CA Water Board Data Management Strategy and Open Data Initiative

Greg Gearheart, Deputy Director
Office of Information Management and Analysis

State Water Board
March 30, 2018

Why do we care about mercury in CA waters?

Who is most affected?

How are they affected?

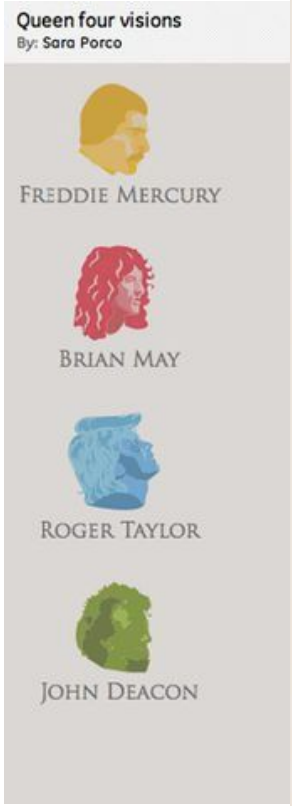
What affects how they are affected?

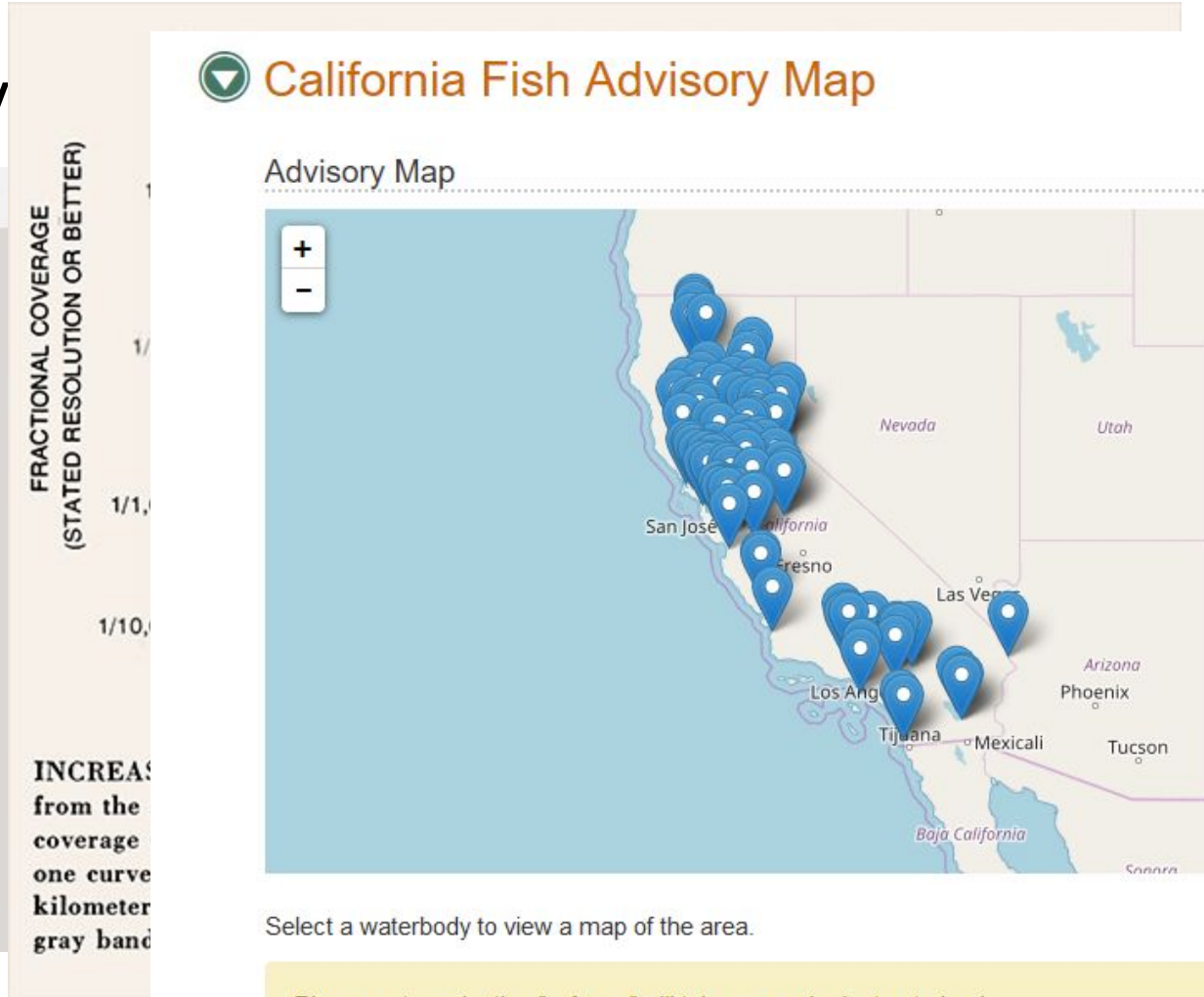
Do we make decisions about this impact transparently?

