DIGITAL TECHNOLOGIES ACTING AS A GATEKEEPER TO INFORMATION AND DATA FLOWS

D2.1 State-of-the-Art & GAP analysis Distributed Data Management, Processing and Storage

Document Identification			
Status	Final	**Due Date**	01/03/2023
Version	1.0	**Submission Date**	06/03/2023

Related WP	WP2	**Document Reference**	D2.1
Related Deliverable(s)	D2.2, D2.3	**Dissemination Level (*)**	PU
Lead Participant	FHG	**Lead Author**	Felix Hermsen (FHG)
Contributors	FHG, DBC, CEA , KUL, UTH, UOM, UPRC, UOG, IDSA, EGI, ECO, LISC, VTT, ANYSOL, UMU	**Reviewers**	Giulia, Giussani (IDSA) Apostolos Apostolaras (UTH) Ignacio Lamata Martínez (EGI)

Keywords:

Distributed Privacy-preserving Data Management, Blockchain, PPML, Trustworthy Data Sharing, Privacy by Design, Self-encryption, Secure data sharing, Self-sovereign Identity Management, Side-channel Attacks, EDA, Energy efficient AI, Privacy Threat Modelling

Disclaimer for Deliverables with dissemination level PUBLIC

This document is issued within the frame and for the purpose of the TANGO project. This project has received funding from the European Union’s Horizon Europe Framework Programme under Grant Agreement No. 101070052. The opinions expressed and arguments employed herein do not necessarily reflect the official views of the European Commission.

The dissemination of this document reflects only the author’s view and the European Commission is not responsible for any use that may be made of the information it contains. This deliverable is subject to final acceptance by the European Commission.

This document and its content are the property of the TANGO Consortium. The content of all or parts of this document can be used and distributed provided that the TANGO project and the document are properly referenced.

Each TANGO Partner may use this document in conformity with the TANGO Consortium Grant Agreement provisions.

Document Information

List of Contributors	
Name	Partner
Tomas Pariente Lobo	ATOS
Jürgen Neises, Claudia Mertinger	FSDE
Sofiane Lagraa, Moussa Ouedraogo	FUJ_LU
Vitalii Demianets	NORB
Konstantinos Kentrotis, Cristina Nichiforov	EXUS
Ioannis Spyropoulos, Antonios Chronakis	SVI
Thanassis Kountzeris	QBE
Elissavet Zogopoulou	SQD
Nejc Bat, Jakob Jenko, Matija Cankar	XLAB
Dolores Ordóñez	ANYS
Felix Hermsen, Avikarsha Mandal	FHG
Ville Ollikainen, Anni Karinsalo, Sami Lehtonen	VTT
Kaitai Liang, Zeshun Shi	TUD/DUT
Antonio Skarmeta, Jesús García Rodríguez	UMU
Theano Karanikioti, Peggy Valcke	KU Leuven
Apostolos Apostolaras, Stavroula Maglavera	UTH
Giulia Giussani	IDSA
Lauresha Memeti	ECO
Hugo Steep, Dries Verhees	FMAKE
Nicolas Belleville	CESGA/ CEA
Vasili Schewelow, Andreas Kopysov	VISAR
Christos Kotselidis, Athanasios Stratikopoulos	UoM
Sakshyam Panda, Emmanouil Panaousis	UoG
Ioannis Katsanakis	UPRC
Chariton Palaiologk	FN
Kiriakopoulou Georgia	MET

Document History			
Version	Date	Change editors	Changes
0.1	07/12/2022	Felix Hermsen (FHG)	Table of contents
0.2	08/01/2023	Felix Hermsen (FHG)	First input of technical partners
0.3	31/01/2023	Felix Hermsen (FHG)	Revised input of technical partners
0.4	13/02/2023	Felix Hermsen (FHG)	Pilot input
0.5	24/02/2023	Felix Hermsen (FHG)	Document ready for internal review
0.6	02/03/2023	Felix Hermsen (FHG)	Internal review collected and incorporated
1.0	03/03/2023	Felix Hermsen (FHG)	Final version ready for submission

Quality Control		
Role	Who (Partner short name)	Approval Date
Deliverable leader	Felix Hermsen (FHG)	02/03/2023
Quality manager	Juergen Neises (FSDE)	06/03/2023
Project Coordinator	Tomás Pariente (ATOS)	06/03/2023

Table of Contents

List of Tables

List of Figures

List of Acronyms

Abbreviation / acronym	Description
ALTAI	Assessment List for Trustworthy AI
CoE	Council of Europe
CP-ABE	Ciphertext-Policy Attribute-Based Encryption
CSI	Channel State Information
CTI	Cyber-Threat Intelligence
DIF	Decentralised Identity Foundation
DP	Differential Privacy
DPIA	Data Protection Impact Assessment
Dx.y	Deliverable number y belonging to WP x
EC	European Commission
ECHR	European Convention for Human Rights
EDA	Exploratory Data Analysis
FL	Federated Learning
GAN	Generative Adversarial Networks
GDPR	General Data Protection Regulation
HE	Homomorphic Encryption
HLEG	High-Level Expert Group
KP-ABE	Key-Policy Attribute-Based Encryption
KPI	Key performance indicator
LSTM	Long Short-Term Memory
MFA	Multi-Factor Authentication
PCS	Principal Component Analysis
PII	Personal Identifiable Information
POS	Proof of Stake
POW	Proof of Work
RSS	Received Signal Strength
SMPC	Secure Multiparty Computation
WP	Work Package
XAI	Explainable AI

Executive Summary

This deliverable, D2.1, presents the current state of the art (SOTA) on distributed data management, processing, and storage, and outlines how the TANGO project could innovate to go beyond the current SOTA. The primary objective of this deliverable is to identify potential key technologies that can be utilized to develop the TANGO platform. This has been accomplished by conducting a GAP analysis on various technologies and then mapping to different pilot cases. This deliverable is a result of the task T2.1 GAP Analysis in Distributed Data Management, Processing & Storage and provides foundation for task T2.2 User Needs and Requirements for Data Management, Processing & Storage (deliverable D2.2).

Upon reading this document, the reader will develop an understanding of relevant technologies for trustworthy data sharing, as well as learn about the current research gaps within the field. These gaps are identified by mapping each technology to at least one real-world use case. For instance, the reader will learn about privacy-preserving mechanisms, efficient AI, blockchain usage for distributed data management and storage, advanced encryption mechanisms and trust management. Beyond that, the reader will also gain an understanding of the current gaps and room for improvement, that has been derived by applying each technology to at least one of the considered pilots.

In the TANGO project, six pilot cases will be considered. The first one is smart hospitality, which is about improving the guest experience in a hotel. The second use case is about safeguarding data that is exchanged and used between a user and an autonomous car, while the third pilot is about improving additive manufacturing. Furthermore, pilot four is about training AI models using federated learning on data of several banks. Pilot five on the other hand is about a platform that helps people relocate to a different country by managing Visas. The objective in this pilot is to safeguard against multiple attack vectors on the distributed platform. Finally, pilot six deals with a large retailer that wants to improve its customer experience while maintaining GDPR compliance. This deliverable briefly introduces each pilot, along with potential technologies to be applied in each pilot.

In summary, the contributions of this deliverable are twofold and will serve as the foundation for future deliverables of the TANGO project. The first contribution is a thorough analysis of the gap for a variety of technologies for data management, processing, and storage. The second contribution is the mapping of these technologies to the six considered pilots and the identification of challenges and benefits.

1 Introduction

1.1 Purpose of the document

This document is the first technical deliverable of the TANGO project. The main goal of the document is to lay out a solid technical foundation on which the project will continue to develop. In addition, this document aims to align the TANGO consortium by providing a comprehensive description of available technologies on distribution data management, processing, and storage. In also includes an initial mapping of these technologies to TANGO pilot cases, to facilitate collaboration and joint exploitation among the consortium members. To achieve this, first, the document includes a state-of-the-art analysis for each technology to be used in the project. The state-of-the-art analysis also includes a general description of each technology. In total, the TANGO consortium considers 16 different technical areas, which fall into three broad categories, mainly: distributed data management and storage (WP3), distributed trust management (WP4), and AI-based techniques for green and trustworthy operations (WP5). Based on this, the document outlines how each technology could be improved from its current state by TANGO. Furthermore, this document gives an overview of five use cases (referred to as pilots) out of six and how the improved technologies could be applied to them. The description of another pilot (no. 4) will be added in the future deliverables. In a nutshell, the purpose of the document is to define how the TANGO platform can go beyond the state of the art and how the innovations will be applied to real world use cases.

1.2 Relation to other project work

As this document provides a state-of-the-art analysis for each technology belonging to the TANGO platform, this document can be viewed as the foundation on which the platform is built upon. Herein, this deliverable will serve as the basis for other deliverables in the project, specifically to: D2.2, D2.3, D3.1 – D3.5, D4.1- D4.5 and D5.1 – D5.6. To move forward, this document establishes the initial framework for potential innovations and limitations that should be considered in the future when making important management decisions.

1.3 Structure of the document

The rest of the document is organised in five main sections:

- **Section 2** presents the SOTA[1] of technologies that are used for *Distributed Privacy-preserving Data Management and Storage*. Specifically, the focus is on Blockchain based technologies, Self-encryption techniques, Dataspaces, ePrivacy Mechanisms, Sticky Policies and other trustworthy data sharing methods like Federated Learning.
- **Section 3** presents the SOTA of technologies that are used to enable Distributed Trust Management. The emphasis is primarily on self-sovereign identity management, user onboarding, behavioural user and device authentication, and defences against side channel attacks.
- **Section 4** discusses the SOTA of an AI-based Framework for Green & Trustworthy operations. It covers exploratory data analysis, energy efficient AI and model training, the dynamic intelligent execution of heterogenous systems, privacy threat modelling, explainable AI and infrastructure management specifically for AI.
- **Section 5** describes use cases (also known as *pilots*) and how the technologies described in Sections 3 to 5 can be applied to them.
- Finally, **Section 6** presents some conclusions on this deliverable and next steps.

[1] State of the art

2 SOTA on Distributed Privacy-preserving Data Management and Storage

2.1 Blockchain-based Data Storage and Sharing

2.1.1 SOTA including Comparison

Blockchain technology provides an immutable and decentralised way of sharing and managing data. It has been applied in different domains such as finance, healthcare and energy [1]. A distributed data model that is based on Blockchain, increases the degree of security and performance efficiency, and provides significant advantages for improved data management and sharing: increased level of security, untampered data streams and integrity and increased level of trust amongst stakeholders [2]. Without means for centralised data management, data security and system integrity are protected, as no single point of failure cannot occur.

Currently distributed ledger technologies [1], [2], have been introduced to provide applications to a wide range of domains ranging from data governance to verticals such as agriculture, healthcare, transport, etc.

Blockchain data storage:

A Blockchain is a decentralized, public ledger for recording transactions and securing the network. To access to data on a Blockchain, is only possible for authorized users of the system.

Blockchain features include:

- Data are openly shared to all nodes of the network and verified through consensus by participants.
- It offers encrypted transactions, time stamping, and proof of work.
- No centralised agency or institution alone controls the data, so its security is not ensured via traditional means, like those adopted by banks with traditional databases. Data can only be accessed by those permitted to do so, who must authenticate themselves as part of the network using private keys.

Traditional data storage method vs Blockchain

- Cloud storage is one of the most widely used form of data storage. The biggest disadvantage of cloud storage is sometimes it is operated/controlled by a central entity/platform and data transactions are not encrypted by default. Furthermore, as data is the most critical asset, its storage, processing, and analysis are often challenging. A platform that is decentralized and only stores encrypted data by default can help mitigate some of these downsides.

Reasons why Blockchain is required for data storage:

- Decentralization: The decentralized nature of Blockchain ensures that there is no single central entity governing data-related decisions.
- Security: Decentralized cloud data are difficult to attack, as there are multiple nodes in the network with the same copy of, so malicious attackers are forced to change data on the majority of nodes on the network to make a change look legitimate.
- Distributed: Blockchain is a distributed ledger where independent computers record, share, and synchronize the transactions instead of keeping data centralized in one location.
- Blockchain data are stored on a decentralized public ledger. The data on the ledger are stored in chunks called blocks, which are chained together using cryptography.
- Every block has a unique cryptographic hash as an identifier along with the previous block in the Blockchain.
- Each transaction inside a block is timestamped and added to the ledger with each block. Each new block records all transactions and adds them to the previous one. The data stored on Blockchain cannot be altered or removed from the blockchain as it would require alterations on every subsequent block.

Data sharing

Blockchain is a digital ledger of transactions that get duplicated and distributed over the entire network. It offers a decentralized system for managing data and transactions in a peer-to-peer network. Unlike the traditional way of sharing data through clouds in Web 2.0, the decentralized system allows the distribution of data divided into blocks. Each block is then protected from intruders attempting to alter the data. Every change in the data needs to be validated by all peers or miners (users) in the network using advanced cryptography techniques.

2.1.2 Innovation through TANGO

TANGO aims to go beyond the current state of the art and pioneer the creation of a Blockchain platform to store and share data and derived intelligence through controllable access mechanisms. This platform will perform trusted data transactions that will ensure data integrity, accountability, transparency, traceability, security and privacy of all the information shared amongst practitioners.

Through the TANGO pilot for Banking and Hospitality, safe, secure, and immutable nature of the sensitive PII can be shared and distributed within the network, to ensure a higher rate of protection and enhanced security.

Key innovation: legally compliant sharing of PII data points, enhanced privacy controls lacking in existing Blockchain solutions.

Consensus data sharing efficiency:

The current data sharing efficiency focuses on the speed and cost of accessing and processing data. On the other hand, consensus efficiency focuses on speed and cost of reaching consensus in a Blockchain network. It measures how quickly and effectively a network can validate transactions and add them to the Blockchain.

Consensus data sharing proves to be more secure and reliable. An efficient consensus mechanism validates transactions faster and at a lower cost, resulting in a more streamlined and reliable network.

Why private blockchain?

Both Proof of Work (PoW) and Proof of Stack (PoS) have their own disadvantages and may not be completely useful for certain use cases. With regards to the TANGO project, the main aim of having an energy efficient solution cannot be achieved by having PoW or PoS as a consensus mechanism.

Energy consumption, slow transaction and centralization are some of the disadvantages that prevent us from using Proof of Work and Proof of Stake. Both consensus mechanisms can lead to centralization as both depend on who has a major stake in the network. Also, PoS is less secure and vulnerable to attacks and PoW has high energy consumption.

All the above has led us to the decision of using a private Blockchain with permission-based architecture. None of the pilot use cases requires an open Blockchain with PoW or PoS consensus mechanism, thus enabling us to use more secure and energy-efficient private Blockchain.

Off-chain data storage

Five major problems prevent utilizing an on-chain data storage approach:

1. **Scalability**: The current technology of Blockchain struggles with high volumes of transactions due to its limited scalability, causing high latency and slow transaction processing.
2. **Cost**: Storing data on a Blockchain network incurs high transaction and storage costs as every piece of data is considered as a separate transaction, and every piece of data is replicated at every Blockchain node.
3. **Data Privacy**: In a Blockchain, data are visible to everyone, which might not be ideal for sensitive information.
4. **Data size limitations**: Blockchains have limited data storage capacity, making it challenging to store large amounts of data.
5. **Interoperability**: Interoperability between different Blockchain networks can be challenging, which could make it difficult to transfer data between networks.

Additionally, due to Blockchain's immutable nature, once data are added to a Blockchain, they are extremely difficult to modify or remove, which could be problematic for some scenarios involving PII data where "right to erasure" should be respected. Therefore, the only viable solution is to store data in a decentralized off-chain data storage while keeping only some integrity ensuring information (e.g. data hashes) on-chain.

2.2 Trustworthy Data Sharing

2.2.1 SOTA including Comparison

Trustworthiness

It will focus on the design and development of a dynamically configurable trustworthiness module based on a set of metrics including characteristics, attributes, and properties such as behavioral, transactional and contextual characteristics. The definition and elaboration of trustworthiness will be based on rules over collected security information, in-line with standards such as IEC 62443 and the ISO 27000 family of standards. Metrics of trustworthiness will be used to drive various technical results of the project such as how and when to filter data but also in the scope of the negotiation and harmonization of diverse policies between heterogeneous data platforms.

Assessing Trustworthiness is a continuous activity in of human life. Trustworthiness is not well defined and the levels of trustworthiness are often assigned in a vague manner. However, in the world of technical systems, communicating and interacting beyond controlled boundaries, such as vehicles or companies, requires Trustworthiness to be assessed based on sound models, metrices and specifications.

Several viewpoints have been defined in the area of industrial IoT especially within the context of data sharing and secure communication across entities. The industrial internet consortium defines Trustworthiness as fulfilling expected requirements in a safe way: *"A satisfactory level of* ***confidence can be established*** *and the* ***partner system*** *(be that a sensor, a machine, or a factory)* ***is what it claims to be, fulfils its tasks and not endangers the business partners*** *by introducing malicious components into the network.*" The Platform Industry 4.0 has a similar view but focuses on Trustworthiness as a quality KPI: *"The term 'trustworthiness' is used to describe the* ***quality of existing and future relationships between companies, people, systems, and components****. A trustworthy system ensures that all of its components* ***behave in an expected manner****."* Finally, in a joint statement the Platform Industry and the Robot Revolution initiative specify *"For* ***supply/value chain security and risk management****, the term 'Trustworthiness' corresponds to the supplier's ability to* ***meet the expectations*** *of the potential contract partner* ***in a verifiable way****"* [3].

Technically, Trustworthiness has been discussed especially in this context of IoT, e.g. industrial Internet / Industry 4.0, and autonomous systems. And the Industrial Internet Consortium laid the foundation of a Trustworthiness model in several papers.

Within the IIC-based model Trustworthiness is a compound of the so-called System Characteristics Security, Reliability, Resilience, Privacy, Safety as illustrated by the Trustworthiness Radar [3]. Based on the discussions with the Japanese Robot Revolution Initiative and the German Platform Industry 4.0 the set of system characteristics was extended, e.g. by Integrity, and a selection of system characteristics was considered. Finally, standardization was pushed forward[2] [3].

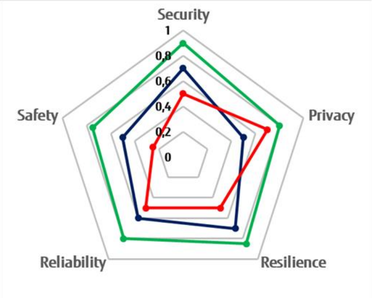


Figure 1- IIC Trustworthiness Radar

[2] EFFRA Innovation Portal - ISO/IEC 30149 - Internet of things (IoT) -Trustworthiness framework https://portal.effra.eu/result/show/4086

[3] ISO/IEC AWI 30149 Internet of things (IoT) — Trustworthiness framework https://www.iso.org/standard/53269.html

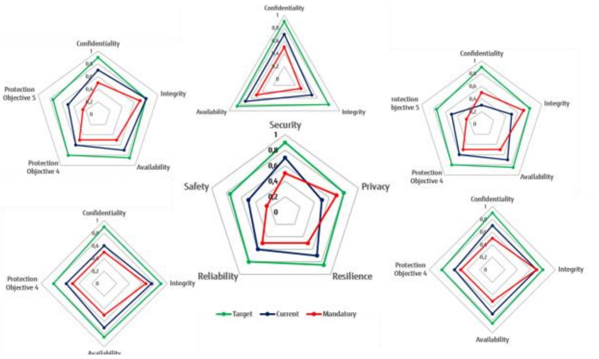


Figure 2- IIC Trustworthiness Radar

In parallel to these discussions within the HORIZON 2020 project SecureIoT the original IIC-model was a bit more formalized as a staggered evaluation of a weighted sum of characteristics consisting of again weighted sums of attributes, and degree of fulfilment of related properties. First metrics for quantification and measurement were outlined in [4]. Trust levels like the security levels, which are described in the IEC 62444, may illustrate the complexity of the specification of Trustworthiness as this depends on the intended use and the selection of the measures. Such levels may be considered as maturity levels of the IoT system related to each system characteristic. With this in mind, the work in that project showed, that a technical evaluation of Trustworthiness requires further research on quantification and measurement of standards, policies and legal descriptions in the application areas.

Another stream of research considered Trustworthiness in the context of autonomous systems. In this area of autonomous vehicles, non-functional aspects, like Availability, Usability, Ethic s, Legal Compliance and Robustness were introduced [5]. Such "emerging" characteristics were rendered not being implementable in the autonomous systems itself but should be considered in a Trustworthiness concept before a system implementation.

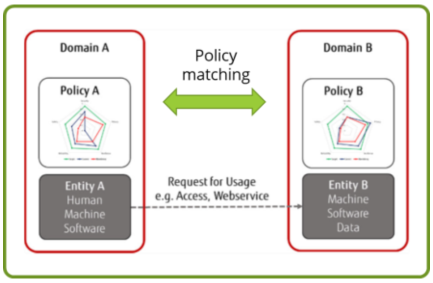


Figure 3 Example policy matching

In both cases of a quantified model as a basis for continuous evaluation of measurable Trustworthiness or as part of a pre-development analysis, the objectives of the characteristics, which represent policies, need to be aligned, and their interdependencies must be resolved free of conflicts. Moreover, characteristics, attributes and properties need to be selected with care and perhaps dynamically.

The quantification and evaluation of the characteristics' attributes and their properties shall take existing methods of assuring trust, such as standards or compliance guidelines, into account. This requires the specification of appropriate metrics, which may include monitoring of device behaviour instead of fixed catalogues. This way, this task is related to tasks T4.1 Self-sovereign Identity Management and T4.4 Device Continuous Behavioural Authentication. Moreover, mechanisms for non-matching policies need to be defined. Metrics, policies as well as conflict resolution probably depend on business requirements and may differ for the different use cases. These gaps shall be tackled within the TANGO project.

Ubiquitous Personal Context Vectors (UPCV)

This task will exploit previous research of so called Ubiquitous Personal Context Vectors (UPCV), a scalable collaborative filtering technology [6], for e.g., service personalization, general recommendations, and advertising. In the context of multiple users and multiple vendors (e.g. shops, hotels or travel agencies), each party can control their own data, which discloses nothing about the others. Therefore, vendors can share or trade their data without breaching users' privacy. Despite of being designed for distributed computing, previous research has only implemented a centralized architecture; distributed architectures with ecosystem aspects have this far only being proposed theoretically [7]. Due to its properties, UPCV is especially applicable in environments where there are several vendors providing same or similar items (e.g. goods, services or attractions) for groups of users. This exploitation will be studied in applicable TANGO pilots at later stages of the project.

Privacy Risk Assessment

Last, Federated Learning techniques will be exploited to allow data, knowledge and insight sharing among users and organizations.

Although Federated Learning favours privacy by minimising data footprint in the network, it is neither immune to attacks nor are current supporting technologies mature enough to be expected to address all privacy issues. From its inception, Federated Learning aims to guarantee users' privacy by using local training model parameters instead of actual data. However, recent research proves that adversaries can partially reveal each participant's training data and gather information about them. In such inference attacks, the adversary misuses the global model to get information on the training data of the users. Some examples on inference attacks on Federated Learning are Membership inference attacks [8] [9] [10] [11] [12] Unintentional data leakage [13] [14] [15] and GAN-based inference attacks [16] [17]. The state-of-the-art algorithms to enhance privacy-preserving and to mitigate threats in federated learning could be broadly categorised into Secure Multi-party Computation and Differential Privacy. Secure Multi-party Computation (SMPC) was introduced to secure the inputs of multiple participants while they jointly compute a model [18] [19]. SMPC secures communication using cryptographic methods. The authors in [20] combined encryption with asynchronous stochastic gradient descent to prevent data leakage from client updates at the central server. The authors in [21] combined homomorphic encryption and differential privacy to mitigate the risk of client data exposure. Focusing on the privacy issues of Federated Learning, this paper implemented client level differential privacy and encrypted the model updates. SMC presents itself as a promising solution, but the balance between efficacy and privacy is a key challenge. Differential Privacy, on the other hand, adds noise to sensitive personal data in order to preserve privacy. In this technique, the data privacy is preserved but the statistical quality of data is lost. In Federated Learning, differential privacy is used to add noise to participant's uploaded parameters. The authors in [22] use Differential Privacy to make GAN-based attacks inefficient in inferencing training data of users in deep learning networks. In a similar approach, [23] [21] combined secure multiparty computation and differential privacy to achieve a secured Federated Learning model with high accuracy. [24] [25] presents a list of algorithms with the performance analysis that could be used for Federated Learning with differential privacy.

2.2.2 Innovation through TANGO

Trustworthiness

TANGO activities will focus on the design and development of a dynamically configurable trustworthiness module based on a set of metrics including characteristics, attributes and properties such as behavioural, transactional and contextual characteristics. The definition and elaboration of trustworthiness will be based on rules over collected security information, in-line with standards such as IEC 62443 and the ISO 27000 family of standards. Metrics of trustworthiness will be used to drive various technical results of the project, such as how and when to filter data, but also in the scope of the negotiation and harmonization of diverse policies between heterogeneous data platforms.

During the project, the existing approach of a Trustworthiness model will be generalized and applied to Data Exchange in the TANGO use cases domains. This will include the definition of appropriate characteristics, attributes, and properties as well as metrices. Specifically, behavioural, transactional and contextual characteristics will be taken into account. They will leverage existing standards in the use case areas, e.g. the ISO 27000 family or IEC62443. The evaluation of Trustworthiness will facilitate the communication and data exchange amongst a wide variety of entities in the TANGO use cases.

UPCV

Consequently, we will demonstrate the privacy-preserving technology for user-item and item-item recommendations and advertising: Vendors will get an advantage of creating 'Google Ads' or 'Facebook Ads' like recommendations independently, without using either these nor other global platforms. This would particularly apply for Smart Hospitality and Retailer pilots, enabling local advertising ecosystems; decisions in these to be made.

In addition, vendors may also present their offerings in personalized relevance for each user individually, find users for empty seats (item-user) or forming groups of users with similar behaviour (user-user), the latter two modes require specific user consent.
Since the technology collects user information in privacy-protecting format, trading behavioural information will also be possible without violating data protection regulations.

2.3 Confidentiality and Privacy by Design

2.3.1 SOTA including Comparison

The true value of data does not arise from connecting devices, but from the use of data that are produced, and how they can be processed to obtain information, and finally knowledge, to offer more valuable services to society. Consequently, it is necessary to move towards data-centric approaches, where security and privacy are tackled in a holistic manner. To protect their rights, users must be empowered with mechanisms to control how their data are shared, to whom, and under what circumstances. In this sense, the Privacy by design [26] principles are relevant: proactive protection instead of remedial action violations happen; default setting is privacy; privacy embedded into the design; full functionality with privacy protection; privacy protection through the complete lifecycle of the data; visibility and transparency; and respect for user privacy. The implementation of these principles, e.g., protecting data throughout its whole lifecycle, is specially challenging in a data sharing ecosystem such as the one considered in this project.

