K-means Clustering for Visualization

Yu-Shuen Wang, CS, NCTU

K-means Clustering

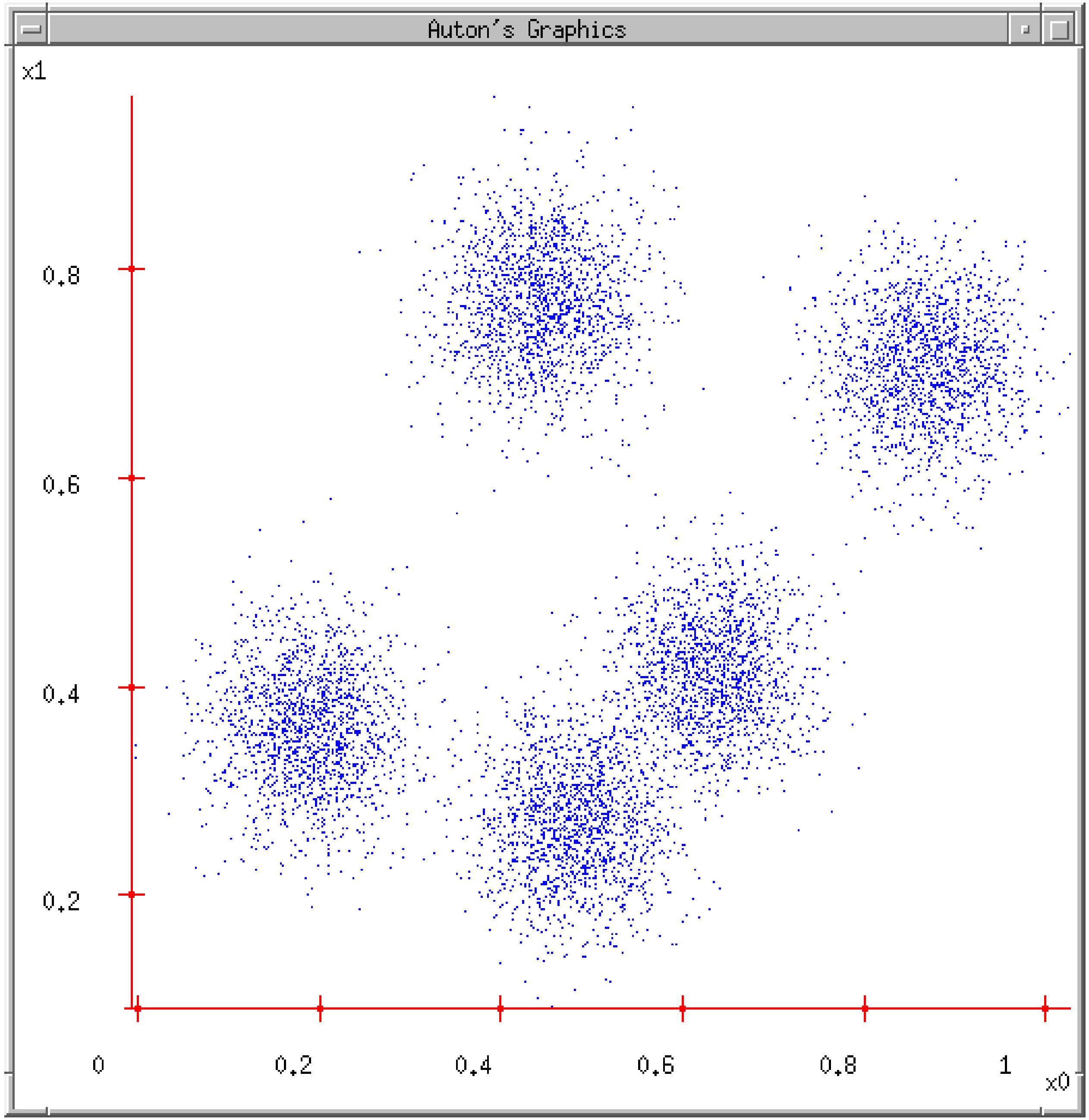
- What is clustering?
- Why would we want to cluster?
- How would you determine clusters?
- How can you do this efficiently?

K-means Clustering

- Strengths
  - Simple iterative method
  - User provides “K”
- Weaknesses
  - Often too simple → bad results
  - Difficult to guess the correct “K”

