QTM 340: Approaches to Data Science with Text

Sandeep Soni

01/18/2024

Question

Question

- Answer the question using text:

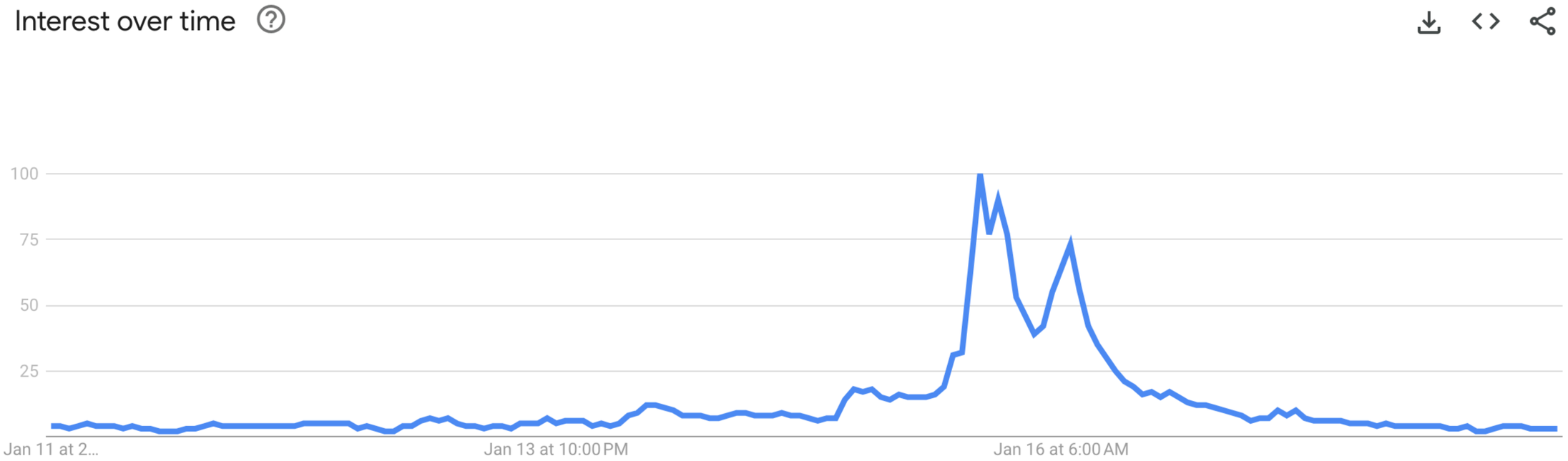
Question

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

Question

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



Data from Google Trends for the search term “Iowa primary”

Question

Question

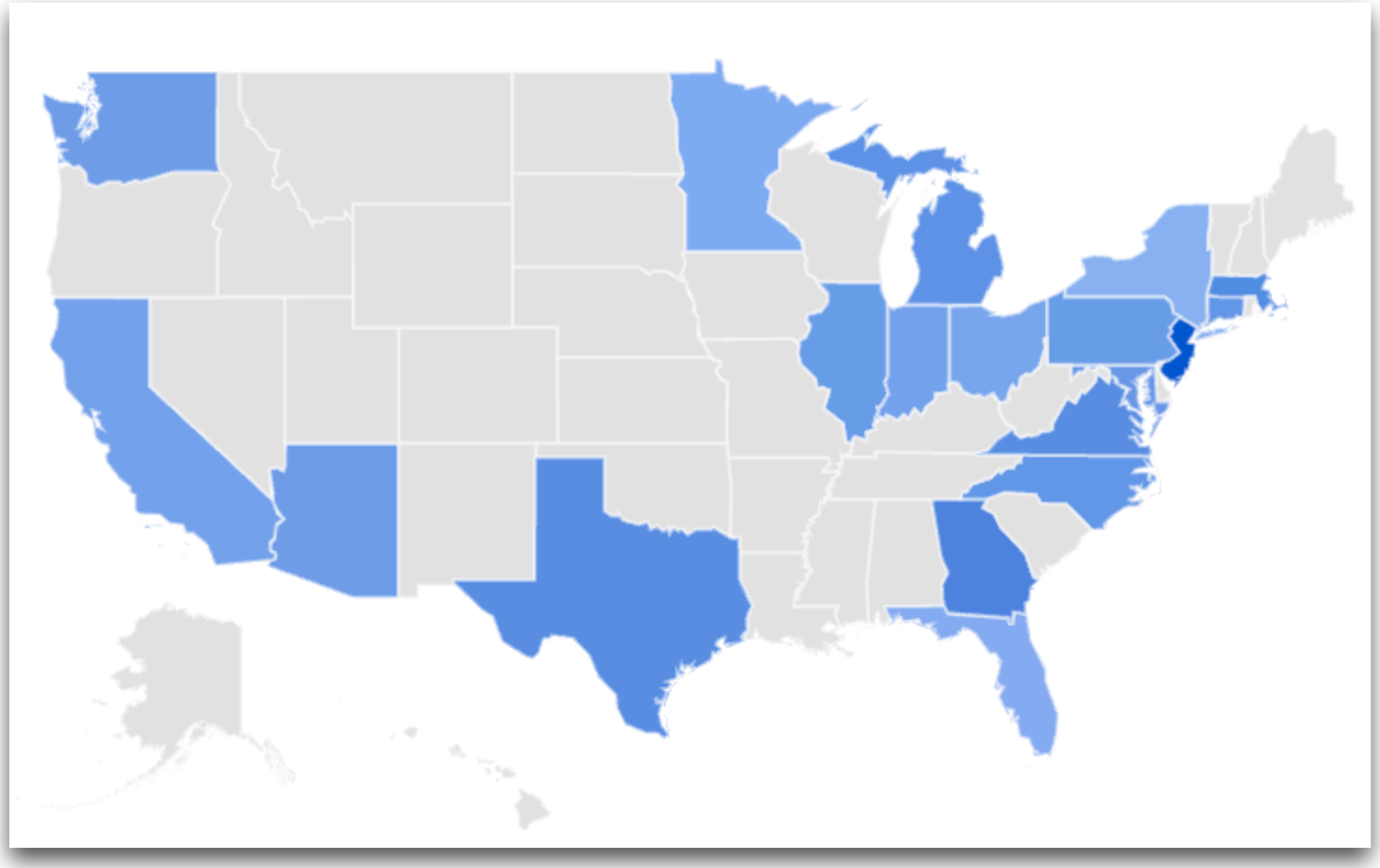
- Answer the question using text:

Question

- Answer the question using text:
  - In August 2023, India became the first country to soft land on the southern pole of the moon. From where did people in the US watch the landing?

From where did people watch the landing?

From where did people watch the landing?



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 Everywhere

Web forums

Social media

Email

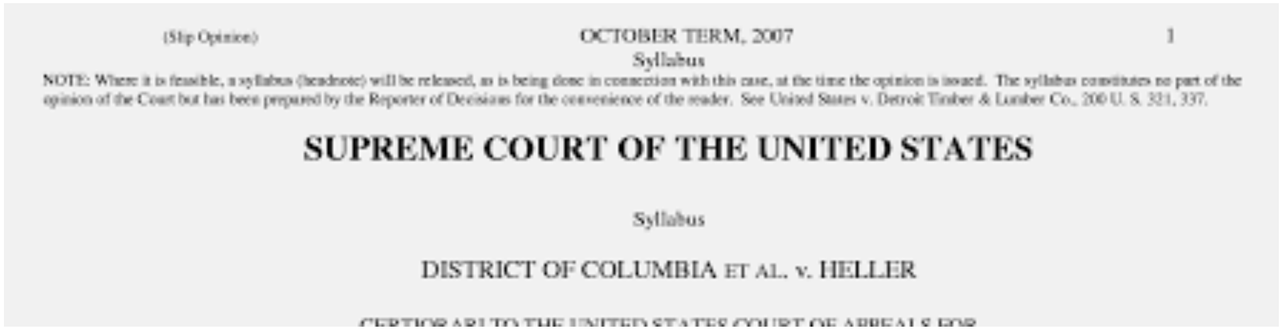
Books

Newspapers

Court opinions

Research papers

.. many more examples


