Social Event Extraction: Inferring International Relations and Police Killings from the News

Brendan O'Connor
College of Information and Computer Sciences
University of Massachusetts Amherst
http://brenocon.com

Data Science Workshop on Computational Social Science
Yale University, October 20, 2017

Joint work with: Katherine Keith, Abram Handler, Michael Pinkham, Cara Magliozzi, Joshua McDuffie, Saul Shanabrook, Brandon Stewart, Noah Smith

Computational Social Science

1. Computationally mediated human social behavior: e.g. crowdsourcing, online auctions
2. Computationally oriented analysis methods: e.g. agent-based simulations
3. Artificial intelligence (ML/Vision/NLP) as a social scientific, data analysis method

Computational Social Science

Official social data

Data collection

100 BCE

Data analysis

1829

1900 2000

Computational Social Science

Official social data

Data collection

100 BCE

Data analysis

1829

Newly available social data

Digitized behavior
Billions of users
Billions of messages/day

Digitized news
Thousands of articles/day

Digitized archives
Millions of books/century

HATHI TRUST Digital Library

Google books

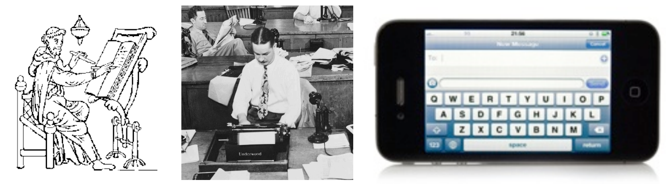
1900 2000

TextGenerator(SocialAttributes) → *Text*

Data generation process

Society *(SocialAttributes)*

Writing *(TextGenerator)*



Text Data *(Text)*

HATHI TRUST Digital Library

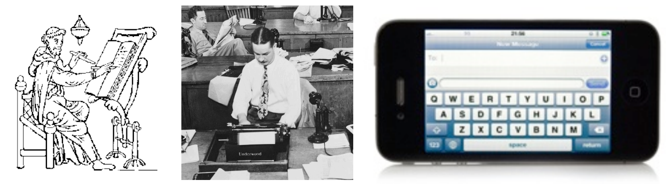
Language generation as social process
P(TextGen | Text, SocAttr)

Language for social measurement
P(SocAttr | Text, TextGen)

TextGenerator(SocialAttributes) → Text

Society *(SocialAttributes)* | Writing *(TextGenerator)* | Text Data *(Text)*

Data generation process



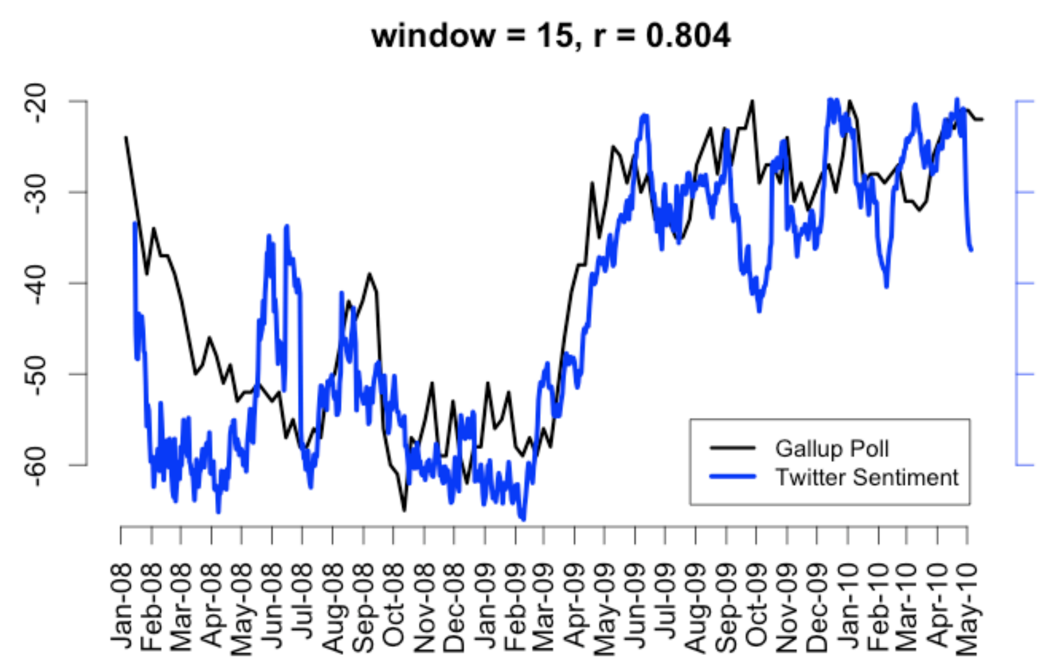
Public opinion

Social media usage

Model assumptions

Language for social measurement

P(SocAttr | Text, TextGen)



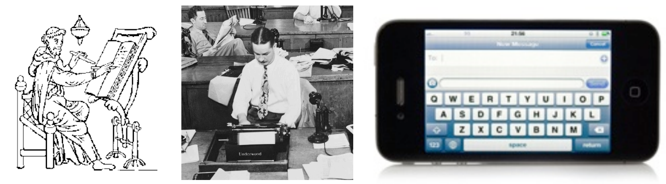
[O'Connor et al., 2010]

TextGenerator(SocialAttributes) → *Text*

Society *(SocialAttributes)*

Data generation process

Writing *(TextGenerator)*



Text Data *(Text)*

HATHI TRUST Digital Library

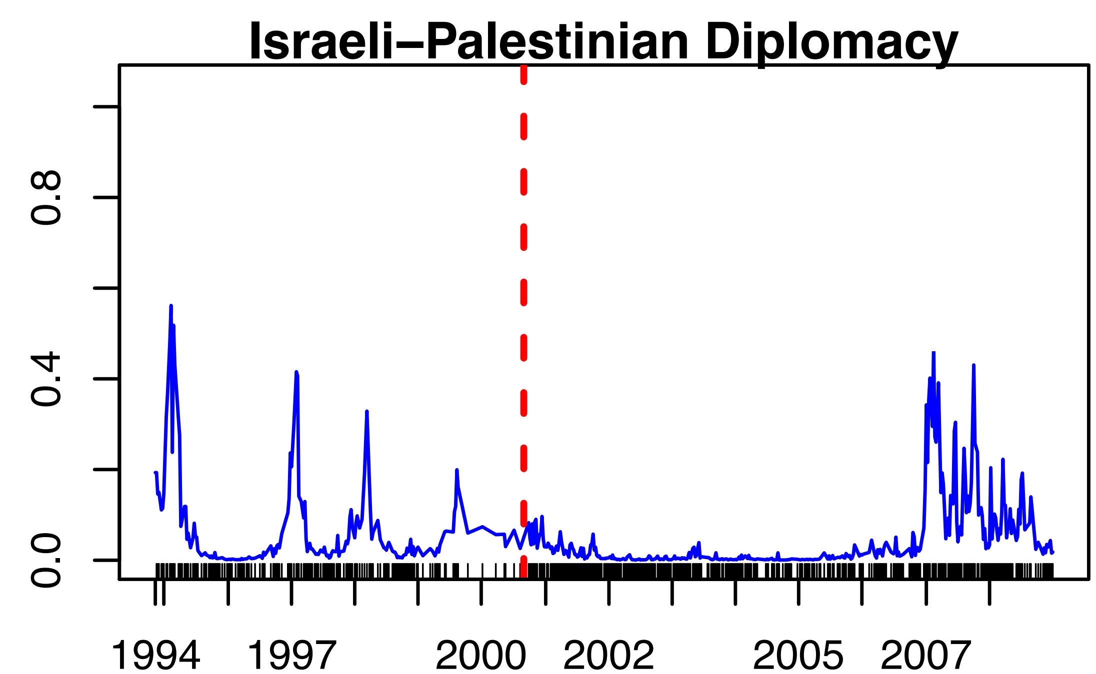
Real-world political events

News media process

Model assumptions

Language for social measurement

P(SocAttr | Text, TextGen)



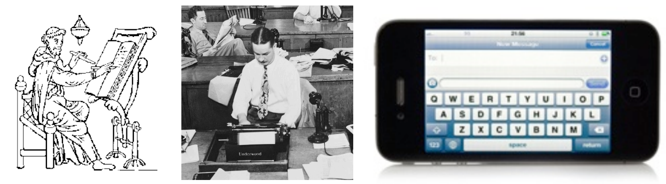
[O'Connor, Stewart, Smith 2013]

TextGenerator(SocialAttributes) → *Text*

Data generation process

Society *(SocialAttributes)*

Writing *(TextGenerator)*



Text Data *(Text)*

Real-world political events

News media process

Language generation as social process

P(TextGen | Text, SocAttr)

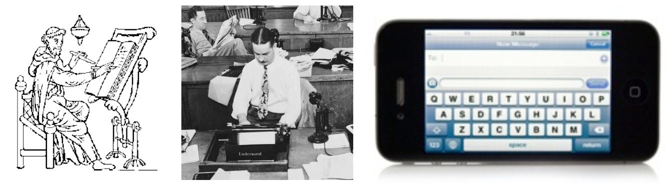
TextGenerator(SocialAttributes) → Text

Data generation process

Society *(SocialAttributes)*

→

Writing *(TextGenerator)*



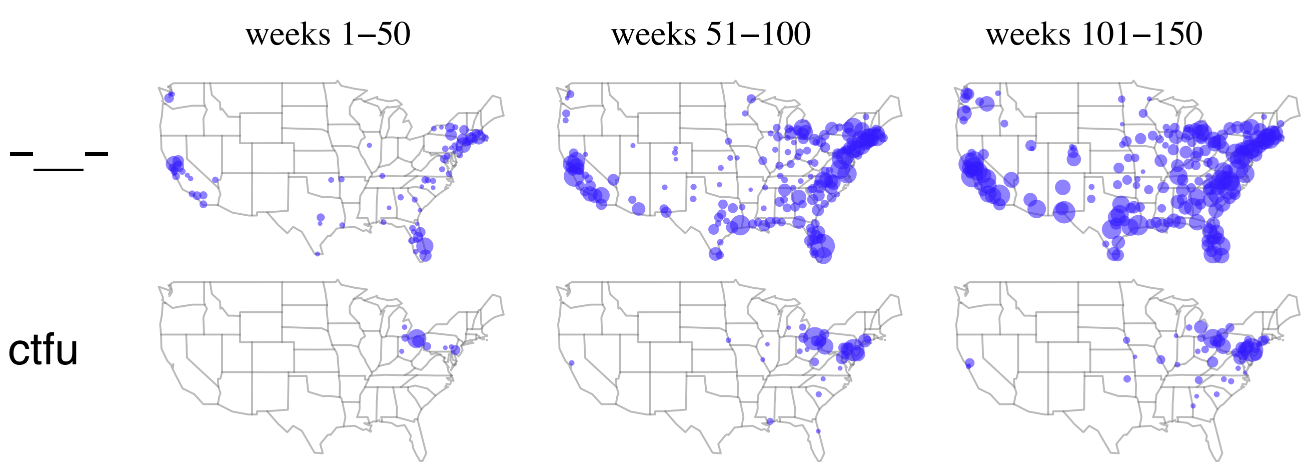
→

Text Data *(Text)*

HATHI TRUST Digital Library

Geography of authors

Social media usage →



Language generation as social process

P(TextGen | Text, SocAttr)

