Content Analysis

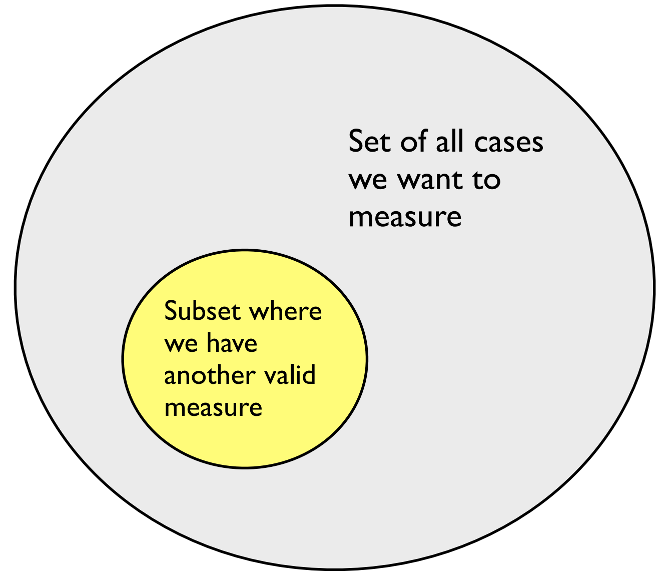
Lecture 2: Validation, clustering, and topic modeling

2 May, 2016

Prof. Andrew Eggers

How we validate, with two examples

Basically, we assess whether a measure works for the subset of cases where we know what it should produce, i.e. where we have another valid measure.



Two examples:

- Measuring implication in 2009 parliamentary expenses scandal with counts of Google News articles (Eggers 2014)
- Measuring political power with mentions in U.S. newspapers (Ban, Fouirnaies, Hall, Snyder 2015)

Example: Eggers (2014) on expenses scandal

Research question: How did local strength of party preference affect degree to which MPs were punished in expenses scandal?

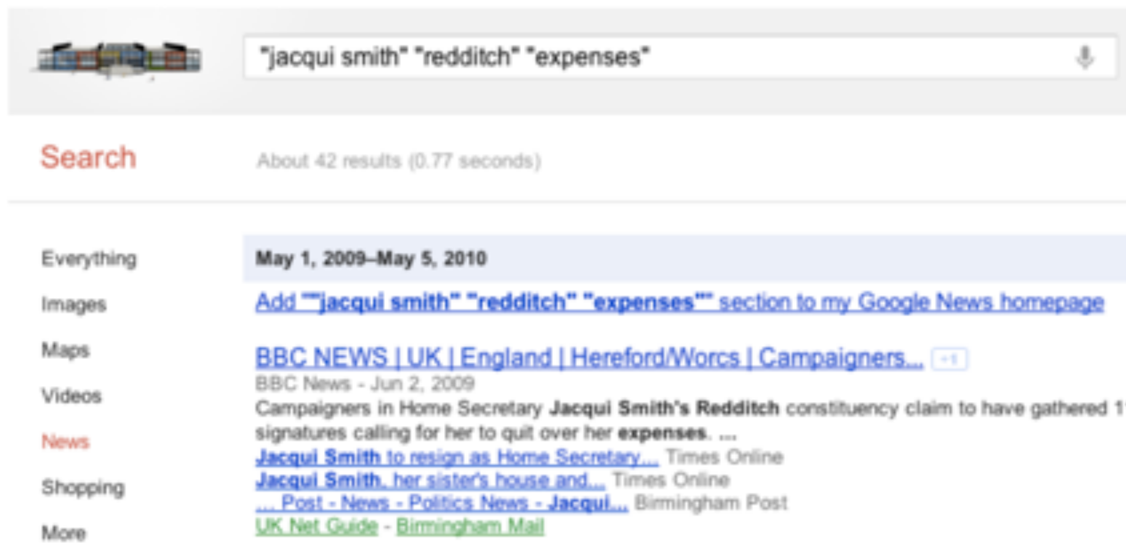
Measurement problem: How much was each MP implicated?

Possible measures:

- Amount of money MP spent
- Amount of money MP was asked to return
- BES survey of voters: "did your MP spend money improperly?"
- Appearance on a list of worst offenders e.g. in the *Telegraph* in May 2009

Step 1: count Google News hits for MP’s name and constituency between scandal and election

Step 2: count hits for for MP’s name and constituency **and the word “expenses”**



Step 3: divide to get implication score

$$\text{Implication}_i = \frac{\#\text{expenses stories}_i}{\#\text{stories}_i + n_0}$$

How to validate?

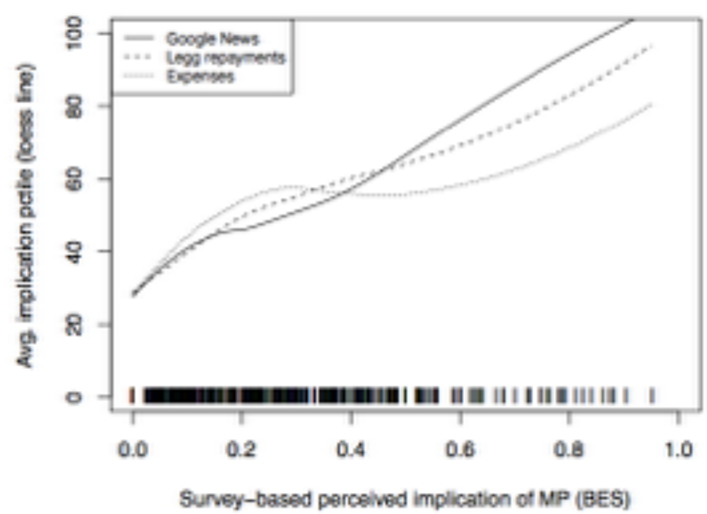
1. Compare with Telegraph's list of "saints" and "sinners"

2. Check list against substantive knowledge

Top 5

MP	Total stories	Expenses stories	Index
Margaret Moran	158	140	0.83
David Chaytor	109	93	0.78
Andrew MacKay	111	89	0.74
Julie Kirkbride	198	147	0.71
Peter Viggers	92	72	0.71

(3. Assess correlation with other possible measures)



Example: Ban, Fouirnaies, Hall, and Snyder (2015) on political power

Research question: Did U.S. Progressive-era reforms weaken state party machines?

Measurement problem: How powerful is the state party machine?

Possible measures:

- Historians' accounts
- Mayhew's measures, which only apply to 1966-1970

"THAT'S WHAT'S THE MATTER."
Boss Tweed. "As long as I count the Votes, what are you going to do about it? say?"

