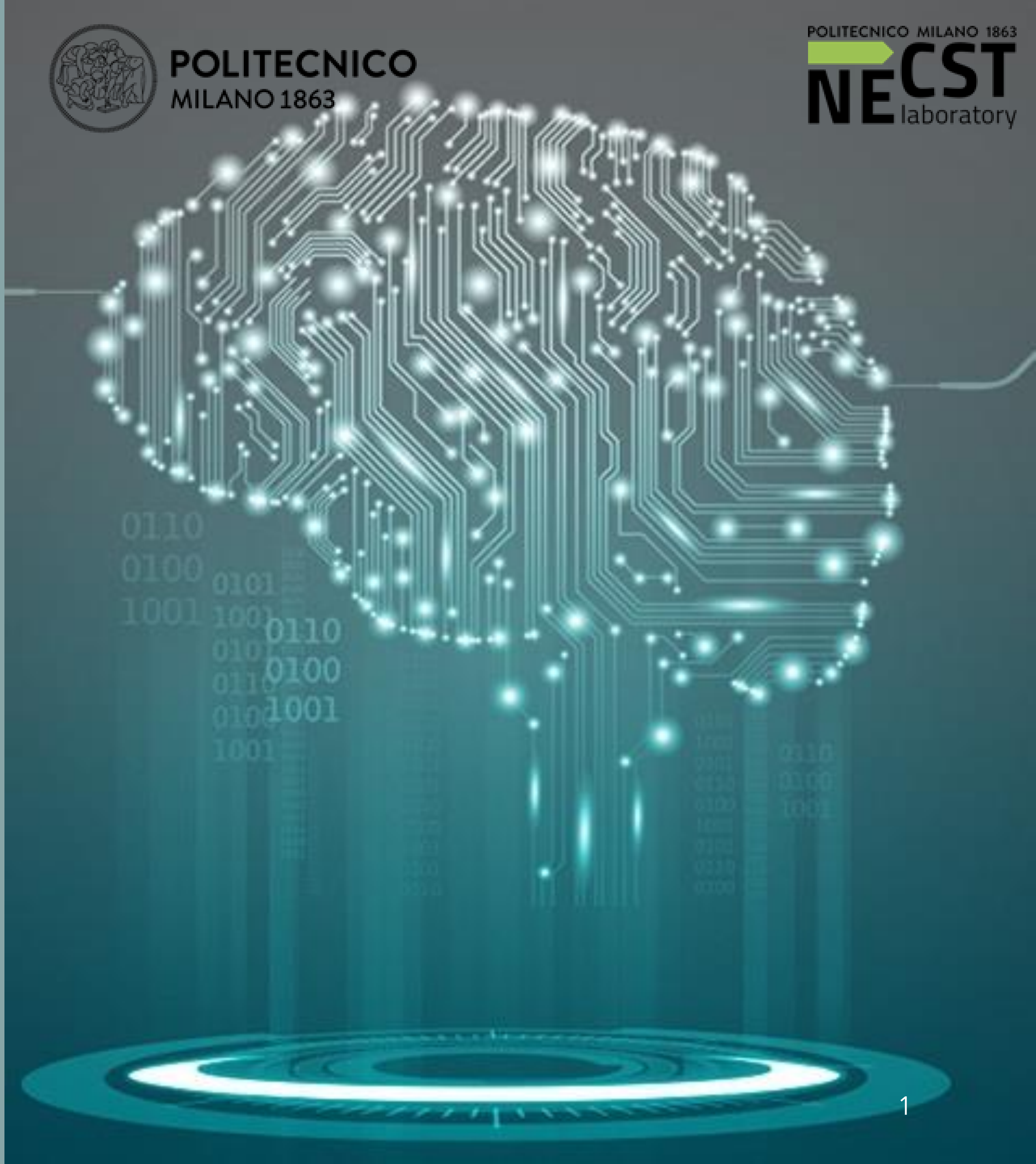
Creativity, Science and **Innovation**

Introduction to Federated Learning

December 1st, 2025

Alessandro Verosimile
alessandro.verosimile@polimi.it



Machine Learning Pipeline

Steps to train a ML model:

1. **Collect a dataset** with the data of interest. Each user (client) collects their own data to build a unique big dataset for the task.

1

X	Y	Z	Label
[2,-1,4,...]	[12,5,8,...]	[9,5,8,...]	Walking (0)
[2,-2,4,...]	[11,2,9,...]	[4,1,5,...]	Running (1)
[1,3,-1,...]	[10,5,7,...]	[4,1,5,...]	Walking (0)
[2,-1,1,...]	[8,11,8,...]	[5,5,7,...]	Walking (0)
[1,5,3,...]	[9,5,8,...]	[8,2,8,...]	Going up (2)
[2,-1,4,...]	[5,5,7,...]	[2,3,4,...]	Walking (0)
[4,5,9,...]	[4,5,9,...]	[4,1,5,...]	Walking (0)

Machine Learning Pipeline

Steps to train a ML model:

1. **Collect a dataset** with the data of interest. Each user (client) collects their own data to build a unique big dataset for the task.

2. **Train** an ML model on such data. The model will learn to predict the Label given the inputs

1

X	Y	Z	Label
[2,-1,4,...]	[12,5,8,...]	[9,5,8,...]	Walking (0)
[2,-2,4,...]	[11,2,9,...]	[4,1,5,...]	Running (1)
[1,3,-1,...]	[10,5,7,...]	[4,1,5,...]	Walking (0)
[2,-1,1,...]	[8,11,8,...]	[5,5,7,...]	Walking (0)
[1,5,3,...]	[9,5,8,...]	[8,2,8,...]	Going up (2)
[2,-1,4,...]	[5,5,7,...]	[2,3,4,...]	Walking (0)
[4,5,9,...]	[4,5,9,...]	[4,1,5,...]	Walking (0)

2

ML model

POLITECNICO MILANO 1863

Machine Learning Pipeline

Steps to train a ML model:

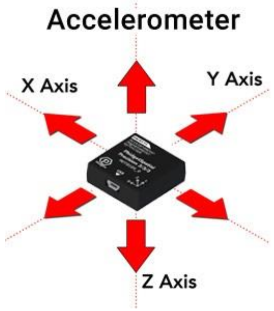
1. **Collect a dataset** with the data of interest. Each user (client) collects their own data to build a unique big dataset for the task.
2. **Train** an ML model on such data. The model will learn to predict the Label given the inputs
3. When you have the final model, you will be able to use it in **inference** mode on new data coming from the device

1

X	Y	Z	Label
[2,-1,4,...]	[12,5,8,...]	[9,5,8,...]	Walking (0)
[2,-2,4,...]	[11,2,9,...]	[4,1,5,...]	Running (1)
[1,3,-1,...]	[10,5,7,...]	[4,1,5,...]	Walking (0)
[2,-1,1,...]	[8,11,8,...]	[5,5,7,...]	Walking (0)
[1,5,3,...]	[9,5,8,...]	[8,2,8,...]	Going up (2)
[2,-1,4,...]	[5,5,7,...]	[2,3,4,...]	Walking (0)
[4,5,9,...]	[4,5,9,...]	[4,1,5,...]	Walking (0)

2

ML model



3

Prediction

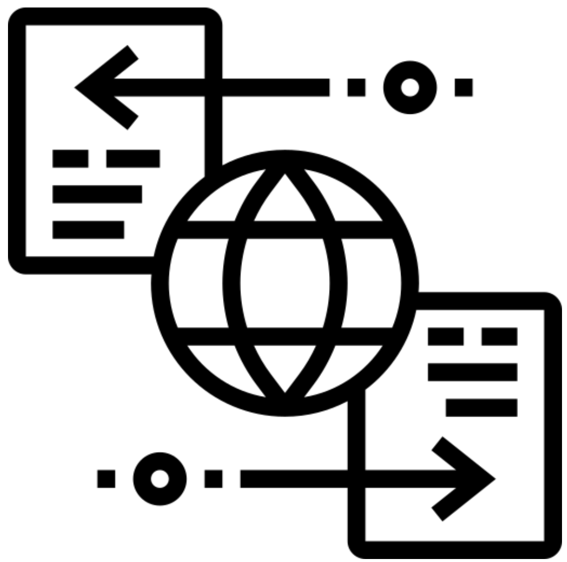
Why Federated Learning?

Why Federated Learning?

Due to new **regulations** (GDPR), to protect users' **privacy**, sensitive data cannot be shared

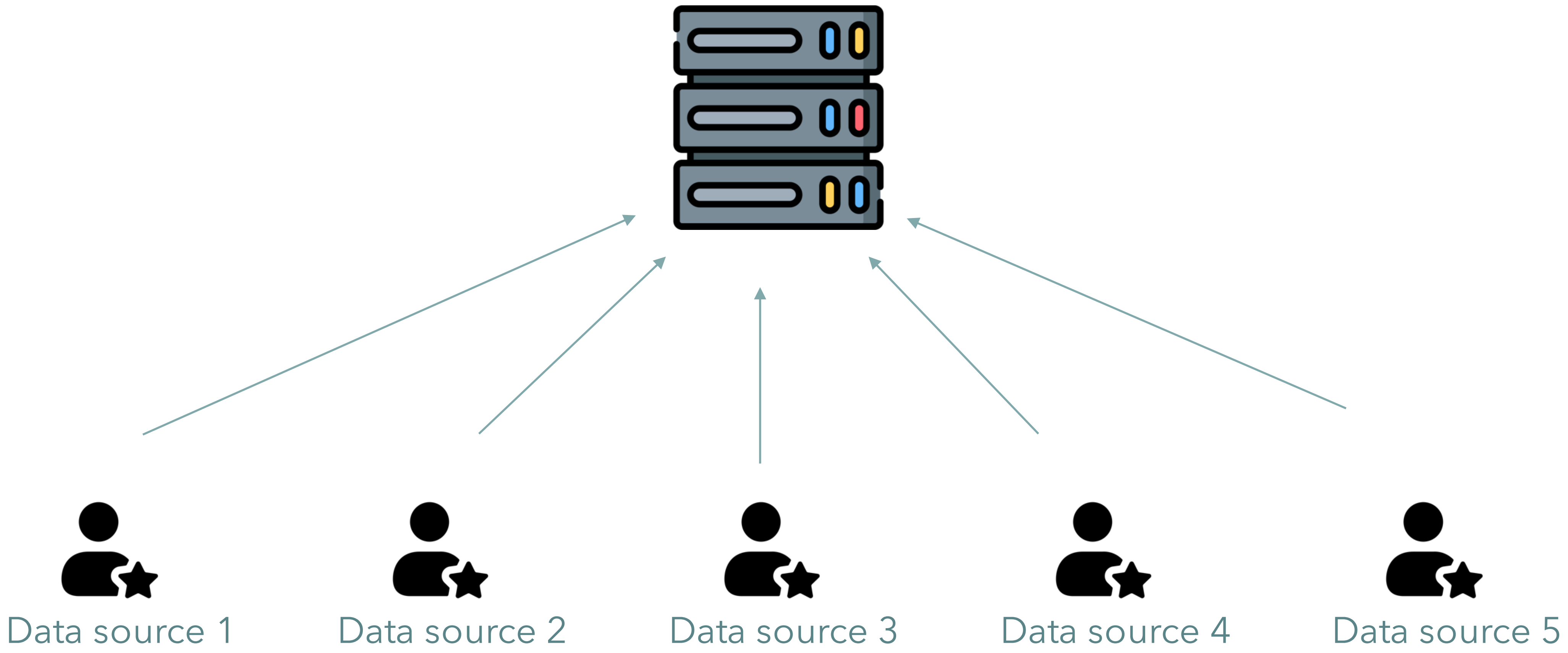
Why Federated Learning?

Due to new **regulations** (GDPR), to protect users' **privacy**, sensitive data cannot be shared

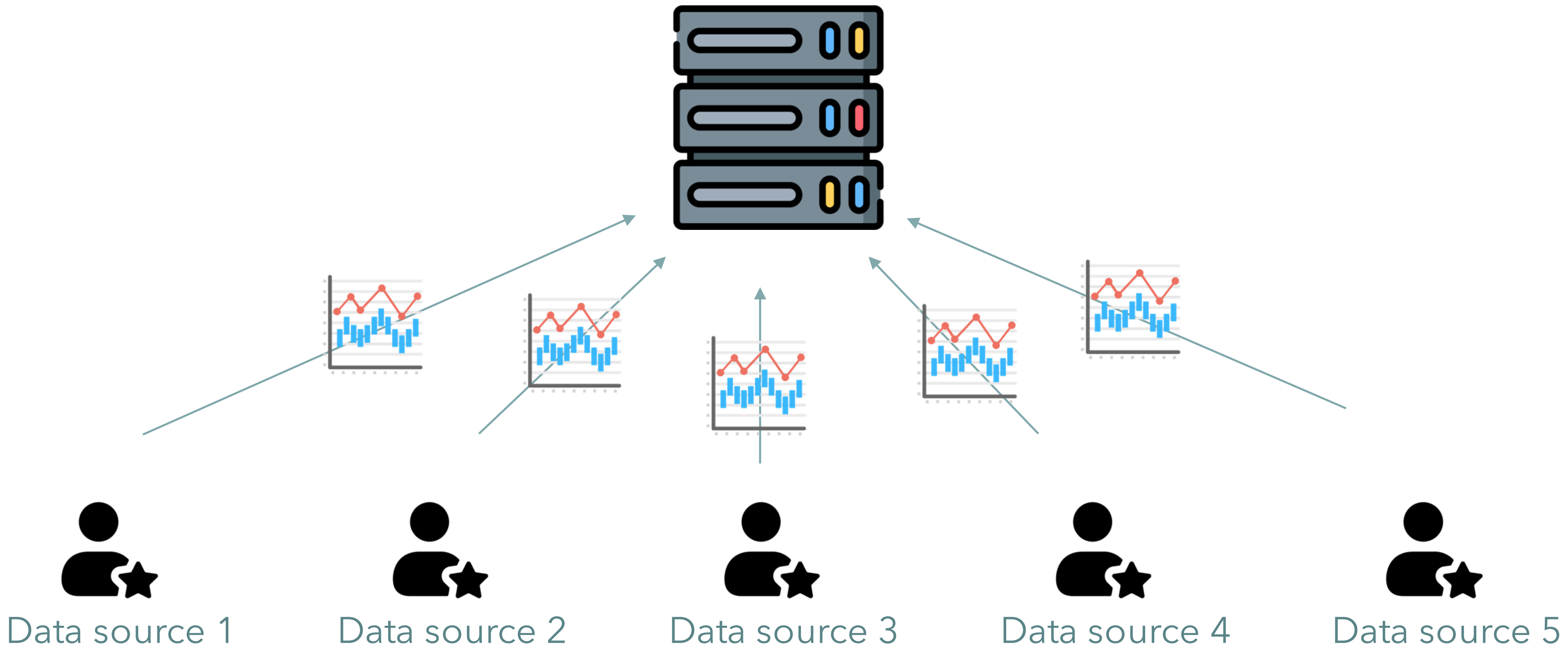


Communication overhead: when dealing with **heavy data**, such as images or volumes for medical data, data transfer is an issue

