Use of data from social networks to obtain statistical and geographical information

Global Conference on Big Data for Official Statistics

October 2015, Abu Dhabi, UAE

INSTITUTO NACIONAL DE ESTADÍSTICA Y GEOGRAFÍA

Purpose of the presentation

Sharing Mexican experiences about the use of tweets to obtain statistical information by the National Institute of Statistics and Geography of Mexico, INEGI.

Pilot Test

CONTEXT

National Institute of Statistics and Geography of Mexico, INEGI

INEGI is a Constitutionally **Autonomous Entity** with:

- **Statistical and geographical** responsibilities
- Around 20,000 employees
- **10 Regional** Offices and **34 State** Offices all around the country
- **5 General Directorates** which generate and integrate statistical data: economic, social, demographic, government, public safety and justice

INEGI is responsible for Mexico's **National System of Statistical and Geographical Information**

INSTITUTO NACIONAL DE ESTADÍSTICA Y GEOGRAFÍA

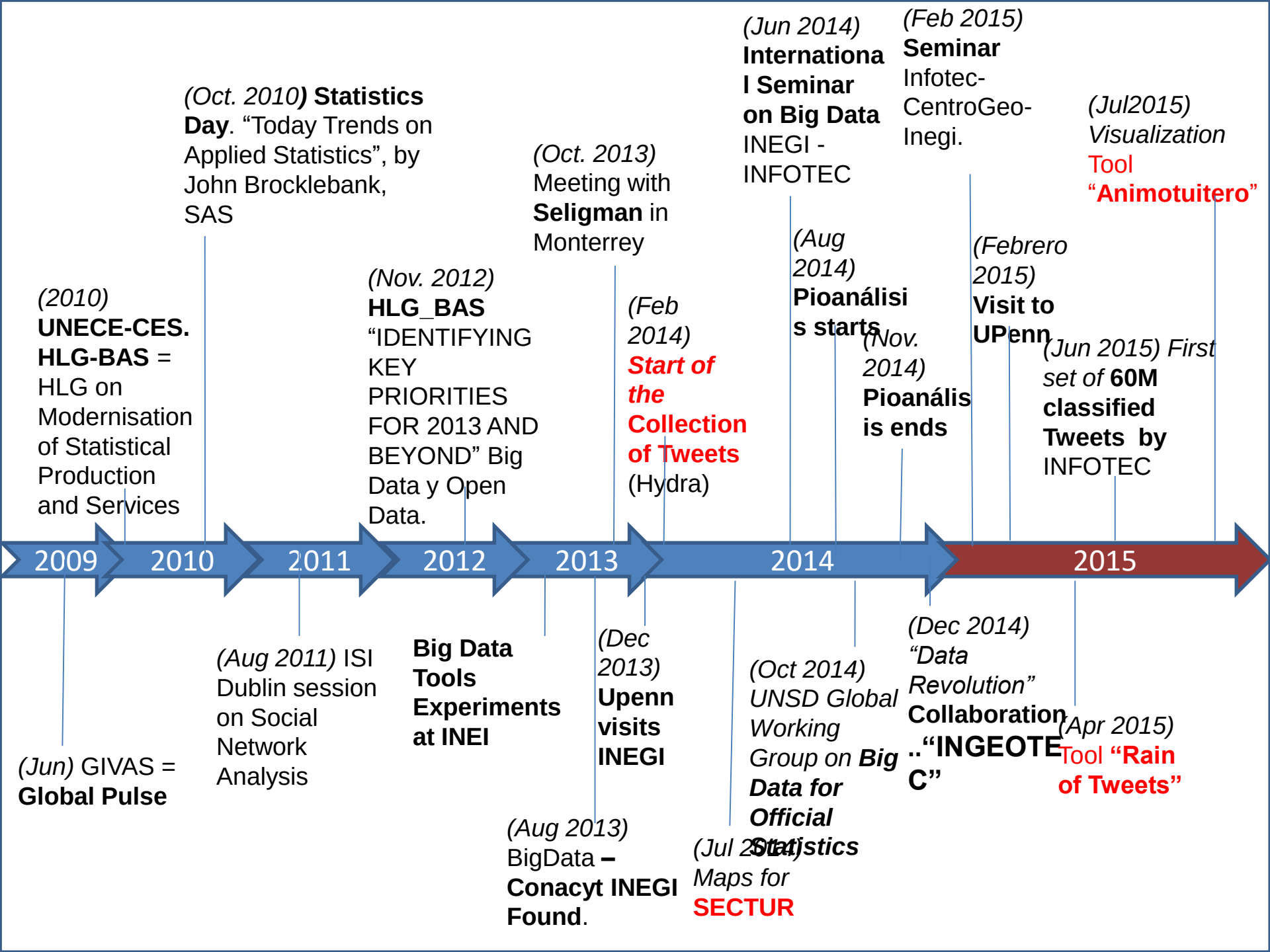
Emergence and Evolution of new information sources

INEGI
INSTITUTO NACIONAL
DE ESTADÍSTICA Y GEOGRAFÍA

Big Data Sources

- Smart meters and sensors
  - Traffic cameras, GPS devices, price scanners, power monitors, smart watches, smart phones, etc.
- Social Interactions
  - Talks and publications on social networks like Twitter, Facebook, FourSquare, etc.
- Business Transactions
  - Movements of credit cards, electronic cash registers, cell phone records, etc.
- Electronic files
  - Documents which are available in electronic formats such as PDF files, websites, videos, audio, digital media broadcasting
- Broadcast media
  - Digital video and audio produced on real time

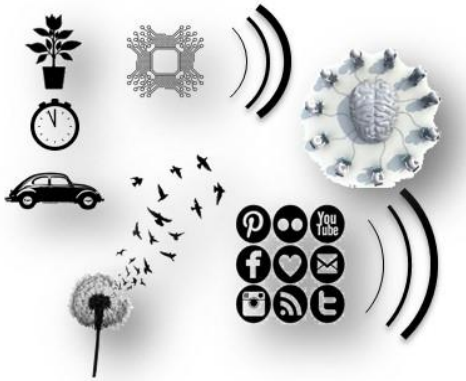
INEGI
INSTITUTO NACIONAL DE ESTADÍSTICA Y GEOGRAFÍ


(2010) UNECE-CES. HLG-BAS = HLG on Modernisation of Statistical Production and Services
(Oct. 2010) Statistics Day. "Today Trends on Applied Statistics", by John Brocklebank, SAS
(Nov. 2012) HLG_BAS "IDENTIFYING KEY PRIORITIES FOR 2013 AND BEYOND" Big Data y Open Data.
(Oct. 2013) Meeting with Seligman in Monterrey
(Feb 2014) Start of the Collection of Tweets (Hydra)
(Jun 2014) International Seminar on Big Data INEGI - INFOTEC
(Aug 2014) Pioanálisis starts
(Nov. 2014) Pioanálisis ends
(Feb 2015) Seminar Infotec-CentroGeo-Inegi.
(Febrero 2015) Visit to UPenn
(Jul2015) Visualization Tool "Animotuitero"
(Jun 2015) First set of 60M classified Tweets by INFOTEC
2009
2010
2011
2012
2013
2014
2015
(Jun) GIVAS = Global Pulse
(Aug 2011) ISI Dublin session on Social Network Analysis
Big Data Tools Experiments at INEI
(Aug 2013) BigData – Conacyt INEGI Found.
(Dec 2013) Upenn visits INEGI
(Jul 2014) Maps for SECTUR
(Oct 2014) UNSD Global Working Group on Big Data for Official Statistics
(Dec 2014) "Data Revolution" Collaboration .."INGEOTEC"
(Apr 2015) Tool "Rain of Tweets"

Governing Board Commitment

Technological Landscape

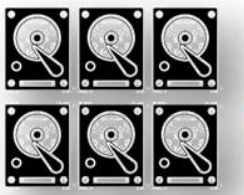
Internet of things, people and ideas

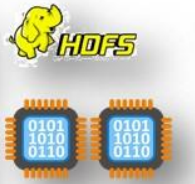


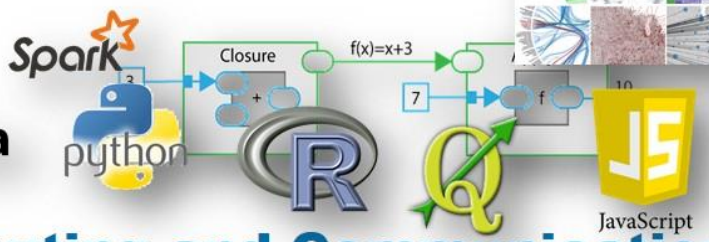
Advanced statistics, mathematics and data Science

Signal processing	Uncertainty modeling
Probability models	Data warehousing
Machine learning	Data compression
Statistical learning	computer
Data mining	Programming
Database	High-performance
Data engineering	computing
Pattern recognition	Geolocation
Learning patterns	Geo-referencing
Predictive analytics	…

Business Knowledge (experts)







IT Infrastructure: Robust Computing and Communications, Specialized software tools for processing, analysis, visualization, etc.

Institutional Cooperation

- National
- International

INFOTEC

CENTROGEO

CIDE CENTRO DE INVESTIGACIÓN Y DOCENCIA ECONÓMICAS, A.C.

MÉXICO
PRESIDENCIA DE LA REPÚBLICA

UNIVERSIDAD TECMILENIO.
Innovación con propósito de vida.

