Crowdsourcing the Measurement of Conflict Data

Vito D'Orazio[*], Michael Kenwick[†], Matthew Lane[‡], Glenn Palmer[§], David Reitter[¶]

[*]Harvard University, Institute for Quantitative Social Science
[†]Pennsylvania State University, Department of Political Science
[‡]Pennsylvania State University, Department of Political Science
[§]Pennsylvania State University, Department of Political Science
[¶]Pennsylvania State University, Dept of Information Science and Technology

Abstract

Much of the data used to measure conflict is extracted from news reports. This is typically accomplished using either expert coders to quantify the relevant information or machine coders to automatically extract data from documents. Although expert coding is costly, it produces quality data. Machine coding is fast and inexpensive, but the data are noisy. To diminish the severity of this tradeoff, we introduce a method for analyzing news documents that uses crowdsourcing, supplemented with computational approaches. The new method is tested on documents about Militarized Interstate Disputes, and its accuracy ranges between about 68 and 76 percent. This is shown to be a considerable improvement over automated coding, and to cost less and be much faster than expert coding.

Keywords: crowdsourcing, measurement, Bayesian networks, data collection, militarized interstate disputes, mechanical turk

Much of the data used in conflict studies is generated from the content analysis of news reports. Thus far, there have been two primary methods for conducting such analyses: using highly-trained, expert coders or using machines and automated processes. Both experts and machines have benefits as well as drawbacks for collecting conflict data. While expert coders are expensive and slow, they are capable of accurate, valid data collection and generally remain the gold standard in much of the discipline. Alternatively, machine coding can quickly classify vast quantities of information at very low cost. The data produced using these automated methods, however, tend to be noisy, limiting the ability to measure concepts of interest accurately. As a result, researchers are faced with the dilemma of either sacrificing cost for quality or sacrificing the quality of data for a reduction in time and cost.

More recently, crowdsourcing has emerged as an alternative to standard methods of data collection. In our application, crowdsourcing refers to a process in which a large group of non-experts is asked a series of questions as a means of gathering information on conflict events from news sources. Like machines, crowdsourcing is considered to be fast, cheap, and capable of collecting data on a large scale. Like experts, crowdsourcing utilizes human intelligence and benefits from the ability to summarize documents and understand the context in which events are taking place.

While crowdsourcing has been shown to be effective in many tasks that were previously believed to require expert knowledge (Benoit et al. 2014; Cooper et al. 2010; Kittur et al. 2008), it has not been applied to measurement in the field of conflict studies. Given the large number of data projects that exist in this field of research, and the even larger number of data projects that *could* exist if not for the costs of expert data collection, an accurate and efficient method of data collection could prove particularly valuable.

We propose and assess a method for coding conflict data that supplements crowdsourcing with computational approaches in an effort to produce accurate data efficiently

(henceforth referred to as the crowdsourcing method). Workers use an online platform to read a news story and answer a line of objective questions structured as a question tree. These questions are supplemented with meaningful information, such as possible international actors, extracted from the news documents using automated methods. Finally, these individual crowdsourced answers are aggregated via machine-learning methods and used to predict all components of the concept of interest.

In this study, crowdsourcing and machine coding methods are presented and compared as alternatives to expert coding and examined for their abilities to measure militarized interstate disputes (MIDs) (Palmer et al. 2015; Jones et al. 1996). These methods are assessed based on accuracy, financial cost, and time to completion. We choose the MID as our construct of interest since the data collection procedures for MIDs are widely known and can be replicated with reasonable accuracy. More importantly, MIDs are representative of many theoretical constructs in conflict research, and political science research more broadly, in that they are complex social processes that are classified with a defined set of coding rules and measured using primarily news reports.

We begin our analysis by having highly trained, expert coders read a set of 450 news documents to identify whether each contains information on a militarized dispute and, if so, which state was responsible for initiating hostilities, which state was the target, and which specific type of militarized action took place according to the MID coding rules. well in terms of intercoder reliability, but that disagreement over the correct classification of militarized disputes occurs with some frequency, demonstrating that the MID is a difficult construct to measure. Each document is then coded using Tabari, open-source software for generating event data from text documents. The use of expert coding provides us with the "correct" classification of each document, as well as an estimated cost for obtaining this coding.[1] The use of Tabari provides us with a benchmark against

[1]For complex concepts, the "correct" coding may sometimes be subjective. Thus, even though we speak as though expert coding has perfect accuracy, it does not.

which the accuracy of crowdsourcing can be assessed. We then outline the crowdsourcing infrastructure developed for this study.

In brief, we use worker recruitment services to obtain six to ten workers to classify each news story. We provide these workers with a questionnaire designed to obtain information that can be used to identify which documents contain information about MIDs and the relevant features of those disputes. Building on this basic infrastructure, we further combine standard crowdsourcing techniques with computational methods, specifically named entity recognition and Bayesian network aggregation, to improve the quality of the data collected. The results indicate crowdsourcing as an increasingly viable option for data collection, as the crowd identifies the pertinent aspects of a news document with significantly greater accuracy than machine coding. Additionally, the crowdsourcing method was significantly cheaper and faster, both in montary cost and time to completion, than comparable expert coding.

This project makes several important contributions to data collection efforts, particularly in the field of conflict studies, but also generalizable to many efforts in political and social science. First, we provide a direct comparison of expert coding, crowdsourcing, and automated data collection methods. Despite the increasing commonality of these latter two approaches, very few studies have endeavored to compare the accuracy of these methods.

Second, we demonstrate the applicability of crowdsourcing techniques to data collection tasks that are conventionally thought to require highly trained experts. Even as crowdsourcing techniques quickly gain traction in numerous research fields, applications in conflict studies and political science are comparatively rare, likely because these methods are often thought to be incapable of supplanting traditional data collection techniques. We hope to dispel these notions by providing consistent evidence that crowdsourcing outperforms automated techniques and can be used to approximate expert classifications

at much lower cost.

Finally, we extend existing crowdsourcing techniques by integrating machine-learning techniques, specifically named entity recognition and error-corrected voting procedures. We expect that each of these techniques will provide a useful means of improving accuracy of crowdsourced classifications in a variety of applications. More broadly, such integration demonstrates that human judgements supplemented with computational technologies provide higher quality than machine coders at lower costs in time and money than expert coders.

Crowdsourcing and Data Collection

Measuring social concepts is a fundamental task in the social sciences (Krippendorff 2013; Goertz 2006; Sartori 1984; Zeller and Carmines 1980). Traditionally, this has been a resource-intensive undertaking in which scholars and their research assistants search and read a vast quantity of documents toward locating events of interest that satisfy their operational definition. Some of the largest projects in political science have devoted immense resources to this task, including the Polity and Correlates of War projects. Certainly, these resources have been put to good use. Nevertheless, given recent technological advances and the increasing prevalence of real-time, big data analyses, costly and time-intensive approaches to data collection are becoming increasingly antiquated.

