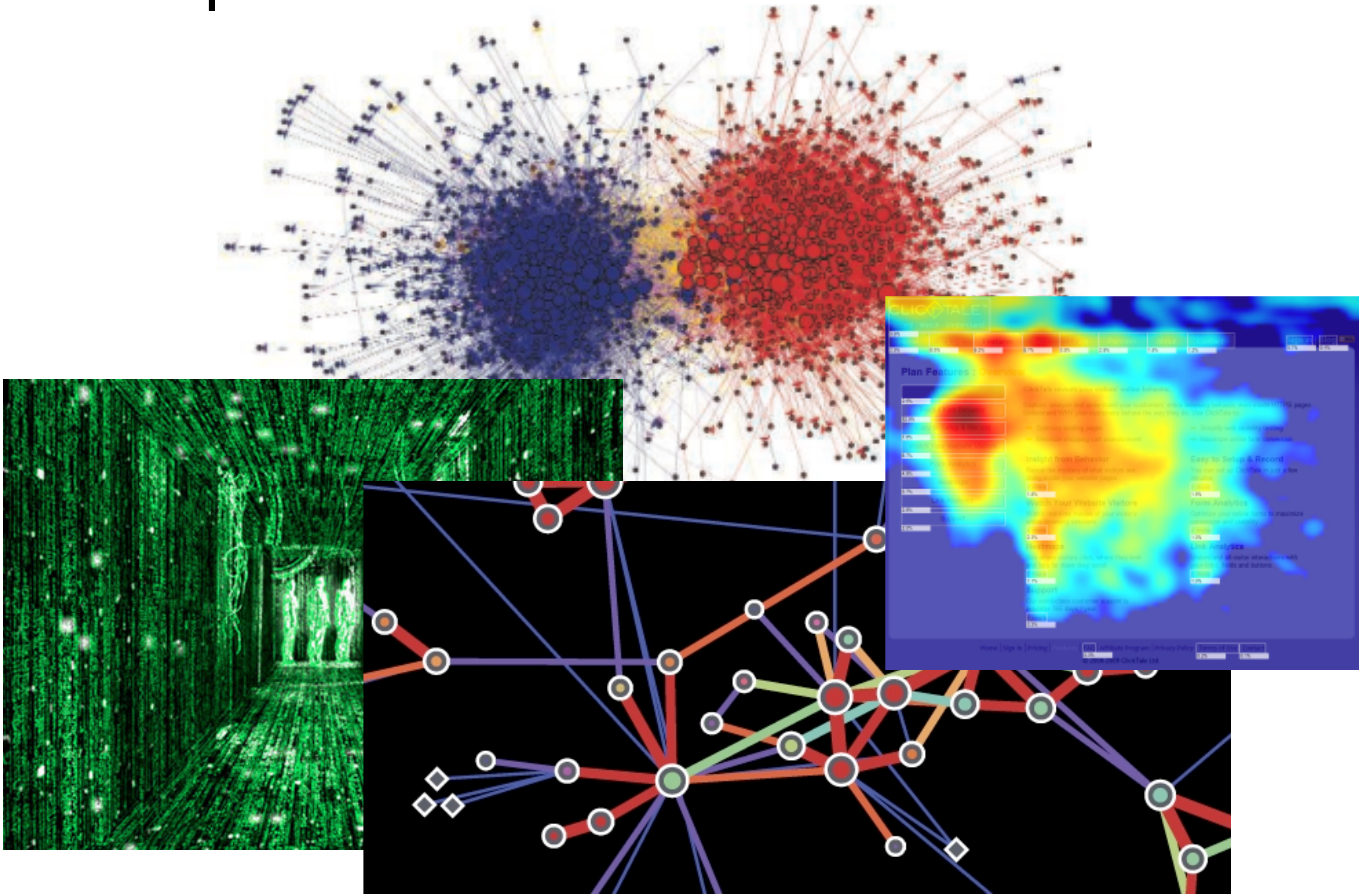
Learning to Extract International Relations from News Text

Brendan O'Connor
Machine Learning Department
Carnegie Mellon University
Presentation: NSF SOCS Doctoral Symposium, June 27, 2013

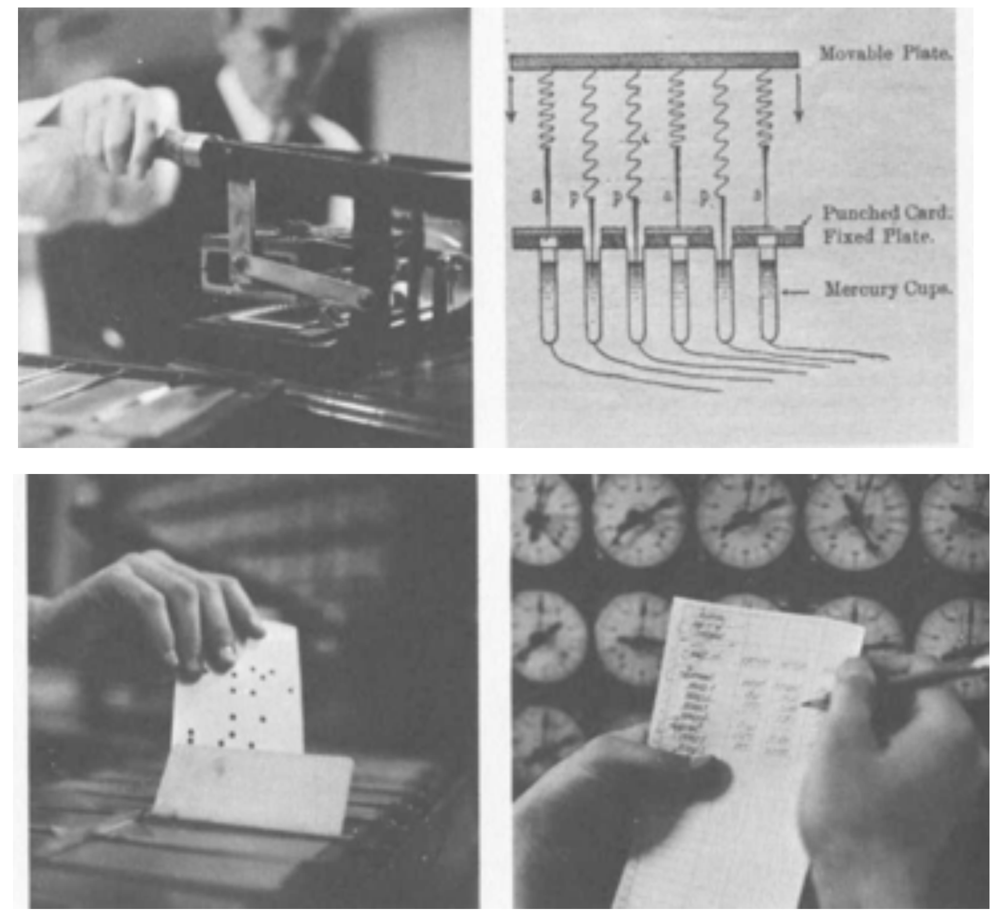
Forthcoming, ACL 2013. Joint work with
Brandon Stewart (political science, Harvard)
Noah Smith (CMU)

