



NYU

DATA
SCIENCE

Language Use Through The Lens of Big Data

Bruno Gonçalves

www.bgoncalves.com





The Internet:
Where people write on walls
and worship cats

The Internet In Real Time



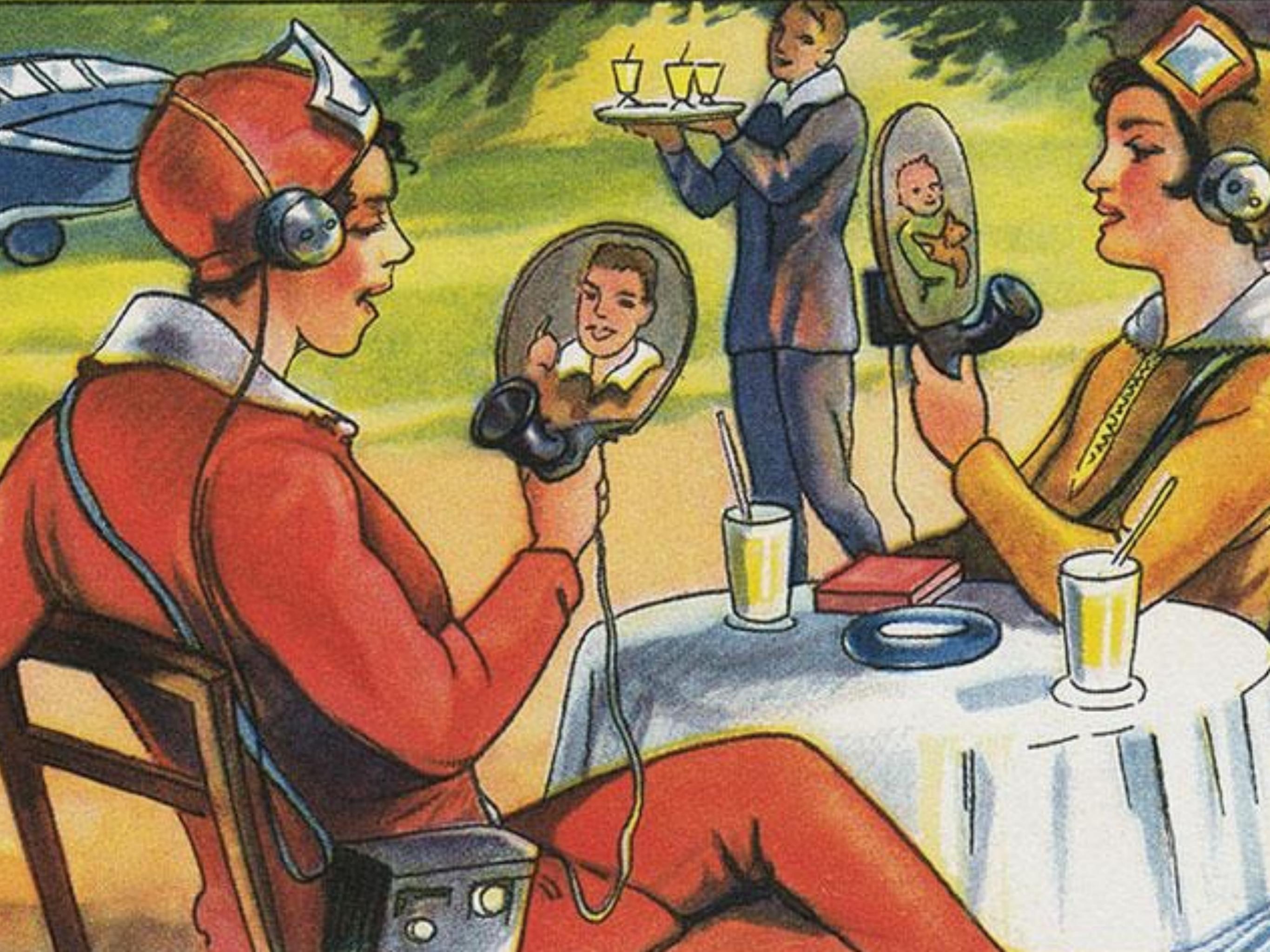
All this technology is making us antisocial



Social Media

SOCIAL MEOWDIA EXPLAINED





JAN
2017

GLOBAL DIGITAL SNAPSHOT

KEY STATISTICAL INDICATORS FOR THE WORLD'S INTERNET, MOBILE, AND SOCIAL MEDIA USERS

TOTAL
POPULATION



7.476
BILLION

URBANISATION:
54%

INTERNET
USERS



3.773
BILLION

PENETRATION:
50%

ACTIVE SOCIAL
MEDIA USERS



2.789
BILLION

PENETRATION:
37%

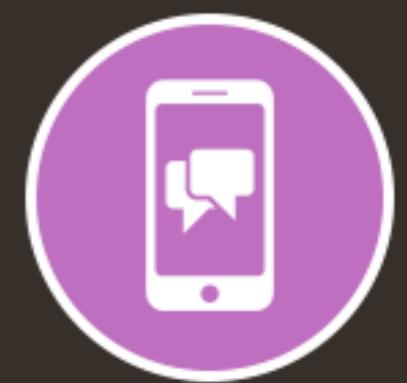
UNIQUE
MOBILE USERS



4.917
BILLION

PENETRATION:
66%

ACTIVE MOBILE
SOCIAL USERS



2.549
BILLION

PENETRATION:
34%

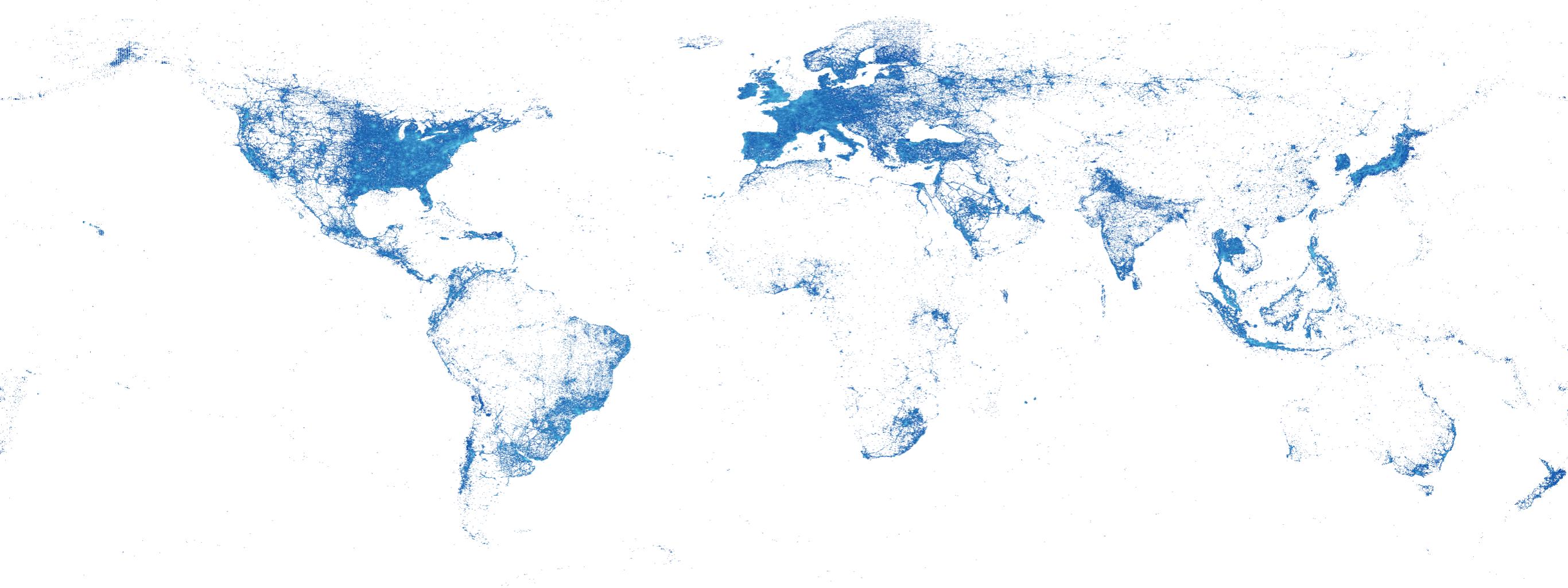
SOURCES: POPULATION: UNITED NATIONS; U.S. CENSUS BUREAU; INTERNET: INTERNETWORLDSTATS; ITU; INTERNETLIVESTATS; CIA WORLD FACTBOOK; FACEBOOK; NATIONAL REGULATORY AUTHORITIES; SOCIAL MEDIA AND MOBILE SOCIAL MEDIA: FACEBOOK; TENCENT; VKONTAKTE; LIVEINTERNET.RU; KAKAO; NAVER; NIKI AGHAEI; CAFEBAZAAR.IR; SIMILARWEB; DING; EXTRAPOLATION OF TNS DATA; MOBILE: GSMA INTELLIGENCE; EXTRAPOLATION OF EMARKETER AND ERICSSON DATA.

 **Hootsuite™**  **we
are.
social**

Geolocated Tweets



GPS Coordinates



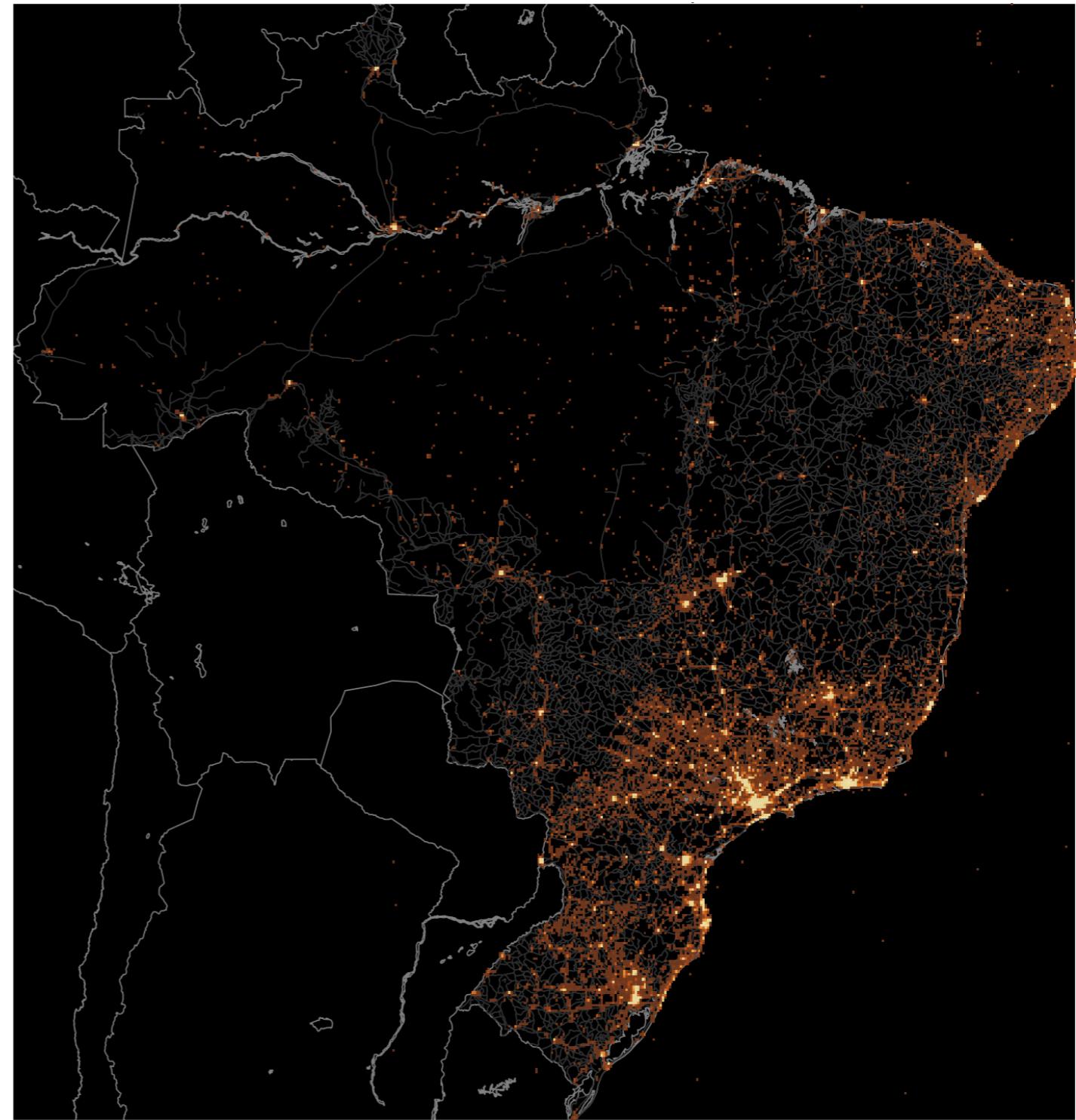
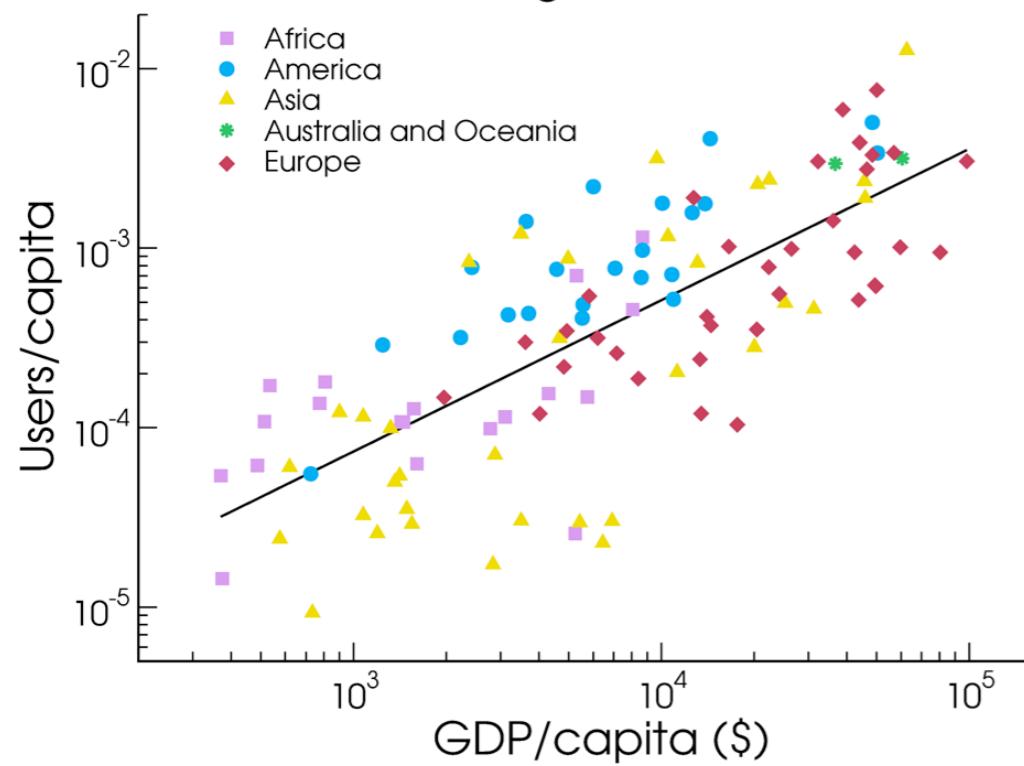
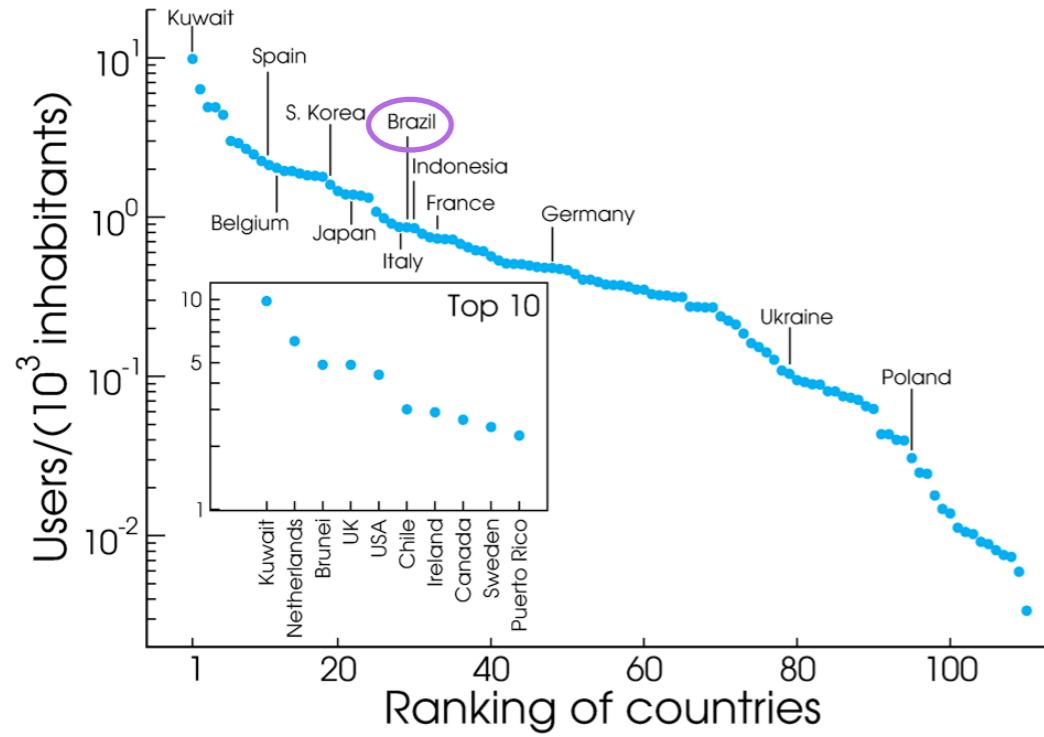
Geolocated Tweets

PLoS One 8, E61981 (2013)

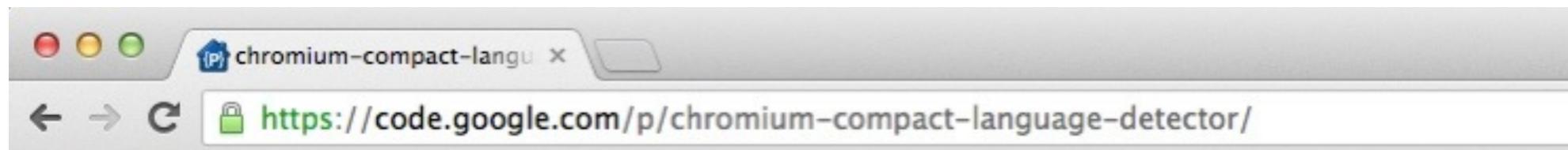
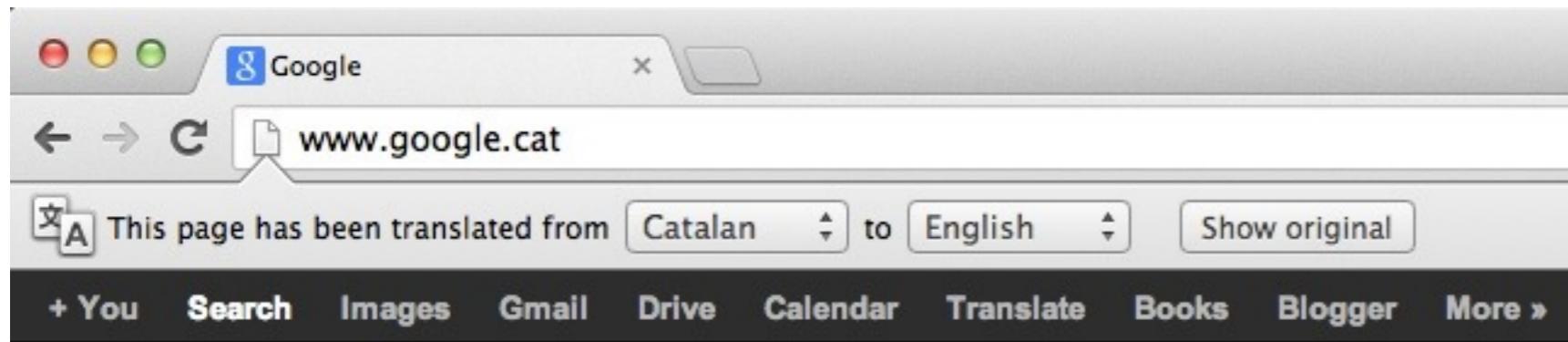


Market Penetration

PLoS One 8, E61981 (2013)



Language Detection



 **chromium-compact-language-detector**

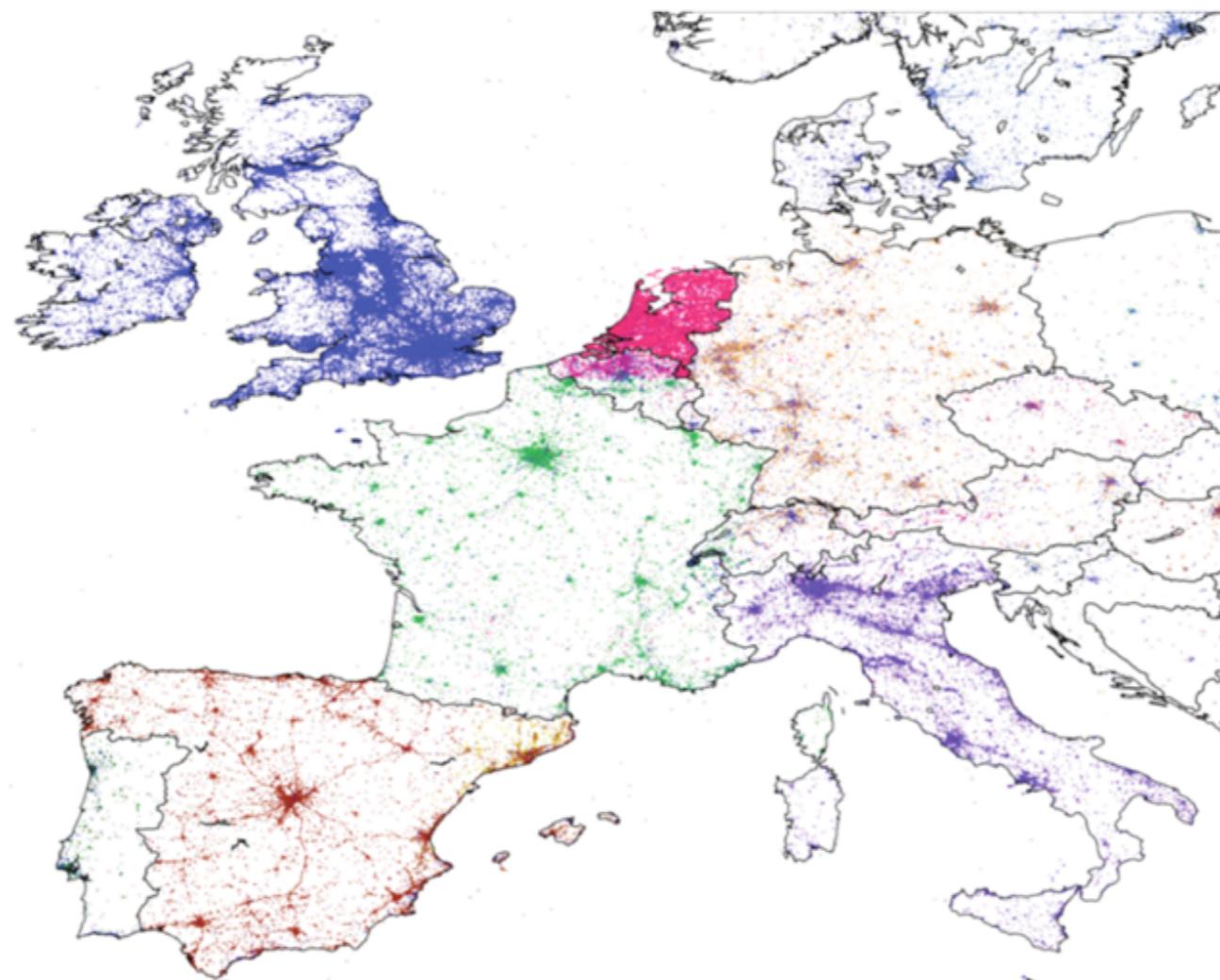
C++ library and Python bindings for detecting language from UTF8 text, extracted from the Chromium browser