Machine coding is a more recent, alternative method for analyzing the content of news reports (Schrodt 1994; King and Lowe 2003; O'Brien 2010). While it is capable of producing data in near-real-time at marginally zero cost (Schrodt 2010), the resulting data have yet to be shown to be reliable measures of concepts of interest. Such "noisy" output is a consequence of construction; machine coded data commonly map subject and predicate phrases in a news document to a dictionary of terms that are associated with a particular concept (O'Connor et al. 2013). This process codes events in isolation from

other events and without any knowledge of the broader political atmosphere.

Crowdsourcing, which uses distributed labor from a community of individuals, is an alternative approach that addresses the primary drawbacks of the expert and machine coded methods. Crowdsourcing has been widely used to find subjects for experimental studies in the behavioral and social sciences (Berinsky et al. 2012; Paolacci and Chandler 2014; Tingley and Tomz 2014). In our application, we are not recruiting subjects for experimentation, but recruiting non-expert workers to perform classification tasks. The community of workers may be accessed through an online platform, such as Amazon's Mechanical Turk (AMT), or it may be more defined, such as the community of undergraduates at a college or university.

Research has shown crowdsourcing to be an effective tool for solving problems that were previously thought to require substantive expertise. For example, many natural language processing tasks are amenable to the use of crowdsourcing in place of trained experts and, in some instances, the data generated is better than that of experts (Berinsky et al. 2012; Chandler et al. 2014; Goodman et al. 2013; Zhai et al. 2013; Paolacci et al. 2010; Sabou et al. 2012; Snow et al. 2008). Most closely related to our application, Benoit et al. (2014) and Honaker et al. (2013) have used crowdsourcing to measure political ideology and regime types, respectively. Each of these studies were successful in using crowdsourcing to provide quality data efficiently.

These findings suggest that crowdsourcing is likely to be cheaper and faster than expert coding and more accurate than machine coding. However, no study has directly compared these methods for data collection, nor demonstrated their utility in the field of conflict studies. Additionally, few, if any, studies have explored the potential for crowdsourcing to accurately and consistently code subjective events from observational reports, such as news documents. As a result, we do not know if crowdsourcing is cheap enough, fast enough, or accurate enough for data collection projects in conflict research,

which commonly relies on data coded from such sources. This study quantifies these comparisons for coding MIDs, a commonly used measure of international conflict.

Coding The Militarized Interstate Dispute

Militarized Interstate Disputes are "united historical cases of conflict in which the threat, display or use of military force short of war by one member state is explicitly directed towards the government, official representatives, official forces, property, or territory of another state" (Jones et al. 1996). Each MID is comprised of at least one and potentially hundreds of Militarized Interstate Incidents (MIIs). An MII "is defined as a single military action involving an explicit threat, display, or use of force by one system member state towards another system member state" (Jones et al. 1996), and are frequently described in a single news report. The traditional method for collecting data on MIIs, as undertaken by the MID3 and MID4 projects (Ghosn et al. 2004; Palmer et al. 2015), is to query LexisNexis and retrieve a document set, manually code the retrieved set with the assistance of highly-trained research assistants, and then for lead researchers to verify the coding—a process common to many subfields.

As our first step, expert coders from the MID4 project replicated this process by individually coding 450 documents of MID-related news stories. To obtain this sample, we drew upon documents identified by the MID4 project as potentially containing information about MIDs. These were gathered by querying LexisNexis using terms related to militarized disputes and then using automated techniques to obtain the subset of stories most likely to be related to MIDs (D'Orazio et al. 2014). We then binned these stories by year (2007, 2009, and 2010), and manually identified those that likely contained information on MIDs and those that likely did not. For each year, we then randomly sampled 100 stories from the positive bin and 50 stories from the negative bin, providing a total of 450 documents for coding. Next, three experts in MID data collection coded each of

the 450 news documents, independently of each other.

The "Gold Standard"

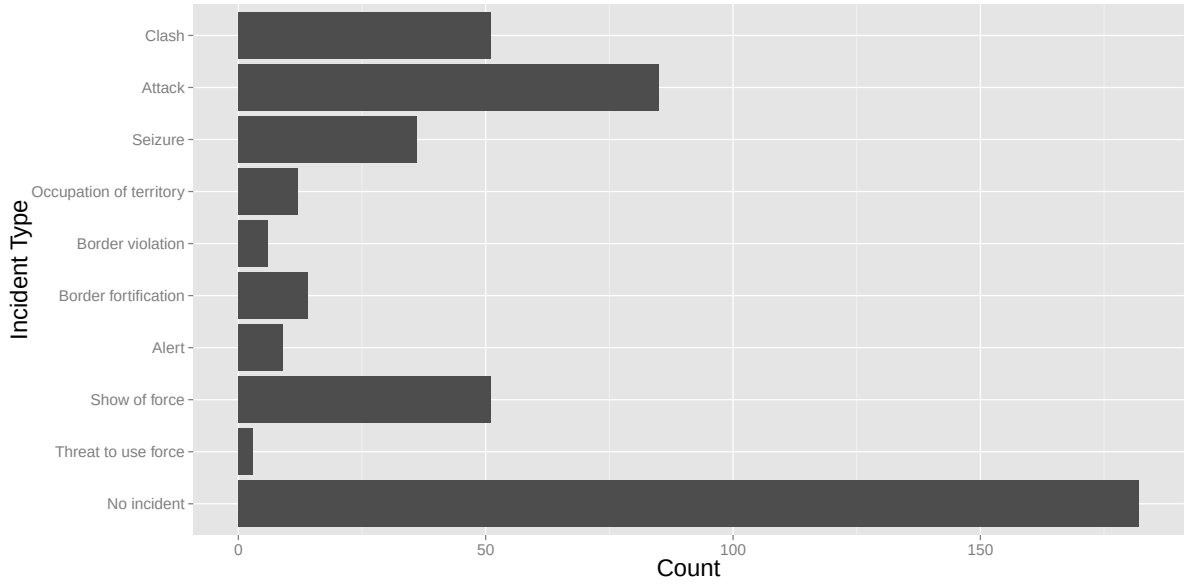
Having established a set of expert codings, we used consensus among the experts and resolved instance of disagreement to generate a"gold standard" classification for assessing the accuracy of the crowdsourcing and machine coding methods. More specifically, the three experts were reconvened *after* the initial coding, evaluated each news document, and mutually determined the "correct" classification for the attributes of interest.[2] The "gold standard" represents the true coding of each news document and is assumed to be inherently correct.[3] Creating the gold standard provides a necessary benchmark against which the accuracy of machine coding and crowdsourcing can be gauged.

