UC Center Sacramento
Policy Brief
Volume 03, Issue 11 • October 30, 2019

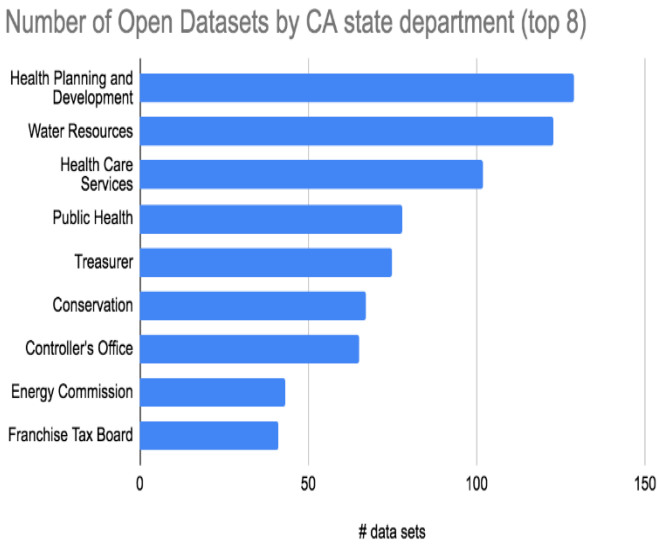
Leveraging Data Science for California Policy

Karen Chapple, Chair and Professor of City and Regional Planning, UC-Berkeley
Amelia Baum, UC-Berkeley

Big data and data science are all the rage in the private sector, but the public sector is a late adopter. Although local and state governments increasingly make their data publicly accessible, policymaking rarely uses the big data and analytics that now drive how businesses make decisions about markets and management. This is a lost opportunity to make government agencies not just more innovative and efficient but more responsive to their publics, thereby strengthening our democracy.

Big data is data of unprecedented volume, often at real-time speed and crowdsourced directly from users, typically analyzed with algorithms rather than traditional inferential statistics, that shed new light on the scale, time, interaction, and content of phenomena. Data science typically uses *found* data sources, rather than the *designed* data – primary or secondary data that is collected according to a subjective framework -- that has underpinned much urban analysis to date. They then use objective algorithms for analysis via machine learning, rather than regression models with carefully selected variables, in the process removing some bias.

Access to organized, comprehensive data sources on a range of topics is a requirement for data science methods in the public sector. Just 17% of California cities[1] and 15 out of 58 counties maintain their own open data portals.[2] The open data landscape is uneven, with far more data available in urban parts of the state. As part of an Open Data policy launched in March 2019, the state of California maintains an open data portal (ca.data.gov) at the state level, which hosts nearly 2200 datasets from 42 state departments. Environmental,. As shown below, health and financial data are the most well-represented sources.



Interactive data visualization and web tools have gained popularity with several public agencies, with tools like Tableau and ArcGIS Online providing low barriers to entry for agencies working with public-facing data projects for the first time. For instance, California's Housing and Community Development Department provides an interactive map that allows users to check the progress that individual jurisdictions are making towards their housing goals.

Yet, the public sector falls short in other ways. For business, enabling widespread access to dynamic data has been critical in spurring innovation; but the use of application programming interfaces (APIs) remains rare among agencies. Creating more dynamic infrastructure is an essential next step for agencies to modernize their data pipelines and keep information widely accessible and up to date.

[1]Cities and Census Designated Places with populations over 30,000.
[2]Jurisdictions that maintain data portals with data with content *other* than GIS shapefiles and zoning maps

Likewise, machine learning remains virtually absent from the purview of public agencies in California, likely in part to ethical concerns and insufficient capacity. Yet, machine learning methods can provide more accurate ways of predicting outcomes that allow agencies to distribute resources more effectively. Recent work from educational and research institutions has demonstrated machine learning approaches to policy questions relevant to public agencies, especially in the context of water access, energy usage, and transportation planning. For example, the San Francisco County Transportation Authority partnered with researchers at Northeastern University to use data collected from Uber and Lyft's APIs to construct a more granular picture of TNC usage in the city.[3]

