International Society of the Learning Sciences

Rivulet: A Framework and Tools for Fetching Pedagogically Generative Scientific Datasets

Michelle H. Wilkerson, Berkeley School of Education, California, USA, mwilkers@berkeley.edu
Adelmo Eloy, Universidade de Sao Paulo, Brazil and Teachers College, Columbia University, New York, USA, adelmo@fablearn.net
Helen L. Fitzmaurice, Berkeley School of Education, California, USA, hfitzmaurice@berkeley.edu

Abstract: *Rivulet* is a framework and collection of Python Jupyter notebooks for fetching pedagogically generative scientific datasets. They provide executable templates for accessing, filtering, and evaluating data from public databases using Application Programming Interfaces (APIs) and expert insights regarding quantitative features of scientific interest. Our primary goal is to support educators and curriculum designers finding data that support meaningful statistical and scientific engagement through relevance to both students and to the domain.

Motivation

The promise of "Goldilocks data" (Rubin, 2019), data that is not too big nor too small but just right for learners, has captured the imagination of educators. The continued use of classic datasets such as Old Faithful Geyser eruptions (Shaughnessy & Pfannkuch, 2002) or the "100 Cats" dataset in (Konold et al., 2007) highlights how conceptually rich data can inspire meaningful statistical questions (e.g., "How faithful is Old Faithful?"). But their repeated use also highlights how difficult finding such generative datasets can be.

At the same time, educators and researchers are calling for more up-to-date, relevant, and authentic datasets for students (NASEM, 2023; Rosenberg et al., 2020; Weiland & Williams, 2023). There are calls for submission and guidelines (Krishnamurthi et al., n.d.) for building curated lists for identifying educational datasets (e.g., DASL, n.d.; *SAS Data Depot*, n.d.). However, static datasets cannot be easily updated or tailored to students' or curricular needs. A more flexible approach introduces teachers or students to Application Programming Interfaces (APIs) to make queries to live databases through code or specially-formatted URLs (Fergusson & Wild, 2021; Israel-Fishelson, Moon, Tabak, et al., 2024; Walker et al., 2024), but it is difficult to ensure that these data offer the resolution and robustness needed for rich statistical questions. And, there are ongoing challenges in finding data that can support connections between data and context (Penuel et al., 2025; Sagrans et al., 2022).

The [Toolname] Project seeks to merge the flexibility of APIs with ways to ensure that these datasets are robust enough to support key statistical and topical learning goals. To work toward this balance, we are building Python notebook utilities, developed with input from domain experts, that allow users to identify relevant constraints (e.g. places or times of interest), and download robust datasets that illustrate generative patterns within those constraints. These interactive documents integrate customizable API calls to robust database systems, a series of Python methods to filter and evaluate fetched data for key patterns, and discussion of how the fetched data address statistical and curricular learning goals, and explanatory text and instructive comments throughout.

Design principles

Our project draws from the learning sciences, statistics education, and data science education to inform the design of tools for fetching educational datasets. Specifically, we emphasize three key design principles.

First, we support *relevance and relationship building* to establish the value of data as a meaningful tool for investigation. Rivulet notebooks demonstrate how queries can be tailored to specific geographies, time periods, events, or comparisons. We encourage and model that such datasets be paired with news stories, local hands-on investigations, or other sources of information that allow students to learn more detail about focal patterns in data.

Second, we support the *statistical generativity* of a dataset to demonstrate patterns that are conducive to statistical analysis. Often, though large quantities of data are available for a given system, the specific indicators that are most likely to exhibit relevant patterns or make clear connections to content explorations can be more difficult to find, or may not be collected at a resolution that is conducive to analysis (e.g., monthly averages versus hourly samples). Rivulet utilities pre-emptively scan datasets for desired characteristics (such as the presence of high-leverage indicators, strong expected "signal" patterns as identified by experts, high data resolution, and/or patterns of variability and comparative measures across groups that are sufficient for meaningful analysis).

Third, we support the *disciplinary generativity* of a dataset, identifying what are the key patterns and related disciplinary concepts that can be examined in depth related to the data available through a given API, and what indicators and quantitative characteristics are necessary to unpack that specific idea. We consult with experts

 © ISLS

regarding what are "signature patterns" that clearly illustrate key disciplinary ideas, support users in reviewing subsets of data to determine whether those patterns are evident and can be meaningfully analyzed, and identify corresponding statistical questions that can motivate investigations of those patterns.

