Supercomputing to Help UC San Diego Researchers Visualize Cultural Patterns

January 9, 2009

Doug Ramsey

High-performance computing and the humanities are connecting at the University of California, San Diego - with a little matchmaking help from the Department of Energy (DOE) and the National Endowment for the Humanities (NEH).

The two agencies have awarded 330,000 hours of time on DOE supercomputers to UC San Diego's Software Studies Initiative to explore the full potential of cultural analytics in a project on "Visualizing Patterns in Databases of Cultural Images and Video." The grant is one of three inaugural awards from a new Humanities High Performance Computing Program established jointly by DOE and NEH.

"Digitization of media collections, the development of Web 2.0 and the rapid growth of social media have created unique opportunities to study social and cultural processes in new ways," said principal investigator Lev Manovich, who directs the Software Studies Initiative. "For the first time in human history, we have access to unprecedented amounts of data about people's cultural behavior and preferences as well as cultural assets in digital form. This grant guarantees that we'll be able to process that data and extract real meaning from all of that information."

The Software Studies Initiative is a joint venture of the UCSD division of the California Institute for Telecommunications and Information Technology (Calit2) and the university's Center for Research in Computing and the Arts (CRCA).

The cultural data to be crunched by the DOE supercomputers include millions of images, paintings, professional photography, graphic design, and user-generated photos, as well as tens of thousands of videos, feature films, animations, anime music videos and user-generated videos. Examples of data sets to be processed include 200,000 images from Artstor.org, as well as 1500 feature films, 2,000 videogame previews and 1,000 video recordings of videogame play, all from Archive.org.

(All the cultural data is in the public domain, published under appropriate Creative Commons license, or is under publishing agreements that allow research and educational use.)

"This will allow us to compare how the proposed approach works with different types of data and also to communicate to different scholarly communities the idea of using high-performance computing for the analysis of visual and media data," said Manovich. "We hope that our project will act as a catalyst encouraging more people to begin their own projects in turning visual data into knowledge."

The Humanities High Performance Computing Program is a one-of-a-kind initiative that gives humanities researchers access to some of the world's most powerful supercomputers in DOE's National Energy Research Scientific Computing Center (NERSC) at Lawrence Berkeley National Laboratory. The program will also fund training time for the humanities researchers on the computers.

The UC San Diego project and two others - based at Tufts University and the University of Virginia - were selected after a highly competitive peer-review process led by the NEH's Office of Digital Humanities.

"A connection between the humanities and high-performance computing communities had never been formally established until this collaboration between DOE and NEH," said Katherine Yelick, NERSC Division Director. "The partnership allows us to realize the full potential of supercomputers to help us gain a better understanding of our world and history."

For approximately three years, Manovich has been developing the broad framework for research in cultural analytics as part of the Software Studies Initiative. The framework uses interactive visualization, data mining and statistical data analysis for research, teaching and presentation of cultural artifacts, processes and flows. Another focus is on using the wealth of cultural information available on the Web to construct detailed and interactive spatio-temporal maps of contemporary global cultural patterns.

"I am very excited about this award because it allows us to undertake quantitative analysis of massive amounts of visual data," explained Manovich.

Researchers will use a number of algorithms to extract image features and structure from the images and video. The resulting metadata will be analyzed using a variety of statistical techniques, including multivariate statistics methods such as factor analysis, cluster analysis, and multidimensional scaling. That statistical analysis and the original data sets will then be used to produce a number of highly detailed visualizations to reveal new patterns in the data.

All the outcomes of the project will be made freely available via a new Web portal, CultureVis, set up by Software Studies Initiative.

Manovich is an expert in the theoretical and historical analysis of digital culture. His previous work will be used to determine which features of visual and media data should be analyzed. "Above all," he noted, "we hope to direct our work towards discovery of new cultural patterns, which so far have not been described in the literature."

Said NEH Chairman Bruce Cole: "Supercomputers have been a vital tool for science, contributing to numerous breakthroughs and discoveries. The Endowment is pleased to partner with DOE to now make these resources and opportunities available to humanities scholars as well, and we look forward to seeing how the same technology can further their work."

In addition to UC San Diego's cultural analytics project, two other groups were awarded time on the DOE supercomputers. The Perseus Digital Library Project, led by Gregory Crane of Tufts University, will use NERSC systems to measure how the meanings of words in Latin and Greek have changed over their lifetimes, and compare classic Greek and Latin texts with literary works written in the past 2,000 years. Separately, David Koller of the University of Virginia will use the high-performance computing to reconstruct ancient artifacts and architecture through processing and analysis of digitized 3D models of cultural heritage.

Media Contacts: Doug Ramsey, Calit2, 858-822-5825 Linda Vu, Lawrence Berkeley National Laboratory, 510-495-2402.