# Data Mining at Duke

("What to do with all of those hard drives")

Molly Tamarkin
Associate University Librarian for Information Technology Services

Joel Herndon
Head, Data & GIS Services

# Today's Talk

- Rise of text analysis questions
- Challenges in providing text analysis services
- Duke University Libraries' response


# Brandaleone Center for Data and GIS Services

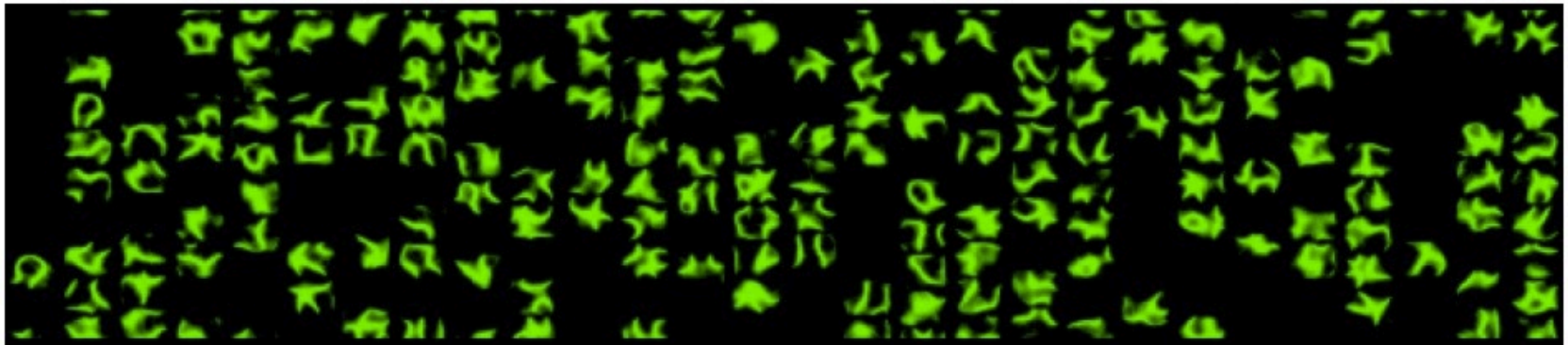
# The Rise of Text as Data

# New Questions for Research Libraries


- How has the North American press covered environmental issues over the last 20 years?
- Can we analyze all (17000) journal articles on German studies in the 20th century?
- What might tweets reveal about the Arab Spring in social media?

# Duke Libraries + Digital Scholarship

*discovering, exploring, and building new forms of scholarship*



Home PhD Lab in Digital Knowledge Text > Data

## Text visualization of presentations? Check.

Posted on 1 November 2012 by em160@duke.edu

If you missed today's Text > DATA presentation by Greg Appelbaum & Elizabeth Beam, "Mapping the Disciplinary Structure of Cognitive Neuroscience through Semantic & Network Analysis," you can check out the presentation slides and storified tweets of their talk. Slides and storifies are available for past presentations in this series (scroll to the presentation you're interested in under Text > Data Schedule + Topics).

**News**

- Text visualization of presentations? Check.
- Stay ahead of the pack with savvy citation management
- NVivo: The free qualitative analysis tool you didn't know you needed
- To Publish or To Make Public?
- The place of social media in one's academic work (and life)

**Archives**

Select Month

**Tweet Blender**

twitter

http://sites.duke.edu/digital/

# Duke Qualitative Research Community


## Overview of 11/16 meeting

Posted on November 26, 2012 by Alexandra Cooper, Ph.D.

**Meeting overview:**

The qualitative working group met 11/16. Two researchers presented works-in-progress; Judith Adams (School of Nursing) presented preliminary findings from interviews with family members of ICU patients, and Clarissa Lee (Literature) presented an overview of her developing plans to explore the creation, transmission, and movement of information within the field of particle physics. With Q&A, each presentation lasted about 30 minutes; discussion included tips on strategies for collecting relevant and useful information and for focusing the research, and on resources at Duke relevant to each of these projects. Clarissa provided a written overview of her research plans that is now available via the

