Kondria
COMPSCI 316 Project Milestone 1 - Progress Report
Emily Mi, Davis Booth, Thara Veeramachaneni, Elliott Bolzan, Jamie Palka

**Description of application**

Recall the last time you were sick. These days, your first instinct is probably to reach into your pocket, pull out your smart phone, open its default internet browser, and type in your symptoms. Google then crawls millions (possibly billions) of websites to return the most relevant results to you. Putting your full trust in Google, you tap one of the first couple of options, which usually contain the common culprits: WebMD, Mayo Clinic, etc. The page loads and you scroll down the page with your thumb, tensing your entire body in anticipation of the result. To your displeasure, WebMD informs you that you have contracted cancer, multiple sclerosis, hepatitis, or some other potentially terminal illness. From here you have two options. First: freak out, say your goodbyes to your closest friends and family, get your will in order, and prepare for your inevitable death. Or second: deem WebMD useless, call your doctor, and schedule an appointment only for your doctor to diagnose you with a common cold. Clearly, neither option leaves you in a better place than before. With all of the amazing research that currently exists in the medical field, it seems that there should be an easy way to at least get a general idea for the type of illness you could have with a simple search of the internet. Kondria puts this incredible ability in your pocket. An iOS application that uses an intelligent keyword filtering algorithm to suggest potential illnesses to you, Kondria flips the conventional model of determining illness. Kondria allows users to specify the symptoms they are experiencing with their illness, and in return it provides a table of viruses sorted from top to bottom by likelihood. Using queries, we want to rank diseases based on the similarities of the symptoms they are related to and the symptoms input by the user. The database will be updated when users create profiles and input their data.

**Plan for getting the data to populate our database and some sample data**

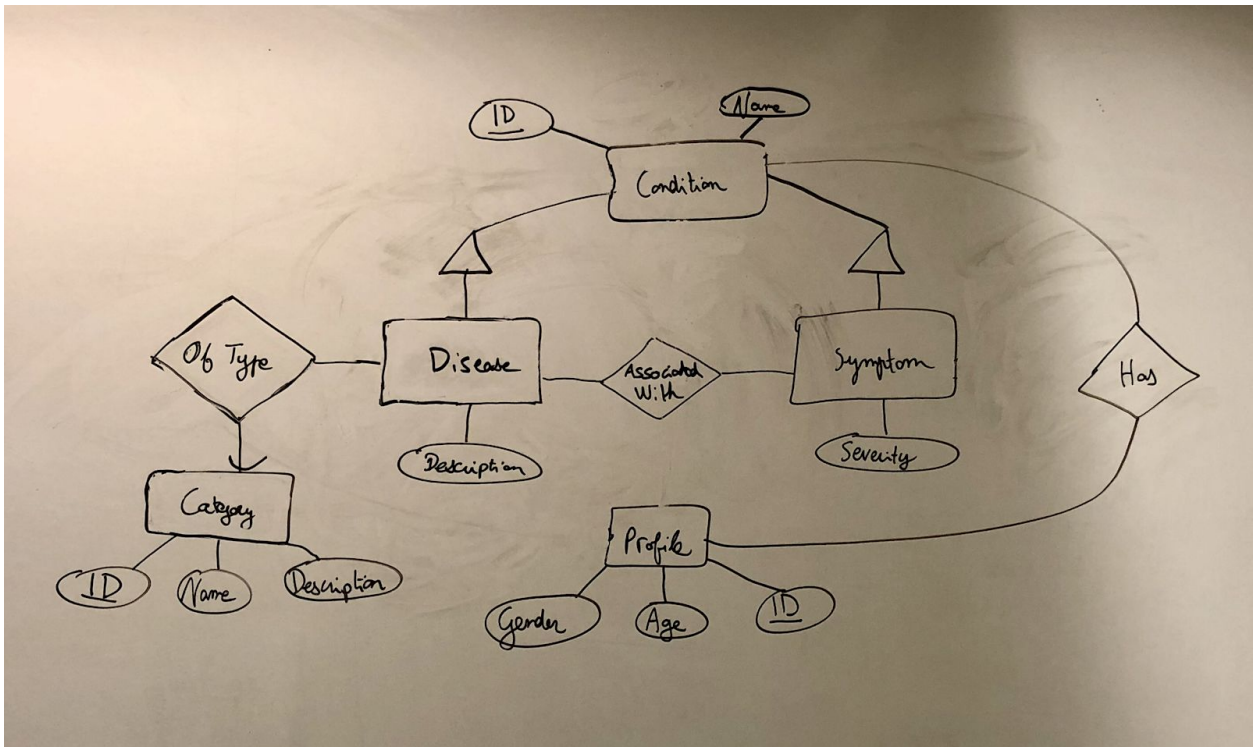
A number of databases provide symptom to diagnosis data. Among these, we have identified two rich sources of data: www.diseasesdatabase.com and medlineplus.gov/encyclopedia.html. To populate our original database, we plan on scraping these two sources to compile information on symptoms and the diseases related to them. To scrape these databases, we plan on using python and the BeautifulSoup module, which is specifically designed for scraping web pages. Of these two databases, the former, Diseases Database, will be easier to parse. Indeed, it lists diseases alphabetically and provides well-defined URLs to the symptoms that cause it and the other conditions it can lead to. The second of these databases, medlineplus.gov, provides more information, but with a less clearly defined structure. To parse this data, we will have to parse freeform descriptions of symptoms and infer keywords from them. Similar operations could be performed on larger, more famous datasets, like WebMD's data.

