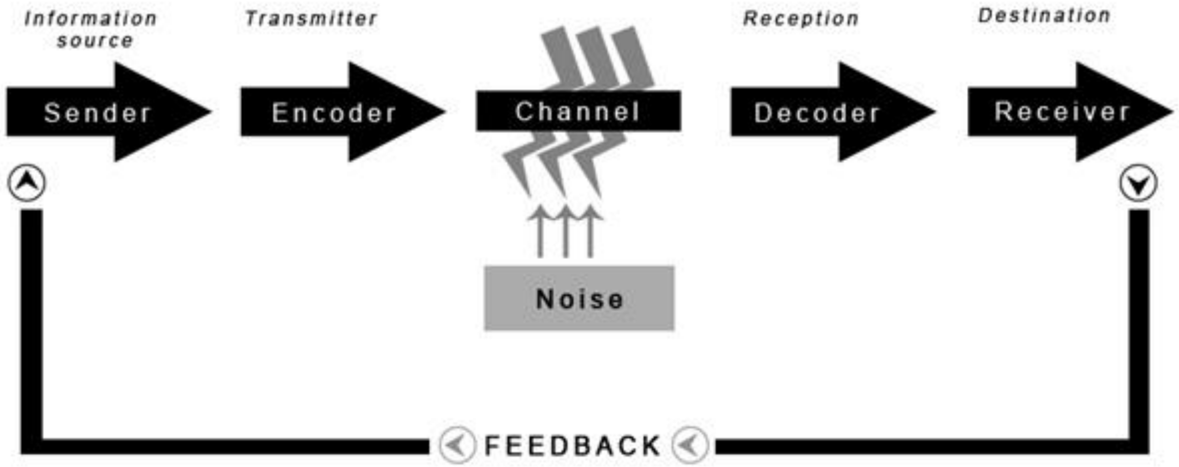
Natural Language Processing with ***Frames, Discourse, and Agendas***

(informal title: descriptive results are not bad)

Josephine Lukito

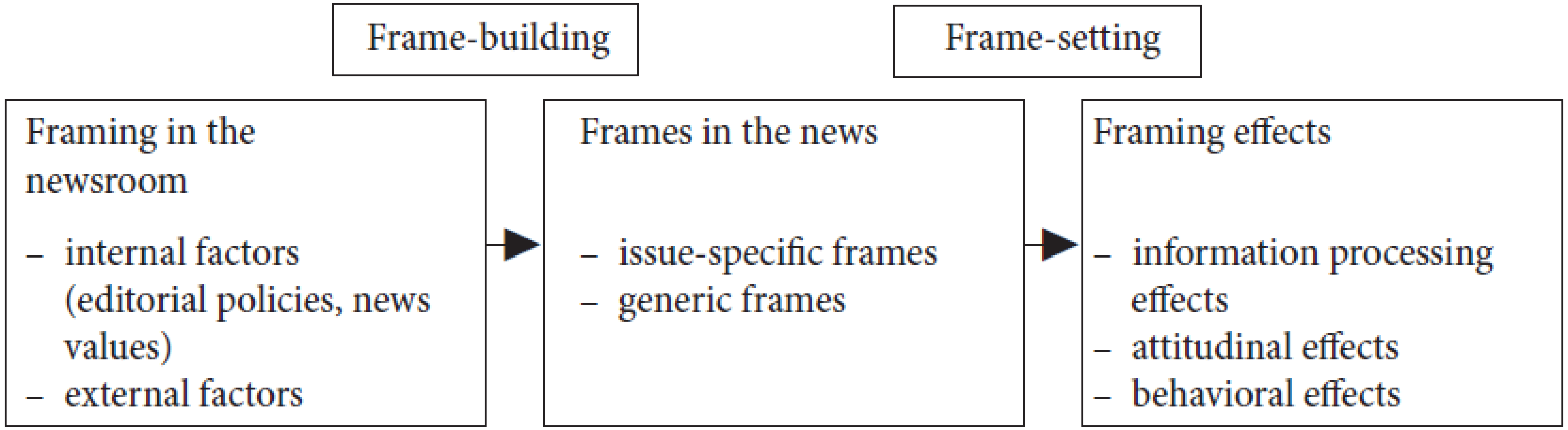
WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Shannon-Weaver Model & Language



