BIG DATA
and
Human Mobility

Maxi San Miguel

IFISC

Universitat de les Illes Balears

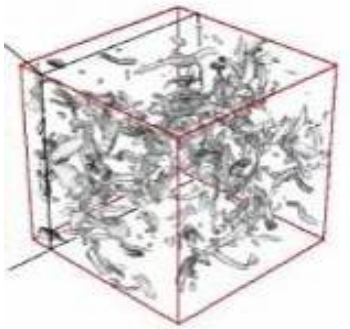
CSIC

Fourth Paradgim, J. Gray, 2007

Thousand years ago:
science was **empirical**
describing natural phenomena

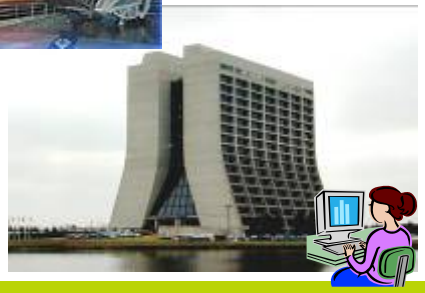
Last few hundred years:
theoretical branch
using models, generalizations

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - \mathrm{K}\frac{c^2}{a^2}$$



Last few decades:
a **computational** branch
simulating complex phenomena

Today:
data exploration (eScience)
unify theory, experiment, and simulation
Information/Knowledge stored in computer
Scientist analyzes database / files
using data management and statistics



BIG DATA

Twitter

Cell Phone

Electronic Transactions

TWITTER DATA

31 days

87,569 users

581,749 messages

"INDIGNADOS"

IFISC

Twitter Data

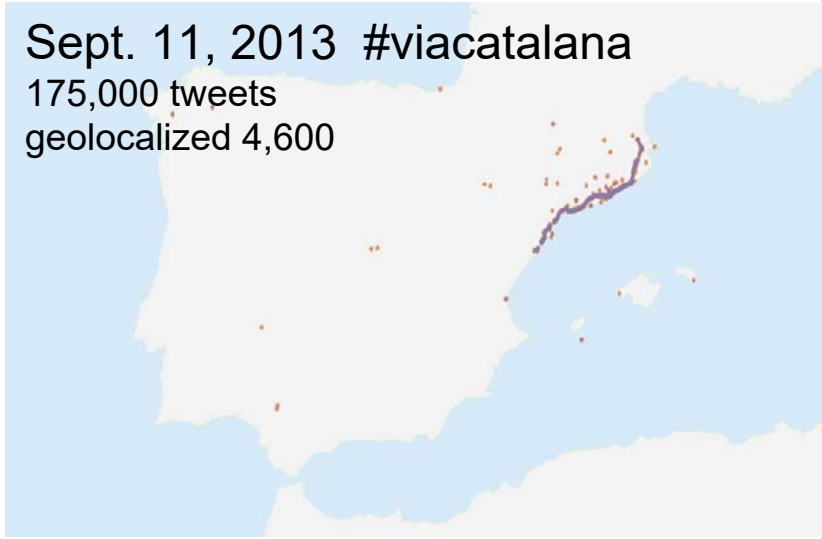
Geolocalized twitter

Mobility

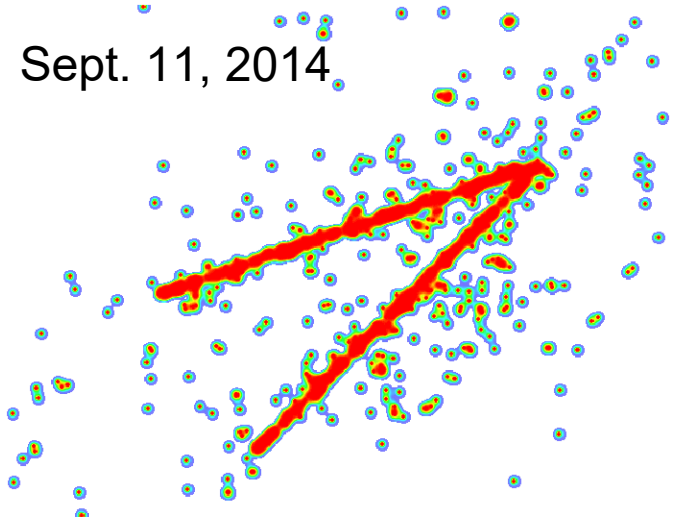
Sept. 11, 2013 #viacatalana

175,000 tweets

geolocalized 4,600



Sept. 11, 2014



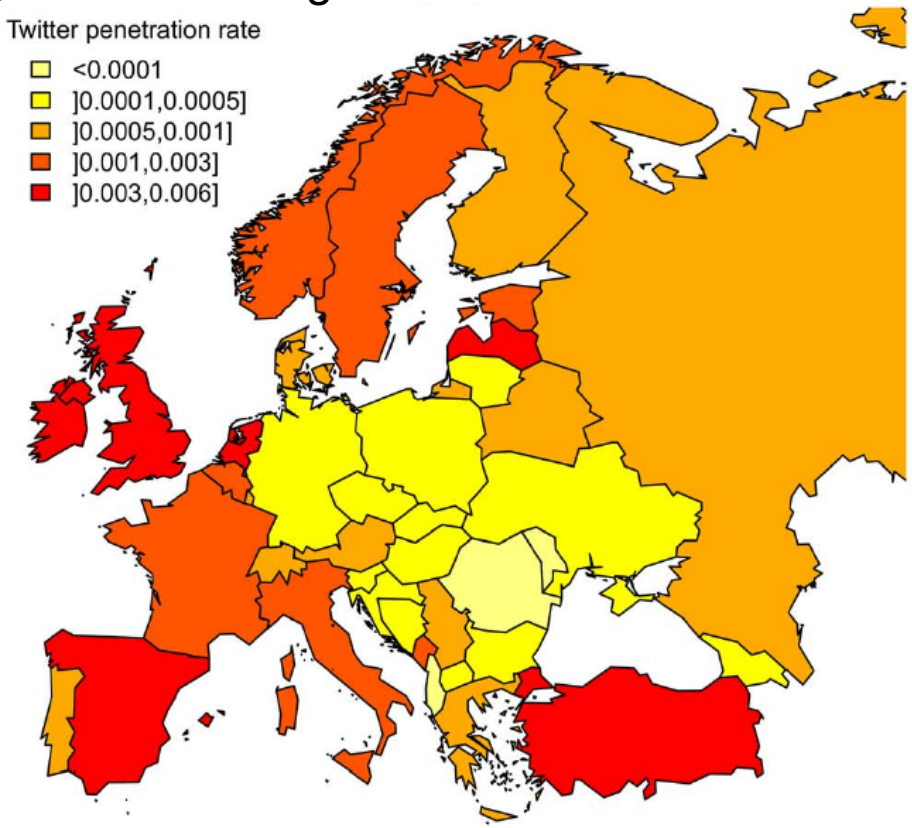
Data Mining: 15-25 Mill. Tweets/day

http://ifisc.uib-csic.es/humanmobility/

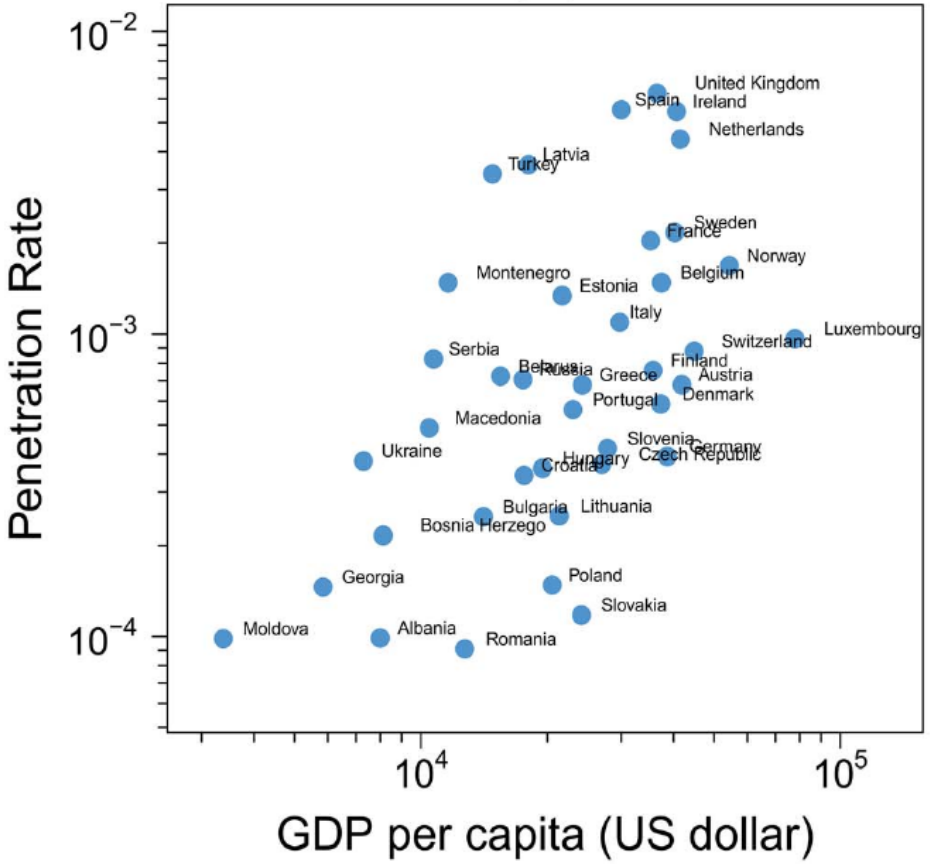
Geolocated twitter penetration rate

M. Lenormand, et al. PLOS ONE 9, e105407 (2014)

geolocated tweets/inhabitants



5.219.539 geo-located tweets
Sept 2012 to November 2103.
1.477.263 users.



