Use of social media data for official statistics

International Conference on Big Data for Official Statistics, October 2014, Beijing, China

Big Data Team

INSTITUTO NACIONAL
DE ESTADÍSTICA Y GEOGRAFÍA

Content

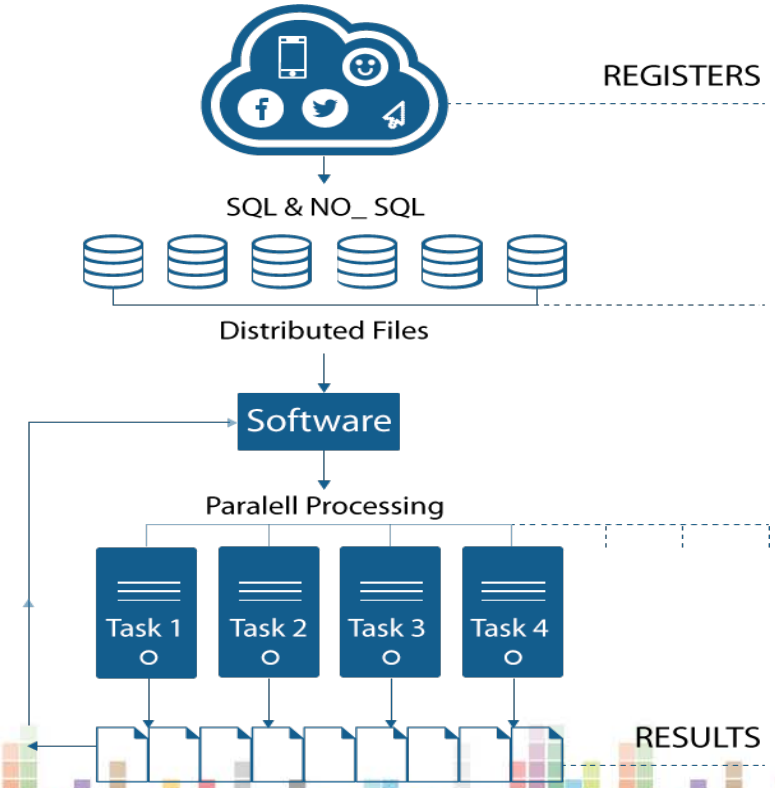
1. Why Twitter
2. Subjective well-being
3. Tourism exercises
4. Mobility studies
5. Next projects
6. Institutional strategy

Why Twitter?

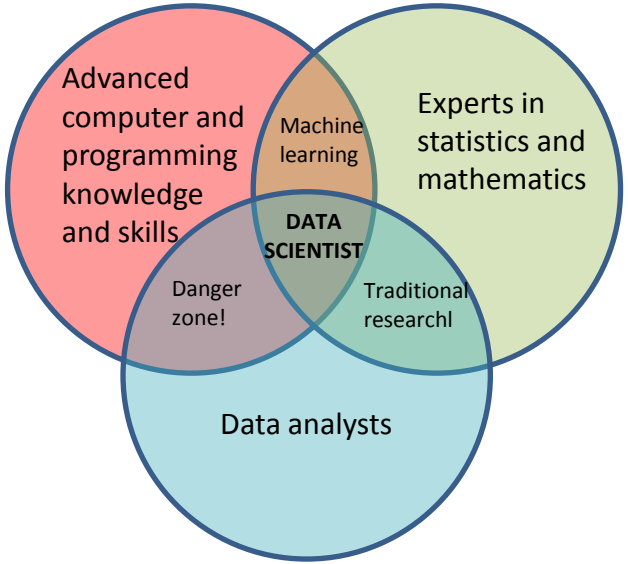
- Readily available
- Up to 1% of global tweets at no cost
- Around 12 M accounts in Mexico
- Geo-located tweets from 700 thousand accounts
- 90 M plus tweets downloaded since January 2014