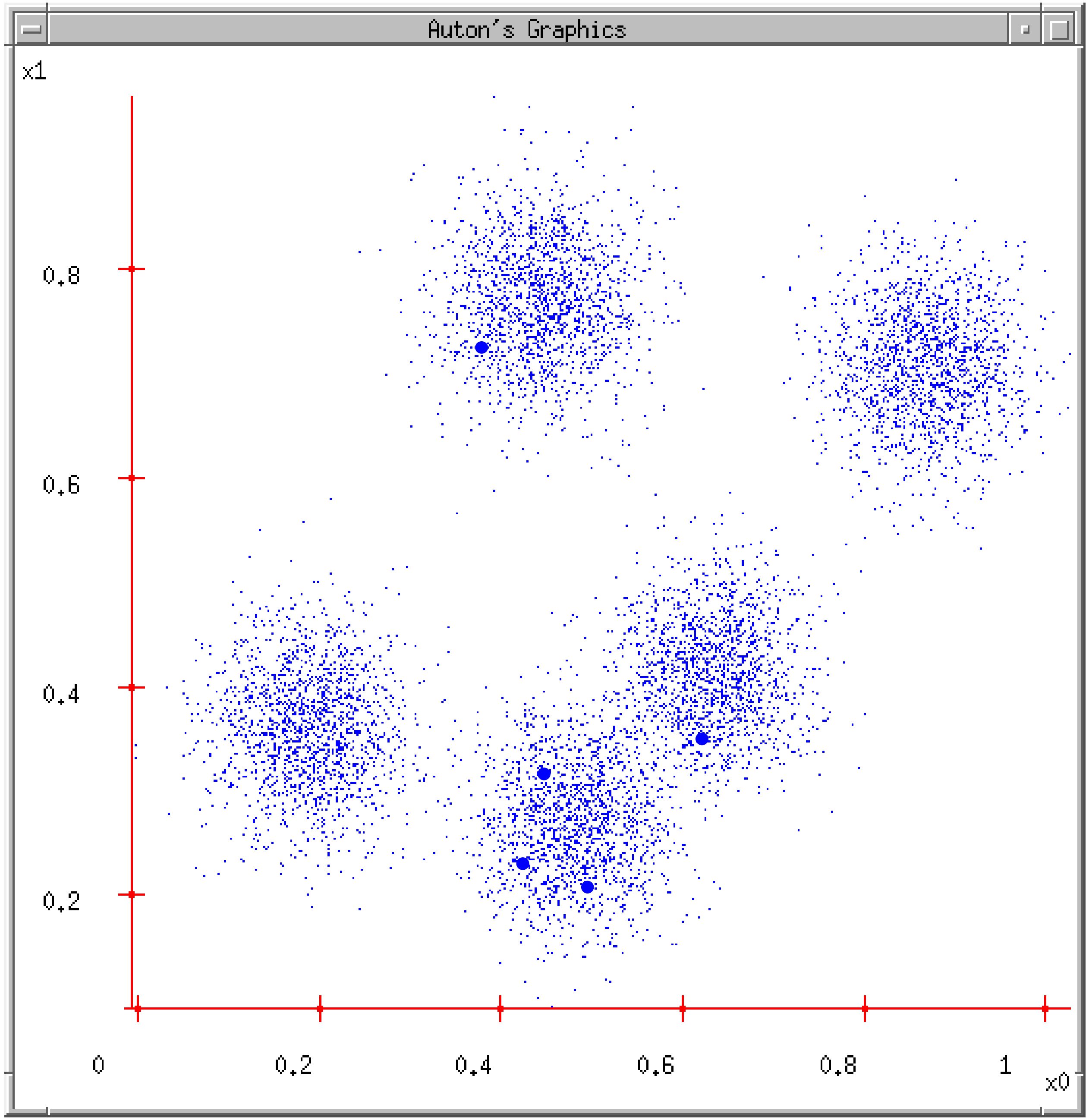
K-means

1. Ask user how many clusters they'd like. *(e.g. k=5)*



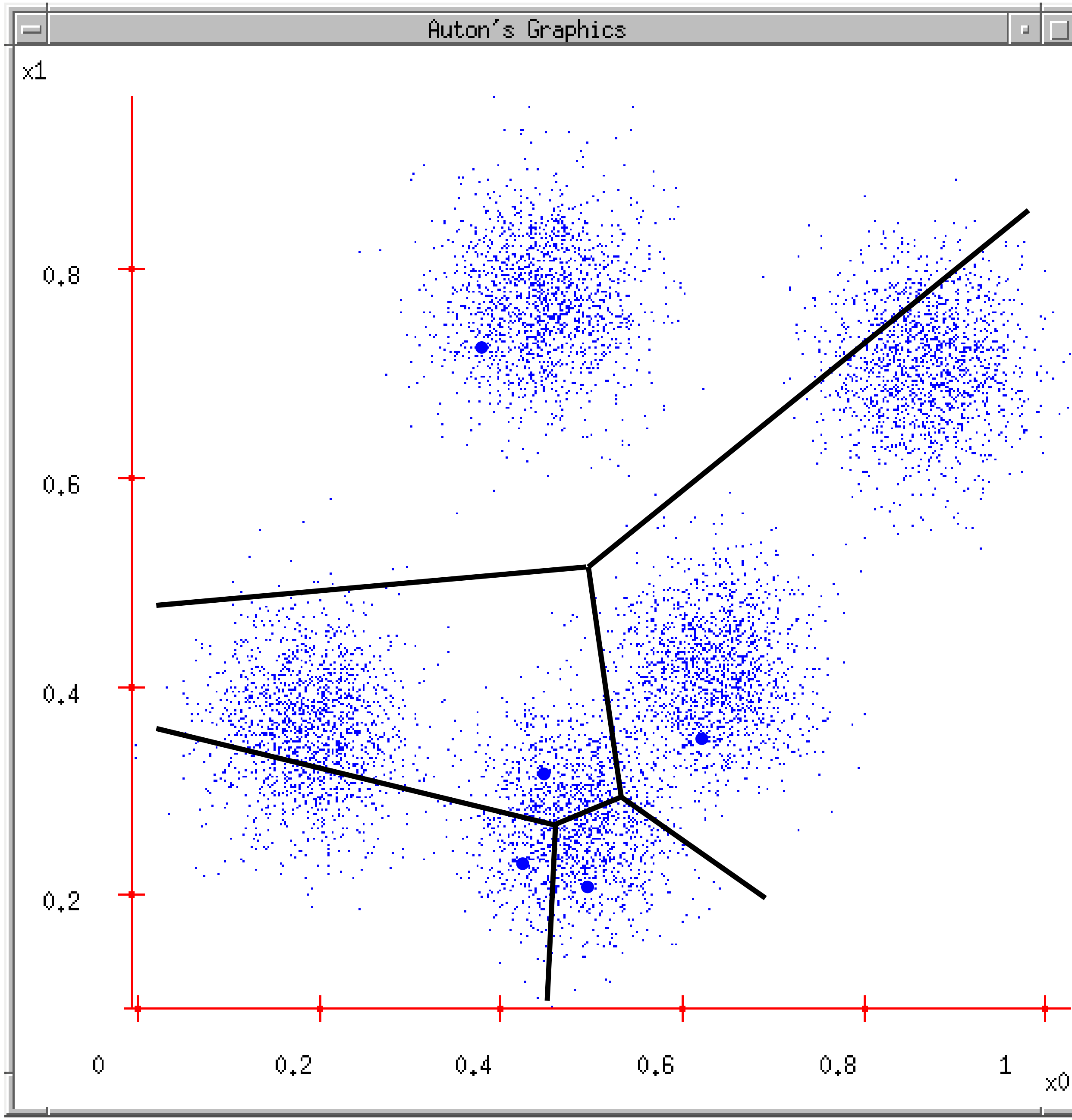
K-means

1. Ask user how many clusters they'd like. *(e.g. k=5)*
2. Randomly guess k cluster Center locations



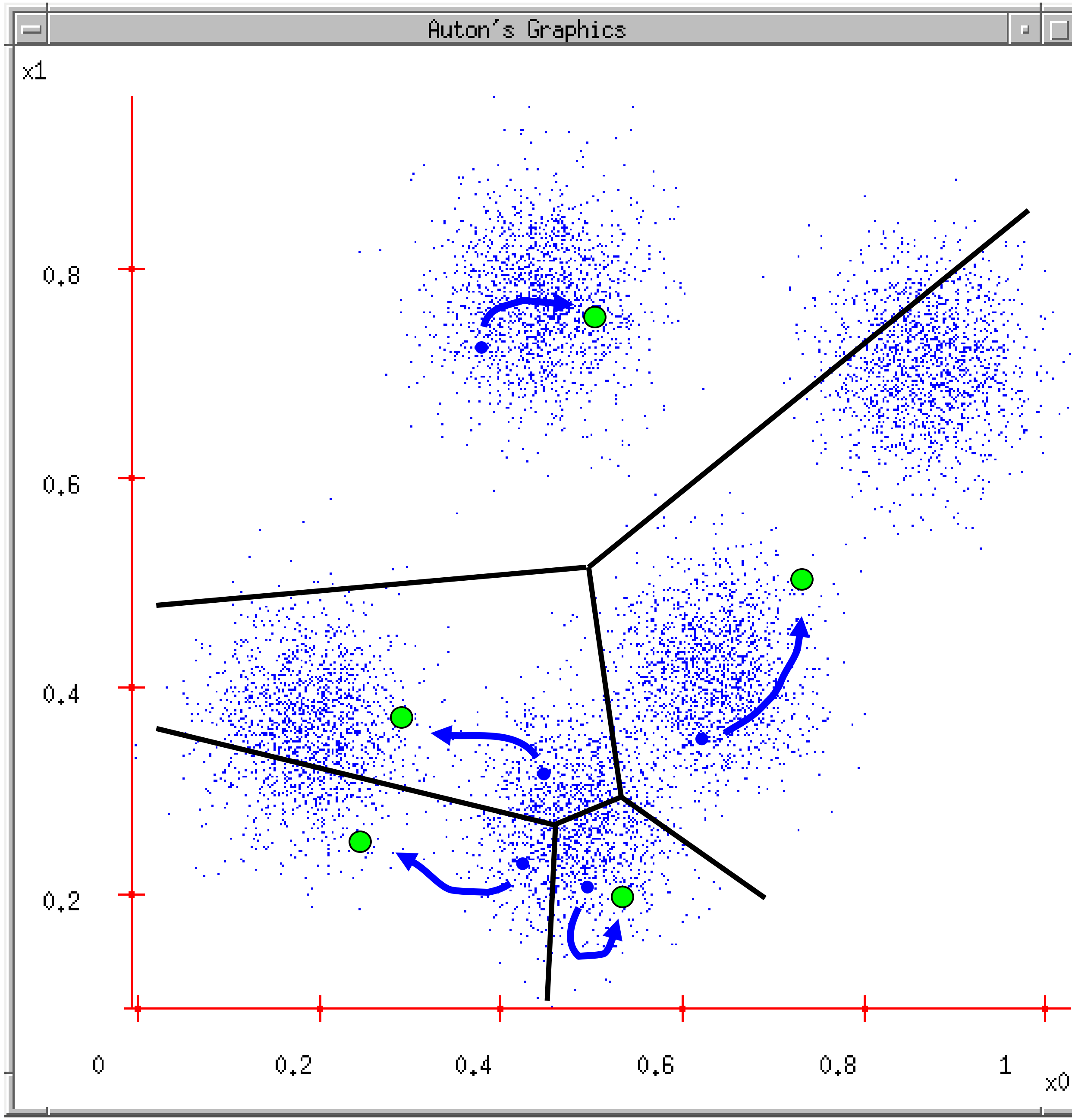
K-means

1. Ask user how many clusters they'd like. *(e.g. k=5)*
2. Randomly guess k cluster Center locations
3. Each datapoint finds out which Center it's closest to. (Thus each Center "owns" a set of datapoints)



K-means

1. Ask user how many clusters they'd like. *(e.g. k=5)*
2. Randomly guess k cluster Center locations
3. Each datapoint finds out which Center it's closest to.
4. Each Center finds the centroid of the points it owns



K-means

1. Ask user how many clusters they'd like. *(e.g. k=5)*

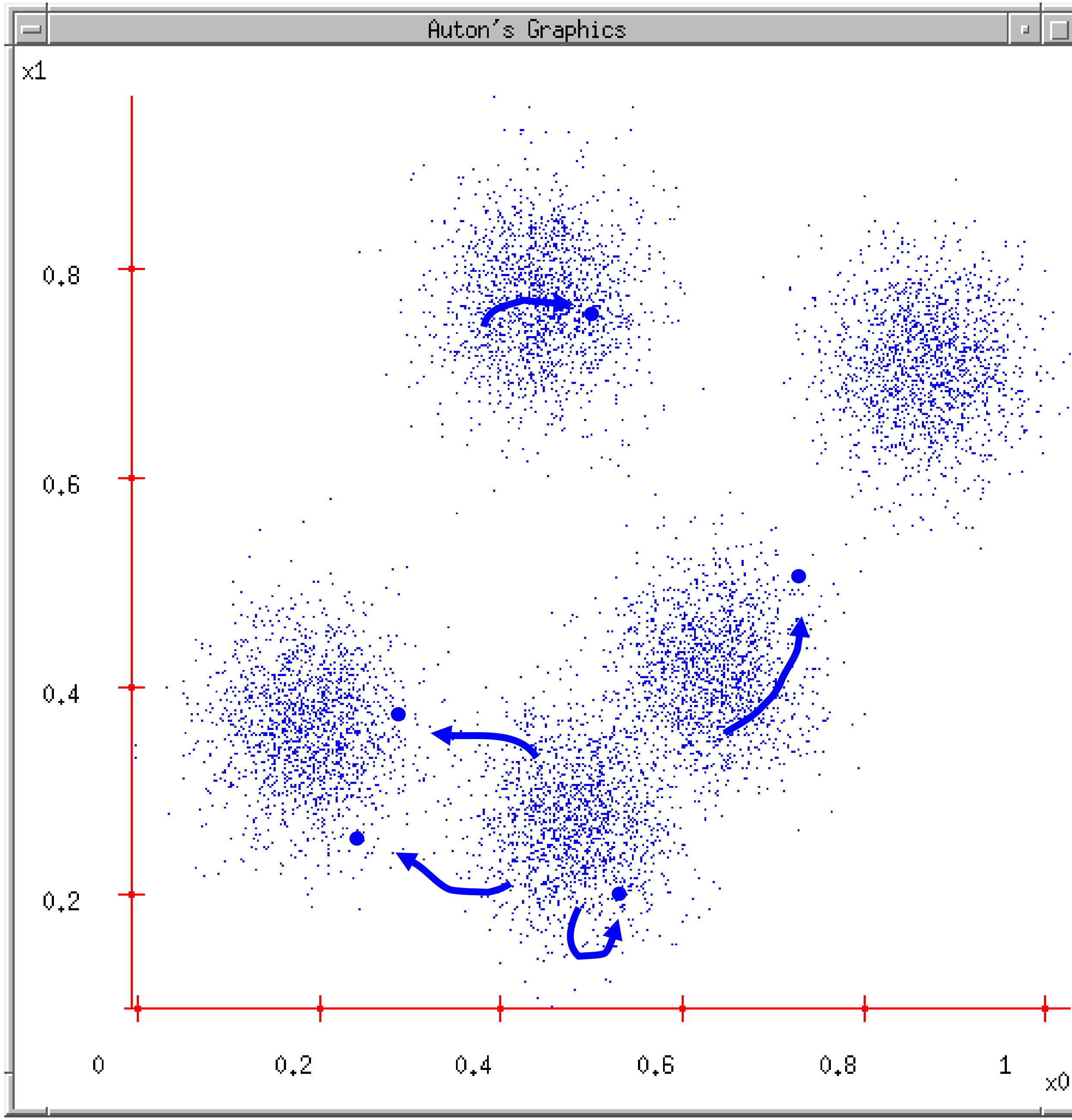
2. Randomly guess k cluster Center locations

3. Each datapoint finds out which Center it's closest to.

4. Each Center finds the centroid of the points it owns…

5. …and jumps there

6. …Repeat until terminated!

