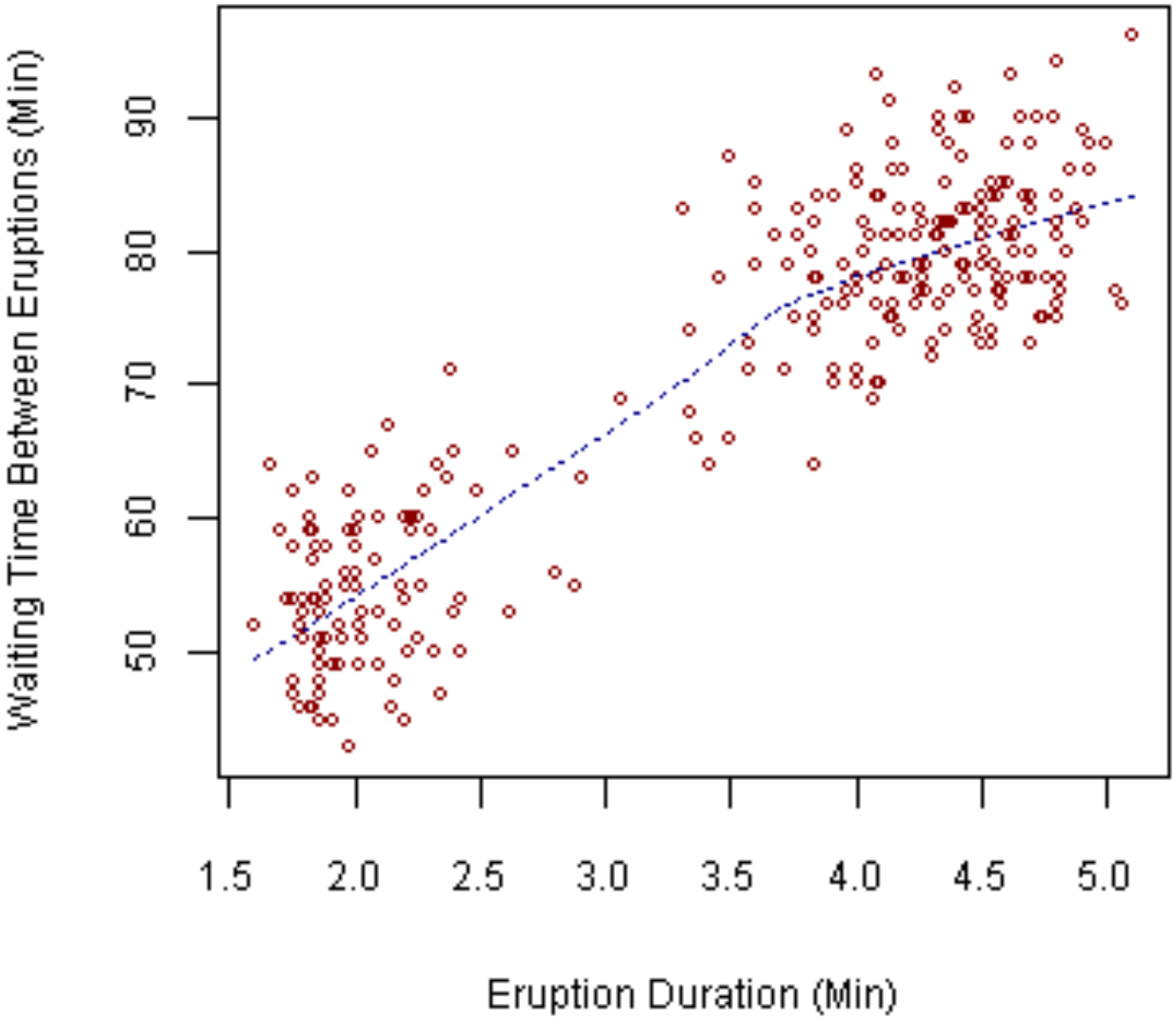
K-means Clustering

10/23/15

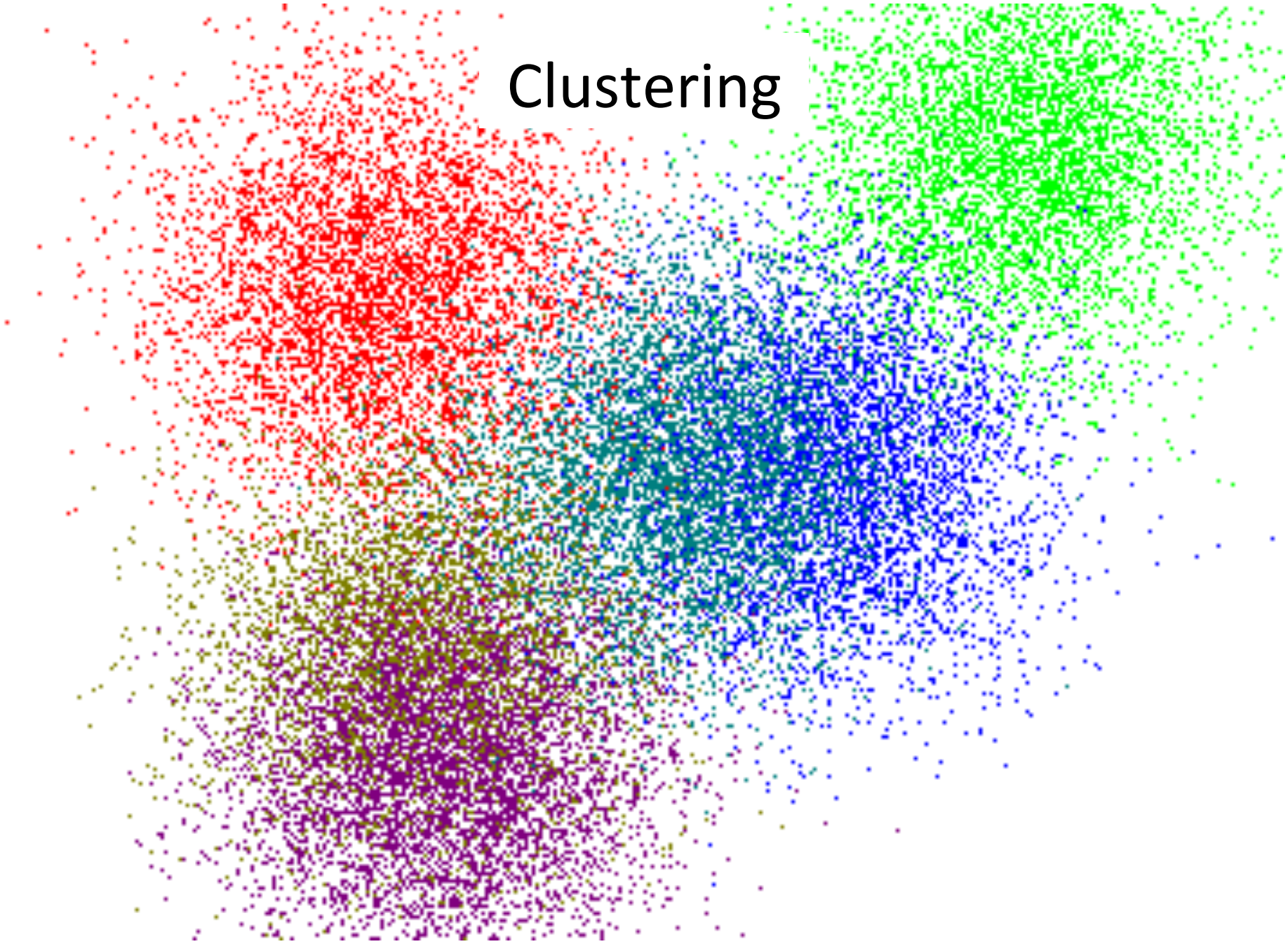
Old Faithful Eruptions



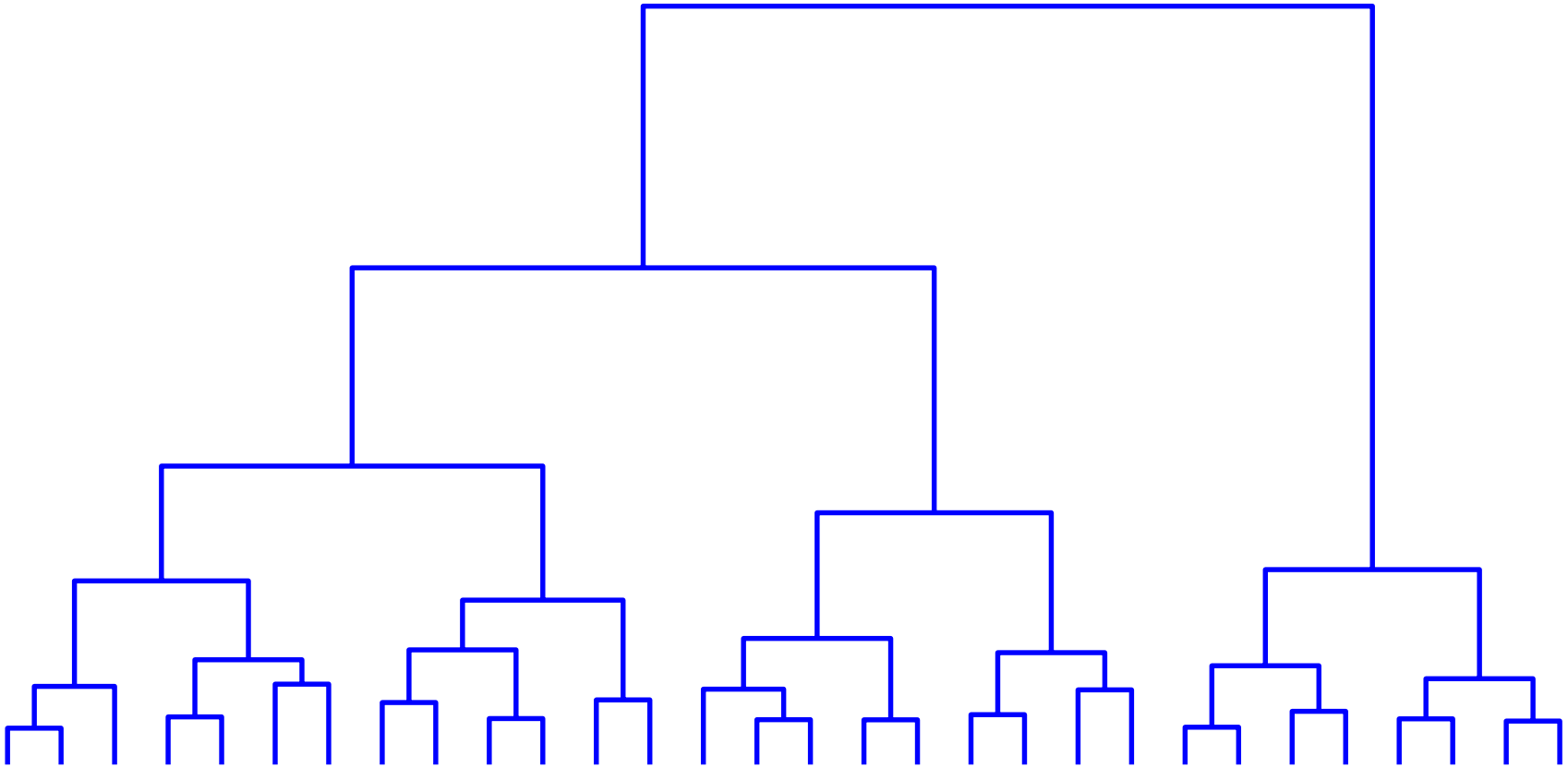


R. Robinson

Clustering



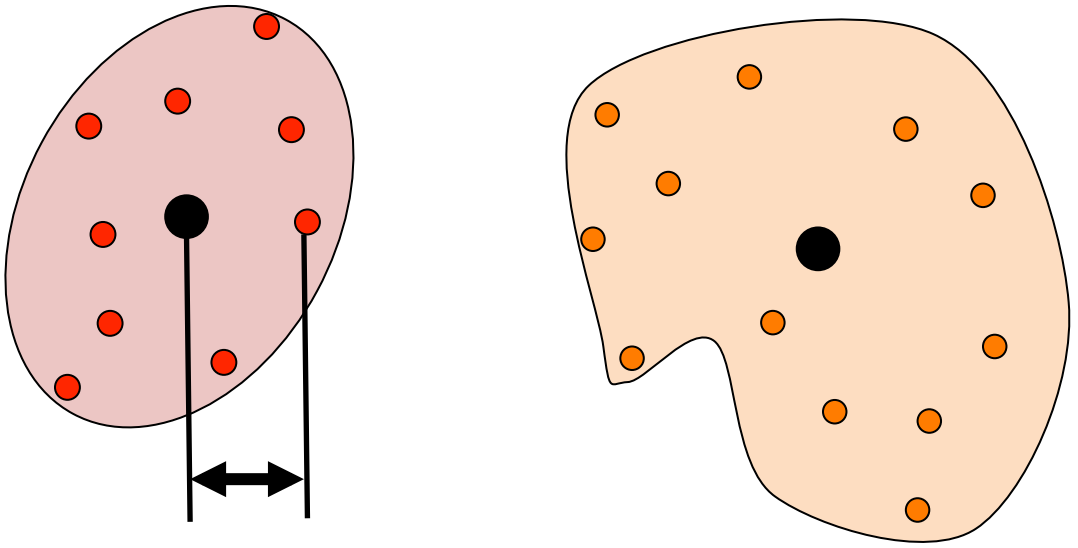
