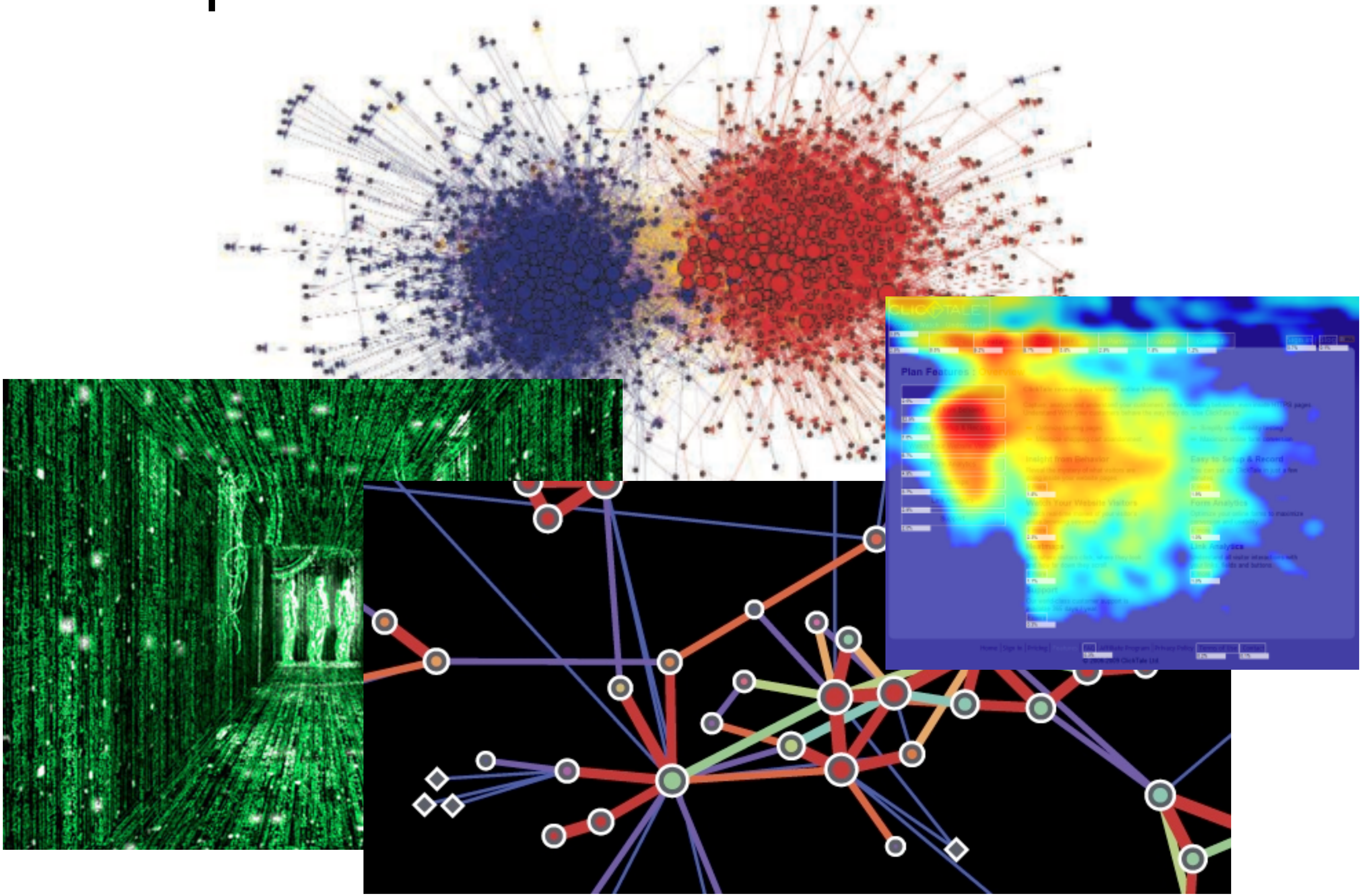
Statistical Text Analysis for Social Science: Learning to Extract International Relations from the News

Brendan O'Connor
Machine Learning Department
Carnegie Mellon University

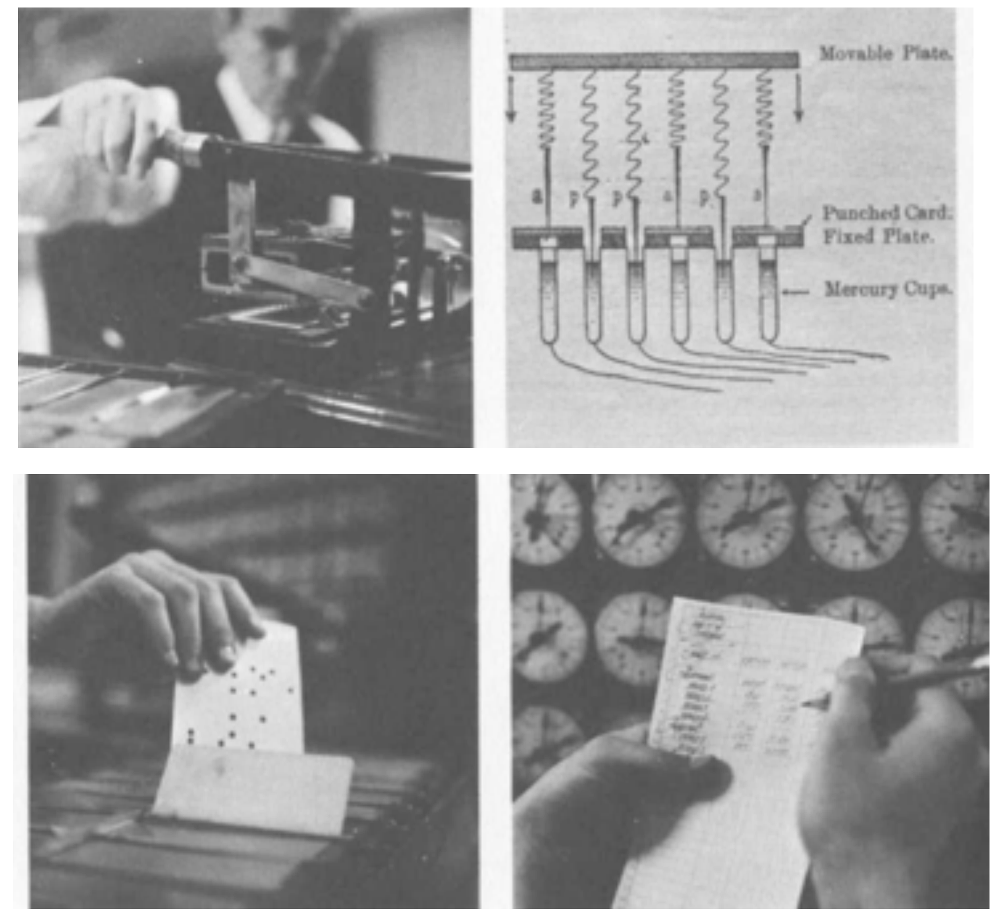
CLIP seminar, University of Maryland
Oct 9, 2013

Materials: http://brenocon.com
Joint work with Brandon Stewart (Harvard Government) and Noah Smith (CMU)

