Content Analysis

Lecture 2: Classification, clustering, and topic modeling

2 May, 2016

Prof. Andrew Eggers

Example: Speed ("Do newspapers now give the news?" 1893)

Characterizing content of New York newspapers (based on 13 topics) on two Sundays 12 years apart.*

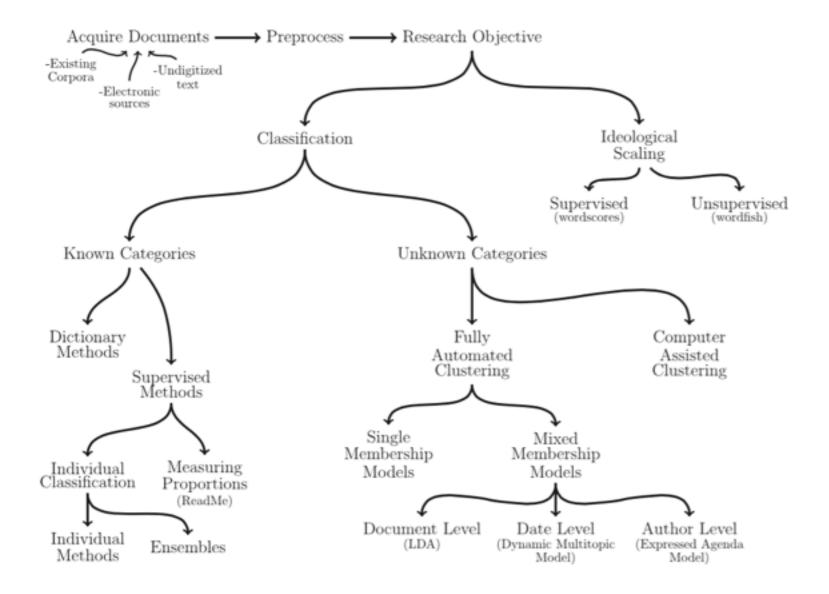
COLUMNS OF READING-MATTER IN NEW YORK NEWSPAPERS, APRIL 17, 1881, AND APRIL 16, 1893.

Subject.	Tribune, 1881.	Tribune, 1893.	World, 1881.	World, 1893.	Times, 1881.	Times, 1893.	Sun, 1881.	Sun, 1893,
Editorial	5.00 2.00	5.00 0.00	4.75 0.75	4.00 0.00	6.00	5.00 0.00	4.00 0.50	4.00 1.00
Scientific	1.00	0.75 3.75	0.00	2.00 10.50	1.00	0.00 4.00	0.00	2.50 3.50
Literary	15.00	5.00	1.00	2.00	18.00	12.00	5.75	6.00
Gossip	0.00	23.00 1.50	1.00	1.50	1.00	16.75 2.50	2.00 0.00	2.00
Sporting	0.00	6.50 7.00	2.50 1.50	16.00 6.50	3.00 1.00	1.50	0.50	17.50 11.50
Historical Music and Drama		2.50 4.00	2.75 1.50	4.00 11.00	2.50 4.00	1.50 7.00	4.25 0.00	14.00 3.50
Crimes and Criminals. Art	0.00 1.00	0.50 1.00	0.00 3.00	6.00 3.00	0.00 2.00	1.00 0.00	0.00	0.00 1.25

Conclusion: "there has been a distinct deterioration and decadence in the New York newspaper press in the last dozen years"

^{*&}quot;I wish to remark here that I selected this date in April merely by chance and not because I was aware of anything in the papers that day making them at all extraordinary."

A classification of "text-as-data" methods



Key categories of automated classification methods

Unsupervised learning

Classification with unknown categories.

"I don't even know where to start with these documents. Can I at least get a summary of what is being discussed?"

Workflow:

- Acquire and process data
- Run classification algorithm
- Try to interpret results (hard)

Supervised learning

Classification with known categories, and some classification done by humans.

"I can't classify all these documents. Can I use my classification of this subset to fill in the rest?"

