

## Index

- ACLED (recent event data set), 110  
 “action level,” and radon remediation, 298, 303  
 Affordable Care Act, 239–40  
 Africa, estimated population of U.S. citizens in, 258, 259, 260  
 Agence France-Presse (AFP), 99, 102  
 aggregations, of topic models, 67–9, 70  
 Agricultural Quarantine Inspection Monitoring (AQIM), 248–50  
 agriculture: and expressed priorities of legislators, 236; social marketing and government policies on invasive species, 247–53, 261–3  
 Airoldi, Edoardo M., 227  
 Akin, Todd, 240  
 Alexander, Rodney, 234  
 alignment, of topic models, 65  
 Allan, J., 110  
 alternative distance measures, of topic models, 76, 78  
 Alvarez, R. Michael, 11, 21n5  
 American National Election Study (ANES), 2, 28, 29, 44n3  
 American Political Science Association (APSA), 19, 137n1, 313  
*American Political Science Review* (journal), 1–2  
*Analytical Methods for Social Research* (journal), 1  
 Anandkumar, Anima, 90n33  
 anchor selection methods, and topic models, 90n30  
 Animal and Plant Health Inspection Service (APHIS), 247–53, 261–3  
 Annenberg School of Communications (University of Pennsylvania), 29  
 anonymity, as critical issue in computational social science, 13–14. *See also* privacy  
 Ansolabehere, Stephen, 31  
 Appropriations Committee (House of Representatives), 236–7  
 Arab Spring Revolutions (2011), 201, 212  
 archival depth, of political event data, 100  
 area-probability sampling, 30  
 Argentina, machine learning algorithms for detection of election fraud in, 275–91  
 Arora, Sanjeev, 83, 84, 85, 86, 90n30  
 audio, progress in data analysis of, 16  
 automated approaches, for review of topic models, 61–2  
 automated dictionary updates, and political event data, 102–105  
 Axelrod, Robert, 13  
 Bafumi, Joseph, 32, 42  
 bagging: and ensemble methods for machine learning algorithms, 274; and random forests, 173  
 Bagozzi, B. E., 106, 109  
 Ball, P., 108  
 ballot box stuffing (BBS), and detection of election fraud, 271, 277, 281, 285, 286, 292n10–11  
 Barbera, Pablo, 9

- Bayes classifiers, and machine learning algorithms, 270–2
- Bayesian model averaging (BMA), 258
- behavior change, and governmental social marketing initiatives, 252
- Beiler, John, 7, 310
- Belmont Report (1979), 13, 17, 18, 21n16
- Benjamin, Yoav, 169
- Bennet, Lance, 202
- Berico Technologies, 111
- Berinsky, Adam J., 40, 44n2
- Berkman, Michael B., 43
- betweenness centrality, of networks, 125, 126, 127
- Biau, Gérard, 189
- big data: and centralized analysis of local data on radon measurements, 299, 301; and concepts of “tall” and “fat” data, 247; costs of, 112; government review of implications, 261; implications of revolution in for future of computational social science, vii–x, 308; and increase in new sources of political data, 51; increasing prominence of in commercial and government space, 262–3; and political event data, 106; prevalence of multimodality in context of, 53; and relationship between social media and protest, 199–219; use of term, 296
- biomedical research, and random forests, 168–90
- Blaydes, Lisa, 229, 231
- Blei, David M., 88n12
- blogs. *See* political blogs
- Bond, Robert M., 137
- Bonneau, Richard, 9
- Bono-Mack, Mary, 240
- boosting, and ensemble methods for machine learning algorithms, 274
- bootstrap sample and bootstrapping, and random forests, 173–4, 175
- Bou-Hamad, Imad, 181
- Boulesteix, Anne-Laure, 175
- Brandt, Patrick T., 7, 101, 106
- Breiman, Leo, 171, 172, 173, 174, 176, 181, 189
- Bühlmann, Peter, 171
- Bush, George W., 69, 76
- California: and health policy data sets, 183–8; and 2010 Cooperative Congressional Election Survey, 153
- California Health Interview Survey (CHIS), 183–8
- CAMEO (event coding ontology), 108–109
- Cantor, Eric, 240
- Cantu, Francisco, 275
- Capuano, Michael, 234
- Carrat, Fabrice, 171
- Caughey, Devin, 39, 40, 45n8
- causal inference, and network analysis, 131–6
- censored data, and survival forests, 180
- Center for Plant Health Science and Technology (CPHST), 249
- centralized analysis, of local data on radon levels in homes, 295–306
- C5.0 algorithm, 292n4
- Chadefaux, T., 106
- Chang, Jeffrey S., 171
- Cheney, Richard B. (Dick), 64–78
- Chicago Tribune*, 304
- Cholesky decomposition, 164n3
- Christakis, Nicholas A., 131, 135
- Chuang, Jason, 78, 89n21
- Ciampi, A., 181
- C-index, and random forests, 182
- city effect, and public opinion on gay marriage, 32–3, 34, 35
- Clark, William Roberts, 5
- classification and regression trees (CARTs), 171–5
- CLIFF (event data set), 112
- Clinton, Hillary, 6–7, 16
- Clinton, Joshua D., 37, 42
- cloud-based file storage, 3
- cluster analysis, 250
- cluster sampling, 30
- coarse topics, and congressional press releases, 230, 231, 232–3, 234, 238, 241
- code-sharing, culture of, 3
- commercial surveys, 28–9, 31
- community detection, and network analysis, 128–31
- Computational Social Science Institute, 313
- computer science, and social science, 12, 58, 61, 82, 307, 308–14. *See also* Internet; software packages
- conditional variables, and random forests, 177
- confidentiality, as critical issue in computational social science, 13–14. *See also* privacy
- confusion matrices, 280, 282
- Congress. *See* House of Representatives
- Conn, Daniel, 8–9, 21n6, 177

