MobiCom 2024

Delta: A Cloud-assisted Data Enrichment Framework for On-Device Continual Learning

Chen Gong[1], Zhenzhe Zheng[1], Fan Wu[1], Xiaofeng Jia[2], Guihai Chen[1]

[1] **Shanghai Jiao Tong University**, [2] Beijing Big Data Center

2024-11-18

Outline

1. **Background**
2. Formulation & Challenges
3. Design of Delta
4. Evaluation Results

On-device Machine Learning

Machine learning models are crucial in modern mobile apps

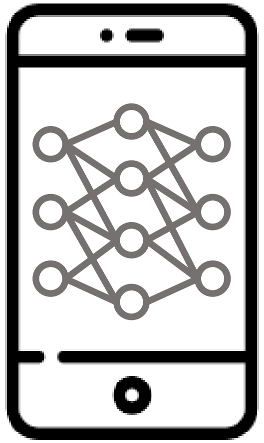
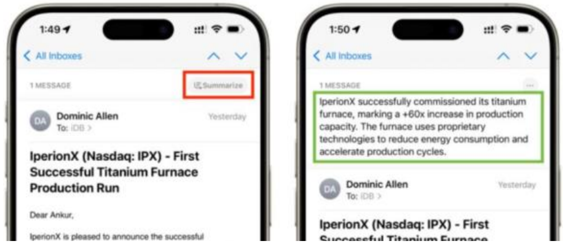


Image Analytics

Activity Recognition



Text Analysis

On-device Continual Learning

Mobile users typically encounter dynamic contexts

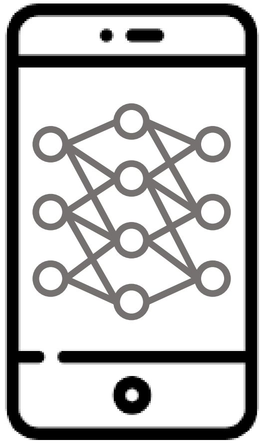
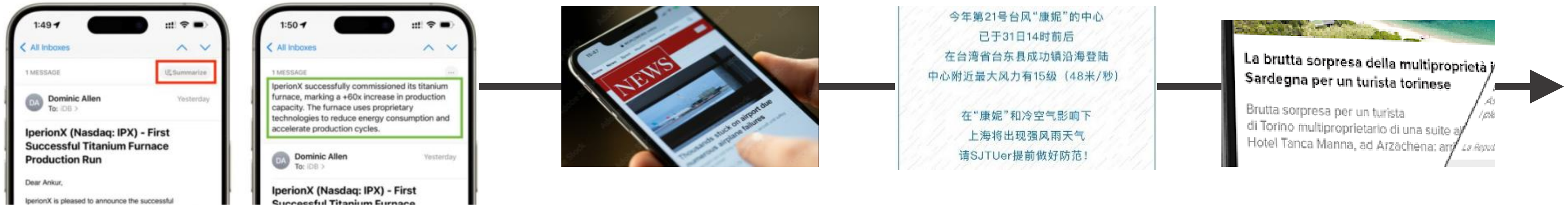


