Tailored IoT & BigData Sandboxes and Testbeds for Smart, Autonomous and Personalized Services in the European Finance and Insurance Services Ecosystem

Infinitech

D4.13 – Encrypted Data Querying and Personal Data Market - I

Lead Beneficiary	**FBK**
Due Date	2020-11-30
Delivered Date	2020-12-14
Revision Number	3.0
Dissemination Level	Public (PU)
Type	Report (R)
Document Status	Final
Review Status	Internally Reviewed and Quality Assurance Reviewed
Document Acceptance	WP Leader Accepted and Coordinator Accepted
EC Project Officer	Pierre-Paul Sondag

HORIZON 2020 - ICT-11-2018

This project has received funding from the European Union’s horizon 2020 research and innovation programme under grant agreement no 856632

Contributing Partners

Partner Acronym	Role[1]	Name Surname[2]
FBK	Lead Beneficiary	Bruno Lepri, Raman Kazhamiakin, Gabriele Santin, Lorenzo Lucchini
IBM	Contributor	Fabiana Fournier, Inna Skarbovsky
GLA	Internal Reviewer	Iadh Ounis
UPRC	Internal Reviewer	Dimosthenis Kyriazis
INNOV	Quality Assurance	John Soldatos

Revision History

Version	**Date**	**Partner(s)**	**Description**
0.1	2020-11-10	FBK	ToC Version
0.2	2020-11-18	FBK	Contributions on Sections 2, 3
0.3	2020-11-23	FBK	Contribution on Section 4
0.4	2020-11-25	IBM	Contribution on Section 6
0.5	2020-12-02	FBK	Contribution on Section 1
0.6	2020-12-04	FBK	Contribution on Section 5
1.0	2020-12-07	FBK	First Version for Internal Review
1.1	2020-12-08	GLA	Internal Review
1.2	2020-12-09	UPRC	Internal Review
2.0	2020-12-10	FBK	Version for Quality Assurance
3.0	2012-12-13	FBK	Version for Submission

[1] Lead Beneficiary, Contributor, Internal Reviewer, Quality Assurance

[2] Can be left void

Executive Summary

This deliverable (D4.13) is the first of three deliverables planned in the scope of Task 4.5 of INFINITECH project. The purpose of this task is to overcome the current limitations of standard data sharing paradigms, and this ambitious goal is achieved by the design and implementation of a framework for securely querying, processing, and analyzing data over the INFINITECH permissioned blockchain infrastructure. This goal will enable decentralized and secure execution of ML algorithms and lay the foundation for a personal data market.

Within this general vision, deliverable D4.13 is designed to analyze the current state of the field and to highlight the shortcomings of the current approaches, to survey interesting existing models, and especially to articulate the details of a newly proposed framework and its components.

The deliverable starts by an account of the current rise of the role of data, in particular personal data, and the importance of its secure storage, sharing, and manipulation, and with the analysis of two models that address these issues and are of partial inspiration for our framework (Section 2). In particular, we argue that the current siloed approaches are responsible for the lack of innovation, but also that any new paradigm needs to be aware of the risks in terms of possible abuses (i.e. privacy violation, discrimination, etc.) in the usage of personal data.

To this end, the core part of the deliverable (Section 3) is devoted to the design of the new framework for secure access, management, and sharing of data in INFINITECH. Our framework assumes an organization model where different institutions collaborate in a shared environment in a federated manner, and it provides secure strategies to access, manage, and share critical and sensitive data. A detailed implementation of the framework is provided, including the description of its components and their interactions towards the provision of the required functionality.

The final part of the deliverable describes the ongoing design, research, and implementation of three of these specific components. Section 4 defines Attribute-based Access Control (ABAC) as a module in the Data Policy Framework. This is a model that allows an organization to define who may access the data, for which purpose (scope), and how (algorithm). Section 5 defines approaches for federated and distributed learning as a component of the Algorithm Runtime. These models allow to run ML algorithms by distributing insights instead of data, and the current research, conducted by FBK within the INFINITECH project, is addressing both the efficiency and the privacy guarantees of these algorithms. The personal data marketplace developed in collaboration between FBK and IBM is discussed in Section 6, which explains how to leverage tokens to realize a data trading infrastructure.

This report provides the foundation for the development of a secure environment for access, management, and sharing of data in INFINITECH. Principles, objectives, and motivations are clearly delineated, and they are implemented in the new data framework, which is a concrete model where multiple components can be combined in an orchestrated manner. Beyond the scope of this deliverable, these principles may serve as a guideline for any data-critical operation within the INFINITECH project.

Table of Contents

List of Figures

Abbreviations

ABAC	Attribute-based Access Control
API	Application Programming Interface
DL	Deep Learning
GDPR	General Data Protection Regulation
MIT	Massachusetts Institute of Technology
ML	Machine Learning
MPC	Multi-Party Computation
OPAL	Open Algorithms
P2P	Peer-to-Peer
PAP	Policy Administration Point
PDP	Policy Decision Point
PDS	Personal Data Store
PEP	Policy Enforcement Point
PII	Personally Identifiable Information
PIP	Policy Information Point
PRP	Policy Retrieval Point
RA	Reference Architecture
SSN	Social Security number
UMA	User-managed Access
XACML	eXtensible Access Control Markup Language

1. Introduction

The current deliverable is the first one of a series of three deliverables whose aim is to describe the activities conducted in Task 4.5 "Secure and Encrypted Queries over Blockchain Data" of the INFINITECH project. The main objective of this task is the design and implementation of a framework for querying encrypted data over the INFINITECH permissioned blockchain infrastructure and for running Machine Learning (ML) algorithms on them.

The inspiration for this framework comes from two recent approaches proposed by MIT Connection Science[3] and other partners (e.g., Imperial College London, Data-Pop Alliance[4]): Open Algorithms (OPAL) [7, 14] and ENIGMA [28]. Both approaches provide a mechanism for the privacy-preserving sharing of data across multiple data repositories. In particular, OPAL (see Section 2.1) introduces the novel concept of moving the ML algorithms to the data repositories, where each data repository participating in the computation performs all its computations behind the firewalls. In this way, OPAL avoids the sharing of raw data. Additionally, ENIGMA (see Section 2.2) introduces the notion of Multi-Party Computation (MPC) [18] that gives the data repositories the ability to collectively perform an algorithm computation that produces some result without revealing the raw data. ENIGMA also encrypts each data item in a repository using a *linear secret sharing scheme* [22, 28]. In this way, a collective computation in a Multi-Party Computation (MPC) instance can be performed on the encrypted datasets without decrypting them.

Task T4.5, in collaboration with Task T4.4 "Tokenization and Smart Contracts Finance and Insurance Services" of the INFINITECH project, has also the objective of laying the foundation of a *personal data market*, where individuals and organizations will be able to trade their data in exchange for tokens or other assets.

