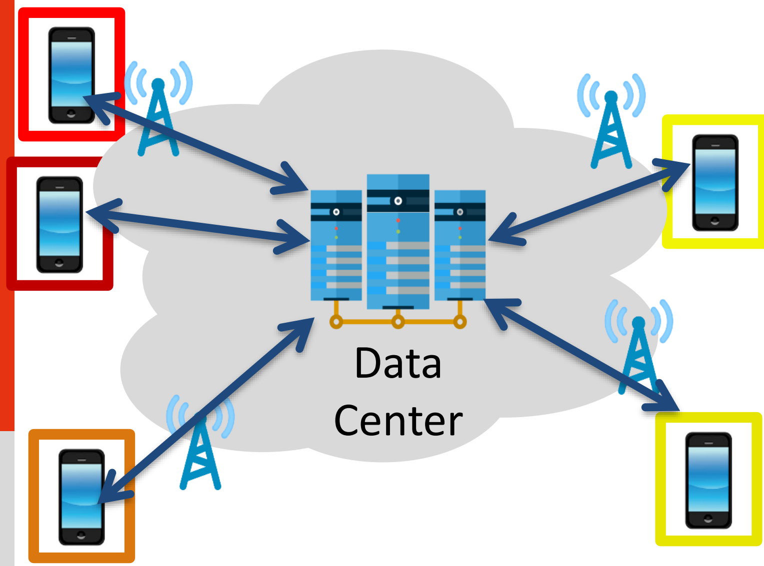
Personalized Federated Learning through Local Memorization

Othmane Marfoq[1, 2] Giovanni Neglia[1] Laetitia Kameni[2]

Richard Vidal[2]

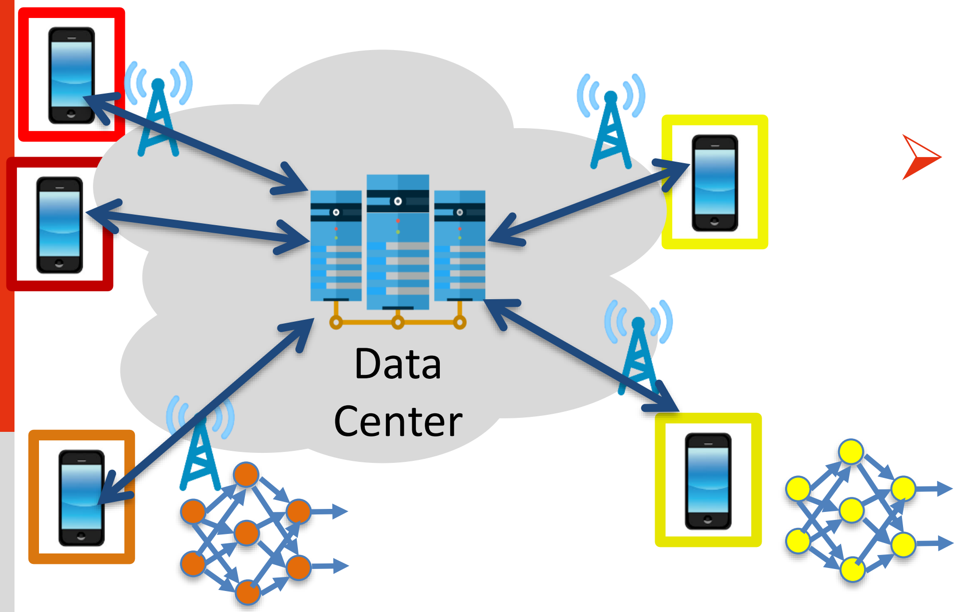
[1]Inria, Université Côte d’Azur and [2]Accenture Labs

Personalization



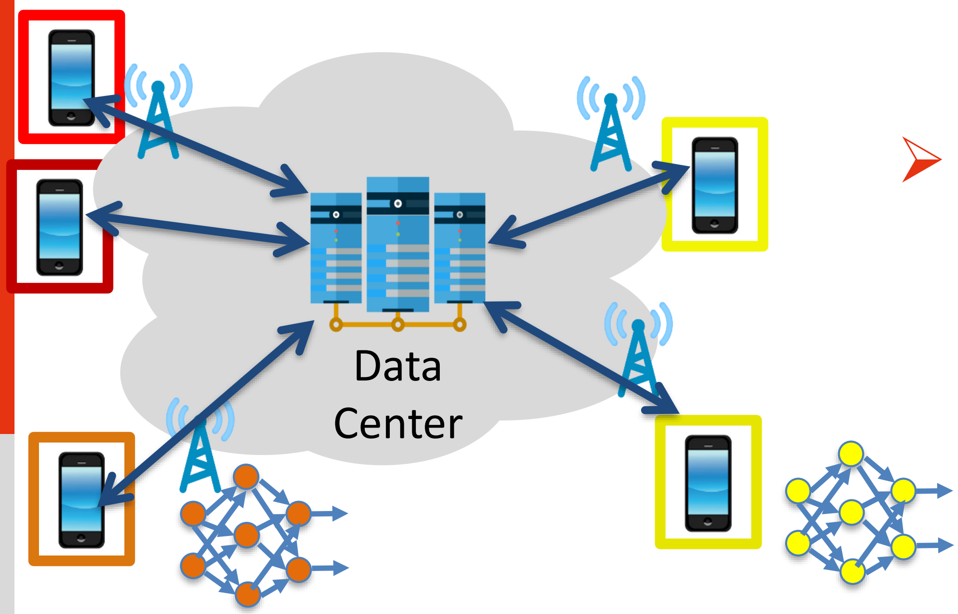
- Why a single model if local dataset come from a different distribution? Statistical heterogeneity

Personalization

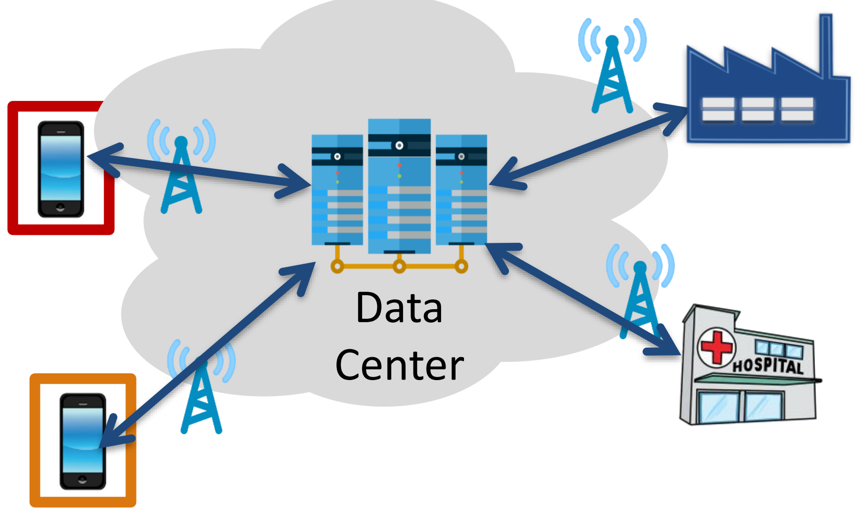


- Why a single model if local dataset come from a different distribution? Statistical heterogeneity

Personalization

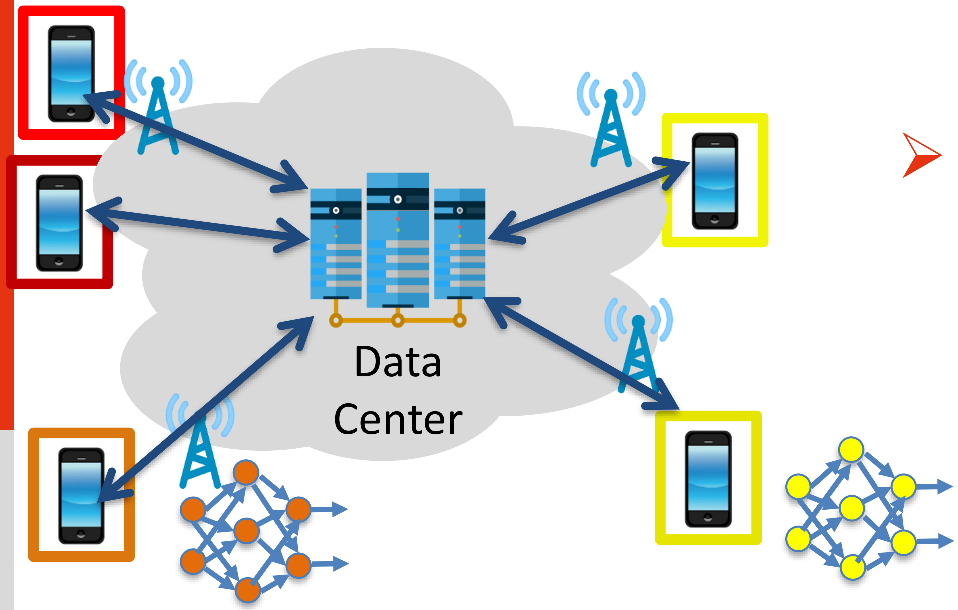


- Why a single model if local dataset come from a different distribution? Statistical heterogeneity