- “connective action,” and social media, 202
- conservatives, asymmetric shift of toward  
Republican identification during 1980s, 40
- consistency, of political event data, 100
- construct validity, of MRP estimates, 36–7
- contextual variables, and identified anomalies  
in machine learning algorithms, 288–91
- convergent validity, of MRP estimates, 36
- convex models, 53–4
- Cooperative Congressional Election Study  
(CCES), 29, 31, 45n6, 45n10, 46n21,  
153–64
- Cordell, Heather J., 171
- correlated features, and random forests,  
176–80
- costs: and benefits of participating in protests,  
218; of political event data, 112–13. *See also* marginal cost
- CountryInfo.txt, 102–103
- covariate effect stability, of structural topic  
models, 71–6
- Cox proportional hazards model, 181
- credit claiming, and press releases of  
legislators, 238–9, 241
- cumulative hazard, and random forests, 182
- customization, of political event data, 113–14
- Data Access and Research Transparency  
(DART) initiative, 19
- data availability, general issues regarding,  
295–6
- “data-mining expeditions,” 255
- Dataverse project (Harvard University), 12,  
19, 21n18
- data visualization, need for improvements in,  
15
- Davenport, C., 108
- De Andres, Sara Alvarez, 171
- Debian (Linux), 311
- decision making, and policy for radon  
exposure levels in homes, 301–302, 303,  
304, 305
- decision tree, 272–3
- de-duplication, of political event data, 110
- degree centrality, of networks, 125, 126
- de Marchi, Scott, 13
- Democratic Party, and changes in expressed  
priorities of legislators after 2008  
presidential election, 225–42
- Denil, Misha, 189
- descriptive analysis, and public opinion  
surveys, 39–41
- Desouza, Kevin, 201
- Diaz-Urriarte, Ramón, 171
- Ding, Weicong, 90n34
- disaggregation, and small area estimation, 31
- discriminant function analysis, 250
- divide-and-conquer strategy, and tree-based  
classifiers, 272–3
- “dorm room” experiments, and experimental  
approaches to network analysis, 133
- dose-response relationship, for radon decay,  
298
- dual-degree programs, in computational social  
science, 314
- Duke University, 2
- economics, measurements of development as  
example of revolution in big data, ix. *See also* costs; political fundraising  
education, and future of computational social  
science, 314
- “Effect of Registration Laws on Voter  
Turnout, The” (Rosenstone and Wolfinger  
1978), 1–2
- Egypt, social media and political protest in,  
203, 206
- eigenvector centrality, of networks, 127
- Elbow Method, 269
- EL:DIABLO coding system, 99, 102,  
114
- “election forensics,” growing literature on,  
266
- elections: machine learning algorithms for  
detecting fraud in, 267–91; and salience in  
statistical models of voter choices, 140–64;  
topic model analysis of political blogs  
written during 2008 presidential, 63–78;  
and voter assistance programs for overseas  
citizens, 253–63. *See also* political science;  
voters and voting
- Electronic Frontier Foundation, 117
- Elmendorf, Christopher S., 41
- ensemble Bayesian model averaging (EBMA),  
255–6
- ensemble methods, and machine learning  
algorithms, 274
- ensemble model averaging (EMA), 256, 258
- Environmental Protection Agency (EPA),  
298–9, 303
- ethics, as important issue in future of  
computational science, 17–18, 19
- Euclidean distance, 269–70
- Euromaidan protests (Ukraine), 202, 203