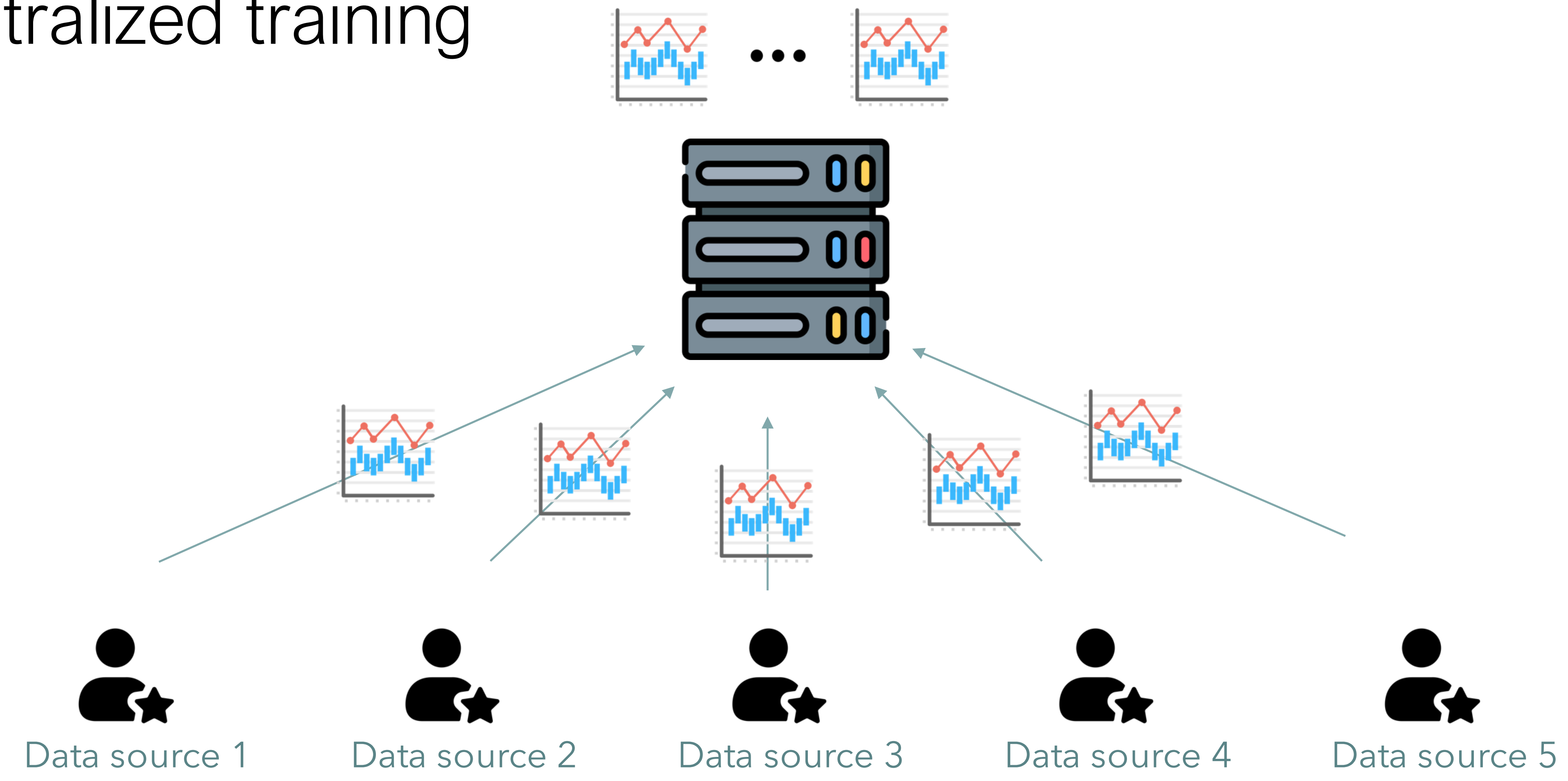
Centralized training



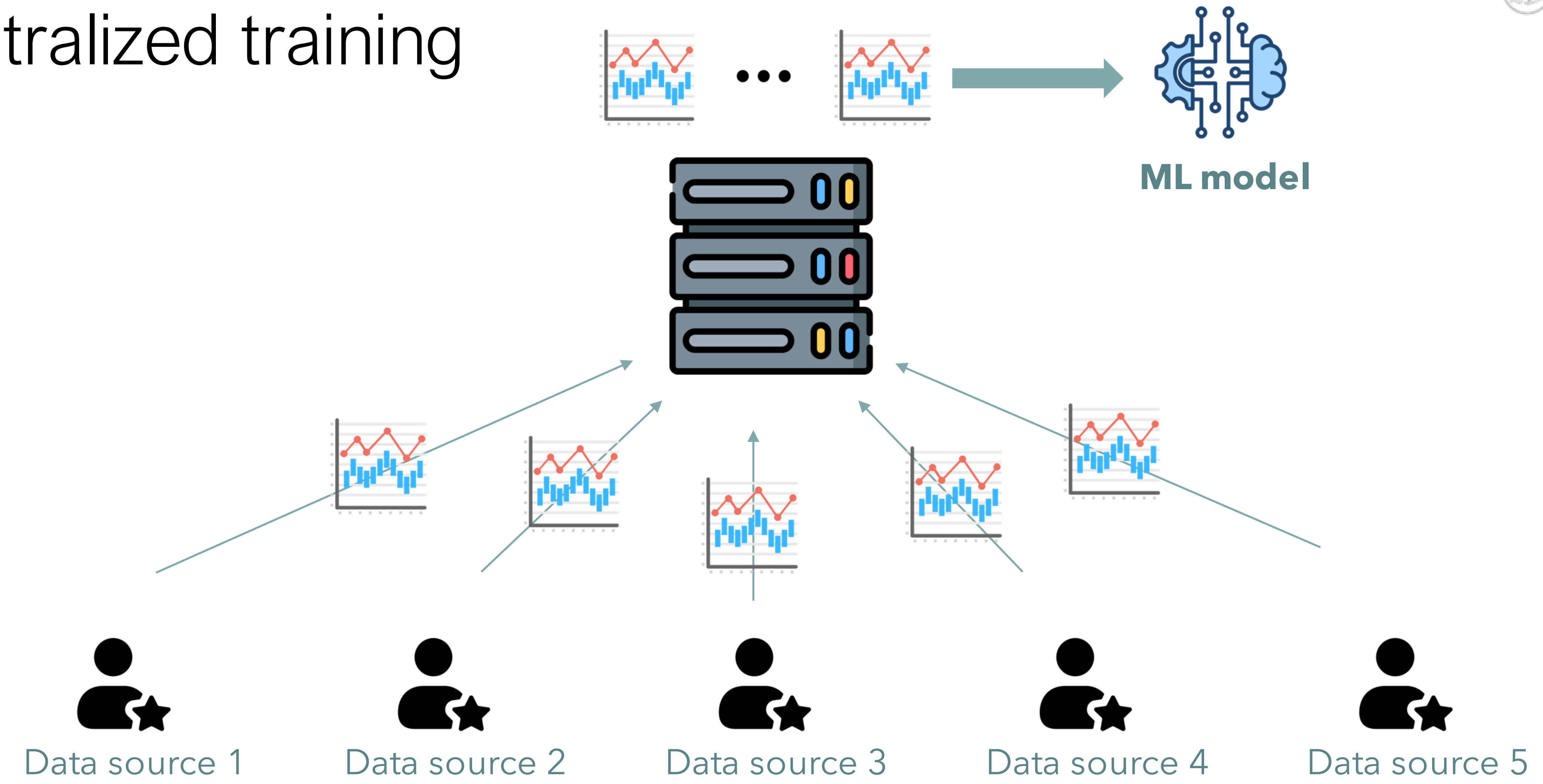
Centralized training



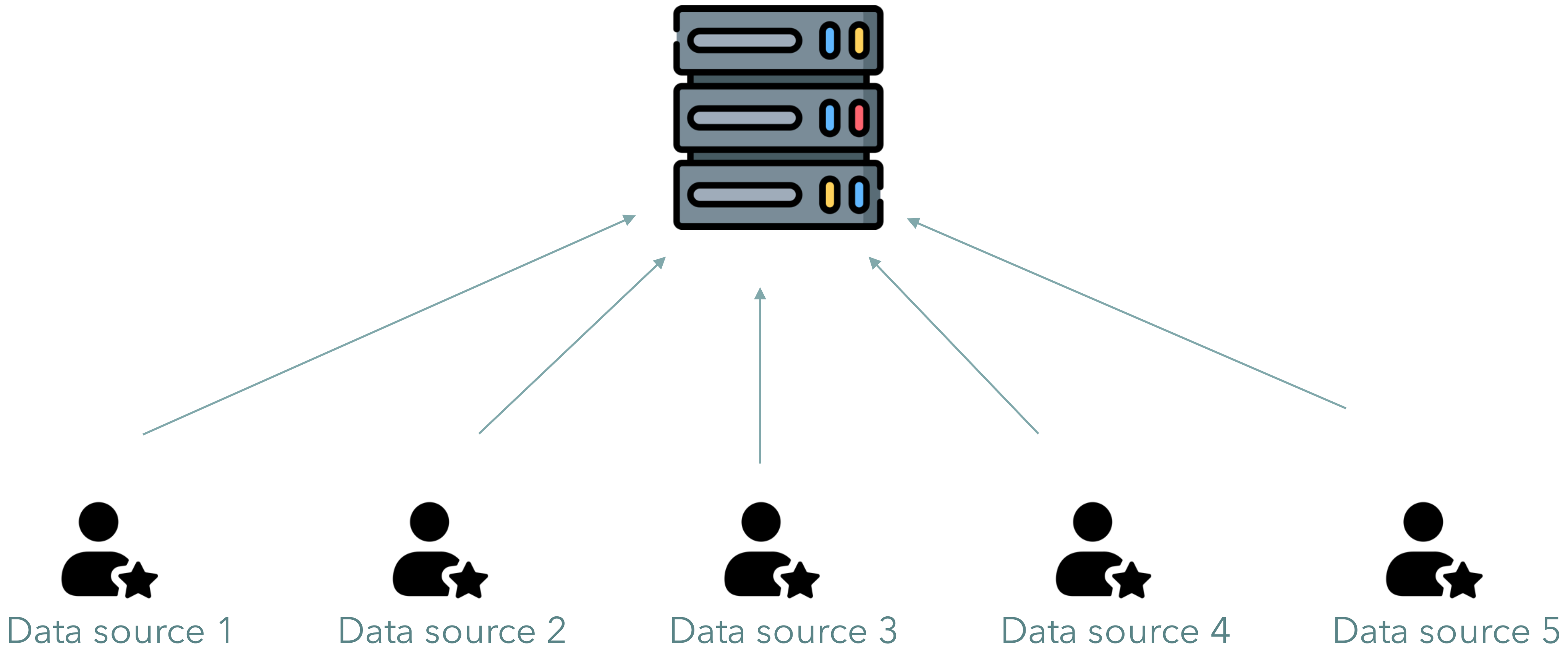
Centralized training



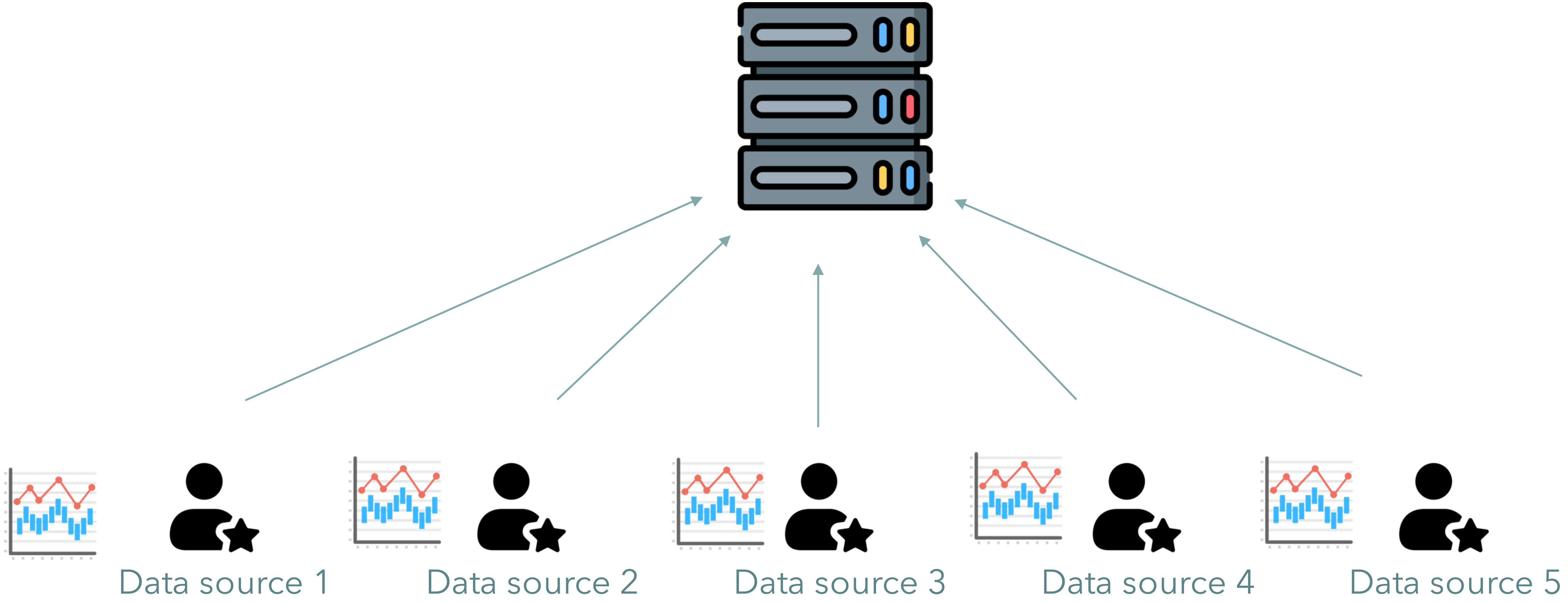
Centralized training



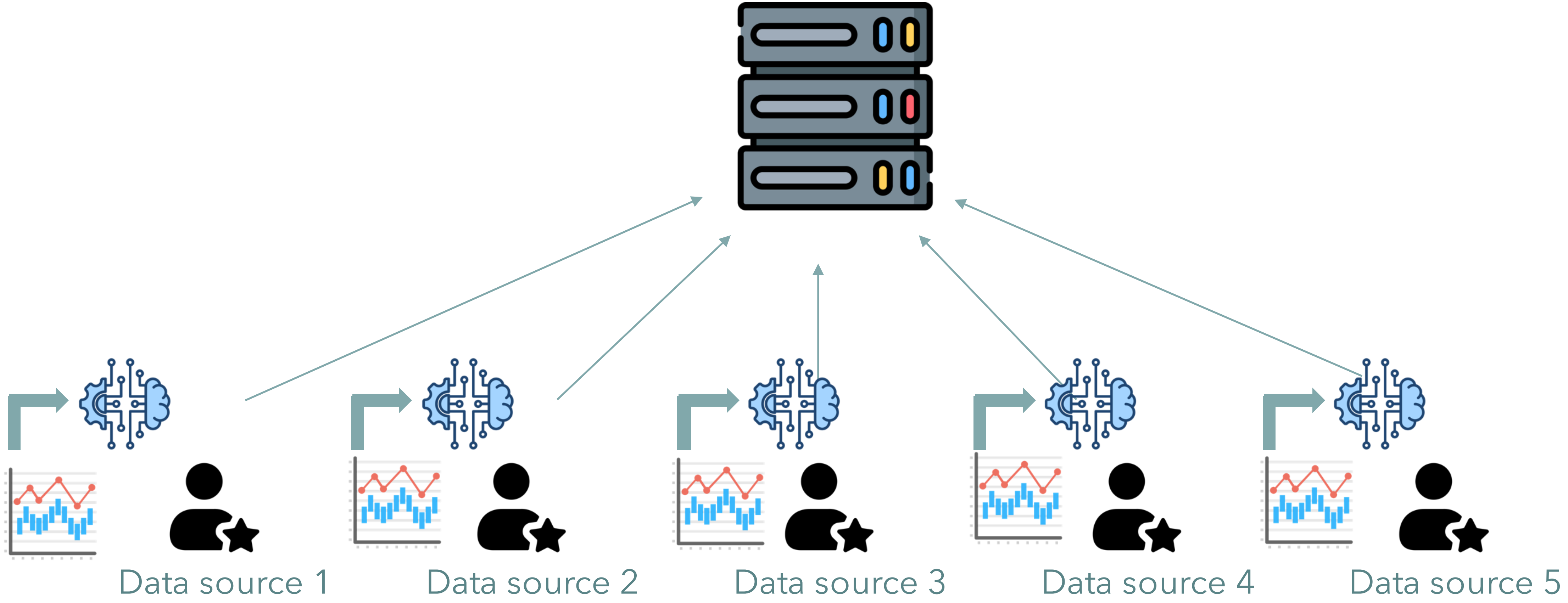
Federated Learning training



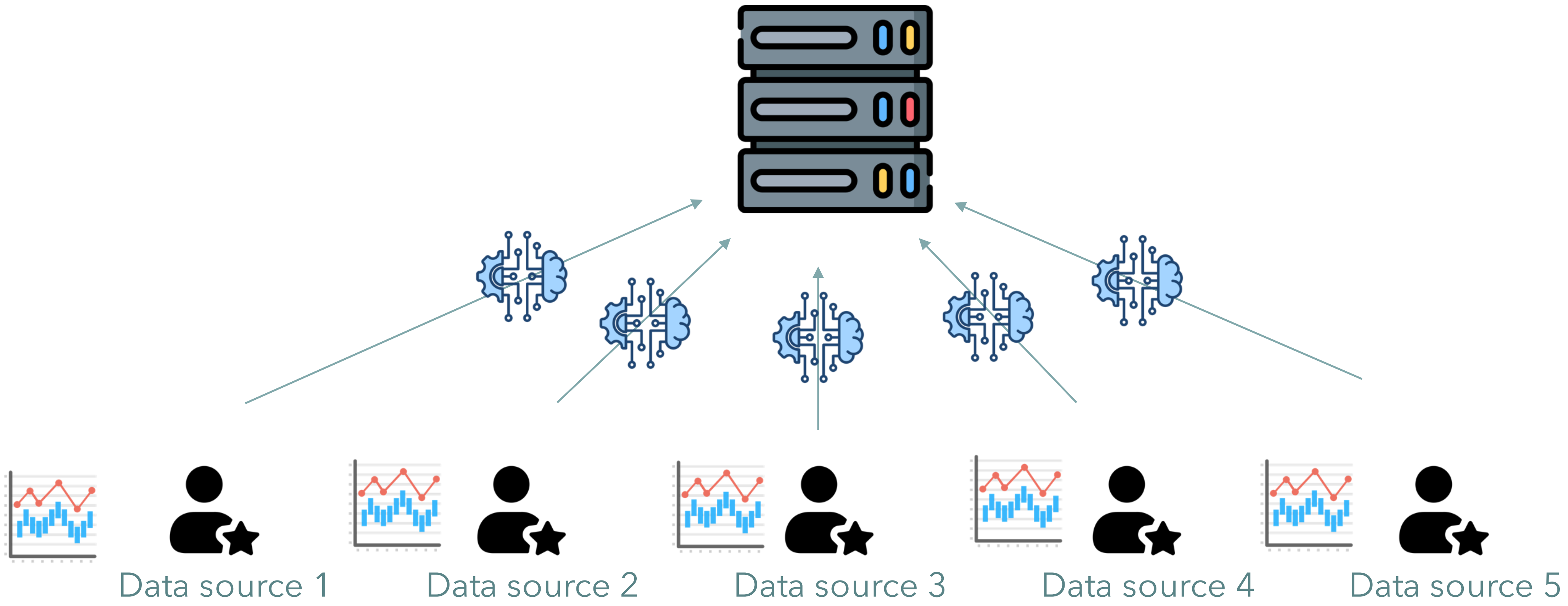
Federated Learning training



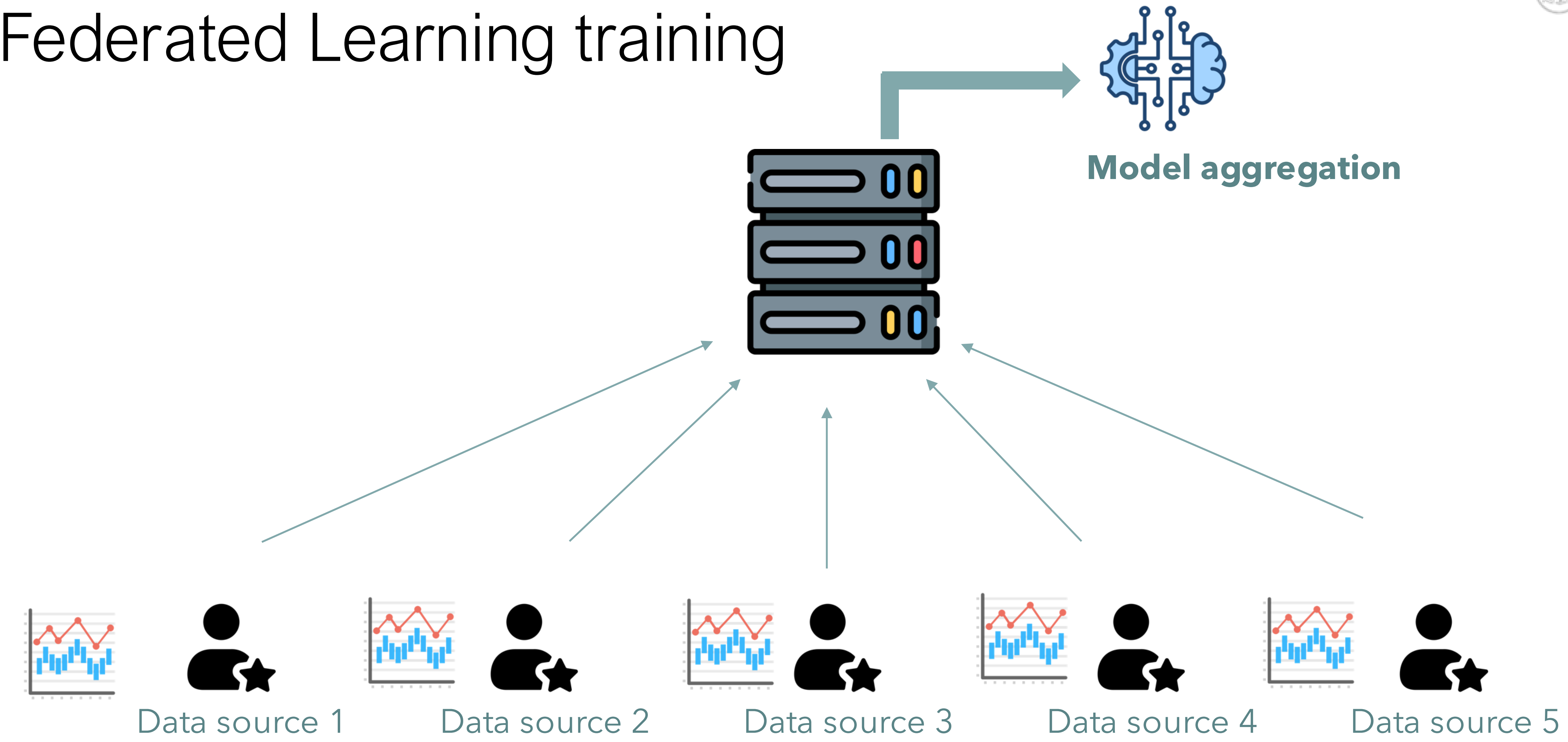
Federated Learning training



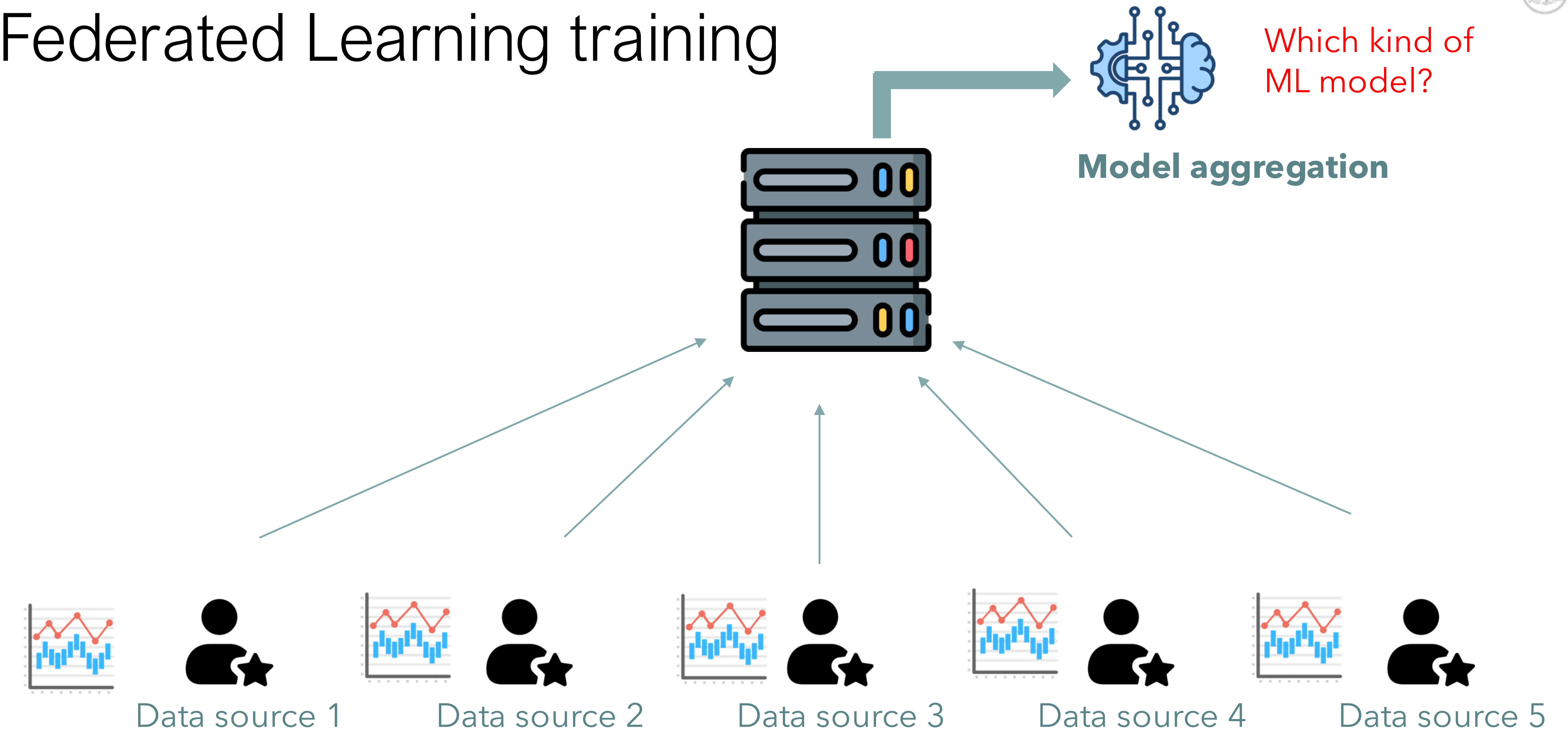
Federated Learning training



Federated Learning training



Federated Learning training



Which are the characteristics of the data over the different clients?