Some of these challenges are provided by ENISA [27], where the use of different anonymization techniques and cryptographic schemes are proposed to ensure the control of how information can be disseminated within such ecosystem. In particular, in order to reconcile the conflicting requirements between privacy and data maximization, advanced techniques need to be applied, for example functional or homomorphic encryption techniques. Works like [28] show the applicability of homomorphic encryption for securing data sharing. However, it limits the data analysis to specific operations (mean, sum, and so on) establishes a central database for access control.

Another type of privacy approach consists in treating the data in advance, so sensitive data are not revealed. For instance, [29] proposes a pseudonym strategy to enable location privacy. Because of its application to vehicular network, the particularity versus other location privacy approaches is that, because of its application to vehicular networks, the change of pseudonym happens when reaching a roadside unit, instead of when a certain number of elements coincide in a specific place. Similarly, [30] presents an algorithm based on minimum instance disclosure risk generalization that aggregates random samplings in groups to preserve privacy. It refers to basic principles such as k-anonymity, l-diversity or t-closeness while preserving correlation information. Another topic related to privacy of individuals comes from applying advanced identity management techniques, as recounted in Section 4.1.

A different matter, and a key topic for achieving privacy by design, is protecting access to data. Particularly, it is necessary to give users the means to establish who can access their data and how. For this, advanced cryptographic techniques can be considered. In functional encryption [31] a trusted entity generates keys for encryptors and decryptors so that the latter can *decrypt* ciphertext, learning a function of the plaintext (i.e., not necessarily the original plaintext). For current applications, only linear or quadratic polynomial functions can be applied. As a subtype, Key-Policy Attribute-Based Encryption (KP-ABE) [32] allows encryptors to encrypt data based on identity attributes.

In a more user centric approach, Sticky Policies [33] present a great opportunity to give users complete control over access to their data throughout its complete lifecycle, even when used in a data sharing platform. They can be instantiated through Ciphertext-Policy Attribute-Based Encryption [34] (CP-ABE), which, in contrast with KP-ABE, allows to attach a policy to a ciphertext so that only entities that have identity attributes (and corresponding cryptographic key) that fulfil that policy can decrypt the message. Figure 4 shows an example of how this works, where a policy (expressed through a logic statement about credentials) is attached to a ciphertext, so that only users that fulfil the policy (User3 in this case) can decrypt the material.

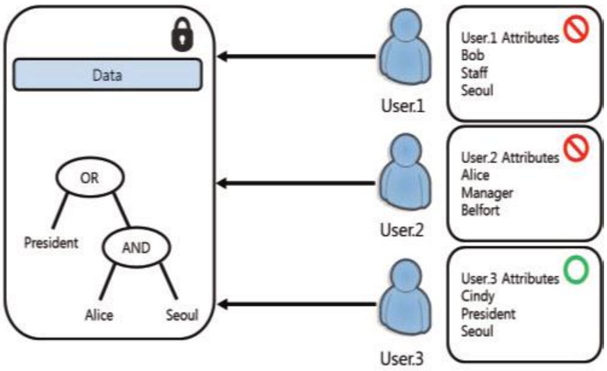


Figure 4 Example of sticky policy through CP-ABE flow

(source: https://www.researchgate.net/figure/CP-ABE-Ciphertext-Policy-Attribute-Based-Encryption_fig2_282500797)

CP-ABE has been applied, for instance, as a means to protect Cyber-Threat Intelligence (CTI) and other security data. For example, [35] proposes the use of CP-ABE to encrypt the extension of the manufacturer’s usage description, achieving very reduced access control, or the system based on CP-ABE to send CTI data between organizations [36].

Apart from the simple functionality application, CP-ABE schemes are introducing extra capabilities suitable for specific scenarios. In [37], the policy for decryption is partially hidden: only the required attributes are revealed, but not their values. It also introduces a feature to trace keys to their original owner, for possible missuses. However, it introduces additional overhead, and does not hide the policy completely, which may leak potentially sensitive information. For its part, MyData[4] allows “revocation” of access, which translates to an update of the policy associated with the ciphertext stored in the sharing server. However, this introduces an efficiency trade-off, and the formal security is not as strong as desirable (not proven against chosen repayable ciphertext).

All in all, these techniques should be further investigated to achieve secure, privacy-respecting, and efficient solutions which are integrated with other components that allow holistic management (e.g., identity management, which would need to abide to privacy principles too for a comprehensive approach) and higher-level methodologies, such as the use of Privacy Impact Assessment tools [5], so data owners are able to manage the lifecycle of their data under a common legal framework. Lastly, while these techniques are practical in many scenarios, their efficiency must be taken into account, especially when working with many attributes. ON general purpose computers, operations may take fractions of a second when the number of attributes is small and even when they start growing (50-100), but the linear increase means that in cases with more attributes, execution time will get into the seconds landmark [6]. In constrained devices, plain CP-ABE schemes might become impractical when the number of attributes grows, taking multiple seconds even with few attributes, and other magnitudes like CPU usage need to be considered[7]. In summary, it is necessary to tackle the gap between the various existing techniques, which fail to fulfil one or various requirements (advanced privacy features, efficiency, short scope), and practical applications where confidentiality and privacy by design are holistically tackled, from empowered user control and GDPR compliance, to technological enforcement of secure and user-centric data sharing. Additionally, such a system would need to not only consider generic scenarios, but also

[4] MyData, *MyData,* https://mydata.org, retrieved 2022-12-20

[5] Evernym, *Evernym,* https://www.evernym.com, retrieved 2022-12-20

[6] Sovrin, *Sovrin*, https://sovrin.org/, retrieved 2022-12-20

[7] Serto, *Serto,* https://www.serto.id/, retrieved 2022-12-20

others where their characteristics (e.g., time or computation constraints) have to lead to adaptation of the solution.

2.3.2 Innovation through TANGO

Supporting end-user empowerment for security and privacy is one of the key objectives of this innovation. Towards this end, TANGO will cover governance aspects to guarantee that end-users are the owners of their data throughout its lifecycle. In particular, it will address security and privacy challenges arising from the integration of the information coming from different sources and devices into central data management platforms. On the one hand, users will be empowered with tools to control how the information is shared. TANGO will address user consent aspects through the use and extension of policy-based approaches to access control. The GDPR regulation will be considered as a legal umbrella for the data sharing ecosystem. Additionally, other aspects will be considered, such as the interaction between data sharing tools and other privacy-related elements (e.g. identity management to achieve a comprehensive private-by-design solution), tackling the difficulties of achieving privacy-by-design in data sharing environments. On the other hand, the work will be complemented with the application of mechanisms based on sticky policies. This will allow TANGO to give sharing entities, whether end-users or businesses, the power to control how and with whom their data are shared, and to be confident that this will be enforced throughout the lifetime of the data on the sharing platform. The application of sticky policies in the project will consider the reflection of expressive "regular-language-like" policy, rather than a simple set of mathematical symbols (e.g., A or B and C), into encryption/decryption and signing to support data sharing, and assist users in protecting their data, avoiding the issues of cumbersome use of the technology that most previous solutions presented. It will also consider the complexity of the algorithms applied and the characteristics of the application scenarios. Thus, it will operate quickly and efficiently, while ensuring security and ease of use. TANGO will also consider the possibility to "hide" the sharing policy from unrelated data parties, maximising the privacy of the data owner and users.

2.4 Self-encryption and Decryption Techniques with Multi-Factor Information Recovery Mechanisms

2.4.1 SOTA including Comparison

Self-encryption is a method of data protection in which the device or system where the data is stored is responsible for both encrypting and decrypting the data, rather than relying on a separate system or software to do so. Self-encryption was introduced as a method of encrypting files that does not requires user intervention or passwords. This algorithm can be used for local encryption of files, whose encrypted chunks will be later uploaded to a cloud-based storage or to a distributed file system (e.g., IPFS [38]). The concept of self-encryption was first introduced by Yu and Ku [39]. The approach of the original paper involves converting a file into a bit stream, extract the key by randomly selecting bits from the stream and then doing the encryption using that key. After the encryption the key and the encrypted file should be stored separately, e.g., the key can be stored locally, while the encrypted file can be sent to a server. The original self-encryption scheme was then extended by the authors in [40]. According to this study, the main improvement is in dividing the plaintext and ciphertext into 1024-bit chunks at XOR process and using the date when encryption process starts as a seed. Their modification also adds the database for the key management function. Storing the key and the encrypted chunks in separate places makes it computationally infeasible to recover the original data. The later industrial adaptation of the self-encryption scheme happened when a team lead by David Irvine made self-encryption the core of his company's (MaidSafe) product - SAFE Network[8]. Irvine's implementation of the self-encryption scheme will be the basis of the TANGO self-encryption method. A a more detailed explanation of the algorithm implementation is given bellow.

[8] https://maidsafe.net/

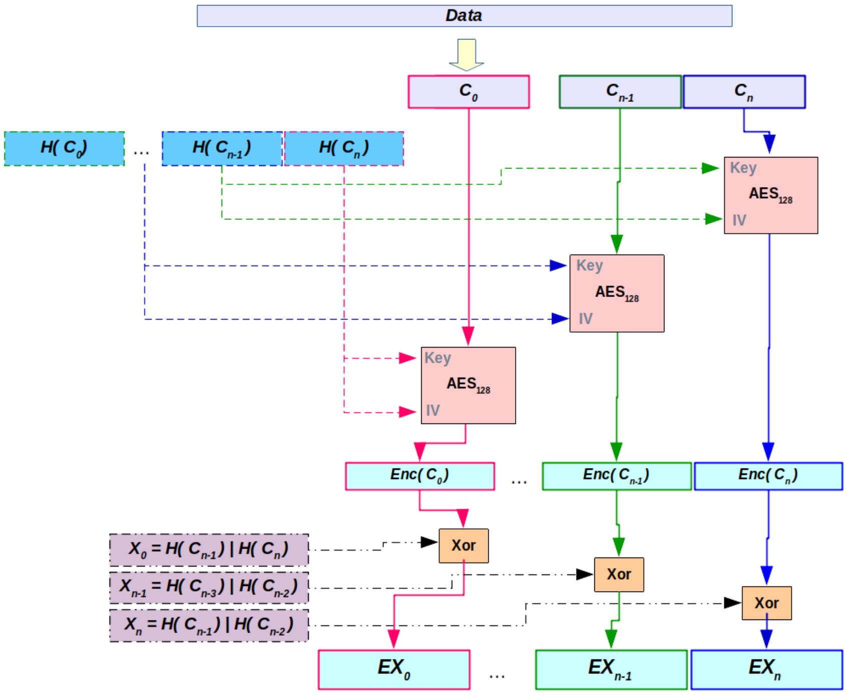


Figure 5 Illustration of self-encryption's principle

The self-encryption designed by David Irvine contains three main steps (see Fig. 6).

1. The first step is the data chunk creation, in which the initial data (plaintext) is divided into identical data chunks (***C_0***, …, ***C_{n-1}***, ***C_n***).
2. The second step contains three main phases.
   I. Calculus of the cryptographic hash value of each data chunk, which is the unique representation or the so-called digital fingerprint [41] of the chunk. In practice, the hash value is at least 256 bits long, and can be calculated with common function families such [42], Keccak [43], or Blake2 [44].
   II. Next, the second phase of this step is the generation of the secret keys and the initialization vectors for chunks' encryption. The AES-128 block cipher [45] is used to encrypt all data chunks, which requires a secret key (Key) of 16 bytes and an initialization vector (IV) of 16 bytes. The hash values computed in the previous phase are used to create the initialization vectors and keys. For encrypting chunk (***C_n***), the previous chunk's (***C_{n-1}***) hash is required for the AES block cipher function. The first 16 bytes of the previous chunk's (***C_{n-1}***) hash value serves as the key and the last 16 bytes as the initialization vector for the AES block cipher. Also, XOR obfuscation values (***X_0***, …, ***X_{n-1}***, ***X_n***) must also be computed for all chunks at this same phase. The obfuscation value corresponding to chunk (***C_n***) is determined by concatenating the hash of the current chunk (***C_n***) and the hash of the first of the two previous chunks (***C_{n-2}***).
   III. Encryption with AES of the chunks using the corresponding keys and initialization vectors. After obtaining the encrypted chunks, obfuscation is applied to each of the AES encrypted chunks (e.g., **Enc**(***C_n***)). In this last step of the self-encryption, the encrypted chunks are XOR-ed with the obfuscation values computed in the previous step in order to obtain the final encrypted chunks (e.g., ***EX_n***). It must be noted that one obfuscation value (e.g., ***X_n***) is 64 bytes long, contrarily to the size of one AES encrypted chunk, which can be higher than 64 bytes. If the AES encrypted chunk's size exceeds 64 bytes, the obfuscation value is rotationally padded by itself until it achieves the same length as the AES encoded chunk's size.

3. The last step is the data map creation, which can be represented as a table. The left column of the table contains the hashes of the data chunks, which are required to determine the keys and initialization vectors. The data chunks' hashes can be considered the secret keys for the self-encrypted data, which means that these values should not be publicly shared. The right column of the table contains the hashes of the final encrypted chunks. The final encrypted data chunks' hash values can be considered the pointer to the final encrypted data chunks allowing their storage in different locations. Storing data chunks in different locations makes it harder to retrieve the totality of the chunks.

Self-encryption technology offers several key benefits, including improved security, improved performance, and ease of implementation and use. By encrypting data on the device or system where it is stored, self-encryption technology helps to protect against attacks and interception, making the data less likely to be exposed. Additionally, since the host system does not need to perform encryption and decryption processes, the overall performance of the system can be improved. Self-encryption technology is also easy to implement and use [46]. To retrieve the original data (plaintext) in self-encryption, all final encrypted data chunks and the keys (hashes of the data chunks) are required. If one of the final encrypted chunks cannot be adequately decrypted or one of the keys is missing, the concatenation of the decrypted chunks will not return the original data.

ID-based encryption is a method of data protection that uses a user's identity as the basis for creating a cryptographic key. One of the main advantages of ID-based encryption is that it allows users to access encrypted data using their ID, rather than a traditional password or passphrase. This can be more convenient for users, as they do not need to remember multiple passwords or passphrases [47]. Additionally, multi-factor recovery mechanism or multi-factor authentication (MFA) is a security measure that requires users to provide multiple forms of identification when accessing a system or service. This helps to protect against unauthorized access by making it difficult for attackers to impersonate a legitimate user. MFA typically involves the use of at least two different forms of identification, such as a password and a one-time code sent to a phone or email [48]. This ensures that even if an attacker obtains a user's password, they cannot access the system or service without also having access to the other form of identification. There are various methods of MFA, including hardware tokens, biometric authentication, and app-based authentication. The specific method used depends on the organization's needs and resources.

It should be noted that combining MFA with self-encryption offers several benefits, including increased security, improved authentication, enhanced privacy, and greater control. For example, MFA adds an extra layer of security by requiring multiple forms of identification to access encrypted data, making it harder for attackers to gain access. It also helps to ensure that only authorized users can access the encrypted data, protecting against unauthorized access and data breaches. Self-encryption protects the privacy of data by encrypting it and making it unreadable without the decryption key. When combined with MFA, this provides an additional level of protection for the data and ensures that it is only accessed by authorized users. Using both MFA and self-encryption allows organizations to have greater control over access to the encrypted data, as they can manage the encryption keys and the forms of identification required for MFA, and revoke access to the data if necessary [49].

2.4.2 Innovation through TANGO

TANGO aims to address the shortcomings in SOTA e.g., identity unlinkability, centralised mode, static key, and lack of interface with other secure tools, and introduces a new self encryption/decryption approach that uses data to be encrypted along with random seed and identity as the encryption key, distributing the keys (via secret sharing) across trusted nodes. There have been some relevant studies in the field of self-encryption and MFA. However, existing self-encryption solutions do not have an ID-based solution to facilitate user access. In TANGO, a new extension to the self-encryption scheme will be deployed by integrating the data owner's identity into the traditional self-encryption process. Such an ID-based integration allows us to preserve ownership of the original file and link it to the person/entity who originally uploaded it [50]. TANGO's self-encryption mechanisms provide strong security guarantees that decryption of a file is computationally impossible under the condition that the encrypted file and the key are safely stored. Besides, combining MFA with self-encryption can provide an extra robust security layer that helps to protect against unauthorized access and data breaches. It is an effective way to ensure that sensitive data remains secure and is only accessed by authorized users.

2.5 ePrivacy, Mechanisms, Protocols and Processes

This task is not *per se* a technical task of the TANGO project. T3.5 will support technical work carried out in the project by developing protocols where personal data and non-personal data containing business-sensitive information will be treated in the same manner, taking into account legal requirements, legislative developments and developments in jurisprudence. Work undertaken under this task will contribute towards setting the framework for achieving a standardized level of security for users, regardless of the context where the technological solution developed will be exploited.

Given its nature, T3.5 does not lend itself to a SOTA and GAP analysis as that carried out for technical tasks of the project and presented in this Deliverable. This task will be reported in Deliverable D3.1 – "Distributed Privacy-preserving Data Management and Storage Intermediate version" (M17) and Deliverable D3.2 – "Distributed Privacy-preserving Data Management and Storage Final Version" (M30).

2.6 Secure data sharing based on IDS and GaiaX standards

2.6.1 SOTA including Comparison

Approximately 80% of industrial data is never shared[9]. Furthermore, similar unused potential is witnessed in all domains. The reason is a lack of interoperability of different systems together with a lack of trust of identities about how the data is handled and used. For these reasons, the data economy is evolving towards federated data-sharing ecosystems, where data sovereignty is ensured [i.e. the ability of a natural or legal person to sovereignly and exclusively decide over the usage of their data as an economic asset]. They are the so-called "**data spaces**". Different initiatives contribute to the definition and the creation of data spaces, but two forerunners are IDSA (International Data Spaces Association) and Gaia-X.

IDSA stands for International Data Spaces Association and is a not-for-profit association which was founded in 2016 in Dortmund Germany. It is a global standardization organization counting 130+ members from 22 countries. Its aim is to create a standard for data sharing in data spaces, tackling both technical, operational, and legal aspects. More information can be found on the IDSA website[10].

Gaia-X was launched in 2019 with the goal of developing a trustworthy and sovereign digital infrastructure for Europe, Gaia-X is a member-regulated digital ecosystem, and it was originally created as a project by the German Ministry of Economic Affairs and Energy, immediately supported by the French Ministry of Economy during the summer of 2019[11].

A first common press release was issued in October 2019[12], while a first Franco-German Position Paper was published February 18, 2020[13].

Following an announcement in June 2020 that an international non-profit organization, Gaia-X Association has been established in the form of an international non-profit association under Belgian law (in French: Association Internationale sans but lucratif, short: AISBL) with its headquarters in Brussels.

[9] https://ec.europa.eu/commission/presscorner/detail/en/ip_22_1113
[10] https://internationaldataspaces.org/
[11] https://www.data-infrastructure.eu/GAIAX/Redaktion/EN/Video/20200604-GAIA-X-Ministerial-talk/20200604-ministerial-talk.html
[12] https://www.data-infrastructure.eu/GAIAX/Redaktion/EN/Press-Releases/20191029-press-release-on-franco-german-common-work-on-a-secure-and-trustworthy-data-infrastructure.html
[13] https://www.bmwk.de/Redaktion/DE/Downloads/F/franco-german-position-on-gaia-x.pdf?__blob=publicationFile&v=10

Many member organizations from all across the world make up the Gaia-X European Association for Data and Cloud AISBL. Companies, associations, research institutions, administrations, and government officials have come together to work on a groundbreaking project for the digital economy of tomorrow: Gaia-X. The original 22 member organizations have now blossomed to more than 340+ member organizations**.**

Gaia-X is an initiative that develops, based on European values, a digital governance that can be applied to any existing cloud/ edge technology stack to obtain transparency, controllability, portability and interoperability across data and services.

Its goal is innovation through digital sovereignty. This is achieved by establishing a decentralised ecosystem in which data is made available, collated, and shared in a trustworthy environment where users always retain sovereignty over their data. Thus, outcome will not be a cloud, but a federated system linking many cloud service providers and users together in a transparent environment that will drive the European data economy of tomorrow.

A **first comparison carried out in 2021**[14] showed that GAIA-X was not as mature as the International Data Spaces (IDS) initiative, but followed the same vision of proliferating data sovereignty and creating an ecosystem of trust for data sharing. In fact, the IDS standard offers various concepts and solutions that contribute to the overall vision of GAIA-X and to the concrete GAIA-X architecture demands. Both initiatives have progressed considerably in the last year. In 2021 IDSA, Gaia-X and other two forerunners in the context of data spaces decided to formally team up and in September 2021 the **Data Spaces Business Alliance was created[15].**

One concrete goal of the DSBA is the alignment of the IDS and Gaia-X architectures. A first result in this direction has been already achieved in September 2022 with the publication of the **Technical Convergence paper**[16]. This represents a remarkable milestone for the harmonization of these complementary approaches and for the proliferation of data spaces.

The IDS framework is described in the Reference Architecture Model (RAM), which is accessible by everyone on the IDSA GitHub[17]. From the RAM, a first official standard has been created: the DIN SPEC 27070 "Requirements and reference architecture of a security gateway for the exchange of industry data and services". The standard is domain agnostic (i.e. it is not bound to any specific domain) and technology agnostic (i.e. it is not related to any specific technology). This means that it can be adopted by anyone interested to develop their own data space or to participate in one.

The standard is currently moving from a research environment to market ready solutions. In fact, the IDS Certification for components and operational environments has been launched in May 2022, and several use cases have been already set-up or are being created in a variety of domains. More than 70 use cases are depicted on the Data Space Radar on the IDSA Website[18].

To ease adoption of the standard and the creation of new data spaces, several component implementations are available on the IDSA Github[19], among which also an open-source testbed[20].

Gaia-X connects isolated data sources in organisations and competing cloud services from different providers in an ecosystem. This enables companies, organisations, authorities and citizens to exchange data securely and, above all, sovereignly. This means that they retain full control over their data and no longer risk of becoming technically dependent on individual platform providers. For this ecosystem,

[14] https://internationaldataspaces.org/wp-content/uploads/dlm_uploads/IDSA-Position-Paper-GAIA-X-and-IDS.pdf
[15] https://data-spaces-business-alliance.eu/
[16] https://internationaldataspaces.org/dsba-releases-technical-convergence-discussion-document/
[17] https://docs.internationaldataspaces.org/ids-ram-4/

[18] https://internationaldataspaces.org/adopt/data-space-radar/
[19] https://github.com/International-Data-Spaces-Association
[20] https://github.com/International-Data-Spaces-Association/IDS-testbed

Gaia-X develops rules, standards, blueprints and technology for data sovereignty, interoperability and freedom of choice in digital services.

Gaia-X fosters software communities and facilitates the interoperability between data spaces.

The Association presents below its Gaia-X Framework, which builds upon the evolution of the X-Model and enables the transition from disjoint data & infrastructure ecosystems, to composable, interoperable & portable cross-sector data sets and services[21]. In addition, Gaia-X focuses on three deliverables: Specifications, Code, Labels[22].

The Gaia-X Federation Services, or GXFS for short, provide the technical foundation for the European data ecosystem. The GXFS are a software framework under an open-source licence, i.e. a collection of open-source programme modules. They can be used to build and manage cloud-based data ecosystems – called federations.

The GXFS provide the technology building blocks and ensure that different federations function according to the same values and that the data ecosystems can later cooperate technically. In doing so, we're not trying to reinvent the wheel but are using existing techniques, solutions and services wherever possible.

The following technical specifications have been drawn up in Spring 2021 and were awarded in an EU-wide tender[23]. On the basis of technical specifications, services based on open-source code (APL2) are being developed through community-driven work on the open-source code via Gitlab[24].

2.6.2 Innovation through TANGO

Considering the relevance and the maturity of both the IDSA standard and the Gaia-X framework, their resources, and the efforts from the whole data space community (both members and non-members of the IDSA and Gaia-X associations), TANGO will use them for project developments.

The purpose of including **IDS** in the project is to enable the TANGO to be potentially interoperable with all present and future data spaces, adding data sources and services to its ecosystem in a sovereign manner. The goal is to adopt the IDS standard to ensure secure data sharing to enable the data provider to maintain control over how the data is accessed and used (e.g. identity management, data usage policies, a certification process for software and operational environments,…)

On the other hand, **Gaia-X** makes it possible to connect isolated data sources in organisations and competing cloud services from different providers in an ecosystem. This facilitates secure data exchange and, especially, sovereignly, so users retain full control over their data and no longer risk becoming technically dependent on individual platform providers.

GXFS can be used to build and manage cloud-based data ecosystems – such as federations. “Federations" by default are just organisational/legal constructs in comparison to Data Spaces which are technical/legal constructs that are linked to each other.

Gaia-X follows –typical for Europe – a federal approach. This means that there will be no centralised European cloud and data platform for all. The technical requirements in the EU countries and economic sectors are too different for this. Moreover, no one could build a new hyper scaler on a greenfield site anymore. Instead, user companies and IT service providers design their own data ecosystems according to their needs along industries, value chains, research topics or geographical spaces. To this end, they organise themselves into self-governing Gaia-X federations. They determine and monitor the rules and technical specifications for their members.

The data space approach can bring **value to TANGO** in multiple ways. The IDS standard is an overarching, domain-agnostic standard which it can be adopted in different sectors. More than 75 use

[21] https://gaia-x.eu/gaia-x-framework/
[22] https://gaia-x.eu/what-is-gaia-x/deliverables/
[23] https://www.gxfs.eu/specifications/
[24] https://gitlab.com/gaia-x/data-infrastructure-federation-services

cases are listed in the Data Space Radar[25] and can provide source of inspiration for the consortium in a variety of domains.

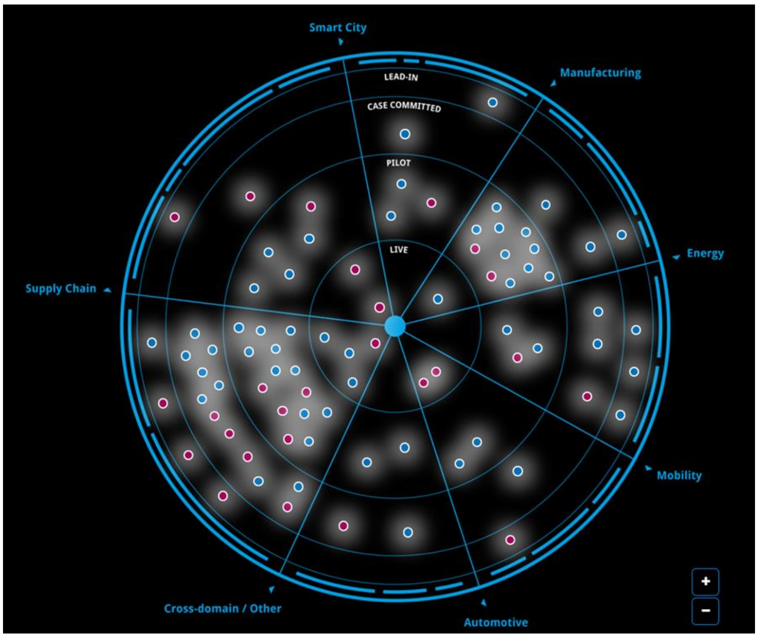


Figure 6 Data Space Radar on the IDSA Website

To provide an example of how the IDS standard can be adopted in TANGO, a few best practices from the manufacturing domain are reported here below.

