Clustering

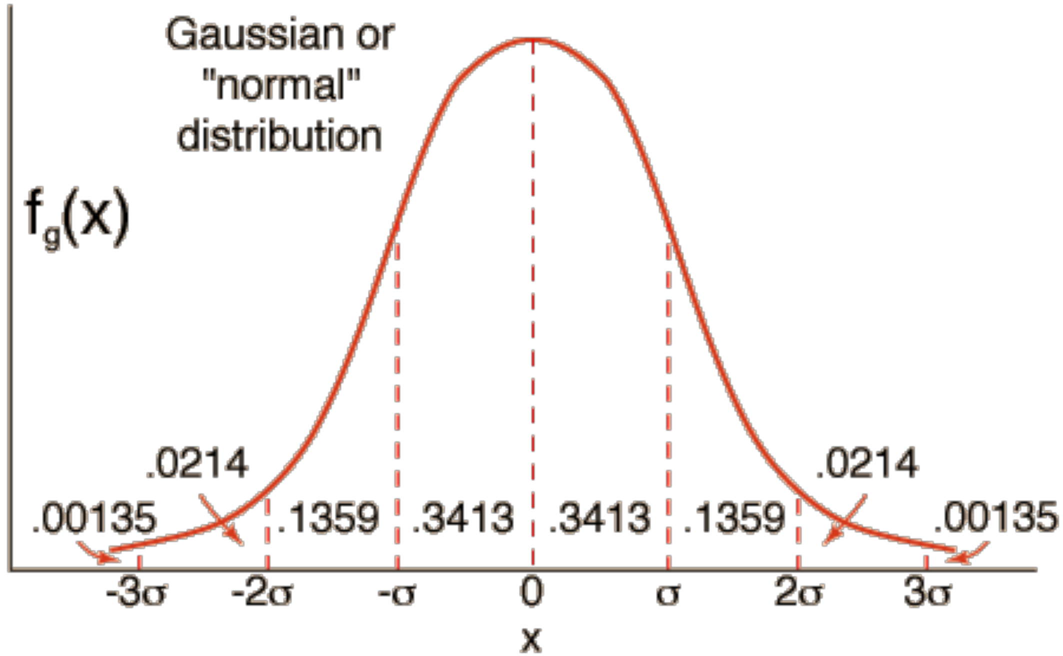
Bradley-Fayyad-Reina (BFR) Algorithm

Mining of Massive Datasets
Leskovec, Rajaraman, and Ullman
Stanford University

BFR Algorithm

- **BFR** [Bradley-Fayyad-Reina] is a variant of *k*-means for very large (disk-resident) data sets
- Assumes each cluster is normally distributed around a centroid in Euclidean space

Normal Distribution

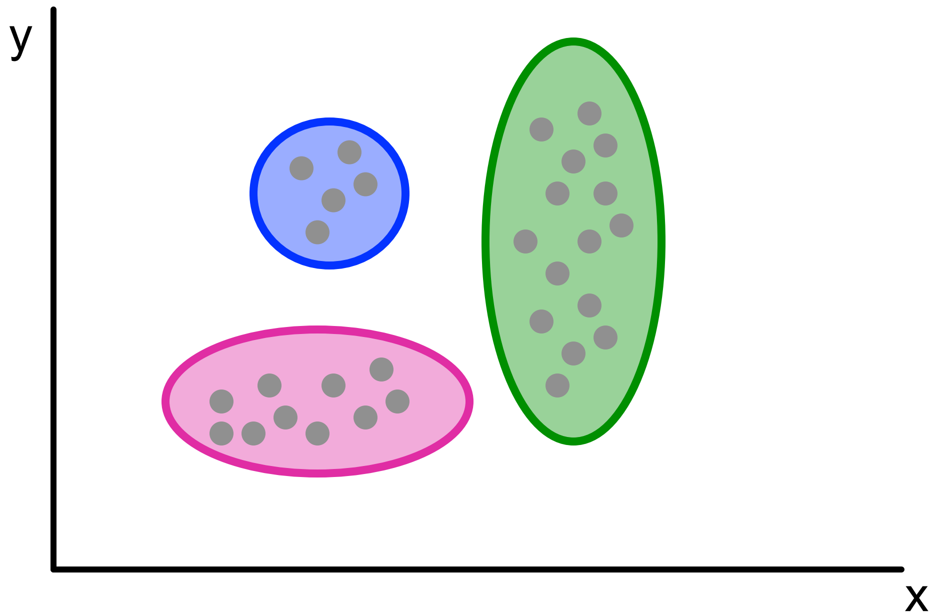


- Can quantify the likelihood of finding a point in the cluster at a given distance from the centroid along each dimension

