Federált tanulás és lehetőségei: COMPLIANCE BY DESIGN IN THE EU CLOUD

Tényi Ákos
E-Group ICT Software Group,
Head of Smart Data Business Line
akos.tenyi@egroup.hu

2023. November 08.
1st NATIONAL FEDERATED BIOBANK IN EU

The **Semmelweis Biobank Network** keeps the data of approx. 100,000 individuals. This data asset is managed by diverse, fragmented, non-interoperable data collection systems.

We create **interoperability** between fragmented databases at the institutional level and to make it possible to jointly analyze them in a way that guarantees the security of the individual’s personal data. We create **federated data assets at the institutional level** that preserves data protection guarantees, which supports interactive cooperation between cross-border health data repositories, the identification of new knowledge, a **better understanding of the pathomechanism of disease**, with the support of clinical research and decision-making. **Data sharing that ensures privacy will ensure optimal interoperability between data collection centers.**

### HIGHLIGHTS

- **12 independent biobank** in the network
- FAIR data and data harmonization activities
- European data standards
- Secure and privacy assuring, distributed execution environment
- Analytical and virtual data exchange system
- Certificate enablement
- Testing and maintenance support
## Why Data is Essential?

<table>
<thead>
<tr>
<th><strong>AI is top priority</strong> strategy across the world</th>
<th><strong>There is no modern and productive industry without data intensive operations</strong></th>
<th><strong>EU-US AI agreement:</strong> The U.S. data stays in the U.S. and European data stays there, together better models</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>AI needs data</strong></td>
<td><strong>Data is distributed</strong> across countries.</td>
<td><strong>Who owns</strong> the data + the data extraction technology (AI) <strong>has the control</strong> as well</td>
</tr>
<tr>
<td><strong>GDPR, EU AI Act:</strong> Training highly restricted, localised on EU data</td>
<td><strong>Emphasis on representative training, testing data sets of AI models</strong></td>
<td></td>
</tr>
</tbody>
</table>
WHY TECHNOLOGICAL PROJECT IS ESSENTIAL?

2017 → 2021
+70%
new data regulations

“Meta hit with record-breaking $1.3 billion fine over Facebook data transfers to the US”

WHAT TYPES OF DATA ARE BLOCKED?

<table>
<thead>
<tr>
<th>Category</th>
<th>Numbers of Countries Blocking</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accounting, Tax, and Financial</td>
<td>18</td>
</tr>
<tr>
<td>Personal</td>
<td>13</td>
</tr>
<tr>
<td>Government and Public</td>
<td>10</td>
</tr>
<tr>
<td>Emerging Digital Services</td>
<td>9</td>
</tr>
<tr>
<td>Other</td>
<td>5</td>
</tr>
<tr>
<td>Telecommunications</td>
<td>4</td>
</tr>
</tbody>
</table>

*ITIF analysis of formal laws or regulations publicly reported as of April 2017.

Learn more at itif.org/databarriers
FROM DATA TO INSIGHT = KNOWLEDGE SPECTRUM

Giving out data, giving out the whole knowledge spectrum

Not "enough" data > knowledge inferior > Is "knowledge licencing" in narrow sense possible?

Monitoring and controlling of Insight-generation
WHAT TO DO?
FEDERATED LEARNING = DIST. KNOWLEDGE GENERATION

BRINGS ANALYTICS TO THE LOCAL DATA AND FEDERATRES EXTRACTED KNOWLEDGE

- Brings the algorithm to the data
- No need to aggregate data
- Machine Learning models aggregated
- Private data kept at the place of origin
- Privacy preserving property of learning process ENSURED
- Secure and fast access to data

MACHINE LEARNING TECHNIQUES

A LOCAL LEARNING

Local data, local learning

B CENTRAL LEARNING

Centralized data, centralized learning

C FEDERATED LEARNING

Train machine learning algorithms across multiple decentralized servers (datasets) without sharing their data

Local data
local learning + global learning
(Trusted ?) AI

(Trusted ?) data
(Trusted ?) data
Trustworthy data?

