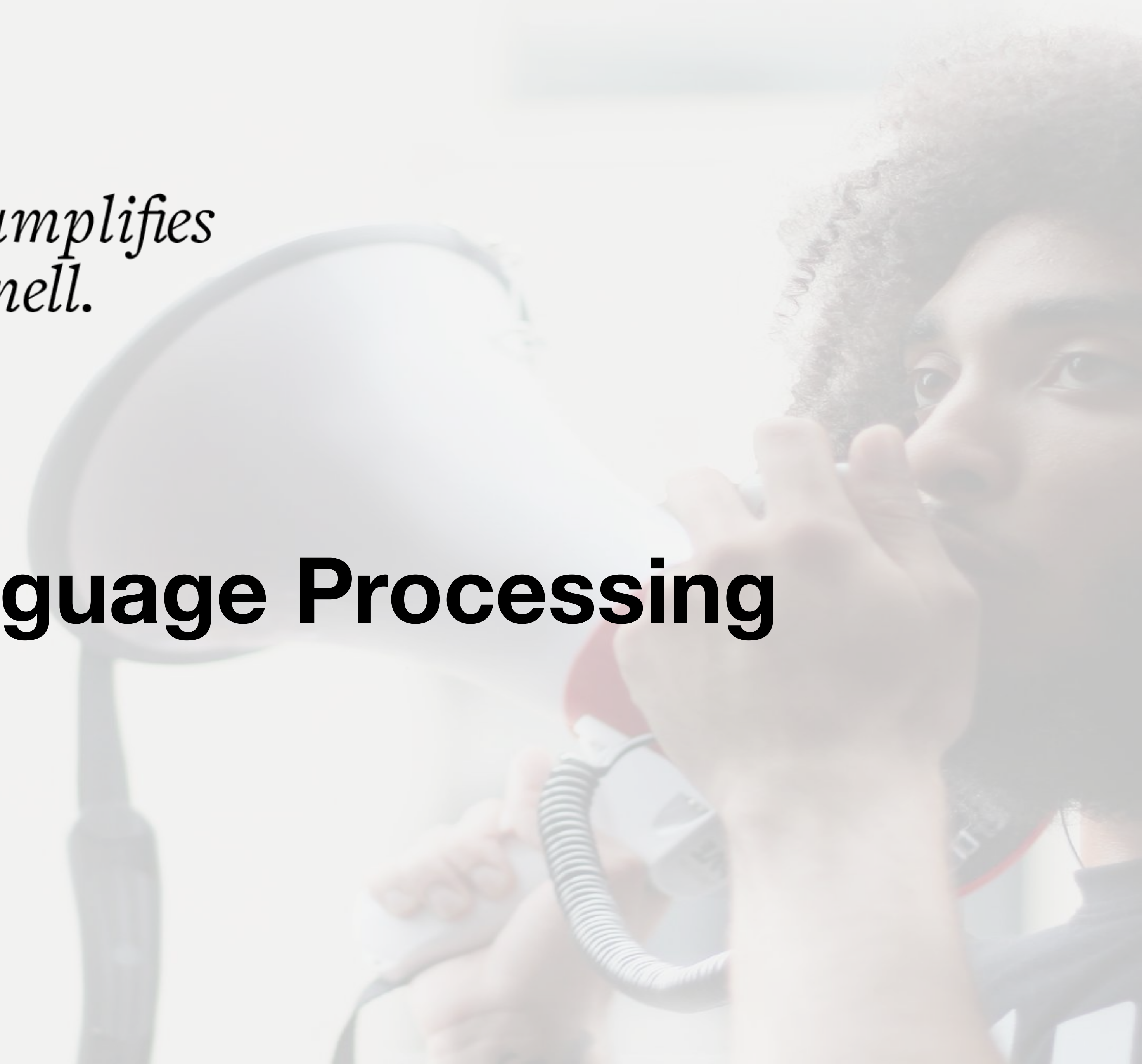
Cornell University

CORNELL CENTER for
SOCIAL SCIENCES

Accelerates, enhances, and amplifies social science research at Cornell.



Natural Language Processing

Remy Stewart
CCSS-ResearchSupport@cornell.edu

CCSS Research Support Code of Conduct

The Cornell Center for Social Sciences provides a welcoming environment for everyone embracing all backgrounds or identities. All instructors and attendees agree to abide by our community norms. We encourage the following behaviors in our workshops:

- Respect differing viewpoints and ideas
- Share your own perspectives and ask any questions
- Accept constructive criticism
- Use welcoming and inclusive language
- Show courtesy and respect for all instructors and attendees

If you believe that an instructor or attendee has violated the code of conduct, please report the violation to CCSS-ResearchSupport@cornell.edu. We take all reported incidents seriously.

