Kmeans – number of clusters vary – depends on k – could wildly change performance – large k to get good representation

e.g. laid out on line – pca would do well, but need lots of clusters

Distinct clusters would perform well

Pca – might perform well even if that’s not the case – if can be linearly reduced – always do well;

If theres no linear reduction where first few components cant capure variance

Low correlation between dimensions

e.g. spherical Gaussian

or sphere – where you need all of it

k means – can represent spherical things

e.g. diagonal Gaussian –

1 cluster would do well

pca wouldn’t do well with low dimension

k means isn’t going to reduce dimensionality

if you can cluster in original dimensions

pca – as long as can project in linear small # of dimensions