USING ADVANCED COMPUTING TO RECOVER BLACK WOMEN’S LOST HISTORY:
EXPLORING SEAMLESS CREATIVITY IN RESEARCH THAT SPANS SOCIOLOGY, AFRICAN AMERICAN STUDIES,
GENDER AND WOMEN’S STUDIES, COMPUTER SCIENCE, ENGLISH AND DIGITAL HUMANITIES

PEARC 18 Conference - Practice and Experience of Advanced Computing
July 25, 2018

Ruby Mendenhall*
Nicole Brown, Michael L. Black, Mark Van Moer, Ismini Lourentzou, Karen Flynn, Malaika Mckee, and Assata Zerai

*Assistant Dean of Diversity & Democratization of Health Innovation

Sociology, African American Studies, Urban and Regional Planning, Gender and Women’s Studies, and Social Work

Faculty Affiliate Woese Institute for Genomic Biology; Institute for Computing in the Humanities, Arts, and Social Sciences; Women and Gender in Global Perspectives; the Cline Center for Advance Social Research; Epstein Health Law and Policy Program; and Family Law and Policy Program
RESEARCH TEAM & COLLABORATORS

Ismini Lourentzou – Research Assistant in Computer Science
Nicole Brown – Former Soc, African and African American Studies, Stanford University
Ruby Mendenhall – Sociology & African American Studies
Karen Flynn – African American Studies, Gender & Women’s Studies
Mark Van Moer – Visualization Programmer, NCSA/XSEDE-ECSS
Malaika McKee – African American Studies
Mike Black – Former I-CHASS/NCSA, University of Massachusetts, Lowell
Assata Zerai – Associate Chancellor for Diversity
Harriett Green - English and Digital Humanities Librarian
Chengxiang Zhai - Computer Science
Michael Simeone – Former I-CHASS/NCSA, Arizona State University
Kevin Franklin – Executive Director of I-CHASS
Marshall Scott Poole – Director of I-CHASS
PRESENTATION OVERVIEW

Demystify - Story Interdisciplinary Research and Seamless Creativity

First Exposure to NCSA – Kevin Franklin and Alan Craig

Start of Project - Michael Simone

Motivation for Study – Recovering Black Women’s History

Findings/Challenges - JSTOR (Journal Storage) and HathiTrust from 1746 to 2014

New Studies and Potential of Big Data
MY CURIOSITY AND BIG DATA

First Exposure to NCSA – Kevin Franklin

• Blank sheet of paper with an image that came to life (Alan Craig’s technology)

• White paper, “Rethinking 21st Century Urban Transformations: Race and the Ecology of Violence” proposed cyber infrastructure to capture unheard stories about violence (return to this)


NSF “The science that these papers collectively envision is data intensive, multidisciplinary, collaborative, and frequently problem-oriented” (p. 16).
Worked with Alex Yahja on K01 proposal to use network analysis to visually map Black mothers’ social networks and how they are affected by violence (e.g., murders, shootings, rapes, etc.) and where it occurs.
BRAINSTORMING ABOUT BIG DATA & SOCIOLOGY

Michael Simeone lectures in methods class

Talked for about 30 minutes

How big data could relate to my sociological research questions
2012 – Proposal Not Accepted

2013 - Visualizing Topic Models about African American Women’s Everyday Experiences and Standpoints

Goal: Search millions of periodicals, books and newspapers in JSTOR AND the HathiTrust to identify conversations and group knowledge (standpoint).
MOTIVATION FOR STUDY
RECOVERING BLACK WOMEN’S HISTORY

• Often, literature by and about African American women is inaccessible.
• Alice Walker’s Search for Zora Neal Hurston’s Grave – Call & Response
• Project’s goal - Recover what was written about their ideas, challenges, actions/agency, and accomplishments
RESEARCH QUESTIONS

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?
CHALLENGE – TIMELY DATA SECURITY AGREEMENTS

HathiTrust Digital Library

• Case study between Illinois and HathiTrust Research Center to make the digital content accessible and usable for research