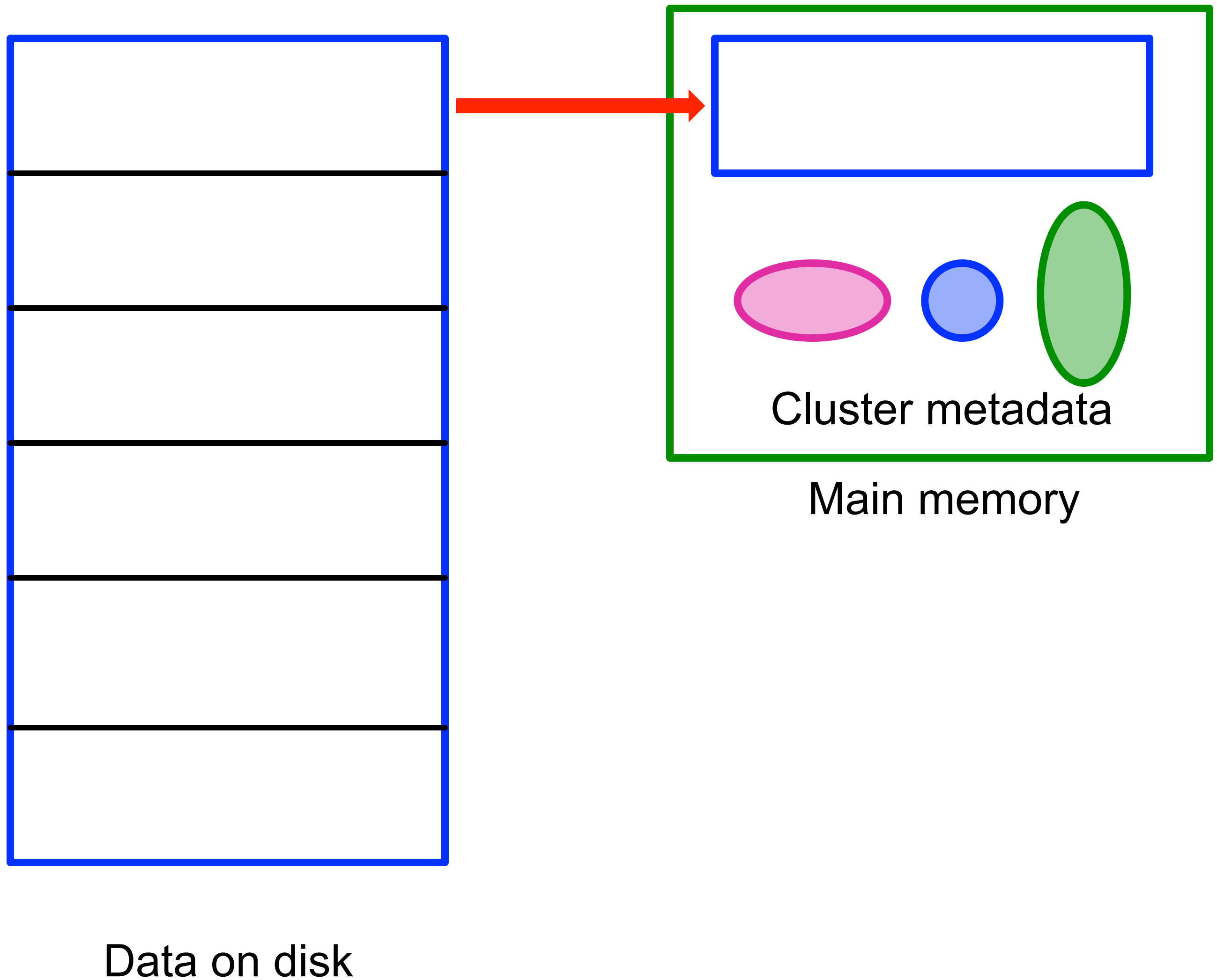
- Standard deviations in different dimenions may vary

