Morning:

Real time event data and EventData R Package (presenter, Patrick Brandt, UT Dallas)

This session introduces a new real-time / Phoenix event dataset that is updated every day producing a new machine coded CAMEO-formatted dataset. This includes data coded from over 300 sources with updating dictionaries and geolocation. We cover how to access and use the datasets' API and how it can be accessed with our new R package.

TwoRavens for Event Data (presenter: Vito D'Orazio)

TwoRavens is a system of interlocking tools for exploring data, building and estimating statistical models, and visualizing and comparing results. It runs in any modern browser, and therefore requires only an Internet connection to use. While metadata are sent client-side, all data remain server-side, enabling TwoRavens to function with large and sensitive datasets. In this paper, we introduce TwoRavens for Event Data. This implementation expands TwoRavens to provide a number of features that improve event data accessibility, usability, and analysis. We have developed two new modes used to subset and aggregate event data, and provide time-series forecasts using back-end algorithms developed through DARPA's Data-Driven Discovery of Models program. Our tool integrates with a number of political event datasets, including Phoenix and ICEWS, and provides tutorials for new users. Subsets and aggregations may be downloaded at any time during a user's session. Data citations are provided with each download. Since all processing occurs server-side, operations are stored to inform and provide recommendations to future users.

Introducing the New TERRER Dataset (presenters: Andy Halterman, Jill Irvine)

This session introduces Terrier (temporally extended regularly reproducible international event records), a new machine coded CAMEO-formatted dataset covering 1979 to 2015. Terrier includes the most comprehensive set of global and regional papers available to date in an event dataset, as well as new dictionary updates and geoparsing. The presentation covers the features of Terrier, how it compares to existing datasets, and how researchers can access and begin working with Terrier.

The session also introduces a new dataset from Arabic language news reports covering 1992 to present. We describe the process of producing Arabic event data using open source tools and a team of coders, and how researchers can make use of the data.

How to Make Your Own Event Data (presenters: Andy Halterman, Jill Irvine)

This session describes the tools for making your own custom data sets. A primary obstacle to researchers creating their own event datasets is the difficulty of making new dictionaries. We discuss the process of creating new CAMEO dictionaries in Arabic and Spanish, including coder

training and management and new, semi-automated approaches to coding. We also give an introduction to the coding platforms developed through the NSF RIDIR project which are freely available to researchers making their own event datasets.

Petrarch Language Integration: A handbook for universal language adaptation (presenters: Jennifer Holmes and Viveca Pavon)

Semi-automated event recognition is a useful tool for identifying events that have not been quantified so far. The tool has gained momentum and continues to develop given technological advances that facilitate event recognition. However, most work to date has focused on English language text. We base this presentation on the experience of working with Spanish and Arabic, to introduce an outline of how to integrate other languages into Petrarch to increase the capabilities of semi-automated event detection. We provide guidelines for new users to adopt for the language of their choosing. Petrarch, like its predecessors, relies on the use of dictionaries for significant verb and noun identification, these would need to be translated into the desired language. In order to maintain language context and reliability, we have developed specific tools for each of these tasks. The Verb Translation App has been designed to incorporate diverse languages ranging from Roman to Semitic with human coder validation of verb translations. The Noun translation tool provides direct translation of included nouns as well as language specific synonyms and variations. This expands the dictionaries and allows for language specific context to be captured. The use of these tools and its overall integration will be further explained in this paper for a better understanding on how to expand the use of Petrarch.

Better Extraction from Text Towards Enhanced Retrieval (BETTER), John Beieler, IARPA

This talk will cover the Intelligence Advanced Research Projects Activity's (IARPA) Better Extraction from Text Towards Enhanced Retrieval (BETTER) program. BETTER aims to push forward the state of the art in complex semantic extraction from text, e.g., extracting structured events, in order to provide more accurate document retrieval and triage. BETTER includes many challenges for performer teams including multi-lingual extraction, extremely fine-grained semantic extraction, and query-by-example semantic information retrieval. Additionally, this talk will provide a brief overview of the current state-of-the-art and challenges in information extraction with a specific focus on event extraction.

