Good afternoon everyone. Today I wanted to present an update on the progress that we have made with our algorithm to predict the stage of a lawsuit using the text from the docket sheet entries.

Problem:

* Automate the prediction of the stage of a lawsuit by solely using the text from docket sheet entries
* Characteristics
  + 5,000 cases, estimate 18,000 – 20,000 entries
  + 3 or four lines of text per entry
  + 11 Stages

Approach:

* Pre-classify about 100 cases into the 11 stages
* Eliminate immaterial docket sheet entries.
* Approach: Attempt to predict the stage of a lawsuit by identifying keywords that appear frequently in a given stage across the 100 cases and are the most unique to that stage relative to the other sages.

Recap – Original Algorithm

* 4 Stages or Steps to our prediction
* Stage 1: Create a frequency distribution of key words by life cycle stage
  + Mechanics:

