ANONOS

Processing Cleartext: A Clear and Present Danger

Statutory Pseudonymisation Permits Processing in Ways Privacy Enhancing Techniques Cannot

August 2022

www.anonos.com

Processing cleartext presents a clear and present danger due to the:

(1) Illegality under Schrems II of many international data transfers, including global decentralised processing in the cloud; and
(2) Inability to establish lawful bases for processing for analytics, machine learning or AI.

Attempts to address these risks using most Privacy Enhancing Techniques (PET) produce results that are unacceptable from both a utility and a data protection perspective.

Fortunately, Anonos enables your organisation to avoid these risks by capture the overwhelming advantages of Statutory Pseudonymisation.

Anonos Data Embassy Software

Enables **Lawful, Value Maximizing, Decentralized Processing**

By Technologically Enforcing **Centralized Controls**

Leveraging **Statutory Pseudonymisation**

To Protect Data **in Use**

Wherever it Travels

What is Lawful, Value Maximizing, Decentralized Processing?

- Compliant with relevant contractual and regulatory requirements
- 100% accuracy and utility relative to processing record-level cleartext for analytics, machine learning and AI
- Processing speed comparable to cleartext
- Readily shared and combined across organizational and jurisdictional boundaries, including hybrid- and multi-cloud processing

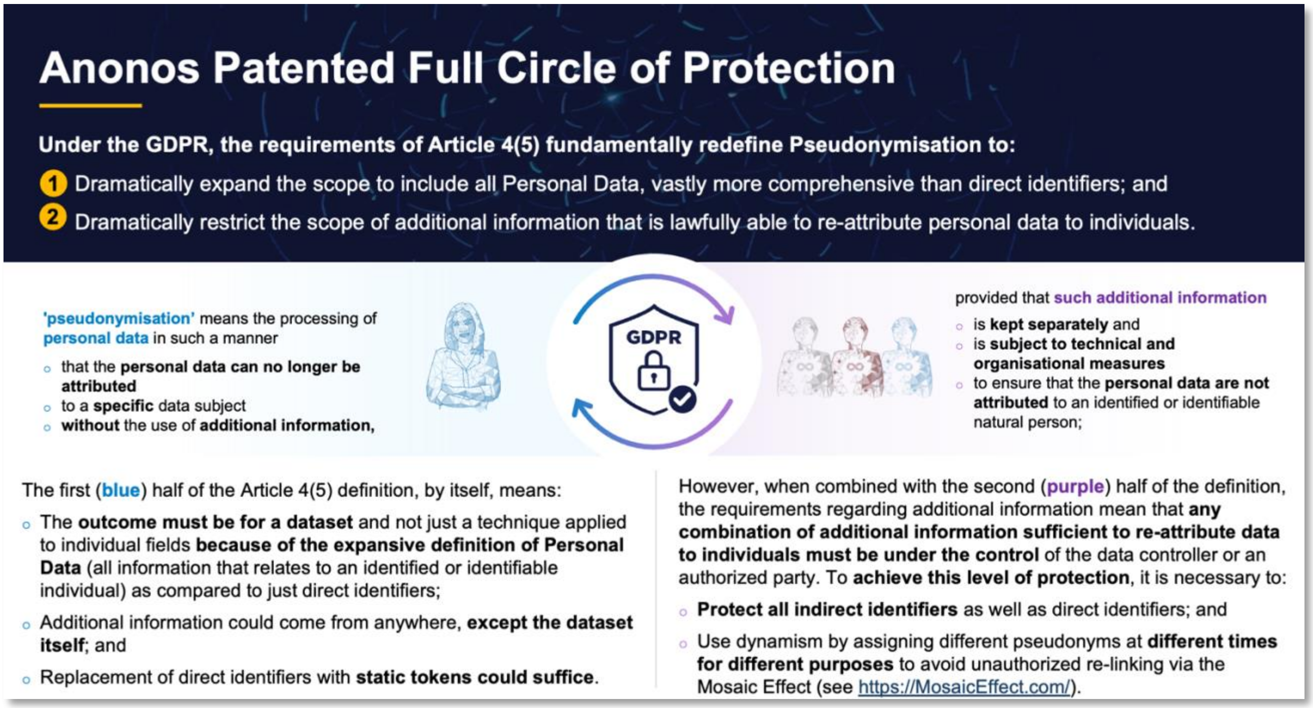
What are Technologically Enforced Centralized Controls?

