Using Alexa voice-activated artificial intelligence to determine the cause of disease outbreaks

Sponsored by the Centers for Disease Control and Prevention

KSU SwE 7903 Project Proposal, Fall 2017

Project Overview and Motivation

Public health departments have a daunting mission: Preventing disease, injury, and disability, promoting health and well-being, and preparing for and responding to disasters. For disease outbreaks in particular, there is a need to rapidly collect and analyze data. One challenge public health departments face in using software tools for data analysis is cost: Products may be too expensive to purchase, or if provided for free, they may be too expensive to operate and maintain. Another is complexity. Most tools are highly generalized and usable across many industries, such as oil and gas, financial, aviation, sales, logistics, and healthcare, among others. Supporting so many features incurs a usability penalty and steepens the software's learning curve.

The U.S. Centers for Disease Control and Prevention (CDC) has recognized that a need exists for a free, fast-to-learn, easy-to-use software suite that is designed around epidemiologists and their public health mission. CDC's Epi Info™ software is designed to address that need. Epi Info™ allows users to rapidly create data entry forms with customized business logic, enter data into those forms, analyze the data, and then display the data on a map. The current technology stack for Epi Info™ spans the Windows desktop environment, iOS and Android, the web, and the cloud. Epi Info™ has over one million users across 180 countries and has been involved in several high-profile public health missions, including in support of the 2014-2015 West Africa Ebola outbreak [1].

However, Epi Info's analysis capabilities remain driven by human intelligence: Humans decide what variables to analyze, which variables should be transformed or filtered, and which analysis routines need to be executed. This is a time-consuming process since outbreak data sets may contain hundreds of variables. Furthermore, the sheer number of visualizations (e.g. tables and graphs) generated by running hundreds or even thousands of analytical routines can be overwhelming. Unfortunately, often lost under the plethora of data points, statistics, and shiny visualizations is the true purpose of epidemiologic data analysis in the first place – finding the source of an outbreak and ending it.

We propose an alternative: Building an open source[1] artificial intelligence component, leveraging Epi Info™, that summarizes any given outbreak data set on demand and perform the relevant analyses to determine the most likely causes of the outbreak. The user interaction will be facilitated by the Amazon Alexa voice service (see https://developer.amazon.com/alexa). The interactions will be conversational and allow the user to ask pointed questions – and follow-up questions – to get deep insights into the outbreak data. A typical Alexa conversation might proceed as follows:

Epidemiologist: Alexa, load my outbreak data set

Alexa: Ok, I've loaded the data. There are 671 records currently. What would you like to know?

Epidemiologist: How many people are ill?

Alexa: 257 people reported feeling ill.

Epidemiologist: What is the profile of the people who are sick?

Alexa: It looks like males in the 30s have the highest incidence for this illness.

Epidemiologist: What's the likely cause of the outbreak?

Alexa: The likely cause of the outbreak is the bean sprouts.

Epidemiologist: Why do you think the outbreak was caused by the bean sprouts?

Alexa: I've noticed that the risk ratio for a cross-tabulation of ill and exposure to beansprouts was the highest among all variables.

Epidemiologist: What are the top five variables with the highest risk ratios?

Alexa: Beansprouts, potato salad, Reuben sandwich, chocolate chip cookies, and brownies.

Epidemiologist: What's the risk ratio for a cross-tabulation of ill and potato salad?

Alexa: The risk ratio is 1.7.

Epidemiologist: What's the risk ratio for a cross-tabulation of ill and bean sprouts?

Alexa: The risk ratio is 3.6.

In the span of perhaps a minute, the user is able to complete an analysis of the data that would otherwise require many steps in a traditional keyboard/mouse-driven user interface. The conversational nature of the interaction mirrors how epidemiologists talk to each other about disease investigations, reducing the reliance on specialized software training and know-how.

The way this is intended to work is straightforward. The user asks a question, which is converted into plain text and then processed by Alexa. Students will use the Alexa SDK to program Alexa to identify specific phrases as questions or commands. Those phrases are then sent to the AI HTTP endpoint where the AI runs a computation based on the phrase, then returns the appropriate text back to Alexa. See Figure 1.

