Substra
Foundation

Towards trustworthy data science

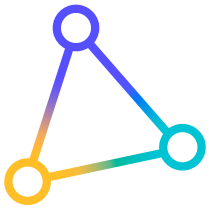
Substra Foundation's **Manifesto**

OBJECTIVE OF THIS DOCUMENT

This document is Substra Foundation's manifesto. It describes its environment, the challenges it tackles, its vision, approach, origin and ambitions, the community of users and contributors it aims to federate and stimulate. The purpose of this manifesto is to provide enough information for anyone to easily get a good sense of what Substra Foundation and the open source Substra Framework are.

Substra Foundation and this manifesto are still young. Remarks, questions and feedback of any kind are more than welcome and will complement and improve future versions: this is an ongoing and collaborative effort.

SOMMAIRE



Substra Foundation

Substra Foundation is an independent research organization dedicated to developing collaborative, responsible and trustworthy data science.

It aims **to foster new scientific and economic collaborations.** It promotes, protects and advances the open source project Substra, a software framework for secure, traceable, distributed ML orchestration.

It aims at **federating a vibrant community of users** and contributors around **Substra Framework.**

The first contributor and a core partner is **Owkin,** a health data AI company that dedicates a complete technical team to the development of the framework.

Towards responsible and trustworthy data science

As of today, **it is difficult to trust** AI

Artificial Intelligence (AI) technologies require **massive amounts of data to achieve high predictive performances.**

The circulation, processing and compilation of datasets by multiple actors **raises concerns about privacy issues and risks of sensitive information leaks.**

Further, many in the tech industry, research sector, and public organizations are growing more preoccupied with AI standards, such as how algorithms **are trained and tested, and how to measure the significance and robustness of AI performance**.

As of today, with clear risks and work-in-progress regulations, it is difficult to trust AI.

The potential of AI **is immense**

Simultaneously, **Machine Learning (ML) and data science keep expanding into research,** business processes, marketing and advertising, products and services in countless industries.

Whether presented as new techniques, specialised tools, or general capabilities, this diffusion of ML fuels a multiplicity of projects aimed at gaining new insights and opens up many possible fields where AI could stimulate innovations.

The potential of AI is immense.

Trustworthy AI by-design **is needed**

The two trends, the growing concerns about privacy, **transparency and quality of AI,** and its expansion into many sectors and organizations, won't disappear in the coming years.

We believe that we need to address both, and in reconciling and combining them we can foster a sound development of trustworthy AI.

New technical and organizational solutions are required for this endeavor, to build up trust, to enable large scale collaborations of citizens, companies and institutions; **ultimately, to create the conditions for responsible, privacy-preserving, and quality data science.**

In short, trustworthy AI by-design is needed, and Substra Foundation is entirely committed to contributing to it.

Substra **Framework**

Privacy researchers and a growing number of interested parties, like ML-driven companies, are developing and testing different privacy-enhancing technologies (PETs). These advance the options for reinforcing the privacy of datasets and models in data science projects, and are becoming increasingly instrumental to develop trustworthy AI projects.