Afternoon:

Alex Hanna, Toronto
MPEDS: A Semi-Automated Approach for the Generation of Protest Event Data

This article introduces the Machine-learning Protest Event Data System (MPEDS), a system for the semi-automated coding of protest event data. MPEDS uses natural language processing and machine learning tools to automate parts of protest event data creation process, is attentive to theoretical concerns of social movement scholarship and is readily available to movement scholars. The system reduces the labor required to generate protest event data in order to maximize temporal and spatial coverage of protest event data while minimizing the biases associated with coding only one or two national-level news sources. Data produced with the MPEDS system is as reliable, if not more so, than that produced by human coders. MPEDS is open, available for replication, and extendable by social movement researchers and computational social scientists.

Zachary Steinert-Threlkeld, UCLA
Protest Activity Detection and Perceived Violence Estimation from Social Media Images

We develop a novel visual model which can recognize protesters, describe their activities by visual attributes and estimate the level of perceived violence in an image. Studies of social media and protests use natural language processing to track how individuals use hashtags and links, often with a focus on those items' diffusion. These approaches, however, may not be effective in fully characterizing actual real-world protests (e.g., violent or peaceful) or estimating the demographics of participants (e.g., age, gender, and race) and their emotions. Our system characterizes protests along these dimensions. We have collected geotagged tweets and their images from 2013-2017 and analyzed multiple major protest events in that period. A multi-task convolutional neural network is employed in order to automatically classify the presence of protesters in an image and predict its visual attributes, perceived violence and exhibited emotions.

Natalie Ahn (University of California, Berkeley)
Flexible Event Extraction Using Knowledge Bases and Active Learning

This paper will introduce a tool to help social scientists extract events from text, combining existing general-domain resources with a simple user interface for event labeling and extraction. The tool is meant for researchers who are not advanced programmers or computational linguists, and who wish to extract new types of events that may not fit within existing event patterns or ontologies. I use SpaCy to dependency parse raw text and extract verb predicates and their noun arguments, as potential actions and participants. I then match similar groups of predicates and arguments across documents, based on similar syntactic and semantic features. For these features, I draw from several external knowledge bases, including FrameNet and WordNet. To incorporate multiple resources into a unified representation of similar

predicate-argument groups, I am encoding these features into a lower-dimensional vector space.
This tool represents an alternative to more established or traditional event extraction systems. In this approach, labor-intensive resources like event templates are at least partially replaced with less supervised low-level representation learning, incorporating general-purpose linguistic resources. The user-facing tool then facilitates a high-level interactive learning process to obtain just enough information to identify the events a user wishes to extract. The paper will include evaluation metrics and discuss potential trade-offs between different approaches. This system may not perform as well as more established methods for extracting well-defined event types within existing ontologies. But it may complement those systems and help extend the application of event data to new topics and domains.

Scott Althaus, Buddy Peyton, and Dan Shalmon.
Cline Center for Advanced Social Research
University of Illinois at Urbana Champaign

Spatial and Temporal Dynamics of Boko Haram Activity across Six Event Generation Pipelines

Recent efforts to test the validity of fully-automated event generation pipelines against human-generated event data (e.g., Wang, Kennedy, Lazer, and Ramakrishnan 2016) have struggled to identify a strong “apples to apples” test case that offers a direct comparison across multiple event extraction systems. This paper presents findings from a case study of Boko Haram activity in and around the countries of Nigeria, Cameroon, Chad, and Niger over a three-month offensive that ran from 22 January to 30 April 2015. We compare the spatial and temporal dynamics of contentious events related to the Boko Haram offensive that were generated by six prominent event extraction pipelines—ACLED, SCAD, ICEWS, GDELT, PETRARCH, and the Cline Center’s SPEED project. Focusing on the subset of events encompassing politically motivated attacks (e.g., assassinations, kidnappings, hostage takings, attacks on personnel), political expression events (e.g., protests, strikes, symbolic acts), and destabilizing state acts (e.g., mutinies by armed forces, proactive arrests/detentions, confiscations of property), this case study clarifies the strengths, weaknesses, and validity challenges associated with different analytical strategies for generating small-scale conflict event data using news reports. By focusing on events associated with a single region, time period, and contentious actor, this case study was designed not only to clarify the impact of human-coded versus machine-coded extraction systems, but also to assess the impact of different news sourcing strategies for documenting small-scale contentious events occurring in remote places.

