Federated Learning under Distributed Concept Drift

Ellango Jothimurugesan
Carnegie Mellon University

Joint work with Kevin Hsieh (Microsoft), Jianyu Wang (CMU), Gauri Joshi (CMU), Phillip B. Gibbons (CMU)

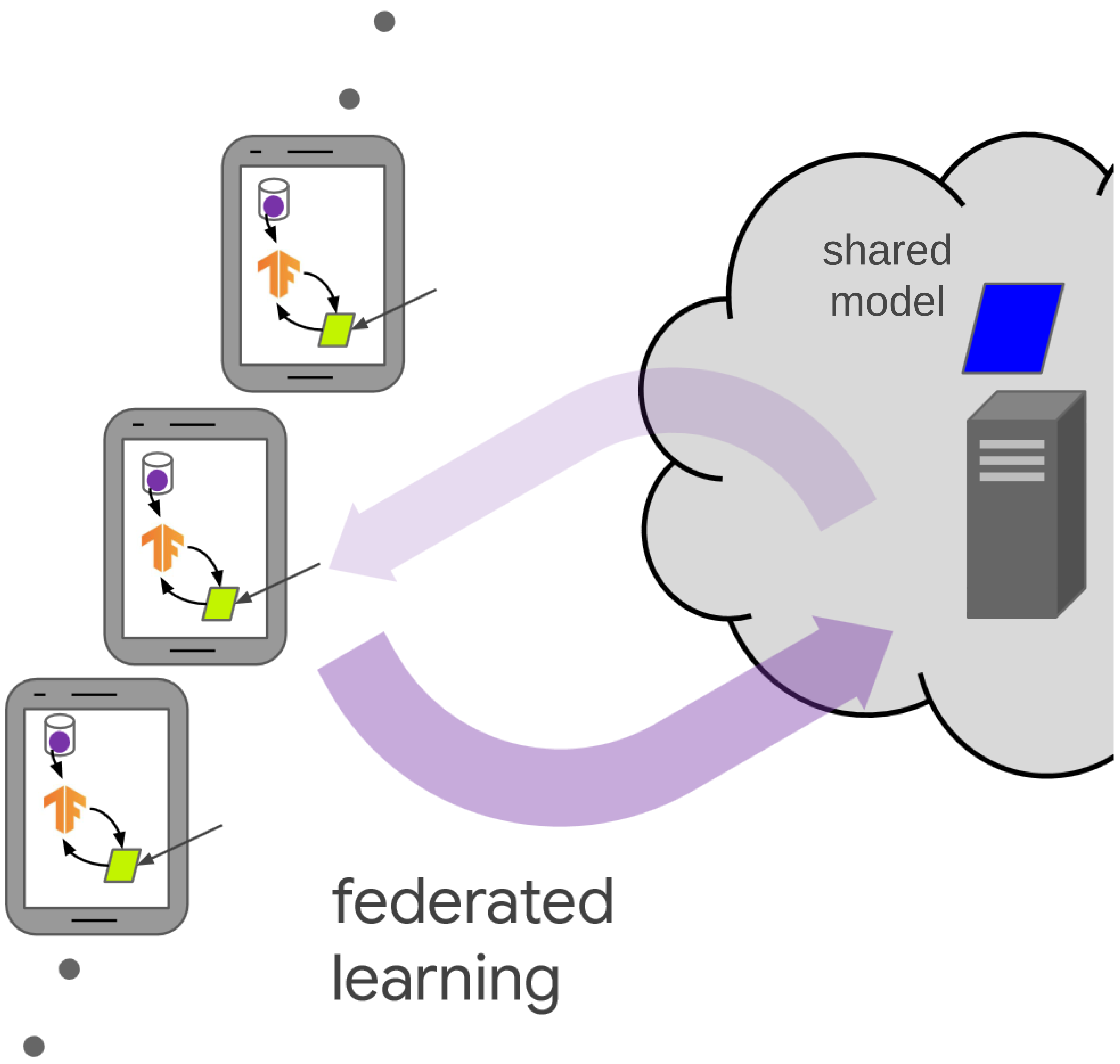
Federated Learning: Continual On-Device Training

Centralized learning:

- Train a model on all data, then deploy

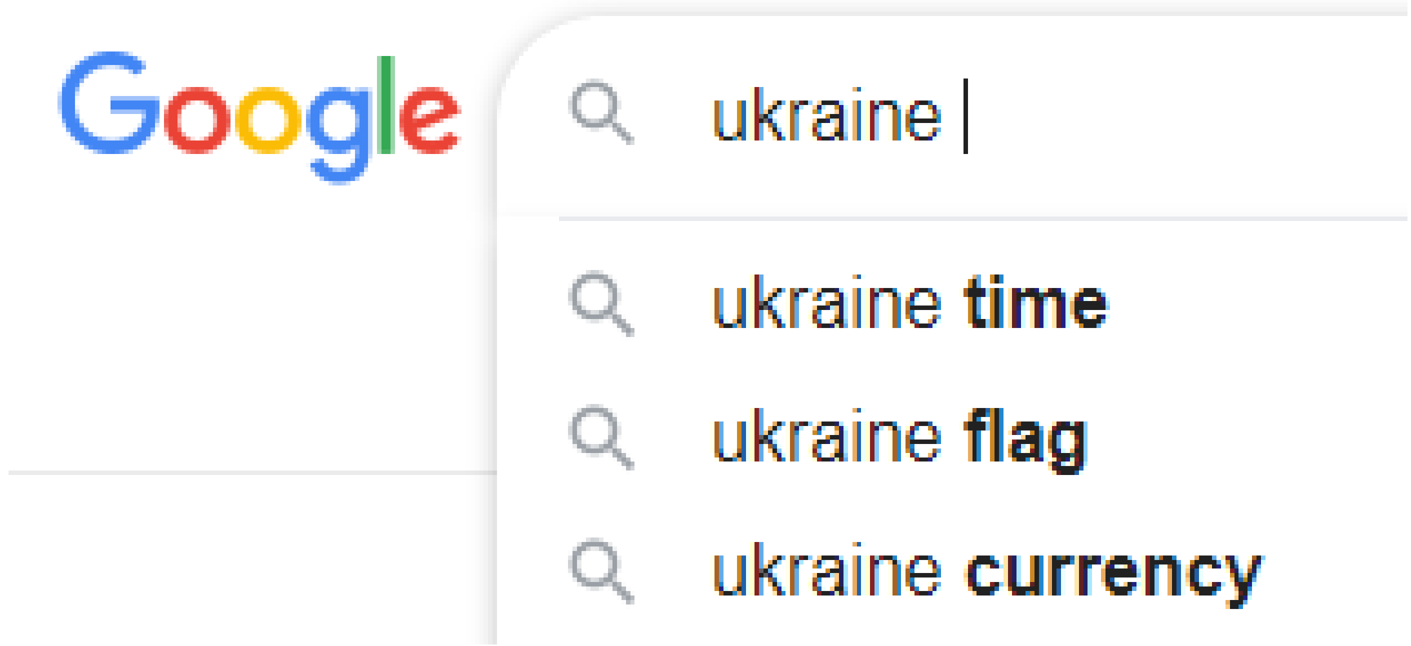
Federated Learning (FL):

