Big Data, Machine Learning, and Data Science for Policy in California: Prospects, Possibilities, Limitations

Henry E. Brady
Monroe Deutsch Class of 1941 Professor of Political Science and Public Policy
Dean, Goldman School of Public Policy, UC Berkeley
Berkeley Institute of Data Science (BIDS)
February 5, 2020
What I Want to Do in This Talk

- **Nomenclature:** Explore meanings of “big data” and “data science” and “artificial intelligence” and “machine learning”
- **The Information Age:** What is happening to us?
- **Big Data Defined:** What is “big data”? Is it a useful phrase?
- **Data Science Defined:** Data Collection, Statistics, Visualization, Curation
- **Artificial Intelligence and Machine Learning:** How do they work?
- **The Limits of Data Science:** What can’t it do?
- **Societal Change:** Overview of Societal Change due to “Big Data” and “AI”
- **Examples of Data Science:** Examples of Big Data/Data Science
- **Conclusions**
NOMENCLATURE
Words Related to The Area

• **Big Data** – Large collections of digitized data
• **Data Science** – Collection, analysis, visualization, and curation of data
• **Artificial Intelligence** – Computer methods such as visual perception, speech recognition, decision-making, language translation
• **Machine Learning** – A specific method of AI and data science
• **Robotics** – Embodied use of big data and data science methods in mechanical/electronic form
• **Cyberinfrastructure** – The computers, sensors, networks, software, and human resources to simplify human and scientific endeavors
2005 National Science Foundation Report on Cyberinfrastructure

Analyzed the future of new technical innovations and the need for more and better cyberinfrastructure
First Half of Talk Drawn from 2019 Article in Annual Review of Political Science:

Annual Review of Political Science

The Challenge of Big Data and Data Science

Henry E. Brady
Department of Political Science and Goldman School of Public Policy, University of California, Berkeley, California 94720, USA, email: hbrady@berkeley.edu

Keywords
big data, data science, artificial intelligence, cyberinfrastructure, causality, prediction, text analysis, internet, smart cities, cyber-warfare, automation

Abstract

Big data and data science are transforming the world in ways that spawn new concerns for social scientists, such as the impacts of the internet on citizens and the media, the repercussions of smart cities, the possibilities of cyber-warfare and cyber-terrorism, the implications of precision medicine, and the consequences of artificial intelligence and automation. Along with these changes in society, powerful new data science methods support research using administrative, internet, textual, and sensor-audio-video data. Bourgeoning data and innovative methods facilitate answering previously hard-to-tackle questions about society by offering new ways to form concepts from data, to do descriptive inference, to make causal inferences, and to generate predictions. They also pose challenges as social scientists must grasp the meaning of concepts and predictions generated by convoluted algorithms, weigh the relative value of prediction versus causal inference, and cope with ethical challenges as their methods, such as algorithms for mobilizing voters or determining bail, are adopted by policy makers.

THE INFORMATION AGE
The Information Age – Important Concepts

Types of Activities

• **Point-to-Point** – telegraph, telephone, e-mail — Person to Person
• **Mass media** – movies, radio, television — Source to Many
• **Computational Power** – ability to calculate — Manipulate Data
• **Networking** – Linkages among people — Connect Groups

Types of Recording of Data

• **Analog versus digital** – e.g., *analog* film or phonograph record versus *digital* audio or video — Digital puts all data in common form — a common “language”
Information Inventions – 1840s to Present

- **Telegraph**: 1840-1860s (point-to-point)
- **Telephones**: 1870-1890s (point-to-point)
- **Phonographs**: 1870-1890s (mass-media)
- **Cinema**: 1890 – 1920s (mass-media)
- **Radio**: 1900 – 1920s (mass-media)
- **Television**: 1940-1950s (mass-media)
- **Mainframe Computers**: 1940-1950s (centralized computing power)
- **Personal Computers**: 1970-1980s (decentralized computing power)
- **Internet & WWW**: 1980-2000s (point-to-point, mass-media, networking)
- **Cell Phones**: 1980-2000s (wide-ranging point-to-point)
- **Amazon, Google, Facebook, Twitter**: 1994-2006 (networking)
- **Smart Phones**: 2000-Now (point-to-point; mass media; computing, networking)
Trends:

-- Greater digitization

-- More point-to-point rather than mass media

-- More information than people can handle

The Physical and Epistemological Features of Information

• **Volume** – Increased enormously (e.g., all sorts of digital information such as the Internet, Videos, etc.)

