Large-Scale Clustering

Scalable Techniques for Billions of Samples

⚠️ **Challenges:** Billions of samples · High dimensionality

Scalability Techniques

🌊 Mini-batch K-means

Process streaming data in small batches

scikit-learn

🔍 Approximate NN

Fast nearest neighbor search

FAISS

🌳 Hierarchical

Multi-level clustering for scalability

scipy

🗂️ Distributed

Parallel processing across clusters

MapReduce • Spark MLlib

⚡ GPU Acceleration

Fast distance computations

CUDA • cuML

📊 Sampling

Representative subset selection

CoreSets

⚖️ Critical Trade-off

Speed vs Accuracy