In the current deliverable, we motivate (Section 2) the need for a paradigm change concerning the currently dominant model of siloed data collection, management, and exploitation, which makes efficient and privacy-preserving sharing of data difficult, and precludes participation to a wide range of actors, most notably the producers of data (i.e., the customers of services). Afterwards, we introduce and describe the aforementioned frameworks: OPAL (Section 2.1), and ENIGMA (Section 2.2).

In Section 3, we illustrate the proposed INFINITECH framework for securely accessing, managing and sharing data, in particular personal data, between customers, financial and insurance institutions. Our framework assumes an organization model where different institutions (e.g., banks, insurance companies, other public and private institutions) work together in a shared environment and a federated manner. The critical and sensitive data managed by individual organizations often cannot be shared with other organizations in their raw form. Hence, we envisage the presence of the "On-Premise Nodes", which represent the infrastructure managed directly by the organizations outside of the shared environment.

Following the presentation of the INFINITECH framework, we focus on the description of the ongoing work in designing and implementing specific components of the framework. First of all, we describe the *Data Policy Framework* (Section 4**),** that allows an organization to define who (i.e., which organizations/app) may access the data, for which purpose (scope), and how (algorithm). In particular, we describe the ongoing work on adopting the *Attribute-Based Access Control* (ABAC) model for

[3] https://connection.mit.edu/

[4] https://datapopalliance.org/

defining the access rules. Then, we introduce our preliminary work on developing federated ML algorithms able to run on secure (encrypted) data and to share insights in the form of parameters (see Section 5). This work is the basis for the characterization of the *Algorithm Runtime* component. Finally, we report the initial collaboration between FBK and IBM on the *Data Marketplace* component (see Section 6), where we propose to implement a *chaincode* for data trading. More specifically, the idea consists of invoking the tokens chaincode for purchasing and trading data between organizations (e.g., insurance companies) as well as individuals and organizations.

The outcomes of this deliverable will provide a basis for:

- Boosting the GDPR Compliance of INFINITECH solutions, through enabling access to insights and outcomes of ML algorithms which keep private sensitive data used to produce these insights.
- Supporting novel trading solutions in conjunction with INFINITECH's work on blockchain tokenization (deliverable D4.7).

1.1. Objectives of the Deliverable

The main goals of Task 4.5 "Secure and Encrypted Queries over Blockchain Data" are the design and implementation of a framework for querying encrypted data over the INFINITECH permissioned blockchain infrastructure and for running ML algorithms on them, as well as creating the foundation for a personal data market where individuals and organizations will be able to trade their data in exchange for tokens or other assets.

These goals encompass, in this first deliverable, the following specific objectives:

- To analyze and describe existing frameworks for securely accessing, managing and sharing data. As described above, our approach is inspired by two recent frameworks, OPAL and ENIGMA, which have been proposed and are currently under development by MIT Connection Science and other partners (e.g, Imperial College London, Data-Pop Alliance). For this reason, the first step of our work was to analyze and document these two frameworks as an inspiring starting point for the design of the INFINITECH framework.
- To describe the INFINITECH framework for securely accessing, managing and sharing data. This is the main contribution of the current deliverable. In Section 3 we illustrate the overall vision of the INFINITECH framework for securely accessing, managing and sharing data, and we describe its components (e.g., *Data Policy Framework*, *Data Agent and Algorithm Runtime*, *Data Usage Audit*, *Data Marketplace*.).
- To describe the ongoing design, research and implementation work on the components of the INFINITECH framework for securely accessing, managing and sharing data. This first deliverable focuses on the description of the ongoing work on the following 3 components: the *Data Policy Framework* component (Section 4), the *Algorithm Runtime* component (Section 5), and the *Data Marketplace* component (see Section 6). The following deliverable D4.14 "Encrypted Data Querying and Personal Data Market - II" will further describe in detail the design, research, and implementation of these components as well as the *Data Usage Audit* component.

1.2. Insights from other Tasks and Deliverables

The deliverable D4.13 is released in the scope of WP4 “Interoperable Data Exchange and Semantic Interoperability” activities, and documents the preliminary outcomes of the work performed within the context of Task 4.5 “Secure and Encrypted Queries over Blockchain Data”. The task is mostly related to Task 4.3 “Distributed Ledger Technologies for Decentralized Data Sharing” and the corresponding deliverable D4.7 “Permissioned Blockchain for Finance and Insurance”, which was submitted in M12 of the project. The designed permissioned blockchain infrastructure devised in the scope of Task 4.3 and the related technological decisions are relevant for the activities performed in Task 4.5.

In addition, Task 4.5 also relates to Task 4.4 “Tokenization and Smart Contracts Finance and Insurance Services”. Indeed, Task 4.4 aims to enhance the permissioned INFINITECH blockchain infrastructure with tokenization features as a way for enabling assets trading (for example, data trading). To this end, we have established a collaboration between IBM (leading partner of Task 4.4) and FBK (leading partner of Task 4.5) on the usage of tokens for enabling the personal data market (see Section 6 for the description of the preliminary work conducted).

Finally, Task 4.4 relates also to Task 5.4 “Library of ML/DL Algorithms for Financial/Insurance Services” that aims at providing a library of Machine Learning (ML) and Deep Learning (DL) algorithms that will be well-aligned and suitable for finance and insurance applications and used by the INFINITECH pilots. In the next period, partners contributing to Task 4.5 and Task 5.4 will collaborate closely to develop innovative ML algorithms that can run on encrypted data and in a federated manner, as well as to test them with some of the ML tasks identified by the INFINITECH pilots.

1.3. Structure

We organized the structure of D4.13 to be efficiently associated with the objectives described in Section 1.1:

- Section 2 motivates the need for a change of paradigm in the currently dominant model of siloed data collection, management, and exploitation, which makes it difficult to efficiently share data across organizations and between individuals and organizations in a privacy-preserving manner. Then, Section 2 describes in detail the two existing frameworks, OPAL and ENIGMA, that have inspired the INFINITECH approach proposed in Section 3.
- Section 3 describes the proposed INFINITECH framework for securely accessing, managing and sharing data, in particular personal data, between customers, financial and insurance institutions.
- Section 4 describes the ongoing work on designing and implementing the *Data Policy Framework* component. In particular, it motivates the adoption of the *Attribute-Based Access Control* (ABAC) model for defining the access rules, and it documents the initial work in defining the rules.
- Section 5 introduces our preliminary work on the *Algorithm Runtime* component. In particular, we document our current activities and plans for developing federated ML algorithms that leverage secure (encrypted) data and share insights in the form of model parameters.
- We conclude the report with Section 6 reporting the state of the collaboration between IBM and FBK on the *Data Marketplace* component, where we propose to implement a token chaincode for purchasing and trading data between organizations as well as individuals and organizations.

