Summary of HackEbola (with Data!)

Motivation and Objective

Many humanitarian organizations have been collecting data regarding the current Ebola outbreak in West Africa. These efforts include collecting information, ranging from socioeconomic indicators to types of infrastructure, about various administrative regions especially within the three hardest hit countries: Liberia, Guinea, and Sierra Leone. Combined with time series of Ebola cases and deaths for each of these administrative regions, these data have the potential to explain what factors may be affecting the outbreak—which in turn may result in actionable insights.

The objective of HackEbola (with Data!) was precisely to address the question:

What factors are affecting the regional and temporal evolution of the Ebola epidemic in West Africa?

Specifically, teams were tasked to:

1. Train a model / mine associations between the time-series and the covariates for the case data through Oct. 1.
2. Test that model / those associations on data from Oct. 1 - Nov. 20
3. Report (a) how well the model performed (prediction accuracy) as well as (b) the key factors in the model. Discuss any factors you think might assist in making decisions to contain the outbreak and limitations due to the models and the data.

The hackathon format created a focused opportunity for talented data scientists to devote time to this important objective. It also allowed for many model variations to be tested quickly. Harvard and UMass Amherst both hosted local events and many others participated remotely. qDatum provided the data management. In all, there were 2395 downloads from 189 unique users. Of those, 14 teams submitted their analyses.

In this document, we summarize both the technical outcomes of this hackathon as well as discuss how to leverage data hackathons for similar events in the future.

The Hackathon Format

The idea behind a hackathon is to get a group of people together to work on a challenge in a shared space over a short (usually a day or weekend) period of time. Working in a shared space creates a shared energy, and by limiting the duration of the event, people can usually commit to an intense period of focused work on the problem. The participants in HackEbola (with Data!) could immediately share issues and insights with other participants in the same room, creating energy and accelerating progress. For

online collaboration and sharing of data and analysis, we used the following websites and platforms: Challenge Post, qDatum, and GitHub, as well as Google docs and Dropbox.

Such a format was excellent for testing a variety of modelling assumptions quickly. Different teams focused on different ways of summarizing the time series of cases and deaths; one could take the results that were consistent across these assumptions for closer scrutiny. The format also provided an excellent opportunity to engage data scientists at all levels of experience and raise awareness around the challenges associated with assisting on these kinds of issues. Several teams reported that they had learned valuable lessons in working with real-world data on real-world problems.

HackEbola (with Data!) was focused on producing insights; this non-competitive atmosphere encouraged participants to post data-cleanup and results in progress for others to use and verify. The non-competitive nature of the event also allowed participants to focus on more general questions like *how* and *what* rather than only on prediction, which are crucial in data exploration. We had a fantastic data management team (qDatum.io) that hosted the data and an exceptional technical staff with knowledge about both Ebola, disease modeling and data analysis. The technical staff helped guide participants with computational but not epidemiological knowledge.

At the same time, because of the short nature of the event, there was little time for outside research and remedying irregularities discovered in the data. Teams came in ready to apply advanced data science expertise on a specific computational challenge; as data -- but not domain -- experts they were not necessarily acquainted with alternate data sources. Even when alternate sources were available, there was not necessarily time within the span of a weekend to obtain and process them. *Future challenges should work with relevant organizations to ensure that high quality, vetted data is available* to address the specific objectives provided.

We opened the Harvard hackathon with a presentation by a representative of the American Red Cross. In addition, observers from the Health Ministry of Haiti attended the Harvard hackathon; in addition to the technical outcomes with potential or indirect relevance to emergency preparedness plans, they also found it useful to see how such an event could be used to quickly derive insights from complex data.

Technical Outcomes

Data

Most participants focused on publicly available datasets from the Humanitarian Data Exchange and the Open Humanitarian Data Repository. These included sub-regional time series of Ebola cases and deaths (34 regions across Liberia, Sierra Leone, and Guinea). These were pre-processed by qDatum.io to have consistent administrative region identifiers across all of the tables. These data are available at:

http://www.qdatum.io/public-sources

Several teams shared processed versions of these data, listed in the Appendix.

