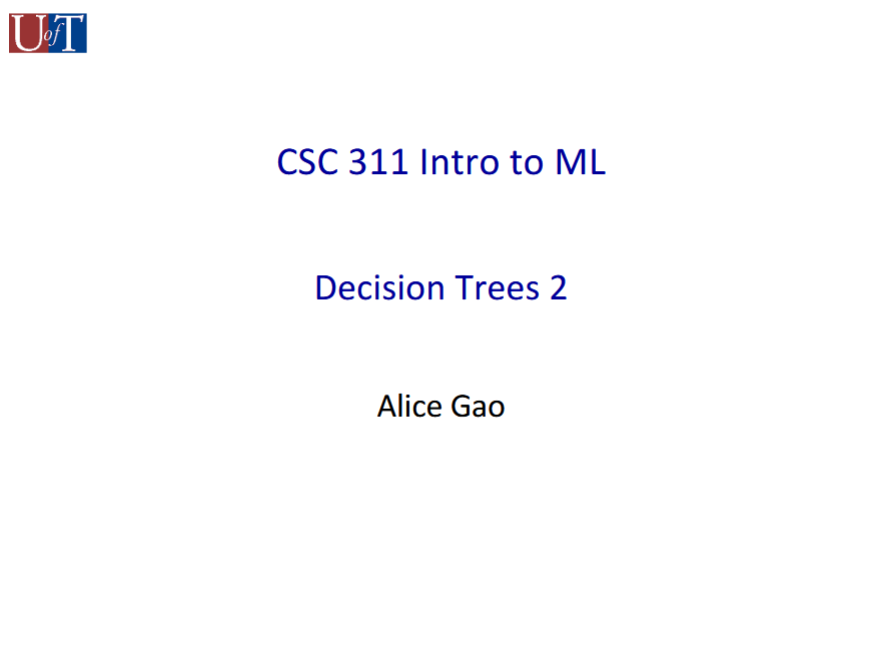
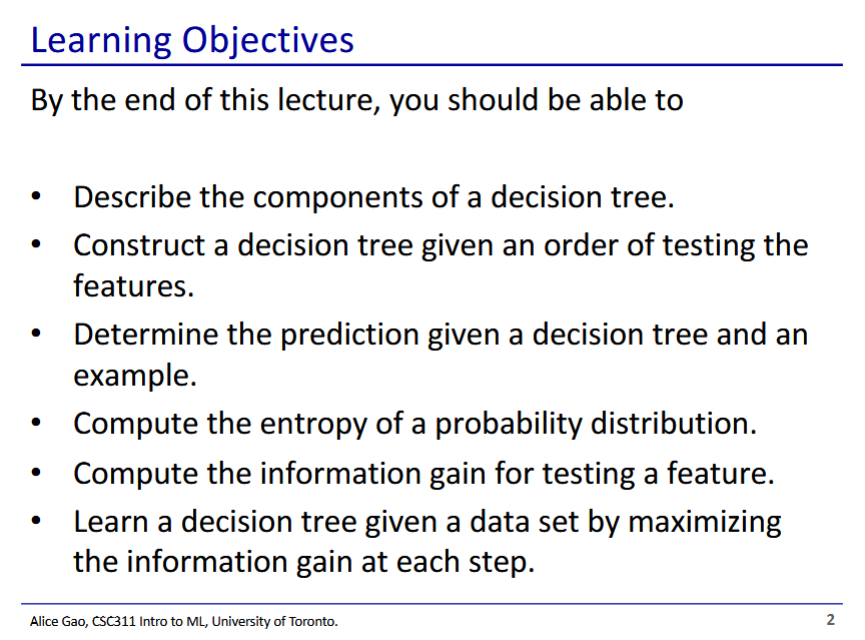
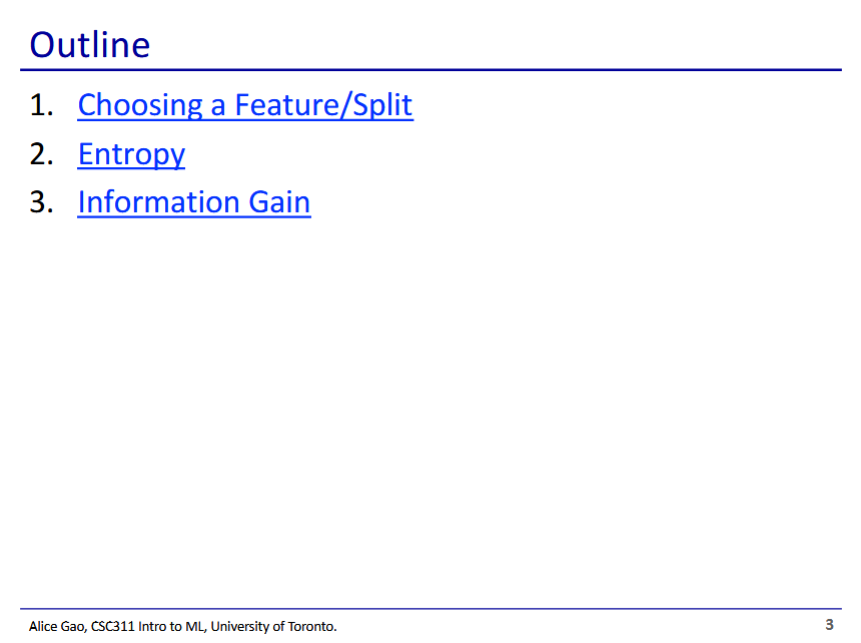
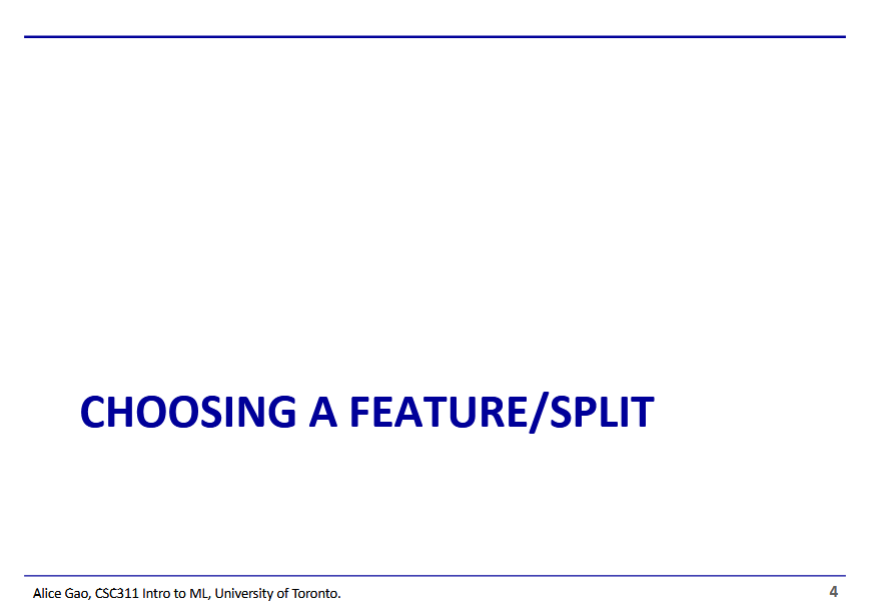
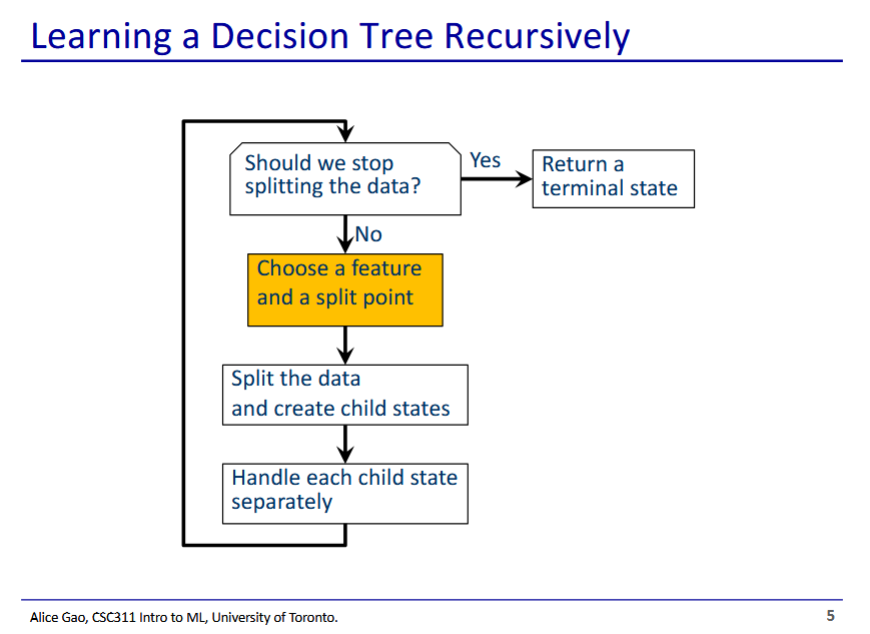
| **Admin stuff**   * Assignment 1 due thursday Jan 25th at **NOON**   + Only covers KNN and decision trees * TA office hours on Monday for extra lab support   **Choosing a feature and split point**   * Theoretical: building the optimal tree   + The optimal tree is the smallest tree that correctly classifies each data point     - Tree that's too small may underfit the data (not capturing important details)     - Tree that's too big may overfit the data (includes too much noise)   + Finding the optimal tree is an NP complete problem, which means it is often impractical to do * Reality: greedy algorithm   + A greedy algorithm can give us an adequate decision tree in a reasonable time   + We pick the best feature to split at each step (which may not be the best split overall)   + Which split is best is evaluated using a loss function     - Many possible loss functions, but we will look at **information gain** in this course     - Best split is along the feature that has the highest information gain for our label   **Entropy (H(X))**   * A method to quantify the uncertainty of a random distribution   + Represents how much certainty is gained about a distribution if we observe a random draw from it * Equation:   + - Flip sign since log is negative for values between 0 and 1 * Entropy is high (1) when probability is distributed evenly over outcomes * Entropy is low (0) when probability is concentrated on a few outcomes   + Entropy is always non-negative * We can also calculate the entropy of Y after we have observed a different random variable X   + We calculate entropy of Y for each X=x separately, then take the weighted sum based on   + Observing X will never increase the uncertainty of Y     - At worst it will not change our uncertainty when X and Y are independent   **Information gain**   * A measure of how much observing one random variable X tells us about a different random variable Y * Equation: * We split the feature with the largest information gain |
| --- |



