What is Big Data?

“Datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze”

-McKinsey Global Institute, Big data: the next frontier for innovation, competition, and productivity, May 2011
Opportunity: The Internet Archive contains the largest single record of the history of the World Wide Web from 1995 to the present—a wealth of untapped research data.

Challenge #1: There is a significant lack of research-ready databases and tools available to the scholarly community

And who really needs 8.5PB of data?
Louisiana and Mississippi declare emergencies

People along the Gulf Coast got some very unwelcome news just before the weekend — Hurricane Katrina appears headed their way and may become a destructive monster by the time it arrives. Governors of Louisiana and Mississippi declared states of emergency Friday. Damages from Katrina are estimated at between $50 million and $2 billion.

'Big Easy' braces for the 'big one'

Residents of New Orleans and the surrounding area jampacked highways Sunday, evacuating from the path of Hurricane Katrina. A direct hit from the huge storm could be "potentially catastrophic" for New Orleans with officials warning of power outages and widespread flooding. But more than 10,000 people opted to remain in the city, taking shelter at the Louisiana Superdome.
Challenges in Tool Development

• Challenging to develop longitudinal data sets on an ad-hoc basis
  - Longitudinal studies are reduced to annual data summaries (Payne & Thelwall, 2007)

• Existing social science methods for handling data and conducting analysis have not proven adequate for large-scale data (Lazer et. Al. 2009)

• IA is structured implicitly, as patterns of associations and interactions are inferred from the nature of content with the sites themselves rather than by formal structures (Arms et. al. 2006)
Opportunity: Scholars from a wide range of fields have the potential to benefit from the advancement of research ready tools to access archival Internet data.

Challenge #2: Expertise in large-scale data analysis resides across multiple disciplines.

HistoryTracker Tool

- Version 2.0
- 20th Century Collection @ RU
- Curated Data Sets
- PIG Scripts in Hadoop Environment
- Link Lists & Other Data
- RU High-Speed Computing Cluster
HistoryTracker: The Data

WATs

- A compressed metadata format
  - Essential metadata for many types of analysis
- Avoids barriers to data exchange

- WATs
  - Metadata format
    - Date archived
    - Content type
    - Outlinks
    - Anchor text for links
    - Size of record
    - Pointer to actual content

Challenge #3: Data are dirty.

Theory is not dead.
Previous Research Sets a Path

- Use of the IA to analyze the frequency of updates based on content analysis (Brock, 2005; Veronin, 2002)

- Measures of age and frequency of updating have been validated against third-party data (Murphy, Hashim and O'Connor, 2005)

- Longitudinal study of e-commerce Web sites to examine development of content (Chu, Leung, Van Hui and Cheung 2007)

- Evolution of organizational hyperlinks in relation to fluctuations in resources (Weber, 2012)
Deciding What Matters

- For a crawl of the 2012 extraction of the Internet:
  - Map output records 1,127,187,232,984 Links

- When does a link matter?
  
<table>
<thead>
<tr>
<th>ELEMENT</th>
<th>COUNT</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;a href=&gt;</td>
<td>3,212</td>
</tr>
<tr>
<td>&lt;img src=&gt;</td>
<td>3,195</td>
</tr>
</tbody>
</table>

- But how do you define what matters?
  - Subcomponents?
  - Text? (anchor text!)
Research Subset: Political Activity

- Database on 109th, 110th, 111th, 112th congresses
  - .gov domain & congressional pages
  - Full snapshots of the full .gov domain taken at the end of each congressional session

- Research Potential
  - Web promises potential to study the growth of political activity in online environments (Adamic & Glance, 2005; Bruns, 2007; Chang & Park, 2012)

Research Subset: Media Organizations

- Database on US Media organizations
  - Major media outlets, blogs, opinion sites, etc.

- Synopsis
  - Dataset extracted from a seed list of 4,891 news media organizations
  - 2008 - 2012
    - 1.7 million URL captures from 2008 – 2012

- Research Potential
  - Previous studies of news media organizations (Greer & Mensing, 2006; Weber, 2012; Weber & Monge, 2011) set forth methodology for examining media organizations via online representations
Research Subset: Social Movements

- Database on social movements
  - Occupy Wall Street
  - Focus on supporting existing avenues of research

- Synopsis
  - 2010 – 2012
  - 528 seed URLs

- Research Potential
  - Previous research on NGOs in the online environment (Bach & Stark, 2004; Shumate, 2003, 2012; Shumate, Fulk, & Monge, 2005)

Research Subset: Disasters

- Database on social movements
  - Superstorm Sandy
  - Hurricane Katrina

- Synopsis
  - 2003 – 2012
  - 672 seed URLs

- Research Potential
  - Online networks as enabling organizational resilience (Chewning, Lai and Doerfel, 2012; Perry, Taylor and Doerfel, 2003)
• Want to get involved?
  – Email me! matthew.weber@rutgers.edu
  – ResearchHub launching in August – go to netsci.rutgers.edu
  – Workshop: Spring 2014 @ Harvard Institute for Quant. Social Science

• Collaborators
  – Kris Carpenter, Internet Archive
  – David Lazer, Northeastern University
  – Peter Monge, University of Southern California

Research support by NSF Award #1244727

Research Example

• Political Networks

<table>
<thead>
<tr>
<th>Subset</th>
<th>Year of Archive</th>
<th>Recorded Links</th>
<th>Reciprocal</th>
</tr>
</thead>
<tbody>
<tr>
<td>109th</td>
<td>2006</td>
<td>54K</td>
<td>10K</td>
</tr>
<tr>
<td>110th</td>
<td>2008</td>
<td>82K</td>
<td>18K</td>
</tr>
<tr>
<td>111th</td>
<td>2010</td>
<td>253K</td>
<td>34K</td>
</tr>
<tr>
<td>112th</td>
<td>2012</td>
<td>346K</td>
<td>43K</td>
</tr>
</tbody>
</table>
Mapping the News Ecosystem

1998

2002

2006

'Digital' organization forms
'Traditional' organization forms