Incorporating Dialectal Variability for Socially Equitable Language Identification

David Jurgens, Yulia Tsvetkov, and Dan Jurafsky



McNamee, P., “Language identification: *a solved problem* suitable for undergraduate instruction” Journal of Computing Sciences in Colleges 20(3) 2005.

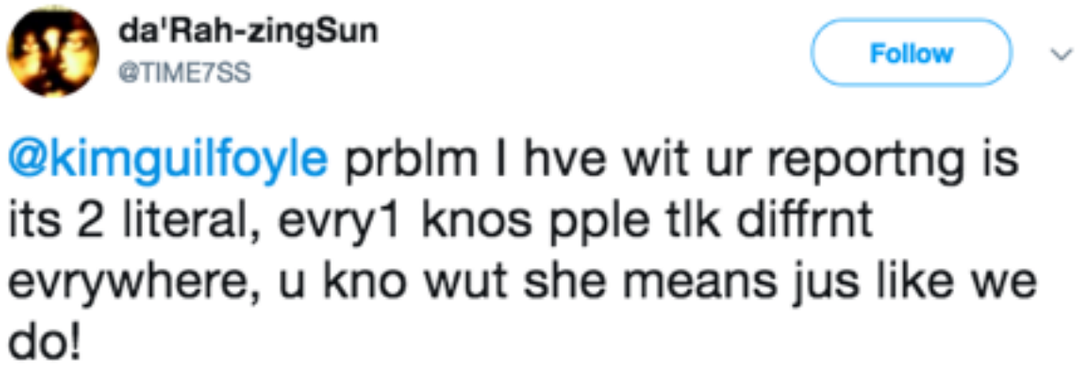
McNamee, P., “Language identification: *a solved problem* suitable for undergraduate instruction” Journal of Computing Sciences in Colleges 20(3) 2005.

“This paper describes […] how even the most simple of these methods **using data obtained from the World Wide Web achieve accuracy approaching 100%** on a test suite comprised of ten European languages”

Whose language are we identifying?

Whose language are we identifying?

Whose language are we identifying?



Whose language are we identifying?

The Royal Family ✓
@RoyalFamily

Follow

Taking place this week on the river Thames is 'Swan Upping' – the annual census of the swan population on the Thames.

da'Rah-zingSun
@TIME7SS

Follow

@kimguilfoyle prblm I hve wit ur reportng is its 2 literal, evry1 knos pple tlk diffrnt evrywhere, u kno wut she means jus like we do!

Mooktar
@bossmukky

Follow

"@Ecstatic_Mi: @bossmukky Ebi like say I wan dey sick sef wlh 'Flu' my whole body dey weak"uw gee...

Ebenezer·
@Physique_cian

Follow

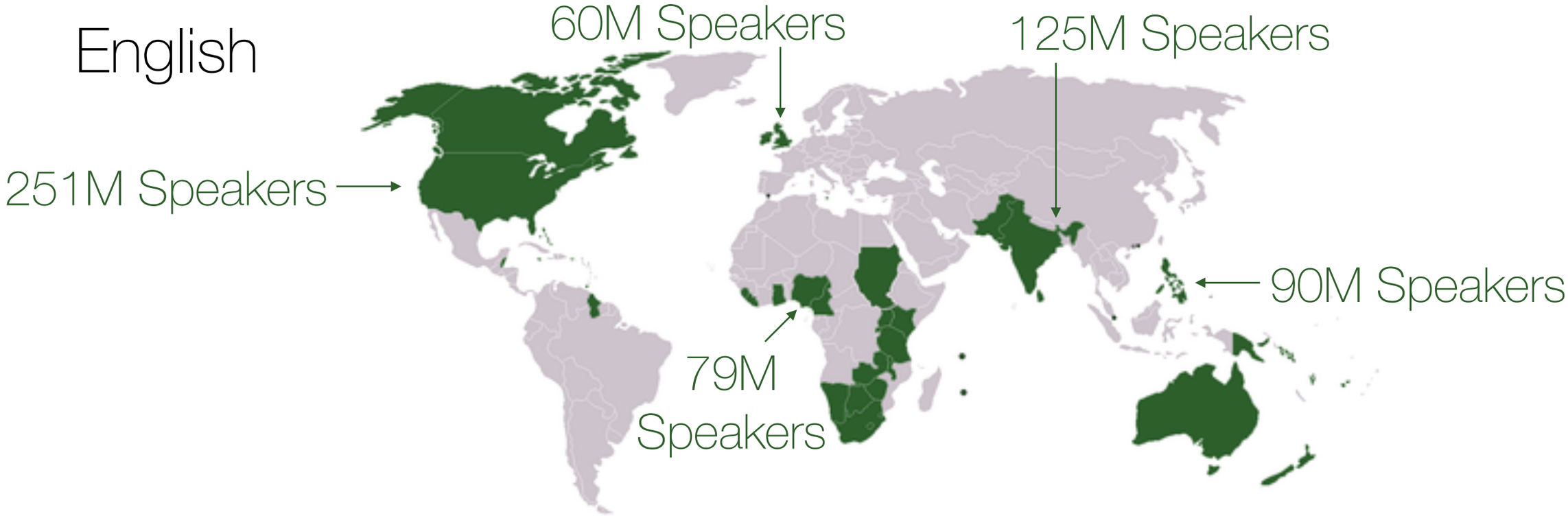
@Tblazeen R u a wizard or wat gan sef : in d mornin- u tweet, afternoon - u tweet, nyt gan u dey tweet.beta get ur IT placement wiv twitter

Global platforms attract global diversity in a language

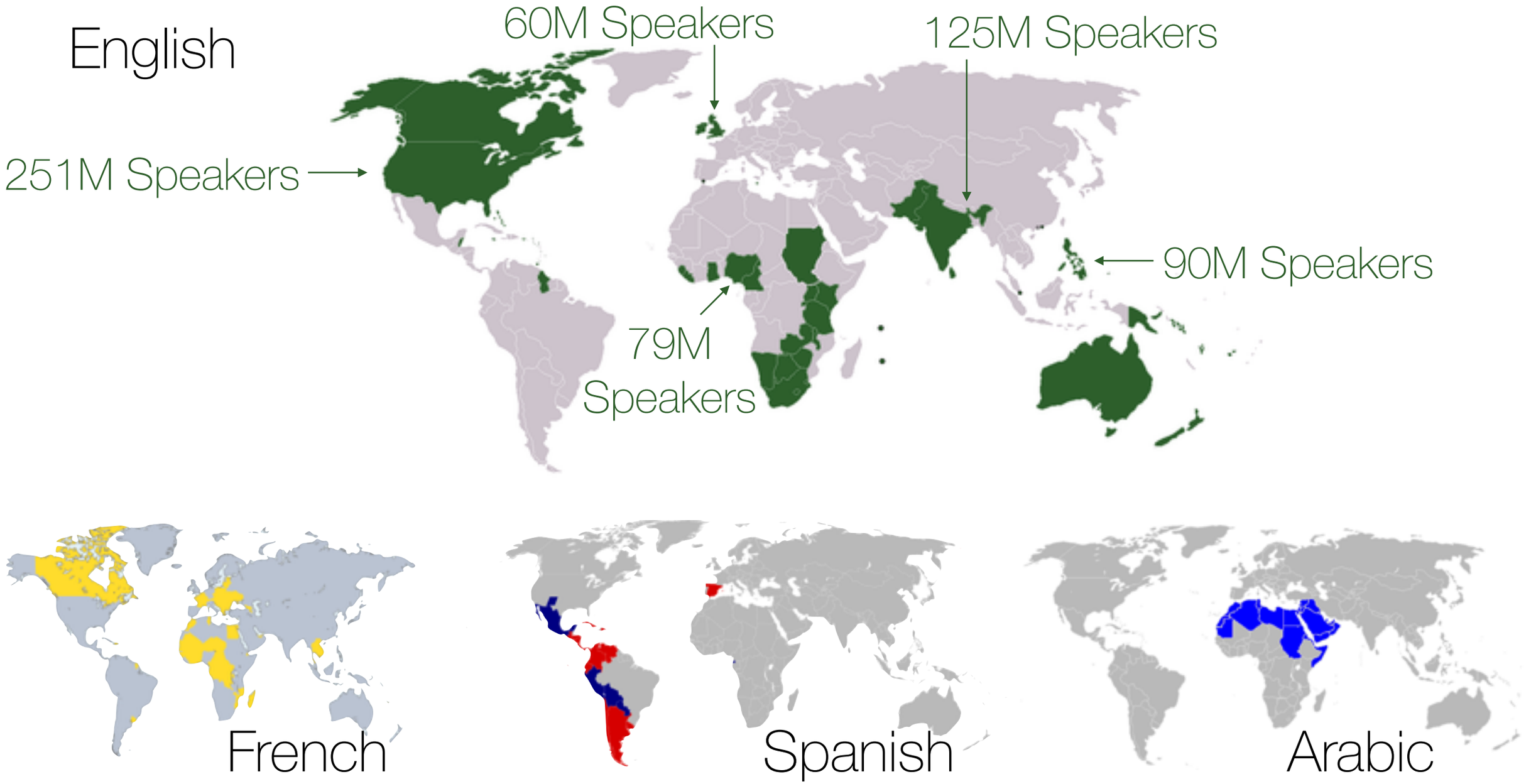
English

Global platforms attract global diversity in a language

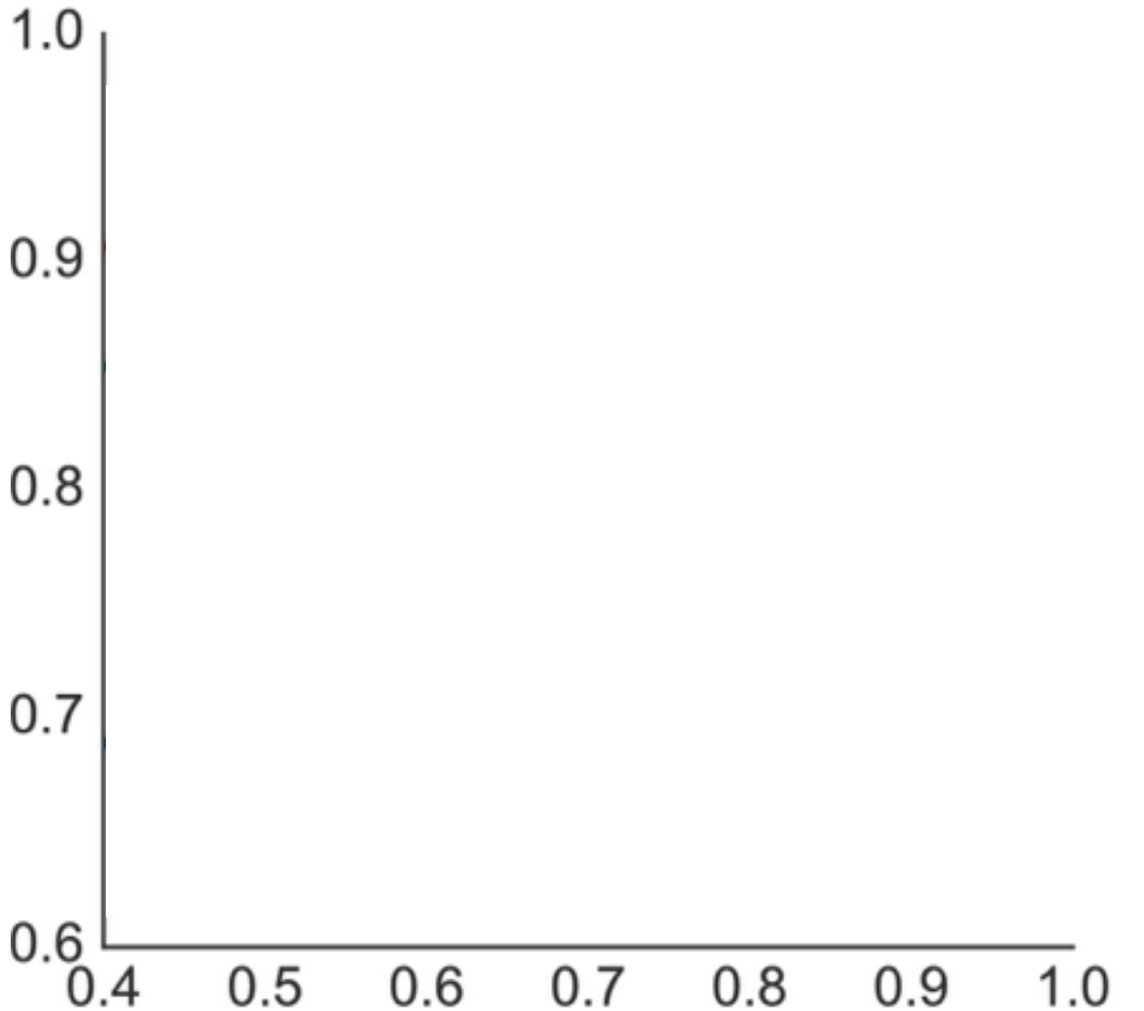
English



Global platforms attract global diversity in a language



Estimated LID accuracy for English tweets

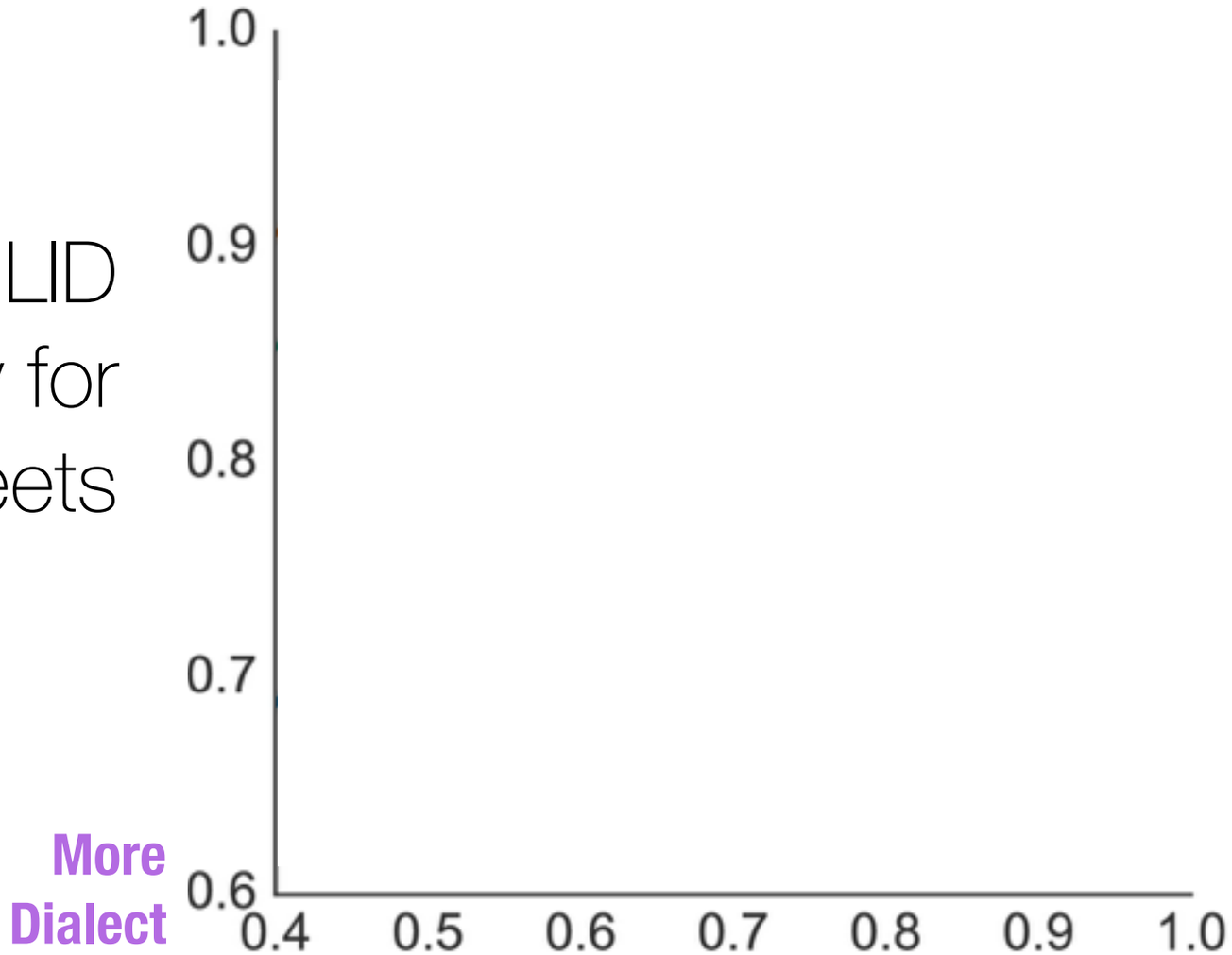


Human Development Index of text's origin country

- Education
- Life expectancy
- Income

(Labov, 1964; Ash, 2002)

Estimated LID accuracy for English tweets



More Dialect

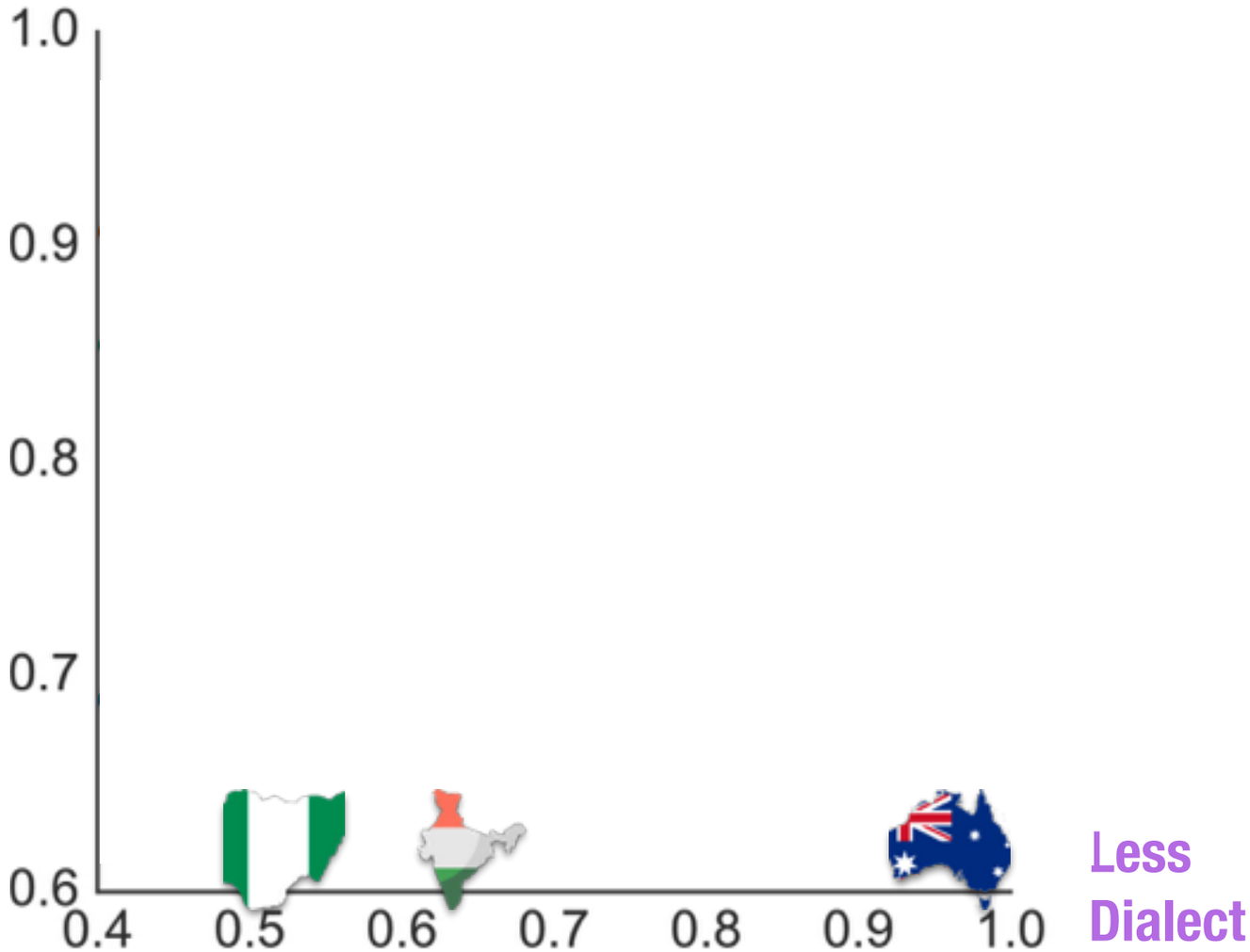
Less Dialect

Human Development Index of text's origin country

- Education
- Life expectancy
- Income

(Labov, 1964; Ash, 2002)

Estimated LID accuracy for English tweets



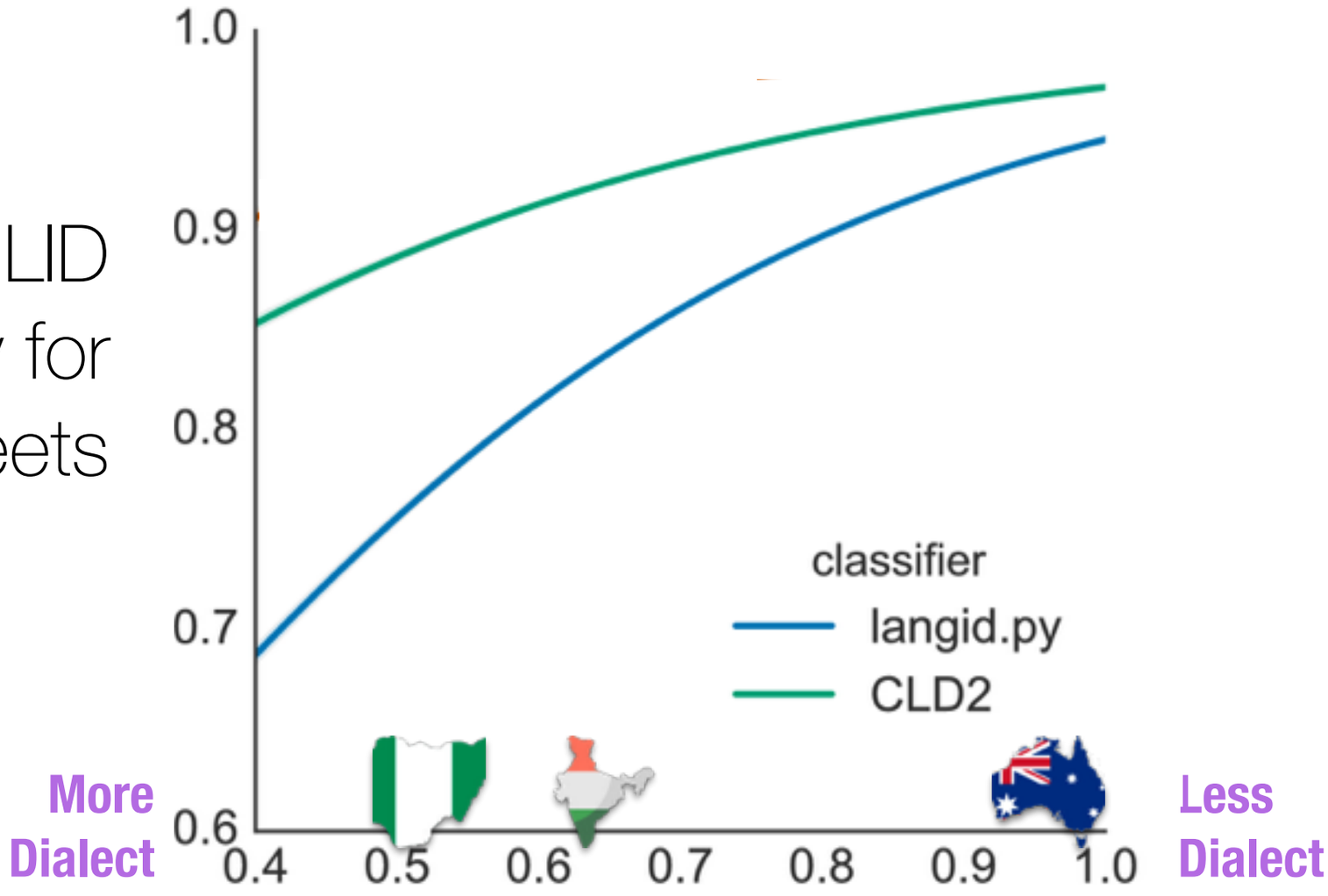
Human Development Index of text's origin country

- Education
- Life expectancy
- Income

(Labov, 1964; Ash, 2002)

Current language detection methods perform significantly worse in less-developed countries

Estimated LID accuracy for English tweets

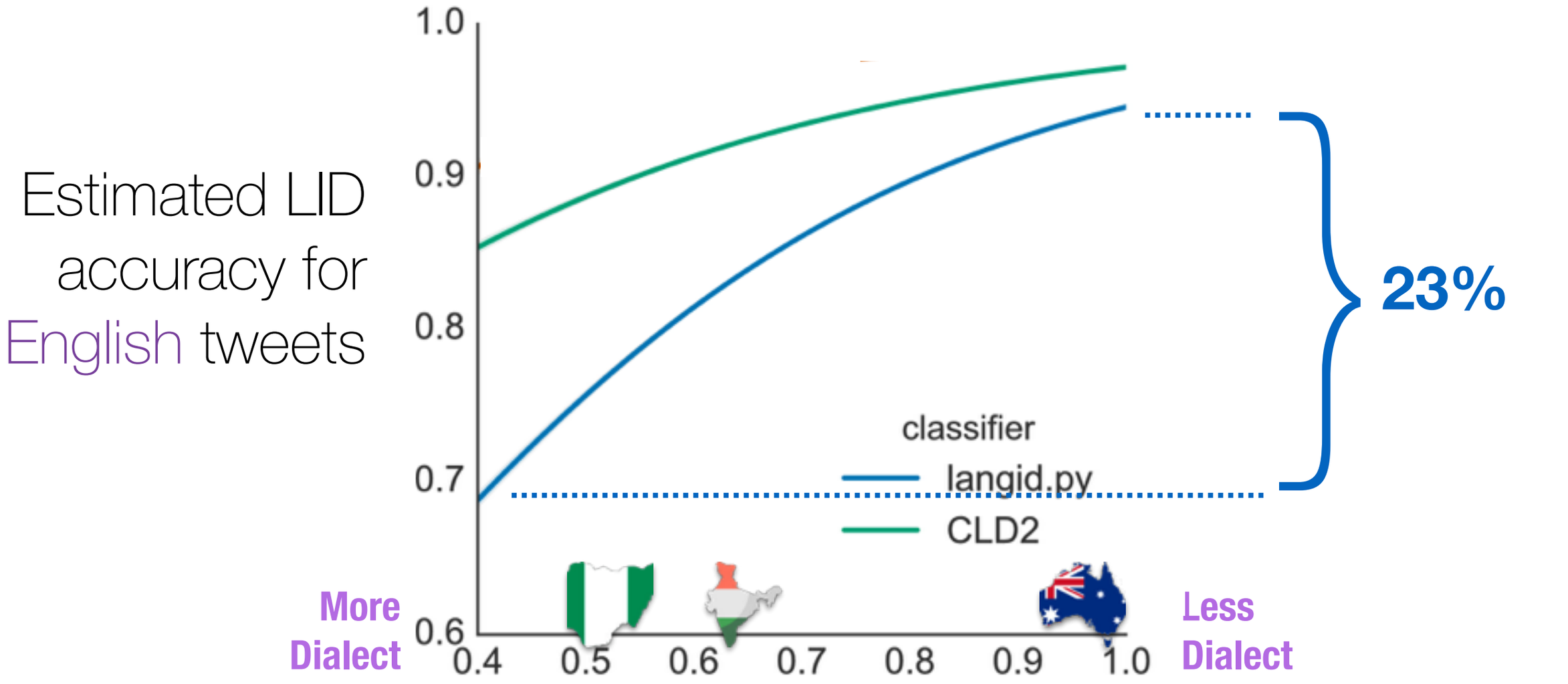


Human Development Index of text's origin country

- Education
- Life expectancy
- Income

(Labov, 1964; Ash, 2002)

Current language detection methods perform significantly worse in less-developed countries



Human Development Index of text's origin country

- Education
- Life expectancy
- Income

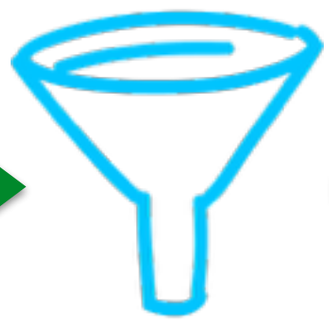
(Labov, 1964; Ash, 2002)

Practical Motivation: Epidemic Detection

Keyword Filter
"flu", "sick"

NLP
Which symptoms?