2. The Rise of (Personal) Data Markets

Data is becoming increasingly relevant for the proper functioning of the current and future digital revolution in the finance and insurance sectors. Furthermore, data is crucial in the day-to-day running of banks, insurance companies, businesses, and governmental institutions. Data also increases in value when combined across verticals, thus allowing unforeseen insights [15]. However, the current reality is that cross-organizational data sharing is challenging from a business, risk, and regulatory perspective. For this reason, the current ecosystems that access and use data, and in particular personal sensitive data, are fragmented and inefficient. Besides, for many participants, the risks exceed the economic returns, and personal privacy concerns are inadequately addressed. In sum, the current technologies and legal systems unfortunately fail in providing the technical infrastructure and the legal framework needed to support a well-functioning digital economy in the finance and insurance sectors.

The core problem, then, is to find ways so that an individual or a company can share sensitive data for a financial or insurance service (e.g., investment advice, money management, access to credit), and be certain that the data will only be used for the intended and approved purposes. This task implicitly recognizes the risks in terms of possible abuses in the usage of the data, and in particular of personal data, and also in terms of the “missed chance for innovation”. This second risk is inherent to the current dominant paradigm of siloed data collection, management, and exploitation, which precludes participation by a wide range of actors, most notably the producers of data (i.e., the customers).

New models, often user-centric, have been proposed for personal data management to empower individuals with more control of their own data’s life-cycle [15]. To this end, researchers and companies are developing repositories, which implement medium-grained access control to different kinds of Personally Identifiable Information (PII), such as passwords, Social Security Numbers (SSNs), health status [27], location [13, 3], and personal data collected through smartphones or other smart devices [3, 26, 17]. Previous research has also introduced the concept of *personal data markets* [1, 23, 24] in which individuals sell their data to entities (e.g., insurance companies, banks, other financial institutions) interested in buying it. While personal data markets or, as more recently introduced, *personal data cooperatives* [8] might be the technological future to support a privacy-preserving and transparent use of personal data, they are still at a research stage. To implement a personal data market, it is necessary to connect potential buyers, namely the demand, with sellers, namely the offer, as well as to provide a trustworthy mechanism for the exchange of value between buyers and sellers. In a personal data market scenario, buyers are likely to be companies, while sellers are individuals who receive compensation for sharing their data.

In the next sections, we introduce in detail two frameworks, OPAL [7, 14] and ENIGMA [28] aiming at providing a mechanism for the privacy-preserving sharing of data across multiple data repositories, belonging to individuals or companies. In particular, OPAL introduces the new paradigm of moving the algorithms to the data repositories, where each data repository participating in the computation performs all its computations behind the firewalls. Raw data thus are never exported and shared, and each data repository has the possibility of returning only safe-answers. ENIGMA adds to the OPAL approach the notion of Multi-Party Computation (MPC) [18] that gives to the data repositories the ability to collectively perform an algorithmic computation that produces some result without revealing the raw data. Moreover, ENIGMA also encrypts each data item in a repository using a *linear secret sharing scheme* [22, 28]. In this way, a collective computation in a MPC instance can be performed on these encrypted data (i.e. shares) without decrypting them. For example, a *linear secret sharing* *scheme* [22, 28] may encrypt a data item into 20 shares where a threshold value of any 12 shares is needed to reconstruct the original data item. Thus, encrypting data into multiple shares has the benefit to eliminate centralized data stores and to disperse shares into physically distributed locations. As a consequence, the physical distribution reduces the risks related to attacks because in a distributed framework an attacker needs to compromise at least N (in our case 12) locations.

As we will discuss in Section 4, OPAL and ENIGMA are an important inspiration source for the proposed INFINITECH framework for task T4.5. The next subsections provide a detailed description of the OPAL and ENIGMA systems, respectively, as a first constitutional block favouring the development and achievement of the T4.5 goals.

2.1. Open Algorithms (OPAL)

The OPAL framework [7, 14] has been introduced and is currently developed by MIT Connection Science and several other partners such as the Computational Privacy Group at Imperial College London[5], Data-Pop Alliance, Orange, and Telefonica. A seed project for the OPAL framework was OpenPDS [3] that has introduced the first implementation of a Personal Data Store (PDS), by incorporating the idea of "safe answers" (i.e., the exact information needed to provide a given service) as the norm for the responses generated by a PDS. However, OpenPDS raises challenges around how large data stores could keep their data secure, safeguard the privacy and comply to regulations (e.g., General Data Protection Regulation-GDPR[6]) as well as at the same time could enable collaborative data sharing.

For example, large centralized data repositories are attractive for hackers. For this reason, the importance of replacing the current approach of ML algorithms running on centralized data stores is evident. As a consequence, a new paradigm of thinking about ML algorithms running on data stored in a distributed manner is emerging. In our opinion, the OPAL paradigm provides a useful starting framework for a secure and privacy-preserving sharing of data that can be adopted by the financial and insurance sectors.

Below, we list and discuss the key principles underlying the OPAL framework [7] that are also a source of inspiration for the INFINITECH framework proposed and illustrated in Section 3:

- Moving the algorithms to the data: Instead of storing raw data into a single centralized location for processing, in OPAL the ML algorithms are sent to distributed data repositories, controlled by data owners, and are processed within the data repositories.
- Raw data must never leave the data repository: Raw data must never be exported from its data repository, and must always be under the control of the data owner.
- Vetted algorithms: ML algorithms must be vetted to be "safe" from biases and discriminations, as well as by privacy violations and other unintended consequences. The data owner must ensure that the ML algorithms have been analysed for fairness, safety and privacy-preservation before their publication.
- Provide only safe answers: When an ML algorithm is executed on a dataset, the data repository must always provide answers that are "safe" from a privacy perspective. This means that responses must not release *personally identifiable information* (PII) without the

[5] https://cpg.doc.ic.ac.uk/

[6] https://en.wikipedia.org/wiki/General_Data_Protection_Regulation

consent of the subject. Thus, it may imply that a data repository sometimes returns only aggregate answers.

- Trust networks (data federation): In a scenario of group-based information sharing - referred to as *Trust Network for Data Sharing Federation* in [7] - ML algorithms must be vetted collectively by the members of the trusted network (for example, banks, insurance companies, and other financial institutions).
- Consent for algorithms' execution: Data repositories holding personal data must obtain explicit consent from the individual for the inclusion of this data in a given ML algorithm execution. Moreover, the people's consent should be explicit, unambiguous and retractable (see Article 7 of GDPR).
- Decentralized data architectures: The OPAL framework supports a decentralized architecture for data stores [16]. Hence, *decentralized data architectures* based on APIs or other standardized interfaces should apply to PDSs as legitimate end-points. Additionally, new architectures based on Peer-to-Peer (P2P) networks should be employed as the basis for new decentralized data stores [2]. The motivation beyond the adoption of a decentralized approach is to increase the resilience of the overall system against attacks, for example, data manipulations and thefts. Along this line, MIT Connection Science is also developing new distributed data security solutions, such as Enigma [28] described in Section 3.2, based on secure MPC [18] and homomorphic encryption (i.e., a type of encryption allowing to perform calculations on encrypted data without having to decrypt it first) [21]. Additionally, this approach increases the relevance of developing a federated ML framework [11] that trains an algorithm across multiple decentralized data stores holding local data sets, without exchanging them. Please consult Section 5 for our current work on developing a federated ML approach for INFINITECH.

