# TAMIDS 2021 - Data Science Competition

Landon Buechner— Dr. Bhattacharya

Team: Cluster One

Texas A&M University

April 14, 2021

# Outline

- Data Aggregation
- Exploratory Data Analysis
- Preprocessing
- Bayesian Hierarchical Modelling
- Campaign Strategy
- Remarks / References

# github.com/magi-1/presidential-elections

magi-1 Update README.md 92df9c8 8 days ago 18 commits

| | | |
|---|---|---|
| Code | Add files via upload | 11 days ago |
| Data | submission | 13 days ago |
| Images | submission | 13 days ago |
| Notebooks | copy of colab notebook | 11 days ago |
| References | submission | 13 days ago |
| README.md | Update README.md | 8 days ago |

README.md

# TAMIDS 2021 Competition

Team: Cluster One

Click to See the Final Report

## Comments

- (4/1/2021) Note that report linked above is the culmination of all of the notebooks in this repository. These notebooks were primarly used for data cleansing and EDA so don't expect them to be the most readable. Additionally I originally worked within the Google Drive/Colab ecosystem so these files are just copies.

# Objectives

**Learning**

Get hands on experience with Bayesian modelling.

**Political Consulting**

*Within each state, how much money should be allocated to each expense category in order to maximize the odds of winning the election?*

# Data Aggregation

All candidates

Candidate master

CANDIDATE MASTER

2021-2022 | 2019-2020 | 2017-2018 | 2015-2016 | 2013-2014 | 2011-2012 | 2009-2010 | 2007-2008 | 2005-2006 | 2003-2004 | 2001-2002 | 1999-2000 | 1997-1998 | 1995-1996 | 1993-1994 | 1991-1992 | 1989-1990 | 1987-1988 | 1985-1986 | 1983-1984 | 1981-1982 | 1979-1980

Data description for this file ▶

Header file

The candidate master file contains one record for each candidate who has either registered with the Federal Election Commission or appeared on a ballot list prepared by a state elections office.

