Modernizing Water and Wastewater Treatment through Data Science Education and Research (MoWaTER)

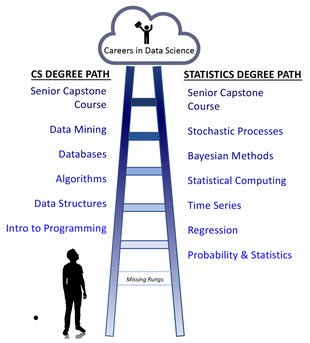
NSF HDR:DSC (#1924146) 09/2020 to 08/2023

- Amanda S. Hering (Baylor, Statistical Science, mandy_hering@baylor.edu)
- Tzahi Y. Cath (Mines, Environmental Engineering, tcath@mines.edu)
- Doug Nychka (Mines, Applied Mathematics & Statistics, nychka@mines.edu)
- Michael Poor (Baylor, Computer Science, michael_poor@baylor.edu)

Overview

This Data Science Corps Award is designed to:

1. Provide an early, motivating entry point where students are exposed to data science.
2. Link students with real datasets from high-impact problems.
3. Improve data-driven decision-making in the water and wastewater treatment (W/WWT) industry



We have built the following programs to meet these needs:

- An entry-level, inquiry-driven introduction to data science course
- A five-week summer data science fellows research program
- An industry workshop for professionals to learn R and data science
- An advanced summer internship program
- A one-week workshop for advanced engineering students

The W/WWT industry has an abundance of data but lacks the human capital to analyze it. Water scarcity can be reduced and water quality improved by improving operations with data-driven decisions.

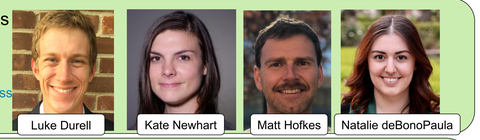
Program Elements

FACULTY

Not shown: Greg Hamerly, Grant Morgan, Jeanne Hill

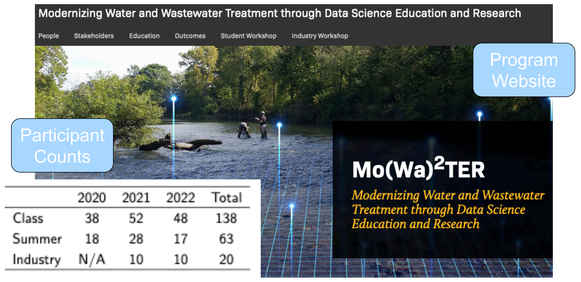
GRAD TAs

Not shown: Aurora Keras, Tayler Elwell, Mason Manross, Sweta Rai



STAKEHOLDERS

SOUTHERN NEVADA WATER AUTHORITY, sentryco, DENVER WATER, Hazen, Colorado Springs Utilities, It's how we're all connected, CITY OF BOULDER, ORANGE COUNTY WATER DISTRICT, SINCE 1933, AQUA-AEROBIC SYSTEMS, INC., A Metawater Company, METRO WATER RECOVERY, RAMEY ENVIRONMENTAL COMPLIANCE, BUREAU OF RECLAMATION, NAVI, NREL



	2020	2021	2022	Total
Class	38	52	48	138
Summer	18	28	17	63
Industry	N/A	10	10	20

Introduction to Data Science Course

The course is

- Developed and team-taught by statistics & computer science faculty and graduate students
- Has no pre-requisites, offered at sophomore level
- Is flipped with videos to watch ahead, and in-class exercises completed in small groups in Zoom break-out rooms
- Inquiry-driven with authentic water-related datasets
- Offered concurrently at both Baylor and Mines
- Students recruited from a diverse array of existing programs
- Goal: Increase the number of students majoring, minoring, or double-majoring in a data science related field

Summer Data Science Fellows Program

The five-week summer program is

- Data-based projects from stakeholders related to water and wastewater treatment, designed as a "pre-REU" experience
- Students work in teams of 3-5 work to solve a real problem
- Goal: Increase the number of students participating in industry internships or REUs in subsequent summers

Weekly Structure:

Pre-program: Faculty solicit data/projects from stakeholders. TAs do initial data cleaning and outline of project goals. R bootcamp for students to improve R skills.

Week 1: Build teams. Teams will load and explore the data, explain what the goal is for each project and give a description of the process that the data were generated from.

Week 2: Identify modeling approaches appropriate for the data. Students may receive reading material or sample code from grad student mentors. Meet with stakeholders to ask clarifying questions.