2.2. ENIGMA

ENIGMA [28], also developed by MIT Connection Science, is a decentralized computation platform based on the Blockchain technology that removes the need for a trusted third party and enables the autonomous control of personal data. Moreover, users can share and sell their private data with cryptographic guarantees regarding their privacy.

From a customer's perspective, ENIGMA is a cloud that ensures both the privacy and integrity of its data. The system also allows any type of computation to be outsourced to the cloud while guaranteeing the privacy of the underlying data and the correctness of the results. A core feature in the system is that it allows the data owners to define and control who can query it, thus ensuring that the approved parties only learn the output. Moreover, no other data is shared in the process to any other party. The ENIGMA cloud is composed of a network of computers that store and execute queries. Using secure MPC [18], each computer only sees random pieces of the data, a fact that prevents information leaks. Furthermore, queries carry a micro-payment for the computing resources as well as to those users whose data is queried, thus providing the foundation for the rise of a *personal data market*.

To illustrate how the platform works, consider the following example: A group of data analysts of an insurance company wishes to test a model that leverages people's GPS location data captured by mobile phones. Instead of sharing their raw GPS data with the data analysts in the insurance company, the users can securely store their data in ENIGMA, and only provide the data analysts with permission to execute their algorithms. The data analysts are thus able to execute their algorithms' code and obtain the results, but nothing else (especially the actual raw data). In the process, the users are compensated for having given access to their GPS location data and the computers in the network are paid for their computing resources.

Three types of entities are defined in ENIGMA (see Figure 1): (i) *owners*, (ii) *services*, and (iii) *parties*. *Owners* are those (e.g., individuals, financial institutions, companies.) sharing their data into the system and controlling who can query it; *services*, if approved, can query the data without learning anything else beyond the answer to their query; and *parties* (or *computing parties*) are the nodes that provide computational and storage resources, but only ever see encrypted or random bits of information. Besides, all entities are connected to a Blockchain as illustrated below.

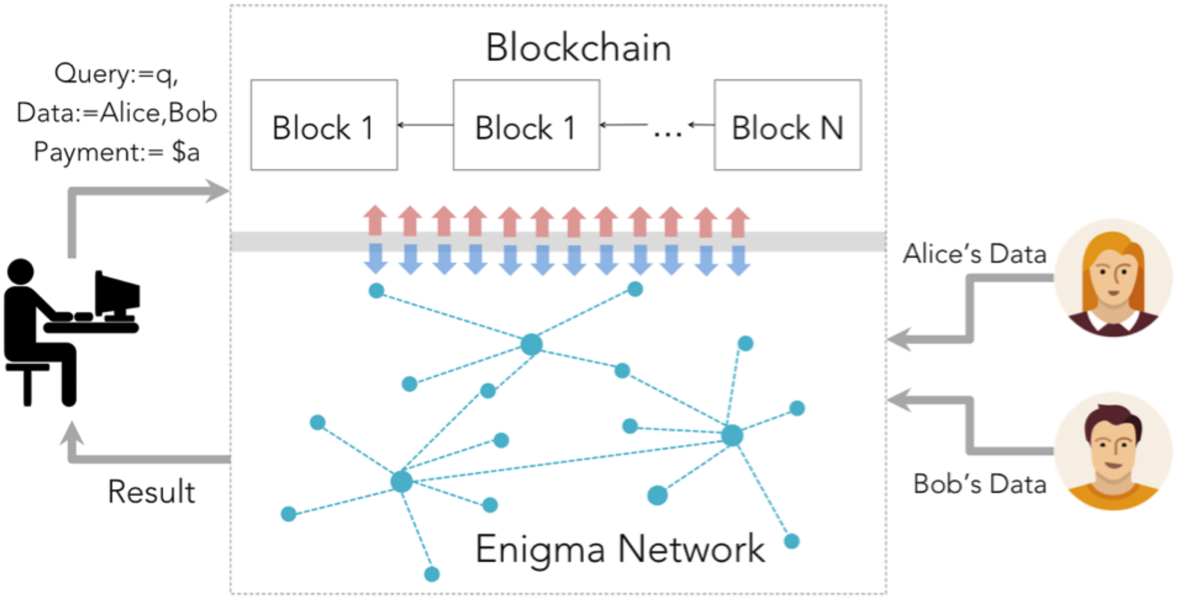

Figure 1 - Overview of the Enigma's decentralized platform.

When owners share data, the data is split into several random pieces called *shares*. *Shares* are created in a process of *secret-sharing* [22] and they perfectly hide the underlying data while maintaining some necessary properties allowing them to later be queried in this masked form. Since users in ENIGMA are the owners of their data, a trusted database to publicly store a proof of who owns which data is needed. However, trusting a centralized database to attest ownership is arguably problematic. For this reason, ENIGMA uses the Blockchain as a decentralized secure database that is not owned by any party [28, 24], and it allows a data owner to designate which services can access its data and under what conditions. This way, when a service requests a computation, parties can query the Blockchain and ensure that it holds the appropriate permissions. Note that parties are only paid when they provide computing resources to an authorized service, so they have no incentive to provide work for unauthorized entities (such behaviour is thus penalized).

3. An INFINITECH Framework for Securely Accessing, Managing, and Sharing Data

In this section, we introduce the proposed INFINITECH framework for securely accessing, managing and sharing data across financial and insurance institutions. This framework assumes an organization model, where different institutions (e.g., companies, public and private institutions) work together in a shared environment (i.e., sandbox) in a federated manner. The critical and sensitive data managed by individual organizations under certain circumstances cannot be shared with third parties in their raw form. Hence, we propose the presence of "*On-Premise Nodes*" that represent the infrastructure managed directly by the organizations outside of the shared environment (c.f. Figure 2 for an architecture representing our vision and the proposed framework).

