Practices of Big Data and Data Science Researchers at the University of Virginia: An Ithaka S+R Local Report

Jennifer Huck
University of Virginia Library
jhuck@virginia.edu

Jacalyn Huband
University of Virginia Research Computing
jmh5ad@virginia.edu

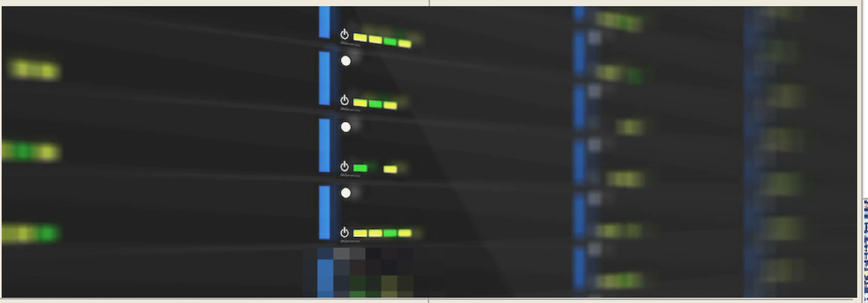
Introduction

We investigated the research practices of Big Data and Data Science Researchers at the University of Virginia. The study was conducted by a team of staff from UVA Library and UVA Research Computing. Ithaka S+R oversaw the study; 20+ other academic institutions participated.

Methodology

We interviewed researchers using Big Data or Data Science methods.

- 11 participants
- Variety of schools
- Variety of career stages
- Participants must have research duties



Data Acquisition

- Quality, Verification and Pre-processing: A considerable amount of time is spent verifying the data, attempting to discover errors, cleaning up non-numeric data, and understanding the dataset as best as they can.
- Challenges:
  - Funding proprietary data purchases
  - Receiving data in inconsistent formats
  - Downloading large datasets in a reasonable amount of time
  - Acquiring datasets from the medical system. Receiving data as periodic dumps, rather than an up-to-date stream

Ethics of Data Handling

- The biggest concern was that the advancement of machine learning algorithms could reveal more about individuals than was intended.
- Other concerns:
  - Are the results sufficient for making life and death decisions?
  - Can findings that are intended to help society be used for bad, instead of good?
  - Can machine learning algorithms have biases that promote gender or race inequality?

Select Findings

Sharing & Collaboration

- Many respondents often share code, but not necessarily data usually because of sensitive or proprietary nature

Challenges to Sharing

- Size of data makes sharing a challenge
- There is some concern about reviewers or other third parties recreating the compute environment, especially in disciplines that are not heavy in data scientists

Collaboration & Community

- The respondents are frequent collaborators, both within and outside of UVA

Strategies for Collaboration and Community

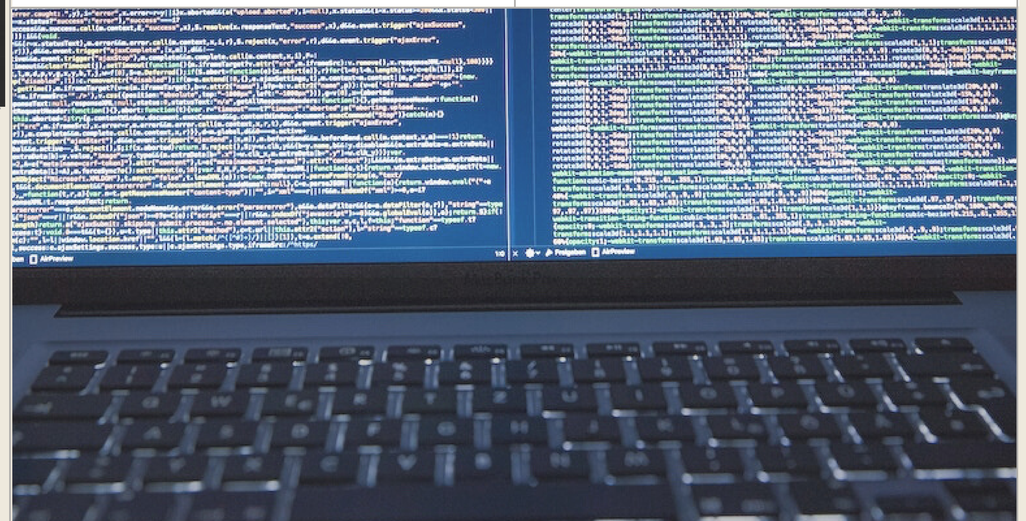
- Use a common IRB across institutions
- Rely on UVA HPC, esp. when outside collaborators do not have access to HPC at home institutions

Storage & Compute Infrastructure

- Wide variety of locations used (e.g., Dropbox, Google Drive, AWS, local servers, leased UVA storage)
- Consensus was that there does not exist a reliable solution for storing and maintaining Big Data at UVA

Challenges:

- Transfer times from storage to compute platform
- Speed of compute platform
- Cost of storage
- Local systems tend to go away when funding runs out or champions leave
- Most participants responded that they would not go outside of their team if they needed it, raising questions about UVA Library's Research Data Services and UVA Research Computing - are they simply not recognized as offering support or are these units are not providing needed services?



Recommendations

For research that involves **sensitive or proprietary data**, we recommend that:

- The Library and Research Computing provide more training for the handling of sensitive data, including what the different types of sensitivity are, where sensitive data can be stored, and where it can be processed.
- Research Computing plan for additional infrastructure where sensitive data can be stored and processed, to handle the future increase in sensitive data.

For **data acquisition**, we recommend that:

- The Library helps acquire proprietary datasets or datasets from industry (e.g., social media companies). The Library has the infrastructure in place to license and store proprietary datasets but would require funding from other units for data acquisition (for example, funding from faculty, Provost, or Vice President for Research).

For data and code **preservation**, we recommend that:

- The Library provides better support for LibraData to accept "big data," or market LibraData as being able to point to another site to find the files.
- The Library markets LibraData as able to accept code. The respondents seemed familiar with LibraData as a place to preserve data, but LibraData was never mentioned as a place to store code. GitHub and Zenodo came up most frequently as sites to share and store code.

For **data processing**, we recommend that:

- The Library reviews the proportion of course time dedicated to data cleaning in the various departments and schools. This is potentially a type of training the Library and Research Computing could offer more intensively. Continue to market the statistical consultation services we already provide.
- The Library and Research Computing explore tools that provide "workflow engines" that can tie together data, code, and computing resources.

For development of **community and collaboration**, we recommend that:

- The Library and Research Computing provide more interdisciplinary methods and data-focused workshops or community building events across Grounds. The Library would be the host organization.
- The Library and Research Computing provide workshops and outreach materials highlighting the strategies that big data researchers and data scientists use for successful collaborations, for example documentation, file naming, organization, version control, and cloud storage and document platforms.
- The Library invites the local IRBs to collaborate on a workshop about designing IRB protocols for collaborations across multiple institutions.

For **general knowledge**, we recommend that:

- The Library highlights journals and conferences that are disciplinary in nature but that embrace Big Data or Data Science methods.
- The Library and Research Computing continue to provide workshops for programming languages, such as R, Python, and MATLAB, but also include workshops and consultation service that show how to efficiently handle Big Data.