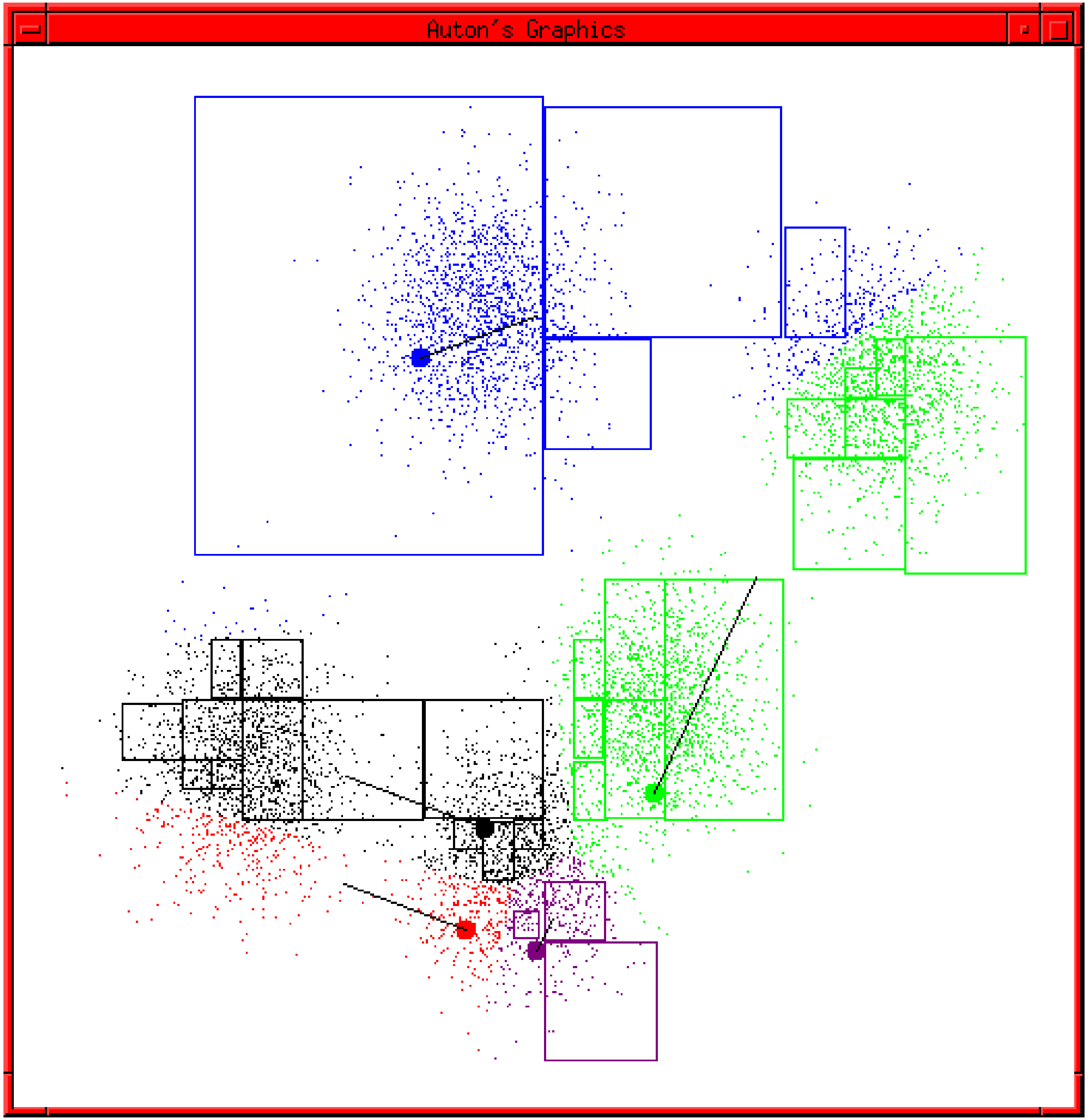


K-means Start

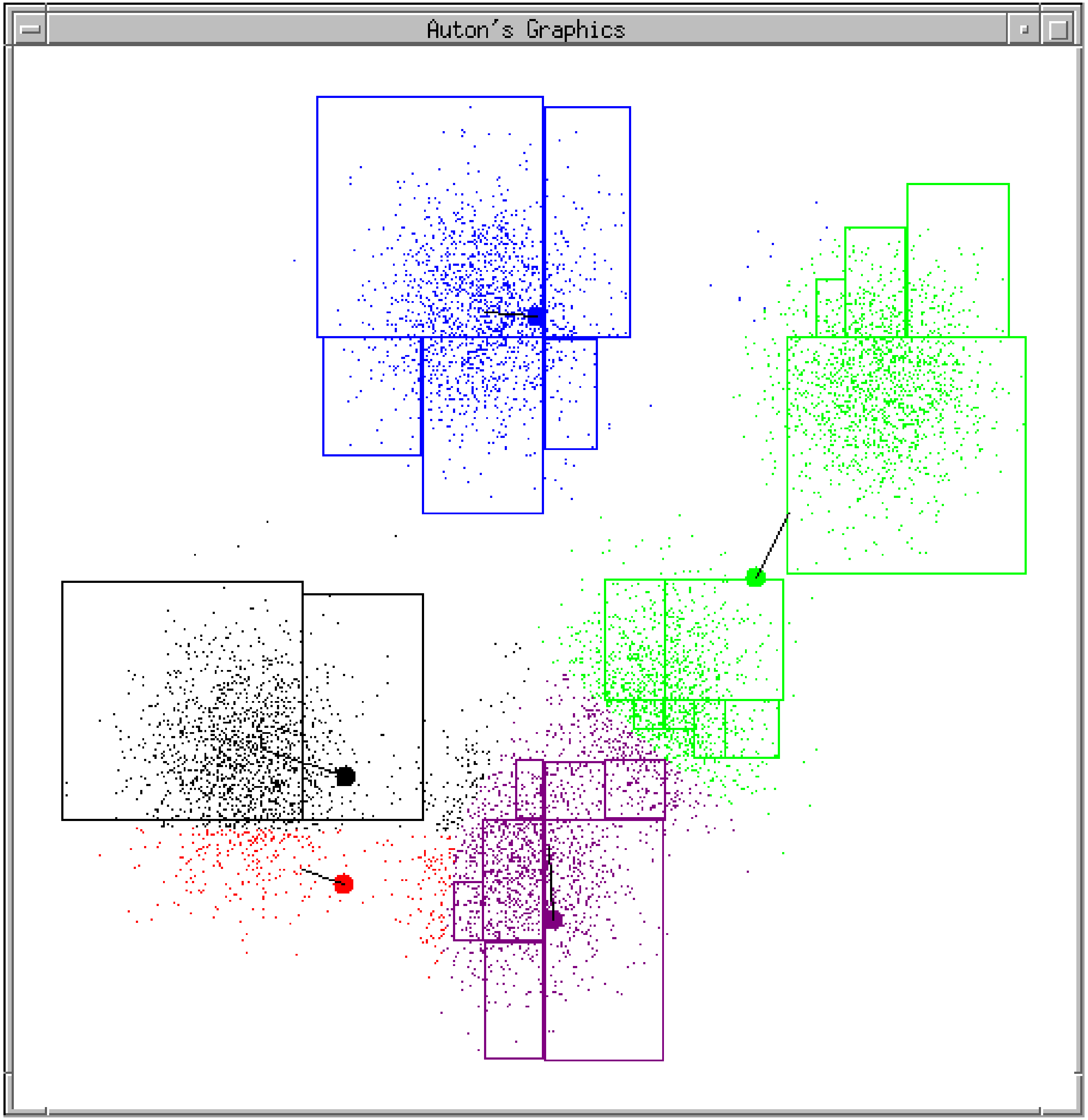
Advance apologies: in Black and White this example will deteriorate

Example generated by Dan Pelleg's super-duper fast K-means system:

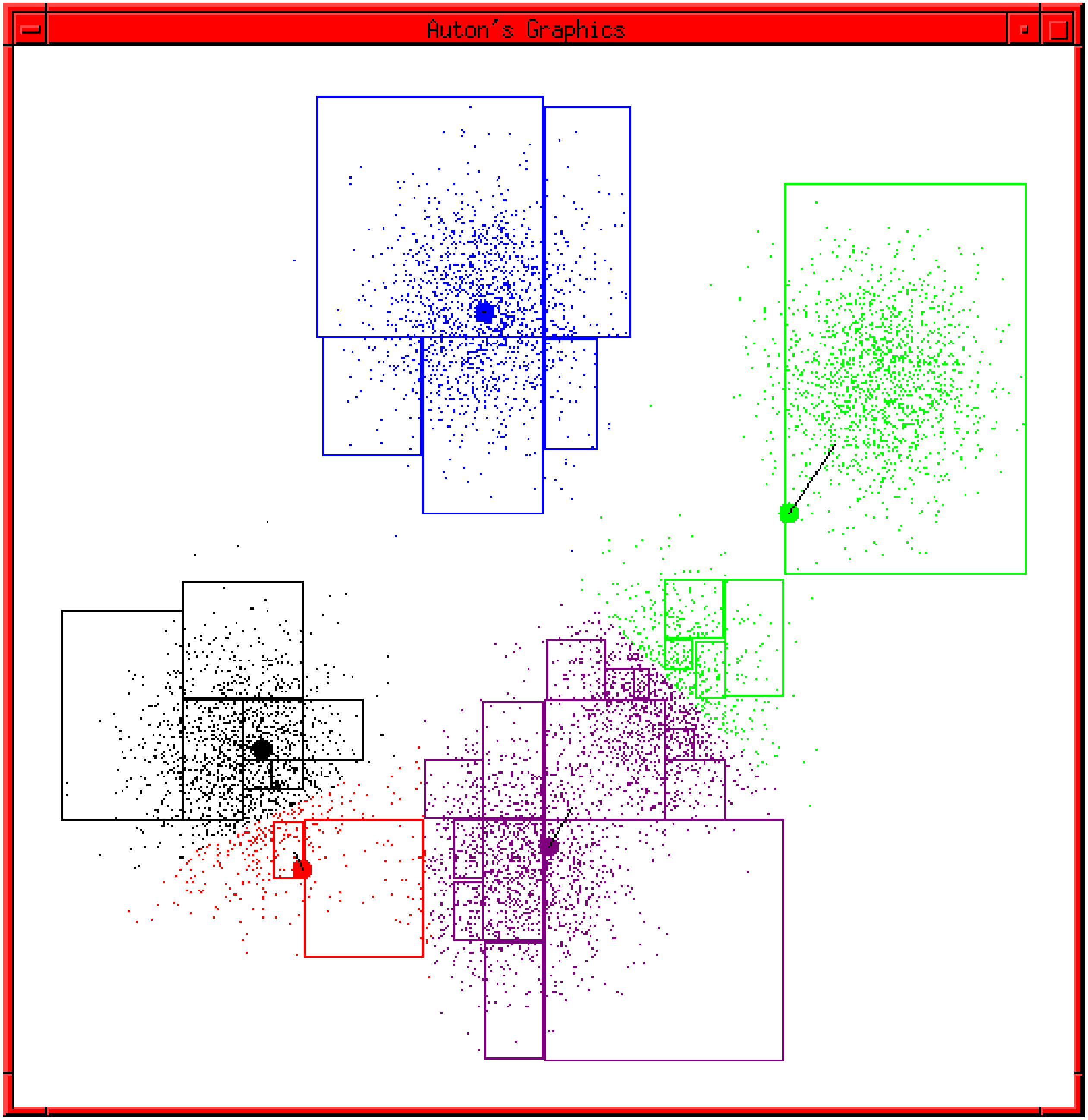
Dan Pelleg and Andrew Moore. Accelerating Exact k-means Algorithms with Geometric Reasoning. Proc. Conference on Knowledge Discovery in Databases 1999, (KDD99) (available on www.autonlab.org/pap.html*)*