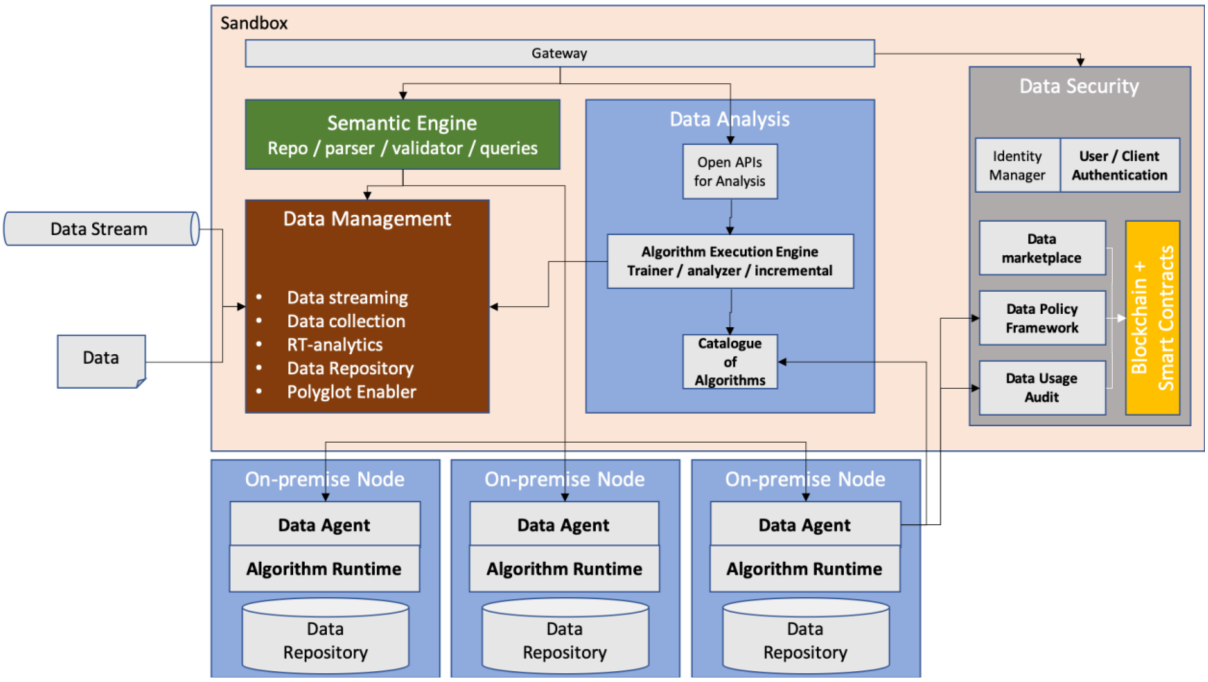

Figure 2 - Representation of the INFINITECH framework for securely accessing, managing and sharing data

The usage of the data may have different forms, where the following two scenarios are probably the most relevant ones: First, an organization works with the data managed by the INFINITECH platform in an aggregated manner, without referring to single individuals. Second, an organization (e.g., an app or a service of the organization) manages data related to a single individual, performing ML activities related for example to customer profiling for credit scoring or to insurance plan offers.

In the former case, the organization providing the data is the owner of that data. In the latter case, the end-user is the owner of the data. Furthermore, the nature of the data may have at least three different forms: (i) open data, (ii) private organization data, and (iii) personal data. Open data may be managed directly in the *Data Management* module (see Figure 2), stored in the centralized repository and shared among the partners for analysis and processing. In some specific scenarios, the private and personal data may also be brought outside of the organizational boundaries but, in most cases, this data should remain in the on-premise storages within the organization boundaries.

Concerning these points, the presented INFINITECH framework focuses only on the aspects related to the scenario where the data (private organizational or personal) should remain within the provider boundaries. More precisely, data access and analytics should be delegated to the corresponding organization. The scenarios where the data is open, or may be managed within the sandbox, may rely on the baseline INFINITECH Reference Architecture (INFINITECH RA) and its implementation (See deliverable D2.13 "INFINITECH Reference Architecture - I" for more details).
Below we describe in detail the main components illustrated in Figure 2:

- Data Management module: It contains components for dealing with the data processing and storage for the scenario, where it is safe to bring the data outside the organization's boundaries. An important requirement here may be to ensure that the module is structured in a multi-tenant way so that single "operators" may elaborate, control, and manage the data independently and in a secure and isolated manner.
- Semantic engine: It is used to model and map the data considering different sources and single representations. The components of the engine are used to mitigate the semantic data queries.
- Gateway: It represents the entry point for accessing the data, and thus for activating the ML algorithm execution. This component relies on the *Data Security* model for controlling and authorizing the access to the underlying modules, as well as performing the *API Management* functionality.
- Data Security module: It defines the components necessary for data/service access, control and monitoring.
- Data Analytics module: It contains the components for defining various ML algorithms and their main activities (e.g., processing, training, analysis) as well as their centralized execution (e.g., incremental, batch).
- APIs for invoking the ML algorithms.

On top of the sandbox, we define the extensions for the federated management of data and ML activities. As previously mentioned, the presented approach is inspired and partly relies on the frameworks described above, OPAL (see Section 2.1) and ENIGMA (see Section 2.2). More precisely, our framework extends the INFINITECH RA as follows:

- Each participating organization that manages critical data is equipped with a module that consists of a *Data Agent* and an *Algorithm Runtime*. The role of the *Data Agent* is to mediate the access to the sensitive data or perform the ML-based training/analyses activities in order to provide safe answers. In this way, the *Data Agent* exposes only the results that are safe with respect to the organization requirements. Specifically, the access to the data or to the ML algorithm API through the *Data Agent* is performed as follows: Each participating organization that manages critical data is equipped with a module that consists of a *Data Agent* and an *Algorithm Runtime*. The role of the *Data Agent* is to intermediate access to sensitive data or perform the ML-based training/analyses activities to provide safe answers. In this way, the *Data Agent* exposes only the results that are safe concerning the organization requirements. Specifically, access to the data or the ML algorithm API through the *Data Agent* is performed as follows:
  - The access request should contain the information about the type of data to be accessed, the subject (i.e., the information about the dataset or the individual), the

operation to be performed (i.e., the algorithm reference as well as the description of the training steps or analyses to perform).

  - Then, the *Data Agent* validates the access with the help of the *Data Policy Framework* (see Section 4) where the constraints about the data access are defined by the data owner (the organization and/or the individual).
  - In the case of policy satisfaction, the algorithm is executed in the local *Algorithm Runtime* and is deployed there if necessary.
  - Finally, the *Data Agent* responds with the algorithm execution's results and logs the data access operation through the *Data Usage Audit* component. The *Data Usage Audit* component should record who is accessing the data (the calling app), which data is used, and which algorithm is being operated.

  It is important to note that, in this approach, only the "controlled/validated" algorithms may be deployed on the local premise to guarantee that no unsafe data or raw data is exposed. The nature of the algorithm may range from the local ML model training to data analysis, to even data transformation/anonymization (if applicable). For the algorithms to be reusable across the different organizations, the approach may rely on similar semantic tools/models used for the INFINITECH RA. The centralized catalogue/library of the ML/DL algorithms of the INFINITECH RA (developed in task T5.4) should also allow for representing (i.e., which dataset/attributes it operates and produces, where it is applicable,.) and validating the federated algorithms.

