Codebook for the Database on Ideology, Money in Politics, and Elections (DIME) (Version 4.0)

December 29, 2024

Description:
The Database on Ideology, Money in Politics, and Elections (DIME) provides a general resource for the study of campaign finance and ideology in American politics. The database was developed as part of an on-going effort to build a comprehensive ideological mapping of politicians, interest groups, political elites, and donors. The construction of the database required a large-scale effort to compile, clean, and process data on contribution records, candidate characteristics, and election outcomes from various sources. The current database contains over 850 million itemized political contributions made by individuals and organizations to local, state, and federal elections covering from 1979 to 2024. A corresponding database of candidates and committees provides additional information on state and federal elections.

Principal Investigator:
Adam Bonica

Contact:
Email: bonica@stanford.edu
Web-page: http://www.stanford.edu/~bonica/

How to Cite this Dataset:

Bonica, Adam. 2024. *Database on Ideology, Money in Politics, and Elections*: Public version 4.0 [Computer file]. Stanford, CA: Stanford University Libraries. <http://data.stanford.edu/dime>.

For CFscore measures, please cite:

Bonica, Adam. 2014. "Mapping the Ideological Marketplace." *American Journal of Political Science*, 58 (2): 367- 387.

For DW-DIME measures, please cite:

Bonica, Adam. 2018. "Inferring Roll-Call Scores from Campaign Contributions Using Supervised Machine Learning" *American Journal of Political Science*, 62, (4): 830-848.

Contents

1 Introduction

A core objective in constructing the database was to make data on campaign finance and elections (1) more centralized and accessible, (2) easier to work with, and (3) more versatile in terms of the types of research questions that can be addressed. To these ends, I have put a great deal of effort into compiling, processing, and augmenting the database. In making the database public, I hope to provide a valuable resource to fellow researchers. A list of the main value-added features of the database is below:

Data processing: Names, addresses, and occupational and employer titles have been cleaned and standardized.

Unique identifiers: Entity resolution techniques were used to assign unique identifiers for all individual and institutional donors included in the database. The contributor IDs make it possible to track giving by individuals across election cycles and levels of government. Additionally, unique identifiers are assigned to candidates, PACs, and others individuals/groups receiving donations.

Geocoding: Each record has been geocoded and overlaid onto congressional districts. The geocoding scheme relies on the contributor IDs to assign a complete set of consistent geocoordinates to donors that report their full address in some records but not in others. This is accomplished by combining information on self-reported address across records. The geocoding scheme further takes into account donors with multiple addresses.

Ideological measures: The common-space DIME scores (CFscores) allow for direct distance comparisons of the ideal points of a wide range of political actors from state and federal politics spanning 45 years. In total, the database includes ideal point estimates for 173,171 candidates and 42,702 political committees as recipients and 41.5 million individuals and 3.3 million organizations as donors. These measures of ideology have been extensively validated across several studies spanning a variety of institutional settings and types of actors. A compendium of these validation results can be accessed here.

Corresponding data on candidates, committees, and elections: The recipient database includes information on fundraising statistics, election outcomes, gender, and other candidate characteristics. All candidates are assigned unique identifiers that track candidates when they run for different offices/districts. The recipient IDs can also be used to match against the database of contribution records. The database also includes entries for PACs, super PACs, party committees, leadership PACs, 527s, state ballot campaigns, and other committees that engage in fundraising activities.

2 Release Notes

2.1 Version 4

- **Expanded Federal Election Coverage:** The database now includes data for all federal elections through the 2024 cycle.
- **Updated DW-DIME Scores:** DW-DIME scores have been recalculated to incorporate new data and provide the most current measures of legislative behavior.
- **New Composite Ideological Scores:** We have added new "composite" scores, which combine information from multiple sources to produce more robust estimates of ideological positions. This new measure leverages campaign contribution records, legislative roll call votes, and other sources

of ideology using multiple overimputation methods. See Section 11 for a detailed explanation of the methodology.

- **Improved Candidate Donation Matching:** The algorithm to match candidates with their personal donation records has been substantially improved using large language models (LLM). This results in significantly higher coverage and more accurate reflection of their full donation histories.
- **Enhanced Candidate Record Tracking:** Large language models (LLMs) have been incorporated to improve the record linkage for candidates. This update provides more accurate tracking of candidates who run for different offices or move across levels of government.
- **Expanded Political Elite Coverage:** Enhanced record linkage now covers a wider range of political elites, such as Federal Judges and individuals included in the Forbes 400 list.
- **Improved Geocoding Accuracy:** We have updated our geocoding process using the Pelias Geocoder, resulting in more accurate location data.
- **State-Level Data Request:** Access to NIMSP/CRP itemized contribution records for recent state-level elections is now available by request. To request access, please contact bonica@stanford.edu.

2.2 Version 3

- Updated FEC data for federal elections through 2022.
- Updated state records through 2022 using NIMSP/CRP data.
- Various improvements made to identity resolution algorithms to enhance accuracy.
- DW-NOMINATE scores updated though 118th Congress.
- DW-DIME scores updated through 2022.
- Added vote shares from the FEC for congressional primary and general elections through 2020.
- Added SQLite version of the database.
- Updated ideal point estimation algorithm for CFscores scores to adjust for over time changes across election cycles. (See section on Adjusted CFscores scores).
- Updated state-to-federal bridging algorithm that uses a supervised machine learning approach as an initial stage.

