Teaching Culture Analytics:

Filling the gap between Humanities and Social Science research and Data Science tools: A Pedagogical Study

Sandra Alvaro (Autonomous University of Barcelona)
Clarisse Bardiot (Université de Valenciennes et du Hainaut-Cambrésis, France)
Stephen DeSalvo (UCLA)
Ekaterina Lapina-Kratasyuk (Higher School of Economics, Moscow)
Mila Oiva (Univ of Turku, Finland)
Mehrdad Yazdani (UCSD)

Introduction

The Humanities and Humanistic Social Sciences are driven by the availability of often highly specialized domain specific data, and are characterized by a significant *ability* to qualitatively analyze that data. An emerging challenge concerns the vast amounts of machine readable cultural data that now characterizes certain research problems in these disciplines, and the inability of these scholars on their own to undertake the data-driven analysis of culture. This disconnect between significant data repositories and well defined theoretical methods for qualitative analysis on the one hand, and the recognition that computational approaches to these data sets can provide considerable value to a complementary quantitative analysis is quickly becoming a pressing concern. A complementary situation holds in the realm of Computer Science, where large numbers of skilled technicians emerge from programs designed to drill them on data structures, algorithmic efficiency, and adherence to project specifications, without any particularly relevant application at the forefront.

Recently, the field of computer programming has undergone a radical transformation. Gone are the days when one would be expected to start a project from scratch, write one's own sort algorithm, or even design one's own data structure. Standard libraries have become the norm, generic programming has become the *de facto* starting point of any complex algorithm, and programming languages such as Python and R greatly simplify the complexities of languages like C.

As user friendly as languages such as Python and R are, there is still a gap which separates many researchers from the computational tools that could unlock their ability to more effectively analyze their data. The gap is closing, however, and at this point the level of expertise demanded of programmers is diminishing. Indeed, many undergraduate students with

basic programming knowledge are ready to help address the computational challenges faced by researchers across many disciplines, particularly those with a traditionally qualitative orientation.

Setup

In an effort to solve practical problems associated with data analysis, and provide a valuable learning experience to undergraduate students, we endeavored to facilitate collaborations between students and researchers. These collaborations were designed to build on each other's strengths. The catalyst was an advanced programming course at UCLA, PIC 10C, where students not majoring in engineering or computer science (i.e. mostly Mathematics and Social Science majors), were instructed differently than their Engineering School counterparts, with an emphasis on C++ language features and the C++ Standard Library rather than abstract language-free constructs.

The students in this course also had a final project requirement, written over the course of 6 weeks, typically in the form of a game written in Qt. For this special term, they were given *the option* to participate in a research project with a participating Culture Analytics long program researcher from the Humanities or Social Sciences. For many students, this was their first experience on a research project. The groups were divided into teams which met regularly with the researchers to discuss progress and future directions.

Before we discuss the projects at a greater level of specificity, we remark on what we believe were several crucial ingredients which contributed to the success of the collaborations:

1. Incentives
2. Milestones
3. Training
4. Ambition

On a semi-weekly schedule, the students had to turn in partially completed work, which forced them to think in terms of breaking the project down into smaller pieces and achieving <u>milestones</u> along the way. It also kept them constantly <u>incentivized</u> to work sooner rather than later. Due to the evolution of their programming <u>training</u> based primarily on learning language features and extracting properties of existing libraries, there was already a culture of self-exploration associated with having an abstract problem and needing to search for an appropriate tool in a library with which to concretely address it. Finally, many, if not most of the students in this experiment, exhibited the <u>ambition</u> to attend graduate school. Therefore, it was appealing to many of them to have a letter of recommendation from a leading researcher in the field attesting to the quality and impact of their work.

Projects

In each of the projects below, we describe briefly the motivation from the Humanities and Social Sciences, and then attempt to describe from a pedagogical point of view the evolution of the projects.

