

Privacy-Preserving Federated Analytics using Multiparty Homomorphic Encryption

David Froelicher

PhD Private Defense, 17.08.2021

Jury President:

Prof. Carmela Troncoso

Advisors:

Prof. Jean-Pierre Hubaux

Prof. Bryan Ford

Experts:

Prof. Bonnie Berger

Dr. Wei Dai

Prof. Martin Jaggi

Outline

- Motivation for Federated Analytics
- Existing Solutions for Federated Analytics
- Thesis Goal
- Thesis Structure
- Common Basis
 - Model & Security Requirements
 - Framework
- Instantiation 1: Verifiable Statistics Computations
- Instantiation 2: Machine Learning Computations
- Practical Use Cases
- Conclusion

Motivation for Federated Analytics

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.

- 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



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**



Motivation for Privacy-Preserving Federated Analytics



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



Comply by-design with regulations, e.g., GDPR.



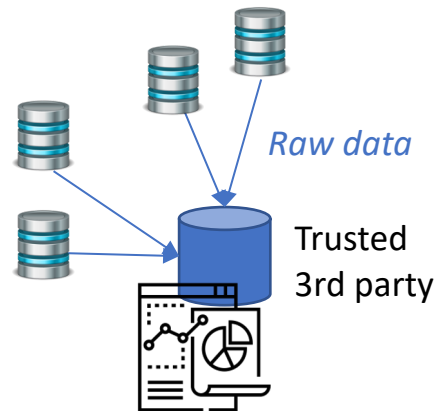
Reduce the need for data-access agreements



Control which information is revealed and **avoid** data leakages

Existing Solutions for Federated Analytics

(a) Fully centralized



❖ Flexible computations

❖ No Bias

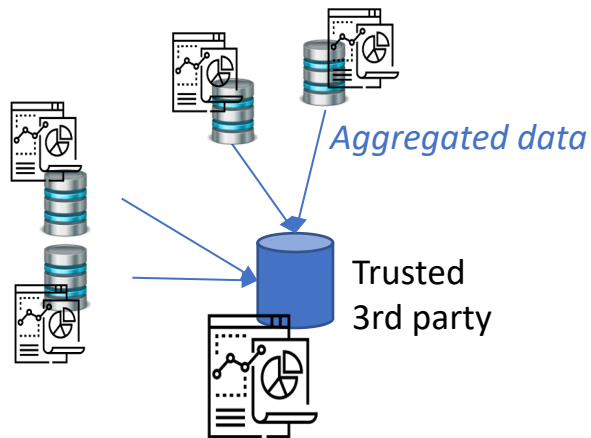
- Single point of failure
- Data leakage
- Data outsourcing (require agreements)
- Scalability with dataset size and number of data providers

Public/private Initiatives:

- All of Us (US National Institute of Health)
- EGA (European Genome-phenome Archive)
- Genomics England

Existing Solutions for Federated Analytics

(b) Meta-Analysis



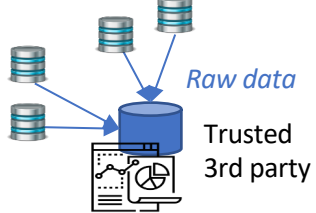
- ❖ Flexible computations
- ❖ No bias
- ❖ Scalability

- Single point of failure
- Data leakage
- Data outsourcing (require agreements)

Public/private Initiatives:

- Consortium for Clinical Characterization of COVID-19 by EHR

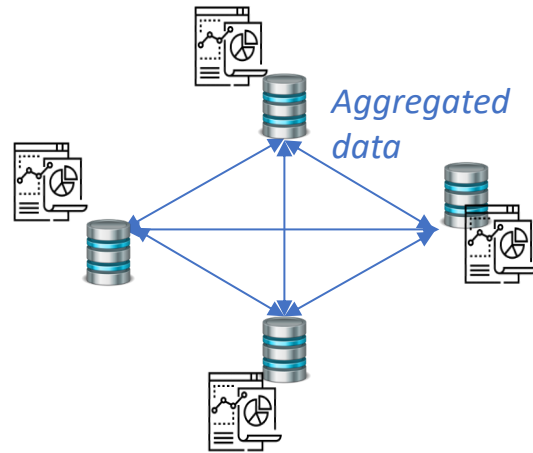
(a) Fully centralized



- ❖ Flexible computations
- ❖ No bias
- Single point of failure
- Data leakage
- Data outsourcing
- Scalability

Existing Solutions for Federated Analytics

(c) Decentralized

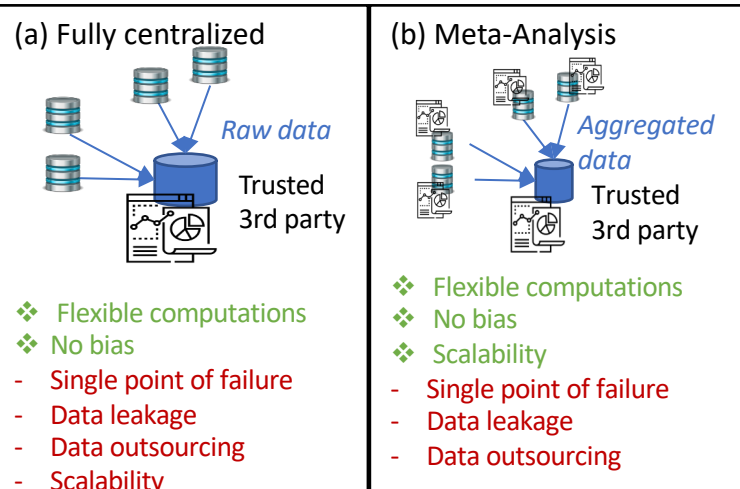


- ❖ Flexible computations
- ❖ No bias
- ❖ Scalability

- Data leakage

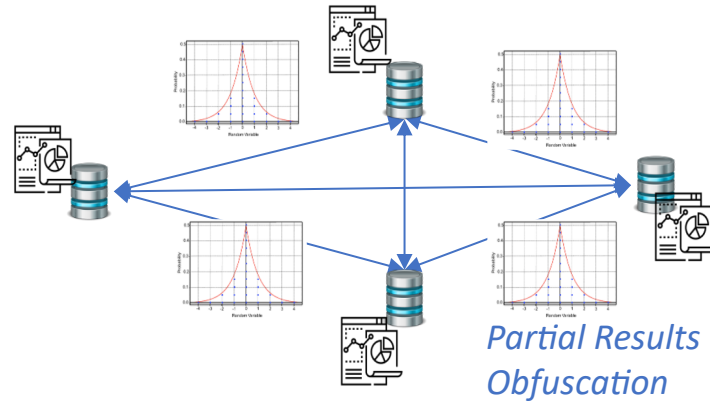
Public/private Initiatives:

- DataSHIELD (UK)
- Personalized Health Train (PHT)
- Vantage6



Existing Solutions for Federated Analytics

(d) Differential-Privacy-based Decentralized

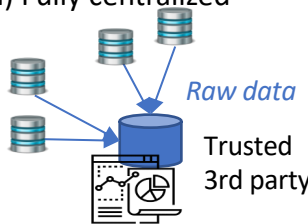
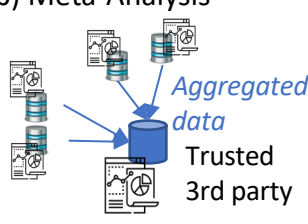
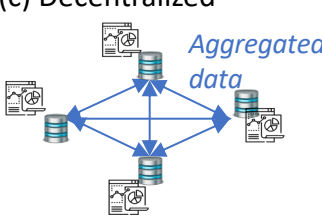


- ❖ Flexible computations
- ❖ Scalability

- Bias introduced by-design
- Trade-off between data leakage and accuracy

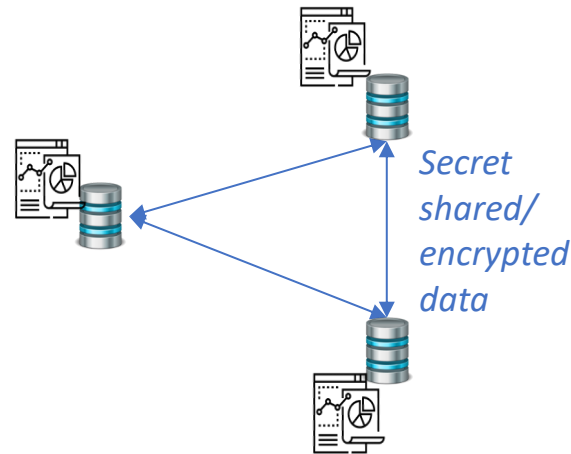
Public/private Initiatives:

- NVIDIA

(a) Fully centralized	(b) Meta-Analysis	(c) Decentralized
		
<ul style="list-style-type: none">❖ Flexible computations❖ No bias- Single point of failure- Data leakage- Data outsourcing- Scalability	<ul style="list-style-type: none">❖ Flexible computations❖ No bias❖ Scalability- Single point of failure- Data leakage- Data outsourcing	<ul style="list-style-type: none">❖ Flexible computations❖ No bias❖ Scalability- Single point of failure- Data leakage- Data outsourcing

Existing Solutions for Federated Analytics

(e) Cryptography-based (SMC, HE) Decentralized



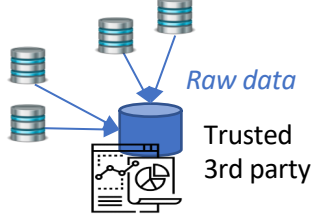
❖ No data leakage

- Difficult to scale with the number of parties
- Data outsourcing

Public/private Initiatives:

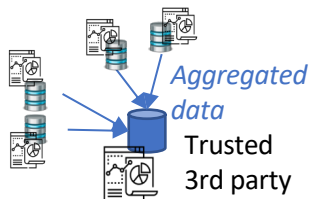
- Sharemind
- Inpher
- Duality

(a) Fully centralized



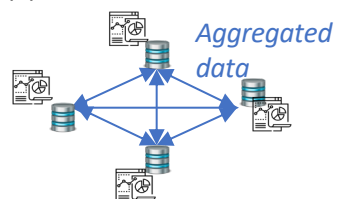
- ❖ Flexible computations
- ❖ No bias
- Single point of failure
- Data leakage
- Data outsourcing
- Scalability

(b) Meta-Analysis



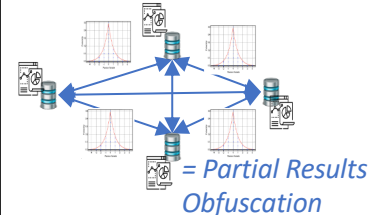
- ❖ Flexible computations
- ❖ No bias
- ❖ Scalability
- Single point of failure
- Data leakage
- Data outsourcing

(c) Decentralized



- ❖ Flexible computations
- ❖ No bias
- ❖ Scalability
- Single point of failure
- Data leakage
- Data outsourcing

