



# Text Analysis



For Digital Humanists and  
Social Scientists



# Agenda

---

- Opening Discussion
- Brief introduction to text analysis in the social sciences and humanities
- What to expect from this workshop

# 0: Opening Discussion

---

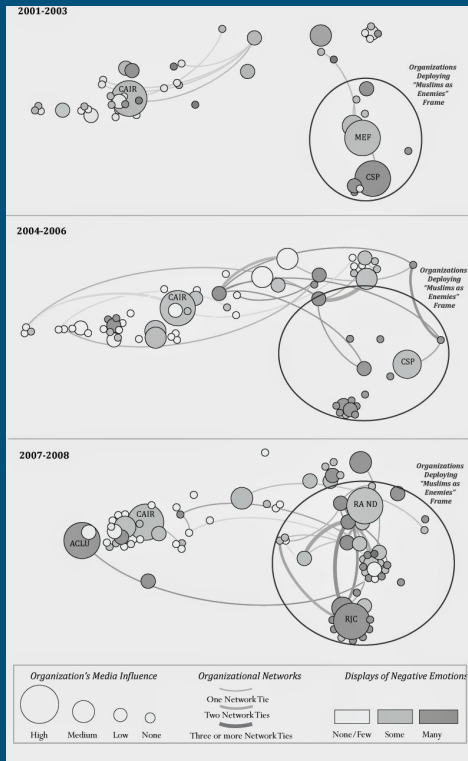
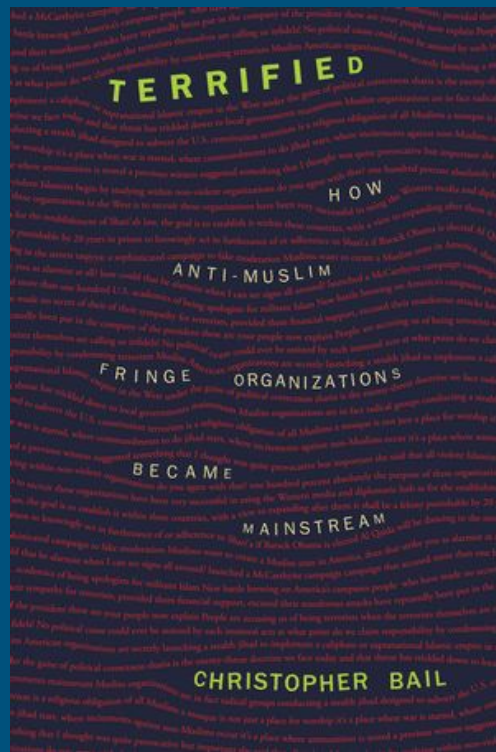
- What would it mean for a computer to successfully read a novel? How about social media post? What about one million social media posts?
- What output would you need for it to convince you that it had successfully read that novel? A social media post? One million social media posts?
- Group brainstorm: What types of social/humanist questions might computational text analysis help us answer?

# Part 1: Introduction to Text Analysis

---

# 1: Introduction to Text Analysis

## Chris Bail (Duke, Sociology)



Bail, Chris. 2016. *Terrified: How Anti-Muslim Fringe Organizations Became Mainstream*. Princeton: Princeton University Press.

# 1: Introduction to Text Analysis

## Justin Grimmer (Stanford, Political Science)

---

- Uncovered fourth political speech category: partisan taunting
  - Grimmer, Justin and Gary King. 2010. "General purpose computer-assisted clustering and conceptualization." *PNAS* 108(7): 2643-2650.
- Shows how and why "aligned" representatives dominate policy debates, and how this leads to artificial ideological extremity
  - Grimmer, Justin. 2013. "Appropriators not Position Takers: The Distorting Effects of Electoral Incentives on Congressional Representation." *American Journal of Political Science* 57(3): 624-642.

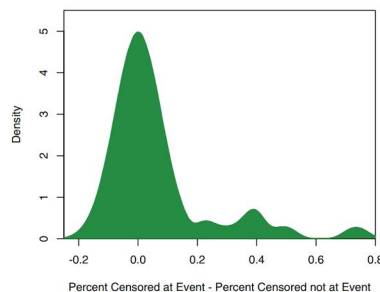
# 1: Introduction to Text Analysis

## King, Pan, and Roberts (Harvard, Political Science)

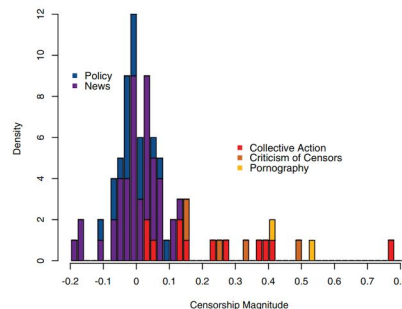
American Political Science Review

May 2013

**Figure 3. “Censorship Magnitude,” The Percent of Posts Censored Inside a Volume Burst Minus Outside Volume Bursts.**



(a) Distribution of Censorship Magnitude



(b) Censorship Magnitude by Event Type

King, Gary, Jennifer Pan, and Margaret E. Roberts. 2013. “How Censorship in China Allows Government Criticism but Silences Collective Expression.” *American Political Science Review* 107(2): 1-18.