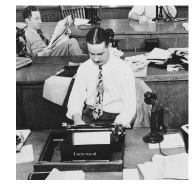
[Eisenstein et al. 2010, O'Connor et al. 2010, Eisenstein et al. 2012]

TextGenerator(SocialAttributes) → Text

Society *(SocialAttributes)*

Data generation process

Writing *(TextGenerator)*



Text Data *(Text)*

HATHI TRUST Digital Library

Racial demographics of authors → Social media usage →

AAE	Ratio	SAE
sholl	1802.49	sure
iont	930.98	I don't
wea	870.45	where
talmbout	809.79	talking about
sumn	520.96	something

Language generation as social process
P(TextGen | Text, SocAttr)

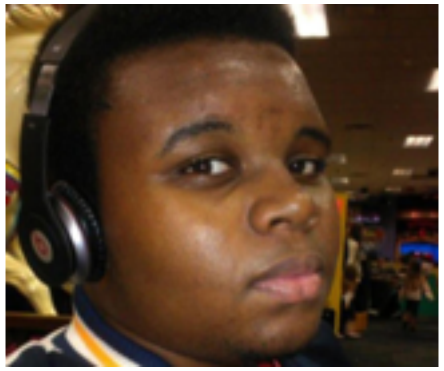
African-American English on Twitter
Dialects and social media NLP

[Blodgett et al. 2016, Blodgett and O'Connor 2017]

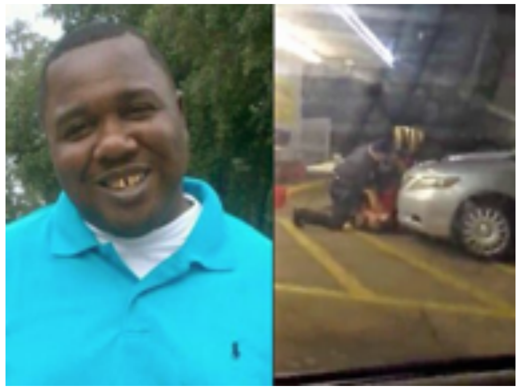
Police killings

July 17, 2014

Aug 9, 2014



July 5, 2016

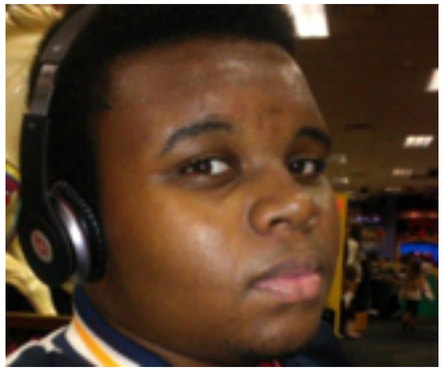


July 6, 2016

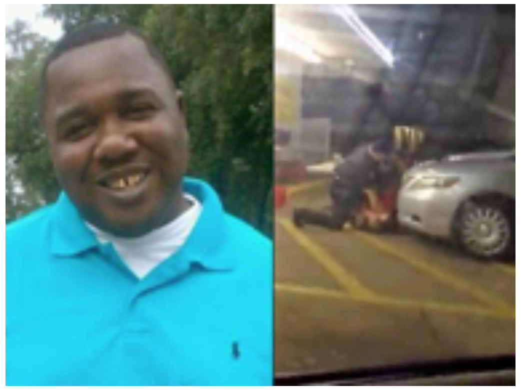
Police killings

July 17, 2014

Aug 9, 2014



July 5, 2016



July 6, 2016

Data?

Eric Garner	New York, NY
Michael Brown	Ferguson, MO
Alton Sterling	Baton Rouge, LA
Philando Castile	Falcon Heights, MN

Police killings

Data?

Eric Garner	New York, NY
Michael Brown	Ferguson, MO
Alton Sterling	Baton Rouge, LA
Philando Castile	Falcon Heights, MN

Police killings

- Are there more or fewer fatalities than last year?

Data?

Eric Garner	New York, NY
Michael Brown	Ferguson, MO
Alton Sterling	Baton Rouge, LA
Philando Castile	Falcon Heights, MN

Police killings

- Are there more or fewer fatalities than last year?
- Is there racial disparity/discrimination?

Data?

Eric Garner	New York, NY
Michael Brown	Ferguson, MO
Alton Sterling	Baton Rouge, LA
Philando Castile	Falcon Heights, MN

Police killings

- Are there more or fewer fatalities than last year?
- Is there racial disparity/ discrimination?
- Which police departments are better or worse? What policing strategies are most effective or safe?

Data?

Eric Garner	New York, NY
Michael Brown	Ferguson, MO
Alton Sterling	Baton Rouge, LA
Philando Castile	Falcon Heights, MN

Police killings

- Are there more or fewer fatalities than last year?
- Is there racial disparity/ discrimination?
- Which police departments are better or worse? What policing strategies are most effective or safe?

- **Need good data for the public interest and social science / policy making**

Data?

Eric Garner	New York, NY
Michael Brown	Ferguson, MO
Alton Sterling	Baton Rouge, LA
Philando Castile	Falcon Heights, MN

Issues in government data

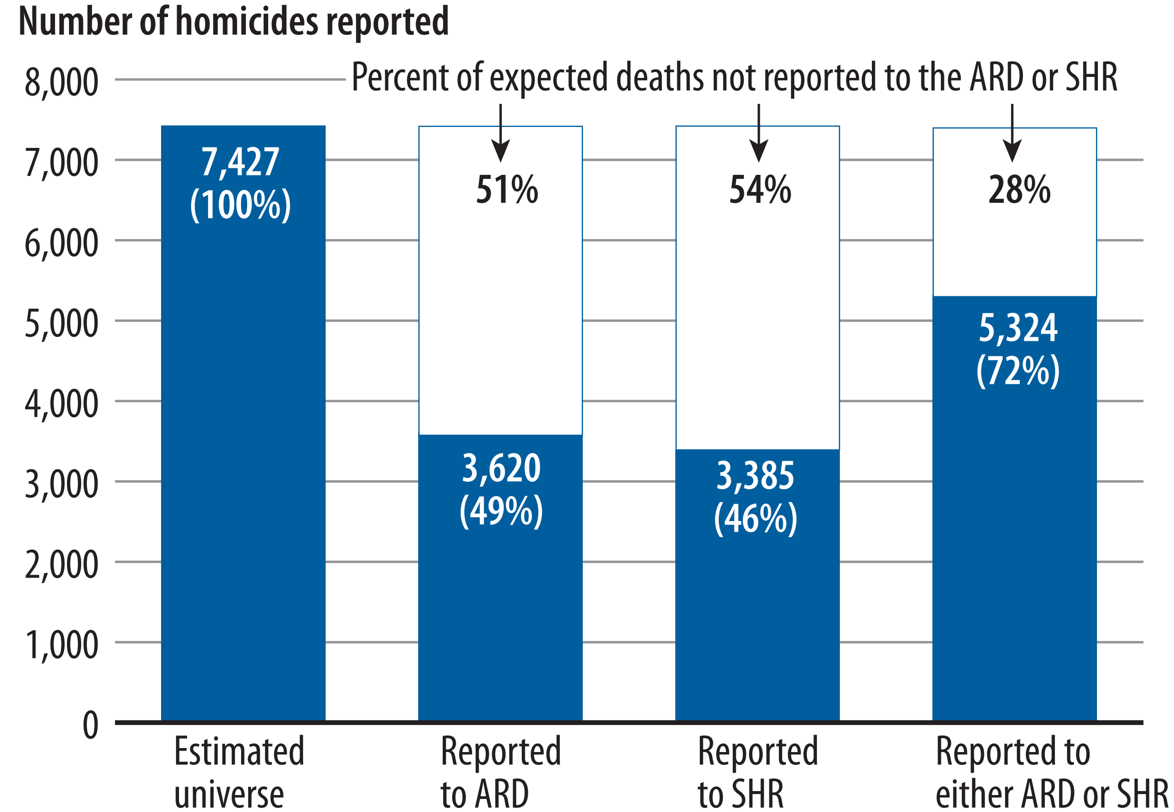
- Washington Post, Oct. 16, 2016:
  "Americans actually have no idea" about how often police use force because nobody has collected enough data.

In a speech to police chiefs on Oct. 16, FBI Director James Comey said videos of police shootings have given the public an inaccurate impression that there's an epidemic of police violence against black people. (Editor's note: This video contains breaks and a facial-recognition square from the source.) (Youtube/fbi)

Issues in government data

- Unreliable partial compliance between local agencies and federal government
    - Massively undercounts deaths *[Banks et al. 2015 (BJS/DOJ), Lum and Ball 2015 (HRDAG, external)]*
- [Compare: voluntary participation approaches, e.g. National Justice Database]

Estimated number of law enforcement homicides and percent not reported, by data source, 2003–2009 and 2011



Alternative: news media reports

- Populate a database by manually reading news articles (filtered by keyword search)
  - FatalEncounters.org, KilledByPolice.net, The Guardian, Washington Post...
    - FE: volunteers have read 2M articles or ledes (!)
    - Augment with open records requests
  - BJS, Dec. 2016: media reports double the count compared to previous government collection efforts
- Secondary vs primary sources

Computational approach

- Goal: extract fatality records from a news corpus
  - Off-the-shelf event extractors work poorly (ACE, FrameNet training/ontologies)
  - Instead, train models for this problem (distant supervision+EM)
- NLP and social analysis
  - Concrete, real-world tasks useful testbed for NLP research
  - Can NLP offer something useful for important tasks?
- Public data and government accountability

Computational approach

- July 17, 2014
- Aug 9, 2014
- July 5, 2016
- July 6, 2016

Eric Garner	New York, NY
Michael Brown	Ferguson, MO
Alton Sterling	Baton Rouge, LA
Philando Castile	Falcon Heights, MN

