Project Summary

Overview

Human Immunodeficiency Virus (HIV) is a disease that affects millions of individuals in the United States. While HIV has been studied for decades and preventative strategies have been in use for decades, its continued persistence has demonstrated the need for new methods both for both monitoring and prevention. Biosurveillance is a new interdisciplinary collaboration between public health and data science that has been developed in the last few decades. It leverages a variety of data mining algorithms performed on large data repositories to identify and monitor disease outcomes and treatments. In this document the PIs describe an approach to identify trends in disease treatment by applying existing and novel machine learning techniques in the field of biosurveillance. The PI's contribution is both in the development of data analysis pipelines for general purpose biosurveilence, and the application of those pipelines for the study of HIV and mental disease monitoring.

Data sources will include text, images, and metadata associated with social media, and will eventually link these data with direct medical data. High level information, analytical tools, and processed data produced through these analyses will be disseminated back into both clinical and social media venues to improve patient's medical outcomes and address any public misconceptions. While we will collaborate with public health and psychology contributors, our research primarily focuses on the data science side aspects of the broader biosurveillance field.

Our efforts in biosurveillance will be addressed in three sections:

Section1: Mining Pre-exposure Prophylaxis trends in social media

One specific application will focus on the use of antiretroviral therapy to preventatively treat individuals who are at-risk for HIV infection. This use of antiretroviral therapy for preventative treatment is referred to as Pre-Exposure Prophylaxis (PrEP) and uses the small molecule nucleotide reverse transcriptase inhibitor trade named Truvada. PrEP is highly effective at preventing HIV, but because it is a new treatment there remain some social and medical obstacles preventing it from achieving its full potential. In this section we use various supervised and unsupervised text mining techniques to identify relevant keywords, hashtages, positive tweets and negative tweets in Twitter data that can be used to gain insight into the social concerns surrounding PrEP adoption. By addressing these concerns, we can make PrEP more effective and improve HIV prevention efforts.

Section2: Semi supervised learning and operations on arbitrary dimension tensors

In order to facilitate our biosurveillance work, we will develop and implement useful computational methods that can run on heterogeneous hardware. Graphical Processing Units (GPUs) have a theoretical computing output at least an order of magnitude more than a similarly priced CPU. By adopting GPU computing hardware we may be able to scale up existing data mining analyses, and still have them run in a reasonable amount of time. One such example, which would be relevant

to text mining and biosurveillance in general, would be a higher order version of the word2vec algorithm implemented for the GPU. In this section we propose this and other similar analyses.

Section3: Using machine learning to determine user personality types associated with disease and treatment responses

Previous research implies that personality type plays an important role in social media interaction dynamics. Specifically, narcissistic personalities tend to be over represented in the highly connected nodes of a social network. In order to understand the social issues surrounding disease treatment, it would be helpful to determine if certain personality types are more likely to react positively or negatively to certain disease treatments. We will use a variety of supervised methods (such as regression) to determine if there are associations between personality type and disease treatment sentiment. In addition we want to take advantage of additional forms of metadata, to see if they correlate with either disease treatment or personality, such as profile picture, network connectivity, geolocation, and embedded URLs in comments. In applicable cases we want to be able to distinguish between valid information and "hoaxes" to protect our analyses from biases due to deliberate misinformation.

The information produced through data mining and machine learning will be useful to provide clinicians and public health professionals feedback concerning successes and challenges associated with disease treatments. Pertinent tools and summary datasets will also be produced and made available to researchers. In addition novel applications of machine learning and data science methods will be developed as necessary, and shared with the data science community to assist and inspire related applications in other domain areas.

Project Description

1 Mining Pre-Exposure Prophylaxis Trends in Social Media

Pre-Exposure Prophylaxis (PrEP) is a recently developed method for the prevention of Human Immunodeficiency Virus (HIV) via the administration of an oral pharmaceutical trade named Truvada. Truvada contains active ingredients tenofovir and emtricitabine, both Nucleotide Reverse Transcriptase Inhibitors (NRTIs). In the last four years, since Truvada was approved for PrEP in 2012, PrEP has shown demonstrated efficacy at preventing HIV for HIV negative individuals in serodiscordant relationships[18]. Though existing methods of effective HIV prevention exist, data shows that they may not be used in case of unexpected sexual contact or personal preference[23]. PrEP is also well suited to individuals in socially or economically underprivileged groups, and for the HIV negative partner in serodiscordant couples[25]. In experimental studies, PrEP has been studied as a method to safely conceive a child[14].

Initial studies of PrEP have shown that it is highly effective[11], however because it is still a new treatment, it is facing a number of medical and social obstacles before it reaches full adoption. Incomplete clinical and patient knowledge, social stigma, and uncertain insurance status have been identified as challenges preventing continued adoption[5]. Also, since Truvada is an oral NRTI, it must be taken daily. Cases where patients do not adhere to their full prescription have led to loss of viral protection, and some patients and clinicians worry that lack of adherence could lead to increased risk of infection with drug resistant strains[2].

The goal presented in this proposal is to improve PrEP adoption and efficacy by reviewing PrEP patients and other HIV community members perceptions of the challenges and successes that have taken place during PrEP's initial adoption. While local clinical monitoring has uncovered some challenges facing PrEP efficacy and adoption[24], large scale data mining has the potential to capture a broader perceptive than local clinical reports can. In addition, by mining the massive data sentiments present on social media platforms, researchers are able to capture unfiltered opinions, and can update disease monitoring analyses in real time. Previous work has used Twitter to predict county-level HIV prevalence[28], and general HIV discussion monitoring[27]. However to our knowledge, prior to the previous work described in the next subsection, data mining social media specifically to determine thoughts and sentiments surrounding PrEP has not been performed.