Search

### Recent Posts

- Overview of 11/16 meeting
- Overview of 11/2 meeting and results
- Qualitative working group forming
- January 26th, Pamela George on Coaching Young Researchers, 12 pm, Old Chem 104
- Schedule for Spring 2012

### Recent Comments

- Yvonne Belanger on Schedule for Spring 2012

### Archives

# Challenges in Providing Text Analysis Services

# Challenges

- Collections
- Licensing
- Infrastructure
- Service model

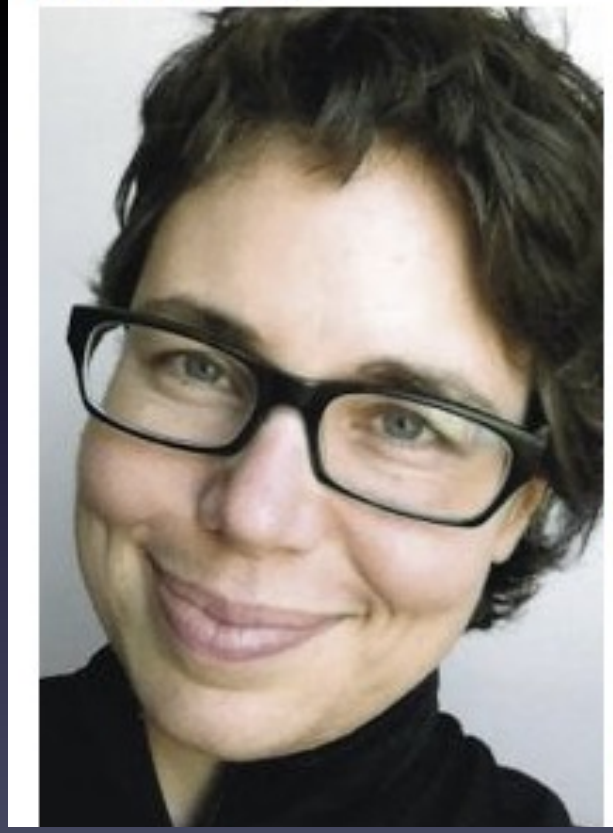
# Open (or mostly open) Access






# Licensing

## Elsevier Experiments With Allowing 'Text Mining' of Its Journals



*By Jennifer Howard*

High-profile scholarly boycotts aren't the only way to get a big publisher's attention. Sometimes all it takes is a tweet.

Not long ago, Heather A. Piwowar, a postdoctoral researcher at the University of British Columbia, found herself on the phone with six high-level employees of the science-publishing giant Elsevier. Ms. Piwowar studies patterns in the sharing and reuse of research data. (Her Twitter handle is @researchremix.) Her work depends on text mining, using computers to automatically pull certain kinds of information from large amounts of text, including databases of journal articles. Many of those are subscription-based, and can be hard to get access to.

http://chronicle.com/article/Hot-Type-Elsevier-Experiments/131789/


"We found some … text mining in fields such as biomedical sciences and chemistry and some early adoption within the social sciences and humanities… however… most text mining in UKFHE is based on Open Access documents or bespoke arrangements." – key findings (p.2)

JISC

http://www.jisc.ac.uk/publications/reports/2012/value-and-benefits-of-text-mining.aspx

# Licensing



Photo from editorsweblog.org

# ECCO Project

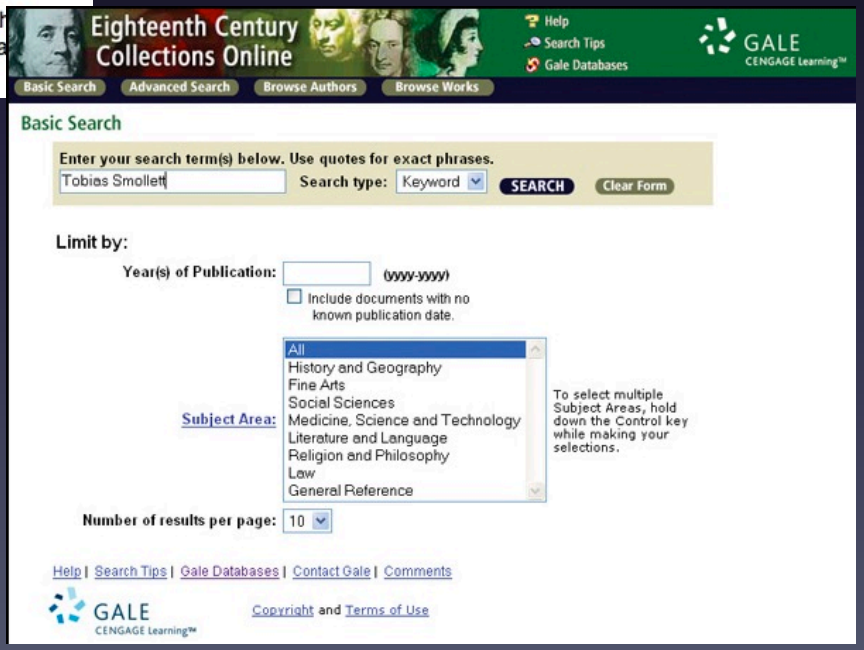
## TCP and Gale Cengage release 2,200 ECCO Texts to the public