We will describe this process by providing some sample data. Consider the entry on radiculitis from the Diseases Database (http://www.diseasesdatabase.com/ddb29521.htm). The website provides a description ("Spinal nerve root inflammation") and lists to symptoms it may cause ("Radiculopathy"), as well as the general category it belongs to ("Inflammatory conditions"). By scrapping Diseases databases, we would obtain the following information: ["Radiculitis", "Spinal nerve root inflammation", "Radiculopathy", "Inflammatory conditions."

**List of assumptions that we are making about the data being modeled**

- People use their phones/the internet to initially diagnose illnesses
- There is no liability involved in suggesting most likely viruses to users
- Databases mapping viruses and diseases to symptoms exist
- The databases specified above are updated regularly in response to new research
- Diseases and viruses can be at least generally identified by a list of symptoms and other data points that don't require active testing (i.e. stethoscope listenings, strep throat tests, etc.)
- This does not require our team to have a member who has a license to practice medicine
- This technology/idea is not under patent or specific restrictions by the medical field
- Diseases are not symptoms for other diseases
- A disease corresponds to exactly one category

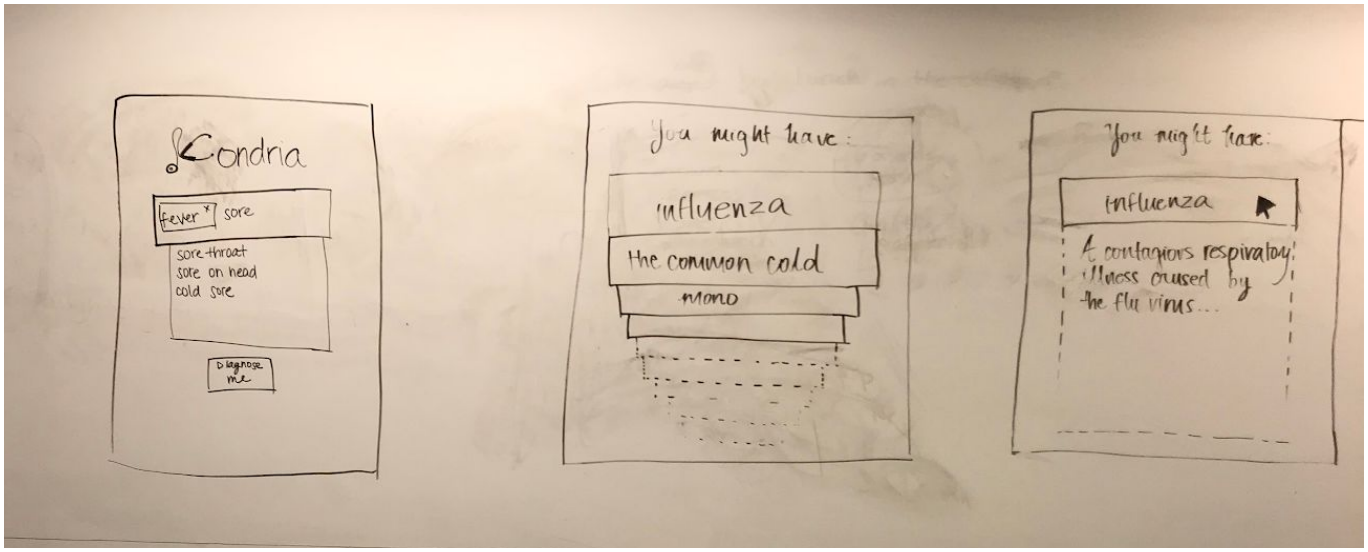
**E/R diagram for our database design**



**List of database tables with keys declared**

Profile(<u>id</u>, age, gender)
Condition(<u>id</u>, name)
Disease(<u>id</u>, description)
Category(id, name, description)
AssociatedWith(symptonID, diseaseID)
OfType(diseaseID, categoryID)
Symptom(<u>id</u>, severity)
Has(<u>profileID</u>, <u>conditionID</u>)

**Description of the Web interface**




**1** **2** **3**

1: What you see when you open the iOS app and begin to type in symptoms (autocomplete function allows you to choose options)

2: What you see when you press “diagnose me” - the top 5 most likely diagnoses corresponding to the input symptoms.

3: What you see when you click on a specific diagnosis - the name of the diagnosis, a description of the diagnosis, and its category.

In addition, we did some research on potential similarity metrics. We plan on using these to compare the set of symptoms provided by the user and the set of symptoms associated with each disease.

Comparing Sets of Symptoms

- https://en.wikipedia.org/wiki/Jaccard_index
- https://en.wikipedia.org/wiki/S%C3%B8rensen%E2%80%93Dice_coefficient
- https://en.wikipedia.org/wiki/Simple_matching_coefficient
- https://en.wikipedia.org/wiki/Tversky_index

Potential Datasets

- This one is fire: http://www.diseasesdatabase.com/
- This one too: https://medlineplus.gov/ency/article/003042.html
- https://www.quora.com/Where-can-I-find-symptom-and-disease-DATASET-to-build-a-disease-prediction-engine
- https://www.kaggle.com/plarmuseau/sdsort
- http://git.dhimmel.com/hsdn/