Data Poisoning

- Local data collection
- Local model training process

Model Poisoning

- Local data collection
- Local model training process

Poisoned model parameter

\[ w' = w + \text{aggregate}(\Delta w_1 + \cdots + \Delta w_n) \]

Central server

\[ \Delta w_1 \]

\[ \Delta w_n \]

(L. Lyu et al. 2020)
Need for trustworthy data and models

Data Poisoning

Local data collection

Model Poisoning

Local model training process

Poisoned model parameter

\[ w' = w + \text{aggregate}(\Delta w_1 + \cdots + \Delta w_n) \]

\[ \Delta w_1 \]

\[ \Delta w_n \]

Central server

Trustworthy Data&AI layer
Powered by Blockchain
(Trusted ?) AI
AI ACT – Risk based classification

AI System (Use case) → Risk-level?

- Unacceptable risk: Use case prohibited
- High risk: Use case may be permitted but must comply with strict rules concerning risk management, data quality, and technical documentation
- Low- or no risk: No action required
AI ACT – CONFORMITY NEEDS ADVANCED DATA STRATEGY

CONFORMITY ASSESSMENT

- Data & data governance
  - training, validation of model based on
    - relevant, representative data
    - appropriate statistical properties as regards to:
      - person or group
      - geographical, behavioural or functional setting
  - bias monitoring, potentially based on personal data > privacy preserving measures

- Model
  - appropriate level of accuracy of model (in light of intended purpose)
  - robustness, fail-safe
  - resilient to errors, unauthorised externals
  - cybersecurity, content adversial attacks

- Intended purpose
  - N x n (sectors x use cases) playbooks + situation specific
  - validation data needed

DATA STRATEGY IMPLEMENTATION REQUIRES ADVANCED DATA & AI LOGISTICS TECHNOLOGY BACKING
AI ACT - DATA REGULATION COMPLEXITY

- General Data Protection Regulation (GDPR)
- Law Enforcement Directive (LED)
- Union law on non-discrimination
- Union competition law
- New Legislative Framework (NLF) (e.g. medical devices, machinery, toys)
- Financial services legislation (e.g. PSD2)
- E-Commerce Directive, Digital Services Act (DSA)
- closely linked to the Data Governance Act, the Open Data Directive and other initiatives under the EU strategy for data re-use, sharing and pooling of data that are essential for the development of data-driven AI models of high quality

DATA REGULATION COMPLIANCE REQUIRES COMPLEMENT/CONSISTENT WITH OTHER REGULATIONS CONTROLLED DATA PROCESSING ENVIRONMENTS
AI ACT - DATA REGULATION COMPLEXITY

**COMPLEMENT/CONSISTENT WITH OTHER REGULATIONS**

- General Data Protection Regulation (GDPR)
- Law Enforcement Directive
- Union law on non-discrimination
- Union competition law
- New Legislative Framework (NLF) (e.g. medical devices, machinery, toys)
- Financial services legislation
- E-Commerce Directive, Digital Services Act (DSA)
- closely linked to the Data Governance Act, the Open Data Directive and other initiatives under the EU strategy for data re-use, sharing and pooling of data that are essential for the development of data-driven AI models of high quality

**DATA REGULATION COMPLIANCE REQUIRES**

**CONTROLLED DATA PROCESSING ENVIRONMENTS**
AI Regulatory Sandbox

Regulatory Sandbox shall provide a controlled data processing environment that facilitates the development, testing and validation of innovative AI systems for a limited time before their placement on the market.