The distribution of incident types in the gold standard is shown in Figure 1. About one-third of the documents are non-incidents, which is consistent with our stratified sampling of the MID4 project's potential incident set. For the documents that contain relevant MII information, the largest category is *attack*, which is consistent with the distribution of MIIs in MID4, as this is the most common type of incident (Palmer et al. 2015). *Clash* and *show of force* are the next most common, with just over fifty of each. *Seizures* are uncharacteristically high, and this resulted from a large number of documents in the sampled set pertaining to a 2007 Iranian capture of the United Kingdom's military personnel. There are very few *threats*, as is also the case in all MID data, and few *alerts*, *border fortifications*, *border violations*, and *occupation of territories*, all of which are consistent with the known distribution of MIIs (Palmer et al. 2015).

[2]A similar process was used in the original coding of the MID4 data. See Palmer et al. (2015) for full details.

[3]Note that in creating the gold standard experts did not employ the same questionnaire architecture used by the crowd, as discussed in future sections. Instead, the method of creating the gold standard represents simple consensus building, which we contend is often used in data collection projects involving small groups of expert coders. This approach was taken because it is straightforward and provides a benchmark of accuracy necessary for comparison with the incumbent method.

Figure 1: **Incident Distribution in Documents**



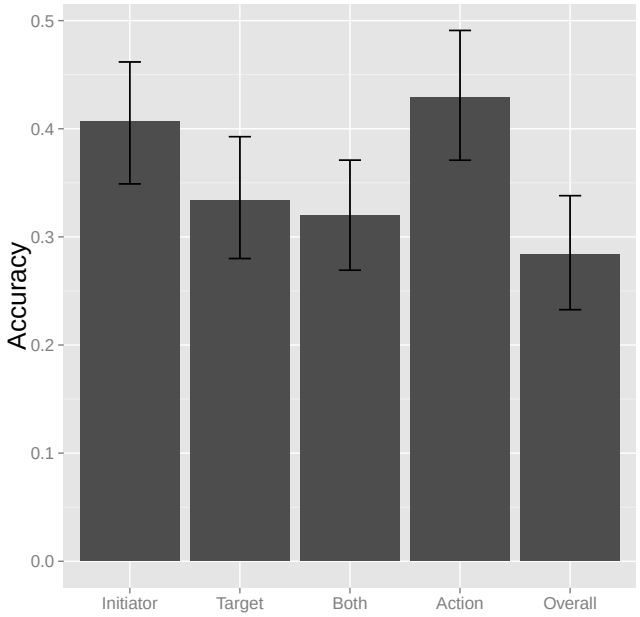
Analysis of Automated Machine Classification

Machine coding is an alternative to expert coding that is scalable, efficient, and can be virtually free of cost (King and Lowe 2003; O'Brien 2010; Schrodt and Gerner 2012). For these reasons, if the method proves accurate in our application, it is likely superior to the crowdsourcing approach which is not, at present, free of cost. We find, however, that the method is not accurate in this case, as its accuracy ranges from about 30 to 45 percent.

To machine code the documents, we used Tabari (Schrodt 2014) and the Conflict and Mediation Event Observations (Cameo) ontology (Schrodt 2012).[4] This produces a database of event data that detail which actor initiated an action, which actor was the target of this action, the CAMEO action type, and the date on which the event occurred. To discern the codeable MII attributes from this dataset, each CAMEO action type was mapped to a MID code using a deterministic set of rules. These rules pertain to both the actors and the actions.

[4]See the Supplemental Appendix for a complete description of the automated coding process.

Figure 2: **Tabari Results**



Note: Figure displays TABARI's accuracy for the 275 documents it could code. Accuracies are reported for initiator, target, both initiator and target, action, and for the combination of initiator, target, and action. 95 percent confidence intervals are calculated from 1000 bootstrapped samples of the data.

TABARI is not very good at coding MIIs. It can code only 275 of the 450 news stories, or about 61 percent.[5] The accuracy of TABARI for the 275 news stories that it could code is plotted in Figure 2. Of these 275 documents, 67 percent (183 documents) contained information about what the gold standard identified as MIIs; TABARI identified 63 percent of those MIIs as militarized incidents. Further, of these 183 MIIs, TABARI correctly identified 36 percent of the initiators, 25 percent of the targets, and 39 percent of the actions. Of the 92 documents that TABARI could code that did NOT contain information about MIIs, TABARI incorrectly identified 49 percent of them as containing information about MIIs. Clearly, utilizing TABARI to code complicated social phenomena such MIIs is not effective.

This application of machine coding has greater difficulty discerning the correct actors

[5]The other 175 documents received a *null* coding. A *null* can be produced if no CAMEO event occurred, if the document lacked the specific verbiage required for TABARI to recognize the CAMEO event, or because of technical issues such as formatting idiosyncrasies.

Table 1: **Tabari Actions**

		True Coding									
		Not MII	Threat to use force	Show of Force	Alert	Border fortification	Border Violation	Occupation of Territory	Seizure	Attack	Clash
Tabari	Null	91	1	15	3	9	4	5	20	18	9
	Not MII: Domestic	28	1	5	1	1		2	4	4	10
	Not MII: International	18	1	14				2	6	11	5
	Show of force	1		3	1						
	Alert				1						
	Mobilization				1	3				1	1
	Occupation of Territory						1	3		1	
	Seizure	9							5		1
	Attack	28		13	2	1			1	45	10
	Clash	7		1			1			5	15

Note: The "null" category pertains to documents where Tabari did not record any event occurring. The "domestic" category pertains to documents Tabari did not classify as taking place between two state actors. Empty cells are analogous to zeros and indicate that no documents were classified in a particular category.

than it does the correct action. On the full dataset, Tabari classified the initiator correctly 45 percent of the time and the target correctly 40 percent of the time. When *nulls* are dropped, the initiator accuracy is just 41 percent, while the target accuracy is 33 percent. Action accuracy, on the other hand, is 46 percent if the nulls are included, and 43 percent if nulls are removed. Finally, overall accuracy is 38 percent if nulls are included, and 28 percent if they are removed.

Table 1 provides a crosstab of the Tabari and the true action code for all documents. The numbers labeling each row and column in the table correspond to MII action levels. *Not MII: Domestic* indicates that Tabari coded both initiator and target as being from the same country and thus not an MII. *Not MII: International* indicates that Tabari coded an international event that is not an MII. *Null* indicates that Tabari did not produce any event data for that document.

The bulk of correct codings came from the 137 true negatives, the 45 true *attacks*, and the 15 true *clashes*.[6] The fact that most of the accuracy was due to true negatives, *attacks*, and *clashes* may be indicative of shortcomings in using CAMEO to classify types of international conflict. Specifically, CAMEO was designed as a codebook for analyzing mediation events, and as such there are events for which CAMEO does not provide action codes (Schrodt 2012; Schrodt and Gerner 2000). For example, all true *border fortifications* will be incorrect because CAMEO has no corresponding event code. Nonetheless, CAMEO is widely regarded as the most sophisticated event ontology, as evidenced by its use in Lockheed Martin's ICEWS Project (O'Brien 2010) and in the Open Event Data Alliance's Phoenix Pipeline (pho 2015).