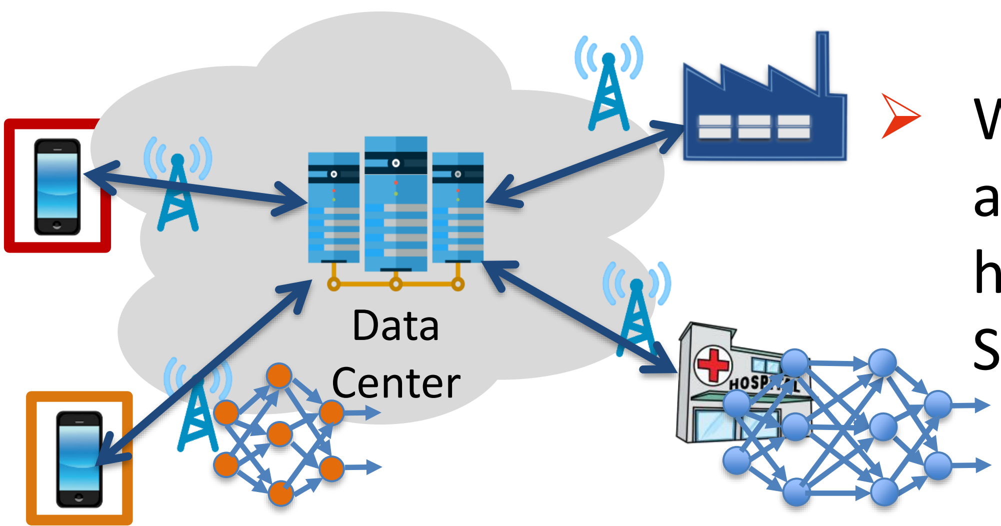


- Why the same model architecture when clients have different capabilities? System heterogeneity

Personalization

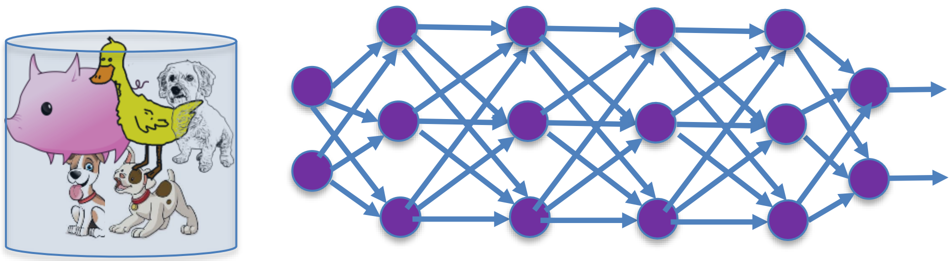


- Why a single model if local dataset come from a different distribution? Statistical heterogeneity



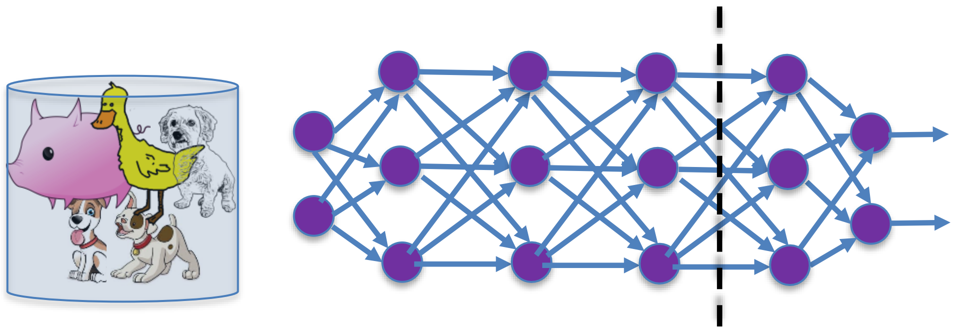
- Why the same model architecture when clients have different capabilities? System heterogeneity

Personalized FL through Local Memorization: background



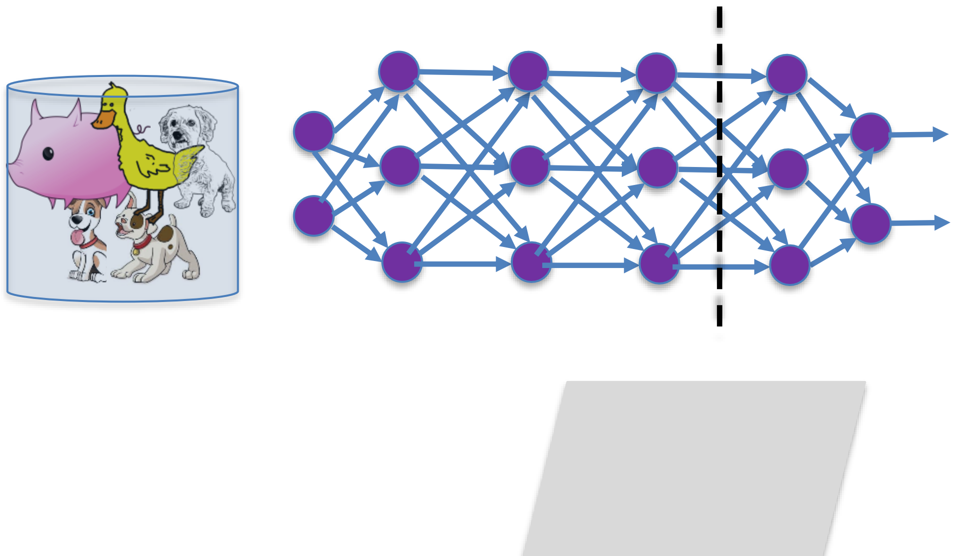
Orhan. A simple cache model for image recognition. NeurIPS, 2018.
Khandelwal, Levy, Jurafsky, Zettlemoyer, Lewis.
Generalization through Memorization: Nearest Neighbor Language Models. ICLR, 2020.

Personalized FL through Local Memorization: background



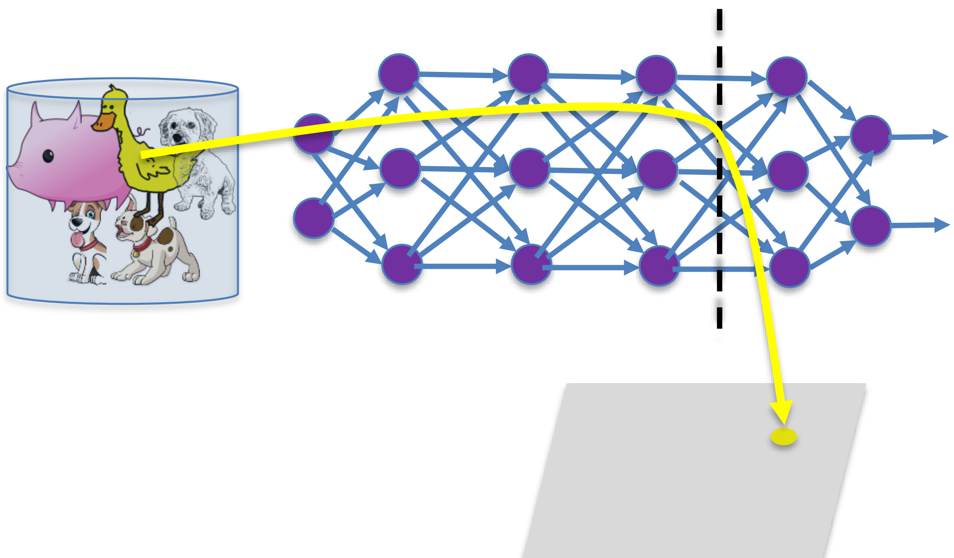
Orhan. A simple cache model for image recognition. NeurIPS, 2018.
Khandelwal, Levy, Jurafsky, Zettlemoyer, Lewis.
Generalization through Memorization: Nearest Neighbor Language Models. ICLR, 2020.

Personalized FL through Local Memorization: background



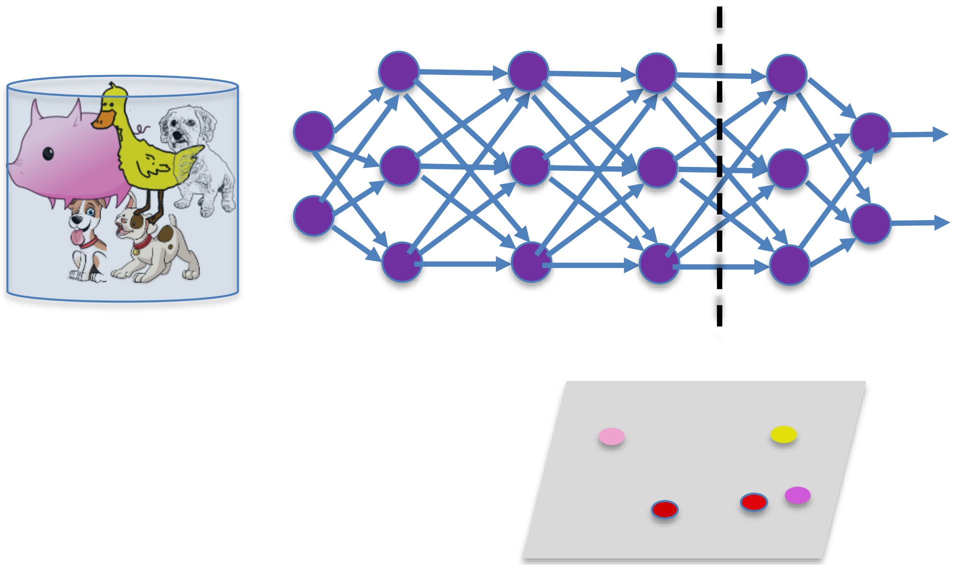
Orhan. A simple cache model for image recognition. NeurIPS, 2018.
Khandelwal, Levy, Jurafsky, Zettlemoyer, Lewis.
Generalization through Memorization: Nearest Neighbor Language Models. ICLR, 2020.

