HYPERION RESEARCH

HPC Profiles In Leadership

Examples of HPC Leadership: Dr. Ruby Mendenhall, University of Illinois at Urbana-Champagne, 2017

Earl Joseph, Steve Conway, Bob Sorensen, and Kevin Monroe

July 2017

HYPERION RESEARCH OPINION

In the 1940s, a group of female scientists were the human computers behind the biggest advances in aeronautics at NASA (see https://www.theguardian.com/lifeandstyle/2016/sep/05/forgot-black-women-nasa-female-scientists-hidden-figures). However, many forgot that some of these women happened to be black. Today, High-performance computing (HPC), also called supercomputers, help design Quantum Artificial Intelligence and fundamental aeronautics at NASA – and uncover stories about black women. Professor Ruby Mendenhall at the University of Illinois is utilizing HPC to help recover black women's history. Dr. Mendenhall is utilizing big data to "Recover Black Women's Lived Experiences." Often, literature by and about African American women is inaccessible. With advances in HPC and Big Data platforms researchers on Dr. Mendenhall's team can use what they consider a Call & Response methodology, which is akin to Alice Walker's Search for Zora Neal Hurston's Grave. The goal of the project is to recover what was written about black women's ideas, challenges, actions/agency, and accomplishments. The research questions are:

- What themes emerge about African American women using topic modeling?
- How can the themes identified be used to recover previously unmarked documents?
- How might we visualize the recovery process?

The Call & Response methodology consists of four procedures (Search, Recognition, Rescue and Recover) and starts with the search (or call) that involves training of the topic model using a subset of 20,000 documents. The next step is recognition intensive and comprising of intermediate and close readings to identify potentially new documents that were not identified before as being by or about Black women's lived experiences. Then a similarity/dissimilarity analysis of 800,000 documents is performed. After confirmations are made, the documents are rescued and placed in the recovered corpus about Black women.

The findings suggest writing and entering historical records are acts of power and privilege. Writings about black women's lived experiences are found in unusual texts and their voices are recovered through the voices of others, often White men. The use of topic modeling led to the recovery of 124 previously unidentified documents related to Black women. Using a cosine similarity test Dr. Mendenhall's team was able to recover an additional 26 previously unidentified documents. It should be noted that they are conducting ongoing research to test model parameters that may lead to the recovery of additional documents. Dr. Mendenhall et al. have shown the application of HPC and Big Data can literally rewrite history by interpreting events from the past in ways that only advancements in hardware and software can make possible.

Note: this page is intentionally blank.

IN THIS HPC PROFILES IN LEADERSHIP REPORT

Finding What Was Hidden: Bringing the Past to Life

Dr. Ruby Mendenhall of the University of Illinois at Urbana-Champaign led a collaboration among social scientists, humanities scholars and librarians, and digital researchers using HPC at the National Center for Supercomputing Applications (NCSA) to search two massive databases of written works from the 18th to 21st century to discover and better understand the historical experiences of black women. They called their project: “Rescuing Lost History: Using Big Data to Recover Black Women’s Lived Experiences.” Dr. Mendenhall was one of the winners of the HPC Innovation Excellence Award at SC16, an annual worldwide supercomputer industry conference.

SITUATION OVERVIEW

National Center for Supercomputing Applications at the University of Illinois at Urbana-Champaign

The NCSA describes itself as a ‘hub of trans-disciplinary research and digital scholarship.” It is a place where University of Illinois faculty, staff, students, and collaborators from around the world come together to research some of the biggest issues and greatest challenges we face today. Using the most sophisticated and advanced HPC technology, they can access information and research to find answers to big questions. Their current research focus areas are Bioinformatics and Health Sciences, Computing and Data Sciences, Culture and Society, Earth and Environment, Materials and Manufacturing and Physics and Astronomy. The NCSA also provides integrated cyber-infrastructure such as computing, software, data, networking and visualization resources and expertise that are essential to the work of the scientists, engineers, and scholars from the University of Illinois at Urbana-Champaign and other universities across the country.

