# Qualitative Methods and Multi-Method Research

Jason Seawright

j-seawright@northwestern.edu

Jan. 25, 2024

# Pure Qualitative Research Is Fine!

# Integrative Research and Qualitative Methods

- Quantitative researchers sometimes, but not always, identify assumptions in the way that quantitative researchers do

# Integrative Research and Qualitative Methods

- Quantitative researchers sometimes, but not always, identify assumptions in the way that quantitative researchers do
- Other times, they talk about boundary conditions, limitations, challenges, needs in their research, etc.

# Some Ideas

1. Enhance testing-based process tracing by reinforcing surprising steps in the process-tracing argument
2. Enhance discovery-based process tracing by broadening the range of subject matters explored
3. Machine learning to help position texts within large collections
4. Quantitative components to help move between levels of analysis

# Enhance Surprising Process-Tracing Steps

- The best process-tracing evidence strongly supports one hypothesis but is very surprising under other hypotheses.
- This motivates people who accept those hypotheses to reinterpret.
- Adding additional evidence with different epistemological properties makes such reinterpretation harder, and therefore speeds up necessary theoretical change

# Party-System Collapse

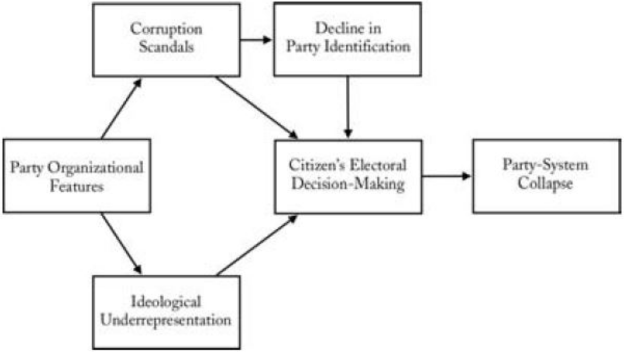



FIGURE 1.1. Causes of party-system collapse: overall structure of the argument

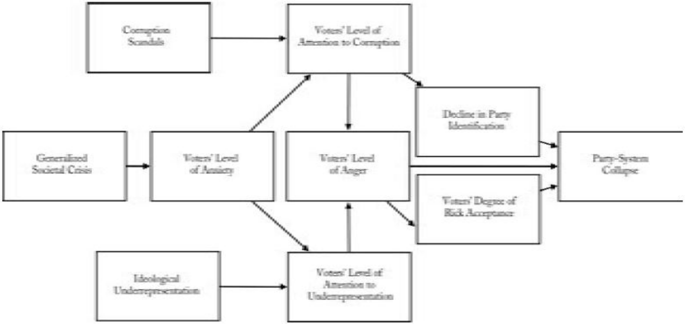



FIGURE 1.2. Voters and party-system collapse: refining the argument

TABLE 3.1. Economic performance of Latin American countries, 1980–2000 (percent)

| *Country* | *Average Annual per Capita Growth* | *Average Annual Inflation* | *Average Unemployment* |
|---|---|---|---|
| Peru | – 0.35 | 544.71 | 7.63 |
| Venezuela | – 1.08 | 32.72 | 10.08 |
| Argentina | 0.38 | 375.74 | 8.11 |
| Bolivia | – 0.31 | 663.83 | 5.31 |
| Brazil | 0.62 | 697.18 | 5.29 |
| Chile | 2.96 | 15.15 | 8.11 |
| Colombia | 1.22 | 23.21 | 11.64 |
| Costa Rica | 0.82 | 23.17 | 5.76 |
| Dominican Republic | 2.88 | 17.54 | 16.65 |
| Ecuador | – 0.55 | 1.09 | 8.69 |
| El Salvador | 0.50 | 5.18 | 9.01 |
| Guatemala | – 0.10 | 13.37 | 2.68 |
| Honduras | – 0.18 | 12.50 | 5.66 |
| Mexico | 0.86 | 42.47 | 3.34 |
| Nicaragua | – 2.09 | 1387.57 | 10.76 |
| Panama | 2.48 | 4.00 | 13.03 |
| Paraguay | 0.45 | 18.28 | 5.70 |
| Uruguay | 1.45 | 49.39 | 10.09 |

TABLE 3.2. Economic models of Latin American elections, 1980–2001

| *Variable Name* | *Economics Only* | *Economics + Unemployment* |
|---|---|---|
| Intercept | 0.596 (0.297) | 1.044 (0.349)* |
| Last opposition vote share | 0.339 (0.139)* | 0.442 (0.145)** |
| Candidate from main opposition | – 0.766 (0.217)** | – 0.859 (0.216)** |
| Peru, 1985 | – 1.384 (0.469)** | – 1.407 (0.462)** |
| Peru, 1990 | 1.262 (0.533)* | 1.704 (1.049) |
| Venezuela, 1993 | – 0.833 (0.460) | – 0.974 (0.457)* |
| Average inflation | – 0.00244 (0.00041)** | – 0.00232 (0.000508)** |
| Change in inflation | – 0.0000821 (0.0000339)* | – 0.0000835 (0.0000334)* |
| Current inflation | 0.000551 (0.000119)** | 0.000460 (0.000256) |
| Average growth | 0.0965 (0.0386)* | 0.119 (0.0409)** |
| Change in growth | 0.0413 (0.0164)* | 0.0390 (0.0165)* |
| Current growth | – 0.0760 (0.0286)* | – 0.0884 (0.0288)** |
| Per capita GDP | 0.00000201 (0.0000302) | 0.00000201 (0.0000308) |
| Unemployment | | – 0.0349 (0.0153)* |
| $R^2$ | 0.744 | 0.763 |
| Degrees of freedom | 48 | 44 |

NOTES:
OLS regression estimates with all vote shares transformed via the logit function.
*($p < 0.05$) **($p < 0.01$).