1.1 Previous Work

Our goal in this previous research was to determine perceptions and sentiments of PrEP using social media, in order to address challenges related to PrEP adoption and efficacy. Though we have investigated the use of multiple social media platforms including Facebook, Grindr, Reddit and Twitter, we have focused primarily on the use of Twitter for past projects and intend to use Twitter for future analyses. Twitter is especially useful because in addition to textual data, various useful metadata is also available including datetime, geolocation, username, hastags, and external hyperlinks. Twitter also features a convenient well documented and free API and has over 300 million monthly active users.

In this analysis we collected over 1 million tweets from the Twitter streaming API filtered on PrEP and HIV related keywords. By using embedding techniques such as Word2Vec and Doc2Vec, we were able to identify new keywords and hashtags that are relevant to the PrEP conversation on

Table 1: Cosine similarity to document-vector "#PrEP"

Related hashtag/tweet	Cosine similarity to #PrEP
#lgbtmedia16	0.739128
#hiv	0.727602
#whereisprep	0.707165
#truvada	0.696113
#hivprevention	0.636068
tweet-702179860983189504	0.630055
user-711275699529764864	0.629254
tweet-708519265540907010	0.628778
tweet-712032637024653313	0.628646
#harrogatehour	0.628547

Twitter including unexpected political connections "NancyReagan" and popular hashtags associated with PrEP that might not have been known to researchers in advance such as "#whereisprep" (see table 1 for examples of hashtags and structurally-related tweets). In addition Doc2Vec can be used to query the top N tweets related to a given hashtag or the top N users related to a given hashtag. This allowed us to identify a small subset of the N tweets most structurally related to "PrEP" that can be human-read without requiring humans to read the full dataset. Reading from this set of structurally important tweets, we uncovered important blog articles exploring some of the fears of drug resistant forms of HIV and concerns as to whether these strains could be caused in part by over use or misuse of PrEP medications. The linked blog articles that we found highlight the usefulness of the hyperlink metadata embedded in tweets. Through these hyperlinks, Twitter acts as an index, providing indirect access to a much larger external social media ecosystem.

Our Doc2Vec results also allowed us to identify the top N users most relevant to PrEP. By querying the Twitter REST API for these users' timelines, and using topic modeling methods like Latent Dirichlet Analysis (LDA), we identified the word-distribution-topics present in the Twitter conversation (see figure 1). This analysis went beyond the simple keyword identification analysis from Word2Vec since it clustered keywords into topics and it operated on all tweets that PrEP-related users tweeted, not just PrEP related tweets. The LDA results showed a variety of related concerns such as other sexually transmitted diseases, LGBT related topics, health insurance and political topics. Neither PrEP or Truvada were present in the top 30 keywords related to HIV/AIDS demonstrating that PrEP is still a nascent rare topic in the online discussion. An extension to LDA, Dynamic Topic Modeling (DTM), was able to capture topic and word frequency over time. The DTM results showed that the keyword "PrEP" is increasing in relative frequency over time, even relative to related words such as "pill", "prevention" and "drug". This demonstrates increased interest in PrEP which may correlate with an increase in PrEP adoption over the data acquisition period.

One of the key terms identified through LDA topic modeling was "stigma", a concern that is a known issue impeding PrEP adoption in previous studies[18]. Stigma is an issue with sexually transmitted diseases in general, including HIV, resulting in many HIV positive indivisuals not knowing their status, and thus may not be aware that they are transmitting the disease when they come in sexual contact with HIV negative individuals. Previous studies have also suggested that

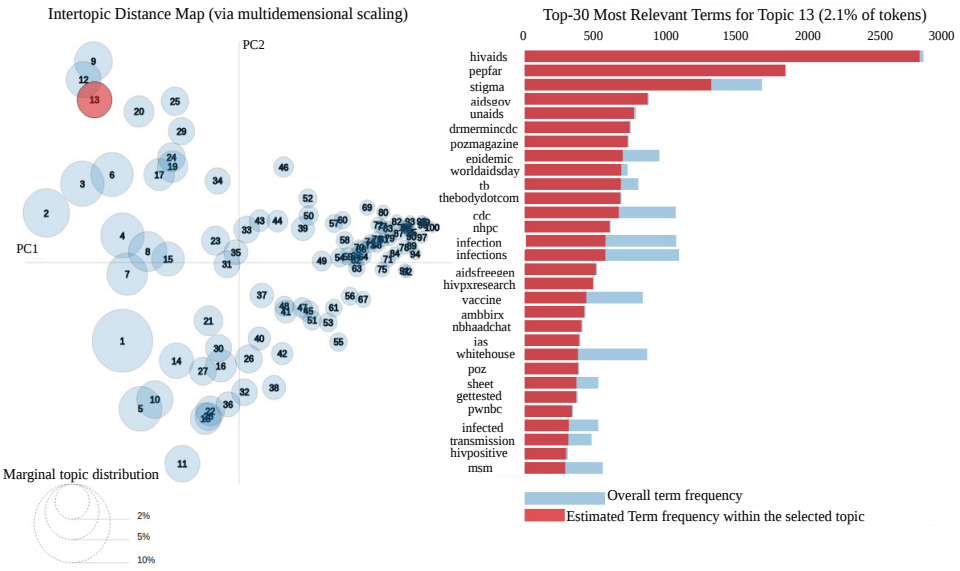

Figure 1: LDA topic modeling for the top 500 users related to PrEP.

stigma may contribute to issues surrounding PrEP adherence[5].

