data in addition to significant penalties under the GDPR. It is unclear how long organisations will have to exercise their one-off opportunity to transition data to support an alternate (non-consent) legal basis as outlined in the Consent Requirements. Data protection authorities (DPAs) are looking for proof that organisations have taken good-faith steps to comply with the GDPR.*
>
> *Due to the uncertain time-sensitive nature of the one-off right to transform data that is otherwise illegal under the GDPR into a new legal format under [the] Consent Requirements, organisations should evaluate their options immediately and take appropriate action.*
>
> *As a result of these requirements, many organisations have simply resorted to either*
>
> - *Adopting blanket data encryption that renders meaningful analytics impossible and **does not address the potential unlawfulness of storing the data**; or*
> - *Deleting data to comply with GDPR requirements because searching, identifying, and classifying personal information and then applying for reconsent is a labour- and resource intensive task.*
>
> *A top 5 global hospitality firm revealed to IDC at a workshop that it had deleted 20 years of loyalty data because of concerns over legality of data processing under GDPR. Those that have the resources and an appetite for consent-based processing are finding the reconsenting process lengthy and depletory to their business data. For instance, a top European financial services organisation said to IDC said that it sought to obtain consent from its customers and obtained only a 60% success rate.*
>
> *In IDC's opinion, these strategies of data deletion or <100% consent results in wasted opportunities. **IDC believes organisations need to balance their risks and opportunities and adopt GDPR-compliant pseudonymisation technologies. This data has huge value and can provide the business with a competitive advantage.***

GDPR-friendly pseudonymisation for data processing is applicable in the following scenarios:

- *Consent is not practical, or may undermine the business;*
- *Statistical analysis to identify broad trends or general conclusions are insufficient; and*
- *Retention of personal data under strict policies for compliance with industry-specific regulations such as healthcare or banking data regulations.*

Data enablement/security start-up Anonos collaborated with storage vendor Hitachi Vantara to launch a solution called SaveYourData to create a legal and technical foundation for legitimate interest processing using privacy-compliant pseudonymisation. The solution offers a deep-freeze state for existing personal data repositories without violating GDPR principles. ***This helps organisations avoid data deletion, blanket encryption, or reconsent exercises.***

SaveYourData provides a means to safely and legally save existing personally identifiable data putting it in deep-freeze — while enterprises implement solutions to address analytics processing issues to comply with GDPR.

The dilemma for data controllers is how to retain valuable Legacy Consent Data when it plays a crucial role in the controller's digital transformation program and data-centric projects like Big Data analytics, AI, ML and DX. **Under the GDPR, a controller can transform data to a new legal basis of Legitimate Interest using SaveYourData GDPR-certified pseudonymisation software as the first step in legally continuing with its data-driven journey.** Further action will be necessary to make use of the data in compliance with Data Safe Havens to maximise the full value of Big Data, but the data controller will have more time to arrange suitable processing and will not be forced to delete valuable data.

Anonos SaveYourData software has been certified under the EuroPrivacy certification scheme as complying with European GDPR requirements. The certification by EuroPrivacy, completed using the "Privacy Flag" certification scheme developed under a research project co-funded by the European Commission and Switzerland, highlights that SaveYourData software meets GDPR requirements for pseudonymisation.[49]

SAFE HAVEN #2: LEGITIMATE INTEREST LEGAL BASIS FOR BIG DATA

Significant questions as to the legality of consent as a valid basis under the GDPR for lawful Big Data analytics, AI, ML and DX has stalled numerous projects affecting the business value and data intelligence extracted from these projects. **Anonos BigPrivacy technology leverages new regulatory requirements as a competitive advantage to balance the increasing demands of data intelligence by leveraging GDPR-certified pseudonymisation capabilities.** Organisations can overcome limitations of consent by using GDPR-compliant pseudonymisation to enable Legitimate Interest processing to support processing:

i. **That cannot be described with required specificity at the time of initial data collection.**

ii. **To avoid having to request re-consent each time a different processing of data is desired.**

In order for iterative Big Data processing to be legal under the GDPR (and evolving "GDPR-like" data protection laws), technology and organisational safeguards are required that support the requirements for Legitimate Interest processing – NOT just in words – but by supporting "dynamism" necessary to satisfy the Legitimate Interest "Balancing of Interest" test required for a valid legal basis. Anonos BigPrivacy is the only technology solution capable of supporting the dynamism required for Legitimate Interest processing and has six granted foundational patents on this capability.

One TechCrunch article[50] included statements by the EU General Data Protection Supervisor that current consent practices constitute "blackmail" and are illegal under the GDPR. In the specific context of "adtech," another TechCrunch article[51] stated "**So if the consent rug it's been squatting on for years suddenly gets ripped out from underneath it, there would need to be radical reshaping of ad-targeting practices to avoid trampling on EU citizens' fundamental rights**."

However, lawful Legitimate Interest processing requires more than mere words and "cannot be equated to the interest of companies to make a profit from our personal data" as made clear in a recent case filed against Acxiom, Oracle, Equifax and Experian.[52] **This court case makes it clear that to serve as a valid legal basis, Legitimate Interest processing must satisfy a three-part test.** The first two tests are relatively easy to satisfy however the third test requires technical and organisational safeguards. The three tests are:

1. **Legitimate Interest** test;

2. **Necessity** test; and

3. **Balancing of Interest** test which requires technical and organisational safeguards that balance the interest of the data controller (or third party) against individual data subjects' rights and freedoms.

Without technical and organisational safeguards that satisfy the Balancing of Interest test, many data processing activities that were commonly practiced for decades are no longer legal.

"Consent" as now defined under the GDPR requires specificity that is impossible to satisfy for iterative Big Data processing, details of which not known at the time of initial data collection. It is not possible to secure legally binding data subject consent for processing activities in the future that are not capable of being described with specificity at the time of data collection. This makes it impossible to rely on "consent" as a legal basis for "Big Data" defined for this purpose as:

- Processing or mere possession (whether in encrypted or decrypted form) of historical data previously collected using "general broad-based consent" that is now illegal, and which exposes organizations to the risk of immediate injunction ordering termination of processing/possession plus GDPR penalties and lawsuits;

- Repurposing of data beyond the original purpose for data collection;

- Sharing of data with external parties; and

- Combining a protected data with other data sets.

The critical first step of the patented BigPrivacy process is to support establishment of a new legal basis of Legitimate Interest so that the creation, use and sharing of standardised BigPrivacy Variant Twin data assets is lawful. All other solutions require a data controller to "Bring Your Own Basis" (BYOB) – necessitating that a valid legal basis already exists which is no longer the case for Big Data using data subject consent. BYOB solutions fail to solve the lack of valid legal basis for (a) processing data collected in the past and (b) performing iterative Big Data processing, both of which now require a new legal basis under the GDPR. As noted previously, without satisfying the Balancing of Interest test, many data processing activities that were legal for decades will no longer be legal to perform.

BigPrivacy's GDPR-certified compliant pseudonymisation technology uniquely supports Legitimate Interest processing as a new (non-consent) legal basis in numerous ways, including:

1. Patented dynamic de-identification functionality that separates information value from identity to defeat unauthorised re-identification between data sets via the Mosaic Effect; and

2. Patented Variant Twin data that can be sourced, curated, combined, shared and processed on-premise and in the cloud in compliance with applicable laws.

Without dynamism, BYOB solutions cannot defeat unauthorised re-identification via the Mosaic Effect. As a result, they cannot satisfy the "Balancing of Interest" test required to support a Legitimate Interest legal basis under the GDPR. **BigPrivacy is the only solution that supports dynamism as necessary to enable new GDPR compliant Legitimate Interest processing**. Anonos holds six granted patents and has 60+ additional patents pending, on the use of dynamism to support privacy-respectful, legally compliant data use without fear of reprisal from jurisdictional regulators for unlawful processing.

Benefits of BigPrivacy-enabled Legitimate Interest processing under the GDPR include the following:

- **Right to Be Forgotten**: If a data controller uses compliant Legitimate Interest processing, under GDPR Article 17(1)(c), so long as they can show they "have overriding legitimate grounds for processing" that are supported using adequate technical and organizational measures to satisfy the Balancing of Interest test to protect the rights of data subjects, they do not have an obligation to comply with Right To Be Forgotten (RTBF) data erasure requests.

- **Right to Restrict Processing**: A data controller processing data using Legitimate Interest processing, under GDPR Article 18(1)(d), does not have an obligation to comply with claims to restrict processing so long as they can show they have technical and organizational measures in place that satisfy the Balancing of Interest test to protect the rights of data subjects such that the legitimate interests of the data controller properly override those of data subjects.

- **Right to Data Portability**: Under GDPR Article 20(1), data controllers using Legitimate Interest processing are not subject to the right of portability which applies to processing based on consent.

- **Right to Object**: Data subjects do not have the right to object to processing under GDPR Article 21(1) when a data controller using Legitimate Interest processing can prove they have technical and organizational measures in place that satisfy the Balancing of Interest test to protect the rights of data subjects such that

the legitimate interests of the data controller properly override those of data subjects. However, data subjects always have the right under Article 21(3) to not receive direct marketing outreach resulting from processing of the data.

Under the GDPR, data controllers are not allowed to retrospectively swap from one legal basis for processing personal data to another.[53] Data subjects must be **informed at the time of initial data collection** of each lawful basis upon which the controller relies for each separate process.

- **Example**: Bank A collects data and puts customers on notice at the time of initial data collection that the Bank A relies on Article 6(1)(a) Consent. Bank A may lawfully process the data as necessary for primary purposes like completing credit and debit transactions. However, Bank A may not lawfully process the data for iterative analytics like determining "Next Best Action" (e.g., cross sell or upsell opportunities) or other internal or external repurposing like combing data to discover new insights and correlations in reliance on any legal basis not disclosed at the time of initial collection.

Anonos state-of-the-art GDPR-certified pseudonymisation helps to supports Legitimate Interest processing of Big Data analytics, AI, ML and DX. Other GDPR legal bases do not provide support for Big Data analytics, AI and ML.

- **Example**: Bank B collects data and puts the customer on notice at the time of initial data collection that the Bank relies on Article 6(1)(a) Consent and Article 6(1)(f) Legitimate Interest processing to perform statistical analysis to improve product and service offerings and to enhance user experience by leveraging BigPrivacy GDPR-certified pseudonymisation capabilities to ensure that the data controller's interests are balanced with the data subjects' interests by enforcing safeguards to limit undue impact on the data subjects by supporting data minimisation and Privacy-Enhancing Techniques (PETs). Bank B may lawfully process the data as necessary to complete primary purposes like completing credit and debit transactions and may lawfully process the data for Big Data analytics, AI, ML and DX consistent with the notice provided to customers. See below for shortcomings of other legal bases for supporting lawful Big Data analytics, AI, ML and DX.

Legal Basis Analysis

As noted above, Data controllers are not allowed to retrospectively swap from one legal basis to another. Data subjects must be **informed at the time of initial data collection** of each lawful basis upon which the controller relies for each separate process.[54] The following considerations are relevant in connection with internal or external repurposing of data for Big Data analytics, AI, ML and DX under the six legal bases established under GDPR Articles 6(1)(a)-(f):

a) **Consent** – "Bundling" of consent to get approval for Big Data analytics, AI, ML and DX within the acceptance of terms and conditions, or "tying" the provision of a contract or a service to consent for such processing, is not lawful because consent is not freely given.[55] Data subjects can only lawfully consent to data uses that are specifically and unambiguously explained at the time of consent.[56] This requirement significantly reduces (some authorities even say prohibits) the ability of data controllers to rely on consent as a legal basis for iterative Big Data analytics, AI, ML and DX since these activities cannot be explained with sufficient detail or clarity at the time of consent. A recent paper by Southampton University (UK) professors entitled "Data Analytics and the GDPR: Friends or Foes? A Call for a Dynamic Approach to Data Protection Law"[57] noted that "…if a controller processes

data based on consent and wishes to process the data for a new purpose, the controller needs to seek a new consent from the data subject for the new processing purpose. **Yet, one could argue, consent is doomed in a data analytics context because the purpose simply cannot be specific…many organisations will find obtaining GDPR-satisfactory (i.e. 'informed') consent for innovative reuse from data subjects impractical."**

b) **Necessary for Contract** – The WP29 *Opinion on the Notion of Legitimate Interests of the Data Controller* and *Guidelines on Automated Individual Decision-Making and Profiling* stipulates that the lawful basis of "necessary for the performance of a contract" is to be interpreted strictly to cover only the minimum data required for contract performance. For these reasons, this legal basis does not generally support lawful Big Data analytics, AI, ML and DX.[58]

c) **Compliance with Legal Obligation of Controller** – This legal basis is "strictly delimited" to compliance with obligations imposed by the laws of the EU or a Member State. Obligations under the laws of third countries are not included in this legal basis. To be valid, a legal obligation of a third country must be officially recognised and integrated in the legal order of the Member State. On the other hand, the need to comply with a foreign obligation may represent a legitimate interest of the controller, subject to the balancing test of Article 6(1)(f). For these reasons, this legal basis does not generally support lawful Big Data analytics, AI, ML and DX.[59]

- **Example**: Bank C may lawfully rely on Article 6(1)(c) to process data for criminal prevention purposes required under EU or Member State laws, like Anti-Money Laundering (AML) and Know Your Customer (KYC) requirements. However, Bank C may not rely on Article 6(1)(c) to process iterative analytics like determining Next Best Action (e.g., cross sell or upsell opportunities) or other internal or external repurposing like combing data to discover new insights and correlations from Big Data analytics, AI, ML and DX. A separate legal basis is required to support lawful Big Data analytics, AI, ML and DX.