Big Data Requirements

Infrastructure

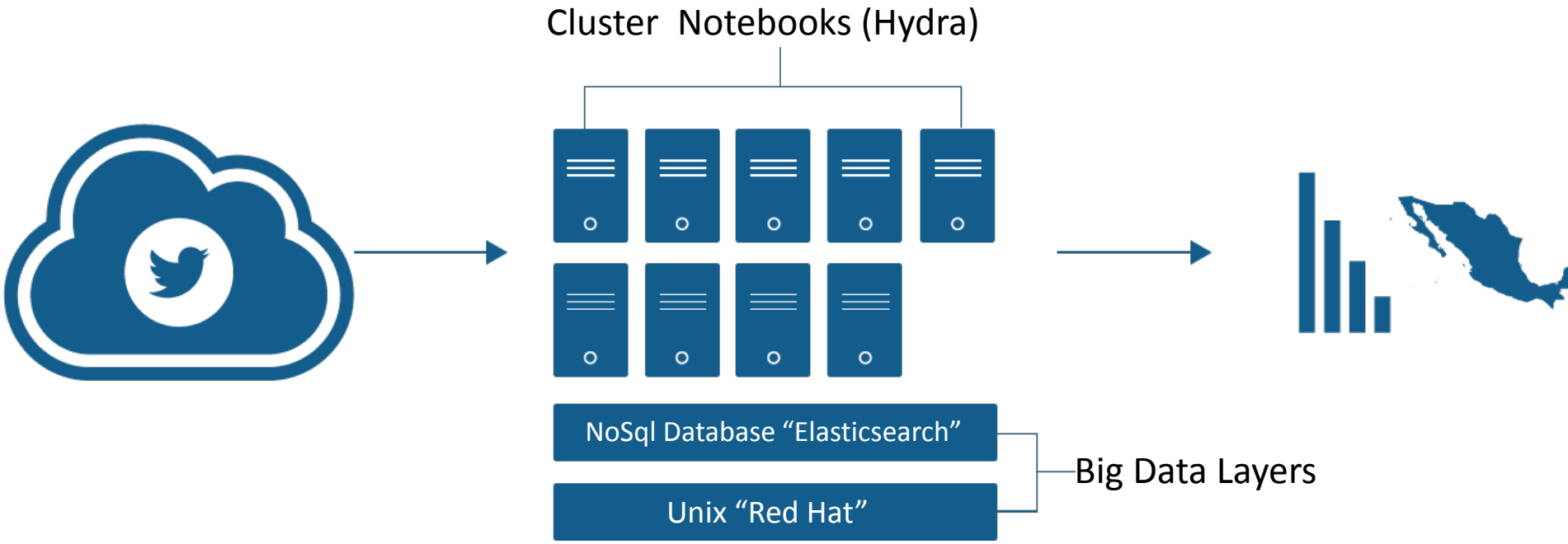


Multidisciplinary task team

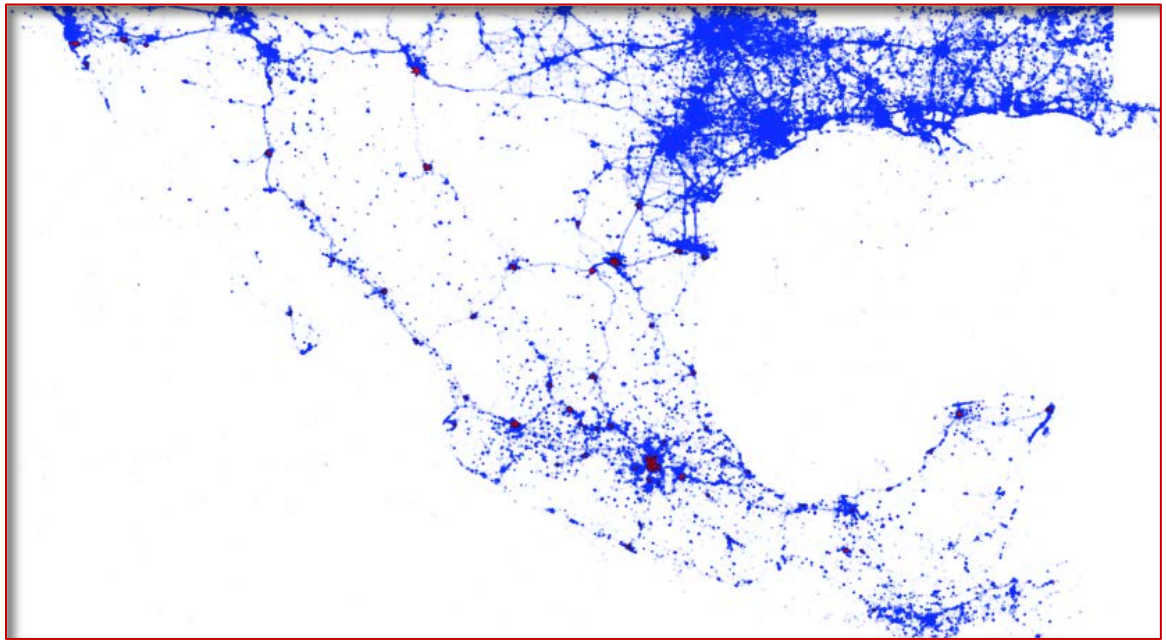


INSTITUTO NACIONAL DE ESTADÍSTICA Y GEOGRAFÍA

Infrastructure for collecting tweets



First trials with 20 M tweets…



Subjective well-being

Twitter for Subjective Well-being

Research to use Twitter:

Data source:

- Structure of the tweet
- Ease of access
- Randomness
- Geographic filters

Studies in other countries:

"Sentiment Analysis" University of Pennsylvania

"Mood of the Nation" UK

"Big Data and Official Statistics" NL

"Workshop on Sentiment Analysis 2013" of the Spanish Society for Natural Language Processing (SEPLN)

Classifications

Naive Bayes

Support Vector Machines (SVM)

KNN

Word Count

Lists of words and dictionaries used for sentiments analysis

Spanish Emotion Lexicon (SEL) KNN

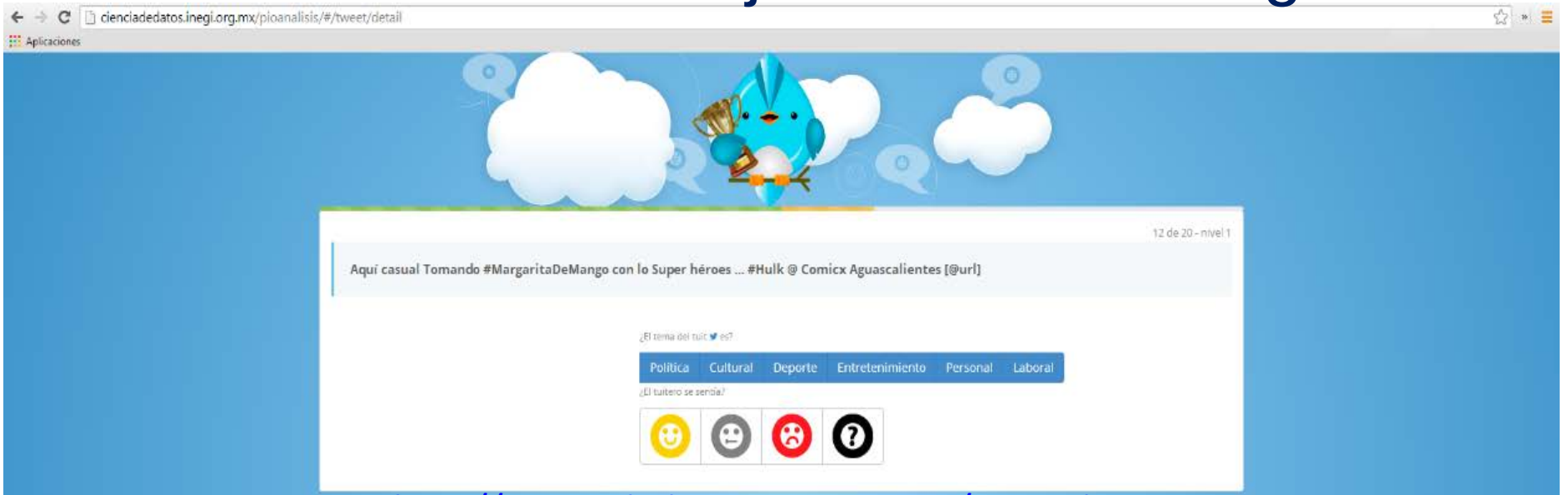
AFINN, WordNet, ANEW

Twitter for Subjective Well-being

Process to analyze tweets:

- Word counting (single or tokenized) will not be used.
- Dictionaries for sentiment analysis will not be used either.
- Instead, supervised learning method will be used.
  - Sentiment in a sample of tweets is graded by humans resulting in a training set.
  - This set is then used to find similarities in order to classify the remaining and future tweets for sentiment.

Twitter for Subjective Well-being



http://cienciadedatos.inegi.org.mx/pioanalisis

An app was developed by INEGI in order to classify tweets for sentiments: positive, negative or neutral, and then allocate them to different domains.

Universidad Tec Milenio supports the exercise with about 3000 students who are classifying the training set

Twitter for Subjective Well-being

Partnerships with industry:

- The Spanish firm Lambdoop offered free of charge, the implementation of the processing software with the method called “supervised learning” to extract the sentiment from the tweets.
- The Dattlas Division of the firm KioNetworks, offered free of charge, the provision of a cluster with enough capacity to carry out our pilot tests and be able to identify, assess and to budget the HW and SW requirements.

Tourist Exercise

Twitter- Tourist Exercise

SECTUR
SECRETARÍA DE TURISMO

Background:

- Collaboration with the Ministry of Tourism.
- 95% of Twitter users in Mexico are in the age group of 18 to 49.
- 87% of domestic tourists are between 18 and 55 years old.