Do we effectively involve (let alone engage) those most affected in our processes?

What can we do better now?

Mercury





California Fish Advisory Map

Advisory Map

Select a waterbody to view a map of the area.

Decisions about “salmon vs mercury” could benefit from more timely, more accessible information

Water (and therefore “water data”) Governance in CA

- Federal agencies (usual suspects)
- State agencies (CA Water Boards and CA Dept. of Water Resources are primary)
- Local agencies (county and city governments in some cases)
- Special districts (lots)

Open Water Data in CA

- Pre September 2016 - voluntary, decentralized, topic driven efforts
- Post September 2016 - the Open and Transparent Water Data Act (AB 1755) has organized CA water agencies
  - State partners and a Data Management Strategic Plan (DWR leading)
  - Stakeholders curating use cases (UC Water)
  - Technical requirements (San Diego Supercomputing Center)

WB Data Management Strategy (2017 update)

Annual Civic Engagement Events

- Data Fair (open house)
- Data Innovation Challenge (hackathons)
- Water / Data Science Synthesis
- Brown Bag Series of Speakers
- Other partnerships

Our Data Management Strategy Framework

- Based on Principles
- Will guide:
  - Data driven management
  - “Water” decisions
  - “Technology” decisions
  - Quality program
- Lists our data management values
- Encourages “data literacy”

Databases and Datasets at the CA Water Boards

- Over 20 enterprise database applications
- Water quality, water rights, drinking water, etc.
- Program data (e.g, facilities, activities) and environmental / ambient data (e.g., surface water and groundwater quality, water use, water conservation, etc.)
- 18 data resources on data.ca.gov → more all the time

Why do we collect, use and produce data?

- To inform our **data-driven management** and planning activities – performance report cards, workplans, resource assignment/augmentation, evaluating program effectiveness, and many others examples;
- To inform our **critical decisions** regarding our mission(s) and water management responsibilities – water allocation and use, water quality planning and "policies," permitting, program prioritization, and many other examples; and
- To provide **transparency** to our many partners and stakeholders for their use, interests and purposes.

Why do we focus on open data?

- Open data is machine readable, well documented, accessible data
- From here, data becomes information SO MUCH EASIER
- To get here (or from here in an iterative mode), data users will help enforce that it must be structured, reviewed for quality, described, and timely
- Provides a perfect vantage point to talk about the whole data lifecycle