Personalized FL through Local Memorization: background



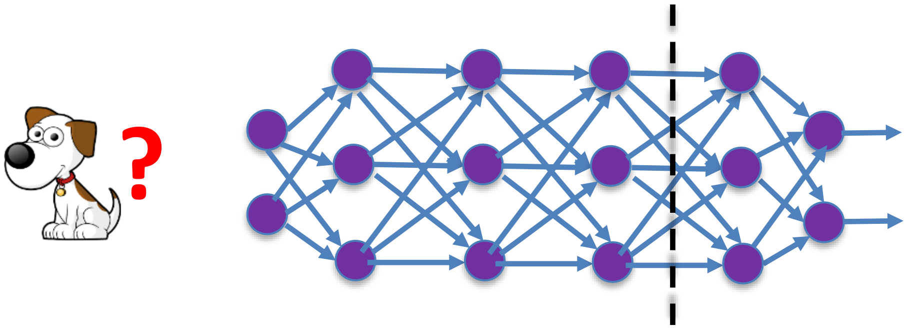
Orhan. A simple cache model for image recognition. NeurIPS, 2018.
Khandelwal, Levy, Jurafsky, Zettlemoyer, Lewis.
Generalization through Memorization: Nearest Neighbor Language Models. ICLR, 2020.

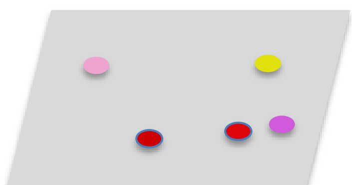
Personalized FL through Local Memorization: background



Orhan. A simple cache model for image recognition. NeurIPS, 2018.
Khandelwal, Levy, Jurafsky, Zettlemoyer, Lewis.
Generalization through Memorization: Nearest Neighbor Language Models. ICLR, 2020.

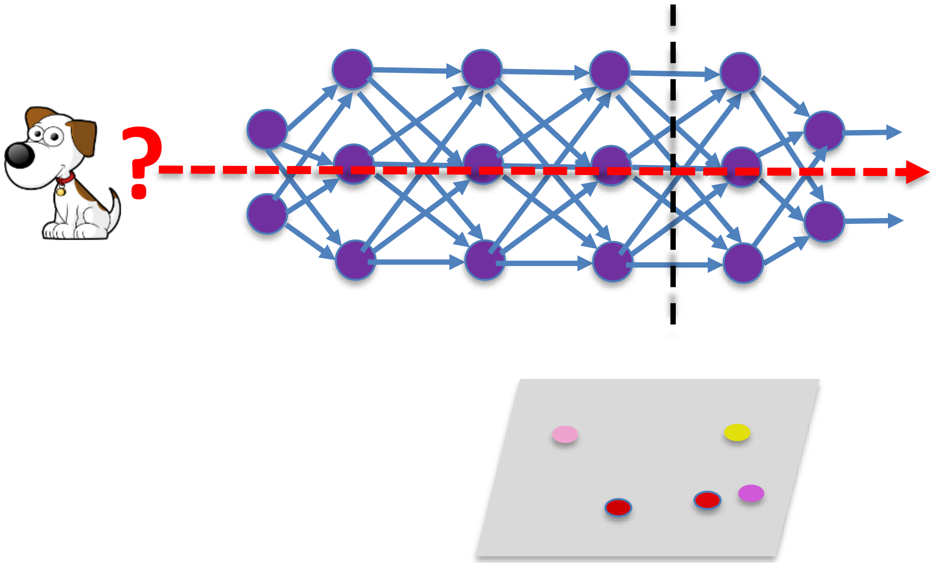
Personalized FL through Local Memorization: background





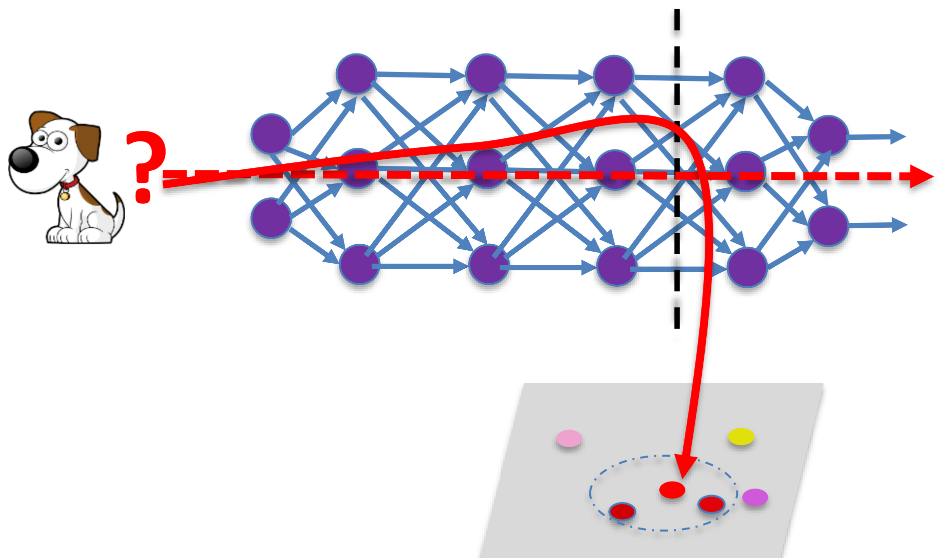
Orhan. A simple cache model for image recognition. NeurIPS, 2018.
Khandelwal, Levy, Jurafsky, Zettlemoyer, Lewis.
Generalization through Memorization: Nearest Neighbor Language Models. ICLR, 2020.

Personalized FL through Local Memorization: background



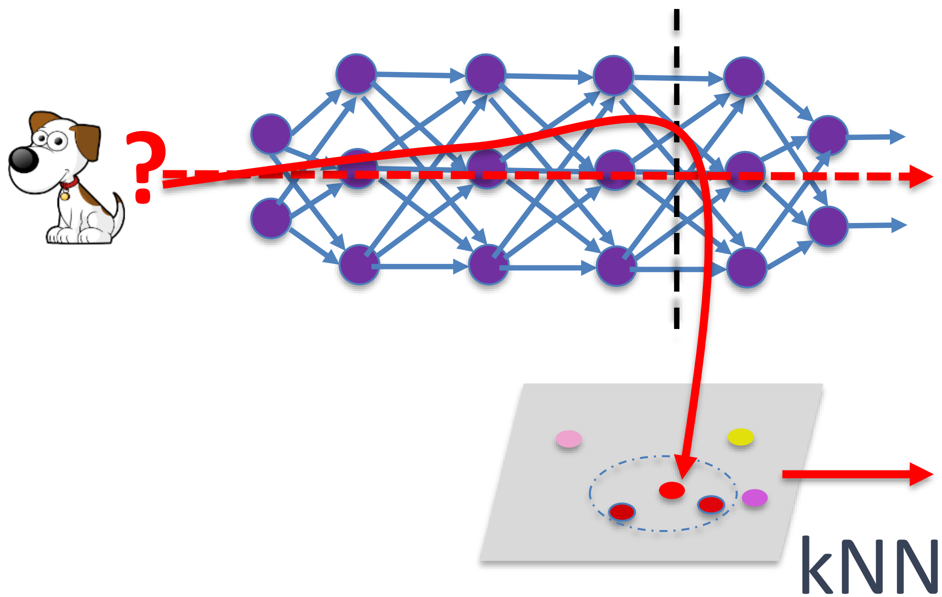
Orhan. A simple cache model for image recognition. NeurIPS, 2018.
Khandelwal, Levy, Jurafsky, Zettlemoyer, Lewis.
Generalization through Memorization: Nearest Neighbor Language Models. ICLR, 2020.

Personalized FL through Local Memorization: background



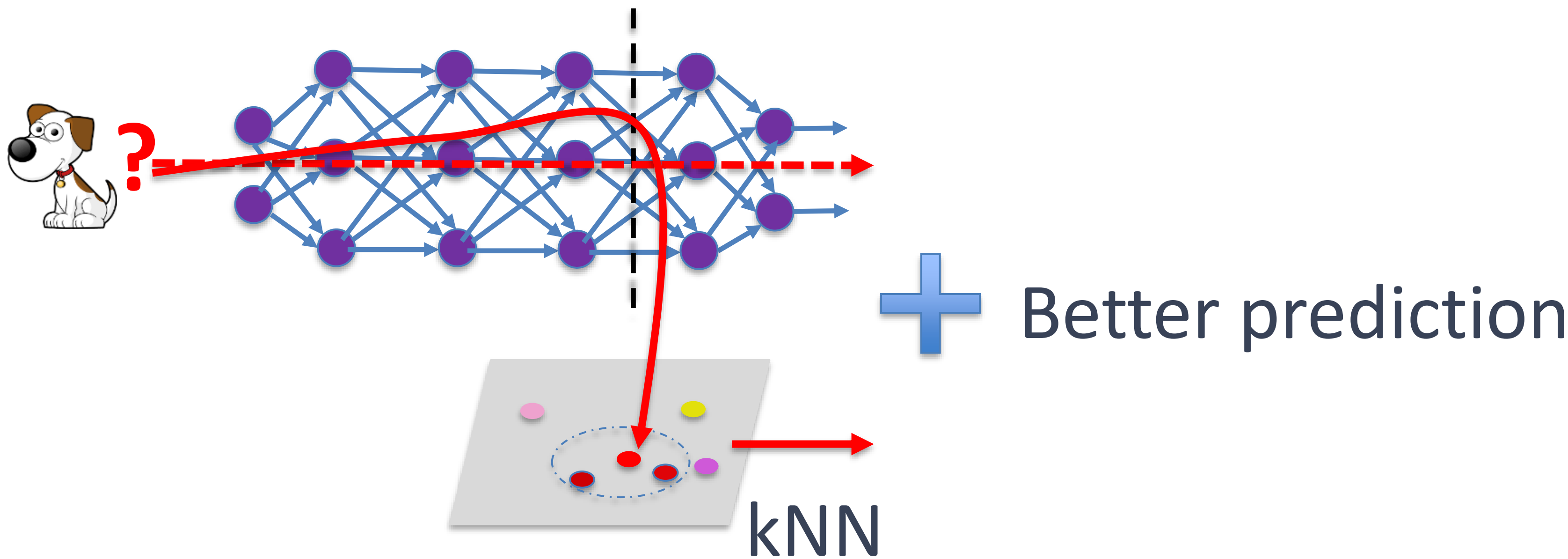
Orhan. A simple cache model for image recognition. NeurIPS, 2018.
Khandelwal, Levy, Jurafsky, Zettlemoyer, Lewis.
Generalization through Memorization: Nearest Neighbor Language Models. ICLR, 2020.

