RESCUING LOST HISTORY: USING BIG DATA TO RECOVER BLACK WOMEN'S LIVED EXPERIENCES

HPC User Forum
September 8, 2016 – Austin, TX

Funder: XSEDE -Extreme Science and Engineering Discovery Environment, ECSS

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

*Sociology, African American Studies, Urban and Regional Planning, and Social Work
Faculty Woese Institute for Genomic Biology
Institution for Computing in the Humanities, Arts, and Social Sciences

RESEARCH TEAM & COLLABORATORS

Ismini Lourentzou – Research Assistant in Computer Science

Nicole Brown – Sociology, Illinois Wesleyan University

Ruby Mendenhall – Sociology & African American Studies

Karen Flynn – African American Studies, Gender & Women Studies

Mark Van Moer – Visualization Programmer, NCSA/XSEDE-ECSS

Malaika McKee – African American Studies

Mike Black – Former I-CHASS/NCSA , University of Massachusetts, Amherst

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 – Associate Director of NCSA, Executive Director of I-CHASS

Marshall Scott Poole – Director of I-CHASS

PRESENTATION OVERVIEW

Workforce Development - First Exposure to NCSA – Kevin Franklin

Start of Project - Michael Simone

Motivation for Study – Recovering Black Women's History

Findings/Challenges

Lessons Learned

10 Copies of Paper on this Topic

MY CURIOSITY AND BIG DATA

First Exposure to NCSA – Kevin Franklin

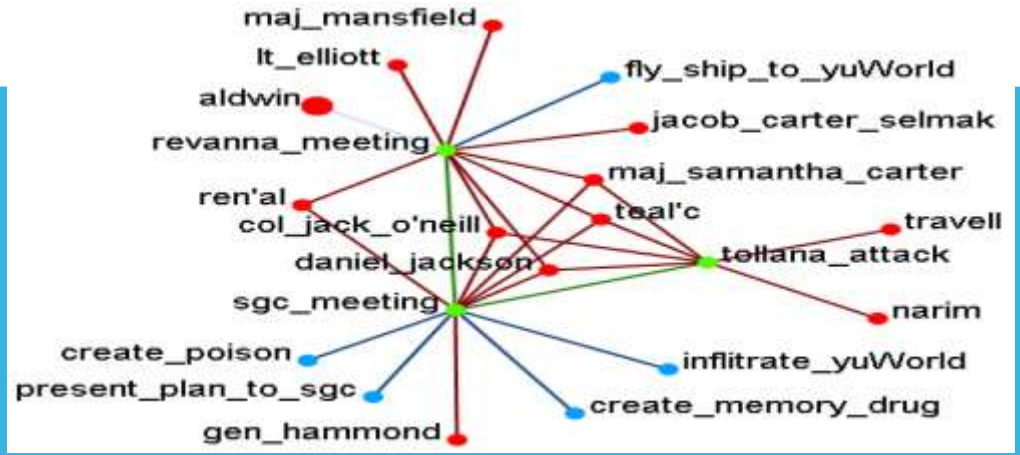
- Blank sheet of paper with an image that came to life (Alan Craig's technology)

White paper, "Rethinking 21st Century Urban Transformations: Race and the Ecology of Violence" proposed cyber infrastructure, proposed to capture unheard stories about violence

(Current documentary about mothers who lost children to violence in Chicago).