Using an open dataset of tweets which were labeled with binary sentiment labels, we performed a semi-supervised classification sentiment analysis. In the sentiment analysis, 5513 tweets were provided by Sanders Analytics. These tweets were human labeled either positive or negative. The sentiment analysis was performed by converting the tweets to numerical vectors using the unsupervised Doc2Vec embedding method. These numerical vectors were then classified according to the sentiment labels, and the trained model was used to infer sentiment labels for each of our PrEP-related tweets.

The sentiment analysis allowed us to identify N PrEP related tweets with the highest sentiment, and N PrEP tweets with the lowest sentiment. After performing the automated sentiment analysis, a human was inserted into the analysis loop to quickly read the top positive and negative tweets to get a sense of the most important successes and concerns present in public PrEP perception. In the positive tweets we found hyperlinks to blogs with positive firsthand accounts from individuals successfully using PrEP to stay HIV negative. In the negative tweets we found concerns of whether Truvada can protect against drug resistant strains of HIV (example negative tweets shown in table 2). Other concerns suggested that over prescription of Truvada could give individuals a false sense of security, and lead to a rise of non-HIV sexually transmitted diseases. These results show that patients, and the public at large need to be educated on the specific risks of drug resistant strains.

The issue of individuals gaining a false sense of security towards other sexually transmitted diseases is being addressed though clinical education for individuals that are prescribed Truvada, though these results show that this information is lacking in some individuals, presumably not taking Truvada. This implies that the dissemination of Truvada educational information through a platform like Twitter may improve overall PrEP-related knowledge and address some of the concerns and stigmas currently associated with PrEP. Together these approaches and results show that we can take raw text and metadata and extract keywords, hastags, temporal trends, and sentiment information. Doc2Vec and sentiment classification allow the researcher to extract a set

Table 2: Negative sentiment tweets.

Category	Text
General	"Also, how f***ing vile of Hillary to say. Reagan did f***ing NOTHING during the AIDS epidemic until it was too late. What a stupid old hag."
General	"I wonder why he beat her a** when she was tryna leave like she wasn't gone be running back when she found out she had HIV & nobody want her"
General	"Aaannd. Hillary Clinton breathes a sigh of relief that Twitter has left its outrage of her AIDS comments behind to tend to Drumpf debacle."
PrEP specific	"RT gaston_croupier #Truvada patent's not expired yet but it is sold online as a generic drug? There's something rotten in internet #PrEP h"
PrEP specific	"Equality_MI Syph & Hep C have gone up 550% in Gay Men bc many feel tht bc they're on PrEP, they don't need condoms. HIV isn't the only STI."
PrEP specific	"Xaviom8 in interviews he says he was adherent. strain was highly resistant, and Truvada wouldn't have blocked it anyways. PrEP didn't fail."
Truvada specific	"not surprised at all that someone got HIV on truvada. people get pregnant on birth control. tomato-condoms are still important-tomahto"
Truvada specific	"Now reading that truvada does not protect against certain strains of the HIV virus. Yet people want to take that risk.."
Truvada specific	"I think I have conjunctivitis unless truvada cured it overnight cuz im not feeling as horrible today as last night"

of the N most highly relevant tweets from a large corpus that can be easily human-readable.

Though our previous research on data mining social media has uncovered important tweets, keywords, sentiments and hashtags related to the Twitter discussion of PrEP, many important questions remain poorly understood. One important issue is determining why patients stop adhering to their PrEP medication. While our LDA results uncovered "stigma" and other related keywords, and some of the critical tweets we identified described uncertainty in the efficacy of PrEP, this question still remains to be fully answered. One of the challenges with using Twitter as a data source is that we cant verify personal information for the people authoring the tweets. For example we don't know if they are taking Truvada, or whether they are HIV negative or positive, or other important details of their medical status. We also don't know if misinformation or excessive negativity is being spread by uninformed individuals, or by nefarious individuals. By incorporating medical data we could potentially identify direct connections between sentiment and written opinions with more concrete medical outcomes, though such connections would fall under The Health Insurance Portability and Accountability Act of 1996 (HIPAA) and require patient consent and/or medical board approval. As of the time this previous work was carried out we were unable to access such data.

1.2 Proposed work

There are many ways that we can continue to analyze social media data to identify specific issues impeding HIV prevention efforts. One extension we propose is to investigate additional diseases or behaviors that could correlate with risk of developing HIV. One of the overall goals of biosurveilance is to predict disease outbreaks before they happen in order to intervene preventatively. An outbreak in Scott County Indiana in early 2016 was thought to have been caused by drug usage. In a town of 4,000 people 135 people were diagnosed with HIV, and about 80% of those diagnosed were codiagnosed with hepatitis C. In theory some online social activity may have been able to indirectly identify this outbreak before it happened[10]. In order to identify and predict these hot-spots we can make use of the geolocation metadata, and also do network analyses to identify subgroups, and how information and via inference, social interactions propagate through these subgroups. Some first steps along these lines have already been made by our collaborators [28], who used drug-related terms to predict HIV prevalence at the county level.

