Privacy-Preserving Federated Analytics using Multiparty Homomorphic Encryption

David Froelicher

PhD Public Defense, 01.10.2021

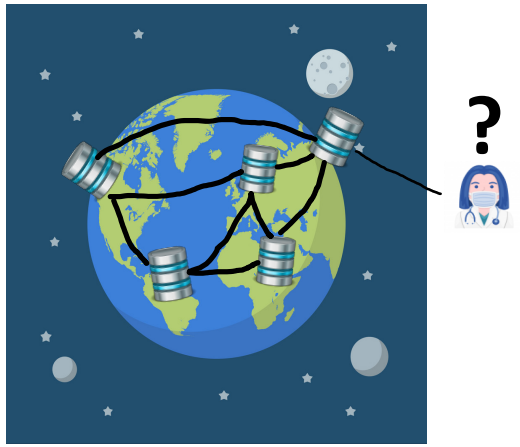
lds.epfl.ch

Advisors: Prof. Jean-Pierre Hubaux

Prof. Bryan Ford

L'analyse fédérée de données en utilisant l'encryption homomorphique distribuée.

Motivation for Federated Analytics



?

- More than **1 billion people worldwide** are fully vaccinated against COVID-19
- Severe (life threatening) reactions are **extremely rare and dispersed around the globe**
- Studying these cases requires the **international sharing** of dispersed sensitive patients' data

De nombreuses études requièrent le partage international de données sensibles, par exemple celles concernant le vaccin contre la COVID-19.

Motivation for Federated Analytics

However, sensitive/personal data are difficult to share because of:

- **Stringent regulations**, e.g., GDPR.
- Complex/costly **data-access agreements**
- High repercussions in case of **data leakage**
- **Competition** among stakeholders

→ **Sensitive data are often siloed**

MIT Technology Review

Why is it so hard to review the Johnson & Johnson vaccine? Data.

The clock is ticking for regulators looking into covid vaccine side effects. But their task is made harder by America's fragmented data systems.

Partager des données sensibles est difficile pour une multitude de raisons.

Motivation for **Privacy-Preserving** Federated Analytics

By ensuring data privacy, one can **enable data analytics among multiple entities** and :

- **Comply** by-design with regulations, e.g., GDPR.
- **Reduce** the need for data-access agreements
- **Control** which information is revealed and **avoid** data leakages

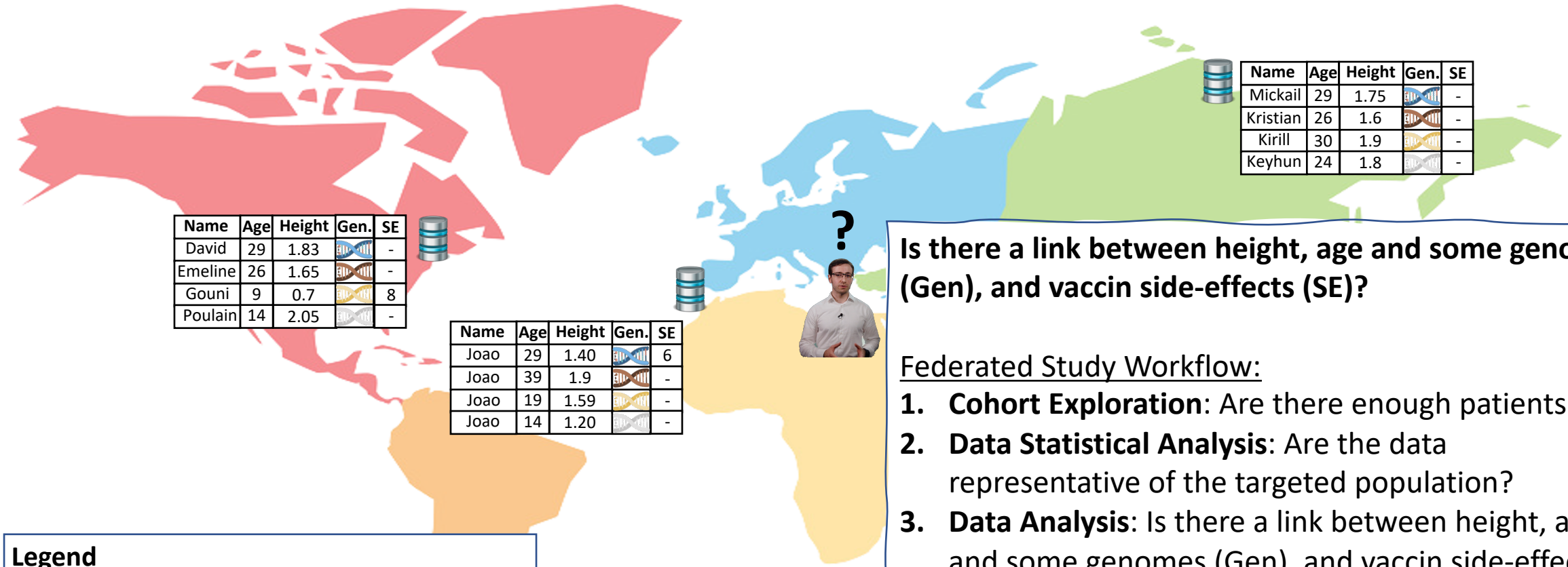
→ **Example of a federated study**

→ **Brief overview of existing solutions and their flaws**

→ **Our solutions**

L'analyse distribuée des données est une solution à ce problème.

Federated Study Workflow



Name	Age	Height	Gen.	SE
David	29	1.83		-
Emeline	26	1.65		-
Gouni	9	0.7		8
Poulain	14	2.05		-

Name	Age	Height	Gen.	SE
Joao	29	1.40		6
Joao	39	1.9		-
Joao	19	1.59		-
Joao	14	1.20		-

Name	Age	Height	Gen.	SE
Mickail	29	1.75		-
Kristian	26	1.6		-
Kirill	30	1.9		-
Keyhun	24	1.8		-

Is there a link between height, age and some genomes (Gen), and vaccin side-effects (SE)?

Federated Study Workflow:

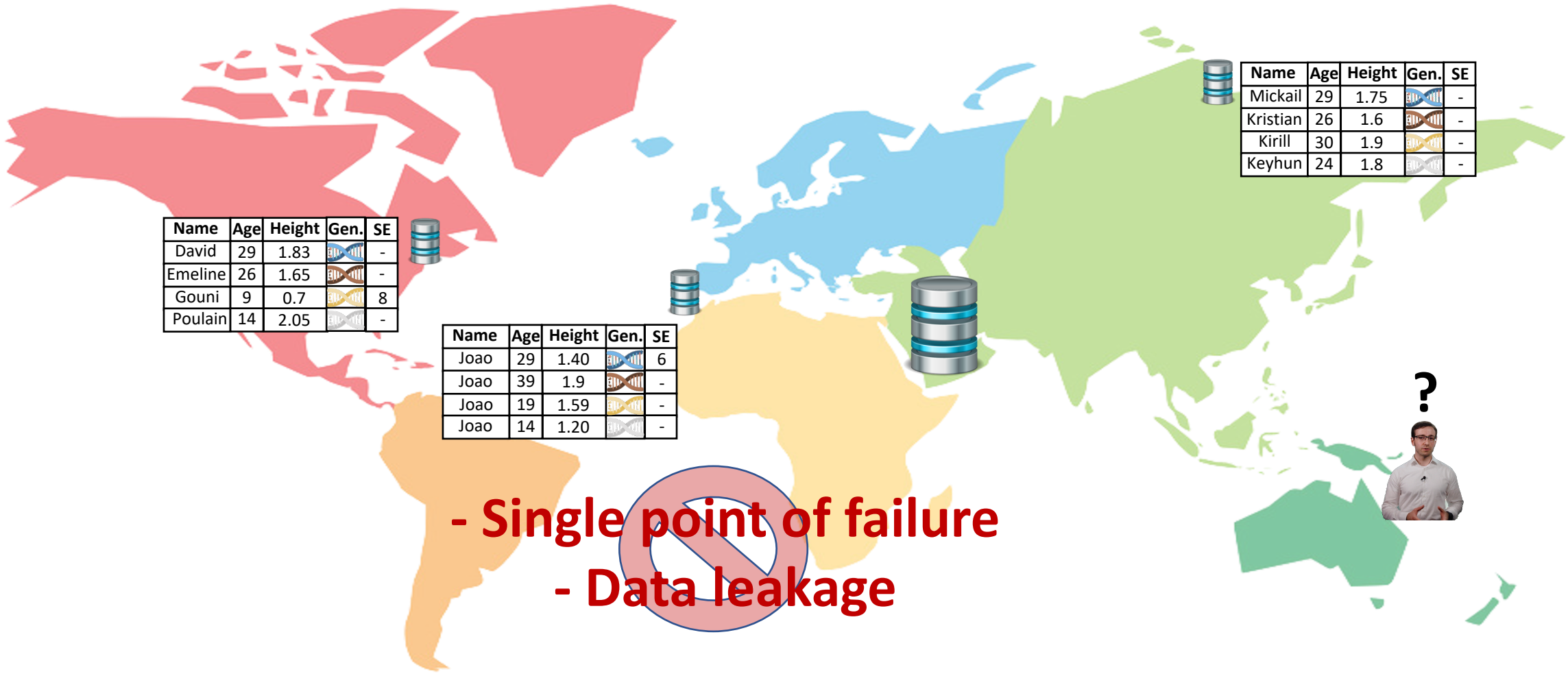
1. **Cohort Exploration**: Are there enough patients?
2. **Data Statistical Analysis**: Are the data representative of the targeted population?
3. **Data Analysis**: Is there a link between height, age and some genomes (Gen), and vaccin side-effects (SE)?

Legend

SE = Side-Effect: intensity between 1-10
Gen. = variants at specific positions in the genome

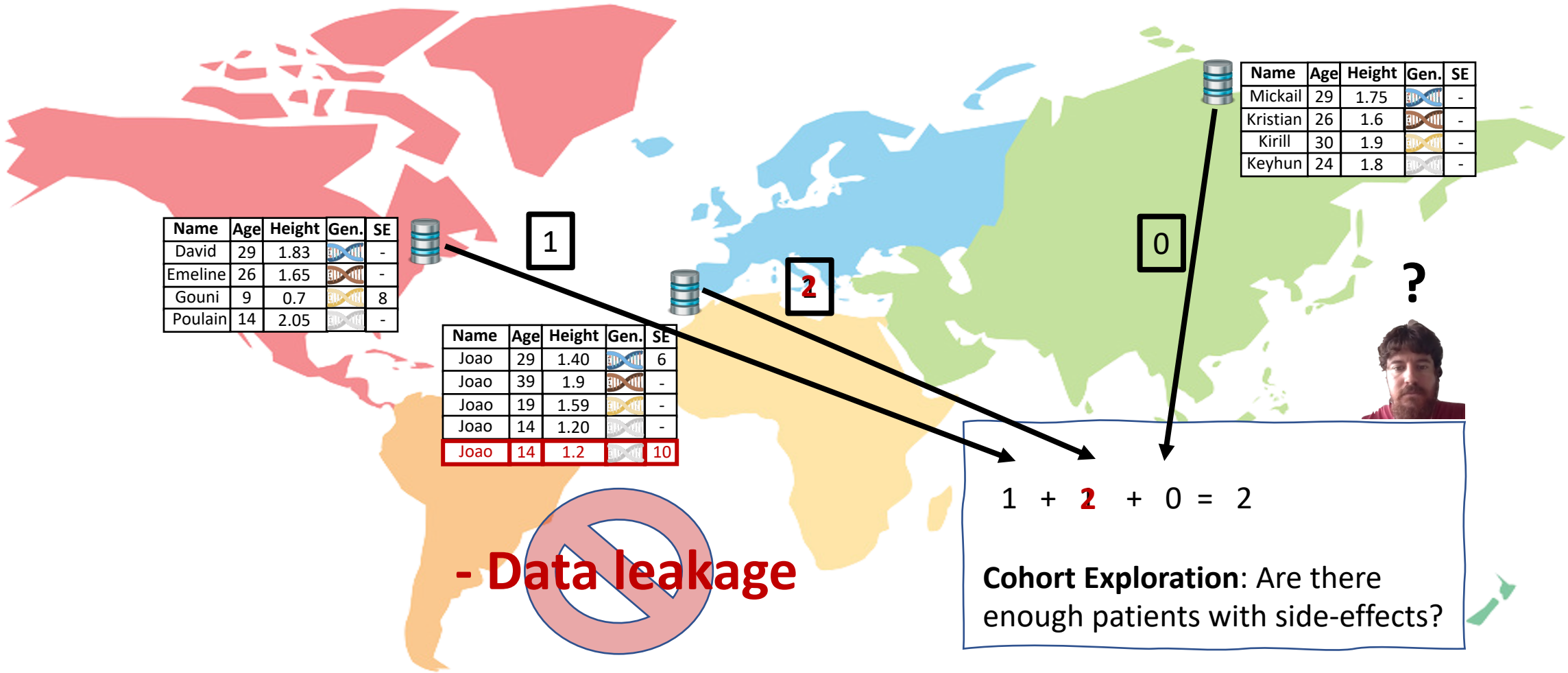
Le flux de travail d'une étude fédérée: exploration, analyse statistique et analyse finale.