- the SCSN data space[26]
- the Catena-X data space[27]
- the Mondragon pilot[28]
- Additive Manufacturing case by IBM, thyssenkrupp and Fraunhofer ISST[29]

Herein, the TANGO pilots could leverage IDS to improve vulnerable data operations, also as far as data sharing with suppliers or service providers is concerned. To know more about manufacturing cases leveraging the IDS, please refer to the the IDSA Industrial community (called IDS-I)[30], and in particular to two IDS-I papers, respectively:

- Paper "Data Sovereignty – Critical Success Factor for the Manufacturing Industry" (IDSA Position Paper 1.0, 2021)[31]
- Paper "Data Sovereignty – Requirements Analysis of Manufacturing Use Cases (IDSA Position Paper 1.0, 2022)[32]

[25] https://internationaldataspaces.org/adopt/data-space-radar/

[26] Reference 1: video https://youtu.be/vapiKD3xzbE, reference 2: https://internationaldataspaces.org/the-smart-connected-supplier-network-by-tno/

[27] Reference 1: https://internationaldataspaces.org/catena-x-network-for-cross-company-data-exchange-in-the-automotive-industry-relies-on-ids/, reference 2: https://catena-x.net/en

[28] Reference 1: video https://www.youtube.com/watch?v=yzPNrxSclk8, reference 2: Position Paper pages 32-34 https://internationaldataspaces.org/wp-content/uploads/IDSA-Position-Paper-Data-Sovereignty-Requirements-Analysis-of-Manufacturing-Use-Cases.pdf

[29] https://internationaldataspaces.org/usecases/ibm-thyssenkrupp-fraunhofer/

[30] https://internationaldataspaces.org/make/communities/

[31] https://internationaldataspaces.org/wp-content/uploads/dlm_uploads/IDSA-Position-Paper-Data-Sovereignty%E2%80%93Critical-Success-Factor-for-the-Manufacturing-Industry.pdf

[32] https://internationaldataspaces.org/wp-content/uploads/IDSA-Position-Paper-Data-Sovereignty-Requirements-Analysis-of-Manufacturing-Use-Cases.pdf

The IDS approach has a strong footprint on **manufacturing and supply chain**, but does not limit to it. Other main fields of adoption are in addition **mobility, automotive and energy**. New initiatives are also being started in domains like the **healthcare, rail, tourism, media, agriculture**. A big potential is still to be unlocked in the **financial, retail and public administration sectors**. With its pilots in these domains, TANGO could be a forerunner, positioning itself as a path leader in the context of data spaces in these fields.

As far as Gaia-X and GXFS is concerned, they also can fit any scenario. Therefore, a great innovative potential also lays in leveraging Gaia-X and IDS in joint use cases. As mentioned, the two approaches combine very well, and an alignment of the architectures is ongoing via the DSBA. As described above, a first milestone is the the technical convergence paper. For this reason, concrete experimentation on joint architectures is to start. TANGO could leverage this window of opportunity and gain recognition in the whole data space community bringing the SOTA to the next level with a **joint IDS and Gaia-X use case or demonstrator**.

3 Distributed Trust Management Framework

3.1 Self-sovereign Identity Management

3.1.1 SOTA including Comparison

TANGO will use a privacy-preserving and self-sovereign identity approach, which improves user authentication with added privacy and trust, and empowers end-users with more control over their own personal data. The concept of SSI is illustrated in Figure 7, and is based on the interaction between the actors Issuer (issuing credentials), Holder (holding the credentials, e.g. in their wallet application), and Verifier (verifying the credentials), in addition to the Data registry (e.g. Blockchain) that includes the functions to verify the credentials of the user. In the field of SSI, solutions like MyData[33] Global empower individuals by improving their right to self-determination regarding their personal data, providing interoperable infrastructure for human-centric personal data management and governance. To achieve distributed identity solutions, Blockchain is a typical technology. Sovrin Network[34] is an open-source project that aims to bring the personal control and ease-of-use of analog ID cards to the Internet. Evernym[35] constitutes a practical implementation of the Self-sovereign Identity described at Sovrin. The Sovrin Foundation has formally endorsed the project of ESSIC – the European Self-Sovereign Identity Consortium, which aims to promote interoperability between different identity networks. Another example is Serto[36] by ConsenSys, which can be used by, e.g., artists to establish a decentralized identifier that can be tied to NTFs of their art platform-independently. In Moreno et al. [51], the extension of the infrastructure with a DLT and smart contracts for trusted public data sharing and auditing is recounted.

To achieve the best results in terms of privacy and security, advanced cryptographic techniques are needed. Privacy-preserving Attribute-Based Credentials (p-ABCs) is a technology that can be used to generate tokens where only a subset of attributes (or even partial information about them) are revealed. In recent years, they have been the subject of extensive research. For instance, Camenish et al. [52] introduces distributed p-ABCs based on multi-signatures. Other notable works are Hébant et al. [53], which presents *aggregatable* and *traceable* p-ABCs for accountability in case of credential abuse (using a tracing authority), and Sanders [54], which uses redactable signatures to propose efficient p-ABCs. Issuer-hiding attribute-based credentials, allowing one to hide the precise issuer of a certificate, have been introduced by Bobolz et al. [55]. Efficient and practical schemes have also been suggested by Hanzlik and Slamanig [56].

Another example of the various SSI technologies existing today is **Hyperledger Aries**[37] an open source evolution from Hyperledger Indy that creates a modular and extensible SSI Framework that is completely independent of any Verifiable Data Registry, be it based on DLT or otherwise. Aries has notably led standards-based interfaces through its work in W3C and the Decentralised Identity Foundation (DIF).

[33] MyData https://mydata.org, retrieved 2022-12-20
[34] Sovrin https://sovrin.org/, retrieved 2022-12-20
[35] Evernym https://www.evernym.com, retrieved 2022-12-20
[36] Serto https://www.serto.id/, retrieved 2022-12-20
[37] https://wiki.hyperledger.org/display/TSC/Hyperledger+Aries+Proposal

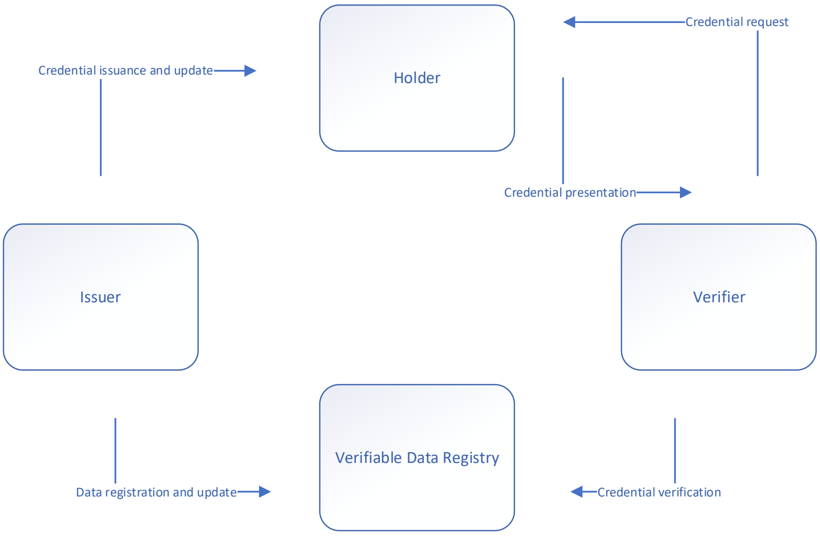


Figure 7 SSI concepts

Digital Identity development in EU:

The EU is working on a new EU Digital Identity Wallet (EU DIW) mobile app, secured by biometrics, that will allow European citizens to use their eID to access public and private services across borders throughout Europe. In this regard, the EC is amending the eIDAS regulation to establish a framework for a European digital identity to enable all EU citizens, residents and businesses to benefit from the new European Digital Identity Wallet, facilitated by qualified electronic attestation of attributes.

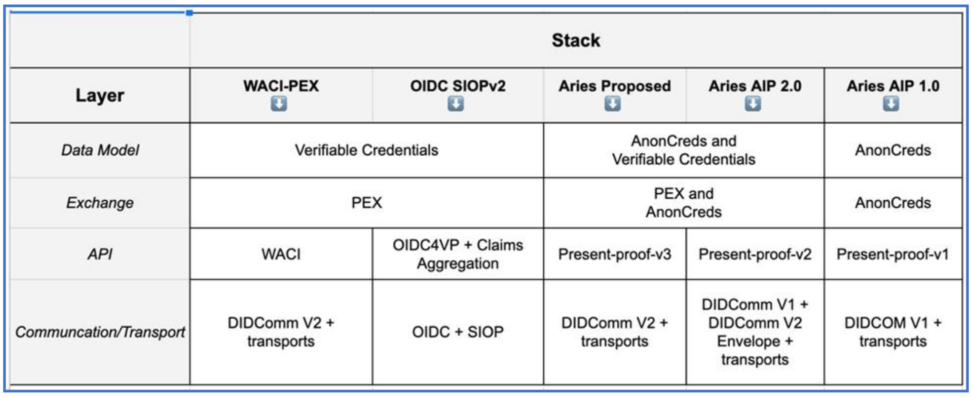
The full technical specification of the EU DIW is as yet not fully published, but is expected that Verifiable Credentials will be employed and that qualified electronic attestation of attributes by qualified providers, as in the current regulation is retained in the regulation.

One of the four digital identity pilots funded by the EC is about to launch in March 2023, the Nordic-Baltic eID Project (NOBID), which will implement a large-scale pilot for a banking payment use case with the EU Digital Wallet with involvement of banks in Financial companies, including banks, in Germany, Norway, Denmark, Italy, and Iceland.

Interoperability

Due to initial SSI developments preceding much of the standards work and differing rival technologies, their implementations were never going to be able to interwork with each other. However, today there is a lot of effort going into interoperability as the standards have matured.

As we can see in the table below from the Decentralized Identity Foundation Interoperability WG there are various protocols supporting different SSI stack approaches for the Verifiable Credentials (VF) data model, exchange, proof presentations and transport.

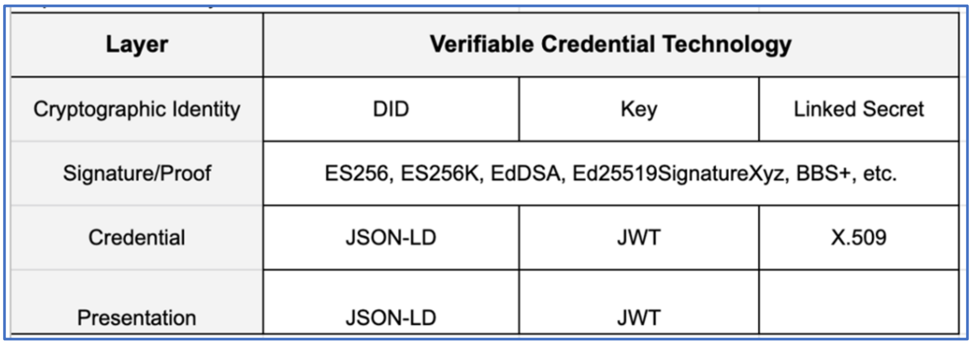


	Stack				
Layer	**WACI-PEX** ⬇	**OIDC SIOPv2** ⬇	**Aries Proposed** ⬇	**Aries AIP 2.0** ⬇	**Aries AIP 1.0** ⬇
Data Model	Verifiable Credentials		AnonCreds and Verifiable Credentials		AnonCreds
Exchange	PEX		PEX and AnonCreds		AnonCreds
API	WACI	OIDC4VP + Claims Aggregation	Present-proof-v3	Present-proof-v2	Present-proof-v1
Communcation/Transport	DIDComm V2 + transports	OIDC + SIOP	DIDComm V2 + transports	DIDComm V1 + DIDComm V2 Envelope + transports	DIDCOM V1 + transports

Figure 8 Protocol support of SSI

The two main layer-2 protocols in SSI for communication between SSI Agents and wallets are DIDComm and SIOP, which have a good vendor support either individually or both of them.

The overview of the technology supported by the Verifiable Credentials is shown in the next table together with the most used cryptographic signature algorithms that are supported today.



Layer	**Verifiable Credential Technology**		
Cryptographic Identity	DID	Key	Linked Secret
Signature/Proof	ES256, ES256K, EdDSA, Ed25519SignatureXyz, BBS+, etc.		
Credential	JSON-LD	JWT	X.509
Presentation	JSON-LD	JWT	

Figure 9 Cryptographic technology

Many different organisations are working on SSI solutions today as well as member state level activities including those at ESSIF.

There are several Self-Sovereign Identity solutions available today based on the evolving standards of Decentralized Identities and Verifiable Credentials. The following is a list of some of these solutions:

- Veramo[38]
- Veres One[39]
- Hyperledger Indy[40]
- Hyperledger Aries[41]
- Jolocom[42]
- MATTR[43]
- SpruceID[44]
- IOTA Identity[45]

[38] https://github.com/uport-project/veramo
[39] https://veres.one
[40] https://www.hyperledger.org/use/hyperledger-indy
[41] https://github.com/hyperledger/aries
[42] https://jolocom.io
[43] https://github.com/mattrglobal
[44] https://github.com/spruceid/ssi
[45] https://github.com/iotaledger/identity.rs/

- AlastriaID[46]

A covering study of SSI ecosystems is presented by Soltani et al. [57]. In the context of EOSC, the Authentication and Authorisation Infrastructure (AAI) is a key concern for the security and trust of any collaboration: it aims to build a foundation for e-science that ensures the long-term availability of aspects of digital identity to build a foundation for e-science that ensures the long-term availability of aspects of digital identity, while enabling the establishment and maintenance of high-trust collaborations with little or no friction for the end user. The EOSC AAI architecture [58] is based on the AARC Blueprint Architecture 2019 [59] and follows the AARC Interoperability guidelines to work with the international community and to meet the evolving needs of EOSC. In the federated environment of EOSC, the AAI is developed in the context of a global marketplace of AAI products and services that typically focus on the consumer-business relationship, where they provide trusted identity information and enable scalable management of roles and entitlements. The Check-in service[47] provided by EGI and available for the TANGO project is an AAI service. Amongst its characteristics are that it is a member of the EOSC federated AAI, compliant with the AARC-BPA architecture, meets all the requirements defined in the EOSC AAI architecture, and implements the EOSC interoperability guidelines.

3.1.2 Innovation through TANGO

TANGO will use a privacy-preserving, distributed, efficient and self-sovereign identity approach, leveraging p-ABC systems, so the authentication of the different users and IoT devices will achieve unlinkability and minimal disclosure in identity data. TANGO will develop obvious properties needed for sustainable personal data management: a) granular personal data management, b) temporal management of personal data disclosures, c) cascaded identity control and d) fingerprinting of the data. This asset will leverage multiple individual IdPs to manage user identities and authentication. It relies on distributed p-ABCs to offer privacy-preserving (minimal disclosure and unlinkability) authentication (presentation of attributes), which may be linked to eIDAS for further identity assurances. A trust framework based on Blockchain to complement the usage of credentials will be introduced. TANGO will progress beyond the state of the art by implementing mature (e.g., embedded databases, symmetric and public key encryption) and nearly mature technologies in a disrupting way, resulting in academic contributions and demonstrable proof-of-concepts.

A leaner implementation of SSI Agents than those that run in the cloud and in wallets is needed to support IoT Devices. TANGO will investigate how to provide more efficient SSI Agents with reduced resource/memory needs, adapted to work with secure instructions from the business app on the device.

In order to support IoT Devices there is needed a leaner implementation of SSI Agents than those that run in the cloud and in wallets, as is supported in the main, today. TANGO will investigate how to provide more efficient SSI Agents with reduced resource/memory needs and that are adapted to work with secure instructions from the business app on the device.

EGI is able to provide a development environment and a production environment managed by Check-in service in which to develop the components for the support of the TANGO SSI model in the EOSC environment.

Regarding the Digital Identity development in EU, it is important to keep track of the current pilot projects:DC4EU[48], and NOBID[49] in terms of their protocol implementation in the EU DIW with use of Verifiable Credentials for Task 4.1, and also Task 4.2 for on-boarding to ensure high level of assurance and also its application in the banking sector.

[46] https://github.com/alastria/alastria-identity

[47] EGI Check-in service: https://www.egi.eu/service/check-in/

[48] https://www.dc4eu.eu

[49] https://www.norden.org/en/project/nordic-baltic-eid-project-nobid

3.2 Seamless Onboarding for Users and Devices

3.2.1 SOTA including Comparison

On 25 May 2018, the General Data Protection Regulation (GDPR) came into force with the sole purpose of protecting the privacy of European citizens. One of the elements defined in the regulation is that users have the right to request the data about them held by an organization in a human-readable format. However, with the lack of security in verifying the user identity, the regulation itself has led to data breaches.

In September 2019, James Pavur a cyber security researcher showcased at Black Hat conference[50], how he sent 150 GDPR requests for data portability, impersonating his fiancé. The requests included only the name, fake email, a real email (from social media) and a phone number. About 24% of the companies provided the requested data with only that input information, leading to a data breach. Some 16% of the companies requested some weak form of ID that could be easily forged, such as a written statement claiming that he was the person he claimed; Pavur chose not to forge the weak ID. About 3% of companies deleted the account without any notice, which led to discontinuing the service and removing the personal data.

Most of the organisations today do not have the expertise and the capacity to handle the required identity verification. A user-centric and privacy preserving identity verification system is not trivial. Initially verifying the identity of a person remotely is a great challenge. Depending on the eIDAS level of assurances, there are different ways that organizations are tackling this challenge, as shown in the table below. However, these are not sufficient to ensure that only the owner has access to the data; data breaches are part of daily headlines.

Type of identity	**Current Situation**	**Security threats**
Account login	*Not all organisations have account login that hold personal data*	*Username/password based authentication can be breached.*
Access to email	*Does not constitute a proof of identity*	*Usually email clients are not password protected*
Device cookie	*Does not constitute a proof of identity*	*Can be acquired from browser*
Government ID	*Organisations do not have the proper means/capacity to store, secure the ID.*	*No cross-check of the document with the actual person.*
Signed Statement	*It is not a strong form of identity.*	*Can be easily forged.*
Knowledge question	*Requires additional knowledge from the organisation.*	*Can be retrieved through social engineering*
Utility bill	*It is not a strong form of identity.*	*Can be easily acquired/forged.*
Phone interview	*It is time-consuming and prone to human error*	*A person easily impersonates someone else*
Credit card number	*Possession of credit card number does not verify the identity of the user.*	*Can be easily acquired physically with social engineering*

Table 1 eIDAS security threats

[50] https://www.blackhat.com/us-19/speakers/James-Pavur.html

End-users do not trust organizations such as a marketing agency to handle sensitive data e.g. a copy of their passport. Thus a coherent solution should have the appropriate infrastructure to handle the identity verification of the end-user. Without the need for the organization to worry about how to store, process, secure and maintain sensitive personal data such as proof-of-identity. It is essential to provide to the end-user the appropriate level of trust regarding the management of the personal data.

On the other side, in order to create a secure on boarding to a particular service, there is a need to on-board both the user and the device through which a person is accessing the particular service. The banking sector, which follows the Payment Service Directive 2 for Strong Customer Authentication, the most common approach followed by the banks and Payment Providers on-board a device is to send and SMS to a smartphone. That SMS is then used generate a unique key that will be used as a digital token to authenticate the device. This approach creates friction for the user considering that they need to manage the process of inputting the SMS [60]. Furthermore, authenticating to a service running on an IoT device has become very challenging [61]. IoT deployment is treated as a programmable infrastructure where multiple services can be deployed. Thus, multiple users can interact with multiple applications that are running on the same IoT device. The fundamental problems that are raised on the IoT domain is the lack of computational power to host sophisticated identity server combined with the problem of an increased attack surface that is exposed by an IoT device.

Furthermore, it is common for IoT devices to perform communication with a service over an unencrypted channels. There are not few cases where eavesdropping was taking place allowing malicious attackers to “listen” on sensitive and personal information that were logged and transmitted by IoT devices. Even in the cases where encryption is used during the transmission process there is no actual way of knowing which particular device is currently connected to the service. Considering that the largest part of end-use energy consumption worldwide is associated with the buildings and the large facilities sector, there is a need for novel and secure-by-design solutions in smart facilities for secure authentication and energy efficient techniques.

3.2.2 Innovation through TANGO

Unlike state-of-the-art techniques, which are prone to security issues and create enormous frustration to end-users when onboarding to a particular service, TANGO will offer novel mechanisms that will allow end-users to seamlessly onboard to particular services with High Level of Assurance based on the eIDAS regulation while also allow IoT devices to easily and securely authenticate to a service. This will allow organisations across all verticals that require to verify their users to perform it through a seamless and secure manner that does require the involvement of a human person in order to ensure that it is the same person that performs the verification with the owner of the identity document. For IoT devices, the technology will ensure that the device that communicates with a particular service is the device the same with the registered device and no tampering has taken place. This will allow the creation of a device registry leveraging encryption tools that will ensure the secure communication between the device and the actual service. Organisations will benefit from the cost reduction due to potential fraud and cyberattacks as well as due to operations as current solutions have increased costs for allowing the onboarding of a person in order to verify their identity.

For end-users, TANGO will offer an AI solution that will verify the identity of a user based on an existing identity document, following a four step process for verifying the authenticity of the document and also confirming that the person in the process of performing the verification is the same as the person that the identity document belongs to. TANGO will help European citizens protect their privacy while helping organizations reduce costs and improve productivity by introducing a user-centric privacy preserving and secure identity verification mechanism. The solution will leverage AI to remotely verify the identity of the user without the need of human validator intervention providing high level of assurance. To ensure user’s security and privacy, the solution will secure the verified identity information through distributed ledger technology, allowing the user to have full control over the data and will only share with the organisation the verification result. The mechanism will allow remote on-boarding of users with Highest Level of Assurance based on eIDAS regulation through a mobile app using a four-step verification process with a proof-of-identity including a) NFC document scan, b) OCR document scan, c) facial comparison including liveness detection and d) cross-validation of the identity document.

For the devices, TANGO will offer a novel solution that ensures the verification process of the device through a common self-registration process that showcases the identity of the device, allowing other TANGO components to identify a device uniquely. The solution will introduce a self-service registry that ensures the privacy and confidentiality of the cryptographic keys that are generated uniquely for the identification of a particular device and the secure communication between the IoT device and the TANGO platform. This way the system is able to prove the ownership of a device through the use of PKI keys that do not require consensus among all the nodes of the network, which is time and energy consuming. This mechanism will allow the interconnection with other protocols and solutions that perform user authentication in order to allow third-parties to perform a fusion of the identification process with various contextual information such as the identity of a user and the identity of a device.

The TANGO User and Device On-boarding mechanism will be split into specific components that will allow ensure the identity verification of the end-users through a mobile SDK while for the on-boarding of the device a collection of components will ensure the authenticity of the device at the server side.

3.3 User Continuous Behavioural Authentication

3.3.1 SOTA including Comparison

Authentication constitutes an important component of online and offline systems, ensuring the access control. The primarily purpose is to ensure that the person trying to access a particular system is exactly who they say they are, safeguarding the security of the private data of the system for the particular user, but also preventing any potential cases of fraud. Literature has explored various ways in order to ensure who is accessing a particular system.

Traditional systems have focused on username/password authentication methods or have added multi-layer authentication mechanisms. However, data leaks are still a very common niche of today’s information systems and platforms. The current state-of-the-art in the authentication field consists mainly of: (1) authentication through a sequence of numbers, (2) authentication through a pattern/graphical based password, (3) two or multi-factor authentication, (4) individual biometrics authentication; (5) computer-learned biometric authentication.

1. Authentication through **passcodes or PINs** (usually a sequence of numbers): users need to remember the passcode and enter it every time they want to access the device, which in turn causes them to use simple passcodes that can be easily guessed, making the device vulnerable;
2. Authentication through **pattern/graphical-based password**: due to the nature of the human skin, the pattern remains on the glass/plastic and is becoming visible to other malicious users.
3. **Two-factor authentication** (constitutes of a second layer of security upon the username/password credentials): while using the username/password credentials is proven to be vulnerable, the second layer of authentication through a different device also has its own limitations, such as the key logging, man-in-the-middle attacks and loss of the device.
4. **Individual biometrics** based authentication (voice, iris, face and fingerprint recognition): each of these individual biometrics have their own disadvantages: the voice recognition is prone to environmental noise; the face recognition requires to hold the device a certain way, while also being affected by the ambient light; the iris recognition does not operate well in environments with strong light; the fingerprint recognition is susceptible to the different angles when the finger is scanned and does not offer the user the possibility of training all the fingers to unlock the device.
5. More **advanced biometrics-based** authentication (the system learns the walking or device usage patterns): walking patterns are not constant, and the device usage patterns are affected by the way the user holds the device; thus, these cannot be considered as a viable individual authentication biometric.

All the authentication methods mentioned above, if breached, compromise the users' privacy. Another important issue is that users do not often have any control over the databases in which the information used to authenticate them is stored; thus, there remains the very pertinent issue a user's authentication information (be it username/ password credentials, or biometric data – fingerprint, iris, face, voice, behavioural patterns) being compromised and violating the users' privacy.

3.3.2 Innovation through TANGO

While state-of-the-art approaches focus on using one biometric feature (fingerprint, iris, voice, face) to authenticate users, the TANGO solution will use biometrics and behavioural patterns from people's daily life to identify individuals, based on location, interaction, speed, movement, walking, smartphone usage and so on. Behavioural patterns constitute a unique characteristic of people in their daily life, which provide valuable information in the process of verifying and identifying users in information systems. Having a secure and frictionless authentication mechanism ensures the security and privacy of individuals while provides a user-friendly approach. Human behaviour paves the way for autonomous, continuous, and multi-modal user authentication, while minimising the potential error and improving the user experience.

The proposed solution will continuously authenticate the user, without the need for any user input, by learning the users' behavioural patterns. The authentication process is based on understanding and learning the behavioural patterns of the users by leveraging data extracted from smartphones, such as sensor and device usage data. The system combines multiple behavioural observations from which the identity of a user can be determined. The combination of different types of observations leads to an extremely high accuracy in authenticating individuals by eliminating the errors of each individual behavioural pattern. Psychologists are involved in the project to understand the grammar of behavioural cues and patterns to identify individuals. Each of these behavioural patterns has proven to be efficient in identifying individuals using a rather limited observation period. Combining these individual observations with machine-learning and psychological models increases the performance of the authentication system and replaces vulnerable authentication methods such as usernames/passwords.

The TANGO Continuous Behavioural Authentication system will consist of the platform, where the processing and the authentication procedure takes place, and the mobile app/library, where the data are being collected to send them to the back-end platform.