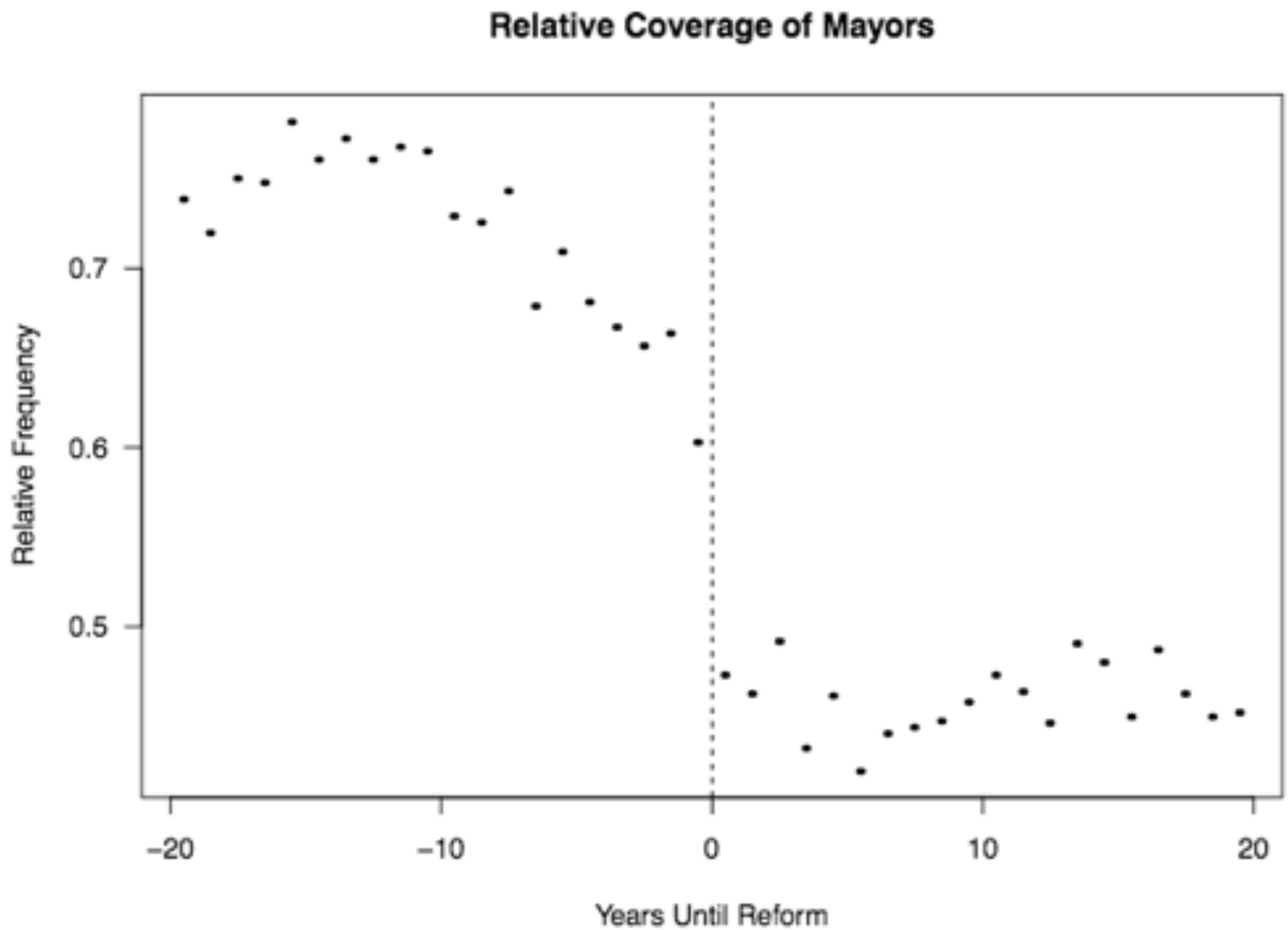
Ban et al (2015): Using newspaper mentions to measure power

Procedure:

- Gather huge newspaper database from online sources
  - 3,000+ newspapers
  - 1877-1977
  - 60+ million pages of text
- Count instances (by state and year) when the word "committee" follows within 5 words of "state", "county", "district", "local" etc **and** "Democratic", "Republican", or "GOP"

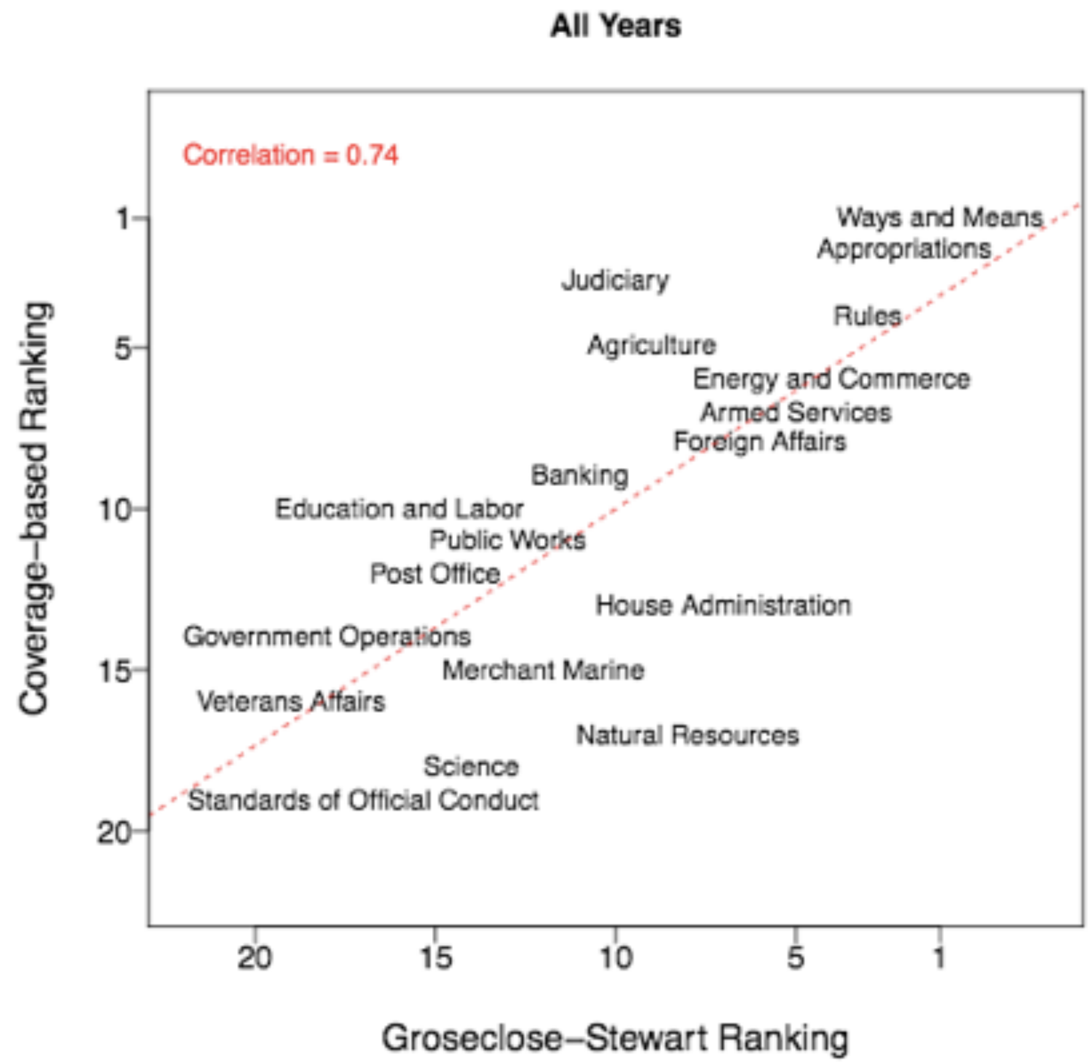
Ban et al (2015): validation: do "mentions" measure power?

