Privacy Enhancing Technology Project Parties

NIH NATIONAL CANCER INSTITUTE

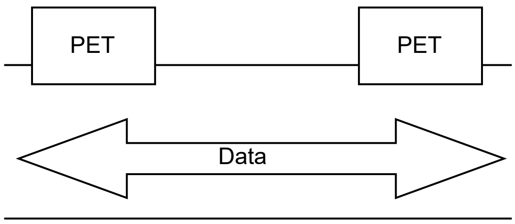
U.S. National Science Foundation

NAACCR North American Association of Central Cancer Registries

OAK RIDGE National Laboratory

NIST National Institute of Standards and Technology U.S. Department of Commerce

eRA



NHS England

NDRS NATIONAL DISEASE REGISTRATION SERVICE

NPIC

Office for National Statistics

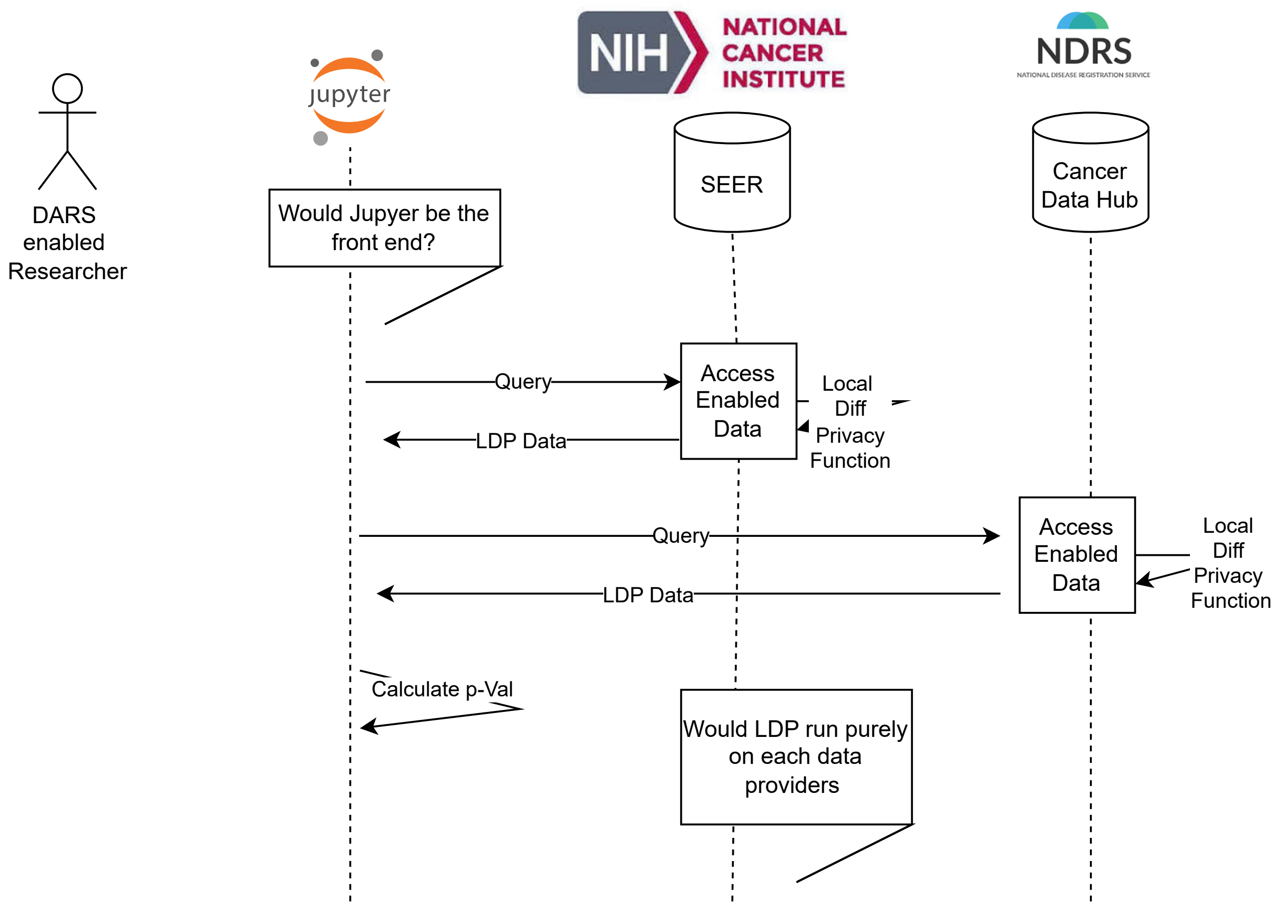
Genomics England

HDRUK Health Data Research UK

Differential Privacy Use Case Flow

https://en.wikipedia.org/wiki/Local_differential_privacy

A permitted Researcher with access to NIH SEER database and a Jupyter Notebook wishes to calculate the p-value between a US rare cancer group and a UK rare cancer group.