Image Analytics

Unseen weathers, objects

Activity Recognition

New device positions, human activities

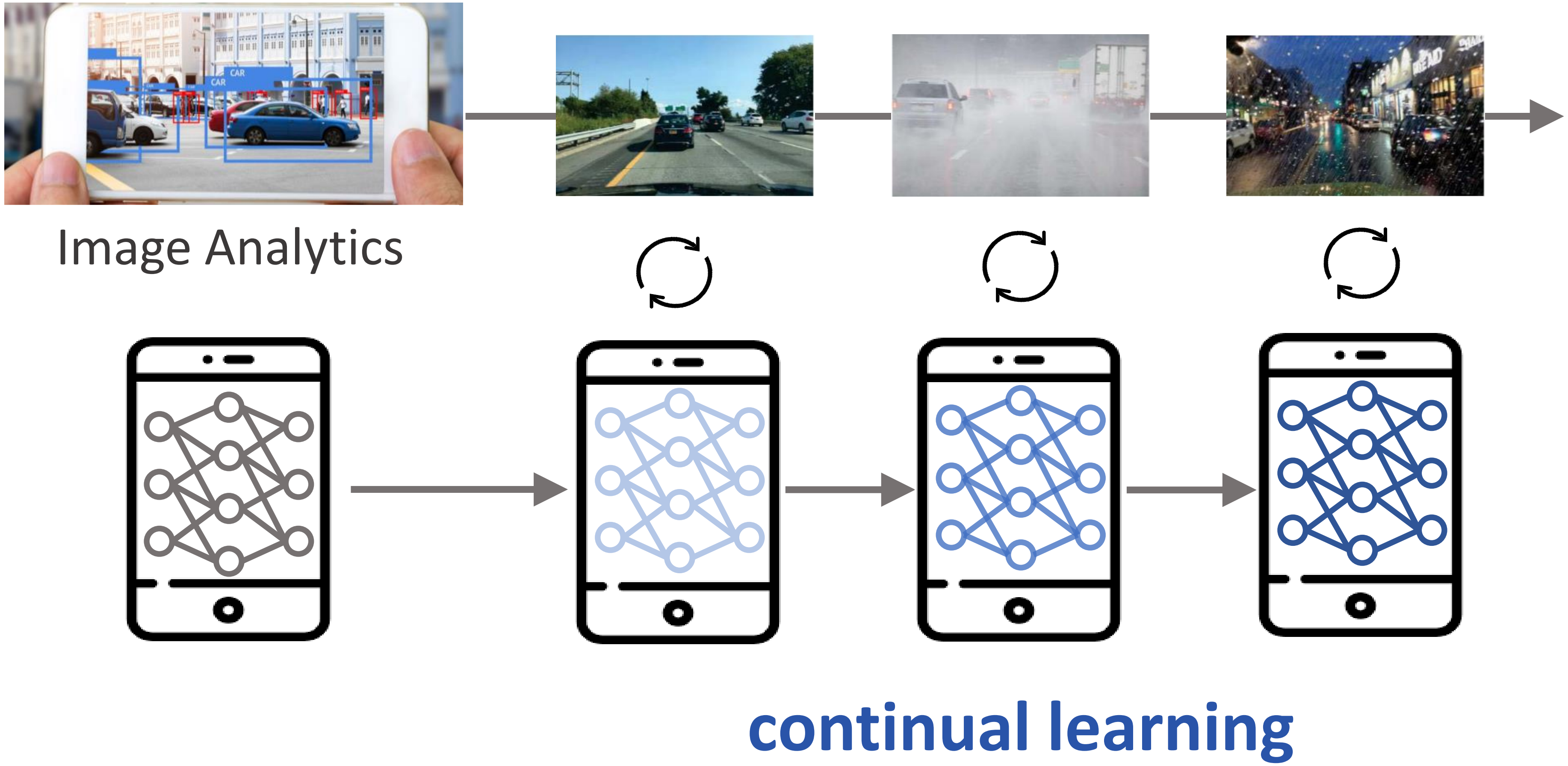


Text Analysis

Different languages, topics, …

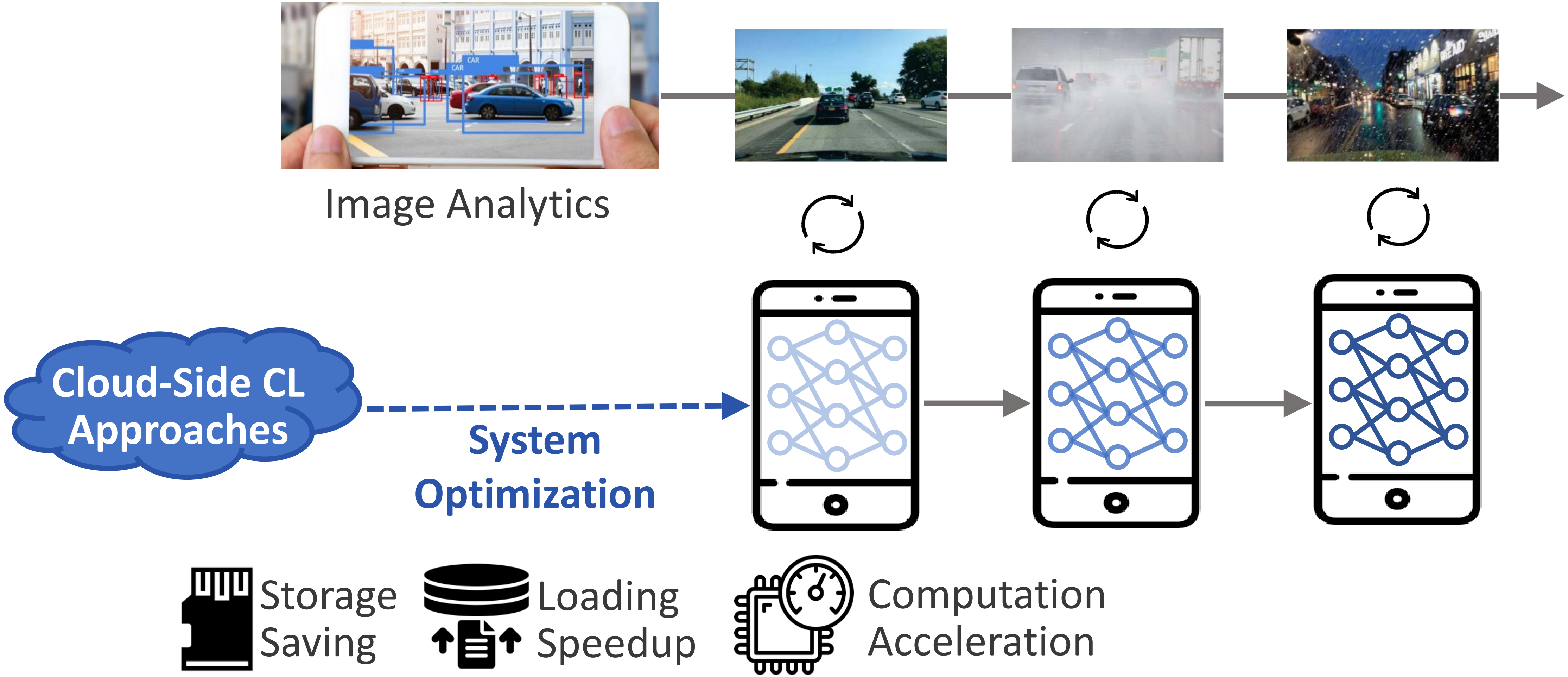
On-device Continual Learning

It is critical to enable continual learning on mobile devices



Prior Focus: System Bottleneck

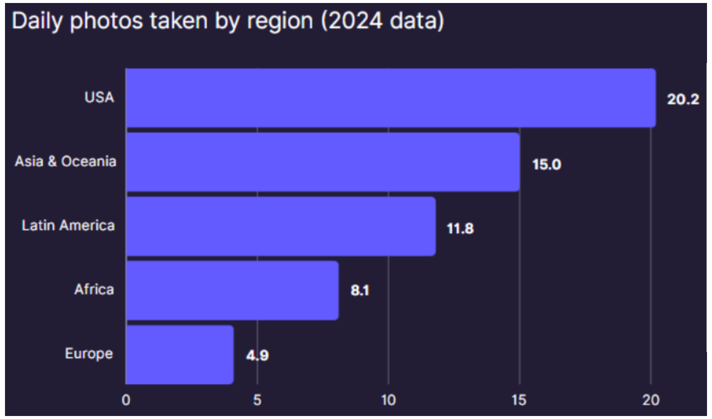
Efficient on-device deployment of cloud-side approaches



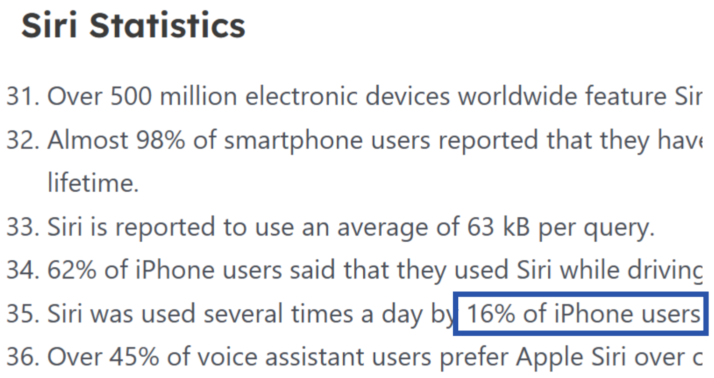
Our Focus: Data Bottleneck

Scarce data resource on mobile devices is a key bottleneck

Data scarcity is a prevalent issue



Average person takes **≈12 photos** daily [1]


Siri Statistics

31. Over 500 million electronic devices worldwide feature Sir
32. Almost 98% of smartphone users reported that they have lifetime.
33. Siri is reported to use an average of 63 kB per query.
34. 62% of iPhone users said that they used Siri while driving
35. Siri was used several times a day by 16% of iPhone users
36. Over 45% of voice assistant users prefer Apple Siri over c

16% of iPhone users use Siri **several times** a day [2]

Data sets the performance ceiling

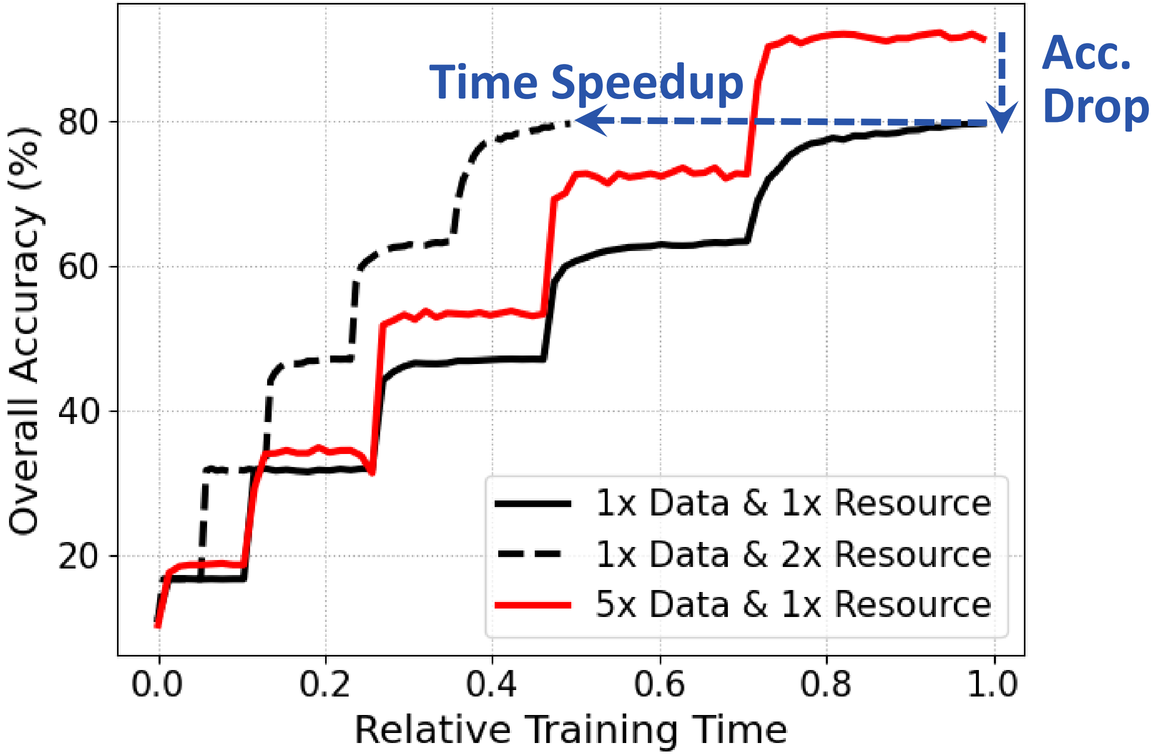
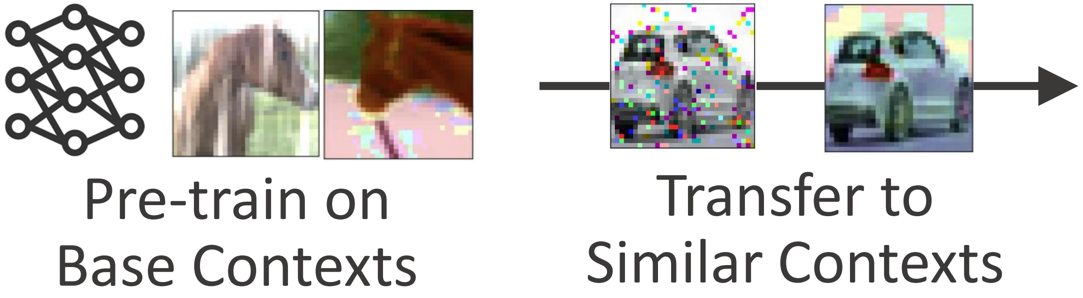


Image classification task with new weathers

Existing Solutions

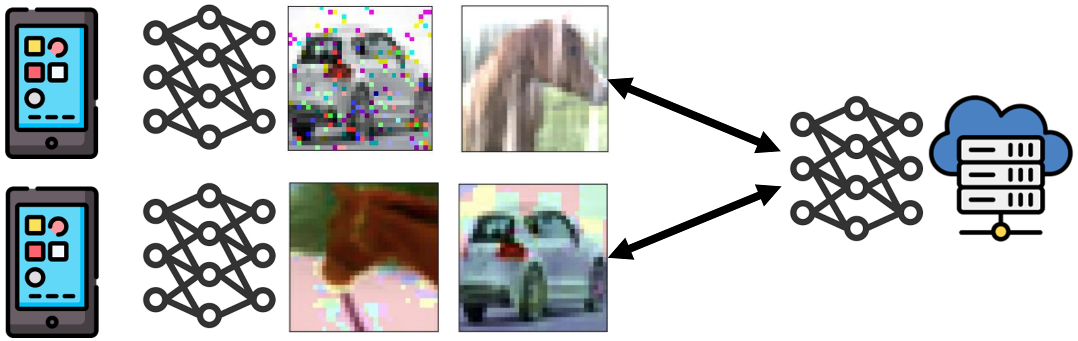
Param-and-model-based methods are ineffective or inefficient

