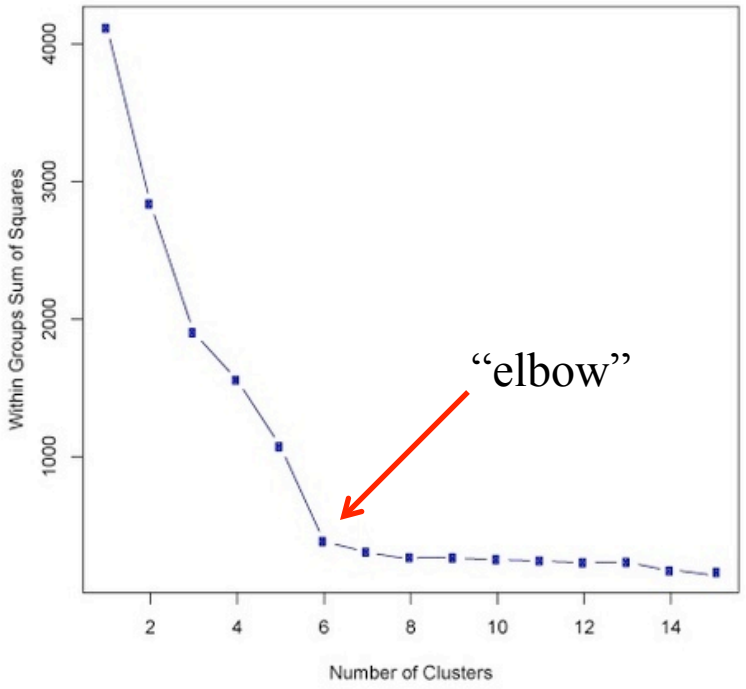
Choosing the *K* in *K*-Means

- Hard problem! Often no "correct" answer for unlabeled data
- Many proposed methods! Here are a few:
- Try several values of *K*, see which is best, via cross-validation.
  - Metrics: sum-squared error, sum-squared separation, penalty for too many clusters
- Start with $K = 2$. Then try splitting each cluster.
  - New means are one sigma away from cluster center in direction of greatest variation.
  - Use similar metrics to above.

- "Elbow" method:
  - Plot *SSE* vs. *K*. Choose *K* at which SSE (or other metric) stops decreasing abruptly.



  - However, sometimes no clear "elbow"

- Assume each cluster is Gaussian
  - Run *K*-means with increasing *K* until a statistical test accepts hypothesis that data in each cluster is Gaussian with respect to the cluster center.

- Many other proposed methods