- Europe, estimated population of U.S. citizens in, 258, 259, 260
- Evans, Sarah, 10
- exclusivity, of topic models, 62
- exercise, research on as example of revolution in big data, viii–ix
- experimental approaches, to network analysis, 133–4
- exponential random graph models (ERGMs), 135–6
- expressed priorities, changes in legislators' after 2008 presidential election, 225–42
- Facebook, 18, 200, 205–19, 221n17, 221n19, 222n26, 296, 312. *See also* social media failure time, and survival forests, 180, 181 fat data, 247, 261–2
- Federal Voting Assistance Program (FVAP), 253–63
- Fenno, F. Richard, Jr., 10
- Fernandez de Kirchner, Cristina, 275
- 15M protests (Spain), 202
- Firearms Control Regulations Act of 1975, 72
- first-stage models, and public opinion surveys, 36
- flexibility, of political event data, 100
- Florence (Italy), and Medici family in 1400s, 122, 125, 126, 136
- Florida, and radon levels in homes, 304
- Foley, Peter, 8
- Foreign Government Estimate (FGE), 254
- Fors Marsh Group, 10
- Foster, Dean P., 90n33
- Fowler, James H., 128, 131, 135
- Framingham Heart Study, 131–3
- Freeman, J. R., 106
- Frente para la Victoria* (FPV) party, 275, 277, 281–90, 292n6, 292n10–11
- Frente Progresista Civico y Social* (FPCS), 276, 278, 281, 287, 288, 292n6
- Frente Renovador* (FR) party, 275, 277, 281–90, 292n6, 292n10–11
- friendship networks, measurement of as example of revolution in big data, ix. *See also* social networks
- Fromentin, Remi, 177
- fuzzy forests, 177–80
- GDELT (recent event data set), 110, 112
- Ge, Rong, 83, 84, 85, 86, 90n30, 90n33
- Gelman, Andrew, 11–12, 32, 36
- General Social Survey (GSS), 43, 44n3
- genomics, and random forests, 170
- geolocation: and political event data, 110–12; of social media posts, 204, 212–13, 221n20–1. *See also* location information
- Gerber, A. S., 140
- Getoor, L., 103
- Geziparkidireni (Facebook page), 210
- Gezi Park protests (Turkey), 202–203, 210
- Ghitza, Yair, 36
- Gillis, Nicolas, 90n30
- Gini importance, 175
- Gini index, 172, 174
- Girvan, M., 129
- Github, 4, 113, 205
- Gladwell, Malcolm, 9, 201–202
- Glasgow, G., 141
- global alignment, of topic models, 65
- global warming, and political blogs, 64–78
- Golder, Matt, 5
- Google, 296
- Gordon, L., 181
- government. *See* Environmental Protection Agency; House of Representatives; political science; public policy; states; Supreme Court; U.S. Census Bureau; U.S. Department of Agriculture; U.S. Geological Survey
- Granovetter, Mark, 201
- granular topics, and congressional press releases, 230, 231, 232–3, 234, 238, 241
- greedy learners, tree- and rule-based classifiers as, 274
- Griepentrog, Brian, 10
- Grimmer, Justin, 10, 89–90n22, 225, 226, 227, 229, 230, 231, 236, 237–8, 242, 309
- Guillory, Jamie E., 18
- Hale, Henry, 201
- Hall, Ralph, 239
- Hall, Thad E., 21n1
- Halpern, Yonatan, 83, 84, 85, 90n30
- Halterman, Andrew, 7
- Hamiltonian Monte Carlo, 153, 155
- Hancock, Jeffrey T., 18
- Hand, D. J., 301
- Hapfelmeier, Alexander, 171
- Harbagiu, S., 110
- Harvard University, 19, 21n18
- Hastie, Trevor, 173, 189
- Hawaii, and invasive species prevention, 249