Paper and other information at: http://brenocon.com/irevents/

Computational Social Science



Computational Social Science



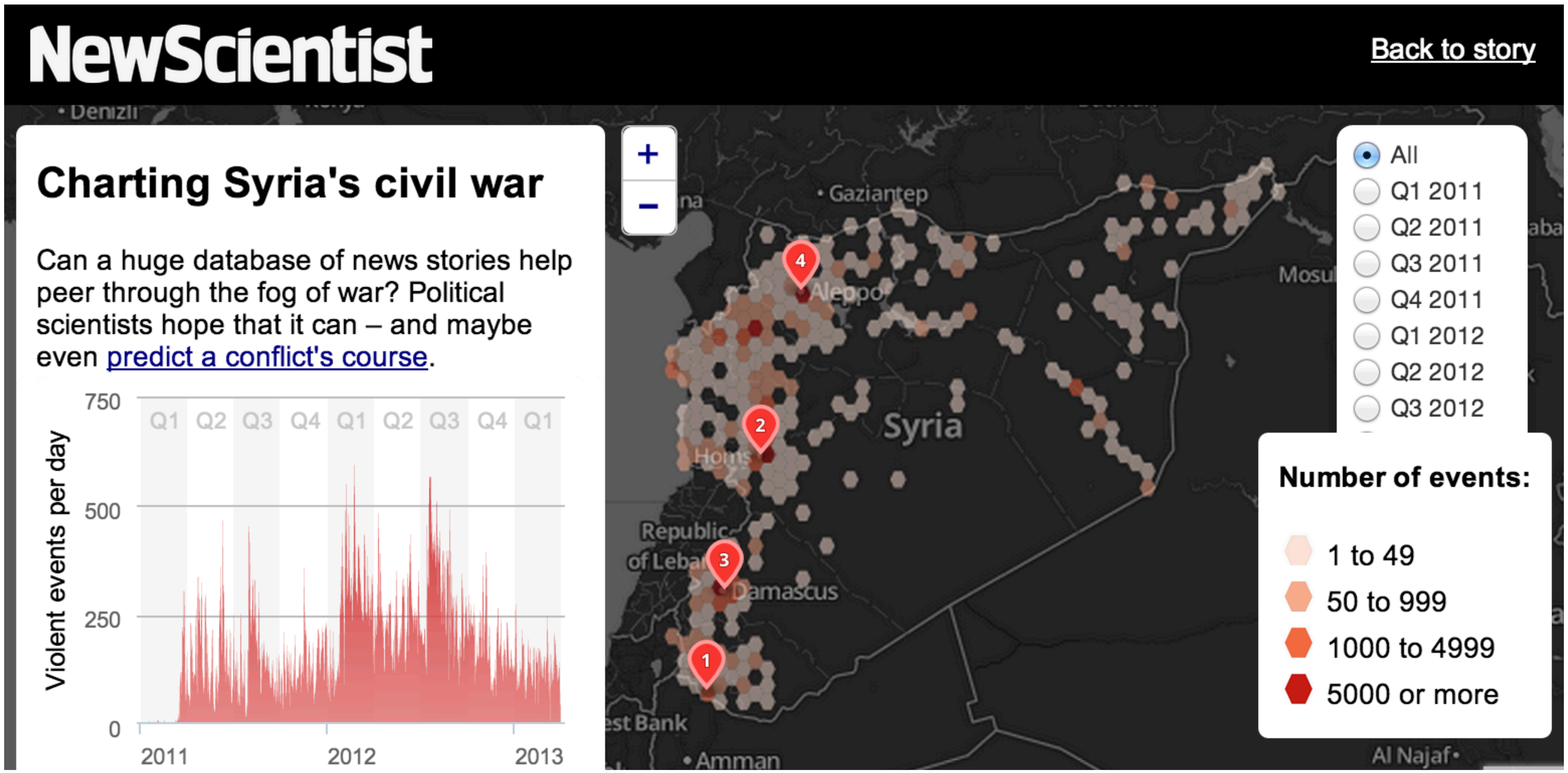
1890 Census tabulator - solved 1880’s data deluge

