Microsoft

AI for Earth Grantee Profile

EcoHealth Alliance

AI-based text analytics for scientific research

Summary

As new infectious diseases emerge and spread in different areas of the world, tracking the outbreaks is an important step in analyzing where they might emerge or spread next. Archives of scientific publications such as PubMed Central present a resource for monitoring this information, but with thousands and thousands of articles published annually without a common standard for presenting data, extracting that data is very challenging. EcoHealth Alliance, an international nonprofit organization dedicated to preventing pandemics and protecting both human lives and wildlife, is turning to AI to meet this challenge. With assistance from a Microsoft AI for Earth grant, EcoHealth Alliance is developing PubCrawler, an AI-based software project that uses natural language processing to produce high-resolution datasets of the locations where research is being done into various diseases. The tools of this project also can be applied more broadly to meet other needs in biodiversity and conservation research.

Tracking diseases through scientific literature with AI

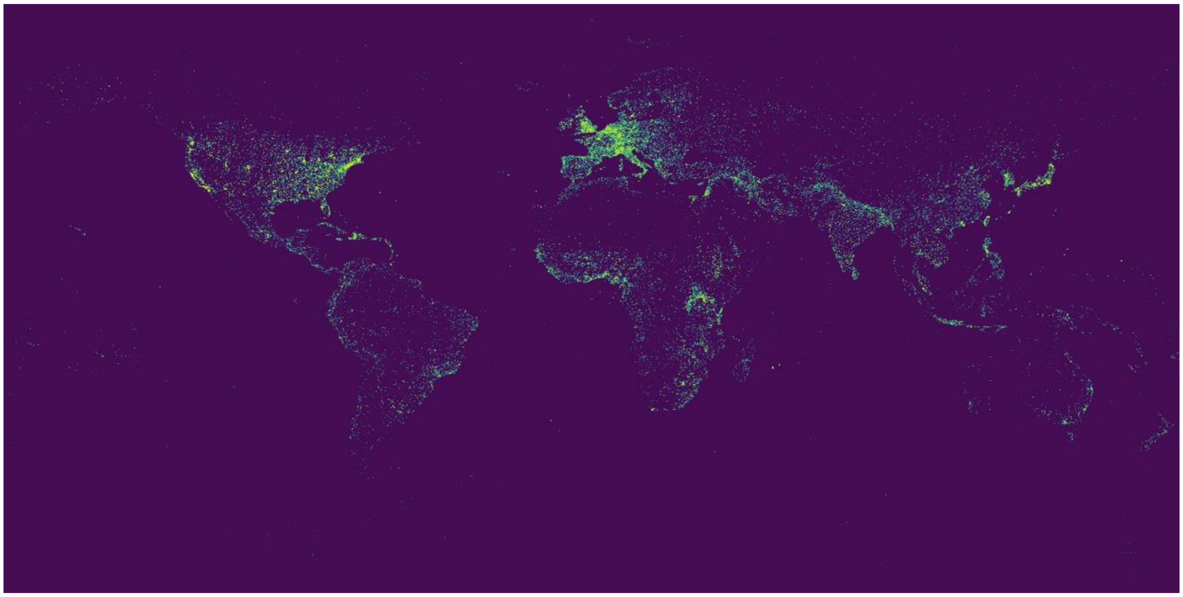
"One of the big problems faced by organizations working on this topic is there's no good consistent global dataset for emerging infectious disease outbreaks," explains Toph Allen. Allen is a Senior Data Scientist with EcoHealth Alliance. EcoHealth Alliance leads research into public health and environmental science with the goals of preventing pandemics and promoting conservation. One aspect of its work is keeping abreast of the scientific literature on various emerging diseases, but as Allen says, this is no easy task.

> "There's no good consistent global dataset for emerging infectious disease outbreaks."—Toph Allen, Senior Data Scientist, EcoHealth Alliance

No single standard format is in use for publishing articles and data, and the amount of information is ever increasing. For example, the US National Library of Medicine maintains PubMed Central, a free full-text archive of biomedical and life sciences journal literature. The archive's Open Access Subset has grown in the past few years from 500,000 articles to nearly 2 million, and more research is being published under open access guidelines all the time.