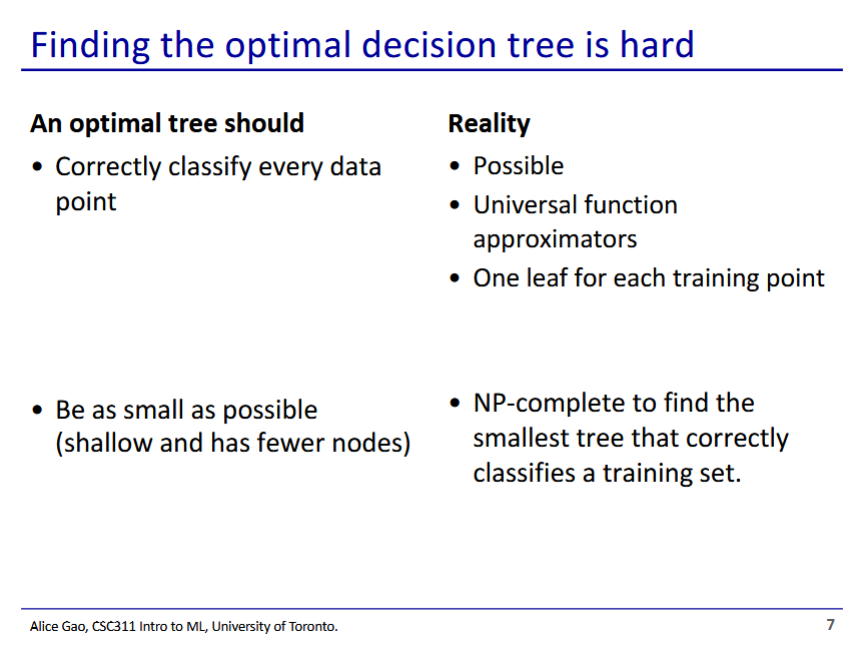




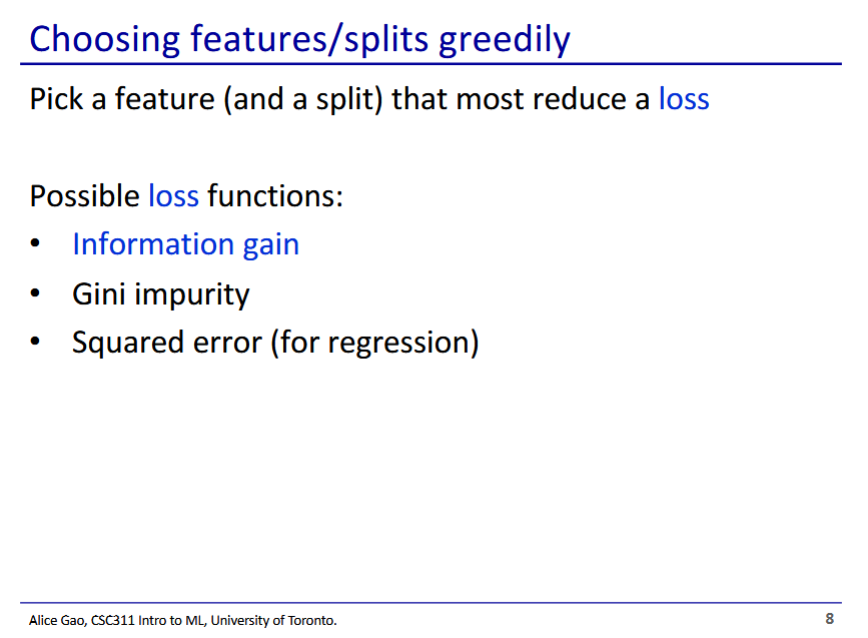




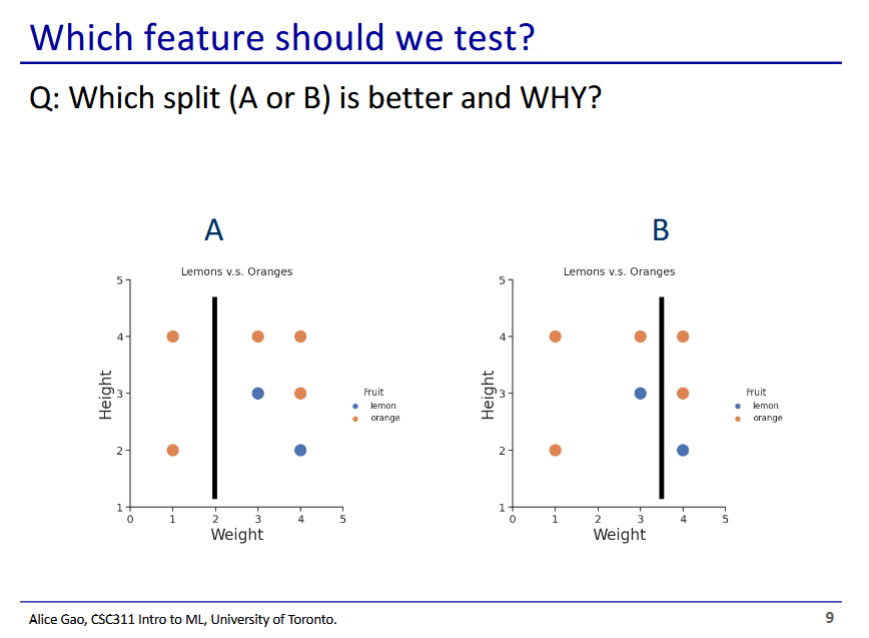
* Choosing what feature and where to split is a very important part of building our decision tree



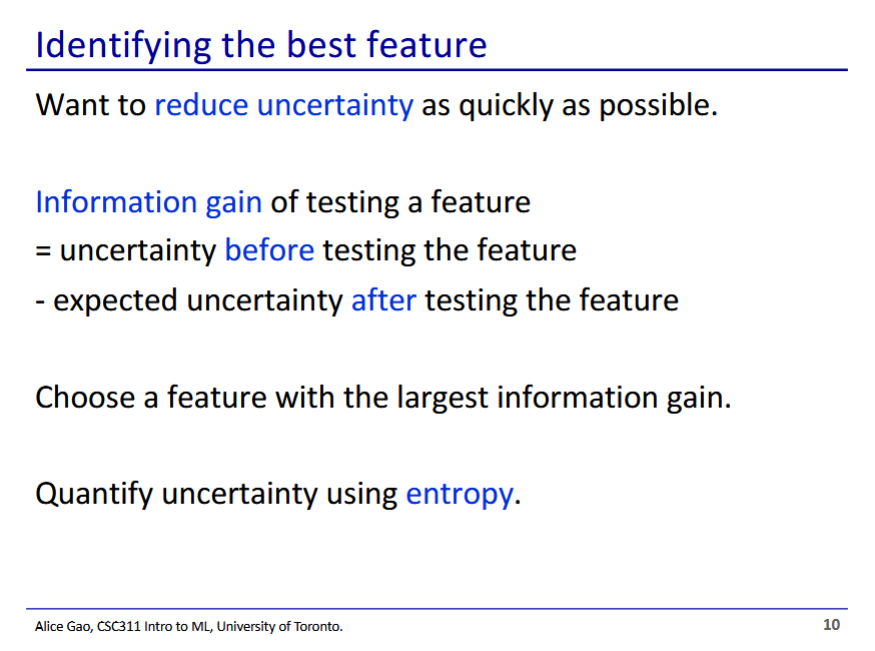
* A good tree is not too small but not too big
  + Needs to be big enough to pick up on important details of data
  + Needs to be small enough to not be too complicated or overfit (large tree includes noise)
* The (theoretically) optimal tree correctly classifies every data point while having the smallest size
* Realistic tree
  + It is possible to correctly classify every data point
    - Decision tree is a very powerful structure (universal function approximator)
  + However finding the optimal tree for the training set is a difficult problem (NP-complete, cannot be done quickly for large datasets)



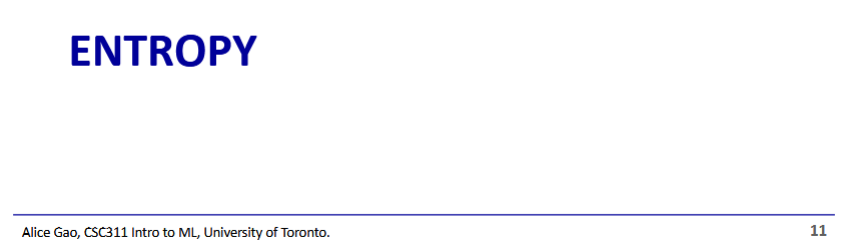
* We can use a greedy algorithm to get a suboptimal but adequate tree in a faster time
  + Generates a tree by making the “best” split at each step, but may not be the best split overall
* Greedy algorithm uses a loss function to gauge how good/bad each step is
  + There are many choices for loss functions
* Information gain also called entropy

