
# Towards Federated Learning on Fresh Datasets

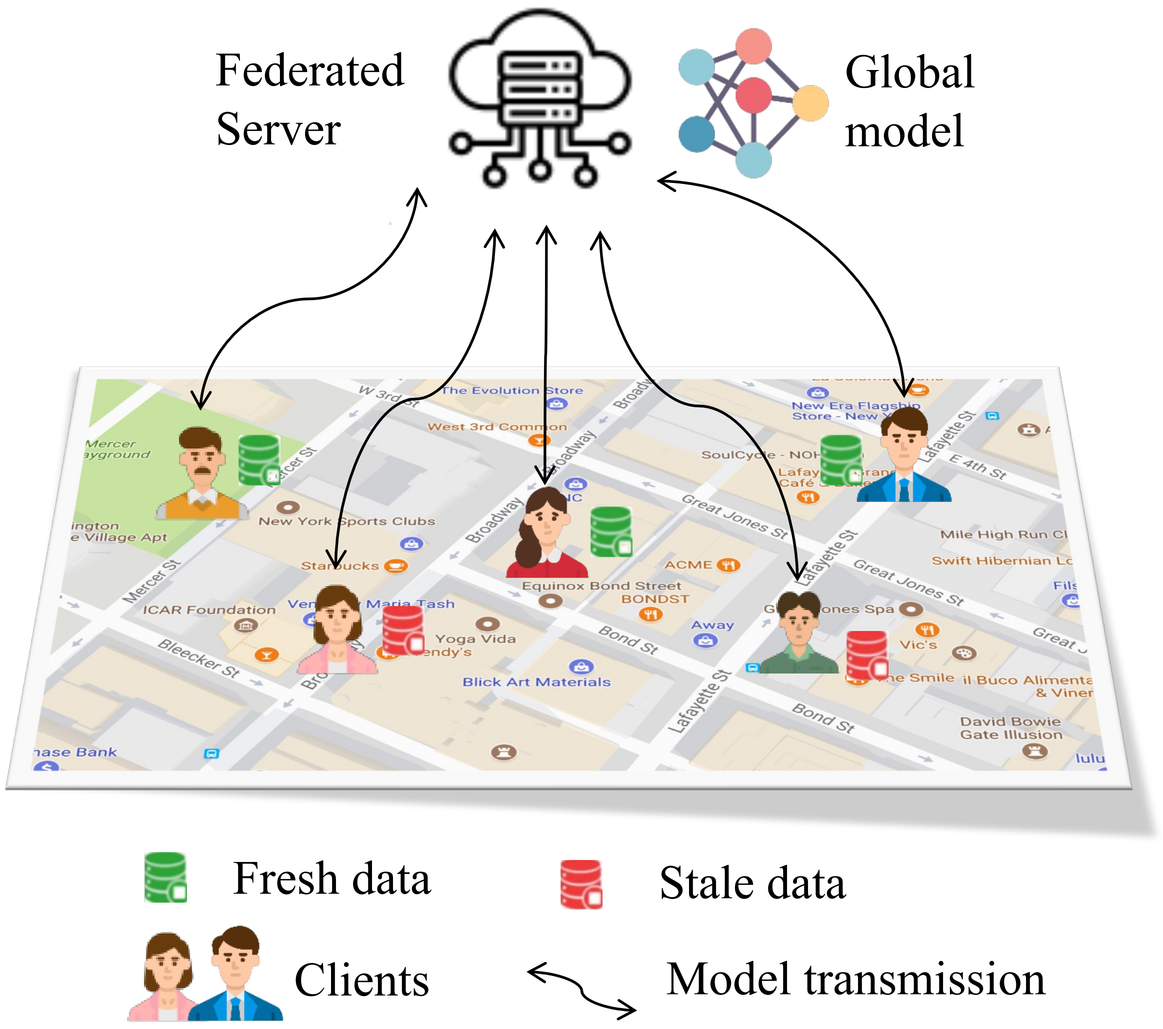
Chen Wu[1], Mingjun Xiao[1], **Jie Wu**[2]
Yin Xu[1], Jinrui Zhou[1], and He Sun[1]

[1]University of Science and Technology of China
[2]Temple University

# Outline

- Motivation & Challenges
- Preliminaries & Problem Formulation
- Basic Idea & Solution
- Evaluation & Conclusion

# Motivation




## Traditional FL

- **Invariability:** clients' local datasets are static;
- **Inadaptability:** data in real-world are continuously generated along with the time.

*Update Datasets* *Train with Fresh Data*

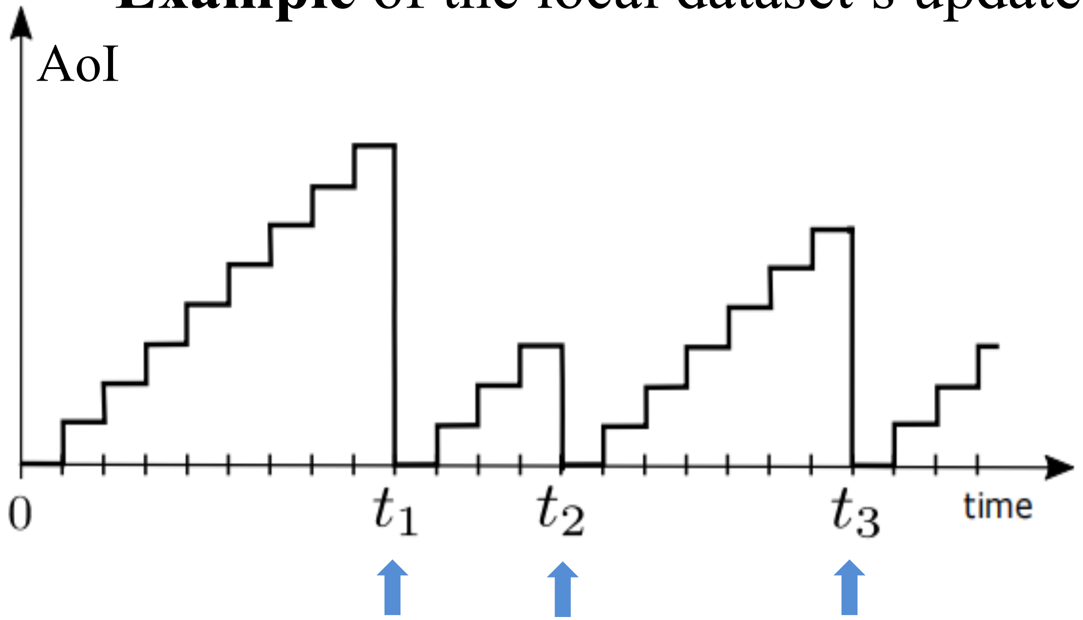
## FL on Fresh Datasets

- **Variability:** clients collect new data periodically;
- **Freshness of Models:** fresh data can accurately characterize the model parameters;
- **Budget Limit**: clients spend some extra costs while the total budget is limited.