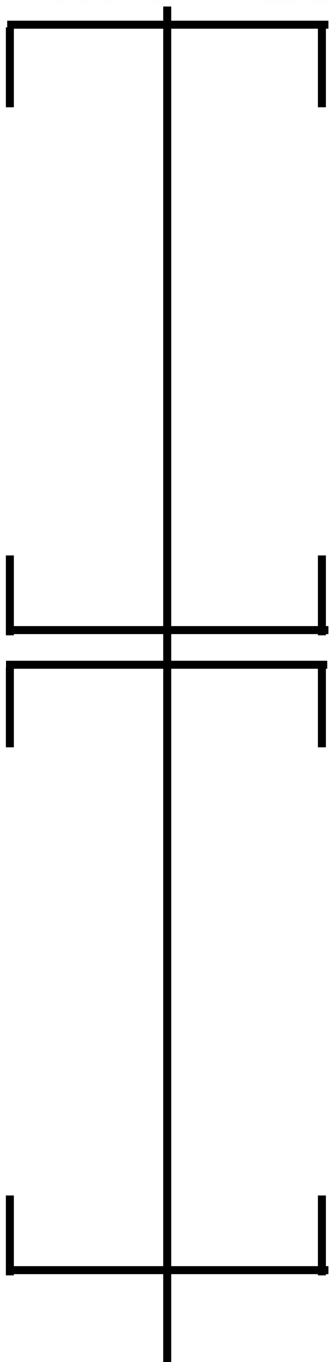
Task: Database Population

Time-delimited corpus

Infer names of persons killed by police during that timeframe

Alton Sterling	Baton Rouge, LA
Philando Castile	Falcon Heights, MN

Task: Database Update

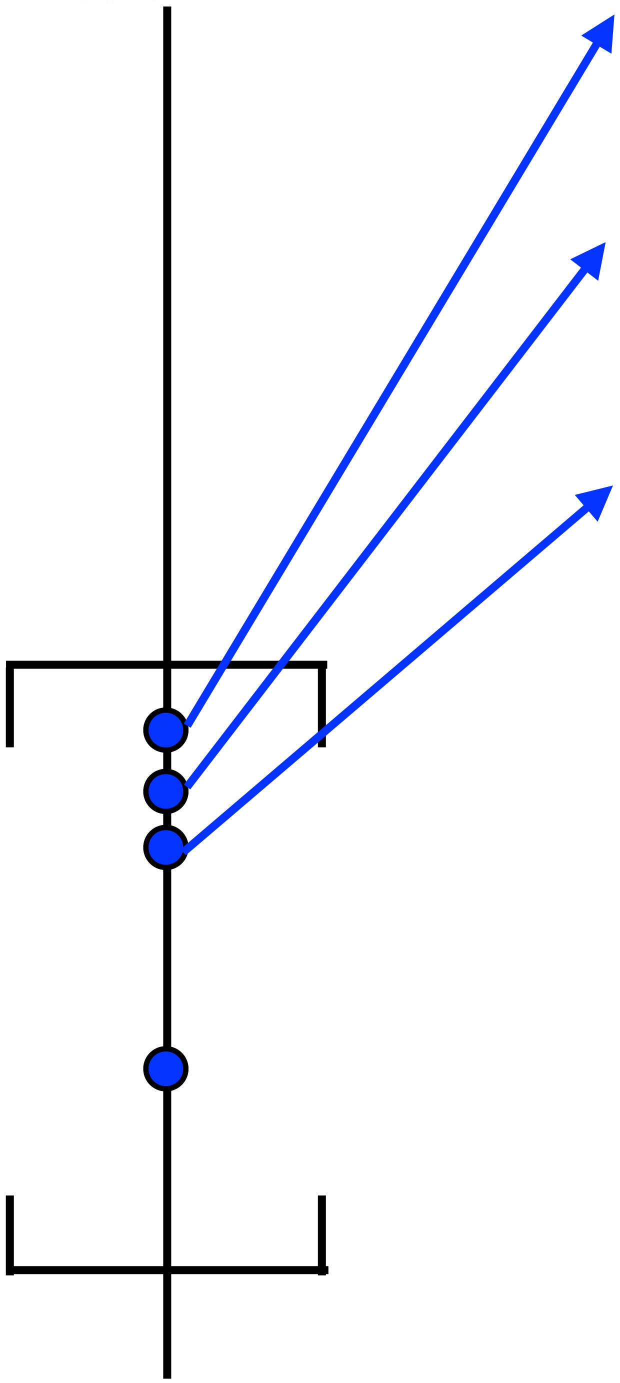


Historical data
(Distant supervision)

Testing/Runtime

Eric Garner	New York, NY
Michael Brown	Ferguson, MO
Alton Sterling	Baton Rouge, LA
Philando Castile	Falcon Heights, MN

Mentions



The Baton Rouge Police Department confirms that confirms **Alton Sterling** , 37 , died during a shooting at the Triple S Food Mart

... the two officers involved in Tuesday 's shooting of **Alton Sterling** ...

... **Alton Sterling** was a resident of Baton Rouge...

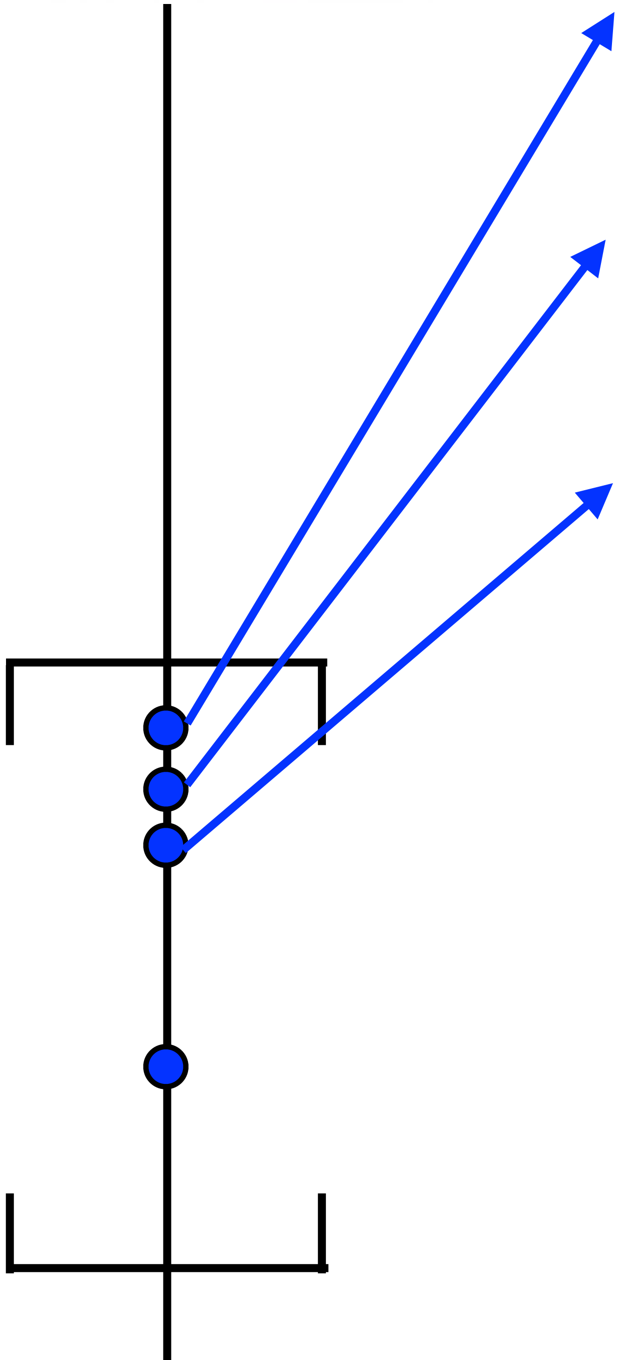
Mentions

predict: describes police fatality?

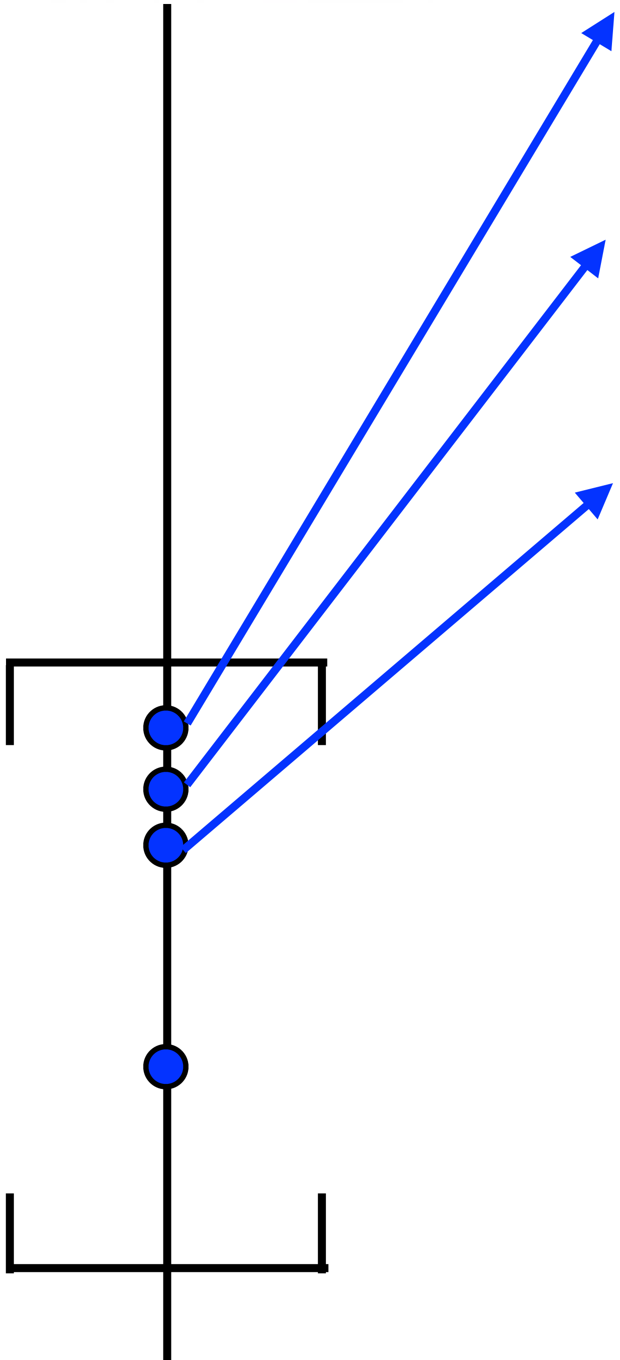
The Baton Rouge Police Department confirms that confirms **Alton Sterling** , 37 , died during a shooting at the Triple S Food Mart — 0.4

... the two officers involved in Tuesday 's shooting of **Alton Sterling** ... — 0.8

... **Alton Sterling** was a resident of Baton Rouge... — 0.01



Mentions



The Baton Rouge Police Department confirms that confirms **Alton Sterling** , 37 , died during a shooting at the Triple S Food Mart

... the two officers involved in Tuesday 's shooting of **Alton Sterling** ...