Hierarchical



K-means

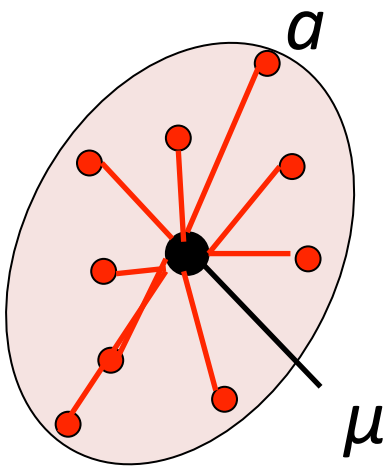
Intuition: K-clusters, with each cluster as “local” as possible

“local” means close to centroid or center of cluster



Practice

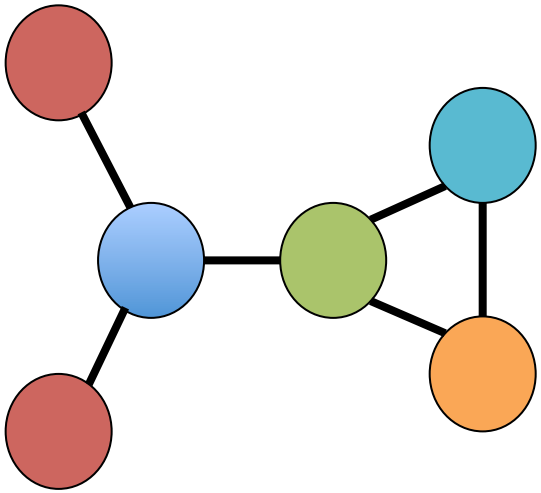