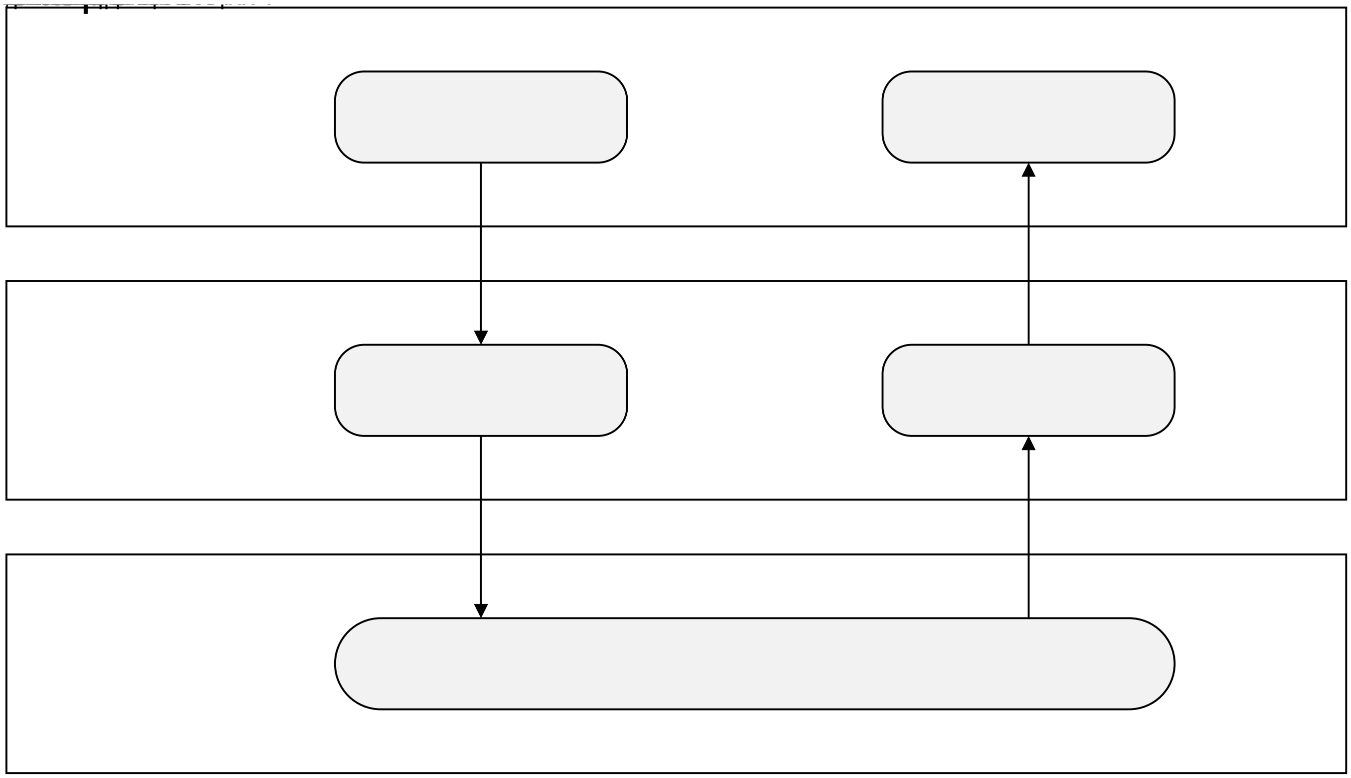

Figure 1: Flow diagram for human-Alexa interaction

The result will be quicker identification of the leading causes of disease outbreaks, in turn allowing government agencies to be more responsive to the public – and perhaps even saving lives.