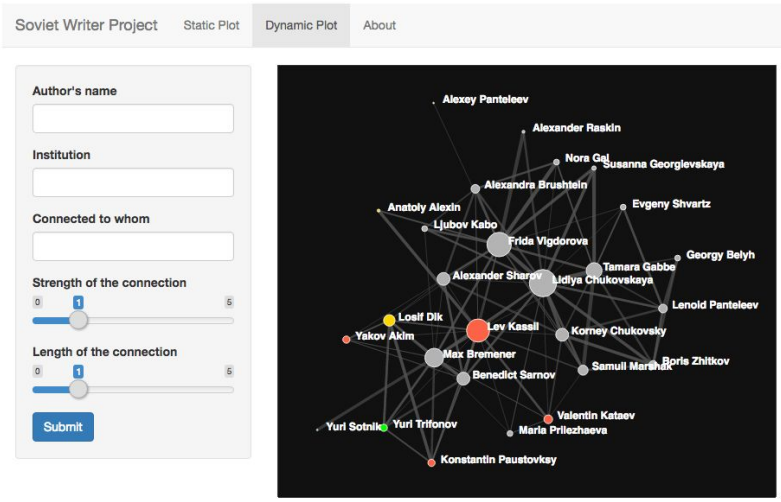
a. Soviet Writers Project

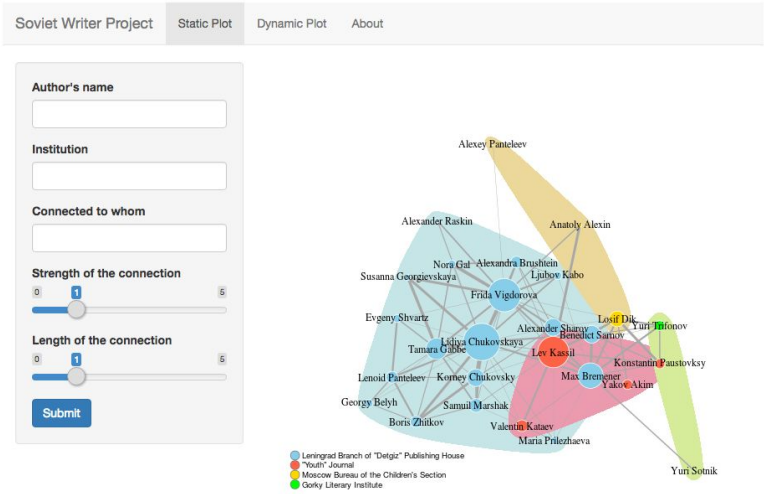
(In Soviet Russia, the computer programs you!)

This project, which involved a small, self-contained, mostly complete, and static data set, the pedagogical emphasis was on comparing two different approaches to interpreting and solving a problem. One group of three students desired to store, analyze, and display the data completely from scratch using only tools from the C++ Standard Library, and using the basic plot tools provided in the Qt framework. Another group of three students chose instead to use the *network* library in R and web tools to embed the resulting images into a html page so that it would be more accessible to researchers in other parts of the world.

Students:
Jing An, Wei Dai, Hengzhi Wu, Natalia Kazimi, Samir Patel, Zhongqi Wang

	A	B	C	D	E	F	G	H
1	Node Writer	Target Writer	Strength	Length	Institution			
2	Tamara Gabbe	Samuil Marshak	5	1	Leningrad Branch of "Detgiz" Publishing House			
3	Tamara Gabbe	Evgeny Shvartz	2	1	Leningrad Branch of "Detgiz" Publishing House			
4	Tamara Gabbe	Lidiya Chukovskaya	5	1	Leningrad Branch of "Detgiz" Publishing House			
5	Tamara Gabbe	Lenoid Panteleev	2	1	Leningrad Branch of "Detgiz" Publishing House			
6	Tamara Gabbe	Korney Chukovsky	2	2	Leningrad Branch of "Detgiz" Publishing House			
7	Tamara Gabbe	Georgy Belyh	1	2	Leningrad Branch of "Detgiz" Publishing House			
8	Tamara Gabbe	Boris Zhitkov	3	1	Leningrad Branch of "Detgiz" Publishing House			
9	Tamara Gabbe	Frida Vigdorova	3	2	Leningrad Branch of "Detgiz" Publishing House			
10	Tamara Gabbe	Nora Gal	2	1	Leningrad Branch of "Detgiz" Publishing House			
11	Tamara Gabbe	Susanna Georgievskaya	4	1	Leningrad Branch of "Detgiz" Publishing House			
12	Boris Zhitkov	Korney Chukovsky	4	1	Leningrad Branch of "Detgiz" Publishing House			
13	Boris Zhitkov	Lidiya Chukovskaya	4	1	Leningrad Branch of "Detgiz" Publishing House			
14	Boris Zhitkov	Lenoid Panteleev	3	1	Leningrad Branch of "Detgiz" Publishing House			
15	Boris Zhitkov	Georgy Belyh	3	1	Leningrad Branch of "Detgiz" Publishing House			
16	Boris Zhitkov	Samuil Marshak	3	1	Leningrad Branch of "Detgiz" Publishing House			
17	Samuil Marshak	Korney Chukovsky	2	1	"Youth" Journal			
18	Samuil Marshak	Lidiya Chukovskaya	3	1	"Youth" Journal			
19	Samuil Marshak	Frida Vigdorova	1	2	"Youth" Journal			
20	Samuil Marshak	Lev Kassil	2	1	"Youth" Journal			