Although this is a validity assessment of machine coded data, and therefore similar in spirit to studies such as Schrodt and Gerner (1994) and Bond et al. (1997), we differ from these early assessments of machine coding in that we do not evaluate the machine's ability to code events against a human's ability to code events. Rather, we evaluate our ability to automatically code MIIs against an MII "gold standard." To automatically code MIIs, we use a machine coder to extract events and then we map those events onto the MID event ontology. Thus, our results do not contradict the findings that say machine coders are at least as good as humans *in coding events*. Rather, these results demonstrate that coding an MII from a machine coded event dataset yields accuracies in the 30 to 45 percent range, which is not sufficiently high to be considered a viable option for use in many research projects.

[6]The true negatives include the 91 *nulls* that were not MIIs, the 28 *domestic* events that were not MIIs, and the 18 *international* events that were not MIIs.

Crowdsourcing

We describe a new method that supplements crowdsourcing with computational tools in an effort to obtain high accuracy *and* efficiency. The findings show that the crowdsourcing-based approach yields an actor accuracy of 76 percent, an action accuracy of 74 percent, and an overall accuracy of 68 percent. The method incurs some financial costs—about 0.62 that of experts—but is also much faster and opens up useful avenues for future research. In the following section we describe the general crowdsourcing architecture used for the coding of MIIs.

Crowdsourcing Architecture

Prior to initiating the questionnaire, natural language processing tools were used to extract metadata from each document, including the date, title, and named entities.[7] These metadata were piped into the questionnaire through an external web service, providing the workers computationally generated information about the document to help guide their decisions. The goal of this additional processing step was to limit user error by eliminating extraneous or unnecessary options and to streamline the overall process for the respondents.

Next, individuals read a news story and were asked a line of objective questions structured as a question tree.[8] That is, their answers were used to determine which question they were asked next. The set of responses were then compiled into the crowd's assessment of each document.

The questionnaire was hosted on Qualtrics, a web-based survey platform, and workers were recruited through Amazon's Mechanical Turk (AMT), a widely used crowdsourcing

[7]In this case named entities refer to relevant state names or individuals representing the state. The list of named entities are contained in Phil Schrodt's CountryInfox.txt file, found at `https://github.com/openeventdata/CountryInfo`.

[8]The questionnaire is available as part of the replication materials.

Figure 3: **MII Example Story**

Georgia accuses Russia of violating its air space
GRG-RUS

March 6, 2009

Two Russian helicopters violated Georgian airspace in the area of the Georgian-Abkhaz border on Thursday, the Georgian Interior Ministry has reported. "The Russian Mi-8 and Mi-24 helicopters flew over the villages of Tanmukhuri and Khurcha, Zugdidi region, where Interior Ministry troops are stationed, at 11:00 a.m. on March 5," says the report.
The Georgian ministry said Russian helicopters also flew over the Zugdidi region on March 4.
Interfax so far has no comments on this information from Russian military officials.

platform.[9] When a worker was completing the questionnaire for the first time, she was required to take a quality-control test.[10] We designed this test in accordance with the general types of error our project was attempting to minimize. More specifically, we attempted to minimize the occurrence of false negatives, or situations in which workers classified a story about a militarized incident as a non-event. The logic underlying this decision is that it is often easier to remove false positives (i.e. stories about non-events classified as MIIs) through additional verification phases than it is to recover false negatives that were screened out in earlier phases of the project. With this in mind, our test required the worker to read and correctly classify a story about a militarized clash between opposing military forces.[11]

[9] Amazon Mechanical Turk is a web service platform hosting many thousands of workers. We utilize this system because of its straightforward and flexible interface. Additionally, the size of the AMT worker population and typical task length means that, even if crowdsourcing becomes a widely utilized data-collection method, a significant increase in available tasks likely would not diminish the speed or efficiency of the process.

[10] To protect the anonymity of respondents, we do not collect any demographic data, such as ethnicity or socio-economic status. While it is perhaps conceivable that variation in such demographic markers leads to variation in the quality of data produced through crowdsourcing, such propositions are outside the scope of this project and, as such, we leave the testing of such propositions to future research.

[11] An additional or alternative approach at this phase would have been to provide the worker with a news story that *did not* contain information about a MII to protect against the well-known proclivity of workers to "over-classify," or produce an abundance of false positives. We elected not to take this approach, as doing so may have produced a sample of workers less likely to produce false positives, but more likely to produce false negatives, which was inconsistent with our project aims. In this way, future projects utilizing crowdsourcing for data collection may vary their screening process to align with the

After passing the test, the worker was given a set of simple instructions asking her to use only the information in the story and to focus only on actions taken by countries against other countries, not non-state actors. She was then prompted with a link to a story, hosted on an external server and an example of which is shown in Figure 3. The worker then proceeded through the questionnaire until completion, when she was given a unique code to submit to AMT for payment. Upon completion, the worker was paid fifty cents plus a twenty-five cent bonus if she was determined to have answered well.[12]

After reading the news story, workers were asked to extract the initiating and target actors. For example, a mockup of the questions pertaining to actors is shown below. Consistent with the example in Figure 3, the named entity recognition tools identified "Russia" and "Georgia" as countries identified in the news document and provided them as options for workers to select as answers to these questions. Had other countries been referred to in the text, these too would have been offered to the workers as choices.

What country or countries (or their military personnel) initiated the action? If civilians or non-state actors initiated an action, focus instead on the action a country took in response.

☐ Russia

☐ Georgia

☐ Other, please enter the name of the initiating country: [_____]

Who is the target of the action by the country/countries selected above? The target may be a country, its civilians, its territory, or a non-state actor operating within the country. Either way, please name the country that the action has been directed against.

☐ Russia

☐ Georgia

☐ Other, please enter the name of the target country: [_____]

---

aims of their specific project.

[12]That is, if the worker answers a random test question correctly, spends sufficient time on the questionnaire to convince us he or she is not randomly clicking, and writes at least three words when asked to summarize the document. Workers are made aware of this incentive structure prior to completing the questionnaire.

Following the actor questions, workers were asked a series of questions about the events that took place between these actors. The questionnaire proceeded from questions about the most hostile event types (uses of force) to the least hostile event types (threats to use force). Upon a categorization of action type, workers were presented with the MID project's operational definition and asked if their categorization was consistent with this definition. For example, if the worker coded the story as a show of force, as the story in Figure 3 should be coded, and coded the actors as Russia and Georgia, they would have been prompted with the following:

> **Russia engaged in a show of force against Georgia. A show of force is defined as a public demonstration by a state of its military forces intended to intimidate another state but not involving actual combat.**
>
> **Examples include non-routine military maneuvers and military exercises, naval patrols immediately outside the territorial waters of another state, and the intentional violation of another state's territorial waters or air space. Is this an accurate description of the event?**
>
> ☐ Yes
>
> ☐ No
>
> ☐ Yes, but the actors are reversed

This question is used as a final validation of the respondent's coding and as a means of providing respondents with an opportunity to holistically review their coding of an event. If respondents answered "No" to this final validation question, they were prompted to complete the questionnaire again until they achieved a coding that they felt accurately records their conceptualization of the event. We utilize responses only where the worker answered "Yes" or "Yes, but the actors are reversed" to this question.

