# **Social media fingerprints** of ***unemployment***

---

Alejandro Llorente (IIC, Spain)
Manuel García Herranz (UAM and United Nations, NY)
Manuel Cebrián (NICTA, Australia)
Esteban Moro (UC3M, Spain)

PLoS ONE 10(5): e0128692

WE
*are*
WHAT WE
*repeatedly*
DO.

-aristotle

# Computational Social Science

**You are** what *you repeatedly do* *[Aristóteles]*

Using BigData to infer behavior or society situation

**Situation**
- Demographics
- Health
- Economy
- Unemployment
- Transportation
- Geography
- Politics

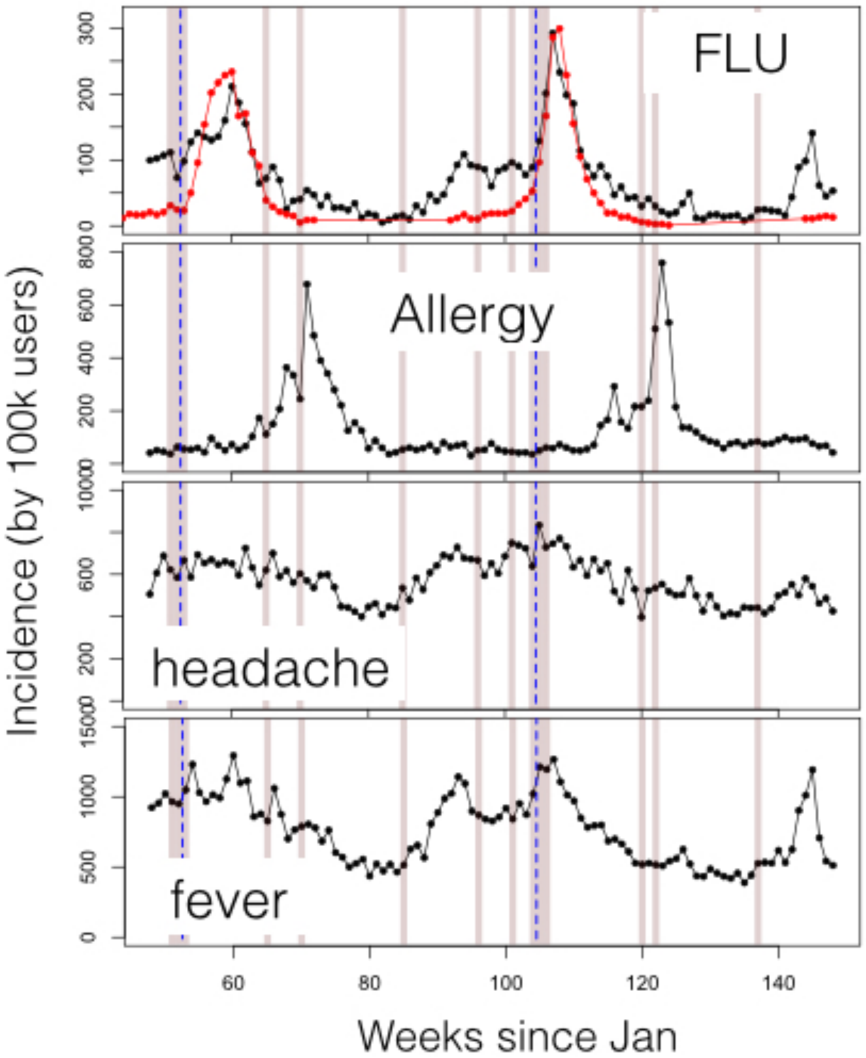
**Behavior**
- Social
- Mobility
- Activity
- Content

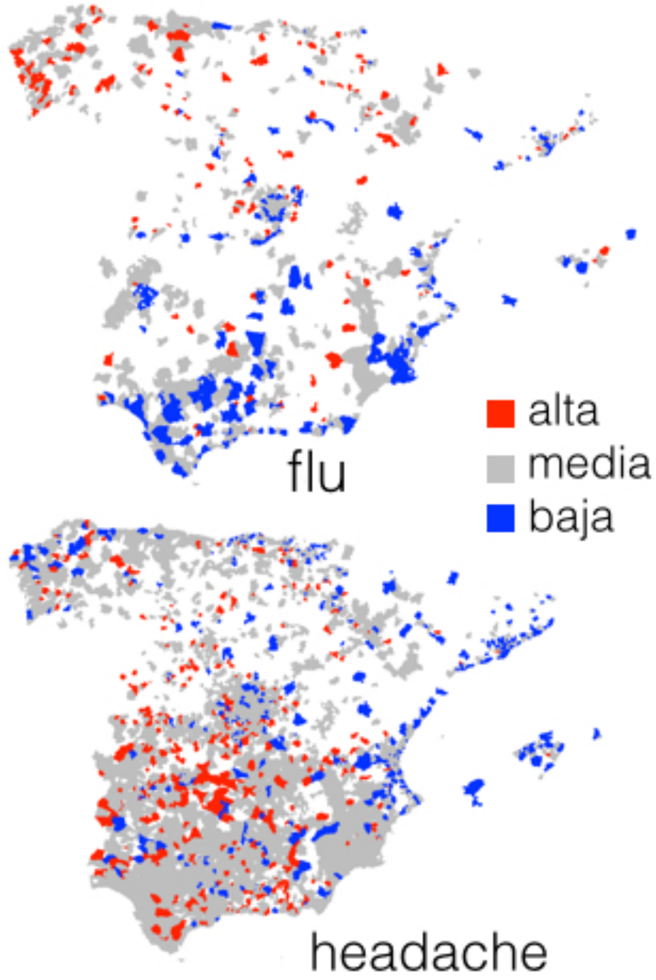
**Observation**
- Surveys
- Credit card
- Mobile phone
- Social media
- Searches
- …

Individual - Group - City

# Computational Social Science

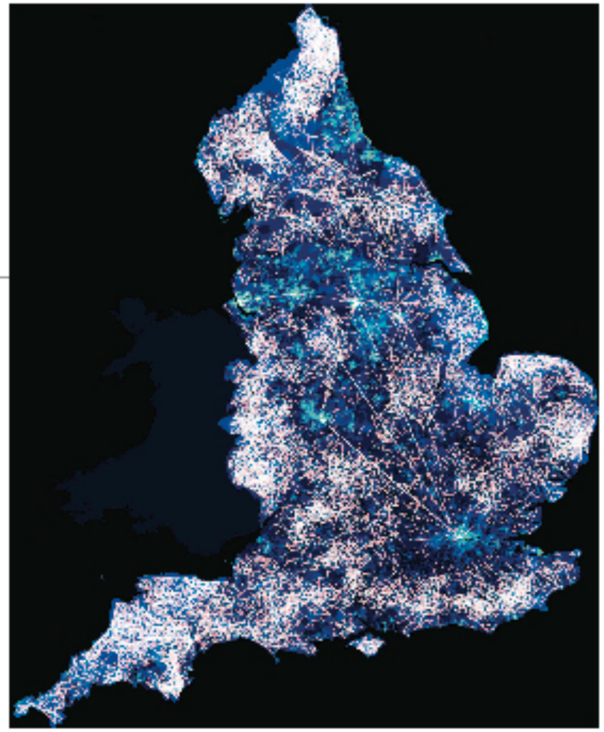
## Health -> Content -> Social Media




Incidence

# Computational Social Science



**Economy** -> **Behavior** -> **Observation**

Eagle et al, Science 2010

Development -> Social/Mobility (city) diversity -> Mobile phones

Soto, V. et al., Prediction of Socioeconomic Levels using Cell Phone Records 2011.

Socio economical level -> Social/Mobility/Activity -> Mobile phones

Antenucci, D. et al., 2014. Using Social Media to Measure Labor Market Flows.

Unemployment -> Content -> Social Media

# Computational Social Science

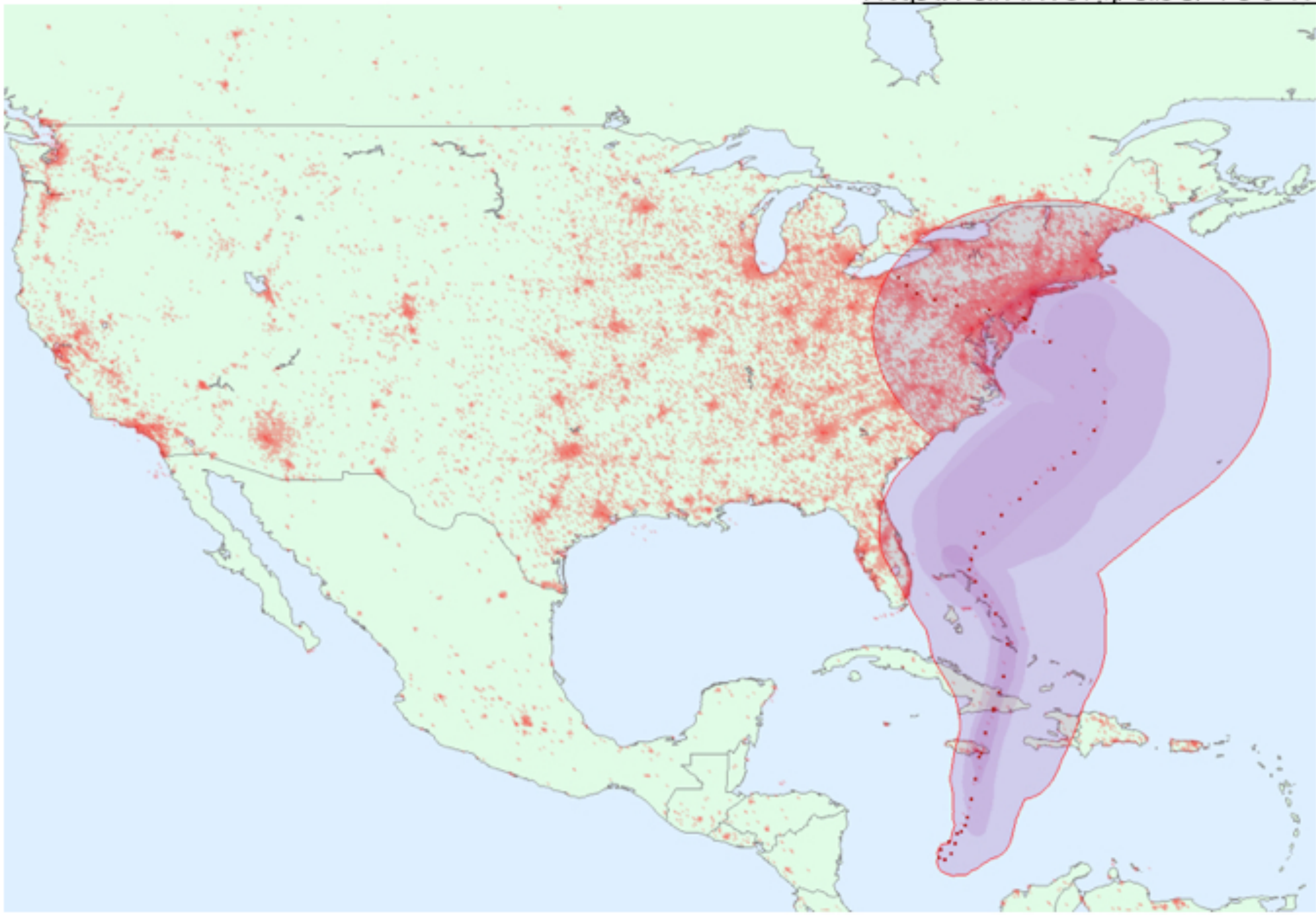
Wealth -> Behavior -> Credit card

Krumme, C. et al., 2013. The predictability of consumer visitation patterns. Scientific Reports

# Computational Social Science

Disaster -> Behavior -> Social Media

Dataset: 52.55 Million messages, 14Million users
Yury Kryvasheyeu, Manuel Cebrián, EM, et al 2015
http://arxiv.org/abs/1504.06827

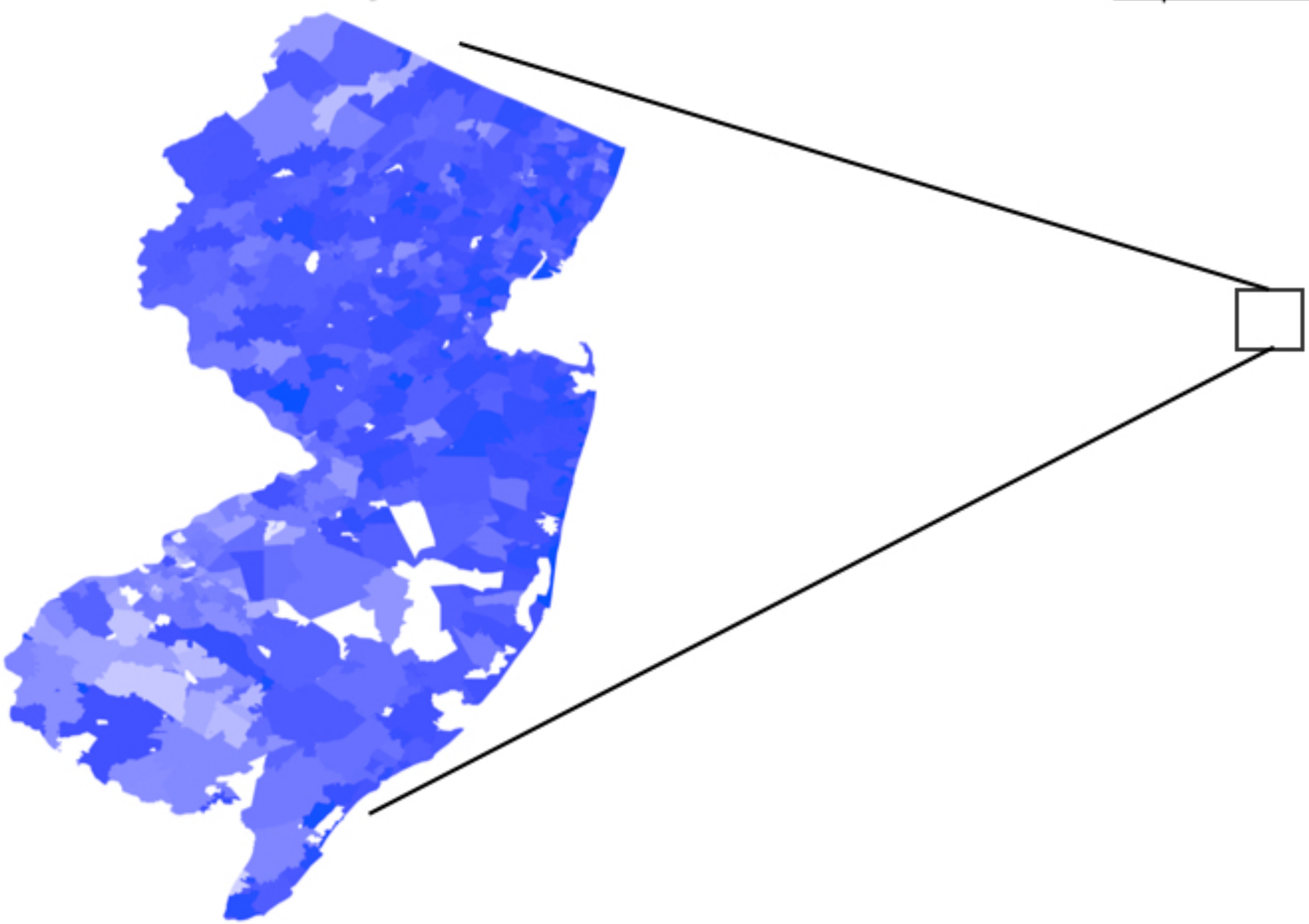


Economical Impact ~ $1Billion

# Computational Social Science

Disaster -> Behavior -> Social Media

Dataset: 52.55 Million messages, 14Million users
Yury Kryvasheyeu, Manuel Cebrián, EM, et al 2015
http://arxiv.org/abs/1504.06827

