Fall 2018: Introduction to Data Science

GIRI NARASIMHAN, SCIS, FIU

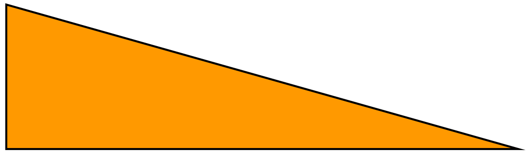
Implementing Clustering

Example High-Dim Application: SkyCat

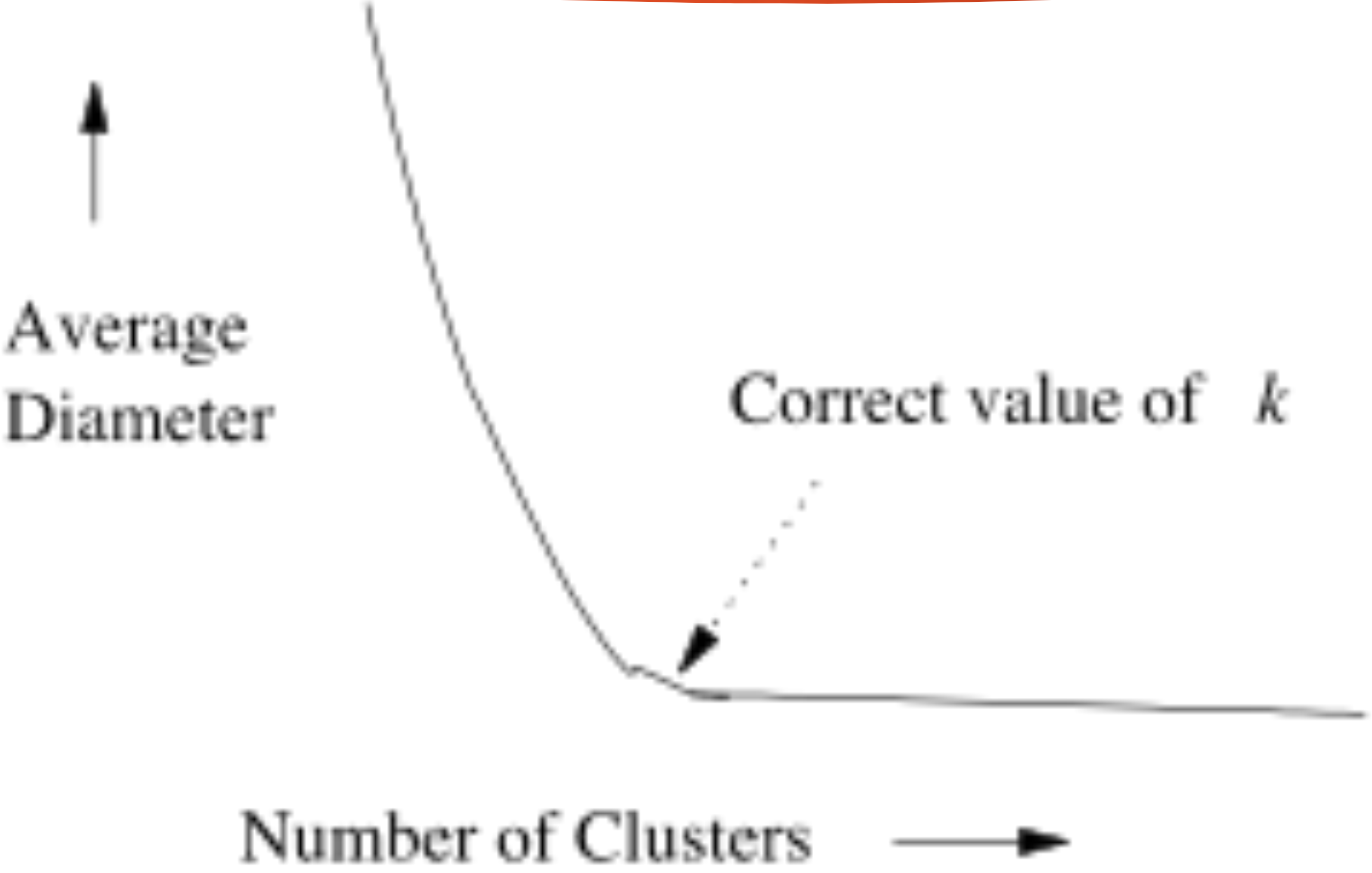
- A catalog of 2 billion “sky objects” represents objects by their radiation in 7 dimensions (frequency bands).
- Problem: cluster into similar objects, e.g., galaxies, nearby stars, quasars, etc.
- Sloan Sky Survey is a newer, better version.

Curse of Dimensionality

- Assume random points within a bounding box, e.g., values between 0 and 1 in each dimension.
- In 2 dimensions: a variety of distances between 0 and 1.41.
- In 10,000 dimensions, the difference in any one dimension is distributed as a triangle.



How to find K for K-means?



