**Zeshan Hussain**

**Using a Machine-Learning Based Approach to Predict the Outbreak of Vector-**

**Borne Diseases in Rural India**

**Abstract:**

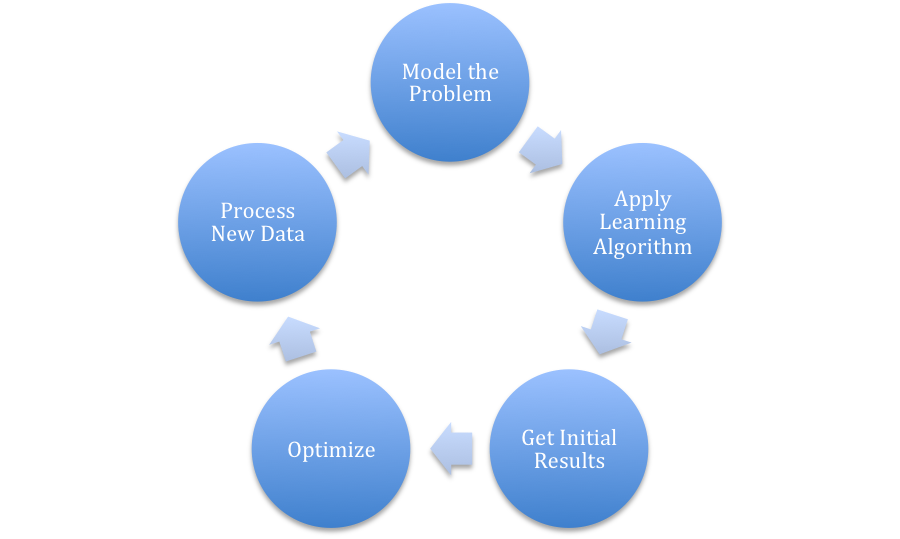
Currently in India, the healthcare space does not effectively utilize the massive amount of data that is being collected via mobile devices, the internet, and hospital instruments, to target inefficiencies and curb disease.[[1]](#footnote-1) In particular, the outbreak of vector-borne diseases in rural India is a recurring problem. For example, in the case of Malaria, which affects around 89% of the local population and costs ~60-90 million dollars to control.[[2]](#footnote-2) Excluding sub-saharan Africa, the death rate due to malaria in India is twice as much as the global average.[[3]](#footnote-3) However, the data that is collected on these diseases by existing health programs, such as the National Vector Borne Disease Control Program, presents an opportunity for clinicians, scientists, and researchers, to predict the outbreak of disease, streamline disease prevention, and improve resource allocation using modern information technology and data analytics. Furthermore, given this opportunity, the objective of this project is two-fold: 1) to create a machine learning system that predicts outbreaks of a vector-borne disease (we will focus on Malaria as a proof of concept), visualizes this information in an understandable way for clinicians/care-providers, and allows health professionals to optimize resource allocation; 2) to serve and volunteer at local clinics to determine what current disease prevention solutions are in India and to use domain knowledge to optimize the learning framework. The ultimate goal of this project is to construct a generalizable system that will not only improve clinical workflow in Malaria prevention/treatment, but also the treatment and prevention of other vector-borne diseases.

**Literature Review:**

The biological mechanisms of the transmission of vector-borne diseases, as well as key environmental factors necessary for outbreak, have been thoroughly studied. For example, Goswami et al. and Mutheneni et al. explore the relative roles of weather variables and seasonal patterns in the modulation of malaria and Japanese encephalitis. However, while there have been basic statistical models that have been applied to vector-borne disease forecasting (Lauderdale et al.) and databases that have been created for disease data storage (Updahyayula et al.), machine learning specifically has not been utilized as a predictive framework. Thus, implementing a complete machine learning pipeline that processes relevant data, extracts useful information, and predicts on queries will be a novel and valuable contribution to the field.

