Data Science: Challenges and Opportunities for State Policymaking

Karen Chapple
Professor and Chair,
Department of City & Regional Planning
What is big data and analytics?

- Found Data
- Unprecedented volumes
- Diverse Sources
- Analyzed with Algorithms
Data Science in the Private Sector

- Lower legal, and economic barriers
- Qualified Practitioners
- Data Infrastructure
The Risks of Outsourcing

- Privacy and Security Risks
- Sustainability Concerns
- Equity in Access and Impact
Key Questions

What should data science applications in the public sector look like?

How can the UC system support California government with data science?
The State of Data

Number of Open Datasets by CA state department (top 8)

- Health Planning and
- Water Resources
- Health Care
- Public Health
- Treasurer
- Conservation
- Controller's Office
- Energy Commission
- Franchise Tax

# data sets
An Uneven Data Landscape

Cities (pop >30,000) without open data portals
Cities (pop > 30,000) with open data portals
Webtools: big data for a diverse audience

Tableau, Microsoft BI and ArcGIS online: user-friendly interactive data visualization

Javascript web tools: customizable for a specific use case and data set(s)
What is Machine Learning?

- Umbrella term for processes that derive information from large datasets using statistical modeling and computing methods
- ML models are “trained” on existing data before being used to interpret new data

**Predict** a continuous value of an attribute based on other data and information
- OLS regression, neural networks, K-means regression

**Classify** a datapoint into one of many discrete categories based on attributes of the datapoint
- Support vector machines, random forest, nearest neighbor clustering
Classification Applications

Support vector machines for predicting distribution of Sudden Oak Death in California

Qinghua Guo\textsuperscript{a,*}, Maggi Kelly\textsuperscript{a,1}, Catherine H. Graham\textsuperscript{b,c}

\textsuperscript{a} Department of Environmental Science, Policy and Management, 151 Hilgard Hall #3110, University of California, Berkeley, CA 94720-3110, USA
\textsuperscript{b} 3101 Valley Life Science Building, Museum of Vertebrate Zoology, University of California, Berkeley, CA 94720-3160, USA
\textsuperscript{c} Department of Ecology and Evolution, 650 Life Sciences Building, Stony Brook University, NY 11794-5245, USA

Received 1 May 2003; received in revised form 22 June 2004; accepted 12 July 2004

Space-Varying Regression Coefficients: A Semi-parametric Approach Applied to Real Estate Markets

Andrey D. Pavlov*
Predicting Arsenic in Drinking Water Wells of the Central Valley, California

Joseph D. Ayotte, Bernard T. Nolan, and Jo Ann Gronberg

‡U.S. Geological Survey, National Center 413, 12201 Sunrise Valley, CA 92677
§U.S. Geological Survey, McKelvey Bldg., 345 Middlefield Road, Menlo Park, CA 94025

Artificial Neural Networks and Long-Range Precipitation Prediction in California

DAVID SILVERMAN AND JOHN A. DRACUP

Civil and Environmental Engineering Department, University of California, Los Angeles, Los Angeles, California

(Manuscript received 18 August 1998, in final form 5 March 1999)

ABSTRACT

Artificial neural networks (ANNs), which are modeled on the operating behavior of the brain, are tolerant of some imprecision and are especially useful for classification and function approximation/mapping problems, to which hard and fast rules cannot be applied easily. Using ANNs, this study maps a 1-yr monthly (January–December) time series of the 700-hPa teleconnection indices and ENSO indicators onto the water year (October–September) total precipitation of California’s seven climatic zones, with different lag times between the inputs and outputs. It was found that the pattern of rainfall predicted by the ANN model matched closely the observed rainfall with a 1-yr time lag for most California climate zones and for most years. This research shows the possibility of making long-range predictions using ANNs and large-scale climatological parameters. This research also extends the use of neural networks to determine important parameters in long-range precipitation prediction by comparing results gained using all the inputs with results from leaving an individual index out of the network training. This comparison gives insight into the physical meteorological factors that influence California’s rainfall.
Citizen Science

the collection and usage of data that is crowdsourced by members of the general public

*More* data, *Less* time and money
Street Story Reports

This information has been collected through Street Story, an online platform that allows residents, community groups and agencies to collect community input about transportation crashes, near-misses, general hazards and safe locations to travel.

To learn more about Street Story, please see the Street Story Starter Guide or visit our information page.

For qualitative data please contact: katembeck@berkeley.edu.

Report Map

Crashes / Near-misses

Hazards / Safe places

Download Data
API: Application Programming Interface

YOUR SYSTEMS

Data

Applications

API PORTAL

Your API Storefront

Developer Community

Apps

Get your APIs to market on a portal
What could the state do? APR example
- *big data*
- *analytics*
- *API*
- *citizen science*

### ANNUAL ELEMENT PROGRESS REPORT

#### Housing Element Implementation

**Table A**

<table>
<thead>
<tr>
<th>Unit Types</th>
<th>Date Application Submitted</th>
<th>Proposed Units - Affordability by Household Incomes</th>
<th>Total Approved Units by Project</th>
<th>Total Disapproved Units by Project</th>
<th>Streamlining</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Very Low-Income Restricted</td>
<td>Very Low-Income Restricted</td>
<td>Low-Income Restricted</td>
<td>Low-Income Restricted</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>5+</td>
<td>R 11/9/2018</td>
<td>396</td>
<td>544</td>
<td>19</td>
<td>4</td>
</tr>
<tr>
<td>5+</td>
<td>R 12/19/2018</td>
<td>254</td>
<td>130</td>
<td>110</td>
<td>2</td>
</tr>
<tr>
<td>5+</td>
<td>R 12/16/2018</td>
<td>110</td>
<td>75</td>
<td>104</td>
<td>2</td>
</tr>
<tr>
<td>5+</td>
<td>R 11/5/2018</td>
<td>75</td>
<td>75</td>
<td>104</td>
<td>2</td>
</tr>
<tr>
<td>5+</td>
<td>R 10/18/2018</td>
<td>54</td>
<td>1</td>
<td>54</td>
<td>1</td>
</tr>
<tr>
<td>5+</td>
<td>R 10/31/2018</td>
<td>101</td>
<td>1</td>
<td>101</td>
<td>1</td>
</tr>
</tbody>
</table>

**Coming in Spring 2020 at UCB:**

Data Science for Housing – A Professional Training Course
Strategic Plan for Data – and Data Science – what about in California?
Data Science in the UC system