Big Data en el INEGI

UNA VISIÓN TECNOLÓGICA

11 de Noviembre 2015

4 mil Empleados en Aguascalientes y 18 mil en todo el País

COLABORACIÓN INTERINSTITUCIONAL

- Nacional

- Internacional

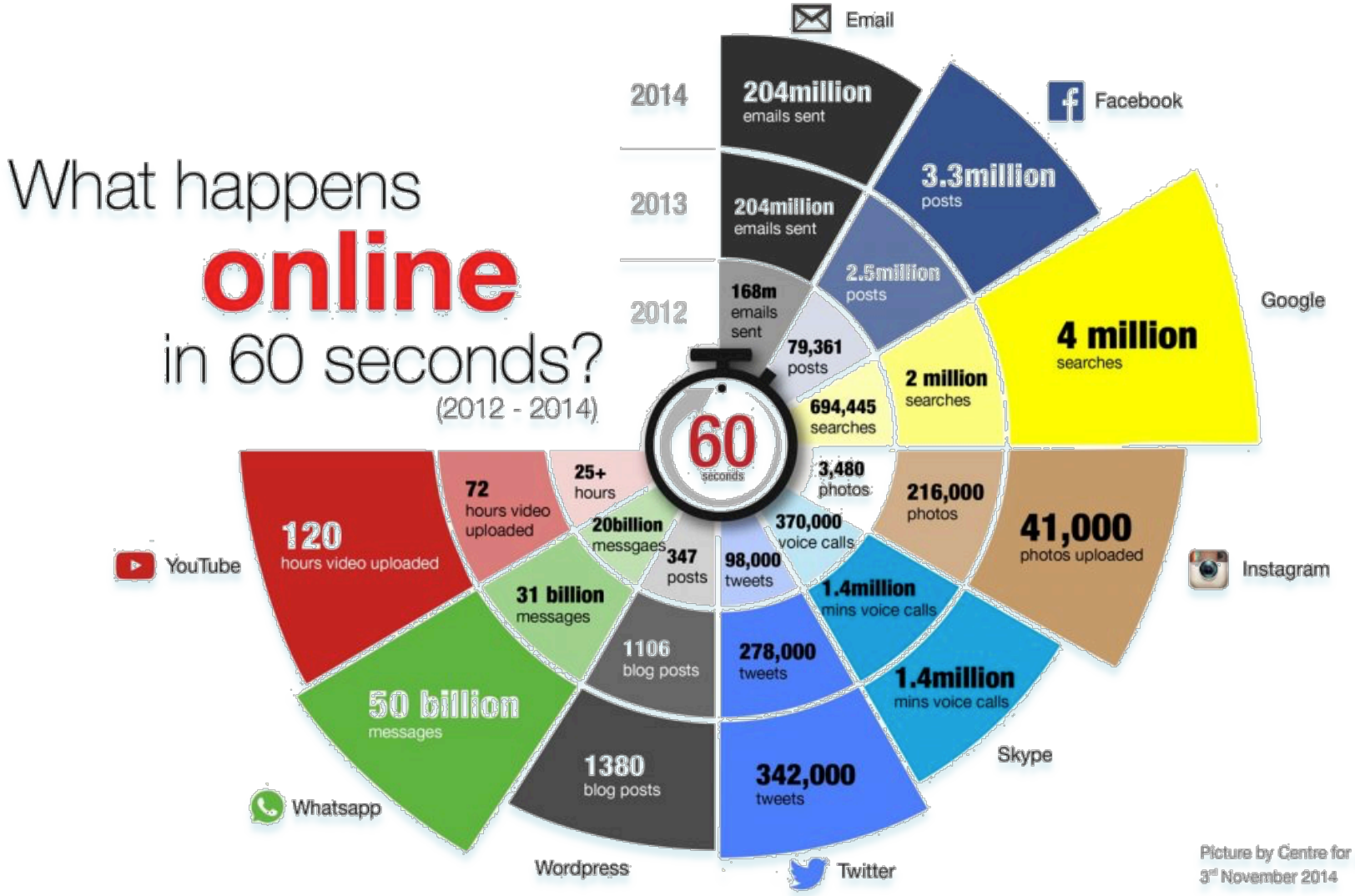
BIGDATA INEGI

Considerar las Nuevas Fuentes de Datos para Complementar a las Tradicionales.

Fuentes sin un diseño original, en una diversidad de fines posibles, normalmente ajenos a las causas que permitieron la generación

@abxda

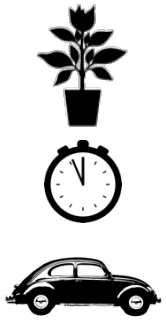
Las Fuentes de Información siguen creciendo

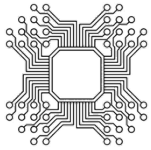


Panorama Tecnológico

BIGDATA INEGI

Internet de las Cosas





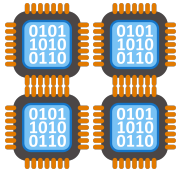
Sensores

{ json }
< xml >
c,s,v

Infraestructura de Cómputo

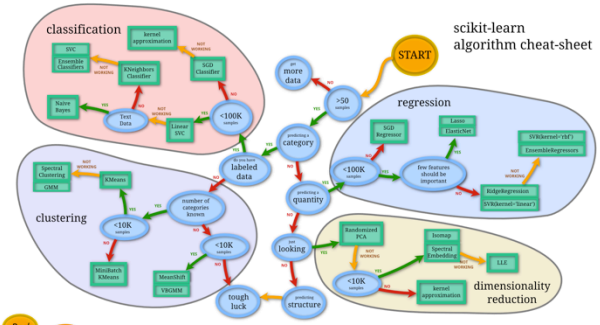
Sistemas de Archivos Distribuidos

Bases de Datos NoSQL



Computo Paralelo y Concurrente

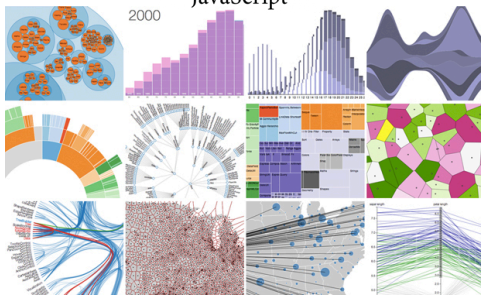
Estadística



Análisis Multivariado
Machine Learning
Análisis de Interacción Espacial

Visualización

JavaScript

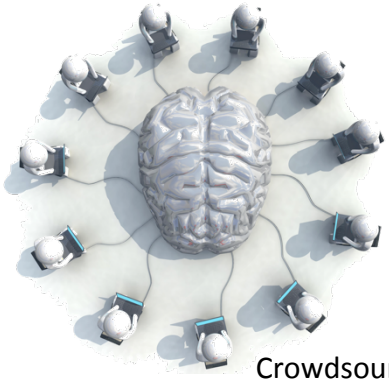


Internet de las Personas

Redes Sociales

{ json }
< xml >
c,s,v

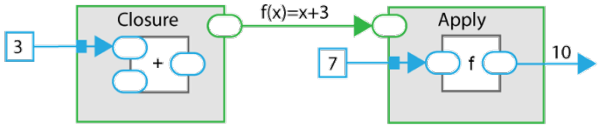
Internet de las Ideas



Crowdsourcing

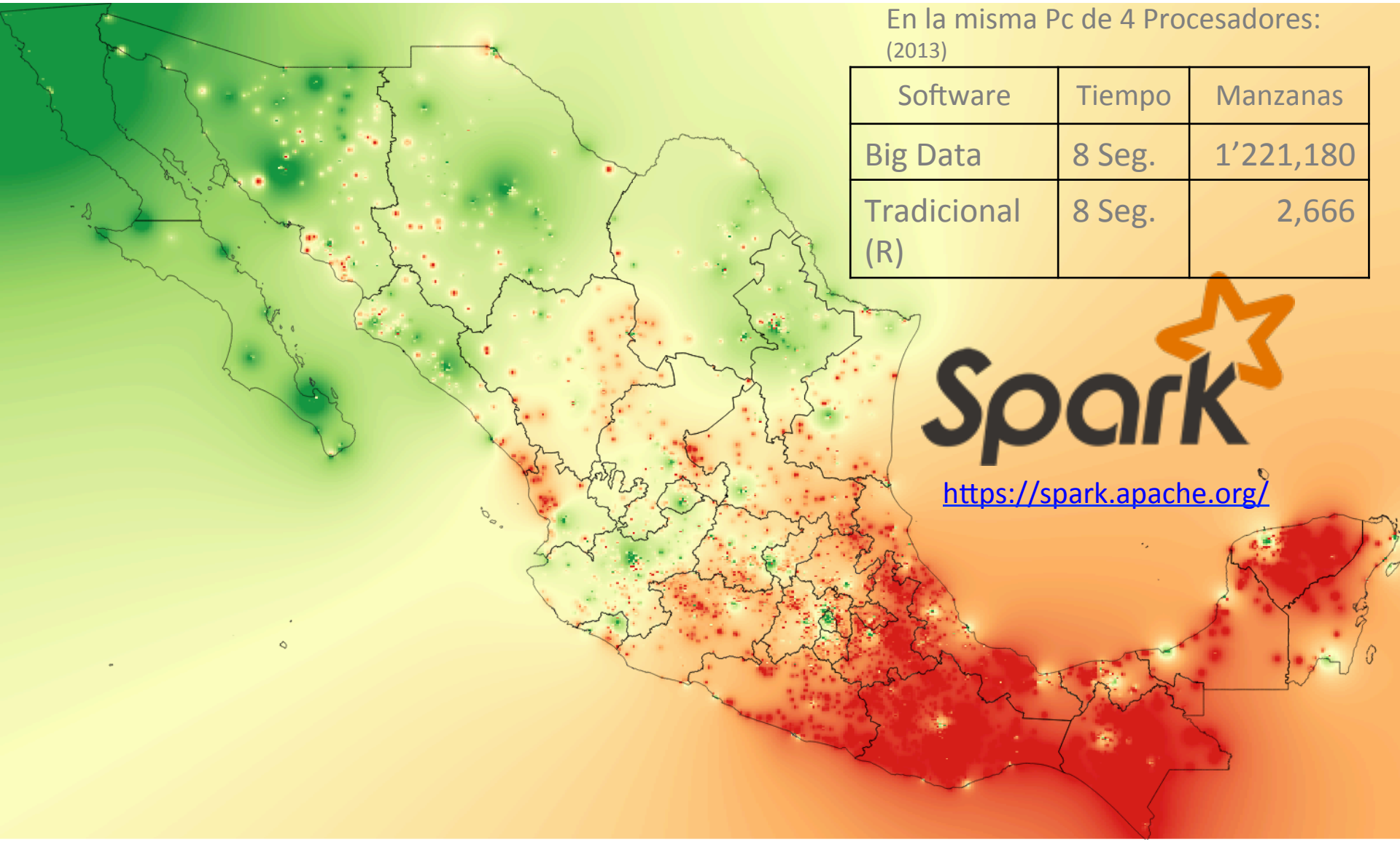
{ json }
< xml >
c,s,v

Programación Funcional



Razonamiento Algebraico

Estratificación de 1.2 M de Manzanas (2013)



En la misma Pc de 4 Procesadores:
(2013)