Jesse Hammond, Naval PGS
Cheap Talk and costly action in the international system

Classical liberals and constructivist scholars have commonly used the term “security community” to describe emergent relationships of trust that arise between states, particularly liberal democracies. Over time, states recognize one another as being fundamentally similar and trustworthy, developing joint identities of friendship that make militarized conflict not just unlikely but virtually unthinkable. This literature has a rich body of theoretical and qualitative analysis that provides both logic and evidence to support the idea that shared identity can explain peace and cooperation in a way that purely functional or material factors cannot. However, it has been difficult to test these theories quantitatively, largely due to the difficulty of operationalizing the underlying concept of “identity” or “friendship”.

I propose a quantitative approach to detect underlying communities of friendship and enmity by examining the day-to-day interactions between states. I develop an R package currently titled “EventNetworks” (https://github.com/jrhammond/EventNetworks) to gather, process, and convert several major event data sets (currently ICEWS, Phoenix, and the new historic Phoenix from UIUC) into flexible temporal network structures. Embedding these dyadic events in a larger network structure of interaction allows me to not only proxy for friendship and enmity between state dyads, but to identify emergent communities of cooperation that exert indirect effects on interstate relationships.

State dyads that are embedded in de-facto communities of positive engagement benefit from a pool of potential mediators to increase communication and lower the likelihood of major conflict. I expect that both direct positive ties and membership in the same de-facto community will be associated with a lower likelihood of major conflicts and an increased likelihood of active cooperation within a given state dyad. By looking at patterns of repeated interaction between states, this approach improves our ability to identify groups of “friends” and “enemies” relative to structural explanations such as shared IGO membership or joint democratic institutions.

CLAUDIO CIOFFI-REVILLA, GMU

ON SOME APPLICATIONS OF COMPLEXITY SCIENCE TO POLITICAL EVENT DATA ANALYSIS

Political event data analysis is undergoing a significant and potentially consequential renaissance driven by big data, new geospatial analytics, and new algorithmic tools, among other significant innovations. However, while this is all good news from a basic and applied science perspective, these and other similar positive developments fail to provide sufficient safeguards against persistent methodological pathologies caused by misused or abused methods from mainstream quantitative political science still being taught in the traditional curriculum and published in leading publications. This paper covers a set of viable methods drawn from theory and research in complexity science applied to political event data analysis. Such methods include nonequilibrium (i.e., non-Gaussian) event processes, nonstationary time-series, and related statistical, mathematical, and computational approaches that have

either demonstrable value in the extant literature, or potential applications for new research that can advance knowledge frontiers.

Nick Dietrich and Kristine Eck, Penn State

Known Unknowns: Explaining Media Bias in the Reporting of Political Violence

How does sourcing affect which events are included in political violence datasets? When collecting data on human rights violations and other political violence events, researchers must make decisions about which sources to include. These data collection efforts often rely on media reports, but various factors affect the likelihood that an event will be reported on. We investigate how sourcing decisions systematically affect which events are observed. We explore sourcing bias using the UCDP GED dataset, which includes events from media reports, NGOs, international organizations, and other sources. Our analysis leverages variation in sourcing of GED events to examine how geographical, technological, and political factors affect the likelihood that an event is recorded in media reports or other sources. The question of bias in political violence data is a particularly consequential one for machine-coded datasets that rely on a corpus of media reports gathered by Lexis-Nexis or other news aggregators.

David Muchlinski, UNSW

Using Social Media to Predict Low-Intensity Political Violence

The study of political violence is commonly concerned with mass violence like civil war onset or rebellion. Such events are relatively easy to code in datasets because most sources of data (e.g., news reports) report on their occurrence. Other important sources of conflict, however, may not rise to a significant level in order to be reported in common textual sources. Using machine learning, this article develops a new method to predict the occurrence of violence by mining social media text. Rather than considering words as atomized features with no relation to each other, we utilize a method that retains the syntactic and semantic relationship of words within text by transforming it into a dense real valued vector. We show that a neural network using these vectors as its input provides superior predictive accuracy of violence compared to other commonly utilized machine learning methods for textual analysis. This increase in predictive accuracy is due to superior clustering of topics made possible by retaining the linguistic features of text-as-data that many researchers consider unimportant. Preserving the lexical features of text provides richer data to machine learning algorithms, enhancing the clarity of clustered events, while also providing distinct borders between different events, enhancing prediction of low-intensity political violence.