Doctoral Program in Informatics

# Towards a Privacy-Preserving Distributed Machine Learning Framework

Cláudia Brito, 2024

Under the supervision of
João Tiago Paulo
Pedro Gabriel Ferreira

INESCTEC

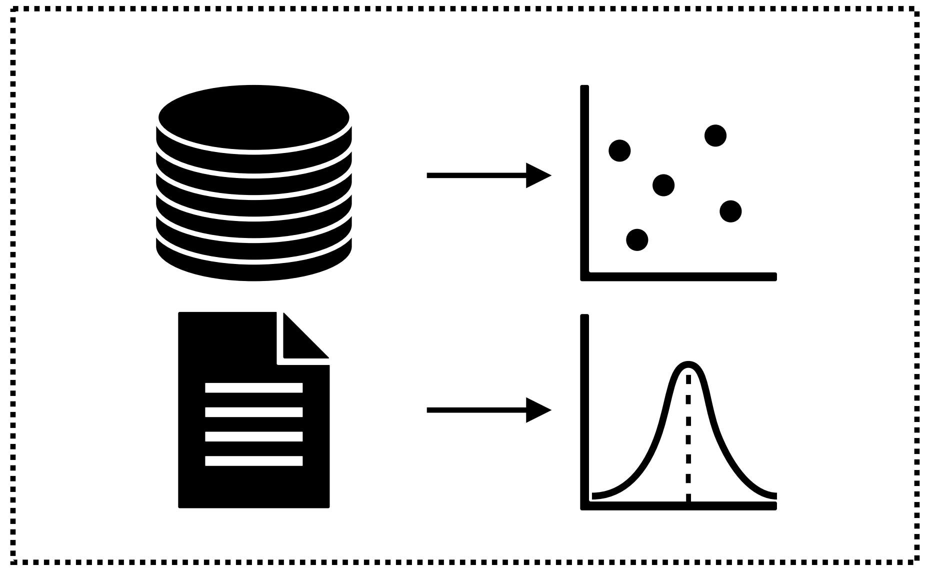
# Privacy and Security in Machine Learning

## Motivation

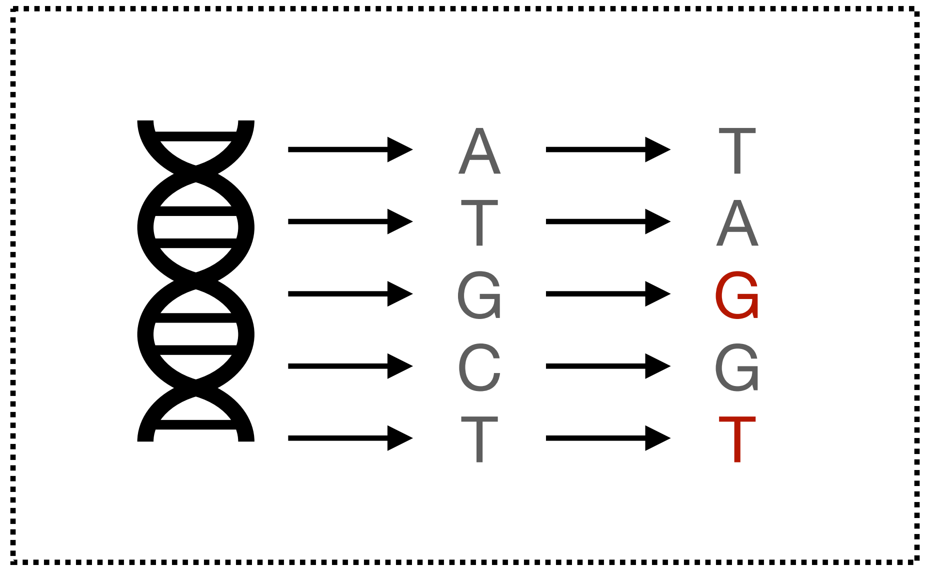
# Privacy and Security in Machine Learning

## Motivation

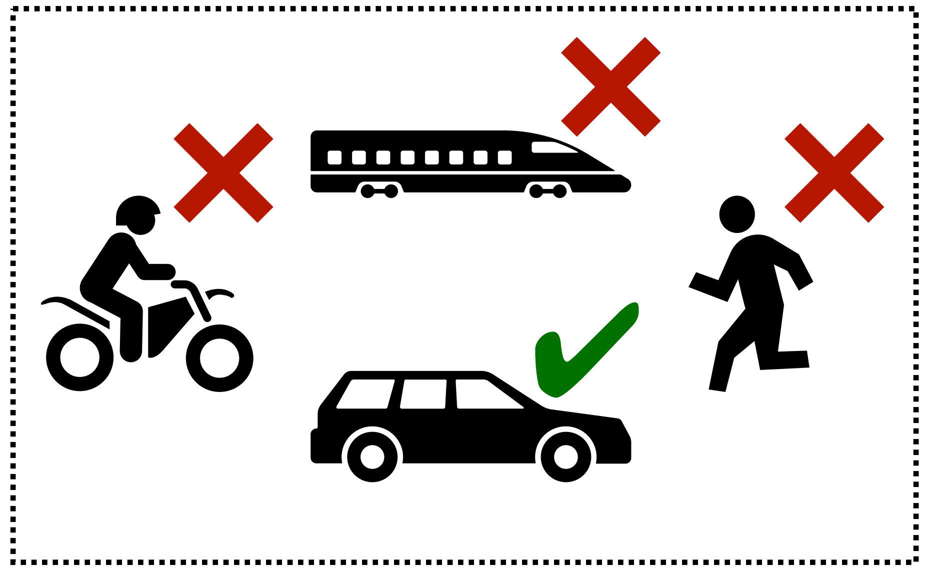
- Machine learning is growing in terms of applicability and complexity (i.e., models and datasets).



Large Datasets




Genomic Analysis

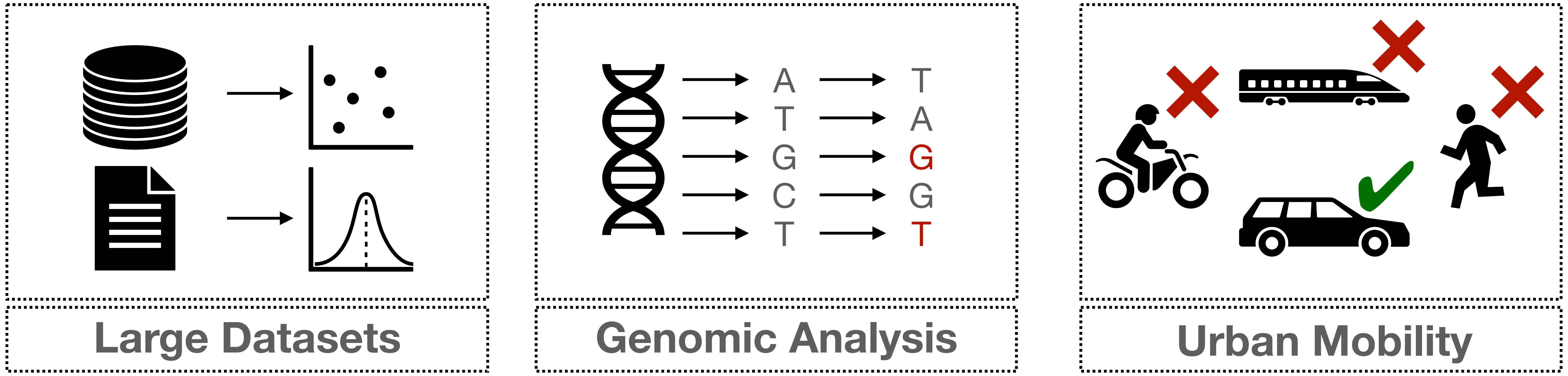


Urban Mobility

# Privacy and Security in Machine Learning

## Motivation

- Machine learning is growing in terms of applicability and complexity (i.e., models and datasets).




- Increasing the need to outsource the computation and storage to untrusted third-parties.

- Current regulations were built to protect user's privacy.

# Privacy and Security in Machine Learning

## Motivation

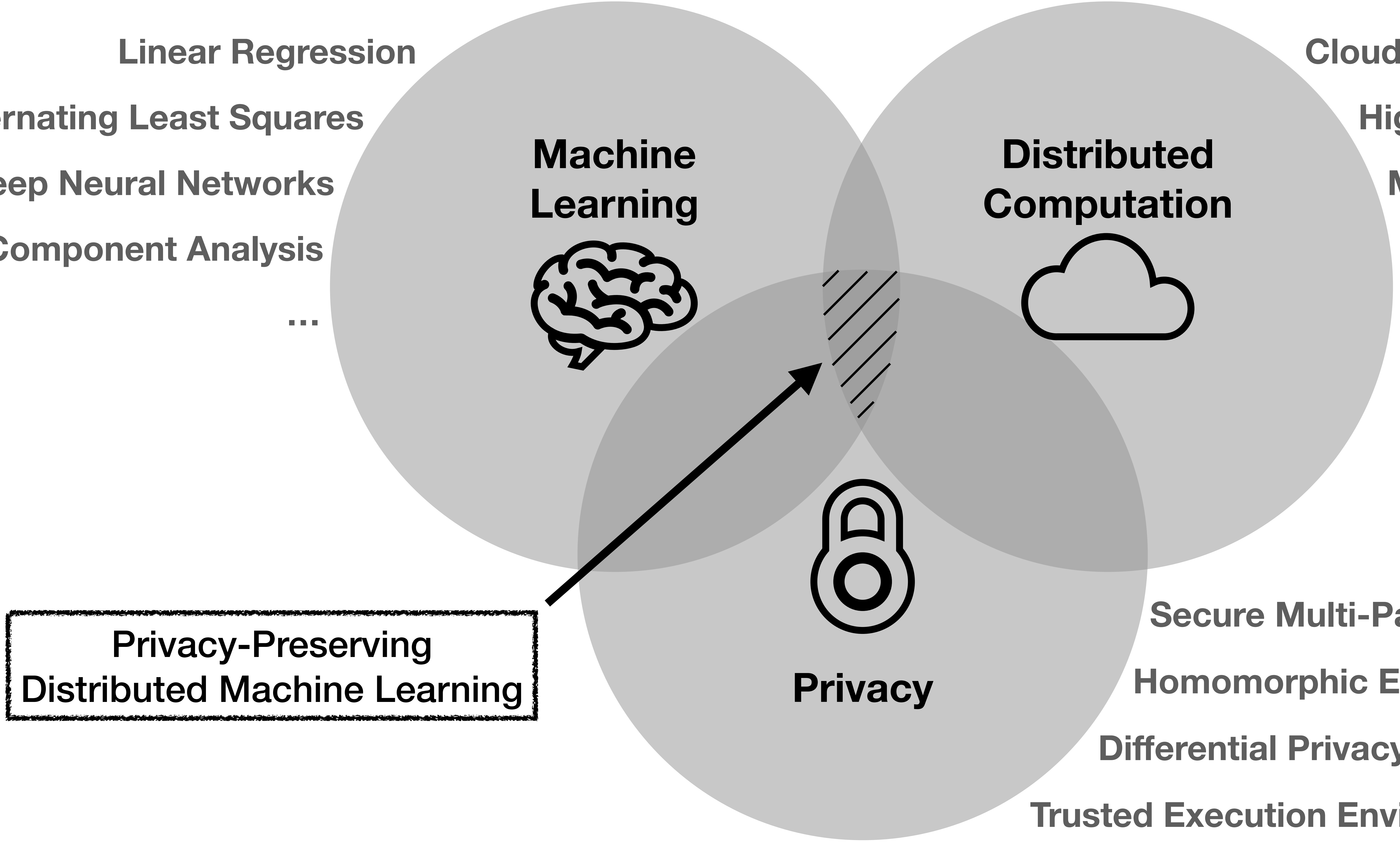
Linear Regression

Alternating Least Squares

Deep Neural Networks

Principal Component Analysis

...




Cloud Computing

High-Performance Computing

Mobile/Edge Devices

Secure Multi-Party Computation

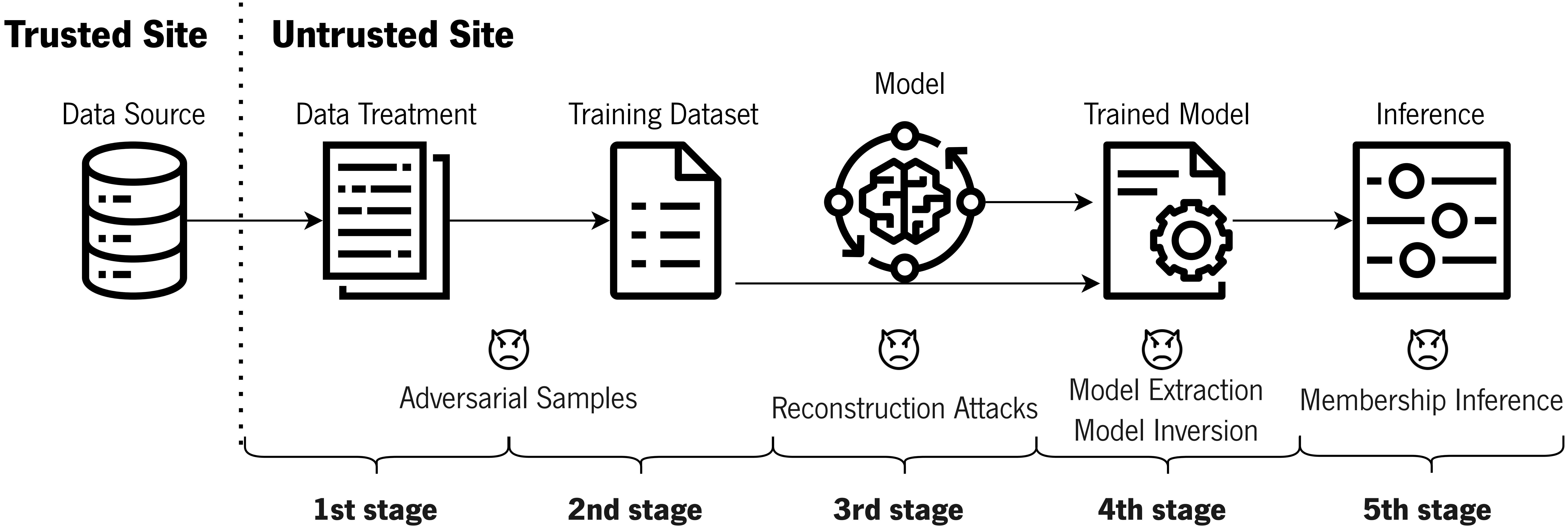
Homomorphic Encryption

Differential Privacy

Trusted Execution Environments

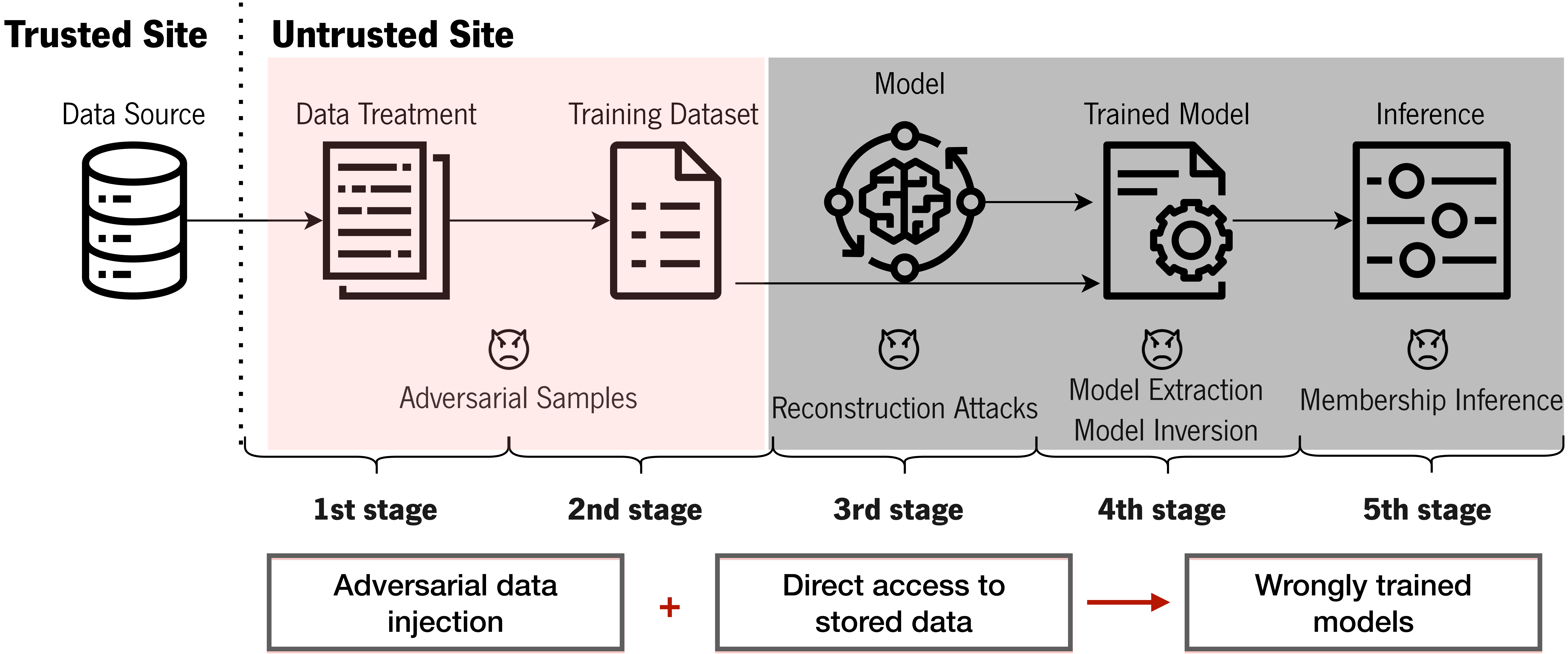
# Privacy and Security in Machine Learning