3 List of Data Files

Candidate/Recipient Files: The main recipient file includes cycle-specific entries for all candidates and committees included in the scaling. The raw candidate/recipient file additionally includes cycle-specific entries for candidates and committees that did not meet the requirements for inclusion in the scaling—including labor, corporate, and trade PACs that were excluded from the estimation stage and candidates that did not raise funds from the required number of contributors to be included in the scaling.

- `dime_recipients_1979_2024.csv` - Candidate/Recipient DIME scores (included in scaling) - State, Federal, 1979-2024 data file
- `dime_recipients_all_1979_2024.csv` - Candidate/Recipient DIME scores (all candidates) - State, Federal, 1979-2024 data file

Contributor Files: Includes entries for each individual and organization that has made political contributions.

- `dime_contributors_1979_2024.csv` - State, Federal, 1979-2024

Contribution Records: Contains the itemized contribution records grouped by election cycle. An additional set of files contains records organized by seat type. These files contain all contribution records associated with (1) gubernatorial, (2) judicial, or (3) presidential candidates.

Itemized contribution records grouped by cycle:

- `ContribDB_1980.csv` (Row count: 447,970)
- `ContribDB_1982.csv` (Row count: 314,178)
- `ContribDB_1984.csv` (Row count: 434,646)
- `ContribDB_1986.csv` (Row count: 475,688)
- `ContribDB_1988.csv` (Row count: 654,047)
- `ContribDB_1990.csv` (Row count: 1,031,190)
- `ContribDB_1992.csv` (Row count: 1,507,890)
- `ContribDB_1994.csv` (Row count: 1,790,543)
- `ContribDB_1996.csv` (Row count: 3,168,963)
- `ContribDB_1998.csv` (Row count: 5,695,396)
- `ContribDB_2000.csv` (Row count: 7,138,968)
- `ContribDB_2002.csv` (Row count: 11,253,509)
- `ContribDB_2004.csv` (Row count: 16,865,847)
- `ContribDB_2006.csv` (Row count: 21,019,196)
- `ContribDB_2008.csv` (Row count: 24,951,708)
- `ContribDB_2010.csv` (Row count: 24,621,247)
- `ContribDB_2012.csv` (Row count: 37,695,364)
- `ContribDB_2014.csv` (Row count: 33,467,487)
- `ContribDB_2016.csv` (Row count: 52,704,059)
- `ContribDB_2018.csv` (Row count: 45,450,271)
- `ContribDB_2020.csv` (Row count: 224,227,056)
- `ContribDB_2022.csv` (Row count: 143,540,873)
- `ContribDB_2024.csv` (Row count: 202,886,167)
- `ContribDB_nimsp.csv` (Row count: 27,352,201) (Available by request)

Itemized contribution records grouped by seat type:

- `contribDB_governor.csv` (Row count: 19,292,776)
- `contribDB_judicial.csv` (Row count: 2,301,859)
- `contribDB_president.csv` (Row count: 116,068,524)

Contingency Matrix of Contribution Amounts: Includes an R list object organizes the contribution data into an n by m contingency matrix of contribution amounts.

- `mimp.rdata` - Sparse matrix of contribution amounts

DIME SQLite Database: A single SQLite file that combines the above files in a single SQL database. This is the best format for working with the database on memory restricted machines. The file will reside on disk rather than memory and can be accessed using the SQLite interface functionality of R or Python. The SQLite database includes pre-built indexes for key variables.

For a primer on how to interface with the SQLite database in R, see the following example code: dime_sqlite_v3_example.R

Extensions: The DIME+ data repository on congressional activity extends DIME to cover detailed data on legislative voting, lawmaking, and political rhetoric. (See `http://data.stanford.edu/dime-plus` for details.)

Identifying sets of important political actors: Contribution records have been matched onto other publicly available databases of important political actors. Examples include:

- Fortune 500 directors and CEOs (Data) (Paper)
- State supreme court justices (Data) (Paper)
- Federal court judges (Data) (Paper)
- Executives appointees to federal agencies (Data) (Paper).
- Medical professionals (Data) (Paper)

Each of the above are available for download. (Note that some of these curated datasets have been compiled using previous versions of DIME.)

4 Data Sources

Federal Elections: Contribution records, candidate and committee filings, and election outcomes for federal elections are from the Federal Election Commission (FEC).

State Elections: Contribution records, candidate and committee filings, and election outcomes for state elections are primarily from state reporting agencies. For a subset of the records in more recent election cycles, data are provided by the Center for Responsive Politics (CRP) and the National Institute for Money in State Politics (NIMSP). The itemized records of CRP/NIMSP are excluded from the publicly available downloads. These account for 27,352,201 records in total. Please contact me at bonica@stanford.edu to request access for strictly academic use. The CRP/NIMSP data are licensed under the Creative Commons

Attribution-Non-commercial-Share Alike 3.0 United States License.

527s: Donation records to 527s are from the IRS.

Other Data:

DW-NOMINATE scores are provided by Keith Poole and Howard Rosenthal and are available for download at http://www.voteview.com.

Cite: Poole, Keith T., and Howard Rosenthal. 2007. *Ideology & Congress.* 2nd rev. ed. New Brunswick: Transaction Publishers.

Data on historical boundaries for congressional districts are from:

Jeffrey B. Lewis, Brandon DeVine, Lincoln Pitcher, and Kenneth C. Martis. (2013) Digital Boundary Definitions of United States Congressional Districts, 1789-2012. Version 1.2 (March 20, 2015). Retrieved from http://cdmaps.polisci.ucla.edu.

Geocoding performed using the opensource Pelias geocoder (https://pelias.io/). Shape files for census tracts are from Census.gov: (http://www.census.gov/rdo/data/).

