



# QTM 340: APPROACHES TO DATA SCIENCE WITH TEXT

Sandeep Soni

08/29/2024

# EXERCISE

# EXERCISE

- Answer the question using text:

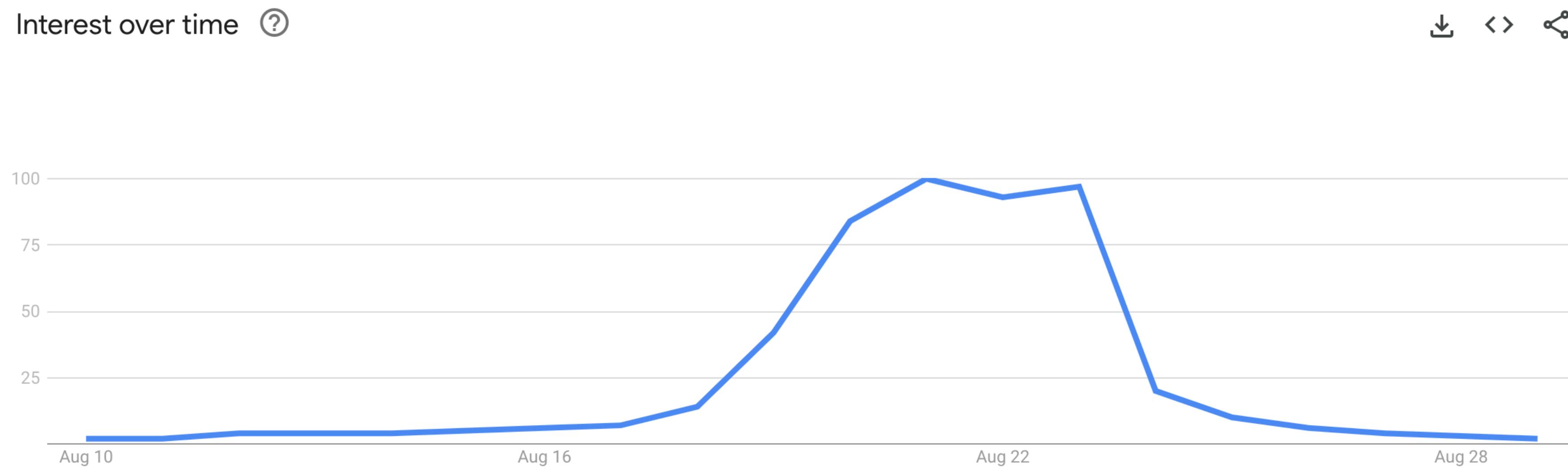
# EXERCISE

- Answer the question using text:
  - When was the most recent DNC held?

# EXERCISE

- Answer the question using text:
  - When was the most recent DNC held?
  - You're free to choose your source of text data. Give an outline of the technique that you'll use

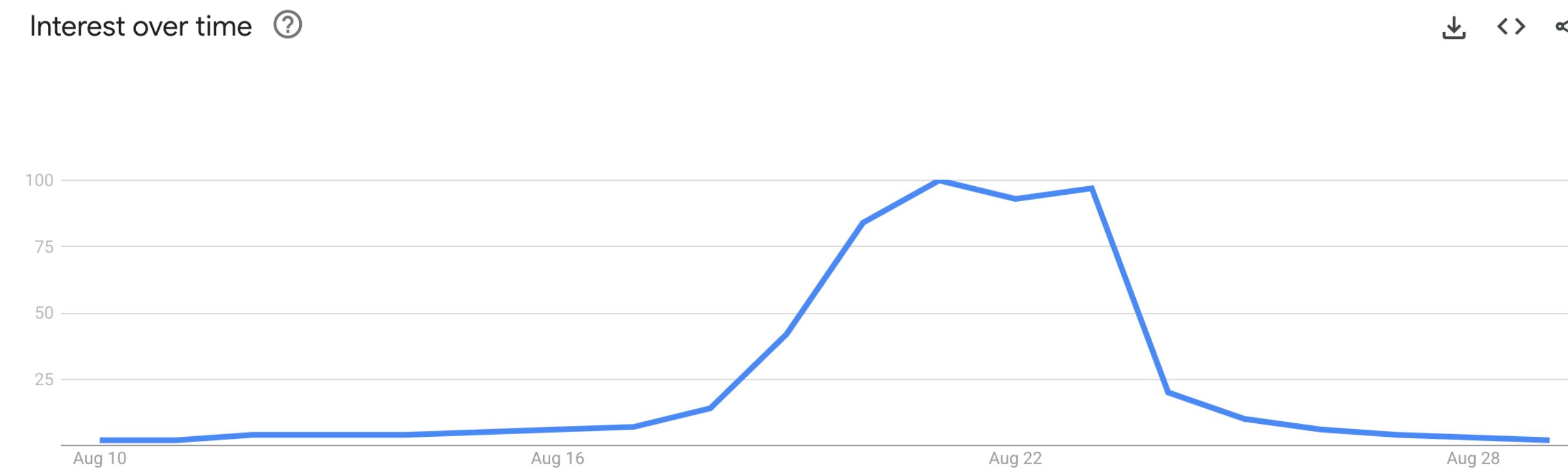
# When was the most recent DNC held?