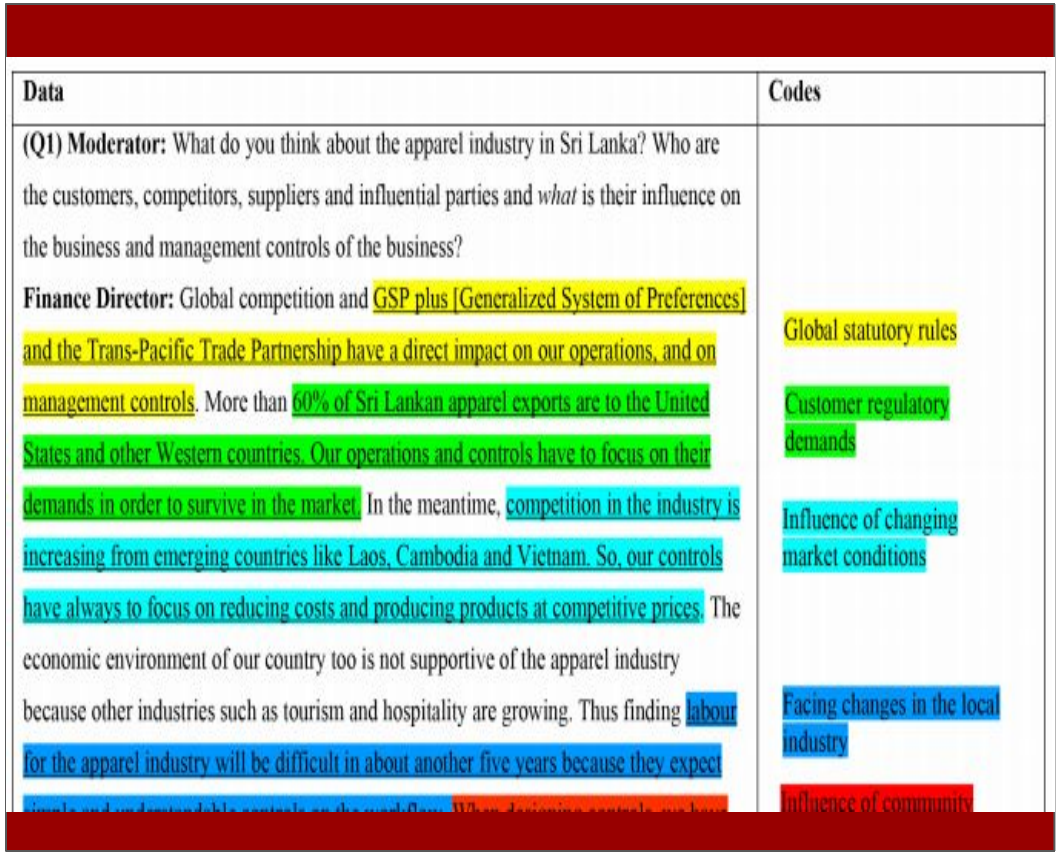
Land Acknowledgement

Cornell University is located on the traditional homelands of the Gayogo̱hó:nǫ' (the Cayuga Nation). The Gayogo̱hó:nǫ' are members of the Haudenosaunee Confederacy, an alliance of six sovereign Nations with a historic and contemporary presence on this land. The Confederacy precedes the establishment of Cornell University, New York state, and the United States of America. We acknowledge the painful history of Gayogo̱hó:nǫ' dispossession, and honor the ongoing connection of Gayogo̱hó:nǫ' people, past and present, to these lands and waters.

Here are additional links for more on the history of Cornell's violent, colonial formation, the movement to return native lands, and about the AIISP program at Cornell.

Consider donating to the Gayogo̱hó:nǫ' sovereignty initiative here.

Text as Data

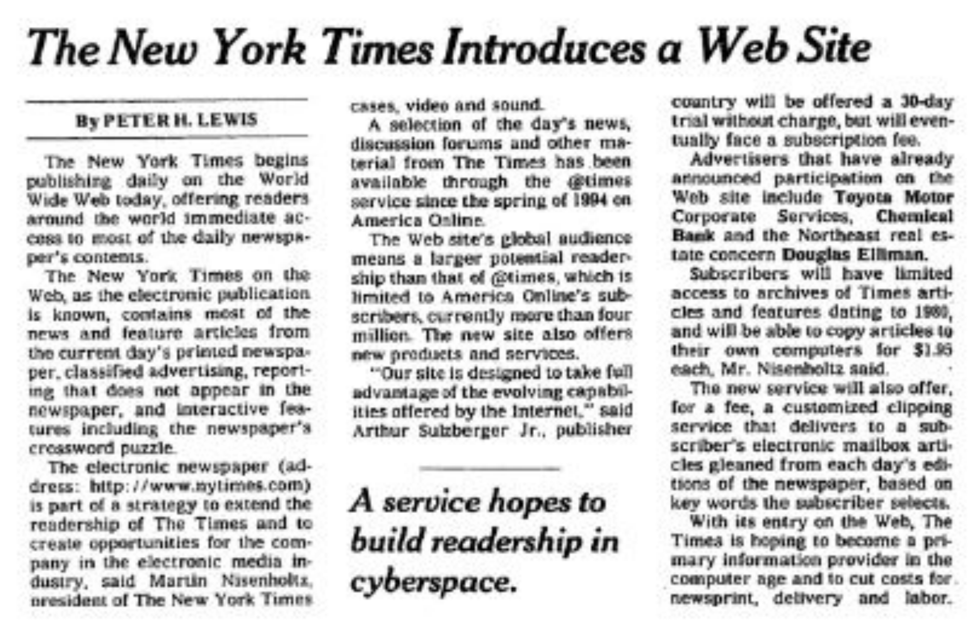


Data	Codes
(Q1) **Moderator:** What do you think about the apparel industry in Sri Lanka? Who are the customers, competitors, suppliers and influential parties and *what* is their influence on the business and management controls of the business? **Finance Director:** Global competition and GSP plus [Generalized System of Preferences] and the Trans-Pacific Trade Partnership have a direct impact on our operations, and on management controls. More than 60% of Sri Lankan apparel exports are to the United States and other Western countries. Our operations and controls have to focus on their demands in order to survive in the market. In the meantime, competition in the industry is increasing from emerging countries like Laos, Cambodia and Vietnam. So, our controls have always to focus on reducing costs and producing products at competitive prices. The economic environment of our country too is not supportive of the apparel industry because other industries such as tourism and hospitality are growing. Thus finding labour for the apparel industry will be difficult in about another five years because they expect simple and understandable controls on the workflow. When designing controls, we have	Global statutory rules Customer regulatory demands Influence of changing market conditions Facing changes in the local industry Influence of community

- Text has been a core data source within social sciences
- Often previously limited to qualitative methods
- Digital Age has rapidly shifted these circumstances
  - Wide range of innovative computational methods
  - Powerful & cheaper compute resources
  - Open-source software that is accessible to learn

Text Data in the Digital Age

- Text data is rapidly created as written behavior is recorded online
- Linked to the growth of social media & online communities
- Digital conversion of printed books, historical records, government documents, etc.



The New York Times Introduces a Web Site

By PETER H. LEWIS

The New York Times begins publishing daily on the World Wide Web today, offering readers around the world immediate access to most of the daily newspaper's contents.

The New York Times on the Web, as the electronic publication is known, contains most of the news and feature articles from the current day's printed newspaper, classified advertising, reporting that does not appear in the newspaper, and interactive features including the newspaper's crossword puzzle.

The electronic newspaper (address: http://www.nytimes.com) is part of a strategy to extend the readership of The Times and to create opportunities for the company in the electronic media industry, said Martin Nisenholtz, oresident of The New York Times

cases, video and sound.