5 Variable Codings

5.1 Candidate/Recipient Database

The candidate/recipient database includes information on voting records, fundraising statistics, election outcomes, and other candidate characteristics. All candidates are assigned a unique identifiers that track them as they move across offices. The recipient IDs can also be used to match against the database of contribution records. The database also includes entries for PACs, super PACs, leadership PACs, 527s, state ballot campaign committees, and other recipient committees that engage in fundraising activities. Entries for candidates with two or more active fundraising committees during a given election cycle have been de-duped. The methodology used to estimate common-space DIME scores (or CFscores) for contributors is described in "Mapping the Ideological Marketplace" (Bonica 2014). The methodology used to estimate the DW-DIME scores is described in "Inferring Roll-Call Scores from Campaign Contributions Using Supervised Machine Learning" (Bonica 2018).

Variable Descriptions:

1) `election:` Election cycle preceded by two-letter state code. Federal candidates have 'fd' as the state code.

2) `cycle:` Four digit number that indicates the two-year election cycle during which the contribution was recorded.

3) `fecyear:` Year listed by the FEC indicating the year of campaign's the target election. The 'election' variable indicates the election cycle during which the contribution was received. But the election can occur in a future cycle—as is the case for senators that fundraise during their first four years in office.

4) `bonica.rid:` Unique ID assigned to candidates/recipients. Each candidate/recipient receives a unique ID that is constant across election cycles, levels of government, and offices sought. With the exception of party-switchers the bonica.rid values have a one-to-one correspondence with the ICPSR2 scores. (Use this variable to merge with the database of contribution records).

5) `bonica.cid:` The unique contributor ID for the candidate. This variable can be used to match candidates with their personal contributions records. (Note: The construction of this variable is not yet complete. It includes a partial set candidates that could be easily linked with their contribution records using an automated matching scheme. Missing values do not necessarily mean that a candidate has not made contributions).

6) `name:` Name of the candidate/recipient.

7) `lname:` Last name of the candidate/recipient.

8) `ffname:` Concatenates first name, middle name, suffix, and title.

9) `fname:` First name of the candidate/recipient.

10) `mname:` First name of the candidate/recipient.

11) `title:` Title of the candidate/recipient. (e.g. Mr., Mrs., Dr., Esq).

12) `suffix:` Suffix of the candidate/recipient.

13) `party:` Party of candidate/recipient (100 = Dem, 200 = Rep, 328 = Ind).

14) `state:` Two-letter state abbreviations.

15) `seat:` Office sought. Committees are listed as federal:committee. See section on seat codes below.

16) `district:` District code: two-letter state code followed by congressional district number. District numbers for senate candidates take on the value of 'S' followed by the year of the seat will be up for election.

17) `ico.status:` Incumbency status. ('I' = Incumbent, 'C' = Challenger, 'O' = Open Seat Candidate, " - not up for election).

18) `cand.gender:` Candidate gender codings. (With the exceptions of candidates that have served in Congress, all gender codes are based on an automated coding scheme that incorporates information gender ratios of first names as reported by the U.S. Census and gender-specific titles (e.g. Mrs., Mr., Jr., Sr.) reported in the contribution records.

19) `recipient.cfscore:` Estimated ideology of candidate/recipient based on donations received.

20) `recipient.cfscores.dyn:` Period-specific estimates of ideology. (Candidate/recipient scores are re-estimated in each election cycle while holding contributor scores constant.)

21) `contributor.cfscore:` Estimated ideology of candidate/recipient based on their personal donations given to other candidates/recipients.

22) `dwdime:` DW-DIME scores for recipients. These scores are described in detail in "Inferring Roll-Call Scores from Campaign Contributions Using Supervised Machine Learning" Bonica (2018).

23) `composite.score:` Composite ideological score for recipients, combining information from multiple sources. See Section 11 for details.

24) `dwnom1:` (voteview.com) First dimension common-space DW-NOMINATE score. (Based on joint scaling of the 1st to the 117th Congresses.)

25) `dwnom2:` (voteview.com) Second dimension common-space DW-NOMINATE score.

26) `ps.dwnom1:` (voteview.com) First dimension Nokken-Poole period-specific DW-NOMINATE score. (Scores for the House and Senate are scaled separately and thus should not be directly compared.)

27) `ps.dwnom2:` (voteview.com) Second dimension Nokken-Poole period-specific DW-NOMINATE score.

28) `irt.cfscores:` Estimates of ideology for recipients/candidates from IRT count-model applied to PAC data. (See "Ideology and Interests in the Political Marketplace" Bonica (2013) for details.)

29) `num.givers:` Number of distinct donors that gave to the candidate during a specific election cycle.

30) `num.givers.total:` Number of distinct donors that gave to the candidate/recipient aggregating over the candidate/recipient's career.

31) `total.receipts:` Sum total of contributions raised during an election cycle.

32) `total.disbursements:` Sum total campaign disbursements during an election cycle..

33) `total.indiv.contribs:` Sum total of itemized contributions from individuals raised during an election cycle.

34) `unitemized:` Sum of unitemized contributions from individuals raised during an election cycle.

35) `total.pac.contribs:` Sum total contributions raised from contributions from PACs and other committees.

36) `total.party.contribs:` Sum total raised from party committees.

37) `contribs.from.candidate:` Sum total raised from candidate self-contributions.

38) `ind.exp.support:` Sum total independent expenditures made to support the candidate.

39) `ind.exp.oppose:` Sum total independent expenditures made to oppose the candidate.

