Latent Gaussian Compression

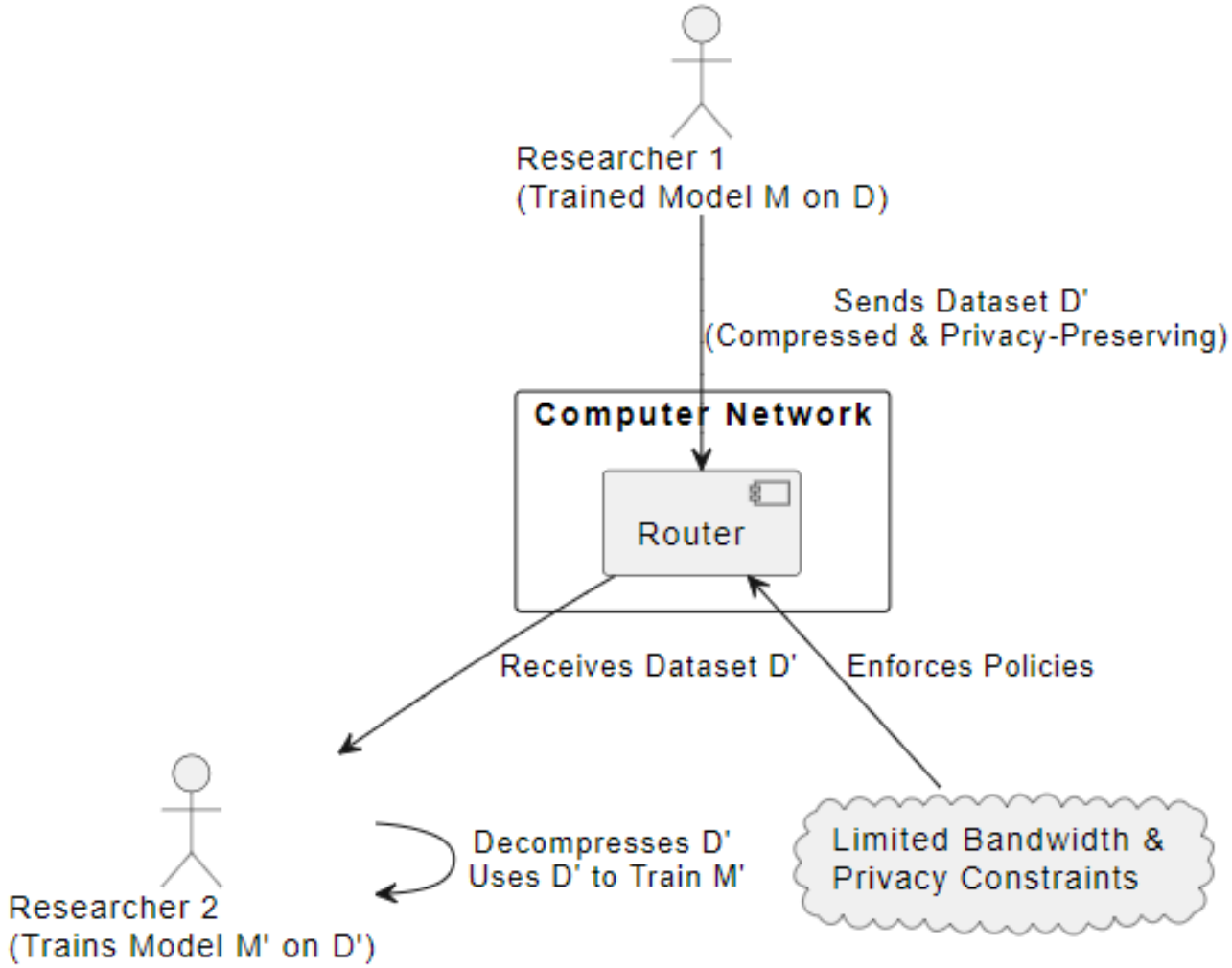
James Zhao, Blaine Arihara, Emily Tang, Terry Weber