- Digitally Encoded Data Privacy Policies
- Protections that are embedded into data to protect it while in use, wherever it travels
- Immutable Audit Logs
- Group-based permissions to enforce Separation of Responsibilities
- Role-based permissions to enforce Segregation of Duties, including approvals
- Relinkability Controls to enforce Need to Know

What is Statutory Pseudonymisation?

Essentially identical statutory language is found in the EU GDPR, the UK GDPR and the Data Protection regulations of Brazil, Japan, South Korea and five US States (CA, VA, CO, UT, and CT). In each case it has been included to provide a means for reconciling conflicts between maximising data value and protection. Other countries and states are looking to adopt similar provisions to incorporate Statutory Pseudonymisation because of its unique ability to simultaneously maximize to both data value and data protection.

The following graphic highlights how significantly different and more demanding the requirements are for Statutory Pseudonymisation than for the privacy enhancing technique (PET) known variously as pseudonymization, hashing, tokenization, and key-coding.



Anonos Patented Full Circle of Protection

Under the GDPR, the requirements of Article 4(5) fundamentally redefine Pseudonymisation to:

1. Dramatically expand the scope to include all Personal Data, vastly more comprehensive than direct identifiers; and
2. Dramatically restrict the scope of additional information that is lawfully able to re-attribute personal data to individuals.

'pseudonymisation' means the processing of **personal data** in such a manner

- that the **personal data can no longer be attributed**
- to a **specific** data subject
- **without** the use of **additional information,**

provided that **such additional information**

- is **kept separately** and
- is **subject to technical and organisational measures**
- to ensure that the **personal data are not attributed** to an identified or identifiable natural person;

The first (**blue**) half of the Article 4(5) definition, by itself, means:

- The **outcome must be for a dataset** and not just a technique applied to individual fields **because of the expansive definition of Personal Data** (all information that relates to an identified or identifiable individual) as compared to just direct identifiers;
- Additional information could come from anywhere, **except the dataset itself**; and
- Replacement of direct identifiers with **static tokens could suffice**.

However, when combined with the second (**purple**) half of the definition, the requirements regarding additional information mean that **any combination of additional information sufficient to re-attribute data to individuals must be under the control** of the data controller or an authorized party. To **achieve this level of protection**, it is necessary to:

- **Protect all indirect identifiers** as well as direct identifiers; and
- Use dynamism by assigning different pseudonyms at **different times for different purposes** to avoid unauthorized re-linking via the Mosaic Effect (see https://MosaicEffect.com/).

Adopted by Regulators Around the Globe:

ANONOS

EUROPEAN UNION	UK	BRAZIL	SOUTH KOREA	JAPAN	US STATES (5)