Data from Google Trends for the search term “dnc”

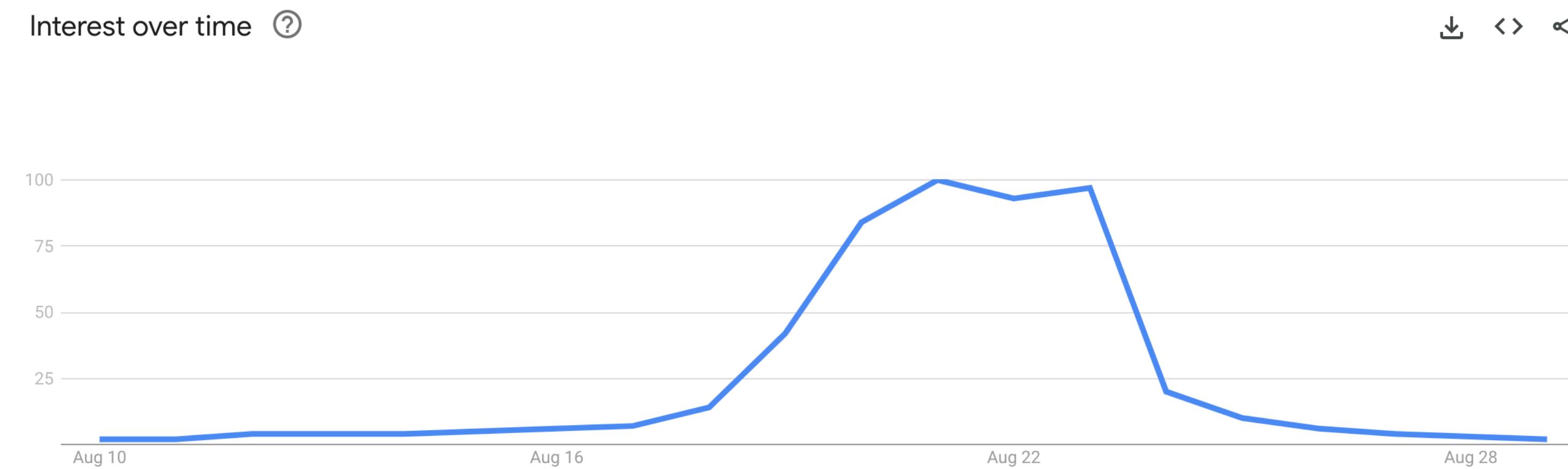
Q. When was the most recent DNC held?

A. Maybe Aug 20-23?



Q. When was the most recent DNC held?

A. Maybe Aug 20-23?



The actual answer  
is Aug 19-22

# LET'S REFLECT!

## LET'S REFLECT!

- What are the strengths of this method?

# LET'S REFLECT!

# LET'S REFLECT!

- What are the limitations?

# LETTERS

---

## Detecting influenza epidemics using search engine query data

Jeremy Ginsberg<sup>1</sup>, Matthew H. Mohebbi<sup>1</sup>, Rajan S. Patel<sup>1</sup>, Lynnette Brammer<sup>2</sup>, Mark S. Smolinski<sup>1</sup> & Larry Brilliant<sup>1</sup>

Seasonal influenza epidemics are a major public health concern, causing tens of millions of respiratory illnesses and 250,000 to 500,000 deaths worldwide each year<sup>1</sup>. In addition to seasonal influenza, a new strain of influenza virus against which no previous immunity exists and that demonstrates human-to-human transmission could result in a pandemic with millions of fatalities<sup>2</sup>. Early detection of disease activity, when followed by a rapid response, can reduce the impact of both seasonal and pandemic influenza<sup>3,4</sup>. One way to improve early detection is to monitor health-seeking behaviour in the form of queries to online search engines, which are submitted by millions of users around the world each day. Here we present a method of analysing large numbers of Google search queries to track influenza-like illness in a population. Because the relative frequency of certain queries is highly correlated with the percentage of physician visits in which a

By aggregating historical logs of online web search queries submitted between 2003 and 2008, we computed a time series of weekly counts for 50 million of the most common search queries in the United States. Separate aggregate weekly counts were kept for every query in each state. No information about the identity of any user was retained. Each time series was normalized by dividing the count for each query in a particular week by the total number of online search queries submitted in that location during the week, resulting in a query fraction (Supplementary Fig. 1).

We sought to develop a simple model that estimates the probability that a random physician visit in a particular region is related to an ILI; this is equivalent to the percentage of ILI-related physician visits. A single explanatory variable was used: the probability that a random search query submitted from the same region is ILI-related, as determined by an automated method described below. We fit a linear

The essential idea was that when people are sick with the flu, many search for flu-related information on Google, providing almost instant signals of overall flu prevalence. The paper demonstrated that search data, if properly tuned to the flu tracking information from the Centers for Disease Control and Prevention, could produce accurate estimates of flu prevalence two weeks earlier than the CDC's data—turning the digital refuse of people's searches into potentially life-saving insights.

—David Lazer

<https://www.wired.com/2015/10/can-learn-epic-failure-google-flu-trends/>

## **Google Flu Trends Still Appears Sick: An Evaluation of the 2013-2014 Flu Season**

David Lazer,<sup>1,2\*</sup> Ryan Kennedy,<sup>1,2,3</sup> Gary King,<sup>2</sup> Alessandro Vespignani<sup>1</sup>

Google's algorithm was quite vulnerable to overfitting to seasonal terms unrelated to the flu, like "high school basketball." With millions of search terms being fit to the CDC's data, there were bound to be searches that were strongly correlated by pure chance, and these terms were unlikely to be driven by actual flu cases or predictive of future trends. Google also did not take into account changes in search behavior over time. After the introduction of GFT, Google introduced its suggested search feature as well as a number of new health-based add-ons to help people more effectively find the information they need.