1. Do mayor's mentions go down when city shifts power to a city manager?



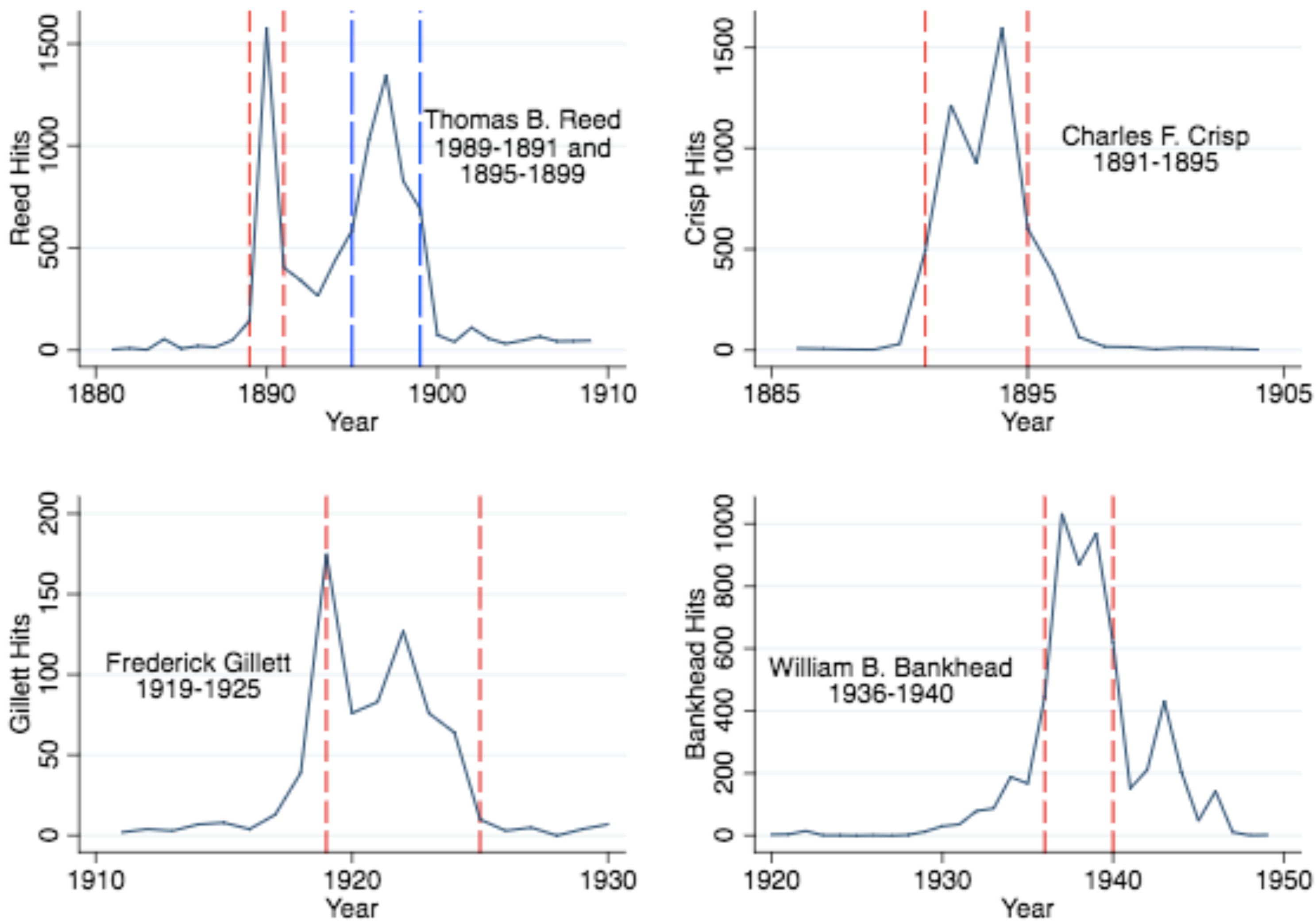
Ban et al (2015): validation: do “mentions” measure power?

2. Do congressional committees recognized as powerful get mentioned more?



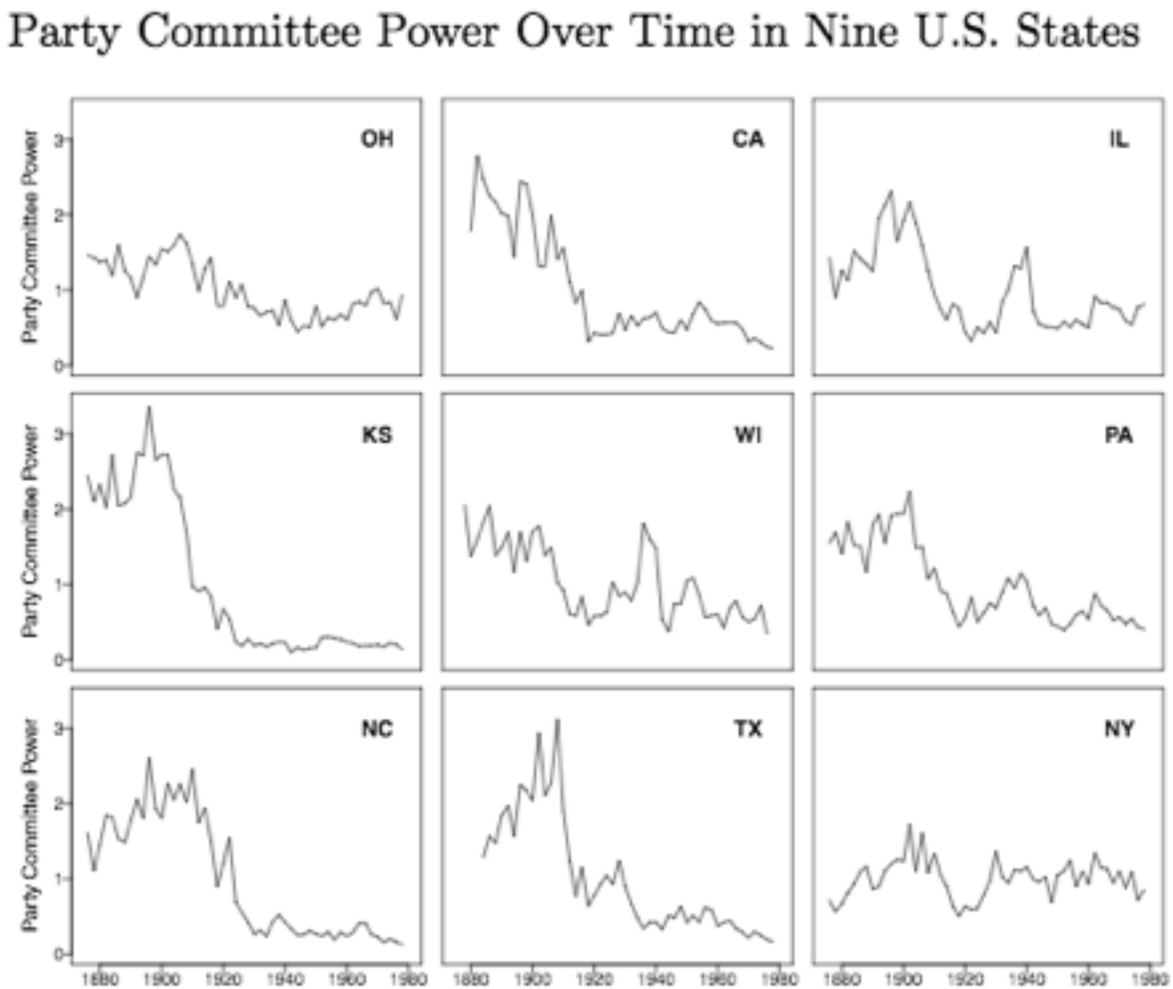
Ban et al (2015): validation: do "mentions" measure power?

3. Do members of Congress get mentioned more when they occupy leadership positions?



Ban et al (2015): validation: do "mentions" measure power?