## ML Pipeline



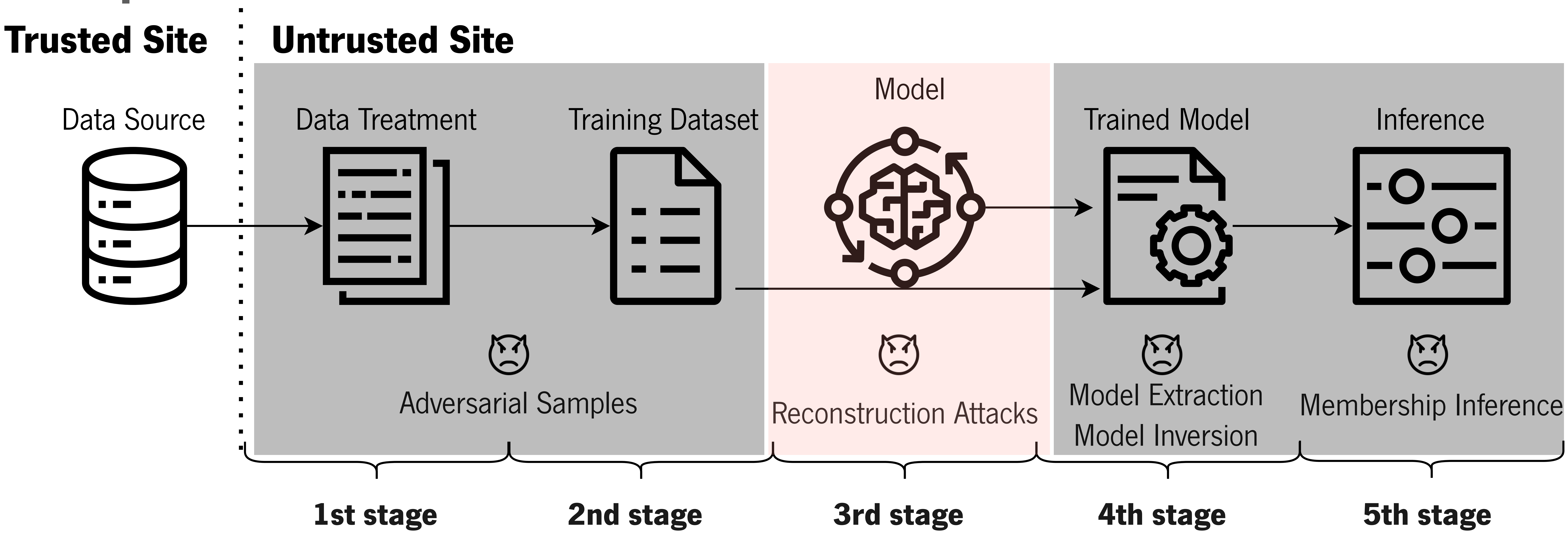
# Privacy and Security in Machine Learning

## ML Pipeline



# Privacy and Security in Machine Learning

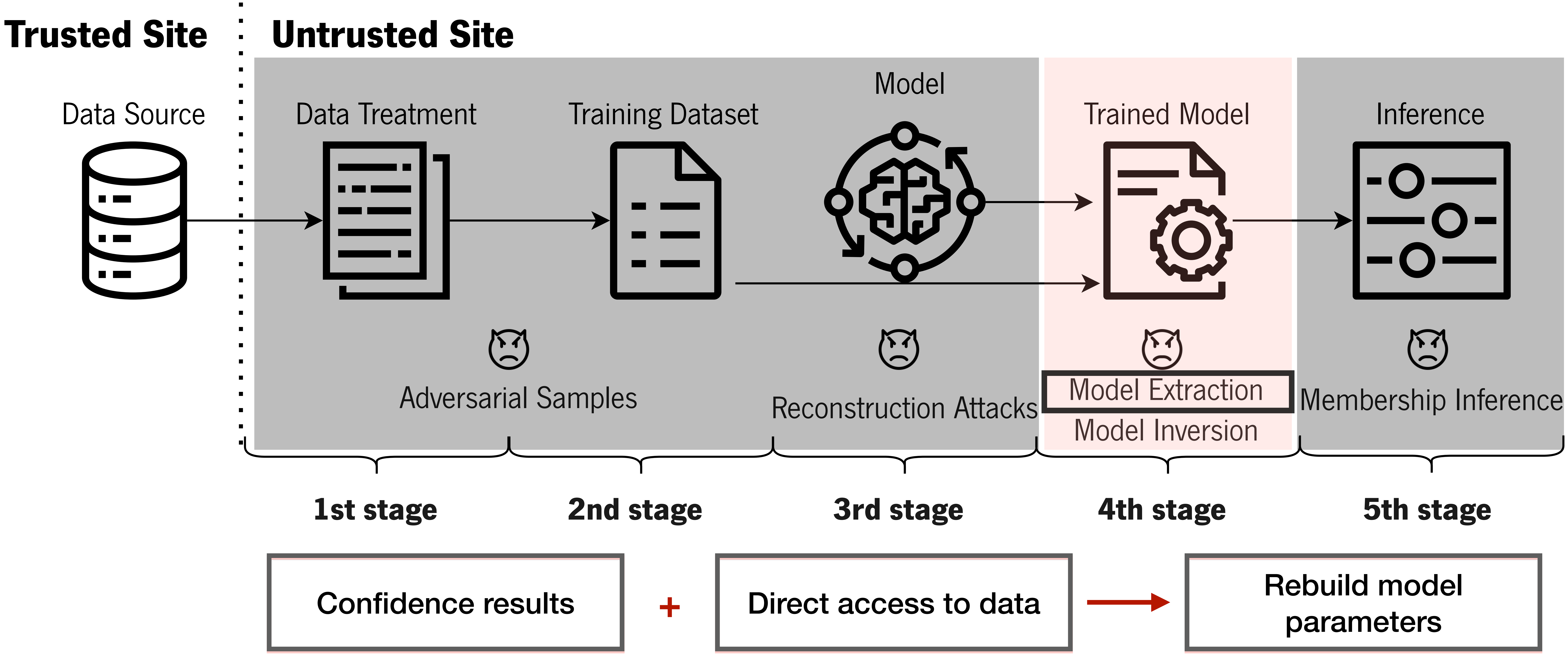
## ML Pipeline




Direct access to feature vectors → Reconstruction of raw data

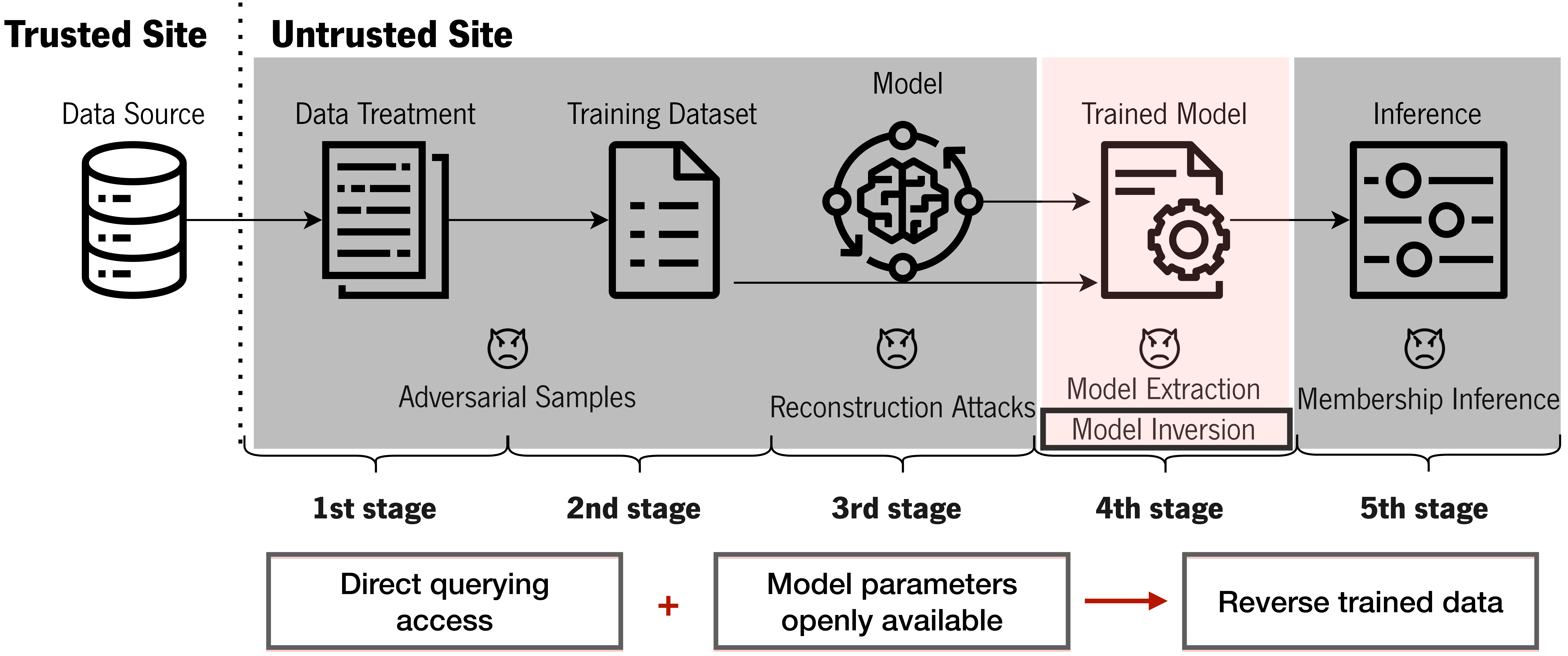
# Privacy and Security in Machine Learning

## ML Pipeline



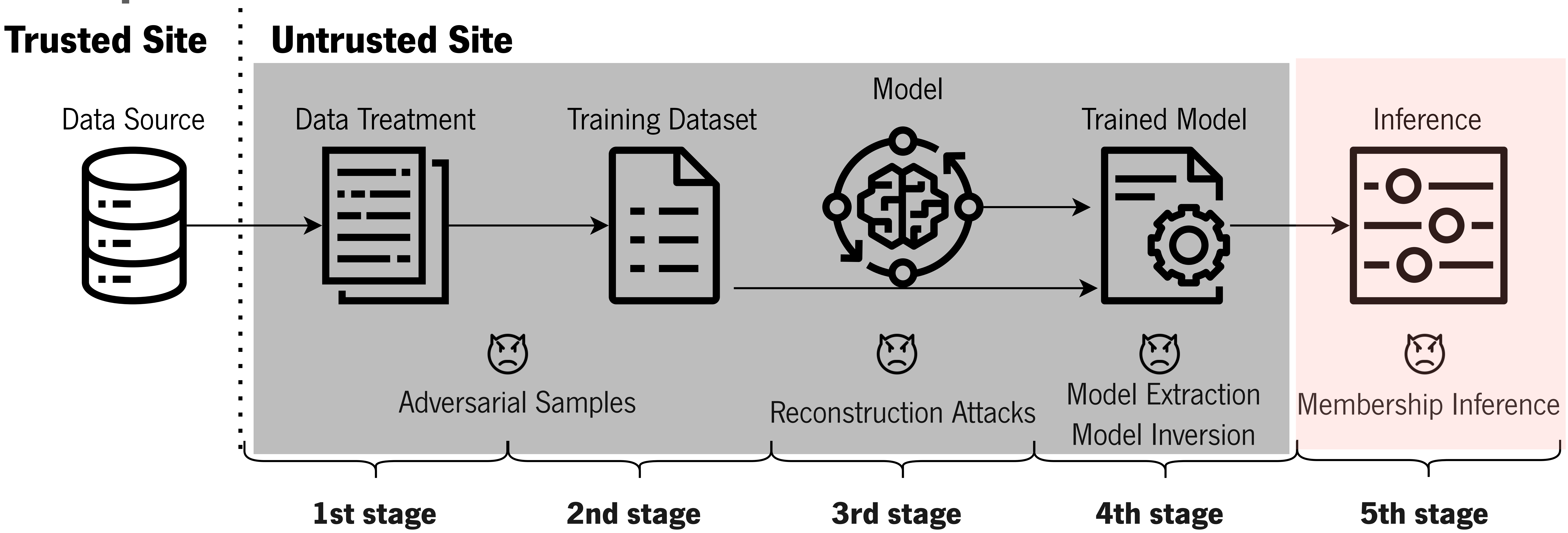
# Privacy and Security in Machine Learning

## ML Pipeline



# Privacy and Security in Machine Learning

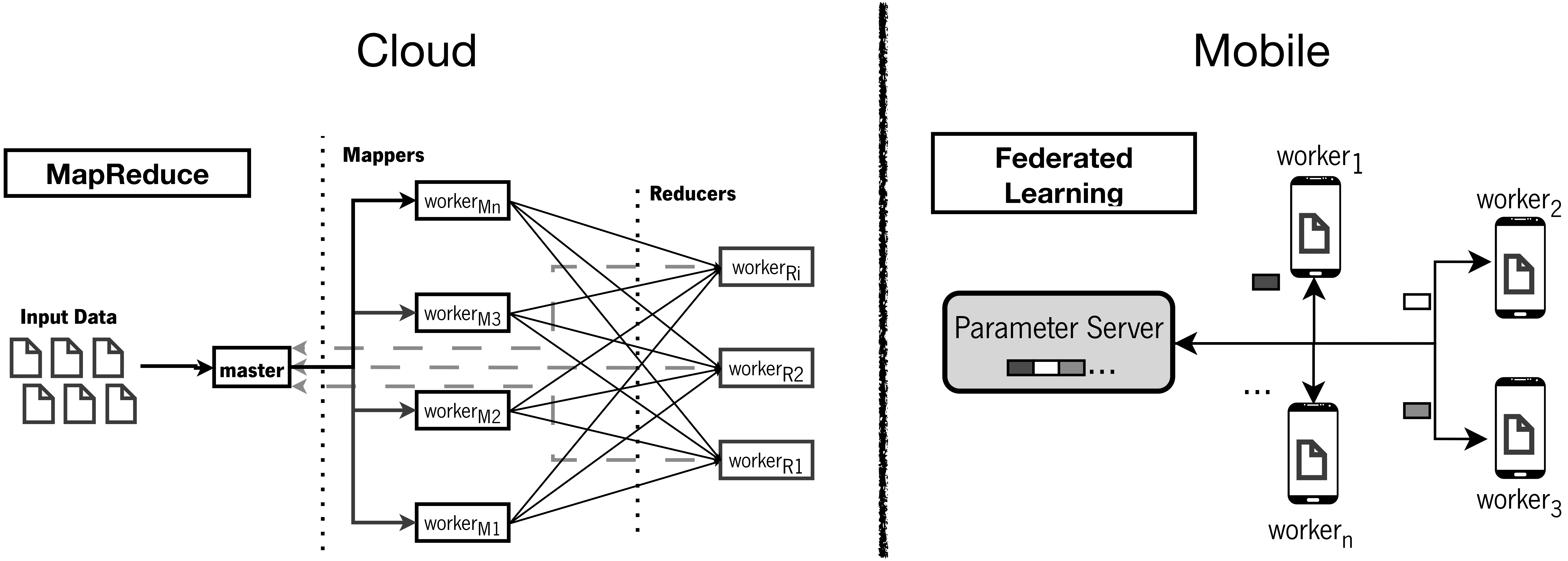
## ML Pipeline




Direct querying access + Model parameters openly available → Check for specific data point on trained dataset

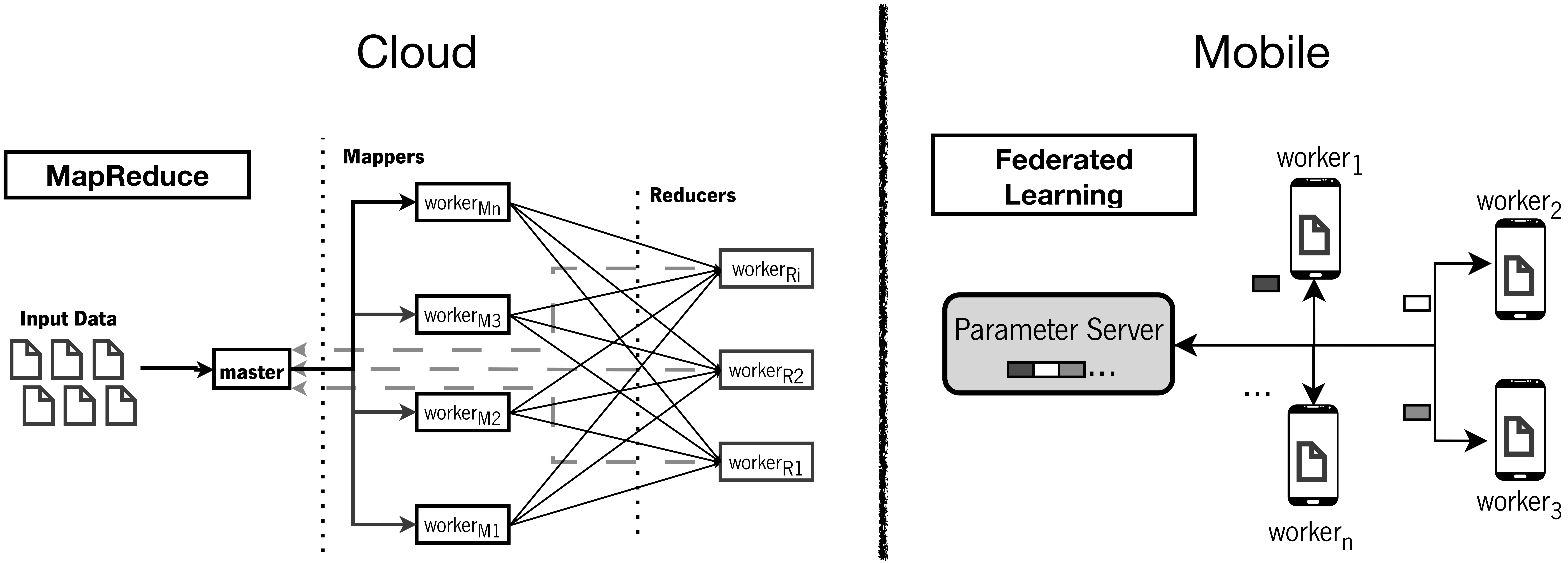
# Privacy and Security in Machine Learning

## Distributed ML



# Privacy and Security in Machine Learning

## Distributed ML




- Efficient and scalable.

# Privacy and Security in Machine Learning

## Privacy-Preserving ML

- Software-based
  - Homomorphic Encryption
  - SMPC
  - Differential Privacy
- Hardware-based
  - Trusted Execution Environments (Intel SGX, AMD SEV, Trustzone)

# Privacy and Security in Machine Learning

## Privacy-Preserving ML

- Software-based
  - Homomorphic Encryption
  - SMPC
  - Differential Privacy
- Hardware-based
  - Trusted Execution Environments (Intel SGX, AMD SEV, Trustzone)

➡ Common cryptographic schemes impose **impractical overheads**.

➡ TEEs' performance depends on the number of **computations**, **I/O** operations and **trusted computing base** (TCB).

# Privacy and Security in Machine Learning

## Challenges

# Privacy and Security in Machine Learning

## Challenges

**Challenge #1 - Privacy**

> Trusting third-parties.

> Rewrite algorithms.

# Privacy and Security in Machine Learning

## Challenges

**Challenge #1 - Privacy**

**>** Trusting third-parties.

**>** Rewrite algorithms.

**Challenge #2 - Utility**

**>** Use case specific.

**>** Accuracy impact.

# Privacy and Security in Machine Learning

## Challenges

**Challenge #1 - Privacy**

> Trusting third-parties.

> Rewrite algorithms.

**Challenge #2 - Utility**

> Use case specific.

> Accuracy impact.

**Challenge #3 - Performance**

> Low application performance.

> High resource consumption.

**Is it possible to balance privacy, performance, and utility in a PPDML solution?**

# Contributions

- ✦ **SOTERIA:** A generic PPDML solution built on top of Apache Spark and MLlib based on computation partitioning[1,2].
- ✦ **GYOSA:** A specialized privacy-preserving solution built on top of Apache Spark and Glow to handle genomic data[3].
- ✦ **TAPUS:** A FL prototype to ensure user's mobility data privacy[4,5].

