Sandbox Project On Social Media Data for Sentiment Analysis/Mobility

3rd International Conference on Big Data for Official Statistics
30 Aug – 1 Sep 2016

BigData UN Global Working Group

INEGI
INSTITUTO NACIONAL DE ESTADÍSTICA Y GEOGRAFÍA

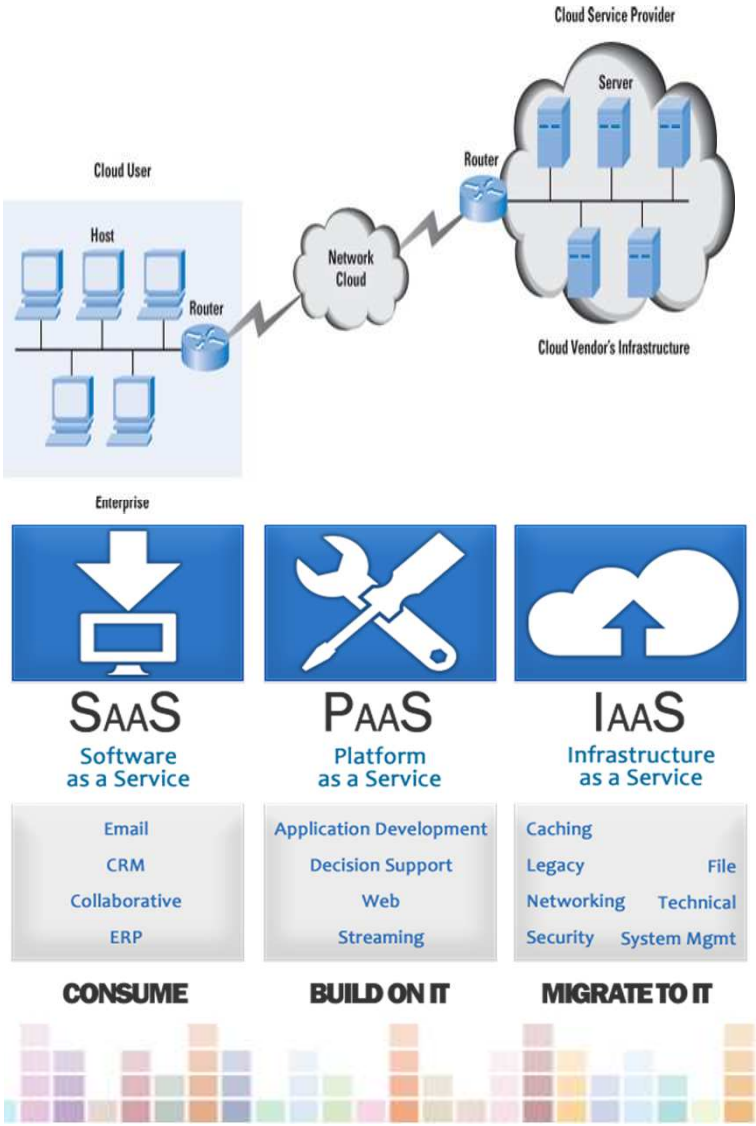
Objective

To share the experience of INEGI in the use of Twitter as a Big Data source and his collaboration in the Sandbox project

Cloud Computing

Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.

US National Institute of Standards (NIST)



Sandbox, since 2014

- Braced by the Commission of European Statisticians (CES)
- Promoted by High Level Group for the Modernization of Official Statistics (HLG-MOS)
- Coordinated by the Statistical Division of the United Nations Economic Commission for Europe (UNECE)
- Implemented in the Irish Centre for High-End Computing

UNECE

ICHEC
Irish Centre for High-End Computing

INEGI
INSTITUTO NACIONAL DE ESTADÍSTICA Y GEOGRAFÍA

Sandbox Initial Aims

- Test feasibility of remote access and processing: Could this approach be used in practice?
- Test whether existing statistical standards / models / methods can be applied to Big Data
- Determine which Big Data software tools are most useful for statistical organisations
- Learn about the potential uses, advantages and disadvantages of Big Data – “learning by doing”.
- Build an international collaboration community on the technical aspects of using Big Data

modern stats

INEGI
INSTITUTO NACIONAL DE ESTADÍSTICA Y GEOGRAFÍA

What can we expect from the Sandbox?

- A state-of the-art shared computing environment, or in other words: a high-end collaborative ICT environment to make experiments
- Hardware:
  - 4 Data/compute nodes, each with 2x10 core Intel Xeon CPUs, 128 GB RAM, 4x4TB disks, 56 Gbit InifiniBand network
  - 2 Service/login nodes , each with 2x10 core Intel Xeon CPUs, 128 GB RAM, 4x4TB disks, 10 Gbit connection to Internet
- Software:
  - Hortonworks Data Platform (Hadoop, Spark, Hive, Pig), R-Studio, RHadoop, ElasticSearch