**Design of Framework**

There will be a few main components to the machine learning system. The first component will be the infrastructure. The infrastructure will be vital for data processing and cleaning, as well as data storage. This part of the system will put the data in a format that can be used by a learning (predictive) algorithm. The second component will be the modeling of the problem. Because of the timeframe of this project, I will first focus on predicting the outbreak of malaria to serve as a proof of concept for other vector-borne diseases. I have chosen malaria as a sufficient candidate for proof of concept since its transmission mechanism is fairly representative of other vector borne diseases. As in any machine learning problem, it is conventional to start with a simple model and add more complexity as initial results are generated and analyzed. Thus, I can begin by modeling malaria outbreak as a supervised classification problem. For some input, such as a particular location or coordinates, the model can predict whether there will be an outbreak of malaria at that location. The final component will be the learning algorithms that will be trained on the data and applied to the model. The final system will be the result of multiple optimizations on this baseline machine learning system.

However, the key to optimization is gathering domain knowledge and being aware of the structure of the data. The most efficient way to gather domain knowledge is to, naturally, visit the actual sites of outbreak, analyze the parasitic load (i.e. the potential for disease outbreak) of rural communities, and discuss this with doctors, healthcare practitioners, and nurses.

Furthermore, by acquiring knowledge on the pain points of vector-borne disease treatment, I will be able to input additional relevant features that will improve the accuracy of my machine learning system.

Being able to observe standard clinical solutions, discussing potential improvements with healthcare providers, and designing a system that makes a nontrivial impact is incredibly valuable. However, these facets of the project cannot be done remotely (at least, it would be very difficult and inefficient to do so); thus, traveling to India is pivotal to this research.

**Resources and Preparation:**

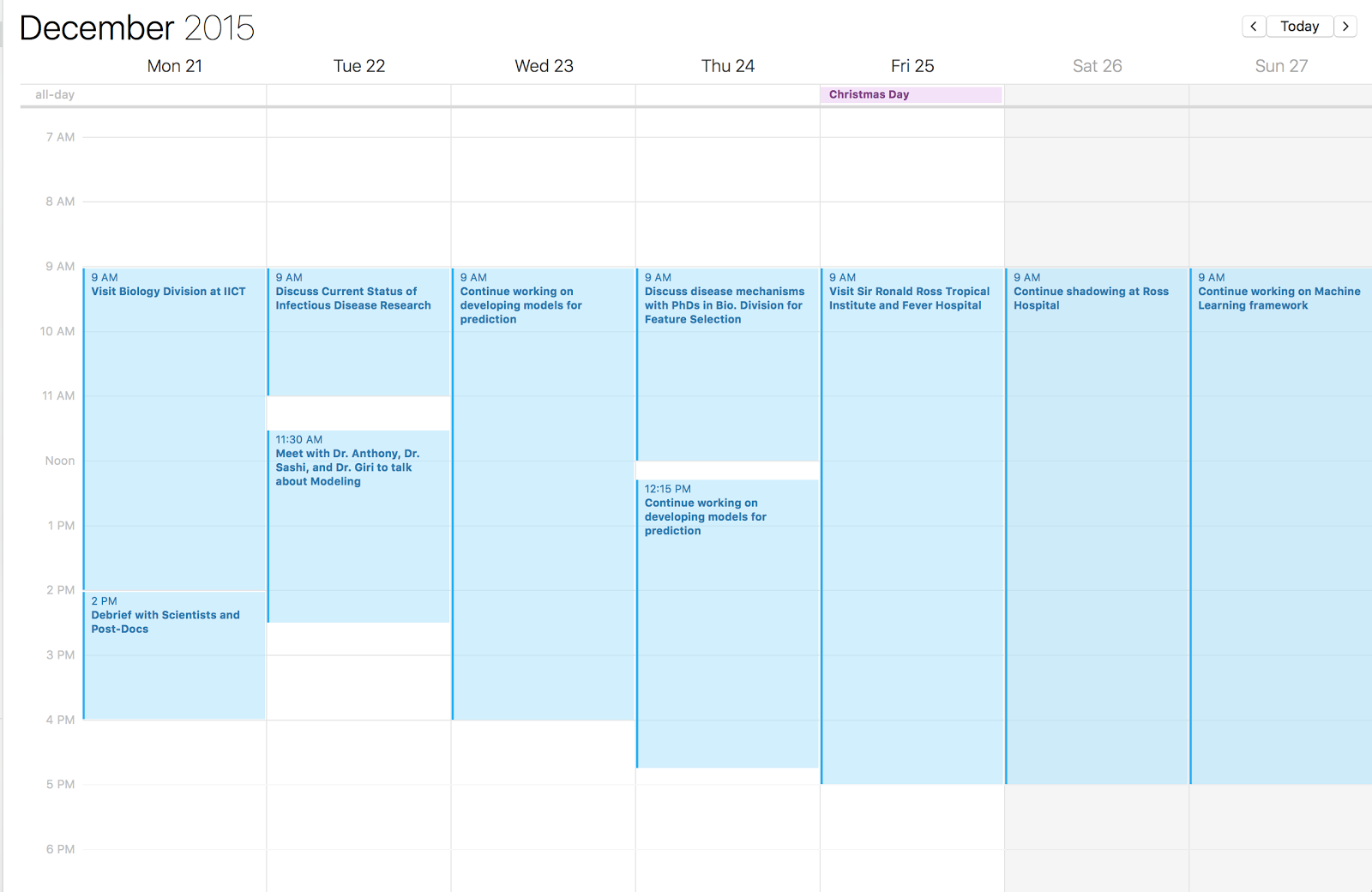
The 2 week-trip will be used primarily for field work. The goal is not necessarily to collect data, but to figure out what weaknesses there are in the data as well as to increase my domain knowledge by working with experts and volunteering at clinics. Again, because the overall goal of the project is to implement a robust machine learning system, it is more important to detect the weaknesses in the data and what features to add than to collect data randomly.