Computational Social Science



Computational Social Science



1890 Census tabulator - solved 1880's data deluge

Computation *as a tool* for social science applications

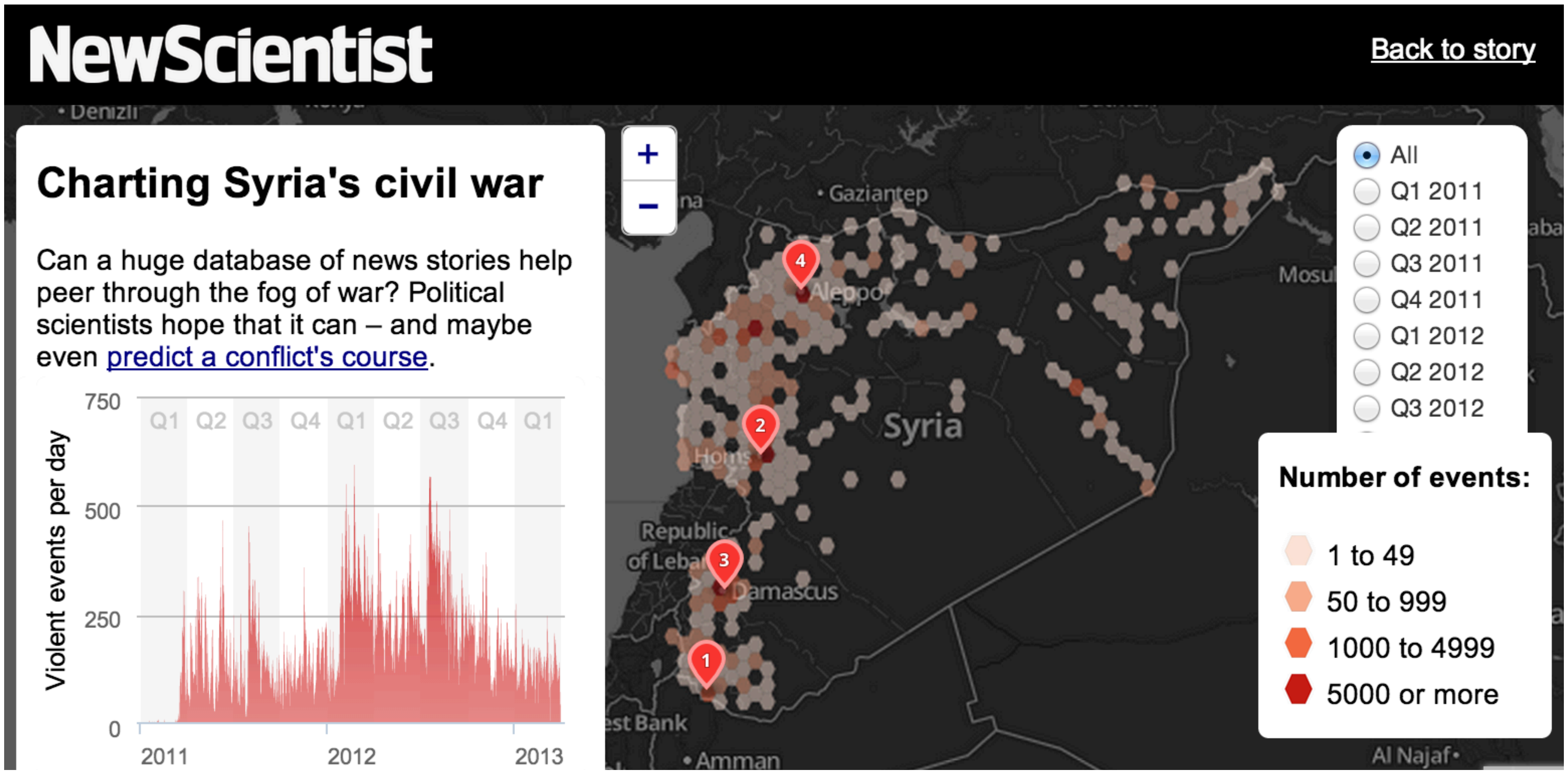
Automated Text Analysis

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

International Relations

- "Democratic peace" hypothesis:
  fewer wars between democracies?
- When do crises escalate or get resolved?
- When and where will future conflicts happen?

International Relations Event Data



GDELT project (Leetaru and Schrodt, 2013)

Extracted from news text

http://gdelt.utdallas.edu

International Relations Event Data

- Goal: Analyze time-series of country-country interactions: *who* did *what* to *whom*?
  - Create historical datasets of diplomatic and military actions between countries, derived from <u>news articles</u>
- 1960's: manual coding of news articles
- 1990's: automated coding (**information extraction**)
  - Rule-based verb pattern extractors
  - Used in dozens of political science studies

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)

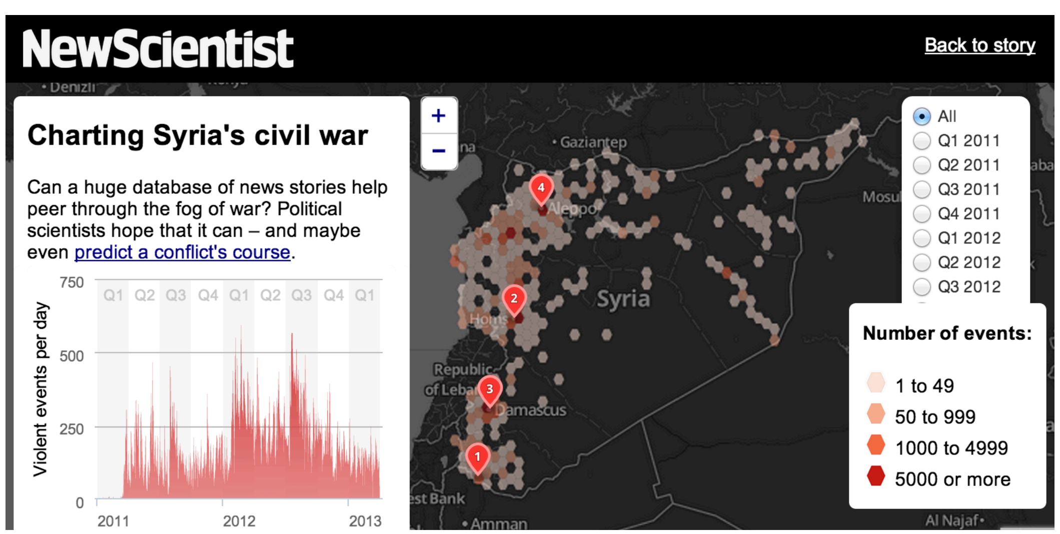
03 - EXPRESS INTENT TO COOPERATE
07 - PROVIDE AID
15 - EXHIBIT MILITARY POSTURE

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

<u>Issue:</u> Hard to maintain and adapt to new domains

Our approach

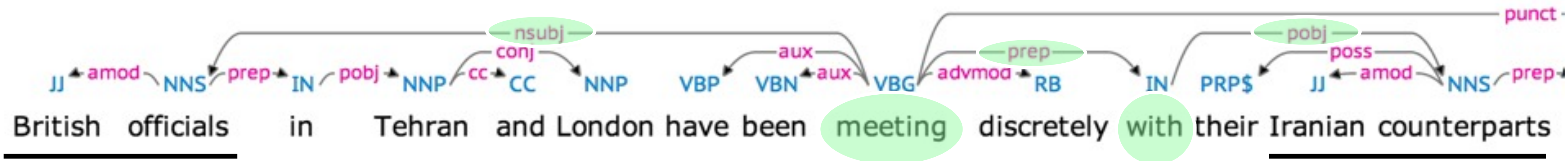
- Joint learning for high-level summary of event timelines
  - 1. Automatically learn the event types
  - 2. Extract events / political dynamics
- Probabilistic methods (Bayesian learning)
- Social context to drive unsupervised learning about language

[O'Connor, Stewart, Smith *ACL 2013*]

Newswire entity/predicate data

- Inputs
  1. 6.5 million news articles, 1987-2008
     - Gigaword corpus, including: AP, AFP, NYT, Xinhua
  2. Named entities: dictionary of country names
- Output: ~350k event tuples
  - Events between two actors
    (*SourceEntity*, *ReceiverEntity*, *Time (week)*, $w_{predpath}$)
- "Pakistan promptly accused India" [1/1/2000]
  => (PAK, IND, 268, *X -nsubj> accuse <dobj- Y*)

Event Extraction: Who did what to whom?



GBR

IRN

Country namelist matches

Verb-based dependency path

Source (*s*): GBR

Recipient (*r*): IRN

Predicate (*w*): *<--nsubj--* meet *--prep-->* with *--pobj-->*

"X meets with Y"

