Experiments on Federated Data Synthesis

CLAIRE LITTLE, MARK ELLIOT, RICHARD ALLMENDINGER

UNIVERSITY OF MANCHESTER

MANCHESTER
1824
The University of Manchester

Questions?

https://tinyurl.com/QuestionsUoM

Federated Learning (FL)

FL (McMahan et al., 2017) is a decentralized approach to training statistical models

- Multiple clients can produce one global model
- Clients do not share or exchange their own data
- Can reduce privacy and security risks (compared to methods that combine multiple data sources)
- Allows models to train on data that is more representative of the whole distribution
- Useful where clients do not possess enough data to generate the required statistical power

Federated Learning (FL)

Central server controls the process (but does not access any client data)

- Initialises model, sends to each client
  - Typically, neural network type models are used

Each client trains the model on their own data

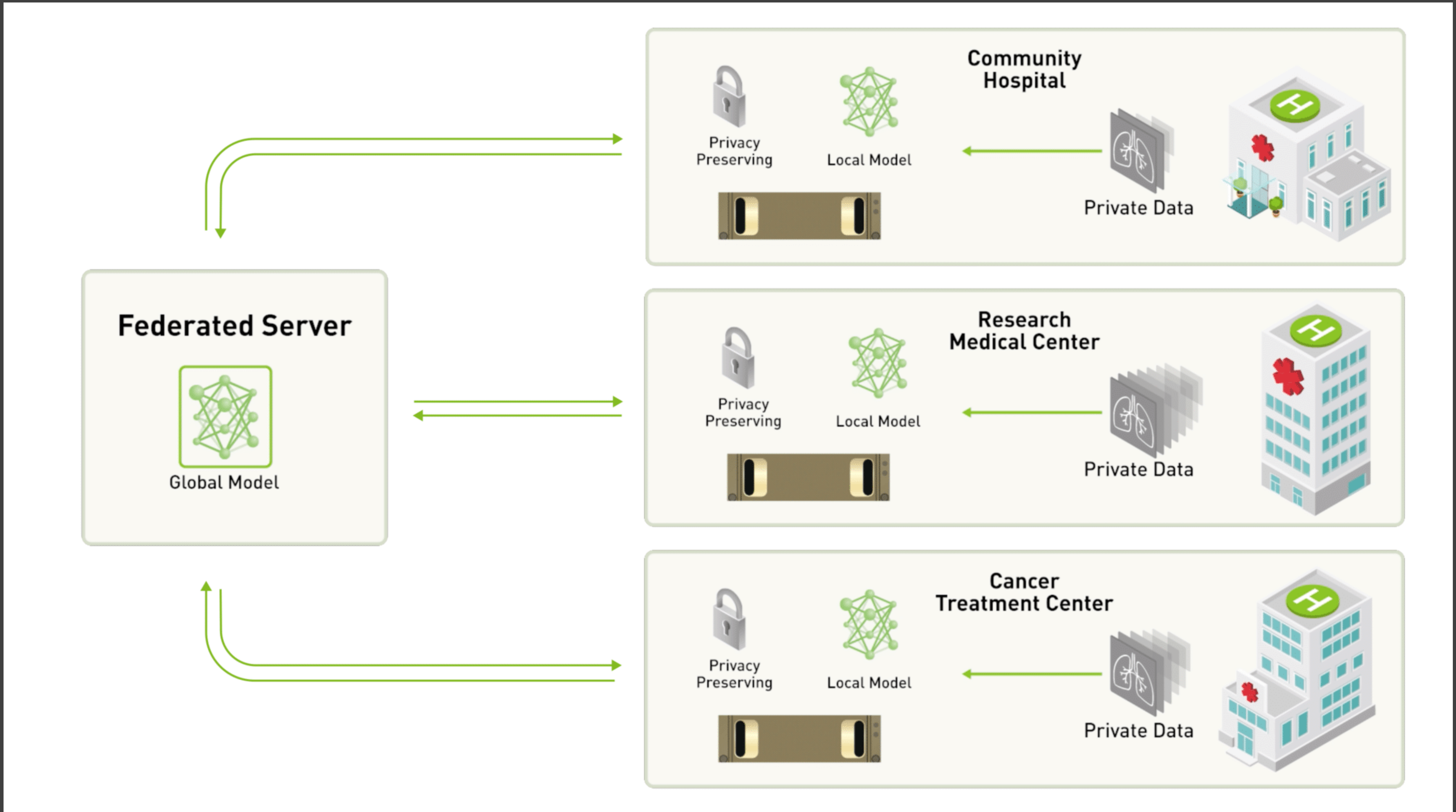
- Send updates (parameters or model weights) back to server

