Towards Federated Foundation Models:
Scalable Dataset Pipelines for Group-Structured Learning

Zachary Charles*, Nicole Mitchell*, Krishna Pillutla*, Michael Reneer, Zachary Garrett

**FL research is inhibited by data scale**
Research datasets for FL are often:
- Small
- Difficult to create and customize
- Unsuitable for LLMs

At foundation scale, FL = training with group-structured data

Need for large-scale, group-structured datasets
* scalable, flexible and efficient pipelines to create them

**Dataset Grouper** - a library for creating group-structured datasets
- **Scalable**: can handle millions of clients
- **Flexible**: any custom partitioning of any TFDS/HuggingFace dataset
- **Platform-agnostic**: works with any existing or future platform, e.g. TF, PyTorch, JAX, NumPy...

**Core Features**

1. **Scalable streaming data loaders**
   - Existing hierarchical formats are much slower
   - > 10x slower!
   - Existing in-memory formats don't scale due to large memory requirements

2. **Flexible partitioning of existing datasets**
   ```python
go to gpy

3. **Platform-agnostic group iterators**
   ```python
go to gpy

**FL simulations at scale**

**Model**: O(100M) and O(1B) transformer

**Train**: FedC4
**Eval**: FedBookCO

1. **FedAvg is a meta-learner!**

2. **FedAvg is robust to server learning rate schedules!**

**Installation**
```
pip install dataset-grouper
```

Pull requests welcome!