Shannon-Weaver's Model of Communication

The Framing Process



Frame-building

Frame-setting

Framing in the newsroom

- internal factors (editorial policies, news values)
- external factors

Frames in the news

- issue-specific frames
- generic frames

Framing effects

- information processing effects
- attitudinal effects
- behavioral effects

Figure 1. An integrated process model of framing

De Vreese (2005)

Computational Text Analysis

- Text-as-Data
- Natural Language Processing
- Computational/Corpus Linguistics
- Computationally-assisted Content Analyses
- Text analytics

Content Analysis and Computational Methods

- Lewis, S. C., Zamith, R., & Hermida, A. (2013). Content analysis in an era of big data: A hybrid approach to computational and manual methods. *Journal of Broadcasting & Electronic Media*, *57*(1), 34-52.
- Boumans, J. W., & Trilling, D. (2016). Taking stock of the toolkit: An overview of relevant automated content analysis approaches and techniques for digital journalism scholars. *Digital Journalism*, *4*(1), 8-23.
- Burscher, B., Odijk, D., Vliegenthart, R., De Rijke, M., & De Vreese, C. H. (2014). Teaching the computer to code frames in news: Comparing two supervised machine learning approaches to frame analysis. *Communication Methods and Measures*, *8*(3), 190-206.
- Opperhuizen, A. E., Schouten, K., & Klijn, E. H. (2019). Framing a Conflict! How Media Report on Earthquake Risks Caused by Gas Drilling: A Longitudinal Analysis Using Machine Learning Techniques of Media Reporting on Gas Drilling from 1990 to 2015. *Journalism Studies*, *20*(5), 714-734.

3 Projects

- **Project 1**: Using dependency parsing to study how protesters and police are framed in U.S. news
- **Project 2**: Using word embeddings to study discourse shifts during the 1st 2016 U.S. Presidential debate.
- **Project 3**: Using structural topic modeling to study news coverage of agendas during the 2016 U.S. Presidential election.

Project 1: Adjectives for Protesters and Police

- **NLP for identifying how actors are framed.**
- **Research Question**: How do news media frame protesters and police in news coverage of four 21st century protests?
- **Corpus**:
  - 2,174 news articles
  - 5 outlets: Breitbart, *CNN*, *Fox*, *MSNBC*, and *The New York Times*
  - 4 Events: Charlottesville, Ferguson, Women's March, and Dakota Pipeline
- **Method**: Dependency parsing, entity-sentiment[ish]

Lukito, J., McLeod, D. & Boyle, M. (2019). Allies and Opponents of the Status Quo: Partisan News Media Descriptions of Protesters and Police in Four 21st Century Protests. **[To be presented at #ica19, Monday 5-6:15]**

