JAE YEON KIM

Overview

- Computational social scientist: using data science to study the politics of diversity and inclusion
- Research software developer: building tools that make digital data collection easier and faster
- Experience: analyzing survey, experimental, administrative, and text data using statistical and machine learning methods

EDUCATION

2016 | Present
University of California, Berkeley
PhD Candidate in Political Science — Berkeley, California, USA

Summer 2019
Summer Institute in Computational Social Science
Participant (10% acceptance rate)
Princeton University, Princeton, USA

2014 | 2016
University of California, Berkeley
MA in Political Science — Berkeley, California, USA

2012
Korea University
BA in Political Science and English — Seoul, South Korea

PROFESSIONAL EXPERIENCE

May 2019 | Present
Senior Data Science Fellow, Instructor, and Statistical Consultant
Data-intensive Social Sciences Lab — UC Berkeley
- Consulted 60+ Berkeley faculty, students, and staff on applied statistics, machine learning, and database management
- Developed five original data science workshops (SQL for R Users, R Package Development, Functional Programming in R, Advanced Data Wrangling in R, and Reproducible Project Management in R)
- Taught programming fundamentals, data wrangling, machine learning, and project management
- Organized and streamlined consulting and workshop production

Fall 2020 | Present
Visiting Fellow
P3 Lab, SNF Agora Institute — Johns Hopkins University,
- Working with Milan de Vries (former Director of Analytics at MoveOn.Org) to build a data infrastructure on civic organizations in the US and their relationships with food security, polarization, and the 2020 racial justice movement

Last updated on 2020-09-24.

CONTACT INFO

jaeyeonkim@berkeley.edu

jaeyk.github.io

github.com/jaeyk

linkedin.com/jae-yeon-kim

+1 510-646-5183

For more information, please contact me via email.

COURSEWORK

Statistical and Causal Inference, Experimental Design, Survey Methods, Game Theory, Computational Social Science

Passed Political Behavior (social and cognitive psychology, survey and experimental design) field exam with distinction

SKILLS

Quantitative: Statistical and causal inference, Experimental and survey design

Computational: Natural language processing, Machine learning, R (tidyverse, tidymodels, statistical packages), Python (pandas, scikit-learn), Git, SQL (PostgreSQL), NoSQL (MongoDB), Linux Command Line

Spring 2020

Data Science Education Program Fellow
Data Science Education Program — UC Berkeley

- Served as research lead for the undergraduate students and project partners involved in 40+ data science discovery projects
- Taught original workshops on project management, computational reproducibility, bias in machine learning, and data visualization
- Published an article on project management in SAGE Ocean, an initiative from SAGE Publishing focusing on computational social science

Fall 2016 | Present

Graduate student instructor
Department of Political Science — UC Berkeley

- Developed and taught original graduate-level courses on computational tools for social science research as lead instructor and digital data collection as co-instructor
- Taught an undergraduate-level applied statistics as a teaching assistant and received the outstanding graduate student instructor award, which is given to less than 10% of Berkeley TAs

SOFTWARE

tidytweetjson: R package for turning Tweet JSON files into a cleaned and wrangled dataset. The package takes average 5 seconds to turn 100 tweets into a tidy dataframe.

tidyethnicnews: R package for turning search results from one of the largest databases on ethnic newspapers and magazines published in the United States into a cleaned and wrangled dataset. The package takes average 0.0005 seconds to turn 1,000 articles into a tidy dataframe.

RESEARCH EXPERIENCE

Summer 2020 | Present

Large-scale Twitter Analysis on COVID-19 and Anti-Asian Climate [GitHub]
PhD Candidate — UC Berkeley

- Developed an R package that automates parsing a large Tweet JSON file (>5GB) into a cleaned and wrangled dataset
- Applied dynamic topic modeling to 1.4 million tweets and traced the rise of anti-Asian sentiment in the post-pandemic US
- Presented at the 2020 American Political Science Association annual meeting

Spring 2020 | Present

Intersectional Bias in Hate Speech and Abusive Language Detection Datasets [GitHub] [Preprint] [Slides]
PhD Candidate — UC Berkeley

- Classified gender, racial, and party identities of the 100k tweets
- Demonstrated African American tweets were up to 3.7 times more likely to be labeled as abusive, and African American male tweets were up to 77% more likely to be labeled as hateful compared to the others
- Published the paper version in *Proceedings of the Fourteenth International Conference on Web and Social Media (ICWSM), Data Challenge Workshop*

Fellowships: Democracy Visiting Fellowship, Ash Center for Democratic Governance and Innovation, Kennedy School, Harvard University (2020 - 2021, declined), D-Lab Data Science Fellowship, UC Berkeley (2020), Data Science Education Program Fellowship, UC Berkeley (2020), American Democracy Project Fellowship, UC Berkeley (2019), California Poverty and Socioeconomic Inequality Fellows Program, the Blum Initiative for Global and Regional Poverty Studies (2017), Berkeley Empirical Legal Studies Graduate Fellowship, Center for the Study of Law and Society, UC Berkeley (2017)

Awards: Don T. Nakanishi Award for Distinguished Scholarship and Service in Asian Pacific American Politics, Western Political Science Association (2020), Outstanding Graduate Student Instructor Award, UC Berkeley (2016)

Fall 2019 | Present

Causal Inference and Machine Learning [GitHub] [Preprint] [Slides]

PhD Candidate — UC Berkeley

- Demonstrated how machine learning assists causal inference by combining text classification and interrupted time series design
- Presented at the joint Political Computational Social Science and Political Network 2020 Conference and the Berkeley Computational Social Science Forum
- Authored a preprint

2018 | Spring 2019

Natural Language Processing and Machine Learning [GitHub] [Preprint] [Slides]

PhD Candidate — UC Berkeley

- Developed an R package that automates parsing unstructured HTML files into a cleaned and wrangled dataset
- Demonstrated unreliable training data generates weak predictions and extreme interpretations using 80k+ historical newspaper articles
- Received the Best Paper Award in Asian Pacific American Politics from the Western Political Science Association (2020)
- Authored a preprint, which was conditionally accepted at the *Journal of Computational Social Science*

2016 | 2018

Statistical Modeling of Time Series Data [GitHub] [Preprint]

PhD Candidate — UC Berkeley

- Examined how social policy influenced community organizing among Asian Americans and Latinos by creating an original organizational dataset and modeling time-series data
- Authored a preprint, which was invited to Revise and Resubmit at *Political Research Quarterly*

2019 | Present

Survey and Experimental Research [GitHub] [Preprint]

PhD Candidate — UC Berkeley

- Designed a within-subject experiment and embedded it in a California-wide survey to investigate how different racial groups interpret questions on racial solidarity differently
- Authored a preprint

Summer 2018

Survey Research [GitHub]

Graduate Student Researcher — UC Berkeley

- Cleaned and wrangled the largest panel survey data on Asian Americans and conducted factor and regression analysis

ORGANIZING EXPERIENCE

Summer Institute in Computational Social Science in the San Francisco Bay Area

Co-organizer

August 2019 - July 2020

- Raised 50k+, reviewed 100+ applicants and selected 20 participants
- Developed close partnerships with nonprofits (e.g., Code for America, DonorsChoose, Hopelab)
- Designed the curriculum, guided the project development and developed the evaluation criteria
- Published a blog post on the *Berkeley Institute of Data Science* website that highlights the key accomplishments of the Summer Program