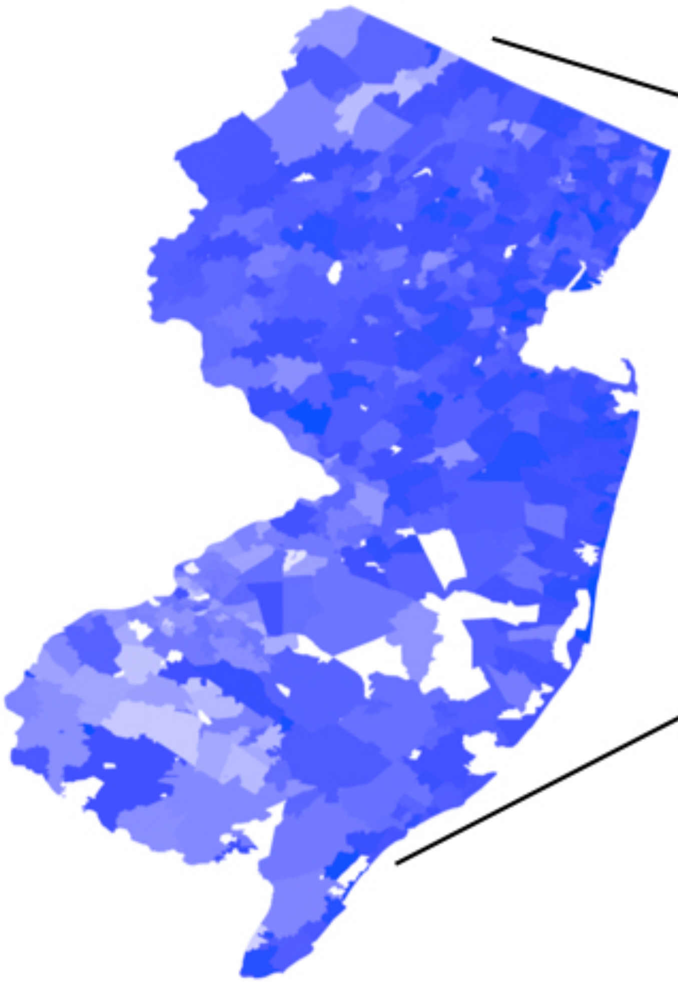


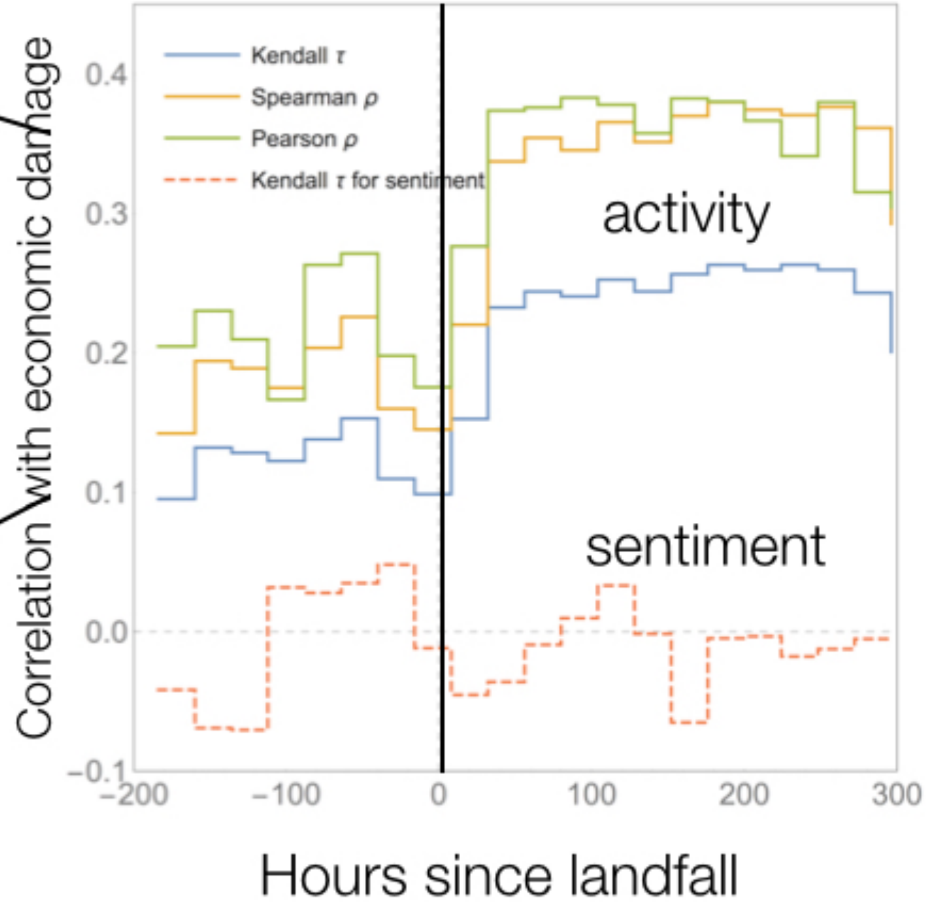
Economical Impact ~ $1Billion

# Computational Social Science

Disaster -> Behavior -> Social Media

Dataset: 52.55 Million messages, 14Million users
Yury Kryvasheyeu, Manuel Cebrián, EM, et al 2015
http://arxiv.org/abs/1504.06827



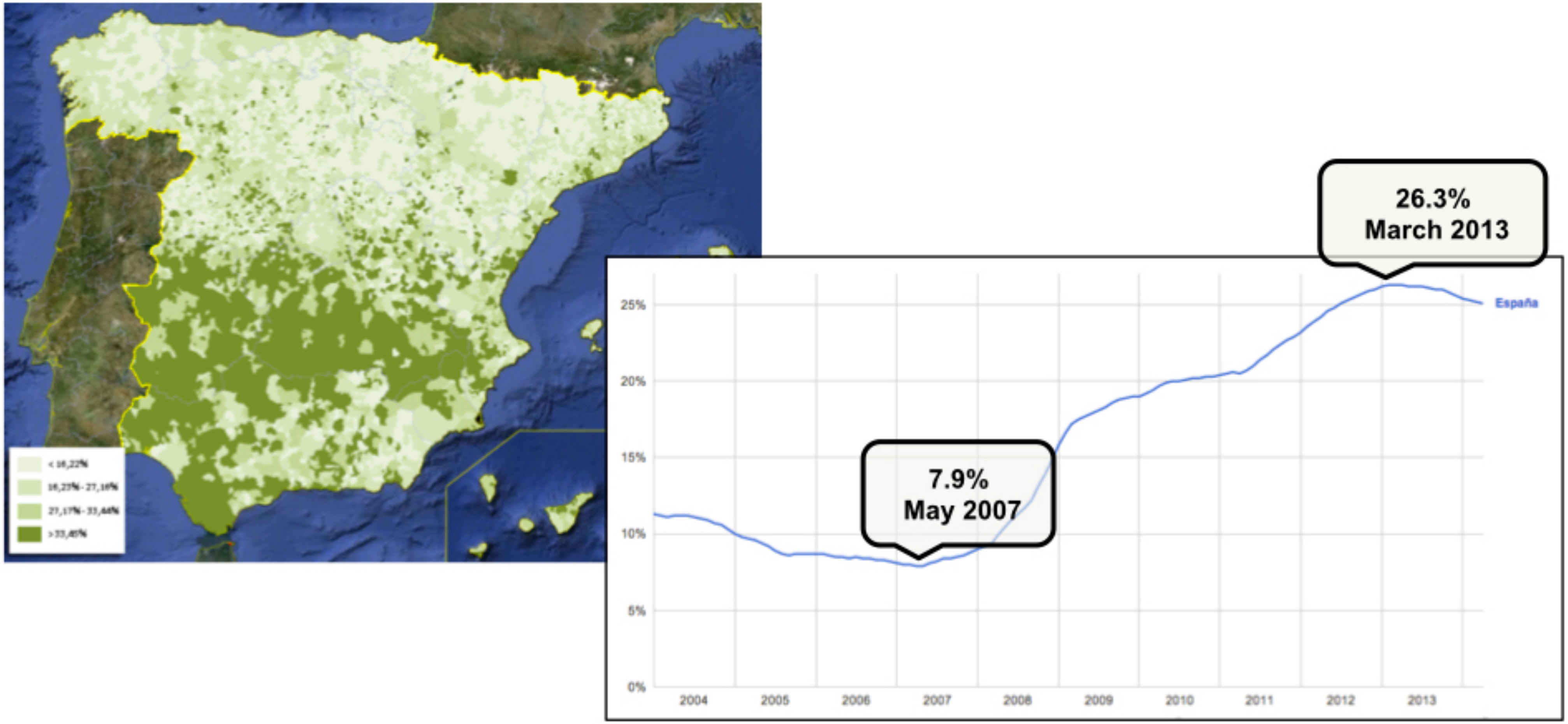



Economical Impact ~ $1Billion

# Our objective

**Unemployment** -> **Behavior** -> **Social media (Twitter)**

- It's not another economical problem in Spain. It is the BIG problem



# Our objective

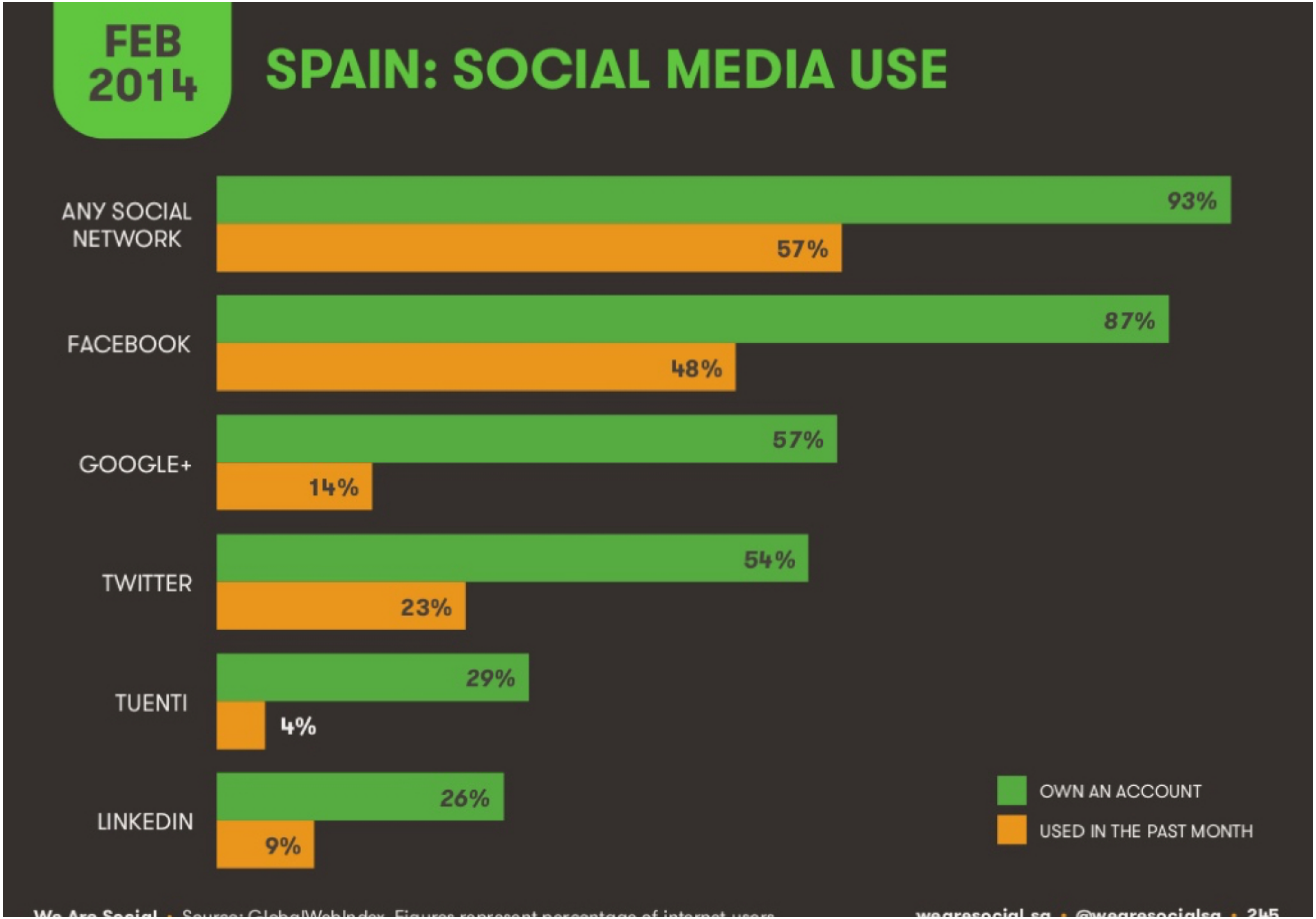
**Unemployment** -> **Behavior** -> **Social media (Twitter)**