... **Alton Sterling** was a resident of Baton Rouge...

predict: describes police fatality?

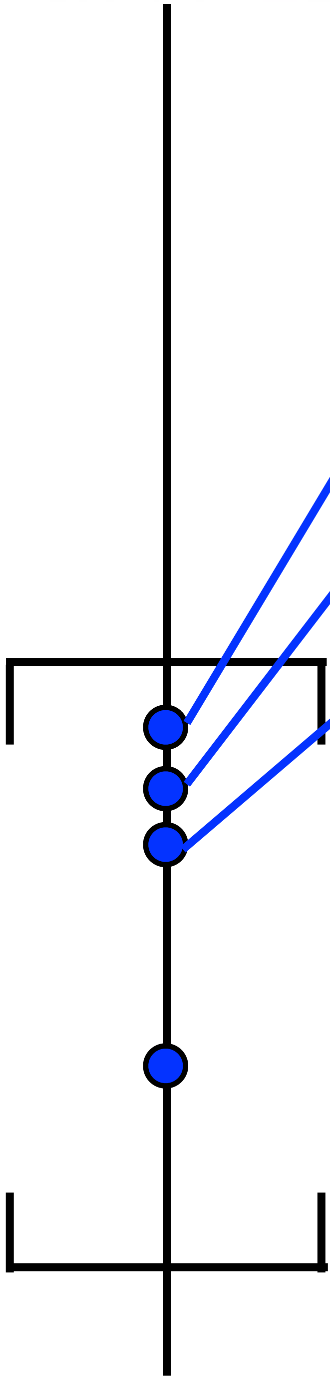
0.4

0.8

0.01

aggregate: add to database?

Mentions



predict: describes police fatality?

The Baton Rouge Police Department confirms that confirms **Alton Sterling** , 37 , died during a shooting at the Triple S Food Mart

0.4

... the two officers involved in Tuesday 's shooting of **Alton Sterling** ...

0.8

... **Alton Sterling** was a resident of Baton Rouge...

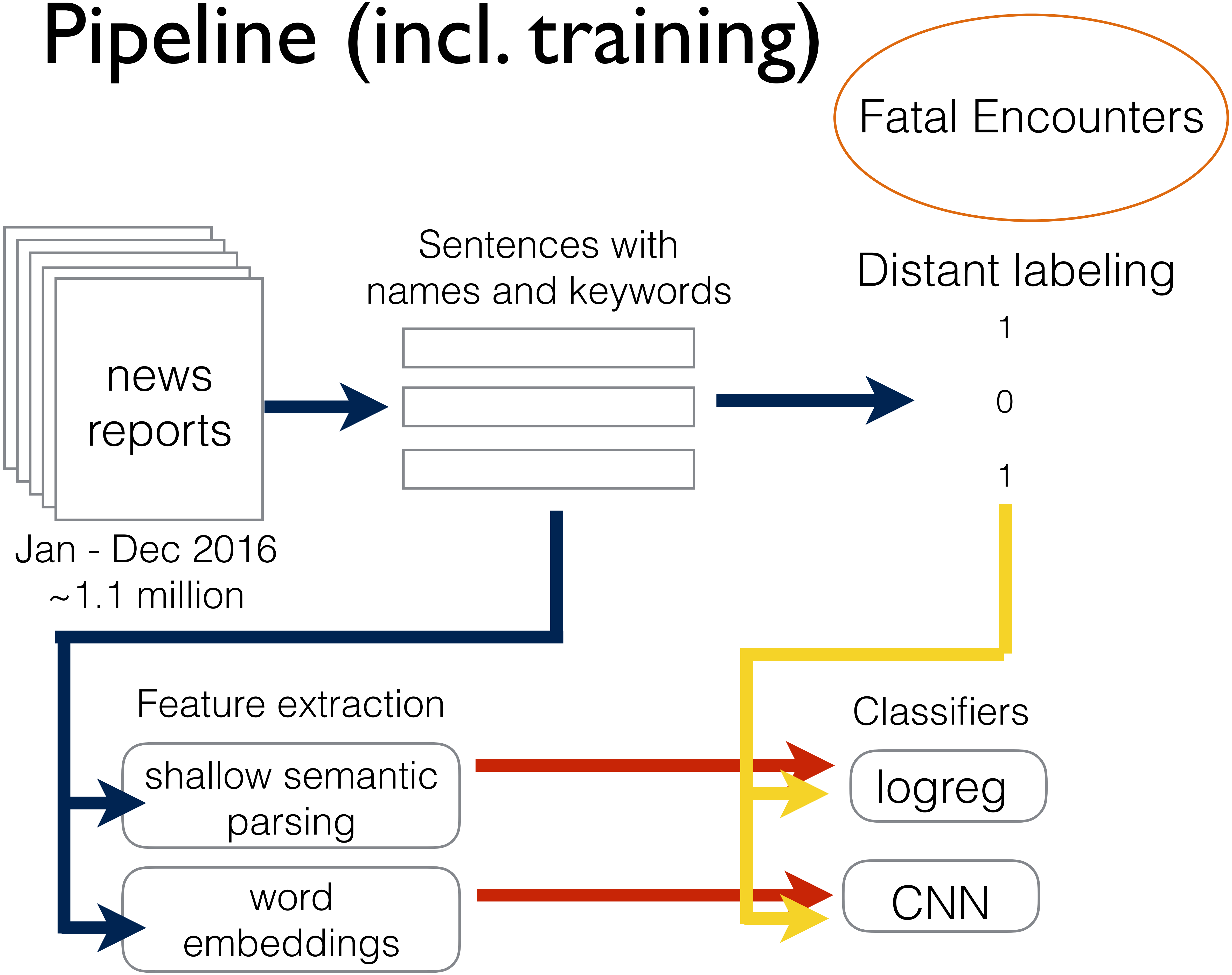
0.01

aggregate: add to database?

Alton Sterling

Entity-level fatality record (name)

Pipeline (incl. training)



Data

FATAL ENCOUNTERS

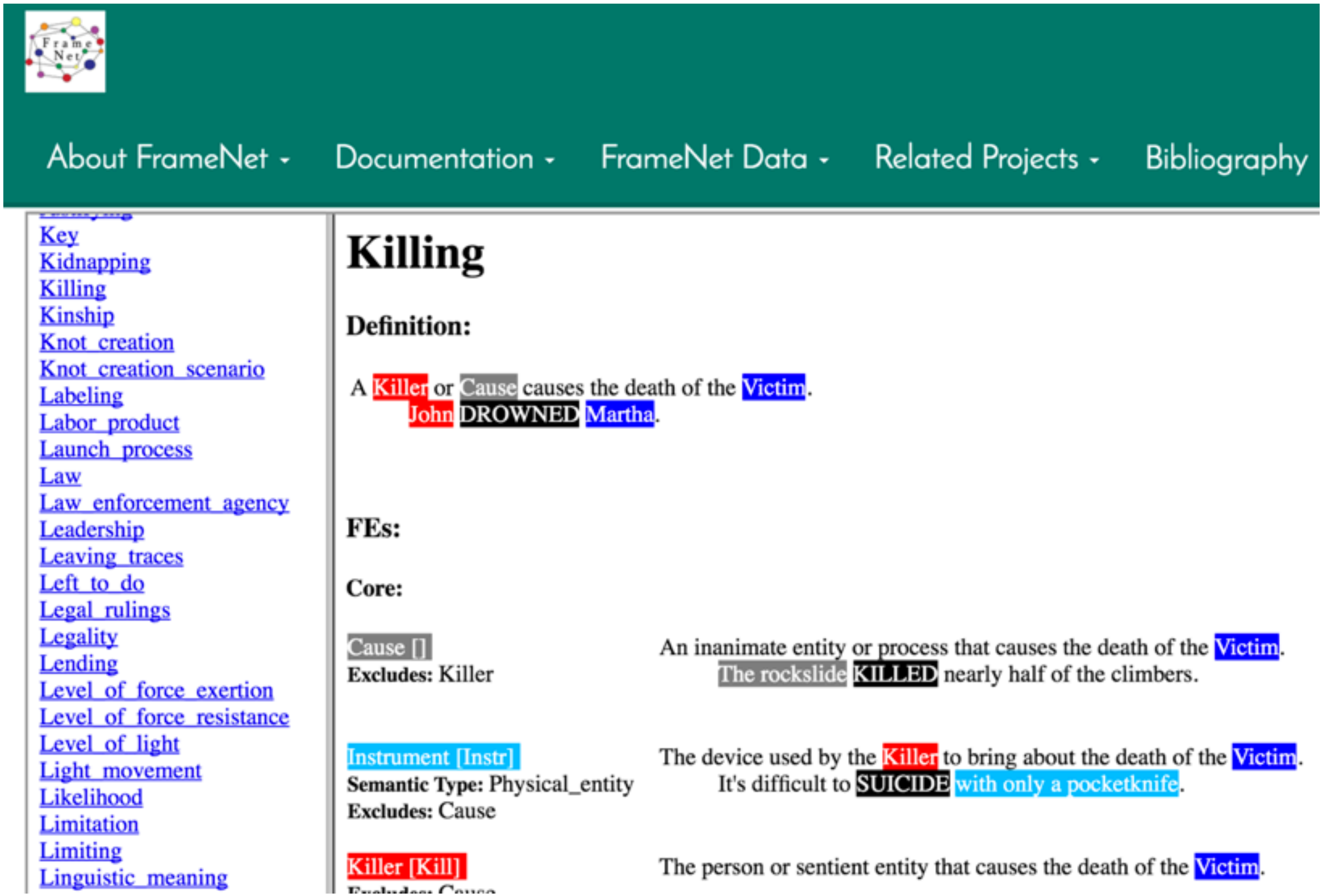
Knowledge base	Historical	Test
FE incident dates	Jan 2000 – Aug 2016	Sep 2016 – Dec 2016
FE gold entities ($\mathcal{G}$)	17,219	452

Google news

News dataset	Train	Test
doc. dates	Jan 2016 – Aug 2016	Sep 2016 – Dec 2016
total docs. ($\mathcal{D}$)	793,010	317,345
total ments. ($\mathcal{M}$)	132,833	68,925
pos. ments. ($\mathcal{M}^+$)	11,274	6,132
total entities ($\mathcal{E}$)	49,203	24,550
pos. entities ($\mathcal{E}^+$)	916	258

- Keyword-querying web scraper running throughout 2016
- Preprocessing: text extraction, NER+parsing

Can NLP help?