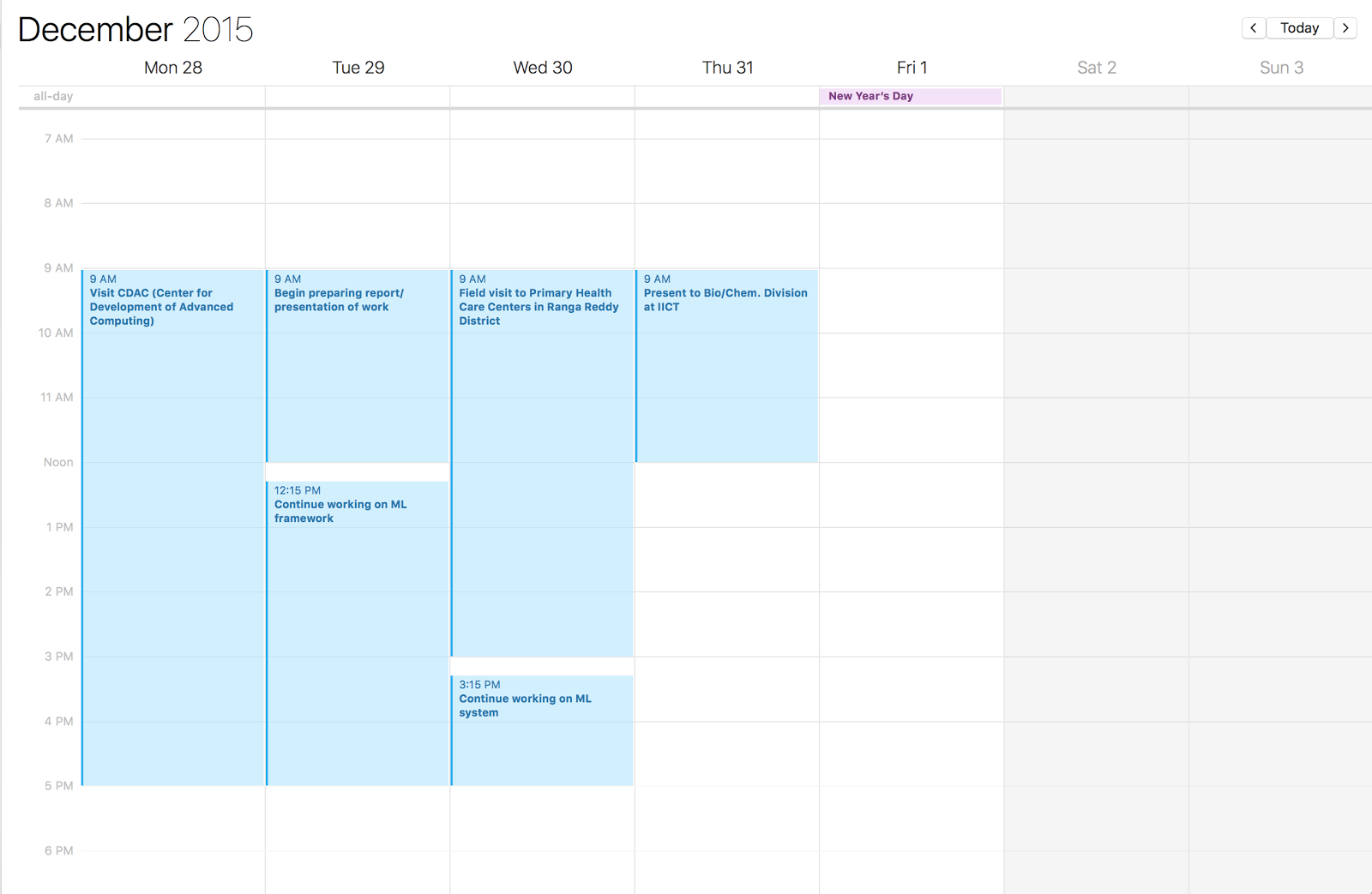
I will be spending a considerable amount of time beforehand developing a pipeline for data processing (infrastructure) and implementing basic models to produce initial results. These initial results will be useful in deciding how to improve the system, whether it be inputting new features by talking to experts, adjusting the model to reflect the actual structure of the data, etc. Because the goal is not to actually collect any data, but rather to work through improving an already implemented baseline system with other researchers, I believe two weeks is enough to accomplish the goal.