STATUTORY PSEUDONYMISATION

SOO ▪ DON ▪ NIM ▪ EYE ▪ SAY ▪ SHN

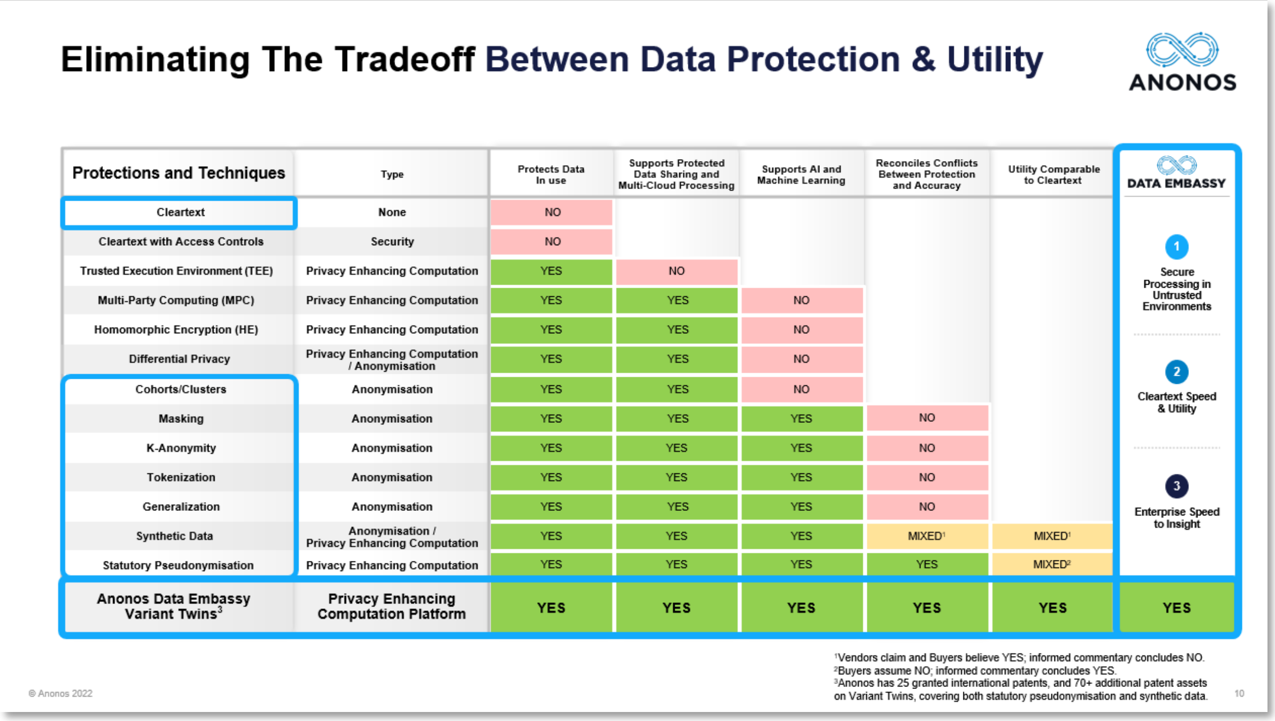
© Anonos 2022

4

See below for a more detailed description of what is necessary to transform cleartext to a protected output that meets the heightened requirements for Statutory Pseudonymisation.

Shortcomings of Privacy Enhancing Technologies (PETs), Including the Failure to Protect Data in Use Wherever it Travels

The following chart evaluates the full range of data protection techniques, including both security-based approaches and traditional PETs against a series of criteria for evaluating the effectiveness of protection and the utility of the protected output. Rather than being a traditional red/green or stoplight chart that evaluates all PETs against all criteria, this is a knockout chart. PETs are evaluated against the criteria sequentially from left to right, and once a PET fails to meet a criterion it is dropped from further consideration. The following analysis also lists prominent vendors offering different technologies.



Eliminating The Tradeoff Between Data Protection & Utility

Protections and Techniques	Type	Protects Data In use	Supports Protected Data Sharing and Multi-Cloud Processing	Supports AI and Machine Learning	Reconciles Conflicts Between Protection and Accuracy	Utility Comparable to Cleartext	DATA EMBASSY
Cleartext	None	NO					1 Secure Processing in Untrusted Environments 2 Cleartext Speed & Utility 3 Enterprise Speed to Insight
Cleartext with Access Controls	Security	NO					
Trusted Execution Environment (TEE)	Privacy Enhancing Computation	YES	NO				
Multi-Party Computing (MPC)	Privacy Enhancing Computation	YES	YES	NO			
Homomorphic Encryption (HE)	Privacy Enhancing Computation	YES	YES	NO			
Differential Privacy	Privacy Enhancing Computation / Anonymisation	YES	YES	NO			
Cohorts/Clusters	Anonymisation	YES	YES	NO			
Masking	Anonymisation	YES	YES	YES	NO		
K-Anonymity	Anonymisation	YES	YES	YES	NO		
Tokenization	Anonymisation	YES	YES	YES	NO		
Generalization	Anonymisation	YES	YES	YES	NO		
Synthetic Data	Anonymisation / Privacy Enhancing Computation	YES	YES	YES	MIXED[1]	MIXED[1]	
Statutory Pseudonymisation	Privacy Enhancing Computation	YES	YES	YES	YES	MIXED[2]	
Anonos Data Embassy Variant Twins[3]	**Privacy Enhancing Computation Platform**	**YES**	**YES**	**YES**	**YES**	**YES**	**YES**