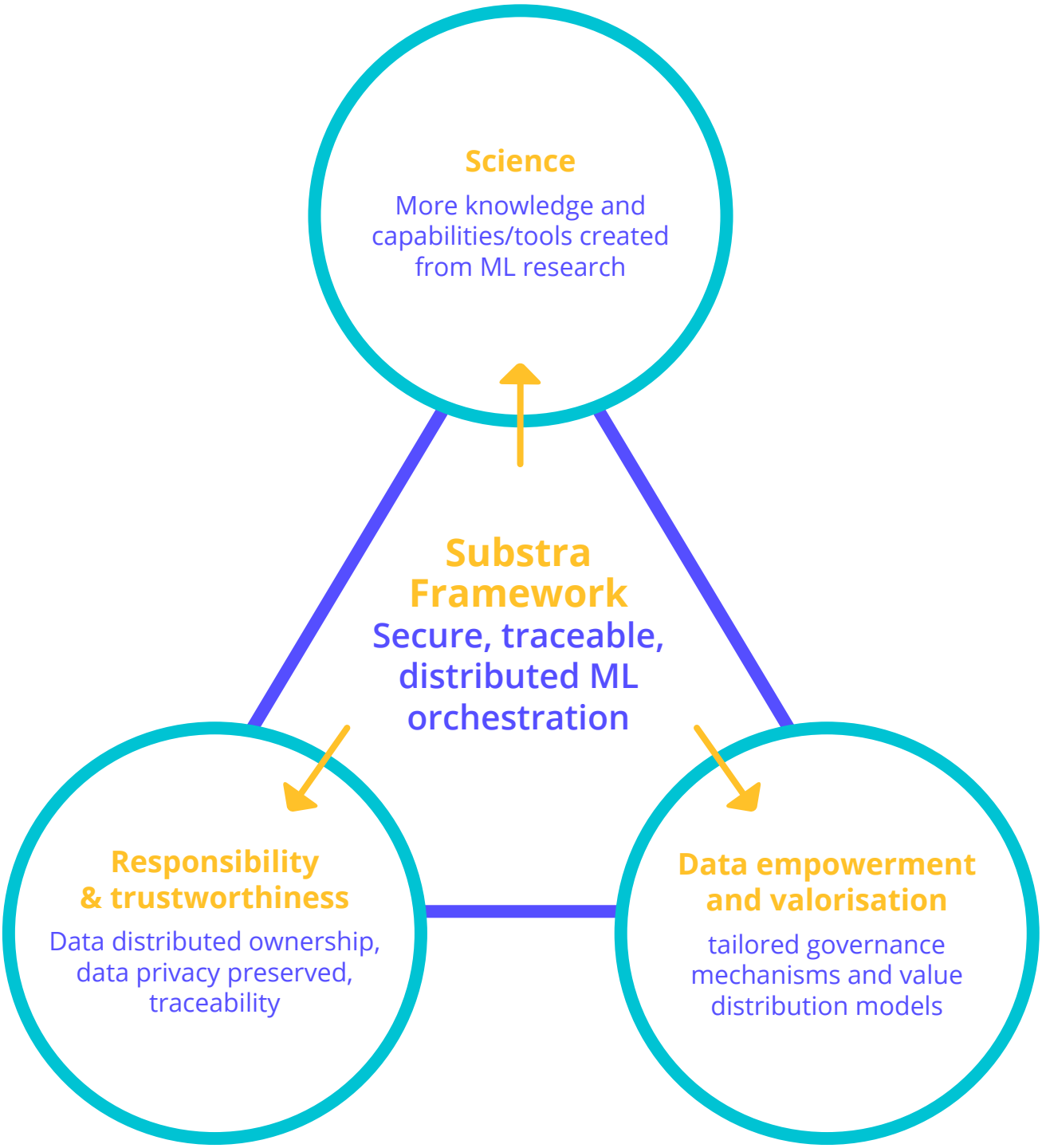
Substra Framework is a low-layer tool, offering secure, traceable, distributed orchestration of machine learning tasks among partners.

It aims at compatibility with PETs to complement their use to provide efficient and transparent privacy-preserving workflows for data science.

Our mission is to make new scientific and economic data science collaborations possible.

