TÉMATERÜLETI KIVÁLÓSÁGI PROGRAM

NEMZETVÉDELEM, NEMZETBIZTONSÁG ALPROGRAM

Quality Inference in Federated Learning with Secure Aggregation

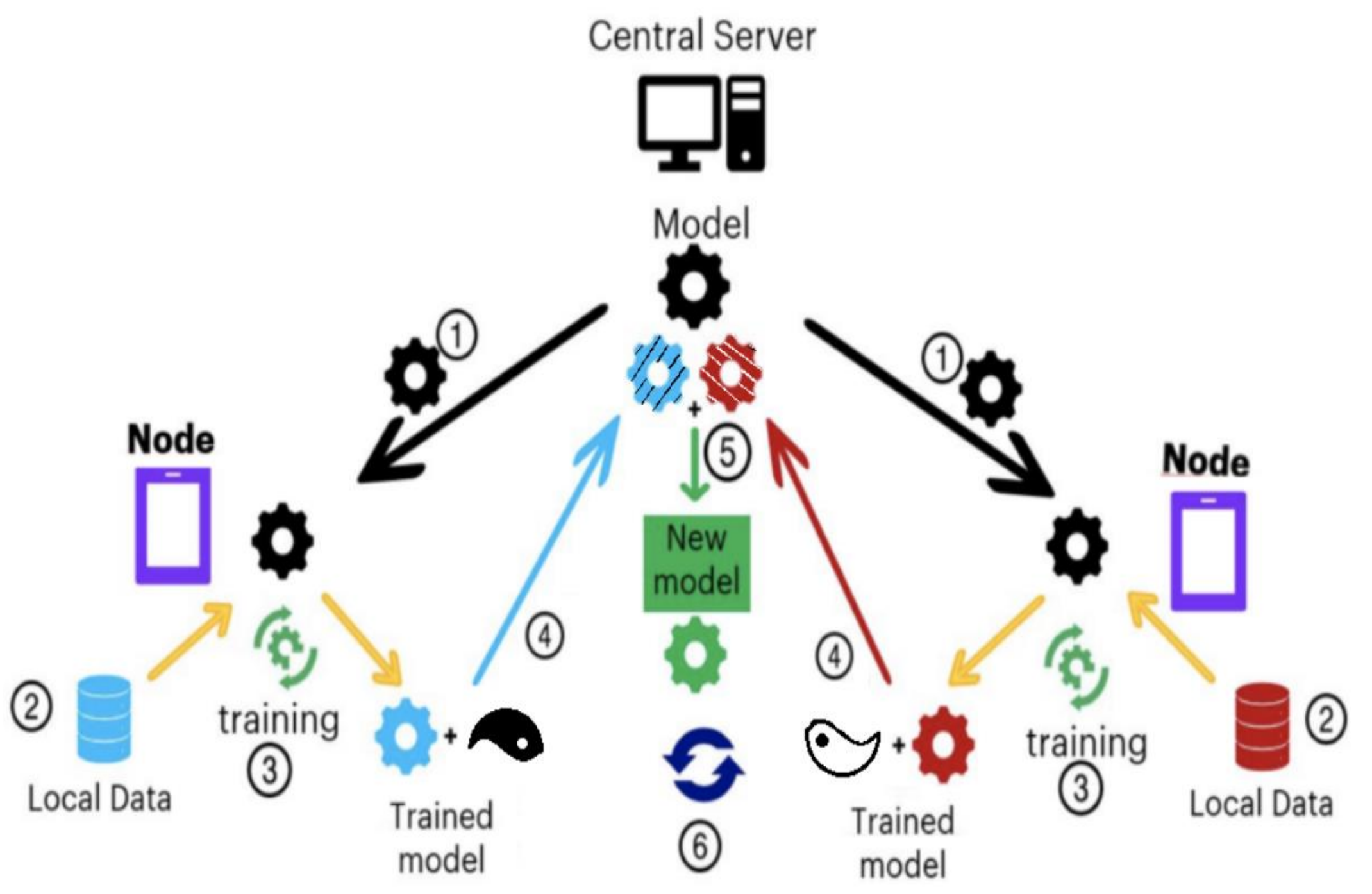
Balázs Pejó & Gergely Biczók - CrySyS Lab