• **Velocity** – Increased enormously (e.g., transactions data)

• **Variety** – Increased enormously – all sorts of information is now digitized such as cell phone video, podcasts, government data, transactions data such as Amazon, etc.

• **Veracity** – Increased opportunities for counterfeiting of data and information
Four Sociological Features and Trends in Information

• **Datification** (Digitization) – (Recording) – Many events are now turned into digital data including e-mails, videos, commercial transactions, personal transactions, etc.

• **Connectedness** – Events can be linked by name, face, government identification numbers, place of residence, web address

• **Networking** – People have formed networks that are ways that groups can communicate and self-mediate their relationships – somewhere between point-to-point and mass media

• **Authoring** – Computers can “author” information as in answers to queries, computer aided design, video games, “bots,” etc.
BIG DATA DEFINED
Big Data Defined

• As challenge to existing technology because of so much data – not a good definition since technology keeps improving

• Best described by trends noted above:
  • Datatification (digitization)
  • Connectedness
  • Networking
  • Authoring

• “Immersive data” a better term – an environment suffused with data which changes our cognitive environment
Types of Big Data and Its Strengths and Weaknesses

- **Administrative Data**
  - **Good**: Data of record; complete populations; sensitive hard-to-get data
  - **Bad**: Lack of denominator or reference data; lack of demographics; low quality of some data; complexities of linkage; problems of privacy

- **Internet Data**
  - **Good**: Network data; dynamic data; unobtrusive measure of culturally disapproved topics
  - **Bad**: Self-selection into platforms; missing data; problems of privacy

- **Textual Data**
  - **Good**: Much faster than human coding; large bodies of data
  - **Bad**: “Bag of words” paradigm for analysis; complicated and “touchy” estimation methods; requires excellent training sets

- **Sensor, Audio, and Video**
  - **Good**: Often real time; geographically broad coverage; very detailed
  - **Bad**: Difficulty of processing—especially video; Modifiable areal unit problem; problems of privacy
ADMINISTRATIVE DATA: Linking Data Across Government Social Programs in US States

Arrow percentages are proportion of states linking those data sets around 2000.

Percentages in circles are proportion of states linking those data over time.

Using these data, economist Seth Stephens-Davidowitz shows that by comparing John Kerry’s 2004 vote with that of Obama in 2008 and 2012, Obama lost about 4% of the vote due to racial animus.

• **Text Analysis Method** -- Texts of millions of social media posts from Chinese media services in China analyzed before Chinese government censors them.

• **Criticism Not Censored** -- Contrary to previous understandings, posts with negative, even vitriolic, criticism of the state, its leaders, and its policies are not more likely to be censored.

• **Mobilization Efforts Censored** -- Censorship stops collective action by silencing comments that spur social mobilization, regardless of content.

SENSOR DATA: GRACE: Gravity Recovery And Climate Experiment Shows Groundwater Depletion and Loss of Snow/Ice Storage

Over Time Depletion of Groundwater in Central Valley During Drought –2011-13

Source of Slide: Rearranged version from Laurel Larsen, BIDS Talk, 2018
DATA SCIENCE DEFINED
Drew Conway’s Venn Diagram

Hacking Skills

Machine Learning

Math & Statistics Knowledge

Data Science

Danger Zone!

Traditional Research

Substantive Expertise

Data Science -- Elements

- Data Gathering Preparation and Exploration  -- **Substantive fields**
- Data Representation and Transformation  -- **Computer Science**
- Computing with Data  -- **Computer Science**
- Data Modeling  -- **Statistics**
- Data Visualization and Presentation  -- **Media Laboratories**
- Data Archiving, Indexing, Search, and Governance  -- **Library & I-Schools**
- Science about Data Science  -- **Technology Studies Programs**

ARTIFICIAL INTELLIGENCE and MACHINE LEARNING
Data Science and Understanding – Neural Networks

Identifying Dogs and Cats

• **Problem:** How do we classify an animal as a dog or a cat?