Server aggregates the client updates

- Sends updated model back to clients

Iterative process

- Training usually terminated when specific criterion is met:
  - E.g., maximum number of iterations



NVIDIA - A centralized-server approach to federated learning. https://blogs.nvidia.com/blog/2019/10/13/what-is-federated-learning/

Federated Synthesis

Using FL to generate synthetic data

- Emerging research field
- Small body of research focussing mostly on image data
- Less research on tabular data
- Methods predominantly use GANs (Generative Adversarial Networks, Goodfellow et al. 2014))

Is it possible to produce useful synthetic microdata in a federated way?

- Proof of concept using Genetic Algorithm (GA)

Genetic Algorithms (GAs)

GAs (Holland, 1992) perform iterative optimisation, training over multiple generations

- Three main biologically inspired operators:
  - Selection, Crossover, Mutation

- Initial population of candidate solutions (candidate solution = synthetic dataset)
- Fitness (similarity to original data) of each candidate calculated
- Select fitter candidates (parents) to reproduce for new population
- Crossover – combines parents to produce new candidates (children)
- Mutation – randomly change some of the candidates features
- Next generation – children, or combination of best (fittest) parents and children (elitism)
- Repeat process multiple times (generations) using fitness to guide

Study Design - Data

A (very) simple binary dataset, randomly sampled from UK 1991 Census microdata (University of Manchester, 2023)

- Small dataset to enable understanding
- 10 rows, 5 binary variables
  - “Original” dataset
- Randomly split into two five-row datasets
  - representing two clients (A and B)

AGE	MSTATUS	SEX	LTILL	TENURE	client
1	2	2	2	2	A
1	1	1	2	2	A
1	1	1	2	2	A
2	2	2	2	1	A
1	1	1	2	1	A
2	2	2	2	1	B
1	2	2	2	1	B
1	1	1	2	1	B
1	1	1	1	2	B
1	1	1	2	1	B

Study Design - Parameters

Huge potential range of variation in the simulation

Three types of parameters:

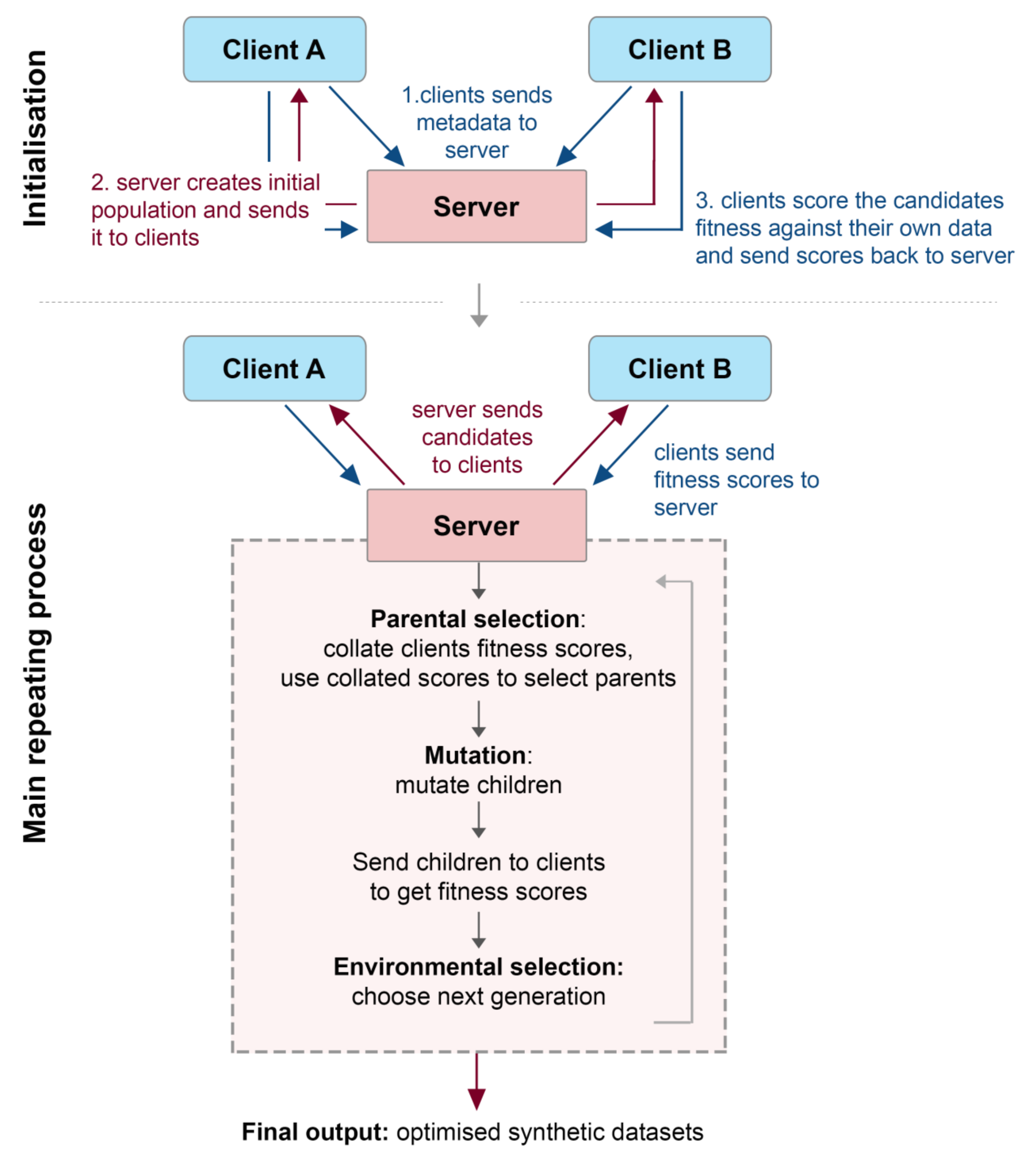
- Model: changeable settings for the GA (e.g., mutation rate)
- Simulation: variations in the scenario being presented (e.g., number of clients)
- Experimental: elements that are not part of the simulation itself (e.g., data choice, number of runs)

Model complexity is kept low to aid with interpreting the results

- Focus only on utility (not risk)
- Small dataset
- GA uses mutation but not crossover
- Two clients for FL