WHY HPC IS IMPORTANT TO DR. RUBY MENDENHALL'S RESEARCH

Dr. Ruby Mendenhall won the HPC User Award last year for her groundbreaking work researching little known writings that shed light on African-American women’s history from the 18th-21st century. She is a modern-day Renaissance woman whose vision and purpose to bring clarity and understanding to this historical period led her to take a deep dive into Big Data.

She turned to this kind of analysis and research to "chart a new course by centering previously marginalized voices and perspectives using Big Data,” Mendenhall explained. “Examinations of how identities are shaped and reshaped by vectors of race, gender, class and sexuality are critical for the production of new knowledge.”

Her work ranges across disciplines. Currently, she is an Associate Professor in Sociology, African-American Studies, Urban and Regional Planning, Gender and Women’s Studies, and Social Work at the University of Illinois at Urbana-Champaign. She is also a faculty member at the Institute for Genomic Biology. She has affiliate appointments in Women and Gender in Global Perspective, the Institute for Computing in Humanities, Arts and Social Sciences and the Cline Center for Democracy. Currently, she is working on creating DREAM, a project which will promote heath and help build safe spaces for women on the south side of Chicago.

Rescuing the Hidden Voices of African-American Women

"Often, literature by and about African American women's ideas, challenges, actions/agency, and accomplishments is inaccessible"

The scope of the research stretched from 1746-2014, studying the themes that emerge about African American women using topic modeling to identify and recover previously unmarked documents about African American women across history, moving among such eras as slavery, the abolitionist movement, world wars, the Civil Rights Movement, Feminist Movement and, most recently, the Black Lives Matter Movement. Her research team encompassed multiple fields and included computer scientists, visualization programmers, and African-American Studies scholars.

"The research we do is important to the world, because valuable materials in journals, books, newspapers and letters are becoming increasingly available electronically," Mendenhall said. "The sheer volume of accessible data calls for new and innovative approaches to surveying this big data to extract knowledge that is useful to how society functions for groups on the margins.

The theoretical underpinnings of Dr. Mendenhall's research are uncovering the pivotal role knowledge plays in reproducing and dismantling social inequality. Standpoint theory involves group knowledge based on shared common experiences such as oppression that links the everyday lived experiences of Black women to interlocking systems of race, class, and gender discrimination.

"Using digitized data to understand African-American women's experience is of particular importance because there is so much still to discover about how they lived, survived and thrived, in some cases, despite the oppression associated with slavery, Jim Crow, segregated housing, integration into predominantly white environments and the actions they took to change society. We're finding some of those answers may be found in the digitized record."

Seeking History Using Algorithms

For her award-winning study, Mendenhall said she used "comparative text mining" (CTM), a generative probabilistic mixture topic model where the algorithm provides outputs that allow for comparisons of sets of common topics produced across the entire corpus (common models) and variations in topics across specific time periods (expert models). The computers she used were the Blacklight, Greenfield, and Bridges systems at NCSA.

She said, "the historical periods for the comparative model are leveraged to create temporal clusters, and topics are produced based on the documents in the corpus written during these specific periods." By using this methodology, Mendenhall said this approach "allows researchers to capture how topics may vary from the common topic over time. The comparative text mining algorithm does not inherently make for comparisons across time; the document collection can be divided into sub-collections based on any available metadata, such as time, location or authorship. We then use the temporal metadata of the corpora to make these comparisons by setting up the dataset so the results can be arranged within time periods."

Mendenhall's team divided the collection of 20,000 documents into different time periods, "trying to keep the proportion of documents stable across all sub-collections, as this is important for keeping the common model general and avoid having an expert model being overrepresented in the common model."

EYES ON THE PRIZE: REDISCOVERING THE LIVES OF BLACK WOMEN

A Success Story

The prize Mendenhall was seeking was to discover the hidden texts about African-American women that may be difficult to find or obscure using traditional research methods – and by so doing to bring to life Black Women's history. "Often, literature by and about African-American women is inaccessible," she said. "The project's goal was to recover what was written about their ideas, challenges, actions/agency, and accomplishments."

