



QTM 340: APPROACHES TO DATA SCIENCE WITH TEXT

Sandeep Soni

08/23/2023

QUESTIONS ABOUT CHANDRAYAAN-3

QUESTIONS ABOUT CHANDRAYAAN-3

- Answer the questions about the mission using text:

QUESTIONS ABOUT CHANDRAYAAN-3

- Answer the questions about the mission using text:
 - When did the lander land?

QUESTIONS ABOUT CHANDRAYAAN-3

- Answer the questions about the mission using text:
 - When did the lander land?
 - From where did people watch it?

When did the lander land?

When did the lander land?

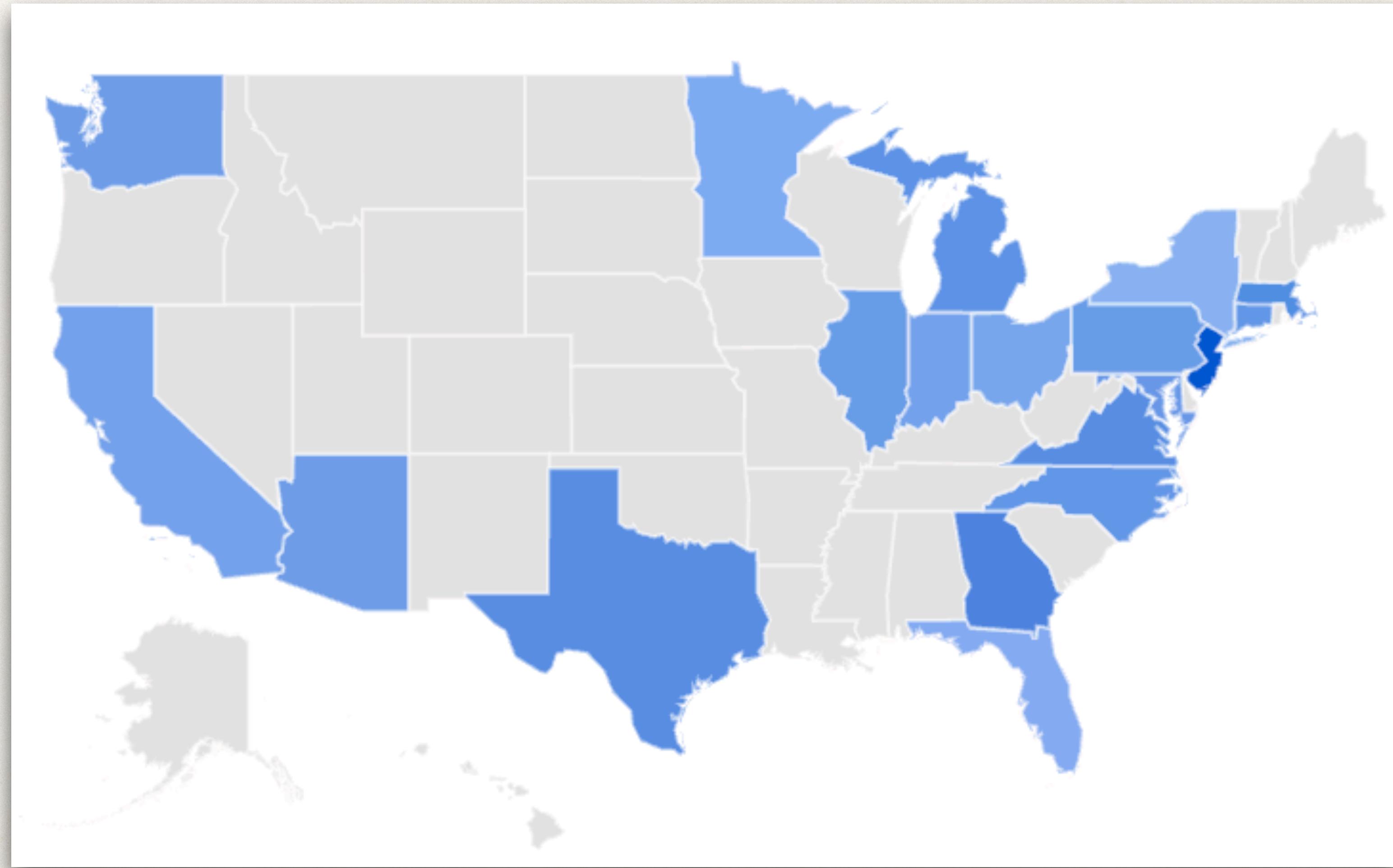


Timeseries of Google searches for the term “Chandrayaan-3 status” in approximately the last 24 hours

[https://trends.google.com/trends/explore?
date=now%201-d&q=Chandrayaan-3%20status&hl=en](https://trends.google.com/trends/explore?date=now%201-d&q=Chandrayaan-3%20status&hl=en)

From where did people watch it?