Practical Motivation: Epidemic Detection

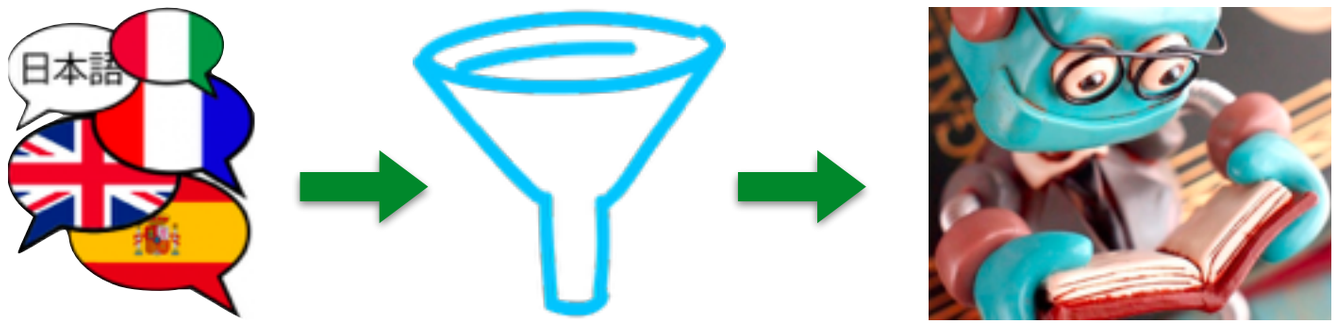


Language Detection

Keyword Filter
“flu”, “sick”

NLP
Which symptoms?

Practical Motivation: Epidemic Detection



Language Detection

Keyword Filter
"flu", "sick"

NLP
Which symptoms?

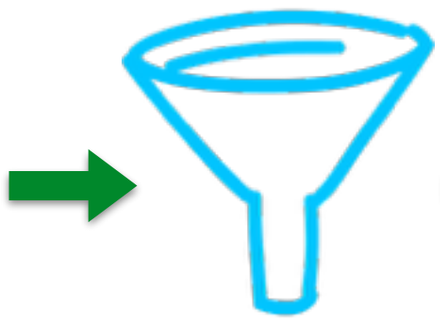
non-English

Practical Motivation: Epidemic Detection

Brooke
@Brookiepoo134

Follow

got the flu over the weekend and I didn't know until today, & I somehow managed to give it to FIVE of my friends!!!!!!



Language Detection

Keyword Filter
"flu", "sick"

NLP
Which symptoms?

non-English

Practical Motivation: Epidemic Detection

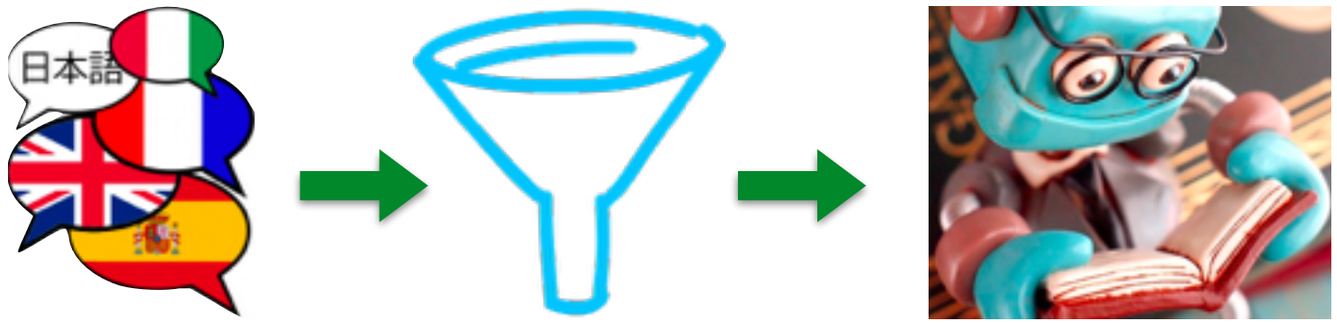
Language Detection

Keyword Filter
"flu", "sick"

NLP
Which symptoms?

non-English

Practical Motivation: Epidemic Detection



Language Detection

Keyword Filter
"flu", "sick"

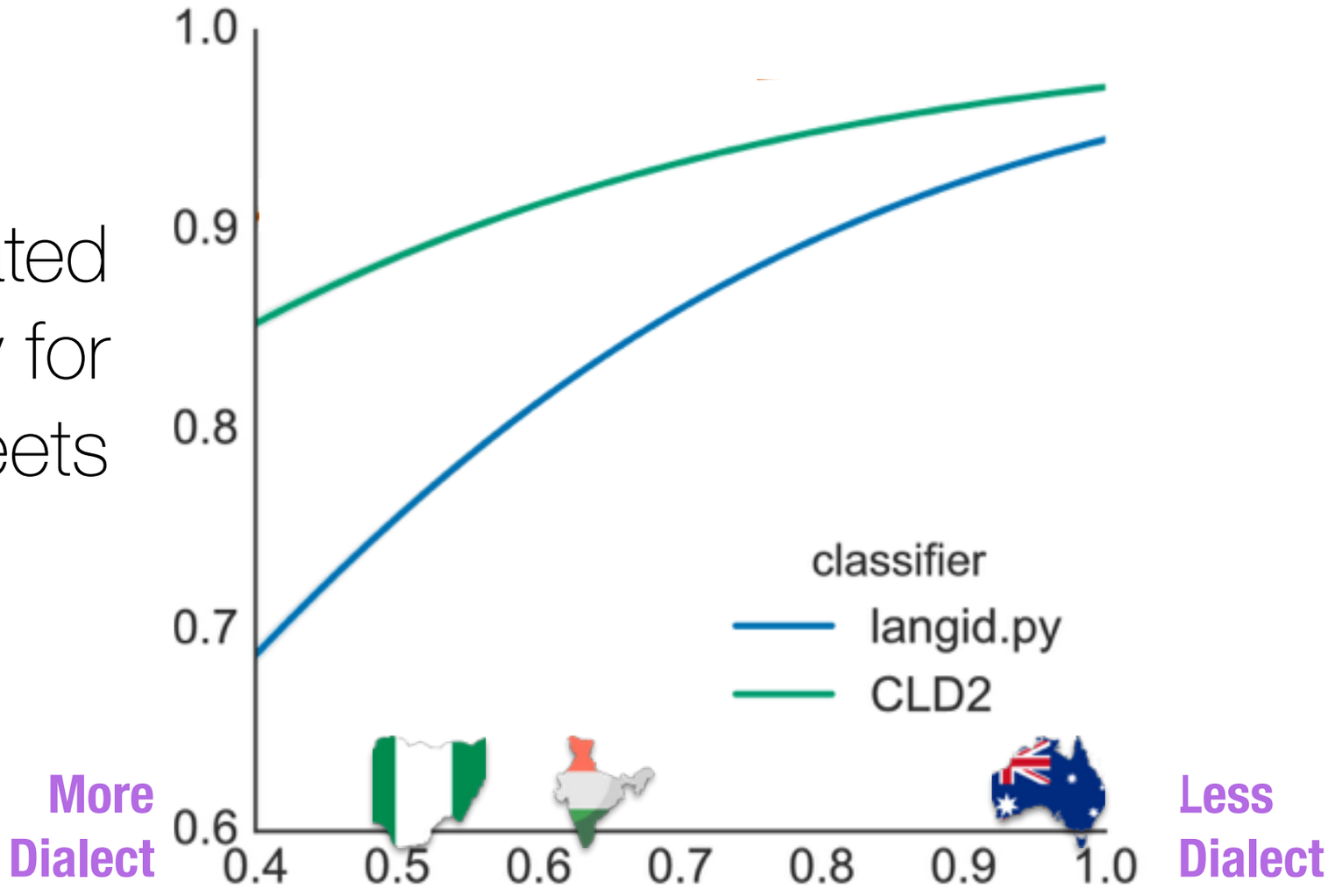
NLP
Which symptoms?

non-English?


Failing to recognize a language silences its speakers’ voices

Current language detection methods perform significantly worse in less-developed countries

Estimated accuracy for English tweets

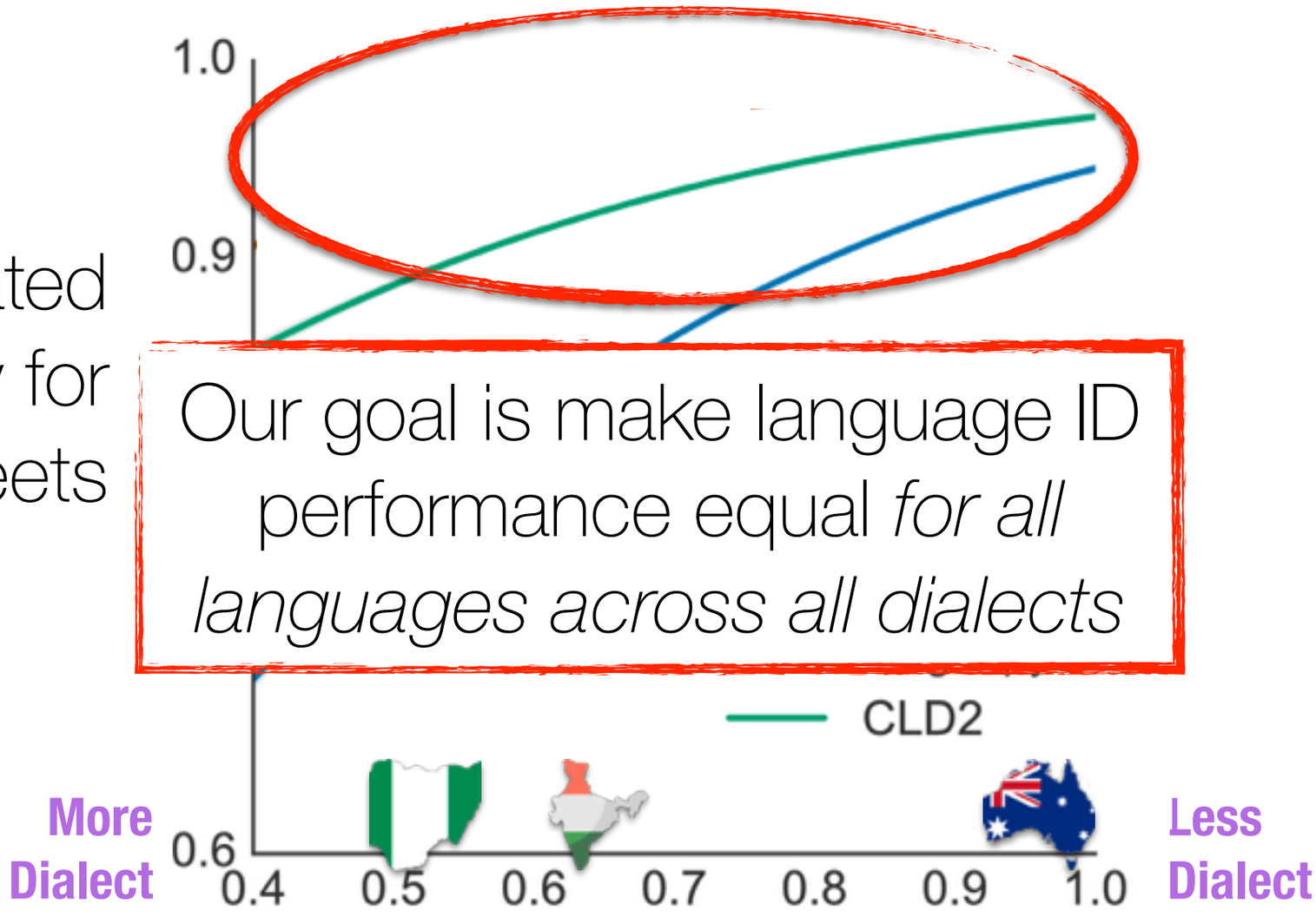


Human Development Index of text's origin country

(Labov, 1964; Ash, 2002)

Current language detection methods perform significantly worse in less-developed countries

Estimated accuracy for English tweets



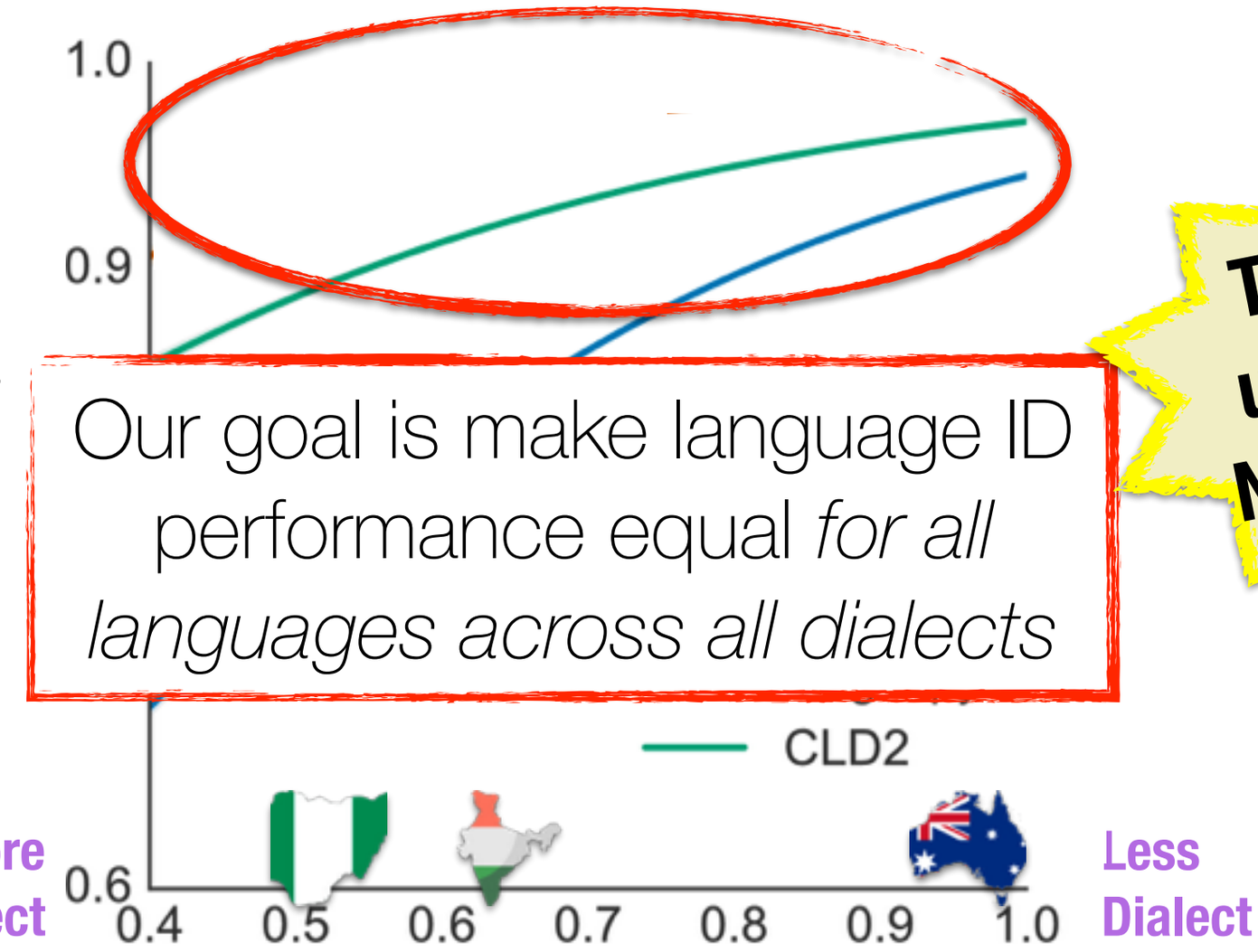
Our goal is make language ID performance equal *for all languages across all dialects*

Human Development Index of text's origin country

(Labov, 1964; Ash, 2002)

Current language detection methods perform significantly worse in less-developed countries

Estimated accuracy for English tweets



Our goal is make language ID performance equal *for all languages across all dialects*

This is a universal NLP issue!