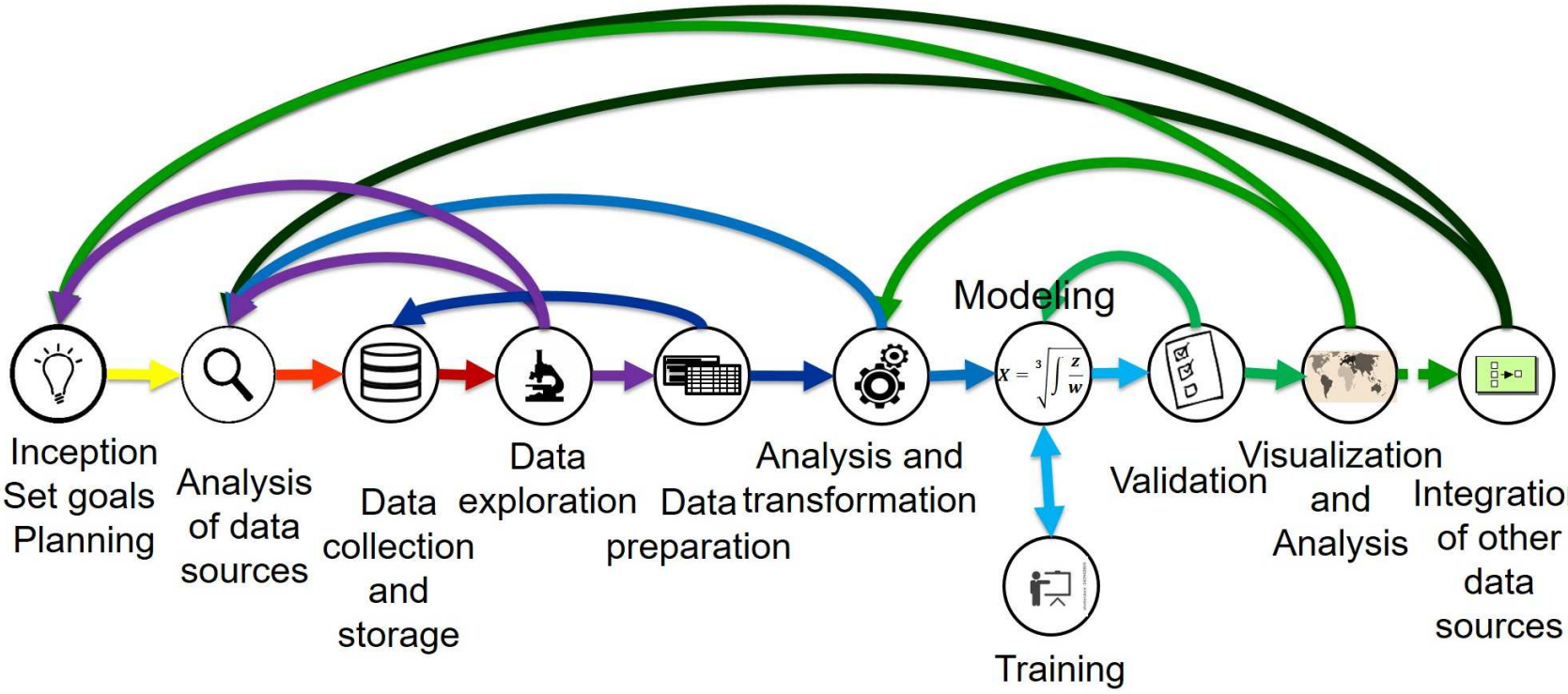
Types of Big Data Sources

- **Meters and smart sensors**. Traffic cameras, GPS devices, power consume meters, IoT, smartwatches, smartphones, etc.
- **Social interactions**. Conversations and publications on social networks like Twitter, Facebook, FourSquare, etc.
- **Business transactions**. Credit cards movements, scanned data, cell phone records, etc.
- **Electronic files**. Documents which are available in electronic formats such as PDF files, websites, videos, audio, images, photos, etc.
- **Broadcast media**. Digital video and audio streamed on real-time

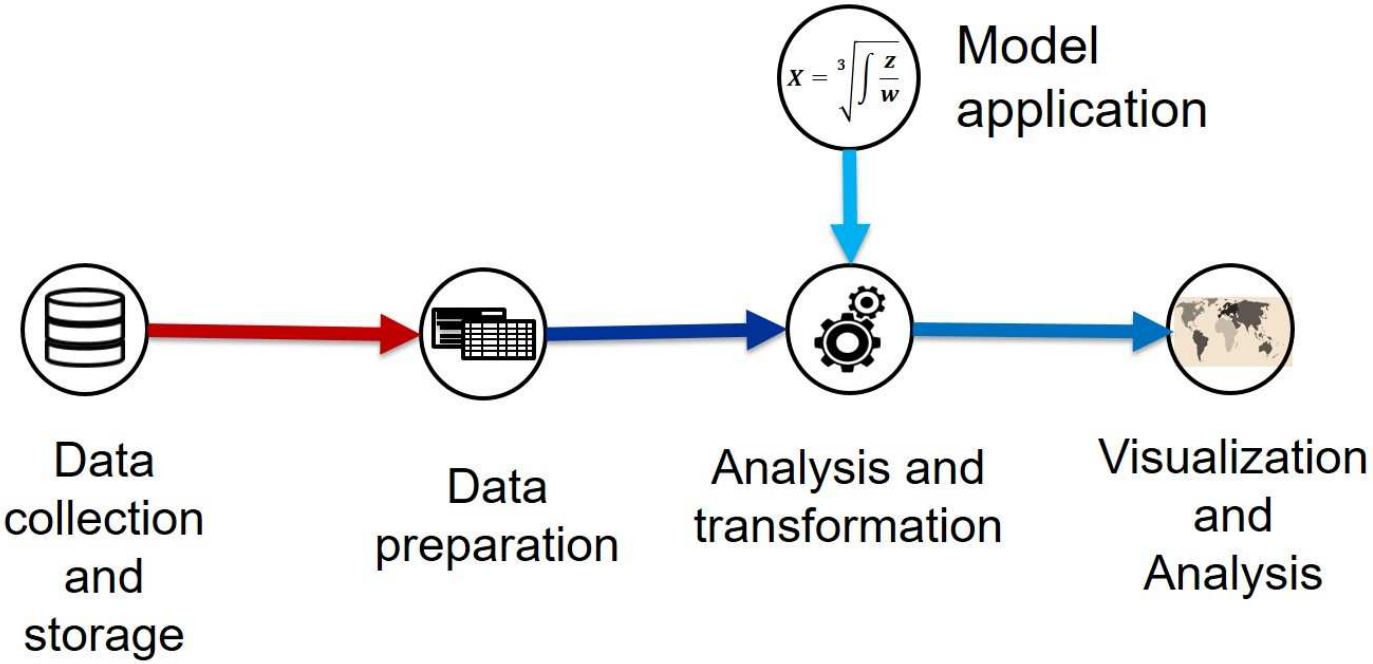
Types of Big Data Sources

- **Meters and smart sensors**. Traffic cameras, GPS devices, power consume meters, IoT, smartwatches, smartphones, etc.
- **Social interactions**. Conversations and publications on social networks like Twitter, Facebook, FourSquare, etc.
- **Business transactions**. Credit cards movements, 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 streamed on real-time

Process Followed by INEGI (until now)