Federated Learning Setup

Model type

- Neural Networks
- Decision-Tree based Ensemble

Based on the model type, the way in which the federated training works is different

Data characteristics

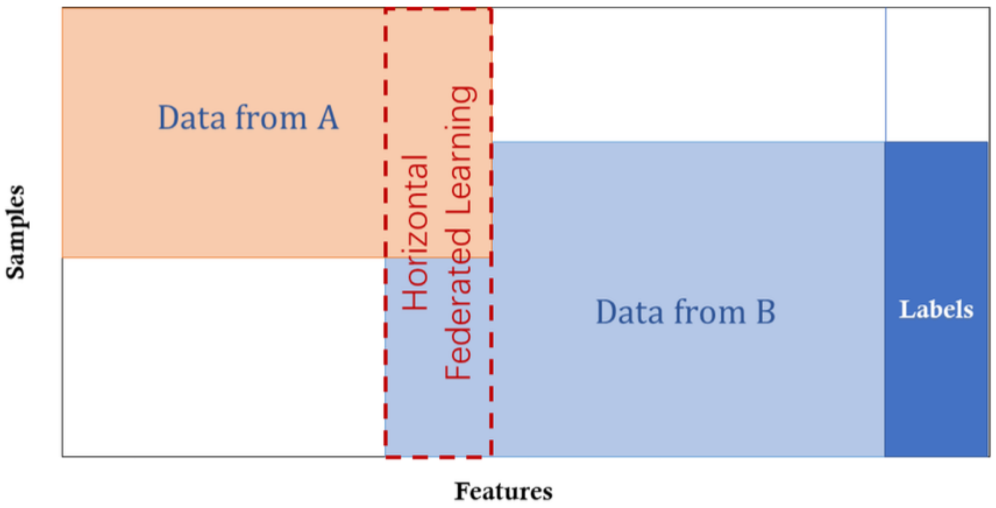
- Different data sources share the same features space: **Horizontal FL**
- Different data sources share the same sample space: **Vertical FL**

Federated Learning Setup

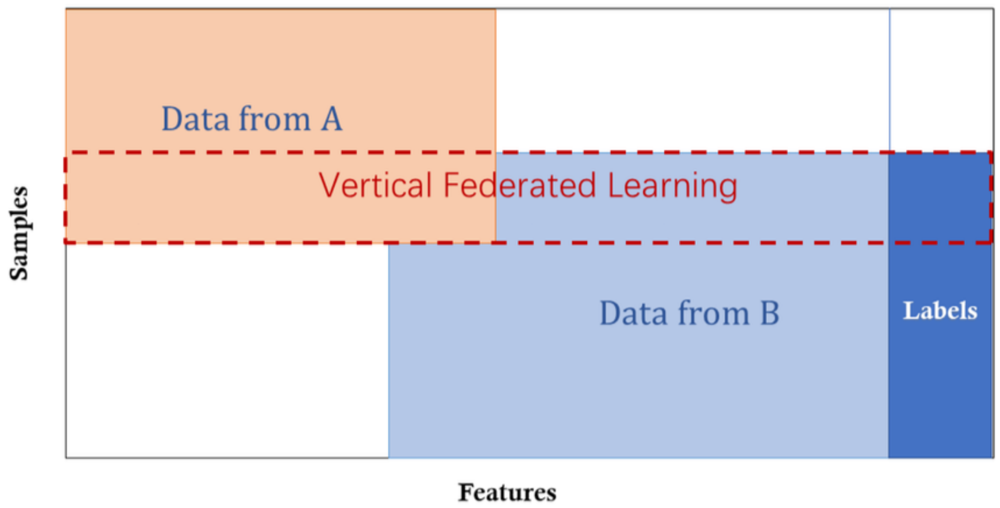
Data characteristics

Horizontal FL : Different data sources share the same features space

Vertical FL: Different data sources share the same sample space



(a) Horizontal Federated Learning



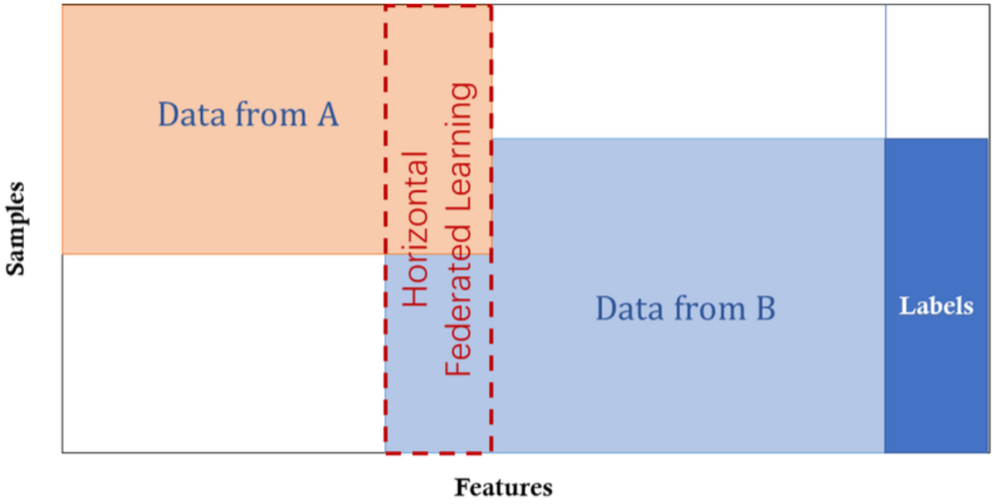
(b) Vertical Federated Learning

[1] Yang, Qiang, et al. "Federated machine learning: Concept and applications." *ACM Transactions on Intelligent Systems and Technology (TIST)* 10.2 (2019): 1-19.

Federated Learning Setup

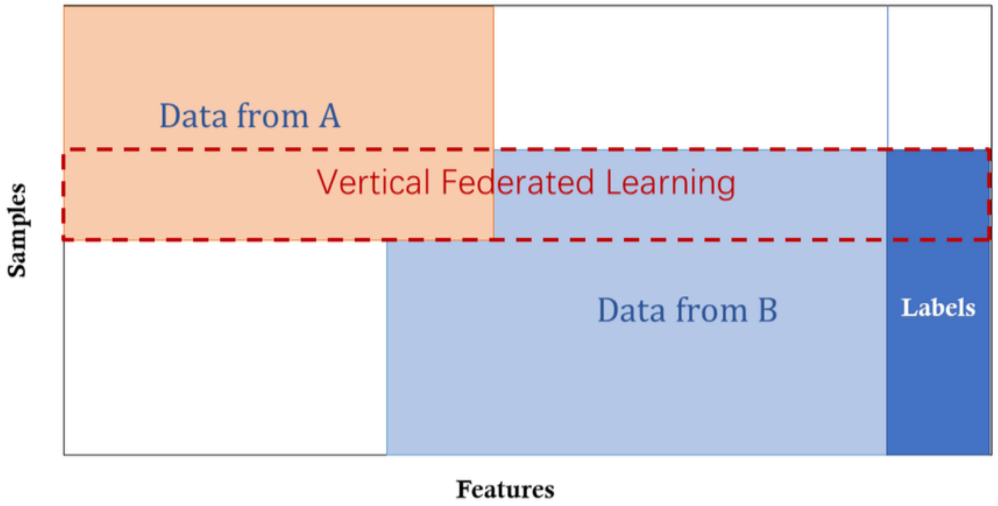
Data characteristics

Horizontal FL : Different data sources share the same features space



(a) Horizontal Federated Learning

Vertical FL: Different data sources share the same sample space



(b) Vertical Federated Learning

[1] Yang, Qiang, et al. "Federated machine learning: Concept and applications." *ACM Transactions on Intelligent Systems and Technology (TIST)* 10.2 (2019): 1-19.

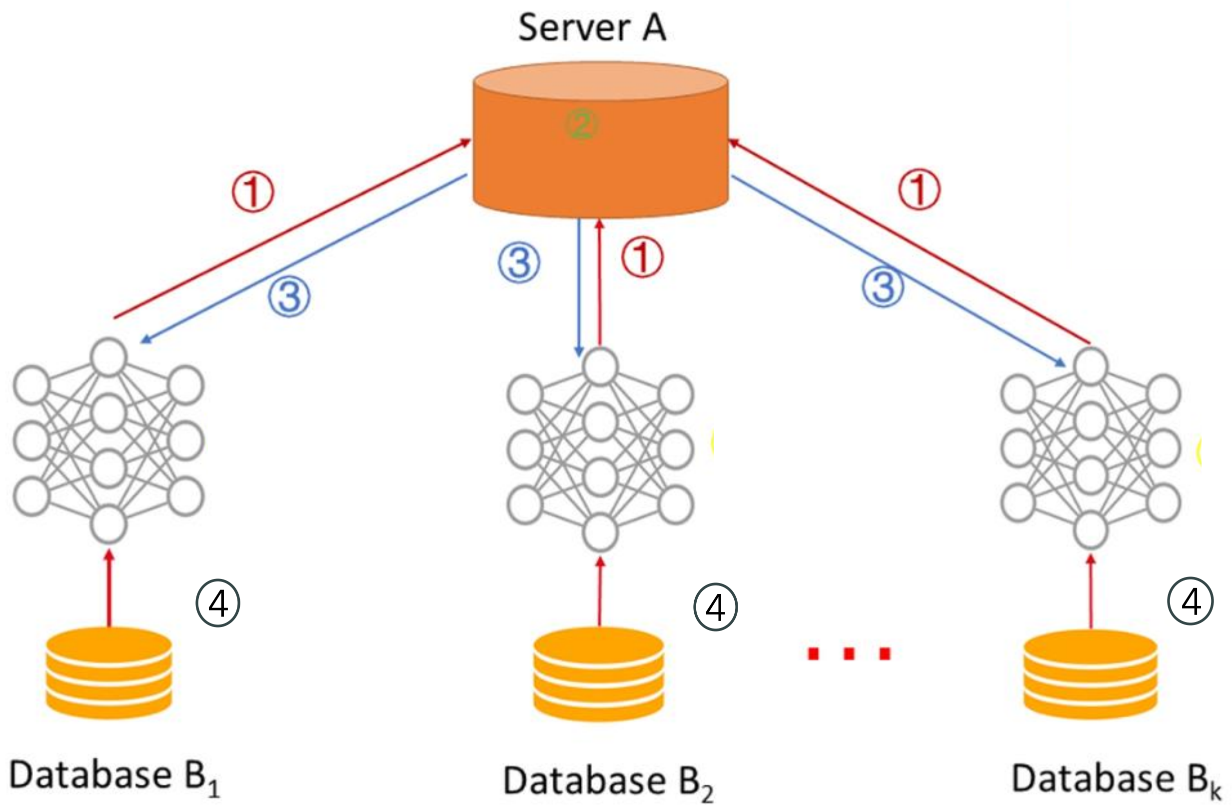
Horizontal Federated Learning – Neural Networks

1. Weights (or only gradients) are **encrypted** and **sent** from each client to the central server

2. The server **aggregates** the models with a pre-defined strategy (FedAvg)

3. The aggregated model is sent back to the clients

4. The clients can **evaluate** the aggregated model over their own data



[1] Yang, Qiang, et al. "Federated machine learning: Concept and applications." *ACM Transactions on Intelligent Systems and Technology (TIST)* 10.2 (2019): 1-19.

Horizontal Federated Learning – Performance

Typically, Federated models improve the performance of models trained only on local data, but they rarely reach the performance of a centralized training.

FL is a trade-off between **privacy** and **model performance**

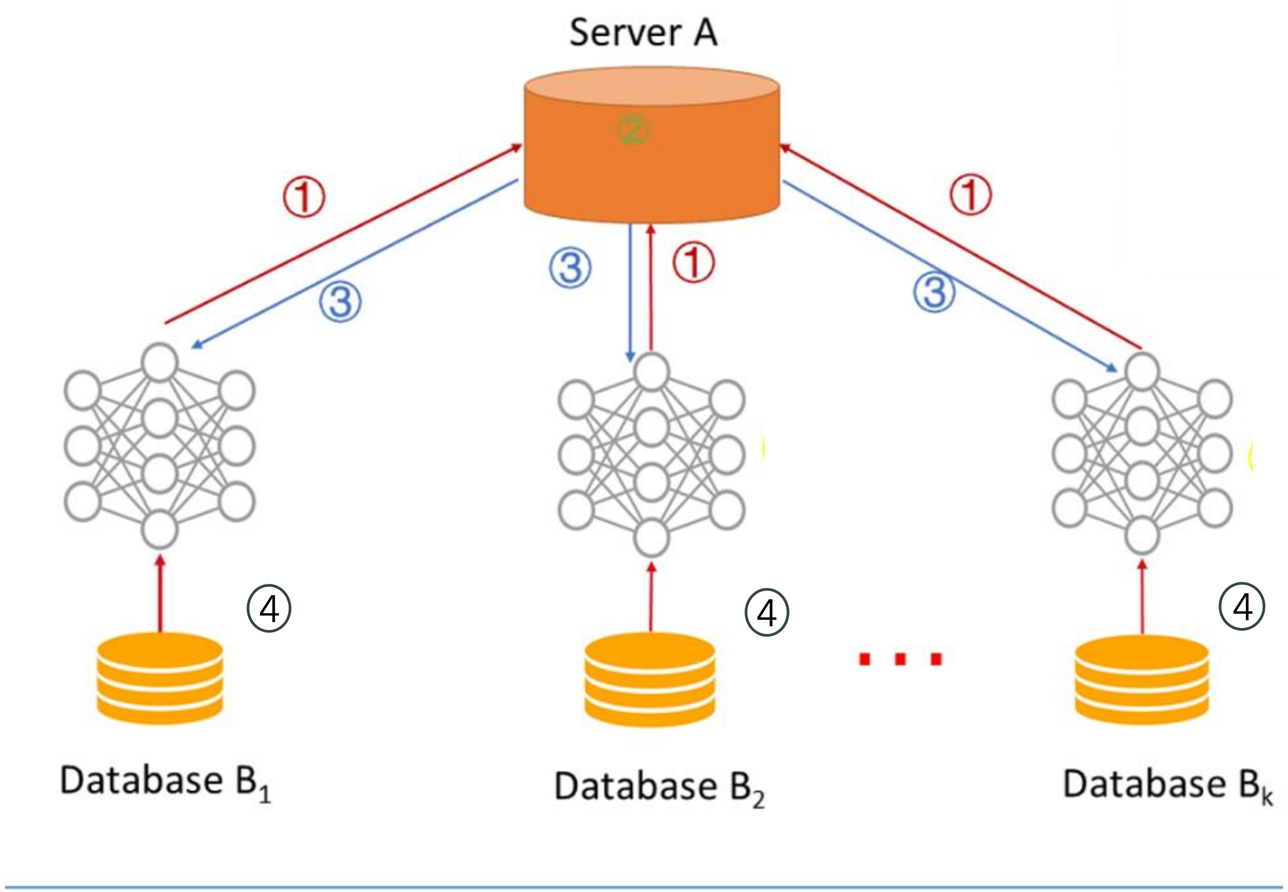
Perf.

Centralized model

Federated model

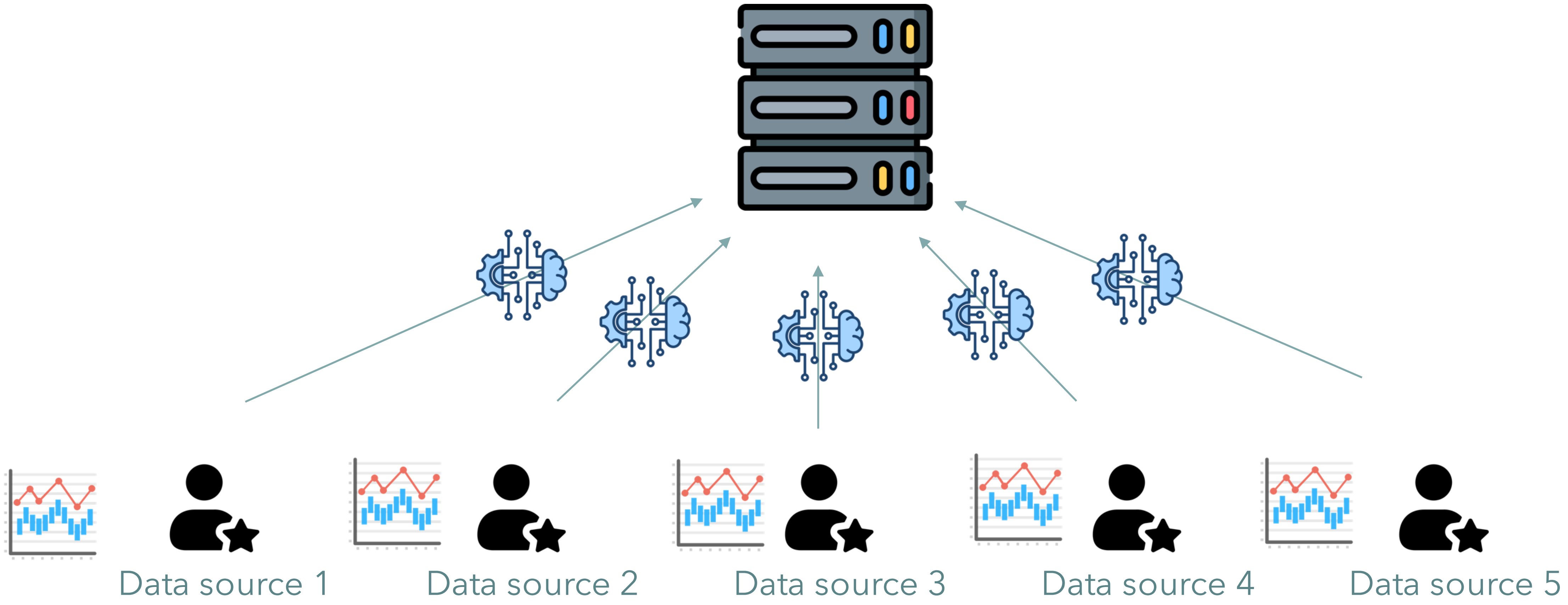
Local-only data model

Privacy

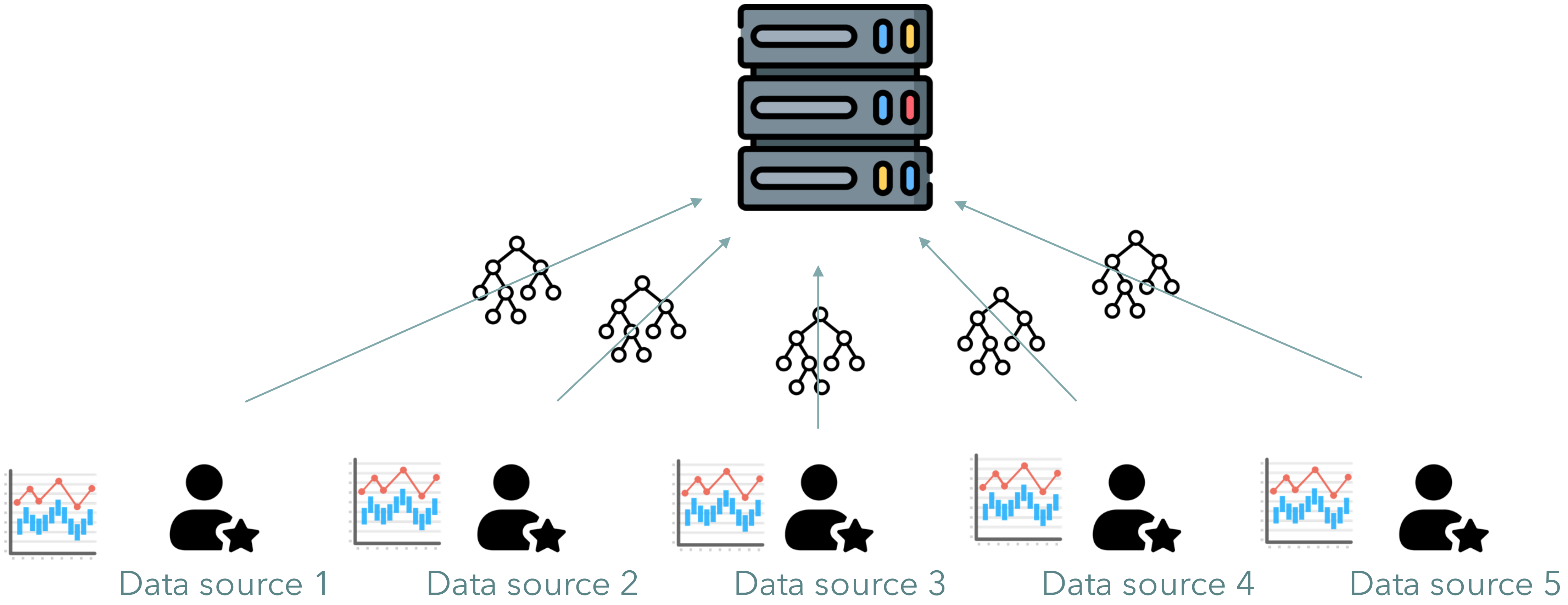


[1] Yang, Qiang, et al. "Federated machine learning: Concept and applications." *ACM Transactions on Intelligent Systems and Technology (TIST)* 10.2 (2019): 1-19.

