FastShap++

Differentially Private FastSHAP for Federated Learning Model Explainability

Background

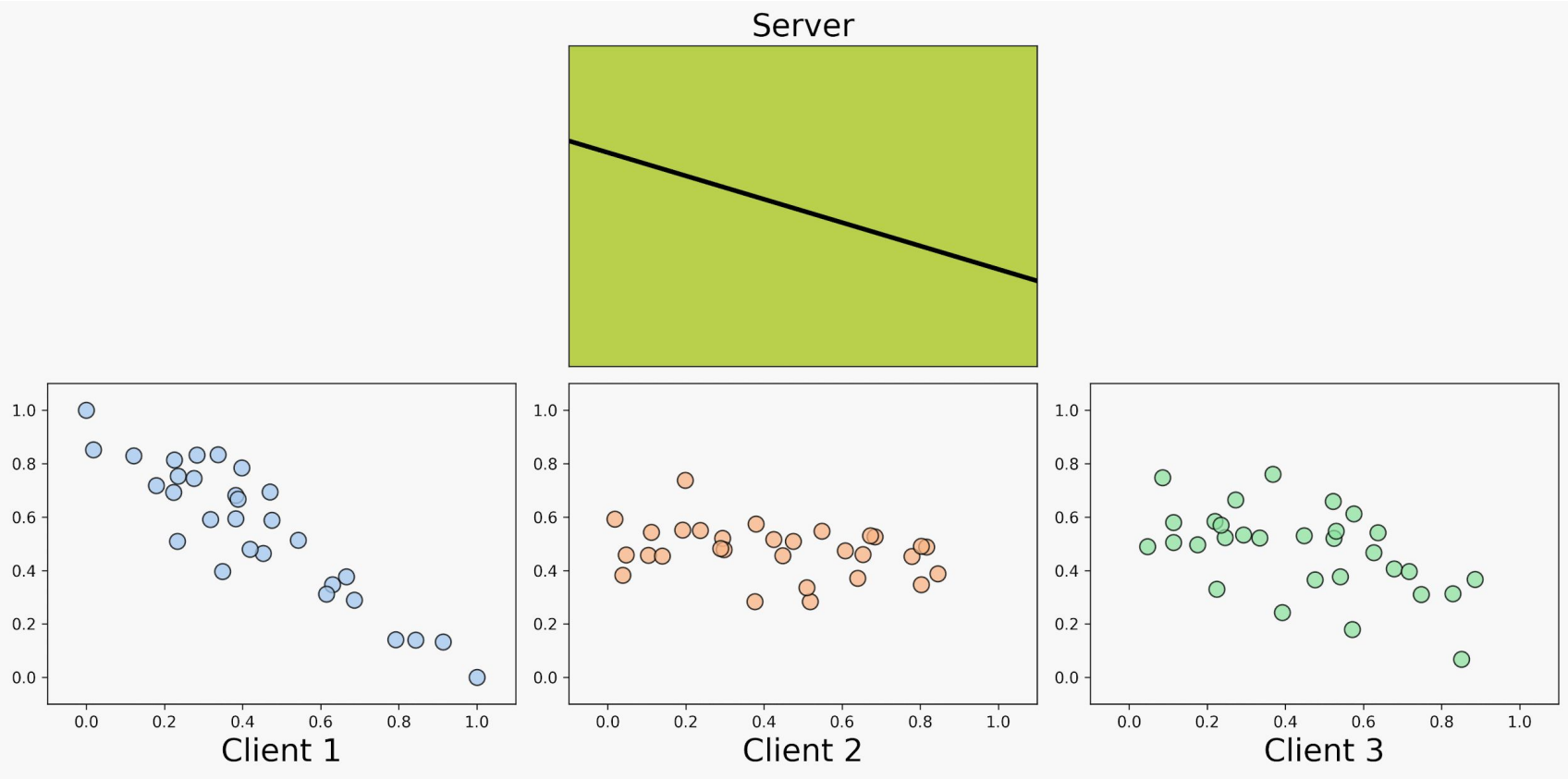
Federated Learning

eXplainability

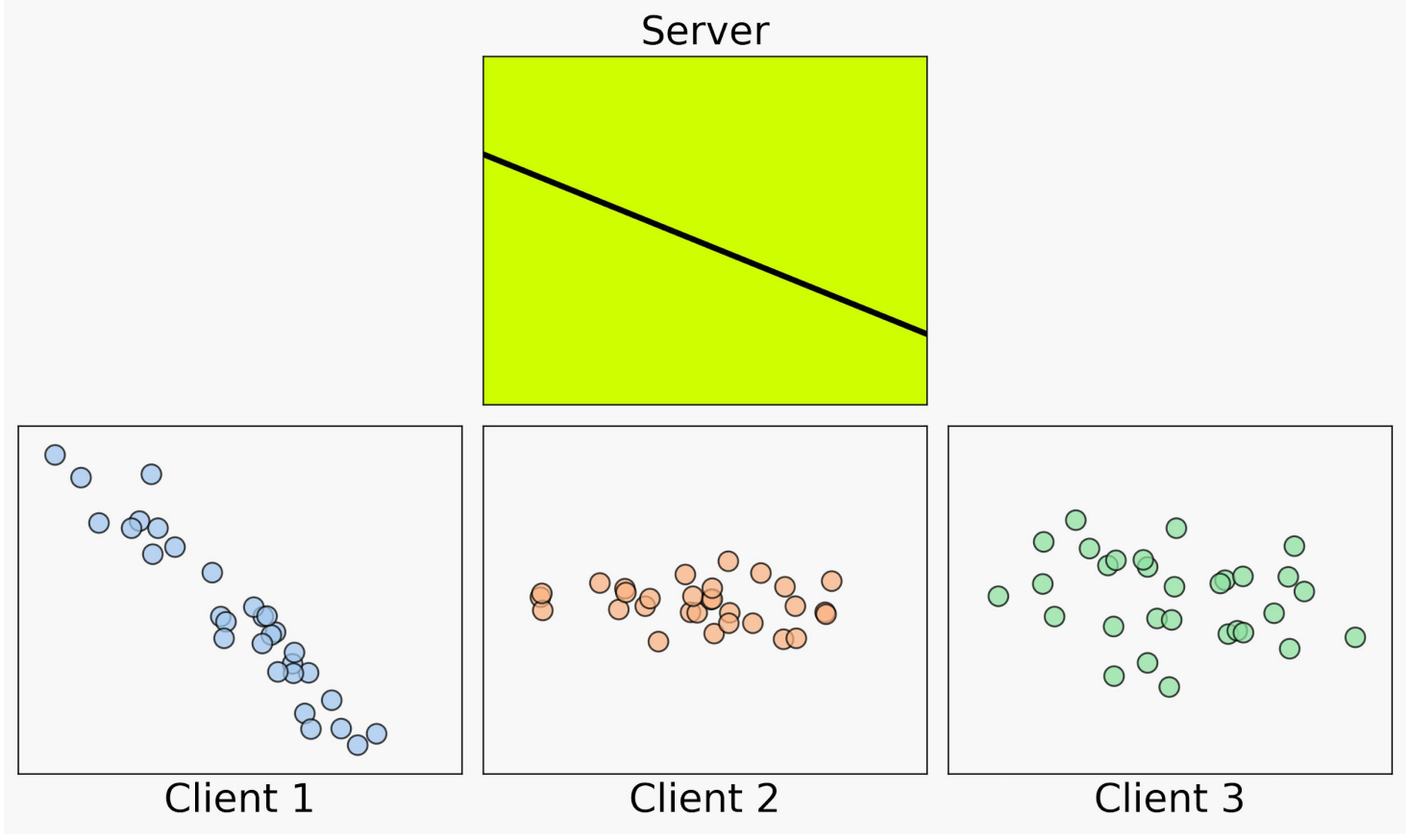
FastSHAP

Federated Learning

Federated Learning - FedAvg



Federated Learning - FedAvg



Flower Framework

Powerful Abstractions

High Flexibility

Fast Prototyping

Fast Deployment

eXplainable Artificial Intelligence

eXplainable Artificial Intelligence

Why explanations are produced?