Software	Tiempo	Manzanas
Big Data	8 Seg.	1’221,180
Tradicional (R)	8 Seg.	2,666

Spark

https://spark.apache.org/

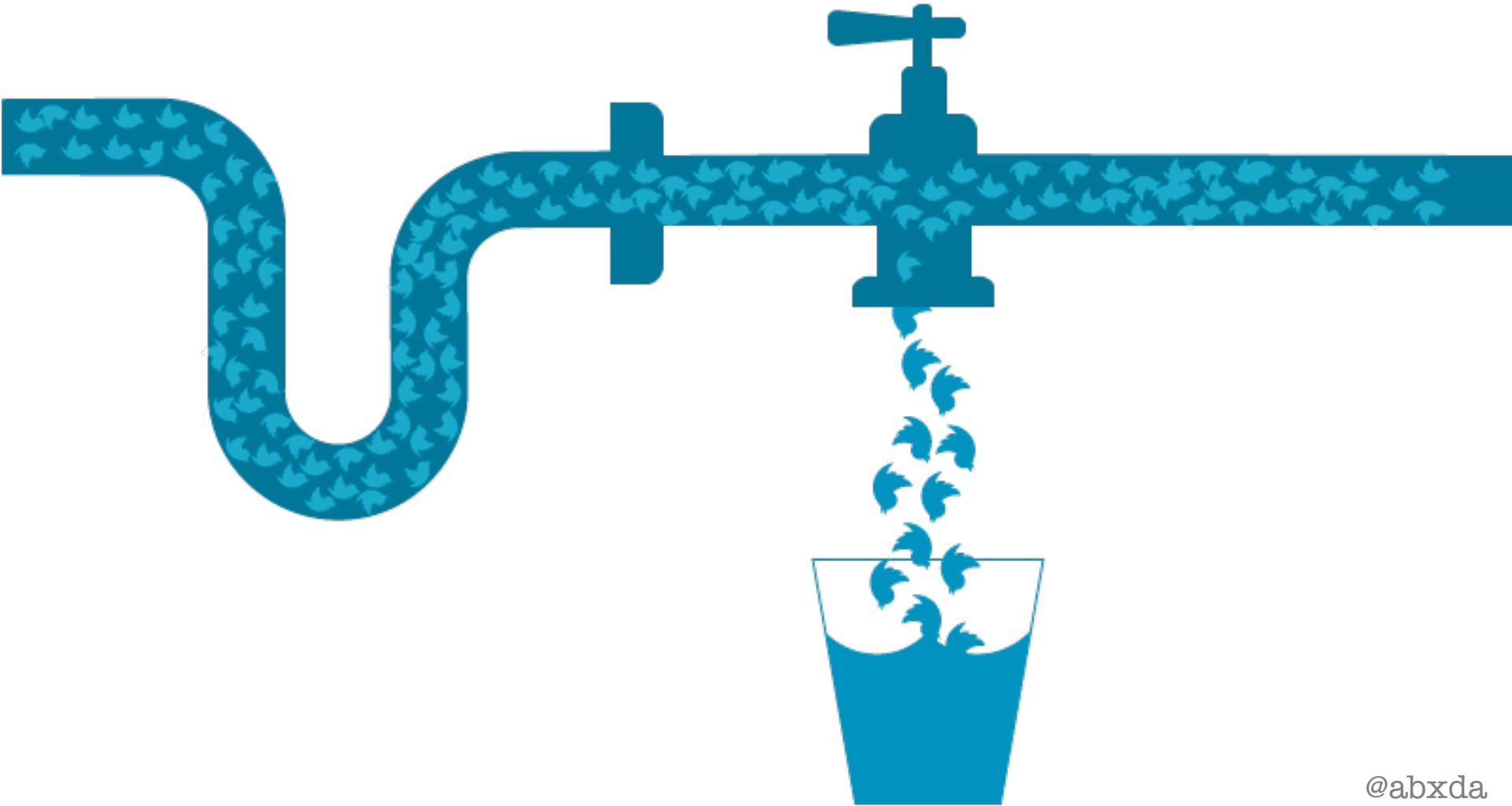
%Acceso a Internet, %Pc, %Telefono Celular, %Automovil

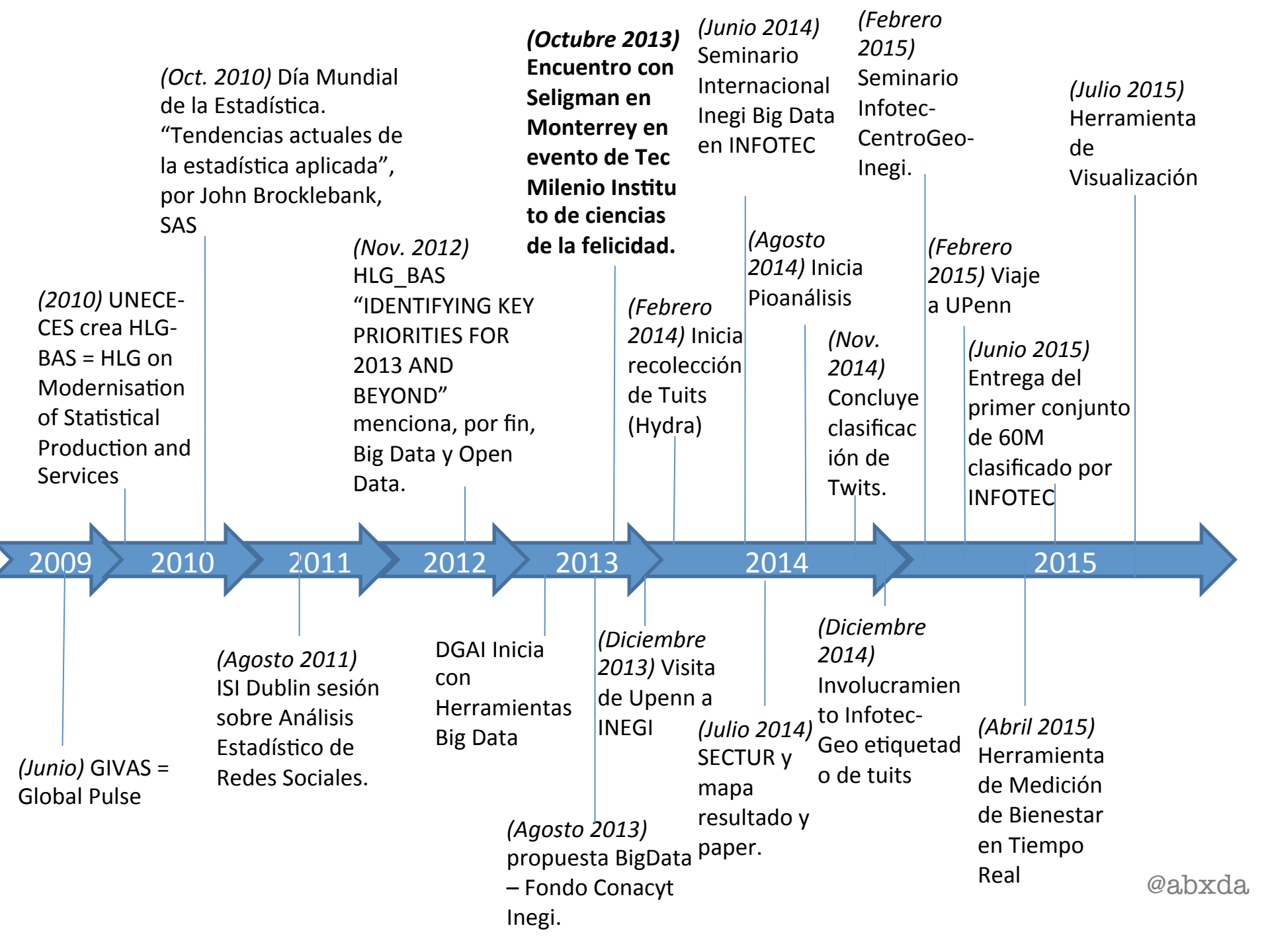
@abxda

TWITTER COMO FUENTE DE BIG DATA

Para medir el pulso emotivo de México

…y mucho más …




2009
2010
2011
2012
2013
2014
2015
(2010) UNECE-CES crea HLG-BAS = HLG on Modernisation of Statistical Production and Services
(Oct. 2010) Día Mundial de la Estadística. "Tendencias actuales de la estadística aplicada", por John Brocklebank, SAS
(Nov. 2012) HLG_BAS "IDENTIFYING KEY PRIORITIES FOR 2013 AND BEYOND" menciona, por fin, Big Data y Open Data.
(Octubre 2013) Encuentro con Seligman en Monterrey en evento de Tec Milenio Instituto de ciencias de la felicidad.
(Junio 2014) Seminario Internacional Inegi Big Data en INFOTEC
(Febrero 2015) Seminario Infotec-CentroGeo-Inegi.
(Julio 2015) Herramienta de Visualización
(Junio) GIVAS = Global Pulse
(Agosto 2011) ISI Dublin sesión sobre Análisis Estadístico de Redes Sociales.
DGAI Inicia con Herramientas Big Data
(Agosto 2013) propuesta BigData – Fondo Conacyt Inegi.
(Diciembre 2013) Visita de Upenn a INEGI
(Febrero 2014) Inicia recolección de Tuits (Hydra)
(Julio 2014) SECTUR y mapa resultado y paper.
(Agosto 2014) Inicia Pioanálisis
(Nov. 2014) Concluye clasificación de Twits.
(Diciembre 2014) Involucramiento Infotec-Geo etiquetado de tuits
(Febrero 2015) Viaje a UPenn
(Junio 2015) Entrega del primer conjunto de 60M clasificado por INFOTEC
(Abril 2015) Herramienta de Medición de Bienestar en Tiempo Real
@abxda

OBJETIVO DEL PROYECTO

Generar indicadores experimentales, nuevos o que complementen los generados por métodos tradicionales, utilizando **técnicas de Big Data** para la extracción, almacenamiento, procesamiento, análisis y visualización de los datos.