—David Lazer

<https://www.wired.com/2015/10/can-learn-epic-failure-google-flu-trends/>

# ITERATIVE (TEXT) DATA SCIENCE

# ITERATIVE (TEXT) DATA SCIENCE

- Ask an interesting question
- Collect the appropriate data
- Apply the methods
- Interpret the results

# ITERATIVE (TEXT) DATA SCIENCE

- Ask an interesting question
- Collect the appropriate data
- Apply the methods
- Interpret the results

We will focus on text!

# TEXT IS EVERWHERE

Web forums

Social media

Email

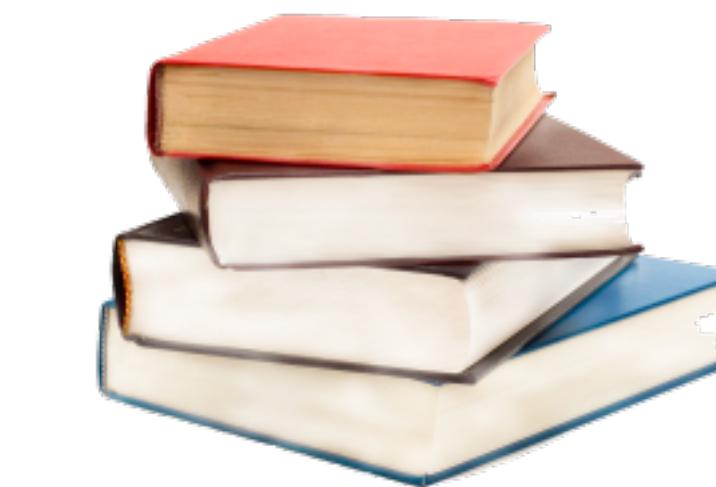
Books

Newspapers

Court opinions

Research papers

.. many more examples



(Sip Opinion)

OCTOBER TERM, 2007

Syllabus

NOTE: Where it is feasible, a syllabus (headnote) will be released, as is being done in connection with this case, at the time the opinion is issued. The syllabus constitutes no part of the opinion of the Court but has been prepared by the Reporter of Decisions for the convenience of the reader. See United States v. Detroit Timber & Lumber Co., 200 U. S. 321, 337.

SUPREME COURT OF THE UNITED STATES

Syllabus

DISTRICT OF COLUMBIA ET AL. v. HELLER

STANDING ARMS IN THE UNITED STATES EXHIBIT INC. ADDITIONAL INFO

# ABSTRACTIONS WITH TEXT

# ABSTRACTIONS WITH TEXT

- Text is “massively addressable at different levels of scale” (Whitmore)
  - Characters, lines, paragraphs, etc
  - Characters, words, phrases, sequences, trees, etc
  - Books, genre, collection, etc
  - Political speeches, bail hearings, quarterly results, etc

# ABSTRACTIONS WITH TEXT

- Text is “massively addressable at different levels of scale” (Whitmore)
  - Characters, lines, paragraphs, etc
  - Characters, words, phrases, sequences, trees, etc
  - Books, genre, collection, etc
  - Political speeches, bail hearings, quarterly results, etc

Text presents unique challenges about storing, cleaning, modeling, and interpreting the data

# QUALITATIVE VS QUANTITATIVE

- Historical text analysis mostly qualitative
- “close reading” Vs “distant reading”
- Qualitative analysis useful in initial exploration and theory building
- Quantitative analysis easy to scale and replicate

← → C chat.openai.com/chat

KC Explain quantum computing in simple terms

Reset Thread

Light Mode

OpenAI Discord

Updates & FAQ

Log out

Quantum computing is a type of computing that uses the principles of quantum mechanics to store and process information. In contrast to classical computers, which use bits to store and process information, quantum computers use quantum bits, or qubits. These qubits can be in a state of "0" and "1" at the same time, a property known as superposition. This allows quantum computers to perform certain operations much faster than classical computers. Additionally, quantum computers can use a phenomenon known as entanglement to process multiple qubits at the same time, further increasing their computational power. Overall, quantum computing has the potential to greatly expand the capabilities of computing, but it is still an emerging technology that is not yet widely available.

≡ Google Translate

Hindi English

◀ HINDI X

शाम को बारिश होने की संभावना है, कृपया अपने साथ छाता लेके चलें

Camera Handwriting Conversation Voice

◀ ENGLISH ☆

There is a possibility of rain in the evening, please take the umbrella with you

Offline

← → C chat.openai.com/chat

KC Explain quantum computing in simple terms

Reset Thread

Light Mode

OpenAI Discord

Updates & FAQ

Log out

Quantum computing is a type of computing that uses the principles of quantum mechanics to store and process information. In contrast to classical computers, which use bits to store and process information, quantum computers use quantum bits, or qubits. These qubits can be in a state of "0" and "1" at the same time, a property known as superposition. This allows quantum computers to perform certain operations much faster than classical computers. Additionally, quantum computers can use a phenomenon known as entanglement to process multiple qubits at the same time, further increasing their computational power. Overall, quantum computing has the potential to greatly expand the capabilities of computing, but it is still an emerging technology that is not yet widely available.

Like Dislike

≡ Google Translate

Hindi English

◀ HINDI X

शाम को बारिश होने की संभावना है, कृपया अपने साथ छाता लेके चलें

Camera Handwriting Conversation Voice

◀ ENGLISH ☆

There is a possibility of rain in the evening, please take the umbrella with you

Offline

Text is also used as data to build natural language processing systems but that's not our focus!