Stage 1 - Collecting Data

CIWQS Data

CIWQS Data Collection

Standards

Data Checkers

Data Checkers

Data

SWAMP Data Collection

Standards

S2 - Storing Data

CIWQS Database

SWAMP / CEDEN Database

Other Databases

Web Services

Web Services

S3 - Publishing Open Data

	C	D	E	F	G	H	I
1	Latitude	Longitude	SampleDate	ProjectCode	LocationCode	CollectionTime	MatrixName
2	34.14699	-118.1633	8-Mar-16	BASMAA_TCT_Midchannel		8:57	samplewater
3	34.14699	-118.1633	8-Mar-16	BASMAA_TCT_Midchannel		10:08	samplewater
4	34.14699	-118.1633	8-Mar-16	BASMAA_TCT_Midchannel		11:53	samplewater
5	34.14699	-118.1633	8-Mar-16	BASMAA_TCT_Midchannel		12:34	samplewater
6	34.14699	-118.1633	8-Mar-16	BASMAA_TCT_Midchannel		12:34	samplewater
7	34.14699	-118.1633	8-Mar-16	BASMAA_TCT_Midchannel		11:18	samplewater
8	34.14699	-118.1633	8-Mar-16	BASMAA_TCT_Midchannel		13:14	samplewater
9	34.14699	-118.1633	8-Mar-16	BASMAA_TCT_Midchannel		13:48	samplewater
10	34.14699	-118.1633	7-Mar-16	BASMAA_TCT_Midchannel		12:47	samplewater
11	34.14699	-118.1633	7-Mar-16	BASMAA_TCT_Midchannel		13:31	samplewater
12	34.14699	-118.1633	7-Mar-16	BASMAA_TCT_Midchannel		13:31	samplewater
13	34.14699	-118.1633	7-Mar-16	BASMAA_TCT_Midchannel		14:07	samplewater
14	34.14699	-118.1633	7-Mar-16	BASMAA_TCT_Midchannel		14:43	samplewater
15	34.14699	-118.1633	7-Mar-16	BASMAA_TCT_Midchannel		15:24	samplewater
16	34.14699	-118.1633	7-Mar-16	BASMAA_TCT_Midchannel		16:03	samplewater

CalData data.ca.gov Portal

CIWQS

SWAMP / CEDEN

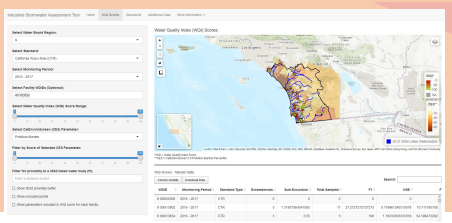
Other Open Data

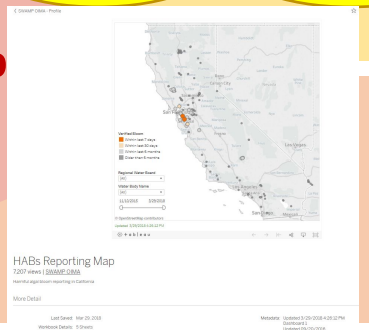
Federal

CA DWR

APIs / Web Services

S4 - Turning Data into Information





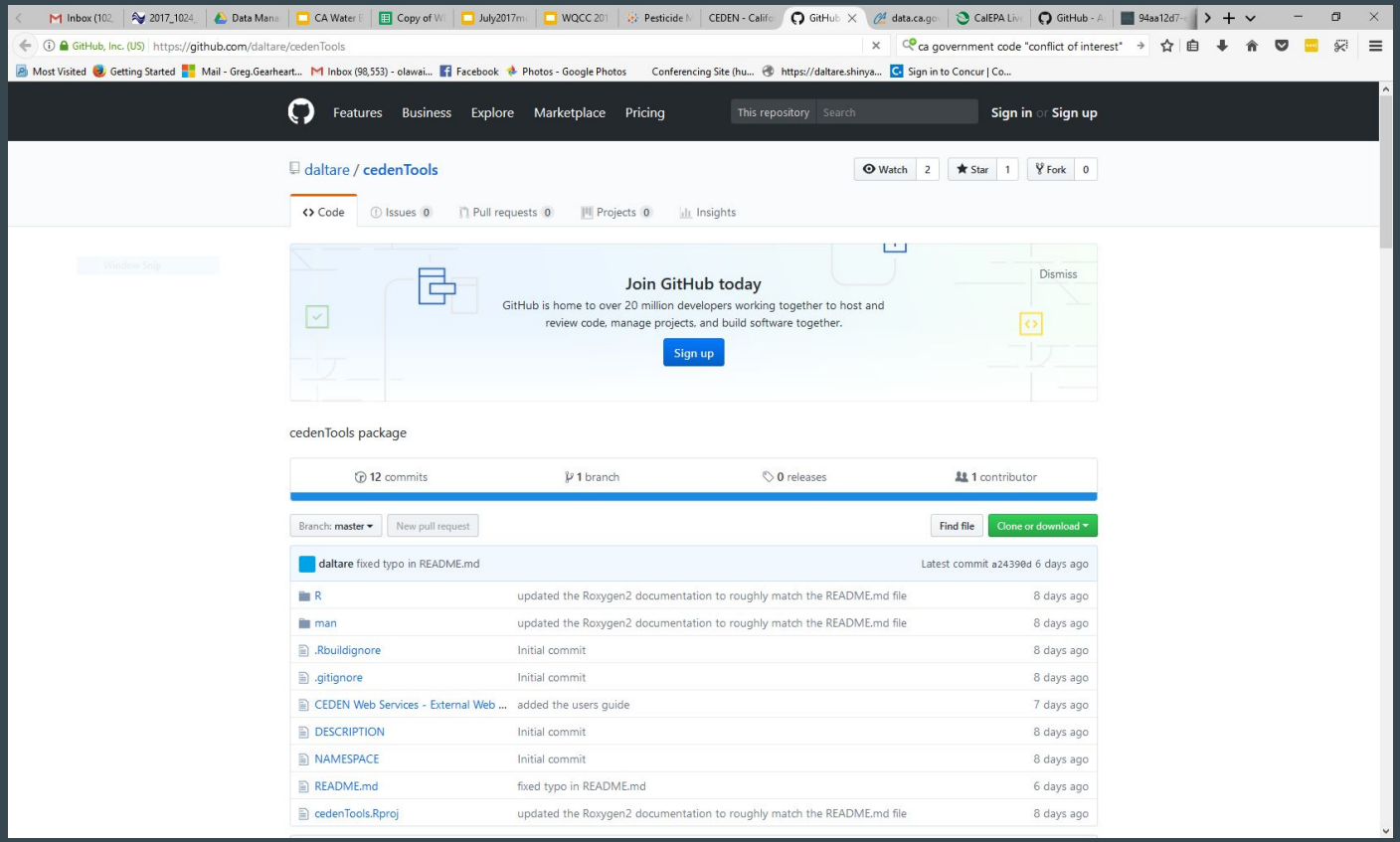
Stage 0 - the driver(s) for data collection

