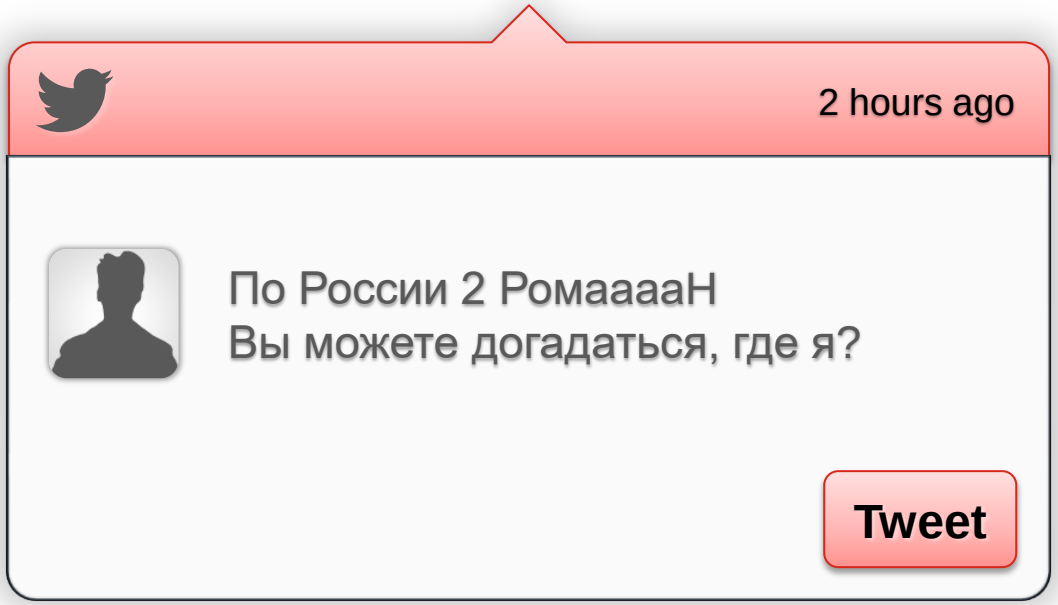
Language Influences on Tweeter Geolocation