[1] Vendors claim and Buyers believe YES; informed commentary concludes NO.
[2] Buyers assume NO; informed commentary concludes YES.
[3] Anonos has 25 granted international patents, and 70+ additional patent assets on Variant Twins, covering both statutory pseudonymisation and synthetic data.

© Anonos 2022

10

Cleartext with Access Controls

Access controls are an essential component of data security. However, no matter how granular they are (e.g., attribute-based, tasked-based or even zero-trust) they are still binary; once granted, access is to clear-text. As a result, they do not provide protection for data in use.

Vendors: Okta, Microsoft, Ping, OneLogin, ForgeRock

Trusted Execution Environment (TEE) / Confidential Computing Environment (CEE):

Perhaps among the newest of new techniques, this approach sets up an on-processor enclave of a portion of system memory, and in some implementations, part of the CPU itself. Data is stored and moved around the processor in encrypted form until inside the enclave, where it is decrypted using a key only available within the enclave. Implementation is technically challenging, and often requires rewriting applications to work in the TEE. Additionally, the enclave is by definition a silo. Thus, this approach is not well-suited for data sharing and combining and multi-cloud or hybrid-cloud applications.

Vendors: Alibaba Cloud; AWS; Fortanix; IBM; Intel; Microsoft; Private Machines

Multi-Party Computing (MPC)

A relatively new technique that is frequently (mis)represented as “encryption in use”, presumably for marketing purposes. The justification seems to be that more precisely, the encoding of data done to enable the shared computations is fairly characterized as a cryptographic technique, as is encryption. But as commonly used, encryption is not understood to be the encoding done in MPC, which results in cleartext values. In any case, MPC requires tremendous bandwidth for the communication and coordination required between the computing parties, which can be both expensive and results in processing speed penalties, limiting its use to niche applications.

Vendors: Baffle; Cybernetica; Inpher; IXUP; LiveRamp; Nth Party; Sepior; Snowflake (CryptoNumerics); Triple Blind; Unbound Security; Ziroh Labs

Homomorphic Encryption (HE)

Also a “newer” technique, though research has been ongoing for many years in an effort to find a way to improve the speed to a level even approaching commercial viability. Most information touting progress talks only of “improvements” and not actual processing throughput results, for good reason. Estimates suggest processing speeds that are 5 to 10 orders of magnitude slower than processing cleartext. That implies that computations take would take 1 millisecond in clear text would take anywhere from 1.5 minutes to nearly 4 months.

Vendors: Duality Technologies; Enveil; IBM; Inpher; IXUP; LiveRamp; Ziroh Labs

Differential Privacy and Cohorts/Clusters

By definition, these techniques provide results that are aggregated, and do not provide the record-level output necessary for the vast majority of uses of data.

Vendors: Immuta; LeapYear; LiveRamp; PHEMI Systems; Privitar; Sarus Technologies

Anonymisation / De-Identification Techniques

The following techniques, Masking, Generalization, Tokenization, K-Anonymity, Noise Introduction, and Synthetic Data all are used, typically by combining several together, in an effort to Anonymise data. However, in the effort to do so, they all fail to resolve the intractable trade-off between privacy and utility that is inherent in anonymisation. Generally, in a big data world, they fail to deliver the privacy promised by anonymisation, and efforts to push them to their limits to do so end up destroying the utility of the protected output.

Masking

This technique protects direct identifiers by masking or overwriting one or more characters. It requires the data, its use, and its users are all restricted/sequestered to prevent other unprotected fields in a record from being combined with the information in additional data sources to enable an individual to be distinguished from others or identified via linkage attacks (see https://MosaicEffect.com). This requirement to restrict access is inconsistent with the architectural requirements of increasingly prevalent use cases that require free flowing data and involve dynamically changing data sources, processes, and processors.