Workflow:

- Acquire and process data
- Decide on classes
- Classify a subset by hand
- Run classification algorithm
- Check accuracy

Like having a robot clean your basement



Unsupervised learning

Tell robot how many piles you want.

Robot tries to put objects in piles with similar objects.

Supervised learning

You put a sample of items into piles.

Robot tries to organize the rest the same way.

The term-document matrix (TDM)

The TDM is the starting point for many text analysis techniques.

Toy corpus

Document I: "This is a document."

Document 2: "This is another document."

Document 3: "When is lunch?"

Term-document matrix

	DI	D2	D3
this	ı	ı	0
is	I	I	I
a	1	0	0
document	I	I	0
another	0	1	0
when	0	0	I
lunch	0	0	I

Choices in making a term-document matrix:

- stemming? ("trying" => "tri")
- lower case? ("This" => "this"?)
- remove "stop words"? (keep "is", "a"?)

```
> require(tm)
```

> stemDocument(PlainTextDocument("stemming is not that difficult honestly"))

```
<<PlainTextDocument (metadata: 7)>> stem is not that difficult honest
```

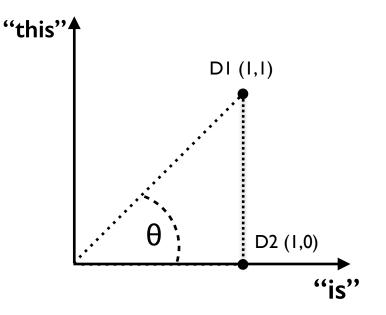
Unsupervised learning: clustering algorithms



How would the robot try to group similar documents together? First, decide on a measure of similarity (or distance).

Potential measures of similarity/distance between two documents (vectors):

- Correlation of column vectors
- Euclidean distance between column vectors (in n-dimensional space)
- Cosine of angle between column vectors



Term-document matrix

	DI	D2	D3
this	I	I	0
is	I	1	I
a	I	0	0
document	I	1	0
another	0	1	0
when	0	0	I
lunch	0	0	I

Unsupervised learning: kmeans clustering



Given a measure of similarity/distance, how do we assign documents to groups?

Intuition for k-means clustering:

- Goal: Assign documents into k clusters based on similarity
- **Input:** The documents, the number of clusters e.g. k=2
- Output: Cluster assignments (e.g. cluster 1: {D1, D2}; cluster 2: {D3})
- Objective function: Minimize sum (over documents & terms) of squared distance between document and its cluster's mean location

Augmented TDM (k=2)

	Assign t	o CI		Assign to C2		Total distance ² from cluster means	
	DI	D2	CI avg	D3	C2 avg		
this	I	I	1	0	0	0	
is	I	I	1	I	1	0	
a	I	0	0.5	0	0	0.5	
document	I	I	1	0	0	0	
another	0	I	0.5	0	0	0.5	
when	0	0	0	I	1	0	
lunch	0	0	0	I	I	0	

Sum:

Unsupervised learning: hierarchical clustering



Given a measure of similarity/distance, how do we assign documents to groups?

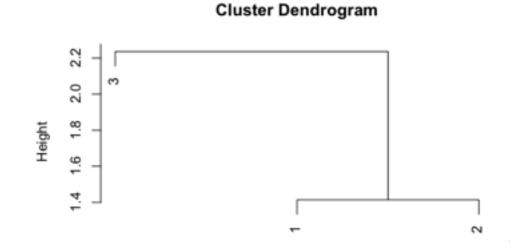
Intuition for hierarchical clustering:

- Goal: Assign documents into clusters based on similarity
- **Input:** The distance matrix for the documents
- Output: Cluster assignments at each stage of the clustering; cluster dendrogram
- Algorithm: Start with each document in its own cluster. Join the most similar clusters together & recalculate distances. Repeat.