Strategy

To assess what factors may be explaining inter-regional variation, most teams took the following general approach:

- **Defining Dependent Variables.** Most teams processed the time-series in each region into a set of summary statistics, such as total cases, total fatalities, case fatality rate (CFR), transmission rate (TR), cases per capita, and fatalities per capita. A variety of assumptions were used to derive these summary statistics: simple counts, exponential and logistic models, and variations on SIR models.
- **Collecting Independent Variables**. While most teams focused on a standard of regional indicators, some teams built additional variables to model regional connectivity and the effect of recent civil wars.
- **Analysis**. Teams used a variety of methods to compare the dependent and independent variables, including correlation computations, generalized linear regression, and other machine learning approaches.

Team [8] incorporated a model of movement patterns into the standard SIR model using the number of roads between districts. Though the model had high prediction errors, forming the model in this way might assist with future efforts. More generally, some teams also investigated other factors, such as the change in food prices in the region and mapping accessibility to current Ebola treatment units (ETUs).

Main Findings

As published in many sources, all teams found that standard SIR models could fit national-level data well. However, data at subnational levels contained many more irregularities. It is important to note that many teams did not find anything significant in the data, and even those trends that were statistically significant are suspect due to the biases in the data (more details in Limitations and Discussion section).

The most consistently reported potential covariate was age and country. Many teams [1, 2, 3, 4, 5] noted that the age distribution seemed to have an impact on the growth rate: rates were higher if the number of adults (20-60 years) was higher and lower among populations with more children (0-9 years) or seniors (over 60 years). [1, 3, 7] found that the country/latitude was one of the most dominant effects when predicting the growth rate, suggesting important country-specific distinctions.

When regressed individually, several teams [2, 4, 6, 7] found that urbanicity and level of education seemed to be positively correlated with case and fatality rates; most likely

these variables may be a proxy for populations that are more likely to be better reported. Similarly, teams found increased case and fatality rates for higher floor quality, cell phones, and flush toilets (and reduced rates for poor floors and poor toilets). While other correlations from [7] such as bad water and small house make sense, it seems that these correlation analyses largely discover that more well-off areas have better reporting.

Teams that performed combined regressions also found that the urbanicity and education variables had the strongest effects. [1, 3] found that higher education had a protective effect, as did electricity. The Poisson regression employed by [1] predicted observed cases after Oct. 1 with 99.8% accuracy; they later presented a polished version of their results to the WHO. However, teasing apart collinearities and confounds is challenging; for example, [1] also found that having tap water had a negative effect.

Additional Directions

Food Prices. [7, 9, 10] also looked at the changes in food prices during the outbreak. Teams also found that food prices, where available, had stayed mostly stable through the outbreak. This was corroborated by [11], although other sources [12] have reported increasing prices (as seen by [13]). However, as seen in the report of [10]:

https://www.dropbox.com/sh/7425tmisb2tyr6e/AACHNkbaZ-zkoyoSL3WcfJTYa?dl=0

The changes in prices during 2014 are within the already broad price variability due to other factors in the region.

Visualizations and Geography. Finally, a set of teams [8, 13] focused on exploring the data through visualizations and geographic analysis. In particular, [14] created a visualization showing locations that were within an hour of an ETU:

http://challengepost.com/software/etu-location-analysis

Limitations and Discussion

In trying to model the Ebola outbreak at a sub-national level, teams were fundamentally limited by the quality of the data:

- **Irregularities in Case Data** All participants reported many irregularities in the case data. Case data had many jumps reflecting the fact that the data indicated when cases were reported, not when cases occurred. In some regions, the number of deaths exceeded the number of cases—which could be explained by movement between regions or gaps in case reporting. Summing cases across regions did not produce numbers that matched national-level statistics. Cases were reported by five different sources with widely varying values; participants had to make difficult choices about which sources to use or how to combine

sources. Different time periods had data at different levels of regional and temporal granularity.

- **Granularity of the Data** Because of the many irregularities in the case data, most teams focused on deriving a few summary statistics of the sub-national time series (in particular, it was impossible to ascertain whether a spike or dip in cases was due to reporting or the effect of an intervention). Thus, the multi-month time series were collapsed into 34 regional data points.
- **Time of Covariate Collection, Missingness** Teams focused on food prices noted that prices were often missing during the outbreak, and the dates when ETUs were opened or closed were also often missing. Covariates (socioeconomic status, education, etc.) were all collected prior to the outbreak, some as early as 2007.