KIO Networks

SECTUR
SECRETARÍA DE TURISMO

dattlas
WORLD OF INFO BIG DATA

Penn
University of Pennsylvania

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

United Nations
Statistics Division

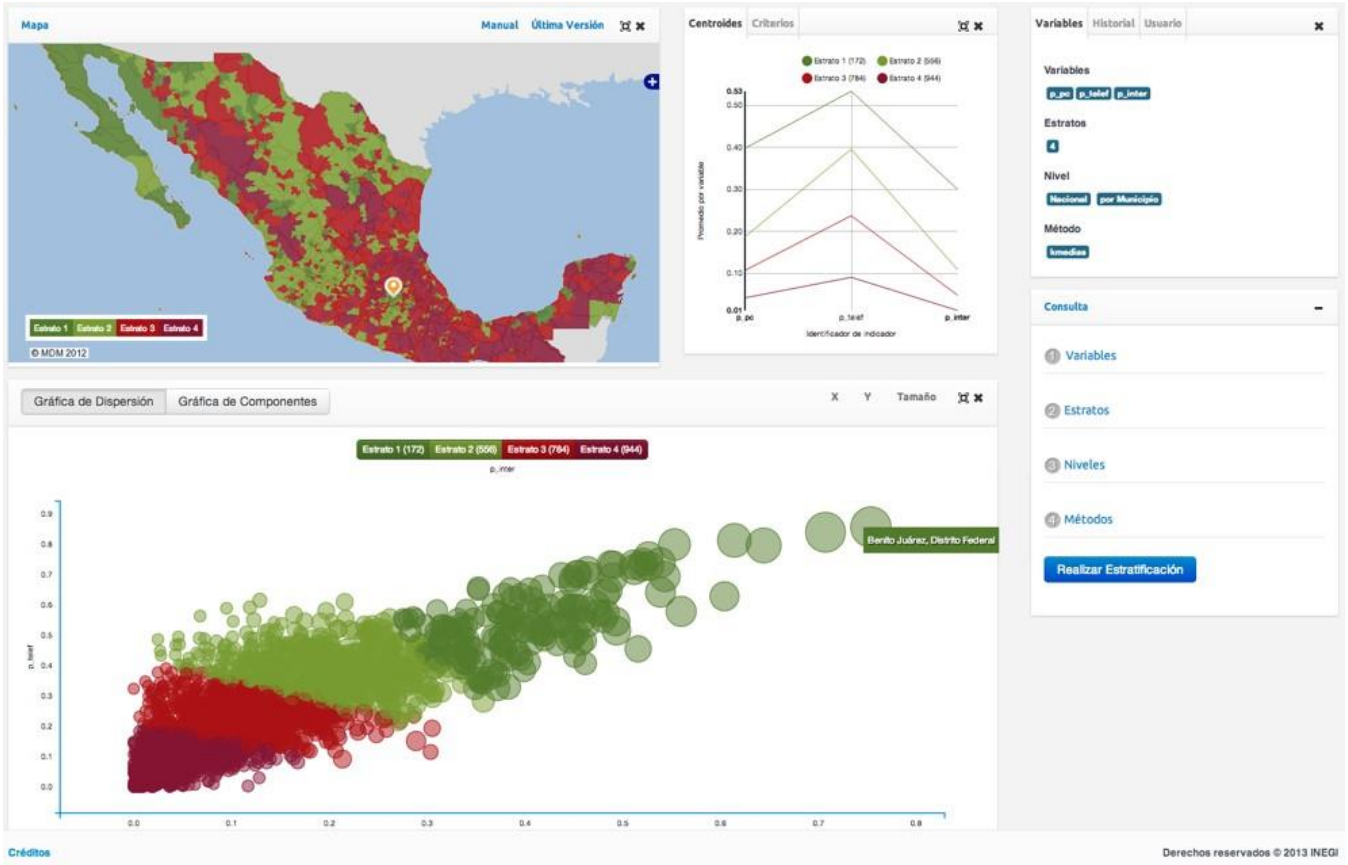
ICHEC
Irish Centre for High-End Computing

LAMBDOOP

INEGI
INSTITUTO NACIONAL
DE ESTADÍSTICA Y GEOGRAFÍA

TWEETS COLLECTION

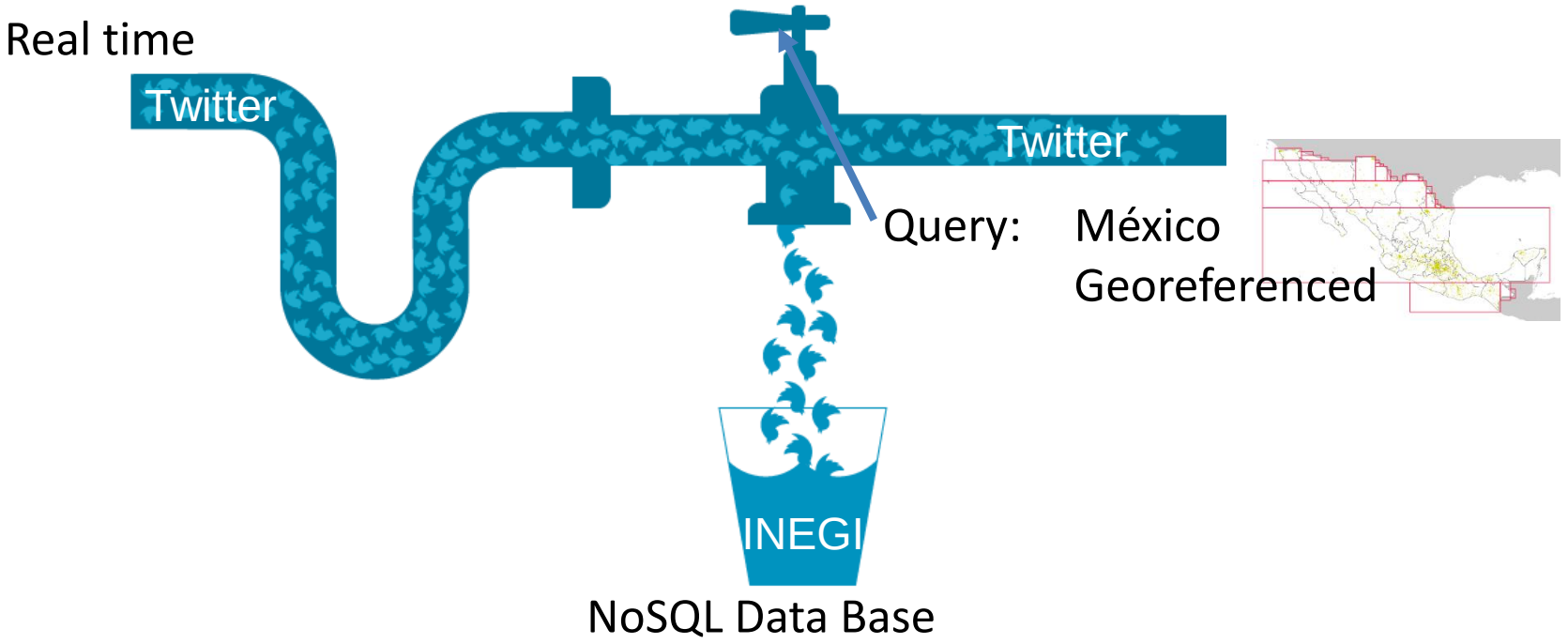
Stratification: First Efforts



www.inegi.org.mx/est/contenidos/Proyectos/estratificador/

INSTITUTO NACIONAL
DE ESTADÍSTICA Y GEOGRAFÍA

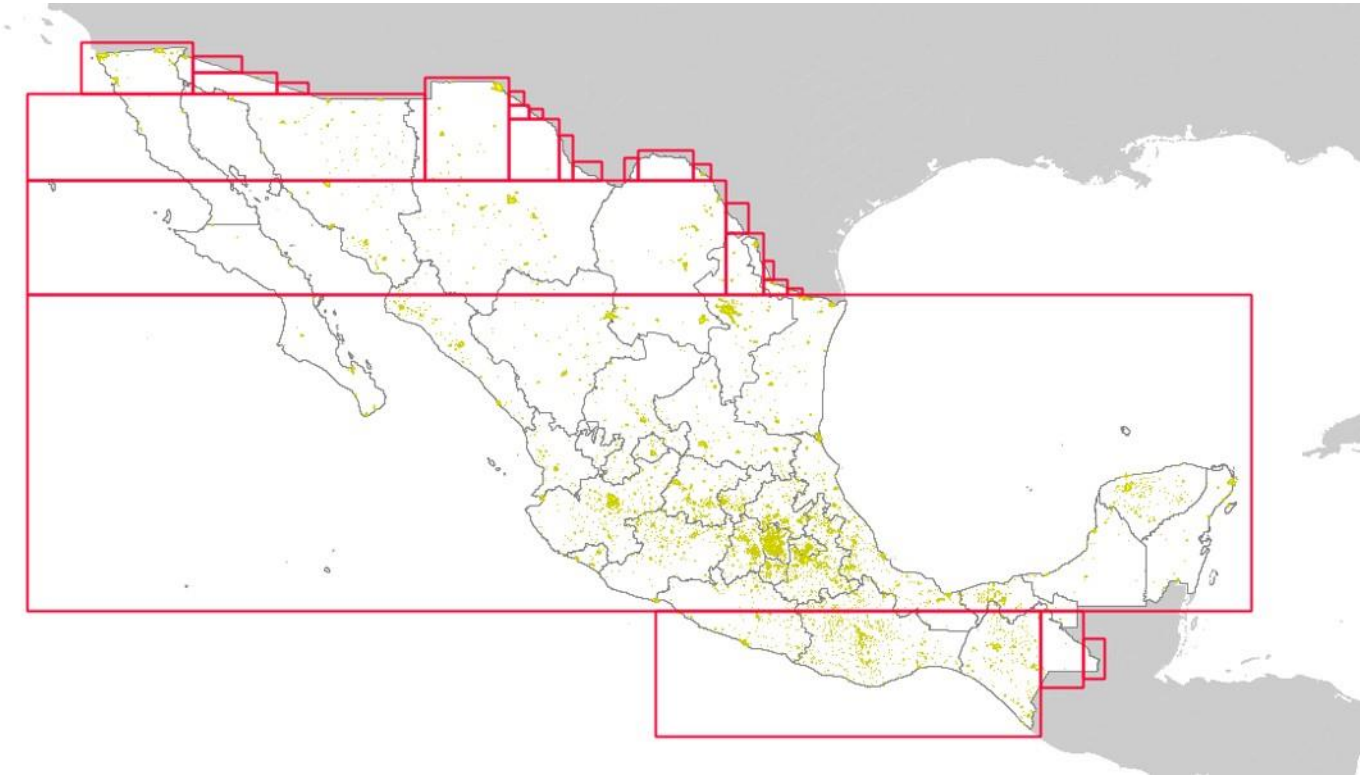
Twitter as a data source



Why Twitter?