$$G = \sum_i \|\mu - a_i\|$$

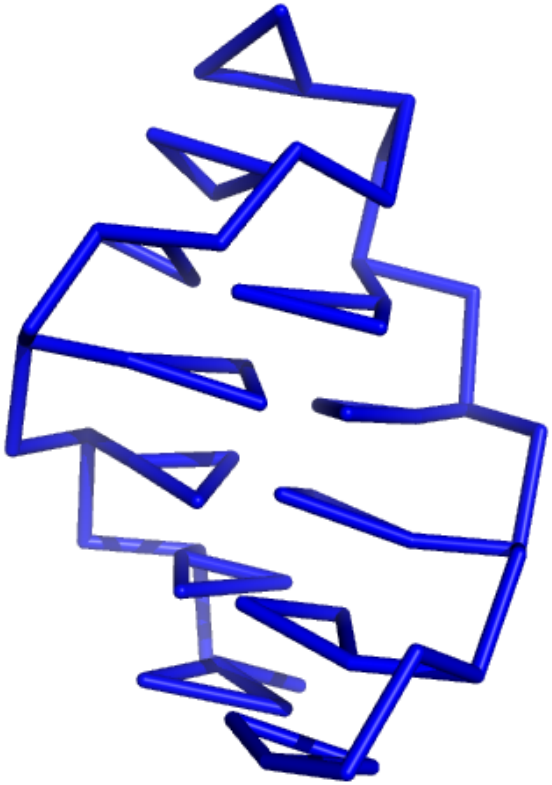


1. Randomly select K clusters μ
2. Select a closest to μ to be in cluster
3. Compute average distance in cluster
4. Select new μ closest to average distance