A selection of the day's news, discussion forums and other material from The Times has been available through the @times service since the spring of 1994 on America Online.

The Web site's global audience means a larger potential readership than that of @times, which is limited to America Online's subscribers, currently more than four million. The new site also offers new products and services.

"Our site is designed to take full advantage of the evolving capabilities offered by the Internet," said Arthur Sulzberger Jr., publisher

A service hopes to build readership in cyberspace.

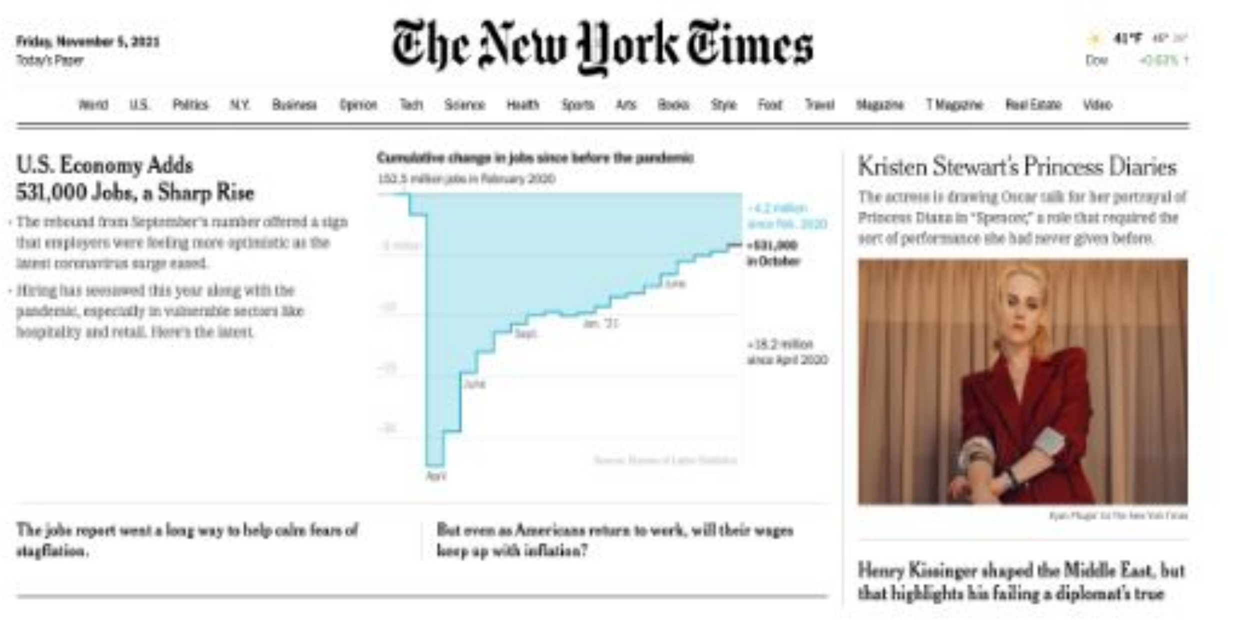
country will be offered a 30-day trial without charge, but will eventually face a subscription fee.

Advertisers that have already announced participation on the Web site include **Toyota Motor Corporate Services**, **Chemical Bank** and the Northeast real estate concern **Douglas Elliman.**

Subscribers will have limited access to archives of Times articles and features dating to 1980, and will be able to copy articles to their own computers for $1.95 each, Mr. Nisenholtz said.

The new service will also offer, for a fee, a customized clipping service that delivers to a subscriber's electronic mailbox articles gleaned from each day's editions of the newspaper, based on key words the subscriber selects.

With its entry on the Web, The Times is hoping to become a primary information provider in the computer age and to cut costs for newsprint, delivery and labor.



Friday, November 5, 2021
Today's Paper

The New York Times

41°F 45° 39°
Dow +0.63% ↑

World U.S. Politics N.Y. Business Opinion Tech Science Health Sports Arts Books Style Food Travel Magazine T Magazine Real Estate Video

U.S. Economy Adds 531,000 Jobs, a Sharp Rise

- The rebound from September's number offered a sign that employers were feeling more optimistic as the latest coronavirus surge eased.
- Hiring has seesawed this year along with the pandemic, especially in vulnerable sectors like hospitality and retail. Here's the latest.

Cumulative change in jobs since before the pandemic
152.5 million jobs in February 2020
-4.2 million since Feb. 2020
+531,000 in October
+18.2 million since April 2020
-5 million
-10
-15
-20
April
June
Sept.
Jan. '21
June
Source: Bureau of Labor Statistics

The jobs report went a long way to help calm fears of stagflation.

But even as Americans return to work, will their wages keep up with inflation?