• **Features** (each dichotomous for simplicity):
  • Large (around 50 pounds) or not-large (around 8 pounds)
  • Bark or no-bark
  • Tail or no-tail
  • White or non-white
  • Meow or no-meow

• **Classifications:**
  • Large, with bark, tail, non-white, no-meow – a Dog
  • Not-Large, no bark, tail, non-white, meow – a Cat
  • Not-large, no bark, tail, non-white, no meow – (Persian Cat? Basenji Dog?)

• **Note:** Some features are highly diagnostic (bark, meow), others somewhat diagnostic (large or small), and others not helpful (tail or no-tail; white or non-white)
Deep Machine Learning

Many Problems are Pattern Detection or Puzzle Solving

• Symptoms and diagnosis of disease
• Features of the road ahead and safety for proceeding
• Words provided for search and relevant websites
• Bits of DNA and entire genetic sequence
• Patterns of behavior and criminal activity
• Translation of a phrase into another language
THE LIMITS OF DATA SCIENCE
Computer Classification Mistakes

School Bus  A Business  Starfish

Source: Gary Smith, 2018, *The AI Delusion*, Oxford University Press,
Why Computers Were Confused

Source: Gary Smith, 2018, *The AI Delusion*, Oxford University Press,
Dangers of Pattern Detection

• **Training Set Problem** -- Depends upon training set or models input to procedure which can be wrong or biased or limited

• **Better at prediction than at causal inference**
  • More police in area typically mean more crime – do police cause crime?
  • More need for steel means higher prices – does that mean that if I produce more steel the price will go up?
  • Higher education associated with higher income – does that mean that more education leads to more income?
  • More CO$_2$ in the air associated with higher temperatures–does that mean that CO$_2$ cause higher temperatures?

• **Novelty Problem** -- Not good at detecting truly novel and new things never seen before.
What AI Can’t Do (Yet? Ever?)

• **Experience Emotions**—Which help us figure out what is important and which impel us to act. (Note how often you may remember how you feel about something but not the details of why. Think about how emotions lead to action).

• **Create Subtle Analogies and Models**—Which help us understand how the world works because saying that B is like A allows us to hypothesize about what might be true of B given what we know about A. (e.g., People are like animals, like machines, like computers; heat is like a fluid, like atoms moving, etc.)

• **Develop Causal Thinking and Inference**—Which help us understand how our actions (or other actions) will have an impact on the world. (e.g., Putting your hand on a stove may cause a burn; hitting a ball with a bat will cause it to travel, etc.)
SOCIETAL CHANGE IN THE INFORMATION AGE
Cyber Warfare – E.g. Stuxnet virus which destroyed Iranian centrifuges for making nuclear material

Is it warfare when there are no deaths and there is no public declaration of war in order to achieve political advantage?

Or is it “just” sabotage, espionage, and subversion?

Will warfare in the future involve robots? (We already have drones.)

Cyber Sabotage – Russian ads about Hillary Clinton

Issues:

-- How do we find out if an ad is fake?

-- Who can “police” the Internet?

-- How does this interact with free speech?

Three Types of Data:

-- Digitized administrative datasets on people and places (records of services for people, taxes paid, land transactions, police encounters)

-- Sensors, wireless networks, video cameras, etc. can monitor people and things throughout a city and the “Internet of Things” makes it possible to control things.

-- Internet data such as Google Street View, Zillow (real-estate), Yelp (reviews of retailers)
Precision Medicine

Issues:

-- What about people’s privacy?
-- Can we optimize the systems?
-- Who controls the systems?
-- What happens when systems fail?

Media and Politics

What is the Role of the Internet? Mixed:

-- Sorts people into homogenous groups that form “echo-chambers” (probably bad)

-- Discourages preference falsification by providing alternative avenues for expressing opinions (probably good)

-- Overcomes costs of collective action (probably good)

-- Provides erroneous information without any editorial review (probably bad)

Source: Twitter: https://twitter.com/zeynep/status/341358623911448576?s=20. @zeynep “Graffiti: ‘Revolution will not be televised. It will be tweeted.’ Not a metaphor given self-censorship in Turkish TV.” 6:00 pm, June 2, 2013.
Robots and Jobs

Issues:

-- What about people’s jobs?