- contains personal data (lawfully collected for other purposes)
- aim to aid development and testing

BLUEPRINT OF A REGULATORY SANDBOX

- **Req 1.** contains highly relevant, private data that needed for innovative AI development
- **Req 2.** monitoring of risks during development
- **Req 3.** functionally separate, isolated, protected data processing environment
- **Req 4.** only authorised person has access
- **Req 5.** no personal data to be transmitted, transferred or otherwise accessed by other participants
- **Req 6.** processing fully preserves data subject privacy
- **Req 7.** access to data can be fully revoked (time period, other reasons)

REGULATORY SANDBOX “BY DESIGN” REQUIRE PRIVACY-FIRST DATA PROCESSING TECH ENVIRONMENT
CREATE FEDERATED AI TECHNOLOGY & SERVICE ON DISTRIBUTED DATA NATIONAL, PAN-EU LEVEL

COMPETENT NATIONAL AUTHORITIES: for conformity assessment -> Regulatory Sandbox!
AI COMPANIES (SME & BIGindustry too): Development, testing, validation -> AI compliance!

AI COMPLIANCE SANDBOX

- Trusted technology platform for trustworthy data and AI development, validation EU level
- Data stays, AI travels
- Compliance with data protection laws and standardized, streamlined automatic processes
  - Privacy by Tech Design
  - Zero-risk “data processing” (data stays)
  - Standardized and streamlined data processing security assessment
  - Data security and personal data rights guarantees embedded
- Ability to support National, Pan EU AI ecosystem
- Reduces complexity and enables new markets to strive
- Supports data enabled industry B2B AI/MI collaboration without risking data sharing issues
FEDX – FEDERATED AI LEARNING PLATFORM

PRIVACY FIRST, CONTROLLED, FEDERATED AI SANDBOX FOR CONFORM AI TRAINING, TESTING AND VALIDATION.

HIGHLIGHTS

- **Federated Client Database**: functionally separate, isolated, protected data processing environment (Req 1., Req 3. see earlier slide AI Act requirement)

- **Trusted Federated Machine Learning Engine**: no personal data transmitted, transferred during model development or validation (Req 5.)

- **Privacy-preserving AI**: processing fully preserves data subject and data provider privacy (Req 6.)

- **Training & Resource Monitoring, Logging**: monitoring of risks during development (Req 2.)

- **User Management**: only authorized person has access, temporary access to data processing can be given (Req 4., Req 7.)

- **FedX SF®**: federated technology without adaptation barrier: zero loss of model accuracy, even on non-IID data, highly energy efficient training
## Tech edge

<table>
<thead>
<tr>
<th>Technology</th>
<th>FedEx SF® Federated Learning</th>
<th>Federated Learning</th>
<th>Centralized Learning</th>
<th>Homomorphic Encryption</th>
<th>Secure Multi-Party Computation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Product using technology</td>
<td>FedEx</td>
<td>Owkin Connect</td>
<td>Amazon SageMaker</td>
<td>ZeroReveal (Enveil)</td>
<td>Sharemind</td>
</tr>
<tr>
<td>Data Privacy</td>
<td>High</td>
<td>High</td>
<td>Low</td>
<td>High</td>
<td>High</td>
</tr>
<tr>
<td>Computational Efficiency</td>
<td>High</td>
<td>Low</td>
<td>High</td>
<td>Low</td>
<td>Low</td>
</tr>
<tr>
<td>Communication Efficiency</td>
<td>High</td>
<td>Low</td>
<td>High</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>Performance on Real-World Data</td>
<td>High</td>
<td>Medium</td>
<td>High</td>
<td>High</td>
<td>High</td>
</tr>
<tr>
<td>Machine Learning/AI Capabilities</td>
<td>High</td>
<td>High</td>
<td>High</td>
<td>Medium</td>
<td>Medium</td>
</tr>
<tr>
<td>Data Analytics Capabilities</td>
<td>High</td>
<td>Low</td>
<td>High</td>
<td>Low</td>
<td>Medium</td>
</tr>
</tbody>
</table>
Our TECH EDGE