## Index

321

- Heagerty, Patrick J., 182
- health care: expressed priorities of legislators on reform of, 239; and network analysis of outcomes, 131–3; random forests and policy data sets on, 183–8. *See also* medicine; public health
- Heller v. District of Columbia*, 72
- Herron, Michael C., 42
- Herseth, Stephanie, 234
- Hersh, Eitan, 137
- hierarchical clustering, 250
- hierarchical modeling, of salience matrices, 149–51
- Higgins, Brian, 234
- high-dimensional regression methods, and biomedical research, 169
- Hill, Seth, 42
- Hinich, M. J., 140, 142
- historical surveys, of public opinion, 40
- Hochberg, Yosef, 169
- homophily, and experimental approaches to network analysis, 133
- Hothorn, Torsten, 181, 182, 183
- House of Representatives: and network study on cosponsorship, 128; study of changes in expressed priorities of after 2008 presidential election, 225–42; study of links between public opinion and political ideology in representation, 41–3
- housing, and radon exposure levels, 295–306
- Hsu, Daniel J., 90n33
- Huang, Furong, 90n31
- Human Rights Campaign, 45n12
- Hungarian algorithm, 89n18
- Hyde, Susan D., 21n5
- hyperplanes, and salience, 143–4, 145, 147, 166
- IDEA (event coding ontology), 108
- identified anomalies, and contextual variables in machine learning algorithms, 288–91
- ideology, and concept of salience, 141. *See also* political ideology
- images, and progress in data analysis of, 16
- Indignados movement (Spain), 201, 202
- information: machine learning algorithms and gain in, 292n4; participation in protests and access to, 218, 219; topic models and visualization of, 61
- initialization, of topic models, 78–82, 85
- Institutional Review Boards (IRBs), 18
- Integrated Crisis Early Warning System, 107
- interdisciplinary programs, and education in computational social science, 314
- Internet: and access to data sources, 3, 31, 99, 101, 102, 116; and changes in methodology of applied social science research, 3. *See also* political blogs; social media; websites
- inter-ocular test approach, to network analysis, 130
- Interuniversity Consortium for Social and Political Research (ICSPR), 2, 19, 101
- Iran: and political blogs in 2008, 64–78; and protests in 2009, 201
- Iraq War, 234, 235, 236
- IRT models. *See* item response theory
- Ishwar, Prakash, 90n34
- Ishwaran, Hemant, 181, 183
- Italy, and Medici family of Florence in 1400s, 122, 125, 126, 136
- item response theory (IRT), 37–9, 88n10, 142, 143–5, 147, 153, 166
- Jackman, Simon, 37, 46n20
- Jackson, Matthew O., 16
- Jags (software), 32, 37
- Jeon, Yongho, 189
- Job Control Language (JCL), 2
- Jost, John, 9
- journals, and publication of research papers in computational social science, 19, 21n12
- Kahneman, D., 112
- Kakade, Sham M., 90n33
- Kannan, Ravindran, 84, 86
- Kansas Event Data System (KEDS), 7
- Kaplan-Meier estimator, 180, 183
- karate club, as example of network analysis, 127–8, 129–30, 137n4
- Kastellec, Jonathan P., 42, 45n9
- k* clusters, and electoral variables, 268–71, 281
- King, Gary, 5, 89–90n22, 99
- k* means algorithms: and machine learning approaches, 268–70; and topic models, 79–80
- k*-nearest neighbors (kNN) classifier, 270–1
- Kogalur, Udaya B., 181
- Koltcov, Sergei, 78, 88n12
- Koltsova, Olessia, 78, 88n12
- Kosorok, Michael R., 181
- Kramer, Adam D., 18
- Kumaran, G., 110