NVA-63: „Az MI adatbiztonsági kérdései" alprojekt

Villamosmérnöki és Informatikai Kar

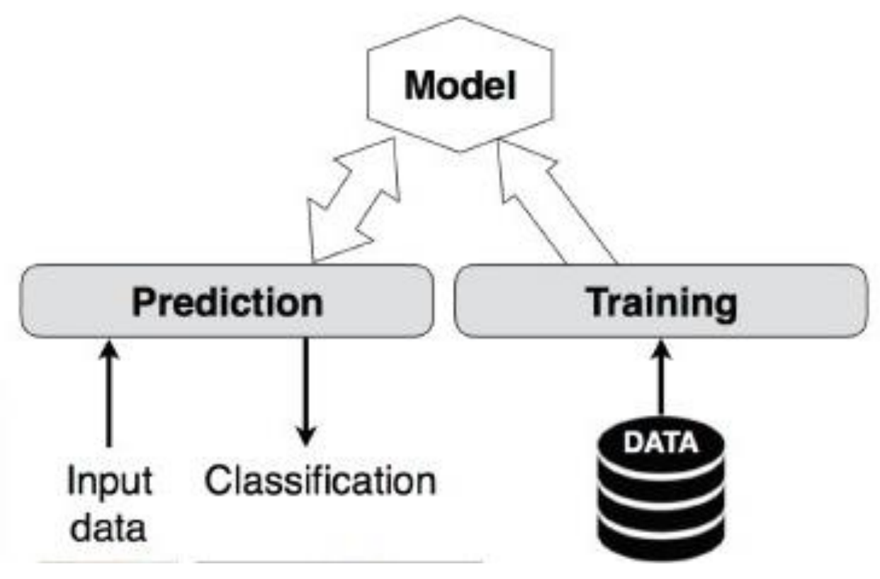
FEDERATED LEARNING WITH SECURE AGGREGATION

- Train locally, share noisy models
- Noise cancels out during aggregation
  - … protect individual privacy
  - … without accuracy loss



MEMBERSHIP INFERENCE

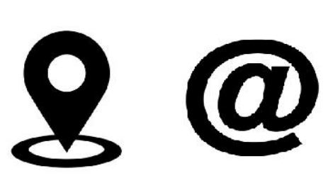
From the model updates it is possible to determine whether a particular data sample was user for training or not.