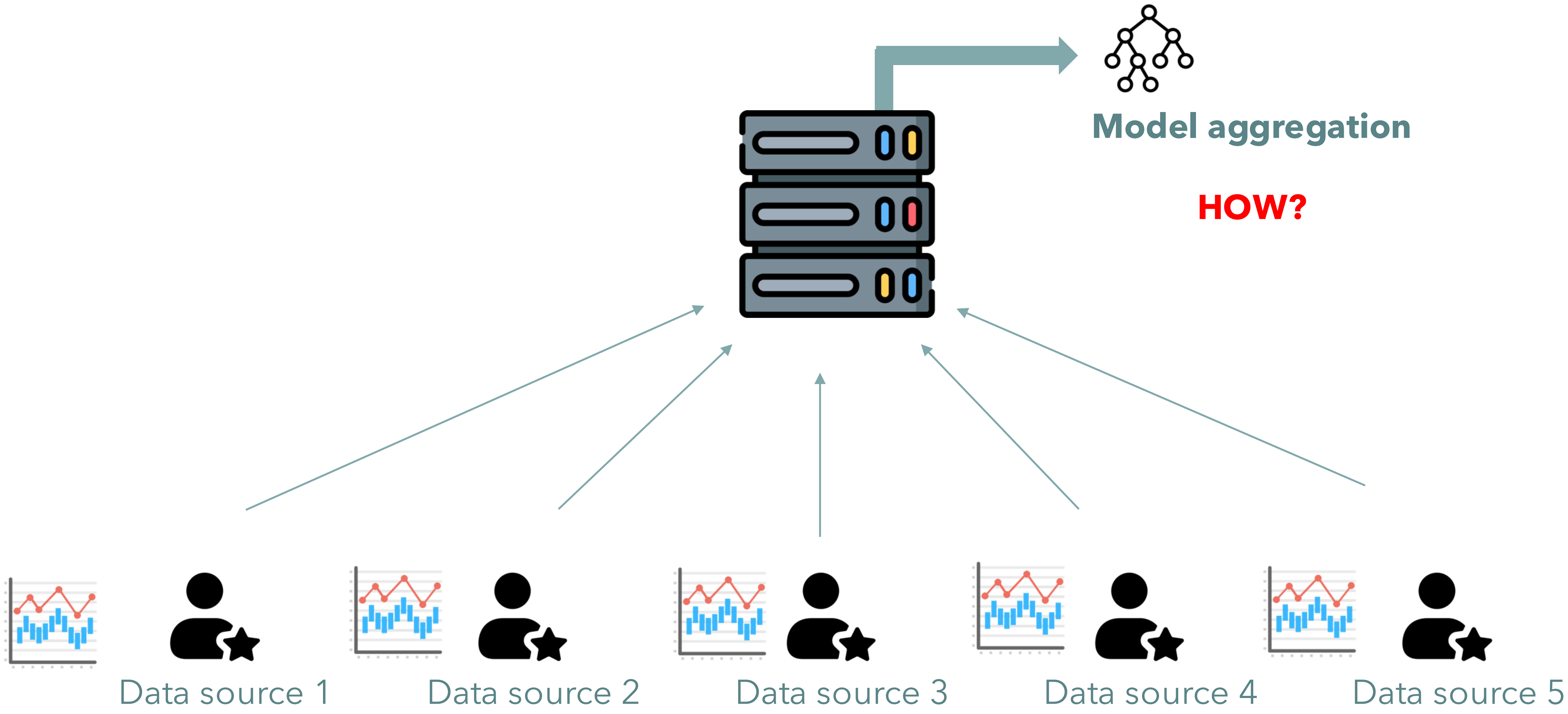
Horizontal Federated Learning – DT Ensembles



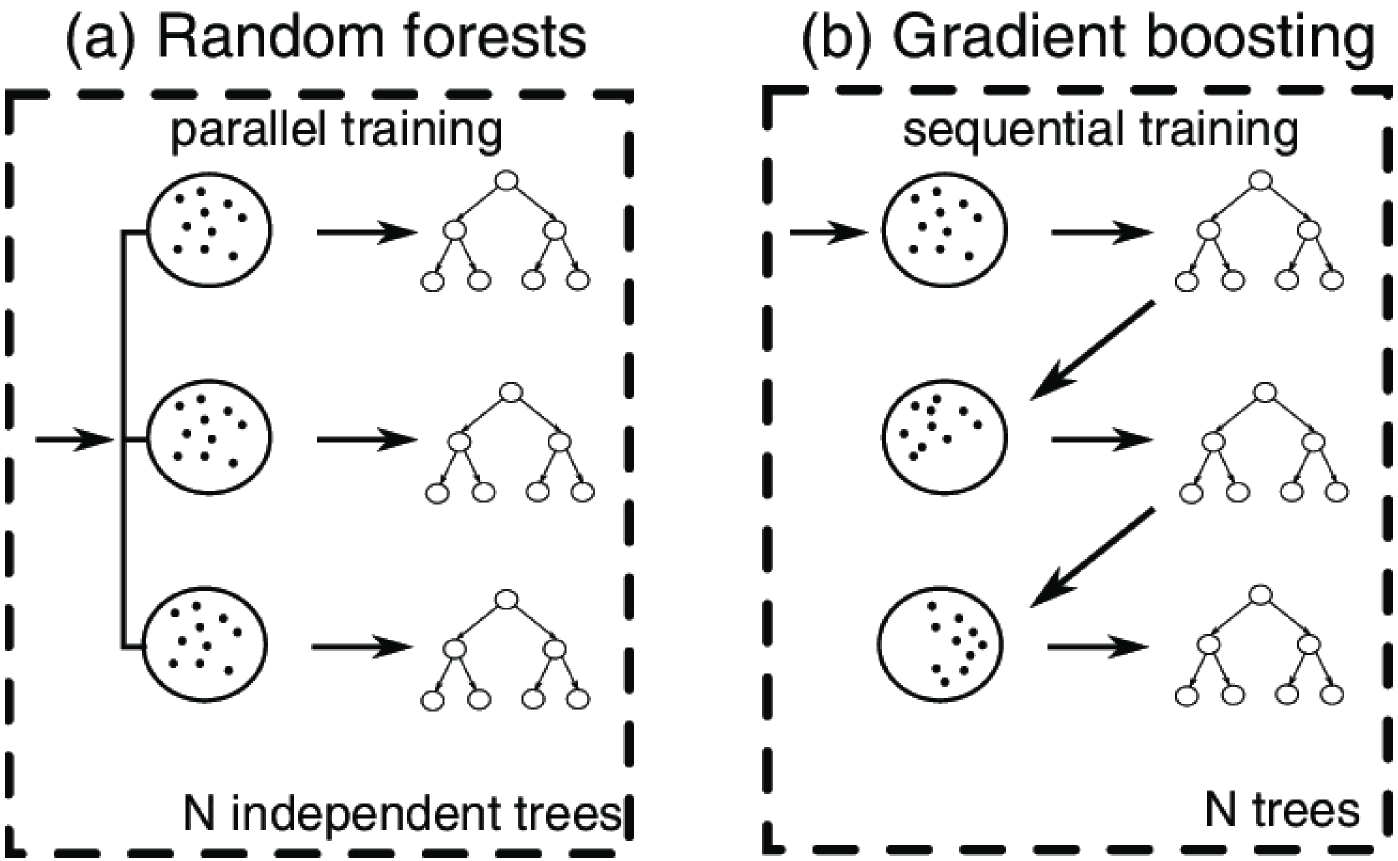
Horizontal Federated Learning – DT Ensembles



Horizontal Federated Learning – DT Ensembles



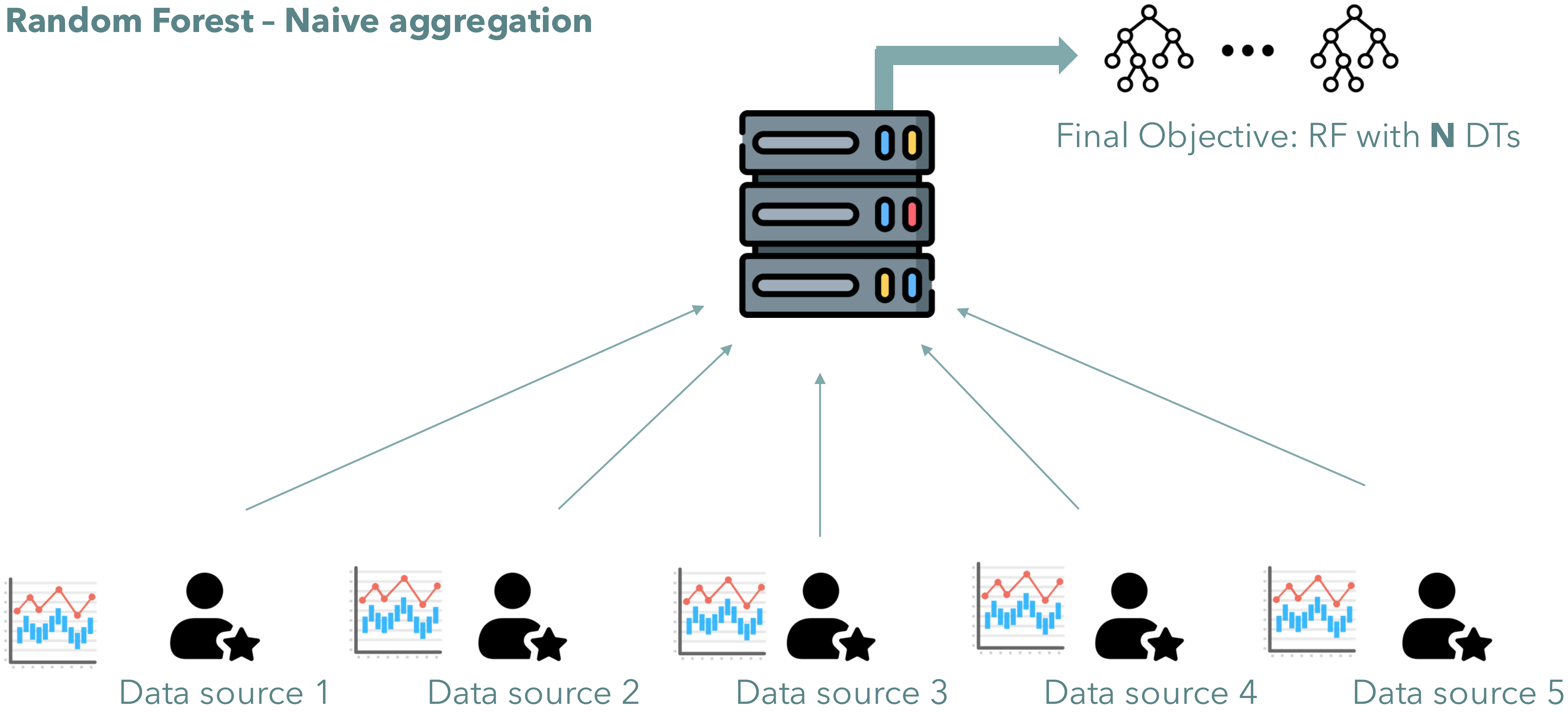
Horizontal Federated Learning – DT Ensembles



[2] Kowalek, P., Loch-Olszewska, H., & Szwabiński, J. (2019). Classification of diffusion modes in single-particle tracking data: Feature-based versus deep-learning approach. *Physical Review E*, *100*(3), 032410.

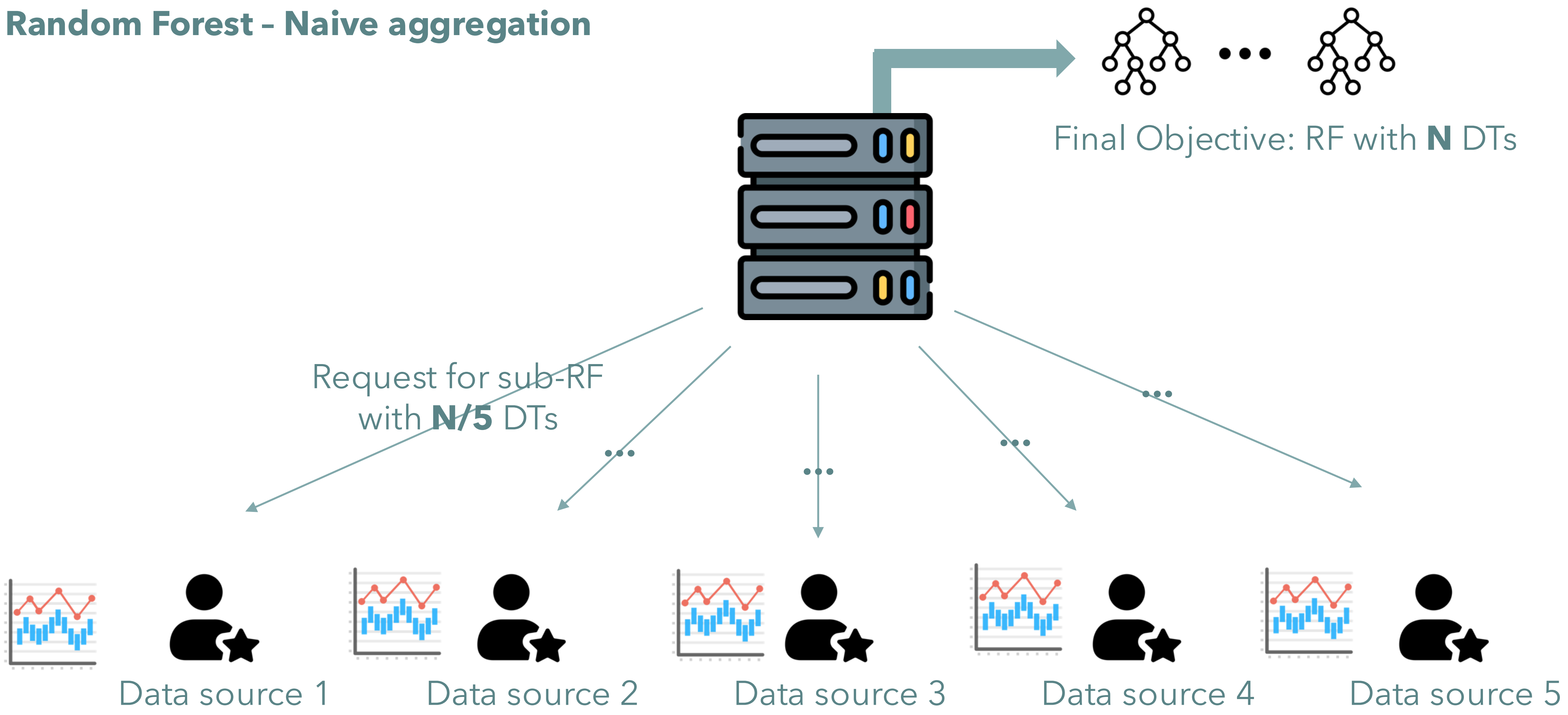
Horizontal Federated Learning – DT Ensembles

Random Forest - Naive aggregation



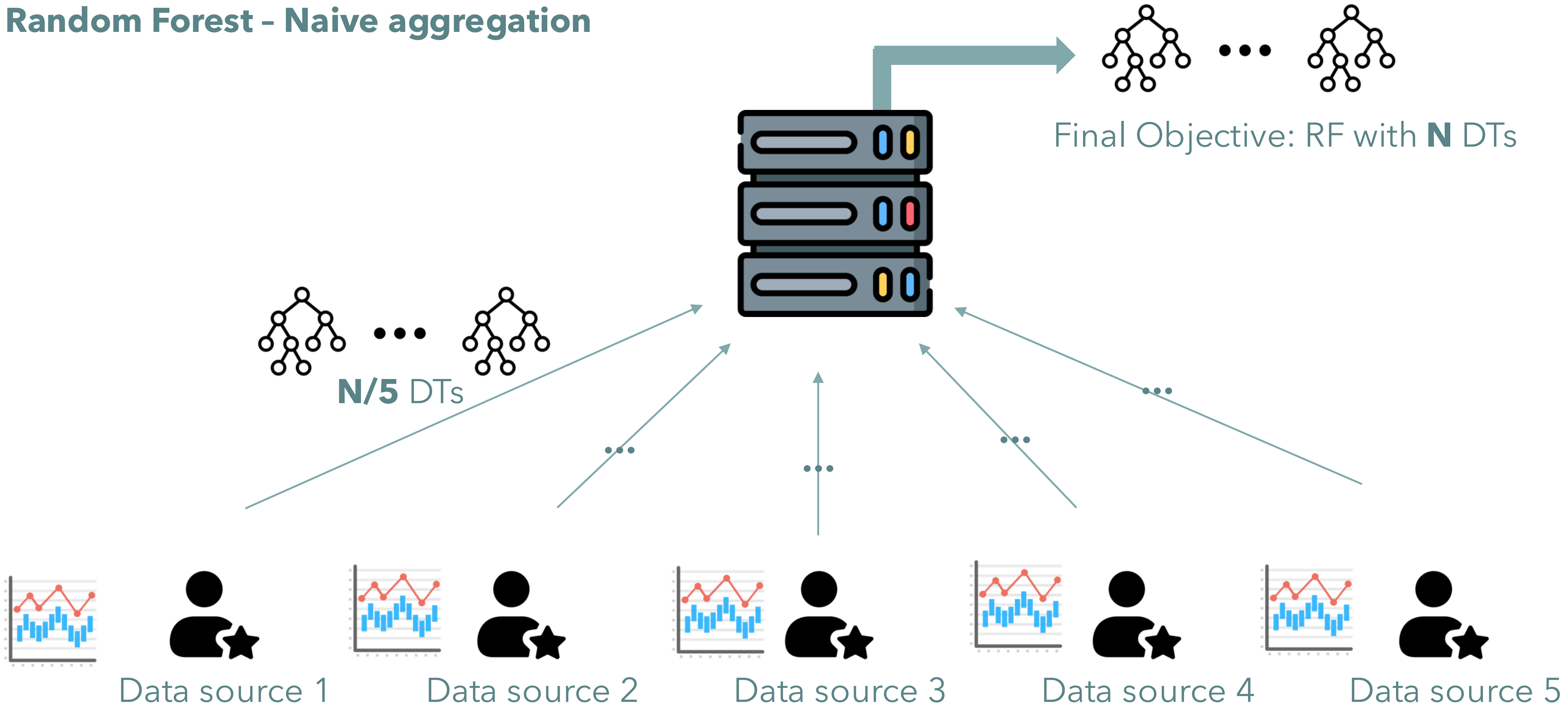
Horizontal Federated Learning – DT Ensembles

Random Forest – Naive aggregation



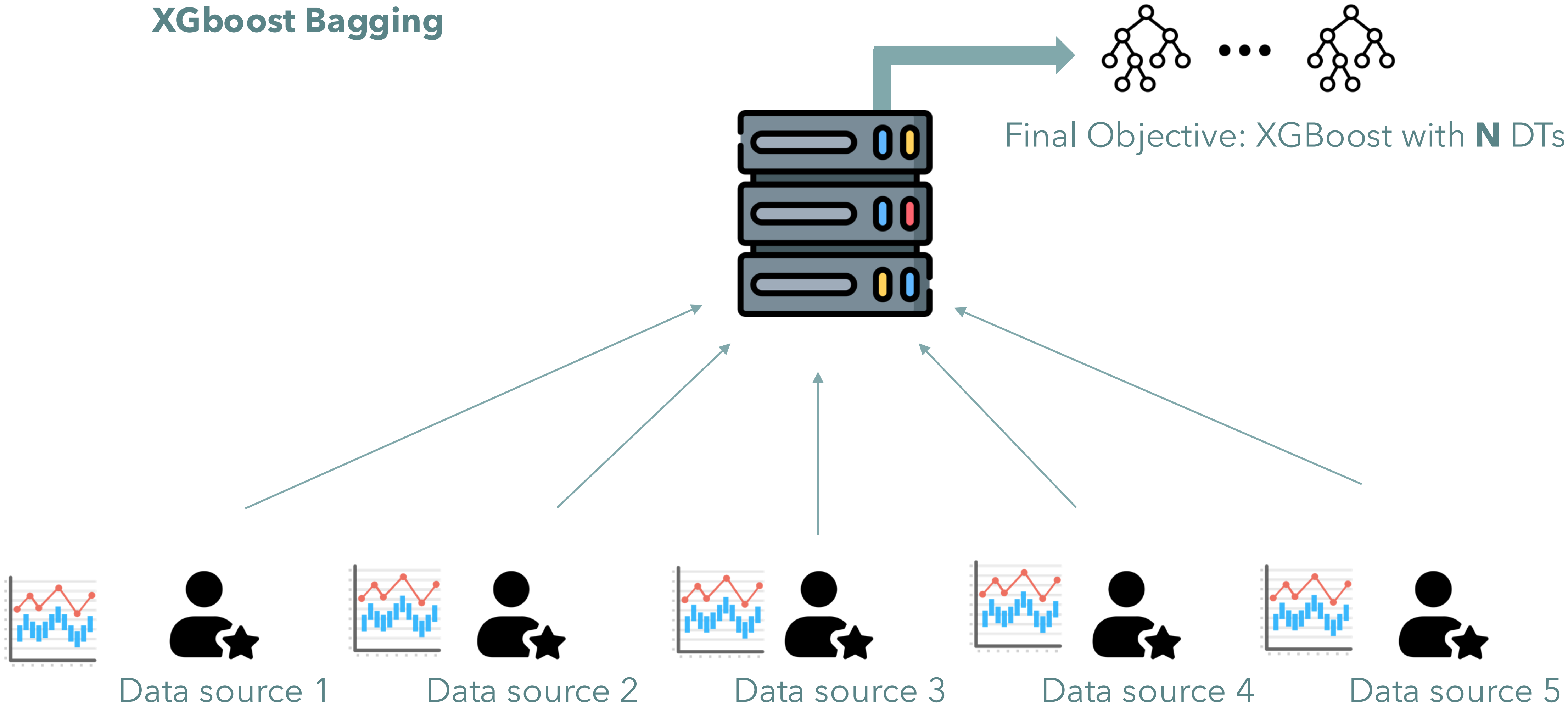
Horizontal Federated Learning – DT Ensembles