TABLE 4.2. 1993 Model of Venezuelan traditional party identification

| *Variable Name* | *Estimate (S.E.)* |
|---|---|
| Intercept | – 1.048** |
| | (0.289) |
| Past party identification | 0.790** |
| | (0.258) |
| Corruption concerns | – 0.905* |
| | (0.428) |
| Economic concerns | – 0.079 |
| | (0.128) |
| Preferred size of state | 0.074 |
| | (0.080) |

TABLE 5.4. 1998 model of Venezuelan presidential vote choice

| *Variable Name* | *Acción Democrática (AD) Estimate* | *PVZ Estimate* | *Movimiento V República (MVR) Estimate* |
|---|---|---|---|
| Intercept | – 4.01 (1.40)** | 0.39 (0.53) | – 1.47 (0.59) |
| Economic evaluations | – 0.18 (0.22) | – 0.08 (0.11) | 0.21 (0.10)* |
| Corruption perceptions | 0.65 (0.30)* | 0.08 (0.15) | 0.29 (0.14)* |
| Ideological leftist | – 0.93 (0.56) | – 0.23 (0.35) | 1.02 (0.17)** |
| AD party ID | 4.30 (0.34)** | 0.49 (0.31) | – 1.71 (0.56)** |
| Social class "C" | 0.29 (1.29) | 0.12 (0.41) | 0.98 (0.50)* |
| Social class "D" | 1.75 (1.19) | – 0.06 (0.39) | 0.74 (0.49) |
| Social class "E" | 1.60 (1.20) | – 0.39 (0.40) | 0.79 (0.49) |

NOTES:
Maximum-likelihood estimates of the parameters in a multinomial logit model of vote choice. Null deviance is 3831.08 on 4350 degrees of freedom; residual deviance is 3183.47 on 4320 degrees of freedom.
*($p < 0.05$) **($p < 0.01$).

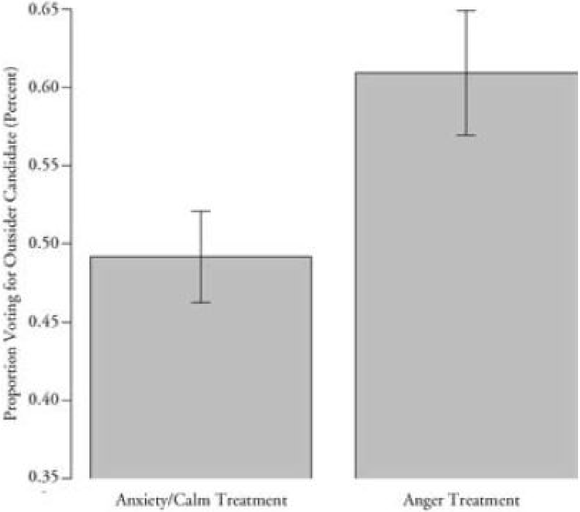



FIGURE 6.1. Affect and vote choice.
SOURCE: These data collected by the author in Lima and Cuzco, Peru, July–August 2009.

TABLE 7.2. Summary of parties' organizational traits

| *Dimension* | *AD* | *COPEI* | *APRA* | *AP* | *IU* | *PJ* | *UCR* |
|---|---|---|---|---|---|---|---|
| Membership size | Large | Large | Medium | Small | Small | Medium | Medium |
| Leadership size | Large | Large | Medium | Medium | Medium | Small | Medium |
| Leadership experience | High | High | Medium | Low | Low | Low | Medium |
| Leadership pragmatism | Medium | Medium | Medium | Medium | Medium | Medium | Medium |
| Membership financing | Medium | High | Medium | Low | High | Medium | Medium |
| Leadership power in local nominations | High | High | High | Medium | Medium | Low | Low |
| Power of local units in national decisions | High | High | Medium | Medium | High | Low | Medium |
| Communication flows | Low | Medium | High | Medium | Medium | Medium | Medium |
| Civil-society ties | High | Medium | Medium | Medium | High | Low | Low |
| Particularistic benefits in activist recruitment | High | High | Medium | Medium | Medium | Low | Low |
| Autonomy from unions | Low | Medium | Medium | High | Medium | High | High |

TABLE 7.3. Regression of local/national ideological distance on party organizational traits

| *Variable* | *Estimate (SE)* | P *Value* |
|---|---|---|
| Intercept | - 0.27 (0.42) | 0.53 |
| Local voters' ideological distance from the national party | 0.17 (0.05) | 0.00 |
| Membership size | 0.04 (0.03) | 0.16 |
| Leadership size | - 0.02 (0.03) | 0.60 |
| Leadership experience | 0.00 (0.04) | 0.96 |
| Leadership pragmatism | 0.05 (0.09) | 0.59 |
| Membership financing | 0.03 (0.03) | 0.38 |
| Leadership power in local nominations | - 0.11 (0.16) | 0.47 |
| Power of local units in national decisions | 0.13 (0.06) | 0.04 |
| Communication flows | - 0.03 (0.07) | 0.65 |
| Civil-society ties | 0.03 (0.04) | 0.49 |
| Particularistic benefits in activist recruitment | 0.06 (0.12) | 0.63 |
| Autonomy from unions | - 0.04 (0.06) | 0.50 |
| Intra-party ideological diversity | 0.51 (0.05) | $< 0.01$ |
| Complexity of membership and outreach organizations | - 0.06 (0.05) | 0.22 |
| Importance of patronage | 0.00 (0.04) | 0.94 |
| $R^2$ | 0.51 | |

# Lacombe (2018)

The NRA deploys ideas to construct a politicized group social identity among gun owners, who are then easy to mobilize into political participation.