40) `prim.vote.pct:` FEC reported vote share in primary election. Primary election outcomes are only coded for federal congressional candidates.

41) `pwinner:` Primary election outcome ('W' = won election; 'L' = lost election). Primary election outcomes are only coded for federal congressional candidates. See `nimsp.candidate.status` for additional information on race outcomes for state elections.

42) `gen.vote.pct:` FEC reported vote share in general election.

43) `gwinner:` General election outcome ('W' = won election; 'L' = lost election). Election outcomes for congressional candidates are from the FEC. Election outcomes for state-level candidates are coded based on `nimsp.candidate.status`.

44) `s.elec.stat:` FEC special election code (W = Win) (L = Lose).

45) `r.elec.stat:` FEC run-off election code (W = Win) (L = Lose).

46) `district.pres.vs:` District-level percentage of the two-party vote share won by the Democratic presidential nominee in the most recent (or concurrent) presidential election.

47) `fec.cand.status:` Indicates the status of the candidate's campaign assigned by the FEC. ('C' = Statutory candidate; 'F' = Statutory candidate for future election; 'N' = Not yet a statutory candidate; 'P' = Statutory candidate in prior cycle).

48) `recipient.type:` ('cand' = candidate, 'comm' = committee).

49) `igcat:` FEC interest group category code (C = Corporation, L = Labor organization, M = Membership organization, T = Trade association, V = Cooperative, W = Corporation without capital stock).

50) `comtype:` FEC code for type of committee.

51) `ICPSR:` Adjusted ICPSR legislator ID. Candidates that have never served in Congress are assigned IDs based off of their candidate IDs assigned by the FEC, NIMSP, or state reporting agencies. The four-digit election cycle is appended to the ID. Candidates that are active in multiple election cycles (or file to run for multiple seats during a single election cycle) will appear multiple times. This variable provides a unique row identifier.

52) `ICPSR2:` Adjusted ICPSR legislator ID. Candidates that have never served in Congress are assigned IDs based off of their candidate IDs assigned by the FEC, NIMSP, or state reporting agencies. Party switchers are assigned new ICPSR2 IDs after switching parties. (See *before.switch.ICPSR* and *after.switch.ICPSR*.)

53) `Cand.ID:` The candidate ID assigned by the FEC or state reporting agency.

54) `FEC.ID:` The committee ID assigned by the FEC or state reporting agency campaign committee.

55) `NID:` (CRP/NIMSP) Unique candidate IDs assigned by the Center for Responsive Politics/National Institute for Money in State Politics.

56) `before.switch.ICPSR:` ICPSR ID prior to switching parties (included for party-switchers only).

57) `after.switch.ICPSR:` ICPSR ID after switching parties (included for party-switchers only).

58) `party.orig:` Original party before switch.

59) `nimsp.party:` (nimsp) three-letter party code assigned by the NIMSP.

60) `nimsp.candidate.ICO.code:` (nimsp) incumbency status assigned by the NIMSP.

61) `nimsp.district:` (nimsp) district number assigned by the NIMSP.

62) `nimsp.office:` (nimsp) state-office sought.

63) `nimsp.candidate.status:` (nimsp) election outcome.

5.2 Contributor Database

Includes rows for individuals and organizational donors included in the database. *Note that all donors are assigned ideal point estimates, including one-off donors who contributed only to single candidate or committee.* The *is.projected* column is used to indicate donors that were excluded from the estimation stage and later projected onto the recovered space as supplementary observations. Contributors who have only given to a single recipient are assigned the ideal point of the recipient. Researchers should aware of this when deciding which donors/ideal points to include in their study. The *num.distinct* column indicates the number of observations that were used to estimate the contributors ideal point. Typically, donating

to eight or more distinct recipients is sufficient to recover a reliable ideal point estimate.

Variable Descriptions:

1) `bonica.cid:` Unique contributor IDs for each donor in the database.
2) `contributor.type:` Contributor type ('I' = individual, 'C' - committee/organization).
3) `num.records:` The number of records in the contribution database by the donor.
4) `num.distinct:` The number of distinct recipients included in the scaling receiving contributions from the donor.
5) `most.recent.contributor.name:` Contributor's self-reported name from most recent record.
6) `most.recent.contributor.address:` Contributor's self-reported street address from most recent record.
7) `most.recent.contributor.city:` Contributor's self-reported name city/municipality on most recent record.
8) `most.recent.contributor.zipcode:` Contributor's self-reported zip-code (5 or 9 digits) from most recent record.
9) `most.recent.contributor.state:` Contributor's self-reported state from most recent record.
10) `most.recent.latitude:` Geocoded latitude from the most recent record.
11) `most.recent.longitude:` Geocoded longitude from the most recent record.
12) `most.recent.contributor.occupation:` Contributor's self-reported occupational title from most recent record.
13) `most.recent.contributor.employer:` Contributor's self-reported employer from most recent record.
14) `most.recent.transaction.id:` transaction.id value of the most recent record from the contribution database.
15) `most.recent.transaction.date:` date of the most recent record from the contribution database.
16) `contributor.gender:` Contributor gender ('F'=Female, 'M'=Male, 'U'=Uncoded/Unknown)
17) `is.corp:` Indicates whether the contributor is identified as either a corporate entity or q trade organization (only applies to committees). Takes on the value 'corp' for corporations and trade organizations and is blank otherwise. These donors are excluded from the scaling.
18) `contributor.cfscore:` Contributor CFscore.
19) `is.projected:` Indicates whether the was excluded from the estimation stage but was later projected onto the recovered space as supplementary observations. This will take on the value of 1 for PACs and organizations directly affiliated with corporations or trade organizations and individual donors who gave to a single recipient.
20) `first.cycle.active:` The first recorded cycle in which the donor was active.
21) `last.cycle.active:` The last recorded cycle in which the donor was active.
22) `amount.cycle.[YYYY]:` Total amount contributed in a given election cycle.