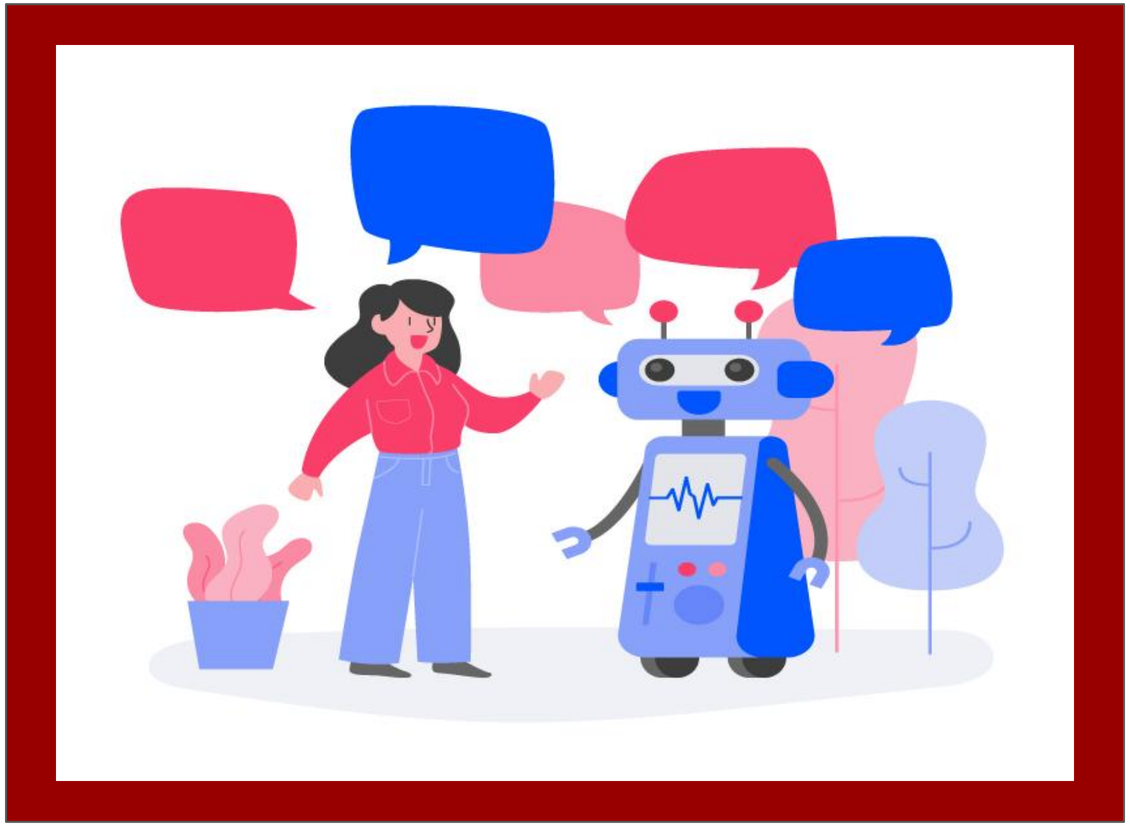
Kristen Stewart's Princess Diaries

The actress is drawing Oscar talk for her portrayal of Princess Diana in "Spencer," a role that required the sort of performance she had never given before.

Ryan Pfluger for The New York Times

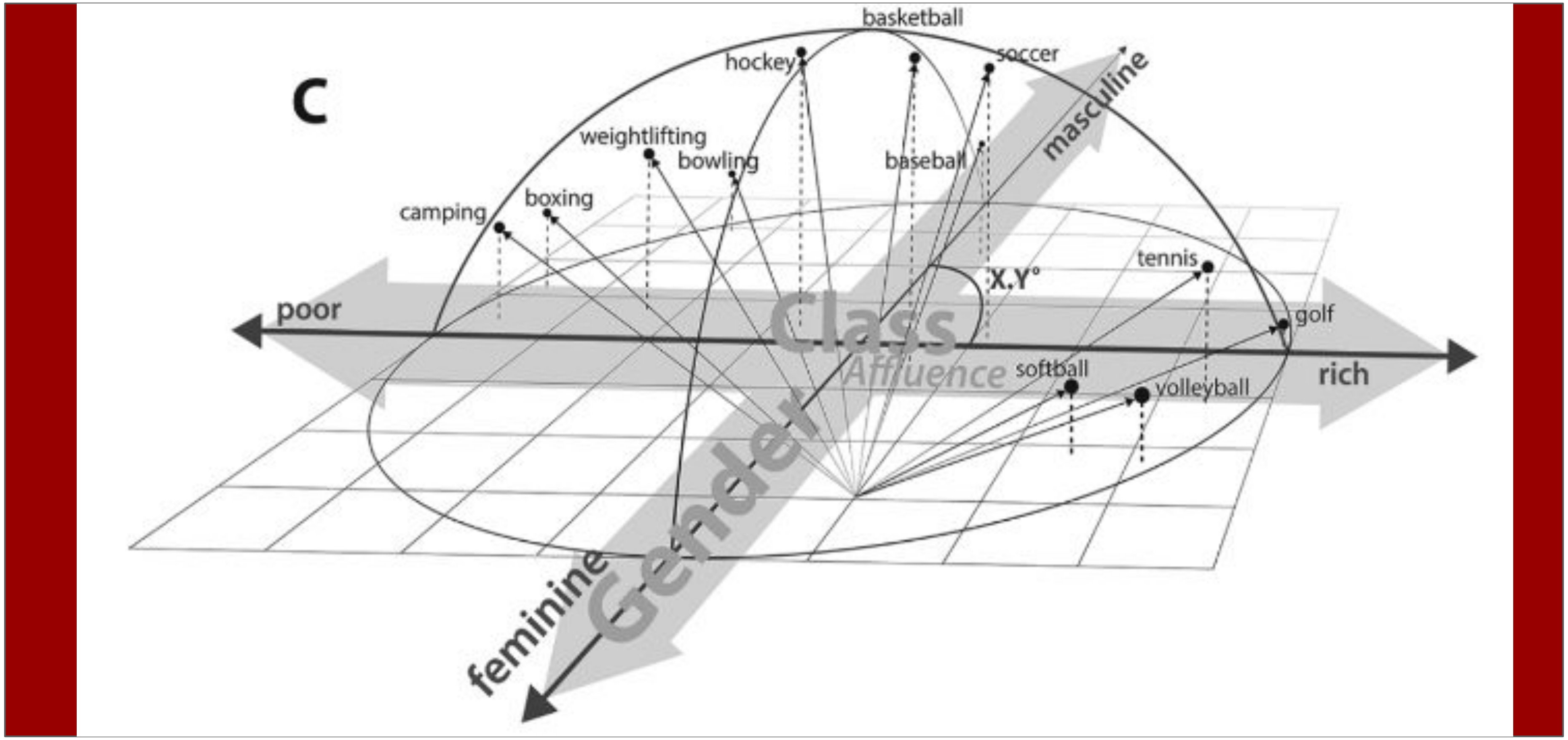
Henry Kissinger shaped the Middle East, but that highlights his failing a diplomat's true

Natural Language Processing (NLP)

