*10.12 Ideas for Fuzzy Joins*

Notes

* Gravitating more toward using it to fill out blank entries
* Essentially, you train it on the rows that already have filled out entries for the column and then run it on the rows
* For each column, the goal is to figure out which other columns influence it. For example, likelihood to vote might be influenced by age and location, but location might be influenced job, etc. The model would figure that out
* There’s essentially going to be a different model for each column
* Mark off the rows with blanks
* Essentially, you take any column (will you vote) which may have 400 filled out and 600 empty. Train the “vote model” on the values that are there, and then use that model to fill out the empties
* It would be easy to do the model accuracy because we can just delete a few actual examples and get the accuracy. We would want more a distance metric (like MSE) if the categories we’re predicting are ordered
* Can experiment with filling in means/medians/modes for certain empties
* Essentially, if you had 4 columns to predict, start with the one that has the fewest empties, then go from there
* Could also raise a warning before filling out a column if the accuracy will be below a certain threshold
* One other thing we could check out is seeing if there’s a certain number of items we can delete/keep before getting a fall off in accuracy. Maybe we only need 200 random ones left before the results are less accurate
* Email Iklaq\*\*