5.3 Contribution Database

The contribution database includes a complete set of contribution records grouped by election cycle. Each row represents and individual transaction between a donor and recipient.

NOTE: There are several columns of ID numbers for recipients. Use bonica.rid and bonica.cid as the default. The other columns are included so that the database can be linked back to other data sources.

Variable Descriptions:

1) `cycle:` Election Cycle.
2) `transaction.id:` A primary key that contains a unique transaction id for each record.
3) `transaction.type:` FEC code for transaction type. (See section below for details.)
4) `amount:` Dollar amount of the contribution.
5) `date:` Transaction date of the contribution.
6) `bonica.cid:` A unique contributor id assigned to each individual and organization in the database.
7) `contributor.name:` Complete name of contributor (last, first); suffix and title removed.
8) `contributor.lname:` Last name of contributor.
9) `contributor.fname:` First name of contributor.
10) `contributor.mname:` Middle name or initial of contributor.
11) `contributor.suffix:` Suffix of contributor (e.g. Jr., Sr.).
12) `contributor.title:` Title of contributor (e.g. Mr., Mrs., Dr., Esq).
13) `contributor.ffname:` Concatenates first name, middle name, suffix, and title.
14) `contributor.type:` ('I' = individual; 'C' = committee or organization).
15) `contributor.gender:` Contributor gender coding ('M' = male; 'F' = female; 'U' = unknown). Gender codes are based on an automated coding scheme that incorporates information gender ratios of first names as reported by the U.S. Census and gender-specific titles (e.g. Mrs., Mr., Jr., Sr).
16) `contributor.address:` Contributor's self-reported street address.
17) `contributor.city:` Contributor's self-reported name city/municipality.
18) `contributor.state:` Contributor's self-reported state.
19) `contributor.zipcode:` Contributor's self-reported zip-code (5 or 9 digits).
20) `contributor.occupation:` Contributor's self-reported occupational title.
21) `contributor.employer:` Contributor's self-reported employer.
22) `is.corp:` Indicates whether the contribution is made by a corporate entity or q trade organization (only applies to committees). Takes on the value 'corp' for corporations and trade organizations and is blank otherwise.

23) `recipient.name:` Name of the recipient candidate or committee.

24) `bonica.rid:` Unique ID for recipients. Can be matched against candidate database which contains more detailed information on candidates, elections, and constituencies.

25) `recipient.party:` Party of the recipient (100=DEM; 200=REP; 328 = IND). (Match against candidate database for more detailed party codings.)

26) `recipient.type:` ('CAND' = candidate; 'COMM' = PAC, organization, or party committee)

27) `recipient.state:` Two-letter state abbreviation of the recipients.

28) `seat:` Elected office sought by candidate.

29) `election.type:` ('P' = primary elections; 'G' = general elections).

30) `latitude:` Geo-location (latitude).

31) `longitude:` Geo-location (longitude).

32) `gis.confidence:` A measure of confidence of the accuracy of the GIS coordinates.

33) `contributor.district.90s:` Contributor's geocode mapping onto a congressional district with respect to boundaries for 1992-2000.

34) `contributor.district.00s:` Contributor's geocode mapping onto a congressional district with respect to boundaries for 2002-2010.

35) `contributor.district.10s:` Contributor's geocode mapping onto a congressional district with respect to boundaries after 2010. Includes updated district boundaries for 2022.

36) `censustract:` Contributor's geocode mapping onto a census tract.

37) `efec.memo:` Memo field from FEC electronic filings.

38) `efec.memo2:` Auxiliary memo field from FEC electronic filings.

39) `efec.transaction.id.orig:` Original transaction id from FEC electronic filings.

40) `bk.ref.transaction.id:` Indicates whether the contribution record previously appeared in the database. The value link back to transaction.id. This can be used to remove duplicate entries.

41) `efec.org.orig:` Original recipient name from from FEC electronic filings.

42) `efec.comid.orig:` Original committee ID from FEC electronic filings.

43) `efec.form.type:` Form type from FEC electronic filings.

44) `contributor.cfscore:` Contributor's ideal CFscore.

45) `candidate.cfscore:` Candidate/recipient's CFscore.

6 Candidate-Contributor Contingency Matrix

The file includes an R list object that contains sparse matrix objects that organize the contribution data into n by m contingency matrices of amounts where the rows index contributors, the columns index candidates/recipients, and each entry R_{ij} stores the total amount contributor i gives to recipient j. Note that the cell values do not represent raw dollar amounts. Rather, they are the transformed values used to recover the common-space CFscores. The transformation is based on a normalization scheme that helps to adjust for variation in contribution limits by converting contribution amounts to count values. The conversion is based on federal contribution limits. Contributions between $1 and $100 are coded as 1, contributions between $101 and $200 are coded as 2, and so on. Contributions of $5,000 or greater are capped at 50. (See "Mapping the Ideological Marketplace" (Bonica 2014) for details.) Another set of sparse matrix objects report the raw dollar amounts for each cell.
The R list object contains four sparse matrices:

- `mimp$contrib.matrix` - Sparse matrix of count values with columns indexed by `bonica.rid.cycle`. (Includes separate columns for each candidate/recipient-cycle observation.)
- `mimp$cm` - Sparse matrix of count values with columns indexed by `bonica.rid`. (Collapses columns such that each candidate/recipient has a single column and cell values are aggregate amounts given across cycles.)
- `mimp$contrib.matrix.am` - Sparse matrix of raw dollar amounts with columns indexed by `bonica.rid.cycle`. (Includes separate columns for each candidate/recipient-cycle observation.)
- `mimp$cm.am` - Sparse matrix of raw dollar amounts with columns indexed by `bonica.rid`. (Collapses columns such that each candidate/recipient has a single column and cell values are aggregate amounts given across cycles.)
- `mimp$cands` - Candidate/recipient database as a data.frame object.
- `mimp$contribs` - Contributor database as a data.frame object.