(Slip Opinion) OCTOBER TERM, 2007 1

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

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

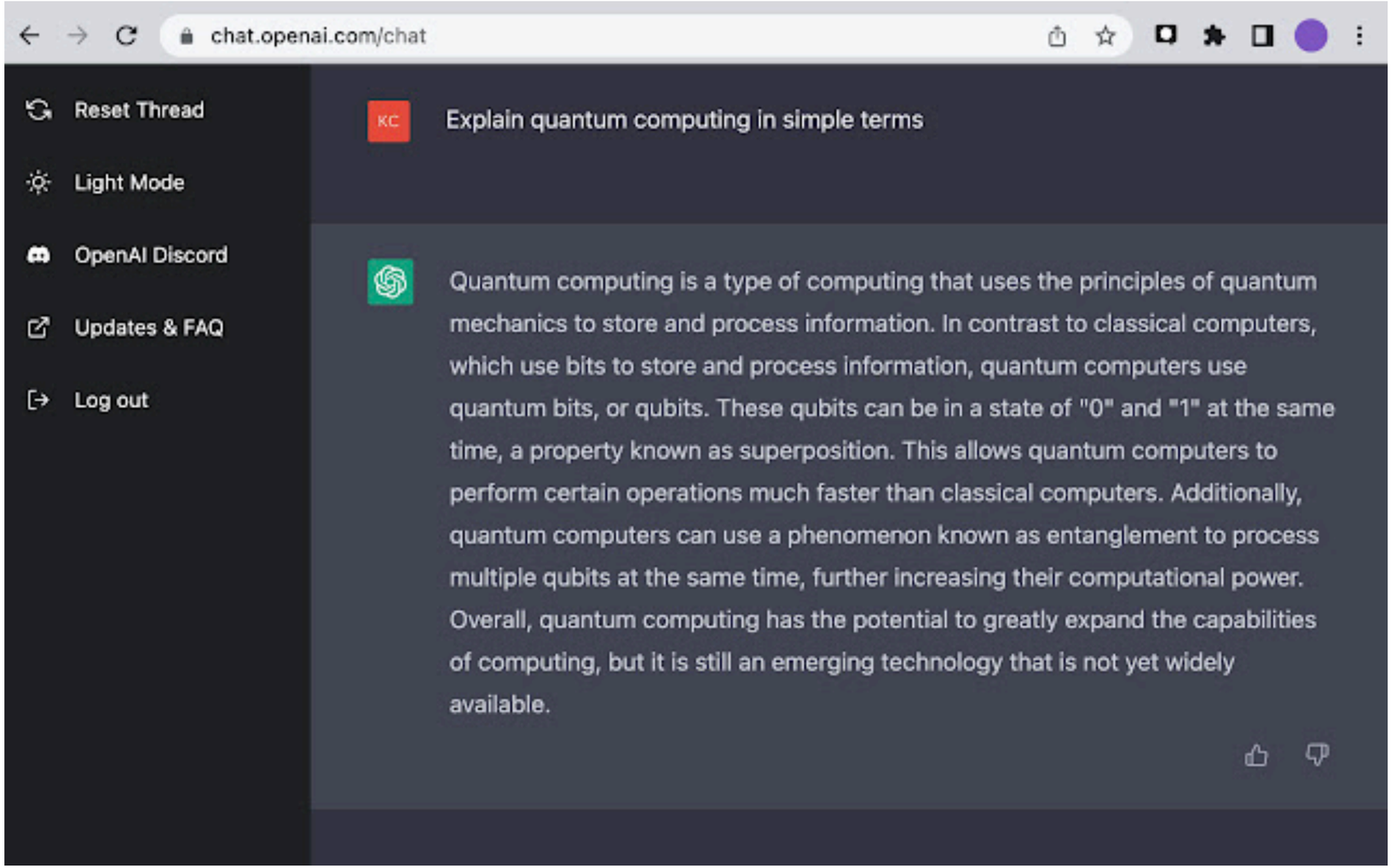
Abstractions with Text

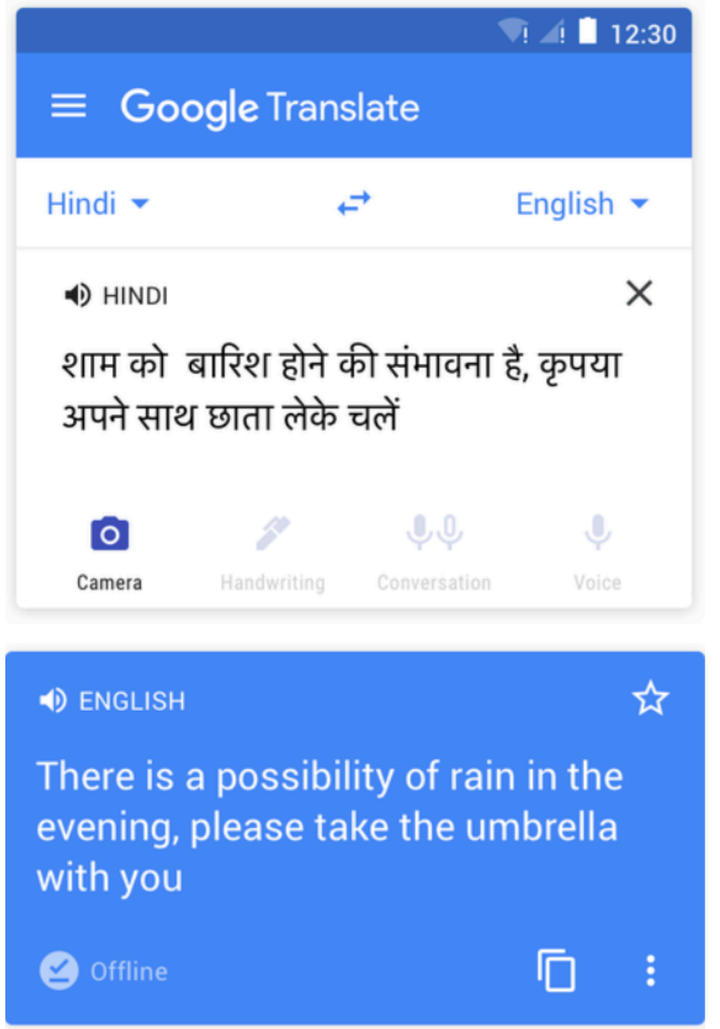
- Text is “massively addressable at different levels of scale” (Whitmore)
  - Characters, lines, paragraphs, etc
  - Characters, words, phrases, sequences, trees, etc
  - Books, genre, collection, etc
  - Political speeches, bail hearings, quarterly results, etc

