Los Angeles County Department of Public Health

Assignment location:
Los Angeles, CA

Los Angeles County Department of Public Health
Acute Communicable Disease Control/Epidemiology and Data Unit

Primary Mentor: Robert Mitchell, Master of Science, Information Science (MSIS)
Data Science Supervisor for Data Science and Systems Team
Los Angeles County Department of Public Health, Acute Communicable Disease Control

Secondary Mentor: Rebecca Fisher, MPH, MA
Chief Epidemiologist
Los Angeles County Department of Public Health, Acute Communicable Disease Control

Work Environment: 100% Virtual

Assignment Description

Agency Description:
The LAC DPH provides public health services to over 10 million residents. It includes 39 programs that provide services in the areas of environmental health, chronic disease and injury prevention, communicable disease control and prevention, substance abuse prevention and control, community health services, health facilities inspection, health assessment, and more. The Communicable Disease Control & Prevention Bureau houses the ACDC program that includes several units, such as the EDU that is led by Chief Epidemiologist, Rebecca Fisher. Within the EDU, several additional teams exist that include the Data Science & Systems (DSS) Team, Data Integration Team, eCR team, Morning Data Team, Lab Data Operations Team, and the Cluster Detection Team. The host team of the fellow would be the DSS Team that works in collaboration with many of the other EDU teams as well as other ACDC units, other programs in the Bureau of Disease Control, and the DPH Public Health Information Services (PHIS) team.

LAC DPH prioritizes improving the health of LAC residents as well as promoting health equity, and the ACDC program accomplishes that through the control of acute communicable diseases (excluding TB, STDs, and HIV). The DSS Team was established to provide additional data science, informatics and modernization support in improving the use of data within LAC for disease surveillance and case management activities. The DSS Team is a 100% remote team that uses the Scrum/Agile framework that includes daily standup meetings, weekly or bi-weekly project-specific meetings, a bi-weekly Learning
Session, a bi-weekly Retrospective meeting, and a bi-weekly Sprint Planning meeting. The team values collaboration, communication, flexibility, creativity, personal/professional growth, and cultivating a learning-centric environment.

Describe Statistical and Data Analysis Support, Such as Databases, Software, and Surveillance Systems Available to the Fellow:

Currently we are using Posit (formerly RStudio) professional tools: Posit Workbench, Posit Connect, and Posit Package Manager. Workbench is our main IDE, which provides RStudio, JupyterLab, Jupyter, and VSCode environments. Connect is where we deploy reports, APIs, dashboards, etc in both R and Python. Things deployed to Connect can also be scheduled and automated. These are deployed on premises on Red Hat Enterprise Linux. We also have development and production Apache Airflow and Apache Spark clusters that are used for more computationally intensive scripts. We have a combination of Microsoft SQL Server and PostgreSQL databases available within the enterprise and are continually exploring other approaches and solutions. Additionally, we are expected to begin a gradual move towards the cloud and are working on deploying our eCR/ELR integration on Azure, which will also leverage Azure’s FHIR converter. This is the project in collaboration with USDS and CDC, which will be an event driven pipeline that functionally expects and produces FHIR payloads. Each element will be containerized, and we’re excited to continue to move towards containerization. Although we’re not at the point where we’re able to move towards Kubernetes it is a goal of the DevOps guild members within DPH (currently made of of folks from PHIS and ACDC and VPDC Data Science members).

Any additional information about the placement:

The DSS Team is a 100% virtual team that utilizes Microsoft Teams and Slack to meet and communicate during typical 8-5pm PST workdays and a typical 40-hour workweek. Team members are located throughout the country on both the east and west coasts, and so most meetings are scheduled to end prior to 3/4pm PST to be mindful of the differences in time zones. The team uses the Scrum/Agile framework with structured meeting times at least once a day to address issues, meet with other teams/stakeholders regarding projects, meet with team members for collaboration, and to plan work for the sprint.