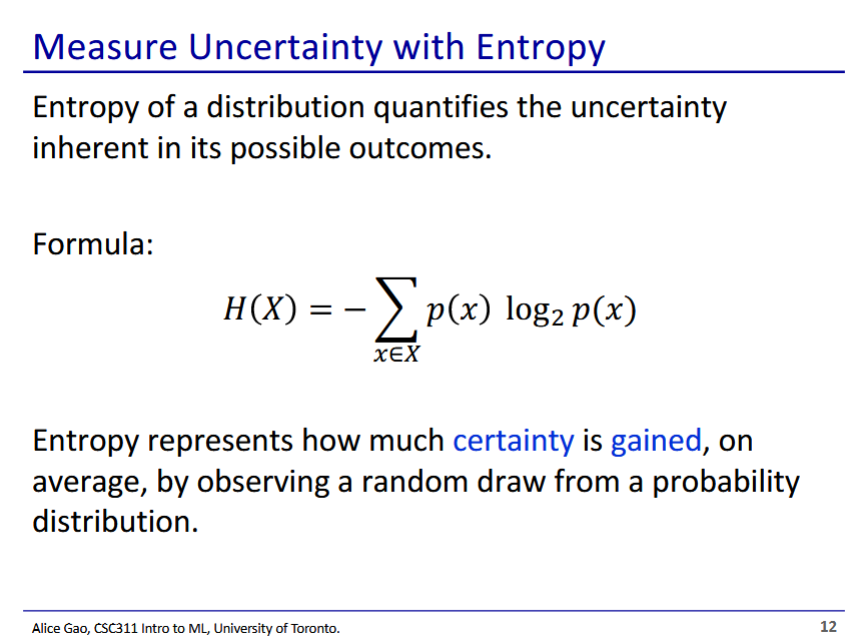


* A seems intuitively better
  + We have only oranges on the left of the threshold and can immediately make a leaf node
  + With B, we still have uncertainty on both sides, as both sides have a mix of both labels
* We can try to quantify which split is better using entropy

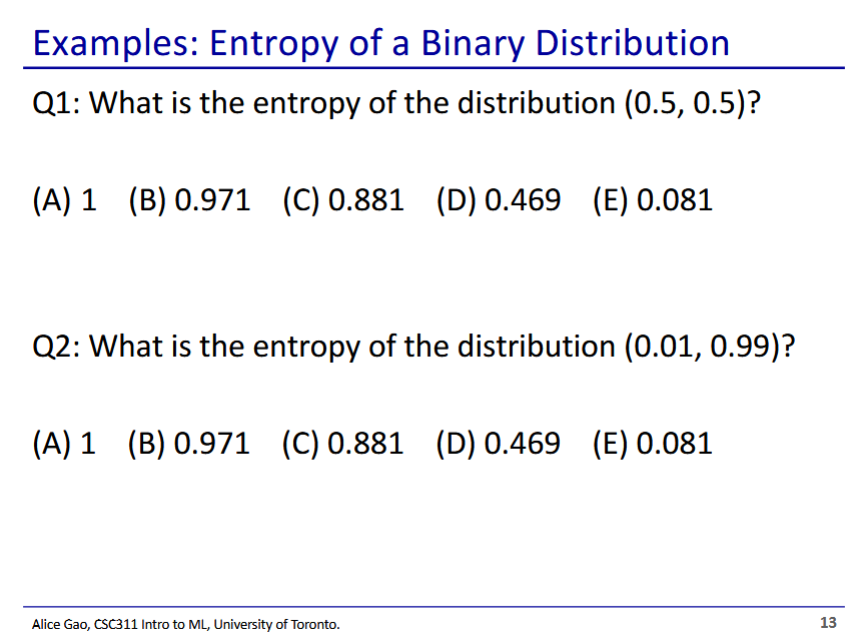


* **Information gain** - change in uncertainty before and after a split

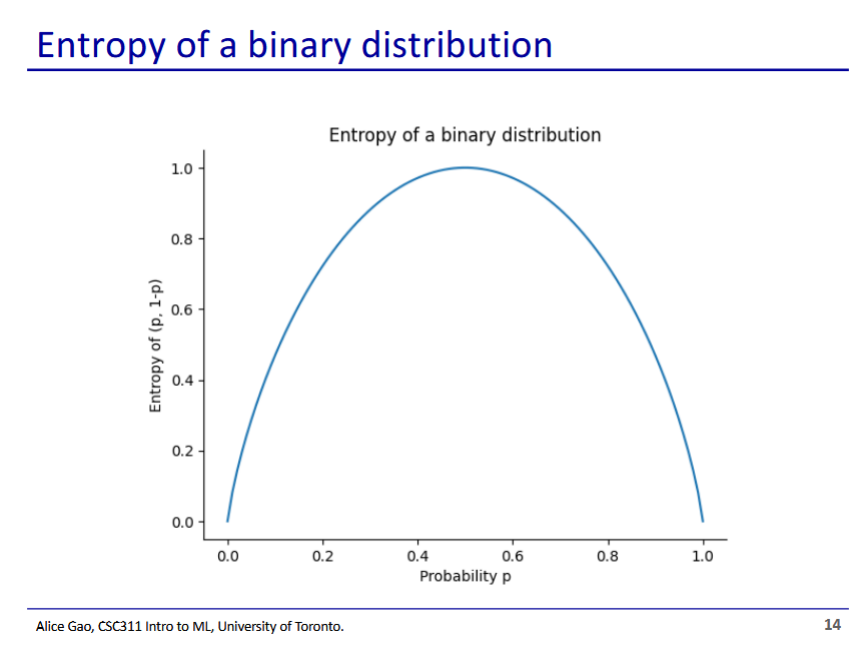




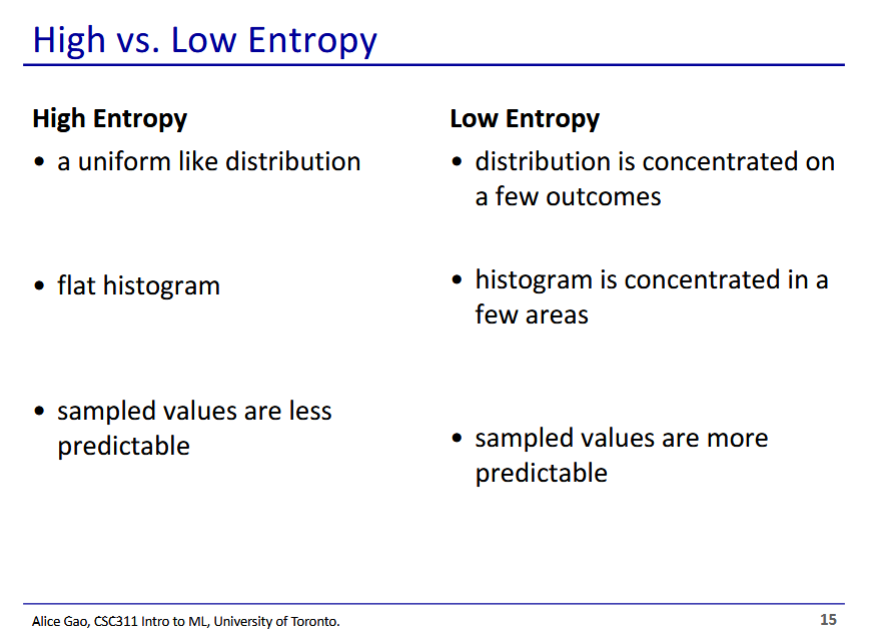
* Entropy is a concept from information theory
  + How much do we learn about the distribution using a random draw?
  + If we have high entropy, a random draw tells us very little
  + If we have low entropy, a random draw tells us a lot
* Calculating entropy
  + Negative of Sum of for each outcome x in the distribution
    - Negative is to flip the sign since log is negative for values between 0 and 1