Project 1: Results

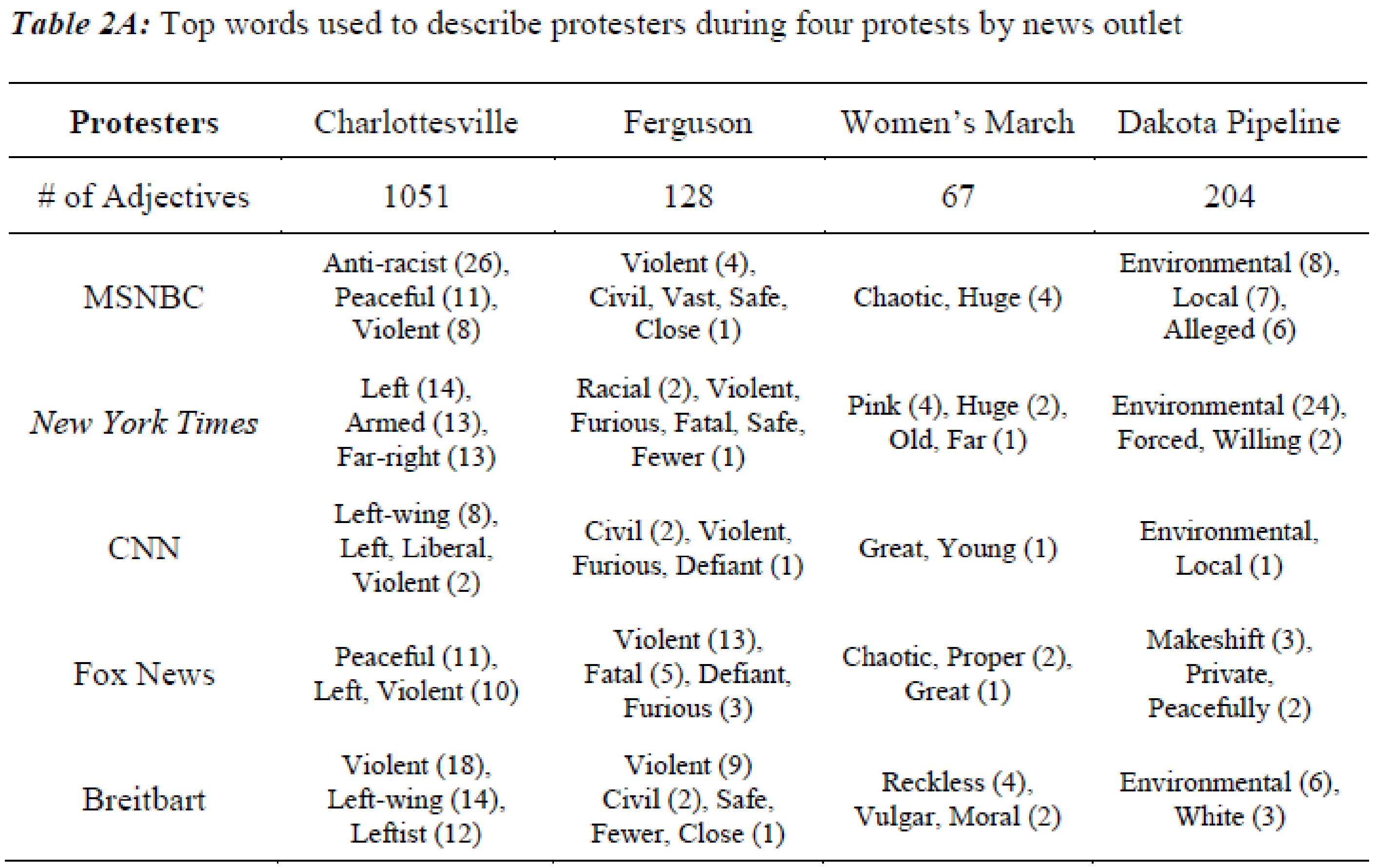


Table 2A: Top words used to describe protesters during four protests by news outlet

Protesters	Charlottesville	Ferguson	Women's March	Dakota Pipeline
# of Adjectives	1051	128	67	204
MSNBC	Anti-racist (26), Peaceful (11), Violent (8)	Violent (4), Civil, Vast, Safe, Close (1)	Chaotic, Huge (4)	Environmental (8), Local (7), Alleged (6)
New York Times	Left (14), Armed (13), Far-right (13)	Racial (2), Violent, Furious, Fatal, Safe, Fewer (1)	Pink (4), Huge (2), Old, Far (1)	Environmental (24), Forced, Willing (2)
CNN	Left-wing (8), Left, Liberal, Violent (2)	Civil (2), Violent, Furious, Defiant (1)	Great, Young (1)	Environmental, Local (1)
Fox News	Peaceful (11), Left, Violent (10)	Violent (13), Fatal (5), Defiant, Furious (3)	Chaotic, Proper (2), Great (1)	Makeshift (3), Private, Peacefully (2)
Breitbart	Violent (18), Left-wing (14), Leftist (12)	Violent (9) Civil (2), Safe, Fewer, Close (1)	Reckless (4), Vulgar, Moral (2)	Environmental (6), White (3)