Problem Setup

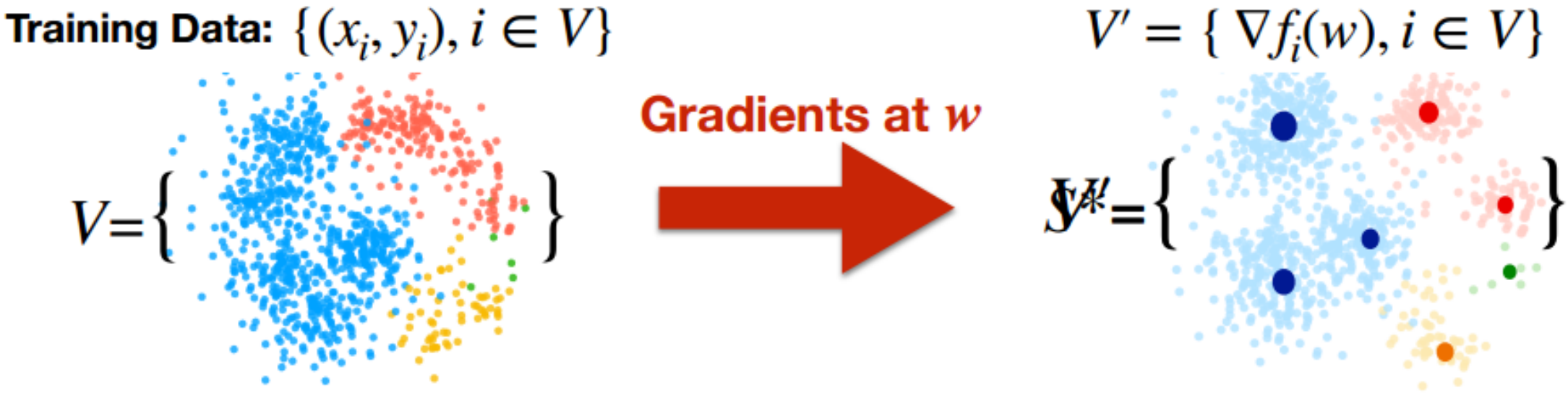
Suppose we have a dataset $D = \{Cat, Dog\}$ with two classes and we want to train a classifier.

- The Problem:
  - Cannot store or transmit full dataset D because of
    - Network bandwidth constraints
    - Space constraints
    - Privacy constraints
- Can we share compressed dataset D' (equivalent to coreset S_k) instead?