Crowdsourcing Costs

In total, the questionnaire was completed 3,899 times for the sample of 446 stories.[13] The average number of responses per story ranged between 6 and 10, with an average of 8.47. Some individuals completed the questionnaire for multiple stories. In total, 1,644 individuals took part in this exercise, coding between 1 and 55 documents each. The modal number of responses per individual was one, with 1,251 workers completing the questionnaire only a single time. The average length of time spent taking the survey was 13.37 minutes. Together, the workers completed a total of about 851 hours of work. As previously stated, each worker was paid between 50 and 75 cents depending on whether the respondent answered a bonus question correctly, provided an adequately long summary of the news document, and spent sufficient time on the survey. With this incentive structure in place, procuring 3,899 responses had the potential to cost between $1,949.50 and $2,924.25, depending on the proportion of respondents that received the 25 cent bonus. In actuality, the costs came to $2,819.5, with about 90% of workers receiving the bonus.

As expected, crowdsourced labor is considerably cheaper than that obtained from highly trained experts. While individuals recruited online worked for about 3.31 dollars an hour, the expert coders were paid $15 per hour, representing relatively typical wages for trained research assistants. Moreover, the expert coders classified news documents at a similar rate to the crowd, indicating that the former are not necessarily more economical in their use of time. Of course, experts are expected to produce more accurate classifications than the average crowd worker, and fewer coders are required to classify each news document as a result. Nevertheless, employing three expert coders to classify each of the 446 documents required roughly 300 hours of work, costing about $4,500, considerably more than that spent on the crowdsourced labor.

[13]Of the 3,899 responses, 111 were removed because of processing issues or because respondents answered "No" to the verification question.

More than the hourly or aggregate economic costs, the greatest advantage of crowdsourced data collection is the relative gains in terms of the time required for data collection. While the crowd completed a total 851 hours of work, the time between the initial posting of this job and its completion was roughly five days.[14] By contrast, it took three expert coders nearly five weeks to complete the same task, working an average of 20 hours per week. Furthermore, crowdsourcing is scalable; as the task becomes larger, one may simply recruit more workers to participate. In short, as has been consistently found in other applications, crowdsourcing yields significant gains in both the costs and time required to complete tasks typically left to expert coders.

Crowdsourcing Accuracy

To improve accuracy and leverage the size of the crowd in data collection, the same story was analyzed by multiple workers, and auxiliary questions were asked to each respondent for additional robustness. Like any crowdsourcing approach, this poses a problem of aggregation: how can the output from several workers be combined most effectively? A simple technique is to code the MII attributes based on a plurality vote, which we termed *naive voting*. As the following results show, such aggregation, though relatively simple, significantly improves the accuracy of the collected data and can in the future serve as a very useful and easily-implemented tool for researchers. This is not to say, however, that naive voting *will always* produce more accurate data. It will, only if an individual is more likely to be correct than incorrect. Poorly or haphazardly designed questions that systematically point workers in the wrong direction will lead to inaccurate data even in the aggregate. Furthermore, provided individuals are more likely to be correct, the higher their expected accuracy, the more quickly responses will converge to the truth, and the fewer workers required to achieve acceptable accuracies. Thus, improvements in data

[14] We posted job requests in intervals during this time in order to monitor crowd responses. Had we posted the full set of job requests at once, this process would have been completed considerably faster.

accuracy is highly correlated with the creation of an intuitive and effective questionnaire.

Although the naive voting aggregation is effective, it is also limited in that it aggregates individual MII characteristics in isolation, rather than utilizing the full spectrum of information about the MII provided by respondents. For example, when identifying the initiator state through naive voting, only classifications of the initiator state were used, ignoring other information provided by the respondent that may aid in correctly identifying the initiator.

We utilized a more advanced aggregation technique, termed *error corrected voting*, using Bayesian network models to leverage the full range of information provided by respondents about the MII (Heckerman et al. 1995).[15] We used the Bayesian network because it outperformed other general machine learning algorithms, including support vector machines and random forests, as well as more specialized crowdsourcing aggregation algorithms such as ZenCrowd (Ororbia et al. 2015). In the error corrected voting algorithm, each MII attribute was *predicted* using a Bayesian network model. The predictors were the individual worker's responses to each question. The model's predictions for each MII attribute of interest were then aggregated using the same voting procedure as with the naive voting method.

An Application of Bayesian Networks

A Bayesian network is a probabilistic machine learning classifier, based on directed acyclic graphs (DAG), which leverages dependencies between variables to uncover meaningful relationships. In this case, each node in the network represents a question asked to the workers, and each connection between nodes represents a probabilistic dependence between those questions. As such, stronger connections correspond to stronger or more clearly identifiable relationships. This method of aggregation is advantageous, as opposed

[15]The analysis was conducted using the Waikato Environment for Knowledge Analysis (WEKA 3.7.10).

to other machine-learning classifiers, in that it is relatively quick to construct, makes minimal structural assumptions about the nature of relationships in the data, and is generally robust to minor variations in the data (Heckerman 1998; Nielsen and Jensen 2009). This approach is also useful because it is self-contained. That is, the network is built entirely from data provided by questionnaire responses, with no outside influence or information, making the approach very flexible to changes in variables of interest or questionnaire design.

"Error corrected" refers to the notion that the Bayesian network model utilizes the worker's responses to construct a network that describes dependencies among the questions, which can then be leveraged to generate accurate predictions pertaining to the initiators, targets, and actions within each story. This is done by first constructing an initial hypothesis about the network of dependencies in the data and then "correcting" this network by examining a set of alternative hypotheses or dependencies about the network structure. The dependencies between variables uncovered through this process are then used to generate weighted decisions pertaining to each variable of interest. For example, a decision rule might be a probabilistic variant of "IF Action Type==Show of Force AND Initiator State==North Korea THEN Target State=South Korea." This hypothetical rule might be a function of the fact that an overwhelming proportion of North Korean shows of force are directed at South Korea, and so the target state in such scenarios is largely assured.

An additional strength of this approach in this application is that Bayesian networks can utilize information from auxiliary and redundant questions asked to each respondent, which increases the breadth of information available for network construction. The auxiliary questions asked about important, but peripheral characteristics of the MII, such as the presence of non-state actors in the story or fatalities documented in the story. The redundant questions were more general, asking simply for broad classifications of

the news event, such as whether the action in the story was cooperative/conflictual in nature and whether the action was material or verbal in nature. Using these additional questions allows the potential to increase accuracy by increasing the scope of information exploited to generate predictions.