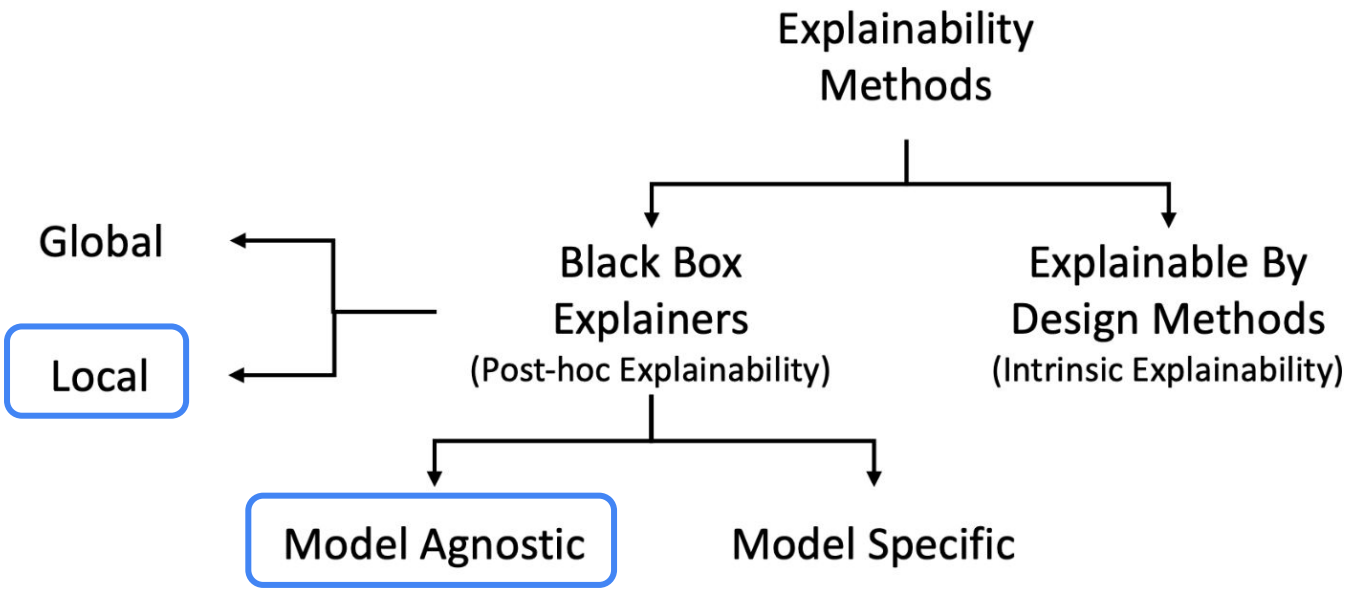
Red Team - Validation Centered

- **R**esearch on Data
- **E**xplore Models
- **D**ebug

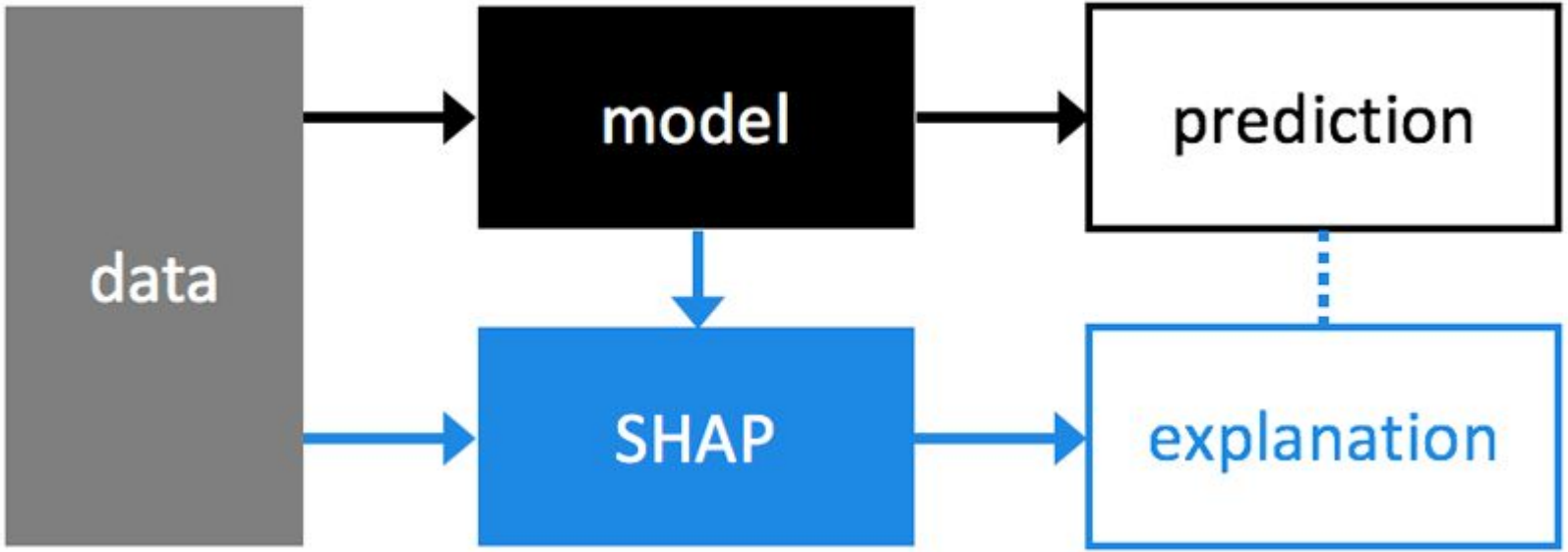
Blue Team - Human Centered

- responsi**B**le
- **L**egal issues
- tr**U**stfulness in predictions
- **E**thical issues

