INTRO TO TEXT ANALYSIS IN R

Sze Yuh Nina Wang

SIPS 2020

SZE YUH NINA WANG
SZEYUHWANG@GMAIL.COM
SZEYUHNINAWANG.COM

Overview

- Zoom basics
- Intro: Text as data
- Language analysis methods
  - Dictionary methods
  - Word embeddings
  - Topic modelling
- Tutorial
  - Sentiment analysis
  - Topic model
- General principles in text analysis
- Additional resources

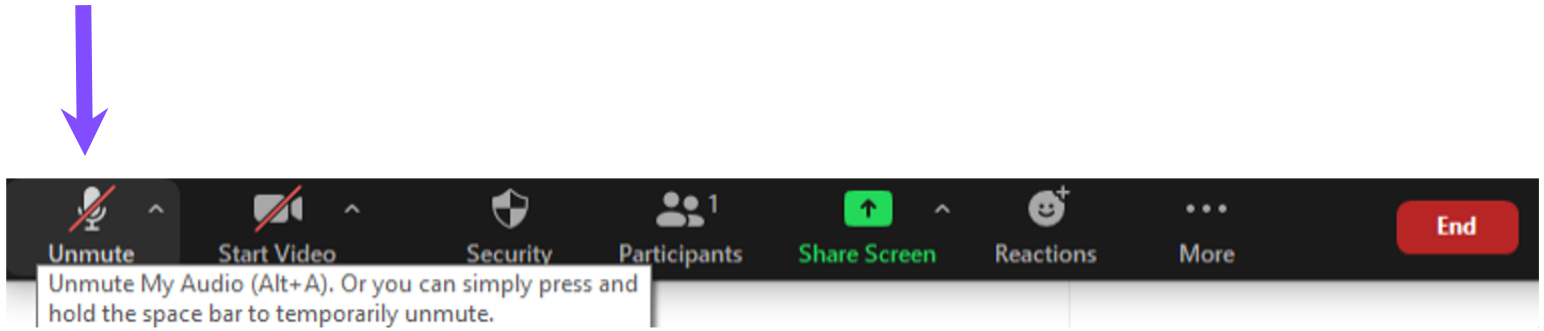
Sorry if we’re loud!

Zoom tutorial for SIPS2020

A good online conference experience

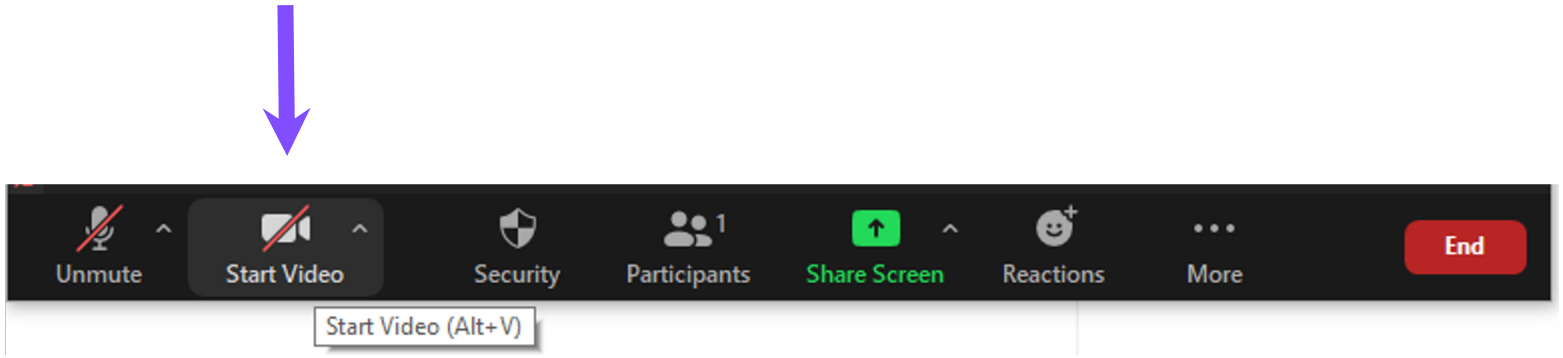
Audio Etiquette

- Keep your mic muted when not talking
- Unmute when you are ready to speak



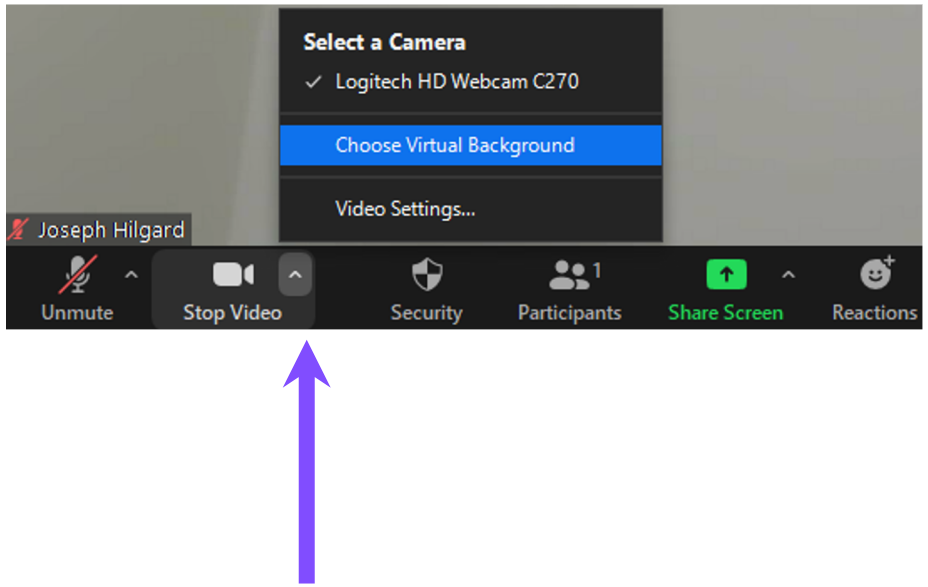
Video Etiquette

- Nobody is required to share their video
  - (but it’s nice to see each other!)



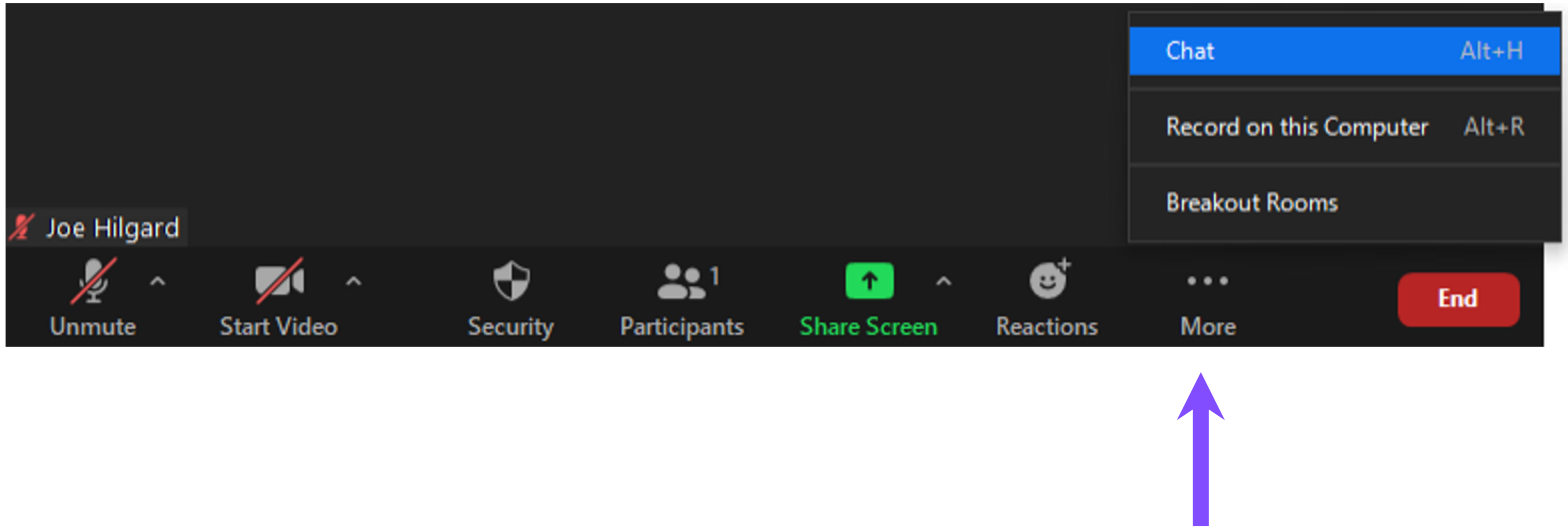
Video Etiquette

- A virtual background can be useful to keep others from seeing your personal living space
  - (May not be available on some computer hardware)



Viewing the chat panel

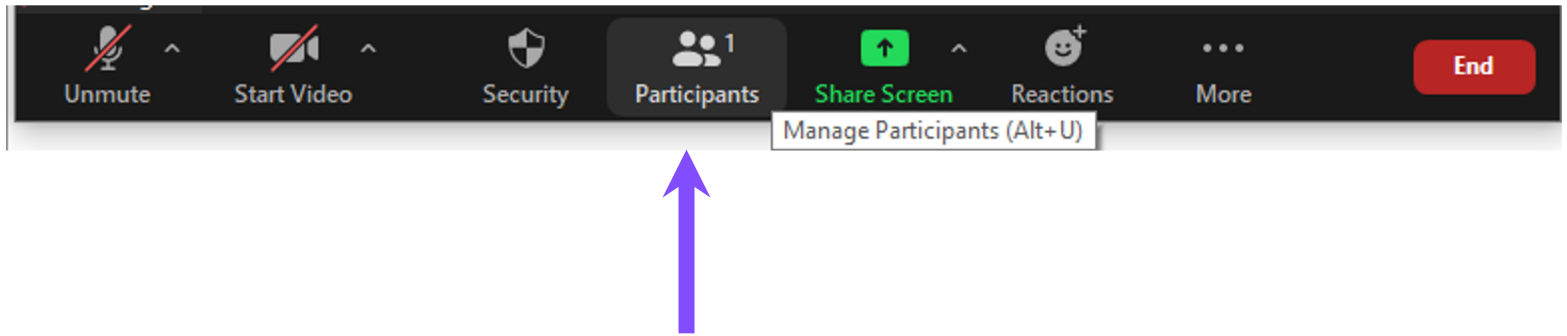
The chat channel can be used for public discussion and private whispers.



Be aware that chat history can be saved offline.

Viewing the participants panel

The participants panel lists everyone in the room and has tools for interaction and management.



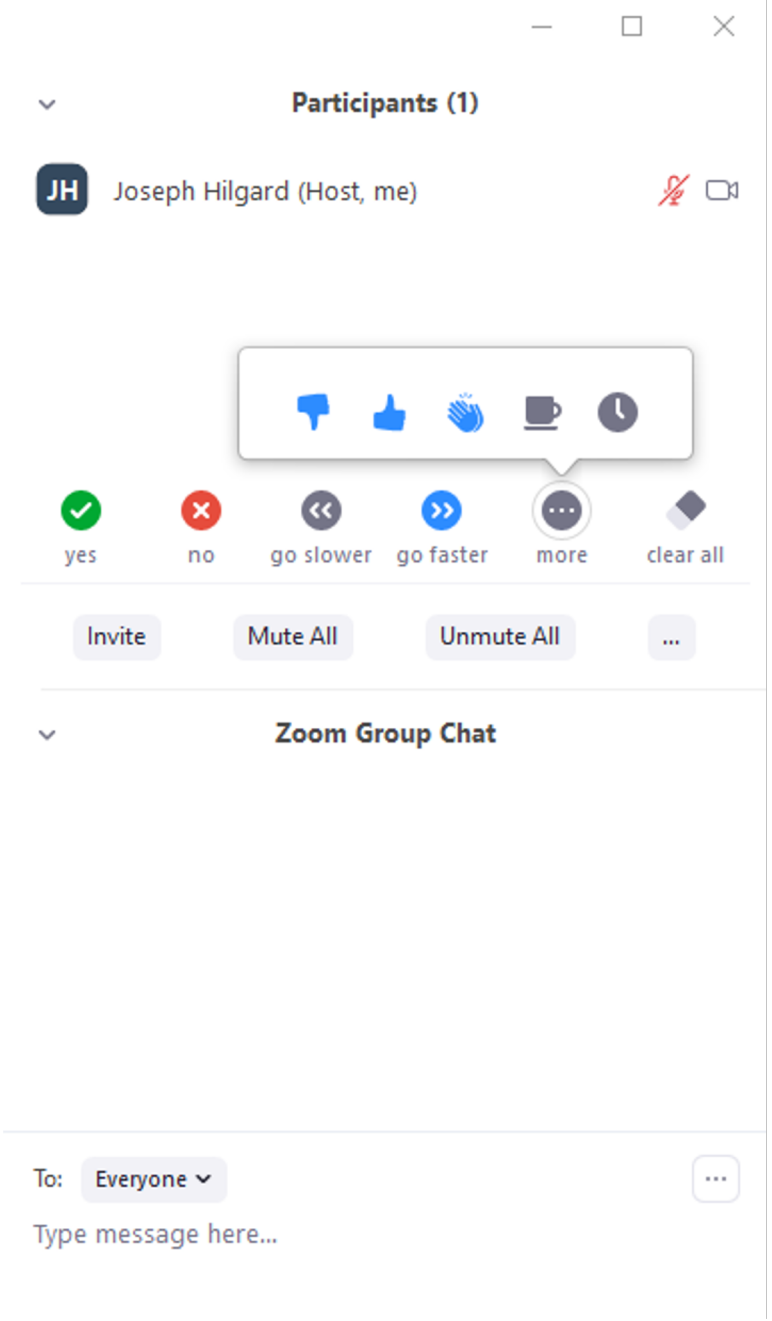
In the participants panel:

Participants can:

- Raise their hand to get the host's attention
- Use nonverbal reactions like "yes", "no", "go slower", "need a break"

Hosts & Co-hosts can:

- Admit registrants from the waiting room
- Turn off participants' video or audio
- Remove disruptive participants



Technical support

Join the SIPS 2020 Slack (here)

Post your question to the Tech Q&A channel (here)

Post content questions on the session Slack channel (here)

Moderator/co-host: Sue Song

Support from organizers

Feel free to use the “private chat” function to message session organizers with any questions or concerns. Tell them:

- If you prefer not to be recorded
- If you experience any difficulties with other participants

DOWNLOAD TUTORIAL FILES

https://github.com/szeyuhninawang/sips-text-analysis

Text as data

- Manual coding → automated analysis
- New and increasingly important sources of text data (e.g., social media)
- New methods for automated analysis
- What kind of info can we extract from text?
  - Emotion
  - Personality
  - Moral sentiment

Some sources of text data ...

- News media
- Social media
- Political speeches
- Blog posts
- Open-response survey data
- Interview transcripts

LANGUAGE ANALYSIS METHODS

Dictionary methods
Word embeddings
Topic modelling

Dictionary methods

- AKA word counting
- LIWC (Linguistic Inquiry and Word Count; Pennebaker, Booth, Boyd, & Francis, 2015)

Dictionary methods

- Use a dictionary of words that represent your concept of interest
- Count number of times they appear in each document
- E.g. a dictionary for negative emotion might include words like “angry”, “dislike”, “hate”
- Important to test robustness of dictionaries

Dictionary methods: applications

- Moral foundations
- Personality
- Analytic language of politicians (Jordan & Pennebaker, 2017)

- Review papers: Tausczik & Pennebaker (2010); Boyd (2017)

Word embeddings

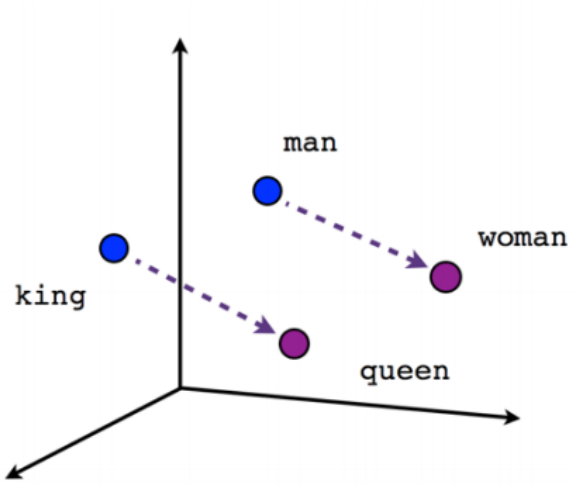
“The meaning of a word is its use in the language”

– Ludwig Wittgenstein

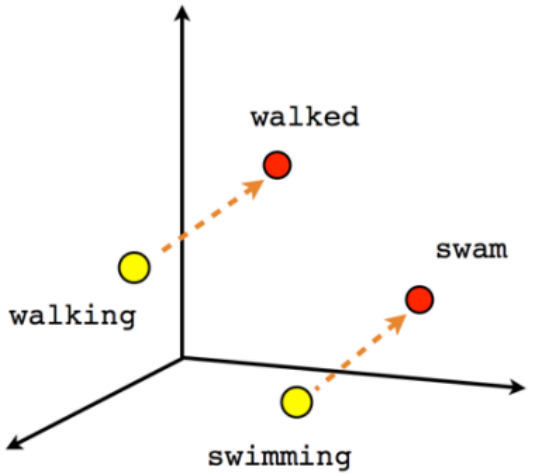
Word embeddings

- Representing meaning of words based on the context in which they appear
  - The **oculist** performed an eye exam.
  - The **eye-doctor** performed an eye exam.
- Distributional hypothesis: words that occur in similar contexts tend to have similar meanings
- Words as vectors in a multidimensional space
  - Words that are close together → semantically similar
  - Words that are far apart → semantically distant

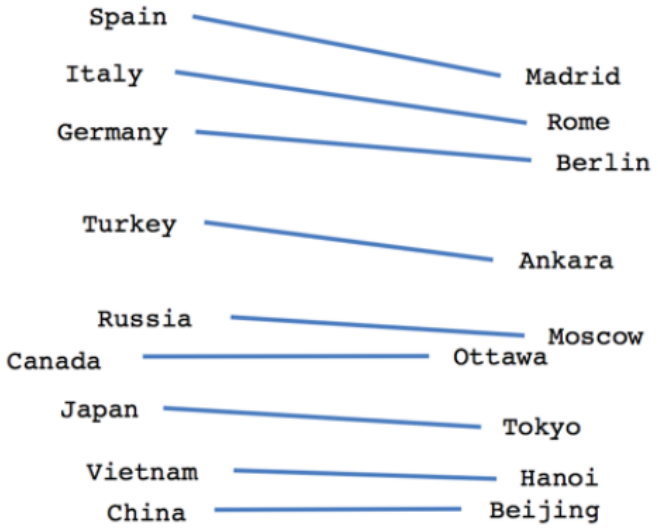
Word embeddings



Male-Female

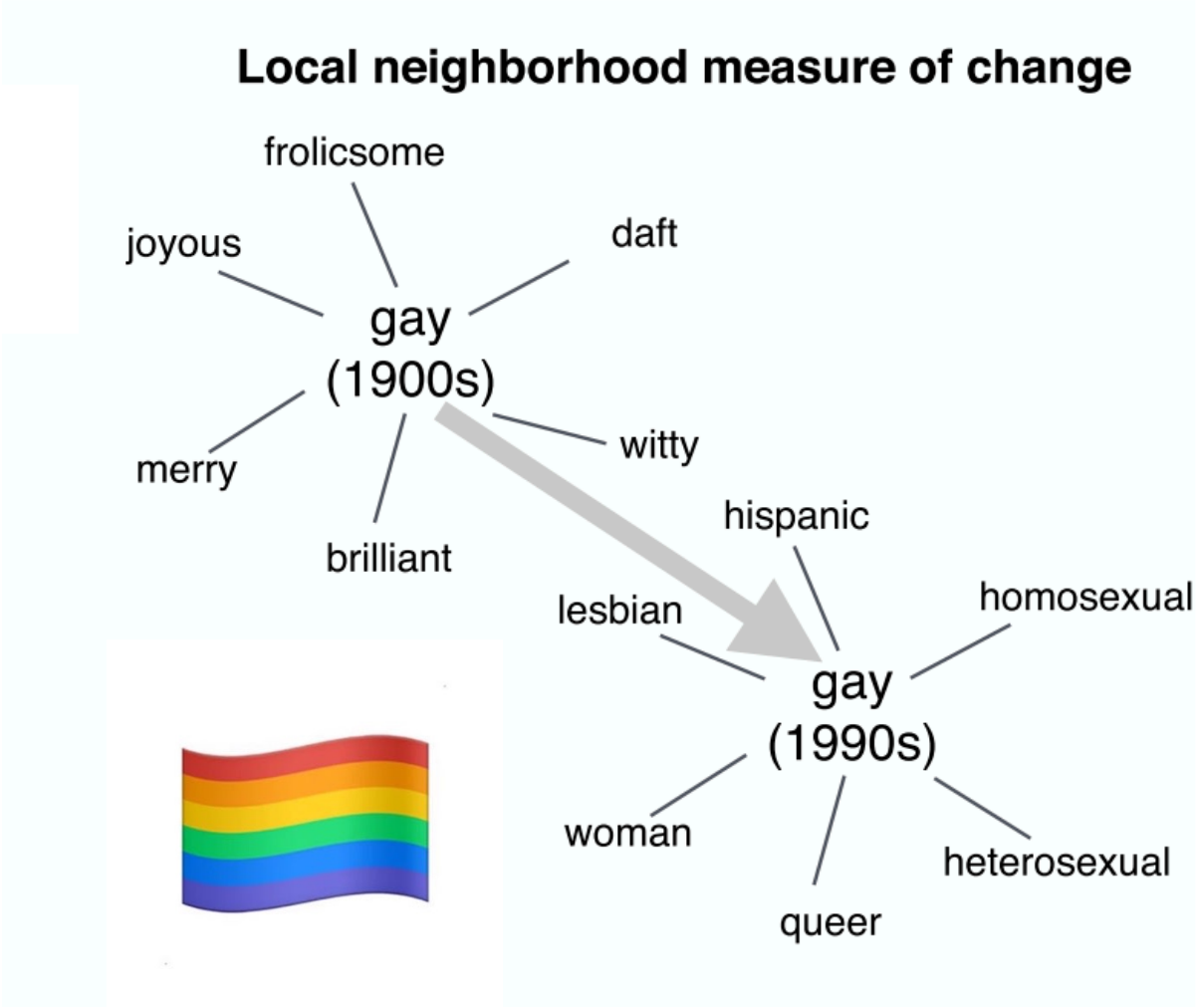


Verb tense



Country-Capital

Word embeddings



Hamilton, Leskovec, & Jurafsky, 2016

Word embeddings

Table 2. Top adjectives associated with women in 1910, 1950, and 1990 by relative norm difference in the COHA embedding

1910	1950	1990
Charming	Delicate	Maternal
Placid	Sweet	Morbid
Delicate	Charming	Artificial
Passionate	Transparent	Physical
Sweet	Placid	Caring
Dreamy	Childish	Emotional
Indulgent	Soft	Protective
Playful	Colorless	Attractive
Mellow	Tasteless	Soft
Sentimental	Agreeable	Tidy

Garg, Schiebinger, Jurafsky, & Zou (2018)

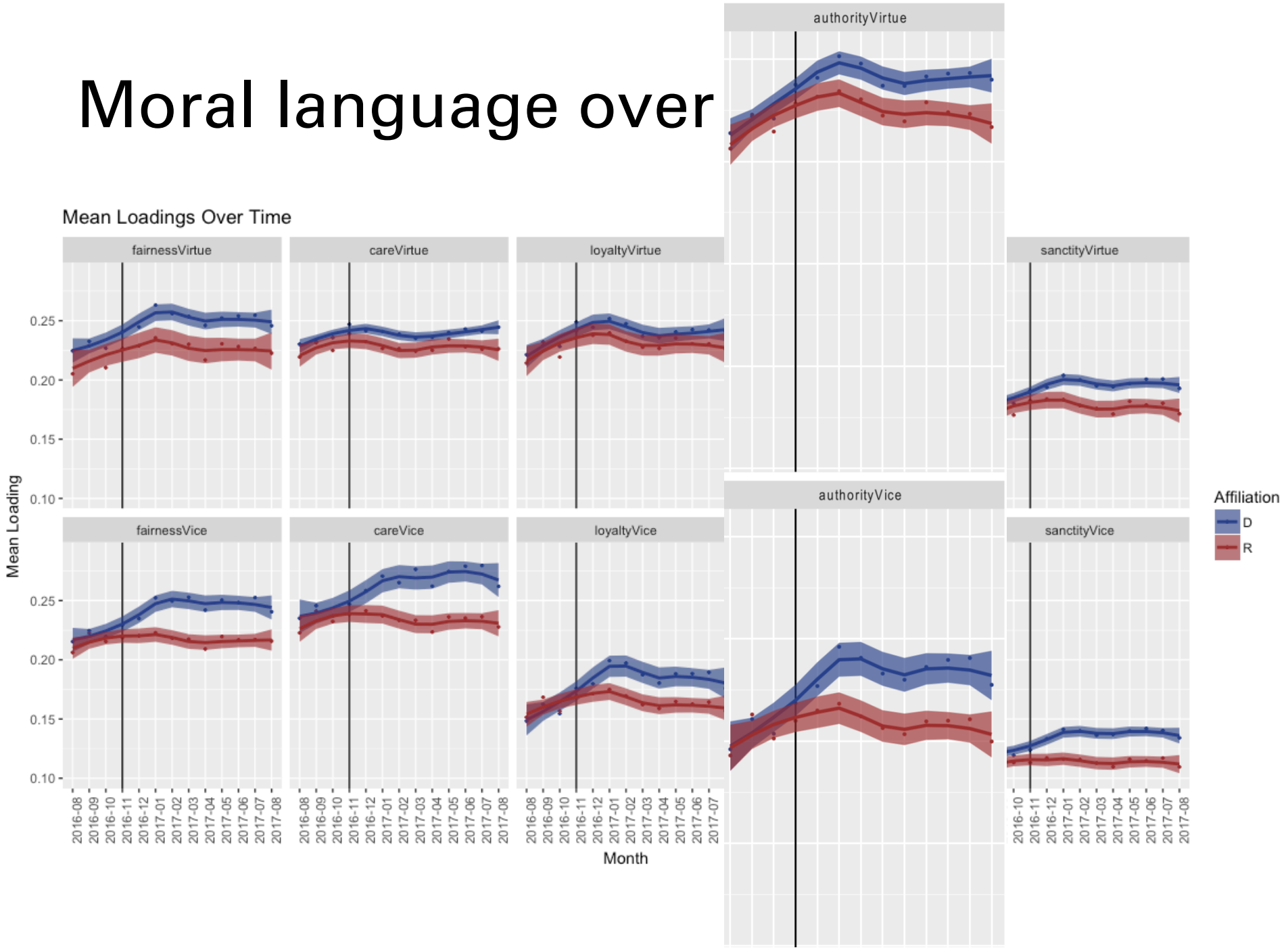
Word embeddings: applications

- Moral language
- Predicting personality from social media posts (Arnoux, Xu, Boyette, Mahmud, Akkiraju, & Sinha, 2017)
- Tracking stereotypes and social attitudes over time (Garg, Schiebinger, Jurafsky, & Zou, 2018)

Word embeddings: example

- Measuring the moral rhetoric of politicians using Distributed Dictionary Representations (Garten, Hoover, Johnson, Boghrati, Iskiwitch, & Dehghani, 2018)
- Use seed words from Moral Foundations Dictionary
- Take cosine similarity (distance) between
  1. Vector representing concept of interest (moral foundation)
  2. Vector representing document
- Output: “moral loading” of how strongly each document is expressing each moral foundation

Moral language over



Topic modelling

- Discovers themes/topics within a collection of documents and assigns each document a probability of belonging to each theme
- There are words that identify a topic, and will tend to occur frequently and uniquely

Topic modelling

”dog”, “bone”

Unique to topic

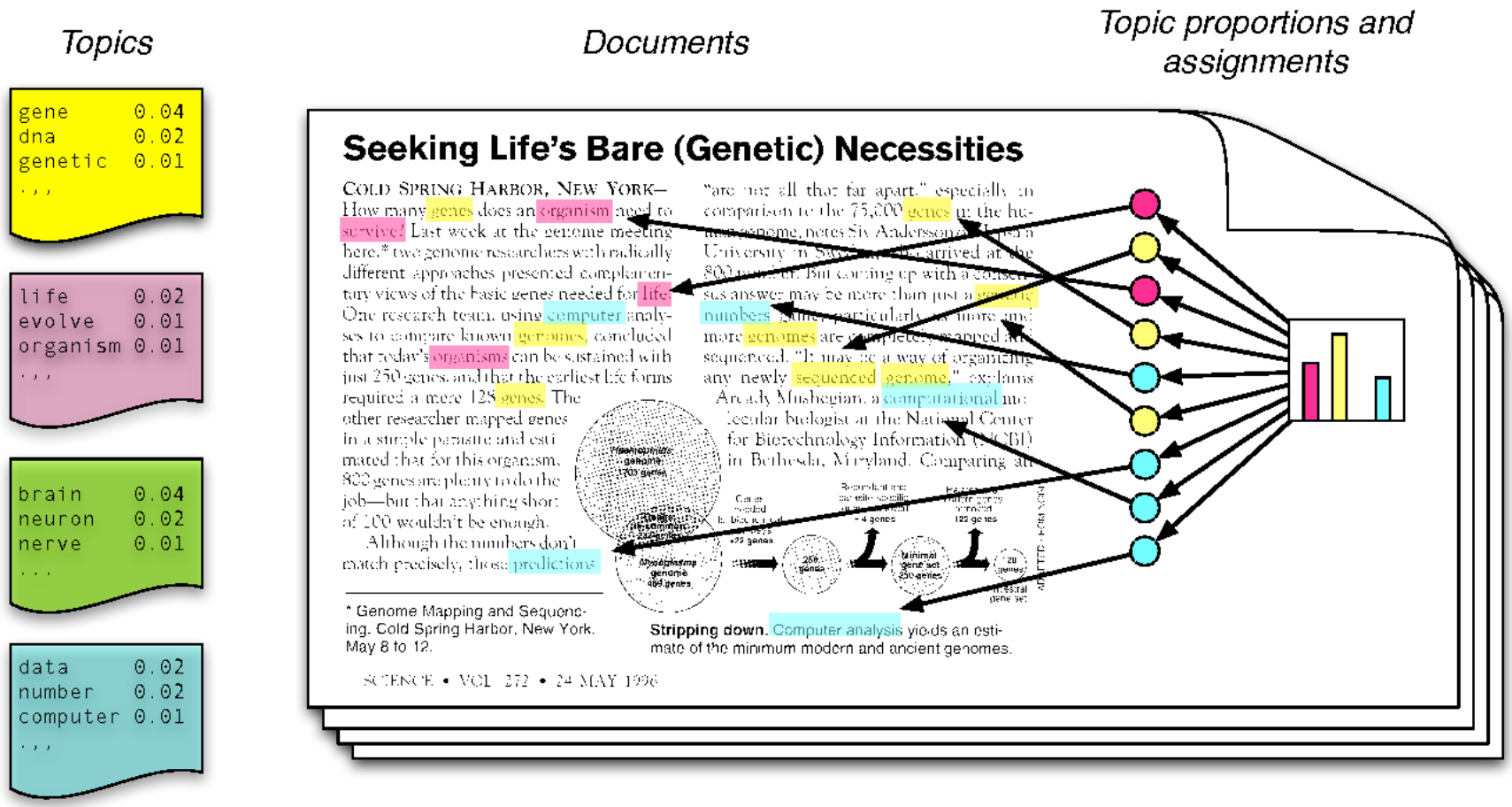
“the”, “is”

Common to both topics

“cat”, “meow”

Unique to topic

Topic modelling



Topic modelling

- Structural topic models
  - Allow for covariates to be added to the model that affect the **prevalence** and **content** of topics
- R package: stm

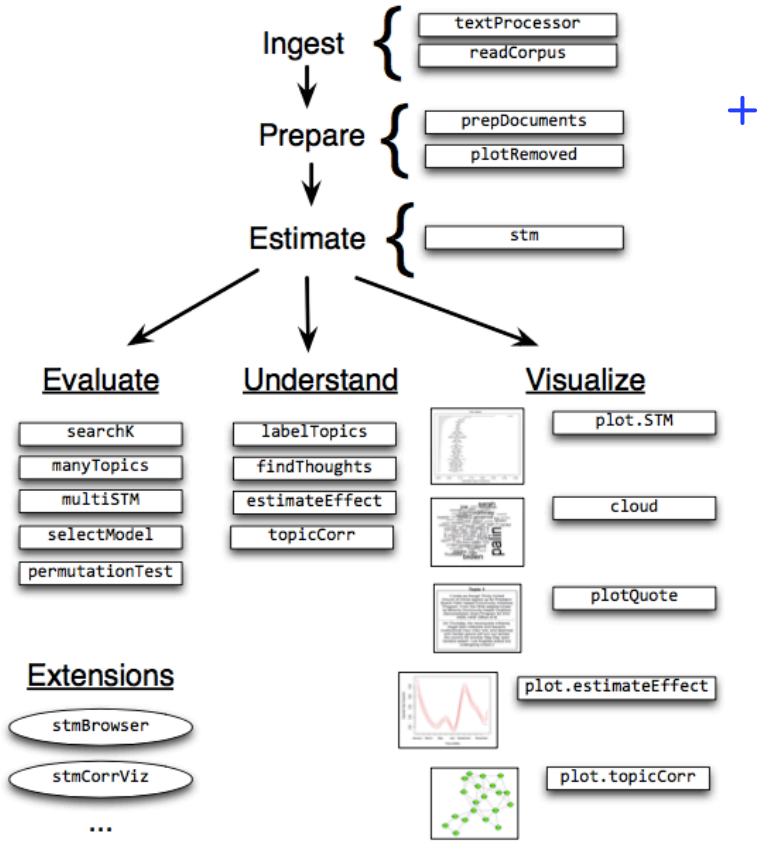


Figure 2: Heuristic description of **stm** package features.

Topic modelling: applications

- Open-ended survey responses (Roberts et al., 2014)
- Analyzing couples therapy transcripts (Atkins et al., 2012) and mental health support group posts (Carron-Arthur, Reynolds, Bennett, Bennett, & Griffiths, 2016)
- Political issues (Barberá, Casas, Nagler, Egan, Bonneau, Jost, & Tucker, 2019)

QUESTIONS?

Tutorial

- Sentiment analysis and topic modelling with political blog posts
- Dataset: ~13000 posts from 6 political blogs from 2008 (collected by Eisenstein & Xing, 2010)

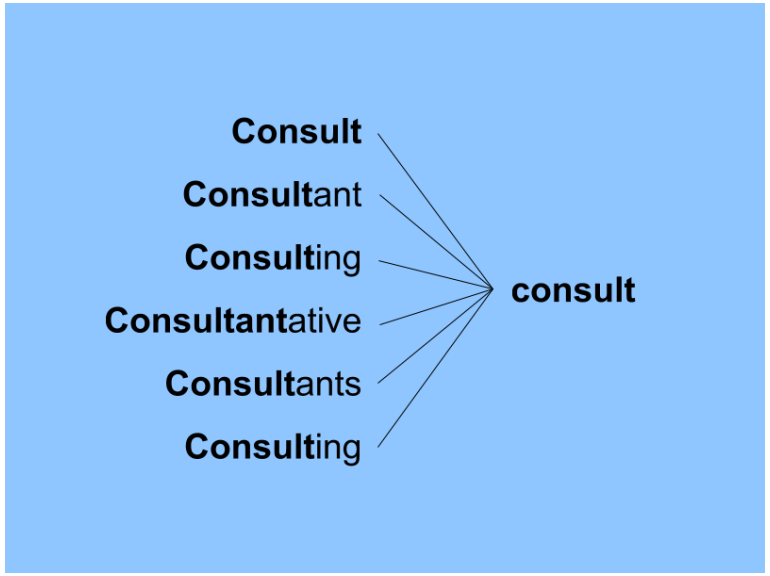
https://github.com/szeyuhninawang/sips-text-analysis

Text analysis: General principles

- Preprocessing
- Validation

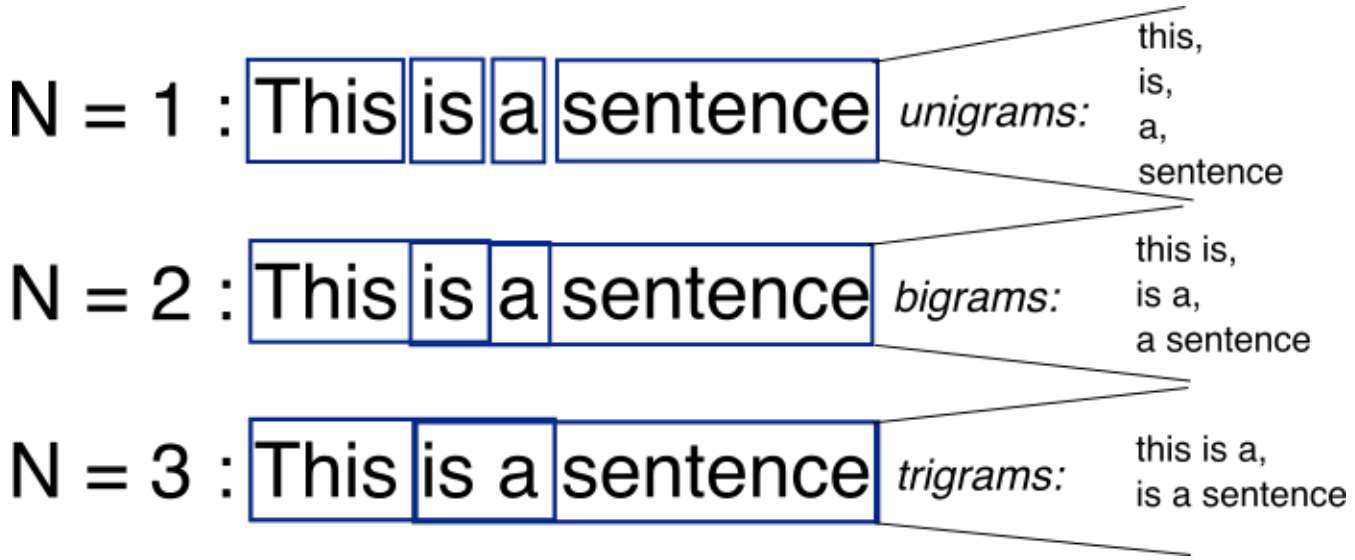
Common preprocessing steps

- Removing **stop words**
  - Common and frequently occurring words that don't contain much semantic info (e.g. "and", "is", "the")
- Stemming