Generalization

This technique attempts to protect against reidentification by reducing the granularity of the original data. Classic examples include converting age to age ranges by range binning, or by rounding numerical values. Masking can also be used for generalization such as masking one or more trailing digits of a zip or postal code to create values that represent larger areas. By itself, this technique does little to protect identity as it generally isn't useful for direct identifiers. It is often put into practice to achieve a specified level of k-anonymity (see below).

Tokenization

(Hashing/Key-Coding/Pre-GDPR Pseudonymization): These techniques: (i) only protect direct identifiers and (ii) protect those direct identifiers by replacing them with a recurring (persistent) token, making them effective only for limited, static use cases. They require that the data, its use, and its users are all restricted/sequestered to prevent other unprotected fields in a record from being combined with the information in additional data sources to enable an individual to be distinguished from others or identified via linkage attacks (see https://MosaicEffect.com). This requirement to restrict access is inconsistent with the architectural requirements of increasingly prevalent use cases that require free flowing data and involve dynamically changing data sources, processes, and processors.

K-Anonymity

K-anonymity techniques are intended to prevent a data subject from being singled out by grouping them with at least “k”-1 other individuals who share the same values for a specified subset of attributes in a data set. This subset of attributes, which are commonly referred to as quasi-identifiers because of their ability to, when used in combination, reveal identity. The quasi-identifiers are generalized as necessary (using techniques such as range binning, rounding, and masking) to ensure that all possible subgroups defined by the values of the quasi-identifiers have at least k individuals in them. In most cases, to achieve that status for all records in the data set, the required generalization severely degrades the utility of the data. To mitigate the degradation, a decision is made to be less aggressive in the generalization, and then suppress values or entire records in those subgroups where k falls short of the specified level. Note however that this also results in degradation of data utility, as a result of distortion in the output dataset statistical properties relative to those in the original source data.

Noise Introduction

This technique involves intentionally changing values in a data set so that they are less likely to be useful in revealing identify while at the same time avoiding excessive degradation in data utility due to distortion of the statistical relationships among attributes. This technique explicitly trades off utility (i.e., accuracy) for privacy, and tends to fall short on both accounts.

Vendors: Immuta, Privacera, Privacy Analytics, Privitar, Protegrity, SecurePi TokenEx

Synthetic Data

The failure of synthetic data to adequately protect against identity disclosure is now well documented in academic papers. The current state of the art appears to be ~ 1% of data subjects at risk of identity disclosure, which is likely to be judged to be far short of the regulatory requirements for anonymous data. Efforts to reduce this risk inevitably come at the expense of accuracy, as maximizing accuracy leads to overfitting and duplicating relatively unique records in the source data. Some organizations report accuracy rates of as low as 70% when attempting to ensure low risks of identity disclosure. An additional challenge relates to the incorporation of incremental records to an existing source data set, or the addition of additional tables. In order to properly preserve the statistical properties between records, fields and tables, these situations almost always will require regenerating the models used to create synthetic data

Vendors: AI.Reverie; Mostly.ai; Replica Analytics; Tonic

Requirements of Statutory Pseudonymisation

Statutory (GDPR) Pseudonymisation requires:

- **Protecting all data elements:** Footnotes 83 and 84 of the EDPB Final Schrems II Guidance[1] highlight that achieving GDPR Pseudonymisation status must be evaluated for a data set as a whole, not just particular fields. This requires assessing the degree of protection for all data elements in a data set, going beyond direct identifiers to include indirect identifiers and attributes. This is underscored by the definition of "Personal Data" under GDPR Article 4(1) as encompassing more than immediately identifying information and extending to "any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person."
- **Protecting against singling out attacks:** Paragraph 85 of the EDPB Final Schrems II Guidance mandates protection against "singling out" of a data subject in a larger group, effectively making the use of either k-anonymity or aggregation mandatory.
- **Dynamism:** Complying with the requirements in Paragraphs 79, 85, 86, 87 and 88 of the EDPB Final Schrems II Guidance to protect against the use of information from different datasets to re-identify data subjects necessitates the use of different replacement tokens at different times for differing purposes (i.e., dynamism) to prevent re-identification by leveraging correlations among data sets without access to the "additional information held separately" by the EU data controller (see https://www.MosaicEffect.com);
- **Non-algorithmic lookup tables:** the requirement of Paragraph 89 of the EDPB Final Schrems II Guidance to take into account the vulnerability of cryptographic techniques (particularly over time) to brute force attacks and quantum computing risk will necessitate the use of non-algorithmic derived look-up tables in certain instances; and
- **Controlled re-linkability:** The combination of the four preceding items are necessary to meet the requirement in Paragraph 85(1) of the EDPB Final Schrems II Guidance that, along with other requirements, the standard of EU GDPR pseudonymisation can be met only if "a data exporter transfers personal data processed in such a manner that the personal data can no longer be attributed to a specific data subject, nor be used to single out the data subject in a larger group, without the use of additional information."

[1] See EDPB Recommendations 01/2020 on Measures that Supplement Transfer Tools to Ensure Compliance with the EU Level of Protection of Personal Data (version 2.0) at https://edpb.europa.eu/system/files/2021-06/edpb_recommendations_202001vo.2.0_supplementarymeasurestransferstools_en.pdf

Benefits of Anonos Data Embassy Software

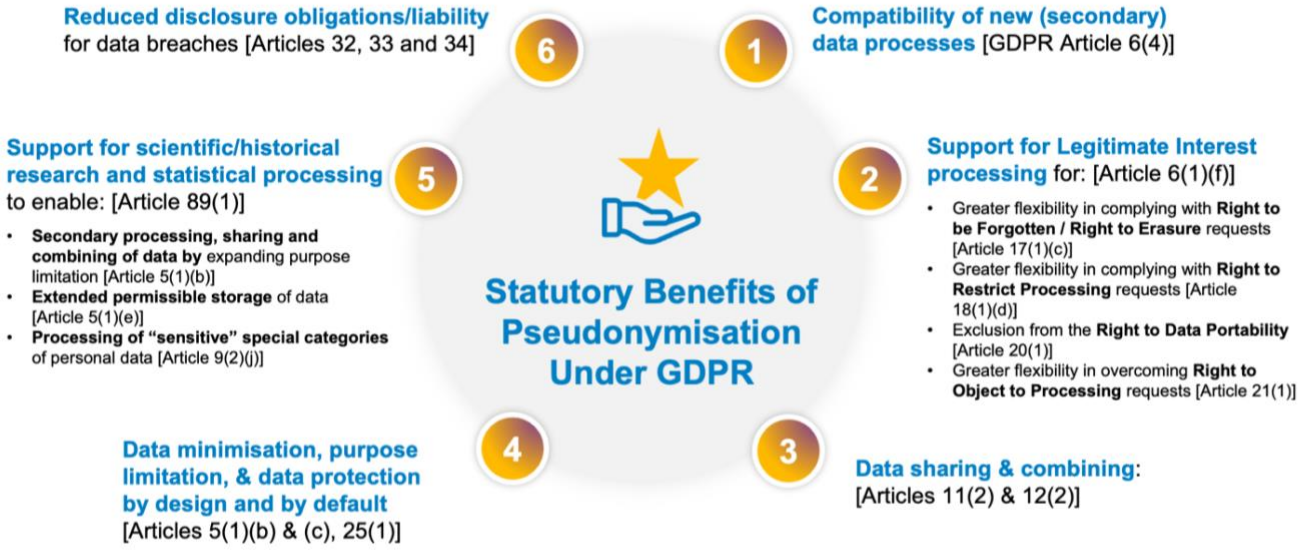
Anonos Data Embassy Software produces protected "Variant Twin" outputs that satisfy the requirements for Statutory Pseudonymisation and **go further by enabling**:

- **Dynamism:** The use of different dynamically assigned replacement pseudonyms at different times for different purposes that are not re-linkable introduces maximum "entropy" (uncertainty) within and between data sets thereby reducing the risk of unauthorized reidentification. If an adversary gains access to a statutorily pseudonymised Variant Twin, there are only a few options for attempting reidentification: (a) the information necessary to reidentify (keys or lookup tables); (b) the details of the transformation(s) used to create the Variant Twin (to allow inference of the information necessary to reidentify), or (c) auxiliary information (i.e., direct or indirect identifiers found in the Variant Twin). To ensure that a Variant Twin satisfies GDPR requirements for statutory pseudonymisation, the (a) keys and lookup tables as well as (b) the details of the transformation(s) remain under "technical and organizational controls that limit access" and are not available to unauthorized parties, including of course an adversary. Any (c) auxiliary information would be cleartext direct or indirect identifiers, however, in the Variant Twin these are either omitted or statutorily pseudonymised. Since attributes are specific to the data subject/data controller relationship, these are also not available to an unauthorised party. As a result, an adversary cannot reattribute information contained in the Variant Twin to a data subject.
- **Controlled Relinkability for 100% Accuracy & Utility:** The controlled relinkable nature of protected data outputs (called "Variant Twins") enable 100% accuracy and utility. Contrast that to (at best) 70% accuracy supported by alternative approaches like processing synthetic data or differential privacy that also suffer from levels of reidentification risk that fail to meet statutory requirements for anonymisation.
- **Scalability:** more expansive data use, sharing, and combining is possible due to:
  - Embedded dynamic de-risking controls that travel with the data and can be processed at the speed of unprotected cleartext without requiring additional processing power or bandwidth.
  - Technological enforcement of policies using standardized, predictable, comparable, and consistently applied controls support:
  - **16X improvements in productivity as a result of getting**
    - **4X as many projects approved**
    - **Each in 25% of the time**
  - Because reviews by privacy/legal are performed on an "exception" basis focused on how controls from new use cases differ from those previously approved.
  - Centralized secure storage of controllably relinkable data necessary to reveal identity in one place (or minimal places) versus innumerable locations throughout an ecosystem. This enables firms to continue using data while complying with data subject deletion requests, by deleting centralized links to identifiable "additional information held separately" but enabling remaining non-relinkable data to be retained and processed after the links have been deleted.

- Native support for high throughput/high availability by leveraging Kubernetes-based architectures enables throughputs exceeding **2 TB/hr (0.56 GB/sec) on clusters of just 12 nodes; *there are no known limits to scalability.***
- Facilitating processing of data in the cloud that was not capable of being processed previously, accelerating access to the benefits of the scalability of cloud-based processing.
- The illegal processing of cleartext EU personal data in US-operated clouds and other third-party infrastructures can in many situations be transformed into lawful processing by using Anonos software to satisfy the requirements for Statutory (GDPR) Pseudonymisation (see Use Case 2: Transfer of Pseudonymised Data as specified in EDPB Final Schrems II Guidance[2]).
- Support for the highest value use cases of advanced analytics, AI, and machine learning, by satisfying Statutory (GDPR) Pseudonymisation requirements to satisfy:
    - Article 6(1)(f) Legitimate Interest processing requirements,
    - Secondary Processing obligations under Article 6.4 (for processing beyond purposes authorised by Consent or Contract),
    - Data Protection by Design and by Default obligations under Article 25, and
    - Security of Processing obligations under Article 32.
- Anonos software leverages well-established technologies and open standards together with patented data/privacy engineering techniques that are quickly grasped by data engineers. The learning curve is short. It is the way we bring together our patented capabilities and open standards that is our secret sauce.
- Anonos software installs quickly on physical servers or VMs running Linux and Docker/Kubernetes, with only a handful of connection points:
    - Docker Hub (temporarily) for container images during install/upgrades,
    - Cryptlex (outbound only) for licensing checks
    - Keycloak to enterprise authorization applications
    - A reverse proxy connection for the browser-based UI used in development and test modes
    - Data connectors for batch (Spark) and streaming (Kafka) data ingest/output.
- Importantly, for production operations, all actions that can be accomplished manually using the browser, and others that cannot, can be automated via the fully documented API for the software, enabling hands off routine/recurring protection operations.