- Lancichinetti, Andrea, 88n12
- languages: and analysis of political event data, 105–106; and social media use, 204, 214–17, 217, 221n22
- LASSO, 53, 88n6, 169, 181
- latent Dirichlet allocation (LDA), and analysis of political blogs during 2008 presidential election, 51, 55–8, 62, 80–2, 83–4
- latent variables, and surveys of public opinion, 37–9
- Lauderdale, B. E., 141
- Lax, Jeffrey R., 41, 42, 45n9
- Lazer, David, 15
- lazy learners, and k-nearest neighbor classifiers, 270
- learning algorithms. *See* machine learning algorithms
- Levin, Ines, 11, 21n6
- Lewis, Jeffrey B., 38–9
- Lewis, Jerry, 234
- Li, Wei, 10
- liberalism, and changes in public policy at state level, 39, 40, 41
- Lightbox Analytics, 254
- Likert response categories, 30
- Lin, Yi, 189
- Linguistic Data Consortium, 116
- Literary Digest, 44n1
- Little, Thomas C., 32
- Liu, Yi-kai, 90n33
- local alignment, of topic models, 65, 75
- local data, and centralized analysis of radon exposure levels in homes, 295–306
- location information, on radon measurements in homes, 299–300. *See also* geolocation
- Lockheed-Martin, 107
- logit model, 1–2
- Lunetta, Kathryn L., 171
- lung cancer, and radon exposure, 297, 298
- Lysenko Volodymyr, 201
- Machanavajjhala, A., 103
- machine-coded political event data, 100–101
- machine learning algorithms, for detection of voter fraud, 267–91
- machine-readable texts, 99
- Mansiaux, Yohann, 171
- manual review, of reference models, 60–1
- Map-Reduce paradigm, 90n24
- marginal cost, of generation and analysis of political event data, 101–102. *See also* costs
- Marsh, Sean, 10
- Marubini, E., 181
- Masket, Seth, 137n2
- Massa, Sergio, 275
- McCallum, Andrew, 10
- McClurg, Scott, 121
- McQueen, Alison, 229, 231
- mean squared error (MSE), 257
- Medici family, in Florence, Italy during 1400s, 122, 125, 126, 136
- medicine, data analysis in evidence-based, 296. *See also* biomedical research; health care
- Messing, Solomon, 225, 236, 238, 242
- Metzger, Megan, 9
- Middle East, estimated population of U.S. citizens in, 258, 259, 260
- Minnick, B., 305
- MIT Center for Civic Media, 112
- mixture models, 54–5
- modeling approaches, to network analysis, 134–6
- model-level aggregations, of topic models, 67–9
- Moitra, Ankur, 84, 86
- Moldova, and protests, 201
- Molinaro, Annette M., 183
- “MoneyBombs” visualization, 15
- Moore’s law, ix
- Morozov, Evgeny, 201, 220n4
- motivation, for participation in protests, 218
- multidisciplinary research: and future of computational social science, 313–14; increased emphasis on and acceptance of in social science, 4
- multilevel regression and poststratification (MRP), 32–7
- multimodality, and topic models, 52–7, 86–7
- multinomial logit, 144
- Munger, M. C., 140, 142
- Municipal Equality Index (MEI), 45n12
- Nagler, Jonathan, 9
- “naïve” Bayes classifiers, 271
- Nall, Clayton, 137
- named entity recognition (NER), 103, 110
- National Health Interview Survey, 21n8
- National Opinion Research Center (NORC), 254
- National Science Foundation, 102
- natural resources, and government policies on invasive species, 247–53

## Index

323

- near-duplicate detection, and political event data, 110
- near-real-time coding, by machine-based systems, 100–101
- nested topics, and study of congressional press releases, 229–31
- network community, 129
- Network Data, 138n5
- networks: and causal inference, 131–6; definition of, 123–4; estimation of, 124–31; need for better approaches for statistical analysis of, 16–17; reasons for study of in social science, 121–3. *See also* friendship networks; political networks; social networks
- neural network models, 88n10
- New Deal, 40
- “New Directions in Text as Data” conference, 313
- new event detection (NED), 109–10
- Newman, M. E. J., 129
- New York, radon mapping programs and information campaigns in, 305
- New York Times*, 110, 303
- Nicodemus, Kristin K., 176
- Nikolenko, Sergey, 78, 88n12
- 9/11 disaster, and networks, 122
- NIPS (machine learning conference), 313
- noise parameterization, and salience, 144–5, 147–9
- nonconvex models, and spectral learning, 82
- nondeterministic polynomial-time hard (NP-hard) problems, 58, 79, 89n14
- non-negative matrix factorization (NMF), 83–4, 90n29
- North Korea, and political blogs, 64–78
- nuclear weapons, and political blogs, 64–78
- Obama, Barack, 64–78, 225–42, 261
- O’Brien, S. P., 101, 107
- Occupy Movement, 201
- O’Connor, B., 113
- Ogburn, E. L., 133
- Oishen, R. A., 181
- OneR rule learner, 274, 292n11
- Online Appendix, 220n10, 222n23
- ontologies, limits of for political event data, 108–10
- Onuch, Olga, 217
- Open Event Data Alliance (OEDA), 7, 99, 109, 112–18
- open-source software, 3–4
- ordinal logit model, 152, 153
- ordinal probit, 152
- Oregon, and duplicate records in voter registration lists, 14
- Organization for Economic Cooperation and Development (OECD) International Migration Database, 254
- “out of bag” (OOB) samples, and random forests, 173–4, 182
- Pacheco, Julianna, 41
- pachinko allocation models, 226
- Page, Scott E., 13
- pairwise similarity, of topic models, 65–7
- Park, David K., 32
- Penfold-Brown, Duncan, 9
- Pennsylvania State University, 111
- “perfect-information action level,” 302
- Perl (software), 4
- Peronist Party (Argentina), 275
- perpendicular shift invariance, and salience, 146
- personally identifiable information (PII), 13, 248
- Phillips, Justin H., 41, 42, 45n9
- phone surveys, rate of nonresponse to, 31
- Plutzer, Eric, 43
- policy space, and item response theory, 143. *See also* public policy
- Polimetrix, 153
- Political Analysis* (journal), 1, 4, 14, 15, 19
- political blogs, use of topic models for analysis of, 63–78
- political event data, generation of in near real time, 98–118
- political fundraising, and development of quantitative tools to study networks, 17
- political ideology: measurement of as variable in public opinion surveys, 37–9; role of in spatial model of voting, 20n4; and influence of same-party constituents on roll-call behavior of legislators, 41–3. *See also* conservatives; Democratic Party; protest; Republican Party
- political networks, 124
- political parties, salience analysis and strategies of, 162–4. *See also* Democratic Party; *Frente para la Victoria*; *Frente Progresista Civico y Social*; *Frente Renovador*; Peronist Party; Republican Party; *Unidos por la Libertad y el Trabajo*