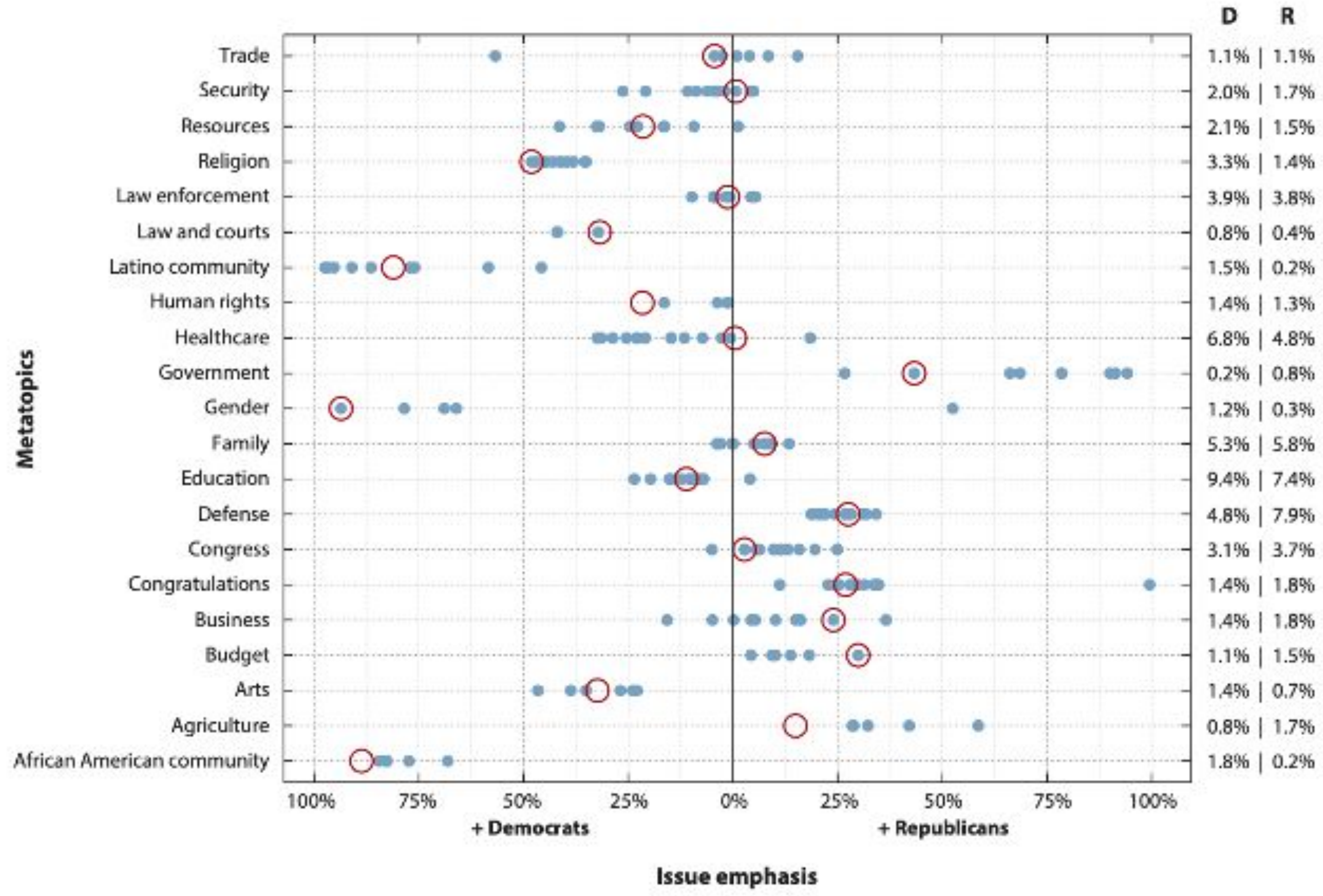


- Interdisciplinary field using computational tools to interpret text & language
- While not exclusively ML based, ML is highly influential within NLP
- Powers the AI behind language translation, audio-to-text transcription, speech recognition (Siri, Alexa)
- NLP as a natural computational approach to incorporate within social sciences

Word embeddings from Kozlowski, Taddy, and Evans (2019)



Structural Topic Models from Wilkerson & Casas (2017)



Core Terminology

1. A **corpus** is the total collection of text being analyzed.

All 400k of our tweets.

2. A **vocabulary** is all of the unique words in a corpus.

All words used at least once across the tweets.

3. A **document** is an individual record within the corpus.

A user tweets “This brings me so much joy.”

4. **Tokens** are the smaller text units within a document. Most commonly refers to words, but can also be sentences, paragraphs, or text characters.

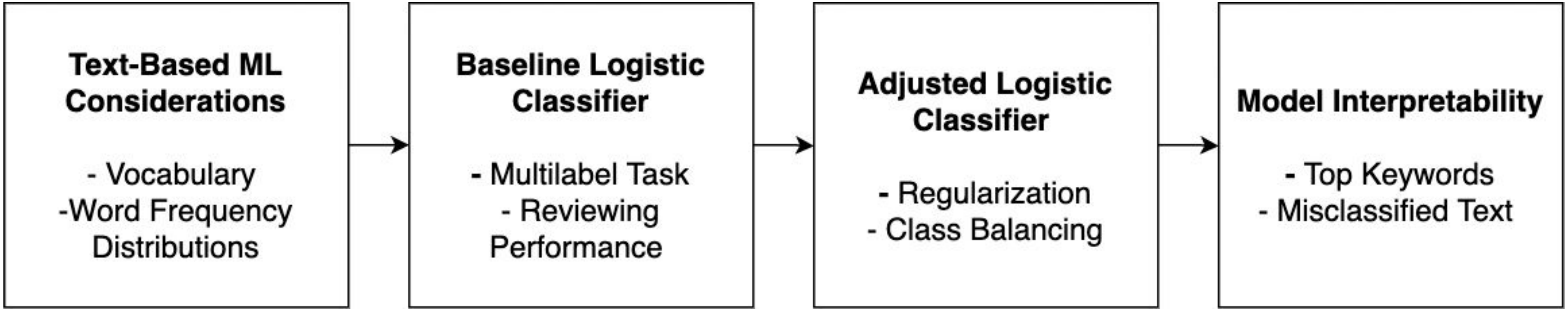
“This”, “brings”, “me”, “so”, …

5. **N-grams** considers multiple word tokens at once with “n” as a specified number.