# Lacombe (2018)

1. Archival research
2. Structural topic models
3. Qualitative and quantitative content analysis
4. In-depth reading
5. Time-series analysis
6. Process tracing

| TABLE 1. Top Words Associated with Each Topic from *Rifleman* Corpus | | |
|---|---|---|
| | **Topic Label** | **Words** |
| 1 | *Shooting Sports and Military Preparedness* | *FREX*: rifl, train, marksmanship, war, program, shooter, match, game, civilian, fire |
| | | *High Prob*: nation, rifl, associ, shoot, program, train, will, war, time, servic |
| 2 | *Membership Programs and Benefits* | *FREX*: nra, member, membership, futur, generat, perri, editori, hold, help, nras |
| | | *High Prob*: nra, member, year, can, one, take, now, will, million, come |
| 3 | *Gun Regulation* | *FREX*: citizen, registr, propos, possess, weapon, regist, purchas, honest, author, govern |
| | | *High Prob*: firearm, citizen, state, arm, gun, use, govern, person, nation, weapon |
| 4 | *Crime, Self-Defense, and Guns* | *FREX*: law, feder, control, crime, handgun, crimin, bill, owner, legisl, court |
| | | *High Prob*: gun, law, feder, legisl, control, polic, crimin, crime, bill, firearm |
| 5 | *Second Amendment* | *FREX*: citi, amend, vote, liberti, hous, presid, second, ban, magazin, declar |
| | | *High Prob*: right, american, will, power, amend, peopl, citi, polit, constitut, bear |
| 6 | *Americanism and Guns* | *FREX*: hunt, men, safeti, board, respons, hunter, educ, cours, recreat, accid |
| | | *High Prob*: america, will, men, hunt, american, safeti, peopl, hunter, respons, one |

**Note:** Words are stemmed.

| TABLE 2. Identity-Forming Language in Gun Control Editorials and Letters to the Editor | | | |
|---|---|---|---|
| | **Identity-Forming Language** | **In-Group Positive** | **Out-Group Negative** |
| *NRA Editorials* | 80% (338/422) | 55% (232/422) | 66% (280/422) |
| *Pro-Gun Letters* | 64% (1366/2135) | 43% (909/2135) | 38% (813/2135) |
| *Anti-Gun Letters* | 39% (401/1018) | 7% (71/1018) | 36% (362/1018) |

**Note:** The "Identity-Forming Language" column depicts the portion of editorials or letters that discuss either in-group positive or out-group negative characteristics, or both. The "In-Group Positive" and "Out-Group Negative" are more specific and depict the extent to which each type of identity-forming language is used.

The perceived opponents of gun rights consist of several distinct groups, the three most prominent of which are politicians, the media, and lawyers.

Politicians are described as: bureaucrat(ic), reformer(s), big city, urban, elitist, special interests, tyrannical, and "F" troop (politicians who have received "F" ratings from the NRA).

The media is described as: liar(s), coward(ly), elitist, phony, cynical, devious, shameless, and propaganda/propagandists.

Lawyers as: greedy, fat-cat, opportunist(s), big city, urban, elitist, phony, cynical, and liar(s).

A set of more general characteristics is used to portray gun regulation proponents as un-American, including: fanatic(s), extreme/extremists, radical(s), hysterical, anti-liberty, Communist(s), tyrannical, globalist, and internationalist. Finally, gun control supporters are described as "anti-gunners" and "the gun ban crowd."

**TABLE 3. Origination of Most Distinctive In-Group/Out-Group Phrases and Results of Lagged Linear Probability Model Predicting the Presence of Each Phrase in Pro-Gun Letters to the Editor**

| Phrase | First appearance in an identity-framing document | Significant Effect in a Lagged Year | *Rifleman* 1 Year Lag | *Rifleman* 2 Year Lag | *Rifleman* 3 Year Lag |
|---|---|---|---|---|---|
| | | | Coefficient (p-value) | Coefficient (p-value) | Coefficient (p-value) |
| *Anti-gunners* | *American Rifleman* (December 1975) | ✓ | – | – | 0.219 (0.020) |
| *Average citizens* | *American Rifleman* (February 1936) | × | – | – | – |
| *Freedom-loving* | *American Rifleman* (May 1944) | × | – | – | – |
| *Ordinary citizens* | *American Rifleman* (March 1948) | ✓ | – | 0.303 (0.083) | – |
| *Law-abiding* | *New York Times* (September 1931) | ✓ | 0.203 (0.050) | 0.249 (0.020) | – |

Note: Dependent variables are binary variables indicating whether a phrase appeared in a pro-gun letter to the editor in a given year for each year in the dataset (1930-2008). Separate models were estimated for each phrase. The independent variables presented in the table for each model are lagged binary variables indicating whether the phrase appeared in a *Rifleman* editorial in each of the three previous years. Also included in each model, as controls, was a binary variable indicating whether a phrase appeared in the *Rifleman* in the same year, as well as lagged binary variables indicating whether the phrase appeared in a pro-gun letter to the editor in each of the three previous years. All coefficients for variables included in the table that are significant at the $p<0.1$ level are included.