*By mandellc on April 25, 2011*

Gale Cengage and the ECCO Text Creation Partnership have agreed to release 2,231 eighteenth-century texts to anyone who wishes to have them. You can search through those 18thCentury texts here at 18thConnect.org by word or phrase: go to the search page, select "ECCO" under "Other Digital Collections" as one of your search facets (by clicking on it), and then scroll down to select "Full-text" only as a search facet as well. Then enter any text (words, phrase) into the search blank, making it a facet as well.

Though we have no formal way of delivering documents, we are happy to be the source for plain text files: simply send Laura Mandell an email (lauraDOTmandellATgmailDOTcom) to request the texts; we can send you all of them, or selected texts.

As always, we are so grateful to the University of Michigan's Text Creation Partnership for all th
work they are doing to insure that we will send into the future the highest-quality digital surroga
our eighteenth-century heritage. And thanks to Gale for its openness to scholarly needs.



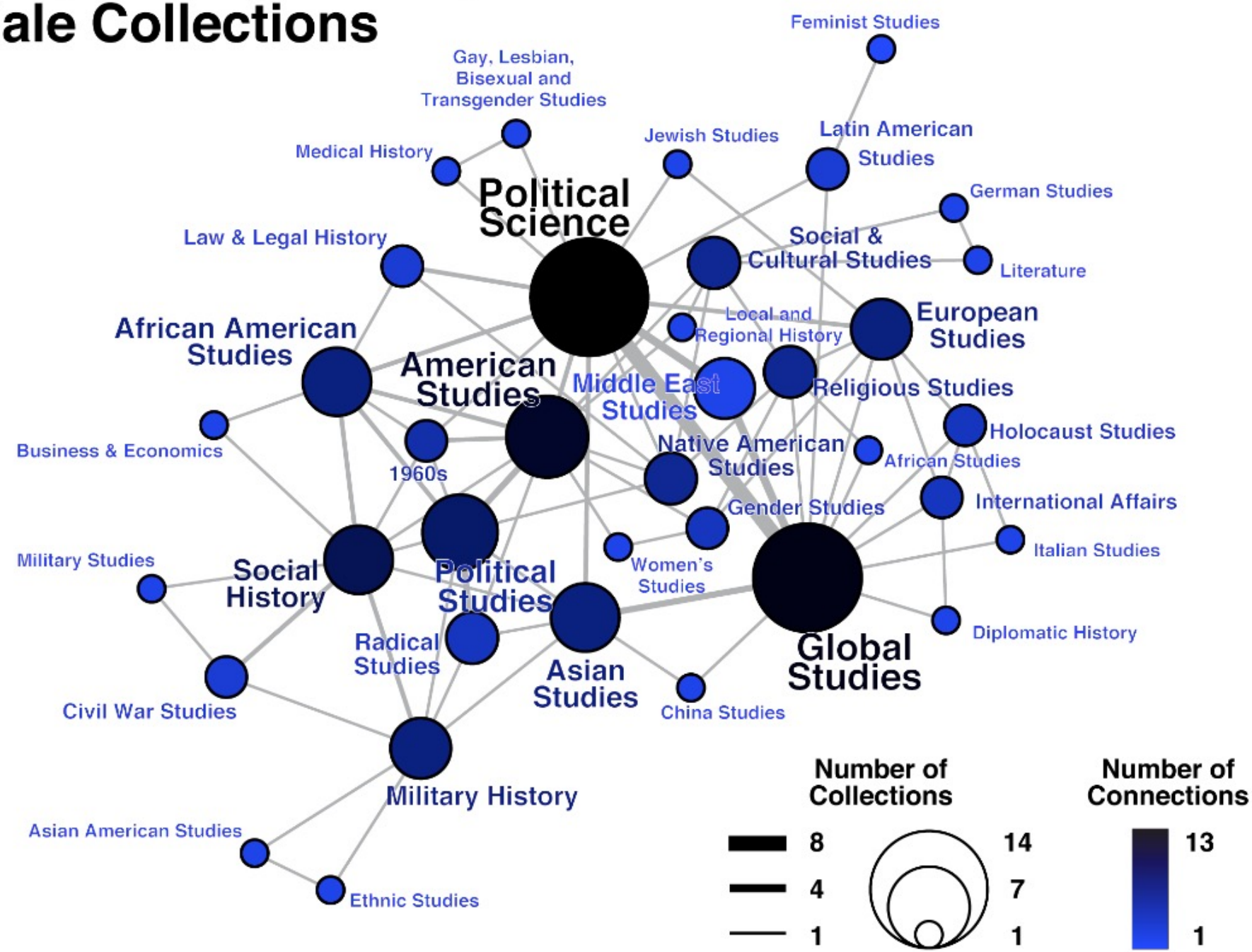
# “Big Data”