From where did people watch it?



Geographical map of Google searches for the term
“Chandrayaan-3 status” on 23rd August

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

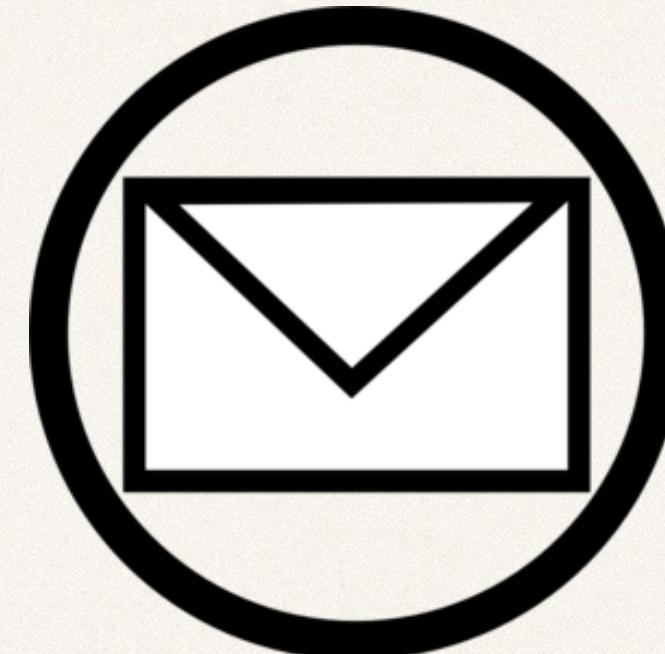
Books

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

PRINTING AREA IN THE UNITED STATES EXHIBIT FIVE ADDITIONAL COPY

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.

Like Dislike

≡ Google Translate

Hindi English

HINDI

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!

MOTIVATION

MOTIVATION

What can we learn about the world by applying natural language processing methods on textual data?

OTHER RELATED CLASSES

- Computational linguistics (CS/QTM/Ling 329)

To learn computational methods for linguistic investigations

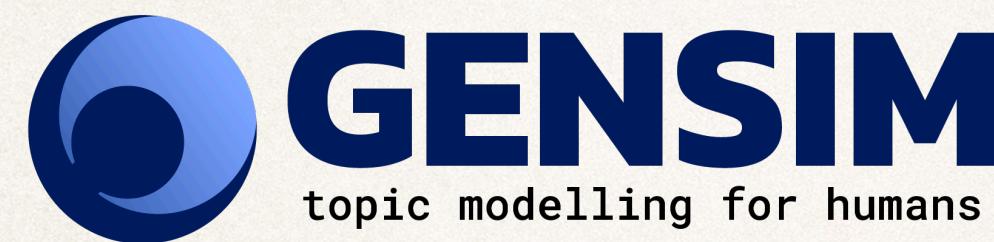
- Natural language processing (CS 571)

To learn computational methods for modeling natural languages

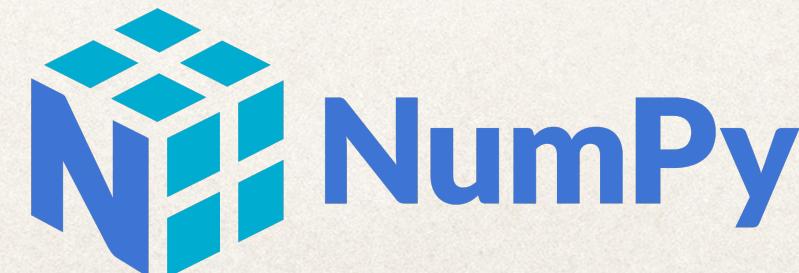
- Data Science for Beginners (SOC 190)

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

TOOLS AND TECHNOLOGIES



PyTorch



spaCy



HUGGING FACE

INTERJECTION

What do you want to take away from this class?