Candidate-committee linkages

House/Senate current campaigns

Committee master

PAC summary

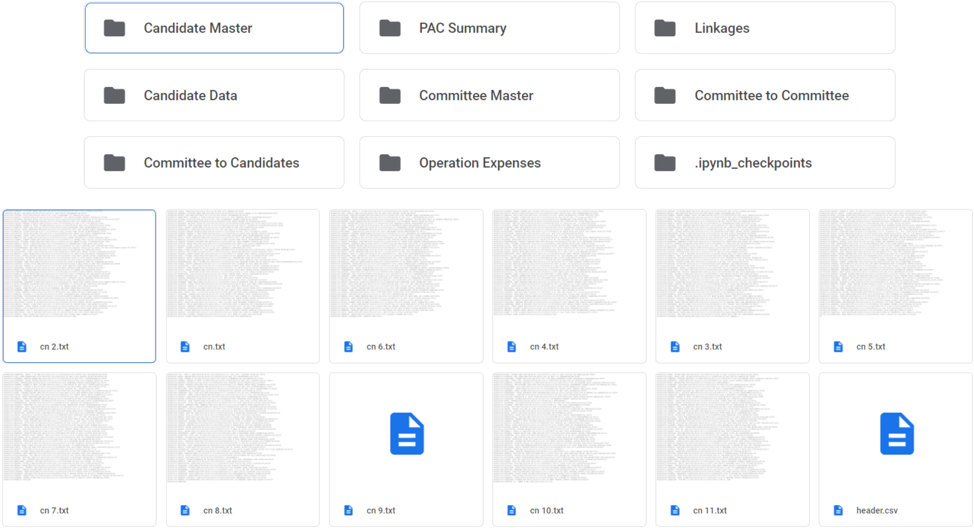
Contributions by individuals

Contributions from committees to candidates & independent expenditures

Any transaction from one committee to another

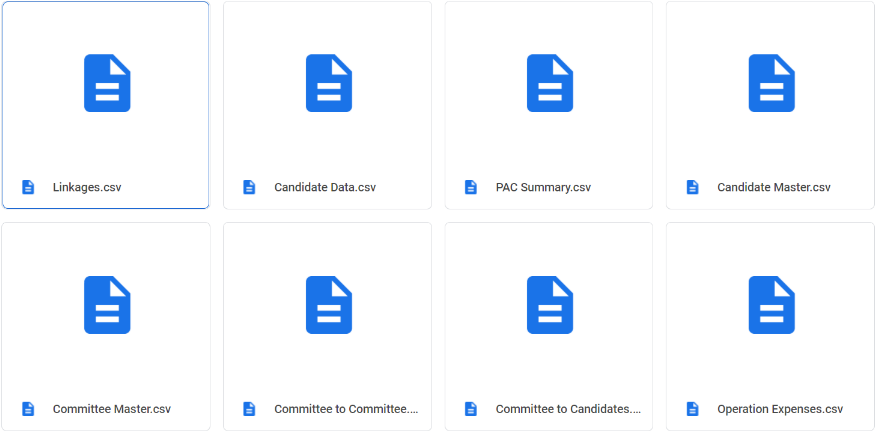
Operating expenditures

# Data Aggregation



# Data Aggregation

Useful for further analysis of election data and is available on my Github.



# Data Aggregation

Source(s):

- The Economist: Potus Model
- FiveThirtyEight

```
[ ] poll_dem_path = os.path.join(data_path, 'Polling and Demographics')
    _path = os.path.join(poll_dem_path, 'FiveThirtyEight')
```

```
[ ] polls = pd.read_csv(os.path.join(_path, 'POLLS.csv'))
    polls['electiondate'] = pd.to_datetime(polls['electiondate'])
    polls['polldate'] = pd.to_datetime(polls['polldate'])
    polls = polls.query("type_simple == 'Pres-P' & polldate < electiondate")

    good_cols = ['pollster_rating_id', 'location', 'polldate','samplesize', 'cand1_name', 'cand1_pct', 'cand2_name', 'cand2_pct', 'electiondate']
    polls = polls[good_cols]
```

| | pollster_rating_id | location | polldate | samplesize | cand1_name | cand1_pct | cand2_name | cand2_pct | electiondate |
|---|---|---|---|---|---|---|---|---|---|
| 273 | 248 | IA | 2000-01-04 | 300.0 | Gore | 45.0 | Bradley | 32.0 | 2000-01-24 |
| 274 | 304 | IA | 2000-01-04 | 600.0 | Gore | 54.0 | Bradley | 33.0 | 2000-01-24 |
| 275 | 304 | IA | 2000-01-04 | 600.0 | Bush | 45.0 | Forbes | 18.0 | 2000-01-24 |
| 276 | 281 | IA | 2000-01-10 | 304.0 | Gore | 52.0 | Bradley | 34.0 | 2000-01-24 |
| 277 | 248 | IA | 2000-01-10 | 300.0 | Bush | 46.0 | Forbes | 17.0 | 2000-01-24 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 7644 | 127 | CA | 2016-05-25 | 412.0 | Clinton | 49.0 | Sanders | 39.0 | 2016-06-07 |
| 7645 | 94 | CA | 2016-05-29 | 571.0 | Clinton | 45.0 | Sanders | 43.0 | 2016-06-07 |
| 7646 | 183 | CA | 2016-05-30 | 557.0 | Clinton | 49.0 | Sanders | 47.0 | 2016-06-07 |
| 7647 | 9 | CA | 2016-06-01 | 400.0 | Clinton | 48.0 | Sanders | 47.0 | 2016-06-07 |
| 7648 | 391 | CA | 2016-06-02 | 674.0 | Clinton | 49.0 | Sanders | 47.0 | 2016-06-07 |

