

# APPLIED ANALYTICS

WARNERMEDIA

Geocoding Without Geotags: A Text-based Approach for *reddit*

Keith Harrigan

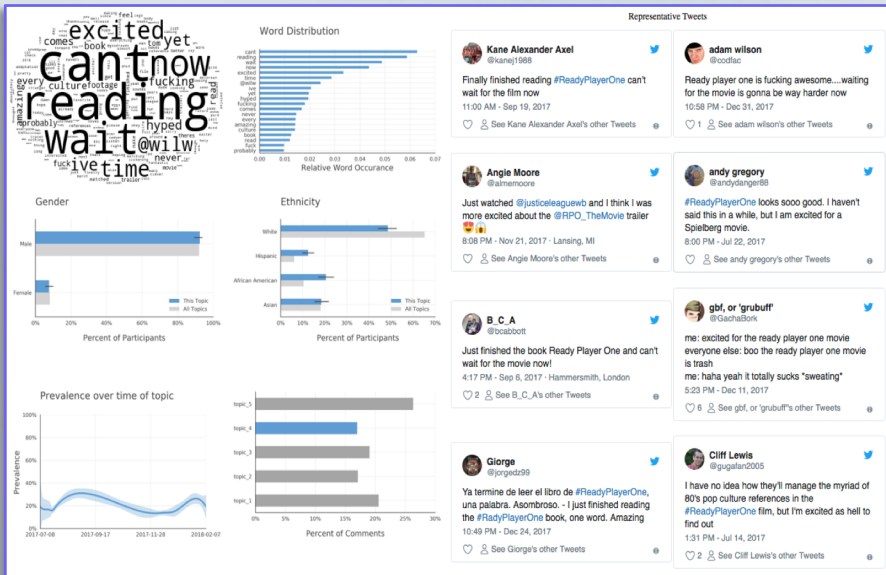
# WARNERMEDIA

The HBO logo, featuring the letters 'HBO' in a bold, black, sans-serif font, with a white circle around the letter 'O'.The HBO GO and HBO NOW logos. HBO GO is in a grey, sans-serif font, and HBO NOW is in a black, sans-serif font, with a white circle around the letter 'O'.The Turner logo, featuring the word 'Turner' in a bold, black, sans-serif font.

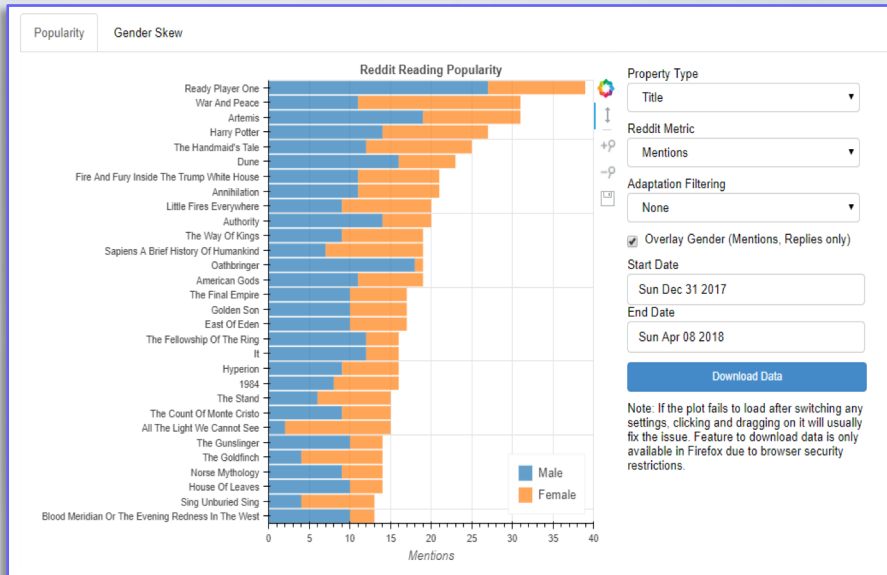
## Applied Analytics

- 15-person Quant Team with backgrounds in the social, physical, and mathematical sciences
- Employ advanced statistical techniques to inform the production and marketing of media properties
- Leverage social media and crowdsourced data to extract insights at scale

## Conversation Segmentation



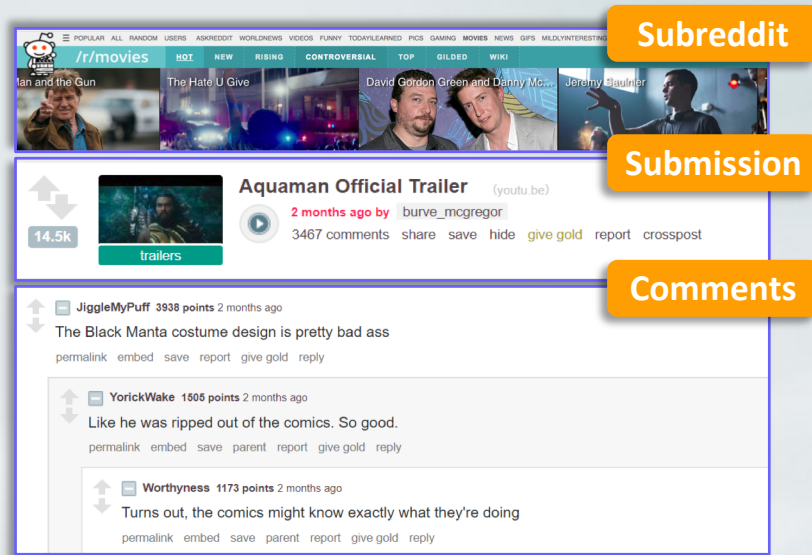
## Pop Culture Awareness



Demographic attribution provides an additional layer of audience understanding and enables data-driven targeted marketing