The data collected from a smartphone app and forwarded to the third-party platform, are then forwarded again to the platform, with an added pseudo-ID. As a first step, the platform will create a record of the pseudo-ID and initiate the learning process for the authentication models. Once the behavioural models have learned the patterns of the user, the same path will be followed for the biometric data. In this case, the authentication service will be also executed. Initially, it will perform authentication followed by learning, in order to adapt to the contextual changes of the person.

3.4 Device Continuous Behavioural Authentication

3.4.1 SOTA on CBA

Continuous Authentication refers to methods that have been used to provide increased security and reliability considering authentication in comparison to traditional authentication methods relying on passwords and certifications. Biometrics and behavioural patterns can offer insights that aim to deliver increased fraud protection that provides system operators with actionable intelligence to create a secure, smooth and frictionless digital customer experience. This review complements the review on the topic "Continuous User Behavioural Authentication". It focuses on characteristics and information that are identical to the device operation and cannot be controlled, imitated and/or changed directly by the end-user (e.g. keystroke dynamics, mouse movement, touch gestures, walking gait, input patterns). An exemplary paradigm is the small IoT devices ubiquitous. Although these devices feature sufficient computing power and can support applications, they lack conventional interfaces such as keyboards, mice, and touchscreens, or embedded sensors such as the mobile devices have (gyroscope, accelerometer, etc.). As a consequence, conventional continuous authentication and authorization methods that exploit user behavioural patterns collected by embedded sensors cannot be applied. Furthermore, the dynamic nature of the IoT devices' operation often restricts users from having physical control on them. Therefore, there is a strong need for users and devices to be continuously authenticated and gain authorized access [62] Currently, protocols such as *WebAuthn* allows servers to register and authenticate users using public key cryptography instead of a password, while OAuth 2.0[51], the industry-standard protocol for delegated authorization, focuses on client developer simplicity while providing specific authorization flows for Web applications, desktop applications, mobile phones, and living room devices. Figure 10 shows how continuous data acquisition enables an endless verification and assessment of collected data (whether they are physiological, biometric or system performance) for feature extraction to validate and authorize a user to get access.

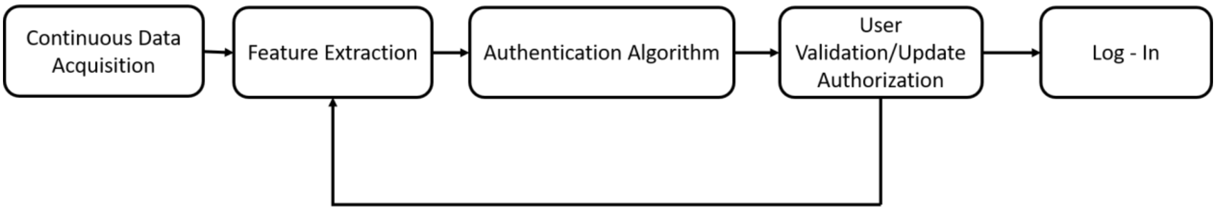

Continuous Data Acquisition
Feature Extraction
Authentication Algorithm
User Validation/Update Authorization
Log - In

Figure 10 Continuous Authentication Diagram.

Device continuous behavioural authentication is a method of authenticating devices based on the way those devices operate and interact with the network and ambient environment. It is based on continuously monitoring a device's operational patterns and using that information to verify its identity. This information can stem from usage patterns but it cannot be inferred directly from them. As users are unique in how they use their devices, the device operation can have unique characteristics based on specific configurational parameters during setup and from the environment where it operates. Device continuous authentication complements the user continuous authentication and provides an extra layer of security.

Services have become more ubiquitous than ever and a constantly increasing number of users access them with their portable devices (e.g. cell phones, laptops, digital pads). The majority of those services are offered in "stores" through applications that can be downloaded by the user. Most of these applications have a client–server architecture. The client runs on the device's operating system, which is most frequently Android or iOS. This client is downloaded to the device from the app distribution platforms, where developers publish their software updates. Unfortunately, the need for optimized functionalities and improved features often comes at the expense of mobile security. Security risks include (i) insecure and unreliable communication that is not only based on self-signed certificates but it also lacks encryption, (ii) weak input validation (iii) insecure data storage (iv) client code security (v) poor authentication and authorization controls (e.g. offline authentication) can result in data theft and compromise of backend services.

Continuous authentication can enhance a system's security by providing an additional layer of protection beyond just a password, certificate or sensor-based biometric authentication. If a device's system operational patterns deviate from their normal patterns, it may indicate that someone else is attempting to use the device and the system can trigger fraud preventive mechanisms, such as prompting the user to re-enter the password, renew certificates and/or permitting or not access to the device. Examples include the assessment of nearby signals, network traffic and signals and/or power consumption, along with user behavioural patterns. Figure 11 presents a classification of the Continuous Authentication showing also the respective device behavioural categories.

[51] https://oauth.net/2/

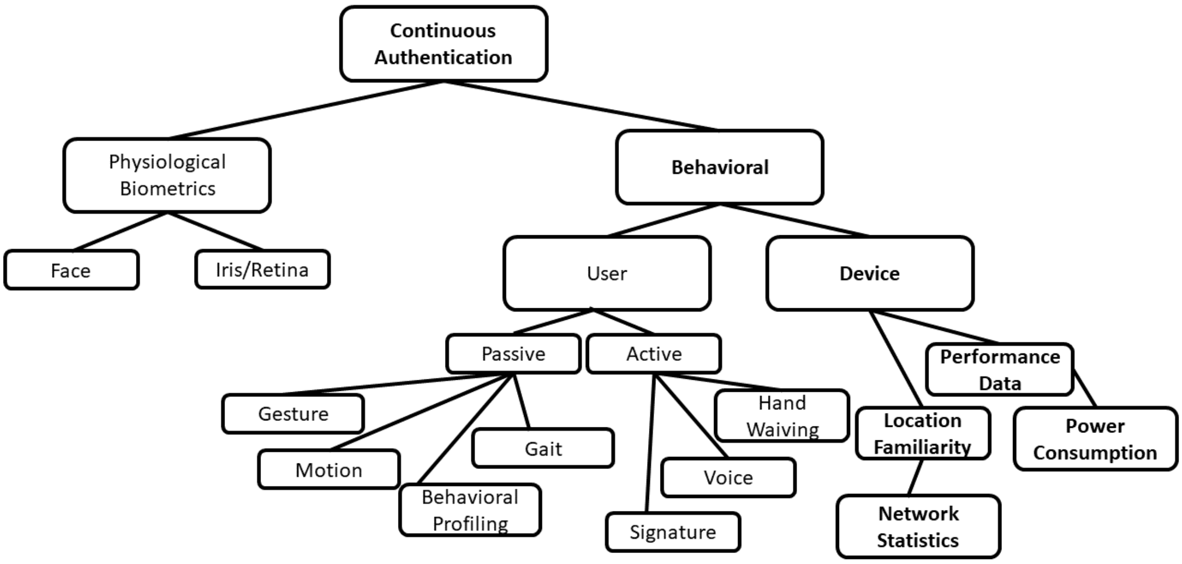


Figure 11 Device Continuous Authentication Classification.

Location familiarity has been used to determine the identity of a device by sensing also nearby electronic device's signals with an aim to establish a legitimate level of user's authenticity [63]. Employing pervasive modalities such as radio frequency (RF) signals, ambient light, and sound, can be used for enabling continuous authorization. In a recent work [64], authors created an RF-based authentication system that exploits the presence of nearby Wi-Fi networks to identify users. The intuition behind RF-based authentication is that the wireless channel metrics — such as channel state information (CSI) and received signal strength (RSS) — change based on a user's presence and movement but it cannot be controlled directly by them. The patterns of change in these metrics not only depend on the way the user moves but also follow the spatiotemporal characteristics of the presence of nearby networks that rely on the network deployment setup and configuration. Users are characterized by different gaits, and as a result they produce different patterns in wireless channel metrics as they move across the different networks. Applying Machine Learning techniques can be used to associate each user with the corresponding patterns of change that are identical to him/her and identify the user. In [65] authors proposed a scheme that uses a Long Short-Term Memory (LSTM) Deep Learning (DL) based classification model to identify an attacker's fraudulent behaviour.

Gupta el al. presented a context profiler [66] that assessed location traces with an aim to detect places of interest for a user's mobile device and created profiles of the nearby devices using Bluetooth and Wi-Fi interfaces in those places to estimate the familiarity of the environment. In their work, they showed how differentiated levels of familiarity scores can be utilized to infer safety and use the safety score indicators to enable access control decisions.

Another choice for enabling continuous authentication relies on the monitoring of the device's **power consumption**. Power consumption of devices is not only highly associated with usage behavioural patterns but it also depends on the device's configurational parameters. In [67] [68] authors have shown that a smartphone's power consumption is strongly related to the state of operation of that particular smartphone's driver. Moreover, they have proposed continuous monitoring of the operational behaviour based on a machine learning-based approach that consisted of three different modalities: power consumption, touch gestures, and physical movement, and verified the functionality of the continuous authentication of the users.

In addition, in [69] authors, conducted an evaluation of the Power Consumption of Co-authentication as a Continuous User Authentication Method in Mobile Systems by measuring the power consumption levels of the Bluetooth interfaces of the wireless devices.

Furthermore**, Blockchain technology** has been also proposed [70] to provide real-time and non-intrusive continuous authentication for the IoT networks, since any transaction occurring is recorded by the Blockchain network, thus offering a database for analysing patterns and behaviours. Blockchain technology provides mechanisms that prevent unauthorized data tampering and improve data integrity and immutability [71]. Using Blockchain technology for continuous authentication of IoT devices though, poses several challenges due to the limited processing, computing and storage capacity of the IoT devices.

Although Blockchain technology provides the means for secure and untampered data integrity, it cannot guarantee the veracity of information. Moreover, scalability is a key issue, as it becomes essential to meet the workload of Blockchain heavy-computational tasks, as users own mobile devices of constrained power and computational capabilities. In addition, continuous monitoring of the device operational characteristics, such as power consumption and network scanning, does not only comprise a data logging activity, but it requires the use of frequent system/drivers calls that might affect the device performance. For instance, the level of granularity of the collection of a complete list and a thorough assessment of the nearby networks depends on the number of scans that a mobile phone should conduct at a given time. This unavoidably introduces trade-offs with power consumption and affects also networking parameters as the device’s wireless interfaces are utilized for scanning.

3.4.2 Innovation through TANGO

TANGO aims to advance the state-of-the-art and enable continuous authentication based on an assessment of the devices’ behavioural patterns. UTH will design and implement a monitoring tool that will assess the behavioural performance of a device based on power consumption and network analytics, taking into account the limitations and constraints of the mobile devices in terms of power consumption and network scanning operations. The tool will be able to detect deviations in comparison to their normal operation based on specific performance efficiency metrics (*e.g.* power consumption, rssi, network traffic metrics). In order to determine the power consumption from the usage activities performed by the user, the Android built-in software drivers that provide measurements for the device power consumption will be used. For collecting statistical information about the nearby networks, the respective networking drivers that scan the wireless medium to detect other devices or wireless networks that operate in the same area will be used, and a solution will be developed to gather and assess this information to create a location familiarity profile for the device.

Based on the assessment of the statistics of the devices power consumption and network information, the aim of this tool is to provide complementary continuous authentication to user behavioural authentication tools. It can also be used as a standalone solution in cases where the use-case requirements limit or do not allow the physical presence of a user to access and use a device.

3.5 Hardening against Side-channel Attacks

3.5.1 SOTA including Comparison

Side-channel attacks represent a dangerous threat against embedded systems and IoT devices. Indeed, they are notoriously efficient at finding encryption keys, in particular for the AES block-cipher [72].

Countermeasures against side-channel attack are divided in 2 categories:

- Masking countermeasures, which make attacks more complicated by splitting any intermediate value that gives information on the secret into n variables (called shares), in a way that forces to know all of the shares to know the secret. The algorithm is adapted in order to manipulate the shares instead of the original variables, carefully avoiding any share recombination that could lead to information leakage.
- Hiding countermeasures that consist in lowering the signal-to-noise ratio.

3.5.1.1 Masking Countermeasures

Several techniques exist to split the secret values into shares. Boolean masking use an *exclusive OR* operation (XOR) to do so [73].The creation of two shares s0 and s1 from the secret s and a random value r is the following: $s0 = r$; $s1 = r \oplus s$. Arithmetic masking follows a similar procedure, but use a modular arithmetic addition instead of an XOR. Affine masking leverages both a finite field multiplication and an XOR to create the shares: two random numbers are drawn, one multiplicative mask m (non-null), and a Boolean mask b [74]. Then, the shares are $(m \otimes s \oplus b)$, m, and b. Inner-product masking leverages the inner-product operation, combining a public vector (for instance <1, 44>) with a vector containing the secret and random numbers (<s, r>) [75]. In our example, this gives the following computation: $1 \otimes s \oplus 44 \otimes r$.

The difficulty for all masking schemes relies in the modification of the code to be secured: while some operations are easy to perform on shares, others require procedure that is more complex.

The presence of such operation usually implies a big performance overhead on the secured code.

Resilience to micro-architectural leakages and target platform change is also a concerning issue for masking countermeasures [76]. Indeed, recombination effects due to the micro-architecture can induce leakage directly dependent on the secret and as such, can dramatically reduce the practical security of the implementation. Such effects are invisible on the assembly code and varies from one device to another.

3.5.1.2 Hiding Countermeasures

Various hiding countermeasures have been proposed as well, such as:

- Instruction shuffling [77] [78] [79] [80],
- Random delays or noise instructions [81] [82] [83] [80],
- Random precharging [84] [85],
- Use of semantic variants [86] [79] [80]

These techniques are generic and can be applied on any program. Yet, their efficiency depends on the program to secure: in the case of instruction shuffling, the program should have enough independent instructions to create an efficient shuffling effect. Best results are obtained by combining several techniques to better consider the variety of possible programs. For instance, [86] use both semantic variants and instruction shuffling as well as some random permutation of table accesses, and [80] leverage register renaming, instruction shuffling, semantic variant and insertion of noise instructions. Combination of several techniques also allows the approach to offer various trade-offs between security and performance.

Hiding techniques suffer less from micro-architectural specificities than masking techniques, as they do not aim at eliminating leakage but at making any leakage (even unexpected ones) harder to exploit.

3.5.1.3 Automation of application of countermeasures

Automation of application of countermeasures represent a promising way to ease the application of countermeasures against side-channel attacks. Indeed, the application of countermeasure by hand is error-prone and tedious. Various approaches have been presented in the state-of-the-art to automate countermeasures applications [84] [85] [86] [79] [87] [88]. While some modify the source code and other the assembly, most works apply the countermeasures within the compilation flow. Such approach enables to interleave tightly the countermeasure application with the performance optimisations performed by the compiler, which reduce the risk that the compiler removes partially the countermeasures. In addition, the compiler offers various representations of the code, enabling even complex countermeasures such as masking to be applied with reasonable effort.

Currently, most automated approaches considered mostly Boolean masking, and various hiding countermeasures.

3.5.2 Gaps and disadvantages of the current technologies

From the state-of-the-art, several needs arise: (1) the need to (at least partially) automate the application of countermeasures, (2) the need to have countermeasures resilient to micro-architectural leakages, and (3) the need to offer various performance vs security trade-offs.

Indeed, most countermeasures are considered in isolation: few works consider the combination of countermeasures. This leads to a situation were very few countermeasures, and a fortiori automated tools, allow to reach various performance vs security trade-offs. The combination of several countermeasures has the potential of filling this gap, and to reach better resilience to micro-architectural leakage by combining the strengths of individual countermeasures.

3.5.3 Innovation through TANGO

Data management is at the heart of TANGO. To protect data from malicious consultation or corruption, encryption is a key feature. In this regard, hardening against side-channel attacks is crucial for any embedded system as side-channel attacks are very effective to attack implementations of ciphers. For instance, the AES block cipher is highly vulnerable to side-channel attacks, leading to key recoveries in a matter of minutes on unprotected systems. As TANGO has multiple use-cases with different needs, having automation for the application of countermeasure and having various trade-offs possible is important.

We consider the combination of countermeasures and their automated application for TANGO. In particular, we aim at combining a loop-shuffling countermeasure with code polymorphism. Both are hiding countermeasures and do not have natural flaws for micro-architectural leakages. Loop-shuffling is a countermeasure offering very low performance overhead, and code polymorphism is highly configurable. We aim at extending the configuration possibilities by adding new code transformations inside the polymorphism engine. This should result in a highly flexible countermeasure combination, with various performance/security trade-offs possible. In addition, having both countermeasures applied automatically will allow the application of countermeasures on diverse implementations depending on the use-case needs.

4 AI-based Framework for Green & Trustworthy Operations

4.1 Exploratory Data Analysis Engine

4.1.1 SOTA including Comparison

Data analysis is defined as the process of analysing raw data to draw out meaningful insights. When it comes to data analysis, there are mainly three popular approaches: 1) the classical, 2) the exploratory and 3) the Bayesian inference approach. In classical analytics, the data collection is followed by the imposition of a model (e.g. normality, linearity), whilst the analysis, estimation, and testing that come after, are focused on the parameters of that model. In EDA[52], the data collection is not followed by a model imposition, but by a process aiming to infer what model would be appropriate instead. Finally, in Bayesian analysis[53], the analyst attempts to incorporate scientific/engineering knowledge/expertise into the analysis by imposing a data-independent distribution on the parameters of the selected model, which, in turn, results into a combination of both the prior distribution on the parameters and the collected data in order to jointly make inferences and/or test assumptions about the model parameters.

EDA, which will be applied in the context of the TANGO project, constitutes an operative approach to data analysis, aiming to improve the understanding and accessibility of results. In more detail, EDA builds upon the soundness of statistical models and hypothesis formulation, which are also used in the classical approach, to reveal hidden and unknown information from data in a form that enables analysts to obtain an immediate, direct and easy-to-understand representation of it. Hence, visual graphs are usually generated when this approach is followed, so that a more direct and trustworthy interpretation of similarities, differences, trends, clusters and correlations are obtained through a picture, rather than through a series of numbers. In reality, EDA forms an analytical framework where the visual examination of data sets, by means of statistically significant representations, plays the pivotal role to support the formulation of hypotheses that could be tested on new data sets. The ability of comparison between two concepts, for instance the dynamic experimentation on data (e.g. evaluating the results on different subsets of a same data set, under different pre-processing conditions), along with the exhaustive visualisation capabilities, empower researchers to identify outliers, trends and patterns in data, upon which new theories and hypotheses can be built.

Thus, EDA has recently become a popular methodology due to the increasing availability of datasets and its advantages in deriving insights from data, being usually the first technique when approaching data, assisting both with the identification of hidden patterns and correlations among features/attributes, but also with the formulation of hypotheses from the data and its validation. It is being used in multiple sectors/fields and its implementation relies on a wide range of techniques and tools, from basic statistical exploration and visualisation to more sophisticated attribute/feature transformations such as Principal Component Analysis (PCA).

EDA usually includes several steps, such as:

1) Identification of attributes/features/variables;

2) Univariate/bivariate/multivariate data analysis to characterise the data in the dataset;

3) Detection of interactions among attributes/features by performing bivariate and multivariate analysis;

4) Detection and minimisation of impact of missing and abnormal values;

5) Detection of outliers or anomalies (further analysis or errors);

6) Dimensionality reduction;

7) Feature engineering, where features are transformed or combined to generate new features [89] [90].

[52] Tukey, J.W.: Exploratory Data Analysis. Addison-Wesley, Reading, MA (1977)

[53] https://huyunwei.github.io/eda_book/exploratory-data-analysis.html, Accessed: 16.12.2022

However, when it comes to the creation of time series subsequences joins, typical EDA displays several disadvantages that stem from the daunting nature of the problem. For even modest sized datasets the obvious nested-loop algorithm can take months, and the typical speed-up techniques in this domain (i.e., indexing, lower-bounding, triangular-inequality pruning and early abandoning) at best produce one or two orders of magnitude speedup. Moreover, current EDA comes with disadvantages that include offering inconclusive results, lack of standardised analysis, small sample population and outdated information that can adversely affect the authenticity of information.

4.1.2 Innovation through TANGO

As the interest in data analysis lies mostly in the detection of anomalies and trends, TANGO's Exploratory Data Analysis Engine (EDAE) will rely on methods and techniques that display significant advantages to this direction. Specifically, EDAE's will implement a data analysis framework that constitutes an extension of the state-of-the-art Matrix Profile (MP) technique, which is applied in a variety of sectors (e.g. such as finance, biology, robotics, healthcare, IoT, geology) but not for exploratory data analysis purposes per se. The MP method, introduced in 2016, can be used to find the top motifs (the best matching sub-sequences in a series) and the top discords (the most distinct sub-sequences in a series), being, thus, well-suited for anomaly and trends detection in a variety of contexts that are characterised by unique behaviours. MP displays numerous advantages as it is exact, robust, computational efficient, scalable and, to a large extent, parameter-free [91]. It also has the capacity of describing a dataset through the identified anomalies or trends/patterns.

EDAE will however go one step beyond the work conducted using the MP (or its current extensions that can be found in the literature), by designing and developing an extended framework that focuses on the implicit distance matrix calculation Series Distance Matrix (SDM). In this framework, distance measures (SDM-generators) and distance processors (SDM-consumers) will be freely combined, leading to a higher degree of flexibility and easier experimentation. Additionally, a Contextual Matrix Profile (CMP) will be developed as a new SDM-consumer that is capable of discovering repeating patterns [92]. This will enable the EDAE to provide intuitive visualisations of the data analysis results and to detect anomalies that are not discords.

In the EDAE, other Machine Learning (ML) techniques (e.g. Linear Regression, Logistic Regression, Decision Trees, Naïve Bayes, etc.) will surround the MP so as to enable the automated analysis of large amounts of information (i.e., data in domains not covered by human expertise) and provide insights into hidden patterns. The combination of well-established data analysis techniques and innovative ones will be exploited in order to develop an AI-driven method to automatically perform the data correlation analysis, visualising the generated results. Last, but not least, EDAE will be designed in a manner so as to become a replicable solution for any kind of input data, thus possibly expanding its application to other use cases on top of the two (i.e. the autonomous vehicle and the retail ones) currently described in the Description of Work. Depending on the number of the uses case in which EDAE will be finally involved and the type of data provided and processed in each case, the novelties concerning the final exploratory data analytics framework that will be developed and implemented may be increased, always in accordance with the specificities of the exploration that is sought per use case.

EDAE will cover all the above steps, putting an emphasis on anomaly detection. Besides the well-known techniques that can be applied to detect and treat outliers/anomalies, EDAE will implement a technique that shows advantages to this direction. More specifically, EDAE will implement the relatively new and innovative technique called Matrix Profile (MP) [93]. The advantages of using MP for most use cases include, among others, a) accuracy (no false positives), b) simplicity and parameter-free learning, c) space efficiency and d) treatment of missing data.

Figure 12 depicts EDAE's position in the process that begins with the definition of a use case and goes up to decision making, whilst it also shows a high-level architecture of the module.

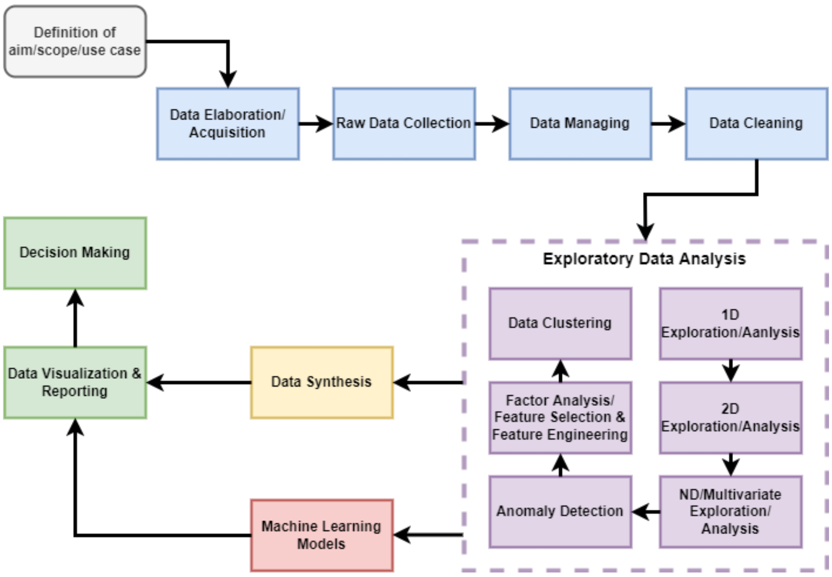

Definition of aim/scope/use case
Data Elaboration/ Acquisition
Raw Data Collection
Data Managing
Data Cleaning
Exploratory Data Analysis
Data Clustering
1D Exploration/Aanlysis
Factor Analysis/ Feature Selection & Feature Engineering
2D Exploration/Analysis
Anomaly Detection
ND/Multivariate Exploration/ Analysis
Decision Making
Data Visualization & Reporting
Data Synthesis
Machine Learning Models

Figure 12 Exploratory Data Analysis Engine high level architecture

The MP algorithms to be implemented in TANGO’s EDA will be simple, fast, parallelisable and parameter-free, while they can be incrementally updated for moderately fast data arrival rates, thus contributing to the system’s robustness, scalability and computational efficiency.

4.2 Energy efficient AI model training

4.2.1 SOTA in AutoML

AutoML stands for Automatic Machine Learning, and includes methods to build and validate ML pipelines minimizing the intervention of a user, which is of great interest for final users, such as enterprises [94]. AutoML may include all stages from the exploratory data analysis of datasets to the deployment of a ML model. However, in many cases AutoML focuses on the automation of specific ML tasks, often relying on specific hardware platforms, mainly feature engineering, hyperparameter optimization, pipeline optimizers or neural architecture search, often offering comparable results with an easier learning curve for non-expert than hand-crafted ML models.