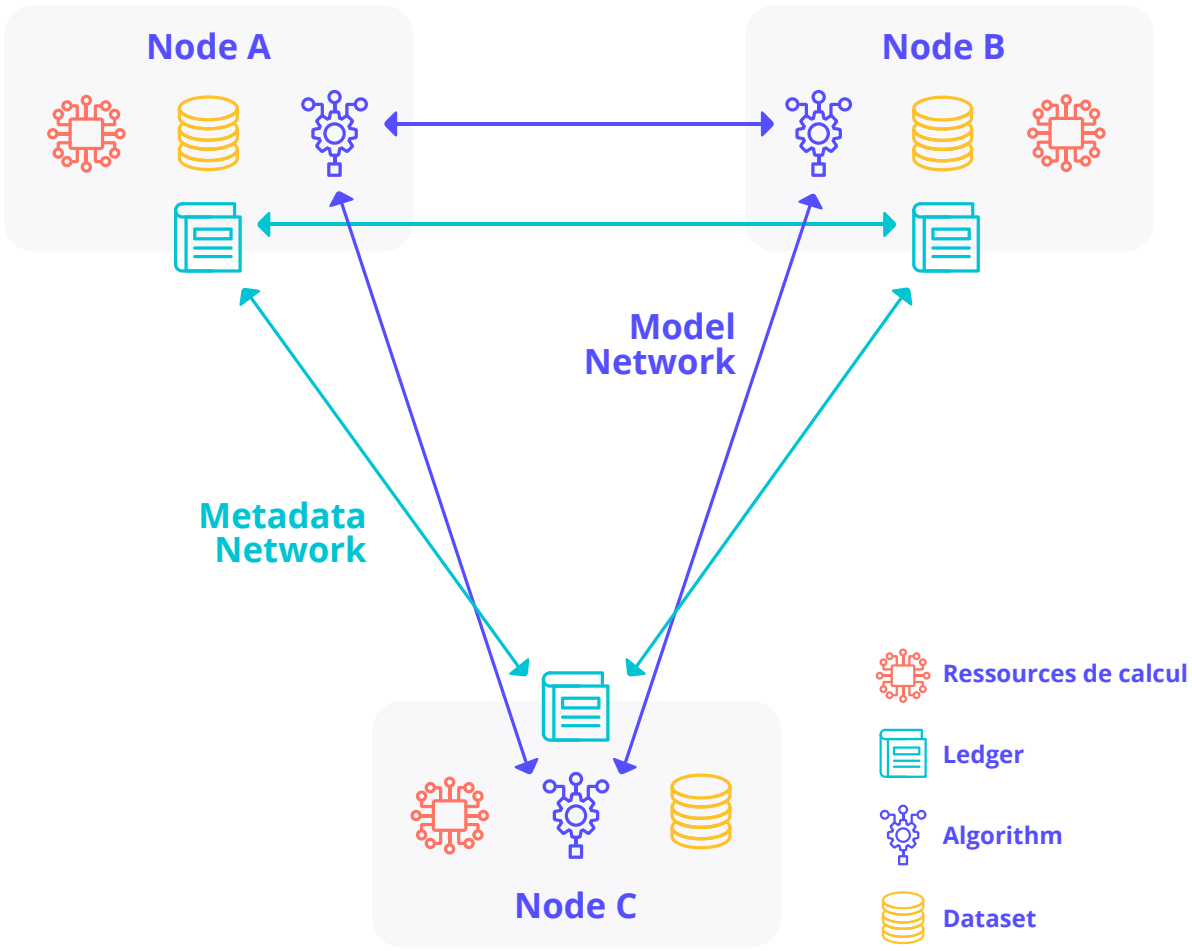
Secure, traceable,
distributed Machine Learning orchestration

We believe in and advocate for a model where **data providers retain full control of their data,** while these data are unlocked for use in concrete and impactful ML projects, in privacy-preserving conditions.



Substra is a software framework building upon the **leading distributed ledger technology, Hyperledger Fabric.** It offers by-design secure, traceable, distributed machine learning orchestration among multiple parties. **Data analysis algorithms travel to distributed training dataset nodes,** and computations are performed on local, disposable, secure computing containers at each node. The trustless nature of the distributed ledger framework **enforces the control of computed analysis and provides an incorruptible traceability of all operations.** It's this feature of control that enables users to deploy the framework **in efforts to preserve data privacy.**

Substra Framework enables multi-party data analysis and machine learning collaborations; it facilitates a large variety of strategies for privacy-preserving data science partnerships.



Substra Foundation's mission is to

facilitate collaborative Machine Learning projects

An independent research nonprofit to host and drive the open source Substra project

Substra Foundation is an independent research nonprofit created in early 2018 by passionate individuals, dedicated to fostering the development of collaborative, responsible and trustworthy data science ecosystems.

It focuses on the following 3 lines of action:

1. **Industrial R&D consortiums**
   Participate in research consortiums dedicated to new data science collaborations on sensitive data.
   Currently: HealthChain, Melloddy

2. **Substra Framework open source initiative**
   Promote, protect and advance the open source Substra project
   Substra is a framework for secure, traceable, distributed ML orchestration

3. **Contribute to the community**
   Participate in the efforts of the trustworthy AI community
   Topics of interest: datasets contributivity to a model, certifying models validation workflows, scoring 'responsible data science' practices

In our core mission to promote, protect and advance the open source project Substra, we aim to:

- **Gather** feedback on Substra usage
- **Synthesize** improvement requests and propose a common development roadmap
- **Host** the Substra Github repository and other collaboration tools
- **Animate** reflections on new collaboration practices
- **Propose** ideas to foster responsible and trustworthy data science

A core objective is to foster the emergence of a vibrant community of users and contributors. We also participate in research consortiums dedicated to new data science collaborations on sensitive data.