Proto-role terminology (Dowty 1991): Agent, Patient

Event Extraction

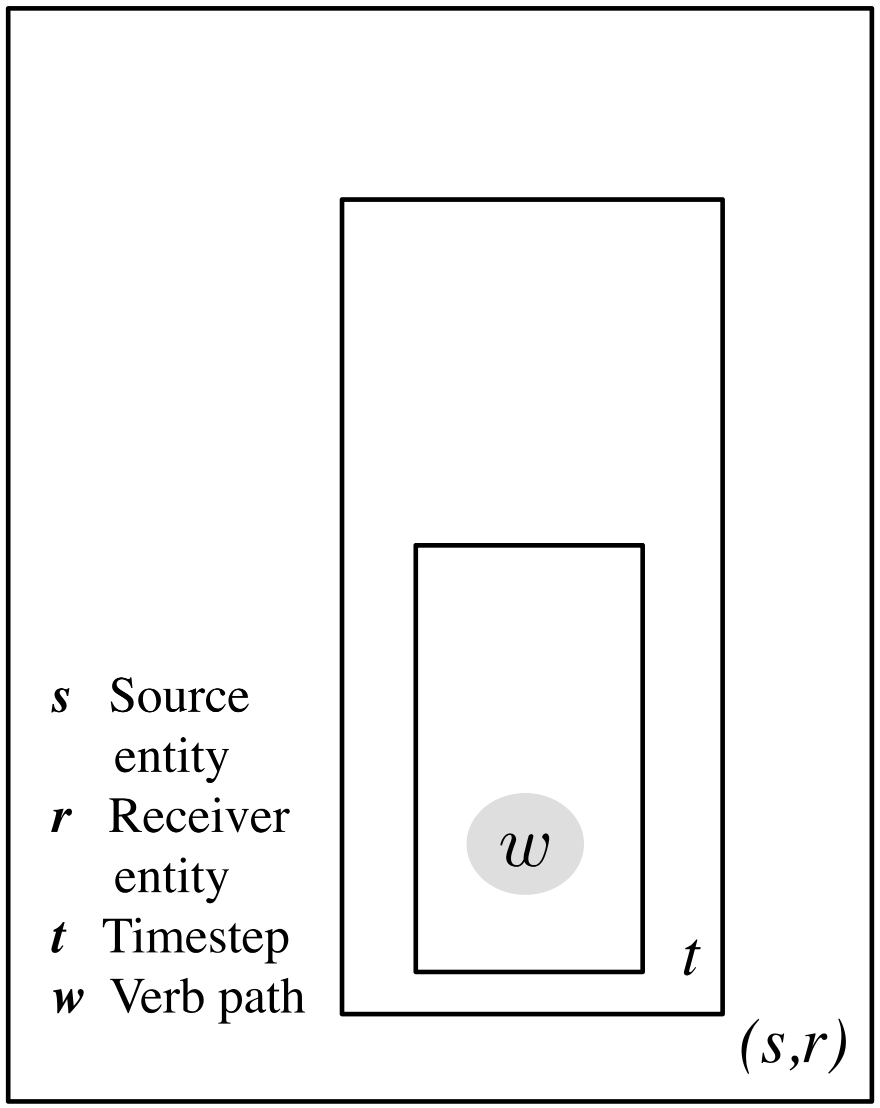
- Parsing and POS preprocessing: CoreNLP
- Fixed list of country names
- Predicates as verb-based dependency paths
- Filters for topics, factivity, verb-y paths, and parse quality

Predicate Paths

Most common	
16312	accuse
9762	visit
7533	arrive in
6206	meet with
6032	send to
6008	meet
5905	urge
5261	tell
5247	call on
4095	warn
3837	join
3823	say
3646	reject
3512	kill in
3402	hold with
2951	condemn

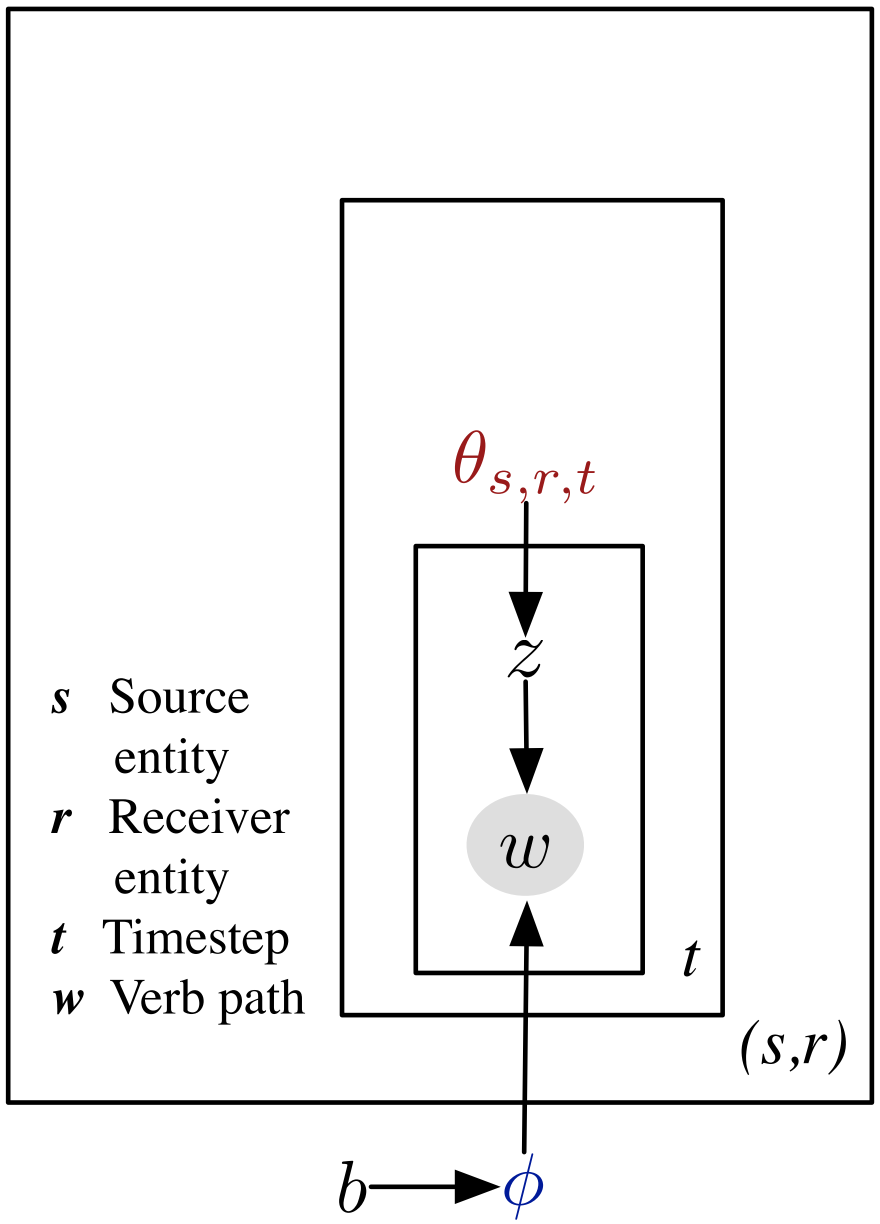
Sample	
21	say *ccomp*-> ask *nsubjpass*->
154	send in
401	put
1000	give to
1564	have troops in
293	gain from
279	launch into
83	arrive *partmod*-> start to
210	attend in
100	assail
13	deny *xcomp*-> support in
454	make in
384	serve in
176	have troops *partmod*-> station in
46	receive to
25	proceed to