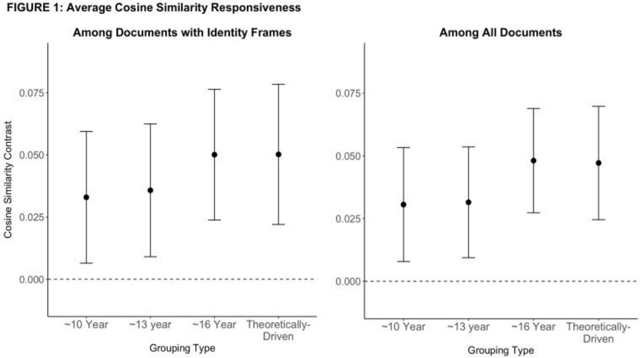


FIGURE 1: Average Cosine Similarity Responsiveness
Among Documents with Identity Frames
Among All Documents
Cosine Similarity Contrast
0.075
0.050
0.025
0.000
~10 Year
~13 year
~16 Year
Theoretically-Driven
Grouping Type

Case studies showing that dissemination and textual influence include broad arguments and not just word-use patterns.

**TABLE 5. Logistic Regression Predicting Calls to Action**

| | B (SE) | | z-value | p-value |
|---|---|---|---|---|
| (Intercept) | -3.007 | (0.551) | -5.454 | <0.001 |
| Threat | 1.975 | (0.324) | 6.104 | <0.001 |
| Identity-Building Language | 0.439 | (0.366) | 1.200 | 0.230 |
| Policy Discussion | 0.627 | (0.496) | 1.265 | 0.206 |

Null deviance: 551.58 on 421 degrees of freedom. Residual deviance: 475.88 on 418 degrees of freedom. AIC: 483.88.

Case studies of letters written to Presidents Johnson and Bush in response to NRA calls to action.

# Enhance Discovery-Based Process-Tracing

- It is much easier to add a new topic/theoretical theme to qualitative research before it begins than after data collection is complete.

# Enhance Discovery-Based Process-Tracing

1. Collect some kind of data set related to the question of interest, that is as inclusive as possible.
2. Use a machine-learning method to identify predictors related to the topics of interest.
3. Treat those predictors as clues about themes to explore in qualitative research.

# Example: Rise of the U.S. Alt-Right

- Alex Jones, InfoWars, and the Alt-Right
- It is likely that whatever is responsible for the growth of the alt-right has left traces in the discussions on InfoWars broadcasts.

# Lasso Regression

OLS regression minimizes: $\sum(Y - X\hat{\beta})^2$

# Lasso Regression

Lasso regression minimizes: $\sum(Y - X\hat{\beta})^2 + \lambda|\hat{\beta}|$

*American Political Science Review* (2018) 112, 2, 358–375
doi:10.1017/S0003055417000570 

# Reading Between the Lines: Prediction of Political Violence Using Newspaper Text

HANNES MUELLER *Institut d'Analisi Economica*
CHRISTOPHER RAUH *University of Montreal*

*This article provides a new methodology to predict armed conflict by using newspaper text. Through machine learning, vast quantities of newspaper text are reduced to interpretable topics. These topics are then used in panel regressions to predict the onset of conflict. We propose the use of the within-country variation of these topics to predict the timing of conflict. This allows us to avoid the tendency of predicting conflict only in countries where it occurred before. We show that the within-country variation of topics is a good predictor of conflict and becomes particularly useful when risk in previously peaceful countries arises. Two aspects seem to be responsible for these features. Topics provide depth because they consist of changing, long lists of terms that make them able to capture the changing context of conflict. At the same time, topics provide width because they are summaries of the full text, including stabilizing factors.*

## READING BETWEEN THE LINES

We now explore why topics provide such useful forecasting power on the time dimension.[23] To do this, we first let a simple machine learning algorithm choose variables to predict conflict within-sample. We use the least absolute shrinkage and selection operator (LASSO) with country fixed effects to choose variables from a pool of over 30 variables, including our 15 topic shares.[24] We base our analysis on the topic model estimated in 2013, as this is the last year of text we can use for estimation. The other variables in the pool are all previously used variables based on Chadefaux (2014), Ward et al. (2013), and Goldstone et al. (2010). This includes a host of standard political and economic variables, word counts based on our text, and two event counts from ICEWS. To this we add the incidence of armed conflict when explaining the onset of civil war a year later. To get a more and less restrictive set of variables, we vary the parameter that captures the weight given to choosing few variables. We pick three levels to show how the chosen model evolves with increasing selectivity.[25]

Table 3 shows the six models selected by the LASSO. Columns 1–3 show variables selected when predicting the onset of civil war and Columns 4–6 when predicting the onset of armed conflict. We report the share of the topic variables in these models in bold at the top of the table. The LASSO always chooses at least 50% of all variables from amongst the topics, and this share is higher in the more selective models.