- Clients continually compute updates to the model with new data
- Server continually aggregates updates

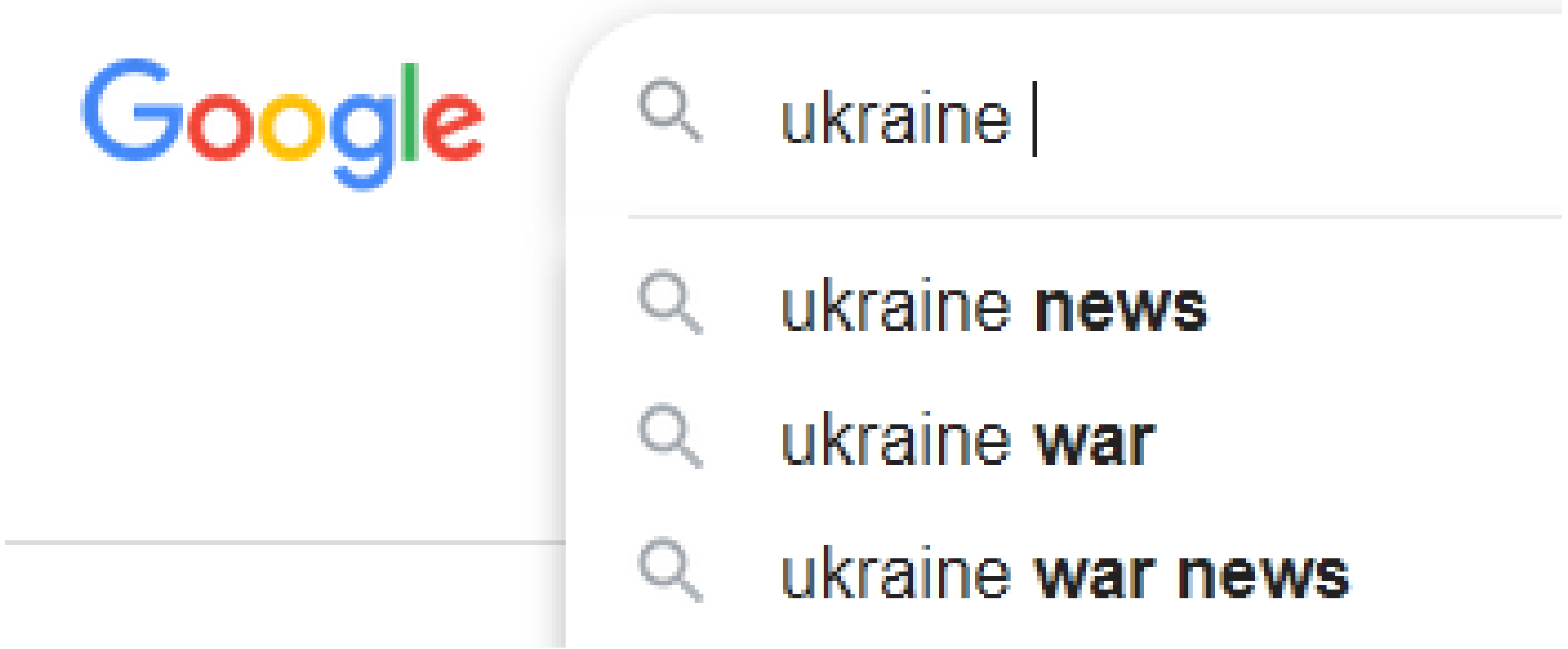


Concept Drift

- The data distribution (concept) can change over time
- Ex: next word prediction