* Q1 entropy of binary distribution X where P(X=0)=0.5 and P(X=1)=0.5
  + We plug in the formula and solve
* This distribution has very high entropy
* Q2: (0.01, 0.99)
* This distribution has very little entropy

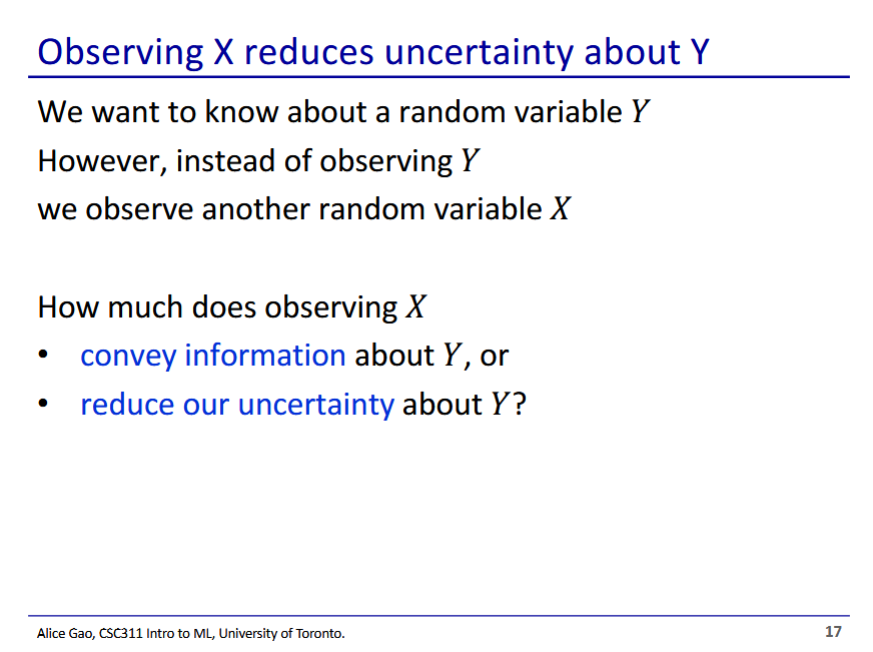


* Graph: shows entropy of each possible binary distribution (p, 1-p)
* Entropy is highest when p is 0.5
  + This is a fair coin
  + We have the most uncertainty about the results
* Entropy is lowest when p approaches 1 or 0

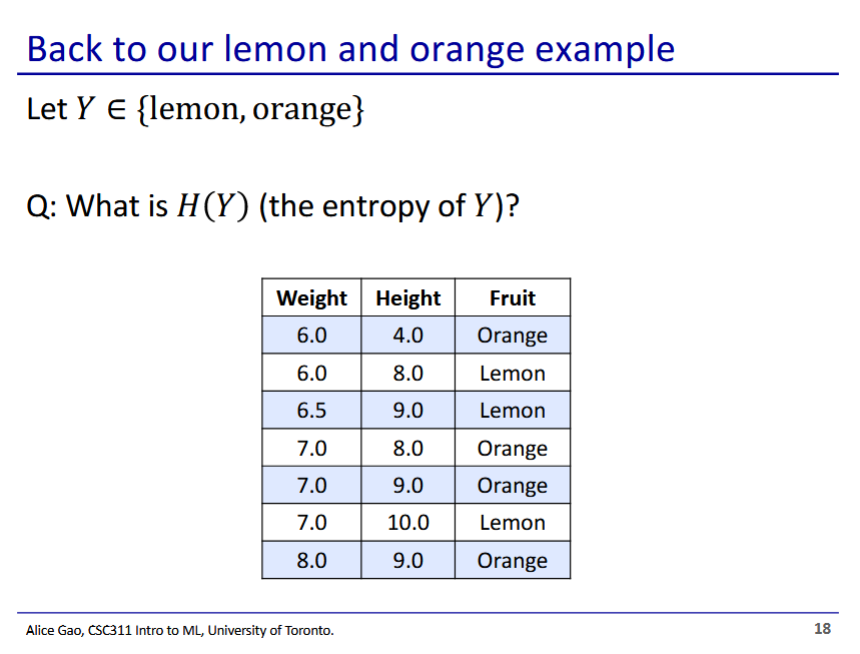


* Entropy is high when probability is distributed evenly over outcomes
* Entropy is low when probability is concentrated on a few outcomes