Human Development Index of text's origin country

(Labov, 1964; Ash, 2002)

Key Problems: Current methods struggle in the global setting because

Key Problems: Current methods struggle in the global setting because

Data: No corpora that captures global variation in lexicon and dialect

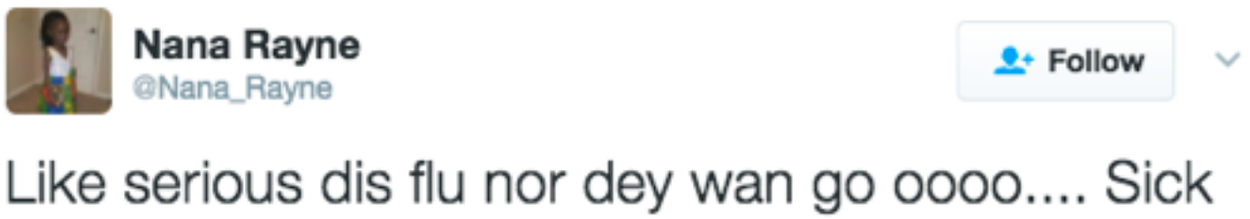
Nana Rayne
@Nana_Rayne

Follow

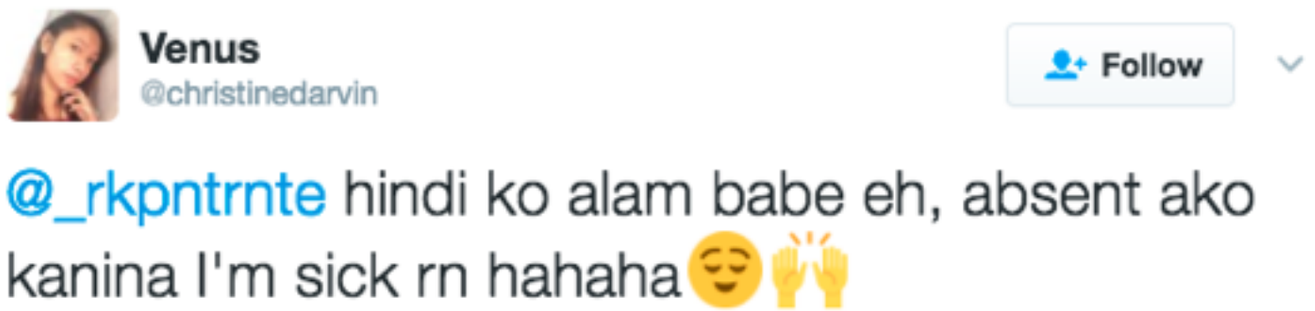
Like serious dis flu nor dey wan go oooo.... Sick

Key Problems: Current methods struggle in the global setting because

Data: No corpora that captures global variation in lexicon and dialect

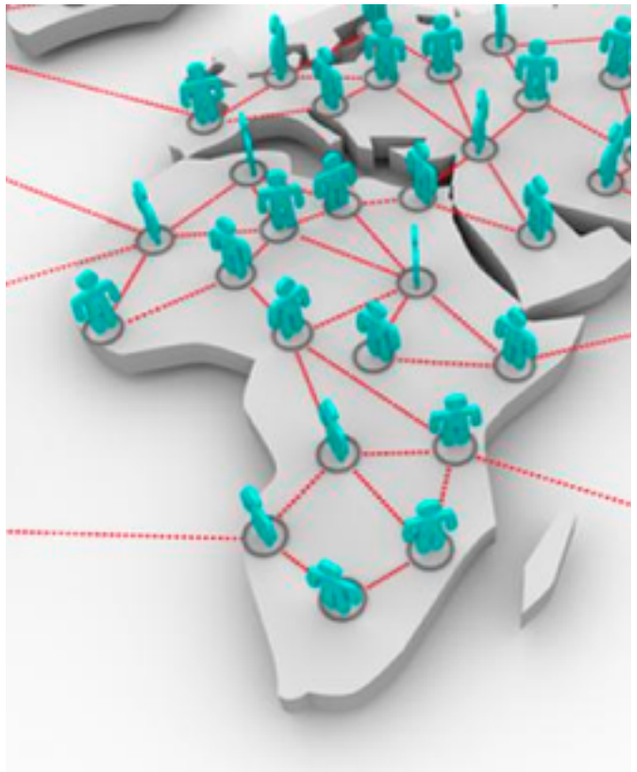


Model: makes simplistic assumptions about how multilinguals communicate

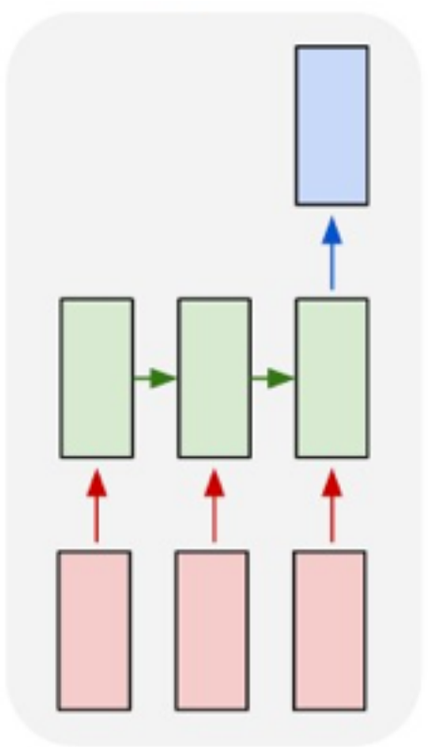


Our approach

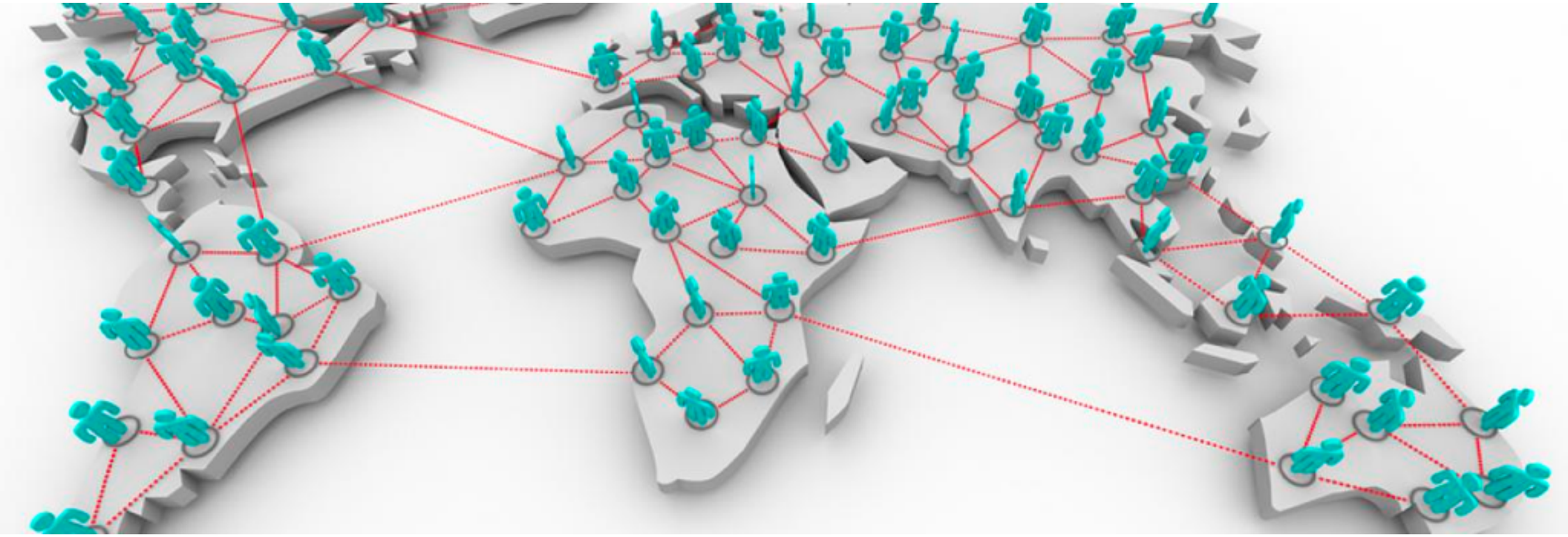
Better **social representation** through network-based sampling



NLP methodologies capable of handling **linguistic variation**



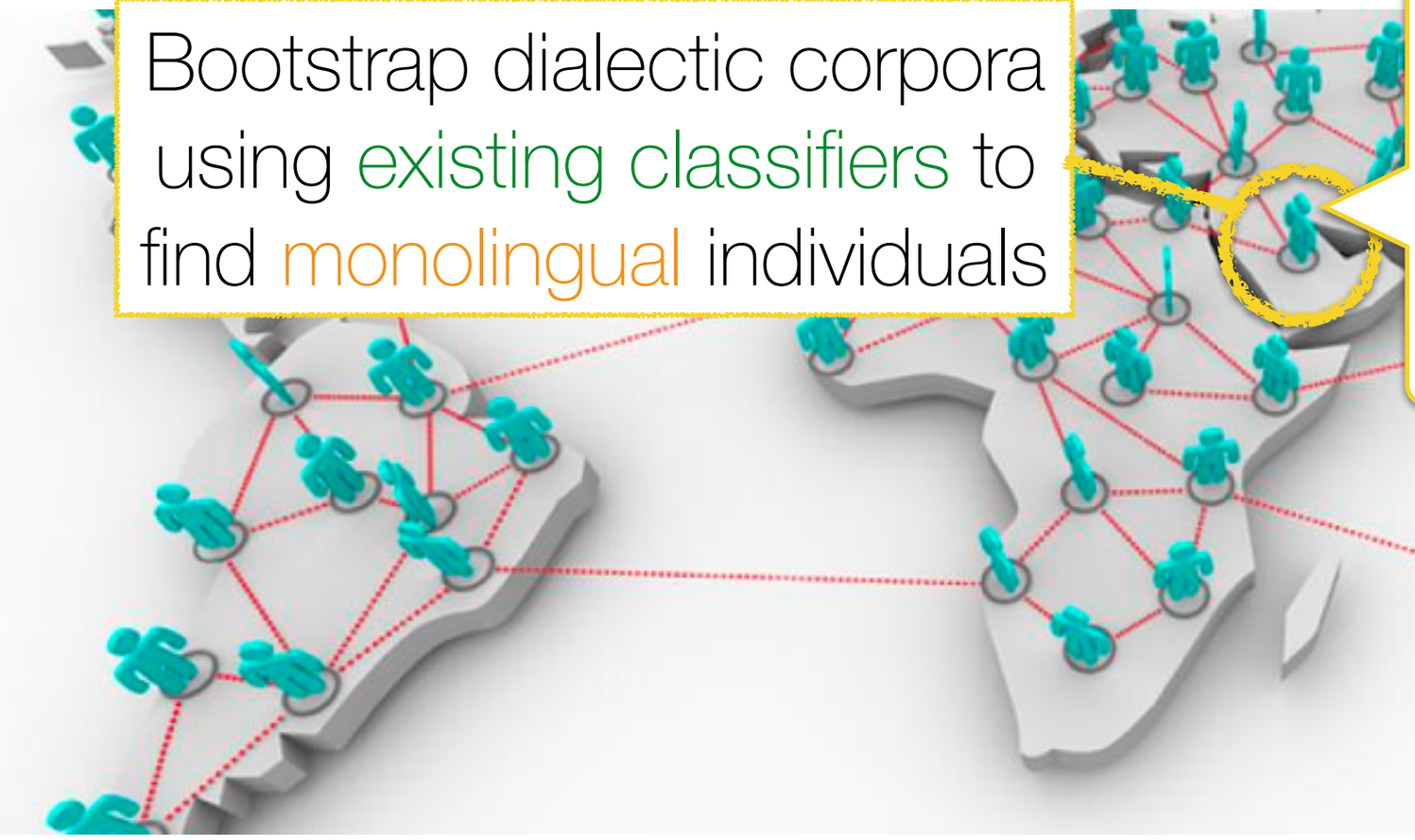
Our Data Solution: Improve linguistic representation through network-based sampling

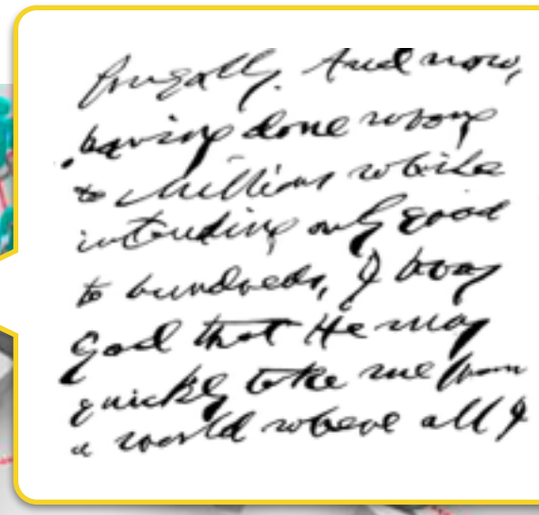


Our Data Solution: Improve linguistic representation through network-based sampling

Our Data Solution: Improve linguistic representation through network-based sampling