Personalized FL through Local Memorization: background



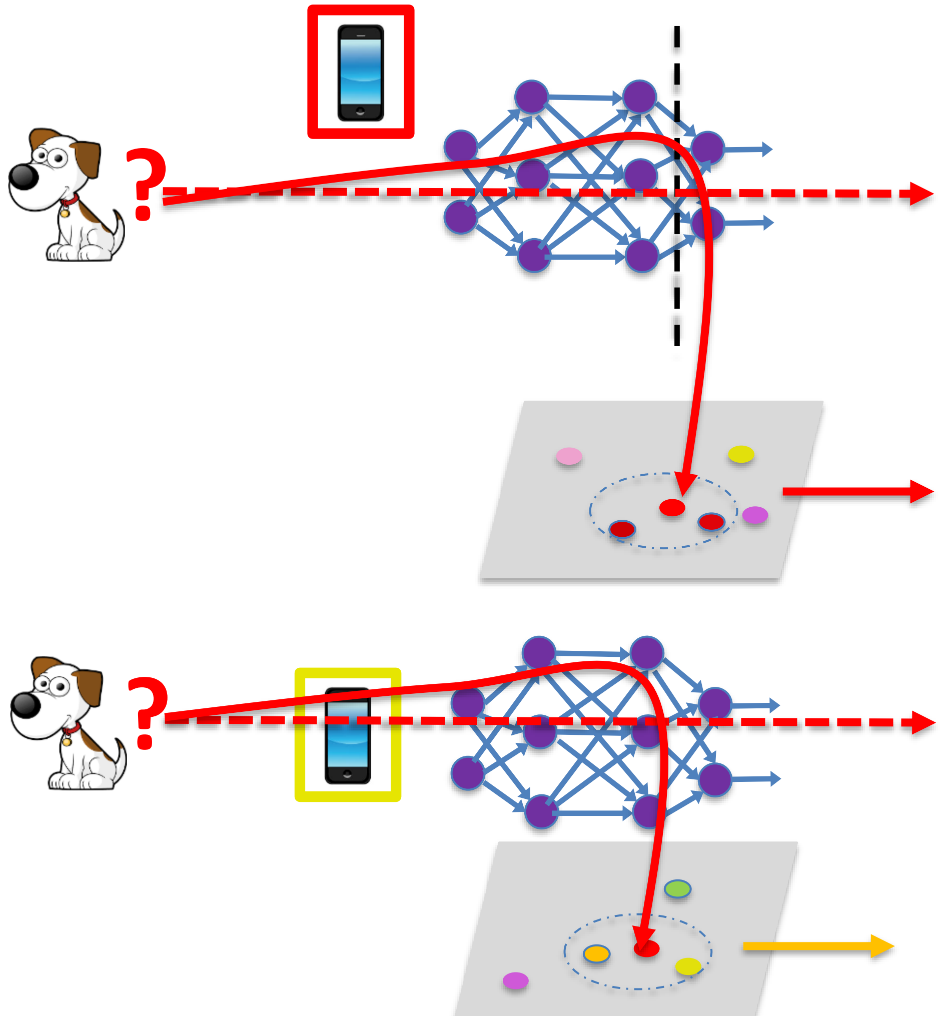
Orhan. A simple cache model for image recognition. NeurIPS, 2018.
Khandelwal, Levy, Jurafsky, Zettlemoyer, Lewis.
Generalization through Memorization: Nearest Neighbor Language Models. ICLR, 2020.

Personalized FL through Local Memorization: background

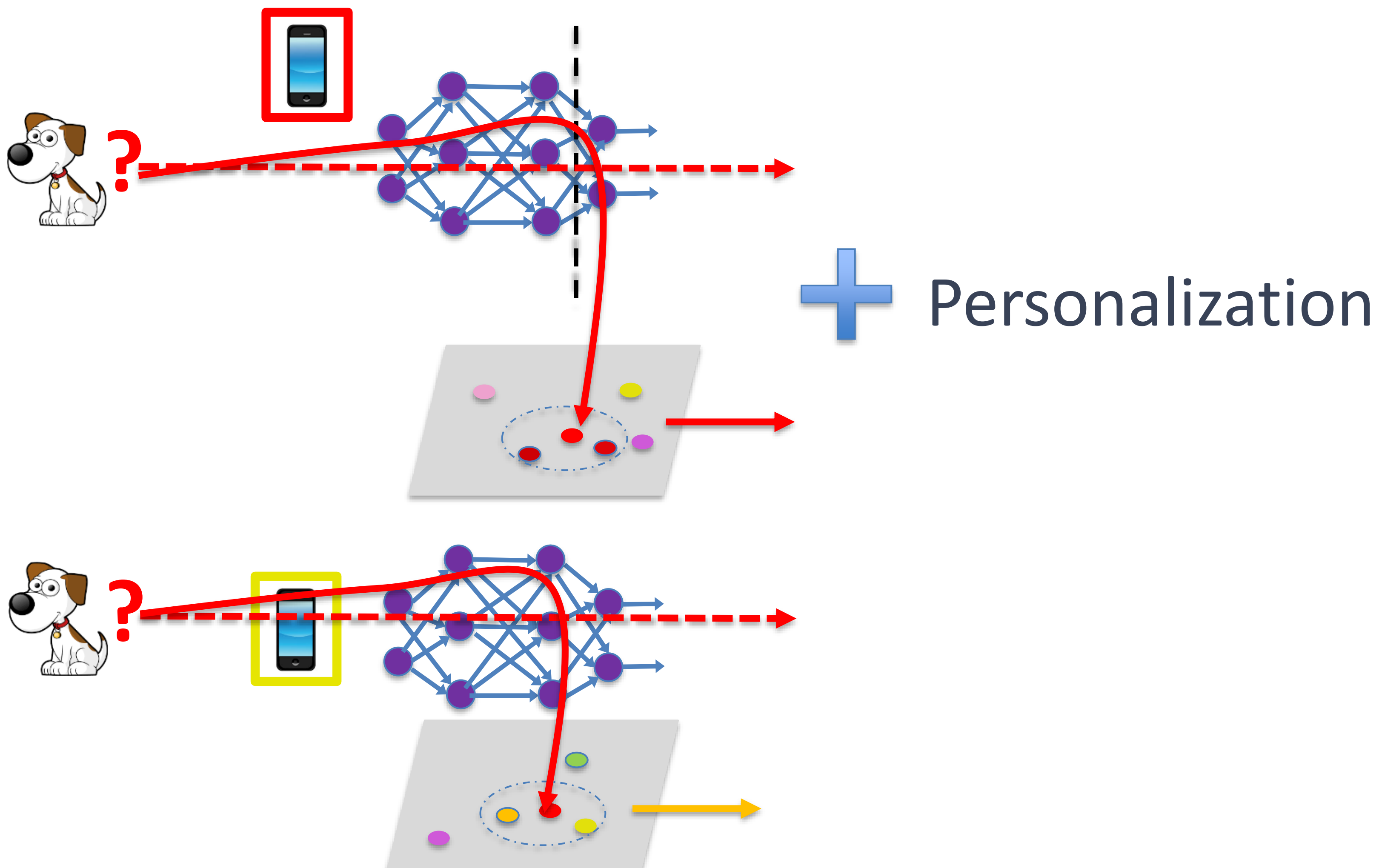


Orhan. A simple cache model for image recognition. NeurIPS, 2018.
Khandelwal, Levy, Jurafsky, Zettlemoyer, Lewis.
Generalization through Memorization: Nearest Neighbor Language Models. ICLR, 2020.

