

# Introducing ICBe: Very High Recall and Precision Event Extraction from Narratives about International Crises

Rex W. Douglass<sup>a,1</sup>, Thomas Leo Scherer<sup>a</sup>, J. Andrés Gannon<sup>b</sup>, Erik Gartzke<sup>a</sup>, Jon Lindsay<sup>c</sup>, Shannon Carcelli<sup>d</sup>, Jonathan Wilkenfeld<sup>d</sup>, David M. Quinn<sup>d</sup>, Catherine Aiken<sup>e</sup>, Jose Miguel Cabezas Navarro<sup>f</sup>, Neil Lund<sup>d</sup>, Egle Murauskaite<sup>d</sup>, and Diana Partridge<sup>d</sup>

<sup>a</sup>University of California, San Diego ; <sup>b</sup>Vanderbilt University; <sup>c</sup>Georgia Institute of Technology; <sup>d</sup>University of Maryland; <sup>e</sup>Georgetown University; <sup>f</sup>Universidad Mayor

This manuscript was compiled on August 10, 2022

How do international crises unfold? We conceptualize of international relations as a strategic chess game between adversaries and develop a systematic way to measure pieces, moves, and gambits accurately and consistently over a hundred years of history. We introduce a new ontology and dataset of international events called ICBe based on a very high-quality corpus of narratives from the International Crisis Behavior (ICB) Project. We demonstrate that ICBe has higher coverage, recall, and precision than existing state of the art datasets and conduct two detailed case studies of the Cuban Missile Crisis (1962) and Crimea-Donbas Crisis (2014). We further introduce two new event visualizations (event iconography and crisis maps), an automated benchmark for measuring event recall using natural language processing (synthetic narratives), and an ontology reconstruction task for objectively measuring event precision. We make the data, online appendix, replication material, and visualizations of every historical episode available at a companion website [www.crisisevents.org](http://www.crisisevents.org) and the github repository.

Diplomacy | War | Crises | International Affairs | Computational Social Science

If we could record every important interaction between countries in all of diplomacy, military conflict, and international political economy, how much unique information would this chronicle amount to, and how surprised would we be to see something new? In other words, what is the entropy of international relations? This record could in principle be unbounded, but the central conceit of social science is that there are structural regularities that limit what actors can do, their best options, and even which actors are likely to survive (1, 2). If so, then these events can be systematically measured, and accordingly, massive effort is expended in social science attempting to record these regularities.\* Thanks to improvements in natural language processing, more open-ended efforts have begun to capture entire unstructured streams of events from international news reports.<sup>†</sup> How close these efforts are to accurately measuring all or even most of what is essential in international relations is an open empirical question, one for which we provide new evidence here.

Our contribution is a high coverage ontology and event dataset for key historical episodes in 20th and 21st-century international relations. We develop a large, flexible ontology

of international events with the help of both human coders and natural language processing. We apply it sentence-by-sentence to an unusually high-quality corpus of historical narratives of international crises (1, 29–32). The result is a new lower bound estimate of how much actually happens between states during pivotal historical episodes. We then develop several methods for objectively gauging how well these event codings reconstruct the information contained in the original narrative. We conclude by benchmarking our event codings against several current state-of-the-art event data collection efforts. We find that existing systems produce sequences of events that do not contain enough information to reconstruct the underlying historical episode. The underlying fine-grained variation in international affairs is unrecognizable through the lens of current quantification efforts.

This is a measurement paper that makes the following argument — there is a real-world unobserved latent concept known as international relations, we propose a method for systematically measuring it, we successfully apply this method producing a new large scale set of measurements, those measurements exhibit several desirable kinds of internal and external validity, and those measurements out-perform other existing approaches. The article organizes that argument into eight sections: task definition; corpus; priors/existing state

## Significance Statement

Countries routinely face crises that risk escalating into full scale war but we do not have systematic measurements of the progression of past crises and what moves and counter moves led to or helped avoid war. Instead policy makers typically rely on one or two historical analogies, chosen through ad hoc selection criteria, and described in unsystematic terms. This paper introduces a new scientific approach to measuring the step by step moves of international crises over the last hundred years, combining subject expertise with state of the art natural language processing and machine learning methods. It serves as a guide for constructing and evaluating large scale measurement collection in the social sciences.

Conceptualization: R.W.D., E.G., J.L.; Methodology: R.W.D., T.L.S.; Software: R.W.D.; Validation: R.W.D., T.L.S.; Formal Analysis: R.W.D., T.L.S.; Investigation: S.C., R.W.D., J.A.G., O.K., N.L., E.M., J.M.C.N., D.P., D.Q., J.W.; Data Curation: R.W.D., D.Q., T.L.S., J.W.; Writing - Original Draft: R.W.D., T.L.S.; Writing - Review & Editing: R.W.D., J.A.G., E.G., T.L.S.; Visualization: R.W.D., T.L.S.; Supervision: E.G.; Project Administration: S.C., R.W.D., J.A.G., D.Q., T.L.S., J.W.; Funding Acquisition: E.G., J.L.

\* See work on crises (3, 4), militarized disputes (5–7), wars (8, 9), organized violence (10, 11), political violence (12), sanctions (13), trade (14), and international agreements (15–17), dispute resolution (17, 18), and diplomacy (19, 20).

<sup>†</sup> See (21); (22); (23); (24); (25); (26). On event-extraction from images and social-media see (27) and (28).

<sup>1</sup> To whom correspondence should be addressed. E-mail: rexouglass@gmail.com

46 of the art; ICBe coding process; internal consistency; case  
 47 study selection; recall; and precision. A final section  
 48 concludes.  
 49 **Task Definition**  
 50 We consider the measurement task of abstracting discrete  
 51 events about a historical episode in international relations.  
 52 The easiest way to convey the task is with an example. Figure  
 53 1 shows a narrative account of the Cuban Missile Crisis  
 54 (1962) alongside a mapping from each natural language sentence  
 55 to discrete machine readable abstractive events. Formally,  
 56 a historical episode,  $H$ , is demarcated by a period of time  
 $[T_{start}, T_{end}] \in T$ , a set of Players  $p \in P$ , and a set of behaviors they undertook during that time  $b \in B$ . International Relations,  $IR$ , is the system of regularities that govern the strategic interactions that world actors make during a historical episode, given their available options, preferences, beliefs, and expectations of choices made by others. We observe neither  $H$  nor  $IR$  directly. Rather the Historical Record,  $HR$ , produces documents  $d \in D$  containing some relevant and true (as well as irrelevant and untrue) information about behaviors that were undertaken recorded in the form of unstructured natural language text. The task is to combine informative priors about  $IR$  with an unstructured corpus  $D$  to produce a series of structured discrete events,  $e \in E$ , that have high coverage, precision, and recall over what actually took place in history,  $H$ .

**Corpus**  
 72 For our corpus,  $D$ , we select a set of unusually high-quality historical narratives from the International Crisis Behavior (ICB) project ( $n = 471$ ) (SI Appendix, Table A1) (33, 34).<sup>†</sup> Their domain is 20th and 21st-century crises, defined as a change in the type, or an increase in the intensity, of disruptive interaction with a heightened probability of military hostilities that destabilizes states' relationships or challenges the structure of the international system (3).<sup>§</sup> Crises are a significant focus of detailed single case studies and case comparisons because they provide an opportunity to examine behaviors in IR short of, or at least prior to, full conflict (3, 35–42). Case selection was exhaustive based on a survey of world news archives and region experts, cross-checked against other databases of war and conflict, and non-English sources (33, 43). Each narrative was written by consensus by a small number of scholars, using a uniform coding scheme, with similar specificity (44). The corpus is unique in IR because it is designed to be used in a downstream quantitative coding project.

**Prior Beliefs about IR, Ontological Coverage, and the Existing State of the Art**  
 91 Next we draw informative prior beliefs about the underlying process of IR that we expect to govern behavior during historical episodes and their conversion to the historical record. We  
 92 organize our prior beliefs along two overarching axes, summarized in detail by Table 1.

