Ultimate Consequences:
The Significance of Deaths in Bolivian Political Conflict, 1982 to Present
INTELLECTUAL MERIT ADDENDUM ON DATA SCIENCE

The Ultimate Consequences database is the heart of this research project. Due to the number of lethal events in the study period, the dataset is both large enough for quantitative research that analyzes patterns and small enough for qualitative, journalistic, and historical examination of the individual deaths involved. To serve these multiple purposes, we are including detailed information about every death, and will make that data available to social scientists, historians, journalists, and human rights advocates. Verifiable documentation of these lethal events, transparency in our coding decisions and analysis, and accessibility of information are all core to our research purpose.

Data Collection, Search Strategies, and Documentation

This project follows in traditions of datasets designed to analyze the frequency of political conflict, and to bring accountability for acts of violence. Many conflict-related datasets were constructed using passive forms of monitoring, such as the World Labor Group database (Silver 2003, 181–97), which enumerated all labor strikes recorded by two major newspapers, and Airwars' (n.d.) monitoring of news coverage of deaths in air strikes. These projects sought to rapidly capture many known cases with roughly constant likelihood over time. When reports conflict, Airwars grades the quality of reporting and documents the minimum number of deaths reported. Compilations by human rights monitors, by contrast, gather more detailed information on causality and responsibility and are motivated by holding states and armed nonstate actors to account.

Our project draws on journalistic, advocacy, and scholarly sources to comprehensively document all deaths in political conflict, including those not readily categorizable as human rights violations. Not only does this approach allow more complete coverage of the topic, but it also allows us to shed light on which kinds of deaths are documented or ignored by different sources. We began the project with journalistic compilations of conflict deaths, and human rights reports by domestic and international human rights bodies. Our iterative research strategy then sought out additional information from two Bolivian NGOs that archive and summarize news about social movement activity, archives of the Bolivian news media, and a range of other sources. The multivariable design of our database, as well as internal rating system for the quality of our descriptions, drove our search for additional sources on each lethal event.

The interview phase of our research, which will begin in 2022, is grounded in the practices of oral history. Oral historians use consented interviews, transcripts reviewed by interviewees, long-term archiving, and the option of embargoing interviews (Oral History Association 2009) to create a lasting and reusable record of the past as understood by participants in it. We believe that this framework is the right way to record an on-the-record account of the significance of deaths in political conflict and the social constraints on political violence.

Transparency and Reproducibility of Our Analysis

The research using the database takes place in two steps: drawing on documentary and other sources to name and code the circumstances of each of the deaths, and analyzing patterns in loss and political violence based on information in the database. The task of providing transparency and reproducibility differs between these two steps. Regarding the first step, we have compiled

bibliographic information about sources, quotes from those sources, and comments about any complications in coding into the database and a narrative summary of "complex cases" in the dataset. Coding decisions are made according to a codebook, which we also share. Criteria for including or excluding deaths from database are also listed in the codebook, and deaths that we have chosen to exclude (for example, due to inaccurate reports and non-conflict accidents) are archived within the database. Each entry in our database currently lists sources consulted, and we are compiling fully-cited narrative descriptions for each lethal event. We will make this sourcing transparent through by posting these cited narratives, maintaining a public Zotero library of referenced sources, and making scanned documents available to other researchers. Coupling our dataset with its sources allows historical and social scientific researchers to validate our conclusions and build their own scholarship on the source materials we provide.

When we analyze the dataset, we will ground our choices in clearly defined criteria, using categorical information in the database wherever possible. By specifying our search criteria for relevant cases, explaining exceptions, and embedding our analysis techniques in R scripts that use the database, we document our choices and create tools that can automatically update results when additional cases are uncovered, errors are corrected, or new information is brought to light. Archiving and versioning (Klump et al. 2020) of the database and these tools allows other researchers to both reproduce our results, and test their robustness against different choices in coding or analysis. Publicly released versions of the dataset will be archived on GitHub, assigned doi locators, and downloadable into R via datastorr (Falster et al. 2019). Our analysis scripts will be shared as an R package on Github, with their results presented in R Markdown and the underlying code available for others researchers' use, critique, and modification.

Sharing Our Data with Other Researchers

Our research process is built around using publicly accessible tools to collaborate within our research team and to share our data, analysis methods, and results. During the funding period, both our data-coding and tool-building work will be linked to sharing our data in a fully documented form via scholarly repositories. This includes an initial release of our dataset in a public version. (All releases will include only those variables that have been checked for accuracy, and refined into categories as specified in the codebook, with as much completeness as can be derived from available sources.) We have chosen formats and data repositories in the interest of open science, with particular consideration of the FAIR (Findable, Accessible, Interoperable and Re-usable) principles (Force11 2014).

In developing the project so far, we have created a variety of ways to summarize and visualize the data, including month-by-month heatmaps of events; summary tables and histograms by year, presidency, protest domain, and state role in deaths; and charts illustrating how deaths are distributed by geographic region. (Figures 1 and 2 in the proposal are two examples.) By scripting each of these visualization tools in R, and making them web-accessibly through an R Shiny interface, we will allow other researchers and the public to see these overviews of the dataset. Further, by including querying tools, we will allow users to browse lists of deaths meeting particular criteria (such as deaths perpetrated by state security forces, or related to peasant movements), or apply any of these visualization tools to selected subsets of the database.

Finally, our open data strategy benefits from participation in, and training from, a community of data scholars at Vanderbilt's Data Science Institute and Center for Digital Humanities.

References Cited

Airwars. n.d. "Methodology." Airwars. Accessed July 21, 2021. https://airwars.org/about/methodology/.

Falster, Daniel S, Richard G FitzJohn, Matthew W Pennell, and William K Cornwell. 2019. "Datastorr: A Workflow and Package for Delivering Successive Versions of 'evolving Data' Directly into R." *GigaScience* 8 (5). https://doi.org/10.1093/gigascience/giz035.

Force11. 2014. "Guiding Principles for Findable, Accessible, Interoperable and Re-usable Data Publishing, Version B1. 0."

Klump, Jens, Lesley Wyborn, Mingfang Wu, Robert Downs, Ari Asmi, Gerry Ryder, and Julia Martin. 2020. "Principles and Best Practices in Data Versioning for All Data Sets Big and Small." *Revised Report of the Research Data Alliance Data Versioning Working Group. Doi* 10.

Oral History Association. 2009. "Principles and Best Practices." https://www.oralhistory.org/about/principles-and-practices-revised-2009/.

Silver, Beverly. 2003. *Forces of Labor: Workers' Movements and Globalization since 1870*. Cambridge: Cambridge University Press.