b. Digitizing Polish newspapers from 1950-1980

(OCR)

The initial problem was that the Optical Character Recognition (OCR) of the photographed images of the Polish economic newspaper “Zycie Gospodarcze” using the open and free OCR program Tesseract, could only achieve approximately 5% accuracy, not nearly sufficient for research purposes. The goal was to devise a much more accurate OCR on the dataset. The researcher proposing this task, Mila Oiva (Cultural History, University of Turku & Aleksanteri Institute, University of Helsinki, Finland), explained that Tesseract worked well for other Polish texts, but this particular dataset was problematic because of the quality of the photographs. Good OCR tools are extremely important for historians because the majority of the sources used in history research are not born-digital. Furthermore, the availability of easy-to-use OCR tools is crucial because the majority of historians do not have a strong computing background. The availability of these tools is also important because institutional archives often concentrate on digitizing documents that are considered to be nationally important. Yet, in order to maintain the variety of the ways in which we understand our past, it is essential that historians can also use documents that are not considered of major importance, and therefore not readily available in digitized form. Many text analysis programs are made initially for English, which poses problems for studying sources in any other language in the world.

This project represented the greatest diversity of programming skills, backgrounds, and mentoring. Part of this was due to the form of the data, i.e., image files, which are not easy to parse directly. Instead, the students invoked the use of languages such as Octave to load in images, extract pixel values, and then apply various thresholding algorithms.

This project also stands out due to the use and timing of mentorship. Initially, the instructor of the programming course seeded the general approach of the project to one of the participants, breaking down the complexity of the problem into the following steps:

1. Load in the file
2. Extract a line of text from the image file
3. Crawl through the line and attempt to extract one character at a time
4. Analyze the character, e.g., compare with dictionary of letters.