## OTHER RELATED CLASSES

- Computational linguistics (CS/QTM/Ling 329)

To learn computational methods for linguistic investigations

- Natural language processing (CS 571)

To learn computational methods for modeling natural languages

- Data Science for Beginners (SOC 190)

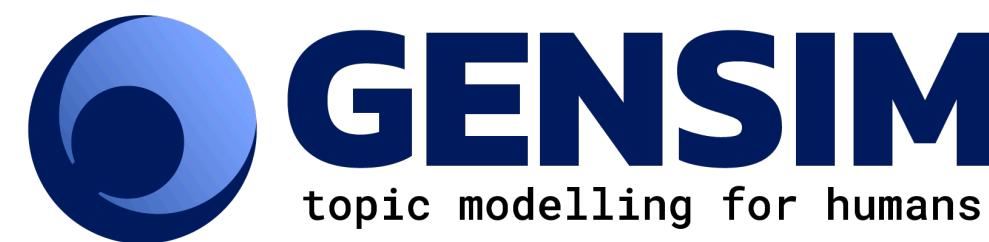
To learn general techniques to mine, model, and analyze data

# CONCURRENT QTM 340 SECTION

- Taught by Lauren Klein
- Focuses more on the human and ethical aspects of text as data
- This course is more statistical and methods focused
- We'll have some joint sessions

# TOOLS AND TECHNOLOGIES

---



spaCy

PyTorch



HUGGING FACE



What is the range of things you might learn in this class?

# SENTIMENT ANALYSIS

# Input:

# Tweets

Output:  
Average  
“happiness”

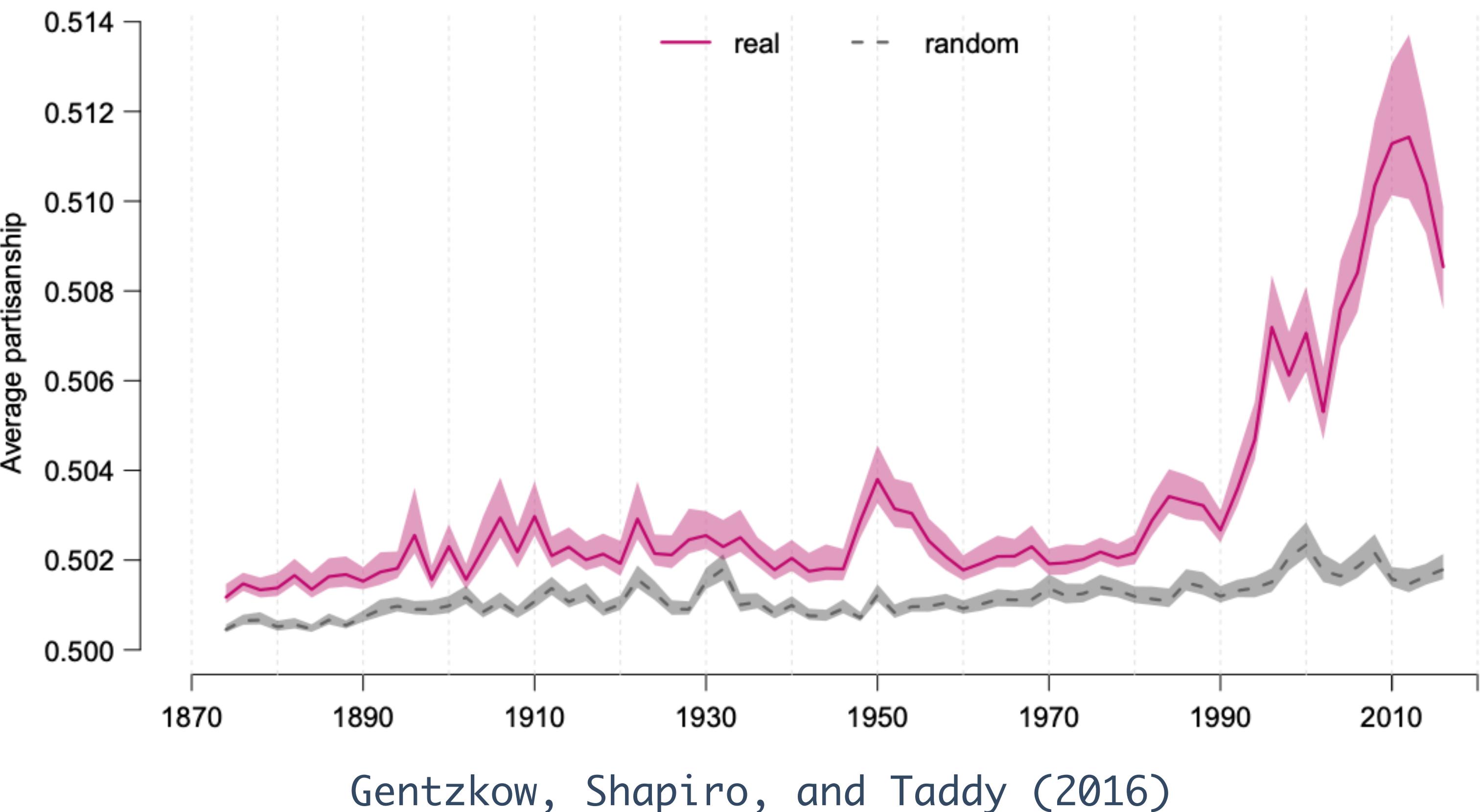


hedonometer.org  
(slide credit Brendan O'Connor and Laure Thompson)