1. **Brito, C.**, Ferreira, P., Portela, B., Oliveira, R. and Paulo, J. "SOTERIA: Preserving Privacy in Distributed Machine Learning." In Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing, 2023.
2. **Brito, C.**, Ferreira, P., Portela, B., Oliveira, R. and Paulo, J. "Privacy-Preserving Machine Learning on Apache Spark." In IEEE Access, 2023.
3. **Brito, C.**, Ferreira, P. and Paulo, J., "A Distributed Computing Solution for Privacy-Preserving Genome-Wide Association Studies." Available as a preprint in bioRxiv and submitted for JBHI.
4. Pina, N., **Brito, C.**, Vitorino, R., Cunha, I. "Promoting sustainable and personalized travel behaviors while preserving data privacy." In Transportation Research Procedia - Proceedings of TRALisbon, 2022.
5. **Brito, C.**, Pina, N., Esteves, T., Vitorino, R., Cunha, I., Paulo, J. "Promoting sustainable and personalized travel behaviors while preserving data privacy." Accepted on Transportation Engineering (TRENG), 2024.

# SOTERIA

## Preserving Privacy in Distributed Machine Learning

# SOTERIA

## Preserving Privacy in Distributed Machine Learning

- **General applicability for ML workloads:**

➡ Several algorithms from Spark's Machine Learning API.

- **Privacy-by-design:**

➡ Plaintext information only inside the enclaves (by resorting to Intel SGX).

- **Balanced overhead:**

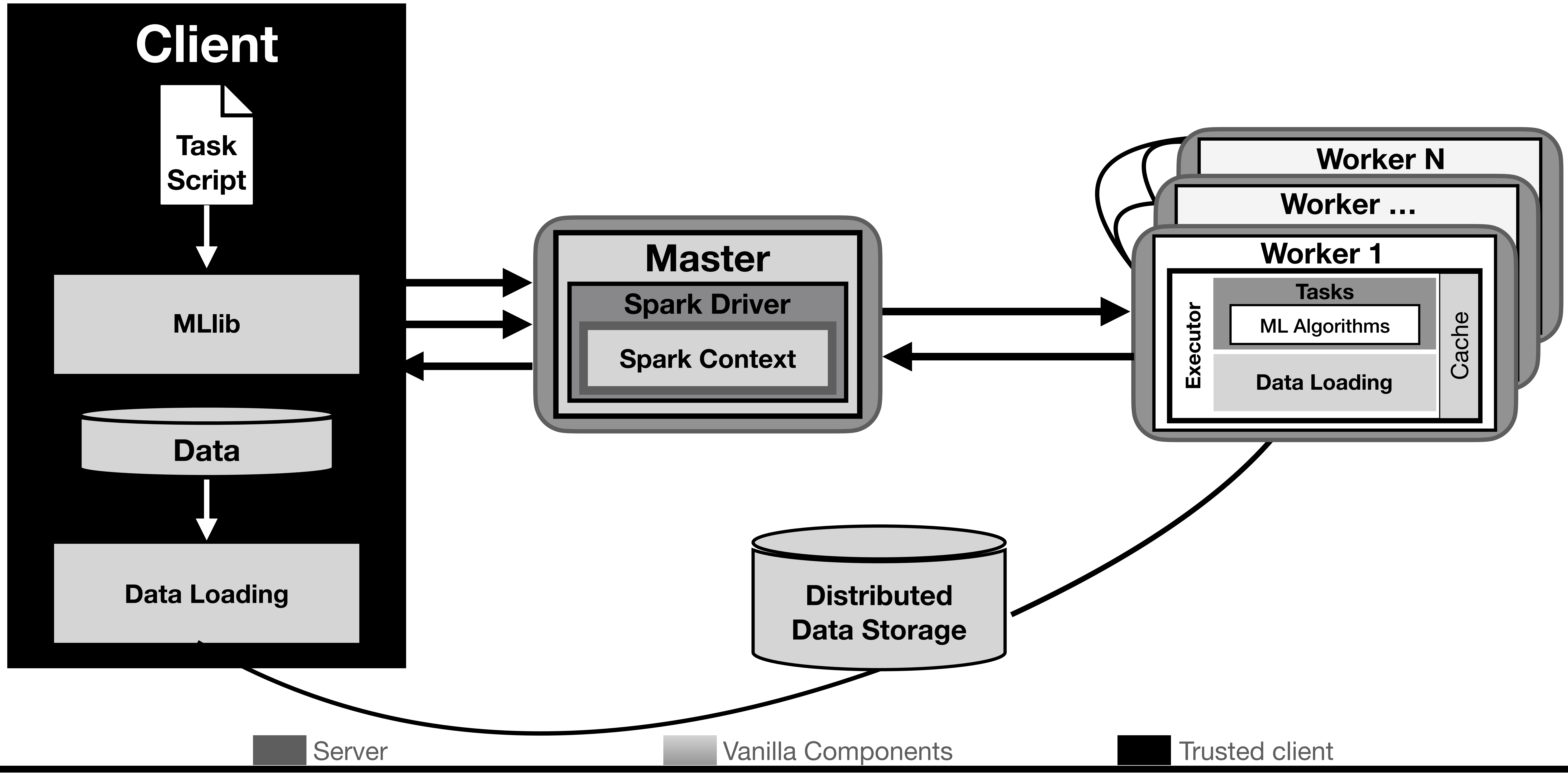
➡ Novel partitioning scheme balancing privacy and performance.

- **Low intrusiveness:**

➡ Processing flow remains unchanged.