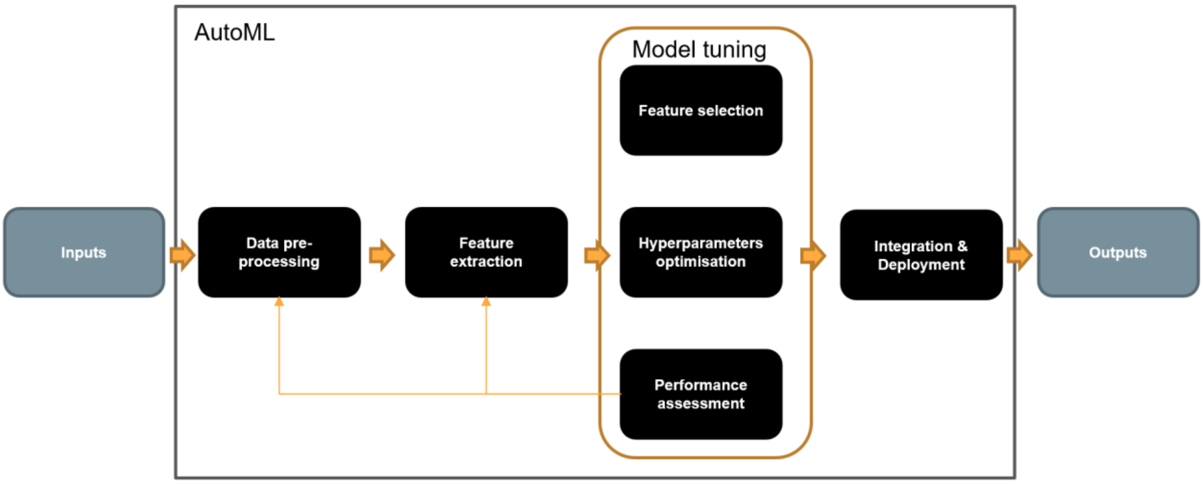


Figure 13 AutoML typical workflow

The authors in [95] present a perspective to make AutoML greener, as criticism has been made that AutoML is expensive in terms of resource consumption, as many approaches rely on large experiments over many datasets, especially while benchmarking many of the existing solutions. They introduce methods to evaluate the carbon footprint of AutoML as well as propose environmentally friendly benchmarking principles and future research outcomes. In [96], an interesting call to action for future improvements of AutoML in the scope of climate change is presented, including the usage of AutoML over spatio-temporal data, not well-represented so far in the literature and tools.

In [97], the authors provide a review of the current state of the art of AutoML solutions for time series forecasting but covers many of the existing tools and frameworks available, such as H2O [98], AutoSKLearn [99], AutoGluon [100], AutoWeka [101], or AutoKeras [102], among others. Other interesting open source AutoML tools are Kubeflow Katib[54] on Kubernetes environments and MLJAR AutoML[55], a python package for AutoML on tabular data. Major vendors offer their own AutoML tools, such as Google (AutoML Tables[56]), Microsoft Azure (AutomatedML[57]) or AWS(AutoGluon and SageMaker Autopilot[58]).

4.2.2 SOTA in MLOps

Similar to the concept of DevOps for shortening and automating the system development, delivery and deployment, the concept of MLOps tries to do the same for the entire life-cycle of the Machine Learning models management.

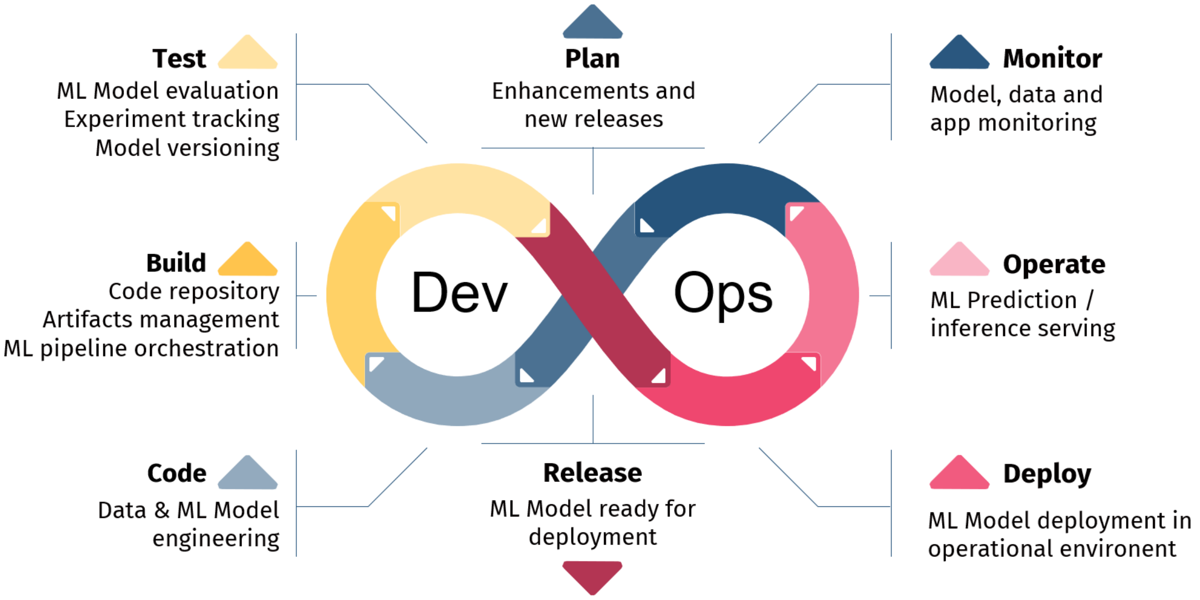

Test
ML Model evaluation
Experiment tracking
Model versioning
Plan
Enhancements and new releases
Monitor
Model, data and app monitoring
Build
Code repository
Artifacts management
ML pipeline orchestration
Dev
Ops
Operate
ML Prediction / inference serving
Code
Data & ML Model engineering
Release
ML Model ready for deployment
Deploy
ML Model deployment in operational environent

Figure 14 MLOps high-level life-cycle

The MLOps site[59] provides and overview of the main principles and current state of play of MLOps. According to this site, the technology stack supporting a MLOps framework should include support for the ML life-cycle over data, model and associated code. In order to develop a model multiple experiments with data should be performed, so data engineering and version control of data, experiments and code are paramount. As in the case of DevOps, support for continuous integration and continuous

[54] https://www.kubeflow.org/docs/components/katib/overview/
[55] https://github.com/mljar/mljar-supervised
[56] https://cloud.google.com/automl-tables/docs
[57] https://azure.microsoft.com/en-us/products/machine-learning/automatedml/
[58] https://aws.amazon.com/sagemaker/autopilot/
[59] https://ml-ops.org/content/state-of-mlops

delivery pipelines (CI/CD) and the automation of deployment is key, as well as further support for model performance assessment, and model monitoring in production, once it is deployed.

Although MLOps began by providing guidelines and best practices, several tools and frameworks have been developed in recent years to cover some of the aspects cited above. As in the case of DevOps, there is no single tool that fits-it-all aspects of the ML life-cycle, so normally organisations tend to build their own MLOps stack. An MLOps framework that cover most of the steps are MLFlow[60] and Kubeflow[61].

MLFlow offers four flavours for the different steps of the ML life-cycle. MLFlow Tracking manages data, experiments and models, MLFlow Projects allows to package reproducible code and runs, MLFlow Models helps with the deployment of the serving models in different environments, and MLFlow Registry allows the management of models. It is also about including MLFlow Pipelines, a tool to create ML pipelines based on YAML files.

Kubeflow is a MLOps framework that supports the development and deployment of ML models and pipelines in Kubernetes. Kubeflow is integrated with Jupyter Notebooks for data science experimentation, TensorFlow model training, TensorFlow model serving container, Kubeflow Pipelines, as well as extensions to work with other frameworks such as PyTorch, MXNET and others.

Current tools for MLOps are still evolving, as the field is relatively new. Enhancements in these tools or new ones are expected in the coming years.

4.2.3 SOTA in Federated Learning

Federated Learning (FL) [103] aims to train Machine Learning models with data distributed among across multiple devices and computing nodes without the need of disclosure them. It is therefore a technique that brings the ML algorithms to the data and then combines the partial training results in a new improved trained model. There are multiple approaches to FL [104], but most of them face similar challenges.

One of the main challenges is to preserve the privacy. FL is in itself a major step towards protecting privacy, as raw data remains in the original node rather than travelling to a central repository to perform the training process. However, sending the results over the network may disclose some sensitive data, such as the originator of the raw data. Several techniques have been proposed to deal with this, such as differential privacy [105], secure multiparty computation [106], or using secure trusted mechanisms such as the nodes being part of a data space following IDS principles[62].

Although the raw data remains in the original infrastructure, hence minimizing the transmission of data and therefore the energy consumption, the communication of the results of potentially thousands or millions of devices might become very expensive. The results of the training process on each node (model updates) have to be either broken into smaller pieces transmitted over time, or sent entirely at the end of the training process. It is therefore necessary to find the right balance between sending the entire model in a single message or sending many messages with successive updates. There are different types of communication, as the orchestration of the final learned model might be centralized (a central node request updates to other nodes and builds the final model), distributed (no central node, but the nodes coordinate to build the final trained model) or heterogeneous (where all nodes share the same global model structure but instead of a single model there are several models learned locally). The centralized approach is simpler if there is no bottleneck in the communication, while the other approaches show advantages in case of orchestration issues in a centralized fashion. The communication can also be reduced by applying model compression techniques [107].

Another aspect to take into consideration is that the training process might occur in very heterogeneous systems, ranging from low-power, CPU or GPU-enabled edge devices with low connectivity, to powerful servers [108]. This imposes to the FL approach to be fault-tolerant to failures in communication, or to be able to use only the available devices at a given time.

[60] https://mlflow.org/
[61] https://www.kubeflow.org/
[62] https://design-principles-for-data-spaces.org/

A good review of the current FL approaches, including a literature overview, applications of FL and future directions is discussed in [109].

4.2.4 Innovation through TANGO

This task aims to unify and leverage the developments made in the previously mentioned areas, offering an integrated solution for energy-aware development and exploitation of machine learning models. There are several synergies between the aforementioned research areas, with the focus of TANGO being their integration and evaluation in customer-centric scenarios.

In the case of AutoML, the overall idea is to study how to design and generate energy/efficient DNNs, balancing the appropriate energy consumption and the performance in terms of accuracy, using distributed environments for training and inferencing of the models, including edge computing devices. Most of AutoML solutions are suited for use cases using centralized repositories, therefore not benefiting of the distributed nature of data to find optimum neural network architectures or their hyper-parameters in terms of carbon footprint. TANGO will leverage on the current SOTA to integrate models to predict the computational power required for the training and serving of the DNNs. The innovation of TANGO will be mainly focused on the optimization of the local designs using feature-partitioned data and then the aggregation of the results to find a global solution collaboratively. TANGO will also target specific hardware platforms as design criteria for AutoML

To this extent, TANGO will explore AutoML and MLOps techniques in combination with Federated Learning. The integration of these technologies will provide a framework for the evaluation of potential energy savings in all the steps of model development and exploitation. One example of this is the study of the usage of data spaces (based on IDS) for training models using federated learning in secure and distributed data sharing spaces, thus connecting to other research areas of the project and benefiting of the energy savings proposed in other tasks.

4.3 Dynamic Intelligent Execution on Heterogeneous Systems

4.3.1 SOTA on Heterogenous Execution on Managed Programming Languages

Regarding Java and managed programming languages, prior work has focused on embedded support for heterogeneous programming inside existing languages targeting GPGPUs, FPGAs, vector units, and multi-core processors. The majority of these efforts rely on translating bytecode into CUDA or OpenCL C to generate code for the GPGPU. To the best of our knowledge, the most complete attempt at enabling the use of hardware accelerators from Java is TornadoVM[63] (originated in The University of Manchester) which uses the GRAAL dynamic compiler to generate OpenCL C code for a wide range of hardware accelerators like GPGPUs, multi-core CPUs and FPGAs. It improves over work like APARAPI by using runtime optimizations to exploit hidden task-parallelism and has the ability to apply a range of different data-parallelization schemes automatically. On the other hand, APARAPI was used as inspiration for the now defunct OpenJDK Sumatra project[64]. Furthermore, Numba [110], Copperhead [111], RiverTrail [112] and ParallelJS [113] are JIT compilers that can offload Python and Javascript code on GPUs. However, to identify the code segments for acceleration, they expose specific annotations and new types to programmers. With regard to FPGAs, several frameworks have recently emerged that enable FPGA-based acceleration for Python applications. PyRTL [114] compiles Python programs to Verilog, while Hot&Spicy [115] compiles Python programs to FPGA binaries by leveraging prior research from open-source projects for Python development on Xilinx Zynq (PYNQ) [65]. However, both frameworks require programmers to add hardware-specific primitives from Python; and in turn, transform the input programs to contain appropriate wrapper bindings for interfacing with the generated hardware design. Currently, the mainstream for GPU acceleration from Python, Java, Scala and other

[63] https://www.tornadovm.org/

[64] The Sumatra Project. https://openJDK.java.net/projects/sumatra/

[65].PYNQ: Python Productivity for Zynq. 2017 https://www.pynq.io/Xilinx

managed languages is by using specific libraries (e.g. CuPy[66], RAPIDS[67]) that invoke pre-compiled code optimized for a specific architecture. Therefore, the state-of-the-art machine learning frameworks (e.g. TensorFlow[68], PyTorch[69], etc.) that are built on top of C/C++, expose bindings to managed languages in order to offer them the benefits of hardware acceleration. TVM [116] is an optimizing compiler for deep learning applications (i.e., Caffe[70], MXNet[71], etc.), to utilize CPUs, PUs and TPUs.

As discussed in previous paragraph, the programming models for hardware accelerators are primarily created for a limited set of programming languages, such as C, C++, and Python.. Other programming languages, especially managed programming languages are less researched. TANGO employs TornadoVM, the state-of-the-art technology for bridging this gap in managed programming languages, and it aims to enhance it to create execution plans based on energy efficiency.

4.3.2 SOTA on ML-based Execution Plan for Heterogenous Systems

Troodon [117] is a load-balancing scheduling heuristic that classifies OpenCL applications as suitable for CPU or GPU execution, based on a speed-up predictor. The Qilin [118] compiler uses offline profiling to create a regression model for predicting the execution time of input applications. Ogilvie et al. [119] introduce a low-cost predictive model for the automatic construction of heuristics that reduce the training overhead for execution on CPU-GPU equipped platform. Furthermore, Grewe et al. [120] leverages predictive modelling to influence the OpenCL code generation from OpenMP programs when speed-ups are predicted. Additionally, Chen et al. [116] combine generic search with learning and benchmarking to find good scheduling methods for execution on heterogeneous hardware, including CPUs, server GPUs, mobile GPUs, and FPGA-based accelerators. However, the supported scheduling mechanism is semi-automated, as the search space must be manually defined by a programmer for each algorithm similar to a template. Wen et al. [121] show that the concurrent execution of OpenCL kernels can increase the GPU utilization and improve performance. This is achieved by applying a decision tree based prediction model to determine whether an application kernel should be scheduled individually or along with other kernels. Baldini et al. [122] use existing OpenMP applications and supervised learning to predict the potential GPU execution speed-up among different vendors. Brown et al. [123] present a model that allows to get accurate predictions of speed-ups using a small set of features, while also being portable portability across Nvidia GPUs with different capabilities. Adams et al. [124] propose a novel scheduling algorithm for the Halide programming language that targets image processing pipelines. Their model combines symbolic analysis with machine learning to predict performance. Finally, TornadoVM has implemented dynamic reconfiguration [125] and multiple-tasks-on-multiple-devices (MTMD) [126]. The former is a technique that can transparently migrate execution from one GPGPU to an FPGA and vice-versa, while the latter is a novel ML-assisted scheduling mechanism that allows the concurrent execution of multiple methods onto multiple heterogeneous devices (e.g. CPUs, integrated GPUs, GPGPUs).

4.3.3 Innovation through TANGO

TANGO aims to advance the state-of-the-art and enable the dynamic intelligent execution on heterogeneous systems. To achieve this goal, the UoM will extend the in-house TornadoVM technology with power measurements into its run time layer and with the employment of ML models for the automatic adaptation of the execution plan across different devices. Thus, TANGO pilots will be enabled to exploit hardware acceleration in an automatic manner, while the execution will be transparently adapted.

[66] https://cupy.chainer.org
[67] https://developer.nvidia.com/rapids
[68] https://www.tensorflow.org
[69] https://pytorch.org
[70] https://caffe.berkeleyvision.org
[71] https://mxnet.apache.org

4.4 Privacy Threat Modelling and Identification for Trustworthy AI

4.4.1 SOTA Privacy Risk Assessment

In the cybersecurity domain, risk assessment is relatively well studied, compared to privacy risk assessment, with commonly recognized risk models for analysis using metrics such as likelihood, vulnerability, threat and impact. Organisations have substantial volume of resources, frameworks and tools to support cybersecurity risk assessment. The privacy domain, on the other hand, lacks development and uptake of uniform concepts of privacy risk assessment, as well as in-depth guidance and tools for assessing and managing privacy risks. As a result, organizations find it challenging to integrate privacy risk assessment into their risk management approaches, access and measure the impacts, and identify measures to mitigate the impacts in an actionable way despite the availability of guidance, standards, practices, and tools to help organizations and individuals protect and manage personal and sensitive information, such as the General Data Protection Regulation (GDPR)[72], UK Data Protection Act 2018[73] and California Consumer Privacy Act[74], through a set of key principles, rights and obligations for processing of personal data. Standards provide real benefits for privacy management by establishing practices to be applied consistently across organizations, while frameworks set up basic guidelines to help fulfil compliance obligations, benchmark against industry best practices and enable global interoperability. Besides popular information security standards such as ISO/IEC 27000 series, ISO/IEC 29100[75], NIST SP 800, ETSI TS 103 485[76] and BS 10012[77], there are variety of guidelines, tools and frameworks to manage information security and privacy concerns and support compliance. Despite the availability of external standards and frameworks, many organizations chose to develop their own privacy risk
assessment methodology to assess and improve their privacy protection program. Improvements could include additional privacy management principles, privacy enhancing technologies or closing gaps in existing data handling processes.

A privacy risk assessment, also known as data protection impact assessment or privacy impact assessment, is the process of identifying and evaluating privacy risks. Better management of privacy risks through the development of effective solutions to protect individuals' privacy when designing or deploying systems, products and services that process data can help organizations build customer trust. The process can also help organizations understand and priorities privacy risks with their broader profile of enterprise risks and drive comprehensive risk management approaches to promote better resource allocation and decision-making. During a privacy risk assessment, organizations identify events leading to possible privacy risks, quantify and rank the risks, and finally decide on whether and how to reduce, remove, transfer or accept the risks. Currently privacy risk is considered as one element of enterprise risk, meaning privacy risk is a secondary consideration that might result in inappropriate or ineffective selection of countermeasures to mitigate or protect against it. Viewing data separately through a privacy risk lens provides better chances of identifying privacy risks and thus the most appropriate countermeasures can be implemented to address the privacy risks. Privacy risk analysis methods are essential for minimizing or avoiding privacy breaches. Tang et al. [127] present a list of existing mechanisms and approaches offering privacy risk analysis. Methods have been developed to measure privacy risk based on the number of records stored in the system [128], the characteristics of the system [129] or based on the entire frameworks for privacy risk assessment [130] [131]. Gharib et al. [132] applied privacy by design concept to identify challenges in understanding privacy and suggested that the vagueness of privacy concepts add ambiguity for designers and stakeholders, hindering the decision making. The authors further identified 38 key concepts and relationships grouped across four categories, creating 17 organizational factors, nine risks, five treatments and seven privacy factors. Privacy risk can

[72] https://gdpr-info.eu/
[73] https://www.legislation.gov.uk/ukpga/2018/12/contents/enacted
[74] https://oag.ca.gov/privacy/ccpa
[75] https://www.iso.org/standard/45123.html
[76] https://www.etsi.org/deliver/etsi_ts/103400_103499/103485/01.01.01_60/ts_103485v010101p.pdf
[77] https://www.bsigroup.com/en-GB/BS-10012-Personal-information-management/

be assessed from a wide range of perspectives. These assessment models include assessment of potential privacy impacts on data subjects' fundamental rights. PRIAM [133] provides a detailed view on privacy risk to the data subjects using privacy harm trees to assess risks related to feared events, risk sources and weaknesses. Sion et al. [134] proposes a methodology to understand privacy risks while assessing the relevance of privacy threats to the system under design using threat modelling techniques. Other privacy risk assessment models ([135], [136], [137]) focus on assessing the risk specific to a single data user or subject. The NIST Privacy Framework[78], on the other hand, enables better privacy engineering practices driven by privacy by design concepts and help organizations protect individuals' privacy through Enterprise Risk Management and facilitates a common language to communicate privacy requirements with entities within the data processing ecosystem.

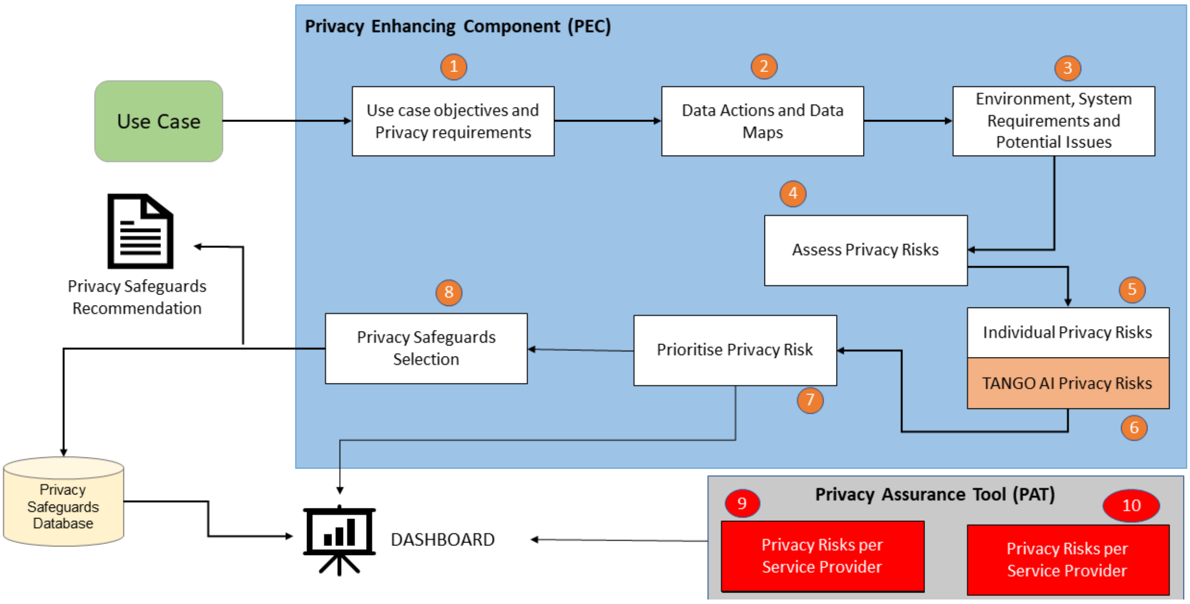

Privacy Enhancing Component (PEC)
Use Case
1
Use case objectives and Privacy requirements
2
Data Actions and Data Maps
3
Environment, System Requirements and Potential Issues
4
Assess Privacy Risks
5
Individual Privacy Risks
TANGO AI Privacy Risks
6
Prioritise Privacy Risk
7
Privacy Safeguards Selection
8
Privacy Safeguards Recommendation
Privacy Safeguards Database
DASHBOARD
Privacy Assurance Tool (PAT)
9
Privacy Risks per Service Provider
10
Privacy Risks per Service Provider

Figure 15 T5.4 Architecture with overview of Privacy Enhancing Component (PEC)

Limitations were identified:

- Existing frameworks predominantly assess data through security risk with little to no ability to include privacy risks.
- Recommendations from existing privacy frameworks does not provide guidance on compliance with regulatory standards such as GDPR.
- Existing privacy frameworks does not assess AI/ML algorithms from a privacy perspective.

4.4.2 SOTA Privacy Assurance Tool

Privacy risk analysis mechanisms are essential for minimizing or avoiding privacy breaches. FairWarning®[79] is an organization-centring for healthcare sector to detect suspicious activities and policy violations. It allows to manage the access to patient healthcare information. FairWarning®'s turn-key privacy auditing solutions are compatible with healthcare applications from every major vendor including Allscripts, Cerner, Epic, GE, McKesson, MEDITECH, Siemens, and many others. FairWarning® is compliance with several healthcare privacy regulations such as ARRA HITECH privacy and meaningful use criteria, HIPAA, EU Data Protection. However, FairWarning® is limited to healthcare sector. In [138], the authors presented Personal Data Analyzer (PDA), a solution to monitor personal data transactions to detect, prevent anomalies, and assess associated privacy risks, misbehaved

[78] https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.01162020.pdf
[79] https://www.netsurion.com/knowledge-packs/fairwarning

transactions, and keeps users engaged along the process. PDA results aimed at enhancing privacy protection: the H2020 PoSeID on project ([139], [140], [140]).

In [141], the authors proposed PrivacyScore, an automated website scanning portal that allows anyone to benchmark security and privacy features of multiple websites. It can also be used by data protection authorities to perform regularly scheduled compliance checks. The first publicly available version of PrivacyScore was announced at the ENISA Annual Privacy Forum in June 2017. Similar project is proposed in [142] called PrivacyMeter, a browser extension that computes on-the-fly, a relative privacy score for any website that a user is visiting. The difference is that PrivacyMeter is running on the client-side and therefore having access to all content that a server-side crawler cannot access. In [143], the authors proposed a conceptual privacy by design and privacy-aware personal data management framework is user-centric. The formwork aiming at putting the human-in-the-loop concept. It assesses privacy risks in using online services as well as the data disclosure behaviour. It used in leisure travellers' case study. The framework runs on the client side, without dependency on service providers. It means that no private or sensitive data are shared with any existing or new remote services. The authors used an ontology described which is an extended version of the one reported in [144]. The ontology is an ontological graph in the leisure travel context. It is used to enable automatic reasoning about data and value flows, for the purposes of joint privacy risk-value analysis and behaviour nudging construction. Behavioural nudging refers to the use of interface elements which aims at guiding user behaviours when people are required to make judgements and decisions [145].

In [146], the authors proposed a user-centric, data-flow graph based semantic model, for a privacy-benefit trade-offs related to data disclosures. The model can show how a given user's personal and sensitive data have been disclosed to different entities and what benefits the user gained through such data disclosures. The authors used their model on two scenarios: online bookings with travel service provider, and data disclosed to other people in online social networks. This work can increase users' awareness of privacy issues and guide them to make more informed decisions by balancing their privacy preferences and expected benefits from data disclosures.