Like any statistical model, it is necessary to ensure that the Bayesian network model is not overfit to idiosyncrasies within the data. To ameliorate this risk, the model is built on 40 random samples of the data and examined for accuracy. This ensures the the model is robust and generalizable to various combinations of the data and does not develop decision rules based only on select cases.

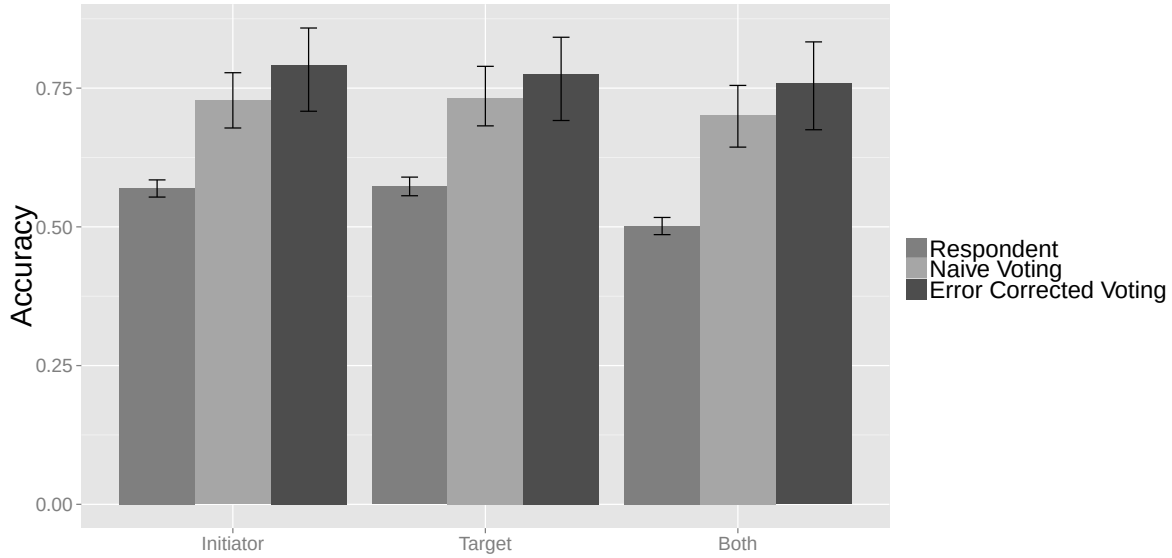
Actors

Figure 4 reports the crowdsourcing accuracy in classifying the actors in each news story.[16] These results are reported at both the respondent and group level, with the latter calculated using both naive and error corrected voting. We find that, while individual respondents may preform relatively poorly, accuracy is improved after aggregating these responses. We also find that *error corrected* voting modestly improves overall accuracy when compared to the more rudimentary *naive voting* form of aggregation. These results are important and underscore two fundamental strengths of crowdsourced data collection. First, while individual classifications of the data may be inaccurate, the size of the crowd can be leveraged to dramatically improve the quality of the data and, as Figure 4 illustrates, even straightforward aggregation techniques offer substantial improvements in accuracy. Second, crowdsourcing allows researchers flexibility in choosing aggregation methods that may be used to further improve the quality of results.

The crowdsourcing method identified actors with relatively high accuracy. Specifically, using the error corrected voting method, we correctly identified the initiator 79 percent

[16]As was the case when computing accuracy using machine coding, we consider the classification of actors to be correct for all cases that respondents *correctly* classify news documents as non-MIIs, since there are no initiators or targets in these cases

Figure 4: **Crowdsourcing Actor Accuracy**



Note: Figure displays crowdsourcing accuracy in identifying the state actors described across the sample of news documents. 95 percent confidence intervals are calculated from 1000 bootstrapped samples of the data.

of the time, the target 78 percent of the time, and both initiator and target 76 percent of the time. These results are a substantial improvement over the accuracy reported for the machine coding approach, which only ranges from 38 to 45 percent in actor categories.

We also examined whether there are any systematic causes of error among respondents. We found that a substantial number of responses with incorrect actor codings were from news documents about non-state actors such as rebel groups and terrorist organizations. Respondents often incorrectly identified these instances as MIIs, even though they do not take place between two state actors, as required by the coding rules. Respondents tended to misclassify the relevant actors for these news documents as a result. This bias is discussed in more detail below.

Another, less systematic source of error occurred when respondents provided a free-text entry too specific for the coding classification, referencing specific individuals or villages instead of countries. In a real-world application, such errors might be handled by post-processing responses to map the free-text entry to the corresponding state.

Actions

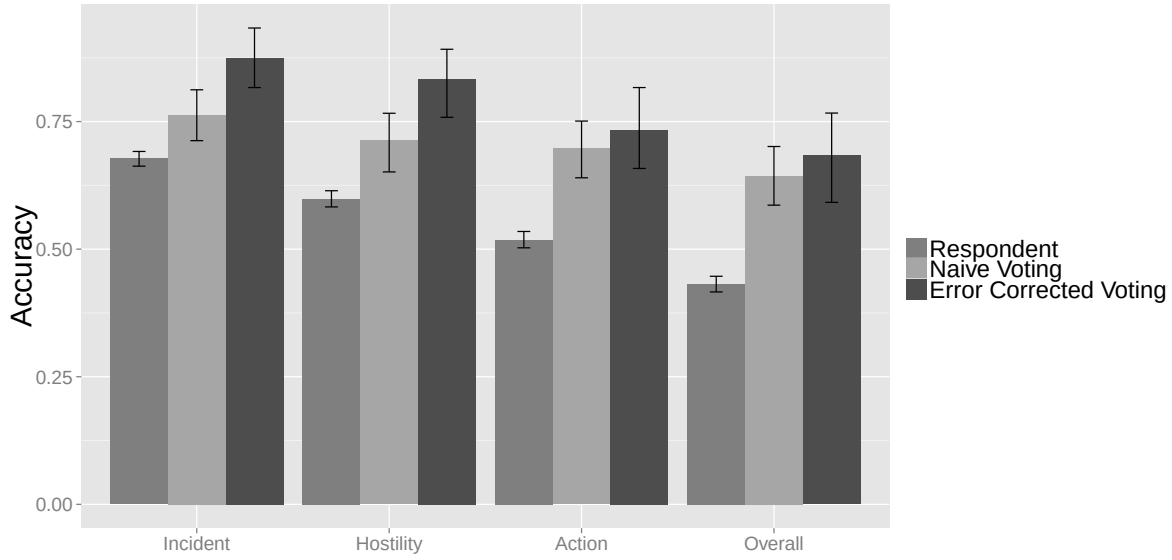
In this section we report accuracies for incident occurrence, hostility level, action type, and the overall accuracy. Incident occurrence is a dichotomous measure of whether the news story presents some type of MII. Hostility levels refer to the broader categories or groupings of militarized actions laid out by the MID coding rules. The five hostility levels are *No Militarized Action*, *Threat to Use Force*, *Display of Force*, *Use of Force*, and *War*. Overall accuracy refers to the correct coding of both actors and action type.