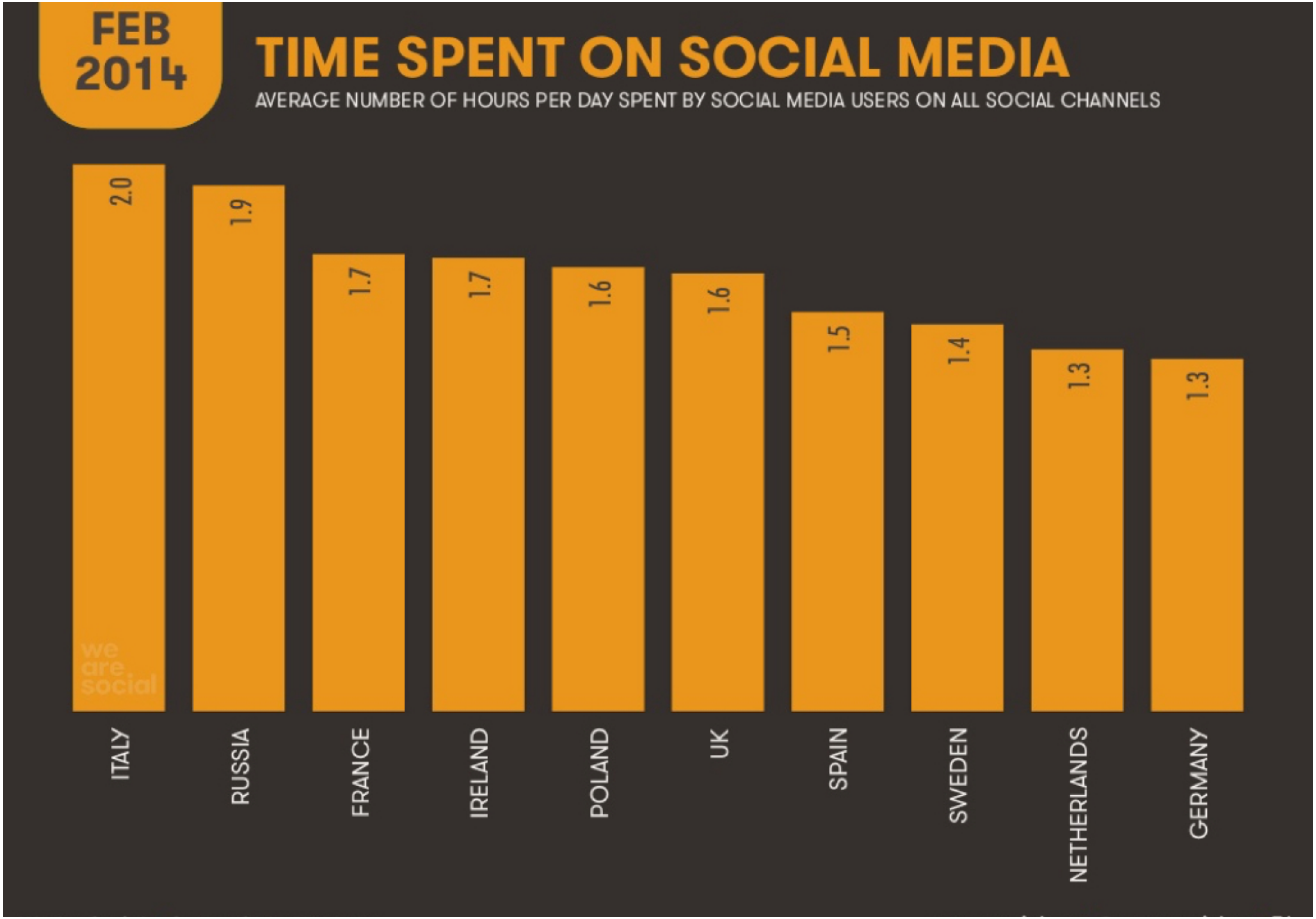
- Twitter allows us to study social/mobility/content behaviors together
  - *But, which behavior(s) are the more relevant with respect to unemployment? (360º view)*
- Twitter is **BigSparseData**: some demographic and geographic areas are not well represented, thus
  - *How much of the observed unemployment can be described by Twitter-detected behaviors?*

FEB 2014

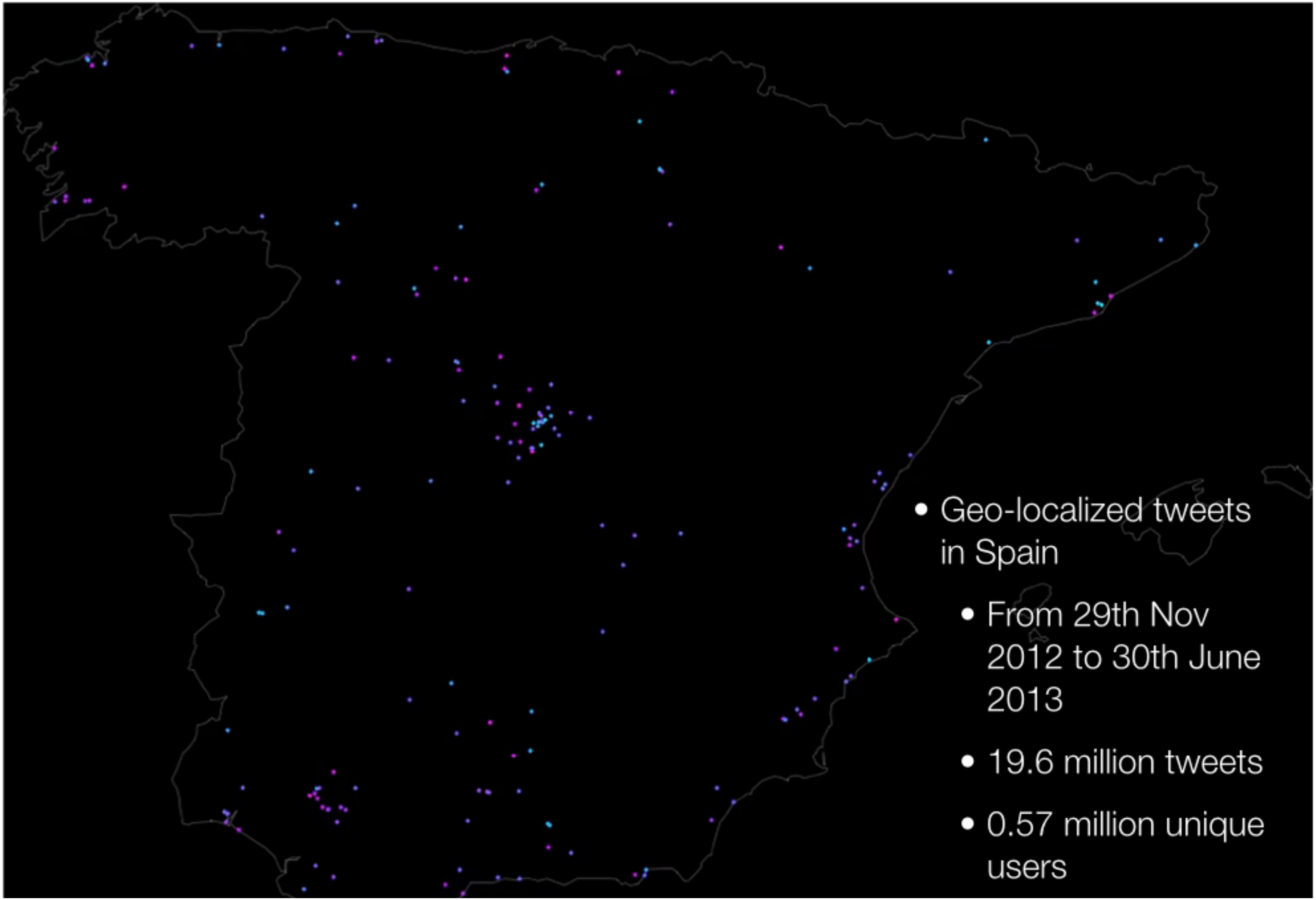
# SPAIN

| | | |
|---|---|---|
| 47,370,542 | TOTAL POPULATION | URBAN 77% / RURAL 23% |
| 33,870,948 | INTERNET USERS | INTERNET PENETRATION 72% |
| 19,600,000 | ACTIVE FACEBOOK USERS | FACEBOOK PENETRATION 41% |
| 55,740,000 | ACTIVE MOBILE SUBSCRIPTIONS | MOBILE SUBSCRIPTION PENETRATION 118% |


FEB 2014
SPAIN: SOCIAL MEDIA USE
ANY SOCIAL NETWORK
93%
57%
FACEBOOK
87%
48%
GOOGLE+
57%
14%
TWITTER
54%
23%
TUENTI
29%
4%
LINKEDIN
26%
9%
OWN AN ACCOUNT
USED IN THE PAST MONTH
We Are Social • Source: GlobalWebIndex. Figures represent percentage of internet users
wearesocial.sg • @wearesocialsg • 245

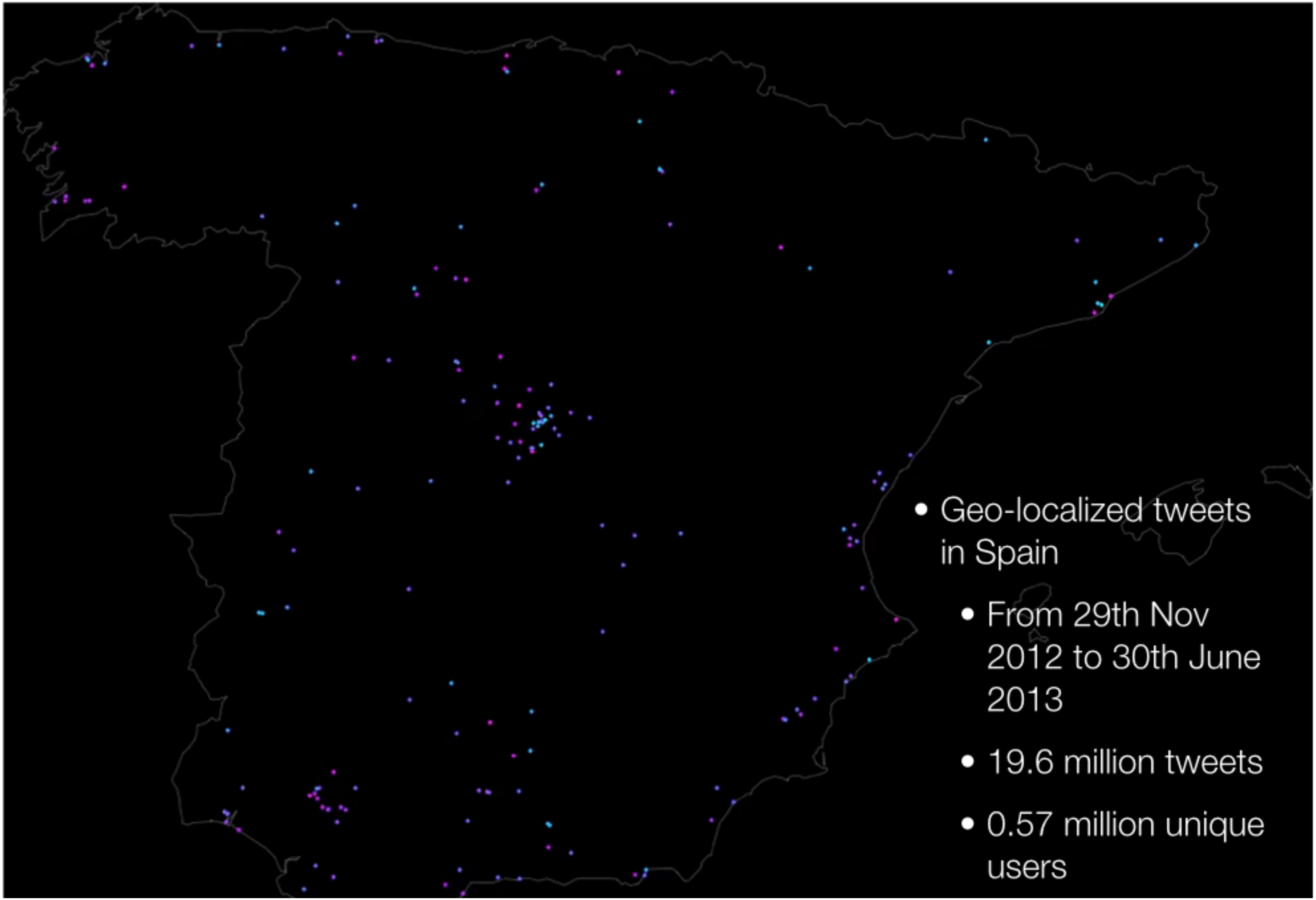

FEB 2014
TIME SPENT ON SOCIAL MEDIA
AVERAGE NUMBER OF HOURS PER DAY SPENT BY SOCIAL MEDIA USERS ON ALL SOCIAL CHANNELS
ITALY 2.0
RUSSIA 1.9
FRANCE 1.7
IRELAND 1.7
POLAND 1.6
UK 1.6
SPAIN 1.5
SWEDEN 1.4
NETHERLANDS 1.3
GERMANY 1.3

# Our database



- Geo-localized tweets in Spain
  - From 29th Nov 2012 to 30th June 2013
  - 19.6 million tweets
  - 0.57 million unique users

# Our database



- Geo-localized tweets in Spain
    - From 29th Nov 2012 to 30th June 2013
    - 19.6 million tweets
    - 0.57 million unique users

# Geographical areas in Spain

- Municipalities:
  - ~8200 municipalities in Spain
  - Very heterogeneous: population ranging from 7 to 3.2
- We propose a functional approach to the definition of the areas based on daily mobility
- *Users' municipality home is where they tweet most*