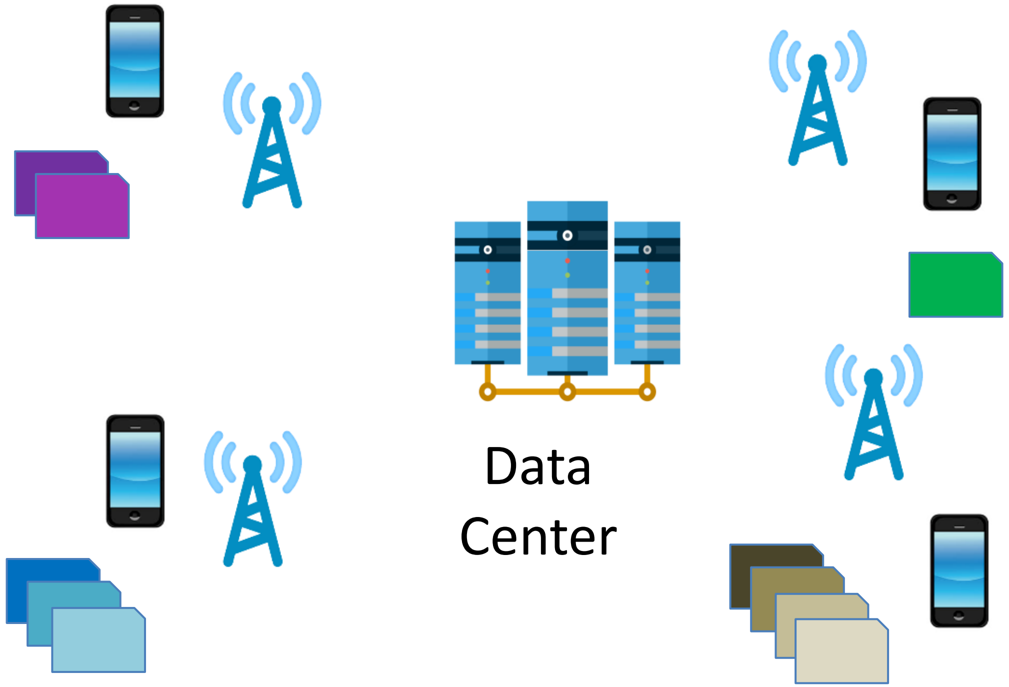
Our idea:
Datastore for Personalization



Our idea:
Datastore for Personalization



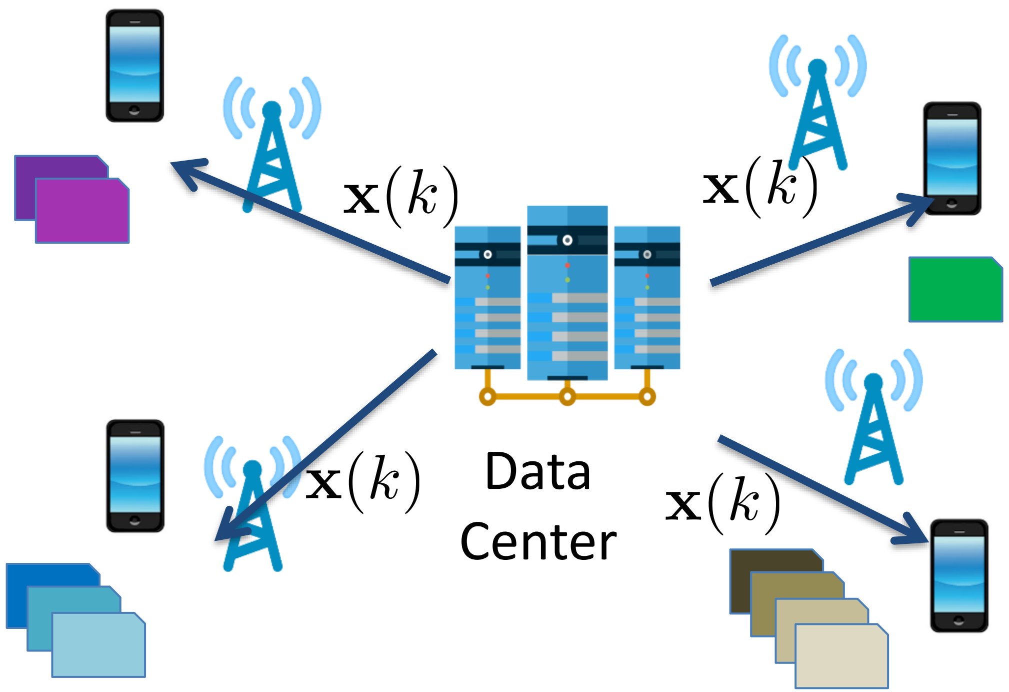
kNN-Per



kNN-Per

1. Clients train a global model using a federated learning algorithm (e.g. FedAvg)

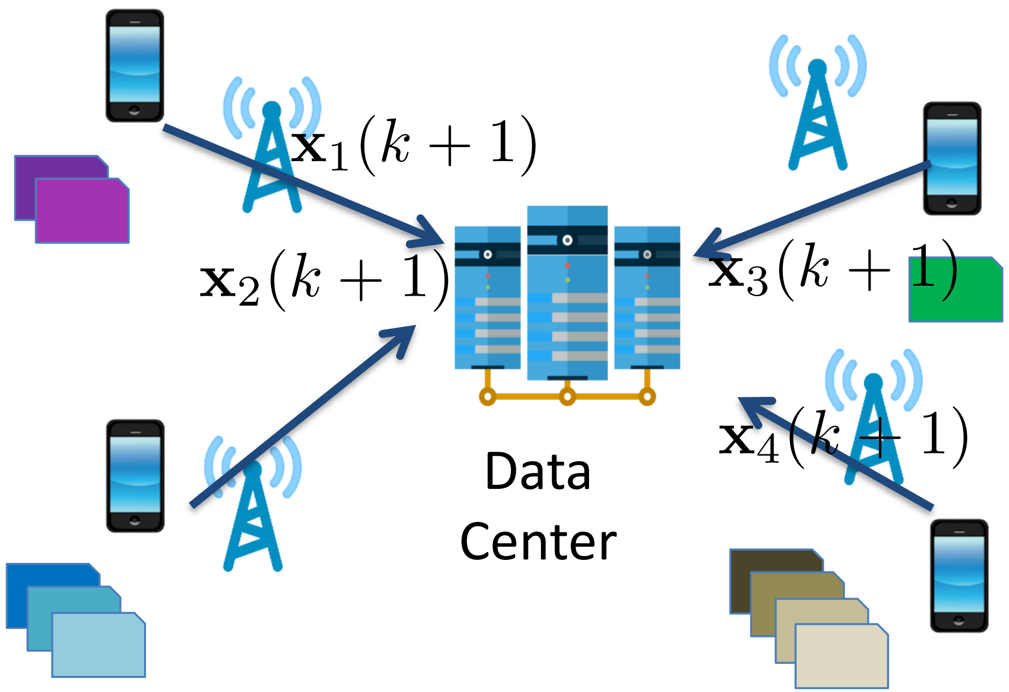
kNN-Per



kNN-Per

1. Clients train a global model using a federated learning algorithm (e.g. FedAvg)

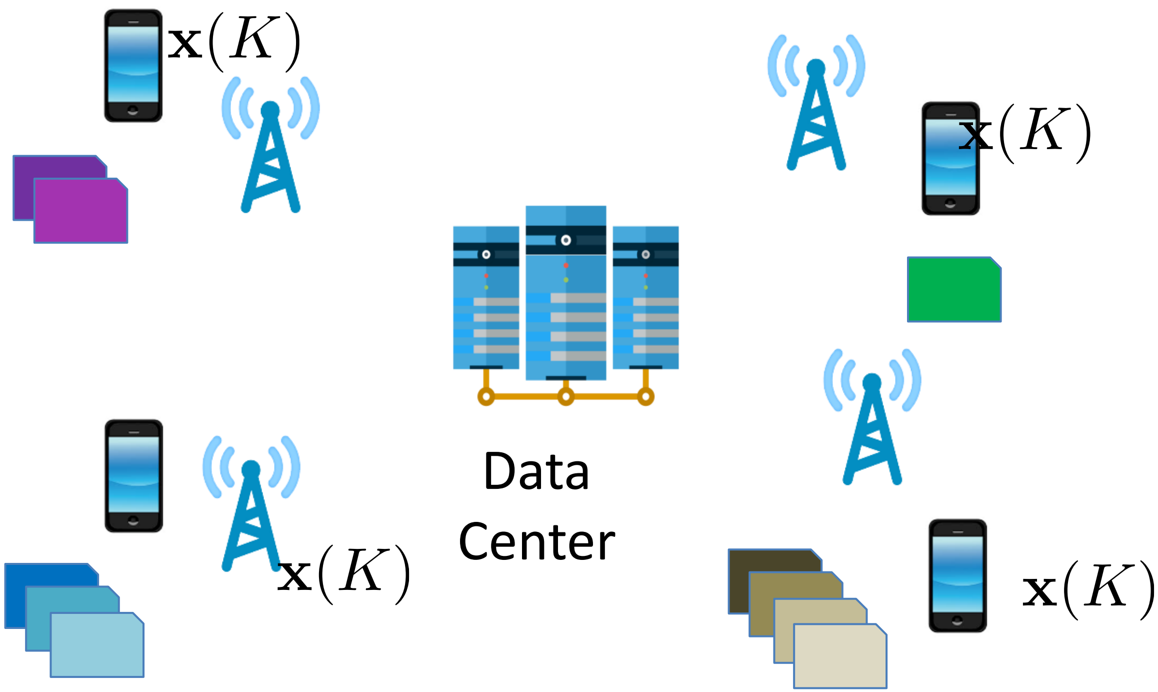
kNN-Per



kNN-Per

1. Clients train a global model using a federated learning algorithm (e.g. FedAvg)

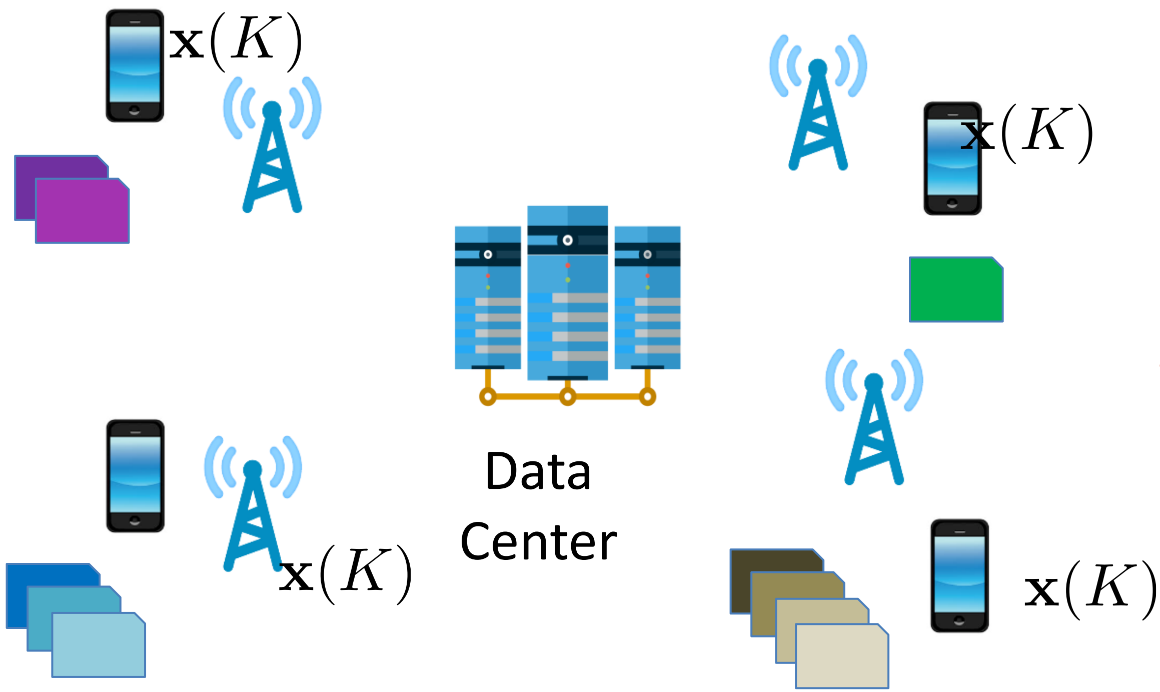
kNN-Per



kNN-Per

1. Clients train a global model using a federated learning algorithm (e.g. FedAvg)

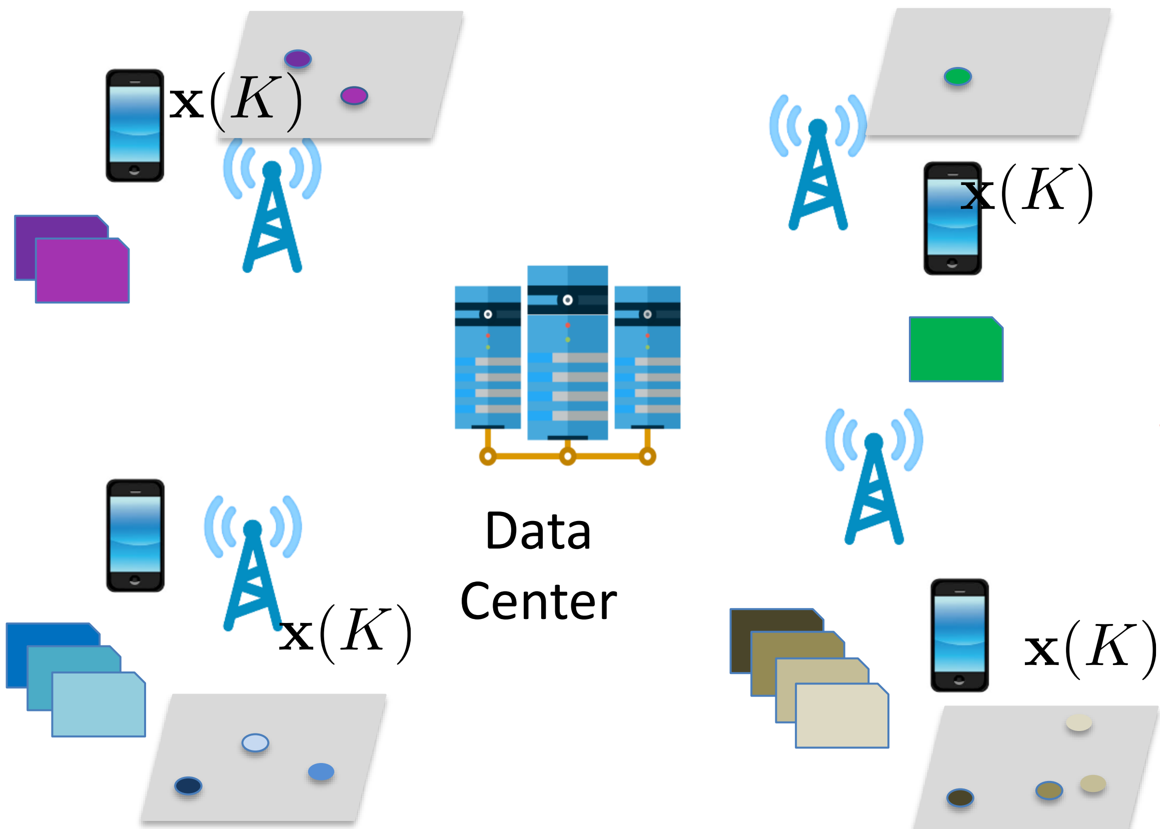
kNN-Per



kNN-Per

1. Clients train a global model using a federated learning algorithm (e.g. FedAvg)
2. Each client creates its local datastore

kNN-Per



kNN-Per

1. Clients train a global model using a federated learning algorithm (e.g. FedAvg)
2. Each client creates its local datastore

kNN-Per

kNN-Per

1. Clients train a global model using a federated learning algorithm (e.g. FedAvg)
2. Each client creates its local datastore
3. A linear interpolation is used at inference

$$(1-\lambda)h_{\text{glob}}(\mathbf{x}(K),\chi)+\lambda h_{i,k\text{NN}}(\mathbf{x}(K),\chi)$$

kNN-Per

- Enjoys global model's convergence properties

kNN-Per

- Enjoys global model's convergence properties

- What about generalization properties?

kNN-Per

- Enjoys global model's convergence properties

- What about generalization properties?

$$\mathop{\mathbb{E}}_{\mathcal{S}\sim\otimes_{m=1}^{M}\mathcal{D}_m^{n_m}}\left[\mathcal{L}_{\mathcal{D}_m}\left(h_{m,\lambda}\right)\right] \leq (1+\lambda)\cdot\mathcal{L}_{\mathcal{D}_m}\left(h_m^{*}\right)$$
$$+ c_1\left(1-\lambda\right)\cdot \underset{\mathcal{H}}{\texttt{disc}}\left(\bar{\mathcal{D}},\mathcal{D}_m\right) + c_3\left(1-\lambda\right)\cdot\sqrt{\frac{d}{n}}\cdot\sqrt{c_4+\log\left(\frac{n}{d}\right)}$$