Term-document matrix

	DI	D2	D3	
this	1		0	-
is	I	I	I	
a	I	0	0	
document	I	ı	0	
another	0	I	0	
when	0	0	I	
lunch	0	0	ı	

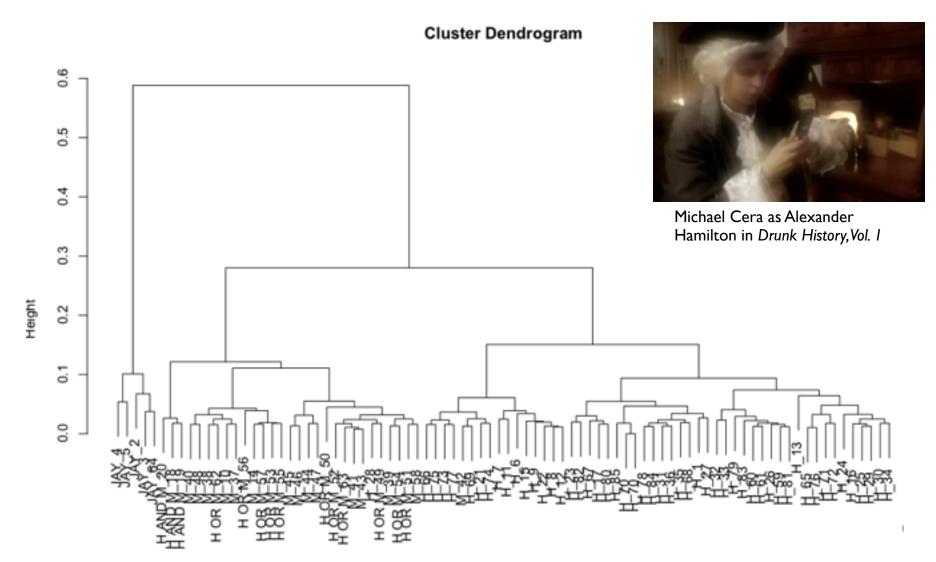
> dtm = rbind(c(1,1,1,1,0,0,0), c(1,1,0,1,1,0,0), c(0,1,0,0,0,1,1))
> plot(hclust(dist(dtm)))



Unsupervised learning: hierarchical clustering (2)



Hierarchical clustering of Federalist Papers based on stop words: solution to an authorship puzzle?



Unsupervised learning: model-based approaches



Simplest model-based methods are directly analogous to kmeans clustering: just add statistics (Bayesian/MLE)!

Think of text as having been produced by a data generating process (generative model) whose parameters we want to estimate.

- In our usual regressions, parameters are the slope coefficients
- In single-membership topic models, parameters are
 - the word frequencies for each topic
 - the topic membership of each document

Same in kmeans clustering!

Data structure and statistical theory of topic modeling

Data (Term Doc. Matrix)

Parameters to estimate

W matrix: N	word fre	•			θ matrix: N word frequencies for K topics			Z matrix: K topic labels for E documents						
	Wı	W ₂	•••	WD		θι	θ_2	•••	θ_{K}		Ζı	Z 2	•••	z D
word I	WII	W 21	•••	W DI	word I	θπ	θ21	•••	θκι	topic I	ZII	Z 21	•••	Z DI
word 2	WI2	W 22	•••	W D2	word 2	θ12	θ_{22} θ_{23}	•••	θ_{K2}	topic 2	Z 12	Z 22		Z D2
word 3	W 13	W 23	•••	W D3	word 3	θ13	θ_{23}	•••	θ_{K3}	topic 3	Z 13	Z 23	•••	Z D3
•••			•••	•••	•••	•••	•••	•••	•••	•••	•••	•••	•••	•••
word N	WIN	W2N	•••	WDN	word N	θιΝ	θ_{2N}	•••	θ_{KN}	topic K	ZIK	Z 2K	•••	Z DK