#1 Param-based: Few-Shot CL



Ineffective for Unpredictable User Contexts

#2 Model-based: Federated CL

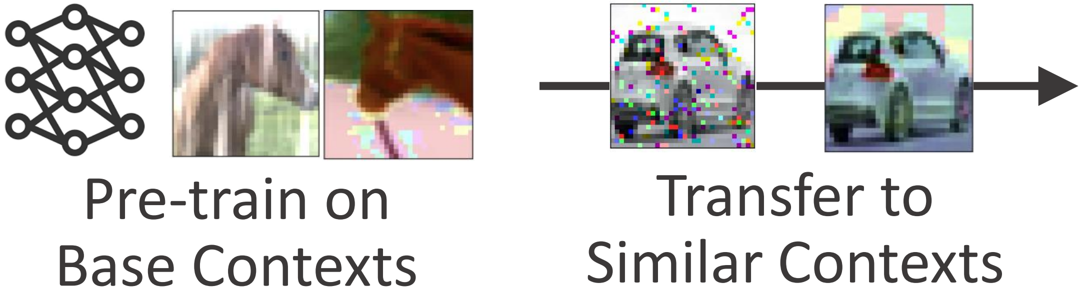


Inefficient for Heterogeneous Cross-Device Contexts

Existing Solutions

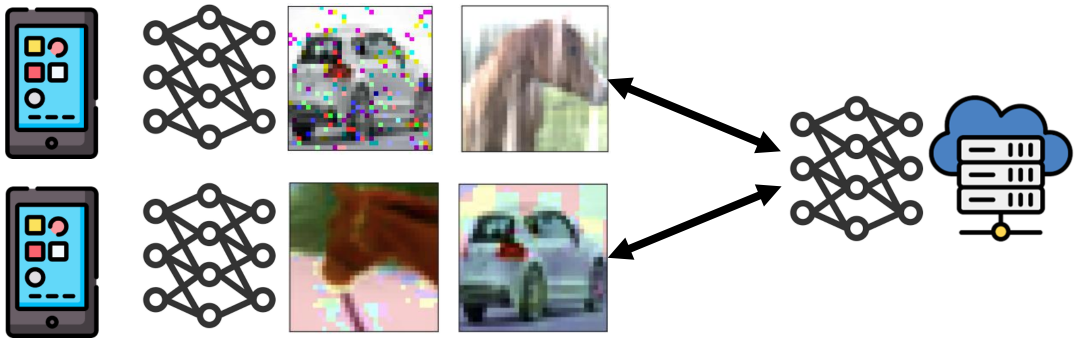
Param-and-model-based methods are ineffective or inefficient

#1 Param-based: Few-Shot CL



Ineffective for Unpredictable User Contexts

#2 Model-based: Federated CL



Inefficient for Heterogeneous Cross-Device Contexts

Fundamental Solution from Data Aspect:
Enrich scarce device data with cloud data !