$$+ c_2\lambda\cdot\frac{\sqrt{p}}{\sqrt[p+1]{n_m}}\cdot \underset{\mathcal{H}}{\texttt{disc}}\left(\bar{\mathcal{D}},\mathcal{D}_m\right) + c_5\lambda\cdot\sqrt{\frac{d}{n}}\cdot\sqrt{c_4+\log\left(\frac{n}{d}\right)}\cdot\frac{\sqrt{p}}{\sqrt[p+1]{n_m}}$$

kNN-Per

Assumption. *Let* $h_m^* \in \arg\min_{h\in\mathcal{H}} \mathcal{L}_{D_m}(h)$*. There exist constants* $\gamma_1, \gamma_2 > 0$*, such that for any dataset* $\mathcal{S}$ *drawn from* $\mathcal{X} \times \mathcal{Y}$ *and any data points* $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$*, we have*

$$\left|\eta_m(\mathbf{x}) - \eta_m(\mathbf{x}')\right| \leq d\left(\phi_{h_\mathcal{S}}(\mathbf{x}), \phi_{h_\mathcal{S}}(\mathbf{x}')\right) \times \left(\gamma_1 + \gamma_2\left(\mathcal{L}_{\mathcal{D}_m}(h_\mathcal{S}) - \mathcal{L}_{\mathcal{D}_m}(h_m^*)\right)\right).$$

kNN-Per

Assumption. *Let* $h_m^* \in \arg\min_{h \in \mathcal{H}} \mathcal{L}_{D_m}(h)$. *There exist constants* $\gamma_1, \gamma_2 > 0$, *such that for any dataset* $\mathcal{S}$ *drawn from* $\mathcal{X} \times \mathcal{Y}$ *and any data points* $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$, *we have*

$$\left|\eta_m(\mathbf{x}) - \eta_m(\mathbf{x}')\right| \leq d\left(\phi_{h_\mathcal{S}}(\mathbf{x}), \phi_{h_\mathcal{S}}(\mathbf{x}')\right) \times \left(\gamma_1 + \gamma_2\left(\mathcal{L}_{\mathcal{D}_m}(h_\mathcal{S}) - \mathcal{L}_{\mathcal{D}_m}(h_m^*)\right)\right).$$

x & x'
in same class?

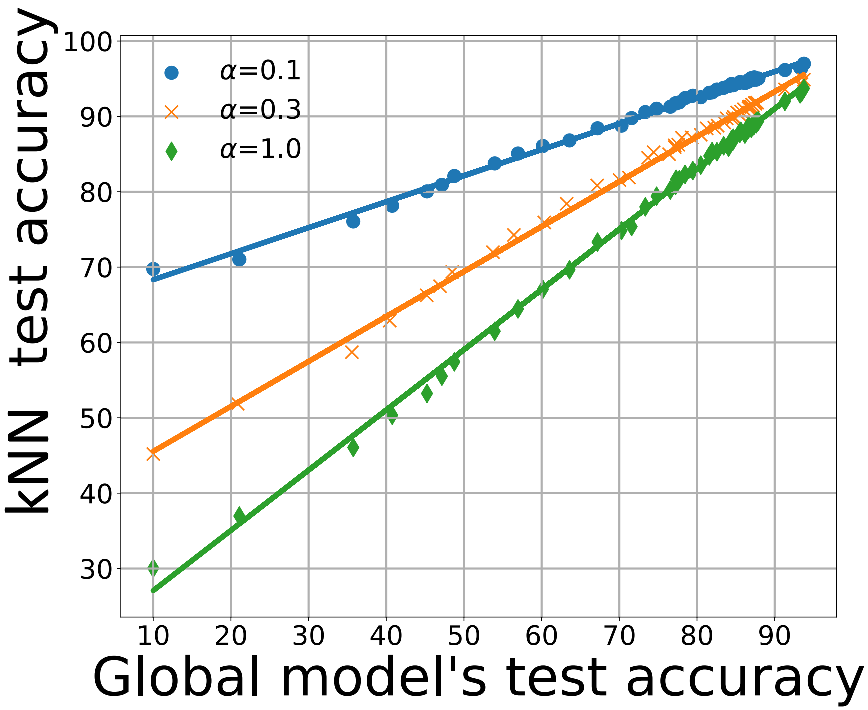
representations'
distance

global model's
quality for client m

kNN-Per

Assumption. *Let* $h_m^* \in \arg\min_{h\in\mathcal{H}} \mathcal{L}_{D_m}(h)$*. There exist constants* $\gamma_1, \gamma_2 > 0$*, such that for any dataset* $\mathcal{S}$ *drawn from* $\mathcal{X} \times \mathcal{Y}$ *and any data points* $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$*, we have*

$$\left|\eta_m(\mathbf{x}) - \eta_m(\mathbf{x}')\right| \leq d\left(\phi_{h_\mathcal{S}}(\mathbf{x}), \phi_{h_\mathcal{S}}(\mathbf{x}')\right) \times \left(\gamma_1 + \gamma_2\left(\mathcal{L}_{\mathcal{D}_m}(h_\mathcal{S}) - \mathcal{L}_{\mathcal{D}_m}(h_m^*)\right)\right).$$