Ahmed Mourad

Assoc. Prof. Falk Scholer

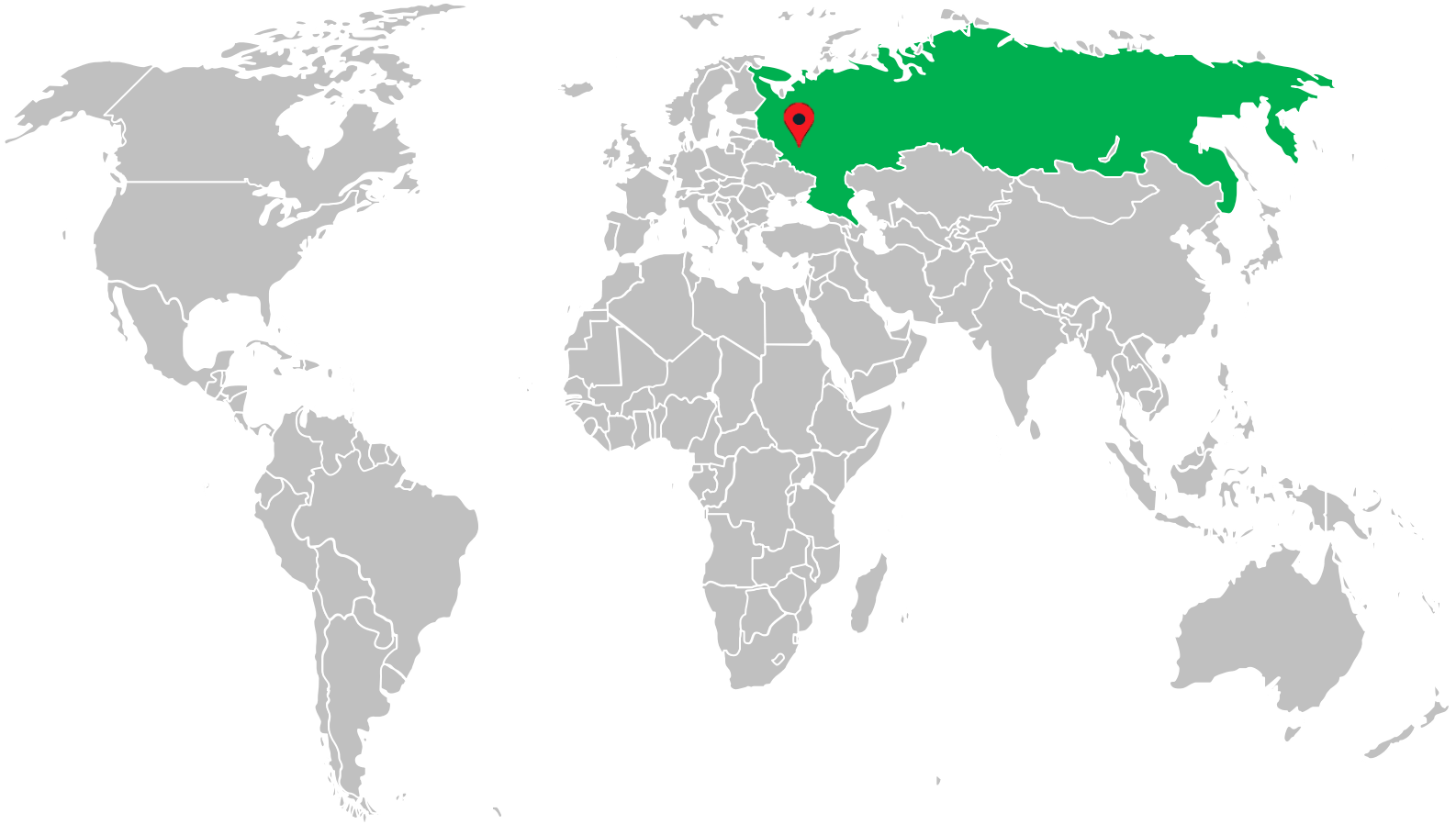
Prof. Mark Sanderson

Where in the world am I?



In Russia 2 RomaaaH
Can you guess where I am?

Russia! Maybe Moscow!



http://www.powerpointslides.net/powerpointgraphics/powerpointmaps.html

Hypothesis

Users tweeting in languages with restricted geographical usage (e.g. Turkish, Arabic, Russian and Indonesian) are much easier to geolocate than languages that are more diverse in usage (e.g. English and Spanish)

Implicit assumption

Spoken languages with restricted geographical usage on the ground will be restricted on Twitter as well