Colaboración

Dr. Enrique Ordaz
enrique.ordaz@inegi.org.mx

Dr. Gerardo Leyva
gerardo.leyva@inegi.org.mx

Dr. Alfredo Bustos
alfredo.bustos@inegi.org.mx

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

Ing. Silvia Fraustro
Silvia.fraustro@inegi.org.mx

Mtro. Abel Coronado
abel.coronado@inegi.org.mx

Ing. Ricardo Olvera
Ricardo.olvera@inegi.org.mx

Lic. Marco Ibarra
Marco.ibarra@inegi.org.mx

Dr. Elio Villaseñor
elio.villaseñor@infotec.com.mx

Dr. Mario Graff
mario.graff@infotec.com.mx

Dr. Eric Tellez
eric.tellez@infotec.com.mx

Dr. Sabino Miranda
sabino.miranda@infotec.com.mx

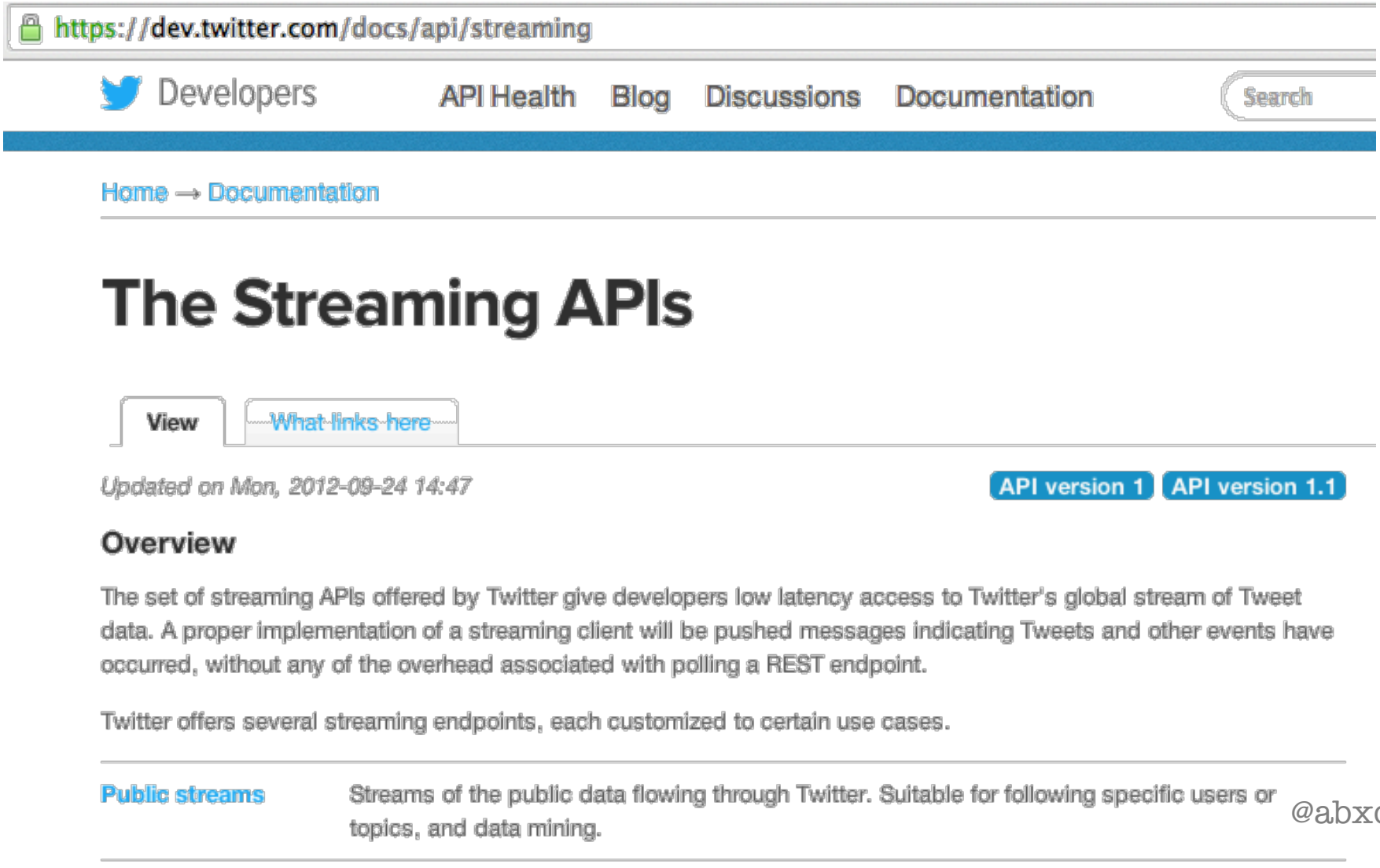
Dr. Oscar S. Siordia
osanchez@centrogeo.edu.mx

Dra. Daniela Moctezuma
dmoctezuma@centrogeo.edu.mx

Y el apoyo de:

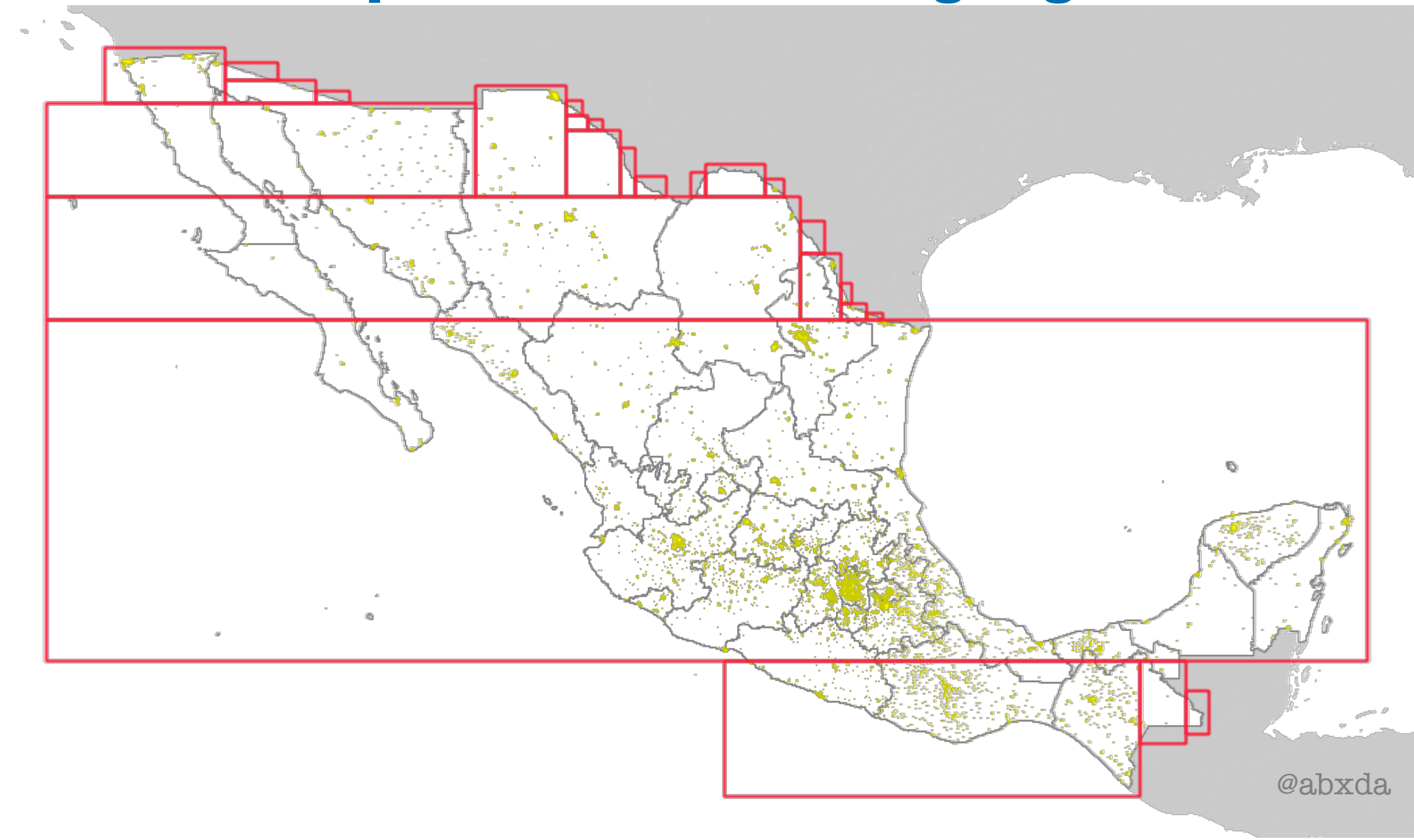
@abxda

Todos los tuits están disponibles para su recolección en tiempo real.