MLE version: choose θ , Z to maximize $Pr(W|\theta, Z)$

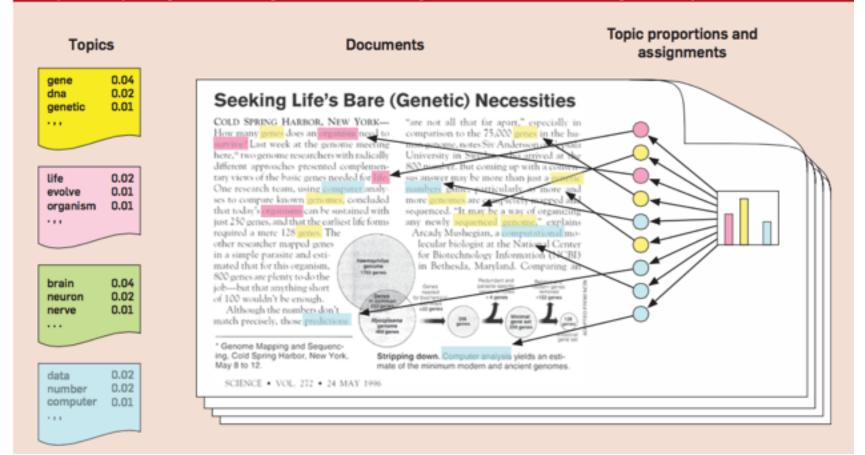
Bayesian version: describe $Pr(\theta, Z|W) \propto Pr(W|\theta, Z)Pr(\theta, Z)$

e.g. $z_2 = [0,0,...1,0,...,0]$ (Because singlemembership model)

Mixed-membership topic models: Latent Dirichlet Allocation (LDA)

Single-membership: draw a single topic for the document; draw the words from that topic Mixed membership: draw a mix of topics for the document; draw a single topic for each word; draw specific word from that topic

Figure 1. The intuitions behind latent Dirichlet allocation. We assume that some number of "topics," which are distributions over words, exist for the whole collection (far left). Each document is assumed to be generated as follows. First choose a distribution over the topics (the histogram at right); then, for each word, choose a topic assignment (the colored coins) and choose the word from the corresponding topic. The topics and topic assignments in this figure are illustrative—they are not fit from real data. See Figure 2 for topics fit from data.



Assumptions

- How many topics?
- Which words? (stop words, stemming, etc: see e.g. work of Hannah Wallach)

Implementation in R

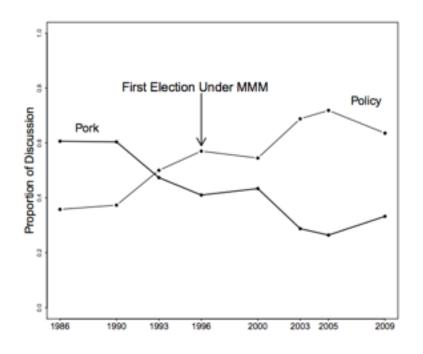
- tm package (text mining)
- lda package
- stm package: topic modeling with covariates, so we can compare topic distribution for e.g. treatment and control group, men and women, etc.
- quanteda package: many techniques in one package

Application of topic models: Catalinac (2014)

Did electoral system reform in 1994 change the incentives for parties to address national security issues?

Uses LDA to summarize topics addressed in 8,000 election manifestos.

Divides topics into "pork" and "policy".



- Sensitivity to number of topics? "we fit the model with 69 topics because this
 was the lowest specification that produced a clear national security topic and
 topics suggestive of pork and policy"
- What about dictionary methods? What about hand-coding a sample and using supervised learning?

Supervised learning: when you know what you want (almost)

		Do you kn categories in want to place o	which you	
		Yes No		
Do you know the rule for	Yes	Dictionary methods	NA	
placing documents in categories?	No	Supervised learning	Topic models	

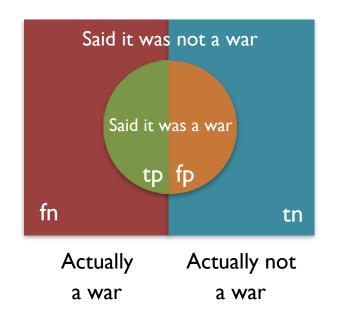
Two basic scenarios:

- You want to classify a corpus of texts. You read and classify a (random) subset. You fit a predictive model, and apply it to the unread documents.
- You want to classify a corpus of texts. A subset is already labeled.
 You fit a predictive model and apply it to the unlabeled documents.