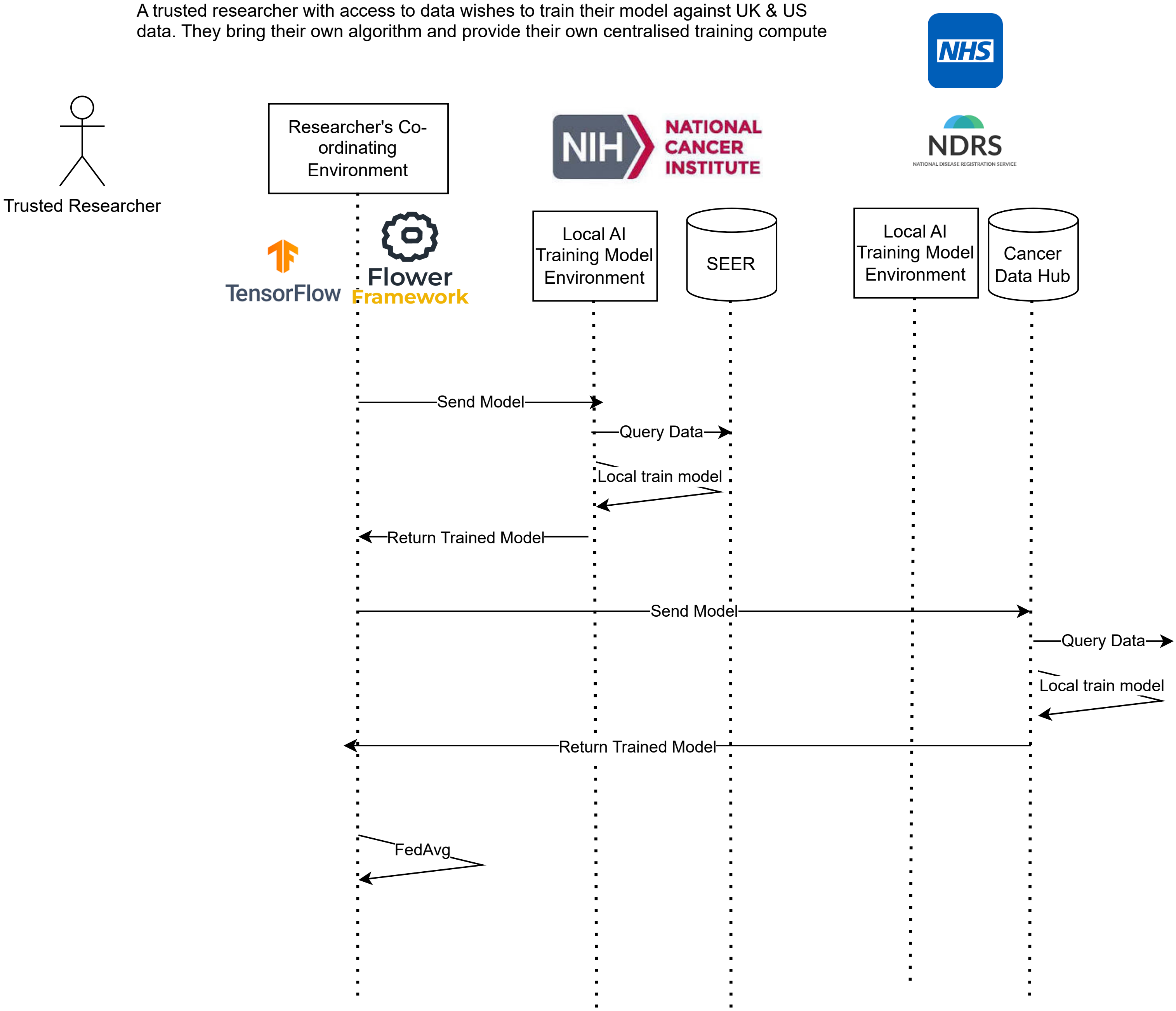


Assumptions:

1. Users will have passed Data Access Request permission flows (on a per user basis) with each data provider
2. Users will conform a 5 Safes style framework
3. ε-differential privacy levels will be set by each data provider
4. SEER will be the data provider for the US in the project
5. NDRS / NHS England will be the data provider for the UK in the project
6. Each data provider will implement noise on their data
7. The researcher will use a Python notebook style tool for the front end which will make remote queries via API?
8. There will be a controlled bastion server on each data provider that the client accesses?
9.

Federated Learning Use Case Flow
(Bring your own model + orchestrator)

A trusted researcher with access to data wishes to train their model against UK & US data. They bring their own algorithm and provide their own centralised training compute