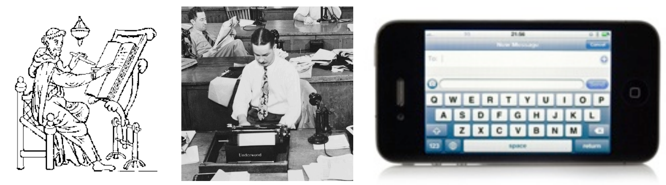

Frame Net
About FrameNet
Documentation
FrameNet Data
Related Projects
Bibliography
Key
Kidnapping
Killing
Kinship
Knot_creation
Knot_creation_scenario
Labeling
Labor_product
Launch_process
Law
Law_enforcement_agency
Leadership
Leaving_traces
Left_to_do
Legal_rulings
Legality
Lending
Level_of_force_exertion
Level_of_force_resistance
Level_of_light
Light_movement
Likelihood
Limitation
Limiting
Linguistic_meaning
Killing
Definition:
A Killer or Cause causes the death of the Victim.
John DROWNED Martha.
FEs:
Core:
Cause []
Excludes: Killer
An inanimate entity or process that causes the death of the Victim.
The rockslide KILLED nearly half of the climbers.
Instrument [Instr]
Semantic Type: Physical_entity
Excludes: Cause
The device used by the Killer to bring about the death of the Victim.
It's difficult to SUICIDE with only a pocketknife.
Killer [Kill]
The person or sentient entity that causes the death of the Victim.

Evaluations

Society *(SocialAttributes)*

Data generation process

Writing *(TextGenerator)*



Text Data *(Text)*

Inference & collection

Inference

rank	name
1	**Keith Scott**
2	**Terence Crutcher**
3	Alfred Olango
4	Deborah Danner
5	Carnell Snell
6	Kajuan Raye
7	Terrence Sterling

Fatality database from FatalEncounters.org (Previous work)

Evaluation: do these correlate?

Can NLP help?

- We tried off-the-shelf event extractors
  - SEMAFOR: trained for FrameNet *[Das et al. 2014]*
  - RPI Joint Info. Extraction: trained for ACE *[Li and Ji 2014]*
    - Found useful for gun violence extraction *[Pavlick and Callison-Burch 2016]*

	Rule	Prec.	Recall	F1
SEMAFOR	R1	0.011	0.436	0.022
	R2	0.031	0.162	0.051
	R3	0.098	0.009	0.016
RPI-JIE	R1	0.016	0.447	0.030
	R2	0.044	0.327	0.078
	R3	0.172	0.168	**0.170**
Data upper bound		1.0	0.57	0.73

Can NLP help?

- We tried off-the-shelf event extractors
  - SEMAFOR: trained for FrameNet *[Das et al. 2014]*
  - RPI Joint Info. Extraction: trained for ACE *[Li and Ji 2014]*
    - Found useful for gun violence extraction *[Pavlick and Callison-Burch 2016]*

	Rule	Prec.	Recall	F1
SEMAFOR	R1	0.011	0.436	0.022
	R2	0.031	0.162	0.051
	R3	0.098	0.009	0.016
RPI-JIE	R1	0.016	0.447	0.030
	R2	0.044	0.327	0.078
	R3	0.172	0.168	**0.170**
Data upper bound		1.0	0.57	0.73

- Hard problem!
- Domain adaptation? Text cleanliness? Training data weirdness?

Model

- (1) Identify sentence-level fatality assertions

- (2) Aggregate to entity (person)-level predictions

Model

- (1) Identify sentence-level fatality assertions

$$P(z_i = 1 \mid x_i) = \sigma(\beta^\mathsf{T} f_\gamma(x_i))$$

describes police killing event?

sentence

e.g. logistic regression, convolutional neural network

Text	Person killed by police?
Alton Sterling was killed by police.	True
Officers shot and killed **Philando Castile**.	True
Officer **Andrew Hanson** was shot.	False
Police report **Megan Short** was fatally shot in apparent murder-suicide.	False

- (2) Aggregate to entity (person)-level predictions

Model

- (1) Identify sentence-level fatality assertions

$$P(z_i = 1 \mid x_i) = \sigma(\beta^\mathsf{T} f_\gamma(x_i))$$

describes police killing event?

sentence

e.g. logistic regression, convolutional neural network

Text	Person killed by police?
Alton Sterling was killed by police.	True
Officers shot and killed **Philando Castile**.	True
Officer **Andrew Hanson** was shot.	False
Police report **Megan Short** was fatally shot in apparent murder-suicide.	False

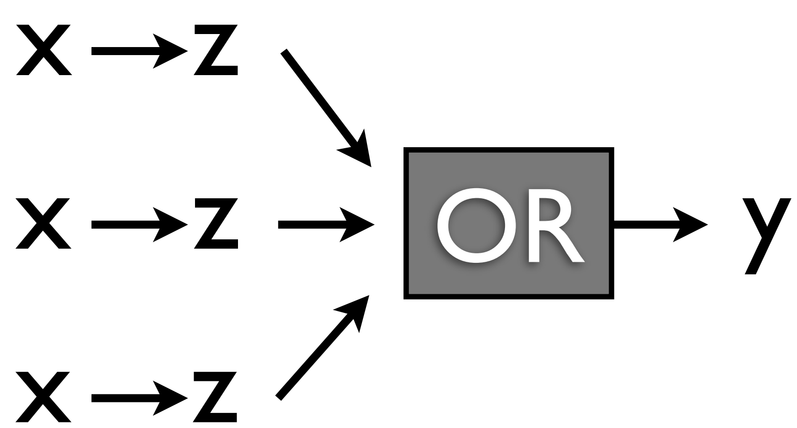
- (2) Aggregate to entity (person)-level predictions

$$P(y_e = 1 | x_{\mathcal{M}(e)})$$

was person **e** killed by police?

all sentences mentioning person **e**

Model



- Prediction through disjunction:
  - Decide an entity was killed by police, if at least one of their sentences asserts they were killed by police
- Integrate over x→z uncertainty: *noisyor* *[e.g. Craven and Kumlien 1999]*

$$P(y_e = 1 | x_{\mathcal{M}(e)}) = 1 - \prod_{i \in \mathcal{M}(e)} (1 - P(z_i = 1 \mid x_i))$$

was person **e** killed by police?

all sentences mentioning person **e**

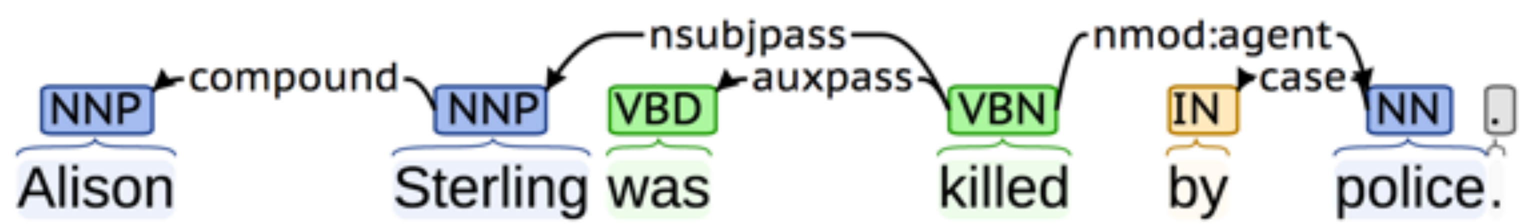
Mention-level models

$$P(z_i = 1 \mid x_i) = \sigma(\beta^{\mathsf{T}} f_\gamma(x_i))$$

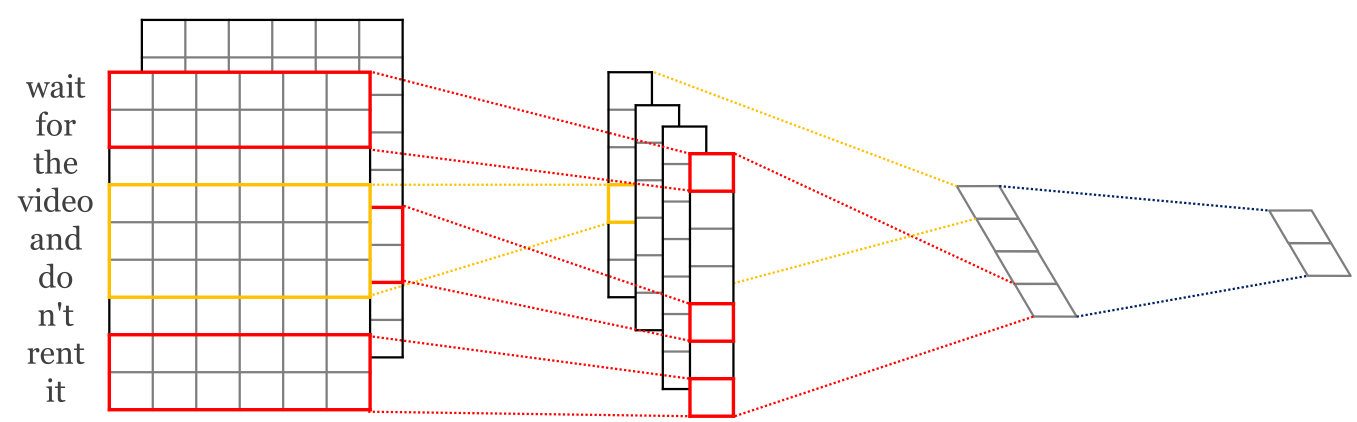
describes police killing event?

sentence

1. Feature-engineered logistic regression
   - Syntactic dependency paths
   - N-grams



2. Convolutional neural network *[e.g. Nguyen and Grishman 2015]*



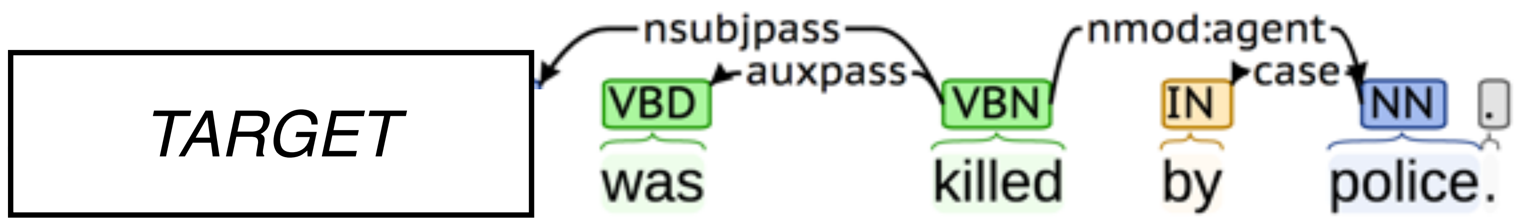
Mention-level models

$$P(z_i = 1 \mid x_i) = \sigma(\beta^{\mathsf{T}} f_\gamma(x_i))$$

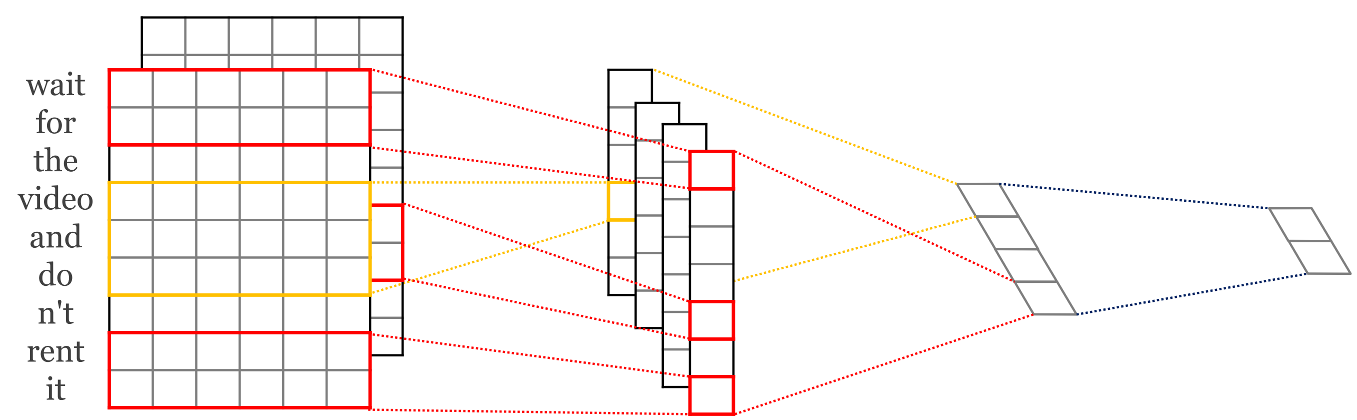
describes police killing event?