Evaluating a classification model: binary case

Confusion matrix

	Said it was a war	Said it was not a war
Actually war	tp : true positive	fn: false negative
Actually not a war	fp: false positive	tn: true negative

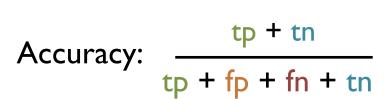


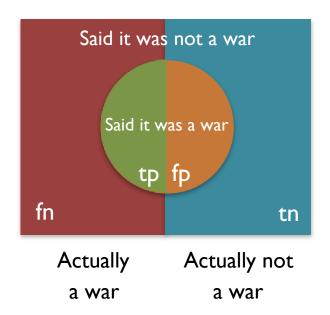
Evaluating a classification model: binary case (2)

Precision:
$$\frac{tp}{tp + fp}$$

Recall/sensitivity:
$$\frac{\mathsf{tp}}{\mathsf{tp} + \mathsf{fn}}$$

Specificity:
$$\frac{tn}{tn + fp}$$





Classification of events from news stories: ICEWS

In 2008, U.S. Defense Advanced Research Projects Agency (DARPA) launched Integrated Crisis Early Warning System (ICEWS) program.



"The overarching technical goal of the program is to automatically monitor, assess, and forecast the consequences of national and sub-national events and interactions that could affect US national security interests, and inform decisions on how to allocate DIME (diplomatic, information, military, and economic) resources to mitigate them. The tools and methodologies developed in ICEWS are designed to allow users to:

- Account for the complexity of interactions between governments and government institutions, the people they govern (or claim to govern), and non-state actors such as al-Qaeda and other similar groups that are not tied to any specific geographic location.
- Identify the generalizable patterns in these interactions (that is, "early warning indicators") that allow users to estimate with a high degree of accuracy the probability that an insurgency will develop, a civil war will occur, one or more countries will attack another with military force, or a military coup will be hatched to dispatch a current set of rulers, to name but a few examples."
- etc

O'Brien (2010), "Crisis Early Warning and Decision Support: Contemporary Approaches and Thoughts on Future Research", International Studies Review

Classification of events from news stories: ICEWS (2)

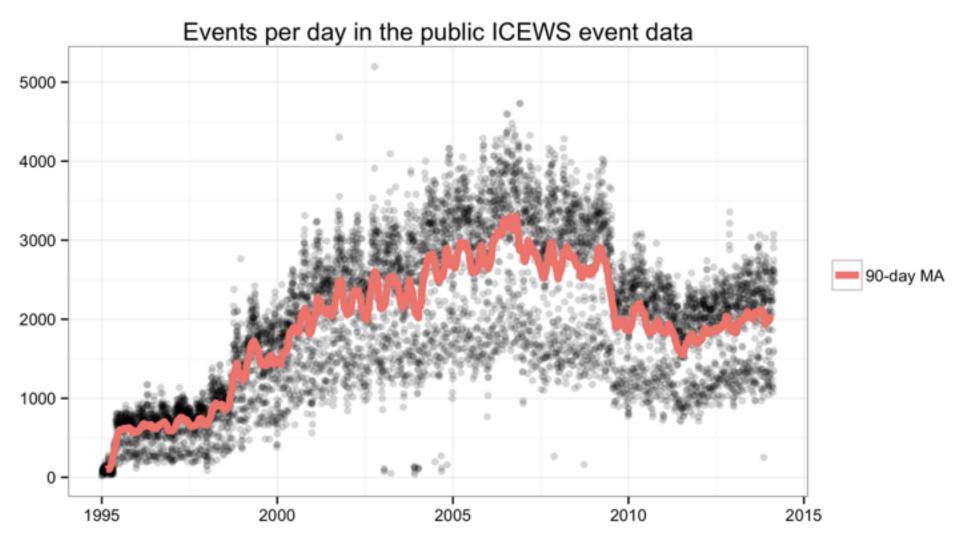


- 2008: Lockheed Martin wins DARPA competition for early warning system
- March 2015: ICEWS releases coded event data: disaggregated (one row per event) and aggregated (one dyad-year or monad-year per event)