The iconoclastic scholar said when teaching social research methods, she wanted students "to understand that they can incorporate working with large-scale data into their research methods to improve their ability to answer questions such as, "what is social inequality; how do interlocking systems work together; how is inequality expressed in the everyday lives of African American women and other groups; and how can entrenched systems of oppression be altered?"

She said she asked Michael Simone (now Director of the Digital Humanities and Trans-Disciplinary Informatics Research Lab at Arizona State University) to give a big data lecture in her research methods course. Simone encouraged her to write the Institute for Advanced Computing Applications and Technologies (IACAT) Fellowship proposal that eventually became this award-winning project.

Mendenhall said she and her team approached the research by asking these questions:

- What types of topics and conversations will emerge about Black women's shared experiences over time (1700s-2000s) in various genres of text (poetry, science, psychology, sociology, African American Studies, public policy, etc.)?
- What expected and unexpected themes emerge when using topic models to examine several centuries of text by and about Black women?
- What is the meaning-making process that is applied to the topics that emerge?
- How can this massive complex data that spans various key historical periods be visualized?

Notable Discoveries

Among the most important discoveries of the research are these:

- Due to the limits of distant reading, they created a digitized technique that is between distant and close reading, called intermediate reading. Intermediate reading provided them with document titles, authors and dates associated with each topic along with the percentage of words in the document assigned to a topic. This allowed them to validate the interpretation of the topic models and gain insight into Black women's standpoints (group knowledge) during key historical periods.

- Many volumes for or about Black women were not tagged properly. After examining 300,000 volumes, the team found that about 27 percent of them (~80,000) did not have subject metadata. Therefore, if a researcher does not know to search for specific documents by or about Black women, they may not have access to a significant amount of data on the topic which interferes with knowledge production.
- The team has started to recover previously unidentified documents by or about Black women that will help researchers and community members better understand the lived experiences of Black women. Their preliminary efforts resulted in the recovery of 124 previously unidentified documents using the KL divergence list and 26 volumes using the cosine similarity list. The team's goal is to examine the entire 800,000 corpus to recover additional documents about Black women's history and America's history.
- Research themes that emerged included how inequality is expressed (or hidden) in the everyday lives of African-American women, and how the women seek to change entrenched interlocking systems of oppression such as racism, classism, or sexism. The team found themes about slavery, the Black Women's Club Movement, and the New Negro Movement, which resulted in increased resistance to oppression and greater cultural expression. These themes verify that the LDA models can capture Black women's experiences that are consistent with what is known historically and from close readings of text.
- A significant challenge to recovering documents that centered Black women lived experiences was the need to examine documents written by white men, white women and Black men. The ability to write and enter the digitized record often reflects power and privilege. Therefore, it is important to understand big data and computation as reflecting social structures that often erase Black women. An example of erasure in the team's research involves the role that Black women and their bodies played in medical advancements in U.S. society. Our cosine results identified strong associations with medical references that were enhanced with data visualization and close readings. These findings are consistent with and expands the close readings of text done by Harriet Washington for her (2006) book *Medical Apartheid: The Dark History of Medical Experimentation on Black Americans from Colonial Times to the Present.*
- Data visualizations using word clouds, topic vectors, histograms, network maps, etc. were critical to intuitively understand complex data that spanned centuries and different types of lived experiences for Black women.

These are important finding for scholars who are interested in using "big data" to answer key questions about the nature of structural oppression but who may be hesitant to use computational tools. Utilizing this type of data to answer key questions can allow researchers to uncover hidden aspects of Black women's experiences that may have been lost due to the texts' rare nature inaccessible physical locations or erasures due to social structures.

"Understanding the intersectional nature of race, class, gender and sexuality oppression helped to contextualize and interpret computational findings as we sought to make meaning of such a large dataset," Dr. Mendenhall explained. "We also trained our algorithm in the African American tradition of 'call and response' as we set out to find previously undiscovered documents within the corpus that exhibited the same discursive patterns as those documents we knew to be for and about Black women."

