PROJECT PROPOSAL – from Mario Salgado, Per Jenson and Gilles Comeau

Name: “Predicting side effects and withdrawals of pharmaceuticals from Online Media”

## Background and Motivation

Every year, commonly available medication is withdrawn from the market due to safety concerns.  
From the market withdrawal of Thalidomide (a sedative causing foetal malformation) in the early 60s, to the withdrawal of Vioxx (a painkiller causing heart-attacks and strokes, marketed for 5 years) in the mid 00s, the need for withdrawal of marketed medication constitutes an important health hazard.

In order to gain marketing authorization for clinically approved medical products, pharmaceutical companies should demonstrate the capability to gather post-marketing reports of side effects through local country offices and telephone hotlines. With the widespread use of social media, the tendency to report side effects to the general public has risen tremendously.

## Project Objectives

Determine whether Online Media should be added as a mandatory source of reported side effects by examining the questions below: (in increasing complexity)

1. To what extent do online media users report side effects?
2. What media types especially appeal to reporters of side effects?
3. Can online media help us identify unlabelled side effects (those not found on the package leaflet)?
4. Can we use sentiment analysis as a feature to predict how risky / how bad a side effect is?
5. Can we predict withdrawals from media posts?

## Data Sources

Our analysis will include the following data sources

1. Pharmaceuticals
   * 20 pharmaceuticals from the withdrawals list at the US FDA Drug Authority  
     <http://www.fda.gov/Drugs/DrugSafety/DrugRecalls/default.htm>  
     In particular, we’ll be looking at drug name, drug name synonyms to create our Boolean queries and withdrawal date for our timeframes.
   * 20 drugs as a control set which are similar to those drugs, but have not been withdrawn.
2. Known lists of side effects for our “truth” set. <http://www.drugs.com/sfx/>  
   SPEC: Drug name and side effect will be used to build the truth set.
3. Media data – We’ll be searching an archive of millions of postings per day for the Boolean queries outlined in #2. In the archive, we have a collection of 500,000 forums, blogs, local and national news and consumer reviews and social media.

## Must Have Features

Our group will definitely be able to determine whether and how often people are reporting side effects to various drugs from our dataset. We will be able to determine whether we find this information more frequently in Blogs (for example, health forums) / Forums / Social Media.

After this, as we are doing data exploration with linguistically complicated problem set, we will determine whether sentiment analysis can be used as a feature – but there is no guarantee that sentiment analysis is accurate enough for this purpose.

Even more complicated is to determine whether withdrawals can be predicted. We would like to answer whether we think this is possible, but cannot guarantee that it will be possible.

## Optional Features

As mentioned above, it is optional as to whether sentiment analysis can be used as a feature for how risky a posting about a pharmaceutical is – and also whether withdrawals can be predicted.

## Design Overview

Platform – it’s difficult to determine what platform is ideal before we determine the size of the dataset. For NLP work, my preference is SOLR as an information retrieval platform. (like elastic search), but then we should extract the data and work on it in Python for analysis capabilities.

Statistical and Computational Methods –

* Boolean queries / taxonomies for the initial extraction of data
* NLP techniques – tokenization / BM25 / TFIDF for determining the degree that the article is about the drug.
* Sentiment Analysis / Classification for predicting Risk
* Classification – with truth of “Withdrawn” and “not withdrawn”, we can use features to predict whether a drug would be withdrawn.

## Verification

We expect to use our truth set of side effects to determine whether online media can uncover existing side effects. After that, we will use a known list of side effects and determine how many additional side effects are present – success is if we uncover side effects which are not known / listed.

As far as predicting pharmaceutical withdrawals, we’ll be looking for mentions of the reason the pharmaceutical was withdrawn in media – if we find these mentions, than this is another proof point for scanning online media for risky side effects.

## Visualization and Presentation

<MARIO TO COMPLETE>

## Schedule Overview

In general, our project should follow the following timeline:

**Start time**: November 13th, 2015

* Week 1 (ending Nov 20th) –
  + Data – Determine dates / drugs / side effects. Build Boolean query to extract media. Extract media. Build Dataframe with relevant information for algorithms team.
  + Algorithms – Begin exploring algorithm used for analysis.
  + Visualization - <MARIO>
* Week 2 (ending Nov 27th)
  + Data – Extract additional information / sources as required.
  + Algorithms – Heavy week of algorithms to determine the answers to the questions.
  + Visualization - <MARIO>
* Week 3 (ending Dec 4th)
  + Data – Extract additional information / sources as required.
  + Algorithms – Complete algorithms analysis of the data.
  + Visualization - <MARIO>
* Week 4 (ending Dec 10th – Submission date)
  + Prepare presentation
  + Prepare video
  + Complete write-up

**End time** December 10th, 2015

## Team Member Contributions

Gilles has access and is skilled at Boolean queries, data extraction for our online media as well as NLP.

Per is an algorithms ninja and will be leading the analysis algorithms and analysis with Gilles / Mario supporting.

Mario is a design wizard who will be building compelling visualizations for demonstrating our results to our audience.