Exploring Figure 5, the incident accuracy measures how well the crowd is able to distinguish between MIIs and non-MIIs.[17] At the respondent level, the incident accuracy is just 68 percent, but it increases to 76 percent using *naive voting* and 88 percent using *error corrected voting*. The crowd performs well when distinguishing hostility levels, and these results, too, see significant improvements using advanced techniques. The hostility accuracy is 60 percent at the respondent level, 71 percent using *naive voting*, and 83 percent using *error corrected voting*. The action accuracy measures the ability of the crowd to correctly identify the correct MID action type. At the respondent level, this is just 52 percent, but the accuracy increases to 69 percent when using *naive voting* and 73 percent using *error corrected voting*.

Table 2 reports the cross tabulation of the crowd's responses (using *naive voting*) against the expert coding with respect to each specific action type. Note that while Table 2 reports *Tie* as a separate category, the following results are reported when one of the tied categories are chosen at random. In total, expert coders classified 266 news stories as containing information about MIIs and 180 news stories as non-incidents. At the incident level, of the 266 MII stories, the crowd classified 234 correctly and 32 incorrectly. When these false negatives occur, the crowd often mistakes shows of force, seizures, and attacks for non-incidents. Of the stories containing no information about MIIs, the crowd

[17]For voting methods, where there is a tie one answer is chosen at random.

Figure 5: **Action Accuracy**



Note: Figure displays crowdsourcing accuracy in identifying the militarized actions described across the sample of news documents. 95 percent confidence intervals are calculated from 1000 bootstrapped samples of the data.

classified 115 correctly and 65 incorrectly. Among these false positives, the crowd is most likely to classify non-events as attacks. At a rate of about three to one, the crowd clearly produces more false positives than false negatives. This is consistent with our previously-stated intent to structure a survey that minimizes the misclassification of stories about MIIs.

Moving beyond these aggregate classifications, Table 2 also shows that non-experts struggle when identifying the third hostility level, *Display of Force*. With respect to *Displays of Force*, many *shows of force* are miscoded, and the distribution of these false codings are dispersed across five different action types. This suggests that non-experts do not understand what constitutes a *show of force*, which may reflect ambiguities in the coding rules or perhaps loose wording in the questionnaire. For example, a *show of force* is "a public demonstration by a state of its military forces intended to intimidate another state but not involving actual combat operations" (Correlates of War 2000, 5). Contrast this with an *attack*, which is "the use of regular armed forces of a state to fire upon

Table 2: **Crowdsourcing MID Actions**

		True Coding									
		Not MII	Threat to use force	Show of Force	Alert	Border fortification	Border Violation	Occupation of Territory	Seizure	Attack	Clash
Crowd	Not MII	108		5	1	2		2	9	6	1
	Threat to use force	3	3								
	Show of Force	6		29		1					
	Border fortification	1		3	1	7					
	Border Violation	1			1	1	2				
	Occupation of Territory	1		1		1	2	2		1	
	Seizure	3						3	24		
	Attack	32		4			1	1		71	4
	Clash	9		2	4					3	44
	Tie	16		6	2	2	1	4	3	4	2

Note: The "tie" category pertains to documents for which respondents were equally divided in classifying a document among two or more action types. Empty cells are analogous to zeros and indicate that no documents were classified in a particular category.

the armed forces, population, or territory of another state" (Correlates of War 2000, 6). Conceptually, it is more difficult to observe the intent to intimidate than it is to observe a state military firing upon another state. This indicates that the accuracy attained when using non-expert data collection is partially dependent on the ease of identifying distinctions in the concept of interest. Thus, some concepts will lend themselves to this method more easily than others.

The overall accuracy, which measures both actions and actors, represents the crowd's ability to construct an MII. To do this, respondents must correctly identify the *Initiator State*, the *Target State*, and the *Militarized Action*. At the respondent level, the overall accuracy is just 43 percent. However, using *naive voting* this accuracy increases to 63 percent, and increases to 68 percent when using the *error corrected voting* model.

Overall, the results of our crowdsourcing method reveal (and confirm) several interesting insights. First, while individual workers may produce noisy or inaccurate data, the

wisdom of the crowd can be leveraged to produce accurate data in the aggregate. Additionally, the results of the naive voting aggregation show that, while more sophisticated aggregation methods may perform better, even simple aggregation techniques provide substantial increases in the accuracy of collected data. However, since the gains from the Bayesian network are considerably large for *Incident* and *Hostility* prediction, pairing with machine learning algorithms is a complimentary future line of research. Finally, as shown in the dispersion of *show of force* judgements, crowdsourcing appears to lend itself well to some concepts and not as well to others.

Conclusion

Incumbent methods of data collection often force researchers into a dilemma in which they must choose between cost and accuracy. While utilizing small groups of expert coders provides accurate data, the process is slow and expensive. Conversely, automated machine coding methods are cheap and fast, but consistently produce noisy data. The results presented here show that crowdsourcing, in many respects, marginalizes these drawbacks while retaining select benefits from each approach.

Our comparison of crowdsourced data collection methods to machine coding suggests that the former is better able to reproduce expert classifications according to the MID coding rules. Focusing on three important and basic characteristics of the MIIs, non-expert workers identified and classified the actors and militarized actions in a series of news documents at between 68 and 76 percent accuracy. Machine coding, on the other hand, produced accuracy measures that range from 30 to 45 percent. Of course, this is but one application comparing the relative advantages of crowdsourcing and machine coding. The fact that crowdsourcing performed so well in this particular application is particularly significant given that the MID is a subjective construct based on several layers of coding rules. While MIDs are complex cases to discern and properly code,

the development of clear and straightforward questionnaires allow non-expert crowds to recreate such data at relatively high accuracy. That is, these results robustly support the notion that, in the aggregate, even non-expert crowds can largely supplant experts for simple data collection, even for concept-driven data structures.

Our analyses also demonstrate novel ways in which crowdsourcing can be interlaced with other methodological advancements. Pre-processing news documents to extract named state entities allowed us to provide workers with valid options when identifying actors in associated documents, which prevented incorrect or incoherent results. Perhaps more importantly, we find that Bayesian modeling techniques can be used to more accurately aggregate responses obtained from multiple respondents.

Crowdsourcing also greatly reduced the time and financial costs and generated data much faster than data collected by small numbers of expert coders. The gains in time are particularly consequential, as reliance on human labor represents a primary bottleneck in many data collection processes (e.g. Palmer et al. (2015), 240). Furthermore, these reductions in cost underscore the scalability of crowdsourcing. That is, traditional data collection methods entail a linear increase in time for collecting more data. On the other hand, the increase in time when crowdsourcing additional data is marginal, meaning that data collection tasks can be expanded or replicated rather efficiently.