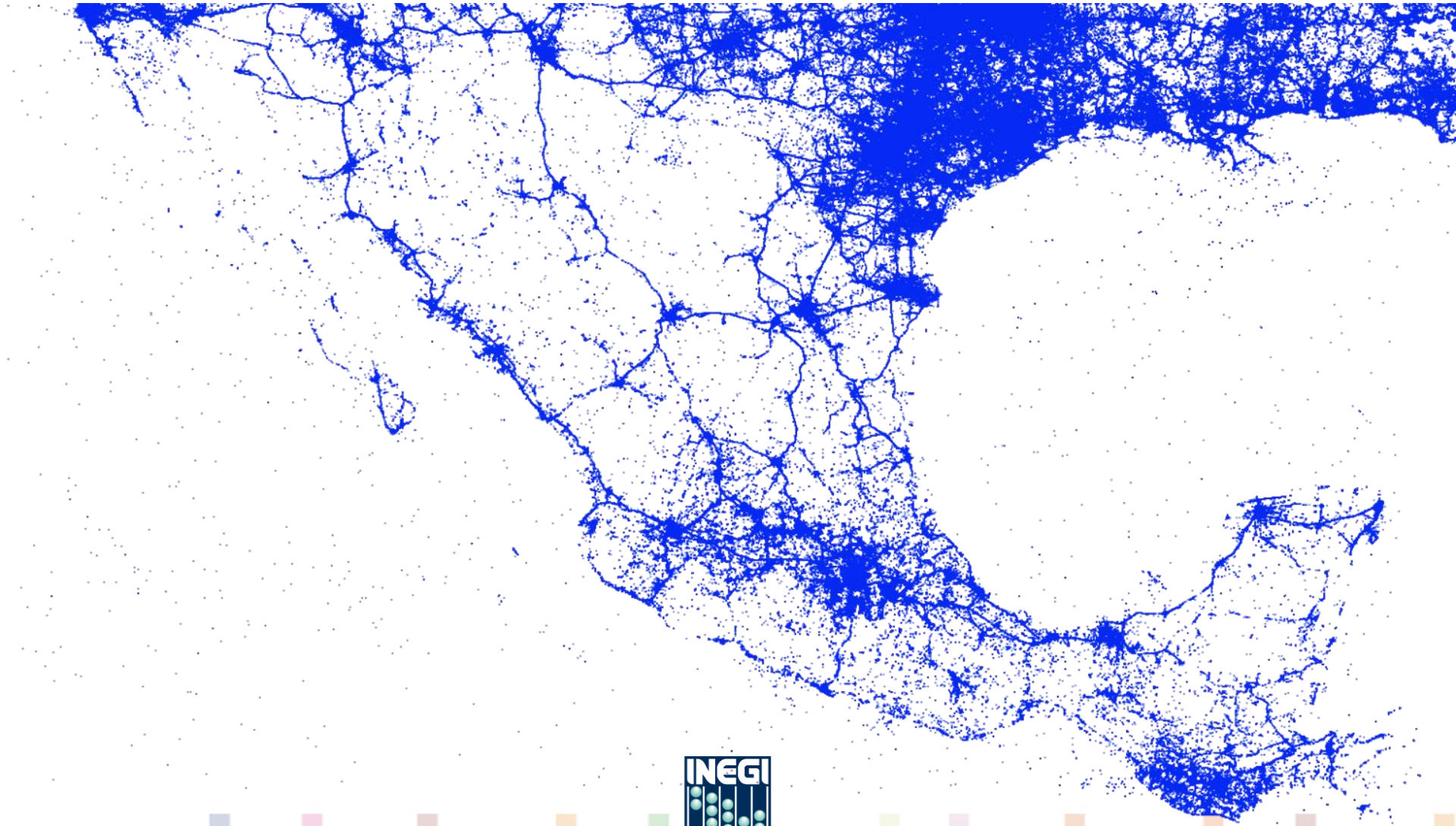
- Readily available
- Up to 1% of global tweets at no cost
- Around 12 M accounts in Mexico
- Geolocated tweets by 700 thousand accounts
- 150 M plus tweets downloaded since January 2014
- Even though its drawbacks: Not documented, not supported by “traditional” statistics methodologies