Project 1: Results

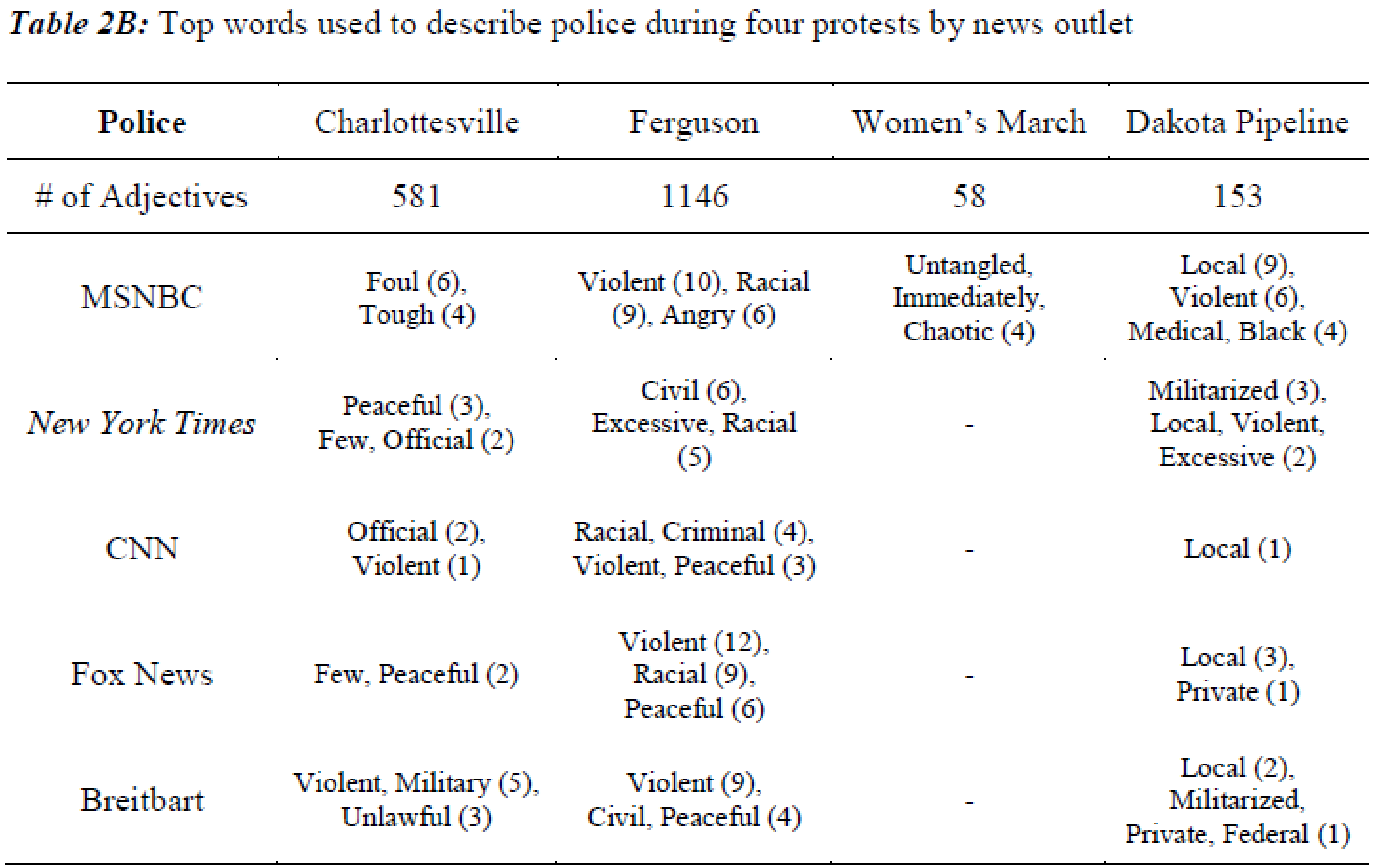


Table 2B: Top words used to describe police during four protests by news outlet

Police	Charlottesville	Ferguson	Women's March	Dakota Pipeline
# of Adjectives	581	1146	58	153
MSNBC	Foul (6), Tough (4)	Violent (10), Racial (9), Angry (6)	Untangled, Immediately, Chaotic (4)	Local (9), Violent (6), Medical, Black (4)
New York Times	Peaceful (3), Few, Official (2)	Civil (6), Excessive, Racial (5)	-	Militarized (3), Local, Violent, Excessive (2)
CNN	Official (2), Violent (1)	Racial, Criminal (4), Violent, Peaceful (3)	-	Local (1)
