Data Reconstruction Attacks in AI models are Inverse Problems

Qi Lei

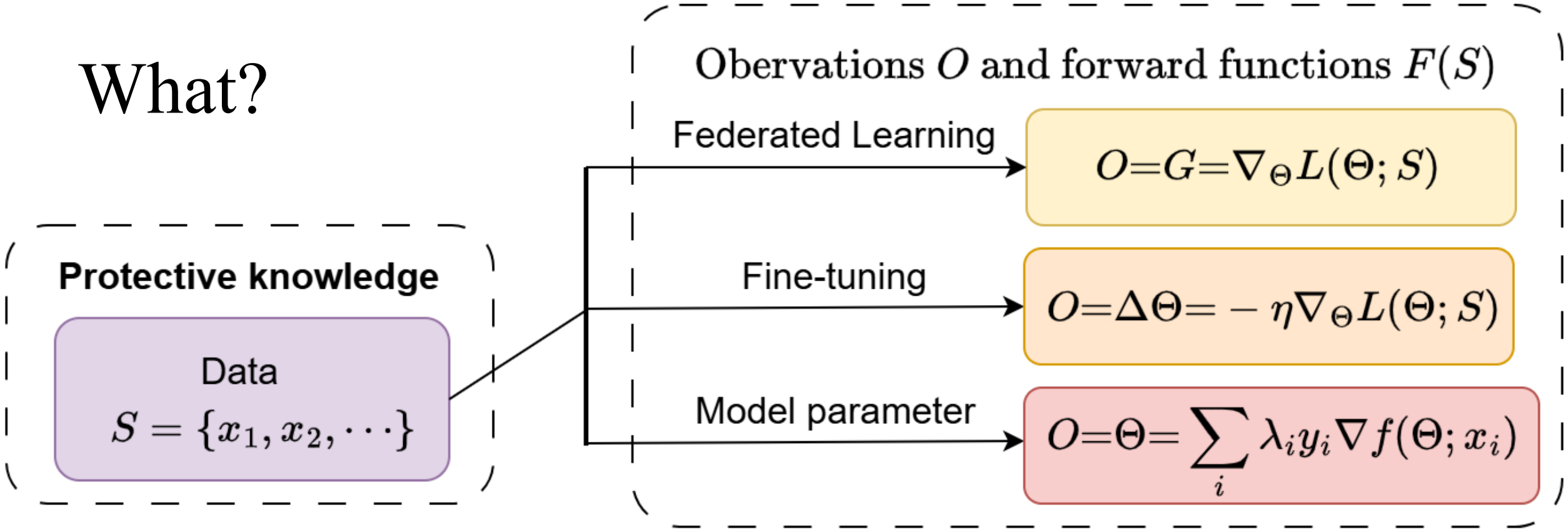
Courant Math & Center for Data Science @ NYU

w/ Benjamin Peherstorfer

Inverse Methods for Complex Systems under Uncertainty Workshop

NYU COURANT INSTITUTE OF MATHEMATICAL SCIENCES

What?