Polygon for the collection of tweets



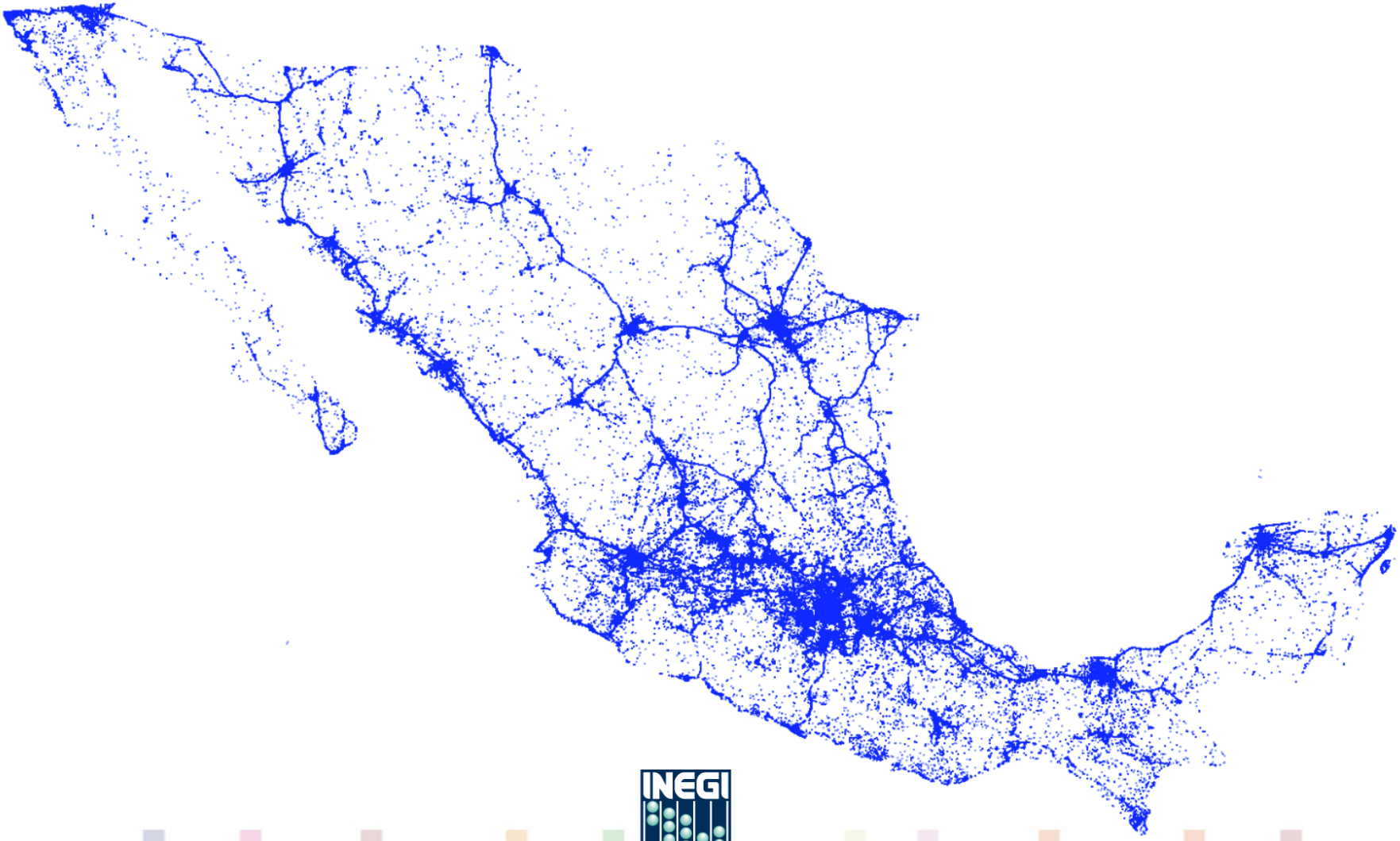
INEGI
INSTITUTO NACIONAL
DE ESTADÍSTICA Y GEOGRAFÍA

150 Millions of Tweets, August 2015



INEGI
INSTITUTO NACIONAL
DE ESTADÍSTICA Y GEOGRAFÍA

70+ Millions of Tweets
August 2015



INEGI
INSTITUTO NACIONAL
DE ESTADÍSTICA Y GEOGRAFÍA

MOBILITY

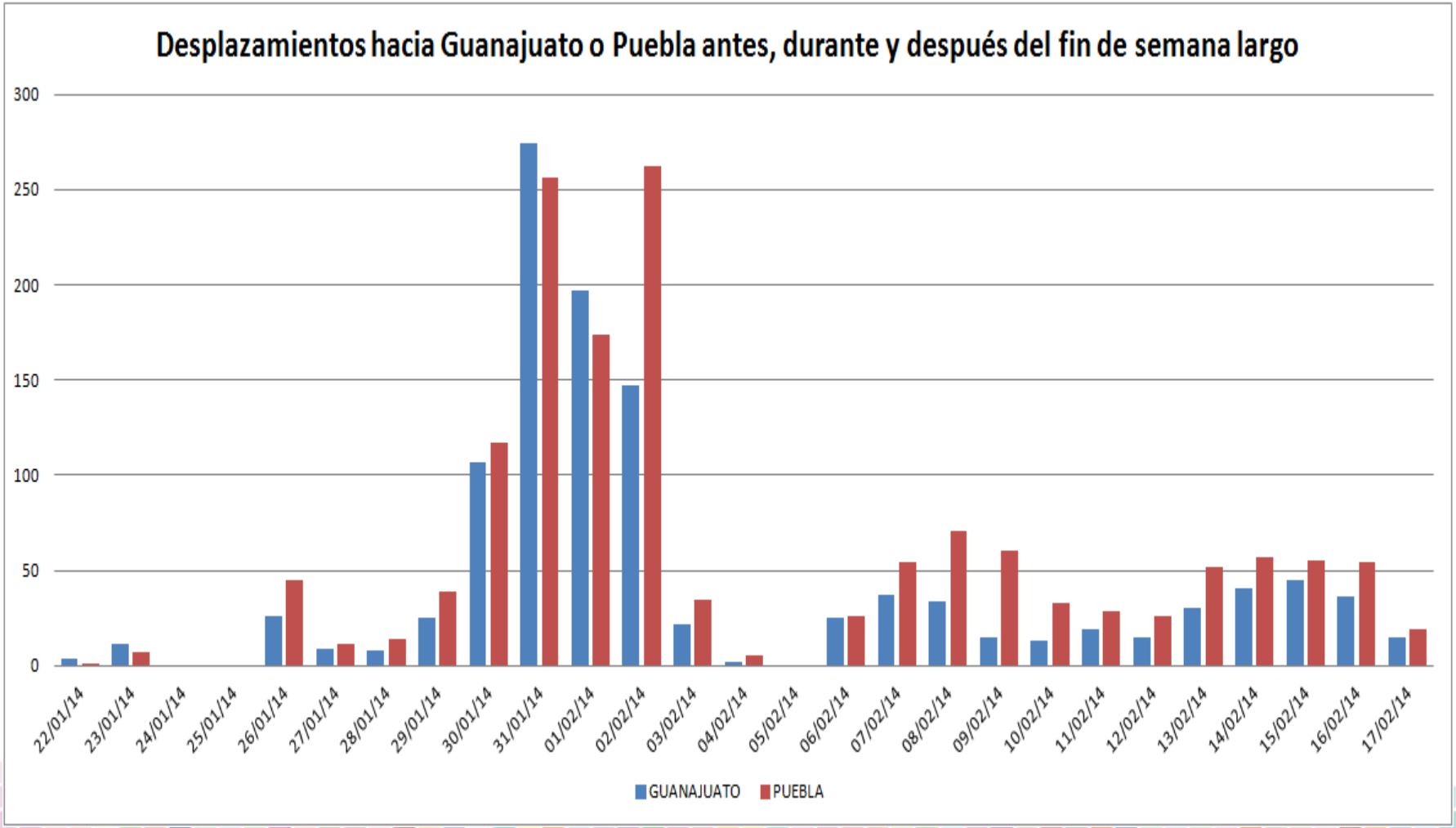
INEGI
INSTITUTO NACIONAL
DE ESTADÍSTICA Y GEOGRAFÍA

Tweeter Users Mobility

4’469,550 displacements
347,157 Tweeter users

INEGI
INSTITUTO NACIONAL
DE ESTADÍSTICA Y GEOGRAFÍA

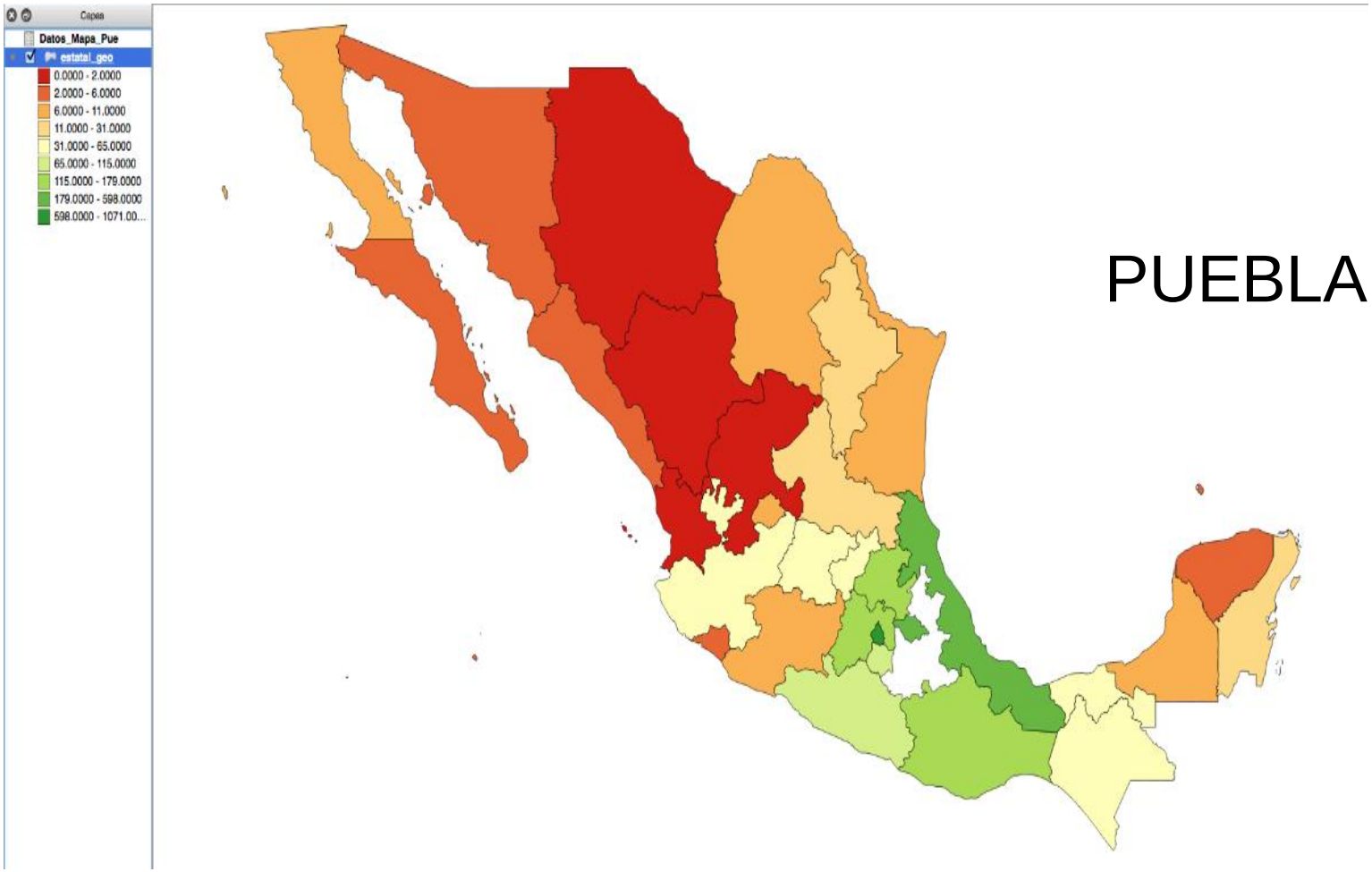
Tweets behavior on a Long-Weekend



Use of Twitter Data for Tourism Studies

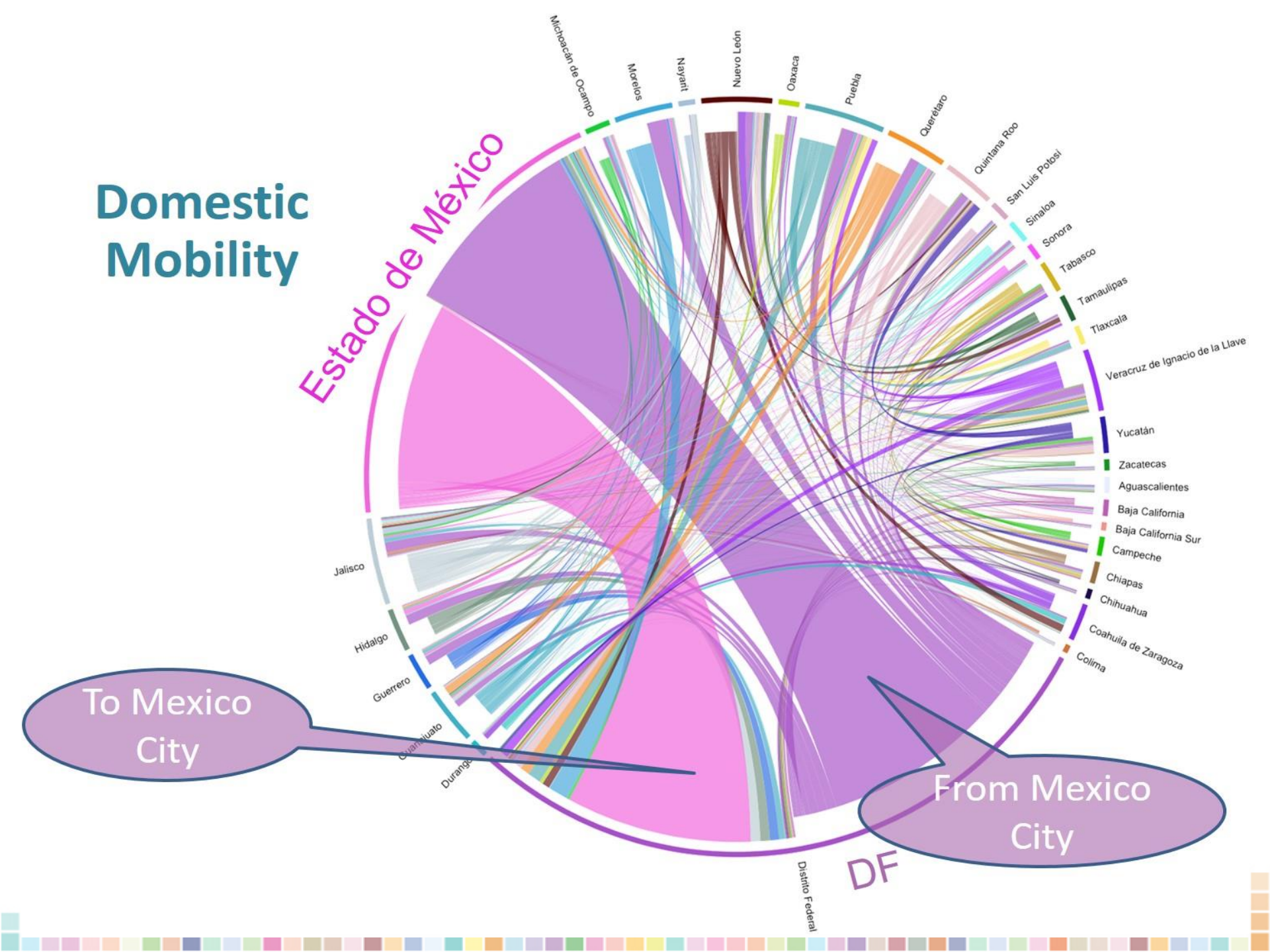
SECTUR
SECRETARÍA DE TURISMO

INEGI
INSTITUTO NACIONAL
DE ESTADÍSTICA Y GEOGRAFÍA


Capas
Datos_Mapa_Pue
estatal_geo
0.0000 - 2.0000
2.0000 - 6.0000
6.0000 - 11.0000
11.0000 - 31.0000
31.0000 - 65.0000
65.0000 - 115.0000
115.0000 - 179.0000
179.0000 - 598.0000
598.0000 - 1071.00...

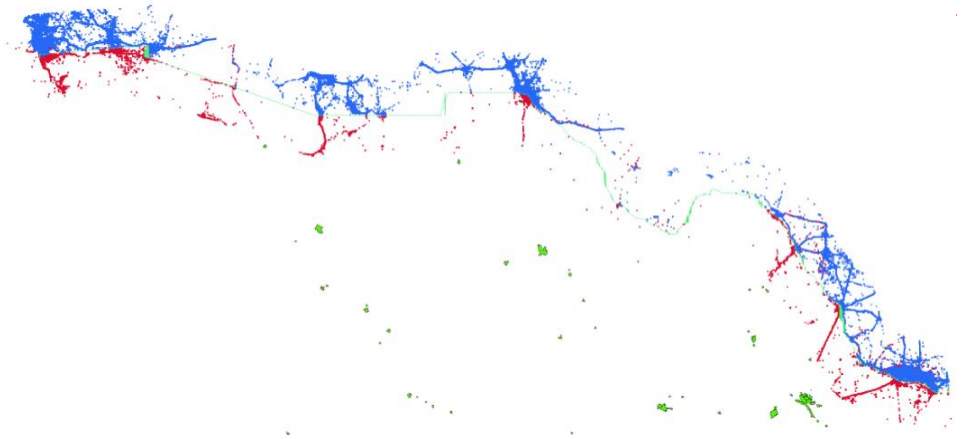
PUEBLA

Domestic Mobility

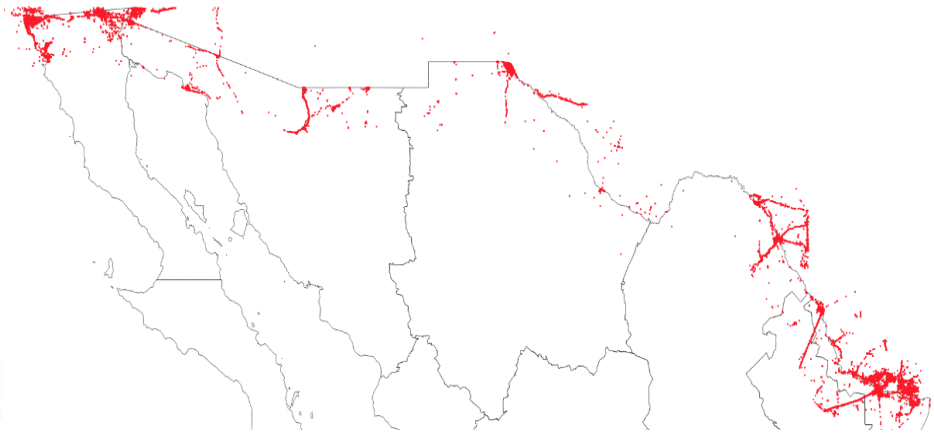


Border Mobility

Research for development of analytical method to measure trans-border mobility trough Geo-referenced tweets.