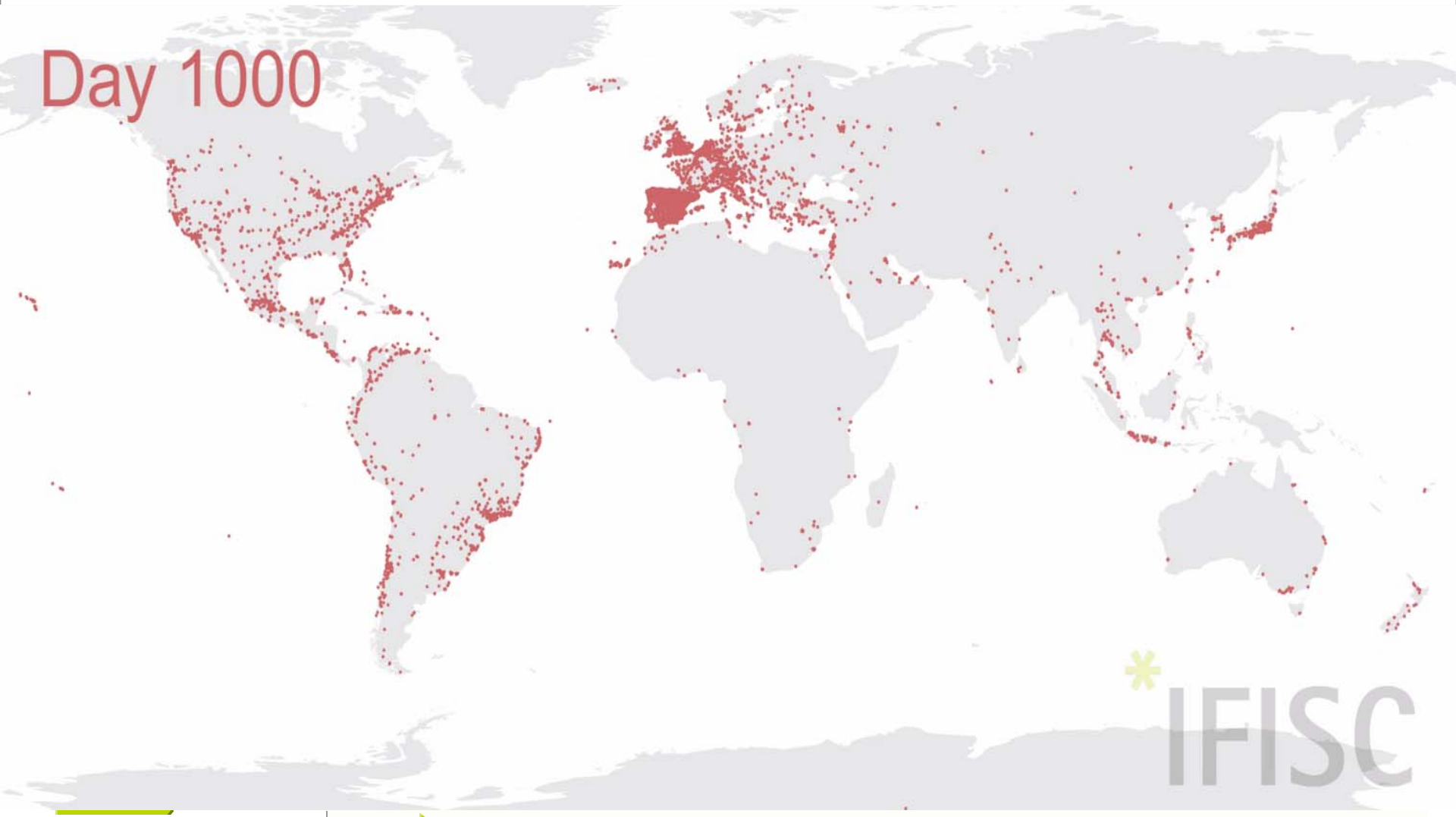
Lower penetration rate in central Europe.

No correlation with GDP per capita.

New type of data on cultural differences

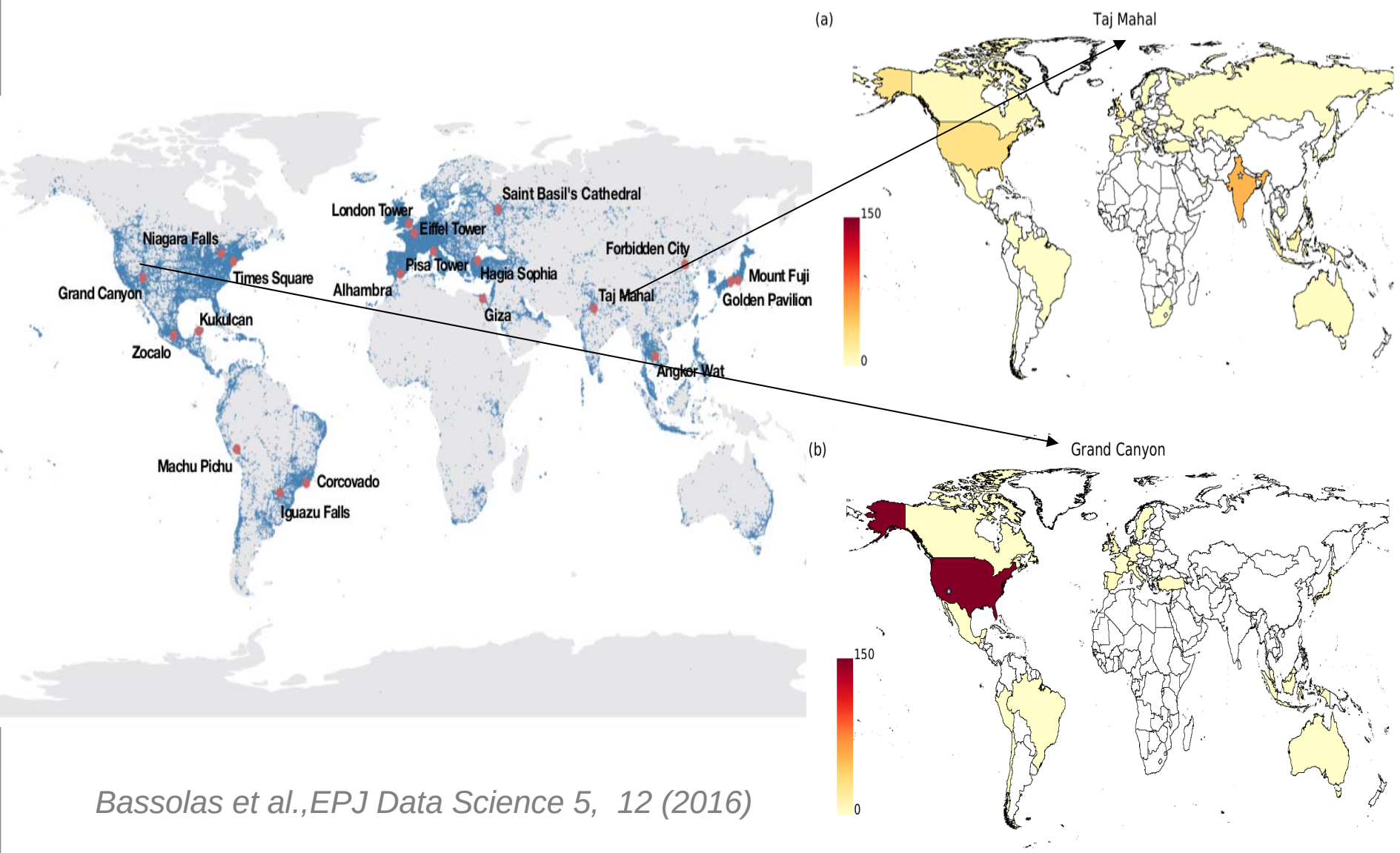
Twitter, Mobility and City Influence

M. Lenormand et l. J. Royal Society Interface 12, 20150473 (2015)



Day 1 Madrid

Touristic site attractiveness



Bassolas et al.,EPJ Data Science 5, 12 (2016)

Tweets on the road

M. Lenormand, et al. PLOS ONE 9, e105407 (2014)

5.219.539 geo-located tweets
Sept 2012 to November 2103.
1.477.263 users.