Procedure for generating coded event data:

- Collect media reports in English, Spanish, Portuguese, French (translate to English where appropriate)
- Remove duplicate stories based on shared trigrams (remember trigrams?)
- Using first 6 sentences of each story, classify according to CAMEO event ontology (Schrodt et al) using (proprietary) ACCENT event coder (some supervised learning, using some grammar parsing)

Classification of events from news stories: ICEWS (3)



ICEWS (4): CAMEO (Conflict and Mediation Event Observations) ontology

Basic idea: An event can be classified by a (standardized) verb and actors (source and target).

Example: "Demonstrators in Ukraine called for the resignation of Prime Minister Mykola Azarov."

Event code [Verb]: 1411 (Demonstrate for leadership change)

Source actor: Protester (Ukraine)

Target actor: Mykola Azarov



Demonstrate
——for leadership
change



ICEWS (5): CAMEO (Conflict and Mediation Event Observations) ontology

Top-level categories

Sub-categories

070:[7.0] Provide aid, not specified below Make Public Statement (01) 071:[7.4] Provide economic aid Appeal (02) 072:[8.3] Provide military aid Express Intent to Cooperate (03) 073:[7.4] Provide humanitarian aid Consult (04) 074:[8.5] Provide military protection or Engage in Diplomatic Cooperation peacekeeping (05)075:[7.0] Grant asylum Material Cooperation (06) 115:[-2.0] Bring lawsuit against Provide Aid (07) -Yield (08) Investigate (09) Demand (10) Disapprove (11) -Reject (12) 110:[-2.0] Disapprove, not specified below Threaten (13) 111:[-2.0] Criticize or denounce Protest (14) 112:[-2.0] Accuse, not specified below Exhibit Military Posture (15) 1121:[-2.0] Accuse of crime, corruption Reduce Relations (16) 1122:[-2.0] Accuse of human rights abuses Coerce (17) 1123:[-2.0] Accuse of aggression Assault (18) -1124:[-2.0] Accuse of war crimes Fight (19) 1125:[-2.0] Accuse of espionage, treason Engage in Unconventional Mass 113:[-2.0] Rally opposition against Violence (20) 114:[-2.0] Complain officially 115:[-2.0] Bring lawsuit against

180:[-9.0] Use unconventional violence, not specified below
181:[-9.0] Abduct, hijack, or take hostage
182:[-9.5] Physically assault, not specified below
1821:[-9.0] Sexually assault
1822:[-9.0] Torture
1823:[-10.0] Kill by physical assault
183:[-10.0] Conduct suicide, car, or other non-military bombing, not spec below
1831:[-10.0] Carry out suicide bombing
1832:[-10.0] Carry out car bombing
1833:[-10.0] Carry out roadside bombing
184:[-8.0] Use as human shield
185:[-8.0] Attempt to assassinate
186:[-10.0] Assassinate