Computation ***<u>as a tool</u>*** for social science applications

Automated Text Analysis

- Textual media: news, books, articles, internet, messages...
- Automated content analysis:
  tools for *discovery* and *measurement* of concepts, attitudes, events
- Natural language processing, information retrieval, data mining, and machine learning as quantitative social science methodology

International Relations Event Data



Extracted from news text

http://gdelt.utdallas.edu

Previous work: knowledge engineering approach

Open-source TABARI software and ontology/patterns

~15000 verb patterns, ~200 event classes

(Schrodt 1994..2012; ontology goes back to 1960's)

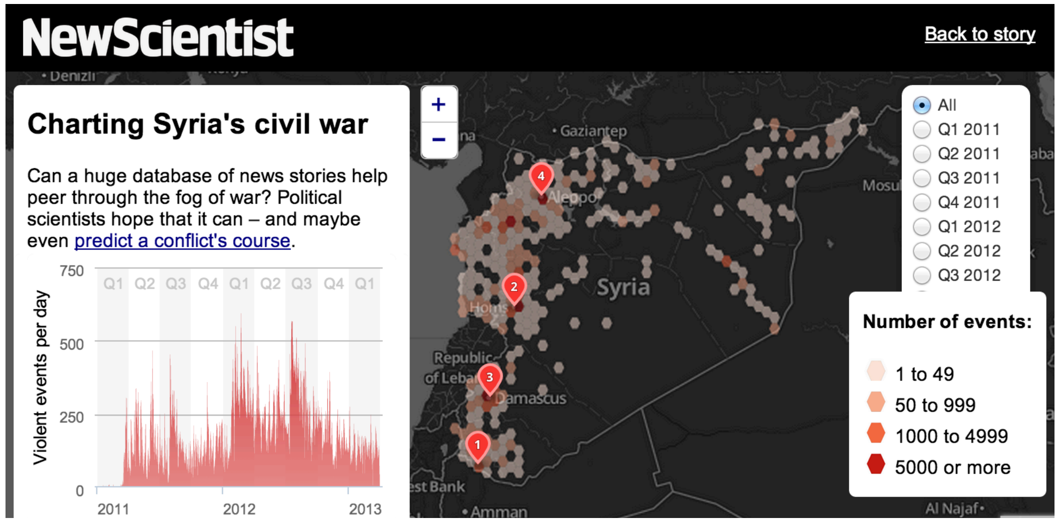
03 - EXPRESS INTENT TO COOPERATE
07 - PROVIDE AID
15 - EXHIBIT MILITARY POSTURE

Event types

191 - Impose blockade, restrict movement

not_ allow to_ enter ;mj 02 aug 2006
barred travel
block traffic from ;ab 17 nov 2005
block road ;hux 1/7/98

Verb patterns per event type



Extract events from news text

Previous work: knowledge engineering approach

Open-source TABARI software and ontology/patterns

~15000 verb patterns, ~200 event classes

(Schrodt 1994..2012; ontology goes back to 1960's)

03 - EXPRESS INTENT TO COOPERATE
07 - PROVIDE AID
15 - EXHIBIT MILITARY POSTURE

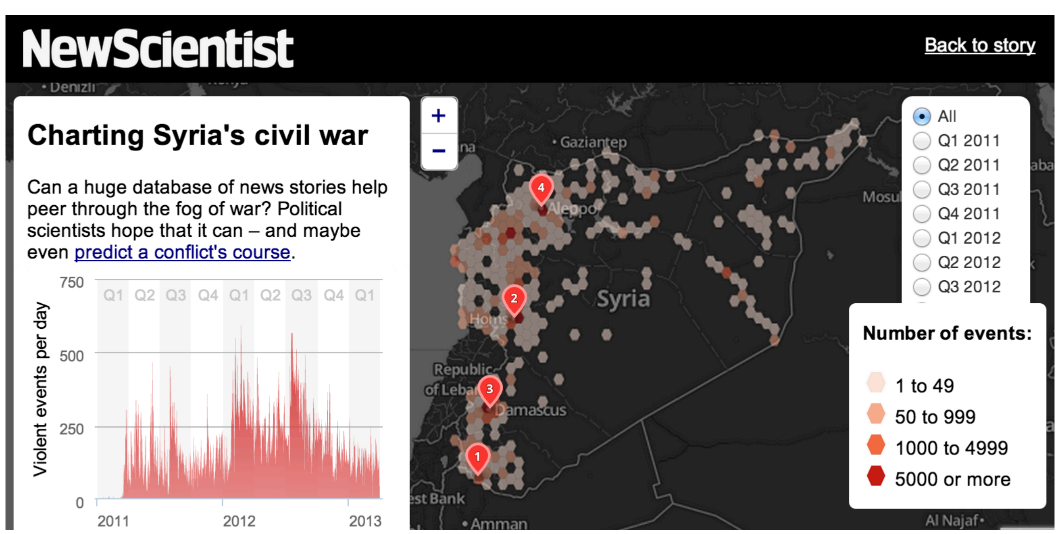
← Event types

191 - Impose blockade, restrict movement

not_ allow to_ enter ;mj 02 aug 2006
barred travel
block traffic from ;ab 17 nov 2005
block road ;hux 1/7/98

← Verb patterns per event type

Extract events from news text



Issues:

1. Hard to maintain and adapt to new domains
2. Precision is low (Boschee et al 2013)

Our approach

- Joint learning for high-level summary of event timelines
    - 1. Automatically learn the verb ontology
    - 2. Extract events / political dynamics
- Social context to drive unsupervised learning about language