SENTIMENT ANALYSIS

Input:
Tweets

Output:
Average
“happiness”

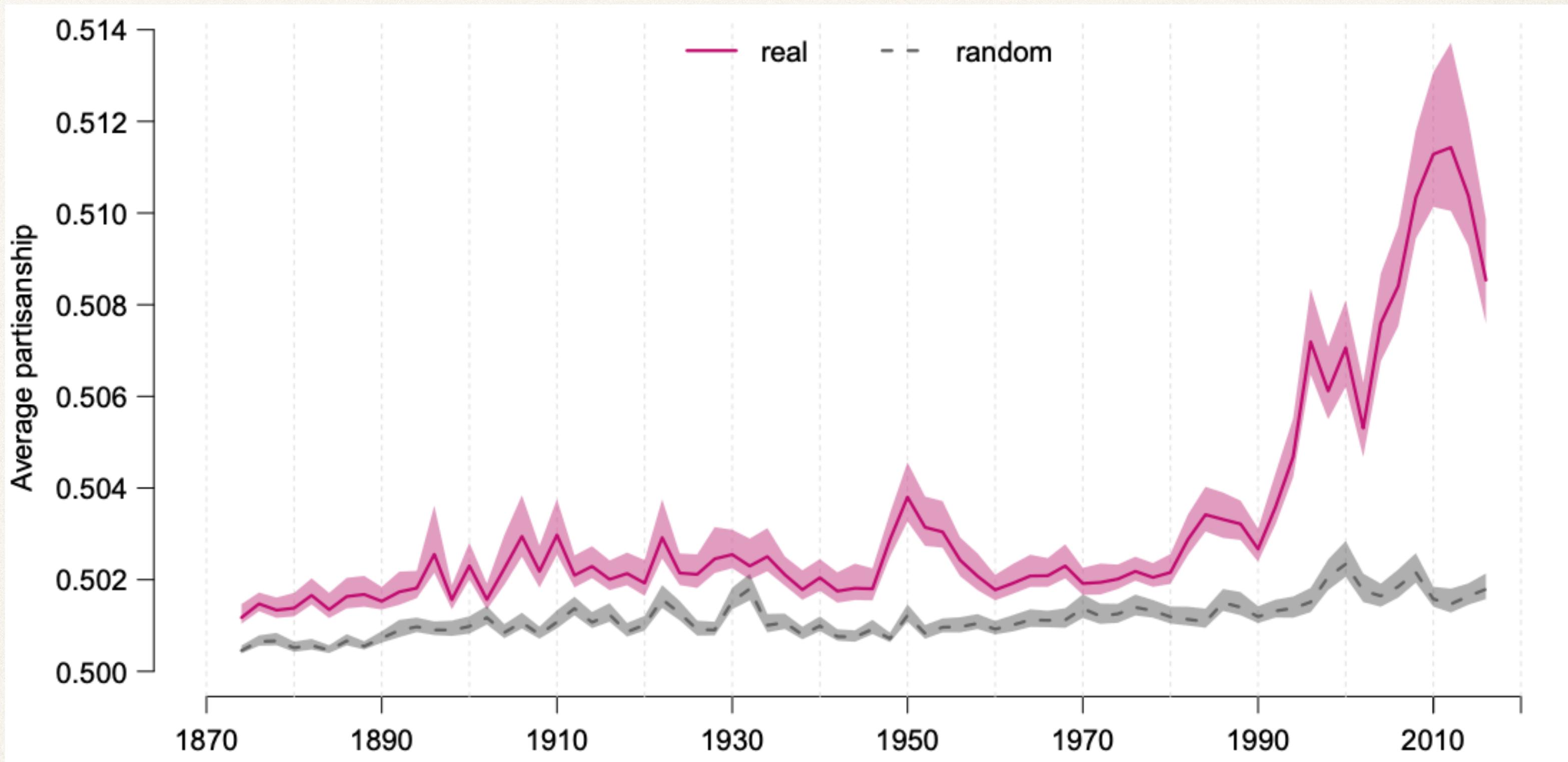


hedonometer.org
(slide credit Brendan O'Connor and Laure Thompson)