(d) Differential Privacy

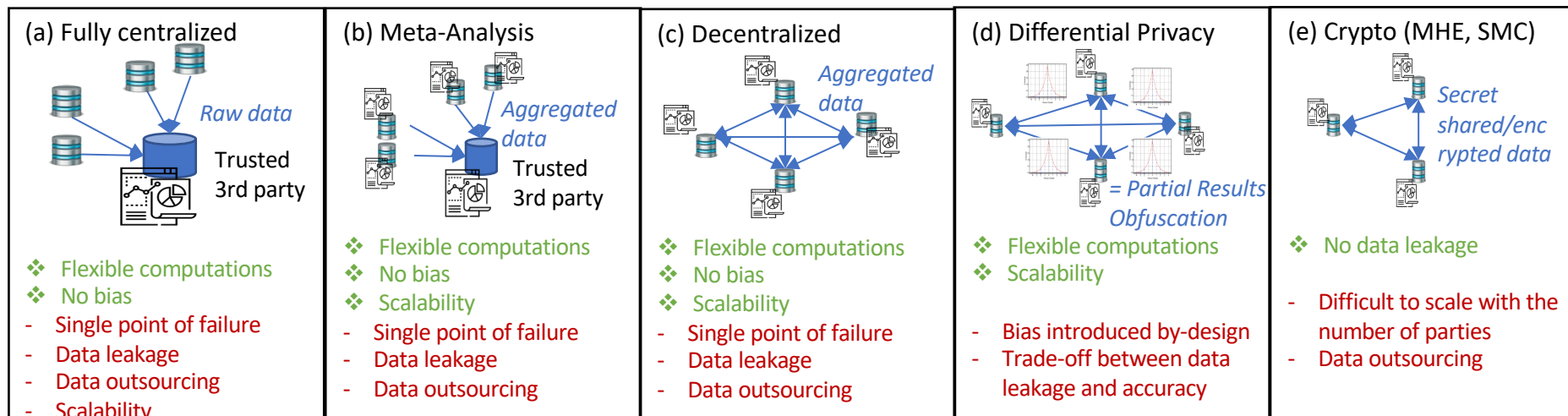


- ❖ Flexible computations
- ❖ Scalability
- Bias introduced by-design
- Trade-off between data leakage and accuracy

Thesis Goal

Modular federated privacy-preserving analytics with:

- ❖ Data confidentiality
- ❖ Flexible computations
- ❖ No bias introduced by design
- ❖ Scalability in all dimensions: number of parties, size of datasets
- ❖ Decentralization, no single point of failure



Thesis Structure

UnLynx (Chapter 3) [1]



SELECT **sum/count** ... FROM DP_1, \dots, DP_n
WHERE ... GROUP BY ...



Data privacy & **Confidentiality**
Computations Verification
Collective protection of local data



Anytrust Model with
passive adversary

Drynix (Chapter 4) [2]



SELECT **statistic()** ... FROM DP_1, \dots, DP_n
WHERE ... GROUP BY ...



Data privacy & **Confidentiality**
Computations & Input Verification



Anytrust Model with
active adversary

SPINDLE (Chapter 5) [3]



Cooperative gradient descent (training)
and model evaluation



Data & Model **Confidentiality**



Anytrust Model with
passive adversary

FAMHE (Chapter 6) [4]



Federated biomedical studies: **survival curves and genome-wide association studies**



Data **Confidentiality**



Anytrust Model with
passive adversary

[1] **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.

[2] **D. Froelicher**, J.R. Troncoso-Pastoriza, J.S. Sousa and J.P. Hubaux, "Drynx: Decentralized, Secure, Verifiable System for Statistical Queries and Machine Learning on Distributed Datasets.", IEEE TIFS, 2020.

[3] **D. Froelicher**, J. R. Troncoso-Pastoriza, A. Pyrgelis, S. Sav, J. S. Sousa, J.-P. Bossuat, and J.-P. Hubaux. "Scalable Privacy-Preserving Distributed Learning." PETS'21.

[4] **D. Froelicher**, J. R. Troncoso-Pastoriza, J. L. Raisaro, M. Cuendet, J. S. Sousa, H. Cho, B. Berger, J. Fellay, and J.-P. Hubaux. "Truly Privacy-Preserving Federated Analytics for Precision Medicine with Multiparty Homomorphic Encryption". Conditionally Accepted in Nature Communications , 2021.

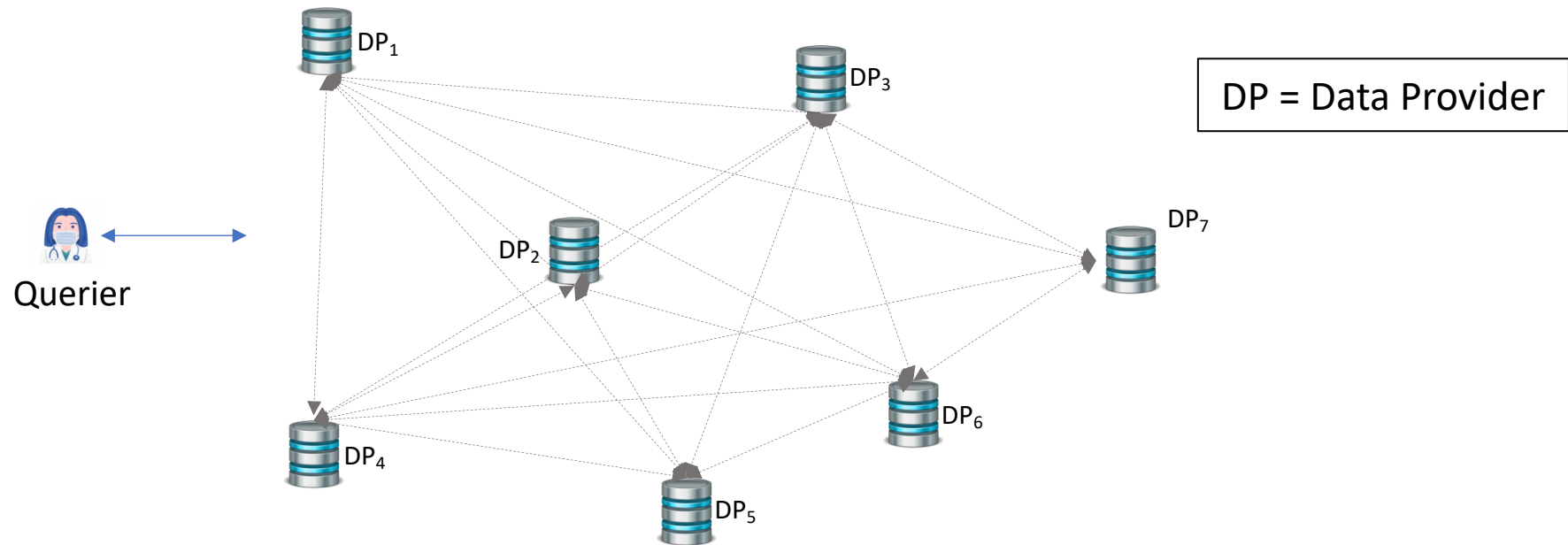
A Common Basis: Model & Security Requirement

System Model:

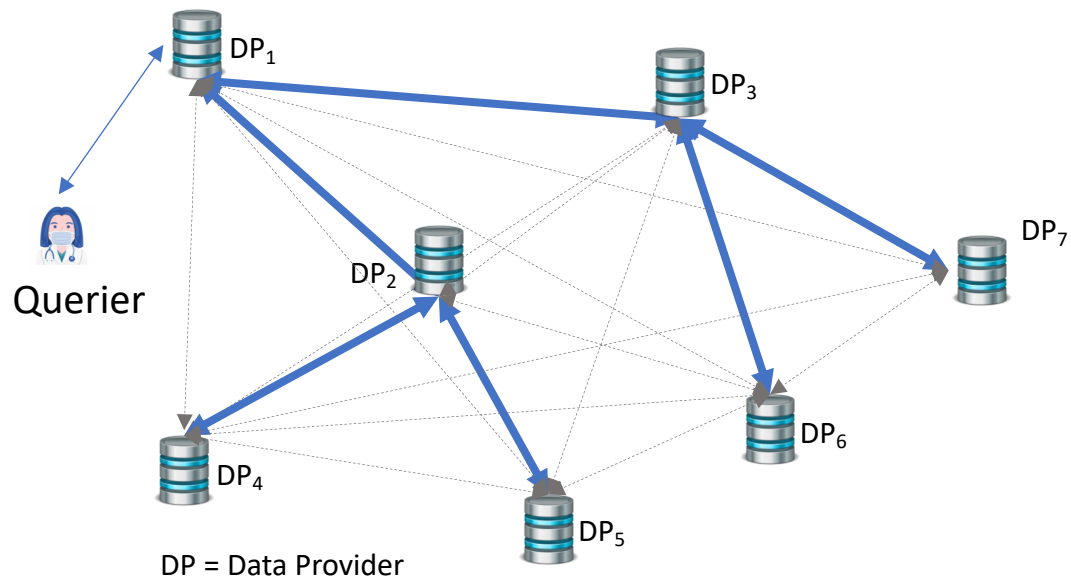
Interconnected **data providers** willing to collaborate but not to share their data.

Minimum Security Requirement:

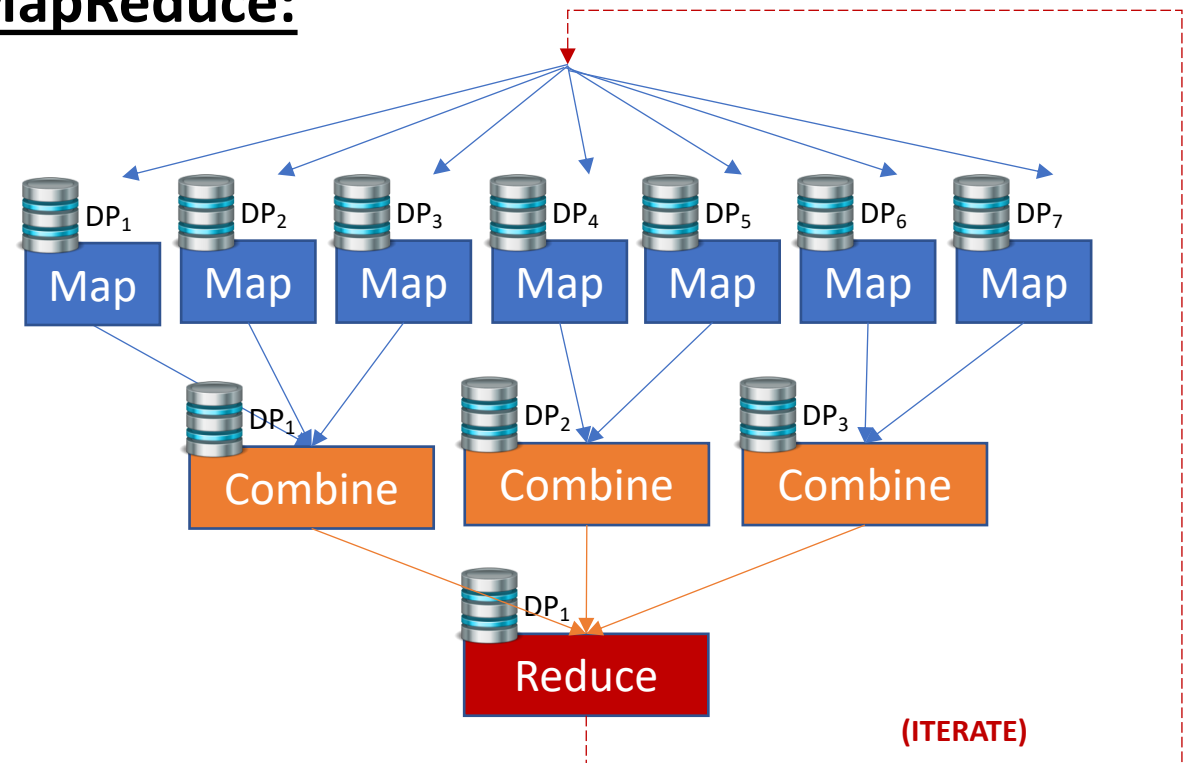
Data Confidentiality must be ensured as long as one DP is honest.