- political protest. *See* protest movements
- political representation, study of links between public opinion and political outcomes at federal level, 41–3
- political science: and concept of salience, 140; and growth in prominence of statistical topic models, 57; reasons for focus on in discussion of new computational methodologies, 5. *See also* elections; House of Representatives; political ideology; political parties; public opinion; voters and voting
- Pomares, Julia, 11
- Popkin, Samuel, 13
- Porter, Mason A., 129
- presentational styles, of members of Congress, 237
- press releases, by legislators after 2008 presidential election, 225–42
- Price, Phillip, 11–12
- Price, Tom, 240
- privacy: as critical issue in computational social science, 13–14; and prisoner's dilemma, 297
- privacy impact analysis (PIA), 249
- probit model, 1–2
- Proceedings of the National Academy of Sciences* (PNAS), 18. *See also* Facebook
- professional societies, and important initiatives in computational social science, 19
- property values, and radon levels in homes, 304
- proteomics, and random forests, 170
- protest movements, and social media, 9, 199–219
- PS: Political Science & Politics* (symposium), 5
- publication: online forms of, 99; of research papers in computational social science, 19, 21n12, 313–14
- public health: and analysis of local data on home radon levels, 295–306; and examples of revolution in big data, ix
- public opinion: application of big data in study of, 28–30, 43; future of research on, 43–4; measurement of ideology and other latent variables in, 37–9; relationship between legislators' roll-call behavior and, 41–3; sampling design for survey of, 30–1; and small area estimation, 31–7
- public policy: examination of linked data sets as source of important information for, 14; liberalism and changes in at state level, 39, 40; on radon levels in homes, 295–306; and prevention of risk to agriculture and natural resources from invasive species, 247–53; random forests and health policy data sets, 183–8; reasons for focus on in discussion of new computational methodologies, 5
- Puerto Rico, and inspections for invasive species, 249
- Putnam, Adam, 239
- Python (software), 4, 9, 102, 205
- quantification, of networks, 122, 125–8
- Quarterly Journal of Political Science* (journal), 19
- Quinn, Kevin M., 89n15, 231
- quota sampling, 30
- racism, and levels of anti-black stereotyping by states, 41
- radon exposure levels, measurement of in homes and policy decisions, 295–306
- Ramirez, Christina, 8–9, 21n6
- random-digit dialing (RDD), 31
- random forests (RFs): applied to biomedical research, 168–90; classification and regression trees, 172–4; and fuzzy forests, 177–9; and health policy data sets, 183–8; and machine learning algorithms, 274, 282, 285, 290; and robustness testing, 258; and survival forests, 179–83; and variable importance measures, 174–7, 185
- Random Jungle, 170
- Raskutti, Garvesh, 169
- Reeves, A., 108
- reference model, and topic models, 60–2, 75
- regression trees, 172–5
- reliability, of political event data, 100
- remediation, and radon exposure levels in homes, 298, 303
- remote sensing, of political entities, 99
- replication: and automated coding efforts generating big data, 113, 118; and policies of scholarly journals, 14, 19, 101
- representation. *See* political representation
- Republican Party: asymmetric shift of conservatives to identification with in 1980s, 40; and changes in expressed priorities of legislators after 2008 presidential election, 225–42
- Reuters (news agency), 99, 102
- ridge regression, 174, 181

