Capturing the Moment: Lightweight Similarity Computations

Georgios Damaskinos, Rachid Guerraoui, Rhicheek Patra

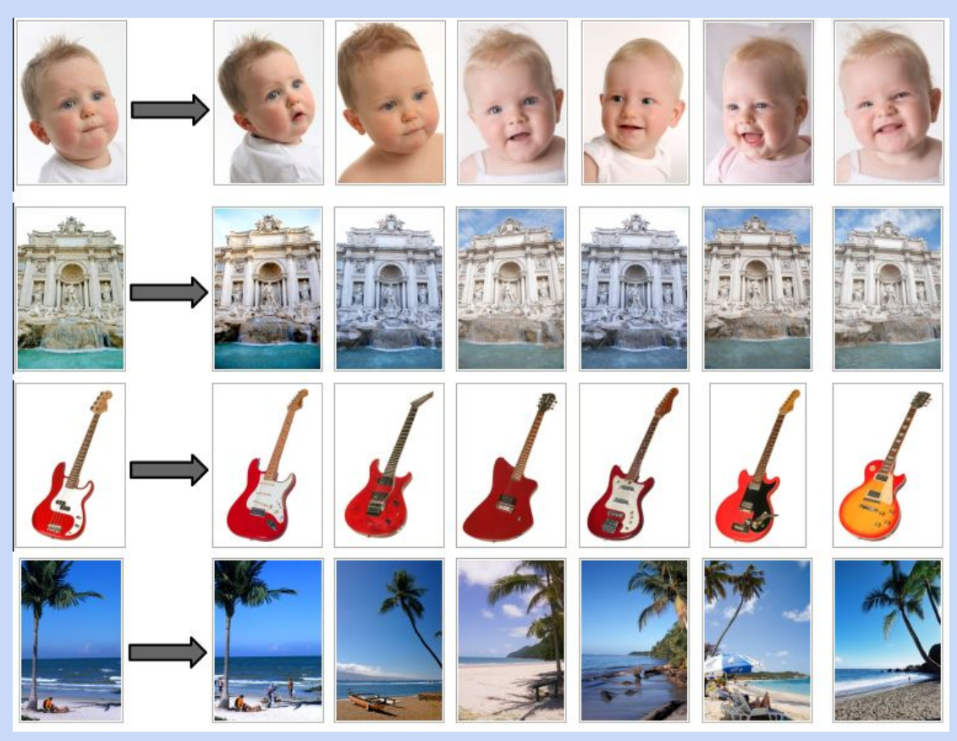
EPFL

Context

Similarity Computing

- Collaborative Filtering
- Similarity Search
- Trust / Distrust

...