NIMH K01 GRANT PROPOSAL

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

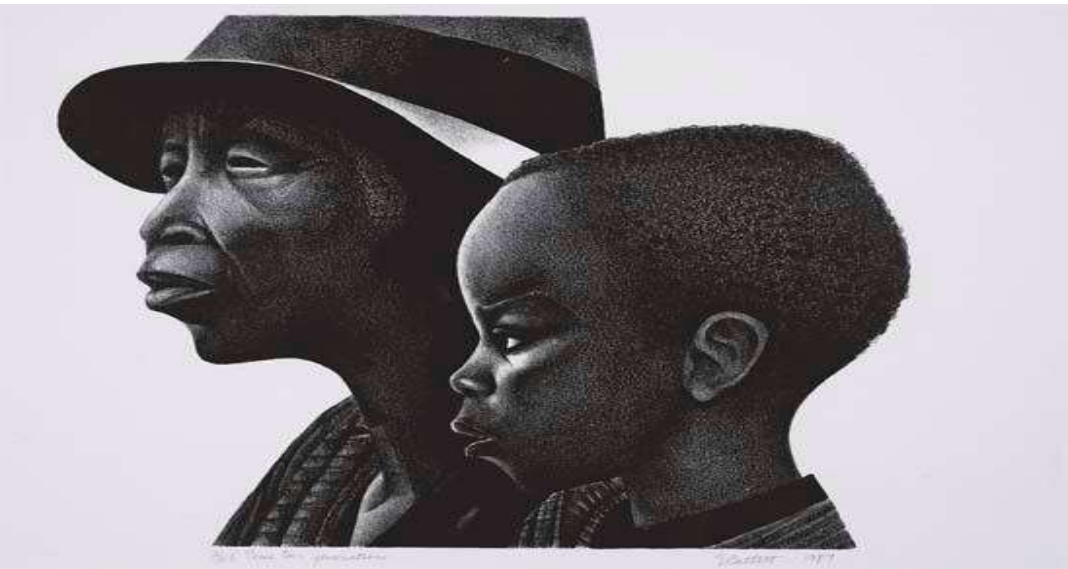
INSTITUTE FOR ADVANCED COMPUTING APPLICATIONS AND TECHNOLOGIES (IACAT) FACULTY FELLOWS PROGRAM PROPOSAL

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
- Partnership of academic & research institutions, offering a collection of millions of titles digitized from libraries around the world.
- Public Domain – prior to 1923

CHALLENGE SEARCH TERMS
TEXT NOT BY OR ABOUT BLACK WOMEN

GROUP A: Race

- Black
- Afr* American
- negr*
- colored
- nig*

GROUP B: Gender

- wom?n
- female?
- girl?
- lady
- ladies

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)

- Changing Racial Labels: from “Colored” to “Negro” to “Black” to “African American” by Tom Smith. *Public Opinion Quarterly* 1992

HathiTrust

- *The Crisis* - W.E.B. DuBois
- *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.

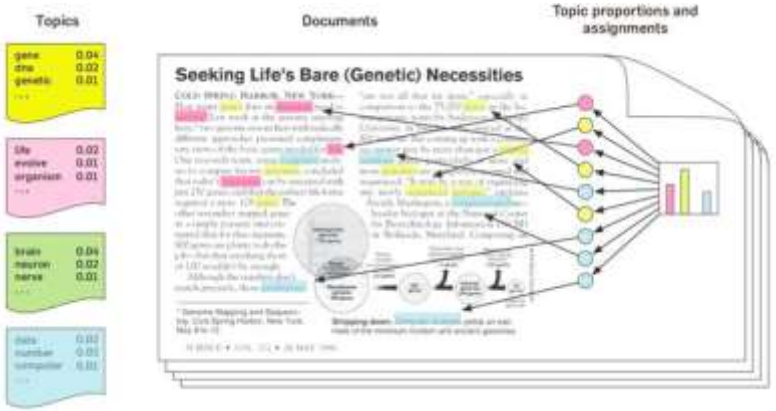
STANDPOINT THEORY

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)

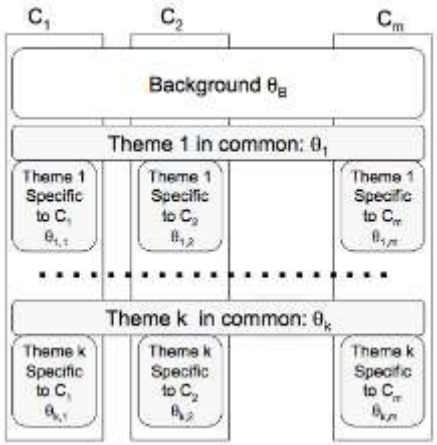
Discover patterns of word distribution

- 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)

NAMING THE TOPIC 20

Topics 20- legal battles played out in court. Are Black female slaves taking cases to court?