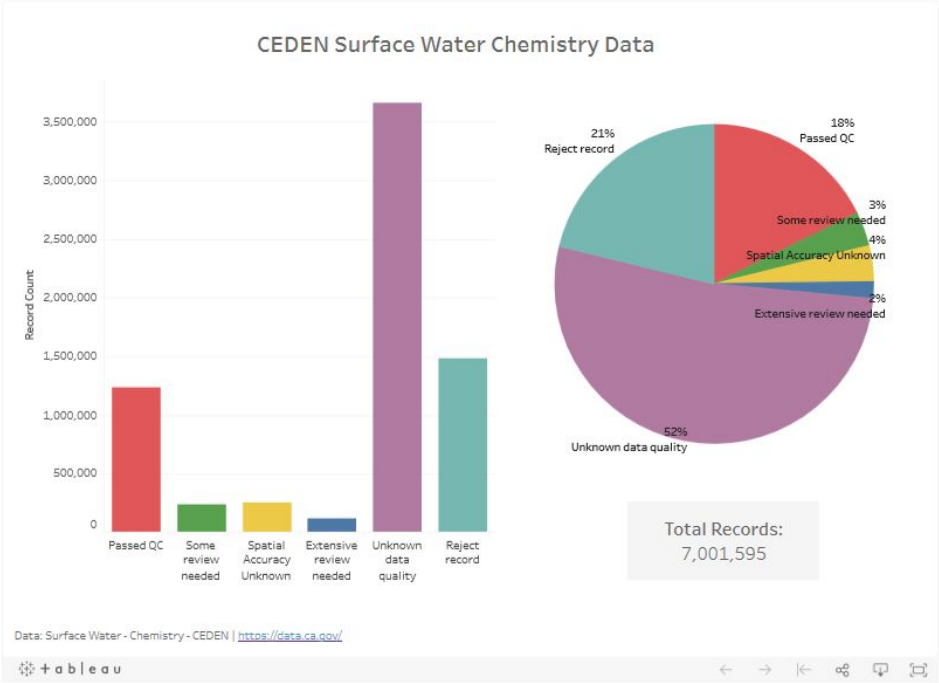
- Why do we collect data in the first place?
  - Permit requirements
  - Special studies
  - Monitoring program (sometimes regional)
  - Other agencies doing data collection!