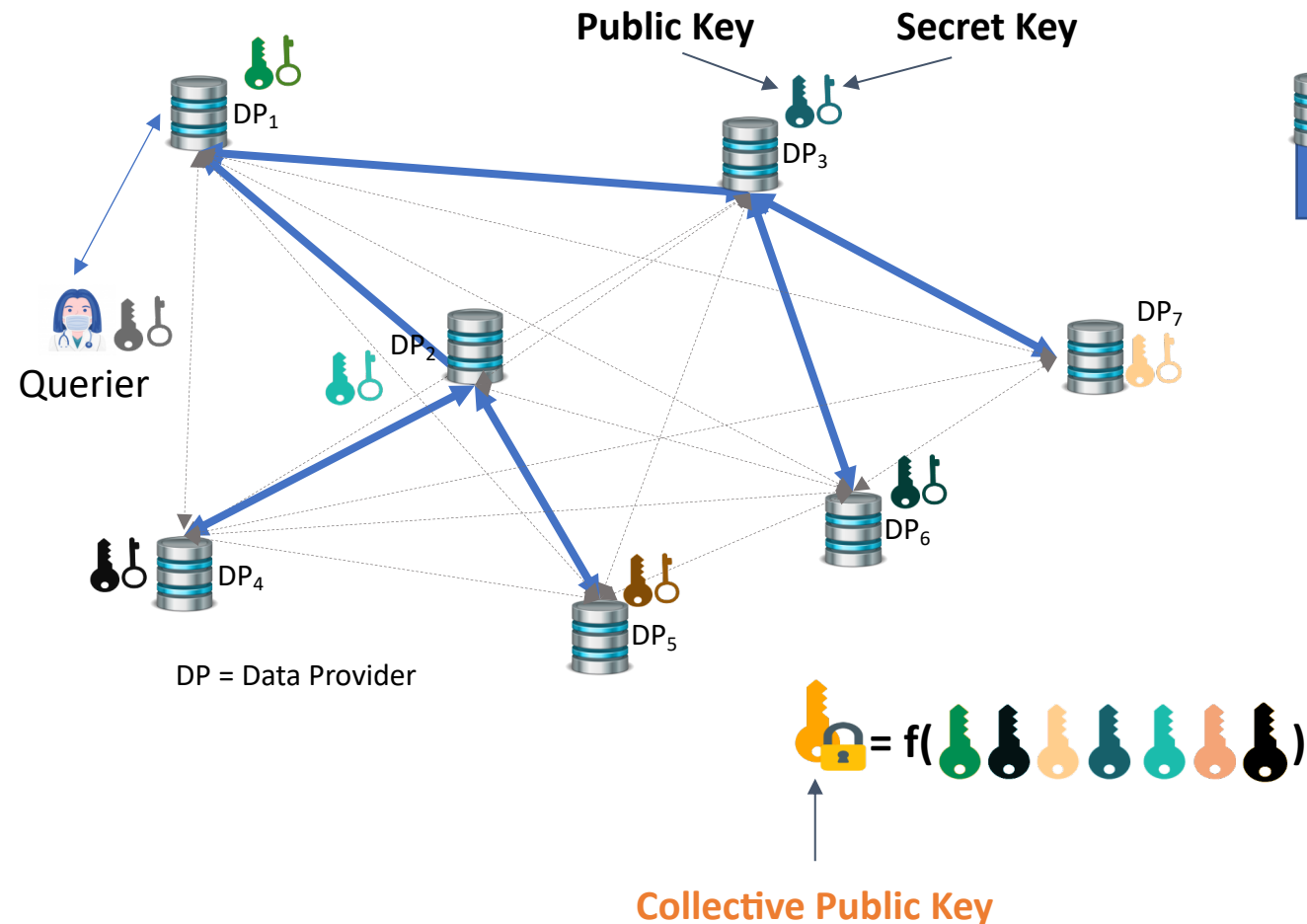
A Common Basis: Framework



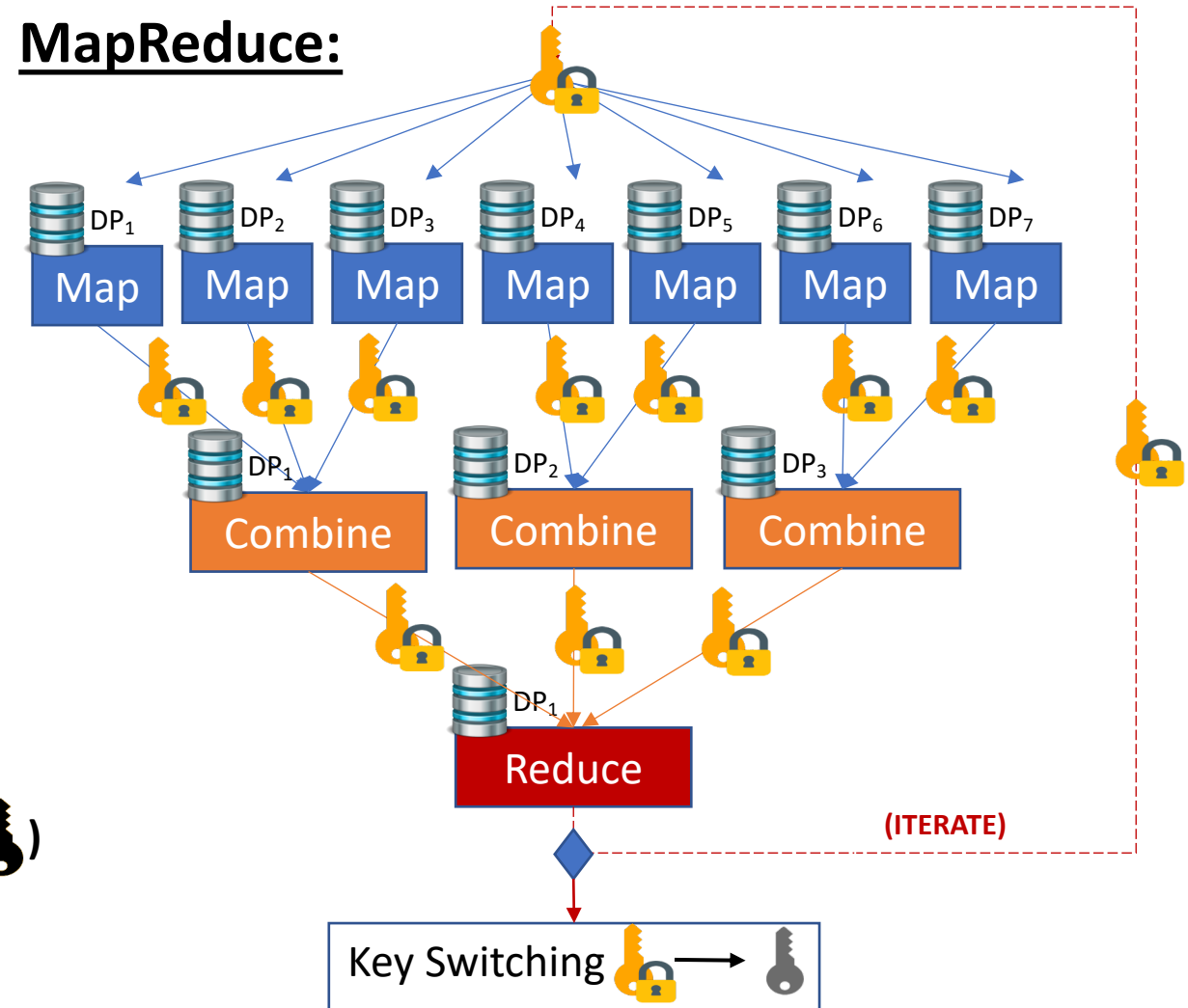
MapReduce:



A Common Basis: Protection Mechanism



MapReduce:







A Common Basis: Two Instantiations

1: Instantiation 1: Verifiable Statistics Computations

Goal: Instantiate (based on Chapters 3 & 4) our framework such that it

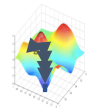

- ❖ Enables **statistical** computations
- ❖ Enables efficient and public **auditability**
- ❖ Remains secure even **against an active adversary**

Multiparty HE 	Encodings 
Proofs 	Diff. Privacy 

Instantiation 2: Machine Learning Computations

Goal: Instantiate (based on Chapter 5) our framework such that it

- ❖ Enables the cooperative execution of a **gradient descent**
- ❖ Enables an **oblivious evaluation** of the trained model
- ❖ Remains secure **against a passive adversary** controlling all but one DPs

Gradient Descent 	Multiparty HE 
---	--

1: Building Blocks

Multiparty Additive ElGamal Homomorphic Encryption

- **Security:** hardness of [discrete logarithm](#) in a finite field
- **Collective Encryption Key:** [sum](#) of the data providers' public keys



- **Encryption:** each ciphertext [encrypts one integer](#)
- **Homomorphic Operations:** [additions](#)
- **Key Switching:** collective protocol in which each data provider uses its secret key to [partially decrypt the value and re-encrypt](#) with another public key (e.g., querier)
- **Decryption:** one party (e.g., querier) uses her [secret key to decrypt](#) a message encrypted with its public key

1: Building Blocks

Publicly-verifiable zero-knowledge proofs of correctness

Correct Computations: proofs for general statements about discrete logarithms [1] and computation transcripts

Correct Input: proof of input-range [2] adapted to the multiparty scenario

Differential Privacy

A public list of noise values satisfying (ϵ, δ) -differential privacy is collectively and verifiably shuffled [3] by the data providers before the noise is added to the computation's result.

[1] J. Camenisch and M. Stadler. Proof Systems for General Statements about Discrete Logarithms. Technical Report, 1997

[2] J. Camenisch, R. Chaabouni, and a. Shelat. Efficient Protocols for Set Membership and Range Proofs. In International Conference on the Theory and Application of Cryptology and Information Security, pages 234–252. Springer, 2008

[3] C. A. Neff. Verifiable Mixing (Shuffling) of ElGamal Pairs, 2004.

1: Building Blocks

Encodings

To enable statistical computations with an additively homomorphic encryption scheme, **data providers locally encode their results** before the collective aggregation.