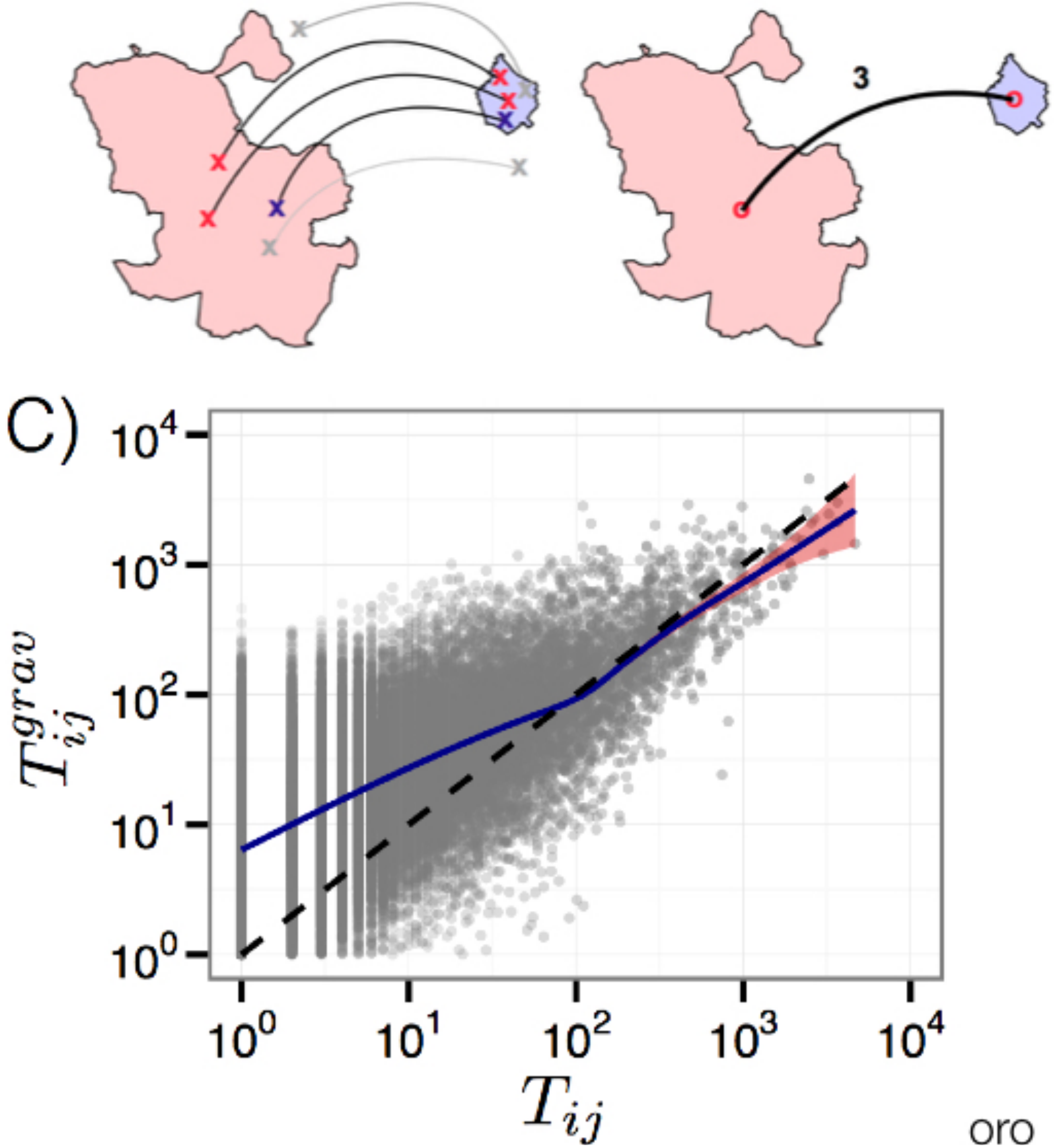
# Geographical areas in Spain

- We take all trips between two municipalities to construct the flow $T_{ij}$ the number of trips between municipalities
- Flow is well described by the gravity model

$$T_{ij} \simeq T_{ij}^{grav} = \frac{P_i^{\alpha_i} P_j^{\alpha_j}}{d_{ij}^{\beta}}$$

$$\alpha_i \approx \alpha_j = 0.42, \beta = 0.89$$

$$R^2 = 0.69$$

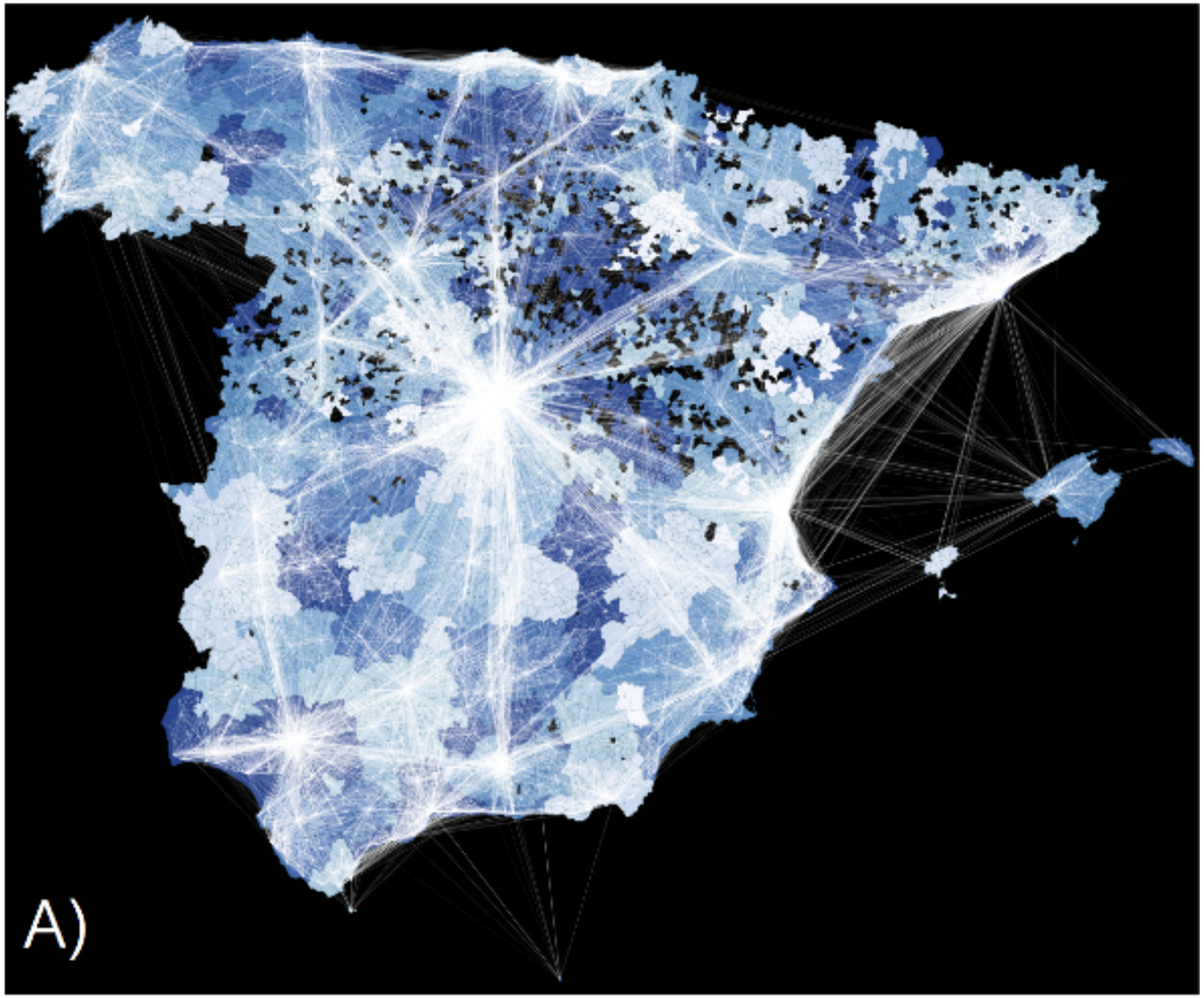



Lenormand, M. et al., 2014. Cross-checking different sources of mobility information.

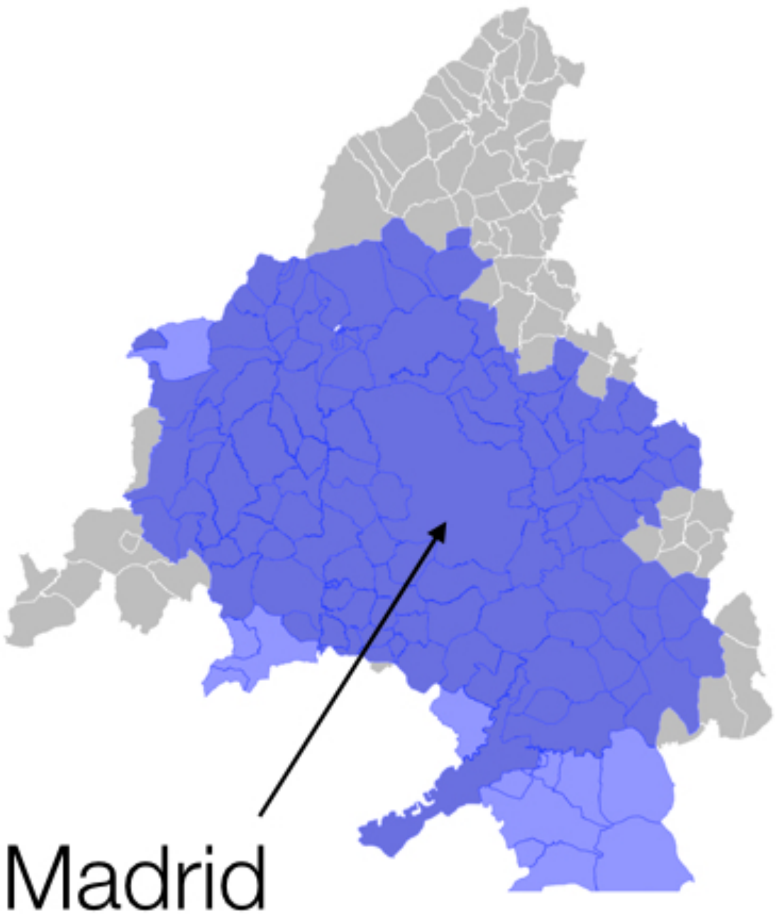
oro

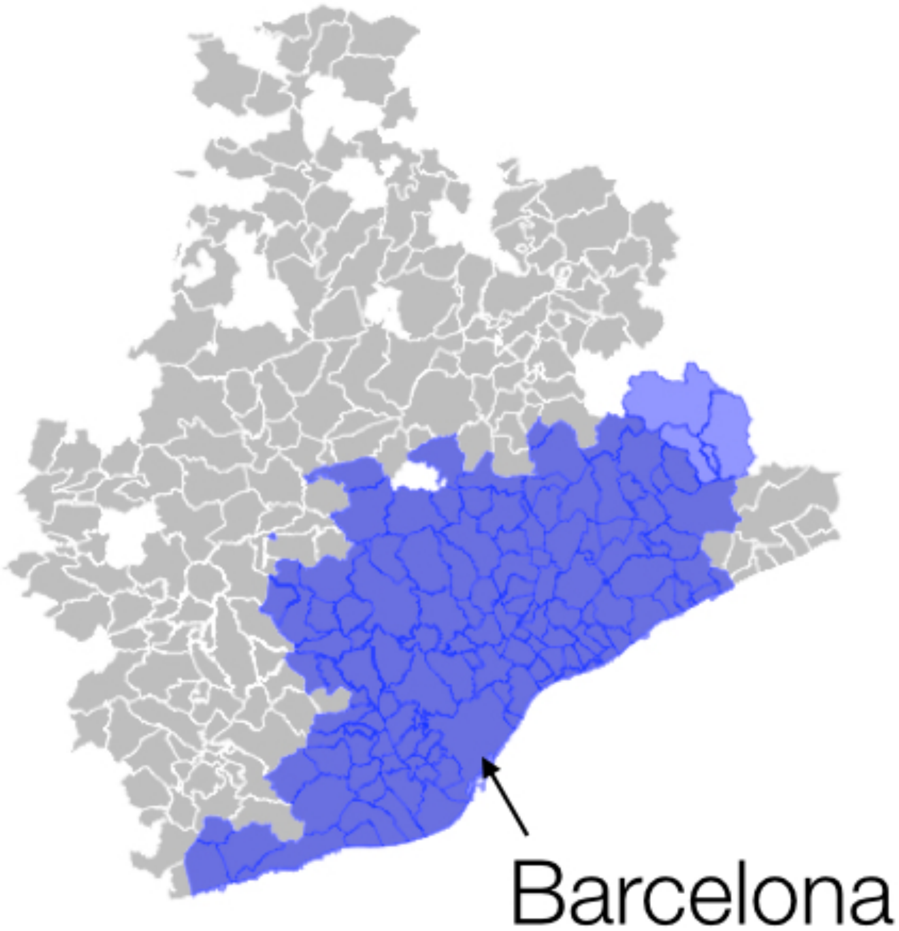
# (Functional) geographical areas in Spain

- We look for communities in the flow graph
- Infomap algorithm
- 340 functional areas were detected:
  - They are cohesive
  - Statistically robust
  - Modularity is high
  - High overlap with "comarcas"