Motivation

Big Data

- Changing patterns
- GBs of data / day

Challenge

Scalable + **Accurate**

Algorithms + Systems

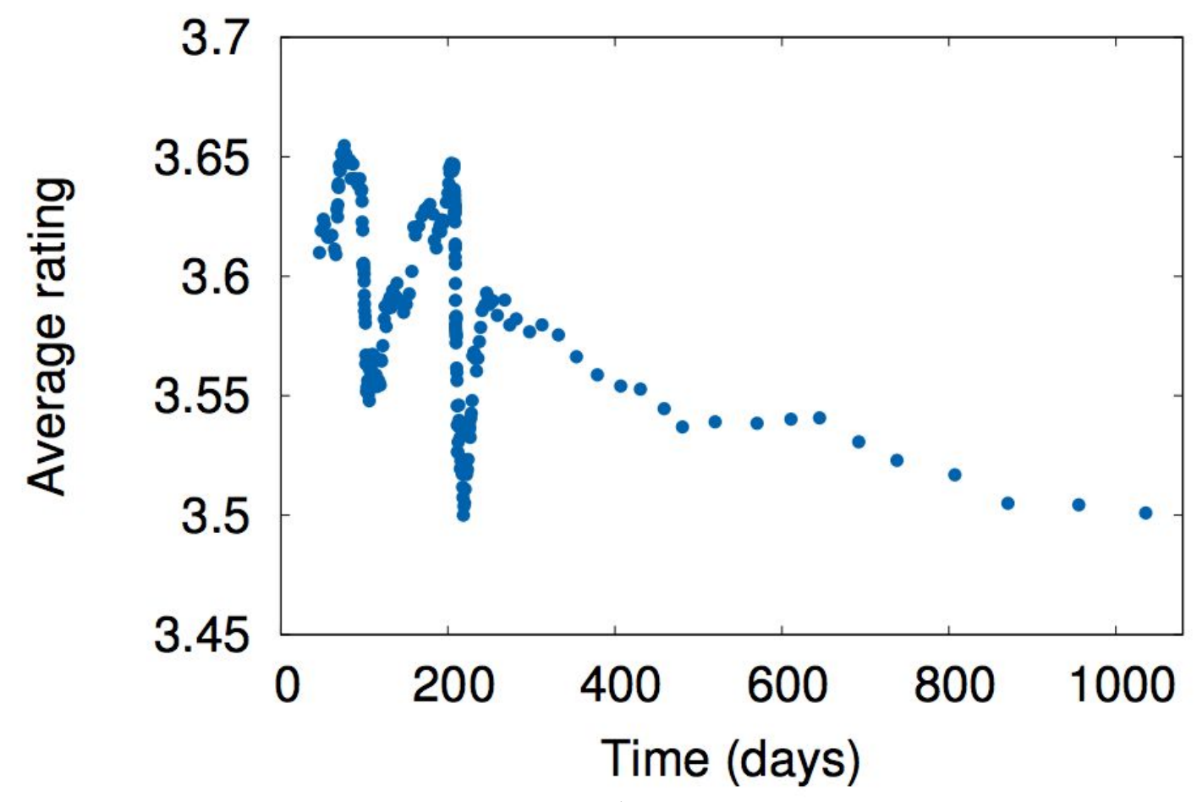
for online Machine Learning

Accuracy

I-SIM

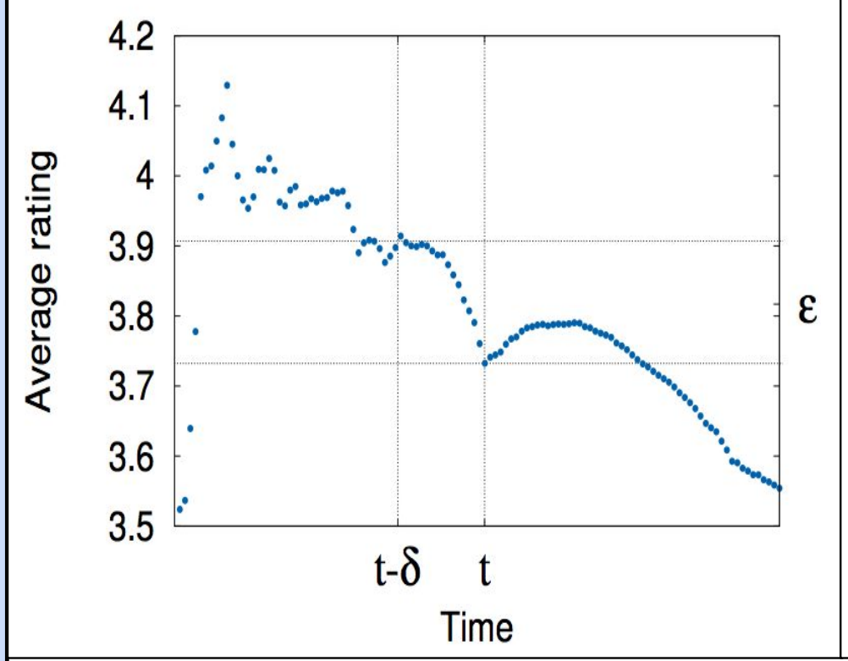
Scalability

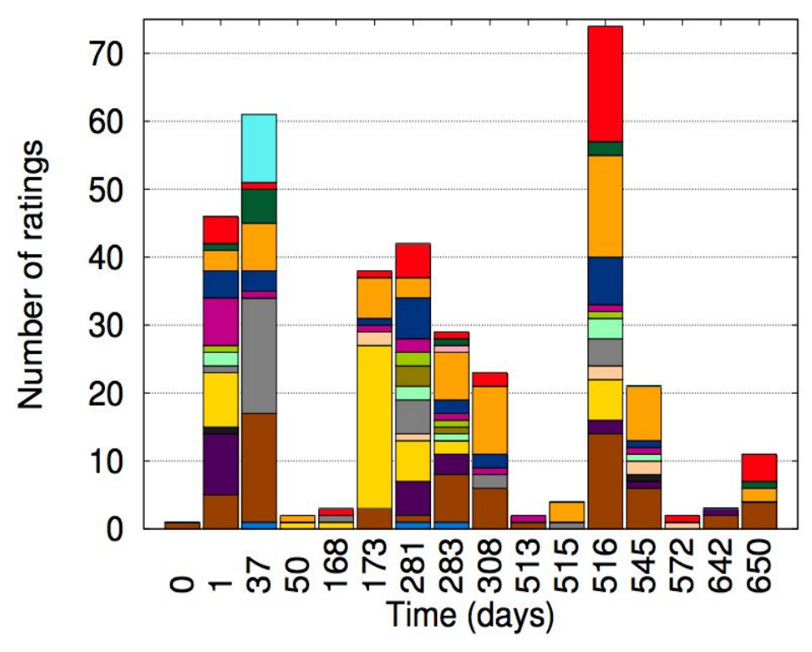
Temporality



Behavioral Drift

Preference Drift





$$S_{ij}(t) = \frac{P_{ij}(t)}{\sqrt{Q_i(t)}\sqrt{Q_j(t)}}$$

$$P_{ij}(t) = \underbrace{\Delta P_{ij}(t) + e^{-2\alpha} P_{ij}(t-1)}_{\text{standard component}} - \underbrace{e^{-2\alpha}[L_{ij}(t-1) - M_{ij}(t-1)]}_{\text{adjustment component}}$$