## *reddit* as a Social Platform

- 18<sup>th</sup> most visited website globally and 5<sup>th</sup> most visited in the United States
- Long-form commentary from the most dedicated fans
- Pseudonymity encourages disinhibition



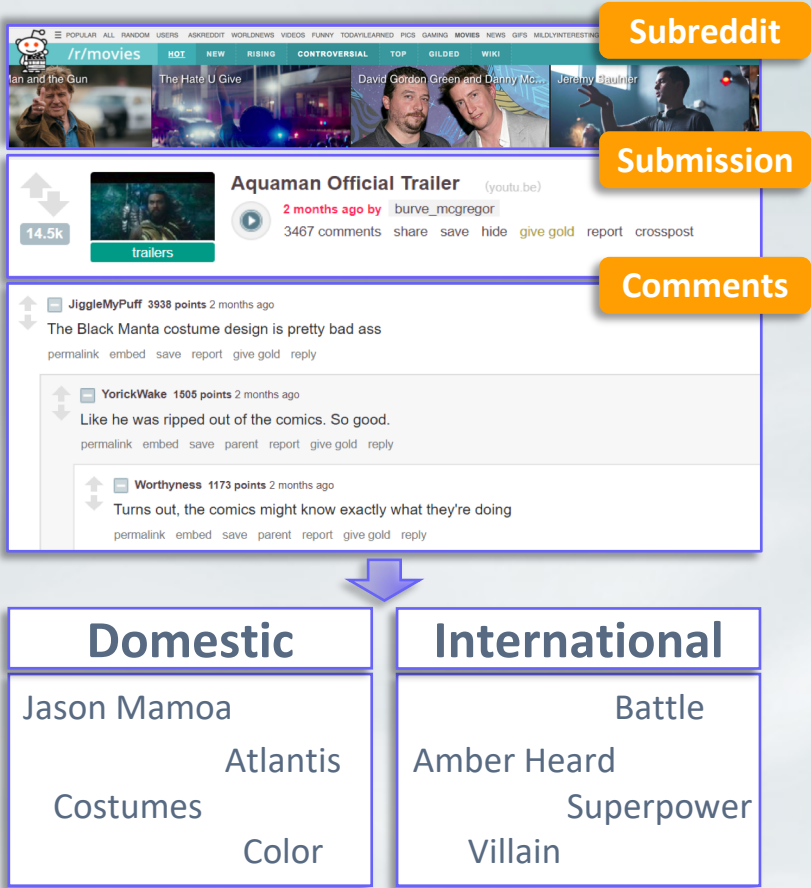


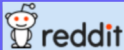
## reddit as a Social Platform

- 18<sup>th</sup> most visited website globally and 5<sup>th</sup> most visited in the United States
- Long-form commentary from the most dedicated fans
- Pseudonymity encourages disinhibition

## Geolocation Attribution

- Estimate global appeal of new media properties
- Inform region-specific marketing strategy (e.g. spend, creative material)





Sign Up

Log In

OVERVIEW POSTS COMMENTS \*\*\*

SORT BY NEW ▼

HuskyKeith commented on a post in r/modeltrains

↑

7

↓

Making a Baseboard with no woodworking tools or expirience (r/modeltrains)

submitted 26 days ago by OOScaleNerdUSA to r/modeltrains

HuskyKeith • 1 point • submitted 26 days ago

Not a traditional approach, but I was in a similar situation and went with some cheap tables from Ikea. I was able to pick up two tabletops and some legs and then configure everything in an L-shape to go into the corner of my apartment.

For me, without access to a car or power tools, this was the best option to get back into the hobby. Besides some slow shopping times, I haven't had any issues with tables (you can check out my post history to see how everything has turned out so far).

There are only two minor disadvantages when going this route. First, the table legs are on the shorter side of the spectrum compared to NMRA standards. If you have a bad back, you may not enjoy bending over to work on things. The second disadvantage is the cost. You'll pay a bit of a premium for relatively cheap material. That said, it might end up being cheaper than more traditional baseboard kits.

HuskyKeith commented on a post in r/pystats

↑

5

↓

Need HELP with building Recommender systems (using python) (r/pystats)

submitted 4 months ago by rbaja1997 to r/pystats

HuskyKeith • 1 point • submitted 4 months ago

Check out the "implicit" python package. It implements Alternating Least Squares regression to perform collaborative filtering. You can use the source code to learn some of the math and construct your own package from scratch thereafter.