[Project Home](#) [Downloads](#) [Wiki](#) [Issues](#) [Source](#)

[Summary](#) [People](#)

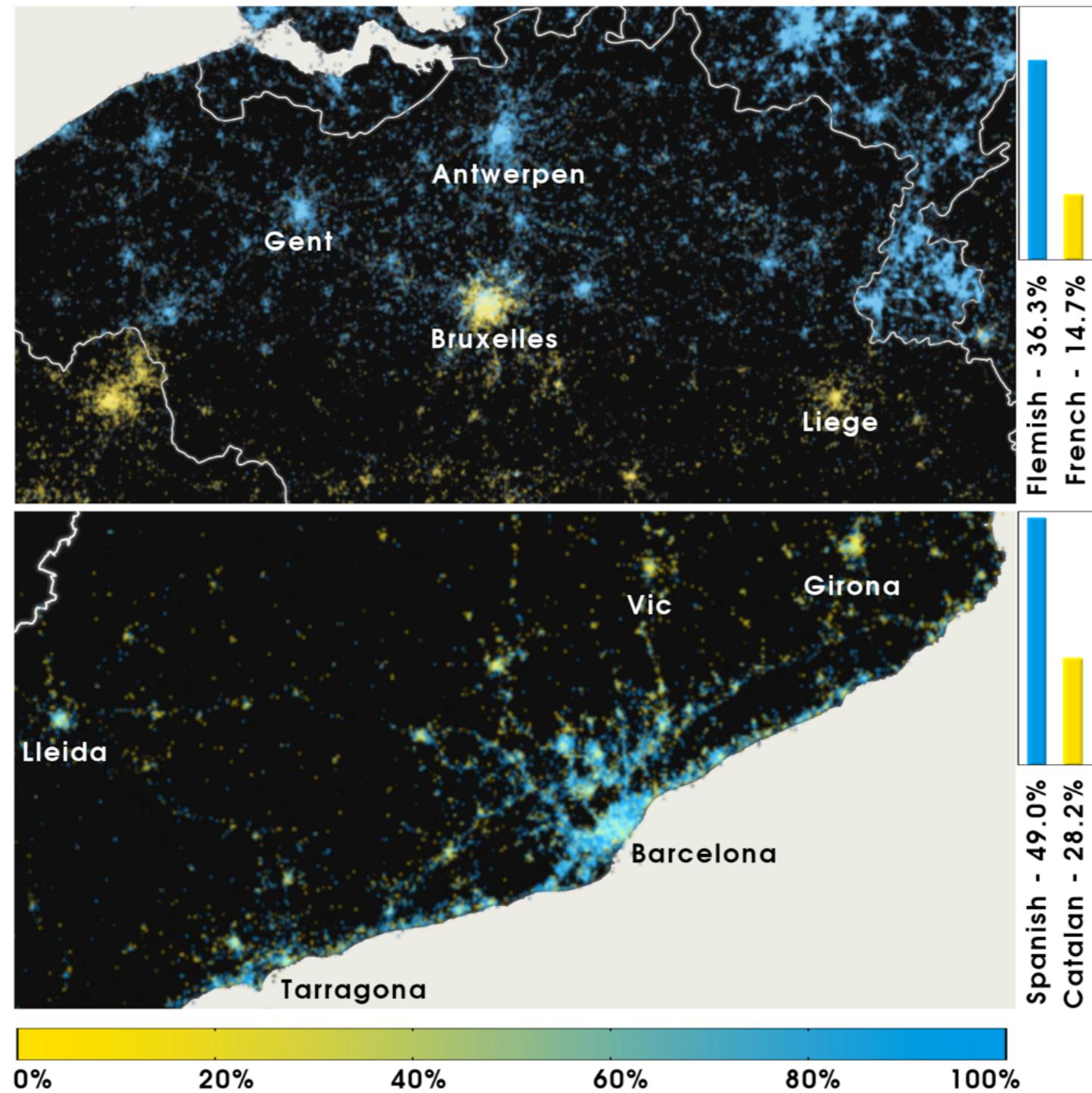
Primary Language Use

PLoS One 8, E61981 (2013)

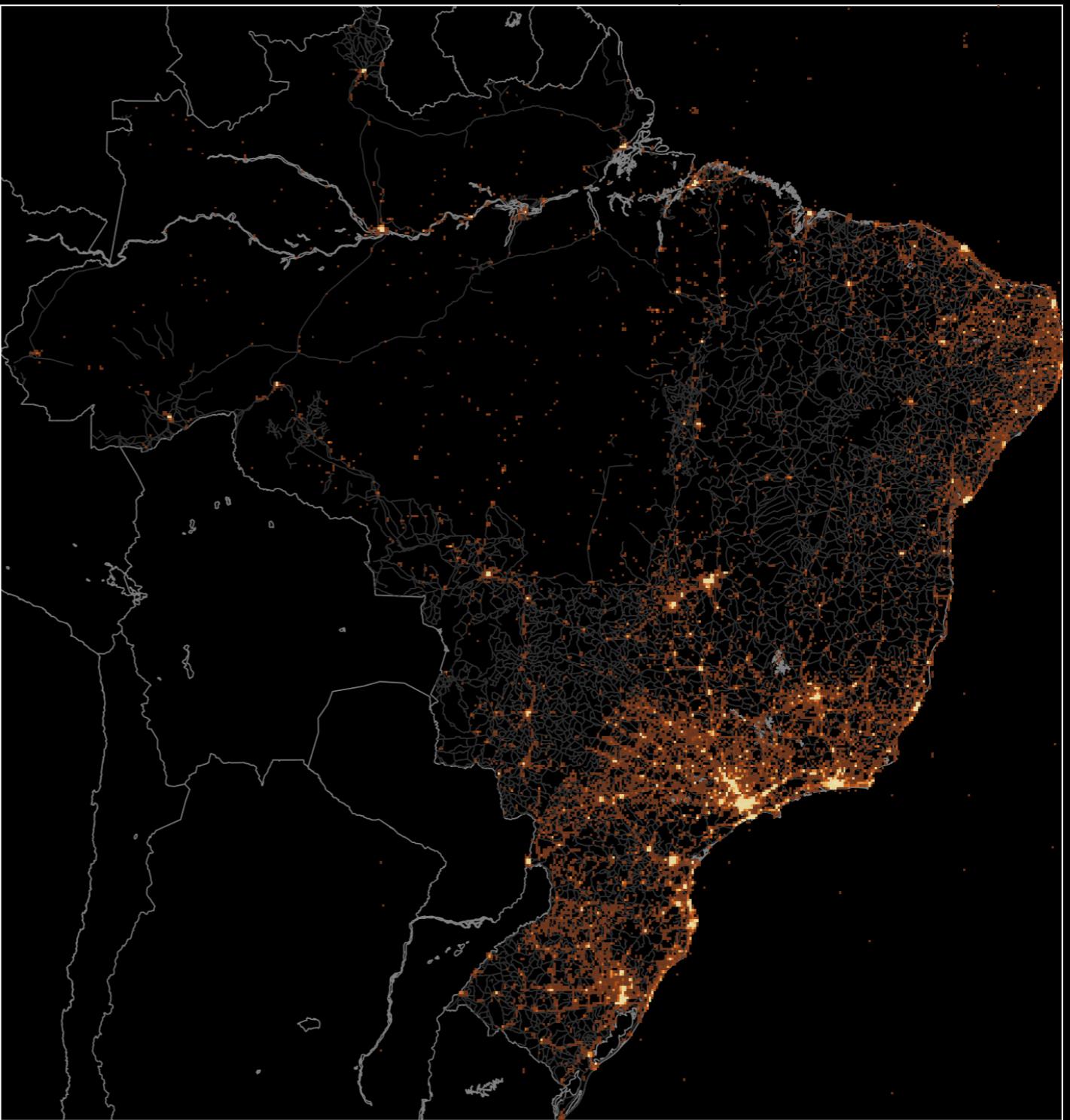


Language Distribution

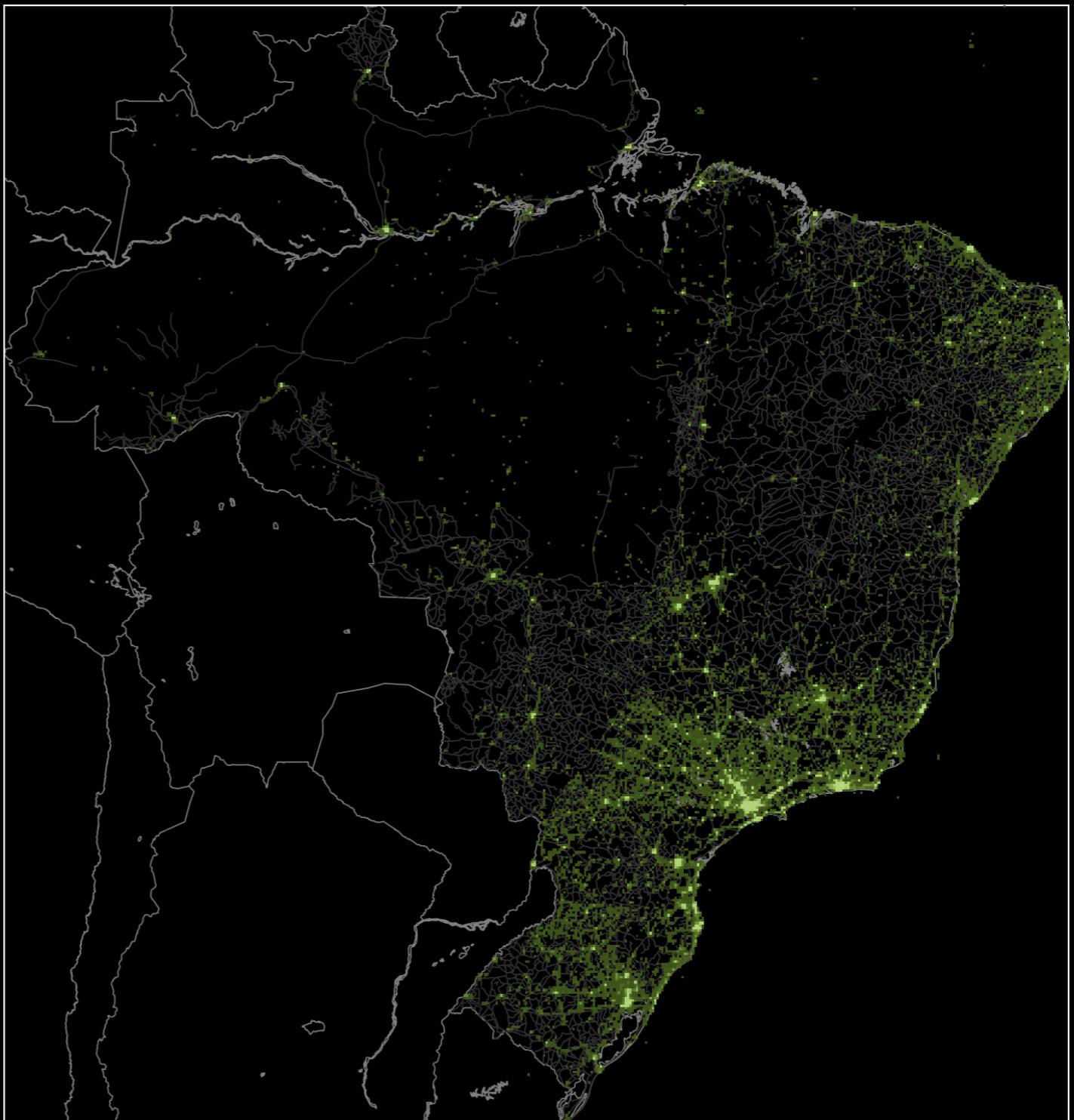
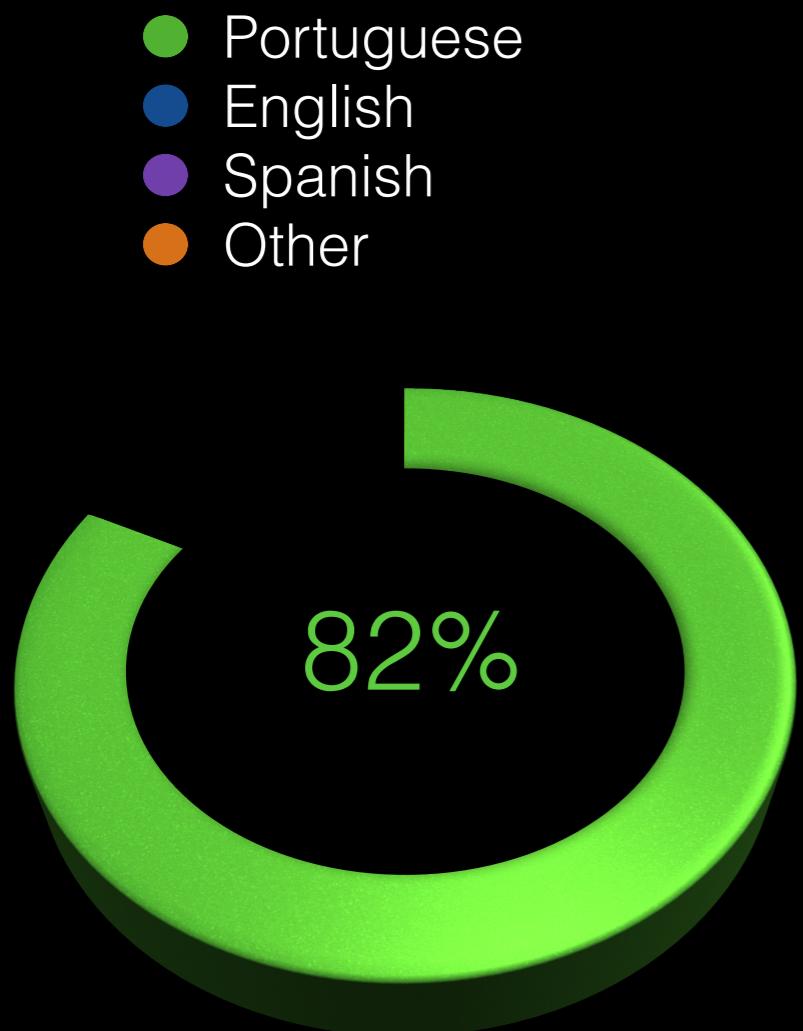
PLoS One 8, E61981 (2013)



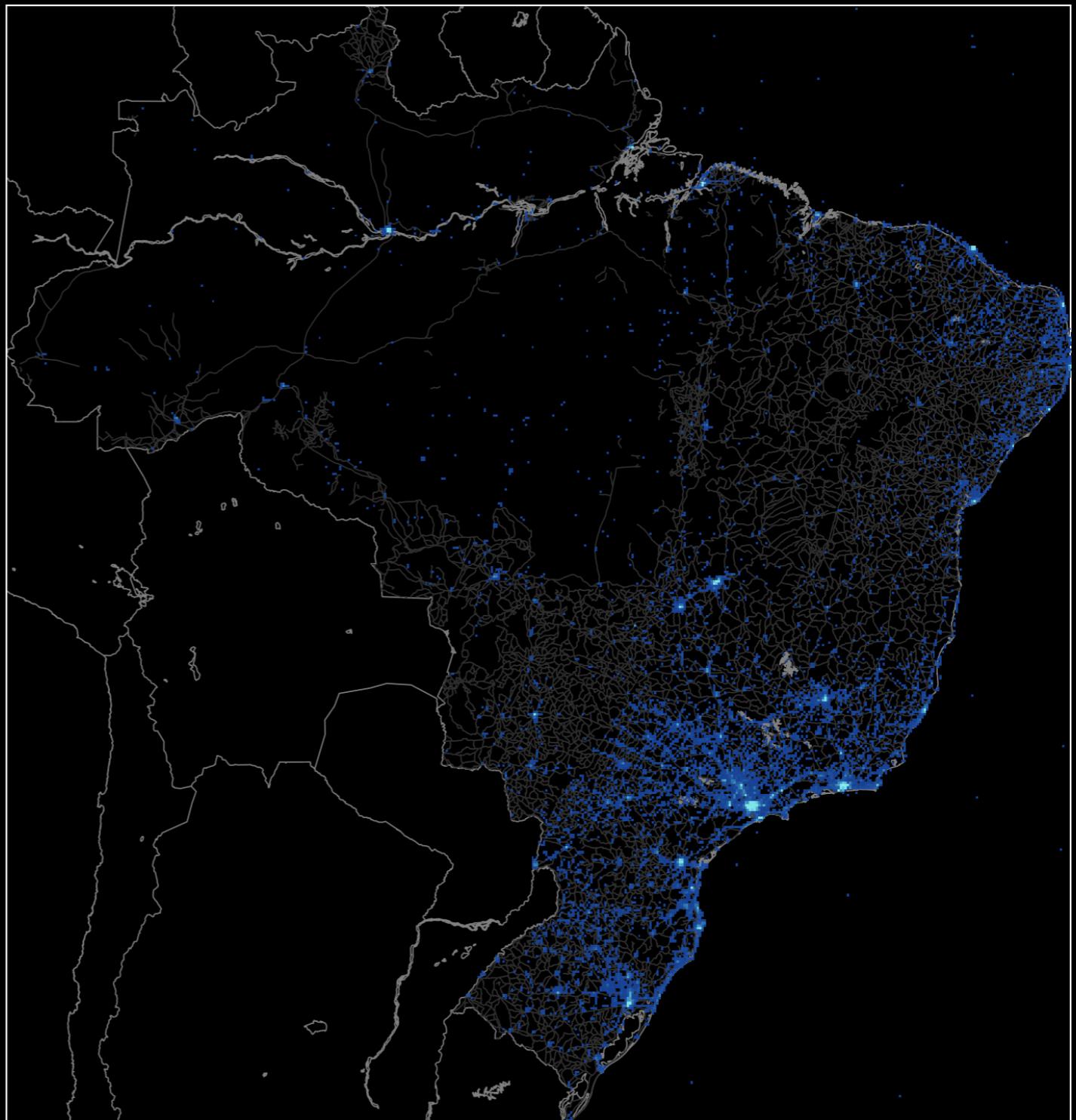
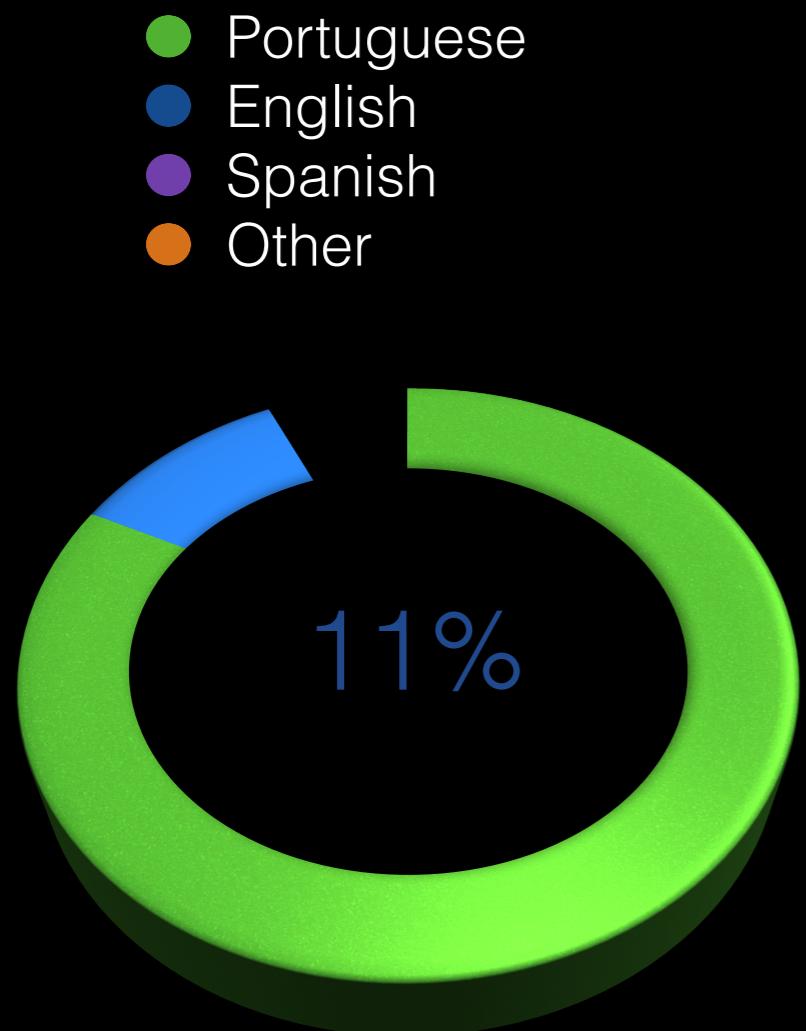
Signal By Language