4. How well does measure of party committee power correlate with Mayhew's TPO scores for 1966-1970? [corr > .5]



Topic labeling example: Speed ("Do newspapers now give the news?" 1893)

Characterizing content of New York newspapers (based on13 topics) on two Sundays 12 years apart.*

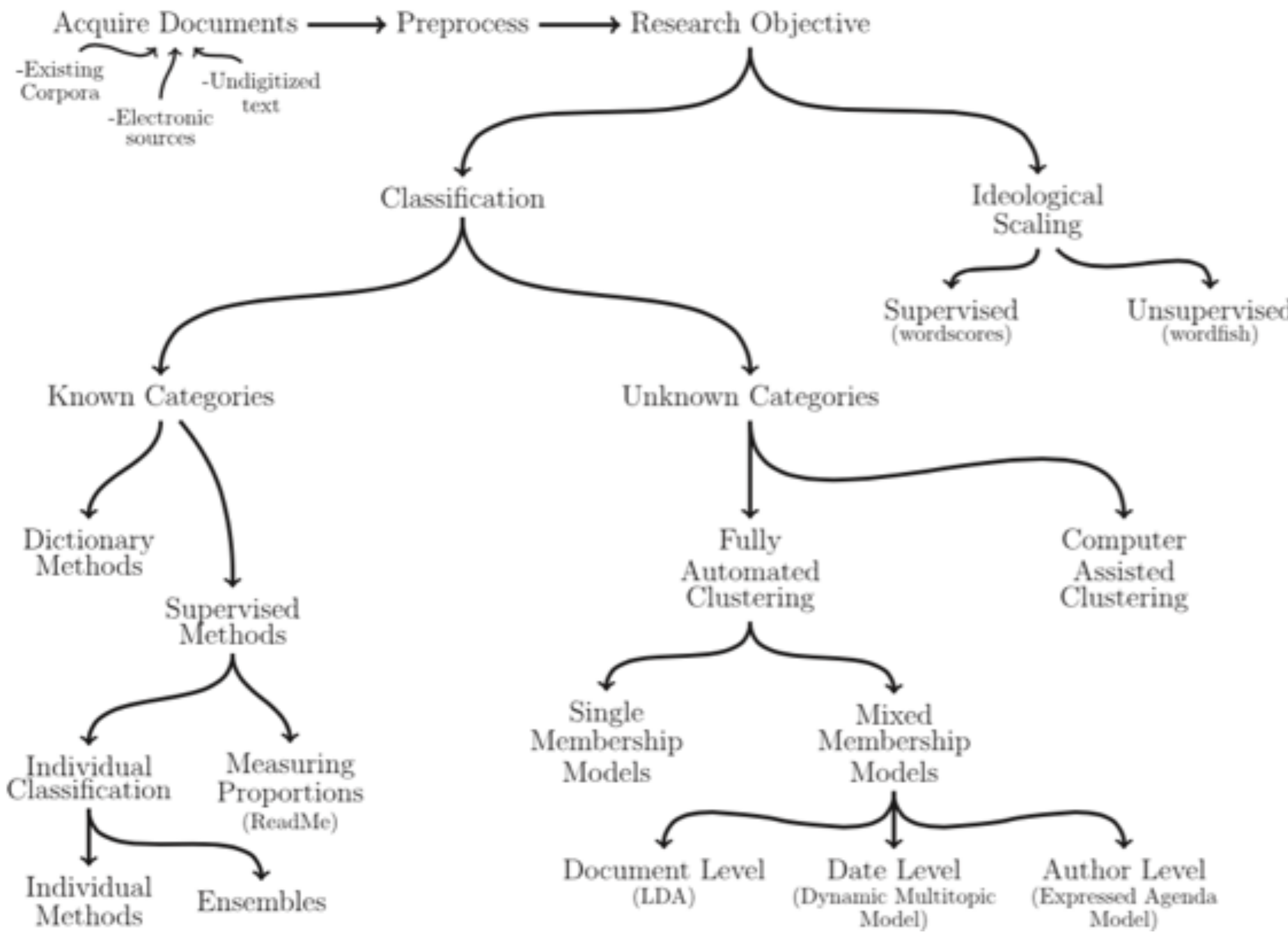
COLUMNS OF READING-MATTER IN NEW YORK NEWSPAPERS, APRIL 17, 1881, AND APRIL 16, 1893.

Subject.	Tribune, 1881.	Tribune, 1893.	World, 1881.	World, 1893.	Times, 1881.	Times, 1893.	Sun, 1881.	Sun, 1893.
Editorial	5.00	5.00	4.75	4.00	6.00	5.00	4.00	4.00
Religious	2.00	0.00	0.75	0.00	1.00	0.00	0.50	1.00
Scientific	1.00	0.75	0.00	2.00	1.00	0.00	0.00	2.50
Political	3.00	3.75	0.00	10.50	1.00	4.00	1.00	3.50
Literary	15.00	5.00	1.00	2.00	18.00	12.00	5.75	6.00
Gossip	1.00	23.00	1.00	63.50	.50	16.75	2.00	13.00
Scandals	0.00	1.50	0.00	1.50	1.00	2.50	0.00	2.00
Sporting	1.00	6.50	2.50	16.00	3.00	10.00	0.50	17.50
Fiction	0.00	7.00	1.50	6.50	1.00	1.50	0.00	11.50
Historical	2.50	2.50	2.75	4.00	2.50	1.50	4.25	14.00
Music and Drama	2.50	4.00	1.50	11.00	4.00	7.00	0.00	3.50
Crimes and Criminals	0.00	0.50	0.00	6.00	0.00	1.00	0.00	0.00
Art	1.00	1.00	3.00	3.00	2.00	0.00	0.25	1.25

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

*"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



Grimmer and Stewart (2013), "Text as Data"

Key categories of automated classification methods

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

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)

See Quinn et al (2010)

Like having a robot clean your basement

Supervised learning

You put a sample of items into piles.

Robot tries to organize the rest the same way.

Unsupervised learning

Tell robot how many piles you want.

Robot tries to put objects in piles with similar objects.

The term-document matrix (TDM)

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

Toy corpus

Document 1: "This is a document."

Document 2: "This is another document."

Document 3: "When is lunch?"

Term-document matrix

	D1	D2	D3
this	1	1	0
is	1	1	1
a	1	0	0
document	1	1	0
another	0	1	0
when	0	0	1
lunch	0	0	1