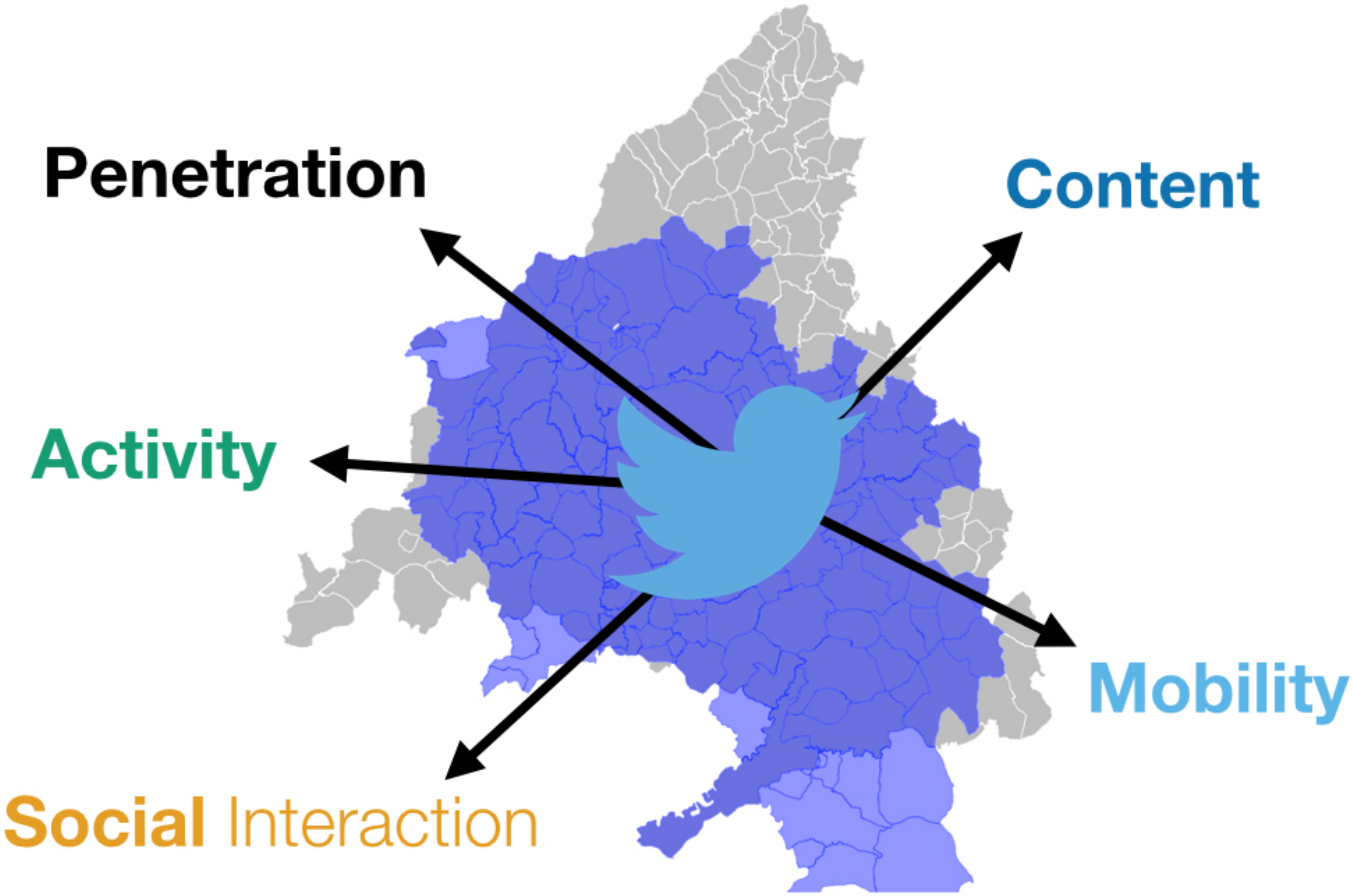


# (Functional) geographic areas in Spain



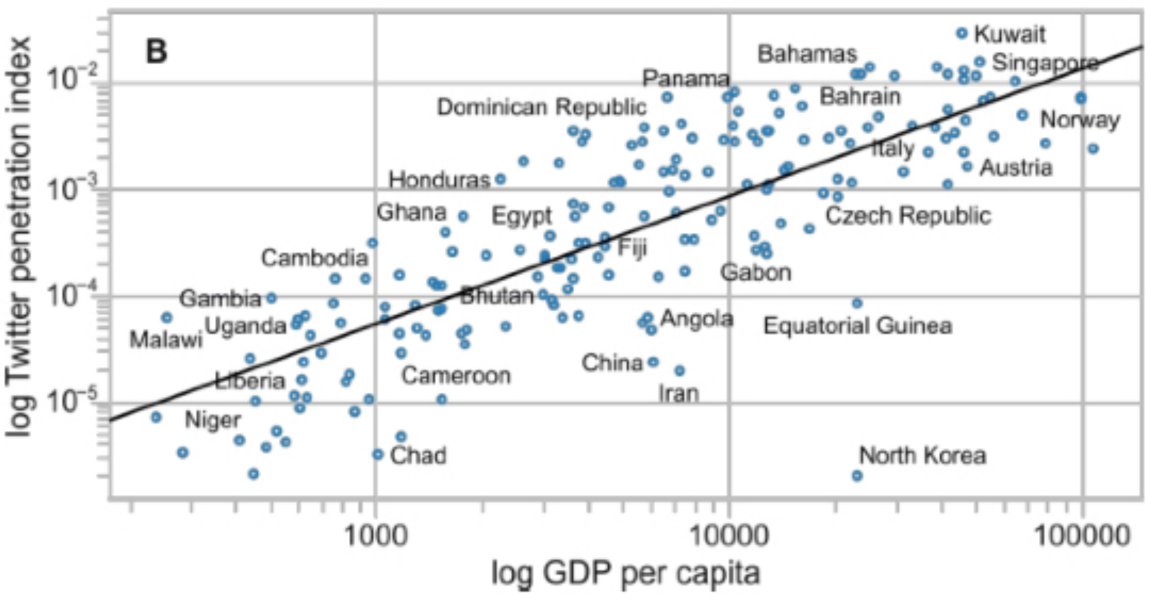




"The piece is absolutely useless, even ridiculous, outside Spain, because the audience cannot hope to understand its significance, nor the performers to play it as it should be played."

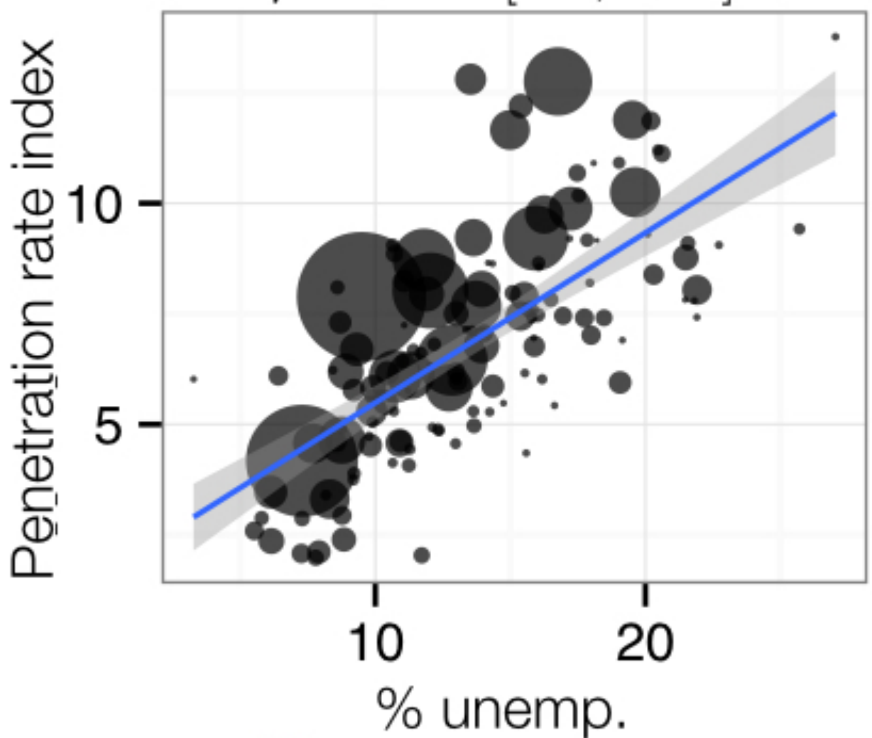

Penetration
Content
Activity
Mobility
Social Interaction

# Twitter **penetration**

- Is Twitter penetration related to economical development of areas?
  - At country scale twitter penetration ~ GDP
    Hawelka, B. et al., 2013. Geo-located Twitter as the proxy for global mobility patterns.
  - At small scale is the opposite! twitter penetration ~ unemployment




$\rho = 0.70\ [0.6, 0.77]$




Universidad Carlos III de Madrid @estebanmoro

# Twitter **social interactions**

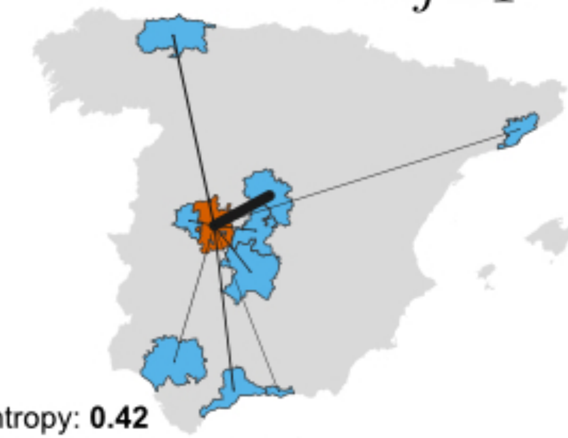
- Granovetter: diversity of interactions yields to more opportunities
- Diversity of interactions between cities is correlated with economical development
  Eagle et al, Science 2010
- We construct the graph of social interactions

$$w_{ij} = \text{number of @ between areas } i \text{ and } j$$

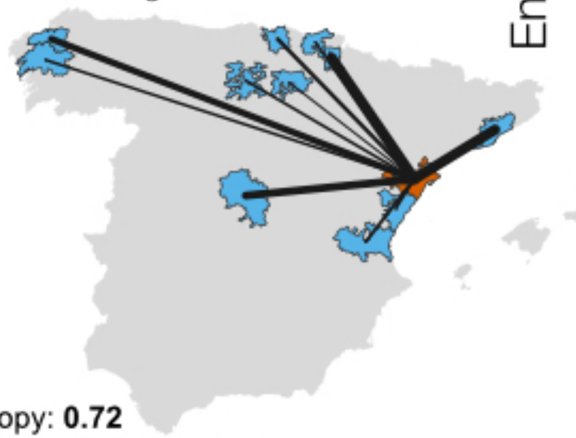
$$p_{ij} = w_{ij} / \sum_{j=1}^{k_i} w_{ij}$$

- Measure diversity with **entropy**

$$S_i = - \sum_{j=1}^{k_i} p_{ij} \log p_{ij}$$

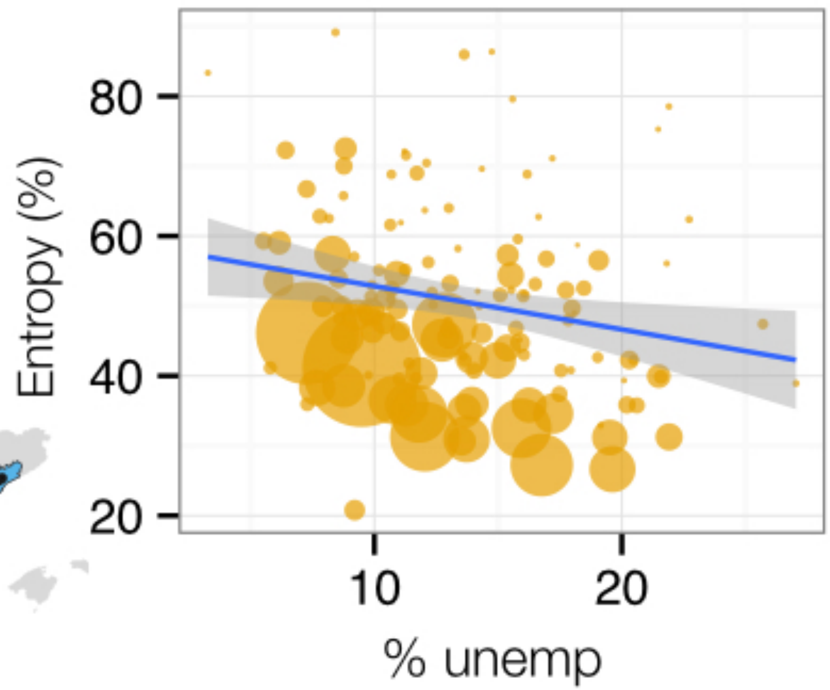


Entropy: **0.42**
Unemployment: **20.3%**



Entropy: **0.72**
Unemployment: **8.8%**

$$\rho = -0.21[-0.37, -0.04]$$



# Twitter **geographical interactions**

- Diversity of geographical mobility is correlated with development

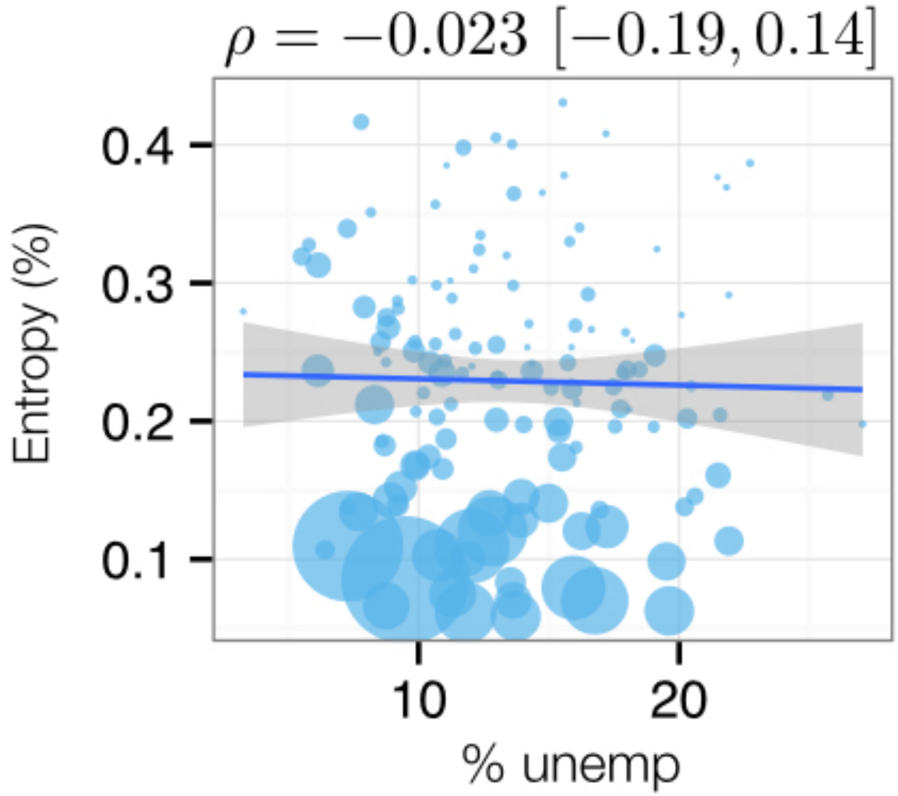
  Smith, C., Mashhadi, A. & Capra, L., 2013. Ubiquitous sensing for mapping poverty in developing countries.
  Smith, C., Quercia, D. & Capra, L., 2013. Finger on the pulse: identifying deprivation using transit flow analysis.