93 The first axis (rows) represents the types of information we  
 94 expect to find in IR and forms the basis for our proposed ontology. We employ a metaphor of a chess game, with players

101 (polities, rebel groups, IGOs, etc.), pieces (military platforms, 102 civilians, domains), and behaviors (think, say, do). Precise sequencing is required to capture gambits (sequences of moves) 103 and outcomes (victory, defeat, peace, etc.), while precise geocoding is required to understand the chessboard (medium of 104 conflict). We find 472 actors and 117 different behaviors, and 105 provide a full codebook in the online material.<sup>¶</sup>  
 106

107 We base our informed priors primarily on two sources of information. The first is the extensive existing measurement 108 efforts of IR which we provide citations to alongside each 109 concept. Second, we performed preliminary natural language 110 processing of the corpus and identified named entities and 111 behaviors mentioned in the text. Verbs were matched to the 112 most likely definition found in Wordnet (45), tallied, and then 113 aggregated into a smaller number hypernyms balancing 114 conceptual detail and manageable sparsity for human coding (SI 115 Appendix, Table A2).

116 The second axis (columns) compares the very high ontological 117 coverage of ICBe to existing state of the art systems in production 118 and with global coverage. They begin with our contribution ICBe, 119 alongside other event-level datasets including CAMEO dictionary 120 lookup based systems (Historical Phoenix (46); ICEWS (24, 25); Terrier (26)), the Militarized Interstate 121 Disputes Incidents dataset, and the UCDP-GED dataset (10, 122 11, 47).<sup>||</sup> The final set of columns compares episode-level 123 datasets beginning with the original ICB project (3, 4, 56); 124 the Militarized Interstate Disputes dataset (5, 6, 57, 58), and 125 the Correlates of War (8). With the exception of large scale 126 CAMEO dictionary based systems, the existing state of the 127 art quantitative datasets ignore the vast majority of the information 128 content found in international relations.<sup>\*\*</sup>

## ICBe Coding Process

132 The ICBe ontology follows a hierarchical design philosophy where a smaller number of significant decisions are made early 133 on and then progressively refined into more specific details 134 (59).<sup>††</sup> Each coder was instructed to first thoroughly read 135 the full crisis narrative and then presented with a custom 136 graphical user interface (SI Appendix, Fig. B1). Coders 137 then proceeded sentence by sentence, choosing the number 138 of events (0-3) that occurred, the highest behavior (thought, 139 speech, or activity), a set of players ( $P$ ), whether the means 140 were primarily armed or unarmed, whether there was an 141 increase or decrease in aggression (uncooperative/escalating 142 or cooperative/de-escalating), and finally one or more 143 non-mutually exclusive specific activities. Some additional 144 details like location and timing information was always 145 collected while other details were only collected if appropriate, 146 e.g. force size, fatalities, domains, units, etc. A unique feature 147 of the ontology is that thought, speech, and do behaviors can 148 be nested into combinations, e.g. an offer for the U.S.S.R. to 149 remove missiles from Cuba in exchange for the U.S. removing 150

<sup>†</sup> See the Github Repository [ICBEEventData](#).

<sup>||</sup> Additional relevant but dated or too small of an overlap in domain include BCOW (48), WEIS (49), CREON (50), CASCON (51), SHERFACS (52), Real-Time Phoenix (23), and COFEE (53) (see histories in (54) and (55)).

<sup>\*\*</sup> See (53) for a recent review of ontological depth and availability of Gold Standard example text.

<sup>††</sup> This process quickly focuses the coder on a smaller number of relevant options while also allowing them to apply multiple tags if the sentence explicitly includes more than one or there is insufficient evidence to choose only one tag. The guided coding process also allows for the possibility that earlier coarse decisions have less error than later fine-grained decisions.

<sup>‡</sup> The Online Appendix is at the [ICBEEventData Github Repository](#).

<sup>§</sup> On near crises see (32).

**Fig. 1.** Case Study 1: Cuban Missile Crisis (1962) - ICB Narrative vs. ICBe Events

5	Natural Language Sentences (ICB Corpus)	Machine Readable Events (ICBe)
4	When the U.S. discovered the presence of Soviet military personnel in Cuba on 7 September 1962 it called up 150,000 reservists.	mobilization  100ks discover fact -  deployment to area  100s; 1ks mobilization  1ks
5	The Soviets mobilized on the 11th.	
6	Although persistent rumors circulated concerning the deployment of Soviet missiles in Cuba, Soviet Ambassador Anatoly Dobrynin denied the charges, and Premier Khrushchev gave his personal assurances that ground-to-ground missiles would never be shipped to Cuba.	
8	The U.S. crisis was triggered on 16 October when the CIA presented to President Kennedy photographic evidence of the presence of Soviet missiles in Cuba.	discover fact;start of crisis -  deployment to area fortify
9	The U.S. responded with a decision on the 20th to blockade all offensive military equipment en route to Cuba.	start of crisis -  deployment to area blockade  coastline 10ks blockade  coastline
10	When this was announced on 22 October, a crisis was triggered for Cuba and the USSR. An urgent meeting of the UN Security Council was requested by both the U.S. and Cuba on the 22nd, and by the USSR the next day.	blockade  coastline start of crisis -  USA
11	On the 23rd as well, the Soviets accused the United States of violating the UN Charter and announced an alert of its armed forces and those of the Warsaw Pact members.	appeal  UN -  meeting  USA raise in alert         100ks
12	That day Cuba responded by condemning the U.S. blockade and declaring its willingness to fight.	accuse  USA -  violate terms of treaty  UN disapprove  USA  coastline
13	A resolution was adopted on the 23rd by the OAS calling for the withdrawal of the missiles from Cuba and recommending that member-states take all measures, including the use of force, to ensure that the government of Cuba would not continue to receive military material.	demand  area -  withdraw from area     1ks express intent  area -  mediation  UN  USA
14	On 24 October the Security Council adopted a resolution requesting the Secretary-General to confer with the parties.	express intent -  mediation  UN  USA mediation  USA
15	On that same day, U Thant began mediation by sending identical letters to Khrushchev and Kennedy which proposed that the Soviet Union and the United States enter into negotiations, during which period both the shipment of arms and the quarantine would be suspended.	offer  USA happens -  withdraw from area     -C will happen  USA end blockade
16	Moscow's major response to the crisis was a letter from Khrushchev to Kennedy on 26 October offering the removal of Soviet offensive weapons from Cuba and the cessation of further shipments in exchange for an end to the U.S. quarantine and a U.S. assurance that it would not invade Cuba.	offer  USA happens -  withdraw from area    -C will happen  UN  USA border violation  airspace  1s  skull offer  USA happens -  withdraw from area    -C will happen  UN  USA border violation  airspace
17	The situation was exacerbated on the 27th when a U.S. U-2 surveillance plane was shot down.	offer  USA happens -  withdraw from area    -C will happen  UN  USA border violation  airspace  1s  skull
18	That day another Khrushchev letter was received in Washington offering the removal of Soviet missiles from Cuba in exchange for the removal of U.S. missiles from Turkey.	offer  USA happens -  withdraw from area    -C will happen  UN  USA border violation  airspace
19	U.S. mobilization and aerial reconnaissance flights were stepped up.	border violation;mobilization mobilization  airspace
20	And on the 27th President Kennedy sent the Soviet premier an acceptance of the proposals contained in the letter of 26 October while making no reference to Khrushchev's second letter of the 27th.	accept  USA -  UN
21	The following day Khrushchev notified the U.S. government that he had ordered work on the missile sites in Cuba stopped.	accept  USA -  UN
23	At the same time he warned Washington that U-2 reconnaissance flights over Cuba must be stopped as well.	demand  area -  withdraw from area    airspace
24	The crisis continued at a lower level of intensity for several more weeks due to Cuban President Castro's demands concerning a U.S. pledge not to invade his country.	demand  area -  de-mobilization;lower alert    body of water;coastline 10ks
25	On 30 October U Thant began talks in Havana, and Kennedy agreed to lift the quarantine for the duration of the talks.	end blockade    coastline 10ks accept  UN -  end blockade    coastline
26	When Cuba rejected UN inspection, the U.S. resumed the quarantine and air surveillance.	reject  UN -  unspecified cooperation  UN blockade  coastline  airspace;coastline 10ks
27	The Kremlin sent Deputy Premier Anastas Mikoyan to Cuba on 2 November to try to persuade Castro to allow UN inspection.	blockade  coastline  airspace;coastline 10ks discussion;meeting  UN
28	When this proved unsuccessful, a U.S.-USSR agreement was reached on 7 November allowing U.S. inspection and interception of Soviet ships leaving Cuba and the photographing of the missiles.	sign formal agreement  UN
29	The following day the superpowers negotiated the removal of the IL-28 bombers which Castro had claimed were Cuban property.	withdraw from area    100s sign formal agreement  UN end of crisis
30	Castro's agreement was conveyed to the U.S. on 20 November 1962, which terminated the Missile crisis for all three actors.	end blockade    body of water;coastline 10ks
31	The U.S. naval quarantine was lifted immediately, but aerial surveillance continued until the agreement was completely carried out.	