- The *Data Security* module is equipped with the component to represent and trace the data access policies both for organizations owning critical data and individuals: the *Data Policy Framework*. This framework allows the data owner to define who (i.e., which organizations/apps) may access the data, for which purpose (scope), and how (algorithm). The realization of the policy framework may rely on existing approaches, such as *Attribute-based Access Control* [9, 10], User-managed Access (UMA), etc. Section 4 will introduce the usage of *Attribute-based Access Control* as the approach for realizing the *Data Policy Framework*.
- To ensure transparency and the traceability of data access, the *Data Usage Audit* component records all the operations appropriately signing and certifying them. The implementation relies on the Blockchain technology and on smart contracts tightly integrated with the *Data Policy Framework*. The individuals/organizations may also be incentivized to provide the access to their data through the mechanism of the *Data Marketplace*, where the individuals may associate a "cost" to the access of their data. When the data access is registered, the corresponding smart contracts will emit the "tokens" defined (See Section 6 for the initial plan and the ongoing collaboration between IBM and FBK on this component).

4. Data Policy Framework: An Attribute-Based Access Control Approach

As previously mentioned, the *Data Policy Framework* component uses an *Attribute-Based Access Control* (ABAC) model for defining the access rules [9, 10]. The ABAC model allows granting access rights according to the attributes of the entities involved, namely *subjects* (e.g., individuals, companies, applications), *actions* (e.g., read, write, update, run), *resources* (e.g., a file, a document, a dataset), and *environments* (e.g., contextual information such as location or time of the day).

Although the concept existed for many years, ABAC is considered a novel authorization model because it provides dynamic, context-aware and risk-intelligent access control to resources (i.e., data in our scenario). In particular, ABAC allows access control policies, including specific attributes from many different organizations. These control policies need to be defined to resolve an authorization and to achieve efficient regulatory compliance, thus allowing flexibility in their implementations based on their existing infrastructures.

Usually, ABAC comes with a recommended architecture, namely the following one:

- The *Policy Enforcement Point* (*PEP*) is responsible for protecting the data on which the ABAC rules are applied. More specifically, the *PEP* module has the role of evaluating a data access request and generating an authorization from this request. The authorization is then sent to the *Policy Decision Point* (*PDP*).
- The *Policy Retrieval Point* (*PRP*) retrieves and stores the deployed policies. Policies in ABAC are statements about attributes and they express what is allowed and what is not allowed. Policies can be local or global and can be written in a way that they override other policies.
- The *Policy Decision Point* (*PDP*) is the component that evaluates the incoming requests against the policies. The *PDP* returns a permit/deny decision, and it may also use the *Policy Information Point* (*PIP*) to retrieve missing metadata.
- The *Policy Information Point* (*PIP*) bridges the PDP to external sources of attributes (e.g., to the INFINITECH Blockchain).
- The *Policy Administration Point* (*PAP*) is the architectural entity used to manage policies. These policies are later evaluated by the *PDP*.

We use an eXtensible Access Control Markup Language (XACML) [20] for writing ABAC rules due to the fact that is a de facto standard for ABAC and because there exist open source frameworks for ABAC such as AuthzForce[7] and WSO2[8] using XACML. XACML is an XML-based standard that supports Boolean logic for combining attributes and for writing the rules. An example of an ABAC rule in our scenario is the following one:

IF User = Insurance Data Analyst AND Request State = In Evaluation AND Role Subscription Level ≥ Subscription Level for Reading Data and Write/Run ML Algorithms
THEN Permit.

To comply with the ABAC recommended structure, we can use the back-end of an organization (e.g., a bank, an insurance company,) as the *PEP*. Moreover, we can introduce a new server for the capabilities of the *PDP*. The *PDP* uses the *PIPs*, located on the same machine. The role of the *PIPs* is to talk directly with the data sources (e.g., the database within the organization premises or the data stored on the Blockchain) to obtain the attributes. Once a data access request, sent by the *PEP*, is

[7] http://authzforce.ow2.org
[8] https://github.com/wso2

successfully accepted from the *PDP* server, an authorized token is sent from the *PDP* to the *PEP*, which is then allowed to let the ML algorithm run on the data (for example, on the data stored in the Blockchain).

The next version of the deliverable, namely D4.14 “Encrypted Data Querying and Personal Data Market - II”, will detail the ABAC rules defined in the *Data Policy Framework* as well as will provide a concrete use case.

5. Algorithm Runtime: A Federated Machine Learning Approach for Sharing Insights and not Data

Our proposed framework for securely accessing, managing, and sharing data (see Section 3) needs to be mirrored by a parallel redesign and rethinking of the algorithmic infrastructure (*Algorithm Runtime*) that builds and runs upon these data.

In a rapidly changing data landscape, *distributed* and *federated* learning have recently emerged as new fields of research to provide decentralized and secure counterparts to traditional ML algorithms [11]. In this setting, both the decentralized and the cooperative aspects are of great importance to manage the data according to the principles of Section 3. The need to perform secure and orchestrated data-sharing, as discussed above, requires algorithms that can exploit multiple sources of data to obtain a global gain that may be shared across multiple nodes. In this way, the performance of the single node can be significantly improved by the cooperative action, in a way that makes it possible to break the limitations faced by each single data owner, while maintaining data integrity.

In this general field, *cooperative multi-agent decision making* is a particularly effective framework to enclose many different practical problems that are faced by the finance and insurance institutions, such as portfolio optimization, fraud detection, and credit scoring. In the classical single-agent situation, this approach models an agent (e.g., a user, or an organization) facing the challenge of finding an optimal solution to a given task under partial observations and incomplete information. The agent builds up knowledge about the environment while solving the iterative optimization process: at each iteration, the agent takes a step into the most promising direction within the environment, and the environment, in turn, gives back information on the quality of the move in terms of a reward. This information may be used to reinforce the knowledge that the agent has about the environment, and to plan the next move. The challenge is to balance the exploitation of promising search directions with the exploration of new ones to reach a quasi-optimal solution of the task.

A cooperative layer may be added on top of this framework by connecting several nodes that are trying to solve similar (but not necessarily equal) problems, so that they can leverage communication over a network to improve their overall performance [12].

As argued in Section 3, different scenarios require different solutions, and these correspond to the different nature of the data (i.e., open data, private organization data, and personal data) and data-sharing levels. A currently successful approach is *Coop-KernelUCB*, which is a nonlinear solver based on Kernel methods that significantly outperforms existing benchmarks [5]. Some of its variants include methods for differentially-private cooperative learning, even if only when restricted to lower-accuracy linear problems (*FedUCB*) [4], and for the modelling of private communications and the stochastic disturbance of the information [6]. Furthermore, several other algorithms are being rephrased in a decentralized, federated, and privacy-preserving manner. We mention as a relevant example the case of *Decentralised Regression* [19], that leverages the multi-agent setting to obtain a linear improvement in memory cost over a single agent, and a linear reduction in computational runtime if a sufficiently large amount of data is available.

