Whats 'Natural' k to use for this data?
Sometimes we don't know which 'k' to use.

Purity

→ intra-Class (intra-cluster) similarity is high
— inter-class similarity is low

in other words, data inside clusters should be consistent, while data among different clusters should differ.

Intra Cluster distance

avg. pairwise intra-cluster distance $G_{ri} = \frac{\sum_{x,y \in C_i} d(x,y)}{(2|C_i|)}$ (cluster i, distance fn d())

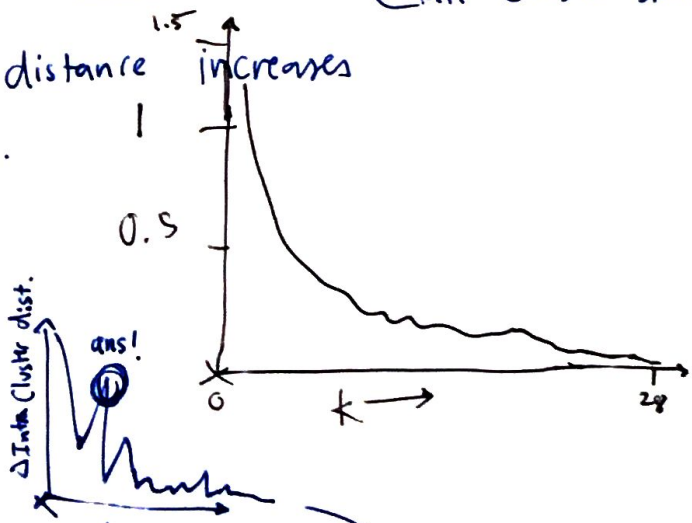
do for all clusters in Cluster Set.
weighted (avg) intra-cluster diff for level j is

$$W_j = \frac{\sum_{i=1}^{j} |C_i| G_{ri}}{N}$$

We want to minimize inter Cluster distance (& maximize inter cluster similarity)

Graph-based Approach:

* when we merge, intra-Cluster distance increases



- Choosing min. of this is not useful.
  will end up getting k=N,
- So, we look @ slope!

$$W_j' = \frac{(W_{j+1} - W_{j-1})}{2}$$

A: — Choose place where there's steepest absolute slope

$$k = \text{argmax}_j (|W_j'|)$$

? → How about where is greatest Δ Slope? — 2nd derivative!
(knee of Graph)