**Table 1. Ontological coverage of ICBe versus existing State of the Art**

		Concept	Events Datasets								Episodes Datasets			
Domain	Player		MIDs	Incidents	UCDP-GED				MIDs	COW	ICB	MIDs		
	ICBe	Phoenix	Terrier	Automated (CAMEO)	1918	1945	1977	1995	1993	1989	1918	1816	1816	
		Start (14) End (5, 6) N Coders				2017	2019	2018	2020	2010	2015	2017	2014	2007
		Corpus Date Location				32k	8.5M	28.4M	17.5M	9.6K	128k	1K	5.9K	1K
					Hand				Hand	Hand	Hand	Hand	Hand	Hand
	Players	States (710) Subnational Actors (1114) IGO/NGO (1518) Civilians (19, 20)			ICB	News		Mix	News		Mix	Mix	Mix	
	Pieces	Fatalities (21, 22) Force Size (2325) Force Domain (2629) Geography (location, territorial change) (30)			Event	Article		Event	Article		Event	Event	Event	
	Think	Alert (Start/End Crisis) (31) Wishes (Desire/Fear) (32, 33) Evaluation (Victory/Defeat) (34) Aims (Territory, Policy, Regime, Preemption) (35) Awareness (Discover, Become Convinced) (3638)			Event	Event		Actor	Event		Actor	Event	Actor	
	Say	React to past event (Praise, Disapprove, Accept, Reject, Accuse) (3941) Request future event (Appeal, Demand) (42) Predict future event (Promise, Threaten, Express Intent, Offer Without Condition) (43, 44) Predict with condition (Offer, Ultimatum) (45)			✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
	Unarmed	Government (Leadership/Institution Change, Coup, Assassination) (4650) By Civilians (Protest/Riot/Strike) (51) Against Civilians (Terrorism, Domestic Rights, Mass Killing, Evacuate) (52, 53) Diplomacy (Discussion, Meeting, Mediation, Break off negotiations, Withdraw/Expel Diplomats, Propaganda) (54) Legal Agreements (Sign Agreement, Settle Dispute, Join War on Behalf of, Ally, Mutual Defense Pact, Open Border, Cede Territory, Allow Inspections, Political Succession, Leave Alliance, Terminate Treaty) (5557) Violate Agreement (Violate Terms of Agreement) (58) Mutual Cooperation or Directed Aid (Economic cooperation or Aid, Military Cooperation, Intelligence Cooperation, Unspecified) (59) Directed Aid (General Political Support, Economic Aid, Humanitarian Aid, Military Aid, Intelligence Aid, Unspecified Aid) (60, 61)			✓	✓	✓	✓						
	Do	Preparation (Alert, Mobilization, Fortify, Exercise, Weapons Test) (62) Maneuver (Deployment, Show of Force, Blockade, No Fly Zone, Border Violation) (63) Combat (Battle/Clash, Attack, Invasion/Occupation, Bombard, Cease Fire, Retreat) (64, 65) Strategic (Declare War, Join War, Continue Fighting, Surrender, End War, Withdraw from War, Switch Sides) (66, 67) Autonomy (Assert Political Control Over, Assert Autonomy Against, Annex, Reduce Control Over, Decolonize) (6870)			✓	✓	✓	✓						

missiles from Turkey. Through compounding, the ontology can capture what players were said to have known, learned, or said about other specific fully described actions.

Each crisis was typically assigned to 2 expert coders and 2 novice coders with an additional tie-breaking expert coder assigned to sentences with high disagreement.<sup>††</sup> For the purposes of measuring intercoder agreement and consensus, we temporarily disaggregate the unit of analysis to the Coder-Crisis-Sentence-Tag (n=993,740), where a tag is any unique piece of information a coder can associate with a sentence such as an actor, date, behavior, etc. We then aggregate those tags into final events (n=18,783), using a consensus procedure (SI Appendix, Algorithm B2) that requires a tag to have been chosen by at least one expert coder and either a majority of expert or novice coders. This screens noisy tags that no expert considered possible but leverages novice knowledge to tie-break between equally plausible tags chosen by experts.

## Internal Consistency

We evaluate the internal validity of the coding process in several ways. For every tag applied we calculate the observed intercoder agreement as the percent of other coders who also applied that same tag (SI Appendix, Fig. B3). Across all concepts, the Top 1 Tag Agreement was low among novices (31%), moderate for experts (65%), and high (73%) following the consensus screening procedure.

We attribute the remaining disagreement primarily to three sources. First, we required coders to rate their confidence which was observed to be low for 20% of sentences- half due to a mismatch between the ontology and the text (“survey doesn’t fit event”-45%) and half due to a lack of information or confused writing in the source text (“more knowledge needed”-40%, “confusing sentence”-6%). Observed disagreement varied predictably with self reported confidence (SI Appendix, Fig. B4). Second, as intended agreement is higher (75-80%) for questions with fewer options near the root of the ontology compared to agreement for questions near the leafs of the ontology (50%-60%). Third, individual coders exhibiting nontrivial coding styles, e.g. some more expressive applying many tags per concept while others focused on only the single best match. We further observed unintended synonymy, e.g. the same information can be framed as either a threat to do something or a promise not to do something.

## Case Study Selection

The remaining two qualities we seek to measure are recall and precision of ICBe events in absolute terms and relative to other existing systems. We provide full ICB narratives, ICBe coding in an easy to read iconographic form, and a wide range of visualizations for every case on the companion website. In this paper, we focus on two deep case studies. The first is the Cuban Missile Crisis (Figure 1) which took place primarily in the second half of 1962, involved the United States, the Soviet Union, and Cuba, and is widely known for bringing the world to the brink of nuclear war (hereafter Cuban Missiles). The second is the Crimea-Donbas Crisis (SI Appendix Figure D1) which took place primarily in 2014, involved Russia, Ukraine,

<sup>††</sup>Expert coders were graduate students or postgraduates who collaboratively developed the ontology and documentation for the codebook. Undergraduate coders were students who engaged in classroom workshops.

and NATO, and within a decade spiraled into a full scale invasion (hereafter Crimea-Donbas). Both cases involve a superpower in crisis with a neighbor, initiated by a change from a friendly to hostile regime, with implications for economic and military security for the superpower, risked full scale invasion, and eventually invited intervention by opposing superpowers. We choose these cases because they are substantively significant to 20th and 21st century international relations, widely known across scientific disciplines and popular culture, and are sufficiently brief to evaluate in depth.

## Recall

Recall measures the share of desired information recovered by a sequence of coded events,  $Pr(E|H)$ , and is poorly defined for historical episodes. First, there is no genuine ground truth about what occurred, only surviving texts about it. Second, there is no *a priori* guide to what information is necessary detail and what is ignorable trivia. History suffers from what is known as the Coastline Paradox (60) — it has a fractal dimension greater than one such that the more you zoom in the more detail you will find about individual events and in between every two discrete events. The ICBe ontology is a proposal about what information is important, but we need an independent benchmark to evaluate whether that proposal is a good one and that allows for comparing proposals from event projects that had different goals. We need a yardstick for history.