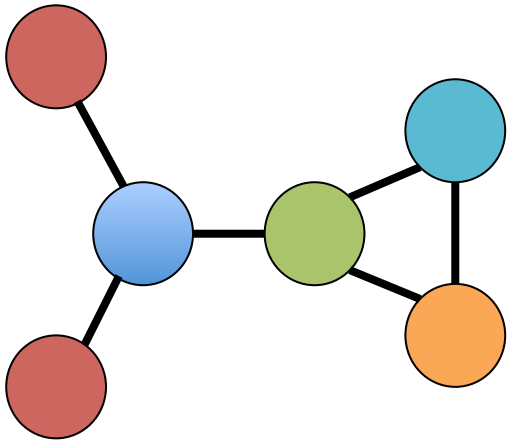
$$\frac{\sum_i^N (\mu - a_i)}{N}$$

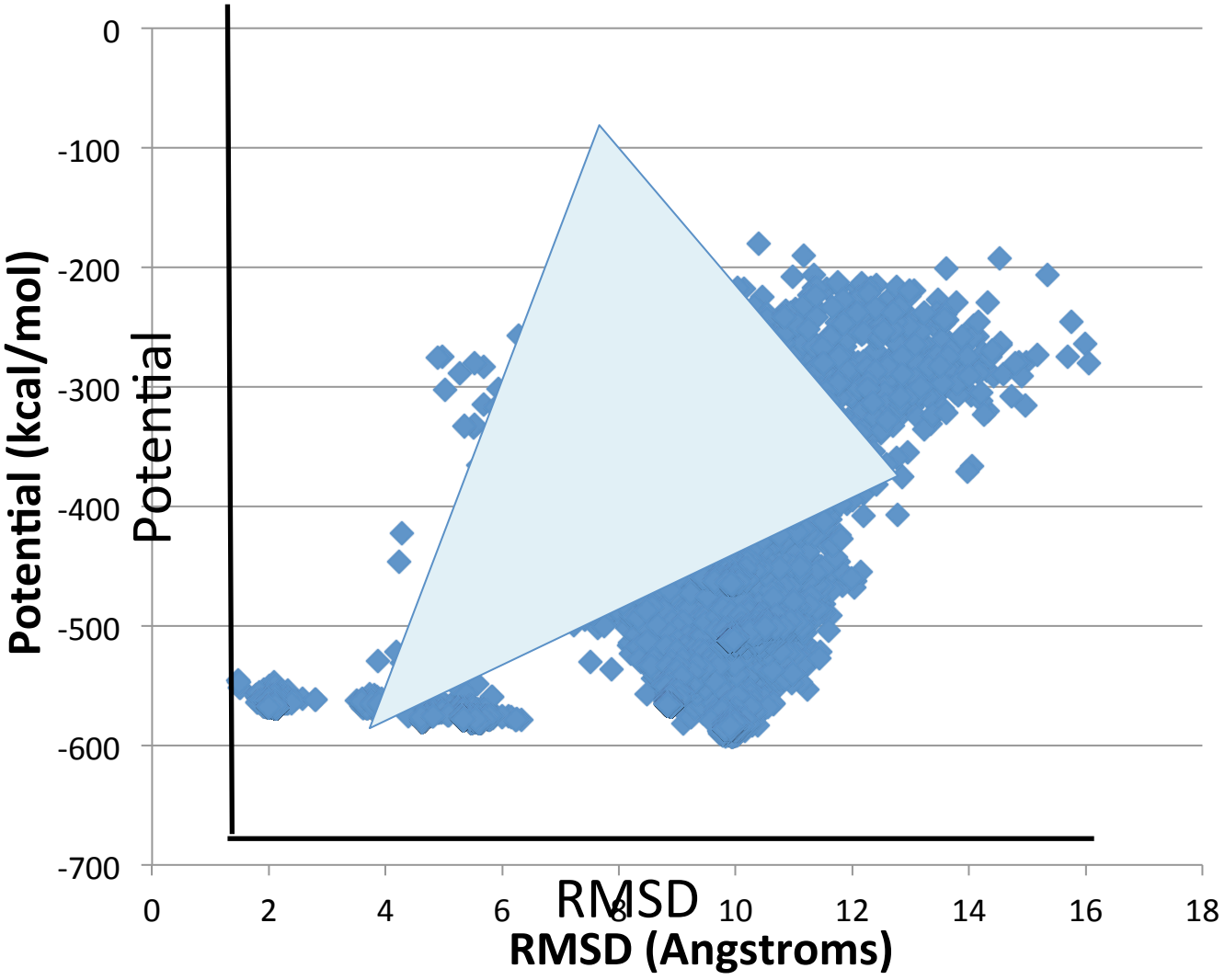
Problem



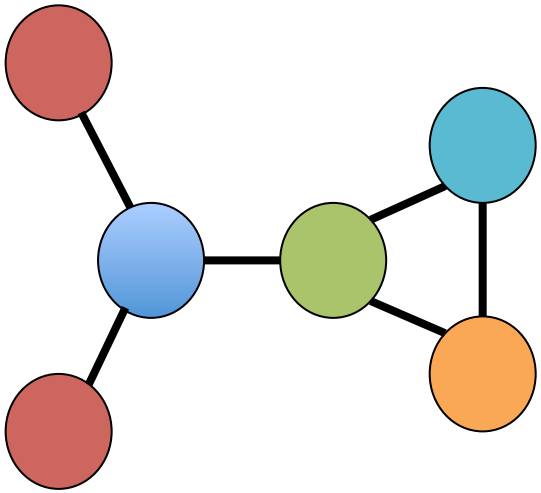


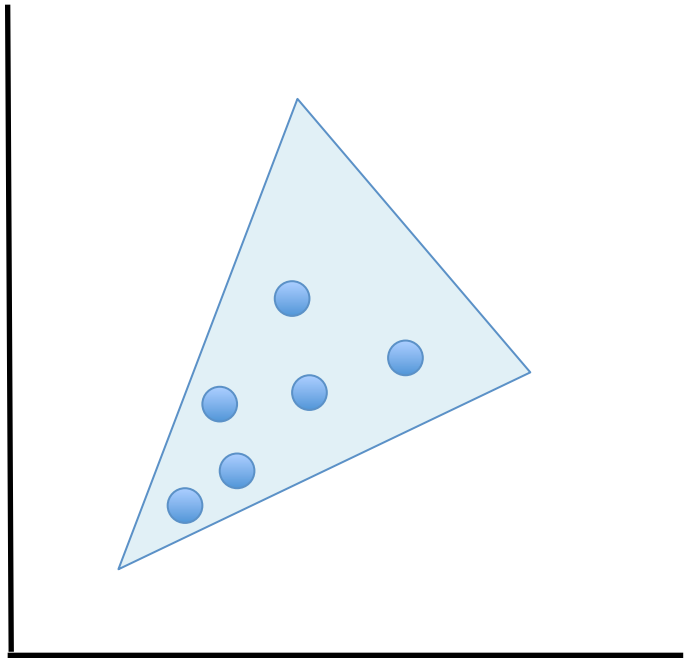
Problem




0
-100
-200
-300
-400
-500
-600
-700
Potential (kcal/mol)
Potential
0
2
4
6
8
10
12
14
16
18
RMSD
RMSD (Angstroms)