- **Vital Interest of the Data Subject or Other Natural Person** – This legal basis applies in situations of life and death, or at the very least, threats that pose a risk of injury or other damage to the health of the data subject. This legal basis does not generally support lawful repurposing of data for Big Data analytics, AI, ML and DX.[60]

- **Task Carried Out in the Public Interest** – This legal basis applies only to processing that is necessary for the performance of a task carried out in the public interest of the EU or of a Member State. This legal basis does not generally support lawful repurposing of data for Big Data analytics, AI, ML and DX carried out for commercial interests.[61]

- **Legitimate Interest** – As more fully described below, GDPR-compliant pseudonymisation helps to support Legitimate Interest processing not based on consent[62] to enable further processing,[63] data minimisation[64] and archiving for statistical purposes[65] by serving as a technical and organisational safeguard that helps to achieve functional separation between analysis and identifying data.[66] Legitimate Interest uniquely helps to support lawful repurposing of data for Big Data analytics, AI, ML and DX.

Disclosure of Legitimate Interest Processing Upon Data Collection

When using Legitimate Interest processing:

- A data controller should put data subjects on notice at the time of initial data collection that: (i) it relies on Legitimate Interest processing to perform statistical analysis to improve product and service offerings and to enhance user experience; (ii) state-of-the-art GDPR compliant pseudonymisation is used to support Legitimate Interest processing to ensure that the data controller's interests are balanced with the data subjects' interests by enforcing safeguards to limit undue impact on the data subject by supporting data minimisation and Privacy-Enhancing Techniques (PETs); and (iii) data subjects have the unconditional right to opt-out of any direct marketing resulting from Legitimate Interest processing-enabled statistical analysis.[67]

- The data controller should also document the results of the three-part test for Legitimate Interest outlined below as evidence of its assessment for greater accountability. [68]

GDPR-Certified Pseudonymisation & Legitimate Interest Processing

Anonos state-of-the-art BigPrivacy data protection leverages GDPR-certified pseudonymisation to support dynamic privacy-respectful and technically enforced processing to defeat the "Mosaic Effect" – i.e., the unauthorized re-identification of personal data by correlating static (or persistent) tokens used in traditional Privacy Enhancing Techniques (PETs) to replace different occurrences of the same data element. Traditional static tokenization (or key coding) techniques fail to satisfy strict GDPR definitional requirements for pseudonymisation, since merely by having access to static (or persistent) tokens it may be possible to enable attribution of personal data to a specific data subject without requiring the use of "additional information kept separately and subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person" as required under the GDPR Article 4(5) definition of pseudonymisation.

At this point in time, the challenge with Legitimate Interest processing is the lack of a well-established body of case law or specific guidance on what is required to support Legitimate Interest processing. One thing that a data controller can do to significantly buttress its support for lawful Legitimate Interest processing, and to avoid both fines and injunctions ordering suspension of data processing, is to implement state-of-the-art safeguards like GDPR-compliant pseudonymisation to balance the risks. By leveraging GDPR-compliant pseudonymisation, a data controller can demonstrate that it has implemented substantial steps to meet its obligation to ensure it has considered the rights and interests of data subjects so that its data processing is being conducted in a privacy respectful manner.

As noted previously, an established requirement for relying on Legitimate Interest processing is the application of the following three-part test to show that data subjects' rights and interests are considered and protected:

1. **Purpose Test:** Are you pursuing a legitimate interest?

2. **Necessity Test:** Is the processing necessary for that purpose?

3. **Balancing of Interests Test:** Are technical and organisational safeguards in place so that data subjects' interests do not override the legitimate interest of the data controller or third party?

In evaluating technical and organisation safeguards,[69] WP29 has stated that pseudonymisation is an effective safeguard that can be used to demonstrate that a data controller has taken steps to minimize the impact on data subjects under the Balancing of Interests Test for Legitimate Interest processing and can "play a role in tipping the balance in favour of the controller." However, the controller must put data subjects on notice prior to processing data that the controller is relying on Legitimate Interest and explain what its interests are and what measures have been put in place to safeguard the rights of data subjects. If a data controller wants to process the personal data for a new purpose, it may continue processing under Legitimate Interest so long as the new purpose is compatible with the original purpose.[70]

SAFE HAVEN #3: LEGAL SECONDARY PROCESSING OF BIG DATA

Anonos BigPrivacy helps enable organisations to ensure secondary processing is "compatible" with the original primary purpose through pseudonymisation and functional separation of the data. Under the GDPR, when an organisation processes personal data obtained for a particular permitted purpose, it cannot process it further except for compatible purposes.[71] However, "further processing" of personal data may be deemed compatible if it satisfies the requirements of (i) **Article 6(4) with respect to further "processing for a purpose other than that for which the personal data have been collected…not based on the data subject's consent"** and/or (ii) **Article 89(1) with respect to processing conducted for** "archiving purposes in the public interest," "scientific or historical research purposes," or "**statistical purposes.**" The GDPR highlights pseudonymisation and functional separation as additional safeguards to help ensure that such further processing is lawful.[72] BigPrivacy's certified and patented pseudonymisation and functional separation capabilities help to support compatible processing in accordance with the GDPR.

A fundamental principle of EU data protection law is purpose limitation (which must be complied with, as well as requiring that there be a valid legal basis, etc.). Under GDPR Article 5(1)b) personal data must be "collected for specified, explicit and legitimate purposes, and not further processed in a manner that is incompatible with those purposes."[73] The analysis in WP29 Opinion 03/2013 on purpose limitation ("WP29 Opinion 03/2013")[74] regarding "further processing" and compatible (or incompatible) purposes under the Directive remains relevant under the GDPR. This opinion identified certain key factors in assessing the compatibility of further processing purposes. One such key factor for consideration is the set of safeguards applied by the controller to ensure fair processing to prevent undue impact on data subjects.

The WP29 Opinion 03/2013 provided that, while all relevant factors must be assessed as a whole, "appropriate additional measures could thus, in principle, serve as 'compensation' for a change of purpose or for the fact that the purposes have not been specified as clearly in the beginning as they should have been. This might require technical and/or organisational measures to ensure functional separation (such as partial or full anonymisation, **pseudonymisation**, and aggregation of data) …." It further stated, "When trying to identify technical and organisational measures that qualify as appropriate safeguards to compensate for the change of purpose, the focus often lies with the notion of isolation. Examples of the relevant measures may include, among other things, full or partial anonymisation, **pseudonymisation**, or aggregation of the data, privacy enhancing technologies, as well as other measures to ensure that the data cannot be used to take decisions or other actions with respect to individuals (**'functional separation'**). These measures are particularly relevant in the context of further use for 'historical, **statistical** or scientific **purposes'**…." **More recently, the European Data Protection Supervisor has encouraged innovative engineering solutions "related to the concept of 'functional separation'**.[75]

Further processing for "archiving purposes in the public interest," "scientific or historical research purposes" or "statistical purposes" is specifically considered not to be incompatible with the initial purposes so long as appropriate safeguards for data subjects are provided to ensure, in particular, data minimisation; measures in that regard may include pseudonymisation.[76] **Pseudonymisation is also an explicitly-recognized safeguard under Article 6(4)(e) to help ensure that any such further processing of personal data "[are] compatible with the purpose for which the personal data are initially collected" in compliance with Article 5(1)(b) ('purpose limitation') requirements.**

BigPrivacy enables GDPR-certified pseudonymisation to help ensure that "further processing" of personal data is a compatible use. BigPrivacy also facilitates functional separation of data. Functional Separation of data involves using technical and organisational measures to ensure that data used for one purpose cannot then be used to "support measures or decisions" with regard to individuals concerned unless specifically authorized by the individuals.[77] In addition to being identified in WP29 Opinion 03/2013, Functional Separation is also recognized in WP29 Opinion 06/2014 on the *Notion of Legitimate Interests of the Data Controller* under Article 7 of Directive 95/46/EC as supporting Legitimate Interest," and in EDPS Opinion 7/2015 on *Meeting the Challenges of Big Data* as playing "a role in reducing the impact on the rights of individuals, while at the same time allowing organisations to take advantage of secondary uses of data."

GDPR Article 5(1)(e) permits archival and use of data for "statistical purposes" if appropriate technical and organisational safeguards outlined in Article 89(1), which specifically include pseudonymisation (such as enabled using BigPrivacy certified and patented technology), are used to help protect the rights and freedoms of data subjects. In this regard, Anonos' certified pseudonymisation capabilities represent a method for safeguarding the rights and freedoms of data subjects by controlling the protection, removal, and restoration of linkages between and among data sets and identifiers for different data processing activities under both Articles 6(4) and 89(1).

If organisations cannot process data for compatible secondary processing and "statistical purposes" to perform predictive analytics, then huge potential benefits are lost for both data subjects and data controllers. Southampton University (UK) professors highlight the benefits of "**a more constructive interpretation of the GDPR…on the basis of a dynamic approach to data protection law**" that distinguishes between three different **"…compliance stages (data collection, data analytics, individual impact)…."**[78] Adopting this three-stage perspective:

1. **Data Collection Stage:** BigPrivacy technology helps support Legitimate Interest-based data collection;

2. **Data Analytics Stage:** BigPrivacy supports creation of non-identifying, dynamically de-identified derivative versions of original data that Anonos refers to as "Variant Twins" for analysis;[79] and

3. **Individual Impact Stage:** after the data controller has gathered, normalized, and analysed non-identifying Variant Twin data "in a way that equally respects their marketing interests and the privacy of users at large,"[80] then the de-identification rules may be reversed under privacy-respectful controlled conditions to enable outreach to applicable customers based on Legitimate Interest or Consent.

    - **Example**: Bank D collects personal data from customers and puts the customers on notice at the time of initial data collection that Bank D relies on Article 6(1)(f) Legitimate Interest processing to perform statistical analysis to improve product and service offerings for customers and to enhance user experience by leveraging GDPR-certified pseudonymisation capabilities to ensure that the data controller's interests are balanced with the data subjects' interests by enforcing safeguards that limit undue impact on the data subjects by supporting

data minimisation and Privacy-Enhancing Techniques (PETs). The Bank desires to use data to improve product and service offerings and to enhance user experience, including data from ex-customers of the Bank. The Bank undertakes the following analysis under GDPR Article 6(4) to evaluate the lawfulness of using the data collected to ensure that use of such personal data is compatible for the purpose for which the data were initially collected:

Article 6(4) Further Processing Analysis:

(a) any link between the purposes for which the personal data have been collected and the purposes of the intended further processing;

There is a direct link between the original use of the data from active customers and the intended further use of the data from past customers to improve Bank product and service offerings and enhance user experience for Bank customers overall.

(b) the context in which the personal data have been collected, in particular regarding the relationship between data subjects and the controller;

The data was initially collected in the context of a Bank-customer relationship which is compatible with the intended further use of data from past customers to improve Bank product and service offerings and enhance user experience for Bank customers overall.

(c) the nature of the personal data, in particular whether special categories of personal data are processed, pursuant to Article 9, or whether personal data related to criminal convictions and offences are processed, pursuant to Article 10;

The nature of the personal data in question does not fit within any of the special categories of data.

(d) the possible consequences of the intended further processing for data subjects;

The intended further use of the data from past customers to improve Bank product and service offerings and enhance user experience for Bank customers overall will not have adverse impacts on the data subjects who were prior customers of the bank.

(e) the existence of appropriate safeguards, which may include encryption or pseudonymisation.

The Bank will leverage BigPrivacy GDPR-certified pseudonymisation capabilities to ensure that the data controller's interests are balanced with the data subjects' interests by enforcing safeguards to limit undue impact on the data subjects by supporting data minimisation and Privacy-Enhancing Techniques (PETs).

Assuming the Bank properly documents the forgoing analysis under Article 6(4), it is reasonable to conclude that the desired further processing is compatible with the purpose for which the personal data were initially collected.

WP29 Profiling Guidelines

The following comments are cross-referenced to page numbers within the WP29 *Guidelines on Automated Individual decision-making and Profiling* ("WP29 Profiling Guidelines" or "Guidance.")[81] The WP29 does not have actual rule making authority and some parties have argued that the WP29 Profiling Guidelines are overly conservative.[82] In the context of profiling and automated decision making, some commentators believe that the WP29 went too far in stating that "A general prohibition on this type of processing exists to reflect the potentially adverse effect on individuals" at the bottom of page 8 of the Guidance. For a strong argument along these lines, see "Did the WP29 Misinterpret the GDPR on Automated Decision-Making."[83]

Page 5 of the Guidance specifically acknowledges that "Profiling and automated decision-making can be useful for individuals and organisations as well as for the economy and society as a whole, delivering benefits such as… increased efficiencies… [and] resource savings" and that they "have many commercial applications, for example, they can be used to better segment markets and tailor services and products to align with individual needs."