Bi-gram of “This_brings”
Tri-Gram of “This_brings_me”

Coding Demo Outline

Supervised Text Classifier



Text-Based ML Considerations

- Vocabulary
-Word Frequency Distributions

Baseline Logistic Classifier

- Multilabel Task
- Reviewing Performance

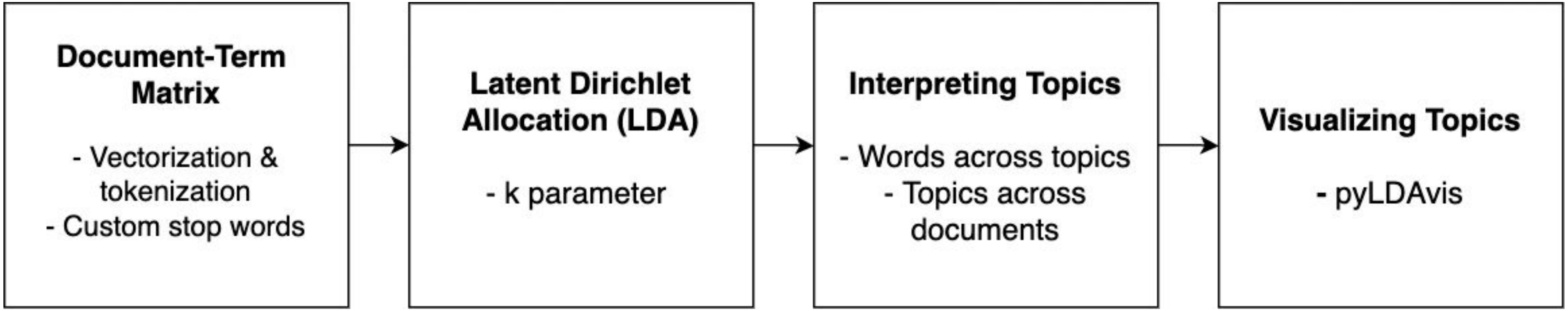
Adjusted Logistic Classifier

- Regularization
- Class Balancing

Model Interpretability

- Top Keywords
- Misclassified Text

Unsupervised Topic Model



Document-Term Matrix

- Vectorization & tokenization
- Custom stop words

Latent Dirichlet Allocation (LDA)

- k parameter

Interpreting Topics

- Words across topics
- Topics across documents

Visualizing Topics

- pyLDAvis

Let’s head to the Github for our final coding demo!

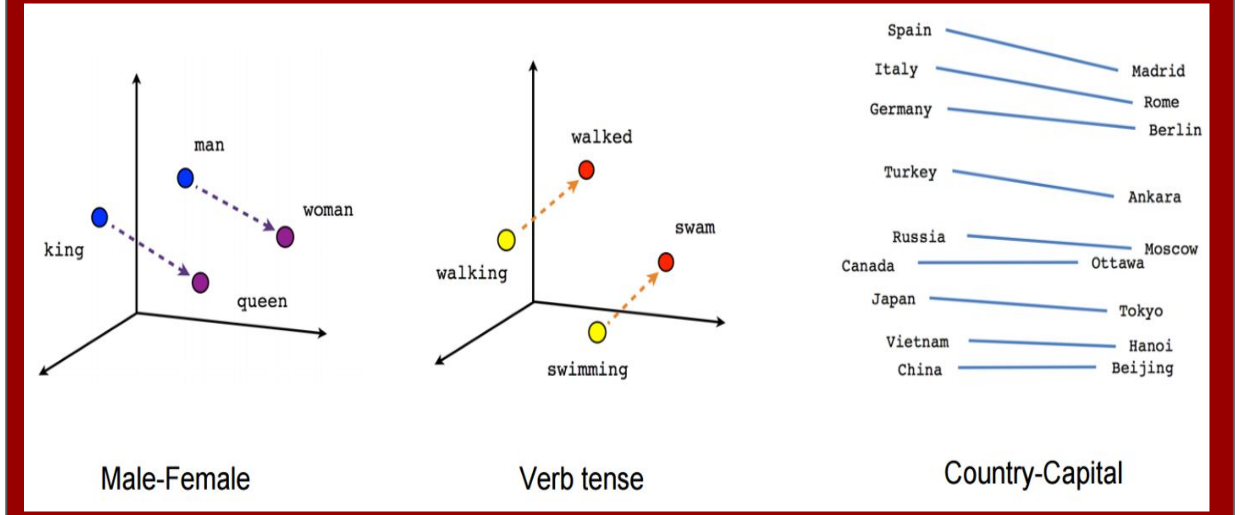
Alternatively, you can access this session’s Colab file directly via the following link:

https://colab.research.google.com/drive/1VwetZHUP75J
I6rKcmKI26skQaJ2J-zvR?usp=sharing