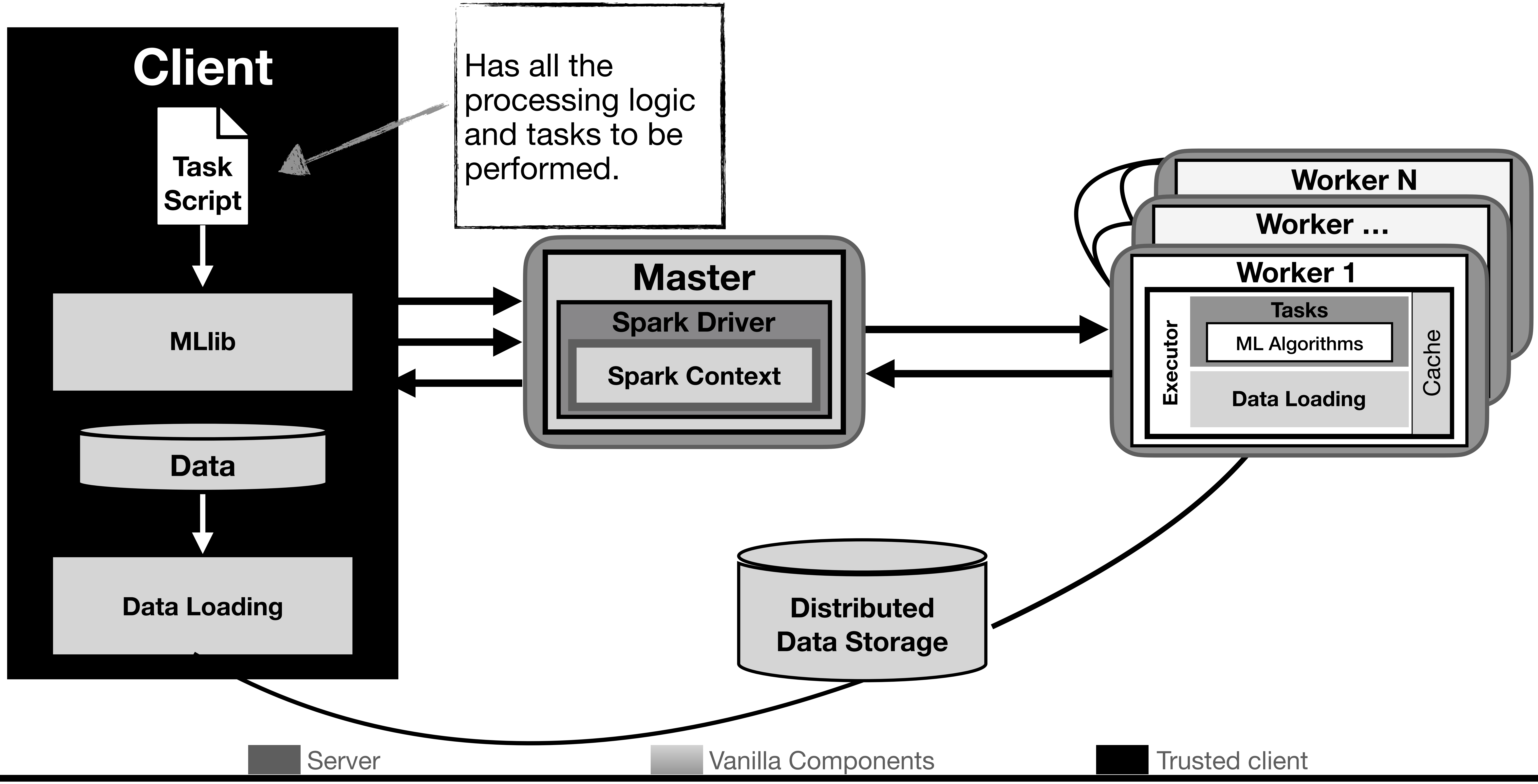
# SOTERIA

## Preserving Privacy in Distributed Machine Learning



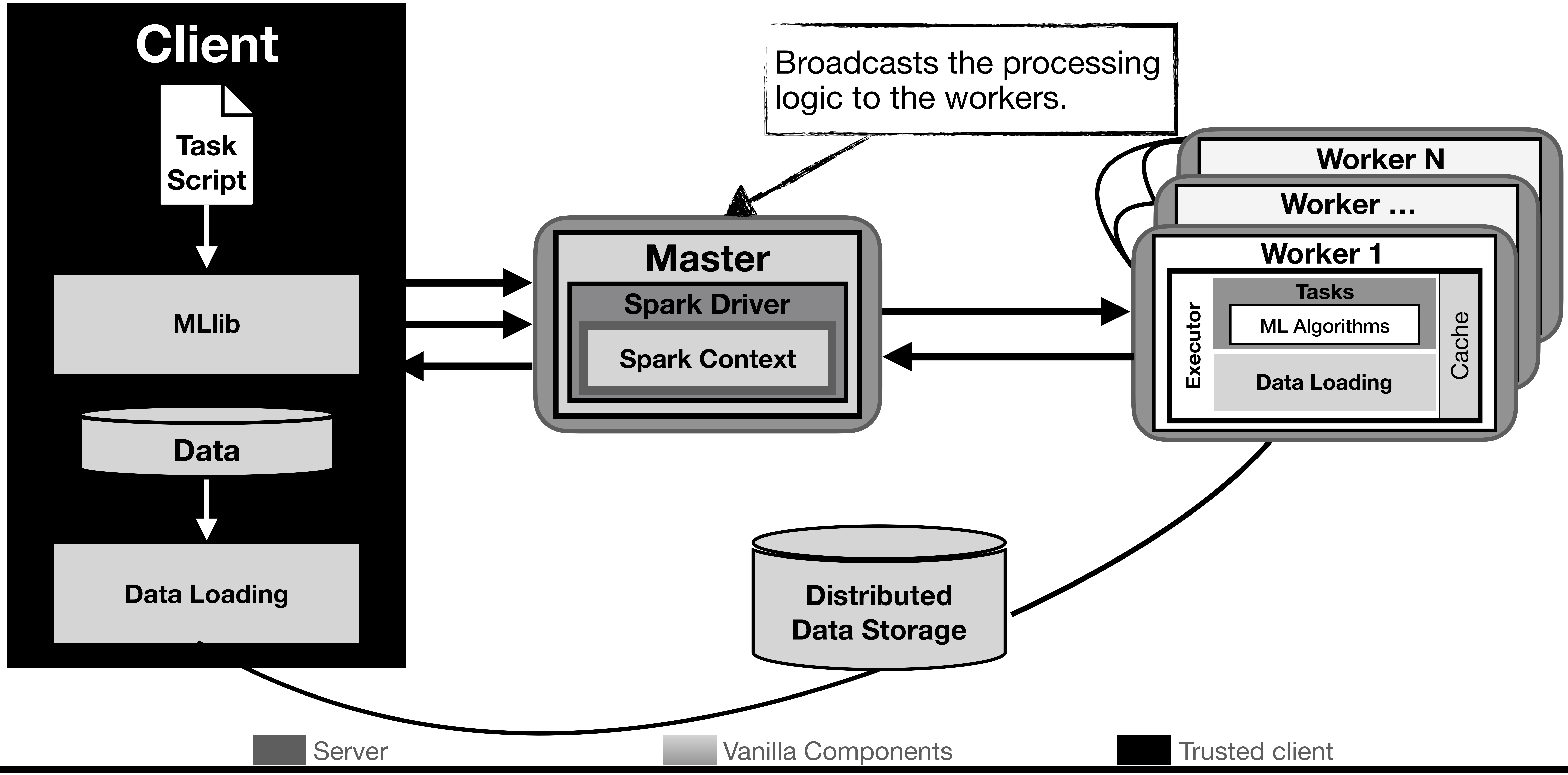
# SOTERIA

## Preserving Privacy in Distributed Machine Learning



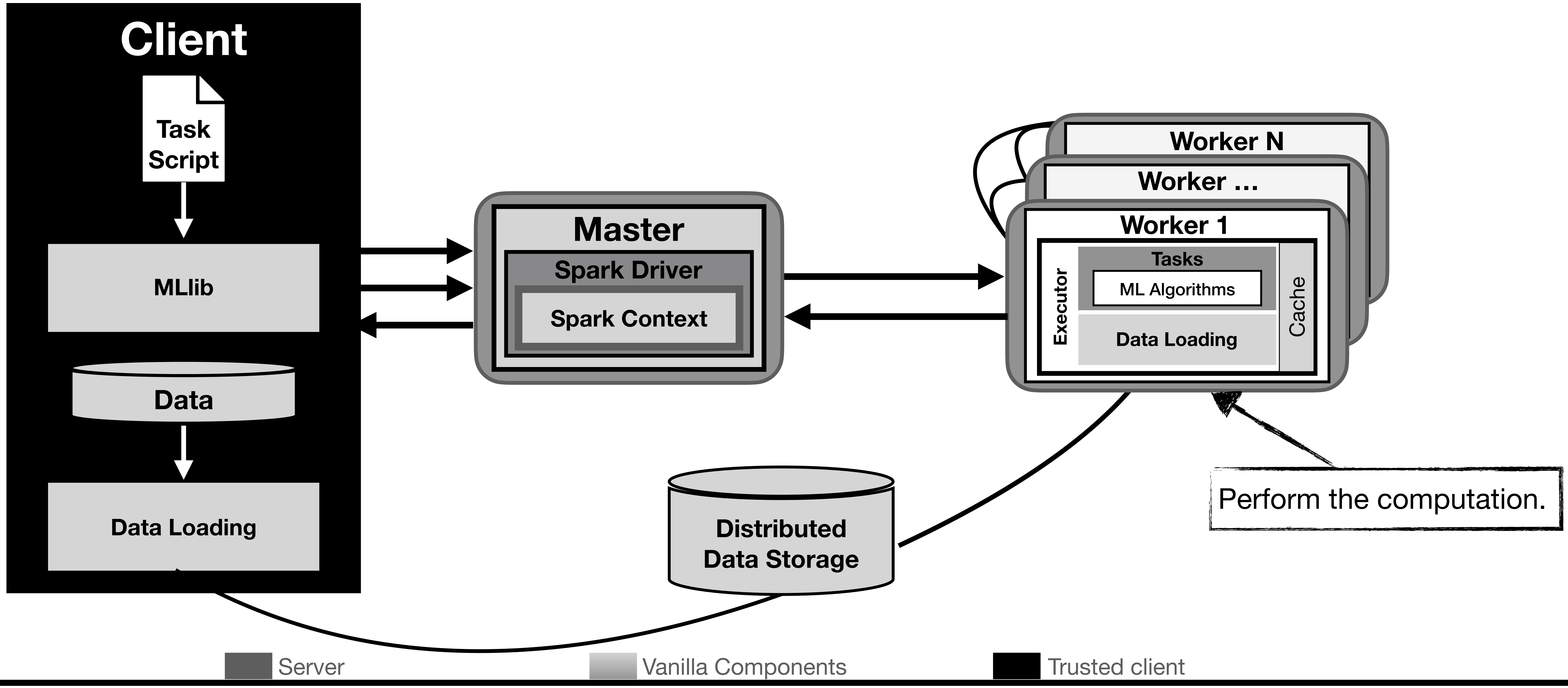
# SOTERIA

## Preserving Privacy in Distributed Machine Learning



# SOTERIA

## Preserving Privacy in Distributed Machine Learning

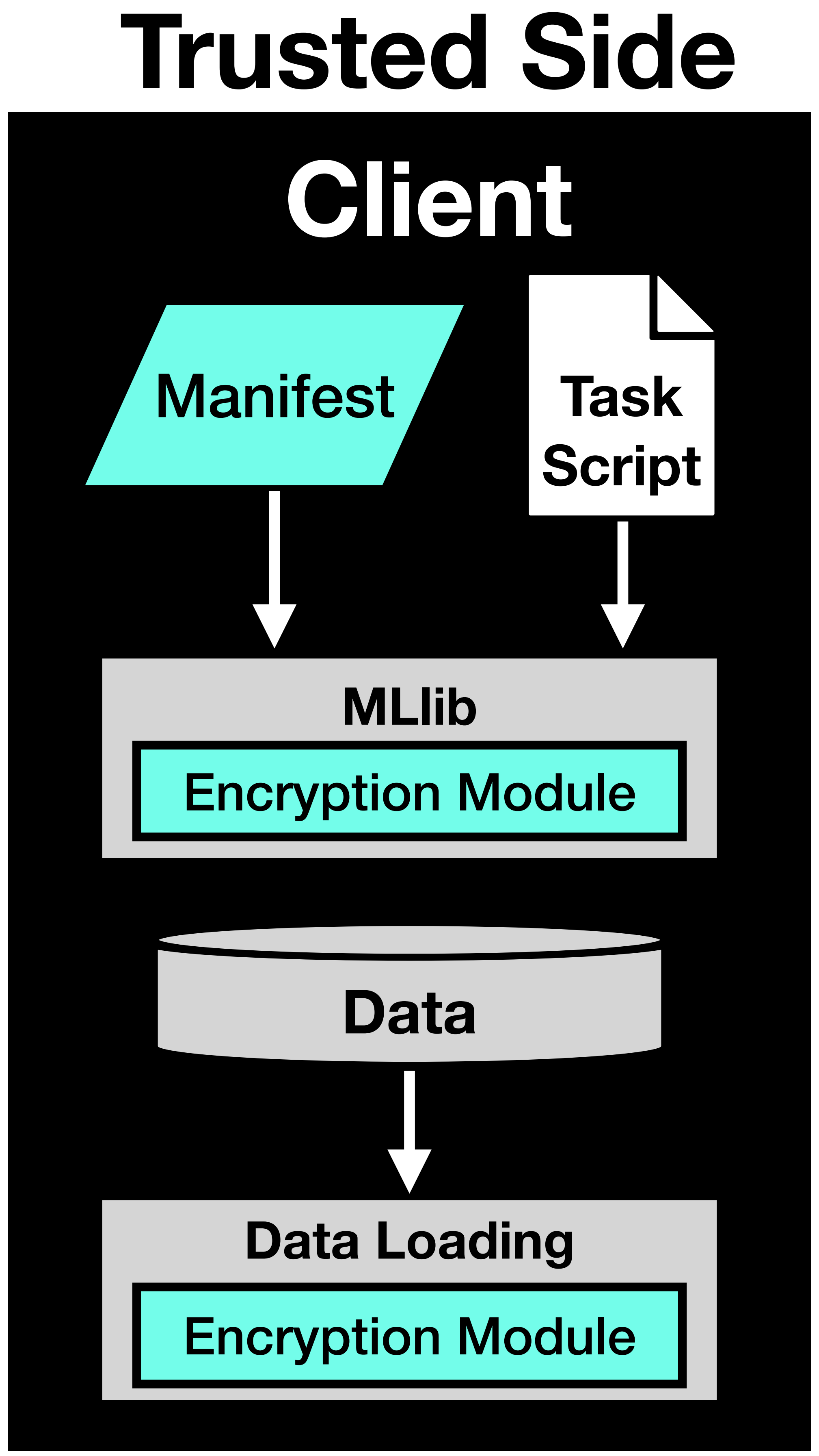


# SOTERIA

## Preserving Privacy in Distributed Machine Learning

# SOTERIA

## Preserving Privacy in Distributed Machine Learning

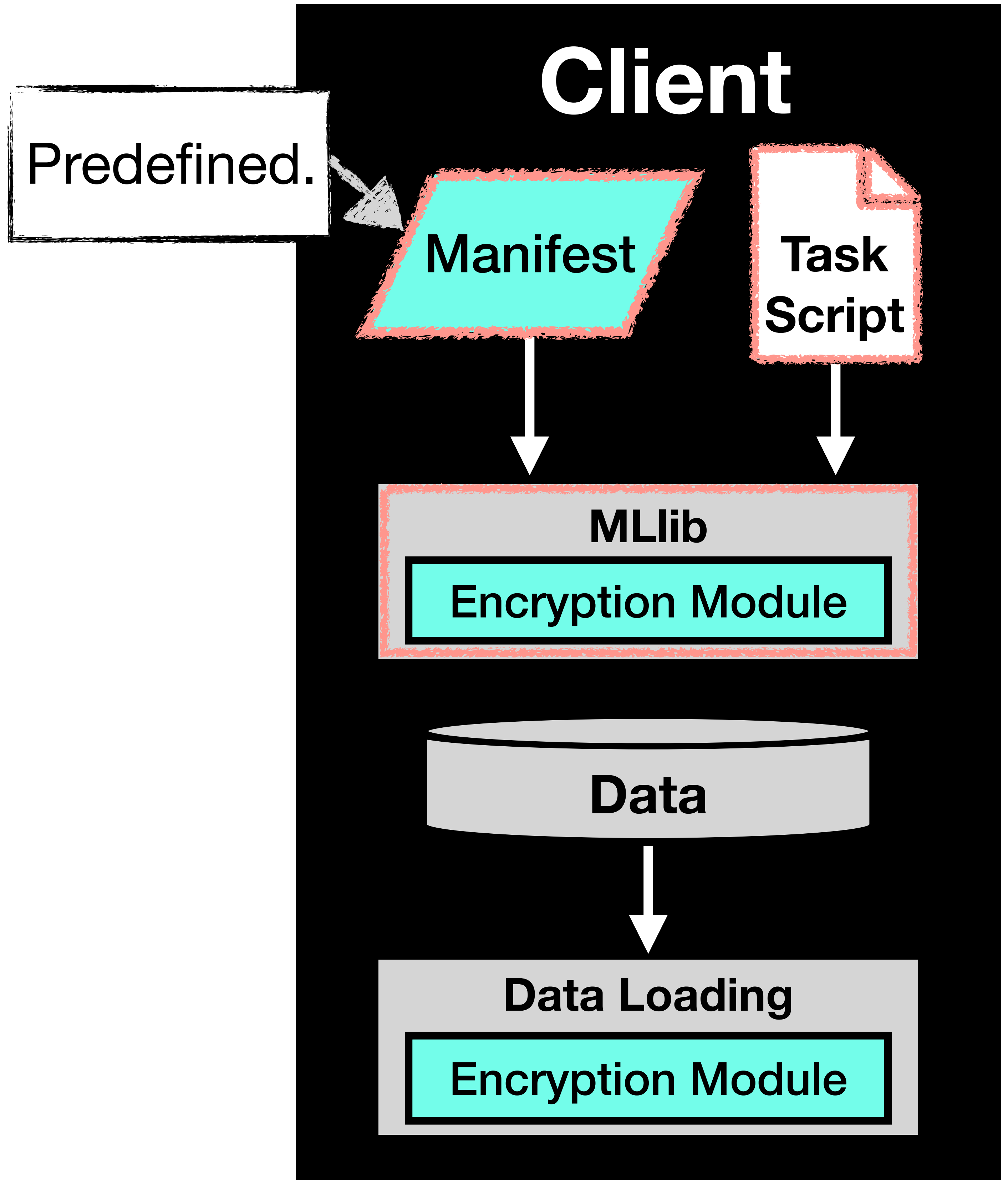



Server New components Vanilla Components Enclave Trusted client Secure Channel

# SOTERIA

## Preserving Privacy in Distributed Machine Learning

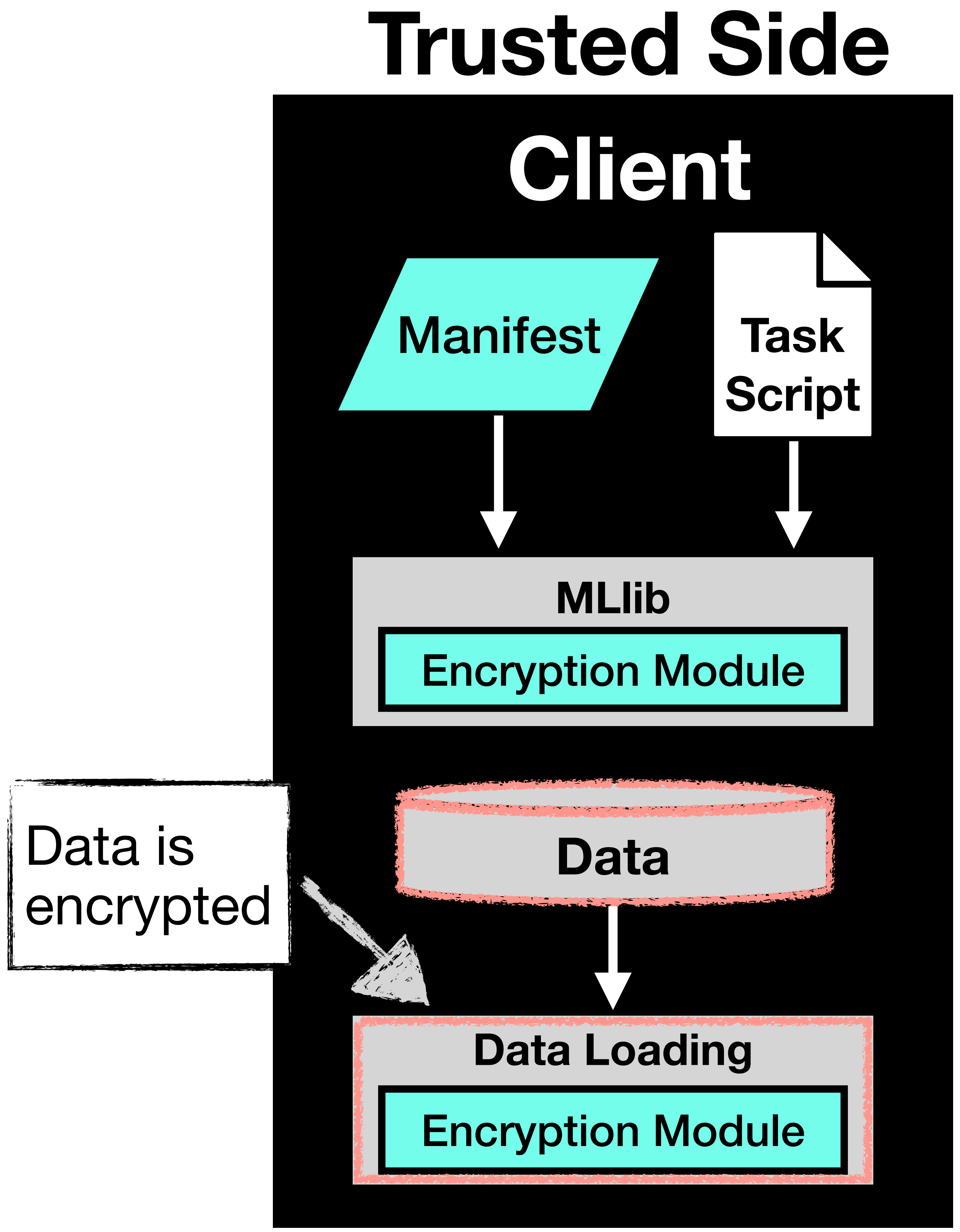
### Trusted Side




Server | New components | Vanilla Components | Enclave | Trusted client | Secure Channel