If you want a more complex system, I'd suggest first choosing a domain you want to model. From there, do a literature search on Google Scholar or arXiv to see what's been implemented for your chosen domain. For example, if you want to work in a domain where a lot of attribute data is available (e.g. movie metadata, clothing descriptions), then a content-based approach might work well.

Python can probably be used for most of the prototyping, but any large-scale implementation may need a faster language or a "cythonized" adaptation.

↑

2

↓

Mixed Effects Linear Model for repeated measures (Statsmodels) (r/AskStatistics)

submitted 4 months ago by HuskyKeith to r/AskStatistics

4 comments share



u/HuskyKeith  
238 Karma

FOLLOW

SEND A PRIVATE MESSAGE

Following this user will show all the posts they make to their profile on your front page.

ACTIVE IN THESE COMMUNITIES

- r/statistics

61,249 subscribers

SUBSCRIBE
- r/AskStatistics

11,551 subscribers

SUBSCRIBE
- r/modeltrains

14,683 subscribers

SUBSCRIBE
- r/boston

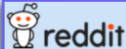
99,348 subscribers

SUBSCRIBE
- r/drums

63,284 subscribers

SUBSCRIBE

View more




Sign UpLog In

OVERVIEWPOSTSCOMMENTS \*\*\*

SORT BY NEW ▼

HuskyKeith commented on a post in r/modeltrains

↑  
7  
↓



**Making a Baseboard with no woodworking tools or expirience** (r/modeltrains)  
submitted 26 days ago by OOScaleNerdUSA to r/modeltrains

HuskyKeith • 1 point • submitted 26 days ago


Not a traditional approach, but I was in a similar situation and went with some cheap tables from Ikea. I was able to pick up two tabletops and some legs and then configure everything in an L-shape to go into the corner of my apartment.

For me, without access to a car or power tools, this was the best option to get back into the hobby. Besides some slow shopping times, I haven't had any issues with tables (you can check out my post history to see how everything has turned out so far).

There are only two minor disadvantages when going this route. First, the table legs are on the shorter side of the spectrum compared to NMRA standards. If you have a bad back, you may not enjoy bending over to work on things. The second disadvantage is the cost. You'll pay a bit of a premium for relatively cheap material. That said, it might end up being cheaper than more traditional baseboard kits.

HuskyKeith commented on a post in r/pystats

↑  
5  
↓



**Need HELP with building Recommender systems (using python)** (r/pystats)  
submitted 4 months ago by rbaja1997 to r/pystats


HuskyKeith • 1 point • submitted 4 months ago

Check out the "implicit" python package. It implements Alternating Least Squares regression to perform collaborative filtering. You can use the source code to learn some of the math and construct your own package from scratch thereafter.

If you want a more complex system, I'd suggest first choosing a domain you want to model. From there, do a literature search on Google Scholar or arXiv to see what's been implemented for your chosen domain. For example, if you want to work in a domain where a lot of attribute data is available (e.g. movie metadata, clothing descriptions), then a content-based approach might work well.

Python can probably be used for most of the prototyping, but any large-scale implementation may need a faster language or a "cythonized" adaptation.

↑  
2  
↓



**Mixed Effects Linear Model for repeated measures (Statsmodels)** (r/AskStatistics)  
submitted 4 months ago by HuskyKeith to r/AskStatistics  
4 comments share



**u/HuskyKeith**  
238 Karma

FOLLOW

SEND A PRIVATE MESSAGE

Following this user will show all the posts they make to their profile on your front page.

ACTIVE IN THESE COMMUNITIES



**r/statistics**  
61,249 subscribers

SUBSCRIBE



**r/AskStatistics**  
11,551 subscribers

SUBSCRIBE



**r/modeltrains**  
14,683 subscribers

SUBSCRIBE



**r/boston**  
99,348 subscribers

SUBSCRIBE



**r/drums**  
63,284 subscribers

SUBSCRIBE

View more

User profiles lack  
location information

The screenshot shows a Reddit interface. At the top, there's a navigation bar with the Reddit logo, 'Sign Up', and 'Log In' buttons. Below this is a sub-header with 'OVERVIEW', 'POSTS', 'COMMENTS', and '\*\*\*'. A 'SORT BY NEW' dropdown is on the right. The main content area shows a comment thread. The first comment is by 'HuskyKeith' on a post in r/modeltrains. The second comment is also by 'HuskyKeith' on a post in r/pystats. The third comment is by 'HuskyKeith' on a post in r/AskStatistics. On the right side, there's a user profile for 'u/HuskyKeith' with 238 Karma. Below the profile is a 'FOLLOW' button and a 'SEND A PRIVATE MESSAGE' button. At the bottom right, there's a section 'ACTIVE IN THESE COMMUNITIES' with a list of subreddits and their subscriber counts, each with a 'SUBSCRIBE' button.

