Reflection - Reading Tea Leaves - Chang et. al
Nicholas Hoernle **May 2, 2018**

The authors propose that standard perplexity and predictive log likelihood do not evaluate the semantic interpretability of probabilistic clustering techniques. I agree with this notion as these techniques are reported to qualitatively mine meaningful topics from corpora but there exists no quantitative method for evaluating how semantically relevant the topics are. The authors propose a method that relies on a human evaluation for the gold standard. This technique has not been widely repeated since the publication in NIPS 2009 (at least as far as my understanding of the literature extends). One major reason for this is the 'human in the loop requirement'. It is difficult and costly to design and implement effective experiments that involve people especially as the topics become more domain specific (e.g. imagine we were clustering gene sequencing - how could Turk workers provide the gold standard for a niche topic such as this).

From Figure 5, it appears that the LDA and pSLI techniques have a higher 'Modal Precision' and 'Topic Log Odds' than the CTM model. I am confused as to why the predictive log odds is different for the left and right charts. I can also conclude that there appears to be an inverse relationship between the number of topics defined in the topic model and the 'interpretability' of those topics. It concerns me that we lack error bars on this plot.