

Developing a Prototype Opioid Surveillance System at a 2-Day Virginia Hackathon

Catherine Ordun, Jessica Bonnie, Jung Byun*, Daewoo Chong, Richard Latham and Joshua Wei

Data Science & Machine Intelligence, Booz Allen Hamilton, Fairfax, VA, USA

Objective

A team of data scientists from Booz Allen competed in an opioid hackathon and developed a prototype opioid surveillance system using data science methods. This presentation intends to 1) describe the positives and negatives of our data science approach, 2) demo the prototype applications built, and 3) discuss next steps for local implementation of a similar capability.

Introduction

At the Governor's Opioid Addiction Crisis Datathon in September 2017, a team of Booz Allen data scientists participated in a two-day hackathon to develop a prototype surveillance system for business users to locate areas of high risk across multiple indicators in the State of Virginia. We addressed 1) how different geographic regions experience the opioid overdose epidemic differently by clustering similar counties by socieconomic indicators, and 2) facilitating better data sharing between health care providers and law enforcement. We believe this inexpensive, open source, surveillance approach could be applied for states across the nation, particularly those with high rates of death due to drug overdoses and those with significant increases in death.

Methods

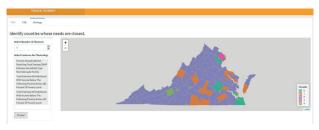
The Datathon provided a combination of publicly available data and State of Virginia datasets consisting of crime data, treatment center data, funding data, mortality and morbidity data for opioid, prescription drugs (i.e. oxycodone, fentanyl), and heroin cases, where dates started as early as 2010. The team focused on three data sources: U.S. Census Bureau (American Community Survey), State of Virginia Opioid Mortality and Overdose Data, and State of Virginia Department of Corrections Data. All data was cleaned and mapped to county-levels using FIPS codes. The prototype system allowed users to cluster similar counties together based on socioeconomic indicators so that underlying demographic patterns like food stamp usage and poverty levels might be revealed as indicative of mortality and overdose rates. This was important because neighboring counties like Goochland and Henrico Counties, while sharing a border, do not necessarily share similar behavioral and population characteristics. As a result, counties in close proximity may require different approaches for community messaging, law enforcement, and treatment infrastructure. The prototype also ingests crime and mortality data at the county-level for dynamic data exploration across multiple time and geographic parameters, a potential vehicle for data exchange in real-time.

Results

The team wrote an agglomerative algorithm similar to k-means clustering in Python, with a Flask API back-end, and visualized using FIPS county codes in R Shiny. Users were allowed to select 2 to 5 clusters for visualization. The second part of the prototype featured two dashboards built in ElasticSearch and Kibana, open source software built on a noSQL database designed for information retrieval. Annual data on number of criminal commits and major offenses and mortality and overdose data on opioid usage were ingested and displayed using multiple descriptive charts and basic NLP. The clustering algorithm indicated that when using five clusters, counties in the east of Virginia are more dissimilar to each other, than counties in the west. The farther west, the more socioeconomically homogenous counties become, which may explain why counties in the west have greater rates of opioid overdose than in the east which involve more recreational use of non-prescription drugs. The dashboards indicated that between 2011 and 2017, the majority of crimes associated with heavy-use of drugs included Larceny/Fraud, Drug Sales, Assault, Burglary, Drug Possession, and Sexual Assault. Filtering by year, county, and offense, allowed for very focused analysis at the county level.

Conclusions

Data science methods using geospatial analytics, unsupervised machine learning, and leverage of noSQL databases for unstructured data, offer powerful and inexpensive ways for local officials to develop their own opioid surveillance system. Our approach of using clustering algorithms could be advanced by including several dozen socioeconomic features, tied to a potential risk score that the group was considering calculating. Further, as the team became more familiar with the data, they considered building a supervised machine learning to not only predict overdoses in each county, but more so, to extract from the model which features would be most predictive county-to-county. Next, because of the fast-paced nature of an overnight hackathon, a variety of open source applications were used to build solutions quickly. The team recommends generating a single architecture that would seamlessly tie together Python, R Shiny, and ElasticSearch/Kibana into one system. Ultimately, the goal of the entire prototype is to ingest and update the models with real-time data dispatched by police, public health, emergency departments, and medical examiners.

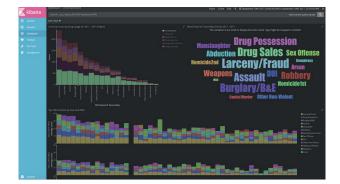




ISDS Annual Conference Proceedings 2018. This is an Open Access article distributed under the terms of the Creative Commons Attribution. Noncommercial 3.0 Unported License (http://creativecommons.org/licenses/by-nc/3.0/), permitting all non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

ISDS 2018 Conference Abstracts





Keywords

clustering; opioid; machine learning; elasticsearch; data science

Acknowledgments

The Booz Allen Opioid Hackathon Team was led by Jessica Bonnie, and consisted of Richard Latham, Jung Byun, Daewoo Chong, Joshua Wei, and Catherine Ordun.

References

https://data.virginia.gov/datathon-2017/ https://vimeo.com/236131006?ref=tw-share https://vimeo.com/236131182?ref=tw-share

*Jung Byun E-mail: smiwjung@gmail.com



ISDS Annual Conference Proceedings 2018. This is an Open Access article distributed under the terms of the Creative Commons Attribution. Noncommercial 3.0 Unported License (http://creativecommons.org/licenses/by-nc/3.0/), permitting all non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.