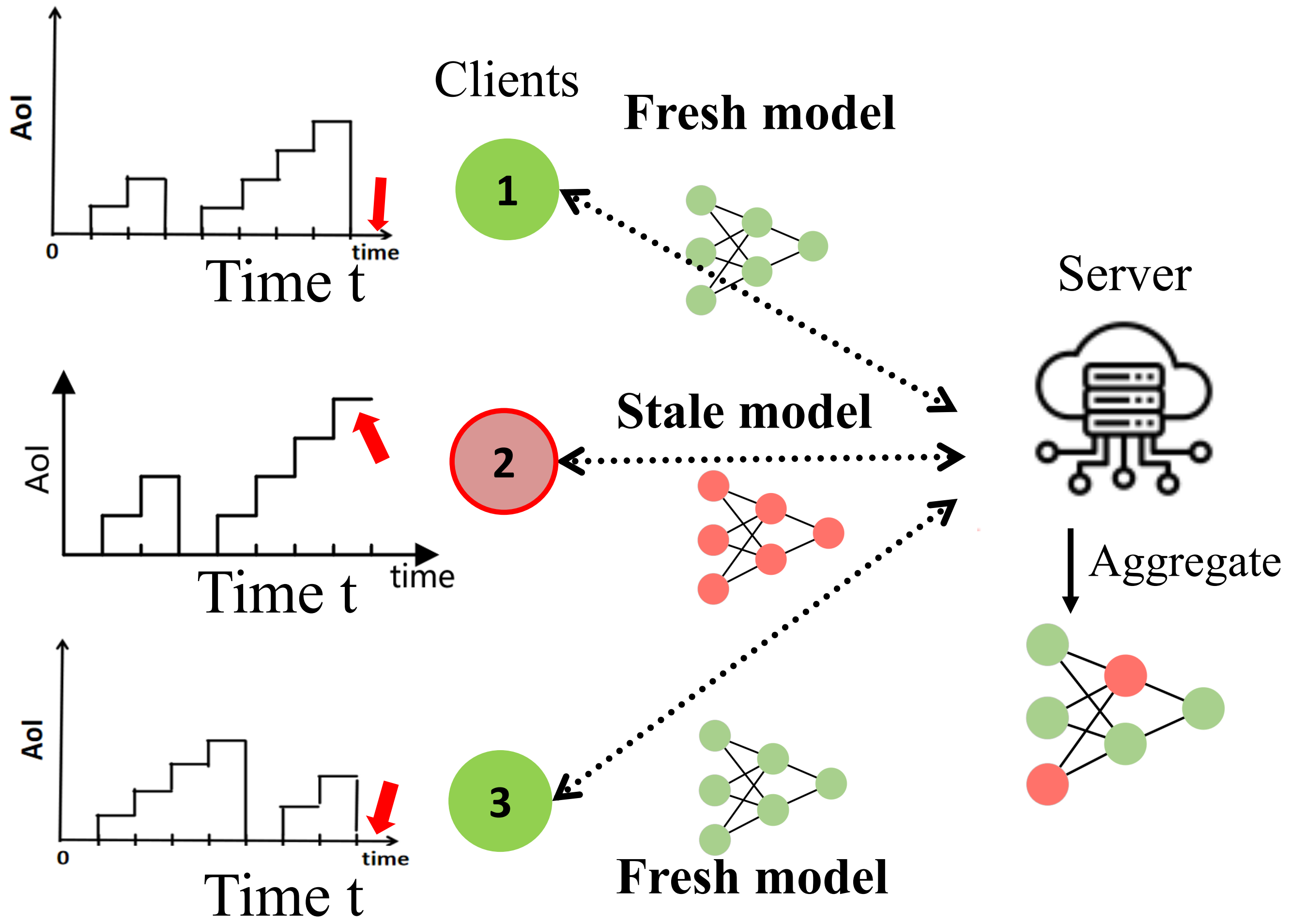
# AoI: Metric for Measuring Data Freshness

***Age of Information (AoI) :*** *freshness of the local dataset --- the time elapsed from the data collection to its usage.*

**Example** of the local dataset's update




AoI evolution with updates at $t_1, t_2$, and $t_3$.



# Challenges

- Selected clients: **update** local datasets & **reduce** the AoI values

→ **Quantify** the impact of AoI on the model training of FL

→ Reveal the **relationship** between the **loss** of global model and the **decrease** of the AoI values of clients' datasets?

- **Dependence**: client selection and the corresponding AoI values

→ Design a client selection strategy to **optimize the performance** of the global model (i.e., **global loss**) within a **budget**?

# Related Work

- **Client Selection:** make decisions under different optimization objectives

  e.g., Huang T, Lin W, Wu W, et al. "An efficiency-boosting client selection scheme for federated learning with fairness guarantee", in IEEE TPDS, 2020, 32(7): 1552-1564.

- **AoI Optimization:** minimize AoI under different scenarios

  e.g., Lim W Y B, et al. "When information freshness meets service latency in federated learning: A task-aware incentive scheme for smart industries", in IEEE TII, 2020, 18(1): 457-466.

- **Restless MAB:** all bandits might evolve stochastically

  e.g., Whittle P. "Restless bandits: Activity allocation in a changing world", in Journal of applied probability, 1988, 25(A): 287-298.

Ignore the importance of data freshness

Ignore the relationship between AoI & loss

We aim to design a clients selection mechanism for FL while considering **data freshness** and **limited budget simultaneously**.