sentence

1. Feature-engineered logistic regression
   - Syntactic dependency paths
   - N-grams



2. Convolutional neural network *[e.g. Nguyen and Grishman 2015]*



Distant supervision

entities ($e \in \mathcal{E}$)	entity label (y_e)	sentences (x_i)	sent. label (z_i)
Katy Perry	0	"Katy Perry reacts to police killings."	0
Alton Sterling	1	"Alton Sterling was killed by police."	?
		"Alton Sterling was a resident of Baton Rouge.	?

e not in database:
enforce hard 0 label

e in database:
assume *at least one* is positive
(latent variable!)

- Multiple instance learning *[Bunescu and Mooney 2007]*
  - Much more accurate than assuming every sentence asserts the event!
- Probabilistic joint training: account for this uncertainty by maximizing marginal likelihood

$$P(y \mid x) = \sum_z P(y \mid z) P_\theta(z \mid x)$$

EM Training *[Dempster et al. 1977]*

E-step: posterior inference given at-least-one disjunction

$$q(z_i) := P(z_i \mid x_{\mathcal{M}(e_i)}, y_{e_i})$$

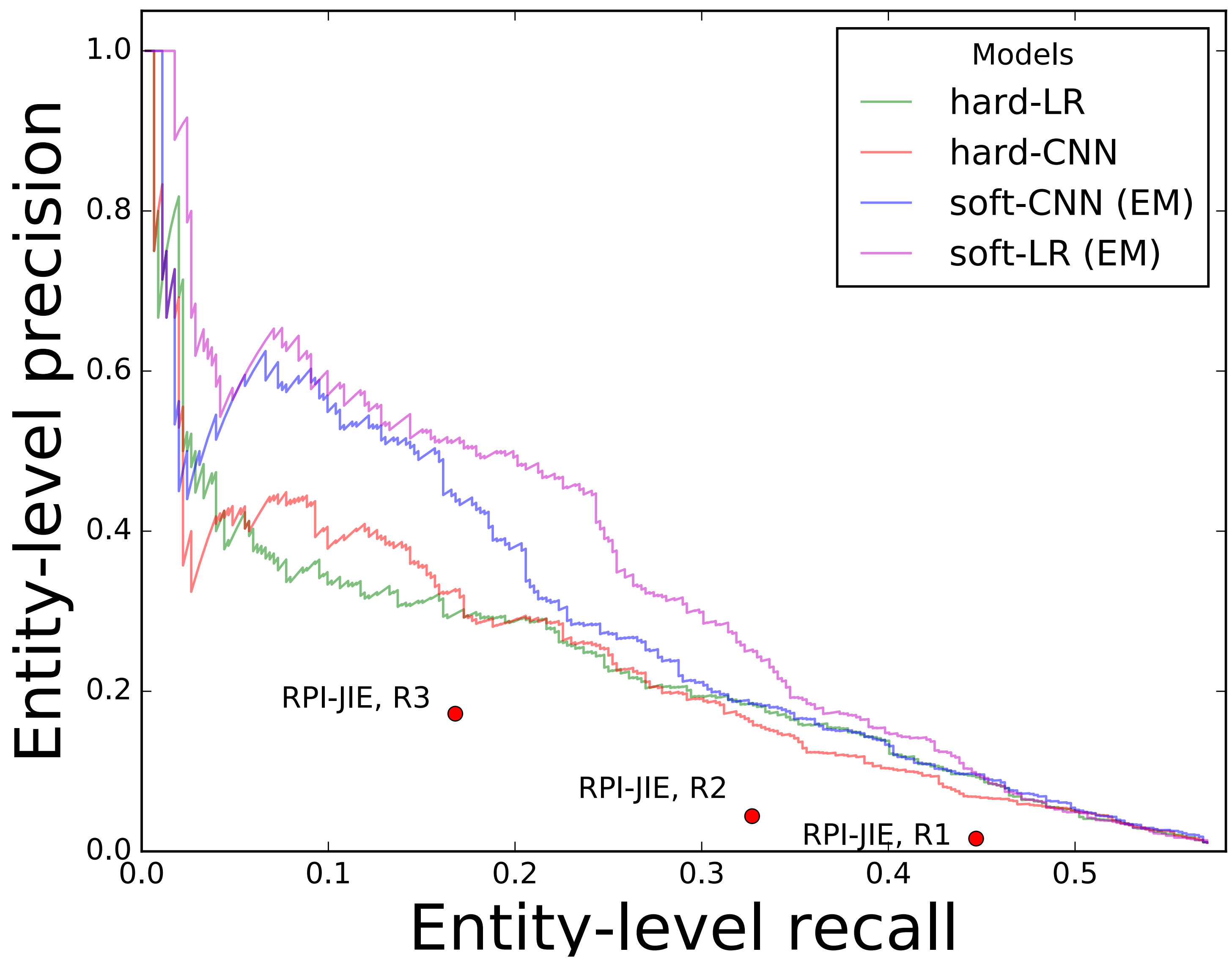
M-step: use soft labels

$$\max_{\theta} \sum_i \sum_{z \in \{0,1\}} q(z_i = z) \log P_\theta(z_i = z \mid x_i)$$

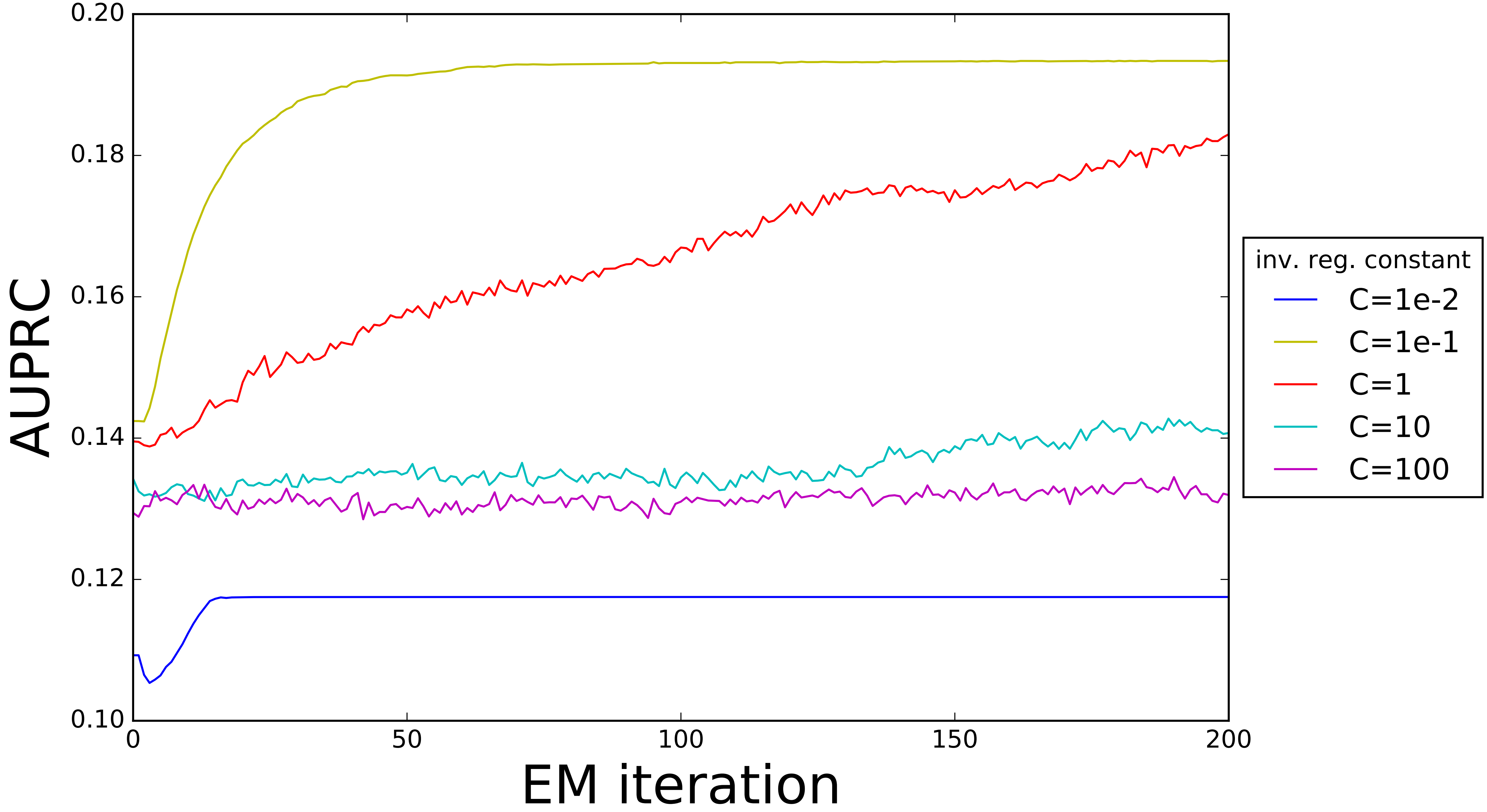
- Logistic regression: full M-step (convex opt., L-BFGS)
- Neural network: several epochs of stochastic gradient descent (Adagrad)
  - Similar to: Expected Conjugate Gradient *[Salakhutdinov et al. 2003]*
- Staged initialization (log.reg. training is nonrandom :))