Random Forest – Naive aggregation



Horizontal Federated Learning – DT Ensembles

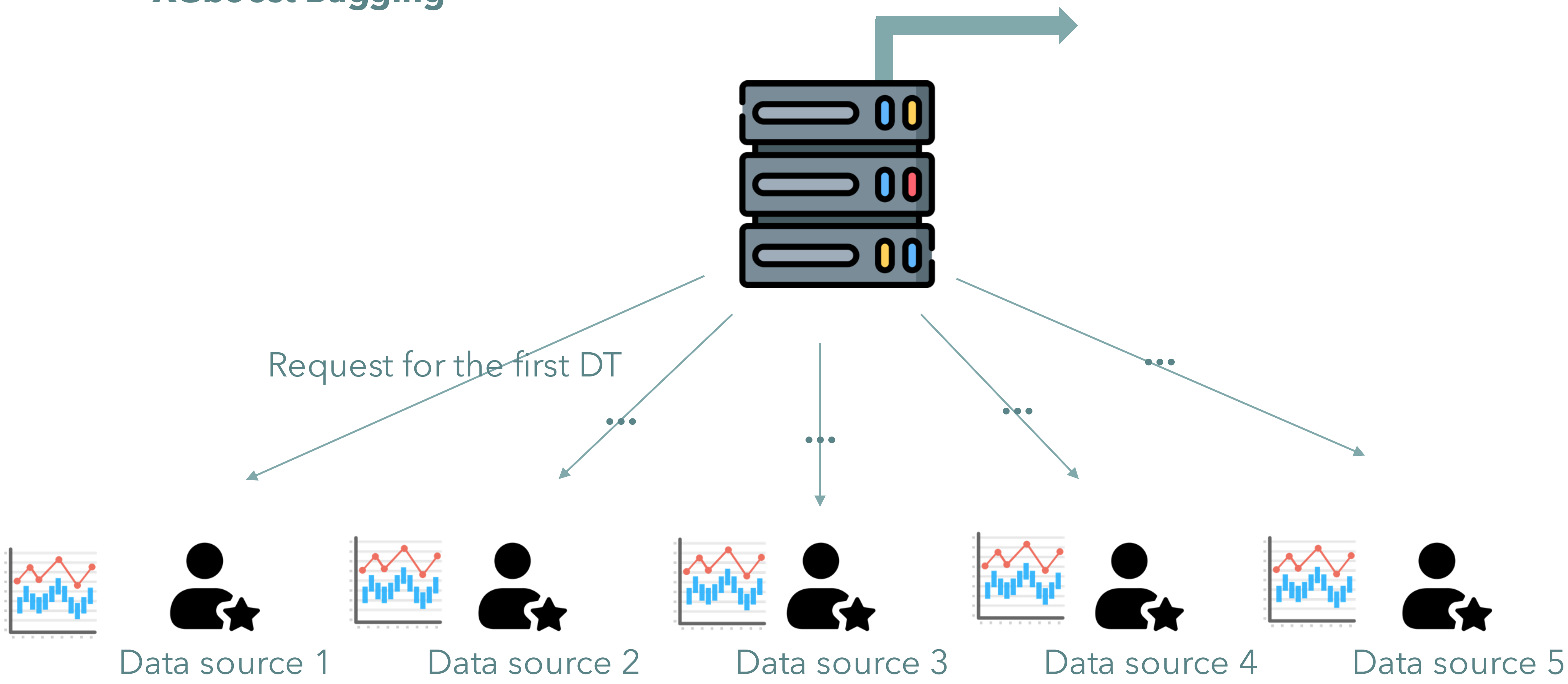
XGboost Bagging



[3] Di Gennaro, Marco, et al. "TimberStrike: Dataset Reconstruction Attack Revealing Privacy Leakage in Federated Tree-Based Systems." *arXiv preprint arXiv:2506.07605* (2025).

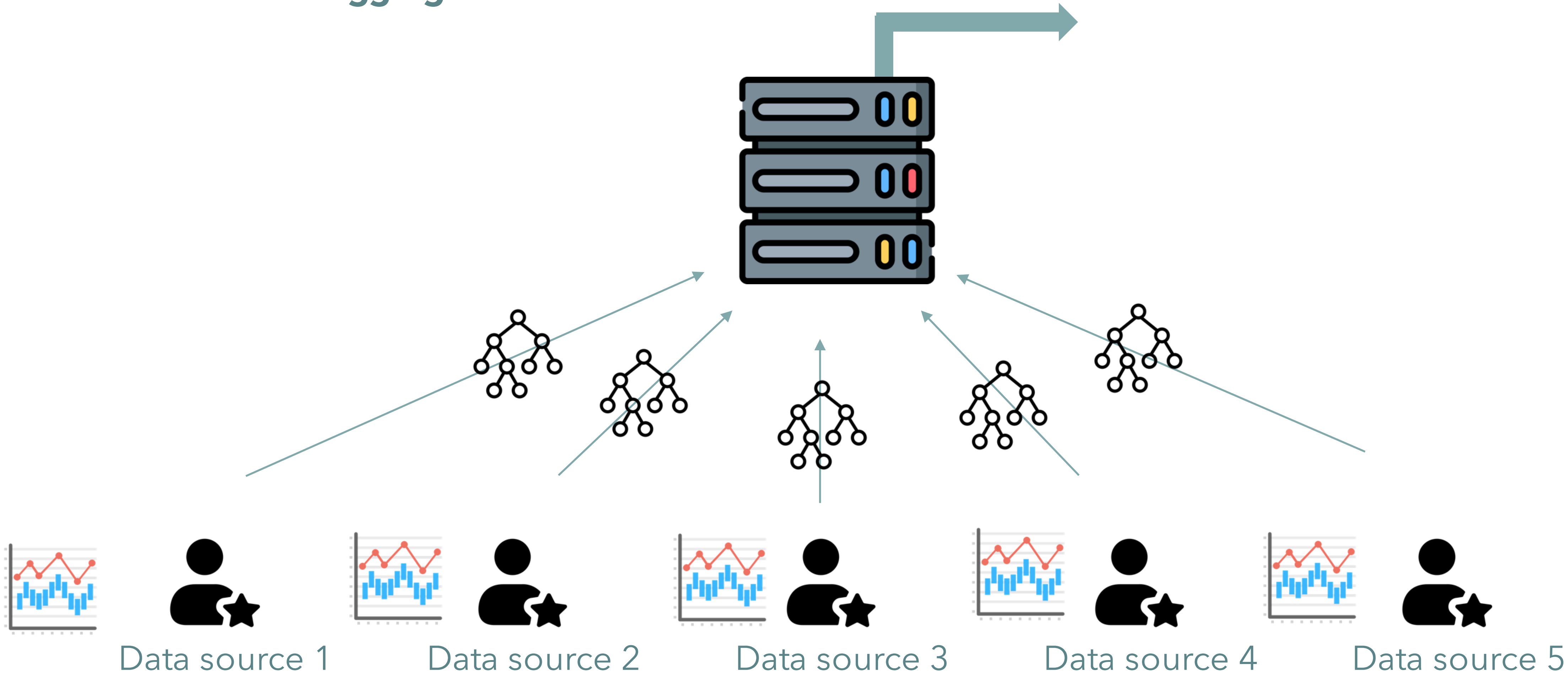
Horizontal Federated Learning – DT Ensembles

XGboost Bagging

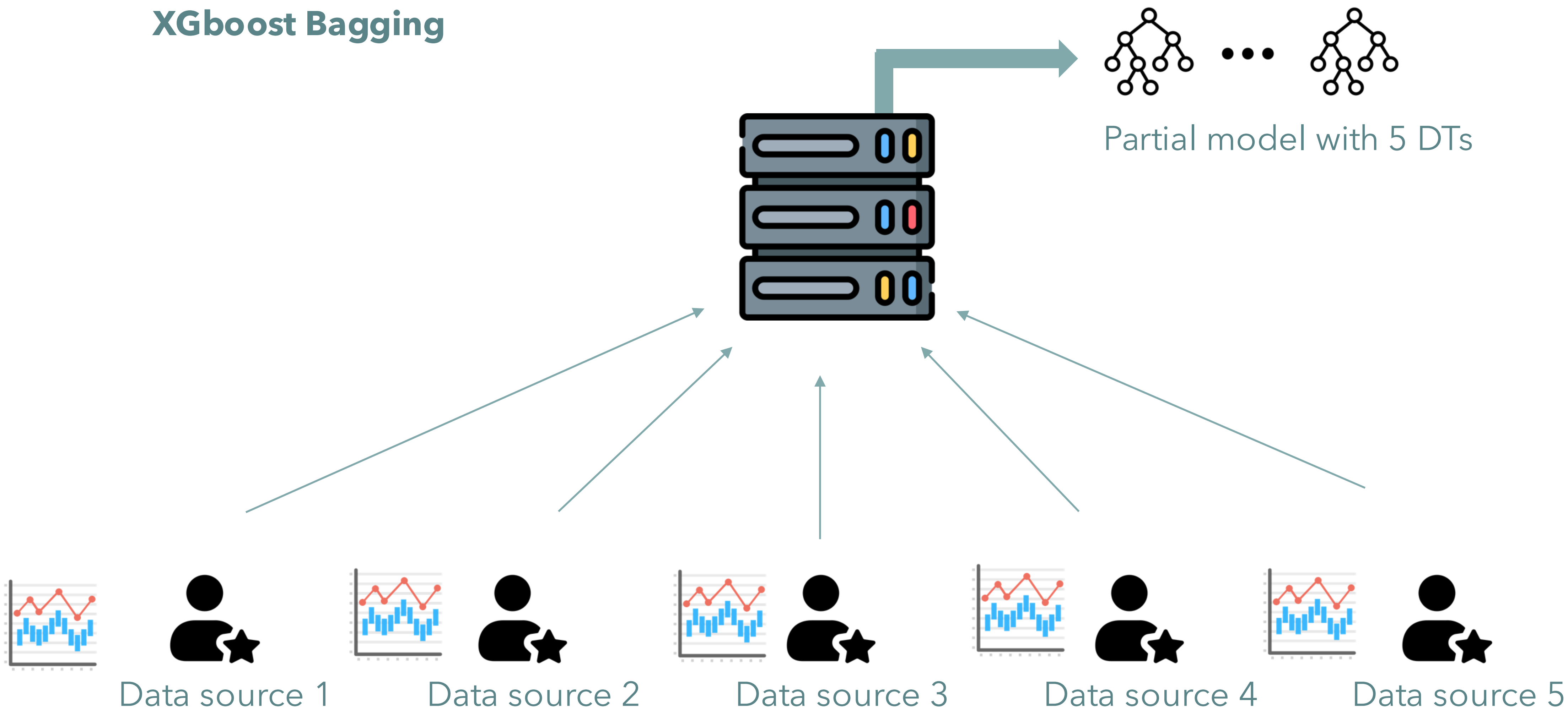


Horizontal Federated Learning – DT Ensembles

XGboost Bagging

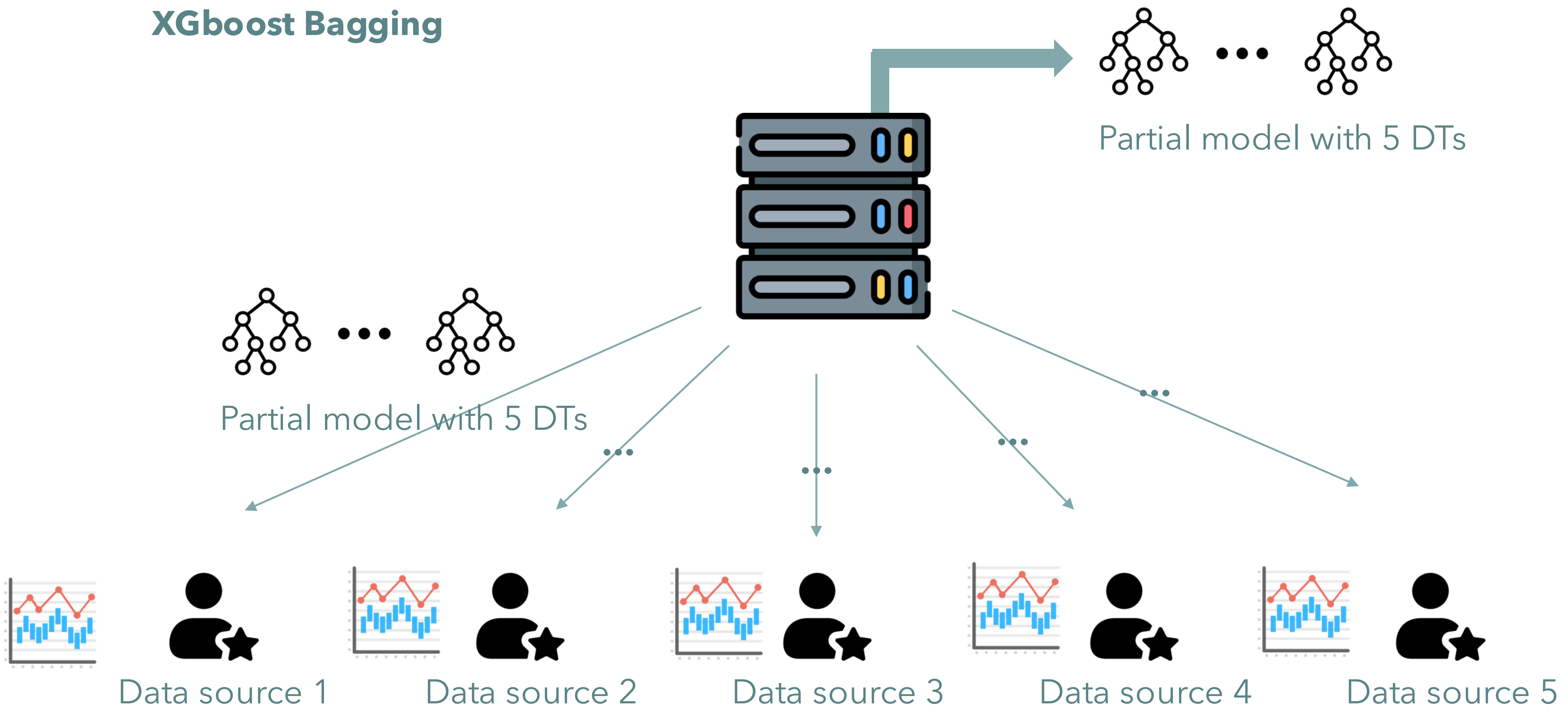


Horizontal Federated Learning – DT Ensembles



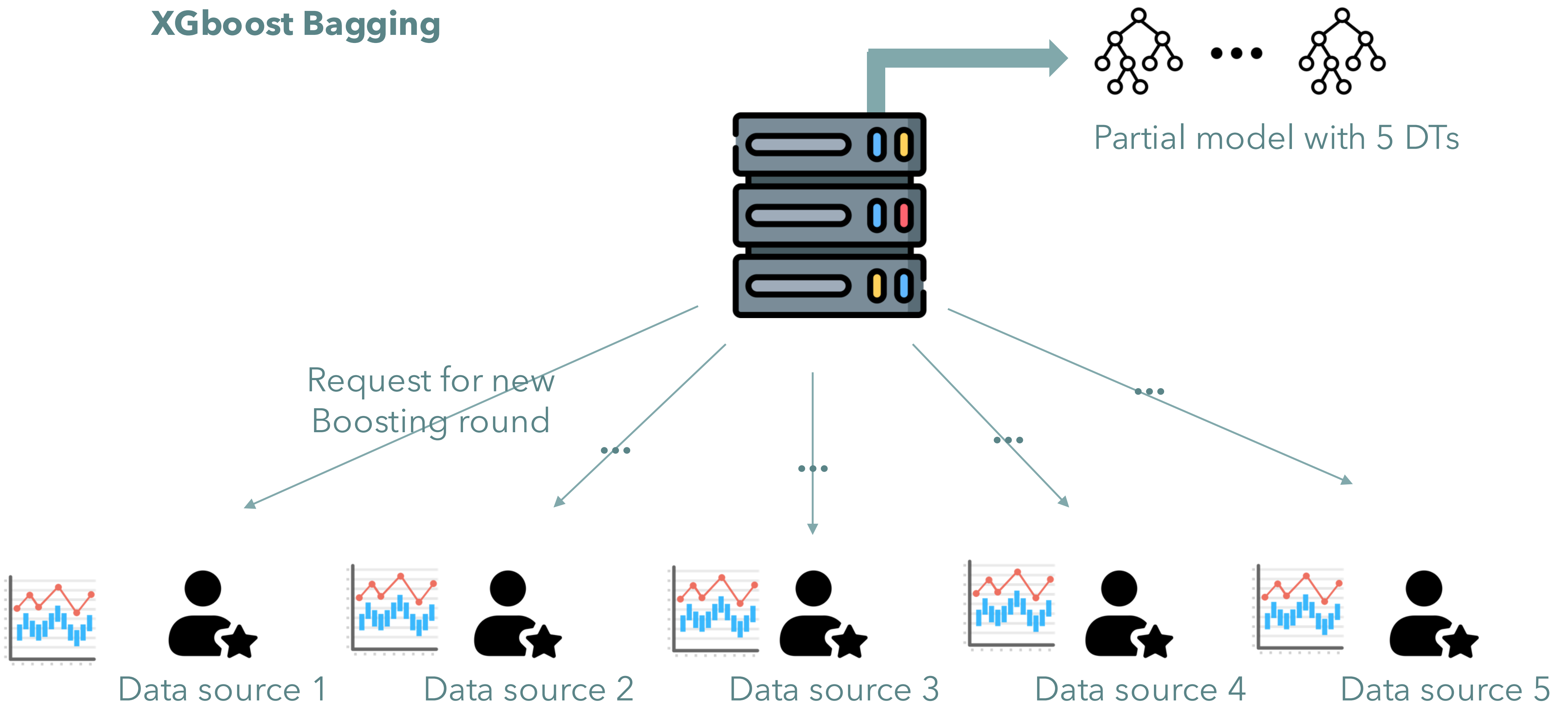
Horizontal Federated Learning – DT Ensembles

