K-Medoids for *K*-Means Seeding

James Newling & François Fleuret

Machine Learning Group,
Idiap Research Institute
& EPFL

December 5th, 2017

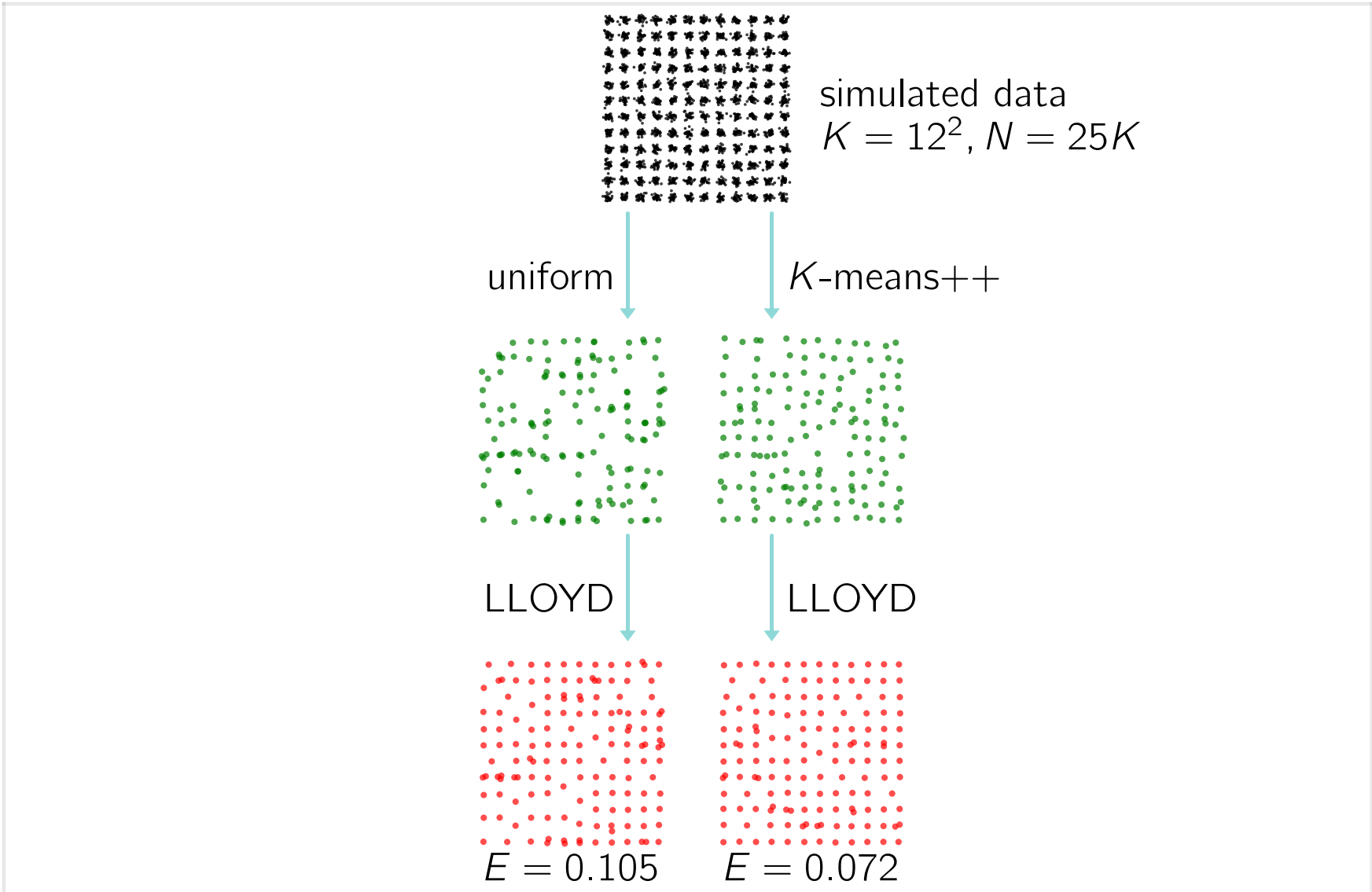
HASLERSTIFTUNG