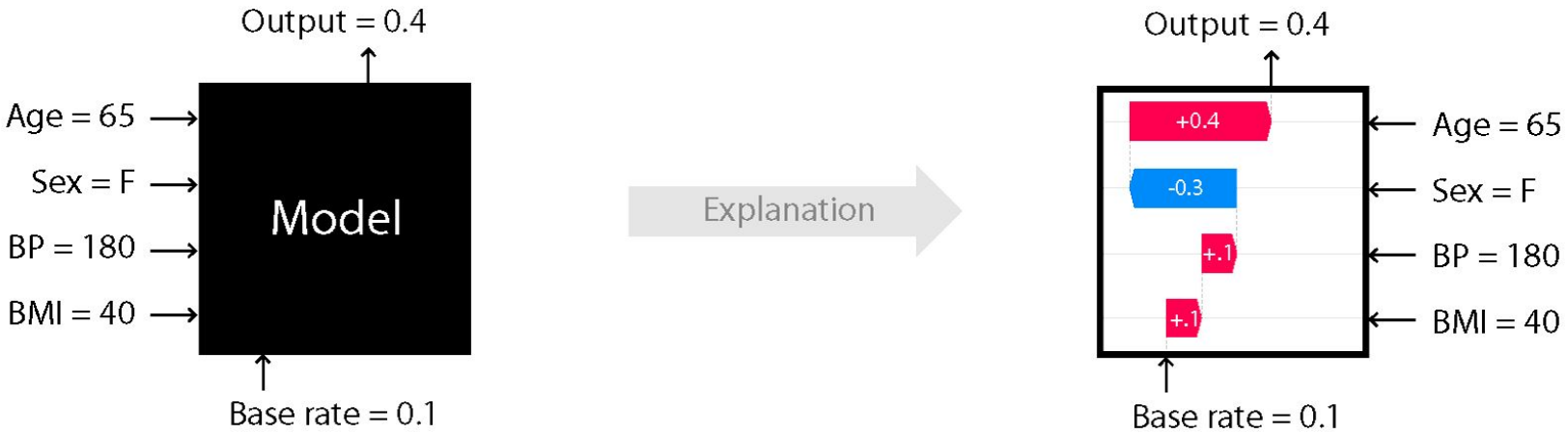
Explanation Taxonomy



Shapley Values as Feature Importance Explanations



Feature Importance Explanation



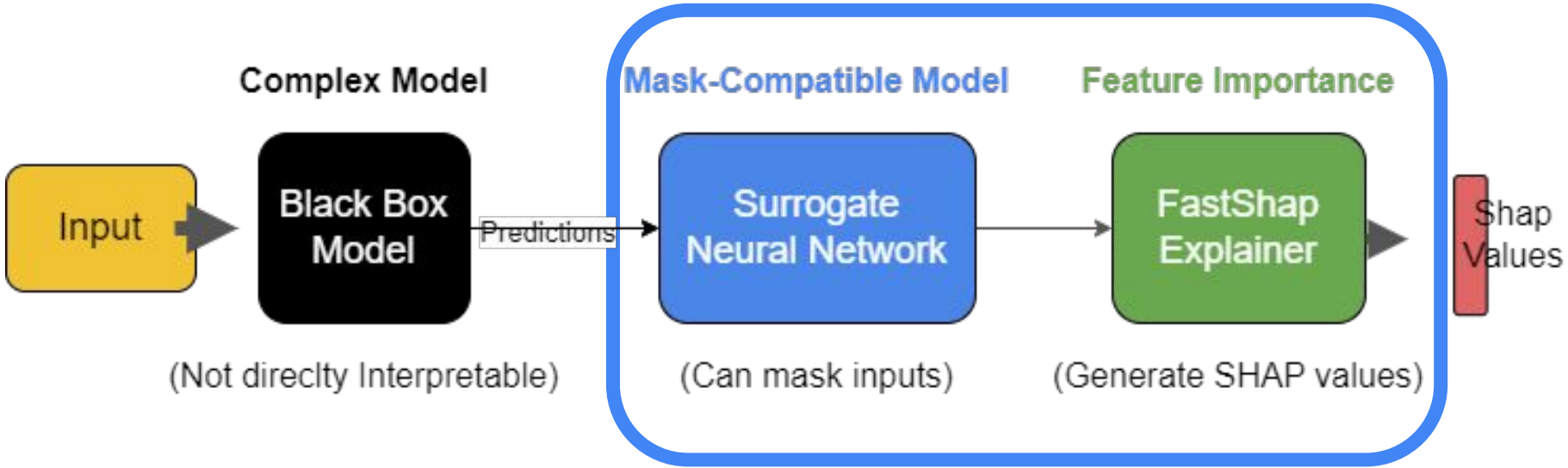
Privacy Enhancing Techniques

Differential Privacy

The higher the epsilon the less noise

FastShap

FastShap Architecture



Shapley values as FI Explanations

Shapley value

$$\phi_i(v) = \frac{1}{d} \sum_{s_i \neq 1} \binom{d-1}{\mathbf{1}^\top s}^{-1} \Big(v(s + e_i) - v(s) \Big)$$

Kernel Shap

$$\phi(v_{x,y}) = \underset{\phi_{x,y}}{\arg\min} \; \mathop{\mathbb{E}}_{p(\mathbf{s})} \Big[\big(v_{x,y}(\mathbf{s}) - v_{x,y}(\mathbf{0}) - \mathbf{s}^\top \phi_{x,y} \big)^2 \Big]$$
$$\text{s.t.} \quad \mathbf{1}^\top \phi_{x,y} = v_{x,y}(\mathbf{1}) - v_{x,y}(\mathbf{0}),$$

FastShap

$$\phi_{\text{fast}}(\mathbf{x}, \mathbf{y}; \theta^*) = \phi(v_{\mathbf{x},\mathbf{y}}) \;\; \text{almost surely in} \;\; p(\mathbf{x}, \mathbf{y})$$

Imitating Black Box with masks (Surrogate)

- D_{KL}:Kullback Leibler Divergence
- gamma: Black Box
- theta: model's parameters
- y: bb prediction
- m: mask function
- x: instance
- b: mask
- beta: surrogate parameters