Looking through the “accessibility” window for data lifecycle management

- Debates about methods → choices for data collection (driven by data quality objectives)
- Burdensome data costs → dialogs about efficient and effective data collection
- Data needs wrangling → solutions to establish data standards or implement data transformation “codes” on the path ro publication
- Prioritizes data that is machine readable, timeliness and is well documented

R package to use web services to access data in CEDEN





CEDEN Water Chemistry Data

50 views | SWAMP OIMA

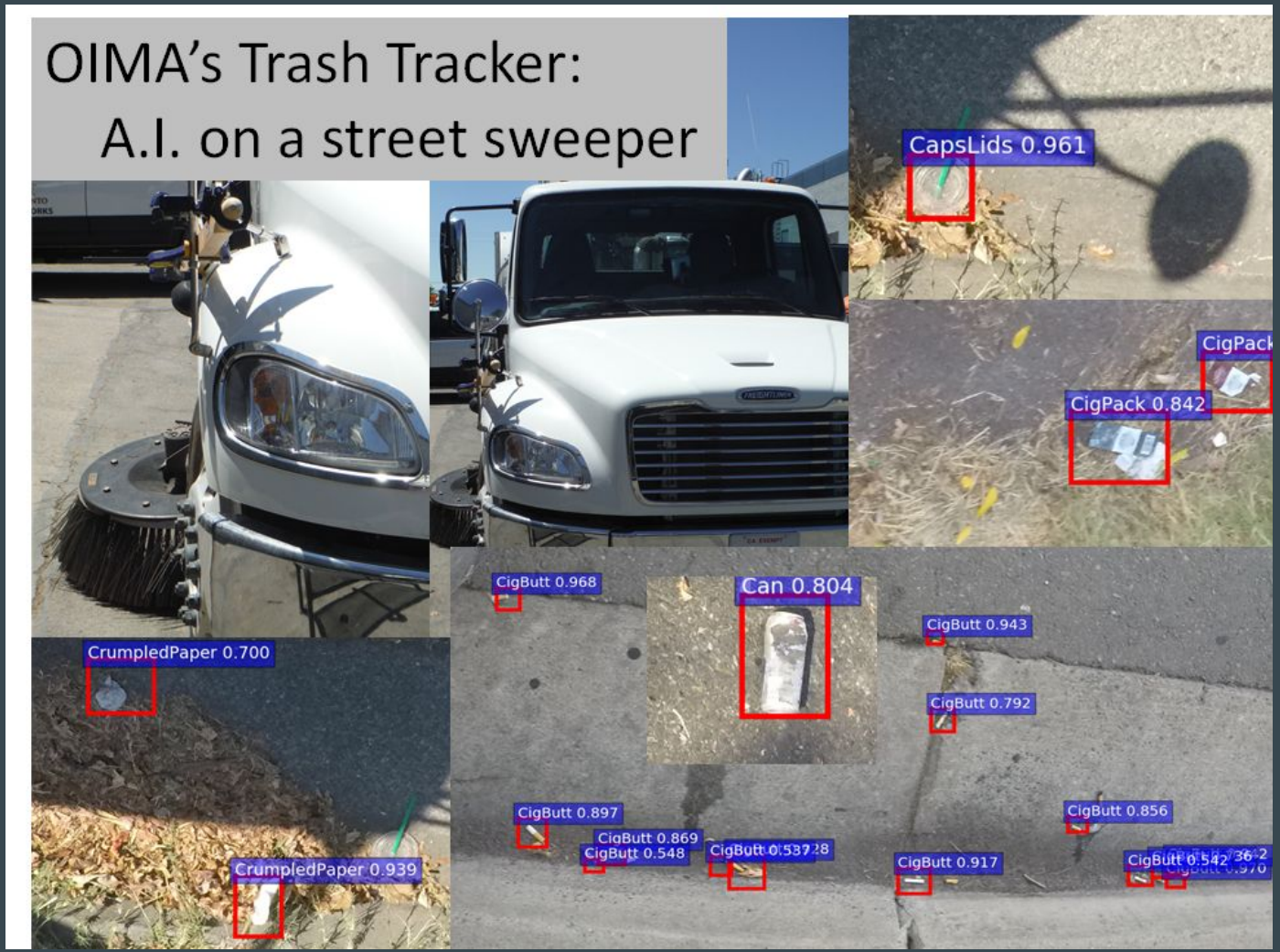
More Detail

Last Saved: Feb 7, 2018

Workbook Details: 3 Sheets

Metadata: Bar
Pie
Dashboard 1

We have proof of concept results using computer vision (a form of artificial intelligence and machine learning) to recognize trash shapes in images, which can be captured via street sweepers, refuse trucks or other means.