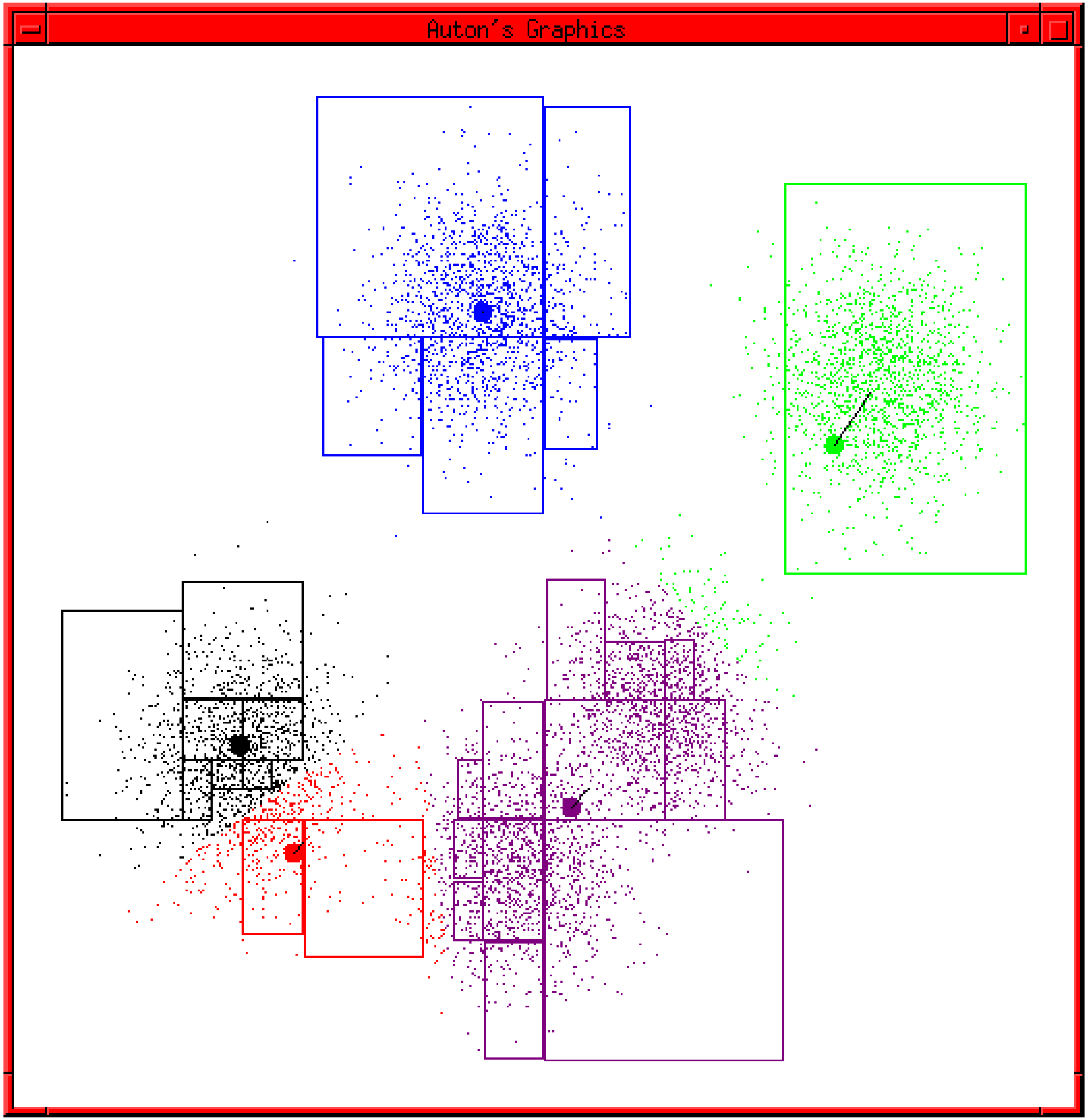
K-means continues …



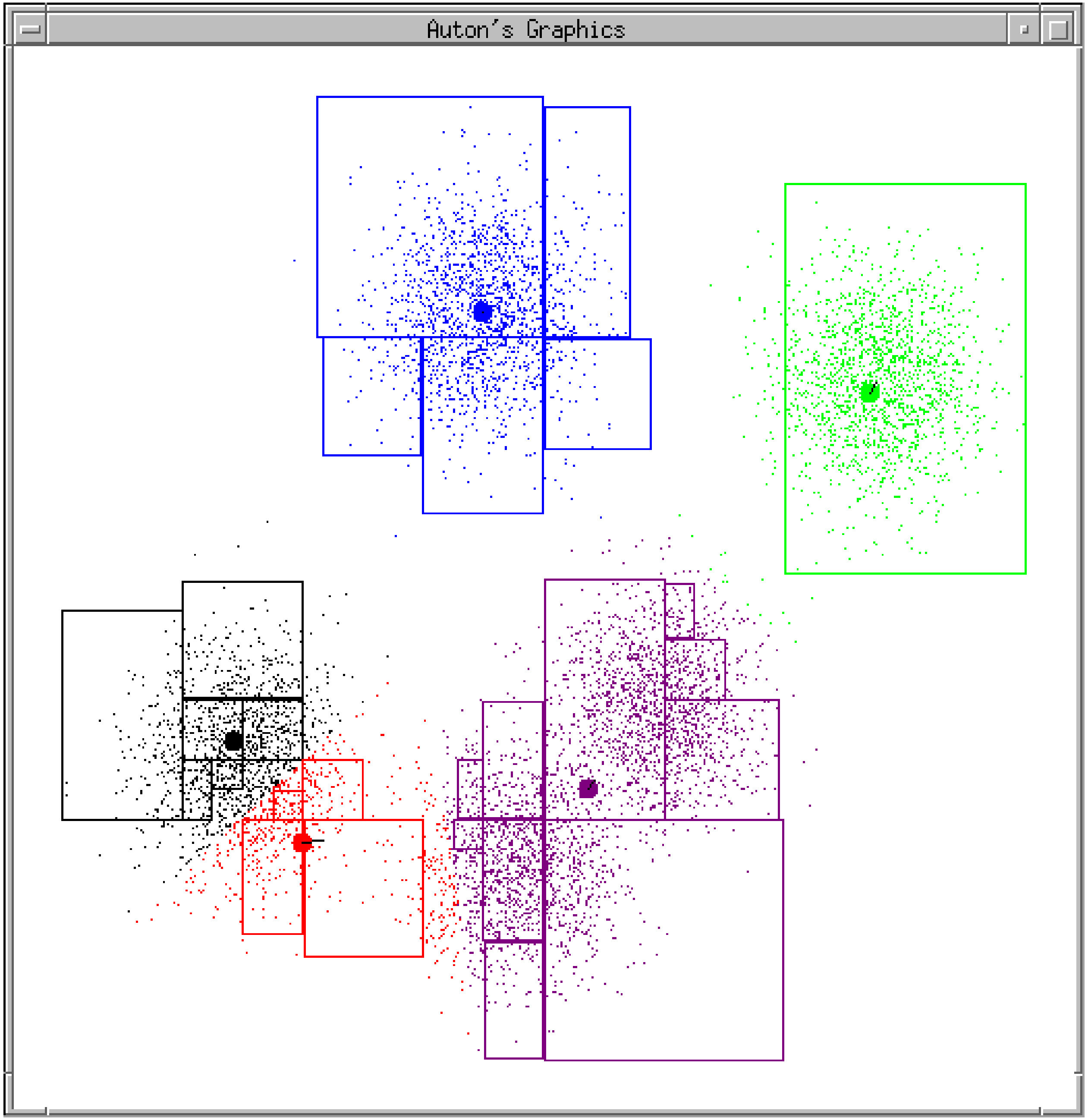
K-means continues …



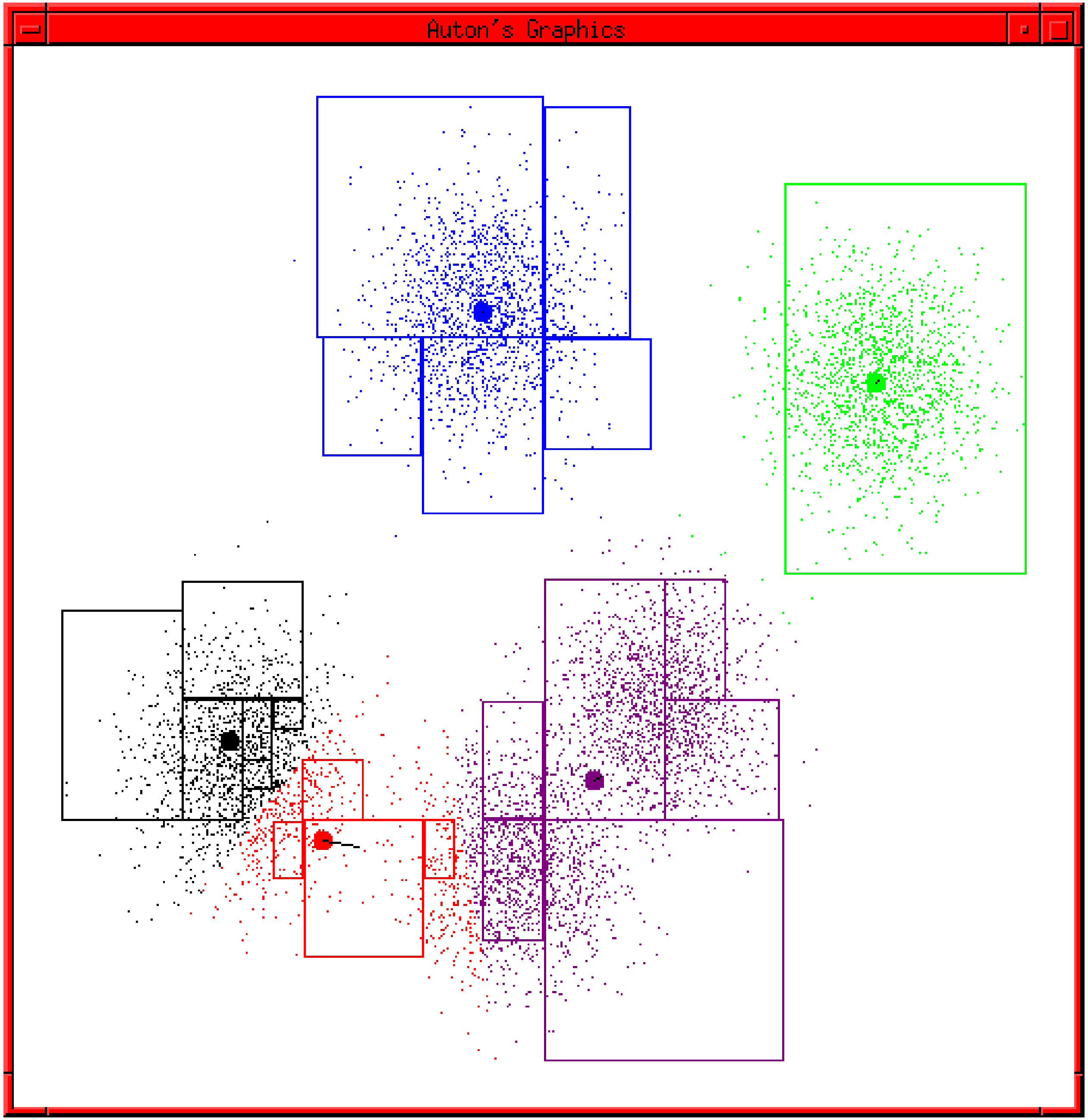
K-means continues …



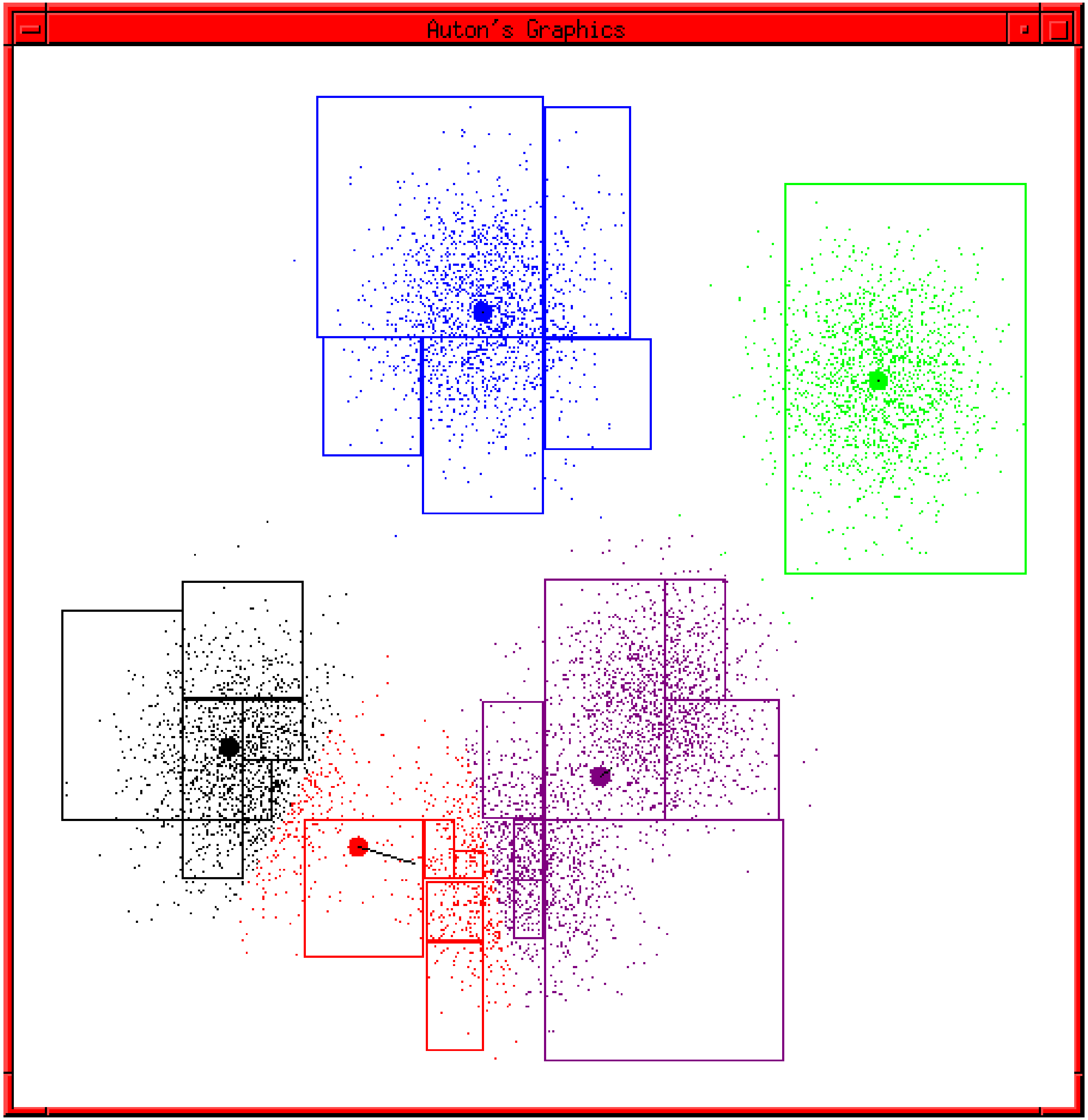
K-means continues …



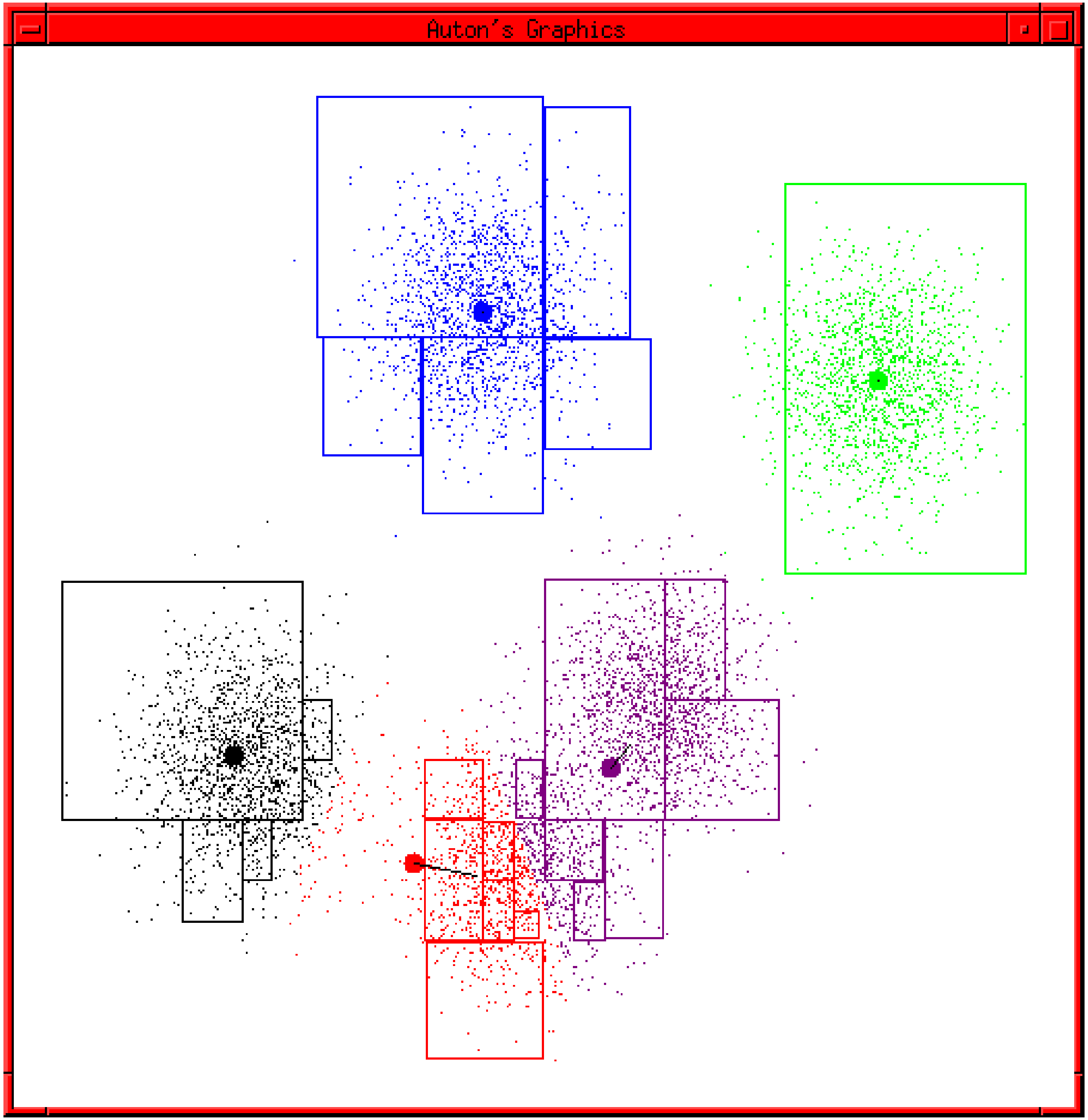
K-means continues …



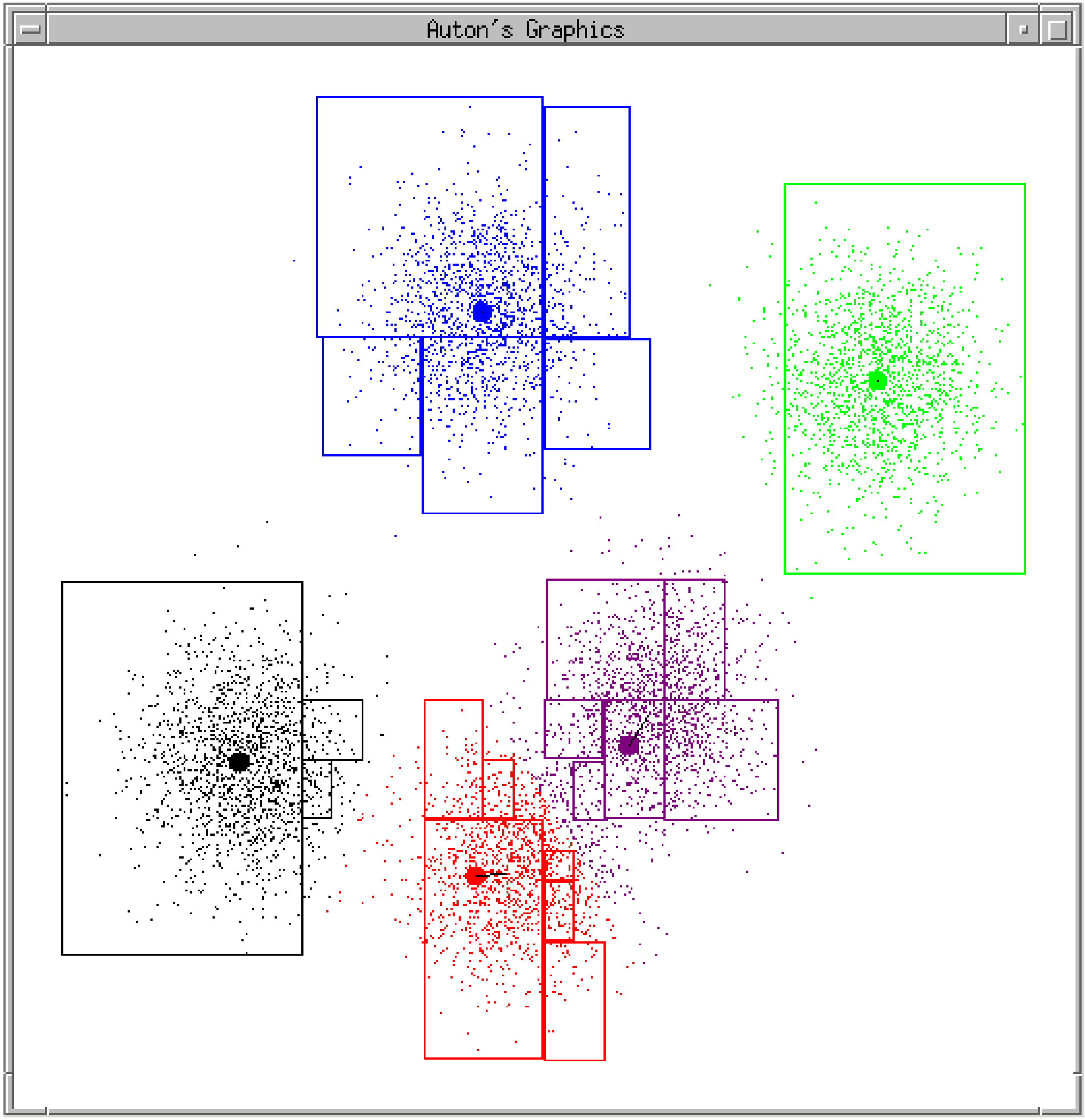