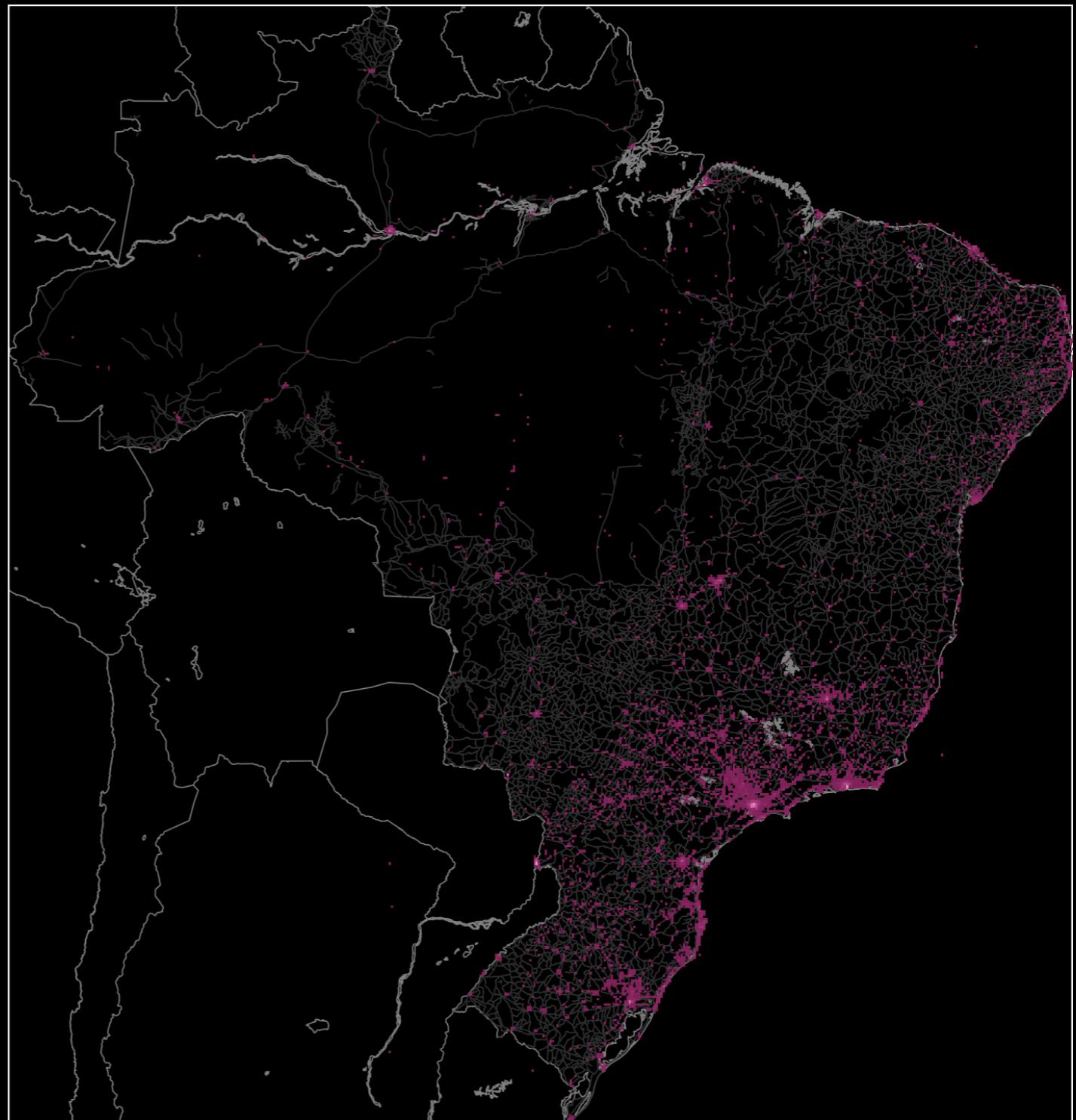
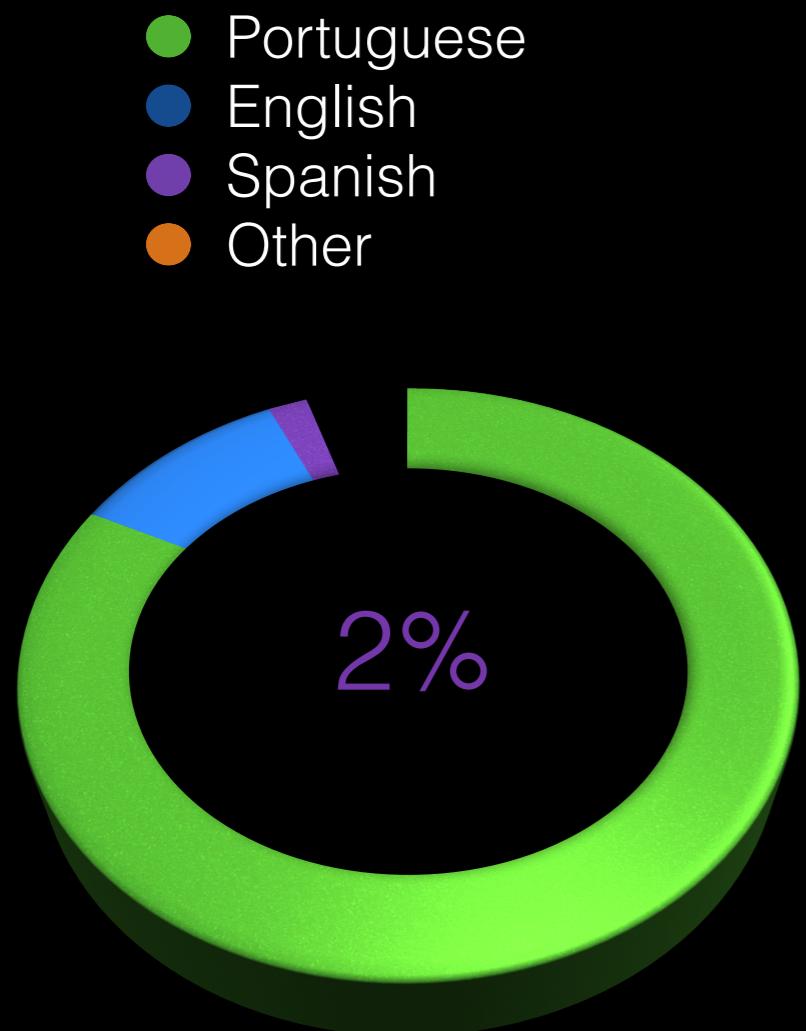
Signal By Language



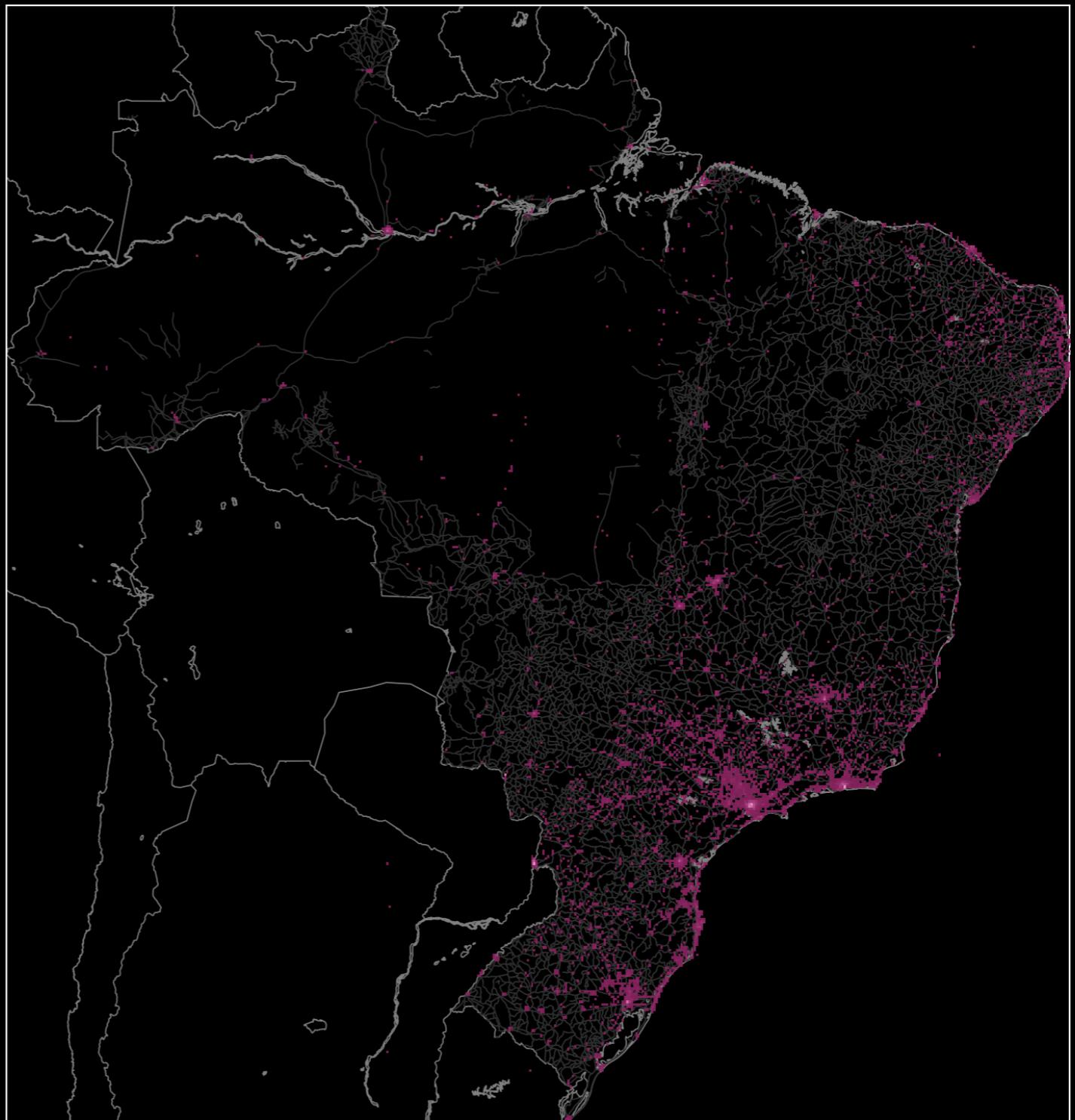
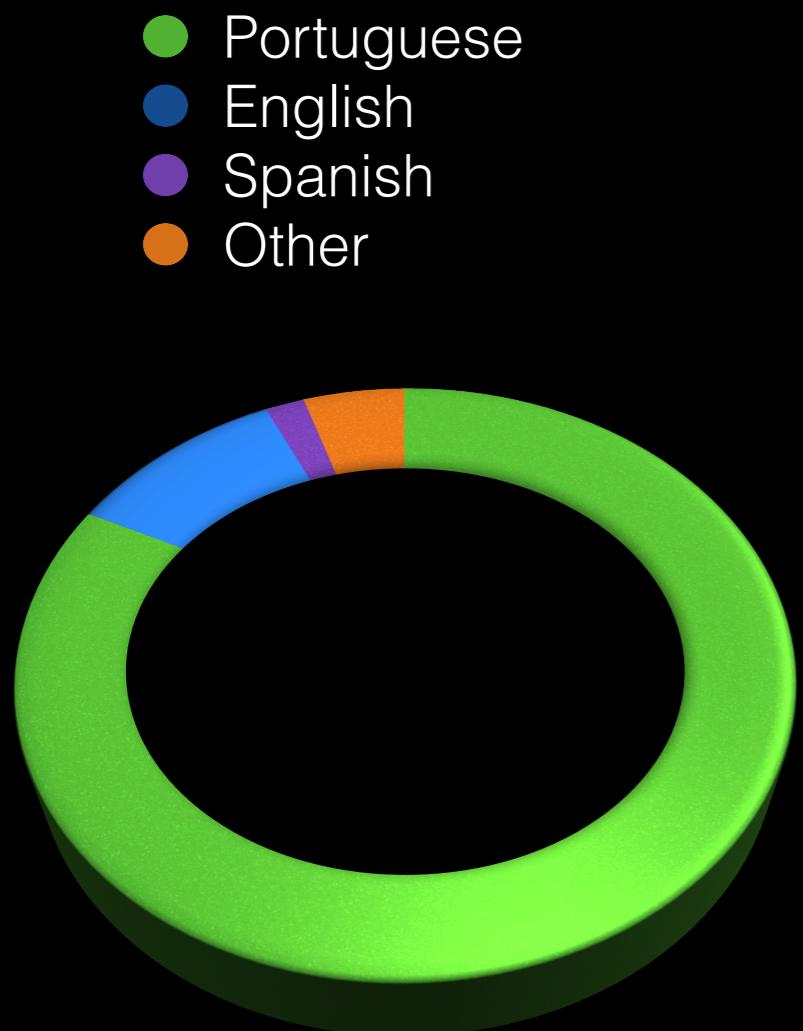
Signal By Language



Signal By Language

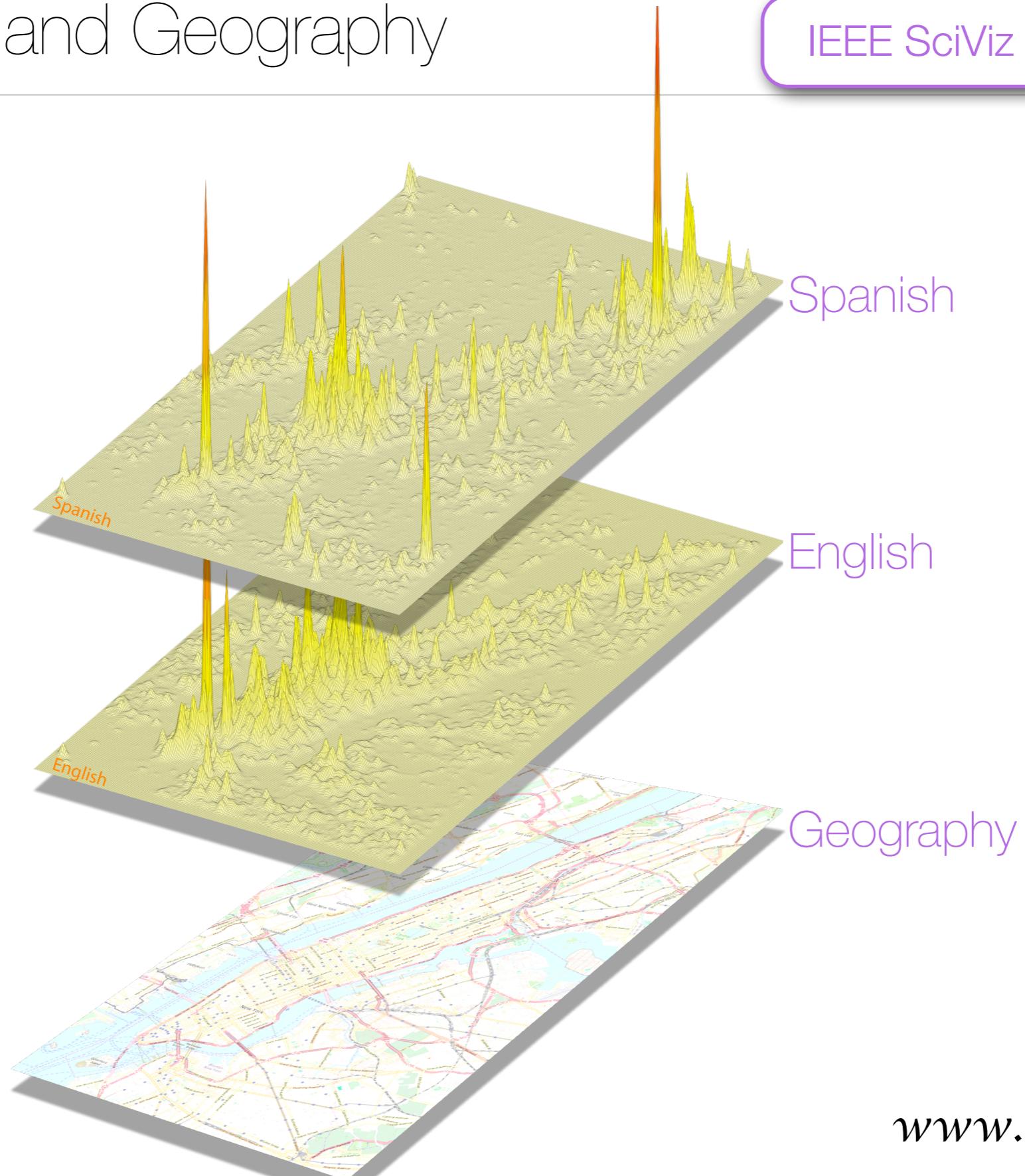


Signal By Language



Language and Geography

IEEE SciVis 23, 791 (2017)



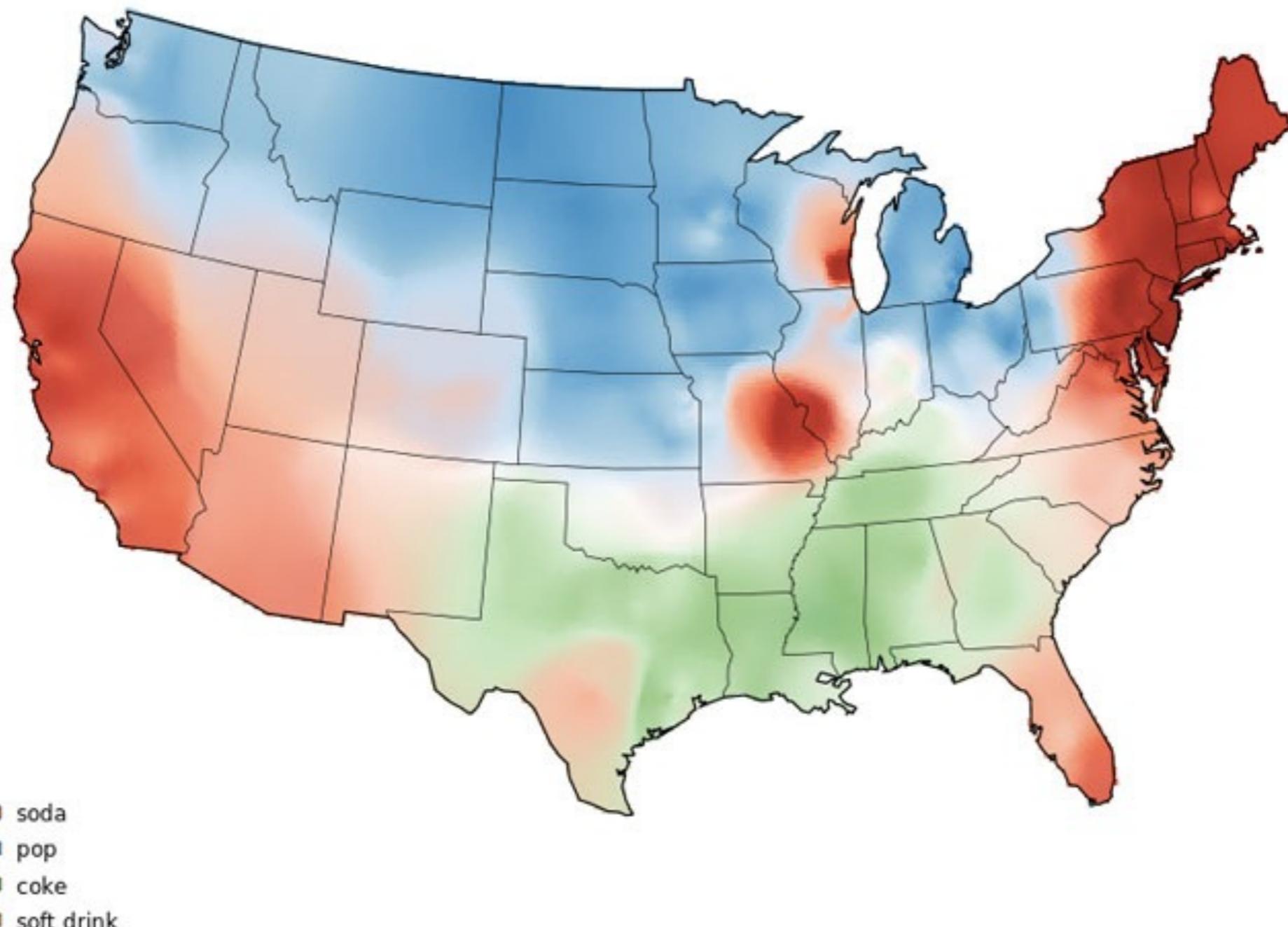


COMPUTADORA
ORDENADOR
COMPUTADOR

**SPANISH
SUPERDIALECTS**

Local Variations

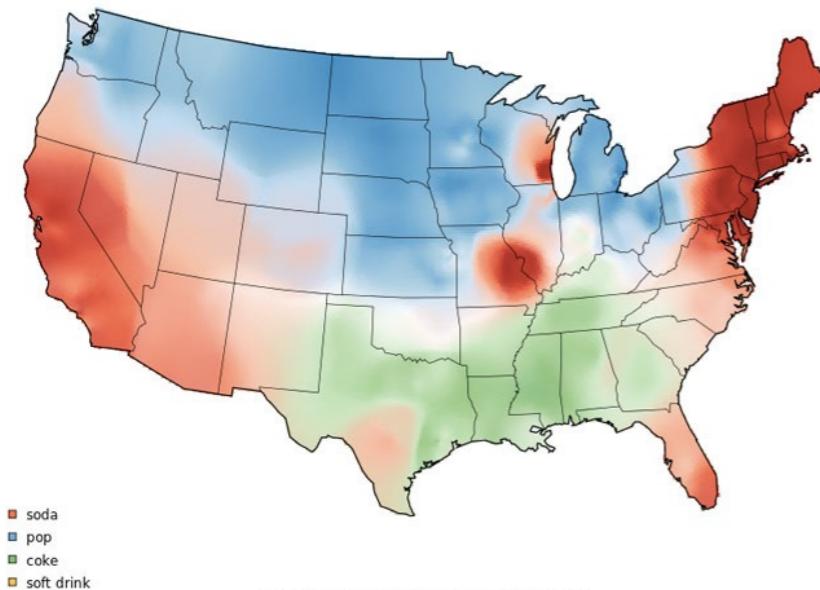
What is your generic term for a sweetened carbonated beverage?



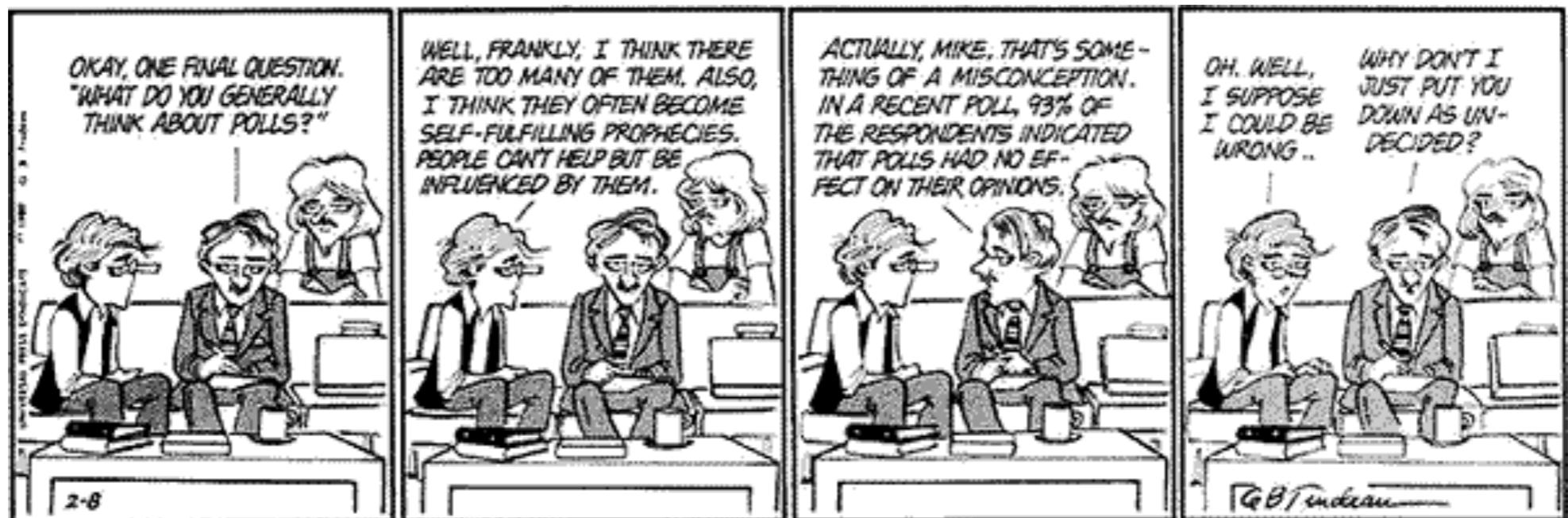
Joshua Katz, Department of Statistics, NC State University

Local Variations

What is your generic term for a sweetened carbonated beverage?

