ABSTRACT
The improper treatment and disposal of hazardous waste can lead to disastrous effects on the environment and human health. The Resource Conservation and Recovery Act (RCRA), enacted in 1976, gives the Environmental Protection Agency (EPA) the authority to regulate hazardous waste “from cradle to grave.” To enforce these regulations, the EPA regularly conducts inspections of facilities that handle hazardous materials. The current inspection process is heuristic, based on prior knowledge or political priorities. Going forward, the EPA wants to adopt a data driven approach to investigation targeting, using historical inspection data to predict the risk of severe violations. Using data on reporting, monitoring, and enforcement, we develop and evaluate predictive models to identify likely violators. These scoring models are weighted by various criteria, including the potential magnitude of the violations, the environmental and public health impact of violations, and the likely outcome of an enforcement action. As a result, the EPA will be able to rank potential violators, better allocate inspection resources, and maximize the impact of each investigation to keep America’s air and water clean. This project will serve as an example for how the EPA can use predictive analytics in the future.

1. THE PROBLEM
The improper treatment and disposal of hazardous waste leads to disastrous effects on the environment and human health. For example, Hooker Chemical Company dumped 22,000 tons of dioxin and pesticides into Love Canal in New York, contaminating groundwater and leading to miscarriages, birth defects, and mental disabilities. To prevent the recurrence of such events, the Resource Conservation and Recovery Act (RCRA) was enacted in 1976, giving the Environmental Protection Agency (EPA) the authority to regulate hazardous waste “from cradle to grave.” The EPA can inspect a RCRA facility at any time to ensure its compliance; however, these inspections are time intensive. Inspections can take a week or more, meaning that with a force of 150 inspectors, it would take more than 50 years to inspect all 400,000 active facilities. Realistically, the EPA can conduct only 1,500 inspections per year. The EPA faces the challenge of prioritizing these inspections in order to catch potential violators, thereby preventing tragedies like the Love Canal incident from happening ever again.

2. THE CURRENT APPROACH
Each year the EPA produces a list of facilities to inspect. These facilities are chosen heuristically, based on prior knowledge or political priorities. The EPA maintains data on facilities across a wide variety of siloed regulatory programs, but currently has no system for effectively sharing data. The EPA would like to allocate its limited resources intelligently by taking a more data driven approach to prioritizing inspections.

3. OUR APPROACH
We frame the task of prioritizing inspections as a machine learning problem. Leveraging data collected by the EPA over the last 15 years, we build a model that predicts the probability that a facility will violate regulatory standards under RCRA. This is the first attempt at prioritizing inspections using machine learning and predictive analytics. By improving regulatory efficiency, our model has the potential to catch catastrophic chemical spills before they happen.

4. DATA SOURCES
Although a given facility may be monitored under multiple programs, RCRA inspectors have access only to the data from their own program. We consolidate these resources, using information across programs to categorize facilities. The primary data sources we use are listed below.

4.1 RCRAInfo
Characterizes facility status, regulated activities, and compliance histories on the generation of hazardous waste from large quantity generators and on waste management practices from treatment, storage, and disposal facilities. RCRAInfo contains data on 800,000 facilities, 450,000 of which remain active in 2015. About 250,000 facilities have ever had an inspection, comprising approximately 850,000 inspections in total.

4.2 Biennial Reporting System (BRS)
Information regarding the generation, management, and final disposition of hazardous wastes regulated under RCRA for odd numbered years. It includes over 690,000 unique facilities that shipped waste between 2001 and 2013, comprising over 9 million total shipment activities.

4.3 Integrated Compliance Information System (ICIS)
Incorporates federal enforcement and compliance (FE&C) case data on 150,000 facilities; discharge monitoring reports (DMR) from 330,000 facilities with National Pollutant Discharge and Elimination System (NPDES) permits; and pollutant monitoring information for 200,000 facilities monitored under the Clean Air Act (CAA).

5. EVALUATION
The EPA only has the resources to inspect about 1,500 facilities every year. Therefore we want to maximize the probability that every inspection that we suggest leads to a found violation. On a test set of approximately 30,000 inspections, we choose the model that maximizes precision on the top 5% (ranked by predicted probability of violation), which corresponds to 1,500 inspections. We use data on inspections performed by both state agencies and the EPA to categorize inspections, but evaluate model performance only on inspections performed by the EPA to mimic the true use case.

6. PRELIMINARY RESULTS
We are currently implementing a number of binary classifiers to predict violations as a function of facility, state, and industry characteristics. We utilize temporal cross validation to ensure that we are correctly estimating our model’s performance on future inspections and not “peeking ahead.” Further, we vary the time window in which we train our model to determine the optimal amount of history to include to make the best predictions on the future.

Our current best performing model is a random forest, with precision in the top 5% that is 2.2 times better than the baseline hit rate in the test set.

7. IMPLEMENTATION
We are working with the EPA to design a field experiment to test the predictive power of the model. We will divide the discretionary inspections into three categories – 100 facilities in the top 1% of our model, 100 facilities in the top 10% of predicted probabilities from our model, and 100 facilities at random from the list – and compare the “hit rate” among all groups. The EPA will select the remaining facilities to inspect in their normal fashion. This will allow us to compare performance at the top 1% and 10% of our model to randomly selected facilities and to the baseline of EPA inspections.

Our model uses insights from historical inspection data, combines various data sources, and reduces manual cross examinations of facilities. The EPA anticipates that this project will serve as a prototype for predictive analytics projects in other regulatory programs. A similar approach can empower government agencies to utilize their data to the greatest possible advantage.

8. ACKNOWLEDGMENTS
This work was done as part of the Eric & Wendy Schmidt Data Science for Social Good Fellowship through the University of Chicago.