# Contributions

- ✓ **System:** Introduce a **novel AoI-aware FL** considering the freshness of the local datasets for client selection.
- ✓ **Analysis:** Derive a **relationship** between the training **loss** of the global model and the **AoI** values of local datasets.
- ✓ **Algorithm:** Propose the **Whittle's Index-based Client Selection (WICS)** algorithm and prove its **approximate optimality**.
- ✓ **Experiments:** Evaluate WICS by using real-world datasets (i.e., MNIST and FMNIST) to verify its performance.

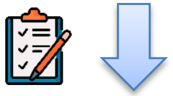
# System Model

## System Model


- **Clients** $\{\mathbf{1}, \ldots, \boldsymbol{i} \ldots, \boldsymbol{N}\}$: each client $i$ collects fresh data and use its local dataset $D_t^i$ to train its local model
- **Cost** $p_i$**:** the payment for fresh data collection to client $i$ from the server
- **Average AoI:** the time elapsed since the client updates this dataset: $\Delta_i\,(t) = t - u_i(t)$

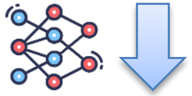
## Procedure

① Server selects a subset of clients $N_t$ to update their local datasets.



② Client $i$: train its local model using local dataset and upload its model.

③ Server aggregrates local models to obtain the global model.



④ Server pays the data collection cost $p_i$ to selected clients.

# Model Training of FL

- Step 1: each client $i$ conducts local training with data size $|D_t^i| = n_i$.

Compute Local Loss $$F_{t,i}(\omega; \mathcal{D}_t^i) = \frac{1}{|\mathcal{D}_t^i|} \sum_{x \in \mathcal{D}_t^i} f(\omega; x),$$

Update Parameters $$\omega_t^{i,k+1} = \omega_t^{i,k} - \eta_t \nabla F_{t,i}(\omega_t^{i,k}; \xi_t^{i,k}),$$

where $f(.)$ is the a server-specified loss function, $\eta_t$ is the learning rate, $k = \{1,2,\dots,\tau\}$ is the index of local iterations, and $\xi_t^{i,k}$ is the $k$-th mini-batch sampled from the dataset $D_t^i$.

- Step 2: the server aggregates received local models.

Aggregate Models $$\omega_t = \sum_{i=1}^{N} \frac{n_i}{n} \omega_t^i, \quad \text{where} \quad n = \sum_{i=1}^{N} n_i$$

Global Loss Function $$F(\omega) \triangleq \frac{1}{T} \sum_{t=1}^{T} \sum_{i=1}^{N} \frac{n_i}{n} F_{t,i}(\omega; \mathcal{D}_t^i).$$

**Goal:** find the optimal global model parameters $$\omega^* = \arg\min_{\omega} F(\omega).$$

# Problem Formulation

## ➢ Original Optimization problem:

$$\mathbf{P1}: \quad \min_{\pi \in \Pi} \ \mathbb{E}[F(\omega_T)] - F^*,$$

$$\text{s.t.} \quad a_i^\pi(t) \in \{0,1\}, \forall i \in \mathcal{N}, \forall t \in \mathcal{T},$$

$$\Delta_i(t) = \mathbb{1}_{\{a_i^\pi(t)=0\}} \left[\Delta_i(t-1)+1\right],$$

$$\sum_{i=1}^N a_i^\pi(t) p_i \le B, \ \forall t \in \mathcal{T}.$$

Goal

Constraints

**Optimization Objective:**
Find a client selection strategy $\boldsymbol{\pi}^*$ that minimizes the gap between the **expected global loss** after T rounds and the **optimal global loss**.

- ☐ **Constraint 1:** client *i* is selected in the $t^{\text{th}}$ time slot. $a_i^\pi(t) = 1$ is selected; otherwise, $a_i^\pi(t) = 0$.
- ☐ **Constraint 2:** the dynamics of each client's AoI, where $1_{\{.\}}$ is an indicator function.
- ☐ **Constraint 3:** the budget constraint of the server in each round of FL.

# Conergence Analysis

**Assumption 1**

For *all* $t, i, F_{t,i}$ is **$\boldsymbol{\beta}$-smooth**, that is, for $\forall \omega_1, \omega_2, F_{t,i}(\omega_2) - F_{t,i}(\omega_1) \leq$ $< \nabla F_{t,i}(\omega_1), \omega_2 - \omega_1 > + \frac{\beta}{2} ||\omega_2 - \omega_1||^2$.

**Assumption 2**

For all $t, i, F_{t,i}$ is **$\boldsymbol{\mu}$-strongly-convex**, that is, for $\forall \omega_1, \omega_2, F_{t,i}(\omega_2) - F_{t,i}(\omega_1)$ $\geq < \nabla F_{t,i}(\omega_1), \omega_2 - \omega_1 > + \frac{\mu}{2} ||\omega_2 - \omega_1||^2$.

**Assumption 3**

For all $t, i$, the stochastic gradients of loss function is **unbiased**, i.e., $E_\xi[\nabla F_{t,i}(\omega; \xi)] = \nabla F_{t,i}(\omega)$.

**Assumption 4**

For all $t, i$, the expected squared norm of stochastic gradients is **AoI-aware bounded**: $E_\xi \left|\left|\nabla F_{t,i}(\omega; \xi)\right|\right| \leq G_i^2 + \boldsymbol{\Delta_i(t)\sigma_i^2}$.

$\Delta_i(t)$-- AoI; $\sigma_i^2$--**sensitivity** of client's local data to freshness; $G_i^2$--client's inherent bound

**Note:** Assumption 4 is an extension of the hypothesis in existing FL, considering the impact of data freshness on training. It is applicable to **mean absolute loss, mean squared loss,** and **cross entropy loss**.

# Step 1: Conergence Analysis

**Theorem 1 (Convergence Upper Bound).** Define $\bar{\eta}$= $\min_t\{\eta_t\}$ and $\tilde{\eta} = \max_t\{\eta_t\}$. Suppose that Assumptions 1 to 4 hold and the step size meets $\bar{\eta} < \frac{2}{\mu}$. Then, the FL training loss after the initial global model $\omega_0$ is updated for T rounds satisfies:

$$\mathrm{E}[\mathrm{F}(\omega_T)] - F^* \leq \frac{\beta}{2}\left(1 - \frac{\mu\bar{\eta}}{2}\right)^2 + \frac{\beta}{2}\sum_{t=1}^{T}\sum_{i=1}^{N}\boldsymbol{\alpha_i}[\boldsymbol{G_i^2} + \Delta_{\mathbf{i}}(\mathbf{t})\boldsymbol{\sigma_i^2}],$$

where $\alpha_i = \frac{\tilde{\eta} n_i}{\mu n} + N\tilde{\eta}(\tau^2\tilde{\eta} + \frac{2(\tau-1)^2}{\mu}\frac{n_i^2}{n^2})$.