Twitter- Tourist Exercise

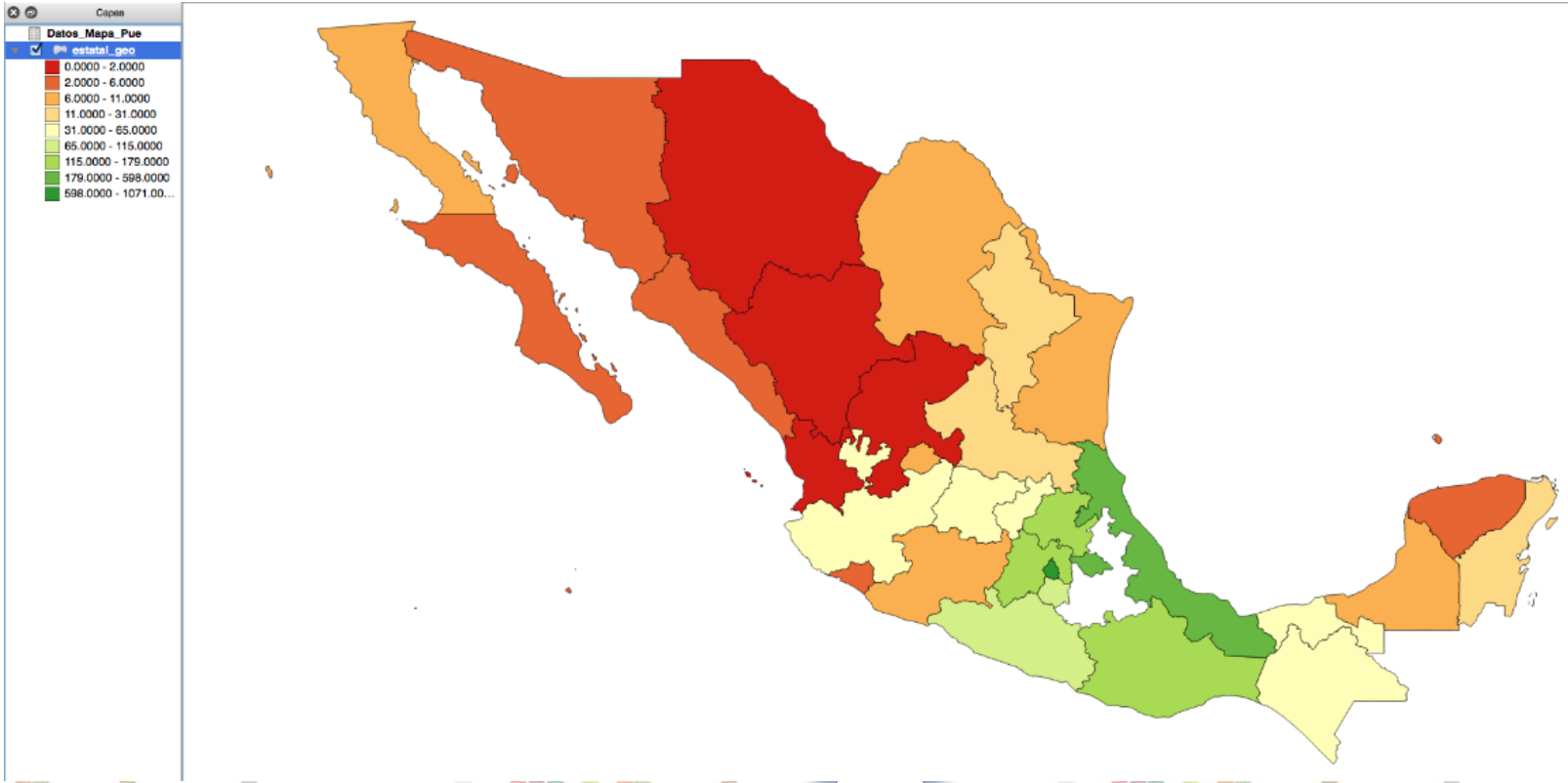
SECTUR
SECRETARÍA DE TURISMO

Production of statistics for the tourist sector:

- 60 M tweets were processed from January to July 2014.
- Guanajuato y Puebla, days February 1st, 2nd y 3rd.
- 7,955 twitters: 827,424 tweets produced from any other state during a 6 month period.
- Find out from which state people twitted and how long had they stayed in that state.
- Stays of 1 to 15 consecutive days in Puebla or Guanajuato.
- With data obtained generate a map for Puebla and another for Guanajuato.