1616 rows × 9 columns

# Inspiration

**Dynamic Bayesian Forecasting of Presidential Elections in the States**

Drew A. Linzer

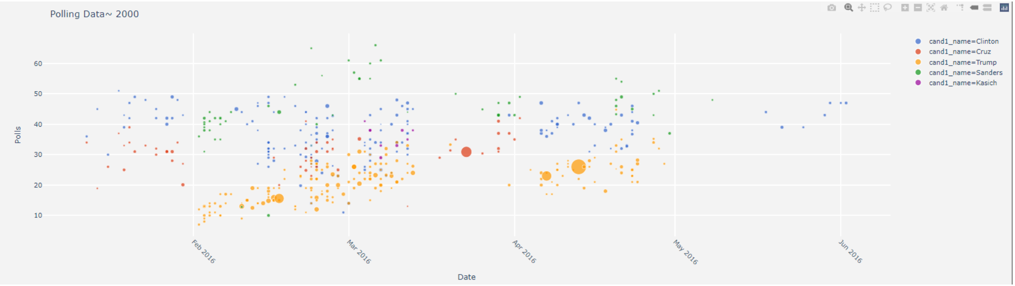
---

An Updated Dynamic Bayesian Forecasting Model for the US Presidential Election

*by Merlin Heidemanns, Andrew Gelman, and G. Elliott Morris*

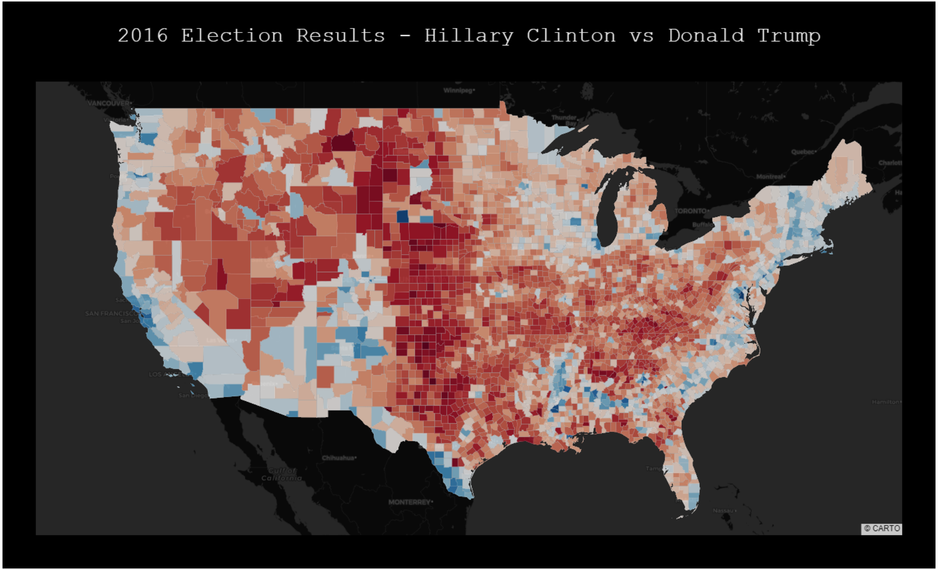
Published on Oct 27, 2020

# EDA



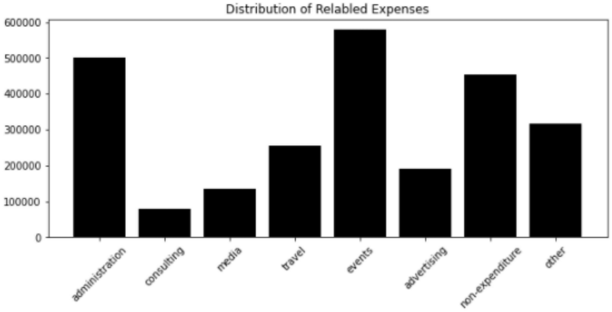
Polling Data~ 2000
60
50
40
30
20
10
Polls
Feb 2016
Mar 2016
Apr 2016
May 2016
Jun 2016
Date
cand1_name=Clinton
cand1_name=Cruz
cand1_name=Trump
cand1_name=Sanders
cand1_name=Kasich

# EDA


2016 Election Results - Hillary Clinton vs Donald Trump
© CARTO

After some basic text processing, the raw FEC data set contains 2,515,044 unique expense types. Created custom mapping to financial categories using key words resulting in only 317,846 unclassified expenses (12.64%).

```
Total key words: 315
{'administration': 43,
 'advertising': 40,
 'consulting': 43,
 'events': 68,
 'media': 50,
 'non-expenditure': 25,
 'other': 0,
 'travel': 46}
```