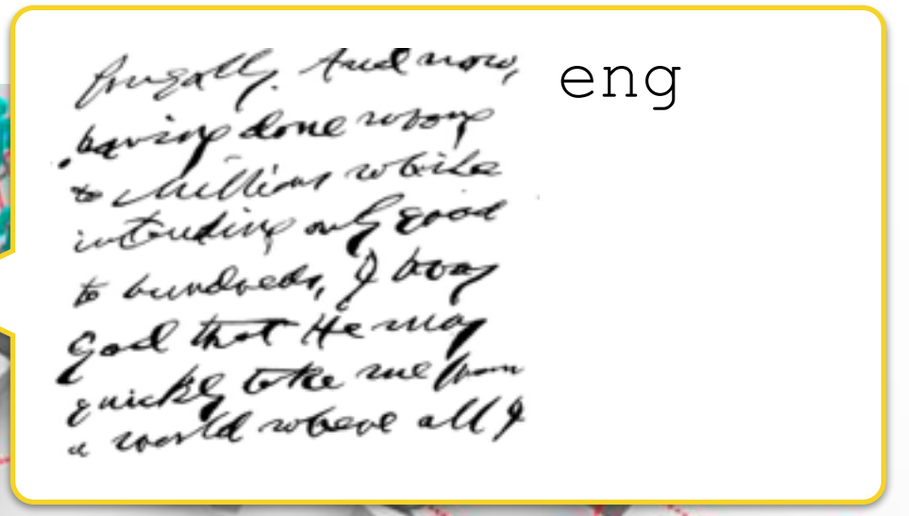
Bootstrap dialectic corpora using existing classifiers to find monolingual individuals





Our Data Solution: Improve linguistic representation through network-based sampling

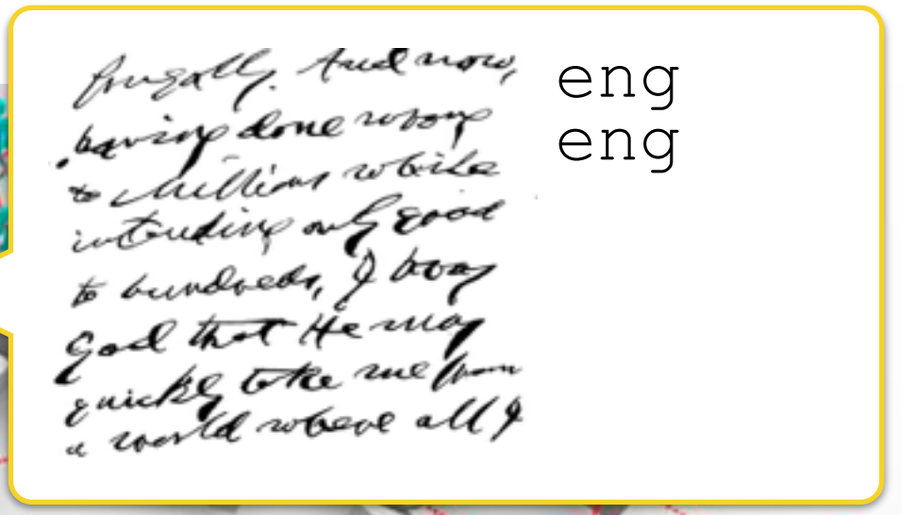
Bootstrap dialectic corpora using existing classifiers to find monolingual individuals



eng

Our Data Solution: Improve linguistic representation through network-based sampling

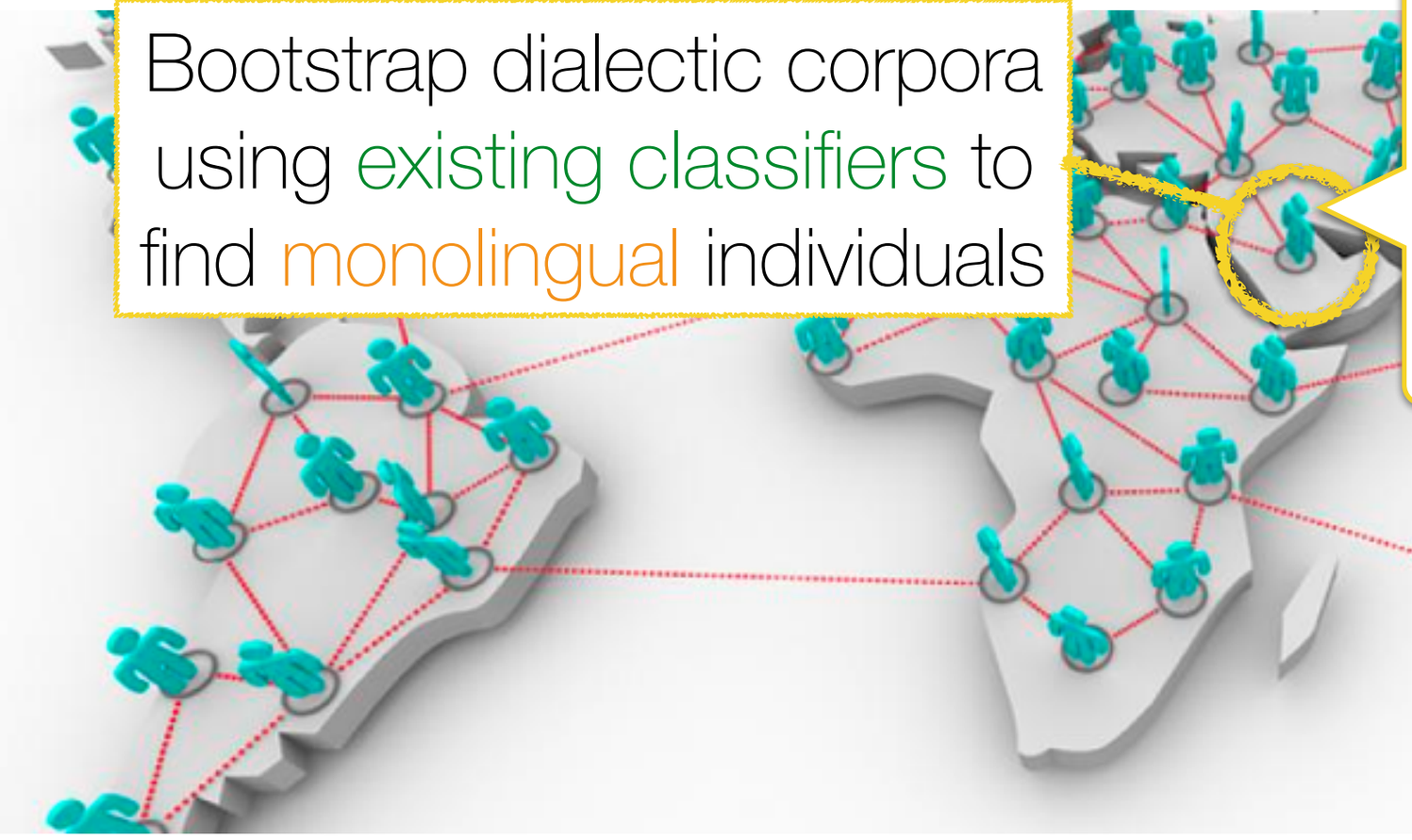
Bootstrap dialectic corpora using existing classifiers to find monolingual individuals

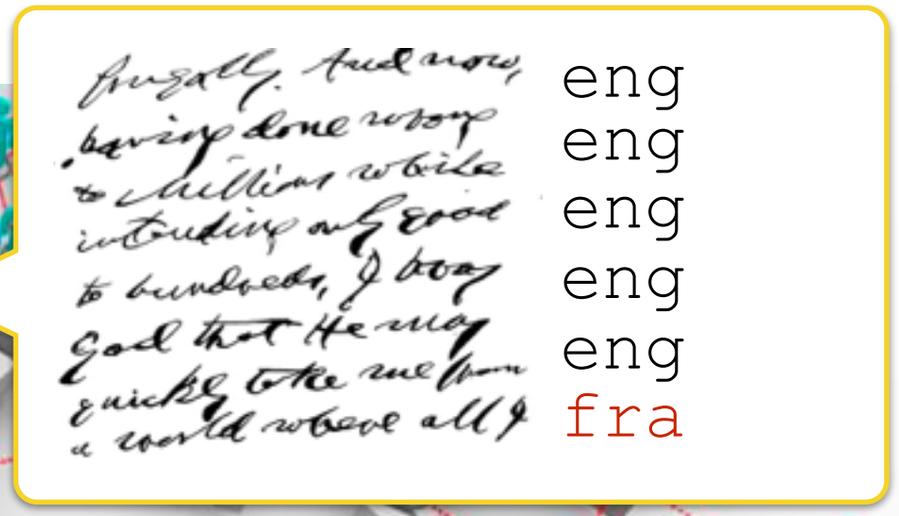


eng
eng

Our Data Solution: Improve linguistic representation through network-based sampling

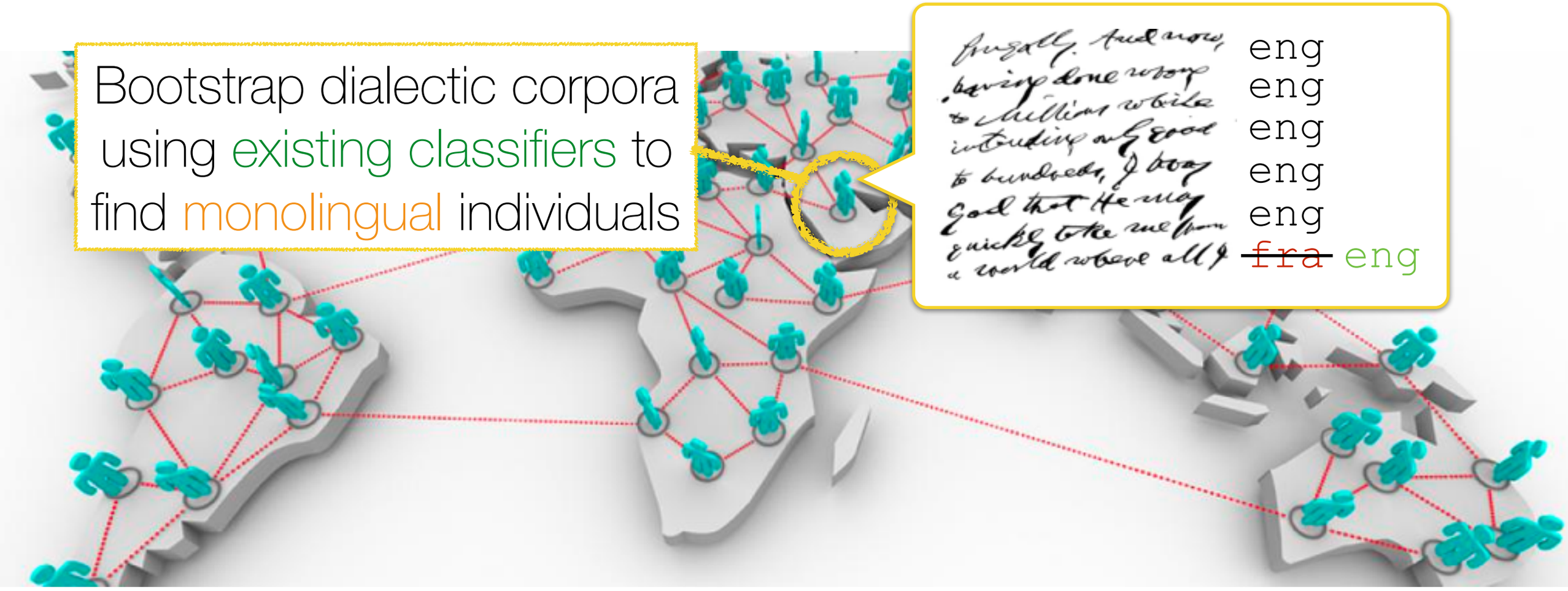
Bootstrap dialectic corpora using existing classifiers to find monolingual individuals





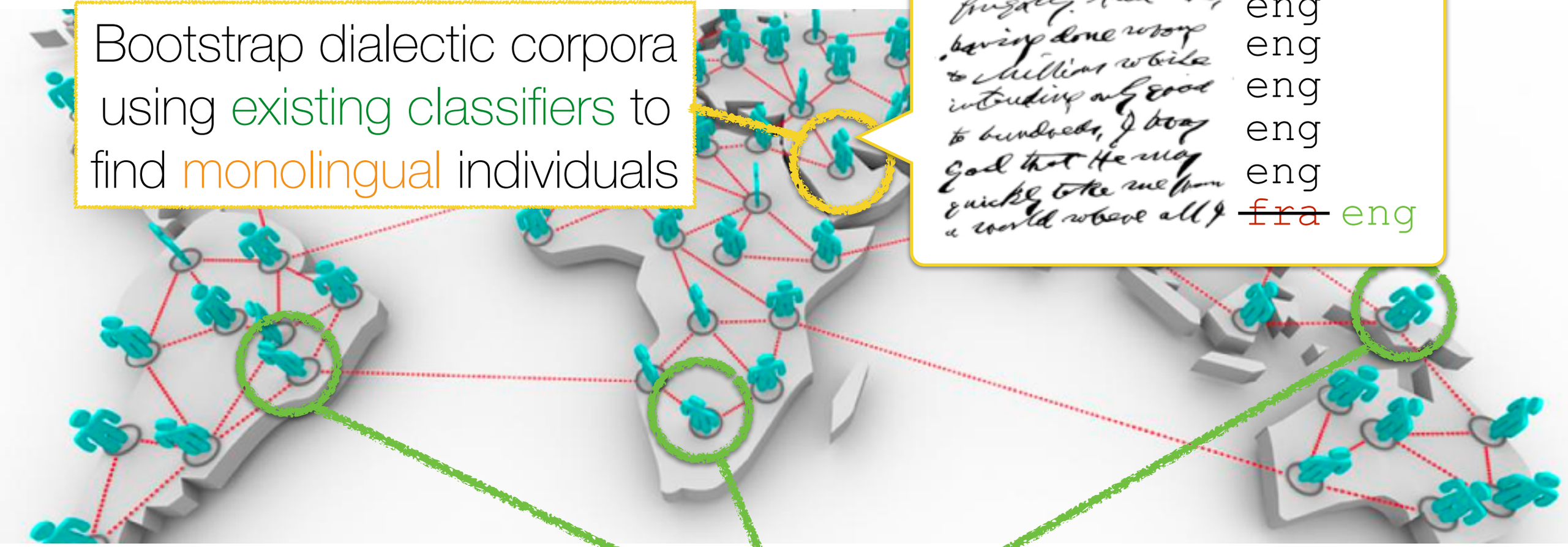
eng
eng
eng
eng
eng
fra

Our Data Solution: Improve linguistic representation through network-based sampling



Our Data Solution: Improve linguistic representation through network-based sampling

Bootstrap dialectic corpora using existing classifiers to find monolingual individuals



	eng
	eng
	eng
	eng
	eng
	~~fra~~ eng

Sample from the geolocated Twitter social network to include text from people at all locations

Build a strategically-diverse corpora

Build a strategically-diverse corpora

Topical

Build a strategically-diverse corpora

Topical

Geographic

Build a strategically-diverse corpora

Topical

Geographic

Social

Build a strategically-diverse corpora and synthesize code-switched examples

Topical

Geographic

Social

Multilingual

Our model solution: treat language identification as a character-based sequence to sequence task.