Mendenhall also realized “how underutilized computational analysis is within the field of sociology, which we hope this research will help to rectify. We also discovered that corpora could be re-imagined as socio-political entities unto themselves, which are created within, and not distinct from, the realms of the social and political. This reinterpretation allows researchers to expose the covert colorblind logic, which can be perpetuated within computational analysis when corpora are presented as apolitical and objective. It also allowed us to explore questions around why some digitized sources are considered ‘legitimate’ in the pursuit of new knowledge while others are not (i.e. Formal digital libraries such as HATHI and JSTOR versus informal online digital sources such as blogs and social media).”

RESEARCH PARTNER S THAT STRETCH CROSS DISCIPLINES

Mendenhall is especially proud of the diverse workforce involved in the project as well the training it offered to undergraduate and graduate students to work within groups that design, build and use HPC and big data,” she said. “Also important is our interdisciplinary team of scholars.” She also praises the infrastructure and resources provided by XSEDE (Extreme Science and Engineering Discovery Environment and its Extended Collaborative Support Service, which is headquartered at the National Center for Supercomputing Applications and the Pittsburgh Supercomputing Center.

Mendenhall’s Team

- Nicole M. Brown: Research Faculty Affiliate at the National Center for Supercomputing Applications (NCSA) at the University of Illinois, Urbana-Champaign. Dr. Brown’s research interests include black feminist theory, computation and political consumerism. Her research utilizes archival and computational analysis to investigate how intersections of race, class and gender influence political consumerism within social movements.
- Michael L. Black: Assistant Professor of English at University of Massachusetts Lowell. His research addresses the cultural history of personal computing, big data, and new media. He recently served as the Associate Director for the Institute for Computing in Humanities, Arts, and Social Sciences (I-CHASS) at the University of Illinois.
- Mark W. Van Moer: Senior Visualization Programmer at the National Center for Supercomputing Applications (NCSA) at the University of Illinois, Urbana-Champaign. He is interested in introducing visualization to new communities and disciplines.
- Ismini Lourentzou: Doctoral student in the Department of Computer Science who does research on comparative text analysis.
- Chengxiang Zhai: Professor of Computer Science. His research interests include intelligent information systems, information retrieval, data mining, natural language processing, and machine learning.
- Assata Zerai: Associate Chancellor for Diversity at the University of Illinois. Research interests include race, class and gender in Africa and its Diaspora.
- Karen C. Flynn: Associate Professor at the University of Illinois, Urbana-Champaign. She holds joint faculty appointments in the Departments of Gender and Women's Studies and African-American Studies. Her research interests include migration and travel, Black Canada, health, popular culture, feminism and diasporic and post-colonial studies.
- Malaika McKee: Visiting Assistant Professor of African American Studies at the University of Illinois, Urbana-Champaign. Research interests include critical pedagogy and big data analysis.
- Harriett Green. English and Digital Humanities Librarian at the University of Illinois, Urbana-Champaign. Her research interests include humanities data curation, digital pedagogy, scholarly communication, use and users of digital humanities resources

The Value of this Project: Integrating Big Data and Lived Experience

“Our goal is to help to de-mystify the integration of topic modeling, visualization and social scientific work and serve as a ‘translator’ across disciplines to help future social science researchers to become more familiar with the language of computational analysis,” Mendenhall explained. “This would greatly benefit the social sciences’ and humanities’ understanding of the power of these tools as well as encourage future researchers to engage large datasets using this technology.

This research will also “help to re-frame our understanding of content analysis and re-shape what is considered knowledge about Black women. In essence, we will provide a theoretical framework (lens) that can provide appropriate context for improved interpretations and inferences about Black women's lived experiences using metadata.” By using these tools, Mendenhall sees great potential for these kinds of studies to address inequality in American life.