**Comment geotagging not supported**

**User profiles lack location information**

reddit

OVERVIEW POSTS COMMENTS \*\*\*

SORT BY NEW ▼

HuskyKeith commented on a post in r/modeltrains

7

Making a Baseboard with no woodworking tools or expirience (r/modeltrains)  
submitted 26 days ago by OOScaleNerdUSA to r/modeltrains

HuskyKeith • 1 point • submitted 26 days ago  
Not a traditional approach, but I was in a similar situation and went with some cheap tables from Ikea. I was able to pick up two tabletops and some legs and then configure everything in an L-shape to go into the corner of my apartment.  
For me, without access to a car or power tools, this was the best option to get back into the hobby. Besides some slow shopping times, I haven't had any issues with tables (you can check out my post history to see how everything has turned out so far).  
There are only two minor disadvantages when going this route. First, the table legs are on the shorter side of the spectrum  
wards. If you have a bad back, you may not enjoy bending over to work on things. The second disadvantage  
of a premium for relatively cheap material. That said, it might end up being cheaper than more traditional

post in r/pystats

5

Need HELP with building Recommender systems (using python) (r/pystats)  
submitted 4 months ago by rbaja1997 to r/pystats

HuskyKeith • 1 point • submitted 4 months ago  
Check out the "implicit" python package. It implements Alternating Least Squares regression to perform collaborative filtering. You can use the source code to learn some of the math and construct your own package from scratch thereafter.  
If you want a more complex system, I'd suggest first choosing a domain you want to model. From there, do a literature search on Google Scholar or arXiv to see what's been implemented for your chosen domain. For example, if you want to work in a domain where a lot of attribute data is available (e.g. movie metadata, clothing descriptions), then a content-based approach might work well.  
Python can probably be used for most of the prototyping, but any large-scale implementation may need a faster language or a "cythonized" adaptation.

2

Mixed Effects Linear Model for repeated measures (Statsmodels) (r/AskStatistics)  
submitted 4 months ago by HuskyKeith to r/AskStatistics  
4 comments share

u/HuskyKeith  
238 Karma ⓘ

FOLLOW

SEND A PRIVATE MESSAGE

Following this user will show all the posts they make to their profile on your front page.

ACTIVE IN THESE COMMUNITIES

r/statistics  
61,249 subscribers SUBSCRIBE

r/AskStatistics  
11,551 subscribers SUBSCRIBE

r/modeltrains  
14,683 subscribers SUBSCRIBE

r/boston  
99,348 subscribers SUBSCRIBE

r/drums  
63,284 subscribers SUBSCRIBE

View more

- Limited understanding of domain transfer in geolocation inference tasks
- Hypothesize that models trained on out-of-domain data will not perform optimally on *reddit*
  1. Demographics vary across platforms
  2. Network-based models require within-domain grounding
  3. Metadata specific to the *reddit* platform may be useful (e.g. subreddit, flair, and hierarchical comment structure)

**Models that generalize between social platforms are limited in the business context without the ability to validate prediction certainty**

## Manually Curate Seed Submissions



- Use Python *reddit* API Wrapper to query for submissions with title similar to “Where do you live?”
- Manually filter down to 1,200 most promising submissions

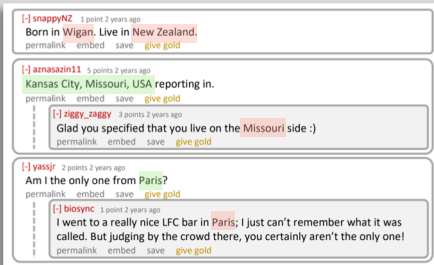


## Manually Curate Seed Submissions



- Use Python *reddit* API Wrapper to query for submissions with title similar to “Where do you live?”
- Manually filter down to 1,200 most promising submissions

## Extract Locations From Noisy Text



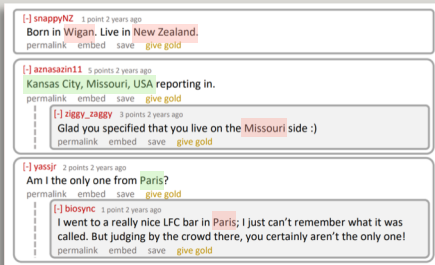
- Isolate top-level comments; remove comments mentioning “born” or “move”
- String-matching and data-informed heuristics (syntax, abbreviations) to identify locations

## Manually Curate Seed Submissions



- Use Python *reddit* API Wrapper to query for submissions with title similar to “Where do you live?”
- Manually filter down to 1,200 most promising submissions

## Extract Locations From Noisy Text



- Isolate top-level comments; remove comments mentioning “born” or “move”
- String-matching and data-informed heuristics (syntax, abbreviations) to identify locations