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
```

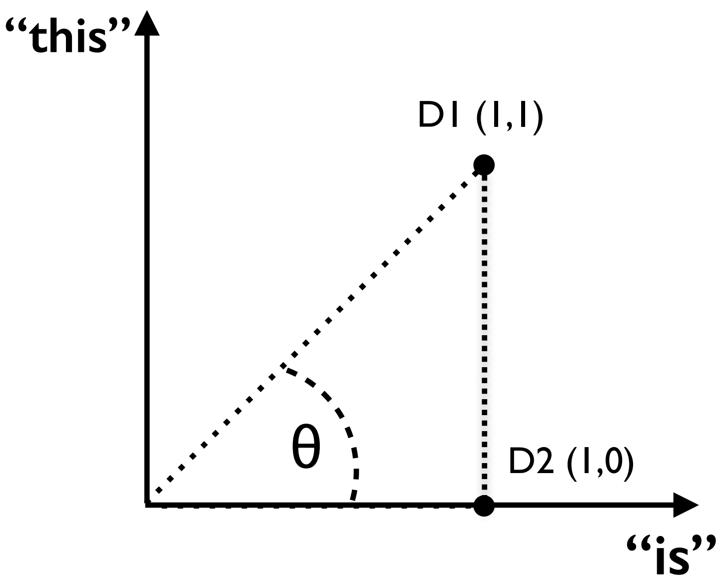
Unsupervised learning: clustering algorithms

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

	D1	D2	D3
this	1	1	0
is	1	1	1
a	1	0	0
document	1	1	0
another	0	1	0
when	0	0	1
lunch	0	0	1

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 to C1			Assign to C2		
	D1	D2	C1 avg	D3	C2 avg	Total distance^2 from cluster means
this	1	1	1	0	0	0
is	1	1	1	1	1	0
a	1	0	0.5	0	0	0.5
document	1	1	1	0	0	0
another	0	1	0.5	0	0	0.5
when	0	0	0	1	1	0
lunch	0	0	0	1	1	0
					Sum:	1

Algorithms do this for us (e.g. kmeans() in R).

Unsupervised learning: hierarchical clustering

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

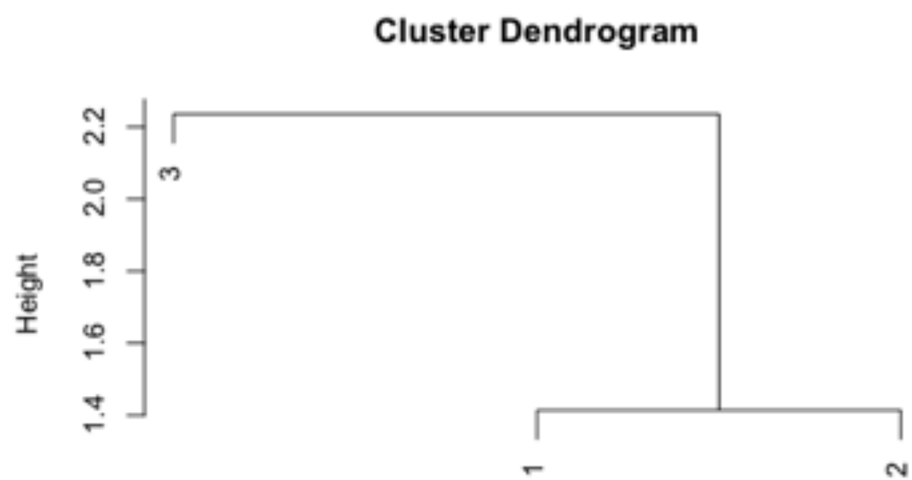
Intuition for hierarchical clustering:

- **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

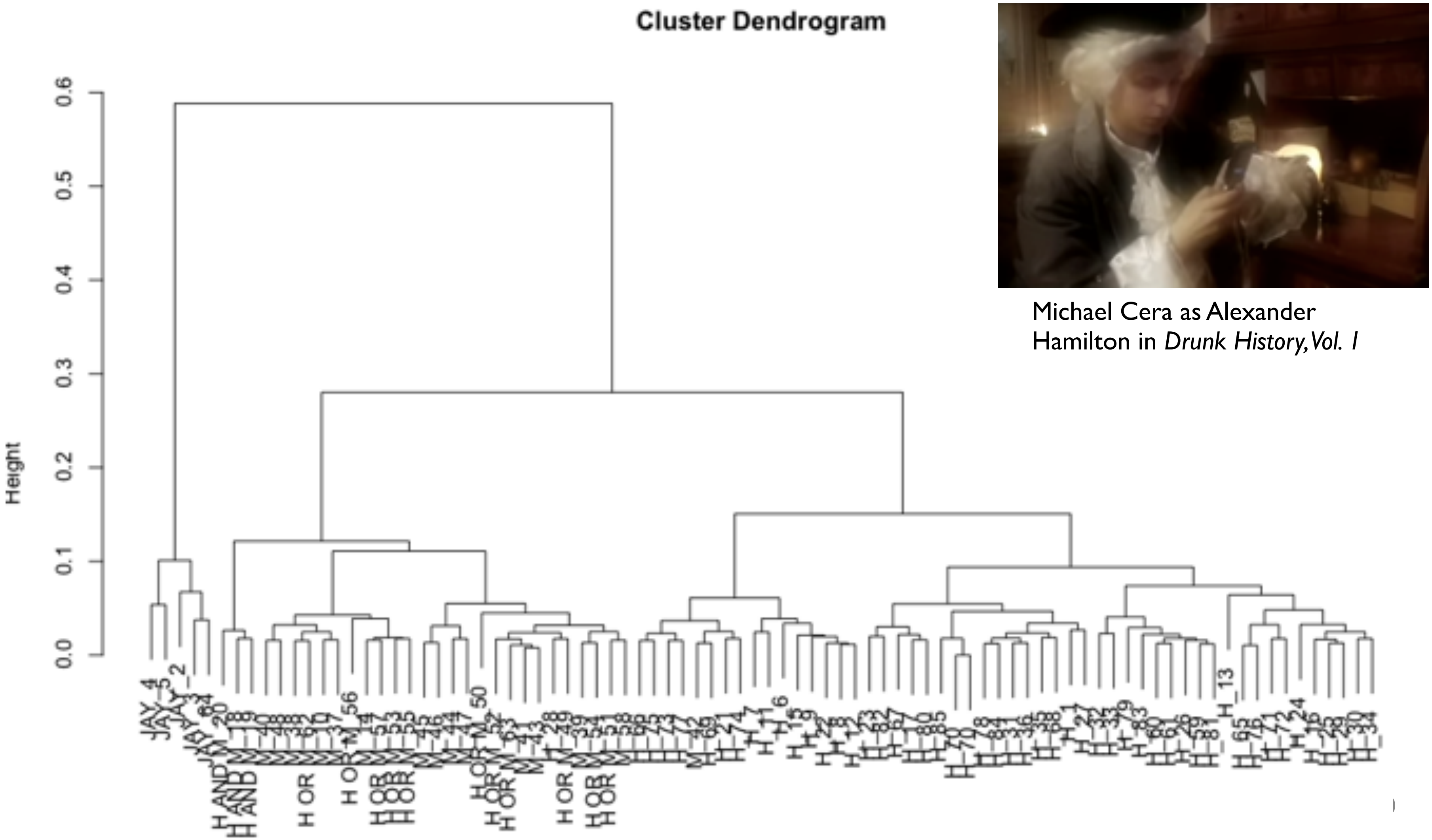
	D1	D2	D3
this	1	1	0
is	1	1	1
a	1	0	0
document	1	1	0
another	0	1	0
when	0	0	1
lunch	0	0	1

```
> 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?



Michael Cera as Alexander Hamilton in *Drunk History, Vol. 1*

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 intercept and slope coefficients
- In topic models, parameters are
  - the word frequencies for each topic
  - the topic membership of each document (e.g. LDA), date (e.g. Monroe et al), or author (e.g. Grimmer 2010)

Same in kmeans clustering!