# Preprocessing


```
{
"administration": ["parking", "office", "desk", "computers", "lease", "rent", "audit", "cleaning", "janitor", "sanitation", "administration", "meeting", "board",
    "investments", "invest", "salary", "payroll", "staff", "administrative", "ink", "cards", "copier", "tools", "communication", "shirts", "clothes",
    "zoom", "trash", "phone", "cell", "supplies", "supply", "equipment", "paraphernalia", "paper", "computer", "office", "rent", "printing", "paper",
    "stapler", "copy", "copies"],
"consulting": ["consulting", "survey", "R&D", "hired", "planning", "consult", "advising", "advisor", "assistance", "aid", "external",
    "corporate", "opposition", "study", "strategy", "consultant", "professional fees", "investigative", "research", "canvassing", "canvass",
    "study", "analysis", "think", "thinktank", "quantitative", "statistics", "statistician", "corporate", "assistance", "assist", "detective",
    "investigator", "survaillance", "inquiry", "counter", "operation", "secret", "fbi", "analytica", "analysis", "reporting", "hunt"],
"media": ["host", "FOX", "MSNBC", "CNN", "server", "network", "footage", "editing", "recoding", "talking", "script", "commercial", "news",
    "broadcast", "show", "historical", "show", "cable", "podcast", "airtime", "lecture", "talkshow", "interview", "spotlight", "facebook", "youtube",
    "forum", "production", "tv", "cameras", "audio", "commercial", "presentation", "documentary", "web", "website", "database", "software", "site",
    "internet", "digital", "softwre", "video", "fox", "cnn", "cable", "network", "tv", "television", "radio"],
"travel": ["passport", "luggage", "storage", "carryon", "suitcase", "briefcase", "backpack", "tickets", "jet", "united", "delta", "spirit",
    "americanairlines", "americanair", "bus", "transportation", "golfcart", "cart", "buggy", "scooter", "moped", "charging", "tesla", "repairs",
    "gas", "bus", "buses", "plane", "aviation", "cab", "taxi", "uber", "lyft", "travel", "airline", "flight", "car", "luggage", "airfare", "mileage",
    "hotel", "motel", "transportation", "housing", "lodging", "auto"],
"events": ["event", "events", "party", "performance", "comedy", "comedian", "guest", "guestlist", "vip", "pamphlet", "speakers", "catering", "bingo",
    "chess", "games", "clown", "chef", "desert", "deserts", "social", "gathering", "parking", "security", "photographer", "photographers", "recording",
    "display", "projector", "function", "celebration", "baloons", "candles", "decorations", "accessories", "gifts", "flowers", "plates", "utensils", "meal",
    "food", "meals", "snack", "sandwiches", "drinks", "cook", "buffet", "catering", "cater", "dinner", "lunch", "breakfast", "restaurant", "charity", "auction",
    "fundraising", "fundraiser", "furdraising", "reception", "events", "event", "convention", "conventions", "conferences", "conference", "debate", "function", "music",
"advertising": ["stamps", "displays", "newspaper", "appealing", "garner", "cultivate", "sticker", "button", "buttons", "paper", "billboard", "board",
    "brochures", "tabloid", "tabloids", "letters", "letter", "attack", "promise", "brochure", "platform", "postage", "post", "signs", "robocalls", "polling",
    "handout", "pamphlets", "marketing", "radio", "media", "advertising", "ad", "email", "emails", "e-mail", "e-mails", "flyers", "mailing", "registration"],
"non-expenditure": ["bitcoin", "btc", "void", "refund", "stopped", "billings", "billing", "expense", "fee", "expenses", "settlement", "pay", "payment",
    "finance", "bank", "banking", "paypal", "credit", "accounting", "card", "return deposit", "fees", "tax", "taxes", "reimbursement"],
"other": []}
```

# Data

Covariates (Standardized):

- Total amount of money spent over the course of the election by both Democratic (DEM) and Republican (REP) campaigns in 7 unique expense categories, leading to 14 financial predictors.
- 10 additional demographic features for each state $i \in \mathcal{S}$ where $\mathcal{S}$ is the index set for all 51 states.

$$X_i = (\text{expenses, demographics})_i \in \mathbb{R}^{24}$$

Response:

- The proportion of DEM voters $\theta_i$ in each state given the observed counts of DEM votes $y_i$ and total turnout $N_i$.