-- Who owns the robots?

-- Who controls the systems?

-- What happens when systems fail?

-- Can robots really do things beyond sophisticated pattern recognition?


The Many Innovations of the Information Age

- High speed and high capacity digital telecommunications
- Semiconductors, Integrated Circuits, Lasers
- Computers and Specialized Chips
- Databases, methods of storing and retrieving information
- Programming languages, operating systems,
- Sensors, satellites, RADAR, LIDAR, sensing methods
- The Internet, World Wide Web, and Internet of Things
- Artificial intelligence, machine learning, data science
- Social media, Facebook, Amazon, Netflix, Google, Twitter, Blockchain

What should we call all of this?
EXAMPLES OF BIG DATA, DATA SCIENCE, AI and MACHINE LEARNING

-- Central Valley Domestic Water Wells
-- Racial Identity and Racial Profiling
  -- Better Placement of Polling Places for Voters
-- Sonoma County Dashboard for Social Services
-- Predicting Homelessness in Los Angeles County
-- Predicting Water Flow to Improve Dam Storage and Use
Central Valley Domestic Water Wells

**Policy Issue:**
- How can we manage groundwater removal sustainably?
- How can we reduce the number of failed domestic water wells in the Central Valley?

**Paper and Modeling**
- Authors from University of California, Davis, Hydrologic Sciences Graduate Group (four authors) and the Department of Environmental Science and Policy (one author); and from the Water Policy Center, Public Policy Institute of California (one author). Lead author is a graduate student at Davis.
- Prize winner at 2018 California Water Data Hackathon at UC Berkeley
Launched by the Governor’s Office of Planning and Research and partners West Big Data Innovation Hub, Water Foundation, Imagine H2O, Bay Area Council, and the State agency stewards of water data such as State Water Resources Control Board and Department of Water Resources.

California Water Data Hackathon
California Safe Drinking Water Data Challenge

HACKING
September 14, 2018 to September 15, 2018
10:00am to 5:00pm
190 Doe Library

GET DIRECTIONS

The Division of Data Sciences at UC Berkeley and the Berkeley Institute for Data Science (BIDS) are hosting the California Water Data Hackathon to help find innovative ways to increase community access to safe drinking water, better understand vulnerabilities, and identify and deploy solutions. This event will immediately follow the Global Climate Action Summit in San Francisco (#GCAS2018), and is just one of the events and efforts supporting this year’s California Safe Drinking Water Data Challenge on June 26 - October 1, 2018 (#CAWaterDataChallenge).
The Basic Problem: Falling Ground Water and Shallow Wells

If ground water (in light blue) drops, then the active well on the left could fail.

Ground Level Falls Because of:
-- Nature: Droughts and Lack of Rain
-- People: Pumping of Water

Failure depends on:
-- Depth of well
-- Ground Level Fall due to Droughts and Pumping of Water

Source: Pauloo et al., 2020
Data Sources

• **Location, Elevation, and Depth of Wells** -- Recently digitized “well completion reports” from 943,469 wells from the California Department of Water Resources – This provides data on all the relevant wells and their depth. [https://data.ca.gov/dataset/well-completion-reports](https://data.ca.gov/dataset/well-completion-reports) (administrative data)

• **Groundwater Levels** -- Seasonal groundwater level measurements -- [https://gis.water.ca.gov/app/gicima/](https://gis.water.ca.gov/app/gicima/)

• **Well Failures Year by Year** – 2012-2016 -- Confidential data obtained from California Department of Water Resources
Three Datasets – Linked by Geography to Develop a Model

Location, Elevation, and Depth of Wells

Changes in Groundwater Over Time

Time of Well-Failure Data

Source: Pauloo et al., 2020

Tyler Johnson and Kenneth Belitz, 2015, Figure 7, p 43.

https://gis.water.ca.gov/app/gicima/
Observed Well Failures Used to “Train” or “Fit” Model

(A) Observed 2012-2016 well failure

(B) Predicted 2012-2016 well failure

Source: Pauloo et al., 2020
Fitted Model Used to Predict Results of Alternative Scenarios

Observed 2012-2016 Drought

Simulated Drought Durations

Cumulative Sum of Well Failures

---

Source: Pauloo et al., 2020
Actual groundwater level


Linear approximation (2003 - 2017)

Linear approximation (2008 - 2017)

Strict sustainability

Glide path

Business as usual

Predicted Results of Long-Term Water Management Policies

Source: Pauloo et al., 2020
Racial Identity and Police Profiling

• Policy Issue:
  • Is there racial profiling by police?
  • Hypothesis:
    • Many studies have shown the existence of implicit racial bias.
    • Does implicit bias give rise to discrimination through discretion?
    • Since police sometimes have high discretion (e.g., stop/search decisions), is it true that disparities increase with discretion?