Jan 2022

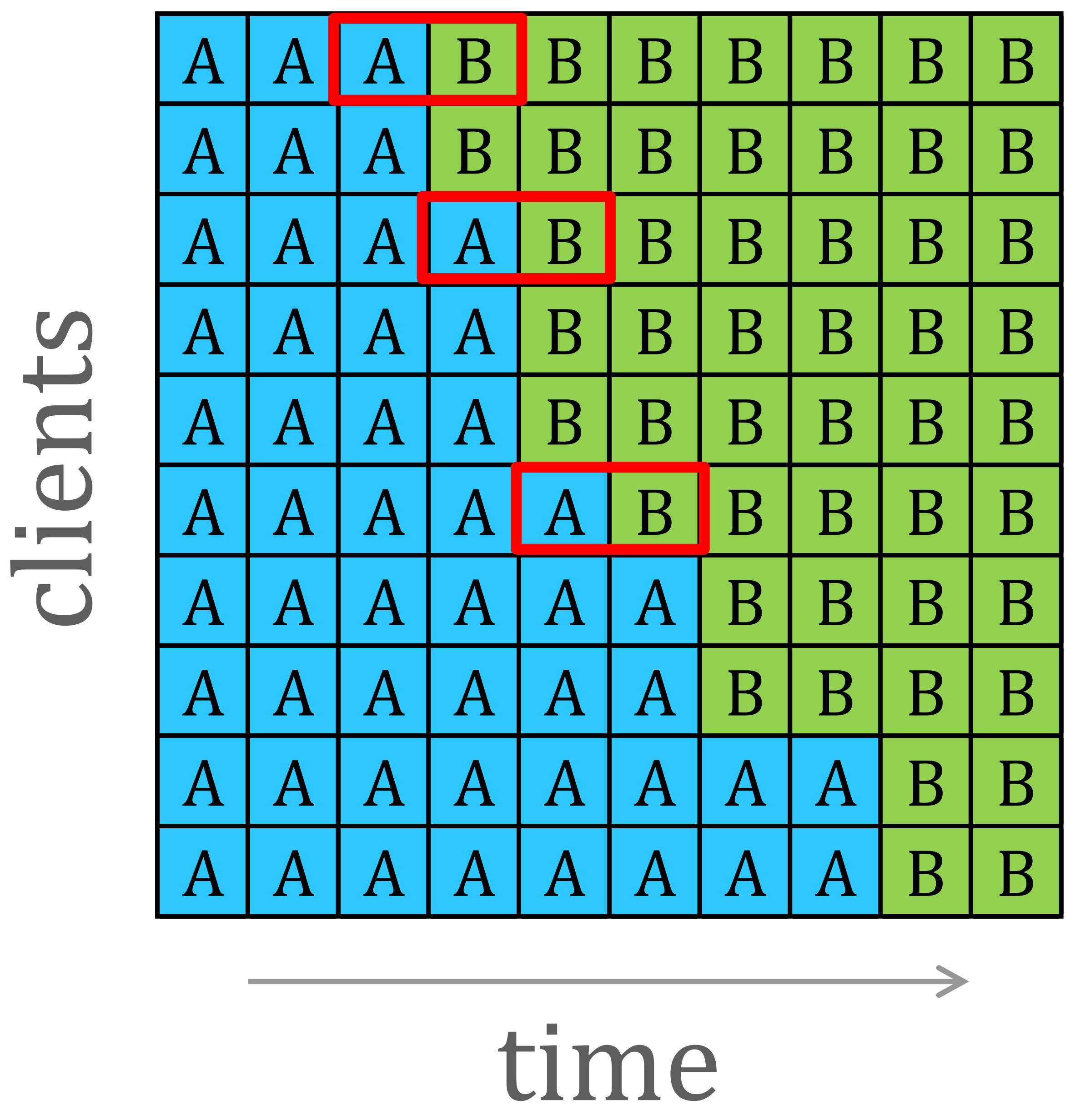


Feb 2022

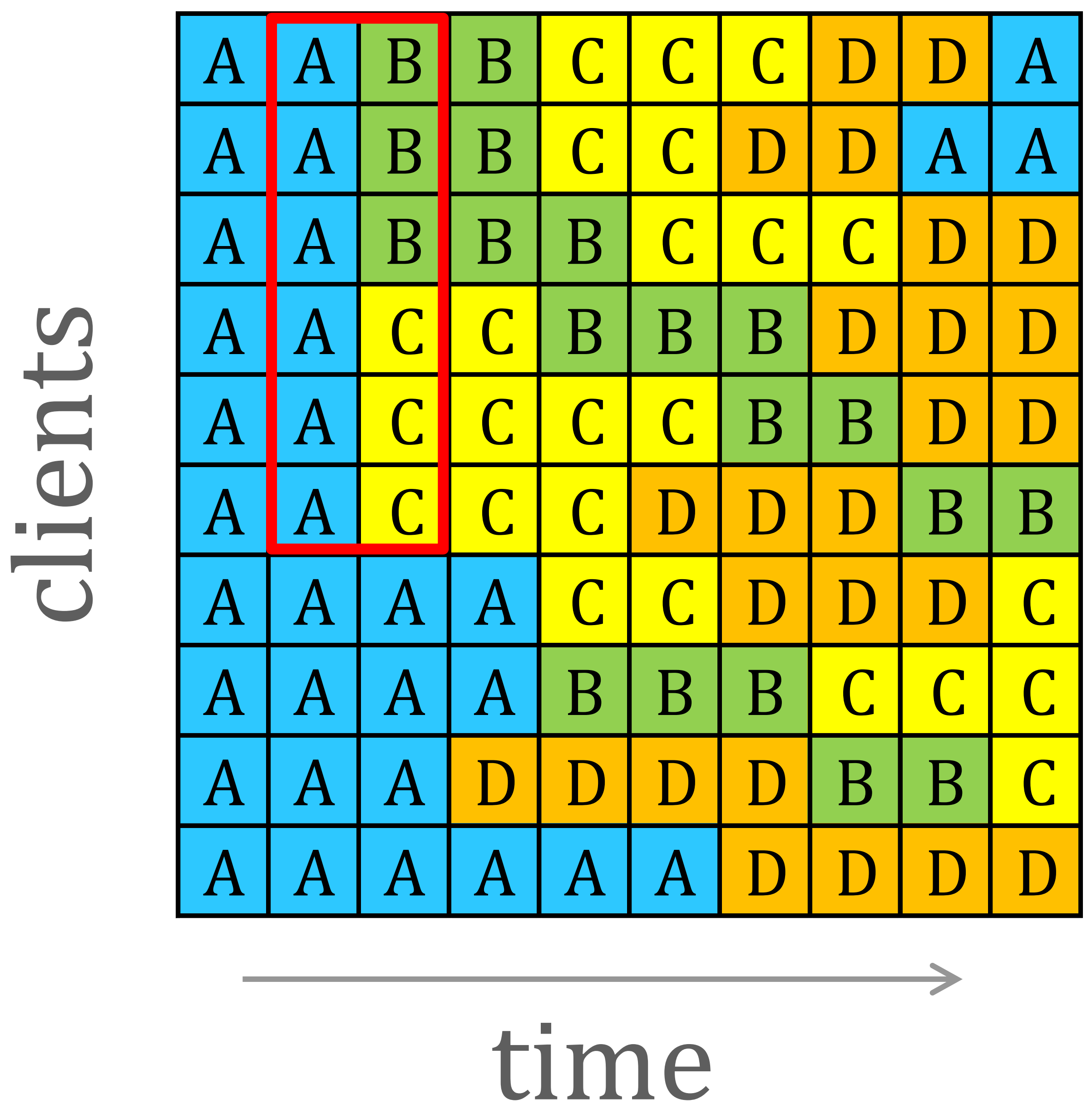
- No single model captures both concepts

Challenges of FL under Distributed Concept Drift

Drifts occur staggered in time across clients



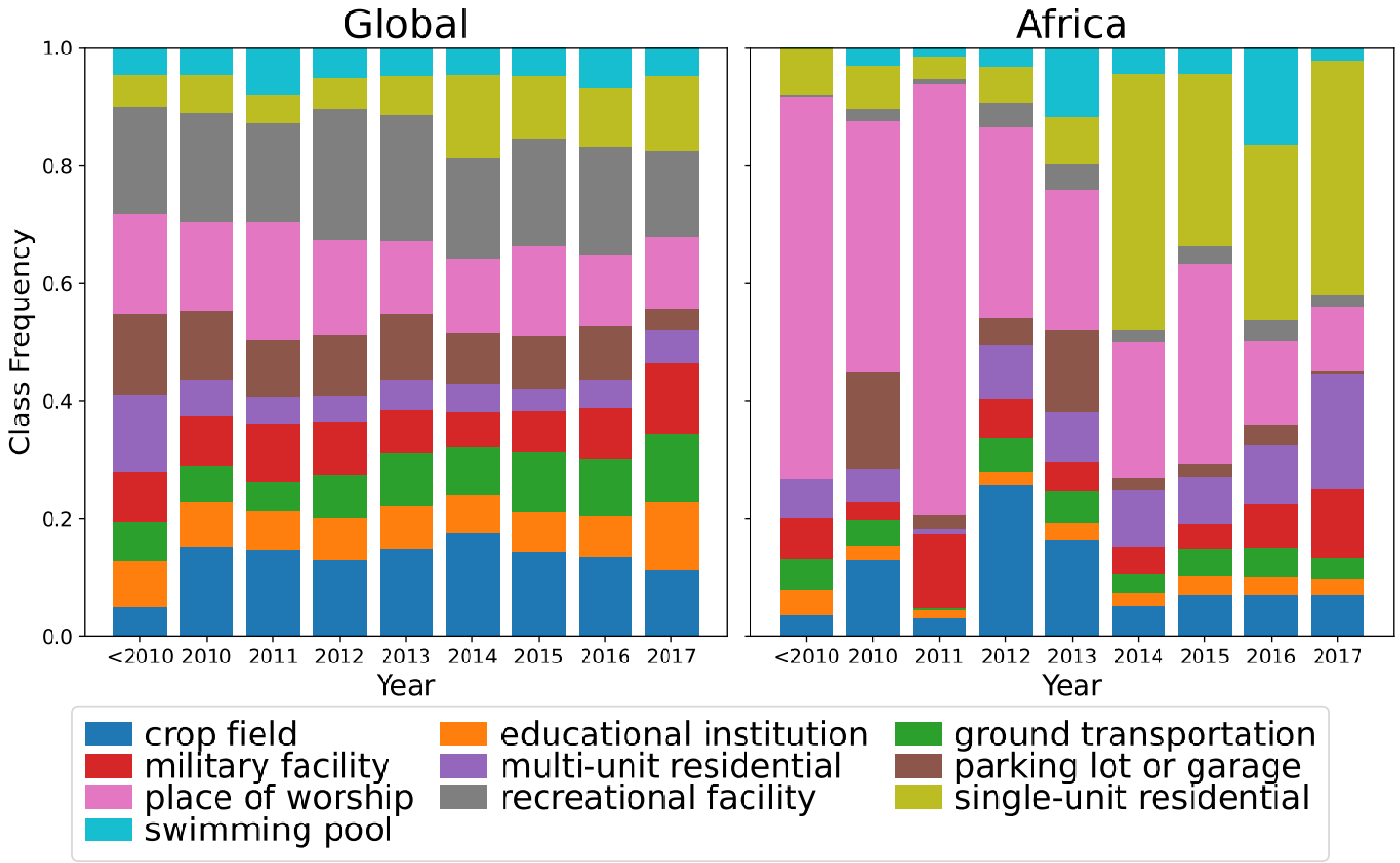
Multiple concepts may arise at the same time