# 1: Introduction to Text Analysis

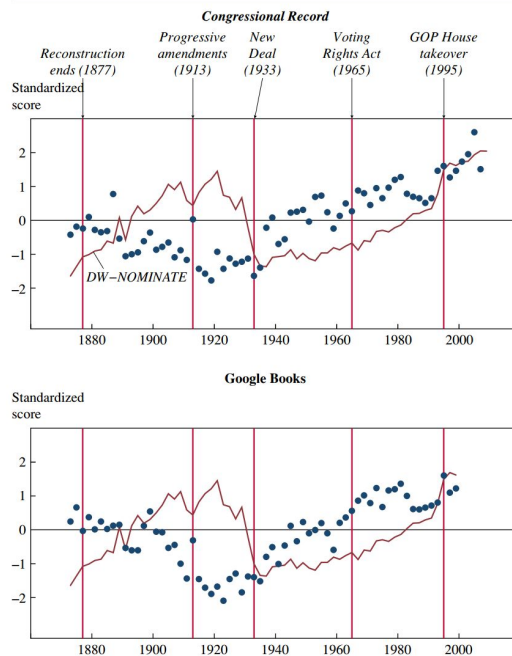
## Jensen, Kaplan, Naidu, and Wilse-Samson (Economics)

Jensen, Jacob, Ethan Kaplan, Suresh Naidu, and Laurence Wilse-Samson. 2012. "Political Polarization and the Dynamics of Political Language: Evidence from 130 Years of Partisan Speech." *Brookings Papers on Economic Activity*.

48

*Brookings Papers on Economic Activity*, Fall 2012

**Figure B.2.** Polarization Measured Using  $t$ -Statistic-Based Threshold and by DW-NOMINATE, 1873–2007<sup>a</sup>



Sources: Authors' calculations using data from the digitized *Congressional Record*, Google Books, and the legislator estimates on [voteview.com/dwnomin.htm](http://voteview.com/dwnomin.htm).

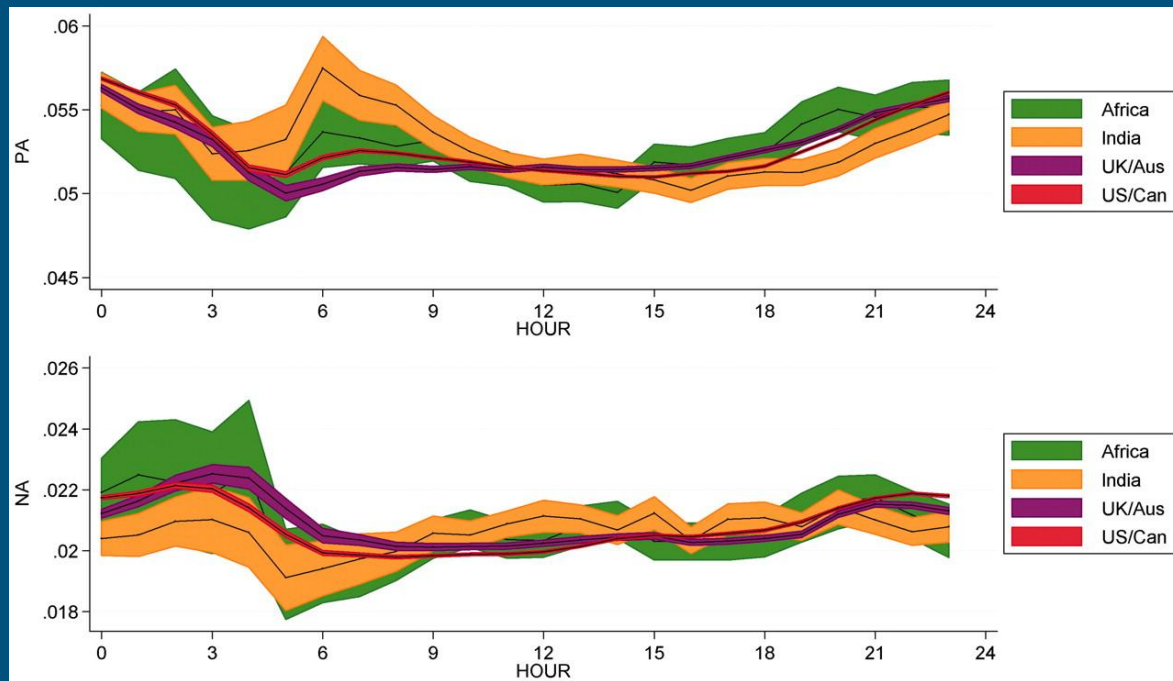
a. All measures are standardized to have a mean of zero and a variance of 1.