Observations

#1 Abundant Cloud Data

Public Datasets

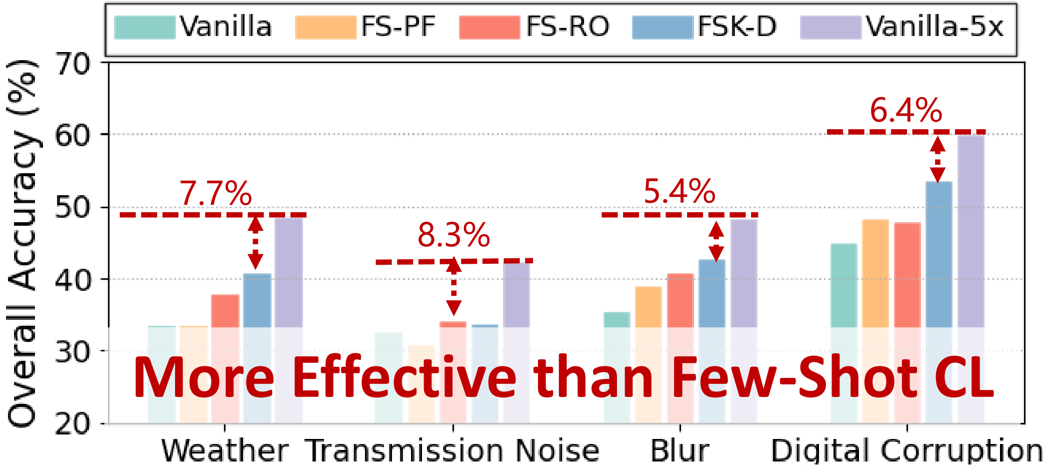
COMMON CRAWL

Crawled Internet Data

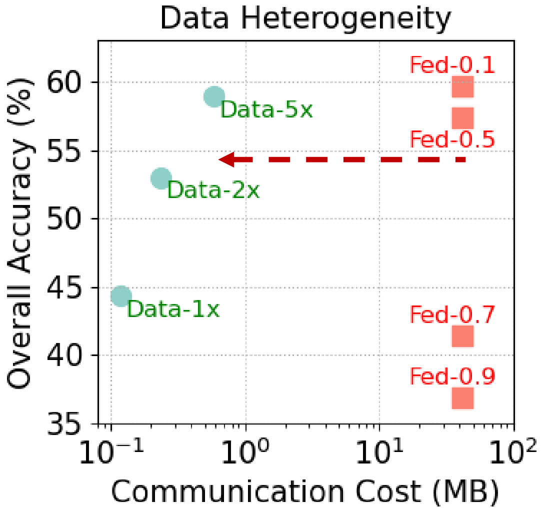
amazon mechanical turk

Crowd-Sourced Data

#2 Large Potential Improvement



More Effective than Few-Shot CL



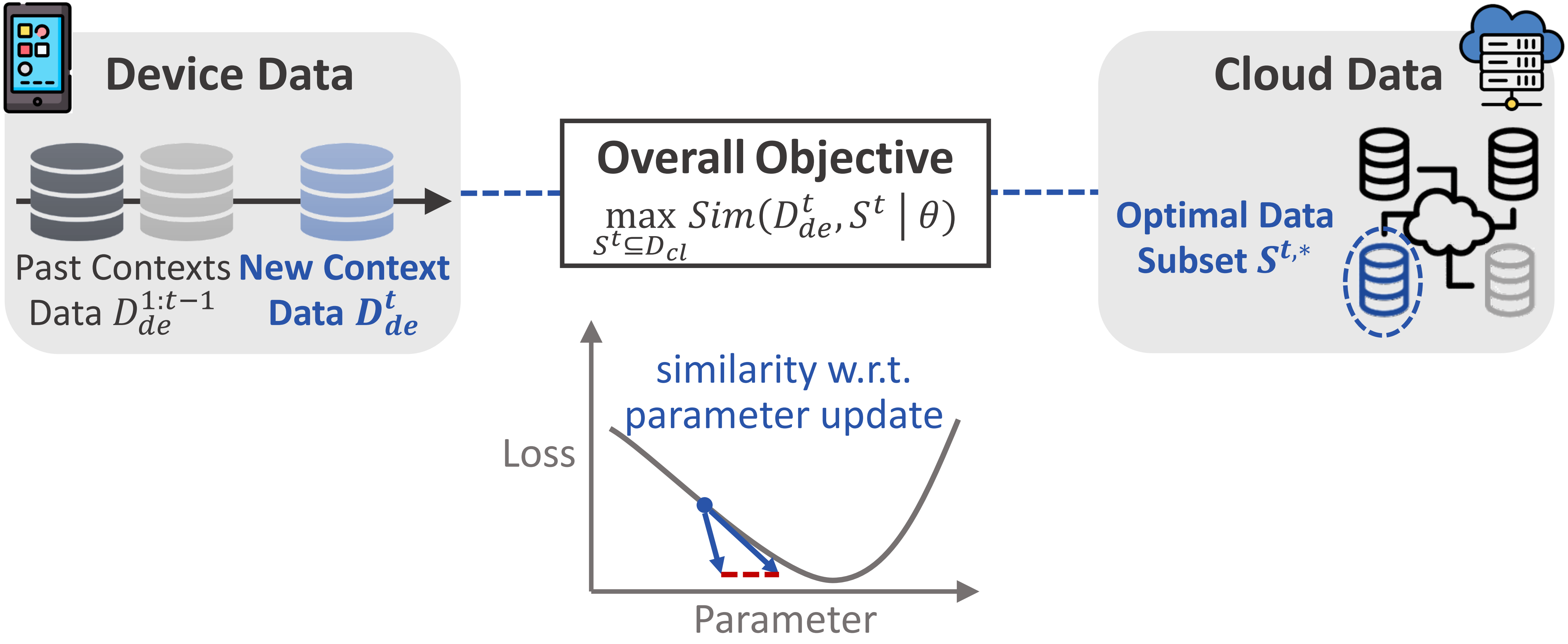
More Efficient than Federated CL

Outline

1. Background
2. **Formulation & Challenges**
3. Design of Delta
4. Evaluation Results

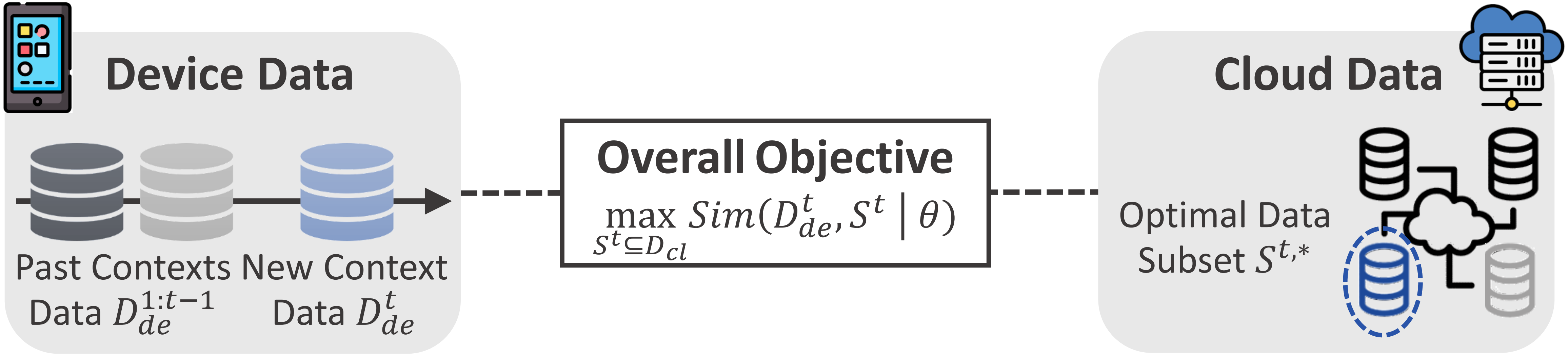
Problem Formulation

Select the cloud data subset most similar to device data



Challenges

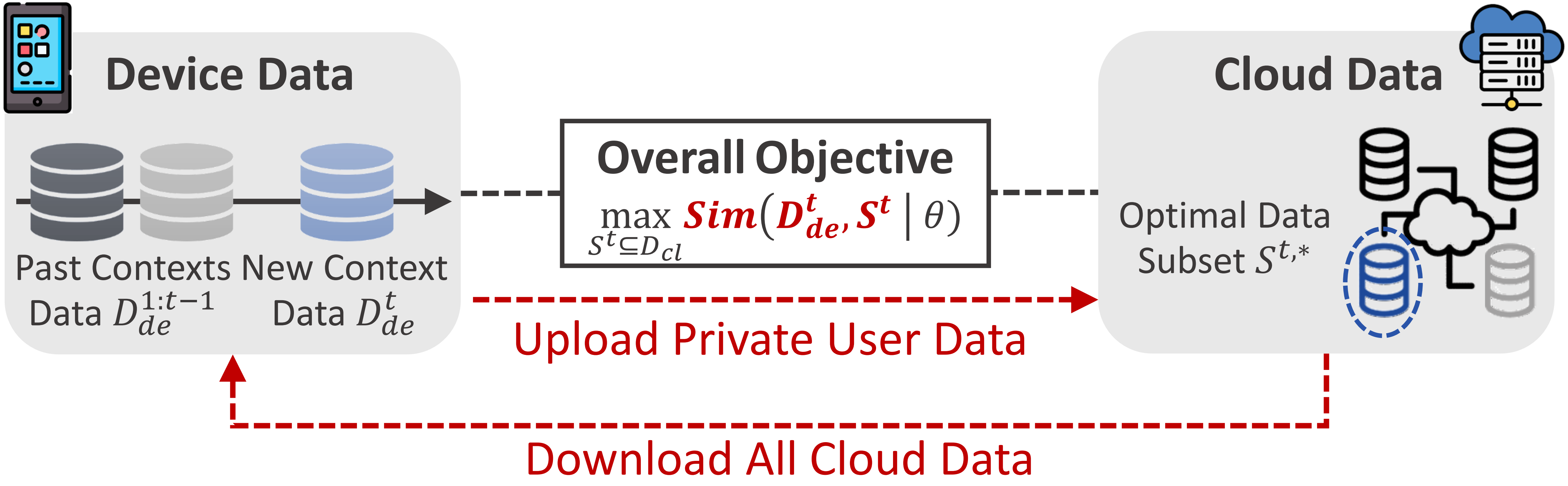
Developing a feasible framework face critical challenges



How to achieve privacy, efficiency and effectiveness simultaneously?

Challenge 1: Privacy and Efficiency

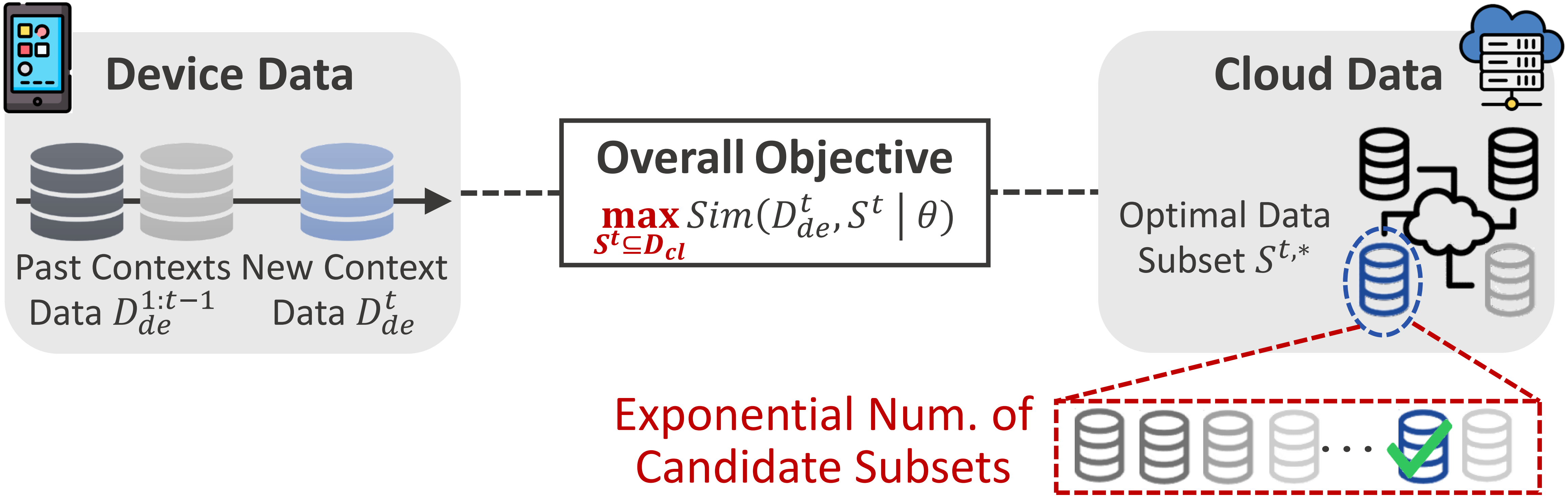
Developing a feasible framework face critical challenges



How to achieve **privacy, efficiency** and effectiveness simultaneously?

Challenge 2: Efficiency and Effectiveness

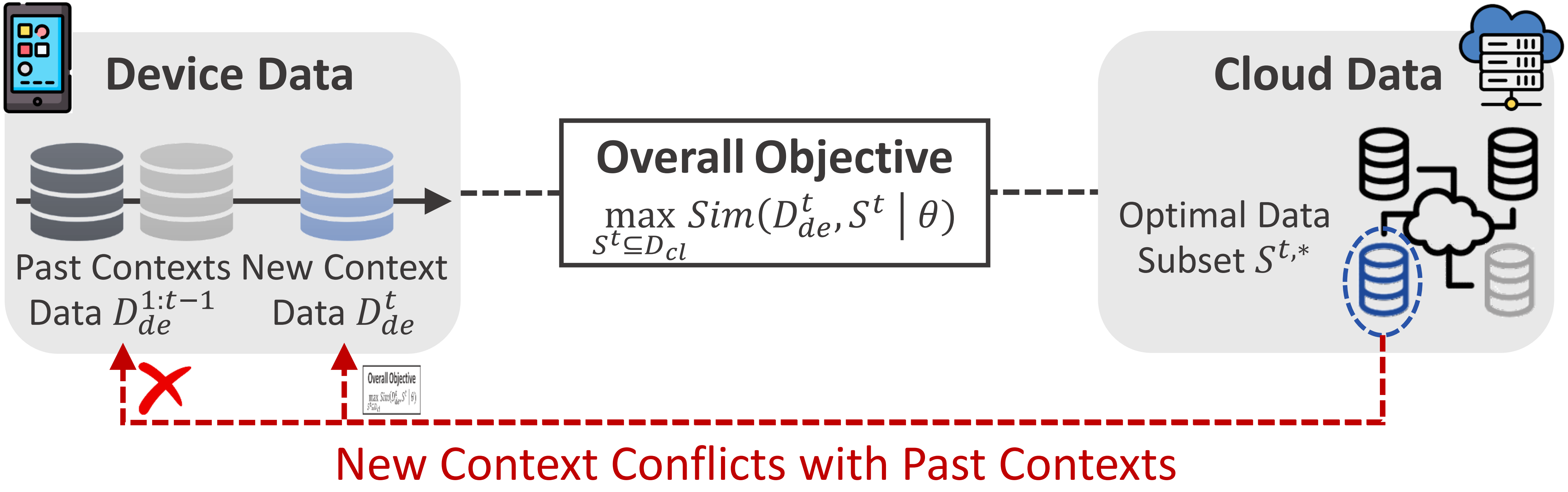
Developing a feasible framework face critical challenges



How to achieve privacy, **efficiency and effectiveness** simultaneously?

Challenge 3: Effectiveness in Continual Learning

Developing a feasible framework face critical challenges



How to achieve privacy, efficiency and **effectiveness** simultaneously?