• Public Domain – prior to 1923
### CHALLENGE SEARCH TERMS

**TEXT NOT BY OR ABOUT BLACK WOMEN**

<table>
<thead>
<tr>
<th>GROUP A: Race</th>
<th>GROUP B: Gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>Black</td>
<td>wom?n</td>
</tr>
<tr>
<td>Afr* American</td>
<td>female?</td>
</tr>
<tr>
<td>negr*</td>
<td>girl?</td>
</tr>
<tr>
<td>colored</td>
<td>lady</td>
</tr>
<tr>
<td>nig*</td>
<td>ladies</td>
</tr>
</tbody>
</table>

Conducted proximity searches (w/5) in the Solr index metadata for the HathiTrust Research Center corpus: Searched for all combinations and variants of Group A and Group B terms.
EXAMPLES OF VOLUMES RESULTING FROM THE SEARCH
1746 - 2014 ~800,000 DOCUMENTS

JSTOR (academic journals & books)


HathiTrust

- Journal of the National Medical Association (Black medical care and disparities from ~1909-current)
- The Negro at Work during the War and during Reconstruction by U.S. Department of Labor 1921.
STANDPOINT THEORY

Seeks to uncover the pivotal role of knowledge in reproducing and dismantling social inequality.

It is group knowledge based on shared common experiences such as oppression.

Links the everyday lived experiences of Black women to interlocking systems of race, class, and gender discrimination (Collins 1998:281).
METHODS – LDA AND CTM

Latent Dirichlet Allocation (LDA)
Discover patterns of word distribution
- within documents
- across a corpus
using Bayesian probability
Per-topic word distributions
Per-document topic distributions

Comparative Text Mining (CTM)
Discover similarities and differences among topics
Comparison of
1. sets of common topics across entire corpus
2. variations inside topics across specific time periods (generative probabilistic model)
Topics 20- legal battles, court, property.

Unclear if property/estate referred to slaves or land

Question: Are Black female slaves taking cases to court?

By 1846, 575 Freedom Suits, ~60% of the time slaves won – “golden age”

- Sojourner Truth’s son, illegally sold
- Went to court & won
TOPIC TO TOPIC CONNECTIONS

Network visualization to show how topics were connected via Pearson’s correlation coefficient.
1808 – Abolition of Slave Importation
1831 – Nat Turner Rebellion
1851 - Uncle Tom’s Cabin published
CTM MODELS INTERACTIVE TREE MAPS

Tree map: words for common models (entire corpus) and proportion of documents in expert models (time periods e.g., temporal clusters like slavery or WWII) that match common models. Click on expert models to see which words are important during this time frame. Goal: Quick visual overview of how much expert models contributed to common models.
METHODS

SeRRR (Search, Recognition, Rescue and Recover)

**Search** (or call) train topic model using a subset of 20,000 documents

**Recognition** intensive intermediate and close readings to identify potentially new documents that were not identified before as being by or about Black women’s lived experiences.

After confirmations, **Rescue** and place them in the **Recovered** corpus about Black women.

We plan to make the recovered documents available to librarians, scholars and community members.
KL Divergence and Cosine Similarity

**Search:** Similarity and dissimilarity of 800,000 documents

Cosine Similarity – range 1 to 0, 1 most similar

KL Divergence – probability distributions, lower numbers more similar

**Recognition:** Close and intermediate readings
1. FINDINGS – MISSING SUBJECT METADATA

Pulled 300,000 HathiTrust volumes, about 80,000 (~27%) did not have subject metadata.

Suggests that if researchers searched for volumes about Black women, they may not have access to a significant amount of documents that may be relevant.

If not tagged properly, need to know the documents exist.
2. FINDINGS – WRITING AS AN ACT OF PRIVILEGE

Challenge to recover documents that centered Black women’s lived experiences

Writing & entering the historical record, acts of power and privilege

Unusual texts contained info on Black women

Often had to recover their voices through the voices of others, often White men

Black children were discussed with limited references to their mothers.

Congenital complications and infant mortality, diseases, and general health issues

Standpoint insights: mothering a sick child, death and grief, their access to medical care, etc.

Black mother and 5 year old son with diarrhea, which he had for one year. Noticed blood in his stool.

Information on social class (her child was undernourished, and she was referred to a charity hospital).