Fully Centralized Approach



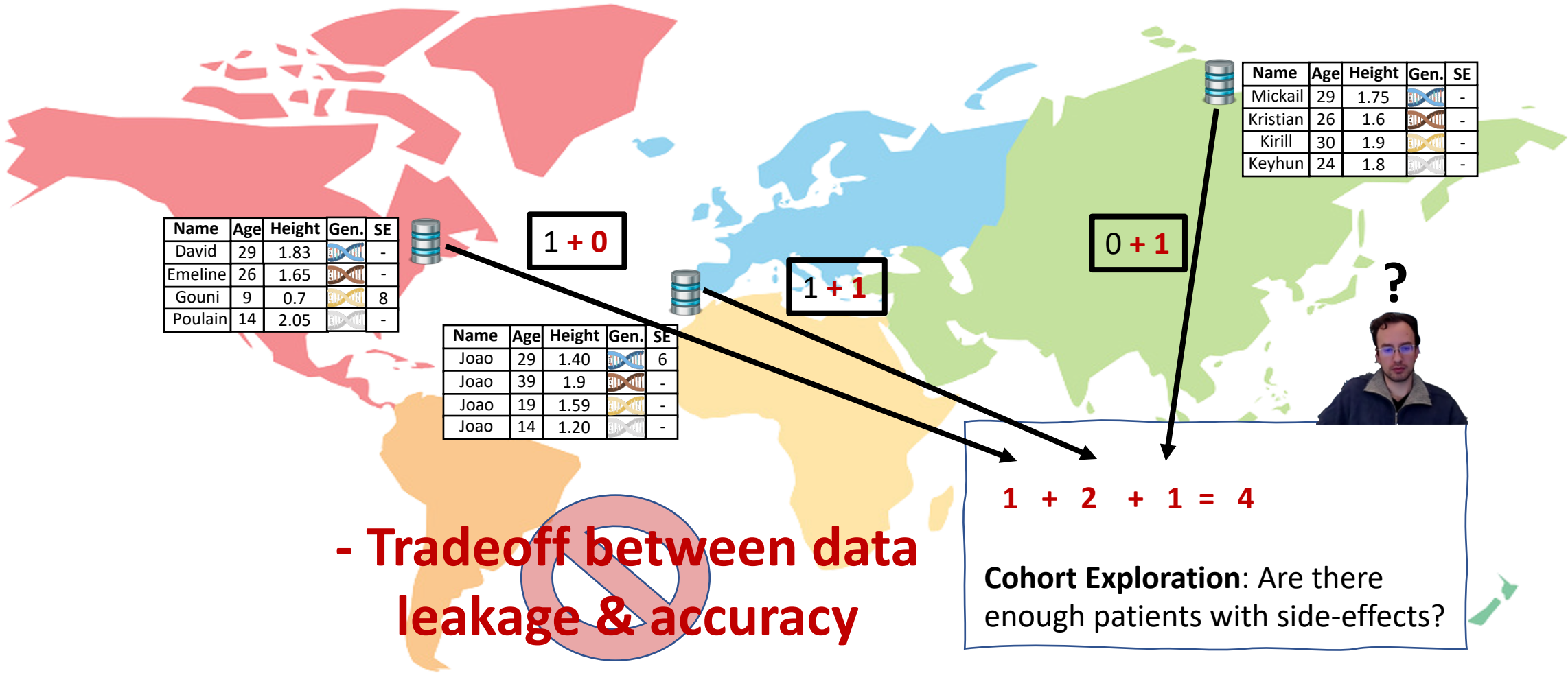
Approche centralisée: point de défaillance unique et fuite de données

Meta-analysis



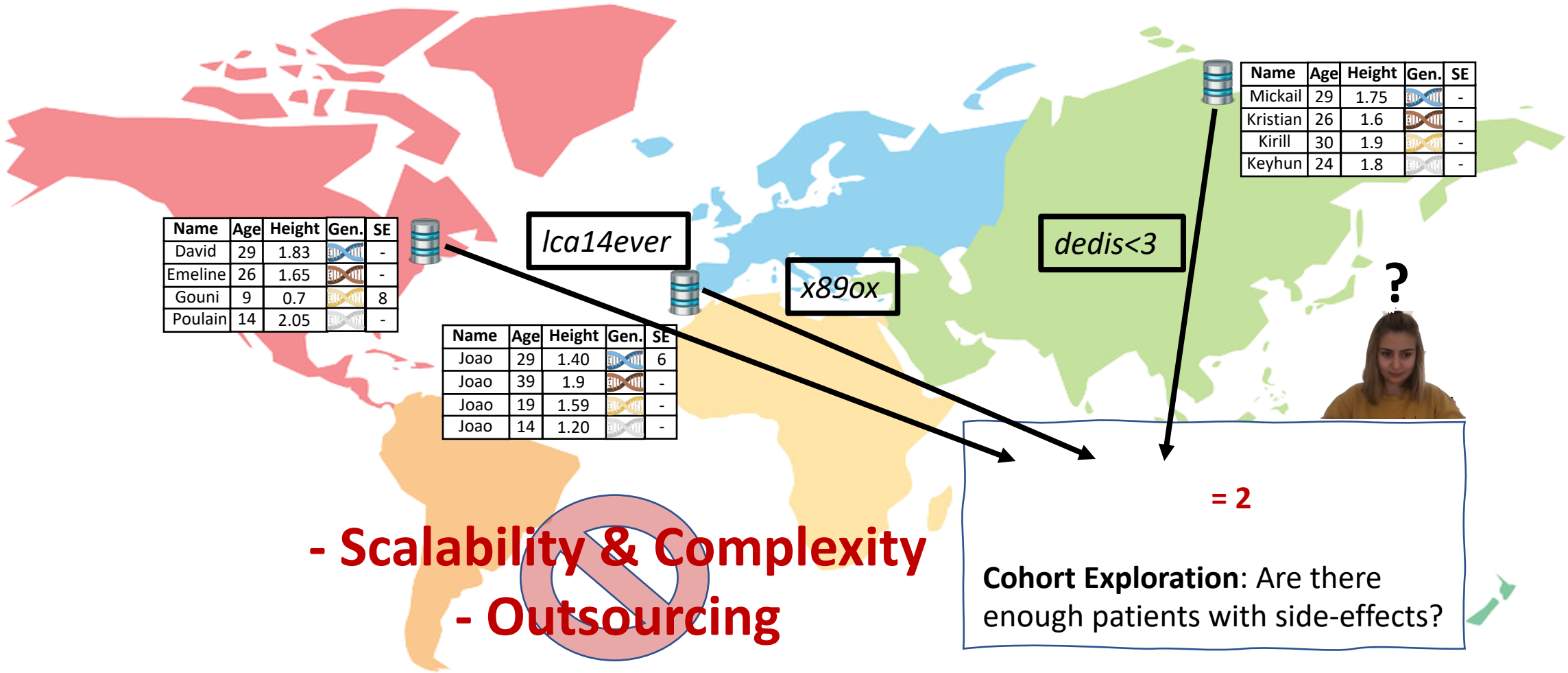
Méta-analyse: point de défaillance unique et fuite de données

Obfuscation-based Techniques



Techniques basées sur l'obscurcissement: compromis entre fuite de données et précision

Crypto-based Techniques

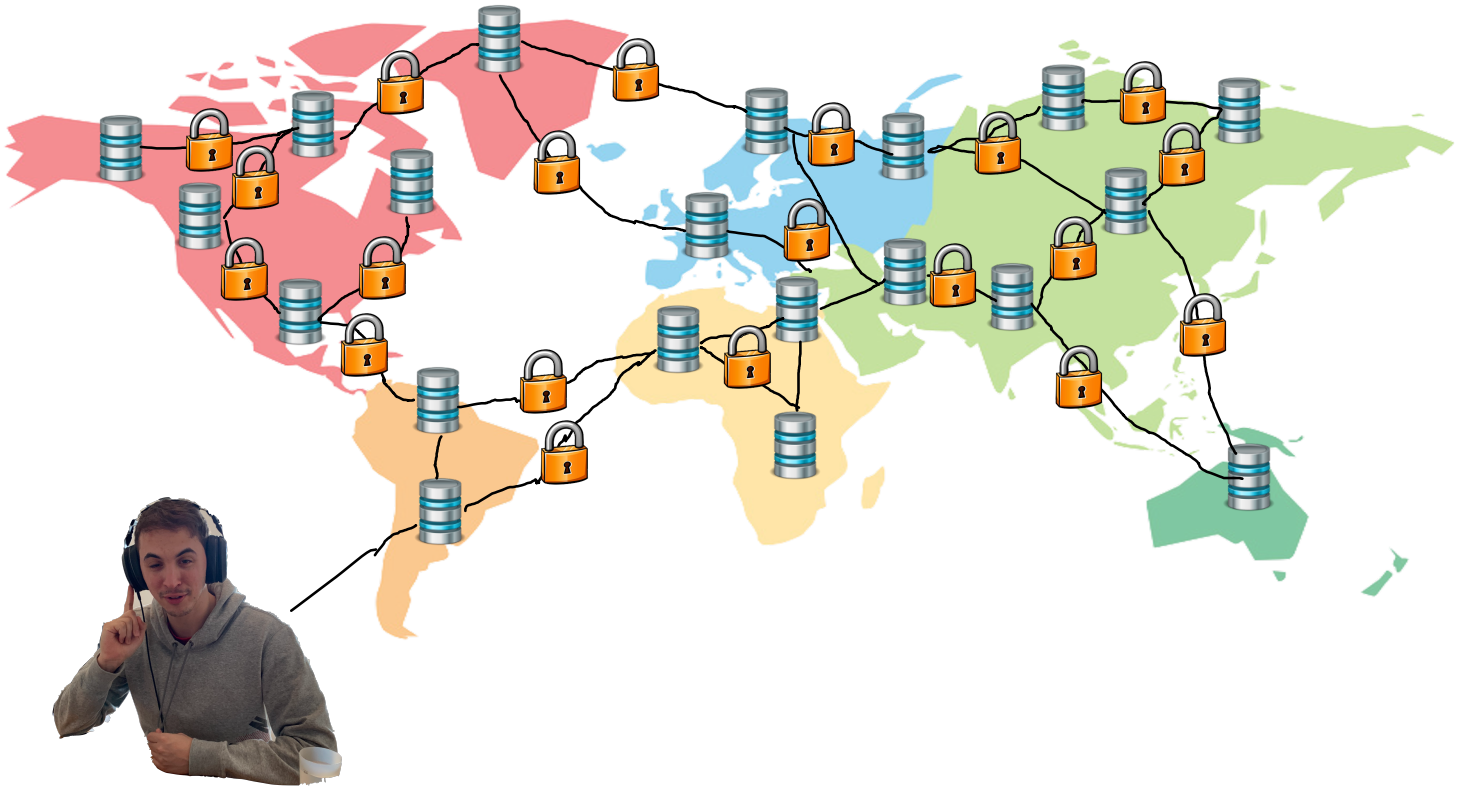


Techniques basées sur la cryptographie: évolutivité et complexité

Thesis Goal

Modular federated privacy-preserving analytics with:

- Scalability
- No data outsourcing
- No data leakage
- No single point of failure
- Computation versatility
- Accuracy



But de la thèse: faire mieux que les solutions précédentes

Thesis Structure

Computation Complexity

Federated Data Exploration

D. Froelicher, P. Egger, J. S. Sousa, J. L. Raisaro, Z. Huang, C. Mouchet, B. Ford, and J.-P. Hubaux: "**UnLynx: A Decentralized System for Privacy-Conscious Data Sharing.**" PETS'17.