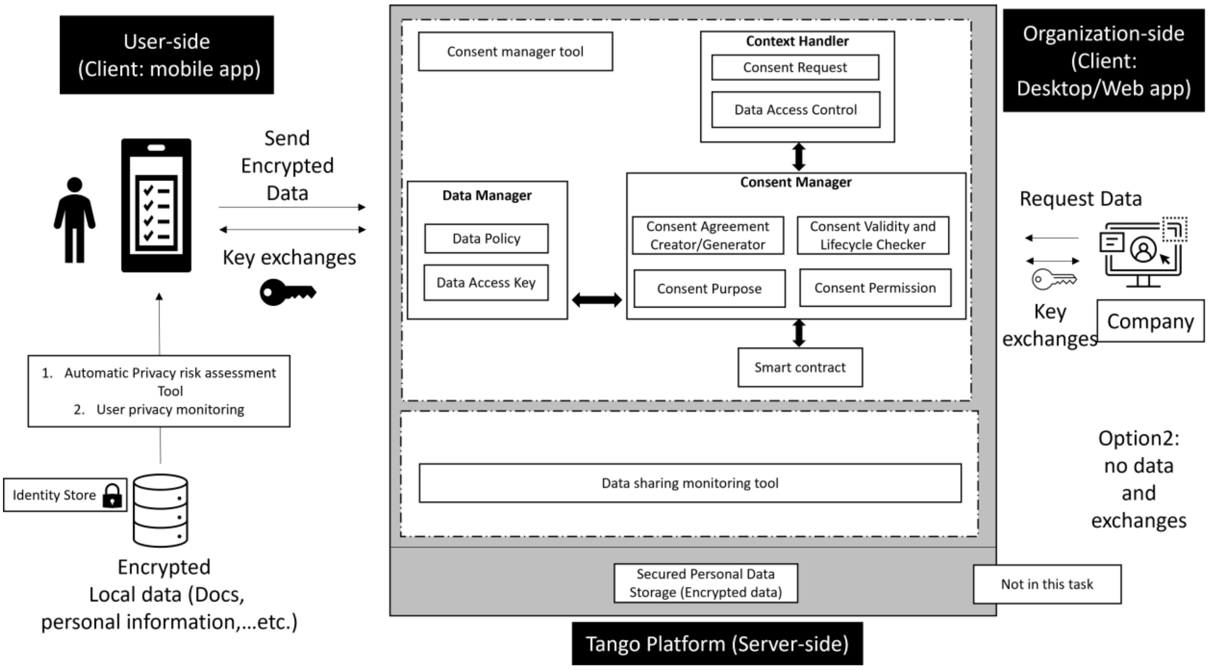


Figure 16 Architecture of Privacy Assurance Tool

In [147], the authors presented a privacy protection framework for IoT applications and systems. It allows the users to control the release of their data and be informed data sharing decisions. They apply their framework on remote patient monitoring with IoT wearables. They used an ontology for privacy risks.

The following limitations were identified:

- Missing encryption protocols for protecting personal and sensitive data shared amongst stakeholders.
- Missing encryption protocols of communications between the user mobiles/IoTs, server, and organizations.
- Lack compliance with the GDPR.
- Missing monitoring and alerting framework highlighting the accessed shared data by organizations.
- The existing frameworks focus on users and they lack focus on organization-centrings.
- The literature shows that privacy preserving data monitoring mechanisms focus on aggregating data to provide different kinds of outputs, such as health, traffic, or location monitoring. These mechanisms are used together with privacy risk assessment tools.

4.4.3 Innovation through TANGO

TANGO aims to advance the state-of-the-art and enable a holistic privacy risk assessment to systematically identify, assess and manage privacy risks. To achieve this goal, T5.4 will apply develop two components to assess privacy from data management perspective and user perspective. The first component, known as Privacy Enhancing Component (PEC), will build upon and extend the NIST privacy risk assessment methodology to illicit privacy concerns and assess related risks for data actions in the TANGO use cases. Along with the data actions which includes data collection, storage, processing and sharing, PEC will develop methodology to assess AI/ML algorithms that uses the data to provide services from a privacy lens. The assessment will lead to the identification of privacy threats, quantification of possible privacy impacts and determination of countermeasures to mitigate the impact. Through this, PEC will not only extend the state-of-the-art by developing a methodology to assess data actions as well as algorithms but also provides an unified framework to identify, assess and mitigate privacy concerns. The second component, known as Privacy Assurance Tool (PAT), will have an automated privacy risk assessment and data monitoring capabilities to assess privacy concerns and provide users greater control on their data. It will offer both users and organizations to granularly manage their sharing of their data and monitor compliance with regulatory standards such as GDPR. Finally, the developed components will be integrated into one in T5.4 providing the users to specify, in simple and graphical way, their privacy requirements for their data handing processes, assess privacy concerns, identify possible impacts and privacy violations, visualize privacy risks and recommendations of countermeasures to address the risks. . Integration of these technologies will enable a framework to evaluate privacy concerns from data processing as well from user behaviour allowing users to prioritize and take apt actions against the concerns.

4.5 X-AI For Privacy and Trust Enhancement

4.5.1 SOTA including Comparison

Explainable AI (XAI) is a collection of complex approaches designed to make the outcomes of a solution intelligible to humans, for the purpose of validating real AI use cases. It stands in stark contrast to the notion of the "black box" in machine learning (ML), where even the developers or designers may be unable to explain why an AI model gets a specific conclusion. Transparency, the antithesis of black-box models, is increasingly sought to prevent the risk of making judgments that cannot be justified and do not permit receiving thorough explanations of their behaviour. In terms of security, the false predictions might make the system vulnerable to attack and lead to the implementation of zero-trust security for vital systems. XAI is the technological equivalent of the social right to explanation, can enhance the user experience of services or goods by convincing end-users that AI will make the correct decision. Consequently, the primary objective of XAI is to explain what activities have been completed, what duties will be performed, and what tasks will be performed in the future, as well as to highlight the explanations or facts upon which the actions are based.

The fundamental characteristics of XAI algorithms are explainability, transparency, and interpretability. Explainability can be viewed as a set of properties of the interpretable domain that leads to a specific result, such as classification or regression. Methods and procedures for explaining AI/ML interpretability can be categorized according to the following criteria: (1) pre-model, in-model, and post-model; (2) intrinsic and post hoc; (3) model-specific and model-agnostic; (4) feature engineering; etc. Interpretability refers to the capacity to interpret the machine learning model and offers the essential basis for user-comprehensible decision-making. Interpretability helps to assure impartiality in decision-making, i.e., to discover and, as a result, correct bias in the training dataset (i.e., an unbalanced dataset); (2) interpretability increases the robustness of AI-based solutions by highlighting potential adversarial perturbations that could alter the prediction; and (3) interpretability increases the trustworthiness of AI-based solutions by providing meaningful variable inference and causality of model reasoning. Transparency describes whether the methods responsible for extracting model parameters from training data and producing labels from test data can be defined and prompted by the approach's creator. Conforming algorithms give a foundation for justifying decisions, tracking, and eventually reviewing them, enhancing algorithms, and investigating facts.

Popular saliency techniques include DeepLIFT and Prediction Difference Analysis, which break down algorithms into discriminative features for classification. The strategies linked with "Verbal Interpretability" offer interpretability in the form of verbal chunks/rules that humans can intuitively comprehend (e.g., sentences that indicate causality). Feature Extraction can also provide insights on the explainability and interpretability of AI models, by identifying the features that are the strongest predictors of the AI outcome. [148] [149] [150] [151]

4.5.2 Innovation through TANGO

TANGO will create and incorporate a potent XAI module that explains AI-based judgments to relevant actors. The module will increase the transparency and credibility of AI-based analytics applied to the information given by various public and private organizations. While retaining a high degree of learning performance (e.g., prediction/classification accuracy), will investigate, customize, and integrate approaches that will enable the platform to generate more explainable and relevant models. Furthermore, a library of XAI approaches that highlight the primary characteristics of cyber-security measures and decisions will be provided. The library will allow stakeholders to not only recognize, perceive, and reproduce AI-based judgments and recommendations, but also intellectually comprehend the context and conditions under which these suggestions are formed.

4.6 Infrastructure Management based on AI

4.6.1 SOTA including Comparison

One of the objective of TANGO is to enable a novel approach for infrastructure management by using the federated AI approach and balancing the trade-off among system responsiveness, energy-consumption, privacy and data-protection. . Although research on energy-efficient infrastructure management based on AI is considerably scarce, a brief literature review followed by a proposed solution through TANGO will be provided.

The publication from the EU research hub [152] highlights the significant role of the ICT sector, including data centres, in contributing to global CO2 emissions. They expose data centres specifically since they have one of the fastest growing emissions. The authors evaluate and analyse current trends in energy consumption and efficiency in data centres in the EU. They used data from companies participating in the Data Centre Energy Efficiency program. The analysis demonstrates that the average Power Usage Effectiveness of the facilities participating decreased over time.

The study by Oro et. al. [153] provides an overview of energy efficiency strategies, as well as renewable energy integration into data centres. The authors highlight the need for dynamic models and metrics to properly understand and quantify energy consumption. This deeper understating would allow us to comprehend the advantages of energy efficiency. Rong et. al. [154] provide a review of energy-saving methods applied in data centres. They also propose strategies to maximize data centres efficiency and minimize their environmental impact. Peoples et.al. [155] propose way to improve energy efficiency of cloud data centres. This is achieved with optimization of communications within and between data centres, by using resources close to the client. Their proposed mechanism is the Data Centre (DC) Energy-Efficient Context-Aware Broker (e-CAB). The Broker automates the selection of a data centre in based on application request. The authors have demonstrated the effectiveness of the broker in reducing carbon emissions and financial cost.

Goir et.al. [156] propose GreenSlot, a parallel batch job scheduler for a data centre powered by a PV solar array and as a backup, the electrical grid. Scheduler predicts the amount of solar energy that will be available and schedules tasks to maximize the use of green energy. The scheduler also uses cheap grid energy to avoid deadline violations. The results for production scientific workloads show that GreenSlot can increase green energy utilization by up to 117% and decrease energy costs by up to 39 %, when compared to a conventional scheduler.

SALINAS-Hilburg et.al. [157] focus on energy-aware task scheduling to improve energy savings. They propose a new approach to estimate the energy consumption of applications without the need for full profiling. They use an "application signature" that allows the estimation of energy consumption. Different scheduling approaches are used in combination with the application signature information to improve the scheduling process. The authors evaluate the accuracy of their approach by comparing it to an oracle method. With an error of less than 1.5 %, and compression ratios between 39.7-45.8.

Another approach to make computing more sustainable is to move tasks geographically, to a location with abundance of renewable energy at the time. This approach does not compromise the time of start of execution, but it may hinder the time to collect results, since it will shift the computing load across locations to accommodate for the availability of appropriate resources in other locations. An example of this is the solution proposed by Carroll et.al. [158] In this paper, genetic algorithms are utilized to optimally shift tasks geographically. The authors propose a solution that prioritizes the renewable energy sources by allowing data centres to share information on renewable energy and by moving services between data centres. To find optimal placement, genetic algorithm was utilized.

The main issue in our focus is the volatility of renewable energy. The goal is to utilize data centres as a buffer, ensuring the stability of the grid. As a consequence data centres could also reduce their carbon footprint and reduce operating costs. To achieve this the production of renewable energy in the near future must be forecasted and use data centres (i.e. shiftable jobs) that can modify the consumption according to the forecast.

4.6.2 Innovation through TANGO

Through TANGO a methodology will be implemented in which the prediction accuracy of this service will be evaluated, using real-world data from well-known and trustworthy baseline on more than seven locations around the world, for a period of half a year. Baseline data will be obtained from *BSRN* (Baseline surface radiation network). This is a well-known and trustworthy baseline used by state-of-the art. Our proposed service is *RENOPS* (Renewable Energy Forecast Production Service), which can provide a forecast of renewable energy availability based on geolocation. The service will obtain the historical irradiation data and weather forecast. The model will process the data and build the response in the form of renewable energy. Thus, the TANGO approach will leverage the AI approach, balancing the trade-off between system responses, energy consumption, privacy and data protection awareness.

5 Pilot Analysis

The Pilot Analysis chapter introduces six use cases (Pilots). Each Pilot has a dedicated section, which is divided into two subsections. The first subsection describes the general idea of the pilot. With this overview of the pilot case, the reader can then understand the second subsection, which maps the technologies that can be used for the pilot.

To present this information in a compact way, the mapping is presented with a table. The first column identifies the organization that leads the development of the associated technology, while the second column identifies the technology. The third column is named feature and gives a very short summary of what the technology can bring to the pilot. In the fourth column, which is called support it is described how the technology aims to solve at least one of the problems described. Finally, the last column describes issues that might arise in applying the technology to the pilot project. Furthermore, it presents potential challenges during the implementation phase.

5.1 Smart Hospitality

5.1.1 Summary

CESGA is a hotel chain in Mallorca that has been internationally recognized thanks to its initiatives in circular economy. Our partner Anysolution SL, on the other hand, has developed a data-driven platform called *NADIA* which is being used to contribute to the digital transformation of smart destinations. The TANGO framework will be implemented in one of the hotels of the CESGA chain in Mallorca, that also uses the NADIA platform.

Although it was planned to evaluate the TANGO framework for guests to check into the hotel through their smartphones, so they can proceed directly to their room without going through the lengthy process at the reception, this is not possible in Spain due to the current regulations, which state that it is compulsory for the tourists/visitors to go through the reception for check-in.

For the pilot, a tablet will be available in the reception desk with a survey that tourists can fill in voluntarily. This survey will include information such as nationality, age, sex, and preferences.

To enhance the tourist experience in the hotel, information from sensors installed in the pilot rooms will be captured. This information will indicate whether guests are in the room or not, and the operational state (ON/OFF) of other room elements, such as lights, HVAC, etc.

Combining all the data gathered, tourists will receive information about the gastronomic offers at the hotel, the touristic activities and attractions that are available (hereinafter also referred as *Offering*) as well as how they can contribute to the circularity principles of the hotel.

Guests' identity data will be securely handled by the TANGO platform according to GDPR requirements, without exposing identity information to staff members or others. Personalized configuration will be loaded at the guests' room without the danger of exposing personal data.

TANGO will contribute to preserving clients' privacy, enabling the delivery of services while ensuring a respectful, secure and trustable use of clients' personal data. The TANGO platform will be used to provide effective resiliency against multiple attack vectors aiming at exploiting a range of vulnerabilities of the host device. TANGO should be able to help this use case improve data security and protect privacy according to the context and the specific conditions.

For the pilot, the operation flow will be as follows:

- A tablet will be installed at the reception containing the survey that customers can fill in.
- Sensors will be installed at the pilot rooms.
- NADIA will be connected to the sensors and to the tablet.
- NADIA will send the information to the TANGO system.
- Within the TANGO system, an instance of UPCV User Behaviour Exchange Module ('UBEM') will be created.

- UBEM will allocate a data repository for the hotel, with a plurality of UPCV Wallets, one Wallet for each Offering (see definition above).

TANGO will contribute to the data flow as follows:

- TANGO will generate a token for each customer, which links the personal data from the CRM and the NADIA platform.
- Each customer will be provided by a personal UBEM, which allocates a data repository for collaborative personalization - UPCV ‘Wallet’.
- Each client will be linked to a token.
- At the reception, the customers will be linked to their respective tokens and they will be able to fill in a survey with their preferences.
- Sensors in the pilot rooms will send data to the NADIA platform.
- Data from NADIA will be sent to the TANGO system.
- Based on the preferences of the customer, the time and hours or by UBEM (similarity of their Wallet and Wallets of the Offerings), the customer will receive recommendations related to gastronomy, tourist experiences and circularity.

The infrastructure that will be used is composed of:

- Sensors.
- The NADIA Platform, a data driven platform based on FIWARE.
- UBEM, for the hotel.
- Personal UBEM for each customer (implementation may be in a mobile APP, if this is decided in the project).

5.1.2 Potential technologies to apply

Identifier	Technique	Feature	Support	Challenges
UoG	Privacy Enhancing Component (PEC)	PEC enables clients to gain insights on possible privacy concerns from their data actions and processes, and countermeasures that could be used to mitigate the concerns.	PEC will enable pilot users to identify potential privacy risk for each data action involved in the use case. Along with a score for each data action, PEC will provide a privacy impact score for AI algorithms used in each pilot. PEC will visualise possible privacy impacts with suggested countermeasures to prioritise and mitigate the risks, ensuring data protection and preserving user privacy.	Pilot hotels does not gather and store client data. It might be challenging to identify risks and preserve privacy. It will be challenging to identify individual data actions with data flow diagrams of the systems as the data is collected and processed by third party.
FUJ_LU	Privacy Assurance Tool (PAT)	Provide a privacy assurance mechanism to a user. Monitor proactively the	PAT will facilitate automated privacy monitoring allowing users to control sharing of their personal information.	Identifying the data types and their attributes as well as data flows in the use case.

Identifier	Technique	Feature	Support	Challenges
		privacy and visualize the privacy scores	Through PAT, users will be able authoritatively control sharing of their personal data. PAT allows to provide privacy awareness mechanisms to a user based on the provided data.	Availability of a data sample (a synthetic data) Identifying and explaining the scenario of use. Any privacy issues related to the use case.
QBE	Seamless user and device onboarding	Enable guests to perform remote identity verification through their mobile phone, to be able to bypass the check-in process at the hotel.	Reduce queuing times at the reception of the hotels as the guests do not have to perform identity verification and check-in at the reception, High security in managing the identity information of the guests.	Challenge in integrating with the guest management system, due to proprietary hotel management software. Lack of NFC enabled doors at the hotel to directly access the room without getting a key from the reception
CEA	Hardening against side-channel attack	Make side-channel attacks harder, to prevent an attacker to find encryption key used by the room sensors.	The use of sensors in the rooms could create a potential way for an attacker to perform a side-channel attack that would allow him to find encryption keys used for the communication between the sensors and the NADIA platform. This could allow an attacker to get information about customer preferences, or even to modify the sensor information transmitted to the platform. As such, hardening against side-channel attacks seems to be relevant for smart hospitality context. CEA will provide countermeasures to increase the security w.r..t these attacks.	

Identifier	Technique	Feature	Support	Challenges
VTT	UPCV-based collaborative personalization	Customers will get a personalized view of Offerings, based on the behaviour of their peers.	Offerings will be presented to customers, based on the interest indicators from pervious customers. Hotels using the TANGO system may also collaborate exchanging their UBEM data, in order to increase user satisfaction in the region, or even sell the data (which is privacy-preserving by nature) to third parties (e.g. information that if customers have been interested in Offering A, they might be interested in Offering B as well)..	Where is the customer instance of UBEM running? Do we have a mobile APP in TANGO, into which it could be integrated. Or a personal cloud for each customer?
VTT	Self-Sovereign-Identity Module	SSI Module enables the user to manage his data and provide only necessary information for the accommodation provider	Interfacing with PEC and PAT The user is aware of what information is shared, to whom, and for how long.	Rules, regulations, and accustoms vary between countries and markets.

Table 2 Potential TANGO technologies applied to the Smart Hospitality pilot.

5.2 Autonomous Vehicles

5.2.1 Summary

IDIADA will provide the appropriate infrastructure to implement the pilot of the Autonomous Vehicles. IDIADA will offer autonomous vehicles to the project to allow the integration of the TANGO distributed identity and trust management as well as the validation of the solution through a pilot. A workshop will be held prior to the pilot to provide participants with hands-on training and state-of-the-art cybersecurity exercises. The TANGO framework will be evaluated in terms of the identity and trust management, such as sharing the autonomous vehicles among trusted users, exploiting onboard IoT devices, loading personalized preferences on the vehicle configuration, etc.

Autonomous vehicles share data with other cars, users and companies, which is necessary to identify passengers and other cars, capture the state of the road and driving conditions, etc. It is therefore very important to ensure that all these data are protected from any leakage.

IDIADA has developed its own “robo-taxi”, called *CAVRide*, which works as a self-driving taxi inside IDIADA’s facilities. A smartphone app is available so the user can request the taxi, and the car drives autonomously to the place where the user is waiting.

There are different data exchanges present in the process: (i) user to car, (ii) car to user and (iii) environment to car. The car uses its sensors to scan the environment, together with HD maps and its GPS positioning to know its location in real-time inside IDIADA’s facilities.

This data exchange must be protected:

- The user calls the car using an app installed on their smartphone.
- The information is sent to the car via cellular and a cloud platform.
- The information goes from the cloud infrastructure to the car.
- The car receives the information and starts sending back information to the user.

Parallelly:

- The car gets the information from the server (user's requests and HD Map).
- The server sends information to the car. The data go to the cloud infrastructure and then, to the car.

While the system is running, the car is updating its position and status to the cloud server. If the car is available, it can be summoned by any user. The user's data are generated on a smartphone app and is sent to the cloud server. The request is evaluated on the cloud and, if the logic is met, it is forwarded to the car.

Once the car gets the request, the car accepts the request, and the server starts to send the car information to the user. The data flows from the car to the server to the end-user, and this flow is continuous until the process has finished.

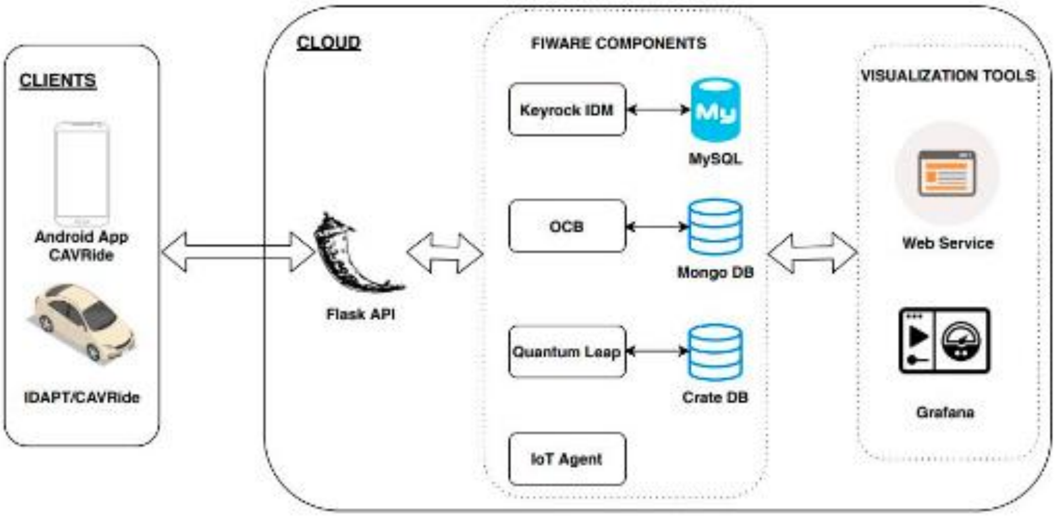


Figure 17 Pilot 2 Autonomous Vehicles

TANGO should protect the data flow (communication user-server, car-server) and the integrity of the HD maps.

The user must be at IDIADA's facilities and can ask for the Autonomous vehicle to pick them up via his/her smartphone. The user must be able to see at any moment where the vehicle is, which means that trustworthy communication must be provided. Once the vehicle has arrived at the requested meeting point, the user selects where they want to go. They must feel safe all along the way. The vehicle must drive giving the user the confidence that it is safe, implying that the user should be able to get all the information on real-time.

5.2.2 Potential technologies to apply

Identifier	Technique	Feature	Support	Challenges
UoG	PEC	Identify and address privacy concerns in autonomous and connect vehicles.	PEC will assure that all privacy-risks related to data actions involved in gathering, sharing and storing of data is identified, privacy impact scores are calculated, and possible countermeasures are defined.	Identifying the data types and data flows in the use case.
UMU		Ensuring privacy during data sharing (e.g., data from users and their preferences)	Users GDPR rights will be ensured through proper access control and user consent. Sticky policies and related identity-based techniques may be used to ensure that only allowed entities may decrypt data. For instance, only (specific) cars may decrypt user's service preferences.	Part of the data shared in the pilot might have strong timing constraints, which may rule out the more complex protection mechanisms. Collaboration with SSI management must be refined to tackle the particularities of the scenario
EXUS	Exploratory Data Analysis	Discover and understand trends, patterns, identify errors as well as detect outliers or anomalous events. Find interesting relations among different variables involved in the use case/ process.	EDA refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations. EDA can be applied whenever there is data available and prior to any other more in-depth data-based investigation. It is a good practice to understand the data first and try to gather as many insights as possible from it.	Availability of the relevant data due to the privacy issues Interpretation of the results
UTH	Device Behavioral Continuous Authentication	Create device behavioral patterns based on operational measurements such	DCBA mechanism will assess the behavioral performance of a device based on power consumption and network analytics --and	Relying the inference and the assessment only on networking measurements increases the risk of false alarms. This

Identifier	Technique	Feature	Support	Challenges
		networking metrics and power consumption	will be able to detect deviations in comparison to their normal operation based on specific performance efficiency metrics (e.g. power consumption, RSSI, network traffic metrics) and infer whether a specific operational behavior is suspicious or not. So, the DCBA mechanism will be able to assess the continuous authentication procedure and enable the authentication of the trusted user while sharing cars among them.	exhibits a high risk when the device deviates from its daily moving routine. (e.g. during a journey or a trip and the user wants to use the mobile services). To eliminate the probability of false alarm the model will incorporate also power consumption measurements to incorporate a second dimension to the inference model and also complement with the user behavioral continuous authentication components. Though DCBA devices will be able to perform continuous authentication.
CEA	Hardening against side-channel attacks	Make side-channel attacks harder, to prevent an attacker to find encryption key used by the car to communicate with other cars or with the cloud.	An attacker could perform a side-channel attack to try to get the encryption keys used in the car, for instance to communicate with the server. This could lead for example to denial of service attack (by creating fake user request send with the same encryption key as the server). Depending on the data exchanged with other cars and on how it is processed and used by the cars, the attack could potentially also lead to dangerous car behaviour, by exchanging fake car positions for instance. As such, hardening against side-channel attacks seems to be relevant in the autonomous vehicle context. CEA will	

Identifier	Technique	Feature	Support	Challenges
			propose countermeasures to make the attacks harder.	
VTT	Self-Sovereign-Identity Module	SSI Module enables the user to manage his data and provide only necessary information for the vehicle (and further to other vehicles/users)	The user is aware of what information is shared, to whom, and for how long.	Rules and regulations vary between countries and markets. Traffic requires near real-time response at times
FSDE	Configurable trustworthiness module	Evaluating the Trustworthiness of the car's system and the CAVRide service	Cars share data among themselves, and Cars communicate with an interface to humans. Establishing Trustworthiness is an important component.	Defining the specific metrics and specifications.

Table 3 Potential TANGO technologies applied to the Autonomous Driving pilot.

5.3 Smart Manufacturing

5.3.1 Summary

Flanders Make (FMAKE) is a research centre aiming to establish the bridge between the academic and industrial know-how in Flanders. For almost more than a decade, FMAKE has been performing research activities for additive manufacturing (AM). FMAKE has been focusing on monitoring and controlling the AM process; with the objective to achieve better productivity and increasing the robustness and print quality; which leads to the reduction in scrap (waste) and costs. FMAKE has a single-laser LPBF industrial printer, which will serve the pilot case in TANGO, that is instrument with advanced in-situ monitoring technologies as well as the ability to control the printing process through different controllers. Using these rich sensory data, FMAKE has been developing advanced AI based algorithms to create a digital twin of the printing process.

In TANGO, Flanders Make will contribute in the context of its digital twin for quality assurance, which runs partially on the edge and partially in the cloud. Flanders Make will focus on the data efficiency in terms of data management/storage/sharing and privacy preserving of their digital twin. Additive manufacturing is used in a variety of industries (aerospace, medical, etc.) where data confidentiality is critical. Managing access rights and privacy is important when printing information is made available externally.

For additive manufacturing, the main operation flow consists of several steps:

1. Design step: the customer designs a part that needs to be printed. This is typically done in design software on the computer. The outcome of this step is a CAD-file, which is sent to the printing company.
2. Print preparation: the printing company combines different designs (CAD-files) into a single print job to increase the productivity of the printing machine and reduce lead times. The outcome of this step is a job file which is uploaded to the printer.
3. Print execution: the print job is executed to print multiple designs in a single build area. While printing, monitoring data is generated and analysed for real-time quality indicators.
4. Print post-processing: the parts are removed from the build area and finished afterwards.