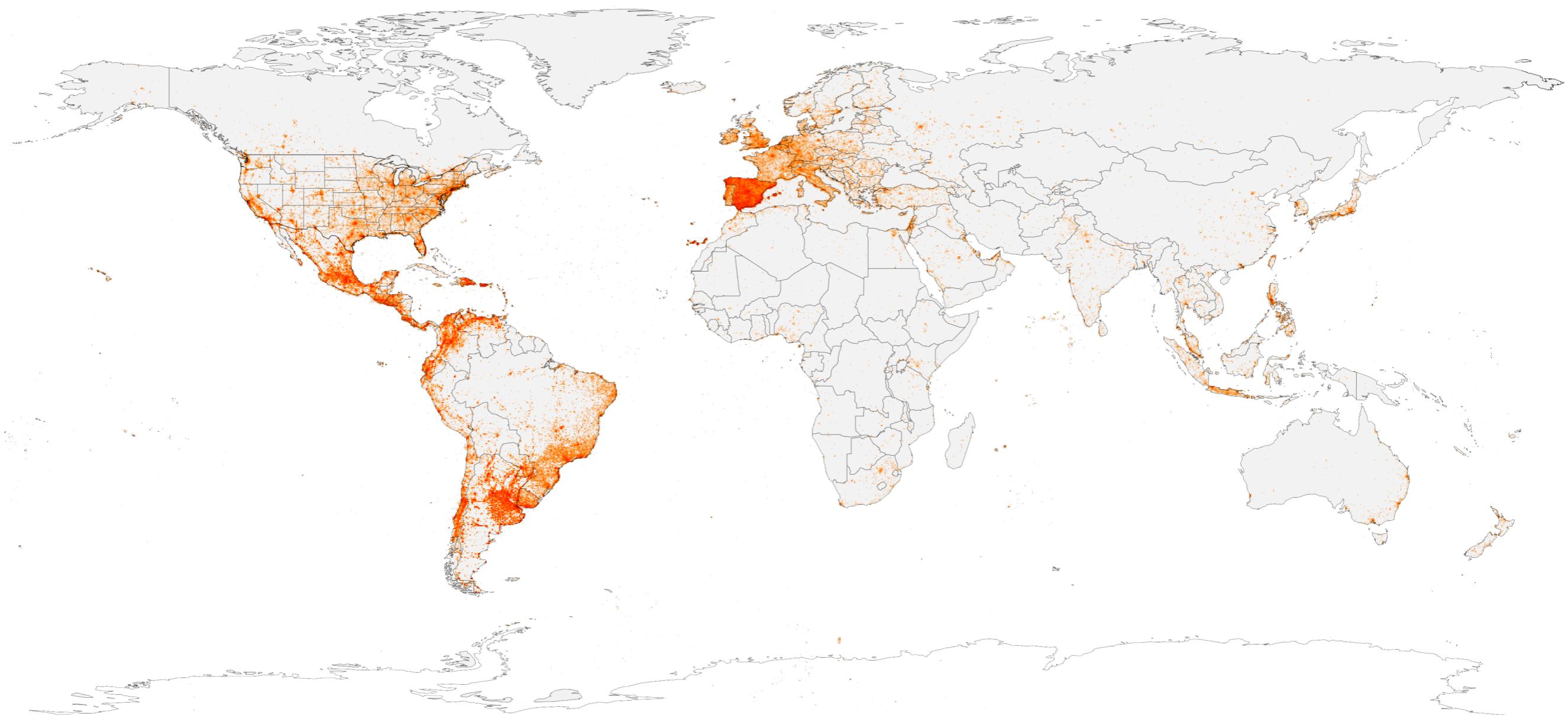


Joshua Katz, Department of Statistics, NC State University



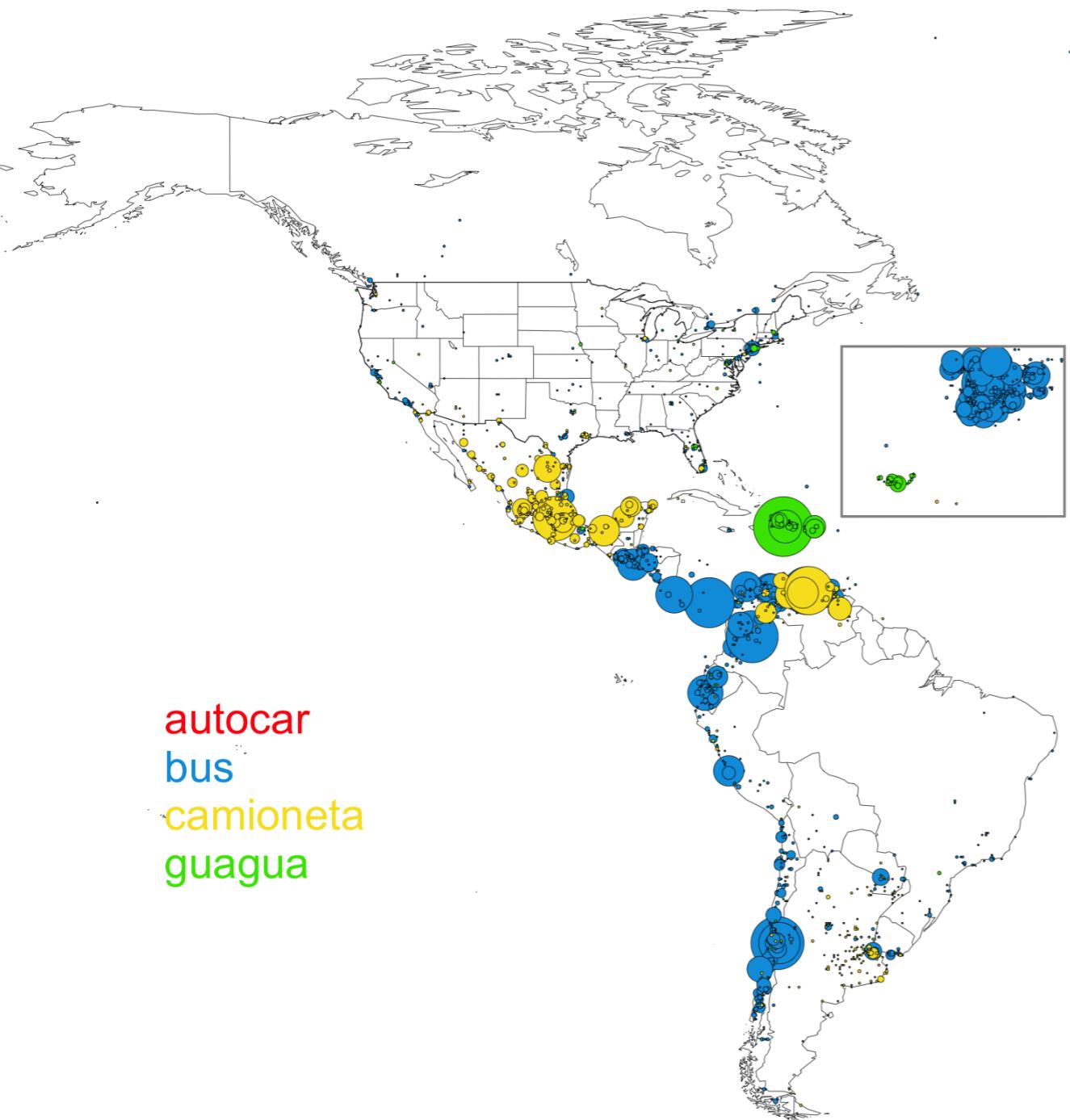
Spanish

PLoS One 9, e112074 (2014)
RLI XVI 2, 65 (2016)

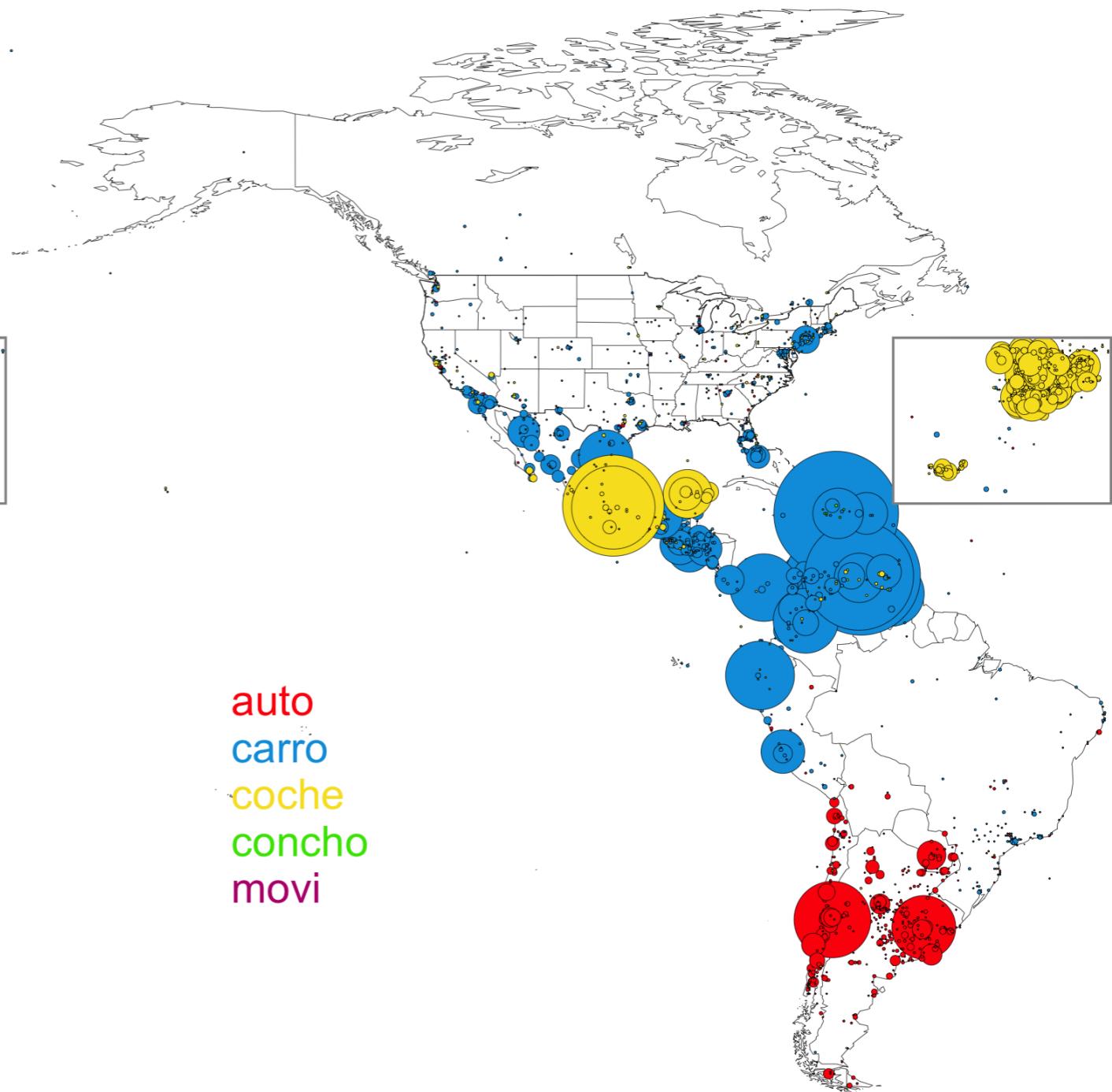


Local Variations

PLoS One 9, e112074 (2014)
RLI XVI 2, 65 (2016)



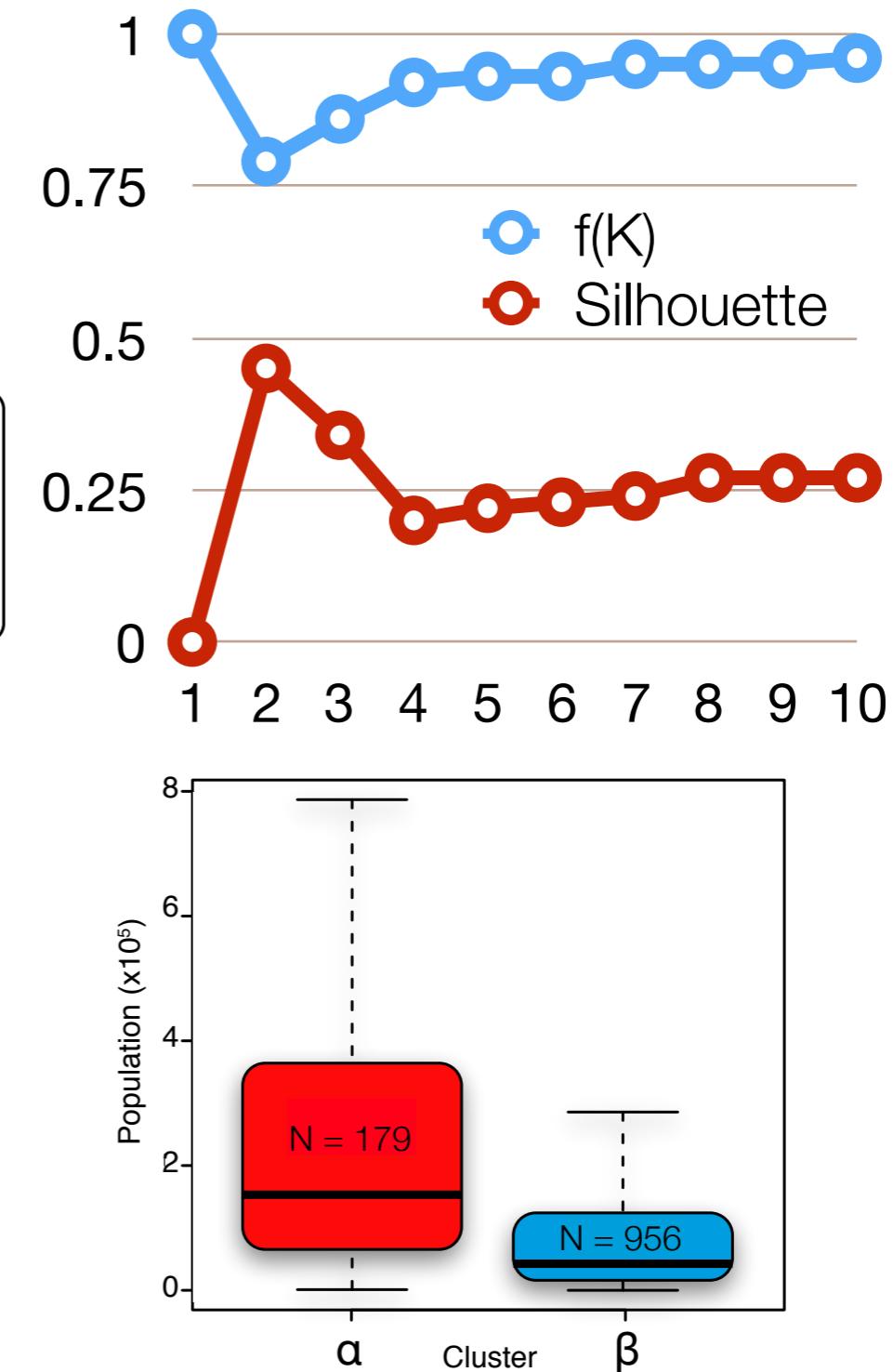
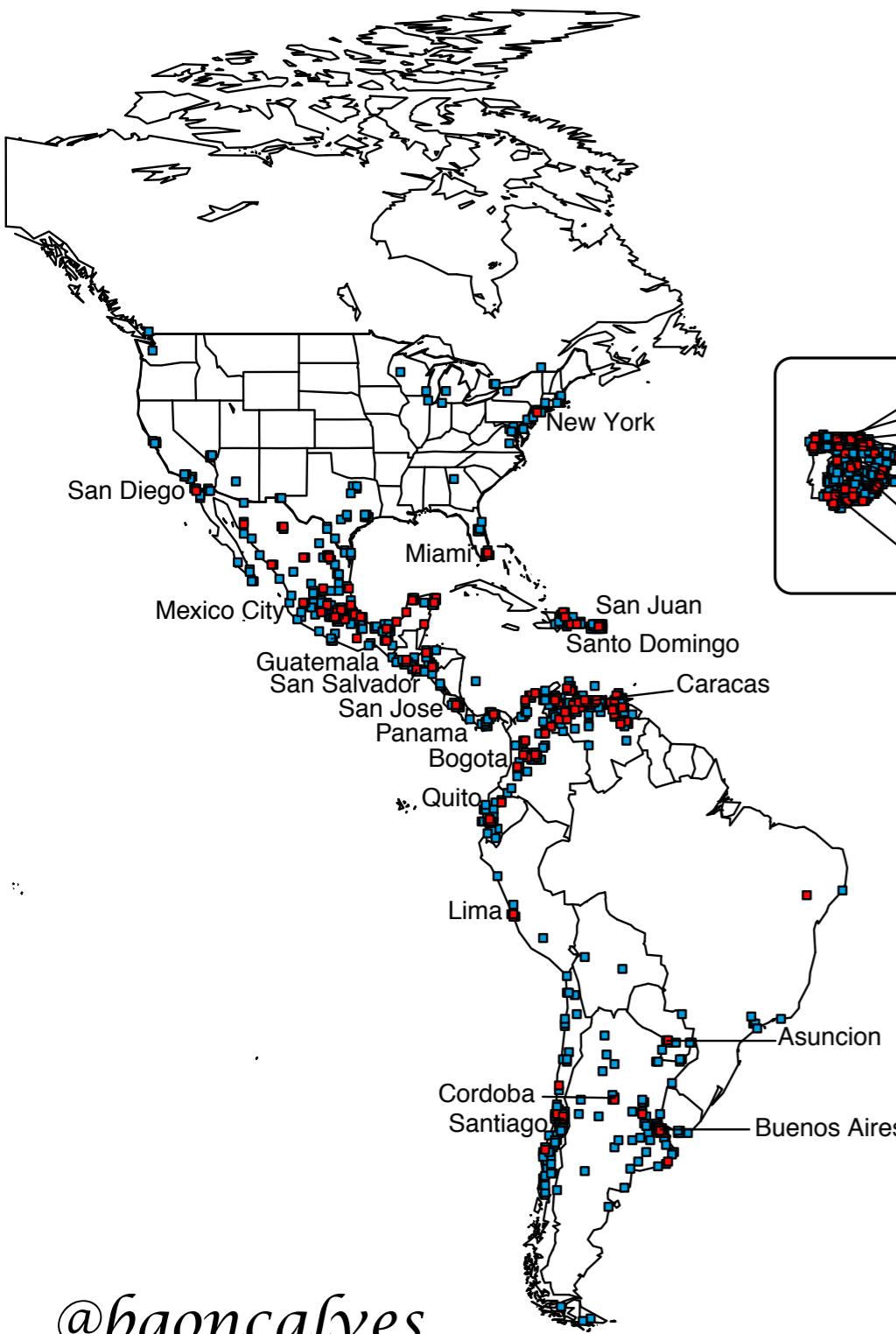
autocar
bus
camioneta
guagua



auto
carro
coche
concho
movi

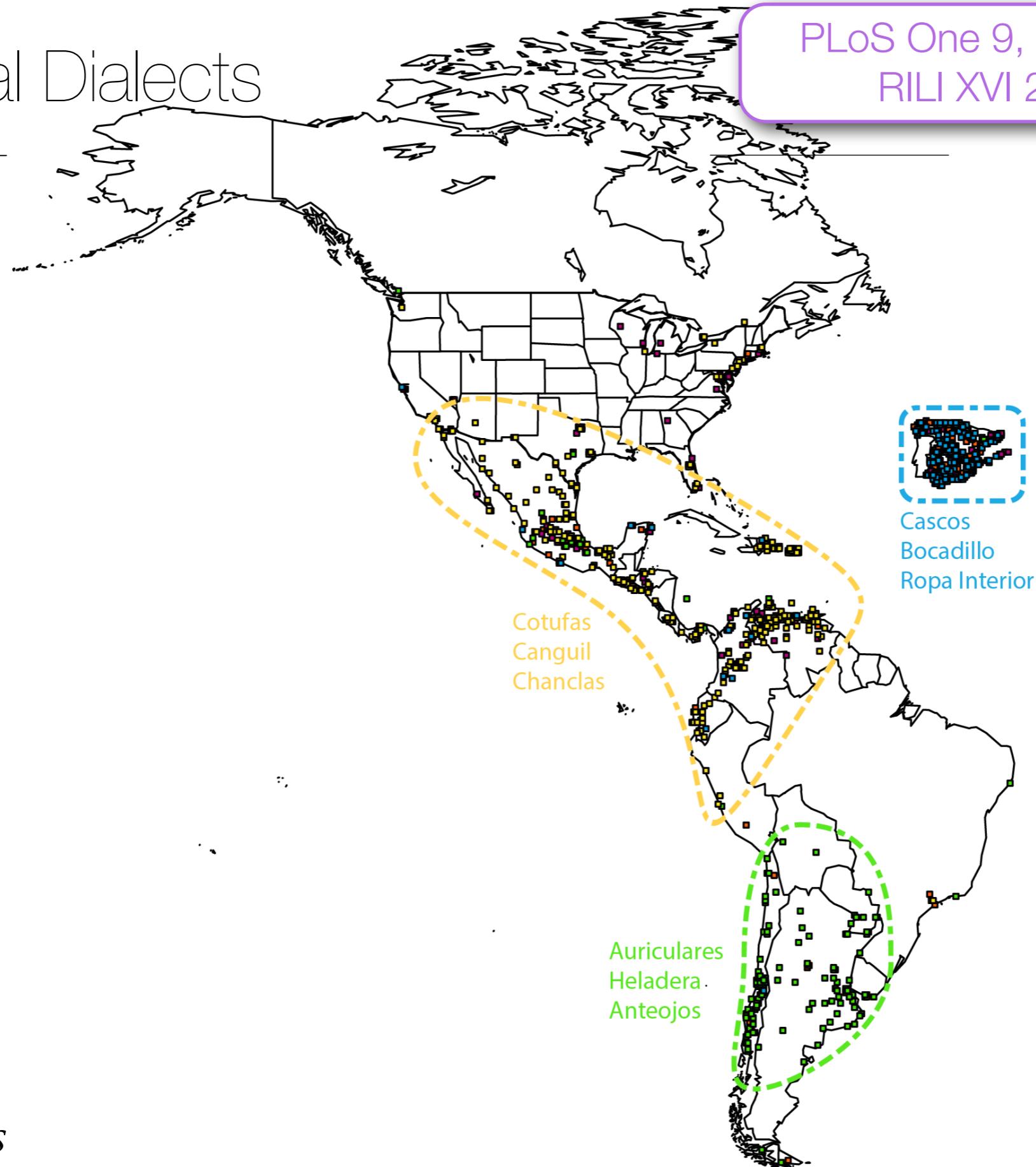
Superdialects

PLoS One 9, e112074 (2014)
RLI XVI 2, 65 (2016)



Regional Dialects

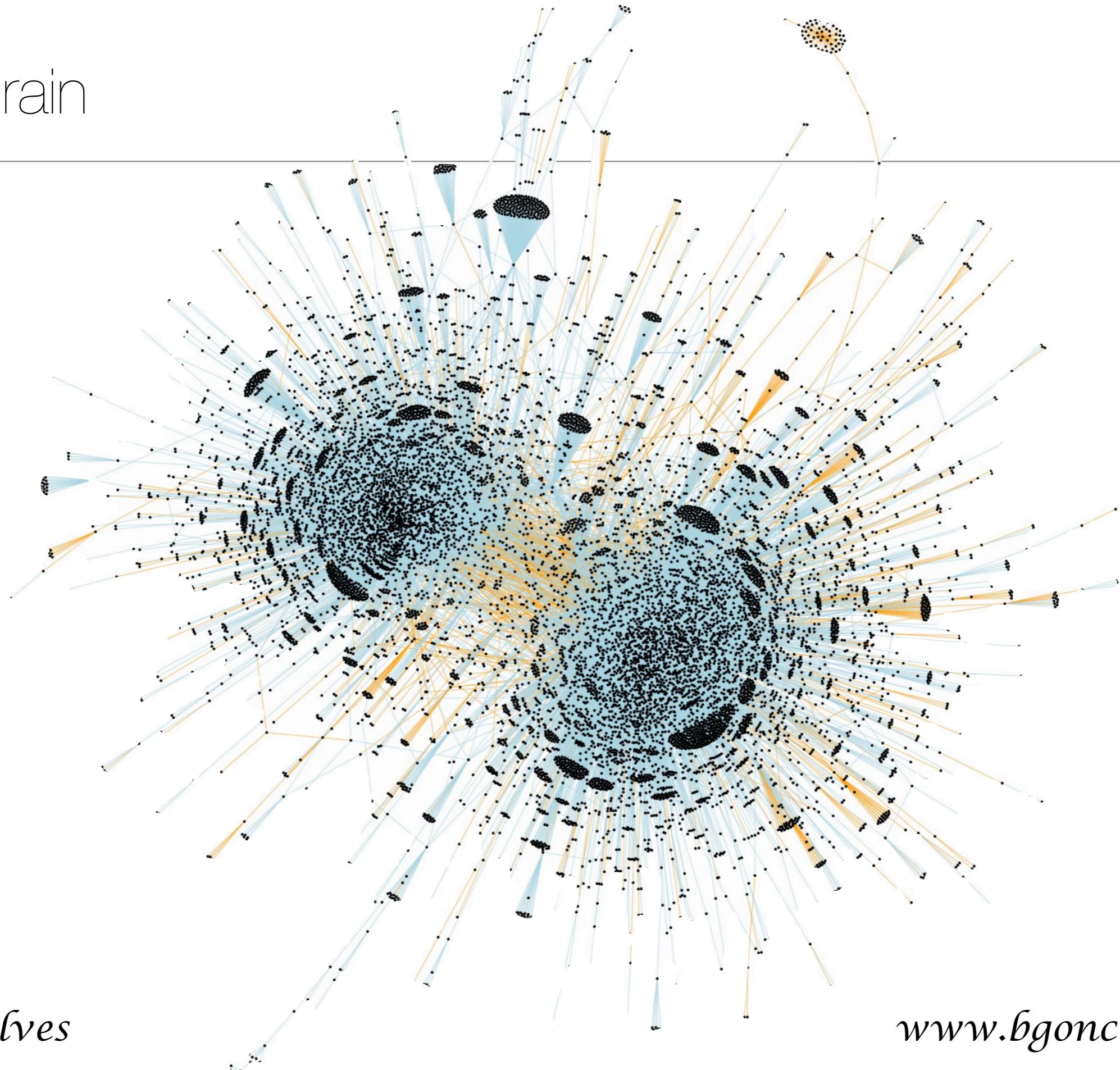
PLoS One 9, e112074 (2014)
RLI XVI 2, 65 (2016)