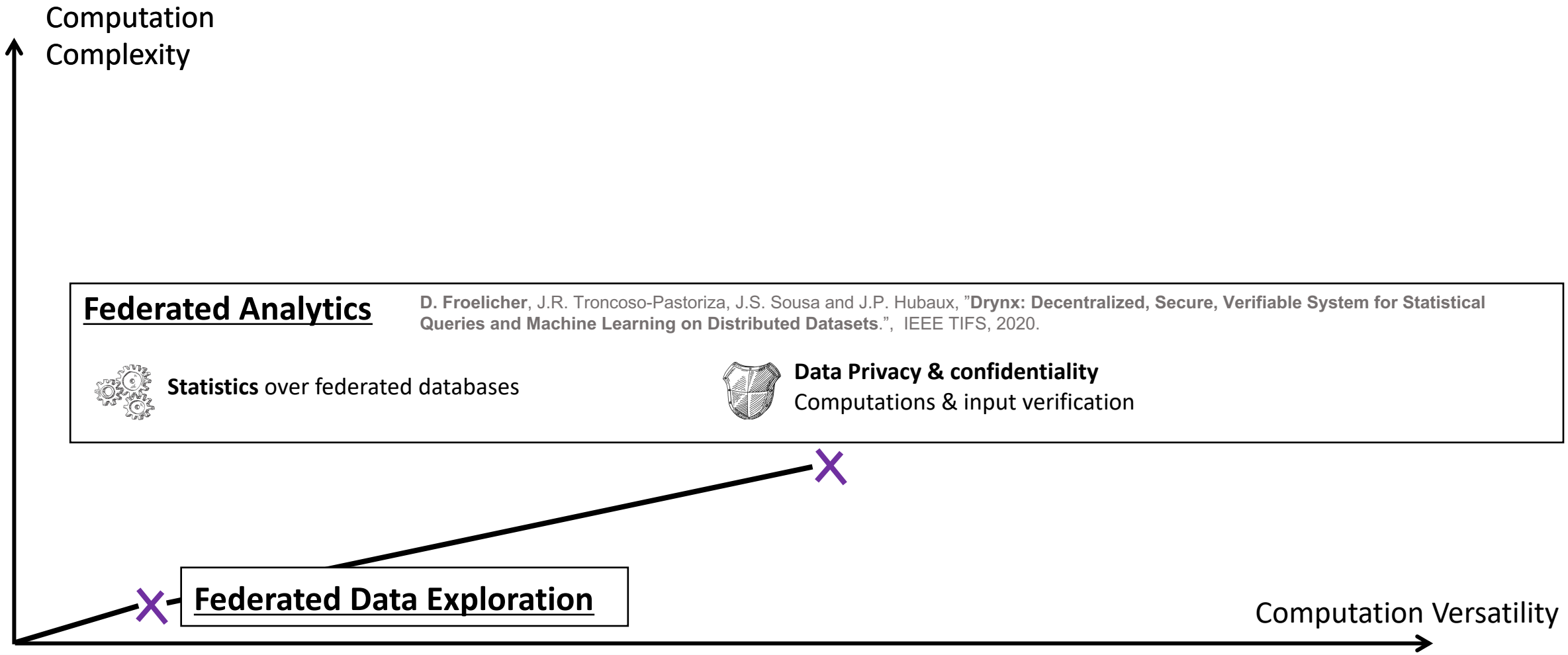
Sum, count and selection over federated databases

Data privacy & confidentiality
Computations verification; Collective protection of local data

Computation Versatility

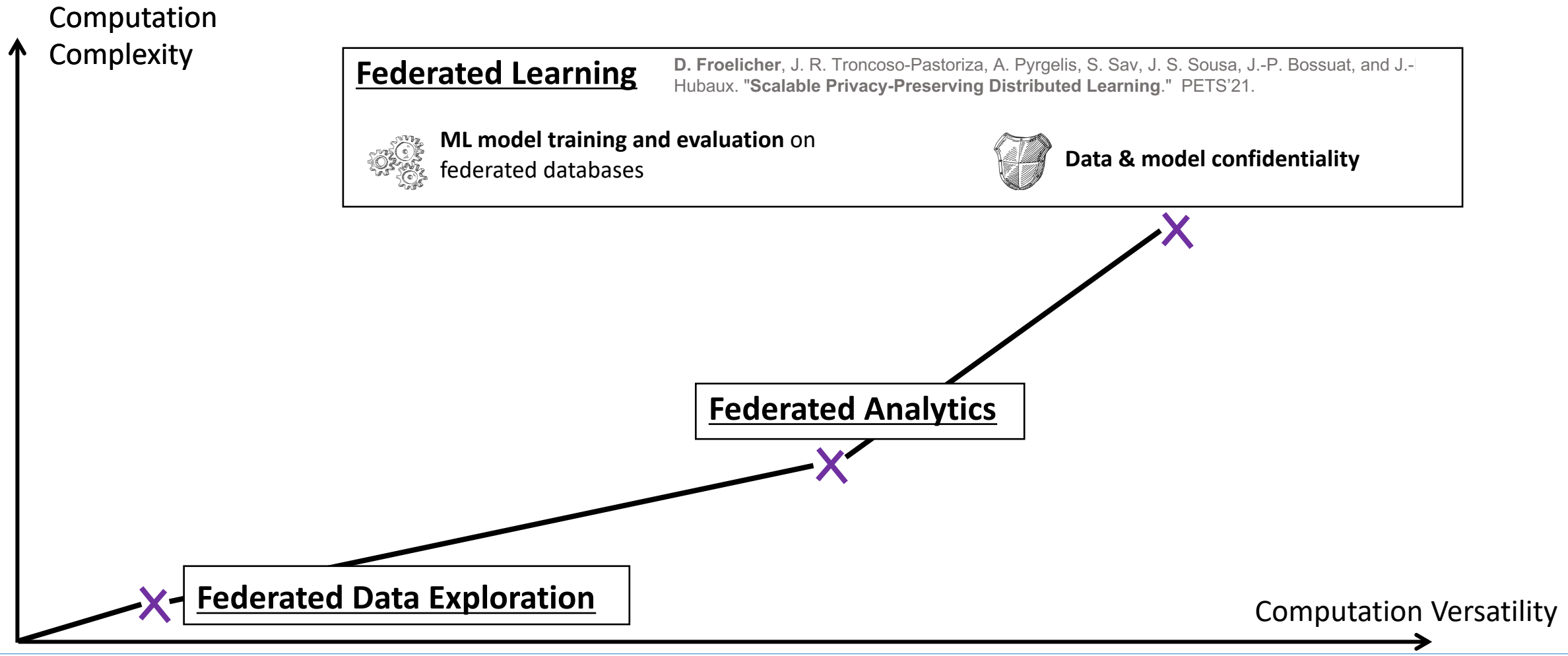
Structure de la thèse: d'abord le comptage sécurisé et fédéré

Thesis Structure



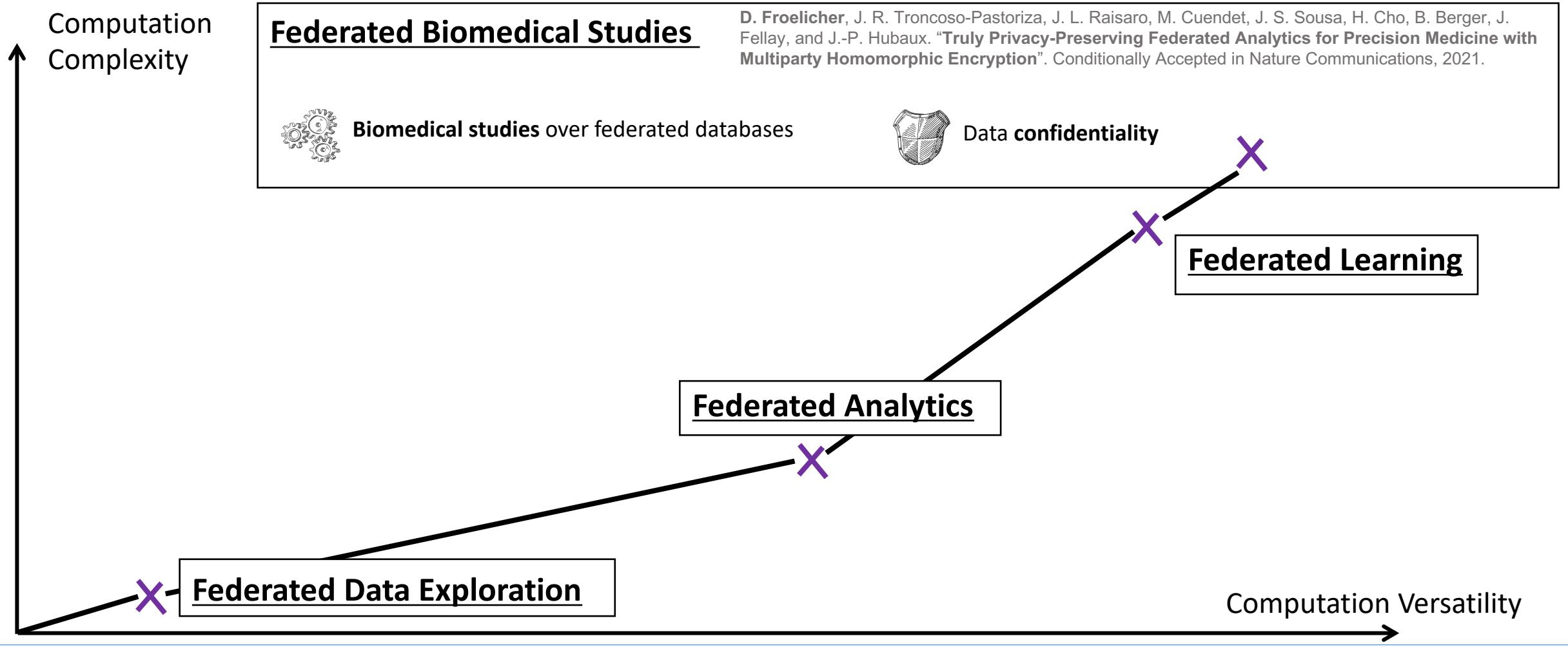
Structure de la thèse: puis des statistiques sécurisées et fédérées

Thesis Structure



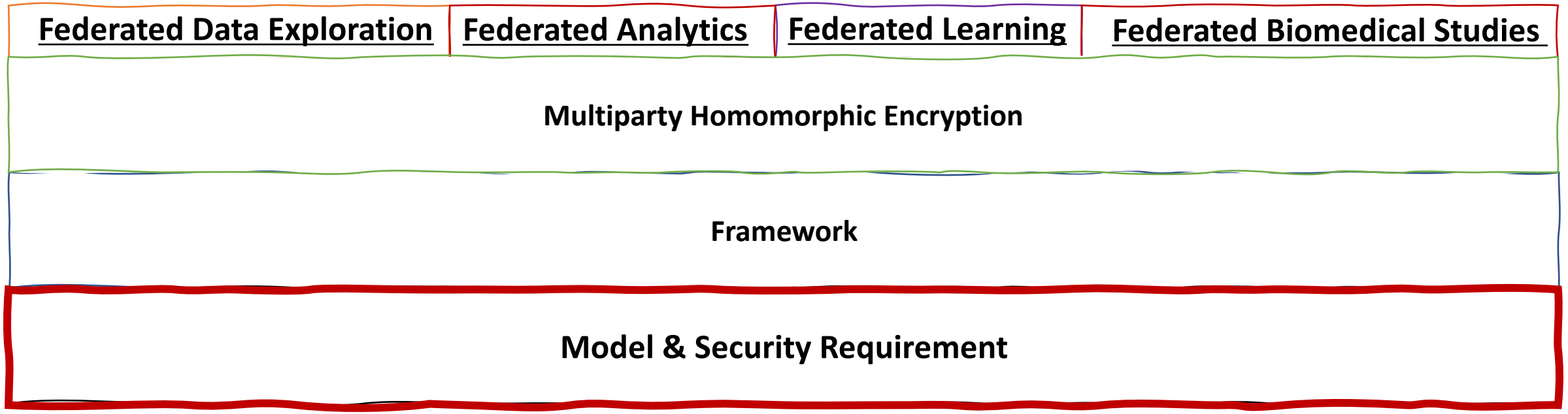
Structure de la thèse: puis de l'apprentissage automatisé, sécurisé et fédéré

Thesis Structure



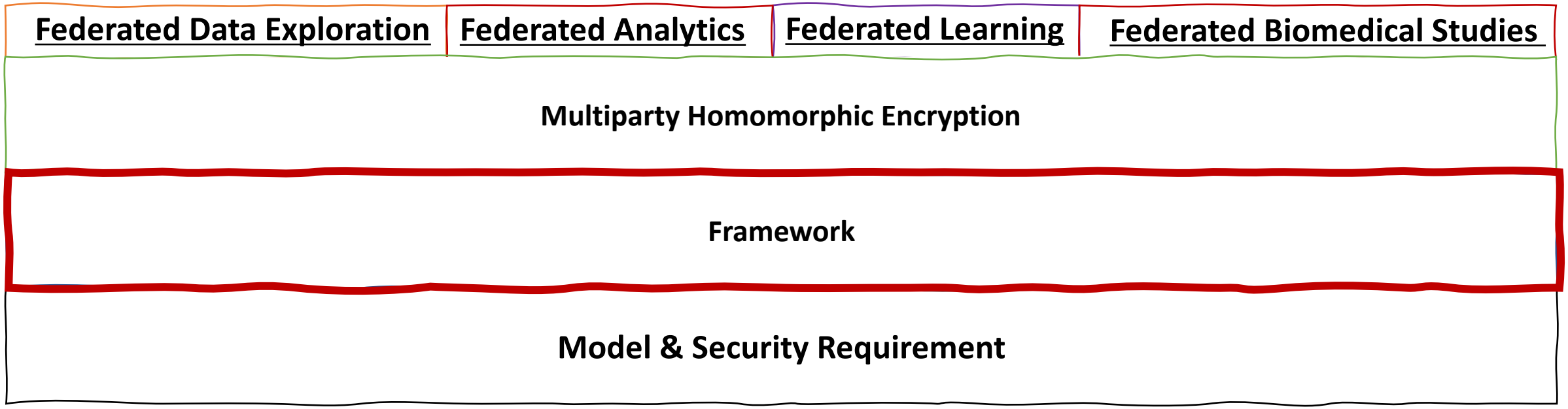
Structure de la thèse: et finalement des études biomédicales de pointe

Presentation Outline



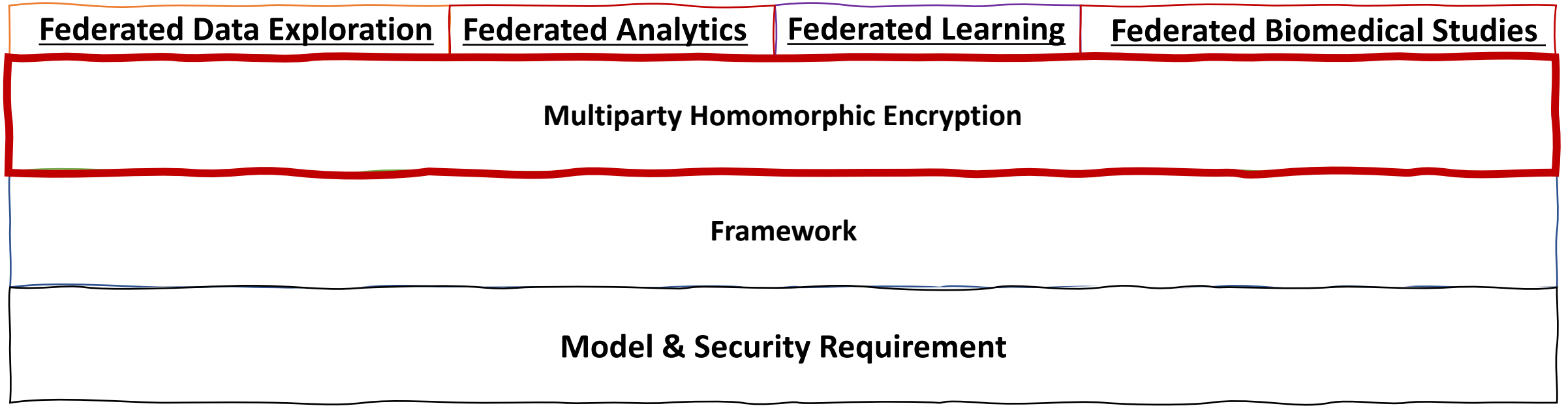
A suivre: présentation du modèle et de la structure pour toutes nos solutions, puis l'encryption homomorphique distribuée et finalement la présentation d'une partie de nos solutions.