Fox News	Few, Peaceful (2)	Violent (12), Racial (9), Peaceful (6)	-	Local (3), Private (1)
Breitbart	Violent, Military (5), Unlawful (3)	Violent (9), Civil, Peaceful (4)	-	Local (2), Militarized, Private, Federal (1)

Lesson: Processing & Annotating Language

Dick [PROPN] is [V] wrong [ADJ]. I [PRON] like [VERB] Rose [PROPN]

Dick is wrong. I like Rose!

dick is wrong. i like rose!

dick is wrong i like rose

dick wrong like rose

dick, like, rose, wrong

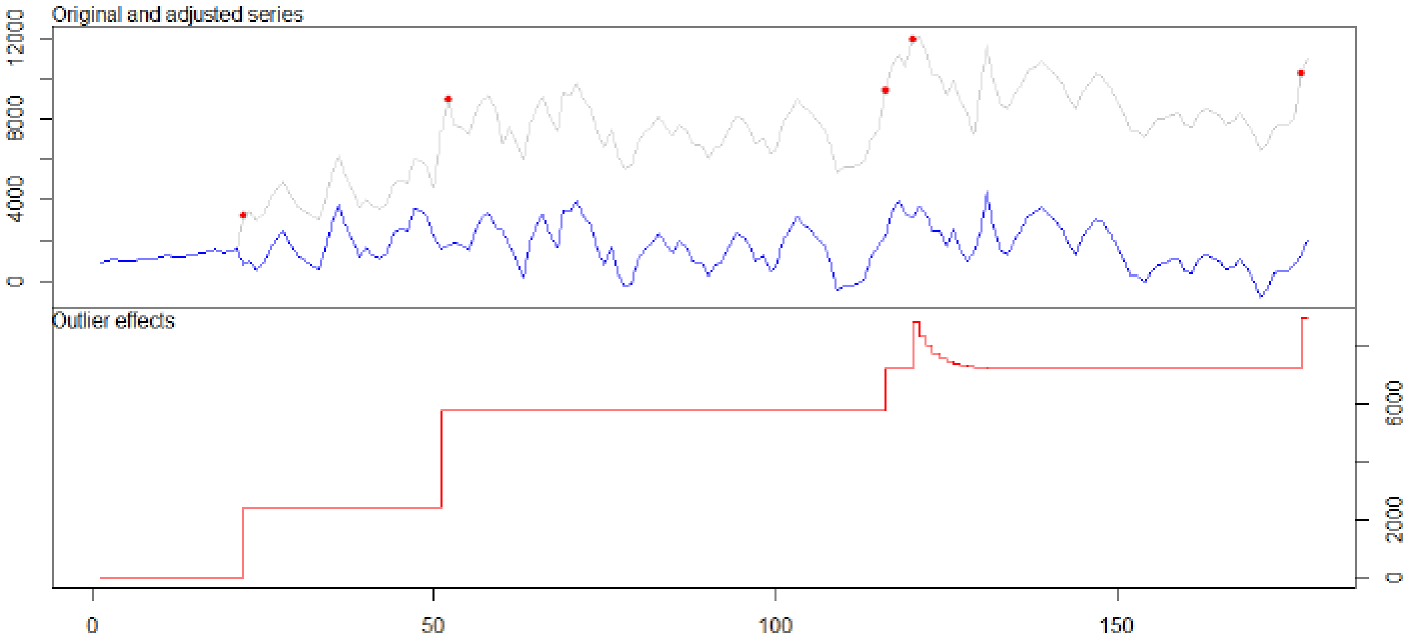
Project 2: Discourse Shifts in Media Events