## Assign Geographic Coordinates



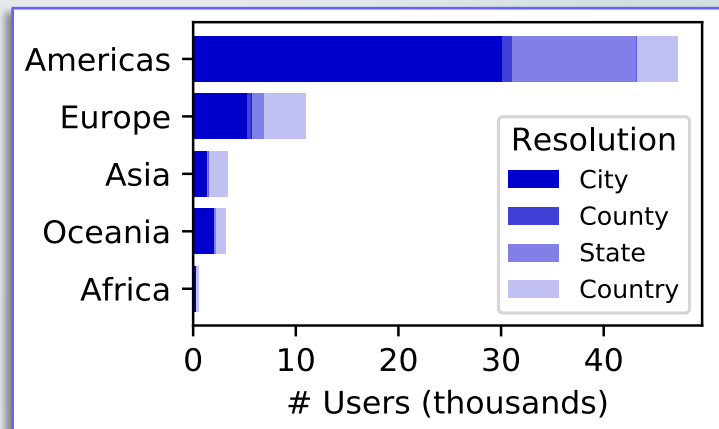
- Leverage Google Geocoding API to assign coordinates to strings
- Bias query results based on source subreddit (e.g. “Scarborough” in *r/Ontario* vs. *r/CasualUK*)

## Location Distribution

- 65,245 labeled users
- Top 5 countries are consistent with Alexa's panel, but over-indexes in North America

## Error Analysis

- 89% of randomly sampled users were labeled within the correct hierarchy and at the appropriate topological resolution
- Accuracy would benefit from improved NLU
  - Disambiguation (e.g. Kansas City, Missouri vs. Kansas City, Kansas)
  - Multiple Locations Mentioned ("From Los Angeles, but currently living in Boston")



Distribution of Labeled Users and Geocoding Resolution

Country	Alexa Traffic	Labeled Users
United States	58.7%	60.1% (n=39,236)
United Kingdom	7.4%	5.4% (n=3,544)
Canada	6.0%	9.4% (n=6,163)
Australia	3.1%	3.5% (n=2,344)
Germany	2.1%	1.7% (n=1,097)

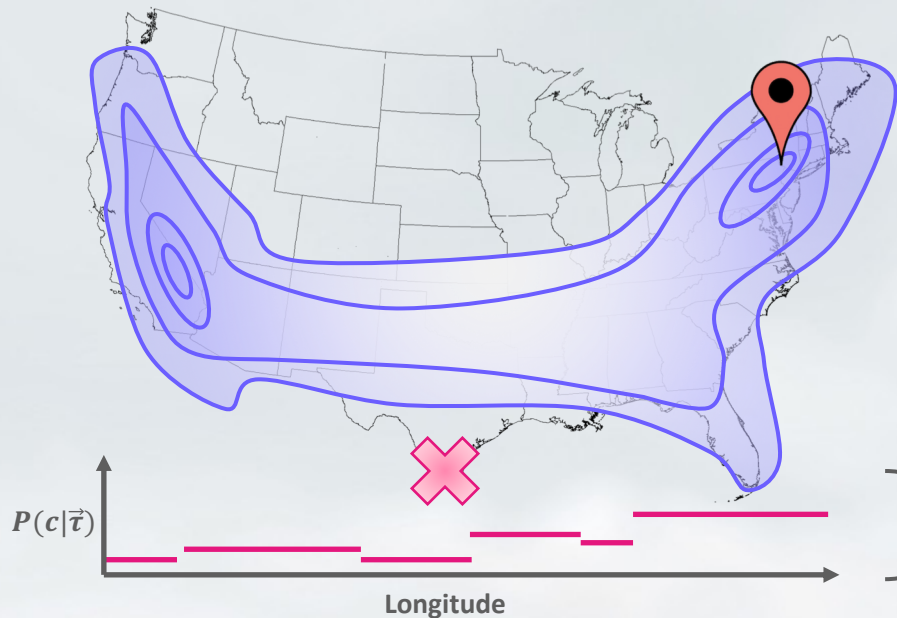
User Distribution vs. Alexa's Propriety Traffic Panel

## Inference Using Imperfect Labels

- **Language usage**  $\vec{w}$ : Bag of words representation of user comments
- **Subreddit membership**  $\vec{s}$ : Frequency distribution of comments amongst subreddits
- **Temporal posting pattern**  $\vec{\tau}$ : Comment counts across 24 hours of the day (UTC)

## Inference Using Imperfect Labels

- **Language usage**  $\vec{w}$ : Bag of words representation of user comments
- **Subreddit membership**  $\vec{s}$ : Frequency distribution of comments amongst subreddits
- **Temporal posting pattern**  $\vec{\tau}$ : Comment counts across 24 hours of the day (UTC)



$$P(c|\vec{u}, \vec{\tau}) \propto P(c|\vec{\tau}) \sum_{u \in \vec{u}} \|u\| P(c|u) P(u)$$

Logistic Regression Estimate Given  $\vec{\tau}$

Dirichlet Process Mixture Model for  $\vec{w}$  &  $\vec{s}$

Model inspired by Cheng et al. (2010)  
and Chang et al. (2012)

Longitudes discretized in percentile-based bins; coordinates in  $C$  take on the probability from assigned longitude bin

## Dimensionality reduction using Non-localness (Chang et al., 2012)

$$NL(f) = \sum_{s \in S} sim_{SKL}(f, s) P(s)$$

$$sim_{SKL}(f_i, f_j) = \sum_{c \in C} P(c|f_i) \log\left(\frac{P(c|f_i)}{P(c|f_j)}\right) + P(c|f_j) \log\left(\frac{P(c|f_j)}{P(c|f_i)}\right)$$