• Report:
  • RACIAL & IDENTITY PROFILING ADVISORY BOARD, 2020 Annual Report
Two Goldman School of Public Policy Faculty Involved:

STEVEN RAPHAEL, Professor of Public Policy, University of California, Berkeley; Board Member, Appointed by the Governor of California

JACK GLASER, Professor, Goldman School of Public Policy, University of California, Berkeley

Assistance in preparing this year’s report.

Source: RIPA, 2020
California’s Racial and Identity Profiling Act of 2015 (RIPA) required collection of data about stops, officer’s perception of person stopped, and the officer. The 2020 report analyzes 1.8 million stops from July 1, 2018 to December 31, 2018. (Eight Largest Police Agencies in Calif.)

### Information Regarding Stop

1. Date, Time, and Duration
2. Location
3. Reason for Stop
4. Was Stop in Response to Call for Service?
5. Actions Taken During Stop
6. Contraband or Evidence Discovered
7. Property Seized
8. Result of Stop

### Information Regarding Officer’s Perception of Person Stopped

1. Perceived Race or Ethnicity
2. Perceived Age
3. Perceived Gender
4. Perceived to be LGBT
5. Limited or No English Fluency
6. Perceived or Known Disability

Source: RIPA, 2020
Who is Stopped?

Does not prove bias because we don’t have “comparison” data to determine “rates” of stops: E.g., “Number of stops per person of that group.”

- Individuals perceived to be Hispanic (39.8%), White (33.2%), or Black (15.2%) comprised the majority of stopped individuals.
Comparison: Percentage of Police Stops (RIPA) of each group in places compared to Census Data (American Community Survey-ACS) on racial/ethnic percentages of places.

Note:
-- More stops than Population for Black and Middle-Eastern/South Asians
-- About Same for Hispanic and White.
-- Fewer stops than population for Asians.

But:
-- Results are very worrisome, but they may not prove bias because we don’t have this comparison: Who should have been stopped?
One Comparison: How many with contraband or evidence?

Consider Search Yield Rates. A search yield rate is the proportion of *individuals that were subject to a search* that officers found to be in possession of contraband or evidence.

SEARCH YIELD RATE = \( \frac{\# \text{ of Searched Individuals with Contraband or Evidence}}{\text{Total Number Searched Individuals}} \times 100 \)

Higher search yield rates indicate that searches were successful and resulted in finding contraband or evidence (a “hit”) more often. Higher search yield rates suggest that there was a good reason (not bias) to perform the search. Lower search yield rates suggest that there might not be a good reason to perform the search.
Somewhat similar Search Yield Rates By Race Ethnicity, although White Search Yield Rates are Highest so that other groups searched with less success which suggests unnecessary searches.
An Important Refinement

• If there is bias, we expect more bias with high discretion searches where the only basis for the search is the officer’s suspicion and consent of those searched.

• With low discretion searches, there is not much leeway for the police officer to decide to search or not to search – the officer is essentially mandated to search so bias is less likely.
When officers have discretion, their searches of Blacks and Hispanics yield less than those of Asians and Whites.

Source: Racial & Identity Profiling Act Annual Report, 2020, Half of Figure 12
Result

• “When officers conducted highly discretionary searches of individuals (only basis was consent), officers had higher yield rates for White persons than for all other racial/ethnic groups.” (RIPA)

• Strongly suggests some bias in searches. Blacks and Hispanics were searched but yields were lower.

• Please see entire report to obtain a complete picture.
Better Placement of Polling Places for Voters

• **Policy Issues**
  • What are the disruption (finding new polling place) and transportation costs of voting?
  • How can we minimize them to not discourage voting?

