Collaborative Research: Elements: Advancing Data Science and Analytics for Water (DSAW)

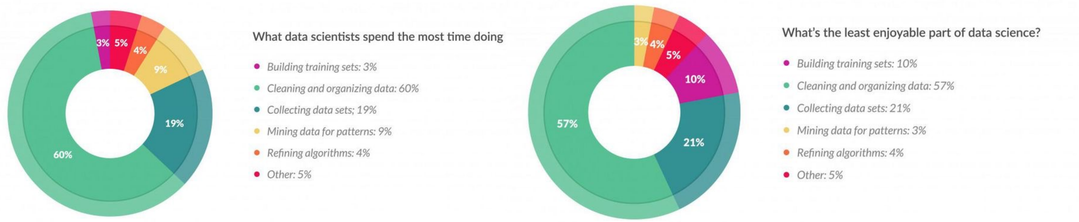
PI: Jeffery S. Horsburgh (jeff.horsburgh@usu.edu)[a], **Co-PIs**: Brian Crookston[a], Alfonso Torres-Rua[a], Tianfang Xu[b], Anthony Castronova[c]
Institutions: [a]Utah Water Research Laboratory, Utah State University; [b]Arizona State University; [c]CUAHSI

Award #: 1931297

1. Motivation

- Challenges in the water domain are multi-disciplinary, requiring data synthesis
- Data manipulation, visualization, and analysis tasks are difficult because datasets are becoming larger, more numerous, and more complex
- Standard data formats for common data types are not always agreed upon or mapped to efficient structures for visualization and/or analysis within an analytical environment.
- Researchers and practitioners lack training in data intensive scientific methods that would enable them to use new and existing data science tools to efficiently tackle large and/or complex datasets
- Overcoming barriers with accessing, organizing, and preparing datasets for data science intensive analyses will be an enabler for transforming scientific inquiries in the water domain.

What do data scientists spend their time doing?



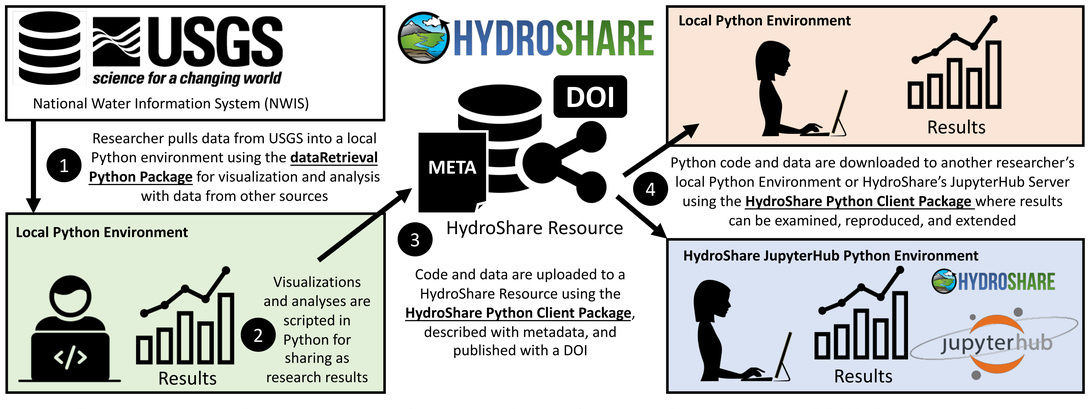
It remains hard to get data and get it into the format you want/need

Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says

https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/?sh=23ad2a316f63

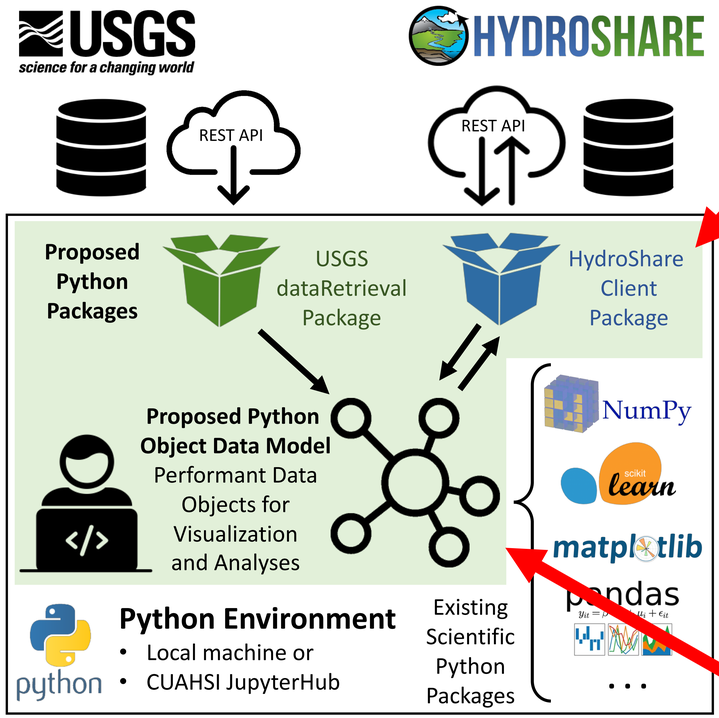
2. DSAW is focused on:

