**HOWARD UNIVERSITY RESEARCH CENTERS FOR MINORITY INSTITUTIONS**

**VIRTUAL APPLIED DATA SCIENCE TRAINING INSTITUTE, VADSTI**

**2023 SPRING TRAINING SERIES**

*Data Science Approaches to Better Understand Health Disparity & Equity Research*

**February 22, 2023 – April 6, 2023**

# About VADSTI

Technological advancements and efficient use of computational tools have made it possible to generate and store large amounts of heterogeneous and complex datasets in many disciplines, including public health, clinical, biomedical, and genomics. There is therefore increased demand for data analytics capabilities to look at trends, predict outcomes, and make better clinical and health policy decisions. Skill sets in data science are particularly critical for advancing the science of minority health and health disparities. The Howard University Research Centers in Minority Institutions, the [AIM-AHEAD program](https://aim-ahead.net/home/about/what), and the [Public Health Informatics Technology for DC (PHIT4DC program](https://www.phit4dc.com/) is pleased to announce VADSTI 2.0, Spring 2023 Training Series to the Howard University community of researchers and beyond. The goal is to enhance data science capability and application by providing training in the foundations of programming and critical data analytic skills for planning and conducting research involving big data pertinent to minority health and health equities. The Spring Training Series **is** project-based and will cover topics including Foundations of Data Science, Python, Data Preparation, Exploration and Visualization, and Cloud Computing, among others.

To register, click the following link. [Register Now](https://nam04.safelinks.protection.outlook.com/?url=https%3A%2F%2Fvadstispring23.eventbrite.com%2F&data=05%7C01%7Crcmi%40Howard.edu%7Cfdd07a0f361d4fa37e6f08db053bb933%7C02ac0c07b75f46bf9b133630ba94bb69%7C0%7C0%7C638109526386416905%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=poXHwmxDepO0L3KqhTcme57vUNhV6iG2ZNHOx3MDGWY%3D&reserved=0)

For questions, contact VADSTI at vadsti@howard.edu or John Kwagyan, Ph.D. at jkwagyan@howard.edu

## **Program Objectives & Competencies**

The primary objective of the 2023 VADSTI Spring Training Series is to provide training in data science fundamentals and cloud computing skills with hands-on application to minority health and health disparity datasets. Over the course of the training program, participants will:

* Be introduced to the foundations of data science.
* Be introduced to Python programming skills.
* Gain practical, hands-on experience with Python and related libraries for accessing data.
* Learn about the underlying concepts of probability and statistics for data analytics.
* Understand the concepts of data partitioning and practice behind supervised and unsupervised learning.
* Be introduced to cloud computing
* Be introduced to tools for applied data science using cloud-based platforms for clinical and genomic research.

**Digital Certificate of Completion:** Participants who complete all the modules and submit their projects in the VADSTI GibHub Data Science Project Portfolio will receive a verified digital certificate of completion.

**Evaluation:** At the end of each training module, you will be requested to complete electronic feedback forms on the extent to which expectations and objectives were met.

**Registration & Fees:** No fees for participation, but **registration is required** to attend.

## **VADSTI Training Program Schedule**

No prerequisite for research knowledge topics. Basic undergraduate knowledge of algebra and probability is recommended for content knowledge topics. The training series consists of the following modules.

**Past Training Recordings**

Participants are encouraged to review the lecture recordings of topics from [2022 Fall Training series.](https://courses.aim-ahead.net/course/c/VADSTI2022)

**Week 1:**

**Module 1**| **Foundations of Data Science with Python**

Wednesday, February 22, & Thursday, February 23, 2023

11:00 AM – 2:00 PM EST
**INSTRUCTOR – Moussa Doumbia, Ph.D.**

**This module will introduce you to the core principles of data science and python programming and associated libraries. You will be introduced to and learn how to use Jupyter notebooks. You will understand what data science and AI can currently do. An overview of the state-of-the-art methods will be introduced and real-life examples from clinical and healthcare data will be used for illustration.**