Describe the Preferred Background and Skills the Ideal Fellow For This Site Would Have:

The ideal candidate would have experience with data engineering, data analysis, and statistical or machine learning knowledge, but not necessarily an expert in any one specific area. The greatest need is in enabling others to be able to do more with new tools and approaches; therefore, our preferred candidate would be self-motivated, hungry to learn and grow, and interested in empowering others. Rather than trying to find someone that is a public health domain expert we find more success in collaboration with epidemiologists to build infrastructure and tooling that works for them and meets their needs.

Projects

Project 1 Title: Data Engineering and Data Science Project
Project objectives and expected deliverables: The purpose of this project is to engineer microservices and to deploy those services within our hybrid cloud/on-prem pipelines and to enhance our continuous integration and continuous delivery (CI/CD) model. These pipelines will be used to process eCR, ELR, FHIR, ADT and other data formats into our microservices architecture.

Expected public health impact from this project: Impact will be immense: there will be significant hours saved from the perspective of computation (code running on servers), individual members’ time (what they are currently spending to produce necessary outputs), and support hours (time required to troubleshoot, maintain, and update the infrastructure) to better support public health policy and programs in serving the residents of LAC.

Project 1 Data Engineering and Data Science Project

The purpose of this project is to migrate our data engineering pipelines from a virtual machine (VM) based architecture to a microservices architecture. This process will include the establishment of a microservices framework, the use of infrastructure as code (IaC), robust documentation with code walkthroughs on GitHub, and additional considerations should we move our computation entirely to the cloud.

The enterprise will be considering its approach in the coming year: to stick with hybrid cloud/on-prem or move towards cloud first, meaning we always deploy cloud unless there is a blocker that does not allow for cloud. Given this uncertainty in connection to our compute environment, this project will ensure that code works in both environments. To ease the burden of provisioning such complex data engineering artifacts we will use IaC to both document the “what” and the “how” our services are provisioned and deployed. This will allow us to automate the establishment of container registries, lifecycle management for containers, and deployment of development, testing, and production environments. Furthermore, the project will look at how we can move away from manual deployments to automatic testing and deployment—helping to establish a roadmap for the organization to close the gap on adoption and implementation of robust continuous integration and continuous delivery (CI/CD). These pipelines will be used to process eCR, ELR, ADT, and other HL7 data formats with considerable effort spent on leveraging FHIR to ensure each step has standardized inputs and outputs—it also means that the pipeline is extensible and able to work for many diseases.

The fellow will collaborate with other DSS Team members and contribute code via GitHub using a combination of these languages: Bash, Python, R, Spark (specifically PySpark), SQL, and Terraform. Our goal is to ensure raw and unprocessed data is validated and transformed in uniform and standard ways from source system to data pipeline ingress to enhancement and enrichment, to analytic focused endpoints. Rather than create a self-service reporting tool we will empower epidemiologists and research analysts to access clean and uniform data that has been enriched so that they can deliver reports, websites, statistical models, and dashboards. Specific deliverables will include:

1. An extract, transform, and load (ETL) pipeline that incorporates ELR, ADT, eCR messages received through our HL7 integration engine pipelines (Rhapsody and Mirth) into our main surveillance system utilizing FHIR; data sources for enhanced surveillance include:
   a. Reproducible scripts for all processes: provisioning, deployment, testing, development, production, et cetera
   b. Automation scripts using Apache Spark and Apache Airflow
c. Unit and integration testing scripts
d. CI/CD process using GitHub Actions
e. Process monitoring functions and alerting
f. Data Model Documentation for data sources (e.g., eCR, ELR, and ADT data sources) using an entity relationship diagram and data flow charts
g. Public code contributions to ongoing efforts, such as DMI, PRIME, CDC, and USDS projects, when applicable or the making of this project’s output publicly available on GitHub for others to leverage and improve.