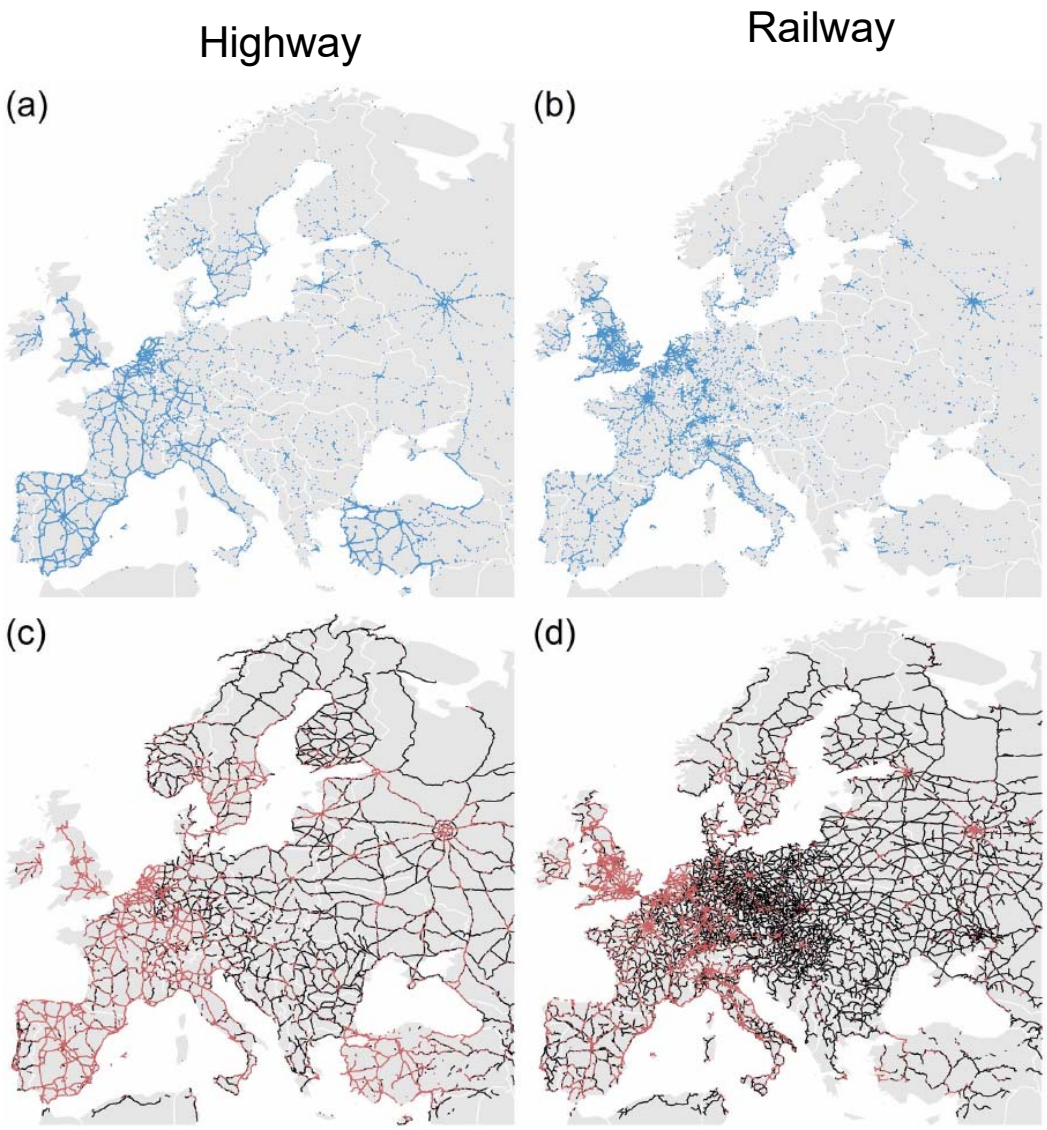
Highway and railway coverage

Tweets geolocated less than 20 from highway/railway

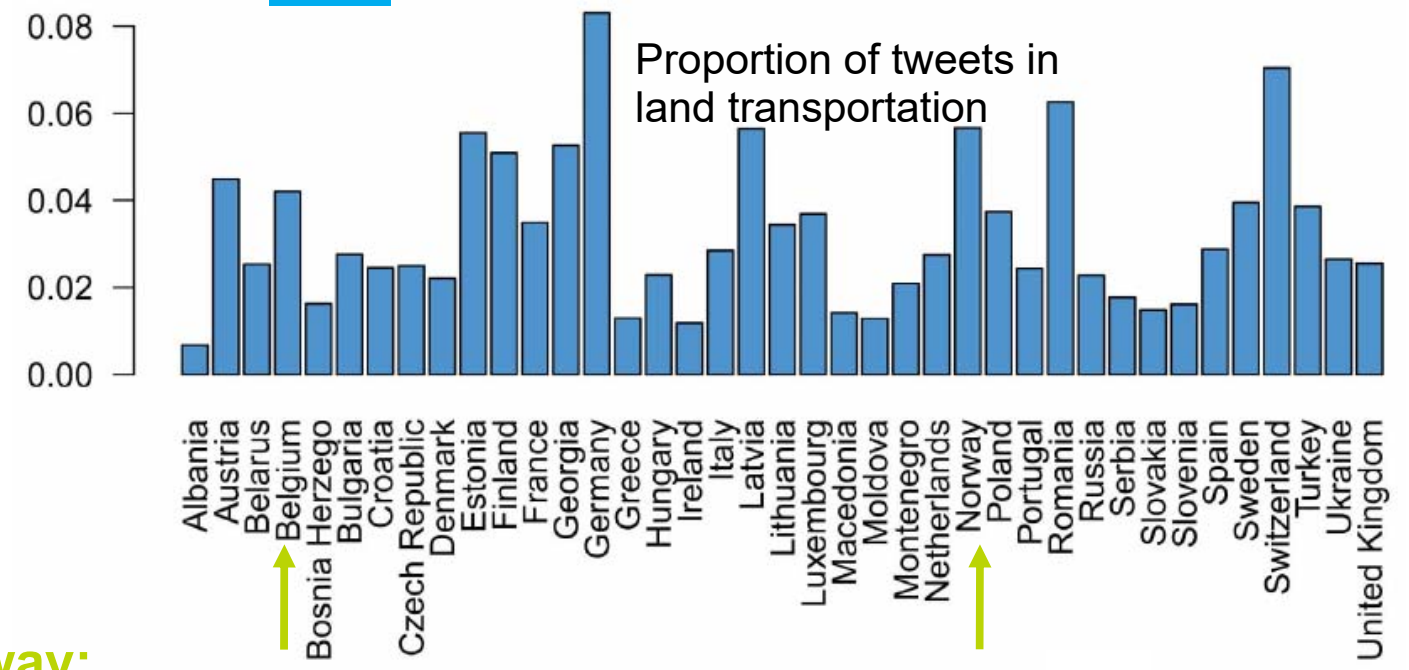
Highway/railway networks divided in segments of 10 Km.