Study Design - Parameters

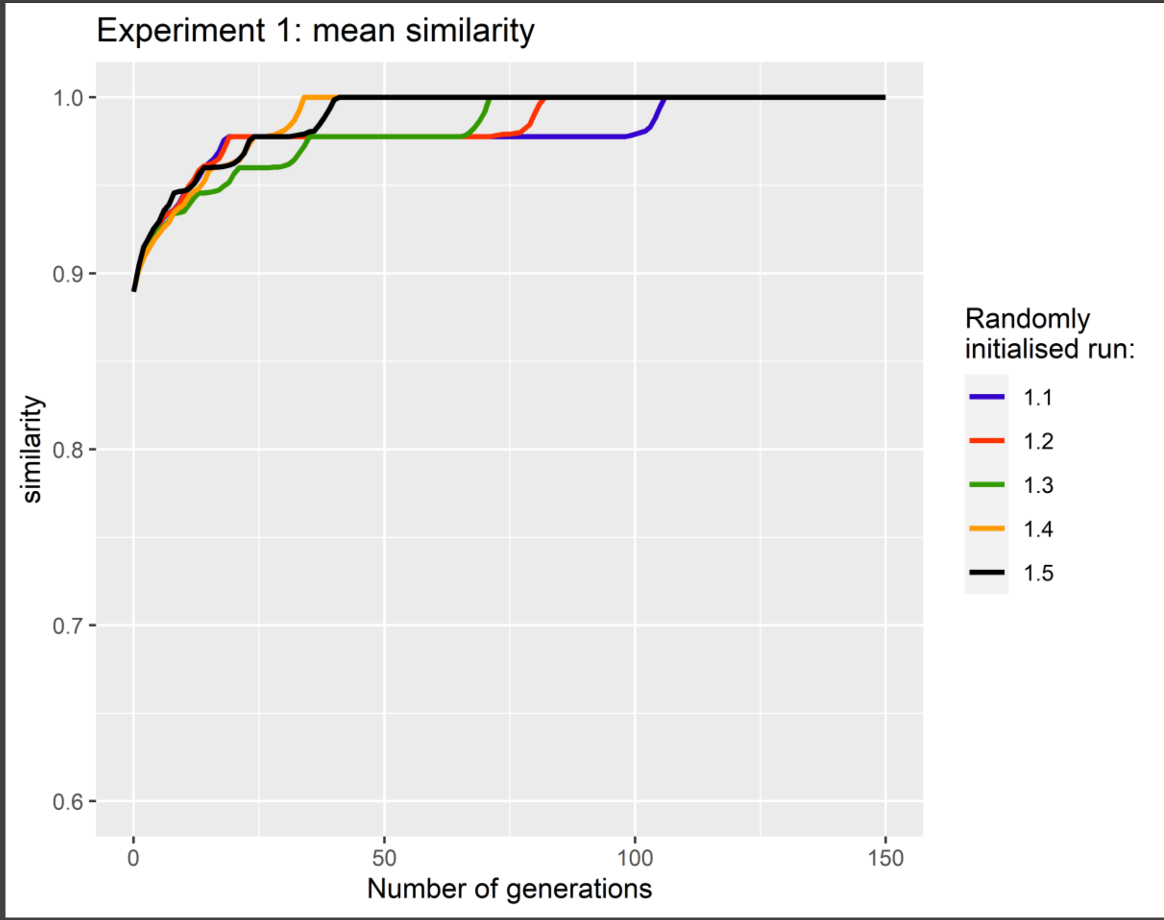
Parameter	Type	Value chosen	Further details
No. of clients	Simulation	VARIES	-
Initial Metadata sent by clients	Simulation	Univariates	-
Combination of client scores	Simulation	VARIES	-
No. of objectives for GA	Simulation	1	Similarity (utility)
SDC applied to the output sent to server	Simulation	None	-
Output passed to client by server	Simulation	VARIES	-
Population size	Model	50	-
Parental selection	Model	Binary tournament	k=2
Mutation rate	Model	0.05	-
Crossover Operator	Model	None	-
Environmental selection	Model	Elitism	-
No. of generations	Experiment	150	-
Choice of Dataset	Experiment	UK Census microdata	1991
No. of rows (per client)	Experiment	VARIES	-
No. of variables	Experiment	5	-
Type of variables	Experiment	Binary	-
No. of runs	Experiment	5	-


Initialisation
Client A
Client B
1.clients sends metadata to server
2. server creates initial population and sends it to clients
Server
3. clients score the candidates fitness against their own data and send scores back to server
Main repeating process
Client A
Client B
server sends candidates to clients
clients send fitness scores to server
Server
Parental selection: collate clients fitness scores, use collated scores to select parents
Mutation: mutate children
Send children to clients to get fitness scores
Environmental selection: choose next generation
Final output: optimised synthetic datasets

Results – Experiment 1

Running GA on original dataset (10 rows)