Stormwater Enforcement Tool

Filters:

Select Water Board Region:

9

Select Standard:

CTR

Select Monitoring Period:

2016 - 2017

Select Facility WDIDs (Optional):

All WDIDs

Select WQI Score Range:

0 100

Select a CES Parameter:

Pollution Burden

Filter by Score of Selected CES Parameter:

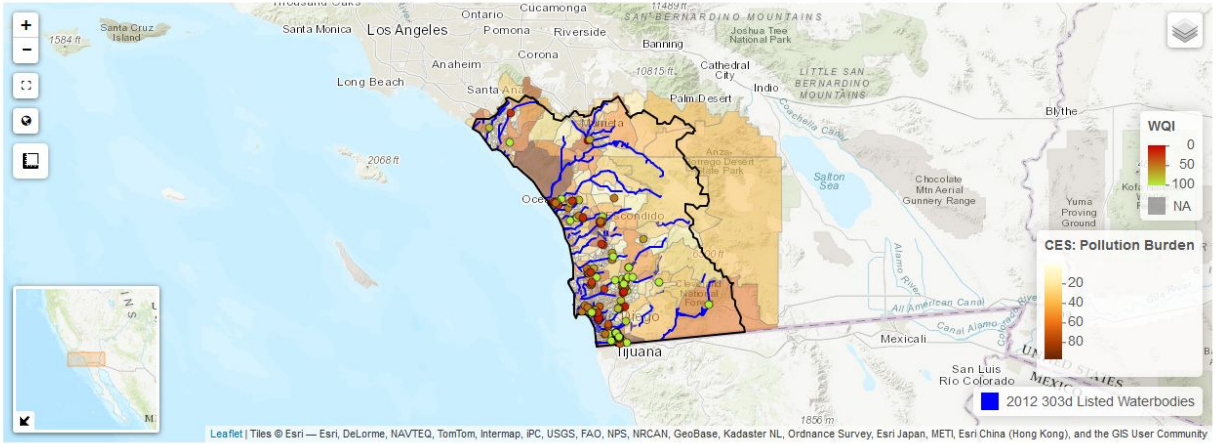
0 100

Filter for proximity to a 303d listed water body (ft):

Enter a distance in feet

- [ ] Show 303d proximity buffer
- [ ] Show excluded points
- [ ] Show parameters included in WQI score for each facility

WQI Scores:



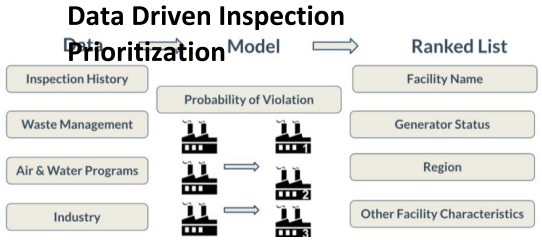
WQI Scores - Tabular Data:

Column visibility | Download Data | Search:

WDID	Monitoring.Period	Standard.Type	Exceedances	Sum.Excursion	Total.Samples	F1	NSE	F
9 301005260	2016 - 2017	CTR	0	0	2	0	0	
9 301013802	2016 - 2017	CTR	3	1.31957364341085	11	27.2727272727273	0.119961240310078	10.711195708
9 301013854	2016 - 2017	CTR	3	3.55	3	100	1.18333333333333	54.198473282
9 331000856	2016 - 2017	CTR	2	1120	2	100	560	99.821746880
9 331023411	2016 - 2017	CTR	2	1.35833333333333	6	33.3333333333333	0.226388888888889	18.459796149
9 371000277	2016 - 2017	CTR	49	236.98	112	43.75	2.11589285714286	67.906470284