BFR Clusters

- Normal distribution assumption implies that clusters “look like” axis-aligned ellipses



BFR Algorithm: Overview



BFR Algorithm

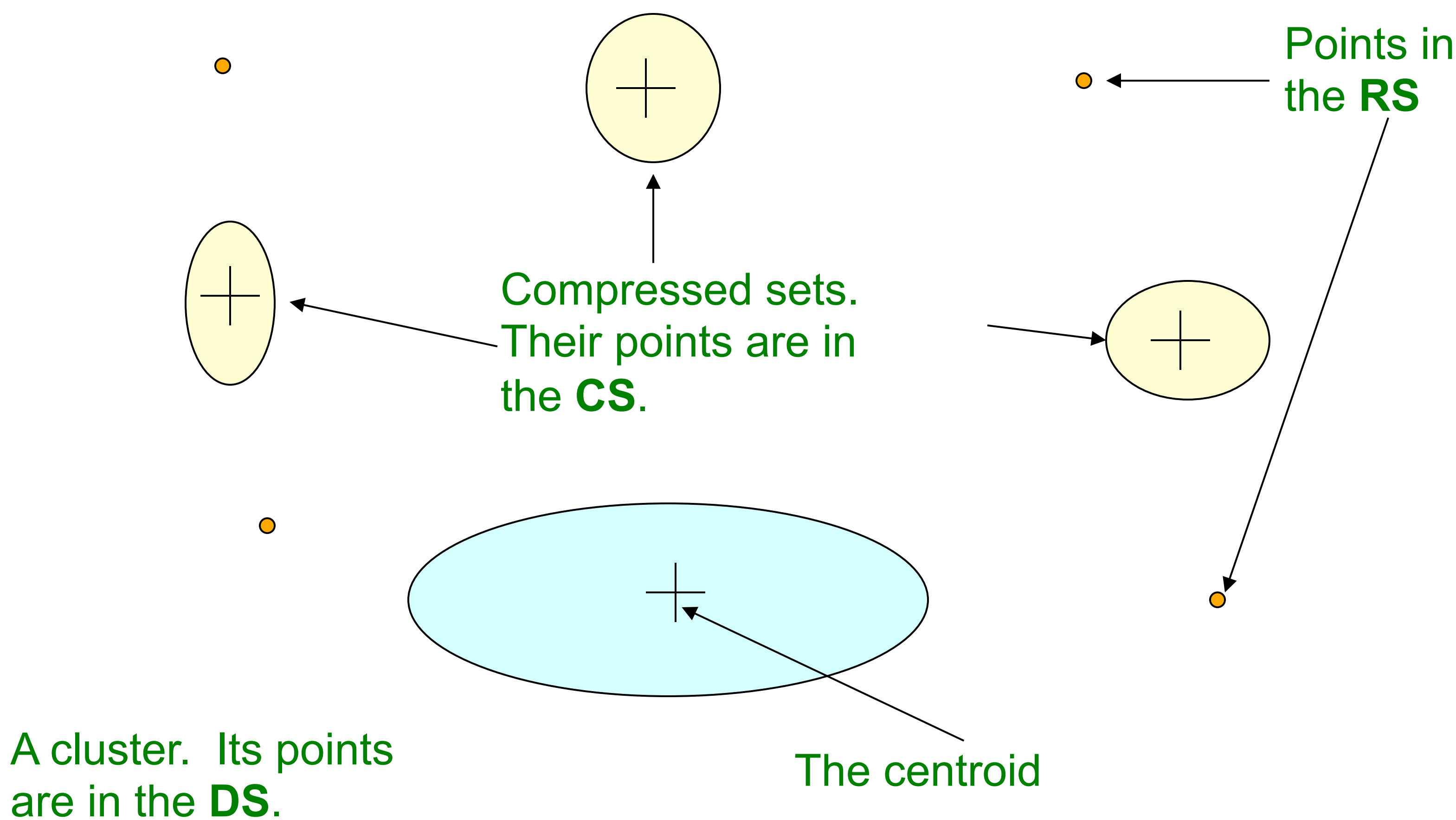
- Points are read from disk one main-memory-full at a time
- Most points from previous memory loads are summarized by **simple statistics**
- To begin, from the initial load we select the initial ***k*** centroids by some sensible approach
  - Using one of the techniques from the k-Means lecture