Segments covered
Segments uncovered

Strong coverage differences between countries

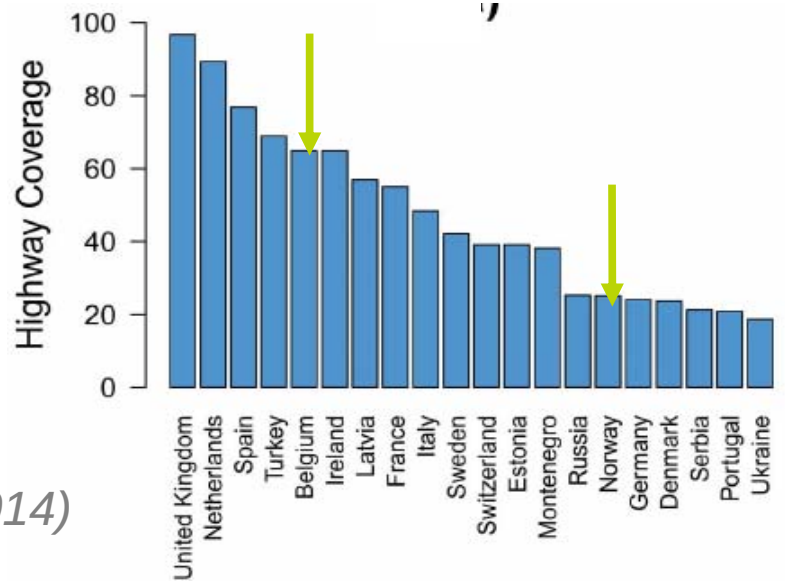


Tweets on the road: sociocultural differences



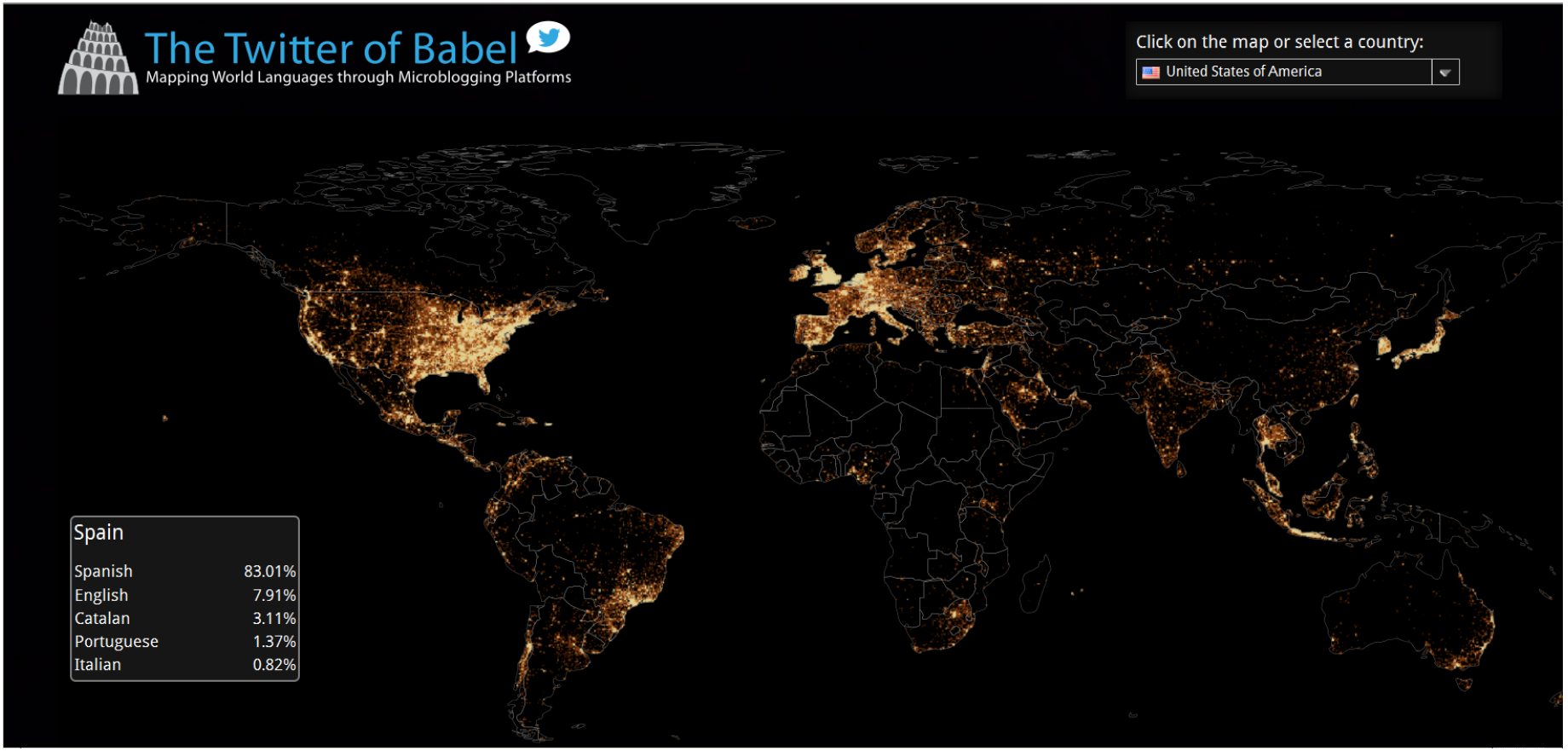
Belgium vs Norway:

- Similar penetration rate.
- **Norway**: larger proportion in land transportation
- **Belgium:** highway coverage three times larger



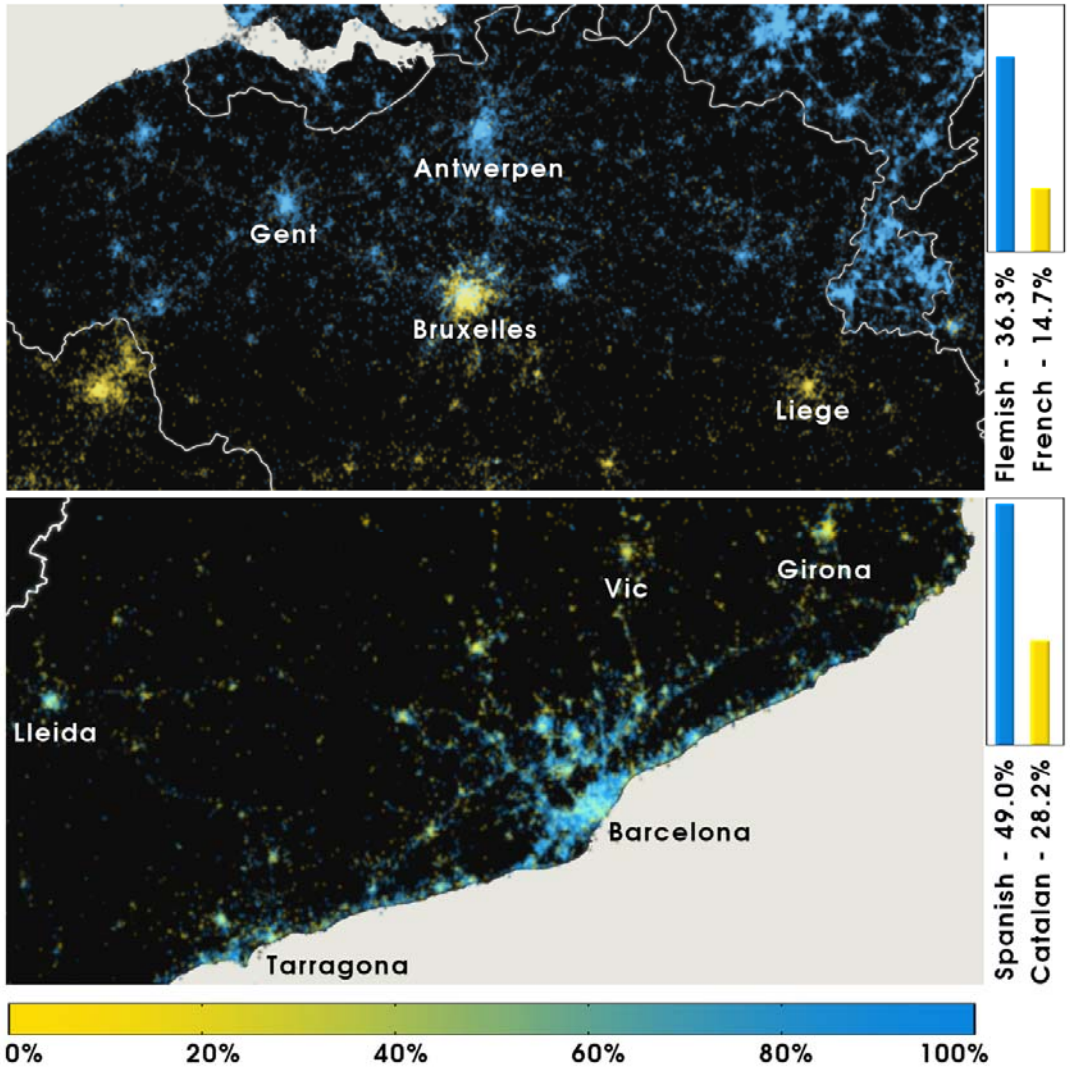
M. Lenormand, et al. PLOS ONE 9, e105407 (2014)

Twitter of Babel



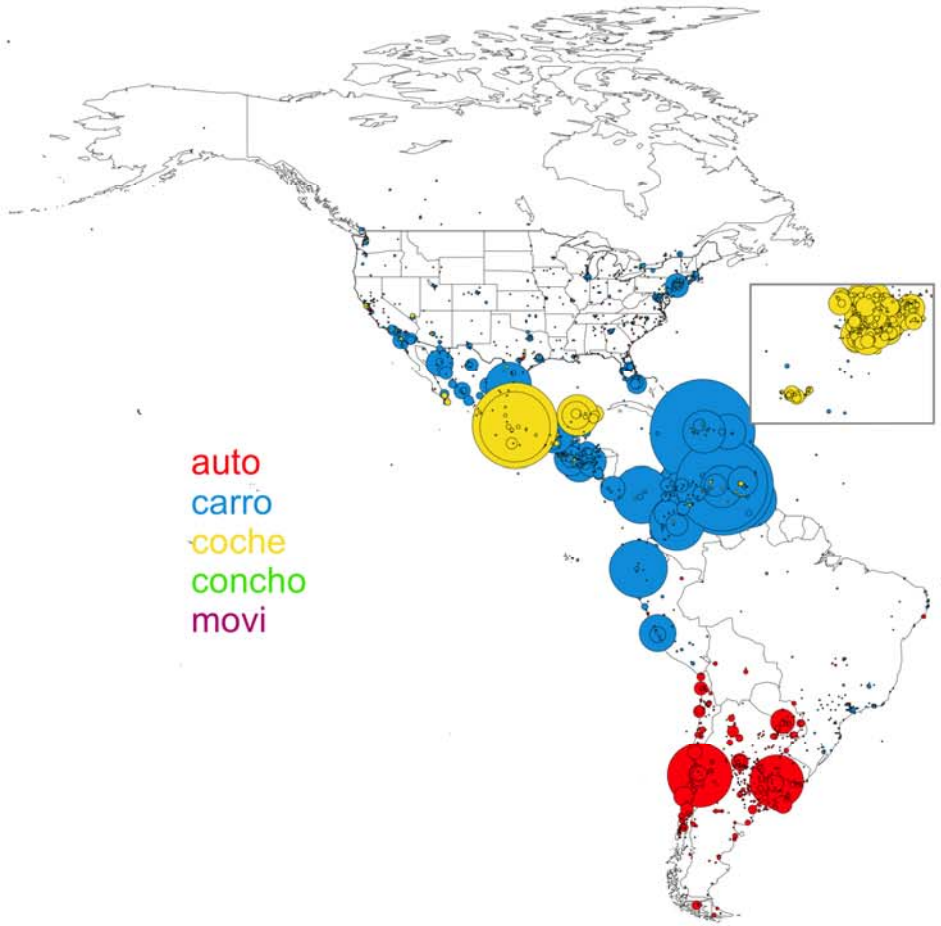
http://www.ccs.neu.edu/home/qianz/MapTwitterLanguage/v1/index.html