CIFAR-10

kNN-Per

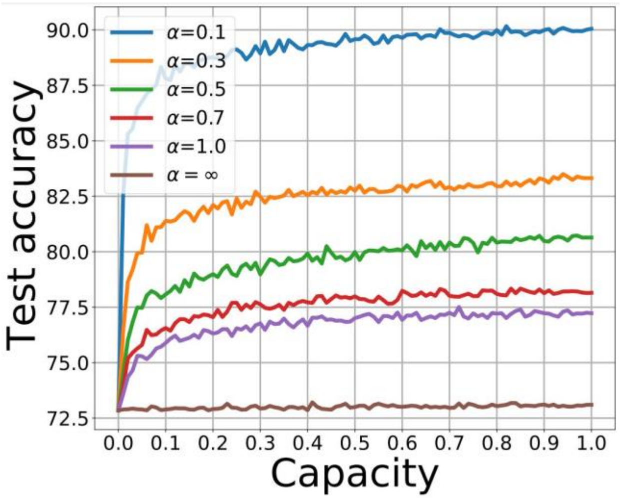
Table 2: Test accuracy: average across clients / bottom decile.

Dataset	Local	FedAvg	FedAvg+	ClusteredFL	Ditto	FedRep	APFL	kNN-Per (Ours)
FEMNIST	71.0 / 57.5	83.4 / 68.9	84.3 / 69.4	83.7 / 69.4	84.3 / 71.3	85.3 / 72.7	84.1 / 69.4	**88.2** / **78.8**
CIFAR-10	57.6 / 41.1	72.8 / 59.6	75.2 / 62.3	73.3 / 61.5	80.0 / 66.5	77.7 / 65.2	78.9 / 68.1	**83.0** / **71.4**
CIFAR-100	31.5 / 19.8	47.4 / 36.0	51.4 / 41.1	47.2 / 36.2	52.0 / 41.4	53.2 / 41.7	51.7 / 41.1	**55.0** / **43.6**
Shakespeare	32.0 / 16.0	48.1 / 43.1	47.0 / 42.2	46.7 / 41.4	47.9 / 42.6	47.2 / 42.3	45.9 / 42.4	**51.4** / **45.4**

kNN-Per

Table 2: Test accuracy: average across clients / bottom decile.

Dataset	Local	FedAvg	FedAvg+	ClusteredFL	Ditto	FedRep	APFL	kNN-Per (Ours)
FEMNIST	71.0 / 57.5	83.4 / 68.9	84.3 / 69.4	83.7 / 69.4	84.3 / 71.3	85.3 / 72.7	84.1 / 69.4	**88.2** / **78.8**
CIFAR-10	57.6 / 41.1	72.8 / 59.6	75.2 / 62.3	73.3 / 61.5	80.0 / 66.5	77.7 / 65.2	78.9 / 68.1	**83.0** / **71.4**
CIFAR-100	31.5 / 19.8	47.4 / 36.0	51.4 / 41.1	47.2 / 36.2	52.0 / 41.4	53.2 / 41.7	51.7 / 41.1	**55.0** / **43.6**
Shakespeare	32.0 / 16.0	48.1 / 43.1	47.0 / 42.2	46.7 / 41.4	47.9 / 42.6	47.2 / 42.3	45.9 / 42.4	**51.4** / **45.4**

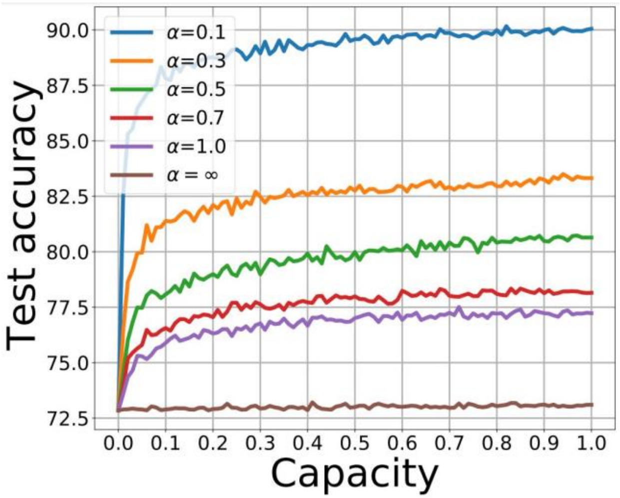


CIFAR-10

kNN-Per

Table 2: Test accuracy: average across clients / bottom decile.

Dataset	Local	FedAvg	FedAvg+	ClusteredFL	Ditto	FedRep	APFL	kNN-Per (Ours)
FEMNIST	71.0 / 57.5	83.4 / 68.9	84.3 / 69.4	83.7 / 69.4	84.3 / 71.3	85.3 / 72.7	84.1 / 69.4	**88.2** / **78.8**
CIFAR-10	57.6 / 41.1	72.8 / 59.6	75.2 / 62.3	73.3 / 61.5	80.0 / 66.5	77.7 / 65.2	78.9 / 68.1	**83.0** / **71.4**
CIFAR-100	31.5 / 19.8	47.4 / 36.0	51.4 / 41.1	47.2 / 36.2	52.0 / 41.4	53.2 / 41.7	51.7 / 41.1	**55.0** / **43.6**
Shakespeare	32.0 / 16.0	48.1 / 43.1	47.0 / 42.2	46.7 / 41.4	47.9 / 42.6	47.2 / 42.3	45.9 / 42.4	**51.4** / **45.4**



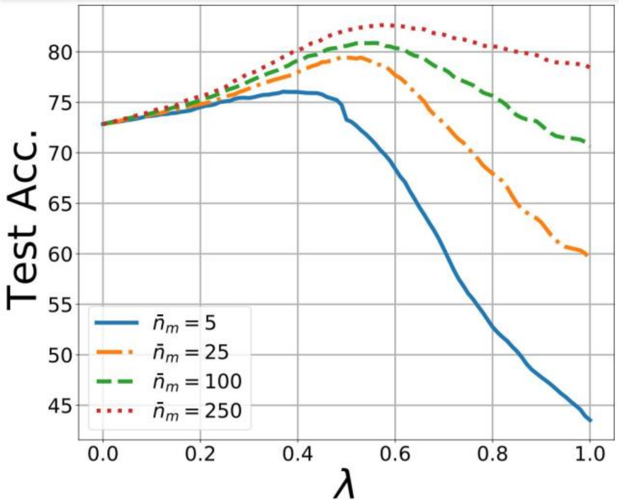
The benefit of kNN-Per is larger when data distributions are more heterogenous

CIFAR-10

kNN-Per

Table 2: Test accuracy: average across clients / bottom decile.

Dataset	Local	FedAvg	FedAvg+	ClusteredFL	Ditto	FedRep	APFL	kNN-Per (Ours)
FEMNIST	71.0 / 57.5	83.4 / 68.9	84.3 / 69.4	83.7 / 69.4	84.3 / 71.3	85.3 / 72.7	84.1 / 69.4	**88.2 / 78.8**
CIFAR-10	57.6 / 41.1	72.8 / 59.6	75.2 / 62.3	73.3 / 61.5	80.0 / 66.5	77.7 / 65.2	78.9 / 68.1	**83.0 / 71.4**
CIFAR-100	31.5 / 19.8	47.4 / 36.0	51.4 / 41.1	47.2 / 36.2	52.0 / 41.4	53.2 / 41.7	51.7 / 41.1	**55.0 / 43.6**
Shakespeare	32.0 / 16.0	48.1 / 43.1	47.0 / 42.2	46.7 / 41.4	47.9 / 42.6	47.2 / 42.3	45.9 / 42.4	**51.4 / 45.4**

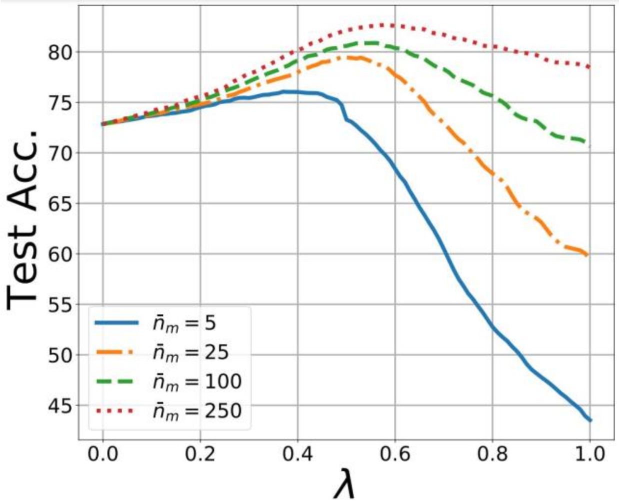


CIFAR-10

kNN-Per

Table 2: Test accuracy: average across clients / bottom decile.