The steps above describe the general high-level flow in a commercial environment. Within TANGO, FMAKE will perform most of these steps in their laboratory.

The use case starts when the CAD-data is available. This data is typically manually sent (file-sharing) to the next step. The print preparation is typically done on a powerful computer, using licensed software. After the preparation, a job file is available which is manually uploaded to the printer controller using SFTP.

During print execution, machine controller data is real-time (<1Hz) available and high-speed monitoring data is sent directly to a processing computer. The monitoring data is analysed in real-time (digital twin) and, after every printing step, key indicators are sent to the cloud; together with the controller data. Due to the high amount of data, only limited parts are sent to the cloud platform for further analysis. After the print, the job file is manually uploaded to the cloud for archiving purpose.

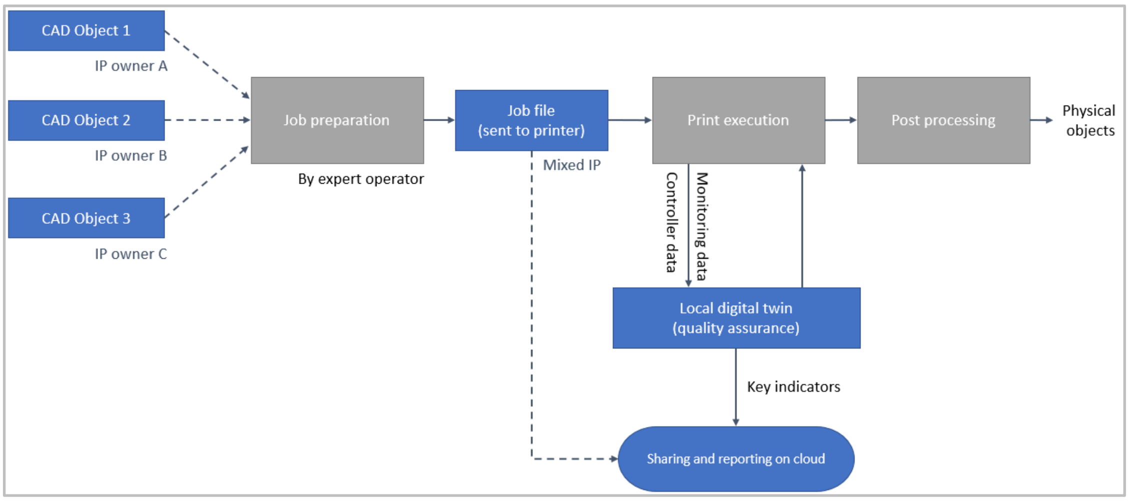


Figure 18 Pilot 3 Smart Manufacturing

Note that the data is aggregated after the print preparation and no data is linked back to their IP owner (CAD).

How to manage the entire data flow, enabling traceability, but preserving confidentially. From design (different IP-owners) to quality reports from digital twin (aggregated information). Using the job file, it is technically possible to link part of the data back to the IP owner, but not in an efficient and secure manner.

Efficient storage and data sharing mechanisms, including a secure manner to combine edge and cloud infrastructure.

RiaStone (RST) is part of "Visabeira Industria" a sub-holding of the https://grupovisabeira.com/en/home conglomerate, RiaStone was created in 2014, after a novel contract being awarded by IKEA Sweden (https://ikea.today) for the Europe based manufacturing of 486 million tableware products in the timeframe 2014-2026. Through that awarded contract, RST manufactures the IKEA Europe wide supply of "Dinera", "Fargrik" and "Flitighet" IKEA tableware families, being these products fabricated through an innovative Industrial ceramics production process: Tableware Automated Single Firing (TASF).

Riastone is applying a systems-level strategy consisting in the integration of new advanced FoF and industry 4.0 techniques to its production line systems and, as a consequence of the introduction of new and advanced externally interconnected data systems, significatively increases the vulnerability of RiaStone to internal and external cyber-attacks.

Riastone has the ambition of under the scope of the TANGO project to be able to implement Secure Data Operations through the implementation of blockchain technologies, and cybersecurity ML models based in behavioral patterns, that enable RiaStone to attain and maintain a high-level Cybersecurity defensive posture.

5.3.2 Potential technologies to apply

Identifier	Technique	Feature	Support	Challenges
UoG	Privacy Enhancing Component (PEC)	Case1: Privacy preserving digital twin. Case2: Privacy preserving AI system.	PEC will assist in identifying and quantifying privacy concerns of the AI/ML techniques used. Defining countermeasures to address privacy concerns in AI.	Identifying the decision making process of the Pilot's proprietary AI software will be challenging.
ATOS	Federated Learning and MLOps	Case 1: Training ML models over data distributed in edge devices Case 2: Train cybersecurity AI models for RIAStone	Case 1: Federated Leaning techniques could be applied to train AI models with data at edge or other computing nodes without having to move or disclose the raw data. This will help in reducing energy consumption and privacy concerns. Case 2: Usage of MLOps for ML Model management life-cycle, potentially AutoML techniques	Case 1: Ensure enough computational capabilities at the edge nodes. Data annotation for training. Handling big amount of data from the machines. Case 2: Data annotation. Identify the case in the pilot
QBE	User continuous behavioural authentication	Perform continuous behavioural authentication for workers and factory staff when accessing infrastructure and equipment	Improved safety in the factory and reduced production discontinuation due to due reduced cases of unauthorised access.	Integration with existing smart factory software may be a challenge depending on the openness of the software. Some equipment does not have many interfaces that the workers interact with, which may not produce enough data to understand workers' behaviour
VTT	Self-Sovereign-Identity Module	SSI Module enables the staff members to manage their data and provide only necessary	For example, Quality Control personnel need to be identified in the end product	Removal of obsolete information after the product is deployed and approved to the customer.

Identifier	Technique	Feature	Support	Challenges
		information for the factory and its customers		

Table 4 Potential TANGO technologies applied to the Smart Manufacturing pilot.

5.4 Banking

5.4.1 Disclaimer

The consortium management is currently in the process of replacing the partner that was leading the Pilot during the proposal time. Thus, the document does not contain a summary for this Pilot. However, the consortium is confident to reach an agreement with ABI Lab.
The focus of the Pilot could be to apply federated learning techniques to train AI models on data that resides in several banks. Given the remaining uncertainty for this PILOT more information on this will be provided in deliverable D2.2. This includes the missing mapping of technologies to this Pilot.

5.5 Public Organizations

5.5.1 Summary

VISAR is a platform that assists third-country nationals with the relocation process to Germany. Applying for a visa in a home country and obtaining the residence permit in Germany are the multi-stage processes in which the correct implementation of all the nuances and requirements is critical. The VISAR platform operation is divided into steps according to the main points of the standard relocation - visa application in a home country, city registration in Germany or application for the long-term residence permit in Germany. Upon successful completion of all steps, the employee receives the long-term residence permit in Germany and VISAR successfully completes the task of assisting the employee in the relocation process.

The platform is primarily intended for employers, who register on the platform. Employers can create cases for their employees on the platform, and employees will receive an activation link for registration. After an employee is registered, an introduction email from the responsible VISAR manager is automatically sent to them. The VISAR platform uses built-in algorithms to assess the eligibility of employees for the preferred visa types. The type of visa depends on the employment requirements for the worker, the educational and professional background, the documents that can be provided by the employee, etc. Once the appropriate visa type has been determined by the VISAR platform, the result must be manually verified by the VISAR manager to eliminate the possibility of technical error.

As the employee moves through the relocation process and all required steps, getting assistance from the VISAR manager is important. The employer can constantly follow the progress on the VISAR platform and be informed in case there are any delays or issues during the procedure. The same feature is available for the employee - the employee gets notified by email if the VISAR manager does any changes in his/her profile (*e.g.* create a checklist for the next appointment or book an appointment at the Diplomatic Mission or the immigration office). The VISAR manager is also able to track if the employer or the employee changes any details in the profiles. The possibility of tracking the progress by all stakeholders ensures the transparency and openness of the VISAR platform.

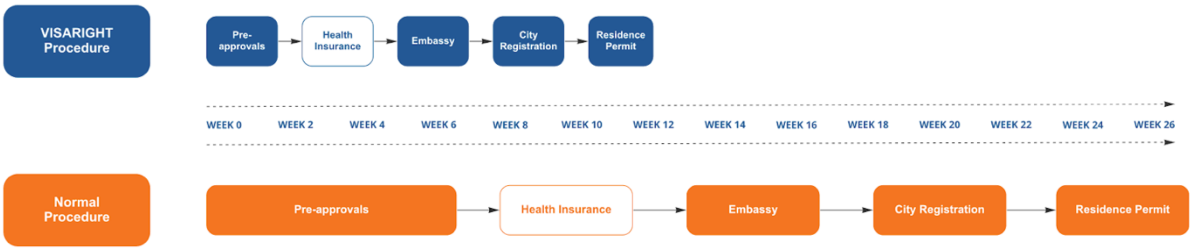


Figure 19 Pilot 5 Public organizations

The TANGO platform will be used to provide effective resiliency against multiple attack vectors that seek to exploit a series of vulnerabilities of the host device. TANGO should be able to help the Pilot improve data security and protect user privacy in what context and under what condition. Furthermore, should technologies such as Blockchain or AI be used, the existing technology can be improved in terms of data transfer and data security.

5.5.2 Potential technologies to apply

Identifier	Technique	Feature	Support	Challenges
UoG	Privacy Enhancing Component (PEC)	Providing a way to assess privacy concerns related to the data handling processes and identify countermeasures to address them.	PEC will assure that all privacy-risks related to data actions involved in gathering, sharing, and storing data is identified, privacy impact scores are calculated, and possible countermeasures are defined. PEC will assist the client to perform due diligence with respected to preserving user privacy and complying with GDPR.	Obtain internal data flow structures of the use case may be a challenge. Identifying existing data protection controls may be against the organisation's disclosure policy.
FUJ_LU	Privacy Assurance Tool (PAT)	Provide an automated way of managing user privacy.	Through PAT, users will be able authoritatively control sharing of their personal data. PAT allows to provide privacy awareness mechanisms to a user based on the provided data.	Identifying the data types and their attributes as well as data flows in the use case. Availability of a data sample (a synthetic data)
UMU	Confidentiality and Privacy by Design	Ensuring privacy principles are held for user data	GDPR should be enforced through proper data access control and user consent. Sticky policies and related identity-based techniques	Different levels of granularity may be necessary, and there are multiple complex scenarios that may require data sharing (from visa application

Identifier	Technique	Feature	Support	Challenges
			may be used to ensure only allowed actors to retrieve data (apart from access control to the database), even with fine-grained control. E.g., only the corresponding	requests to visa procedures to further services like accommodation), so technologies have to be applied in a careful way to cover the different needs.
QBE	Seamless user and device onboarding	Providing a strong way of verifying the identity of the citizens with a high level of assurance when interacting with a government entity.	Allowing citizens to interact with public administration such as visa application remotely, without the need to visit a government entity physically to verify their identity.	Integrating with the existing Public Administration software responsible for the visa application process. Lack of legislation/acceptance in some countries to perform identity verification remotely.
VTT	Self-Sovereign-Identity Module	SSI Module enables the organization members to manage their data and provide only necessary information.	The public role and related data of an organization member are separated from his personal identity/data.	Public officials and their identity are bound to their public role.

Table 5 Potential TANGO technologies applied to the Public Organisations pilot.

5.6 Retailers

5.6.1 Summary

Metro operates in Greece and Cyprus and has presence in e-shop for supermarkets and for professional wholesalers. Metro’s goal is to understand better its customer needs and become customer-centric, by exploiting and leveraging the huge amount of existing environmental data with new technologies, such as Machine Learning and Big Data.

There are three major groups of application platforms in Metro’s system that will be involved in this project:

1) **Stores Applications System**. There is a store checkout application and a store back-office system.
2) **Central Offices Applications**. A custom-made ERP System where transactions from all stores are gathered.
3) **Data Warehouse**. There are two different data warehouse systems, a legacy SYBASE IQ for MCC, and an Oracle Data warehouse Database for Metro.

Analytical sale data are collected at physical stores. Each transaction is stored at the store back-office systems, called Metrisys. By the end of the day each store has produced a custom EDI file containing analytical raw data for all transactions. The file is transferred through secure transfer protocols to the Central Offices of Metro.

Afterwards, an ETL procedure triggered by Central ERP system loads the files to the OLTP database (Oracle) where all transactions are stored.

Then, different scheduled ETL procedures transfer data from OLTP to the Data warehouse (OLAP).

The following must be considered:

- Through the Metro system it is difficult to apply a customer segmentation procedure and extract results for customers and their buying patterns.
- System users need to extract and combine data from many different reports.
- The time needed for aggregating the data for all customers and all stores in our stores network is relevant.
- The retail activity and operating capability in Cyprus is not the same.
- It is difficult to compare results of two systems from two different companies to identify similarities and differences in the segmentations.
- The meaningful market basket analysis for the applied segments is currently done on a pilot basis and for specific customers.
- Combination and consolidation of different reports require enough effort to identify instant variances (differences and similarities).

TANGO can help this Pilot with a categorisation of customers and generation of emails that can be sent to customers to increase the interactivity and have a more personalised shopping experience. The GDPR must be complied with. Moreover, due to the huge amount of data, TANGO will provide the retail sector with data privacy and security, for example via smart contracts, so that partners have the absolute control. Traceability of products will contribute to the environmental impact and the transparency of transactions to the social pillar of the sustainability, as it will increase the trust between the stakeholders.

In summary, the TANGO platform will support the following:

- Achieve effective and efficient data management for customer and product analysis.
- Provide insightful conclusions by entering data inputs from a system (retailer/database) to translate in the same format and summarize and unify product solution from two different countries.
- Group the results in an Executive summary format to analyse trends.
- Provide security and anonymity of data for possible data sharing with both internal and external use-case scenarios.
- The platform could be a future solution for data sharing of sensitive information with strategic partners.
- Current stakeholders, mainly commercial managers, marketing specialist and finance professionals.
- Identify and correct possible disruption of IT operations through the whole process

(i.e., the platform should work efficiently in parallel mode so that daily routines are not interrupted).

- Secure backups and use encryption mechanisms to protect sensitive data inflow and outflow, using AI technology.

5.6.2 Potential technologies to apply

Identifier	Technique	Feature	Support	Challenges
UMU	Confidentiality and Privacy by Design	Ensuring confidentiality of data and partners' absolute control over	Sticky policies and related identity-based techniques may be used to ensure only allowed entities may decrypt data. This offers a further step of protection for	Data from customers must be properly protected according to regulations, avoiding predatory business tactics. As the

Identifier	Technique	Feature	Support	Challenges
		how it is shared	data at rest (apart from access control to cloud sharing platforms) and gives partners that share data complete control over their data, establishing policies for who can access it. That way, data shared by partners through the platform, like analytical data from specific physical stores, is always protected according to their wishes.	scenario involves cross-border sharing, issues about regulations may arise..
EXUS	Exploratory Data Analysis	Dividing customers based on shopping frequency, amounts spent and recency to define their loyalty levels. Understanding shopping patterns and preferences between customers of different countries using different systems and structures. Exposing data to analysts of suppliers, new customer prospects that are setting up their business case, other markets, the tourism industry	The Exploratory Data Analysis (EDA) can be used to provide a deep analysis of the sales data, product data, customer data. EDA can later be used by business/store managers to spot weak areas in a store in order to suggest areas that can be targeted for increased revenue. EDA can be also used to provide a range of insights into customer behaviour. For example, exploring large datasets pertaining to customer purchasing patterns can indicate regional taste preferences, such as what the most popular foods is in each country.	Availability of the relevant data due to the privacy issues. Interpretation of the results.
ATOS	Federated Learning, AutoML and MLOps	Case 1: Training ML models over data distributed in locations (and countries)	Case 1: Federated Leaning techniques could be applied to train AI models with data of the different two countries where METRO operates, hence allowing train models without having	Case 1: Creation of an IDS-compliant data space (or similar), including data connectors in the countries/centres. Data annotation for

Identifier	Technique	Feature	Support	Challenges
		where METRO operates Case 2: Usage of AutoML techniques to derive new ML models based on the available data	to disclose the data from the two countries (a problem that METRO currently faces). This could be achieved for instance creating IDSA connectors and moving the algorithms to the data for training using federated learning. The concept can be extended to different METRO stores/data centres acting as a node of a data space and using federated learning among all the stores, independently of the country. Case 2: Usage AutoML techniques to create new ML models required by METRO.	training. Having annotated data for training purposes Case 2: Data annotation. Identify the case in the pilot
VTT	Self-Sovereign-Identity Module Endorsement functionality	SSI Module enables the customer to manage his data and provide only necessary information to the retailer.	Customer information can be shared to the retailer and transferred if the customer so wishes. Endorsing capabilities help other customers to choose products to their liking.	Retailers are also competitors; customer information is an asset.

Table 6 Potential TANGO technologies applied to the Retailers pilot.

6 Conclusions

In this TANGO deliverable D2.1, we gave a comprehensive analysis of the state of the art (SOTA) of current available technologies for distributed data management, processing and storage, with their challenges and limitations. Our SOTA mainly focuses on technologies that will be highly relevant to development of the TANGO platform. Particularly, the SOTA covers three broad range of categories (as per technical work packages) from privacy-preserving distributed data management, trust management technologies, to AI technologies for green and trustworthy operations. The SOTA of each technological area is provided with a comparison of different methods/techniques, their shortcoming, current gaps, and further, how we see the TANGO project's potential innovations could go beyond current SOTA results.

In addition, we provided an initial description of real-world pilot cases (five out of six) and each technological area has received an initial mapping to at least one pilot case.

Our technical partners then conducted a high-level analysis each pilot case, not only mapping the technologies, as well as potential challenges that might occur when pilot cases will be carried out, which gives the project stakeholders the option to define a clear path forward.

In conclusion, this work is performed within the scope of the task T2.1 GAP Analysis in Distributed Data Management, Processing & Storage, and partly the task T2.2 User Needs and Requirements for Data Management, Processing & Storage. This deliverable should establish the basis for the selection of applications, development tools, required components, related technologies, and standards used in the TANGO Platform. The results will serve as input for the deliverable D2.2 User Needs and Requirements & Use Case Scenarios.

7 References

[1] A. A. Monrat, O. Schelén and K. Andersson, “A survey of blockchain from the perspectives of applications, challenges, and opportunities,” *IEEE Access,* vol. 7, p. 117134–117151, 2019.

[2] M. A. Khan and K. Salah, “IoT security: Review, blockchain solutions, and open challenges,” *Future generation computer systems,* vol. 82, p. 395–411, 2018.

[3] J. Neises and T. Walloschke, “Trustworthiness as Key Enabler for Connected Services in Mobility,” 12 February 2021. [Online]. Available: https://standards.ieee.org/wp-content/uploads/import/documents/other/e2e-presentations/feb-2021/02-Trustworthiness_as_Key_Enabler_Connected_Services_in_Mobility.pdf.

[4] T. W. C. G. J. Neises and B. Popovici, “Trustworthiness as facilitator of Policy and Access Management in Supply Chains,” 12 February 2021. [Online]. Available: https://www.plattform-i40.de/PI40/Redaktion/DE/Downloads/Publikation/2021-conference-volume-industrie40-security.pdf?__blob=publicationFile&v=5.

[5] H. J. Putzer and E. Wozniak, “Trustworthy Autonomous/Cognitive Systems – A Structured Approach,” *Whitepaper,* 2020.

[6] V. Ollikainen, “Clustering Enhancement for a Token-Based Recommender.,” in *CIKM Workshops*, 2018.

[7] V. Ollikainen, “Networked Collaborative Recommendation Architecture,” in *IBC2019 Conference*, 2019.

[8] S. Truex, L. Liu, M. E. Gursoy, L. Yu and W. Wei, “Demystifying membership inference attacks in machine learning as a service,” *IEEE Transactions on Services Computing,* vol. 14, p. 2073–2089, 2019.

[9] M. Nasr, R. Shokri and A. Houmansadr, “Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning,” in *2019 IEEE symposium on security and privacy (SP),* 2019.

[10] L. Lyu, H. Yu, X. Ma, C. Chen, L. Sun, J. Zhao, Q. Yang and S. Y. Philip, “Privacy and robustness in federated learning: Attacks and defenses,” *IEEE transactions on neural networks and learning systems,* 2022.

[11] H. Hu, Z. Salcic, L. Sun, G. Dobbie and X. Zhang, “Source inference attacks in federated learning,” in *2021 IEEE International Conference on Data Mining (ICDM),* 2021.

[12] Y. Liu, J. Peng, J. Kang, A. M. Iliyasu, D. Niyato and A. A. Abd El-Latif, “A secure federated learning framework for 5G networks,” *IEEE Wireless Communications,* vol. 27, p. 24–31, 2020.

[13] L. Melis, C. Song, E. De Cristofaro and V. Shmatikov, “Exploiting unintended feature leakage in collaborative learning,” in *2019 IEEE symposium on security and privacy (SP)*, 2019.

[14] B. Hitaj, G. Ateniese and F. Perez-Cruz, “Deep models under the GAN: information leakage from collaborative deep learning,” in *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security*, 2017.

[15] Y. Lu, X. Huang, Y. Dai, S. Maharjan and Y. Zhang, “Federated learning for data privacy preservation in vehicular cyber-physical systems,” *IEEE Network,* vol. 34, p. 50–56, 2020.

[16] Z. Wang, M. Song, Z. Zhang, Y. Song, Q. Wang and H. Qi, “Beyond inferring class representatives: User-level privacy leakage from federated learning,” in *IEEE INFOCOM 2019-IEEE conference on computer communications*, 2019.

[17] J. Zhang, J. Zhang, J. Chen and S. Yu, “Gan enhanced membership inference: A passive local attack in federated learning,” in *ICC 2020-2020 IEEE International Conference on Communications (ICC),* 2020.

[18] R. Canetti, U. Feige, O. Goldreich and M. Naor, “Adaptively secure multi-party computation,” in *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing,* 1996.

[19] R. Cramer, I. Damgård and U. Maurer, “General secure multi-party computation from any linear secret-sharing scheme,” in *Advances in Cryptology—EUROCRYPT 2000: International Conference on the Theory and Application of Cryptographic Techniques Bruges, Belgium, May 14–18, 2000 Proceedings 19,* 2000.

[20] Y. Aono, T. Hayashi, L. Wang, S. Moriai and others, “Privacy-preserving deep learning via additively homomorphic encryption,” *IEEE Transactions on Information Forensics and Security,* vol. 13, p. 1333–1345, 2017.

[21] M. Hao, H. Li, G. Xu, S. Liu and H. Yang, “Towards efficient and privacy-preserving federated deep learning,” in *ICC 2019-2019 IEEE international conference on communications (ICC),* 2019.

[22] L. Xie, K. Lin, S. Wang, F. Wang and J. Zhou, “Differentially private generative adversarial network,” *arXiv preprint arXiv:1802.06739,* 2018.

[23] S. Truex, N. Baracaldo, A. Anwar, T. Steinke, H. Ludwig, R. Zhang and Y. Zhou, “A hybrid approach to privacy-preserving federated learning,” in *Proceedings of the 12th ACM workshop on artificial intelligence and security*, 2019.

[24] K. Wei, J. Li, M. Ding, C. Ma, H. H. Yang, F. Farokhi, S. Jin, T. Q. S. Quek and H. V. Poor, “Federated learning with differential privacy: Algorithms and performance analysis,” *IEEE Transactions on Information Forensics and Security,* vol. 15, p. 3454–3469, 2020.

[25] A. El Ouadrhiri and A. Abdelhadi, “Differential privacy for deep and federated learning: A survey,” *IEEE Access*, vol. 10, p. 22359–22380, 2022.

[26] M. Langheinrich, “Privacy by design—principles of privacy-aware ubiquitous systems,” in *International conference on ubiquitous computing*, 2001.

[27] G. D'Acquisto, J. Domingo-Ferrer, P. Kikiras, V. Torra, Y.-A. de Montjoye and A. Bourka, “Privacy by design in big data: an overview of privacy enhancing technologies in the era of big data analytics,” *arXiv preprint arXiv:1512.06000,* 2015.

[28] S. Halder and T. Newe, “Enabling secure time-series data sharing via homomorphic encryption in cloud-assisted IIoT,” *Future Generation Computer Systems,* vol. 133, p. 351–363, 2022.

[29] S. Bouchelaghem and M. Omar, “Secure and efficient pseudonymization for privacy-preserving vehicular communications in smart cities,” *Computers & Electrical Engineering,* vol. 82, p. 106557, 2020.

[30] J. Bethencourt, A. Sahai and B. Waters, “Ciphertext-policy attribute-based encryption,” in *2007 IEEE symposium on security and privacy (SP'07),* 2007.

[31] A. Anjum, T. Ahmed, A. Khan, N. Ahmad, M. Ahmad, M. Asif, A. G. Reddy, T. Saba and N. Farooq, “Privacy preserving data by conceptualizing smart cities using MIDR-Angelization,” *Sustainable cities and society*, vol. 40, p. 326–334, 2018.

[32] S. Pearson and M. Casassa-Mont, “Sticky policies: An approach for managing privacy across multiple parties,” *Computer,* vol. 44, p. 60–68, 2011.

[33] S. N. Matheu, A. Robles Enciso, A. Molina Zarca, D. Garcia-Carrillo, J. L. Hernández-Ramos, J. Bernal Bernabe and A. F. Skarmeta, “Security architecture for defining and enforcing security profiles in DLT/SDN-based IoT systems,” *Sensors,* vol. 20, p. 1882, 2020.

[34] D. Preuveneers, W. Joosen, J. Bernal Bernabe and A. Skarmeta, “Distributed security framework for reliable threat intelligence sharing,” *Security and Communication Networks,* vol. 2020, 2020.

[35] P. Zeng, Z. Zhang, R. Lu and K.-K. R. Choo, “Efficient policy-hiding and large universe attribute-based encryption with public traceability for internet of medical things,” *IEEE Internet of Things Journal,* vol. 8, p. 10963–10972, 2021.

[36] C. Ge, W. Susilo, J. Baek, Z. Liu, J. Xia and L. Fang, “Revocable attribute-based encryption with data integrity in clouds,” *IEEE Transactions on Dependable and Secure Computing,* 2021.

[37] D. Wright and P. De Hert, *Introduction to Privacy Impact Assessment. Privacy Impact Assessment, Law, Governance and Technology Series 6,* Springer, Dordrecht, 2012.

[38] J. Benet, “Ipfs-content addressed, versioned, p2p file system,” *arXiv preprint arXiv:1407.3561,* 2014.

[39] Y. Chen and W.-S. Ku, “Self-encryption scheme for data security in mobile devices,” in *2009 6th IEEE Consumer Communications and Networking Conference,* 2009.

[40] M. R. D. Rahardjo and G. F. Shidik, “Design and implementation of self encryption method on file security,” in *2017 International Seminar on Application for Technology of Information and Communication (iSemantic),* 2017.