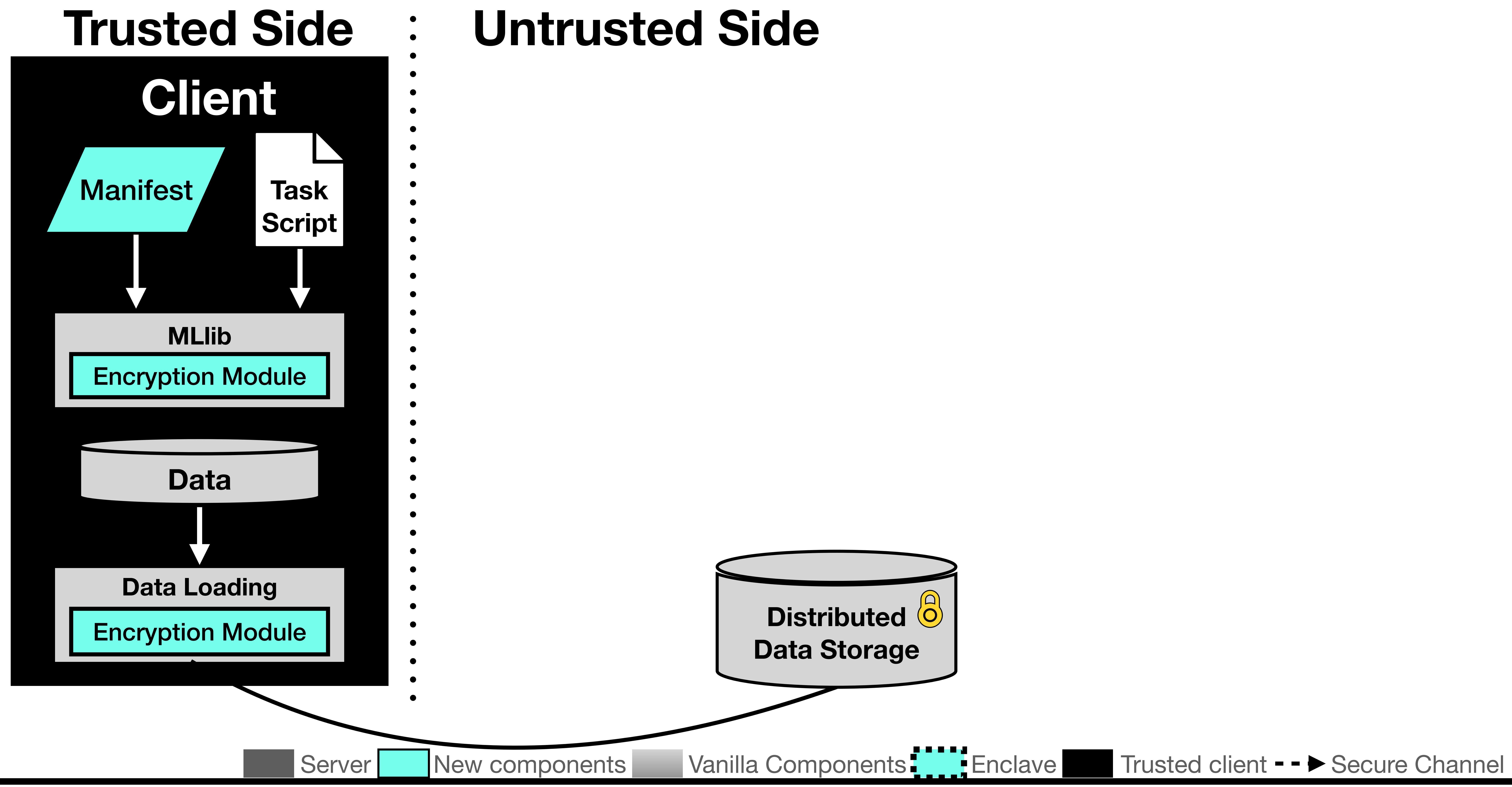
# SOTERIA

## Preserving Privacy in Distributed Machine Learning



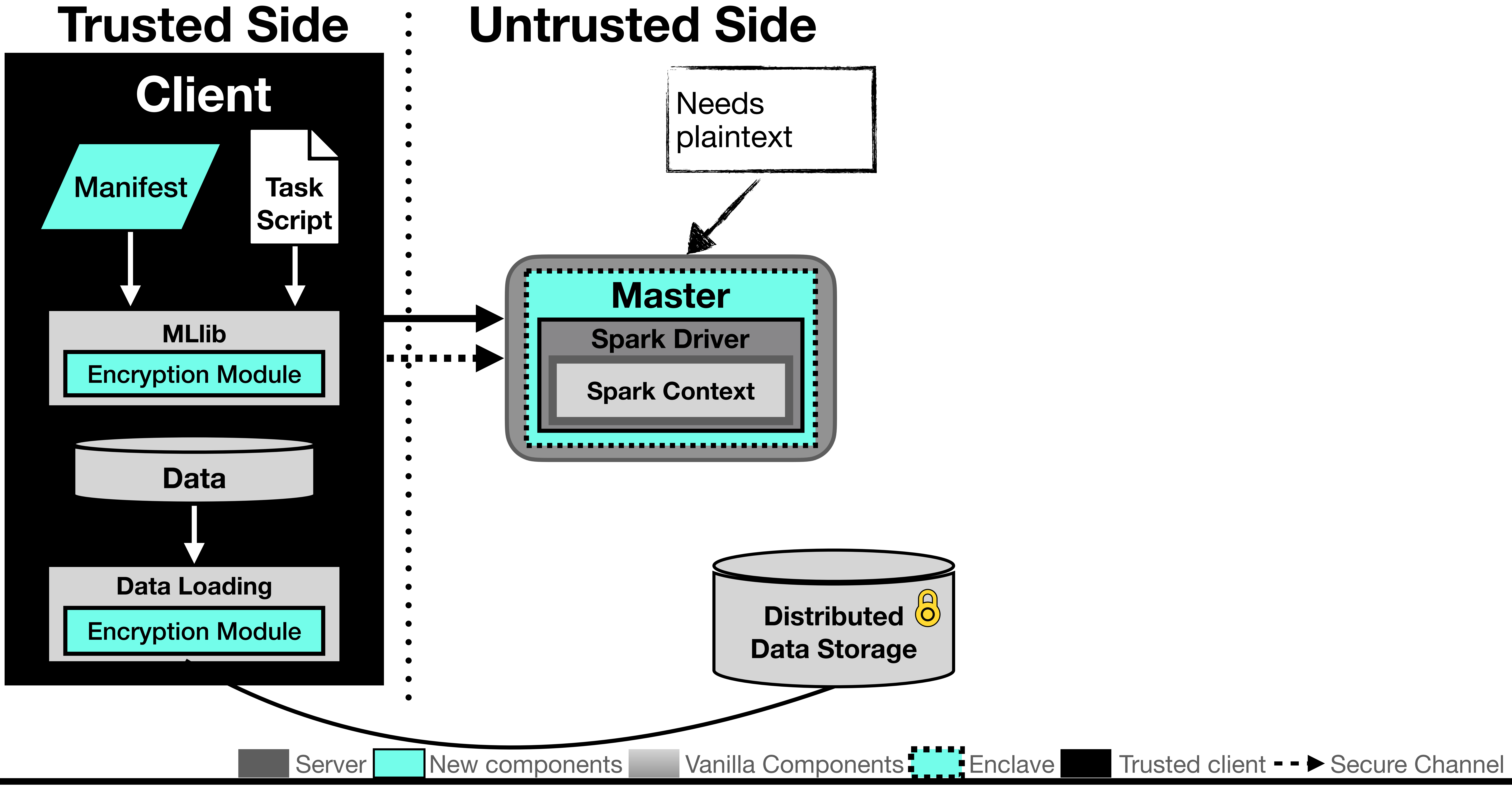

# SOTERIA

## Preserving Privacy in Distributed Machine Learning



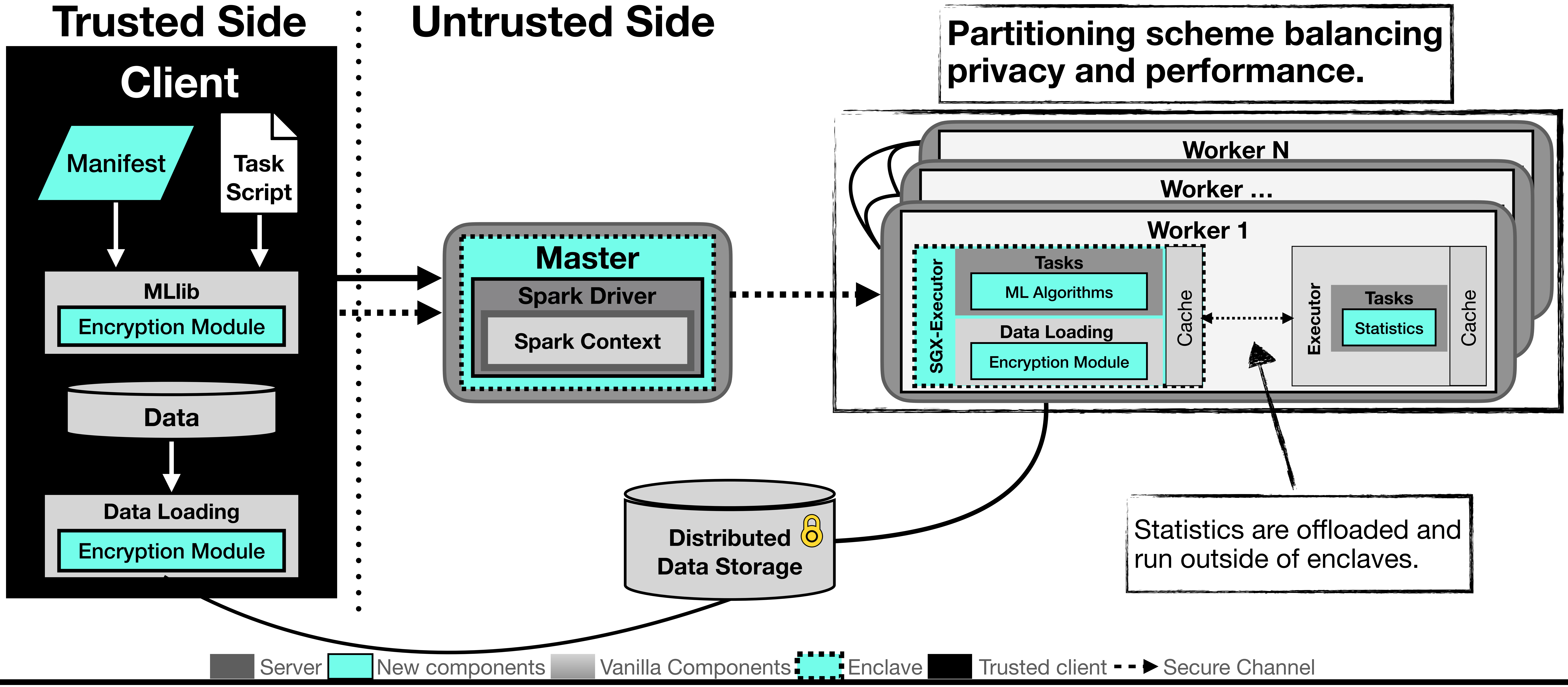
# SOTERIA

## Preserving Privacy in Distributed Machine Learning



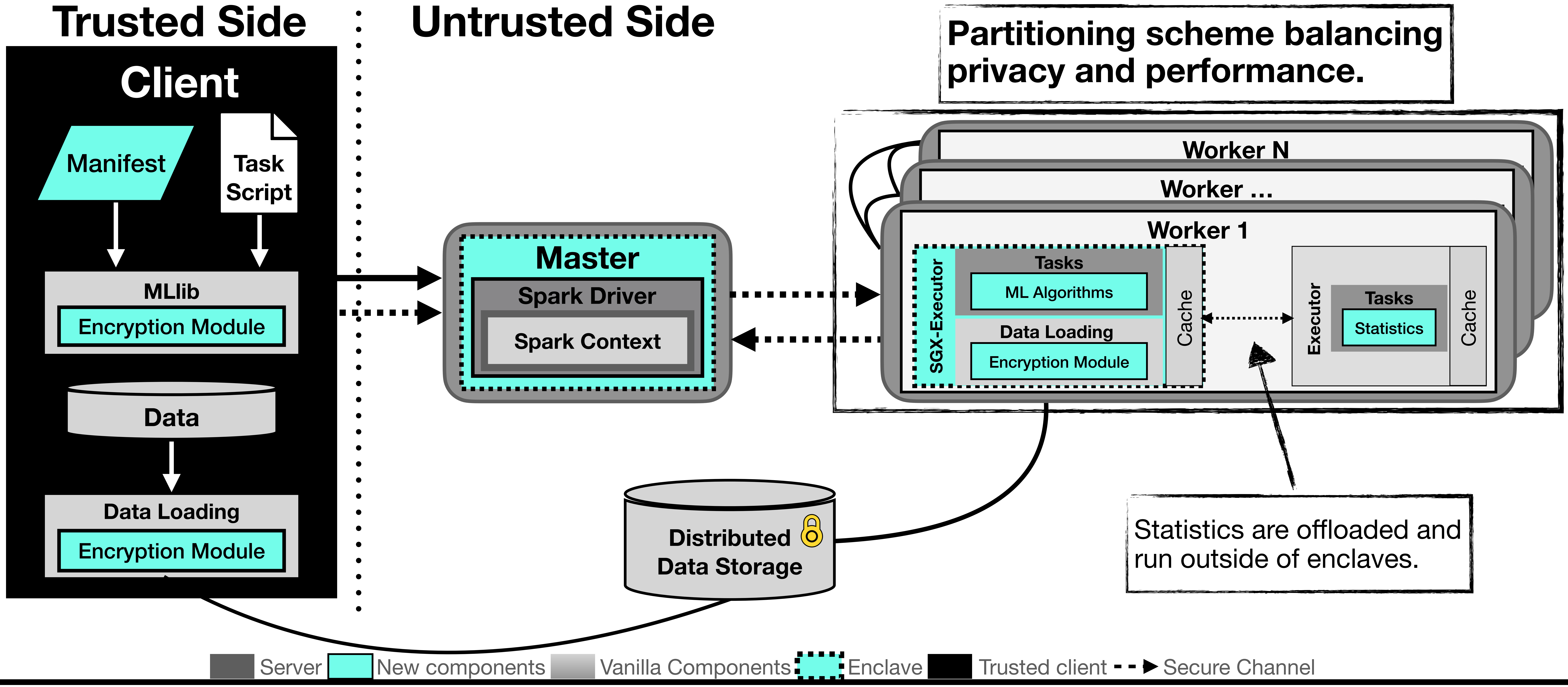
# SOTERIA

## Preserving Privacy in Distributed Machine Learning



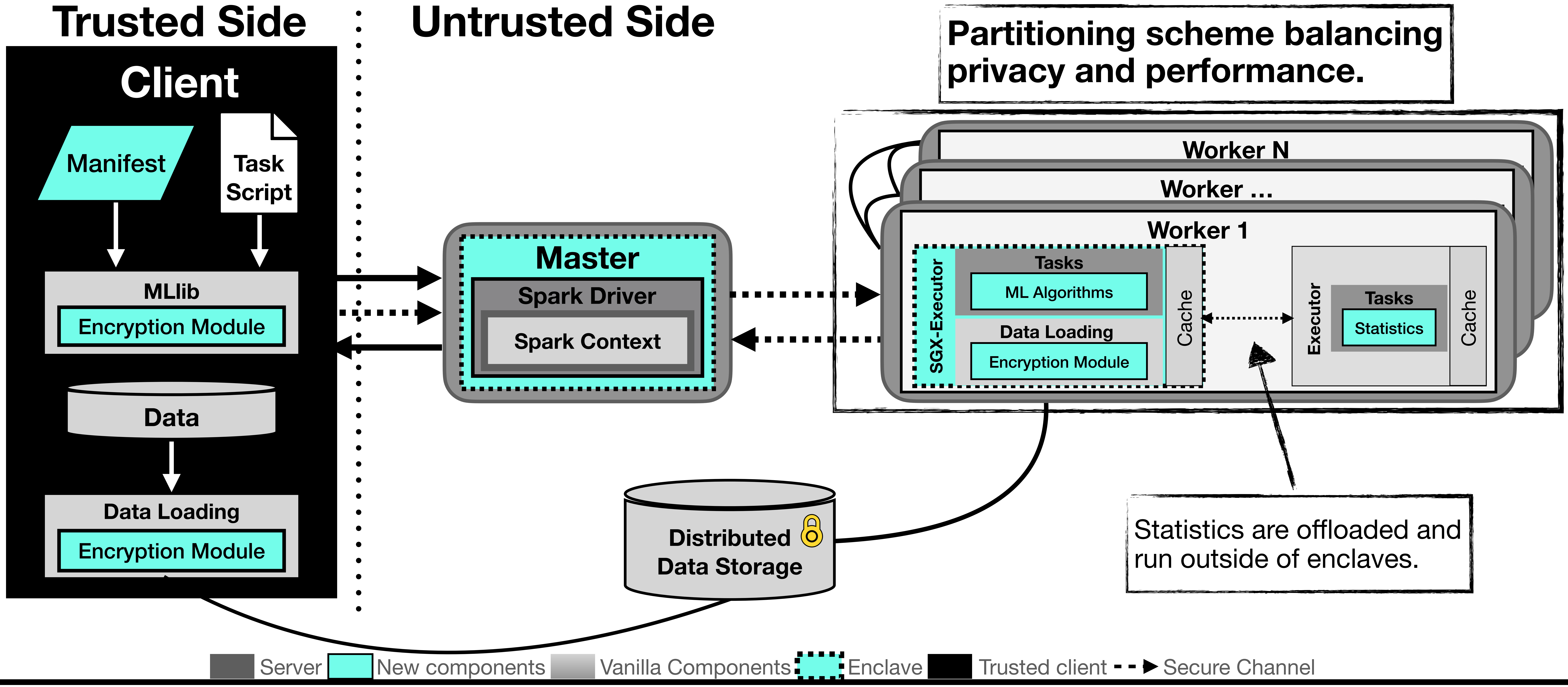
# SOTERIA

## Preserving Privacy in Distributed Machine Learning



# SOTERIA

## Preserving Privacy in Distributed Machine Learning



# SOTERIA

## Results

### Setup

8 servers, Intel Core i5-9500 CPU, 16 GB RAM, and a 256GB NVMe
1 Master/Client, 7 Workers

### Evaluation

**HiBench**

Machine Learning Algorithms
- LR
- PCA
- GBT
- KMeans
- Naive Bayes
- ALS
- LDA

Data Sizes:
- Tiny
- Large
- Huge
- Gigantic

Vanilla
Soteria-B
Soteria-P
SGX-Spark

# SOTERIA

## Results

### Evaluation

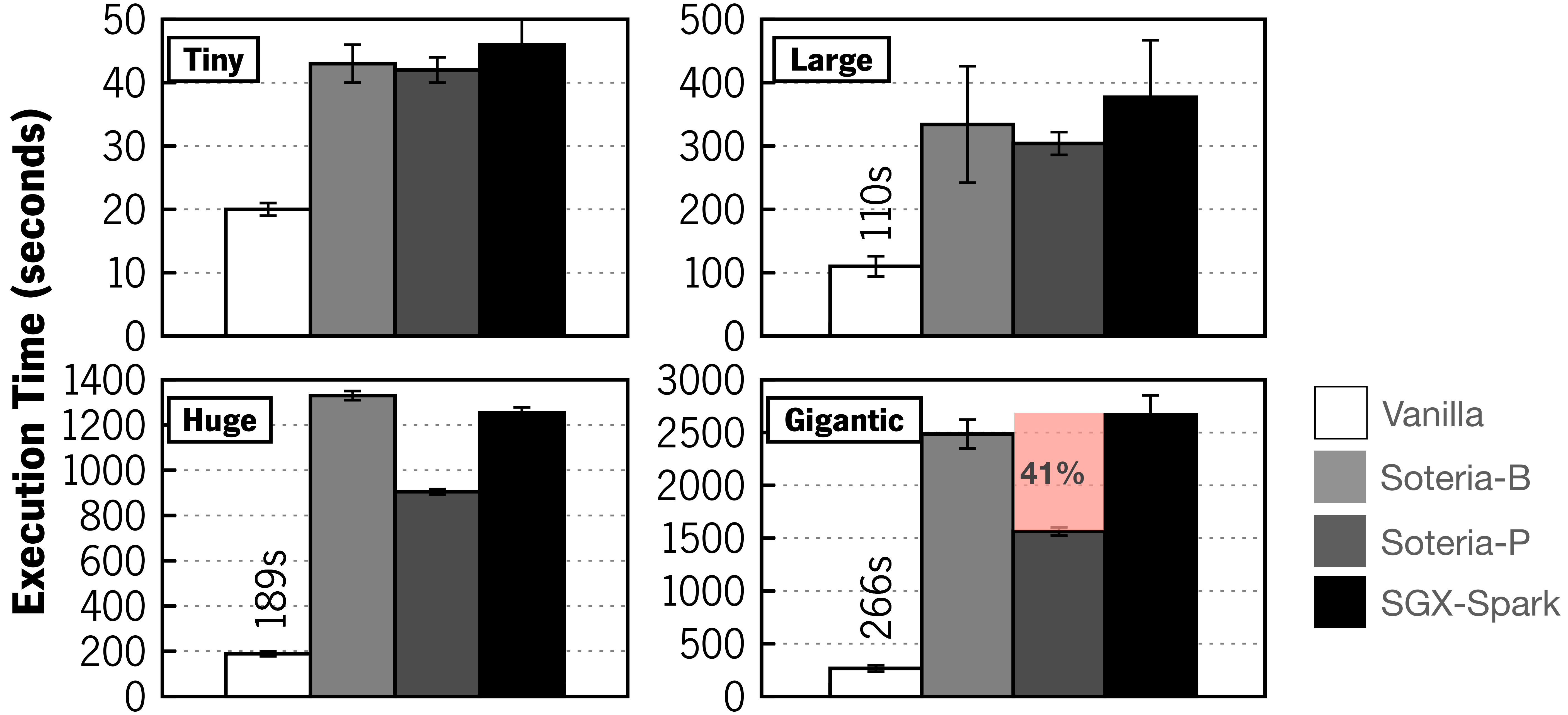
**HiBench**

Machine Learning Algorithms
- LR
- PCA
- GBT
- KMeans
- Naive Bayes
- ALS
- LDA

Data Sizes:
- Tiny
- Large
- Huge
- Gigantic

### Setup

8 servers, Intel Core i5-9500 CPU, 16 GB RAM, and a 256GB NVMe
1 Master/Client, 7 Workers




GBT

- Soteria-P deals better with the data volume increase.

# SOTERIA

## Results

### Evaluation

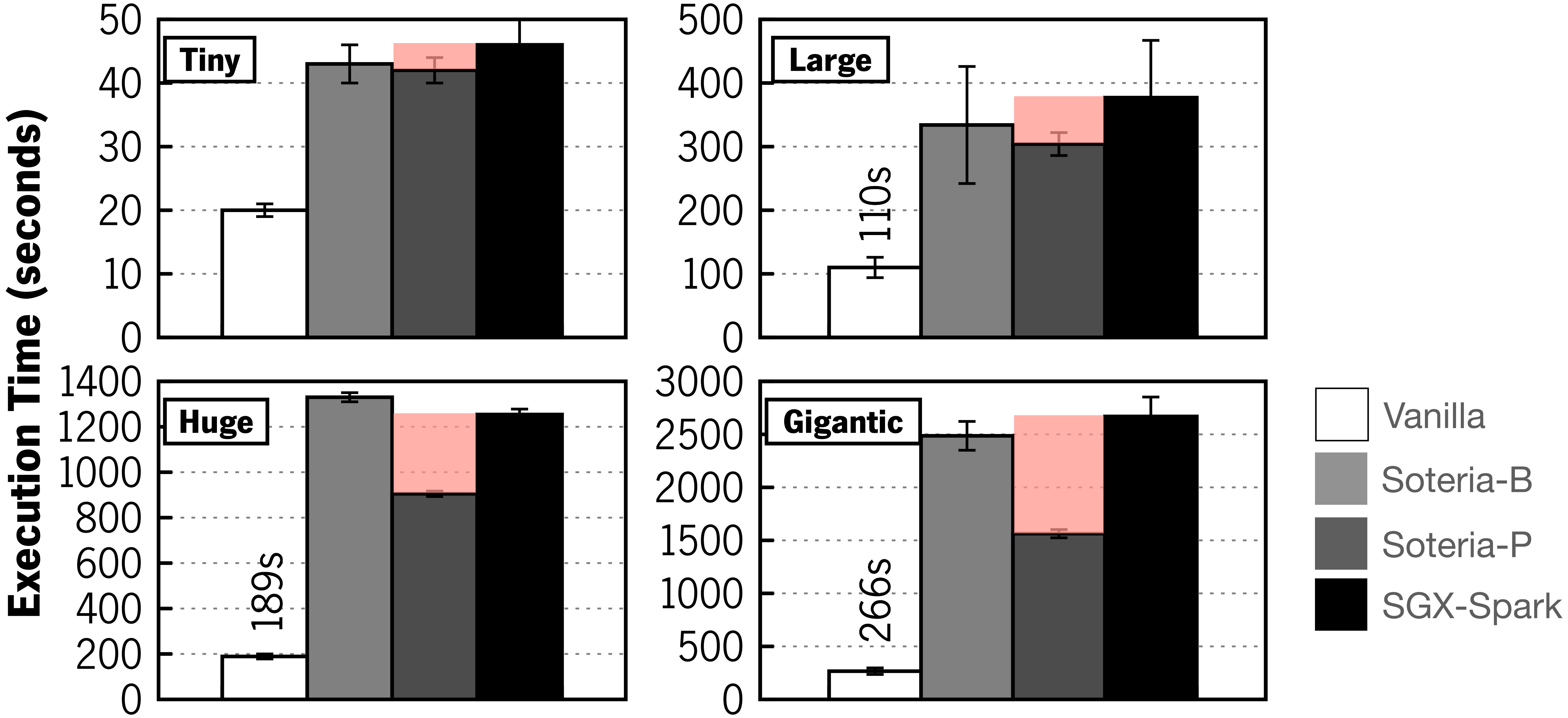
**HiBench**

Machine Learning Algorithms
- LR
- PCA
- GBT
- KMeans
- Naive Bayes
- ALS
- LDA

Data Sizes:
- Tiny
- Large
- Huge
- Gigantic

### Setup

8 servers, Intel Core i5-9500 CPU, 16 GB RAM, and a 256GB NVMe
1 Master/Client, 7 Workers




GBT

- Soteria-P deals better with the data volume increase.
- Soteria-P consistently outperforms SGX-Spark and Soteria-B.

# SOTERIA

## SUMMARY

- **SOTERIA** introduces a novel **partitioning scheme** (Soteria-P) allowing specific ML operations to be deployed outside trusted enclaves.
  - **Offloading non-sensitive operations** from enclaves while covering several ML attacks.
  - **Support** of numerous **ML algorithms**.
  - **Non-intrusive** to the **clients** flow.

➡ Can **SOTERIA** be applied to a specific use case such **genomic analysis**?

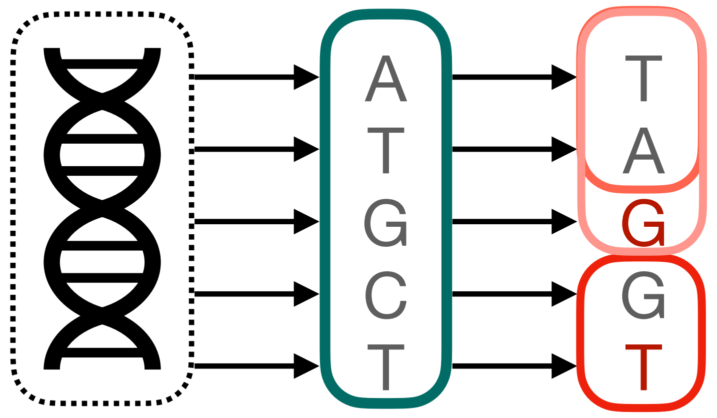
# GYOSA

## Privacy-Preserving Machine Learning for Genome-Wide Association Studies

# GYOSA

## Privacy-Preserving Machine Learning for Genome-Wide Association Studies

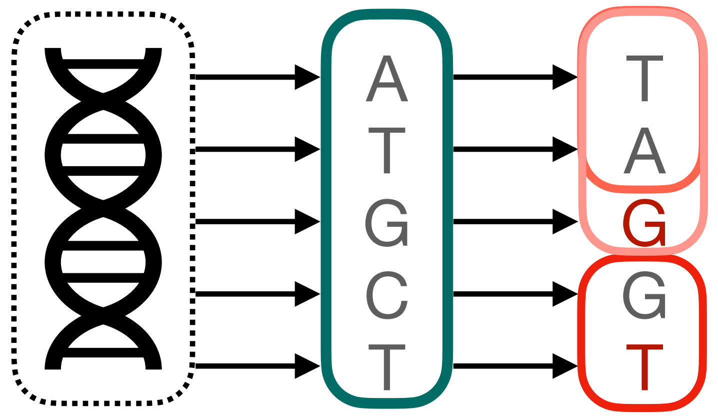
- Genomic data is extremely sensitive and presents a different analysis pipeline.
    - Different algorithms, different data types.