Our strategy for dealing with both problems is a plausibly objective yardstick called a synthetic historical narrative. For both case studies, we collect a large diverse corpus of narratives spanning timelines, encyclopedia entries, journal articles, news reports, websites, and government documents. Using natural language processing (fully described in SI Appendix, Algorithm C1), we identify details that appear across multiple accounts. The more accounts that mention a detail, the more central it is to understanding the true historical episode. The theoretical motivation is that authors face word limits which force them to pick and choose which details to include, and they choose details which serve the specific context of the document they are producing. With a sufficiently large and diverse corpus of documents, we can vary the context while holding the overall episode constant and see which details tend to be invariant to context. Intuitively, a high quality event dataset should have high recall for context invariant details both because of their broader relevance and also because they are easier to find in source material.

Synthetic historical narratives for Cuban Missiles (51 events drawn from 2020 documents) and Crimea-Donbas (30 events drawn from 971 documents) appear in Figure 2. Each row represents a detail which appeared in at least five documents along with an approximate start date, a hand written summary, the number of documents it was mentioned in, and whether it could be identified in the text of our ICB corpus, in our ICBe events, or any of the competing systems.

From them, we draw several stylized facts. First, there is substantial variation in which details any one document will choose to include. Our ground truth ICB narratives included 17/51 and 23/30 of the events from the synthetic narrative, while including other details that are not in the synthetic narrative. Second, mentions of a detail across accounts is expo-

267 nentially distributed with context invariant details appearing  
268 dozens to hundreds of times more than context dependent  
269 details. Third, crisis start and stop dates are arbitrary and  
270 the historical record points to many precursor events as nec-  
271 essary detail for understanding later events, e.g. the U.S. was  
272 in a *de facto* grey scale war with Cuba before it invited Soviet  
273 military protection (61) and Ukraine provided several secu-  
274 rity guarantees to Russia that were potentially undone, e.g. a  
275 long term lease on naval facilities in Crimea. Fourth, we find  
276 variation between the two cases. Cuban Missiles has a cleaner  
277 canonical end with the Soviets agreeing to withdraw missiles  
278 while Crimea-Donbas meekly ends with a second cease fire  
279 agreement (Minsk II) but continued fighting. The canonical  
280 narrative of Cuban Missile also includes high level previously  
281 classified details, while the more recent Crimea-Donbas case  
282 reflects primarily public reporting.

283 We find substantive variation in recall across systems. Recall  
284 for each increases in the number of document mentions which  
285 is an important sign of validity for both them and our bench-  
286 mark. The one outlier is Phoenix which is so noisy that it's  
287 flat to decreasing in mentions. The two episode level datasets  
288 have very low coverage of contextual details. The two other  
289 dictionary systems ICEWs and Terrier have high coverage,  
290 with ICEWs outperforming Terrier. ICBe strictly dominates  
291 all of the systems but ICEWs in recall though we note that  
292 the small sample sizes mean these systems should be consid-  
293 ered statistically indistinguishable. Importantly our corpus of  
294 ICB narratives has very high recall of frequently mentioned  
295 details giving us confidence in how those summaries were con-  
296 structed, and ICBe lags only slightly behind showing that it  
297 left very little additional information on the table.

## 298 Precision

299 The other side of event measurement is precision, the degree  
300 to which a sequence of events correctly and usefully describes  
301 the information in history,  $Pr(H|E)$ . It does little good to  
302 recall a historical event but too vaguely (e.g. MIDs describes  
303 the Cuban Missile crisis as a blockade, a show of force, and  
304 a stalemate) or with too much error (e.g. ICEWS records  
305 263 “Detonate Nuclear Weapons” events between 1995-2019)  
306 to be useful for downstream applications. ICBe’s ontology  
307 and coding system is designed to strike a balance so that the  
308 most important information is recovered accurately but also  
309 abstracted to a level that is still useful and interpretable. You  
310 should be able to lay out events of a crisis on a timeline, as  
311 in Figure 3, and read off the macro structure of an episode  
312 from each individual move. We call this visualization a crisis  
313 map, a directed graph intersected with a timeline, and provide  
314 crisis maps for every event dataset for each case study (SI  
315 Appendix, Fig. D3 and D4) and all crises on the companion  
316 website.

317 We further want to verify individual event codings, which we  
318 can do in the case of ICBe because each event is mapped to  
319 a specific span of text. We develop the iconography system  
320 for presenting event codings as coherent statements that can  
321 be compared side by side to the original source narrative as  
322 for Cuban Missiles (Figure 1), Crimea-Donbas (SI Appendix  
323 Table D1), and for every case on the companion website. We  
324 further provide a stratified sample of event codings alongside  
325 their source text (SI Appendix Table D2).

We find both the visualizations of macro structure and head-  
26 to-head comparisons of ICBe codings to the raw text to  
27 strongly support the quality of ICBe, but as with recall we  
28 seek a more objective detached universal benchmark. Our  
29 proposed measure is a reconstruction task to see whether  
30 our intended ontology can be recovered through only unsup-  
31ervised clustering of sentences they were applied to. Figure  
32 4 shows the location of every sentence from the ICBe corpus  
33 in semantic space as embedded using the same large language  
34 model as before, and the median location of each ICBe event  
35 tag applied to those sentences.<sup>88</sup> Labels reflect the individual  
36 leaves of the ontology and colors reflect the higher level co-  
37 erce branch nodes of the ontology. If ICBe has high precision,  
38 substantively similar tags ought to have been applied to sub-  
39 stantively similar source text, which is what we see both in  
40 two dimensions in the main plot and via hierarchical cluster-  
41 ing on all dimensions in the dendrogram along the righthand  
42 side.<sup>¶¶</sup>

Finally, how does ICBe’s precision compare to the existing  
344 state of the art? The crisis-maps reveal the episode level  
345 datasets like MIDs or the original ICB are too sparse and  
346 vague to reconstruct the structure of the crisis (SI Appendix  
347 Figure D3 and D4). On the other end of the spectrum, the  
348 high recall dictionary based event datasets like Terrier and  
349 ICEWs produce so many noisy events (several hundreds thou-  
350 sand) that even with heavy filtering their crisis maps are  
351 completely unintelligible. Further, because of copyright is-  
352 sues, none of these datasets directly provide the original text  
353 spans making event level precision difficult to verify.