It is important to note that "profiling" and "solely automated processing" are two separate concepts under the GDPR. concept of Profiling defined in GDPR Article 4(4) as:

> *"…any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements."*

While profiling by definition will involve some form of "automated processing," it does not have to involve "solely automated processing." GDPR Article 22 makes significant distinctions (with significant differences in legal rights, responsibilities and obligations) between "profiling" that involves automated processing on the one hand and "solely automated processing" on the other hand.

On page 8, the Guidance identifies three potential ways in which profiling may be used:

i. general profiling;

ii. decision-making based on profiling; and

iii. *solely* automated decision-making, including profiling (Article 22).

It is important to note that "solely" is actually italicized in (iii) above in the Guidance. While BigPrivacy can facilitate profiling involving "solely automated decision-making" as described in (iii) above, many clients use BigPrivacy for "general profiling" and "decision-making based on profiling" as described in (i) and (ii) above in lieu of solely automated decision-making, including profiling as described in (iii) above.

It is also important to note that BigPrivacy's patented dynamic pseudonymisation capabilities support the ability to use Legitimate Interest as a valid – and much more flexible – (non-consent) lawful basis for processing personal data. Page 21 of the Guidance specifically discusses the availability of Article 6(1)(f) Legitimate Interest as an available legal basis for profiling and decision-making based on profiling (that is not solely automated decision-making). Also, it is worth noting that page 30 of the Guidance **specifically enumerates pseudonymisation as a "good practice" suggestion to consider when profiling.**

Please reference the discussion in Data Safe Haven #5 – Non-Identifying Big Data Exclusion from Data Subject Rights – below for information regarding the importance of providing customers the means to "opt out" from receiving direct marketing that results from Variant Twin profiling analysis.

SAFE HAVEN #4: DATA MINIMISATION; DATA PROTECTION BY DESIGN AND BY DEFAULT IN SUPPORT OF BIG DATA

Another principle that is continued and made even more restrictive under the GDPR is "data minimisation."[84] The GDPR further imposes new requirements for Data Protection by Design and by Default Data[85] which means organisations must integrate or 'bake in' significant data protection capabilities into processing practices, from the design stage right through the full data lifecycle. Previously known as 'privacy by design', this concept has always been part of data protection law, however, **two key changes which are newly mandated under the GDPR are:**

i. **It is now a legal mandate to support more than just privacy by design – Data Protection by Design and by Default requires the most stringent implementation of privacy by design; and**

ii. **It has new heightened requirements, including the need to support the GDPR principles of data minimisation and purpose limitation to limit data use to the minimum extent and time necessary to support each specific product or service authorized by a data subject.**

The obligation to support Data Protection by Design and by Default as newly defined under the GDPR obligates each organisation "to be clear in advance about what its plans for secondary processing of personal data intends to achieve…[including] the upfront design of data processing to demonstrate that this thinking has taken place and to ensure safeguards measures can be implemented to mitigate any notable risk areas identified…. data minimisation should be engineered relative to purposes before the start of processing, at the time of the determination of the means." [86] **This essentially means that less, rather than more, personal data must be provided, used or disclosed or otherwise processed for a given purpose.** How much less? Only the minimum amount needed to achieve the authorized purpose. BigPrivacy supports both pseudonymisation, which dynamically masks the actual personal data, and data minimisation, which enforces fine-grained access controls using Controlled Linkable Data,[87] to enable the disclosure of only "minimum identifying data" to those with a need-to-know, all on a case-by-case basis.

While the focus of data minimisation has usually been on minimising the amount of personal data collected at the acquisition stage, the data minimisation requirement also applies to the post-collection use or other processing of personal data. Thus, the GDPR states, "In order to create incentives to apply pseudonymisation when processing personal data, measures of pseudonymisation should, whilst allowing general analysis, be possible within the same controller when that controller has taken technical and organisational measures necessary to ensure, for the processing concerned, that this Regulation is implemented, and that additional information for attributing the personal data to a specific data subject is kept separately. The controller processing the personal data should indicate the authorised persons within the same controller." [88]

BigPrivacy helps to enforce data minimisation and Controlled Linkable Data by technologically enabling the use of the minimum level of linkable (identifiable) data necessary for each authorised process, to protect personal data on a per-authorised-use basis by technically controlling the linkability of data to limit access to linkable (identifying) data and protect against unauthorised use. Accordingly, BigPrivacy helps support data minimisation within an organisation by enforcing selective access to data, ensuring that an individual employee only has access to the data

required for them to do their job and no more. And, when personal data is shared between organisations, BigPrivacy can enforce selective access controls to ensure that data is shared only as authorised. In summary, BigPrivacy helps to ensure that only discrete data elements are made available to support minimal use.

Article 25(1) stipulates that, taking into account the risks to data subjects and certain other factors, controllers must, "both at the time of the determination of the means for processing and at the time of the processing itself, implement appropriate technical and organisational measures, **such as pseudonymisation**, which are designed to implement data-protection principles, such as data minimisation, in an effective manner and to integrate the necessary safeguards into the processing in order to meet the requirements of this Regulation and protect the rights of data subjects." It can be seen that this provision specifically notes that pseudonymisation, such as made possible by BigPrivacy, is an appropriate measure to help implement Data Protection by Design and by Default.

Article 25(2) requires controllers to implement "appropriate technical and organisational measures for ensuring that, by default, only personal data which are necessary for each specific purpose of the processing are actually processed. That obligation applies to the amount of personal data collected, the extent of their processing, the period of their storage and their accessibility. In particular, such measures shall ensure that by default personal data are not made accessible without the individual's intervention to an indefinite number of natural persons." BigPrivacy's support for selective access controls and Controlled Linkable Data uniquely enables organisations to limit accessibility to personal data in a fine-grained, risk-appropriate fashion, and to help comply with Article 25(2). Indeed, in urging internal policies and measures to meet the principles of Data Protection by Design and by Default, Recital 78 of the GDPR notes that such measures could consist, inter alia, of "minimizing the processing of personal data [and] **pseudonymising** personal data as soon as possible…."

BigPrivacy technology enforces fine-grained, dynamic control over both who has access to data and what level of resolution of identifying (or non-identifying) data is disclosed to a user for authorized purposes. BigPrivacy leverages the most current information about a user, data and environment at the time of disclosure to dynamically enforce the appropriate level of resolution for data, as if viewing the data through a "lens." The lower the magnification, the less identifying the disclosed data is (while still conveying necessary information value), whereas with higher magnification, the more "identifying" the disclosed data becomes. By leveraging BigPrivacy technology, only the minimum required level of identifying data is revealed for each authorized purpose. This unique resolution level of data disclosed via such "lens" is a Variant Twin of the original underlying data. **With BigPrivacy, productive data use can continue by disclosing Variant Twins of identifying data that convey the necessary information value to accomplish permitted processing of data in a privacy-respectful and non-identifying manner that enforces the principles of data minimisation and Data Protection by Design and by Default.**

SAFE HAVEN #5: EXPANDED BIG DATA USE AND SHARING OPPORTUNITIES

The GDPR clarifies and enhances the privacy rights of individual data subjects with new well-known rights such as the "right to be forgotten," the "right to data portability" and more. However, under GDPR Articles 11(2) and 12(2), if the purposes for which an organisation processes personal data do not or no longer require identification of an individual, and the organisation can show **that it is not in a position to identify the data subject**, then it is absolved from having to comply with these data subject rights. If personal data is pseudonymised using BigPrivacy GDPR-certified technology so that a given controller or processor cannot identify the individuals concerned, such organisations can avoid the application of these privacy provisions. **BigPrivacy helps to limit the risk of using personal data by data controllers and data processors down the "data chain" enabling the data to be used going forward in a "de-risked" manner which dramatically reduces the risk of data subjects being re-identified.**

By enforcing GDPR-certified pseudonymisation capabilities that dynamically transform identifying personal data into non-identifying Variant Twin data, the identity of data subjects may not be realized without the use of additional information, keys or a mapping table, which can be kept separate and secure.[89] **The re-linking of the pseudonymised data back to the identity of the data subject may therefore be limited to only authorised personnel within the original data controller only for authorized purposes.**[90] In the hands of data controllers and processors who do not have access to required additional information necessary to re-link to identity, data subjects cannot be re-identified by that organisation, so lesser legal obligations are owed to data subjects, enabling greater lawful use of secondary processing applications like Big Data analytics, AI, ML and DX.

Pseudonymised non-identifying Variant Twin versions of data processed under Legitimate Interest may become proprietary assets of an organisation, with respect to which there are no obligations to provide copies to competitors since the right of data portability under Article 20 does not apply**.**[91] And, so long as a controller can demonstrate "compelling legitimate grounds for processing which override the interests, rights and freedoms" of data subjects due to the use of state-of-the-art technical and organisational safeguards (or "for the establishment, exercise or defence of legal claims"), objections by data subjects under Article 21 to using Variant Twin data for Big Data analytics, AI, ML and DX Legitimate Interest processing will not be successful.[92] However, if a data controller relies on Legitimate Interest processing for direct marketing purposes, the right of data subjects to object to and not participate in such direct marketing processing must always be strictly enforced.[93] The data controller may continue with other (non-direct marketing) processes so long as it can show that its legitimate interests are compelling enough and that adequate safeguards are in place to protect the rights of data subjects.[94]

In addition, GDPR-compliant pseudonymisation can also impact data subjects' rights of access under Article 15, rectification under Article 16, and erasure ("right to be forgotten") under Article 17. Article 11 provides an exemption from these rights if "the controller is able to demonstrate that it is not in a position to identify the data subject." Since the GDPR does not require a controller to hold additional information "for the sole purpose of complying with this Regulation," a data controller may use pseudonymisation techniques and subsequently delete information that would enable the reversal of the pseudonymisation to identify individual data subjects.[95]

The GDPR expands individuals' rights in relation to their personal data, including, inter alia, the right to restrict processing under Article 18, and notification obligations to data recipients of any rectification, erasure or restriction of processing under Article 19. Furthermore, Article 22 provides that data subjects have the right not to be subject to automated individual decision-making, including profiling, which produces "legal effects" concerning them or similarly significantly affects them. However, under Articles 11(2) and 12(2), if the purposes for which an organisation processes personal data do not or no longer require identification of an individual, and the organisation can show that it is not in a position to identify the data subject, then the organization is absolved from complying with such rights.

SAFE HAVEN #6: COMPLIANT CLOUD PROCESSING FOR BIG DATA

Anonos BigPrivacy technology supports GDPR-certified pseudonymization, which can be used to dynamically transform personal data into non-identifying Variant Twin data, so that the identity of a data subject cannot be realized without the use of additional information, keys or a mapping table. BigPrivacy pseudonymisation-enabled Variant Twin data can put control over the re-linkability of data solely in the hands of the original data controller. Data controllers are able to "de-risk" personal data by sharing only Variant Twin data with third party "as-a-service" processors thereby retaining control over re-linking to identity and reducing the risk of unauthorized re-identification to reduce the complexity and burden of contractual negotiations and relationships.

Anonos BigPrivacy technology enables organisations to more effectively leverage the value of cloud processors providing "Infrastructure as a Service" (IaaS), "Platform as a Service" (PaaS), "Software as a Service" (SaaS), and other processing services while still protecting privacy and potential risks to data subjects in order to remain GDPR compliant. Prior to the GDPR, only data controllers had direct liability for noncompliance under EU data protection laws. That all changes under the GDPR, which for the first time introduces direct obligations, liability and exposure for data processors which **cannot** be negotiated away in contracts with data controller customers. Under the GDPR, all data processors providing IaaS, PaaS, SaaS and other data processing services involving EU personal data have direct obligations, liability and exposure under the GDPR.

Data controllers face liability for using cloud and other "as-a-service" vendors and other data processors that fail to provide "sufficient guarantees to implement appropriate technical and organisational measures in such a manner that processing will meet the requirements of this Regulation and ensure the protection of the rights of the data subject." As a result, data controllers must select cloud-based and other data processor partners that comply with the GDPR, or risk liability and exposure themselves. Lastly, data controllers and processors bear (effective) "joint and several liability" to compensate data subjects for their material and non-material (non-monetary losses like damage to reputation, emotional distress, pain and suffering, etc.) damage, even if other parties in the supply chain were more at fault, as the aim of the GDPR is to ensure data subjects are made whole for any loss or damage they suffer.[96]

GDPR Recital 28 specifically highlights the benefits of GDPR-compliant **pseudonymisation**, such as is enabled by BigPrivacy certified technology, including in the context of cloud-based and other "as-a-service" processor offerings, in its pronouncement that "the application of **pseudonymisation** to personal data can reduce the risks to the data subjects concerned and help controllers and processors to meet their data-protection obligations." Recital 78 goes even further by stating, "When developing, designing, selecting and using applications, services and products that are based on the processing of personal data or process personal data to fulfil their task, producers of the products, services and applications should be encouraged to take into account the right to data protection when developing and designing such products, services and applications and, with due regard to the state of the art, to make sure that controllers and processors are able to fulfil their data protection obligations."