In the setting of multi-agent decision making, we are currently working on two complementary directions, namely to improve the efficiency of the local nodes (our *Algorithm Runtime* nodes controlled by organizations or individuals) and to increase the privacy guarantees of their data exchanges. Recent work [25] has shown that it is possible to employ state-of-the-art sampling processes in Kernel spaces to simplify the single-users kernelized decision problem by obtaining a very accurate linear approximated problem that can be solved significantly faster. We are extending this approach to the multi-agent setting of *Coop-KernelUCB* to improve the performances of the overall collaborative effort among the agents/nodes. Moreover, by moving to an approximated linear problem, we will be able to extend the differential privacy guarantees of *FedUCB* to the high-accuracy kernelized setting.

These extensions go in the direction of increasing the gain that an individual or an organization (e.g., a bank, an insurance company.) should expect by accepting to collaborate with its peers while providing provable data-privacy guarantees. Both of these conditions are crucial for the acceptance of these methods in the financial and insurance communities.

A further goal in this part of the Task 4.5 is the identification of application scenarios with the INFINITECH pilots. We are particularly looking for well-defined concrete problems, and for some pilots that are willing to test the newly developed frameworks to improve their decision-making algorithms for example in the fields of portfolio optimization, fraud detection, and credit scoring. The identification of such collaborations will be a major occupation in the next months. The successful implementation of this task will also be of interest for the scientific community: the field of federated learning at large still lacks a set of well-established benchmarks with clearly defined evaluation metrics [11], and the real-world problems faced by the INFINITECH pilots offer a great opportunity to contribute towards addressing this challenge.

6. Data Marketplace: Chaincode for Data Trading

One of the goals of the task T4.5 “Secure and Encrypted Queries over Blockchain Data” is to create a foundation for a *personal data market* where users and companies will be able to trade their data in exchange for tokens or other assets. As mentioned in Section 3, this is the role played by the *Data Marketplace* component.

To do this, task T4.5 will leverage the implementation of the token realized in task T4.4 “Tokenization and Smart Contracts Finance and Insurance Services”. For this reason, during October and November 2020 we have started a collaboration between tasks T4.4 and T4.5, and in particular between IBM (leading partner of T4.4) and FBK (leading partner of T4.5).

The initial brainstormed idea (described also in deliverable D4.10 “Blockchain Tokenization and Smart Contracts - I”) is that a chaincode for data trading (T4.5) will invoke the tokens chaincode (T4.4) to perform operations such as transfer of tokens between accounts, checking the balance of accounts, approving a third party to transfer tokens on behalf of an account owner, and more. In this way, organizations can purchase and trade data using a trading platform.

Figure 3 depicts this idea described by the data flow below. Let’s assume we have a Blockchain network with three peers for three insurance companies (insurance company A, insurance company B, and insurance company C). The flow will be composed of the following 2 steps, i.e. write data and read data.

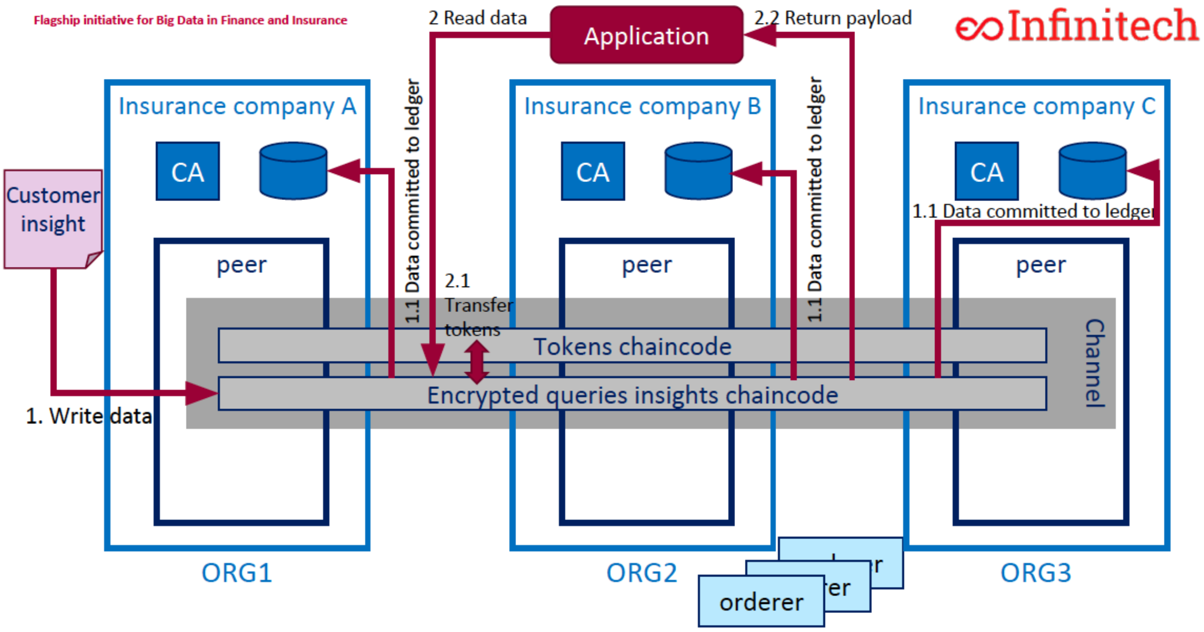

Figure 3 - Tokens Chaincode for trading data

Write data: Insurance company A invokes the data sharing chaincode to write insights on the Blockchain. The data are committed to the ledger of companies A, B, and C

Read data: Insurance company B attempts to read data. The data sharing chaincode checks if insurance company B has permissions to read and, if the condition is met, inquires the number of tokens this data costs and subsequently invokes the token chaincode to transfer tokens to insurance company A. Following these operations, the read query returns with the requested payload.

The next version of the deliverable, namely D4.14 “Encrypted Data Querying and Personal Data Market - II”, will detail the implementation steps of the token chaincode. To this end, IBM and FBK will define a clear use case and will look for the involvement of INFINITECH pilots for the testing of the developed *Data Marketplace* component.

7. Conclusions

The current deliverable has been developed around the three main objectives that have been outlined in Section 1.1.

As the first objective, this deliverable has analysed and described two existing frameworks for securely accessing, managing, and sharing data across organizations and between individuals and organizations: (i) OPAL (Section 2.1) and (ii) ENIGMA (Section 2.2). These two approaches are particularly relevant given the inspiration they provide for the INFINITECH framework we have proposed in Section 3.

The second objective that has been pursued is the description of the INFINITECH framework and of its main components: (i) the *Data Policy Framework*, (ii) the *Data Agent and Algorithm Runtime*, (iii) the *Data Usage Audit*, and (iv) the *Data Marketplace*.

As a final objective, this deliverable has described in detail the ongoing work on the following components: the *Data Policy Framework* (Section 4)**,** the *Algorithm Runtime* (Section 5), and the *Data Marketplace* (Section 6).

