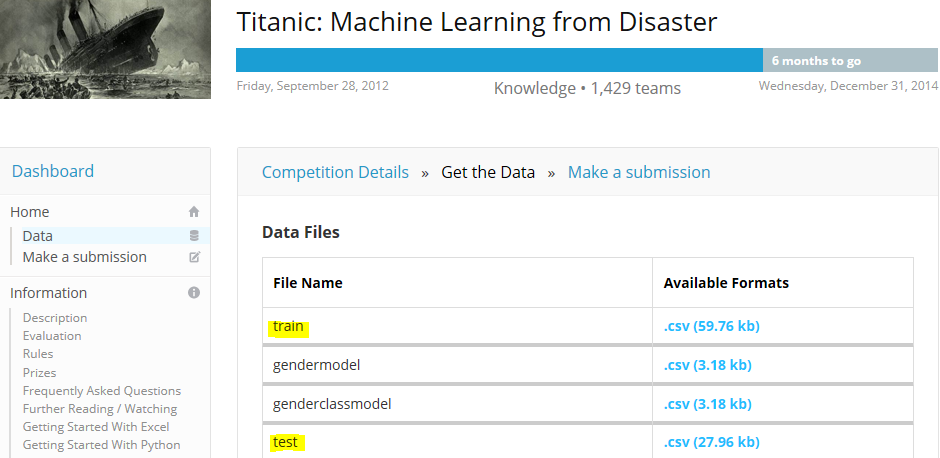
IST565 Data Mining  
Professor Bei Yu

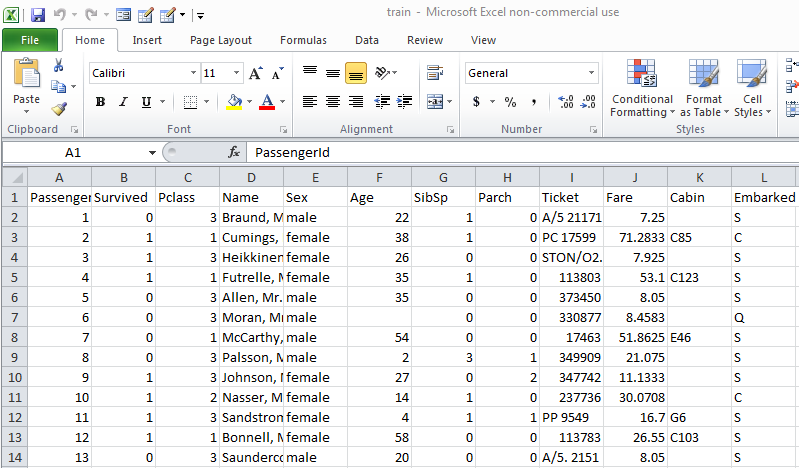
Tutorial on Building Decision Tree for Kaggle Titanic Competition Using Weka

This tutorial will walk you through the process of preparing Kaggle Titanic data for Weka, and then using Weka's J48 algorithm to build a decision tree to predict Titanic survivors, and how to visualize the decision tree to observe what patterns/rules the decision tree model has learned from the training data. This tutorial will also demonstrate how to use information gain as feature ranking methods to explore the relevance between features and prediction target.