$$y_i \sim \text{Binomial}\big(\text{logit}^{-1} X_i^T \beta_i, N_i\big)$$

# Data

| | year | state_po | candidate | candidatevotes | totalvotes | lastname |
|---|---|---|---|---|---|---|
| 0 | 2000 | AK | Al Gore | 79004.0 | 285530 | Gore |
| 1 | 2000 | AK | George W. Bush | 167398.0 | 285530 | Bush |
| 2 | 2000 | AL | Al Gore | 695602.0 | 1672551 | Gore |

| STATE | DEM administration | DEM advertising | DEM consulting | DEM events | DEM media | DEM non-expenditure | DEM travel | REP administration | REP advertising |
|---|---|---|---|---|---|---|---|---|---|
| AK | -0.168236 | -0.147651 | -0.402103 | -0.192996 | -0.400482 | -0.170038 | -0.297039 | -0.224358 | -0.21374 |
| AL | -0.169202 | -0.095724 | -0.381720 | -0.187772 | -0.400482 | -0.170201 | -0.292454 | -0.215521 | -0.21374 |
| AR | -0.166139 | -0.147662 | -0.334160 | 6.950017 | -0.388596 | -0.169912 | -0.253917 | -0.223641 | -0.21374 |

Refer to *processing.py* to see exactly how the data was prepared for modelling.

# Hierarchical binomial-logit regression model with Gaussian prior

**Model**

$$y_i \sim \text{Binomial}\big(\theta_i, N_i\big)$$

$$\text{logit}(\theta_i) = X_i^T \beta_i$$

$$\beta_i \sim \mathcal{N}(\, \beta^{\mu},\ \lambda_2 \Sigma\, )$$

$$\beta^{\mu} \sim \mathcal{N}(\, 0\,,\ \lambda_1 \mathbb{I}\, )$$

**Posterior**

$$\pi(\, \beta \,|\, X,\ Y) \propto \prod_{i=1}^{51} \text{Binomial}\big(\, y_i \,|\, \theta_i,\ N_i \,\big)\, \mathcal{N}(\, \beta_i \,|\, \lambda_2 \Sigma\, )\, \mathcal{N}(\, \beta^{\mu} \,|\, \lambda_1 \mathbb{I}\, )$$

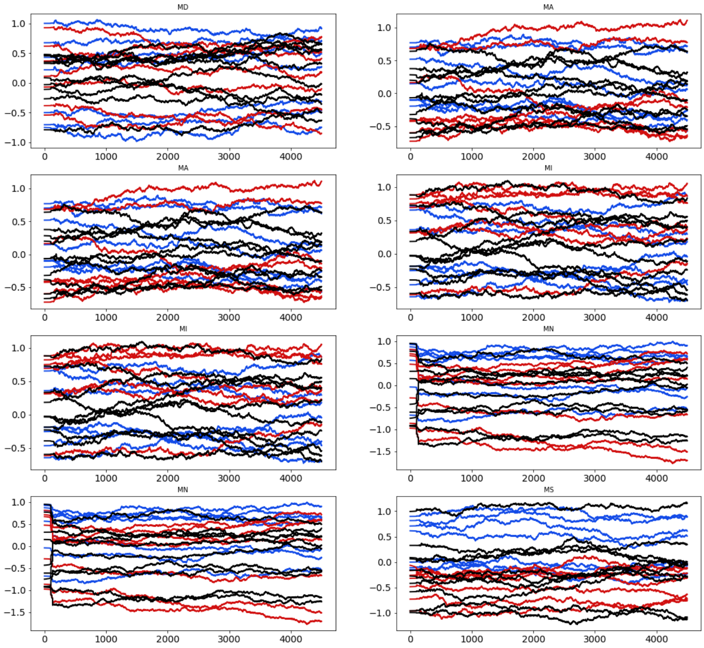
# MCMC (pymc3)