K-means continues …



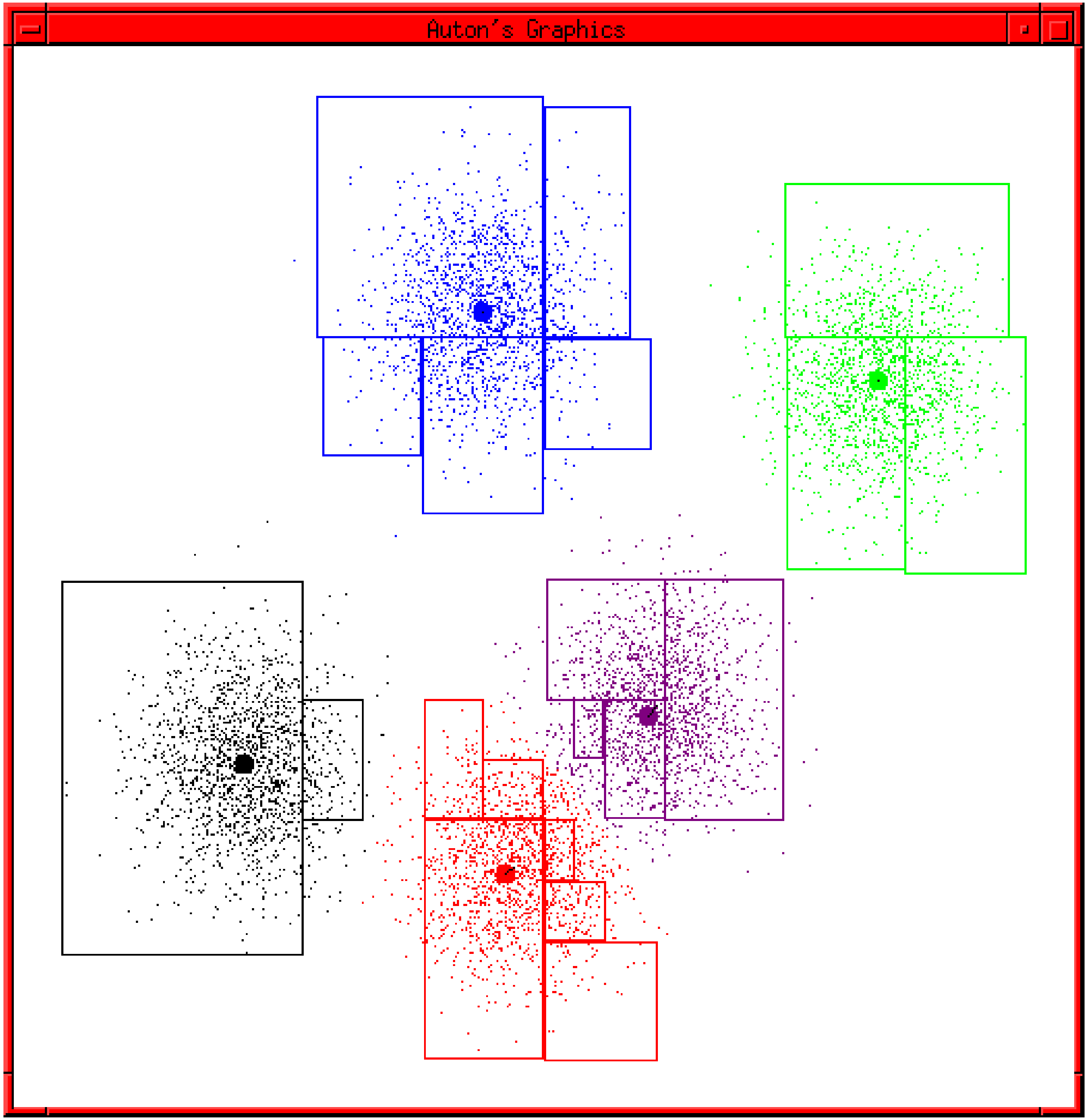
K-means continues …



K-means continues …



K-means terminates



K-means Clustering

- Iterate:
  - Calculate distance from objects to cluster centroids.
  - Assign objects to closest cluster
  - Recalculate new centroids
- Stop based on convergence criteria
  - No change in clusters
  - Max iterations

K-means Issues

- Distance measure is squared Euclidean
  - Scale should be similar in all dimensions
    - Rescale data?
  - Not good for nominal data. Why?
- Approach tries to minimize the within-cluster sum of squares error (WCSS)
  - Implicit assumption that sum of square error (SSE) is similar for each group