Insight racial context in which the mother was raising her child (they were farmers and the doctor reported the child’s diet reflected a typical diet for Blacks).
3. FINDING – BLACK WOMEN’S BODIES AND MEDICAL ADVANCES

Black women’s complicated relationship with the field of medicine is critical to understanding advances in general medicine, OBGYN, and anesthesia in the United States.

Given that there are multiple texts on this subject, it suggests that there is a collective (group knowledge/standpoint), as opposed to individual experiences that requires articulation.

Texts are from period when American Medical Association was established (1847), exploitation in testing/medical procedures
4. FINDING – FINDING NEEDLES IN BIG DATA HAYSTACKS TO RECOVER LOST HISTORY

Reviewed Documents

- Intermediate Readings 5,000 (metadata): 50% - 70%
  - Recovered ~ 2,500 – 3,500 documents

- Close Readings of Entire Documents: 70% - 90%
  - Recovered 485 documents
Looking for the needles

- *Memoir of Salome Lincoln* by Almond H. Davis in 1843, “flourishing society on the island [Nantucket], made of coloured people” and singers in integrated church (p. 104).
Looking for the needles

LESSON LEARNED – LOTS OF SUPPORT ON CAMPUS

Data agreement delays affected our ability to complete the study within the fellowship period.

XSEDE: Extreme Science and Engineering Discovery Environment

Extended Collaborative Support Service (ECSS)

Team at Pittsburgh Supercomputing Center

Chengxiang Zhai in Computer Science offered to help – grad student
COMPUTATIONAL PROCESSING TIMES

CTM more complex LDA – Common model and expert models, need more computational time. Took 5 days on Greenfield to create 25 topics and 8 expert models for each topic.

Inferencing/testing – 2 days on Greenfield and Bridges supercomputers at University of Pittsburgh
- Greenfield used lot of processing units, so we exhausted resources
- Bridges lets you define memory needs, so lowers computing costs
Parallelized the inference and ranking procedures, took 1.5 days.

- If we used sequential processing of the document collection, it would have taken 90 days to finish ranking of the 800,000 documents using one metric and one topic.

Training step was not easy to parallelize and it did not produce any speed ups since the algorithm has to wait for all expert model calculations to finish before comparing among iterations to check for convergence.
### PROCESSING TIMES CONT.
### INFERENCEING CTM MODELS

<table>
<thead>
<tr>
<th>Supercomputer</th>
<th>Time</th>
<th>Service Units of Time (SUs_)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Greenfield</td>
<td>All Models – 168 hrs</td>
<td>~5,000 (terminated, exceeded wall time)</td>
</tr>
<tr>
<td>Greenfield</td>
<td>1 Model – 75 hrs</td>
<td>2,253</td>
</tr>
<tr>
<td>Bridges</td>
<td>1 Model - 81 min</td>
<td>77</td>
</tr>
</tbody>
</table>
CREATION OF NEW KNOWLEDGE

How is inequality expressed (or hidden) in the everyday lives of African American women?

How do they seek to change entrenched interlocking systems of oppression (racism, classism, sexism, etc.)?

Slavery  Reconstruction  Lynching  Civil Rights  Black Lives Matter
“In the 1940s, a group of female scientists were the human computers behind the biggest advances in aeronautics. Hidden Figures...book and film tells their remarkable, untold story.”

Margot Lee Shetterly’s Book called *Hidden Figures* (2016)

Christine Darden, 1975

Imagine in 2016 Film
Black Women’s Contributions to the Du Bois-Atlanta School and to American Sociology


Recovering the genius of Du Bois and the erasure of Black women’s contributions to Du Bois-Atlanta school and American sociology
WOMEN IN DU BOIS’ SOCIAL NETWORKS

How (in)visible are the black women who were contemporaries of Du Bois (Anna Julia Cooper, Ida B. Wells-Barnett and Mary Church Terrell) compared to their white women counterparts (Jane Addams and Mary White Ovington) and compared to Du Bois.
DU BOIS’ NETWORK FINDINGS

800,000 documents by or about Black women, subset of 183 documents

Du Bois towers over all the women with 42% of the citations

Jane Addams, 23%
Mary White Ovington, 4%

Black Women
Ida B. Wells-Barnett, 8%
Anna Julia Cooper, 7%
Mary Church Terrell, 4%

White Women
Dorothy Gautreaux (1927-1968)

Chicago Freedom Movement: The Gautreaux lawsuit argued that CHA and HUD violated the equal protection guarantee under the Constitution and Title VI of the 1964 Civil Rights Act. Education activist too.

