Infectious Disease Surveillance, Big Data, and the HealthMap Platform

Sumiko R Mekaru, sumiko.mekaru@childrens.harvard.edu; John S Brownstein,
john.brownstein@childrens.harvard.edu

When HealthMap launched in 2006, our goal was to integrate the vast but separate online information about infectious disease events by providing a freely accessible, automated platform that organized data about outbreaks according to geography, time, and disease agent. Since then, the percentage of Internet users has grown from 18% to 39% of the global population. As more news services, ministries of health, and international agencies provide disease data online, we can see the global picture of disease more clearly. Additionally, as individuals increasingly use the Internet to learn more about diseases of concern, they leave trails of metadata that can serve as proxies for disease incidence, helping researchers and public health officials respond to health threats more efficiently. Big Data offers opportunities to investigate new research questions, but methods to manage the data and to create sound study designs are critical; computer scientists and epidemiologists must collaborate to capitalize on the full potential of Big Data. Here, we describe the HealthMap platform's approach to organizing large volumes of data as well as research applications for the resulting dataset.

The HealthMap automated text processing system proceeds in four stages: (1) Acquisition (2) Characterization (3) Filtering (4) Clustering. Trained public health analysts then review content and correct misclassifications, providing edited data to improve the automated algorithms through a training feedback loop.

The first step is **data acquisition**. The system collects data via five main channels: news aggregators, specific RSS feeds, email subscriptions, custom-parsed HTML scraping or CSV/XML feeds, and user submitted reports. The framework supports new feed acquisition through a custom PHP class, where the feed-specific code fetches the content, parses it and then extracts a set of common fields.

The common outputs of the feed framework then pass to the **characterization** engine which extracts disease and location entities, flags the report as not-disease-related by rule (if applicable), and identifies the source publication. This module loads large language-specific dictionaries of terms into memory and then matches them against the input text using a rapid matching algorithm. The dictionaries currently support identification of over 250 diseases. HealthMap has an internal location database of over 25,000 locations.

Once disease and location are assigned, the document passes to a **Bayesian filtering module** which breaks the it into unigrams, digrams, trigrams (and tetragrams in some languages), and then computes a likelihood score based on training data from tens of thousands of previously recorded documents. It then assigns the document to one of five categories based on relevance; documents marked Breaking News and Warning are posted to the map, whereas documents in the Old News, Context, or Not Disease Related categories are simply stored in the database for use in future research.

Finally **document clustering** groups duplicate reports together. The clustering module compares the given document to each previously collected document over a fixed time window, and computes a similarity score based on six factors: (1) character-level headline comparison, (2) word-level headline comparison—fraction of words in common, with inverse term frequency taken into account, (3) disease and place comparison, from the output of the named-entity extraction engine (4) country-level comparison, i.e., if the two reports have different locations within the same country, boost similarity score somewhat, (5) timestamp—reports closer in time are more likely to be similar, (6) filter category comparison, based on the output of the Bayesian filtering module.

Following automated processing, HealthMap alerts are posted directly to the public map, partner pages and RSS feeds. HealthMap analysts review all incoming data and correct misclassifications and assign additional higher resolution metadata classifications such as case counts and outbreak description tags (e.g. school, mass gathering, travel-associated).

The resulting database has been used to investigate numerous research questions. For example, in Fisher *et al*'s "Emerging fungal threats to animal, plant and ecosystem health," the authors utilized the HealthMap database (which includes ProMED records) to assess worldwide reporting trends for fungal diseases and assess fungal threats to health. In a current HealthMap-CDC collaboration, researchers are comparing media and official foodborne outbreak reporting in the United States to identify gaps in the official record. While news reports and other publicly available Internet resources undeniably fail to gather the full range of outbreak data, official data also has omissions. Through combining official data with information mined from publicly available Internet sources, a more complete picture emerges.

The complementary nature of the described research projects addresses a significant concern about Big Data, that the data collected reflects unmeasured biases and may not be generalizable. Given the constantly evolving demographics of Internet users and content providers, Big Data studies should as often as possible include comparisons to gold standard data (or at the least data for which the underlying biases or distributions of confounding variables are understood). HealthMap researchers assisted in the development of Google Dengue Trends. The final product is available for only a subset of all countries with dengue because a viable model requires not only Google users who search for dengue-related terms but also gold standard dengue data against which the model could be developed and tested. The biases and confounders of user behavior vary by country, are unmeasured, and are poorly understood. Therefore, Google Dengue Trends was only implemented for locations where there could be a reasonable expectation of validity.

Big Data offers unprecedented data sets for researchers, but it requires platforms which can efficiently and correctly organize large and complicated data as well as caution in the study designs that utilize it. Blind faith in automated methods may result in unchecked assumptions and fundamental misinterpretation of results, especially when data characteristics evolve over time but the automated methods do not address the changes in input. Computer scientists and epidemiologists must collaborate to achieve the promise of Big Data.