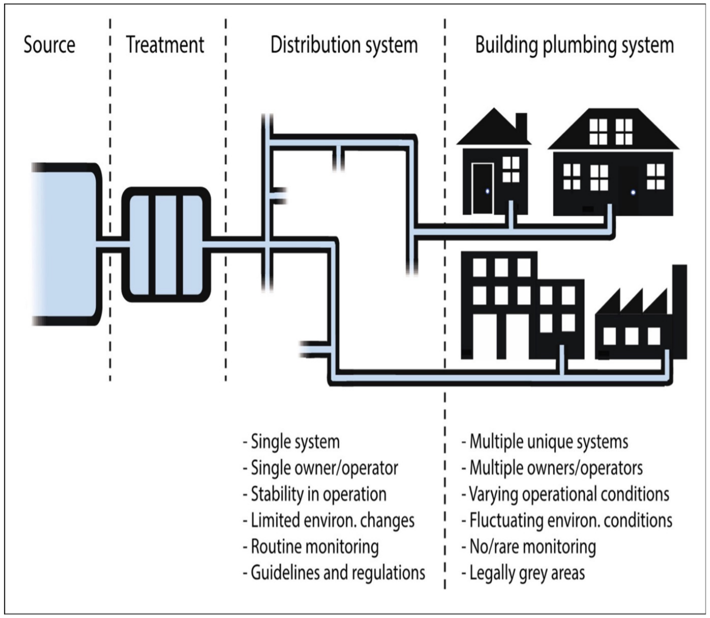
Using Data Science to Protect Residential Water Quality

R. Nerenberg, M. Sisk, M.D. Lemmon, E. Clements, Y. Duan



Motivation and Need

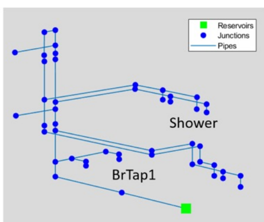
- Water utilities capture raw water, treat it to EPA standards, and distribute it to users via a piped network.
- While utilities must comply with EPA standards up to the user's connection, conditions in residential distribution networks can degrade water quality out of the tap
- These conditions include chlorine dissipation, disinfection by-product formation, leaching of toxic metals from pipes, leaching of organic chemical from plastic pipes/fittings, and the growth of microbial biofilms on pipe walls
- These negative impacts are correlated to "water age" or water residence time, which will in turn will be a function of the age, usage, and condition of the residential water distribution network.

Problem Statement

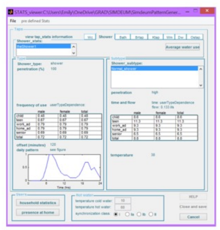
Can data science methodologies identify neighborhoods with the greatest risk of poor water quality and then use this knowledge to develop practical community strategies for mitigating this risk.

Using Data Science to Protect Residential Water Quality

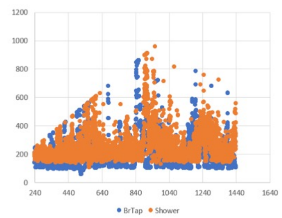
R. Nerenberg, M. Sisk, M.D. Lemmon, E. Clements, Y. Duan



EPANET plumbing network model



SIMDEUM stochastic water demand generator



Simulated water ages for tap and shower, as a function of time

Study 1- Identifying Potential Risk

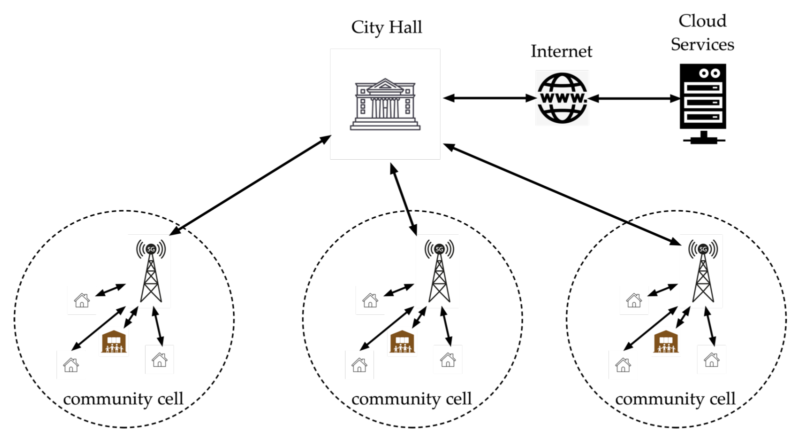
(Nerenberg, Sisk, and Clements)

- Use GIS (Geo. Info Sys) records of home age/size and water use records from water utilities to identify homes with high water age.
- Use selected homes to validate the correlation of our data sources (home age, size, water usage) with empirical measurements of residential water age/quality

Study 2- Mitigating Risk

(Nerenberg, Lemmon, Clements, Duan)

- Use **federated learning** techniques to train models for water age as a function of residence and occupant profiles. Datasets for these models will be based on simulation modeling of user demand and household plumbing flows.
- Challenges involving non i.i.d. sampling of neighborhoods and privacy of residential data will be addressed using **fair federated learning framework** realized through Generative Neural Networks.
- Cloud side model to be used to develop community wide risk mitigation policies that are statistically fair.