ICEWS (6): The dataset & reliability

The first few observations in the 2013 dataset

```
> head(d[,c("Event.Date", "Source.Name", "Source.Country", "Event.Text", "CAMEO.Code", "Intensity", "Target.Name", "Target.Country")])
                         Source.Name Source.Country
                                                                                      Event.Text CAMEO.Code Intensity
                                                                                                                                  Target.Name Target.Country
1 2013-01-01 Citizen (United States) United States
                                                                     Use unconventional violence
                                                                                                                             Citizen (Yemen)
                                                                                                                                                       Yemen
2 2013-01-01
                Police (Afghanistan)
                                        Afghanistan
                                                                 Use conventional military force
                                                                                                                          Militant (Taliban)
                                                                                                                                                 Afghanistan
3 2013-01-01
                Police (Afghanistan)
                                        Afghanistan Arrest, detain, or charge with legal action
                                                                                                         173
                                                                                                                  -5.0 (itizen (Afghanistan)
                                                                                                                                                 Afahanistan
4 2013-01-01
                                                                        Make pessimistic comment
                                               China
                                                                                                                                       Syria
                                                                                                                                                       Syria
5 2013-01-01
                          Wen Jiabao
                                               China
                                                                                                          17
                                                                                                                   0.0
                                                                                                                                                       China
                                                                          Engage in symbolic act
                                                                                                                             Citizen (China)
6 2013-01-01
                          Wen Jiabao
                                               China
                                                                       Make an appeal or request
                                                                                                                          Government (China)
                                                                                                                                                       China
```

The last few where the source country is UK

```
> tail(d[d$Source.Country == "United Kingdom",c("Event.Date", "Source.Name", "Source.Country", "Event.Text", "CAMEO.Code", "Intensity", "Target.Name", "Target.Country")], 18)[3:8,]
                                       Source.Name Source.Country
                                                                                 Event.Text CAMEO.Code Intensity
                                                                                                                                                Target.Name
                                                                                                                                                                             Target.Country
733914 2013-12-31
                                    United Kingdom United Kingdom
                                                                                     Consult
                                                                                                     40
                                                                                                                                              United States
                                                                                                                                                                              United States
733937 2013-12-31 High Commission (United Kingdom) United Kingdom
                                                                                                     48
                                                                                                                                        Syed Ashroful Islam
                                                                                                                                                                                 Bangladesh
734191 2013-12-31
                           Scottish National Party United Kingdom Make an appeal or request
                                                                                                     28
                                                                                                                                   Citizen (United Kingdom)
                                                                                                                                                                             United Kingdom
734423 2013-12-31
                                           Reuters United Kingdom
                                                                       Discuss by telephone
                                                                                                     41
                                                                                                                                   City Mayor (South Sudan)
                                                                                                                                                                                South Sudan
734582 2013-12-31
                                               BBC United Kingdom Return, release person(s)
                                                                                                                7 Citizen (Palestinian Territory, Occupied) Occupied Palestinian Territory
734653 2013-12-31
                                        Nick Clegg United Kingdom
                                                                                                                                   Citizen (United Kingdom)
                                                                                                                                                                             United Kingdom
```

Seems more likely that Israel was releasing prisoners than BBC...

```
tail(d[d$Source.Country == "Isroel",c("Event.Date", "Source.Name", "Source.Country", "Event.Text", "CAMEO.Code", "Intensity", "Target.Name", "Target.Country")])
       Event.Date Source.Name Source.Country
                                                                                  Event.Text CAMEO.Code Intensity
                                                                                                                                                                              Target, Country
734728 2013-12-31
                        Isroel
                                       Israel Express intent to release persons or property
                                                                                                    353
                                                                                                              7.0 Citizen (Palestinian Territory, Occupied) Occupied Palestinian Territory
734739 2013-12-31
                        Israel
                                                                                                     40
                                       Israel
                                                                                                                                      Foreign Affairs (Iran)
                                                                                                     48
734751 2013-12-31
                        Isroel
                                       Israel
                                                                                                              1.0
                                                                                                                                      Foreign Affairs (Iran)
                                                                                                                                                                                        Iron
                                                                                                    311
734932 2013-12-31 Jew (Israel)
                                       Israel
                                                   Express intent to cooperate economically
734939 2013-12-31 Tzipi Livni
                                       Israel
                                                           Engage in diplomatic cooperation
                                                                                                     50
                                                                                                                            Palestinian Territory, Occupied Occupied Palestinian Territory
734959 2013-12-31
                                       Israel
                                                                   Return, release person(s)
                                                                                                              7.8 Citizen (Palestinian Territory, Occupied) Occupied Palestinian Territory
```



Israel releases 26 Palestinian prisoners

ICEWS (7): Is it trustworthy?