$$\mathcal{L}(\beta) = \mathop{\mathbb{E}}_{p(x)} \mathop{\mathbb{E}}_{p(b)} \left[D_{KL}(\gamma(x;\theta) || \hat{\gamma}(y|m(x,b);\beta) \right]$$

Predicting SHAP Values (Explainer)

- Unif(y): uniform distribution over the classes

$$\mathcal{L}(\eta) = \mathop{\mathbb{E}}_{p(x)} \mathop{\mathbb{E}}_{\text{Unif}(y)} \mathop{\mathbb{E}}_{p(b)} \left[\left(v_{x,y}(b) - v_{x,y}(0) - b^T \phi_{fast}(x, y; \eta) \right)^2 \right]$$

Predicted: What the explainer thinks will happen when using certain features

Actual: What actually happens when those feature combinations are tested

Challenges And Current approaches

XAI in Federated Learning Context

Current Solutions and Challenges

Challenges

The majority of Explanation methods (specifically **SHAP** based) needs some data

- Kernel Shap Explainers need a reference dataset

- Data cannot be shared in FL
- Every client has its data, therefore local Shap Explanations differs
- Having a fixed, shared reference dataset makes the computation easier
  - Yet does not protect the local data against privacy attacks

Current Approaches

Specifically to SHAP

- Federated Fuzzy C Means as Background Data Points
  - Shares Centroids
- Aggregation of explanations or Explanations *as a Service*
  - Share Explanations