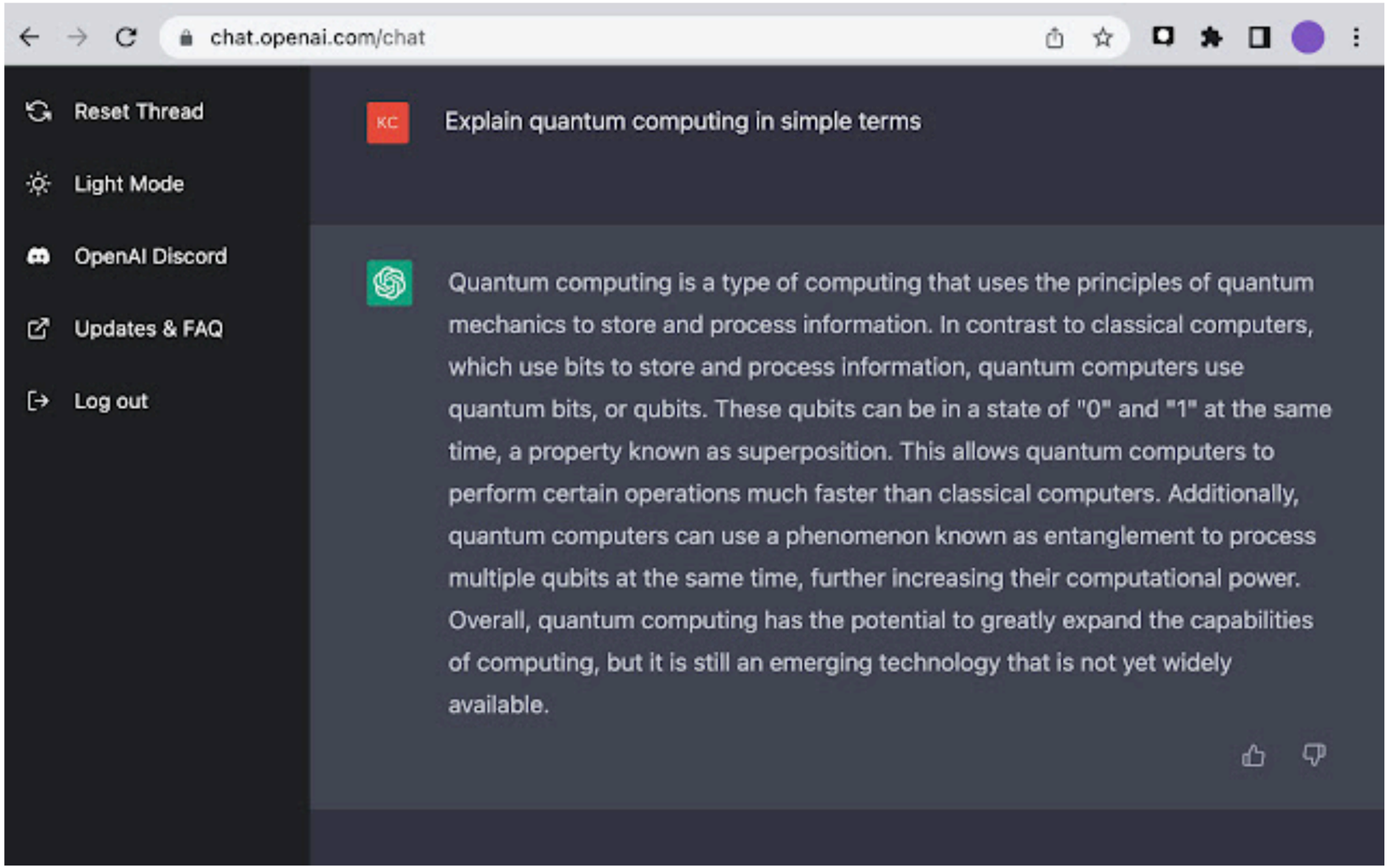
Text presents unique challenges about storing, cleaning, modeling, and interpreting the data