Presentation Outline



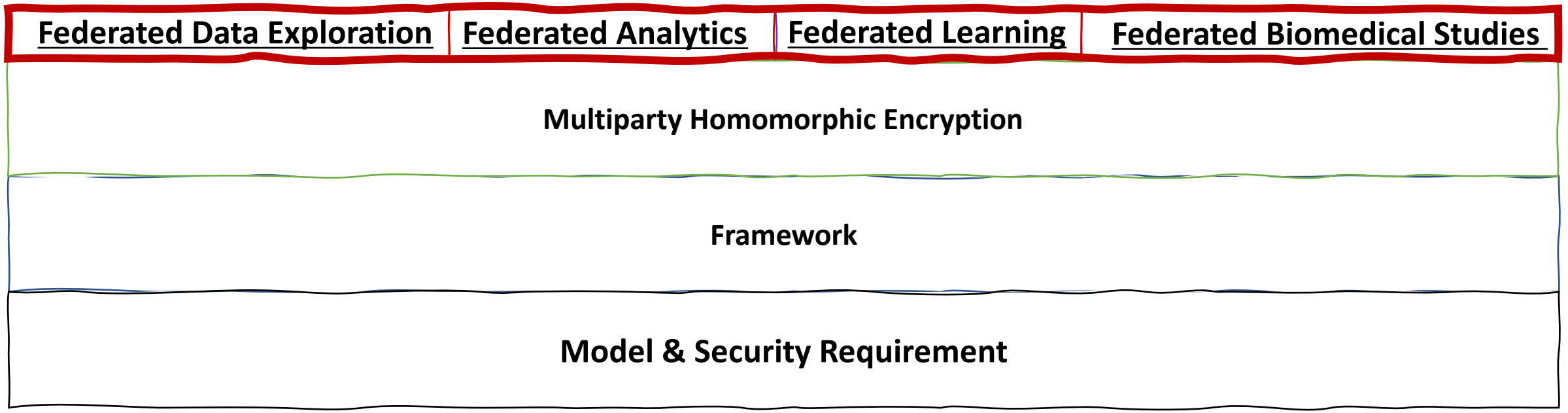
A suivre: présentation du modèle et de la structure pour toutes nos solutions, puis l'encryption homomorphique distribuée et finalement la présentation d'une partie de nos solutions.

Presentation Outline



A suivre: présentation du modèle et de la structure pour toutes nos solutions, puis l'encryption homomorphique distribuée et finalement la présentation d'une partie de nos solutions.

Presentation Outline



A suivre: présentation du modèle et de la structure pour toutes nos solutions, puis l'encryption homomorphique distribuée et finalement la présentation d'une partie de nos solutions.

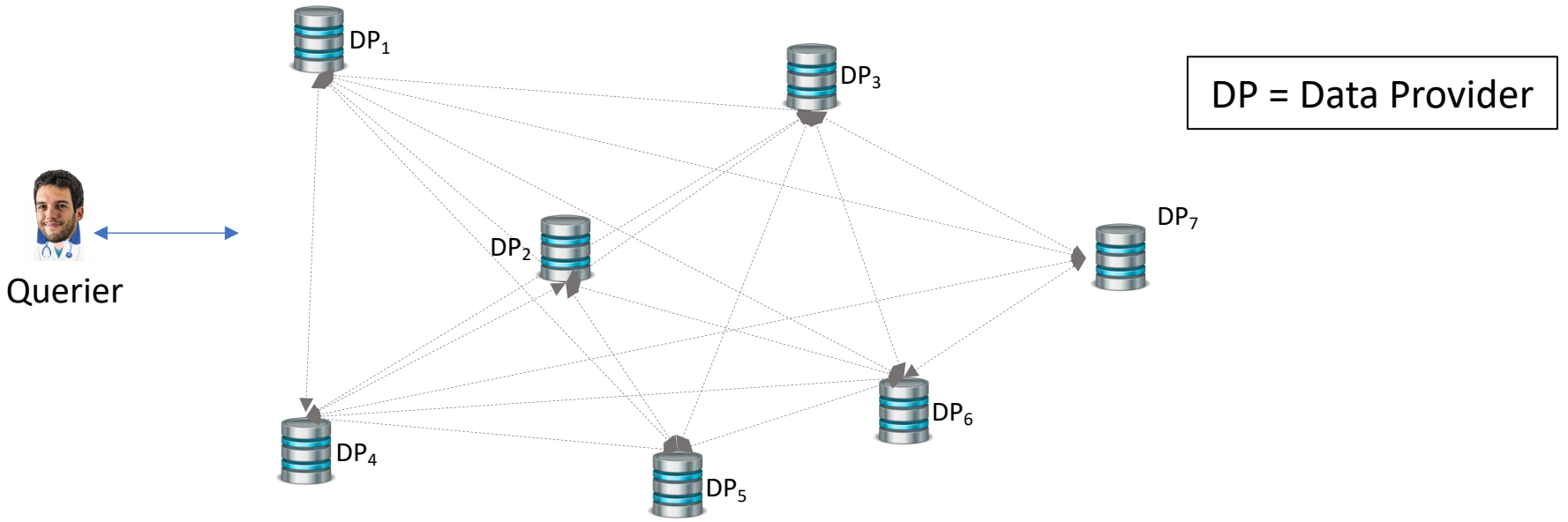
Model & Security Requirement

System Model:

Interconnected **data providers (DPs)** willing to collaborate but not to share their data. Each data sample is fully-owned by one DP.

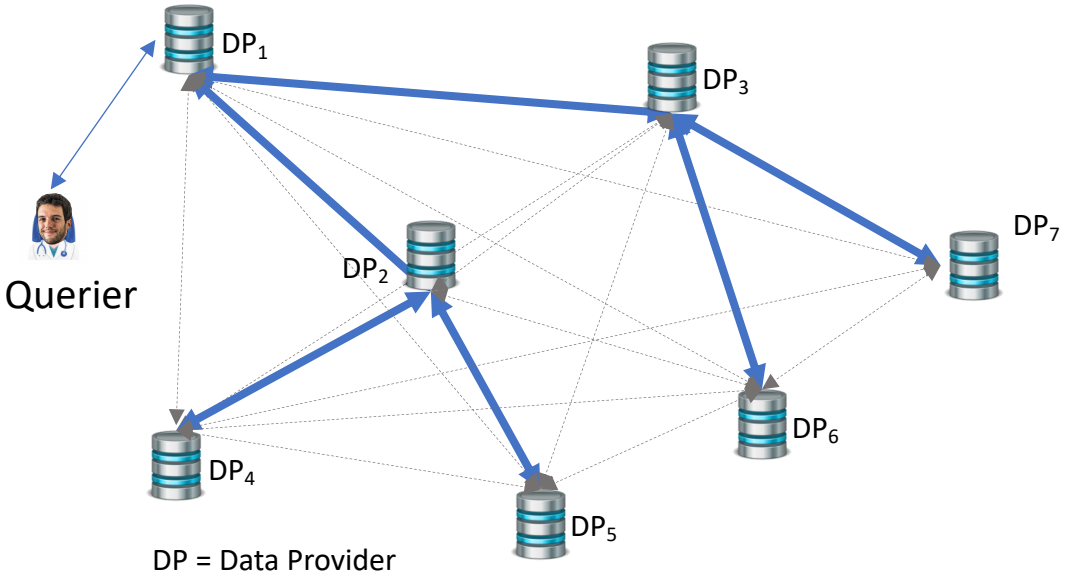
Minimum Security Requirement:

Confidentiality of the data providers' data must be ensured as long as one DP is honest.

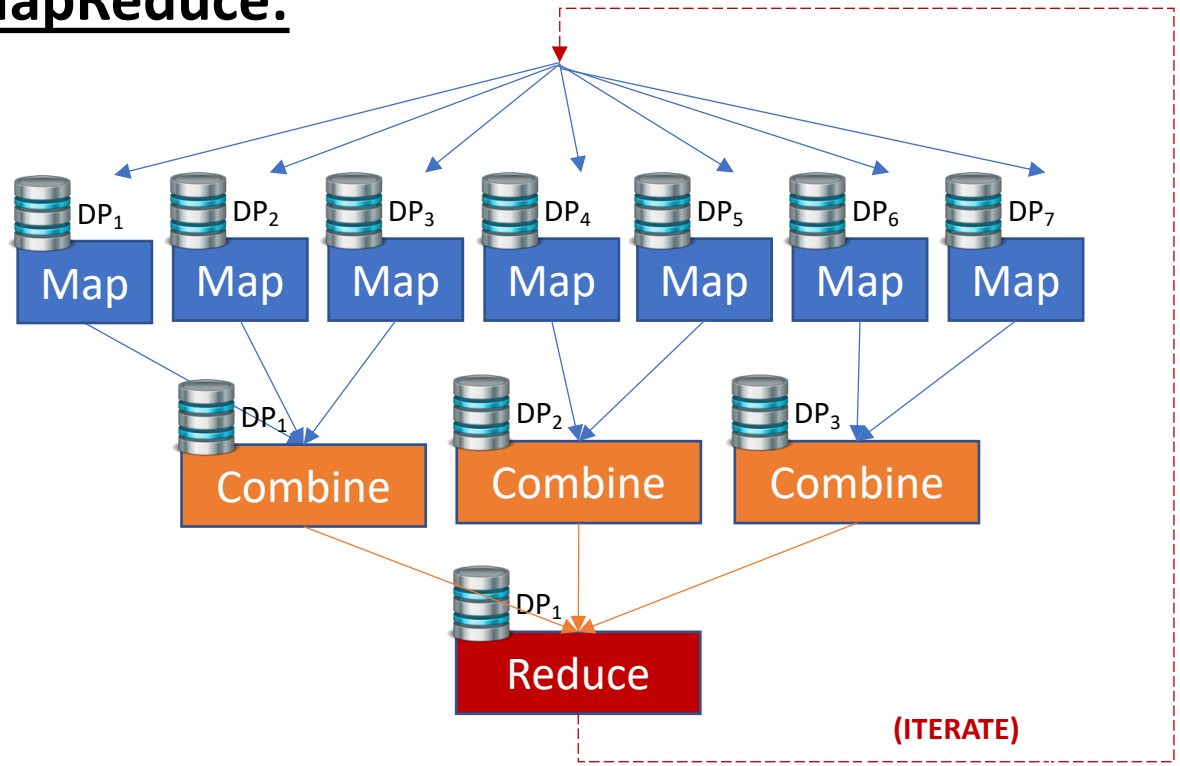


Modèle: Fournisseurs de données interconnectés (DP) disposés à collaborer mais pas à partager leurs données.
Sécurité: confidentialité des données tant que l'un des possesseurs de données est honnête.