# PARTISANSHIP

Input:  
Congressional  
speech

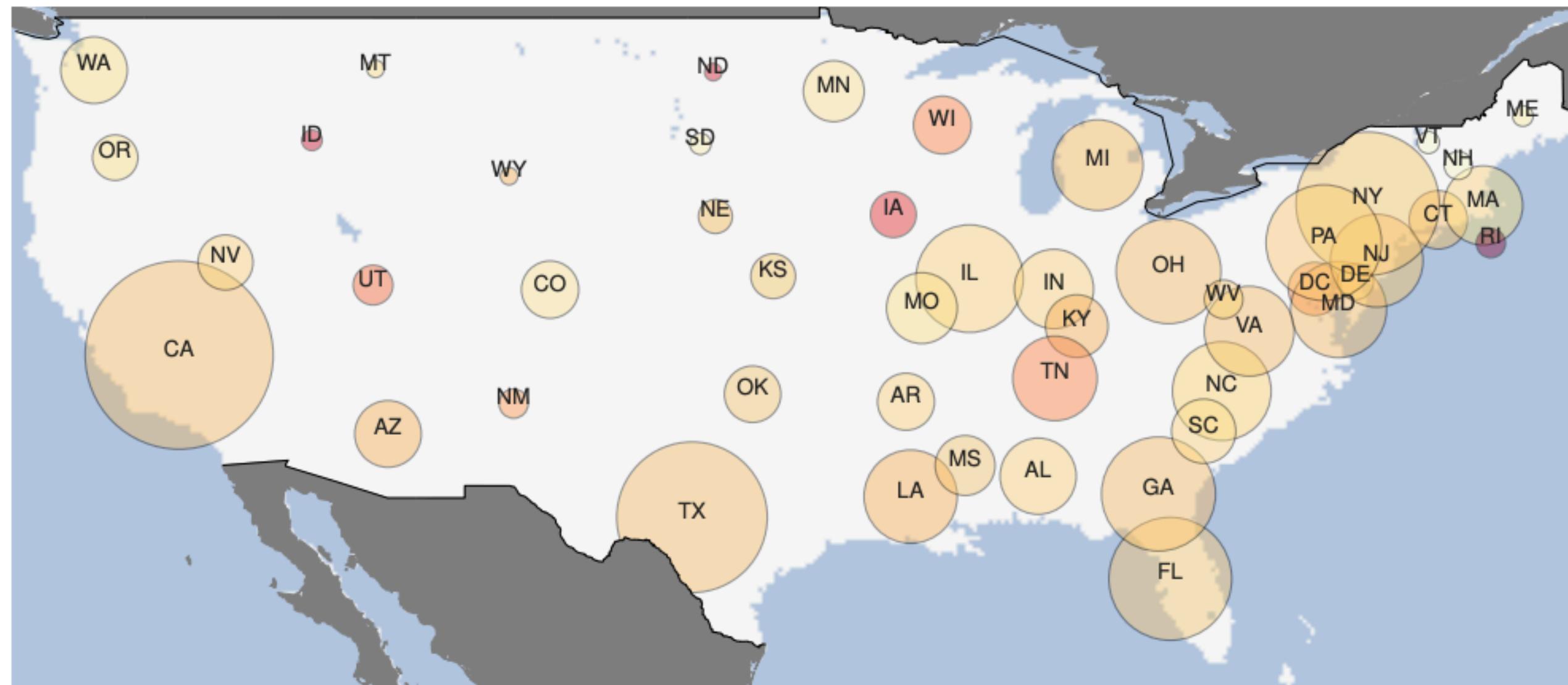
Output:  
Partisanship or  
party  
affiliation



Gentzkow, Shapiro, and Taddy (2016)