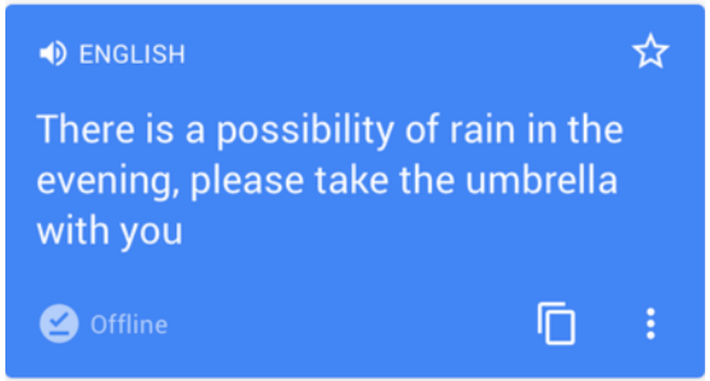
Qualitative Vs Quantitative

- 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


chat.openai.com/chat
Reset Thread
Light Mode
OpenAI Discord
Updates & FAQ
Log out
KC
Explain quantum computing in simple terms
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.


12:30
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





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

HUGGING FACE

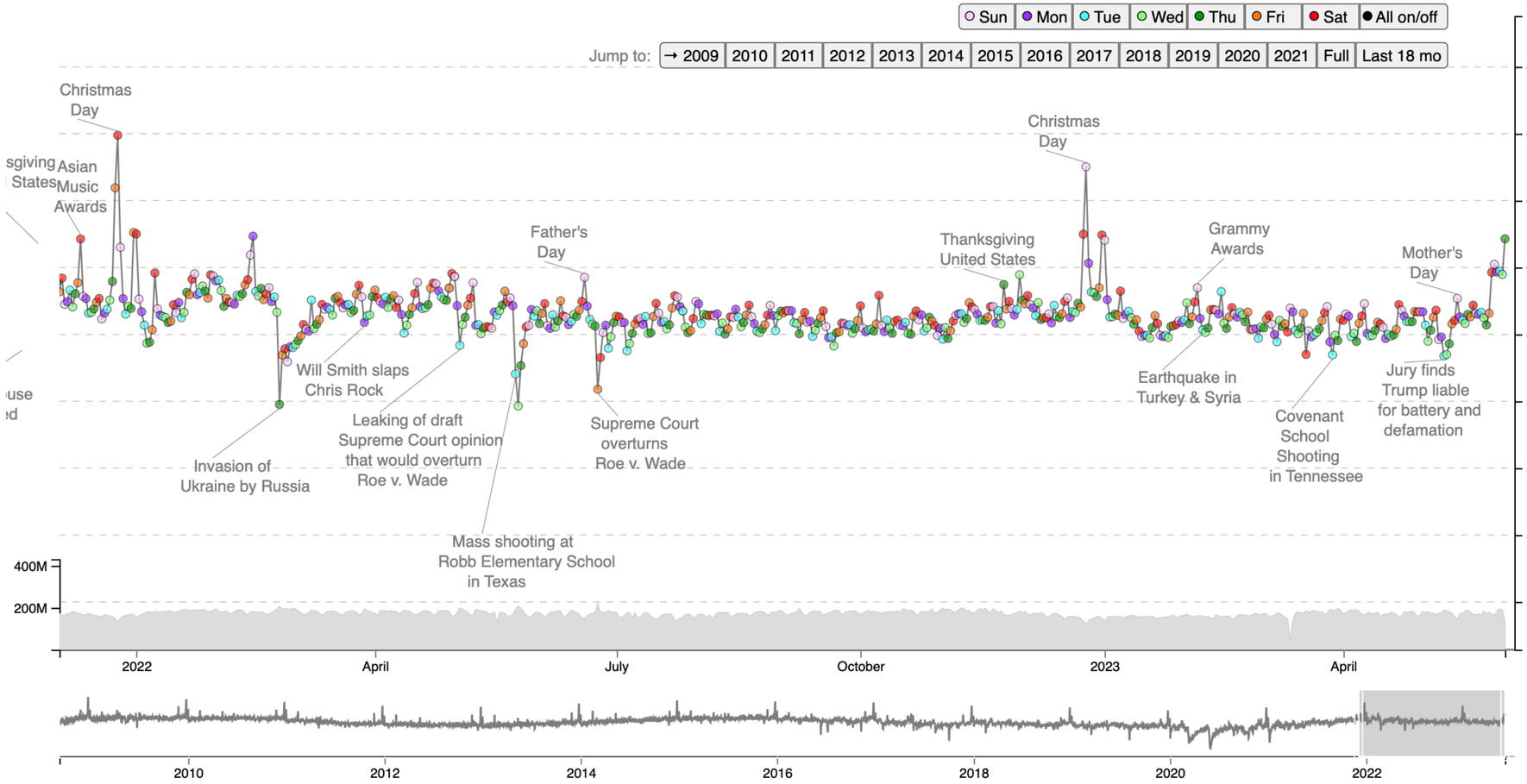
Interjection

What do you want to take away from this class?