These principles are integrated and communicated through *computational notebooks*, a type of interactive document that mixes executable and editable code with rich text. Because the code presented is editable, users can customize their queries (Hüsing, et al., 2024). The design takes advantage of this flexibility to instruct users on how to tailor queries to access datasets that are relevant, and then to help users filter their results to identify datasets that have appropriate resolution, variability, and/or differences to warrant deeper analysis.

An example: Sourcing pedagogically useful air quality data with AQS.ipynb

The Air Quality Index (AQI) is important, but simply demonstrating that AQI goes up or down over time does not highlight how data can lend new insights. The AQS notebook was developed with a consulting scientist whose area of specialty involved neighborhood-scale air quality and guides users to fetch datasets that allow for deeper engagement with air quality, related environmental and atmospheric science, and social impacts.

The U. S. Environmental Protection Agency's Air Quality System or "AQS" is a longitudinal database of air quality data collected by U. S. agencies that has been evaluated and verified for accuracy. In Part I of the AQS notebook, users are instructed on how to acquire an API, connect to the AQS service, and query the status of the service to make sure it is online and able to provide requested data. Next, users identify a bounding box using latitude and longitude coordinates that defines a geographic region from which data should be fetched. They also identify a "target date," which will serve as the center of a time interval of interest. The default date is June 6, 2023, when several large cities across United States experienced historically high levels of air pollution from a complex of wildfires in Canada. Users are encouraged to edit the date to make sense for their context.

Once users have successfully connected to the API and defined their preferred settings, the notebook offers several query templates focused on particular scientific or statistical learning goals. These include:

(1) *Investigating Pollutants.* Exploring the dynamics of air pollutants can motivate investigations of the underlying chemical and behavioral complexity of air pollution. For example, ozone (O3) often demonstrates a daily cyclical variation in concentration as a secondary pollutant that is generated when other pollutants interact with heat and light. In contrast, particulate matter levels (PM) rise after events that release particles into the air. The notebook models a query that identifies air quality monitors within the specified location and time range that report *both* Ozone (O3) and particulate matter (PM2.5) levels at high resolution, so that differences can be identified in the data. Users can edit the code to specify other pollutants, though this combination often yields dramatic differences in pattern that can support robust hypothesis generation ("Why does Ozone go up and down while PM follows a different pattern?" or "Why did PM rise more than Ozone during the wildfire?") and examination of primary versus secondary air pollution and other key atmospheric chemistry concepts.

(2) *Exploring Differences in Air Quality.* Air quality can vary dramatically across even close geographic neighbors, often due to a complex combination of geographic features, weather patterns, human behaviors, zoning and development, and other social and scientific causes. The AQS notebook allows users to specify whether they are interested in composite AQI or a specific pollutant for a given region. They are then given a list of monitoring sites within the region that report these values at least daily, ordered by concentration so that data that demonstrate stark differences can be identified and fetched for comparison and investigation.

(3) *Unpacking AQI.* The Air Quality Index is a composite index that tracks six main pollutants–ground level ozone, particulate matter (PM2.5 and PM10), carbon monoxide, sulfur dioxide, and nitrogen dioxide levels. The reported AQI reflects only the maximally hazardous pollutant concentration at a given time; this may be different (e.g. Ozone versus PM2.5) within the same day. The AQS notebook identifies monitoring sites within the specified day and time that report as many of the AQI-constituent pollutants as possible. Once a site is selected, it fetches the dataset and calculates the corresponding hourly AQI, supporting investigations into the composite measure ("Is the air really equally healthy if two sites that report the same AQI? Why or why not?").

Accessing Rivulet notebooks and next steps

Rivulet notebooks are available for free under the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) License and can be accessed at https://github.com/CalCoRE/rivulet-utils. The notebooks are undergoing iterative development as we consult with expert advisors and as they are used as a part of active curriculum development. In addition to national and internationally-focused notebooks on air quality, we are developing notebooks focused on water quality using the United States Geological Survey's Water Quality Exchange or WQX API, and weather data using the U. S. National Weather Service API. We look forward to learning what might be other areas of need or opportunity for this work, and to inviting collaboration.