- **NLP for identifying changes in discourse.**
- **Goal**: Use time series + NLP-identified discursive shifts to identify viral moments in media events (2016 U.S. presidential debate)
- **Corpus**:
  - 4,121,760 tweets
  - Tweets posted about “Clinton” or “Trump” in first 2016 debate
- **Method**: Time Series + Word Embeddings

Lukito, J., Sarma, P., Foley, J. & Abhishek, A. (2019). Using time series and natural language processing to identify viral moments in the 2016 U.S. Presidential Debate. **[To be presented at NAACL 2019]**

Project 2: Clinton Time Series Results

***Figure 4a**: Outlier Detection during Viral Moments*



Project 2: Clinton Discourse Shifts

- Construct 2 corpora

- **Corpora A**: All tweets posted 2 minutes prior to the viral moment
- **Corpora B**: All tweets posted 2 minutes after the viral moment

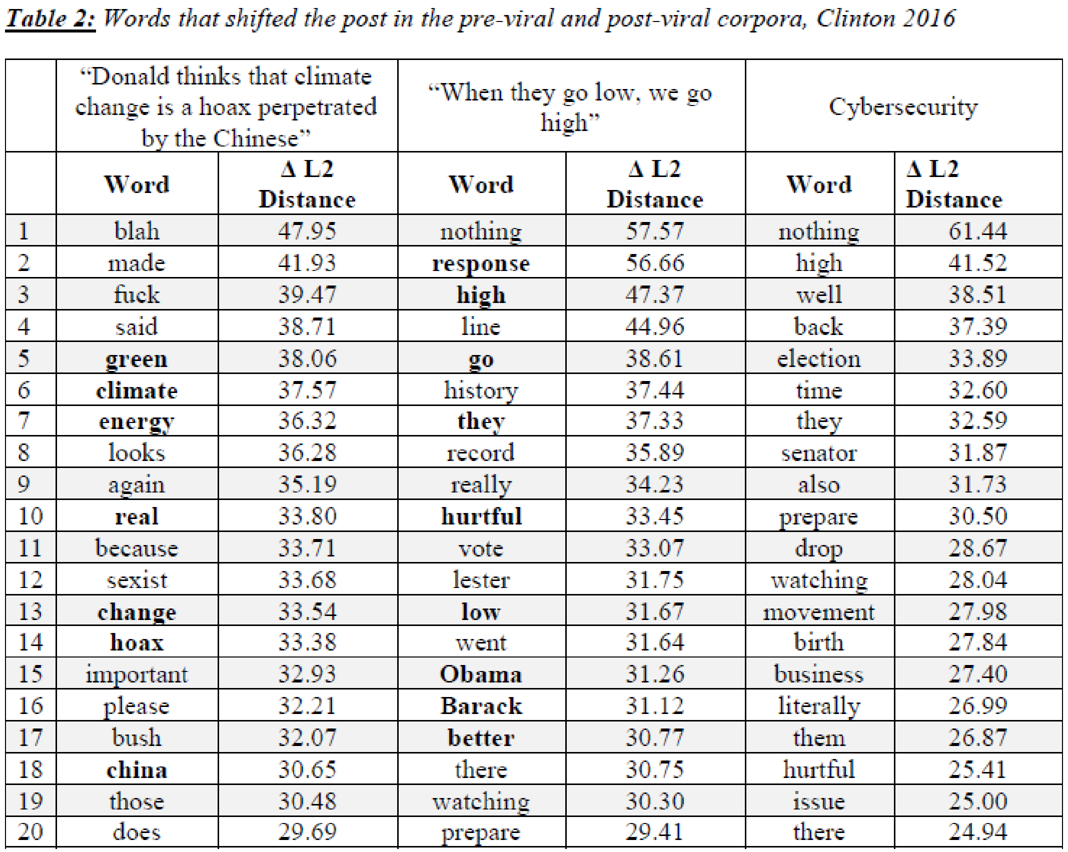


Table 2: *Words that shifted the post in the pre-viral and post-viral corpora, Clinton 2016*