As mentioned in the Introduction Section, the final objectives of the task T4.5 and the three associated deliverables (D4.13, D4.14, and D4.15) are the design and implementation of a framework for querying encrypted data over the INFINITECH permissioned blockchain infrastructure and for running ML algorithms on them, as well as the creation of the foundation for a personal data market where individuals and organizations will be able to trade their data in exchange for tokens. To this end, the second deliverable of T4.5 will provide some concrete application scenarios and a more detailed description of the design, research, and implementation of the *Data Policy Framework*, the *Algorithm Runtime*, and of the *Data Marketplace* components as well as a description of the *Data Usage Audit* component.

Appendix A: Literature

[1] Adar, E., and Huberman, B. (2001). A market for secrets. First Monday, 6(8).

[2] De Filippi, P., and McCarthy, S. (2012). Cloud computing: Centralization and data sovereignty- European Journal of Law and Technology, 3(2), available at SSRN: https://ssrn.com/abstract=2167372.

[3] de Montjoye, Y.-A., Shmueli, E., Wang, S.S., and Pentland, A. (2014). OpenPDS: Protecting the privacy of metadata through SafeAnswers. PloS One, 9 (7), e98790.

[4] Dubey, A., and Pentland, A. (2020a). Differentially-private federated linear bandits. In Proceedings of the Conference on Neural Information Processing Systems (NeurIPS 2020).

[5] Dubey, A., and Pentland, A. (2020b). Kernel methods for cooperative multi-agent contextual bandits. In Proceedings of the 37th International Conference on Machine Learning (ICML), 119:2740-2750.

[6] Dubey, A., and Pentland, A. (2020c). Private and byzantine-proof cooperative decision-making. In Proceedings of the International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS 2020), 357-365.

[7] Hardjono, T., and Pentland, A. (2017). Open algorithms for identity federation. arXiv:1705.10880.

[8] Hardjono, T., and Pentland, A. (2019). Data Cooperatives: Towards a foundation for decentralized personal data management. arXiv: 1905.08819.

[9] Hu, V., Ferraiolo, D., Kuhn, D., Friedman, A., Lang, A., Cogdell, M.M., Schnitzer, A., Sandlin, K., Miller, R., and Scarfone, K. (2014). Attribute-Based Access Control Guide to Attribute Based Access Control (ABAC) Definition and Considerations. National Institute of Standards and Technology, NIST SP 800-162.

[10] Hu, V., Kuhn, D., and Ferraiolo, D. (2015). Attribute-based access control. Comput. J. 48, 864–866. doi: 10.1109/MC.2015.33.

[11] Kairouz, P., McMahan, H.B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A.N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., D'Oliveira, R.G.L., El Rouayheb, S., Evans, D., Gardner, J., Garrett, Z., Gascón, A., Ghazi, B., Gibbons, P.B., Gruteser, M., Harchaoui, Z., He, C., He, L., Huo, Z., Hutchinson, B., Hsu, J., Jaggi, M., Javidi, T., Joshi, G., Khodak, M., Konecný, J., Korolova, A, Koushanfar, F., Koyejo, S., Lepoint, T., Liu, Y., Mittal, P., Mohri, M., Nock, R., Özgür, A., Pagh, R., Raykova, M., Qi, H., Ramage, D., Raskar, R., Song, D., Song, W., Stich, S.U., Sun, Z., Suresh, A.T., Tramér, F., Vepakomma, P., Wang, J., Xiong, L., Xu, Z., Yang, Q., Yu, F.X., Yu, H., and Zhao, S.. (2019). Toward trustworthy AI development: Mechanisms for supporting verifiable claims. arXiv:1912.04977.

[12] Landgren, P., Srivastava, V., and Leonard, N.E. (2016). On distributed cooperative decision-making in multi armed bandits. In Proceedings of the IEEE European Control Conference (ECC), 243–248.

[13] Mun, M., Hao, S., Mishra, N., Shilton, K., Burke, J., Estrin, D., Hansen, M., and Govindan, R. (2010). Personal data vaults: A locus of control for personal data streams. In Proceedings of the 6th International Conference Co-NEXT '10, 1–12.

[14] Oehmichen, A., Jain, S., Gadotti, A., and de Montjoye, Y.-A. (2019). OPAL: High performance platform for large-scale privacy-preserving location data analytics. In Proceedings of BigData 2019: 1332-1342.

[15] Pentland, A. (2012). Society's Nervous System: Building Effective Government, Energy, and Public Health Systems. IEEE Computer, 45(1): 31-38.

[16] Pentland, A. (2014) Saving Big Data from itself. Scientific American, 65–68.

[17] Perentis, C., Vescovi, M., Leonardi, C., Moiso, C., Musolesi, M., Pianesi, F., and Lepri, B. (2017). Anonymous or not? Understanding the factors affecting personal mobile data disclosure. ACM Trans. Internet Techn. 17(2): 13:1-13:19.

[18] Pinkas, B., and Lindell, Y. (2009). A proof of security of Yao's for two-party computation. J Cryptol 22, 161-188.
[19] Richards, D., Rebeschini, P., and Rosasco L. (2020). Decentralised learning with random features and distributed gradient descent. In Proceedings of the 37th International Conference on Machine Learning (ICML), 119: 8105-8115.
[20] Rissanen, E. (2013). Extensible Access Control Markup Language (XACML) Version 3.0. OASIS Standard. Available online at: http://docs.oasis-open.org/xacml/3.0/ xacml-3.0-core-spec-os-en.html
[21] Sen, J. (2013). Homomorphic encryption: Theory and applications. In J. Sen (ed.) Theory and practice of cryptography and network security protocols and technologies. INTECH Publishers.
[22] Shamir, A. (1979). How to share a secret. Communications of the ACM, 22 (11): 612–613.
[23] Staiano, J., Oliver, N., Lepri, B., de Oliveira, R., Caraviello, M., and Sebe, N. (2014). Money walks: A human-centric study on the economics of personal mobile data. In Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing, 583-594.
[24] Staiano, J., Zyskind, G., Lepri, B., Oliver, N., and Pentland, A. (2019). The rise of decentralized personal data markets. In T. Hardjono, D. Shrier, and A. Pentland (eds.) Trust :: Data. A new framework for identity and data sharing. MIT Press.
[25] Takemori, S., and Sato, M. (2020). Approximation methods for kernelized bandits, arXiv 2010.12167.
[26] Vescovi, M., Lepri, B., Perentis, C., Moiso, C., and Leonardi, C. (2014). My data store: Toward user awareness and control on personal data. In Proceedings of UbiComp Adjunct: 179-182.
[27] Want, R., Pering, T., Danneels, G., Kumar, M., Sundar, M., and Light, J. (2002). The personal server: Changing the way we think about ubiquitous computing. In Proceedings of 4th International Conference on Ubiquitous Computing, 194–209.
[28] Zyskind, G., Nathan, O., and Pentland, A. (2015). Decentralizing privacy: Using blockchain to protect personal data. In Proceedings of IEEE Symposium on Security and Privacy Workshops: 180-184.