Data structure and statistical theory of topic modeling

Data (Term Doc. Matrix)

W matrix: N word frequencies for D documents

	w_1	w_2	…	w_D
word 1	w_{11}	w_{21}	…	w_{D1}
word 2	w_{12}	w_{22}	…	w_{D2}
word 3	w_{13}	w_{23}	…	w_{D3}
…	…	…	…	…
word N	w_{1N}	w_{2N}	…	w_{DN}

Parameters to estimate

θ matrix: N word frequencies for K topics

	θ_1	θ_2	…	θ_K
word 1	θ_{11}	θ_{21}	…	θ_{K1}
word 2	θ_{12}	θ_{22}	…	θ_{K2}
word 3	θ_{13}	θ_{23}	…	θ_{K3}
…	…	…	…	…
word N	θ_{1N}	θ_{2N}	…	θ_{KN}

Z matrix: K topic labels for D documents

	z_1	z_2	…	z_D
topic 1	z_{11}	z_{21}	…	z_{D1}
topic 2	z_{12}	z_{22}	…	z_{D2}
topic 3	z_{13}	z_{23}	…	z_{D3}
…	…	…	…	…
topic K	z_{1K}	z_{2K}	…	z_{DK}

e.g. $z_2=[0,0,\ldots 1,0,\ldots,0]$

(If single-membership model)

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

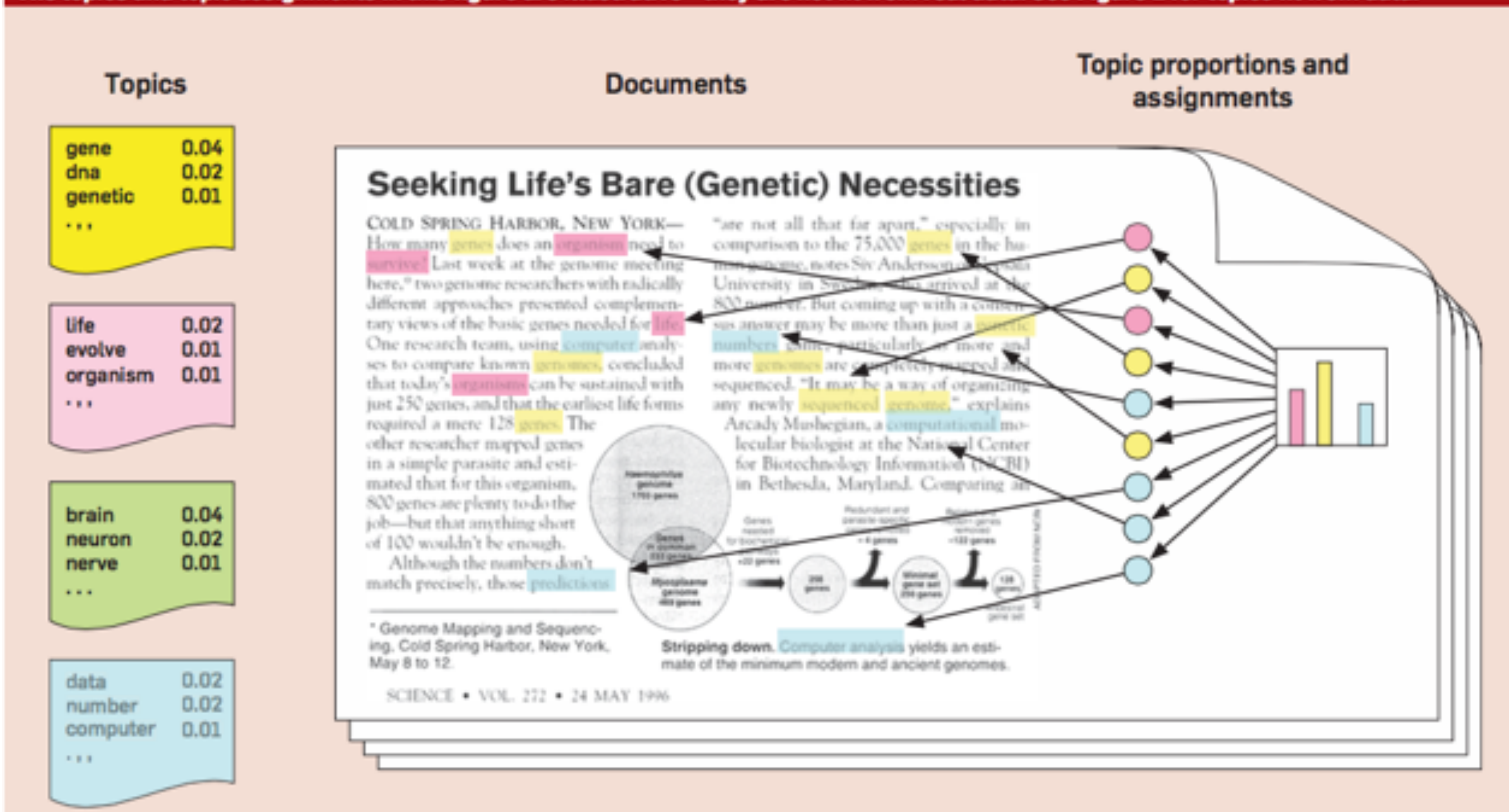
Bayesian version: describe $\Pr(\theta, Z|W) \propto \Pr(W|\theta, Z)\Pr(\theta,Z)$

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.



Blei (2012) "Probabilistic topic models"

Assumptions (1): How many topics?

- Literature has struggled with this.
- One reasonable statistical approach (Wallach et al 2009, "Evaluation Methods for Topic Models"): see if "held-out" documents are "likely" given model; compare results across models estimated with different *k*
- Semi-manual approach (Chang et al 2009, "Reading Tea Leaves: How Humans Interpret Topic Models"):
  - show human evaluators words that define estimated topics; see if they recognize an "intruder" word
  - show human evaluators a document and topics assigned to the document; see if they recognize an "intruder" document
- Recent progress? (Cheng et al (2015) "Model Selection for Topic Models via Spectral Decomposition")

Assumptions (2): Which features?

- typical approach: "bag of words".
- variation in:
  - unigrams? bigrams? trigrams?
  - "stop words"
  - stemming

See

- work of Hannah Wallach
- Denny & Spirling (2017) "Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It"

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: Grimmer (2010)

Goal: describe what Senators talk about in their press releases — their "expressed agenda"

Some of the topics:

Description	*Stems*
FEMA	disast,fema,storm,damag,declar,emerg,flood,recoveri,rebuild,recov
Food safety	food,fda,agricultur,contamin,recal,inspect,product,nutrit,drug,consum
Worker rights	worker,employe,wage,employ,labor,workplac,job,minimum,fair,workforc
AG/Justice	gener,justic,gonzal,judiciari,confirm,resign,investig,million,polit,nomine
Agriculture	farmer,agricultur,crop,produc,rancher,usda,livestock,nutrit,conserv,food
SCHIP	children,insur,uninsur,schip,kid,enrol,chip,reauthor,parent,incom
Public land	land,forest,manag,fish,wildlif,public,recreat,area,natur,speci
Pres. Veto/SOTU	presid,bush,veto,depart,iraq,announc,speech,union,democrat,facil
Loan crisis	mortgag,loan,lender,borrow,homeown,lend,bank,crisi,rate,market
Border security	border,homeland,immigr,patrol,secur,cross,agent,mexico,illeg,dh
Illegal immgr.	immigr,border,illeg,reform,legal,debat,enforc,broken,alien,citizenship
Honorary	honor,provid,rememb,friend,program,celebr,depart,prayer,tribut,legaci
Global warming	climat,warm,emiss,greenhous,global,carbon,chang,pollut,reduct,environment

Validation:

- committee leaders focus on topics that fit their committee's jurisdiction
- "expressed agendas" cluster geographically
- members who paid more attention to appropriations more likely to oppose earmark reform

Application of topic models: Grimmer (2010) (cont'd)

Finding: "Senators who represent the same state have more similar expressed agendas than senators from other states"