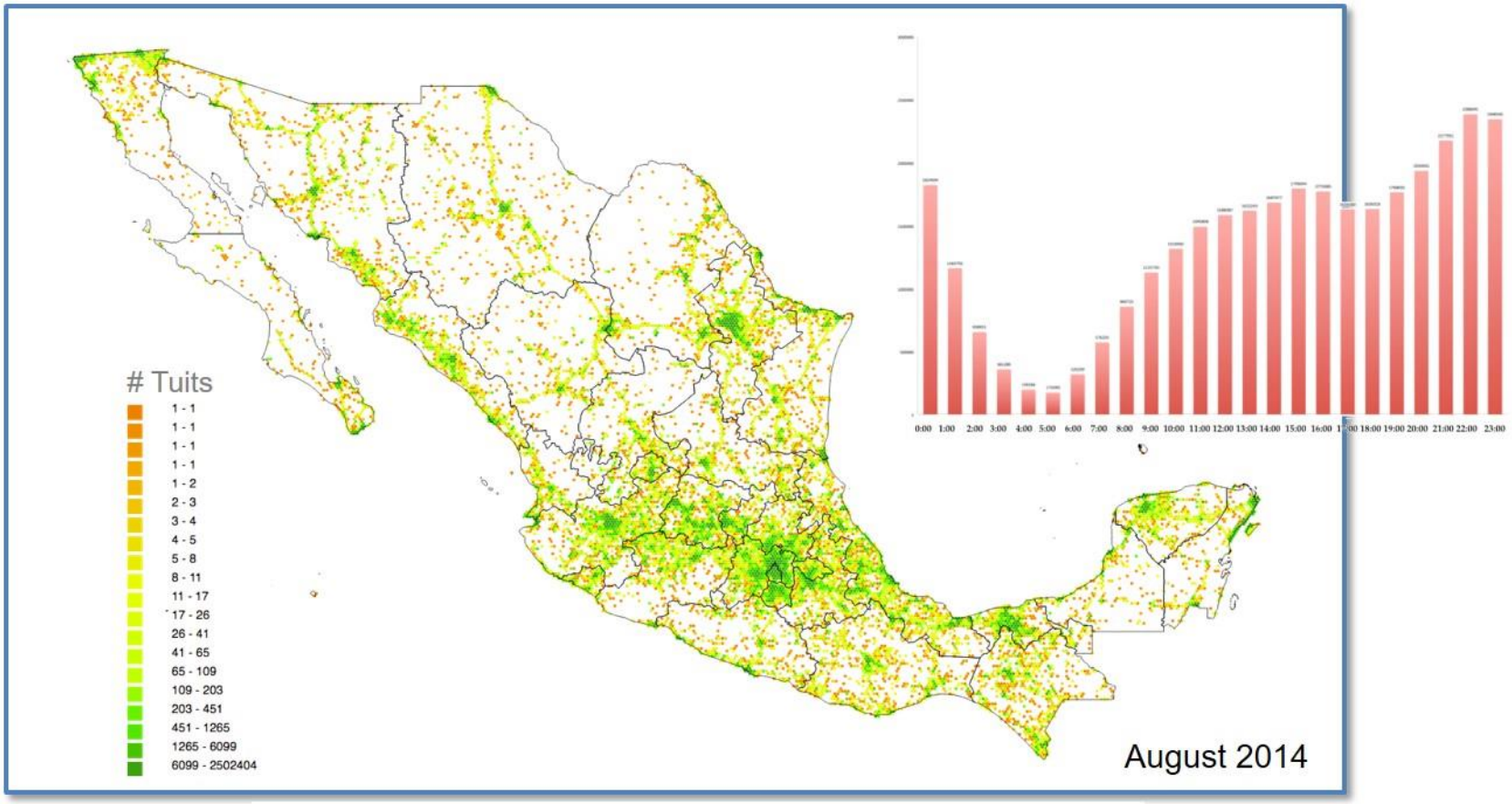


Mexican (red) and US(blue) tweets



Mexican tweets

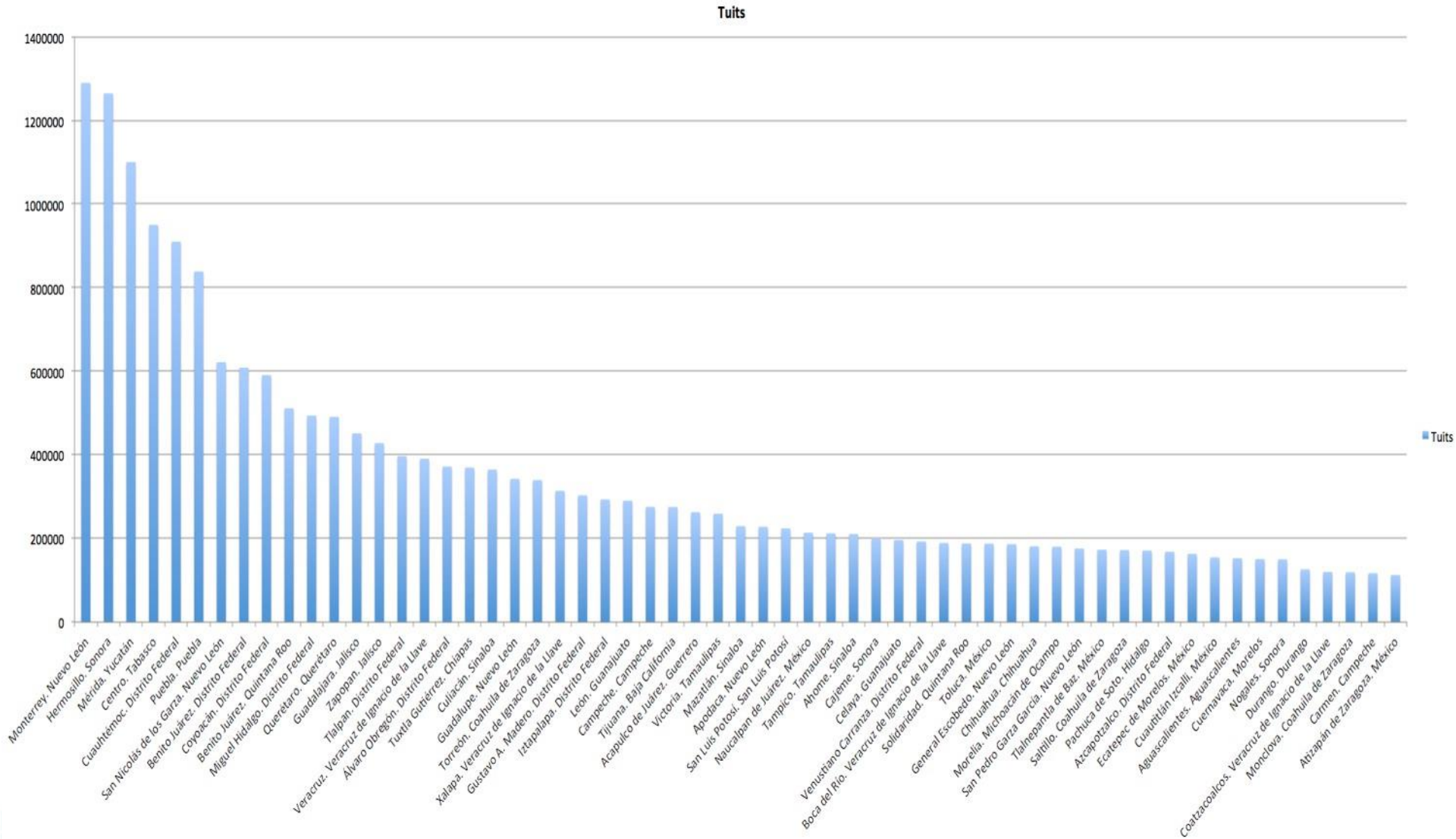
Frequency of Tweet Generation in the country