- We use the graph of flows

$$\tilde{p}_{ij} = T_{ij} / \sum_{j=1}^{\tilde{k}_i} T_{ij}$$

- Measure diversity with **entropy**

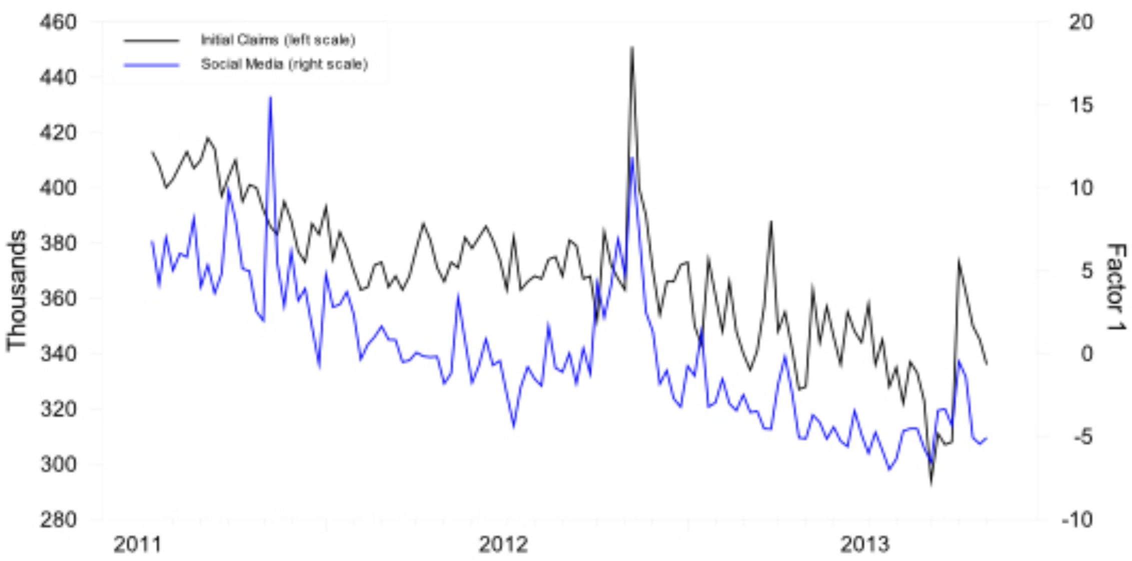
$$\tilde{S}_i = - \sum_{j=1}^{\tilde{k}_i} \tilde{p}_{ij} \log \tilde{p}_{ij}$$

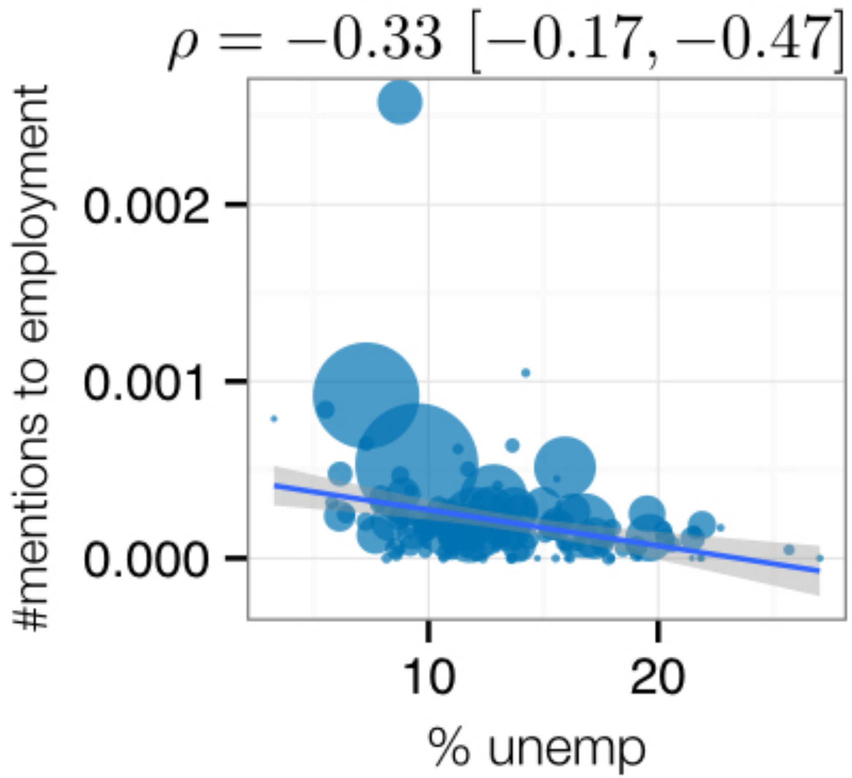


# Twitter **content**

- Two different approaches
    - Classical approach: NLP applied to detect mentions to “unemployment”, “job”, “economy”, …

Antenucci, D. et al., 2014. Using Social Media to Measure Labor Market Flows.

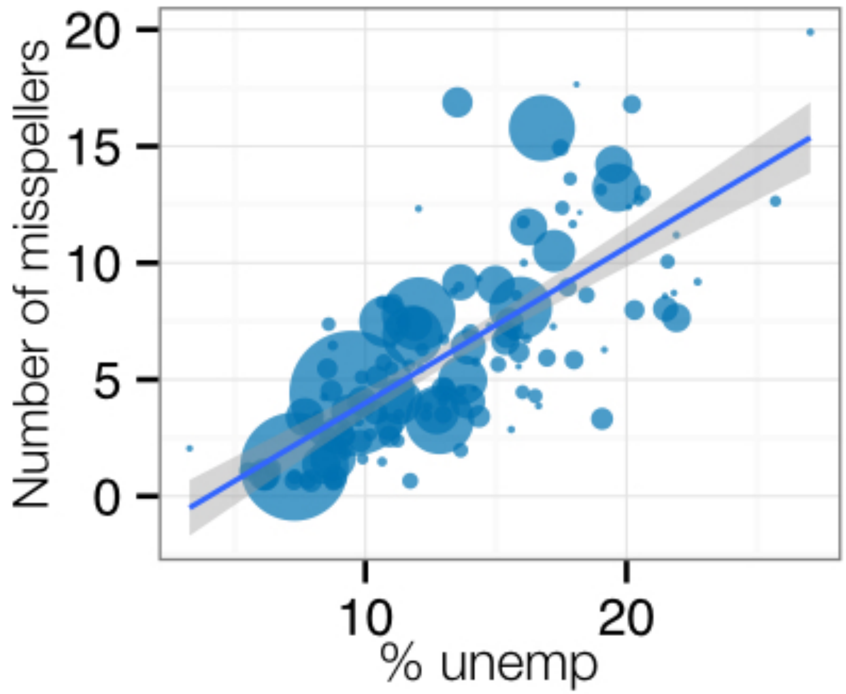


# Twitter **content**

- Our approach: NLP applied to detect **lexical complexity** (as a proxy for educational level)
  - Readability (Gunning index) *J. Davenport et al, The Readability of Tweets and their Geographic Correlation with Education, arXiv:1401.6058, 2014*
  - **Serious misspellings**

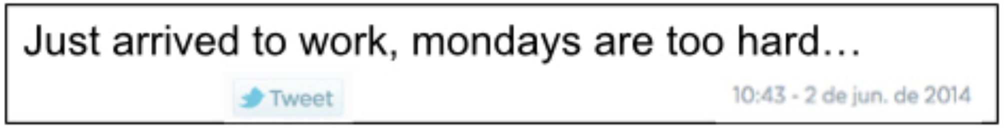
| Tweet | Correct spelling |
|---|---|
| Alguien se viene **con migo aver** la vida de PI?? | Alguien se viene **conmigo a ver** la vida de PI?? |
| La quiero mucho y la **hecho de menos** | La quiero mucho y la **echo de menos** |
| Yo **llendo** a trabajar con este tiempo | Yo **yendo** a trabajar con este tiempo |

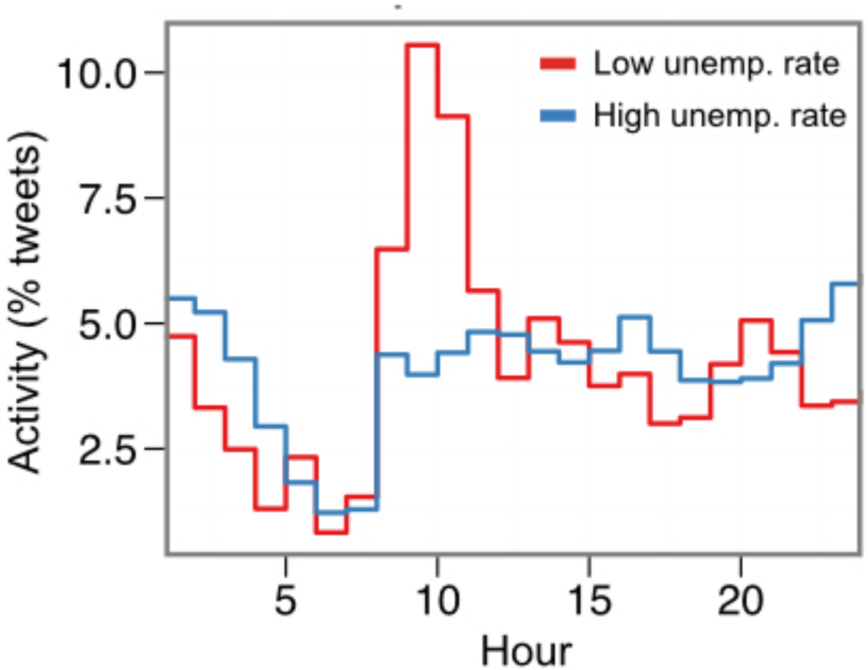



- We construct a list of more than 600 incorrect expressions of this type validated by spanish language linguistic experts.
- We do not take into account misspellings due to different Spanish accents and IM abbreviations
- We compute for each area the fraction of users that make a number of serious misspellings

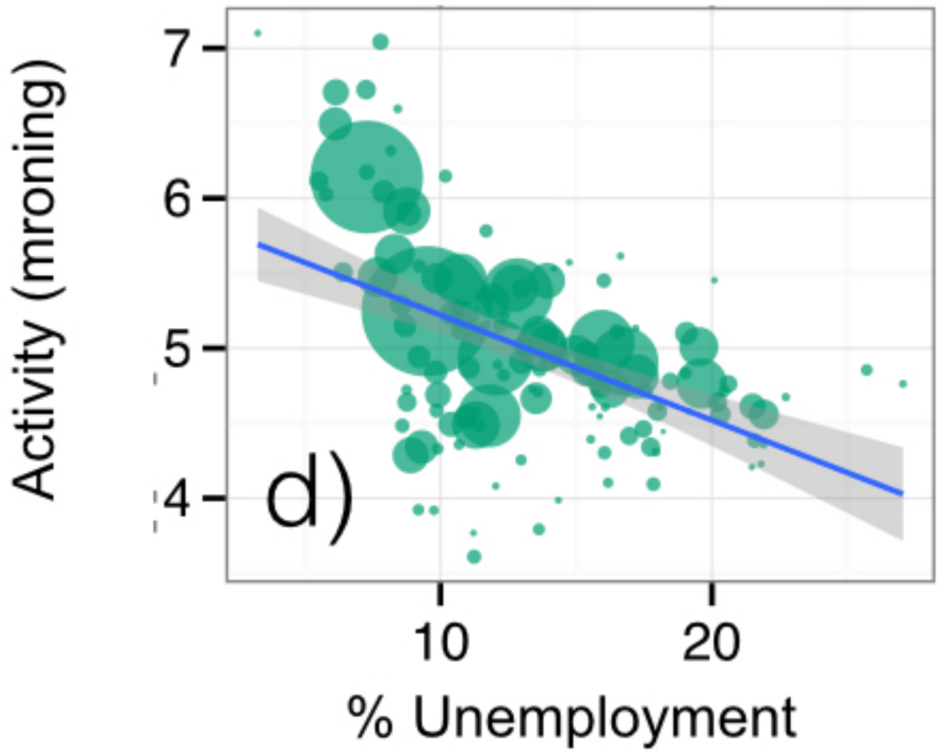
# Twitter **activity**

- Is unemployment reflected in twitter daily patterns?



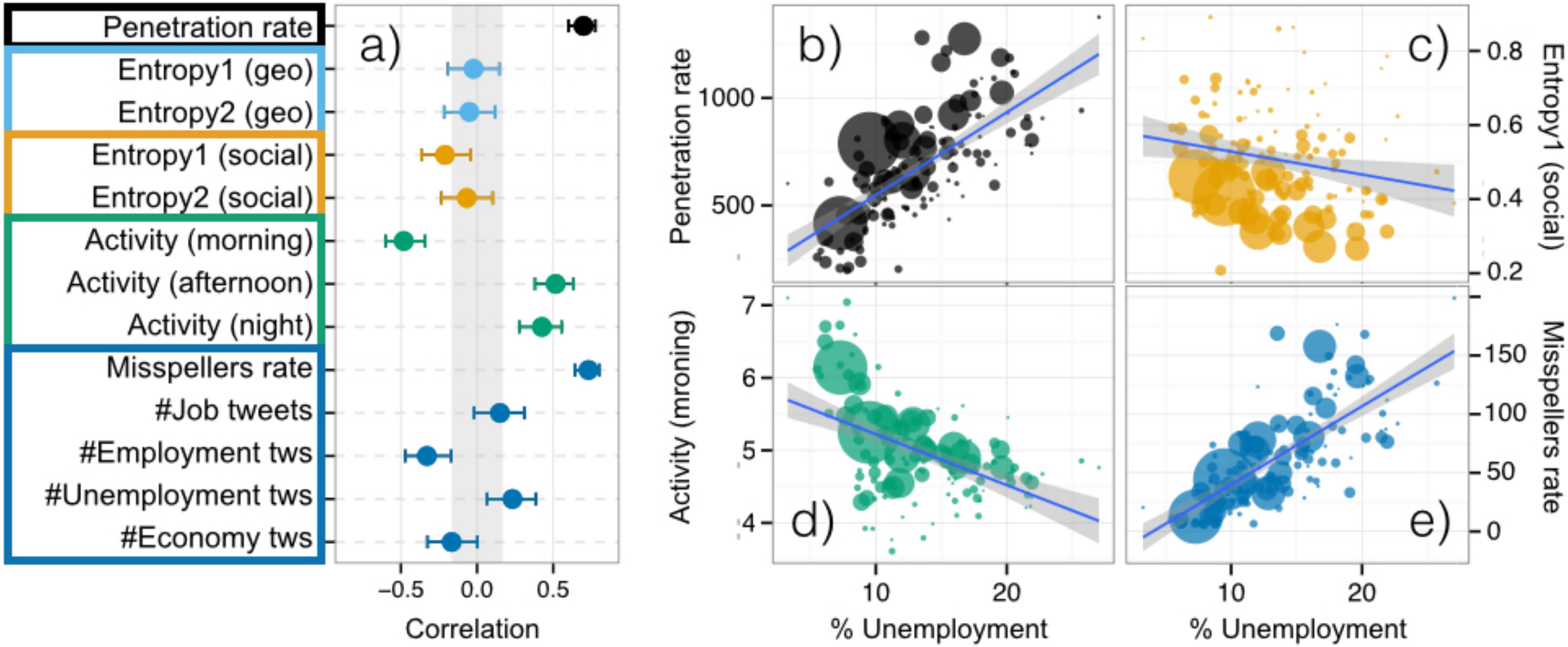

$\rho = -0.48\ [-0.34, -0.60]$



# Summary of the variables

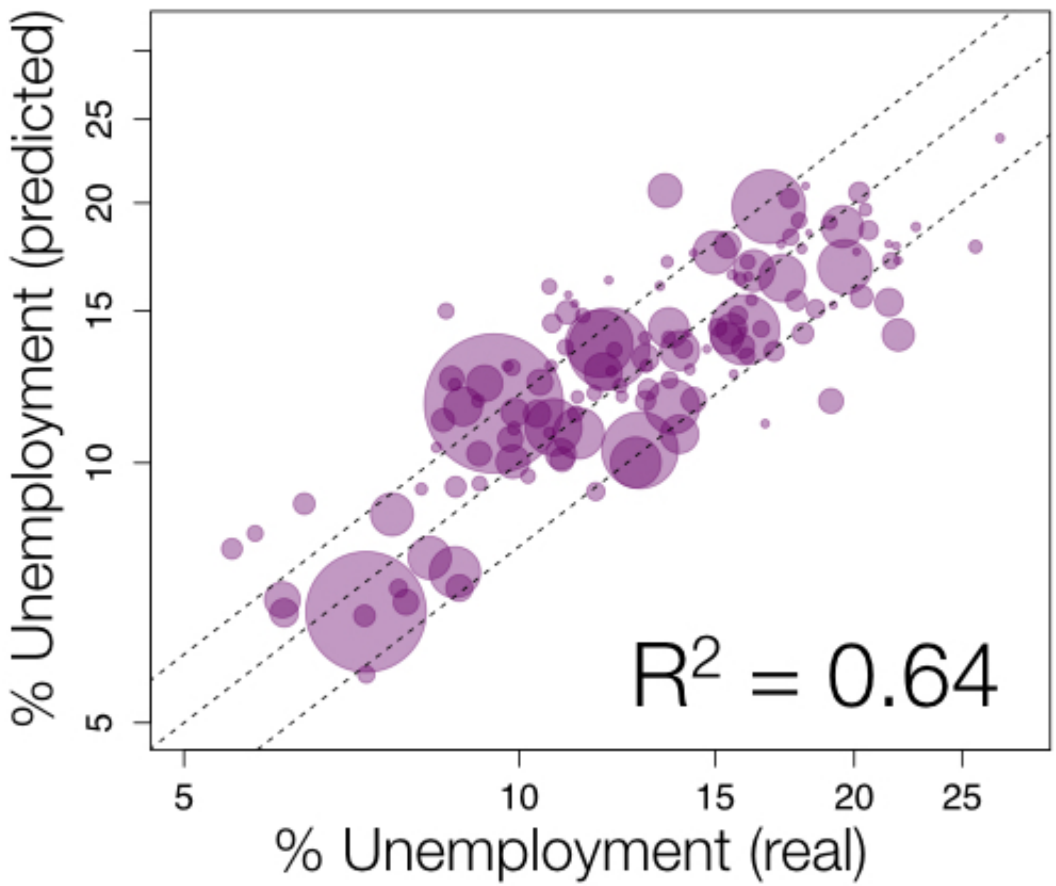
Social/geo variables have low correlation