# GEOLOCATION

Input:  
Social  
media posts



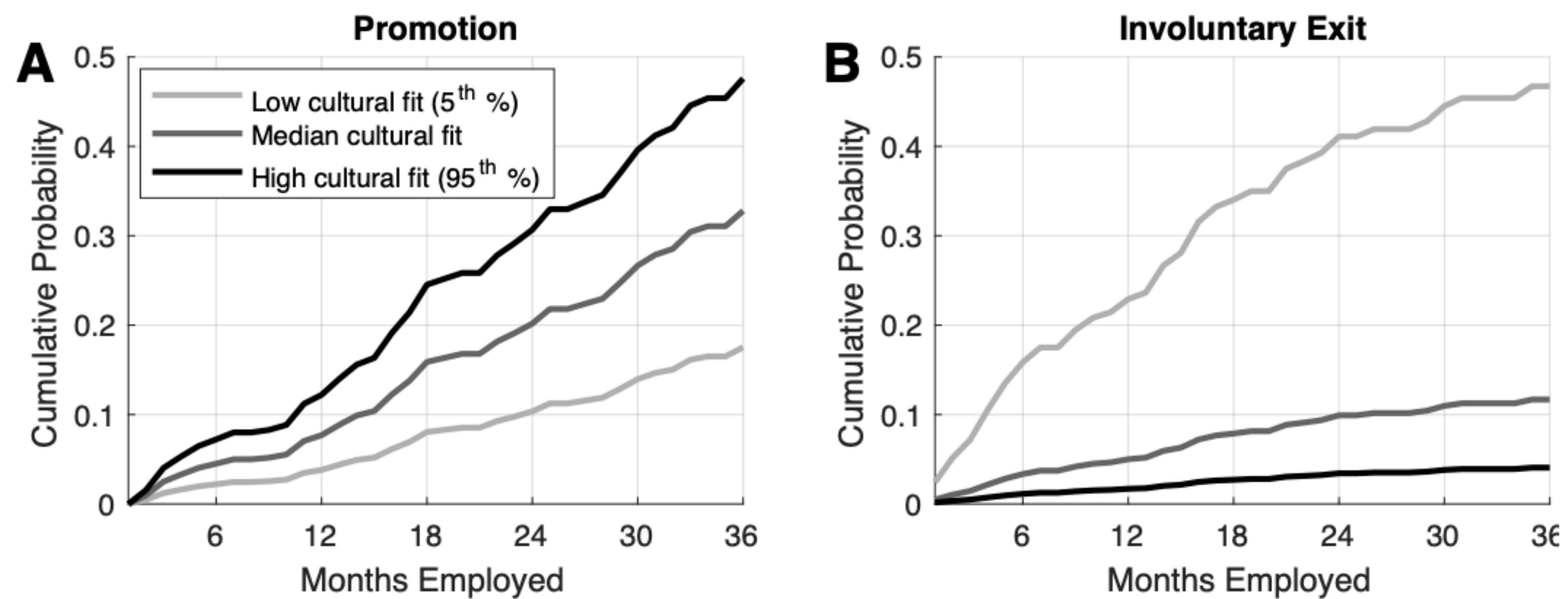
Output:  
Place name

Seattle, WA	Austin, TX	Jacksonville, FL	Columbus, OH	Charlotte, NC	Phoenix, AZ	New Orleans, LA	Baltimore, MD
#goseahawks	stubb	unf	laffayette	#asheville	clutterbuck	mcneese	bhop
smock	gsd	ribault	#weareohio	#depinga	waffles	keela	#dsu
traffuck	#meatsweats	wahoowa	#arcgis	batesburg	bahumbug	pentecostals	chestertown
ferran	lanterna	wjct	#slammin	stewey	iedereen	lutcher	aduh
promissory	pupper	fscj	#ouhc	#bojangles	rockharbor	grogan	umbc
chowdown	effaced	floridian	#cow	#occupyraleigh	redtail	suela	lmt
ckrib	#austin	#jacksonville	mommyhood	gville	gewoon	cajuns	assistly
#uwhuskies	lmfbo	#mer	beering	sweezy	jms	bmw	slurpies