Three Classes of Points

3 sets of points which we keep track of:

- **Discard set (DS):**
  - Points close enough to a centroid to be summarized
- **Compression set (CS):**
  - Groups of points that are close together but not close to any existing centroid
  - These points are summarized, but not assigned to a cluster
- **Retained set (RS):**
  - Isolated points waiting to be assigned to a compression set

BFR: "Galaxies" Picture

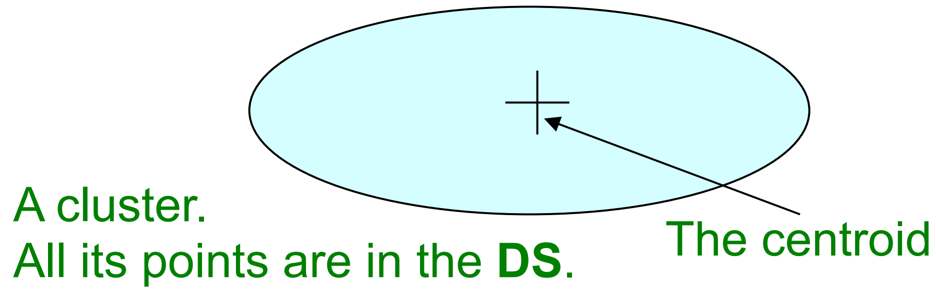


Discard set (DS): Close enough to a centroid to be summarized
Compression set (CS): Summarized, but not assigned to a cluster
Retained set (RS): Isolated points

Summarizing Sets of Points

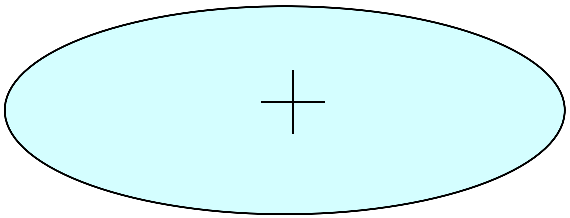
For each cluster, the discard set (DS) is <u>summarized</u> by:

- The number of points, ***N***
- The vector ***SUM***, whose i^{th} component = sum of the coordinates of the points in the i^{th} dimension
- The vector ***SUMSQ***: i^{th} component = sum of squares of coordinates in i^{th} dimension