Real-world Example: Localized Drift in FMoW

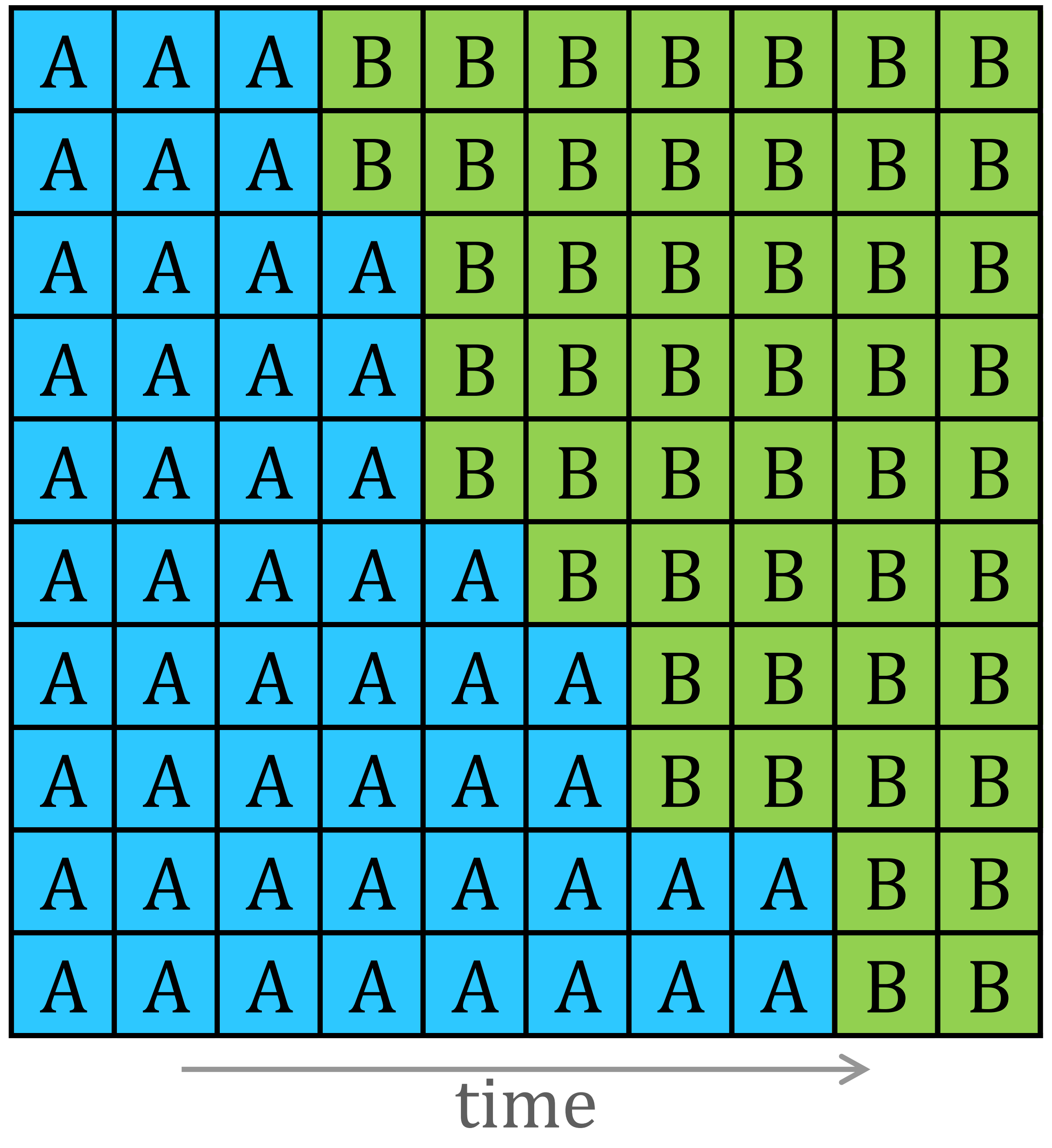
Functional Map of the World (FMoW): identify building type from satellite images

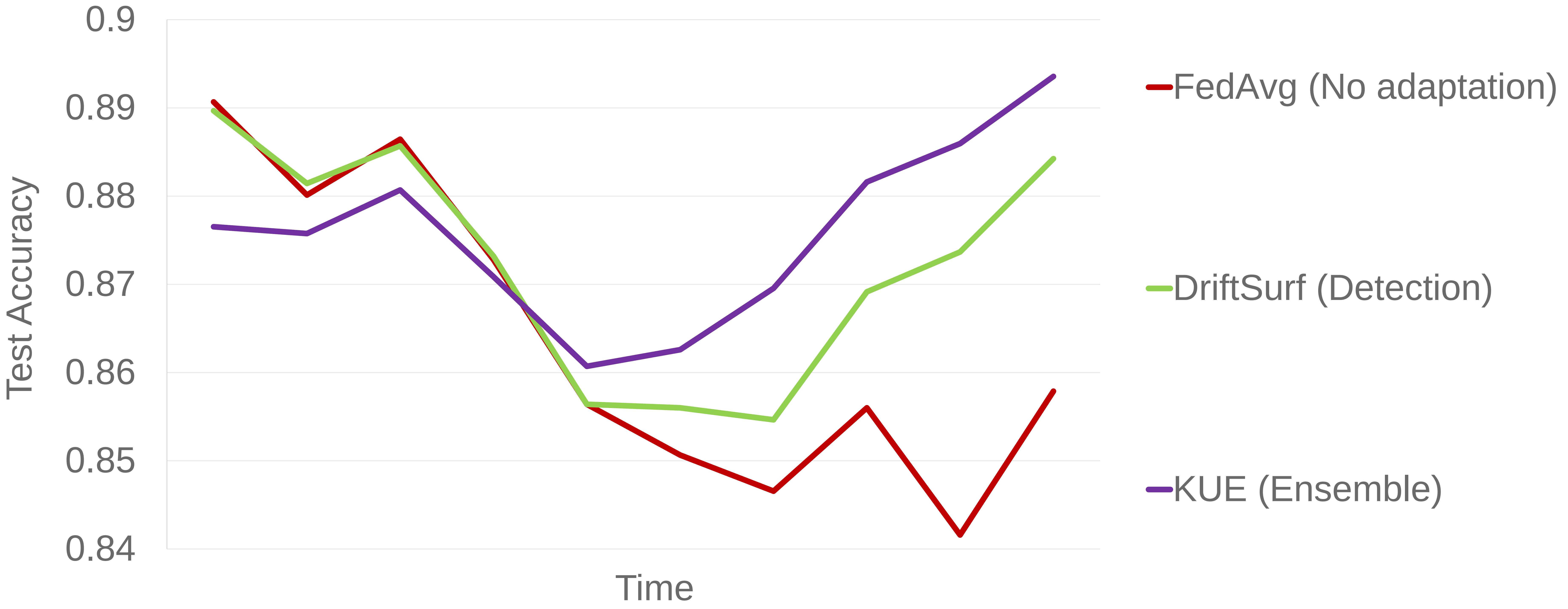
- Globally, drift is small compared to local drift for Africa
- Global model has only 48% accuracy on Africa post-2014, compared to 66% on rest of the world