# GYOSA

## Privacy-Preserving Machine Learning for Genome-Wide Association Studies

- Genomic data is extremely sensitive and presents a different analysis pipeline.
  - Different algorithms, different data types.




- **GYOSA** extends **SOTERIA** allowing the computation of GWAS:
  - Updated encryption module to support genomic data types (i.e., VCFs).
  - Extended support for Glow, allowing the partitioning of regression-based algorithms built for GWAS.

# GYOSA

## Privacy-Preserving Machine Learning for Genome-Wide Association Studies

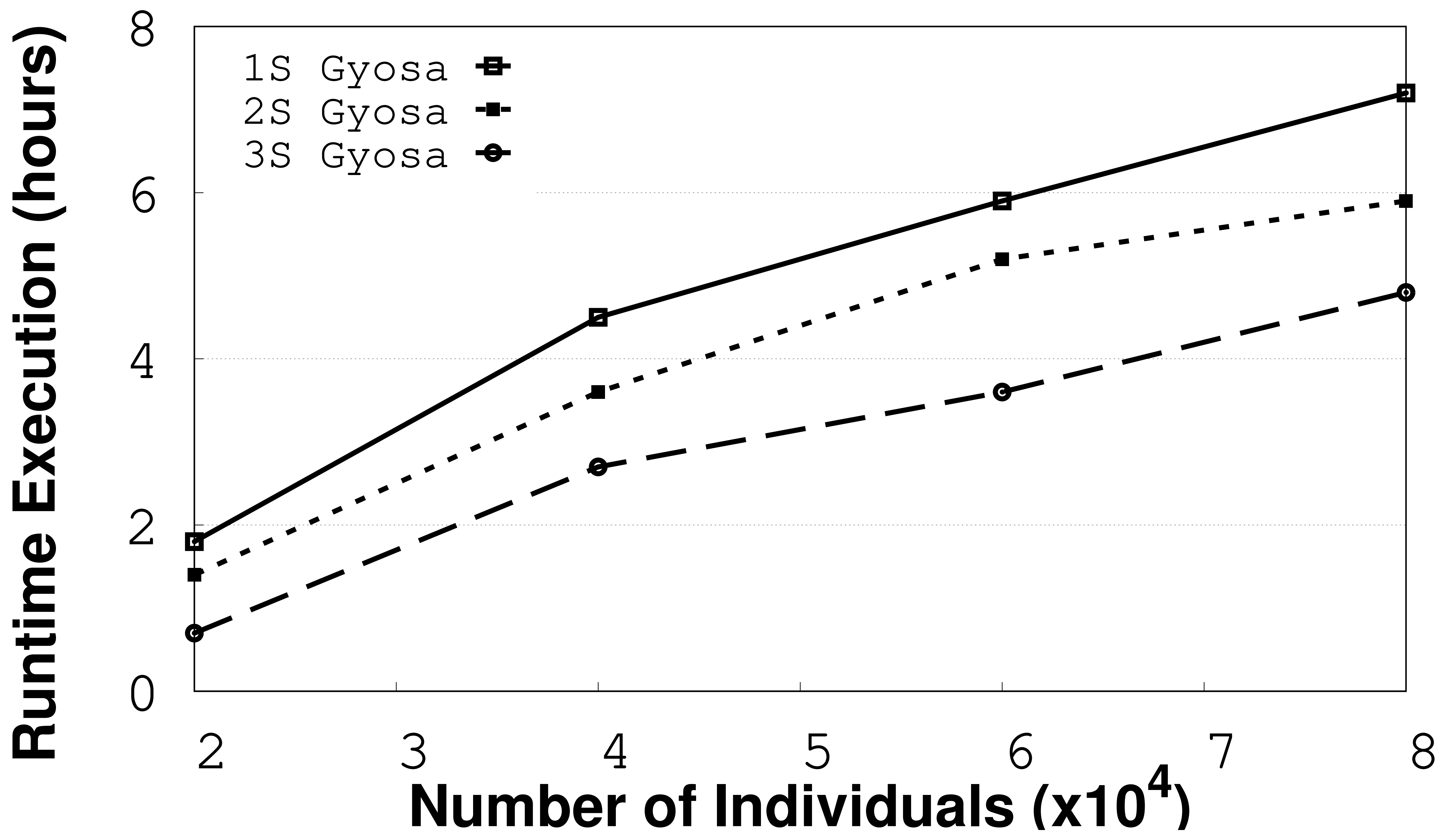
**Setup**

4 servers: Intel Core i5-9500 CPU, 16 GB RAM, and a 256GB NVMe
1-3 workers / 1 master/client

**Evaluation**

**Algorithms:**
- Linear Regression
- Logistic Regression
- $X^2$ Test

**Data Sizes:**
- 1GB - 32GB
- $2x10^4$ - $8x10^4$ individuals, with $1x10^6$ SNPs.



# GYOSA

**Setup**

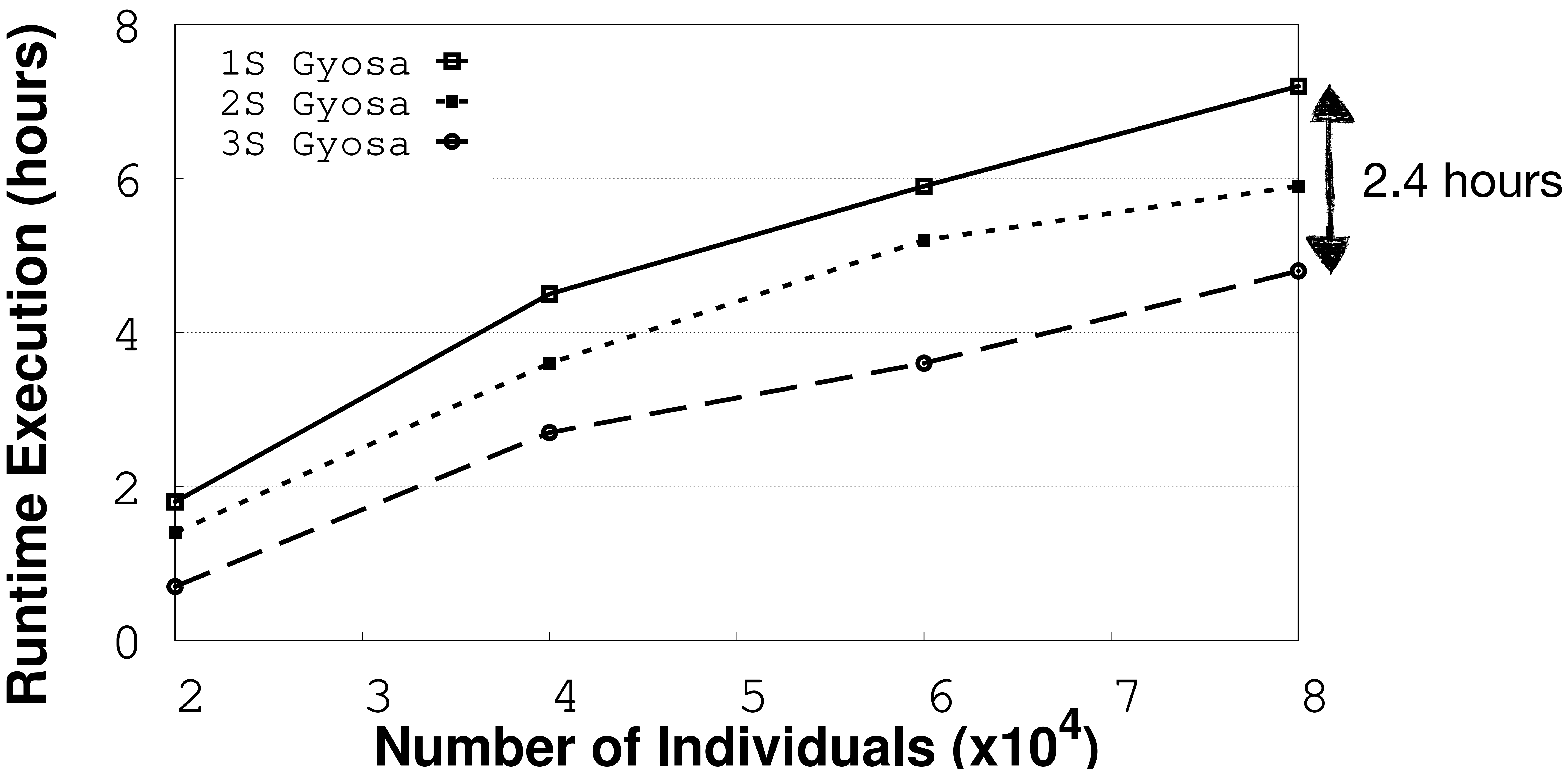
4 servers: Intel Core i5-9500 CPU, 16 GB RAM, and a 256GB NVMe
1-3 workers / 1 master/client

## Privacy-Preserving Machine Learning for Genome-Wide Association Studies

**Evaluation**

**Algorithms:**
- Linear Regression
- Logistic Regression
- $X^2$ Test

**Data Sizes:**
- 1GB - 32GB
- $2x10^4$ - $8x10^4$ individuals, with $1x10^6$ SNPs.




- The runtime execution decreases by 2.4 hours.

# GYOSA

## Privacy-Preserving Machine Learning for Genome-Wide Association Studies

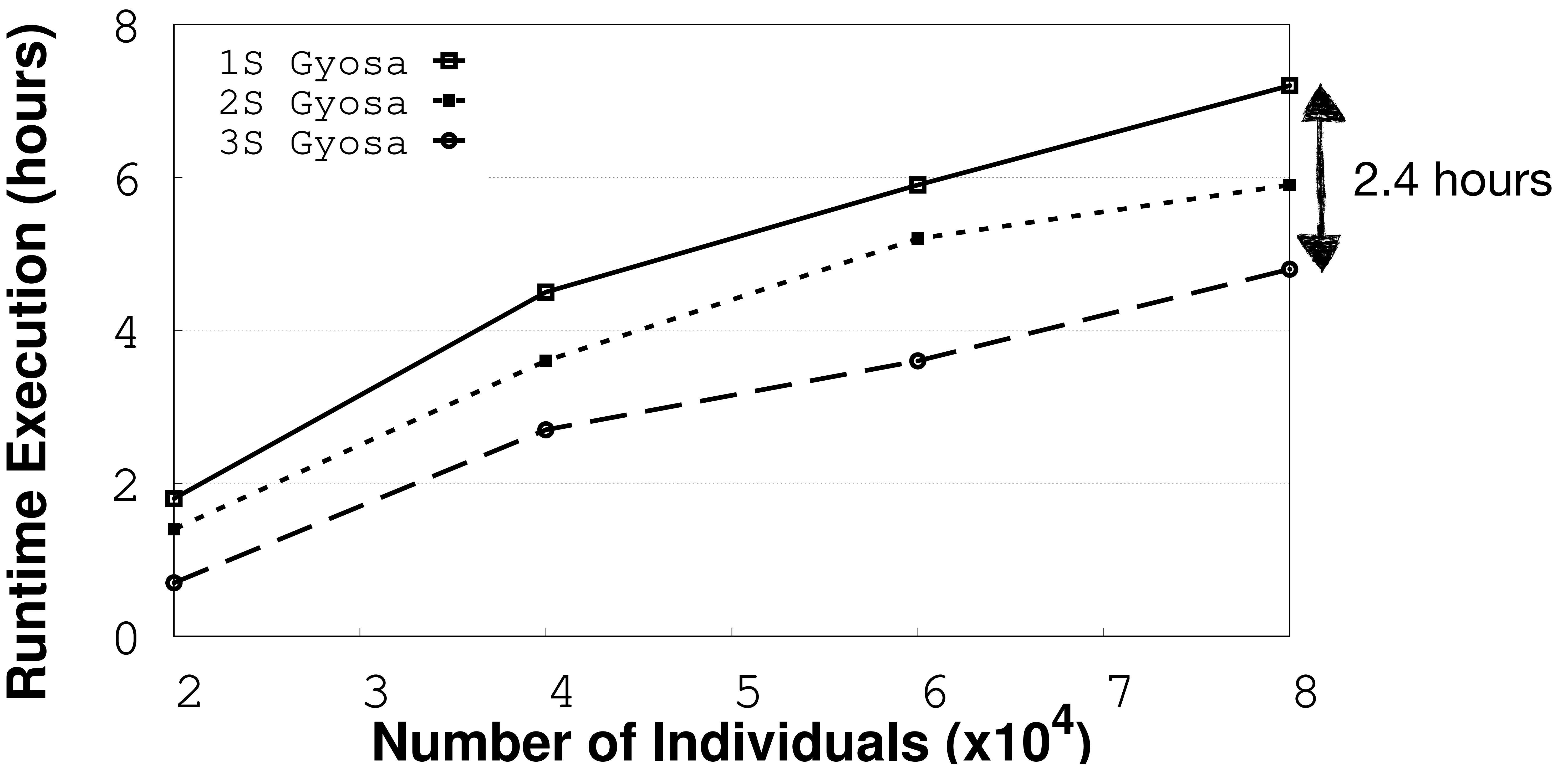
**Setup**

4 servers: Intel Core i5-9500 CPU, 16 GB RAM, and a 256GB NVMe
1-3 workers / 1 master/client

**Evaluation**

**Algorithms:**
- Linear Regression
- Logistic Regression
- $X^2$ Test

**Data Sizes:**
- 1GB - 32GB
- $2x10^4$ - $8x10^4$ individuals, with $1x10^6$ SNPs.




- The runtime execution decreases by 2.4 hours.
- No accuracy impact.

# GYOSA

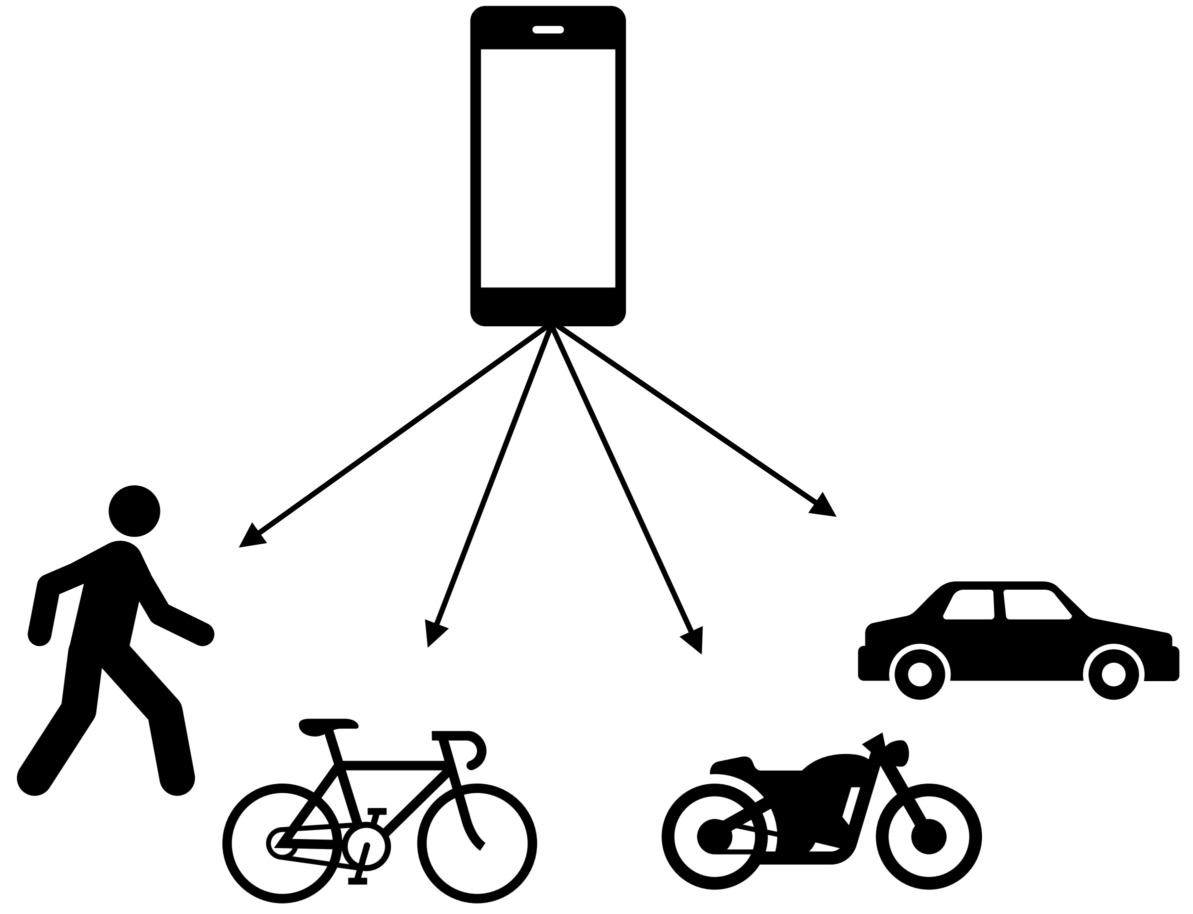
## SUMMARY

- **GYOSA** extends **SOTERIA’s** applicability with a tailored pipeline processing, e.g., for genomic association tests.
  - Offers the first **distributed SGX-based** solution for **genomic data**.

➡ How can we privacy-preserve data when considering **mobile devices** and **restricted hardware**?