Sentiment Analysis

Input:
Tweets

Output:
Average
“happiness”



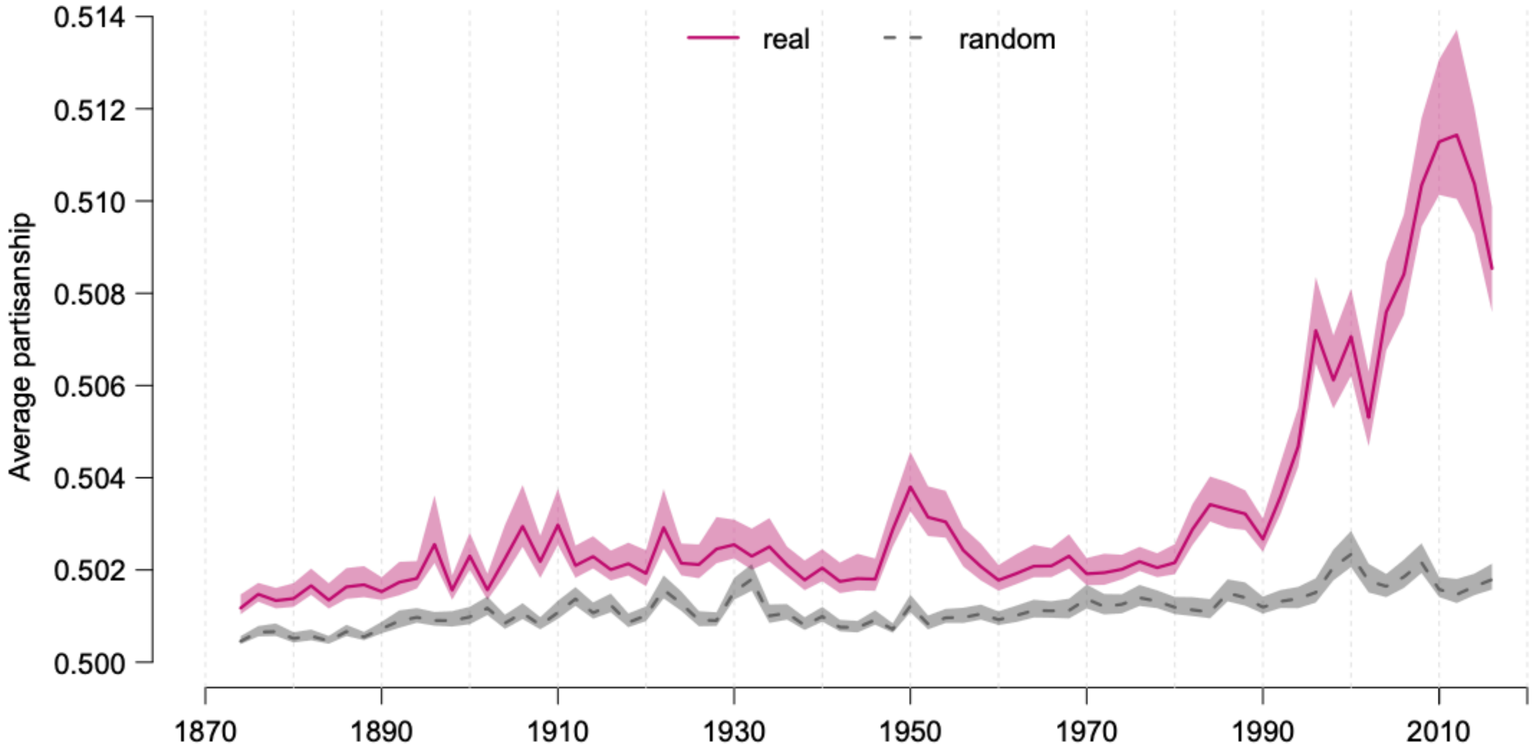
hedonometer.org

(slide credit Brendan O’Connor and Laure Thompson)