ICEWS data release includes Raytheon-BBN's* test of the precision** of ACCENT coder.

Judged correct if human coders agreed with both

- event code, or close (5/6/7, 11/12/16, 18/19)
- actors, or close

Result: encouraging! 3/4 of classifications deemed correct.

See Phil Schrodt's <u>critique</u> of the evaluation (and project in general): "Seven observations on the newly released ICEWS data"

*Producer/owner of ACCENT event coder.

**They claimed to be testing accuracy.

Event Code	BBN ACCENT Accuracy
01: Make Public Statement	71.1%
02: Appeal	71.4%
03: Express Intent To Cooperate	74.8%
04: Consult	80.6%
05: Diplomatic Cooperation	81.1%
06: Material Cooperation	65.9%
07: Provide Aid	73.9%
08: Yield	62.0%
09: Investigate	70.2%
10: Demand	58.7%
11: Disapprove	65.2%
12: Reject	74.6%
13: Threaten	66.0%
14: Protest	84.5%
15: Exhibit Force Posture	70.9%
16: Reduce Relations	69.9%
17: Coerce	88.1%
18/19: Assault/Fight	73.8%
20: Unconventional Mass Violence	83.6%
ALL (weighted by code frequency)	75.6%

ICEWS (8): Is it trustworthy enough?

Any classification produces errors. The question is how those errors relate to your research question.

Cases:

I. Your goal is to describe the total extent of armed conflict in the world over time.

Do errors in ICEWS lead to over-counting or under-counting of armed conflict? Does the degree of over-counting/under-counting vary over time?

2. Your goal is to assess whether signing a bilateral trade agreement improves relations between two countries.

Are errors in ICEWS (as manifested in your measure of bilateral relations) correlated with the treatment (signing PTA)? How does measurement error affect magnitude of estimated effects?

See larger literature on measurement error.

Example of research using ICEWS

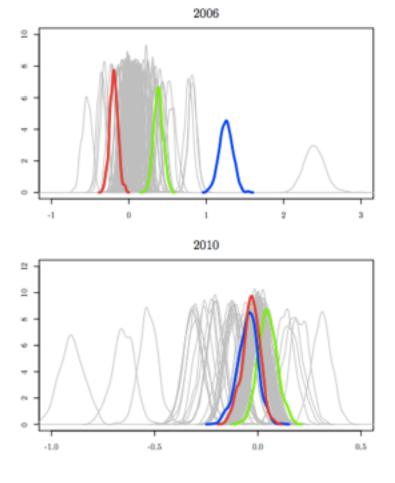
Simon Weschle, "The Impact of Economic Crises on Political Representation in Public Interactions: Evidence from the Eurozone" (working paper)

Research question: Does economic crisis cause political parties to "put politics aside" and cooperate?

Measurement problem: How do we measure extent of cooperation/conflict among political parties (and other societal actors)?

Measurement strategy:

- code all domestic interactions (party-party, partyother, other-other) as cooperative or conflictual (based on CAMEO categories)
- put log(#cooperative/#conflictual) for each pair in a country-year in an NxN symmetric matrix
- some statistics to derive a unidimensional scaling, such that actors close together are cooperative and far apart are conflictual



Results for Greece (parties in color, other societal actors in gray)

Final thoughts: description & measurement; learning and exploring

Two conflicting observations about the purpose of research:

- I. Description is a valuable part of what social scientists do.
- 2. Most political scientists are primarily interested in causal questions.

Two conflicting pieces of advice about the **practice of research**:

- I. When grappling with a measurement problem, ask yourself, "Suppose I had a perfect measure; what would I do with it?"
- 2. If you find a problem interesting, pursue it for a while even if you're not sure where it's headed.

Thanks and keep in touch!