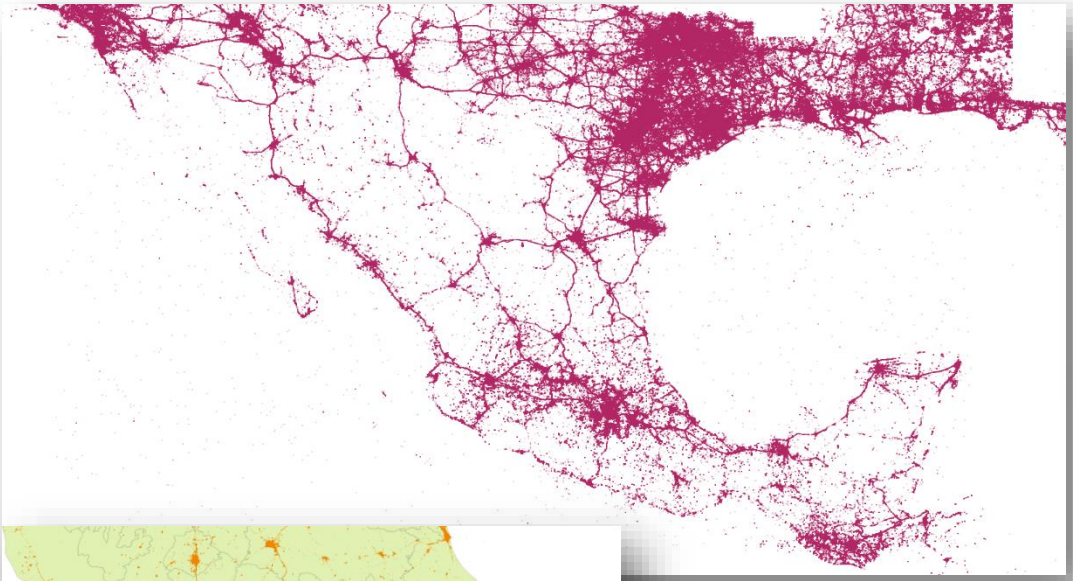
882,007 Twitter Users

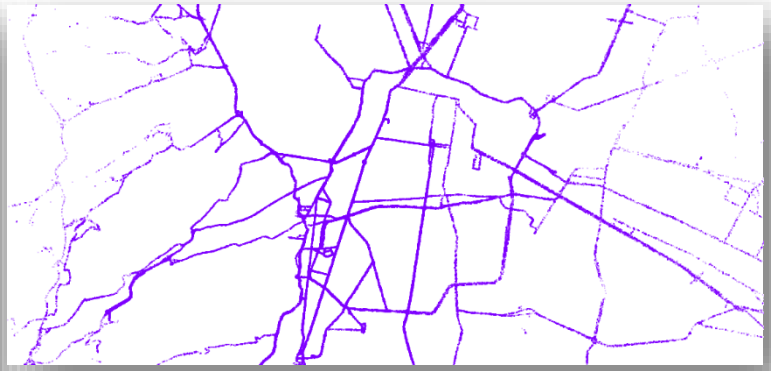
43’079,312 Geo-Referenced Tweets

Geo-referenced Tweets in the Municipalities



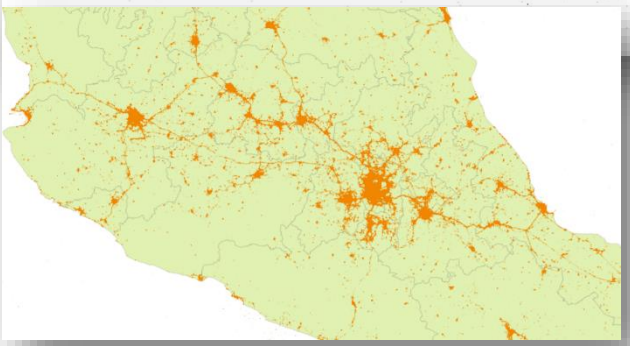
Use of the Mexican Roads Network





121 millions of geo-referenced tweets

January 2015



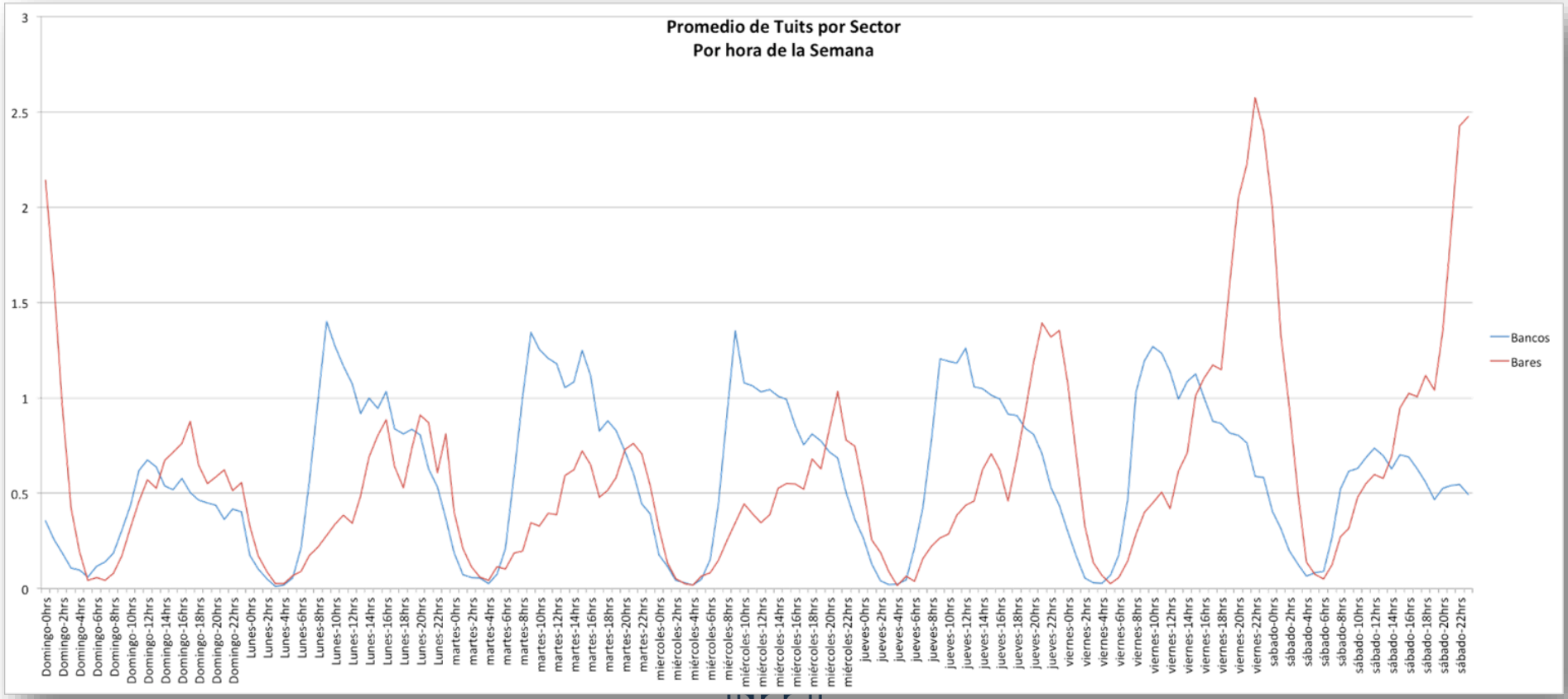
70 millions of geo-referenced tweets

October 2014

INEGI

INSTITUTO NACIONAL DE ESTADÍSTICA Y GEOGRAFÍA

Average Tweets on Banks and Bars in a Week



INSTITUTO NACIONAL
DE ESTADÍSTICA Y GEOGRAFÍA

Next studies on mobility…

- “Feria Nacional de San Marcos” visitants
- Internal tourism in all the country
- Mobility in our Borderlines
- Urban-Rural Systems

SUBJECTIVE WELLNESS

Project Objective

Generation of experimental indicators, new or complementary to traditional methods using data science technologies for the extraction, storage, processing and visualization of big data.

INEGI
INSTITUTO NACIONAL
DE ESTADÍSTICA Y GEOGRAFÍA

Expected benefits

- **New indicators obtained from Big Data Sources**
- **Correlation of results with traditional methodologies information**
- **Scientific production**

Process

Inception

- **Supervised Learning Method**
  - Humans put qualifications on a training set

  - The system uses similarities to qualify other tweets

Knowledge gattering…

(5000+ students)

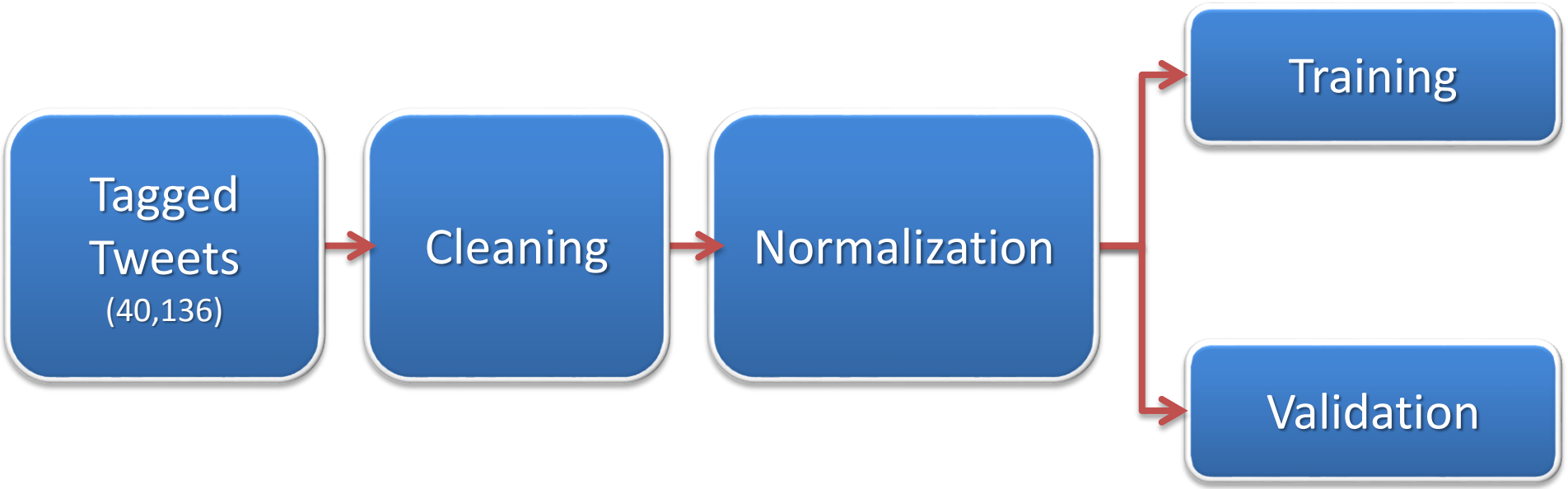
Knowledge

Set of tagged tweets from PioAnálisis (Tweet Analysis)

- 5000+ volunteers
- 374 000 tagged tweets
- 40 000 different tagged tweets (each tweet have been revised 9 times on average)

Automatized Analysis and Qualification

With the manually tagged set of tweets we built a training set to teach the system to recognize and use similarities to qualify other tweets


Tagged Tweets (40,136)
Cleaning
Normalization
Training
Validation

Cleaning process

Information Pre-Processing (normalization)

Word correction
(dictionary, statistics and heuristics)

rojooooooo → rojo

jajaja

Coding
(diacritic symbols)

caña → na

perdón → perdon

amig → am

Lematizing
(FreeLing)

rompí → romper

llegó → llegar

subió → subir

Polarity of Emoticons
(tag of polarity)

+1 0 -1

Q-Grams
(q=4)

romper → romp

perder → perd

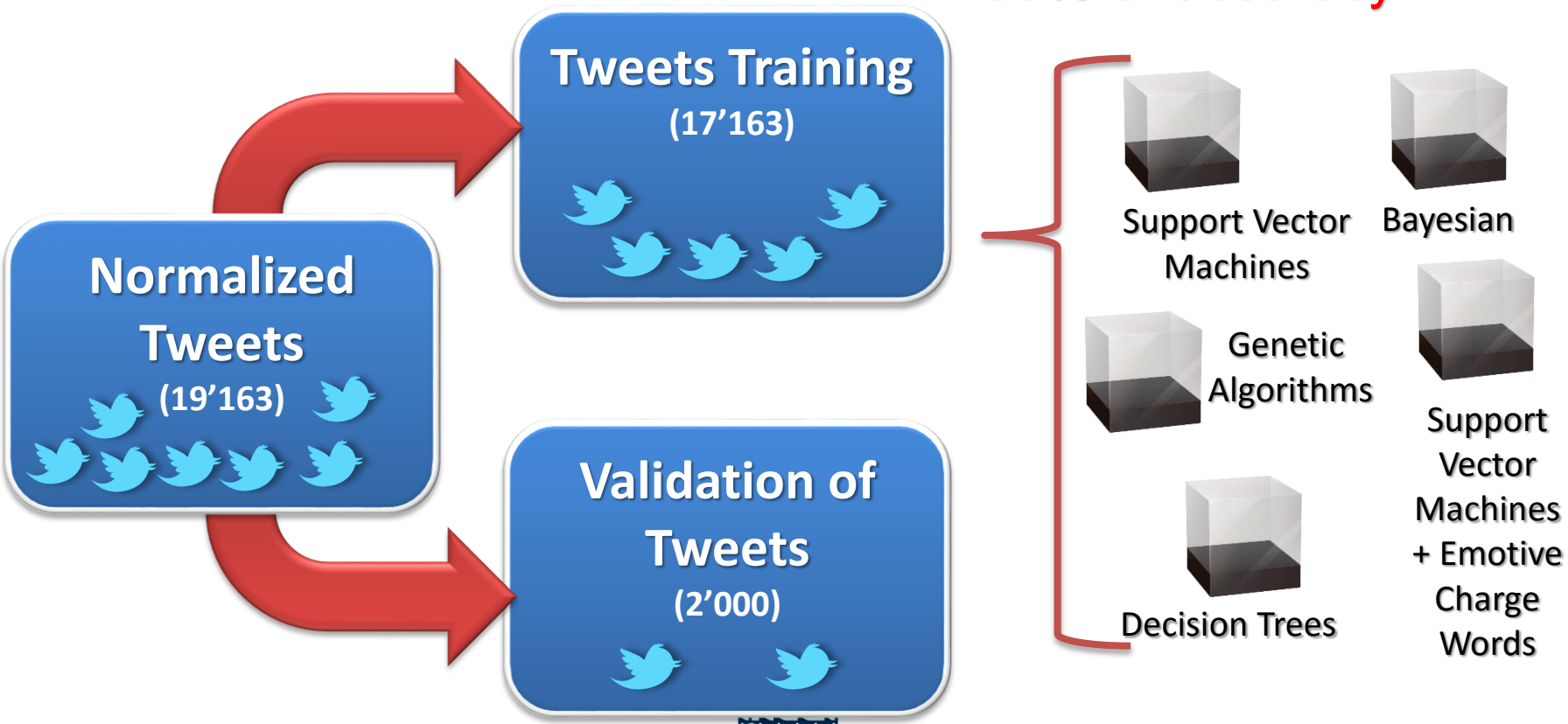
amigs → amig

Filtering
(Suppressing of stop words)