Problem Assumptions

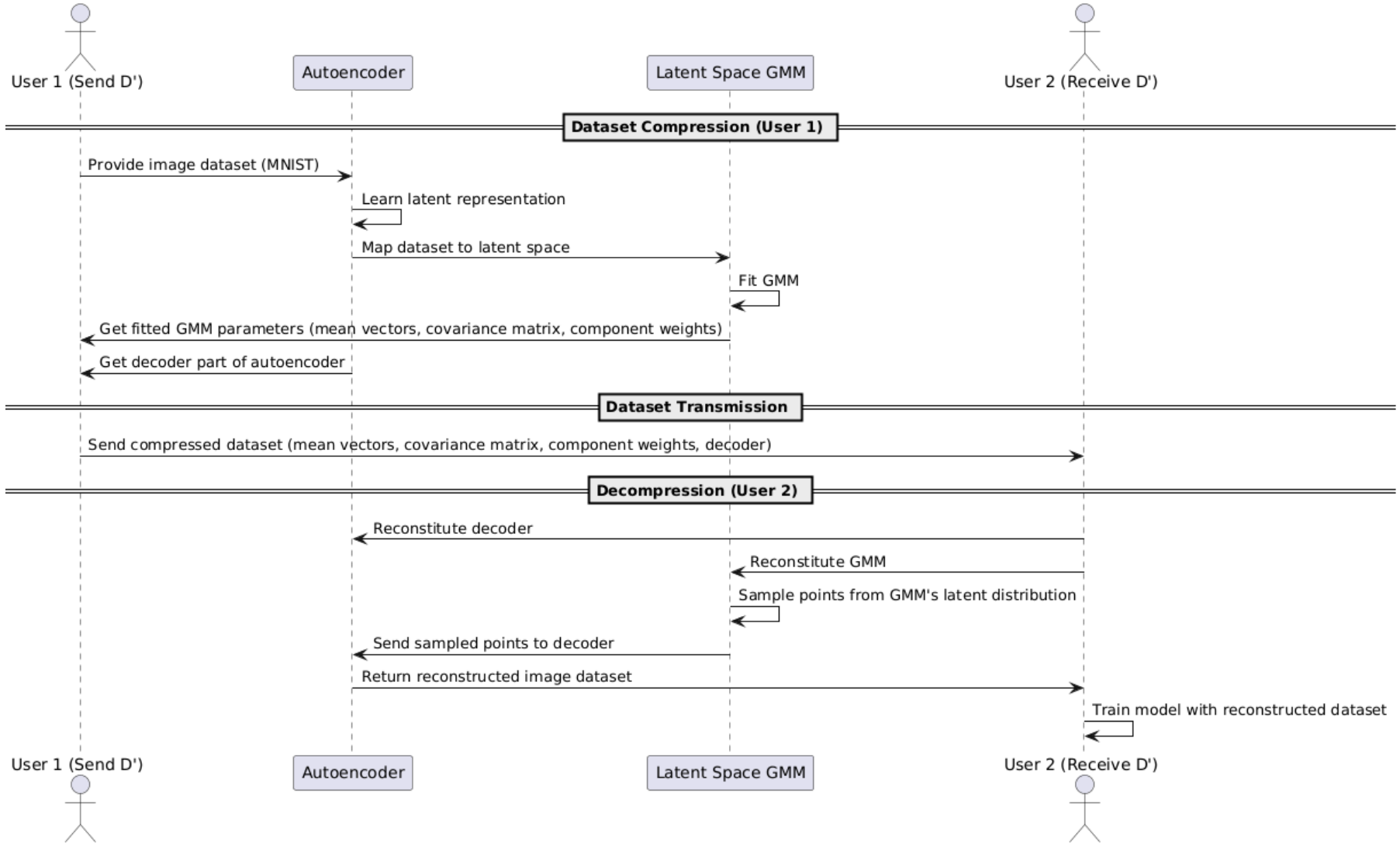


Existing Approaches



- Select subset $S*$ and obtain a $\frac{|V|}{|S*|}$ speedup and compression factor

Workflow


User 1 (Send D')
Autoencoder
Latent Space GMM
User 2 (Receive D')
Dataset Compression (User 1)
Provide image dataset (MNIST)
Learn latent representation
Map dataset to latent space
Fit GMM
Get fitted GMM parameters (mean vectors, covariance matrix, component weights)
Get decoder part of autoencoder
Dataset Transmission
Send compressed dataset (mean vectors, covariance matrix, component weights, decoder)
Decompression (User 2)
Reconstitute decoder
Reconstitute GMM
Sample points from GMM's latent distribution
Send sampled points to decoder
Return reconstructed image dataset
Train model with reconstructed dataset
User 1 (Send D')
Autoencoder
Latent Space GMM
User 2 (Receive D')

Gaussian Mixture Modeling

- Map original data in $R^n(A, B)$ to simpler latent space $R^l(A', B')$ where $l << n$.
- We can approximate the class distributions using Gaussian Mixture Models (GMMs):
  - Represent each class distribution $C' \in (A', B')$ as linear combinations of k Gaussian distributions:

$$P(z) = \sum_{i=1}^{k} \pi_i \mathcal{N}(\mu_{k_{C'}}, \Sigma_{k_{C'}}), \quad z \in R^l$$