Freedom Suits. 575 by 1846, and ~60% of the time slaves won – "golden age"

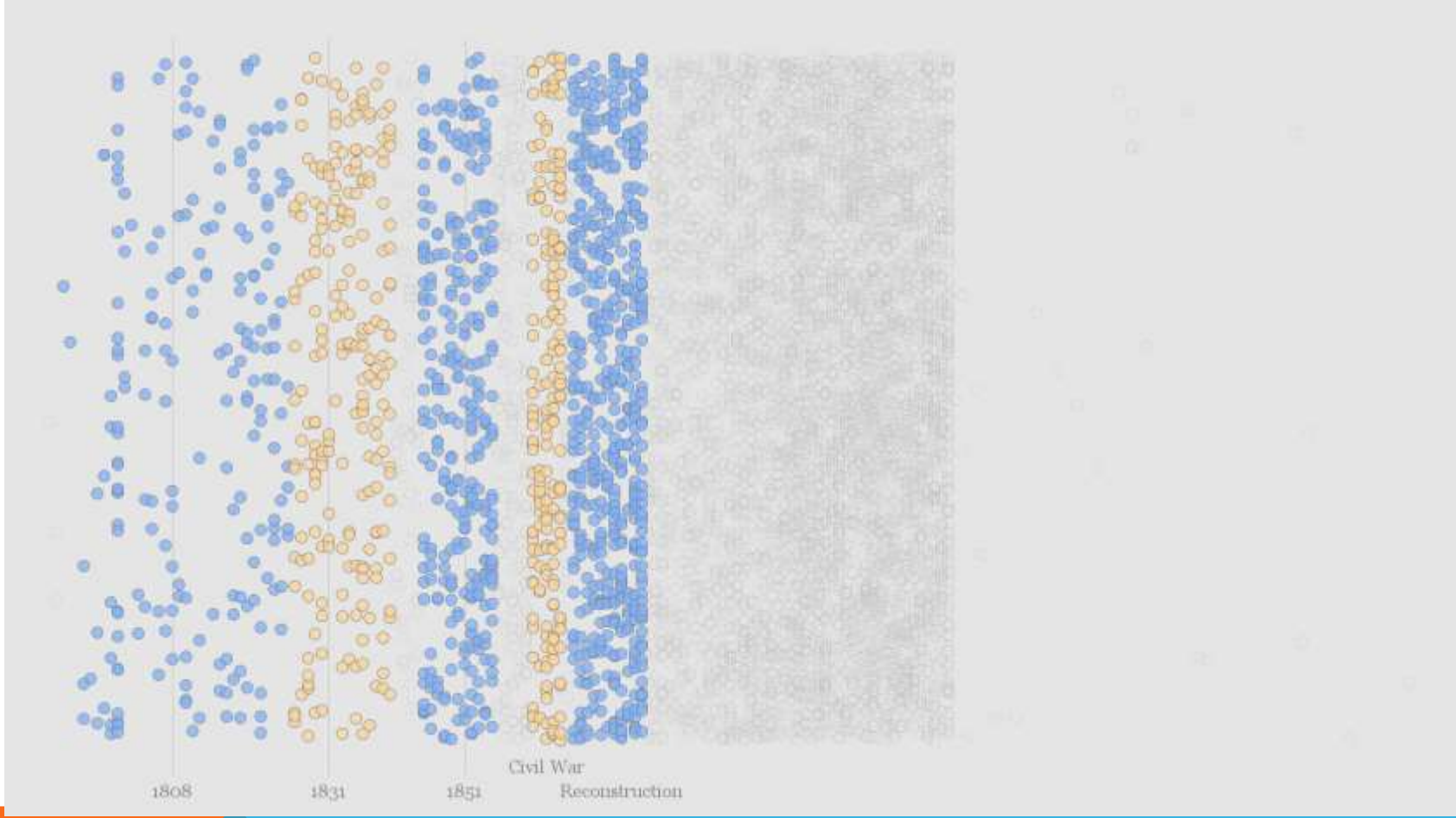
Unclear if property/estate refers to slaves or land (historical period will tell)