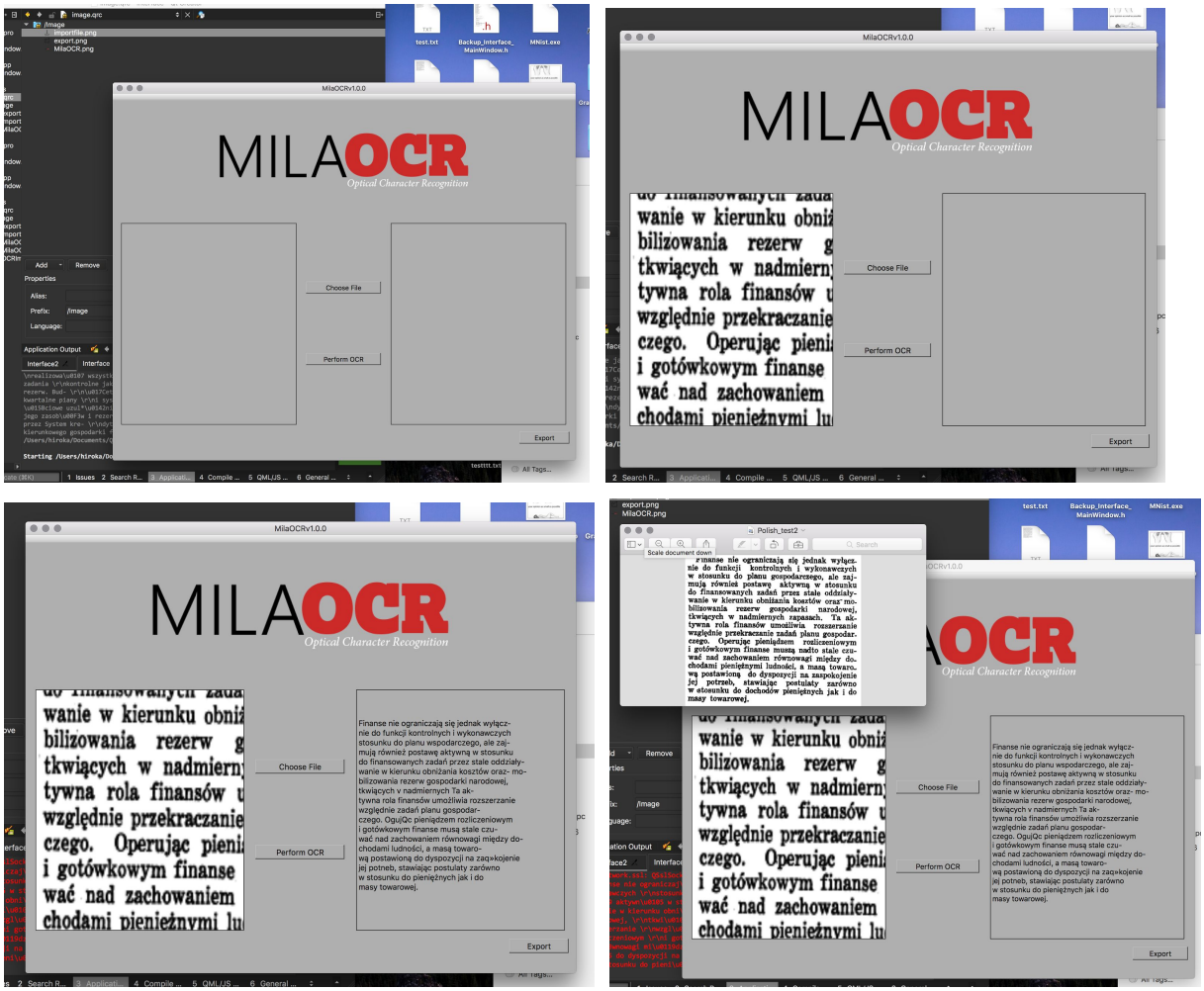
The tasks were then divided up among the remaining participants in a self-organizing manner, without any added input from the instructor. While the first three approaches were ultimately adopted and implemented, the final analysis of character recognition was the main bottleneck until the final week of the project. At that point, further mentorship was provided by the teaching assistant, an Applied Mathematics PhD student with a moderate background in image processing and an extensive background in programming. At his suggestion, they implemented a recent neural network algorithm which led to significant boosts in performance.

The researcher and the student group held meetings once a week, where the students reported how they had been proceeding, and what kinds of problems they faced. The role of the researcher was mainly of a client, keeping up at which phase the project was. She provided the group with short texts she typed from the newspapers to use as a training set.

The final result was a program which loaded an image file, allowed the user to select a subset of the image, and then extract the set of characters with a much higher accuracy than previous software packages. It does not allow batched OCR, which is often helpful when working with large quantities of text. However, the program is extremely useful for selecting particular articles in the images depicting the whole newspaper page. The output text has underlining of the misspelled words, which makes the cleaning process easier.

The students produced an introductory video of the project: https://www.youtube.com/watch?v=3CDqex8LSI8&feature=youtu.be

Students: Jennifer Okamura, Jack Wu, Xinting Zhao (Bess), Zane Karl, Simon Hua



c. Providing an Interface using PyQt for Image Study

(Who doesn't like the terminal?)