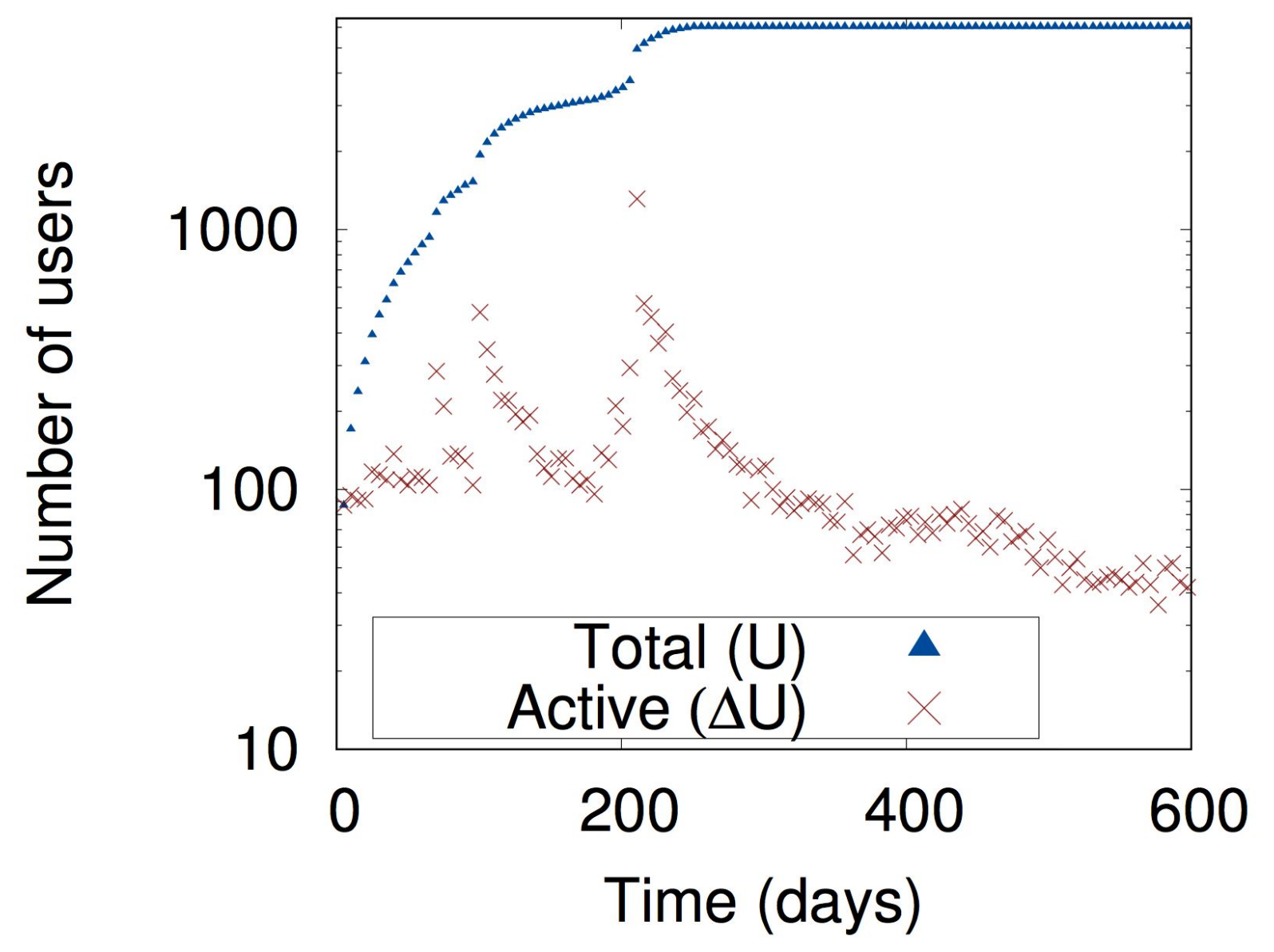
$$\mathcal{O}(|\Delta\mathcal{U}_i^t| + |\Delta\mathcal{U}_j^t|)$$

$$Q_i(t) = \underbrace{\Delta Q_i(t) + e^{-2\alpha} Q_i(t-1)}_{\text{standard component}} - \underbrace{e^{-2\alpha}[L_i(t-1) - M_i(t-1)]}_{\text{adjustment component}}$$

$$\mathcal{O}(|\Delta\mathcal{U}_i^t|)$$

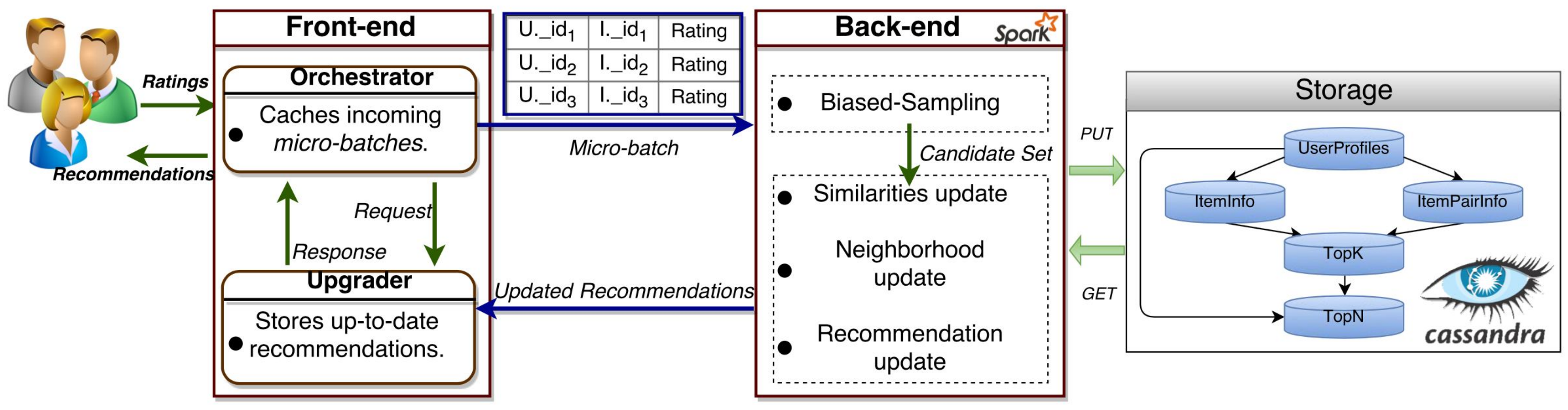
Trade-off : Storage for *L* and *M* terms