idiap RESEARCH INSTITUTE

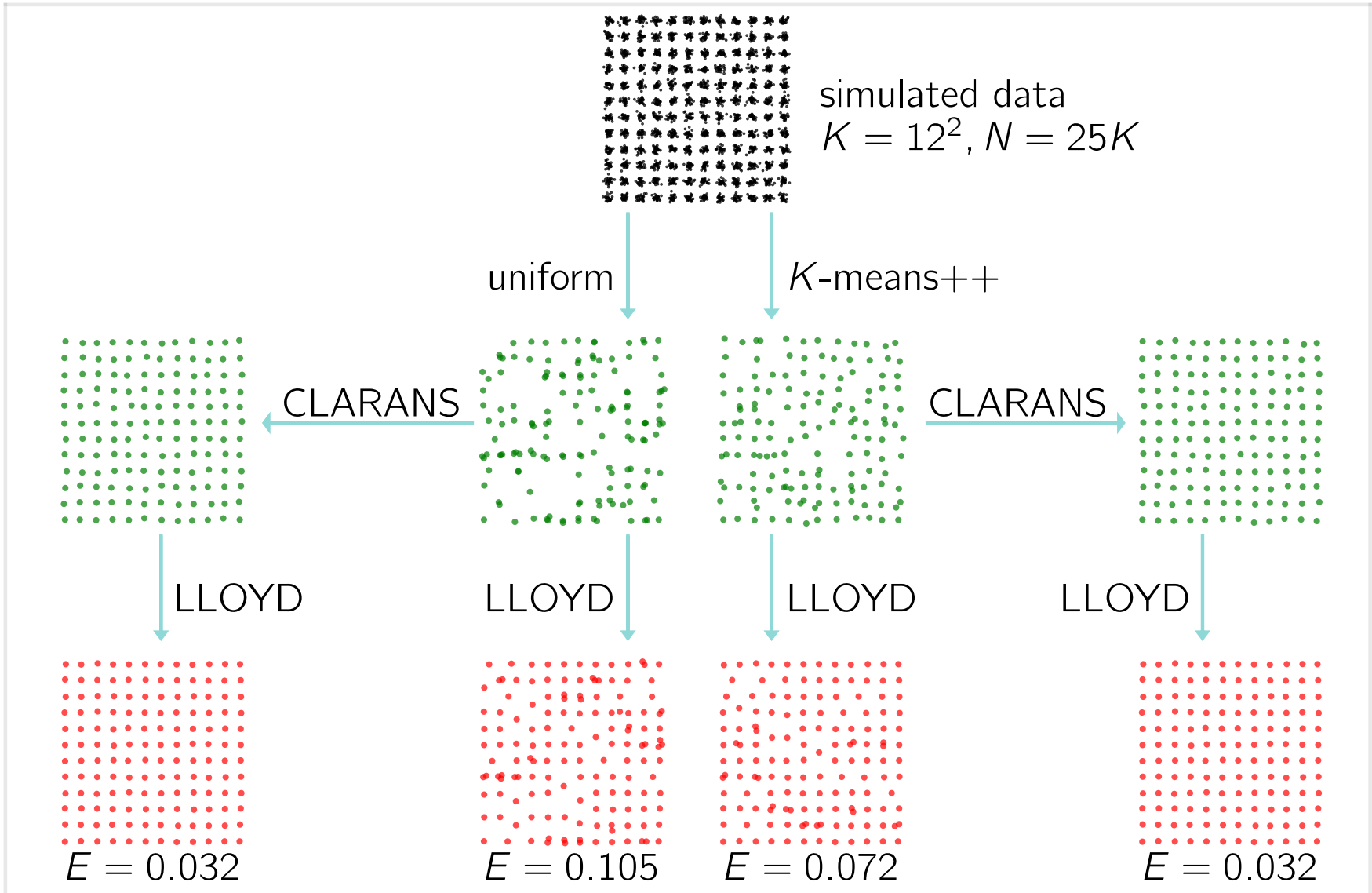
ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

The standard K-means pipeline

First: Seeding. Second: Lloyd's (a.k.a. K-means) algorithm.