PARTISANSHIP

Input:
Congressional
speech

Output:
Partisanship or
party
affiliation

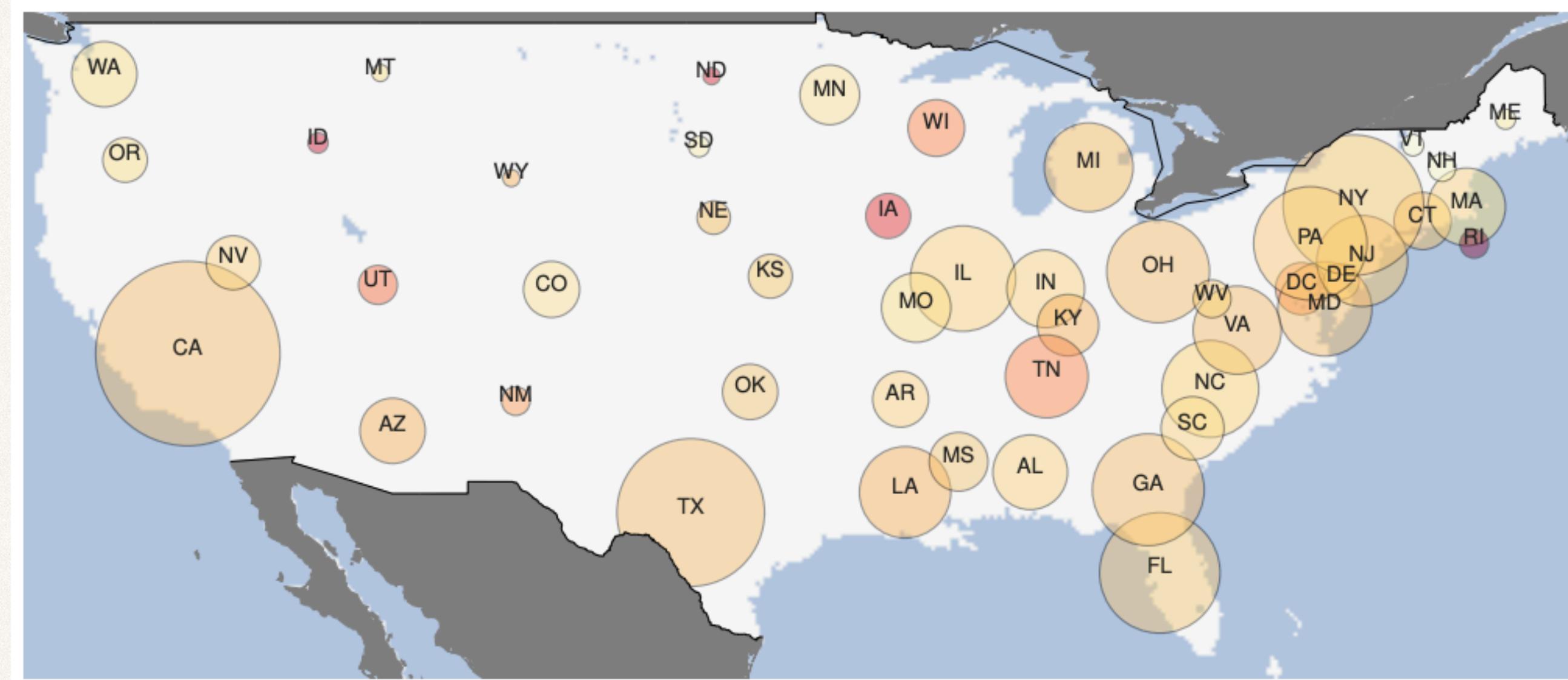


Gentzkow, Shapiro, and Taddy (2016)