**NOTE:** controlling $\sum_{t=1}^{T}\sum_{i=1}^{N}\alpha_i\,\Delta_i(t)\sigma_i^2$ can control the convergence of the model.

# Restless Multi-Armed Bandit

- **Modeling:** a **Restless Multi-Armed Bandit** (RMAB) --- a generalization of MAB problem
- **Characteristic:** any number of bandits (more than 1) can be made **active** and all bandits might **evolve stochastically**.

| RAMB | Our problem |
|---|---|
| Restless bandit | Each client |
| State | AoI value |
| Reward | Fresh local model |

# Step 2: Convert Problem

➢ **Converted Optimization problem:**

$$\mathbf{P2}: \quad \min_{\pi\in\Pi} \frac{1}{TN}\sum_{t=1}^{T}\sum_{i=1}^{N}\phi_i\Delta_i(t),$$

$$\text{s.t.} \quad a_i^\pi(t)\in\{0,1\}, \forall i\in\mathcal{N}, \forall t\in\mathcal{T},$$

$$\Delta_i(t) = \mathbb{1}_{\{a_i^\pi(t)=0\}}\left[\Delta_i(t-1)+1\right],$$

$$\sum_{i=1}^{N} a_i^\pi(t)p_i \le B, \ \forall t\in\mathcal{T}.$$

Goal

Constraints

**Optimization Objective:**
According to Theorem 1, finding the optimal strategy $\boldsymbol{\pi}$ for **Problem P1** can be converted for **Problem P2**.

**Note:** $\phi_i = \frac{\alpha_i\sigma_i^2\beta NT}{2}$

- **Constraint 1:** client $i$ is selected in the $t^{\text{th}}$ time slot. $a_i^\pi(t)=1$ is selected; otherwise, $a_i^\pi(t)=0$.
- **Constraint 2:** the dynamics of each client's AoI, where $1_{\{.\}}$ is an indicator function.
- **Constraint 3:** the budget constraint of the server in each round of FL.

# Step 3: Relaxation and Decoupling

- **Relax** Constraint 3: $\sum_{i=1}^{N} a_i^{\pi}(t) p_i \leq B$ ➡ $\frac{1}{TN}\sum_{t=1}^{T}\sum_{i=1}^{N} a_i^{\pi}(t)\frac{p_i}{B} \leq \frac{1}{N}$
- Transform Problem **P2** into the **Lagrangian Dual Problem P3:**

**P3 :** $\boldsymbol{max_{\lambda} min_{\pi}}\ \boldsymbol{\mathcal{L}(\pi,\lambda)} = \frac{1}{TN}\sum_{t=1}^{T}\sum_{i=1}^{N}\phi_i \Delta_i(t) + \lambda[\frac{1}{TN}\sum_{t=1}^{T}\sum_{i=1}^{N} a_i^{\pi}(t)\frac{p_i}{B} - \frac{1}{N}]$

s.t.

$$\Delta_i(t) = \mathbb{I}_{\{a_i^{\pi}(t)=0\}}[\Delta_i(t-1)+1],$$

$$a_i^{\pi}(t) \in \{0,1\},\ \ \lambda \geq 0.$$

- Solve $\boldsymbol{min_{\pi}\mathcal{L}(\pi,\lambda)}$: finding the optimal strategy $\boldsymbol{\pi}$ for any given $\boldsymbol{\lambda}$;
  Problem **P3** can be **decoupled** to Problem **P4:**

**P4 :** $\min_{\pi\in\Pi}\left\{\lim_{T\to+\infty}\frac{1}{T}\sum_{t=1}^{T}\left[\frac{B\phi_i}{p_i}\Delta_i(t)+\lambda a_i^{\pi}(t)\right]\right\}$

s.t.

$$a_i^{\pi}(t)\in\{0,1\},\forall i\in\mathcal{N},\forall t\in\mathcal{T},$$

$$\Delta_i(t) = \mathbb{1}_{\{a_i^{\pi}(t)=0\}}\left[\Delta_i(t-1)+1\right],$$

$$\lambda \geq 0.$$

# Step 4: Solving Problem P4

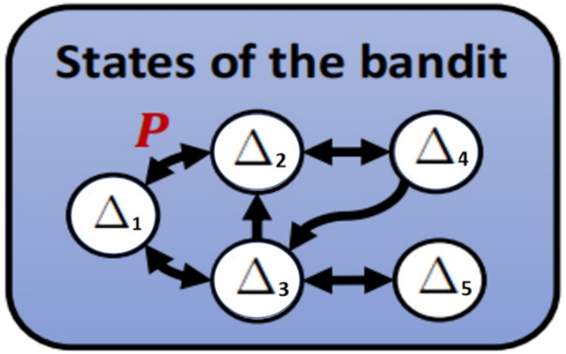
- **Formulation:** The decoupled problem can be formulated as a **Markov Decision Process (MDP)** with AoI state $\Delta_i(t)$, control variable $a_i^\pi(t)$, state transition $P(\cdot)$, and cost function $C(\cdot)$.


**State Transition**

$$\mathbb{P}(\Delta_i(t+1) = \Delta_i(t) + 1 | a_i^\pi(t) = 0) = 1;$$
$$\mathbb{P}(\Delta_i(t+1) = 0 | a_i^\pi(t) = 0) = 0;$$
$$\mathbb{P}(\Delta_i(t+1) = \Delta_i(t) + 1 | a_i^\pi(t) = 1) = 0;$$
$$\mathbb{P}(\Delta_i(t+1) = 0 | a_i^\pi(t) = 1) = 1;$$

**Cost Function**

$$C_i(\Delta_i(t), a_i^\pi(t)) \triangleq \frac{B\phi_i}{p_i}\Delta_i(t) + \lambda a_i^\pi(t)$$




**NOTE**: the Lagrange multiplier λ is a kind of service charge for client $i$ under the MDP model, generated only when $a_i^\pi(t) = 1$.

