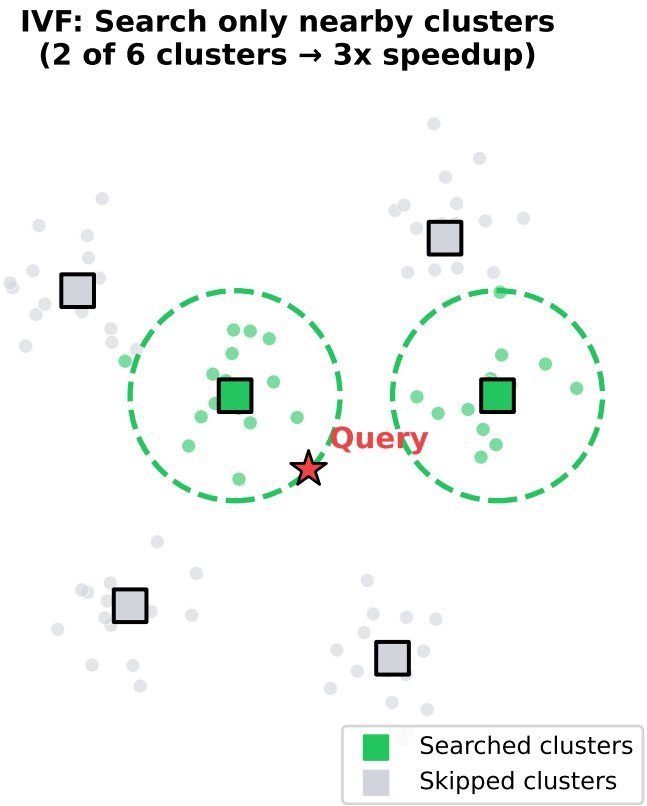
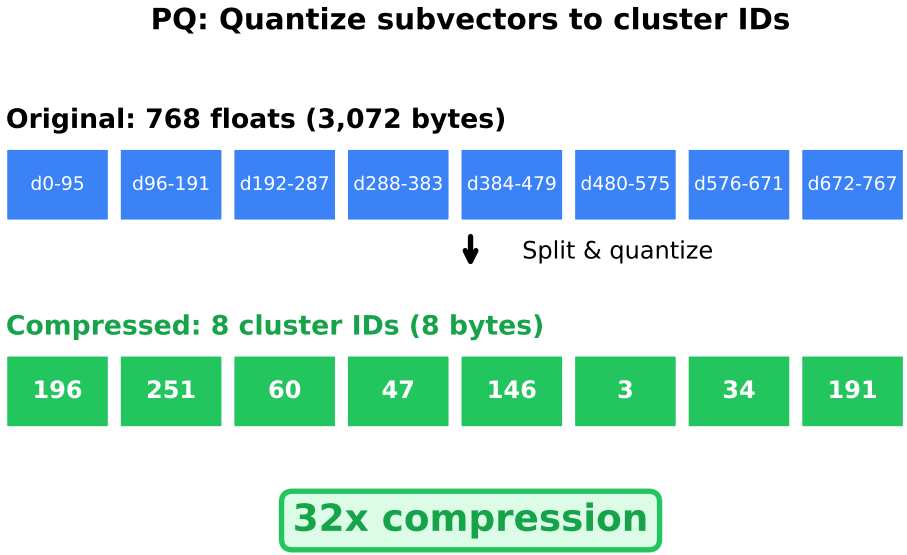

IVF: Search only nearby clusters
(2 of 6 clusters → 3x speedup)
Query
Searched clusters
Skipped clusters


PQ: Quantize subvectors to cluster IDs
Original: 768 floats (3,072 bytes)
d0-95
d96-191
d192-287
d288-383
d384-479
d480-575
d576-671
d672-767
Split & quantize
Compressed: 8 cluster IDs (8 bytes)
196
251
60
47
146
3
34
191
32x compression