# High Performance Computing for Photogrammetry and OCR Made Easy

**Quinn Dombrowski**
quinnd@berkeley.edu
UC Berkeley, United States of America

**Tassie Gniady**
ctgniady@iu.edu
Indiana University, United States of America

**Megan Meredith-Lobay**
megan.lobay@ubc.ca
University of British Columbia, Canada

**John Simpson**
john.simpson@computecanada.ca
University of Alberta, Canada

Computationally-intensive research methods have seen increasing adoption among digital humanities scholars, but for scholars outside R1 institutions with robust computing environments, techniques like photogrammetry or text recognition within images can easily monopolize desktop computers for days at a time. Even at institutions with a research computing program, systems are configured for scientific applications, and IT staff may be unaccustomed to working with humanities scholars, particularly those who are not already proficient at using the command line. National compute infrastructures in North America (Compute Canada and XSEDE) are a compelling alternative, providing no-cost compute allocations for researchers and offering support from technical staff interested in and familiar with humanities computing needs. This workshop will start by introducing participants to Compute Canada and XSEDE, cover how to obtain a compute allocation (including for researchers outside of the US and Canada), and proceed through two hands-on tutorials on research methods that benefit from the additional compute power provided by these infrastructures: 1) photogrammetry using PhotoScan and 2) using OCR via Tesseract to extract metadata from images.

## Photogrammetry

Photogrammetry (generating 3D models from a series of partially-overlapping 2D images) is quickly gaining favor as an efficient way to develop models of everything from small artifacts that fit in a light box to large archaeological sites, using drone photography. Stitching photographs together, generating point clouds, and generating the dense mesh that underlies a final model are all computationally-intensive processes that can take up to tens of hours for a small object to weeks for a landscape to be stitched on a high-powered desktop. Using a high-performance compute cluster can reduce the computation time to about ten hours for human-sized statues and twenty-four hours for small landscapes. Generating a dense cloud, in particular, sees a significant performance when run on GPU nodes, which are increasingly common in institutional HPC clusters and available through Compute Canada and XSEDE.

One disadvantage of doing photogrammetry on an HPC cluster is that it requires use of the command line and Photoscan's Python API. Since it is not reasonable to expect that all, or even most, scholars who would benefit from photogrammetry are proficient with Python, UC Berkeley has developed a Jupyter notebook that walks through the steps of the photogrammetry process, with opportunities for users to configure the settings along the way. Jupyter notebooks embed documentation along with code, and can serve both as a resource tool for researchers who are learning Python, and as a stand-alone utility for those who want to simply run the code, rather than write it. Indiana University, on the other hand, has developed a workflow using a remote desktop interface so that all the GUI capabilities and workflows of PhotoScan are still available. A python script is still needed so that the user may avail herself of the compute nodes, but the rest of the workflow is very similar traditional PhotoScan usage. Finally, both methods offload the processing the HPC cluster, allowing users to continue to work on a computer that might normally be tied up by the processing demands of photogrammetry.

The workshop will give participants hands-on experience creating a 3D model using two different approaches: first, by accessing the Photoscan graphical user interface on a virtual desktop running on XSEDE's Jetstream cloud resource; and second, by using a Jupyter notebook running on an HPC cluster.

## OCR

Optical Character Recognition (OCR) is a tool used for extracting text from images and is perhaps most well known as a core technology behind the creation of the Google Books and HathiTrust corpora. OCR continues to open historical texts for analysis at large scale, fuelling a significant portion of research work within the digital humanities to the point that it would be difficult to think of the "million books problem" existing without this technology. While there are many OCR tools available the most popular tool that is also free and open source is Tesseract.

This portion of the workshop will also make use of Jupyter Notebooks to provide templates for learning the development process and that can then be taken away to speed development of future code. We will feature two projects for participants to practice with. A "traditional" OCR task that will have workshop participants processing images from the London Times in a demonstration of the improvements in OCR over the past few years and a task focusing on processing historical photographs to find text that can be added to the associated metadata to improve the searchability of an index.

## Target Audience

We anticipate that this workshop will appeal particularly to scholars who work with cultural heritage materials (a field where photogrammetry is an increasingly common method for generating digital surrogates), as well as those who work with archival photographs, and scholars with large corpora of photographs. It will also be relevant for scholars who already engage in computational analysis of primary sources, who wish to increase the efficiency of their analysis by leveraging high-performance compute environments. No previous experience with HPC environments is necessary. This workshop can accommodate 25 participants.

## Instructors

### *Quinn Dombrowski*

Quinn is the Humanities Domain Expert at Berkeley Research Computing. At UC Berkeley, Quinn works with humanities researchers and research computing staff at Research IT to bridge the gap between humanities research questions and campus-provided resources for computation and research data management. She was previously a member of the program team for the Mellon-funded cyberinfrastructure initiative Project Bamboo, has led the DiRT tool directory and served as the technical editor of DHCommons. Quinn has an MLIS from the University of Illinois, and a BA and MA in Slavic linguistics from the University of Chicago.

### *Tassie Gniady*

Tassie manages the Cyberinfrastructure for Digital Humanities group at Indiana University. She has a PhD in Early Modern English Literature from the University of California-Santa Barbara where she began her digital humanities journey in 2002 under the wing of Patricia Fumerton. She coded the first version of the NEH-funded English Broadside Ballad Archive, making many mistakes and learning much along the way. She now has an MIS from Indiana University, teaches a digital humanities course in the Department of Information and Library Science at IU, and holds regular workshops on text analysis with R and photogrammetry.

### *Megan Meredith-Lobay*

Megan Meredith-Lobay is the digital humanities and social sciences analyst, as well as the Vice President, for Advanced Research Computing at the University of Briitsh Columbia. She holds a PhD from the University of Cambridge in medieval archaeology where she used a variety of computing resources to investigate ritual landscapes in early medieval Scotland Scotland. Megan has worked at the University of Alberta where she supported research computing for the Faculty of Arts, and at the University of Oxford where she was the programme coordinator for Digital Social Research, an Economic and Social Research Council project to promote advanced ICT in Social Science research.

### *John Simpson*

John Simpson joined Compute Canada in January 2015 as a Digital Humanities Specialist and bringing a diverse background in Philosophy and Computing. Prior to Compute Canada, he was involved in a research-intensive postdoctoral fellowship focusing on developing semantic web expertise and prototyping tools capable of assisting academics in consuming and curating the new data made available by digital environments. He has a PhD in Philosophy from the University of Alberta, and an MA in Philosophy and BA in Philosophy & Economics from the University of Waterloo. In addition to his role at WestGrid, John is also a Member-at-Large of the Canadian Society for Digital Humanities (CSDH-SCHN), a Programming Instructor

with the Digital Humanities Summer Institute (DHSI),
and the national coordinator for Software