1. Enabling water-data scientists to more easily share and collaborate around data and analyses
2. Providing data management, visualization, and analysis tools that advance scientists' data science capabilities
3. Promoting more consistent data workflows, data reuse, and reproducibility of scientific results.



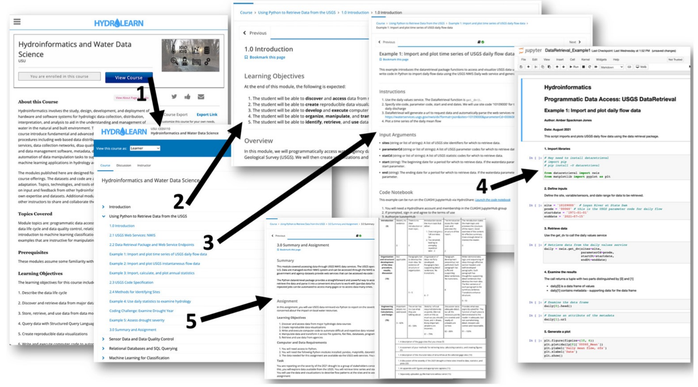
Example reproducible water-data science workflow using HydroShare and new DSAW Python packages.

3. Objectives and Outcomes

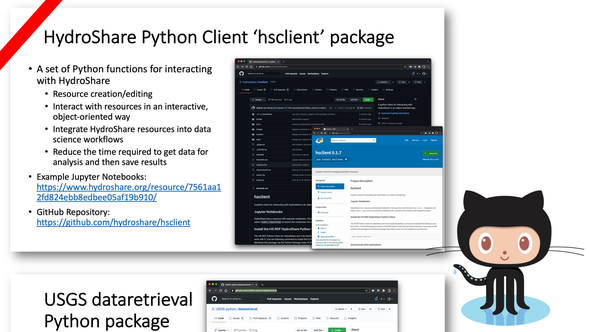


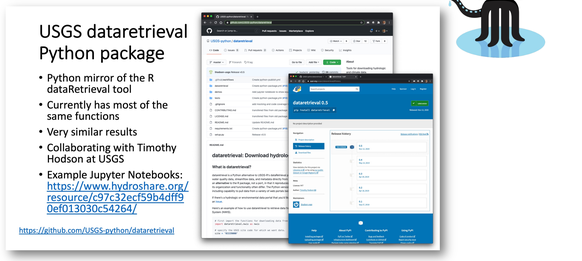
DSAW software architecture. New software elements are highlighted in green.

<u>Water-data science applications and educational modules</u> that demonstrate advanced analytical workflows using the tools we develop and help deliver the software to the water science community for use by educators and students

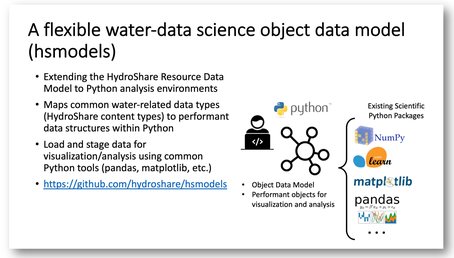


<u>New Python packages</u> that integrate performant data structures with data science capabilities in Python and advanced data access, collaboration, and data archival capabilities of HydroShare





<u>An advanced object data model</u> that maps common water-related data types to performant data structures within the object-oriented Python language and analytical environment



HYDROLEARN

Educational module implementation in HydroLearn: Numbered steps indicate the workflow and location of essential module elements: (1) course landing page and links to course outline, (2) learning objectives, (3) module narrative, (4) code examples as interactive notebooks in the CUAHSI JupyterHub linked from HydroLearn, and (5) the technical assignment and associated rubric.