Newswire entity/predicate data

- 6.5 million news articles, 1987-2008
- Focus on events between two actors:
  (*SourceEntity*, *ReceiverEntity*, *Time*, $w_{predpath}$)
- “Pakistan promptly accused India” [1/1/2000]
  => (PAK, IND, 268, *SRC -nsubj> accuse <dobj- REC*)
- Named entities: dictionary of country names
- Predicate paths: where verb dominates *Source* in subject position. *Receiver* most commonly directobj, prepobj constructions (some others too)

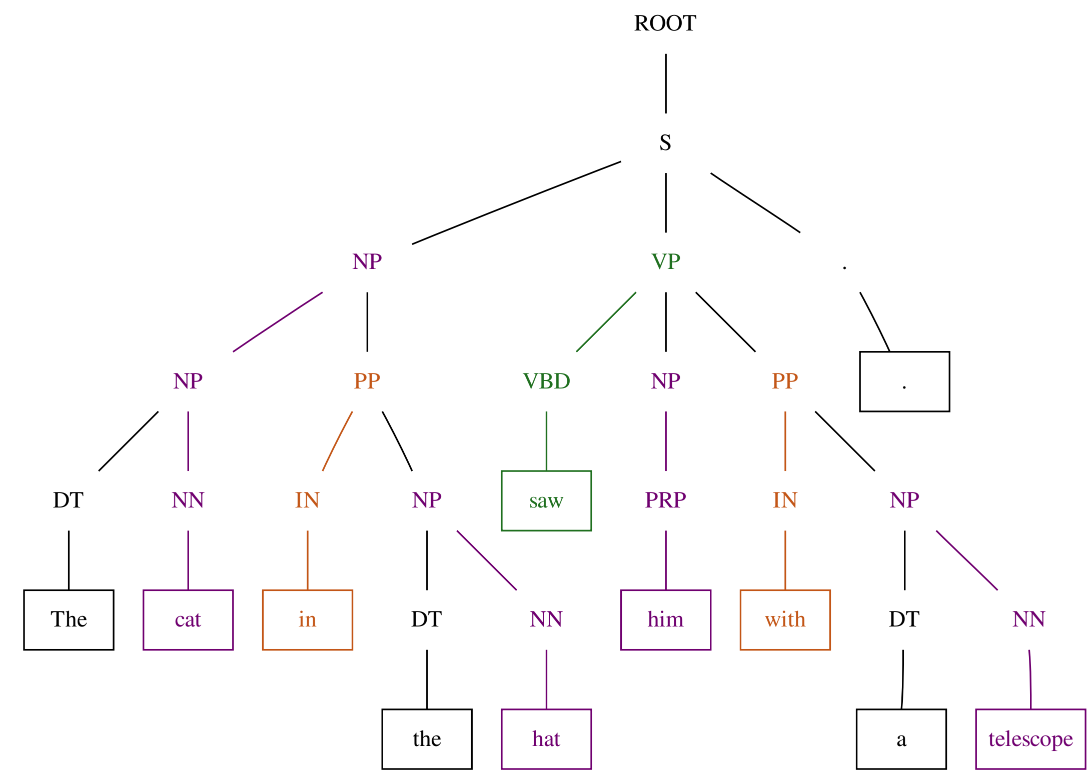
APW_ENG_20080306.0529 -
In other violence Thursday , Israeli forces attacked a rocket-launching site in the northern Gaza Strip , killing one military , Palestinian medics said .

S31 src=ISR rec=PSE pred=[["A","semagent","->"], ["W","attack","verb"],["A","prep_in","<-"]]

Newswire entity/predicate data

Very rare to see parsers in text-as-data studies.
Parsers are slow, hard to use, and make errors.

- Entities: as noun phrases
- Events: as verbs and arguments
- Co-occurrence has low precision
- Preprocess with Stanford CoreNLP for part-of-speech tags and syntactic dependencies
- Filters for topics, factivity, verb-y paths, and parse quality
- Makes unsupervised learning easier: verb-argument information decides which words represent the event