Additional NLP Methods

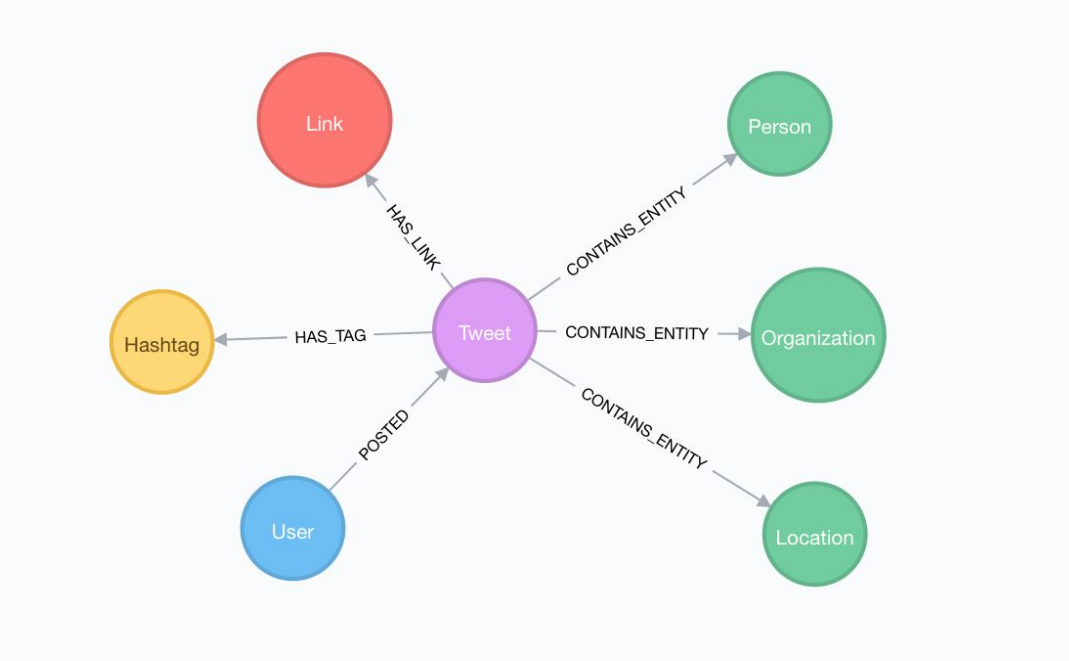
Word Embeddings

Mapping word similarity within vector spaces



Text Networks

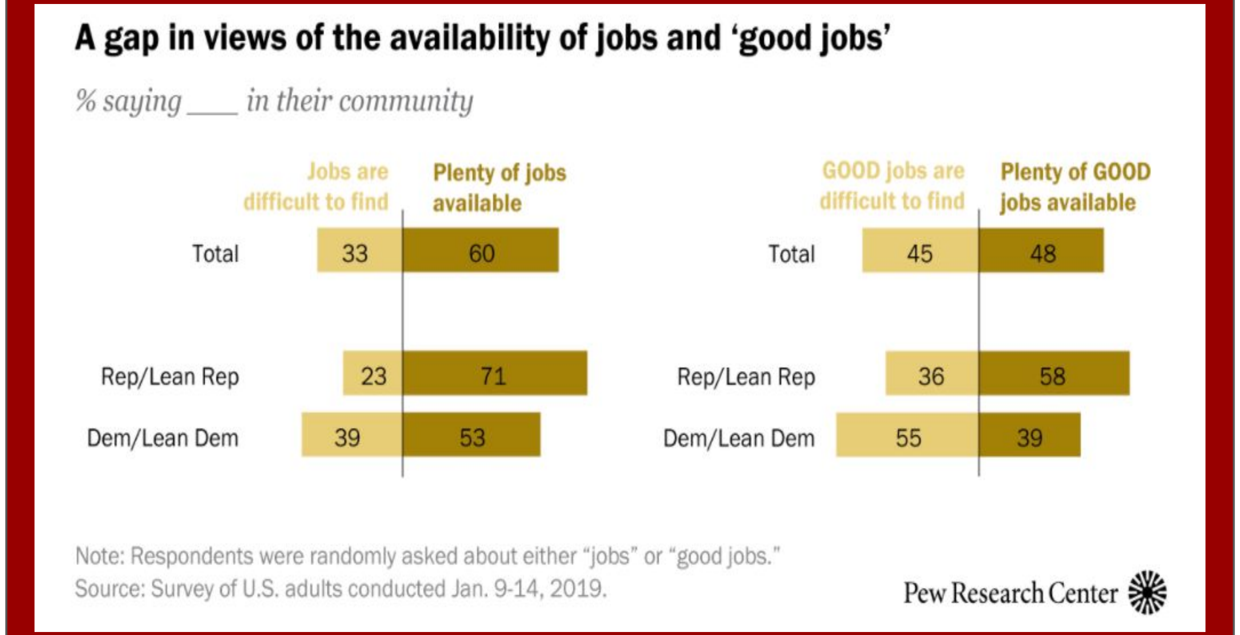
Node/edge relationships between text



Additional NLP Methods

Text & Causality

Text as treatment vs. text as outcome



A gap in views of the availability of jobs and 'good jobs'

% saying ___ in their community

Note: Respondents were randomly asked about either "jobs" or "good jobs."
Source: Survey of U.S. adults conducted Jan. 9-14, 2019.

Pew Research Center

Neural Networks

Classification, question answering, summation, and beyond

Transformers

PyTorch Lightning

OpenAI

Ethics Within ML

- Data is a fundamentally social product embedded within structural inequities

- ML has great ability to reinforce bias & perpetuate harm

- Many ethical concerns within ML only engaged with after problems are retroactively discovered

why are black women so
why are black women so angry
why are black women so loud
why are black women so mean
why are black women so attractive
why are black women so lazy
why are black women so annoying
why are black women so confident
why are black women so sassy
why are black women so insecure
ALGORITHMS OF OPPRESSION
HOW SEARCH ENGINES REINFORCE RACISM
SAFIYA UMOJA NOBLE
"This book is downright scary—but…you will emerge smarter and more empowered to demand justice." —NAOMI KLEIN
AUTOMATING INEQUALITY
HOW HIGH-TECH TOOLS PROFILE, POLICE, AND PUNISH THE POOR
VIRGINIA EUBANKS
WEAPONS OF MATH DESTRUCTION
HOW BIG DATA INCREASES INEQUALITY AND THREATENS DEMOCRACY
CATHY O'NEIL

Ethics within ML

Ethical Questions within ML Research

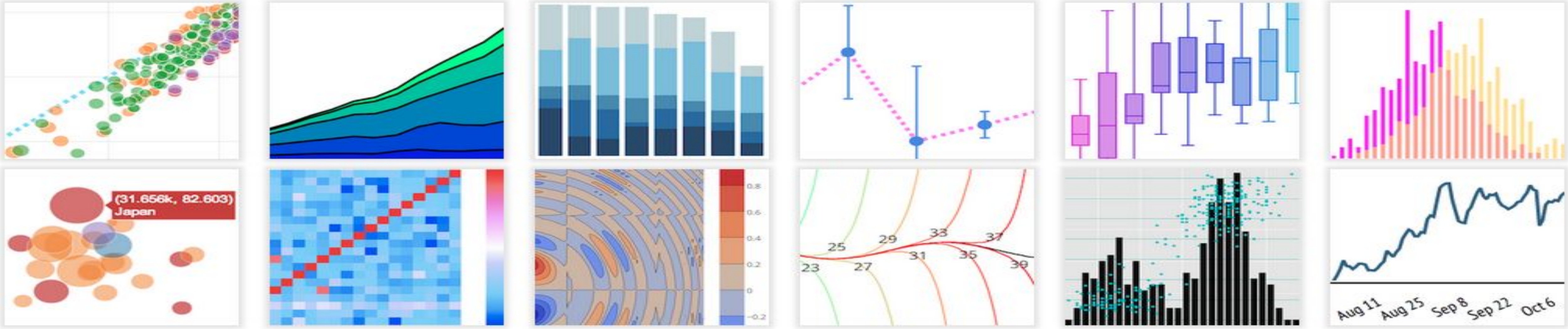
- What are the potential risks associated with the model?
- What is the broadest range of potential consequences/outcomes of this research?
- How could vulnerable groups be impacted?
- What are the relationships between researcher, participants, & third parties?
- What are expectations around user autonomy, privacy, informed consent?