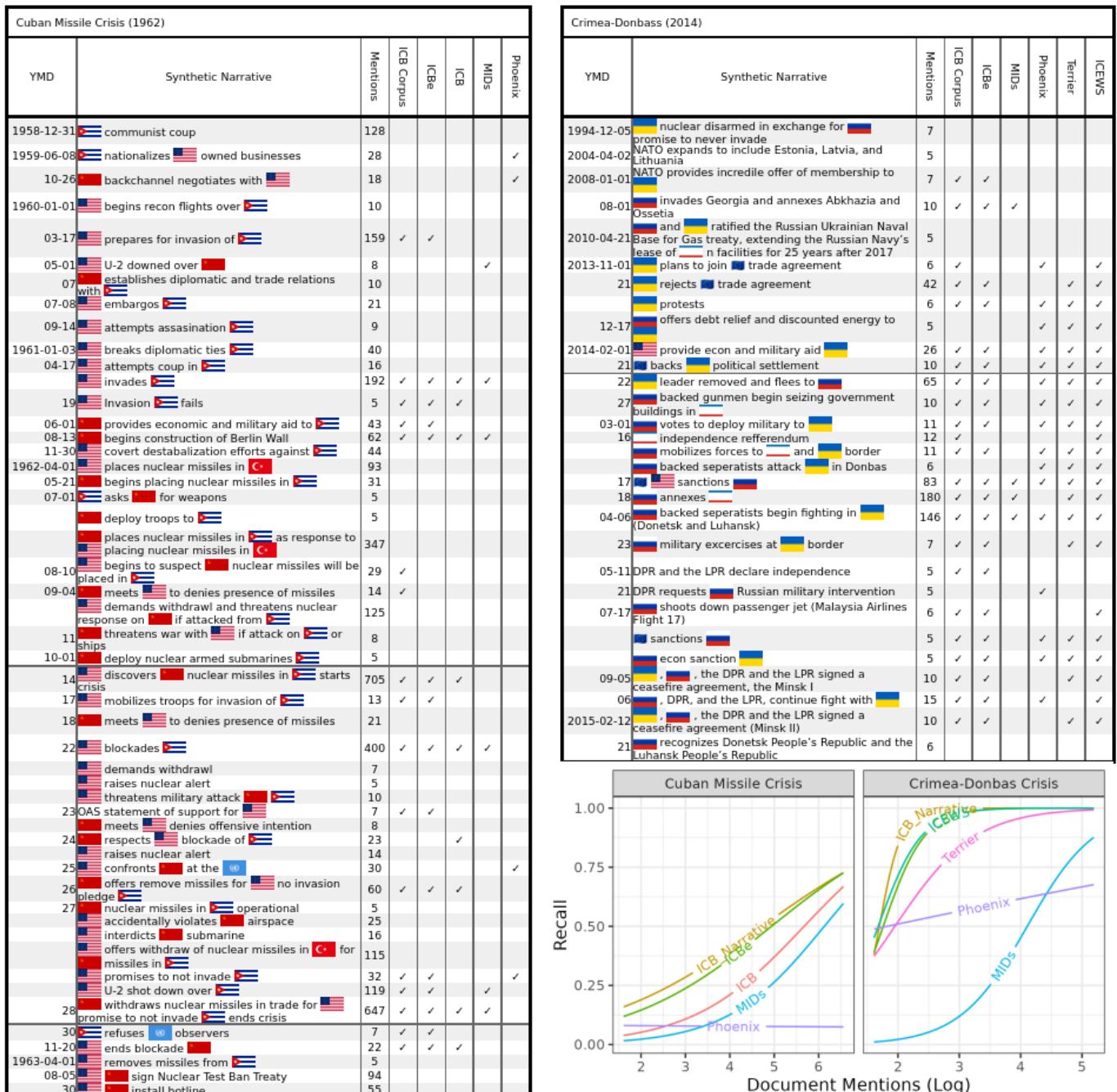
However, given their high recall on our task and the global and  
355 real-time coverage of dictionary based event systems, we want  
356 to take seriously the possibility that some functional transfor-  
357 mation could recover the precision of ICBe. For example, (62)  
358 attempts to correct for the mechanically increasing amount of  
359 news coverage each year by detrending violent event counts  
360 from Phoenix using a human coded baseline. Others have  
361 focused on verifying precision for ICEWs on specific subsets  
362 of details against known ground truths, e.g. geolocation (63),  
363 protest events (80%) (64), anti-government protest networks  
364 (46.1%) (65).

We take the same approach here in Figure 5, selecting four  
366 specific CAMEO event codings and checking how often they  
367 reflect a true real world event. We choose four event types  
368 around key moments in the crisis. The start of the crisis re-  
369 volves around Ukraine backing out of trade deal with the EU  
370 in favor of Russia, but “sign formal agreement” events act  
371 more like a topic detector with dozens of events generated by  
372 discussions of a possible agreement but not the actual agree-  
373 ment which never materialized. The switch is caught by the  
374 “reject plan, agreement to settle dispute”, but also continues  
375 for Victor Yanukovych for even after he was removed from  
376 power because of articles retroactively discussing the cause of  
377 his removal. Events for “use conventional military force”  
378 capture a threshold around the start of hostilities and who the  
379 participants were but not any particular battles or campaigns.  
380 Likewise, “impose embargo, boycott, or sanctions” captures  
381 the start of waves of sanctions and from who but are effectively  
382 constantly as the news coverage does not distinguish between

<sup>88</sup>We preprocess sentences to replace named entities with a generic Entity token.

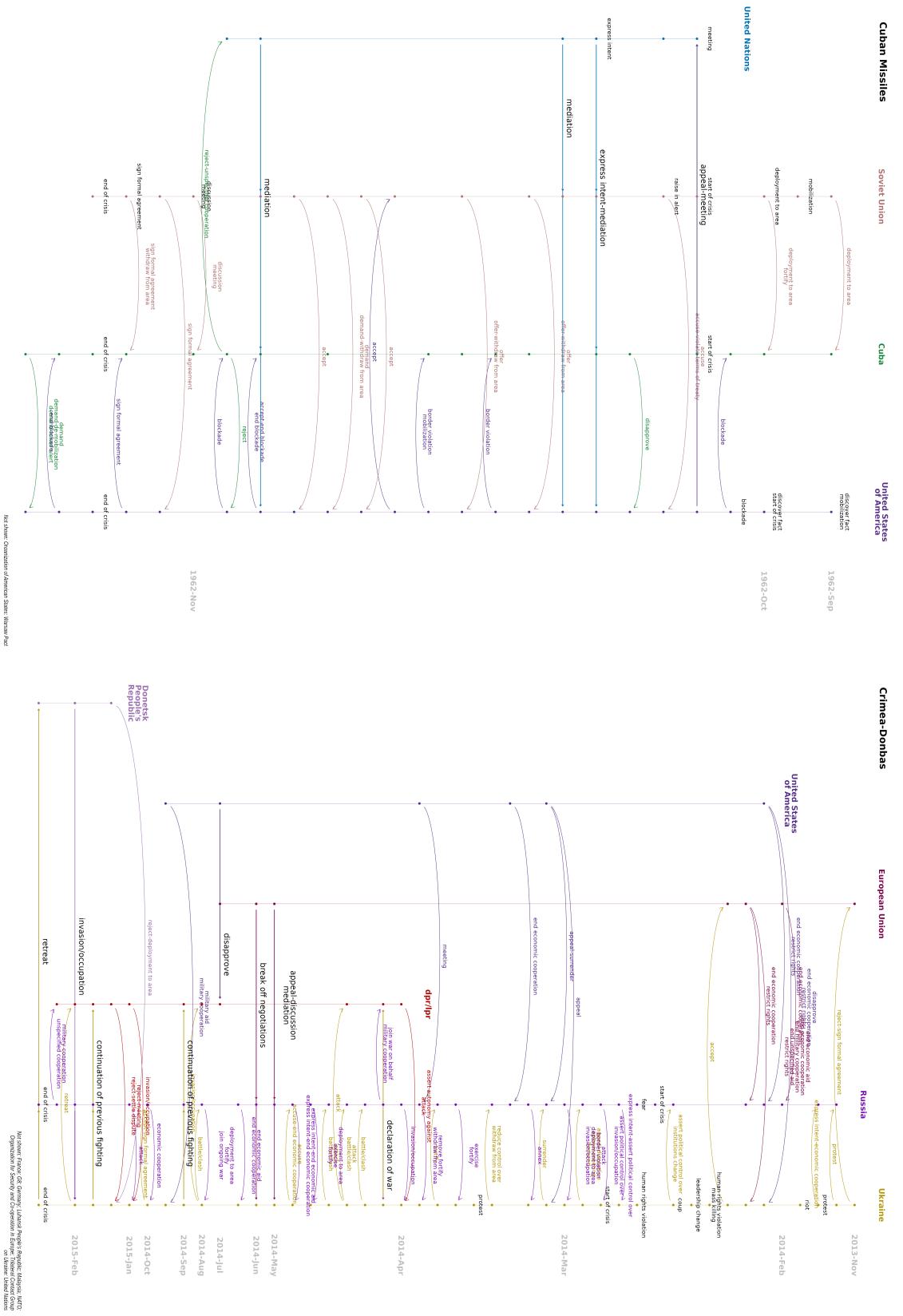
<sup>¶¶</sup>Hierarchical clustering on cosine similarity and with Ward’s method.

**Fig. 2.** Measuring Recall with Synthetic Historical Narratives



Notes: Synthetic narratives combine several thousand accounts of each crisis into a single timeline of events, taking only those mentioned in at least 5 or more documents. Checkmarks represent whether that event could be hand matched to any detail in the ICB corpus, ICBe dataset, or any of the other event datasets.