## Dimensionality reduction using Non-localness (Chang et al., 2012)

$$NL(f) = \sum_{s \in S} sim_{SKL}(f, s) P(s)$$

$$sim_{SKL}(f_i, f_j) = \sum_{c \in C} P(c|f_i) \log\left(\frac{P(c|f_i)}{P(c|f_j)}\right) + P(c|f_j) \log\left(\frac{P(c|f_j)}{P(c|f_i)}\right)$$

## Top Words

## Top Subreddits

Massachusetts, USA

allston, mbta, waltham, saugus, brookline, masshole

r/PokemonGoBoston, r/bostonhousing

Ohio, USA

ohioan, cincinnatis, jenis, clevelander, graeters

r/uCinci, r/ColumbusSocial

Germany

zeigen, dennoch, wenige, zeigt, solltest

r/FragReddit, r/de\_IAmA, r/rocketbeans,

Belgium

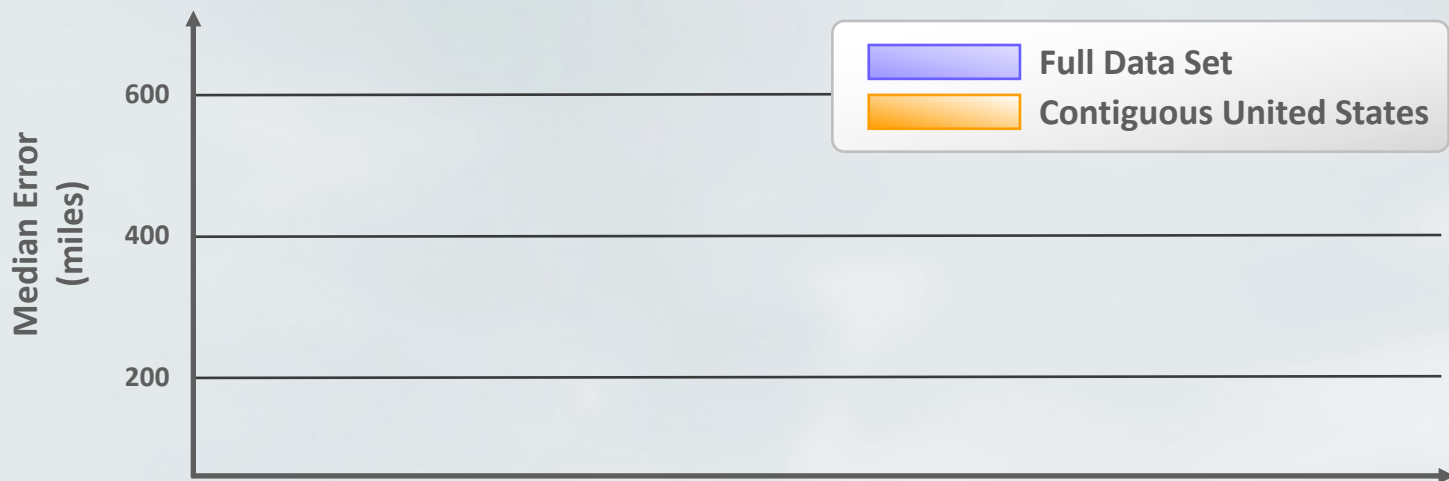
telenet, walloon, vlaams, jupiler, leuven, vlaanderen

r/belgium, r/brussels, r/Vivillon, r/ecr\_eu

Feature selection procedure validates data set construction, reduces computational expense, and improves prediction accuracy