# Federated Learning and Mobile Devices

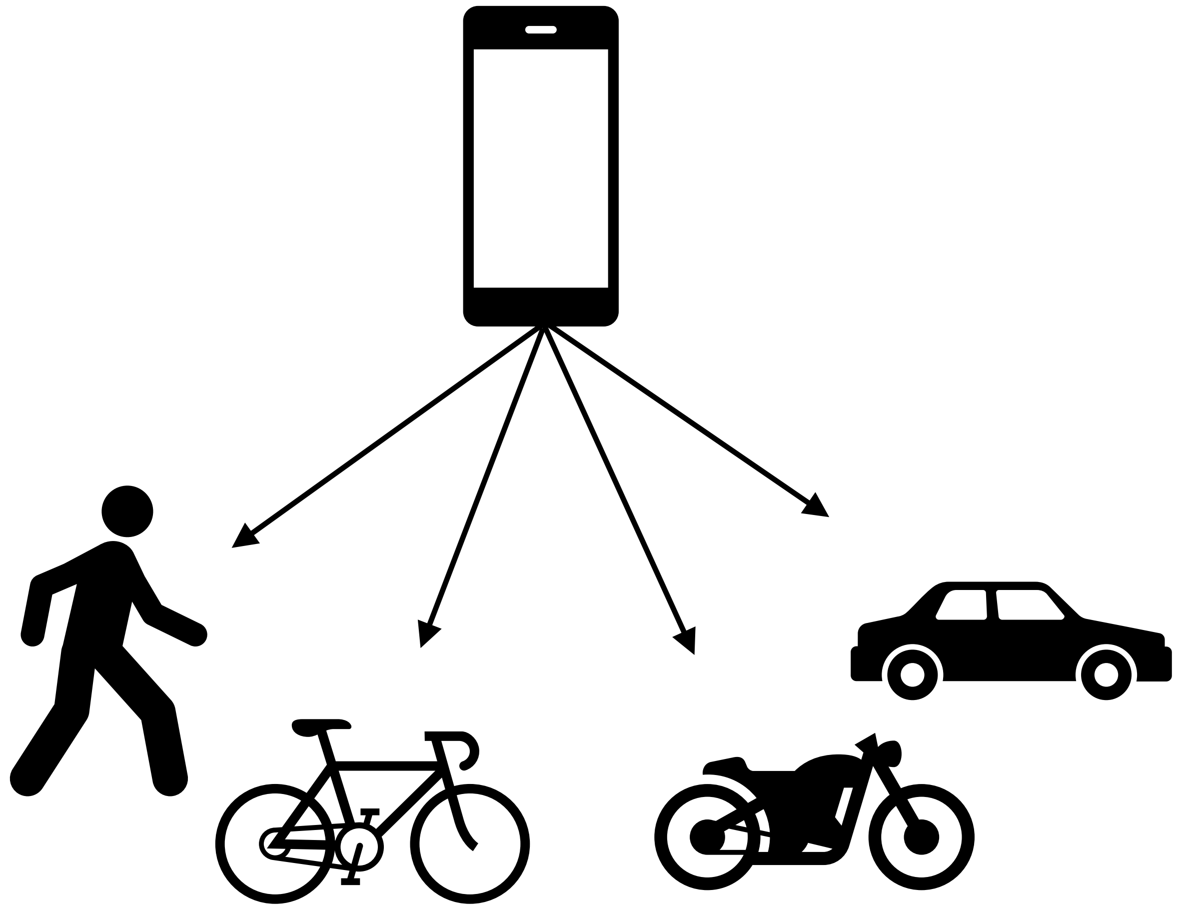
## Protecting User’s Mobility Patterns with Differential Privacy



# Federated Learning and Mobile Devices

## Protecting User's Mobility Patterns with Differential Privacy

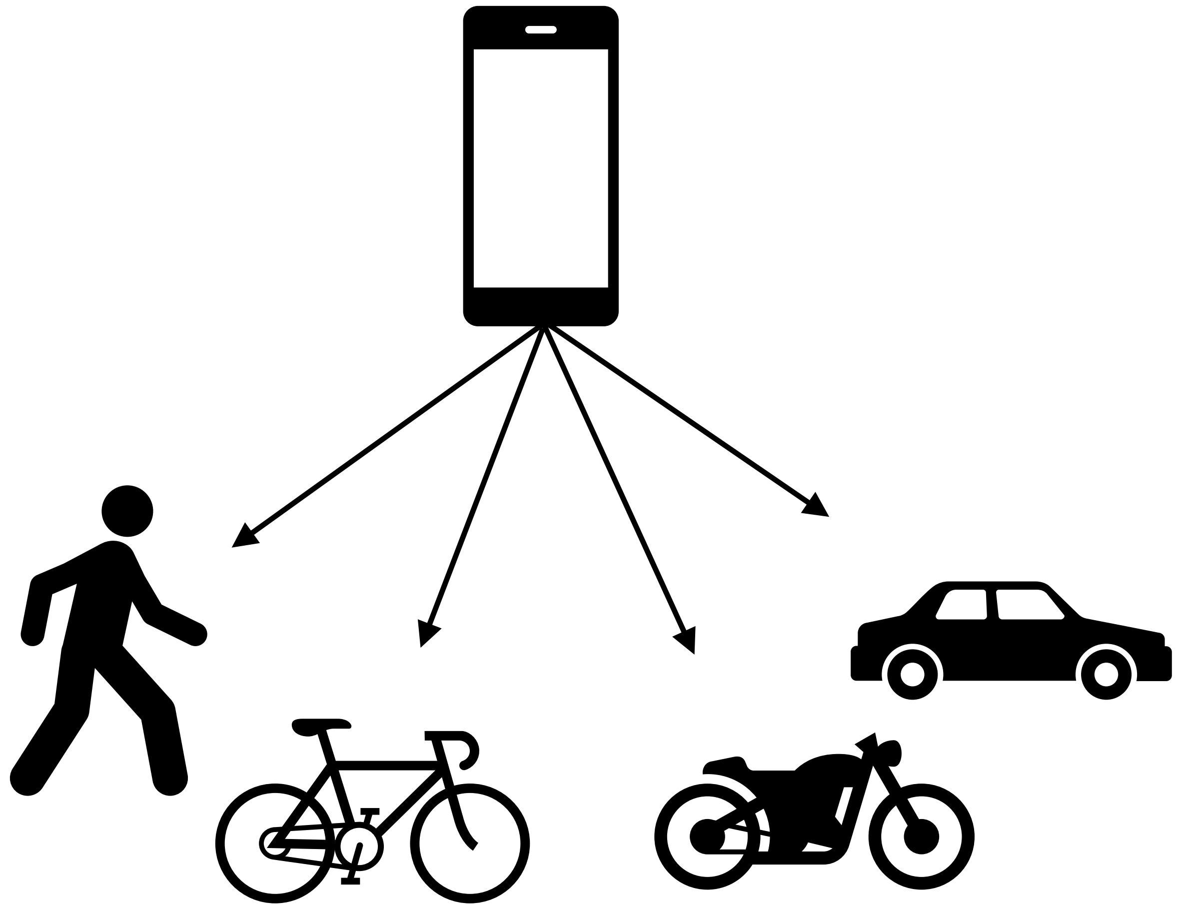
- Mobile devices collect several user-specific data.



# Federated Learning and Mobile Devices

## Protecting User's Mobility Patterns with Differential Privacy

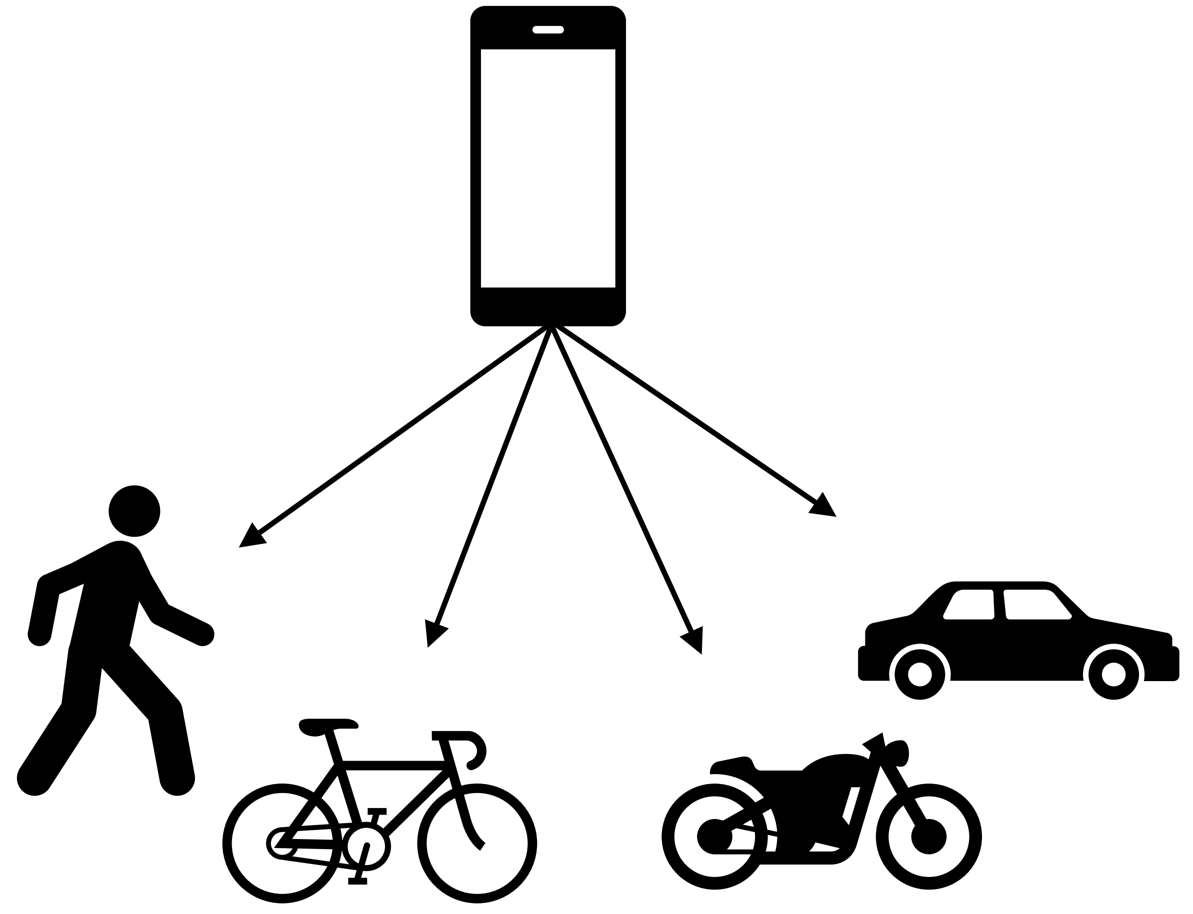
- Mobile devices collect several user-specific data.
- Mobile sensors can be used to understand user's mobility patterns.



# Federated Learning and Mobile Devices

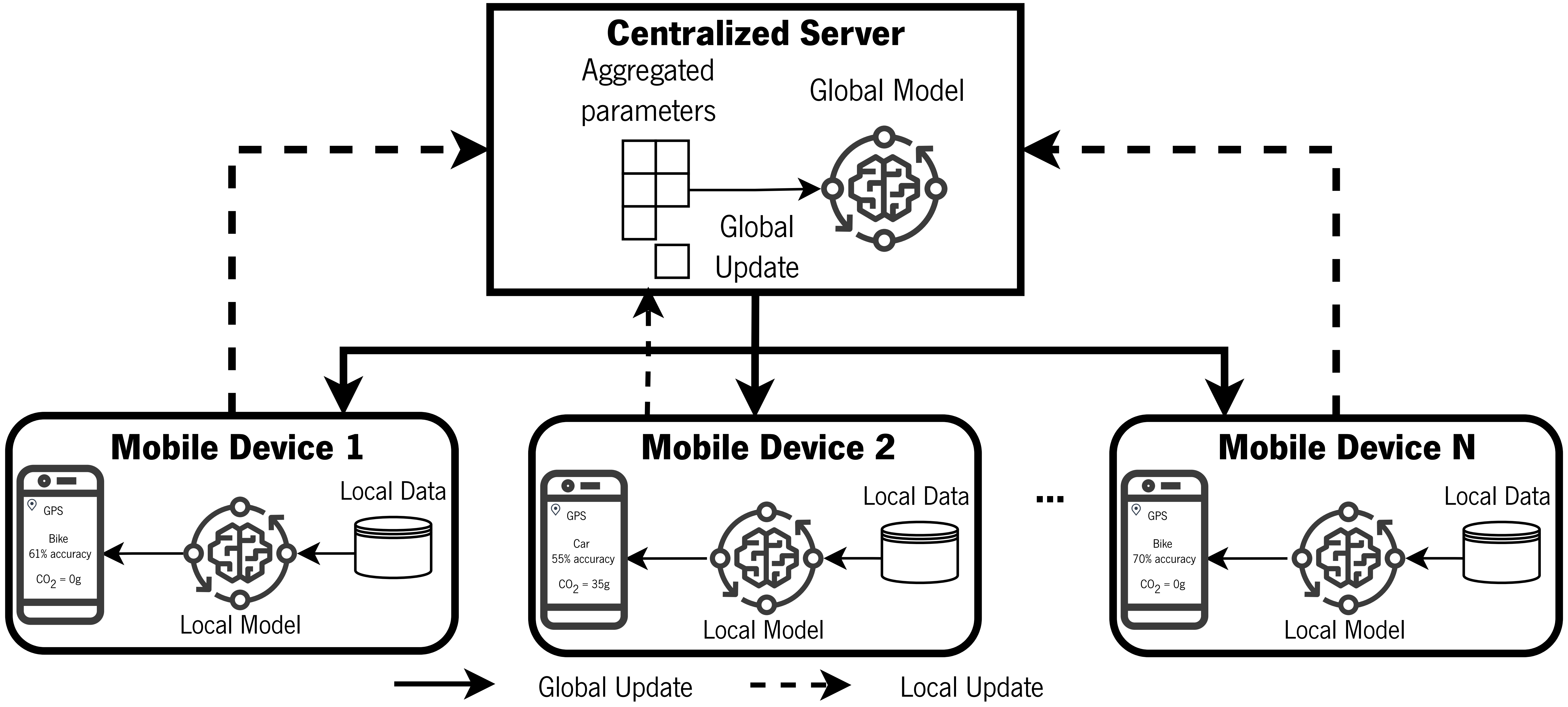
## Protecting User’s Mobility Patterns with Differential Privacy

- Mobile devices collect several user-specific data.
- Mobile sensors can be used to understand user’s mobility patterns.
- Lack of privacy measures.



# TAPUS

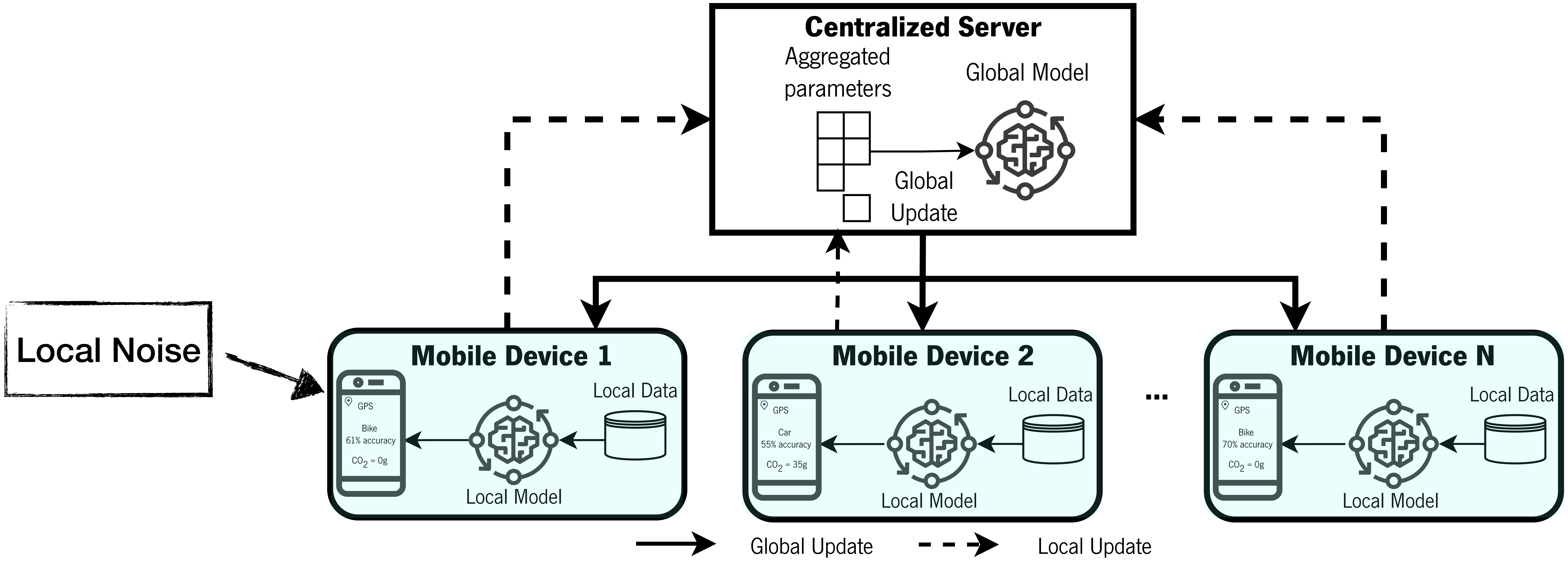
## Protecting User's Mobility Patterns with Differential Privacy




- AI model to identify transportation mode.
- DP-based noise added to local data or gradients.

# TAPUS

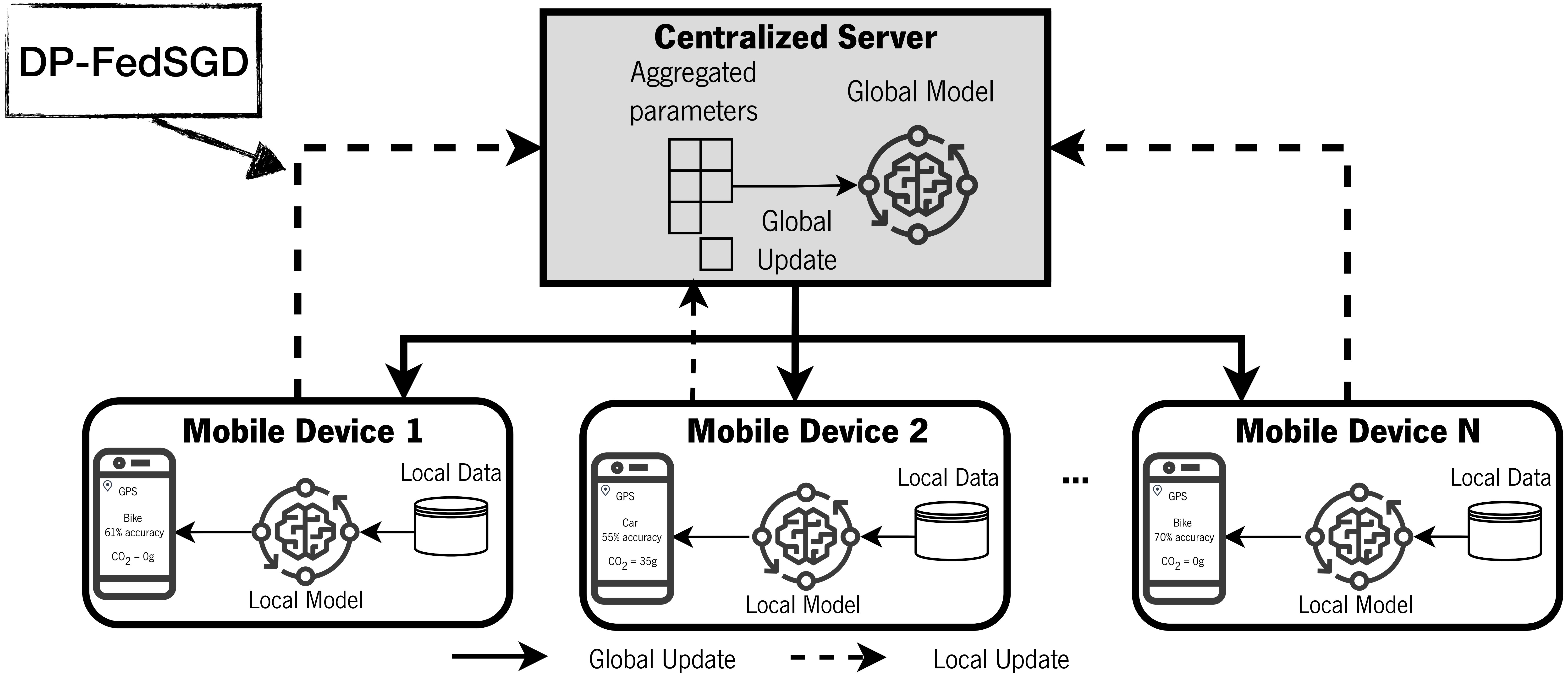
## Protecting User’s Mobility Patterns with Differential Privacy




- AI model to identify transportation mode.
- DP-based noise added to local data or gradients.