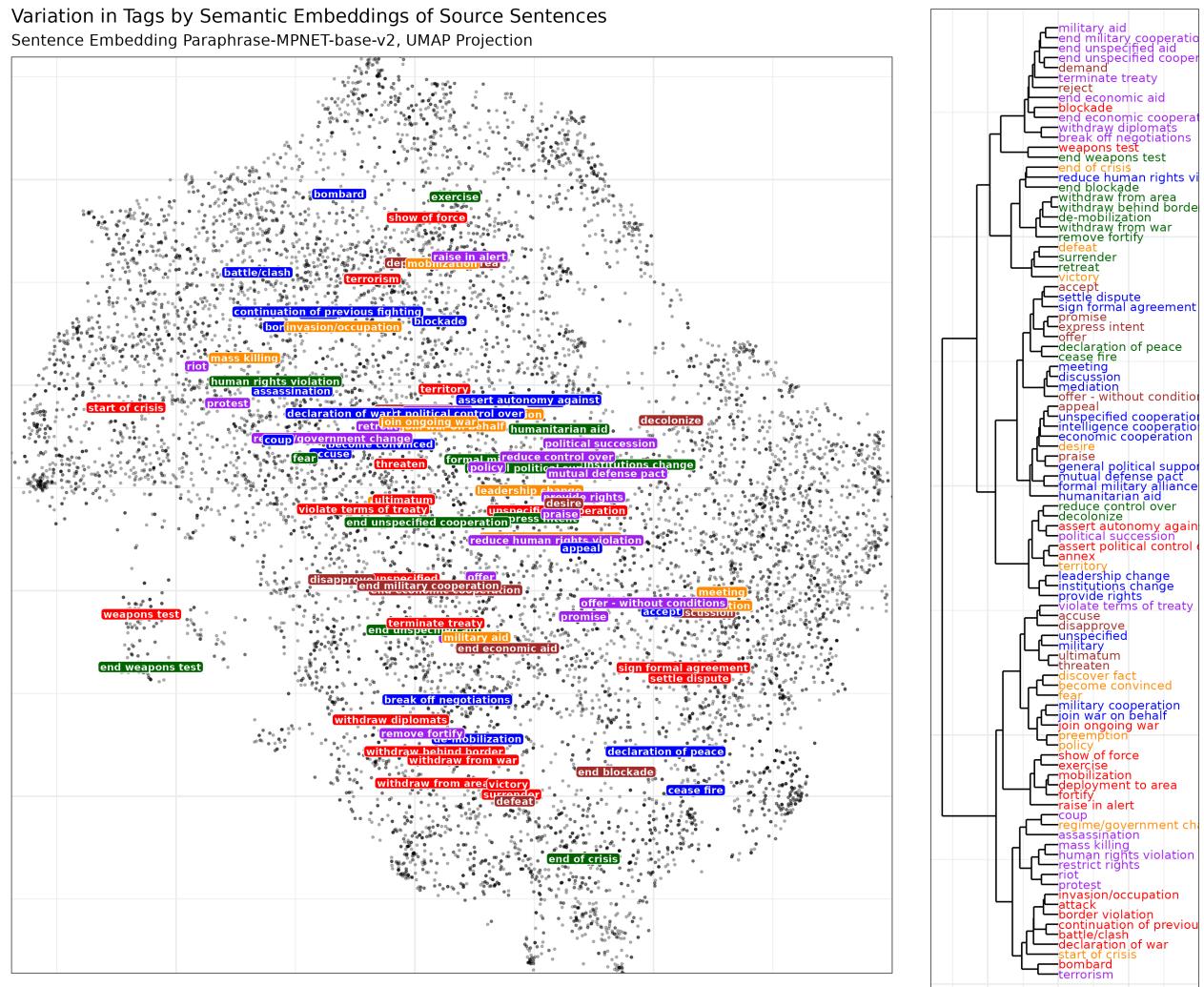
**Fig. 3.** Crisis Maps



**Fig. 4.** ICBe event codings in comparison to Semantic Embeddings from source sentences

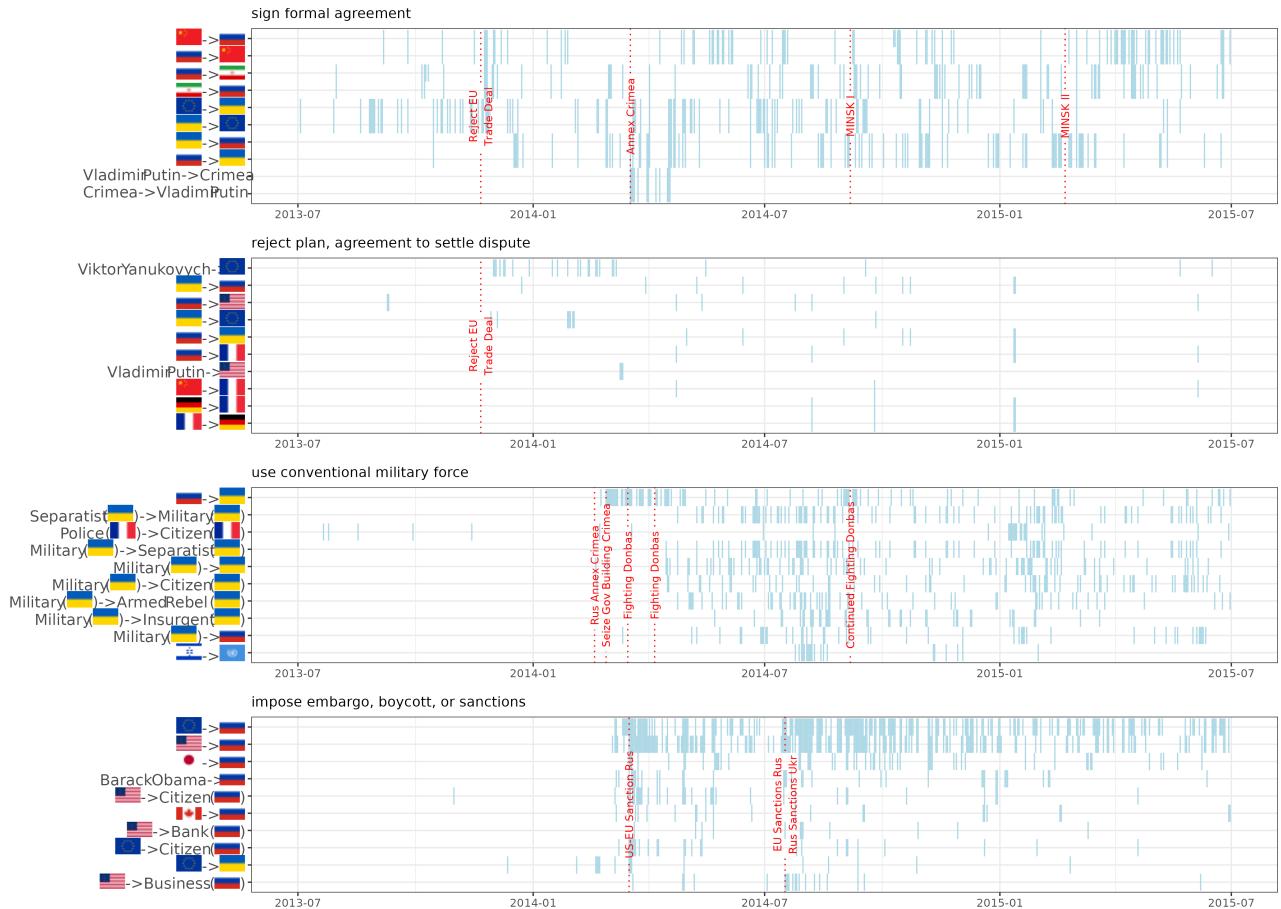
## Variation in Tags by Semantic Embeddings of Source Sentences

Sentence Embedding Paraphrase-MPNET-base-v2, UMAP Projection



*Notes: Dots represent individual ICB narrative sentences, as embeded by the Paraphrase-MPNET-base-v2 large language model and flattened into two dimensions with UMAP. Text labels reflect individual leaves of the ICBe ontology, and colors represent intermediate branches of the ontology. Label placement is the median of all of the sentences that tag was applied to by the coders. The dendrogram shows hierarchical clustering of the tags. If ICBe precision is high, the sentences tags were applied to ought to say similar things, and the intended shape of the ontology ought to be visually recognizable.*

**Fig. 5.** ICEWs Events by Day by Type during the Crimea-Donbas Crisis



Notes: Unit of analysis is the Dyad-Day. Edges  $<->$  indicates undirected dyad and  $->$  indicates directed dyad. Top 10 most active dyads per category shown. Red text shows events from the synthetic narrative relative to that event category. Blue bars indicate an event recorded by ICEWs for that dyad on that day.

subtle changes or additions. In sum, dictionary based methods on news corpora tend to have high recall because they parse everything in the news, but for the same reason their specificity for most event types is too low to back out individual chess like sequencing that ICBe aims to record.