[41] C. Paar and J. Pelzl, Understanding cryptography: a textbook for students and practitioners, Springer Science & Business Media, 2009.

[42] D. Eastlake 3rd and T. Hansen, “US secure hash algorithms (SHA and SHA-based HMAC and HKDF),” 2011.

[43] G. Bertoni, J. Daemen, M. Peeters and G. Van Assche, “Keccak sponge function family main document,” *Submission to NIST (Round 2),* vol. 3, p. 320–337, 2009.

[44] M. J. Saarinen and J.-P. Aumasson, “The BLAKE2 cryptographic hash and message authentication code (MAC),” 2015.

[45] J. Daemen and V. Rijmen, “AES proposal: Rijndael,” 1999.

[46] K.-F. Hwang and C.-C. Chang, “A self-encryption mechanism for authentication of roaming and teleconference services,” *IEEE Transactions on Wireless Communications,* vol. 2, p. 400–407, 2003.

[47] D. Yao, N. Fazio, Y. Dodis and A. Lysyanskaya, “ID-based encryption for complex hierarchies with applications to forward security and broadcast encryption,” in *Proceedings of the 11th ACM conference on Computer and communications security,* 2004.

[48] A. Ometov, S. Bezzateev, N. Mäkitalo, S. Andreev, T. Mikkonen and Y. Koucheryavy, “Multi-factor authentication: A survey,” *Cryptography,* vol. 2, p. 1, 2018.

[49] K. Ezirim, W. Khoo, G. Koumantaris, R. Law and I. M. Perera, “Trusted Platform Module–A Survey,” *The Graduate Center of The City University of New York,* p. 11, 2012.

[50] I. Grishkov, R. Kromes, T. Giannetsos and K. Liang, “ID-based self-encryption via Hyperledger Fabric based smart contract,” *arXiv preprint arXiv:2207.01605,* 2022.

[51] R. T. Moreno, J. García-Rodríguez, J. B. Bernabé and A. Skarmeta, “A trusted approach for decentralised and privacy-preserving identity management,” *IEEE Access,* vol. 9, p. 105788–105804, 2021.

[52] J. Camenisch, M. Drijvers, A. Lehmann, G. Neven and P. Towa, “Short threshold dynamic group signatures,” in *International Conference on Security and Cryptography for Networks,* 2020.

[53] C. Hébant and D. Pointcheval, “Traceable Attribute-Based Anonymous Credentials.,” *IACR Cryptol. ePrint Arch.,* vol. 2020, p. 657, 2020.

[54] O. Sanders, “Efficient redactable signature and application to anonymous credentials,” in *IACR International Conference on Public-Key Cryptography,* 2020.

[55] J. Bobolz, F. Eidens, S. Krenn, S. Ramacher and K. Samelin, “Issuer-hiding attribute-based credentials,” in *International Conference on Cryptology and Network Security,* 2021.

[56] L. Hanzlik and D. Slamanig, “With a little help from my friends: constructing practical anonymous credentials,” in *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security,* 2021.

[57] R. Soltani, U. T. Nguyen and A. An, “A survey of self-sovereign identity ecosystem,” *Security and Communication Networks,* pp. 1--26, 2021.

[58] K. a. J. L. a. K. C. a. G. D. a. V. D. a. L. N. a. S. M. v. d. a. R. J. a. A. S. a. L. M. a. o. Wierenga, “EOSC Authentication and Authorization Infrastructure (AAI): Report from the EOSC Executive Board Working Group (WG) Architecture AAI Task Force (TF),” 2021.

[59] M. a. L. N. a. S. U. a. S. M. a. C. K. a. J. J. a. M. S. a. G. D. a. L. M. a. P. S. Hardt, “AARC Blueprint Architecture,” 2019.

[60] G. E. Suh and S. Devadas, “Physical unclonable functions for device authentication and secret key generation,” in *2007 44th ACM/IEEE Design Automation Conference,* 2007.

[61] I. Ali, S. Sabir and Z. Ullah, “Internet of things security, device authentication and access control: a review,” *arXiv preprint arXiv:1901.07309,* 2019.

[62] P. Nespoli, M. Zago, A. H. Celdran, M. G. Perez, F. G. Marmol and F. J. G. Clernente, “A dynamic continuous authentication framework in IoT-enabled environments,” in *2018 Fifth International Conference on Internet of Things: Systems, Management and Security,* 2018.

[63] O. Riva, C. Qin, K. Strauss and D. Lymberopoulos, “Progressive authentication: deciding when to authenticate on mobile phones,” in *21st USENIX Security Symposium (USENIX Security 12),* 2012.

[64] M. Shahzad and M. P. Singh, “Continuous authentication and authorization for the internet of things,” *IEEE Internet Computing,* vol. 21, p. 86–90, 2017.

[65] A. K. Sahu, S. Sharma and R. Raja, “Deep Learning-based Continuous Authentication for an IoT-enabled healthcare service,” *Computers and Electrical Engineering,* vol. 99, p. 107817, 2022.

[66] A. Gupta, M. Miettinen, N. Asokan and M. Nagy, “Intuitive security policy configuration in mobile devices using context profiling,” in *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing,* 2012.

[67] R. Murmuria, J. Medsger, A. Stavrou and J. M. Voas, “Mobile application and device power usage measurements,” in *2012 IEEE Sixth International Conference on Software Security and Reliability,* 2012.

[68] R. Murmuria, A. Stavrou, D. Barbará and D. Fleck, “Continuous authentication on mobile devices using power consumption, touch gestures and physical movement of users,” in *International Symposium on Recent Advances in Intrusion Detection*, 2015.

[69] B. Corn, A. J. Perez, A. Ruiz, C. Cetin and J. Ligatti, “An evaluation of the power consumption of coauthentication as a continuous user authentication method in mobile systems,” in *Proceedings of the 2020 ACM Southeast Conference*, 2020.

[70] F. H. Al-Naji and R. Zagrouba, “CAB-IoT: Continuous authentication architecture based on Blockchain for internet of things,” *Journal of King Saud University-Computer and Information Sciences*, 2020.

[71] M. A. Khan, M. M. Jamali, T. Maksymyuk and J. Gazda, “A blockchain token-based trading model for secondary spectrum markets in future generation mobile networks,” *Wireless Communications and Mobile Computing*, vol. 2020, 2020.

[72] S. Mangard, E. Oswald and T. Popp, Power analysis attacks: Revealing the secrets of smart cards, vol. 31, Springer Science & Business Media, 2008.

[73] Y. Ishai, A. Sahai and D. Wagner, “Private circuits: Securing hardware against probing attacks,” in *Annual International Cryptology Conference*, 2003.

[74] G. Fumaroli, A. Martinelli, E. Prouff and M. Rivain, “Affine masking against higher-order side channel analysis,” in *International Workshop on Selected Areas in Cryptography*, 2010.

[75] J. Balasch, S. Faust, B. Gierlichs, C. Paglialonga and F.-X. Standaert, “Consolidating inner product masking,” in *International Conference on the Theory and Application of Cryptology and Information Security*, 2017.

[76] A. Beckers, L. Wouters, B. Gierlichs, B. Preneel and I. Verbauwhede, “Provable Secure Software Masking in the Real-World,” in *International Workshop on Constructive Side-Channel Analysis and Secure Design*, 2022.

[77] P. Luo, L. Zhang, Y. Fei and A. A. Ding, “Towards secure cryptographic software implementation against side-channel power analysis attacks,” in *2015 IEEE 26th International Conference on Application-specific Systems, Architectures and Processors (ASAP)*, 2015.

[78] M. Rivain, E. Prouff and J. Doget, “Higher-order masking and shuffling for software implementations of block ciphers,” in *International Workshop on Cryptographic Hardware and Embedded Systems*, 2009.

[79] G. Agosta, A. Barenghi, G. Pelosi and M. Scandale, “Enhancing Passive Side-Channel Attack Resilience through Schedulability Analysis of Data-Dependency Graphs,” in *Network and System Security*, Berlin, 2013.

[80] D. Couroussé, T. Barry, B. Robisson, P. Jaillon, O. Potin and J.-L. Lanet, “Runtime code polymorphism as a protection against side channel attacks,” in *IFIP International Conference on Information Security Theory and Practice*, 2016.

[81] M. Tunstall and O. Benoit, “Efficient use of random delays in embedded software,” in *IFIP International Workshop on Information Security Theory and Practices*, 2007.

[82] J.-S. Coron and I. Kizhvatov, “An efficient method for random delay generation in embedded software,” in *International Workshop on Cryptographic Hardware and Embedded Systems*, 2009.

[83] J.-S. Coron and I. Kizhvatov, “Analysis and improvement of the random delay countermeasure of CHES 2009,” in *International Workshop on Cryptographic Hardware and Embedded Systems*, 2010.

[84] A. G. Bayrak, N. Velickovic, P. Ienne and W. Burleson, “An architecture-independent instruction shuffler to protect against side-channel attacks,” *ACM Transactions on Architecture and Code Optimization (TACO),* vol. 8, p. 1–19, 2012.

[85] A. G. Bayrak, F. Regazzoni, D. Novo, P. Brisk, F.-X. Standaert and P. Ienne, “Automatic application of power analysis countermeasures,” *IEEE Transactions on Computers,* vol. 64, p. 329–341, 2013.

[86] G. Agosta, A. Barenghi and G. Pelosi, “A code morphing methodology to automate power analysis countermeasures,” in *Proceedings of the 49th Annual Design Automation Conference*, 2012.

[87] N. Belleville, D. Couroussé, K. Heydemann and H.-P. Charles, “Automated software protection for the masses against side-channel attacks,” *ACM Transactions on Architecture and Code Optimization (TACO),* vol. 15, p. 1–27, 2018.

[88] N. Belleville, D. Couroussé, K. Heydemann, Q. Meunier and I. B. El Ouahma, “Maskara: Compilation of a Masking Countermeasure With Optimized Polynomial Interpolation,” *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems,* vol. 39, p. 3774–3786, 2020.

[89] A. Ghosh, M. Nashaat, J. Miller, S. Quader and C. Marston, “A comprehensive review of tools for exploratory analysis of tabular industrial datasets,” *Visual Informatics,* vol. 2, p. 235–253, 2018.

[90] J. C. Roberts, “State of the art: Coordinated & multiple views in exploratory visualization,” in *Fifth international conference on coordinated and multiple views in exploratory visualization (CMV 2007)*, 2007.

[91] A. Athira, D. Dondorp, J. Rudolf, O. Peytral and M. Chatzigeorgiou, “Comprehensive analysis of locomotion dynamics in the protochordate Ciona intestinalis reveals how neuromodulators flexibly shape its behavioral repertoire,” *Plos Biology,* vol. 20, p. e3001744, 2022.

[92] D. De Paepe, S. V. Hautte, B. Steenwinckel, F. De Turck, F. Ongenae, O. Janssens and S. Van Hoecke, “A generalized matrix profile framework with support for contextual series analysis,” *Engineering Applications of Artificial Intelligence,* vol. 90, p. 103487, 2020.

[93] C.-C. M. a. Z. Y. a. U. L. a. B. N. a. D. Y. a. D. H. A. a. S. D. F. a. M. A. a. K. E. Yeh, “Matrix profile I: all pairs similarity joins for time series: a unifying view that includes motifs, discords and shapelets,” in *IEEE*, 2016.

[94] M. &. J. S. K, “Automated machine learning.,” in *Int. J Adv. Res. Innov Ideas Educ. 6*, 2021.

[95] T. Tornede, A. Tornede, J. Hanselle, M. Wever, F. Mohr and E. Hüllermeier, “Towards Green Automated Machine Learning: Status Quo and Future Directions,” *arXiv preprint arXiv:2111.05850,* 2021.

[96] R. Tu, N. Roberts, V. Prasad, S. Nayak, P. Jain, F. Sala, G. Ramakrishnan, A. Talwalkar, W. Neiswanger and C. White, “AutoML for Climate Change: A Call to Action,” *arXiv preprint arXiv:2210.03324,* 2022.

[97] A. Alsharef, K. Aggarwal, M. Kumar, A. Mishra and others, “Review of ML and AutoML solutions to forecast time-series data,” *Archives of Computational Methods in Engineering,* vol. 29, p. 5297–5311, 2022.

[98] E. LeDell and S. Poirier, “H2o automl: Scalable automatic machine learning,” in *Proceedings of the AutoML Workshop at ICML*, 2020.

[99] M. Feurer, A. Klein, K. Eggensperger, J. T. Springenberg, M. Blum and F. Hutter, “Auto-sklearn: Efficient and Robust Automated Machine Learning,” in *Automated Machine Learning: Methods,*

Systems, Challenges, F. Hutter, L. Kotthoff and J. Vanschoren, Eds., Cham, Springer International Publishing, 2019, p. 113–134.

[100] N. Erickson, J. Mueller, A. Shirkov, H. Zhang, P. Larroy, M. Li and A. Smola, “Autogluon-tabular: Robust and accurate automl for structured data,” *arXiv preprint arXiv:2003.06505,* 2020.

[101] L. Kotthoff, C. Thornton, H. H. Hoos, F. Hutter and K. Leyton-Brown, “Auto-WEKA: Automatic model selection and hyperparameter optimization in WEKA,” in *Automated machine learning*, Springer, Cham, 2019, p. 81–95.

[102] H. Jin, Q. Song and X. Hu, “Auto-keras: An efficient neural architecture search system,” in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019.

[103] B. McMahan, E. Moore, D. Ramage, S. Hampson and B. A. y Arcas, “Communication-efficient learning of deep networks from decentralized data,” in *Artificial intelligence and statistics*, 2017.

[104] T. Li, A. K. Sahu, A. Talwalkar and V. Smith, “Federated learning: Challenges, methods, and future directions,” *IEEE Signal Processing Magazine,* vol. 37, p. 50–60, 2020.

[105] H. B. McMahan, D. Ramage, K. Talwar and L. Zhang, “Learning differentially private recurrent language models,” *arXiv preprint arXiv:1710.06963,* 2017.

[106] K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan, S. Patel, D. Ramage, A. Segal and K. Seth, “Practical secure aggregation for privacy-preserving machine learning,” in *proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security,* 2017.

[107] N. Agarwal, A. T. Suresh, F. X. X. Yu, S. Kumar and B. McMahan, “cpSGD: Communication-efficient and differentially-private distributed SGD,” *Advances in Neural Information Processing Systems,* vol. 31, 2018.

[108] K. Bonawitz, H. Eichner, W. Grieskamp, D. Huba, A. Ingerman, V. Ivanov, C. Kiddon, J. Konečnỳ, S. Mazzocchi, B. McMahan and others, “Towards federated learning at scale: System design,” *Proceedings of Machine Learning and Systems,* vol. 1, p. 374–388, 2019.

[109] S. Banabilah, M. Aloqaily, E. Alsayed, N. Malik and Y. Jararweh, “Federated learning review: Fundamentals, enabling technologies, and future applications,” *Information Processing & Management,* vol. 59, p. 103061, 2022.

[110] S. K. Lam, A. Pitrou and S. Seibert, “Numba: A llvm-based python jit compiler,” in *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC,* 2015.

[111] B. Catanzaro, M. Garland and K. Keutzer, “Copperhead: compiling an embedded data parallel language,” in *Proceedings of the 16th ACM symposium on Principles and practice of parallel programming*, 2011.

[112] S. Herhut, R. L. Hudson, T. Shpeisman and J. Sreeram, “River Trail: A path to parallelism in JavaScript,” *ACM SIGPLAN Notices,* vol. 48, p. 729–744, 2013.

[113] J. Wang, N. Rubin and S. Yalamanchili, “Paralleljs: An execution framework for javascript on heterogeneous systems,” in *Proceedings of Workshop on General Purpose Processing Using GPUs*, 2014.

[114] J. Clow, G. Tzimpragos, D. Dangwal, S. Guo, J. McMahan and T. Sherwood, “A pythonic approach for rapid hardware prototyping and instrumentation,” in *2017 27th International Conference on Field Programmable Logic and Applications (FPL)*, 2017.

[115] S. Skalicky, J. Monson, A. Schmidt and M. French, “Hot & spicy: Improving productivity with python and HLS for FPGAs,” in *2018 IEEE 26th Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM),* 2018.

[116] T. Chen, T. Moreau, Z. Jiang, H. Shen, E. Q. Yan, L. Wang, Y. Hu, L. Ceze, C. Guestrin and A. Krishnamurthy, “TVM: end-to-end optimization stack for deep learning,” *arXiv preprint arXiv:1802.04799,* vol. 11, p. 20, 2018.

[117] Y. N. Khalid, M. Aleem, U. Ahmed, M. A. Islam and M. A. Iqbal, “Troodon: A machine-learning based load-balancing application scheduler for CPU–GPU system,” *Journal of Parallel and Distributed Computing,* vol. 132, p. 79–94, 2019.

[118] C.-K. Luk, S. Hong and H. Kim, “Qilin: exploiting parallelism on heterogeneous multiprocessors with adaptive mapping,” in *2009 42nd Annual IEEE/ACM International Symposium on Microarchitecture (MICRO),* 2009.

[119] W. F. Ogilvie, P. Petoumenos, Z. Wang and H. Leather, “Fast automatic heuristic construction using active learning,” in *Languages and Compilers for Parallel Computing: 27th International Workshop, LCPC 2014, Hillsboro, OR, USA, September 15-17, 2014, Revised Selected Papers 27,* 2015.

[120] D. Grewe, Z. Wang and M. F. P. O'Boyle, “Portable mapping of data parallel programs to opencl for heterogeneous systems,” in *Proceedings of the 2013 IEEE/ACM International Symposium on Code Generation and Optimization (CGO),* 2013.

[121] Y. Wen, Z. Wang and M. F. P. O'boyle, “Smart multi-task scheduling for OpenCL programs on CPU/GPU heterogeneous platforms,” in *2014 21st International conference on high performance computing (HiPC),* 2014.

[122] I. Baldini, S. J. Fink and E. Altman, “Predicting gpu performance from cpu runs using machine learning,” in *2014 IEEE 26th International Symposium on Computer Architecture and High Performance Computing,* 2014.

[123] L. Braun, S. Nikas, C. Song, V. Heuveline and H. Fröning, “A simple model for portable and fast prediction of execution time and power consumption of GPU kernels,” *ACM Transactions on Architecture and Code Optimization (TACO),* vol. 18, p. 1–25, 2020.

[124] A. Adams, K. Ma, L. Anderson, R. Baghdadi, T.-M. Li, M. Gharbi, B. Steiner, S. Johnson, K. Fatahalian, F. Durand and others, “Learning to optimize halide with tree search and random programs,” *ACM Transactions on Graphics (TOG),* vol. 38, p. 1–12, 2019.

[125] J. Fumero, M. Papadimitriou, F. S. Zakkak, M. Xekalaki, J. Clarkson and C. Kotselidis, “Dynamic application reconfiguration on heterogeneous hardware,” in *Proceedings of the 15th ACM SIGPLAN/SIGOPS International Conference on Virtual Execution Environments,* 2019.

[126] M. Papadimitriou, E. Markou, J. Fumero, A. Stratikopoulos, F. Blanaru and C. Kotselidis, “Multiple-tasks on multiple-devices (MTMD): exploiting concurrency in heterogeneous managed runtimes,” in *Proceedings of the 17th ACM SIGPLAN/SIGOPS international conference on virtual execution environments,* 2021.

[127] J. Tang, Y. Cui, Q. Li, K. Ren, J. Liu and R. Buyya, “Ensuring security and privacy preservation for cloud data services,” *ACM Computing Surveys (CSUR),* vol. 49, p. 1–39, 2016.

[128] A. J. Grosso, G. W. Leffard and J. G. O'dwyer, *System and method for analyzing privacy breach risk data,* Google Patents, 2014.

[129] S. Todd, R. Baldwin, D. Dietrich and W. A. Pauley Jr, *Privacy scoring for cloud services,* Google Patents, 2016.

[130] C. Dhasarathan, V. Thirumal and D. Ponnurangam, "Data privacy breach prevention framework for the cloud service," *Security and Communication Networks,* vol. 8, p. 982–1005, 2015.

[131] R. G. Pensa and G. Di Blasi, "A privacy self-assessment framework for online social networks," *Expert Systems with Applications,* vol. 86, p. 18–31, 2017.

[132] M. Gharib, P. Giorgini and J. Mylopoulos, "An ontology for privacy requirements via a systematic literature review," *Journal on Data Semantics,* vol. 9, p. 123–149, 2020.

[133] S. J. De and D. Le Métayer, "PRIAM: a privacy risk analysis methodology," in *Data Privacy Management and Security Assurance: 11th International Workshop, DPM 2016 and 5th International Workshop, QASA 2016, Heraklion, Crete, Greece, September 26-27, 2016, Proceedings 11*, 2016.

[134] L. Sion, D. Van Landuyt, K. Wuyts and W. Joosen, "Privacy risk assessment for data subject-aware threat modeling," in *2019 IEEE Security and Privacy Workshops (SPW)*, 2019.

[135] F. Karegar, N. Gerber, M. Volkamer and S. Fischer-Hübner, "Helping john to make informed decisions on using social login," in *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*, 2018.

[136] S. J. De and A. Imine, "To reveal or not to reveal: balancing user-centric social benefit and privacy in online social networks," in *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*, 2018.

[137] S. J. De and D. Le Métayer, "Privacy risk analysis to enable informed privacy settings," in *2018 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, 2018.

[138] P. Silva, C. Gonçalves, N. Antunes, M. Curado and B. Walek, "Privacy risk assessment and privacy-preserving data monitoring," *Expert Systems with Applications,* vol. 200, p. 116867, 2022.

[139] M. Friedewald, M. Önen, E. Lievens, S. Krenn and S. Fricker, Privacy and identity management: data for better living: AI and privacy, Springer, 2019.

[140] P. Silva, R. Casaleiro, P. Simões, N. Antunes, M. Curado and E. Monteiro, "Risk management and privacy violation detection in the PoSeID-on data privacy platform," *SN Computer Science,* vol. 1, p. 1–10, 2020.

[141] M. Maass, P. Wichmann, H. Pridöhl and D. Herrmann, "Privacyscore: Improving privacy and security via crowd-sourced benchmarks of websites," in *Annual Privacy Forum*, 2017.

[142] O. Starov and N. Nikiforakis, "Privacymeter: Designing and developing a privacy-preserving browser extension," in *International Symposium on Engineering Secure Software and Systems*, 2018.

[143] Y. Lu, S. Li, A. Ioannou and I. Tussyadiah, "From data disclosure to privacy nudges: a privacy-aware and user-centric personal data management framework," in *International Conference on Dependability in Sensor, Cloud, and Big Data Systems and Applications*, 2019.

[144] Y. Lu, L. Shujun and others, "From data flows to privacy issues: A user-centric semantic model for representing and discovering privacy issues," in *Proceedings of the 53rd Hawaii International Conference on System Sciences*, 2020.

[145] M. Weinmann, C. Schneider and J. vom Brocke, *Digital nudging. Bus Inf Syst Eng 58 (6): 433–436,* 2016.

[146] Y. Lu and S. Li, "From data flows to privacy-benefit trade-offs: A user-centric semantic model," *Security and Privacy,* p. e225, 2022.

[147] M. N. Alraja, H. Barhamgi, A. Rattrout and M. Barhamgi, “An integrated framework for privacy protection in IoT—Applied to smart healthcare,” *Computers & Electrical Engineering,* vol. 91, p. 107060, 2021.

[148] A. L. Buczak and E. Guven, “A survey of data mining and machine learning methods for cyber security intrusion detection,” *IEEE Communications surveys & tutorials,* vol. 18, p. 1153–1176, 2015.

[149] D. K. Sharma, J. Mishra, A. Singh, R. Govil, G. Srivastava and J. C.-W. Lin, “Explainable artificial intelligence for cybersecurity,” *Computers and Electrical Engineering,* vol. 103, p. 108356, 2022.

[150] M. Van Lent, W. Fisher and M. Mancuso, “An explainable artificial intelligence system for small-unit tactical behavior,” in *Proceedings of the national conference on artificial intelligence*, 2004.

[151] B. Mahbooba, M. Timilsina, R. Sahal and M. Serrano, “Explainable artificial intelligence (XAI) to enhance trust management in intrusion detection systems using decision tree model,” *Complexity,* vol. 2021, 2021.

[152] M. Avgerinou, P. Bertoldi and L. Castellazzi, “Trends in data centre energy consumption under the european code of conduct for data centre energy efficiency,” *Energies,* vol. 10, p. 1470, 2017.

[153] E. Oró, V. Depoorter, A. Garcia and J. Salom, “Energy efficiency and renewable energy integration in data centres. Strategies and modelling review,” *Renewable and Sustainable Energy Reviews,* vol. 42, p. 429–445, 2015.

[154] H. Rong, H. Zhang, S. Xiao, C. Li and C. Hu, “Optimizing energy consumption for data centers,” *Renewable and Sustainable Energy Reviews,* vol. 58, p. 674–691, 2016.

[155] C. Peoples, G. Parr and S. McClean, “Energy-aware data centre management,” in *2011 National Conference on Communications (NCC)*, 2011.

[156] Í. Goiri, K. Le, M. E. Haque, R. Beauchea, T. D. Nguyen, J. Guitart, J. Torres and R. Bianchini, “Greenslot: scheduling energy consumption in green datacenters,” in *Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis*, 2011.

[157] J. C. Salinas-Hilburg, M. Zapater, J. M. Moya and J. L. Ayala, “Energy-aware task scheduling in data centers using an application signature,” *Computers & Electrical Engineering,* vol. 97, p. 107630, 2022.

[158] R. Carroll, S. Balasubramaniam, D. Botvich and W. Donnelly, “Application of genetic algorithm to maximise clean energy usage for data centres,” in *Bio-Inspired Models of Network, Information, and Computing Systems: 5th International ICST Conference, BIONETICS 2010, Boston, USA, December 1-3, 2010, Revised Selected Papers 5*, 2012.

[159] G. Agosta, A. Barenghi, G. Pelosi and M. Scandale, “The MEET approach: Securing cryptographic embedded software against side channel attacks,” *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems,* vol. 34, p. 1320–1333, 2015.

[160] I. I. Consortium and others, “The industrial internet of things: Managing and assessing trustworthiness for iiot in practice,” *Whitepaper] Retrieved June,* vol. 10, p. 2020, 2019.

[161] H. Levenstein, Revolution at the table: the transformation of the American diet, vol. 7, Univ of California Press, 2003.

[162] A. I. HLEG, *High-level expert group on artificial intelligence*, European Commission. Available at: https://ec. europa. eu/digital-single …, 2019, p. 6.

[163] J. W. Tukey and others, Exploratory data analysis, vol. 2, Reading, MA, 1977.

[164] T. Lyons, “EU Blockchain Observatory and Forum,” in *Workshop Report. Government Services and Digital Identity. Brussels, July 5*, 2018.

[165] T. a. M. T. a. J. Z. a. S. H. a. Y. E. Q. a. W. L. a. H. Y. a. C. L. a. G. C. a. K. A. Chen, “TVM: end-to-end optimization stack for deep learning,” 2018.