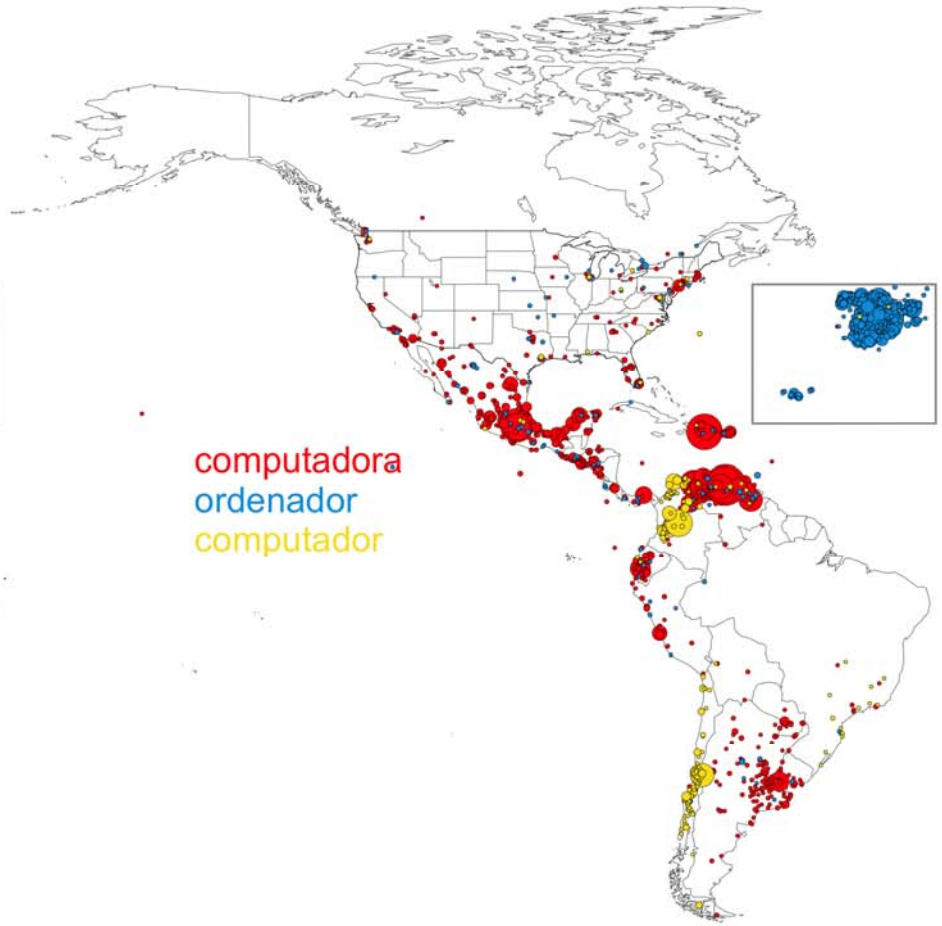
Twitter of Babel



D. Mocanau et al. PLOS ONE (2013)

Spanish Dialects





D. Sánchez, PLoS ONE 9, e112074, (2014)

Ghettos from language use in twitter

ArXiv 1611.01056

354 Million geolocalized tweets, 1+ Milion twitter users, 53 world cities

Immigrant community integration in world cities

Fabio Lamanna,[1,*] Maxime Lenormand,[2] María Henar Salas-Olmedo,[3]
Gustavo Romanillos,[3] Bruno Gonçalves,[4] and José J. Ramasco[1]

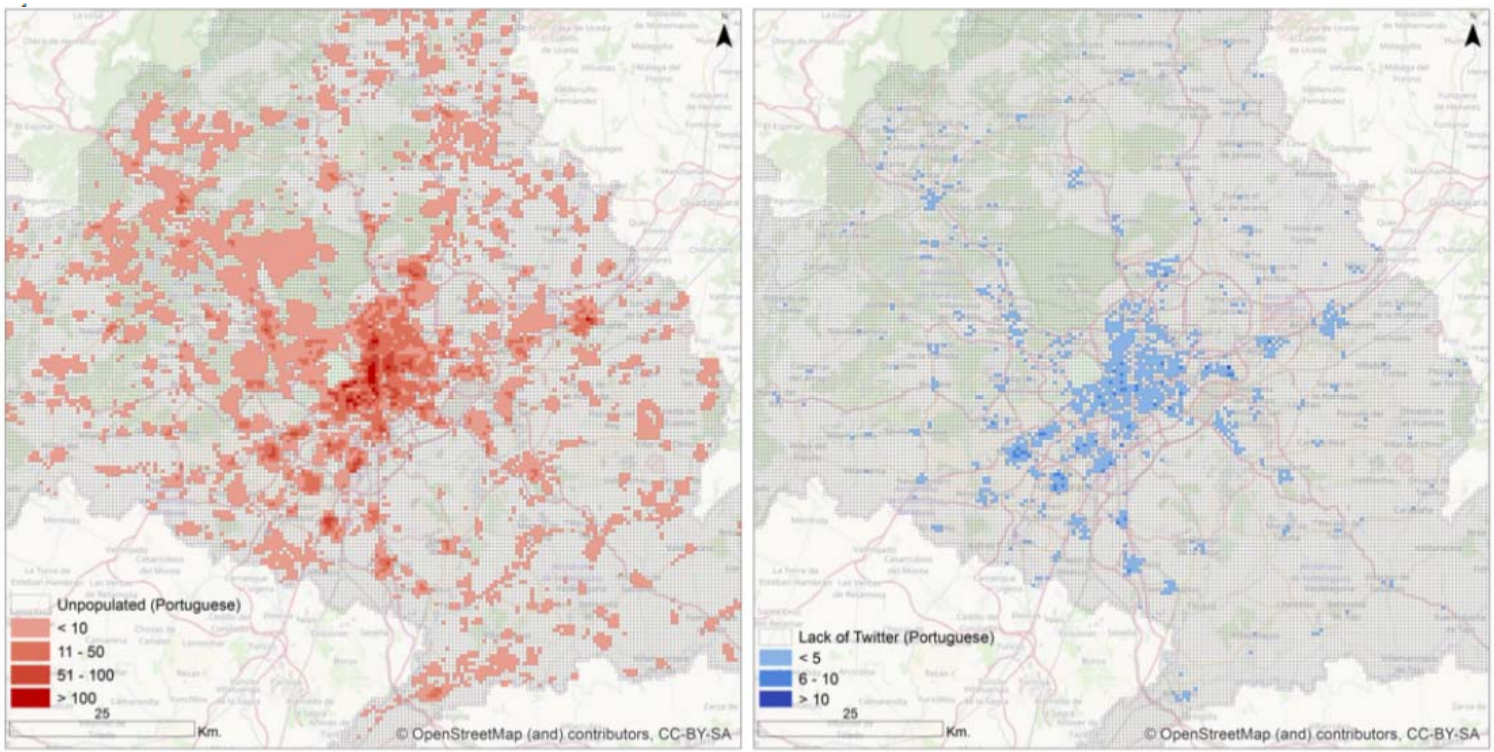
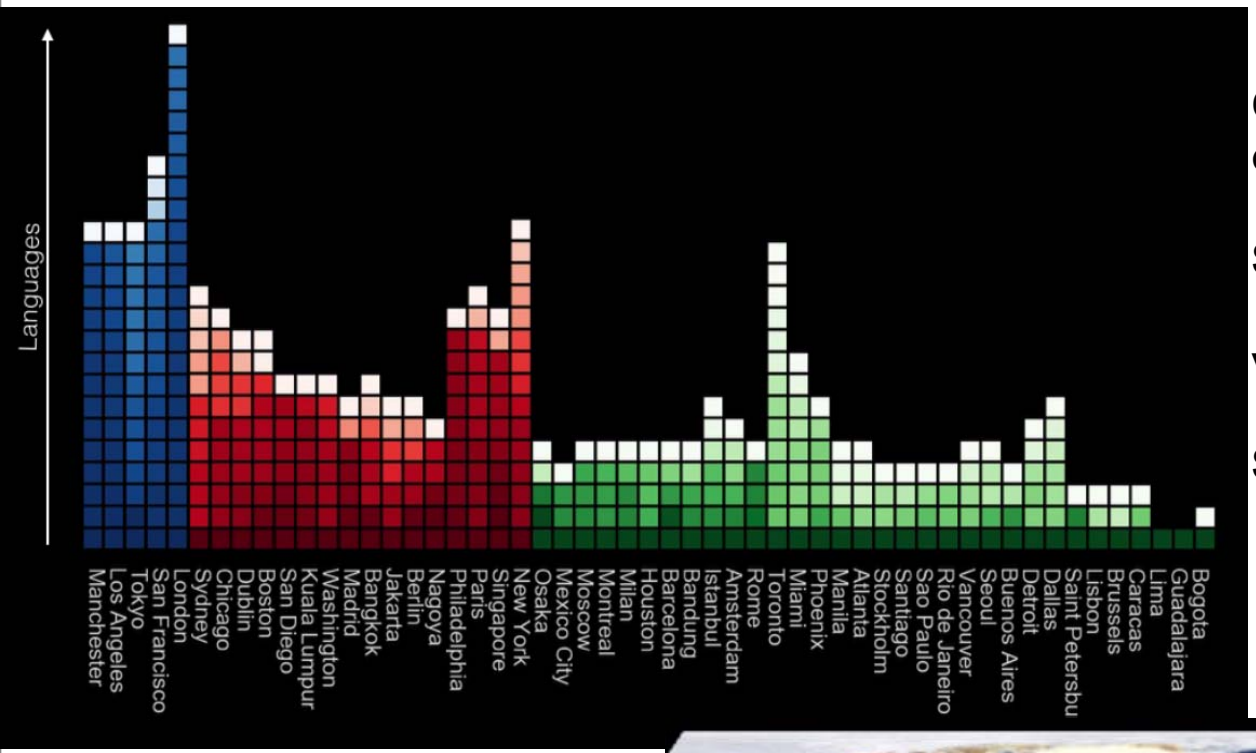


Figure S3. Data Validation (2/2). Distribution of Portuguese native users in Madrid, according to official statistics and to our framework of language detection process.