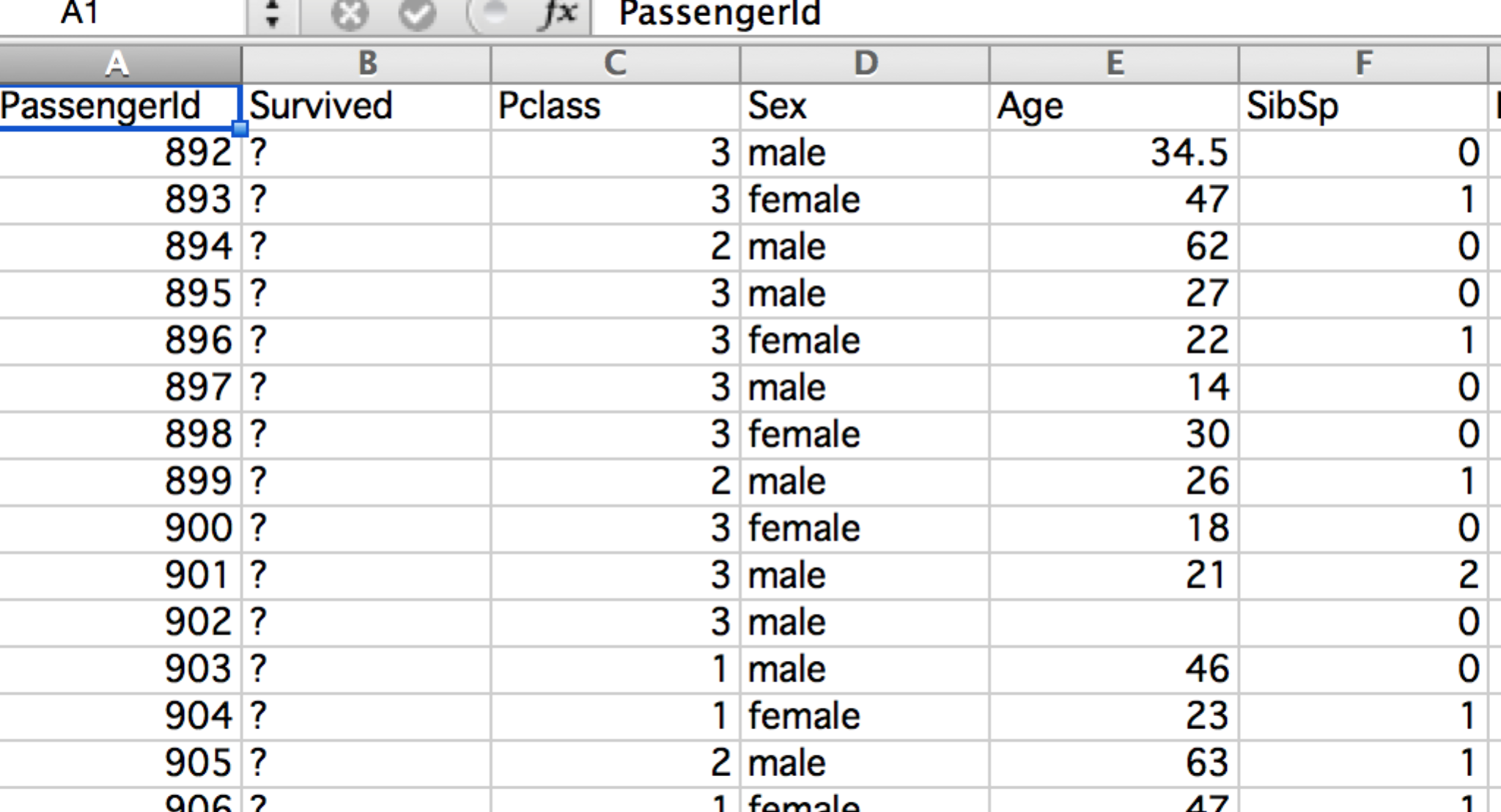
1. Go to [Kaggle](http://www.kaggle.com/c/titanic-gettingStarted/data). Download **test** and **training** data CSV files.



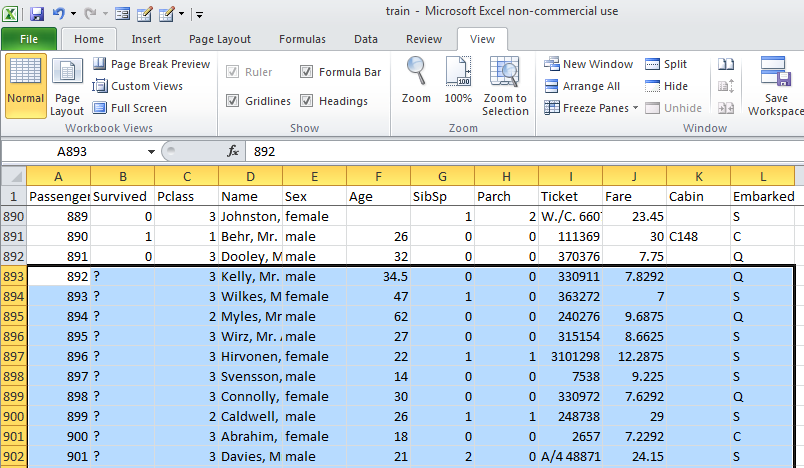
1. Prepare data files.
   1. Open **training** CSV data file in MS Excel.
   2. There are 12 columns (attributes).