@abxda

Incluso permite consultas geográficas



BIGDATA INEGI

Hydra

1 año 8 meses

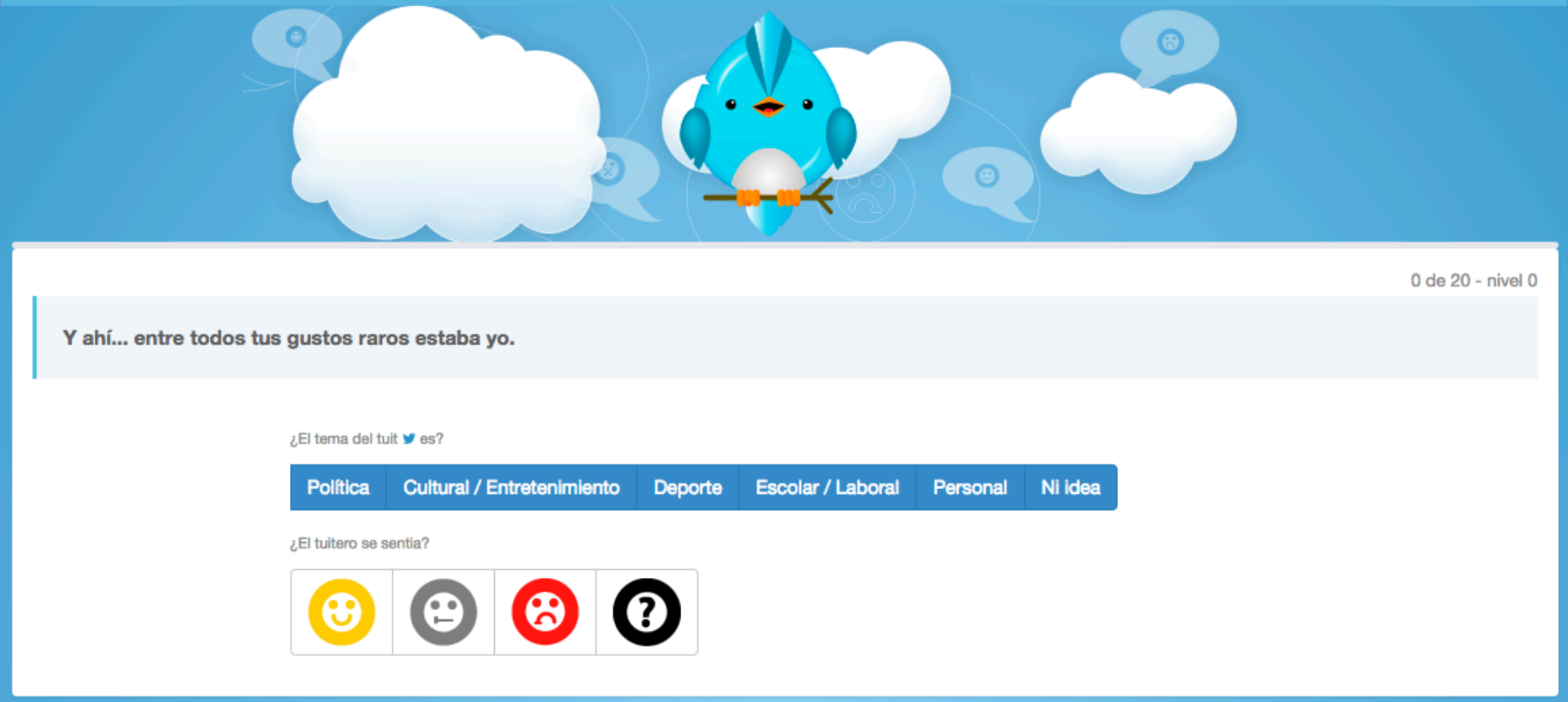
< ESCALABILIDAD HORIZONTAL >

elasticsearch

https://www.elastic.co

@abxda

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



@hbcolectivo **@ricardoaolvera**

@abxda

Proceso de Machine Learning

BIGDATA INEGI

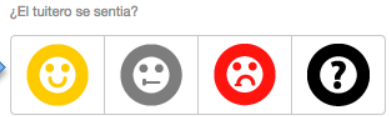
INFOTEC CENTROGEO

Entrenamiento

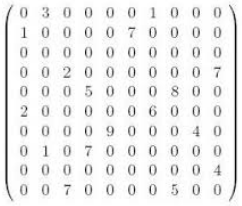
Muestra de Tuits

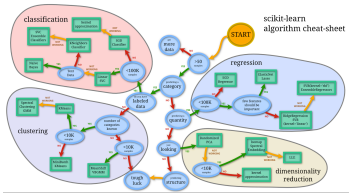
$f(x)$

Representación numérica



Etiquetado Manual

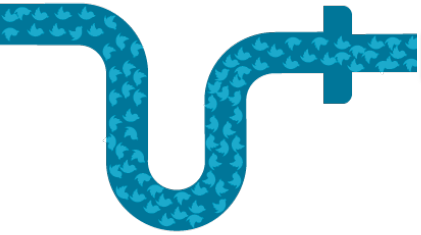




http://scikit-learn.org/
http://www.r-project.org

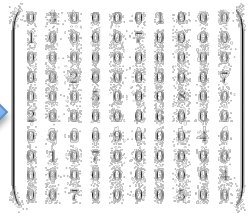
Machine Learning

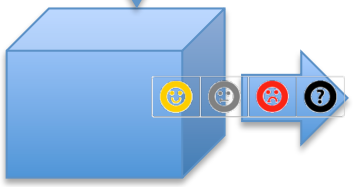
Producción



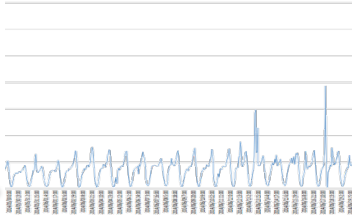
Tuits en Tiempo Real

$f(x)$





Clasificador

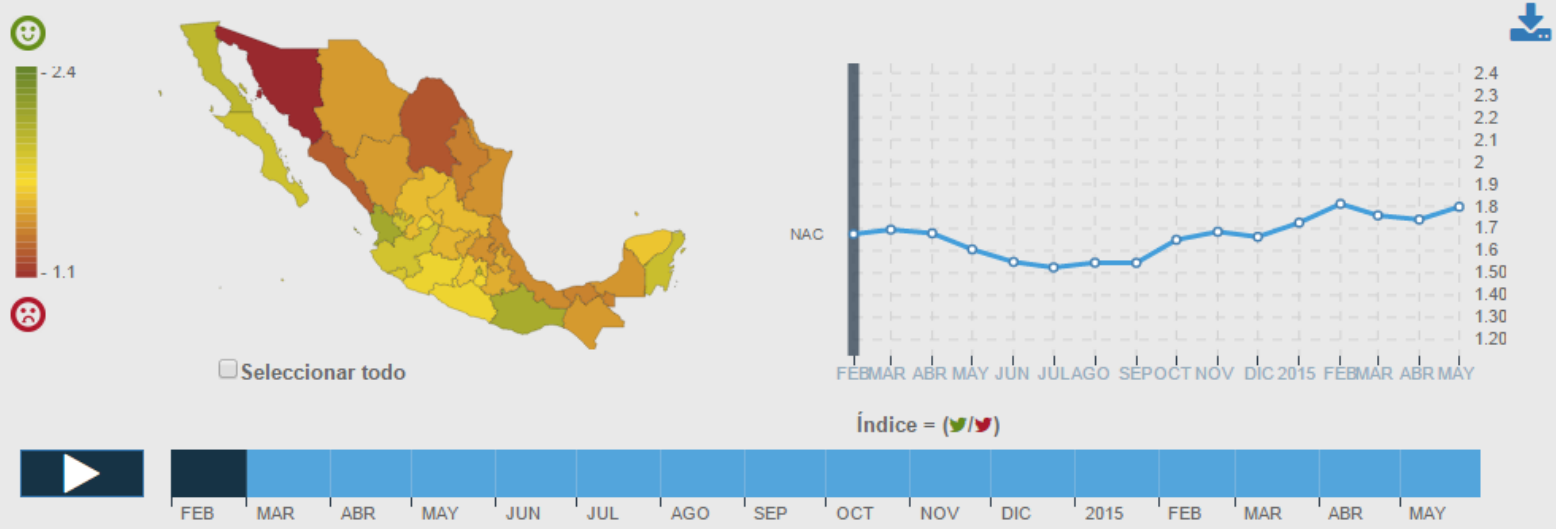


Indicador
de sentimiento

INSTITUTO NACIONAL DE ESTADÍSTICA Y GEOGRAFÍA

SOCIAL BIGDATA INEGI

Estado de ánimo de los tuiteros en México



Nube de palabras

Positivo - Nacional - Febrero

Negativo - Nacional - Febrero

Original

Entrenamiento

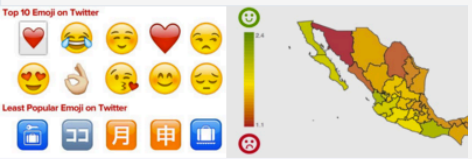
Normalizada

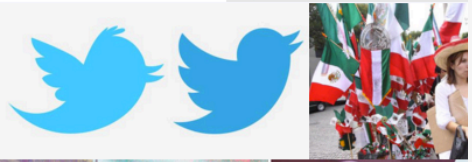
En colaboración con:

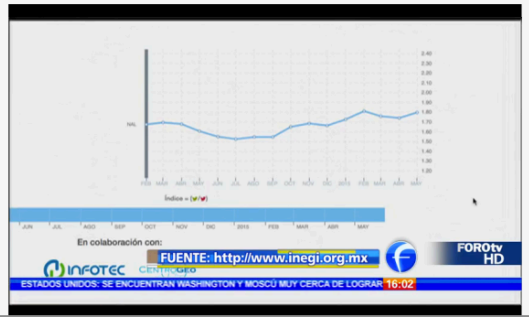
INFOTEC CENTROGEO

@abxda

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







EXPANSIÓN
CEO Responsabilidad, ¿para qué?
Edición 1177
SUSCRÍBETE

CNN EXPANSIÓN

Inicio | Empresas | Economía | Mercados | Finanzas Personales | Emprendedores | Carrera | Tecnología | Videos

¿Feliz o triste?, el INEGI analiza el ánimo de los tuiteros

noticieros.televisa.com/programas-primero-noticias/1510/primero-futuro-estado-animo-tuiteros-me

Primero Futuro: Estado de ánimo de tuiteros de México

Aura López comenta que gracias a un estudio que hizo el INEGI, ya es posible saber el estado de ánimo de los usuarios de Twitter en México

Temas Relacionados
Recomendados, Primero Noticias, Carlos Loret de Mola

Coberturas Especiales | Por el Planeta | Infografías | Lo que debes saber sobre el

INEGI usará Twitter para un ambicioso proyecto de estadística

Pretende dar seguimiento en tiempo real al estado de ánimo de los tuiteros en México

Twittear 8 | G+1 0 | Share | Comentarios (0)

MILENIO.COM

TENDENCI

Firmas | Estados | Policía | Tendencias | Negocios | Cultura | Datal

El Polígrafo | Emprendedores | Fotogalerías | Moneros | Laberinto

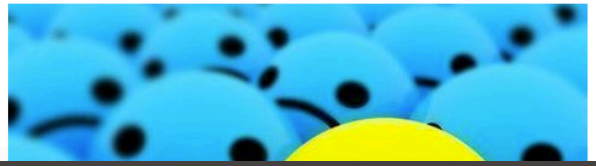
áles son los estados más es, según sus tuits?

ır millones de tuits, el Inegi concluyó que Guerrero, Oaxaca y Michoacán los estados más felices; Sonora y Coahuila son de los más tristes.