Model



Data:
Each dyad has a sequence of timesteps
Each timestep has a number of events

Model



$$z \sim \text{Mult}(\theta_{s,r,t})$$

$$w \sim \text{Mult}(\phi_z)$$

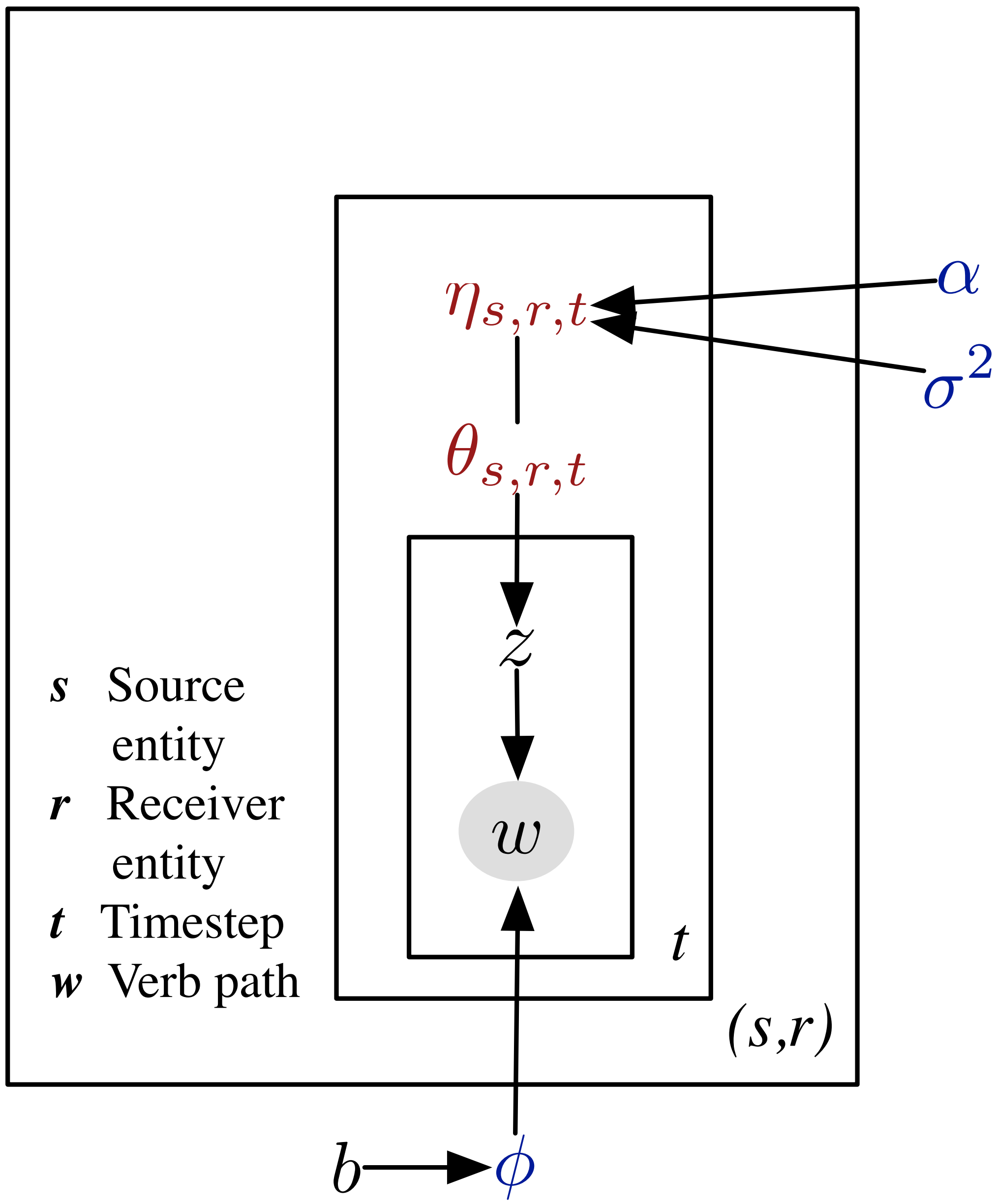
K event types: verb distributions

$$\phi_k \sim Dir(b) \qquad \in \text{simplex}(V)$$

Model

Key assumption: **dyadic** and **temporal** coherence

Model 1: independent contexts



$$\eta_{s,r,t} \sim N(\alpha + \beta_{s,r,t},\ \Sigma)$$

$$(\theta_{s,r,t})_k \propto \exp(\eta_{s,r,t,k})$$

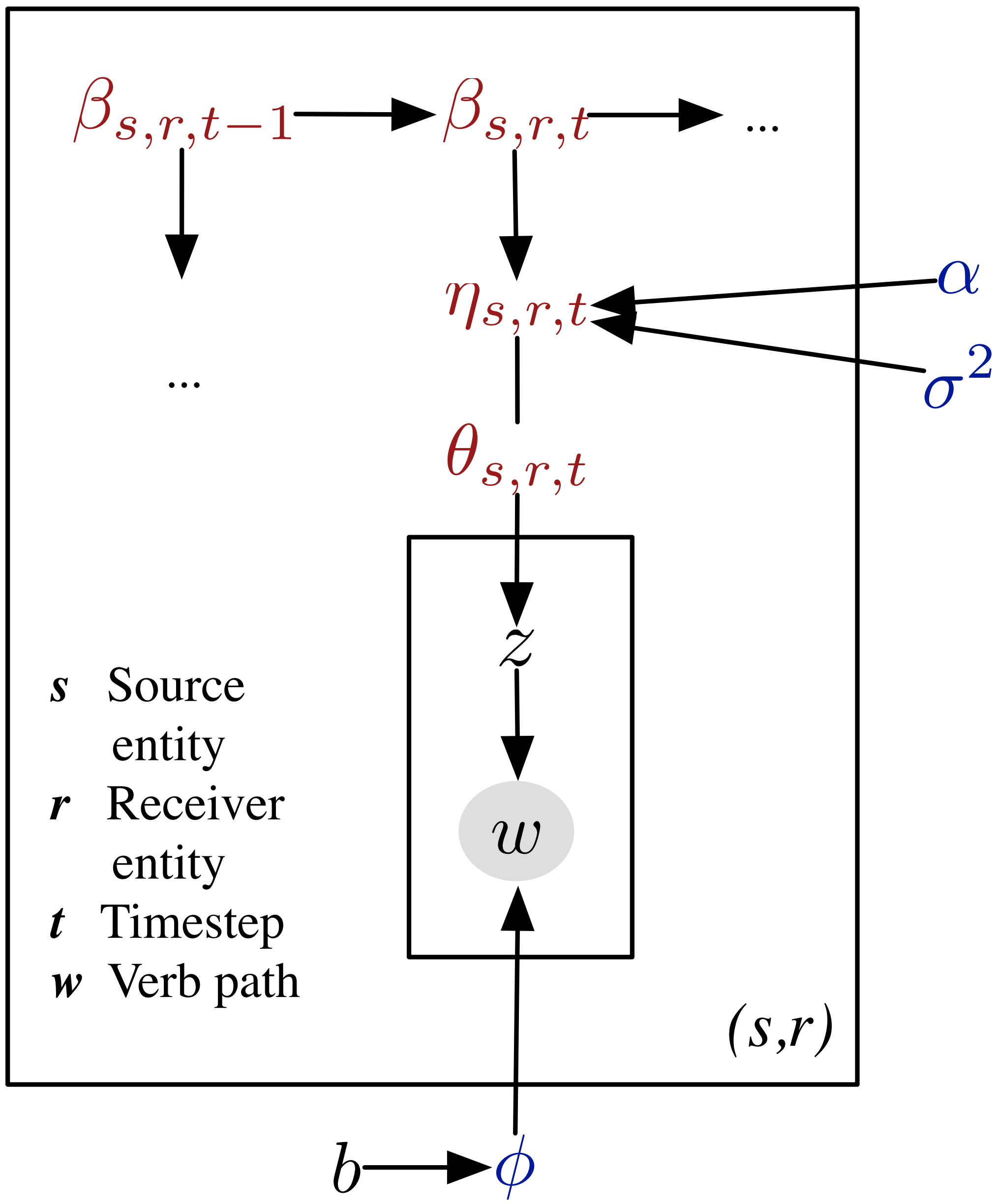
$$z \sim \mathrm{Mult}(\theta_{s,r,t})$$

$$w \sim \mathrm{Mult}(\phi_z)$$

K event types: verb distributions

$$\phi_k \sim Dir(b) \qquad \in \mathrm{simplex}(V)$$

Model



Key assumption: **dyadic** and **temporal** coherence

Model 1: independent contexts
Model 2: temporal smoothing

$$\beta_{s,r,t} \sim N(\beta_{s,r,t-1},\ \tau^2 \mathbb{I})$$

$$\eta_{s,r,t} \sim N(\alpha + \beta_{s,r,t},\ \Sigma)$$

$$(\theta_{s,r,t})_k \propto \exp(\eta_{s,r,t,k})$$

$$z \sim \mathrm{Mult}(\theta_{s,r,t})$$

$$w \sim \mathrm{Mult}(\phi_z)$$

K event types: verb distributions

$$\phi_k \sim Dir(b) \qquad \in \mathrm{simplex}(V)$$

Inference: blocked Gibbs sampling

(beta | eta, alpha, tau)
Linear dynamical system
(Gaussian HMM):
Forward filter backward sampler

(eta,theta | beta,alpha,z)
Logistic normal:
Metropolis-within-Gibbs,
Laplace approximation proposal