BFR Algorithm

- BFR (Bradley-Fayyad-Reina) – variant of *K* -means for very large (disk-resident) data sets.
- Assumes that clusters are normally distributed around a centroid in Euclidean space.
  - SDs in different dimensions may vary

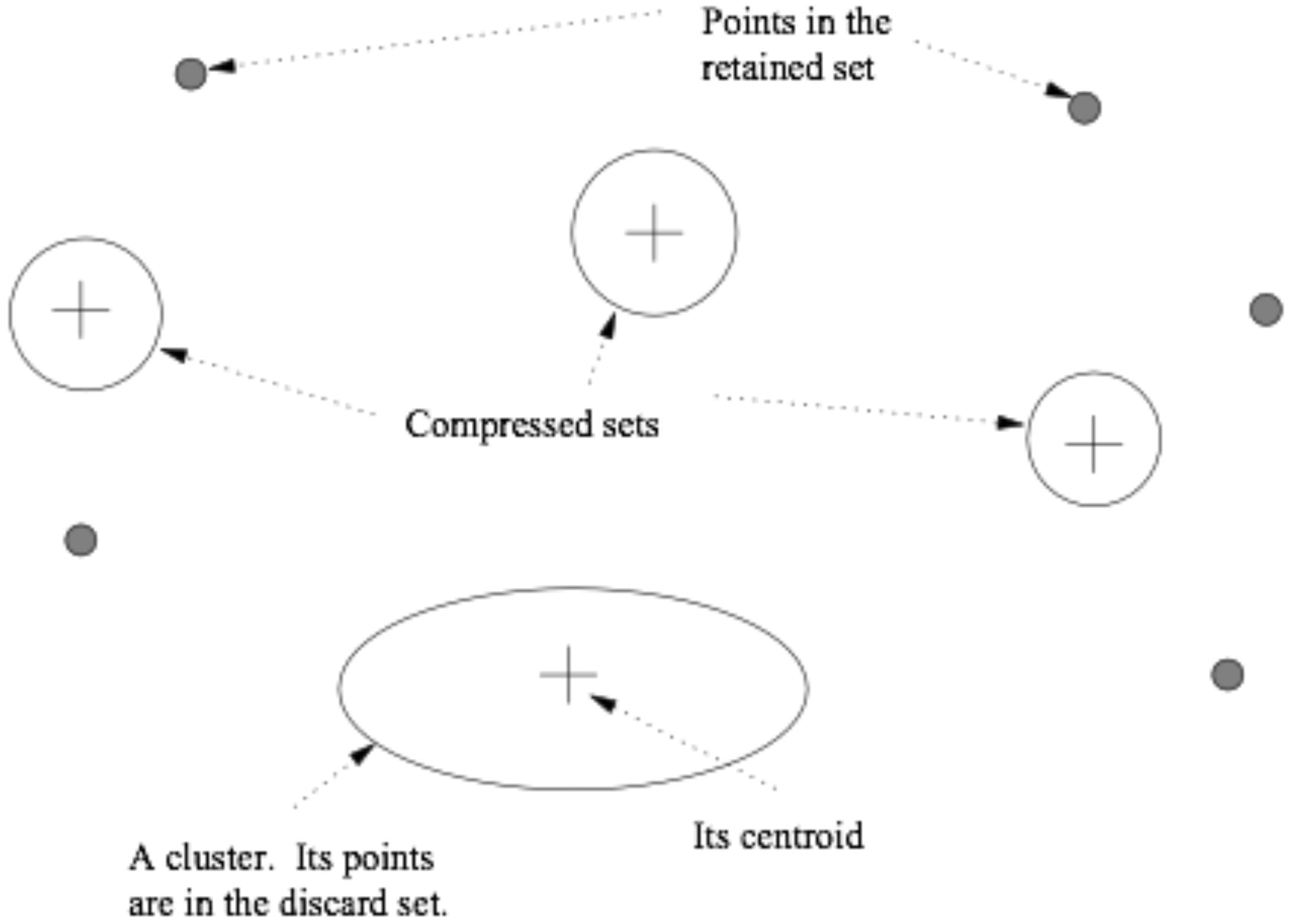
BFR … 2

- Points read "chunk" at a time.
- Most points from previous chunks summarized by simple statistics.
- First load handled by some sensible approach:
    1. Take small random sample and cluster optimally.
    2. Take sample; pick random point, & *k* – 1 more points incrementally, each as far from previously points as possible.

BFR … 3

1. *Discard set* : points close enough to a centroid to be summarized.
2. *Compression set* : groups of points that are close together but not close to any centroid. They are summarized, but not assigned to a cluster.
3. *Retained set* : isolated points.

BFR … 4


Points in the retained set
Compressed sets
A cluster. Its points are in the discard set.
Its centroid

BFR: How to summarize?

- Discard Set & Compression Set: N, SUM, SUMSQ
- 2d + 1 values
- Average easy to compute
  - SUM/N
- SD not too hard to compute
  - VARIANCE = (SUMSQ/N) – $(\text{SUM/N})^2$

BFR: Processing

- Maintain N, SUM, SUMSQ for clusters
- Policies for merging compressed sets needed and for merging a point in a cluster
- Last chunk handled differently
  - Merge all compressed sets
  - Merge all retained sets into nearest clusters
- BFR suggests **Mahalanobis Distance**

Mahalanobis Distance

- Normalized Euclidean distance from centroid.
- For point ($x_1,\ldots,x_k$) and centroid ($c_1,\ldots,c_k$):

1. Normalize in each dimension: $y_i = (x_i - c_i)/\sigma_i$
2. Take sum of the squares of the y_i's.
3. Take the square root.

- For Gaussian clusters, ~65% of points within SD dist

GRPGF Algorithm

GRPGF Algorithm

- Works for non-Euclidean distances
- Efficient, but approximate
- Works well for high dimensional data
  - Exploits orthogonality property for high dim data
- Rules for splitting and merging clusters

Clustering for Streams

- BDMO (authors, B. Babcock, M. Datar, R. Motwani, & L. O'Callaghan)
- Points of stream partitioned into, and summarized by, buckets with sizes equal to powers of two. Size of bucket is number of points it represents.
- Sizes of buckets obey restriction that <= two of each size. Sizes are required to form a sequence -- each size twice previous size, e.g., 3,6,12,24,... .
- Bucket sizes restrained to be nondecreasing as we go back in time. As in Section 4.6, we can conclude that there will be O(log N) buckets.
- Rules for initializing, merging and splitting buckets