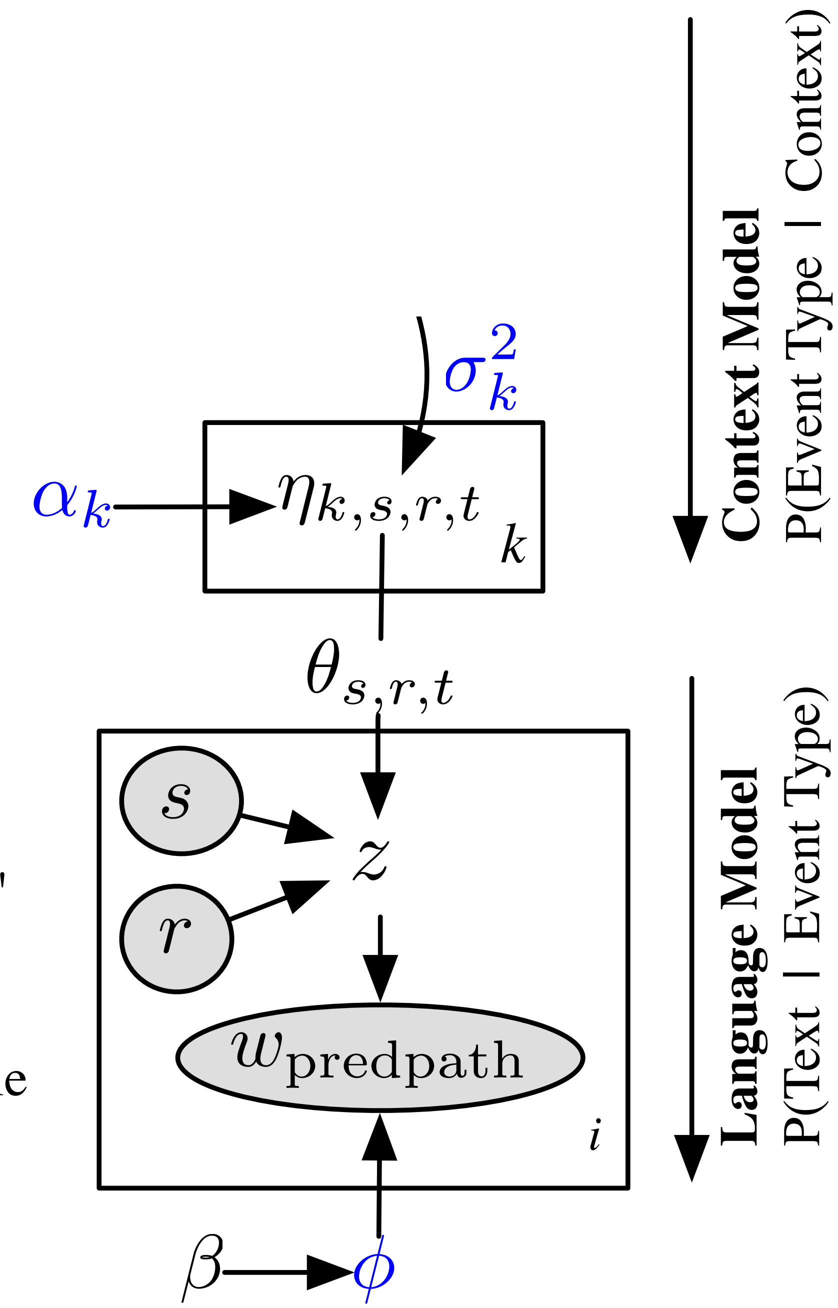
Model

Vanilla Model

Independent contexts

Frame learning from verb co-occurrence within contexts

One (s,r,t) slice of graphical model



s "Source" entity
r "Receiver" entity
t Timestep
i Event tuple
k Frame

Frame prior log-odds in this context

Overall prevalence of frame (event class)

$$\eta_{k,s,r,t} \sim N(\alpha_k,\ \sigma_k^2)$$

Frame prior in this context

$$\theta_{k,s,r,t} = \frac{\exp \eta_{k,s,r,t}}{\sum_{k'=1}^{K} \exp \eta_{k',s,r,t}}$$

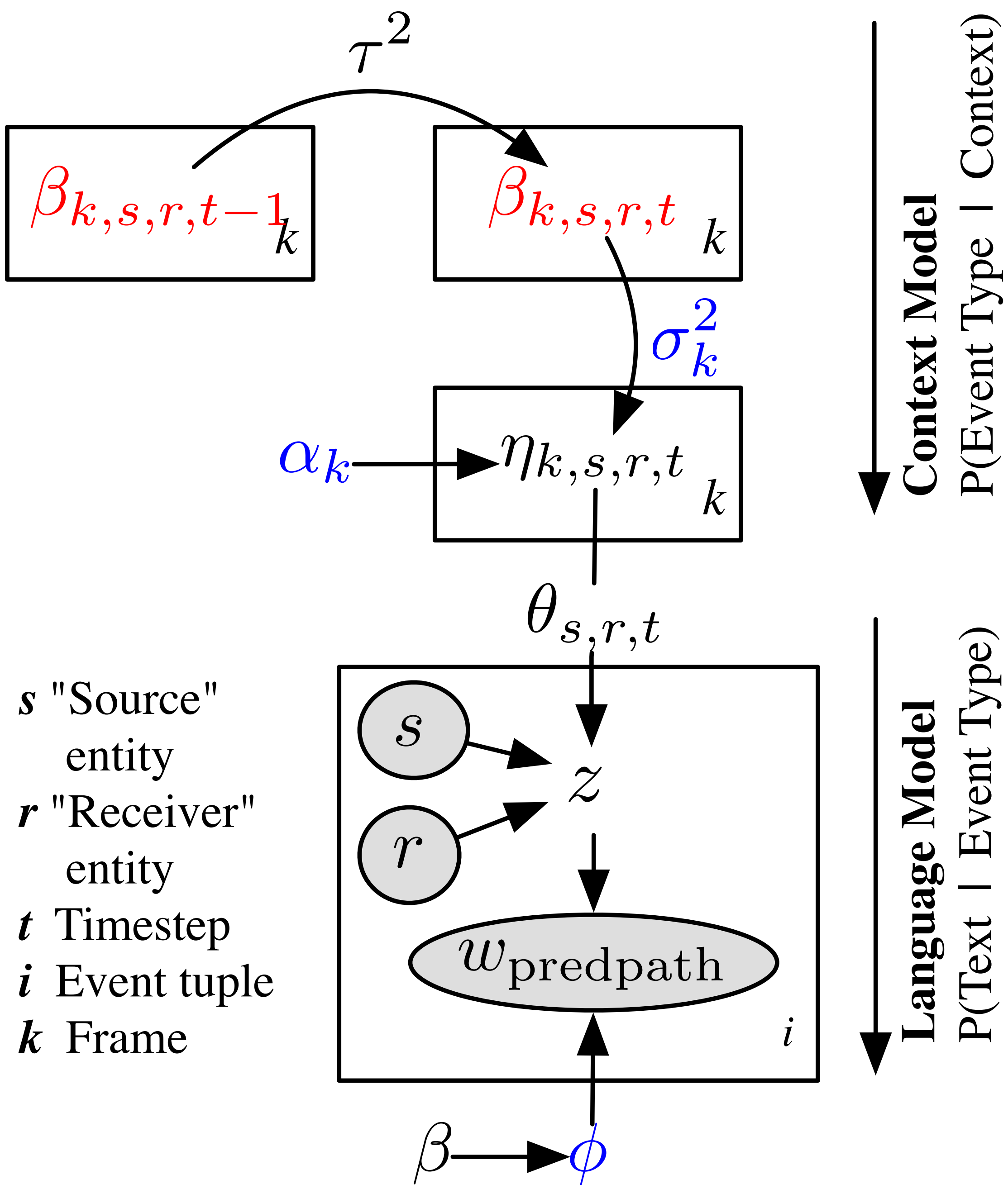
$$z \sim \theta_{s,r,t}$$

$$w \sim \phi_z$$

Training: blocked Gibbs sampling
(Markov Chain Monte Carlo)