Results

Model	AUPRC	F1
hard-LR, dep. feats.	0.117	0.229
hard-LR, n-gram feats.	0.134	0.257
hard-LR, all feats.	0.142	0.266
hard-CNN	0.130	0.252
soft-CNN (EM)	0.164	0.267
soft-LR (EM)	**0.193**	**0.316**
Data upper bound (§4.6)	0.57	0.73



EM Training Logistic regression



EM Training Neural network

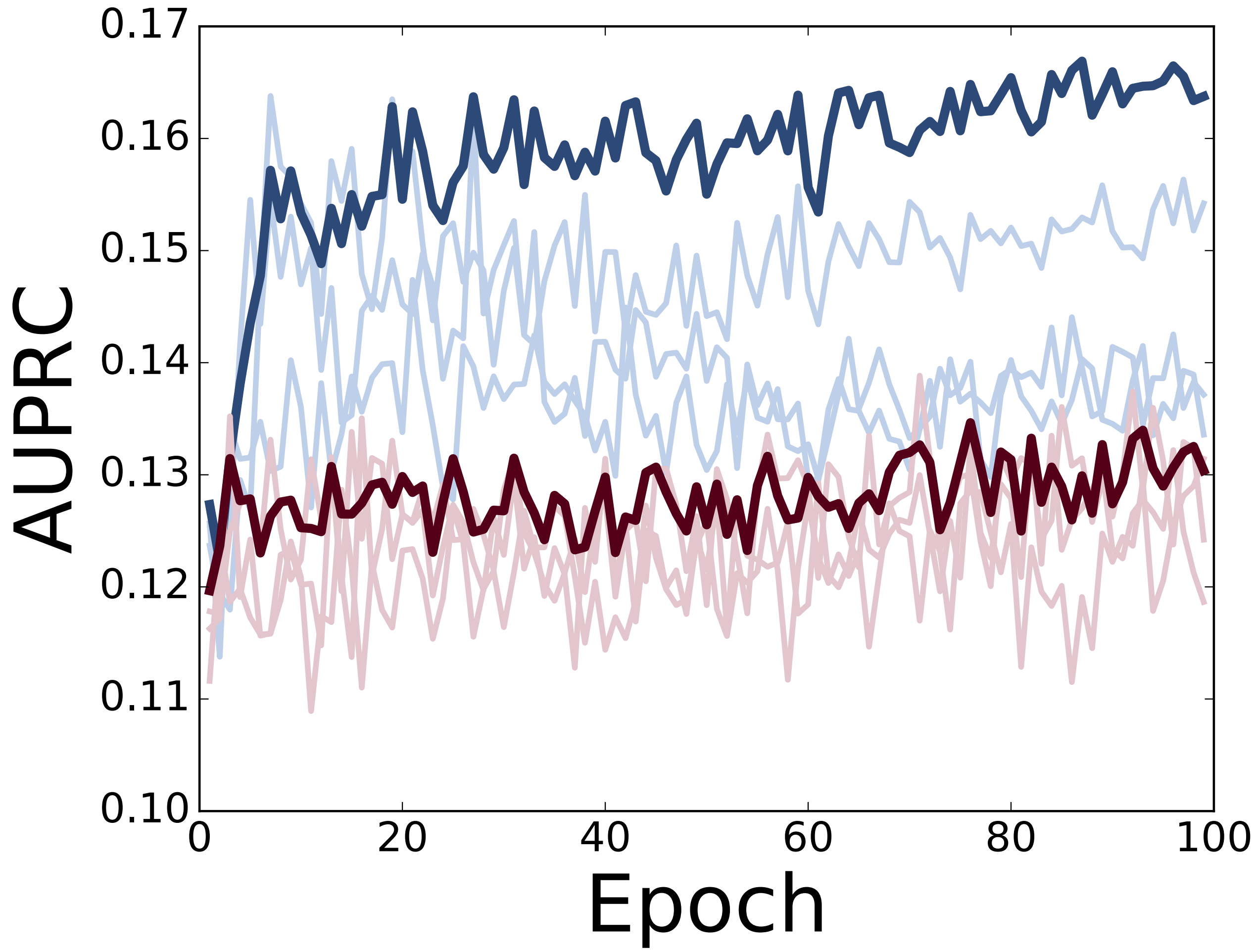
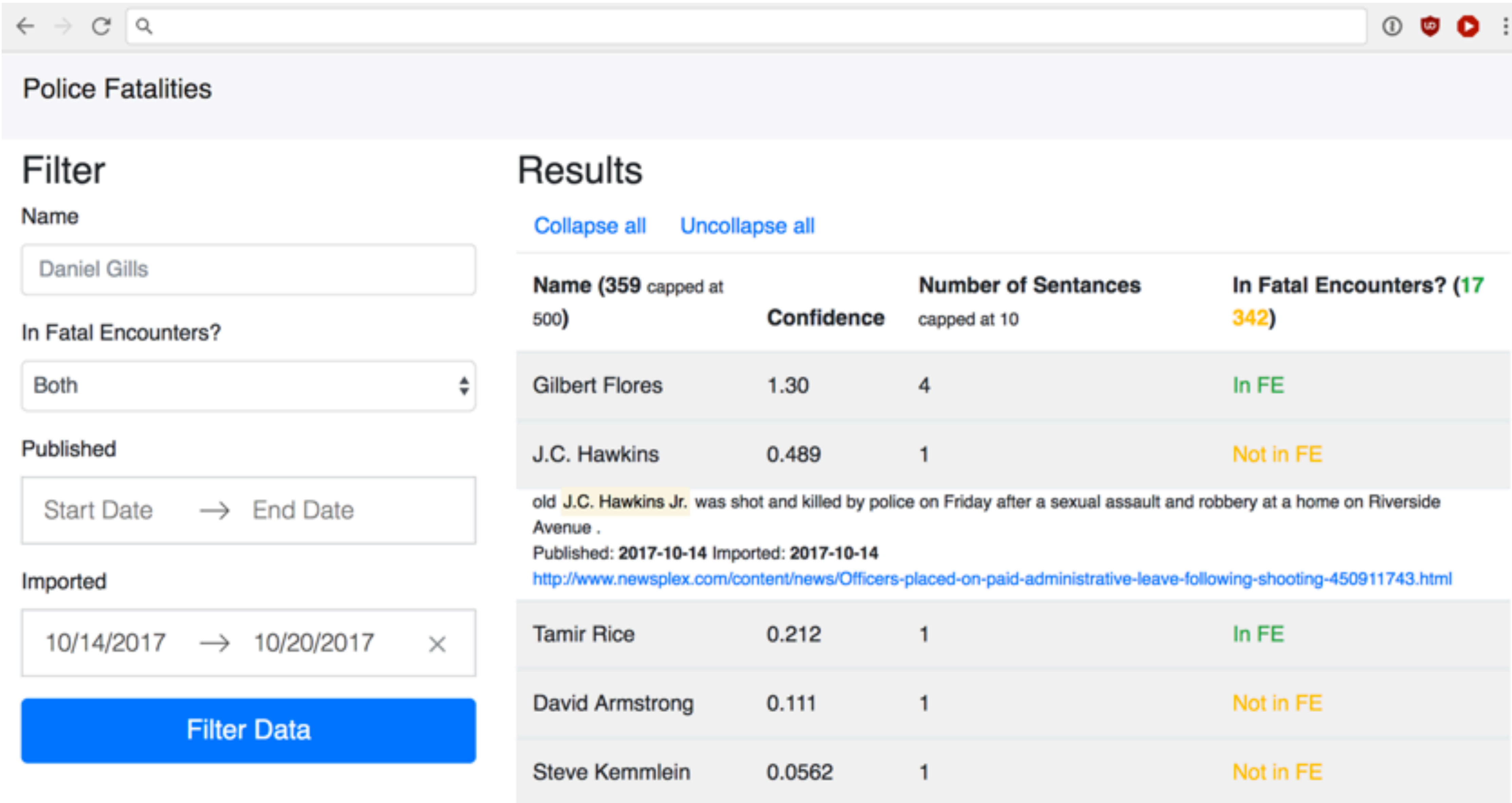


Figure 3: Test set AUPRC for three runs of soft-CNN (EM) (**blue**, higher in graph), and hard-CNN (**red**, lower in graph). Darker lines show performance of averaged predictions.

Interface for practitioners



- Fatal Encounters has been using our monitoring system for weekly updates -- ongoing work
- Dozens of cases and updates found

Predictions

entity (e)	ment.(i) prob.	ment. text (x_i)
Keith Scott (true pos)	0.98	Charlotte protests Charlotte's Mayor Jennifer Roberts speaks to reporters the morning after protests against the police shooting of **Keith Scott**, in Charlotte, North Carolina .
Terence Crutcher (true pos)	0.96	Tulsa Police Department released video footage Monday, Sept. 19, 2016, showing white Tulsa police officer Betty Shelby fatally shooting **Terence Crutcher,** 40, a black man police later determined was unarmed.
Mark Duggan (false pos)	0.97	The fatal shooting of **Mark Duggan** by police led to some of the worst riots in England's recent history.
Logan Clarke (false pos)	0.92	**Logan Clarke** was shot by a campus police officer after waving kitchen knives at fellow students outside the cafeteria at Hug High School in Reno, Nevada, on December 7.

Table 7: Example of highly ranked entities, with selected mention predictions and text.