~63 Drives

~63 terabytes

>40 Topics

# Subjects Covered by Gale Collections



Thank you for your purchase of a Gale product. The backfiles you purchased are enclosed here on SDLT tape or on Disc. Also, for applicable products, a complete file list is included on Disc.

We recommend that you store this media in the same manner in which you would store any valuable media you wish to preserve, observing recommended temperature, lighting and humidity levels.

Your library should now have online access to the resources represented by these backfiles. If not, please contact your Gale Sales Representative at 1-800-877-4253.

Sincerely,
Gale, a part of Cengage Learning

*Visit* ***gale.cengage.com*** *for more information or to access our full product catalog.*


# Gale Backup Drive Collection

# Infrastructure

# Six Methods of Text Analysis

- Reading
- Counting Words
- Human Coding (researchers coding events/texts)
- Dictionary Methods (sentiment analysis)
- Supervised machine learning (using corpora)
- Unsupervised Machine Learning (topic modeling)

http://aeshin.org/textmining/

http://dx.doi.org/10.1111/j.1540-5907.2009.00427.x

# Infrastructure Issues

- Storage/ scratch space
- Processing power
- Tools for analytics

# Our Workstations

- 16 gigs of memory
- 1 TB of storage
- 64 bit computing
- Intel Xeon 3.5 GHz, 4 core
- Scanner available
- Fast networking

# Swappable Drives?

# General Software

NVIVO10

R

python™

# Specialized Software

MALLET

machine learning for language toolkit

topic-modeling-tool

A graphical user interface tool for topic modeling

# Service Model

# Services - Staffing

# Expert on Visualization

**Angela Zoss**

Data Visualization Coordinator
**Phone:** 919-684-8186
**Location:** Room 226-A Perkins (Second Floor Perkins 226)
**Email:** angela.zoss@duke.edu
**Twitter:** duke_vis
**Expertise:** data processing, data visualization, text analysis, network analysis, web visualizations, Excel, Tableau, python