V. CONCLUSION

BigPrivacy patented, GDPR certified, award-winning, state-of-the-art certified technology provides substantial competitive advantages in two ways:

First, BigPrivacy helps to harmonize the three perspectives of Business, Technology and Compliance, enabling GDPR-specific advantages to accrue to each.

Second, BigPrivacy harmonizes two objectives which had previously been in opposition: ***it delivers data value through data use, sharing and combination while simultaneously enabling GDPR-compliant data protection.***

Endnotes

1 M. Gary LaFever is CEO and Co-Founder at Anonos Inc. ("Anonos"), former law partner at the international law firm of Hogan Lovells, and former Management Information Consultant at the international advisory firm of Accenture.

2 See Article 29 Data Protection Working Party Guidelines on Consent under Regulation 2016/679 dated 10 April 2018, http://ec.europa.eu/newsroom/article29/document.cfm?action=display&doc_id=51030 at 3.

3 Id, at 23.

4 Id, at 31

5 Anonos has been actively engaged in research and development to advance the state-of-the-art in data protection, privacy and security technology since 2012. Anonos BigPrivacy and Variant Twin systems, methods and devices are covered by foundational granted patents (including, but not limited to, 10,043,035 issued in 2018; 9,619,669 issued in 2017; 9,361,481 issued in 2016; and 9,129,133; 9,087,216; and 9,087,215 issued in 2015) and a portfolio of over 60 pending domestic and international patent applications. BigPrivacy is a registered trademark of Anonos Inc. ("Anonos"). Additional protected trademarks of Anonos include, but are not limited to: Anonos; Anonos BigPrivacy Unlocks Big Data; Anonosizing; Anonosphere; Building a Layer of Trust with Transparency; Circle of Trust; CoT; Data-Privacy-as-a-Service; DDID; De-Risk; De-Risk Data. Discover Value; De-Risk Data to Maximize Value; Do Not Record; Do Not Remember; DRMD; DRM for the Individual; DRMI; Dynamic Anonymity; Dynamic Data Protection by Default; Dynamic De-Identifier; Have Your Cake and Eat It Too; In-Line Privacy; Just-In-Time-Identity; JITI; Privacy for The Interconnected World; Privacy Is the New Security; Privacy. Context. Control; Privacy Rights Management; PRM; Taking The "Personal" Out of Personal Data; There is No Undo in Privacy; Unlocking the Full Value of Data; Unlocking the Future Value of Data; Unlocking the Value of Data; and Variant Twin.

6 Data Protection by Design and by Default is defined in GDPR Article 25(1) and (2) as follows: "(1) Taking into account the state of the art, the cost of implementation and the nature, scope, context and purposes of processing as well as the risks of varying likelihood and severity for rights and freedoms of natural persons posed by the processing, the controller shall, both at the time of the determination of the means for processing and at the time of the processing itself, implement appropriate technical and organisational measures, such as pseudonymisation, which are designed to implement data-protection principles, such as data minimisation, in an effective manner and to integrate the necessary safeguards into the processing in order to meet the requirements of this Regulation and protect the rights of data subjects. (2) The controller shall implement appropriate technical and organisational measures for ensuring that, by default, only personal data which are necessary for each specific purpose of the processing are processed. That obligation applies to the amount of personal data collected, the extent of their processing, the period of their storage and their accessibility. In particular, such measures shall ensure that by default personal data are not made accessible without the individual's intervention to an indefinite number of natural persons."

Data Protection by Design and by Default and data minimisation are obligations imposed on data controllers (see Note 15, infra) under the GDPR. Data controllers are required to select software, hardware and processors (see Note 16, infra) that enable them to comply with their GDPR obligations. Data processors who enable controllers to comply with these obligations will have a competitive advantage and are more likely to be engaged by data controller customers than processors who do not assist controllers in complying with Data Protection by Design and by Default, data minimisation and other obligations under the GDPR.

7 See https://hbr.org/sponsored/2017/03/how-organisations-can-thrive-in-the-digital-economy

8 See https://www.idc.com/promo/thirdplatform/innovationaccelerators

9 See https://en.wikipedia.org/wiki/Infonomics

10 See https://www.itgovernance.co.uk/blog/gdpr-how-the-definition-of-personal-data-will-change/

11 See http://www.mondaq.com/ireland/x/636336/data+protection/ massively+expanded+enforcement+toolkit+and+potentially

12 See by analogy https://www.bna.com/kpmg-reaches-62m-n73014463153/

13 See http://www.wablegal.com/focus-gdpr-civil-claims/

14 See https://www.financierworldwide.com/roundtable-risks-facing-directors-officers-aug17/#.WhIeQktrzwd

15 https://www.ftc.gov/system/files/documents/reports/federal-trade-commission-staff-report-november- 2013-workshop-entitled-internet-things-privacy/150127iotrpt.pdf

16 See https://www.itif.org/publications/2015/01/27/ftc%E2%80%99s-internet-things-report-misses-mark

17 See data scientist expert determination for Anonos BigPrivacy technology attached as Appendix 1 to this blueprint.

18 See GDPR Recitals 26, 28, 29, 75, 78, 85, and 156 and Articles 4(5), 6(4)(e), 25(1), 32(1)(a), 40(2)(d), and 89(1).

19 GDPR Article 4(7) provides that "controller" means the natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of personal data; where the purposes and means of such processing are determined by Union or Member State law, the controller or the specific criteria for its nomination may be provided for by Union or Member State law.

20 GDPR Article 4(8) provides that "processor" means a natural or legal person, public authority, agency or other body which processes personal data on behalf of the controller. This can include third-party service providers such as hosting, cloud and analytics providers. For example, a vendor could offer processing services to its customers based on BigPrivacy technology.

21 The "Mosaic Effect" occurs when a person is indirectly identifiable due to a phenomenon referred to by the EU Article 29 Working Party as "unique combinations" where, notwithstanding the lack of identifiers that directly single out of a particular person, the person is still "identifiable" because that information may be combined with other pieces of information known to relate to the same individual (whether the latter is retained by the data controller or not), enabling the individual to be distinguished from others. See http://ec.europa.eu/justice/data-protection/article-29/documentation/opinion-recommendation/files/2007/wp136_en.pdf

22 Ted Myerson, Co-Founder of Anonos, presented a TED Talk on how BigPrivacy technology "stands DRM on its head." A video of, and the transcript for, this TED Talk is at https://anonos.com/TEDTalk - "TED Talk" is a trademark of Ted Conferences, LLC

23 See https://techcrunch.com/2006/08/09/first-person-identified-from-aol-data-thelma-arnold

24 See https://bits.blogs.nytimes.com/2010/03/12/netflix-cancels-contest-plans-and-settles-suit/

25 See http://dataprivacylab.org/projects/identifiability/paper1.pdf

26 See https://ec.europa.eu/programmes/horizon2020/

27 See https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3034261 at pg. 34.

28 See https://en.wikipedia.org/wiki/Data_binning

29 See discussion in Section IV regarding Safe Haven #3 for further discussion on functional separation.

30 See http://www.computerweekly.com/news/450299333/Most-European-firms-ill-equipped-for-insider- security-breaches

31 See https://en.oxforddictionaries.com/definition/dereference

32 The phrase "temporally unique" means the time of initial assignment of a dynamic de-identifier to a data subject, activity, process or trait is known, but the time period of assignment is of any duration, from limited to perpetual.

33 Data perturbation is a form of privacy-preserving data mining. See https://www.techopedia.com/definition/25013/data-perturbation

34 https://en.wikipedia.org/wiki/Star_Trek:_The_Original_Series

35 See Lindell and Pinkas, "Secure Multiparty Computation for Privacy Preserving Data Mining," The Journal of Privacy and Confidentiality athttp://repository.cmu.edu/cgi/viewcontent.cgi?article=1004&context=jpc

36 "k-anonymity" requires that each equivalence class (i.e., a set of records indistinguishable from each other with respect to certain "identifying" attributes) contains at least "k" records. See http://ukanon.net/wp-content/uploads/2015/05/The-Anonymisation-Decision-making-Framework.pdf

37 "l-diversity" requires that each equivalence class has at least "l" well-represented values for each sensitive attribute. See http://ukanon.net/wp-content/uploads/2015/05/The-Anonymisation-Decision-making- Framework.pdf

38 See https://en.wikipedia.org/wiki/Interoperability#Syntactic_interoperability

39 See https://en.wikipedia.org/wiki/Interoperability#Semantic_interoperability

40 See https://www.quora.com/What-are-the-limitations-of-Apple%E2%80%99s-%E2%80%9Cdifferential-privacy%E2%80%9D-approach-to-collecting-user-data

41 Ladjel Bellatreche and Sharma Chakravarthy. Big Data Analytics and Knowledge Discovery: 19th International Conference, DaWaK 2017, Lyon, France, August 28–31, 2017, Proceedings (Lecture Notes in Computer Science) (Kindle Location 1623). Springer International Publishing. Kindle Edition.

42 See https://fpf.org/2017/05/26/homomorphic-encryption-signals-future-socially-valuable-research-private-data/

43 See Note 41, *supra*, at 8025-8026.

44 Id, at 8055-8056. See also, Laécio, A & Costa, Laécio & Ruy, B & G B De Queiroz, J. (2014), The Use of Fully Homomorphic Encryption in Data Mining with Privacy Preserving at https://www.researchgate.net/publication/267507936_The_Use_of_Fully_Homomorphic_Encryption_in_Data_Mining_with_Privacy_Preserving

45 This blueprint represents general business advice; it is not, nor shall it be construed as, providing any legal opinion or conclusion, nor is this blueprint a substitute for obtaining professional advice from qualified legal counsel.

46 See WP29 Guidelines on Consent WP259 rev.01 at https://iapp.org/media/pdf/resource_center/20180416_Article29WPGuidelinesonConsent_publishpdf.pdf

47 See announcement of EuroPrivacy certification of SaveYourData software at https://www.prnewswire.com/news-releases/anonos-saveyourdata-software-officially-certified-by-europrivacy-meets-the-requirements-of-the-eu-general-data-protection-regulation-gdpr-300741945.html

48 The IDC report is available at www.anonos.com/DoNotDeleteYourData.

49 See Note 47, *supra*

50 See https://techcrunch.com/2018/10/03/europe-is-drawing-fresh-battle-lines-around-the-ethics-of-big-data/

51 See https://techcrunch.com/2018/11/20/how-a-small-french-privacy-ruling-could-remake-adtech-for-good/

52 See https://privacyinternational.org/advocacy-briefing/2434/why-weve-filed-complaints-against-companies-most-people-have-never-heard-and

53 See Note 46, *supra*, at 23

54 See GDPR Articles 13 and 14

55 See GDPR Recital 43 and Article 7(4). and quotes from Giovanni Buttarelli, European Data Protection Supervisor, at https://techcrunch.com/2018/10/03/europe-is-drawing-fresh-battle-lines-around-the-ethics-of-big-data/

56 See GDPR Recital 32.

57 See Data analytics and the GDPR: friends or foes? A call for a dynamic approach to data protection law by Sophie Stalla-Bourdillon and Alison Knight at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3248976 at page 12.

58 See WP29 Opinion 06/2014 on the *Notion of Legitimate Interests of the Data Controller* at https://iapp.org/media/pdf/resource_center/wp217_legitimate-interests_04-2014.pdf and also *Guidelines on Automated Individual Decision-Making and Profiling* at https://iapp.org/media/pdf/resource_center/W29-auto-decision_profiling_02-2018.pdf

59 See WP29 Opinion 06/2014 on the *Notion of Legitimate Interests of the Data Controller* at https://iapp.org/media/pdf/resource_center/wp217_legitimate-interests_04-2014.pdf

60 Id.

61 Id.

62 See GDPR Article 6(1)(f)

63 See GDPR Articles 5(1)(b) and 6(4)

64 See GDPR Article 5(1)(c)

65 See GDPR Article 89(1)

66 See Note 59, *supra, and Note 75, infra.*

67 See GDPR Article 21(3)

68 See Note 59, *supra, and Note 75, infra.*

69 Id.

70 See GDPR Recital 50 and Articles 5(1)(b) and 6(4)

71 See GDPR Article 5(1)(b)

72 See GDPR Articles 6(4)(e) and 89(1)

73 See GDPR Article 5(1)(b)

74 See https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2013/wp203_en.pdf

75 See Meeting the Challenges of Big Data at https://edps.europa.eu/sites/edp/files/publication/15-11-19_big_data_en.pdf on page 15.