# Step 4: Solving Problem P4

➢ **Solving MDP →** Get the optimal strategy for the decoupled problem (P4)

**Theorem 2 (Optimal Strategy for MDP):** Consider the decoupled model over an infinite time-horizon. Given $\boldsymbol{\lambda}$, the optimal strategy $\pi^*$ is selecting client $i$ in each time slot $t$ to update its local dataset only when $\Delta_i(t) > H_i - 1$, where

$$H_i = \left\lfloor -\frac{1}{2} + \sqrt{\frac{1}{4} + \frac{2\lambda p_i}{B\phi_i}} \right\rfloor.$$

$\boldsymbol{\pi}^*$:

client $i$

If $\Delta_i(t) > H_i - 1$ → Selected ✓ YES

If $\Delta_i(t) \leq H_i - 1$ → Not Selected ✗ NO

# Step 5: Approximately Solve Problem P3 (and P2)

- Solving $\boldsymbol{max_{\lambda}\,\mathcal{L}(\pi^{*},\lambda)}$: finding the optimal $\boldsymbol{\lambda}$ is difficult.
- Using the **Whittle's approximation method:**

  Find a $\lambda_i$ to maximize the objective function for each decoupled problem separately;
  Each $\lambda_i$ also follows Theorem 2;

$$WI_{i,t} \triangleq \lambda_i(\Delta_i(t)) = \frac{(\Delta_i(t)+1)(\Delta_i(t)+2)B\phi_i}{2p_i}$$

$WI_{i,t}$ is the **Whittle's index** for client $i$.

**Whittle's Index-based Client Selection (WICS)** → $\boldsymbol{P_3}$(and $\boldsymbol{P_2}$ based on duality)

**Basic idea:** Select the clients with higher WI values in each time slot under budget constraint $B$.

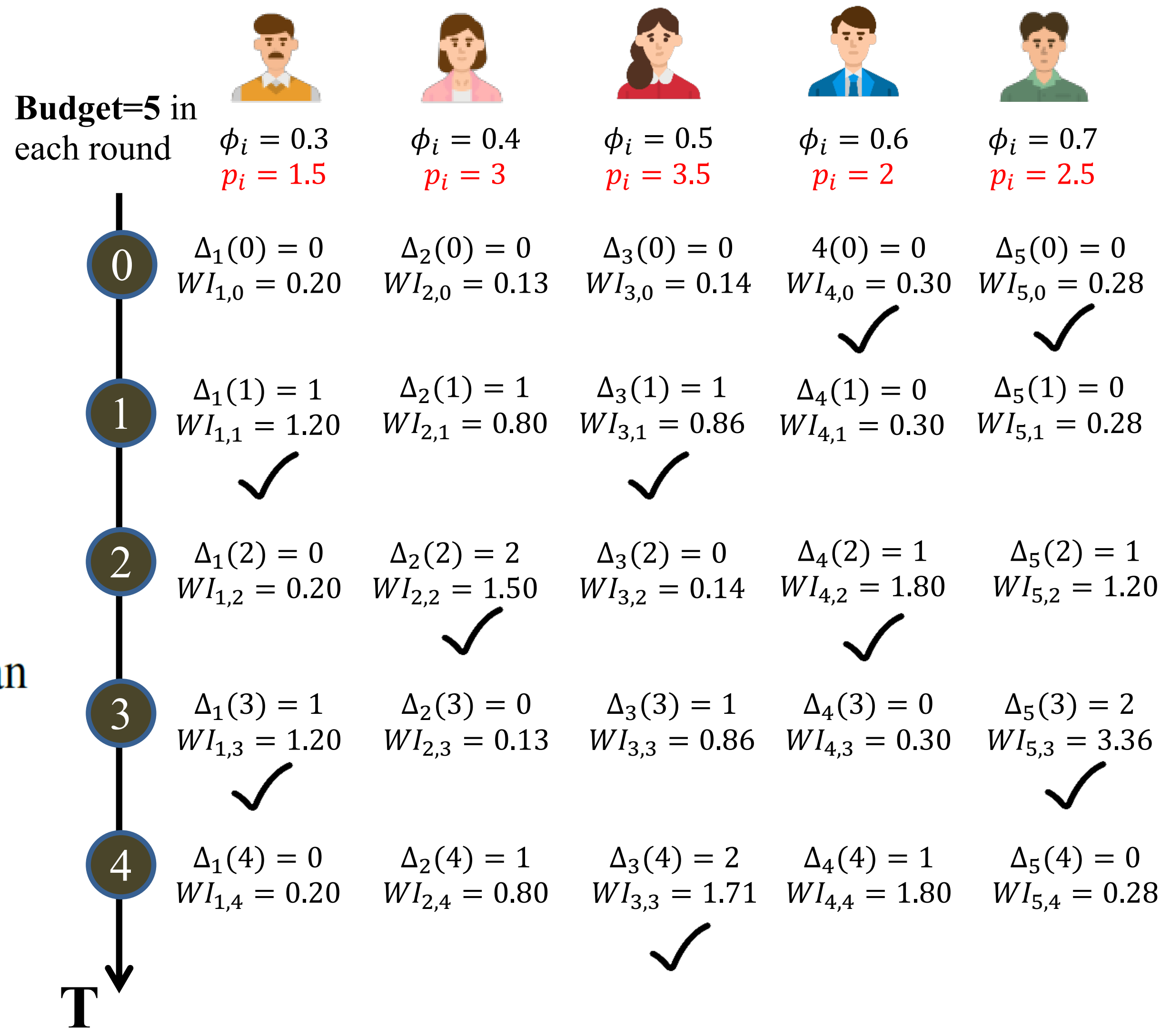
# Algorithm

**Algorithm 1:** Whittle's Index based Client Selection

**Input:** AoI value of each client $\{\Delta_1(t), \cdots, \Delta_N(t)\}$, weight of each client $\{\phi_1, \cdots, \phi_N\}$, payment of each client $\{p_1, \cdots, p_N\}$, budget $B$