## Conclusion

We investigated event abstraction from narratives describing key historical episodes in international relations. We synthesized a prior belief about the latent unobserved phenomena that drive these events in international relations and proposed a mapping to observable concepts that enter into the observed historical record. We designed an ontology with high coverage over those concepts and developed a training procedure and technical stack for human coding of historical texts. Multiple validity checks find the resulting codings have high internal validity (e.g. intercoder agreement) and external validity (i.e. matching source material in both micro-details at the sentence level and macro-details spanning full historical episodes). Further, these codings perform much better in terms of recall, precision, coverage, and overall coherence in capturing these historical episodes than existing event systems used in international relations.

We release several open-source products along with supporting code and documentation to further advance the study of IR, event extraction, and natural language processing. The first is the International Crisis Behavior Events (ICBe) dataset, an event-level aggregation of what took place during the crises identified by the ICB project. These data are appropriate for statistical analysis of hard questions about the sequencing of events (e.g. escalation and de-escalation of conflicts). Second, we provide a coder-level disaggregation with multiple codings of each sentence by experts and undergrads that allows for the introduction of uncertainty and human interpretation of events. Further, we release a direct mapping from the codings to the source text at the sentence level as a new resource for natural language processing. Finally, we provide a companion website that incorporates detailed visualizations of all of the data introduced here ([www.crisisevents.org](http://www.crisisevents.org)).

## Materials and Methods

This paper's data, online appendix, and replication material as well as visualizations of every historical episode available at a companion website [www.crisisevents.org](http://www.crisisevents.org) and the github repository [ICBEEventData](https://github.com/ICBEEvent/ICBEEventData).

**ACKNOWLEDGMENTS.** We thank the ICB Project and its directors and contributors for their foundational work and their help with this effort. We make special acknowledgment of Michael Brecher for helping found the ICB project in 1975, creating a resource that continues to spark new insights to this day. We thank the many undergraduate coders for their patience and dedication. Thanks to the Center for Peace and Security Studies and its membership for comments. Special thanks to Rebecca Cordell, Philip Schrod, Zachary Steinert-Threlkeld, and Zhanna Terechshenko for generous feedback. Thank you to the cPASS research assistants that contributed to this project: Helen Chung, Daman Heer, Syeda ShahBano Ijaz, Anthony Limon, Erin Ling, Ari Michelson, Prithviraj Pahwa, Gianna Pedro, Tobias Stodieck, Yiyi 'Effie' Sun, Erin Werner, Lisa Yen, and Ruixuan Zhang. This project was supported by a grant from the Office of Naval Research [N00014-19-1-2491]

and benefited from the Charles Koch Foundation's support for the Center for Peace and Security Studies.

1. Brecher M (1999) International studies in the twentieth century and beyond: Flawed dichotomies, synthesis, cumulation: ISA presidential address. *International Studies Quarterly* 43(2):213–264.  
443  
444
2. Reiter D (2015) [Should We Leave Behind the Subfield of International Relations?](#) *Annu Rev Polit Sci* 18(1):481–499.  
445  
448  
449
3. Brecher M, Wilkenfeld J (1982) [Crises in World Politics](#). *World Politics* 34(3):380–417.  
450  
451
4. Beardsley K, James P, Wilkenfeld J, Brecher M (2020) The International Crisis Behavior Project. doi:[10.1093/acrefore/9780190228637.013.1638](https://doi.org/10.1093/acrefore/9780190228637.013.1638).  
452  
453
5. Palmer G, et al. (2021) [The MID5 Dataset, 2011–2014: Procedures, coding rules, and description](#). *Conflict Management and Peace Science*:0738894221995743.  
454  
455
6. Gibler DM (2018) [International Conflicts, 1816–2010: Militarized Interstate Dispute Narratives](#) (Rowman & Littlefield) Available at: [https://books.google.com?id=\\_4VTDwAAQBAJ](https://books.google.com?id=_4VTDwAAQBAJ).  
456  
457
7. Maoz Z, Johnson PL, Kaplan J, Ogunkoya F, Shreve AP (2019) [The Dyadic Militarized Interstate Disputes \(MIDs\) Dataset Version 3.0: Logic, Characteristics, and Comparisons to Alternative Datasets](#). *Journal of Conflict Resolution* 63(3):811–835.  
458  
459
8. Sarkees MR, Wayman F (2010) [Resort to war: 1816–2007](#) (CQ Press).  
460  
461
9. Reiter D, Stam AC, Horowitz MC (2016) [A Revised Look at Interstate Wars, 1816–2007](#). *Journal of Conflict Resolution* 60(5):956–976.  
462  
463
10. Ralph Sundberg, Mihael Croicu (2016) [UCDP GED Codebook version 5.0](#) (Department of Peace and Conflict Research, Uppsala University).  
464  
465
11. Pettersson T, Eck K (2018) [Organized violence, 1989–2017](#). *Journal of Peace Research* 55(4):535–547.  
466  
467
12. Raleigh C, Linke A, Hegre H, Karlsen J (2010) Introducing ACLED: An armed conflict location and event dataset: Special data feature. *Journal of peace research* 47(5):651–660.  
468  
469
13. Felbermayr G, Kirilakha A, Syropoulos C, Yalcin E, Yotov YV (2020) [The global sanctions data base](#). *European Economic Review* 129:103561.  
470  
471
14. Barari S, Kim IS Democracy and Trade Policy at the Product Level: Evidence from a New Tariff-line Dataset. 16.  
472
15. Kinne BJ (2020) [The Defense Cooperation Agreement Dataset \(DCAD\)](#). *Journal of Conflict Resolution* 64(4):729–755.  
473  
474  
475
16. Owsiaik AP, Cuttner AK, Buck B (2018) [The International Border Agreements Dataset](#). *Conflict Management and Peace Science* 35(5):559–576.  
476  
477
17. Vabulas F, Snidal D (2021) [Cooperation under autonomy: Building and analyzing the Informal Intergovernmental Organizations 2.0 dataset](#). *Journal of Peace Research* 58(4):859–869.  
478  
479
18. Frederick BA, Hensel PR, Macaulay C (2017) [The Issue Correlates of War Territorial Claims Data, 1816–20011](#). *Journal of Peace Research* 54(1):99–108.  
480  
481
19. Moyer JD, Turner SD, Meisel CJ (2020) [What are the drivers of diplomacy? Introducing and testing new annual dyadic data measuring diplomatic exchange](#). *Journal of Peace Research*:0022343320929740.  
482  
483
20. Sechser TS (2011) [Militarized Compellent Threats, 1918–2001](#). *Conflict Management and Peace Science* 28(4):377–401.  
484  
485
21. Li Q, et al. (2021) A Comprehensive Survey on Schema-based Event Extraction with Deep Learning. Available at: <http://arxiv.org/abs/2107.02126> [Accessed September 10, 2021].  
486  
487