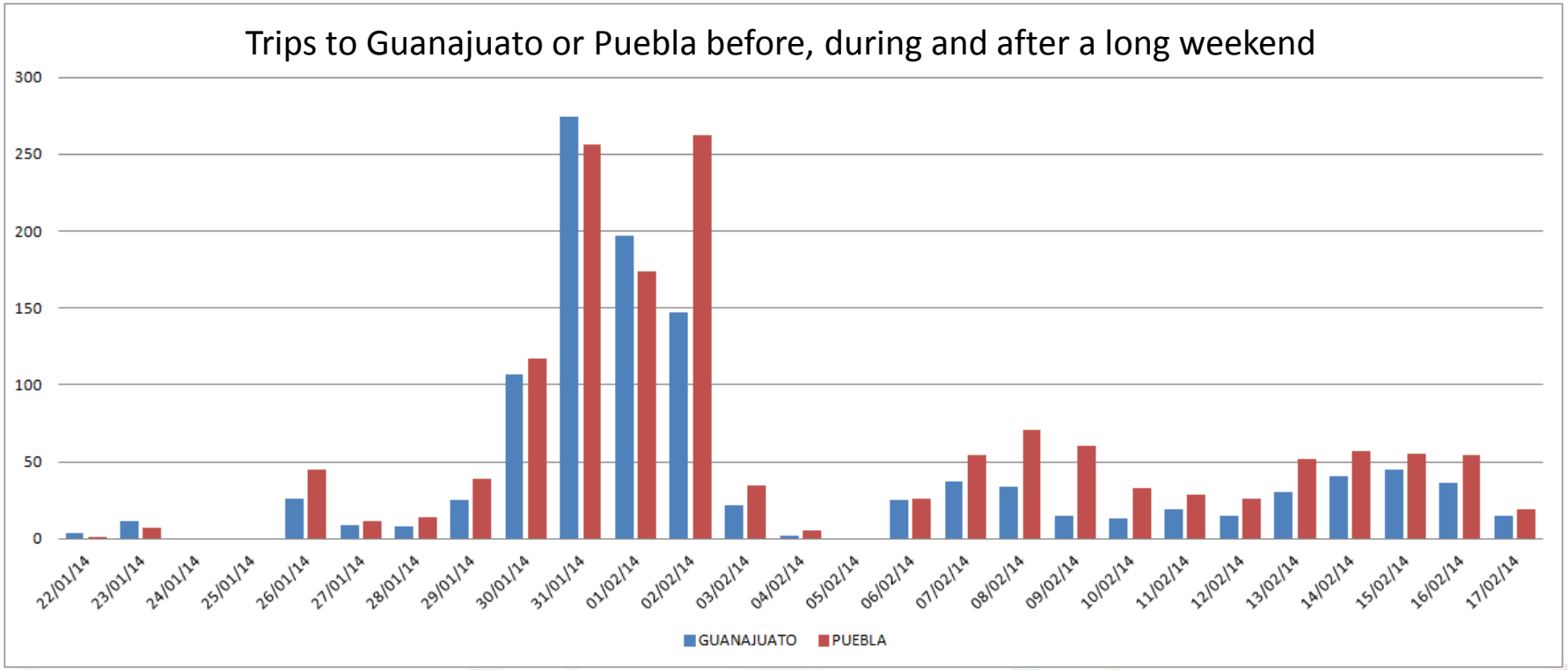
INEGI
INSTITUTO NACIONAL DE ESTADÍSTICA Y GEOGRAFÍA

A. Twitter- Tourist Exercise



B. Twitter- Tourist Exercise

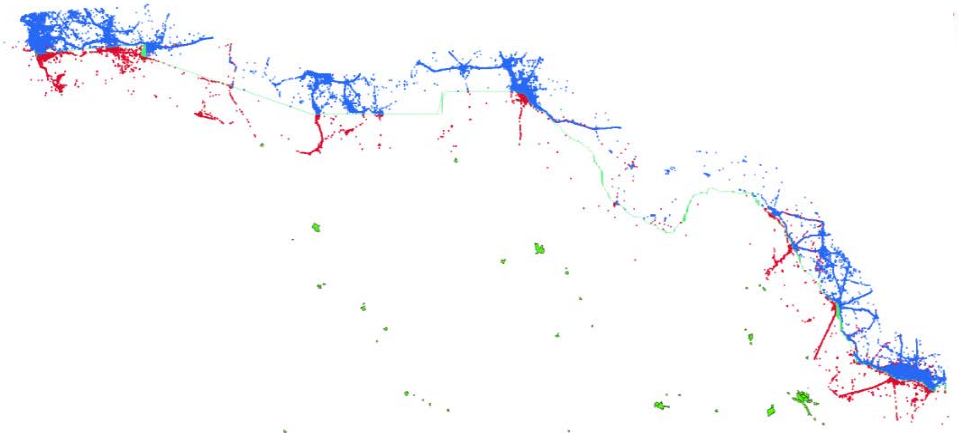
B. Twitter- Tourist Exercise



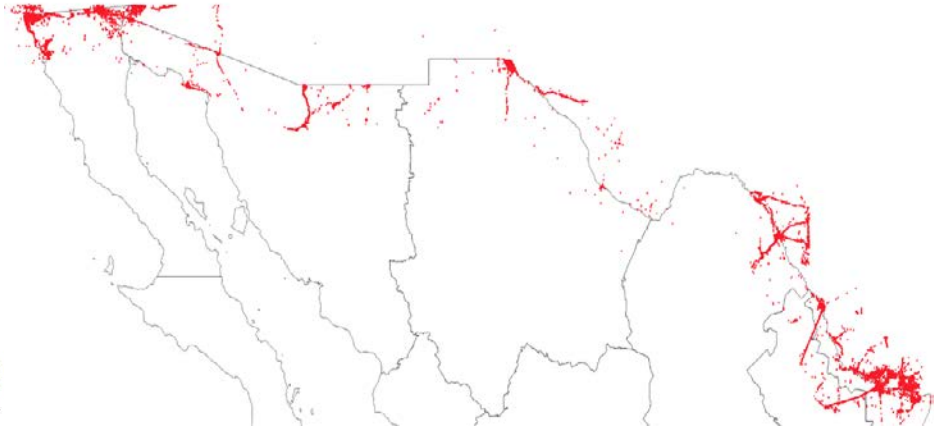
Mobility studies

Mobility studies

Research for development of an analytical method to measure trans-border mobility trough GPS tweets.