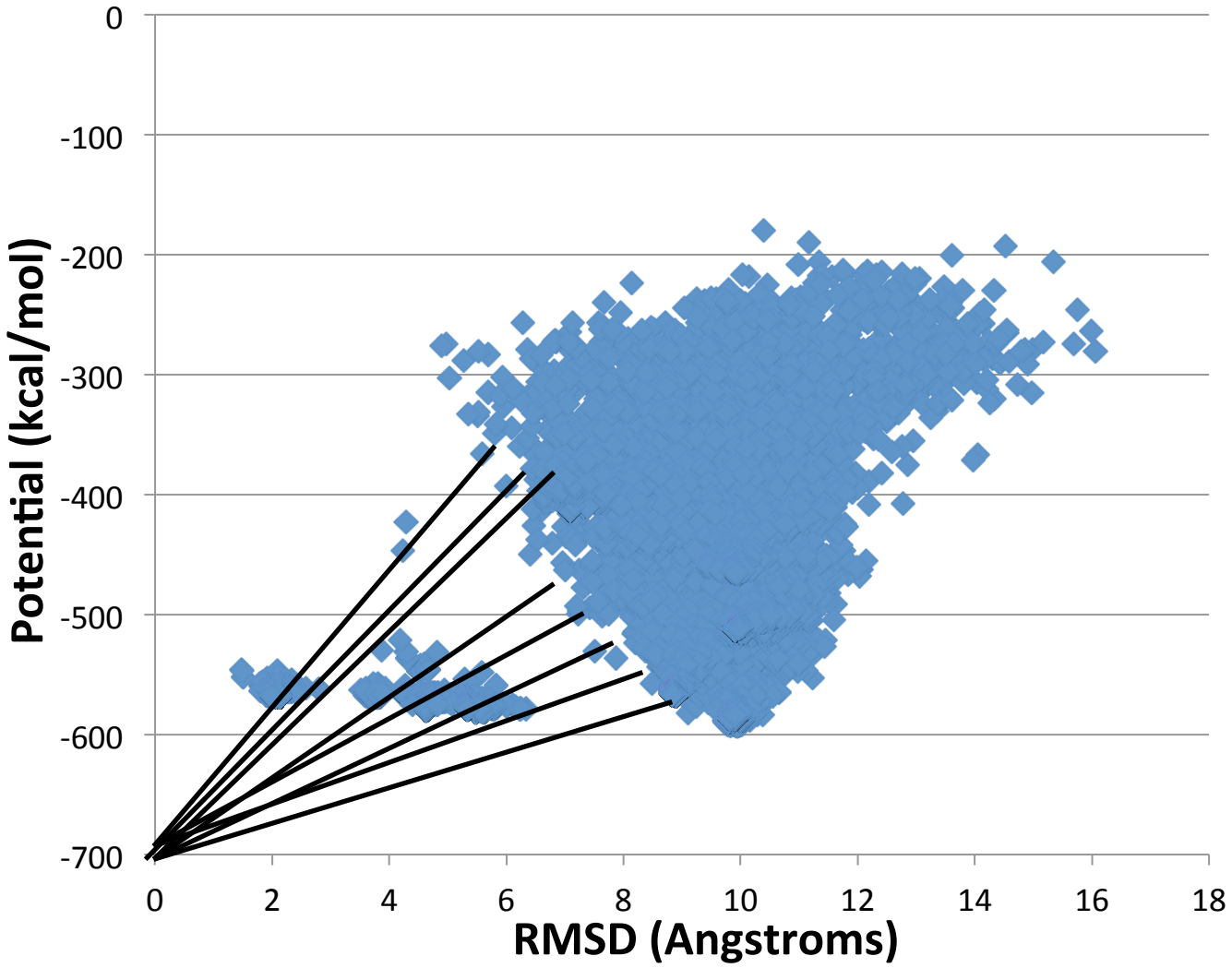
Problem



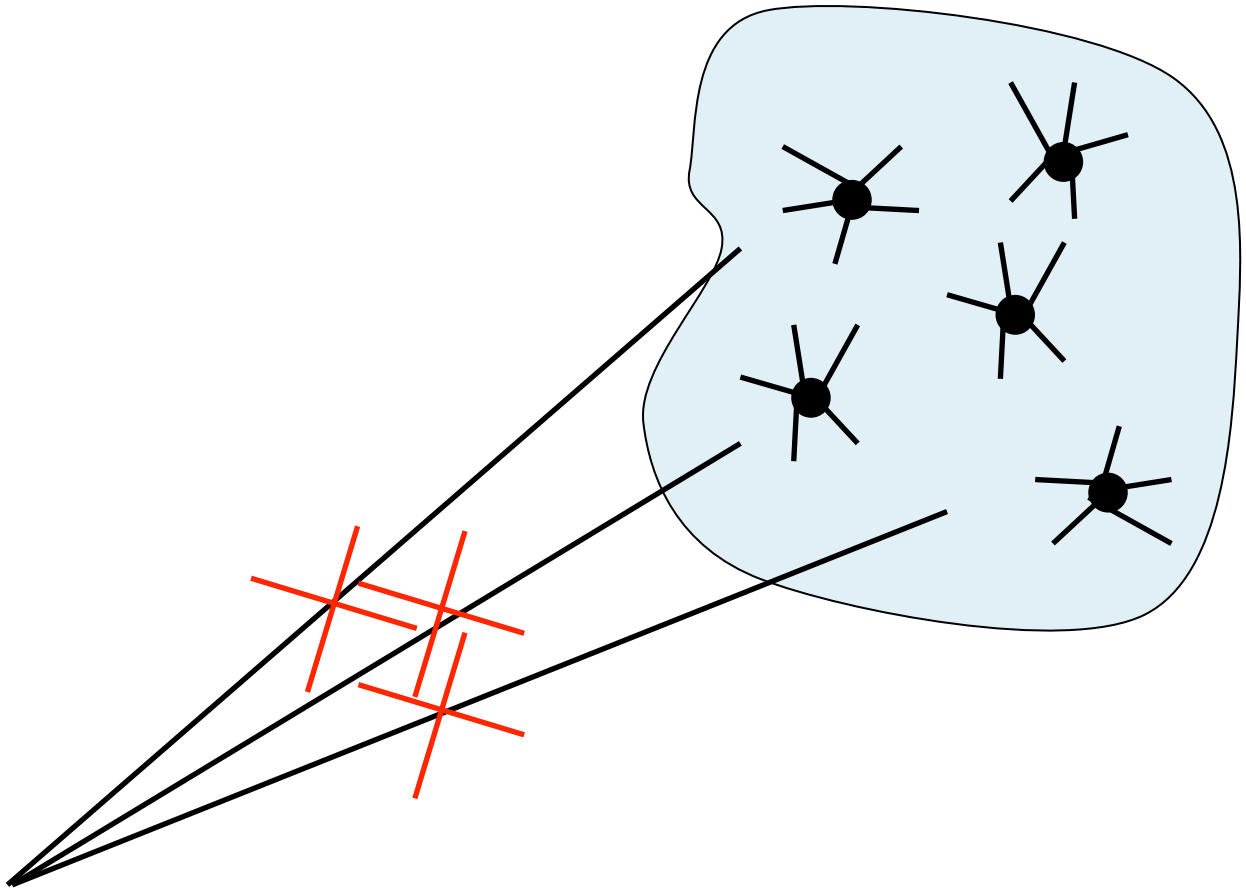


Cluster RMSD

RMSD from Crystal Structure

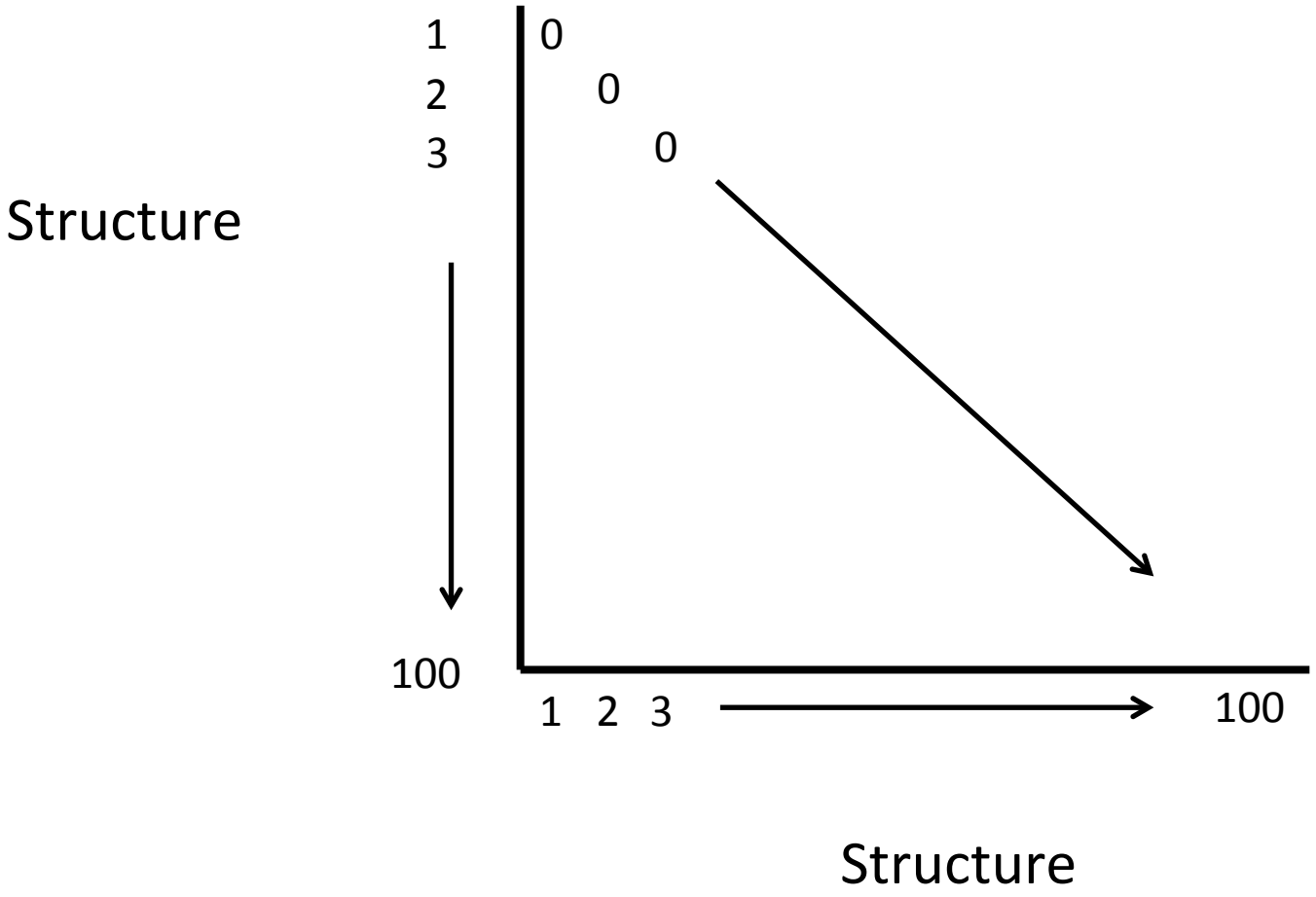