Requirements

The software will need to meet the following requirements:

1. The application shall correctly identify the cause of an outbreak for a given set of outbreak data
   1.1. "Correctly" in this context means the application shall identify the same cause as a human epidemiologist would
   1.2. All "outbreak data" used to validate this requirement will be provided to students by the CDC
   1.3. All "outbreak data" provided by CDC will consist of de-normalized (i.e. flat) data
2. The application shall correctly explain to users why it made a decision to identify a given cause of an outbreak
   2.1. "Correctly" in this context means the application shall provide the same explanation as a human epidemiologist would.
3. The application shall answer questions about the data when asked by the user. Questions include:
   3.1. How many records meet certain criteria? For example, "How many people are ill?"
   3.2. What is the output of a certain statistical method? For example, "What is the odds ratio for ill and bean sprouts?" Valid output requests include:
      3.2.1.Odds ratio
      3.2.2.Risk ratio
      3.2.3.P-value
4. The application shall return an identifying response to the user in no more than 10 seconds after the application has received a question.
   4.1. "Identifying response" in this context refers to a response that correctly identifies the cause of an outbreak.
5. CDC would also strongly prefer the application be extensible so that future developers can add more supported phrases, statistical routines, and question types without much trouble. The purpose is to ensure that what students build can be used as a *framework* that CDC or future KSU project teams can continue to build on and improve.

Solution Design

The application is likely to need three major components:

1) An artificial intelligence module. The module should select statistical methods to run, make a decision about what looks like the most reasonable cause of the outbreak are given those statistics, and then return that data to the caller. It should also be capable of explaining its reasoning.
2) A statistical module. This module will contain the math needed to compute various statistical outputs such as risk ratios, odds ratios, and p-values. Note that Epi Info™ already contains the needed statistical code in C# which can be found on its open source project website (epiinfo.codeplex.com). Depending on the implementation, this code may be taken as-is or translated into other languages such as Kotlin, Java, or Ruby.
3) An Alexa interface module. This module contains the logic that parses the plain text input from the Alexa service into one of several supported Alexa "phrases." Note that Alexa has a robust API and SDK that greatly assist in building phrases and interacting with Alexa.
4) A component that interfaces with a database. Note that the application will not know what the database schema looks like at compile time or design time, so programming against a defined schema is not possible. This database component only needs to support read operations.

Given some outbreaks generate large data sets and that running an AI routine across many variables simultaneously is compute-intensive, some form of parallel processing, distributed computing, and/or cloud computing may be required for the AI to return the correct result within a reasonable amount of time. Students are highly encouraged to explore ways to maximize performance of the AI while still keeping overall computing costs minimal.

Importantly, no hardware is needed to begin. Students can interact with Alexa using plain-text inputs, essentially pretending that an Amazon Echo (or other similar voice-to-text device) has converted spoken words into plain text for them. This will allow KSU to proceed with the project without spending money on hardware devices.

The format of the outbreak data will be flat. That is, students do not need to worry about joining tables in a relational database. All data CDC will provide will be formatted in rows and columns in a single table.

Why choose this project?

There are many reasons to choose this topic. Primarily, you will be learning about AI, statistics, and voice-driven human-computer interaction, all of which are hot topics in the technology industry.

Second, you will learn about how disease investigations are conducted and how CDC supports those efforts. You'll also get a glimpse into how epidemiologists in your own communities investigate and respond to outbreaks and how software can help. As CDC has indicated they will provide assistance and guidance to students, you will have several opportunities to interact with CDC employees throughout the semester.

Third, this project will be a challenge: It is unlikely most students have used the Alexa SDK, worked with artificial intelligence and statistics, and programmed software to work against a database where the database schema is unknown at compile time. However, these challenges speak to potential employers about how well you can rise to difficult situations, learn new topics quickly, and apply that learning to come up with a solution.

Finally, developing free and open source software is a way to contribute to your community. Epi Info was used in several major public health initiatives in 2015 alone and has millions of users [1]. Should some or all of your product be picked up by CDC, rest assured that you will have made a lasting contribution to helping people all over the world.

1 – Epi Info™ is licensed under the Apache License 2.0

References

[1] U.S. Centers for Disease Control and Prevention, "Epi Info 2015 Annual Report," 2015. [Online]. Available: https://www.cdc.gov/epiinfo/pdfs/annualreport/2015_annualreport_epiinfo.pdf.