Training a Single Global Model is Suboptimal

Observations from single concept change on SEA dataset staggered over time

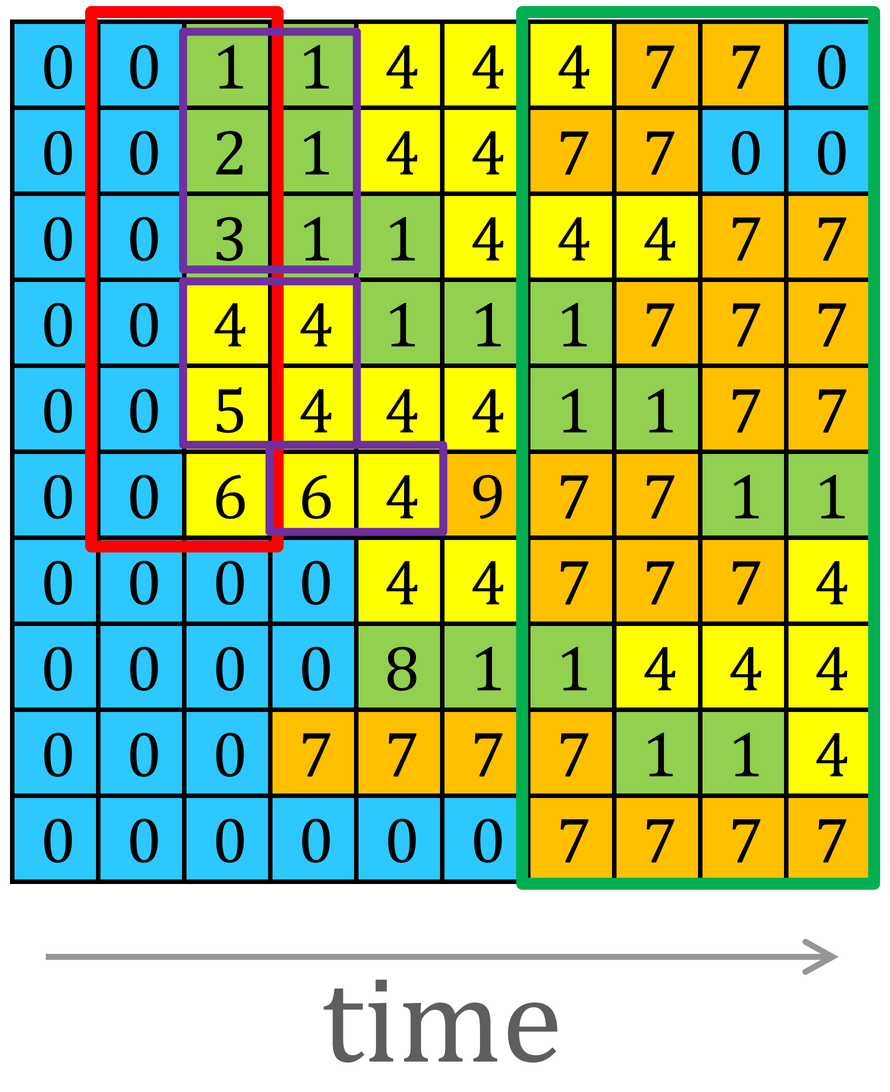




- Locally: Drift occurs abruptly
  Globally: Performance degrades slowly & drift is harder to detect
- Any single global model cannot fit both concepts during the transition

FedDrift Learns the Clustering of Clients

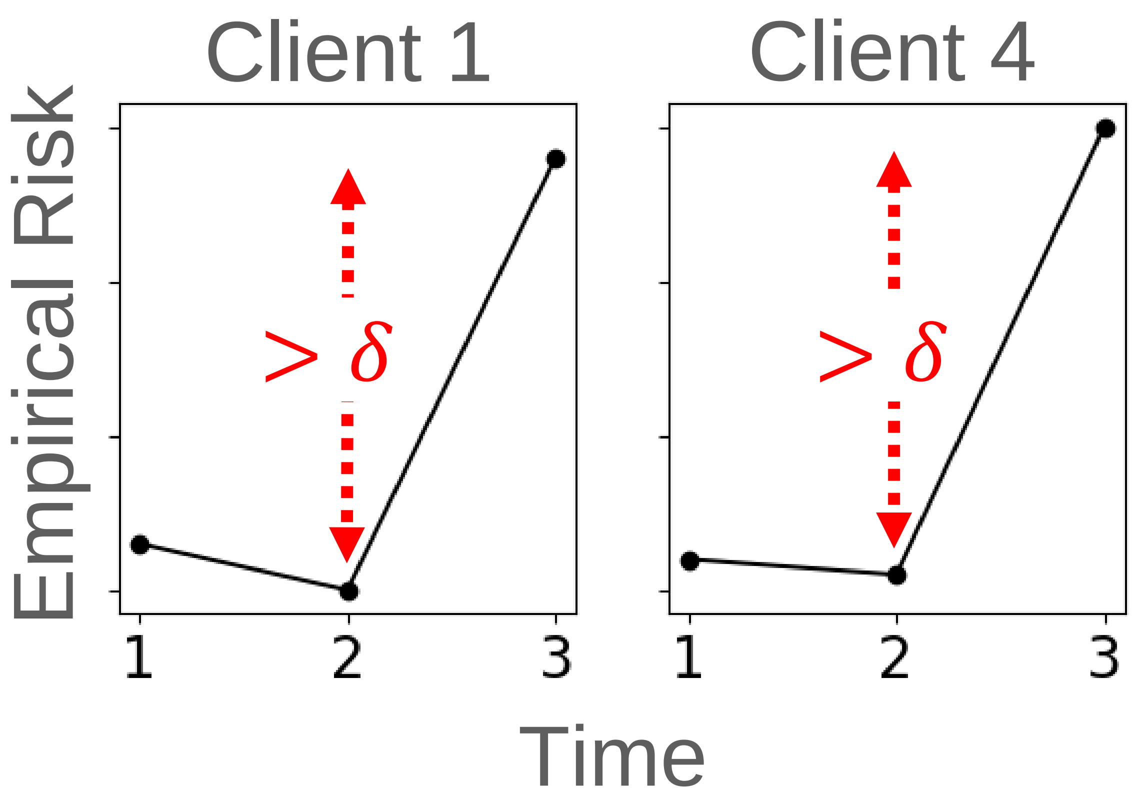
- FedDrift employs multiple models, each trained by a time-varying cluster of clients
- Challenge: determining the right number of clusters