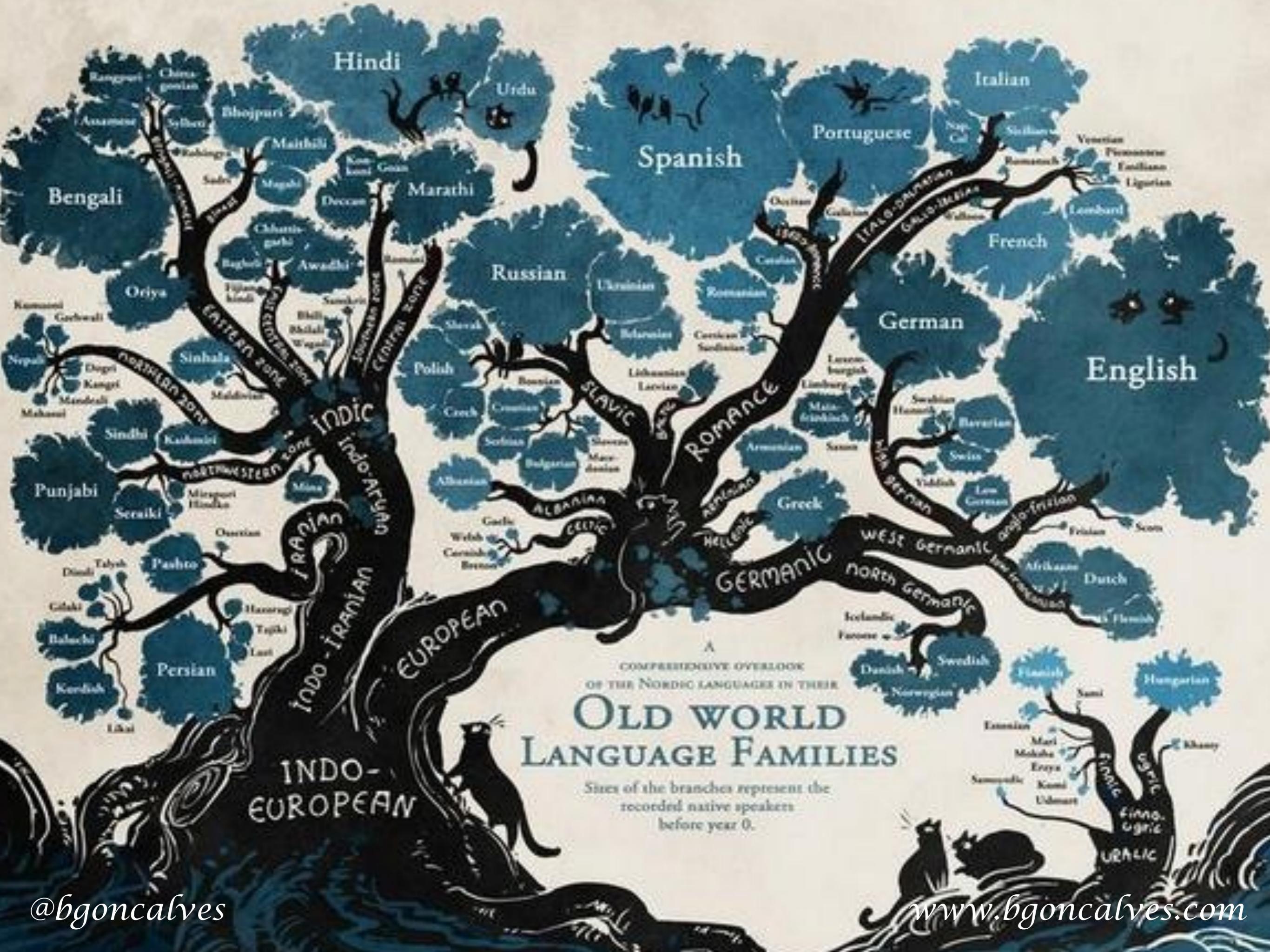


Language Connections

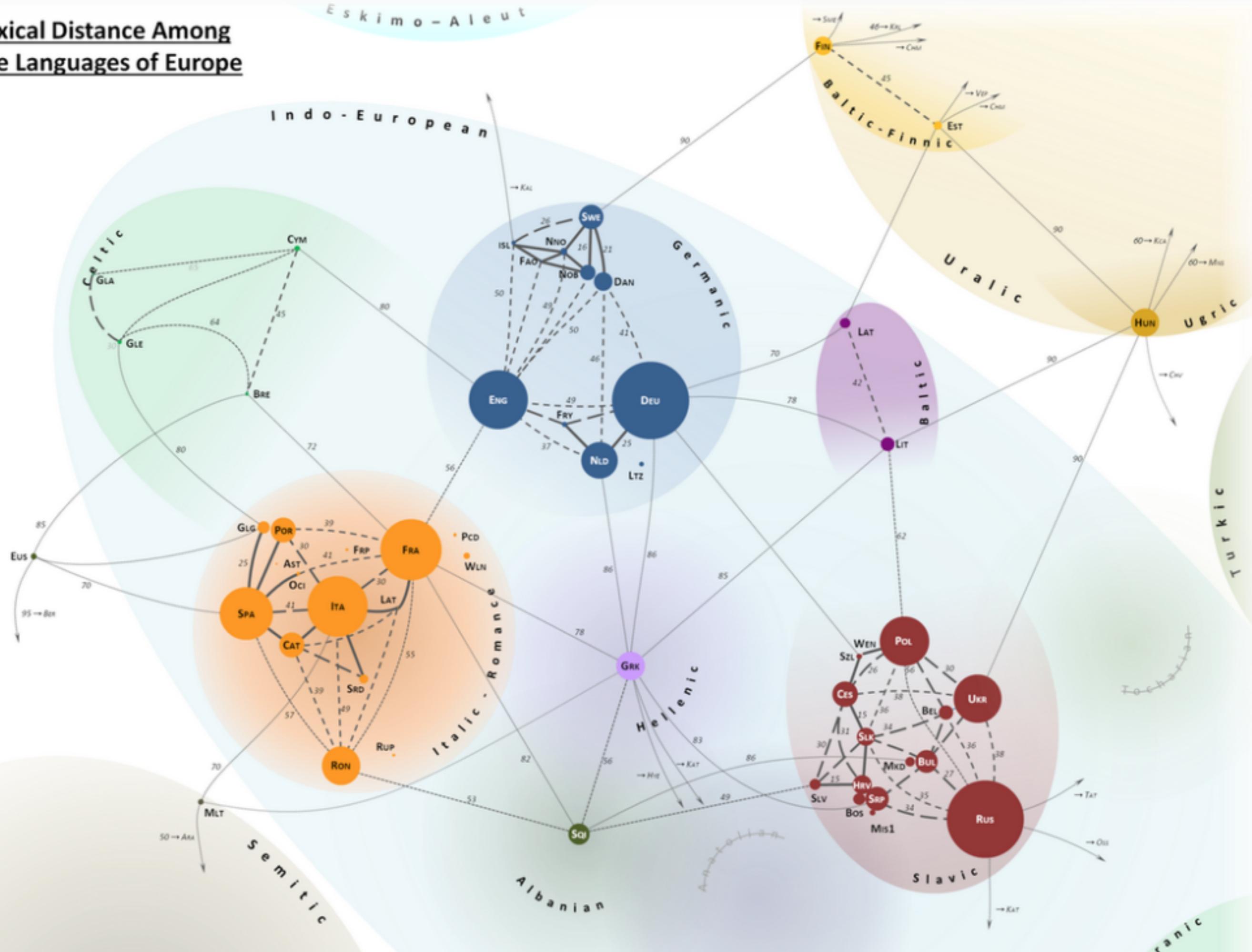


#bahrain



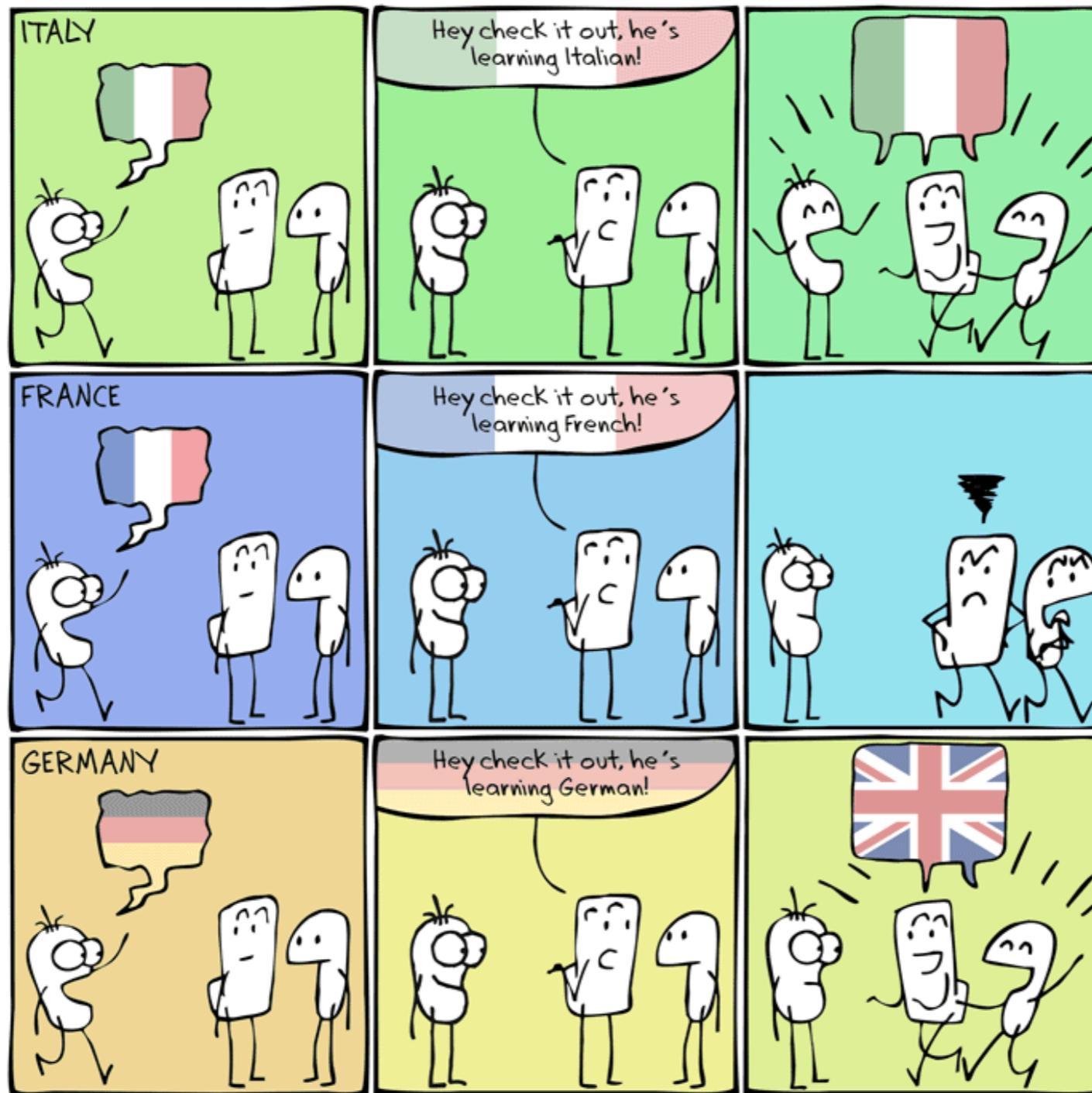


Lexical Distance Among the Languages of Europe



Bilingualism

ITCHY FEET



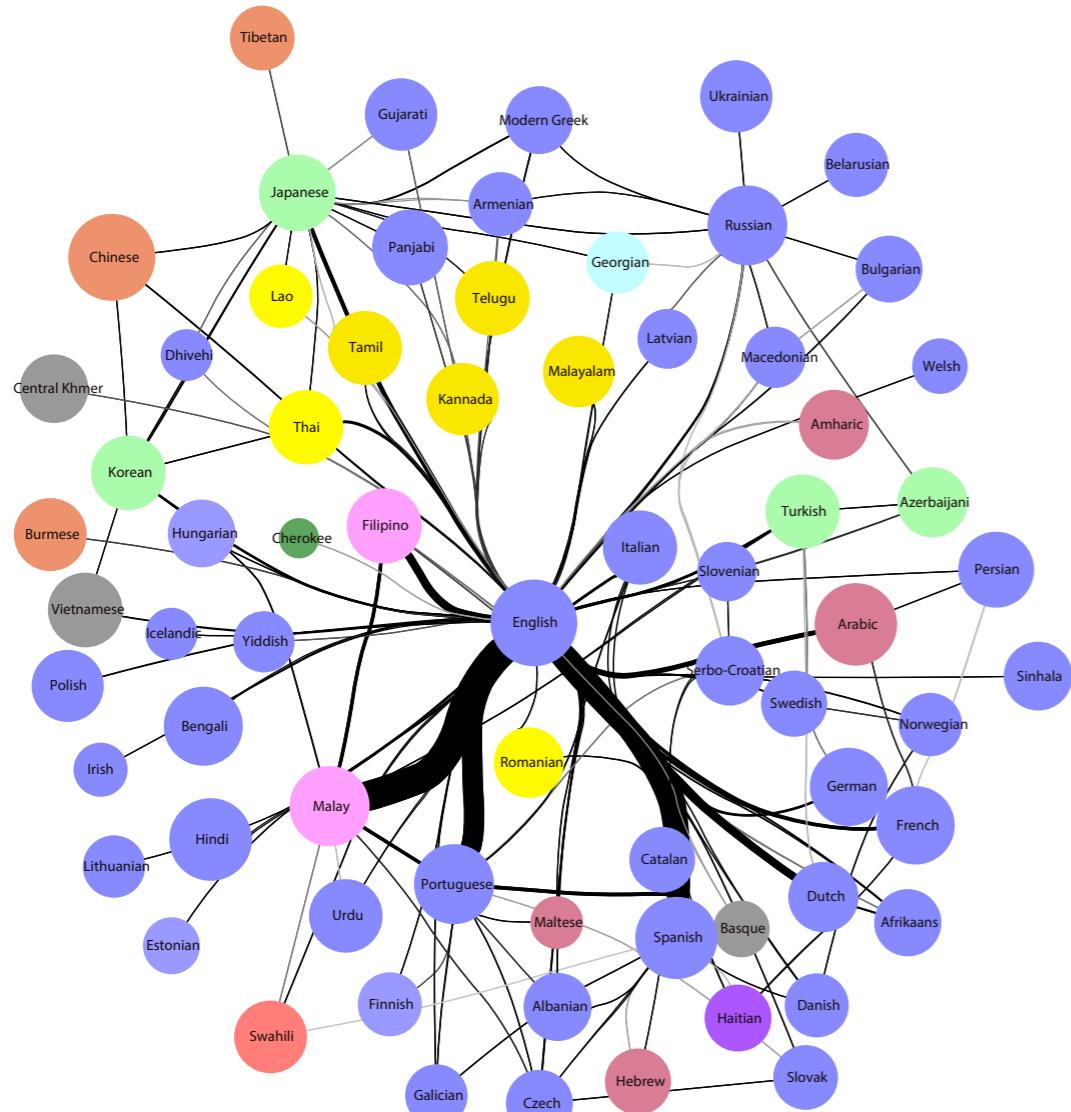
© 2014 - Malachi Ray Rempen

www.itchyfeetcomic.com

Global Language Network

PNAS 111, E5616 (2014)

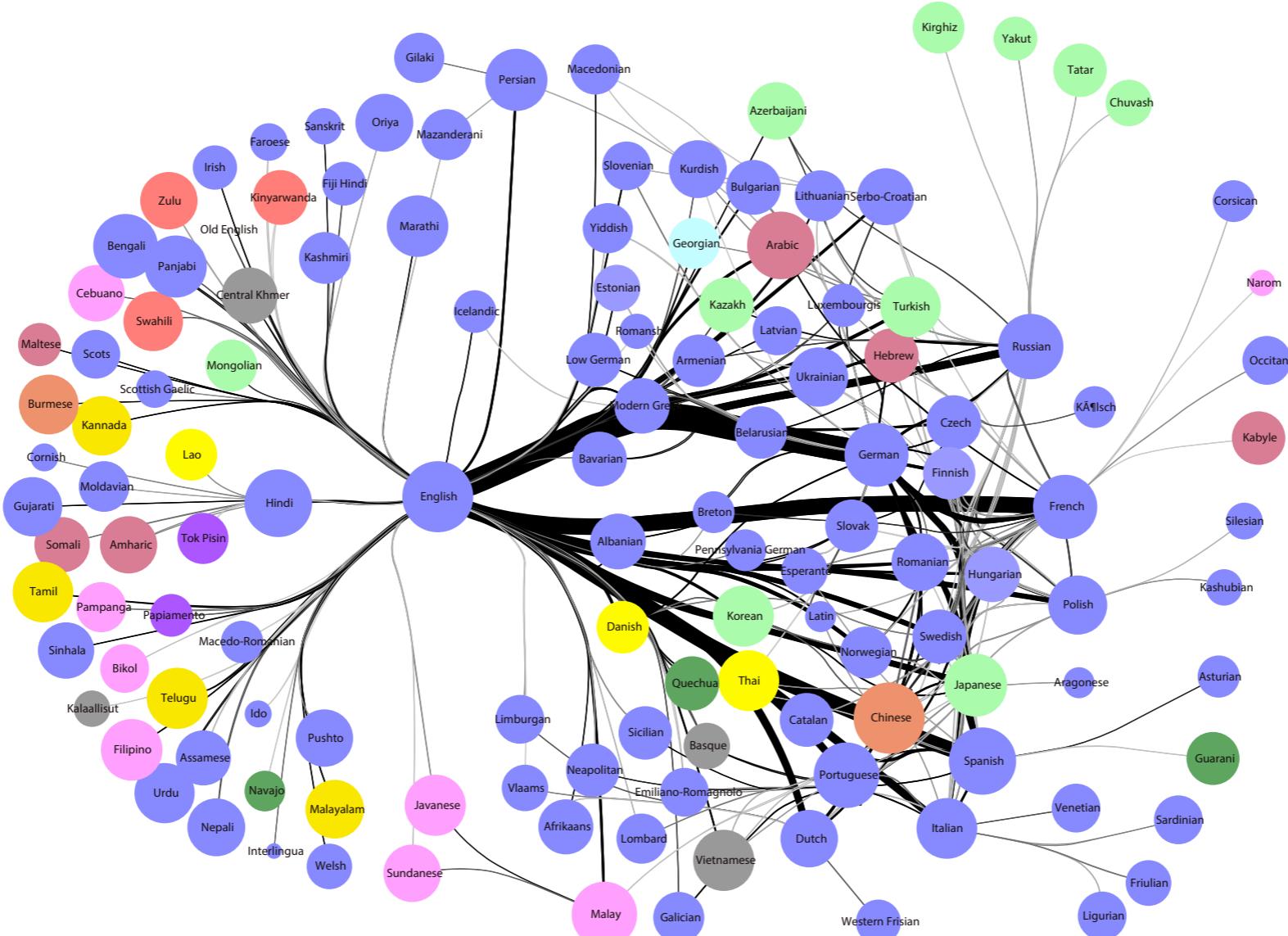
Twitter



Global Language Network

PNAS 111, E5616 (2014)

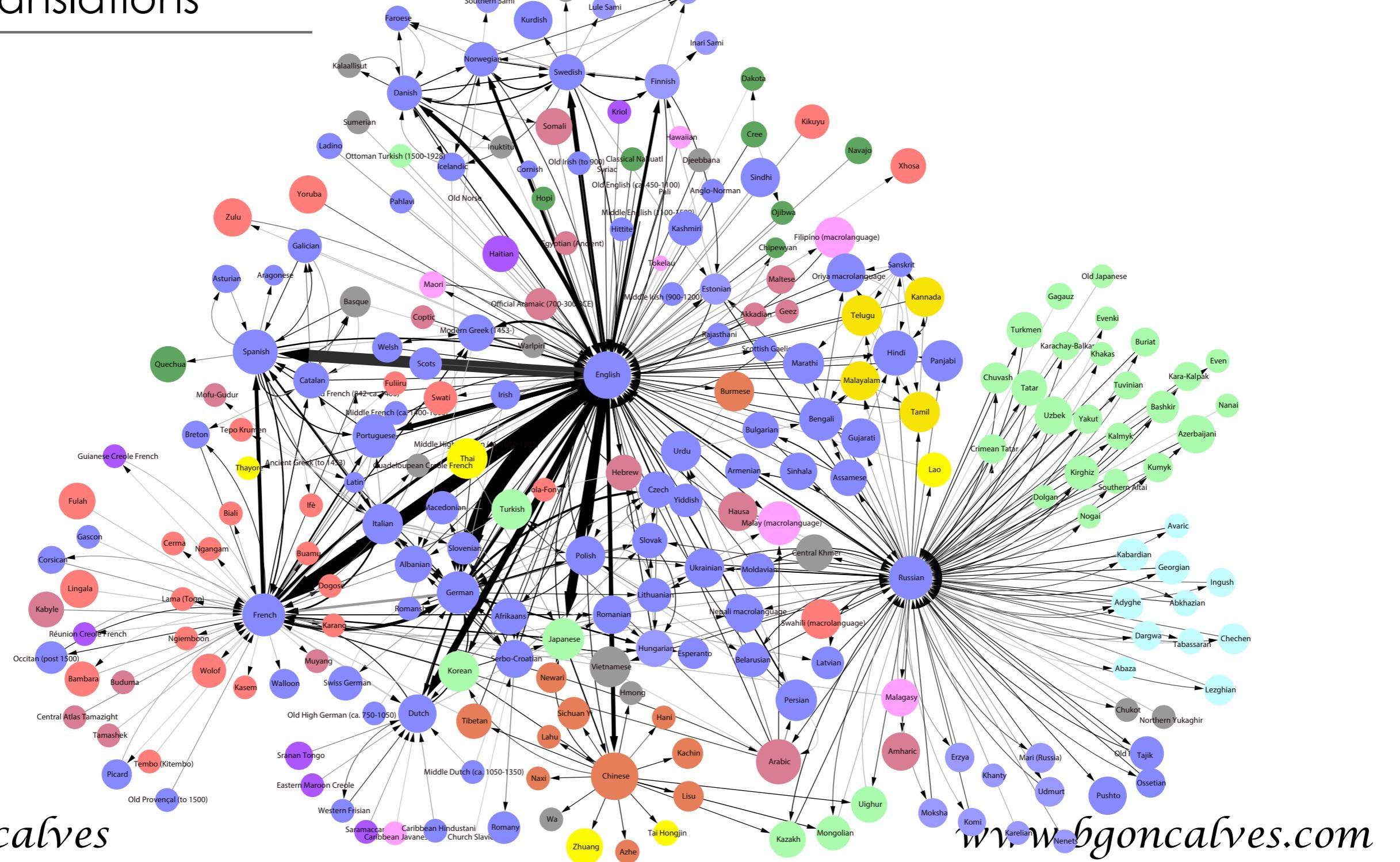
Wikipedia



Global Language Network

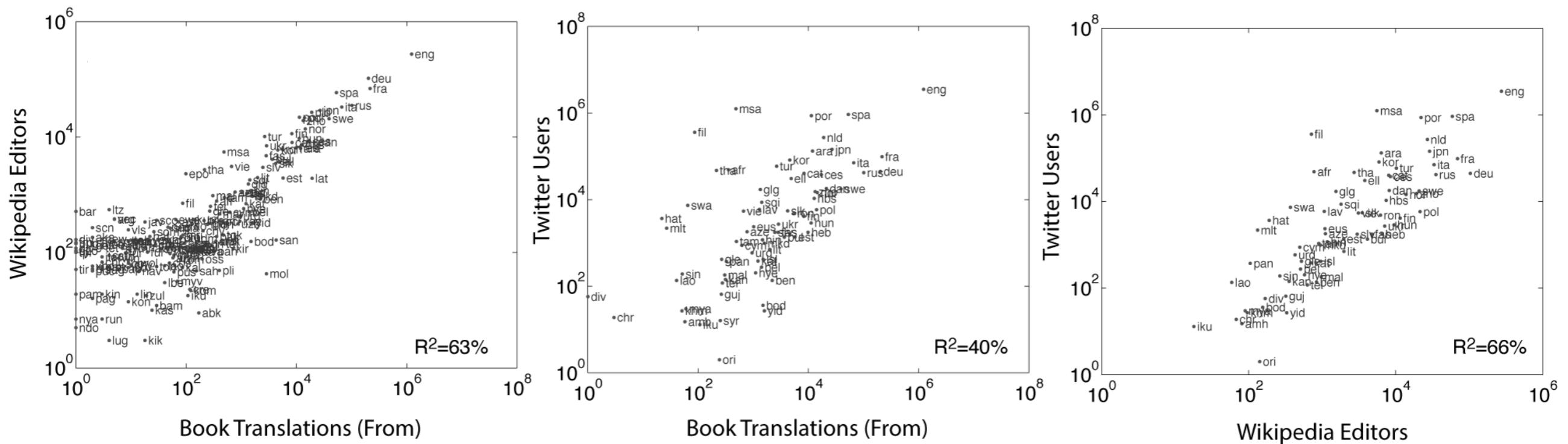
PNAS 111, E5616 (2014)