Ghettos from language use in twitter

ArXiv 1611.01056



Cities form 3 clusters depending on language integration:

Spatial distribution of language

vs

Spatial distribution of population

BIG DATA

Twitter

Telefónica
Telefónica I+D

Cell Phone

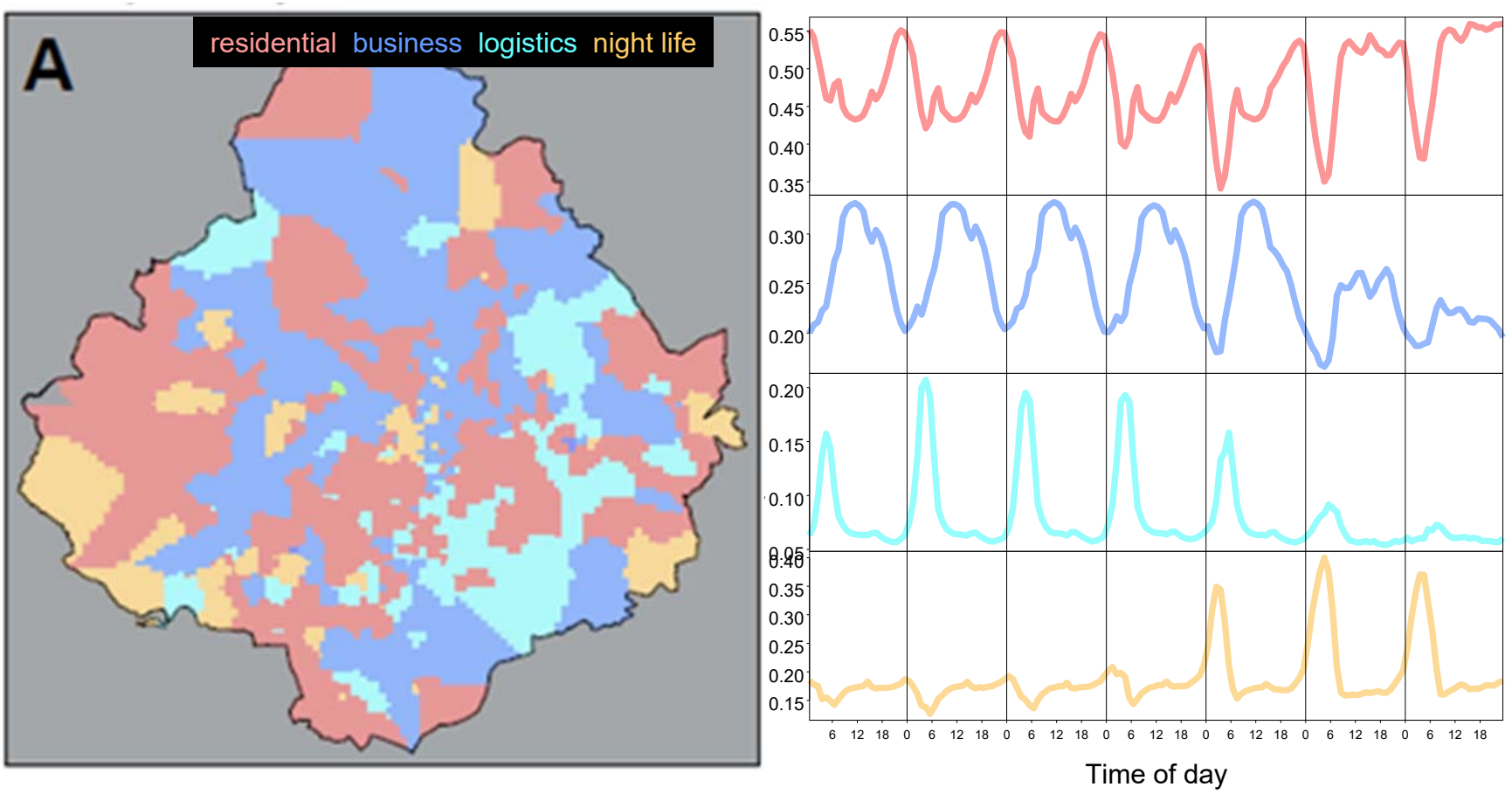
BBVA

Electronic Transactions

Segregation in land use

M. Lenormand et al., Royal Soc. Open Science 2, 150449 (2015)

Landuse from the functional network of the city



Segregation in land use

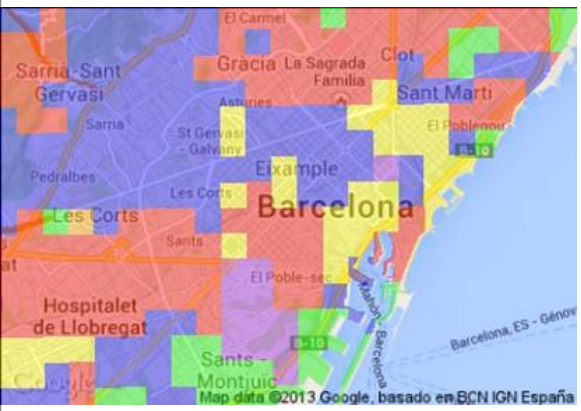
M. Lenormand et al., Royal Soc. Open Science 2, 150449 (2015)

Landuse from the functional network of the city

Network approach to determine land-uses from mobile phone data

Automatic detection of 4 main land use whose relative proportions are very close from one city to another

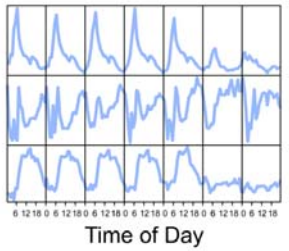
Telefónica I+D



business
residential
logistics
night life

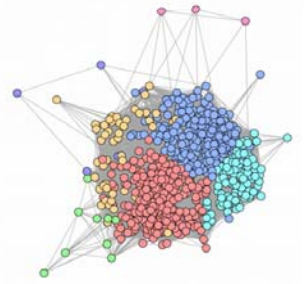
Metropolitan Area

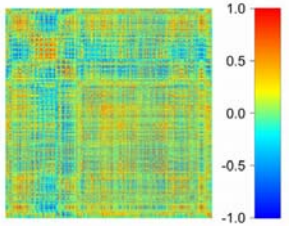
Recordings sites



Signals

Functional Network





Correlation Matrix

M. Lenormand et al., Royal Society Open Science 2, 150449 (2015)

Segregation model *a la* Schelling

Satisfaction S_i of a cell based on the fraction of land use type among its neighbors

p_t^i is the fraction of neighbours of *i* of type *t*

$$\text{if } t_i = L, \quad S_i = \delta_{p_L^i,1},$$

$$\text{if } t_i = N, \quad S_i = p_N^i \delta_{p_L^i,0},$$

$$\text{if } t_i = R, B, \quad S_i = \begin{cases} \delta_{p_L^i,0} & \text{with probability } \gamma, \\ p_{R,B}^i \delta_{p_L^i,0} & \text{with probability } 1-\gamma, \end{cases}$$

Logisitcs gives repulsive forces: logistic cells S_i=1 only if they are surrounded by cells of the same type, and that cells of other types have zero satisfaction if surrounded by any logistic one. Residence and business also attract each other with the only adjustable parameter

Global satisfaction:

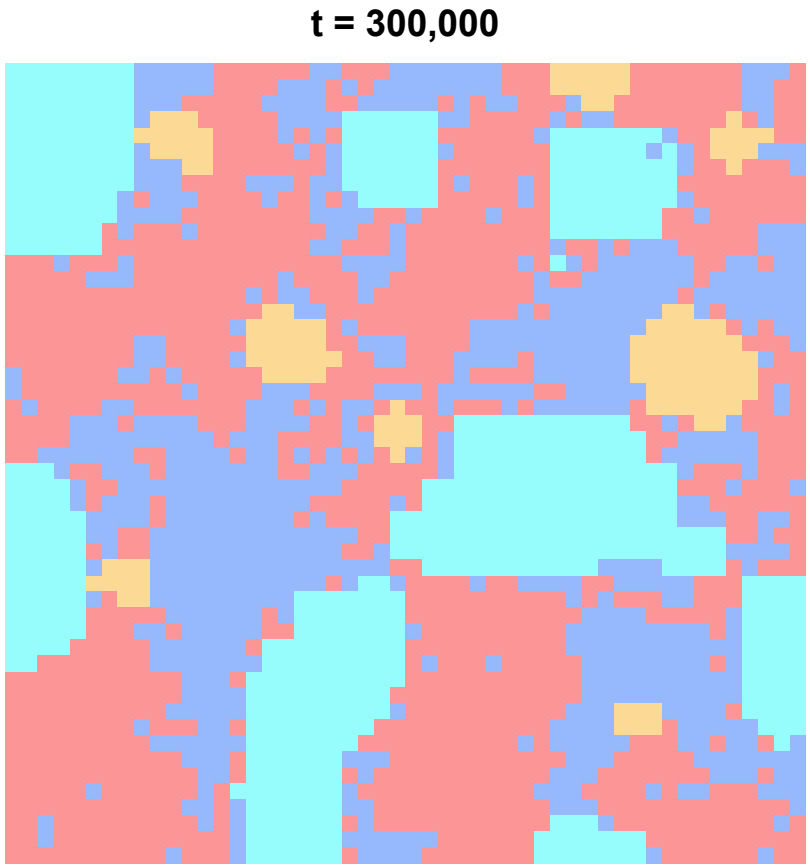
$$S = \sum_i S_i$$

Dynamics:

The model is updated by choosing random pairs of cells and interchanging their land use if the exchange increases *S*. This process is repeated until the satisfaction reaches a stationary state.