Outline

1 Background

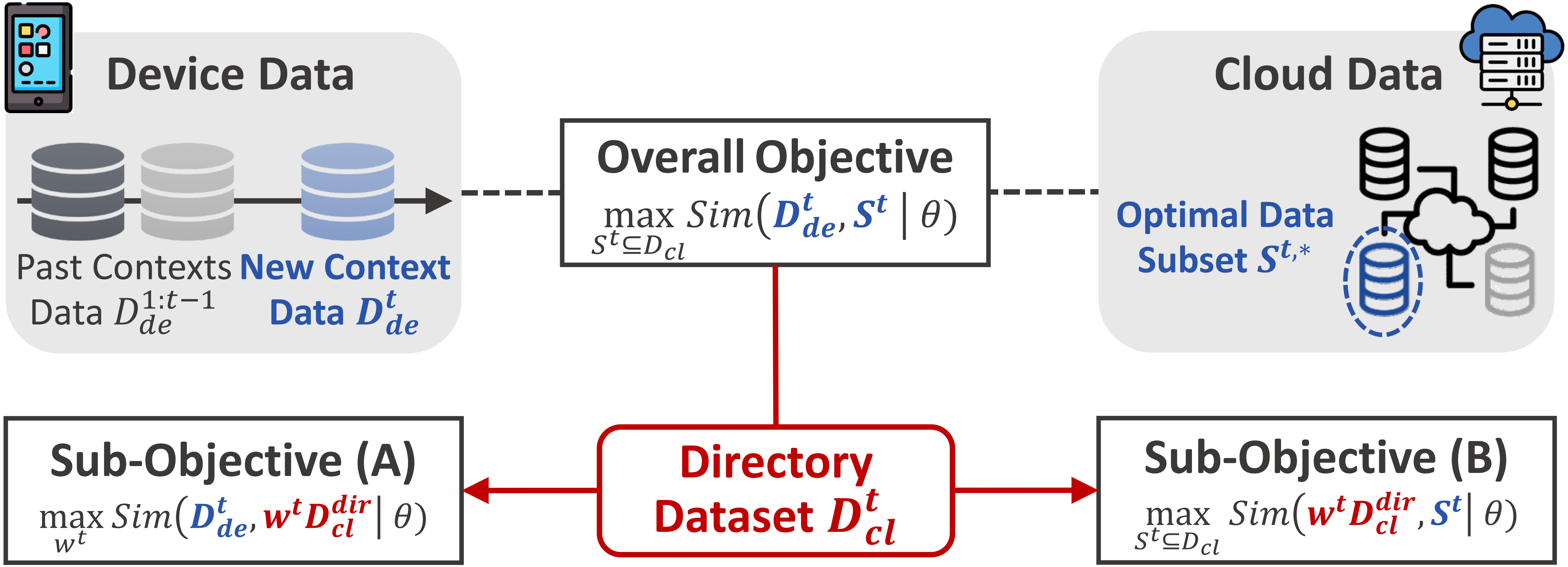
2 Formulation & Challenges

3 **Design of Delta**

4 Evaluation Results

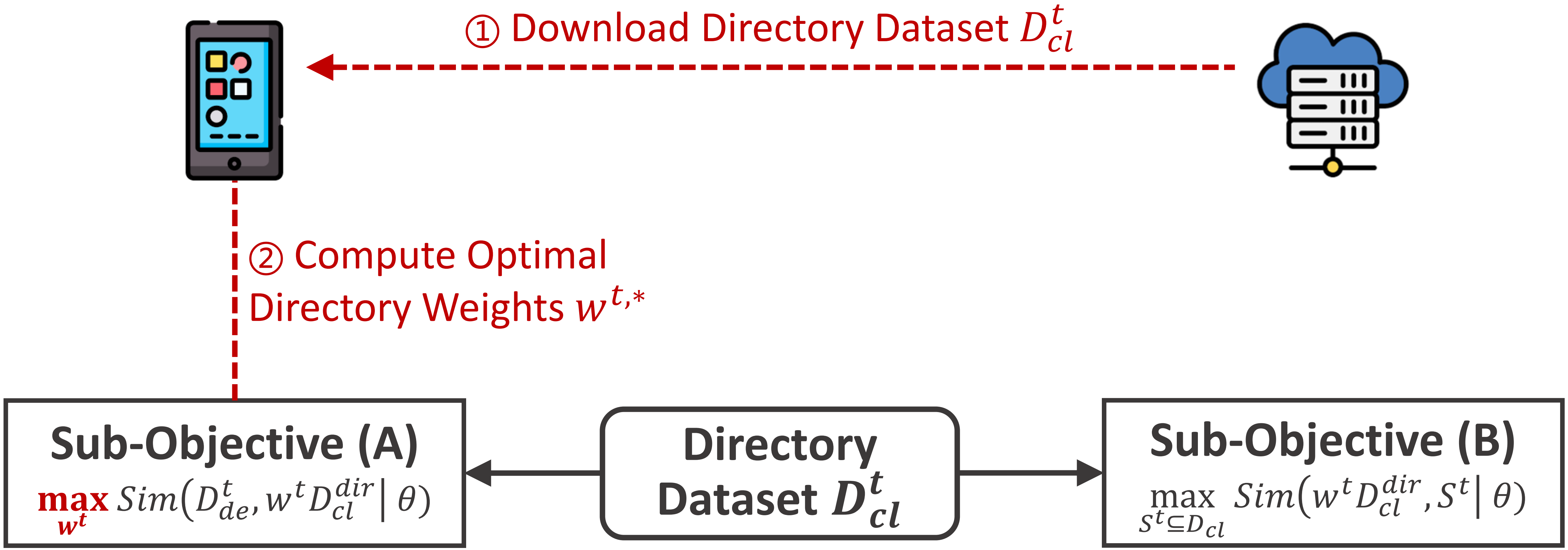
Privacy: Problem Decomposition

Introduce cloud directory dataset for problem decomposition

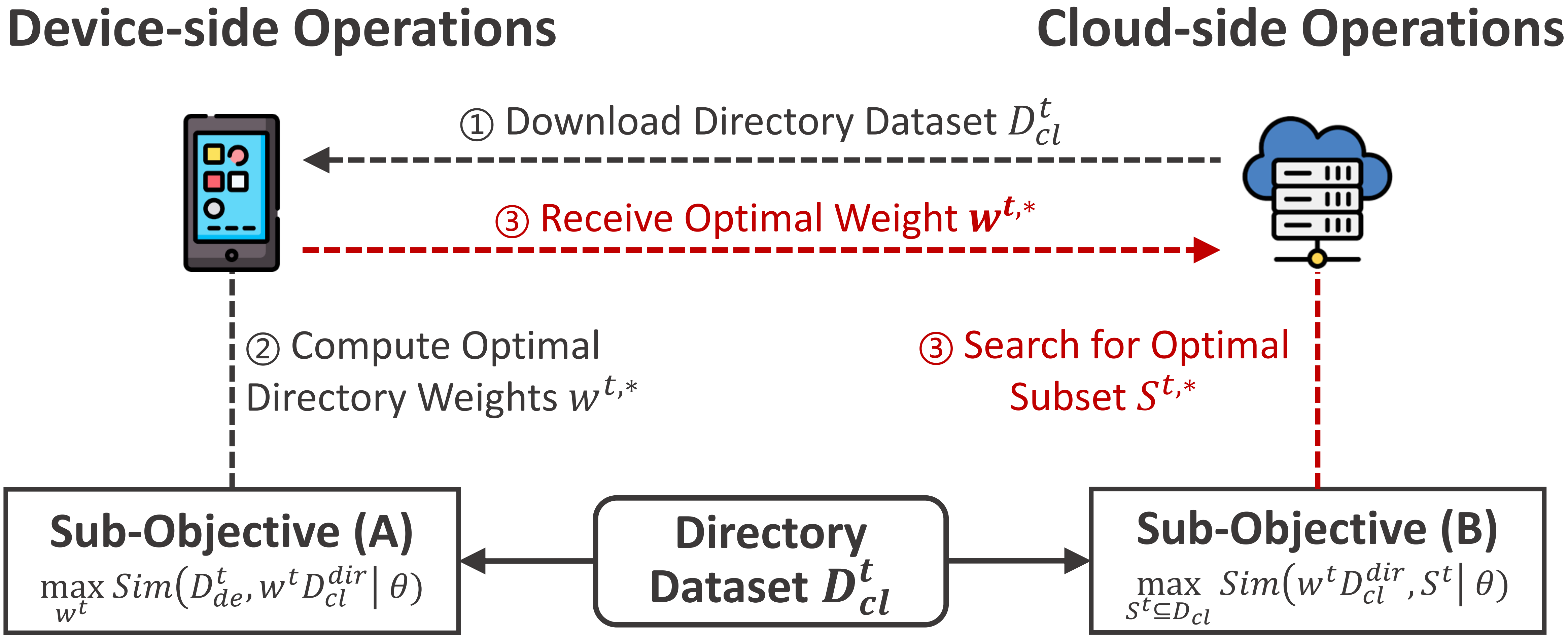


Privacy: Device-Cloud Collaboration

Device-side Operations



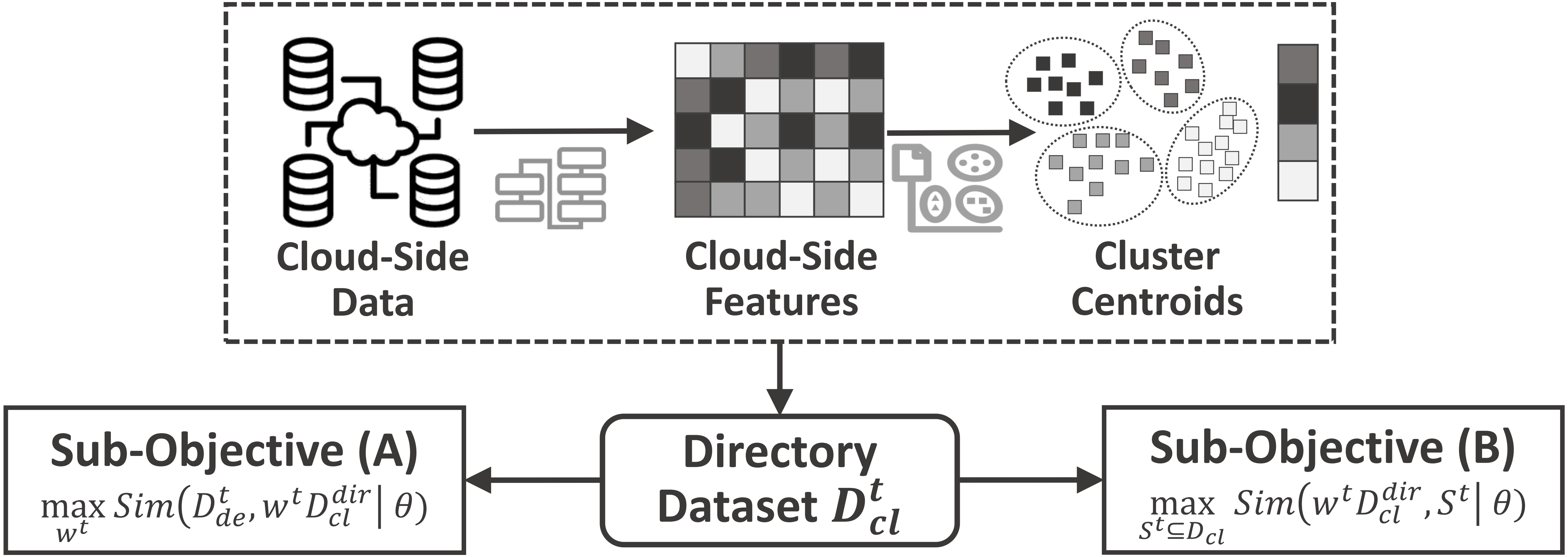
Privacy: Device-Cloud Collaboration



w/o Sharing Raw User Data Samples

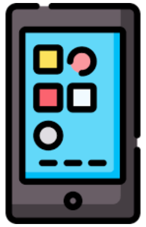
Privacy: Directory Dataset Construction

How to construct a representative directory dataset?



Efficiency: Failure of Naïve Solutions

Naïve solutions are inefficient for device and cloud sub-objectives

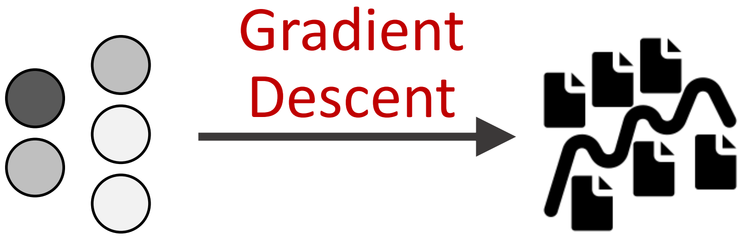


Sub-Objective (A)

$$\max_{w^t} Sim\left(D_{de}^{t}, w^t D_{cl}^{dir} \middle| \theta\right)$$

Sub-Objective (B)

$$\max_{S^t \subseteq D_{cl}} Sim\left(w^t D_{cl}^{dir}, S^t \middle| \theta\right)$$



Scarce Data Overfitted Weight

Exponential Subsets Optimal Subset

Efficiency: Device-Side Soft Matching

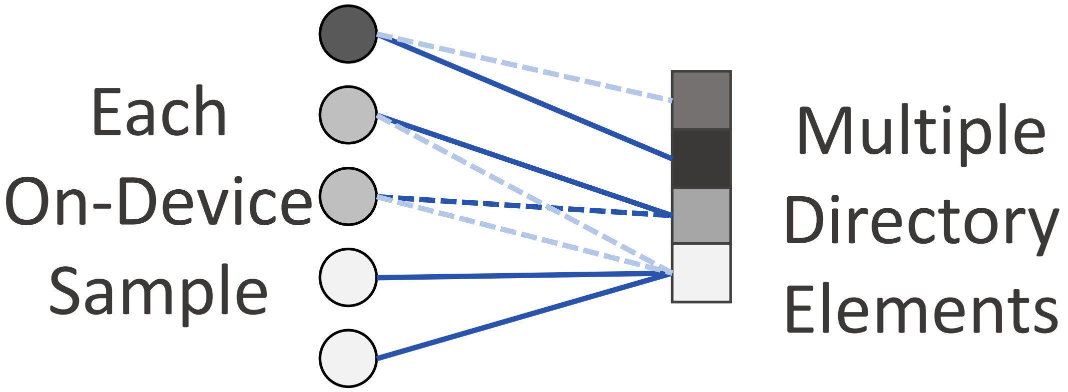
Device-side soft matching strategy for representative weight

Sub-Objective (A)

$$\max_{w^t} Sim(D_{de}^t, w^t D_{cl}^{dir} \mid \theta)$$

Sub-Objective (B)

$$\max_{S^t \subseteq D_{cl}} Sim(w^t D_{cl}^{dir}, S^t \mid \theta)$$



Soft Matching

$$w_c^t \leftarrow w_c^t + Softmax\left(\frac{Sim\left((x,y),(\bar{x}_c,\bar{y}_c) \mid \theta^{t-1}\right)}{\tau}\right),$$

Efficiency: Cloud-Side Optimal Sampling

Cloud-side optimal sampling with constant time complexity

Sub-Objective (A)

$$\max_{w^t} Sim\left(D^t_{de}, w^t D^{dir}_{cl} \middle| \theta\right)$$

Sub-Objective ($\widehat{B}$)

$$\max_{P^t_{cl}} \boldsymbol{E}_{S^t \sim P^t_{cl}} Sim\left(w^t D^{dir}_{cl}, S^t \middle| \theta\right)$$

Scarce Data

Directory Dataset

Soft Matching

$$w_c^t \leftarrow w_c^t + Softmax\left(\frac{Sim\left((x,y),(\bar{x}_c,\bar{y}_c) \mid \theta^{t-1}\right)}{\tau}\right),$$

Optimal Sampling Strategy

Optimal in Expectation

Efficiency: Cloud-Side Optimal Sampling

Cloud-side optimal sampling with constant time complexity

Sub-Objective (A)

$$\max_{w^t} Sim(D^t_{de}, w^t D^{dir}_{cl} \mid \theta)$$

Soft Matching