Book Translations



Global Language Network

PNAS 111, E5616 (2014)

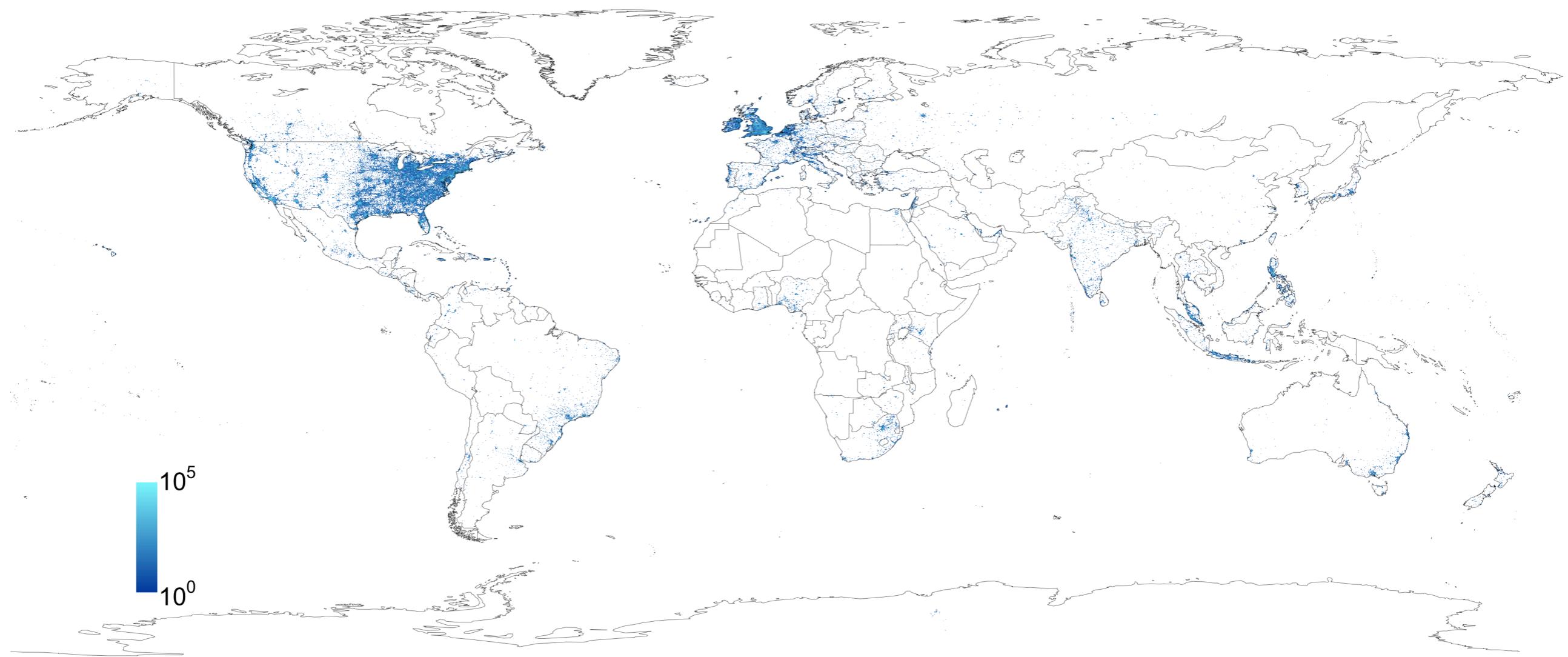


A graphic of the United States flag is positioned at the top left of the slide. It features the stars in the upper left corner and the red and white stripes running horizontally across the rest of the flag.

Linguistic Change

Is English becoming American?

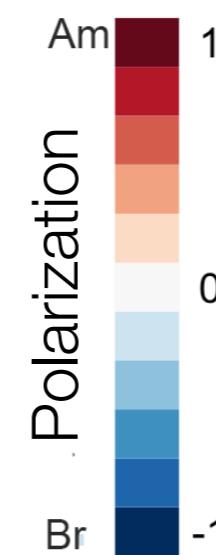
arXiv:1707.00781 (2017)



Is English becoming American?

arXiv:1707.00781 (2017)

Spelling	Vocabulary
color colour	lift elevator
rumor rumour	rubbish trash

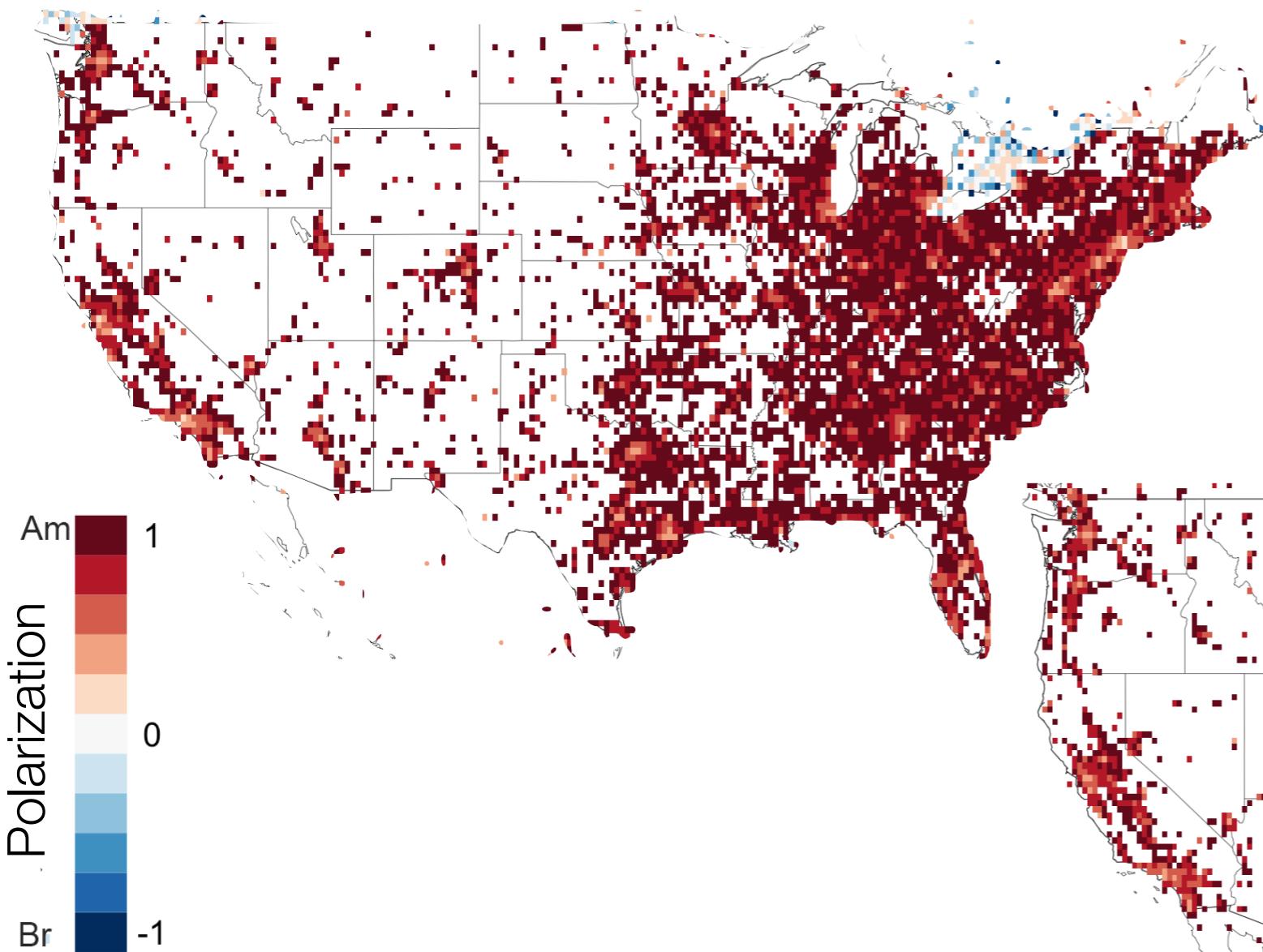


$$V_w^c = \frac{A_w^c - B_w^c}{A_w^c + B_w^c}$$

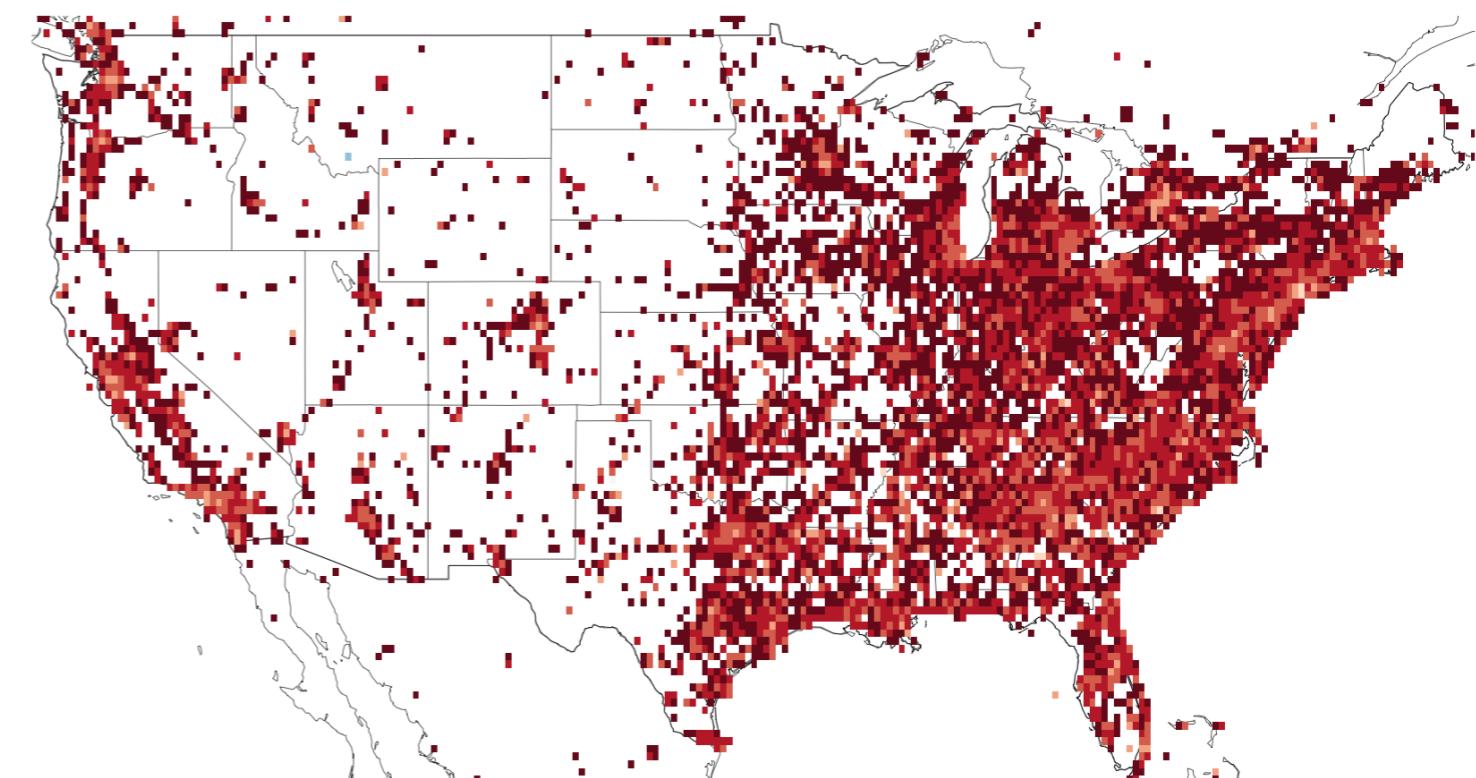
Geographical Variation

arXiv:1707.00781 (2017)

Spelling



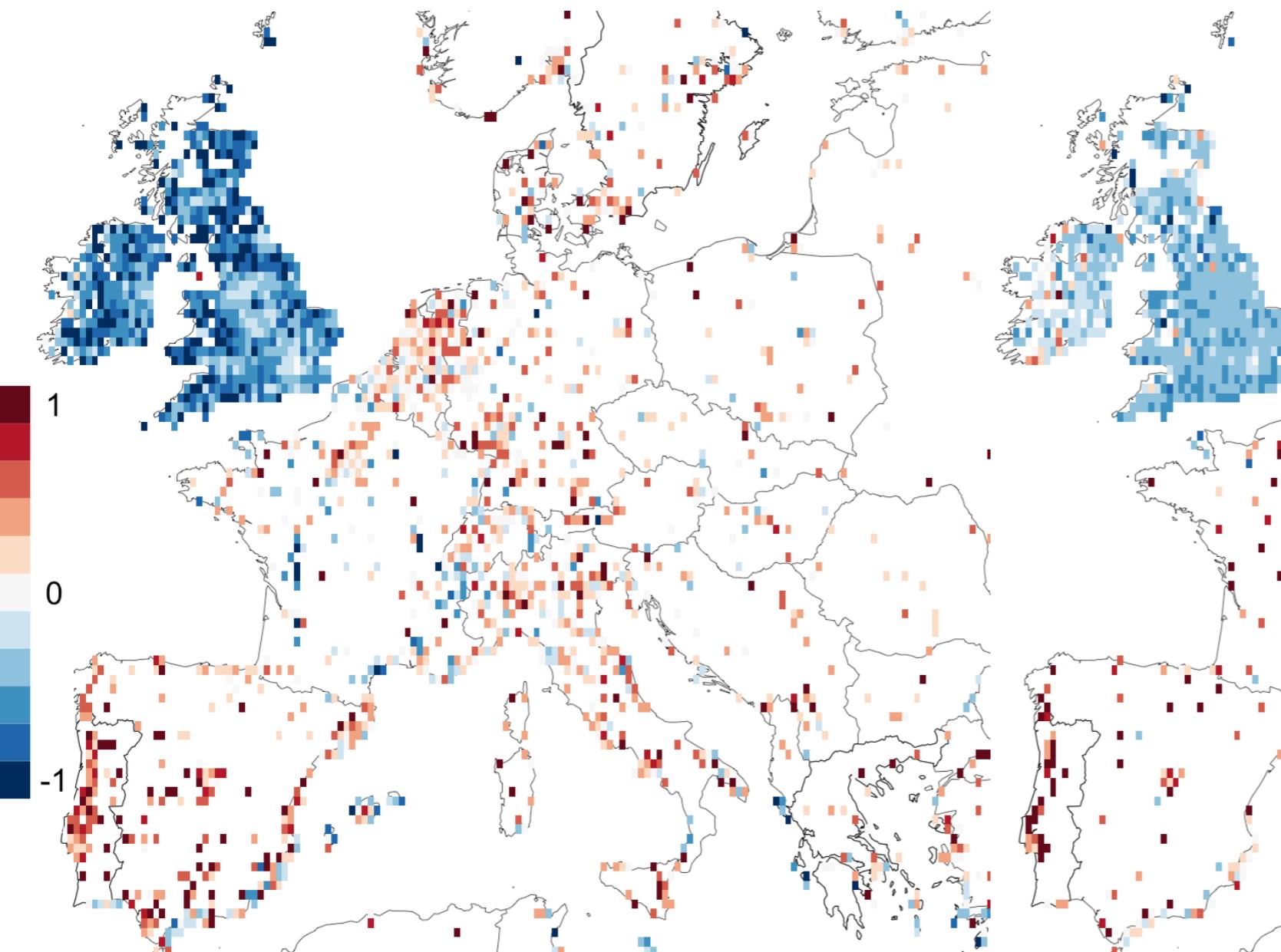
Vocabulary



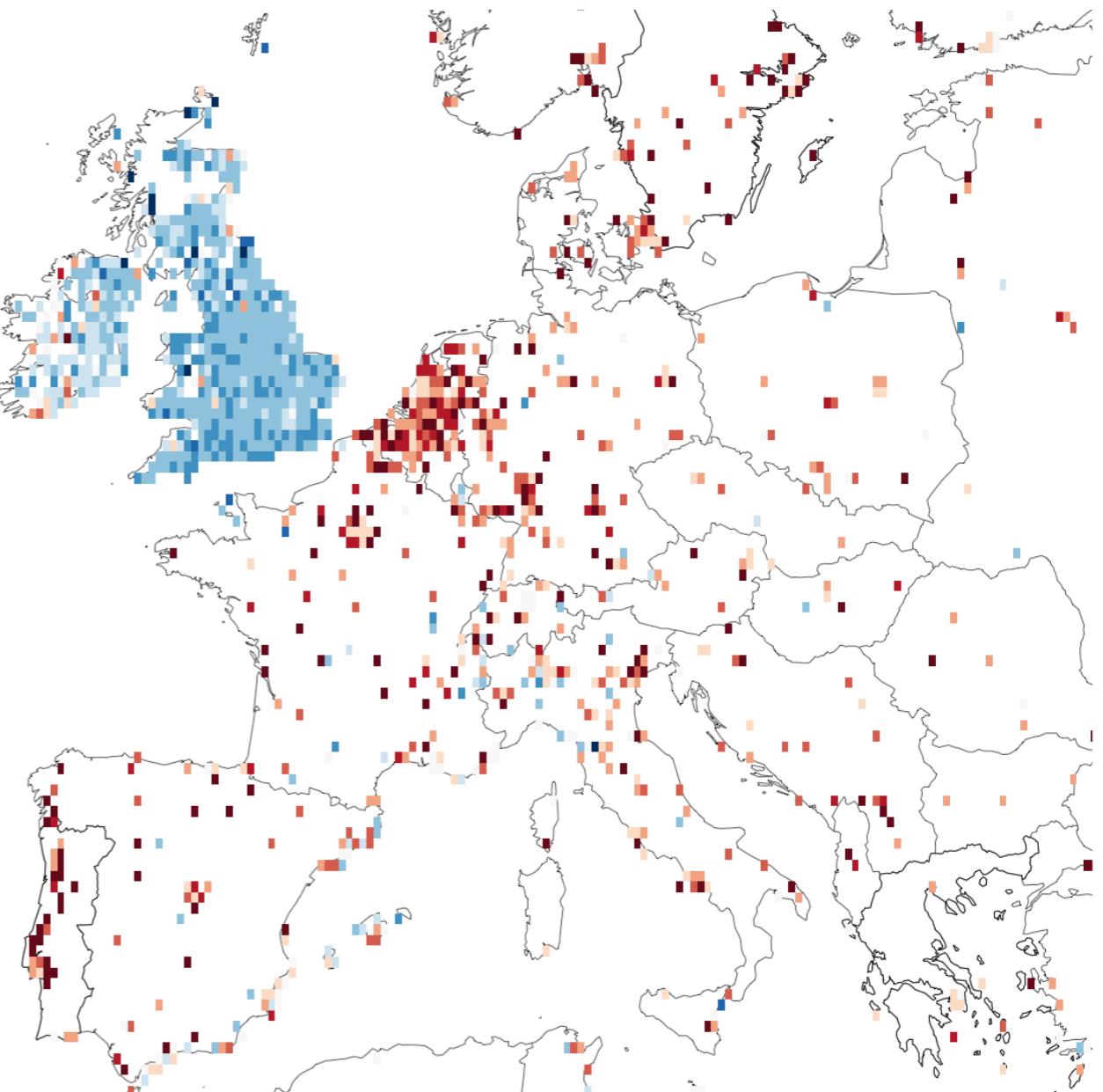
Geographical Variation

arXiv:1707.00781 (2017)