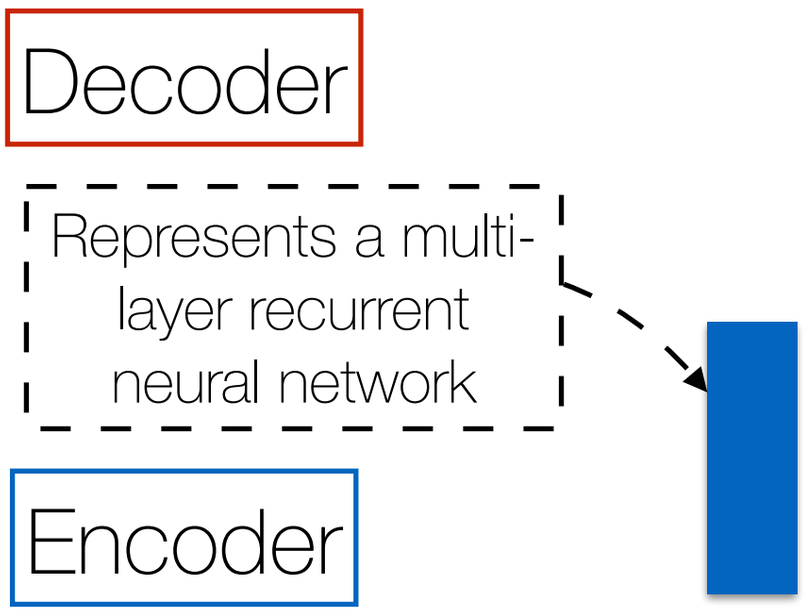
Decoder

Encoder

Je vais commander à emporter. I'm too lazy to cook.

Jaech et al. 2016; Samih et al. 2016

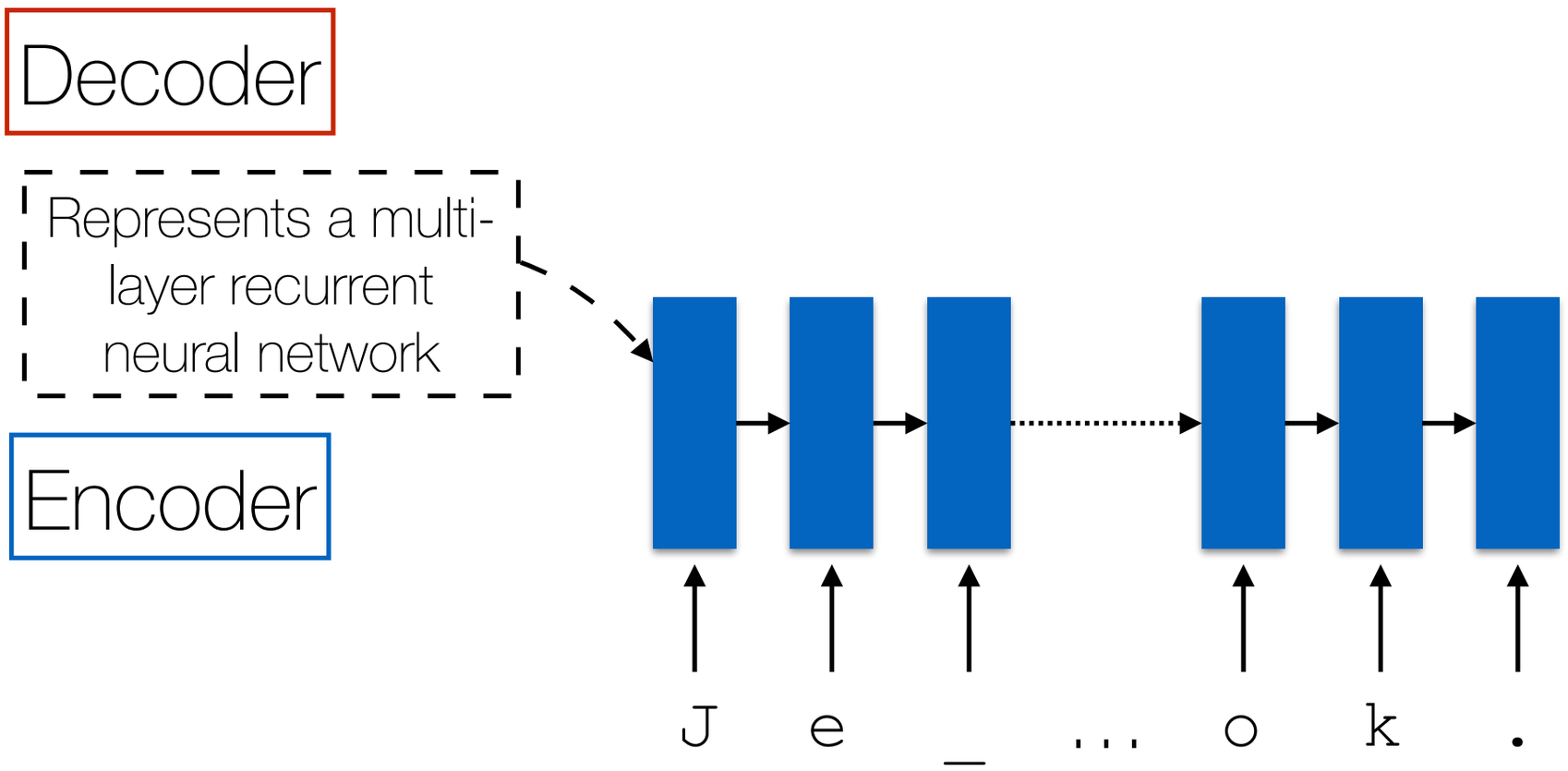
Our model solution: treat language identification as a character-based sequence to sequence task.



Je vais commander à emporter. I'm too lazy to cook.

Jaech et al. 2016; Samih et al. 2016

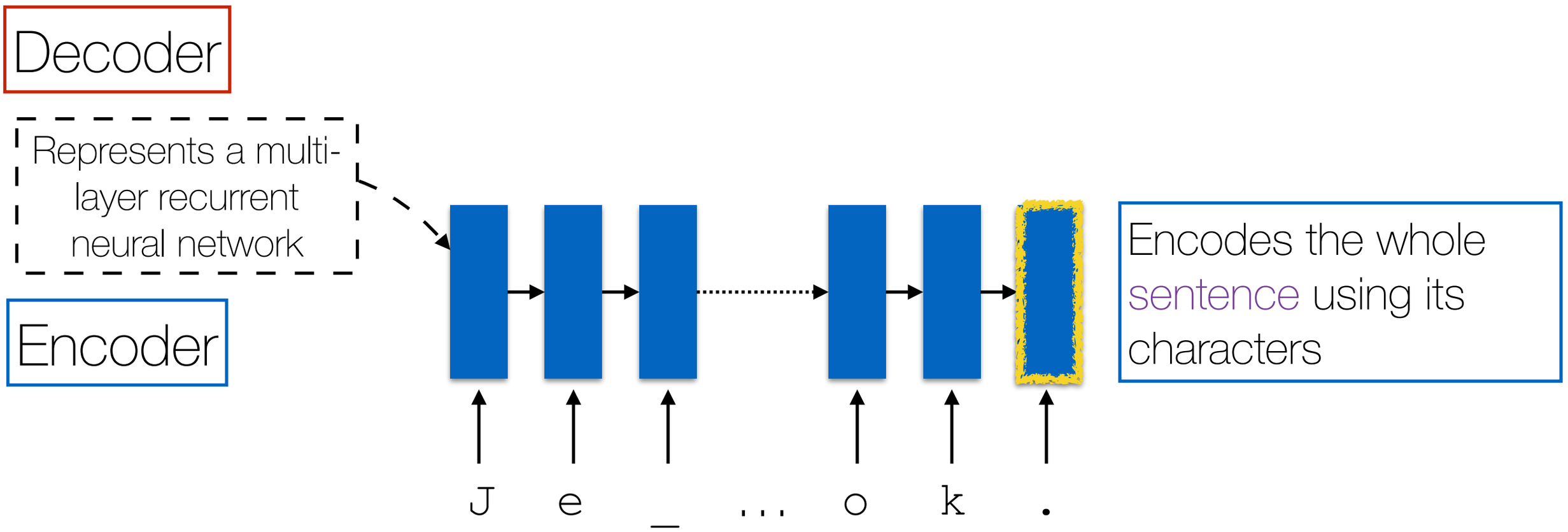
Our model solution: treat language identification as a character-based sequence to sequence task.



Je vais commander à emporter. I'm too lazy to cook.

Jaech et al. 2016; Samih et al. 2016

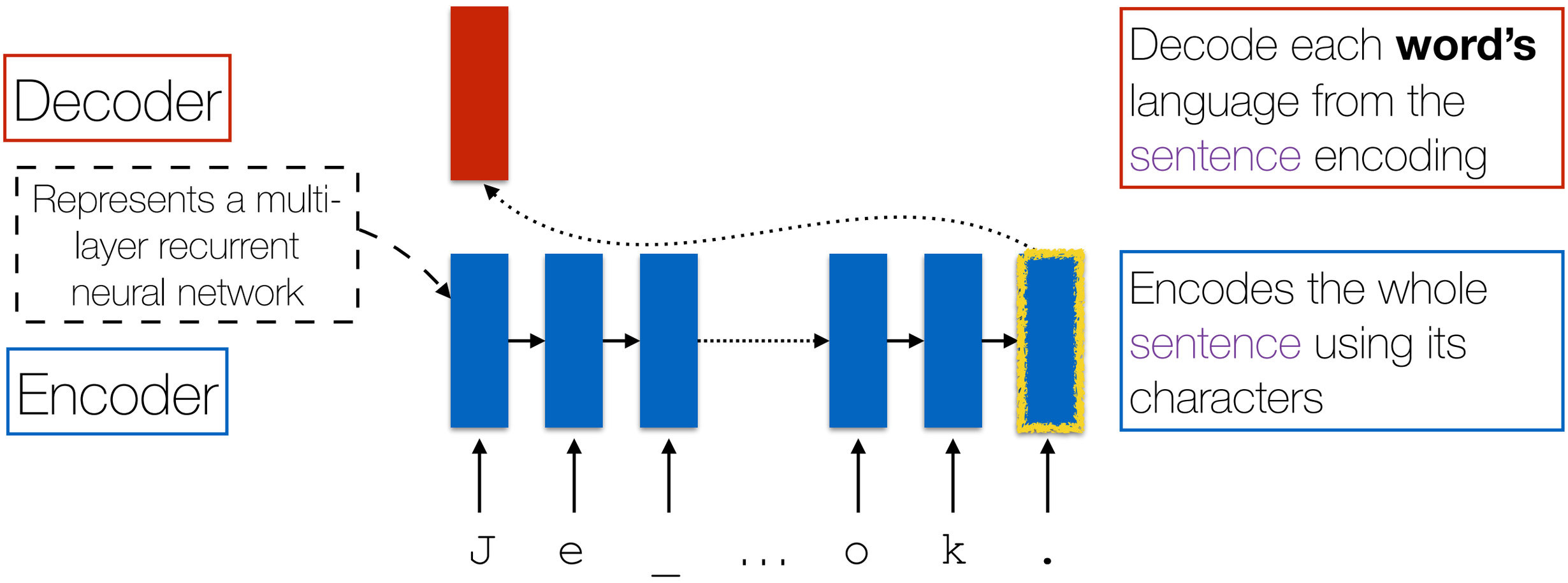
Our model solution: treat language identification as a character-based sequence to sequence task.



Je vais commander à emporter. I'm too lazy to cook.

Jaech et al. 2016; Samih et al. 2016

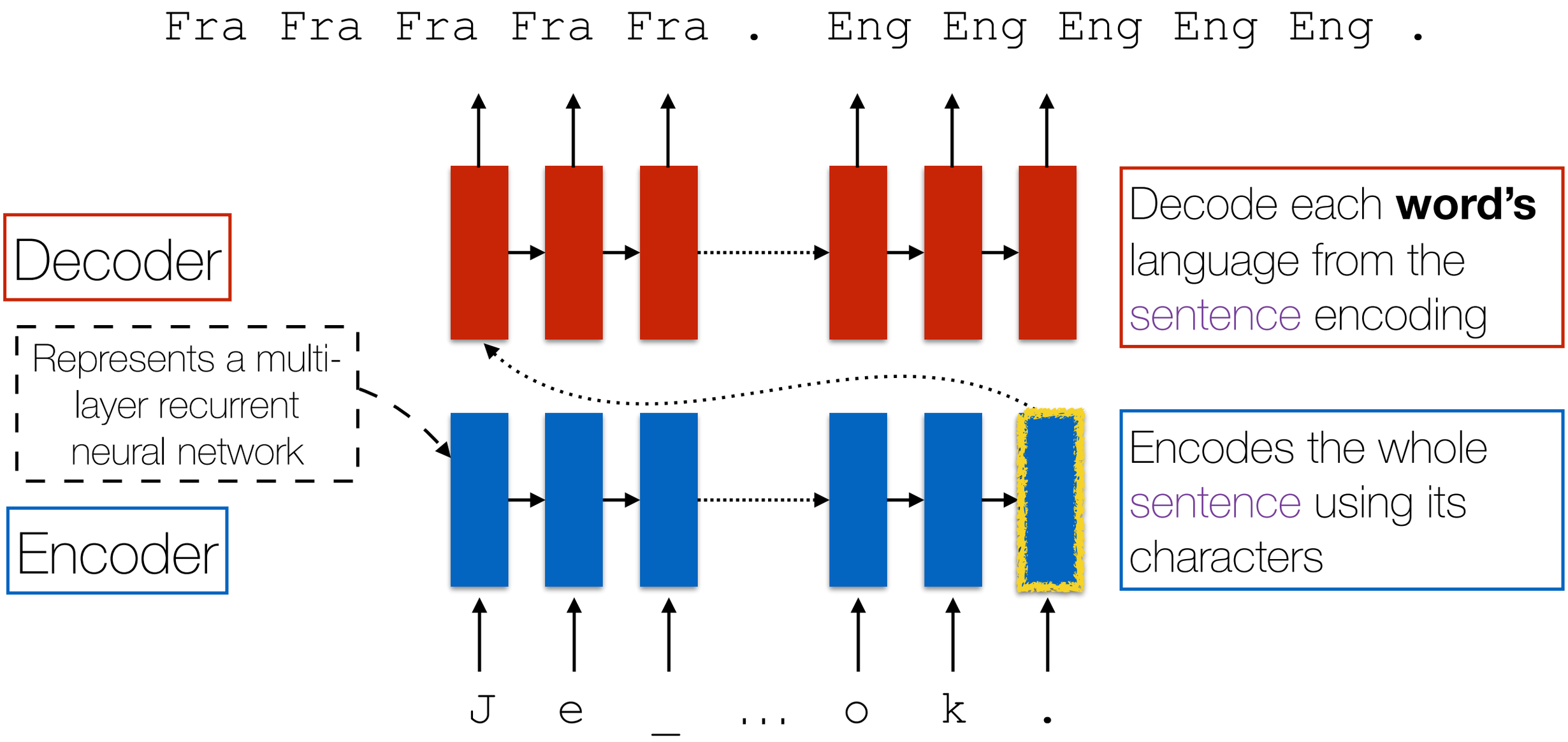
Our model solution: treat language identification as a character-based sequence to sequence task.