rios 114 | Me gusta 301 | 1 | Compartir 416 | G+1 0

i medirá nimo de a través

luién es la

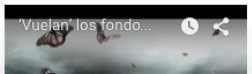


Indigo

Nics | Piensa | Deportes | Ideas | El Wiken | Versión Impresa | Vivo

MÁS SOBRE EL TEMA VIDEOS

os encontrados en

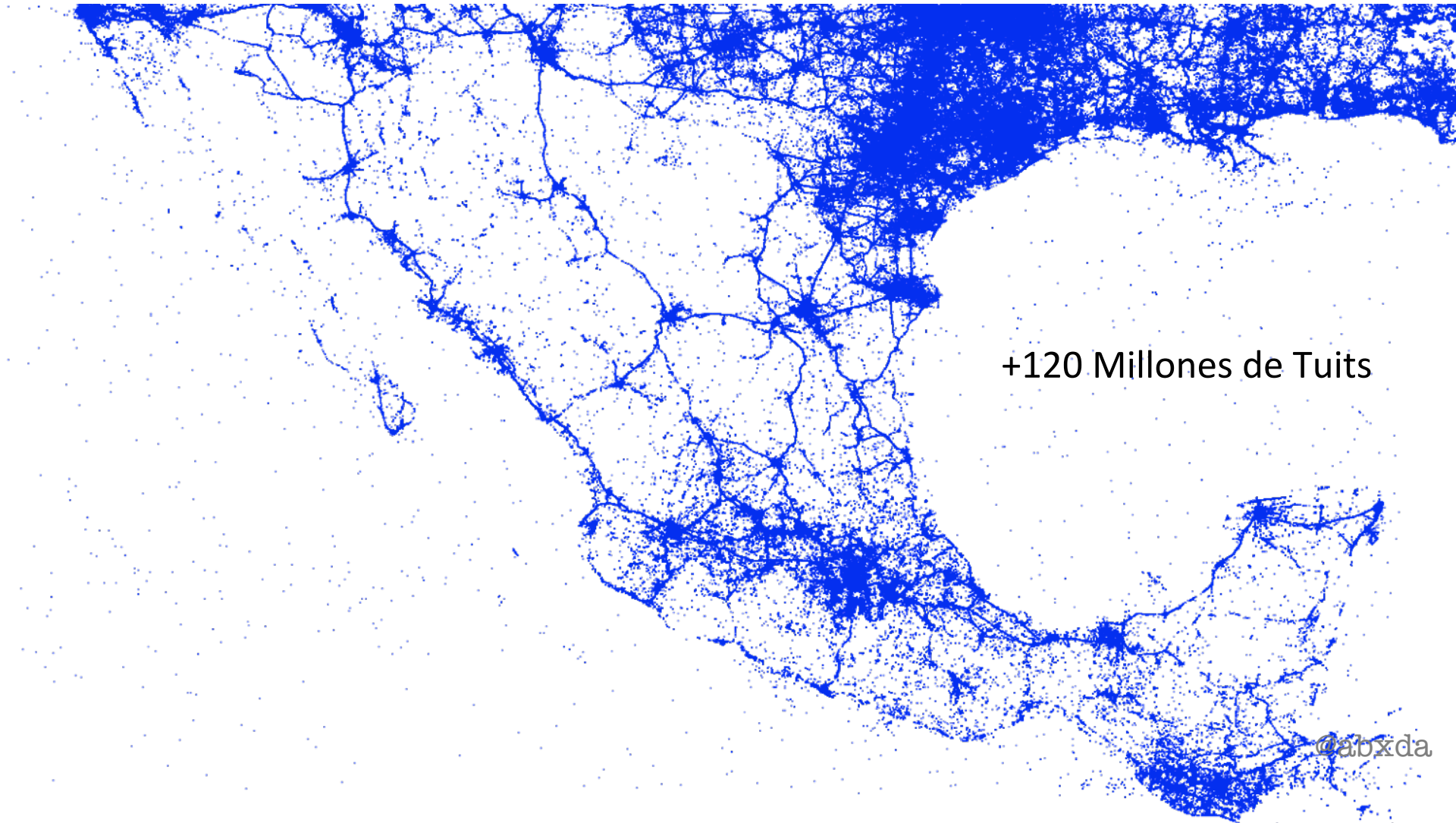


MAS ALLÁ DEL ANÁLISIS DEL SENTIMIENTO

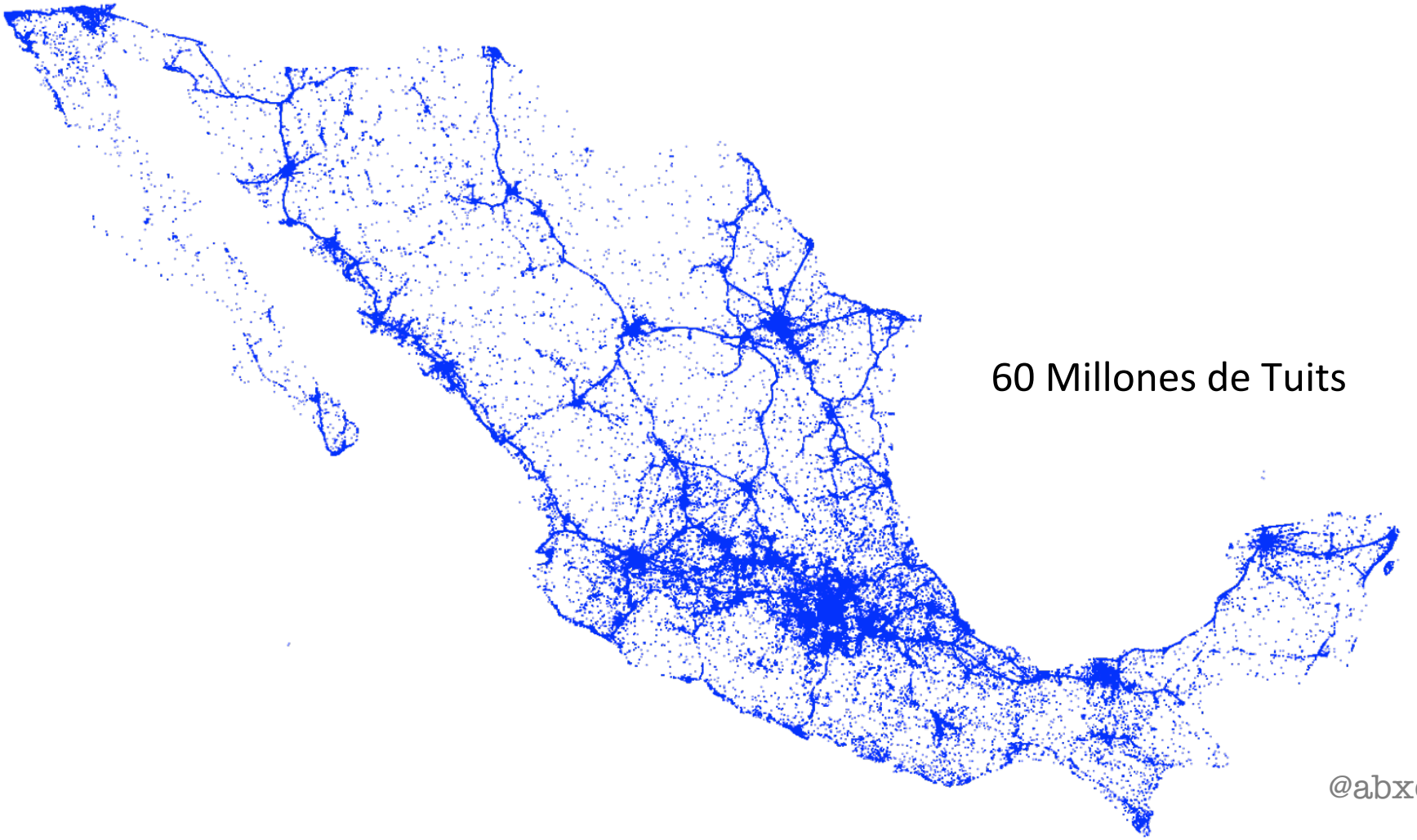
Apache Spark

http://spark.apache.org/

Visualización de la Base de Datos



Visualización de la Base de Datos

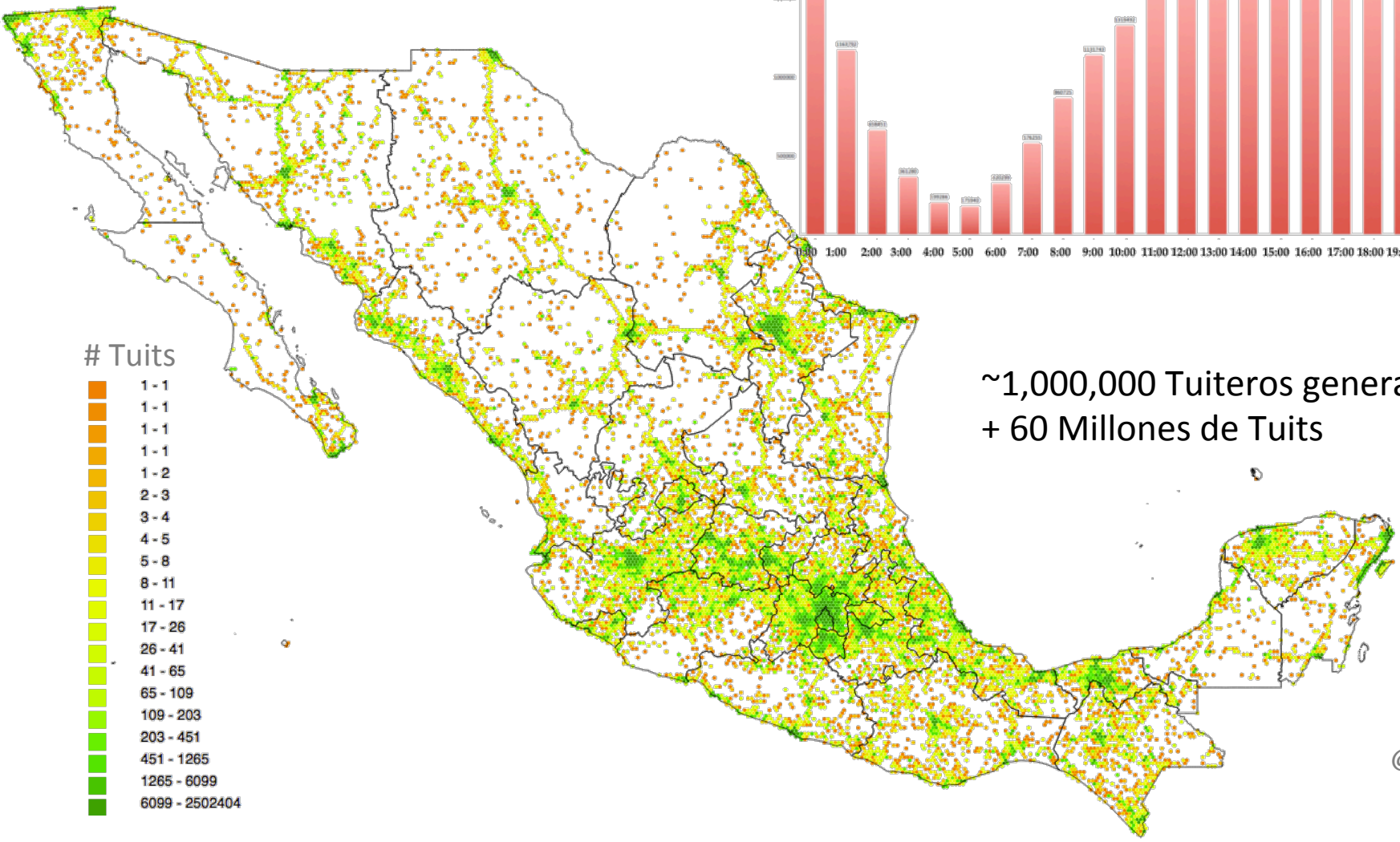