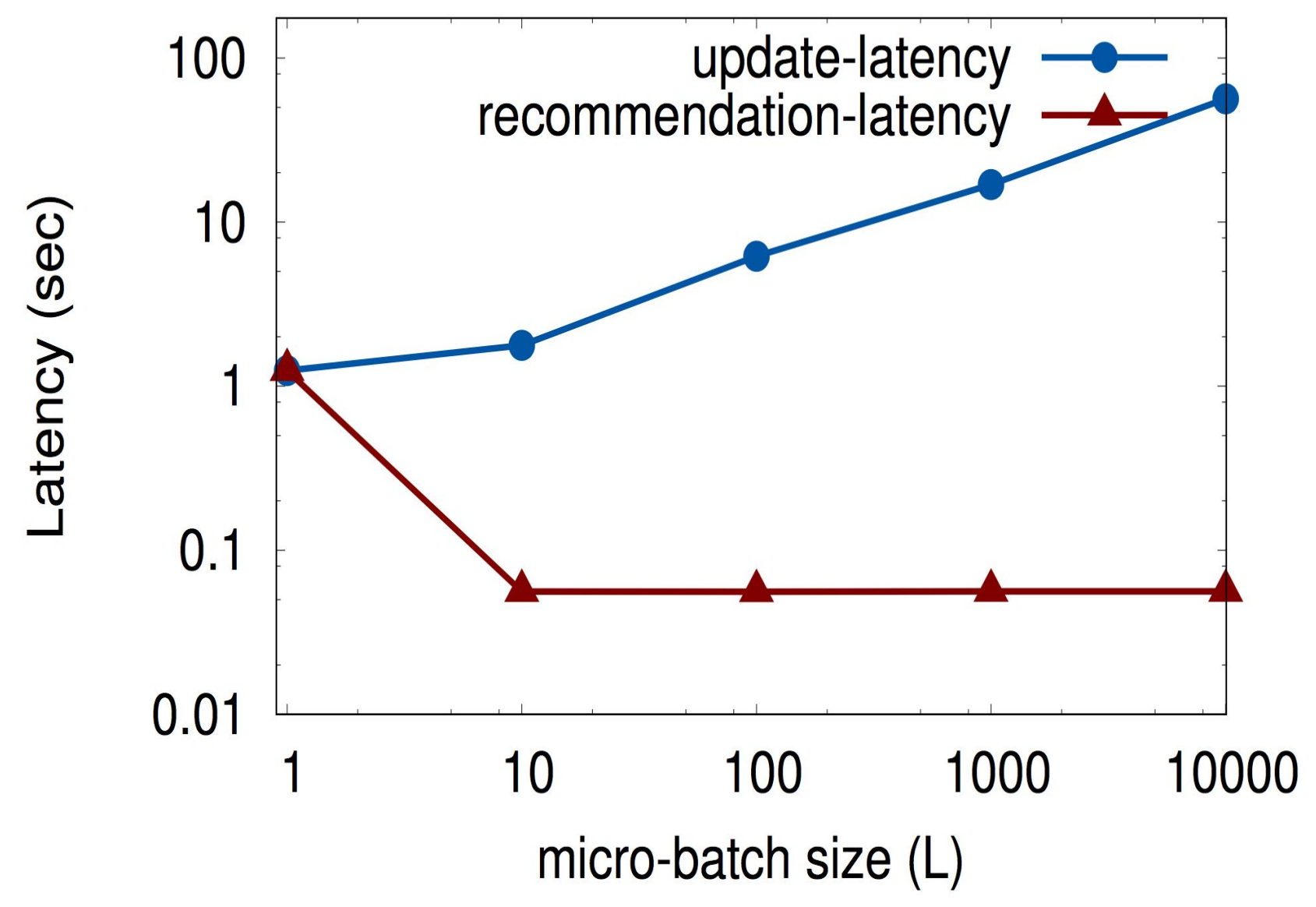
Incrementality



$$\mathcal{O}(\Delta Users)$$

SwIFT





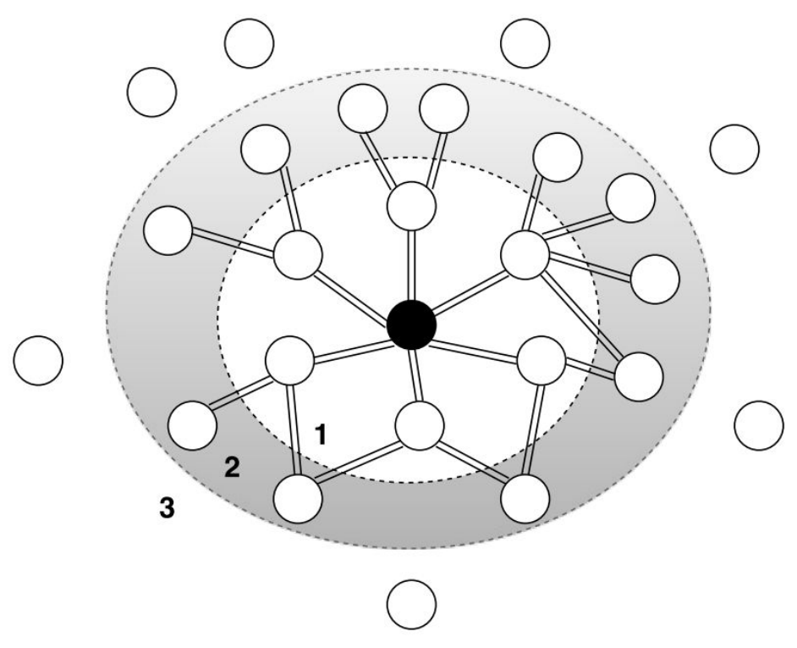
MAE

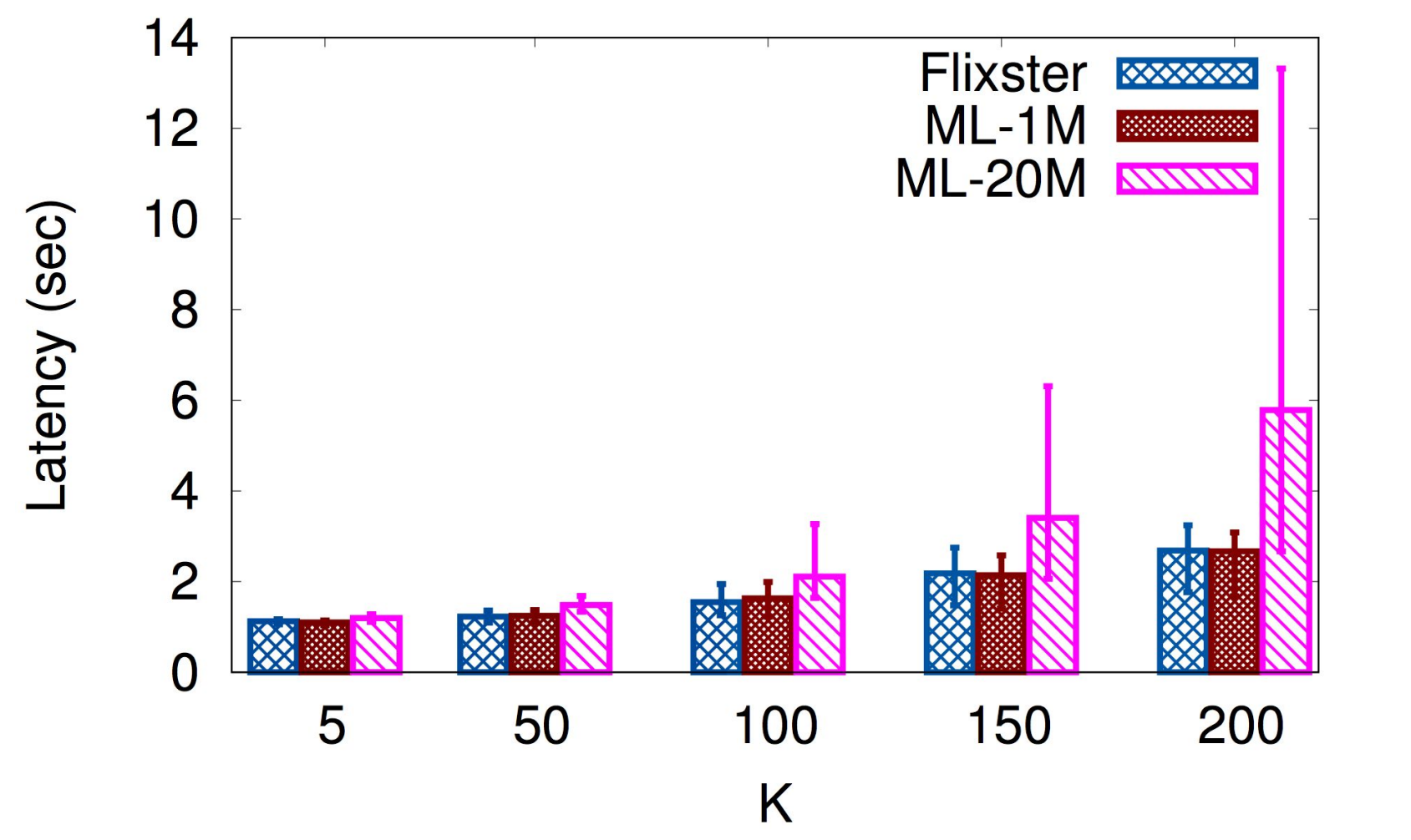
Dataset / Approach	*ML-1M*	*ML-20M*	*Flixster*
FISM	0.731	0.873	0.713
TIMESVD	0.806	0.892	0.73
ALS	0.707	0.746	**0.629**
SWIFT	**0.686**	**0.662**	0.669
TENCENTREC	0.784	0.721	0.684

- Algorithm: K-NN based
- Similarity computation: I-SIM
- Output: Top-N recommendations

✓ **Biased Sampling** ⟶ O(K^2) updates / event

✓ **Micro-batch** ⟶ *Stream* Vs *Batch* processing





I-Trust

Approach	Classification Accuracy
C-TRUST	79.21%
I-TRUST	**80.75%**

- Algorithm: K-NN based
- Similarity computation: I-SIM
- Output: Binary Classification (Trust / Distrust)

Approach	Runtime
C-TRUST	421.2 s
I-TRUST	**11.66 s**

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

Take away

Temporality → Accuracy

Incrementality → Scalability

International Conference on Data Engineering (ICDE)

IEEE, 2017, San Diego CA

Contact:

georgios.damaskinos@epfl.ch, rhicheek.patra@epfl.ch

LPD Laboratoire de Programmation Distribuée Distributed Programming Laboratory