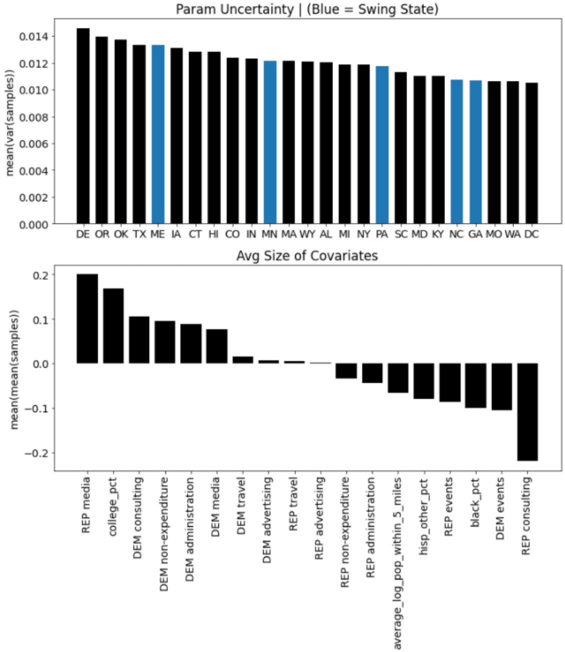
```python
# Specifying Hierarchical Model and performing MCMC
with pm.Model() as Induv_Model:
  beta_mu = pm.MvNormal('beta_mu', mu = np.zeros(num_vars), cov = Lambda1*np.eye(num_vars), shape = (num_vars,))
  beta_offset = pm.MvNormal('offset', mu = np.zeros(num_vars), cov = Lambda2*np.eye(num_vars), shape = (num_vars,))
  beta = pm.MvNormal('beta', mu = beta_mu+beta_offset, cov = Lambda2*np.eye(num_vars), shape = (num_states, num_vars))
  thetas = pm.math.invlogit(pm.math.sum(X.values*beta, axis = 1))
  likelihood = pm.Binomial("likelihood", observed = y, n = N, p = thetas)
```

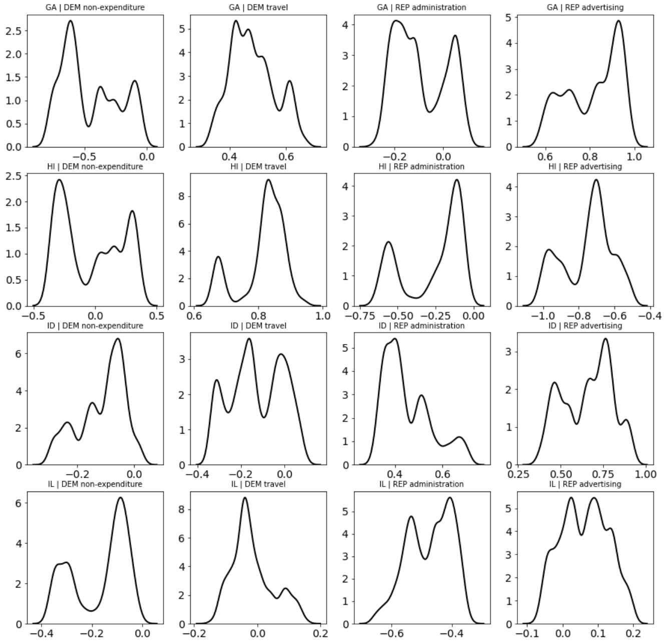
Non-Centered Parameterization (Thomas Wiecki)

FEB 08, 2017

**Why hierarchical models are awesome, tricky, and Bayesian**


MD
MA
MA
MI
MI
MN
MN
MS


Param Uncertainty | (Blue = Swing State)
mean(var(samples))
0.014
0.012
0.010
0.008
0.006
0.004
0.002
0.000
DE OR OK TX ME IA CT HI CO IN MN MA WY AL MI NY PA SC MD KY NC GA MO WA DC
Avg Size of Covariates
mean(mean(samples))
0.2
0.1
0.0
−0.1
−0.2
REP media
college_pct
DEM consulting
DEM non-expenditure
DEM administration
DEM media
DEM travel
DEM advertising
REP travel
REP advertising
REP non-expenditure
REP administration
average_log_pop_within_5_miles
hisp_other_pct
REP events
black_pct
DEM events
REP consulting


GA | DEM non-expenditure
GA | DEM travel
GA | REP administration
GA | REP advertising
HI | DEM non-expenditure
HI | DEM travel
HI | REP administration
HI | REP advertising
ID | DEM non-expenditure
ID | DEM travel
ID | REP administration
ID | REP advertising
IL | DEM non-expenditure
IL | DEM travel
IL | REP administration
IL | REP advertising