https://daltare.shinyapps.io/Stormwater_Enforcement_Tool/

Machine Learning Examples



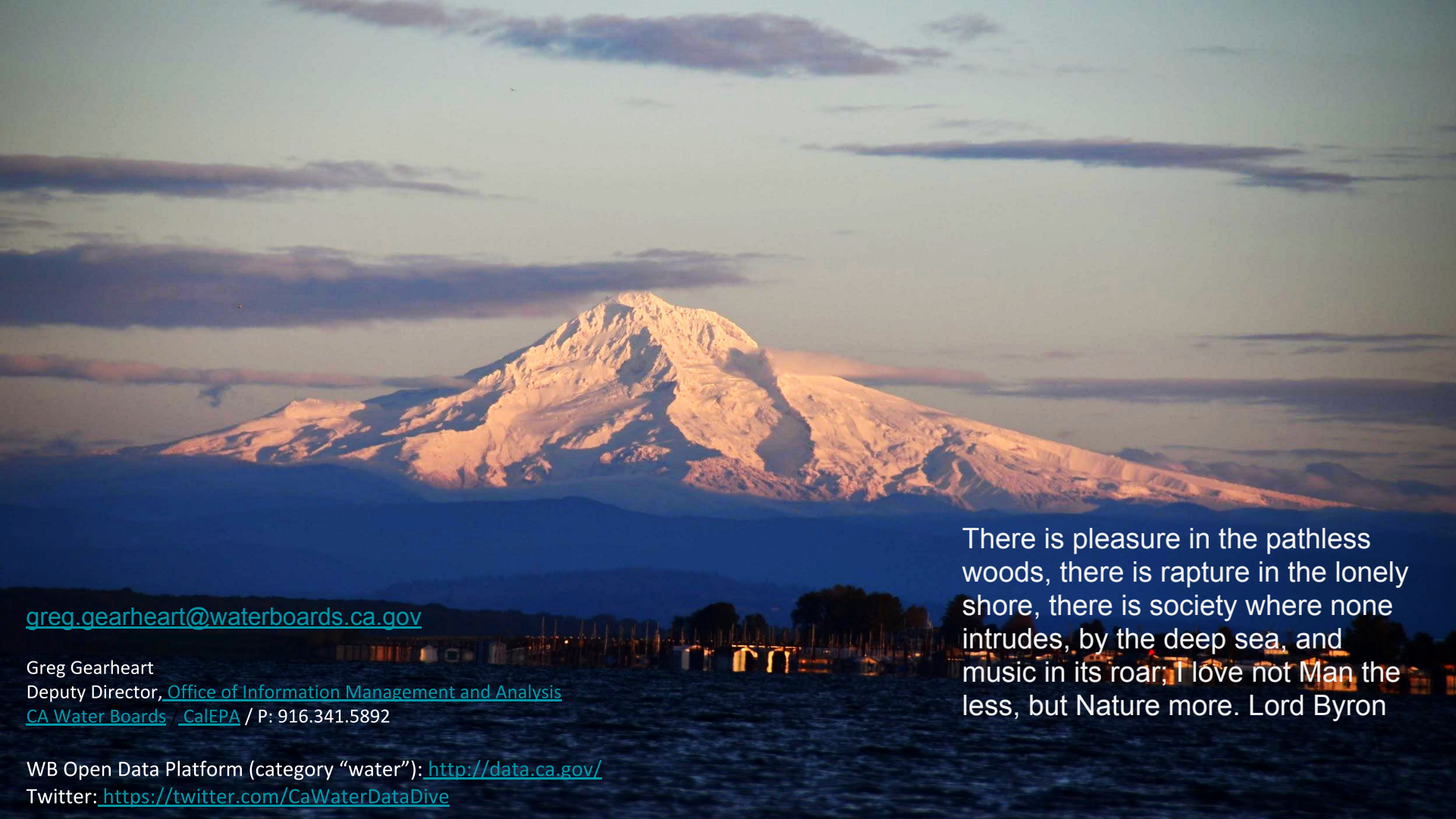
- **University of Chicago's Center for Data Science & Public Policy**:
  - Predictive Enforcement of Pollution and Hazardous Waste Violations (w/ US EPA)
    - EPA wants to conduct more targeted investigations (only 4% of facilities inspected per year)
      - Goal: better allocate inspection resources to maximize the impact of each investigation
    - Developed and evaluated predictive models to identify likely violators using historical EPA data on reporting, monitoring, inspection, & enforcement
      - Results weighted by multiple criteria (e.g., likely outcome of an enforcement action, magnitude, and potential impact of violation on environmental and public health)
    - Predicted 620,000 tons of pollution per year could be prevented by data driven approach
      - From 340,000 tons currently prevented to 960,000 tons prevented by improved inspection approach
      - Increase in inspection hit rate (violations found per inspection) from 28% to 79%
  - Predictive Enforcement of Pollution and Hazardous Waste Violations in New York State (w/ New York State Department of Environmental Conservation)
    - Can consider geographic features like flood zones, population density, etc.
    - Predicted increase from 400 to 750 violations per 1000 inspections using the model
  - Early Warning System for Water Infrastructure Problems
    - Created a predictive model that the city of Syracuse can use to replace or repair water mains before they fail
    - Can also help the city make decisions about the kinds of replacement mains that are best suited for different locations and environments, and help coordinate activities between departments to get the most infrastructure work done in a single dig.
  - Data-Driven Digital Engagement for Environmental Causes
    - Built engagement models for an environmental non-profit that predict which individuals are likely to take particular actions and the best way to communicate with those people (to improve rates of volunteering, donation, and advocacy)
  - Using Sensor Data to Inform and Evaluate Environmental Initiatives