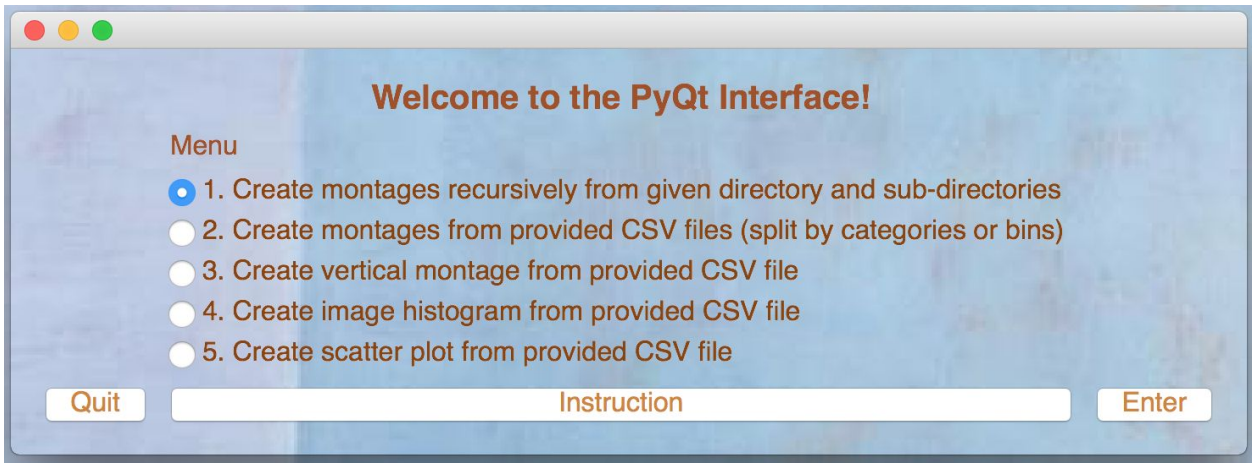
The challenge in this project was to develop a system that, given a large collection of images, could produce various types of montages of the images, allowing an art historical researcher the opportunity to discern latent patterns in the image collection based on various sorts and displays of the images.

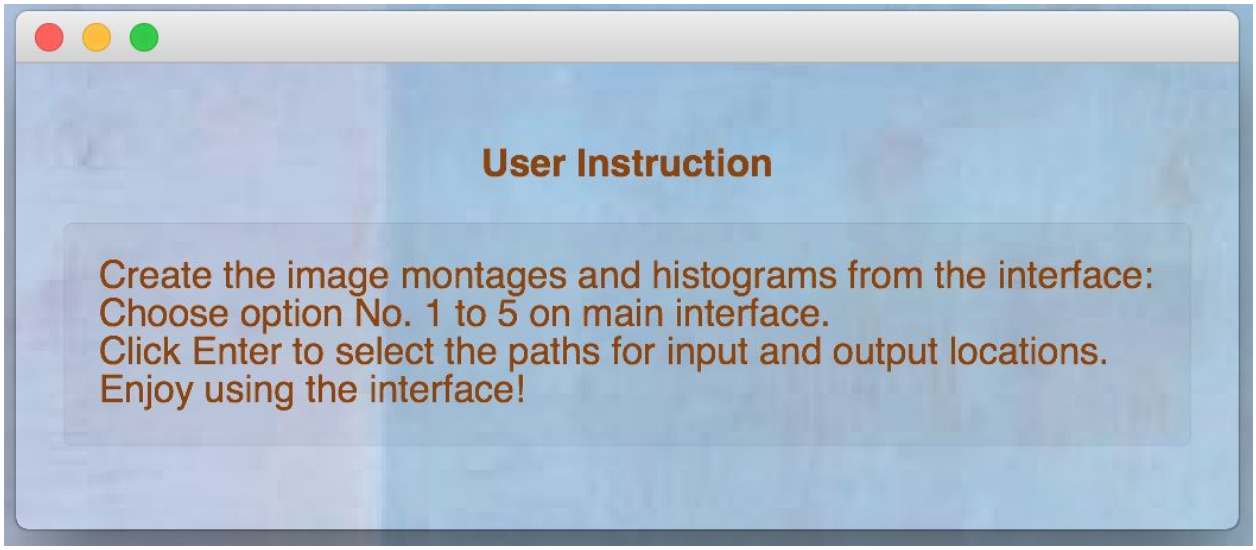
There is nothing worse than to devise and implement a tremendously useful application, only to be thwarted by an interface which deters potential users. Such was the case with this project, which demanded its users utilize the terminal in order to load in images for analysis. Consequently, there was a pressing need for a more intuitive and pleasant interface.