Example: standard deviation

Each data provider i locally computes: sum of values: r_i , sum of squared values r_i^2 and count of values c_i

Such that the **final result can be aggregated over all DPs'** values as:

$$\text{standard deviation} = \sqrt{\frac{\sum r_i^2}{\sum c_i} - \left(\frac{\sum r_i}{\sum c_i}\right)^2}$$

1: Framework Instantiation

Map

Local computations



Encoding



Encryption



Proof of range



Combine

Homomorphic Aggregation



Proofs of Correct Computations



Reduce

Homomorphic Aggregation



Proofs of Correct Computations



Key Switching

Key Switching



Proofs of Correct Computations



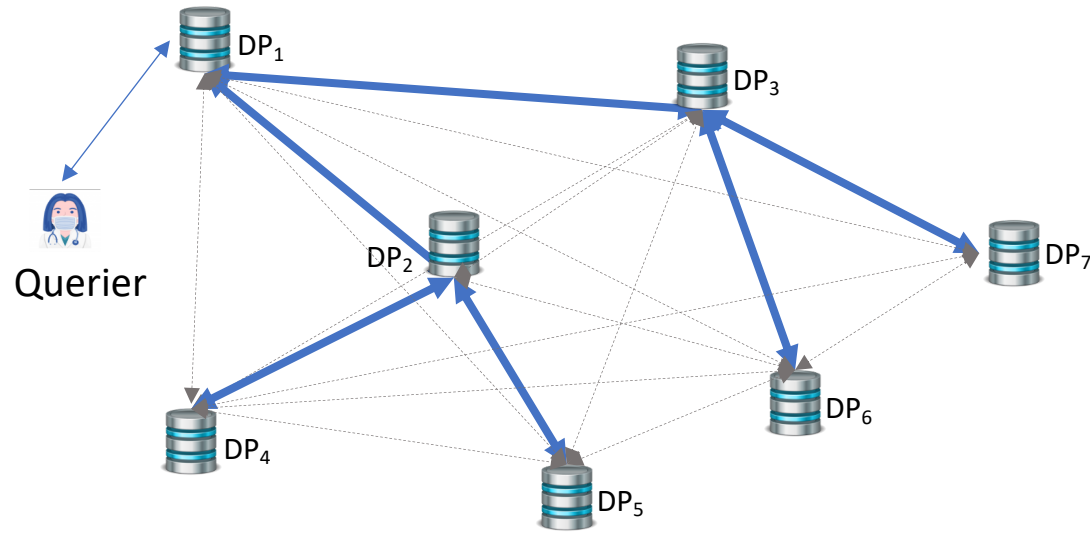
Querier's Decryption

Decryption

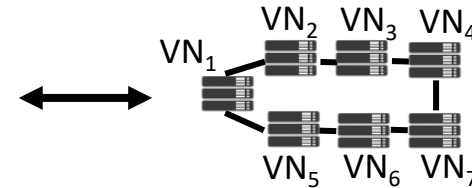


Final Computation

I1: Auditability through Verifying Nodes



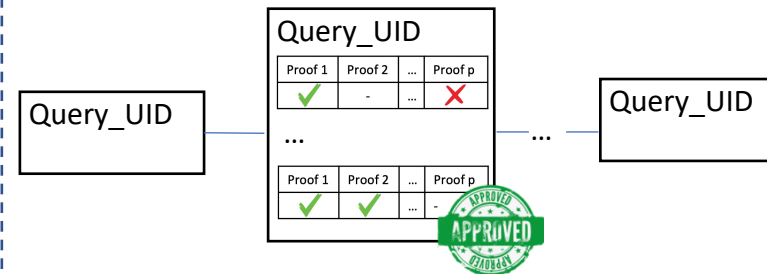
DP = Data Provider
VN = Verifying node



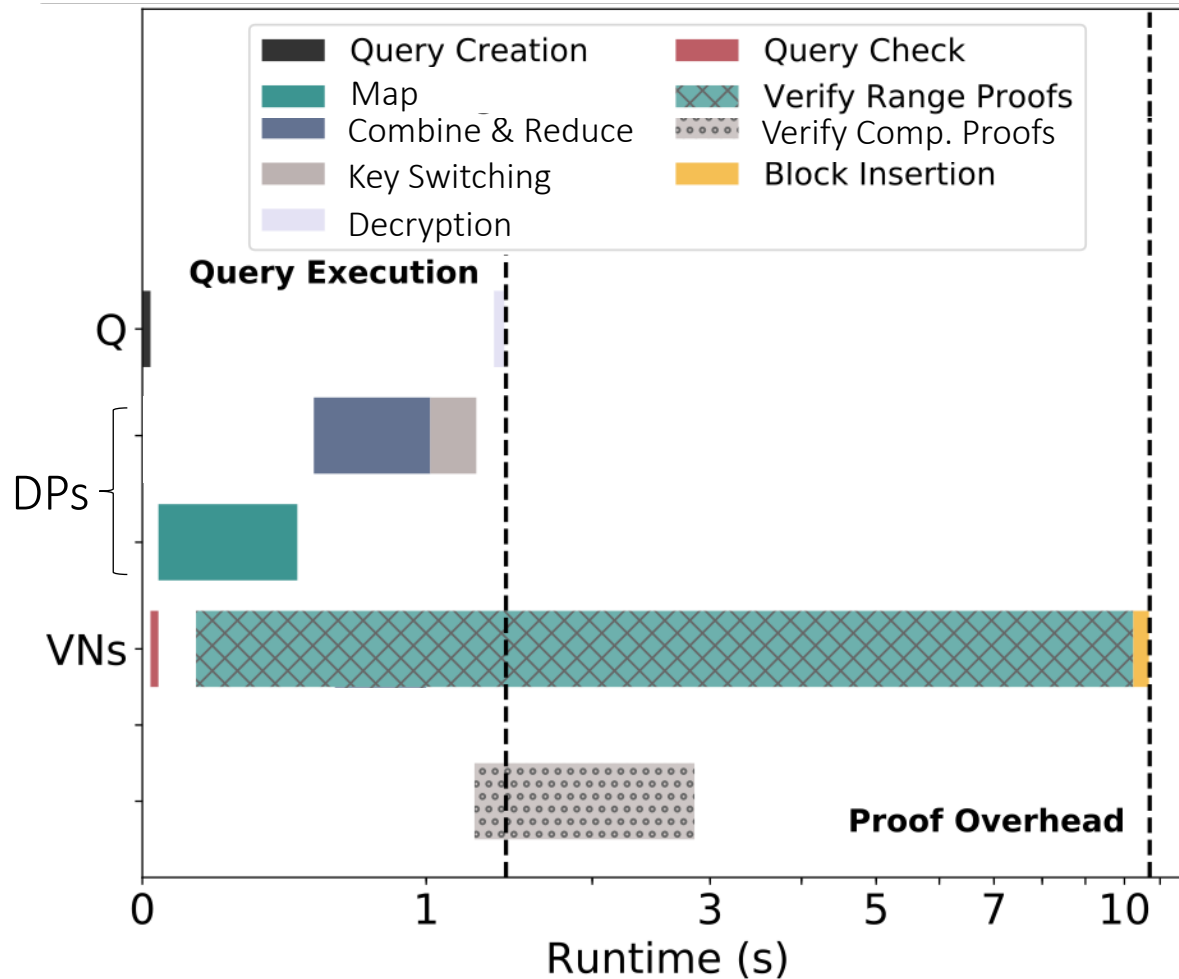
Each verifying node probabilistically verify the proofs:

Proof 1	Proof 2	...	Proof p
✓	-	...	✗

Together the verifying nodes maintain an immutable log, which can be publicly verified.



1: Evaluation



Parameters:

- 13 machines: 2 Intel Xeon E5-2680 v3 CPUs, 2.5GHz frequency, 24 threads on 12 cores, 256GB RAM.
- Operation: Variance of 30 different attributes
- 6000 records split among 60 *DPs*
- Input range of $[0, 2^{20}]$
- 7 *VNs*

Scalability:

- Linear with the number of data providers and with the *DPs'* datasets

A Common Basis: Two Instantiations

I1: Instantiation 1: Verifiable Statistics Computations

Goal: Instantiate (based on Chapters 3 & 4) our framework such that it

- ❖ Enables **statistical** computations
- ❖ Enables an efficient and public **auditability**
- ❖ Remains secure even **against an active adversary**

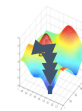
- **Limited by the encryption scheme**
- **Limited scalability with the number of features**
- **Not quantum resistant**

Instantiation 2: Machine Learning Computations

Goal: Instantiate (based on Chapter 5) our framework such that it

- ❖ Enables the cooperative execution of a **gradient descent**
- ❖ Enables an **oblivious evaluation** of the trained model
- ❖ Remains secure **against a passive adversary** controlling all but one DPs

Gradient Descent



Multiparty HE



12: Building Blocks

Multiparty Homomorphic Encryption

adaptation to CKKS [1] of the multiparty scheme proposed by Mouchet et al. [2]

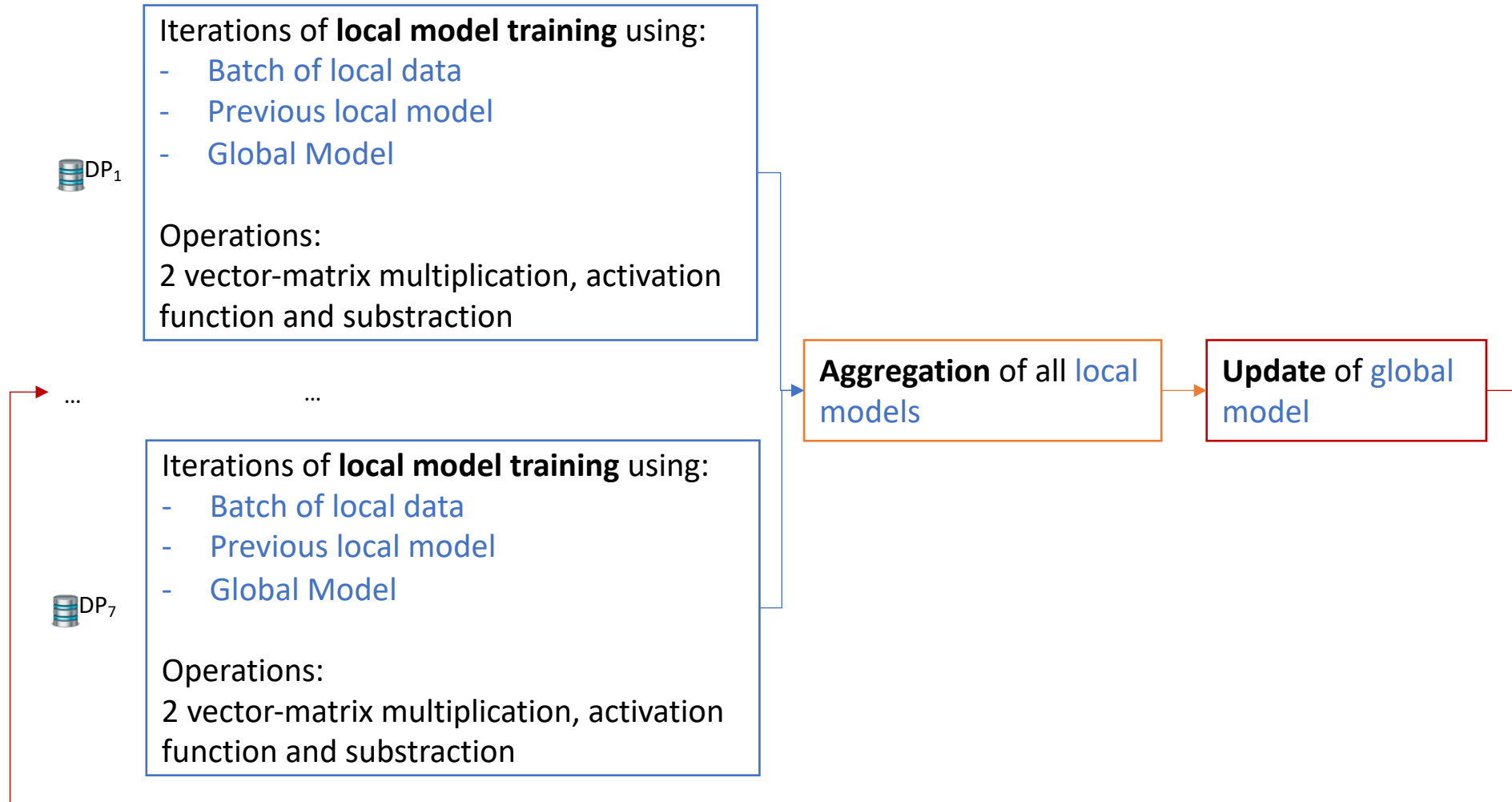
- **Security:** hardness of the [ring learning with errors \(RLWE\)](#) problem
- **Collective Encryption Key:** created in an [interactive protocol](#) among all data providers
- **Encryption:** each ciphertext contains [a vector of \$N\$ values](#): $E_{\text{pk}}(v_1, \dots, v_N)$
- **Homomorphic Operations:** additions, [multiplications](#), rotations
- **Collective Bootstrapping:** interactive protocol to [refresh a ciphertext](#), required after a certain number of operations
- **Key Switching:** collective protocol in which each data provider uses its secret key to [partially decrypt the value and re-encrypt](#) with the querier's public key
- **Decryption:** one party (e.g., querier) uses her [secret key to decrypt](#) a message encrypted with its public key

[1] J. H. Cheon, A. Kim, M. Kim, and Y. Song. Homomorphic encryption for arithmetic of approximate numbers. In ASIACRYPT, 2017.

