400-dimensional dynamic stochastic problem in [9].

Our analyses also required extensive uncertainty quantification, which fortunately is well-suited to Blue Waters. Our applications of uncertainty quantifications used thousands of parameter specifications in [2], and hundreds in [1].

In the next year, we will study the impact of multiple interacting tipping points on the social cost of carbon in a fifteen-dimensional version of DSICE. We will also examine the effects of learning. The high-dimensional nature of both models will require Blue Waters' capabilities.

PUBLICATIONS

Lontzek, T. S., Y. Cai, K. L. Judd, and T. M. Lenton, Stochastic integrated assessment of climate tipping points indicates the need for strict climate policy. *Nat. Clim. Change*, 5 (2015), pp. 441–444, doi:10.1038/nclimate2570.

Cai, Y., K. L. Judd, T. M. Lenton, T. S. Lontzek, and D. Narita, Environmental tipping points significantly affect the cost-benefit assessment of climate policies. *Proc. Nat. Acad. Sci.* U.S.A., 112:15 (2015), pp. 4606–4611, doi:10.1073/pnas.1503890112

COMPUTATIONAL STRATEGIES FOR APPLYING QUALITY SCORING AND ERROR MODELING STRATEGIES TO EXTREME-SCALE TEXT ARCHIVES

Allocation: Illinois/50 Knh
PI: Scott Althaus[1,2]
Co-PIs: Loretta Auvil[3], Boris Capitanu[3], David Tcheng[3], Ted Underwood[1]

[1]University of Illinois at Urbana–Champaign
[2]Cline Center for Democracy, University of Illinois at Urbana–Champaign
[3]Illinois Informatics Institute, University of Illinois at Urbana–Champaign

EXECUTIVE SUMMARY:

An important barrier to extreme-scale analysis of unstructured textual data digitized from printed copies using optical character recognition (OCR) techniques is the uncertain quality of the textual representations that have been made from scanned page images. We used Blue Waters to evaluate OCR errors on the HathiTrust Public Use dataset, which is the world's largest collection of digitized library volumes in the public domain, consisting of 3.2 million zipped files totaling nearly 3 TB. We also used Blue Waters to assess the impact of OCR errors on event-detection algorithms using a collection of 16 million articles from The New York Times. Our aim is to develop error quality scoring and correction strategies that can enhance the ability of data analytics researchers to work with digitized textual data at extreme scales.

INTRODUCTION

Researchers in the humanities and social sciences often analyze unstructured data in the form of images and text that have been scanned and digitized from non-digital sources. For this type of research, the most important barrier to conducting extreme-scale analysis of unstructured data is the uncertain quality of the textual representations of scanned images derived from optical character recognition (OCR) techniques. Our project is using Blue Waters to detect, score, and correct OCR errors in the HathiTrust Public Use dataset, which is the world's largest corpus of digitized library volumes in the public domain, consisting of over 1.2 billion scanned pages of OCR text. We are also using Blue Waters to assess the impact of OCR errors on natural language processing algorithms using a corpus of 16 million historical newspaper articles from The New York Times (NYT).

METHODS & RESULTS

In collaboration with the HathiTrust Research Center (HTRC), we performed two major computations with their data. Each of these implementations leveraged Akka [1] to provide a high-performance JVM-based framework featuring simple concurrency and distribution. The first HTRC-related computation on our Blue Waters allocation supported the evaluation of OCR errors in the HathiTrust public domain volumes (which at the time was 3.2 million volumes). We applied new error-detection algorithms produced by a Mellon-funded project called eMOP under Laura Mandell at Texas A&M. Our team's analysis of text-level quality problems will be compared and correlated with image-level quality analysis undertaken by Paul Conway of the University of Michigan on the page images from which the OCR text was extracted. Our aim is to identify how OCR errors are correlated with specific types of image distortion so that digital librarians can better anticipate quality problems from their digitization efforts.

The second HTRC-related computation on our Blue Waters allocation supported the creation of the HTRC Extracted Features (EF) dataset [2], where "features" are notable or informative characteristics of the text [3–5]. The dataset is derived from 4.8 million HathiTrust public domain volumes, totaling over 1.8 billion pages, 734 billion words, dozens of languages, and spanning multiple centuries. We processed a number of useful features at the page level including part-of-speech tagged token counts, header and footer identification, and a variety of line-level information. The EF dataset was used in the creation of the HathiTrust Bookworm [6], which is a tool that visualizes language usage trends in repositories of digitized texts in a simple and powerful way.

Our Blue Waters allocation is also supporting a third set of computations designed to clarify the impact of OCR error on natural language processing algorithms. The Cline Center for Democracy has access to the entire population of NYT articles from 1980 to 2005 in two different forms: "born-digital" content that contains pristine textual data, and ProQuest's digitized version of the same content that was derived from microfilm images using OCR. We used Blue Waters to deploy Phoenix civil unrest event identification software [7] produced by the Open Event Data Alliance (OEDA) over the combined 16 million articles in this NYT corpus. Phoenix requires making computationally intensive parse trees for the articles in order to identify events, and we have a prototype of this deployed, again using Akka, to distribute the work of creating parse trees, identifying events, and writing results. We also plan to compare 25 years of NYT OCR articles to the born-digital NYT articles using software developed for the eMOP project. This will help us evaluate OCR quality and determine the level and type of correction that most closely restores noisy OCR to resemble the born-digital content so that automated correction algorithms can be optimized for use with OCR data from ProQuest's historical newspaper collection.

WHY BLUE WATERS?

The computational demands of using natural language processing, machine learning, or rule-based scoring strategies on the scale of the HathiTrust Public Use corpus would severely tax the capabilities of other HPC platforms. Only Blue Waters offers the computational scale required to carry out the necessary quality scoring and error correction strategies in a timely fashion on an unstructured text corpus of this size.