# TAPUS

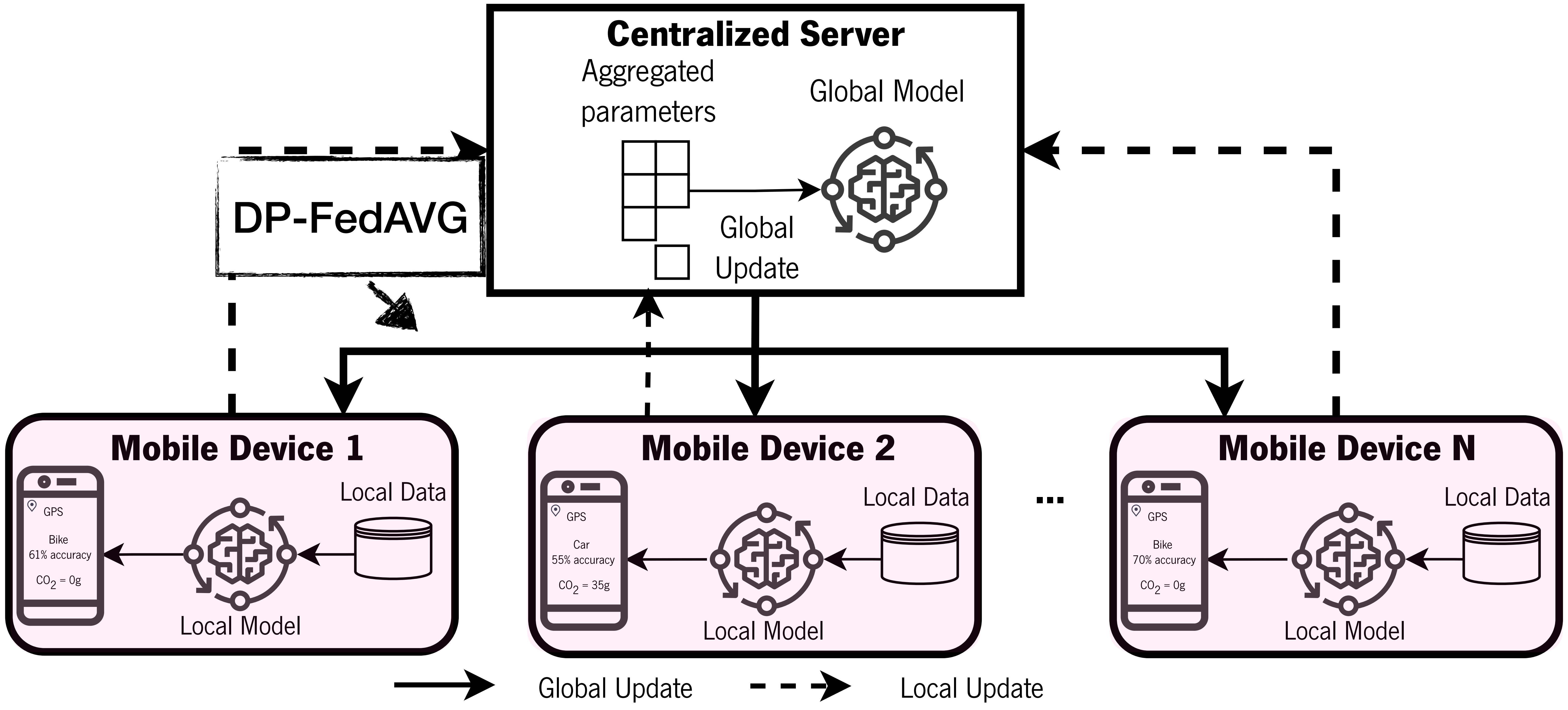
## Protecting User's Mobility Patterns with Differential Privacy




- AI model to identify transportation mode.
- DP-based noise added to local data or gradients.

# TAPUS

## Protecting User's Mobility Patterns with Differential Privacy



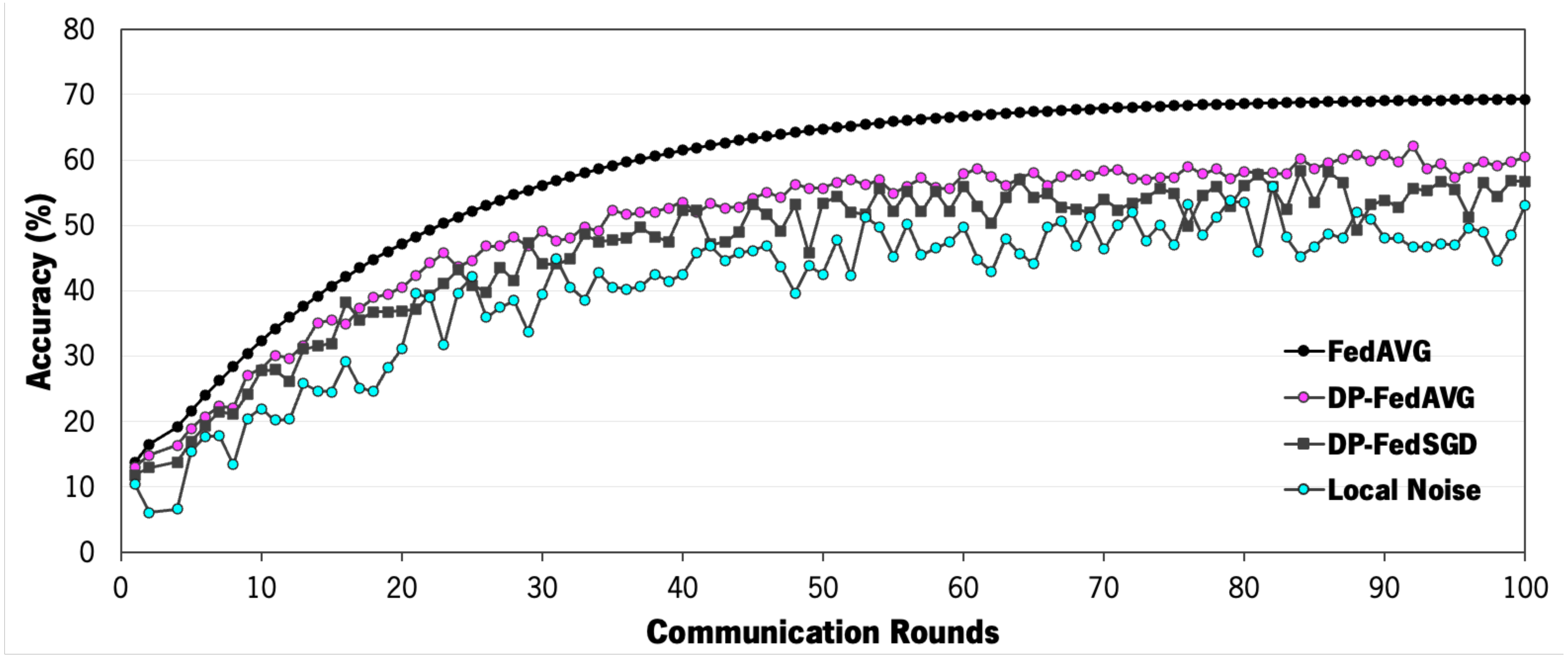

- AI model to identify transportation mode.
- DP-based noise added to local data or gradients.

# TAPUS

## RESULTS

**Setup**

50 clients: two vCPUs and 8 GB memory
1 Parameter: eight vCPUs and 32 GB memory
Epsilon: 0.3; Gradient Clipping: 0.1

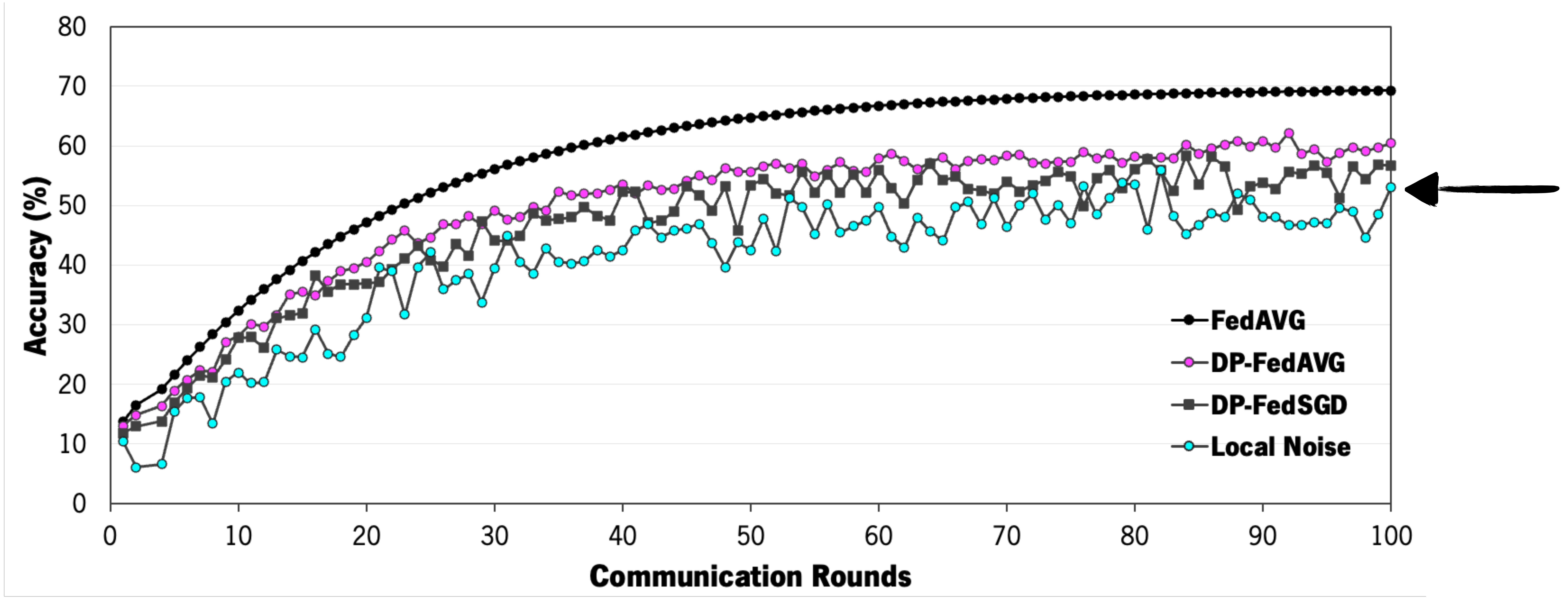


# TAPUS

## RESULTS

**Setup**

50 clients: two vCPUs and 8 GB memory
1 Parameter: eight vCPUs and 32 GB memory
Epsilon: 0.3; Gradient Clipping: 0.1



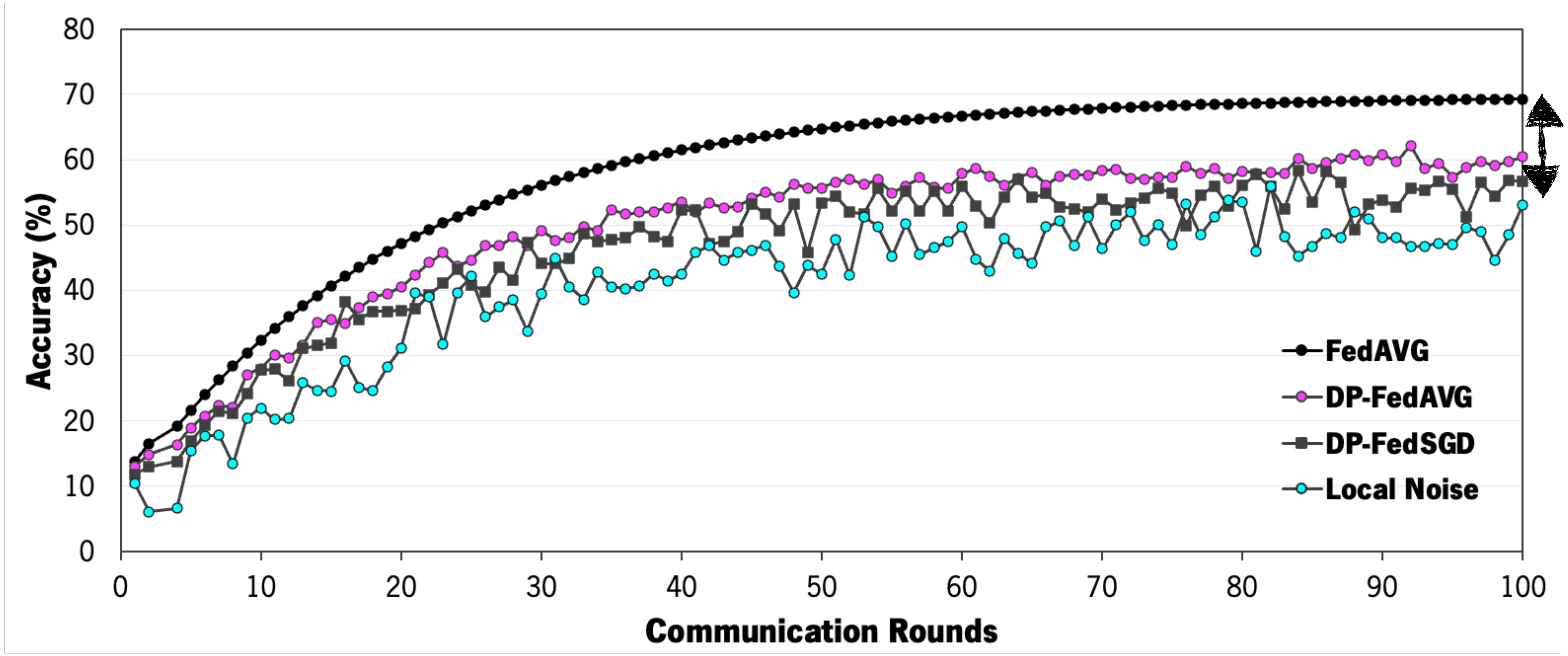

- Local noise significantly impacts the accuracy and quality of the model.

# TAPUS

## RESULTS

**Setup**

50 clients: two vCPUs and 8 GB memory
1 Parameter: eight vCPUs and 32 GB memory
Epsilon: 0.3; Gradient Clipping: 0.1




16% accuracy impact

- Local noise significantly impacts the accuracy and quality of the model.
- Increasing the number of clients, increases the overall noise and decreases the model's accuracy.

# TAPUS

## SUMMARY

- Relies on **DP-based mechanisms** to safeguard users from attacks.
- **Increasing** the amount of **noise** added to the gradients of the model's parameters **decreases** the model's **accuracy**.
- Although the **convergence rate** of the model is **maintained** with the increase in the number of clients, the **model's accuracy decreases** (up to 16%).

# Conclusion

# Conclusion

**Is it possible to balance privacy, performance, and utility in a PPDML solution?**

# Conclusion

**Is it possible to balance privacy, performance, and utility in a PPDML solution?**

1. **SOTERIA** presents a novel partitioning scheme for a distributed framework that guarantees the privacy of data while decreasing the performance overhead in cloud environments.

# Conclusion

**Is it possible to balance privacy, performance, and utility in a PPDML solution?**

1. **SOTERIA** presents a novel partitioning scheme for a distributed framework that guarantees the privacy of data while decreasing the performance overhead in cloud environments.

2. **GYOSA** shows that SOTERIA can be extended to specific use cases without impacting the accuracy of results.

# Conclusion

**Is it possible to balance privacy, performance, and utility in a PPDML solution?**

1. **SOTERIA** presents a novel partitioning scheme for a distributed framework that guarantees the privacy of data while decreasing the performance overhead in cloud environments.

2. **GYOSA** shows that SOTERIA can be extended to specific use cases without impacting the accuracy of results.

3. **TAPUS** explores different levels of privacy and their impact on the accuracy and quality of the models for mobile environments.

# Publications

**Core Publications:**

- **Brito, C.**, Ferreira, P., Portela, B., Oliveira, R. and Paulo, J. “SOTERIA: Preserving Privacy in Distributed Machine Learning.” In Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing, 2023.
- **Brito, C.**, Ferreira, P., Portela, B., Oliveira, R. and Paulo, J. “Privacy-Preserving Machine Learning on Apache Spark.” In IEEE Access, 2023.
- **Brito, C.**, Ferreira, P. and Paulo, J., “A Distributed Computing Solution for Privacy-Preserving Genome-Wide Association Studies.” Available as a preprint in bioRxiv and submitted for JBHI.
- Pina, N., **Brito, C.**, Vitorino, R., Cunha, I. “Promoting sustainable and personalized travel behaviors while preserving data privacy.” In Transportation Research Procedia - Proceedings of TRALisbon, 2022.
- **Brito, C.**, Pina, N., Esteves, T., Vitorino, R., Cunha, I., Paulo, J. “Promoting sustainable and personalized travel behaviors while preserving data privacy.” Accepted on Transportation Engineering (TRENG), 2024.

**Complementary publications:**

- Cepa, B., **Brito, C.** and Sousa, A. “Generative Adversarial Networks in Healthcare: A Case Study on MRI Image Generation.” In IEEE 7th Portuguese Meeting on Bioengineering (ENBENG), pp. 48-51, 2023.
- Alves, J., Soares, B., **Brito, C.**, and Sousa, A. “Cloud-Based Privacy-Preserving Medical Imaging System Using Machine Learning Tools.”. In EPIA Conference on Artificial Intelligence (pp. 195-206), 2022.
- Macedo, R., Correia, C., Dantas, M., **Brito, C.**, Xu, W., Tanimura, Y., Haga, J., Paulo, J. “The Case for Storage Optimization Decoupling in Deep Learning Frameworks.” In 1st Workshop on Re-envisioning Extreme-Scale I/O for Emerging Hybrid HPC Workloads, co-located with IEEE International Conference in Cluster Computing, 2021.
- **Brito, C.**, Machado, M. and Sousa, A. “Electrocardiogram Beat-Classification Based on a ResNet Network.” In MedInfo (pp. 55-59), 2019.