**TABLE 3. Lasso Model**

| Selectivity Level | Mild | Regular | Very | Mild | Regular | Very |
|---|---|---|---|---|---|---|
| | | Civil war onset next year | | | Armed conflict onset next year | |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| *Topic shares* | | | | | | |
| conflict1 | 0.0366 | 0.0564 | | 0.306** | 0.259** | 0.275*** |
| | (0.0685) | (0.0599) | | (0.121) | (0.103) | (0.0999) |
| conflict2 | 0.256** | 0.300*** | 0.281*** | 0.304** | | |
| | (0.104) | (0.103) | (0.0961) | (0.117) | | |
| justice | −0.158** | −0.115* | −0.117** | −0.256*** | −0.215*** | −0.206*** |
| | (0.0664) | (0.0617) | (0.0541) | (0.0826) | (0.0712) | (0.0705) |
| international relations2 | −0.236** | | | −0.130 | −0.0554 | |
| | (0.102) | | | (0.0992) | (0.0909) | |
| civic life2 | −0.0869* | −0.00783 | −0.0247 | −0.0196 | −0.0679 | |
| | (0.0518) | (0.0370) | (0.0298) | (0.0671) | (0.0520) | |
| asia | −0.180** | −0.151** | −0.142** | | | |
| | (0.0803) | (0.0734) | (0.0650) | | | |
| sports | −0.0490 | | | | | |
| | (0.0365) | | | | | |
| politics | −0.141*** | | | | | |
| | (0.0472) | | | | | |
| business | −0.136** | | | | | |
| | (0.0549) | | | | | |
| economics | | | | −0.0256 | | |
| | | | | (0.0891) | | |
| *Other variables* | | | | | | |
| 25+ battle death | 0.0699*** | 0.0713*** | 0.0749*** | | | |
| | (0.0163) | (0.0164) | (0.0165) | | | |
| democracy score | 4.81e-05 | | | | | |
| | (0.000198) | | | | | |
| partial autocracy | | | | 0.0244 | 0.0270* | |
| | | | | (0.0151) | (0.0145) | |
| partial dem. with factionalism | | | | −0.00845 | −0.00163 | −0.00888 |
| | | | | (0.0124) | (0.0104) | (0.00981) |
| partial dem. w/o factionalism | 0.0154 | | | | | |
| | (0.0105) | | | | | |
| full democracy | 0.0174* | | | 0.00183 | 0.00442 | |
| | (0.0102) | | | (0.0165) | (0.0118) | |
| 4+ neighbouring conflicts | 0.0247 | | | | | |
| | (0.0396) | | | | | |
| child mortality rate | | | | −3.86e-05 | | |
| | | | | (0.000212) | | |
| ln (child mortality rate) | 0.00707 | | | 0.00376 | | |
| | (0.00531) | | | (0.00852) | | |
| % pop. discriminated | 0.111* | 0.108* | | | | |
| | (0.0604) | (0.0616) | | | | |
| % pop. excluded from power | | | | −0.0488 | | |
| | | | | (0.0442) | | |
| Country fixed effects | yes | yes | yes | yes | yes | yes |
| Observations | 4,561 | 4,644 | 4,931 | 3,991 | 4,226 | 4,226 |
| R-squared | 0.039 | 0.034 | 0.030 | 0.012 | 0.008 | 0.006 |
| Number of countries | 140 | 141 | 143 | 138 | 139 | 139 |
| **% topics in model** | **56%** | **71%** | **80%** | **50%** | **57%** | **67%** |

*Notes*: Robust standard errors in parentheses. *** $p<0.01$, ** $p<0.05$, * $p<0.1$. The table displays the selected variables using LASSO with parameter $\lambda$ equal to 100 (columns 1 and 4), 150 (columns 2 and 5), 200 (columns 3 and 6) from 15 topics and 30 variables from other models. Topics are from the year 2013. The most prominent words of each topic in 2013 are displayed in Table I.1. Summary statistics of all variables are displayed in Table C.1.

```r
library(tm)

infowarscorpus <- VCorpus(DirSource(directory="D:/Jan 6th/alexjonestranscripts/infowars", "UTF-8"))

ajmeta <- function(initialfilename){
  c(substr(str_split(initialfilename,"/")[[1]][5], 1,4),
    substr(str_split(initialfilename,"/")[[1]][5], 5,6),
    substr(str_split(initialfilename,"/")[[1]][5], 7,8),
    str_split(str_split(str_split(initialfilename,"/")[[1]][5],"_")[[1]][3],"\\.")[[1]][1]
    )
}

temp <- DirSource(directory="D:/Jan 6th/alexjonestranscripts/infowars", "UTF-8")

infometa.temp <- matrix(nrow=length(temp$filelist), ncol=4)

for (i in 1:length(temp$filelist)){
  infometa.temp[i,] <- ajmeta(temp$filelist[i])
}

meta(infowarscorpus, c("year", "month", "day", "show")) <- infometa.temp
```