RMSD pairwise between all structures



RMSD pairwise between all structures

Distance Matrix



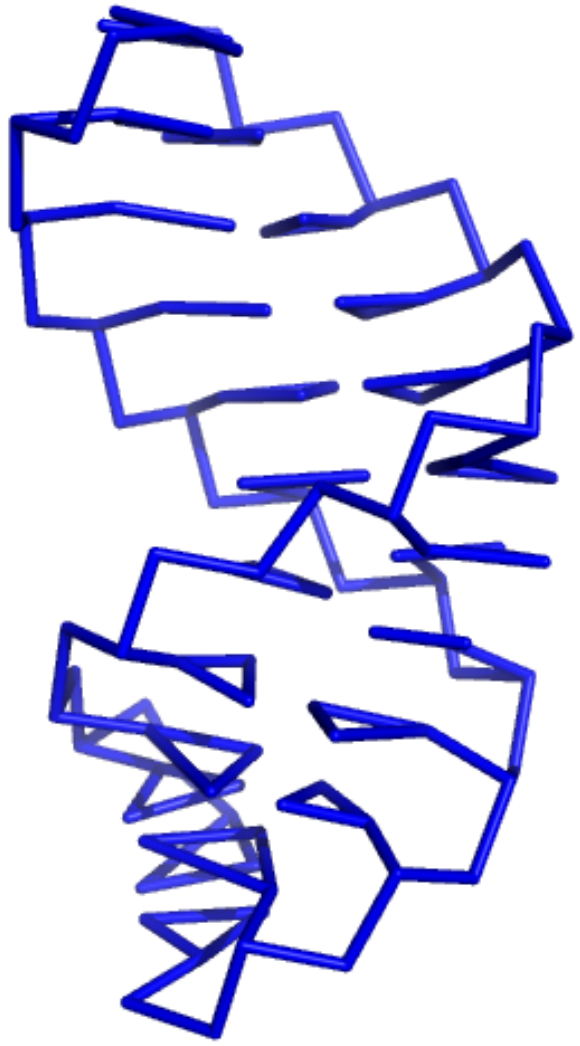
Size Problem

120,000 structures per pdb