M. Lenormand et al., Royal Society Open Science 2, 150449 (2015)

Segregation model *a la* Schelling



M. Lenormand et al., Royal Society Open Science 2, 150449 (2015)

Segregation model *a la* Schelling

Entropy index at different scales. City divided in cells of size DxD

f_i : fraction of area of use α in cell i in city subdivison at scale D

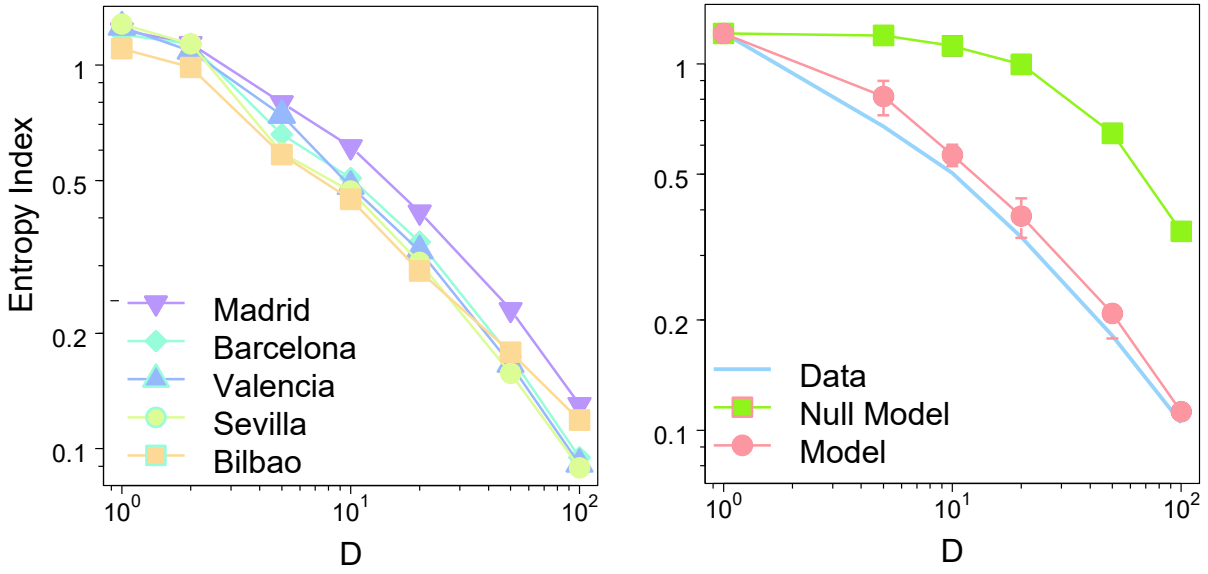
$$E_i = -\sum_{\alpha} f_i^{\alpha} \ln(f_i^{\alpha})$$

α: L,R,B,N

E(D): average of

E_i over cells i

at scale D



BIG DATA

Twitter

Cell Phone

Electronic Transactions

ELECTRONIC TRANSACTIONS

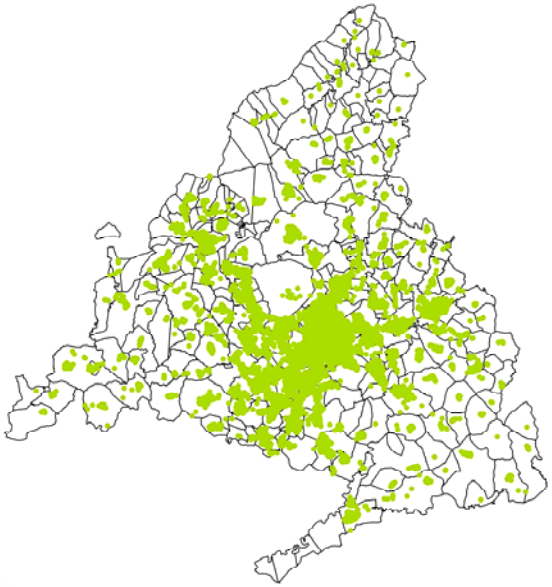
BBVA

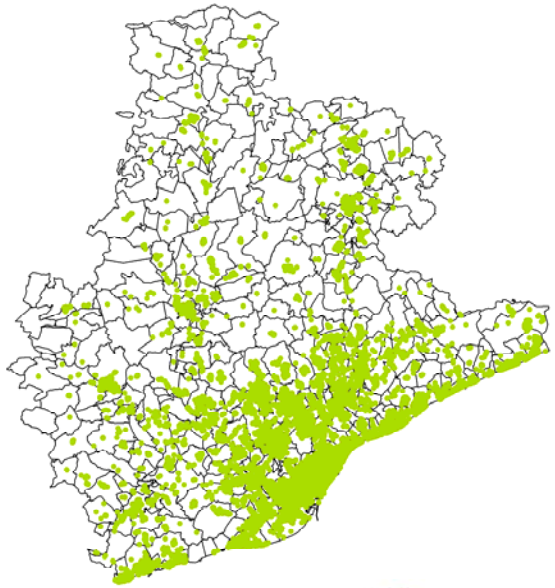
http://www.youtube.com/watch?v=Zel6wych9p0