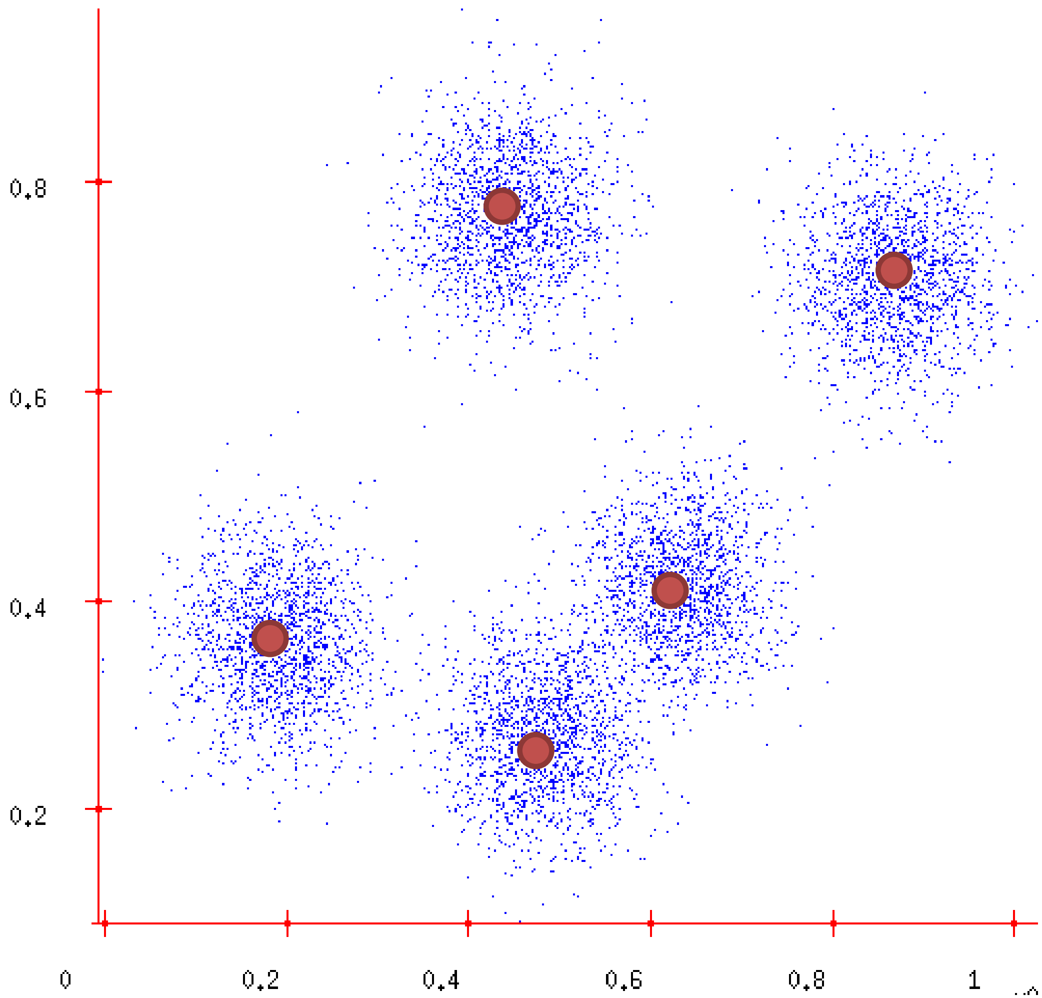
WCSS

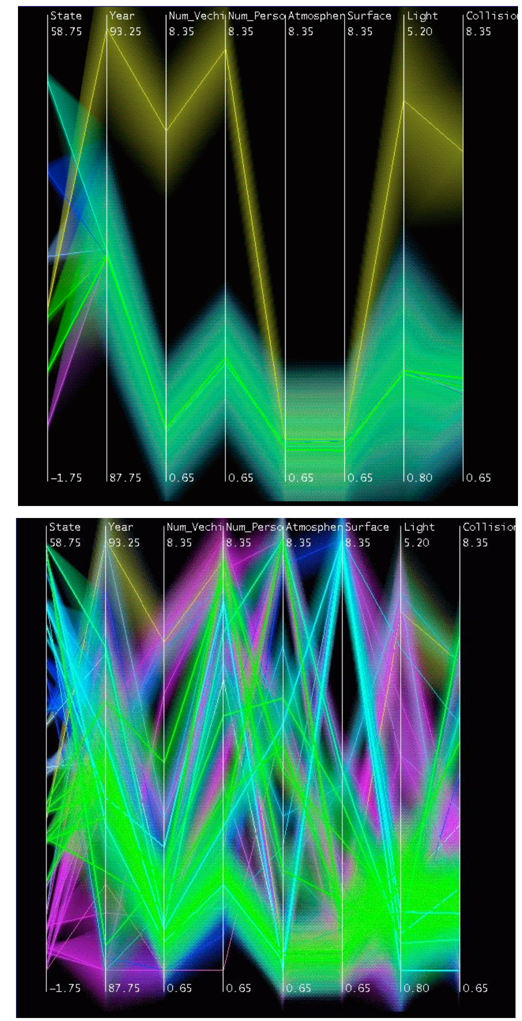
- The over all WCSS is given by:
  $$\sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2$$
- The goal is to find the smallest WCSS
- Does this depend on the initial seed values?
- Possibly.

Bottom Line

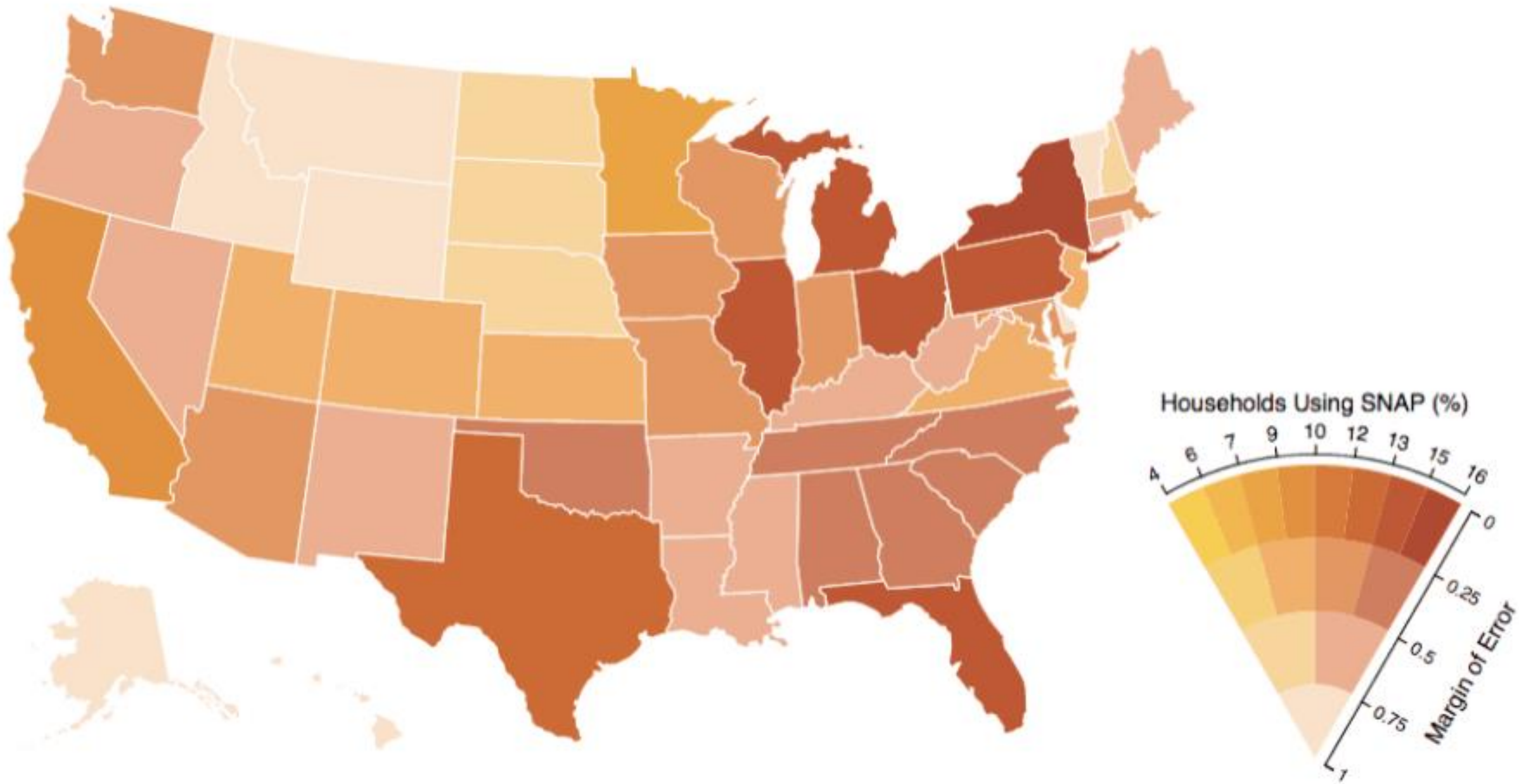
- K-means
  - Easy to use
  - Need to know K
  - May need to scale data
  - Good initial method
- Local optima
  - No guarantee of optimal solution
  - Repeat with different starting values