76 See GDPR Article 89(1)

77 See Note 74, *supra*

78 See Note 57, *supra,* at page 16

79 See data scientist expert determination for Anonos BigPrivacy technology attached as Appendix 1 to this blueprint.

80 See https://www.exchangewire.com/blog/2017/12/04/probabilistic-key-unlocks-new-markets/

81 See https://iapp.org/media/pdf/resource_center/W29-auto-decision_profiling_02-2018.pdf

82 See https://www.sciencedirect.com/science/article/pii/S026736491730376X

83 See https://iapp.org/news/a/did-the-wp29-misinterpret-the-gdpr-on-automated-decision-making/

84 See GDPR Article 5(1)(c)

85 See GDPR Article 25

86 See Note 57, *supra,* at page 15

87 The concept of Controlled Linkable Data was presented at an International Association of Privacy Professionals (IAPP) program entitled General Data Protection Regulation (GDPR) Big Data Analytics featuring Gwendal Le Grand, Director of Technology and Innovation at the French Data Protection Authority—the CNIL, Mike Hintze, Partner at Hintze Law and former Chief Privacy Counsel and Assistant General Counsel at Microsoft, and Gary LaFever, CEO and Co-Founder at Anonos and former law partner at Hogan Lovells (see https://anonos.com/GDPR_Industry_ FAQ.pdf) and explained in a White Paper co-authored by Messrs. Hintze and LaFever entitled Meeting Upcoming GDPR Requirements While Maximizing the Full Value of Data Analytics (see https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2927540)

88 See GDPR Recital 29

89 See GDPR Article data scientist expert determination for Anonos BigPrivacy technology attached as Appendix 1 to this blueprint.

90 GDPR Recital 29 specifically provides that "In order to create incentives to apply pseudonymisation when processing personal data, measures of pseudonymisation should, whilst allowing general analysis, be possible within the same controller when that controller has taken technical and organisational measures necessary to ensure, for the processing concerned, that this Regulation is implemented, and that additional information for attributing the personal data to a specific data subject is kept separately. The controller processing the personal data should indicate the authorised persons within the same controller."

91 See GDPR Article 20(1)

92 See GDPR Article 21(1)

93 See GDPR Articles 21(2) and (3)

94 See GDPR Articles 11(2), 12(2) and 21(1)-(3). Under GDPR Articles 11(2) and 12(2) a data subject must provide additional information enabling his or her re-identification to exercise rights under Articles 15-22, however, this may not be possible if a data controller uses pseudonymisation techniques and deletes all information that could be used to reverse the pseudonymisation process so that the data controller "is not in a position to identify the data subject."

95 Id

96 See, *inter alia*, GDPR Recitals 13, 18, 22, 23, 24, 28, 36, 77, 78, 79, 80, 81, 82, 83, 95, 101, 108, 114, 115, 122, 124, 126, 127, 131, 145, 146, and Articles 3(1), 3(2), 4(8), 27, 28, 29, 30(2), 31, 32, 33, 35, 36, 37, 38, 39, 44, 46, 58, 79, 82, and 83. See also Hon, W. Kuan. *Data Localization Laws and Policy: the EU Data Protection International Transfers Restriction through a Cloud Computing Lens*. Edward Elgar Publishing, 2017.

Anonos®

APPENDIX 1

DATA SCIENTIST EXPERT OPINION ON BIGPRIVACY

TITLE: DATA PRIVACY IN AN AGE OF INCREASINGLY SPECIFIC AND PUBLICLY AVAILABLE DATA: AN ANALYSIS OF RISK RESULTING FROM DATA TREATED USING ANONOS' BIGPRIVACY METHODOLOGY.

December 2, 2017

Author: Sean Clouston, PhD, #3-096, Stony Brook University, Health Sciences Center, 101 Nicholls Rd., Stony Brook, NY, 11794.

Conflicts of interest: The original version of this report was started by Dr. Clouston when he was an external and independent consultant for Anonos, at which time the author had no conflicts of interest to report. However, during the process of the finalization of the initial, and this updated version of the report, but before publication of the initial version of this report, the author became an Anonos shareholder. Anonos played no role in the study conclusions, and did not direct the analysis. Anonos provided editorial suggestions relating to the implications and implementations of the analytic results provided herein.

Note: Changes from the original 2015 report, as reflected in this updated December 2, 2017 report, are limited to clarifications that the efficacy of the BigPrivacy methodology applies equally in situations where an equivalence class is comprised of five members.

Abstract

The public's interest in the reasonable expectation of privacy is met when personal data (PD) remains private: within reasonable limits, the data cannot be used to single out, or to inferentially identify or link PD to a particular data subject. Historically, privacy has been maintained by reducing access to identifiable data, while ensuring that the likelihood of re-identification, largely interpreted through equivalence classes, is reduced. However, PD is increasingly measured outside of traditional situations, contains increasingly detailed information, and is being utilized by new entities in new purposes. Yet, the public maintains the same interest in privacy. As a result, static de-identification tactics, which delete known identifiers by replacing them with one token used consistently for an individual, have been increasingly susceptible to critique.

One way forward is to use new temporally dynamic obscurity protocols that actively minimize the risk of re-identification. This report analyzes the ability of BigPrivacy to minimize re-identification under different circumstances and thus to ensure data privacy. Analyses provided in this report aid in assessing privacy and security risks and maintaining privacy and security when data incorporates detailed and even longitudinal information. PD is kept private within an acceptable level of risk, subject to some constraints on oversight and sample size. Data stored or transmitted pursuant to these methods is de-identified in the traditional sense; further, data storage and transmission are more robust using these methods for a number of reasons outlined in the report. Moreover, because data security is dynamic, privacy policies can be flexibly implemented to ensure security is consistently and completely ensured.

Introduction

Data privacy is important for many reasons. As just one example, health data, once released, could be used to reveal realities about a person's health that he or she may not want to share, including diagnoses of societally stigmatized diseases (e.g., PTSD, HIV/AIDS, Schizophrenia, etc.) and health issues having financial implications for their families or their health insurance carriers (e.g., physical activity, blood pressure, financially discriminatory actions, etc.). On the other hand, as Matthews and Harel (2011) highlight in their review of data privacy issues and solutions, researchers must ensure that data are accessible for potentially valuable research applications. Moreover, data are increasingly collected as by-products of private sector innovation, but while these data need to be protected, it is not in the public interest to stifle innovation requiring that data. Static de-identification, which we here define as de-identification that maintains the structure and linkages of data and achieves de-identification by replacing identification data with a randomized static token, was previously deemed to be sufficient for reducing the risk of re-identification. However, new data, more powerful types of data analytics, and the increasing number of data sources have made researchers, policymakers, and software developers sceptical this can continue (Chen & Zhao, 2012; de Montjoye, Radaelli, Singh, & Pentland, 2015).

As one example of how regulations are affected by the issues surrounding data minimisation, a U.S. Federal Trade Commission report noted that while HIPAA traditionally only pertains to a small number of people handling health information, such as doctors or hospitals, "health apps are [increasingly] collecting this same information through consumer-facing products, to which HIPAA protections do not apply…" and goes on to state that "consumers should have transparency and choices over their sensitive health information, regardless of who collects it" (Federal Trade Commission, 2015). The conclusion of the FTC report was twofold: the majority decision supports the need for "data collection minimisation," or the wholesale deletion of collected information from the information ecosystem, while the minority decision notes that this form of data minimisation might negatively impact health-related research and decision making (Federal Trade Commission, 2015).

The FTC dissent highlights the contrast between data value and data privacy. A non-partisan research firm (the Information Technology and Innovation Foundation or ITIF), further highlights problems with data collection minimisation in the private sector: "the FTC's report correctly recognizes that the Internet of Things offers potentially revolutionary benefits for consumers and that the industry is still at an early stage, [but the report] unfortunately attempts to shoehorn old ideas on new technology by calling for broad-based privacy legislation"; further, "in calling for companies to reduce their use of data, the FTC misses the point that data is the driving force behind innovation in today's information economy" (Castro, 2015). These dissenters each view such data collection and analysis efforts as *serving the individual and public interests*, even at the cost of privacy. *Wired* magazine concretizes these dissents, reporting that though IoT devices currently being developed are geared towards gathering "reams of largely superficial information for young people whose health isn't in question, or at risk" (Herz, 2014), "the people who could most benefit from this technology—the old, the chronically ill, the poor—are being ignored… [primarily because] companies seem more interested in helping the affluent and tech-savvy sculpt their abs and run 5Ks than navigating the labyrinthine world of… HIPAA."

Three Main Limitations

There are three main limitations with the current approach to data privacy. First, static de-identification is not robust. Second, transmission is particularly problematic. Third, an increasing number of entities are involved in providing guidance about privacy in a way that is increasingly difficult to maintain. These are discussed in detail in the following section.

First, data are not truly de-identifiable. Specifically, a recent article in *Science* observed the relative ease with which one can uniquely identify individuals using only small amounts of financial information (de Montjoye et al., 2015): indeed, for 90% of the cited sample only four pieces of information were needed to achieve "unicity" – i.e., development of unique identifying profiles derived from traditionally de-identified financial data. As noted by (El Emam, 2015), unicity in the dataset does not mean that any person has successfully re-identified each individual; however, once de-identified and made available to the public, data are subject to "data fusion", which is the linking of multiple different datasets together in order to broaden our understanding of the people in a dataset. The risk of data fusion has led to this finding being highly publicized, for example the *Wall Street Journal* noted that unicity in financial data meant one could readily "find the name of the person in question by matching their activity against other publicly available information such as LinkedIn and Facebook, Twitter, and social-media check-in apps such as Foursquare" (Hotz, 2015). The *Harvard Business Review* concludes the implications of this work "are profound. Broadly, it means that static anonymity doesn't ensure privacy" (Berinato, 2015).

Second, current data de-identification tactics when data are being transmitted are especially questionable, such transmission increasingly occurring through devices considered to be within the "Internet of Things" (IoT) (Rivera & van der Meulen, 2014). During transmission, the normal "size" of the dataset is curtailed, further weakening the assumptions on which we rely for data security. At the same time, the amount and specificity of data is increasing with data available such as the person's everyday progression from home to work or the number of calories, types of food, and restaurants that they ate in during their last week. Furthermore, IoT devices increase the importance of information transmission; for example, in the case of healthcare information, clinicians might be able to use interconnected devices to monitor a patient's health, including vital signs or physical activity, potentially raising new concerns regarding data privacy and an increased risk of data breach (Tyrrell, 2014).

Finally, de-identification, proceeding in a static manner, must be implemented under one specific policy regime to the detriment of others. For example, it may be that data collected under one policy are made more or less secure than are necessary under a newer or different structure so that data managers either must redo their work to a different standard, resulting in substantial inefficiency, or may simply choose not to allow access to data because the cost is too large to ensure compliance. In such a circumstance, having pre-approved levels of access could help to ensure that data are both accessible to individuals from varying regions or policy regimes, and at varying levels of security.

Data Construction

Without a solution that responds to these concerns and truly de-identifies PD, a broad range of individuals including, but not limited to, software developers and information technology specialists, will have access to non- de-identified PD data. Static de-identification, as noted above, is not working. Dynamically obscuring data may be one way to retain data privacy while reducing the risk involved in collecting, storing, and analysing such data (Warren, 2014). Examining new ways requires a basic knowledge of dataset construction, different types of data, and existing de-identification protocols. The following sections provide an introduction to topics underlying these issues before moving on to a more formal analysis of the risk of re-identification.

De-identification

The process of de-identification decreases privacy risks to individuals by removing identifying information from protected or personal data (PD). Thus, in the dataset presented in Table 1 below, data from the first column (identifying information, here an IP address) would need to be removed from the dataset. We should note, however, that de-identification usually references two somewhat separate processes: the removal of the certainty that any particular individual is part of the observed dataset, and the removal of the certainty that any particular observation might, in the correct circumstances, be sufficiently unique to be re-identified with other available data. Thus, while it is often believed that removing these IP addresses renders similar datasets (usually with more observations) "de-identified" in the traditional sense, as discussed above many of these observations can be uniquely identified using data characteristics that can lead to "unicity" within the database, rendering them unique in the data and thereby at risk of re-identification (de Montjoye et al., 2015).

The problem of de-identification has historically been addressed in a number of temporally static ways. Matthews and Harel (2011) list the following techniques used for de-identification: 1) limitation of detail, 2) top/bottom coding, 3) suppression, 4) rounding, 5) adding noise, and 6) sampling. *Limitation of detail* works through categorizing or collapsing information to reduce the possibility of characteristic re-identification. *Top/bottom coding* characterizes the replacement of observational data with a "top-code", an upper limit on all published values of a variable, and/or a "bottom-code", a lower limit on all published values of a variable, the replacement of which reduces the likelihood of re-identification of data that are more likely to be unique, such as very high incomes, by recoding them so that outlying observations are grouped together. *Suppression* works by removing potentially identifiable data from the publicly available dataset. *Rounding* introduces noise by randomly re-assigning rounded values to all the individuals in a dataset, and is often used for ages because though we may know that individuals are 45.67 years old (i.e., 45 years and 8 months), we recode that information into yearly (as age 45) or into age groupings (such as 45-49). *Addition of noise* uses a randomization routine to change the values in each cell by some random amount, an approach often used with geographic data such as that in the *Demographic and Health Surveys*, which have randomly dispersed geographic residential locations by some distance less than five kilometers (Measure DHS & ICF International, 2013). Finally, *sampling* resolves de-identification by requiring that data released be only a subset of the data available, with the convention that between 95-97% of the data collected could be released; however, sampling also resolves another issue, notably that individuals should not be known to have been a part of the dataset, because it removes, at random, entire individuals from a dataset so that you may not be certain that any particular person who was originally contained within the dataset are also contained within the dataset released.