* 1. Open **test** CSV data file in MS Excel.
  2. Add column for “Survival” in **test** data file.
  3. Populate “Survived” column with question marks (“?”) in **test** data.



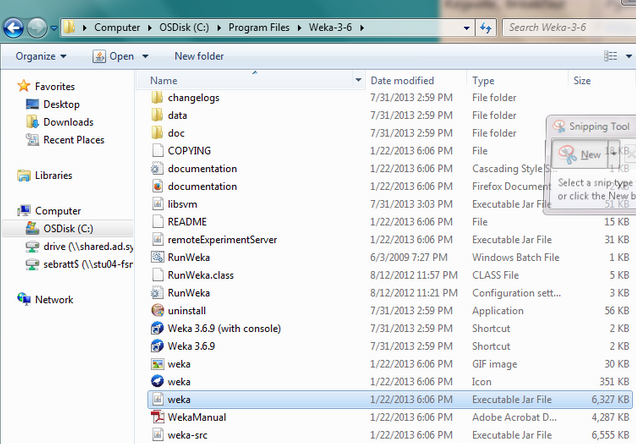
1. Merge training and test data.
   1. Weka will allow us to read in CSV files. But when we read in one training data and one test data Weka will report an error their formats are not compatible, even though they appear identical. In order to allow Weka to recognize both of them as the same data format, we can copy both training and test examples into one file and let Weka read it in, and convert it to its own arff file. Then we separate it again back into train and test in Weka.
   2. Now copy all records from **test** CSV(but **not** the first row of headings) into the **training** CSV file.



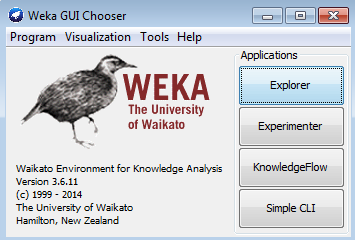
* 1. Delete unnecessary columns (optional). I got rid of “Name” and “Ticket” because these are not relevant or helpful attributes to our classification task.

Now we have preprocessed and merged both files. Save this as a new file (I call mine: “TitanicAll.csv).

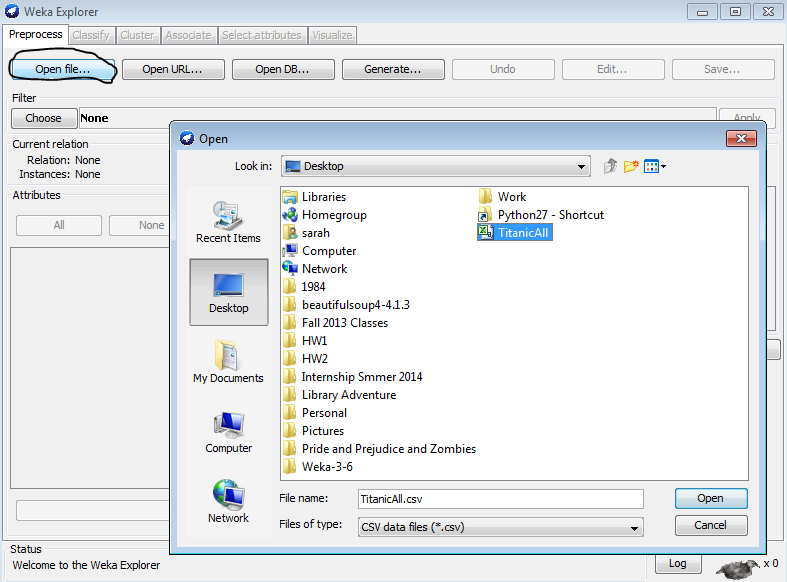
1. Read CSV into Weka
   1. You can use remote lab to access Weka or you can [download Weka](http://sourceforge.net/projects/weka/?source=dlp) (free) onto your own device. On iSchool computers, you can find Weka by going to Computer-> OSdisk\_> program files-> weka 3-6: then select the "executable jar file.”



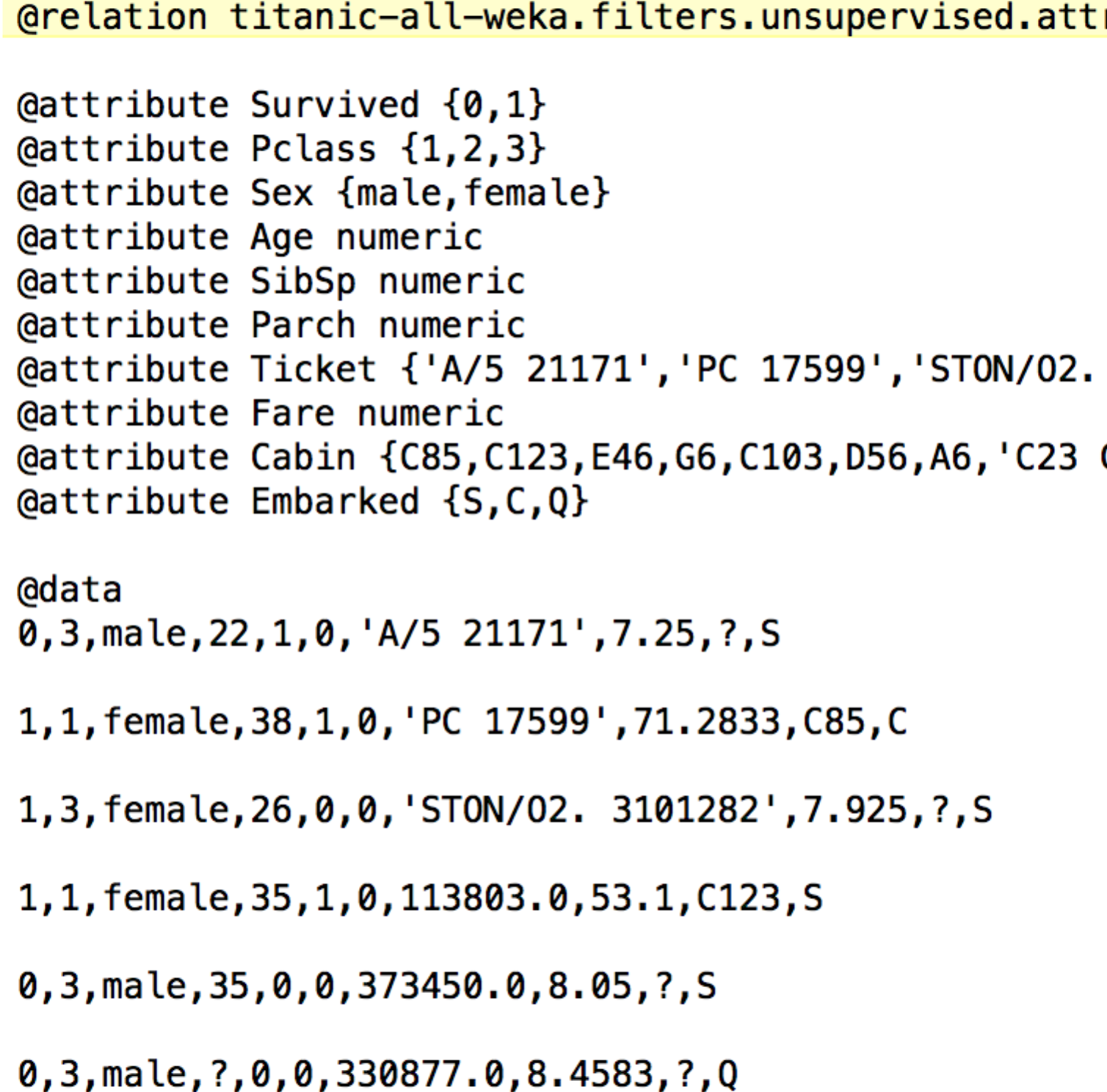
* 1. Open Weka explorer.