| | | | |
|---|---|---|---|
| 20150104_Sun_Alex.mp3 | 2/2/2023 10:51 AM | Text Document | 99 KB |
| 20150105_Mon_Alex.mp3 | 2/2/2023 11:12 AM | Text Document | 170 KB |
| 20150106_Tue_Alex.mp3 | 2/2/2023 11:30 AM | Text Document | 157 KB |
| 20150107_Wed_Alex.mp3 | 2/2/2023 11:50 AM | Text Document | 162 KB |
| 20150108_Thu_Alex.mp3 | 2/2/2023 12:07 PM | Text Document | 152 KB |
| 20150109_Fri_Alex.mp3 | 2/2/2023 6:15 PM | Text Document | 155 KB |
| 20150111_Sun_Alex.mp3 | 2/2/2023 6:29 PM | Text Document | 98 KB |
| 20150112_Mon_Alex.mp3 | 2/2/2023 6:48 PM | Text Document | 162 KB |
| 20150113_Tue_Alex.mp3 | 2/2/2023 7:09 PM | Text Document | 159 KB |
| 20150114_Wed_Alex.mp3 | 2/2/2023 7:28 PM | Text Document | 155 KB |
| 20150115_Thu_Alex.mp3 | 2/2/2023 7:50 PM | Text Document | 163 KB |
| 20150116_Fri_Alex.mp3 | 2/2/2023 8:08 PM | Text Document | 164 KB |
| 20150118_Sun_Alex.mp3 | 2/2/2023 8:22 PM | Text Document | 101 KB |
| 20150119_Mon_Alex.mp3 | 2/2/2023 8:41 PM | Text Document | 150 KB |
| 20150120_Tue_Alex.mp3 | 2/2/2023 9:03 PM | Text Document | 167 KB |

```r
infowarsprocessed <- tm_map(infowarscorpus,content_transformer(tolower))

infowarsprocessed <- tm_map(infowarsprocessed,removeWords,stopwords("english"))

infowars.dtm <- DocumentTermMatrix(infowarsprocessed)
```

```r
library(gtrendsR)
library(dplyr)
library(tidyr)

extremism.trends1 <- gtrends(c("Alex Jones","Infowars","Oath Keepers",
                              "QAnon","Proud Boys"),time="all")#since2004
```

```
extremismwords <- as.data.frame(as.matrix(infowars.dtm), col.names=infowars.dtm$dimnames$Terms)

extremismwords$date <- ISOdate(as.numeric(infowarsprocessed$dmeta$year), as.numeric(infowarsprocessed$dmeta$month),
                              as.numeric(infowarsprocessed$dmeta$day), hour=0)

extremismwords <- left_join(extremismwords, extremism.trends.daily, by=c("date"))
```

```r
library(glmnet)

lambda <- 0.01

aj.lasso <- glmnet(extremismwords[1:292140], extremismwords$Alex.Jones, lambda=lambda,
                   family="gaussian",
                   intercept = T, alpha=1)

ajlassocoefs <- data.frame(varname = rownames(aj.lasso$beta), varcoef = as.numeric(aj.lasso$beta))

ajlassocoefs$varname[order(abs(ajlassocoefs$varcoef), decreasing=TRUE)[1:40]]

ajlassocoefs$varcoef[order(abs(ajlassocoefs$varcoef), decreasing=TRUE)[1:40]]
```

Table: Words and Coefficients Associated With Alex Jones and the Proud Boys

| Outcome | Word | Coefficient |
|---|---|---|
| Alex Jones | Pappert | 6.70 |
| Alex Jones | Sanctimoniously | 6.63 |
| Alex Jones | 87778925398777892539 | 4.34 |
| Alex Jones | Bolsonaro | 4.22 |
| Proud Boys | Zionist | 0.11 |
| Proud Boys | loudmouth | 0.08 |
| Proud Boys | collusion | 0.01 |

# Positioning Texts in a Collection

- Qualitative researchers often discuss texts drawn from large collections, as a way of characterizing both the individual text and the broader collection.

# Positioning Texts in a Collection

1. Run a simple text-as-data model on the text collection as a whole.
2. Summarize the topics selected from the model using qualitative readings of selected texts.
3. Use statistics related to the model to show how individual texts selected for close reading relate to the collection as a whole, and also to describe the overall population.

# Example: January 6th Legal Documents

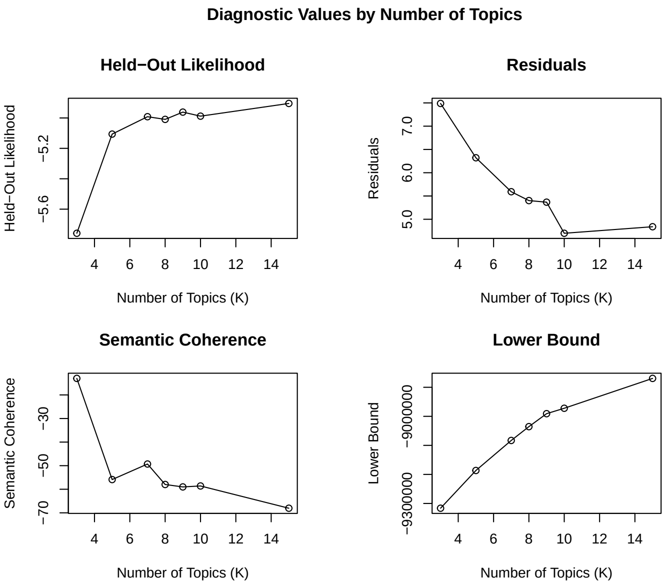
- The main DOJ Jan. 6th website lists about 2,700 documents related to cases against people involved with Jan. 6th.

```
processed_jan6th <- textProcessor(documents=db$text)

prep_jan6th <- prepDocuments(processed_jan6th$documents, processed_jan6th$vocab, processed_jan6th$meta)
```

```
testmodelsize<-searchK(prep_jan6th$documents, prep_jan6th$vocab, K = c(3,5,7,8,9,10,15),
                       verbose=FALSE)

plot(testmodelsize)
```


