A2 In-Class Discussion

Chris Kauffman

Last Updated:
Tue Feb 24 01:44:35 PM EST 2026

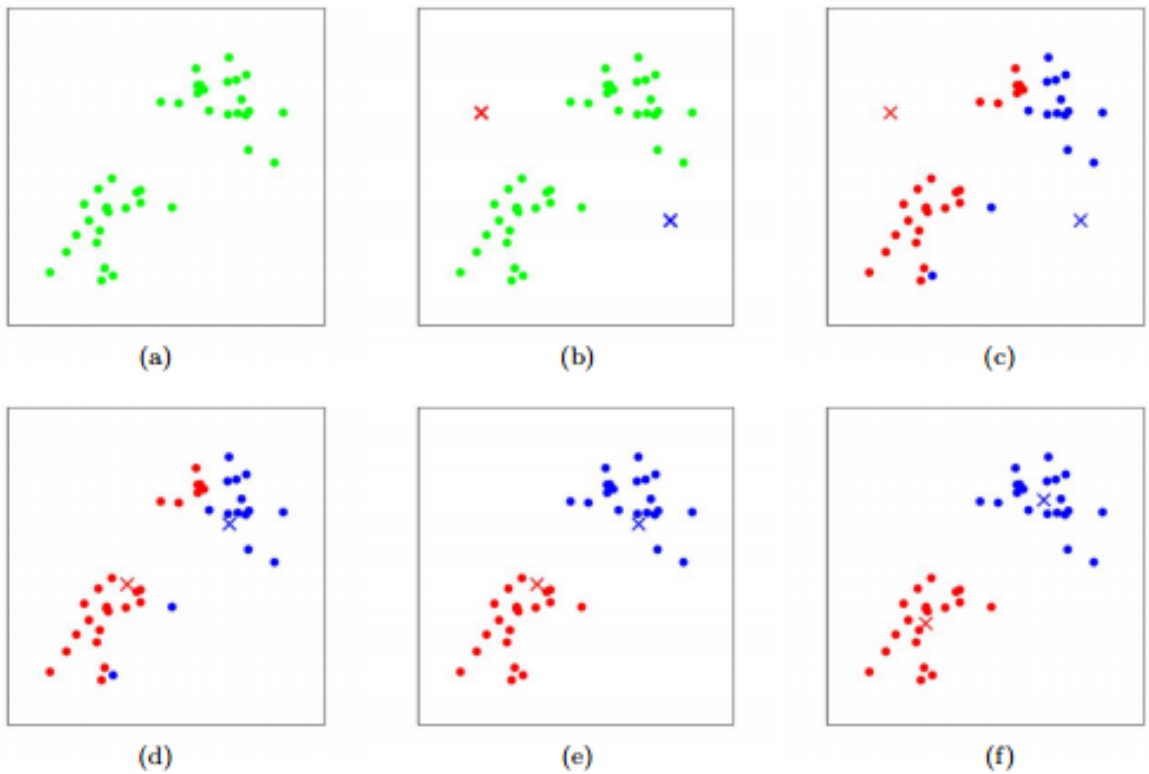
Two MPI Problems

1. Heat Simulation
2. K-means clustering

Previously discussed strategies to parallelize heat problem for distributed memory machine in lecture, will discuss K-Means now

K-Means Clustering

- A standard ML / Data Mining / Stats problem
- Input: data + #of clusters desired
- Output: assignment of each data to a cluster + cluster centers
- Algorithm: Iterates between
  1. Calculate cluster centers
  2. Calculate cluster assignments



Source: K-Means by Chris Piech. Based on a handout by Andrew Ng.

Overall

<MINTED>

Initial Assignment to Random Clusters

<MINTED>

Calculate Cluster Centers

<MINTED>

Assign Data to Clusters

<MINTED>

Distributed Memory Parallel Versions

- Algorithm deals with Data and Clusters, each a matrixy thing
- How would you divide up this data in a distributed parallel version?
- Would data redistribution be required in your scheme?
- What information needs to be exchanged at each iteration?
- Do processors need to communicate for the initial cluster assignment? Or can data be assigned to initial clusters without communication?