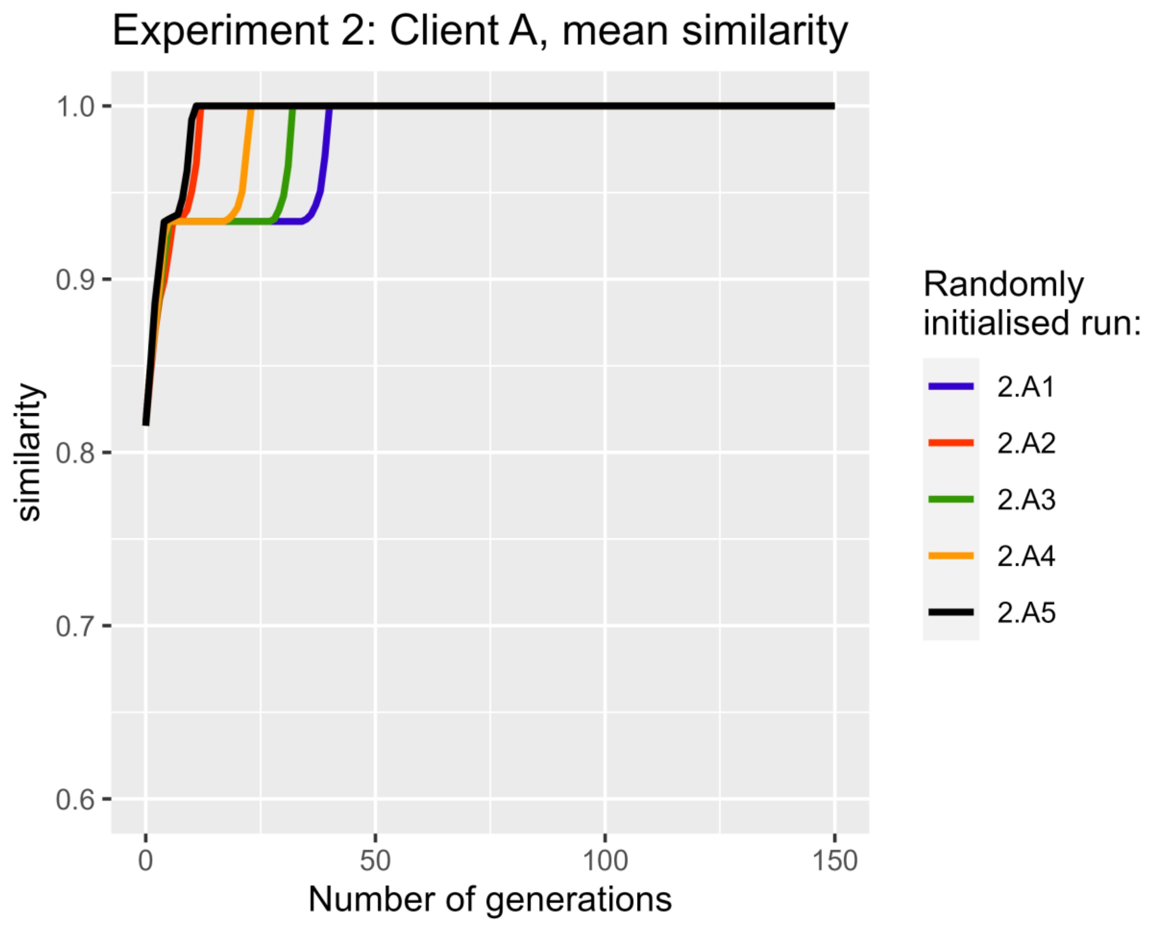
- All five randomly initialised runs converged
  - i.e., they reproduced the original dataset