Week 3: Apply methods to project. Code review: swap code with another team and review.

Week 4: Apply methods to project. Meet with stakeholders to show preliminary results and receive feedback.

Week 5: Prepare 4-page tech brief and final presentation. Catalogue and archive project data and R code. Prepare data and sample code for archiving.

Quotes from Students:

"I loved this class so much that I am now interested in going to grad school to get a masters in data science."

"I initially thought I might not learn from the class much, as it was structured with no prerequisites, and I'm a senior math major with a statistics minor. I was very wrong though! I got crucial information really on how to take uncleaned data and get it manipulated in order to do complex statistical analysis. I would say in all my stats classes, we already work with pre-cleaned data sets that are pretty much ready to go already. Before this class, I would have no idea what to do with data that hadn't been cleaned yet.

"The lecture R-markdown files and video lectures helped immensely with exam preparation. The exams cover exactly what is talked about in each of the lectures and I really appreciated that! The professors were all so helpful in breakout rooms and I really enjoyed the aspect of breakout rooms for this class. Online courses are very hard to manage, so being able to have this opportunity to work with students altogether is very special. The professors' dynamic in this class was wonderful, and it made every class period fun and interactive."

Survey: Was this a Valuable Experience?

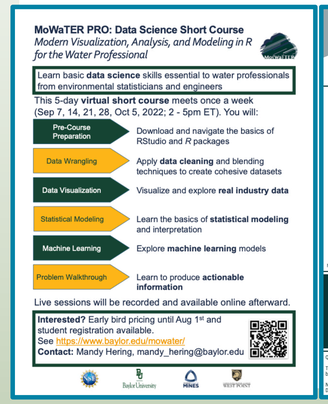
*On a scale of 1 to 10 (10 highest), the average ranking from 47 students was **9.45/10**!*

Sample Schedule
Intro to R
Intro to R
Intro to R
R Markdown
R Markdown
EDA
EDA
EDA
EDA
Teamwork
Data Wrangling
Data Wrangling
Programming
Programming
Graphics
GitHub
Written Presentations
Graphics
Shiny Apps
Shiny Apps
Regression
Regression
Project
Variable Selection
Feature Generation
Clustering
Classification
Model Validation

New Programs

Three new program components have or are being developed:

1. A $1,000 short course in data science for industry professionals
2. A one-week workshop for environmental engineering grad students
3. An advanced eight-week internship with stakeholder partners



Publications in Process

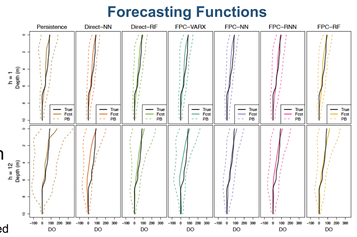
Education Publications:

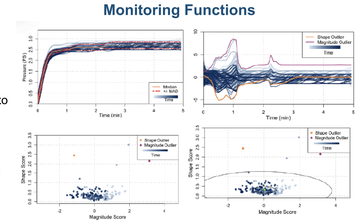
1. Course development and evaluation over 3 years
2. Textbook aggregating our course notes, class exercises, & exams
3. Summer research program structure and evaluation over 3 years
4. Data pipeline: acquisition, wrangling, and archiving
   - We have over 20 datasets from stakeholder partners.
   - We have created a dataverse on the Harvard Data Repository.
   - Metadata, R scripts, and clean datasets are being shared.
5. Lessons learned regarding the nuts and bolts of multi-university teaching
6. Student evaluation through self-written letters of recommendation

HARVARD Dataverse — mowateR — Mo(Wa)²TER Datasets (Baylor University)

Research Publications:

Functional data analysis has repeatedly proved useful because many W/WWT processes have variables that are "cyclic," and supported TAs are developing novel research methods.





1. Durell, L., Scott, J. T., Nychka, D., and Hering, A. S. (2022) "Functional forecasting of dissolved oxygen in high-frequency vertical lake profiles," Now Online in *Environmetrics*.
2. Kuras, A., Cath, T. Y, and Hering, A. S. "Functional data analysis approach for detecting faults in cyclic water and wastewater treatment processes," Submitted to *Environmental Science and Technology: Water*.
3. Durell, L., Scott, J. T., and Hering, A. S. "Hybrid forecasting for functional time series of dissolved oxygen profiles," Revision Submitted to *Data Science in Science*.