14.4 billion rmsd computations

Each rmsd computation 1s,
≈45 computing years

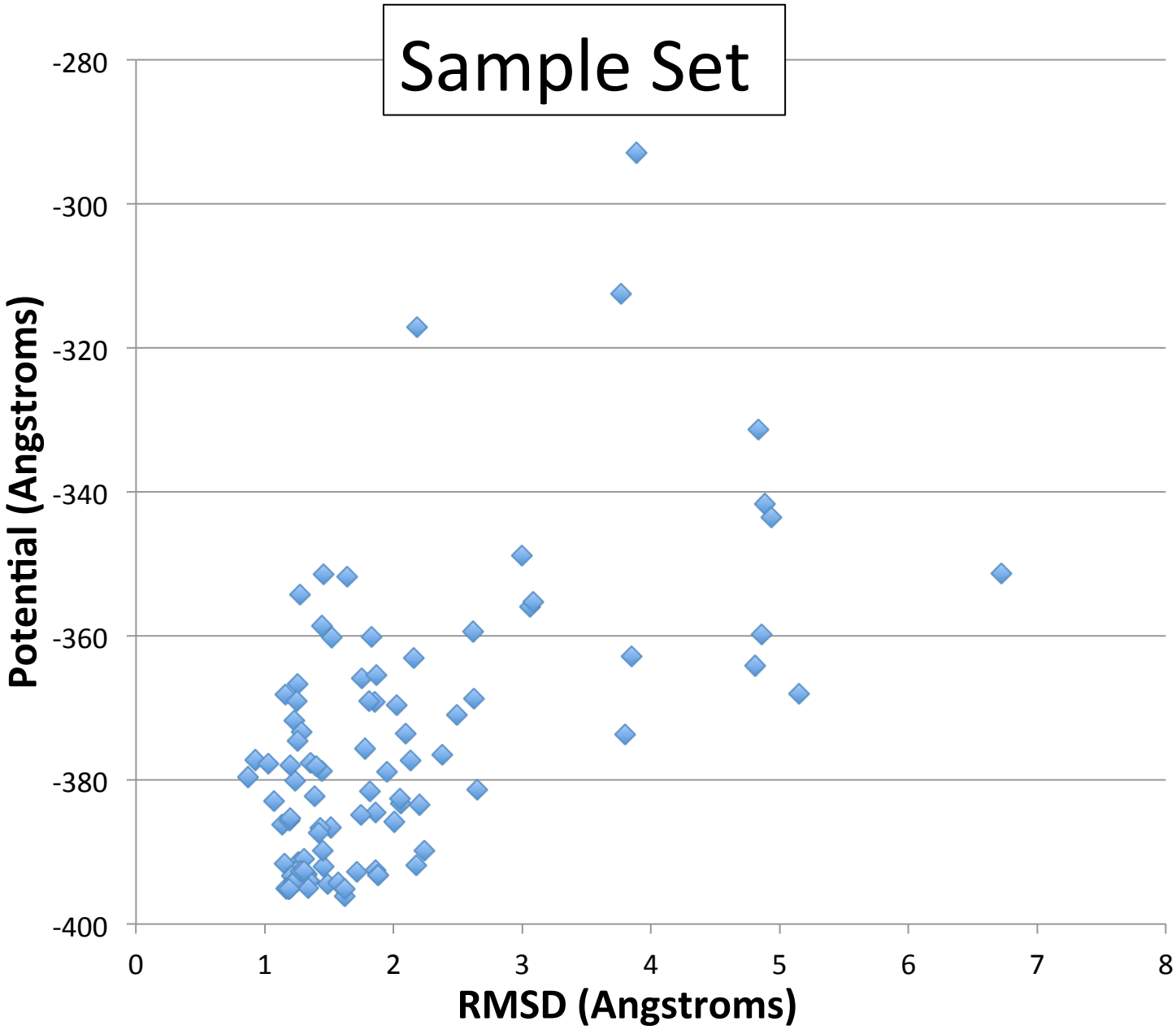
Solution (lack of): take 1 out of 100,
≈16 computing days (2 in practice)