$$w_c^t \leftarrow w_c^t + Softmax\left(\frac{Sim((x,y),(\bar{x}_c,\bar{y}_c) \mid \theta^{t-1})}{\tau}\right),$$

Sub-Objective ($\hat{B}$)

$$\max_{P^t_{cl}} \mathbf{E}_{S^t \sim P^t_{cl}} Sim(w^t D^{dir}_{cl}, S^t \mid \theta)$$

Optimal Sampling Strategy

Optimal in Expectation

Intra-Cluster Prob. (Pre-computed)
Inter-Cluster Size (Real-Time Updated)

Effectiveness: Theoretical Analysis

Theorem. *The impact of enriched data on overall continual learning performance is determined by*

(1) new-context representativeness

(2) past-contexts proximity

(3) cross-context heterogeneity

$$\mathbb{E}_{S^t\sim P^t_{\mathcal{D}_{cl}}}\Big[\underbrace{L(\mathcal{D}^{1:t}_{de},\theta^{t,m+1}) - L(\mathcal{D}^{1:t}_{de},\theta^{t,m})}_{\text{loss reduction in } m\text{-th model update}}\Big]$$

$$\le \frac{1}{2}(H\eta^2-\eta)L_\psi \underbrace{\boxed{\mathbb{V}_{S^t\sim P^t_{\mathcal{D}_{cl}}}\left[\phi(\mathcal{D}^t_{de})-\phi(S^t)\right]}}_{\text{representativeness to new context } t} +$$

$$\frac{\eta L_\psi}{2}\underbrace{\boxed{\mathbb{V}_{S^t\sim P^t_{\mathcal{D}_{cl}}}\left[\phi(\mathcal{D}^{1:t-1}_{de})-\phi(S^t)\right]}}_{\text{proximity to past contexts } 1\sim t-1} + \frac{\eta L_\psi}{2}\underbrace{\boxed{\left\|\phi(\mathcal{D}^t_{de})-\phi(\mathcal{D}^{1:t-1}_{de})\right\|^2,}}_{\text{heterogeneity across contexts}}$$

Effectiveness: Theoretical Analysis

Theorem. *The impact of enriched data on overall continual learning performance is determined by*

(1) new-context representativeness

(2) past-contexts proximity

(3) cross-context heterogeneity

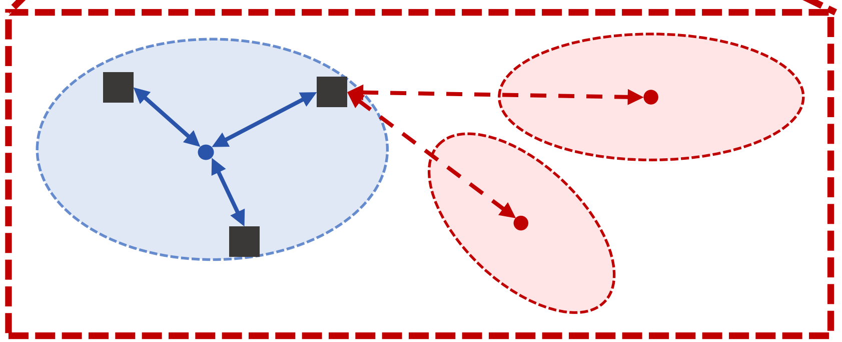
$$\mathbb{E}_{S^t\sim P^t_{\mathcal{D}_{cl}}}\Big[\underbrace{L(\mathcal{D}^{1:t}_{de},\theta^{t,m+1}) - L(\mathcal{D}^{1:t}_{de},\theta^{t,m})}_{\text{loss reduction in } m\text{-th model update}}\Big]$$

$$\le \frac{1}{2}(H\eta^2-\eta)L_\psi \underbrace{\boxed{\mathbb{V}_{S^t\sim P^t_{\mathcal{D}_{cl}}}\left[\phi(\mathcal{D}^t_{de})-\phi(S^t)\right]}}_{\text{representativeness to new context } t}+$$

$$\frac{\eta L_\psi}{2}\underbrace{\boxed{\mathbb{V}_{S^t\sim P^t_{\mathcal{D}_{cl}}}\left[\phi(\mathcal{D}^{1:t-1}_{de})-\phi(S^t)\right]}}_{\text{proximity to past contexts } 1\sim t-1}+\frac{\eta L_\psi}{2}\underbrace{\boxed{\left\|\phi(\mathcal{D}^t_{de})-\phi(\mathcal{D}^{1:t-1}_{de})\right\|^2}}_{\text{heterogeneity across contexts}},$$