**Week 2:**

Wednesday, March 1, & Thursday, March 2, 2023

11:00 AM – 2:00 PM EST

**Module 2**| **Data Preparation, Exploration, and Visualization**
**INSTRUCTOR – Ebelechukwu Nwafor, PhD**
This module provides recipes for data preparation, exploration, and visualization, which are critical steps in any data science project. The goal of this module is for participants to learn how to visualize and perform initial investigations of the data to discover patterns, spot anomalies, test hypotheses, and check assumptions with the help of summary statistics and graphical representations. We will be using python to explore, filter, and manipulate various datasets; identify data anomalies and missingness; learn how to impute missing data; identify highly correlated variables.

**Week 3:**

**SPRING BREAK**| **SPRING BREAK**

 March 4 - March 11, 2023

**Week 4:**

**Module 3**| **Seminal Presentation on Health Disparity and Equity Research:**

Wednesday, March 15, & Thursday, March 16, 202311:00 AM – 2:00 PM EST

3.1: **Social Determinants of Health Data**| Wednesday, 11:00- 12:15 PM

 **Presenter:** Teletia Taylor, PhD

3.2: **Community Data Ownership**| Thursday, 11:00-12:15 PM

 **Presenter**: TBD

3.3: **Health Disparities, Inequities & Inequalities**| Wednesday, 12:30-2:00 PM

 **Presenter:** Kimberly Henderson, PhD

3.4: **Collaboration to Expand Health Equity Data to Improve Community Healthcare Outcomes|**

 Thursday, 12:30-2:00 PM **Presenter**: TBD

**Week 5:**

**Module 4A**| **Cloud Computing I**

Wednesday, March 22 & Thursday, March 23, 202311:00 AM – 2:00 PM EST
**INSTRUCTOR: Guy M. Lingani, PhD**
In this module you will learn about AWS Cloud concepts, core AWS services, security, architecture, pricing, and support to build your AWS Cloud knowledge. Summarize the working definition of AWS. Differentiate between on-premises, hybrid-cloud, and all-in cloud models. Describe the basic global infrastructure of the AWS Cloud. Explain the six main benefits of the AWS Cloud. Describe and provide an example of the core AWS services, including compute, network, database, and storage services. Identify an appropriate solution using AWS Cloud services for various use cases.

**Week 6:**

**Module 4B**| **Cloud Computing II**

Wednesday, March 29 & Thursday, March 30, 202311:00 AM – 2:00 PM EST
**INSTRUCTOR: Guy M. Lingani, PhD**
Cloud Computing II builds upon Cloud Computing I, and has participants, describe the AWS Well-Architected Framework. Explain the AWS Shared Responsibility Model. Describe the core security services within the AWS Cloud. Describe the basics of AWS Cloud migration. Articulate the financial benefits of the AWS Cloud for your organization’s cost management.

**Week 7:**

**Module 5**| **Tools for Applied Data Science Using Cloud-Based Platforms**

Thursday, April 5, & Friday, April 6, 202311:00 AM – 2:00 PM EST
**INSTRUCTOR – AnVIL Team**

The NHGRI Analysis, Visualization, and Informatics Lab-space (AnVIL) is a cloud-based platform that supports the management, analysis and sharing of biomedical data for the NHGRI research community and beyond. It aims to advance our basic understanding of the genetic basis of complex traits and accelerate discovery and development of therapies, diagnostic tests, and other technologies for diseases like cardiovascular disease or autism spectrum disorders. The platform currently hosts more than 150,000 whole human genome data sets, and offers a variety of analysis capabilities including: Terra for large scale computing and managing, analyzing, harmonizing, and sharing large datasets; Dockstore for sharing Docker-based analysis workflows; Jupyter notebooks for organizing live code, equations, visualizations and narrative text into a single document; RStudio for interactive machine learning, statistical computing, and visualizations; Bioconductor for community-driven interactive genomics with R; and Galaxy, for accessible, reproducible, and transparent genomic science. In this module, you will be introduced to the platform, tools and functionality for data science projects.