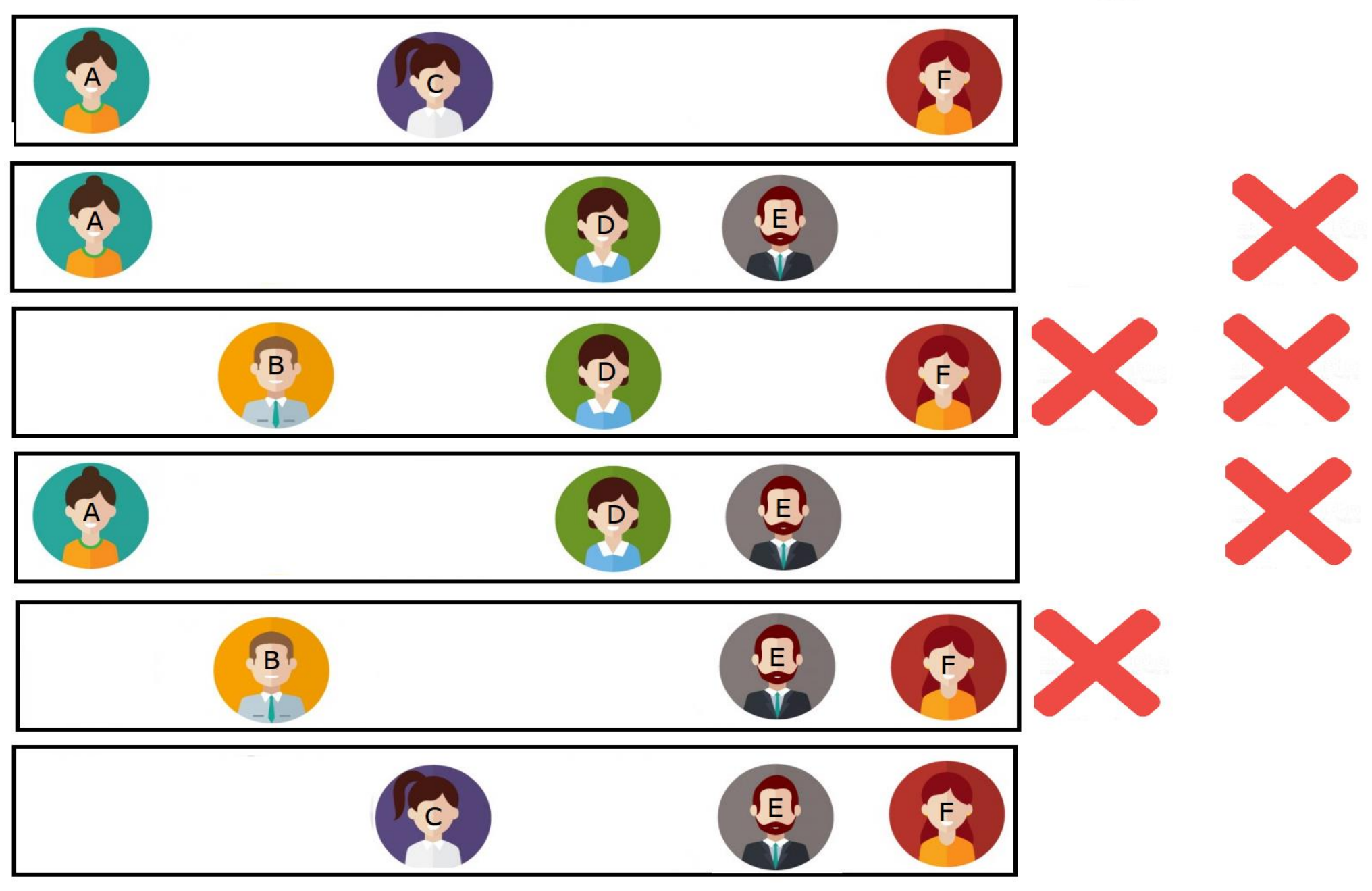


MEMBERSHIP INFERENCE FOR FL WITH SA COULD LEAD TO ATTRIBUTION

Training an NLP Model

- Mail Address ⇒ D
- Location ⇒ B or F





GOAL

Study the possibility of inferring the quality of the individual datasets when Secure Aggregation is in place.