Conjugate normal

$$\beta_{s,r,t} \sim N(\beta_{s,r,t-1},\ \tau^2 \mathbb{I})$$

$$\eta_{s,r,t} \sim N(\alpha + \beta_{s,r,t},\ \Sigma)$$

$$(\theta_{s,r,t})_k \propto \exp(\eta_{s,r,t,k})$$

(z | w,theta,b)
Dirichlet-multinomial:
Collapsed sampling

$$z \sim \mathrm{Mult}(\theta_{s,r,t})$$

$$w \sim \mathrm{Mult}(\phi_z)$$

$$\phi_k \sim Dir(b) \qquad \in \mathrm{simplex}(V)$$

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

$$\phi_k \sim Dir(b) \qquad \in \text{simplex}(V)$$

Evaluation

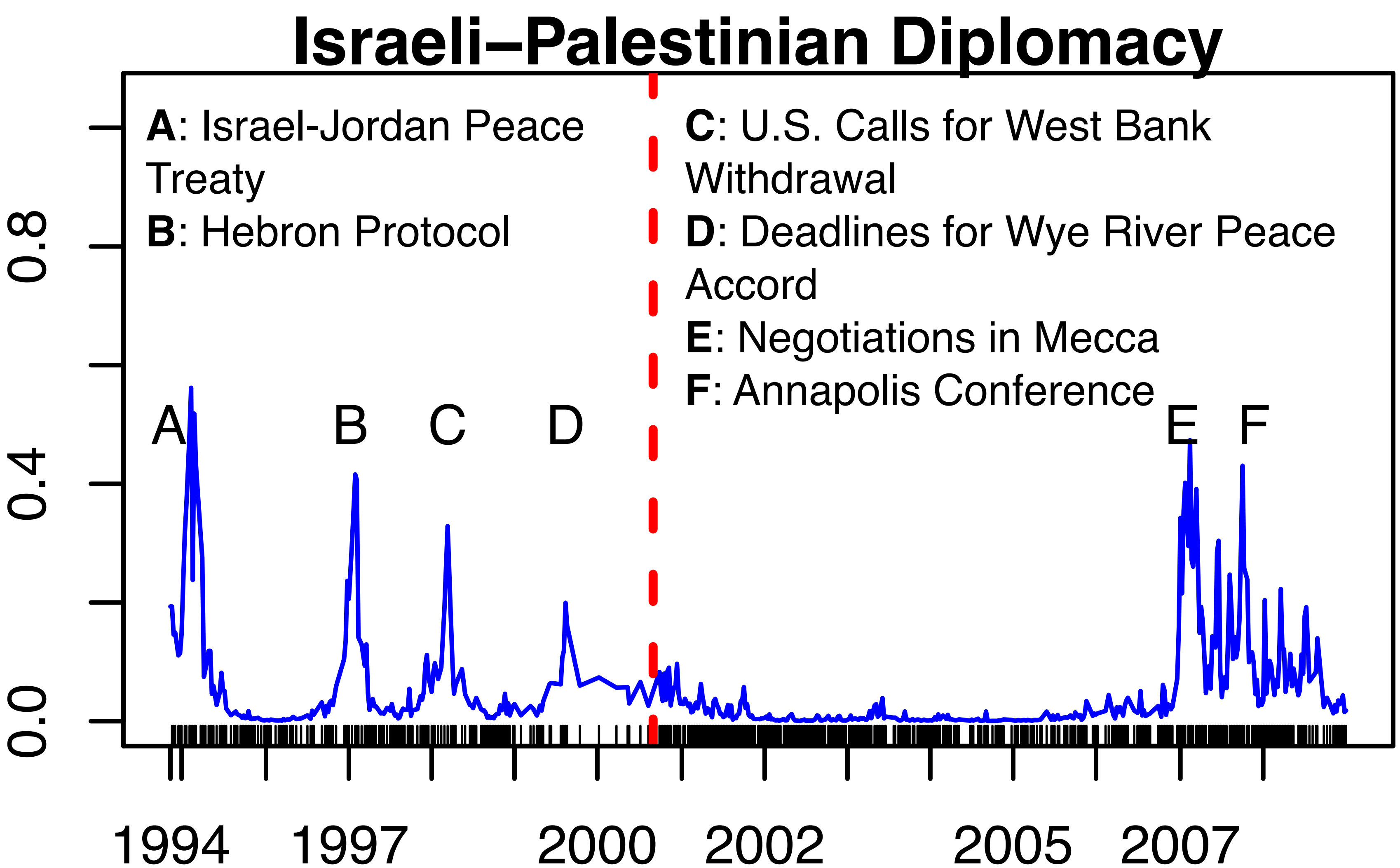
- Unsupervised model evaluation: need multiple checks of reasonableness
- Qualitative case study (face validity)
- Quantitative
  - Recovering a pre-existing ontology
  - Conflict prediction
- [Future work: do actual political science]

Case study

- Israeli-Palestinian conflict, 1994-2008
    - ISR-PSE is most frequent dyad
    - Militarized Interstate Dispute database has *no* data
    - Can our system give a useful analysis?