Both have strong points and weak points

Our solution approaches the challenge differently

FastShap++

Differentially Private FastSHAP for Federated
Learning Model Explainability

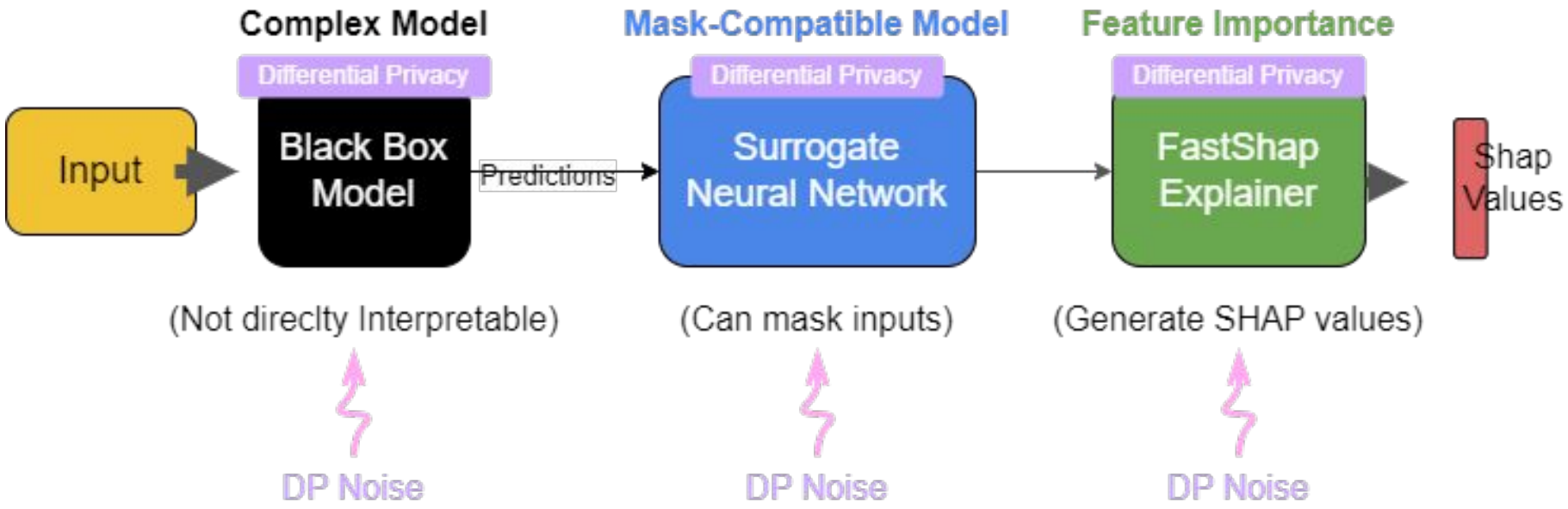
Quick explanations

Privacy Enhancing Technology mechanism

Share only and solely the model's weights

FastShap +

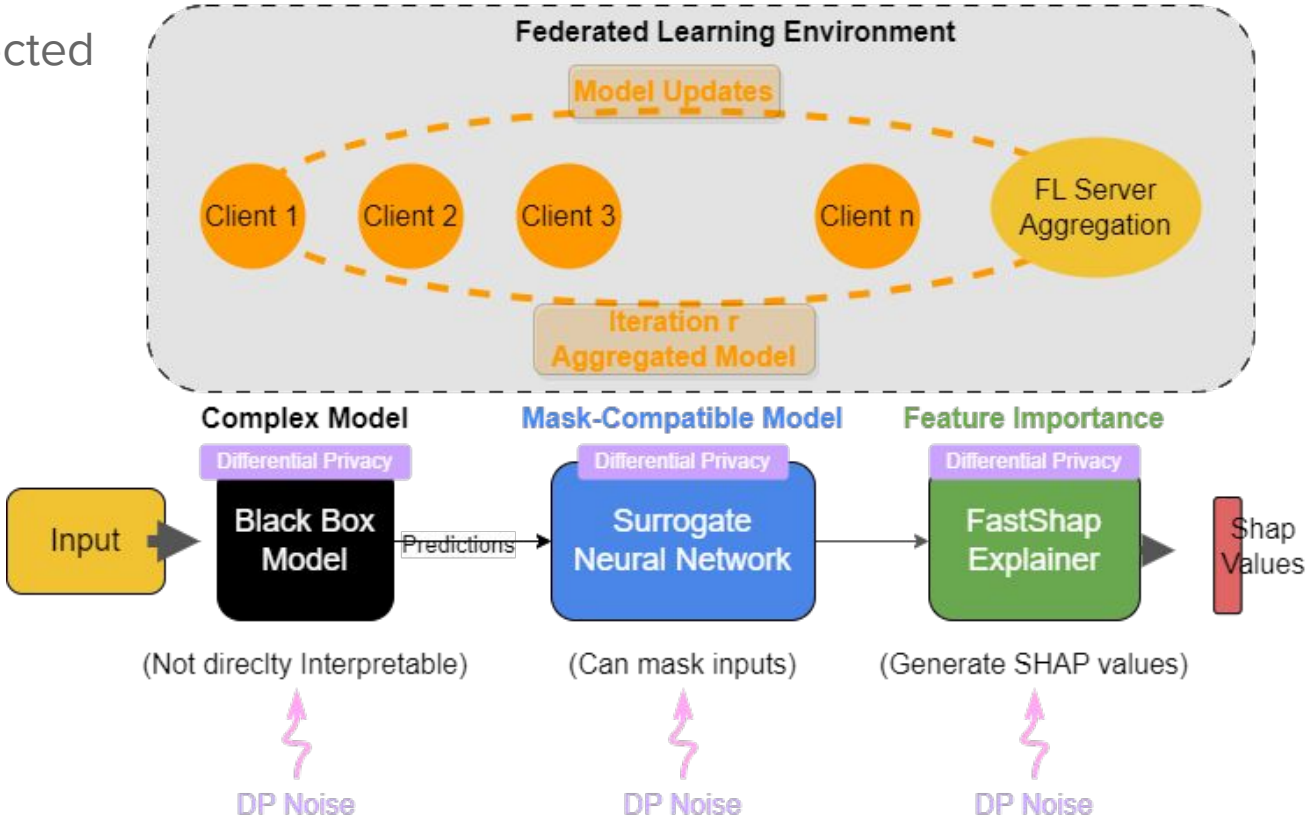
Privacy Protected - Differential Privacy



FastShap ++

Privacy Protected

Federated



Experiments

Experiments

Dataset

- Dutch: 70 Clients
- Income: 51 Clients
- Employment: 51 Clients

Modalities

- *Vanilla* Pipeline benchmark
- Semi-private (only S)
- Full-Private
  - Multiple levels of privacy (only E)

Measures

- BlackBox
  - Accuracy
- Surrogate
  - Fidelity
- Explainer
  - ShapGaps
    - L2-distance
    - Cosine Similarity
    - Feature Agreement
    - Sign Agreement
    - Ranking Correlation
    - Delta Faithfulness

Results

DP impact in Centralized

	Accuracy BB	Fidelity Surrogate
Dutch	0.83±0.01	0.97±0.01
Dutch (DP)	0.82±0.01	0.91±0.01
ACSI	0.77±0.01	0.97±0.01
ACSI (DP)	0.77±0.02	0.95±0.01
ACSE	0.80±0.01	0.95±0.01
ACSE (DP)	0.74±0.03	0.90±0.01

average and standard deviation on three runs

Degradation in performance for the Black Box (using DP as PET)

Small yet visible Impact of DP on Surrogate model

Those results say that epsilon-privacy guarantee come at the cost of accuracy or fidelity