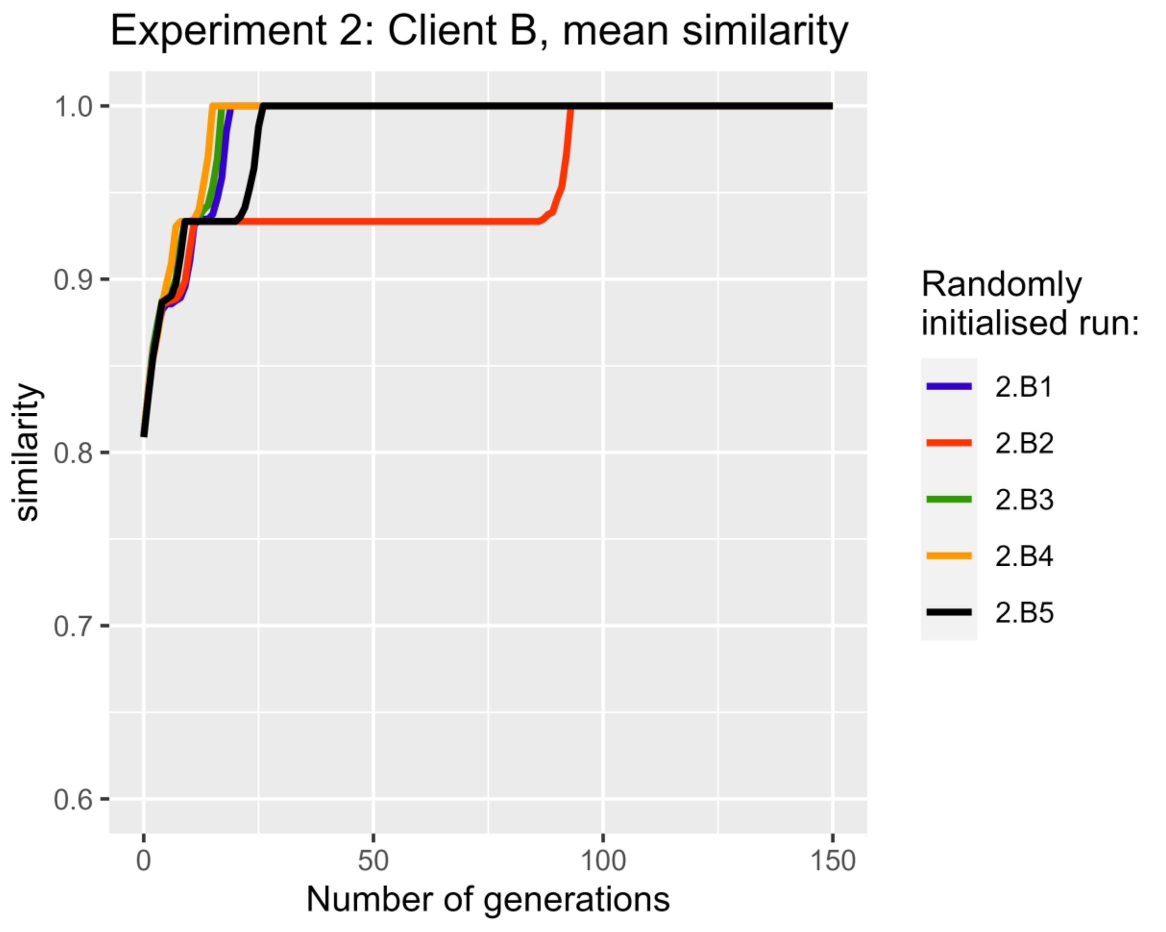


Results – Experiment 2

Running GA separately on client A and B datasets (5 rows each)

- For each, all five randomly initialised runs converged and reproduced the original dataset