• **Natural Experiment:**
  • Los Angeles County had 5000 polling places in 2002 general election
  • And only 2000 polling places in 2003 Recall Election

• **For many People this Increased Costs of:**
  • Finding polling place
  • Traveling to polling place

Los Angeles County Voters

Source: Brady McNulty, 2011
2002 Polling Places

Source: Brady McNulty, 2011
2003 Polling Places

Source: Brady McNulty, 2011
Increased Distance to Polling Place from One-Third of a Mile to Half a Mile on Average

Source: Brady McNulty, 2011
What was done to get data (abbreviated)

Polling Places 2002
N: 5,000

Precinct Link Data

Voter File 2002
N: 3.8 million

Linked by Unique Voter ID

Polling Places 2003
N: 2,000

Precinct Link Data

Voter File 2003
N: 3.6 million

Merged File for Analysis, 2002-2003

Voters in both data sets at same address both years, with polling place data both years

N: 3,111,067

Source: Brady McNulty, 2011
Results – Los Angeles Consolidation

Polling Place Voting
-3.1% -3.1%

Absentee Voting
1.5% 1.2%

Non-Voting
1.6% 1.9%

Unadjusted
Matching

Source: Brady McNulty, 2011
Decrease in Polling Place Voting with:
Location Change (-2.8%) and
Distance Change (-.25% per tenth mile)

Net Decrease in Voting With:
Location Change (1.8%) and
Distance Change (.05% per tenth mile)

Net Increase in Absentee Voting With:
Location Change (1.0%) and
Distance Change (.20% per tenth mile)

Source: Brady McNulty, 2011
Sonoma County Dashboard for Social Services

• **Policy Issue:**
  • Can we manage services better across agencies?
  • **Hypothesis:**
    • Some recipients of county services may interact with multiple agencies and require very costly services, and
    • These multiple agencies may work at cross-purposes.
    • Can we identify these people and design better ways to help them which will cost less money?

• **Report:**
  • California Policy Laboratory, UC Berkeley and UCLA, and Sonoma County
The ISSUE: The High Cost for the Top 5%

Each person in the top 5% of those experiencing homelessness in Santa Clara County costs over $100,000 in services:

-- The top 5% require 57% of resources;

-- The top 10% require 74%.

Steps Taken in Sonoma County

• **ACCESS** -- Accessing Coordinated Care & Empowering Self-Sufficiency – Created by Sonoma Count to identify most vulnerable and to improve services.

• **Problem** – Those who cycle through hospitals, homeless shelters, and jail cells also use other county services – How can they be better served?

• **California Policy Lab** – The Lab helped the County link information on 500,000 people across health, mental health, substance abuse, housing, criminal justice, and human services systems for fiscal years 2014 through 2018.

• **Identifying High Utilizers** – They developed a score measuring prevalence and intensity of utilization in those systems.
Percentage of People in System with Various Mental Health Diagnoses

Percentage of People in System with Various Substance Abuse Issues

Percentage of People in System with Various Criminal Charges

Percentage of People in System with Various Housing Statuses

Source: California Policy Lab
Top Half of Utilizers

Share of Top Utilizers Present in Each System

Source: California Policy Lab
Predicting Homelessness in Los Angeles County

• Policy Issue:
  • Homelessness is a very rare event: just 1.7% of about 1.9 million single adult LA County clients—33,600 people—experienced a new homeless spell in 2017.
  • Give the rarity of homelessness, can we find a way to predict it and make efforts to prevent it?
  • Hypothesis:
    • Those recipients of county services may interact with multiple agencies and there may be some pattern of this interaction that predicts homelessness.
    • This pattern may be hard to discern – can machine learning find the pattern?

• Report:
  • Till von Wachter, Marianne Bertrand, Harold Pollack, Janey Rountree, and Brian Blackwell, September 2019, “Predicting and Preventing Homelessness in Los Angeles, California Policy Laboratory, UC Berkeley and UCLA; University of Chicago Poverty Lab.
The Problem and the Data

County interacts with 1.9 million people per year in:
-- Emergency rooms
-- Inpatient and outpatient medical care
-- Mental health and substance abuse
-- Safety net programs (General Relief, CalFRESH)
-- County Jails and Probation

A fraction Experience Homelessness:
-- 76,000 single adults total with...
  -- 42,000 ongoing episode
  -- 34,000 new spell

Can we use the Los Angeles County Enterprise Linkage Project with 85 million service records on 1.9 million adults from seven agencies to predict homelessness?