Since then, more complex mathematical routines have been used to ensure data is kept confidential and that this confidentiality is unlikely to be broken. The most useful of these build on the randomization approach because it is the most secure and removes the least value from the data. *Matrix masking*, for example, codifies the data by multiplying them by a form of encryption key that researchers must know about and account for when analysing data (Cox, 1994). Another particularly interesting method, called *synthetic data,* uses data matching methods originally built to provide results for missing data to effectively swap characteristics between individuals in a dataset, thereby retaining some of the statistical uniqueness and value while reducing the risk of re-identification (Rubin, 1993). Note that these methods work by *increasing uncertainty* about the identity of any one individual without unnecessarily modifying the data itself.

De-identification

Current static de-identification methods generally rely on data being purposefully collected into one dataset before being anonymized, using usually one or two of the previously mentioned techniques, to protect privacy. These datasets are then shared with the public either at large or through limited access policies depending, largely, on the level of specificity provided in the data, or kept hidden from the public entirely.

This data fusion may be increasingly easy to do, especially as individuals "check in" on publicly available social networking sites. These data fusion techniques are big, flexible, fast, constantly changing, and being made more specific; they also use data that are being increasingly used, held, or transmitted by an ever larger number of individuals. Thus, the conclusions of de Montjoye et al. (2013, 2015) are likely true: if data are unique, those data may be readily identifiable among a substantial portion of users.

Estimating risk of re-identification

The probability of re-identification can be estimated (El Emam, Dankar, Vaillancourt, Roffey, & Lysyk, 2009). Re-identification requires access to the full datasets, which contain information for five separate individuals, and that re-identification is being undertaken purposefully to find a particular individual who may be in the dataset. We provide an example of a blood-pressure monitoring application on an internet-connected wearable device (i.e., a phone, watch, shoe, or other wearable device), which 1) identifies a user, 2) monitors health information, and 3) is linked to geographic positioning systems (GPS) data. Because this is a useful and clearly risky subject, we will rely on this example throughout this analysis. To protect the privacy of individuals associated with these data, we are faced with the following difficulties: de-identification may be subject to difficulties in both data storage and in data transmission, and further may propose a risk to multiple governing bodies since it collects data that may include health information, could easily integrate different types of data including longitudinal data, and may also have clinical applications if clinicians are interested in monitoring patients' everyday health in thisway.

Below we define a data matrix or dataset ***H*** (Table 1), which for simplicity is a 5 x 6 matrix that contains 30 unique data points (called cells). Different rows may contain different information for the same individual if that person is followed over time or is observed by different people (in longitudinal or higher-dimensional data). Note that there are therefore both explicit and implicit identifiers within most datasets: the IP address is explicit while the row number, when not longitudinal data, is an implicit identifier.

Table 1. Hypothetical dataset (**H**) collected from multiple smartphones on the same network by a blood pressure application in December 2014

IP address	Latitude	Longitude	Age	Sex	High blood pressure
192.168.0.0	40.13	-79.85	32	M	No
192.168.101.201	40.13	-79.86	45	F	Yes
192.168.4.57	40.15	-79.54	39	M	No
192.168.22.40	40.29	-79.54	56	M	No
192.168.1.220	40.29	-79.56	42	F	No

Note: IP addresses are assumed static for period of observation; Latitude and Longitude are "rounded off" to two decimal places and so are not precise to the level of a specific house.

Using the example in Table 1 above, we will suppose that *Alice* is the person with high blood pressure corresponding to IP address 192.168.101.201. Let us assume we know the type of information a neighbour, co-worker, or employer might know about *Alice*. Suppose, for example, we know she is female, that she was born approximately 40-50 years before the data were collected, that we know that she lives in Belle Vernon, Pennsylvania (Latitude, Longitude = +40.13, -79.86). However, we want to know further whether *Alice* has high blood pressure, and thus we also need to re-identify her using the data provided.

We follow previous reviews of re-identification risk assessment (El Emam et al., 2009) that define an "acceptable risk" as one that is at most = 0.20; further, for reasons that will become clear later, we further clarify that an "unacceptable risk" is one known to be greater than = 0.20. Then, the whole dataset's risk of re-identification (*r*) can be defined as: , where *f* is the number of individuals with equivalent characteristics to any one particular person, including here *Alice*, and min_j is the minimum number of individuals in a subset of categories (*j;* sometimes called an equivalence class, the basis for the "unicity" argument) that fit *Alice*'s known characteristics.

The risk has also been specified in the following way, including the measurement of how many equivalence classes that there are in a dataset who are expected to be distinct. Along those lines, Benitez and Malin (2010) provide the following definition of total risk: where *k* references the number of individuals in *b* possible equivalence classes for a sample of size *n*. Because the total risk revolves around the risk within a particular equivalence class, we thus begin by briefly overviewing equivalence classes in re-identification, before explaining why data parsing helps secure privacy.

Re-identification in practice

In table 1 above, our first efforts would be to rely on Anonos BigPrivacy to mask and eliminate IP addresses, replacing that information with Anonos BigPrivacy patented dynamically pseudonymised tokens (referred to herein as "Dynamic De-IDentifiers" or "DDIDs"). Prior to this, the risk of identification is perfect: $r = 1/1 = 1$. This is an unacceptable risk because $r = 1 > = 0.20$. However, after removing IP addresses, the risk is reduced because we cannot rely on identifiable information. In particular, the risk becomes $r = 1/n = 1/5 = 0.20$, a borderline but acceptable risk. We still, however, know that *Alice* is a woman aged 40-50 who lives in Belle Vernon, Pennsylvania. The risk of re-identification as a woman is defined as the inverse of the number of people in the equivalence class: in this case, the inverse of the number of women in the dataset, and thus . Because 0.5 is larger than 0.2 (as defined above by our categorization of acceptability,), we note that this is already an unacceptable level of risk. However, for clarification as to the nature of data linkages we push this data further to use more characteristic and specific data. We examine the data and note that there are two women aged 40-50 in the data. Therefore, we calculate r = ½ = 0.50; since r > = 0.20 this remains an unacceptable risk. We further know that *Alice* lives in Belle Vernon (latitude ranges from 40.12 to 40.14, longitude ranging from -79.84 to -79.86). This shows us that there are two people in these data living in Belle Vernon, and thus we calculate r = ½ = 0.50; since r > = 0.20 we define this as an unacceptable risk. Linked together, we can further see that, of those people living in Belle Vernon, only one is a female aged 40-50. Thus, data linking increases our risk of re-identification to $r = 1$, an unacceptable risk resulting in certain re-identification.

Table 2. Hypothetical BigPrivacy data (**H**) collected from multiple smartphones on the same network by a blood pressure application in December 2014

DDID	IP address	DDID	Lat.	DDID	Long.	DDID	Age	DDID	Sex	DDID	High BP
5657	192.168.4.57	5934	40.13	5049	-79.86	4958	42	5141	F	6878	No
5854	192.168.101.201	3030	40.29	3060	-79.85	3938	32	6236	M	4948	No
3938	192.168.0.0	1234	40.13	9090	-79.54	5010	45	7747	M	4094	No
5910	192.168.1.220	1410	40.15	8974	-79.54	7079	56	8585	M	0967	No
2039	192.168.22.40	4040	40.29	9030	-79.56	7078	39	9999	F	0847	Yes

Note: We have shortened title names to limit table size. Lat.: Latitude; Long.: Longitude; BP: Blood Pressure; DDID: Dynamic de-identifiers. Each DDID references the value in the column to its right. For ease of reference, the above table is represented to include both DDIDs and corresponding data; in an actual implementation of the BigPrivacy method, this table would not contain BigPrivacy keys that would be stored in a highly secure master look-up database which contains information necessary to, under technically controlled conditions, relink all connections between direct and indirect (structured and unstructured) identifiers and DDIDs.

Removing information from the implicit linkages is novel because it removes the possibility of linking data together to contextualize information. Thus, data are not saved in Table 1 above, but in a way that more closely resembles Table 2 above. For ease, the following discussions reference the "worst case scenario," wherein an entire dataset with small sample size is observed. We have here kept the variables in the same order: as would be obvious to any knowledgeable party, each column references similar data within variables and different data between variables; and further, the order of variables is rarely meaningful. Also, four-digit numerical characters were used as DDIDs for simplicity alone; this referencing does not reflect the method through which Anonos BigPrivacy derives or defines its DDIDs. We assume, for conservative estimates, that we know which indicator each observation references.

Using this method, both explicit and implicit data linkages are broken, because the parsing process reassigns DDIDs to each individual observation. This effectively removes the real (explicit) *and* contextual (implicit) identifiers from the dataset, and thus eliminates the risk presented by such equivalence classes, while also masking unicity in the dataset. Specifically, it is not clear, without accessing the identification maps, whether the high blood pressure (DDID=0847) is assigned to a female person (DDID=9999). Furthermore, we cannot use that context to link data together to identify with certainty that DDID=0847 is *Alice's* blood pressure reading, as compared to DDID=6878. In this dataset with n=5 individuals, we now know that the risk of re-identification is $r = 1/n = 1/5$ (random) and since $1/5 = 0.20$, this was an acceptable level of risk. Put more strongly, any dataset with at least five observations would be compliant using the Anonos BigPrivacy method. However, even if we uniquely found a single individual with high blood pressure (as is the case above), doing so does not improve our ability to link that to any individual nor to access other information using that knowledge.

Dynamism and uncertainty

While this dataset is currently seen as time-invariant (the most conservative case for analysis, and likely when a third-party gains access to a full dataset only once, perhaps through capture of a transmission); this may not actually be the case when using Anonos' BigPrivacy method over time. Specifically, because Anonos BigPrivacy does not use temporally-static identifiers, downloading the same dataset a second time could easily lead us to reasonably, but incorrectly, conclude that the new dataset references new data because the new DDIDs are dynamic and thus new identifiers reference new data *which is also reordered*. Thus, it may be possible that the second time data are transmitted, the sample size no longer seems to reference only five people, but instead might be seen as incorporating different data and thus reference a sample that may be as large as 10. Doing so would effectively reduce the risk of re-identification so that $1/10 \leq r \leq 1/5$.

Deceptive replication

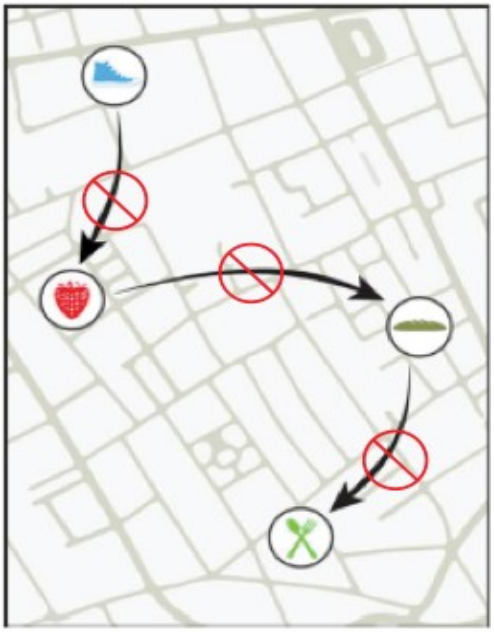
Similarly, you could mask the data in this table by adding in random information (with requisite DDIDs) and similarly mask the sample size and unicity in the data. These types of deceptive replication efforts may be further compounded because new data are randomly sorted and may incorporate more newly integrated observations. If this is the case, the duplicated or incorrect observations may *give the appearance of a larger sample size*, reducing the risk to a range ($1/10 \leq r \leq 1/5$), so that, on average assuming a uniform or normal risk distribution within that range, $r = 3/20$, a reduced risk of re-identification. However, this reduction depends on the assumption that the dataset as a whole does not contain any unique information that is not anonymized (such as the IP address above), which would be obvious once replicated and thus retain the more conservative $r = 1/n$ level of risk.

A secondary, and robust, gain from Anonos' BigPrivacy method is that we no longer know whether the two women referenced in the table above are two women or the same person measured twice. ***Challenging these assumptions can be uniquely beneficial because it forces us to question whether our basic assumptions about equivalence classes are correct, and further highlights the role of data that changes over time.*** Specifically, while we may note that here we have two people with different blood pressures, having two women the same age actually requires us to either assume that each outcome references different people (with the level of risk noted above) or to wrongly posit that these two observations reference the same person and either their health status changed over time (forcing us to question the nature of re-identification itself, since we can no longer determine when a person had the health outcome), or more likely assuming (incorrectly again) that the woman aged 42 did not have high blood pressure because there is only one observation of high blood pressure but two observations of that person.

shop	user_id	time	DDID	price
	6730G	09/23	G!022	$97.30
	S iMX	09/23	015M	$4.33
	3092fcl0	09/23	15166	$43.78
	Z3O	09/23	11A	$35.81
	4c7af72a	09/23	^99M	$15.13
	89c0829c	09/24	0099S	$12.29
	E2Grs	09/24	56KHJ	$3.66

Derived from:
Science 30 January 2015:
Vol. 347 no.6221 pp.536-539
DOI: 10.1126/science.1256297
www.sciencemag.org/content/347/6221/536

Fig. 1 Anonos Just-In-Time-Identity (JITI) enables dynamic protection at the data element level.
The universal "no symbols" highlight that dynamically obscuring data linkages that could be aggregated by parsing recognizable static "anonymous" identifiers breaks the assumptions necessary for re-identification.