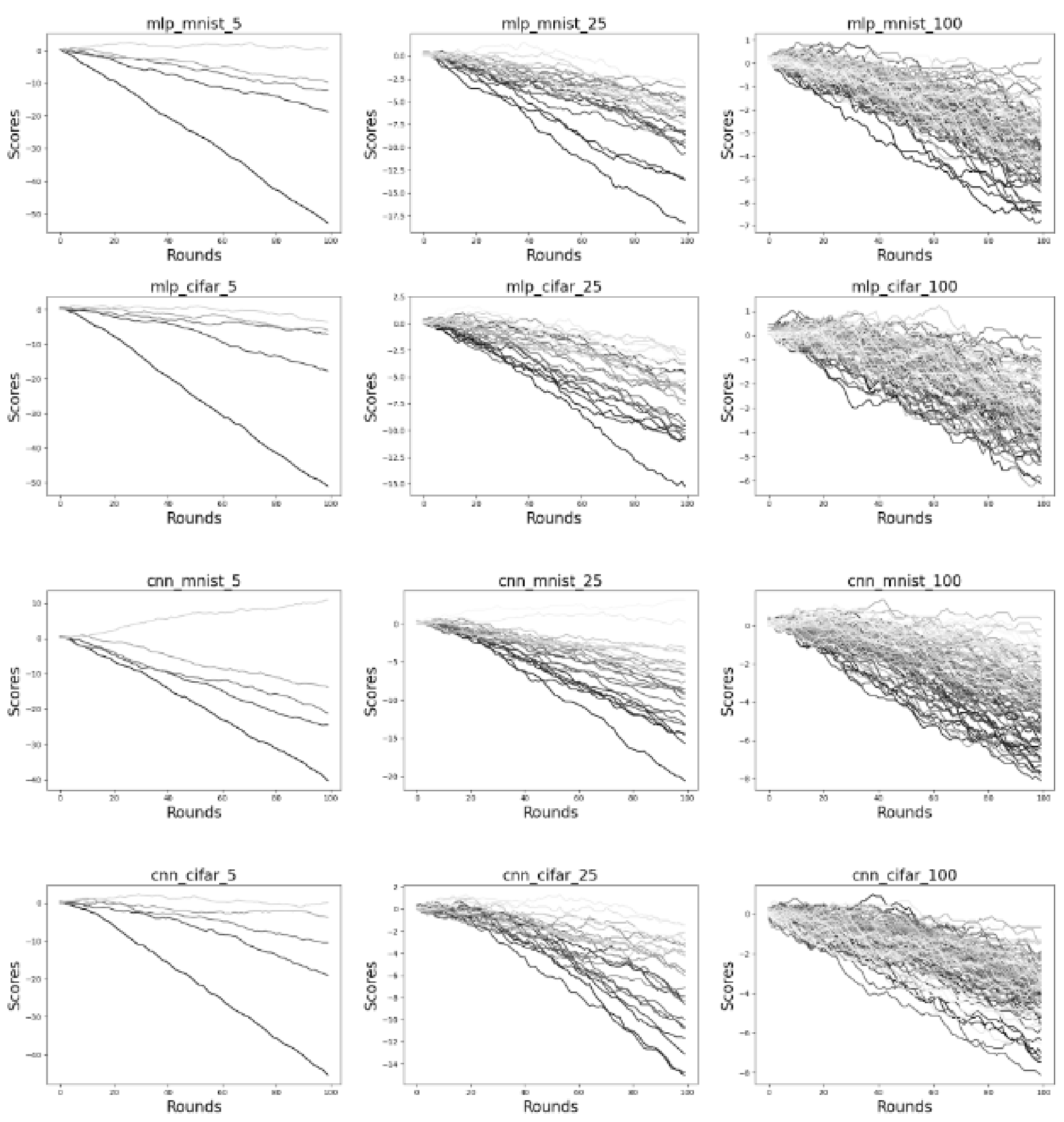
- Quality inference is different from poisoning attack detection, as that merely interested in classifying participants as malicious or benign, while our goal is to enable the fine-grained differentiation of the honest participants with respect to input quality.