GEOLOCATION

Input:
Social
media posts

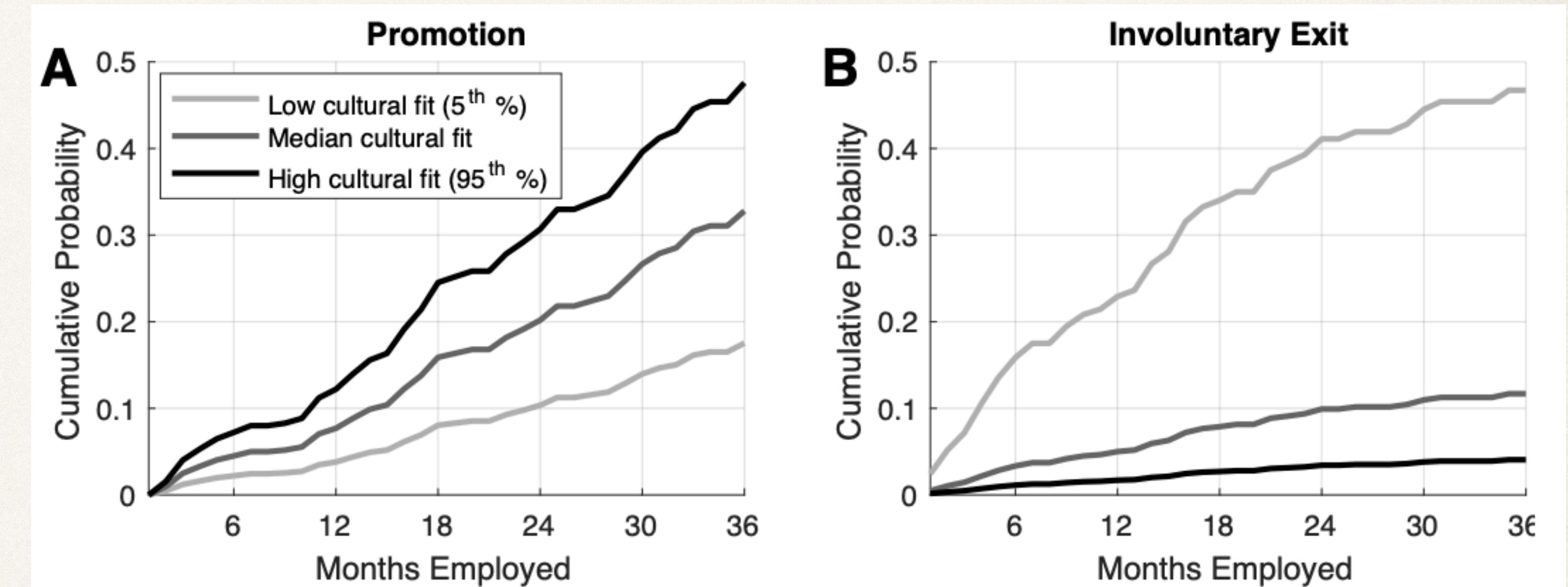
Output:
Place name



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)

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



ADMINISTRIVIA

- Sandeep Soni (sandeep.soni@emory.edu)
- Office hours:
 - Tuesday 11am-12pm (PAIS 588)
 - If you want to book an appointment, just email me!
- Course website:
 - <https://sandeepsoni.github.io/classes/qlm340.html>

CLASS FORMAT

- Every class session divided into two sections:
 - Lecture, to give an overview of the topic for the day
 - Lab exercise, to get hands-on experience
- Students are expected to finish the readings before each class and participate in the class discussion

GRADING

- Participation (10%)
- Homeworks (16%)
- Problem sets (24%)
- Project (50%)

PARTICIPATION: WHAT'S EXPECTED?

- Read the required readings and engage in the classroom discussion
- Homeworks are likely to contain questions that are related to the material that we discuss in class
- Answer questions on piazza/canvas

HOMEWORKS: WHAT'S EXPECTED?

- There will be a total of 10 short homeworks.
- You can drop 2 homeworks.
- Each homework should not take more than 1-2 hours.
- No collaboration and no late homework submissions allowed

PROBLEM SETS: WHAT'S EXPECTED?

- Problem sets are designed to help you gain practical insights
- Students will be given scaffolded code and will be expected to fill in the parts
- Expect to code up to 200-300 lines of code for each problem set.

PROJECT: WHAT'S EXPECTED?

- Semester-long project (1-2 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!

GITHUB

- <https://github.com/sandeepsoni/QTM340-Fall23>
- The repository will contain slides, data, and notebooks

IN CLASS

- Introduce Google Colab
- Explore this tool: <https://voyant-tools.org/>