Model

One (s,r,t) slice of graphical model



Smoothed Model

Linear dynamical system (Random walk)

$$\beta_{k,s,r,1} \sim N(0, 100)$$
$$\beta_{k,s,r,t} \sim N(\beta_{k,s,r,t-1},\ \tau^2)$$
$$\eta_{k,s,r,t} \sim N(\alpha_k + \beta_{k,s,r,t},\ \sigma_k^2)$$

Vanilla Model

Independent contexts

Frame learning from verb co-occurrence within contexts

Frame prior log-odds in this context

Overall prevalence of frame (event class)

$$\eta_{k,s,r,t} \sim N(\alpha_k,\ \sigma_k^2)$$

Frame prior in this context

$$\theta_{k,s,r,t} = \frac{\exp \eta_{k,s,r,t}}{\sum_{k'=1}^{K} \exp \eta_{k',s,r,t}}$$

$$z \sim \theta_{s,r,t}$$
$$w \sim \phi_z$$

Training: blocked Gibbs sampling (Markov Chain Monte Carlo)

Learned Event Types

"diplomacy"

arrive in, visit, meet with, travel to, leave, hold with, meet, meet in, fly to, be in, arrive for talk with, say in, arrive with, head to, hold in, due in, leave for, make to, arrive to, praise

"verbal conflict"

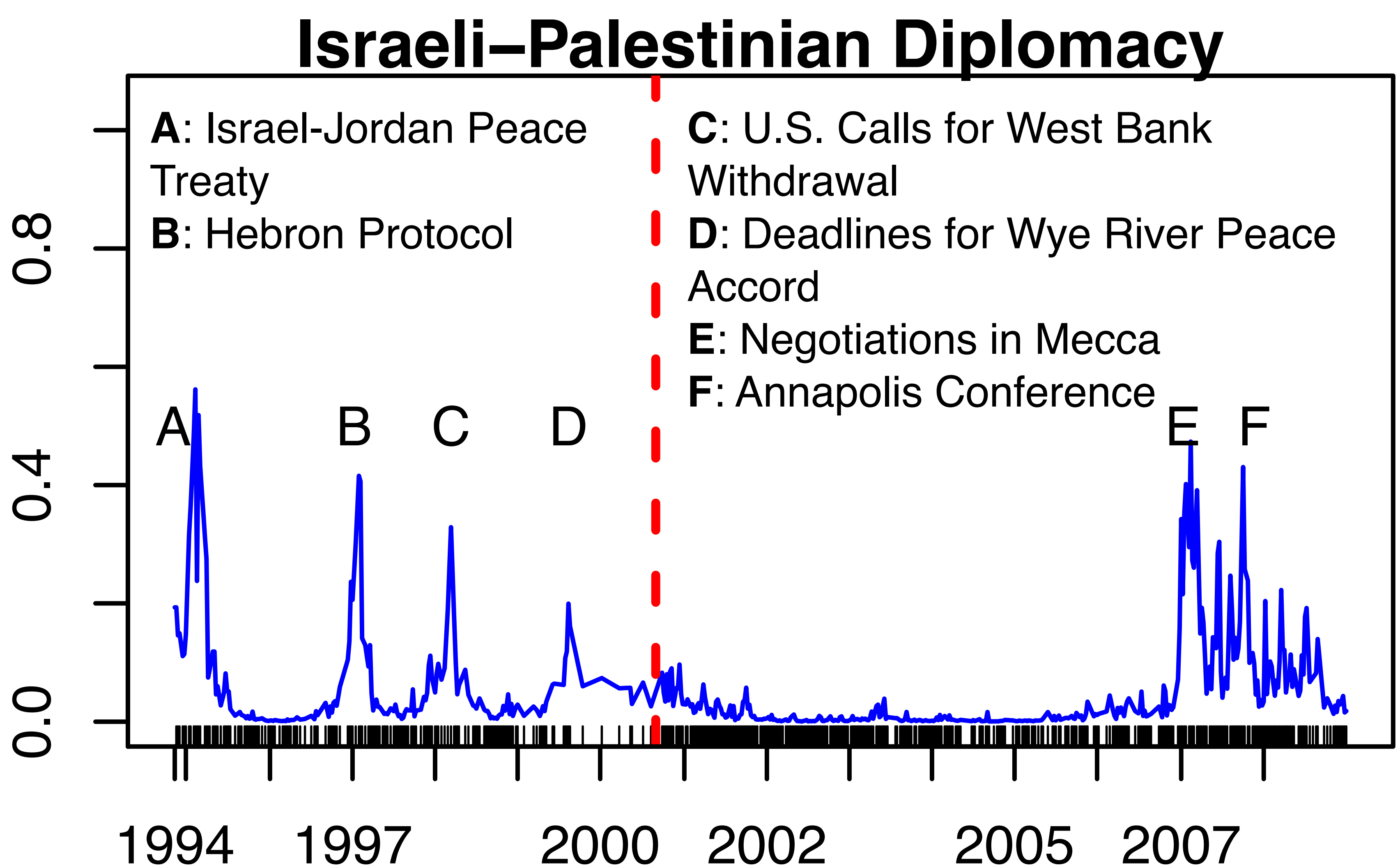
accuse, blame, say, break with, sever with, blame on, warn, call, attack, rule with, charge, say←ccomp come from, say ←ccomp, suspect, slam, accuse government ←poss, accuse agency ←poss, criticize, identify