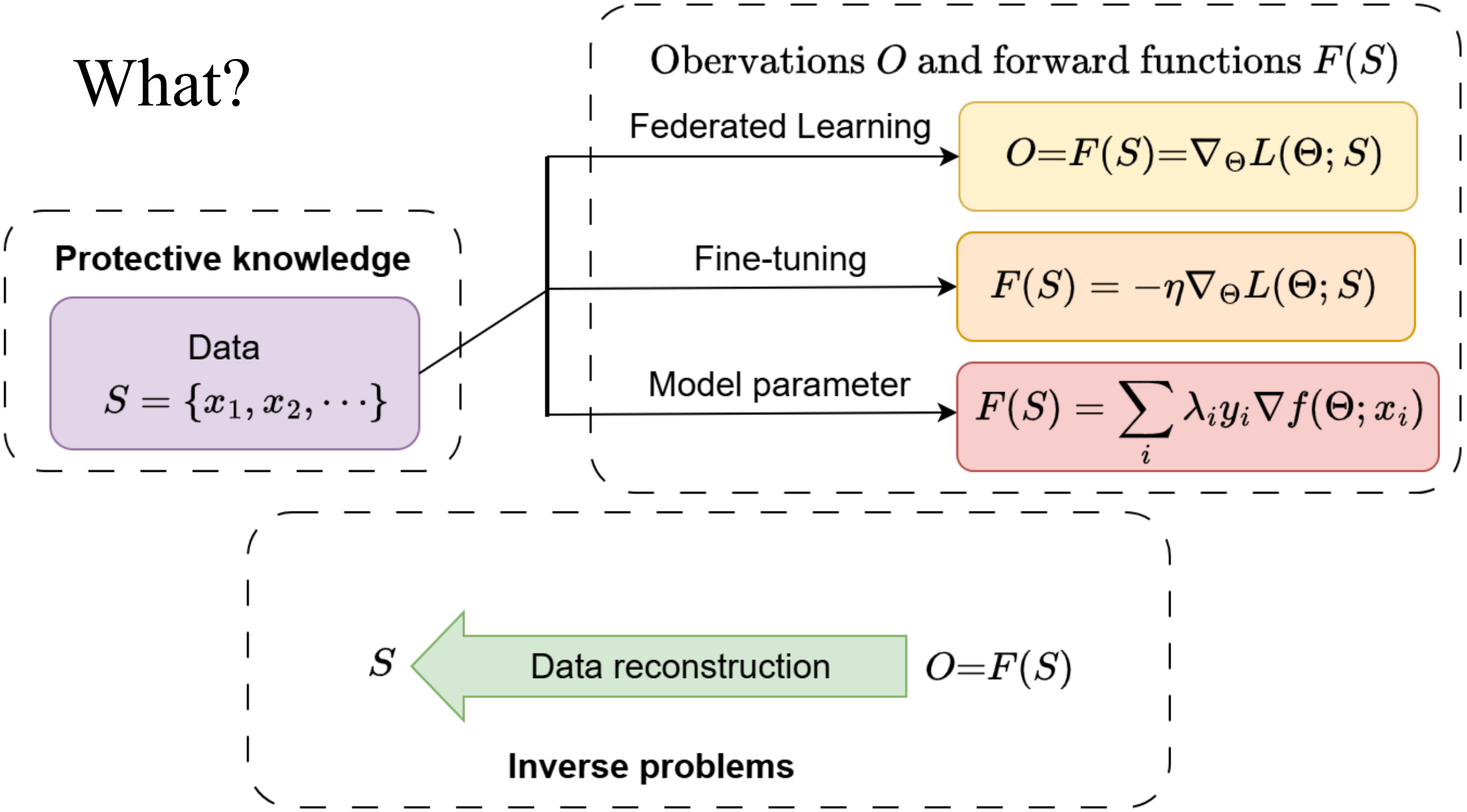


Neural network f parameterized by Θ

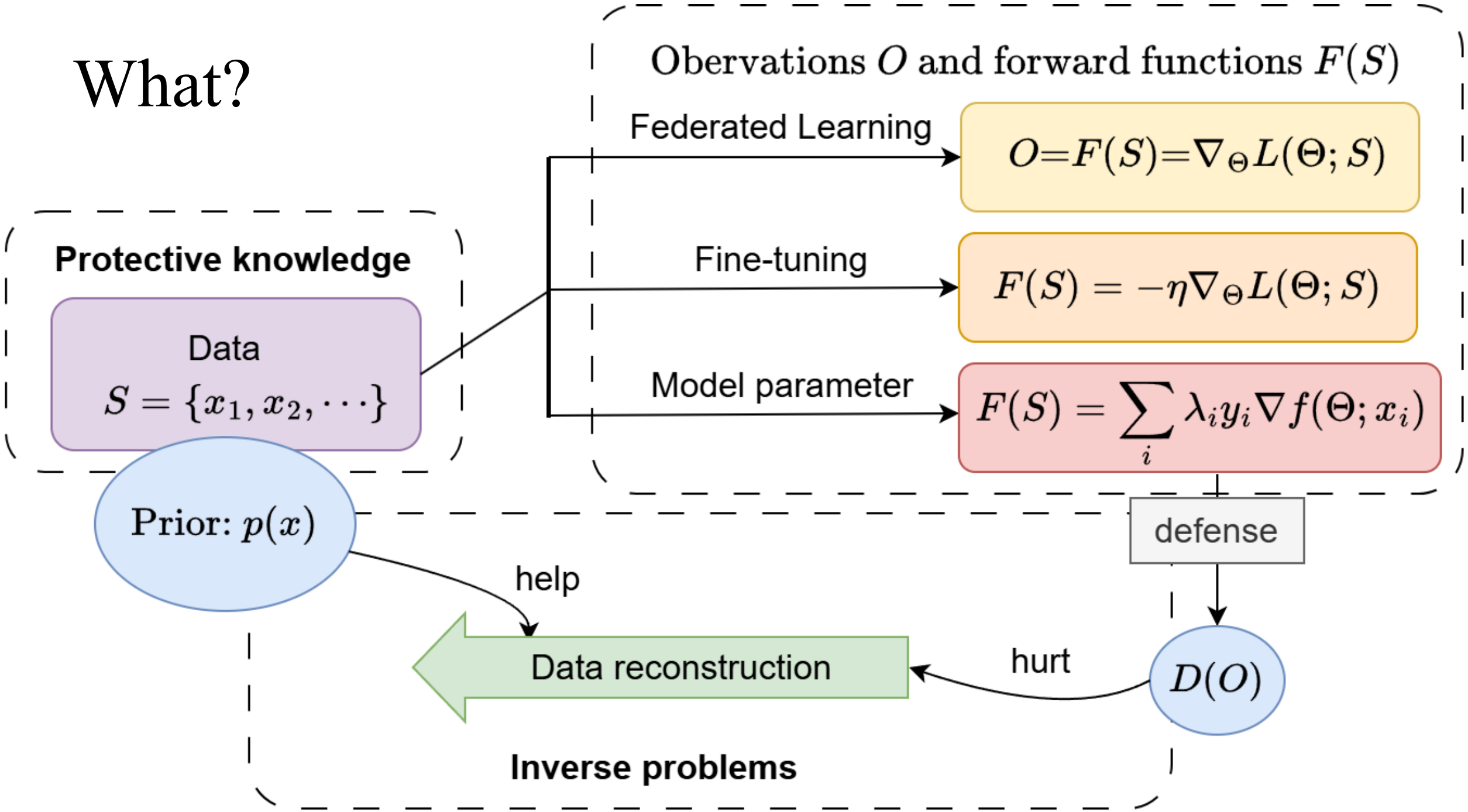
Loss function L

Gradient G

What?


Obervations O and forward functions $F(S)$
Federated Learning
$O=F(S)=\nabla_\Theta L(\Theta;S)$
Fine-tuning
$F(S)=-\eta\nabla_\Theta L(\Theta;S)$
Model parameter
$F(S)=\sum_i \lambda_i y_i \nabla f(\Theta;x_i)$
Protective knowledge
Data
$S=\{x_1,x_2,\cdots\}$
S
Data reconstruction
$O=F(S)$
Inverse problems

What?


Obervations O and forward functions $F(S)$
Federated Learning
$O=F(S)=\nabla_\Theta L(\Theta;S)$
Fine-tuning
$F(S)=-\eta\nabla_\Theta L(\Theta;S)$
Model parameter
$F(S)=\sum_i \lambda_i y_i \nabla f(\Theta;x_i)$
Protective knowledge
Data
$S=\{x_1,x_2,\cdots\}$
Prior: $p(x)$
help
Data reconstruction
hurt
$D(O)$
defense
Inverse problems

Why?

- Linking inverse problem to AI safety (privacy) threats
- Understand memorization in trained neural networks

- Trillions of parameters
- Tend to memorize high quality data

Why not DP?

- **Too strong**: not revealing one bit of information
- **Not necessary**: reconstructing data verbatim is more of copyright violation
- **Unpractical**: industry set ε=1000 in ε-Differential Privacy

How?

Challenges:

- Computational difficulty:
  - Highly non-convex landscape

- Statistical difficulty:
  - With defense: partial/noisy observation

- Multimodal
  - Applications: medical, vision, NLP
  - Incorporate prior knowledge

Existing work:

- Gradient inversion attack
- Tensor/PCA method

- Reconstruction error theoretical analysis

- Incorporate generative priors for time series, graphs, images, language.

Data Reconstruction Attacks in AI models are Inverse Problems:

- Reconstructing Training Data from Model Gradient, Provably. Zihan Wang, Jason Lee, Qi Lei. AISTATS 2023
- Data Reconstruction Attacks and Defenses: A Systematic Evaluation. Sheng Liu, Zihan Wang, Yuxiao Chen, Qi Lei. AISTATS 2025
- Beyond Gradient and Priors in Privacy Attacks: Leveraging Pooler Layer Inputs of Language Models in Federated Learning Jianwei Li, Sheng Liu, Qi Lei. NeurIPS Federated Learning workshop 2023