Production process



Study Case

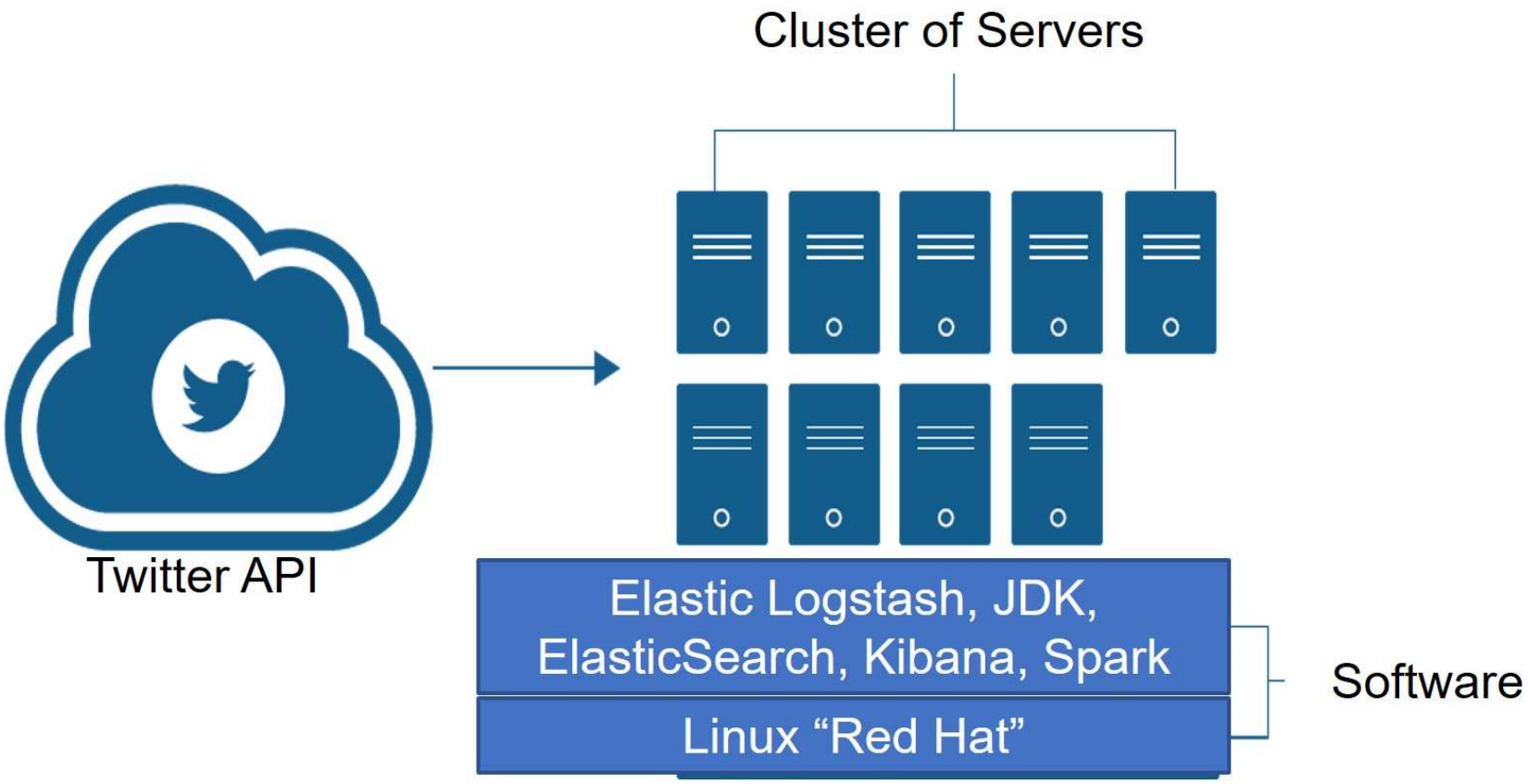
- Initial objective of INEGI's Big Data Project: To generate experimental indicators using Big Data techniques with social media data, to complement statistical information obtained from traditional methods and sources.
- Initial Goal: To obtain indicators of subjective wellbeing from social media data sources.

Why did we choose Twitter as a data source?

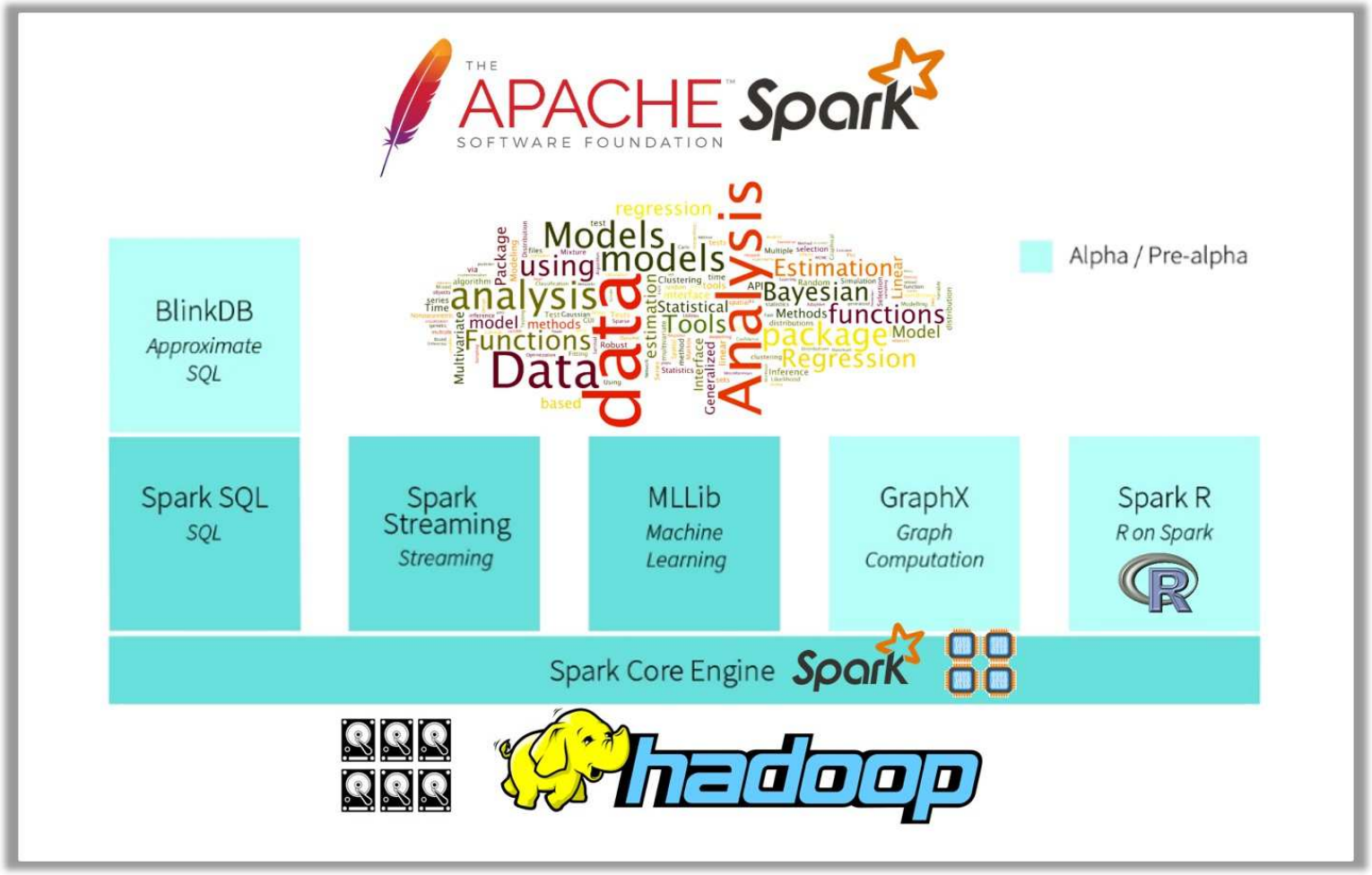
- It's a widely adopted social network where you can find content written by common people
- Tweets are public, so we can use them without concerns about privacy
- There is a free API which allows to get 1% of the tweets that are being produced on real time (https://dev.twitter.com/rest/public)