Uncertainty Visualization

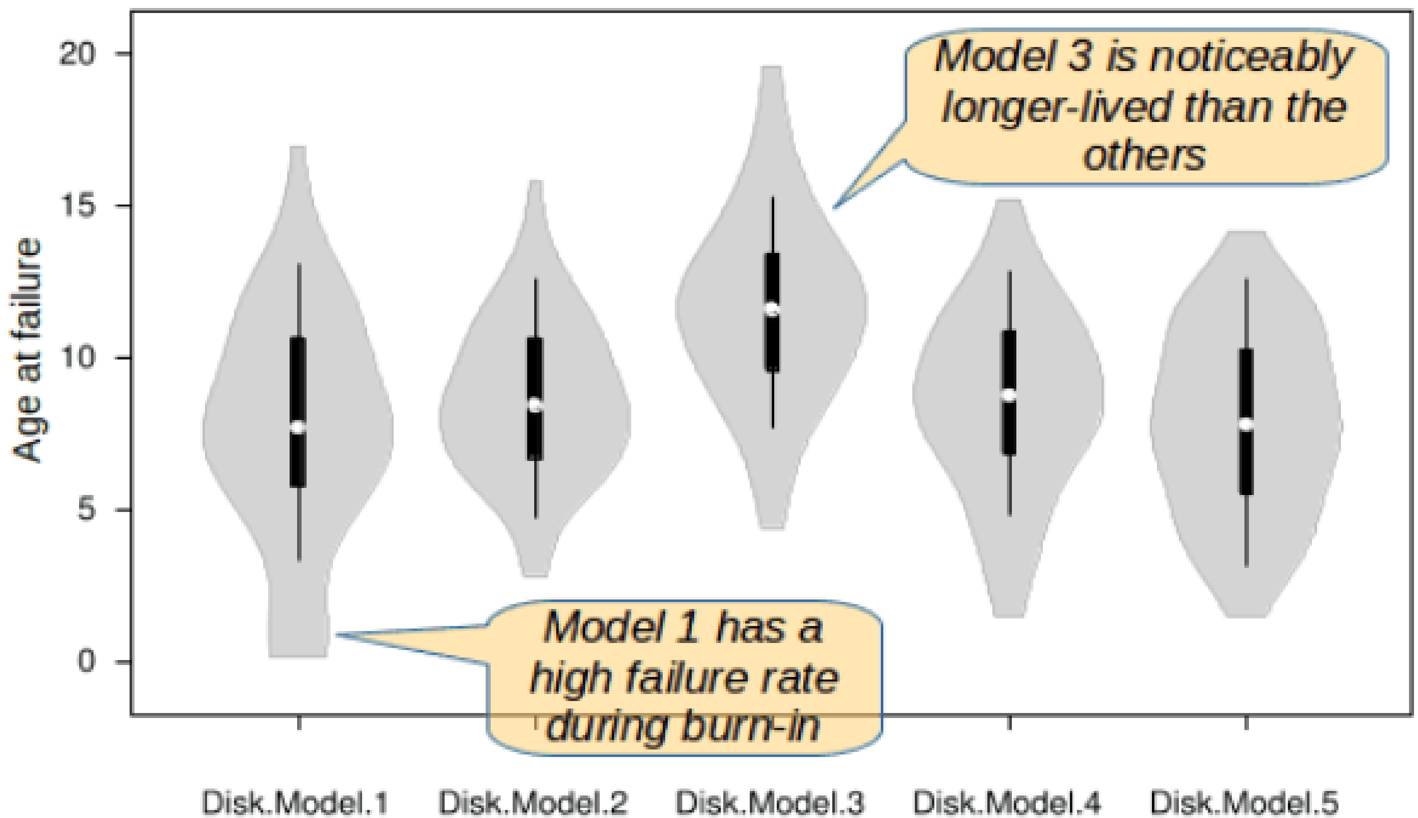




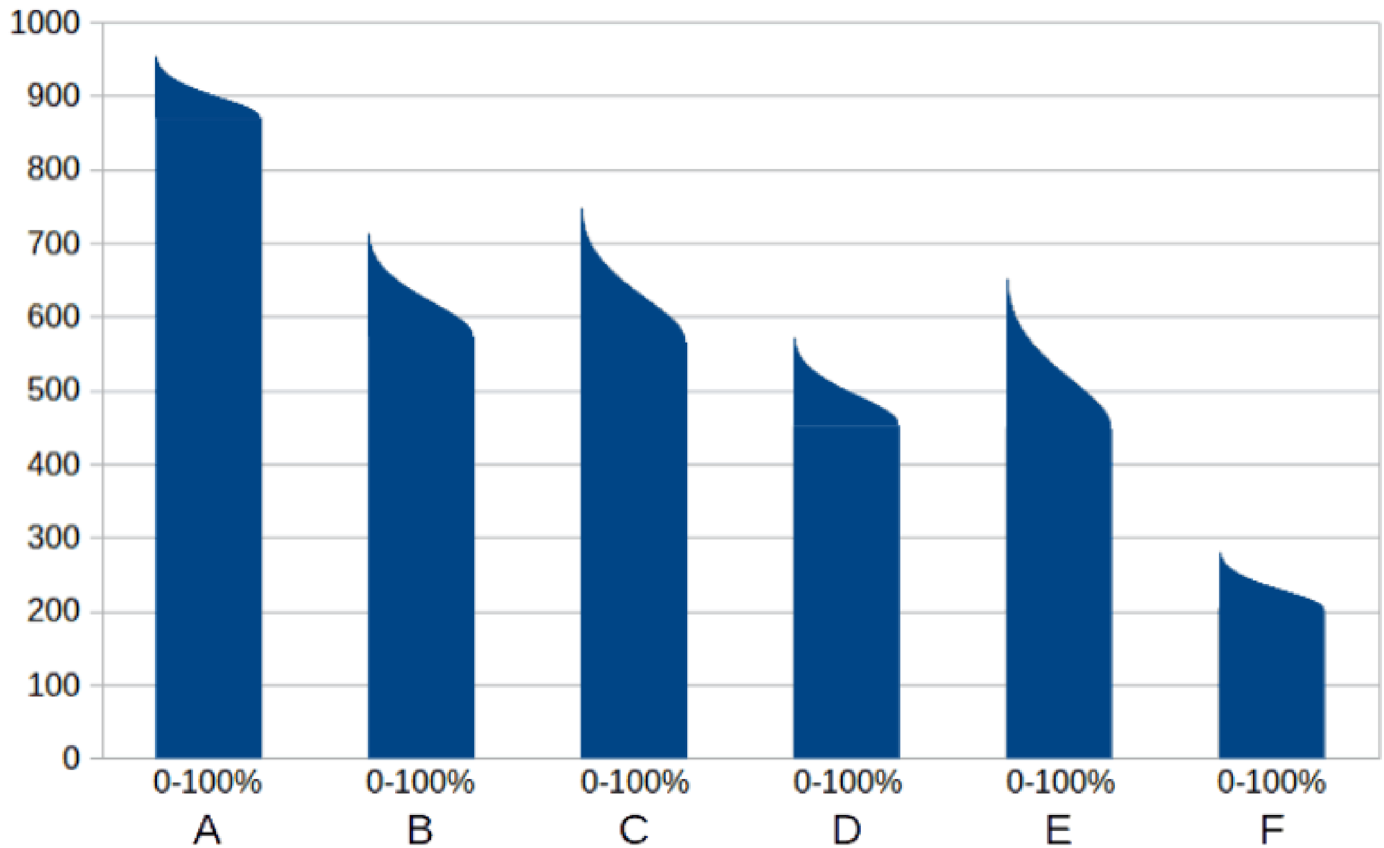
Uncertainty with color



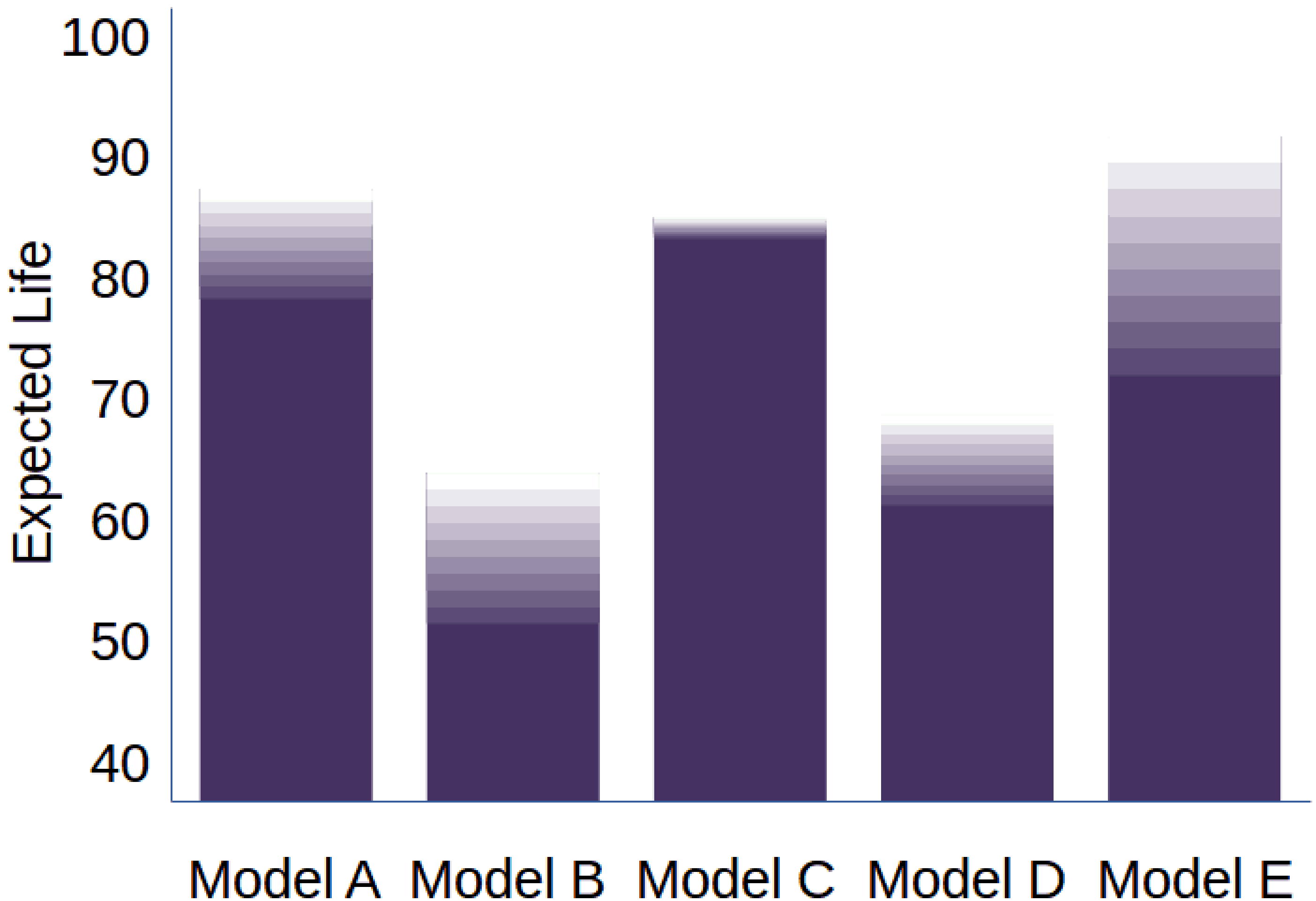
Uncertainty – life time expectancy



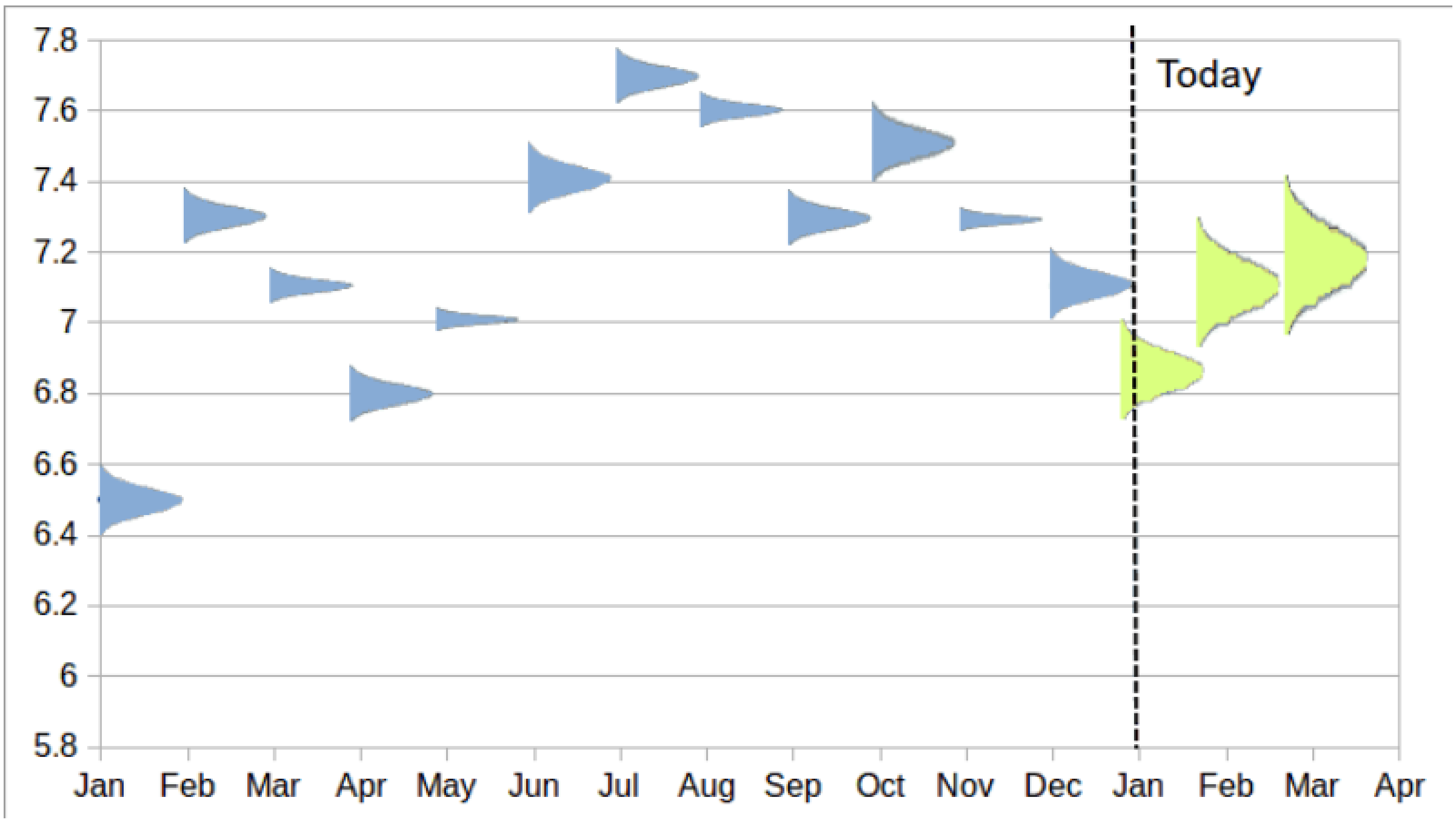
Uncertainty – life time expectancy



Uncertainty – life time expectancy



Uncertainty – consumer satisfaction



Uncertainty – weather forecast