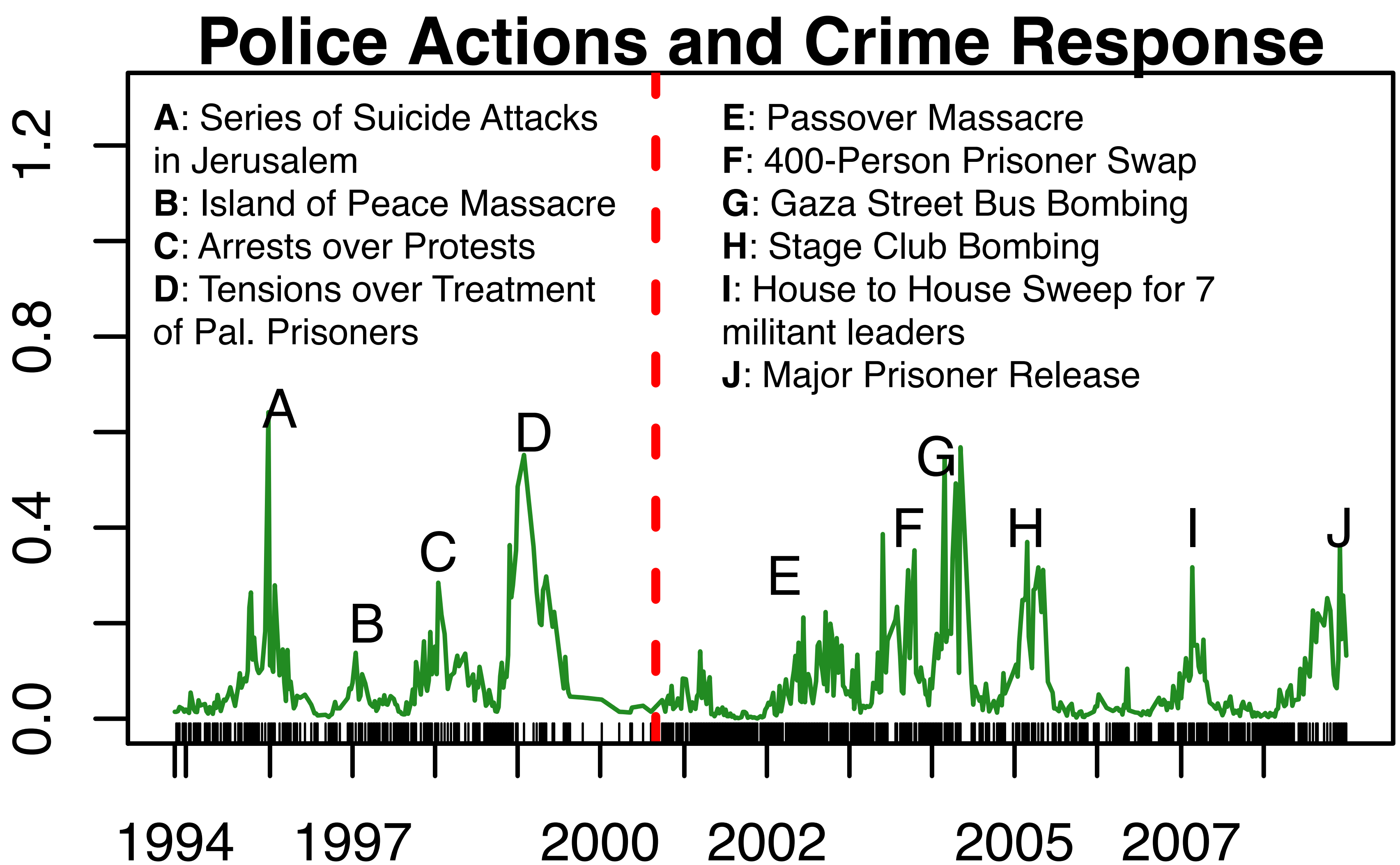
Case study

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

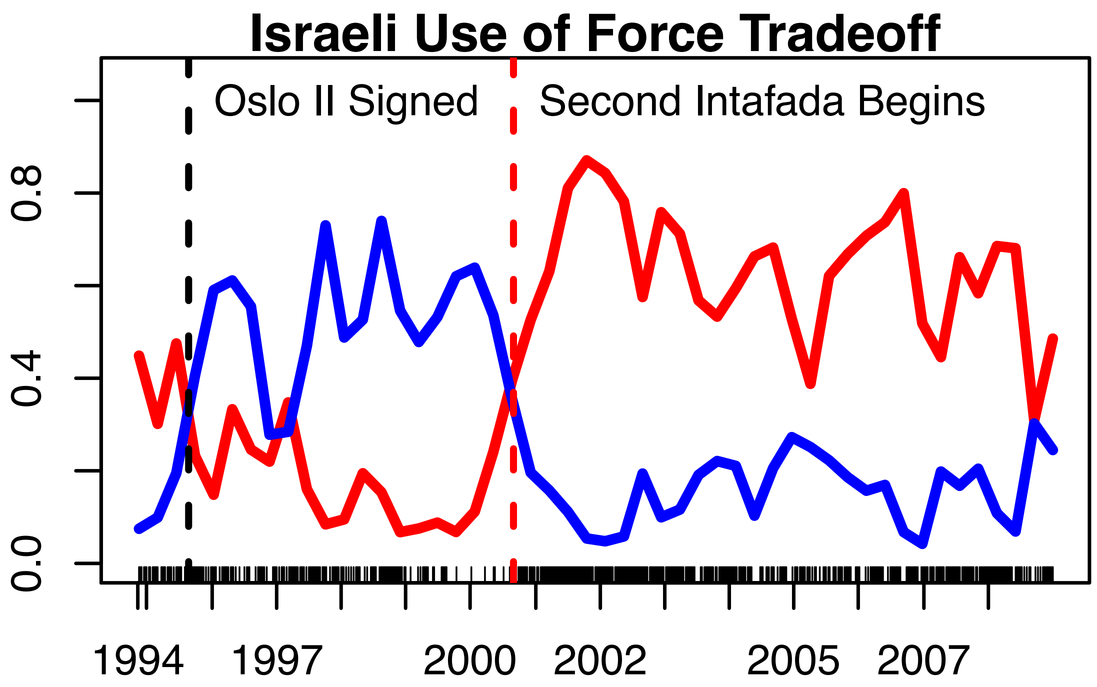


Case study

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



Case study



impose on, seal, capture from, seize from, arrest, ease closure of, close, deport, close with, release

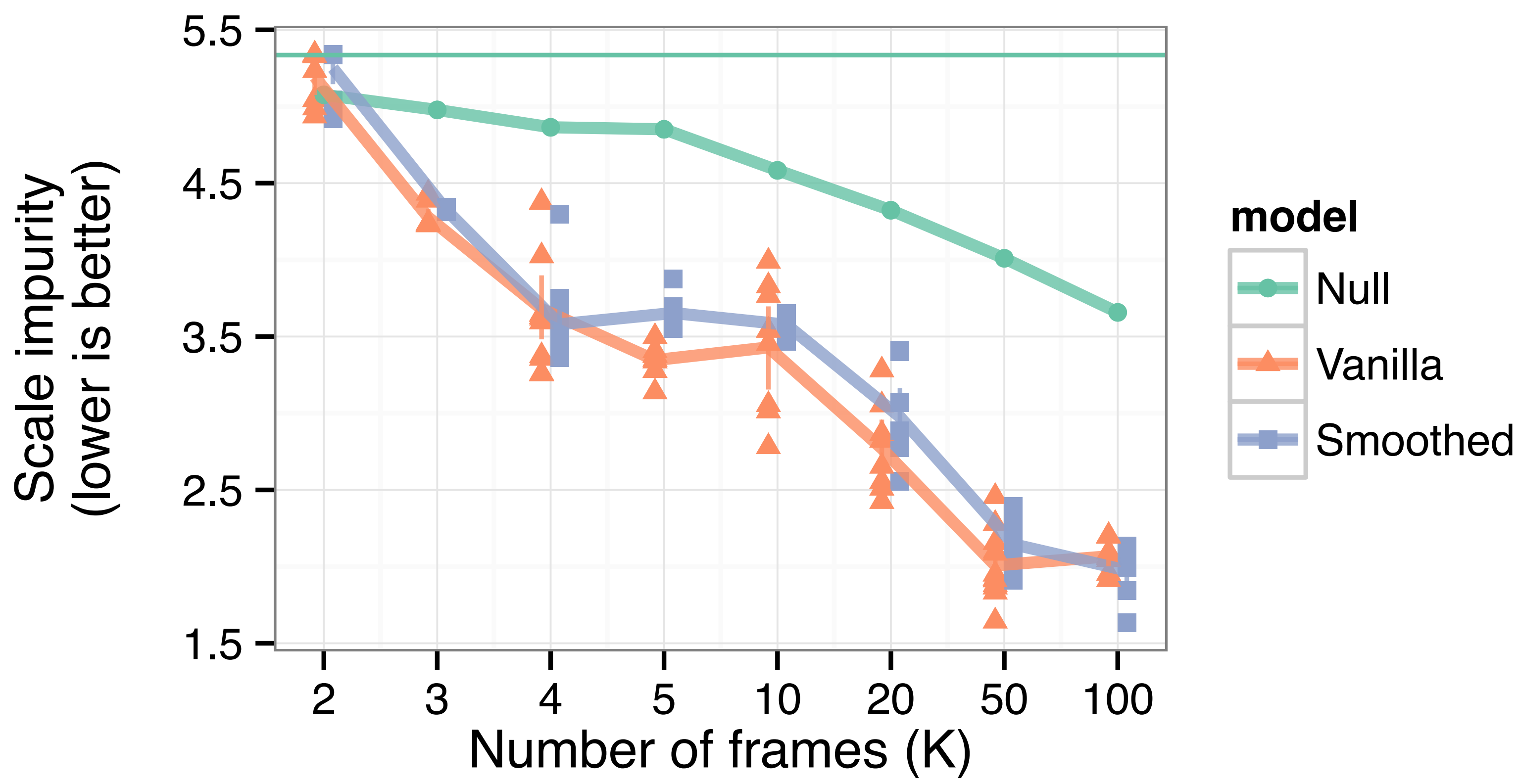
kill, fire at, enter, kill in, attack, raid, strike in, move into, pound, bomb

Correlates to conflict?

Semantic coherence?