INEGI
INSTITUTO NACIONAL DE ESTADÍSTICA Y GEOGRAFÍA

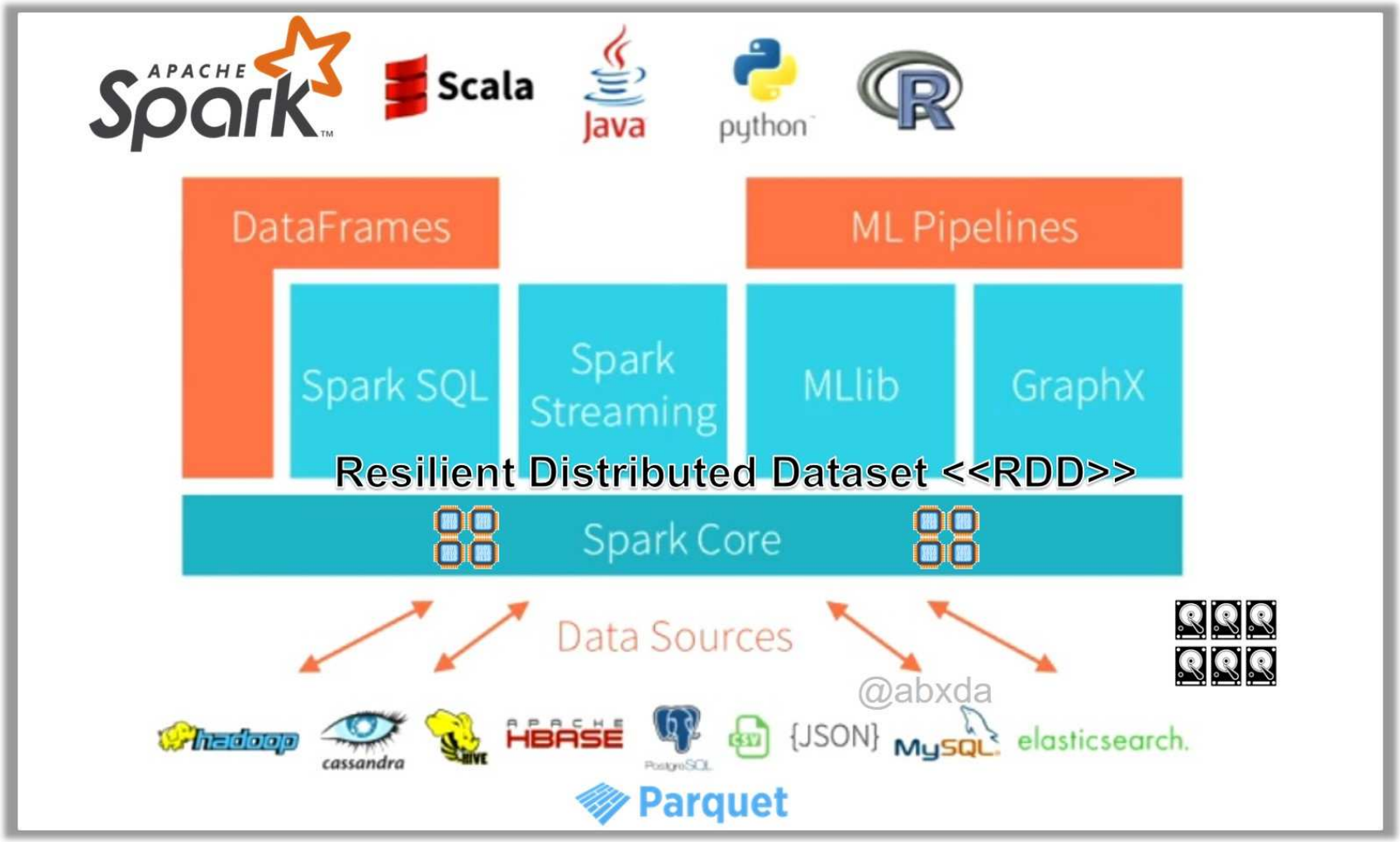
Collection infrastructure



Software Stack (2013)

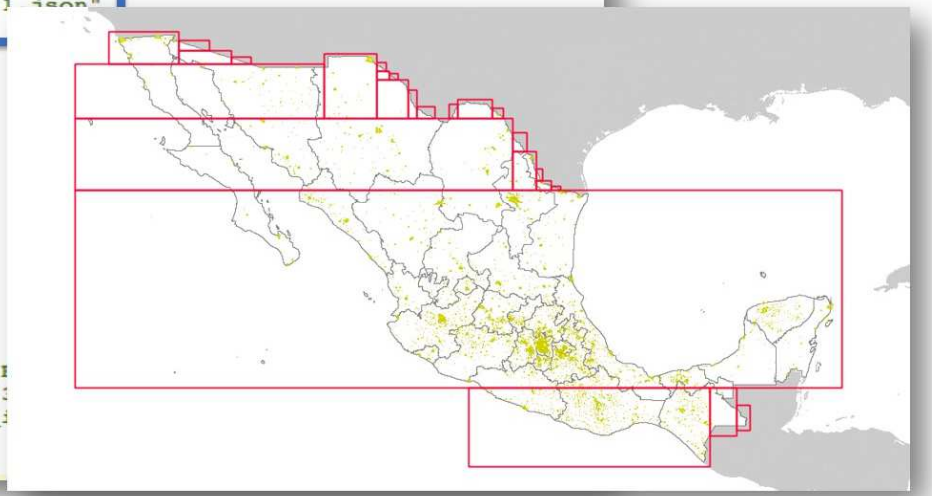


Software Stack (2016)



Tweet Structure


```
{
  "text": Todo listo para la presentación de #BigData en el @FSLmx
  "created_at": "2014-02-21T17:00:50.000Z",
  "truncated": false,
  "mention": [],
  "retweet_count": 0,
  "hashtag": [],
  "location": {
    "lat": 19.39617897,
    "lon": -99.22636055
  },
  "place": {
    "id": "3ad512d283f67a11",
    "name": "Aguascalientes",
    "type": "city",
    "full_name": "Aguascalientes, Aguascalientes",
    "street_address": null,
    "country": "México",
    "country_code": "MX",
    "url": "https://api.twitter.com/1.1/geo/id/3ad512d283f67a11.json"
  },
  "link": [
    {
      "url": "http://t.co/AUNXLVSimQ",
      "display_url": "4sq.com/1p1LyUL",
      "expand_url": "http://4sq.com/1p1LyUL",
      "start": 28,
      "end": 50
    }
  ],
  "user": {
    "id": 205760874,
    "name": "Abel Coronado",
    "screen_name": "abxda",
    "location": "",
    "description": "Filósofo,Desarrollador de Software,M.C.en
    "profile_image_url": "http://pbs.twimg.com/profile_images/
    "profile_image_url_https": "https://pbs.twimg.com/profile_
  }
}
```