Predictions: top-ranked

rank	name	positive	analysis
1	**Keith Scott**	**true**	
2	**Terence Crutcher**	**true**	
3	Alfred Olango	true	
4	Deborah Danner	true	
5	Carnell Snell	true	
6	Kajuan Raye	true	
7	Terrence Sterling	true	
8	Francisco Serna	true	
9	Sam DuBose	false	name mismatch
10	Michael Vance	true	
11	Tyre King	true	
12	Joshua Beal	true	
13	Trayvon Martin	false	killed, not by police
14	**Mark Duggan**	**false**	**non-US**
15	Kirk Figueroa	true	
16	Anis Amri	false	non-US
17	**Logan Clarke**	**false**	**shot not killed**
18	Craig McDougall	false	non-US
19	Frank Clark	true	
20	Benjamin Marconi	false	name of officer

Int’l relations events via knowledge engineering

[Schrodt 1994, Leetaru and Schrodt 2013]

Event classes
(~200)

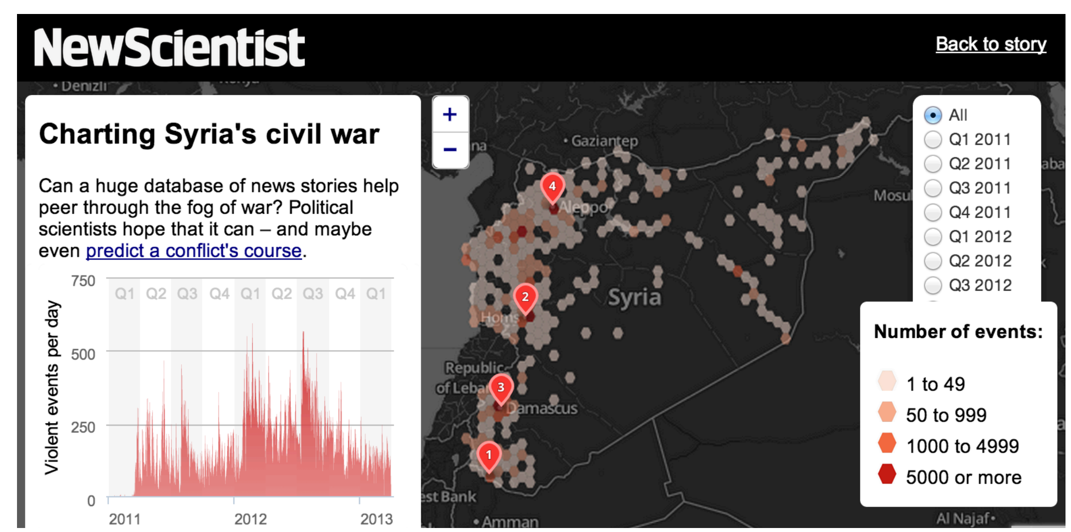
Dictionary:
Verb patterns per event class
(~15000)

Extract events from news text
for pairs of countries

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

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

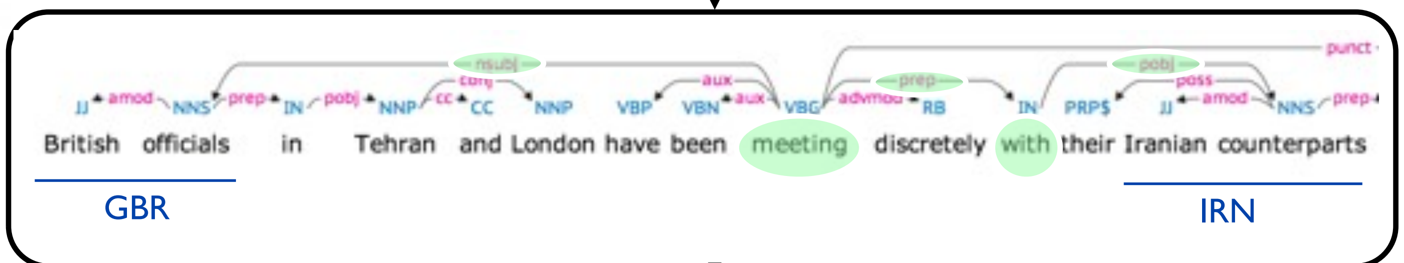


Issue: Hard to maintain and adapt to new domains

Unsupervised learning for int'l relations

Data: twenty years of news articles

Natural Language Processing



Event phrases of actor interactions

Probabilistic Graphical Model

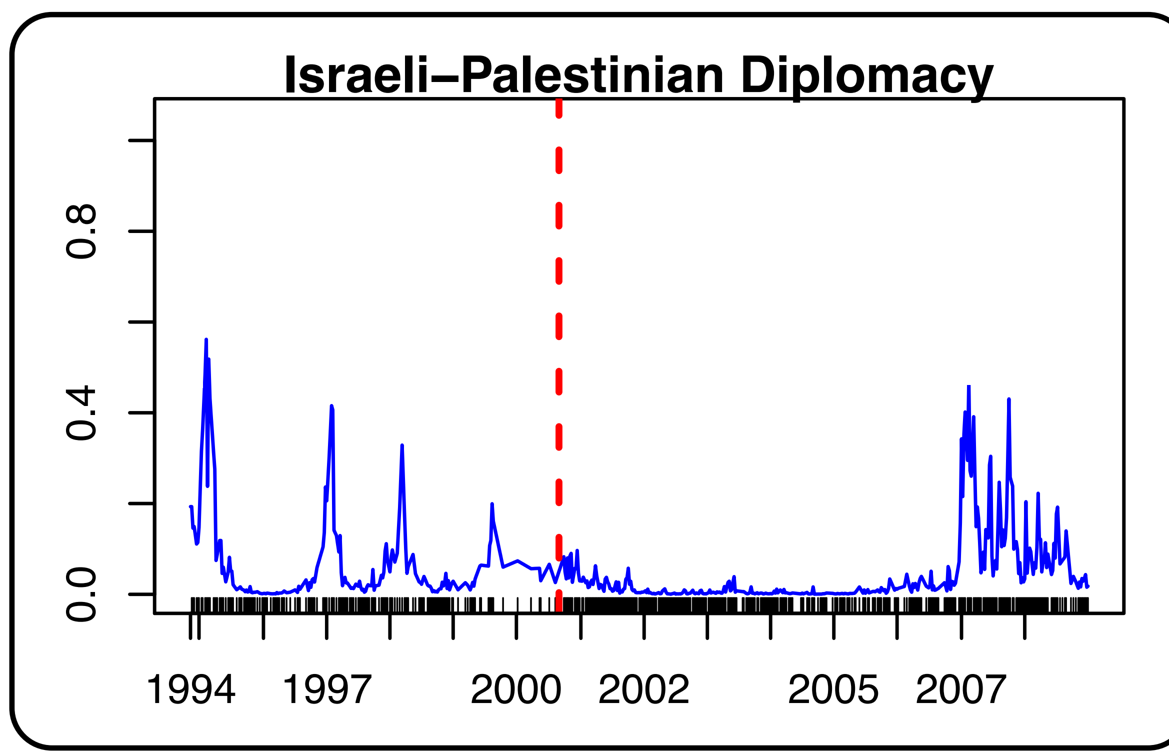
Purely from textual data, jointly learns both

(1) **Event class dictionaries**

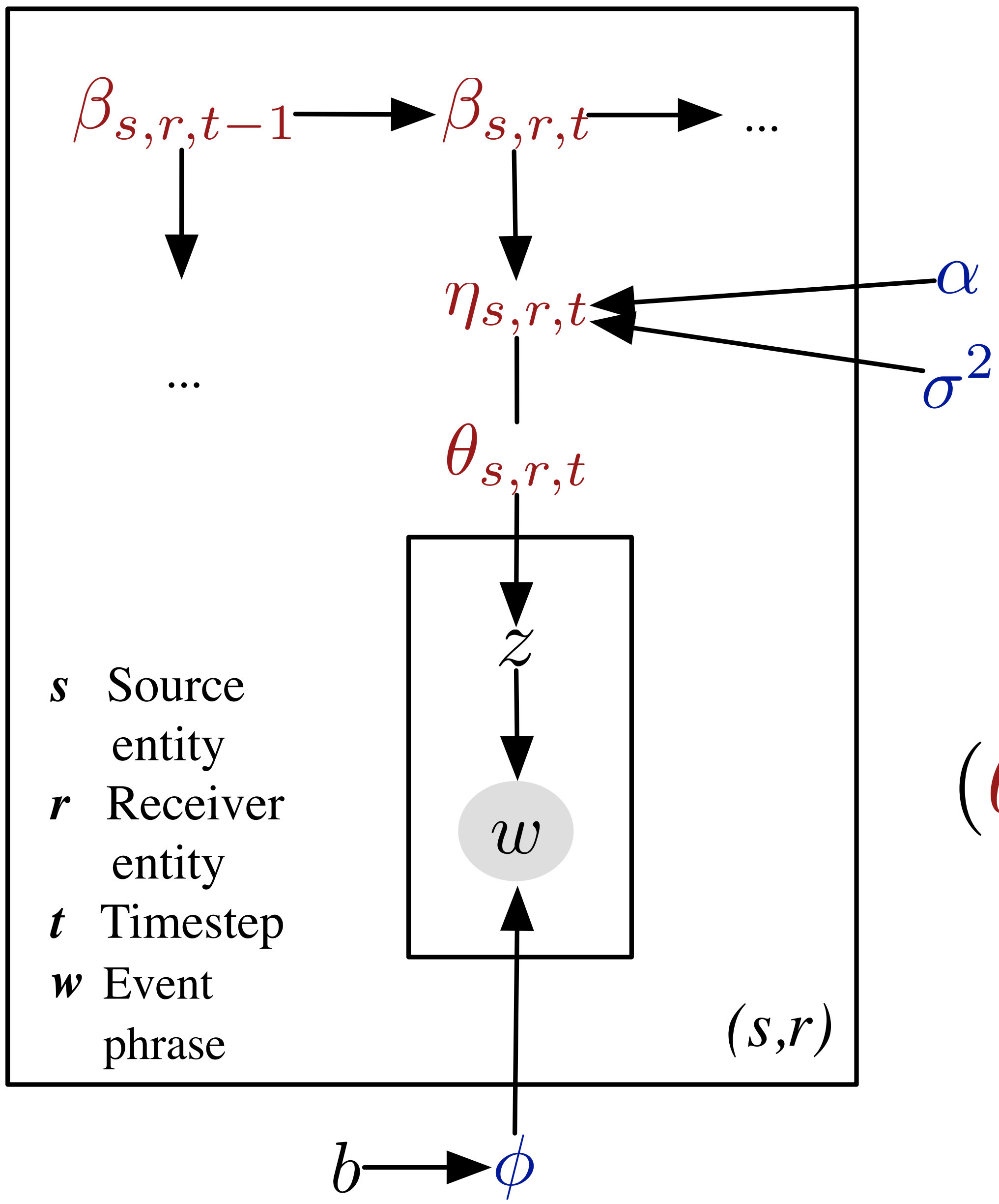
(2) **Political dynamics**

"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,
"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,
"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

↔



Model



Event prior models

M1: independent contexts

M2: temporal smoothing

[Blei and Lafferty 2006, Quinn and Martin 2002]

Adjacent timestep similarity

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

$$\eta_{s,r,t} \sim N(\alpha + \beta_{s,r,t}, \mathrm{Diag}[\sigma_1^2..\sigma_K^2])$$

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

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

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

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

$$\phi_k \sim \mathrm{Dir}(b)$$

K=100 ⟶ 80 million parameters

Social event data extraction

- Natural language processing can help *acquire more behavioral data* from news
  - Police killings
  - International relations
  - Protests *[Hanna 2017]*
  - Gun violence *[Pavlick et al. 2016]*
  - Europe Media Monitor *[Piskorski et al. 2011]*
- Assumes media production reflects reality
  - Alternative: analyze e.g. media bias/attention, as in political science or literature analysis
- NLP and social analysis
  - Concrete, real-world tasks useful testbed for NLP research
  - NLP could offer something useful for important tasks!

Thanks!

- Police Killings project:
  http://slanglab.cs.umass.edu/PoliceKillingsExtraction/

- Others:
  http://brenocon.com/