Dataset	Local	FedAvg	FedAvg+	ClusteredFL	Ditto	FedRep	APFL	kNN-Per (Ours)
FEMNIST	71.0 / 57.5	83.4 / 68.9	84.3 / 69.4	83.7 / 69.4	84.3 / 71.3	85.3 / 72.7	84.1 / 69.4	**88.2** / **78.8**
CIFAR-10	57.6 / 41.1	72.8 / 59.6	75.2 / 62.3	73.3 / 61.5	80.0 / 66.5	77.7 / 65.2	78.9 / 68.1	**83.0** / **71.4**
CIFAR-100	31.5 / 19.8	47.4 / 36.0	51.4 / 41.1	47.2 / 36.2	52.0 / 41.4	53.2 / 41.7	51.7 / 41.1	**55.0** / **43.6**
Shakespeare	32.0 / 16.0	48.1 / 43.1	47.0 / 42.2	46.7 / 41.4	47.9 / 42.6	47.2 / 42.3	45.9 / 42.4	**51.4** / **45.4**



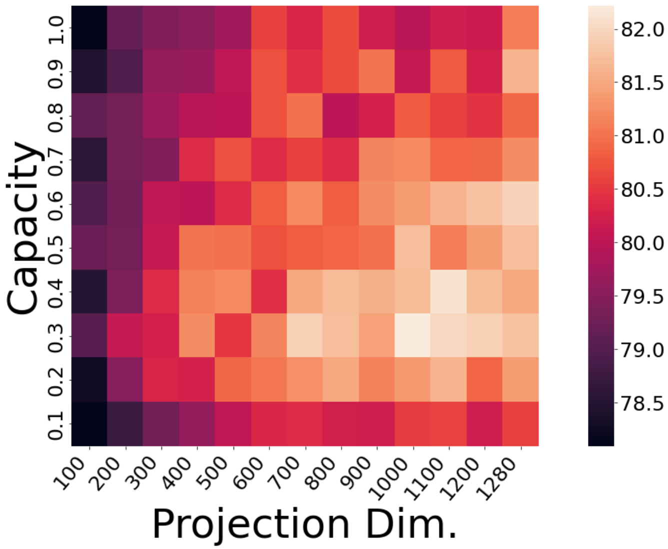
kNN-Per relies mostly on kNN for datasets with more than 100 samples

CIFAR-10

kNN-Per

Table 2: Test accuracy: average across clients / bottom decile.

Dataset	Local	FedAvg	FedAvg+	ClusteredFL	Ditto	FedRep	APFL	kNN-Per (Ours)
FEMNIST	71.0 / 57.5	83.4 / 68.9	84.3 / 69.4	83.7 / 69.4	84.3 / 71.3	85.3 / 72.7	84.1 / 69.4	**88.2** / **78.8**
CIFAR-10	57.6 / 41.1	72.8 / 59.6	75.2 / 62.3	73.3 / 61.5	80.0 / 66.5	77.7 / 65.2	78.9 / 68.1	**83.0** / **71.4**
CIFAR-100	31.5 / 19.8	47.4 / 36.0	51.4 / 41.1	47.2 / 36.2	52.0 / 41.4	53.2 / 41.7	51.7 / 41.1	**55.0** / **43.6**
Shakespeare	32.0 / 16.0	48.1 / 43.1	47.0 / 42.2	46.7 / 41.4	47.9 / 42.6	47.2 / 42.3	45.9 / 42.4	**51.4** / **45.4**



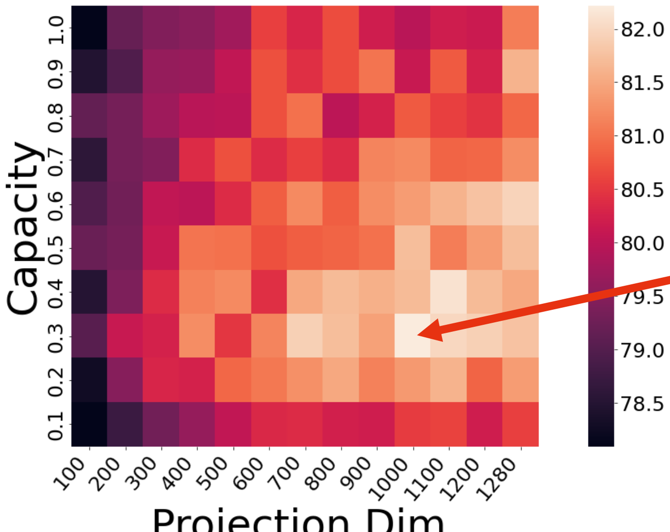
ProtoNN-like datastore compression

CIFAR-10

kNN-Per

Table 2: Test accuracy: average across clients / bottom decile.

Dataset	Local	FedAvg	FedAvg+	ClusteredFL	Ditto	FedRep	APFL	kNN-Per (Ours)
FEMNIST	71.0 / 57.5	83.4 / 68.9	84.3 / 69.4	83.7 / 69.4	84.3 / 71.3	85.3 / 72.7	84.1 / 69.4	**88.2** / **78.8**
CIFAR-10	57.6 / 41.1	72.8 / 59.6	75.2 / 62.3	73.3 / 61.5	80.0 / 66.5	77.7 / 65.2	78.9 / 68.1	**83.0** / **71.4**
CIFAR-100	31.5 / 19.8	47.4 / 36.0	51.4 / 41.1	47.2 / 36.2	52.0 / 41.4	53.2 / 41.7	51.7 / 41.1	**55.0** / **43.6**
Shakespeare	32.0 / 16.0	48.1 / 43.1	47.0 / 42.2	46.7 / 41.4	47.9 / 42.6	47.2 / 42.3	45.9 / 42.4	**51.4** / **45.4**



ProtoNN-like datastore compression

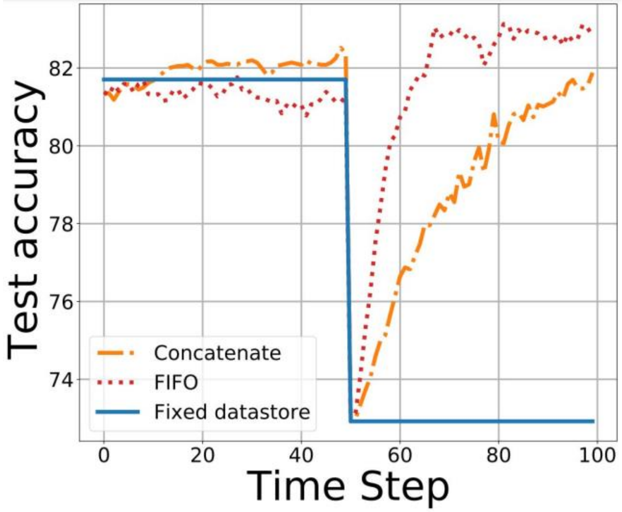
4x memory savings with limited accuracy loss (0.7pp)

CIFAR-10

kNN-Per

Table 2: Test accuracy: average across clients / bottom decile.

Dataset	Local	FedAvg	FedAvg+	ClusteredFL	Ditto	FedRep	APFL	kNN-Per (Ours)
FEMNIST	71.0 / 57.5	83.4 / 68.9	84.3 / 69.4	83.7 / 69.4	84.3 / 71.3	85.3 / 72.7	84.1 / 69.4	**88.2** / **78.8**
CIFAR-10	57.6 / 41.1	72.8 / 59.6	75.2 / 62.3	73.3 / 61.5	80.0 / 66.5	77.7 / 65.2	78.9 / 68.1	**83.0** / **71.4**
CIFAR-100	31.5 / 19.8	47.4 / 36.0	51.4 / 41.1	47.2 / 36.2	52.0 / 41.4	53.2 / 41.7	51.7 / 41.1	**55.0** / **43.6**
Shakespeare	32.0 / 16.0	48.1 / 43.1	47.0 / 42.2	46.7 / 41.4	47.9 / 42.6	47.2 / 42.3	45.9 / 42.4	**51.4** / **45.4**

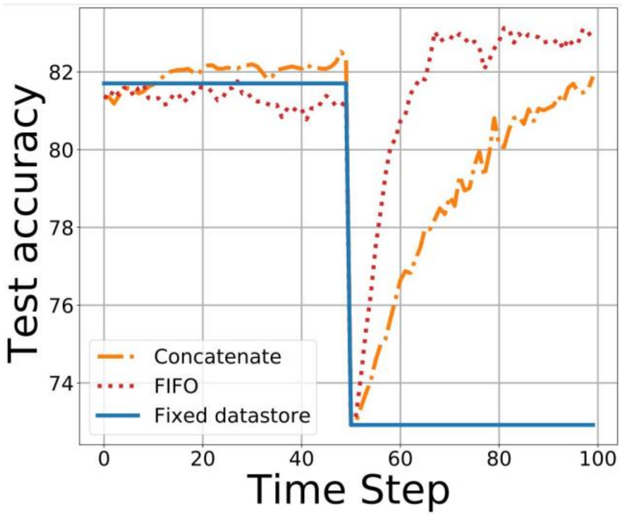


CIFAR-10

kNN-Per

Table 2: Test accuracy: average across clients / bottom decile.

Dataset	Local	FedAvg	FedAvg+	ClusteredFL	Ditto	FedRep	APFL	kNN-Per (Ours)
FEMNIST	71.0 / 57.5	83.4 / 68.9	84.3 / 69.4	83.7 / 69.4	84.3 / 71.3	85.3 / 72.7	84.1 / 69.4	**88.2** / **78.8**
CIFAR-10	57.6 / 41.1	72.8 / 59.6	75.2 / 62.3	73.3 / 61.5	80.0 / 66.5	77.7 / 65.2	78.9 / 68.1	**83.0** / **71.4**
CIFAR-100	31.5 / 19.8	47.4 / 36.0	51.4 / 41.1	47.2 / 36.2	52.0 / 41.4	53.2 / 41.7	51.7 / 41.1	**55.0** / **43.6**
Shakespeare	32.0 / 16.0	48.1 / 43.1	47.0 / 42.2	46.7 / 41.4	47.9 / 42.6	47.2 / 42.3	45.9 / 42.4	**51.4** / **45.4**



kNN-Per is robust to distribution shift

CIFAR-10

Questions?

Paper

Code