adverbs
interjections
verbs
adjectives
nouns
hashtags

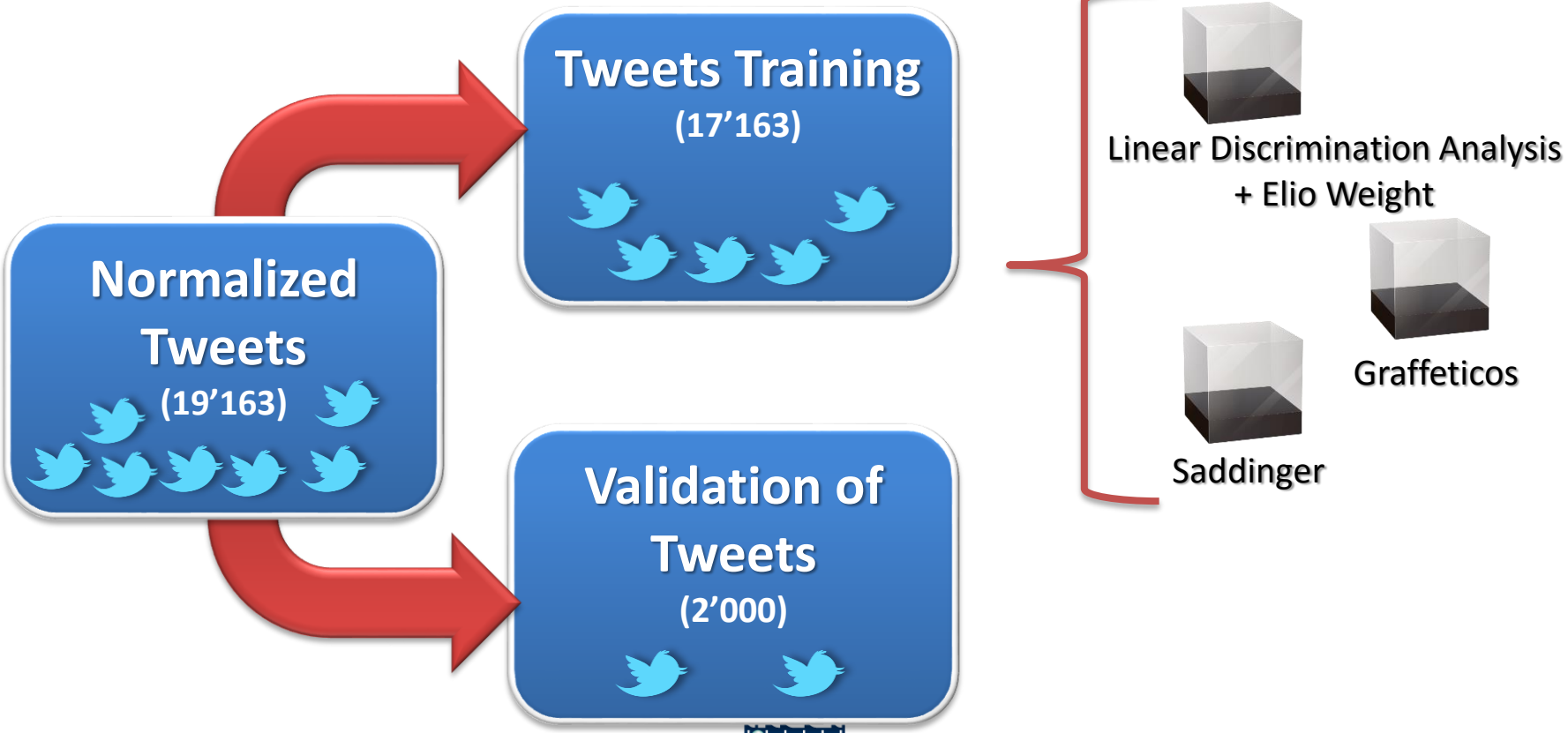
Training with algorithms from the state of the art

≈ 50% of accuracy



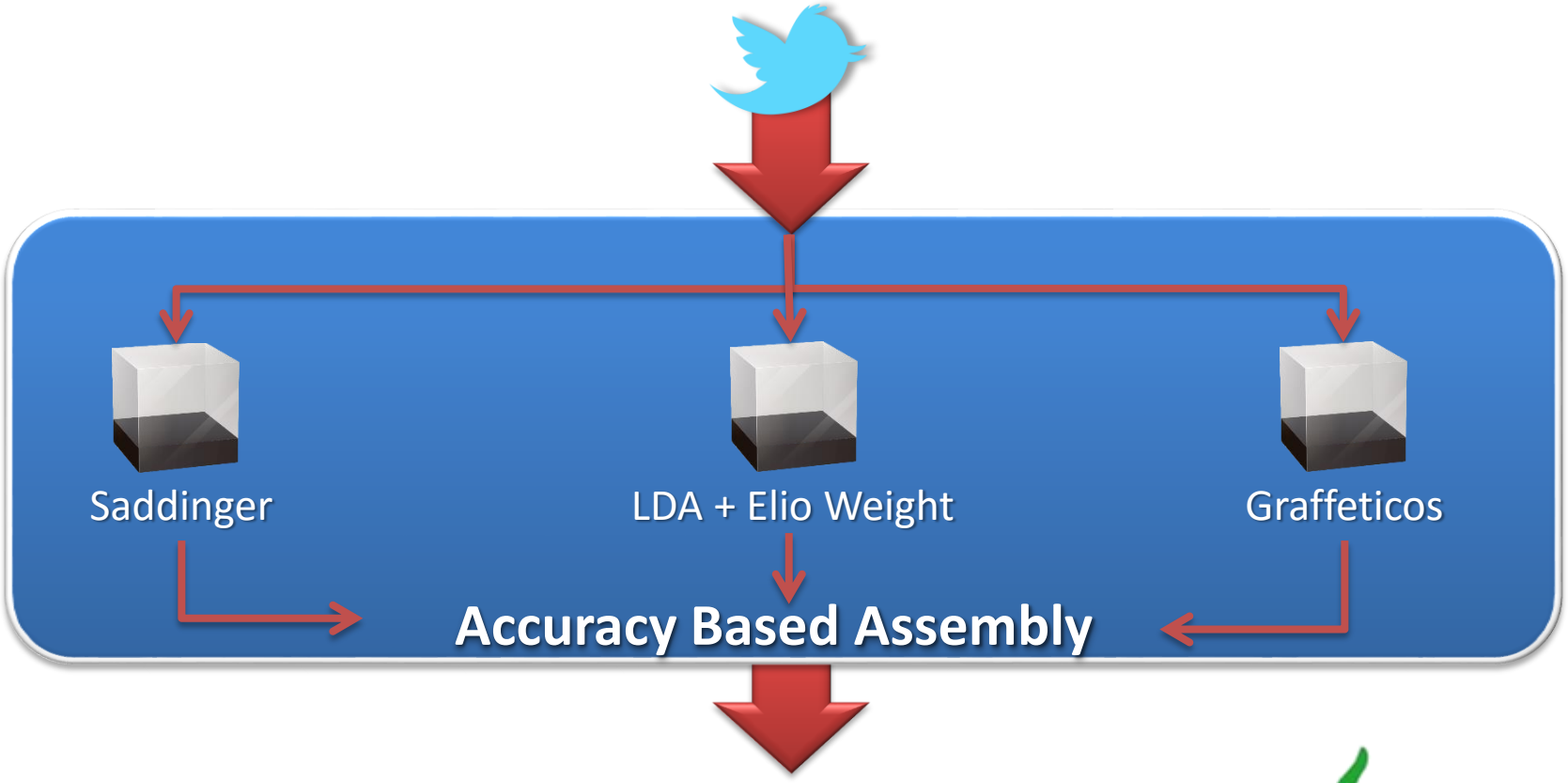
Development of new Classification Algorithms