[2] C. Mouchet, J. R. Troncoso-pastoriza, J.-P. Bossuat, and J. P. Hubaux. Multiparty Homomorphic Encryption from Ring-Learning-With-Errors. In PETS'21.

I2: Building Blocks

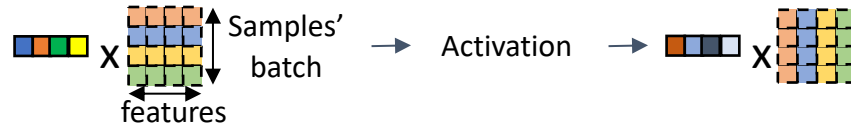
Cooperative Gradient Descent



I2: Framework Instantiation

Map

Stochastic Gradient Descent Operations:

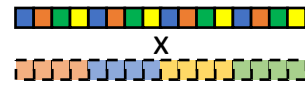


□ Encrypted
□ Cleartext

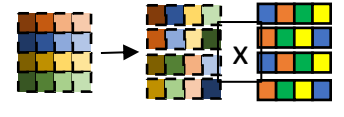
X

$f(\text{features, samples' batch})$
Input Dimension

Row-based Approach



Diagonal Approach

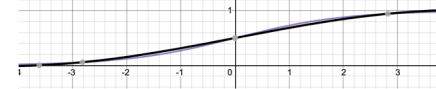


Activation

$$\frac{1}{1+e^{-\theta^T x}}$$



Least Square Approx.



Combine

Homomorphic Aggregation

Reduce

Homomorphic Aggregation

Key Switching

Key Switching

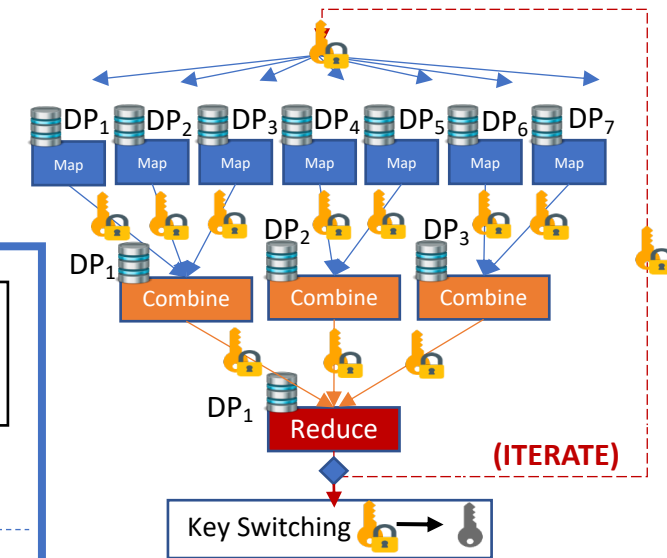


Querier's Decryption

Decryption

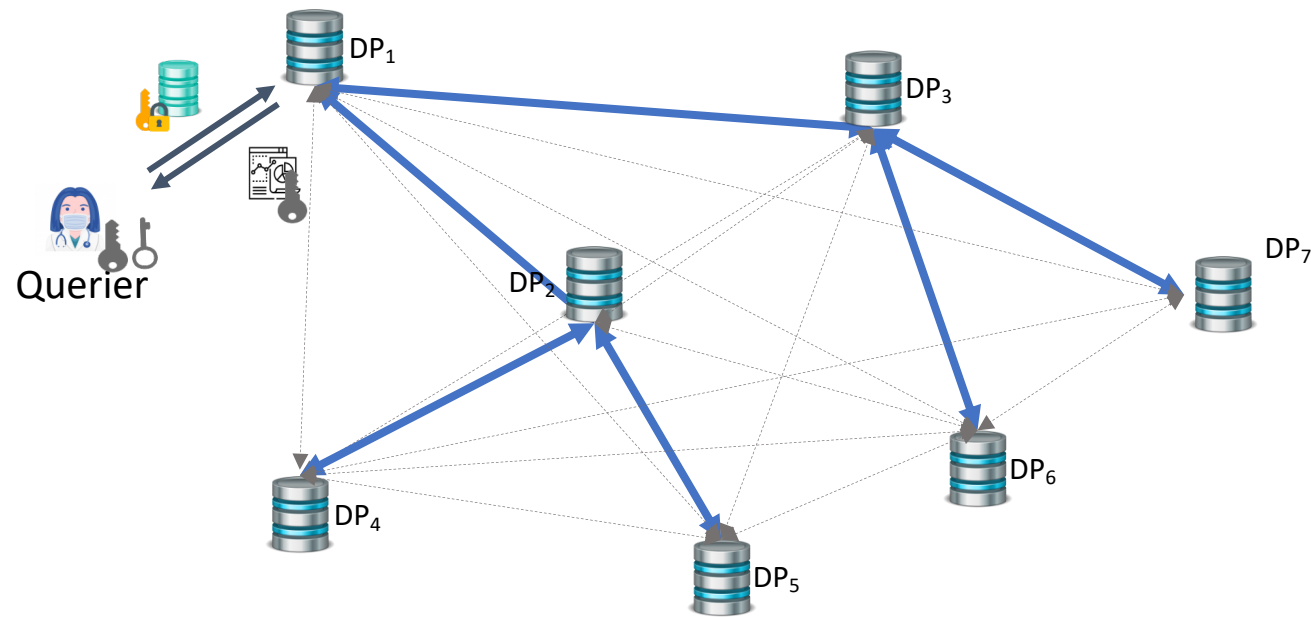


Final Computation



I2: Framework Instantiation

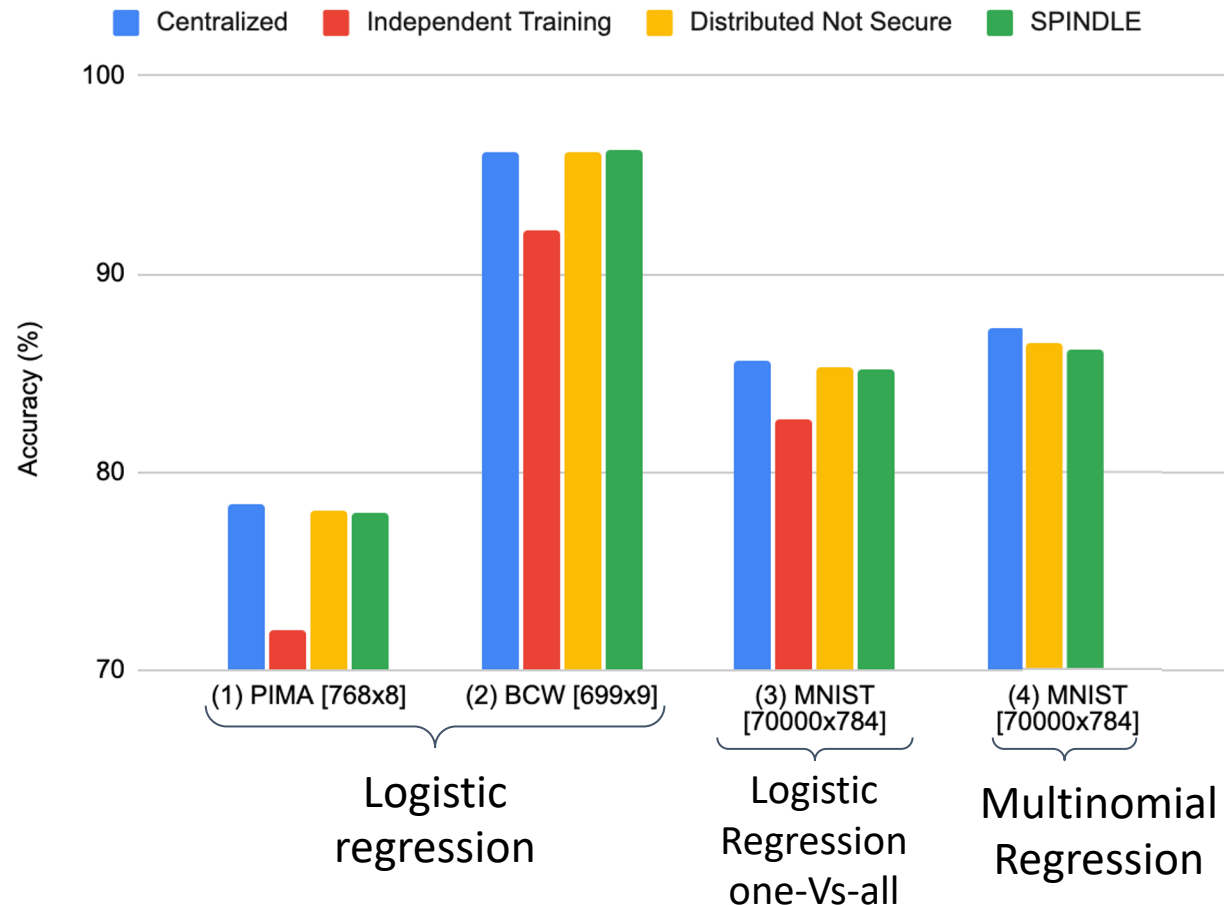
To cover the **entire ML workflow**, the trained model can remain collectively encrypted and be used for **oblivious evaluation**



I2: 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)