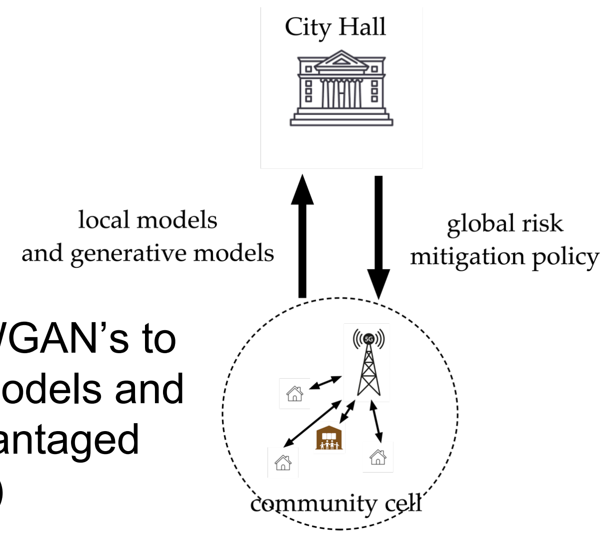


Fair Federated Learning Framework

(study 2 - Duan, Lemmon)

The **Fair Federated Learning** framework has edge devices learn

1. Local classifier, $\boldsymbol{\eta}_k$, that predicts water age based on residence profile.
2. Local generative neural network (WGAN), G_k, for the community's data distribution



The community cloud server uses the data, generated by the WGAN's to train a model, $\boldsymbol{\eta}$, for water quality that minimizes MSE of local models and the statistical fairness (risk difference) between the socially advantaged community (SAC) and socially disadvantaged community (SDC)

$$\boxed{\overbrace{\frac{1}{N}\sum_{k=1}^{N}(\eta_k(\hat{x}) - \eta(\hat{x}))^2}^{\text{MSE}}} + \boxed{\overbrace{\alpha\,|P(\eta(\hat{x}) = 1\,|\,\mathrm{SAC}) - P(\eta(\hat{x}) = 1\,|\,\mathrm{SDC})|}^{\text{Risk Difference}}}$$

Fair Federated Learning Framework – preliminary results

(study 2 - Duan, Lemmon)

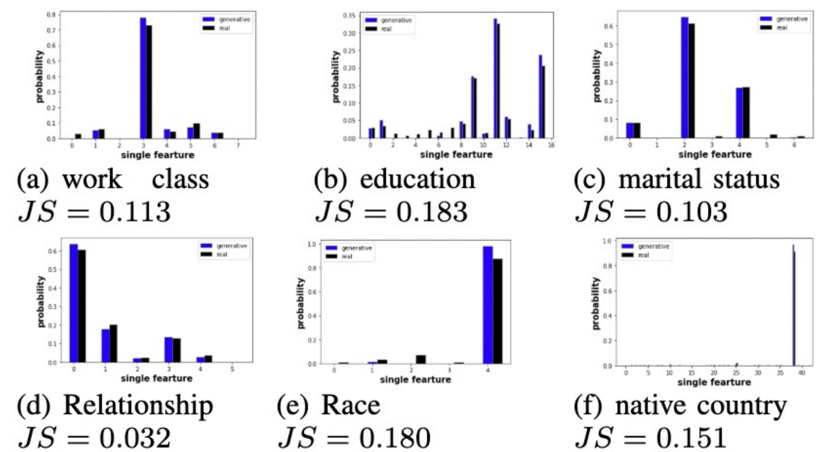
Preliminary results were obtained with our Fair Federated Learning Framework and the UCI adult database

- 14 categorical features (workclass, education, etc.)
- 1 binary label income (high,low)

Dataset split into a SAC and SDC group

- Male group (SAC)
- Femail group (SDC)

Local WGAN Accuracy



(a) work class $JS = 0.113$ (b) education $JS = 0.183$ (c) marital status $JS = 0.103$

(d) Relationship $JS = 0.032$ (e) Race $JS = 0.180$ (f) native country $JS = 0.151$

without fairness regularization

Accuracy = 85% **Risk Difference = 0.18**

with fairness regularization

Accuracy = 81% **Risk Difference = 0.01**