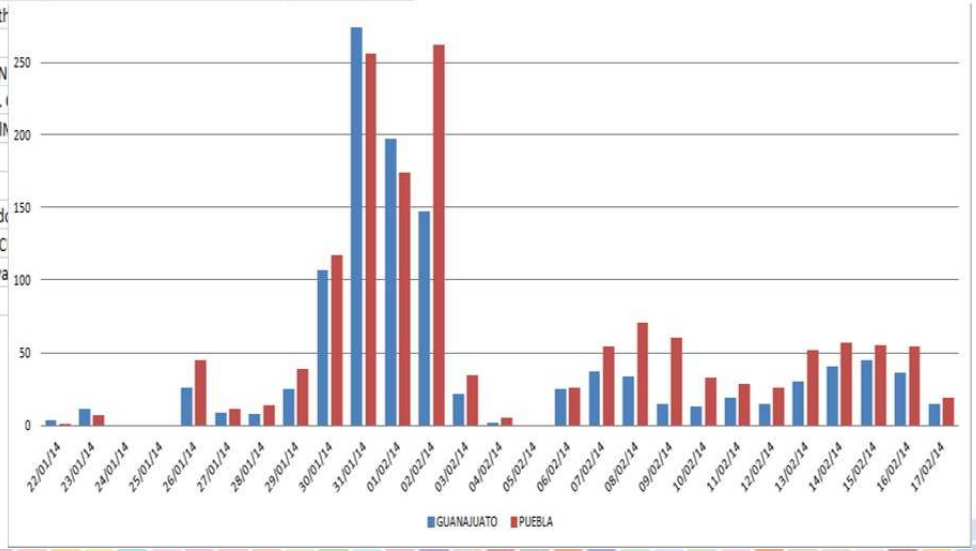



We found that

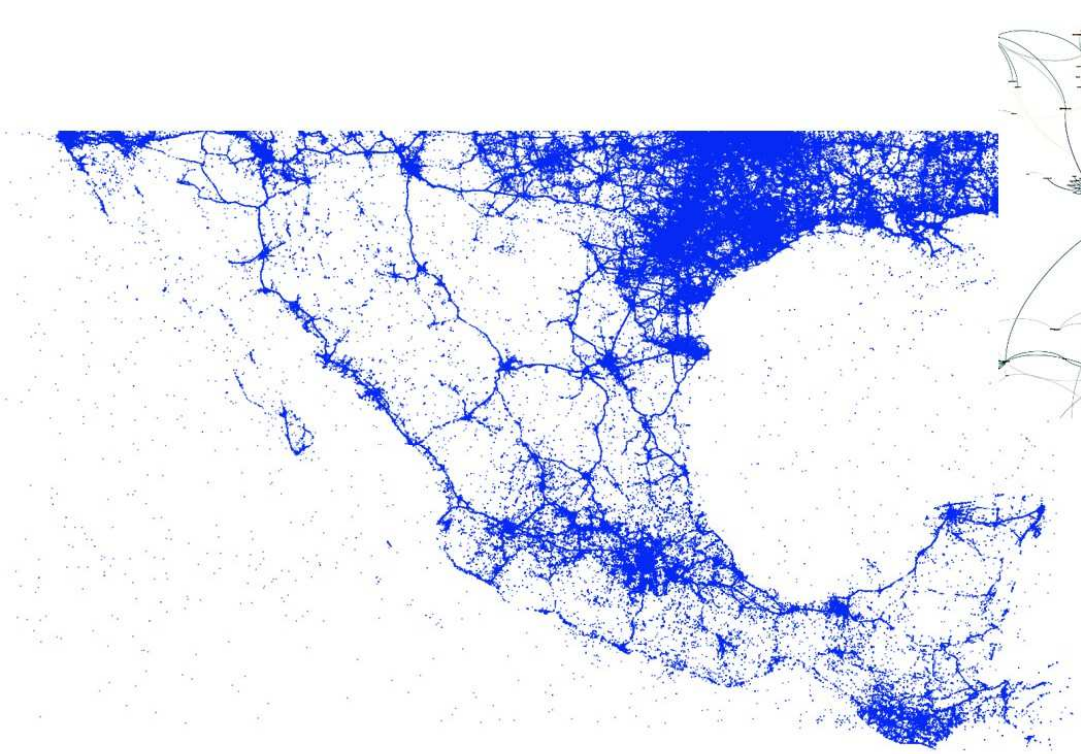
- The JSON structure is easy to process
- The content is text that we can examine to make the sentiment analysis
- Geographical coordinates can be used to filter the tweets and obtain only those of interest (warning: not all the tweets are geo referenced)
- We can make mobility analysis, based in Tweets' location and time