Partisanship

Input:
Congressional speech

Output:
Partisanship or party affiliation

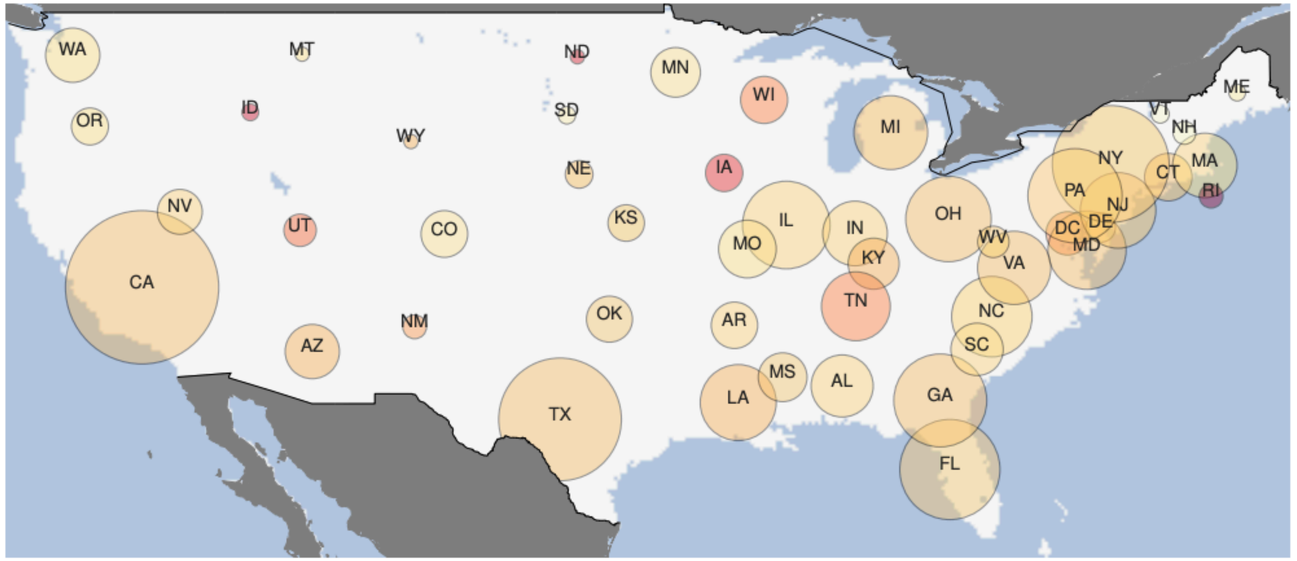


Gentzkow, Shapiro, and Taddy (2016)

Geolocation

Input:
Social
media posts

Output:
Place name



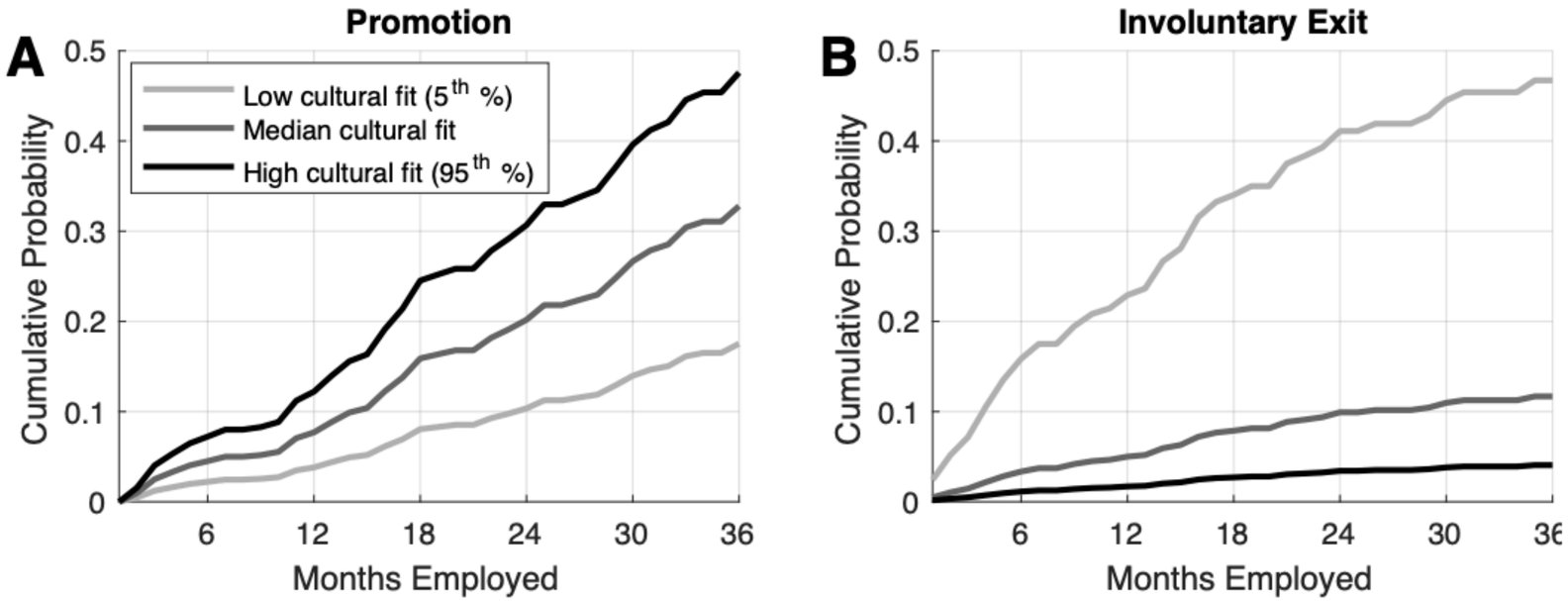
Seattle, WA	Austin, TX	Jacksonville, FL	Columbus, OH	Charlotte, NC	Phoenix, AZ	New Orleans, LA	Baltimore, MD
#goseahawks	stubb	unf	laffayette	#asheville	clutterbuck	mcneese	bhop
smock	gsd	ribault	#weareohio	#depinga	waffels	keela	#dsu
traffuck	#meatsweats	wahoowa	#arcgis	batesburg	bahumbug	pentecostals	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	bmu	slurpies