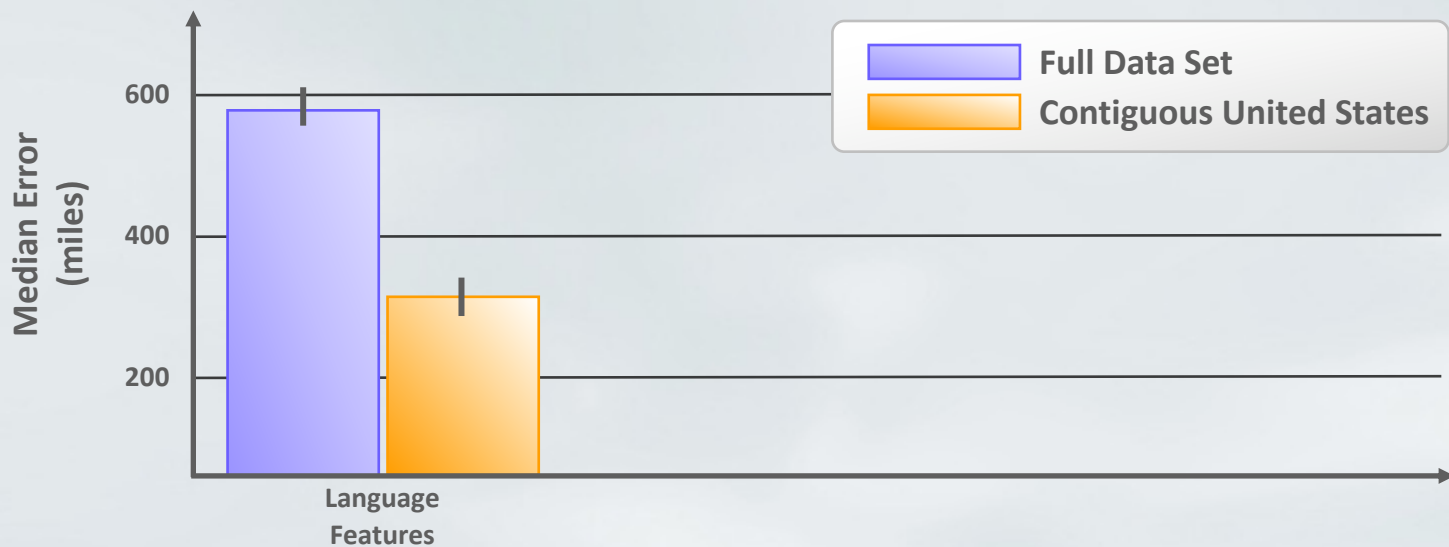
## Evaluation Procedure

- 5-fold Cross Validation with hyperparameter optimization (# features, temporal classifier)



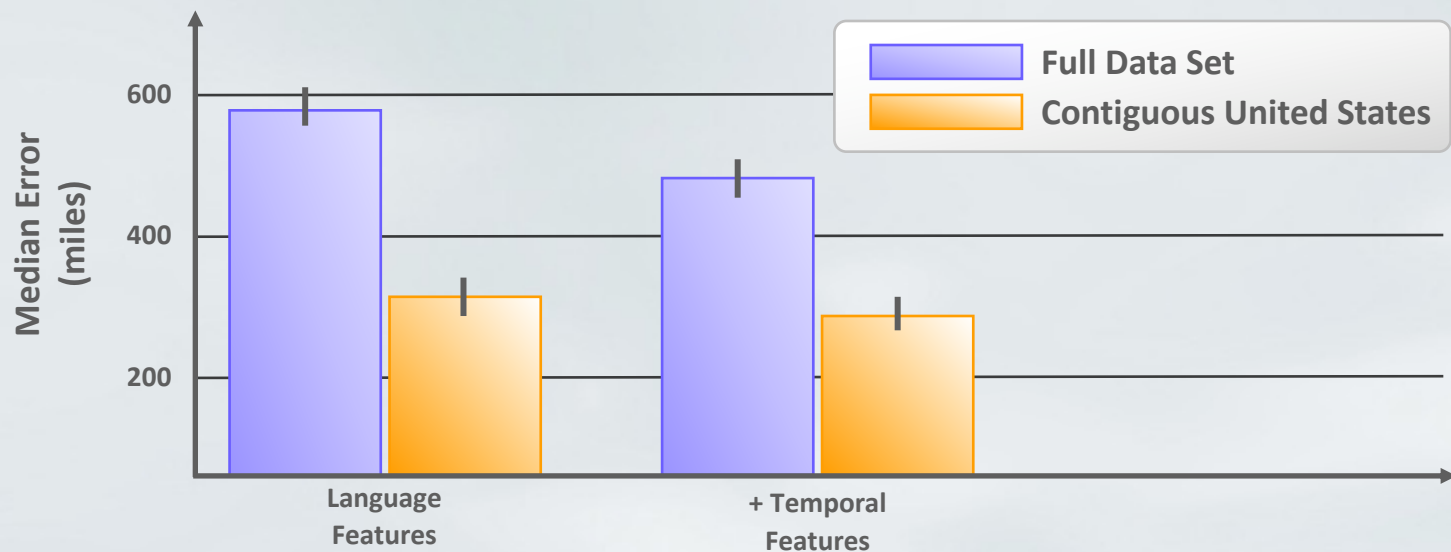
## Evaluation Procedure

- 5-fold Cross Validation with hyperparameter optimization (# features, temporal classifier)