Diagnostic Values by Number of Topics
Held−Out Likelihood
Held−Out Likelihood
−5.2
−5.6
Residuals
Residuals
7.0
6.0
5.0
Semantic Coherence
Semantic Coherence
−30
−50
−70
Lower Bound
Lower Bound
−9000000
−9300000
4
6
8
10
12
14
Number of Topics (K)

```
jan6th8.stm <- stm(prep_jan6th$documents, prep_jan6th$vocab, 8)

labelTopics(jan6th8.stm)
```

```
> labelTopics(jan6th8.stm)
Topic 1 Top Words:
         Highest Prob: capitol, state, unit, build, defend, crowd, senat
         FREX: christensen, repair, expenditur, plea–, bartow, holli, linwood
         Lift: -cr--svh, check–congression, dennison, holli, repair, -cr---rbw, -law
         Score: crowd, presid, senat, chamber, christensen, member, short
Topic 2 Top Words:
         Highest Prob: case, page, document, file, -cr--apm, -cr--bah, -cr--tjk
         FREX: -cr--bah, ÿÿÿ, ÿÿÿÿÿ, -cr--abj, -cr--apm, -cr--rcl, rhr
         Lift: aiii, cda, ced, crl, deh, dlb, dtu
         Score: -cr--apm, file, document, page, -cr--bah, case, -cr--tnm
Topic 3 Top Words:
         Highest Prob: wkh, dqg, dslwro, zlwk, wkdw, cramer, dqxdu
         FREX: dqg, dslwro, wkdw, dqxdu, waynick, munn, portlock
         Lift: “badass, “brian”, “captain, “free”, “gid, “group, “healion
         Score: dqg, wkh, dslwro, wkdw, zdv, iurp, rwkhu
Topic 4 Top Words:
         Highest Prob: state, build, unit, conduct, capitol, restrict, usc
         FREX: rev, yazdani-isfehani, xxxxxx, loammi, xxxxxxxxx, padilla, complaint
         Lift: -cr--kbj, -mj---, -mj--jeg, -offic, “bugziethedon”, “chad, “elijah
         Score: disrupt, wkh, build, cordoned-, roxpeld, disord, vice
Topic 5 Top Words:
         Highest Prob: capitol, build, state, video, januari, unit, senat
         FREX: witness-, bolo, afo-, fbi’, anonym, monm, affiant
         Lift: --job, ‘’m, ‘fuck, “absolutely”, “ami, “anthoni, “benjamin
         Score: affiant, presid, figur, fbi, crowd, chamber, individu
Topic 6 Top Words:
         Highest Prob: januari, capitol, state, offic, messag, member, facebook
         FREX: denney, watkin, rhode, ashlock, crowl, niemela, -mj--rmm-zmf
         Lift: -destruct, -obstruct, –obstruct, “’d, “assault, “certification”, “destruct
         Score: watkin, messag, megg, denney, facebook, oath, parker
Topic 7 Top Words:
         Highest Prob: page, case, file, document, imag, januari, capitol
         FREX: brodi, rockholt, mlynarek, cantrel, preller, balhorn, lovley
         Lift: “kyle, balhorn, bonenberg, brown-color, carhart, cronin, degregori
         Score: brodi, rockholt, preller, cantrel, mlynarek, lovley, balhorn
Topic 8 Top Words:
         Highest Prob: client, agreement, sentenc, court, agre, right, will
         FREX: agreement, client’, impos, guidelin, waiv, appeal, withdraw
```

```
db$filename_original[as.integer(names(prep_jan6th$documents[findThoughts(
        jan6th8.stm)$index$'Topic 1'[1]]))]
```

- Topics 3 and 7 are various kinds of judicial procedural documents, shared across various kinds of cases.
- Topic 2 is mostly plea bargains, while topic 5 is mostly charge sheets.

- For Topic 1, the most representative text is a stipulation of facts in the case of Jonathan Davis Laurens.

The Atlanta Journal-Constitution

ion | Georgia Politics | AJC Peachtree Road Race | Local News | Sports | Investigations | Opinion

Stay informed on topics that matter to you. **Get 3 months for 99¢!**

# Selfie-taking Duluth man pleads to Jan. 6 charge

Jonathan Davis Laurens laughed it off when friends on Facebook suggested he should not be inside the U.S. Capitol as a pro-Trump mob swamped the building on Jan. 6, 2021.

“We got into the capitol, walked around, chanted some slogans and stuff,” he wrote. “We weren’t there to tear (expletive) up, just disrupt the system. All in all, I had fun! LOL.”

Using security camera footage and Laurens' own social media postings, investigators were able to trace his route through the Capitol from the Senate side, through the Rotunda and Statuary Hall, and over to the House where he stood as a mob of rioters attempted to force their way into the House chamber. Then Laurens entered the Rayburn Reception Room near the House chamber where he snapped a selfie underneath a painting of George Washington.

- The most-representative document for topic 4 is an FBI affidavit about the activities of Brandon Nelson and Abram Markofski.

CRIME

# Two more Wisconsin men — one of them a National Guard member — charged with entering U.S. Capitol during Jan. 6 riot

Elliot Hughes
Milwaukee Journal Sentinel

Published 1:15 p.m. CT May 3, 2021 | Updated 4:20 p.m. CT May 3, 2021

After attending Trump's rally and arriving at the Capitol, the two men said police made no attempt to remove them from the building. Nelson said police were guiding people inside. Markofski told agents an officer said to him, "I can't make you guys leave. However, for your safety, you should leave."



Surveillance cameras captured images of both men. They said they spent about 40 minutes inside the building and drove back to Wisconsin after leaving, according to the complaint.

- Topic 6 is most typically represented by a statement of offense for Daniel Ray Caldwell.

NEWS › CRIME

# North Texas man gets over 5 years in for assaulting Capitol officers during riot

Daniel Ray Caldwell doused a group of officers with bear spray during the Jan. 6 uprising at the U.S. Capitol, a prosecutor said.

Daniel Caldwell was caught on video spraying a chemical irritant at police officers defending the Capitol Building on Jan. 6, 2021. (Justice Department)

“During the riot, Caldwell taunted police officers by asking them to spray [pepper spray] and asking if they were ‘scared,’” Peterson said in a sentencing memo.

Later that day, Caldwell gave an interview to an undisclosed person in which he admitted to shouting at officers who were spraying rioters with chemical irritant, saying "Dude, do it again and we'll spray you back," Peterson said in the court filing.