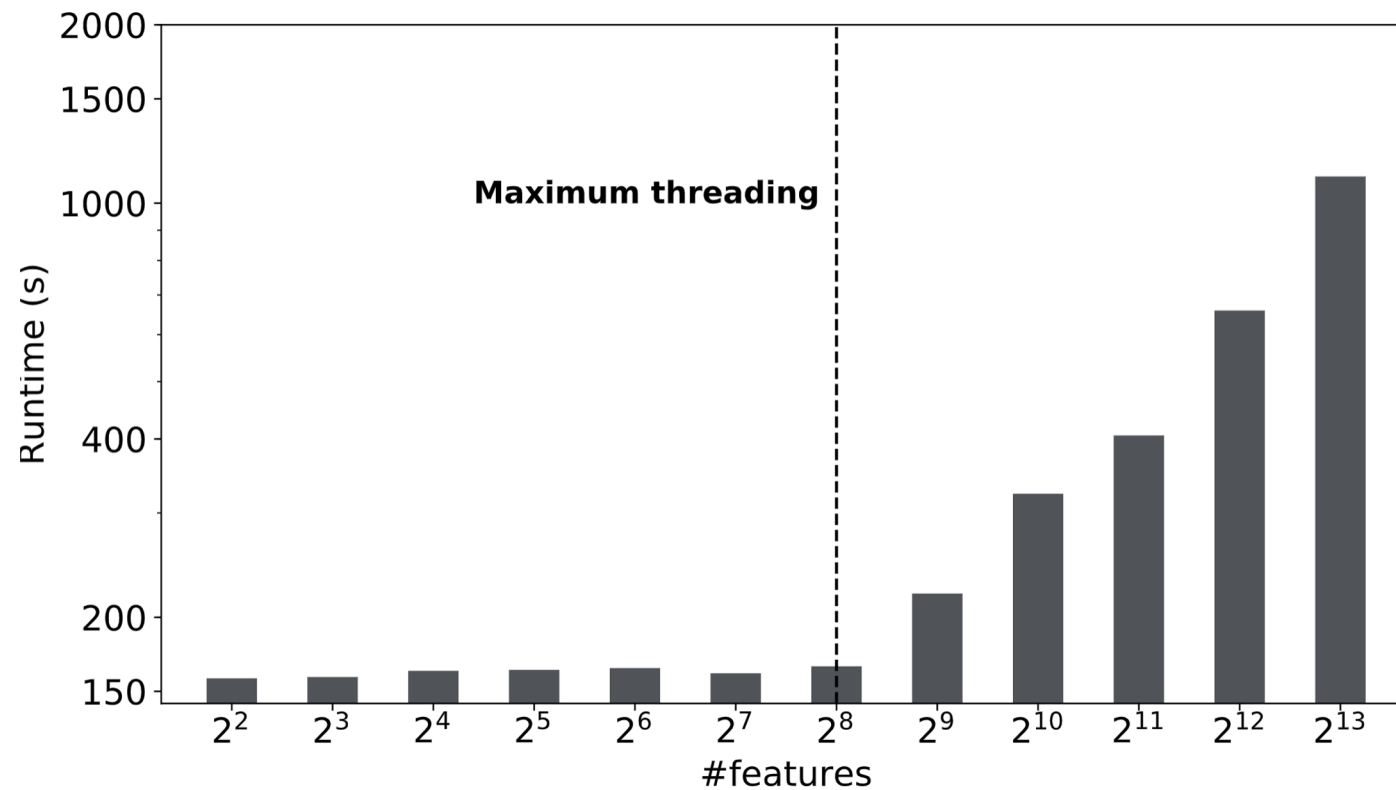
[https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+\(original\)](https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original))

(3,4) MNIST

Y. LeCun and C. Cortes. Handwritten digit database. 2010.

I2: Performance Evaluation

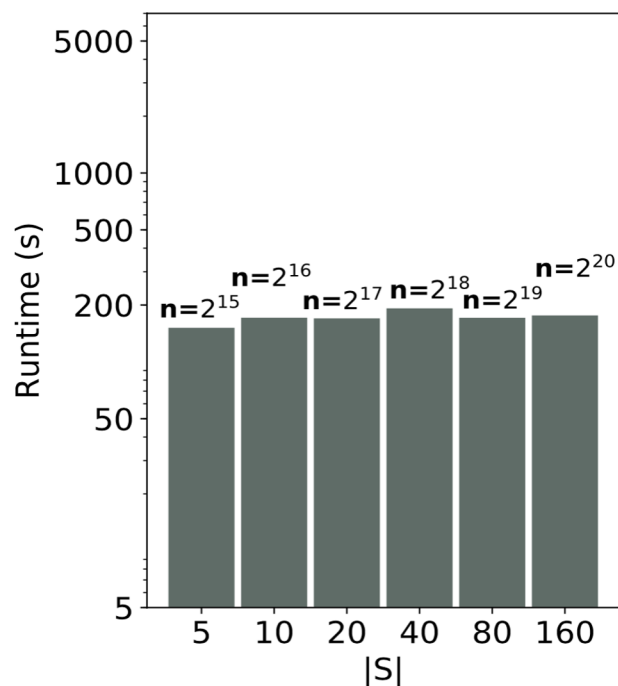
Better than logarithmic increase with the number of features



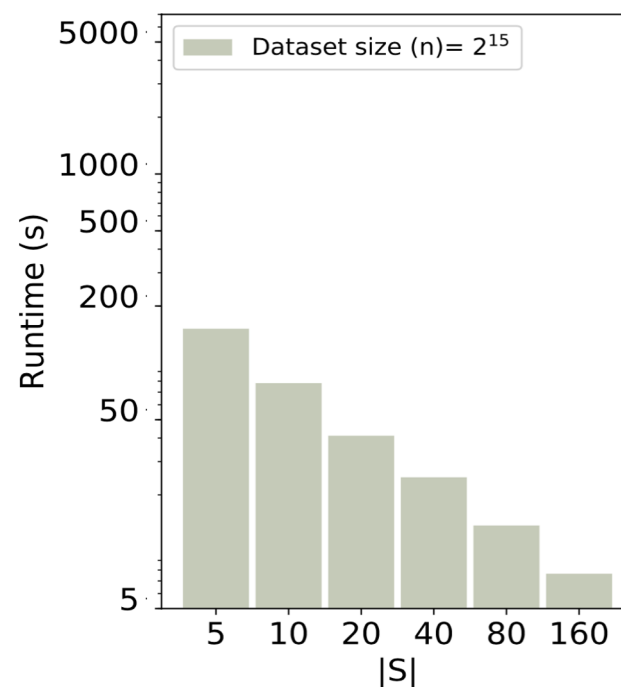
5 data providers, 25600 record, 128-bit security; Each data provider: 2 Intel Xeon E5-2680 v3 CPUs, 2.5GHz frequency, 24 threads on 12 cores, 256GB RAM. Communication: 100Mbps, delay 20ms

I2: Performance Evaluation

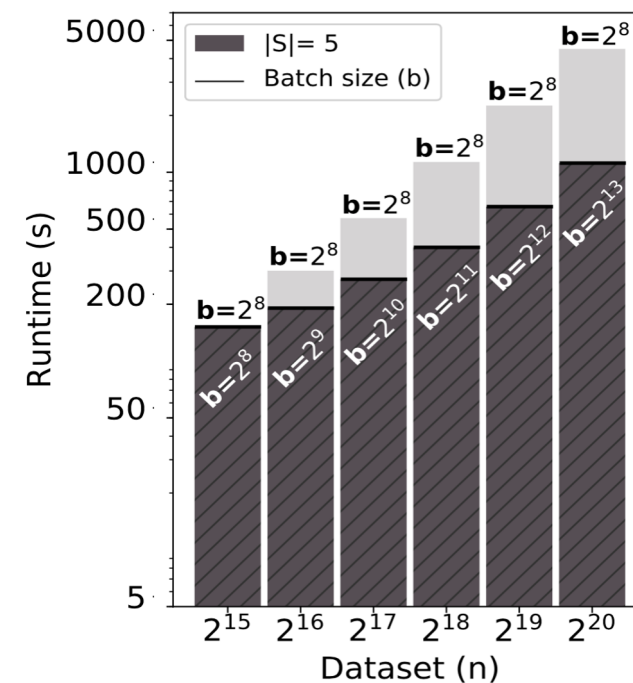
Scales almost independently with the number of data providers $|S|$



Efficient workload distribution



Scales linearly with the size of the data providers datasets n



128-bit security level; Default number of features = 32, $|S|$ = # data providers; n = global dataset size, b is the batch size used in the stochastic gradient descent, One data provider: 2 Intel Xeon E5-2680 v3 CPUs, 2.5GHz frequency, 24 threads on 12 cores, 256GB RAM. Communication: 100Mbps, delay 20ms

Our Framework for Practical Use Cases

Goal: Reproduce in our framework (based on Chapter 6) two biomedical studies that originally relied on data centralisation.

→ demonstrate that we obtain **accurate results** while keeping the data **decentralized and protecting patients' privacy**

Study 1: Survival Curve:

Samstein et al. [1] computed Kaplan-Meier overall survival curves on 1662 advanced-cancer patients to study the effect of a specific treatment.

Tumor mutational load predicts survival after immunotherapy across multiple cancer types

Robert M Samstein^{1 2}, Chung-Han Lee^{3 4}, Alexander N Shoushtari^{3 4},
Matthew D Hellmann^{3 4}, Ronglai Shen⁵, Yelena Y Janjigian^{3 4}, David A Barron^{1 2},
Ahmet Zehir⁶, Emmet J Jordan², Antonio Omuro⁷, Thomas J Kaley⁷,
Sviatoslav M Kendall^{2 8}, Robert J Motzer^{3 4}, A Ari Hakimi⁹, Martin H Voss^{3 4},
Paul Russo⁹, Jonathan Rosenberg^{3 4}, Gopa Iyer^{3 4}, Bernard H Bochner⁹,
Dean F Bajorin^{3 4}, Hikmat A Al-Ahmadie⁶, Jamie E Chaft^{3 4}, Charles M Rudin^{3 4},
Gregory J Riely^{3 4}, Shrujal Baxi^{3 4}, Alan L Ho^{3 4}, Richard J Wong⁹, David G Pfister^{3 4},
Jedd D Wolchok^{3 4}, Christopher A Barker¹, Philip H Gutin⁹, Cameron W Brennan⁹

Study 2: Genome-Wide Association Study:

McLaren et al. [2] studied the link between HIV viral load and specific genome variants.

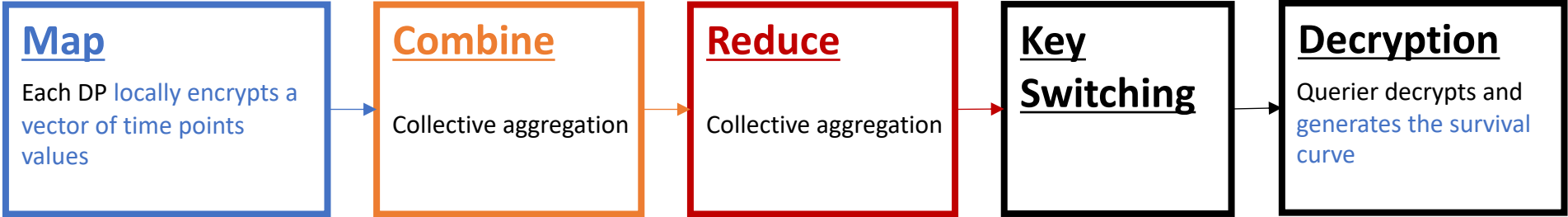
Polymorphisms of large effect explain the majority of the host genetic contribution to variation of HIV-1 virus load

Paul J McLaren¹, Cedric Coulonges², István Bartha¹, Tobias L Lenz³, Aaron J Deutsch⁴,
Arman Bashirova⁵, Susan Buchbinder⁶, Mary N Carrington⁷, Andrea Cossarizza⁸,
Judith Dalmau⁹, Andrea De Luca¹⁰, James J Goedert¹¹, Deepti Gurdasani¹²,
David W Haas¹³, Joshua T Herbeck¹⁴, Eric O Johnson¹⁵, Gregory D Kirk¹⁶,
Olivier Lambotte¹⁷, Ma Luo¹⁸, Simon Mallal¹⁹, Daniëlle van Manen²⁰,
Javier Martinez-Picado²¹, Laurence Meyer²², José M Miro²³, James I Mullins²⁴