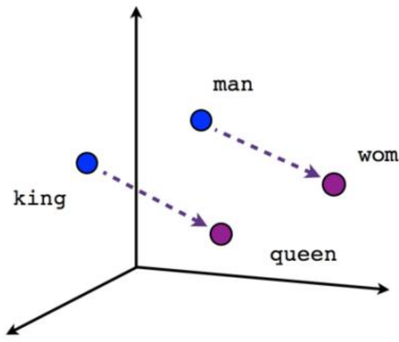
	"Donald thinks that climate change is a hoax perpetrated by the Chinese"		"When they go low, we go high"		Cybersecurity	
	Word	**Δ L2 Distance**	**Word**	**Δ L2 Distance**	**Word**	**Δ L2 Distance**
1	blah	47.95	nothing	57.57	nothing	61.44
2	made	41.93	**response**	56.66	high	41.52
3	fuck	39.47	**high**	47.37	well	38.51
4	said	38.71	line	44.96	back	37.39
5	**green**	38.06	**go**	38.61	election	33.89
6	**climate**	37.57	history	37.44	time	32.60
7	**energy**	36.32	**they**	37.33	they	32.59
8	looks	36.28	record	35.89	senator	31.87
9	again	35.19	really	34.23	also	31.73
10	**real**	33.80	**hurtful**	33.45	prepare	30.50
11	because	33.71	vote	33.07	drop	28.67
12	sexist	33.68	lester	31.75	watching	28.04
13	**change**	33.54	**low**	31.67	movement	27.98
14	**hoax**	33.38	went	31.64	birth	27.84
15	important	32.93	**Obama**	31.26	business	27.40
16	please	32.21	**Barack**	31.12	literally	26.99
17	bush	32.07	**better**	30.77	them	26.87
18	**china**	30.65	there	30.75	hurtful	25.41
19	those	30.48	watching	30.30	issue	25.00
20	does	29.69	prepare	29.41	there	24.94

Lesson: Word Embeddings

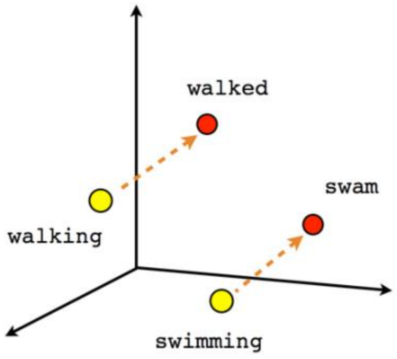
“a word is categorized by the company it keeps” (Firth, 1957)

Words expressed as vectors.

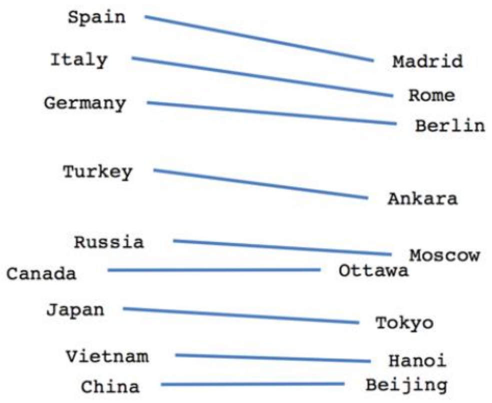
- GloVe
- word2vec
- ELMO (and BERT)



Male-Female



Verb tense



Country-Capital

Project 3: 2016 Election Topics

- **NLP for identifying topics/agendas.**
- Research Question: Did news outlets granger-cause one another to talk about political topics/agendas in the 2016 U.S. election?
- **Corpus**:
  - 125,039 news articles
  - 22 outlets
  - July 1, 2016 to November 8, 2016
- **Method**: Time Series + Structural Topic Modeling