Je vais commander à emporter. I'm too lazy to cook.

Jaech et al. 2016; Samih et al. 2016

Our model solution: treat language identification as a character-based sequence to sequence task.



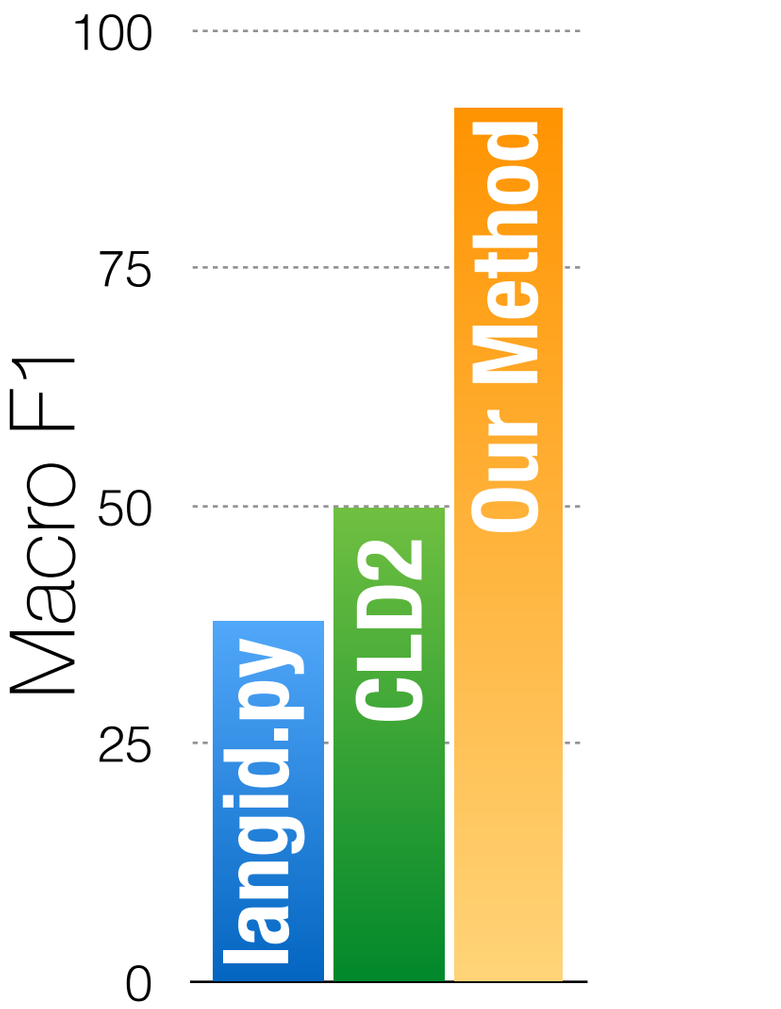
Je vais commander à emporter. I'm too lazy to cook.