Note: Figure 1 above shows that DDIDs 6730G, SiMX, Z3O and E2Grs may be used to refer to the same user at various times. And further provides DDIDs for prices, which have been randomly ordered. This obscures data linkages that could otherwise be aggregated by parsing recognizable static "anonymous" identifiers like the identifier "7abc1a23" that was used to refer to "Scott" for each transaction in de Montjoye et al. (2015).

In either case, these reasonable assumptions force us to make incorrect conclusions, making longitudinal data *less useful to re-identification than cross-sectional data*. Specifically, the risk of re-identification is no longer inversely proportional to the number of people, but instead to the number of observations: $r \leq 1/5$. This challenges the assumptions made by de Montjoye et al. (2015), who used longitudinal and specific data linkages to uniquely identify individuals by, as is noted in Figure 1 above provided by Anonos, breaking the assumptions required for unicity. Specifically, Figure 1 notes that while multiple data-points can be used to uniquely differentiate individuals, that the Anonos BigPrivacy method breaks the implicit and explicit linkages between data points, and thus effectively removes the ability to represent data in this way. Specifically, we may know that an individual shopped at shop A, but not how much they paid at that shop nor which store they shopped at next (if any).

Put conservatively, Anonos' BigPrivacy method does not preclude the possibility that data are unique in a dataset (for example, the high blood pressure reading is unique in Tables 1 and 2 above), but makes that information useless in determining anything else about those data. It prevents unique data from being used to attempt data fusion.

Transmission

By assigning to each person a DDID, Anonos replaces individual identities with temporally dynamic random information, replacing usual static re-identification protocols that we otherwise rely on in making assumptions when re-identifying. The ability to re-identify individuals by intercepting transmitted information is predicated on data linkages – various quasi-identifiers or variables that can be aggregated by parsing the identifiers. During transmission, these data are described by Anonos as being dynamically obscured via the DDID process, and transmitted as the data available above. This corresponds to the following data stream being sent from Table 2 above: DDID=0967; High BP=No; DDID=3030; Lat.=40.29; DDID=4958; Age=42; etc.

If Anonos BigPrivacy fails, the risk that this blood pressure indicates *Alice's* blood pressure is r=1/n. However, the calculated risk is variable and, though dependent on sample size, the risk remains unknown to those interested in re-identification (because the total sample size is unknown). Moreover, deferring the transmission of this information and sending it separately increases uncertainty about the context in which the data is being delivered; because data are being de-identified, we cannot be assured, without making assumptions about this equivalency, even if only longitudinal data referencing a single person's data were being delivered over a period of time, that this data referenced multiple observations of a single person who moved around between different locations, rather than of multiple people living near each other. For example, while the above represents five people because each category provides different locations and IP addresses for a wearable dynamic, if we replaced the table above with a table referencing two women aged 45 followed up for three time points, the new data would *appear* identical to data referencing one woman aged 45 but followed up for six time points. This would be especially confusing if this person moved around between areas. Again, however, given that we know where a person was at time 1, we cannot use that information to derive information about her health or location at that time. As long as the total number of application users equals or exceeds five, and no assumptions can be made about the number and types of information available during a particular transmission (i.e., we cannot know that only one person's information is being transmitted), the risk of re-identification remains acceptable even during transmission.

Let us assume the worst-case scenario: that such a transmission was caught and *revealed in its entirety*. Then we are left with the case, explicated above, where the number of users is known. Thus, , r=1/n=1/5=0.20, which we deem to be an acceptable risk (because 0.2 ≤ 0.20). However, if the transmission is not entirely caught (for example, if blood pressure is not caught or sample sizes differ between observed variables), then the risk of re-identification must be derived from information known about the number of users of this particular blood-pressure application at this particular site. Because the number of users must be at least five (since there are five in our known dataset), we know that the risk of re-identification becomes bounded by the dataset and user-base specifics, and is thus at most 1/5 but could be as small as the actual number of potential users (n_p) so that the risk is potentially much smaller ($r=1/n_p$) since n_p is at least five but may include a huge number of users, and thus we would say that the risk of re-identification is $r \leq 1/5$.

In this case, transmission effectively replicates the "sampling" method detailed by Matthews and Harel (2011) above; a common de-identification technique in itself. Formally, this means that Anonos BigPrivacy efforts serve to increase n by adding to it an unknown amount (k), where $k \in Z$. With the addition of k, re-identification then relies on the probability $1/n = 1/(n' + k)$, where k is *unobserved*. Notably, k could easily be made up of data that is similar to the real data, but is fake, or by replicating randomly sorted data that is not differentiable from its copied counterpart. As a result, the risk decreases rapidly by the inverse of the total number of observed and *unobserved* users (n). More concretely, if we know that at *Alice's* place of work that there are 20 users of the particular application then the risk = 1/20 = 0.05, which is less than 0.20. If, however, all employees (say 350) have been provided access to the application, then the data specific risk = $1/N_{employees}$ = 1/350 = 0.0026 < 0.20. In either case, the risk is acceptable as long as the number of users in the full, accessible, dataset does not allow for $r = 1/n \geq 0.20$ (i.e., $n \geq 5$).

Optimization and efficiency

Regulatory compliance specifically requires that PD are subject to a reasonably low risk of re-identification. We have shown above that the BigPrivacy method can reduce that risk. However, it may be inefficient to dynamically identify every piece of information at all times, so understanding the necessity of such levels of security to maintaining BigPrivacy compliancy may be useful. Above, we suggested that we could parse data by cell into randomized data with unique DDIDs. However, we could maintain many of the same levels of security by viewing the main dataset as a matrix of matrices (i.e., that each matrix ***H*** contains '***j***' matrices within which the data reside). As such, DDIDs could be used to secure data not by providing DDIDs to each cell in the dataset, but instead to each matrix of cells in the dataset. This would provide much of the same level of security discussed above, but would be much more computationally efficient.

Specifically, modifying Table 1 above we provide the following dataset as a set of groups that are defined by the DDID to create Table 3 below, where each level of gray (of which there are j=6 made up of 3 rows and between 1 and 2 columns of data) signifies a different dataset formed of sequential or random pieces of information that could only be reassembled, like a puzzle, using the key. Here, the blocks were predefined to overlap with each type of variable in part because this overlap is easier to see and manage, but is also in many ways more secure.

Table 3. Hypothetical dataset (**H**) collected from multiple smartphones on the same network by a blood pressure application in December 2014

IP address	Latitude	Longitude	Age	Sex	High blood pressure
192.168.0.0	40.13	-79.85	32	M	No
192.168.101.201	40.13	-79.86	45	F	Yes
192.168.4.57	40.15	-79.54	39	M	No
192.168.22.40	40.29	-79.54	56	M	No
192.168.1.220	40.29	-79.56	42	F	No

It would be evident if one happened to receive one particular sub-matrix, that there was a respondent with high blood pressure (the darkest gray). However, it would be impossible to ensure that this respondent was our fictional individual, "Alice" as above. Nevertheless, it would be entirely feasible to know this if certain types of data were contained within that dataset, and thus security would only be ensured if types of data were contained separately from each other, and if the matrix mapping were not unique in itself (i.e., if matrix ***H*** could be reasonably made of by assembling these ***j*** pieces in a number of ways). Here we noted that data could be differentiated in this way: the IP address column is white because we assume it would be deleted, while the data in the blood pressure chart are held in pieces that could be made up of information in *n* (here 6) ways. As such, this provides similar security as does the method above with one caveat: if data are longitudinal and variables are stored in concert, and the outcome are sufficiently specific, then there is a small chance of matching data types together. Nevertheless, this would be reasonably solved by implementing variation in the levels of security promoted by the types of data so that publicly available data are stored in *j* cells while more sensitive data are stored in single cells without linked data. In this example, let us suggest that, under the worst-case scenario, we received the data in full but separated by shade of gray into six datasets. In this scenario, we would know because of our mapping only that one respondent had high blood pressure, resulting in a risk of re-identification of 1/5, which we have defined as acceptable. However, if this were not the worst-case scenario and only a subset of the data were received (the darkest gray box, for example) then the risk of re-identification is *at most* 1/5 and is at least $1/n_p$ where n_p includes the entire potential user base.

As noted above, this would be further secured by the occlusion tactics described above, and could be modified to secure other types of information than health data, subject to risk analysis about the specificity of that data. This application of this type of analysis has two added notes regarding specific data. Specific data could be considered to be largely "identifiable" information and would need to be separated from other forms of information to maintain privacy. Finally, longitudinal data (a form of specific data) can be stored in two ways: wide, with each variable noting an observation; or long, with each observation identified within the implicit structure underlying a dataset (as in Figure 1 above). In either case, this method could be made robust to long or wide data depending on mapping techniques and, if necessary, random sorting. Crucially, in this scenario PD would still 1) not be subject to data fusion and 2) be kept private even if unicity were achieved in the data itself.

Differential levels of security

One benefit of this type of optimization in conjunction with dynamic capabilities is that it facilitates the ability for users to manage security flexibly. Specifically, if some data were considered a greater risk than other data, it would be possible to vary the level of security used to secure different types of data and to secure data for different purveyors. For example, let us imagine that age and sex needed less security levels than a medical diagnosis of schizophrenia. It would be possible to differentiate them and use different mechanisms to organize them. Such variation matches how comfortable individuals might feel sharing information in person. For example, one could keep the age and sex variables integrated but randomly sorted. The diagnoses could, on the other hand, be differentiated and dynamically de-identified and excluded from the dataset's user base unless specifically provided by the security policy. This differentiation would both eradicate the risk of data fusion and would minimize the risk of re-identification. However, it could allow easier access to basic demographic information for accepted purposes. In this way, users could, with the guidance of policymakers *or with the explicit permission of individuals from whom they have collected data,* readily provide a sliding scale of coverage where different types of data are differentially secure.

In practice, this would imply that Anonos BigPrivacy could implement what could be termed a "programmatic policy", or a digitized representation of policy decisions that defines,* a priori, *how data are shared by specifying which data are shared when and with whom.

Unknown Third Parties

The above analyses are sufficient to ensure de-identification in the traditional, static, sense. However, we live in an increasingly demanding and dynamic world, with increasingly opaque privacy protocols. We may therefore reasonably assume that the end-user is not yet defined and that more complex associations may arise. We also may encounter the real outcome that an end-user may try to re-identify individuals in their own data surreptitiously without the knowledge of an implementation of the Anonos BigPrivacy method. We may thus be interested in knowing whether an unknown third party (U3P), not bound by data privacy standards and in possession of substantial resources (human, financial, or political), could *surreptitiously* manipulate the data to facilitate re-identification. If this is the case, then interested parties might have strong incentives to try to *find or create* a circumstance where re-identification could be made easier. These third parties may be internationally based and thus not easily dis-incentivized by standard legal considerations.

To examine this possibility, we asked the following hypothetical question: *could an end-user, with a previously specified user base, ask specific questions in order to facilitate re-identification?* Put more specifically, in an attempt to identify a target user, could a U3P modify an existing membership's data collection routine, containing the targeted user, to modify their data collection routine (but not their user base or Graphical User Interface / GUI) to clandestinely determine that user while incurring an unacceptable level of risk (> 0.20) that health data refer to a particular individual (for ease, *Alice*)?

The most readily available technique is to add questions or indicators that would easily facilitate such re-identification. For example, the risk of re-identification could be increased by defining multiple non-threatening questions that overlap in order to increase unicity and facilitate the unique identification a particular person, or by linking smartphone data or metadata (including, for example, GPS information) to publicly available information. However, because identifiable information and characteristic indicators, which might be easily added to the application in order to expressly identify the individual of interest (i.e., to maximise the risk of re-identification) are

subject to Anonos' BigPrivacy method, these linkages are readily dealt with as noted above. We must therefore assume the U3P could simply access the full dataset with the identifiers from an authorized user; thus, the worst-case scenario is that they would know a person was part of the data collected. It may be possible then to gain the full dataset, but using Anonos' BigPrivacy method, these data will not be linked and thus the result will not be more informative than that – you would know that a person was part of the data, and that there is a $1/n$ risk that any indicator, including high blood pressure, a relatively low probability. Because these data are not linked, we know that asking identifiable or characteristic questions could only be used to determine the health of a particular individual with a risk of re-identification of $1/n$.

If identifiable or characteristic data are not useful, it may still be possible to determine/create a situation in which information is both 1) interesting in its own right, and 2) sufficiently specific to determine with risk ($r >= 0.20$) that a person fits the outcome suggested. This quest is trivial if the person of interest does not, or is not known to, use the application during the period of time under examination, since the risk will then always be $0/n = 0$. However, in our hypothetical situation, the U3P would know that the user's information was contained within the dataset provided. Thus, the data requested must, in a single unlinked variable, reference an outcome where its specificity and risk could be sufficient to identify an individual's information solely on its specifics. This is easiest when the potential outcome is strange rare (either the disease or lack of disease) since the risk of identification relies on assumptions and unicity in that dataset.

