# K-Medoids for K-Means Seeding

James Newling & François Fleuret

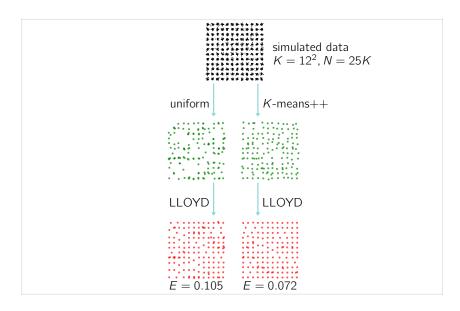
Machine Learning Group, Idiap Research Institute & EPFL

June 20th, 2016

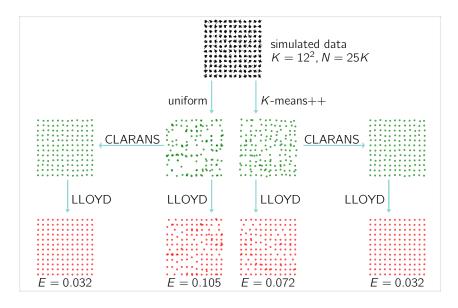




## K-Means seeding, adding CLARANS



# K-Means seeding, adding CLARANS



## CLARANS of Ng and Han (1994)

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

Avoids many local minima through,

- 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** swap decreases *E* **then**
- 4: implement swap
- 5: end if
- 6: end while

Avoids many local minima through,

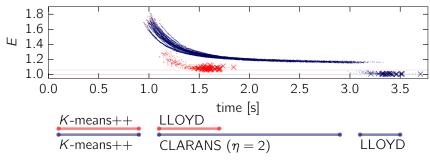
- long-range swaps
- updating centers and samples simultanously.

We present several algorithmic improvements, which make

- swap proposal O(N/K)
- swap implementation O(N)

#### Results

- RNA dataset, d = 8,  $N = 16 \times 10^4$ , K = 400
- Several runs with and without CLARANS.



• On 16 datasets, geometric mean improvent is 3%.

CLARANS with Levenshtein metric for sequence data,  $l_0, l_1, \ldots, l_{\infty}$  for sparse/dense vectors, fast K-means++, LLOYD, many others, on github



iames newling@idian.ch