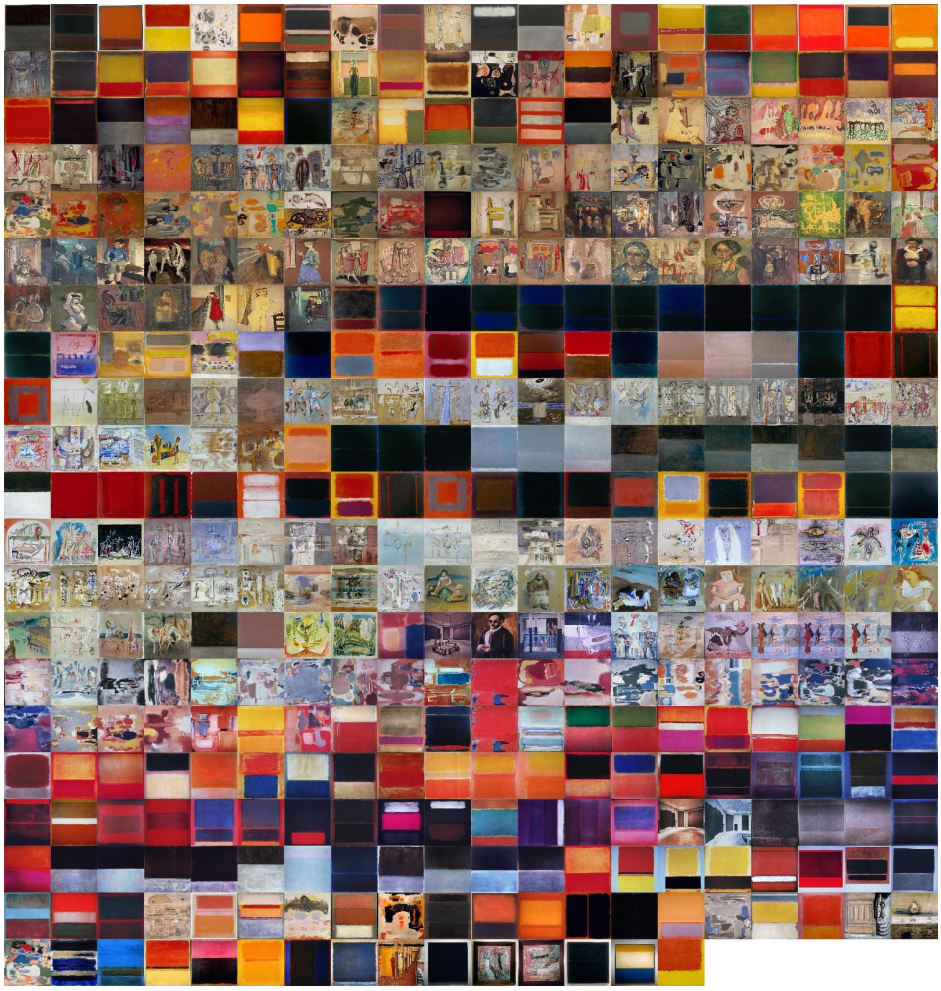
This project also had the added *benefit* of communication difficulties, as the participating researcher from the project provided all of the necessary initial information and data for the students to begin the project, after which they were unable to make contact with the researcher

and get desired feedback. Nevertheless, they were advised to simply continue and explore which parts of the project were able to be completed without further input and given *carte blanche* to develop it as they saw fit. Quite admirably, they saw fit to work around the absence of certain CSV files, providing everything needed in principle so that once such properly formatted files were provided, the program should work verbatim without any extra programming input.

Students:
Shiyun (Lily) Qiu, Siwen (Selina) Tang




User Instruction
Create the image montages and histograms from the interface:
Choose option No. 1 to 5 on main interface.
Click Enter to select the paths for input and output locations.
Enjoy using the interface!



Conclusions

Culture Analytics as a translational field that works at multiple scales and multiple resolutions is already capturing the imagination of students. Our work with advanced undergraduates who have a background in both computer programming and qualitative approaches to the study of culture has revealed a ready and eager group of students who in the years to come will develop the field. Our pedagogical experimentation holds great promise for the future. We were particularly buoyed by the success of bringing students together with researchers whose problems were not typical computer programming problems. As the students became aware of the particular focus of the cultural researchers, they began to approach their programming differently, recognizing that the domain expertise of all members of the team could lead to work that was truly interdisciplinary in scope and geared toward solving a problem in a field where such computational approaches to problem solving were previously intractable.

As a result of our initial pedagogical experiments, we recommend that additional focus be placed, already at the undergraduate level, of the development of research teams focused on specific problems in the data-driven analysis of culture. Support for team-taught courses that are project focused are appealing to researchers, teachers, and students alike, and hold great promise for rapidly advancing the creation of knowledge.