Evaluation 1: number of topics

The results presented in this paper use the variational algorithm derived in Appendix B and assume there are 43 topics present in the data. I varied the number of assumed topic from only five topics, up to 85 different topics. Assuming too few topics resulted in distinct issue being lumped together, whereas too many topics results in several clusters referring to the same issues. During my tests, 43 issues represented a decent middle ground. I corroborated

Evaluation 2: What about

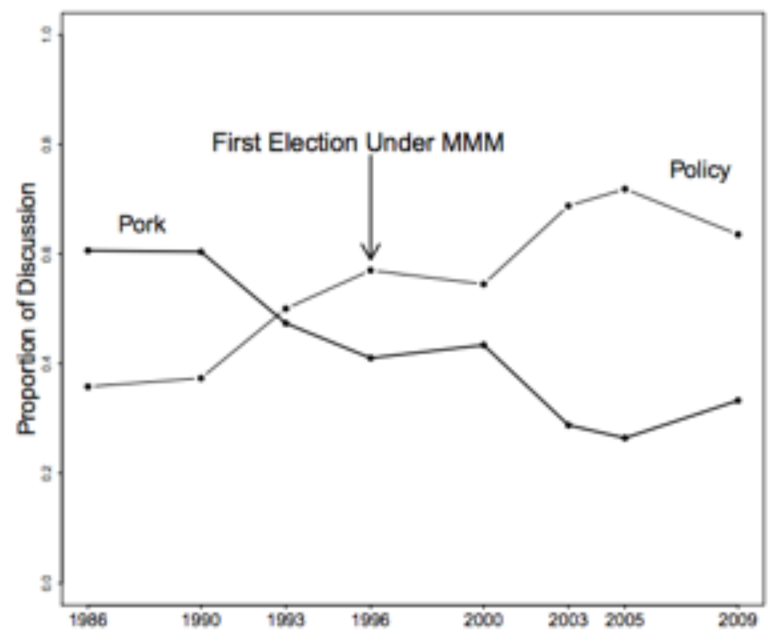
- dictionary methods?
- hand-coding a sample and using supervised learning?

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?
  - hand-coding a sample and using supervised learning?

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

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

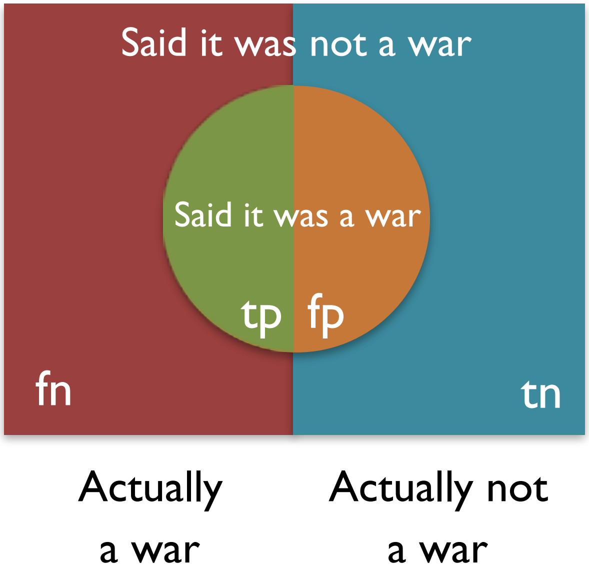
Two basic scenarios:

- 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



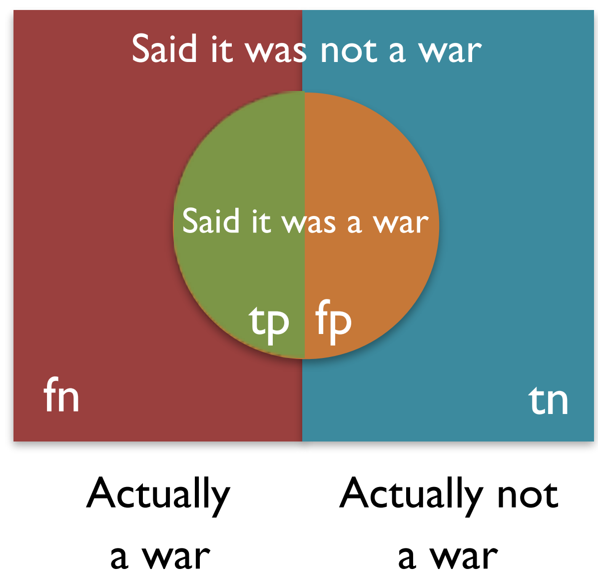
Evaluating a classification model: binary case (2)

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

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

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

Accuracy: $\frac{tp + tn}{tp + fp + fn + tn}$



Final thoughts: description & measurement; learning and exploring

Two conflicting observations about the **purpose of research**:

1. 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**:

1. 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.

Extra slides

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*

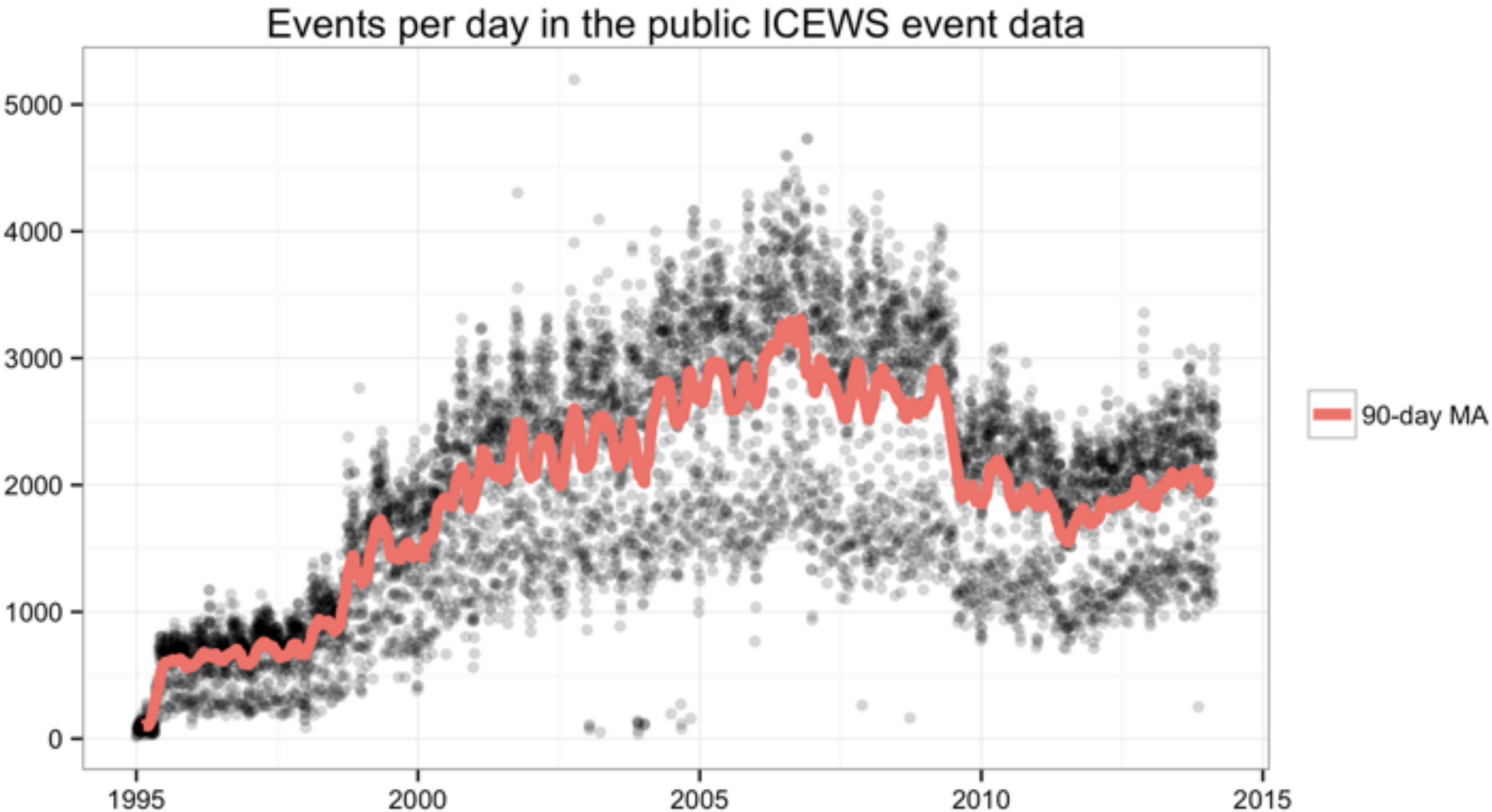
Classification of events from news stories: ICEWS (2)