60 Millones de Tuits

@abxda

BIGDATA INEGI

Frecuencia de Tuiteo

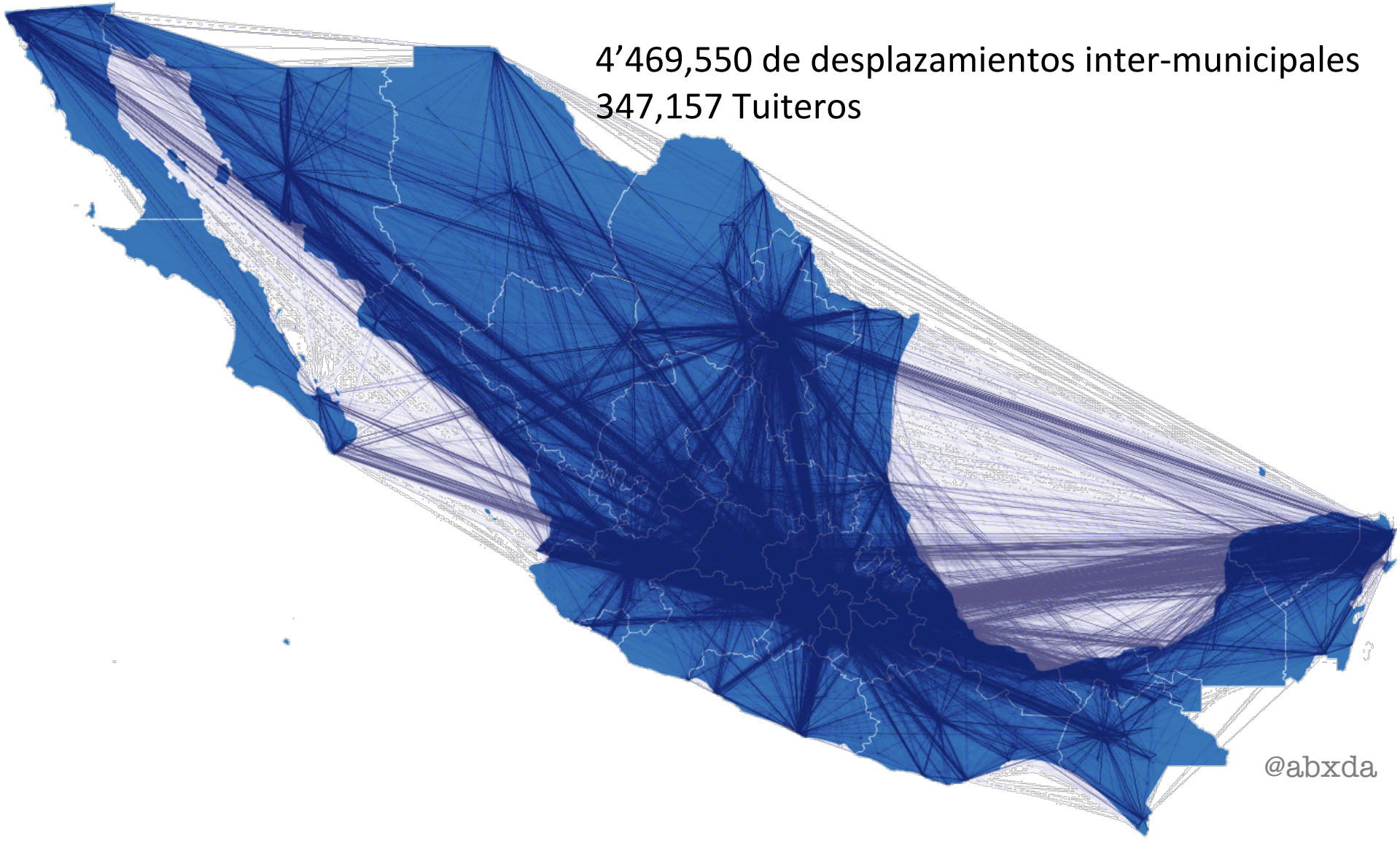


~1,000,000 Tuiteros generaron
+ 60 Millones de Tuits

@abxda

Movilidad de los Tuiteros

4’469,550 de desplazamientos inter-municipales
347,157 Tuiteros



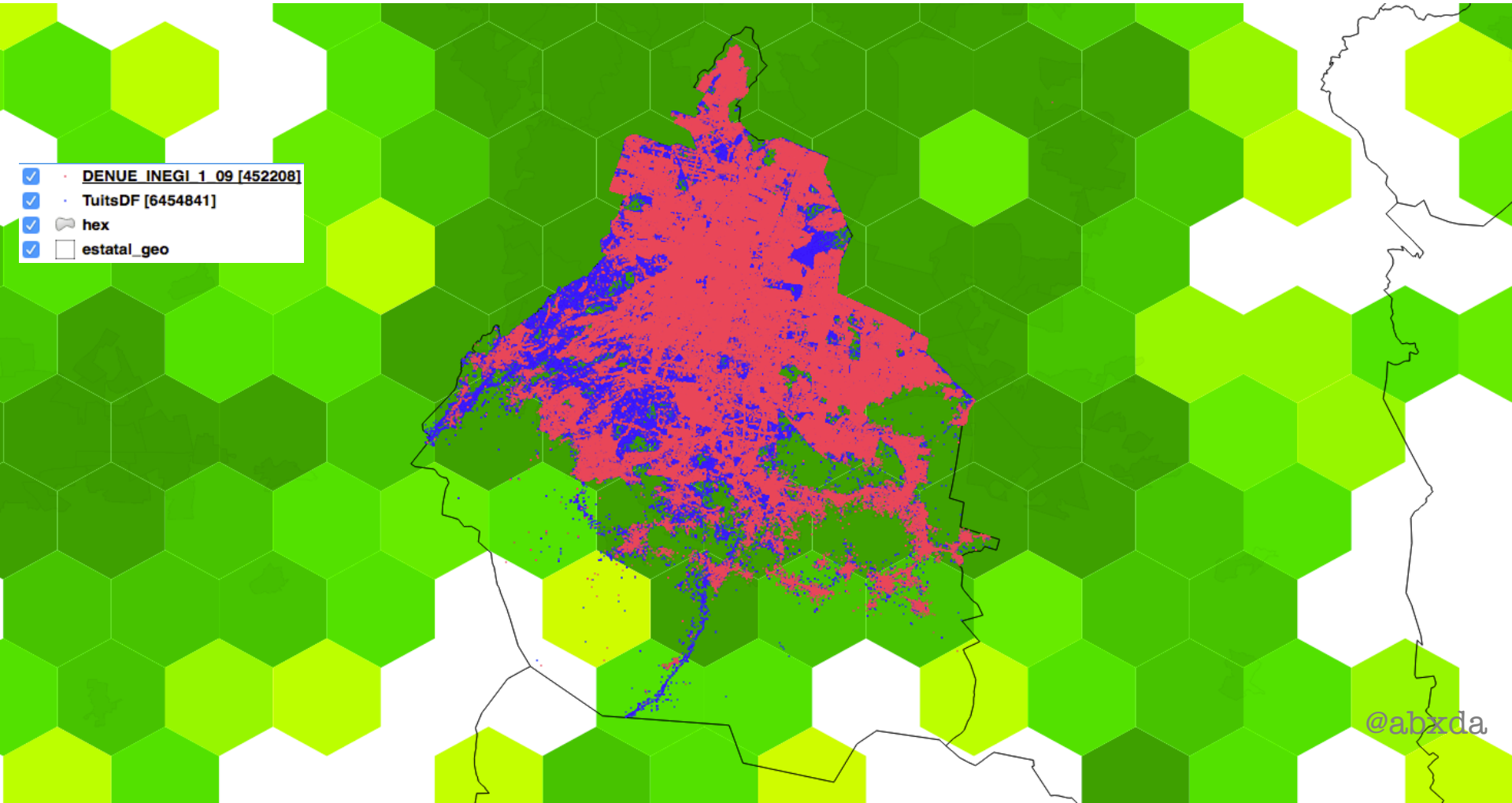
@abxda

Red Nacional de Caminos (Open Data) y Twitter

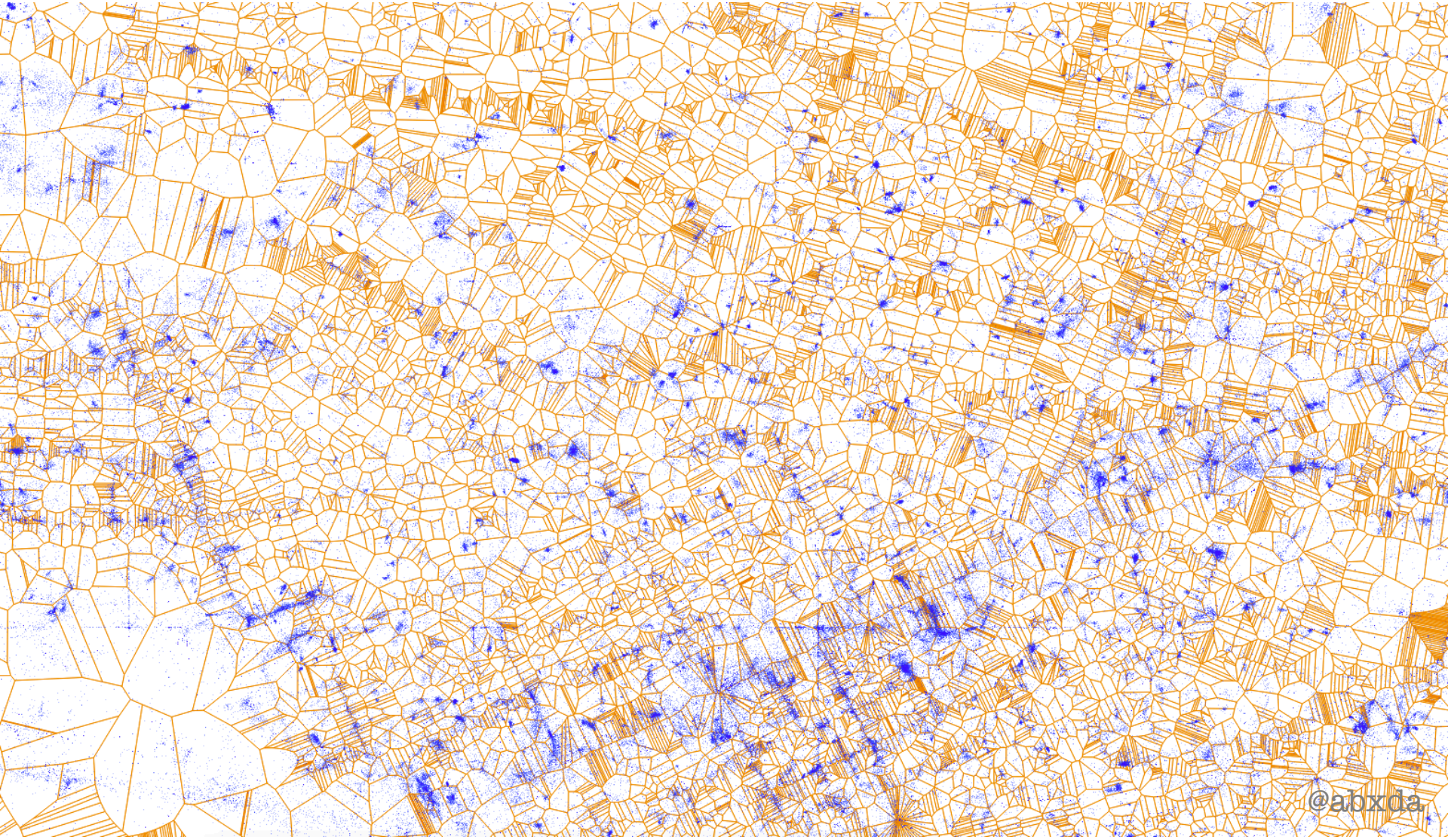
Red Nacional de Caminos y Twitter

@abxda

DENUE & Twitter

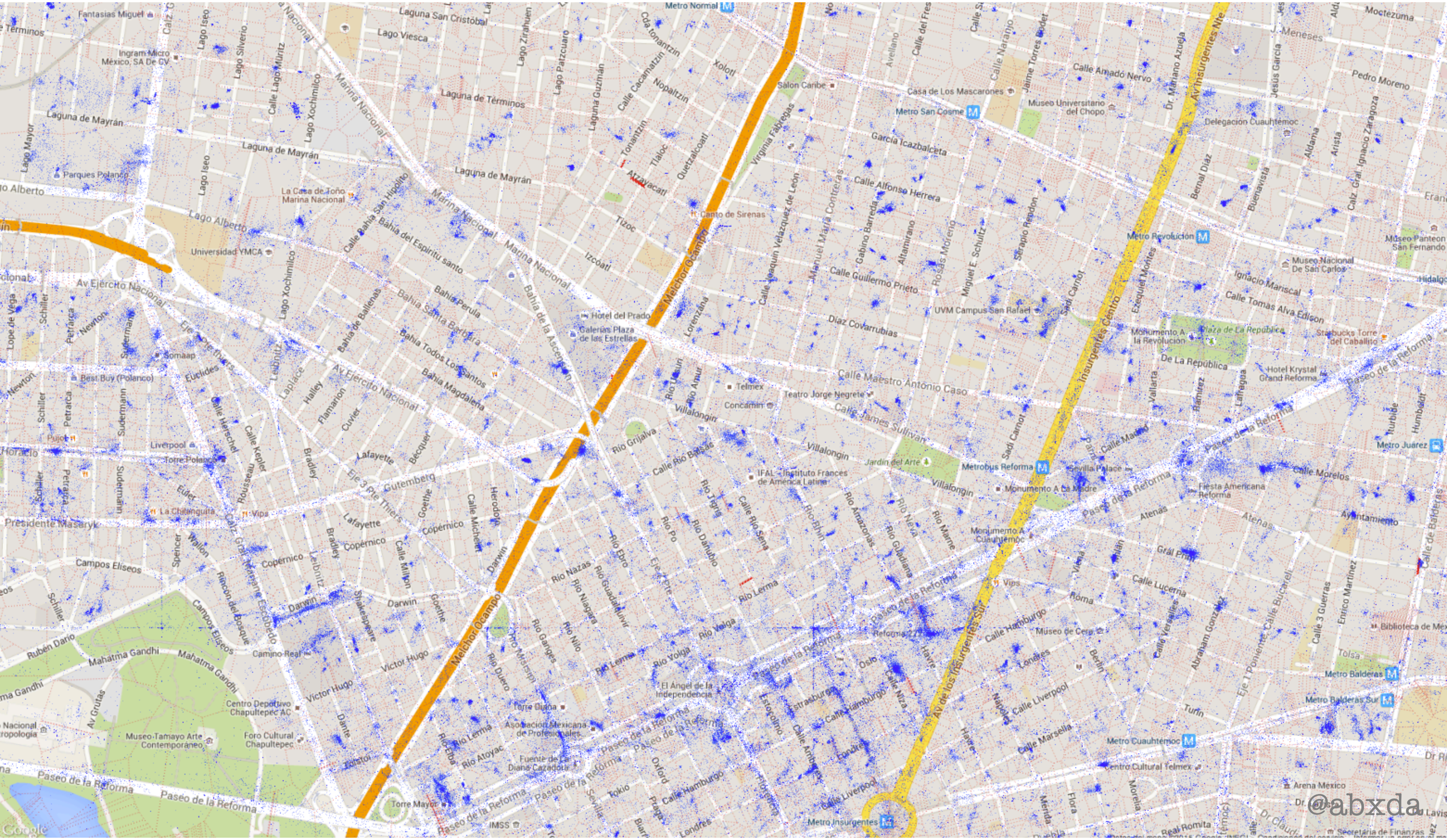


DENUE & Twitter



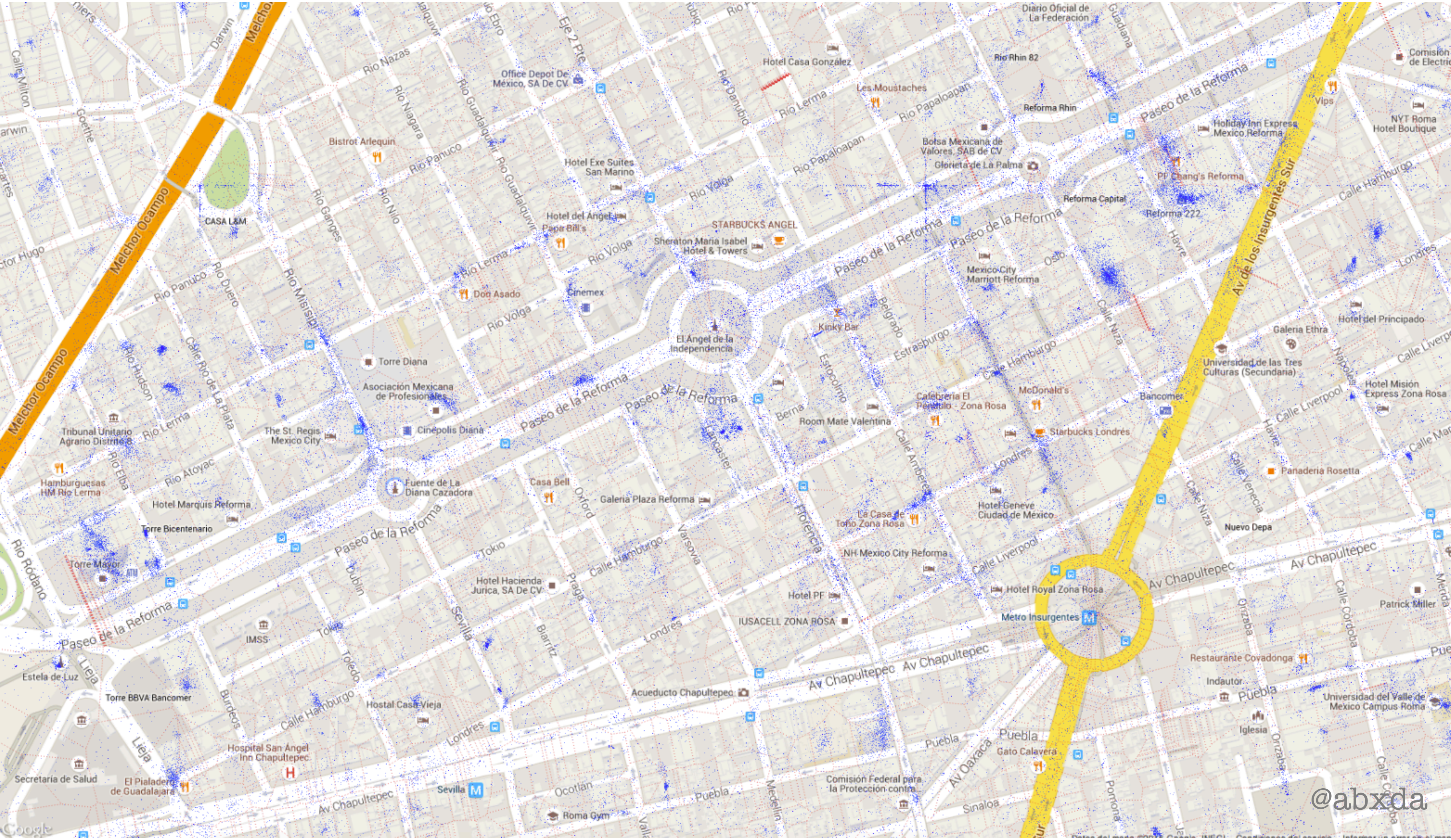