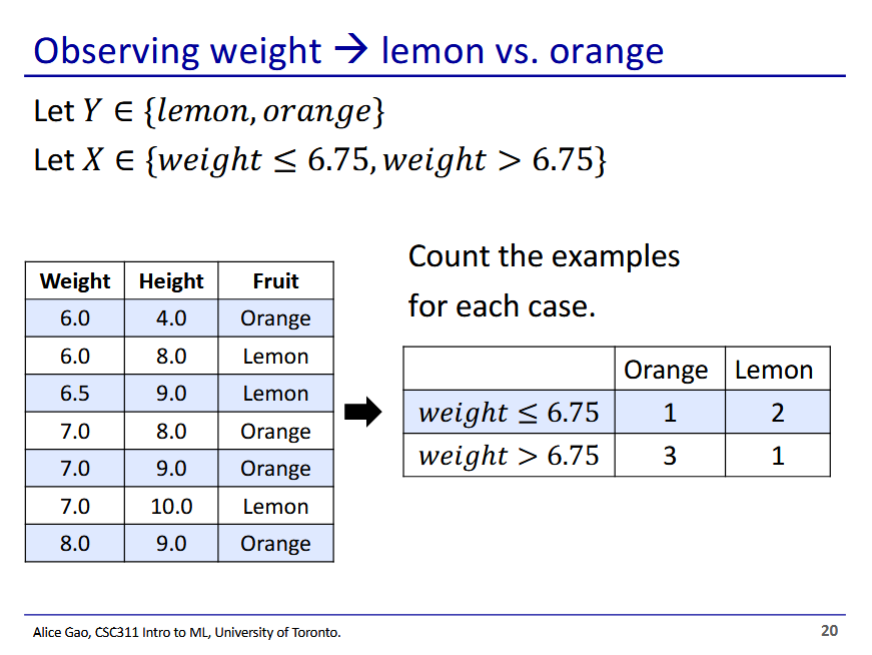




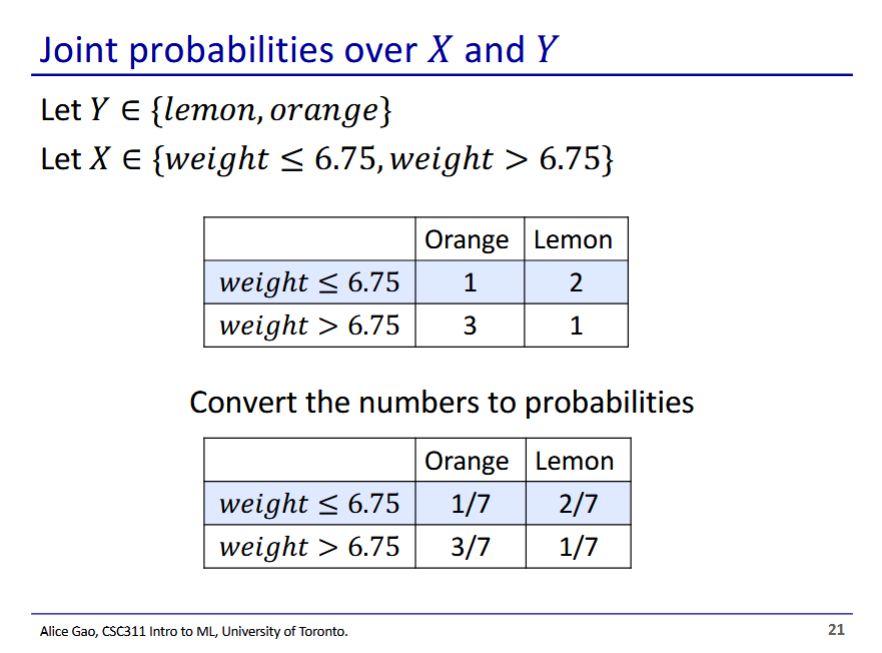
* Information gain comes into play when we have 2 random variables in our data
* How much does observing random variable X reduce our uncertainty of Y?
  + This is useful when we can’t observe Y



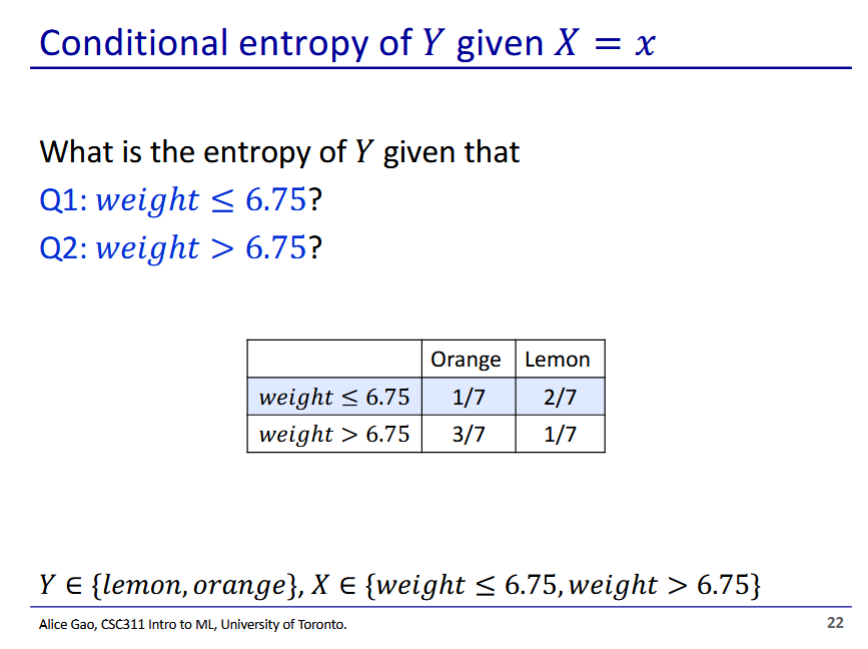
* We have 4 oranges and 3 lemons



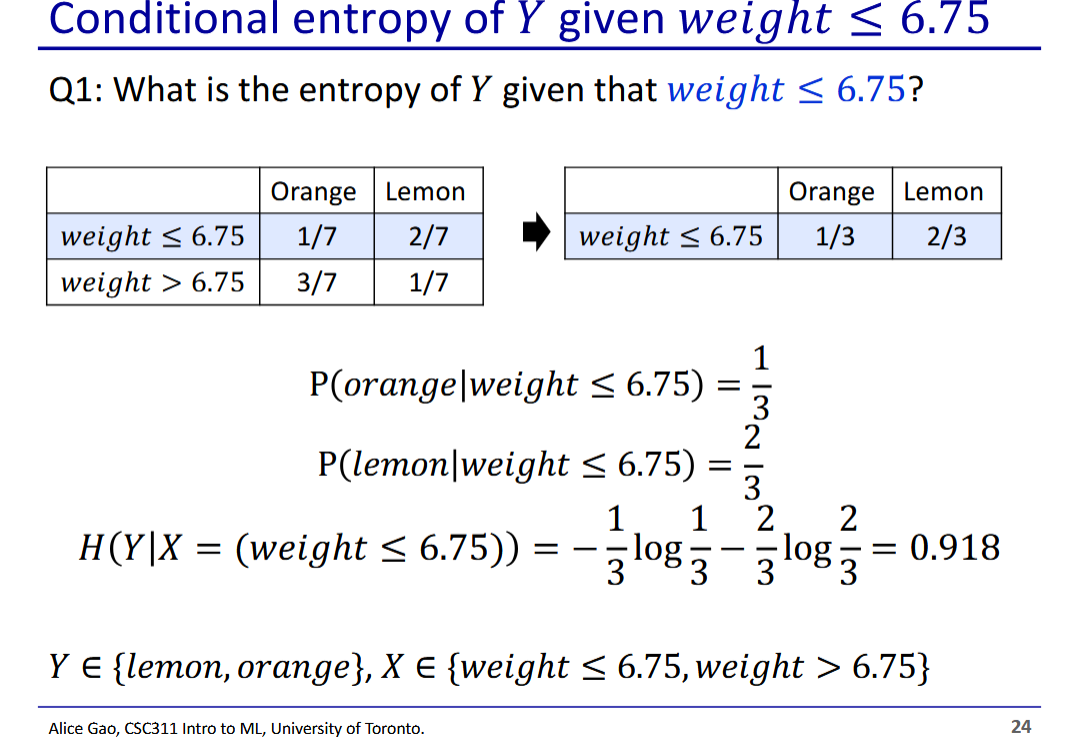
* Let’s see how observing weight can change the entropy of lemon/orange