- Sojourner Truth's son, illegally sold
- Went to court & won

TOPIC TO TOPIC CONNECTIONS

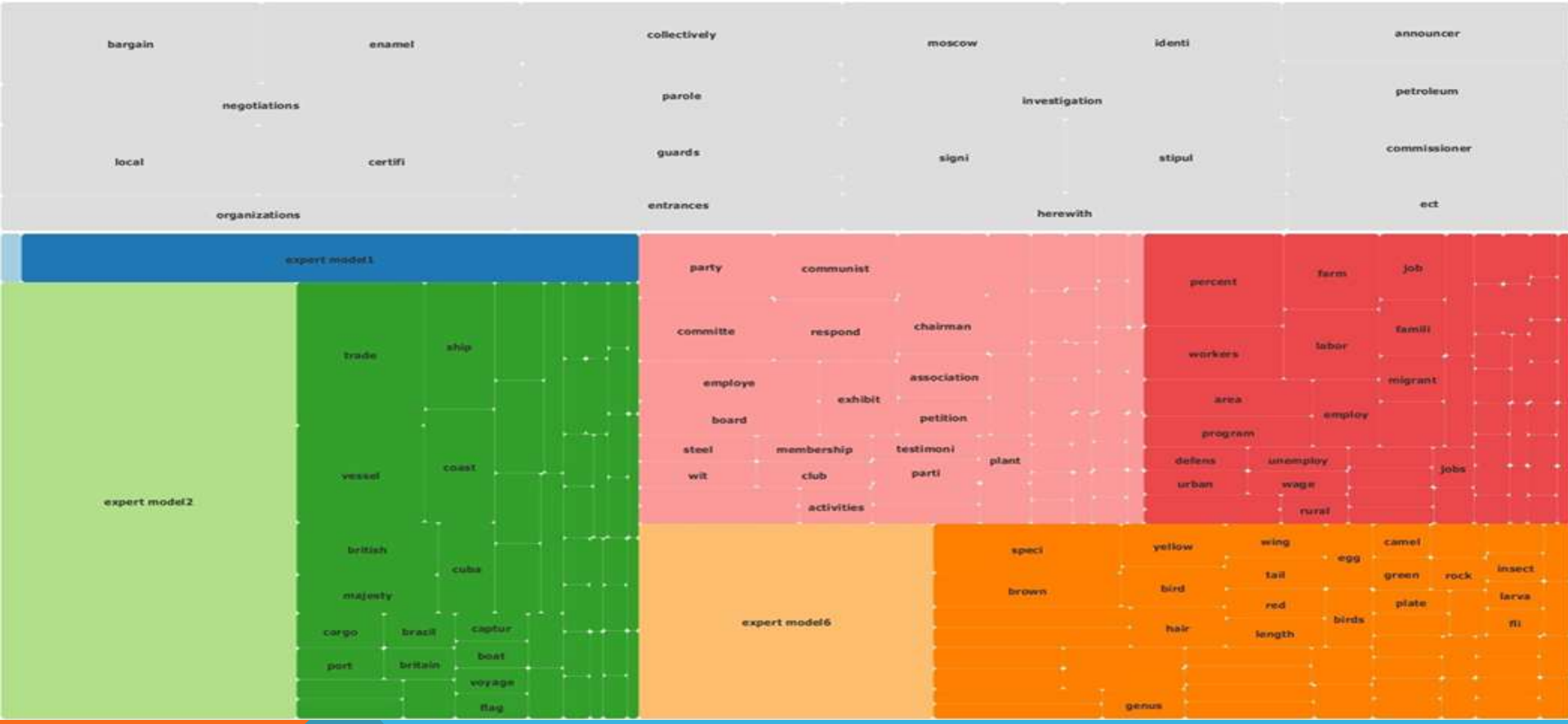
Network visualization to show how topics were connected via Pearson's correlation coefficient

TIME FRAMES OF INTEREST



1808 – Abolition of slave importation
1831 – Nat Turner Rebellion
1851 - Uncle Tom’s Cabin published

CTM MODELS INTERACTIVE TREE MAPS



Tree map: proportion of terms within common (corpus) and expert models (time periods) and proportion of expert models within clusters. Click on subdivisions to see content. Quick visual overview of how much expert models contributed to containers.