FEDERATED VS. CENTRAL TRAINING ACCURACY

<table>
<thead>
<tr>
<th>Accuracy Loss (compared to central)</th>
<th>Benchmark Data</th>
<th>Real World Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>State of the Art Federated Learning</td>
<td>3%</td>
<td>9%</td>
</tr>
<tr>
<td>FedX SF technology®</td>
<td>0%</td>
<td>0%</td>
</tr>
</tbody>
</table>

0% LOSS ON ACCURACY

-→ eliminates technology adaptation barriers
-→ QoS comparable to non-federated approach
-→ Works seamlessly with non-IID (independent and identically distribute, ~ Real world data)

FEDERATED COMPUTATION COST

<table>
<thead>
<tr>
<th>TOTAL COMPUTATIONAL FACTOR</th>
<th>Benchmark Data</th>
<th>Real World Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>State of the Art Federated Learning</td>
<td>18</td>
<td>54</td>
</tr>
<tr>
<td>FedX SF technology®</td>
<td>4</td>
<td>5</td>
</tr>
</tbody>
</table>

90% ENERGY SAVINGS vs Federated learning solutions

-→ Low cost OPEX
-→ Competitive pricing
-→ Green Incentives

©FedX SF techNOLOGY is submitted to PCT/EPO patenting process
EUROPEAN FEDERATED CLOUD DATA INFRASTRUCTURE

THE FEDEU.AI PROJECT CANDIDATE

- Privacy preserving technology platform for ML training
- Scalable, cloud-native federated learning system
- Advanced Federated Learning features, data catalogue and libraries
- Advanced security & privacy protocols
- Reducing cost and energy needs of federated training
- Technology platform for trustworthy data and models across
- Reduces complexity and enables new market

IPCEI-CIS

IMPORTANT PROJECTS OF COMMON EUROPEAN INTEREST

IPCEI ON NEXT GENERATION CLOUD INFRASTRUCTURE AND SERVICES

12 EU Member States join forces to create a common cloud and edge infrastructure and its associated smart services for the future
DOME EU Project (E-Group FedX POC)
A DISTRIBUTED OPEN MARKETPLACE FOR EUROPE CLOUD AND EDGE SERVICES (2023-2025, Project)
E-Group is a science and idea-driven software technology and digital knowledge manufacturing corporation taking great pride in its deep academic roots.

We strive to build a better future leveraging knowledge and wisdom derived from data.

Our successful local and international projects in the FinTech, GovTech, HealthTech and EnergyTech segments prove that with our dedication and expertise, we create outstanding value for our clients and partners.
ALMOST 30 YEARS OF COMPETENCY GROWTH

#TRUST

#XCHANGE

#DATA

Data Sharing & Knowledge Generation Services, Federated Learning
SmartData (Datalakes & Advanced Analytics, applied AI)
Digital Identity and attributes, Trust Management
Digital Banking and Payment processing
Security & Cryptography PKI, DigitalSignature, applications

E-Group ICT Software was founded in 1993


ALLIANCES AND CERTIFICATES

Under NSPA qualification process
We appreciate our cooperation with all of our academic, technology and innovation partners.

# Key Industry Partners

- Thales
- F-Secure
- Roche
- Bittium
- Atos
- ZSL
- CA
- OTP Bank
- Liberbank
- UnionPay

# EU Network: Data AI and Innovation Ecosystem Partners

- BBMRI: Biobanking and Biomolecular Resources Research Infrastructure
- IPCEI: CIS & Health
- QEHDEN: European Health Data & Evidence Network (QEHDEN)

# R&D Academic Partners

- EIT: European Institute of Innovation & Technology
- University of Veterinary Medicine Budapest
- Semmelweis University
- ELTE Eötvös Loránd Tudományegyetem
- Technische Universität Berlin
- Stockholm University
- École Polytechnique Fédérale de Lausanne (EPFL)

Partnerships Are Essential To E-Group's Success