**Output:** The index set of selected clients $\mathcal{N}_{t+1}$

**for** each client $i$ in $\mathcal{N}$ **do**
  Calculates its WI value $WI_{i,t}$ according to Eq.(18) and sends it to the server
**end for**
The server sorts the clients into $(i_1, i_2, \cdots, i_N)$ such that $WI_{i_1,t} \geq WI_{i_2,t} \geq \cdots \geq WI_{i_N,t}$, and initializes an empty set $\mathcal{N}_{t+1}$, an initial index $k = 1$
**while** $\sum_{i \in \mathcal{N}_{t+1}} p_i + p_{i_k} < B$ **do**
  $\mathcal{N}_{t+1} \leftarrow \mathcal{N}_{t+1} \cup \{i_k\}$, $k = k + 1$
**end while**



# Algorithm Analysis

**Theorem 3 (Approximate Optimality):** The solution produced by the WICS algorithm for Problem P2 over an infinite time-horizon is $\boldsymbol{\rho^{WI}}$**-optimal**, where

$$\rho^{WI} < \frac{18N-2}{M-1}$$

Here, $M = \left\lfloor \frac{B}{p_{max}} \right\rfloor$ *and* $p_{max} = \max_i\{p_i\}$.

**NOTE**: $\rho^{WI}$ will not be too large.

# Experimental Settings

## Dataset and Model

- **Dataset**: MNIST and FMNIST (60,000 samples for training and 10,000 for test, IID)
- Model: LR (**convex**) and CNN (**non-convex**, two 5×5 convolution layers**)**

## Parameter settings

- The number of clients N ranges from [10, 40]
- The budget $B$ ranges from [25, 70]
- The learning rate $\eta$ =0.001
- The number of time slots $T = 200$

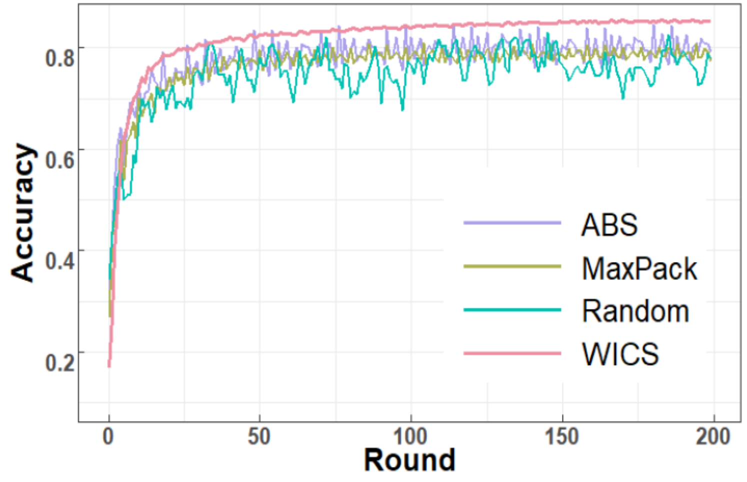
## Compared Algorithms

- WICS : our proposed algorithm
- Random
- MaxPack: based on AoI values
- ABS: based on the time of last selection

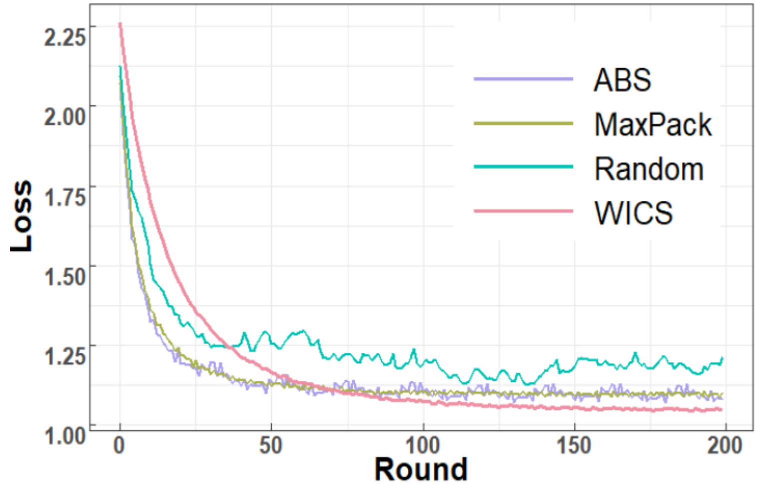
## Evaluation Metrics

- **Accuracy**: the number of correct predictions
- **Loss:** diff. between predicted and actual output
- **Average AoI** of all clients