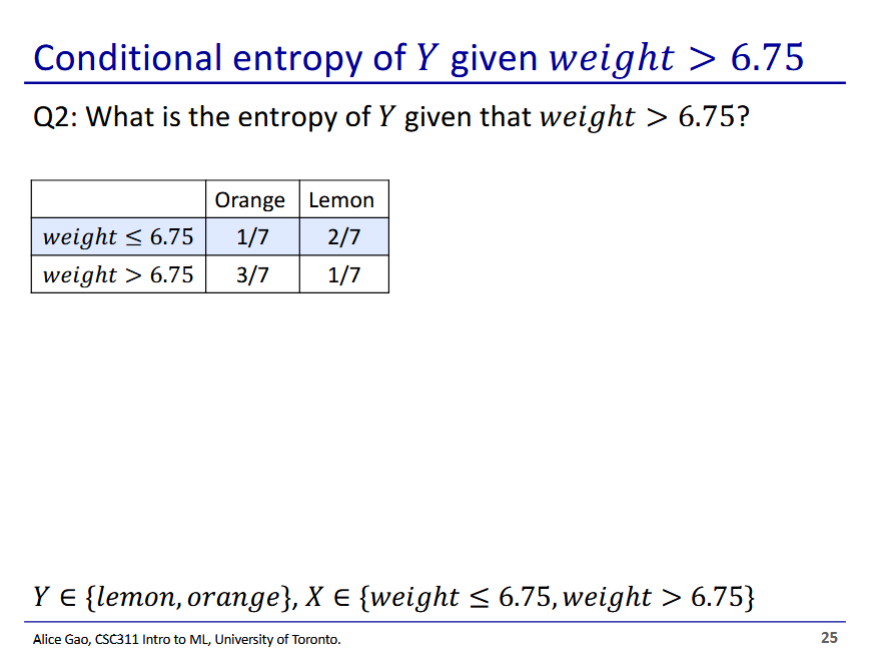
* This probability represents the joint probability between X and Y
  + I.e. the probability that an orange has a weight <= 6.75 is 1/7



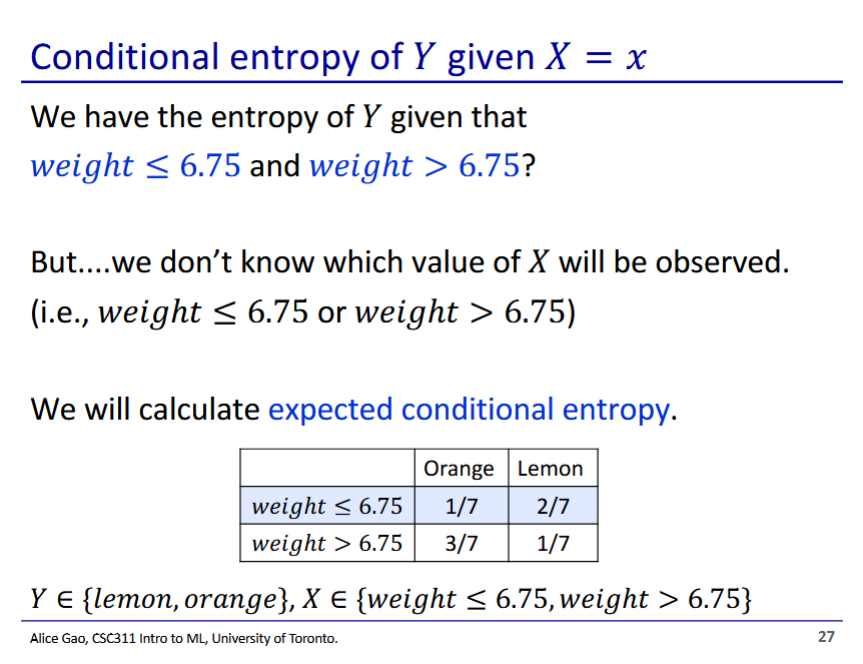
* Answers on subsequent slides



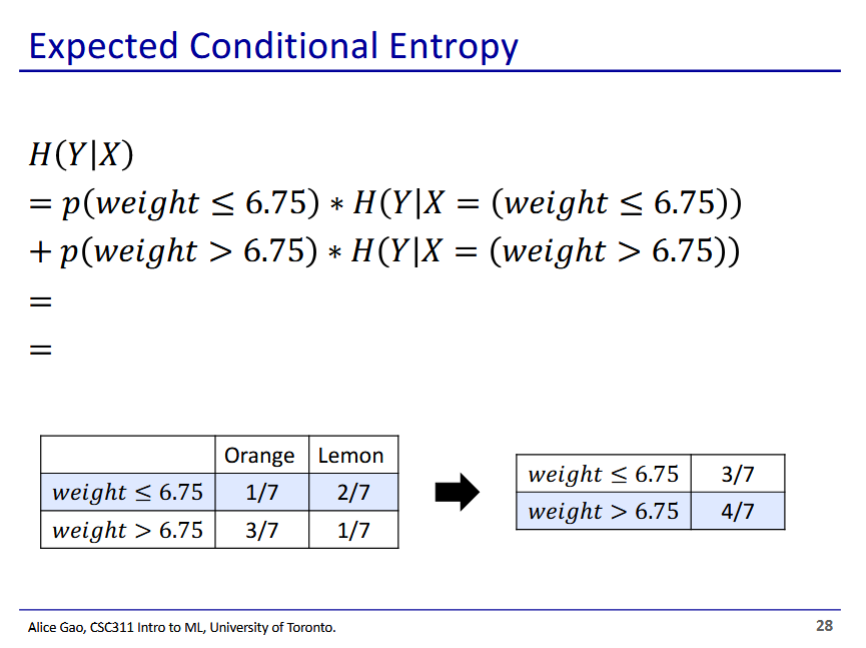
* Q1 answer: 0.918
  + We only look at the top row of the table
  + We renormalise the probabilities on the top row to get a distribution
    - Orange = ⅓, lemon = ⅔
  + Then we plug in the formula



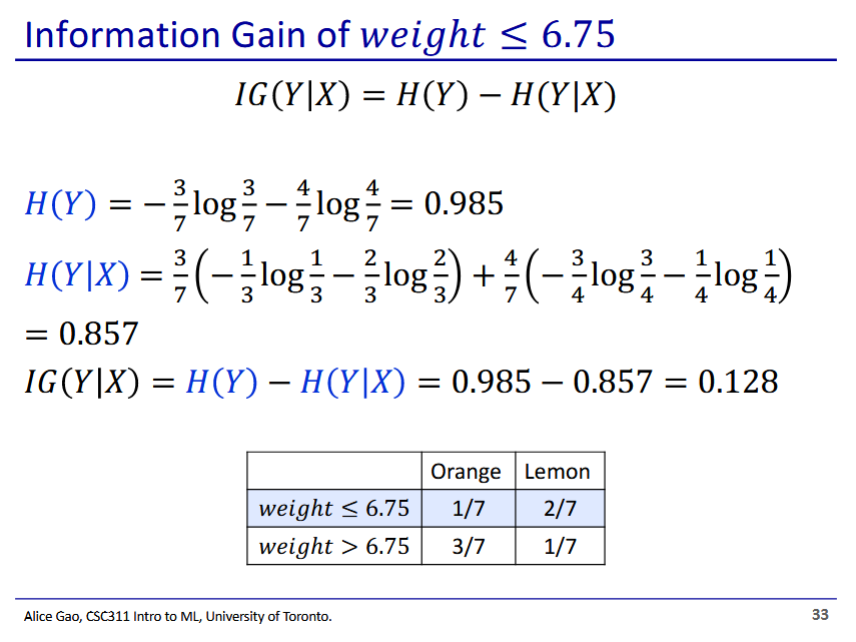
* Q2 answer: 0.811
  + We only look at bottom row
  + Renormalise: orange = ¾, lemon = ¼
  + Plug into formula