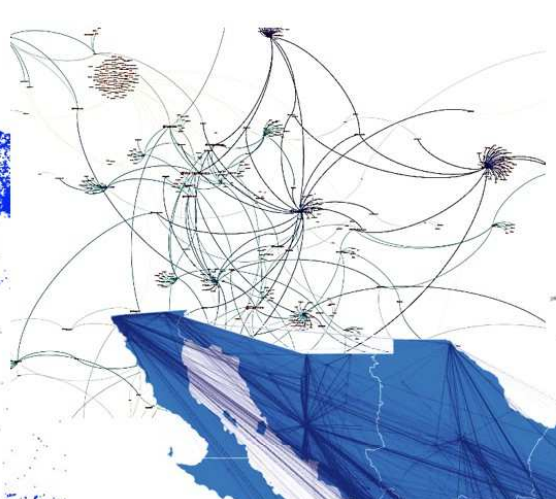
Tweets preparation and analysis

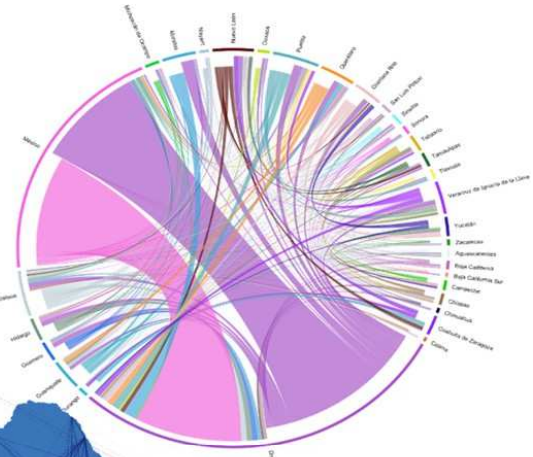
4.3691E+17	ayyekaylla	30.792072	-86.67873	US	<a href="htt	2014-02-21T1	Winter in the senior meeting rn ðŸ˜´ðŸ˜´,
4.3691E+17	_OnlyOneJa	31.328173	-89.329953	US	<a href="htt	2014-02-21T1	Kickback in my room tonight to get my mind off the bs
4.3691E+17	ww_cesar_w	19.396179	-99.226361	MX	<a href="htt	2014-02-21T1	I'm at CafÃ© Punta del Cielo http://t.co/AUNXLVSimQ
4.3691E+17	Ericikee	27.391716	-82.477106	US	<a href="htt	2014-02-21T1	*pets* @2chaynez
4.3691E+17	jennyyylove	29.709114	-95.185322	US	<a href="htt	2014-02-21T1	Somebody is eating the chocolates David have Nd I only have 1 left ðŸ˜”"
4.3691E+17	fets	26.073966	-80.147174	US	<a href="htt	2014-02-21T1	Aaagr moeten deplanen omdat we niet kunnen landen in philly.. :-(minstens 3 uur delay...
4.3691E+17	brettwbailey	30.109563	-91.942463	US	<a href="htt	2014-02-21T1	@ShotsPrincess you gotta be like @PANDAmonium_07 and go straight cutthroat lol
4.3691E+17	MiOfiPolanc	19.435647	-99.185377	MX	<a href="htt	2014-02-21T1	6 frases que un lÃ-der nunca debe decir: El trabajo de un lÃ-der es complicado; las oportunidades para errar est..
4.3691E+17	kleptolovestory			US	<a href="htt	2014-02-21T1	#TrueDetectiveSeason2 cc @mattfraction @zdarsky http://t.co/5HvJB58YFC
4.3691E+17	maxoyervide	26.211199	-98.21623	US	<a href="htt	2014-02-21T1	See ya school!
4.3691E+17	balehorabue	25.671282	-100.313277	MX	<a href="htt	2014-02-21T1	I'm at Facultad de Comunicacion UMM (Monterrey; Nuevo Leon) http://t.co/0A9zli1dOk
4.3691E+17	YoaniitaFran	21.144447	-86.906771	MX	<a href="htt	2014-02-21T1	Ambre..
4.3691E+17	nicole_rend	24.02279	-104.672112	MX	<a href="htt	2014-02-21T1	I'm at Tribunal Superior de Justicia (Durango; Durango; MÃ©xico) http://t.co/2lBUseY6Yf
4.3691E+17	MikelCass	30.699126	-97.864629	US	<a href="htt	2014-02-21T1	Damn it feels great today ðŸ'ŒðŸŽ£ Besides th
4.3691E+17	alezavaliis	15.466293	-87.996081	MX	<a href="htt	2014-02-21T1	@Alexramosjr teamoooâ™¥
4.3691E+17	carlyy_g	30.199626	-97.773748	US	<a href="htt	2014-02-21T1	@LaurenWarwas @sheila_paolaa uh duhhh. N
4.3691E+17	ElColver			MX	web	2014-02-21T1	@CATEDRAODONNELL: CÃ¡tedra Mexicana G.
4.3691E+17	lrobertonava	23.194018	-106.425074	MX	<a href="htt	2014-02-21T1	I'm at SAT (MazatlÃ¡n; Sinaloa) http://t.co/iolM
4.3691E+17	SheFellForM	32.558248	-97.057816	US	<a href="htt	2014-02-21T1	I can't wait for summer Olympics though !!
4.3691E+17	EhmJayyOrN	30.423385	-91.173801	US	<a href="htt	2014-02-21T1	@mylovekhills @kkinnd what u talking bout
4.3691E+17	_Kenahday_I	32.500404	-86.471853	US	<a href="htt	2014-02-21T1	@MadyHendley ðŸ˜´,ðŸ˜´,ðŸ˜´,ðŸ˜´, no ma'am d
4.3691E+17	HaleyMathe	30.609897	-96.333299	US	<a href="htt	2014-02-21T1	@ColinMathews1 OMG YOUR TWIT PIC IS SO C
4.3691E+17	enlacedelacosta			MX	web	2014-02-21T1	@SriaSegOax @PoliciaEstOax @AlbertoEsteva
4.3691E+17	karisalinas32	25.690377	-100.276162	MX	<a href="htt	2014-02-21T1	Compermisooooo ðŸ˜´—ðŸš¬