Framework



MapReduce:



Structure dans laquelle on définit les calculs

Multiparty Homomorphic Encryption (MHE)

Homomorphic Encryption

Secret Key → *Public Key* → *Encryption* $E_{\bullet}(m)$ → *Homomorphic operations* → *Decryption*

$E_{\bullet}(m_1) + E_{\bullet}(m_2) = E_{\bullet}(m_1 + m_2)$

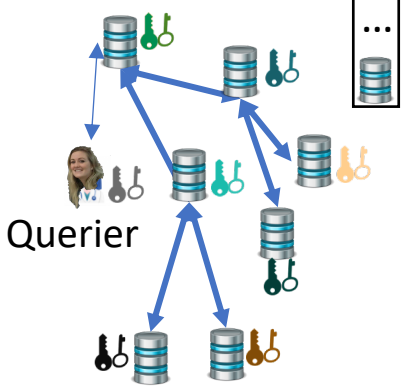
…

$E_{\bullet}(m_1) \times E_{\bullet}(m_2) = E_{\bullet}(m_1 \times m_2)$

$D_{\bullet}(E_{\bullet}(f(m_1,m_2)))$

m is a value, e.g., integer

Multiparty Homomorphic Encryption [Mouchet et al. PETS'21]



Collective *Public Key* = $f(\bullet\bullet\bullet\bullet\bullet\bullet)$ → *Encryption* $E_{\bullet}(m)$ → *Homomorphic operations*

$E_{\bullet}(m_1) + E_{\bullet}(m_2) = E_{\bullet}(m_1 + m_2)$

…

$E_{\bullet}(m_1) \times E_{\bullet}(m_2) = E_{\bullet}(m_1 \times m_2)$

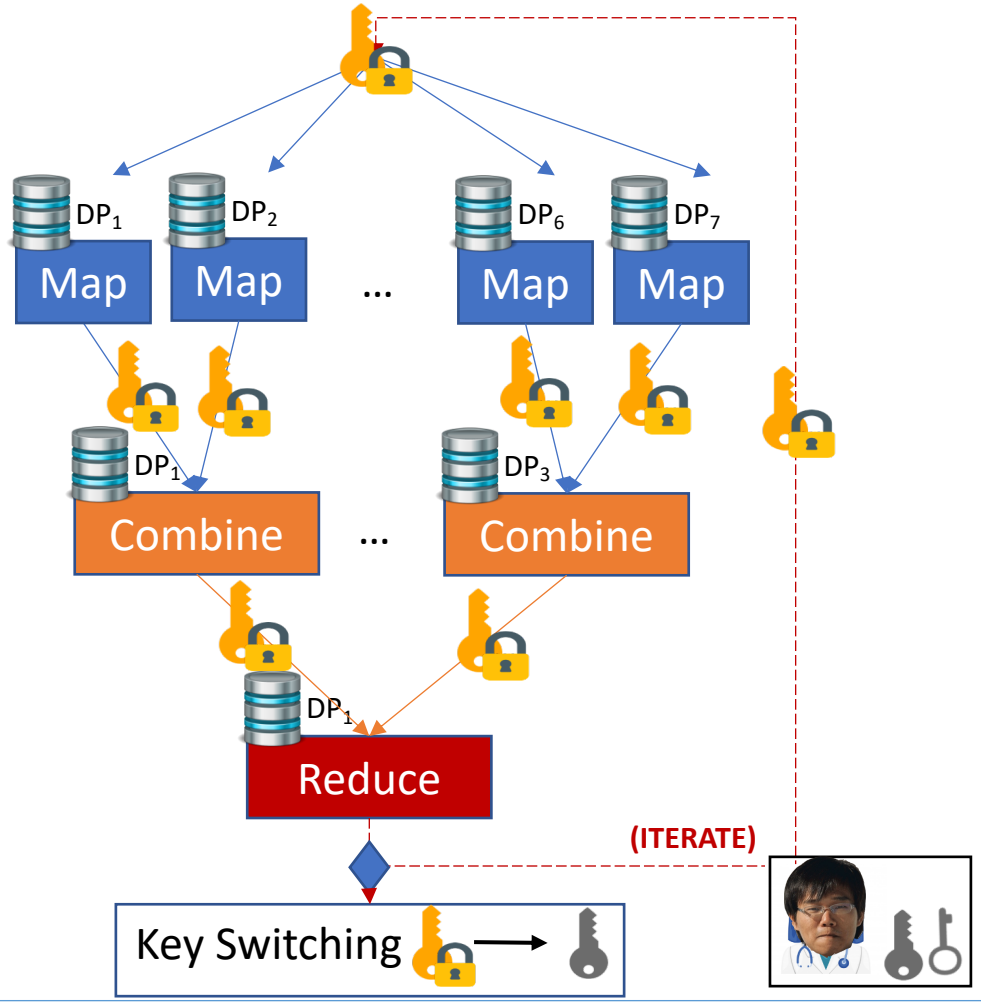
Collective *Key Switching*

$E_{\bullet}(f(m_1,m_2)) \rightarrow E_{\bullet}(f(m_1,m_2))$ → Decryption

$D_{\bullet}(E_{\bullet}(f(m_1,m_2)))$

Méthode d'encryption qui permet de faire des calculs directement sur les données encryptées.

MHE in our Framework

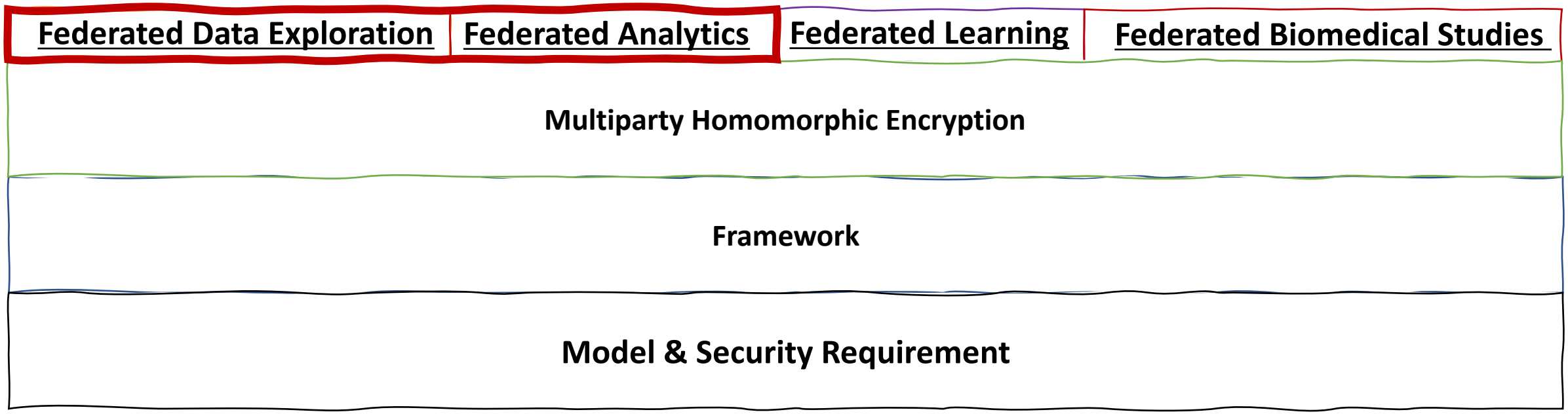


- **Information exchanged** between data providers are **collectively encrypted**
  - Cannot be decrypted unless all DPs participate
- Final result can be switched from the collective key to the querier's public key

→ **Only the querier can decrypt only the final result**

Nous encryptons les données échangées afin de préserver la confidentialité de toutes les données.

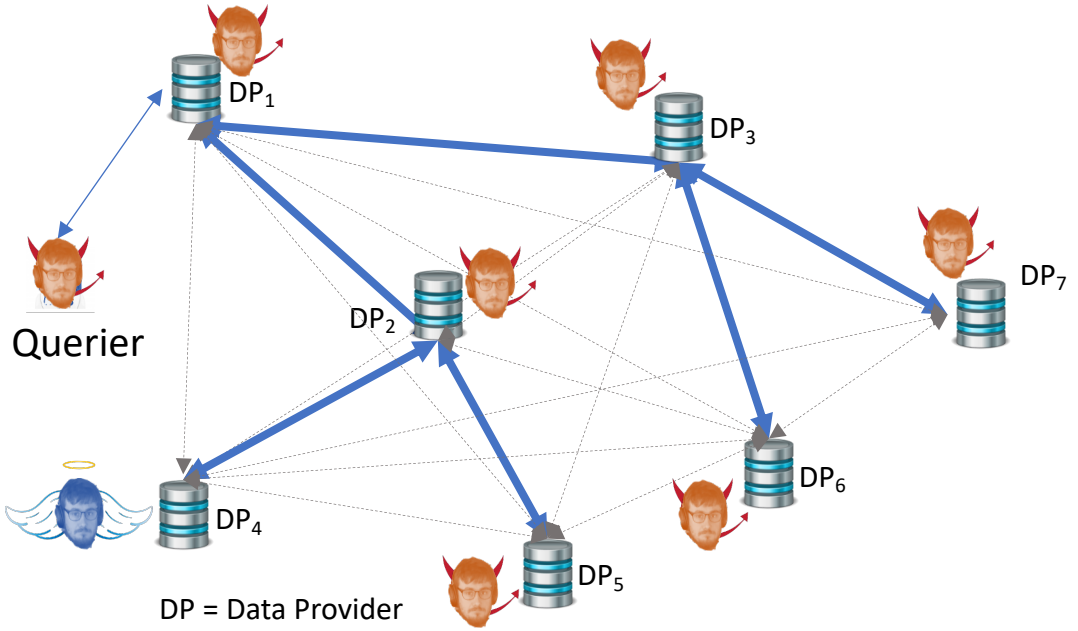
Presentation Outline



A suivre: brève présentation d'une partie de nos solutions.

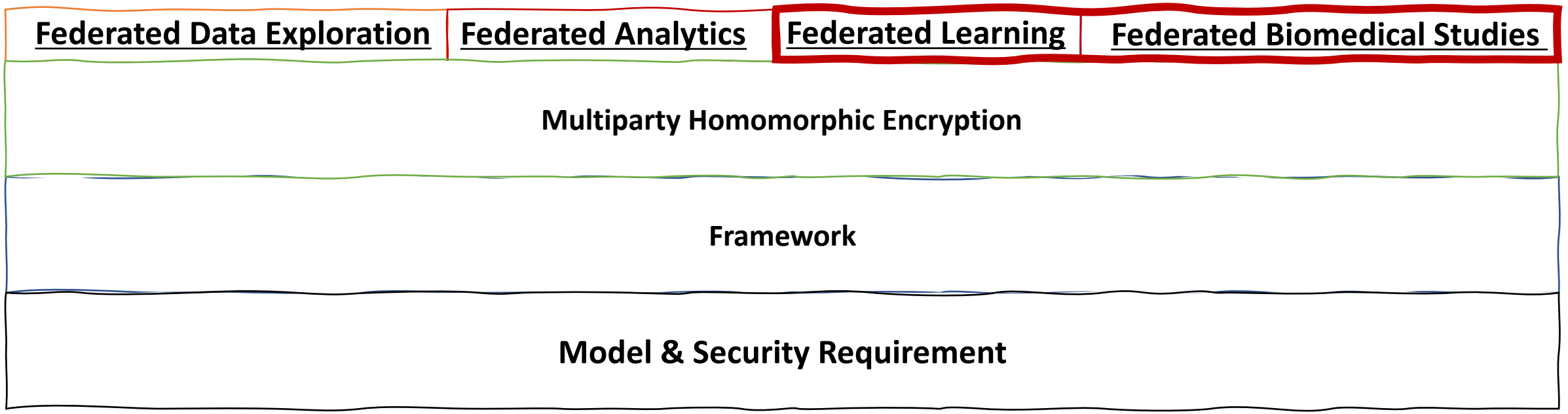
Federated Data Exploration & Analysis

- Enables the computation of **statistics**
- Preserves **data confidentiality and privacy**
- Collectively ensures **differential privacy**
- Ensures **computation correctness & results robustness without delaying the query execution**
  - **As long as one DP is honest**



Brève présentation de nos solutions pour l'exploration et l'analyse statistique de données distribuées.

Presentation Outline



A suivre: présentation de l'autre partie de nos solutions.

Federated Learning

Machine Learning:

Name	Age	Height	Gen.	SE
David	29	1.83		-
Emeline	26	1.64		-
Gouni	9	0.7		8
Poulain	14	2.05		-

→ **Training**: Find weights **w** for *Age* (a) and *height* (h) such that they can be used to predict Side-effect (SE):

$a \cdot w_a + h \cdot w_h = SE$

→ **Prediction**: Predict SE in new patients

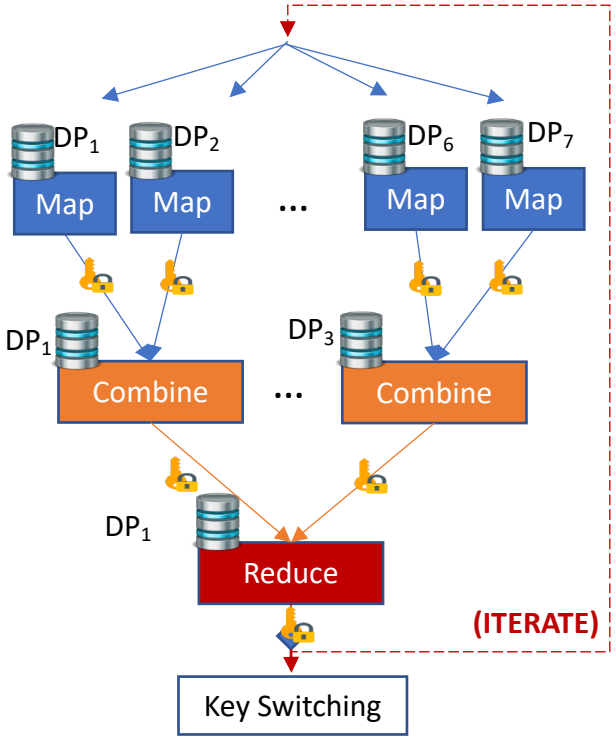
Marius	29	4.04		?