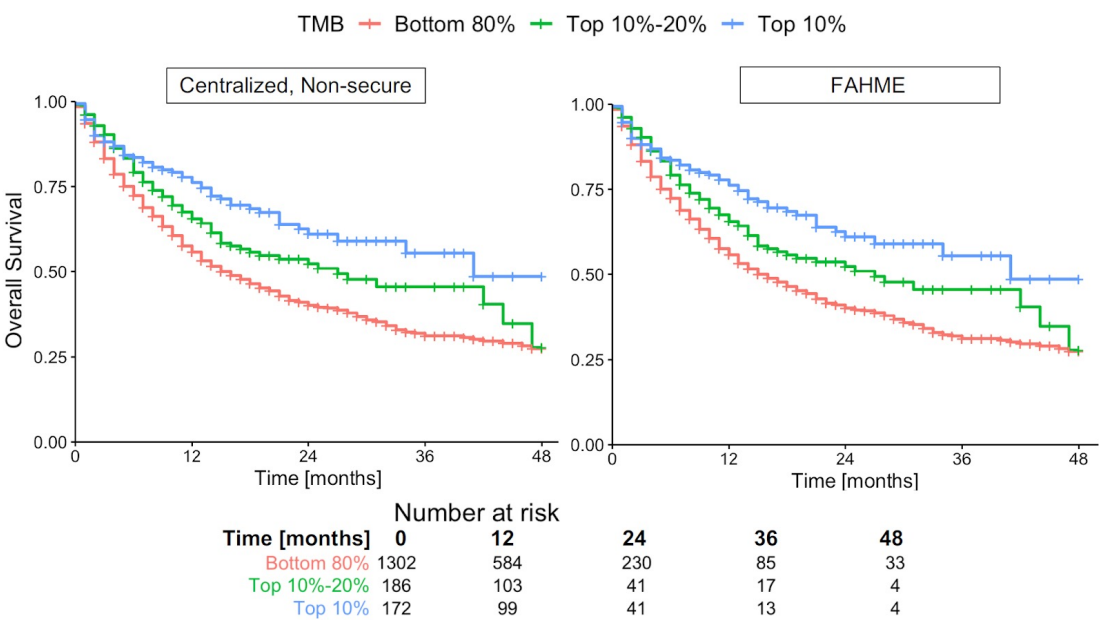
[1] R. M. Samstein et al. Tumor Mutational Load Predicts Survival after Immunotherapy across Multiple Cancer Types. Nature Genetics, 2019.

[2] P. J. McLaren et al. Polymorphisms of Large Effect Explain the Majority of the Host Genetic Contribution to Variation of HIV-1 Virus Load. Proceedings of the National Academy of Sciences of the United States of America, 2015.

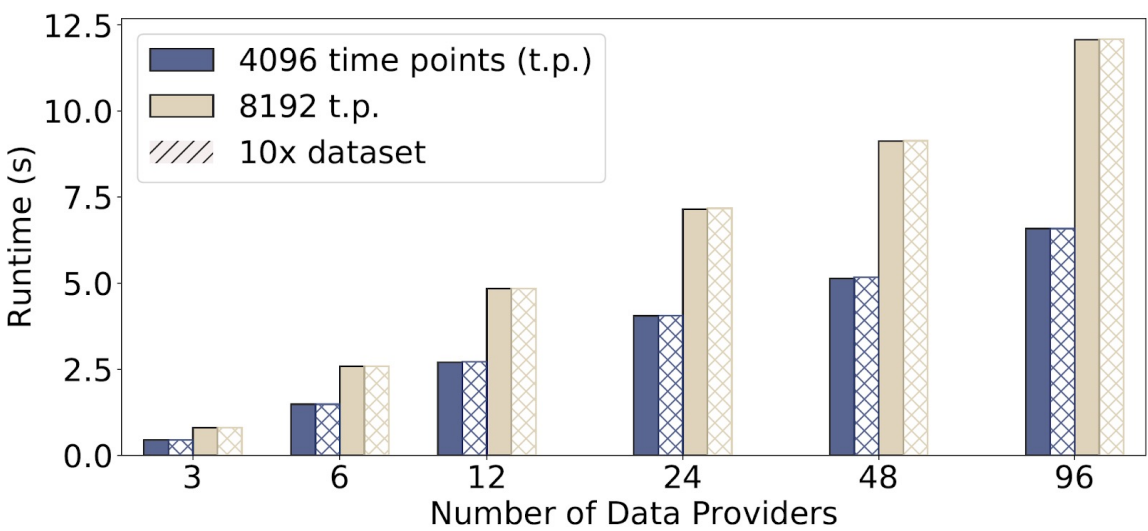
Study 1: Survival Curve



Exact Results

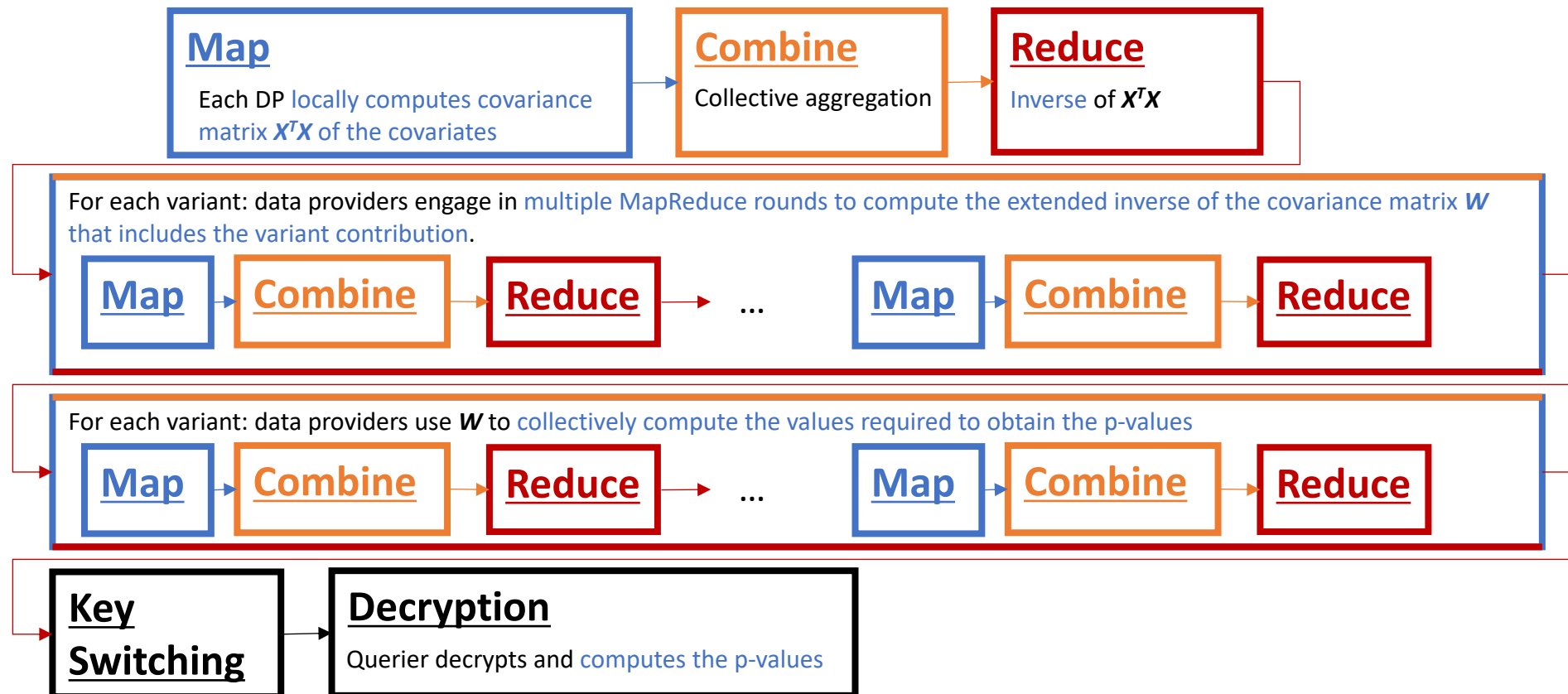


Scale with the number of data providers and with the dataset's size



Study 2: Genome-Wide Association Study

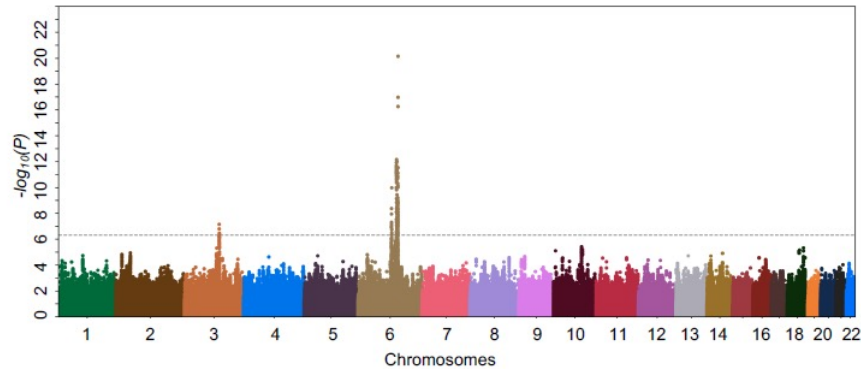
GWAS Method: The **p-value** is computed from the **variant weight**, the **mean-squared error** and the **standard error of the variant's weight**. A small p-value indicates a link.



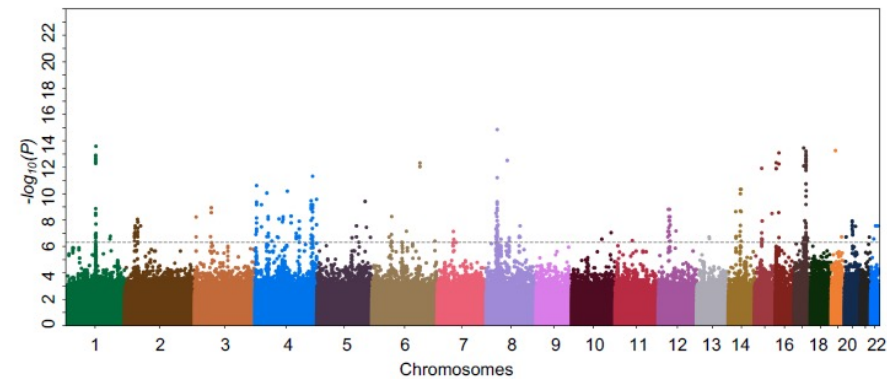
Study 2: Genome-Wide Association Study

Setting: 1857 patients spread among 12 data providers.