In order to extend this work, we could use the presence of other disease keywords known to coinfect with HIV as a basis for predicting HIV prevalence. This proposal is facilitated since we already have a Twitter corpus that contains tweets mentioning HIV, hepatitis B and C and about 30 other infectious diseases. Searching the HIV literature shows that the use of Twitter mentions of Hepatitis C (codiagnosed at a rate of 80% in Scott County) and/or other STDs to predict HIV prevalence are lacking. Investigation of the Hepatitis-HIV connection is also particularly important because it allows us to test the concern that we identified in our previous work, that overuse, or misuse of PrEP could lead to increased prevalence of Hepatitis C. One way to test and measure this criticism of PrEP, would be to try to predict Hepatitis C prevalence using mentions of PrEP on Twitter at a county level. If this prediction succeeds, then we have evidence that confirms and quantifies this criticism, and can seek to address how PrEP is administered to prevent coinfection with Hepatitis, however if this prediction fails, we have shown that PrEP is indeed not leading to increased levels of Hepatitis, and thus we will have strengthened the case for PrEP adoption.

The second specific extension to our previous work that we make, is the use of genomic methods to help determine the network spread of HIV. Genomic evidence such as pylogenetic trees/networks can be used to determine a specific path of virus spreading through a community. A recent study of the San Diego regional heath system reviewed sequencing data collected from patients diagnosed with HIV over a period from 1996 to 2011[17]. Without any epidemiological patient-patient contact information, the researchers inferred a patient-patient network infection model. The authors do not provide details, but imply that a multiple sequence alignment, followed by some measure of sequence-distance was used to infer the phylogenic network.

If we had access to a similar regional genomic dataset, we could infer a similar phylogenic network on anonymized patient data in a different region of interest such as the north Georgia area. Coupling such an analysis with an investigation of social media at the same time and geographical location would produce a hybrid approach to identify the connections between HIV transmission and HIV social media discussion. To our knowledge at the writing of this proposal, no single study has incorporated both an HIV transmission phylogeny and an investigation of HIV on social media. If successful this combined approach could show more directly how a regional outbreak is captured in social media. This proposal can fail most clearly due to lack of access to genomic sequence data, though such genomic sequence data does not need to be associated with personally identifying information. This removes the need for HIPPA procedures, or medical review board

approval making this proposal relatively feasible. Also, the author has used phylogenic methods in the past, further facilitating the feasibility of this approach.

Finally, we would like to mention that our quantitative approaches and computational pipelines for mining qualitative sentiments surrounding disease treatment provide an important contribution by themselves to the larger data science community. In our previous work we have shown a specific application where we mine social sentiment to identify what is working and what the challenges are for PrEP, but a similar framework could easily be taken and applied to improve the social barriers surrounding some other disease treatment like cancer and chemotherapy. Pharmaceutical companies, academic researchers, and hospitals can use our open source code with minimal modification to monitor their disease and treatment of interest to monitor and improve the outcomes and happiness of their patients. We anticipate that the computational approaches produced during the course of the proposed work will also demonstrate useful methods that can be applied to other areas of public health research.

2 Semi supervised learning and operations on arbitrary dimension tensors (working with Arvind Ramanathan, Oak Ridge National Laboratory)

In January 2017 I will start working at Oak Ridge National Laboratory on a data science related fellowship. The goals of the grant supporting me seeks to use novel CPU/GPU hybrid chips to develop novel computational models in the area of semi-supervised deep learning. While I don't understand all of the details of the research proposal, or the specifics of the hardware at the time of writing (early November 2015), I will attempt to give a broad explanation of the research area and some possible directions.

Semi-supervised learning refers to a situation where both labeled and unlabeled data are used to train a discriminative model[29]. In contrast supervised learning uses only labeled data to train a discriminative model and unsupervised learning uses only unlabeled data to train model that captures some structure of the data distribution ie clustering or generative. Supervised learning is often on of the most used and useful types of machine learning since the model is directly predicting a label that has external research value. However, producing that labeled data can often require human annotation or physical experimentation with can be costly. Because labeled data is costly, machine learning models in many domains are error-limited by a lack labeled data to train on[12], which is perhaps ironic in a world of so called "big data". By taking advantage of both an unsupervised model of the data-distribution using cheap unlabeled data, and a discriminative model of the labels given the data on some small labeled dataset, semi supervised learning can produce results better than a n approach that uses only supervised learning.

Let me give an example of a semi supervised learning method in action, and for added consistency, I'll use an example from my own work described in the previous section (Mining Pre-Exposure Prophylaxis Trends in Social Media). When we did the sentiment analysis in the previous section we actually did 2 discrete steps. First we used Doc2Vec to transform each tweet into a dense high dimensional doc-vector, then we used a very simple logistic regression model to classify the tweet doc vectors into positive or negative sentiments. As a human reading the tweets in the results, you might be surprised at the relative accuracy of the classification, especially considering that logistic regression is relatively speaking one of the simplest supervised classifiers, and the natural

language present in the tweets is very terse and complicated for such a simple supervised logistic regression model to work so well. Doc2Vec, an unsupervised embedding method produced dense document vectors that were able to capture the distribution of the tweet data, mapping tweets that had co-occurring word semantics to be "close" in a high dimensional space. Thus the combination of unsupervised embedding and supervised classification yielded better results than a simple supervised classification could have done on its own (we didn't try doing classification without using Doc2Vec in the PrEP paper, but see the original Doc2Vec paper for relative quantification of accuracy gained by embedding[15]). Embedding, which is just one example of semi-supervised learning has become very widely used as a single step in a larger deep learning model (see the last section in this proposal below where we use embedding in a bioinformatics deep learning context) also others[26].