2. A pipeline that builds data products from the raw source tables into data tables to be used as analytic endpoints, including:
   a. Reproducible scripts
   b. Automation scripts using Apache Spark and Airflow
c. Unit and integration testing scripts
d. CI/CD process using GitHub Actions
e. Process monitoring functions and alerting
f. Data Model Documentation for data products using an entity relationship diagram and data flow charts

3. Microservices (i.e. processes built into an application-like service) to incorporate standardizations and geospatial enhancements to ACDC ETL pipelines that includes deduplication processing, geocoding, and spatial enrichment processing.
   a. Process-specific scripts
   b. Containerfile/Dockerfile scripts and potentially Kubernetes deployment YAML for pods
c. REST API for data pipeline steps
d. Unit and integration testing scripts
e. CI/CD process using GitHub Actions
f. Process monitoring functions and alerting
g. Data process flow charts
h. Documentation that includes instructions for deployment and use

Project 2 Title: Research Software Engineering (RSE) Project

Project objectives and expected deliverables: This project will build collaborative environments in which the best software engineering practices are built into workflows that can be used by epidemiologists and analysts that haven’t been exposed to the software engineering domain or otherwise feel lost in the changing landscape of available tooling. We will work to formalize methodological approaches to reporting and analytics within ACDC into generalizable ad-hoc functions to be used by epidemiologists and analysts that enable them to get from analytic endpoints to feature rich reporting and analysis. The project will also focus on teaching and trainings working on building our analytics staff into RSEs that are comfortable leveraging the fruits of the data engineering and DevOps work in either R or Python.

Expected public health impact from this project: Standardizing methodological approaches and building data structures that can be passed between functions that mirror an in-house API design will help each and every analyst that spends time thinking of how to get from an analytic endpoint to typical
summarized variables near instantly. Rather than thinking of how to implement the logic in code, the inhouse API will be designed to accommodate the work they need to accomplish and therefore will provide functions that can arrive at insight faster. We anticipate that this work will save time in critical ad-hoc questions that often come from leadership, the media, or researchers being able to quickly modify and implement core utility functions on new data using validated methods will have an enormous impact on the organization at many levels.

### Project 2 Research Software Engineering (RSE) Project

This project will build collaborative environments in which the best software engineering practices are built into workflows that can be used by epidemiologists and analysts that haven’t been exposed to the software engineering domain or otherwise feel lost in the changing landscape of available tooling. We will work to formalize methodological approaches to reporting and analytics within ACDC into generalizable ad-hoc functions to be used by epidemiologists and research analysts that enable them to get from analytic endpoints to feature rich reporting and analysis. The project will also focus on teaching and training—working on building our analytics staff into RSEs that are comfortable leveraging the fruits of the data engineering and DevOps work in either R or Python.

The fellow will work with members of the Data Product & Project Management unit of the DSS Team to research business needs and technical requirements for developing reusable functions within packages for Python and R to complete standardized processes across ACDC Epidemiology and Data teams. This project will encompass product planning, development, testing, deployment, and adoption/training for users. The completed work can optionally also be presented at a Data Science Conference, pending work completion and acceptance into a conference to share the packages more broadly to the data science, epidemiology, informatics and analytic community.

1. Assessment of current analytic processes and methodologies used across ACDC Epidemiology and Data teams
2. Final Research Report documenting current analytic processes and methodologies used across ACDC Epidemiology and Data teams (e.g., crude and age-adjusted rates, varying time-period rolling averages, sampling algorithms, etc.)
3. Identify analytic processes and methodologies that are candidates for standardization and develop standardized functions within R and Python
   a. Project and scoping plan document
   b. Business and technical requirements documentation for all processes/methodologies
   c. Process scripts containing standardized functions in R and Python
   d. Testing plan, test cases, and acceptance criteria
4. Standardized functions to be included in the ACDC R and Python packages for analytic and reporting processes/methodologies
   a. Scripts/modules written in R/Python for each functionality
   b. R Package metadata description file
   c. R Object documentation file
   d. Python directory files
   e. Unit testing incorporated into scripts
f. Making the binaries available in the internal Posit Package Manager for easy download in both R and Python

5. Documentation for utilizing the functions within the ACDC R package and for it to be made publicly available for others to use.
   a. Documentation for describing the logic and use of each of the developed functions
   b. Training modules to help train analysts how to use the developed functions.

6. (Optional) Presentation of package functionalities at a Posit/Data Science Conference