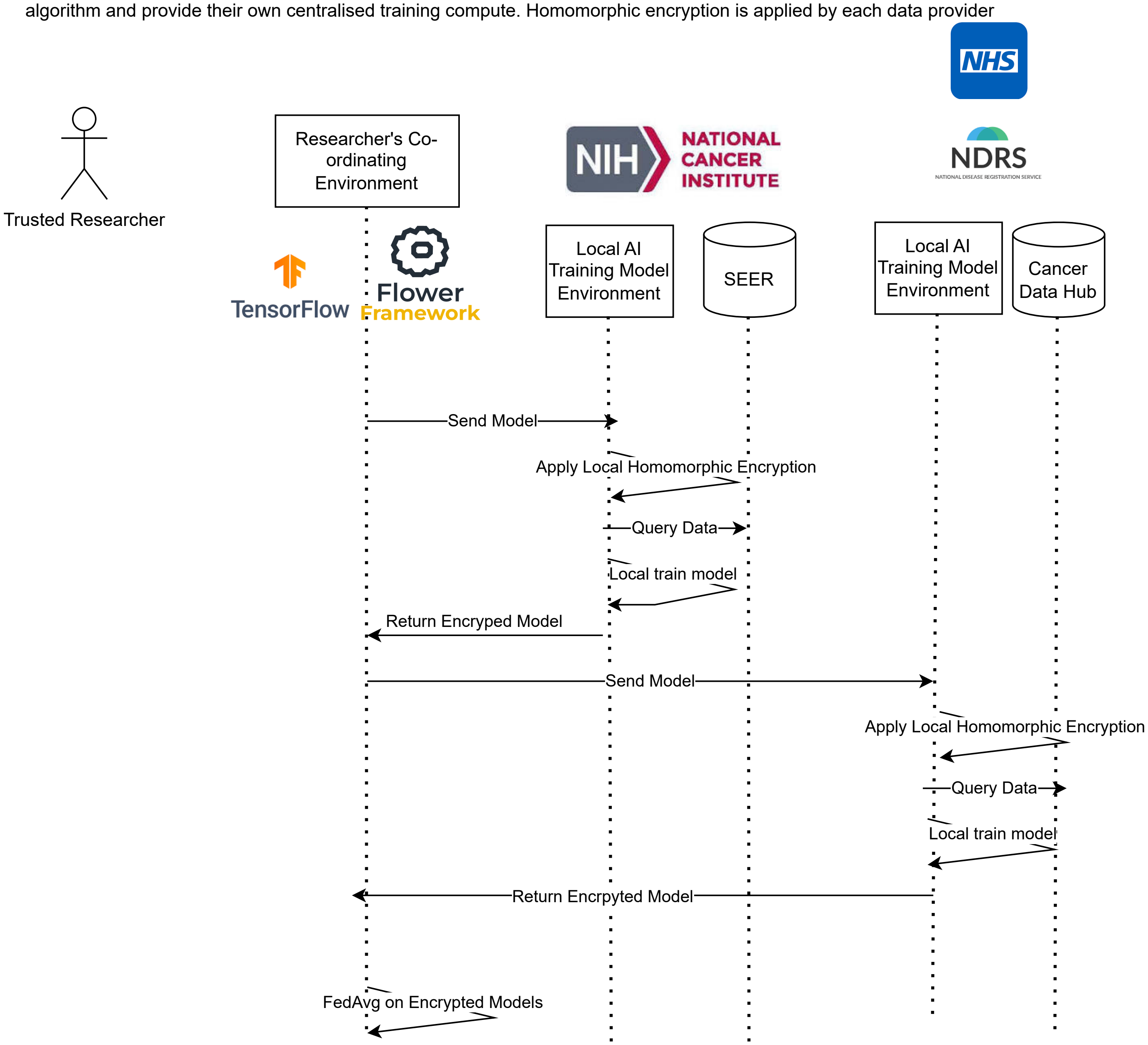


Assumptions:

1. Tensor Flow Federated or Flower framework could both be used as orchestrators of federated learning use case
2. Each data provider would have to provide a local AI training model environment where code would be deployed
3. Models are trained to work against local data so need to be deployed to local training envrionments
4. Orchestration feeds back to the trusted researcher where FedAvg or other algorithm is applied across the models

Federated Learning + Homomorphic Encryption Use Case Flow

A trusted researcher with access to data wishes to train their model against UK & US data. They bring their own algorithm and provide their own centralised training compute. Homomorphic encryption is applied by each data provider