Thus, the results above must be treated with extreme caution. In a detailed analysis after the event, [15] showed how while models could be trained to fit the national-level data well, potentially artifactual variability in the case count trends in the subregions made it challenging to draw conclusions. In particular, small changes in training choices -- type of interpolation, choice of smoothing technique, and whether to train on incidence of new cases or cumulative counts -- resulted in large differences in the parameters. Teams found that changing some of the assumptions used to compute the dependent variables resulted in spurious correlations ([9] provided several examples)—for example, a positive correlation between education and case fatality rates. These effects are in addition to the challenges due to confounds (more urban areas are likely to have better reporting) and collinearities (more urban areas also have more educated citizens).

Conclusions

HackEbola (with Data!) created a considerable interest in the data science community to help address an important need: understanding how various regional factors might be affecting the temporal and geographic evolution of the epidemic. Even created as a “small” event, it resulted in over 300 registrations, of which 189 downloaded data. The event proved that the hackathon format was an excellent format for testing many approaches quickly: each team used different assumptions when wrangling and modeling the data. However, in the end, teams were still limited by the quality of the data. In the future, “deep dive” events such as these might benefit from additional steps to check whether the data are amenable for such fine-grained analysis.

Acknowledgments

We would like to thank the crisis experts who consulted teams during the Hackathon: Colby Wilkason, Dr. Marie Ghislaine Adrien, and Christine Duchatellier Fowler.

Staff

F. Doshi-Velez, Harvard Organizer. D. Lasry, Y.E. Marshall, R. Gong, G. Harling, Q. Wang, Harvard Technical Staff. M. Kapp, J. Ming, A. B. Bind, Harvard Site Staff. N. Reich, K. Gourgoulias, E. Ramos, Amherst Organizers. E. Leschem, I. Maltz, Data Management Team.

Participants

[1] M. A. Testa, M. Su, S. Konate, J. Torres, and E. Savoia.
[2] G. Khimulya, D. Sampas, and R. G. Cutbill.
[3] Y. Gurmu, G. Harling, S. Vardhanabhuti, S. Chin
[4] D. Schoenfeld.
[5] M. Duanmu.
[6] C. Yang and S. Okuda.
[7] X. An, Y. Chang, J. Chen, L. Hou, and X. Li.
[8] Y.E. Marshall, S. Mahimkar, R. A. Zelaya, T. Fussell, and A. Low.
[9] G. Bridgman, Q. Truong, A. Amin, A. Jacobovits, and E. Jospe.
[10] C. Valentine and D. Roscoe.
[13] S. Wang and W. Liu.
[14] A. Low.
[15] G. Sabran, K. Altenburger, K. Fodouop, S. Wang, and M. Andere.
[16] J. Winget.

References

[11] http://www.theigc.org/news-item/the-economic-impact-of-ebola-october-2014-report/
[12] http://documents.wfp.org/stellent/groups/public/documents/ena/wfp268882.pdf

Appendix: Processed Data Sets

Several teams created processed data sets for other teams to use.

[4] created a cleaned up set of data containing the date of the start of the epidemic in the region relative to the start of the epidemic, duration of the epidemic in that location, the total number of cases and the total number of deaths. He also estimated a parameter beta which is the log of number of cases over the duration (estimates the rate of spread):

https://www.dropbox.com/sh/kniy9s8rcuvxu1b/AACpzPnkpNu7IrQMdyTBRL_Va?dl=0

[7] and [3] also shared data sets with gaps interpolated and files merged with indicators:

https://www.dropbox.com/sh/ief4x9yshmd6619/AADJgB49a_xQcwl9aYynv2Pua?dl=0

https://drive.google.com/folderview?id=0B915sMG77RJYblBTS2hhdURaMms&usp=sharing

[2] estimated the number of treatment beds per region and shared the results:

https://www.dropbox.com/sh/qtibq5tv2yful8l/AADP4tuhvD1HRYG3bLAorM_la?dl=0