"material conflict"

kill in, have troops in, die in, be in, wound in, have soldier in, hold in, kill in attack in, remain in, detain in, have in, capture in, stay in, about ←pobj troops in, kill, have troops ←partmod station in, station in, injure in, invade, shoot in

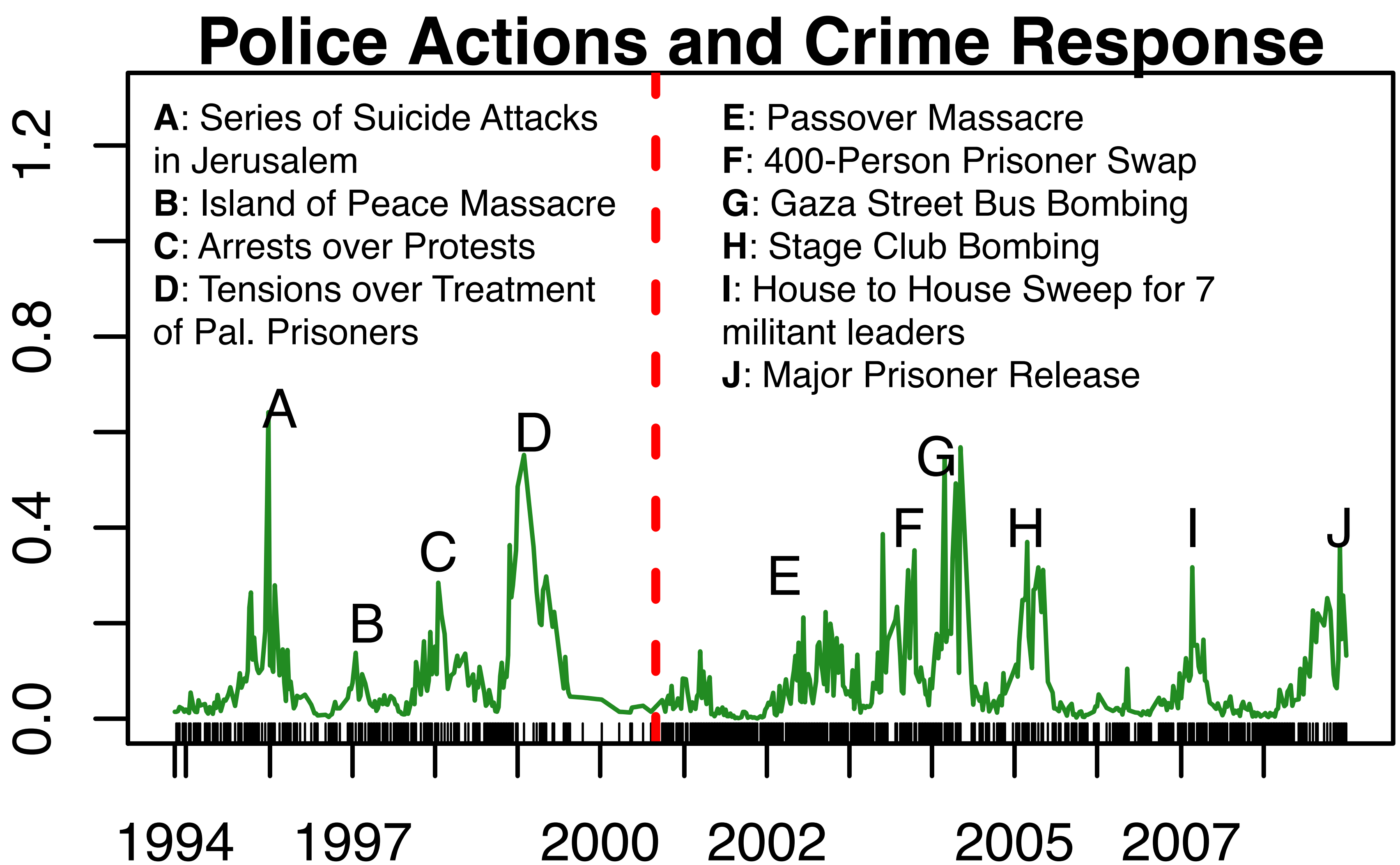
Case study

meet with, sign with, praise, say with, arrive in, host, tell, welcome, join, thank, meet, travel to, criticize, leave, take to, begin to, begin with, summon, reach with, hold with