Example clustering
Color: Ground-truth
Number: Cluster ID

1. Eager Splitting

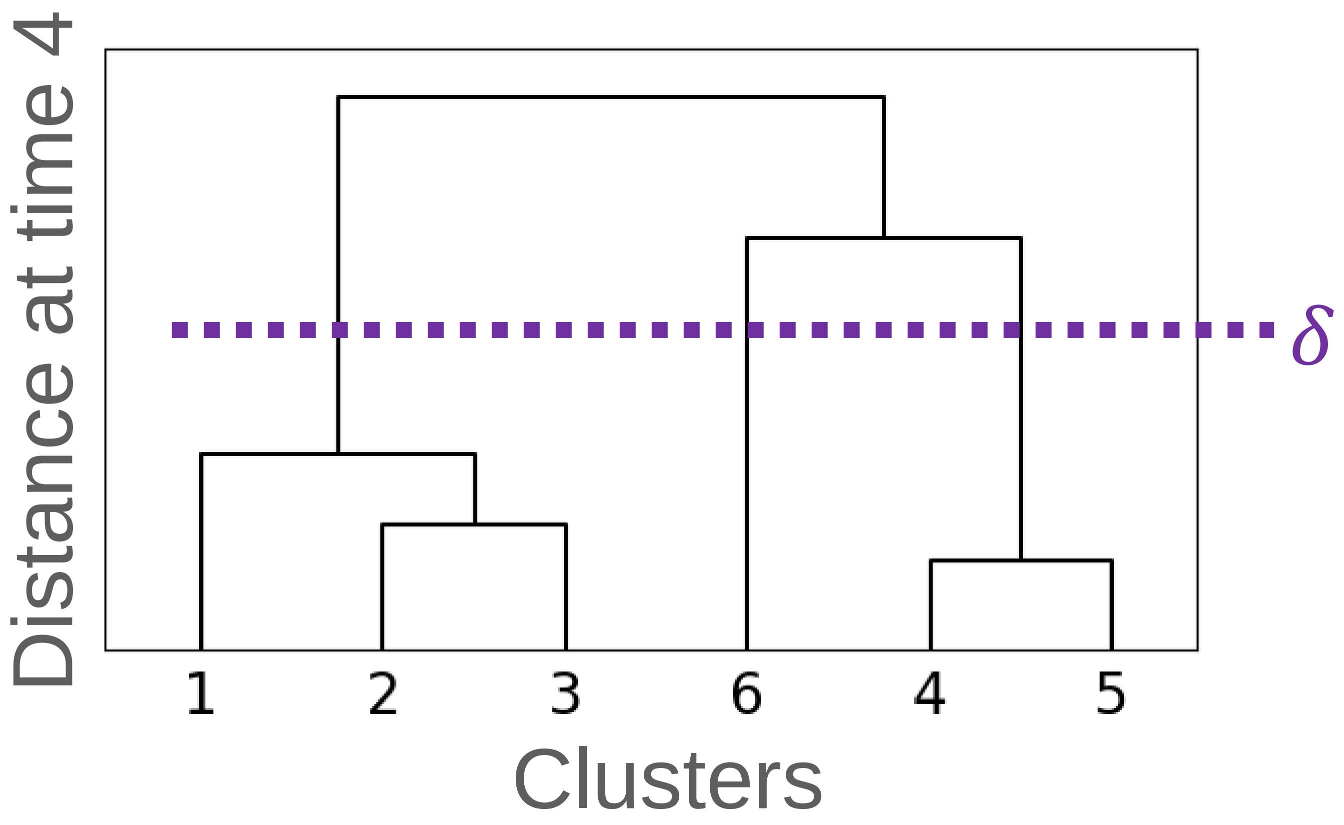
Isolate clients into individual clusters via local drift detection of size δ



2. Lazy Merging

Merge clusters via hierarchical clustering up to distance δ

Cluster distances are the pairwise drift



$d(i, j)$ = risk of model trained by cluster i over data of cluster j, relative to data of cluster i

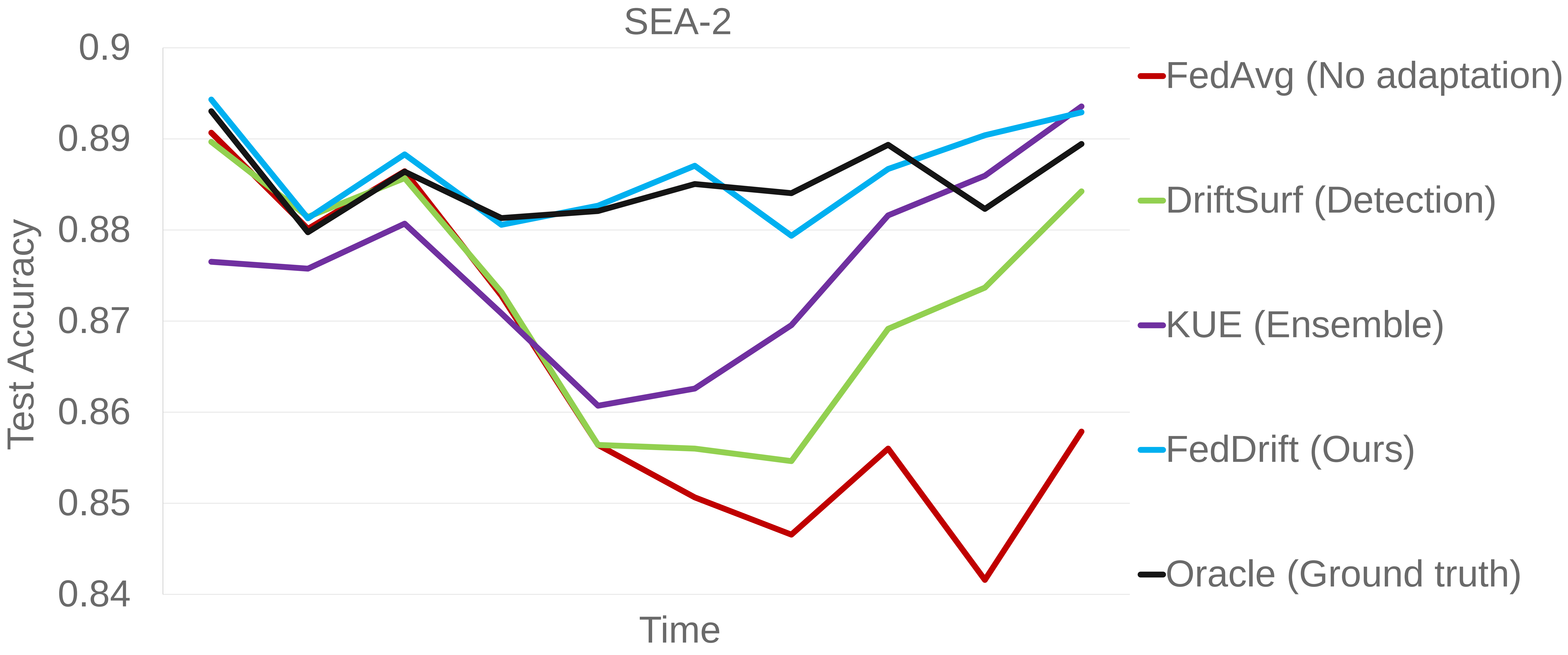
Ideally, clusters correspond 1-to-1 with concepts