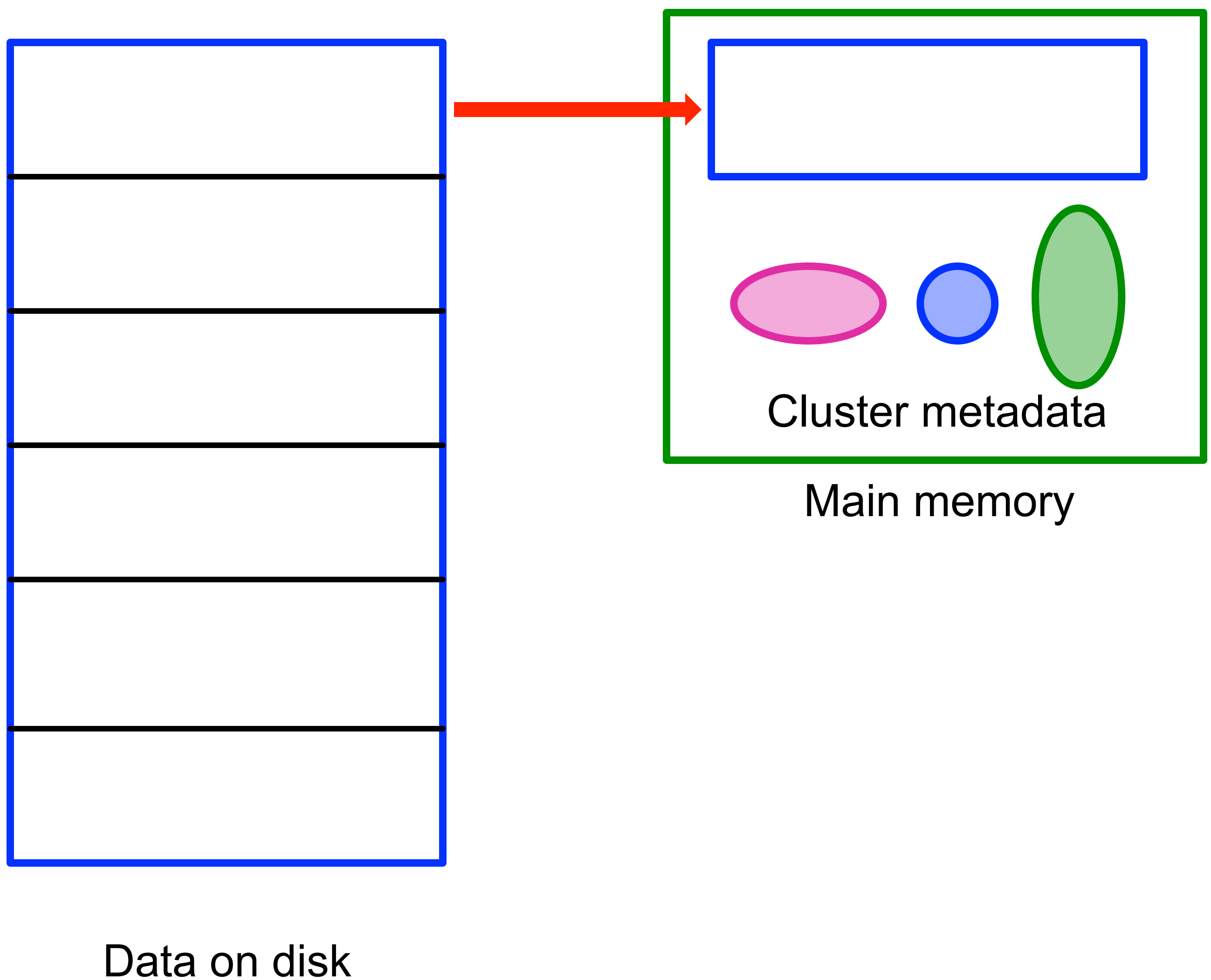


Summarizing Points: Comments

- **2*d* + 1** values represent any size cluster
  - ***d*** = number of dimensions
- Average in **each dimension** (**the centroid**) can be calculated as $\mathbf{SUM}_i\,/\,N$
  - $\mathbf{SUM}_i = i^{th}$ component of SUM
- Variance of a cluster's discard set in dimension i is: $(\mathbf{SUMSQ}_i\,/\,N) - (\mathbf{SUM}_i\,/\,N)^2$
  - And standard deviation is the square root of that
- **Next step: Actual clustering**



BFR Algorithm: Overview



Processing a chunk of points (1)

- Find those points that are "**sufficiently close**" to a cluster centroid
- Add those points to that cluster and the **DS**
  - Then discard the point
- **DS set:** Adjust statistics of each cluster to account for newly added points
  - Add ***N***s, ***SUM***s, ***SUMSQ***s

Discard set (DS): Close enough to a centroid to be summarized.
Compression set (CS): Summarized, but not assigned to a cluster
Retained set (RS): Isolated points

Processing a chunk of points (2)

- The remaining points are not close to any cluster

- Use any main-memory clustering algorithm to cluster these points and the old **RS**
  - Clusters go to the **CS**; outlying points to the **RS**

Discard set (DS): Close enough to a centroid to be summarized.
Compression set (CS): Summarized, but not assigned to a cluster
Retained set (RS): Isolated points

Processing a chunk of points (3)

- Consider merging compressed sets in the **CS**

- If this is the last round, merge all compressed sets in the **CS** and all **RS** points into their nearest cluster

Discard set (DS): Close enough to a centroid to be summarized.
Compression set (CS): Summarized, but not assigned to a cluster
Retained set (RS): Isolated points

A Few Details…

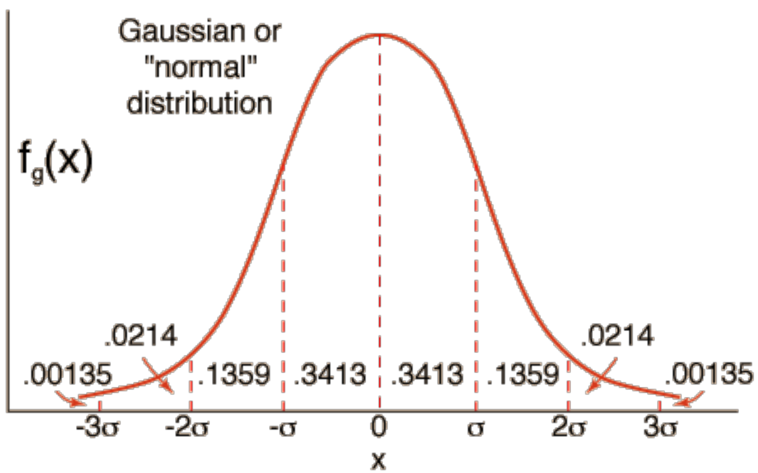
- Q1) How do we decide if a point is "close enough" to a cluster that we will add the point to that cluster?

- Q2) How do we decide whether two compressed sets (CS) deserve to be combined into one?

How Close is Close Enough?

- **Q1) We need a way to decide whether to put a new point into a cluster (and discard)**

- **BFR approach:**
  - The **Mahalanobis distance** is less than a threshold
  - **High likelihood of the point belonging to currently nearest centroid**



Mahalanobis distance

- Cluster *C* has centroid $(c_1,\ldots,c_d)$ and standard deviations $(\sigma_1,\ldots\sigma_d)$
- Point P = $(x_1,\ldots,x_d)$

- Normalized distance in dimension i:

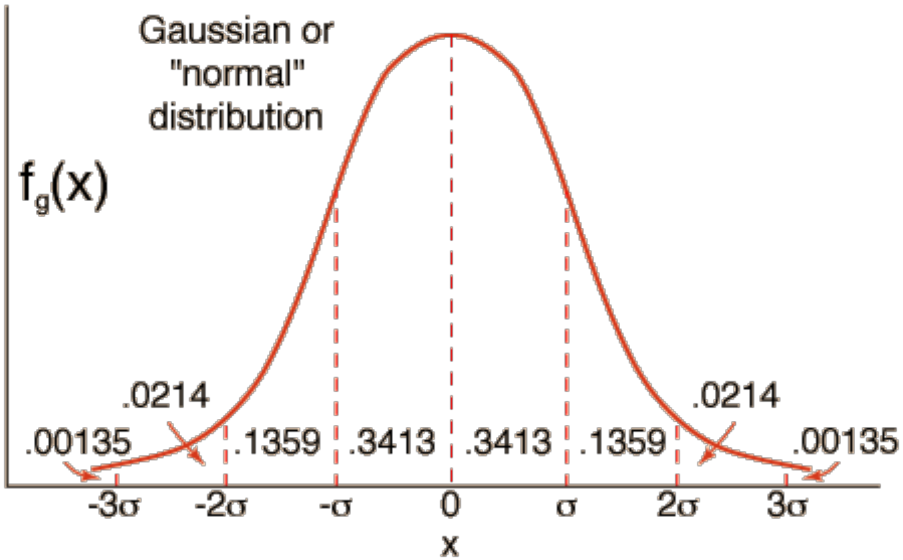
  $y_i = (x_i - c_i)/\sigma_i$

- MD of point P from cluster C:

$$\sqrt{\sum_{i=1}^{d} y_i^2}$$

Mahalanobis Acceptance Criterion

- Suppose point P is one standard dimension away from centroid in each dimension
  - Each y_i = 1 and so the MD of P is $\sqrt{d}$



68% of points have MD $\leq \sqrt{d}$

95% of points have MD $\leq 2\sqrt{d}$

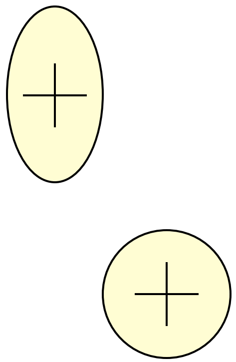
99% of points have MD $\leq 3\sqrt{d}$

Accept point P into cluster C if its MD from cluster centroid is less than a threshold e.g., $3\sqrt{d}$

Should 2 CS clusters be combined?

Q2) Should 2 CS subclusters be combined?

- Compute the variance of the combined subcluster
  - ***N***, ***SUM***, and ***SUMSQ*** allow us to make that calculation quickly
- Combine if the combined variance is below some threshold



- **Many alternatives:** Treat dimensions differently, consider density