Briefly, there has been precedent for information technology solutions in India; one example is the Indira Gandhi Conservation Monitoring Center, which attempted to be a national information provider based on a set of core environmental information systems. Unfortunately, this project failed due to its infeasible implementation and specifications. Sustaining such a project depends on what the goals of the project are, the resulting design derived from those goals, and if that design is in-line with reality (i.e. is it feasible?) (Information Technologies). My project has reasonable goals: it is very possible to implement a machine learning system prototype in 2-3 weeks, evidenced by my previous work in CS229 (Machine Learning) and CS373 (Machine Learning and Statistical Methods in Genomics). The project also has backing from the Indian Institute of Chemical Technology (IICT), the institution that I will be partnering with, to maintain it after my departure. Finally, with both the resources from IICT and Stanford as well as my background in CS, I believe that it is possible to implement a sustainable and robust machine learning system at scale.

[Indira Gandhi Information Technology Book](https://books.google.com/books?id=6ryRITfh-FMC&pg=PT195&lpg=PT195&dq=Indira+Gandhi+Conservation+Monitoring+Center+failure&source=bl&ots=PR9Ci2wh20&sig=dlRBHdT6HDxQWYm621hAmLfeOME&hl=en&sa=X&ved=0ahUKEwjwo8CKlKHJAhVKOIgKHRMLDaUQ6AEIKjAC#v=onepage&q=Indira%20Gandhi%20Conservation%20Monitoring%20Center%20failure&f=false)

Calendar:





Budget:

|  |  |
| --- | --- |
| **Travel Components** | **Travel Budget** |
| Flight (Chicago to Hyderabad) | $1,500 |
| Flight (Hyderabad to Chicago) | $1,500 |
| Total | $3,000 |

1. Two examples of data sources are the eHealthCenter datastream and the OpenEMR database (<http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1001468>) [↑](#footnote-ref-1)
2. <http://www.who.int/malaria/publications/country-profiles/profile_ind_en.pdf> [↑](#footnote-ref-2)
3. <http://www.theguardian.com/news/datablog/2012/feb/03/malaria-deaths-mortality> [↑](#footnote-ref-3)