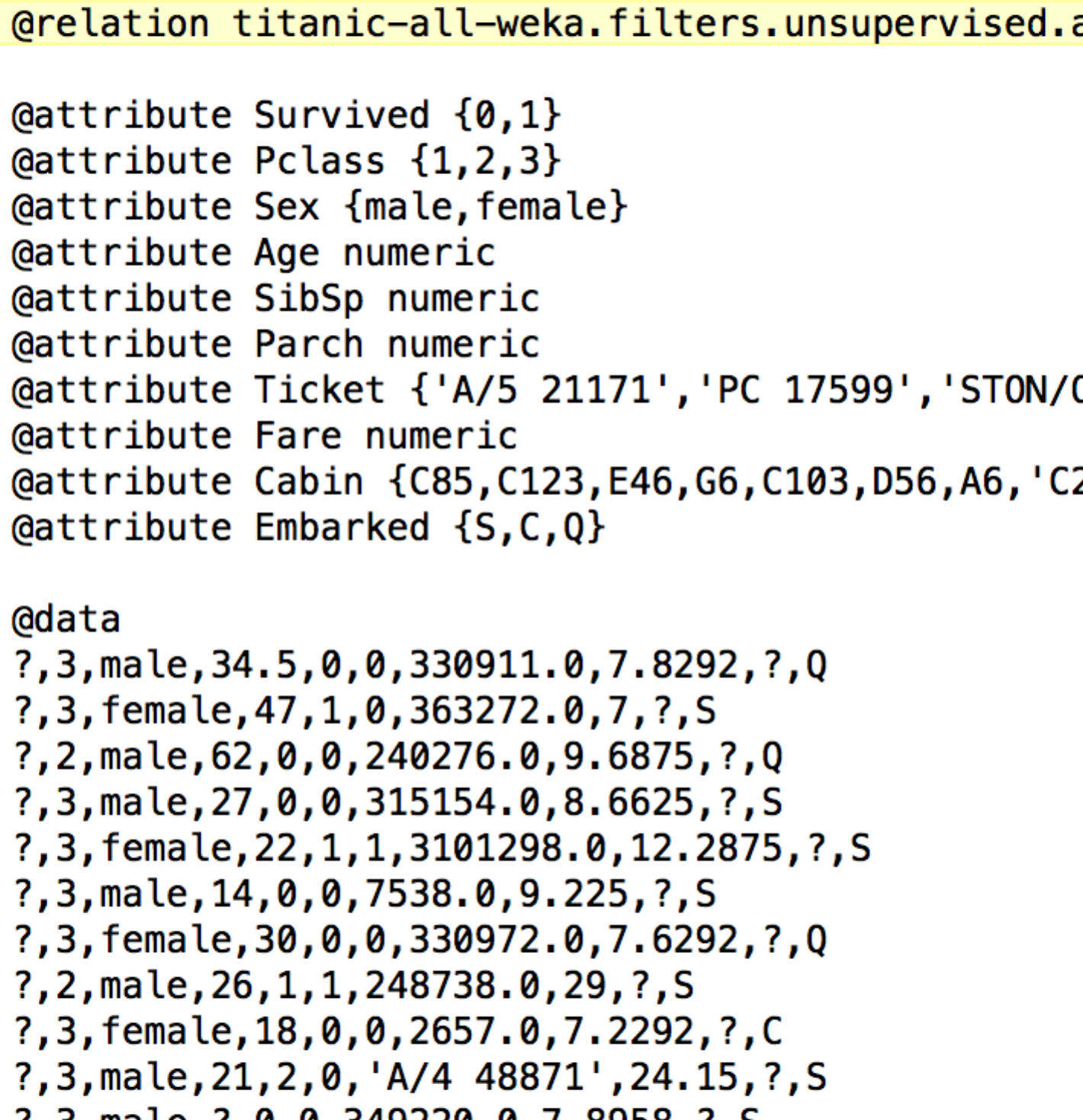


* 1. Click “Open File.” Select “CSV” from “files of type.” Select TitanicAll.csv.



Now you have successfully read your merged training and test data file into Weka. Save this file in the .arff format, and then separate the train and test examples into two files. The two files should have the same heading and data description. Therefore, you need to make a copy of the merged file and name it "Titanic-train.arff", and then delete all test example lines in it. Now you have your training arff file ready. Then make another copy of the merged file and name it "Titanic-test.arff", and then delete all training example lines in it. Now you have your test arff file ready.