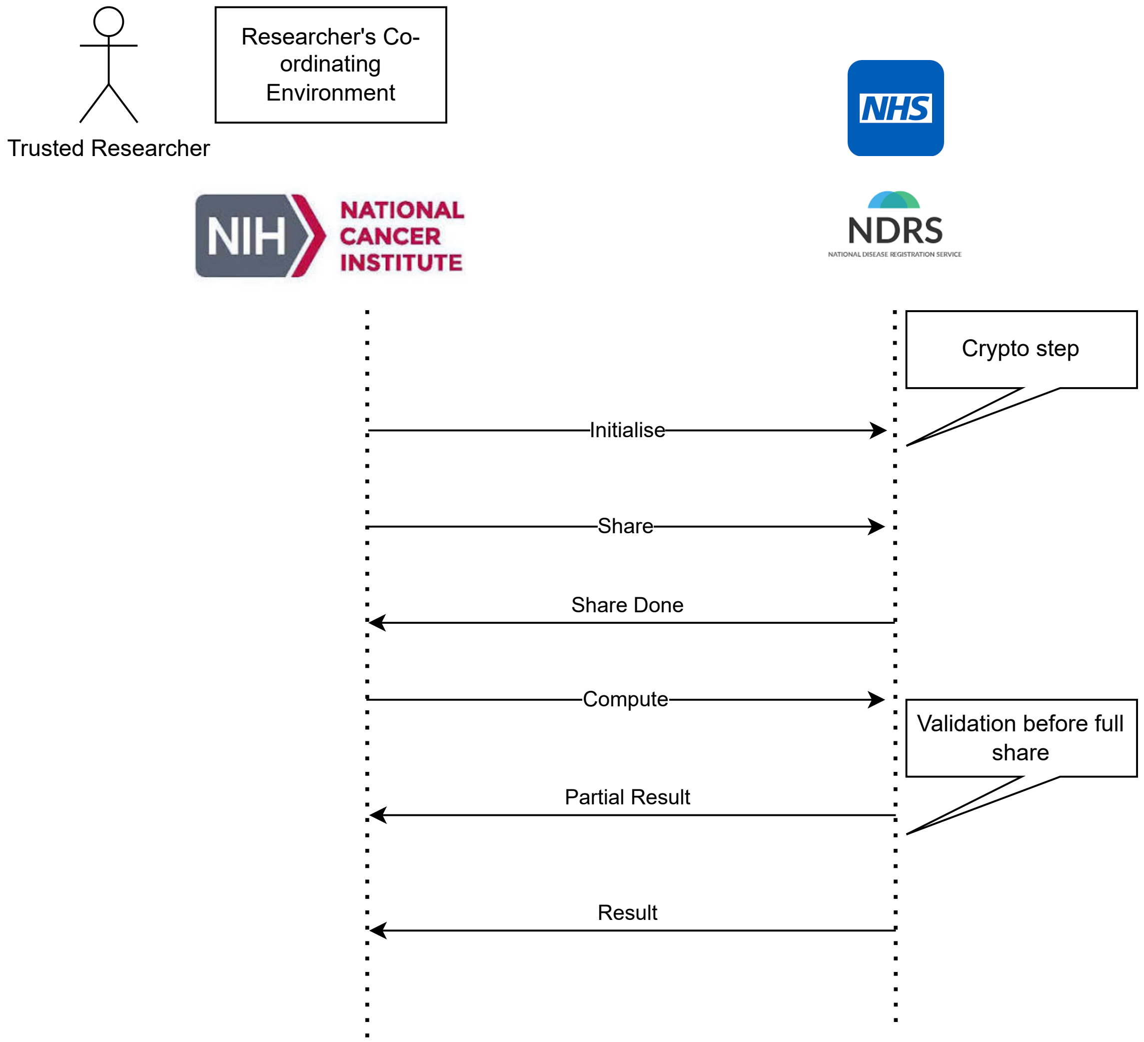


Assumptions:

1. Tensor Flow Federated or Flower framework could both be used as orchestrators of federated learning use case
2. Each data provider would have to provide a local AI training model environment where code would be deployed
3. Models are trained to work against local data so need to be deployed to local training envrionments
4. Each data provider enforces their version of homomorphic encryption
5. Orchestration feeds back to the trusted researcher where FedAvg or other algorithm is applied across the models in encrypted form

Secure Multi Party Computation Use Case Flow

in a Multi Party Computation flow the user accesses via one of the parties in the SMPC pair-wise relationships. In this example the access starts via the US.



Assumptions:

1. Initialize instantiates the agreed protocol

2. Participant A and Participant B share their private inputs with each other, typically in a secret-shared form.

3. Once both participants have shared their inputs, they acknowledge to each other that the sharing process is complete. This ensures that both parties are ready to proceed with the computation.