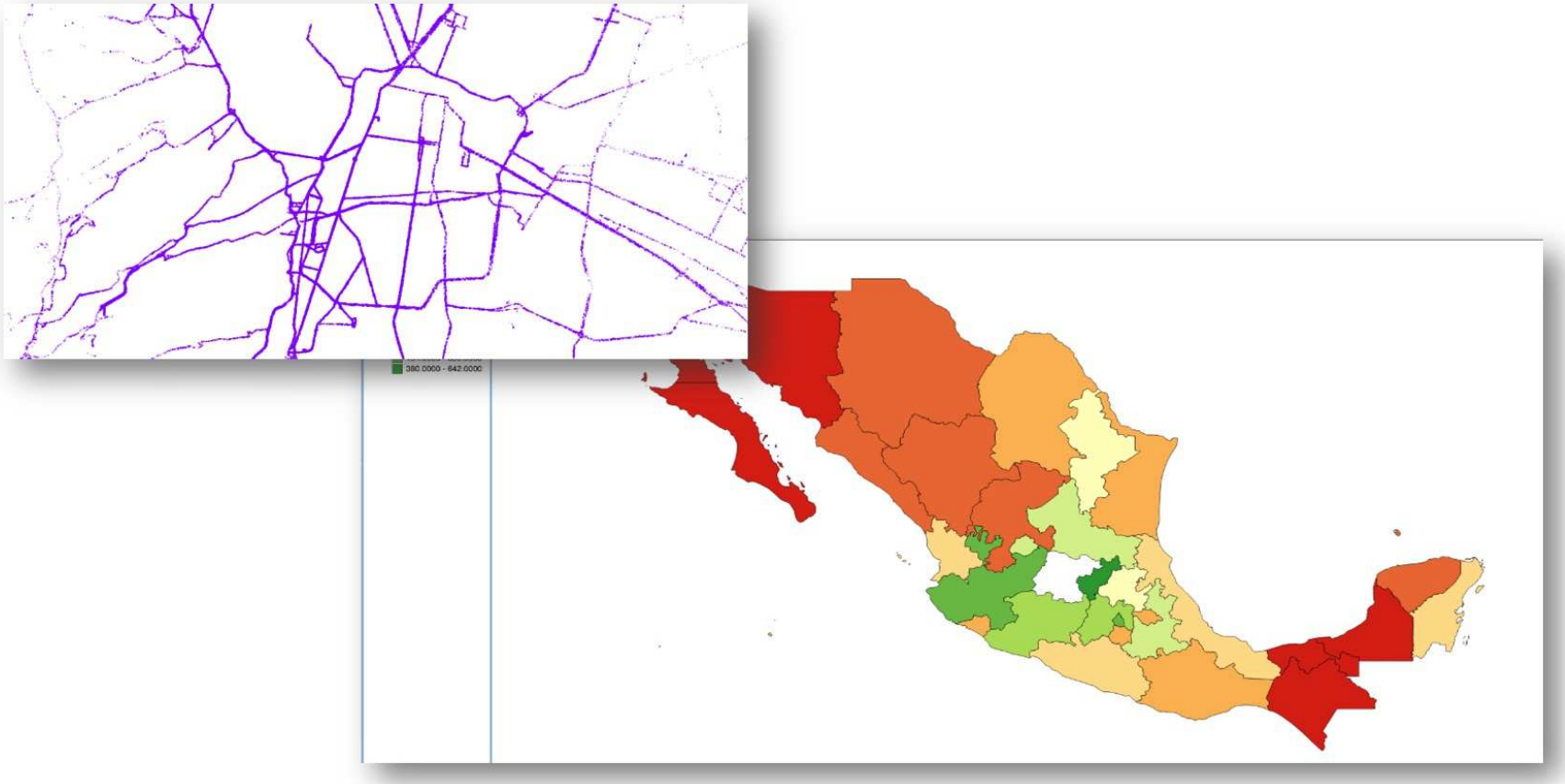
Modeling and Validation





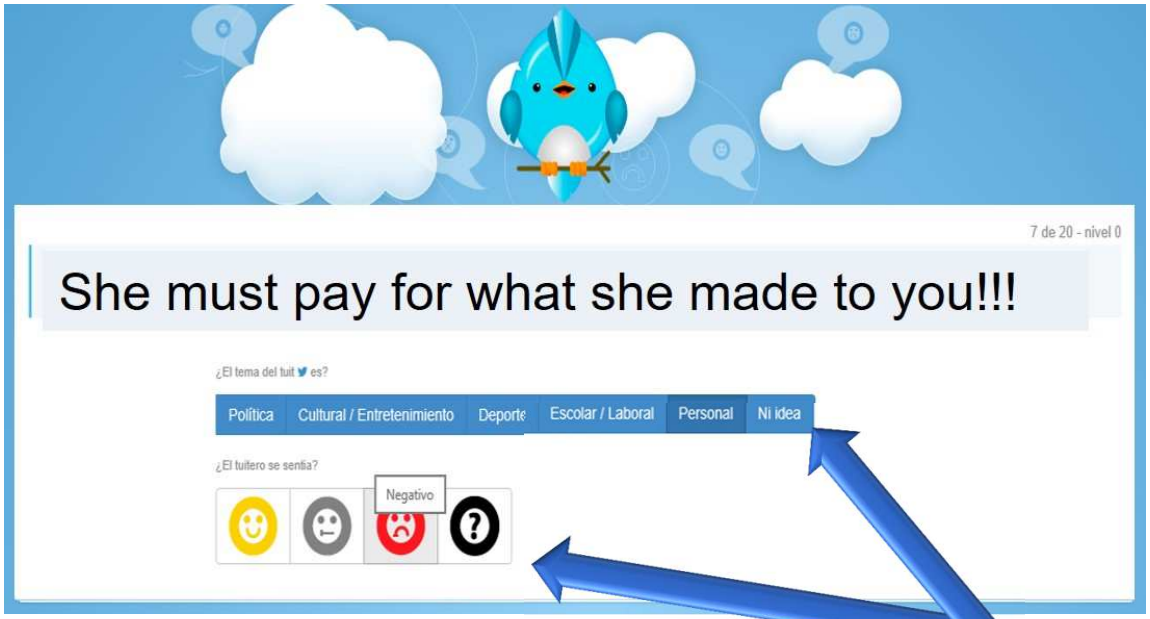


Visualization

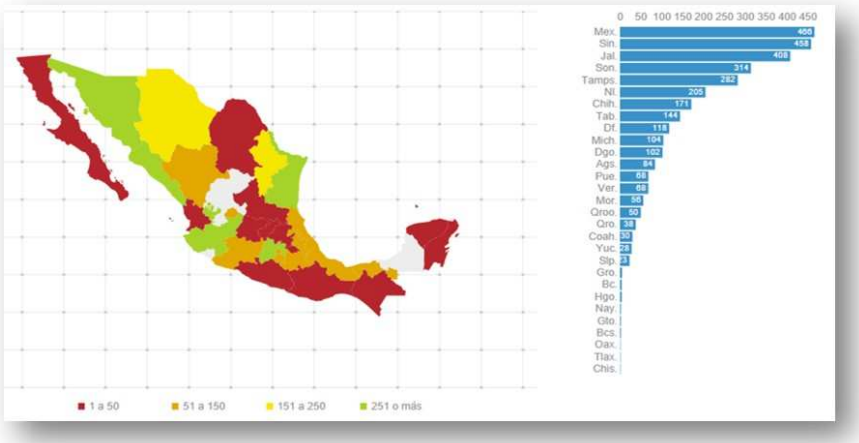


Supervised Training

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



Manual Tagging



5000 people (TecMilenio students), 100 tweets tagged by each one, each tweet was tagged nine times, about 40,000 different tagged tweets, interpreted accordingly to regional idioms

INSTITUTO NACIONAL DE ESTADÍSTICA Y GEOGRAFÍA

Training and modeling

- Latent Dirichlet Allocation
- Decision Trees
- Saddinger, Graffetics, Elio Weight