**Model**

$$y_i \sim \text{Binomial}\big(\theta_i, N_i\big)$$

$$\text{logit}(\theta_i) = X_i^T \beta_i$$

$$\beta_i \sim \sum_{h=1}^{K} w_h \mathcal{N}_h(\, \beta^{\mu},\ \lambda_2 \Sigma\,)$$

$$w \sim \text{Dirichlet}(\, \mathbf{1}_K\,)$$

$$\beta^{\mu} \sim \mathcal{N}(\, \mu,\ \lambda_1 \mathbb{I}\,)$$

**Posterior**

$$\pi(\,\beta,\ w \,|\, X,\ Y) \propto \prod_{i=1}^{51} \text{Binomial}\big(\, y_i \,|\, \theta_i,\ N_i\,\big)\, \mathcal{N}(\,\beta_i \,|\, \lambda_2 \Sigma\,) \left[\, \sum_{h=1}^{K} w_h \mathcal{N}_h(\,\beta^{\mu} \,|\, \lambda_1 \mathbb{I}\,)\right] \pi(w)$$

# Interpretation of Multimodal Posterior

- In any given state the democratic leaning sub-population will have a positive response to democratic advertisements while republicans would be less enthusiastic.
- There likely exists a rural and urban population with republican and Democratic leanings respectively with each state.
- In reality, there exists wide range of voter preferences leading to multiple underlying sub-populations all each a unique response to spending from both parties.

# Latent Voter Classes

*With a clear understanding of the existing voter classes and how they are distributed across states, a campaign can devise more targeted campaign strategies and optimize expenditures / investments.*

- After inference, the posterior state level effects $\beta_i$ can be decomposed in terms of the components of the mixture prior $\sum_{h=1}^{K} w_h \mathcal{N}_h(\beta^\mu, \lambda_2 \Sigma)$.
- Interpret components as latent voter classes with distinct preferences. Intuition tells us that there should be two dominating weights $w_i$ corresponding to strong DEM/REP voter classes.

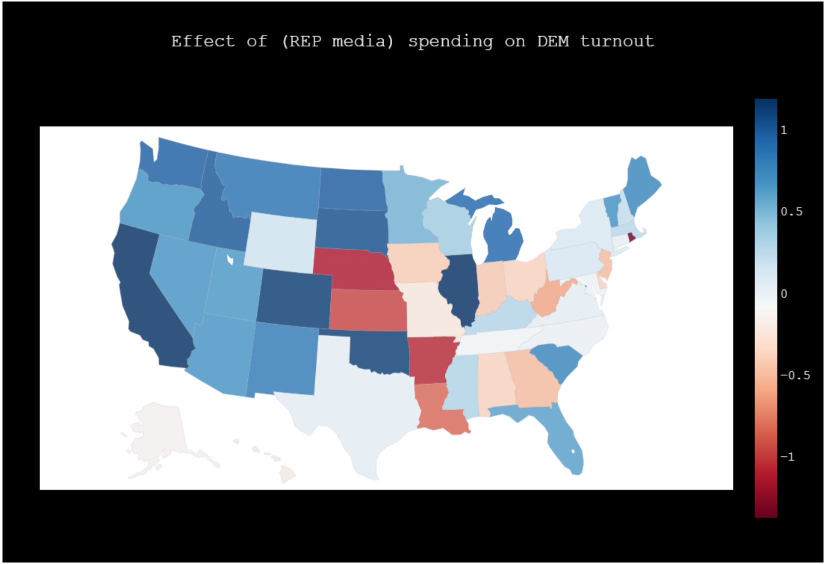
# Compromise

I was unable to successfully perform inference on the model with the mixture prior so I ended up studying the initial model.