Goal: Instantiate our framework such that it

- ❖ Enables the **cooperative training of an encrypted model**
- ❖ Enables an **oblivious evaluation** of the encrypted trained model
- ❖ Ensures data provider's **data confidentiality & model confidentiality**
- ❖ Remains secure **against a passive adversary** controlling all-but-one DPs

Très brève introduction à l'apprentissage automatisé. Notre but est de permettre l'entrainement et l'évaluation de modèles en protégeant la confidentialité des données qui sont utilisées.

Multiparty Homomorphic Encryption Instantiation



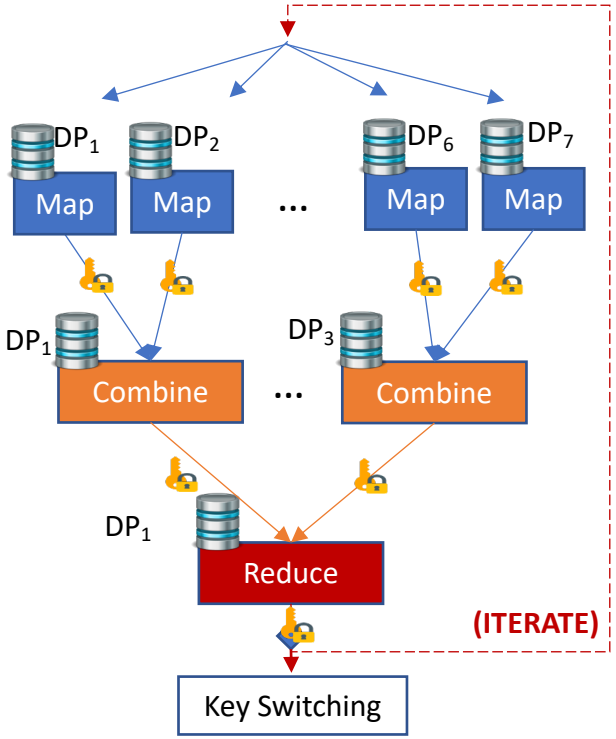
Instantiate MHE framework with:

Fully Homomorphic Encryption scheme [CKKS, ASIACRYPT 2017; Mouchet et al. PETS'21]

- A ciphertext encrypts **a vector of *N* values**: $E_{pk}(v_1, ..., v_N)$
- Enables **homomorphic vector additions, rotations and multiplications**
- Enables **Single Instruction, Multiple Data** operations
- Interactive protocols replace computionally heavy cryptographic operations

Nous utilisons une méthode d'encryption qui permet d'encrypter des vecteurs de valeurs dans un seul message chiffré.

Privacy-Preserving Model Training



Stochastic mini-batch Gradient Descent (SGD): model is **iteratively updated** by computing the gradient **with a batch of data**

Cooperative mini-batch SGD [Wang et al. ICML 2019]: DPs locally and iteratively update their local model and **combine them in a global model**

Map:
Each DP update its local model with:
- its local **cleartext** data
- previous local model **encrypted**
- global model **encrypted**

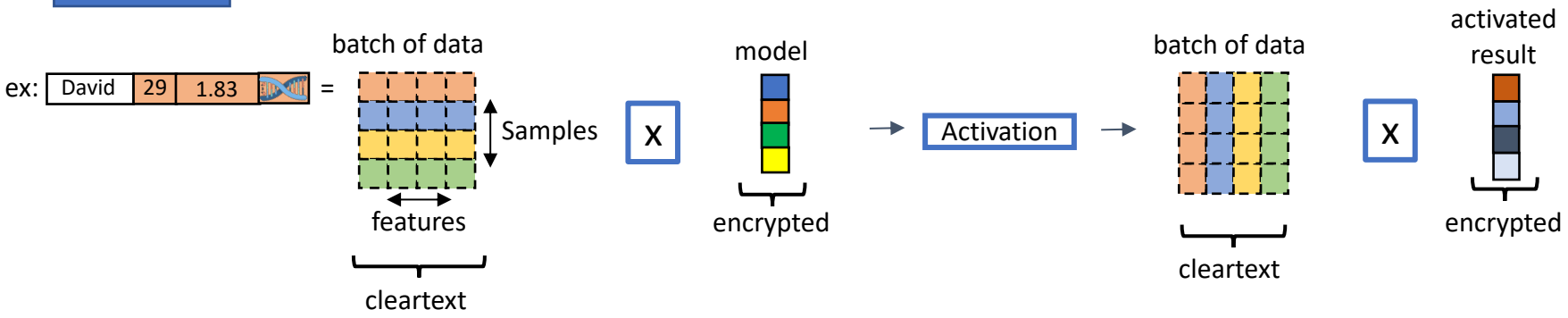
Combine:
Collective Aggregation of all DPs' **encrypted** updated models

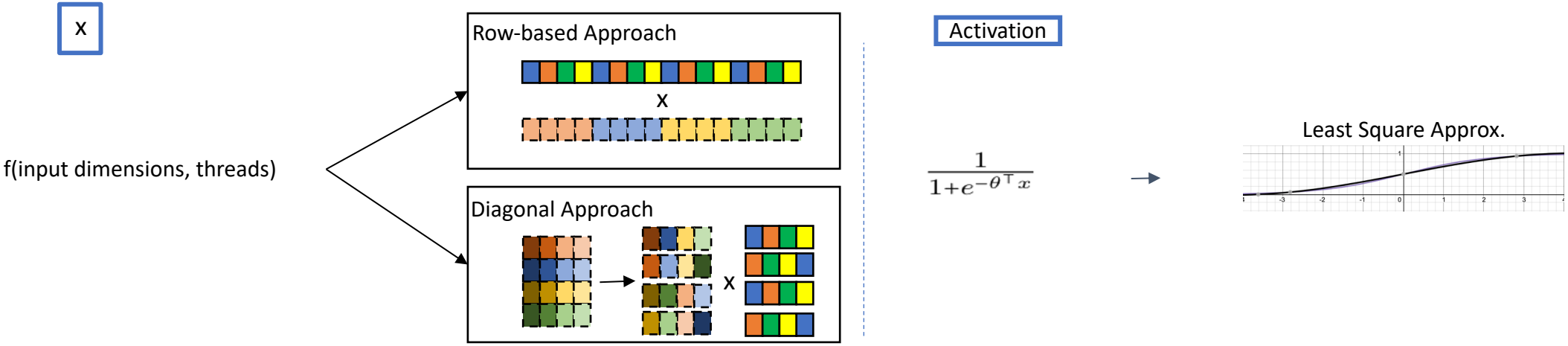
Reduce:
Update of **encrypted** global model

Nous utilisons la technique de la descente de gradient afin d'entrainer un modèle de façon distribuée.

Model Update in Map

Map Stochastic Gradient Descent for Generalized Linear Model (e.g., linear, logistic, and multinomial regressions)

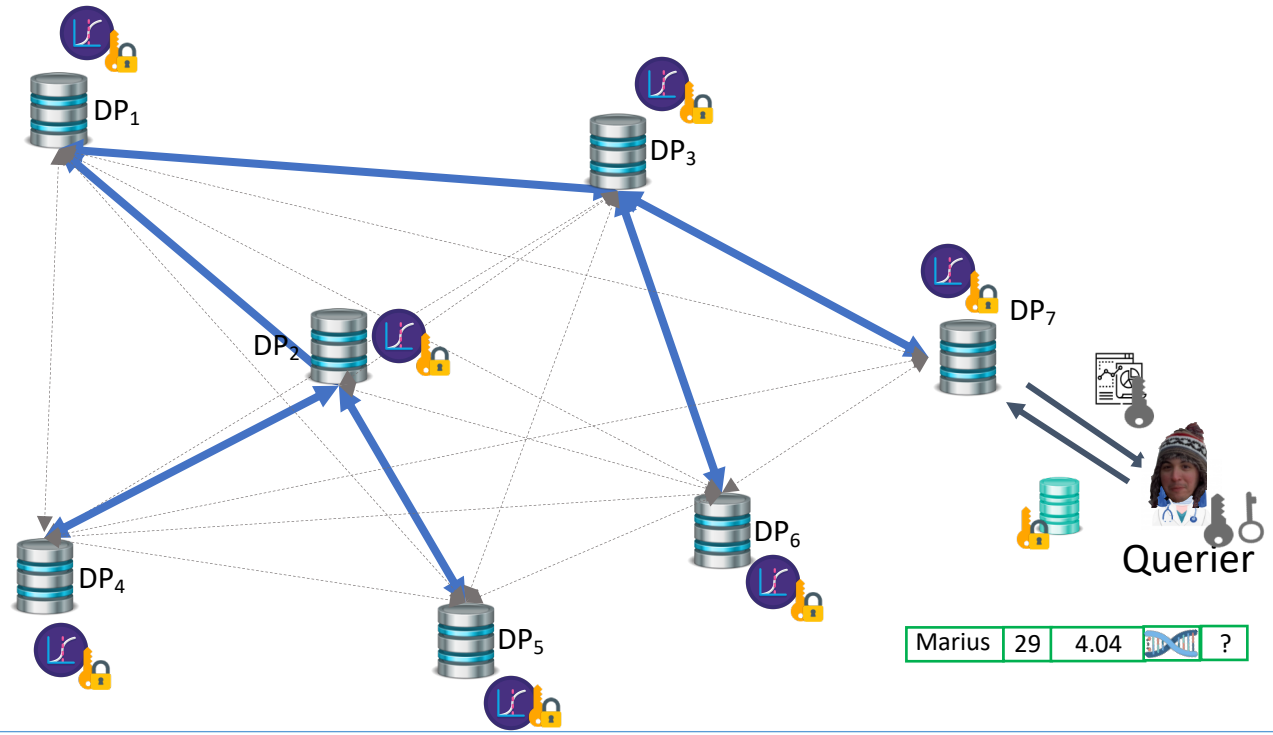


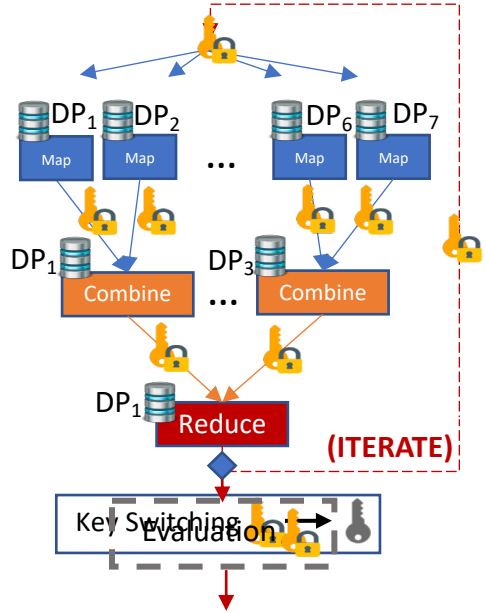


Afin d'entrainer le model efficacement nous «vecteurizons» les données et le système choisit automatiquement la meilleur approche de calcul en se basant sur les dimensions des données.

Oblivious Model Evaluation

To cover the **entire ML workflow**, the trained model can remain collectively encrypted and be used for **oblivious model evaluation**.



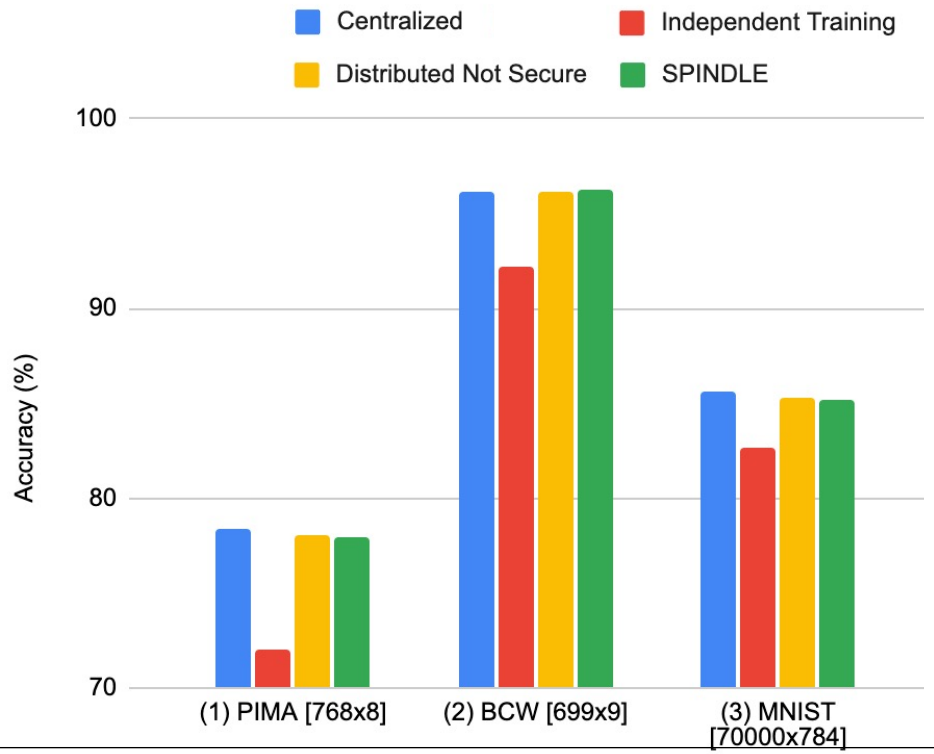


Le model entrainé peut être gardé secret afin de faire des prédictions sur des données confidentielles avec un modèle qui n'est jamais révélé.