## Index

325

- RIPPER (Repeated Incremental Pruning to Produce Error Reduction), 274, 282, 285
- Rivers, Douglas, 37, 141
- Roberts, K., 110
- Roberts, Margaret E., 6–7, 62, 89n19–20, 227, 310–11
- robustness testing, 258
- Rodden, Jonathan, 45n9
- Rohban, Mohammad H., 90n34
- Roper Center for Public Opinion, 28, 30
- Rosenblatt, A., 140
- Rosenstone, Steven J., 1–2, 20n1
- R package psc1 (software), 37, 182
- rule-based classifiers, and machine learning algorithms, 273–4
- Saiegh, Sebastien M., 275
- salience, and statistical models of political ideology and voter choice, 140–64. *See also* political ideology
- Saligrama, Venkatesh, 90n34
- same-sex marriage, and public opinion surveys, 32–3, 34, 35, 45n11
- sampling design, for public opinion surveys, 30–1
- Santos Silva, J. M. C., 256
- SCAD, 170, 181
- Schaffner, Brian F., 31
- Schickler, Eric, 40
- scholastic testing, 296
- Schrodt, Philip A., 7, 109
- Scornet, Erwan, 189
- seeding strategy, and  $k$ -means algorithms, 80
- Seegerberg, Alexandra, 202
- semantic coherence, of topic models, 62
- semi-automated analysis, of topic models, 61
- sensitivity analysis, and networks, 132–3
- separability condition, and non-negative matrix factorization, 84
- separate-and-conquer strategy, and rule-based classifiers, 273–4
- Shalizi, Cosma, 132
- sharing, of political event data, 100
- simple random sample (SRS), 31
- Simpson, Erin, 7
- Sinclair, Betsy, 7–8, 134
- single nucleotide polymorphism (SNP), 168
- “slacktivism,” 201
- small area estimation, 31–7
- Smith, Adam, 246
- Snijders, Tom A. B., 16
- social marketing: and policies on protection of agriculture and natural resources from invasive species, 247–53, 261–3; use of term, 263–4n1; and voting assistance programs for overseas citizens, 253–63
- social media: definition of, 220n7; as example of revolution in big data, viii; and political protest, 9, 199–219; rapid rise in use of, 199. *See also* Facebook; Twitter
- Social Media and Political Participation Laboratory (NYU), 9
- social networks, 124, 219. *See also* friendship networks
- social psychology, and participation in protests, 219
- social science: changes in methodology of applied research in, 1–4; data analysis and theory in, 12–13; future of computational, 307–14; and network analysis, 121–37; and revolution in big data, vii–x; and topic models for study of large collections of texts, 227–9
- software packages, development, adaptation, and use of, 2, 4, 19, 32, 37, 101–102, 111–17, 171, 262, 303
- Sola Pool, Ithiel de, 13
- Southeast Asia, estimated population of U.S. citizens in, 258, 259, 260
- Spain, social media and protests in, 201, 202
- specialized online tools, for data analysis, 296
- spectral learning, 82–3, 90n28
- Spencer, Douglas M., 41
- stability, of topic models, 62–3, 73–5
- Stan (Stan Development Team 2013), 32, 37, 152, 153, 155
- Stanford CoreNLP system, 105, 113
- Stanford Large Network Data set Collection, 138n5
- Stanford University, 111
- states: effects of on public opinion surveys on gay marriage, 33; and levels of anti-black racial stereotyping, 41; liberalism and changes in public policy, 39, 40, 41. *See also specific states*
- Stekhoven, Daniel J., 171
- Stewart, Brandon M., 6–7, 227, 236
- Storey, John D., 169
- Strobl, Carolin, 176, 177
- structural topic model (STM), 52, 63–78, 227
- Stupak, Bart, 234
- subsampling, and bootstrapping of random forests, 175–6