7 Seat Labels

List of seat labels assigned to candidates with respect to their target office:

Candidates	
federal:house	U.S. House of Representatives
federal:senate	U.S. Senate
federal:president	U.S. President
state:upper	upper chamber of state legislature
state:lower	lower chamber of state legislature
state:office	state-wide office (see `nimsp.seat` detailed codes)
state:governor	governor
state:ltgov	lt. governor
state:secofstate	secretary of state
state:treasurer	treasurer/comptroller
state:attorneyg	attorney general
state:judicial	state courts
state:attorney	district attorney
state:education	board or education/superindendent
state:other	misc. offices
state:unknown	unknown
Committees	
federal:committee	federal committee
federal:527	527 organization
state:committee	state committee
state:ballot	state-level ballot measure committee
state:other:convention	convention committee

Note: Various local election seats are labelled as "local:" followed by a label for the type of office. These have not been standardized.

8 Transaction Codes

10	NON-FEDERAL RECEIPT FROM PERSONS LEVIN (L-1A)
11	TRIBAL CONTRIBUTION
12	NON-FEDERAL OTHER RECEIPT LEVIN (L-2)
13	INAUGURAL DONATION ACCEPTED
15	CONTRIBUTION
15C	CONTRIBUTION FROM CANDIDATE
15E	EARMARKED CONTRIBUTION
15F	LOANS FORGIVEN BY CANDIDATE
15I	EARMARKED INTERMEDIARY IN
15J	MEMO (FILER'S \% OF CONTRIBUTION GIVEN TO JOIN
15T	EARMARKED INTERMEDIARY TREASURY IN
15Z	IN-KIND CONTRIBUTION RECEIVED FROM REGISTERED
16C	LOANS RECEIVED FROM THE CANDIDATE
16F	LOANS RECEIVED FROM BANKS
16G	LOAN FROM INDIVIDUAL
16H	LOAN FROM CANDIDATE/COMMITTEE
16J	LOAN REPAYMENTS FROM INDIVIDUAL
16K	LOAN REPAYMENTS FROM CANDIDATE/COMMITTEE
16L	LOAN REPAYMENTS RECEIVED FROM REGISTERED EN
16R	LOANS RECEIVED FROM REGISTERED FILERS
16U	LOAN RECEIVED FROM UNREGISTERED ENTITY
17R	CONTRIBUTION REFUND RECEIVED FROM REGISTERED
17U	REF/REB/RET RECEIVED FROM UNREGISTERED ENTITY
17Y	REF/REB/RET FROM INDIVIDUAL/CORPORATION
17Z	REF/REB/RET FROM CANDIDATE/COMMITTEE
18G	TRANSFER IN AFFILIATED
18H	HONORARIUM RECEIVED
18J	MEMO (FILER'S \% OF CONTRIBUTION GIVEN TO JOIN
18K	CONTRIBUTION RECEIVED FROM REGISTERED FILER
18S	RECEIPTS FROM SECRETARY OF STATE
18U	CONTRIBUTION RECEIVED FROM UNREGISTERED COMMI
19	ELECTIONEERING COMMUNICATION DONATION RECEIVE
19J	MEMO (ELECTIONEERING COMMUNICATION \% OF DONAT
20	DISBURSEMENT - EXEMPT FROM LIMITS
20A	NON-FEDERAL DISBURSEMENT LEVIN (L-4A) VOTER R
20B	NON-FEDERAL DISBURSEMENT LEVIN (L-4B) VOTER I
20C	LOAN REPAYMENTS MADE TO CANDIDATE
20D	NON-FEDERAL DISBURSEMENT LEVIN (L-4D) GENERIC
20F	LOAN REPAYMENTS MADE TO BANKS
20G	LOAN REPAYMENTS MADE TO INDIVIDUAL
20R	LOAN REPAYMENTS MADE TO REGISTERED FILER
20V	NON-FEDERAL DISBURSEMENT LEVIN (L-4C) GET OUT
22G	LOAN TO INDIVIDUAL
22H	LOAN TO CANDIDATE/COMMITTEE
22J	LOAN REPAYMENT TO INDIVIDUAL
22K	LOAN REPAYMENT TO CANDIDATE/COMMITTEE
22L	LOAN REPAYMENT TO BANK
22R	CONTRIBUTION REFUND TO UNREGISTERED ENTITY