≈ 70% of accuracy



INSTITUTO NACIONAL DE ESTADÍSTICA Y GEOGRAFÍA

Classification Algorithms Assembly to Improve Accuracy

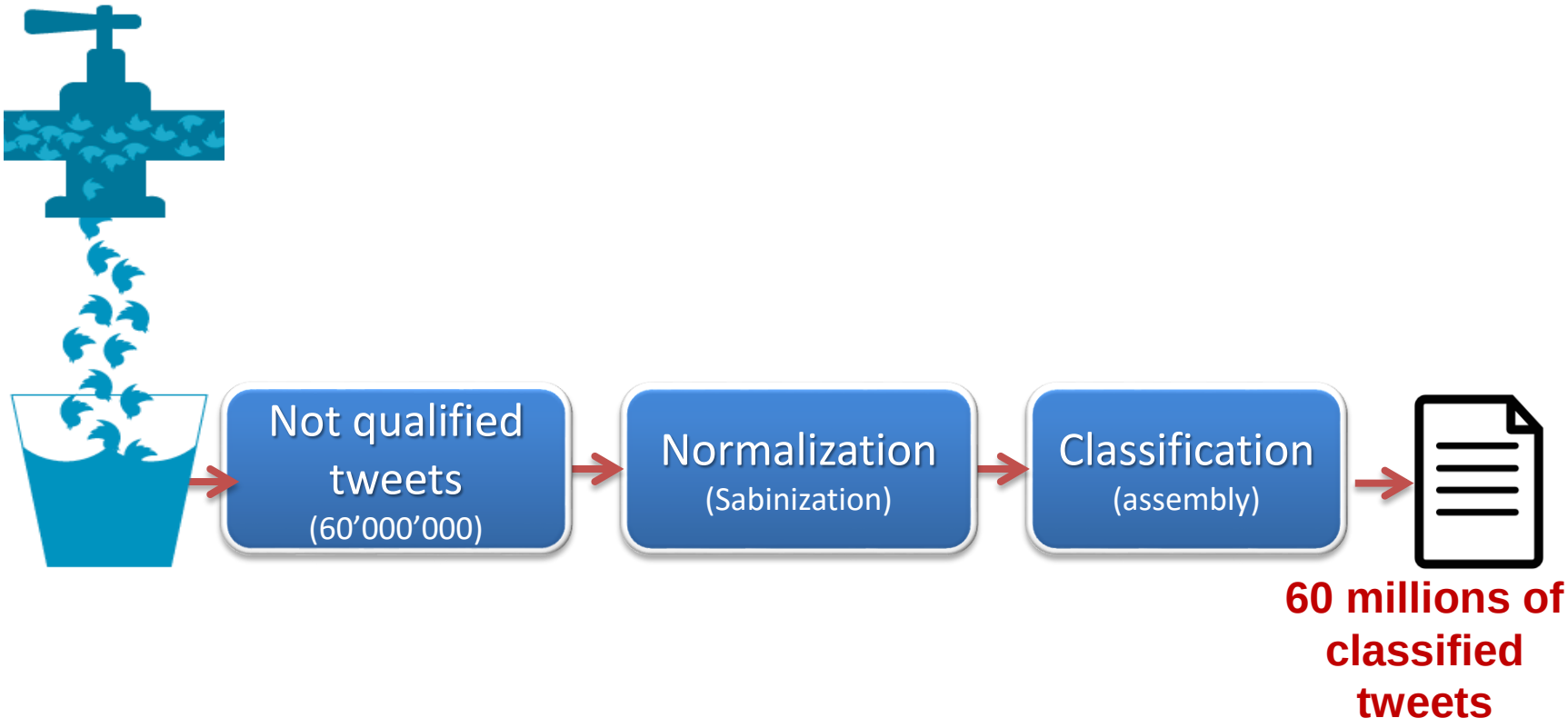


≈ 80% of accuracy

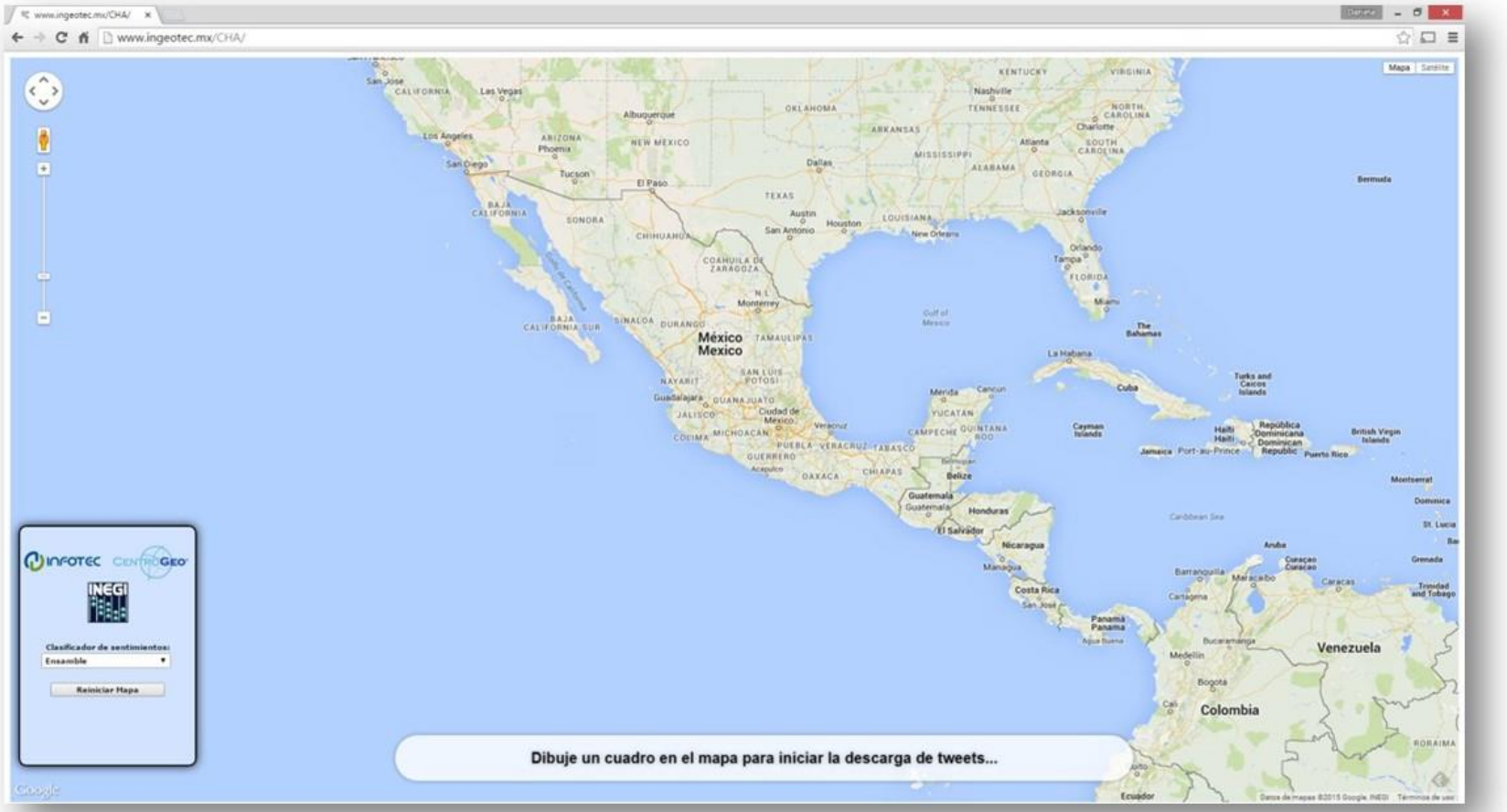
INEGI
INSTITUTO NACIONAL DE ESTADÍSTICA Y GEOGRAFÍA

Automatic classification of other tweets

Classification of 60 millions of tweets with the Assembly



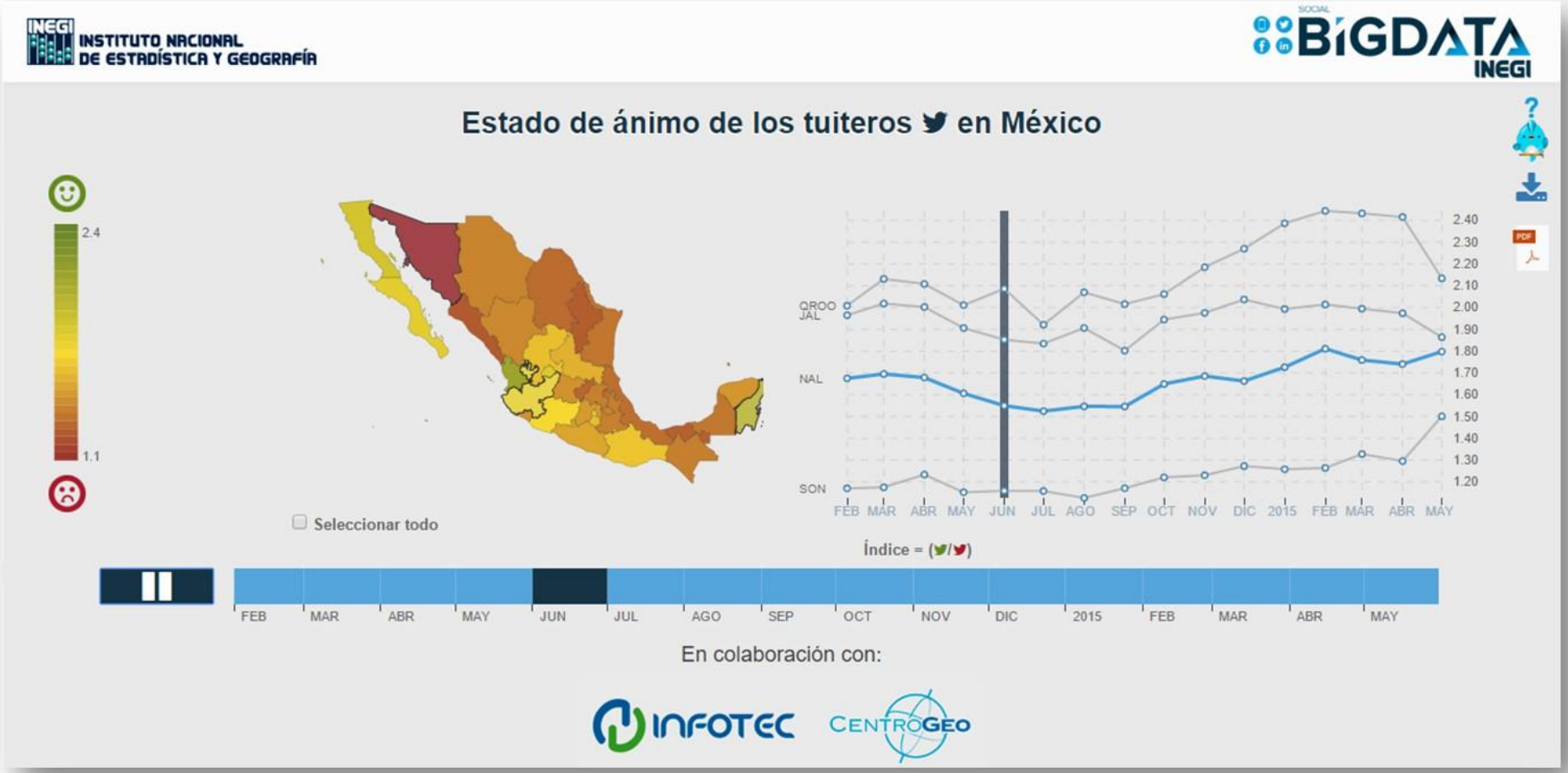
On line demo: www.ingeotec.mx



INSTITUTO NACIONAL
DE ESTADÍSTICA Y GEOGRAFÍA

Visualization

http://cienciadedatos.inegi.org.mx/animotuitero/index.html



High Level Meeting
INEGI, INFOTEC, Centro-GEO, CIDE and CIMAT

Results: Collaboration INEGI, INFOTEC, Centro-GEO, CIDE y CIMAT

- Collaboration lines:
  - Common research
  - Seminars
  - Internships
  - academic programs
  - Spaces exchange
  - Micro data access

NEW PROJECTS

Social Networks monitoring for INEGI

Goal: Publication, following and monitoring of social networks

- Evaluation of new tools: Semantic Web Builder developed by INFOTEC
- Living Lab workshop: Dissemination staff
- Implementing on internal environment

Mental Health of Teenagers (Data2X)

Objective: “Generation of information about mental health on teenager women in Mexico from Tweeter messages”

- INEGI-Data2X agreement (Data2X is a ONG supported by UN)

Mental Health of Teenagers (Data2X)

Statistics on Security and Safety

Research on the possibility to use the tweets database to get information about:

- Collection of dues in urban areas
- Information about natural disasters

Conociendo México

01 800 111 46 34

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

Thank you!!

@inegi_informa

INEGI Informa

INSTITUTO NACIONAL
DE ESTADÍSTICA Y GEOGRAFÍA

Juan.Munoz@inegi.org.mx