FUTURE OUTLOOK

LOOKING AHEAD

SEARCH AND RESCUE: A MAP FOR OTHER EXPLORERS

With the hope that her work will be used by other scholars and researchers, Mendenhall and her team created a “blueprint for conducting research using big data and analyzing the research. Often, the literature on topic modeling is concerned with the efficiency and accuracy of the algorithms' use, but lacks in explanations around the steps for interpreting the results of the data.”

She calls the subsequent blueprint the team created SeRRR (Search, Recognition, Rescue and

Recover) and said it follows the Black call and response tradition. “It is also what Alice Walker used as she searched for the unmarked grave of [African-American writer] Zora Neale Hurston.

The Call is the subset algorithm used for Searching for documents that have similar subjects to the one that we identified as important to the lived experiences of black women: club movement, labor, religion, etc.).”

The Response is the new documents that emerge with topics related to Black women that were not Recognized as being important prior to our research. After intermediate and/or close readings to confirm the document is related to Black women, we Rescue the document and place it in the Recovered corpus about black women. We then mark the documents/site/grave by making the corpus publicly available to librarians and other scholars with copyright considerations.”

WHAT MENDENHALL WOULD LIKE TO SEE MOST IMPROVED IN FUTURE

Challenges and Surprises

Mendenhall, who plans to always use HPC in her work in the future, said some of the biggest surprises and challenges in the work were “the simultaneous effectiveness and ineffectiveness of computational tools in identifying cohesive themes within such a large and varied corpus. Although some output was disjointed, other results provided clear and coherent themes around African-American women’s experiences (i.e., maternal health, children and education, Black resistance, and Black women’s agency such as the club movement). It was a reminder that computation should be used in conjunction with subject matter expertise and not as a substitution for researchers’ expert contextualization and interpretations.”

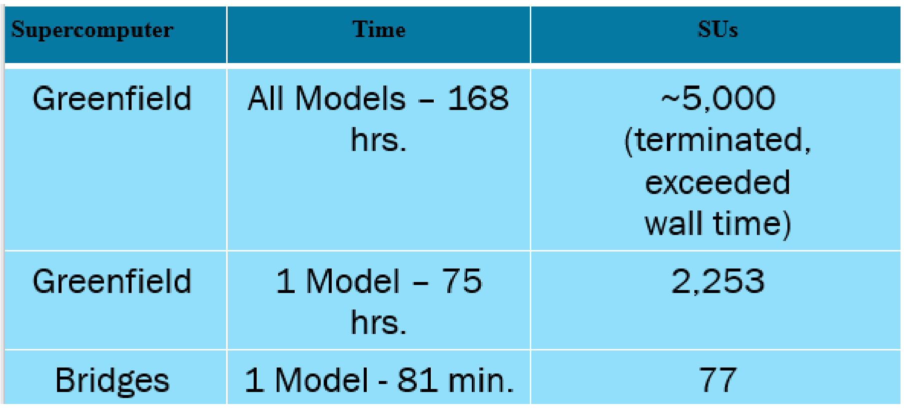
The team was surprised, as well, she said “at the limited amount of text identified as by and about Black women in the HathiTrust and JSTOR digital libraries. Therefore, one of the goals of this project is to center the experiences of Black women within digital humanities research, from corpora creation through the process of analysis. We want to use Black feminist theory as a framework to dismantle embedded privilege and reification and (digitally) center the experiences of Black women within digital humanities research.”

It was also hard, she said, to analyze some of the topics just from the distant readings. To ameliorate this issue, she said the team developed a technique they called “intermediate reading to reflect a process of reading the text that is situated between the close readings associated with traditional research and the distant readings of text associated with topic modeling output.”

“While the distant reading method of interpretation identified several topics of interest,” she explained, “we felt somewhat removed from the collection itself and wanted to contextual information. Topics serve more or less an index of linguistic patterns in a collection, but like an index, these models can be used both to reveal and navigate the complex structure of a collection of texts. To evaluate the relationship between the patterns identified by the model and the texts used to generate them, Michael Black developed a simple tool that scanned the word to topic assignments for each document and returned a list of titles that had a given percentage of their words assigned to topics of interest. This gave us more information about the name and gender of the author, the date of text, and title of the document. We looked up several of the titles and did close readings when we felt that we needed even more context or information to decide if the article was by or about Black women.”