Provinces of Madrid and Barcelona in 2011-12

- 130 M of transactions
- 3.5 M of **BBVA** customers
- 320,000 businesses

BBVA INNOVATION CENTER





eunoia EVOLUTIVE USER-CENTRIC NETWORKS FOR INTRAURBAN ACCESSIBILITY

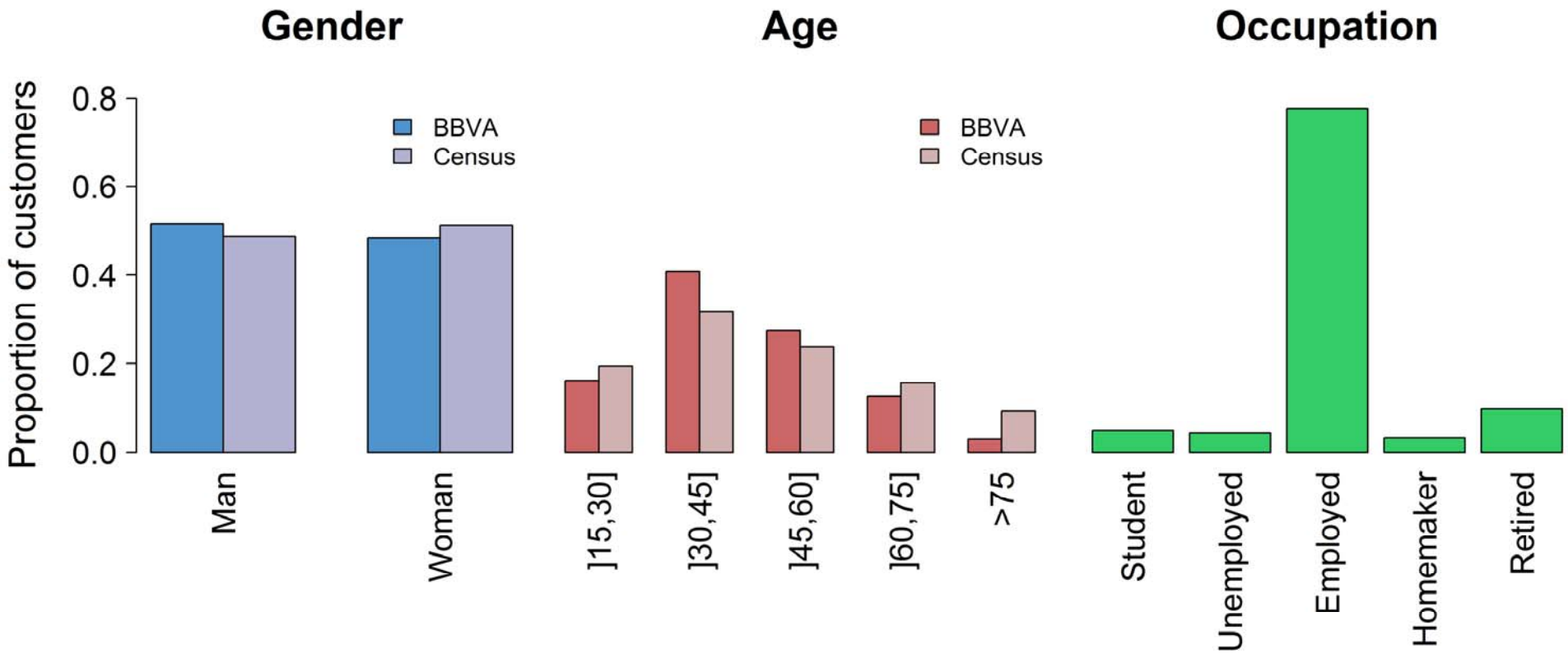
insight

Mobility from electronic transactions

Influence of sociodemographic characteristics on human mobility

M. Lenormand, et al. Sci. Rep. 20 (2015)

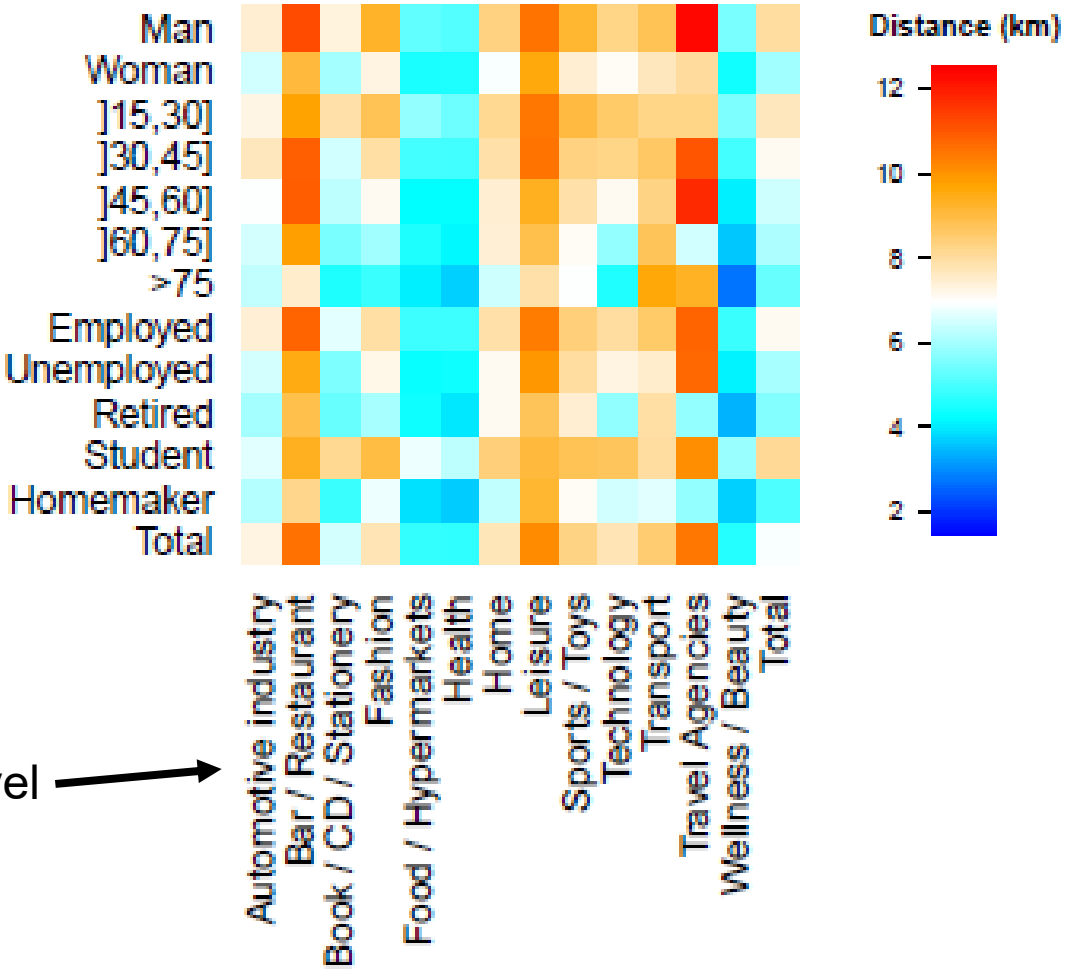
BBVA Data vs Census



Mobility from electronic transactions

M. Lenormand, et al. Sci. Rep. (2015)

Influence of sociodemographic characteristics on human mobility



The end of science/politics?

Do not care to vote, an algorithm is in charge!

Data Science

-Opportunity and Challenge

-Data → Information → Knowledge

-What do we understand when we know everything?:

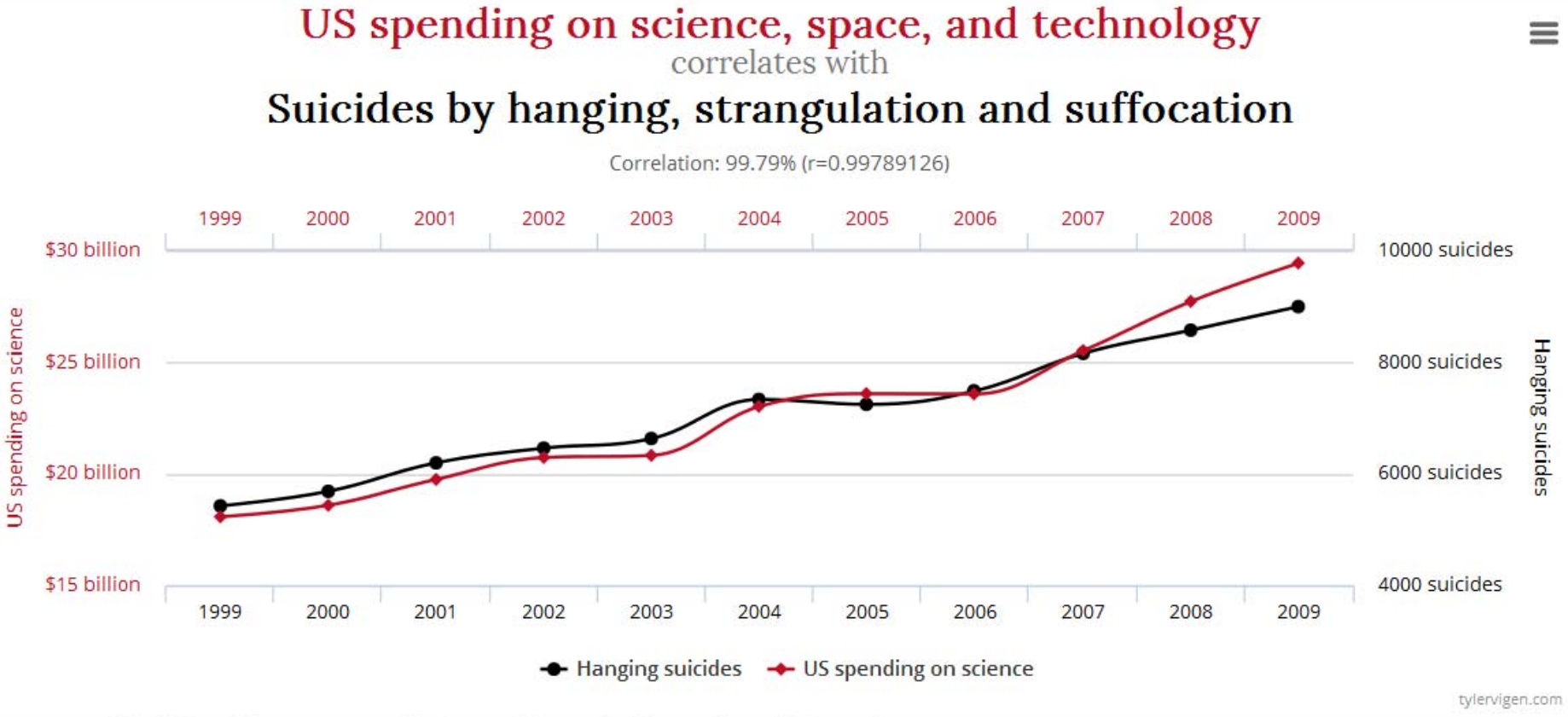
On the face of this 'data deluge', it has been argued we are witnessing the end of theory and that the scientific method is becoming obsolete:
"The new availability of huge amounts of data, along with the statistical tools to crunch these numbers, offers a whole new way of understanding the world. Correlation supersedes causation, and science can advance even without coherent models, unified theories, or really any mechanistic explanation at all."

C. Anderson (2008) The end of theory: The data deluge makes the scientific method obsolete. Wired Magazine.

It is nice to know that the computer understands the problem. But I would like to understand it too (E. Wigner)

http://www.tylervigen.com/

SPURIOUS CORRELATIONS



Data analytics, Machine learning, Deep learning

vs

Modelling

Science as the art of abstraction:

"What do you consider the largest map that would be really useful?" "About six inches to the mile." "Only six inches!" exclaimed Mein Herr. "We very soon got six yards to the mile. Then we tried a hundred yards to the mile. And then came the grandest idea of all! We actually made a map of the country, on the scale of a mile to the mile!" "Have you used it much?" I enquired. "It has never been spread out, yet," said Mein Herr: "The farmers objected: they said it would cover the whole country, and shut out the sunlight! So now we use the country itself, as its own map, and I assure you it does nearly as well (From Lewis Carroll)

Questions and answers:

Computers are useless: They only provide answers! (Pablo Picasso)

Some practical challenges

- **Data:** -Multilevel and multiscale
  -Data mining vs Data analytics
  -From Data to Information to Knowledge
  -Data driven vs. Question driven
  -Data privacy
  -Data sharing: Academics, Companies, Government

- **Limits of prediction/forecasting vs decision making**

- **Difficult dialogue:** Across disciplines, across policy sectors

- **Science in support of policy options:**

  -Policy makers are generally not interested in discussing challenges unless they come with solutions.

  -Furthermore they are interested in a single answer, but a single answer often is not found

CLASH OF TITANS

INDUCTIVISM
F. Bacon

Data driven
Big Data
Inference

DEDUCTIVISM
Carl Popper

Model driven

Complexity science is the discipline in which

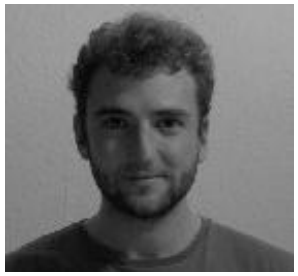
both these approaches merge as one.

Peter Sloot: Visions for Complexity (2016)

THANKS!

José J. Ramasco

Pere Colet



Maxime Lenormand

Fabio Lamanna

Antònia Tugores

Aleix Bassolas