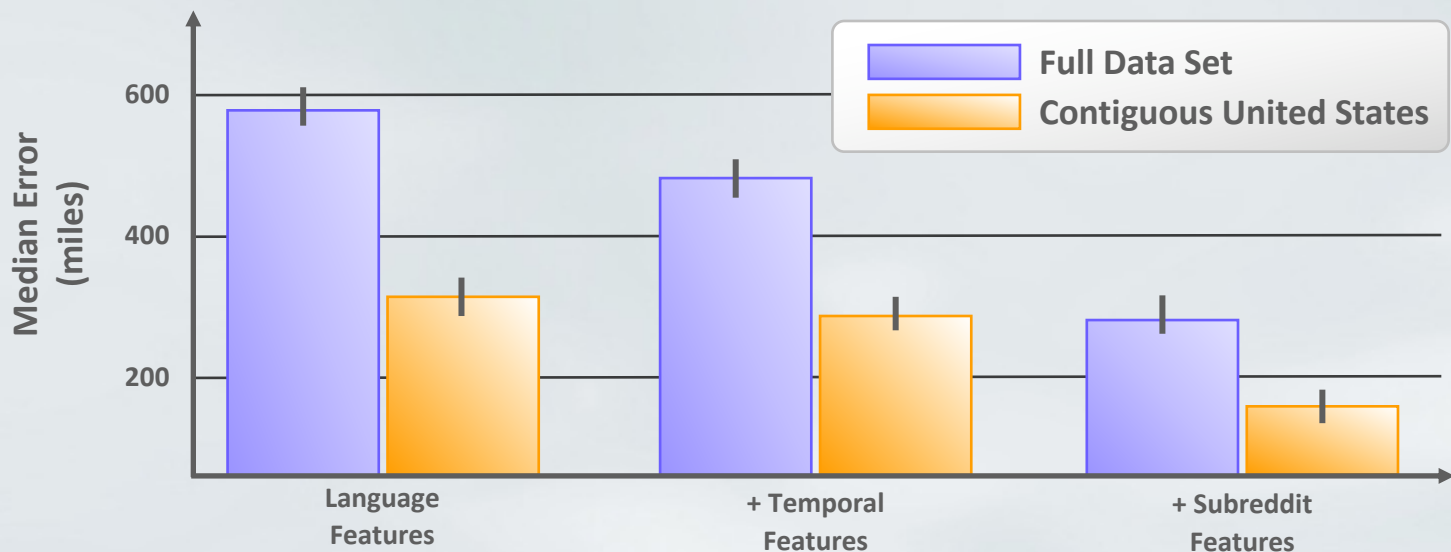
## Evaluation Procedure

- 5-fold Cross Validation with hyperparameter optimization (# features, temporal classifier)



## Evaluation Procedure

- 5-fold Cross Validation with hyperparameter optimization (# features, temporal classifier)



Temporal features reduce error in global data set, while platform-specific subreddit metadata improves performance in both data sets

## Twitter Data Sets

- *Geotext* (Eisenstein et al., 2010), *Twitter-US* (Roller et al., 2012), *Twitter-World* (Han et al., 2012)

## Systematic Comparison

Systematic Comparison

		Testing Set					
		reddit-US	reddit-Full	Geotext	TW-US	TW-World (US Subset)	TW-World (Full)
Training Set	reddit-US						
	reddit-Full						
	Geotext						
	TW-US						
	TW-World (US Subset)						
	TW-World (Full)						



## Twitter Data Sets

- *Geotext* (Eisenstein et al., 2010), *Twitter-US* (Roller et al., 2012), *Twitter-World* (Han et al., 2012)

## Systematic Comparison

		Testing Set					
		reddit-US	reddit-Full	Geotext	TW-US	TW-World (US Subset)	TW-World (Full)
Training Set	reddit-US						
	reddit-Full					5-Fold Cross Validation Hyperparameter Opti	
	Geotext						
	TW-US						
	TW-World (US Subset)						
	TW-World (Full)						

5-Fold Cross Validation With  
Hyperparameter Optimization

## Twitter Data Sets

- *Geotext* (Eisenstein et al., 2010), *Twitter-US* (Roller et al., 2012), *Twitter-World* (Han et al., 2012)

## Systematic Comparison

		Testing Set					
		reddit-US	reddit-Full	Geotext	TW-US	TW-World (US Subset)	TW-World (Full)
Training Set	reddit-US						
	reddit-Full						
	Geotext						
	TW-US						
	TW-World (US Subset)						
	TW-World (Full)						

Optimize Hyperparameters For Training Data

Optimize Hyperparameters  
For Training Data Set

## Twitter Data Sets

- *Geotext* (Eisenstein et al., 2010), *Twitter-US* (Roller et al., 2012), *Twitter-World* (Han et al., 2012)

## Systematic Comparison

Systematic Comparison

		Testing Set					
		reddit-US	reddit-Full	Geotext	TW-US	TW-World (US Subset)	TW-World (Full)
Training Set	reddit-US	157		479	358	592	
	reddit-Full		266				1329
	Geotext	1019		271	755	755	
	TW-US	294		304	220	582	
	TW-World (US Subset)	717		311	563	584	
	TW-World (Full)		817				1405
		Median Error Within-Domain			Median Error Between-Domain		

## Executive Summary

- To the best of our knowledge, this is the first geolocation approach for *reddit*
- Pseudonymity is not an exhaustive barrier to supervised learning
- Metadata specific to the reddit platform critically improves performance
- Significant loss in performance incurred during domain transfer

## Executive Summary

- To the best of our knowledge, this is the first geolocation approach for *reddit*
- Pseudonymity is not an exhaustive barrier to supervised learning
- Metadata specific to the reddit platform critically improves performance
- Significant loss in performance incurred during domain transfer

## Future Directions

- Examine robust natural language understanding systems to improve labeling
- Explore biases introduced during labeling procedure (e.g. activity, topicality)
- Re-run analysis using DNN or more complex model architecture

**Distant supervision provides a viable option to obtaining demographic labels at scale and enables downstream predictive modeling**



APPLIED  
ANALYTICS  
WARNERMEDIA

T H A N K   Y O U