Substra Foundation's leading principle is to maintain its independence from other organizations and its agility as a small and innovative organization. Thus, to welcome new members in its governance over time, only individuals (no organizations or representatives of organizations), accomplished contributors, trustworthy data science enthusiasts, who wish to participate in the administrative management of the nonprofit will be accepted.

Core partnership with Owkin to prepare the open source collaborations

Substra Foundation works in close partnership with Owkin, a health data AI company that dedicates a complete technical team to the development of the framework and is the first contributor to it. The public release of the framework marked the start of the open source collaborations, animated by Substra Foundation.

The objectives on our horizon

Substra Foundation is engaged in large collaborative research projects in France and in Europe in the public and private health sector.

Our goal is to expand the organization's reach and develop the adoption of secure, traceable, distributed ML orchestration and of the Substra Framework in multiple industries and geographies. But our current most challenging ambition is **to federate a vibrant community of contributors and users** *(see next section).*

Over the long run, we desire to:

- **Demonstrate** the security of the framework and its compliance with data protection regulations (e.g. GDPR), establish it as a piece of the solution to privacy preservation requirements;
- **Leverage** the framework mechanisms to foster scientific rigour in ML model development and evaluation;
- **Integrate** value repartition bricks/features in the framework;
- **Cultivate** the organization's independence and elevate it as a trustworthy third-party;
- **Develop** a sound revenue model to ensure our autonomy and sustainability.

An active community formed around contributors

Fostering an engaged and impactful community is a challenge essential to the success of the project.

Currently, all of our objectives are a work in progress. Areas for further collaboration and input include, for example: the animation of a bug bounty program, the deployment of a public 'testnet' instance, or a support program for researchers and developers building tech components and services on top of Substra Framework; all of this alongside further development and features improvement of the framework.

We're currently working on how best to facilitate and stimulate the emergence of these communities, to organize and animate their respective work, and to ensure good coordination between all parties. The objective is to create the conditions for an open, fruitful, enthusiastic collaboration.

ANNEX

Definitions and additional information

Hyperledger Fabric: the world-leading private and permissioned blockchain framework. Hyperledger Fabric is one of the Hyperledger open source projects hosted by the Linux Foundation. It has been widely adopted as a reference framework for implementing blockchain-based services in business ecosystems. Substra Framework is built upon Hyperledger Fabric and its core components (distributed ledger, identities and membership mechanisms, smart contracts, consensus mechanisms, etc.). Substra Foundation is an Associate Member of the Hyperledger Project. Link

Distributed ledger: a distributed ledger is a consensus of replicated, shared, and synchronized digital data geographically spread across multiple sites, countries, or institutions. There is no central administrator or centralized data storage. A peer-to-peer network is required as well as consensus algorithms to ensure replication across nodes is undertaken (source: Wikipedia).

Machine learning orchestration: in contexts where multiple parties collaborate for elaborating machine learning models, the different operations (e.g. algorithms transfers, training computations, model evaluations, predictions...) need to be orchestrated in time and space. Such an orchestration is done over a network connecting the parties, and requires complete traceability of all operations, identities certifications, security (among others). Substra Framework enables the implementation of applications or services requiring secure, traceable, distributed machine learning orchestration.

Trustless: Substra Framework is a 'trustless' ML orchestration framework. The word 'trustless' might be ambiguous in certain circumstances. We believe it should be used as 'doesn't require trust a priori between parties': the code implementation of the software enables parties to collaborate without trusting each other, it technically guarantees that actions and transactions will be performed as defined in the rules agreed upon. What is required is to 'trust the code': it might not be straightforward and even require some audit effort, but in many cases it is easier than trusting a number of other independent organisations.

Privacy-preserving: Substra Framework is a tool in the quest for 'privacy-preserving' ML (with the word 'privacy' referring to both the privacy of the dataset for the organisation managing it, or the privacy of personal data for the individuals these data refer to). It enables data analysis and machine learning computations on data without transferring the data to anyone and without giving data scientists read access to these data. It has to be combined with privacy enhancement approaches in ML algorithms (contractual requirements, algorithms audits...) and data pre-processing (differential privacy, anonymization of PII...).

hello@substra.org
www.substra.org

Substra
Foundation

Startup Palace
4 rue Voltaire
44000 Nantes

Start-up studio Iconosquare
12 rue Armand Barbès
87000 Limoges

La Paillasse
226 rue Saint-Denis
75002 Paris