[2] See *Supra*, Note 1.

Additional Benefits Under the GDPR of Statutory Pseudonymisation



Statutory Benefits of Pseudonymisation Under GDPR

1 **Compatibility of new (secondary) data processes** [GDPR Article 6(4)]

2 **Support for Legitimate Interest processing** for: [Article 6(1)(f)]

- Greater flexibility in complying with **Right to be Forgotten / Right to Erasure** requests [Article 17(1)(c)]
- Greater flexibility in complying with **Right to Restrict Processing** requests [Article 18(1)(d)]
- Exclusion from the **Right to Data Portability** [Article 20(1)]
- Greater flexibility in overcoming **Right to Object to Processing** requests [Article 21(1)]

3 **Data sharing & combining**: [Articles 11(2) & 12(2)]

4 **Data minimisation, purpose limitation, & data protection by design and by default** [Articles 5(1)(b) & (c), 25(1)]

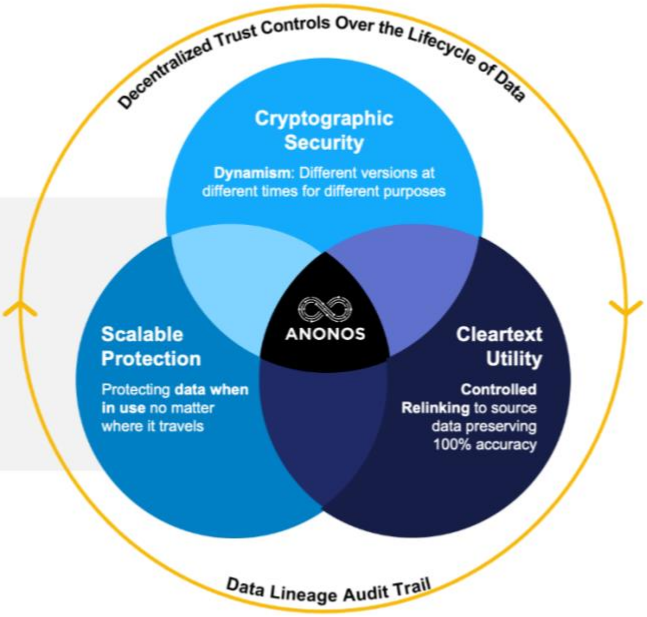
5 **Support for scientific/historical research and statistical processing** to enable: [Article 89(1)]

- **Secondary processing, sharing and combining of data by** expanding purpose limitation [Article 5(1)(b)]
- **Extended permissible storage** of data [Article 5(1)(e)]
- **Processing of "sensitive" special categories** of personal data [Article 9(2)(j)]

6 **Reduced disclosure obligations/liability** for data breaches [Articles 32, 33 and 34]

Anonos Variant Twins Deliver Statutory Pseudonymisation

Anonos is the **only** solution that minimizes risk and maximizes the utility of data globally


Decentralized Trust Controls Over the Lifecycle of Data
Cryptographic Security
Dynamism: Different versions at different times for different purposes
Scalable Protection
Protecting data when in use no matter where it travels
ANONOS
Cleartext Utility
Controlled Relinking to source data preserving 100% accuracy
Data Lineage Audit Trail

Clients Say Anonos is the **ONLY** technology that:

Minimizes risk and maximizes the utility of data globally

Maintains 100% accuracy & utility when processing protected data

Enables Lawful repurposing of data for secondary use processing

Supports surveillance-proof data processing and compliant international transfers (Schrems II, GDPR)

Does not require additional processing overhead – same speed as processing unprotected clear text

Delivers centralized controls over decentralized processing

Solves the limitations of consent and contract for analytics

Is available today; protected by 25 granted patents and 70+ patent assets

Unlocks access, analytics, sharing, and combining of data sets that are otherwise inaccessible

ANONOS

LearnMore@Anonos.com

WORLD ECONOMIC FORUM
Global Innovator

Gartner
Cool Vendor

www.anonos.com