BIGDATA INEGI

DENUE & Twitter



BIGDATA INEGI

DENUE & Twitter

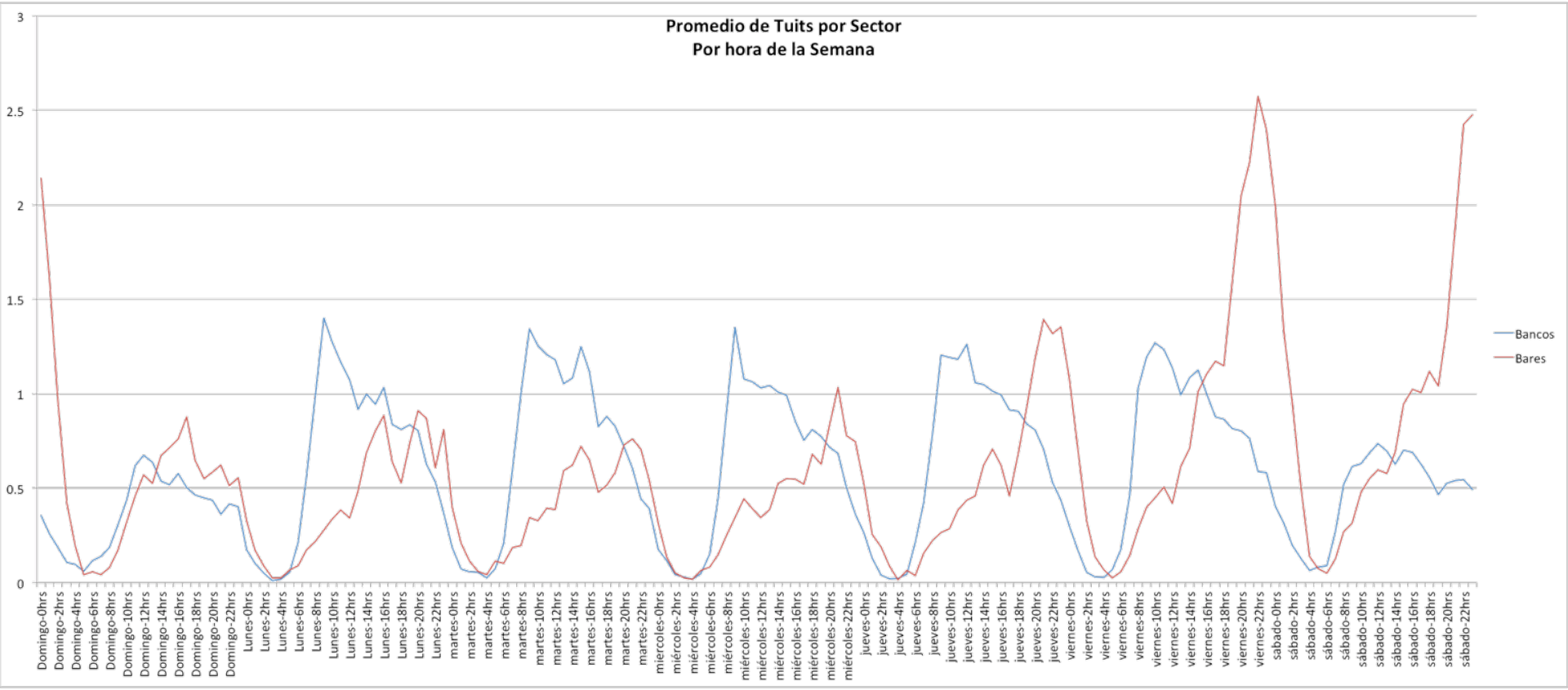


¿De dónde provienen los tuiteros de los hoteles?

Qué se dijo cerca de los Hoteles

BIGDATA INEGI

Horarios de Tuiteo cerca de algún sector



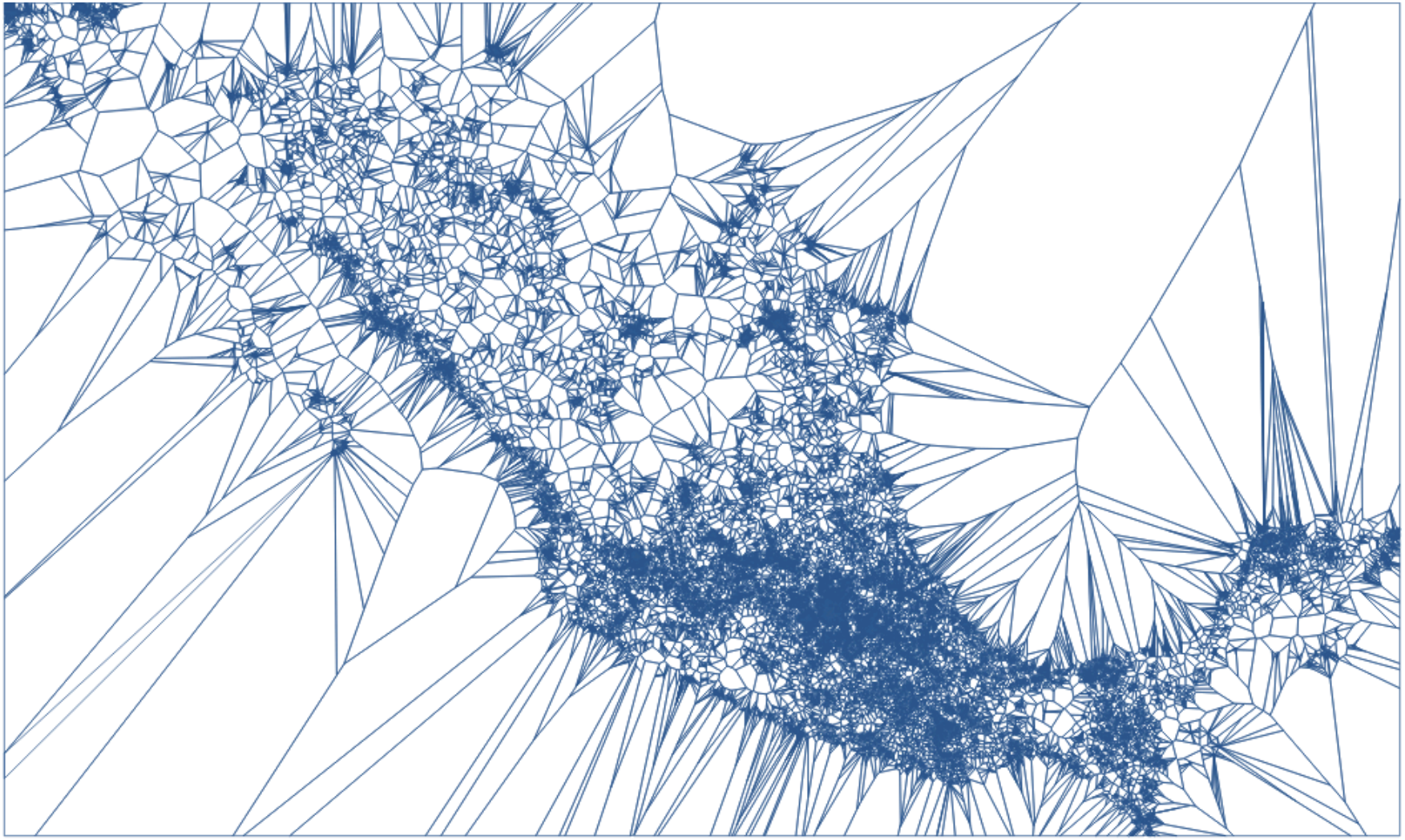
@abxda

Que se dijo cerca de los Bares

PRIMER EJERCICIO NACIONAL

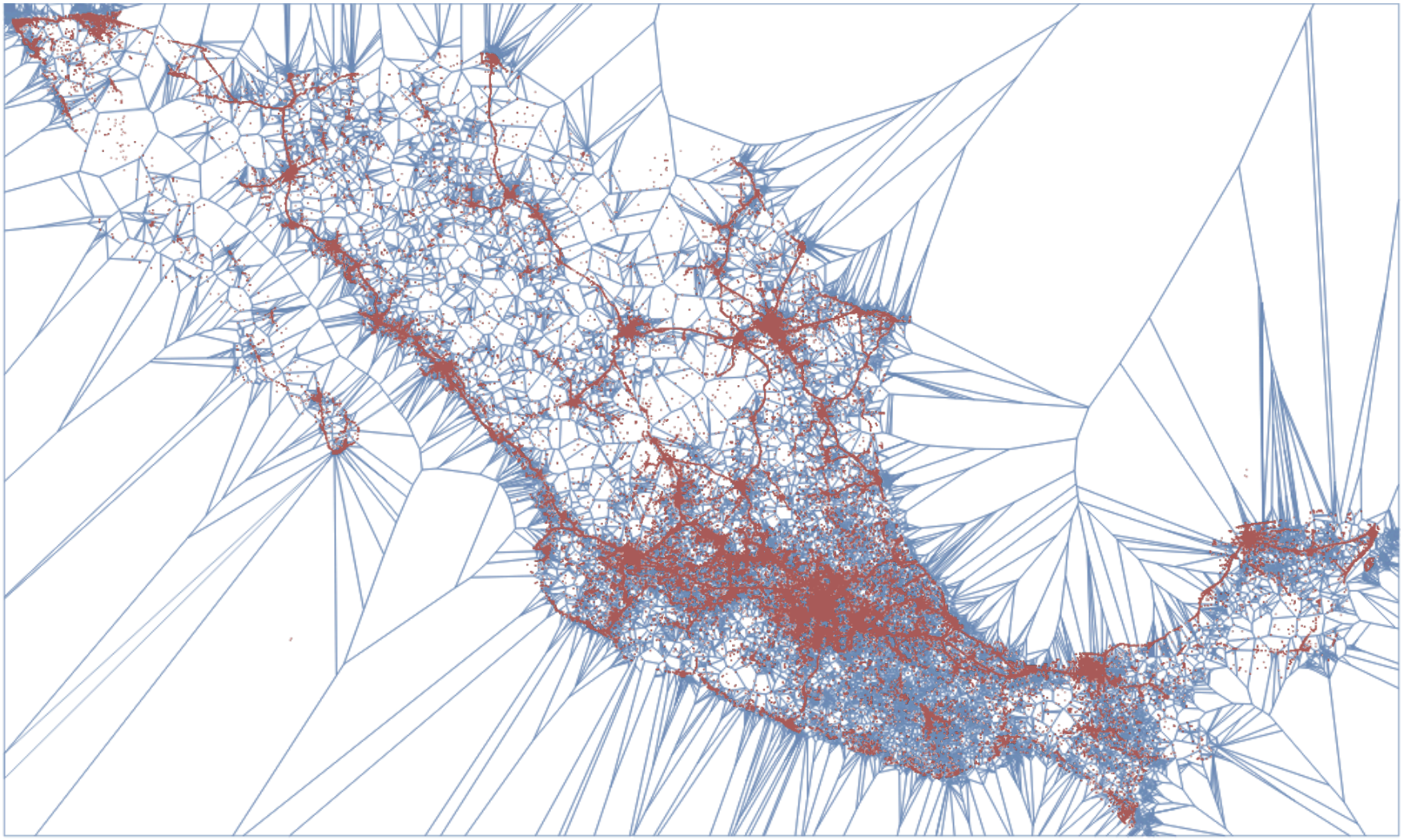
DENUE - TWITTER

4.9 M de Polígonos de Voronoi (DENUE)



BIGDATA INEGI

Big Spatial Join
(4.9 M DENUE +60 M Tweets)