They did and Caldwell sprayed the officers back. "I got like 15 of them," he reportedly said in the interview.

Jenkins said his client was not involved in any organized plan to attack the Capitol. He asked for a four-year prison term.

- With respect to topic 8, the most representative document is a statement of offense for Matthew Capsel.

# Illinois man gets 18 months in prison for fighting with National Guard during Jan. 6 U.S. Capitol riot

Matthew Capsel was arrested just weeks after the Capitol riot for fighting with members of the National Guard. He later wrote on social media that "on the 6 good men had to do a bad thing."

By Jon Seidel | Dec 16, 2022, 5:03pm CDT

SHARE


Federal authorities say this image depicts Matthew Capsel. | Federal court records

Capsel recorded TikTok videos while outside the Capitol on Jan. 6, 2021. In the videos, he said, "they only got so much mace and we got all these patriots we aren't going to run out, they are going to run out. Hold the line, don't run, go down with your eyes out and get some water to drink and hold your ground."

He also joined a mob that ascended the stairs toward the inauguration bleachers on the building's northwest side. He and others overcame officers there and stood on the bleachers.

Later, after a 6 p.m. curfew took effect, Capsel and a mob confronted a line of National Guardsmen. Capsel was at the front of the group that charged the troops, pushing against their riot shields. Capsel retreated only after the troops defended themselves with pepper spray.


TikTok
TikTok
TikTok
TikTok

Between the riot and his arrest in late January 2021, Capsel kept posting videos to TikTok and Facebook. In one, he made a slideshow featuring pictures of President Joe Biden, U.S. House Speaker Nancy Pelosi and Senate Majority Leader Chuck Schumer.

Music also played in the video, with lyrics that said, "Wish you would die, just f---ing die. Ever had a boss that was so damn bad, when you get to work it would drive you f---ing mad. If this place burned down it would be kind of sad. But to be away from him I would f---ing be kind of glad. To do all the work, he takes all the cred, about to blow up on this b------."

- With this framework, we can answer questions about relative prevalence of different degrees of violence.

Table: Texts' Average Degree of Topic Membership

| Topic | Average Membership |
|---|---|
| 4 (Least Violence) | 28% |
| 1 (Near Violence) | 10% |
| 6 (Dramatic Violence) | 3% |
| 8 (Violence, Media) | 9% |

- We can also situate selected texts within the broader context. To pick an actually randomly selected example consider the statement of facts connected with Joseph Howe. The model tells us that Howe belongs in topic 8 (96% membership).

On Thursday, January 7, 2021, a tipster, who will be referred to as T-1, called the FBI's National Threat Operations Center (NTOC) to report information about individuals who participated in the riot at the U.S. Capitol. T-1 identified Michael Sparks (charged elsewhere, *see* D.D.C. Case No. 21-CR-87-TJK) as being the first individual to climb through a broken window into the U.S. Capitol building on January 6, 2021—that is, the very first rioter to breach the U.S. Capitol building. T-1 also knew that Sparks traveled to Washington, DC with co-workers from the Elizabethtown, KY area, including JOSEPH HOWE. (The agent who documented the tip and later interview incorrectly spelled the name "Howell," but he will be referred to throughout using his true name HOWE).

The next day, an FBI agent interviewed T-1 by telephone. T-1 explained that T-1 is an acquaintance of HOWE and overheard HOWE discuss plans with Sparks to travel to Washington, D.C. on January 6, 2021 to attend the pro-Trump rally. T-1 heard Sparks tell HOWE, "This time we are going to shut it down." After the incidents at the U.S. Capitol on January 6, 2021, HOWE's wife, who is also a co-worker of HOWE and Sparks, had a video of her husband being pepper sprayed at the U.S. Capitol.

On January 12, 2021, an FBI agent interviewed T-1 again. The agent provided T-1 with photographs of individuals inside the U.S. Capitol on January 6, 2021 to review, including a photograph of an unknown individual the agent believed at the time might be HOWE (though it was later determined that this photograph showed a different individual wearing similar clothing, including goggles and a dark beanie hat, but was not in fact HOWE). T-1 indicated that the man in the photograph "looks like Joe Howe." T-1 knew that HOWE sent photographs and videos of himself inside the U.S. Capitol to his wife, including one after being pepper sprayed.

On March 22, 2021, an FBI agent interviewed a witness, referred to as W-1. W-1 told the FBI that he traveled to Washington, DC with Sparks, HOWE, and others who worked at the same company in Elizabethtown, KY, in a vehicle rented by Sparks. Upon arriving near the U.S. Capitol on January 6, 2021, W-1 was separated from HOWE based on the large size of the crowd.

CO: You think we're getting in that building?

HOWE: We're getting in it.

CO: You think so?

HOWE: Oh yeah.

Unknown Male ("UM"): Tell you what, you go. We'll follow you.

HOWE: We're getting in it.

UM: [Unintelligble]

HOWE: We're getting in it.

At that point, another individual, who based on my review of the video and my knowledge of Sparks' voice I believe was Sparks, chimed in:

Sparks: All it's gonna take is one person go. The rest is following.

HOWE: Let it go south in there.

# Moving Between Levels of Analysis

- In any kind of research, it can be tough to move between levels of analysis.

# Moving Between Levels of Analysis

- In any kind of research, it can be tough to move between levels of analysis.
- Embedding a quantitative design component can sometimes save time, effort, and money in covering another level of analysis.

# Two More Challenging Designs

1. Combining Game Theory and Qualitative Research
2. Statistically Testing Generalizability of Qualitative Research