22U LOAN REPAID TO UNREGISTERED ENTITY
22X LOAN MADE TO UNREGISTERED ENTITY
22Y CONTRIBUTION REFUND TO INDIVIDUAL
22Z CONTRIBUTION REFUND TO CANDIDATE/COMMITTEE
23Y INAUGURAL DONATION REFUND
24A INDEPENDENT EXPENDITURE AGAINST
24C COORDINATED EXPENDITURE
24E INDEPENDENT EXPENDITURE FOR
24F COMMUNICATION COST FOR CANDIDATE (C7)
24G TRANSFER OUT AFFILIATED
24H HONORARIUM TO CANDIDATE
24I EARMARKED INTERMEDIARY OUT
24K CONTRIBUTION MADE TO NON-AFFILIATED
24N COMMUNICATION COST AGAINST CANDIDATE (C7)
24P CONTRIBUTION MADE TO POSSIBLE CANDIDATE
24R ELECTION RECOUNT DISBURSEMENT
24T EARMARKED INTERMEDIARY TREASURY OUT
24U CONTRIBUTION MADE TO UNREGISTERED
24Z IN-KIND CONTRIBUTION MADE TO REGISTERED FILER
29   ELECTIONEERING COMMUNICATION DISBURSEMENT(S)

NON-FEC CODES
15S   CONTRIBUTION TO STATE ELECTIONS (CATCHALL)
15L   CONTRIBUTION TO LOCAL ELECTIONS (CATCHALL)
15PD CONTRIBUTION MADE AS PAYROLL DEDUCTION
PF   PUBLIC FUNDING (STATE LEVEL)
PFR   PUBLIC FUNDSING RETURNED (STATE LEVEL)

9 Improvements to Scaling Algorithm

9.1 Adjusted CFscores scores

During the 42-year period covered by DIME, the campaign finance landscape has changed dramatically. The volume of contributions per election cycle has increased from a few hundred thousand itemized contributions per election cycle in the 1980s to over two-hundred million itemized contributions in the 2020 election cycle alone. This represents more than 1000-fold increase in data volume. This increase occurred alongside the rise of online giving and increasingly sophisticated fundraising operations, which has skewed the information entering the model towards more recent election cycles.

To account for these developments, the DIME CFscore model has been adjusted. It does so by taking advantage of candidates continuing across election cycles. First, period-specific ideal point are estimated for candidates, while holding the contributor ideal points constant. This yields a vector of ideal points, Y_{it}, for candidate i in cycle t. The vector of cycle-specific ideal points is regressed on fixed effects for cycle and candidate, controlling for incumbency, seat competitiveness, and percentage of contributions raised online through conduits, such as ActBlue and WinRed. The regression equation is:

$$Y_{it} = cand_i + cycle_t + incumbent_{it} + competitive\_seat_{it} + pct\_conduit_{it},$$

weighted by:

$$\omega_{it} = \frac{ln(num\_donations_{it})}{ln(num\_donations_i)} \times ln(num\_donations_i).$$

The adjusted candidate CFscores are set equal the fitted values for candidate fixed effects, $cand_i$. The regression is estimated separately for each party.

10 State-to-Federal Bridging

The state-to-federal bridging algorithm now applies a supervised machine learning approach (specifically, a random forest model) similar to that developed in Bonica (2018). For each state, the contributor matrix is subset donors who have given to state-level candidates in that state and federal candidates at large. A random forest model (a machine learning algorithm used for prediction) is trained to predict the candidate scores recovered from federal elections. Score for state-level candidates are then imputed based on their donors who have also given to federal elections. The imputed scores are limited to state-level candidates with at least 15 donors who had also given to federal candidates. This threshold was selected to balance the need for an adequate sample with the need to include as many state-level candidates as possible. The scores for state-level candidates that do not meet the threshold are recovered using the bridging algorithm outlined in Bonica (2014). A key difference is that scores for state-level candidates that have been imputed via random forest are now treated as fixed values, similar to federal candidates, by the bridging algorithm.

11 Composite Ideological Scores

To provide users with a more comprehensive and robust measure of candidate ideology, the DIME database now includes a composite ideological score, which combines multiple established measures. This composite score is designed to address concerns about relying on any single measure of ideology and to capture a broader range of candidate behaviors and activities.

The composite score is generated using the multiple over-imputation (MO) framework developed by Blackwell, Honaker, and King (2017). This framework is specifically designed to handle measurement error and missing data across multiple variables. The MO framework operates in two steps: (1) it imputes missing values for each variable while accounting for measurement error in observed values; and (2) it over-imputes the observed values to also account for error. Principal component analysis (PCA) is subsequently applied to the set of imputed and over-imputed values to combine them into a single unified dimension, which can be interpreted as a left-right ideological continuum. This approach collapses information from various sources into a single estimate that is strongly correlated with each input measure.

The composite score integrates eight distinct measures of ideology, drawing on different data sources and methodologies:

1. **Recipient CFscores:** Ideological scores based on campaign contributions received by the candidate (Bonica, 2014).
2. **Donor CFscores:** Ideological scores based on the personal donations given by the candidate to other candidates and committees (Bonica, 2014).
3. **DW-DIME scores:** Scores inferred from campaign contributions that predict legislative roll call voting behavior (Bonica, 2018).
4. **DW-NOMINATE scores:** Common-space scores derived from congressional roll call votes (McCarty, Poole, and Rosenthal, 2016).
5. **Contributor DW-NOMINATE scores:** DW-NOMINATE scores calculated from the roll call votes of legislators who received donations from a candidate.
6. **Hall and Snyder (2013) scores:** Ideal point estimates based on a count model that use information on PAC contributions.
7. **GGUM End-against-the-middle roll call scores**: Scores based on generalized graded unfolding model that model candidates as either supporting the extremes of each vote. (Duck-Mayr and Montgomery, 2023)
8. **Shor and McCarty's 2011 state legislative roll call scores:** Ideal point estimates from state legislative roll call votes.