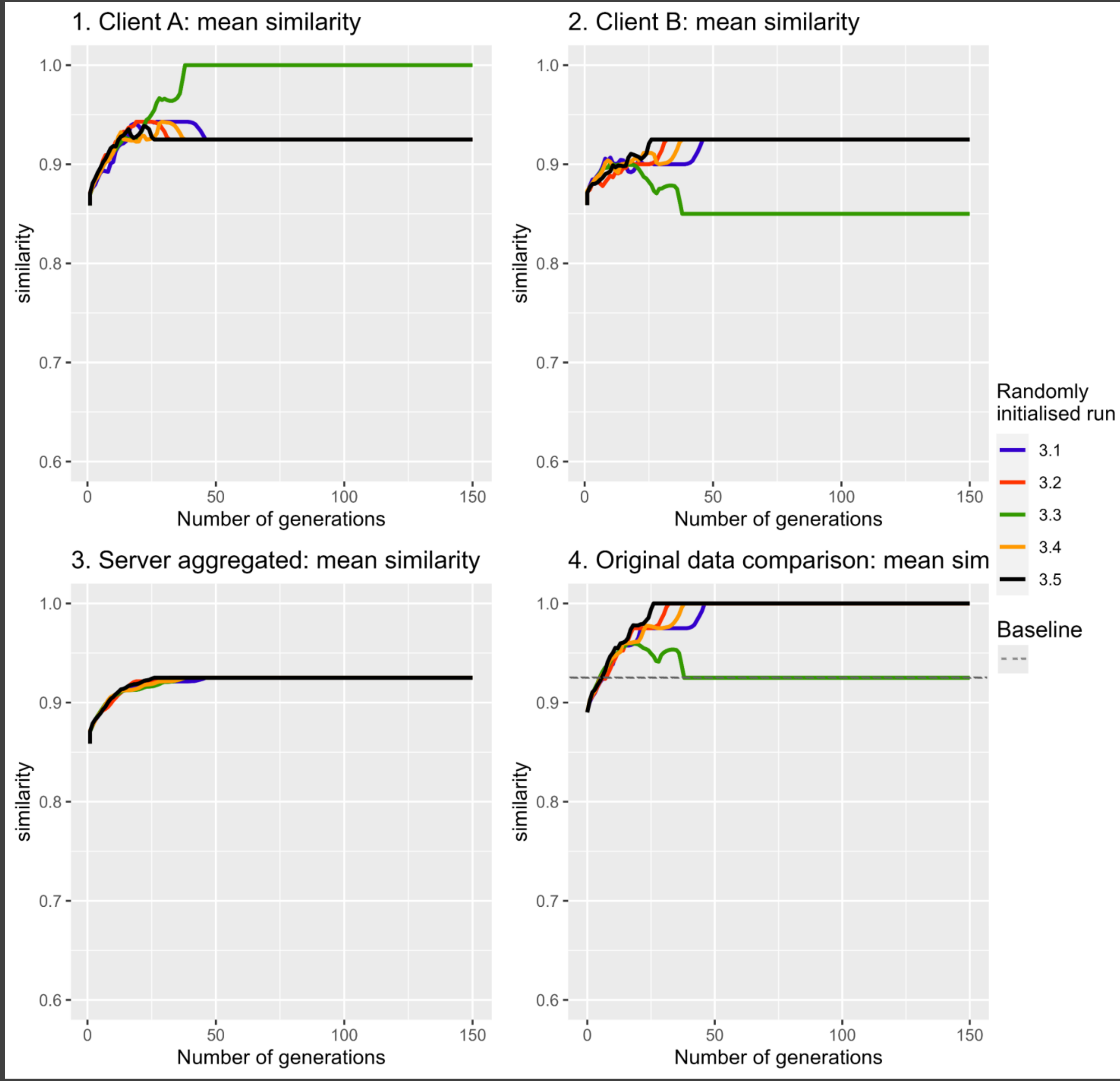




Results Experiment 3

FL with two clients (A and B)

- All but one of the randomly initialised runs converged and reproduced the original datasets
- Panel 4 would not be available in reality – used for evaluation
- Convergence achieved despite the evaluations from clients, and the server aggregated score indicating suboptimality



Discussion

Experiment 3 demonstrates proof of concept

- Analytically useful datasets were synthesised across distributed datasets

It was not clear on the server that the original data had been reproduced

- Might be useful in terms of disclosure risk
- Means we cannot rely on server-side restraint to minimise risk

Caveats and future work

Experiments conducted on small sample of binary Census microdata

- May not scale to larger, more complex data
- Very large datasets may be computationally impractical

Would need to consider different parameters

- More than 2 clients

Single-objective focus on utility

- In a real-life scenario, the goal would not be to reproduce the original data
- Risk would need to be factored in
  - A multi-objective approach within the GA could be used
  - Deep learning methods also a possibility

Questions?

https://tinyurl.com/QuestionsUoM

Email: claire.little@manchester.ac.uk

References

McMahan, B., E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas (2017). Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. PMLR. http://proceedings.mlr.press/v54/mcmahan17a/mcmahan17a.pdf

Goodfellow, I., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio (2014). Generative Adversarial Nets. In *Proceedings of the Advances in Neural Information Processing Systems*, Volume 27. https://papers.nips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf

Holland, J. H. (1992). *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*. MIT press.

University of Manchester, Cathie Marsh Centre for Census and Survey Research, Office for National Statistics, Census Division. (2023). *Census 1991: Individual Sample of Anonymised Records for Great Britain (SARs)*. [data collection]. UK Data Service. SN: 7210, DOI: http://doi.org/10.5255/UKDA-SN-7210-1