# Services - Staffing

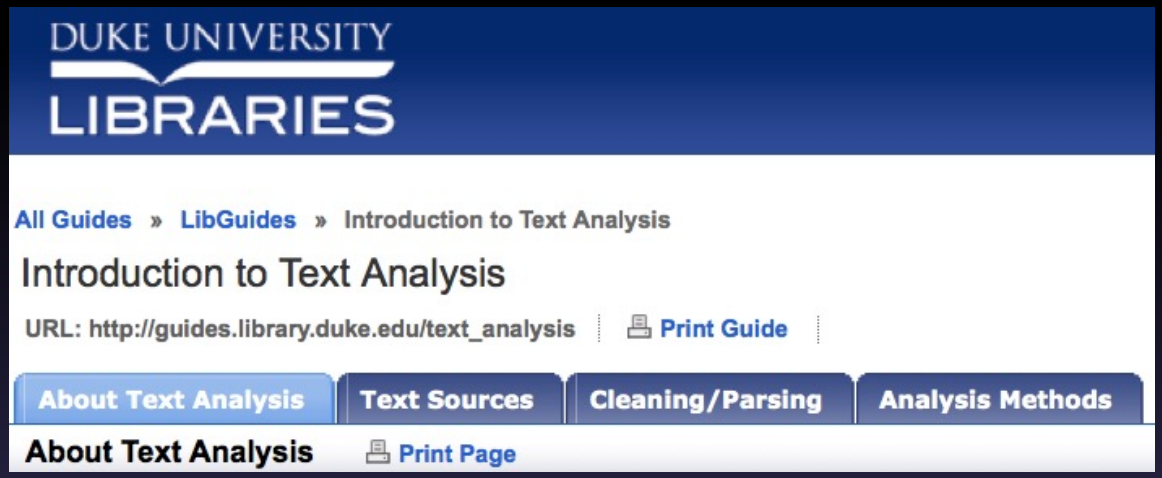
## Preparing texts

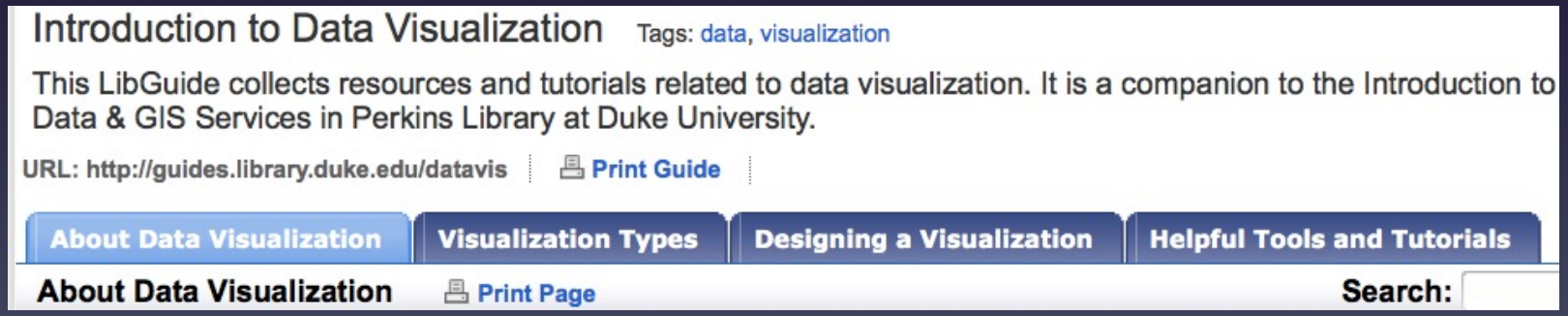
- The **bulk of your time** will be spent acquiring and preparing your texts
- Worth your time to **learn a scripting language** (such as Python)
- **Command-line text-processing tools** on Mac OS and Unix also very useful

@rybesh #duketext 17

http://aeshin.org/textmining/

# Services – Guides




http://library.duke.edu/data/guides/index.html

# Services – Workshops

## Events and Registration

| Details | Date | Time | Name | Location |
|---|---|---|---|---|
| Details | Fri, Dec 14, 2012 | 12:30 PM - 2:00 PM | Introduction to Text Analysis | Bostock Library Room 023 (Library Classroom) |

## HIGH-LEVEL TEXT ANALYSIS AND TECHNIQUES

Angela Zoss
Data Visualization Coordinator
226 Perkins Library
angela.zoss@duke.edu

text > DATA

DUKE UNIVERSITY LIBRARIES

Thursdays
2:00-3:30 PM
Perkins 217
*Open to everyone*

# In Summary

- Lots of research potential
- Licensing may be an issue for some
- Easy way to get started text mining with little investment but maybe some risk?

# Questions?

Joel Herndon – joel.herndon@duke.edu

Molly Tamarkin – tamarkin@duke.edu