By combining these distinct measures—including information derived from campaign contributions, roll call votes, and PAC support—the composite score offers a robust, unified estimate of candidate ideology that is less sensitive to the limitations of any single measure. This approach leverages the information from measures constructed in different settings and reflecting different incentive structures. The resulting composite measure offers a more comprehensive and holistic conceptualization of ideology.
Table 1 displays the correlation between the composite scores and the input measures. As shown in this table, the composite score strongly correlates with each input. Furthermore, the composite measure is not overly reliant on any single input. Importantly, our measure captures nearly all of the within-party variation in both CFscores and DW-NOMINATE, addressing concerns about the observed divergence between these measures for Democrats in recent years (Tausanovitch and Warshaw, 2017; Barber, 2022).

Table 1: Correlations between Composite Measure of Ideology and Input Measures

	Overall	Democrats	Republicans
Recipient DIME	0.978	0.800	0.775
Contributor DIME	0.982	0.790	0.786
Contributor DWNOM	0.948	0.614	0.679
Hall and Snyder	0.966	0.839	0.840
DW-DIME	0.991	0.927	0.957
DW-NOMINATE	0.988	0.840	0.935
GGUM	0.979	0.800	0.877
Shor-McCarty	0.935	0.671	0.650

Note: The table reports Pearson correlation coefficients between each input measure and the composite measure, both overall and within each party. To calculate these correlations, we restrict the sample to candidates with scores from at least four of the input measures. This ensures that the correlations reflect variation in scores across multiple data sources, rather than candidates with only one or two scores.

12 Notes on Data Usage

Although I strive for completeness and accuracy in compiling the database, it is not guaranteed to be 100% comprehensive or without error. The donor and candidate IDs were assigned using an automated entity resolution framework that relies on probabilistic record-linkage algorithms. Researchers interested in collecting contribution records for a specific group of donors are encouraged to hand-check the IDs assigned by the identity resolution algorithm to screen for possible errors. This database is intended to provide the most accurate and complete information possible, while acknowledging the inherent complexity of the political data landscape. Corrections are welcome and can be submitted to bonica@stanford.edu. Duplicate contribution records are also known to occur in the database. This is often due to cross-reporting by one or more agency. For example, contributions to federal-connected 527s will at times be reported by the FEC under the transaction code 19. Contributions given through conduits also appear the FEC data twice with different recipients listed. The current release includes a column named `bk.ref.transaction.id` to aid in removing duplicated records.

Note also that estimated ideal points for candidates and contributors are based on varying amounts of data. For easy reference, the contributor and candidate/recipient files include columns indicating the number of contribution observations that were used to calculate each score. These files also indicate which donors/recipient were included in the scaling. In many cases, a donor may have only contributed to a single recipient and thus was excluded when estimating the ideal point model. Researchers should take this into account when deciding which donors to include in a study and may consider defining a minimum threshold for the number of contributions required for inclusion.

13 Tips for Working with Large Data Files

Due to the size of the contribution files, working with a statistical software package such as R is often necessary. The delimited CSV files will load into most spreadsheet oriented applications but will usually be truncated due to software limitations files with more than 1 million rows. Even when using a statistical software package, some of the contribution files from more recent election cycles can run up against memory limits on machines with less than 16GB of memory. Fortunately, there are several workarounds for accessing the data from disk or using publicly available cloud-based applications. One approach that

has worked well is to import the CSV files into a SQLite database. This can be done within R using the RSQLite package. Alternatively, there are several freely available SQLite applications with built-in support for working with large CSV files. Another user-friendly approach is Google Fusion Tables, a cloud-based platform designed for working with large data files. Google provides the service free-of-charge and provides detailed documentation for getting started and taking advantage of its more advanced features.

14 Reverse Compatibility With Earlier Releases

The database has undergone substantial changes and revisions from the original release. Improvements and other modifications were made to the identity resolution algorithms to take advantage of addition information and increased computational resources. As a result, the `bonica.cid` values for individual records may differ due to improvements in the record linkage algorithms. The `transaction.id` has remained unchanged and can be used as a crosswalk for tracking changes in *bonica.cid* values since the original release of the database.

References

Barber, Michael. 2022. "Comparing campaign finance and vote-based measures of ideology." *The Journal of Politics* 84(1): 613–619.

Blackwell, Matthew, James Honaker, and Gary King. 2017. "A Unified Approach to Measurement Error and Missing Data: Overview and Applications." *Sociological Methods and Research* 46(3): 303–341.

Bonica, Adam. 2014. "Mapping the ideological marketplace." *American Journal of Political Science* 58(2): 367–386.

Bonica, Adam. 2018. "Inferring roll-call scores from campaign contributions using supervised machine learning." *American Journal of Political Science* 62(4): 830–848.

Duck-Mayr, JBrandon, and Jacob Montgomery. 2023. "Ends against the middle: Measuring latent traits when opposites respond the same way for antithetical reasons." *Political Analysis* 31(4): 606–625.

Hall, Andrew B, and James M Snyder. 2013. "Candidate ideology and electoral success." *Working Paper* .

McCarty, Nolan, Keith T Poole, and Howard Rosenthal. 2016. *Polarized America: The dance of ideology and unequal riches*. MIT Press.

Shor, Boris, and Nolan McCarty. 2011. "The Ideological Mapping of American Legislatures." *American Political Science Review* 105(3): 530–551.

Tausanovitch, Chris, and Christopher Warshaw. 2017. "Estimating Candidates' Political Orientation in a Polarized Congress." *Political Analysis* 25(2): 167–187.