XGboost Bagging

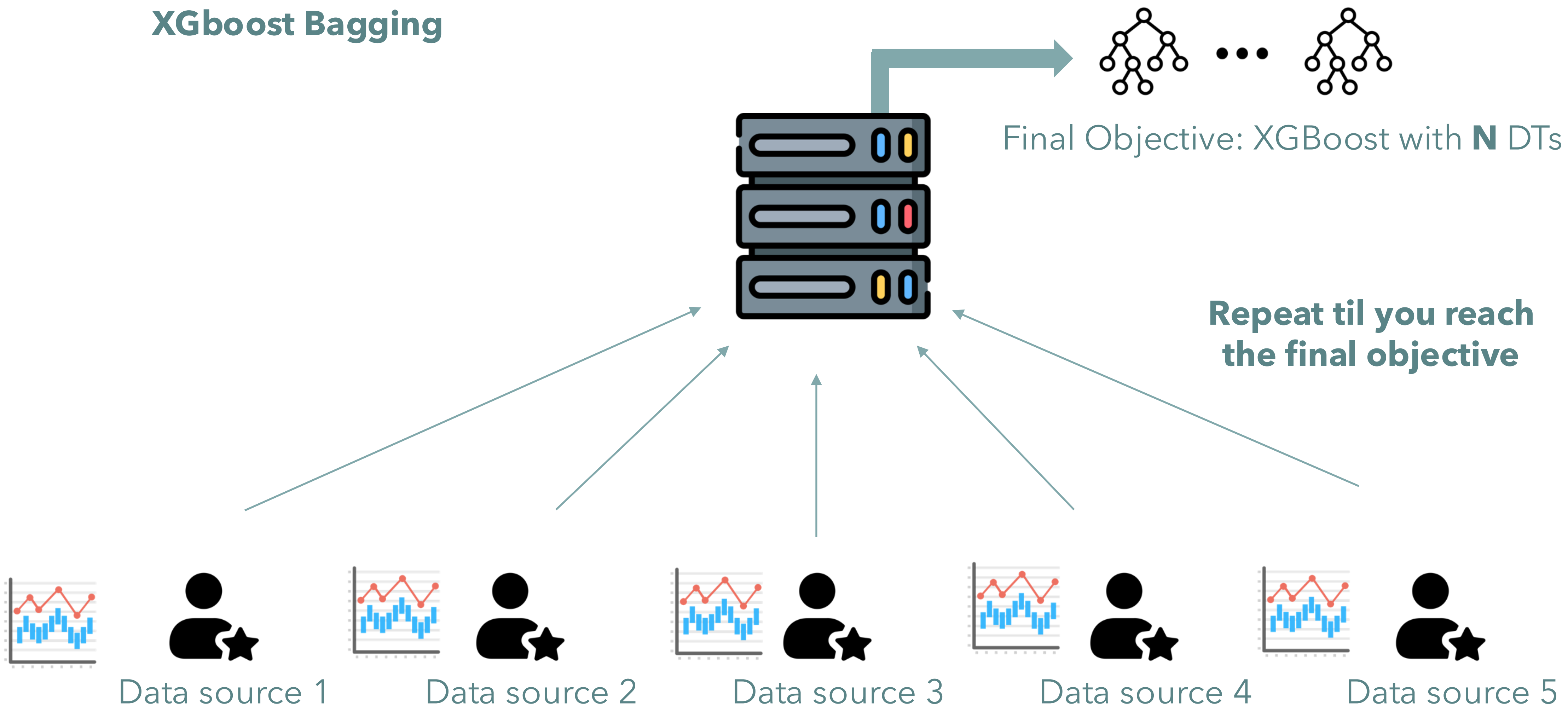


Horizontal Federated Learning – DT Ensembles

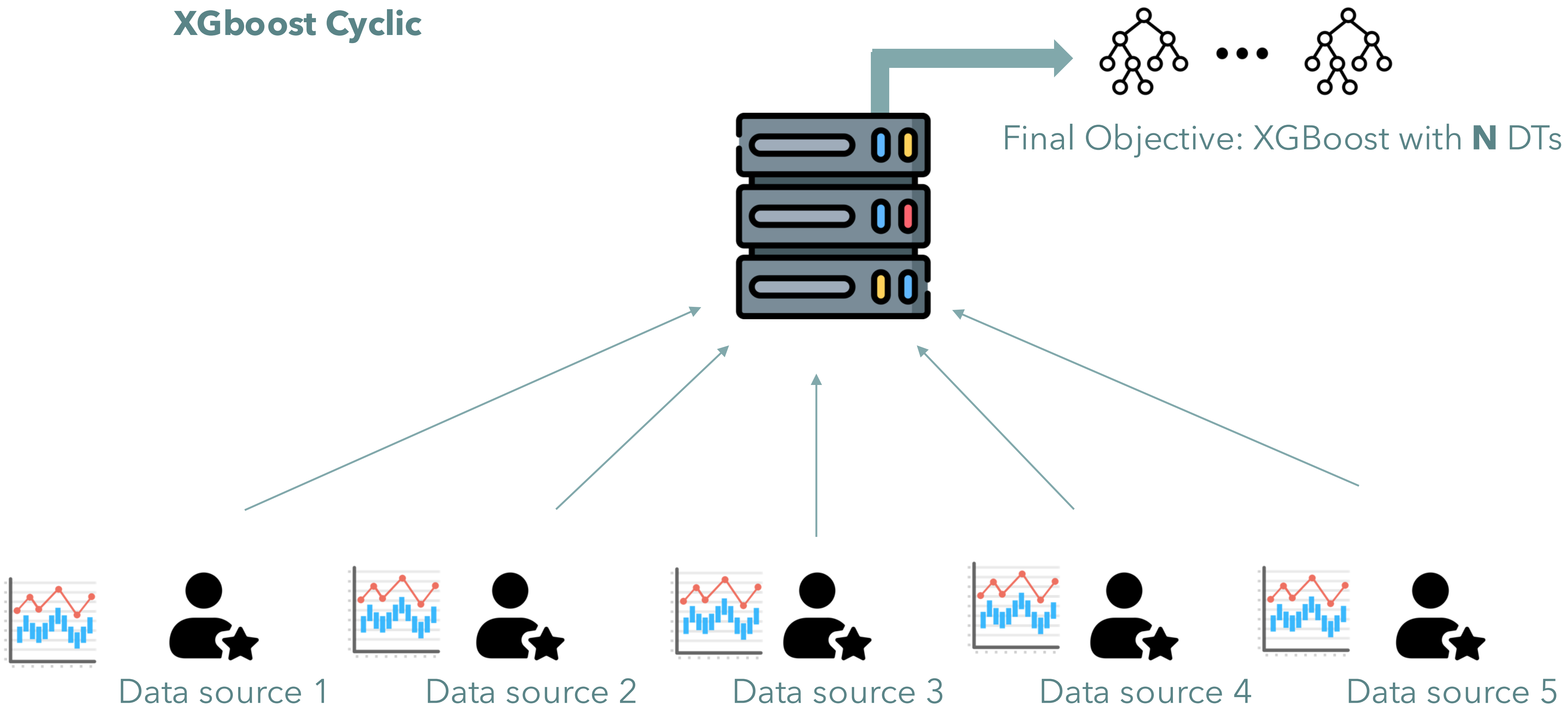
XGboost Bagging



Horizontal Federated Learning – DT Ensembles



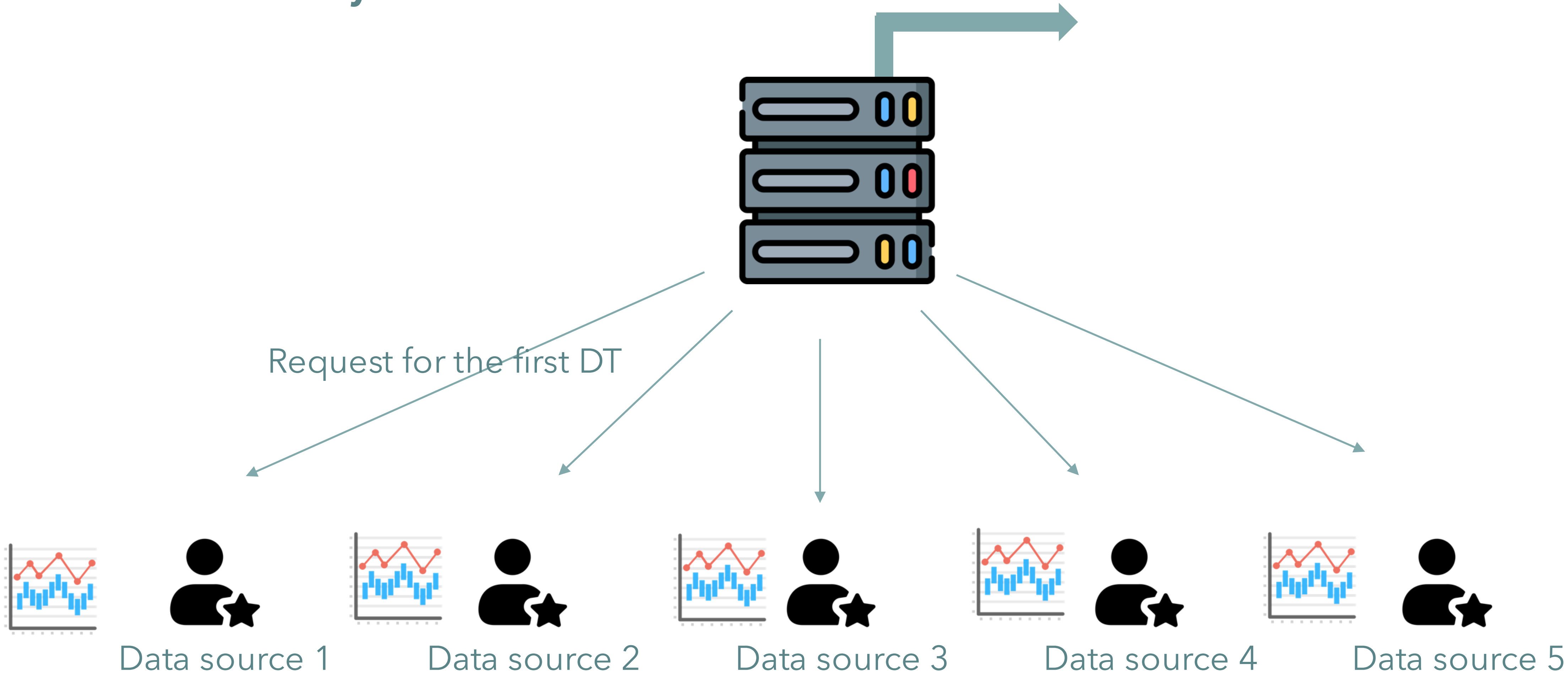
Horizontal Federated Learning – DT Ensembles



[3] Di Gennaro, Marco, et al. "TimberStrike: Dataset Reconstruction Attack Revealing Privacy Leakage in Federated Tree-Based Systems." *arXiv preprint arXiv:2506.07605* (2025).

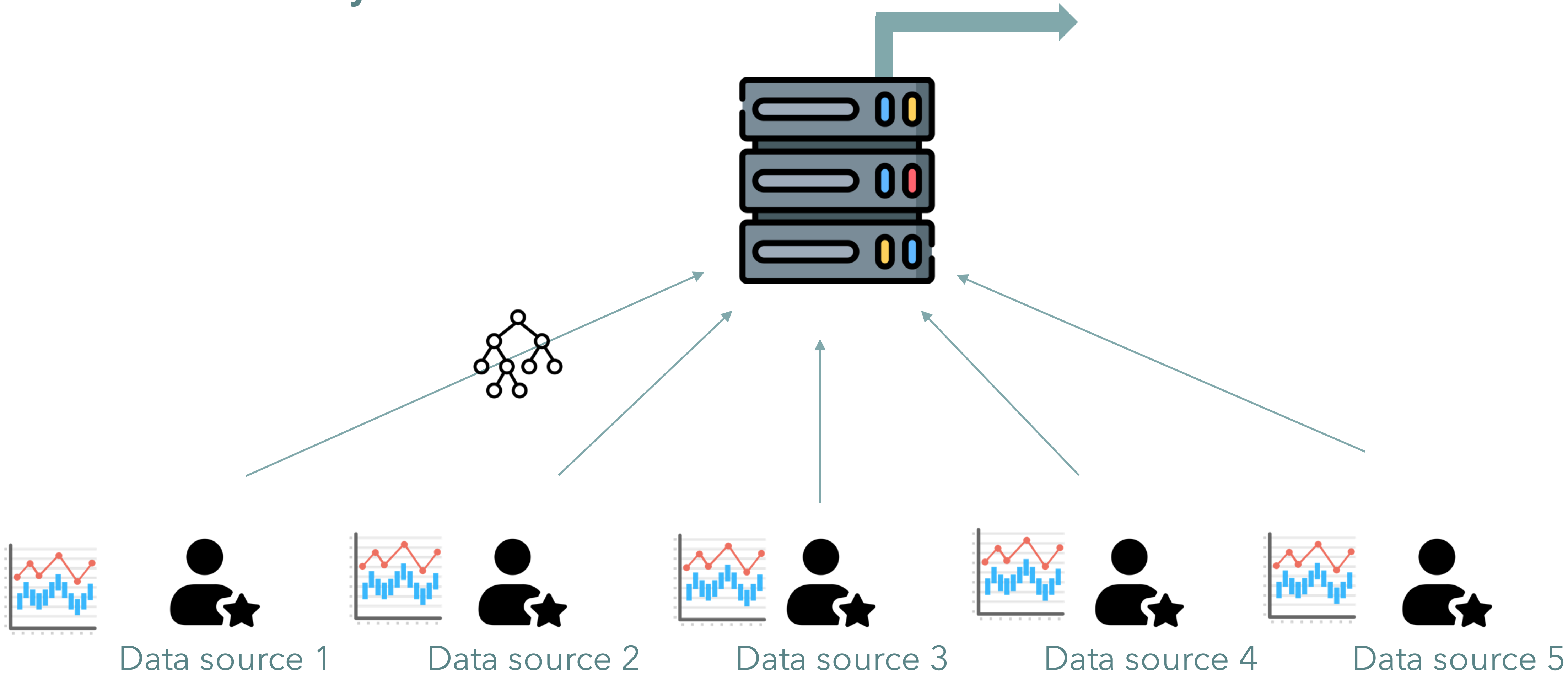
Horizontal Federated Learning – DT Ensembles

XGboost Cyclic



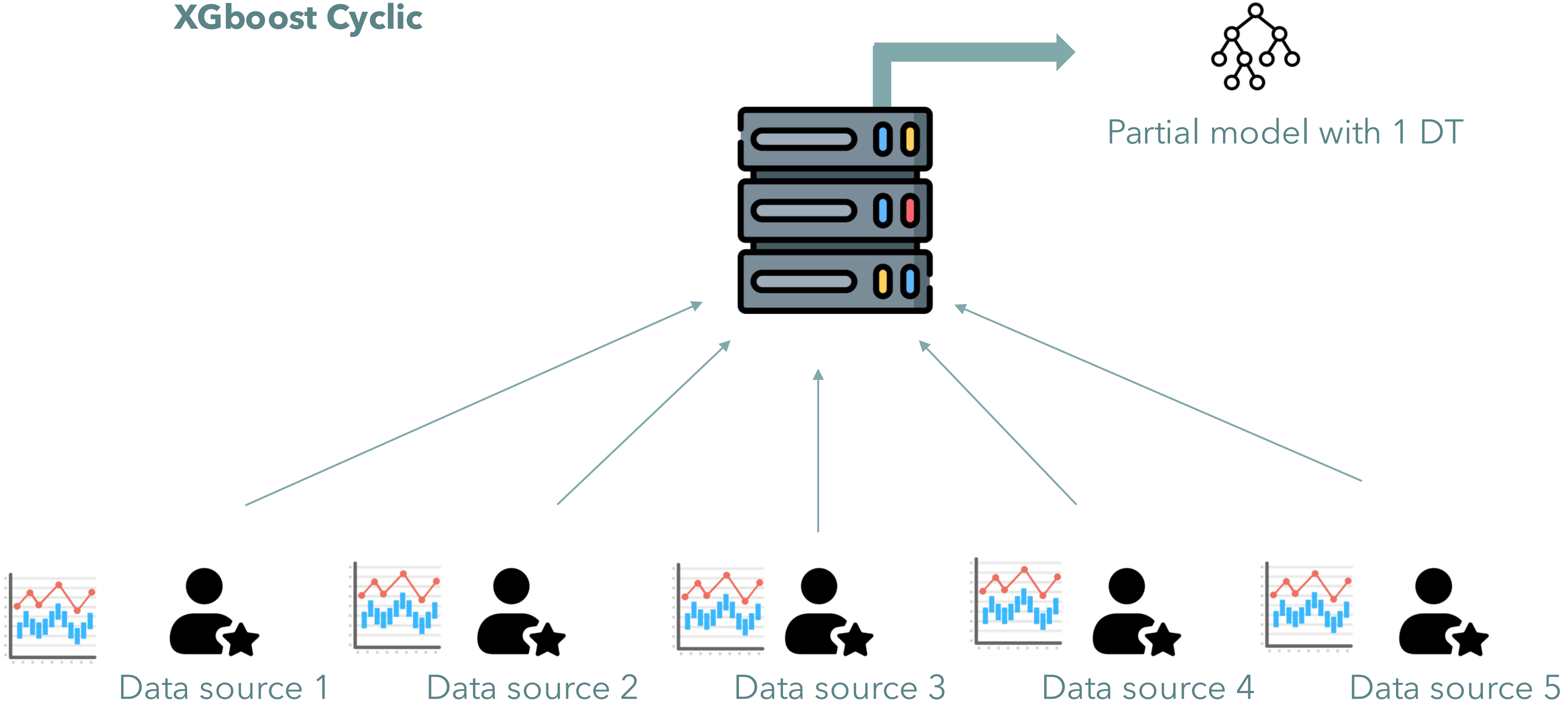
Horizontal Federated Learning – DT Ensembles

XGboost Cyclic



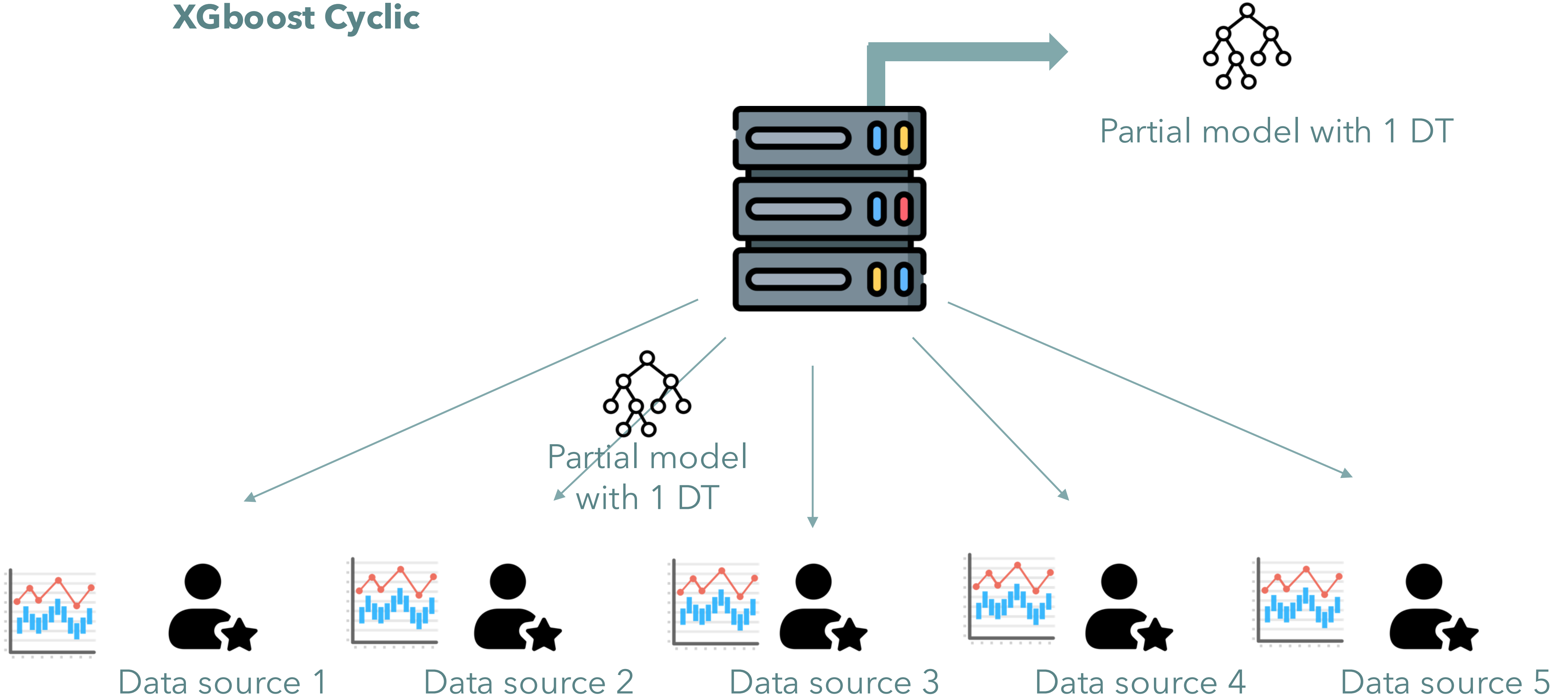
Horizontal Federated Learning – DT Ensembles

XGboost Cyclic



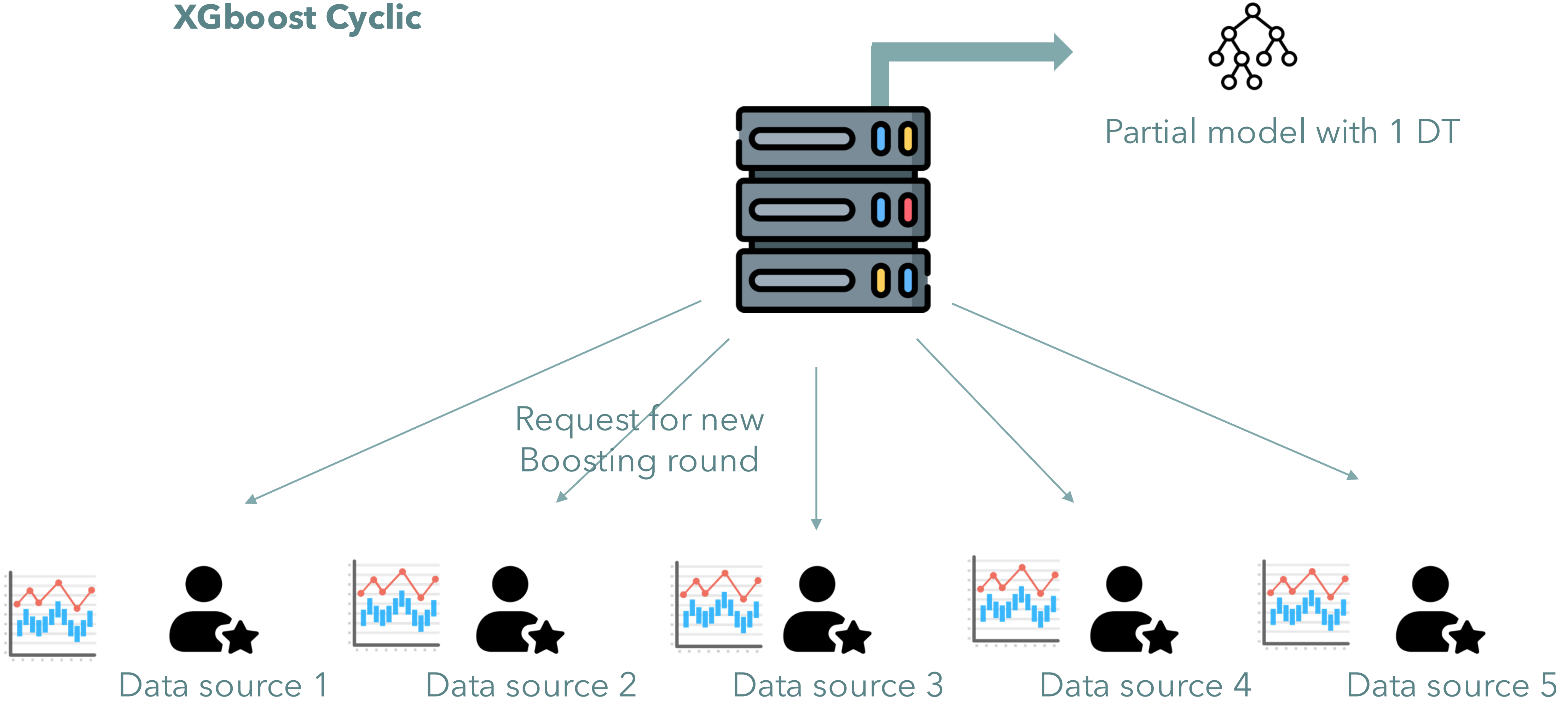
Horizontal Federated Learning – DT Ensembles