Though we know that we will be using novel hardware at Oak Ridge National Laboratory we aren't currently sure of the hardware details. We know that we will have access to new CPU / GPU hybrid chips that are being produced by Nvidia. GPUs have become increasingly popular in machine learning in the last couple of years, especially in the area of deep learning. This is largely because general purpose programming APIs such as CUDA have allowed researchers to take advantage of cost efficient computing power present in GPU's for computation-bound tasks (such as deep learning). On Nvidia's Blog an article titled "Accelerating AI with GPUs: A New Computing Model" describes how typical deep learning models can be trained 50x faster on a Tesla M40 GPU than on a typical CPU. Though Nvidia typically provides a compiler toolchain supporting a language called CUDA with basic C-like syntax, Nvidia has also provided a series of highly optimized routines for things like linear algebra, matrix multiplication, convolution and manipulation that allow researchers to think in terms of linear algebra primitives on matrices instead of low level array-wise manipulation[6]. This makes construction of machine learning models in low-level CUDA code very analogous to using popular higher level machine learning libraries in languages such as Python.

In order to provide support for semi-supervised learning on these novel CPU/GPU hybrid chips, we want to develop computational operations and building blocks that allow for arbitrary dimension tensors. Though most machine learning models use operations on matrices or rank 2 tensors, some theoretical and practical applications could take advantage of operations on tensors of arbitrary dimension. Think of the embedding example above. We embedded a 1 dimensional sequence of word-tokens (a document) into a dense 1 dimensional real-valued vector representation. In other situations we may want to produce arbitrary dimensional embeddings as a building block to construct semi supervised deep learning models. We may want to measure or encode the N-way similarity of words, requiring a rank N tensor. Several papers show examples at a theoretical level, where operations, projections and decompositions on high rank tensors can be used to extract feature information from the original data[7, 9]. These operations such as Principle Components Analysis, Multi-Dimensional Scaling, Locally Preserving Projection, Neighborhood Preserving projection etc. could be used to provide embeddings/decompositions to be used in a larger deep learning network either by preprocessing, or by composing the objective function of the embedding kernel with the other parts of the deep learning network in order to train the whole network through back propagation.

Many deep learning neural networks are analogous to matrix and tensor decompositions. For example a single layer autoencoder learns an equivalent encoding to PCA when a linear activation function is used[13]. We may be able to identify efficient tensor decompositions to use as our

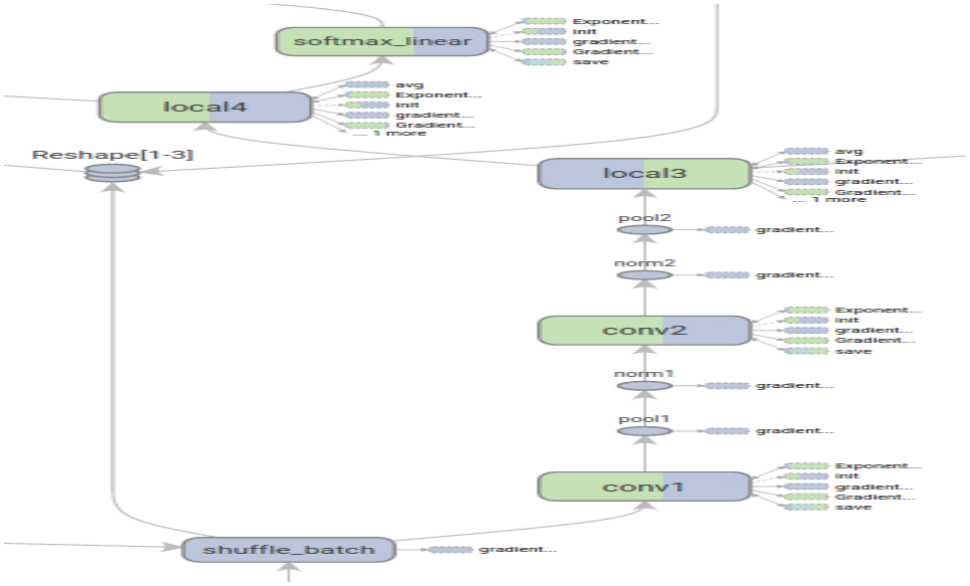

Figure 2: Example Tensorflow graph showing operation-nodes and tensor-edges. In the right panel green operations are tied to CPU and Blue operations are tied to the GPU.

encoding or feature learning step in our semi-supervised learning. These tensor decompositions could be analyzed more formally using theoretical analysis than simple deep neural networks have allowed.

Practically we could in theory build small programs that implement these tensor operations in a small CUDA or C-like program. However, recent development of deep learning frameworks gives us another practical implementation option. Frameworks such as Caffe, Theano, Torch and Tensorflow allow machine learning researchers to build and share efficient machine learning models that take advantage of a simple scripting-level API, and take advantage of highly parallel efficient machine implementations. Of these frameworks, the one that I am most familiar with is Tensorflow, an open source project supported by Google [1].

Tensorflow is currently under fast paced development, but already it offers support for data-flow parallelism, distributed and multi-device computing, automatic differentiation, and it offers C++ and Python based APIs for specifying models. Creating a deep learning model in Tensorflow is done by composing building blocks, or "layers", into a graph. In the graph operations are nodes and tensors are edges that feed out of one node and into another node. In a typical supervised learning situation data flows in a directed acyclic path from the input node(s) to the objective or classification node(s). Through automatic differentiation, the error or "loss" in the objective function is used to produce parameter updates through gradient decent. These updates are then propagated backwards through the graph in the direction opposite data-flow. This iterative process repeats many times as the models parameters are trained.