Source: von Wachter et al, 2019
The Solution: Predictive Analytics Using AI and Machine Learning

• **TRAIN OR ESTIMATE:** Use past service utilization data for which we know who became homeless to identify those risk factors leading to homelessness

• **SIMULATE AND PREDICT:** Apply this model to service utilization data from 2012 to 2016 to predict likelihood of new homelessness for 2017.

• **VALIDATE:** Identify a list of likely people, say 3000. Of this list, how many became homeless?
  • Almost half became homeless: 45.7%

Source: von Wachter et al, 2019
Predicting Water Flow To Improve Dam Storage and Water Use

• **Policy Issues:**
  • Climate change means keeping reservoirs full will help avoid water shortages
  • Full reservoirs can lead to floods downriver when there is heavy precipitation and water must be released; overfull reservoirs can lead to catastrophic failures for spillways or dams
  • Can we manage water better with improved hydrologic models?

• **Source:** “Frontiers in Water Forecasting: Merging Data-Intensive Approaches with First Principles,” Laurel G. Larsen, Depts. of Geography and Civil and Environmental Engineering, UC Berkeley, Presentation at BIDS, 2018. All slides from this talk.
The Catastrophic Possibility

Oroville Dam, February 27, 2017

Image credit: KCRA via AP
Current reservoir operations are typically prescribed by time of year.

Can hydrological knowledge help us manage reservoirs better?

Problem: Earth systems are nonstationary and nonlinear. How to predict the future?

Question: How to properly represent critical interactions and feedbacks in our models?

MODELING: Is modeling all feedbacks the answer?


Image source: http://www.cesm.ucar.edu
DATA: Is analyzing data from new global hydrologic sensor networks the answer?

The Hydrologic Modeling Pendulum: Balance between physical models and data-driven efforts

- Knowledge and **physical models**
  - First principles
  - Conservation of mass
  - Lumped or distributed

- Learning through **data**
  - Regression
  - Artificial neural networks
  - Support vector machine

*How can we optimize learning using physically based and data driven models? We need a bridge:*

- Information theory for causal inference and delineation of critical time and spatial scales
- Sparse regression to “discover” governing equations from data
- Formulate empirical forecasts constrained by physics

Conclusions

- Water storage and balancing direct/indirect needs for water a salient challenge for future water sustainability.
- Smarter water resource management requires improved hydrologic forecasting and understanding of hydro-meteorological teleconnections.
- Hydrology benefits from advances in other disciplines relevant to resolving causal interactions, feedbacks, and critical timescales and spatial scales.

CONCLUSIONS
## CONCLUSIONS

<table>
<thead>
<tr>
<th></th>
<th>Water Wells</th>
<th>Racial Profiling</th>
<th>Polling Places</th>
<th>Sonoma Dashboard</th>
<th>Predicting Homelessness</th>
<th>Water Flow to Dams</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Type of Data</strong></td>
<td>Admin</td>
<td>Admin</td>
<td>Admin</td>
<td>Admin</td>
<td>Admin</td>
<td>Sensor</td>
</tr>
<tr>
<td><strong>Linkages</strong></td>
<td>Geo</td>
<td>Geo</td>
<td>Geo, Identifiers</td>
<td>Names</td>
<td>Names</td>
<td>Geo</td>
</tr>
<tr>
<td><strong>Physical/Social Model</strong></td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Advanced Statistics</strong></td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Policy Goals</strong></td>
<td>Saving Water Wells</td>
<td>Making Policing More Fair</td>
<td>Easier Voting</td>
<td>Managing High Utilizers Better</td>
<td>Predicting Homelessness &amp; Intervening</td>
<td>Managing Reservoirs Better</td>
</tr>
</tbody>
</table>
Conclusions

• Cyberinfrastructure is changing the way we run societies
• Society must wrestle with more data, reduced privacy, computer algorithms, and perhaps robots replacing humans in some jobs
• But there are tremendous opportunities for improving public sector decision-making, management, and service delivery.