Re-Optimize Sampling Strategy

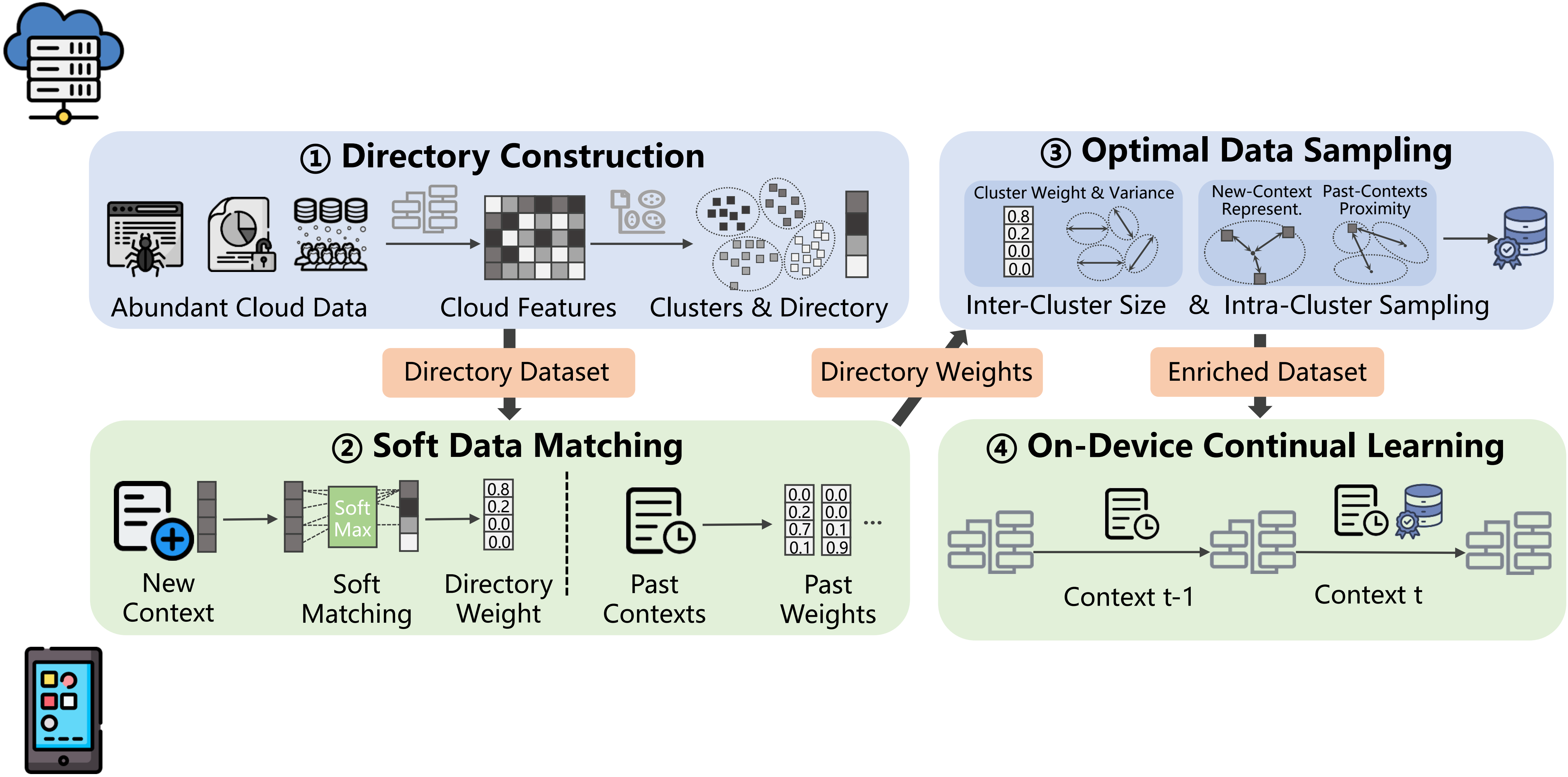
Optimal in Expectation



Consider Impacts on New & Past Contexts

Refer to our paper for more details!

Overall Workflow



Outline

1. Background
2. Formulation & Challenges
3. Design of Delta
4. **Evaluation Results**

Evaluation Setup

- **Implementation**
  - Device: Jetson Nano
  - Cloud: NVIDIA 3090Ti
- **Baselines**
  - 3 **few-shot CL** algorithms
  - **Federated CL**
  - Random data enrichment
- **Tasks & Datasets**
  - **4 tasks** & data modalities
  - Each with **≥2 categories of ≥5 contexts**
  - 4 ML models
- **Configurations**
  - Cloud data: random 50% samples
  - Device data: 5 samples/context
  - Directory: 20 x num. of classes

Modality	Context Category	Dataset	Model(#params)
Image	Object (O), Weather (W), Noise (N), Blur (B), Digital Corruption (D)	Cifar10-C	ResNet18(11.2M)
IMU	Activity (A), Physical Condition (P), Device Placement (D)	HHAR, UCI, Motion, Shoaib	DCNN(17.3K)
Audio	User Command (C), Tone (T), Environmental Noise (N)	Google Speech	VGG11(9.75M)
Text	Article Topic (T), Language (L)	XGLUE	BERT(0.178B)

Overall Performance

Higher overall CL performance compared with few-shot CL:

- 15.1%, 12.4%, 1.1%, 5.6% accuracy improvement for visual, IMU, audio, textual tasks

Tasks	Context Category	Vanilla CL	Few-Shot CL			Federated CL			Data Enrichment		ΔAcc.	ΔComm.
			FS-KD	FS-RO	FS-PF	Fed-0.1	Fed-0.2	Fed-0.4	Random	Delta		
IC	O+W	32.7±1.49	41.7±1.78	39.2±2.13	36.9±2.87	31.8±0.24	46.4±1.65	55.1±0.42	42.5±2.42	57.7±0.54	16.0% ↑	93.7% ↓
	O+N	31.3±1.74	36.2±2.34	35.5±1.65	32.3±1.25	31.1±0.04	40.4±0.51	45.0±0.12	35.8±1.00	50.9±1.66	14.8% ↑	93.5% ↓
	O+B	35.6±0.94	43.7±1.12	40.6±0.24	39.2±0.06	32.6±0.16	39.6±0.24	50.1±0.31	39.9±1.69	57.7±0.98	14.0% ↑	91.1% ↓
	O+D	45.0±2.57	55.1±1.17	51.5±2.66	52.2±3.10	36.9±0.04	49.0±0.51	61.7±0.34	53.7±2.24	72.3±2.27	17.1% ↑	92.2% ↓
	O+W+N+B+D	77.3±0.49	81.2±1.53	80.4±0.81	75.3±0.41	30.0±0.05	39.8±0.71	50.8±0.41	47.8±6.64	94.8±2.74	13.6% ↑	95.3% ↓
HAR	A	52.4±3.67	55.0±3.93	52.9±2.55	48.3±2.69	54.0±0.64	60.0±0.21	61.3±0.55	58.4±0.35	69.3±1.96	14.3% ↑	99.6% ↓
	A+P	51.2±4.53	53.3±3.20	50.1±3.52	49.4±2.95	60.5±1.28	61.1±1.89	63.1±0.85	58.5±0.75	66.6±1.78	13.3% ↑	99.8% ↓
	A+P+D	81.0±4.75	80.3±2.35	78.7±4.37	71.0±4.27	62.2±3.58	66.8±3.97	70.1±4.28	61.1±3.25	90.3±5.09	10.0% ↑	99.7% ↓
AR	C	93.6±0.16	93.5±0.07	92.9±0.65	94.2±0.28	88.1±1.65	88.3±0.83	88.5±1.78	90.4±0.19	94.3±0.17	0.2% ↑	99.9% ↓
	C+T	89.0±0.41	89.4±0.57	89.4±0.38	90.3±0.79	86.5±0.24	88.5±0.62	88.7±0.25	90.3±0.26	91.1±1.17	0.8% ↑	99.9% ↓
	C+T+N	84.7±0.64	84.8±1.52	86.2±0.79	86.9±0.40	87.5±0.54	87.7±0.31	88.0±0.61	88.5±1.45	89.2±1.60	2.3% ↑	99.9% ↓
TC	T	73.2±2.15	73.5±1.35	75.7±4.07	73.3±2.56	79.6±0.37	79.6±0.19	79.8±0.14	73.9±2.69	83.1±2.26	7.3% ↑	99.8% ↓
	T+L	77.7±3.19	82.2±0.29	80.1±3.02	80.0±1.89	84.3±0.14	84.4±0.18	84.7±0.09	79.7±2.21	86.2±2.16	4.0% ↑	99.4% ↓