For high throughput and flexibility the compute graph's operations can be assigned to various "devices" such as CPU or GPU. These devices need not be co-located on the same machine, and during the evaluation of the graph tensors that are output from one operation that are needed on another machine or device will be transported over a bus or over a network connection. The same operation can be instantiated on different devices by instantiating a "kernel" for the given

device. The kernel is usually specified in C, C++ or CUDA-C for performance and for compatibility with the device. Once these operations and kernels are instantiated, they can be made available along with the other kernels present in Tensorflow to provide yet another building block for other researchers to use to build their own compute-graphs. Thus once we implement our high rank tensor operation(s), we can relatively easily use our operation to extend Tensorflow which makes our operations available for many other researchers in the community.

Ultimately in addition to building embedding and arbitrary tensor operations to extend computational capabilities for other machine learning researchers, we also want to produce some specific applications to help solve our surveillance problems. Since we don't know the exact details of the hardware at this point, it is unclear exactly what our capabilities will be. However I think that a basic re-implementation of existing methods like Word2Vec or other deep learning components, perhaps in a way that uses high-rank tensors would provide a familiar embedding-like method that could be used on social media data such as described in the first section of this proposal. Word2Vec is analogous to an implicit matrix factorization of the pairwise mutual information between word-context combinations[16]. Singular Value Decomposition (SVD) could alternatively be used to provide the factorization of this matrix. Alternatively, we could theoretically extend this to higher rank tensors by using a higher order version of SVD on an N-wise mutual information tensor.

By using our newly developed algorithm to do embedding or sentiment classification on textual twitter data, we can perhaps uncover new trends and mechanisms underlying the spread of disease. This would help facilitate the growing need for data analysis mechanisms in the biological and medical sciences[8]. Ultimately we will use these new methods to predict disease outbreaks, prevent disease, and identify social issues and concerns surrounding existing diseases and treatments with the goal of improving the social outcomes in addition to medical outcomes for patients.

3 Using machine learning to determine user personality types associated with disease and treatment responses

For the third section of this proposal, I would like to extend biosurveillance measures to other disease/treatment combinations that were not covered in the PrEP section, and also take into account social network connectivity (number of followers and followees), and other forms of media related to social profiles such as images. Connectivity is important in part due to its association with certain personality traits. Previous research in psychology has determined that certain narcissistic personality traits correlate with connectivity in social networks[4]. In this section we propose an exploration of whether personality traits correlate with disease/treatment discussions in social media.

Because personality traits have been shown previously to correlate with profile pictures and other posted pictures[19], for our investigation of personality traits we wanted to use a social media platform in which users would post images of themselves. While Instagram, a social media platform that specializes in images and short videos, seemed like a great choice for this analysis, unfortunately their public API feed was deactivated indefinitely in June 2016. Facebook also does not have a free API that can be used by data mining researchers, in part because its privacy policy limits many comments to only be viewable by friends. These public API limitations left us with Twitter as the only remaining platform with millions of monthly active users who have profile pictures available for datamining. Twitter is also uniquely useful because it offers a variety of meta data including

datetime, hashtag and geolocation data. Twitter also offers a public streaming API and various REST APIs that can be used to query all public tweets and unlike social media platforms like Facebook, tweets are public by default. Twitter users each have a profile picture available through the REST API, and images posted in tweets can be easily filtered out based on the image format (i.e. *.jpeg or *.png).

A recent psychology study has annotated over 1 million tweets by Myers-Briggs personality type and gender[20]. We propose the use of this labeled tweet corpus to perform supervised label inference on our public health dataset of interest in a similar way to how we did the sentiment analysis in our PrEP-related research (see section one of this document). In this way we can transpose personality labels onto our disease-related tweets. We can also see if there are any aspects of user profile pictures that correlate with either personality or disease topics, or finally whether follower or followee counts associate with disease or personality traits. In addition using network-based semi supervised learning methods, such as graph-laplacian[21], we can do unsupervised learning wherein we use unlabeled twitter users's graph connections to help us better classify labeled user's personalities.

For this project I am proposing the use of several tweet corpuses. Firstly we have the tweets with labeled personality score and gender provided as an open dataset[20]. Next we have a corpus of general tweets that mention an infectious disease, a corpus that mentions at least one of the top 100 prescription pharmaceuticals, and a corpus that mentions at least one of the top 100 hospitals in the US. In addition to these disease related corpuses, I also want to repeat any analyses on a random "background" corpus of general tweets from Twitter to make sure that trends and patterns identified in disease related corpuses are disease-specific and not trends that hold over all tweets.

I want to use a variety of simplistic models such as linear regression and decision trees to determine the features that contribute to whether a tweet is associated with a certain personality or mentions a certain disease negatively or positively. These simple models are useful because they can more easily be interpreted. Next I want to do a variety of hypothesis tests to determine if certain diseases or personality traits correlate significantly with each other. This correlation will have to take into account whether or not the disease or personality scores are continuous or discrete, though this will be decided on at the time. Significance tests will use the background tweet corpus described above as an empirical "null distribution".