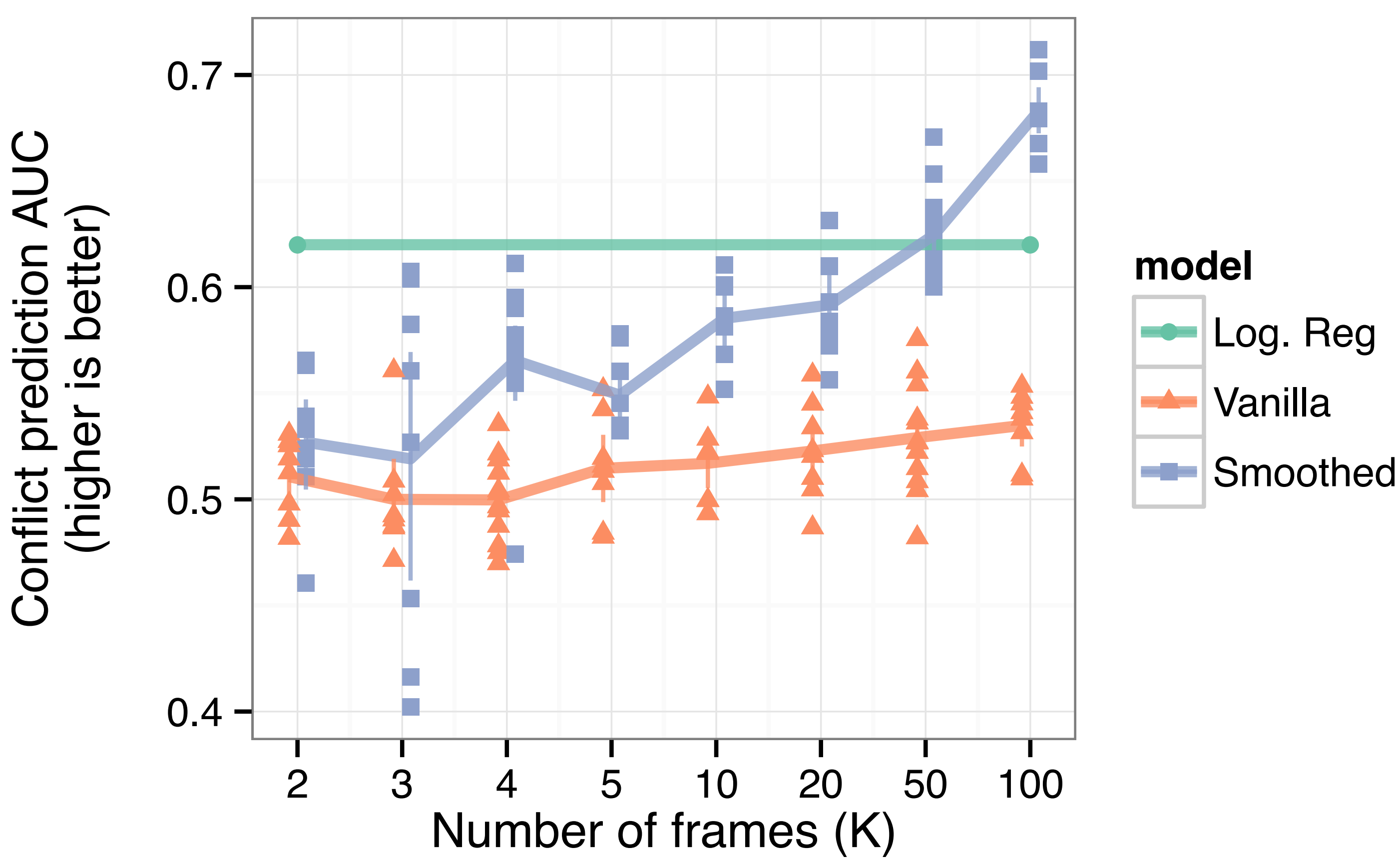
Lexical scale evaluation

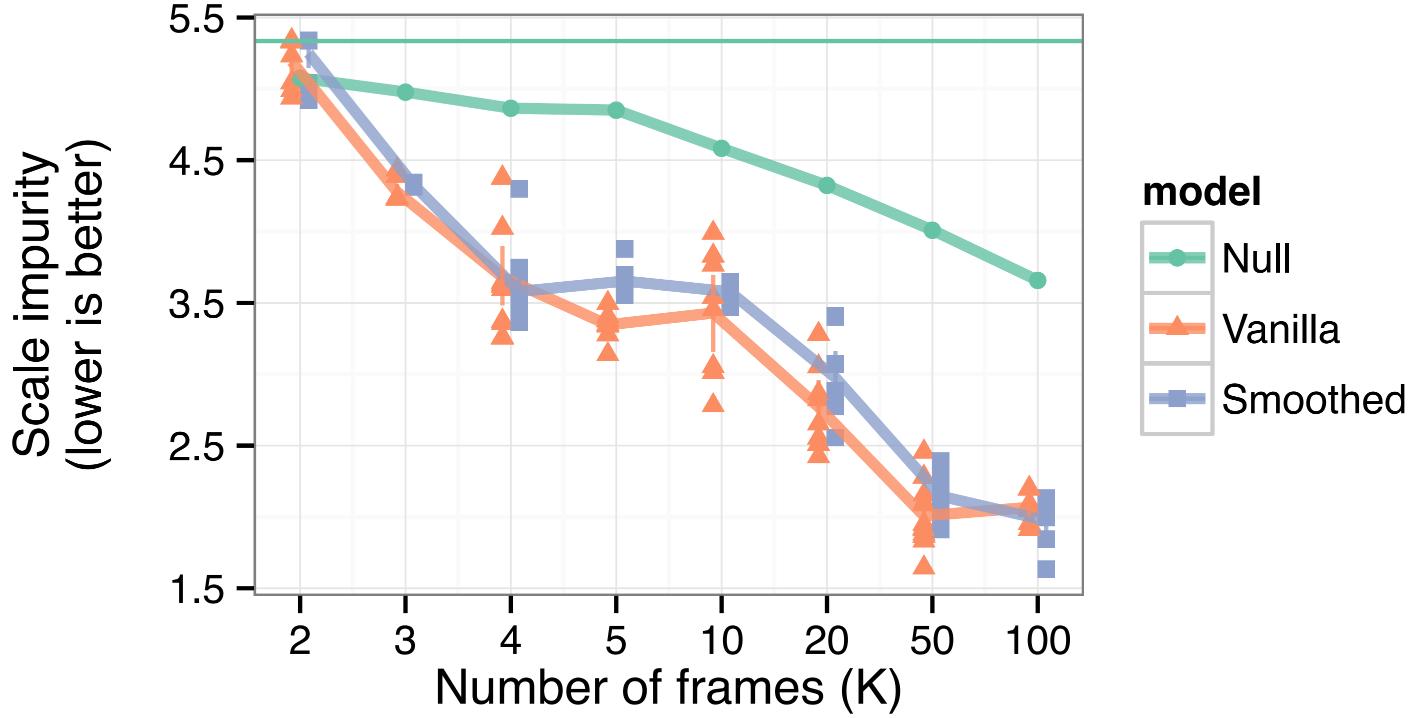
- Do our event types (verb clusters) match the manually defined ontology?
- Match dependency paths against TABARI patterns (536 / 10k)
- Granularity invariance: use expert-assigned scale score (-10 to 10) [controversial?]
- Lexical scale impurity: average difference between randomly chosen words
- Random clusters baseline

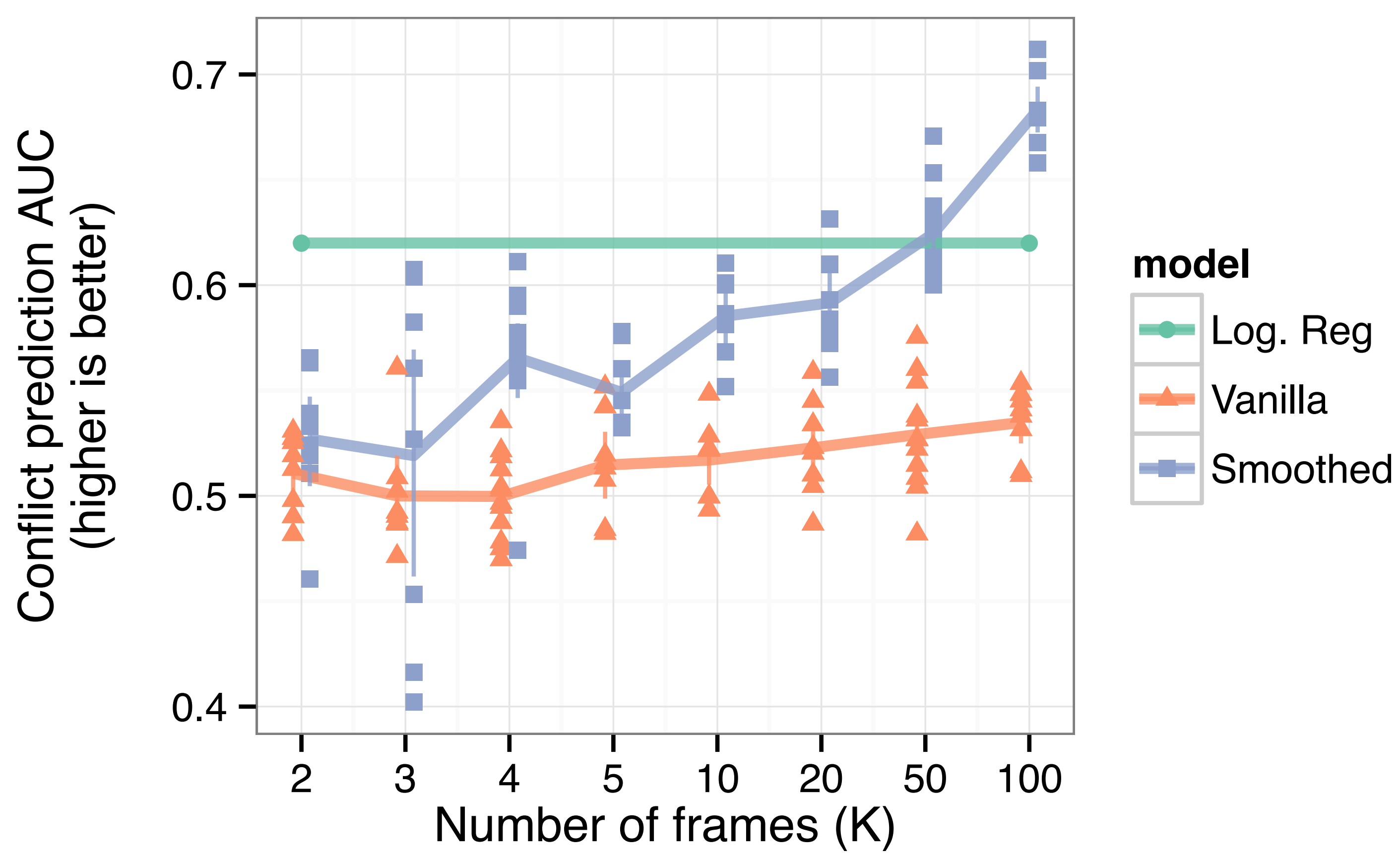


Conflict prediction/correlation

- Do our event types correspond to real-world conflict?
- “Gold” standard: Militarized Interstate Dispute dataset (from Correlates of War project)
- Regularized logistic regression from *theta* (event probs per dyad-time slice)
- Baseline: regularized logistic regression from path counts