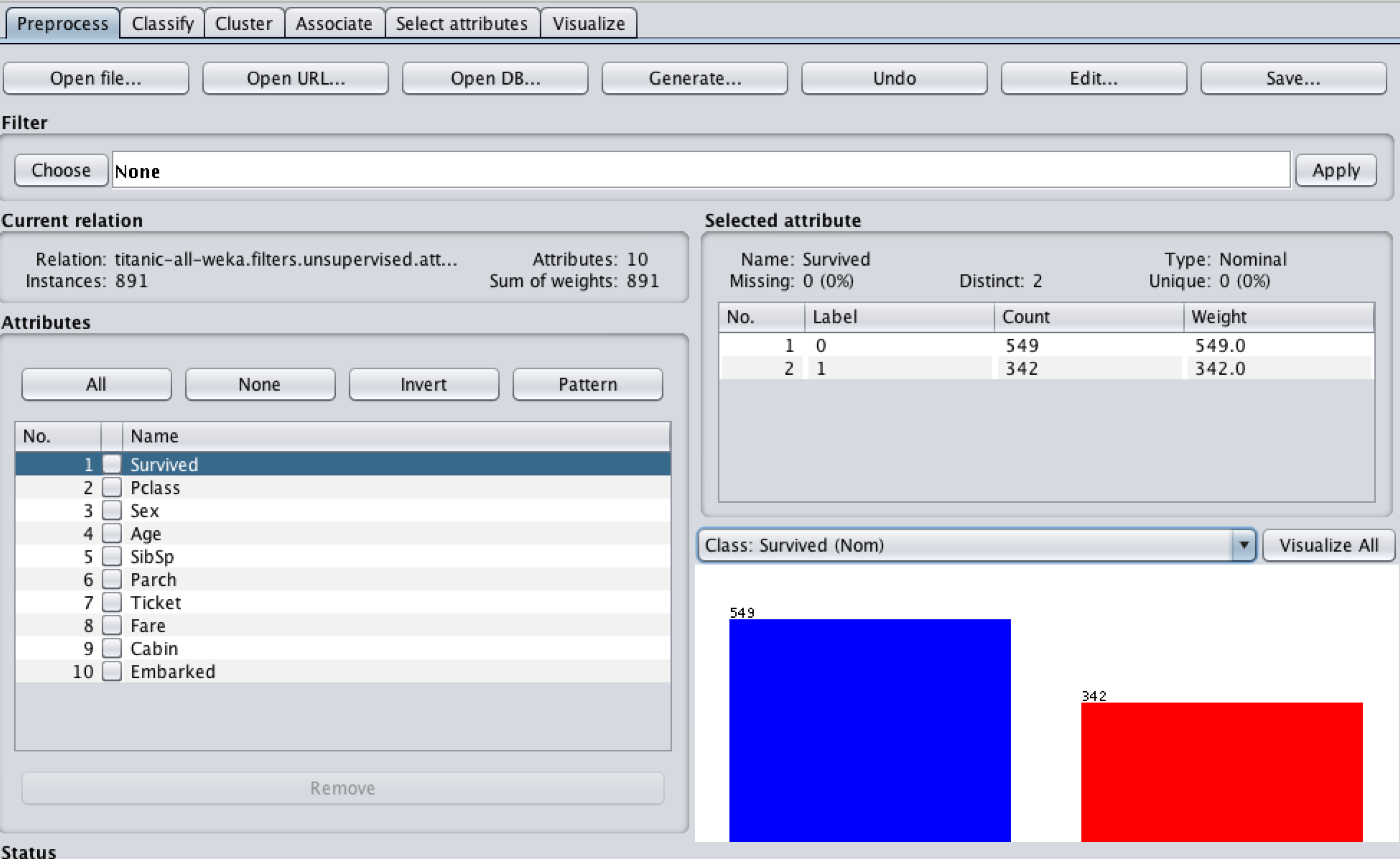




Now the Titanic training data is ready to be loaded into Weka to build decision tree. First, use the "open file" in "Preprocess" tab to load in the training data.

The upper right part of the interface shows some basic statistics of the feature highlighted on the left side. In the following screenshot, it is "Survived".

On the lower right part of the interface, choose "Class Survived (Nom)" in the pull down menu of all features, and you will see the category distribution, i.e. how many survived (Label "1" color red) and how many died (Label "0" color blue)



Now go to "Classify" tab, under "Classifiers", click the "Choose" button, and find J48 algroithm under the "trees" folder.

Maybe just add a screen flow