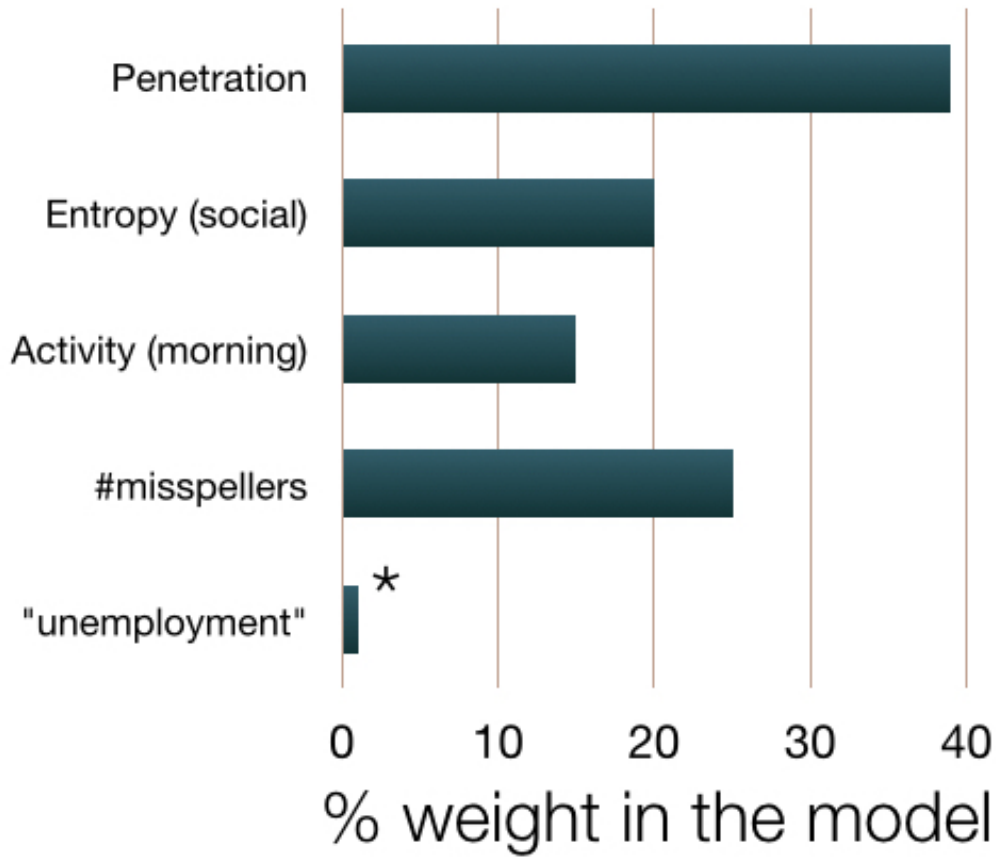
Penetration rate/activity and content are highly correlated with unemployment



# Explanatory power of Twitter variables

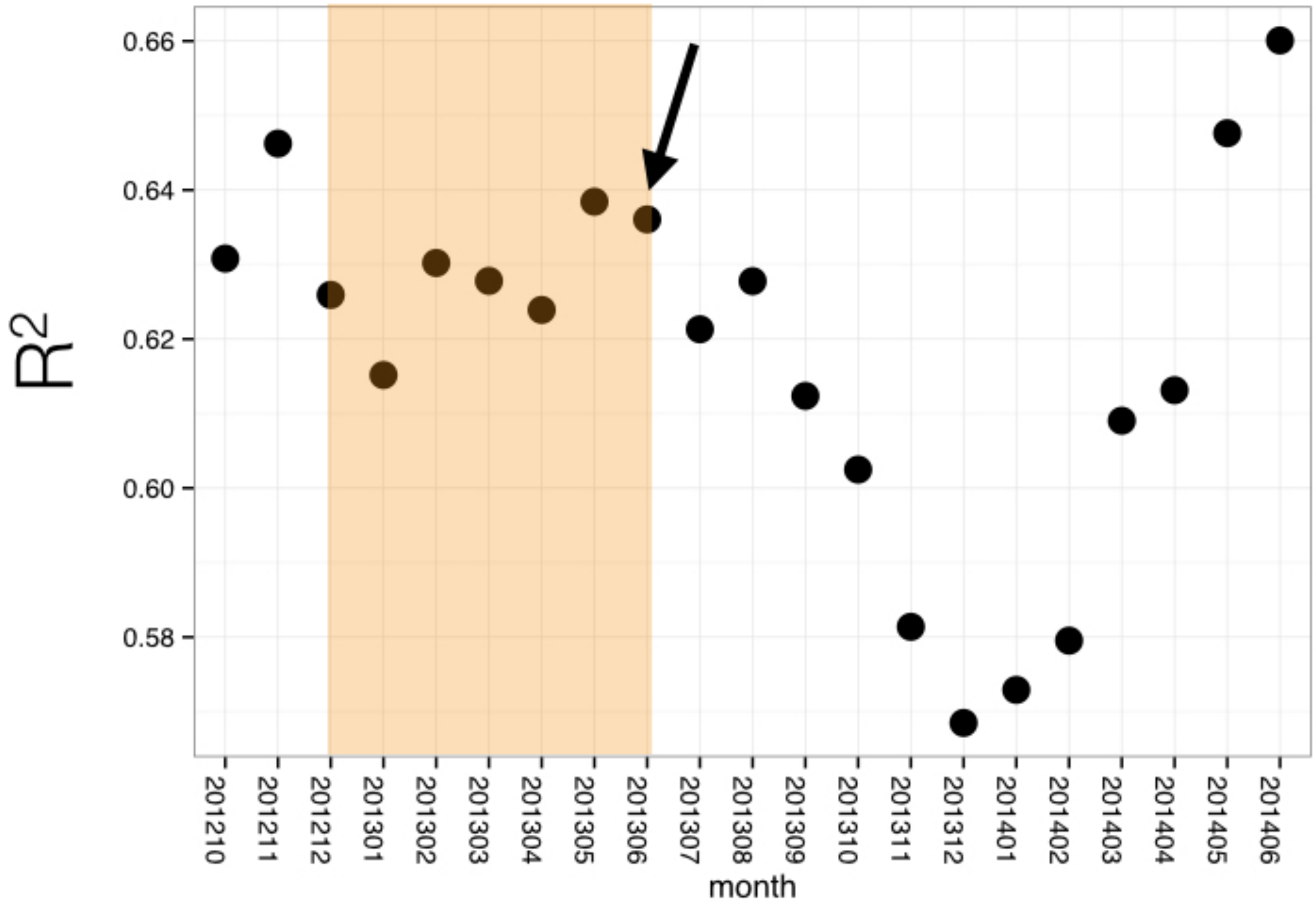
- Simple linear regression



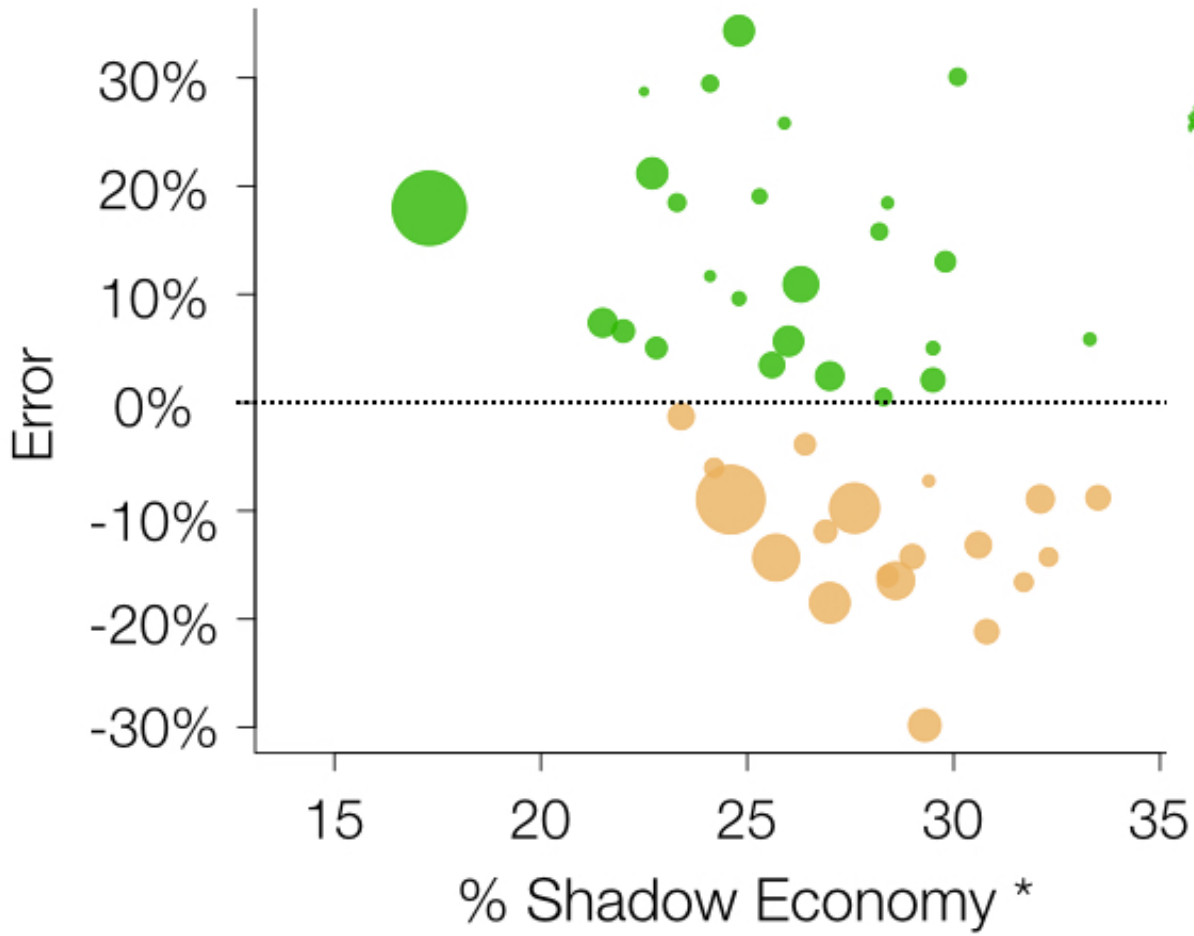
# Explanatory power as a function of time

- At what time of the year the model has more explanatory power?

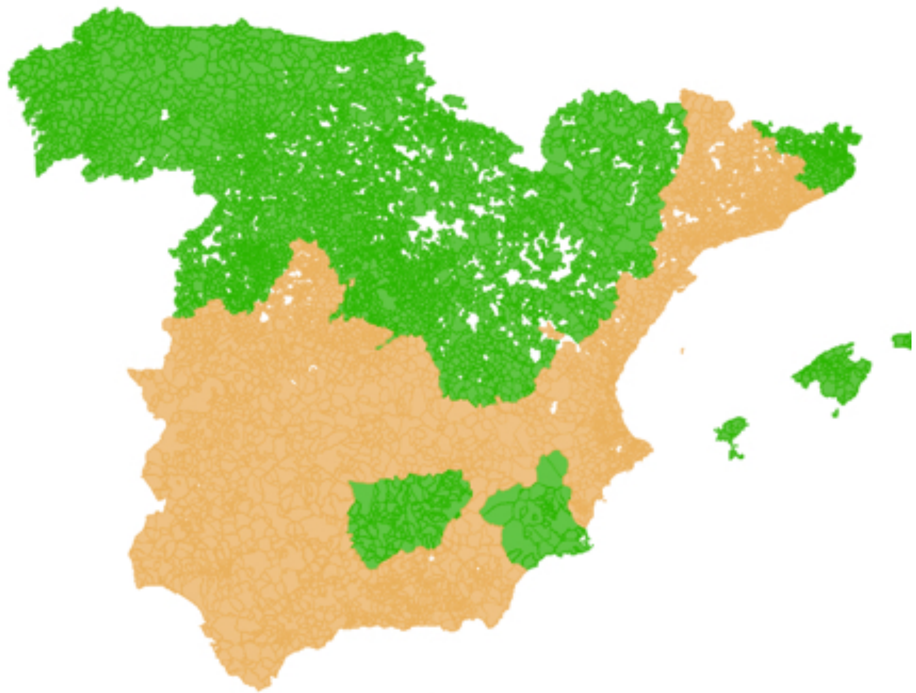


# Are we really wrong?

Model Error = Model[variables] - Official unemployment




(* GESTHA report 2012)



# Summary

# Summary

- **Economical development -> Behavior -> Social Media**

# Summary

- **Economical development -> Behavior -> Social Media**
- Can Twitter be used to infer economical development? YES
  - Areas with different behavior in Twitter show different levels of unemployment
  - Applications to marketing, media, planning
  - Twitter is everywhere! and the API is free! Prove us wrong, please!