Jaech et al. 2016; Samih et al. 2016

Equilid vs off-the-shelf

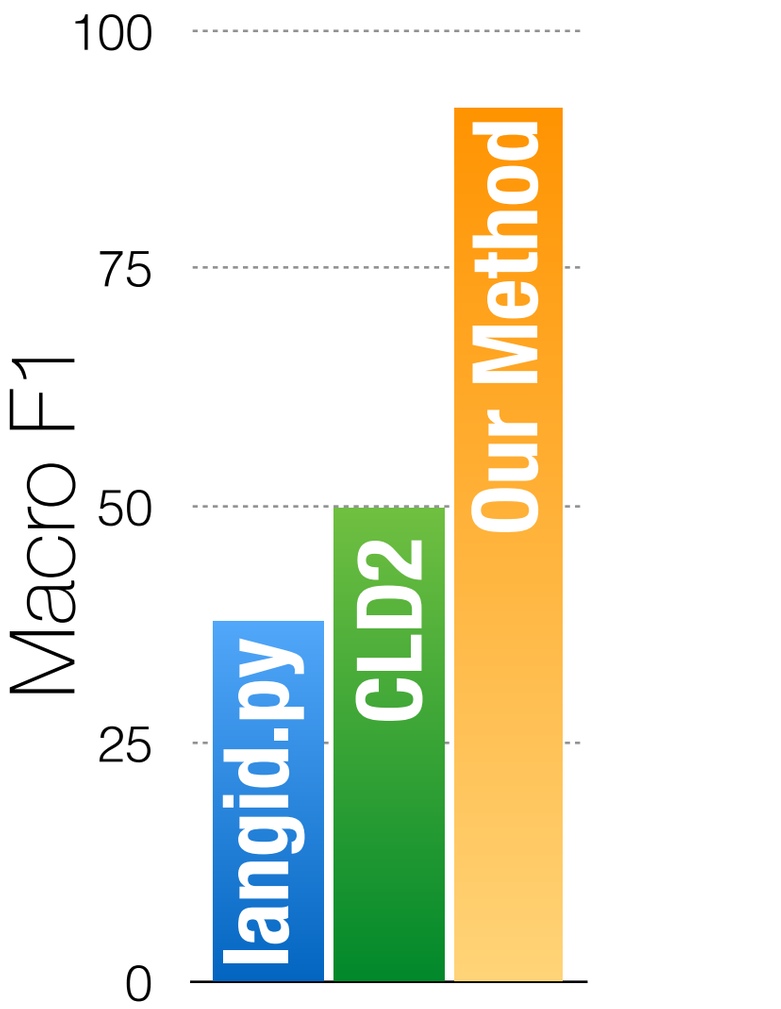
Lui et al. 2013, 2014

Equilid vs off-the-shelf

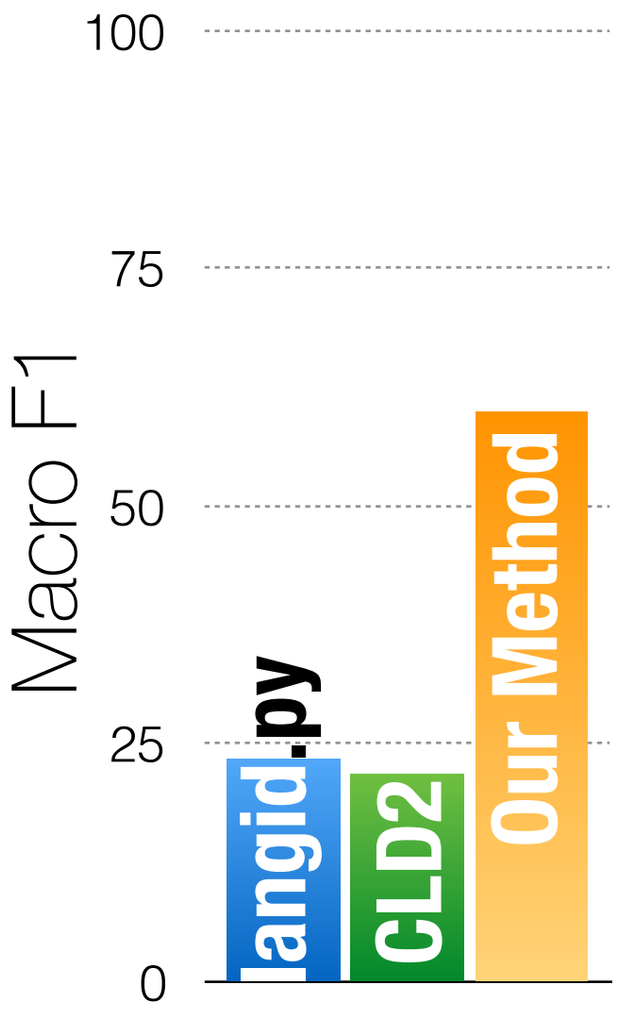


Lui et al. 2013, 2014

Equilid vs off-the-shelf



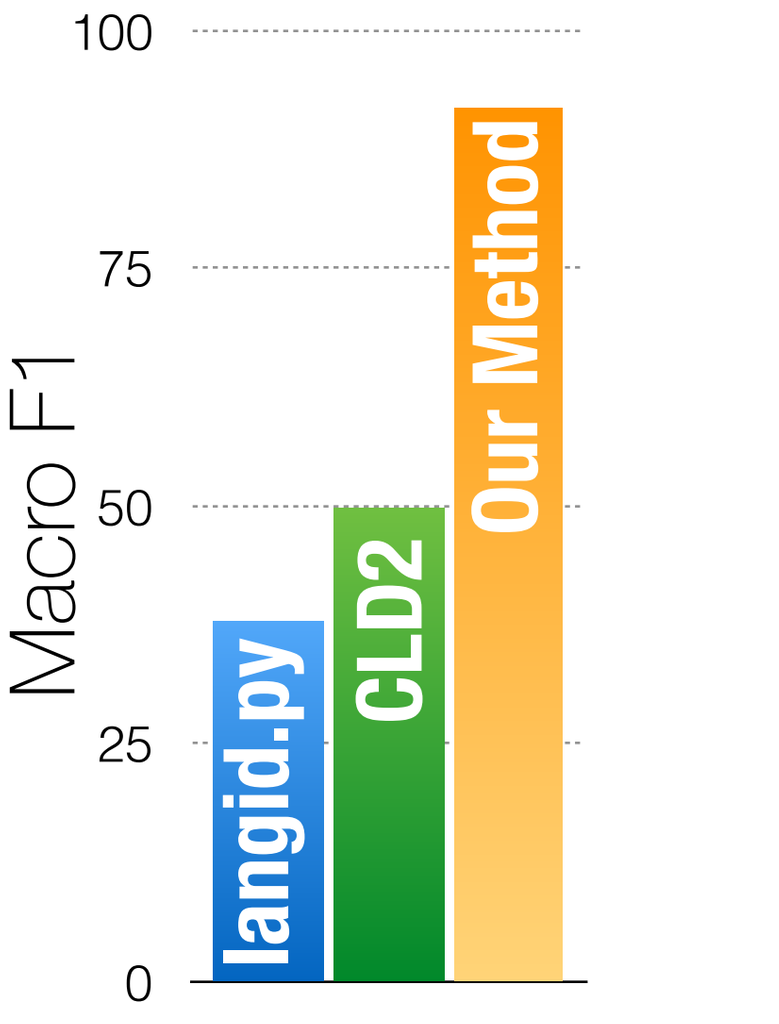
70 Languages on Twitter



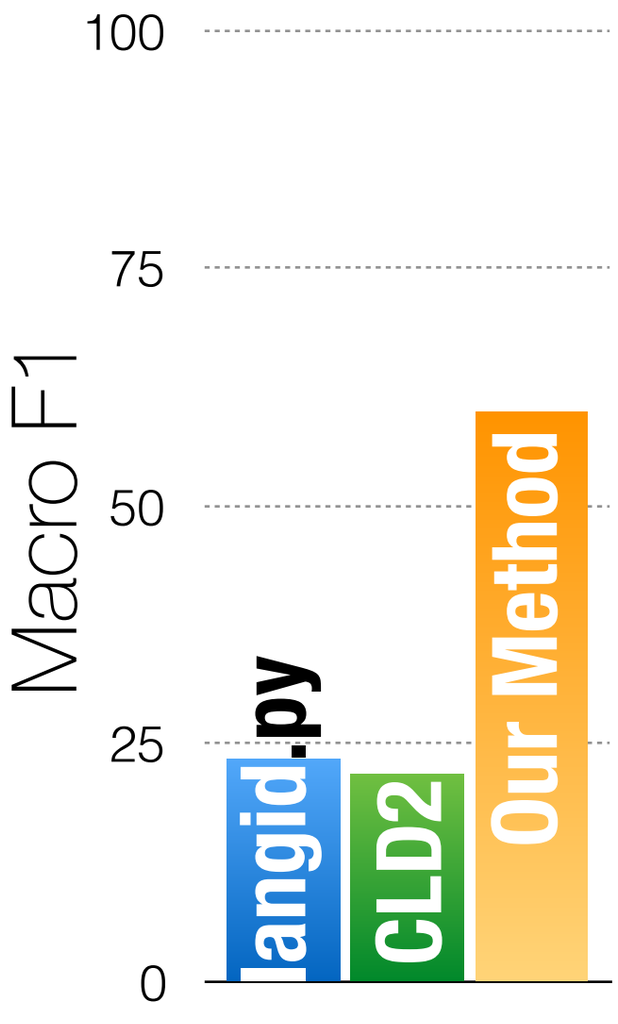
Geo-diverse Tweets

Lui et al. 2013, 2014

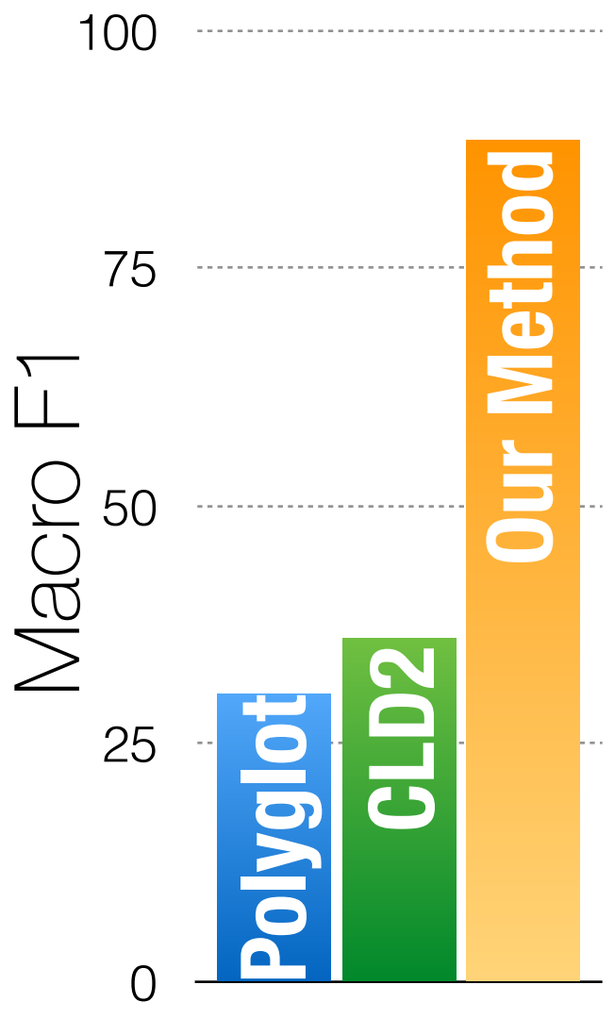
Equilid vs off-the-shelf



70 Languages on Twitter



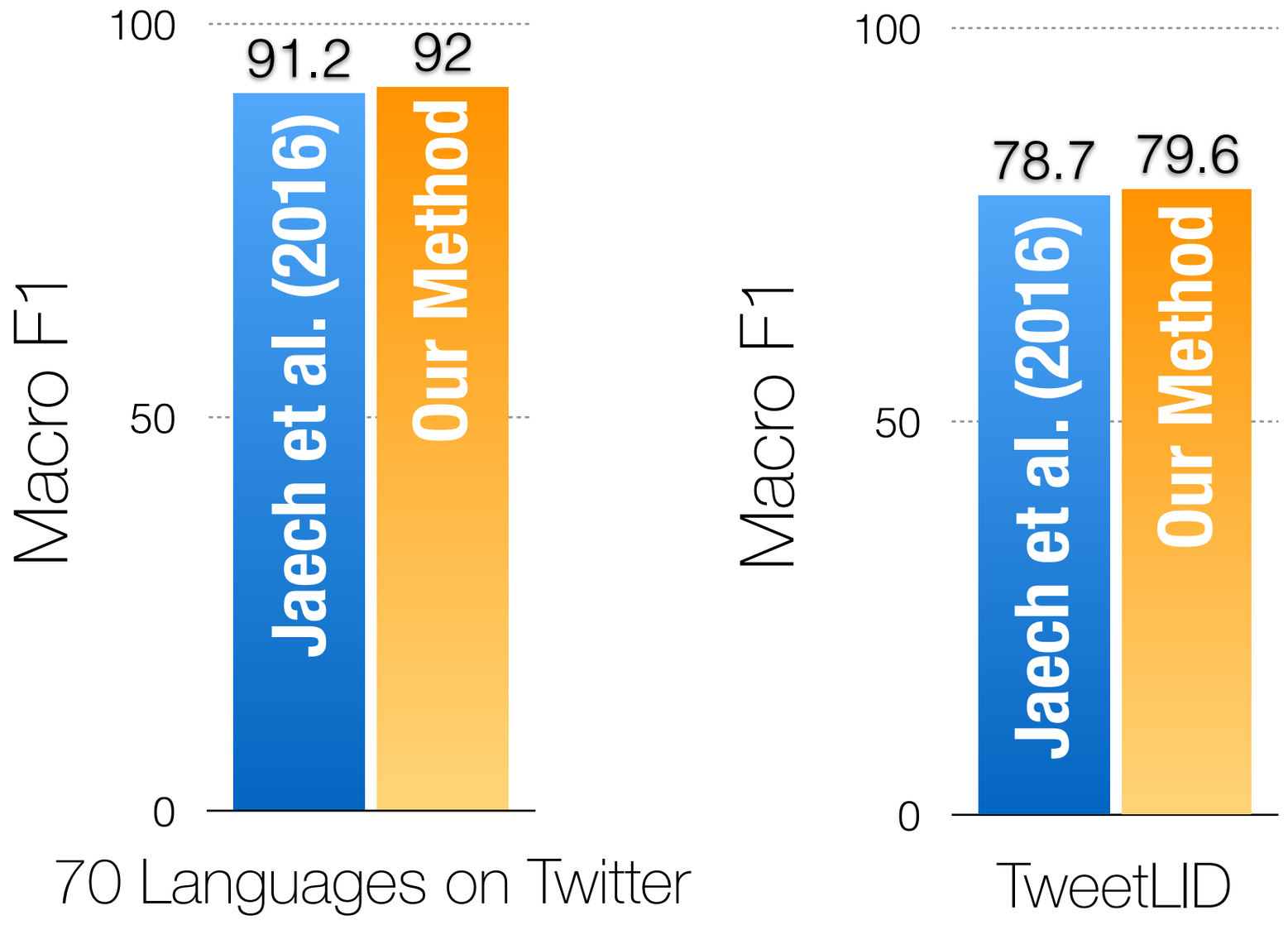
Geo-diverse Tweets



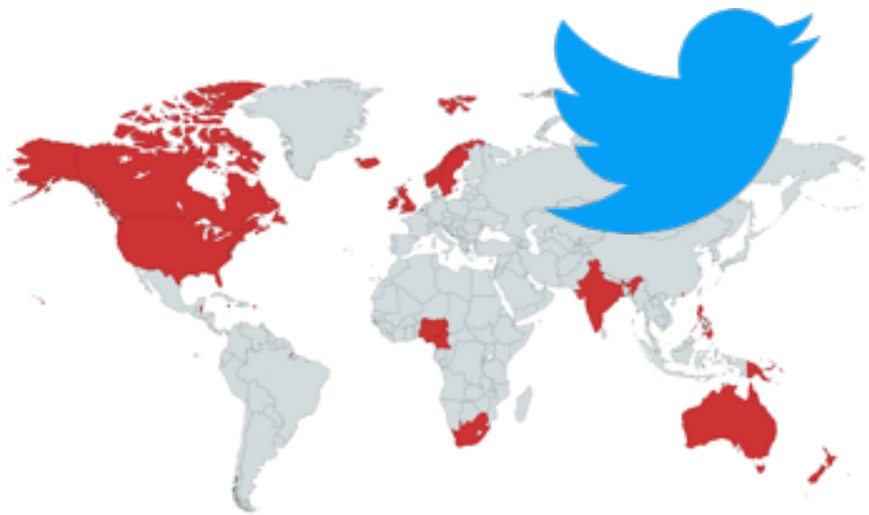
Multilingual Tweets

Lui et al. 2013, 2014

Equilid even outperforms system specifically tuned for each dataset

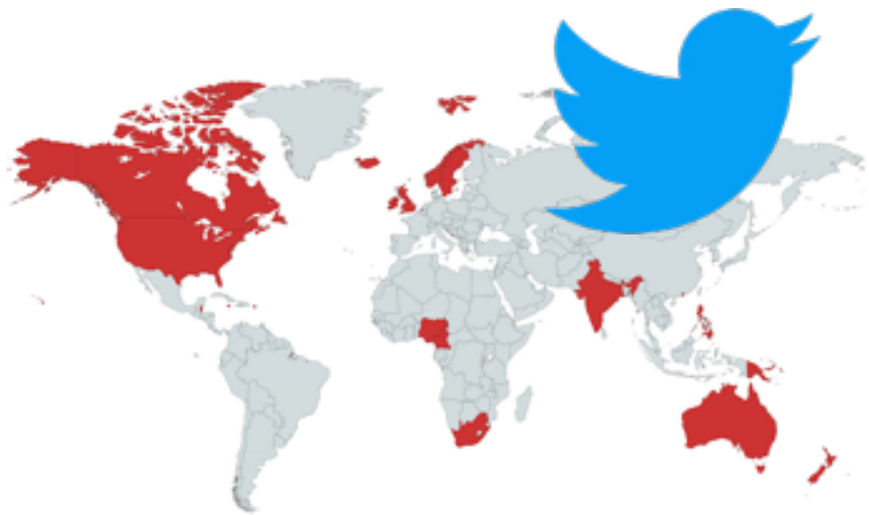


Case Study: Do our solutions provide socially-equitable language identification for health-related queries?



1M Tweets with any of 385 English terms from established lexicons for influenza, psychological well-being, and social health

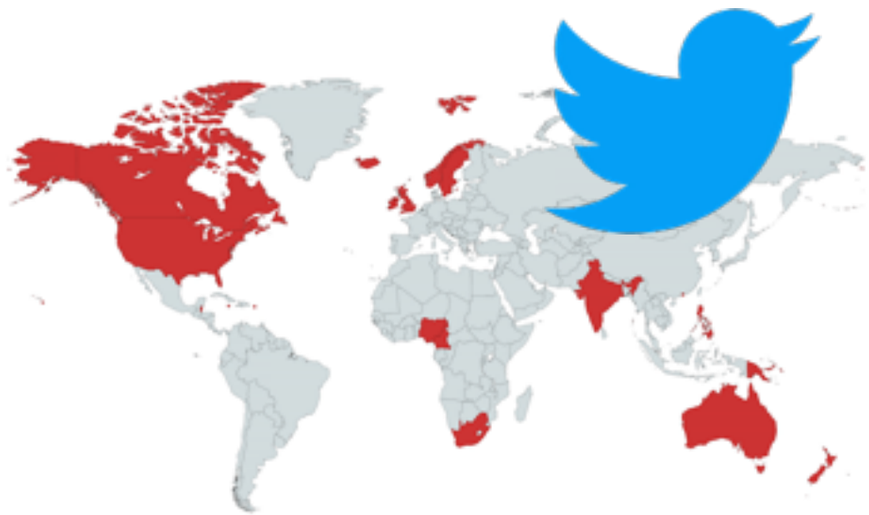
Case Study: Do our solutions provide socially-equitable language identification for health-related queries?



1M Tweets with any of 385 English terms from established lexicons for influenza, psychological well-being, and social health

Lamb et al., (2013); Smith et al., (2016); Preotiuc-Pietro et al., (2015); Park et al., (2016)

Case Study: Do our solutions provide socially-equitable language identification for health-related queries?

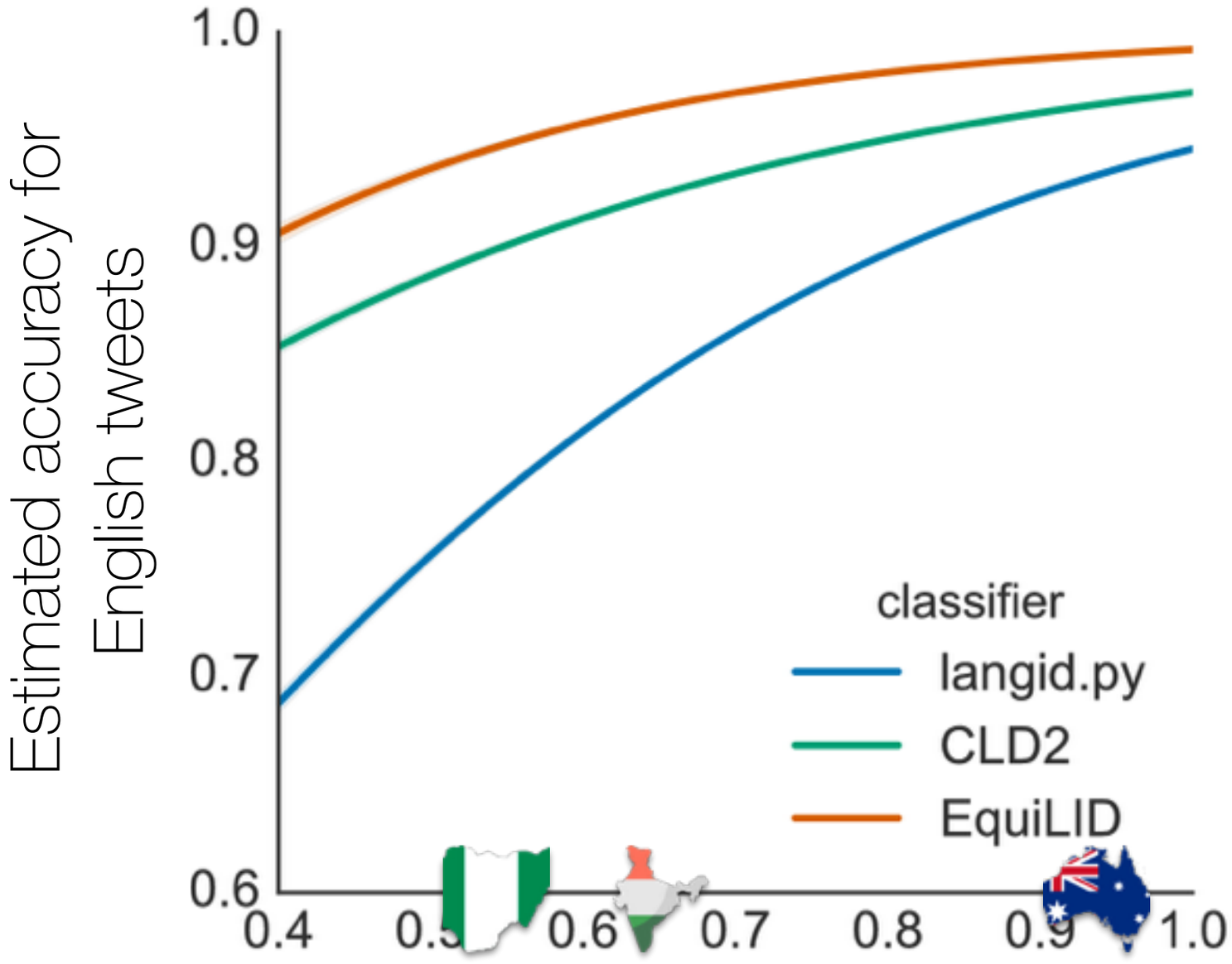


1M Tweets with any of 385 English terms from established lexicons for influenza, psychological well-being, and social health

Task: does the language identification system recognize every tweet as English?

Lamb et al., (2013); Smith et al., (2016); Preotiuc-Pietro et al., (2015); Park et al., (2016)

Equilid raises the bar for socially-equitable language identification



Social Equality doesn’t stop at Language Identification

Better **social representation** in our data

Methodologies capable of handling **language as it is used**

Social Equality doesn’t stop at Language Identification

Better **social representation** in our data

Methodologies capable of handling **language as it is used**

Be equitable!

https://github.com/davidjurgens/equilid

David Jurgens, Yulia Tsvetkov, and Dan Jurafsky