# ENCULTURATION

Input: Employee emails

Output:  
Promotion, time  
to separation

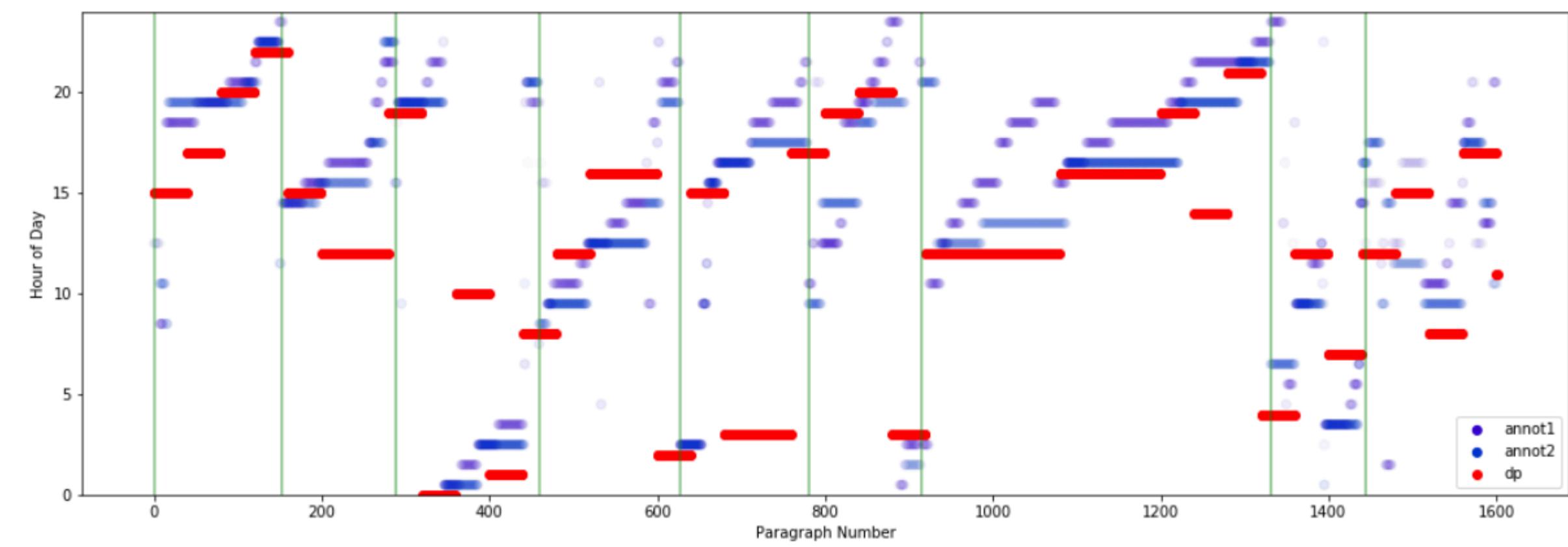
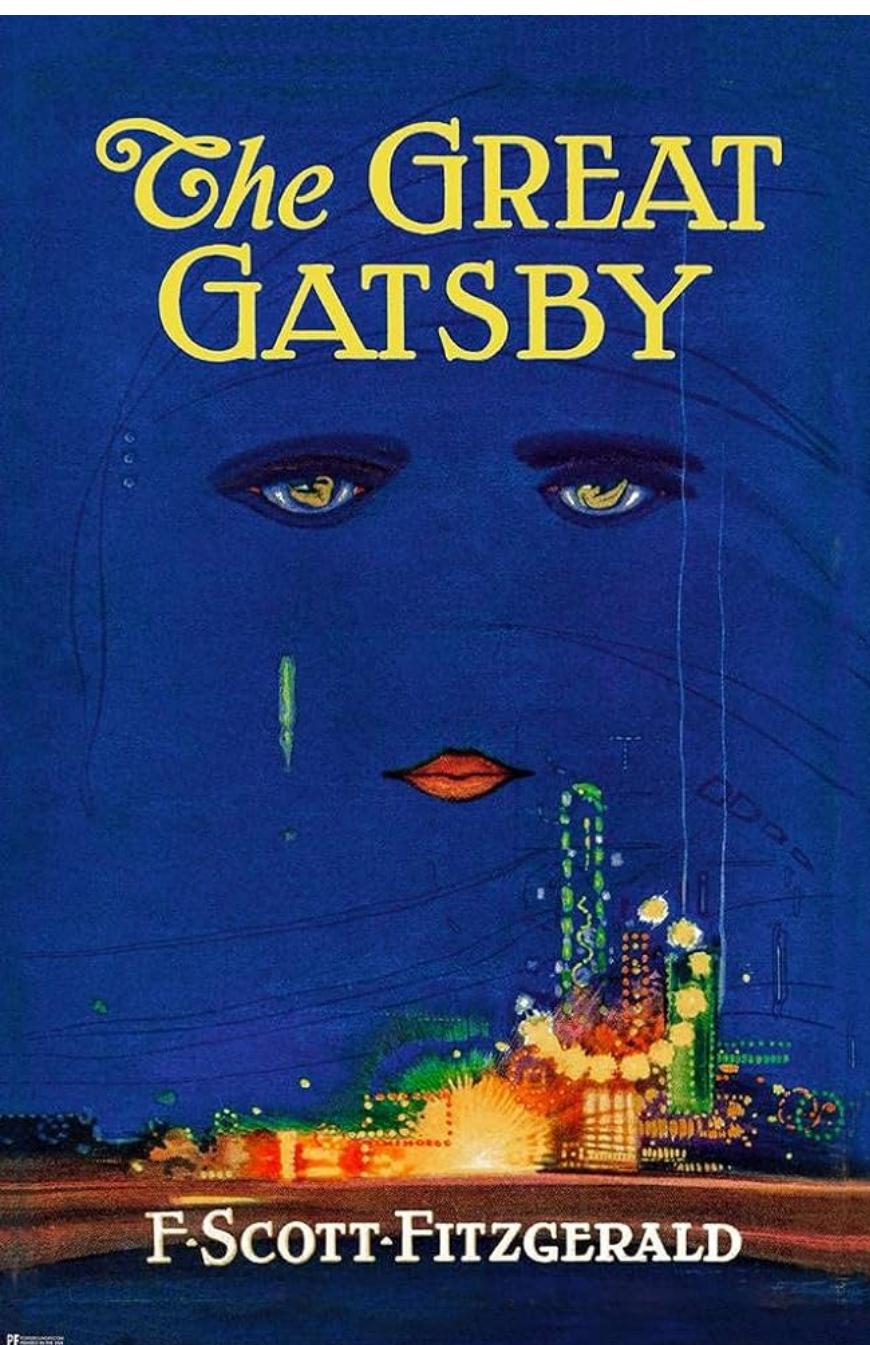


Srivastava et. al. (2018)  
(Slide credit David Bamman)

# TIME OF THE DAY

Input: Passages  
from a novel

Output: Time of  
the day in the  
story



Kim et. al. (2020)

# MEASUREMENT

How to build algorithmic instruments to measure a quantity of interest from text?

# WHY IS THIS DIFFICULT?

- Language is ambiguous
- Variation and change
- Dependence on context

# AMBIGUITY

---

- Language is inherently ambiguous
- This ambiguity is seen at various linguistic levels
- Text modeling and analysis aims at creating robust inference techniques that can handle such ambiguities