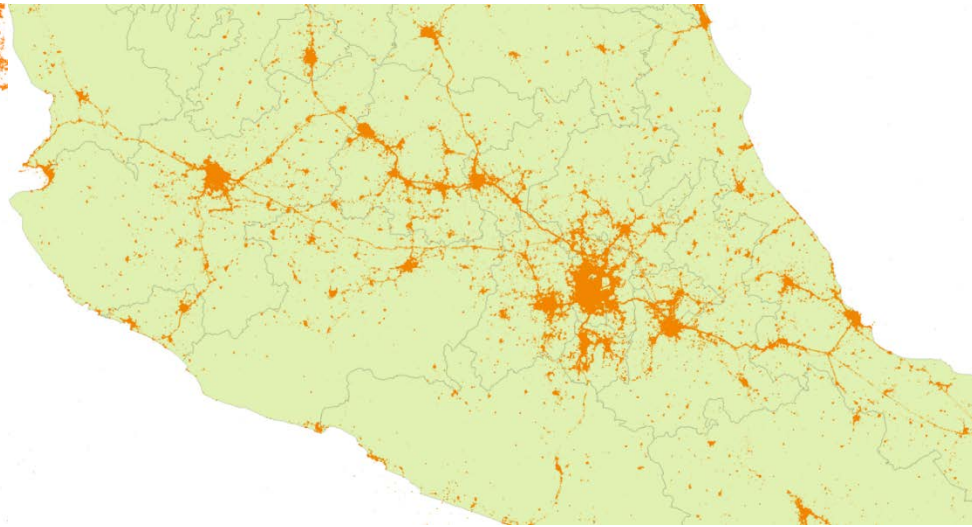
Mexican (red) and US(blue) tweets



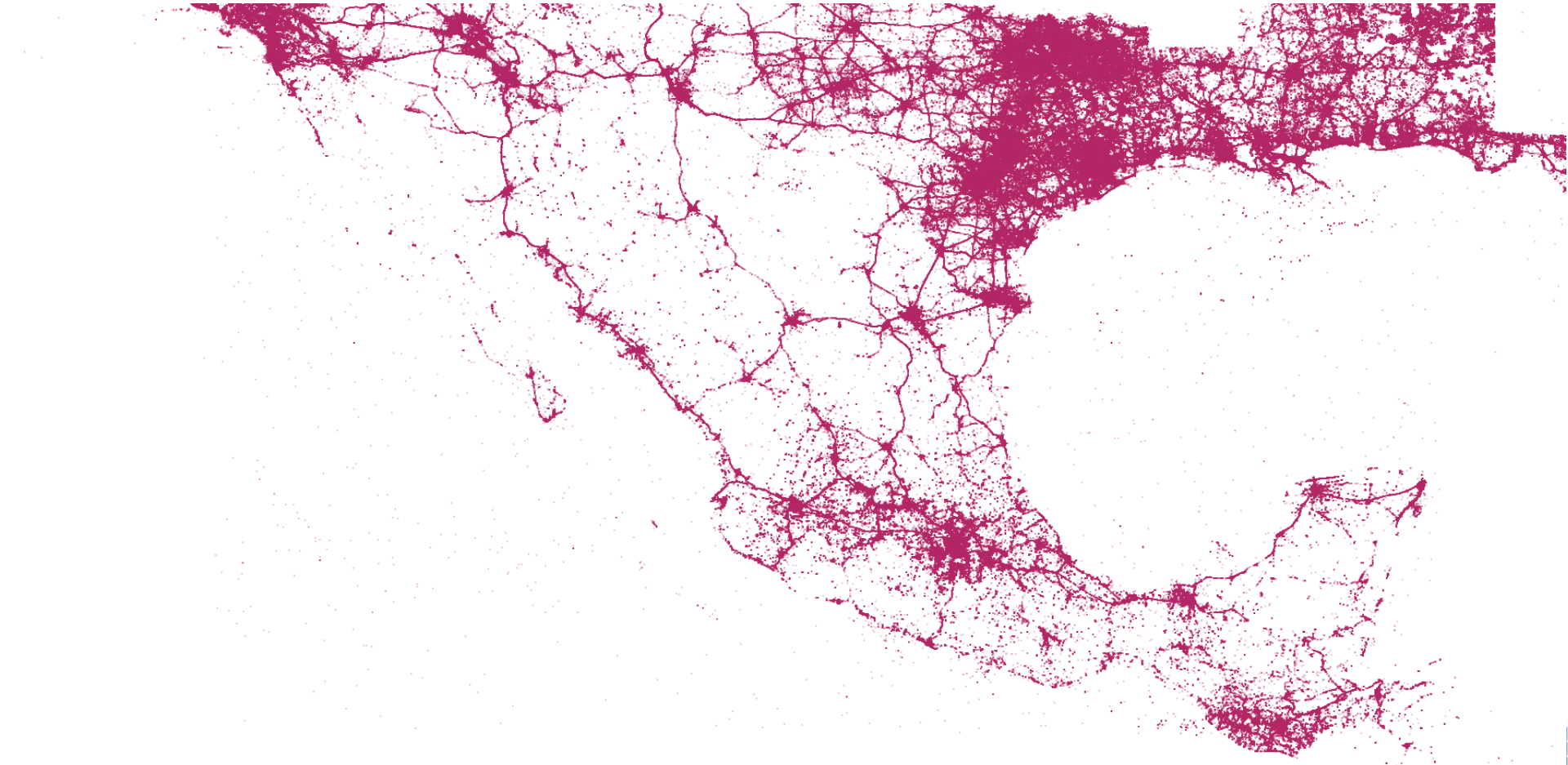
Mexican tweets

Mobility studies

Using the National Roads Network to find out whether domestic mobility analysis can be conducted, current work. (Plot of 70 M tweets)



Mobility studies



Next projects and the institutional strategy

Other potential projects from Twitter

- Consumer confidence
- Indicators on public safety
- Definition of metropolitan areas
- Quality of tourist services
- Conformation of regions

Institutional steps:

- Framework: **Data Revolution**: get everyone involved
- **Big Data**
  - ✓ Participate in the main national and international **initiatives**
  - ✓ **Partnerships** with the government, research centers and universities
  - ✓ **Explore** methodologies, data sets,…
  - ✓ Approach the **private sector**
  - ✓ Publish results as **experimental data**
  - ✓ Form an **inter-disciplinary** task team

INEGI
INSTITUTO NACIONAL DE ESTADÍSTICA Y GEOGRAFÍA

Institutional steps:

- **Specific actions**:
  - ✓ Big Data project jointly funded by CONACYT and INEGI
  - ✓ Partnerships with research centers:
    Infotec (programing, data processing, infrastructure);
    Centro Geo (geographical research);
    CIMAT (mathematics).
  - ✓ Establish a Big Data Laboratory in Aguascalientes

Conociendo México

01 800 111 46 34

www.inegi.org.mx
atencion.usuarios@inegi.org.mx

@inegi_informa

INEGI Informa

INSTITUTO NACIONAL
DE ESTADÍSTICA Y GEOGRAFÍA