Identifying locations with natural language processing

While it'd be impractical to have a team of people searching through the entire PubMed archive for relevant articles, computers with modern AI are well-suited for the task. EcoHealth Alliance has been developing algorithms that use natural language processing to search through articles, identify and tag relevant terms—such as mentions of an infectious disease—and create a dataset of this information. Of particular interest for the organization's purposes is knowing not just whether a disease has been reported, but also where and how many cases. Correlating this information would allow EcoHealth to map outbreaks and create informed analyses on where a disease might emerge or spread to next. With the support of Microsoft, including a grant from Microsoft AI for Earth, EcoHealth is applying its technology to meet this need.



A world map of relevant disease terms by location. [Image courtesy of EcoHealth Alliance]

Through the AI for Earth grant, EcoHealth Alliance was able to collect a better set of training data for one of its algorithms. EcoHealth had one of its applications, the GeoName Curator, process nearly 1,000 PubMed articles to annotate place names, and then worked with a team from iMerit to review and correct the annotations. Through this review, the algorithm was refined to distinguish the context of a location's mention, such as being referenced as the site of field work versus the home university of the researchers. Likewise, the algorithm needed to disambiguate between similarly named locations—determining whether "New York" referred to the state or the city, for instance, or whether "Portland" meant the city in Oregon or in Maine (or one of numerous other locations by that name). With the new data from the grant, the algorithm will have better precision and recall, catching more locations with greater accuracy and returning fewer false positives.

Mapping disease outbreaks—and other research—with AI

Work on the GeoName Curator expands the work of EcoHealth Alliance's PubCrawler project, an ongoing effort to scale up useful scientific datasets using the information in open-access scientific literature like PubMed Central. Down the road, EcoHealth plans to manage the greater scale by moving parts of its workflow to Microsoft Azure. Cloud computing would allow the organization to run its processor-intensive algorithms at will, updating its models quickly, and keep up with the quickly growing archive of articles. EcoHealth would also be able to extend its research to the Europe PubMed Central, adding another 4.6 million open-access articles to the dataset. Although most articles in the two archives are in English, by connecting to the Azure Translation Text API service, PubCrawler will be able to process those articles which are in other languages as well.

> "There are a lot of things you could do by applying these sorts of models.... You could get a much better picture of what research is being done, where."—Allen

With Azure Cognitive Services, EcoHealth Alliance can also expand the scope of PubCrawler to other research topics. "There are a lot of things you could do applying these sorts of models," says Allen. "Let's say we take the place name extractor that we've trained, and then run that on the whole set of open access research articles. You could get a much better picture of what sort of research is being done where in the world." By using the Azure Text Analytics and Academic Knowledge API services, PubCrawler would be able to make connections for other topics in biodiversity and resource conservation, and generate new metrics and maps for that research. The capability of these algorithms to correlate topics such as disease outbreaks to where research is being done has another important effect: it can reveal inherent bias in funding and efforts, and help scientists correct for that bias.

About EcoHealth Alliance

Building on over 45 years of groundbreaking science, EcoHealth Alliance is a global environmental health nonprofit organization dedicated to protecting wildlife and public health from the emergence of disease. The organization's unique "One Health" approach uses a multidisciplinary method to solve health challenges caused by global changes and human-animal interactions. With this science, EcoHealth Alliance works with local governments, in-country scientists, and policymakers around the world to develop solutions that prevent pandemics and promote conservation.

About Toph Allen

Toph Allen, Senior Data Scientist at EcoHealth Alliance, applies his expertise in epidemiology, statistical modeling, and data science to improve our understanding of disease emergence. His work on EcoHealth Alliance' IBIS project identifies outbreaks in news reports and maps their potential spread through the flight network via infected travelers, showing the data in an interactive web app. His other projects include using machine learning to link drivers of disease emergence to past disease emergence events and developing a new method to correct for bias when analyzing large sets of scientific studies. Toph received his MPH (Master of Public Health) in epidemiology at Columbia University.

Resources

Websites

EcoHealth Alliance home site
GeoName Curator project code on GitHub