XGboost Cyclic



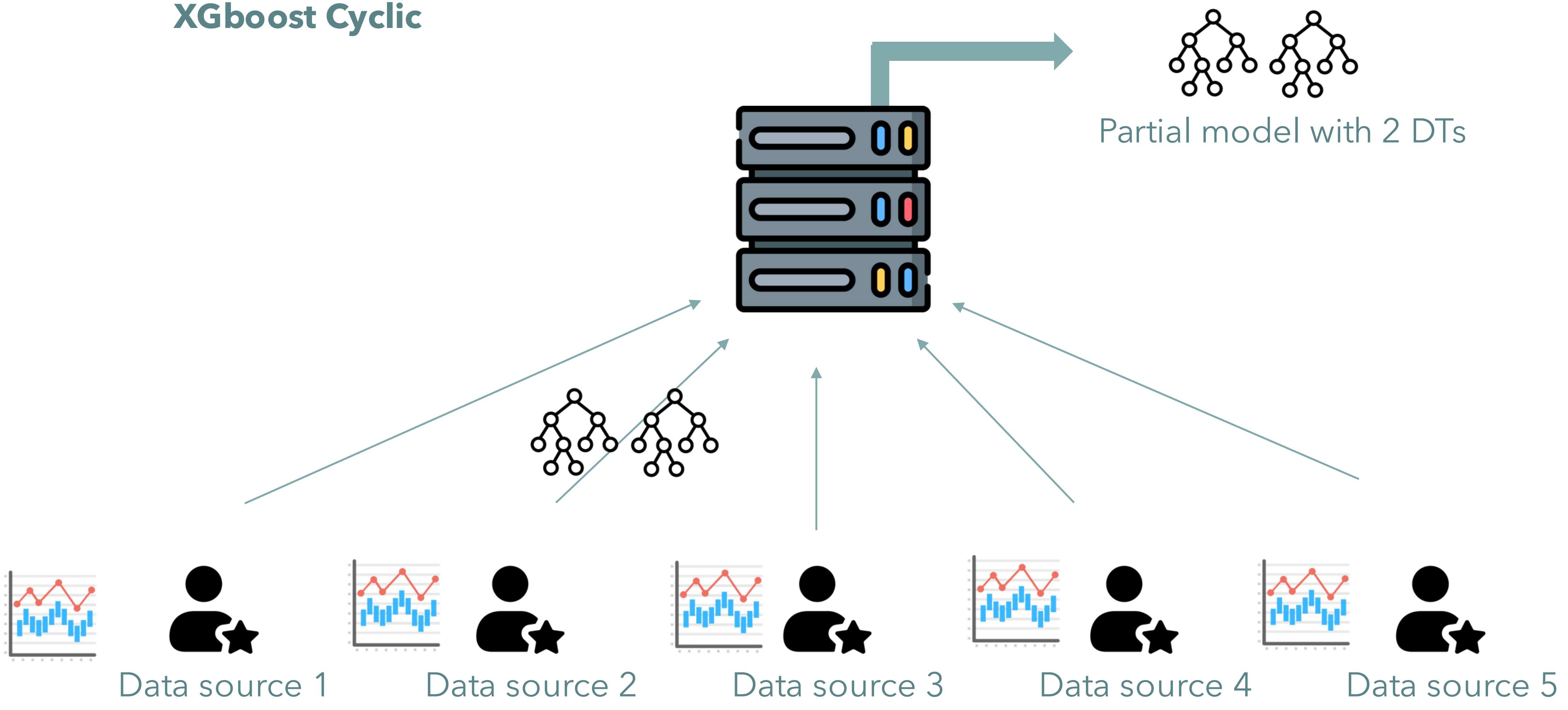
Horizontal Federated Learning – DT Ensembles

XGboost Cyclic



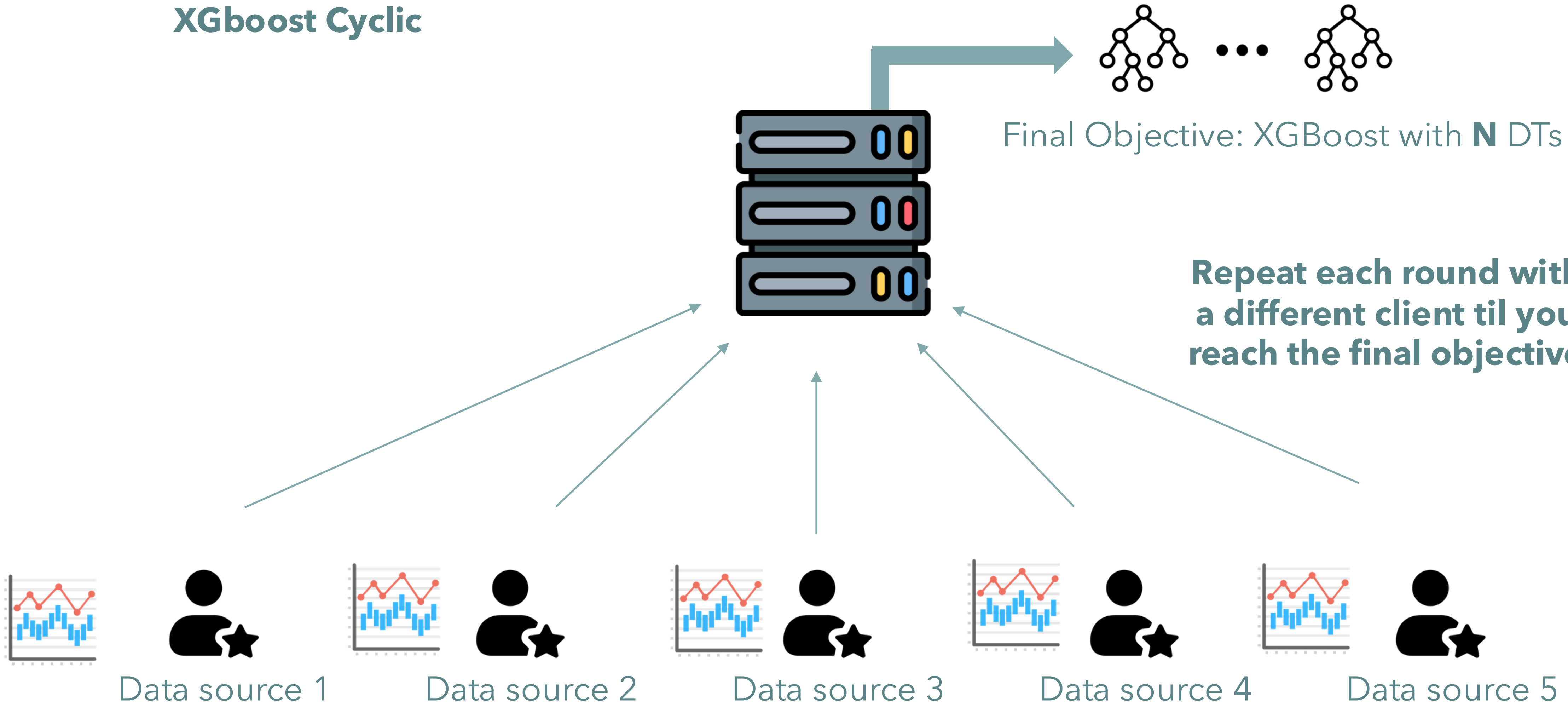
Horizontal Federated Learning – DT Ensembles

XGboost Cyclic



Horizontal Federated Learning – DT Ensembles

XGboost Cyclic

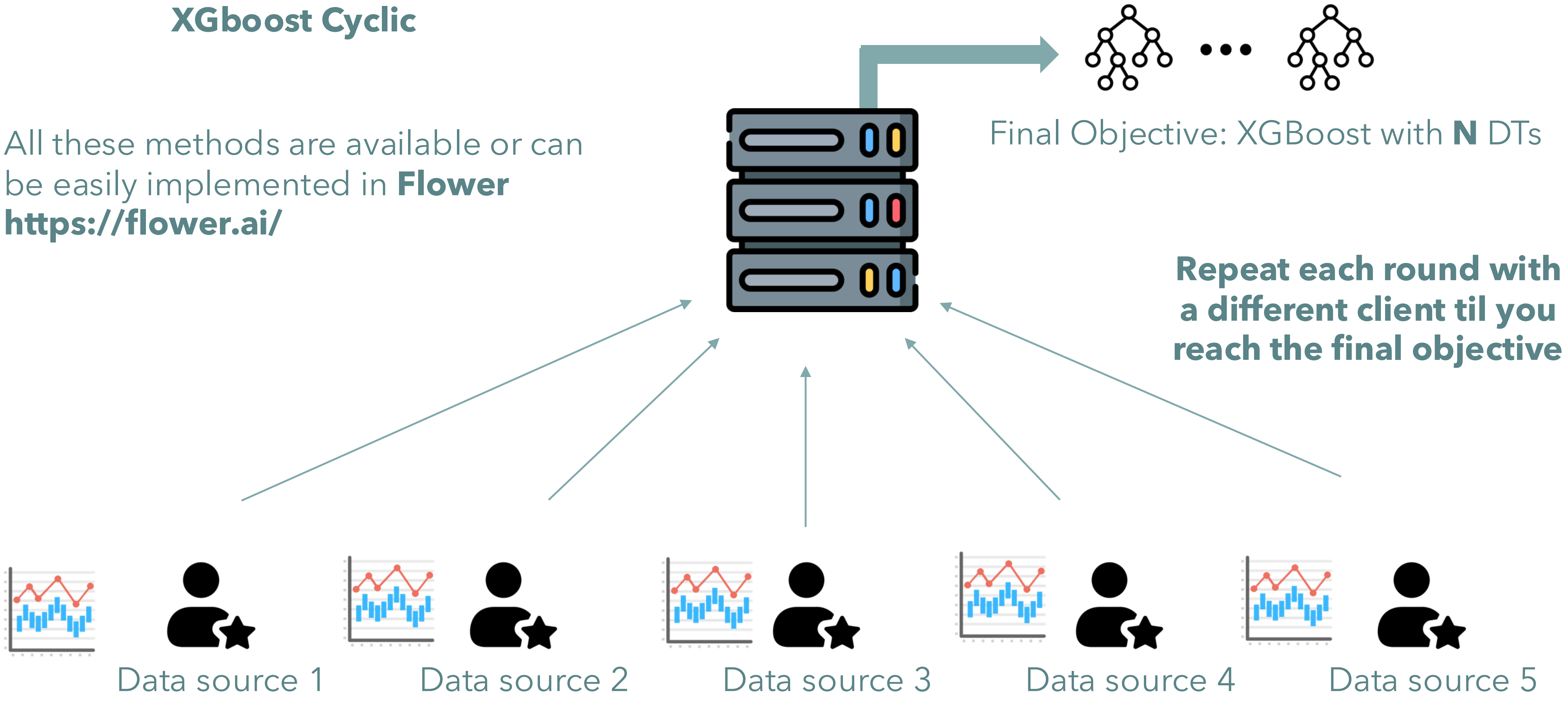


Repeat each round with a different client til you reach the final objective

Horizontal Federated Learning – DT Ensembles

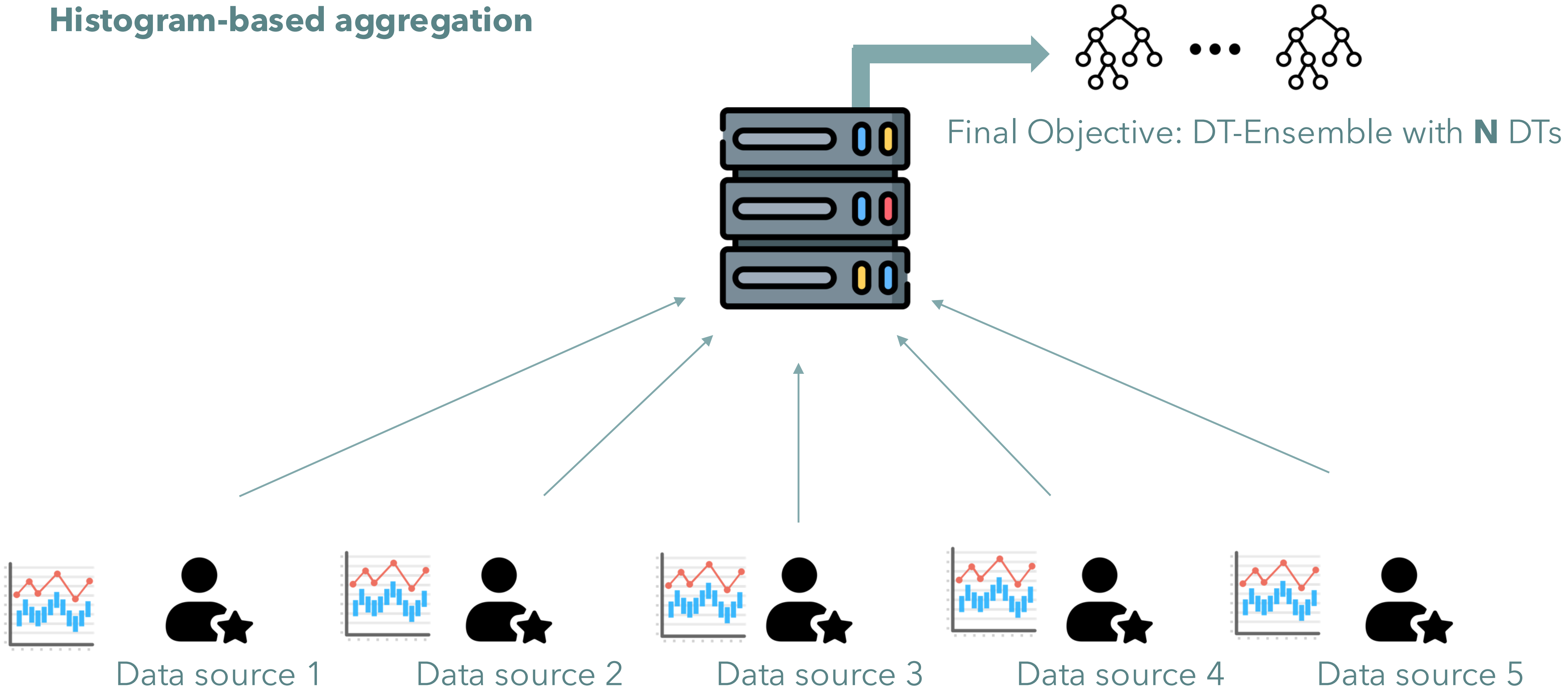
XGboost Cyclic

All these methods are available or can be easily implemented in **Flower** **https://flower.ai/**



Horizontal Federated Learning – DT Ensembles

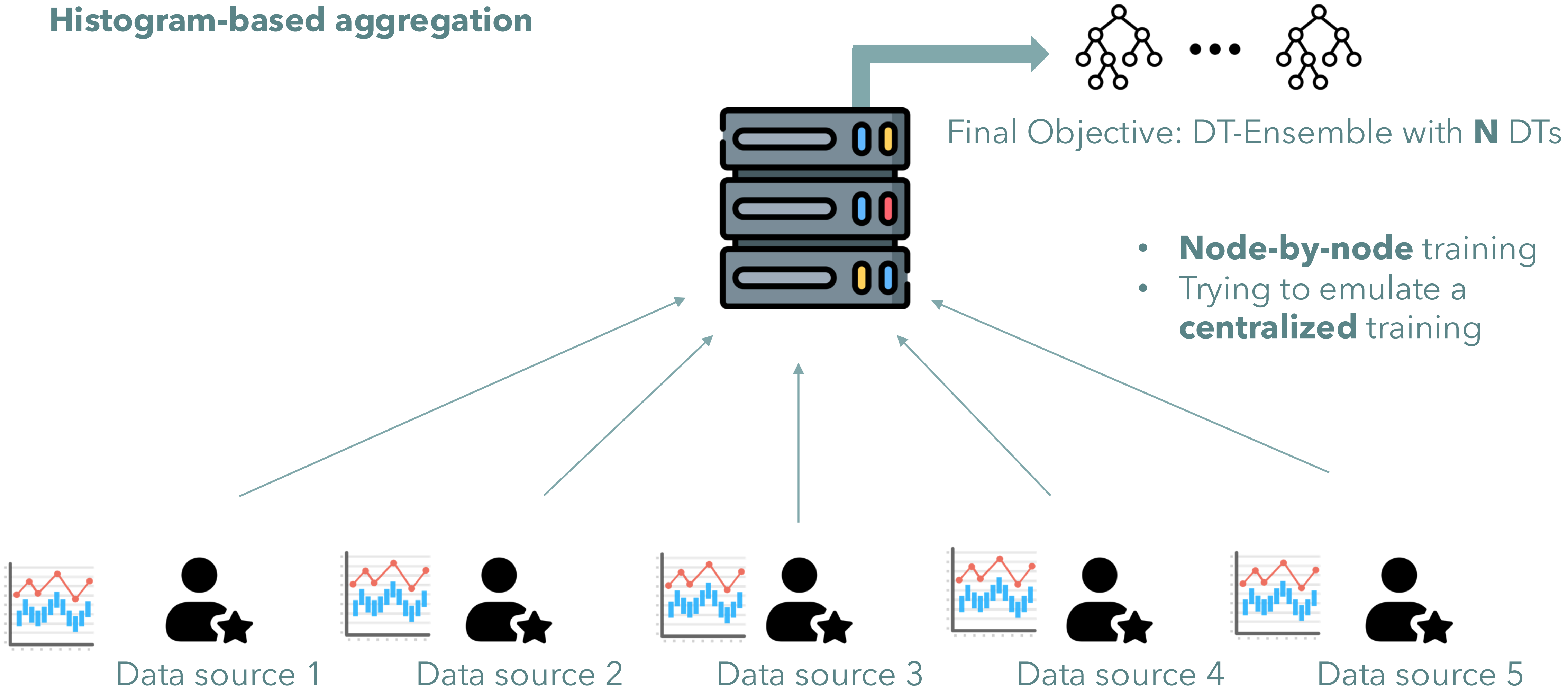
Histogram-based aggregation



[3] Di Gennaro, Marco, et al. "TimberStrike: Dataset Reconstruction Attack Revealing Privacy Leakage in Federated Tree-Based Systems." *arXiv preprint arXiv:2506.07605* (2025).