Accuracy Evaluation

SPINDLE = instantiation of our solution for Generalized Linear Models (linear, logistic, multinomial regressions)

→ achieves accuracy close to centralized solution and **(almost) same accuracy as non-secure distributed solutions**



Evaluation Parameters

10 Data providers

128-bit security level

Legend

Dataset: *Name [#samples x #features]*

(1) Pima = Pima Indians Diabetes
https://www.kaggle.com/uciml/pima-indians-diabetes-database

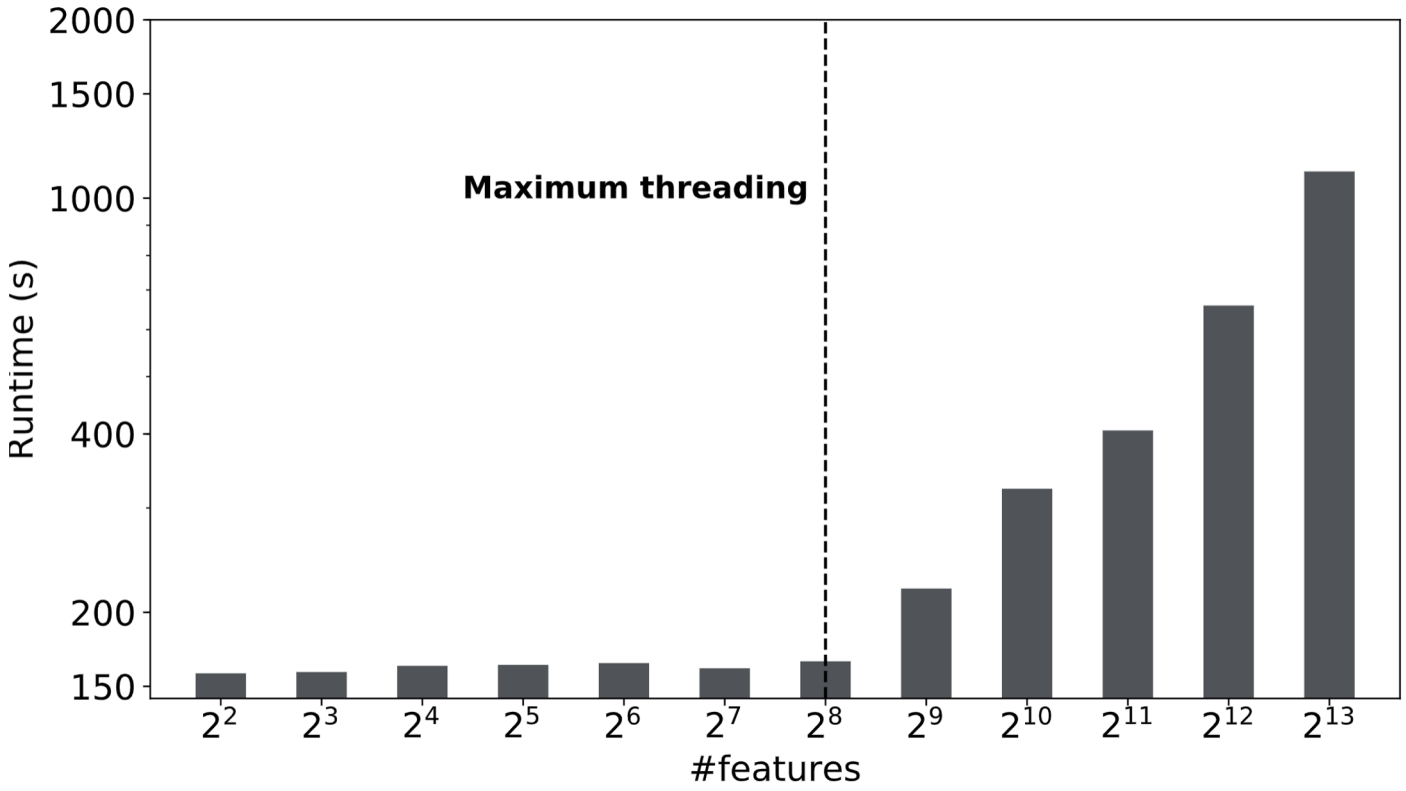
(2) BCW = Breast cancer Wisconsin (original)
https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original)

(3,4) MNIST
Y. LeCun and C. Cortes. Handwritten digit database. 2010.

Nous obtenons la même précision (ou presque) que les solutions non sécurisées.

Performance Evaluation

Better than logarithmic increase with the number of features (model size)

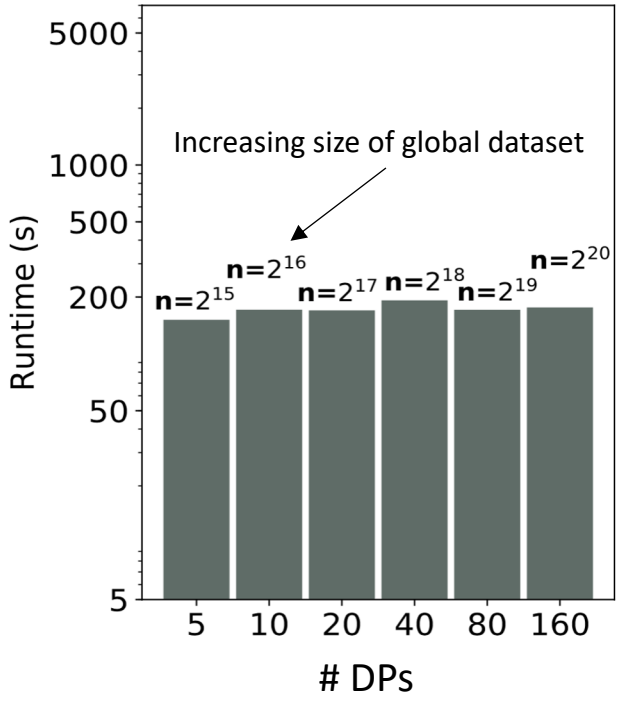


5 data providers, 25600 record, 128-bit security; Each data provider: 2 Intel Xeon E5-2680 v3 CPUs, 2.5GHz frequency, 24 threads on 12 cores, 256GB RAM. Communication: 100Mbps, delay 20ms

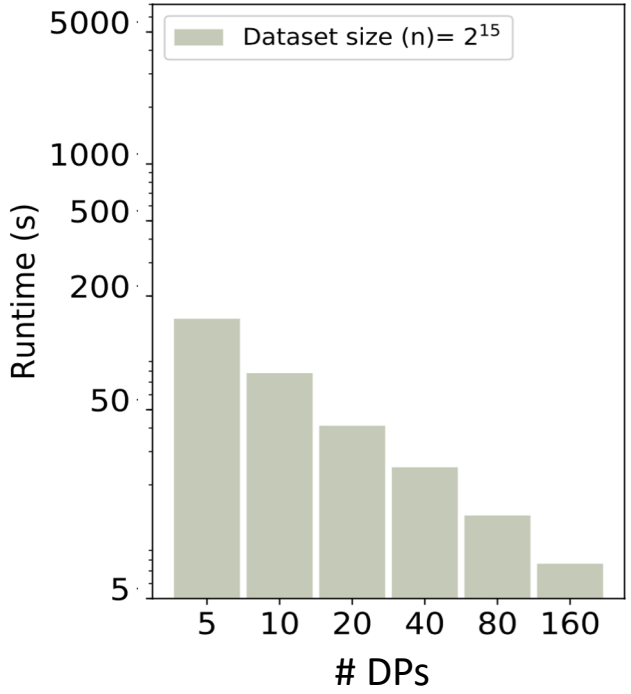
Notre solution permet l'entrainement de modèles qui ont un grand nombre de paramètres.

Performance Evaluation

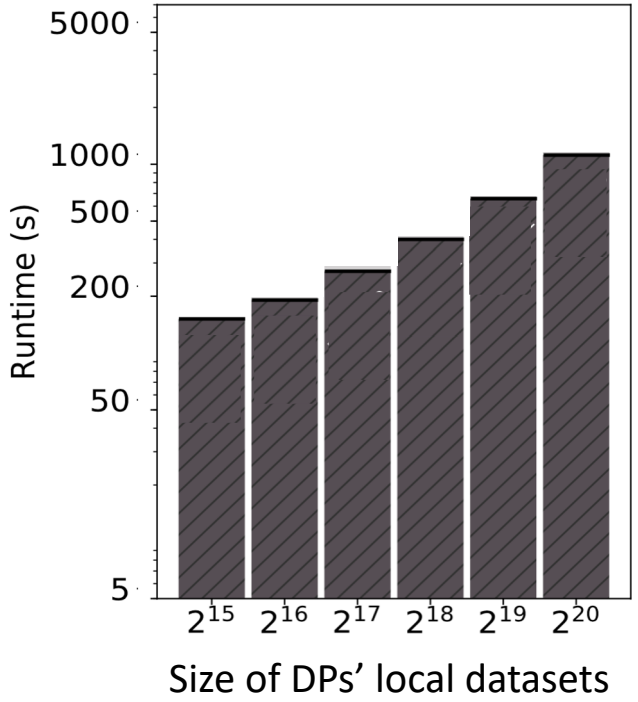
Scales almost independently with the number of data providers



Efficient workload distribution



Scales linearly with the size of the data providers datasets



Notre solution permet l'entrainement de modèles avec un grand nombre de fournisseur de données et avec de grand ensembles de données.

Machine Learning Computations

We have proposed a scalable privacy-preserving system that:

- Enables the **efficient, iterative and distributed execution of a gradient descent on an encrypted model**
- Enables an **oblivious evaluation** of the trained model
- Remains secure **against a passive adversary** controlling all but one DPs

Computation versatility
Accuracy
Scalability
No single point of failure
No data leakage
No data outsourcing

On atteint tous les objectifs fixés. Youpi !

Our Framework for Practical Use Cases

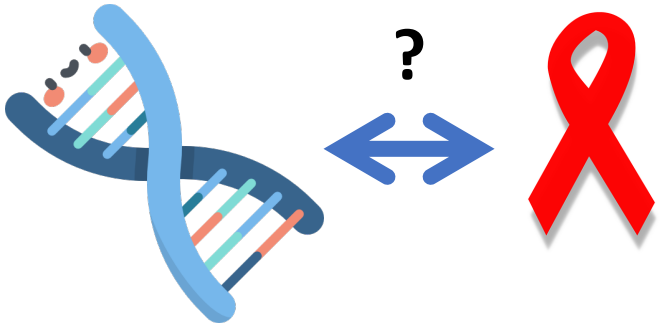
Goal: Reproduce in our framework biomedical studies that originaly relied on data centralisation.

Study : Genome-Wide Association Study:

McLaren et al. [PNAS 2015] studied the **link between HIV viral load and specific genome variants**.

Polymorphisms of large effect explain the majority of the host genetic contribution to variation of HIV-1 virus load

Paul J McLaren [1], Cedric Coulonges [2], István Bartha [1], Tobias L Lenz [3], Aaron J Deutsch [4], Arman Bashirova [5], Susan Buchbinder [6], Mary N Carrington [7], Andrea Cossarizza [8], Judith Dalmau [9], Andrea De Luca [10], James J Goedert [11], Deepti Gurdasani [12], David W Haas [13], Joshua T Herbeck [14], Eric O Johnson [15], Gregory D Kirk [16], Olivier Lambotte [17], Ma Luo [18], Simon Mallal [19], Daniëlle van Manen [20], Javier Martinez-Picado [21], Laurence Meyer [22], José M Miro [23], James I Mullins [24],

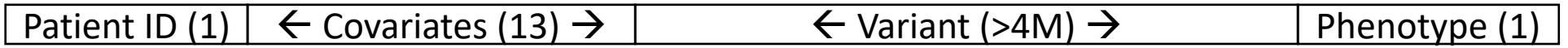


On veut maintenant montrer que nos solutions peuvent être utilisées «en vrai», notamment pour des études biomédicales.