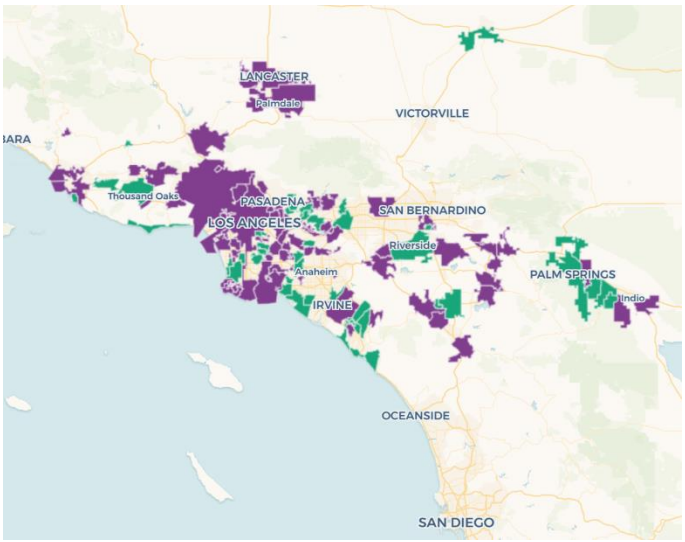
Finally, the expansion of "citizen science" applications can increase the amount of and quality of big data available to agencies, while also making them more accountable. Citizen science refers to the collection and usage of data that is crowdsourced by members of the general public, a process that often yields higher volumes and reduces staff time spent on data collection. For example, nonprofit Ecotrust worked with over 14,000 fishermen off the California coast to identify potential marine protected areas for consideration by the State.

What can we do?

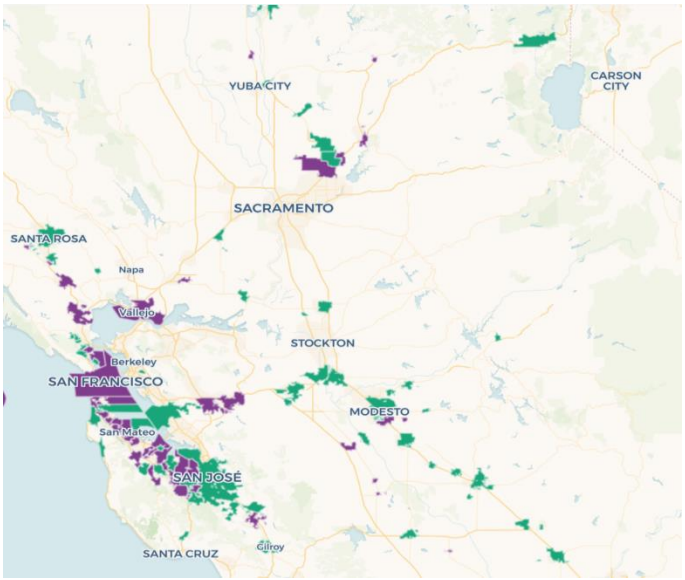
To date, we have largely outsourced innovation in domains such as transportation, environmental quality, and health to the private sector, with significant costs in terms of privacy and equity. Yet, technical capacity building in the public sector can provide a counterbalance--and there is a role here for the University of California. Data science degree programs are blossoming in several UCs; for instance, UC Berkeley already has some 750 data science majors, placing it among the top ten. The California Policy Lab provides an example of UC academics partnering with agencies using rich confidential -- and previously inaccessible -- data.

Following the lead of others, such as the City of New York, the time is right for the State of California to outline its data strategy and communicate the importance of civic data -- and data science -- in government.

Southern California Open Data Portal Cities



Northern California Open Data Portal Cities



Cities (pop >30,000) without open data portals
Cities (pop > 30,000) with open data portals

[3]San Francisco County Transportation Authority and Northeastern University, "TNCs Today: A Profile of San Francisco Transportation Network Company Activity."