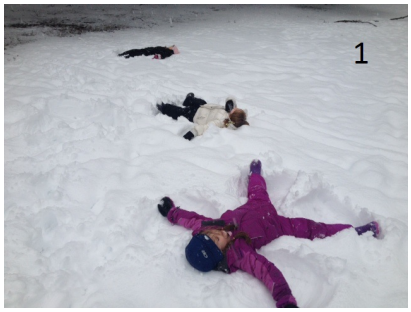
# CPSC 320 Notes, Clustering

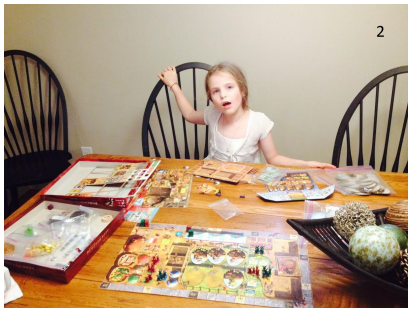
January 22, 2015

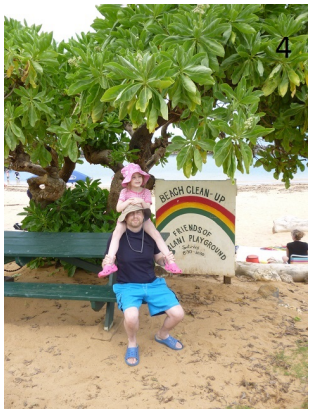
You're working on software to manage people's photos. You get a bunch of uncategorized photos, a number of categories to group them into, and a "similarity metric". A 0 similarity indicates two photos are nothing like each other; a 1 indicates two photos are exactly the same. All other similarities are in between. Your job is to create a "categorization": the requested number of categories, where a category is just a list (of non-zero length) of the photos contained in that category. Every photo belongs to some category, and no photo belongs to more than one category. (I.e., a categorization is a "partition".)

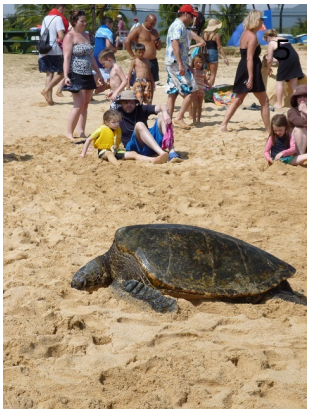
1. Explain how we can interpret the input (except the number of categories) as a graph.

2. Draw a reasonable graph for this set of pictures (if it helps, the pictures are numbered):










3. Divide these into the best three categories you can based on the similarities you chose. (That is, solve the problem for 3 categories.)

4. What metric should we use to measure the "goodness" of a categorization—like yours from the previous part? Give at least two possibilities.

5. From here on, we'll all use the same "goodness" measure.

   First, define the similarity between two categories $C_1$ and $C_2$ to be the maximum similarity between any pair of photos $p_1 \in C_1$ and $p_2 \in C_2$.

   Then, the "goodness" of a categorization is the negation of the maximum similarity between any two of its categories. (So, the best "goodness" is 0.)

   Measure the "goodness" of your categorization from the previous page.

6. Find the edge in your graph on the previous page with the highest similarity. Should the two photos incident on that edge go in the same category?

7. Propose a greedy algorithm to create your categorization.

# 1 Challenge Problem

Using the algorithm we created, think of a **principled** way to decide how many categories there should be given no more input than the similarity graph.