- 488 22. Halterman A (2020) Extracting Political Events from Text  
489 Using Syntax and Semantics. 538
- 490 23. Brandt PT, D'Orazio V, Holmes J, Khan L, Ng V (2018)  
491 Phoenix Real-Time Event Data. Available at: <http://eventdata.utdallas.edu>. 539
- 492 24. Boschee E, et al. (2015) ICEWS coded event data. *Harvard*  
493 *Dataverse* 12. 540
- 494 25. Hegre H, Croicu M, Eck K, Höglbladh S (2020) Introducing  
495 the UCDP Candidate Events Dataset. *Research & Politics*  
7(3):2053168020935257. 541
- 496 26. Grant C, Halterman A, Irvine J, Liang Y, Jabr K (2017)  
497 OU Event Data Project. Available at: <https://osf.io/4m2u7/>  
[Accessed September 1, 2021]. 542
- 498 27. Zhang H, Pan J (2019) **CASM: A Deep-Learning Approach**  
499 for Identifying Collective Action Events with Text and Im-  
500 age Data from Social Media. *Sociological Methodology*  
49(1):1–57. 543
- 500 28. Steinert-Threlkeld ZC (2019) **The Future of Event Data Is**  
501 **Images**. *Sociological Methodology* 49(1):68–75. 544
- 502 29. Brecher M, James P, Wilkenfeld J (2000) Crisis escalation  
503 to war: Findings from the International Crisis Behavior  
504 Project. *What Do We Know About War*. 545
- 505 30. Wilkenfeld J, Brecher M (2000) Interstate crises and vio-  
506 lence: Twentieth-century findings. *Handbook of war studies*  
II:282–300. 546
- 507 31. James P (2019) **What do we know about crisis, escalation**  
508 and war? **A visual assessment of the International Crisis**  
509 **Behavior Project**. *Conflict Management and Peace Science*  
36(1):3–19. 547
- 510 32. Iakhnis E, James P (2019) **Near crises in world politics:**  
511 **A new dataset**. *Conflict Management and Peace Sci-  
512 ence*:0738894219855610. 548
- 513 33. Brecher M, Wilkenfeld J, Beardsley KC, James P, Quinn D  
514 (2017) *International Crisis Behavior Data Codebook* Avail-  
515 able at: <http://sites.duke.edu/icbdata/data-collections/>. 549
- 516 34. Brecher M, Wilkenfeld J (1997) *A Study of Crisis* (Univer-  
517 sity of Michigan Press). 550
- 518 35. Holsti OR (1965) **The 1914 Case**. *The American Political*  
519 *Science Review* 59(2):365–378. 551
- 520 36. Paige GD (1968) *The Korean Decision, June 24–30, 1950*  
521 (Free Press). 552
- 522 37. Allison GT, Zelikow P (1971) *Essence of decision: Explain-  
523 ing the Cuban missile crisis* (Little, Brown Boston). 553
- 524 38. Snyder GH, Diesing P (1977) *Conflict among nations: Bar-  
525 gaining and decision making in international crises* (Princeton  
526 University Press). 554
- 527 39. Gavin FJ (2014) **History, Security Studies, and the July**  
528 **Crisis**. *Journal of Strategic Studies* 37(2):319–331. 555
- 529 40. George AL, Smoke R (1974) *Deterrence in American for-  
530 eign policy: Theory and practice* (Columbia University  
531 Press). 556
- 532 41. Gaddis JL (1987) **Expanding the Data Base: Historians,**  
533 **Political Scientists, and the Enrichment of Security Studies**.  
534 *International Security* 12(1):3–21. 557
- 535 42. Brecher M, James P (1988) Patterns of crisis management.  
536 *Journal of Conflict Resolution* 32(3):426–456. 558
- 537 43. Kang DC, Lin AY-T (2019) US bias in the study of Asian  
538 security: Using Europe to study Asia. *Journal of Global*  
539 *Security Studies* 4(3):393–401. 560
- 540 44. Hewitt JJ (2001) **Engaging International Data in the Class-  
541 room: Using the ICB Interactive Data Library to Teach**  
542 **Conflict and Crisis Analysis**. *Int Stud Perspect* 2(4):371–  
543 383. 561
- 544 45. Miller GA (1995) **WordNet: A lexical database for English**.  
545 *Commun ACM* 38(11):39–41. 562
- 546 46. Althaus S, Bajjaleh J, Carter JF, Peyton B, Shalmon DA  
547 (2019) Cline Center Historical Phoenix Event Data Variable  
548 Descriptions. *Cline Center Historical Phoenix Event Data*. 563
- 549 47. Sundberg R, Melander E (2013) Introducing the UCDP  
550 georeferenced event dataset. *Journal of Peace Research*  
551 50(4):523–532. 564
- 552 48. Leng RJ, Singer JD (1988) **Militarized Interstate Crises: The BCOW Typology and Its Applications**. *International Studies Quarterly* 32(2):155–173. 565
- 553 49. McClelland C (1978) World event/interaction survey, 1966–  
554 1978. *WEIS Codebook ICPSR* 5211. 566
- 555 50. Hermann C (1984) Comparative Research on the Events  
556 of Nations (CREON) Project: Foreign Policy Events, 1959–  
557 1968: Version 1. doi:[10.3886/ICPSR05205.V1](https://doi.org/10.3886/ICPSR05205.V1). 567
- 558 51. Bloomfield LP, Moulton A (1989) CASCON III: Computer-  
559 aided system for analysis of local conflicts. *MIT Center for*  
560 *International Studies, Cambridge*. 568
- 561 52. Sherman FL (2000) SHERFACS: A Cross-Paradigm,  
562 Hierarchical, and Contextually-Sensitive Interna-  
563 tional Conflict Dataset, 1937–1985: Version 1.  
564 doi:[10.3886/ICPSR02292.V1](https://doi.org/10.3886/ICPSR02292.V1). 565
- 565 53. Balali A, Asadpour M, Jafari SH (2021) COFEE: A  
566 Comprehensive Ontology for Event Extraction from text.  
567 doi:[10.48550/arXiv.2107.10326](https://doi.org/10.48550/arXiv.2107.10326). 568
- 568 54. Merritt RL (1994) Measuring events for international polit-  
569 ical analysis. *International Interactions* 20(1–2):3–33. 570
- 570 55. Schrot PA, Hall B (2006) Twenty years of the Kansas event  
571 data system project. *The political methodologist* 14(1):2–8. 572
- 572 56. Brecher M, Wilkenfeld J, Beardsley K, James P, Quinn D  
573 International Crisis Behavior Data Codebook, Version 12. 574
- 574 57. Braithwaite A (2010) **MIDLOC: Introducing the Militarized**  
575 **Interstate Dispute Location dataset**. *Journal of Peace Re-  
576 search* 47(1):91–98. 577
- 577 58. Braithwaite A (2009) Codebook for the Militarized Inter-  
578 state Dispute Location (MIDLOC) Data, v 1.0. *University*  
579 *College London*. 580
- 580 59. Brust C-A, Denzler J (2020) Integrating domain knowledge:  
581 Using hierarchies to improve deep classifiers. Available at:  
582 <http://arxiv.org/abs/1811.07125> [Accessed September 7, 2021]. 583
- 583 60. Mandelbrot BB (1983) *The fractal geometry of nature* (Free-  
584 man, New York). 585
- 585 61. Cormac R, Aldrich RJ (2018) **Grey is the new black: Covert**  
586 **action and implausible deniability**. *International Affairs*  
587 94(3):477–494. 588
- 588 62. Terechshenko Z (2020) **Hot under the collar: A latent mea-  
589 sure of interstate hostility**. *Journal of Peace Research*  
590 57(6):764–776. 591
- 591 63. Cook SJ, Weidmann NB (2019) **Lost in Aggregation: Im-  
592 proving Event Analysis with Report-Level Data**. *American*  
593 *Journal of Political Science* 63(1):250–264. 594
- 594 64. Wüest B, Lorenzini J (2020) External validation of protest  
595 event analysis. *Contention in times of crisis: Recession*  
596 *and political protest in thirty European countries*:49–78. 597
- 597 65. Jäger K The Limits of Studying Networks with Event Data:  
598 Evidence from the ICEWS Dataset. 599