Rahimi, Cohn, and Baldwin (2018)

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

Ambiguity

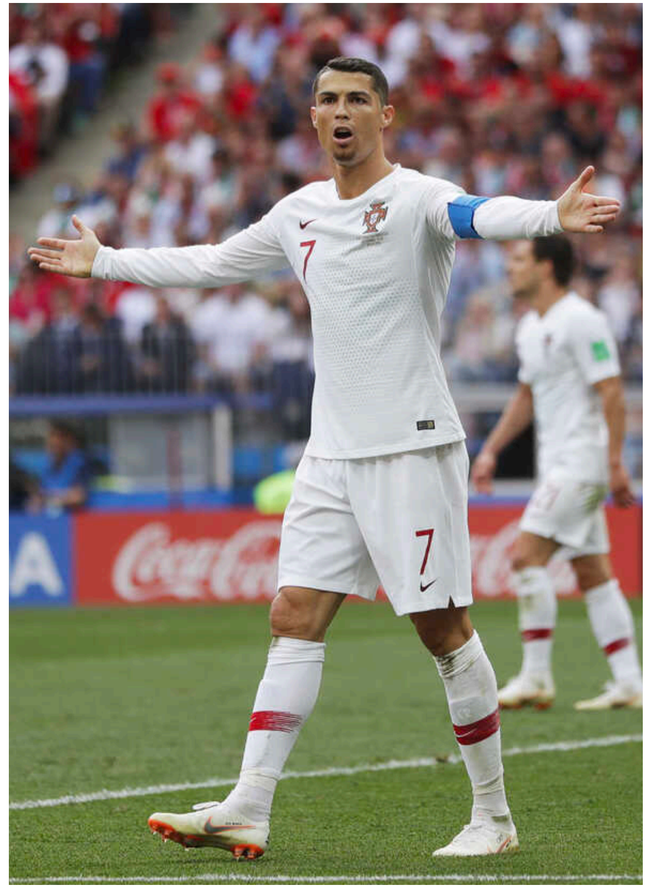
- Language is inherently ambiguous
- This ambiguity is seen at various linguistic levels
- Text modeling and analysis aims at creating robust inference techniques that can handle such ambiguities

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)
  - Please call me by my first name!
- Office hours:
  - Wednesday 11am-12pm (PAIS 588)
  - Friday 11am-12pm (Zoom)
  - If you want to book an appointment, just email me!
- Course website:
  - https://sandeepsoni.github.io/classes/qtm340.html
  - Syllabus and policies page: http://tinyurl.com/mkxmwh86

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 generally expected to finish the readings before each class and participate in the class discussion

Grading

Component	Grade %
Class participation	10%
Reading responses	20%
Problem sets	30%
Group project	40%

Participation: What's expected?

- Attend the class
- Read the required readings and engage in the classroom discussion
- Answer questions on piazza/canvas
- 1% grade is reserved for filling course eval

Reading Response: What's expected?

- There will be 5 reading responses due that answer some questions from the required readings
- You're expected to submit 4 responses (highest 4 graded responses will be counted towards final grade)
- The reading response will synthesize the main takeaways from the required readings
- No collaboration and no late 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 (3-4 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 (in files)

In class

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