- supervised learning, and machine learning algorithms, 282–8
- support vector machines (SVMs), 170
- Supreme Court, and political blogs, 64–78
- survey data, on social media use and protests, 216, 218
- survival forests, and random forests, 180–3
- Svetnik, Vladimir, 171
- Sweeney, Latanya, 13
- TABARI system (National Science Foundation), 19, 102, 113, 118n6
- TaksimDayanismasi (Facebook page), 210
- tall data, 247, 261, 262
- Tausanovitch, Chris, 38–9, 43, 45n11, 46n21
- Tcherger, E. J., 133
- Tea Party movement, and changes in expressed priorities of legislators after 2008 presidential election, 225–42
- Telgarsky, Matus, 90n33
- Tenreiro, S., 256
- term frequency-inverse document frequency (TF-IDF), 109–10
- “Text as Data” conference, 313
- text as data method, for analysis of legislators’ press releases after 2008 presidential election, 225–42
- theory, and data analysis in social science, 12–13
- Thomas, Andrew C., 132
- Tingley, Dustin, 6–7
- Topic Detection and Tracking (TDT) initiative, 109
- topic-level aggregations, of topic models, 69, 70
- topic models: and analysis of political blogs, 63–78; assessing stability of, 62–3; definition of, 6; and evaluation of local modes, 58–60; finding reference model for, 60–2; and global solutions, 82–6; initialization of, 78–82; and multimodality, 52–7, 86–7; and study of large collections of texts, 227–9
- transparency: as important issue in future of computational science, 17, 18–19; and political event data, 100, 112
- Transue, J. E., 140, 141
- tree-based classifiers, and machine learning algorithms, 272–3
- Trounstine, Jessica, 42
- Tucker, Joshua, 9, 309–10
- Tufte, Edward, 15, 21n10
- Turkey, social media and protests in, 202–203, 205–19
- Turner, Sidney Carl, 10
- Twitter, 107, 200, 201, 204–219, 220n11, 221n22, 222n24–5. *See also* social media two-step clustering, 250
- Ubuntu (Linux), 311
- Ukraine: social media and protests in, 203, 205–19; voter fraud and call for new elections in 2003, 291
- uncertainty, and public opinion estimates, 36
- Unidos por la Libertad y el Trabajo* (ULT), 276, 278, 287, 292n6
- Uniformed and Overseas Citizen Absentee Voting Act (UOCAVA), 258
- U.S. Census Bureau, 254, 296
- U.S. Department of Agriculture (USDA), 247–53, 261–3
- U.S. Geological Survey, 298
- University of Chicago, 254
- University of Kansas, 101
- University of Massachusetts, 313
- University of Pennsylvania, 29
- unsupervised learning, and machine learning algorithms, 279–82
- urban dictionary (online), 201
- Vagrant (software), 114
- validation, of political event data, 115
- validity, of political event data, 100
- Van De Geer, Sara A., 169
- VanderWeele, T. J., 132–3
- variable importance measures (VIMs), and random forests, 174–7, 184
- Verikas, Antanas, 189
- versioning, of political event data, 113–14
- video, and progress in analysis of data, 16
- visualization, of networks, 122, 124–5, 137n3
- Vitter, David, 234
- vKontakte (Russian-language social media site), 217
- vote fraud. *See* ballot box stuffing; vote stealing
- voters and voting: and assistance programs for overseas U.S. citizens, 253–63; and identification of duplicate records in registration lists, 14; and machine learning

*Index*

327

- algorithms for detection of fraud, 267–91;  
 salience and statistical models of voter  
 choices, 140–64. *See also* elections  
 vote stealing, detection of as form of election  
 fraud, 271, 277, 279–80, 281, 285, 286,  
 288, 292n10–11
- Wallach, Hanna, 12, 228  
 Wang, Wei, 29, 45n8  
 Warsaw, Christopher, 5–6, 39, 43, 45n9,  
 45n11, 46n21, 310  
 Washington, and duplicate records in voter  
 registration lists, 14  
 Watts, Duncan, 307  
 websites, and data analysis, 296. *See also*  
 Internet  
 weighted genetic co-expression network  
 analysis (WGCNA), 177–8
- Westwood, Sean J., 225, 236, 238, 242  
 Wilkins, Arjun S., 40  
 Willows (random forest), 170  
 Wilson, Christopher, 222n24  
 Wolfinger, Raymond E., 1–2, 20n1  
 World Health Organization, 184  
 Wright, Jeremiah, 64–78
- Yanukovych, Viktor, 203  
 Yatsenyuk, Arseniy, 206, 207  
 Yonamine, J., 106  
 Young, J. K., 121  
 YouTube, 16
- Zeitsoff, T., 107  
 ZeroR, 274  
 Zheng, Yingye, 182  
 Zhu, Ruoqing, 181