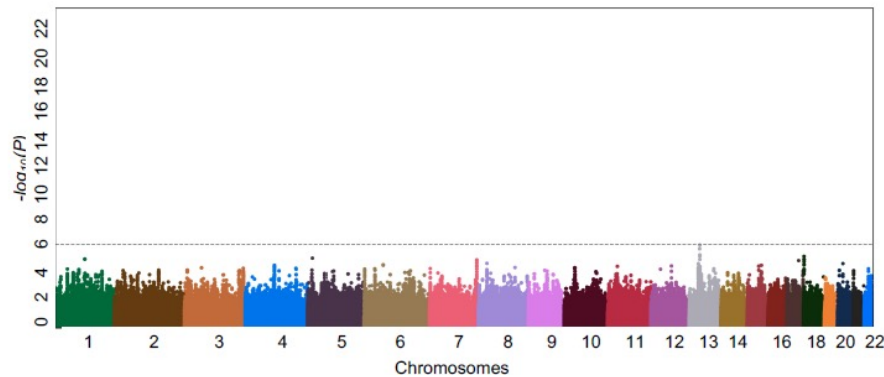
Original Study (Centralized, Non-Secure)



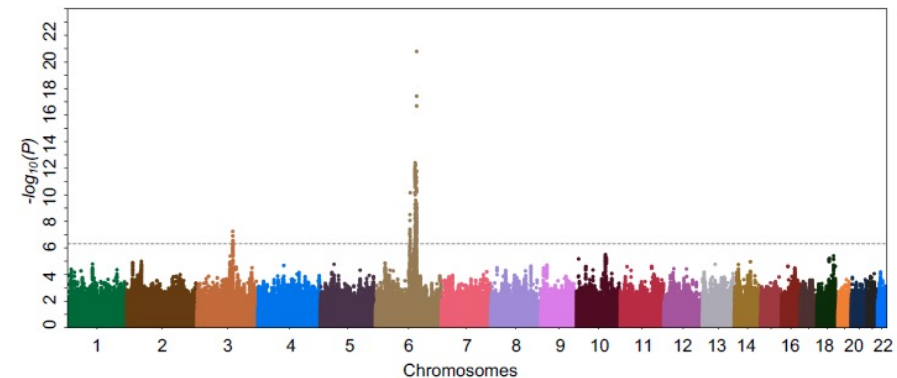
Meta-Analysis (Decentralized, Non-Secure)



Independent (1 single data provider)



Our Approach (Decentralized, Secure)



Almost Exact Results

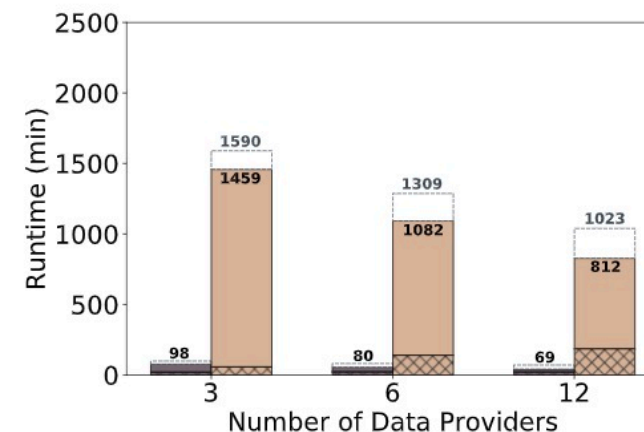
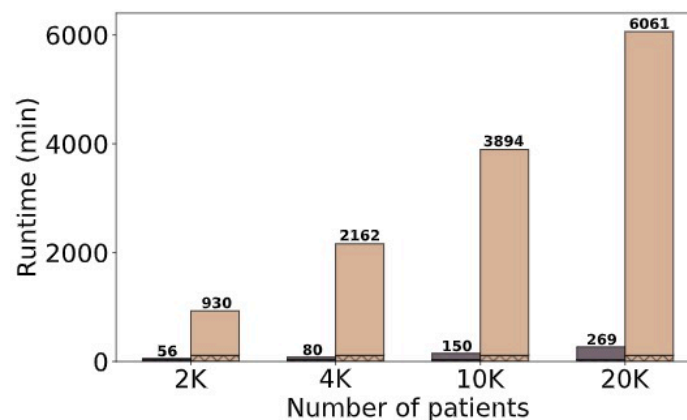
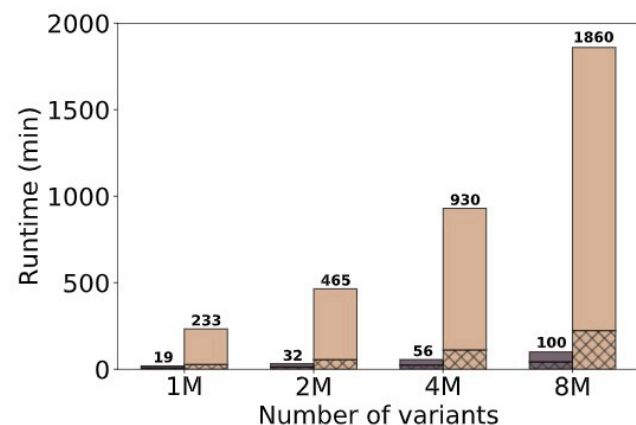
Study 2: Genome-Wide Association Study

Default Setting: 1857 patients spread among 12 data providers.

Performance-optimized Approach: instead of computing the complete inverse of covariance

matrix, we estimate the covariates weights through an efficient gradient descent and only compute the standard error of the variant weight, i.e., one diagonal element.

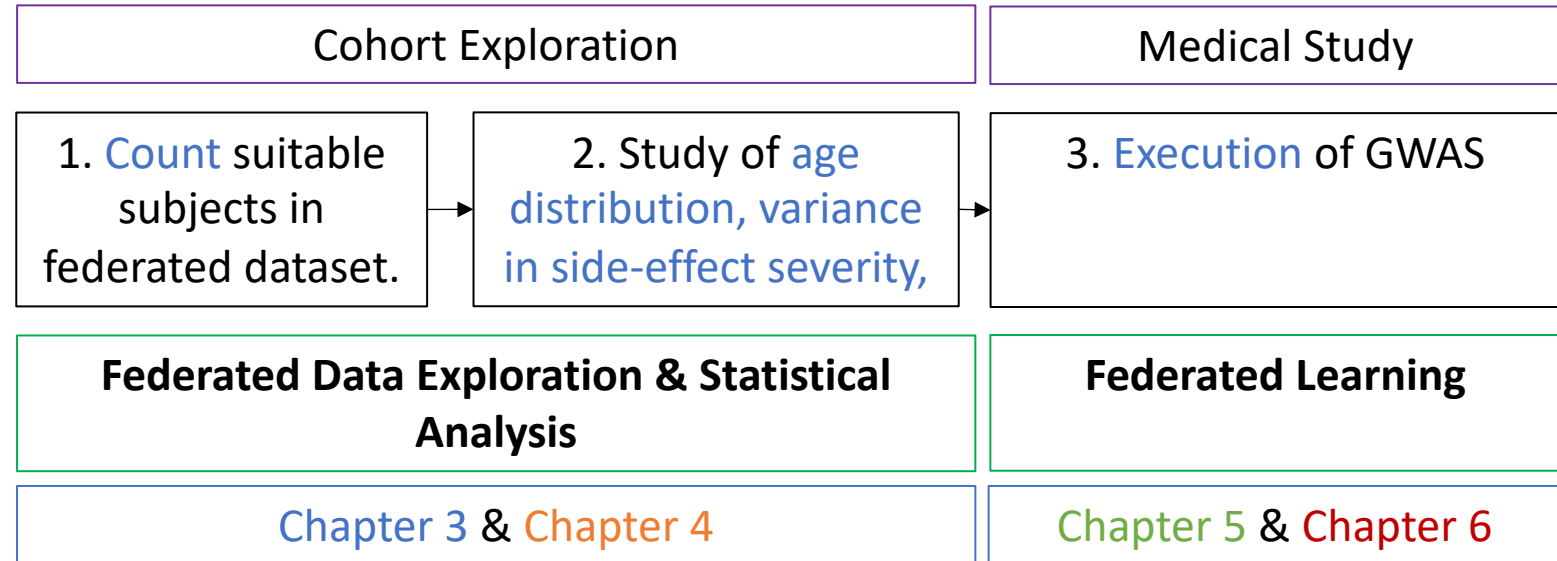
Exact Approach Time dedicated to communication **Overhead in WAN setting**



One data provider: 2 Intel Xeon E5-2680 v3 CPUs, 2.5GHz frequency, 24 threads on 12 cores, 256GB RAM.
Communication: 1Gbps, delay 20ms

Federated Analytics in the Medical Domain

Hypothetical Example: Study link between COVID-vaccine severe side-effect and specific variants.



Impact



Cohort Exploration Tool (based on Chapter 3)

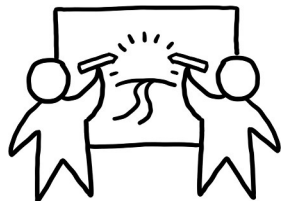
- Deployed Network between multiple hospitals and universities in Switzerland
- Currently being deployed in Netherlands, Italy, USA, ...



Startup building a tool for secure collaboration and federated analytics



2 patents filed based on our work in Chapter 5 and subsequent work



DPPH: Data Protection in Personalized Health funded by the Strategic Focus Area Personalized Health and Related Technologies (PHRT) of the ETH Board.

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

Conclusion

UnLynx (Chapter 3) & Drynx (Chapter 4)



SELECT **sum/count/statistic()** ... FROM
DP₁, ..., DP_n WHERE ... GROUP BY ...



Data privacy & Confidentiality
Computations & **Input Verification**
Collective protection of local data



Anytrust Model with
passive/active
adversary

- Encryption scheme less computationally powerful but enabling the **use of well-established zero-knowledge proofs**
- Rely on **edge-computing to enable flexible operations**
- **Modular security properties**, efficient proof creation & verification with low-effect on query response time
- **Linear scaling** with the number of samples & with the number of data providers

SPINDLE (Chapter 5) & FAMHE (Chapter 6)



Cooperative gradient descent (training)
and model evaluation, **and federated**
biomedical studies



Data & Model Confidentiality



Anytrust Model with
passive adversary

- State-of-the-art encryption scheme enabling **flexible operations and packing**
- Rely on edge-computing to **optimize the balance between operations on encrypted data and on cleartext data**
- Logarithmic scaling with the number of features, linear with the dataset size and **almost independent of the number of data providers**
- **Accurate reproduction of existing results**

Future Directions

- **More Complex Operations:**
 - Training of neural networks
 - Federated principal component analysis
- **Federated parametrization on distributed datasets**
- **Computation Flexibility**
 - Example of survival Curve vs. Genome-Wide Association Study
- **Combination of multiparty homomorphic encryption with other techniques**

Future Directions

- **Adoption of secure solutions**

- Solutions with similar properties as existing non-secure solutions (scaling, tradeoff between accuracy and performance)
- Limited overhead brought by security mechanisms

- **Multidisciplinary challenge**

- Integration in existing tools
- Parametrization, cryptographic \leftrightarrow domain specialised
- Data harmonization



Conclusion

UnLynx (Chapter 3) & Drynx (Chapter 4)



SELECT **sum/count/statistic()** ... FROM
DP₁, ..., DP_n WHERE ... GROUP BY ...



Data privacy & Confidentiality
Computations & **Input Verification**
Collective protection of local data



Anytrust Model with
passive/active
adversary

- Encryption scheme less computationally powerful but enabling the **use of well-established zero-knowledge proofs**
- Rely on **edge-computing to enable flexible operations**
- **Modular security properties**, efficient proof creation & verification with low-effect on query response time
- **Linear scaling** with the number of samples & with the number of data providers

SPINDLE (Chapter 5) & FAMHE (Chapter 6)



Cooperative gradient descent (training)
and model evaluation, **and federated**
biomedical studies



Data & Model Confidentiality



Anytrust Model with
passive adversary

- State-of-the-art encryption scheme enabling **flexible operations and packing**
- Rely on edge-computing to **optimize the balance between operations on encrypted data and on cleartext data**
- Logarithmic scaling with the number of features, linear with the dataset size and **almost independent of the number of data providers**
- **Accurate reproduction of existing results**