The standard K-means pipeline (+CLARANS)



CLARANS of Ng and Han (1994)

1: **while** not converged **do**
2: randomly choose 1 center and 1 non-center
3: **if** swapping them decreases E **then**
4: implement the swap
5: **end if**
6: **end while**

CLARANS of Ng and Han (1994)

1: **while** not converged **do**
2: randomly choose 1 center and 1 non-center
3: **if** swapping them decreases E **then**
4: implement the swap
5: **end if**
6: **end while**

Avoids local minima of LLOYD by,

- long-range swaps
- updating centers and samples *simultanously*.

CLARANS of Ng and Han (1994)

1: **while** not converged **do**
2: randomly choose 1 center and 1 non-center
3: **if** swapping them decreases E **then**
4: implement the swap
5: **end if**
6: **end while**

Avoids local minima of LLOYD by,

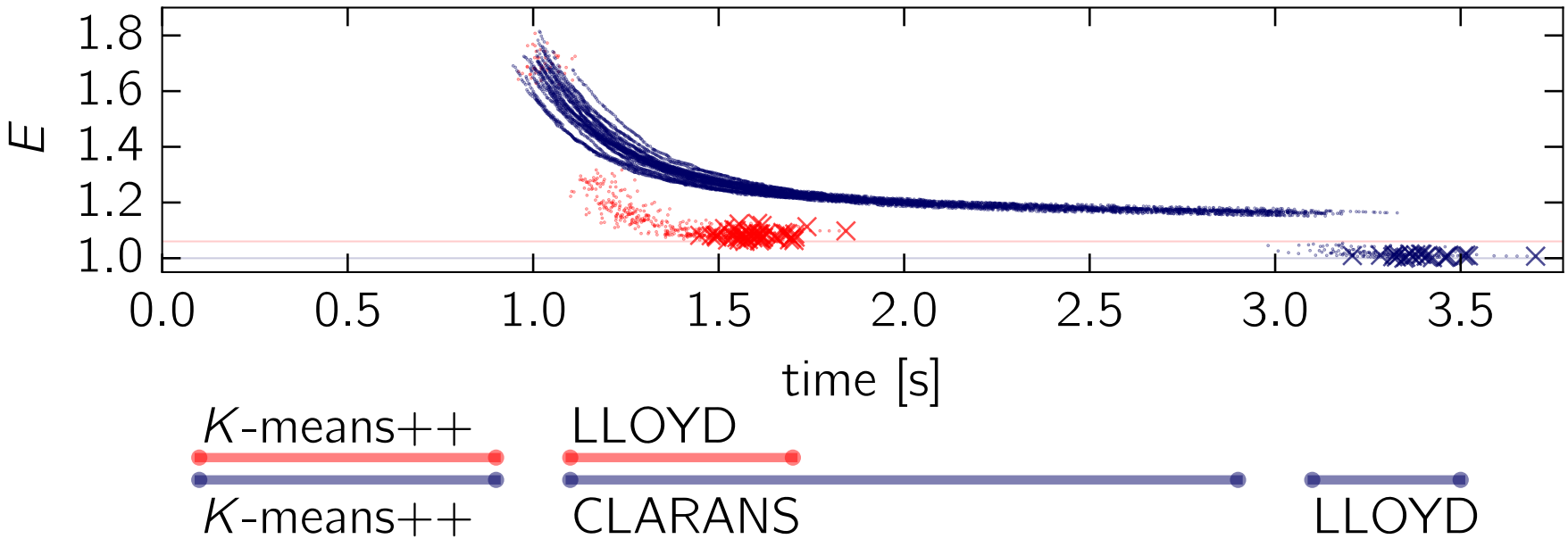
- long-range swaps
- updating centers and samples *simultanously*.

We present algorithmic improvements, where

- computing new E is $O(N/K)$
- implementing swap is $O(N)$.

Results

- RNA dataset, $d = 8$, $N = 16 \times 10^4$, $K = 400$
- 50 runs without CLARANS (red), 24 runs with (blue).



- On 16 datasets, geometric mean improvement is 3%.

CLARANS with Levenshtein metric for sequence data, $l_0, l_1, \ldots, l_\infty$ for sparse/dense vectors, many others, on github.

The end

james.newling@idiap.ch