Spelling

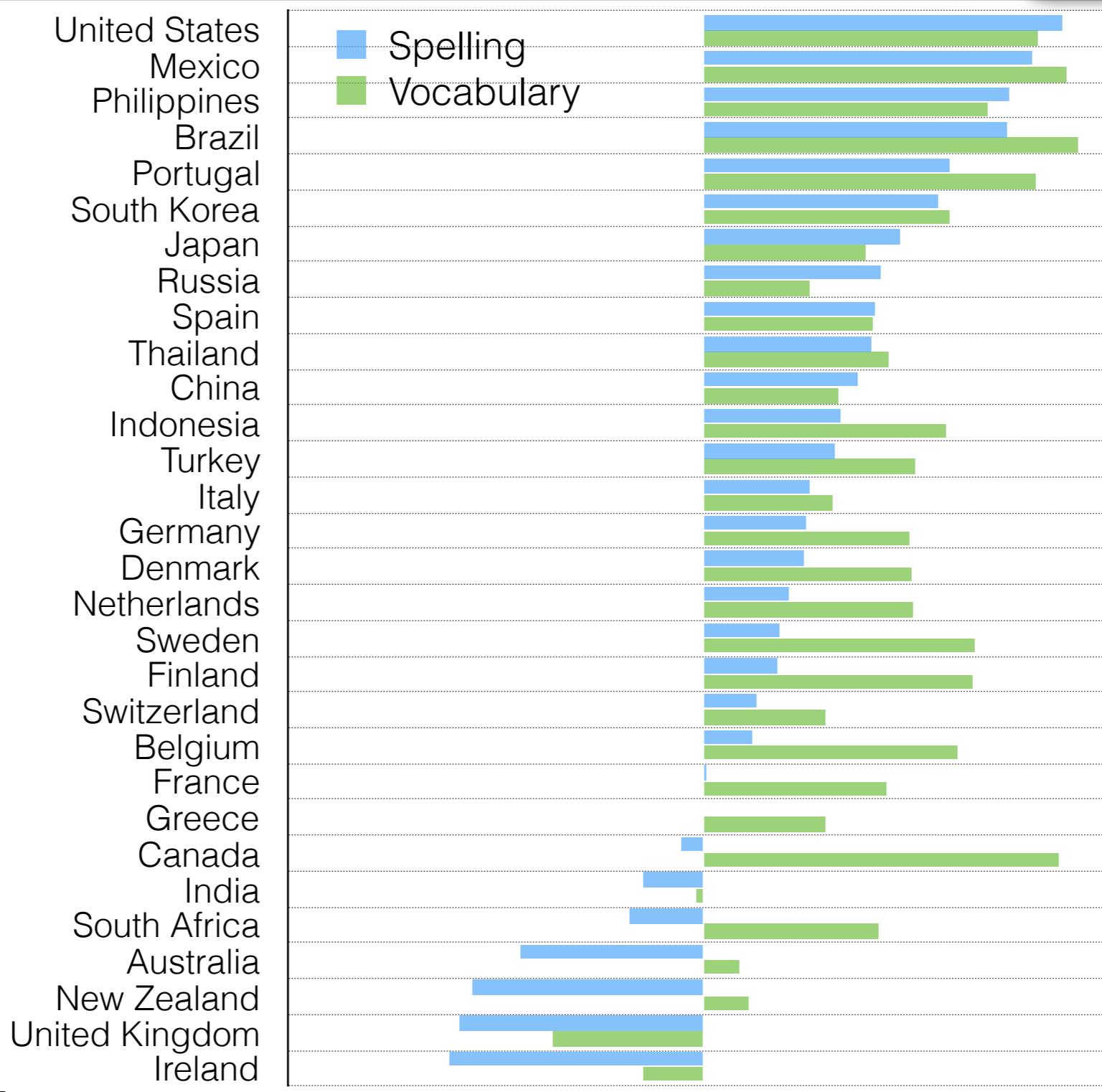


Vocabulary



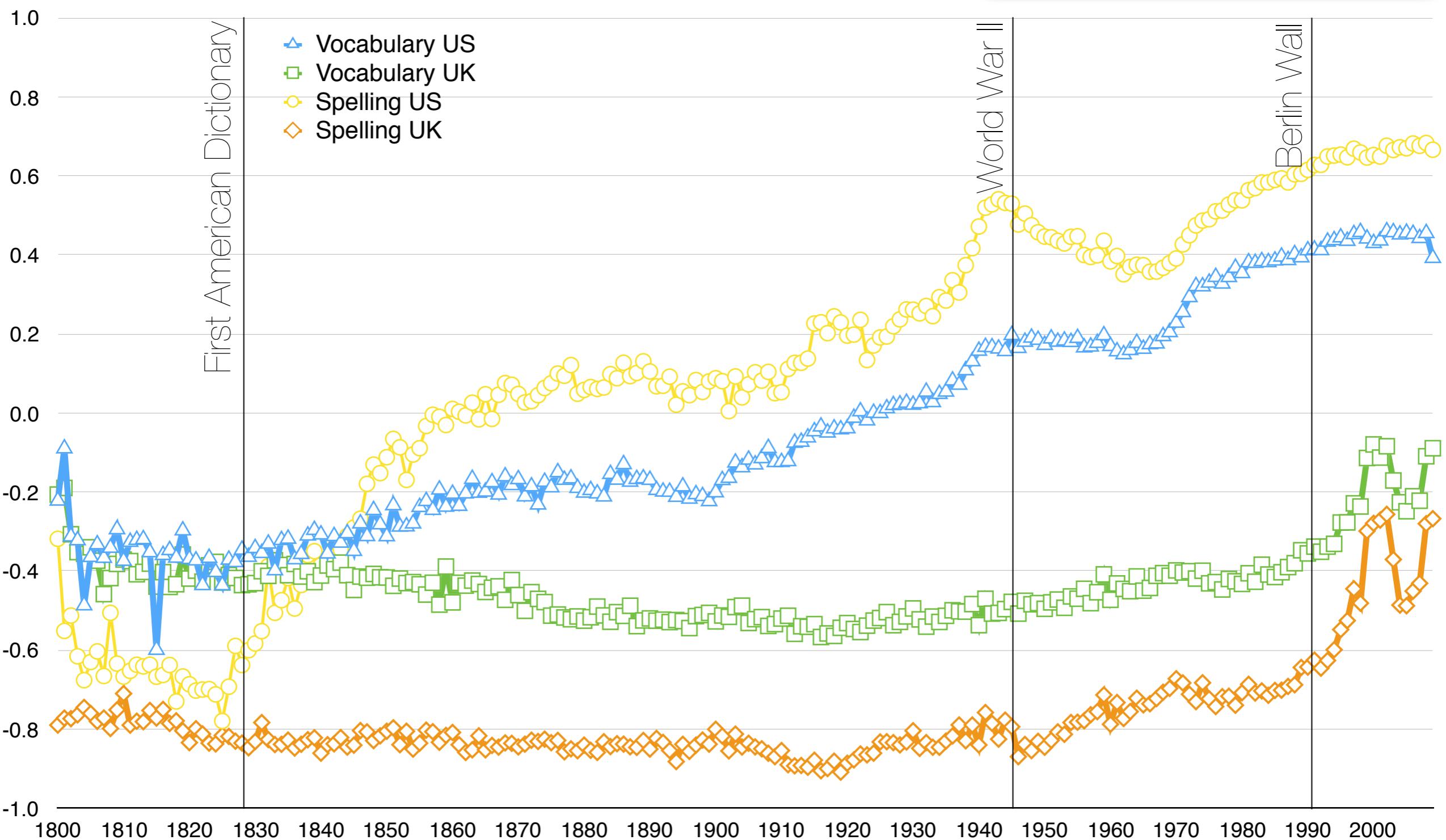
Analysis by country

arXiv:1707.00781 (2017)



Google Books

arXiv:1707.00781 (2017)



Community Integration



Migration is a global phenomenon
of increasing proportions in
our globalized world.

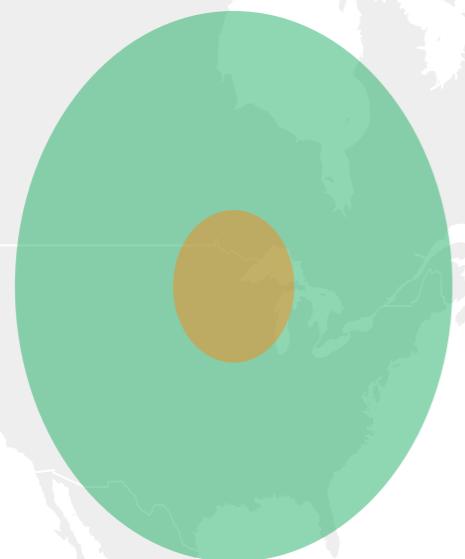
According to the United Nations estimate, in 2015
there were 244 Million immigrants in the world,
spread across all continents.

International Migrants Stock Dataset in 2015

- Indicates where international migrants live
- Indicates where international migrants come from
- The size of the circles is proportional to the number of migrants

North America

54 million
4 million



244 million

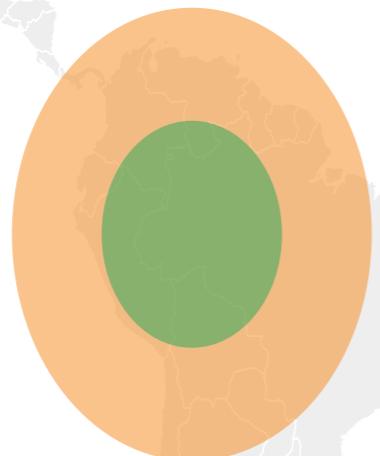


The number of international migrants reached 244 million in 2015.

An increase of 71 million since 2000.

Latin America and the Caribbean

9 million
37 million



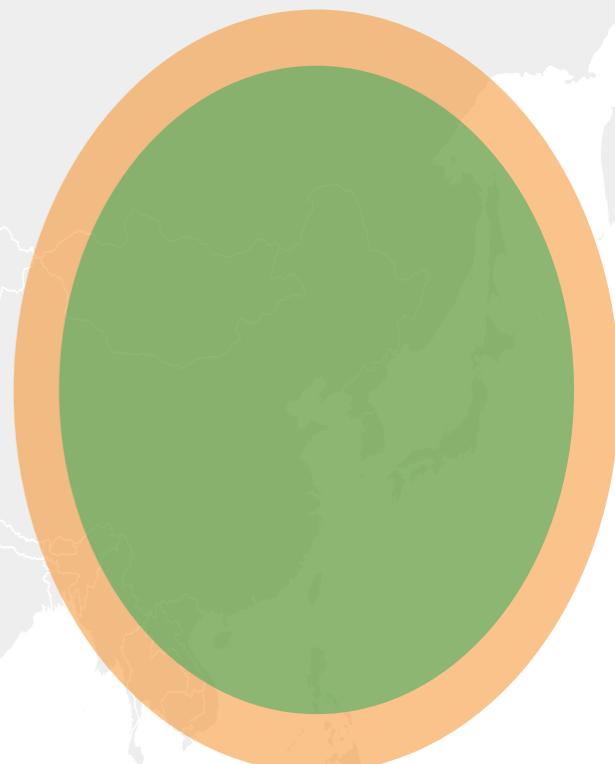
Europe

76 million
62 million



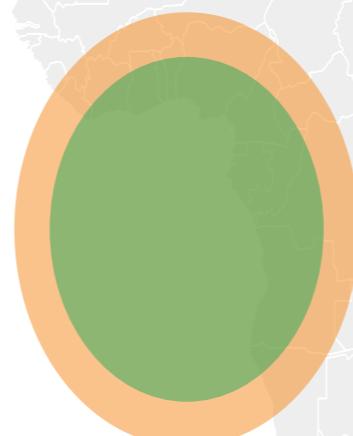
Asia

75 million
104 million



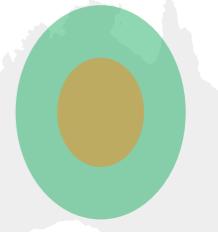
Africa

21 million
34 million



Oceania

8 million
2 million



Notes:

- All numbers are millions of people.
- Unknown residuals were redistributed proportionally to the size of groups for which data on international migrants were available by origin.

48%
are women

39
median age

15%
are below 20
years old

Source: United Nations, Department of Economic and Social Affairs, Population Division (2015).
Trends in International Migrant Stock: The 2015 revision.
(United Nations database, POP/DB/MIG/Stock/Rev.2015).
For more information visit: www.unmigration.org

International Migrants Stock Dataset in 2015

- Indicates where international migrants live
- Indicates where international migrants come from
- The size of the circles is proportional to the number of migrants

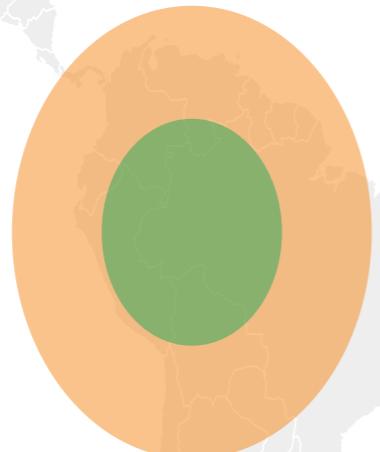
North America
54 4
million million

244 million

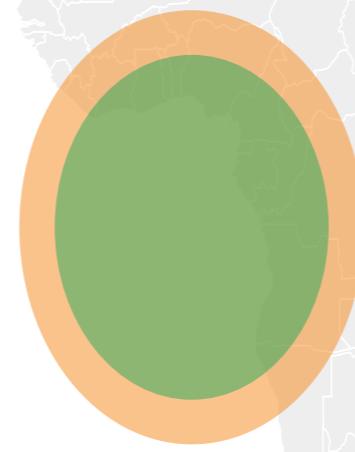


But how are migrants from different backgrounds integrating into their host communities?

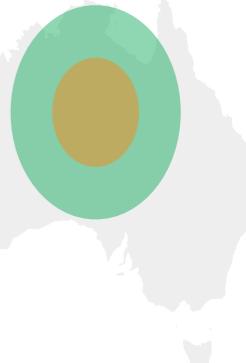
Latin America and the Caribbean
9 37
million million



Africa
21 34
million million



Oceania
8 2
million million



Notes:

- All numbers are millions of people.
- Unknown residuals were redistributed proportionally to the size of groups for which data on international migrants were available by origin.

48%
are women

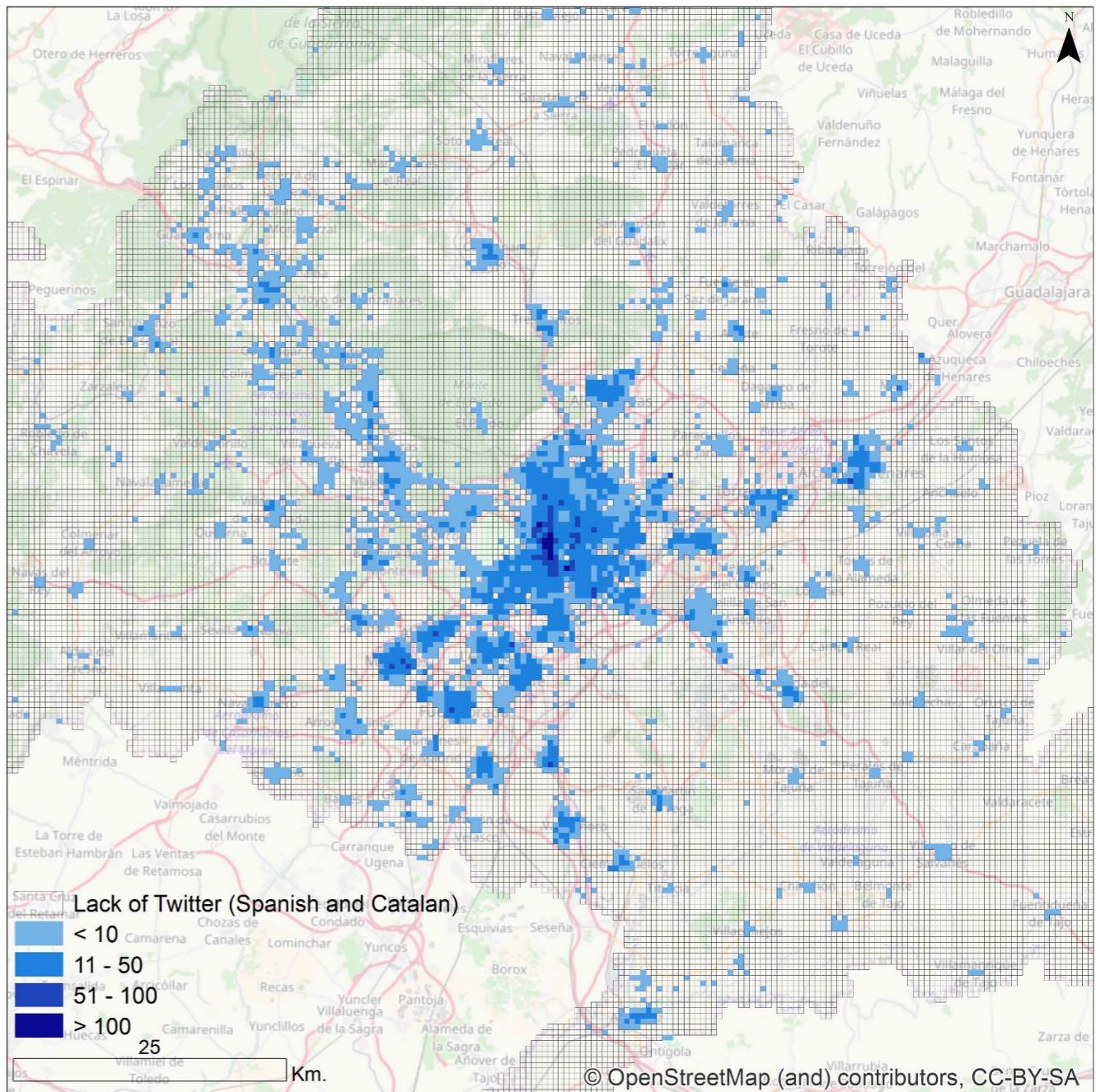
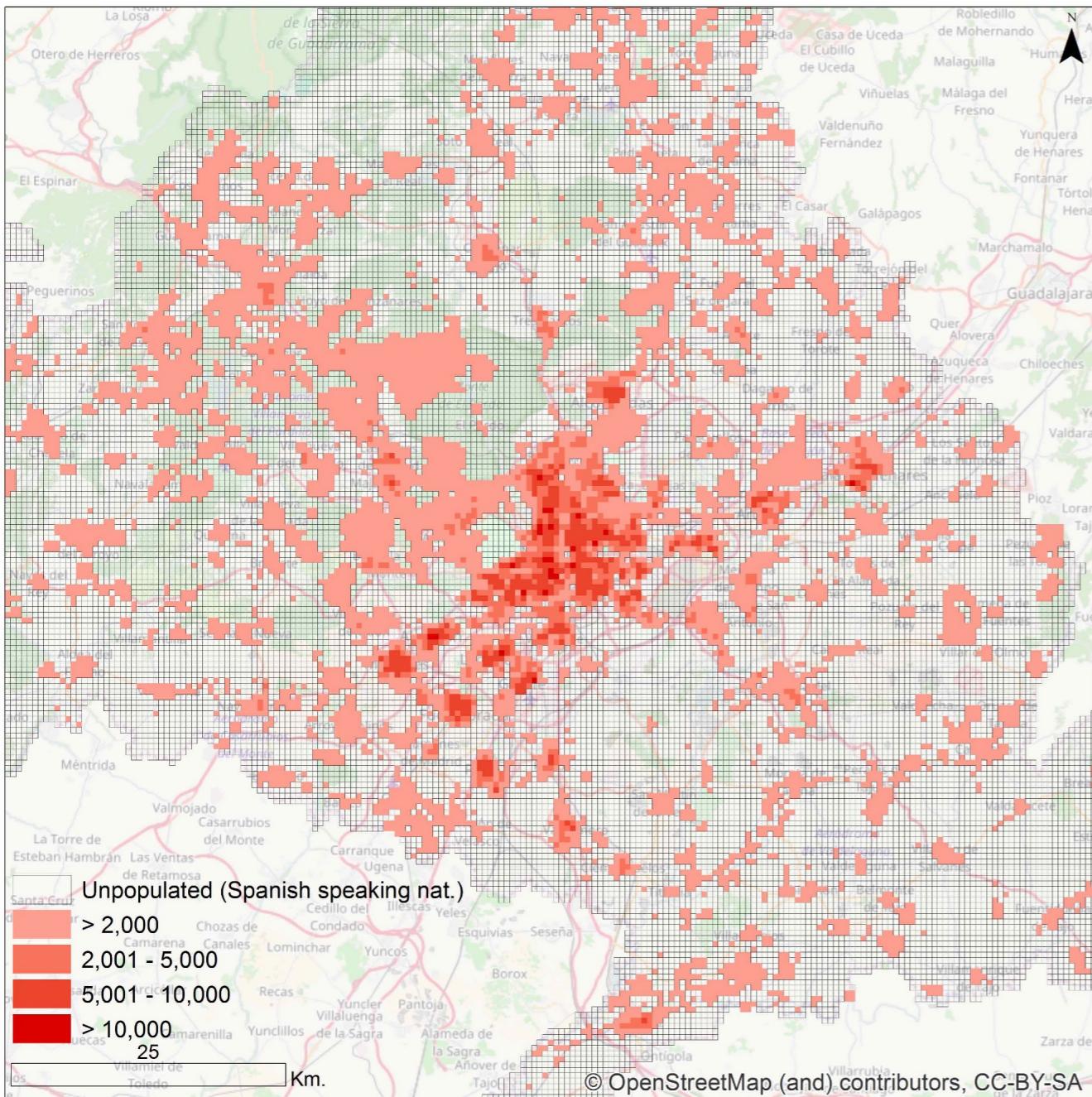
39
median age

15%
are below 20
years old

Source: United Nations, Department of Economic and Social Affairs, Population Division (2015).
Trends in International Migrant Stock: The 2015 revision.
(United Nations database, POP/DB/MIG/Stock/Rev.2015).
For more information visit: www.unmigration.org