References

Arnold, P., Bargagliotti, A., Franklin, C., & Gould, R. (2022). Bringing Complex Data Into the Classroom. *Harvard Data Science Review*, *4*(3). https://doi.org/10.1162/99608f92.4ec90534

Arnold, P., & Franklin, C. (2021). What Makes a Good Statistical Question? *Journal of Statistics and Data Science Education*, *29*(1), 122–130. https://doi.org/10.1080/26939169.2021.1877582

DASL. (n.d.). *Data and Story Library*. Retrieved October 20, 2025, from https://dasl.datadescription.com/

Fergusson, A., & Wild, C. J. (2021). On traversing the data landscape: Introducing APIs to data-science students. *Teaching Statistics*, *43*(S1), S71–S83. https://doi.org/10.1111/test.12266

Hüsing, S., Schulte, C., Sparmann, S., & Bolte, M. (2024). Using Worked Examples for Engaging in Epistemic Programming Projects. *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1*, 443–449. https://doi.org/10.1145/3626252.3630961

Israel-Fishelson, R., Moon, P., Tabak, R., & Weintrop, D. (2024). Understanding the Data in K-12 Data Science. *Harvard Data Science Review*, *6*(2). https://doi.org/10.1162/99608f92.4f3ac3da

Konold, C., Harradine, A., & Kazak, S. (2007). Understanding Distributions by Modeling Them. *International Journal of Computers for Mathematical Learning*, *12*(3), 217–230. https://doi.org/10.1007/s10758-007-9123-1

Krishnamurthi, S., Schanzer, E., Schneider, D., & Drozda, Z. (n.d.). *Datasets for K-12 Data Science*.

[NASEM], N. A. of S., Engineering, and Medicine. (2023). *Foundations of Data Science for Students in Grades K-12: Proceedings of a Workshop*. The National Academies Press. https://doi.org/10.17226/26852

Penuel, W., Rubin, A., Puttick, G., Henson, K., & Deverel-Rico, C. (2025). Balancing design goals for supporting students' work with extant data in science. *Educational Designer*, *19*.

Rosenberg, J., Edwards, A., & Chen, B. (2020). Getting messy with data: Tools and strategies to help students analyze and interpret complex data sources. *The Science Teacher*, *87*(5), 30–35.

Rubin, A. (2019, April 1). Facebook or Instagram? Teens Explore Data About Technology Use. *Hands On! Magazine*. https://www.terc.edu/facebook-or-instagram-teens-explore-data-about-technology-use/

Sagrans, J., Mokros, J., Voyer, C., & Harvey, M. (2022). Data Science Meets Science Teaching. *The Science Teacher*, *89*(3), 64–69. https://doi.org/10.1080/00368555.2022.12293671

SAS Data Depot. (n.d.). Retrieved October 20, 2025, from https://curiosity.sas.com/en/resources/data-depot.html

Shaughnessy, J. M., & Pfannkuch, M. (2002). *How Faithful Is Old Faithful? Statistical Thinking: A Story of Variation and Prediction*. https://doi.org/10.5951/MT.95.4.0252

Walker, J. T., Barany, A., Acquah, A., Reza, S. M., Guzman, K., Johnson, M., Badreldin, O., & Barrera, A. (2024). *Sandbox Data Science: Culturally Relevant K-12 Computing*.

Weiland, T., & Williams, I. (2023). Culturally Relevant Data in Teaching Statistics and Data Science Courses. *Journal of Statistics and Data Science Education*, 1–16. https://doi.org/10.1080/26939169.2023.2249969

Wilkerson, M. H., & Polman, J. L. (2020). Situating data science: Exploring how relationships to data shape learning. *Journal of the Learning Sciences*, *29*(1), 1–10. https://doi.org/10.1080/10508406.2019.1705664

Authorship Statement

MHW initiated, conceptualized, and wrote the paper. MHW and AE built the notebooks. HF served as the science expert for the air quality notebook described. All authors reviewed and approved the paper.

Acknowledgments

This work is partially supported by a grant from the National Science Foundation (DRL-2445609). This project emerged in response to data needs and discussions during collaborative curriculum development with Tamar Fuhrmann, Leah Rosenbaum, Lucas Coletti, and Aditi Wagh and has benefited from additional input from Melanie Boudreault and Paulo Blikstein.