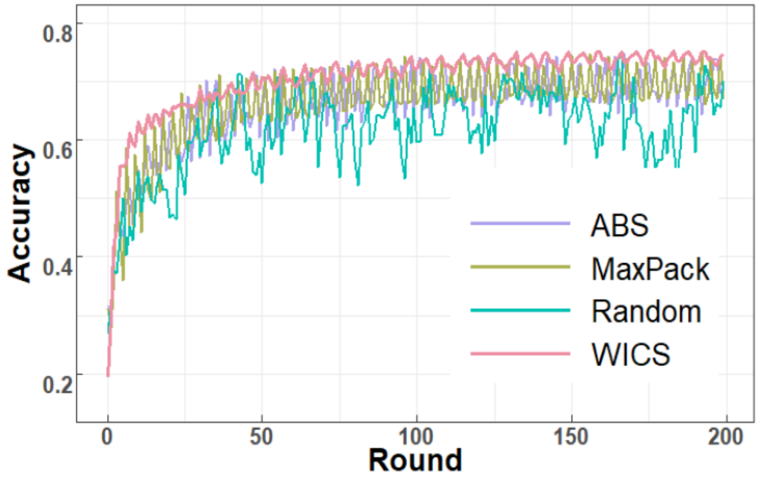
# Performance of LR on MNIST and FMNIST



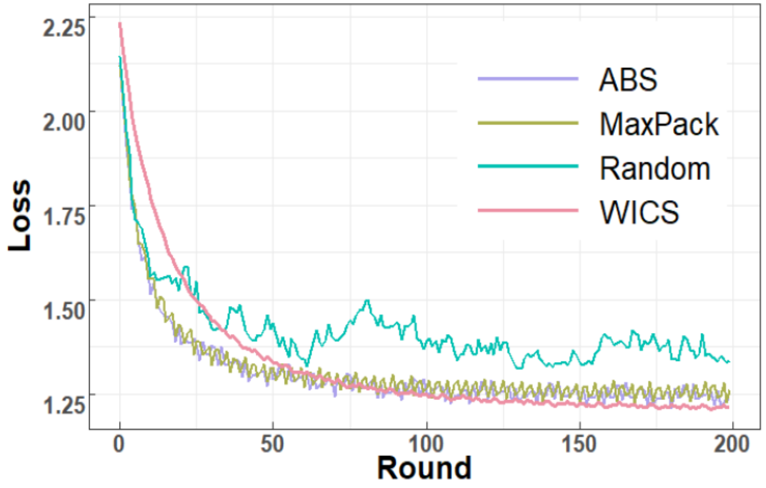

(a) Accuracy of LR on MNIST




(b) Loss of LR on MNIST



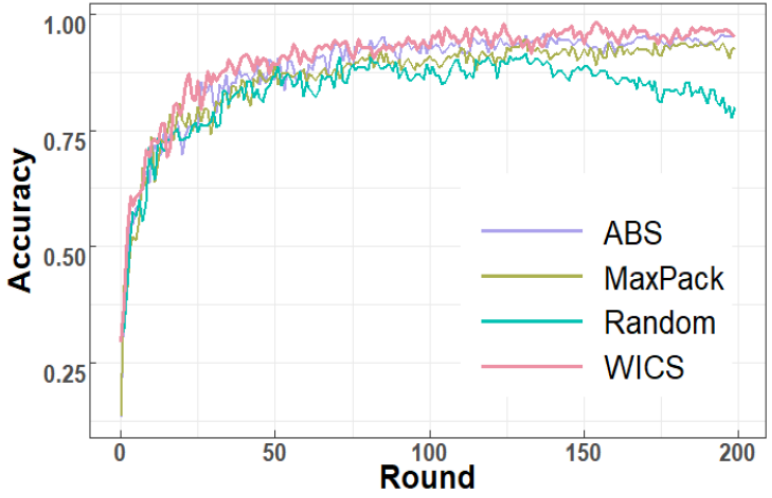

(a) Accuracy of LR on FMNIST



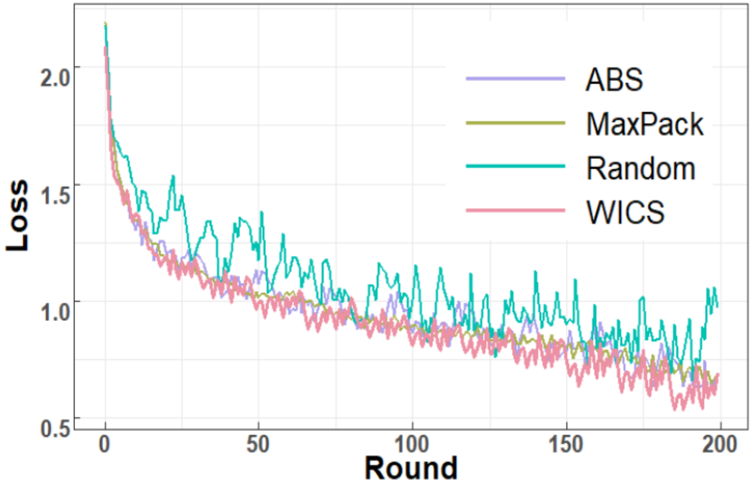

(b) Loss of LR on FMNIST

- The **accuracy** of four algorithms **rises** along with the increase of rounds;
- The **loss** of four algorithms **descends** with the increase of rounds;
- WICS is **better** (in terms of accuracy and loss) than the three compared algorithms.

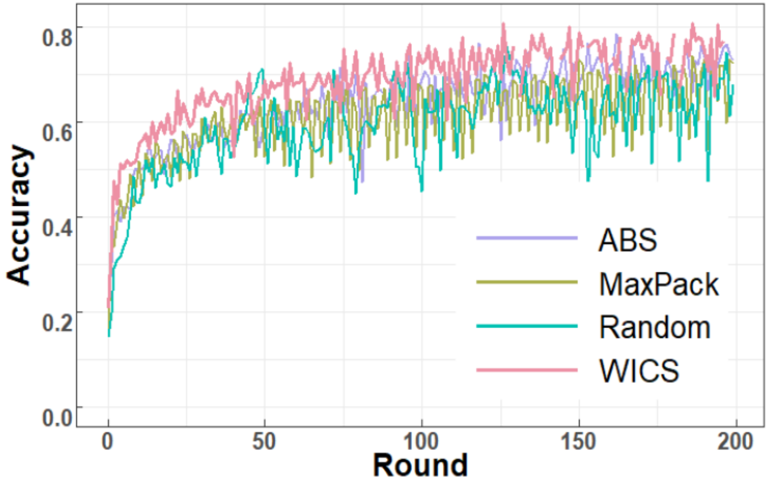
# Performance of CNN on MNIST and FMNIST