Wells, C., Lukito, J. & Sun, Z. (2018). Three ways of looking at a media system: Attention, agenda and tone in the last months of Election 2016 **[Presented at IJPP Conference]**

Project 3: Horserace/Election Topics

1. Topic 1 - Trump sexual allegations
2. Topic 2 - (International) trade
3. Topic 3 - Rallies and protests, general
4. Topic 4 - Democrats/DNC
5. Topic 5 - Trump, Manafort & Lewandowski
6. Topic 6 - Voting / voter fraud
7. Topic 7 - Gender (and education)
8. Topic 8 - Polls
9. Topic 9 - RNC in Cleveland / GOP speeches
10. Topic 10 - [EXCLUDE] Spanish articles
11. Topic 11 - Warren & Biden
12. Topic 12 - [EXCLUDE] Noise
13. Topic 13 - Battleground/swing states
14. Topic 14 - VP picks
15. Topic 15 - Social media + anti-Semitic talk
16. Topic 16 - Clinton Email Leak
17. Topic 17 - Media appearances (mostly Fox)
18. Topic 18 - Muslim + Terrorism
19. Topic 19 - Income tax
20. Topic 20 - [EXCLUDE] Theatre articles
21. Topic 21 - Trump properties
22. Topic 22 - Clinton Foundation
23. Topic 23 - Abortion / LGBTQ + Christianity
24. Topic 24 - Labor unions
25. Topic 25 - Police violence (+BLM)
26. Topic 26 - Other elections
27. Topic 27 - Immigration
28. Topic 28 - GOP Primary candidates
29. Topic 29 - Judges and courts [SCOTUS]
30. Topic 30 - Healthcare
31. Topic 31 - Dem Primary candidates [Sanders]
32. Topic 32 - Gun rights / regulation + 2A
33. Topic 33 - Administration + Congress
34. Topic 34 - DNC + Kahn's speech
35. Topic 35 - Race relations (Black + Latinx)
36. Topic 36 - Clinton FBI Investigation
37. Topic 37 - Political advertising
38. Topic 38 - [EXCLUDE] Transcripts/short stories
39. Topic 39 – Russia
40. Topic 40 - Environmental + Climate Change
41. Topic 41 - Republican Party + Paul Ryan
42. Topic 42 - Republican primaries (Kasich + Cruz)
43. Topic 43 - Iraq, military, war
44. Topic 44 - Bush, past presidents, endorsements
45. Topic 45 - Bill Clinton Sex Scandals
46. Topic 46 - [EXCLUDE] Jokes
47. Topic 47 - [EXCLUDE] rummel + pewdiepie
48. Topic 48 - Debates
49. Topic 49 - Obama Birther Scandal
50. Topic 50 - [EXCLUDE] Theatre articles

Lesson: Value & Limitations of Methods

- Structural Topic Modeling is a multi-class, unsupervised strategy
  - Good for inductive exploration
  - Not good when looking for specific things
- There is no one method, algorithm, or model that will do everything you want with language.
- Complicated methods != best strategy
  - Choose the computational method that fits your research question.
  - Choose the computational method that captures your linguistic phenomenon.

Goal → Strategy

- **Authorship** → stylometrics
- **Topics/Issues/Agendas** → Unsupervised Machine Learning
- **Stakeholder Analysis** → Entity Recognition, Anaphoric Resolution
- **Framing** → Dictionaries, Supervised Machine Learning, Syntax Annotation
  - **Valence/Sentiment** → Sentiment Analysis
- **Corpus/Genre Comparisons** → Syntax Annotation, Word Embeddings

PROCESING IS EQUALLY IMPORTANT: How you process your data should be informed by (1) your RQ and (2) your method

New Avenues of Research

- Mixed Methods
    - Surveys of Journalists + Computational Analyses
    - Computational Analyses + Experiments
- Non-English Computational Linguistics
    - The need to service other languages
    - Computational morphology
- Tools for field-specific language analysis
    - medium : message :: communication : language

Natural Language Processing with ***Frames, Discourse, and Agendas***

Josephine Lukito

https://github.com/jlukito/computational-comm-rg-guide

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON