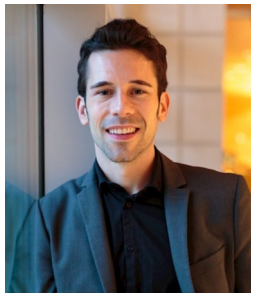
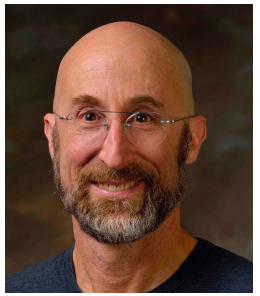
Measuring Regional Variation in Culture Through Embedding-Based Lexica



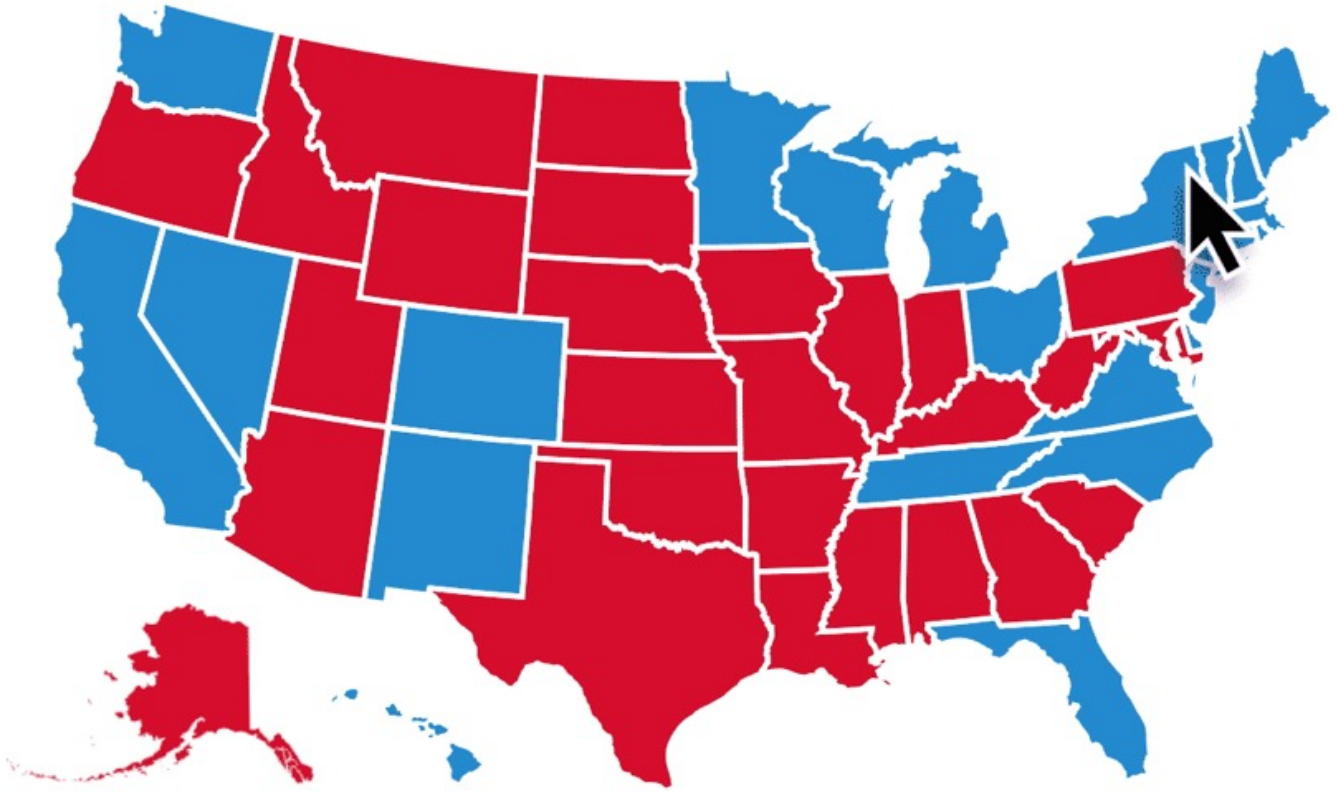




Shreya Havaldar, Salvatore Giorgi, Thomas Talhelm, Sharath Chandra Guntuku, Lyle Ungar

University of Pennsylvania

What is cultural variation?

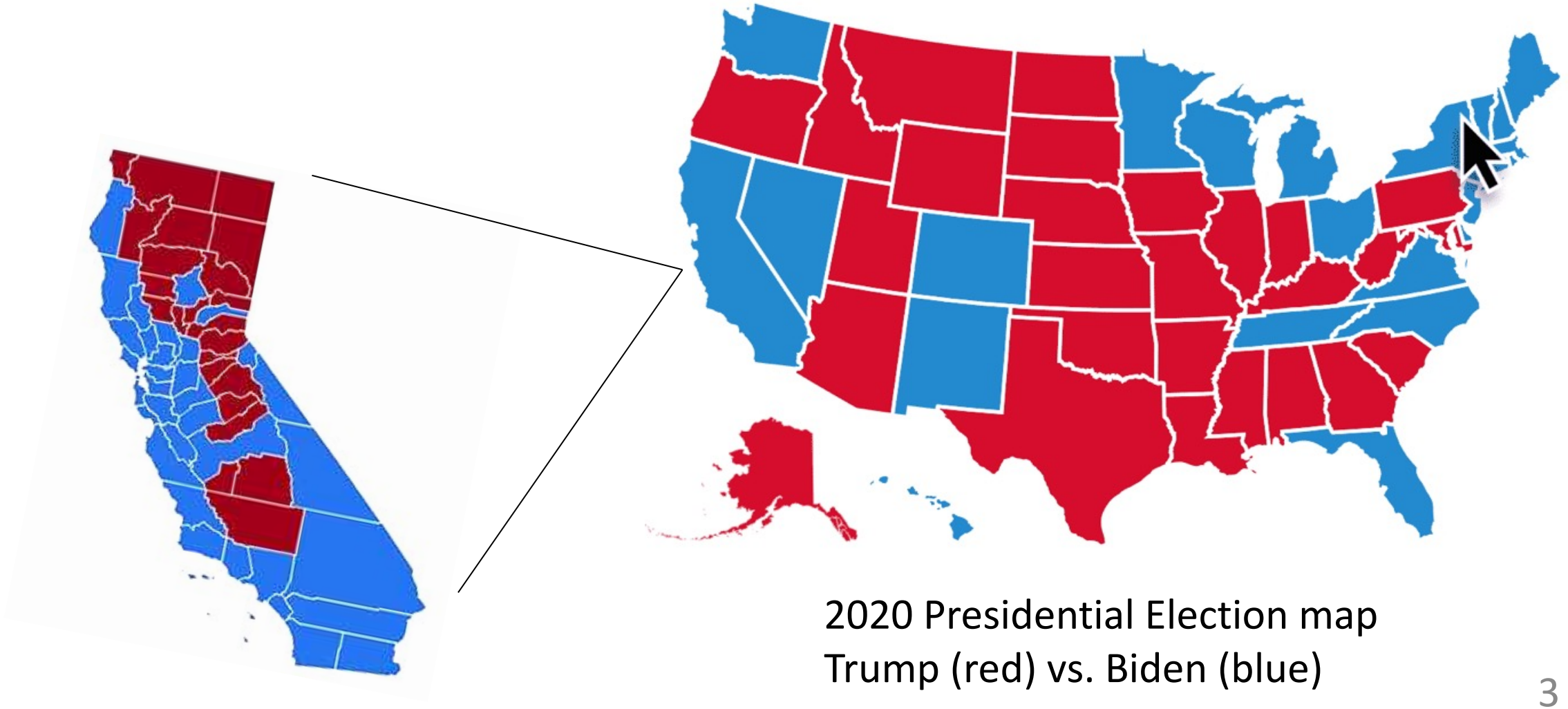


2020 Presidential Election map
Trump (red) vs. Biden (blue)

The differences among individuals that exist because they have acquired different behavior as a result of some form of social learning.

- Cambridge Dictionary

Cultural variation exists both *between* regions and *within* regions

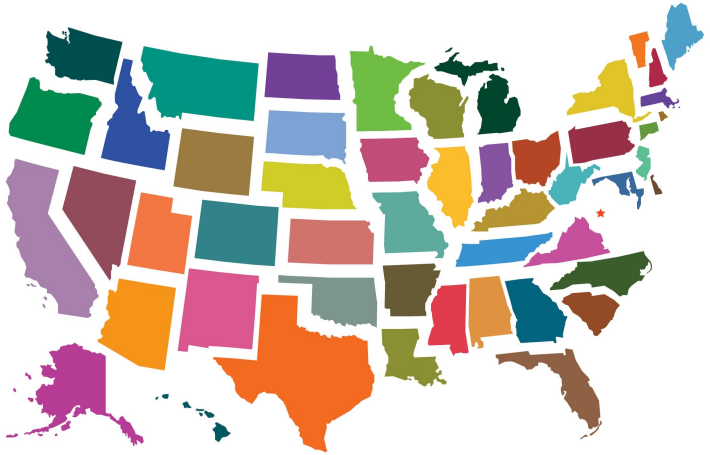


2020 Presidential Election map
Trump (red) vs. Biden (blue)

Why do we want to measure regional variation in culture?

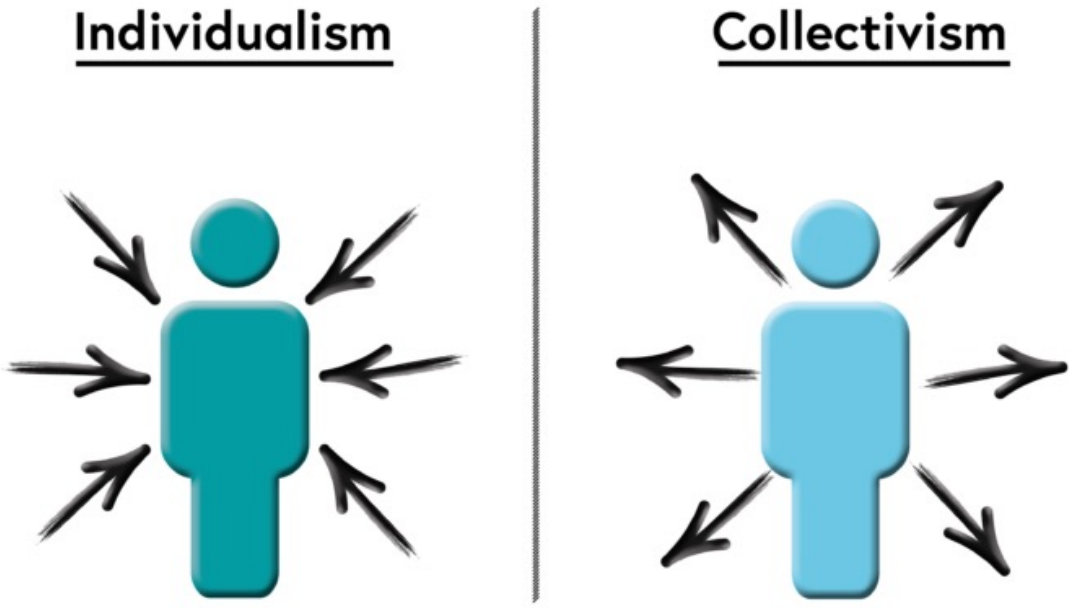
- We can better understand **people and communities**

- We can study how cultural constructs inform **behaviors** (voting, spending habits, charitable giving, etc.)



Individualism vs. Collectivism

Collectivism stresses the importance of the **community**, while **individualism** is focused on the rights and concerns of **each person**.



Are you more individualistic or collectivist?

Do you agree with the following statements?

1. We should keep our aging parents with us at home
2. Children should be taught to place duty before pleasure
3. You should loan a good friend money with no questions asked

Are you more individualistic or collectivist?

Do you agree with the following statements?

1. We should keep our aging parents with us at home
2. Children should be taught to place duty before pleasure
3. You should loan a good friend money with no questions asked

"Yes" → You are likely more **collectivist** than **individualist**

Current methods of measuring cultural constructs are not scalable

Non-Computational Methods

1. Questionnaires/Surveys
   - Too much overhead for researchers
   - We don't want to use a small sample of people when measuring **regional variation**

Idea: We can use peoples' language to measure cultural constructs

Current methods of measuring cultural constructs are not scalable

Computational Methods

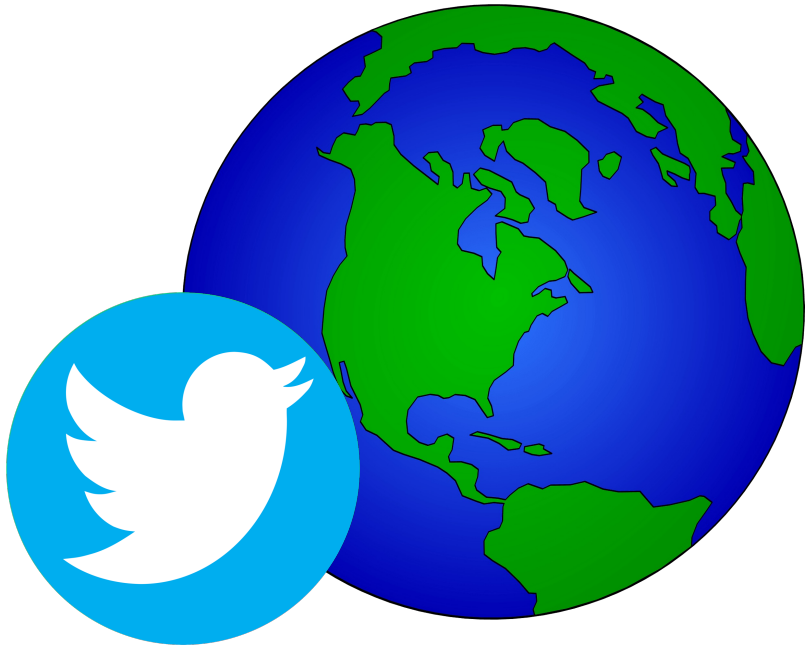
1. Labelling data is expensive
   - Labeling 1 billion Tweets using GPT-4 would cost ~$300,000
2. Prediction takes time
   - Running 1 billion Tweets through a fine-tuned LM would take ~1 GPU year
3. Not all utterances are relevant

This is not a classic ML Problem!

Embedding-based lexica provide a **scalable** way to measure regional variation in culture

We use word embeddings to generate sets of words (lexica) that characterize a cultural construct

- **No need to label data**
- **Computationally inexpensive**
- **Can utilize a massive dataset**



Using our generated lexica, we can analyze ~1.5 billion Tweets from 6 million users in 20 minutes!

Method Overview

1. Curated seed words from a domain expert

2. Embedding-based expansion
   a) Synonym expansion
   b) Concept expansion

3. Lexica Purification
   a) Frequency-based purification
   b) Correlation-based purification

We will apply this method to measure variation in individualism/collectivism across the United States

Curated seed words to measure individualism and collectivism

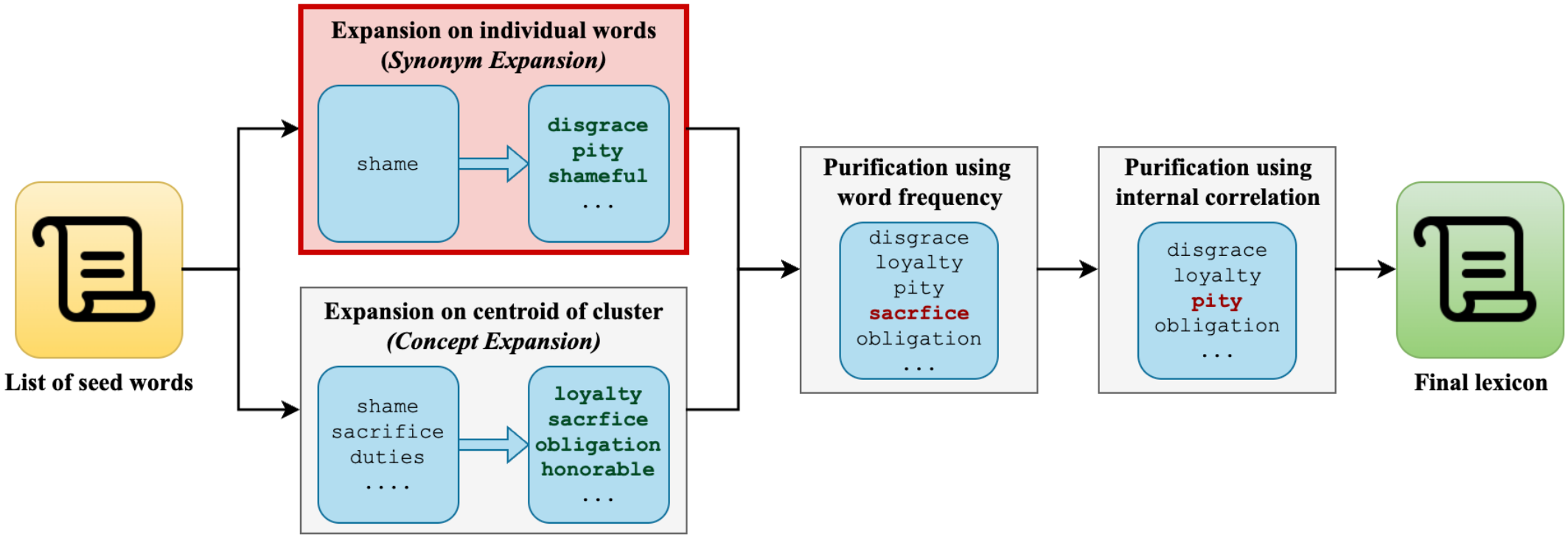
Humanity
Worldwide
Global
Equity
Cooperation
Identity
Guilt
Diversity

Individualism seed words

Duties
Responsibilities
Community
Sacrifice
Shame
Rules
Honor
Obedience

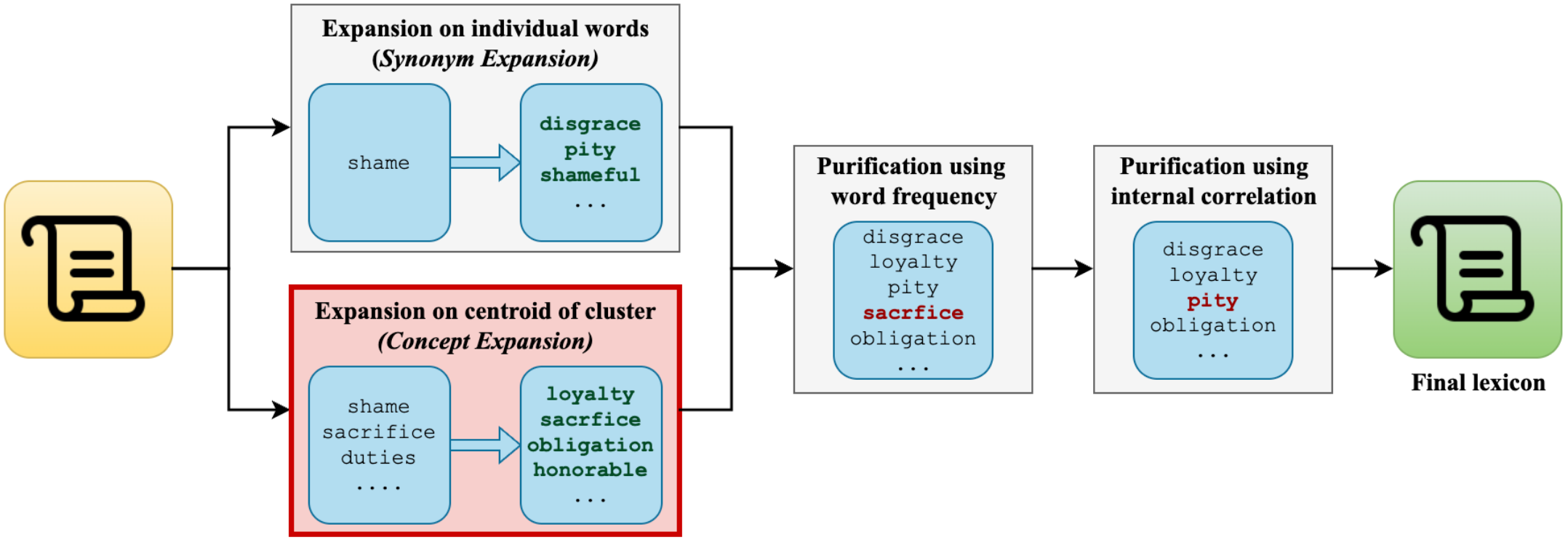
Collectivism seed words

Synonym Expansion



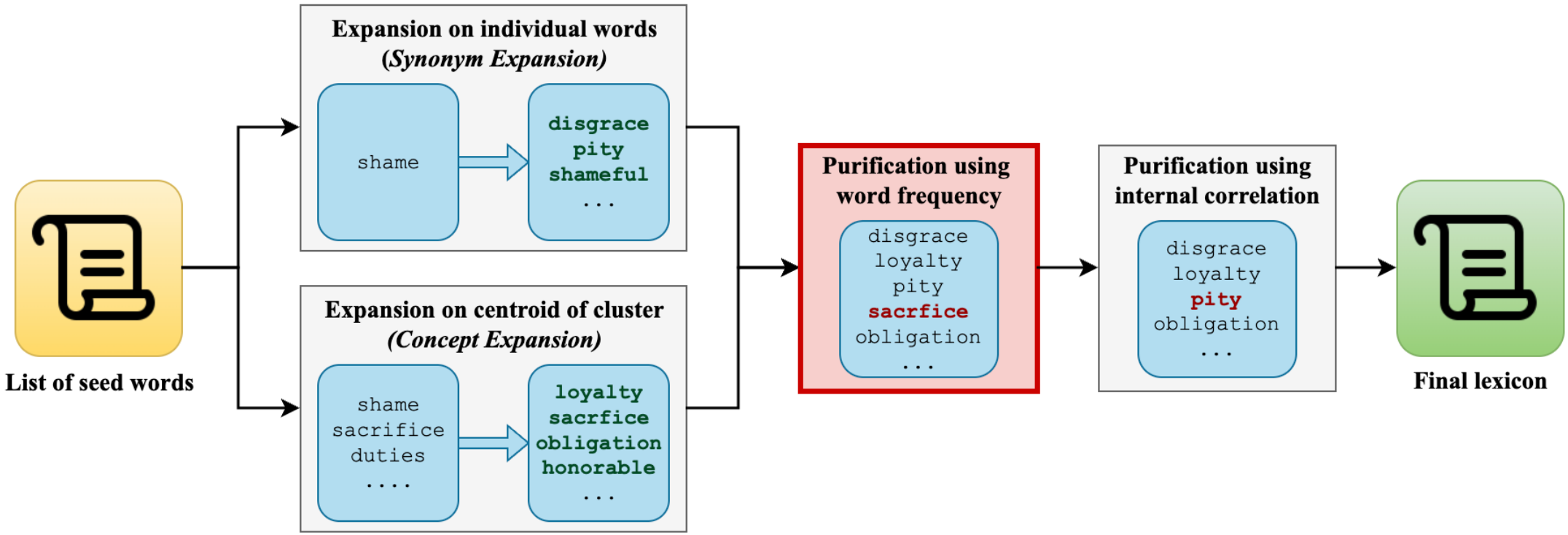
Find words closest to each individual word

Concept Expansion



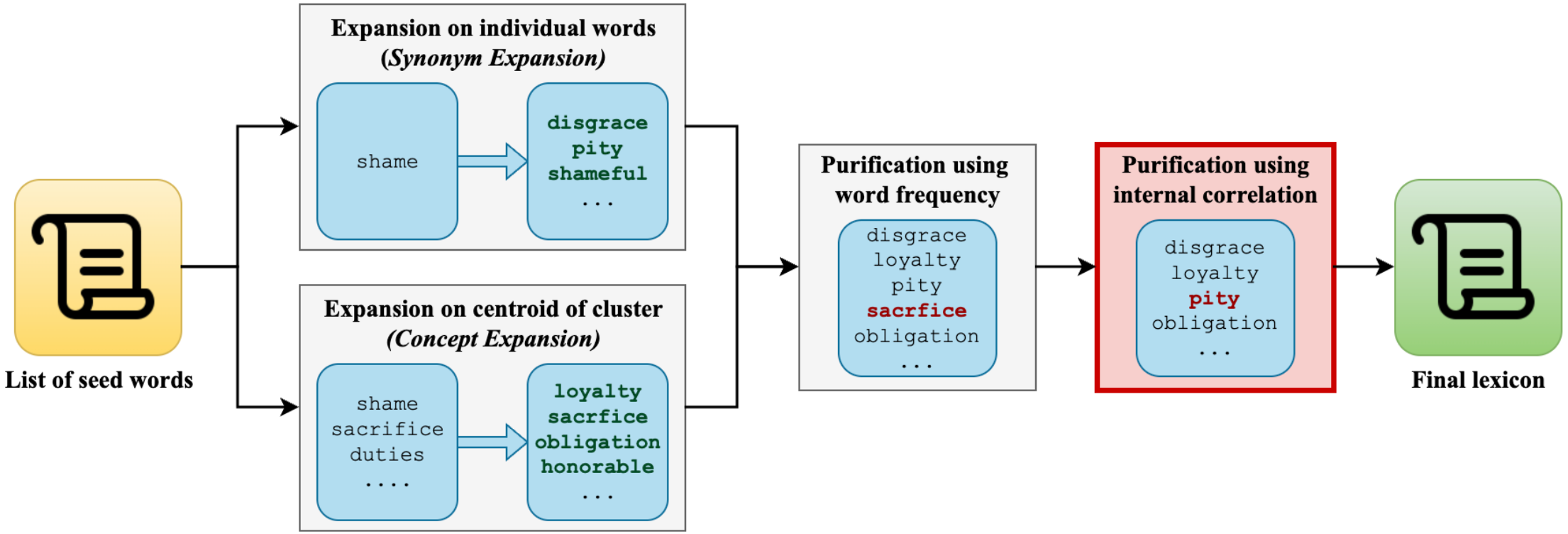
Find words closest to each category's centroid

Frequency-based purification



Remove rare words

Correlation-based purification

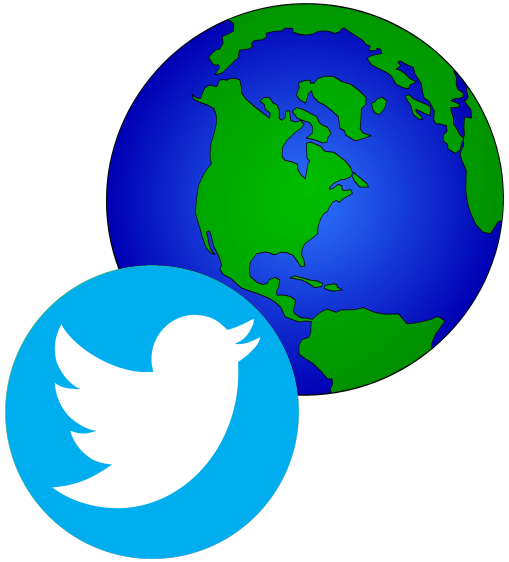


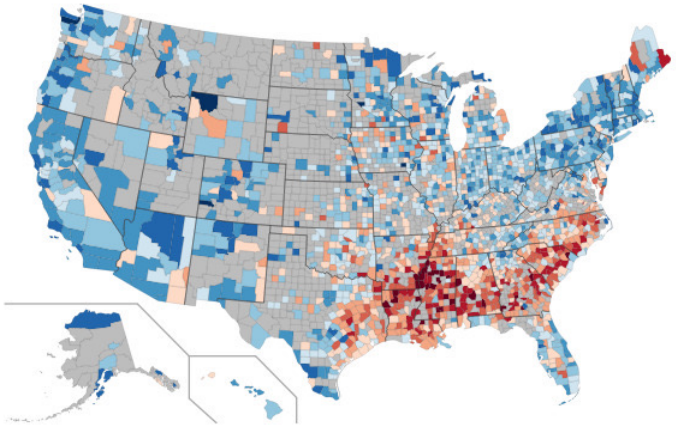
Remove words that do not co-occur with their category

10 seed words → ~200 word lexica

Individualism Lexica

Collectivism Lexica



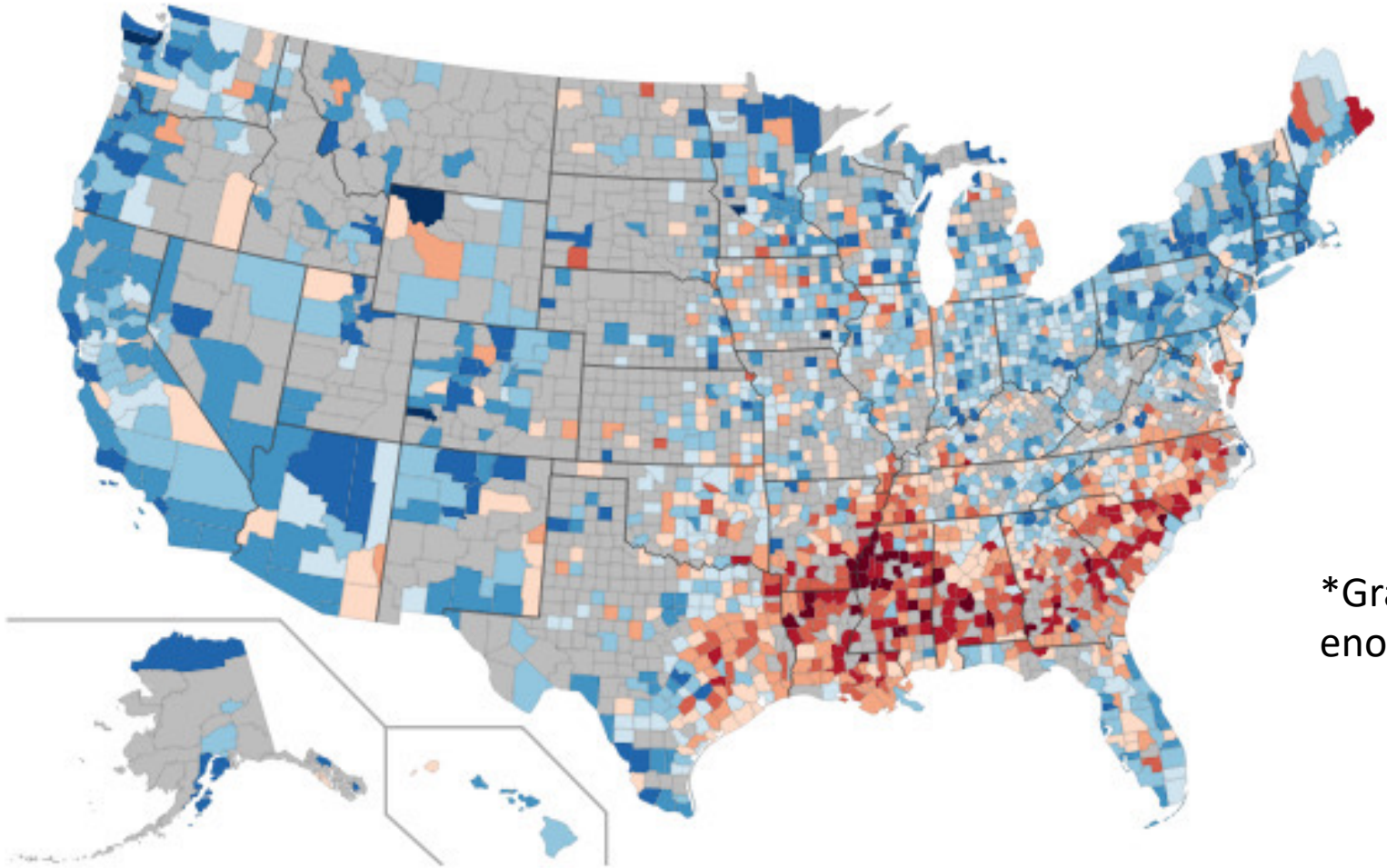


Embedding-based Individualism and Collectivism Lexica

1.5 billion Tweets from 6 million geolocated Twitter users

County-level individualism and collectivism scores

Individualism (blue) and Collectivism (red) across American counties



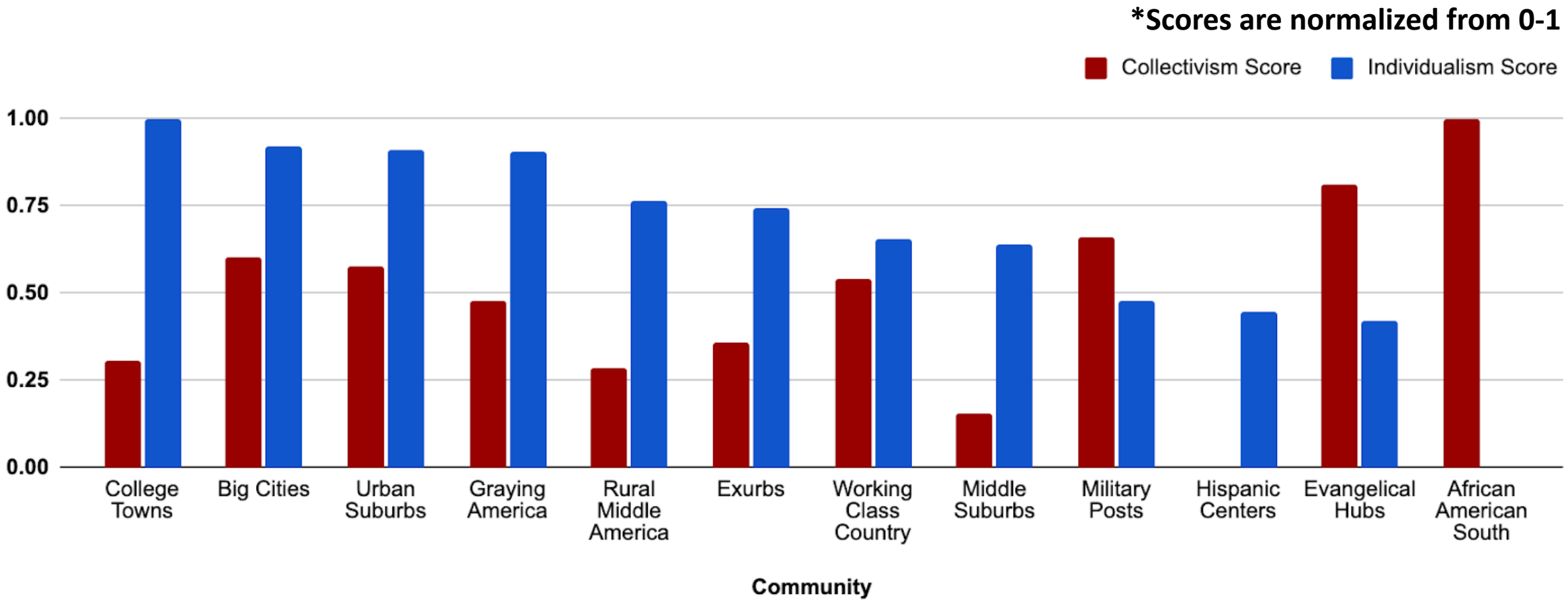
*Grayed out counties do not have enough Twitter data to apply our lexica

Resulting lexica capture cultural variation

- We validate against state-level indicators of collectivism
  - The GCI *(Pelham et. al, 2020)*
  - Collectivism scores *(Vandello and Cohen, 1999)*

- Strong correlation with previously established indicators of collectivism
  - Living arrangements
  - Religiosity
  - Ingroup bias (compatriotism)
  - Scores based on survey data

Individualism (blue) and Collectivism (red) across ACP communities



Future work

Language reflects culture!

- Easily extendible to a multilingual setting
- Analyze more cultural differences
  - Power distance
  - Looseness/tightness
- Understand what *drives* differences in behavior across regions

Thank you!