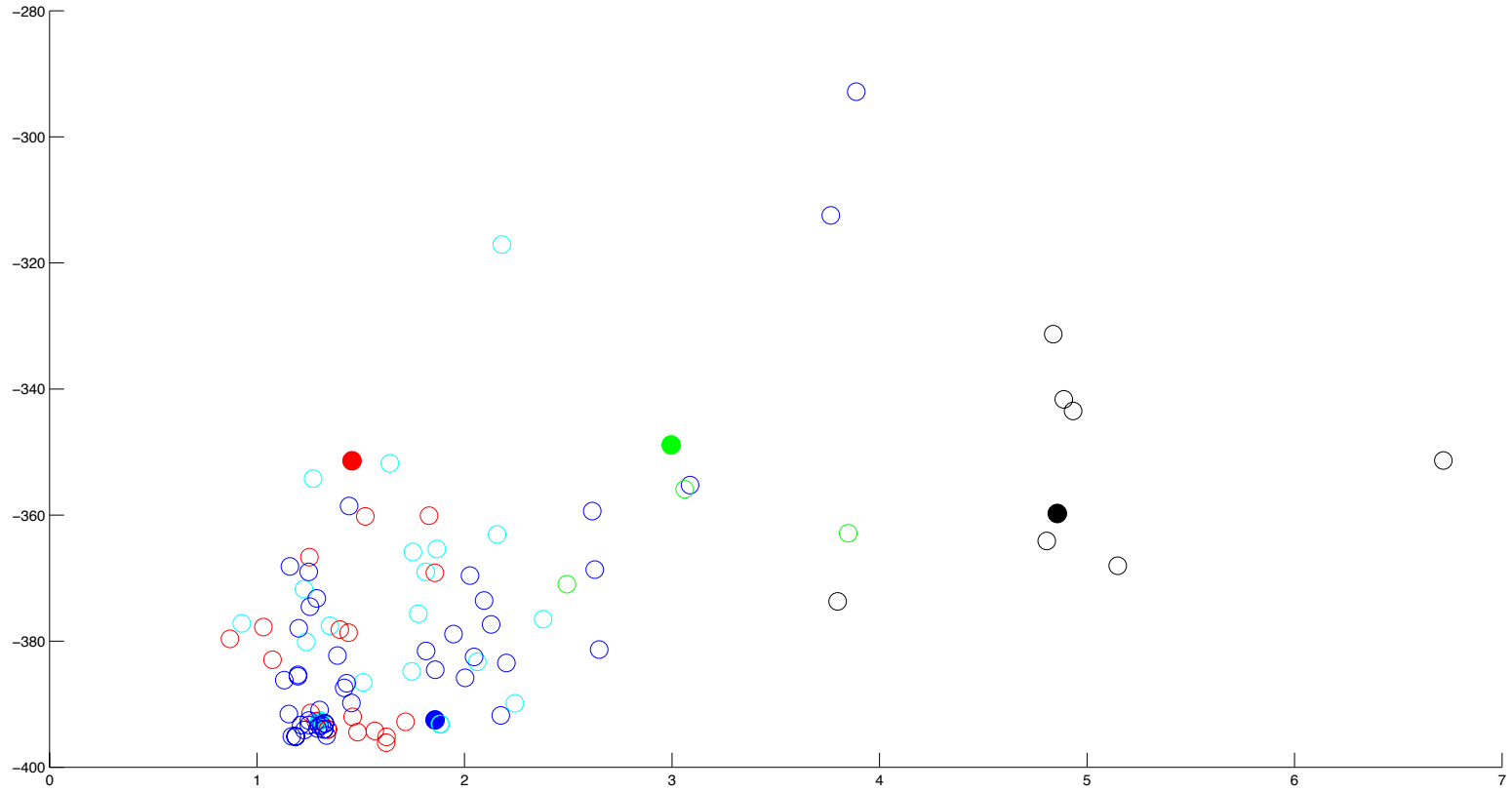


My Kmeans

Input RMSD distance matrix (lower triangle)

Output centroid structures and cluster grouping (iteratively)


Sample Set
-280
-300
-320
-340
-360
-380
-400
Potential (Angstroms)
0
1
2
3
4
5
6
7
8
RMSD (Angstroms)


−280
−300
−320
−340
−360
−380
−400
0
1
2
3
4
5
6
7