Horizontal Federated Learning – DT Ensembles

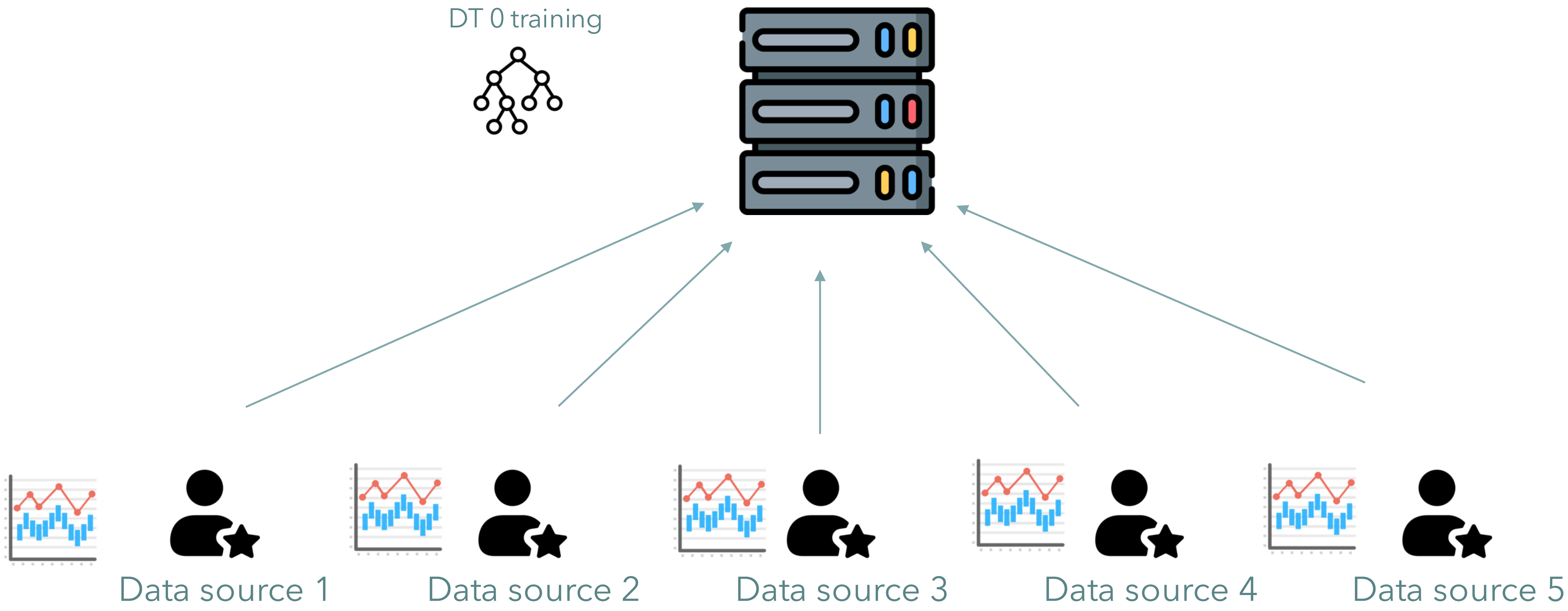
Histogram-based aggregation



- **Node-by-node** training
- Trying to emulate a **centralized** training

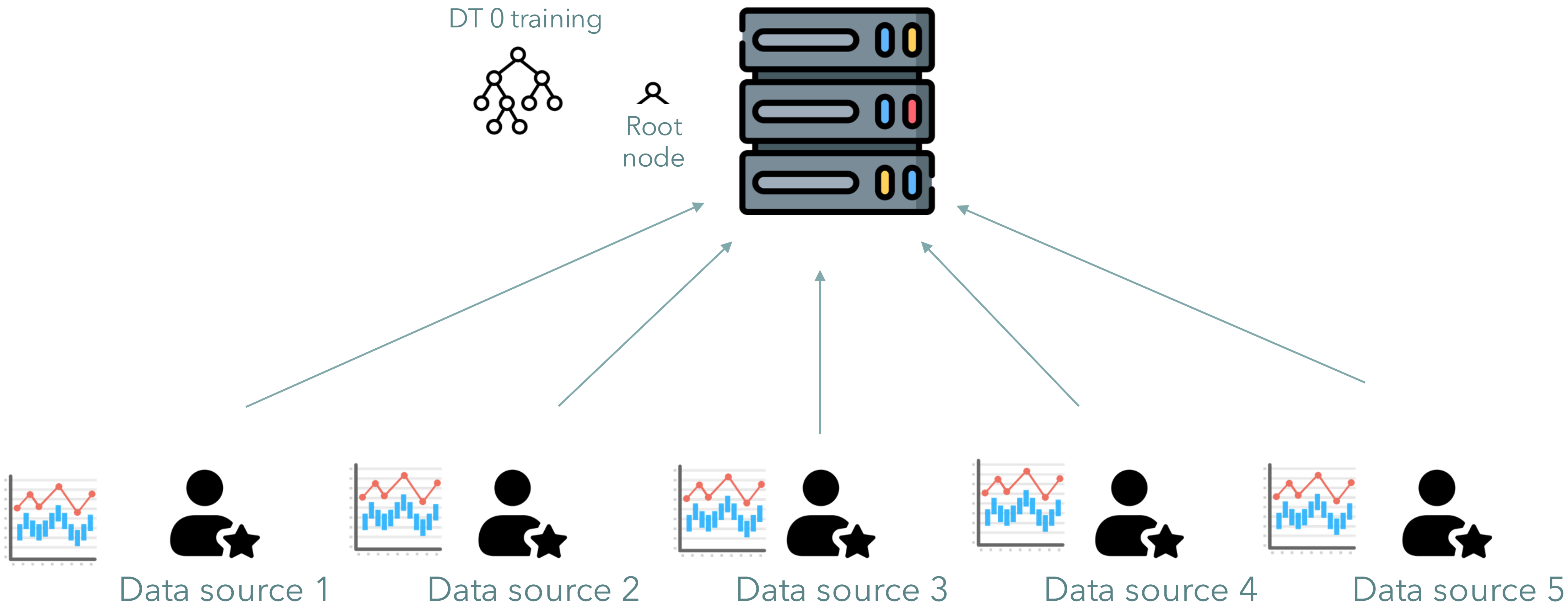
Horizontal Federated Learning – DT Ensembles

Histogram-based aggregation



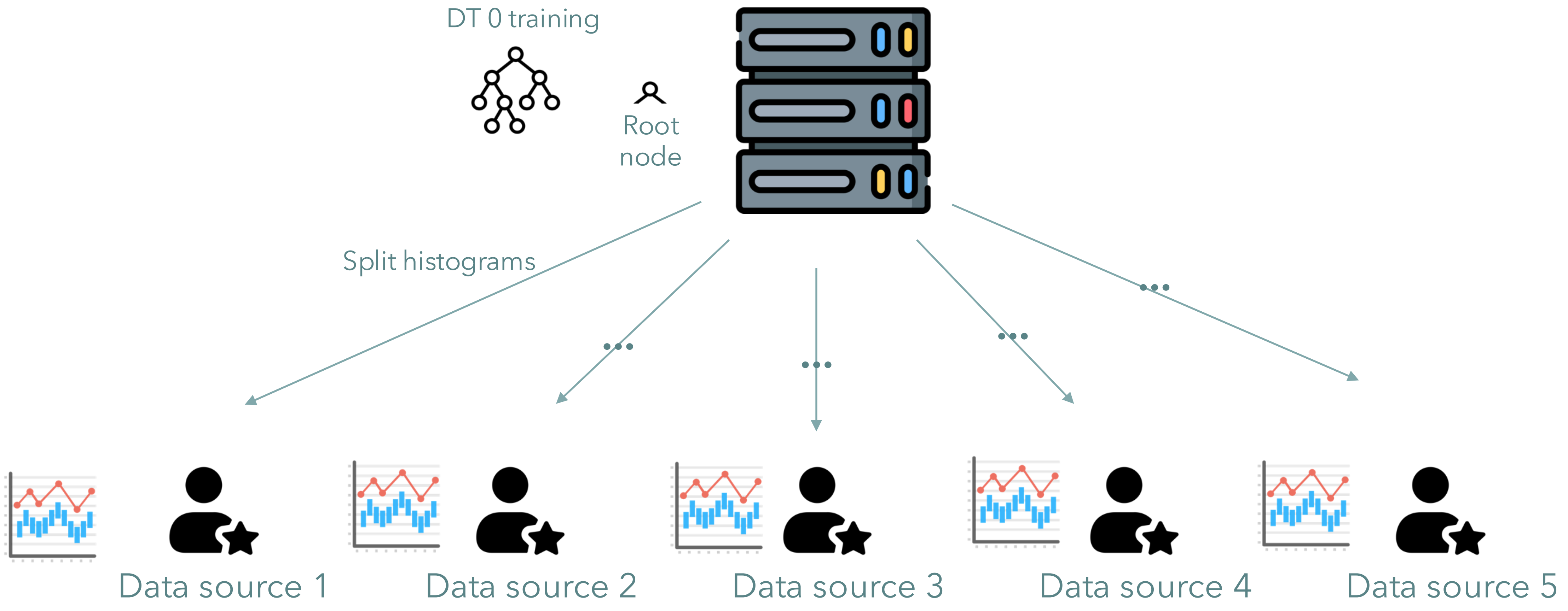
Horizontal Federated Learning – DT Ensembles

Histogram-based aggregation



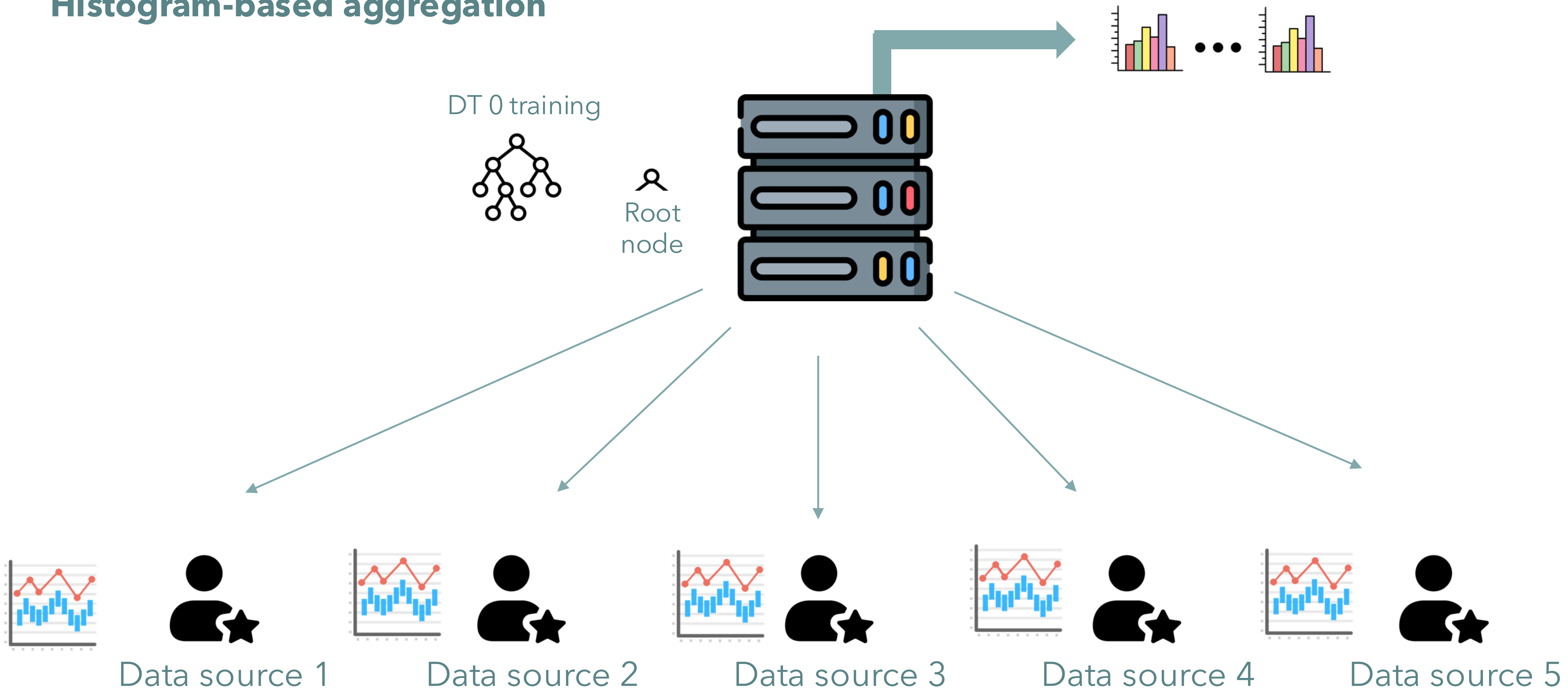
Horizontal Federated Learning – DT Ensembles

Histogram-based aggregation



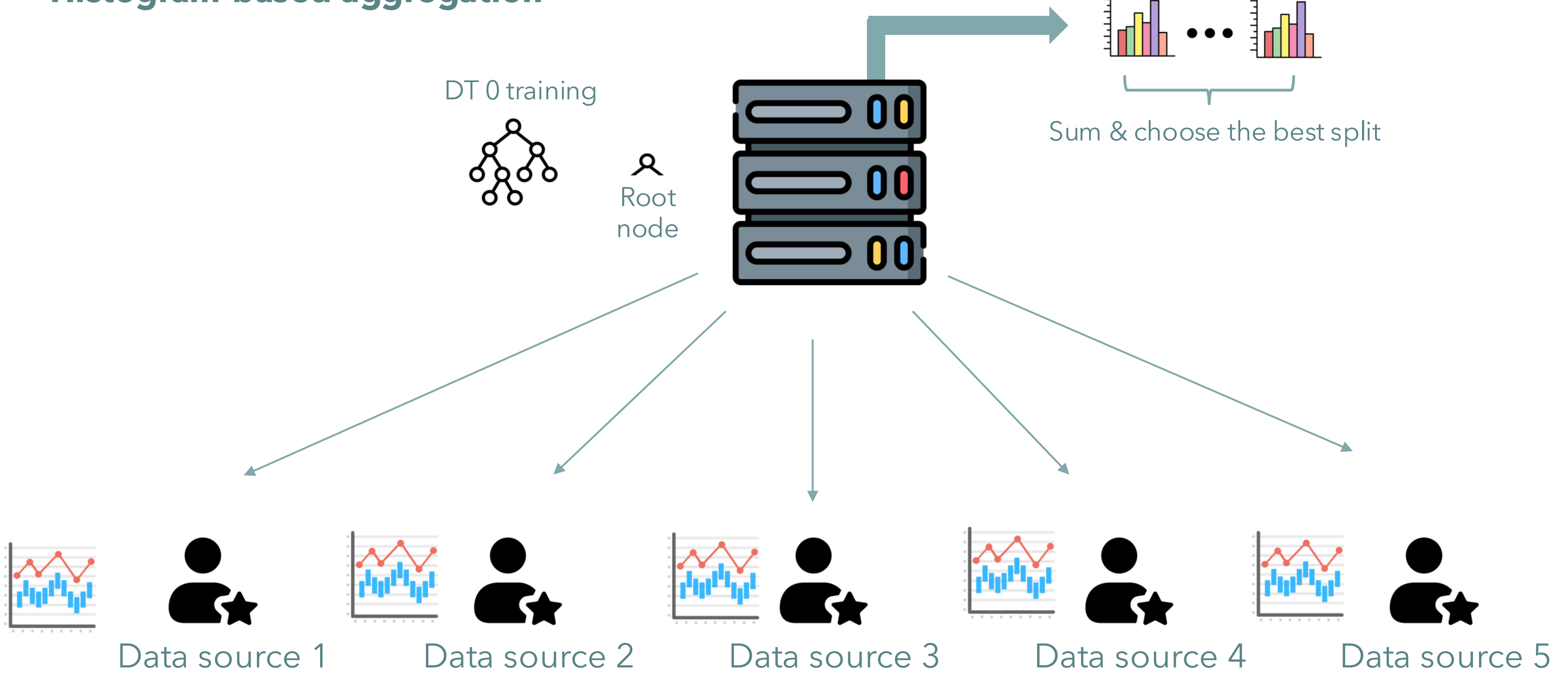
Horizontal Federated Learning – DT Ensembles

Histogram-based aggregation



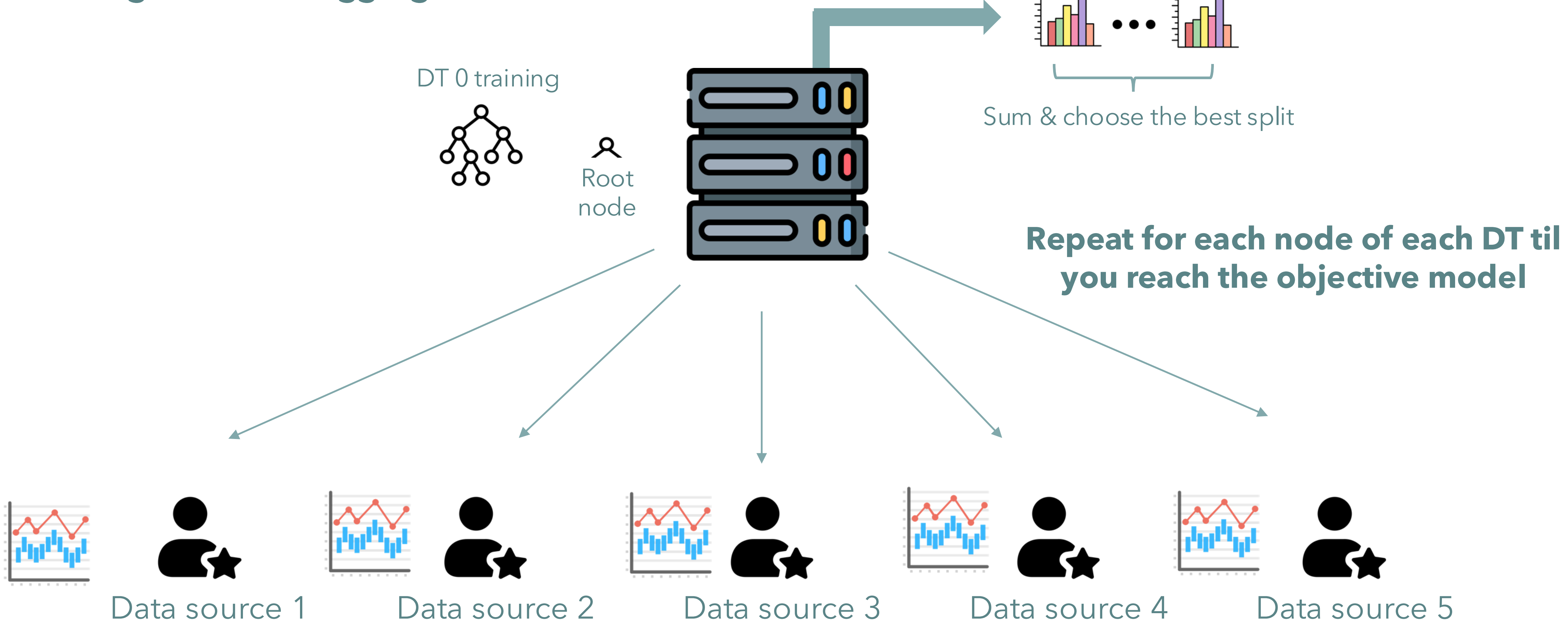
Horizontal Federated Learning – DT Ensembles

Histogram-based aggregation



Horizontal Federated Learning – DT Ensembles

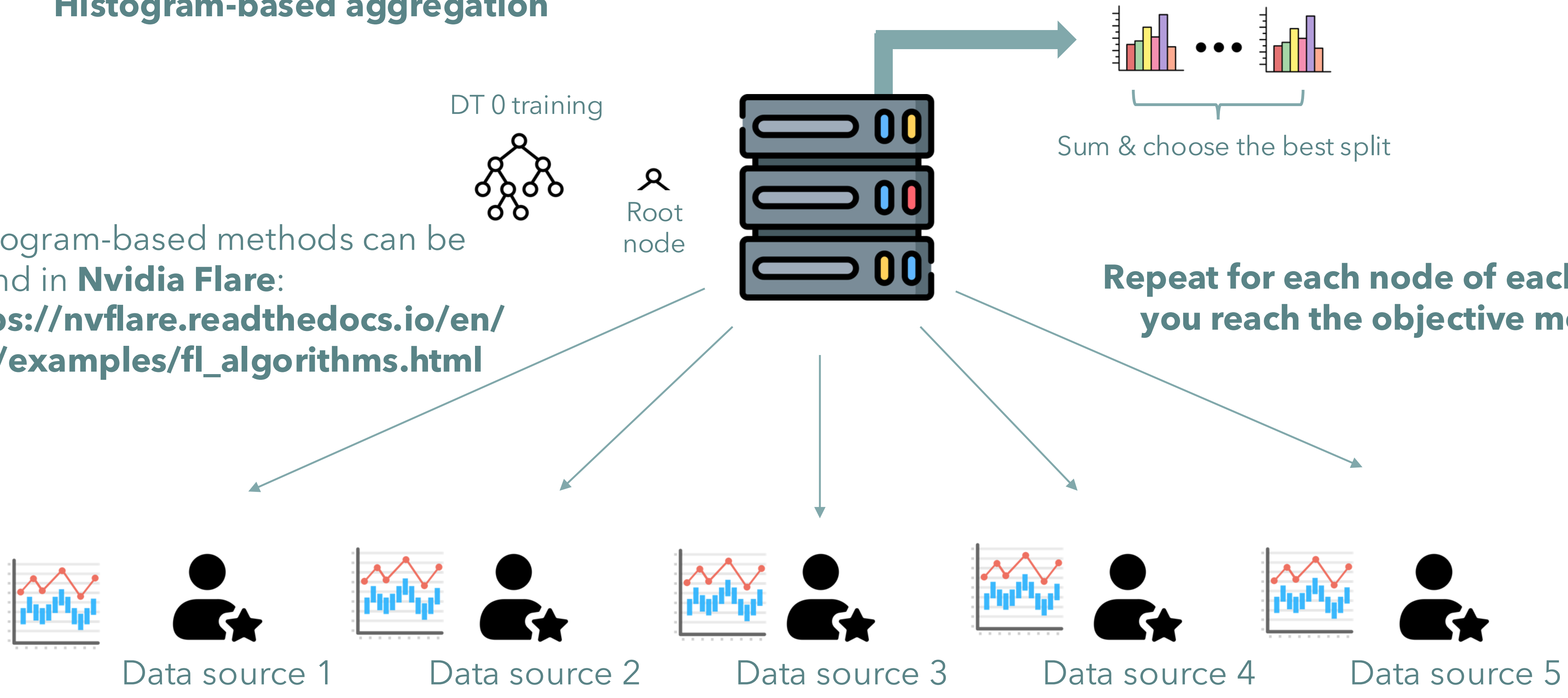
Histogram-based aggregation



Horizontal Federated Learning – DT Ensembles

Histogram-based aggregation

Histogram-based methods can be found in **Nvidia Flare**: **https://nvflare.readthedocs.io/en/2.4/examples/fl_algorithms.html**



Repeat for each node of each DT til you reach the objective model

Horizontal Federated Learning – DT Ensembles

Traditional aggregation methods

Pros
- Fast and simple implementation
- Low communication overhead

Cons
- Accuracy decrease
- Low security w.r.t. malicious attacks

Histogram aggregation methods

Pros
- Accuracy close to centralized training
- High security w.r.t. malicious attacks

Cons
- Harder to implement
- High communication overhead

[3] Di Gennaro, Marco, et al. "TimberStrike: Dataset Reconstruction Attack Revealing Privacy Leakage in Federated Tree-Based Systems." *arXiv preprint arXiv:2506.07605* (2025).

Creativity, Science and **Innovation**

Thank you for your attention

December 1^{st}, 2025

Alessandro Verosimile
alessandro.verosimile@polimi.it