UNIVERSITY of VIRGINIA | Biocomplexity Institute & Initiative

DATA SCIENCE FOR THE PUBLIC GOOD: WORKFORCE DEVELOPMENT

The program goal is to build data science skills in young scholars across a broad set of fields of study and to create a data science workforce pipeline that is interested in civic engagement, this program focuses on "public good" projects. We engage young scholars in finding solutions to some of the most pressing social issues of our time. The program has operated since 2014. Below are summary data that describe the students and project sponsorship.

Table 1. DSPG Summary Data, 2014-2019

	2014	2015	2016	2017	2018	2019*	2014-19*
# DSPG Students	8	6	8	20	16	12	70
Bachelors	38%	67%	75%	80%	63%	58%	66%
Masters	38%	0%	13%	5%	19%	17%	14%
Doctorate	25%	33%	13%	15%	19%	25%	20%
# Male students	63%	67%	25%	50%	56%	33%	49%
# Female Students	38%	33%	75%	50%	44%	67%	51%
Math/Stat/Computer Science	50%	83%	50%	50%	50%	50%	53%
Social Behavioral Sciences	13%	0%	25%	10%	31%	50%	23%
Engineering	25%	17%	0%	10%	13%	0%	10%
Science	13%	0%	25%	15%	0%	0%	9%
Business	0%	0%	0%	15%	6%	0%	6%
# DSPG projects	7	4	12	14	9	9	55
# DSPG posters	7	4	14	15	11	9	60
# DSPG Sponsors	5	2	6	8	6	7	34

*estimated for 2019

Over the 6 years of the program, two-thirds of the students are undergraduates, 14% are masters level, and 20% are doctoral students. Although there has been some yearly variation, overall half the students are male and half are female. Over half of the students were majoring in mathematics, statistics, or computer science and about one-fourth of the students were majoring in social and behavioral sciences. The remainder were in science, engineering, or business. When the 2014-2016 cohort graduated**, 25% took civic positions in government or nonprofit organizations**, 20% went onto graduate school, and the remainder took jobs in the private sector, many with titles of data science.

The vertically integrated DSPG teams work across skill levels (undergraduate and graduate students, post-docs, faculty, and community sponsors) and are horizontally integrated across disciplines (social and behavioral sciences, science, and engineering). The DSPG teams have **completed 55 projects, sponsored by 23 unique local, state, and federal agencies and non-profit organizations (34 total sponsors, counting a sponsor more than once), producing 60 posters (estimated for 2019) presented at the DSPG Annual Research Symposium**. Arlington County, Fairfax County, NSF National Center for Science & Engineering Statistics, US Army Institute for Behavioral & Social Science Research, and US Census Bureau have sponsored more than one project, and some, many projects each year. Our **annual DSPG Symposium** that showcases the students and the team research has **attracted increasing numbers of participants from local, state, and federal agencies and non-profit organizations, companies, and individuals interested in our research.**

Training

DSPG trainings focus on learning data science methods and tools. The students engage in 47 classroom training modules covering the topics listed in Table 2. By the end of the program, the students have learned about and used the skills learned in their projects. All students have learned or increased their proficiency in using data science software and tools.

Table 2. DSPG Training Classes, 2019

Getting Ready	**Data Management**
1 SDAD Data Science Process	24 Storage, Security, Destruction
2 Data Security Requirements	25 Metadata
3 Introduction to the SDAD Computing Platforms	**Data Profiling**
	26 Structure
Code Management with Git	27 Quality
4 Naming Things	28 Metadata & Provenance
5 Project Setup, Git Setup, Git on your own	**Data Preparation**
6 Git with branches for collaboration	29 Restructuring
R Basics Boot Camp	30 Cleaning
7 R Notebooks	31 Transformation
8 Strings, Factors, Numerics, Date, Times	32 Joining
9 Data Objects	33 Deduplication
10 Iteration (Loops and Vectorization), Functions	**Data Exploration & Analysis**
11 Regex	34 Exploration & Visualization
Data Discovery	35 Mapping
12 FIELD TRIP: Library of Congress LABS	36 Text Mining
13 Literature Review	**Media Interaction**
14 Data Inventory & Screening	37 Media Writing
Research Ethics & Behavioral Experiments	38 Media Speaking
15 Research Ethics	**Data Modeling**
16 Brown Bag: Data Privacy & Stakeholder Expectations	39 Modeling: Social Network Analysis
	40 Modeling: Regression / GLM
17 Behavioral Experiments	41 Modeling: Classification & Clustering
Data Ingestion	42 Modeling: Decision Trees
18 From Files & APIs	**Data Presentation**
19 From Databases & Spatial Scraping Data	43 APIs
	44 Graphics Cookbook
20 Web-Scraping: HTML	45 Latex
21 Web-Scraping: Javascript	46 Shiny Documents & Dashboards, part 1
Transforming (& Creating) Data	47 Shiny Documents & Dashboards, part 2
22 Restructuring on Ingest	
23 Creating Data from Satellite Imagery	

Testimonials from DSPG Students

At the end of the summer, students are asked for their feedback about the program. Below are selected excerpts:

Favorite Part of the Program

- "My **favorite part of the DSPG program** was how the faculty, graduate students, and undergraduate students create a great atmosphere that everyone can work together very well, even they are not on the same projects."
- "My favorite part was the **relationship and the community we made during this summer with fellows, students, and faculty** … This community made data science more fun and engaging. I made very valuable connections and relationships with brilliant people in my own field of study."
- "Working with real, useful data. I loved the fact that my work was going to be used by someone else to help others."

Learning

- "I think the most useful for me was **learning how to code**"
- "I really enjoyed being able to **work with the undergrads**. I didn't have a lot of project management experience before and I felt that it was a very good **opportunity to think about how to teach others**."
- "I liked having the **freedom to direct the projects**."
- "Teamwork, time management, and organization."
- "I learned **how to use R, how to use Git, and how to web scrape** which became kind of something I turned out to be really good at because it took my prior knowledge of programming logic concepts and let me apply them to real problems."
- "I learned so much about **visualization and mapping**. I didn't think I would be producing any graphics this summer but I am glad that I learned so much about it."
- "Latex, SQL, Using piping in R, Dplyr, making prettier ggplots and web-scraping."

Team Science

- **"Working in teams was the best thing**. I liked how each individual had their own role or some roles could be shared among two people."
- "A **collaborative environment is an excellent approach for conducting research, especially with an interdisciplinary team**. People on my projects freely shared their ideas and through teamwork and discussion we were able to choose the best ideas and develop them into a conclusive product."