Taking additional advantage of metadata available in tweets, we can determine if there are any correlations or relevant summary statistics that correlate between disease or personality variables. Given two personality types or two sets of positive or negative disease-related tweets, we can measure the most frequent domain name prefixes in the hyper links. We can also measure if there are non-random temporal or spacial (using geolocation metadata) distribution of these tweets relative to some background distribution (hence the need for a set of background tweets).

A final goal that we want to explore is the phenomenon of hoaxes or fake news. With the advent of the internet and various social media platforms, news can spread quickly from unvetted sources. This has produced an issue of misinformation, including misinformation that is deliberatly spread to gain clicks or spread discrediting information. Some research has shown that spread of hoaxes can be modeled using ecological infection models[22]. Recent research has used supervised machine to classify reviews as authentic or fake[3] by using simple models (logistic regression, decision tree, naive bayes, etc.). It may be possible to use more complicated models, such as deep learning models and network-based models, to identify fake news in twitter public health data. This will likely require acquisition of a dataset that has labeled tweets of positive or negative news.

If successful, this work would be significant on its own, and it would also help ensure our other biosurvelience research is resistant to fake news biases.

If successful, this research would help public health officials determine the interactions between social media network connectivity, infectious disease, and personality traits. This would continue to inform our understanding developed from research described in section 1 that attempts to determine the social aspects relevant to disease treatment outcomes.

References Cited

[1] Martın Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*, 2016.

[2] Emily A Arnold, Patrick Hazelton, Tim Lane, Katerina A Christopoulos, Gabriel R Galindo, Wayne T Steward, and Stephen F Morin. A qualitative study of provider thoughts on implementing pre-exposure prophylaxis (prep) in clinical settings to prevent hiv infection. *PloS one*, 7(7):e40603, 2012.

[3] Snehasish Banerjee, Alton YK Chua, and Jung-Jae Kim. Using supervised learning to classify authentic and fake online reviews. In *Proceedings of the 9th International Conference on Ubiquitous Information Management and Communication*, page 88. ACM, 2015.

[4] Laura E Buffardi and W Keith Campbell. Narcissism and social networking web sites. *Personality and social psychology bulletin*, 34(10):1303–1314, 2008.

[5] Sarah K Calabrese and Kristen Underhill. How stigma surrounding the use of hiv preexposure prophylaxis undermines prevention and pleasure: a call to destigmatize truvada whores. *American journal of public health*, 105(10):1960–1964, 2015.

[6] Sharan Chetlur, Cliff Woolley, Philippe Vandermersch, Jonathan Cohen, John Tran, Bryan Catanzaro, and Evan Shelhamer. cudnn: Efficient primitives for deep learning. *arXiv preprint arXiv:1410.0759*, 2014.

[7] A Cichocki, N Lee, IV Oseledets, AH Phan, Q Zhao, and D Mandic. Low-rank tensor networks for dimensionality reduction and large-scale optimization problems: Perspectives and challenges part 1. *arXiv preprint arXiv:1609.00893*, 2016.

[8] Andrzej Cichocki. Era of big data processing: A new approach via tensor networks and tensor decompositions. *arXiv preprint arXiv:1403.2048*, 2014.

[9] Andrzej Cichocki. Tensor networks for big data analytics and large-scale optimization problems. *arXiv preprint arXiv:1407.3124*, 2014.

[10] Caitlin Conrad, Heather M Bradley, Dita Broz, Swamy Buddha, Erika L Chapman, Romeo R Galang, Daniel Hillman, John Hon, Karen W Hoover, Monita R Patel, et al. Community outbreak of hiv infection linked to injection drug use of oxymorphoneindiana, 2015. *MMWR Morb Mortal Wkly Rep*, 64(16):443–444, 2015.

[11] Sarit A Golub, Kristi E Gamarel, H Jonathon Rendina, Anthony Surace, and Corina L Lelutiu-Weinberger. From efficacy to effectiveness: facilitators and barriers to prep acceptability and motivations for adherence among msm and transgender women in new york city. *AIDS patient care and STDs*, 27(4):248–254, 2013.

[12] Matthieu Guillaumin, Jakob Verbeek, and Cordelia Schmid. Multimodal semi-supervised learning for image classification. In *CVPR 2010-23rd IEEE Conference on Computer Vision & Pattern Recognition*, pages 902–909. IEEE Computer Society, 2010.

[13] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

[14] Margaret A Lampe, Dawn K Smith, Gillian JE Anderson, Ashley E Edwards, and Steven R Nesheim. Achieving safe conception in hiv-discordant couples: the potential role of oral preexposure prophylaxis (prep) in the united states. *American Journal of Obstetrics and Gynecology*, 204(6):488–e1, 2011.

[15] Quoc V Le and Tomas Mikolov. Distributed representations of sentences and documents. In *ICML*, volume 14, pages 1188–1196, 2014.

[16] Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. In *Advances in neural information processing systems*, pages 2177–2185, 2014.

[17] Susan J Little, Sergei L Kosakovsky Pond, Christy M Anderson, Jason A Young, Joel O Wertheim, Sanjay R Mehta, Susanne May, and Davey M Smith. Using hiv networks to inform real time prevention interventions. *PloS one*, 9(6):e98443, 2014.

[18] Albert Liu, Stephanie Cohen, Stephen Follansbee, Deborah Cohan, Shannon Weber, Darpun Sachdev, and Susan Buchbinder. Early experiences implementing pre-exposure prophylaxis (prep) for hiv prevention in san francisco. *PLoS Med*, 11(3):e1001613, 2014.