Methods assembly

Tweets sample

Manual tagging

Cleaning and normalization

Mathematical modelling

http://scikit-learn.org/
http://www.r-project.org/

Machine Learning Methods

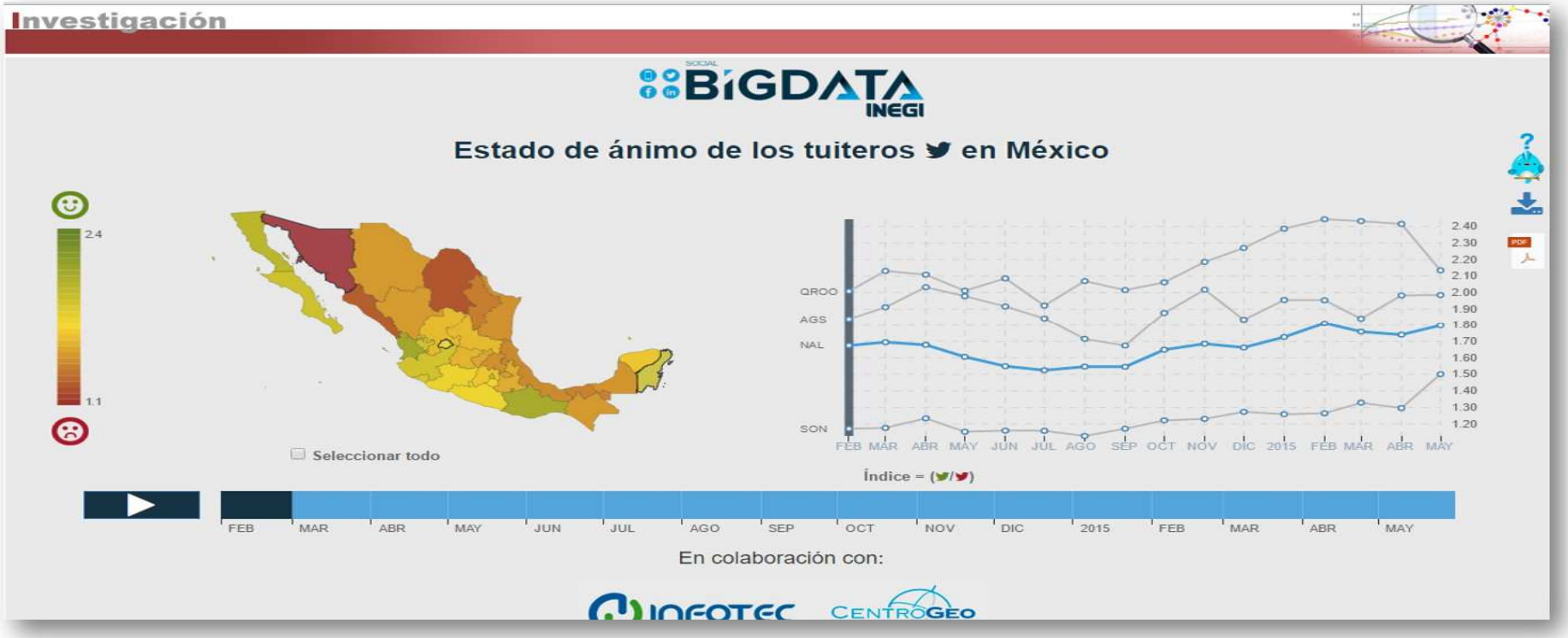
All other collected tweets

Classification model

Training set + Methods assembly

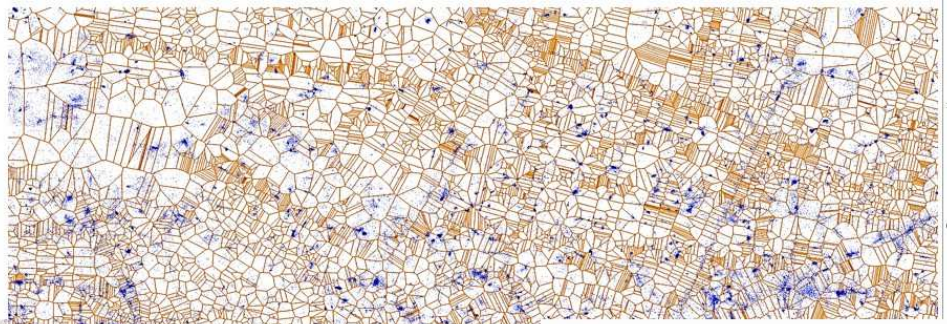
Sentiment indicator

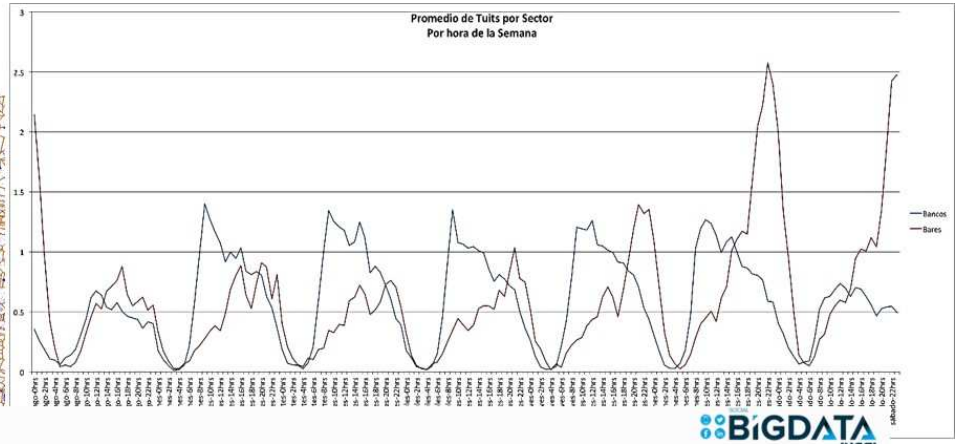
Sentiment Visualization

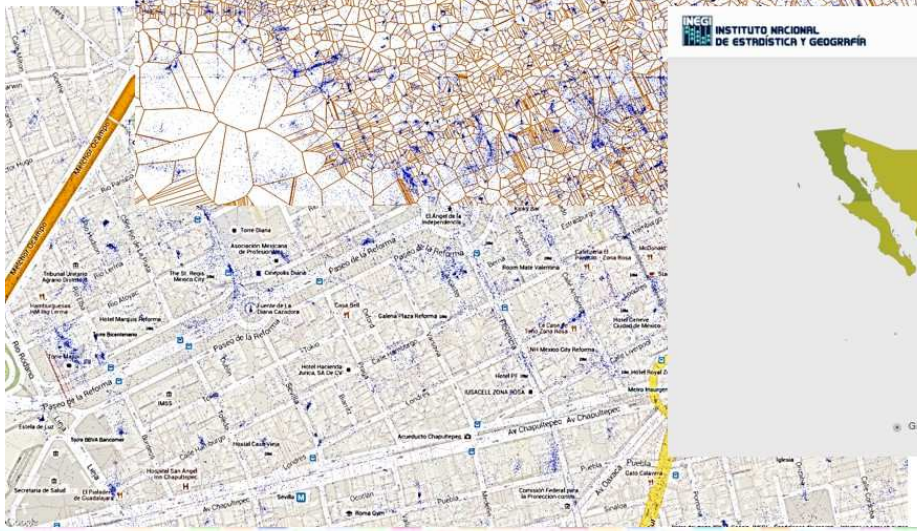


http://www.inegi.org.mx/inegi/contenidos/investigacion/Experimentales/animotuitero/default.aspx

Integration of other sources

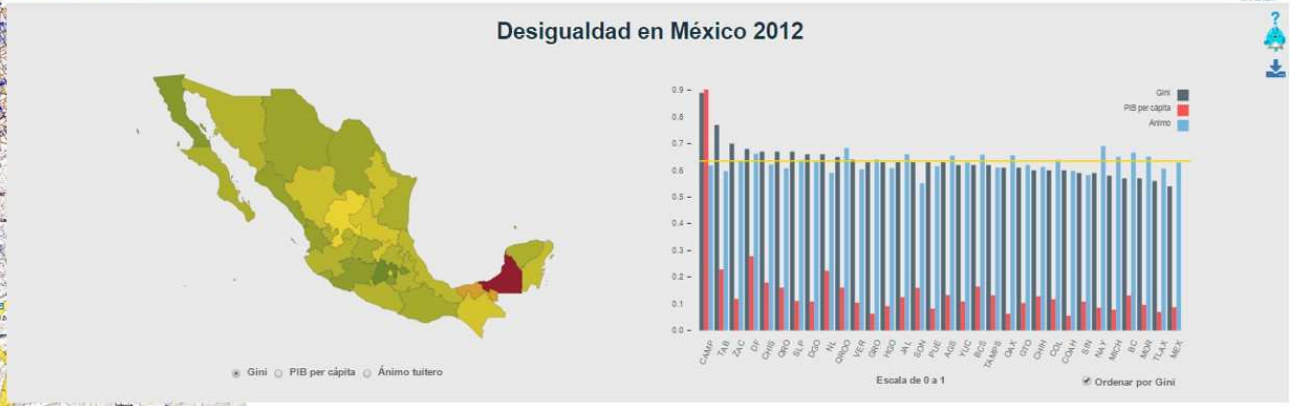






INSTITUTO NACIONAL DE ESTADÍSTICA Y GEOGRAFÍA

BIGDATA INEGI



Some applications

- Tourism
- Migration
- Use of roads
- Regional influence of big cities
- Mobility patterns
- Business activity patterns
- Subjective wellbeing
- Inequities impact
- Impact analysis of relevant news
- Mental health
- Misogynist/discriminatory language use
- SDG indicators?

Collaboration

- International
  - UNECE
    - ICHEC
  - UNSD
  - LAMBDoop
  - University of Pensylvania
- National
  - KioNetworks
  - Dattlas
  - TecMilenio
  - INFOTEC
  - Centro Geo
  - CIDE
  - CIMAT
  - Sectur
- Internal
  - INEGI General Directorates

Innovación con propósito de vida.

INSTITUTO NACIONAL
DE ESTADÍSTICA Y GEOGRAFÍA

Questions?

Juan.Munoz@inegi.org.mx
Dr. Juan Muñoz López

INEGI
INSTITUTO NACIONAL
DE ESTADÍSTICA Y GEOGRAFÍA

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