Madrid

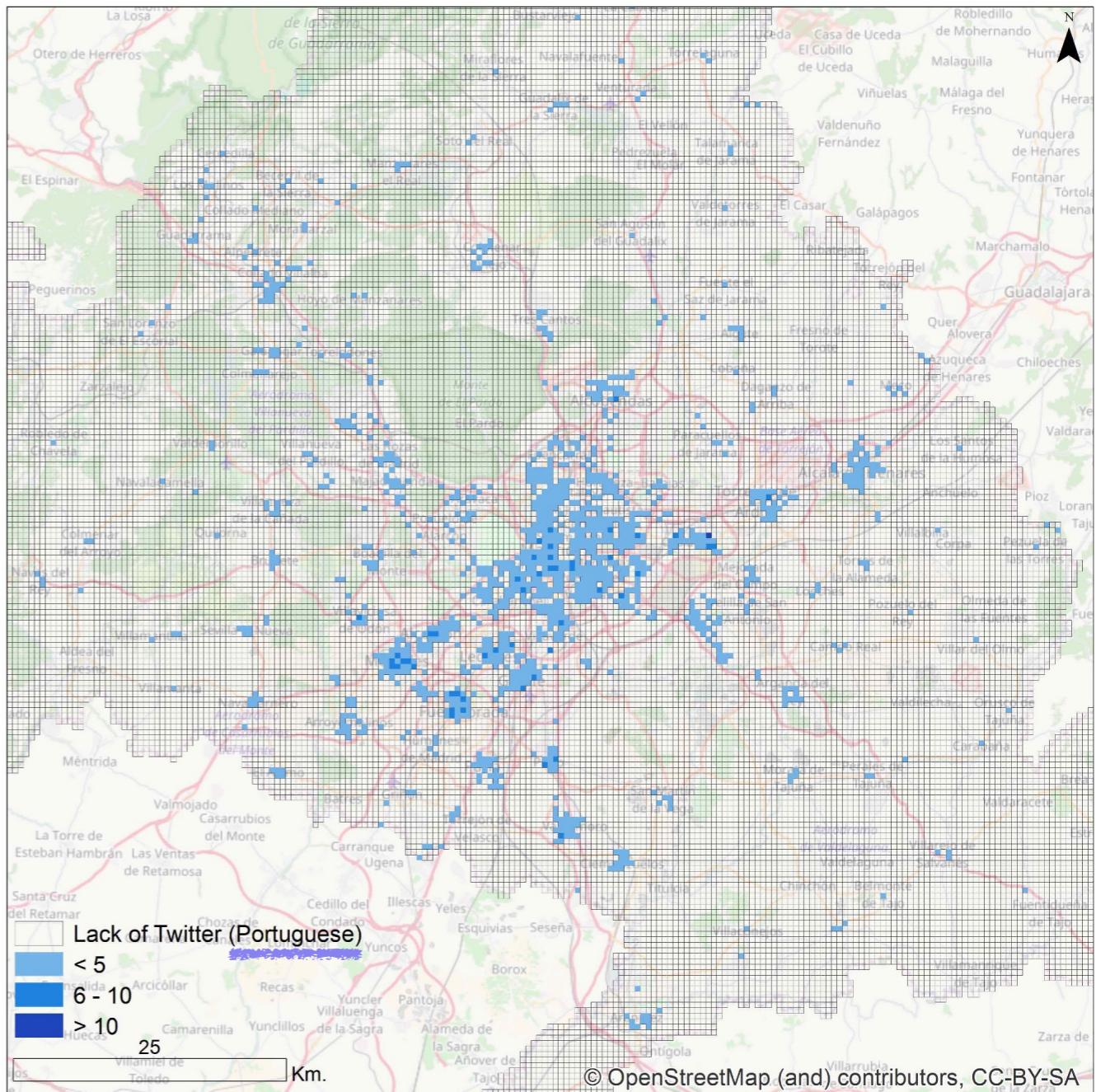
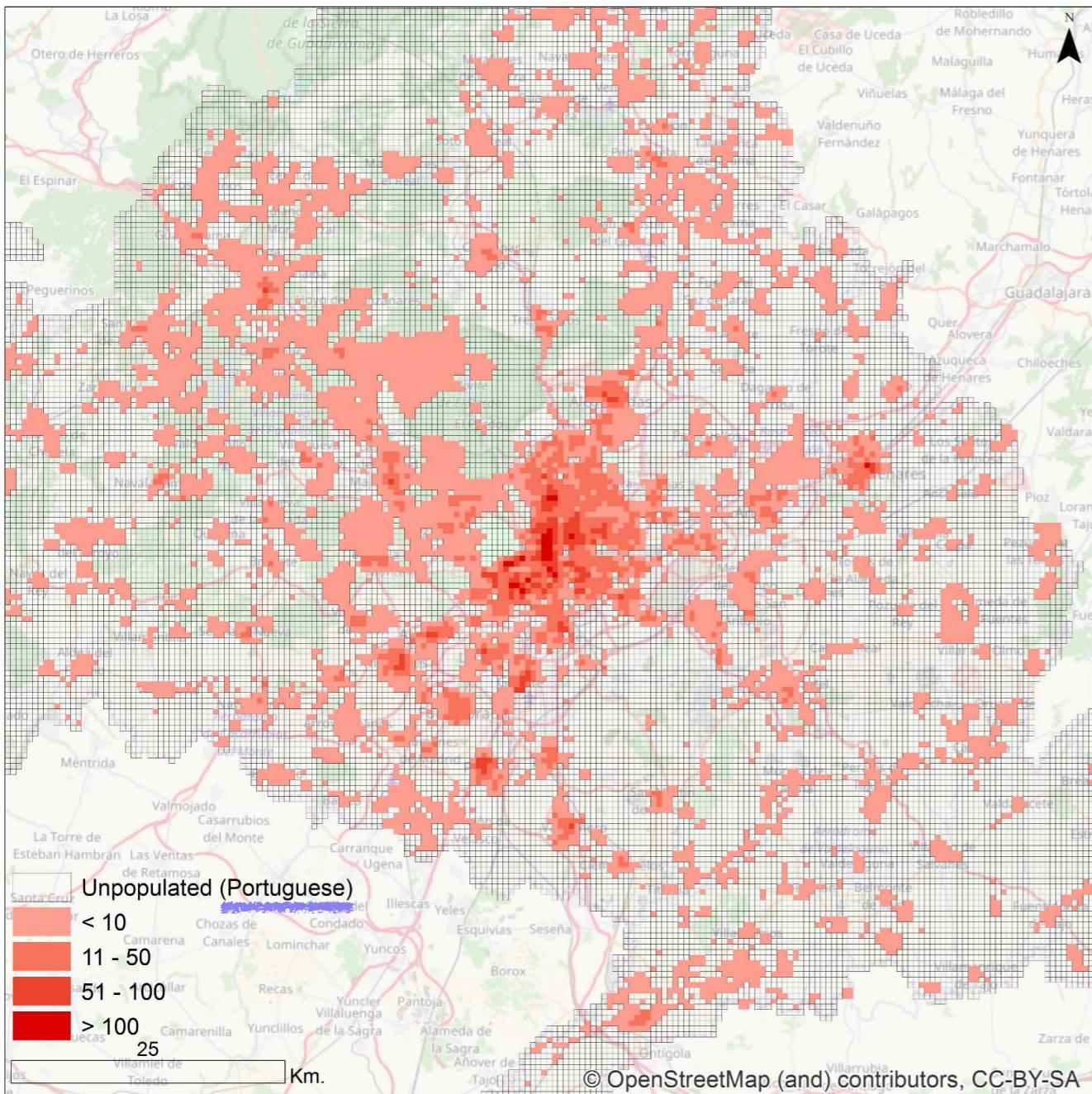
arXiv:1611.01056 (2016)



Twitter activity follows
population density

Madrid

arXiv:1611.01056 (2016)

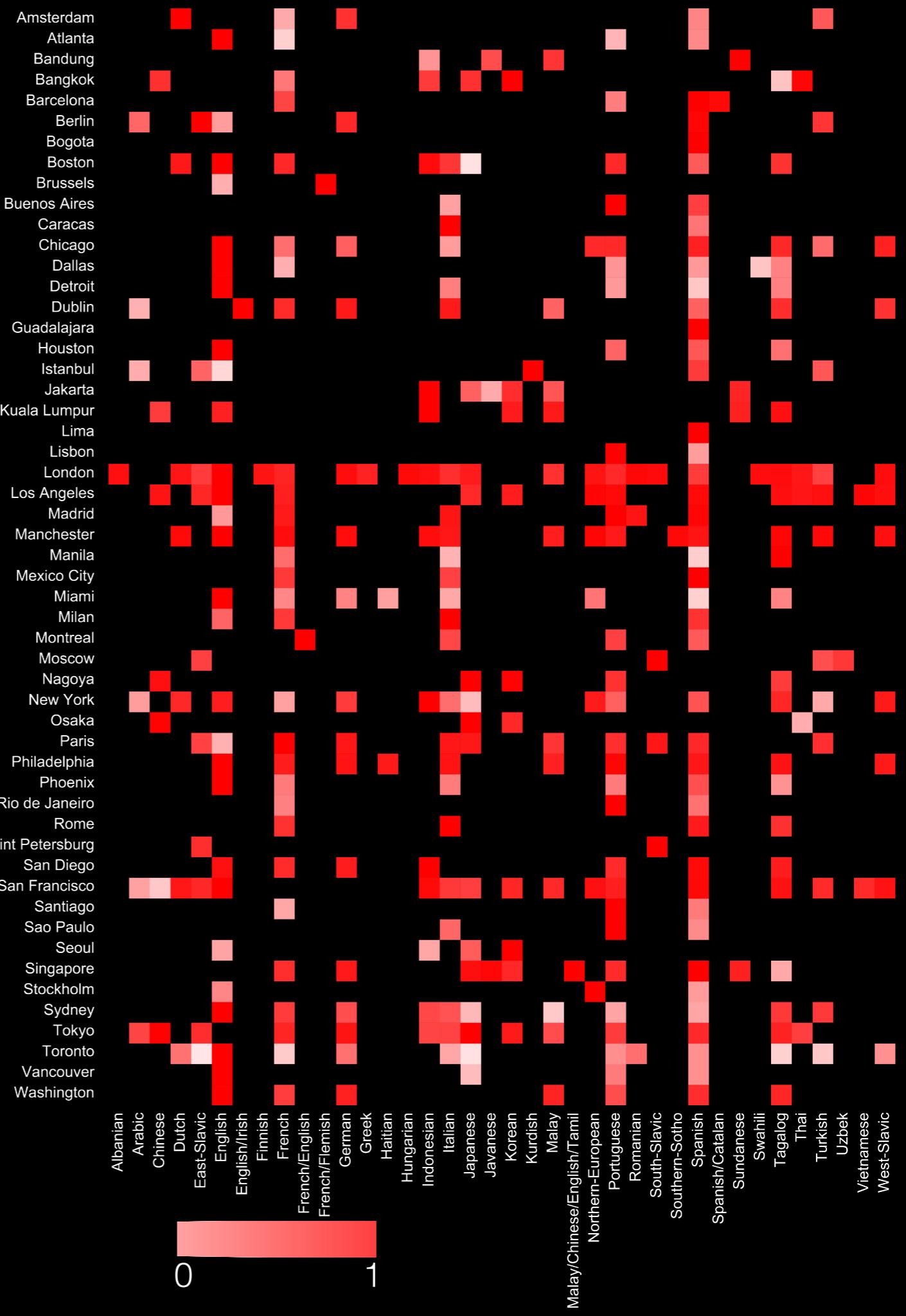


Language identifies
cultural communities

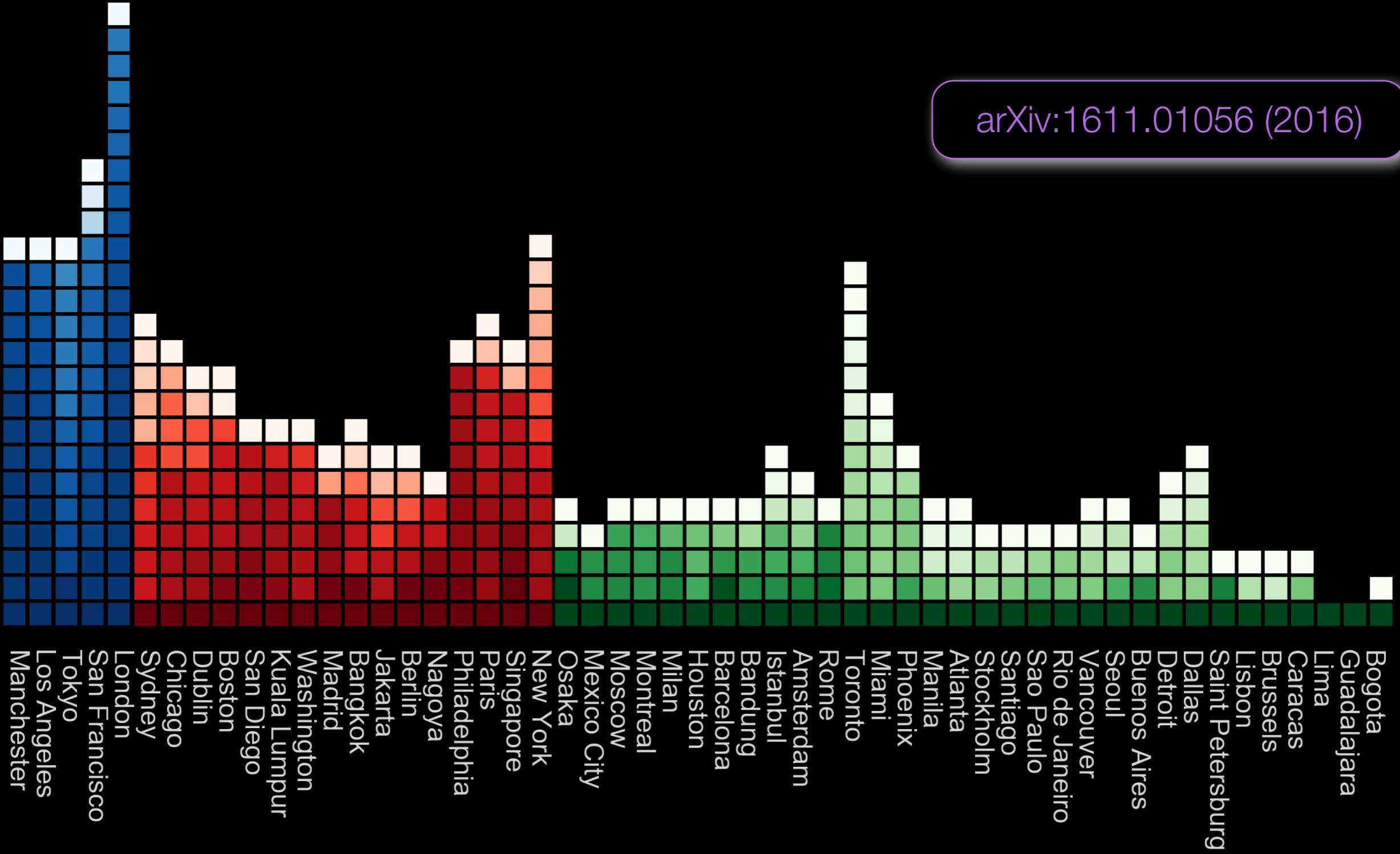
Each city can have multiple languages

Darker shades represent higher levels of **spatial** integration.

Are residents of a given culture **concentrated** in specific neighborhoods or **integrated** across all areas of the city?



Latent dimensions



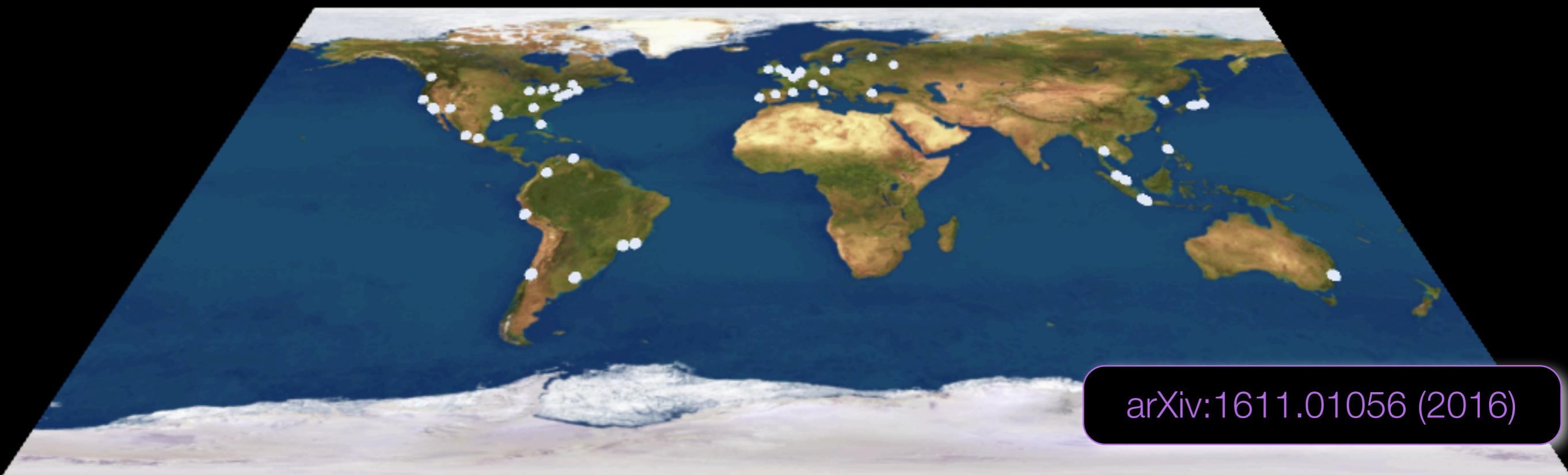
Cities naturally form
3 clusters

Cities naturally form 3 clusters

Amsterdam
Atlanta
Bandung
Bangkok
Barcelona
Berlin
Bogota
Boston
Brussels
Buenos Aires
Caracas
Chicago
Dallas
Detroit
Jakarta
Dublin
Guadalajara
Houston
Istanbul
Kuala Lumpur
Lima
Lisbon
London
Los Angeles
Madrid
Manchester
Manila

Mexico City
Miami
Milan
Montreal
Moscow
Nagoya
New York
Osaka
Paris
Philadelphia
Phoenix
Rio de Janeiro
Rome
Saint Petersburg
San Diego
San Francisco
Santiago
Sao Paulo
Seoul
Singapore
Stockholm
Sydney
Tokyo
Toronto
Vancouver
Washington

That are based on cultural
integration and not on Geography!



At the **country** level we find the Top integrators:

Top 10%
Top 20%



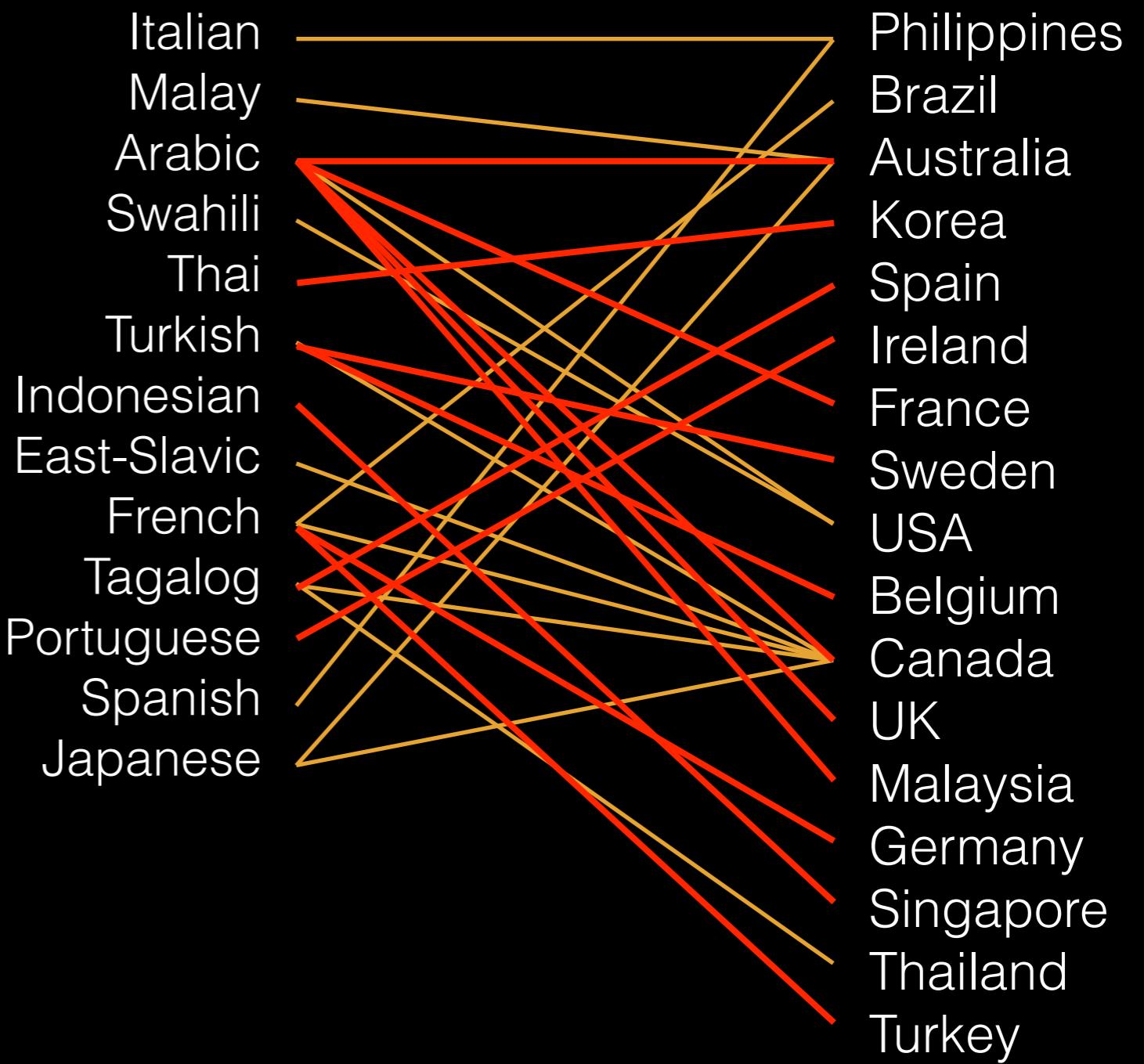
At the **country** level we find the Top integrators:

London and Manchester play the dominant role in shaping the UK's strong Power of Integration



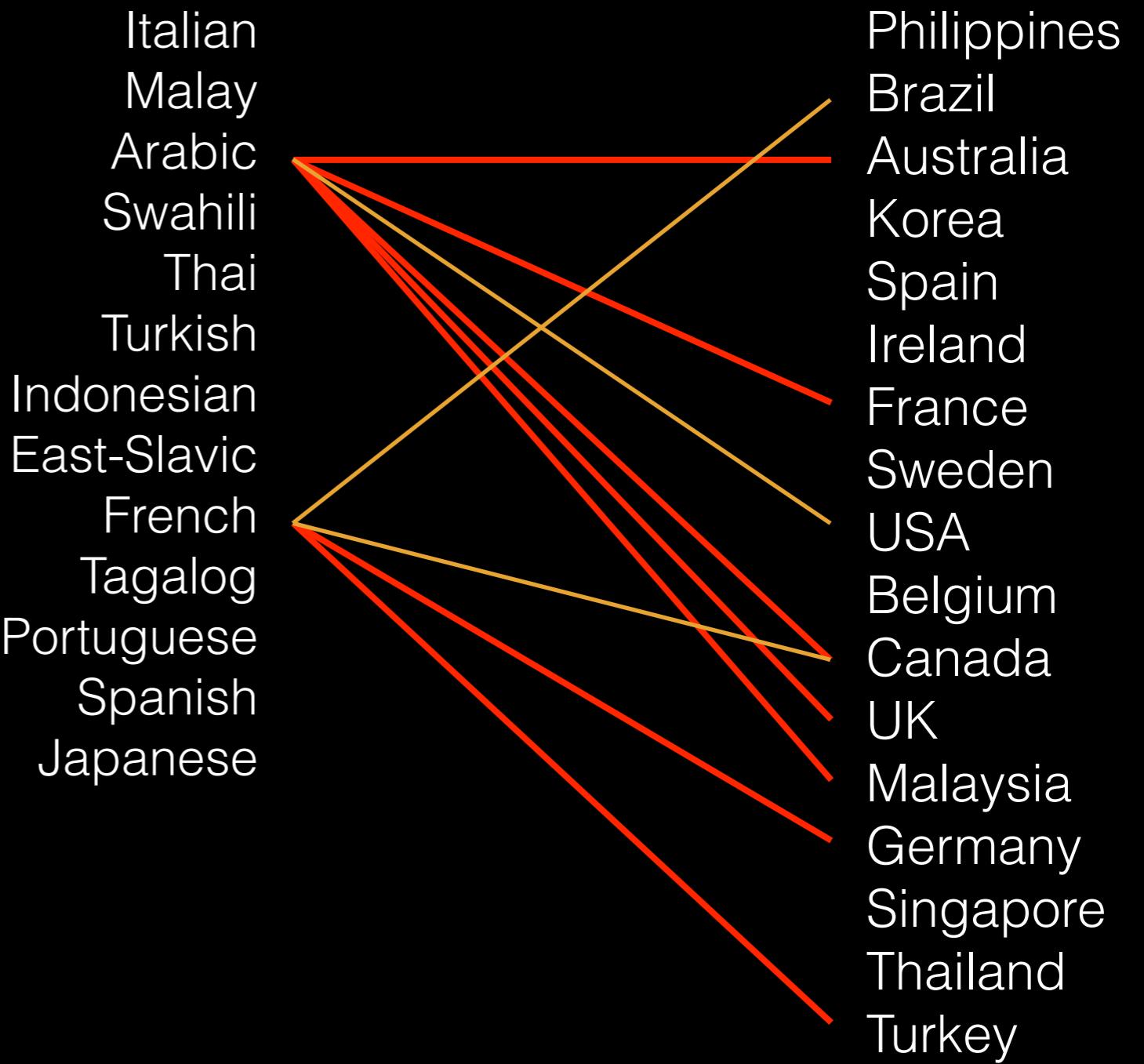
At the **language** level we find the Bottom integrators:

Bottom 10%
Bottom 20%



At the **language** level we find the Bottom integrators:

French and Arabic communities consistently concentrate in specific areas



Discussion

- Online Social Networks generate unprecedented amounts of data on Human Behavior
- The massification of GPS-enabled devices allows us to observe geographical variations
- Language Detection tools and Language restricted datasets provide a unique view on language use in the Real World
- The geographical variation of Spanish synonym use in Twitter allows us to empirically define dialects
- Language co-use is consistent across diverse datasets and highlights the existence of a Global Language Network
- English is increasingly becoming americanized as a result of Global Historical Events
- Languages can be used as proxies for Communities to measure Integration in global cities