Overall Performance

Lower communication overheads compared with federated CL:

- More than 91% communication cost reduction for different tasks

Tasks	Context Category	Vanilla CL	Few-Shot CL			Federated CL			Data Enrichment		ΔAcc.	ΔComm.
			FS-KD	FS-RO	FS-PF	Fed-0.1	Fed-0.2	Fed-0.4	Random	Delta		
IC	O+W	32.7±1.49	41.7±1.78	39.2±2.13	36.9±2.87	31.8±0.24	46.4±1.65	55.1±0.42	42.5±2.42	57.7±0.54	16.0% ↑	93.7% ↓
	O+N	31.3±1.74	36.2±2.34	35.5±1.65	32.3±1.25	31.1±0.04	40.4±0.51	45.0±0.12	35.8±1.00	50.9±1.66	14.8% ↑	93.5% ↓
	O+B	35.6±0.94	43.7±1.12	40.6±0.24	39.2±0.06	32.6±0.16	39.6±0.24	50.1±0.31	39.9±1.69	57.7±0.98	14.0% ↑	91.1% ↓
	O+D	45.0±2.57	55.1±1.17	51.5±2.66	52.2±3.10	36.9±0.04	49.0±0.51	61.7±0.34	53.7±2.24	72.3±2.27	17.1% ↑	92.2% ↓
	O+W+N+B+D	77.3±0.49	81.2±1.53	80.4±0.81	75.3±0.41	30.0±0.05	39.8±0.71	50.8±0.41	47.8±6.64	94.8±2.74	13.6% ↑	95.3% ↓
HAR	A	52.4±3.67	55.0±3.93	52.9±2.55	48.3±2.69	54.0±0.64	60.0±0.21	61.3±0.55	58.4±0.35	69.3±1.96	14.3% ↑	99.6% ↓
	A+P	51.2±4.53	53.3±3.20	50.1±3.52	49.4±2.95	60.5±1.28	61.1±1.89	63.1±0.85	58.5±0.75	66.6±1.78	13.3% ↑	99.8% ↓
	A+P+D	81.0±4.75	80.3±2.35	78.7±4.37	71.0±4.27	62.2±3.58	66.8±3.97	70.1±4.28	61.1±3.25	90.3±5.09	10.0% ↑	99.7% ↓
AR	C	93.6±0.16	93.5±0.07	92.9±0.65	94.2±0.28	88.1±1.65	88.3±0.83	88.5±1.78	90.4±0.19	94.3±0.17	0.2% ↑	99.9% ↓
	C+T	89.0±0.41	89.4±0.57	89.4±0.38	90.3±0.79	86.5±0.24	88.5±0.62	88.7±0.25	90.3±0.26	91.1±1.17	0.8% ↑	99.9% ↓
	C+T+N	84.7±0.64	84.8±1.52	86.2±0.79	86.9±0.40	87.5±0.54	87.7±0.31	88.0±0.61	88.5±1.45	89.2±1.60	2.3% ↑	99.9% ↓
TC	T	73.2±2.15	73.5±1.35	75.7±4.07	73.3±2.56	79.6±0.37	79.6±0.19	79.8±0.14	73.9±2.69	83.1±2.26	7.3% ↑	99.8% ↓
	T+L	77.7±3.19	82.2±0.29	80.1±3.02	80.0±1.89	84.3±0.14	84.4±0.18	84.7±0.09	79.7±2.21	86.2±2.16	4.0% ↑	99.4% ↓

System Scalability

Latency (ms)

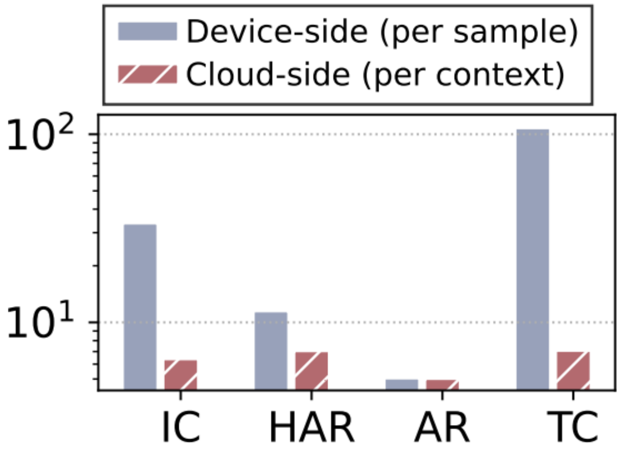
- **Device-Side:** 1.05 – 109 **ms**/sample
- **Cloud-Side:** 2.56 – 7.15 **ms**/context

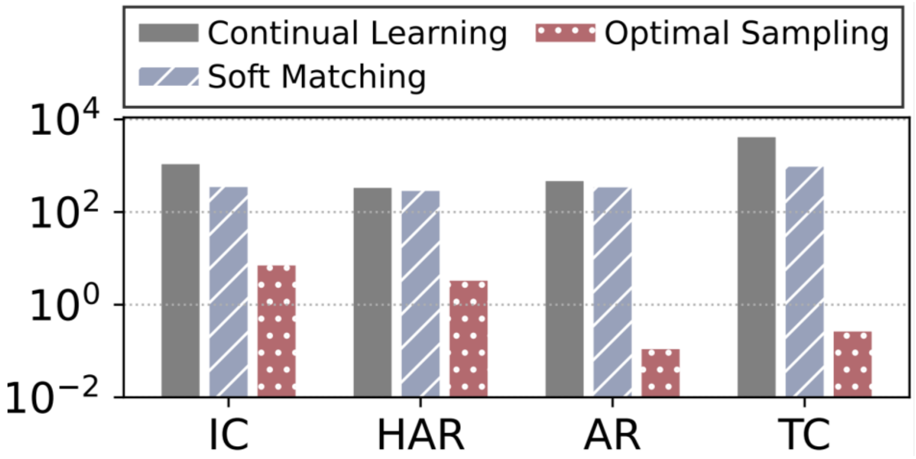
Memory (MB)

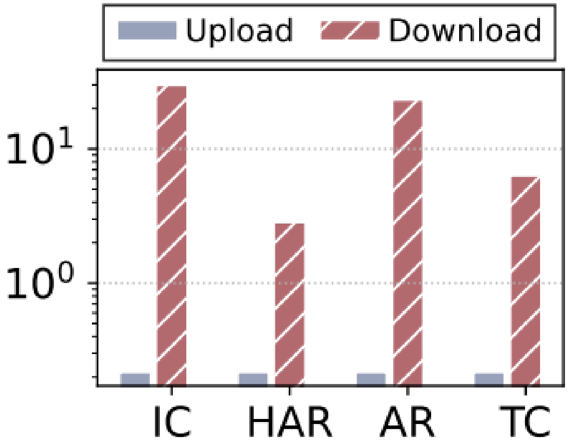
- **Device-Side: No increased peak memory footprint**
- **Cloud-Side:** 0.12 – 7.8 MB extra memory cost

Communication (KB)

- **Upload:** ≤1**KB** for directory weights
- **Download:** 2.89 – 30.4 **KB** for enriched data







System Scalability

Latency (ms)

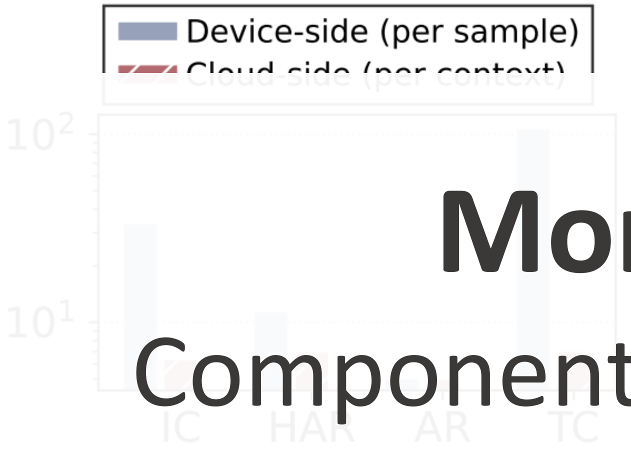
- **Device-Side:** 1.05 – 109 **ms**/sample
- **Cloud-Side:** 2.56 – 7.15 **ms**/context

Memory (MB)

- **Device-Side: No increased peak memory footprint**
- **Cloud-Side:** 0.12 – 7.8 MB extra memory cost

Communication (KB)

- **Upload:** ≤1**KB** for directory weights
- **Download:** 2.89 – 30.4 **KB** for enriched data



More Details in Our Paper:

Component-Wise Analysis, Sensitivity Analysis, Different Impacts on New and Past Contexts

Conclusion

Problem

- The **data bottleneck** in on-device continual learning
- Existing solutions show ineffectiveness and inefficiency

Solution

- Delta, a cloud-assisted data enrichment framework that simultaneously achieves **privacy, efficiency and effectiveness**

Result

- Delta shows **superior continual learning performance** in different tasks with varied data modalities with **marginal system overheads**

Conclusion

Problem

- The **data bottleneck** in on-device continual learning
- Existing solutions show ineffectiveness and inefficiency

Solution

- Delta, a cloud-assisted data enrichment framework that simultaneously achieves **privacy, efficiency and effectiveness**

Result

- Delta shows **superior continual learning performance** in different tasks with varied data modalities with **marginal system overheads**

Thank You for Your Attention !

Chen Gong
gongchen@sjtu.edu.cn