# VARIATION

---

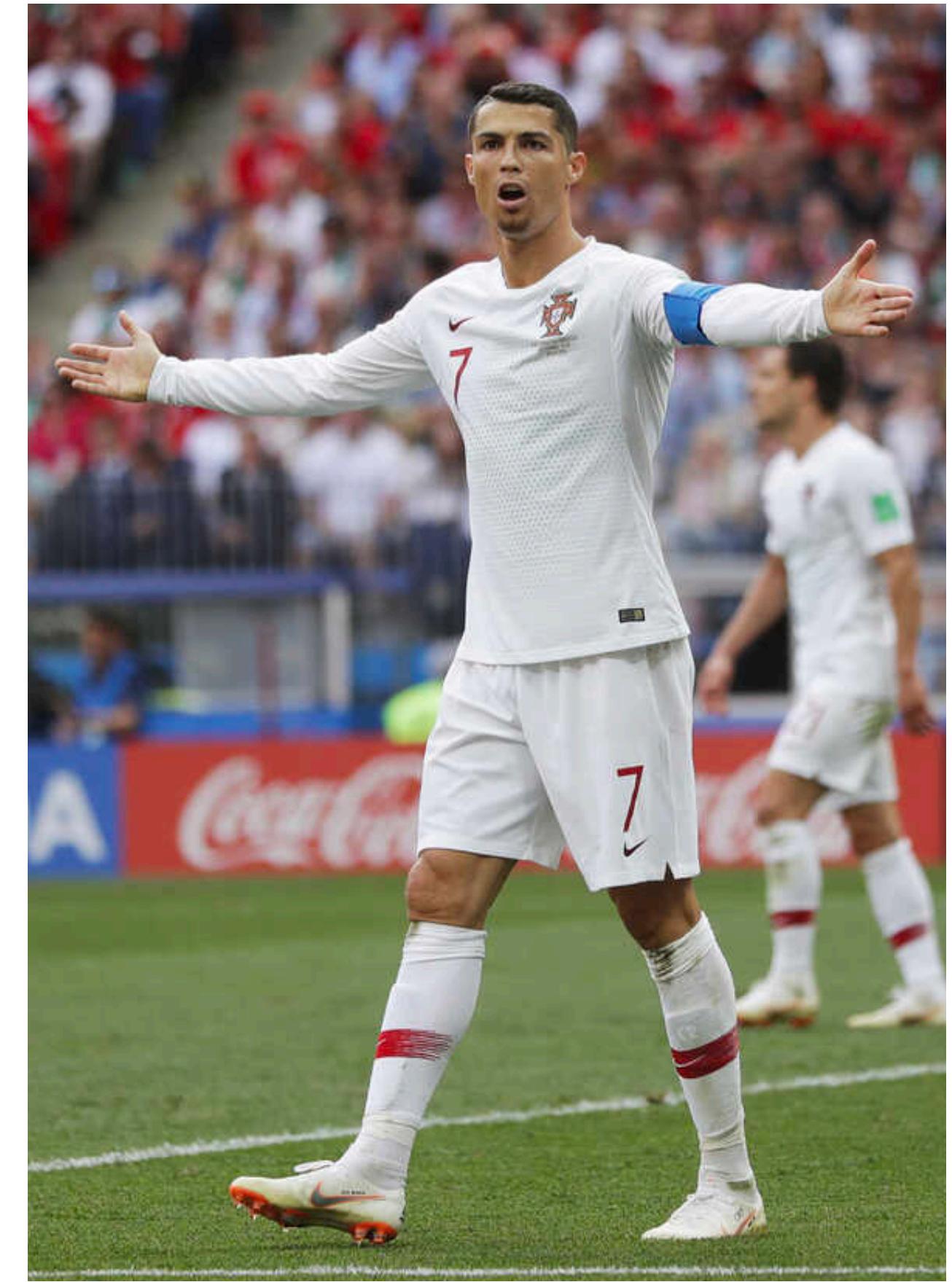
- Diversity in society, demographics, and culture are often encoded in language
- Analysis of text helps decode these socio-cultural layers



# CONTEXTUAL DEPENDENCE

---

- Cristiano Ronaldo plays for -----



Year=2015



Year=2023

# ADMINISTRIVIA

- Sandeep Soni ([sandeep.soni@emory.edu](mailto:sandeep.soni@emory.edu))
  - Please call me by my first name! No exceptions to this rule
- Office hours:
  - Wednesday 11am-12pm (PAIS 588)
  - Monday 11am-12pm (Zoom and only by appointment)
- Course website:
  - <https://sandeepsoni.github.io/classes/qlm340.html>
  - Syllabus and policies page: <https://tinyurl.com/4bnnvex3>



# CLASS FORMAT

- Typically, a class session will have two sections:
  - Lecture, to give an overview of the topic for the day
  - Lab exercise, to get hands-on experience
- Students are generally expected to finish the readings before each class and participate in the class discussion during the lecture part

# GRADING

Component	Grade %
Class participation	10%
Reading responses	18%
Problem sets	32%
Group project	40%

# PARTICIPATION: WHAT'S EXPECTED?

- Attend the class
- Read the required readings and engage in the classroom discussion
- 1% of total course grade is reserved for filling course eval at the end of the semester

# READING RESPONSE: WHAT'S EXPECTED?

- There will be 4 reading responses due that answer some questions from the required readings
- You're expected to at least 3 responses (If you submit 4, highest 3 graded responses will count towards final grade)
- The response will be for specific questions (mix of multiple choice and free-form) about a bunch of readings
- No collaboration and no late submissions allowed

# PROBLEM SETS: WHAT'S EXPECTED?

- Problem sets are designed to act as mini-projects
- For Problem set 1, lots of scaffolded code will be given along with instructions on how to execute the project and interpret the key findings
- For problem set 2 and 3, less scaffolded code and instructions will be given.

## SOME MORE POINTS

- You're expected to know and code in Python!
- You must have taken a linear algebra or probability class
- Knowledge about humanities, linguistics is a plus but not necessary
- Please talk to me if you have any concerns

# PROJECT: WHAT'S EXPECTED?

- Semester-long project (2-4 students) that involves an empirical investigation of a research question using text data
- Milestones include:
  - Project proposal
  - Midterm report
  - Final report
  - Class presentation
- More details will be shared soon!

# CANVAS

- All slides, data, and class notebooks (if any) will be on Canvas (under files)

## IN CLASS

- Introduce Google Colab
- Let's play! <https://research.google.com/semantris/>
- Explore this tool for fun: <https://voyant-tools.org/>