To maximise re-identification risks, we must therefore create *specific* data that are both the health variable of interest and sufficiently unique to successfully identify the information desired. This is a tall order and highly unlikely in any normal dataset, so an interloper asking these types of questions might be obvious to the respondents. In such a situation, we might reasonably assume most individuals to be free of the disease, and that we have reason to believe that the risk that *Alice* has the disease is M times greater than the normal population. Nevertheless, we want then to know what the likelihood is that *Alice* (A) has condition R, given that R is observed in the dataset (denoted, $P(A|R)$). This calculation can be solved using Bayes' Theorem. The probability *Alice* has the disease is: $P(A|R) = P(R|A)*P(A)/P(R)$. These other probabilities are either known or can be guessed. For example, the probability that any observation is *Alice's* is $P(A) = 1/n$. The probability that any sample contains an observation of that is P(R particular disease) $= (R*(n-1)+M*R)/n = R*(n-1+M)/n$, where R (such that $0 \leq R \leq 1$) is the risk of the disease. We believe, from our expectations derived from external observation, that if *Alice* has a risk M times the normal risk (R) of observing the outcome such that $Q = 1-R$, then the probability of a positive outcome given that *Alice* is in the sample is $P(R|A) = MR$. Thus, we have from Bayes' Theorem that $P(A|R) = P(R|A)*P(A)/P(R) = (MR*1/n)/(R*(n-1+M)/n) = M/(n-1+M)$.

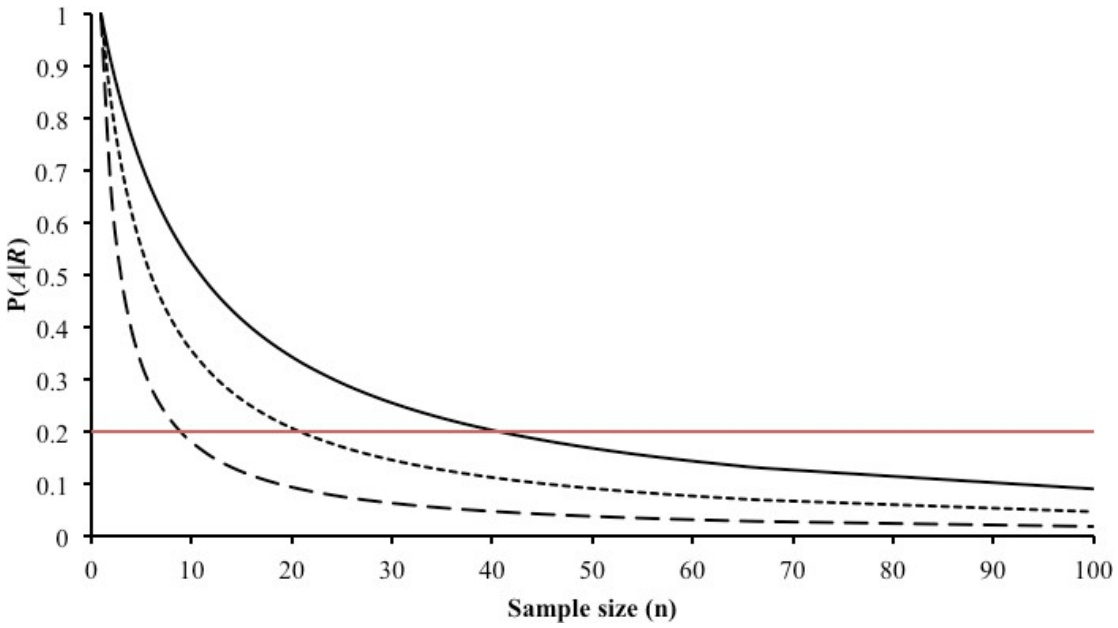


Figure 2. Likelihood of re-identification under the hypothetical condition that data are manipulated in order to engineer the best conditions possible to identify individuals, with an estimated *M* of 2 (long dashes), 5 (dotted), and 10 (solid) by sample size.

Simulations estimating the risk of re-identification given a single positive observation follows Figure 2 above. We have here assumed a range of relative risks ranging from conservative (*M* = 2) to medium (*M* = 5) to very large (*M* = 10). This range of relative risks (*M*) was allowed to range from 2-to-10 to reflect the range often seen for predictors in epidemiological research, and because at most *M* references the difference between a risk of 2% (a rare outcome) and 20% (the risk necessary to be reasonably certain = 0.20). Notably, risks much higher than 2% become decreasingly likely to enable an *M* = 10 outcome because when the population's risk approaches 10%, the personal risk must approach 100% (i.e., 10*10%), a known certainty, and thus the need for re-identification is unnecessary to begin with.

Figure 2 above provides a more conservative estimate of the risk of re-identification than traditional methods. This estimate suggests that in the worst possible situations that Anonos' BigPrivacy method is robust to intentional privacy intrusions by a U3P undertaken with express knowledge of the Anonos BigPrivacy method, as long as total sample size exceeds 41 individuals (the point where the solid black line (*M* = 10) crosses = 0.20). Notably, while it is unlikely that all data are going to need this level of privacy, it is reasonable to suggest that when data are treated in this manner that they achieve or surpass this stringent level of security.

Discussion

In this analysis, we described Anonos' BigPrivacy method as starting with the premise of blending existing methods of de-identification, including for example sampling, suppression and the potential addition of noise, with novel temporally dynamic identifiers and data parsing protocols. We have analysed the risk of re-identification, finding that the Anonos BigPrivacy method can drastically reduce the risk of re-identification, even for specific data. Moreover, these analyses we found that, using the Anonos BigPrivacy method, data were kept private during both transmission and storage, even from the application developers. Specifically, we found that re-identification risks were minimized and could be reduced under the generally accepted statistical and scientific principles and methods for rendering information not individually identifiable (threshold value (here defined as = 0.20)), when the total sample size equalled five analytic units (e.g., individuals, households, online identities, IP addresses, etc.). Moreover, we discussed the potential for BigPrivacy processes to be applied to blocks of data rather than individual observations. We further found that the level of security could be managed by variable differentiation and de-linkage, so that some information, such as basic demographic information was not de-identified but other information was at the same time subjected to the BigPrivacy process. We further discussed the potential for both policymakers and for individuals from whom the data are collected to help define the level of security of particular data.

Risk of re-identification

The risk of re-identification is limited by constraining assumptions that can be made about data contents and structure (Kifer & Machanavajjhala, 2011). *Anonos works by breaking the assumptions that are encoded in datasets and used by others to achieve re-identification.*

Breaking these assumptions has a number of benefits, but the most important one is that it makes the both re-identification and the risk of re-identification difficult to ascertain with any level of certainty without further gaining access to the complete, unadulterated, dataset. Anonos BigPrivacy does this in a few main ways discussed below.

Being dynamic helps. DDIDs provide a level of protection from data and the misuse of data that are not available now. For example, DDIDs necessarily re-integrate randomized follow-up information from the same individuals if data were downloaded later, and thus serve to increase sample size and reduce re-identification risks while reducing our ability to make assumptions about the completeness of the data. Secondly, the data differentiate instances from one another, making assumptions about the completeness of data, and their reference population, less clear. Third, Anonos BigPrivacy efforts work well during transmission to effectively occlude shared information and to maintain security even with characteristic and specific data. Finally, the method can be made robust even to those who are engaged in collecting the data, making data privacy clear and enforcing data use agreements even when unknown third parties are engaged in using the data.

Flexibility of privacy and security

This technology enforced decoding of DDIDs could apply broadly, within a single deployment of the Anonos BigPrivacy method, but it would be possible to overlap multiple cascading rule sets, with an agreed-upon hierarchical relationship, to govern usage of any given primary data table. In practice, this could mean that a lead country's Data Protection Authority (DPA) might define the highest-ranking set of PD access rules, but another concerned party might also insert its own set of PD access rules that may be more stringent. These rules might be applied differently to different types of data within the same dataset. In this event, Anonos can be configured to ensure that no PD access is possible unless both cascaded sets of DPA access rules are enforced when the query is made. Conversely, BigPrivacy could provide flexible controls necessary to support hierarchical handling of various data privacy requirements.

Programmatic policy

The ability to deliver on the many promises of big data in linking together individuals with institutions, clinicians, or researchers, for example, is predicated on this ability to support differing privacy requirements depending on the nature and source of data. Anonos BigPrivacy provides a way to automatically and digitally enforce such privacy policies. For example, consumer health data collected using electronic health records, mobile health applications, and social networking sites may be accessed and data may be useful and available. At the same time, financial data may be transcribed into the same data using the same devices. However, PD is at the same time regulated by privacy and security requirements under a given country's privacy and health privacy acts and may further be subject to specific privacy policies and terms and conditions depending on user preferences for specific websites, devices and applications. The BigPrivacy key itself encodes both the rules necessary to recover the source value from at least one DDID and flexible programmatic policies, or a digitized representation of the privacy policy that is subject to observation, enforcement, and audit. Therefore, if necessary rules and constraints are not being observed and enforced, either because there is 1) a mismatch between a user's permissions and the query that user is trying to submit or 2) access was once granted but has since been revoked or expired, then no DDIDs may be decoded.

Conclusion

The Anonos BigPrivacy invention and protocol mitigates the risk of re-identification by repudiating assumptions about explicit and implicit data linkages. It can therefore ensure privacy even when the dataset as a whole contains characteristic or specific data, such as when single individuals are followed over time or specific details such as geographic location are observed.

The flexibility of the BigPrivacy key mechanism ensures that the Anonos policy-driven BigPrivacy data management platform, even in cases where a single data element of PD must be protected by a single BigPrivacy key, programmatically enforces granular rules for access. We also found that, even when individuals worked to design a situation favouring re-identification, Anonos' BigPrivacy method continued to minimize the risk of re-identification by first removing the risk of characteristic re-identification, while repudiating the ability to make assumptions about the structure of the data, and also by limiting the risk of specific re-identification to acceptable levels given sample size limitations. We then identified opportunities to both: further occlude data in cases of small numbers of observations, and optimize occlusion to facilitate large-scale data management.

It is the author's opinion from the analyses conducted and described herein that, subject to oversight and sample size limitations, Anonos' BigPrivacy method substantially mitigates to a statistically acceptable level the ability to single out, infer about, or link data to an individual so that personal data remains private.

Author Biosketch

Sean Clouston, PhD, is a *Fulbright* scholar who earned his Bachelor's in Arts in Mathematics and Sociology. He did his doctoral work at *McGill University* in statistics, epidemiology, and demography and is currently employed as Core Faculty in the Program in Public Health, and an Assistant Professor of Family, Population, and Preventive Medicine at *Stony Brook University*, part of the State University of New York. Dr. Clouston uses quantitative analysis on high dimensional data to examine questions relating to the distribution and determinants of disease both inthe United States and globally.

References

Benitez, K., & Malin, B. (2010). Evaluating re-identification risks with respect to the HIPAA privacy rule. *Journal of the American Medical Informatics Association, 17*(2), 169-177.

Berinato, S. (2015, February 9th). There's no such thing as anonymous data. *Harvard Business Review*.

Castro, D. (2015). FTC's Internet of Things Report Misses the Mark [Press release]

Chen, D., & Zhao, H. (2012). *Data security and privacy protection issues in cloud computing.* Paper presented at the Computer Science and Electronics Engineering (ICCSEE), 2012 International Conference on.

Cox, L. (1994). Matrix masking methods for disclosure limitation in microdata. *Survey methodology, 20*(2), 165-169.

de Montjoye, Y.-A., Radaelli, L., Singh, V. K., & Pentland, A. S. (2015). Unique in the shopping mall: On the reidentifiability of credit card metadata. *Science, 347*(6221), 536-539. doi:10.1126/science.1256297

El Emam, K. (2015). Is it safe to anonymize data?

El Emam, K., Dankar, F. K., Vaillancourt, R., Roffey, T., & Lysyk, M. (2009). Evaluating the risk of re-identification of patients from hospital prescription records. *The Canadian journal of hospital pharmacy, 62*(4), 307.

Federal Trade Commission. (2015). *Internet of things: Privacy & security in a connected world*. Washington, DC: Federal Trade Commission.

Herz, J. (2014). Wearables are totally failing the people who need them most. *Wired*.

Hotz, R. L. (2015, January 29). Metadata Can Expose Person's Identity Even When Name Isn't; Researchers Use New Analytic Formula. *Wall Street Journal*.

Kifer, D., & Machanavajjhala, A. (2011). *No free lunch in data privacy.* Paper presented at the Proceedings of the 2011 ACM SIGMOD International Conference on Management of data.

Matthews, G. J., & Harel, O. (2011). Data confidentiality: A review of methods for statistical disclosure limitation and methods for assessing privacy. *Statistics Surveys, 5*, 1-29.

Measure DHS, & ICF International. (2013). *Demographic and Health Surveys*. MD5.

Rivera, J., & van der Meulen, R. (2014). Gartner says the Internet of Things will transform the data center. *Gartner*. Retrieved from http://www.gartner.com/newsroom/id/2684616

Rubin, D. B. (1993). Statistical disclosure limitation. *Journal of Official Statistics, 9*(2), 461-468.

Tyrrell, C. (2014). Countering HITECH privacy risks from internet of things products. *HITECH*

Answers. Warren, N. (2014). Dynamic Data Obscurity. *Category Archives*.