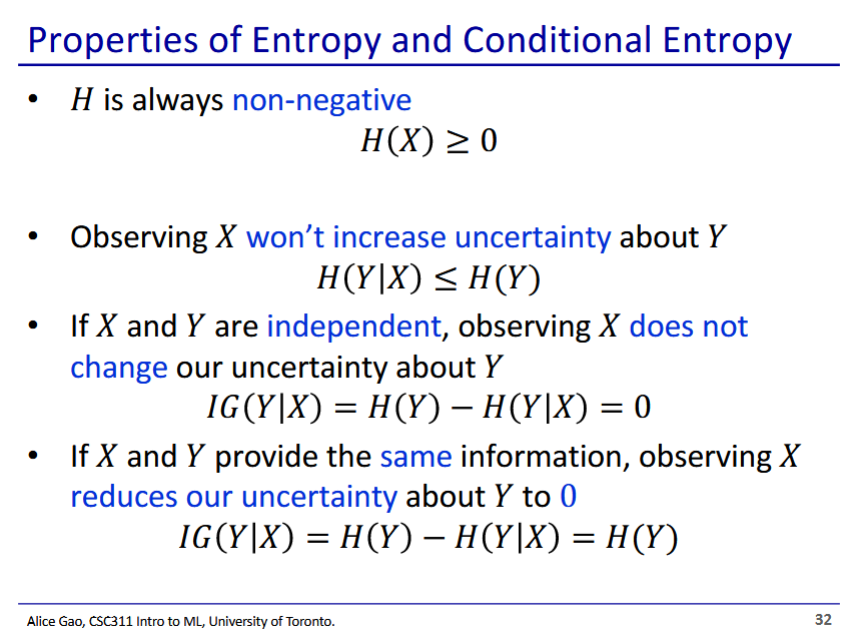
* We need to weigh the different cases
  + The weight of each case is the prior probability of ending up on either side of the split

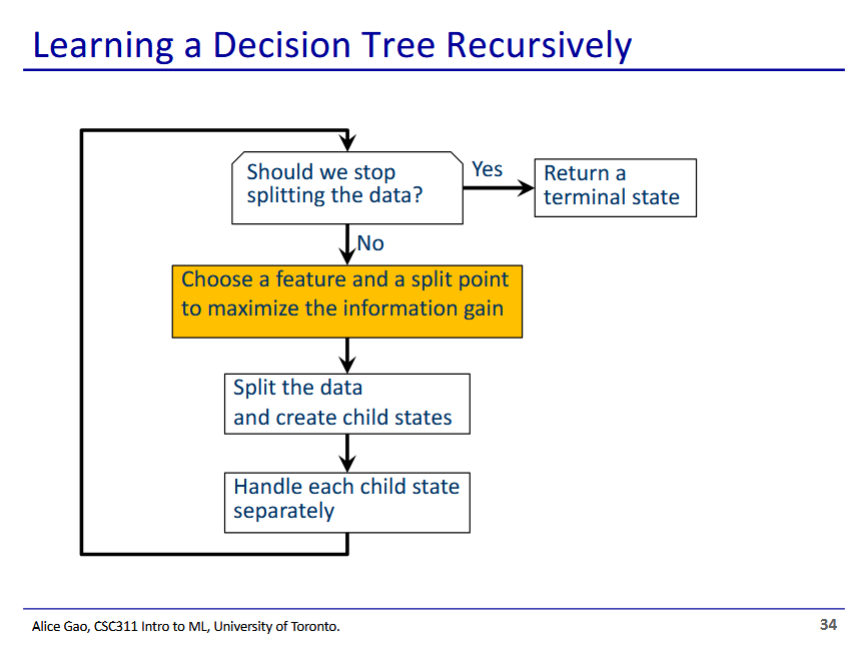


* We then take the weighted average of entropy change for each side of the split to get the expected entropy change

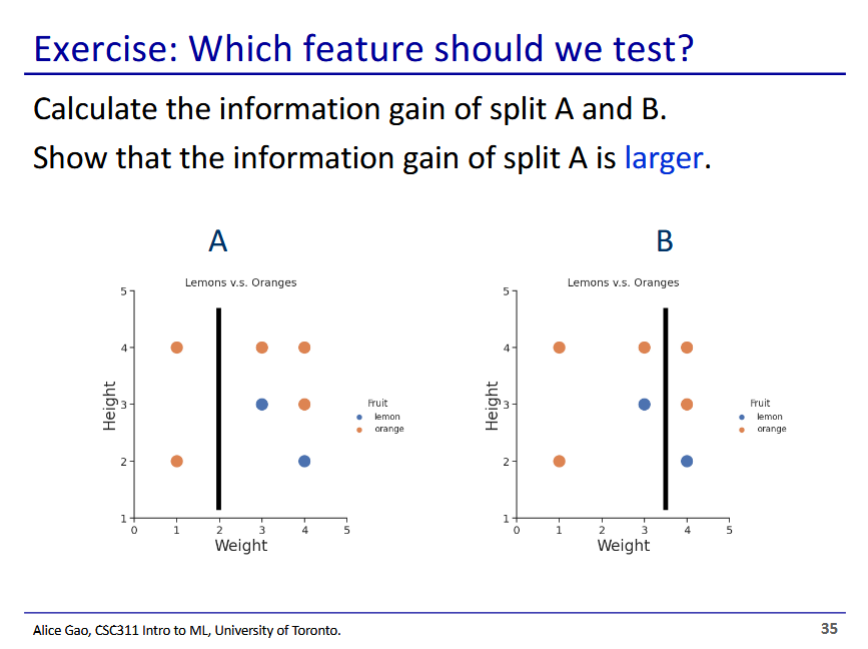


* Now we can calculate the information gain
  + We subtract the entropy of just Y, with the entropy of Y after observing X





* We choose the split point and features that gives the most information gain



* **Solution posted in complete slides on Quercus**
* [**https://q.utoronto.ca/courses/337286/pages/week-2-slides?module\_item\_id=5441188**](https://q.utoronto.ca/courses/337286/pages/week-2-slides?module_item_id=5441188)