Experimental Results: Single Staggered Drift

E.g., Single concept change on SEA dataset staggered over time

A	A	A	B	B	B	B	B	B	B
A	A	A	B	B	B	B	B	B	B
A	A	A	A	B	B	B	B	B	B
A	A	A	A	B	B	B	B	B	B
A	A	A	A	B	B	B	B	B	B
A	A	A	A	A	B	B	B	B	B
A	A	A	A	A	A	B	B	B	B
A	A	A	A	A	A	B	B	B	B
A	A	A	A	A	A	A	A	B	B
A	A	A	A	A	A	A	A	B	B

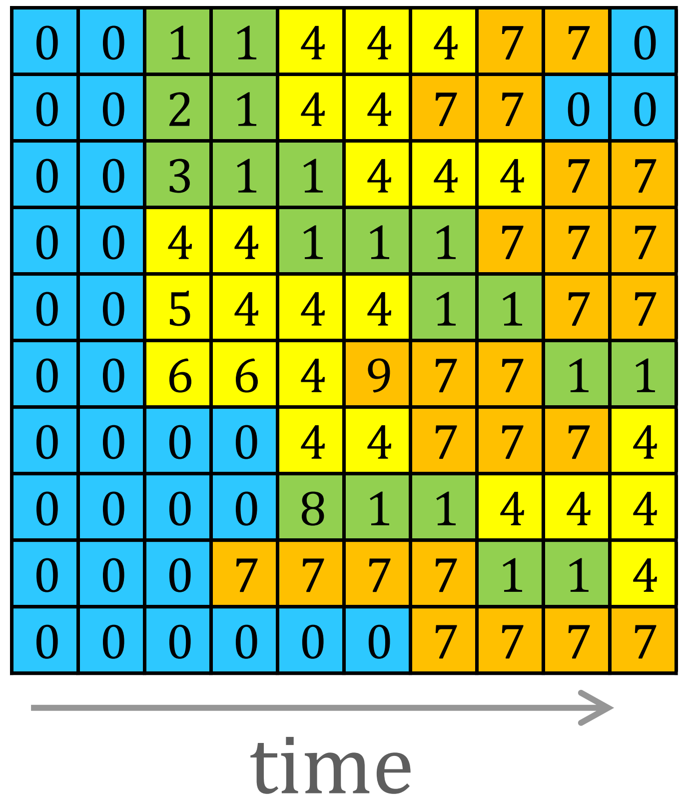
time →

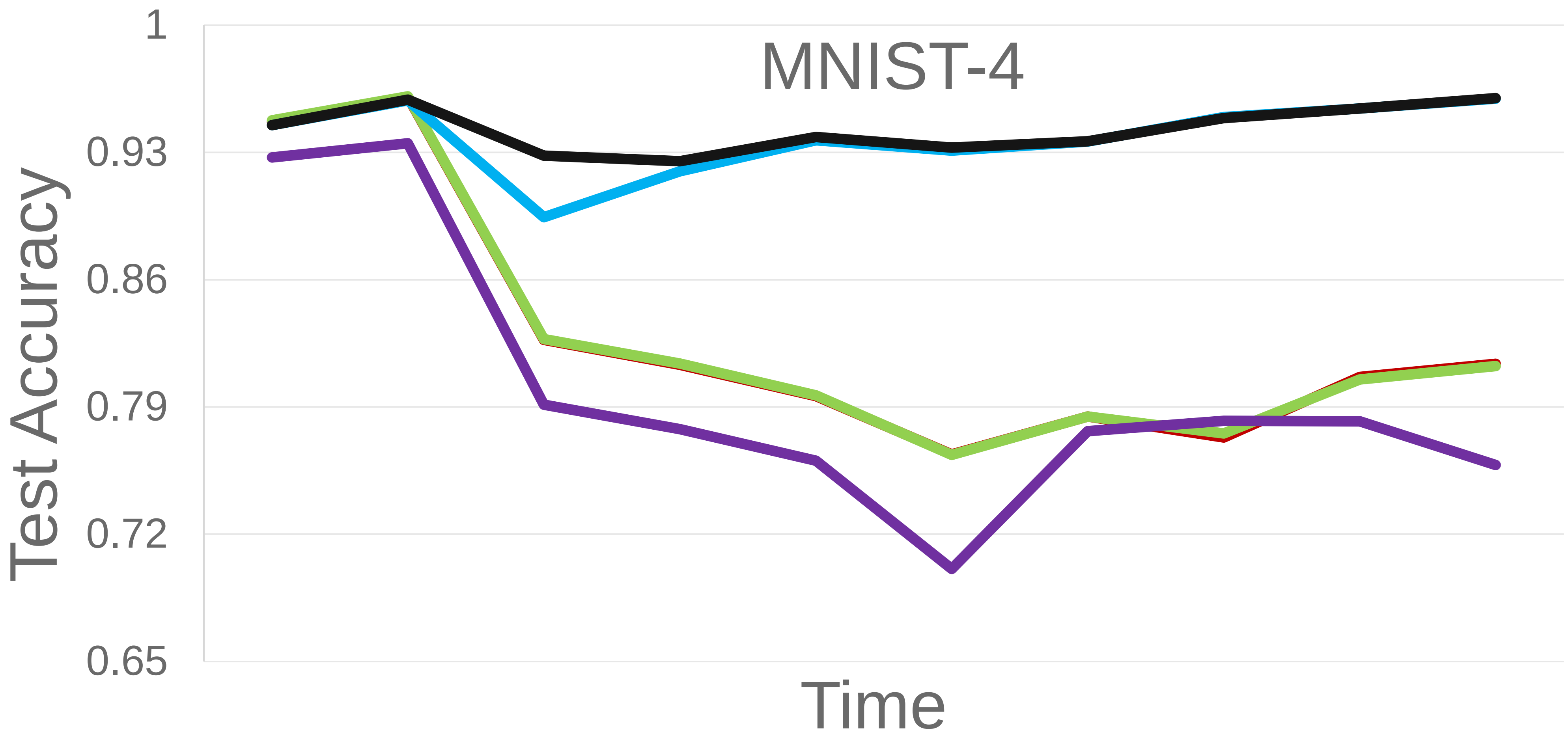


- Oracle has access to matrix, trains a model specialized for each concept
- FedDrift's accuracy is stable and comparable to Oracle