CA Water Data Profile #1 - Lean Water Conservation Data

- Urban Water Supply Data - used to motivate Californians to conserve water (see Stanford https://news.stanford.edu/2017/10/25/media-attention-drought-produced-water-savings/)
- Extremely lean resources assigned to program
- Data driven messaging from top CA water leaders
- Data is now published (automatically) via data.ca.gov

CA Water Data Profile #2 - “Not Lean” Regulatory Data

- The National Pollutant Discharge Elimination System permit system regulates over 20,000 facilities in CA
- Two data systems manage the program - CIWQS (traditional) and SMARTS (stormwater)
- Self reporting data gets loaded regularly
- Over 200 staff are employed to run the program and use the data
- Millions of dollars spent to build and maintain the systems



There is pleasure in the pathless woods, there is rapture in the lonely shore, there is society where none intrudes, by the deep sea, and music in its roar; I love not Man the less, but Nature more. Lord Byron

greg.gearheart@waterboards.ca.gov

Greg Gearheart
Deputy Director, Office of Information Management and Analysis
CA Water Boards / CalEPA / P: 916.341.5892

WB Open Data Platform (category “water”): http://data.ca.gov/
Twitter: https://twitter.com/CaWaterDataDive

Stage 1 - Collecting Data

CIWQS Data

CIWQS Data Collection

Standards

Data Checkers

Data Checkers

Data

SWAMP Data Collection

Standards

S2 - Storing Data

CIWQS Database

SWAMP / CEDEN Database

Other Databases

Web Services

Web Services

S3 - Publishing Open Data

	C	D	E	F	G	H	I
1	Latitude	Longitude	SampleDate	ProjectCode	LocationCode	CollectionTime	MatrixName
2	34.14699	-118.1633	8-Mar-16	BASMAA_TCT_Midchannel		8:57	samplewater
3	34.14699	-118.1633	8-Mar-16	BASMAA_TCT_Midchannel		10:08	samplewater
4	34.14699	-118.1633	8-Mar-16	BASMAA_TCT_Midchannel		11:53	samplewater
5	34.14699	-118.1633	8-Mar-16	BASMAA_TCT_Midchannel		12:34	samplewater
6	34.14699	-118.1633	8-Mar-16	BASMAA_TCT_Midchannel		12:34	samplewater
7	34.14699	-118.1633	8-Mar-16	BASMAA_TCT_Midchannel		11:18	samplewater
8	34.14699	-118.1633	8-Mar-16	BASMAA_TCT_Midchannel		13:14	samplewater
9	34.14699	-118.1633	8-Mar-16	BASMAA_TCT_Midchannel		13:48	samplewater
10	34.14699	-118.1633	7-Mar-16	BASMAA_TCT_Midchannel		12:47	samplewater
11	34.14699	-118.1633	7-Mar-16	BASMAA_TCT_Midchannel		13:31	samplewater
12	34.14699	-118.1633	7-Mar-16	BASMAA_TCT_Midchannel		13:31	samplewater
13	34.14699	-118.1633	7-Mar-16	BASMAA_TCT_Midchannel		14:07	samplewater
14	34.14699	-118.1633	7-Mar-16	BASMAA_TCT_Midchannel		14:43	samplewater
15	34.14699	-118.1633	7-Mar-16	BASMAA_TCT_Midchannel		15:24	samplewater
16	34.14699	-118.1633	7-Mar-16	BASMAA_TCT_Midchannel		16:03	samplewater

CalData data.ca.gov Portal

CIWQS

SWAMP / CEDEN

Other Open Data

Federal

CA DWR

APIs / Web Services

S4 - Turning Data into Information