This project provides a baseline analysis of the utility of crowdsourcing and serves as a springboard for further exploration and improvements in this burgeoning method of data collection. While we find that crowdsourcing outperforms machine coding in this application, future studies will be necessary to determine whether this result pertains only to the classification of complex constructs like MIDs, or whether our results are generalizable to a broader set of classification tasks. Probing these issues further will be critical in developing a stronger understanding of measurement strategies available to political science researchers.

References

(2015). Phoenix pipeline. `http://phoenix-pipeline.readthedocs.org/en/latest/#`. Accessed: 2015-05-13.

Benoit, K., D. Conway, B. E. Lauderdale, M. Laver, and S. Mikhaylov (2014). Crowd-sourced text analysis: Reproducible and agile production of political data. Working Paper.

Berinsky, A. J., G. A. Huber, and G. S. Lenz (2012). Evaluating online labor markets for experimental research: Amazon.com's mechanical turk. *Political Analysis 20*(3), 351–368.

Berinsky, A. J., M. F. Margolis, and M. W. Sances (2012). Separating the shirkers from the workers? making sure respondents pay attention on self-administered surveys. *American Journal of Political Science 58*(3).

Bond, D., J. C. Jenkins, C. L. Taylor, and K. Schock (1997). Mapping mass political conflict and civil society issues and prospects for the automated development of event data. *Journal of Conflict Resolution 41*(4), 553–579.

Chandler, J., P. Mueller, and G. Paolacci (2014). Nonnaivete among amazon mechanical turk workers: Consequences and solutions for behavioral researchers. *Behavioral Research Methods 46*.

Cooper, S., F. Khatib, A. Treuille, J. Barbero, J. Lee, M. Beenen, A. Leaver-Fay, D. Baker, Z. Popovic, and F. Players (2010). Predicting protein structures with a multiplayer online game. *Nature 466*(August), 756–760.

Correlates of War (2000). *MID Incident Coding Manual*.

D'Orazio, V., S. T. Landis, G. Palmer, and P. Schrodt (2014). Separating the wheat from the chaff: Applications of automated document classification using support vector machines. *Political Analysis 22*(2), 224–242.

Ghosn, F., G. Palmer, and S. A. Bremer (2004). The mid3 data set, 1993-2001: Procedures, coding rules, and description. *Conflict Management and Peace Science 21*(2), 133–154.

Goertz, G. (2006). *Social Science Concepts: A User's Guide*. Princeton, NJ: Princeton University Press.

Goodman, J. K., C. E. Cryder, and A. Cheema (2013). Data collection in a flat world: The strengths and weaknesses of mechanical turk samples. *Journal of Behavioral Decision Making 26*.

Heckerman, D. (1998). *A tutorial on learning with Bayesian networks*. Springer.

Heckerman, D., D. Geiger, and D. M. Chickering (1995). Learning bayesian networks: The combination of knowledge and statistical data. *Machine learning 20*(3), 197–243.

Honaker, J., C. Ojeda, M. Berkman, and E. Plutzer (2013). Sorting algorithms for qualitative data to recover latent dimensions with crowdsourced judgments: Measuring state policies for welfare eligibility under tanf. Meeting of the Society for Political Methodology, Charlottesville.

Jones, D. M., S. A. Bremer, and J. D. Singer (1996). Militarized interstate disputes, 1816-1992: Rationale, coding rules, and empirical patters. *Conflict Management and Peace Science 15*(2), 163–213.

King, G. and W. Lowe (2003). An automated information extraction tool for international conflict data with performance as good as human coders: A rare events evaluation design. *International Organization 57*(3), 617–642.

Kittur, A., E. H. Chi, and B. Suh (2008). Crowdsourcing user studies with mechanical turk. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pp. 453–456. ACM.

Krippendorff, K. (2013). *Content Analysis: An Introduction to its Methodology* (Third ed.). Thousand Oaks, CA: Sage.

Nielsen, T. D. and F. V. Jensen (2009). *Bayesian networks and decision graphs*. Springer.

O'Brien, S. (2010). Crisis early warning and decision support: Contemporary approaches and thoughts on future research. *International Studies Review 12*(1), 87–104.

O'Connor, B., B. M. Stewart, and N. A. Smith (2013). Learning to extract international relations from political context. In *ACL (1)*, pp. 1094–1104.

Ororbia, A. G., Y. Xu, V. D'Orazio, and D. Reitter (2015). Error-correction and aggregation in crowd-sourcing of geopolitical incident information. Manuscript.

Palmer, G., V. D'Orazio, M. Kenwick, and M. Lane (2015). The mid4 data set, 2002-2010: Procedures, coding rules, and description. *Conflict Management and Peace Science Forthcoming*.

Paolacci, G. and J. Chandler (2014). Inside the turk understanding mechanical turk as a participant pool. *Current Directions in Psychological Science 23*(3), 184–188.

Paolacci, G., J. Chandler, and P. G. Ipeirotis (2010). Running experiments on amazon mechanical turk. *Judgment and Decision Making 5*(5).

Sabou, M., K. Bontcheva, and A. Scharl (2012). Crowdsourcing research opportunities: lessons from natural language processing. In *Proceedings of the 12th International Conference on Knowledge Management and Knowledge Technologies*, New York, pp. 17:1–17:8. ACM.

Sartori, G. (1984). *Social Science Concepts: A Systematic Analysis*. Beverly Hills: Sage.

Schrodt, P. A. (1994). Statistical characteristics of events data. *International Interactions 20*(1-2), 35–53.

Schrodt, P. A. (2010). Automated production of high-volume, near-real-time political event data. Meeting of the American Political Science Association, Washington.

Schrodt, P. A. (2012). *Conflict and Mediation Event Observations (CAMEO) Codebook*.

Schrodt, P. A. (2014). *TABARI: Textual Analysis By Augmented Replacement Instructions*.

Schrodt, P. A. and D. J. Gerner (1994). Validity assessment of a machine-coded event data set for the middle east, 1982-1992. *American Journal of Political Science 38*, 825–854.

Schrodt, P. A. and D. J. Gerner (2000). Cluster-based early warning indicators for political change in the contemporary Levant. *American Political Science Review 94*(4), 803–817.

Schrodt, P. A. and D. J. Gerner (2012). Analyzing international event data: A handbook of computer-based techniques.

Snow, R., B. O'Connor, D. Jurafsky, and A. Y. Ng (2008). Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, pp. 254–263. Association for Computational Linguistics.

Tingley, D. and M. Tomz (2014). Conditional cooperation and climate change. *Comparative Political Studies 47*(3).

Zeller, R. A. and E. G. Carmines (1980). *Measurement in the Social Sciences: The Link Between Theory and Data*. New York: Cambridge University Press.

Zhai, H., T. Lingren, L. Deleger, Q. Li, M. Kaiser, L. Stoutenborough, and I. Solti (2013). Web 2.0-based crowdsourcing for high-quality gold standard development in clinical natural language processing. *Journal of medical Internet research 15*(4).