[19] Eileen YL Ong, Rebecca P Ang, Jim CM Ho, Joylynn CY Lim, Dion H Goh, Chei Sian Lee, and Alton YK Chua. Narcissism, extraversion and adolescents self-presentation on facebook. *Personality and individual differences*, 50(2):180–185, 2011.

[20] Barbara Plank and Dirk Hovy. Personality traits on twitterorhow to get 1,500 personality tests in a week. In *Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 92–98, 2015.

[21] Vikas Sindhwani, Partha Niyogi, and Mikhail Belkin. Beyond the point cloud: from transductive to semi-supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pages 824–831. ACM, 2005.

[22] Marcella Tambuscio, Giancarlo Ruffo, Alessandro Flammini, and Filippo Menczer. Fact-checking effect on viral hoaxes: A model of misinformation spread in social networks. In *Proceedings of the 24th International Conference on World Wide Web*, pages 977–982. ACM, 2015.

[23] S Wade Taylor, Christina Psaros, David W Pantalone, Jake Tinsley, Steven A Elsesser, Kenneth H Mayer, and Steven A Safren. life-steps for prep adherence: Demonstration of a cbt-based intervention to increase adherence to preexposure prophylaxis (prep) medication among sexual-minority men at high risk for hiv acquisition. *Cognitive and Behavioral Practice*, 2016.

[24] Elisabeth Maria Van der Elst, Judie Mbogua, Don Operario, Gaudensia Mutua, Caroline Kuo, Peter Mugo, Jennifer Kanungi, Sagri Singh, Jessica Haberer, Frances Priddy, et al. High acceptability of hiv pre-exposure prophylaxis but challenges in adherence and use: qualitative insights from a phase i trial of intermittent and daily prep in at-risk populations in kenya. *AIDS and Behavior*, 17(6):2162–2172, 2013.

[25] Norma C Ware, Monique A Wyatt, Jessica E Haberer, Jared M Baeten, Alexander Kintu, Christina Psaros, Steven Safren, Elioda Tumwesigye, Connie L Celum, and David R Bangsberg. What's love got to do with it? explaining adherence to oral antiretroviral pre-exposure prophylaxis (prep) for hiv serodiscordant couples. *Journal of acquired immune deficiency syndromes (1999)*, 59(5), 2012.

[26] Jason Weston, Frédéric Ratle, Hossein Mobahi, and Ronan Collobert. Deep learning via semi-supervised embedding. In *Neural Networks: Tricks of the Trade*, pages 639–655. Springer, 2012.

[27] Sean D Young and Devan Jaganath. Online social networking for hiv education and prevention: a mixed methods analysis. *Sexually transmitted diseases*, 40(2), 2013.

[28] Sean D Young, Caitlin Rivers, and Bryan Lewis. Methods of using real-time social media technologies for detection and remote monitoring of hiv outcomes. *Preventive medicine*, 63:112–115, 2014.

[29] Xiaojin Zhu. Semi-supervised learning. In *Encyclopedia of machine learning*, pages 892–897. Springer, 2011.

Biographical Sketch: Patrick Breen

(a) Professional Preparation

Bowdoin College BA Biochemistry and Mathematics, Brunswick Maine, 2013
University of Georgia PhD Bioinformatics, Athens Georgia, (attending)

(b) Appointments

Oak Ridge National Laboratory Fellowship (2016)
President of Bioinformatics Graduate Student Association (2015)
University of Georgia Presidential Graduate Fellowship (2013)

(c) Products

Mining Pre-Exposure Prophylaxis Trends in Social Media. Patrick Breen, Jane Kelly, Timothy Heckman, Shannon Quinn. DSAA2016.
P2Y6 receptor antagonist, MRS2578, inhibits neutrophil activation and aggregated NET formation induced by gout-associated monosodium urate crystals. Payel Sil ... Patrick Breen ...

(d) Synergistic Activities

Interacted with the local scientific community by judging at high school science fair (2014)

Data Management Plan

Multiple filtered streams of twitter data relevant to diseases studied will be acquired using Twitter's public streaming API. This data includes tweets related to Pre Exposure Prophylaxis, Truvada, HIV, and AIDS as well as other infectious diseases, and commonly used prescription medications. In addition to twitter data we will also acquire other social media data such as Reddit or Facebook, either from an open data repository, or in the case of Reddit through the public API. Direct medical data will be acquired from Oak Ridge National Laboratory and other sources and will be used in accordance with institutional and national laws and regulations (including HIPAA) governing appropriate use of medical data.

Data will be analyzed on local researcher's desktop machines and on research server clusters including the Quinn research group cluster, the Georgia Advanced Computing Resource Center, and Oak Ridge's institutional computing resources. Code developed for analyses will be made openly available on github.com under open source licenses.

Most of the primary information acquired through Twitter's API or direct medical information cannot be shared directly due to Twitter's terms of service and federal regulations such as HIPAA. However anonymized summary information can and will be disseminated in the form of research article publications, and perhaps also through blog articles and other non-technical mediums.

Collaborators and Other Affiliations Information

Collaborators and Co-Editors

Shannon Quinn (University of Georgia)
Jane Kelly (Georgia Department of Public Health)
Timothy Heckman (University of Georgia)
Arvind Ramanathan (Oak Ridge National Laboratory)

Graduate Advisors and Postdoctoral Sponsors

Shannon Quinn (University of Georgia)

Committee Members

Shannon Quinn (University of Georgia)
Jan Mrazek (University of Georgia)
Timothy Heckman (University of Georgia)
Keith Campbell (University of Georgia)