Scale impurity
(lower is better)
5.5
4.5
3.5
2.5
1.5
2
3
4
5
10
20
50
100
Number of frames (K)
model
Null
Vanilla
Smoothed


Conflict prediction AUC
(higher is better)
0.7
0.6
0.5
0.4
2
3
4
5
10
20
50
100
Number of frames (K)
model
Log. Reg
Vanilla
Smoothed

International Relations Event Data

- Jointly learn
  - *linguistic event types* (= verb clusters)
  - *political context* (= dyad's eventtype 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
    - Related: seed words in topic models
  - Annotation evaluation? (standard IE approach)

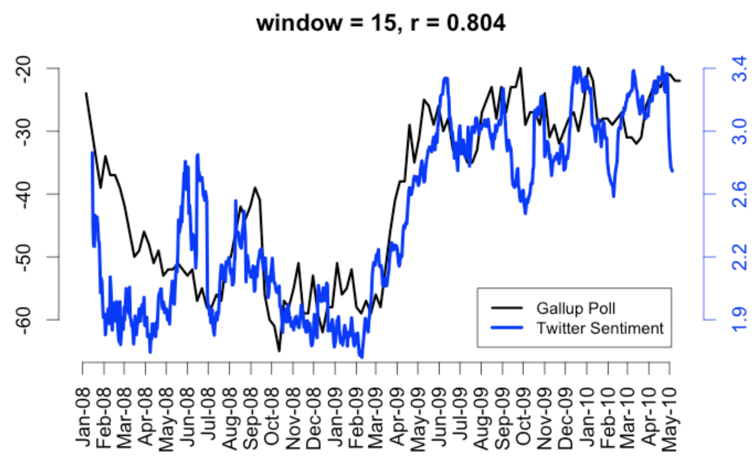
International Relations Event Data

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

Text Analysis for Social Science

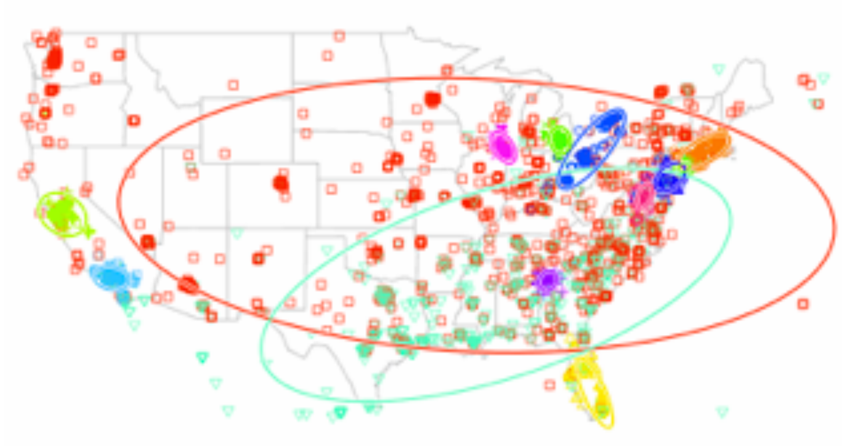
- Automated content analysis:
  tools for *discovery* and *measurement* of concepts, attitudes, events
- Applications to social science areas:
  how to use previous work?
  Interdisciplinary collaboration
- Social contextual factors -- e.g. *who* and *when* -- can drive linguistic learning
- Expert ontologies give evaluations, or hypotheses to test and/or expand

Discovery and measurement in social media text



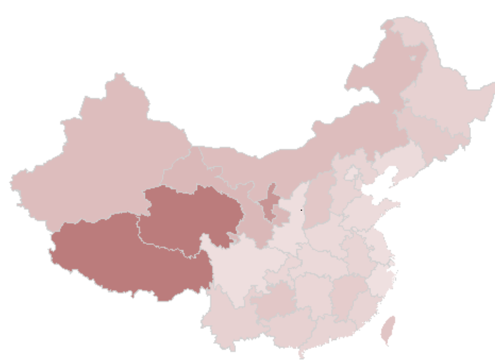
Opinion polls and sentiment analysis
[ICWSM 2010]

Geographic and demographic factors in slang and language change
[EMNLP 2010, work-in-submission]



ikr	smh	he	asked	fir	yo	last	name
!	G	O	V	P	D	A	N

Linguistic analysis tools
[ACL 2011, NAACL 2013]



Censorship in Chinese social media [FM 2011]

Discovery in fictional narratives

[Bamman, O'Connor, Smith ACL 2013]

- From movie plot summaries:
  model of characters' attributes and actions

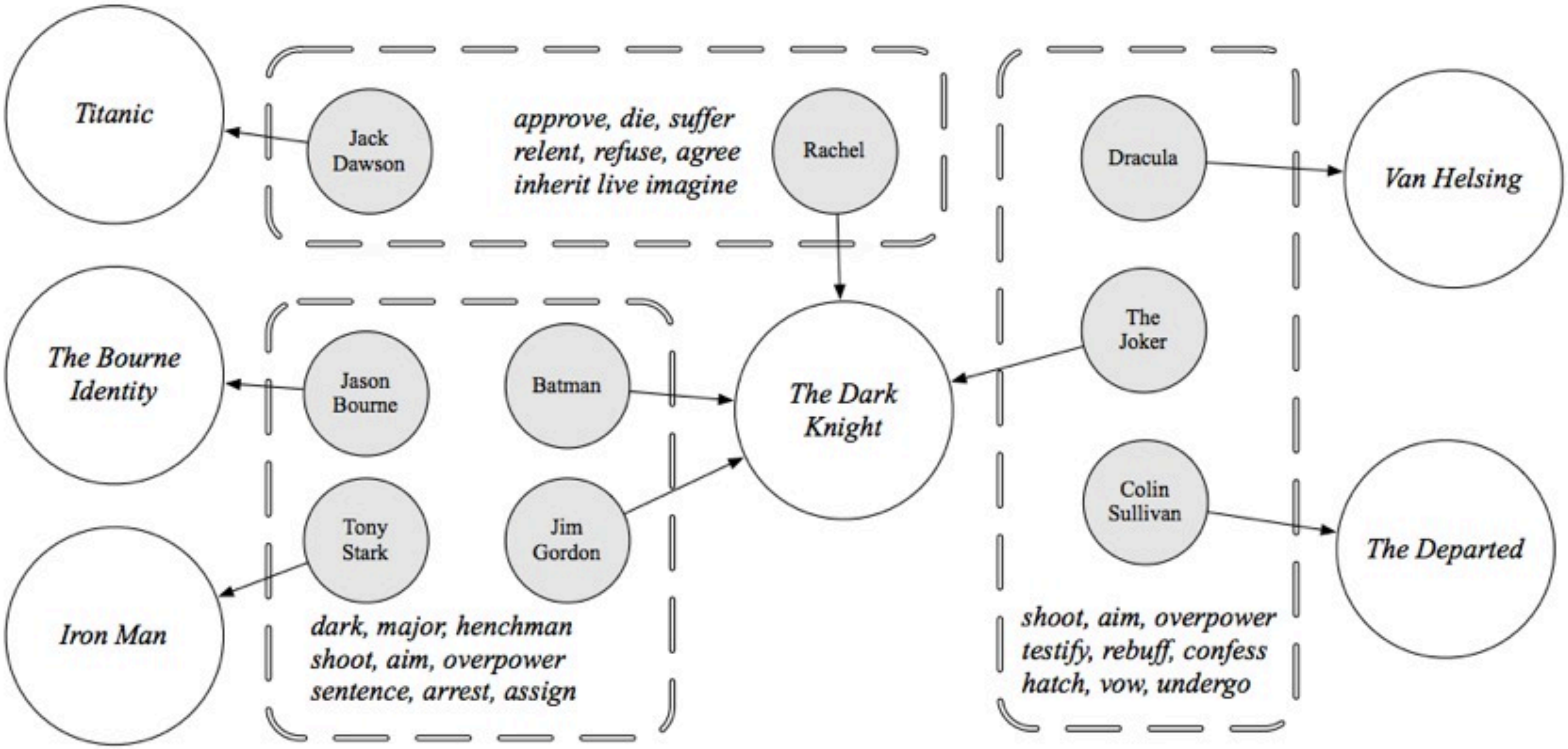


Figure 3: Dramatis personae of *The Dark Knight* (2008), illustrating 3 of the 100 character types learned by the persona regression model, along with links from other characters in those latent classes to other movies. Each character type is listed with the top three latent topics with which it is associated.

Thanks

- Materials, etc: http://brenocon.com
  Feedback welcome!