Core Ethical Considerations

- Model interpretability & transparency
- Participant agency & community engagement
- Data privacy

Key Bootcamp Takeaways

- Learned core ML concepts
- Processed and modeled structured and unstructured data
- Applied both supervised and unsupervised methods
- Thought through key decision points within ML analyses
- Compared models for different use cases



Conclusion

Social scientists are in a unique position to address essential questions with ML given our disciplinary perspectives.

This series provided a broad overview to extend through further learning.

The world of ML can be intimidating- we truly believe that you can master these skills.

We hope this bootcamp series can serve as a source of inspiration for your own research!

CORNELL CENTER for SOCIAL SCIENCES

Research Services

Computing

- Shared & Dedicated Cloud Computing Resources
- Secure Data Enclave
- Software
  - Stata, SPSS, SAS
  - Python
  - R
  - Matlab
  - ArcGIS
  - Atlas.ti
  - MaxQDA

Data Services

- Data Access
- Secure Data Services
- Data Sharing & Archiving
- Results Reproduction

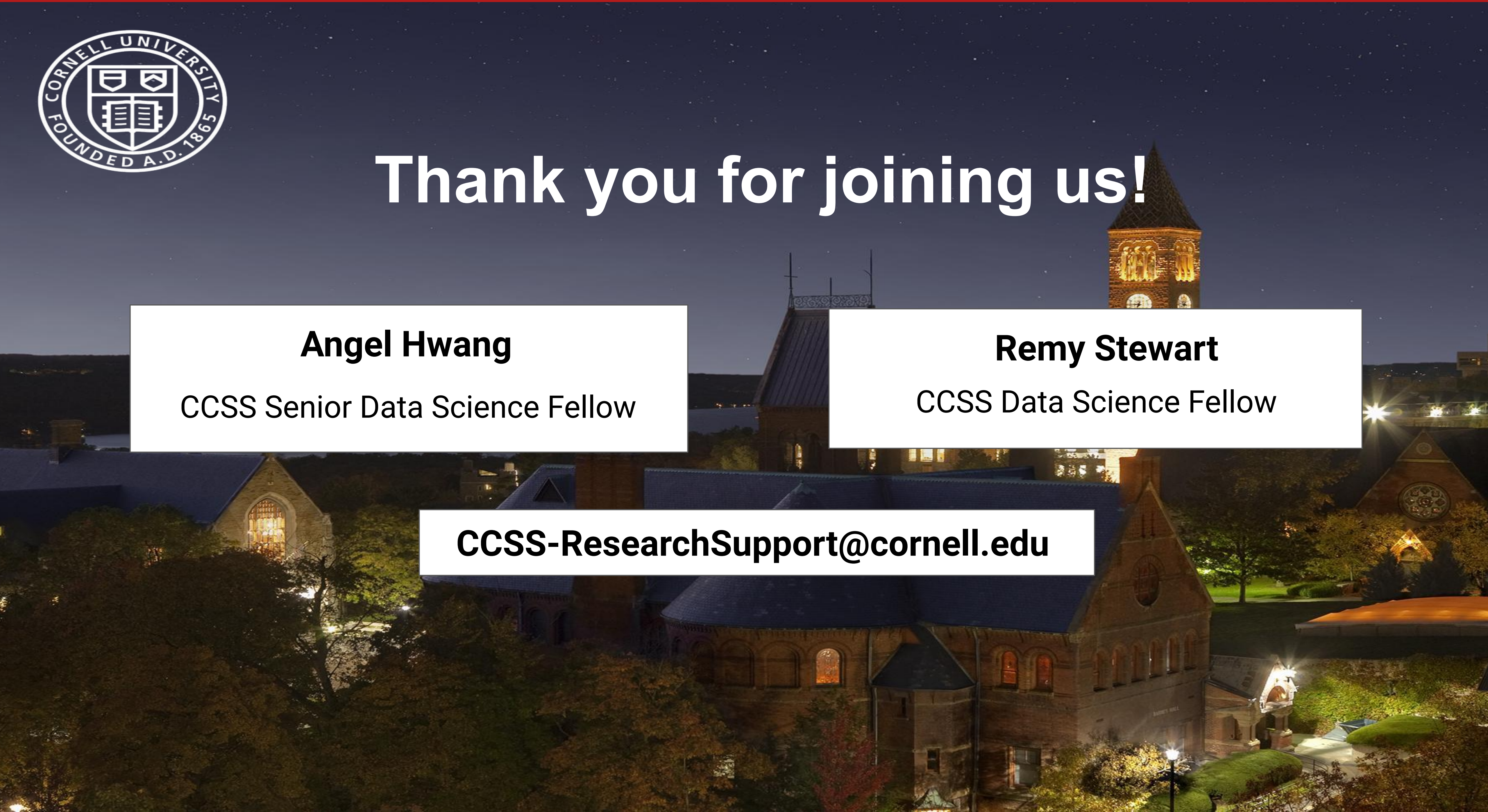
Consulting

- Popular Topics:
  - Qualitative Methods
  - Data Cleaning
  - Visualizations
  - Survey Analysis
  - Finding Data
  - Web Scraping
  - Text Analysis

CORNELL CENTER for
SOCIAL SCIENCES

Please Fill Out Our Evaluation Forum!

We’d love to hear your feedback to improve this Bootcamp series.



Thank you for joining us!

Angel Hwang

CCSS Senior Data Science Fellow

Remy Stewart

CCSS Data Science Fellow

CCSS-ResearchSupport@cornell.edu