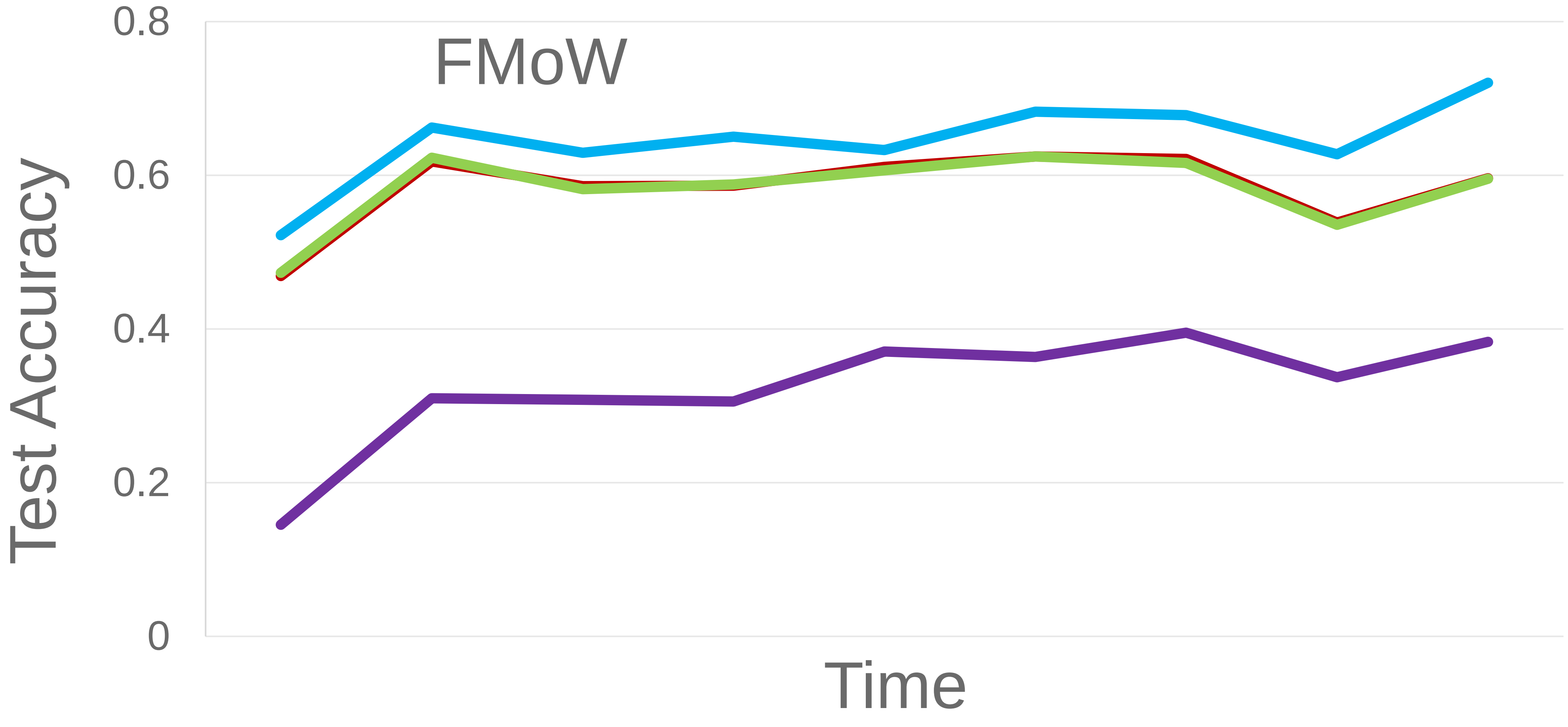
Experimental Results: More General Drifts

- On the 4-concept drift, FedDrift is comparable to Oracle





- On the real-world drift in FMoW, FedDrift outperforms the best baseline (64% to 58%)



(no Oracle for real data)

FedAvg (No adaptation) DriftSurf (Detection) KUE (Ensemble) FedDrift (Ours) Oracle (Ground truth)

FedDrift's Accuracy Higher Than Prior Work

Table 2: Average accuracy (%) across all clients and time (over 5 trials)

	SINE-2	CIRCLE-2	SEA-2	MNIST-2	SEA-4	MNIST-4	FMoW
FedAvg	52.11 ± 1.79	88.38 ± 0.17	86.46 ± 0.22	87.37 ± 0.16	85.40 ± 0.09	82.95 ± 0.03	58.57 ± 0.07
DriftSurf	84.18 ± 1.40	92.34 ± 0.38	87.20 ± 0.27	93.26 ± 0.52	85.55 ± 0.13	82.97 ± 0.09	58.45 ± 0.19
KUE	86.86 ± 0.17	93.71 ± 0.14	87.25 ± 0.94	90.44 ± 0.44	85.09 ± 0.86	79.89 ± 0.26	33.11 ± 6.09
AUE	86.00 ± 0.95	92.84 ± 0.19	87.48 ± 0.07	92.22 ± 0.05	85.47 ± 0.12	82.07 ± 0.47	54.23 ± 0.14
Window	86.28 ± 0.64	93.72 ± 0.14	87.94 ± 0.10	92.34 ± 0.07	85.72 ± 0.13	81.43 ± 0.44	58.88 ± 0.15
Adaptive-FedAvg	74.10 ± 10.03	86.26 ± 0.00	86.77 ± 0.53	92.18 ± 0.05	85.25 ± 0.27	81.64 ± 0.04	52.82 ± 0.21
IFCA+Window	$\mathbf{98.49 \pm 0.13}$	94.31 ± 1.62	$\mathbf{88.04 \pm 0.17}$	91.76 ± 0.50	86.17 ± 1.00	81.27 ± 0.43	49.40 ± 0.76
CFL+Window	96.92 ± 1.84	96.04 ± 1.56	87.81 ± 0.32	90.66 ± 0.35	86.06 ± 0.11	80.51 ± 0.72	58.82 ± 0.11
FedDrift	97.43 ± 0.06	$\mathbf{97.82 \pm 0.19}$	87.29 ± 0.75	95.48 ± 0.08	$\mathbf{88.13 \pm 0.76}$	$\mathbf{93.80 \pm 0.08}$	$\mathbf{64.84 \pm 0.33}$
Oracle	98.45 ± 0.03	97.84 ± 0.22	87.76 ± 0.98	95.54 ± 0.11	88.79 ± 0.41	94.30 ± 0.08	-

Takeaways

- Our work is the first to study drifts distributed both over time and across clients in federated learning
- Existing centralized solutions fail on staggered drifts
- FedDrift's eager splitting and lazy merging accurately clusters
- FedDrift achieves high accuracy on variety of drifts
  - Comparable to an idealized oracle algorithm on synthetic datasets
  - Outperforms the best baseline (64% to 58%) on the real-world FMoW

0	0	1	1
0	0	2	1
0	0	3	1
0	0	4	4
0	0	5	4
0	0	6	4
0	0	0	0
0	0	0	0