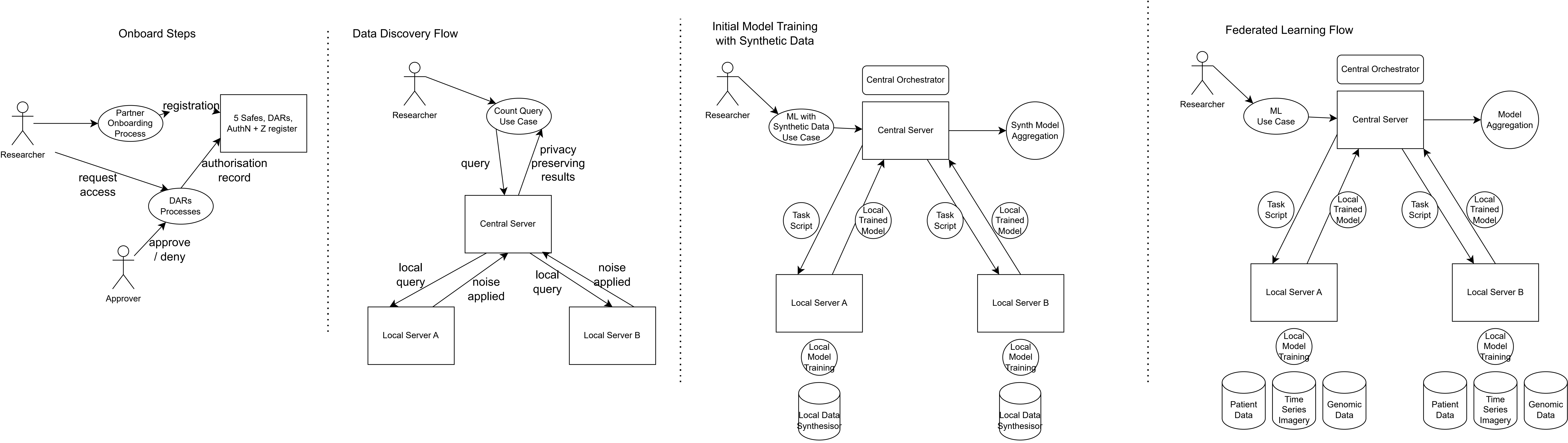
4. Participant A and Participant B independently perform their local computations on the shared inputs. Each participant computes their partial results based on the agreed-upon computation function.

5. Participant A and Participant B exchange their partial results with each other. This allows each participant to verify the correctness of their computation and combine the partial results to obtain the final result

Privacy Enhancing Technologies and Test Driven Security

Agree testing framework in advance of the technology implementation

	What	How	Who
Functional Testing	- Test each PET individually - Test data anonymisation - Check data access controls & consent applied properly		
Security Testing	- Test against security threats - Pen testing - Verify encryption algorithms		
Data Minimisation Tests	- Confirm only minimum necessary data is collected - Check for unneccesary data fields		
Privacy Impact Assessments	- Conduct PIA for implemented PETs & small datasets - Mitigate identified risks		
User Acceptance Testing	- Involve end-users to test functional use cases are achieved - Collected feedback and review		
Compliance Auditing	- Verify that implemented PETs comply with relevant privacy regulations & standards - Ensure documentation is complete		
Third Party Auditing	- Engage external experts to validate implementation - Engage patient privacy groups		
Anonymity Tests	- Assess whether anonymisation is effective in avoiding re-identification - Test for all countries with added external social data, attempt reconstruction with allowed personage		
Performance Tests	- Assess performance impacts especially around encryption - Measure processing time and resource utilisation		


Onboard Steps
Researcher
Partner Onboarding Process
registration
5 Safes, DARs, AuthN + Z register
request access
DARs Processes
authorisation record
approve / deny
Approver
Data Discovery Flow
Researcher
Count Query Use Case
query
privacy preserving results
Central Server
local query
noise applied
local query
noise applied
Local Server A
Local Server B
Initial Model Training with Synthetic Data
Researcher
ML with Synthetic Data Use Case
Central Orchestrator
Central Server
Synth Model Aggregation
Task Script
Local Trained Model
Task Script
Local Trained Model
Local Server A
Local Server B
Local Model Training
Local Model Training
Local Data Synthesisor
Local Data Synthesisor
Federated Learning Flow
Researcher
ML Use Case
Central Orchestrator
Central Server
Model Aggregation
Task Script
Local Trained Model
Task Script
Local Trained Model
Local Server A
Local Server B
Local Model Training
Local Model Training
Patient Data
Time Series Imagery
Genomic Data
Patient Data
Time Series Imagery
Genomic Data