# 1: Introduction to Text Analysis

Scott A. Golder and Michael W. Macy (Cornell, Sociology)

Hourly changes in individual affect in four English-speaking regions. Each series shows mean affect (black lines) and 95% confidence interval (colored regions).



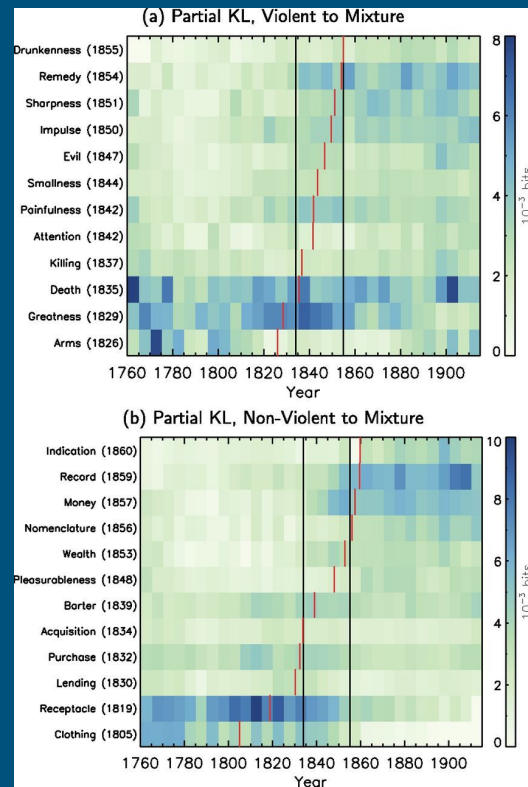
Golder, Scott A. and Michael Macy. 2011. "Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures." *Science* 333(6051): 1878-1881.

# 1: Introduction to Text Analysis

## Klingenstein, Hitchcock, and DeDeo (History)

Distinguishing features of violent and nonviolent trials in the OBC. Shown here is the partial KL,  $KL_i$ , for the 12 most distinctive synonym sets for violent indictments (A) and nonviolent indictments (B).

Klingenstein, Sara, Tim Hitchcock, and Simon DeCeo. 2014. "The civilizing process in London's Old Bailey." *PNAS* 111(26): 9419-9424



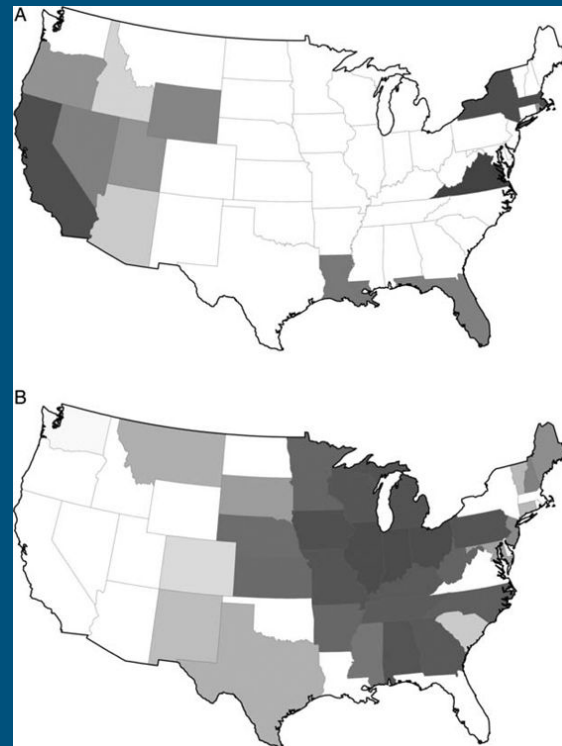
# 1: Introduction to Text Analysis

## Matthew Wilkens (English)

---

Fig. 7: Dunning log-likelihood values for named-location counts in the full corpus measured against mean state populations, 1850–80. (a) States overrepresented relative to their populations; (b) underrepresented states. Darker shades indicate larger absolute values, hence greater under- or overrepresentation.

Wilkens, Matthew. 2013. "The Geographic Imagination of Civil War-Era American Fiction." *American Literary History* 25(4): 803-840.



# 1: Introduction to Text Analysis

## Benjamin Schmidt (English, Northeastern)

### Gendered Language in Teacher Reviews

This interactive chart lets you explore the words used to describe male and female teachers in about 14 million reviews from RateMyProfessor.com.