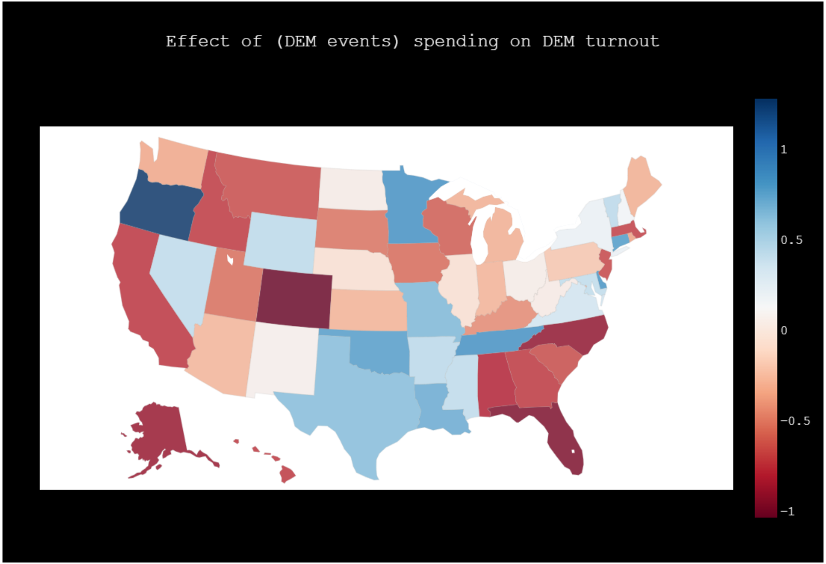
Posterior means $E[\beta_i]$:

| STATE | DEM administration | DEM advertising | DEM consulting | DEM events | DEM media | DEM non-expenditure | DEM travel |
|---|---|---|---|---|---|---|---|
| AK | 1.272752 | -0.056417 | -0.457699 | -0.918774 | -0.069939 | -0.925169 | 0.748361 |
| AL | 0.597561 | 1.165061 | 0.065351 | -0.809965 | 0.592967 | 0.584202 | 0.382308 |
| AR | 0.027215 | 1.002322 | 0.681456 | 0.429429 | 0.491575 | -0.088991 | 0.697438 |
| AZ | 0.891648 | -0.843566 | 0.985341 | -0.289751 | 0.263400 | 0.746861 | -0.527222 |
| CA | -0.232016 | 0.456978 | 0.886225 | -0.743411 | -0.020625 | -0.013290 | 0.062355 |

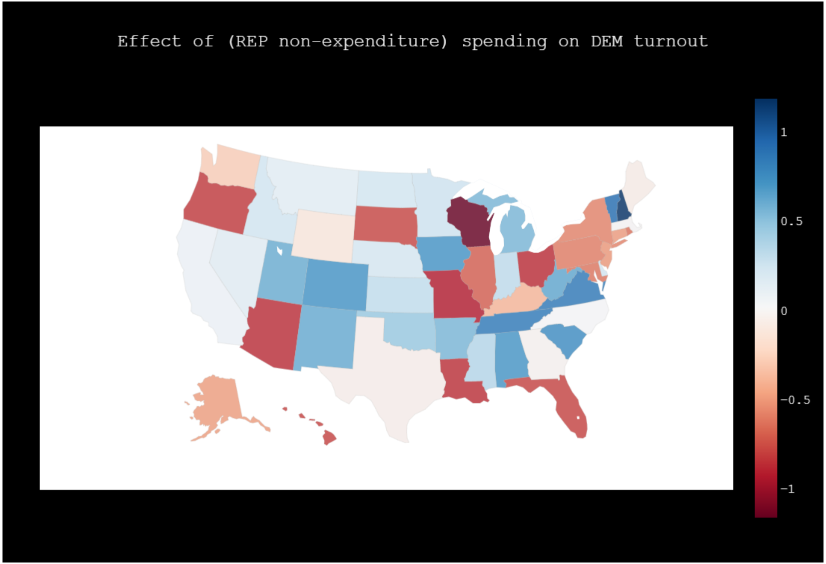
# Insights / Strategy



# Insights / Strategy



# Insights / Strategy



## Issues / Incomplete

- Create a ’utility’ curve that displays various asset allocations and their respective expected voter turnout $E[\theta_i]$ (uniform, model informed, etc).
- Did not validate model out of sample on previous elections. T
- My predictors are extremely watered down in terms of the diversity of transactions that make up each financial category.
- I did not model the joint distribution across elections (complexity explodes).
- So much information not included (Re-elections, NLP, economic factors).
- MCMC diagnostics are questionable.

# References / Data Sources


An Updated Dynamic Bayesian Forecasting Model for the US Presidential Election

Dynamic Bayesian Forecasting of Presidential Elections in the States

Post-Election Interview with Andrew Gelman and G. Elliott Morris

Why hierarchical models are awesome, tricky, and Bayesian

FEC Bulk Data

FiveThirtyEight Polling Data