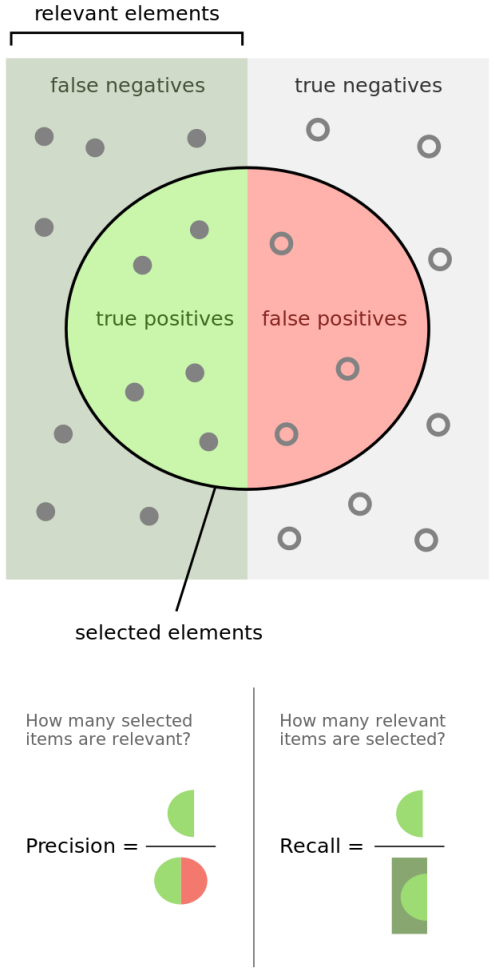
Common preprocessing steps

- N-grams



Validation

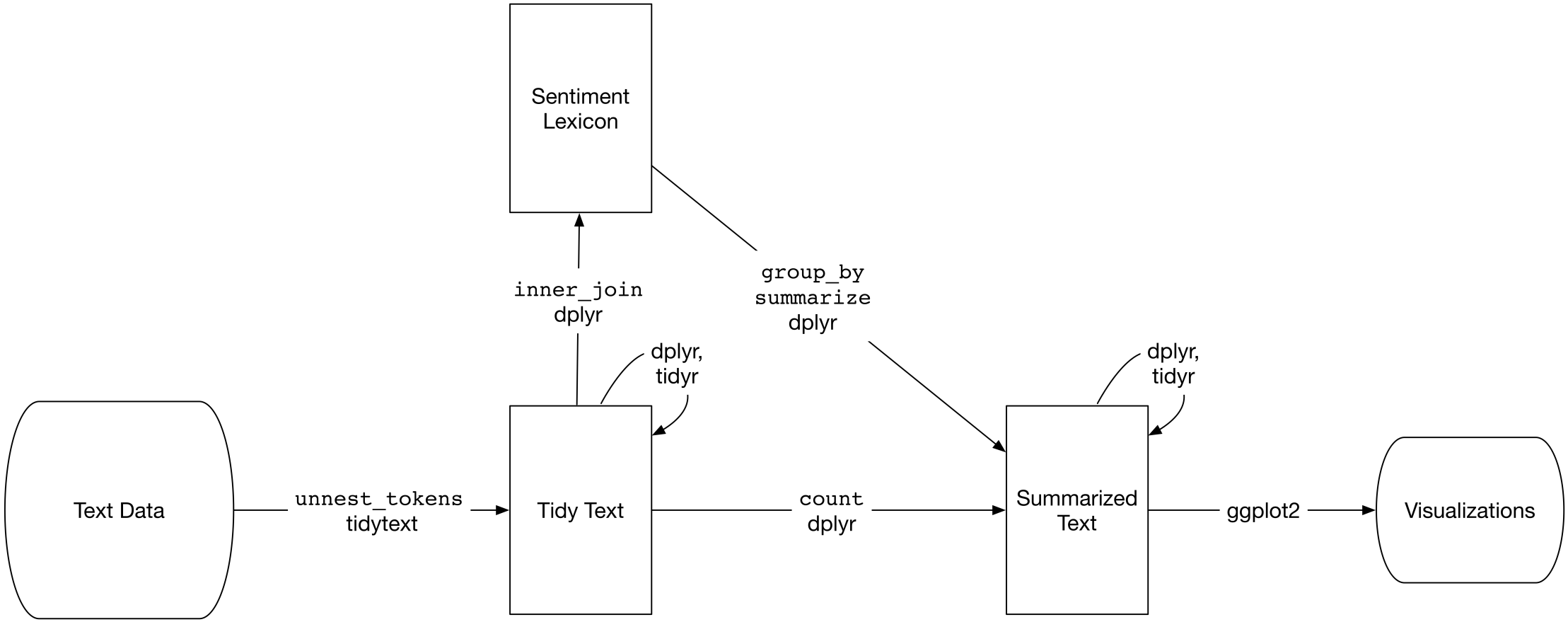
- Important to make sure these automated methods are performing well
- Compare to human coders
  - E.g. for topic modelling: have coders evaluate top words/documents for each topic (Chang, Boyd-Graber, Gerrish, Wang, & Blei, 2009)
- Split into train/test sets
  - E.g., if you decide to edit dictionary/stop word lists, or change pre-processing steps
- Compare methods



Additional resources

- Useful R packages:
  - tidytext
  - stm (Structural Topic Models)
  - quanteda
  - text2vec
  - spacyR
  - topicmodels
  - streamR and twitteR (access to Twitter API via R)

Tidytext pipeline



ADDITIONAL RESOURCES

https://szeyuhninawang.com/text-analysis-resources

THANK YOU!