Università degli Studi di Milano - Bicocca
Dipartimento di Informatica, Sistemistica e Comunicazione
Corso di Laurea Magistrale in Informatica

Università della Svizzera italiana

Federated Learning Approaches for Human Mobility Models

Presented by:
Alex Caselli

Relatore USI: Prof. Marc Langheinrich
Relatore UNIMIB: Prof. Giovanni Denaro
Co-relatore USI: Prof.ssa Silvia Santini

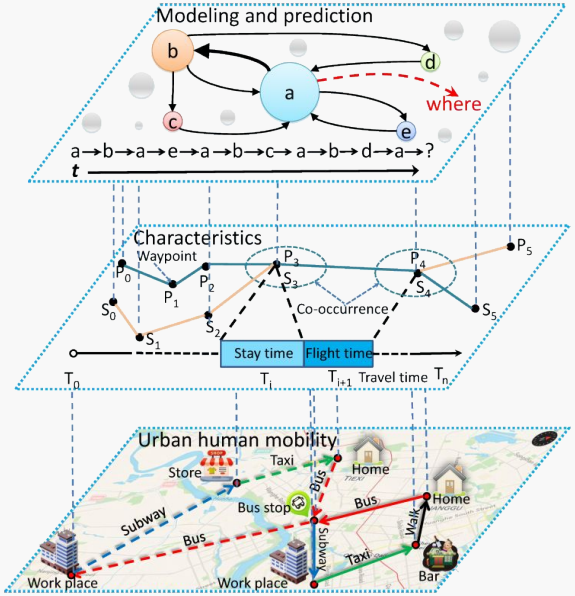
Where will you go next?

Introduction to human mobility modelling

Human Mobility Models

“A simplified representation of the movement of single or groups of mobile entities in a given context, primarily the spatial environment” - Hess et al.

Private data needs to be centralized



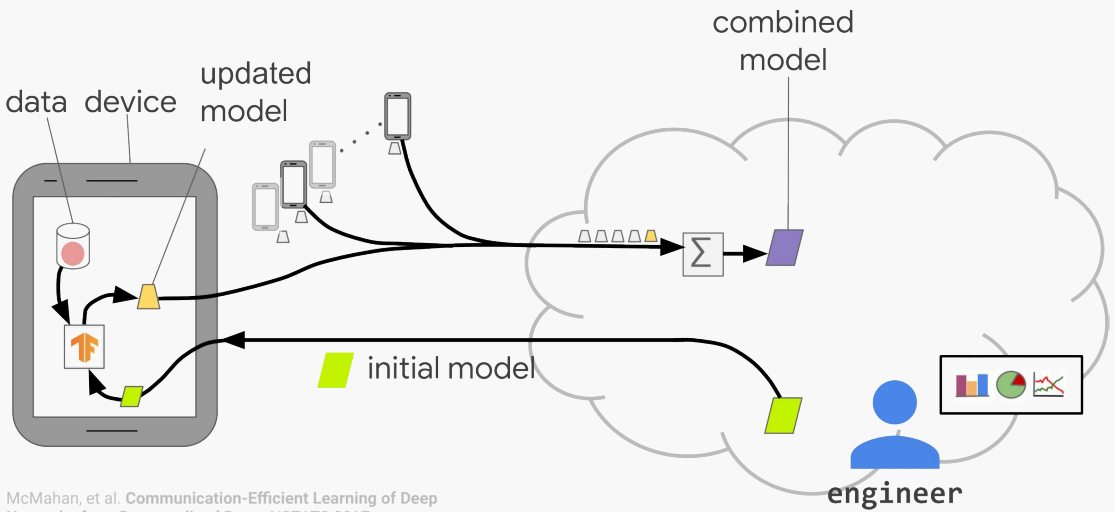
An overview of data-driven modeling and prediction for human mobility. [3]

Threatening the privacy of the users

[3] Urban Human Mobility: Data-Driven Modeling and Prediction, Wang et al. 2019

Federated Learning

Enable edge devices to do state of the art **machine learning without centralizing data** and with privacy by default.



McMahan, et al. **Communication-Efficient Learning of Deep Networks from Decentralized Data.** AISTATS 2017.

- **Privacy** by design
- **User's data** never leave the device
- Massively **decentralized**
- Model is **trained locally** on the devices
- **Hub-and-spoke** topology

How it works

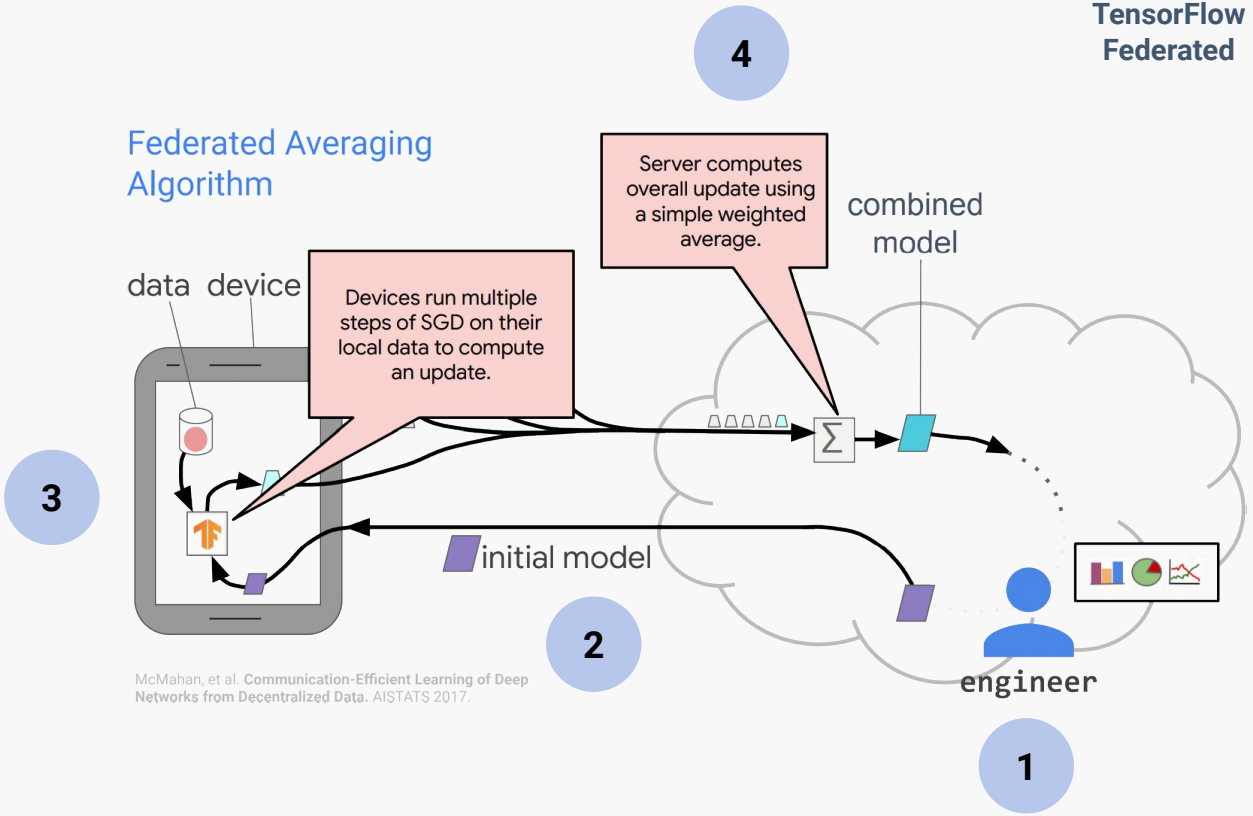
TensorFlow Federated

PHASE 1
Server selects a subset of clients eligible for the training round

PHASE 2
The global model is downloaded by the selected clients

PHASE 3
Clients compute several training steps with their local data

PHASE 4
Updated models are sent back to the server which updates the global model with the aggregated update



Key research questions

1. Efficient model's architecture?
2. Influence of demographic data?
3. Centralized and federated models performance?
4. Influence of sparse clients availability?
5. Impact of differential privacy on model's performance?
6. Advantages of pre-trained models?

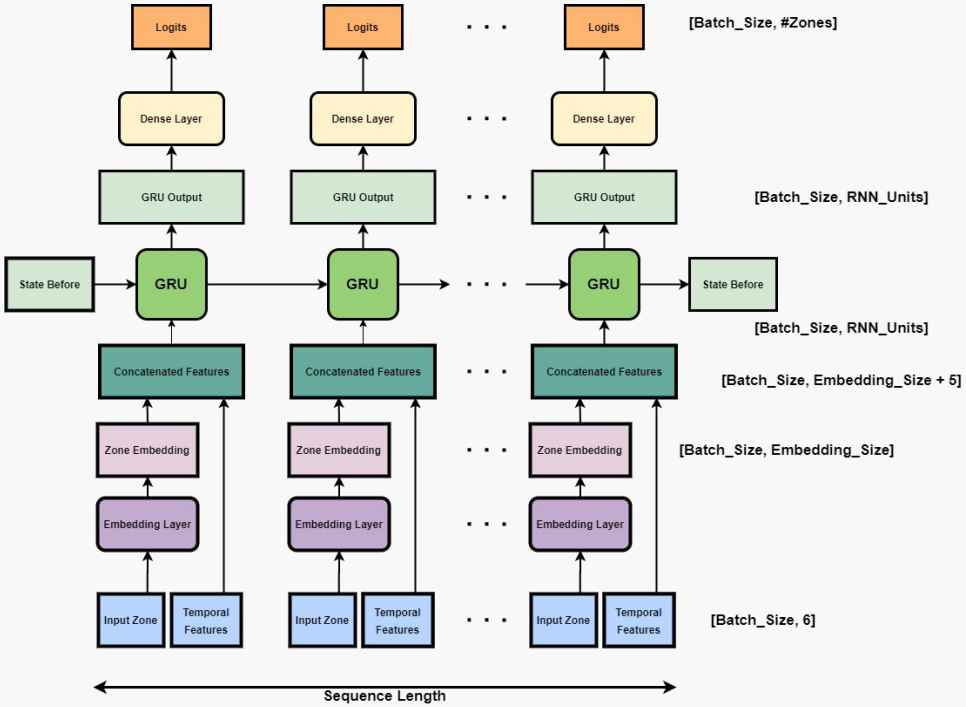
Datasets used

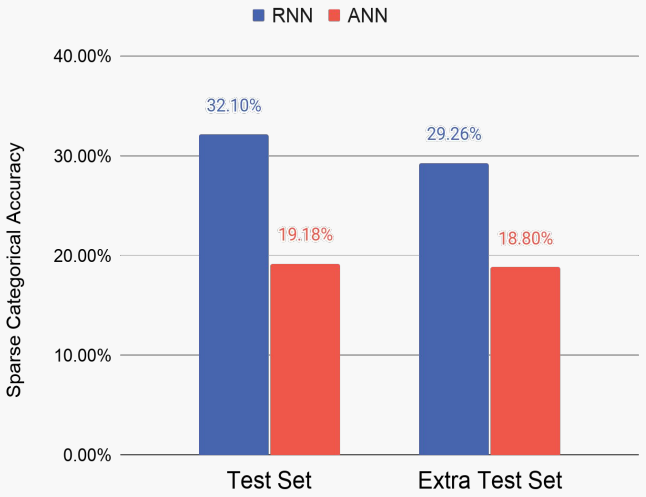
Our work makes use of three mobility datasets to perform the experiments. These datasets include: Persona, NYC taxi dataset and Nokia LDCC dataset. While PERSONA has been developed in the context of this thesis, the latter two publicly available datasets.

	City	Duration	Users	Locations	Records	Density
PERSONA	-	1 month	60	13	41K+	1 h
NYC Taxi Dataset	New York City	1 month	13,250	263	14M+	50 min
Nokia LDCC Dataset	Lausanne	18 months	185	423	89K+	16 h

Summary of the three used human mobility dataset after the pre-processing and data cleaning phases.

1. Model architecture





Comparison of the evaluation results of the two tested neural networks classes in terms of sparse categorical accuracy.

Key research questions

1. Efficient model's architecture?
2. Influence of demographic data?
3. Centralized and federated models performance?
4. Influence of sparse clients availability?
5. Impact of differential privacy on model's performance?
6. Advantages of pre-trained models?

2. Influence of demographic data

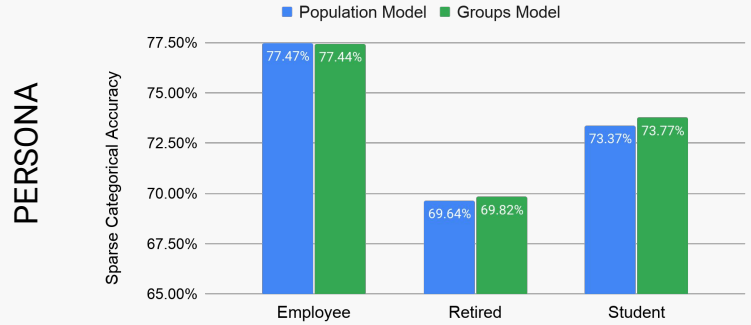
Our results confirm the observation of a recent work by P. Baumann et al. [5] in which the authors concluded that only a little advantage can be achieved by developing different models for different demographic groups.

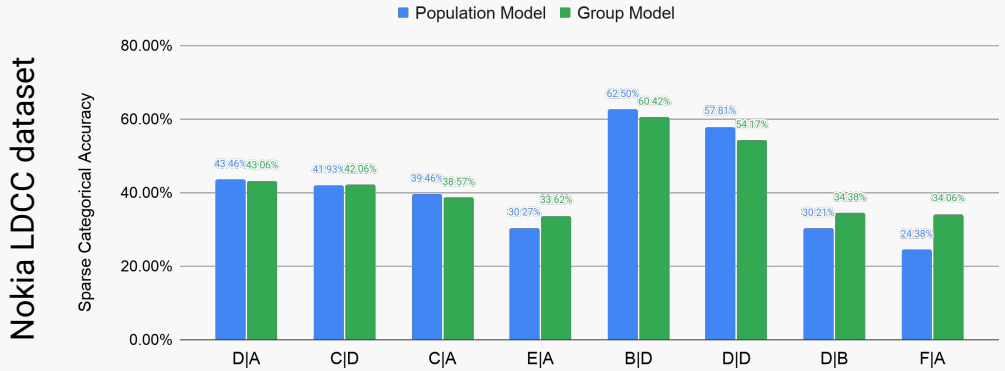
Average improvement provided by group models

PERSONA	**+0.24%**
Nokia LDCC	**+3.12%**

Number of demographic groups

PERSONA	**3**
Nokia LDCC	**8**



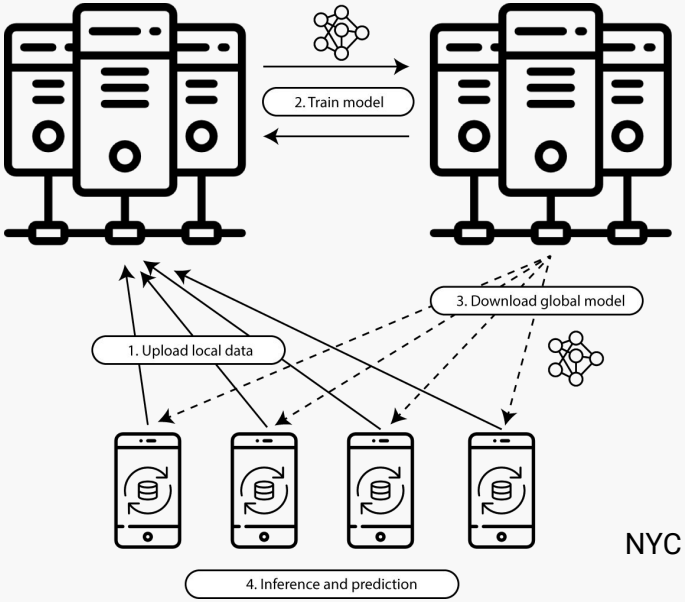


[5] Selecting Individual and Population Models for Predicting Human Mobility, P. Baumann et al. 2017

Key research questions

1. Efficient model's architecture?
2. Influence of demographic data?
3. Centralized and federated models performance?
4. Influence of sparse clients availability?
5. Impact of differential privacy on model's performance?
6. Advantages of pre-trained models?

3. Centralized vs federated models



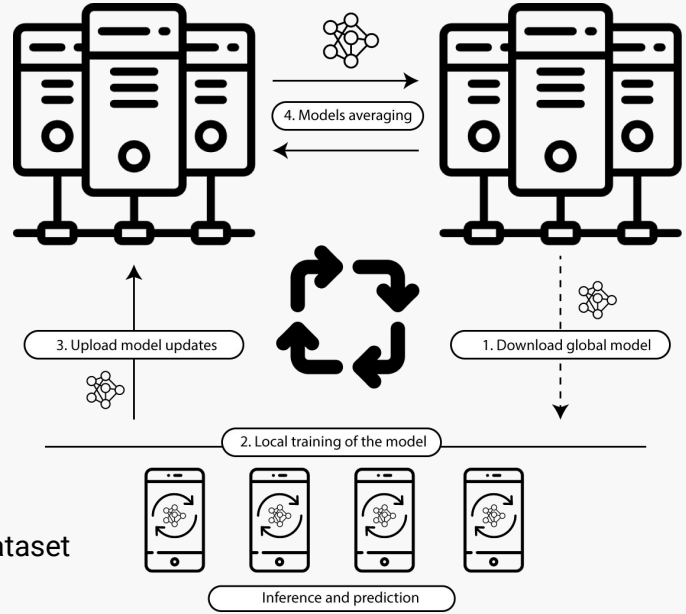
Workflow of the centralized model.

NYC Taxi dataset

100 Taxis

Nokia LDCC dataset

70 Users



Workflow of the federated model.

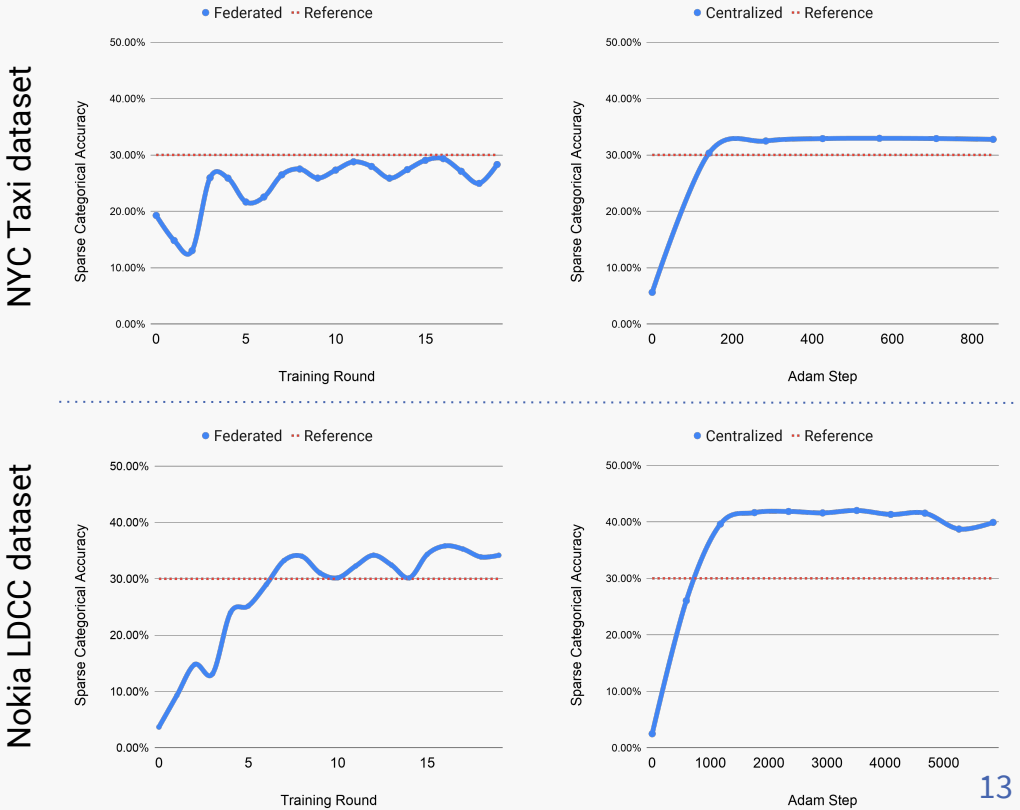
3. Centralized vs federated models

100 Taxis

70 Users

Dataset	Federated model	Centralized model
NYC Taxi	28.85%	**30.68%**
Nokia LDCC	36.95%	**42.09%**

Average sparse categorical accuracy achieved by the models on the two test sets used for evaluation.

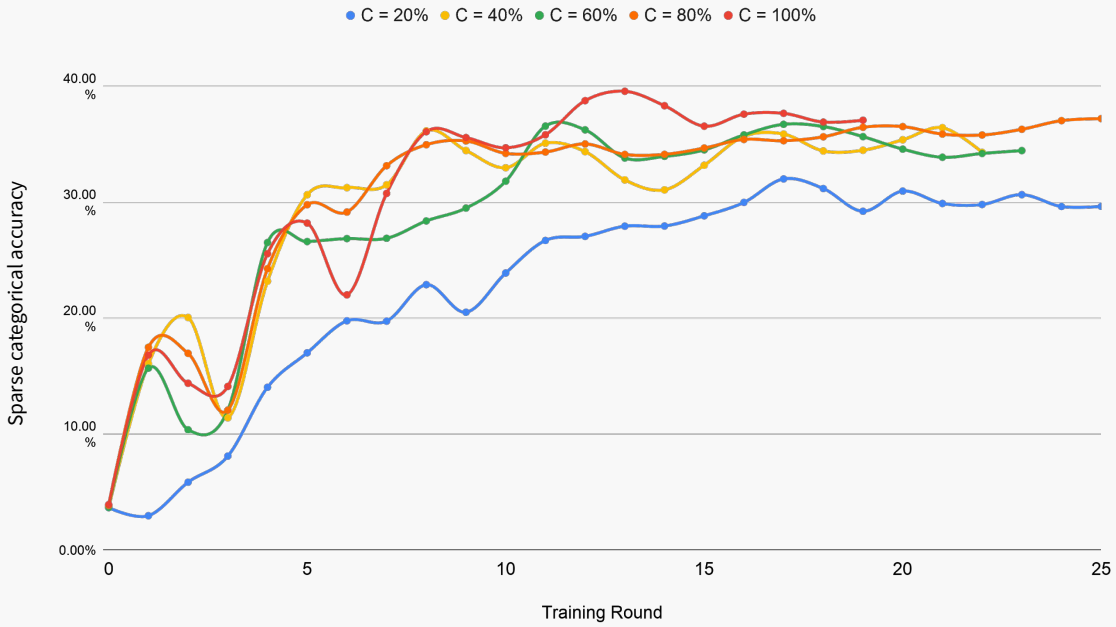


Key research questions

1. Efficient model's architecture?
2. Influence of demographic data?
3. Centralized and federated models performance?
4. Influence of sparse clients availability?
5. Impact of differential privacy on model's performance?
6. Advantages of pre-trained models?

4. Influence of clients availability

One of the main challenges FL has to face is the sparse availability of the clients, and so, the availability of the data. To simulate this characteristic, at each round a fraction of *C* clients are selected randomly to participate.



Sparse categorical accuracy on the validation set as function of training round during federated training.

Rounds to reach 30% of accuracy:

C = 20%: 17
C = 40%: 5
C = 60%: 7
C = 80%: 5
C = 100%: 7

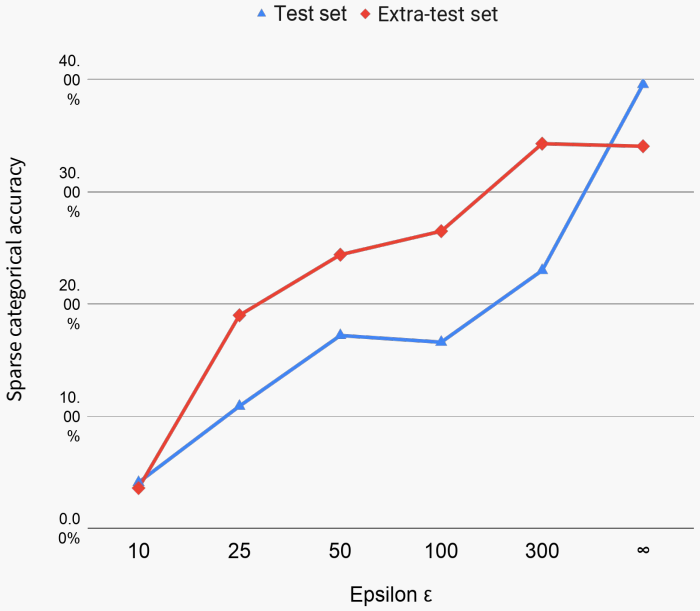
C	Number of users
20%	14
40%	28
60%	42
80%	56
100%	70

Key research questions

1. Efficient model's architecture?
2. Influence of demographic data?
3. Centralized and federated models performance?
4. Influence of sparse clients availability?
5. Impact of differential privacy on model's performance?
6. Advantages of pre-trained models?

5. Impact of differential privacy

One of the most common privacy-preserving mechanisms for sharing information of groups while withholding information about individuals.



Sparse categorical accuracy on both the test sets as function of epsilon.

Differential privacy acts by:

- Clipping the model updates' delta
- Multiplying the clipped updates by random noise

Delta	Clipping Threshold	Number of Clients	Noise Multiplier	Rounds	Epsilon	Accuracy
-	-	70	-	14	∞	36.83%
1e-2	1.0	70	0.3	40	300	28.65%
1e-2	1.0	60	0.5	35	100	21.55%
1e-2	1.0	70	1.0	60	50	20.8%
1e-2	1.0	60	1.3	40	25	14.95%
1e-2	1.0	70	1.7	25	10	3.85%

Differential privacy parameters used for the six models. Accuracy refers to the average sparse categorical accuracy achieved by each model on the test set and the extra-test set.

Key research questions

1. Efficient model's architecture?
2. Influence of demographic data?
3. Centralized and federated models performance?
4. Influence of sparse clients availability?
5. Impact of differential privacy on model's performance?
6. Advantages of pre-trained models?

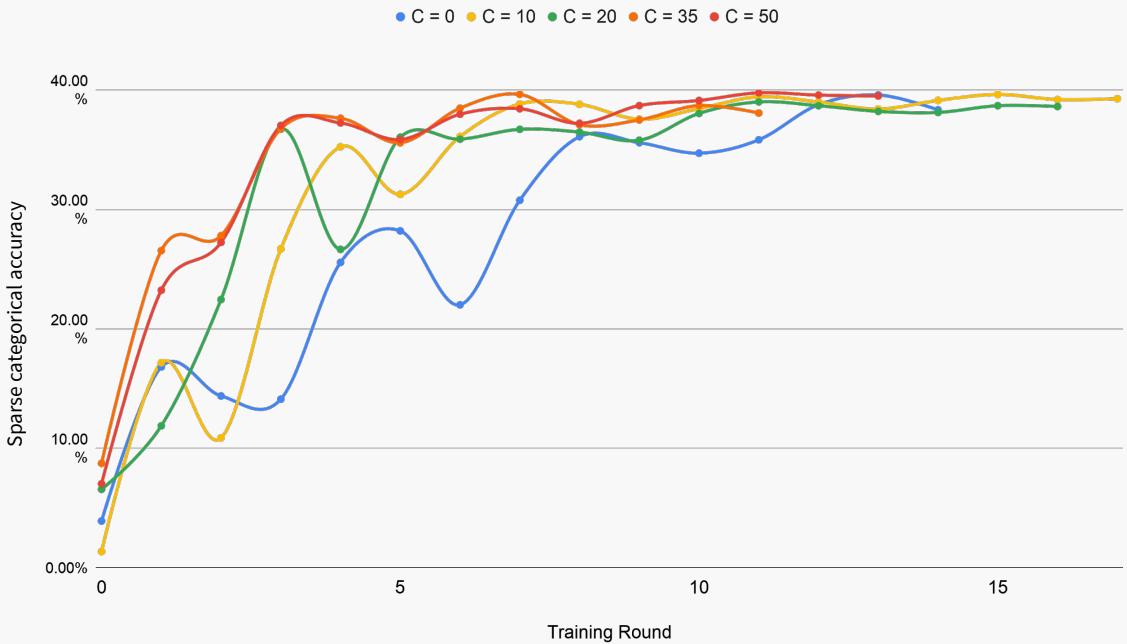
6. Advantages of pre-trained models

Our results show that having pre-trained models can speed up the federated learning process by both reducing the number of training rounds required to converge (also reducing the communication cost) and providing better overall performance.

Rounds to reach 37% of accuracy:

No pre-train:	12
Pre train (C=10):	7
Pre train (C=20):	3
Pre train (C=35):	3
Pre train (C=50):	3

4x Decrease in communication rounds



Sparse categorical accuracy on the validation set as function of training round during federated training.

Conclusions and limitations

Human mobility models via federated learning:

- Encouraging performance
- Privacy by design
- Useful to adapt existing models

Limitations:

- Small loss in performance when compared to centralized methods
- High computational cost in a simulated environment
- Frameworks are in an early stage
- Differential privacy casues degradation of performance

Future work

Federated learning in practice

- Android app collects mobility data
- Central server orchestrate the federated processes
- TensorFlow Lite on the devices to train the federated model

Attacks on federated learning

- Back door federated learning
- Simulate targeted attacks
- Malicious clients with falsified data

Analysis of recently published related works

- Feng et al. (2020): PMF: A Privacy-preserving Human Mobility Prediction Framework via Federated Learning.
  PACM IMWUT, Vol. 4, Issue 1, Art. 10, March 2020.

THANKS for your attention!

Università della Svizzera italiana

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

Main Contributions

- **Development of PERSONA**, a flexible generator of synthetic human mobility dataset.
- **Study** and testing **of the implications of** building different mobility models for different **demographic groups.**
- **Comparison** and investigation **of different neural network architectures** applied to the next-place prediction task.
- **Development of a federated mobility model** able to predict the next place visited by a user.
- **Study of the effects of additional privacy-preserving measures** such as differential privacy.
- **Study of the characteristics and limitations of federated learning** in different scenarios.

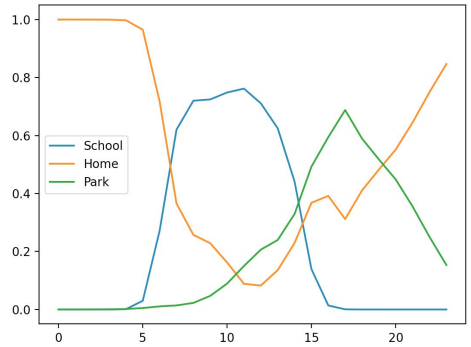
Backup Slides

PERSONA dataset

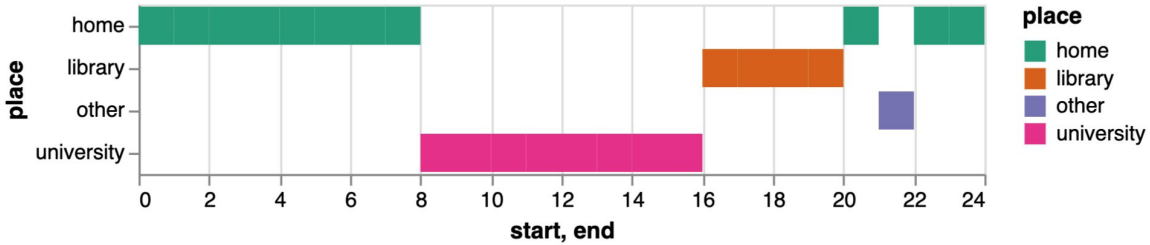
A synthetic human mobility dataset with demographic groups.

		Home	School	Park	University	Library	Station	Work	Parents' House	Grocery Store	Gym	Restaurant	Church
Student	Type A	X	X	X	-	-	-	-	-	-	-	-	-
	Type B	X	-	X	X	X	-	-	-	-	-	-	-
	Type C	X	-	-	X	-	X	-	-	-	-	-	-
Employee	Type A	X	-	-	-	-	-	X	X	X	-	-	-
	Type B	X	-	-	-	-	X	X	-	-	X	-	-
	Type C	X	-	-	-	-	-	-	X	-	X	-	-
Retired	Type A	X	-	X	-	-	-	-	-	X	-	X	-
	Type B	X	-	-	-	-	-	-	-	X	-	X	X
	Type C	X	-	X	-	-	-	-	-	-	X	X	-

Main classes, sub-classes and their relevant places.



Normalized place probability distributions.



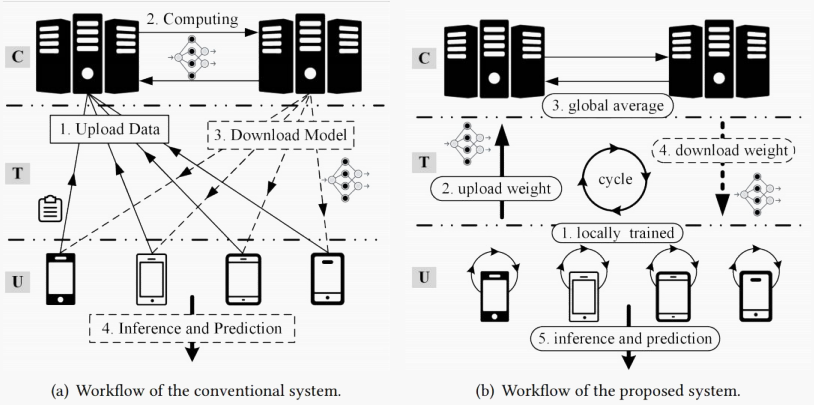
Simulation of a synthetic slotted mobility trace produced by a student.

Name	Description	Values range
users_id	Unique user's id	[0, N-1]
users_class	Main class type of the user	[0, 2]
start_hour	Hour of the day when the user is at start_place	[0, 23]
end_hour	Hour of the day when the user is at end_place	[0, 23]
start_place	Initial relevant place of the user	[0, 12]
end_place	Final relevant place of the user	[0, 12]
day_type	Whether is a weekday or a weekend	[0, 1]

Dataset's schema.

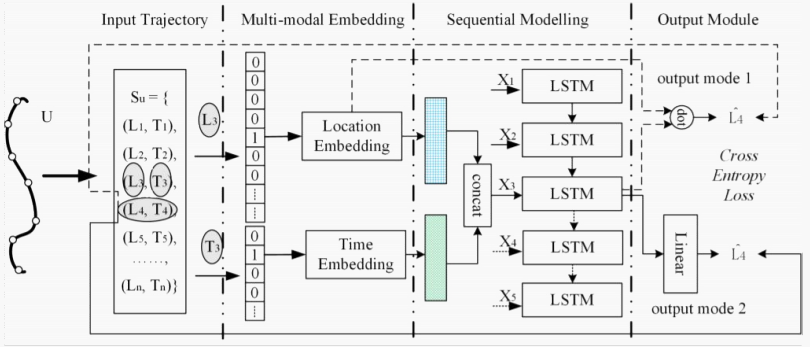
Related work: PMF

To our knowledge, the only work related to federated learning applied to the human mobility field in literature.



(a) Workflow of the conventional system. (b) Workflow of the proposed system.

Workflow comparison between the conventional system and the system proposed in [3]. There are three environments in the figure: “C” denotes the cloud servers, “T” denotes transfer environment, and “U” denotes mobile devices



The framework of the basic mobility prediction model, which includes three parts: multi-modal embedding, sequential modelling, and output module.

PMF: A Privacy-preserving Human Mobility Prediction Framework via Federated Learning, J. Feng et al. 2020

Data is born at the edge

Billions of mobile devices, such as smartphones, tablet and wearables, generate data continuosly.