There were also some issues of timing, until they moved to NCSA’s Bridges supercomputer, Mendenhall said. She explains: “Comparative Text Mining (CTM) is a more complex form of LDA. The more common model and expert models need more computational time. It took 5 days on Greenfield to create 25 topics and 8 expert models for each topic.”

“During the inferencing/testing phase, we used the Greenfield and Bridges supercomputers at University of Pittsburgh. Greenfield uses a great deal of processing units so we often exhausted the resources. Bridges lets you define memory needs, lowers computing costs.”



Supercomputer	Time	SUs
Greenfield	All Models – 168 hrs.	~5,000 (terminated, exceeded wall time)
Greenfield	1 Model – 75 hrs.	2,253
Bridges	1 Model - 81 min.	77

ESSENTIAL GUIDANCE TO OTHER HPC SITES

Best Practices and Advice to Other HPC Sites

Looking ahead, Mendenhall sees long-term infrastructure support as important for researchers using HPC. “The XSEDE start-up grant was critical to keep the project going after my year as a fellow and the delays with getting security negotiation with the project’s two data sources. We then received an XSEDE allocation and Extended Collaborative Support Service (ECSS) program. Having access to Drew Schmidt and Mark Van Moer and other cyber-infrastructure experts was key to the project’s success,” she said. “This support allowed a group from very different backgrounds to meet and have conversations and analyze the data for about three years, which was extraordinary.”

This collaboration was essential, Mendenhall says. “It allowed the social scientists to focus on our area of expertise and to learn about big data methodology without having to start from ground zero. Also, we could say what we wanted with the analysis and visualizations and Mark Van Moer was great at figuring out how to do it, even when what we asked for was challenging.”

FUTURE IMPROVEMENTS

PhD Candidate Ismini Lourentzou, who was a researcher on the project, pointed to areas for future improvements.

“Our work utilized supercomputing to speed up the computation. However, some parts could still be improved ‘for example the Comparative Text Mining algorithm is not fully parallelizable. Extending this work would improve the performance in terms of time complexity. Another limitation was the queue waiting time for a job to start.”

“Comparing hardware between supercomputers, there was a high variation in job durations, for example Bridges was exceptionally faster compared with Greenfield. Having a stable environment with short waiting time would help in creating more advanced retrieval models.

“Moreover, we did not leverage supervised machine learning methods to learn author properties, such as gender or additional metadata. Being able to develop and implement such a model would increase our accuracy and reveal a more representative (document) collection sample regarding experiences and views of black women. Finally, in the future we could also generalize our work as a general toolkit for retrieving a subset of documents with specific metadata constraints."

CURRENT BIG DATA PROJECT

In May 2017, Ruby Mendenhall received a Faculty Fellowship at the National Center for Supercomputing Applications (NCSA) for her project "Using Wearable Sensors and Affective Diaries to Document How Violence Affects Public Life and Public Health." This study seeks to examine the physiological effects of exposure to nearby gun crimes such as shootings as a way to document the lived experiences of African American mothers. We seek to gather one month of physiological data from African American mothers using wrist-worn wearable biosensors as well their physical movements as collected by their smartphone's GPS. Researchers will collaborate with the Chicago Police Department to analyze the big data in Chicago around crime and other measures of lived experiences.

About Hyperion Research, LLC

Hyperion Research, consisting of the former IDC high performance computing (HPC) analyst team, provides HPC information, analysis, and recommendations based on technology and market trends. Research includes market sizing and forecasting, share tracking, segmentation, technology and related trend analysis, and both user & vendor analysis for multi-user technical server technology used for HPC and HPDA (high performance data analysis). We provide thought leadership and practical guidance for users, vendors and other members of the HPC community by focusing on key market and technology trends across government, industry, commerce, and academia.

Headquarters

365 Summit Avenue
St. Paul, MN 55102
USA
612.812.5798
www.hpcuserforum.com

Copyright Notice