A big problem in internet advertising, especially from my experience in the video segment, is ad fraud. A fraudulent impression is when an advertisement loads but is not viewable by a live person. This can happen for a few reasons including: loading an ad below the fold, loading an ad on a non-viewable window/tab, or a combination of both while displayed on a computer controlled by a bot network.

After the AOL acquisition of major ad exchange Adap.tv, analysts made estimates that higher than 30% of the Adap.tv’s inventory was suspect. This represents millions of dollars wasted by advertisers on bogus inventory.

I would like to work on an unsupervised categorical machine learning algorithm to classify inventory on an ad exchange into high quality, suspect, and fraudulent traffic, and ultimately create a recommendation system of which problem publishers to drop.

The hardest part of this will be finding a good data set, as most of this data is kept under heavy lock and key.