# Summary

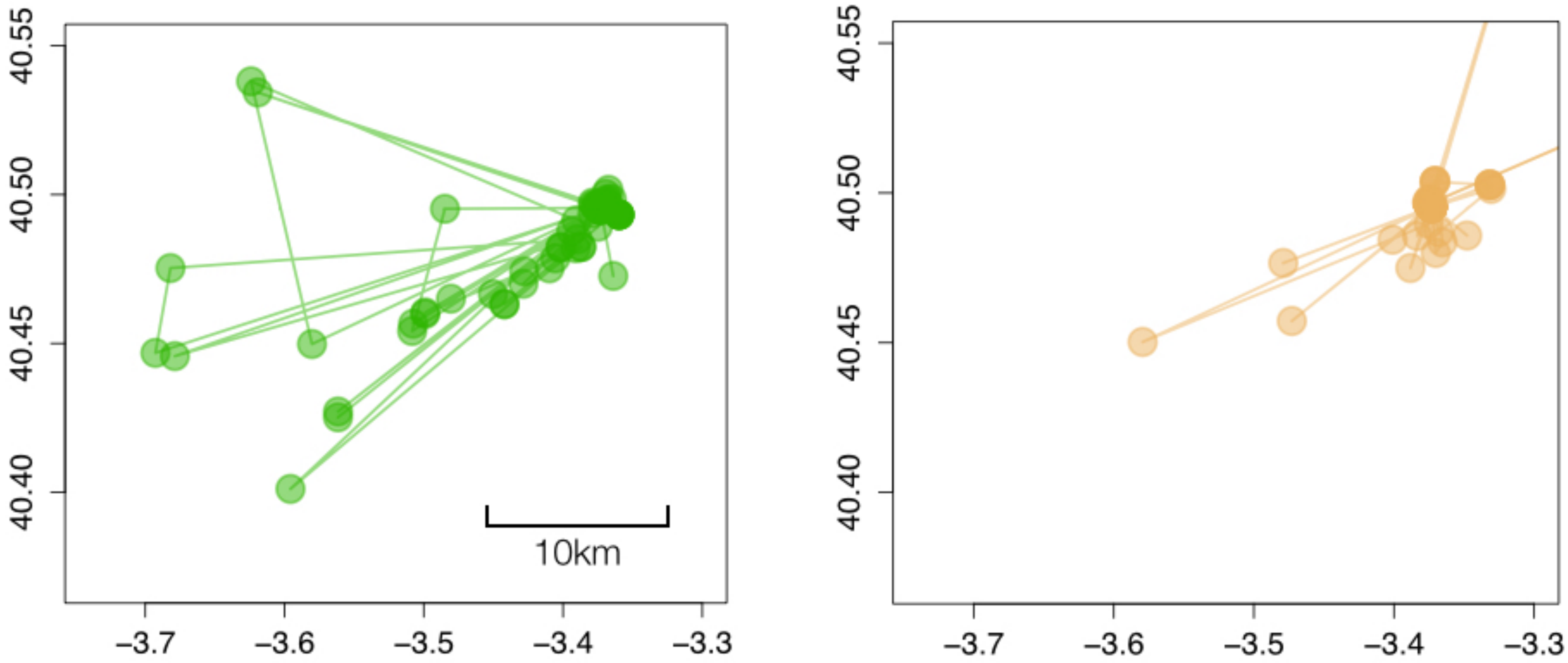
- **Economical development -> Behavior -> Social Media**
- Can Twitter be used to infer economical development? YES
  - Areas with different behavior in Twitter show different levels of unemployment
  - Applications to marketing, media, planning
  - Twitter is everywhere! and the API is free! Prove us wrong, please!
- Can Twitter variables explain the unemployment per area? YES with $R^2 = 0.64$
  - Activity, penetration and content account for 80% of the variance explained
  - Diversity on geographical and social interaction amount only for the 20%

# Summary

- **Economical development -> Behavior -> Social Media**
- Can Twitter be used to infer economical development? YES
    - Areas with different behavior in Twitter show different levels of unemployment
    - Applications to marketing, media, planning
    - Twitter is everywhere! and the API is free! Prove us wrong, please!
- Can Twitter variables explain the unemployment per area? YES with $R^2 = 0.64$
    - Activity, penetration and content account for 80% of the variance explained
    - Diversity on geographical and social interaction amount only for the 20%
- **What we are doing now:**
    - Can we use the model to forecast future unemployment? Or now-casting?
    - What are the behavioral changes behind socio-economical changes?
    - Use the model in under-developed countries.

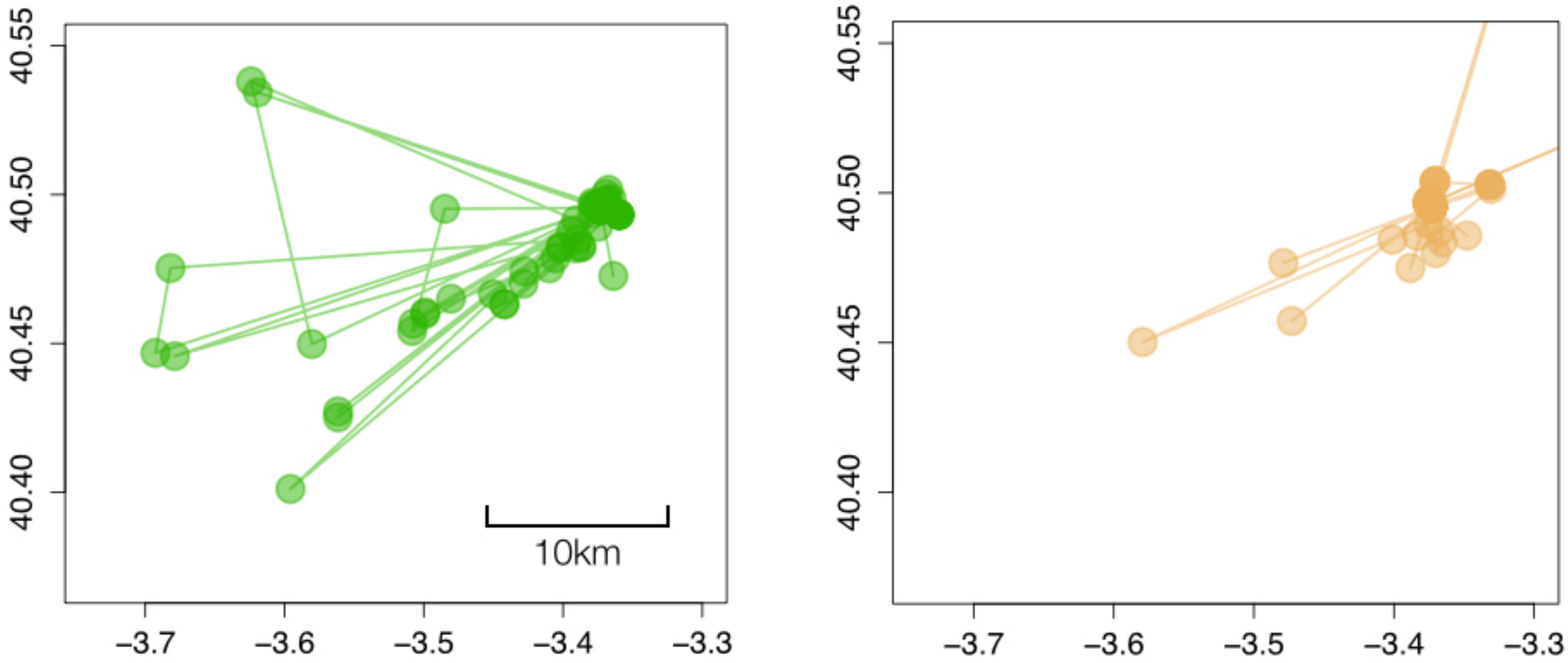
# Behavioral changes behind socio-economical changes

(Geolocalized tweets)



# Behavioral changes behind socio-economical changes

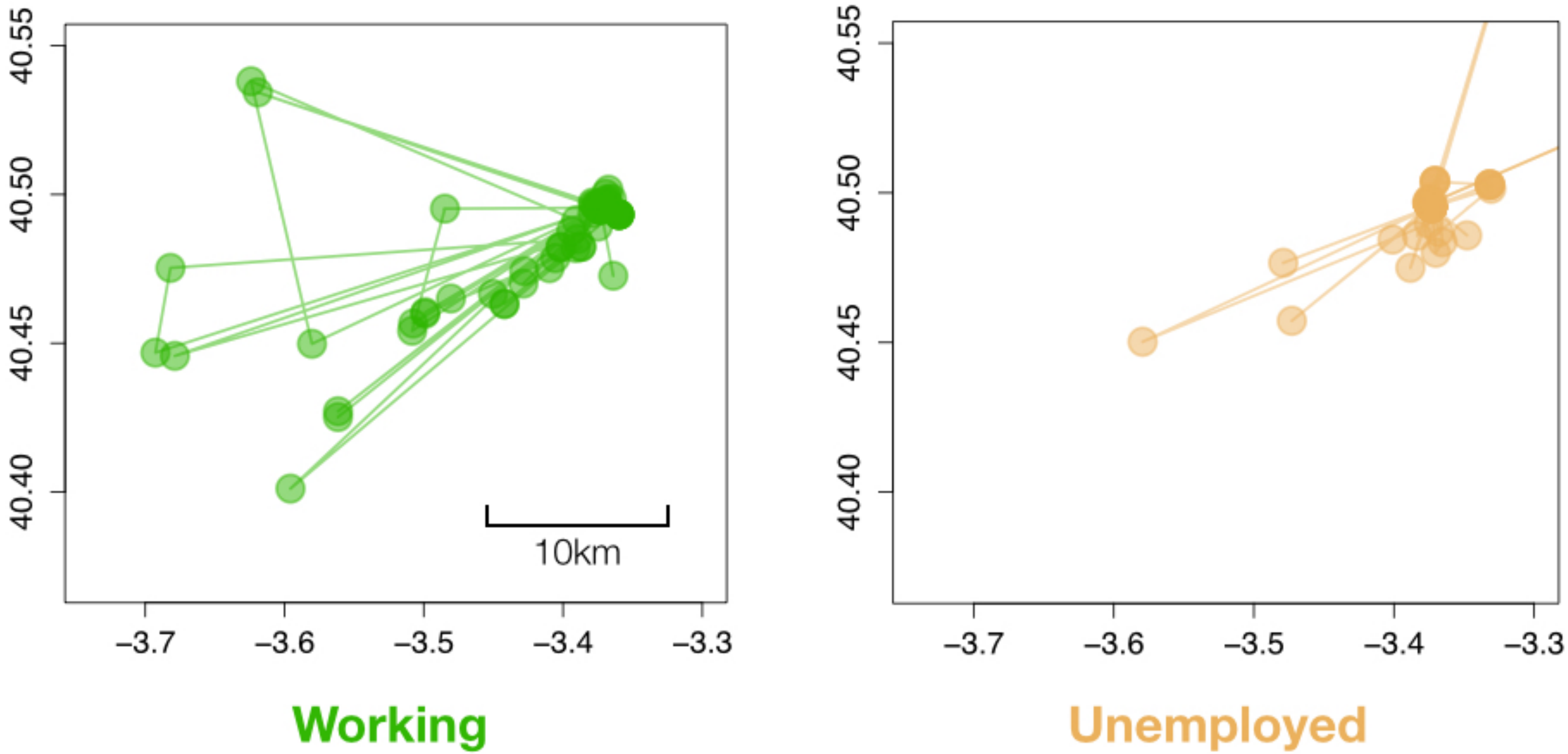
(Geolocalized tweets)




**Working**

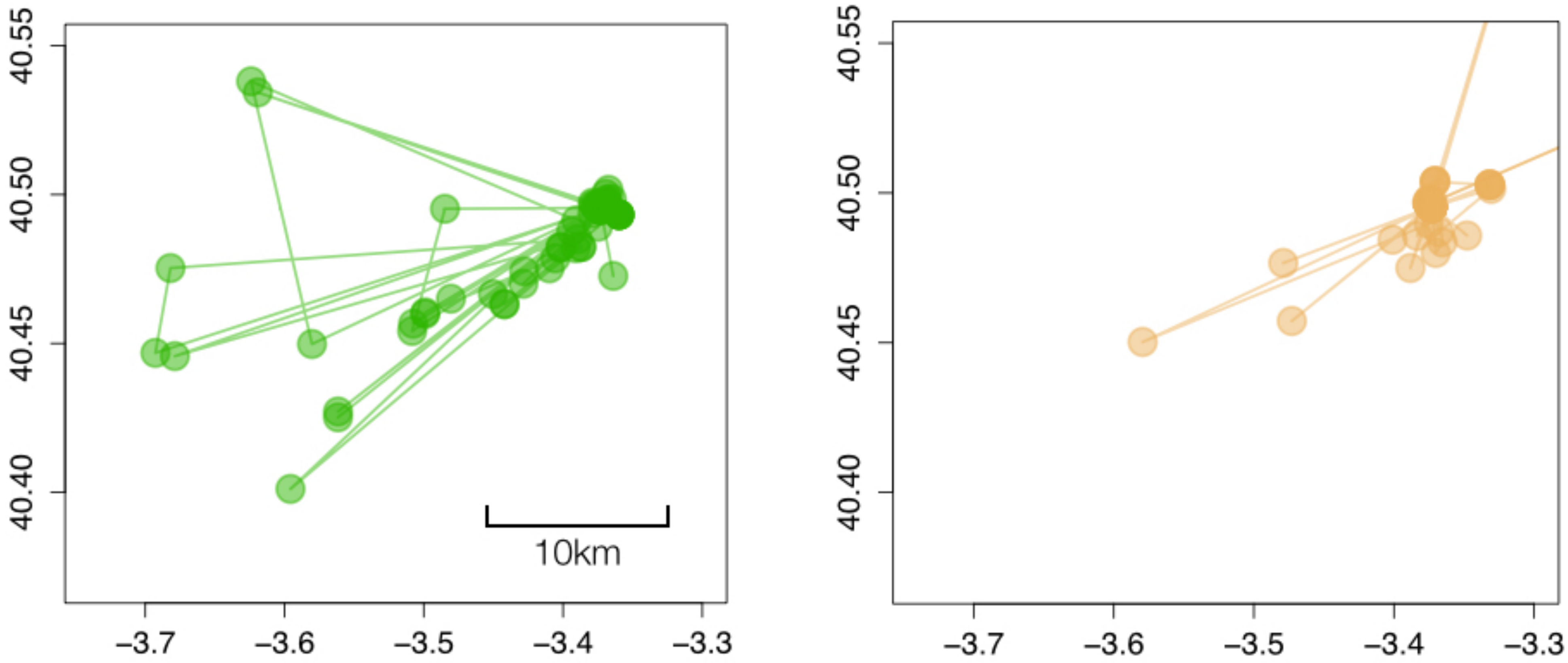
# Behavioral changes behind socio-economical changes

(Geolocalized tweets)



# Behavioral changes behind socio-economical changes

(Geolocalized tweets)




**Working** **Unemployed**

Less geographical mobility, more probability to be unemployed