- 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

Make Public Statement (01)
Appeal (02)
Express Intent to Cooperate (03)
Consult (04)
Engage in Diplomatic Cooperation (05)
Material Cooperation (06)
Provide Aid (07)
Yield (08)
Investigate (09)
Demand (10)
Disapprove (11)
Reject (12)
Threaten (13)
Protest (14)
Exhibit Military Posture (15)
Reduce Relations (16)
Coerce (17)
Assault (18)
Fight (19)
Engage in Unconventional Mass Violence (20)

Sub-categories

070:[7.0] Provide aid, not specified below
071:[7.4] Provide economic aid
072:[8.3] Provide military aid
073:[7.4] Provide humanitarian aid
074:[8.5] Provide military protection or peacekeeping
075:[7.0] Grant asylum
115:[-2.0] Bring lawsuit against

110:[-2.0] Disapprove, not specified below
111:[-2.0] Criticize or denounce
112:[-2.0] Accuse, not specified below
 1121:[-2.0] Accuse of crime, corruption
 1122:[-2.0] Accuse of human rights abuses
 1123:[-2.0] Accuse of aggression
 1124:[-2.0] Accuse of war crimes
 1125:[-2.0] Accuse of espionage, treason
113:[-2.0] Rally opposition against
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")])
  Event.Date            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        180      -9.0        Citizen (Yemen)          Yemen
2 2013-01-01    Police (Afghanistan)    Afghanistan       Use conventional military force        190     -10.0     Militant (Taliban)    Afghanistan
3 2013-01-01    Police (Afghanistan)    Afghanistan Arrest, detain, or charge with legal action        173      -5.0 Citizen (Afghanistan)    Afghanistan
4 2013-01-01                  China          China            Make pessimistic comment         12      -0.4                  Syria          Syria
5 2013-01-01             Wen Jiabao          China               Engage in symbolic act         17       0.0        Citizen (China)          China
6 2013-01-01             Wen Jiabao          China             Make an appeal or request         20       3.0     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")], 10)[3:8,]
       Event.Date                       Source.Name Source.Country                Event.Text CAMEO.Code Intensity                              Target.Name                  Target.Country
733914 2013-12-31                    United Kingdom United Kingdom                   Consult         40         1                            United States                   United States
733937 2013-12-31 High Commission (United Kingdom) United Kingdom                   Consult         40         1                        Syed Ashraful Islam                      Bangladesh
734191 2013-12-31           Scottish National Party United Kingdom Make an appeal or request         20         3                 Citizen (United Kingdom)                  United Kingdom
734423 2013-12-31                           Reuters United Kingdom     Discuss by telephone         41         1                  City Mayor (South Sudan)                     South Sudan
734582 2013-12-31                               BBC United Kingdom Return, release person(s)        841         7 Citizen (Palestinian Territory, Occupied) Occupied Palestinian Territory
734653 2013-12-31                        Nick Clegg United Kingdom                    Accuse        112        -2                 Citizen (United Kingdom)                  United Kingdom
```

Seems more likely that Israel was releasing prisoners than BBC…

```
> tail(d[d$Source.Country == "Israel",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.Name                  Target.Country
734728 2013-12-31       Israel         Israel Express intent to release persons or property        353       7.0 Citizen (Palestinian Territory, Occupied) Occupied Palestinian Territory
734739 2013-12-31       Israel         Israel                                   Consult         40       1.0                    Foreign Affairs (Iran)                            Iran
734751 2013-12-31       Israel         Israel                                   Consult         40       1.0                    Foreign Affairs (Iran)                            Iran
734932 2013-12-31 Jew (Israel)         Israel        Express intent to cooperate economically        311       5.2                                    Israel                          Israel
734939 2013-12-31  Tzipi Livni         Israel           Engage in diplomatic cooperation         50       3.5          Palestinian Territory, Occupied Occupied Palestinian Territory
734959 2013-12-31       Israel         Israel                Return, release person(s)        841       7.0 Citizen (Palestinian Territory, Occupied) Occupied Palestinian Territory
```

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 critique 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:

1. 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

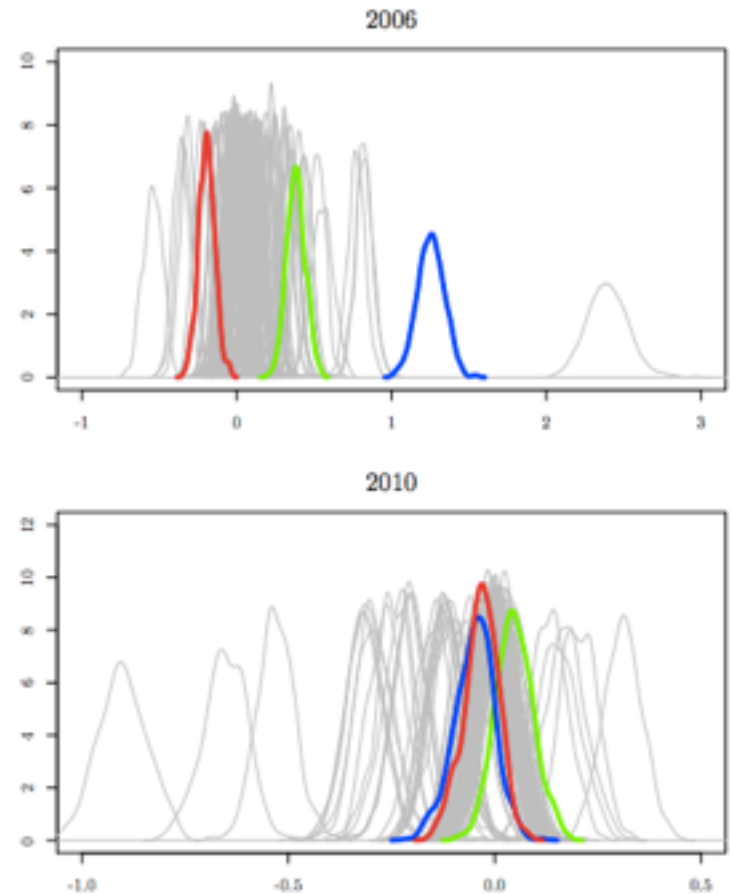
Simon Weschle, "The Impact of Economic Crises on Political Representation in Public Interactions: Evidence from the Eurozone" (working paper)

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, party-other, 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)