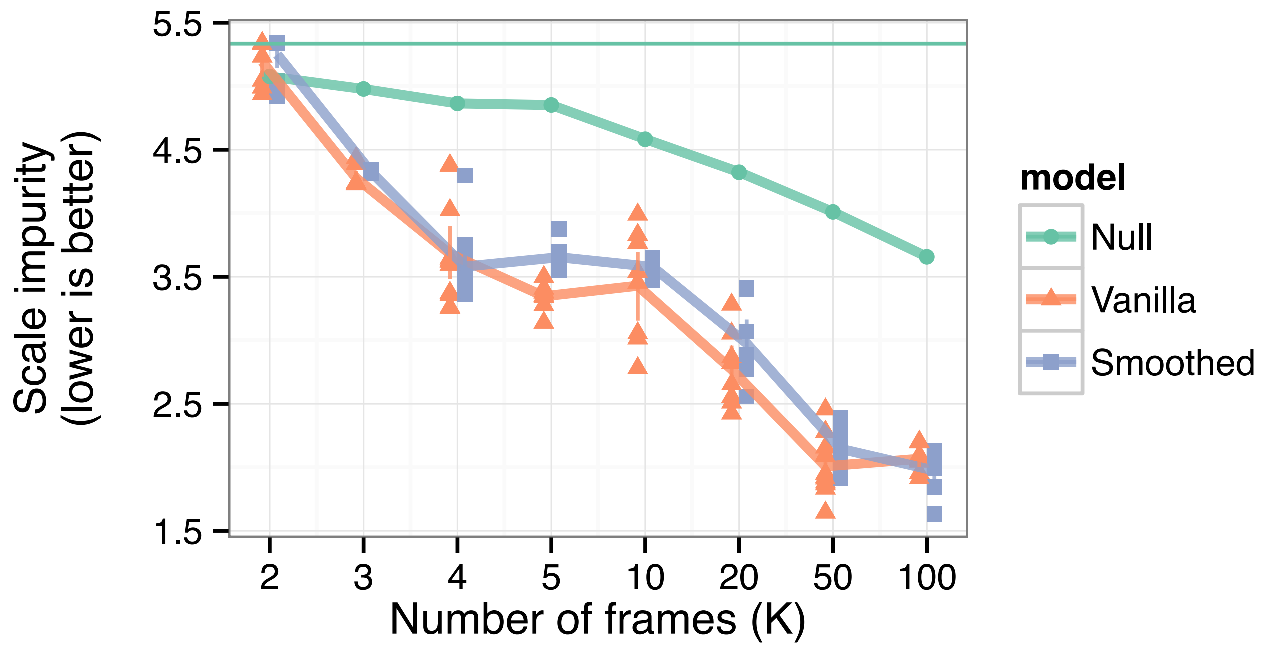
Case study

accuse, criticize, reject, tell, hand to, warn, ask, detain, release, order, deny, arrest, expel, convict, free, extradite to, allow, sign with, charge, urge

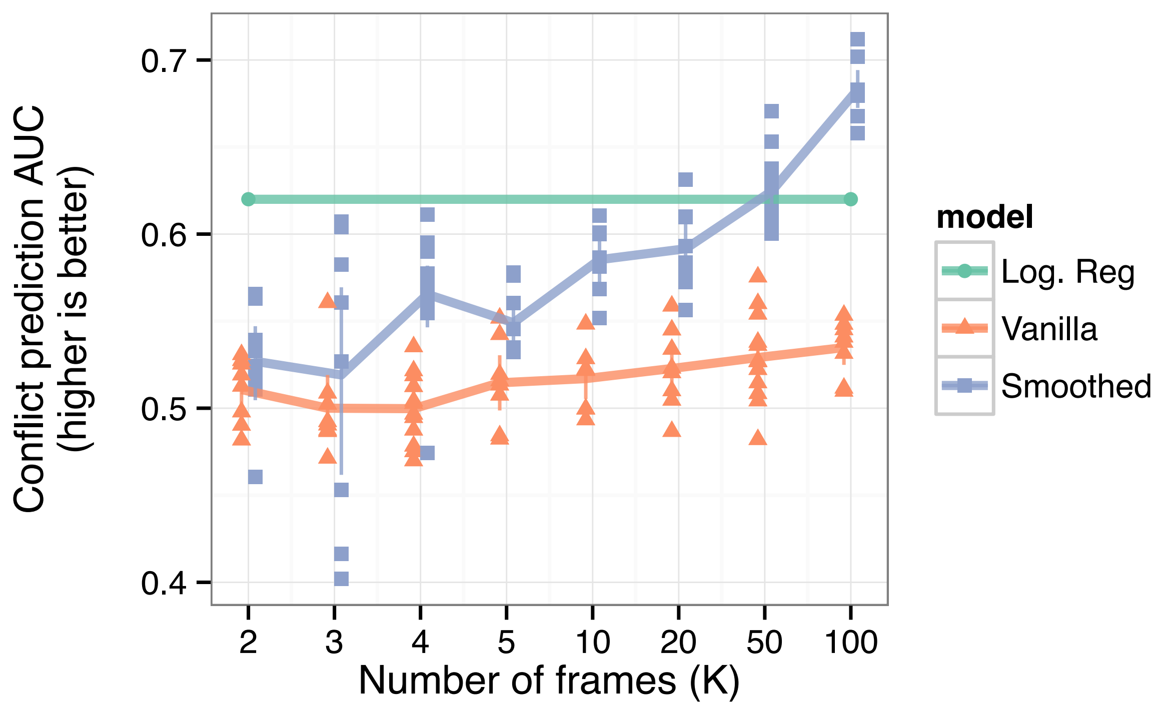


Quantitative Evaluation

Does the automatic ontology match one designed by experts? Compare verb clusters to manually defined ones in previous work (TABARI).



Does the model predict conflict? Use the model's inferred political dynamics to predict whether a conflict is happening between countries, as defined by the Militarized Interstate Dispute dataset.



International Relations Event Data

- Jointly learn
  - *linguistic event classes* (= verb distributions)
  - *political context* (= dyad's event class probs over time)
  - Examples seem consistent with the historical record
- Immediate ongoing work:
  need better semantic quality
  - Semi-supervision with lexicons
    - Extend huge amount of prior work
    - Identifiability helps analysis
  - Annotation evaluation (standard IE approach)

International Relations Event Data

- Goal: use the model to learn *new* facts about international politics
- Future work
  - More data; deeper historical analysis
  - Learning the entity database: domestic politics, other domains
  - Hierarchy and valences on the event types
  - Location and temporal properties of events
  - Network model
  - Temporal dynamics