Study: Genome-Wide Association Study

GWAS principle: Find whether each variant (combined with the covariates) have a link with the phenotype

1 data sample:



example: age, height, …

example: 0, 1, 0, …., 1, 2

example: HIV viral load

Basic approach: **Train > 4 million regression models** and compute the p-value of the variant's weight

Optimized approach: Include the covariates contribution in the phenotype and train 4M **univariate models**

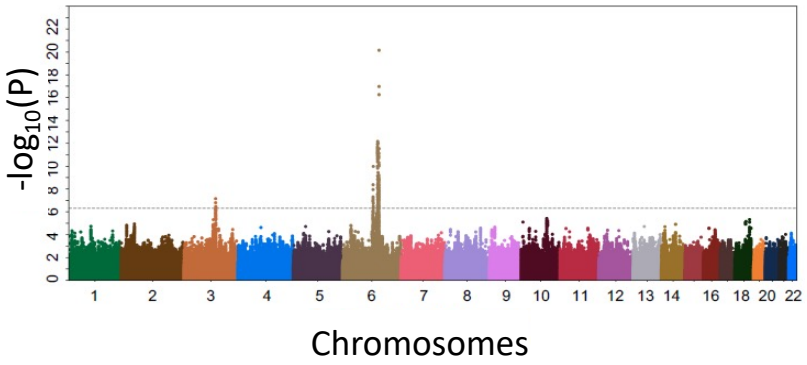
Challenges: Requires encrypted-matrix inversion and multiple matrix multiplications

Étude d'association pangénomique. Nous optimisons l'approche de calcul afin de le faire efficacement.

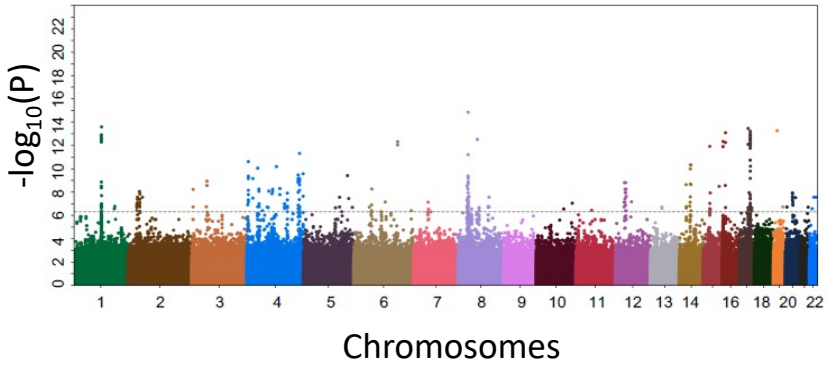
Study 2: Genome-Wide Association Study

Setting: 1857 patients spread among 12 data providers.

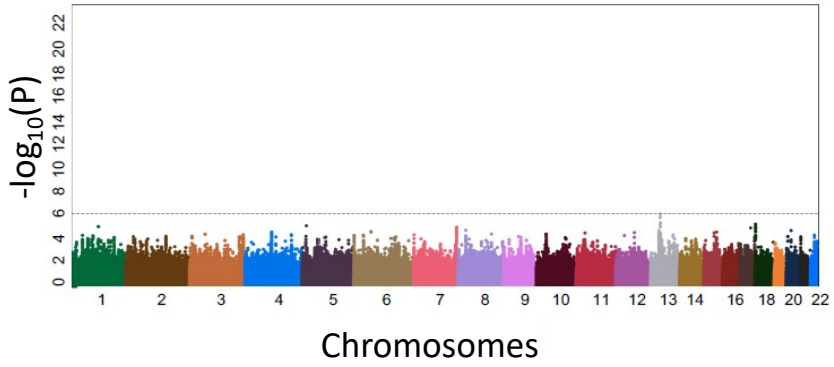
Original Study (Centralized, Non-Secure)



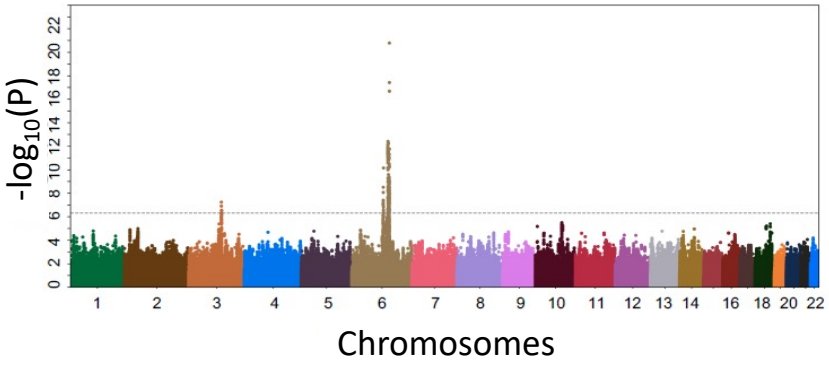
Meta-Analysis (Decentralized, Non-Secure)



Independent (1 single data provider)



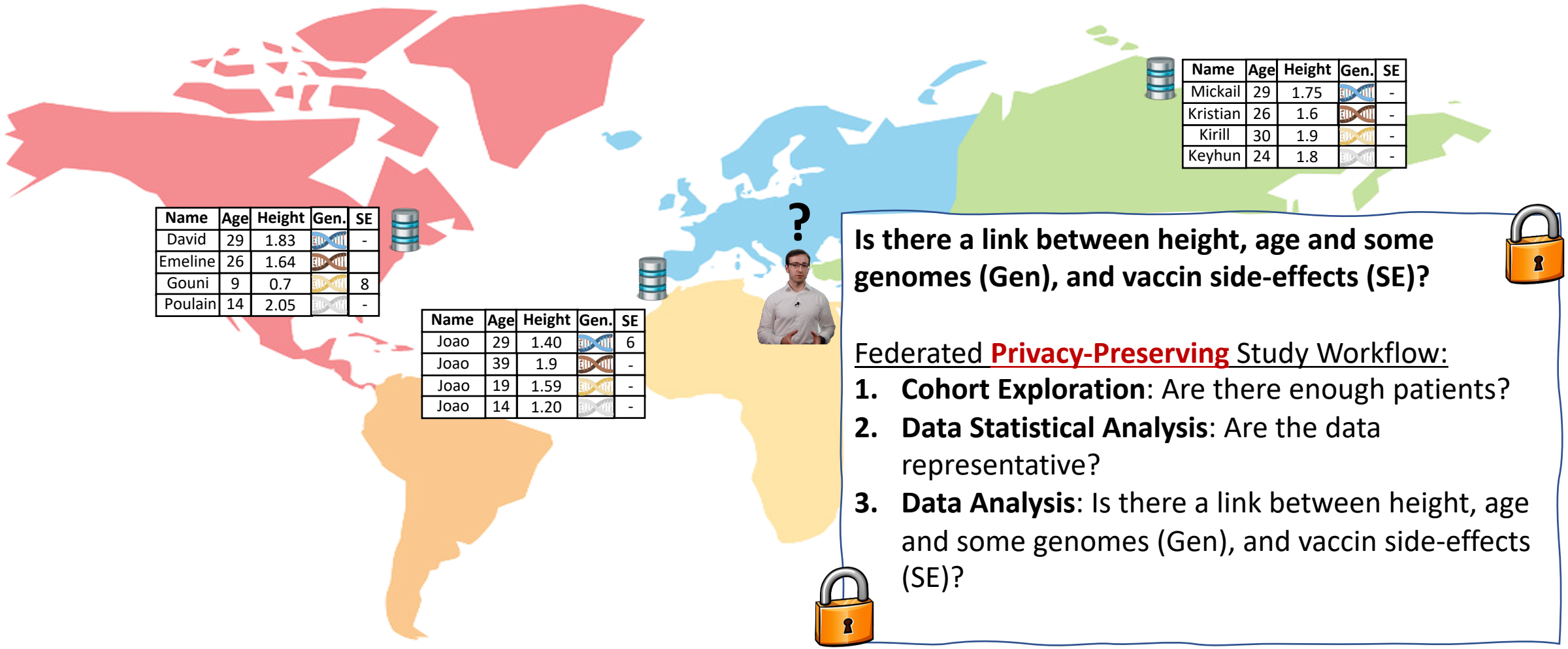
Our Approach (Decentralized, Secure)



Almost Exact Results
+
Same scalability as before

On obtient les même résultats que l'étude originale.

Federated Analytics in the Medical Domain



Name	Age	Height	Gen.	SE
David	29	1.83		-
Emeline	26	1.64		
Gouni	9	0.7		8
Poulain	14	2.05		-

Name	Age	Height	Gen.	SE
Joao	29	1.40		6
Joao	39	1.9		-
Joao	19	1.59		-
Joao	14	1.20		-

Name	Age	Height	Gen.	SE
Mickail	29	1.75		-
Kristian	26	1.6		-
Kirill	30	1.9		-
Keyhun	24	1.8		-

Is there a link between height, age and some genomes (Gen), and vaccin side-effects (SE)?

Federated **Privacy-Preserving** Study Workflow:

1. **Cohort Exploration**: Are there enough patients?
2. **Data Statistical Analysis**: Are the data representative?
3. **Data Analysis**: Is there a link between height, age and some genomes (Gen), and vaccin side-effects (SE)?

On remarque que l'étude décrite au début peut-être exécutée en entier avec les solutions proposées.

Impact

MedCO
Collective protection of medical data
https://medco.epfl.ch/

Cohort Exploration Tool
- Deployed network between multiple hospitals and universities in Switzerland
- Currently being deployed in Netherlands, Italy, USA, …

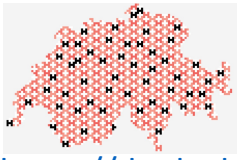
Secure Collective COVID-19 Research
https://securecovidresearch.org/

TUNE INSIGHT
https://tuneinsight.com/

Startup building a tool for secure collaboration and federated analytics

3 patents filed based on our work and subsequent work

DPPH
Data Protection in Personalized Health



https://dpph.ch/

DPPH: Data Protection in Personalized Health funded by PHRT.
2018-2021 | Budget: CHF 3M

MedCo: Enabling the Secure and Privacy-Preserving Exploration of Distributed Clinical and *Omics Cohorts in the Swiss Personalized Health Network (SPHN) funded by the PHRT and the SPHN. 2019-2021 | Budget: CHF 0,5 M

Un software déployé internationalement, une startup et deux projets à l'échelle Suisse.

Conclusion

Federated Statistical Analytics

*Enabled **federated analytics** by building on an **additive MHE** scheme while ensuring **auditability** in a **strong threat model** without influencing the **query execution time***

Federated Machine Learning

*Enabled the **efficient federated execution** of the **iterative** gradient descent on a model encrypted with a **fully MHE** scheme on data held by **large numbers of data providers***

Federated Biomedical Studies

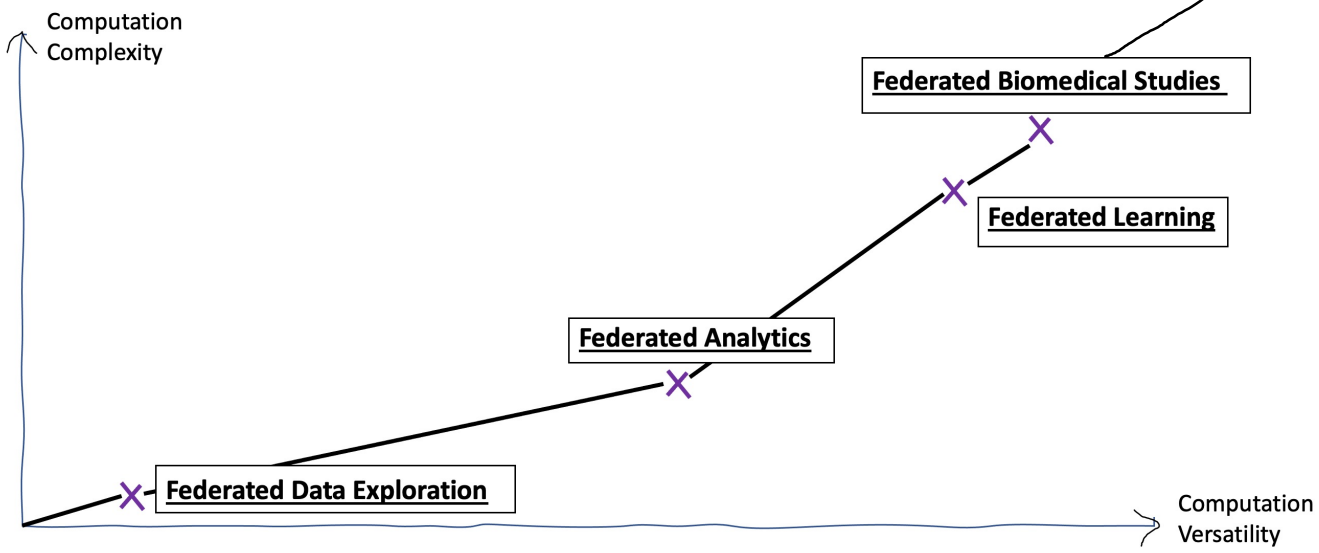
*Efficient, secure, federated, and accurate reproduction of **complex and large-dimensional-inputs workflows***

Merci d'avoir écouté jusque là, c'est presque fini.

Future Work

Complex and high-dimensional computations

Adoption of secure solutions



MIT Technology Review

April 15, 2021

Why is it so hard to review the Johnson & Johnson vaccine? Data.

The clock is ticking for regulators looking into covid vaccine side effects. But their task is made harder by America's fragmented data systems.

MIT Technology Review

Healthcare

It took a pandemic, but the US finally has (some) centralized medical data

Covid exposed the fragmented reality of US health records. Now an effort to bring together data from millions of patients is starting to show results.

by **Cat Ferguson** June 21, 2021

Pour après l'apéro…