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.

Search: KL Divergence and Cosine Similarity

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

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 document that may be relevant.

If not tagged properly, need to know they 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

Recover their voices through the voices of others, often White men

Article: Brown, N.M. Mendenhall, R. Black, M.L. Van Moer, M., Zerai, A., Flynn, K. 2016. Mechanized Margin to Digitized Center: Black Feminism's Contributions to Combatting Erasure within the Digital Humanities. *International Journal of Humanities and Arts Computing* 10(1): 110-125.

ACT OF PRIVILEGE CONT - EXAMPLE

American Journal of Diseases of Children - 1918

- 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 information, etc.
- Black mother brought her 5 year old son in for 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.

Text are from period when American Medical Association was established (1847)

Exploitation and testing medical procedures

4. FINDING – FINDING NEEDLES IN BIG DATA HAYSTACKS TO RECOVER LOST HISTORY

The use of topic modeling led to the recovery of 124 previously unidentified documents related to Black women from the KL divergence list.

Using cosine similarity, we were able to recover an additional 26 previously unidentified documents.

Ongoing research to test model parameters may lead to the recovery of additional documents.

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)

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. Takes 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 uses lot of processing units so exhaust resources
- Bridges lets you define memory needs, so lowers computing costs

PROCESSING TIMES CONT.

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.
INFERENCING CTM MODELS

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

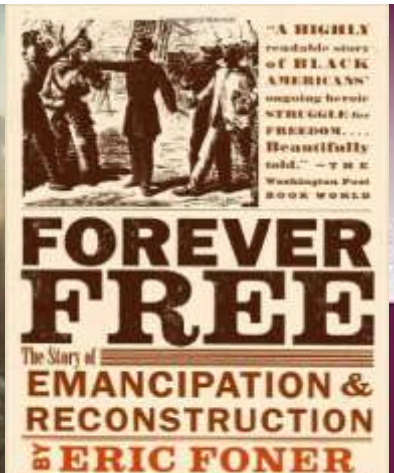
CREATION OF NEW KNOWLEDGE

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

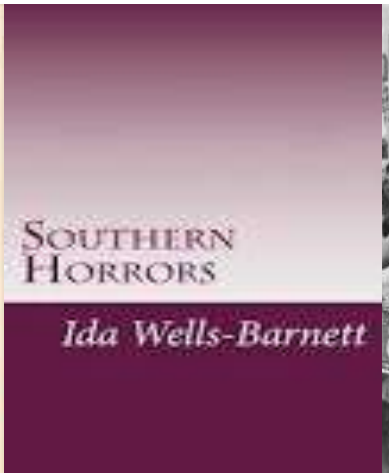
How do they seek to change entrenched interlocking systems of oppression (racism, classism, sexism, etc.)?

Slavery

Reconstruction



Lynching



Civil Rights

Black Lives Matter

DIVERSE WORKFORCE DEVELOPMENT

Nicole Brown – Former graduate student who is now doing her own excellent big data projects

Working on a technology product to provide university administrators and industry leaders with real-time information about racial and campus climate. We expect to have big data that can facilitate welcoming environments that improve recruitment, retention, and promotion.

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

HOW HISTORY FORGOT THE BLACK WOMEN BEHIND NASA'S SPACE RACE

"In the 1940s, a group of female scientists were the human computers behind the biggest advances in aeronautics. Hidden Figures, an upcoming book and film tells their remarkable, untold story." (quote from website: https://www.theguardian.com/lifeandstyle/2016/sep/05/forgot-black-women-nasa-female-scientists-hidden-figures

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

Christine Darden,1975

Imagine in 2016 Film

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: http://www.nsf.gov/discoveries/disc_summ.jsp?cntn_id=137797

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/

Rescued Lost History - Extreme Science and Engineering Discovery Environments 2016 (XSEDE16) Conference Proceedings (Paper): http://dl.acm.org/citation.cfm?id=2949642&CFID=665151917&CFTOKEN=74793502


Thank You