SCORING RULES

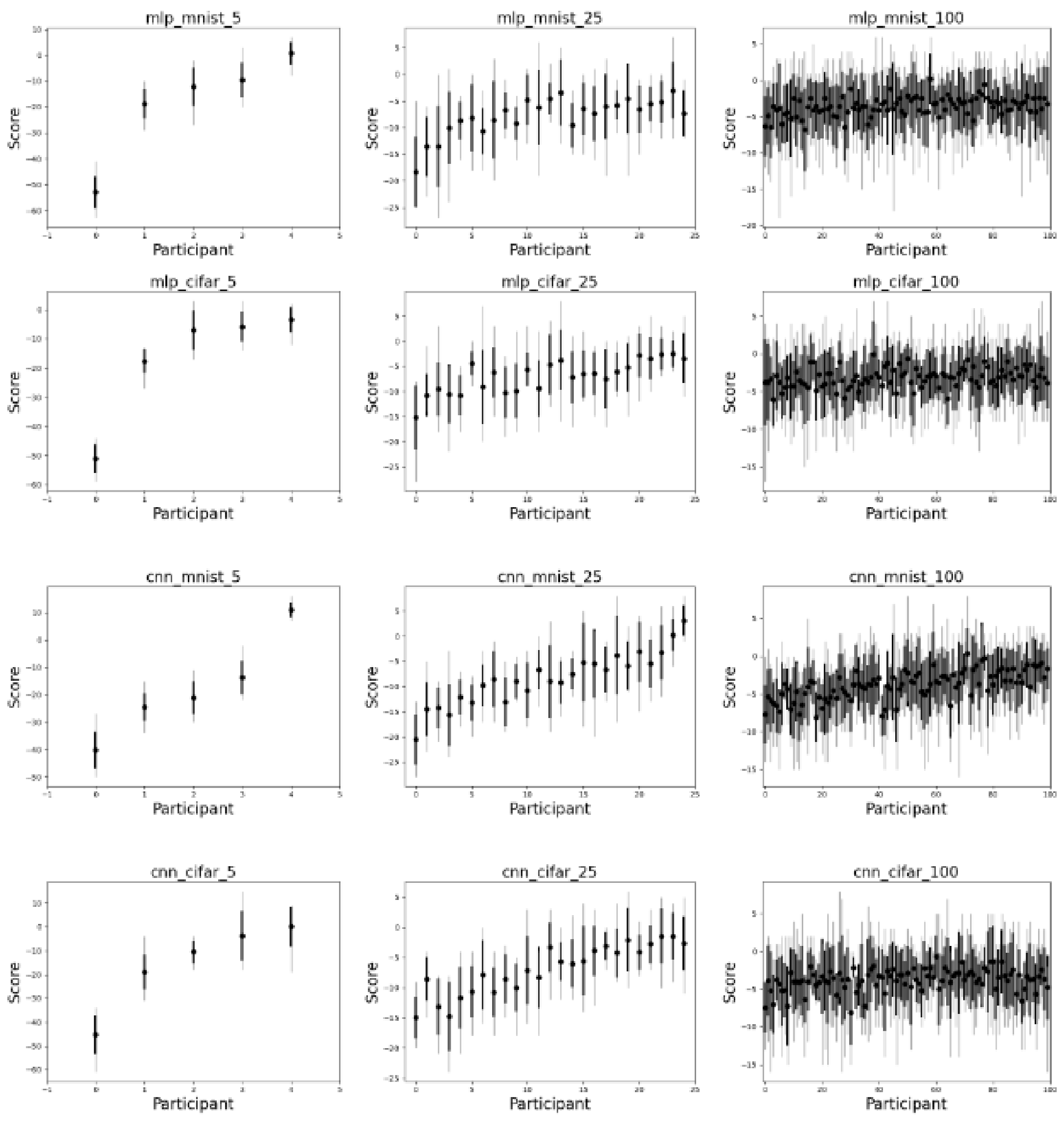
- **The Good**: each participant contributing in a round which improves the model more than the previous round receives +1.
- **The Bad**: each participant contributing in a round which improves the model less than the following round receives -1.
- **The Ugly**: each participant contributing in a round which does not improve the model receives -1.

RESULTS

The round-wise change of the participants' scores: the lighter the better (the darker the worse) corresponding dataset quality.



Quality scores of the participants after 50 rounds where the data quality grows with x axis.



IEEEXPLORE.IEEE.ORG/DOCUMENT/10138056

APPLICATIONS

- **On-the-fly performance boosting**: carefully weighting the participants based on the inferred quality smooths the learning curve as well as improves the trained model's accuracy.
- **Misbehavior detection**: the scores can be used to detect both malicious misbehavior and free-riding.
- **Shapley-Value Approximation**: The scoring rules might be used for contribution score computation, which is currently not solved when Secure Aggregation is enabled.

HEADLINE

Due to the design of federated learning, naïve secure aggregation is not safe:
a few simple quality scoring rules were able to successfully recover the relative ordering of the participant's dataset qualities.

Budapesti Műszaki és Gazdaságtudományi Egyetem

MŰEGYETEM 1782

Nemzeti Kutatási, Fejlesztési és Innovációs Hivatal

AZ NKFI ALAPBÓL MEGVALÓSULÓ PROJEKT