She died in 1968 in her prime, at the age of 41 (stroke, kidney failure, too sick to carry last baby, etc.)
NEW NADIR IN BLACK HISTORY (CHA-JUA 2009)

- Public housing demolition and gentrification of many inner-city neighborhoods
- Great Recession and worst housing crises in U.S. history and under-funded education
- Prison industrial complex
- Shrinking (welfare) safety net
- Sustained postindustrial unemployment
- Police Shootings
- Neighborhood violence
HOW GUN VIOLENCE AFFECTS PUBLIC LIFE & PUBLIC HEALTH

White paper: cyber infrastructure to capture unheard stories about violence

How Gun Violence Affects Public Life and Public Health

The exhibit is about a study that used wearable sensors to document the impact of violent crime on residents from Chicago’s Englewood community and its relationship to public health.

Exhibit on Sat, Aug 4, 2018
2 – 5 pm
5139 S. Ashland, Chicago

Journal about Lived Experiences
Heart Rate and Sleep Data

A Typical Month in 2010

Partners:
100,000 BLACK WOMEN CITIZEN SCIENTISTS

Double V Campaign: Victory Home & Abroad

Double H Campaign: Health & Healing

How is inequality expressed (or hidden)?

How do they engage in social change?

Forefront of social & health solutions such as Black maternal and infant mortality (disparities similar to 1850s), diabetes, cancer, depression, PTSD, etc.
11,000 higi centers
6.9+ million account holders
271+ million tests
Data: pulse, weight, heart rate, blood pressure, steps, BMI, body fat, gym check-in
higi’s survey tool provides a flexible, tune-able data collection mechanism

### Sample Social Determinant Survey Questions:

In the last 12 months, have you **needed to see a doctor, but could not because of cost?**

In the last 12 months, did you ever **eat less than you felt you should** because there wasn’t enough money for food?

Are you worried that in the next 2 months, you **may not have stable housing?**

In the last 12 months, has your utility company shut off your service for **not paying your bills?**

In the last 12 months, have you ever had to **go without health care because you didn’t have a way to get there?**

Do you ever **need help reading hospital materials?**

Are you afraid you **might be hurt in your apartment building or house?**

Do problems getting **child care** make it difficult for you to work or study?

---

Longitudinal health and social data from “outside the walls of healthcare”

How often do you hear the sounds of gun shots?
Are you afraid that your children will witness someone being shot?
Are you afraid your child will be killed or injured by a gun violence?

Relationships to stroke, hypertension, infant and maternal mortality, depression, etc.
TRAINING STUDENTS OF COLOR AND OTHER DIVERSE GROUPS

Train undergraduate and graduate students to work in interdisciplinary groups that design, build, and use HPC and big data

Advance Computing for Social Change – Daring Greatly! – 3rd year at Super Computing Conference (Dallas, TX, Nov 2018)

Extreme Science and Engineering Discovery Environment (XSEDE) and the Texas Advanced Computing Center (TACC)
ADDITIONAL RESOURCES

Recovering Lost History Podcast by Mendenhall et al. (Tennessee Supercomputing):
https://soundcloud.com/tennessee-supercomputing/recovering-lost-history

Rescued History by Ken Chiacchia and Aaron Dubrow. NSF Where Discovery Begins:

An Illinois Sociologist Uses Supercomputing to Recover the Lost History of Black Women by Karis Hustad: http://chicagoinno.streetwise.co/2016/03/16/a-supercomputer-helps-uiuc-researchers-recover-lost-history/

http://dl.acm.org/citation.cfm?id=2949642&CFID=665151917&CFTOKEN=74793502
THANK YOU

XSEDE - Extreme Science and Engineering Discovery Environment

Extended Collaborative Support Services

Chengxiang Zhai - Computer Science