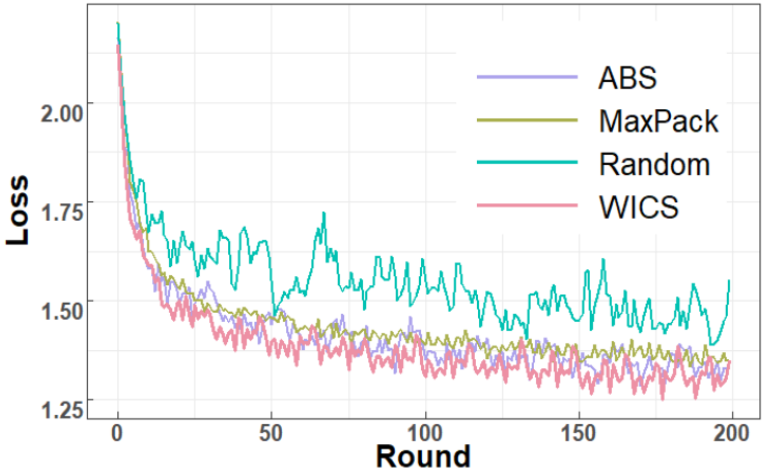

(a) Accuracy of CNN on MNIST




(b) Loss of CNN on MNIST



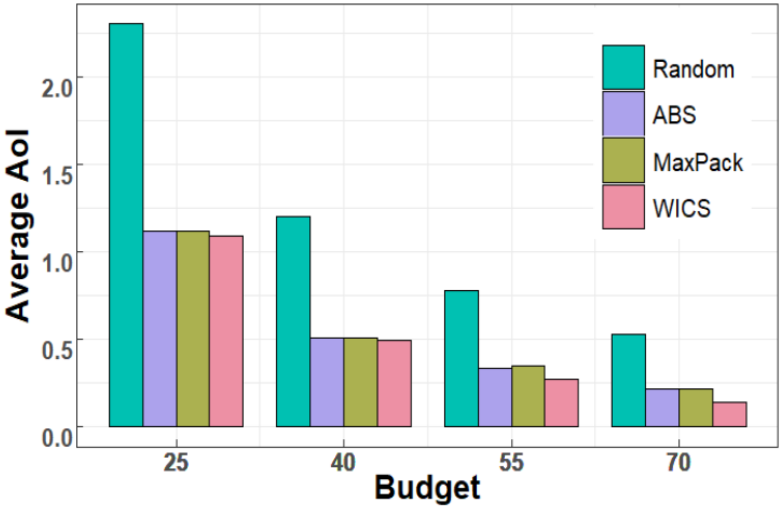

(a) Accuracy of CNN on FMNIST



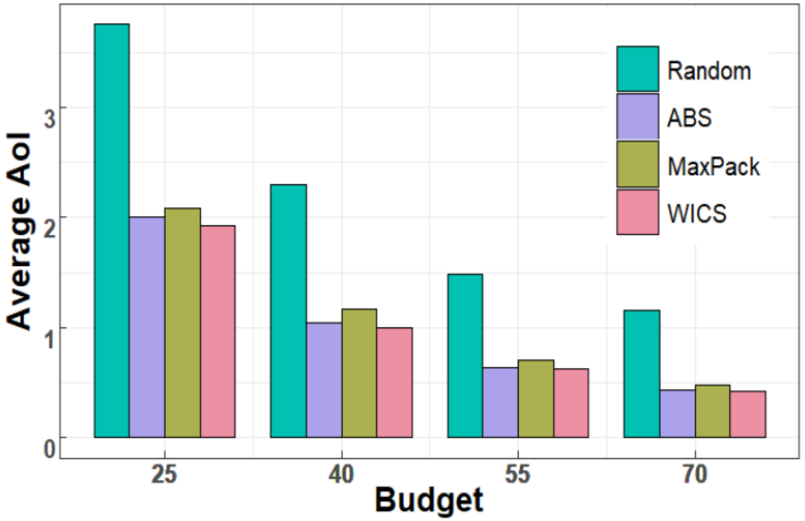

(b) Loss of CNN on FMNIST

When the loss function is **non-convex** (i.e., CNN), the performances of WICS are still **better**.

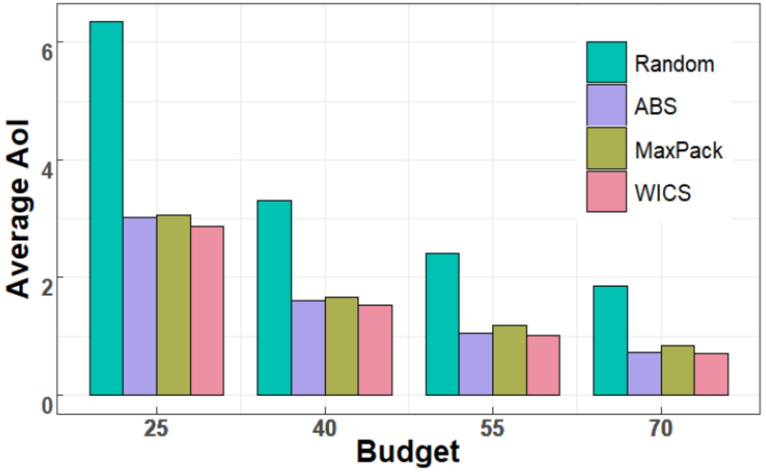
# Average AoI with Different Budget and Client Number



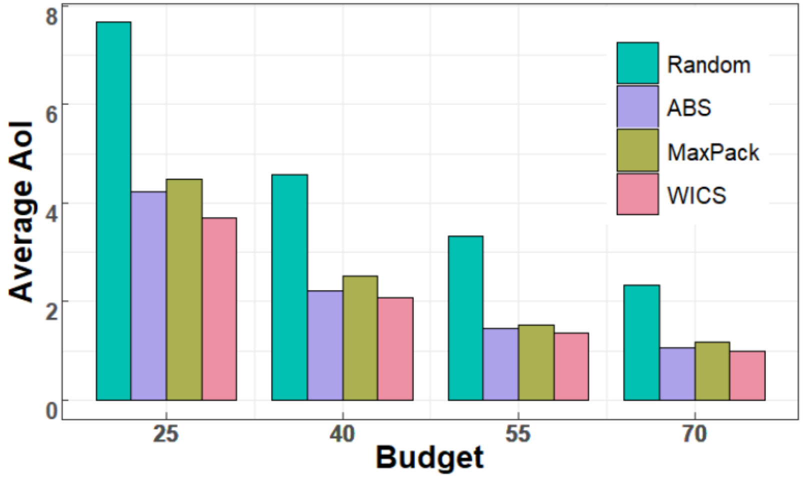

(a) $N$=10




(b) $N$=20




(a) $N$=30




(b) $N$=40

- WICS can achieve the **lowest** weighted average AoI;
- The weighted average AoI exhibits an **uptrend** with the increase of the number of clients $N$.

# Conclusions

- Introduce a **novel AoI-aware FL system**, where the server tries to select suitable clients to provide fresh datasets for local model training.

- Model the **client selection** problem as a **restless multi-armed bandit**, and propose the WICS algorithm by applying Whittle's Index.

- Prove the **approximate optimality** of WICS and evaluate the algorithm performance via simulations.

## Future work:

- Extend using **discount factor** based on time -- more weight on fresh information.
- Investigate on fine-grained integration of **fresh data** and **stale data.**

# Q & A