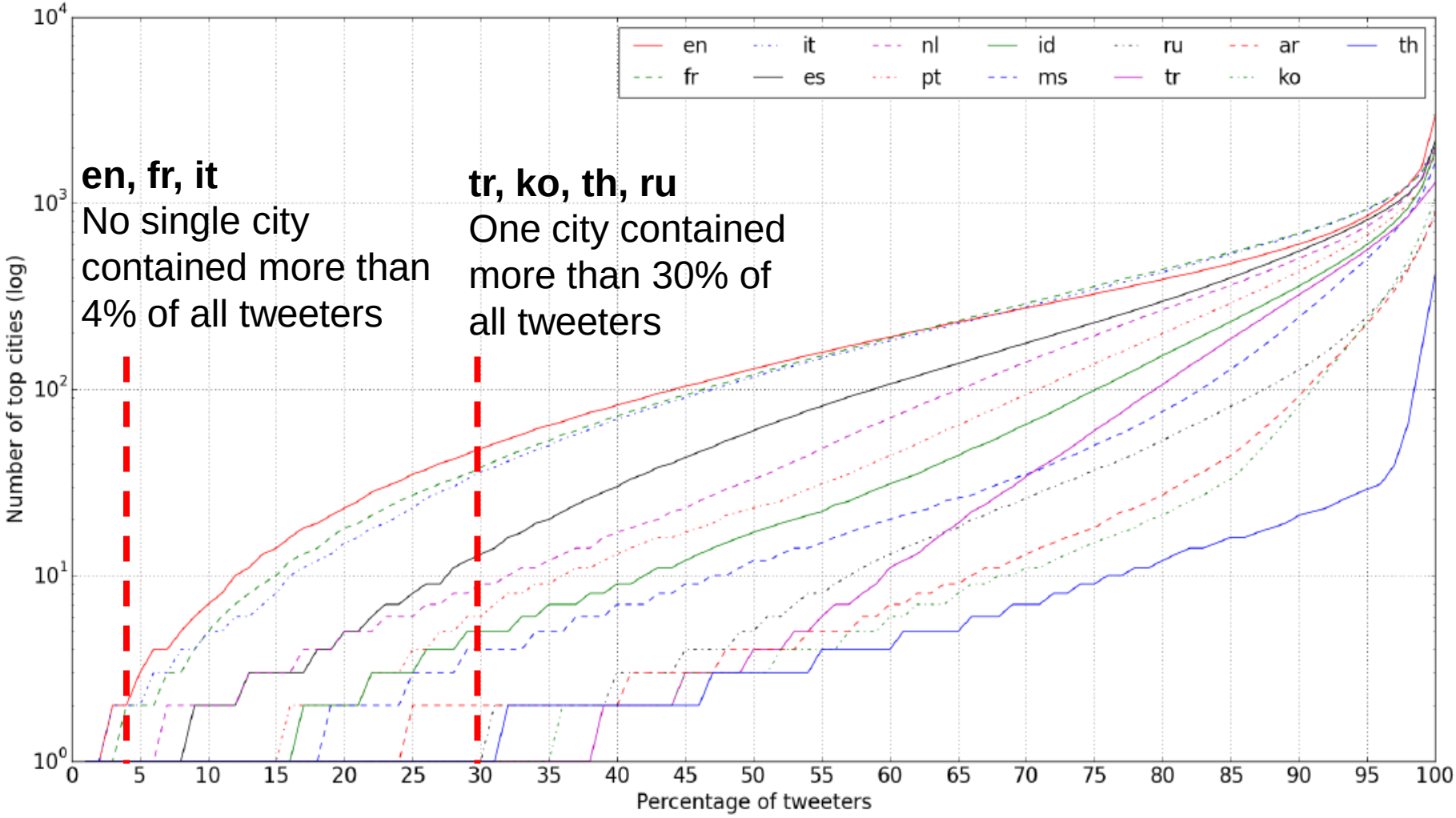
B. Han, P. Cook, and T. Baldwin. Text-based twitter user geolocation prediction. JAIR, 2014.

Every Language is Global

In terms of number of cities (k)

	en	es	it	pt	id	nl	fr	ms	ko	ru	ar	th	tr
WORLD	2.9	2.2	2.1	1.8	1.9	2.0	2.0	1.6	1.1	0.9	0.9	0.4	1.3
TwArchive	3.2	2.3	2.2	1.9	2.0	2.0	2.2	1.7	1.7	1.0	1.6	0.7	1.6

Some Languages are More Imbalanced than Others



Examined Features

- Dataset size

Experimented with different sizes of tweeters per language

- Language

Measured the degree of agreement between the ranking of languages based on accuracy in the collections

- Correlation with individual features

Quantified the influence of 12 different collection properties on the quality of geolocation prediction using common statistical measures

- Alternative measures

Considered alternatives to accuracy to evaluate the impact of imbalance (Precision and Recall using micro and macro averaging)

Thank you
Questions?

See you in the Posters Session! ☺