Centralized vs Federated

Measuring the differences when explaining the Black Box Federated using

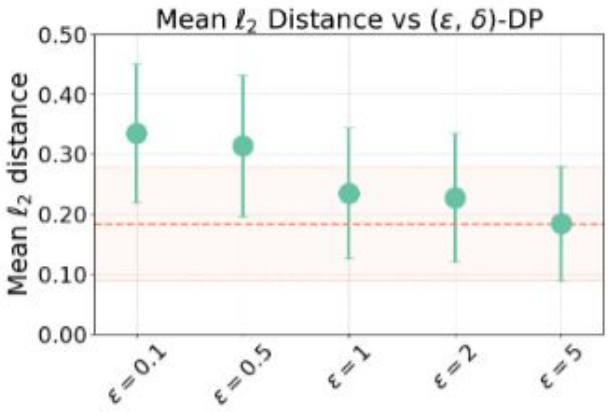
- Centralized FastShap

Compared with

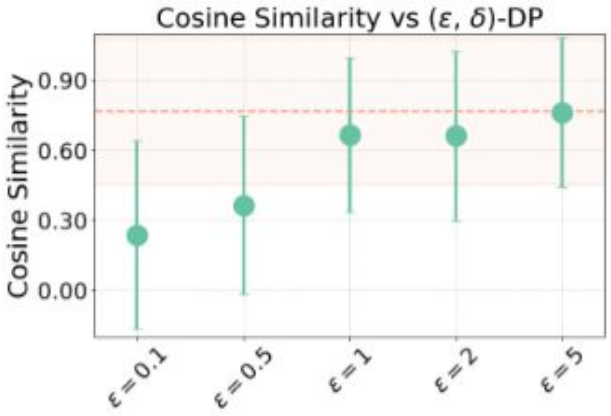
- Federated FastShap

	ℓ_2 Dist. (↓)	Cosine Sim. (↑)	Feat. Agr. (↑)	Sign Agr. (↑)	Rank Corr. (↑)	Δ Faith (↓)
Dutch	0.03±0.02	0.99±0.01	0.87±0.13	0.87±0.13	0.84±0.11	0.02±0.03
ACSI	0.09±0.05	0.81±0.22	0.77±0.14	0.70±0.17	0.70±0.17	0.13±0.10
ACSE	0.10±0.05	0.80±0.19	0.66±0.16	0.64±0.17	0.65±0.16	0.21±0.18

Explanation Measures - 1



(a)



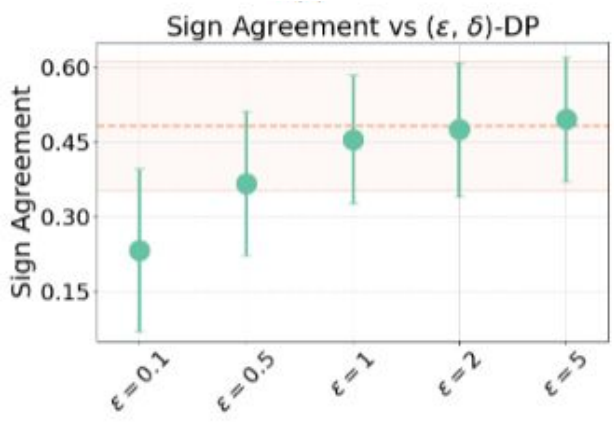
(b)

(c)

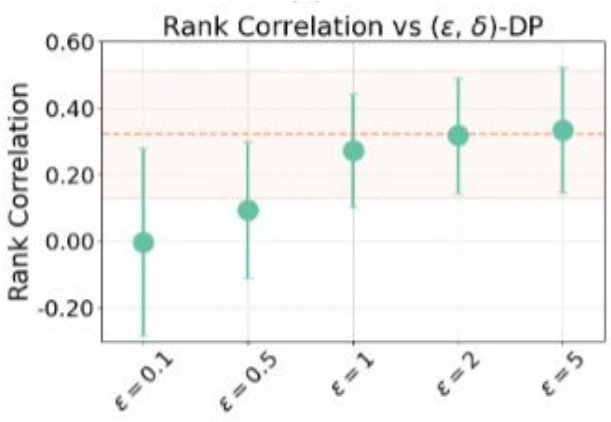
W/o Privacy VS Semi Private

W/o Privacy VS Full Private

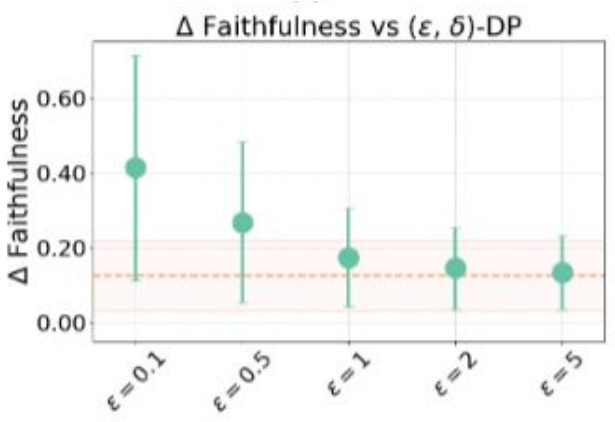
Explanation Measures - 2



(d)



(e)



(f)

W/o Privacy VS Semi Private W/o Privacy VS Full Private

Future Work

Next experiments with

- For image data
- Introduction of Fairness
  - Private, Fair and Federated
    - Fair explanations metrics
    - Fairness as transferable, stable portable property
      - Measured with surrogate
        - Fidelity
      - And explainer
        - Specific measures

Thanks - QA