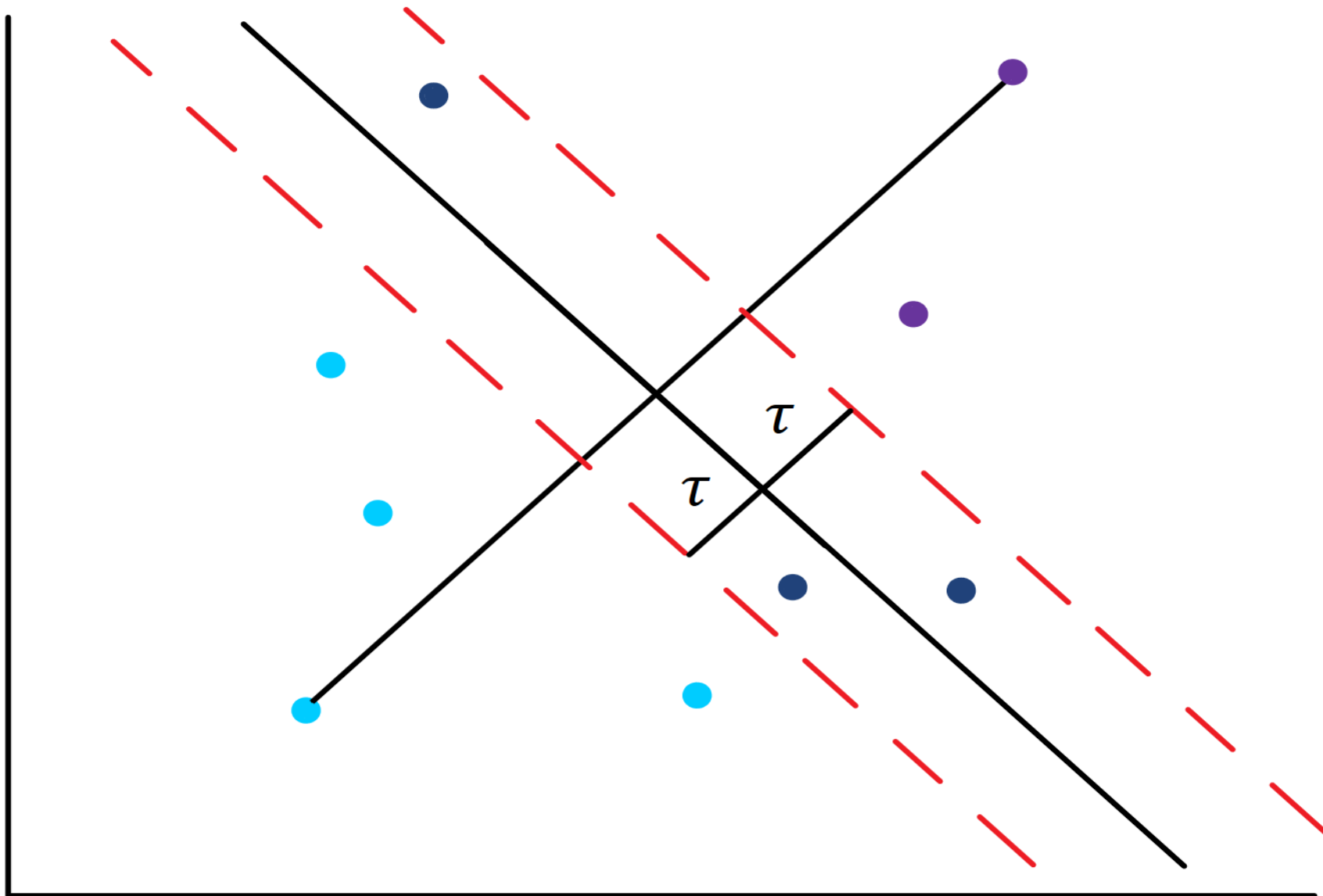# Distributed KNN

Distributed KNN

Data (size n) —*sample with p = 1/m*→ **Sampled Data (size 1/m)**

$O(n)$

Where:
d = *depth of Top HST*
p = *feature space size*
m = *defined s.t. n/m fits on a single machine*
u = *defined s.t. um fits on a machine*
c = *number of children datasets*
u' = *terminal node size bound*
$d \ll n$
$um \ll n$
$u' > k$

use sampled data to create HST

$O((n/m)^2)$

Top Hybrid Spill Tree

Use tree to classify full dataset

$O(ndp)$

Child Datasets (size um) —*Send each child to one machine*→ **Emit c Datasets, Each E(size) = um**

One-to-all communication

Each $O((um)^2)$

On c Machines Compute HST

Each $O((u')^2)$

On Same c Machines KNN Each Terminal Point

If n fits on one machine, do all-to-one communication, then merge duplicate KNN lists and pick the best k

If not, we must use a clever ID system