You can enter any other word into the box below to see how it is split across gender and discipline: the x-axis gives how many times your term is used per million words of text.

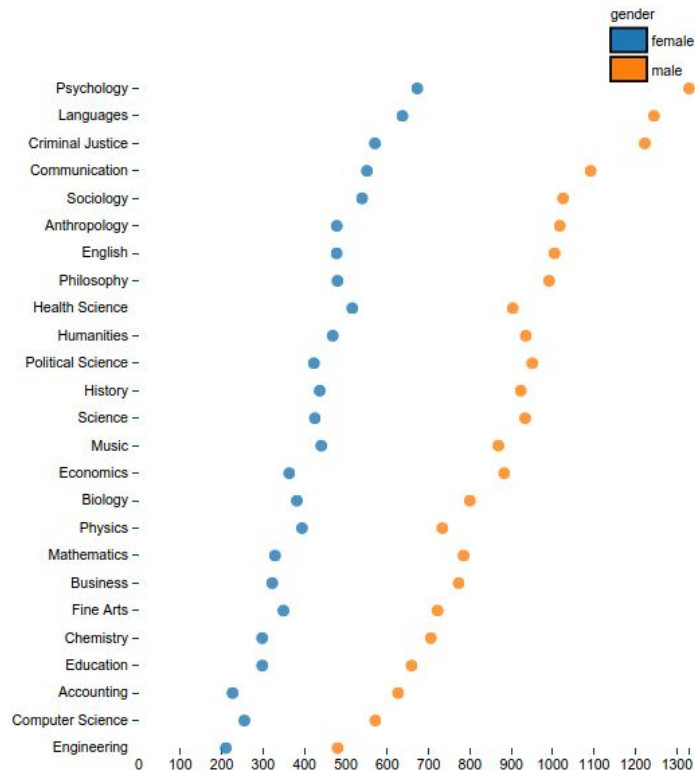
Not all words have gender splits, but a surprising number do. Even things like pronouns are used quite differently by gender.

**Search term(s) (case-sensitive):**  
**use commas to aggregate multiple terms**

All ratings

Only positive

Only negative

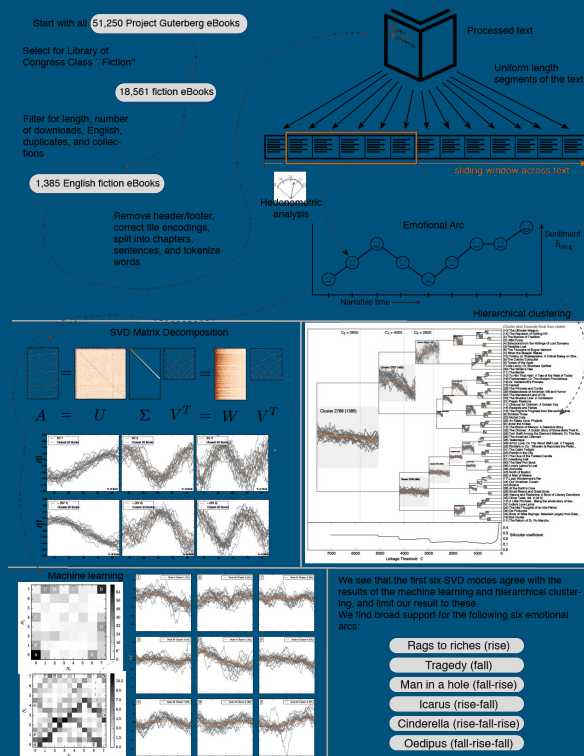


# 1: Introduction to Text Analysis

## Computational Story Lab (University of Vermont)

Reagan, Andrew J., Lewis Mitchell, Dilan Kiley, Christopher M. Danforth, and Peter Sheridan Dodds. 2016. "The emotional arcs of stories are dominated by six basic shapes." *EPJ Data Science* 5(31).

(Click [here](#) for enlarged graphic)



## Part 2: What to expect from this workshop

---

# 2: What to expect from this course

## Overview

---

- Scholars are increasingly using computational methods to analyze these new sources of text in order to ask, and answer, a diverse array of questions about the social world.
- In this workshop you will learn cutting-edge methods to analyze large amounts of texts, at scale, to explore questions fundamental to the humanities and social sciences.
  - Computers won't read for us. We will harness the power of computers to augment our own interpretive and critical thinking abilities.

# 2: What to expect from this course

## What you will learn

---

- By the end of the workshop you will understand:
  - A number of text analysis techniques available
  - what kind of evidence different techniques produce
  - how this evidence can be used to better understand the social world
- Practical skills:
  - How to ask and answer novel questions about the social world using text analysis techniques
  - Practical skills in Python to carry out a text analysis project
  - How to continue to build on the skills learned in the workshop