SpatialSpark

Large-Scale Spatial Join Query Processing in Cloud

Simin You
Dept. of Computer Science
CUNY Graduate Center
New York, NY, USA
syou@gc.cuny.edu

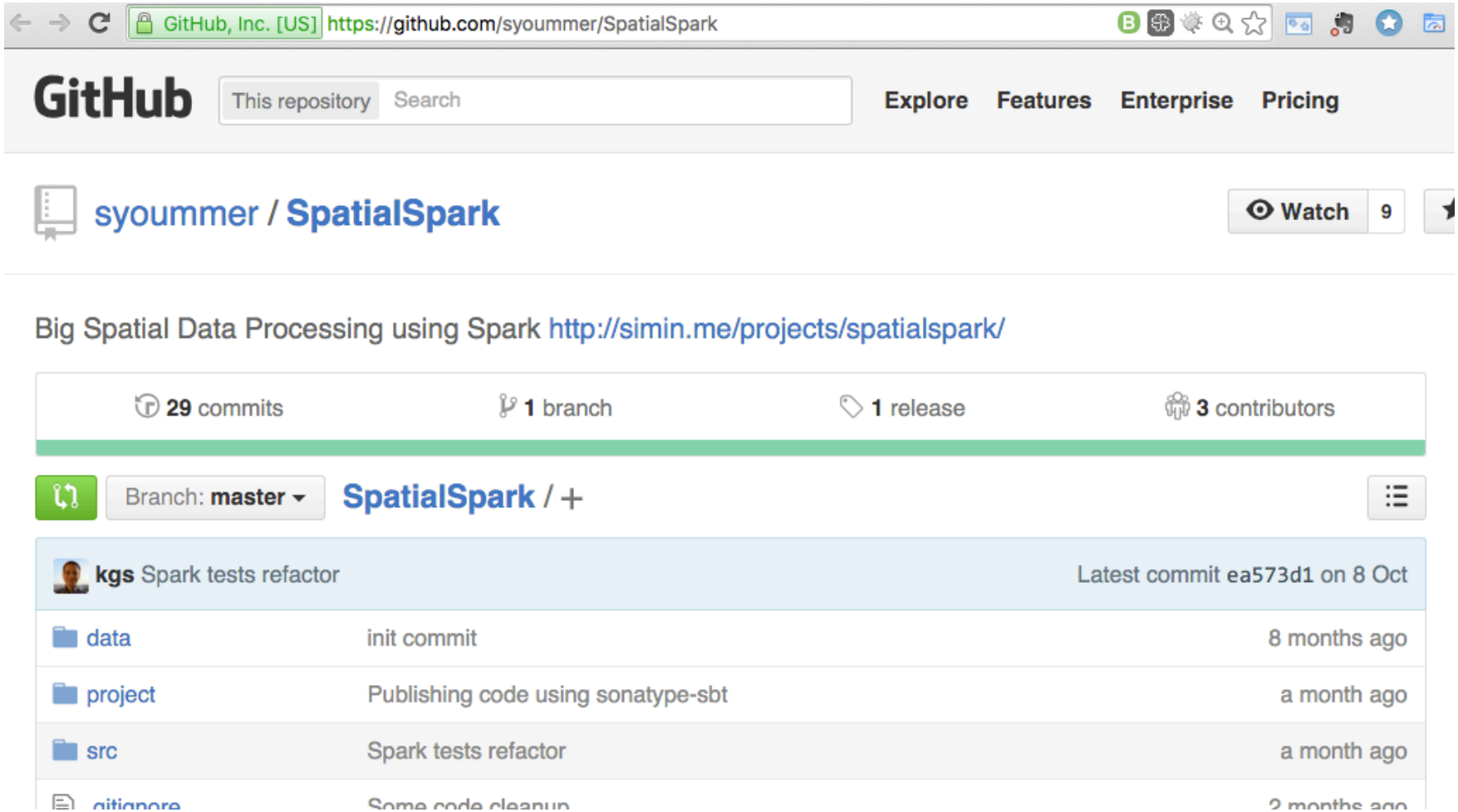
Jianting Zhang
Department of Computer Science
The City College of New York
New York, NY, USA
jzhang@cs.ccny.cuny.edu

Le Gruenwald
Dept. of Computer Science
The University of Oklahoma
Norman, OK, USA
ggruenwald@ou.edu

***Abstract*— The rapidly increasing amount of location data available in many applications has made it desirable to process their large-scale spatial queries in Cloud for performance and scalability. We report our designs and implementations of two prototype systems that are ready for Cloud deployments: SpatialSpark based on Apache Spark and ISP-MC based on Cloudera Impala. Both systems support indexed spatial joins based on point-in-polygon test and point-to-polyline distance computation. Experiments on the pickup locations of ~170 million taxi trips in New York City and ~10 million global species occurrences records have demonstrated both efficiency and scalability using Amazon EC2 clusters.**

Existing works on processing large-scale spatial join query processing mainly fall into two categories: a) improving single-node efficiency by exploiting the massive data-parallel computing power of hardware accelerators, such as Graphics Processing Units (GPUs) that are capable of general computing, and b) exploiting scalability provided by Hadoop/MapReduce based systems. Here MapReduce is referred to both as a computing model and as a component in Hadoop (together with the Hadoop Distributed File System–HDFS). Our previous work on GPU-based spatial joins [2,3] have demonstrated that it is quite possible to achieve orders of magnitude of performance improvements by re-designing and

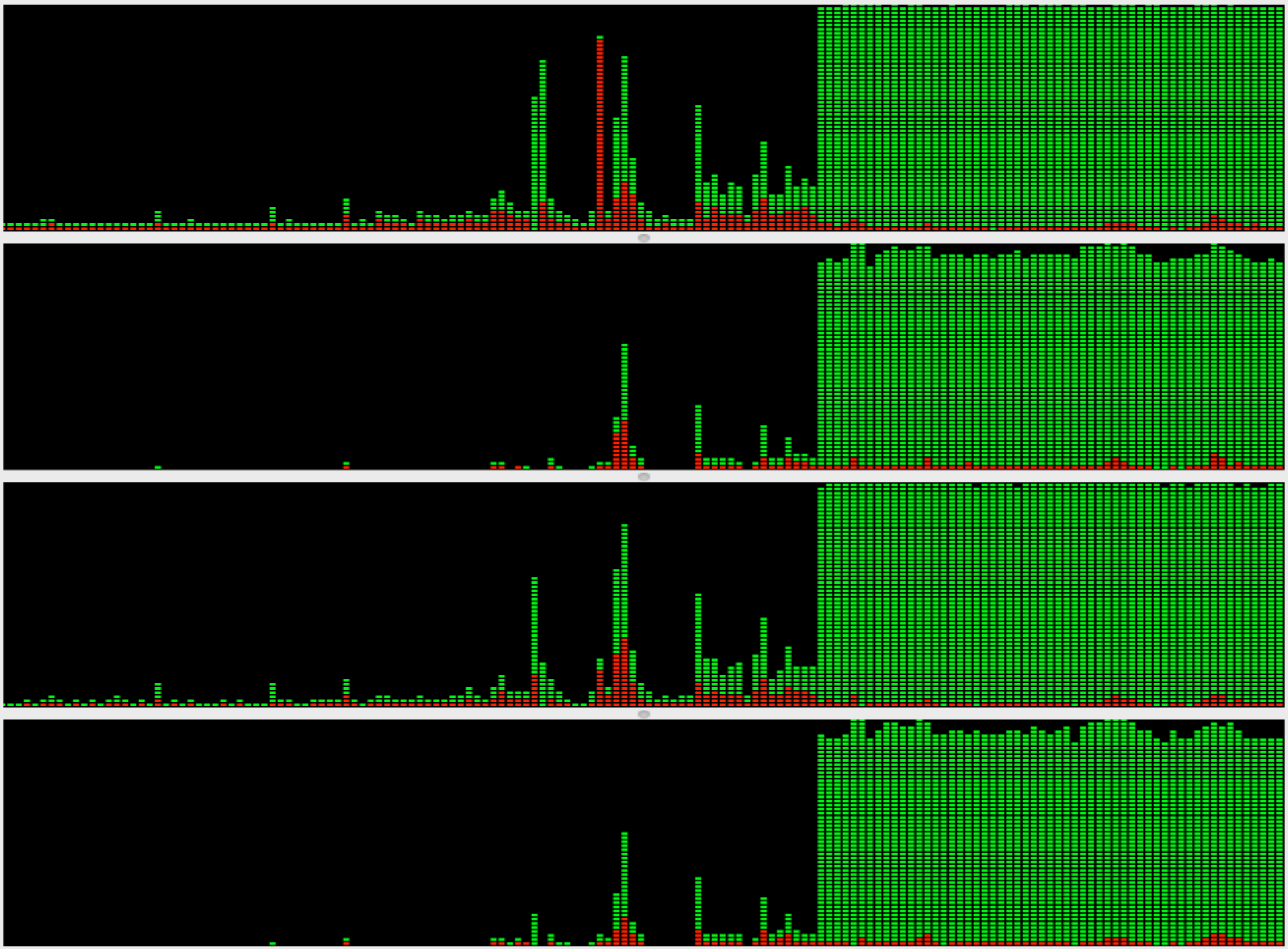
SpatialSpark: Open Source



Runing Code into Local Apache Spark

```
MacBook-Pro-de-Abel:~ abxda$  bin/spark-submit --driver-memory 1800M
 --executor-memory 1800M  --jars /Users/abxda/Development/BigData/Spar
k/GeoSpark-0.1-GeoSpark.jar --class spatialspark.main.SpatialJoinApp
 /Users/abxda/Development/BigData/Spark/SpatialSpark-master/target/sca
la-2.10/spatial-spark_2.10-1.1-SNAPSHOT.jar --left /Users/abxda/Devel
opment/BigData/Spark/tuits_32_ent_mun.tsv --geom_left 2 --right /User
s/abxda/Development/BigData/Spark/VORONOI_MINI_ZAC.csv --geom_right 0
 --output /Users/abxda/Development/BigData/Spark/salidaZacatecas.tsv
 --predicate within --parallel_part true --method stp --conf 32:32:0.1
  --partition 100
```

Historial de la CPU



DENUE - Twitter

Actividad Económica	Total de Tuits
Comercio al por menor en tiendas de abarrotes, ultramarinos y misceláneas	7,741,777
Salones y clínicas de belleza y peluquerías	2,378,443
Comercio al por menor de artículos de papelería	1,470,637
Restaurantes con servicio de preparación de alimentos a la carta o de comida corrida	1,445,646
Comercio al por menor en minisupers	1,246,794
Restaurantes con servicio de preparación de antojitos	1,088,333
Restaurantes con servicio de preparación de tacos y tortas	1,080,876
Reparación mecánica en general de automóviles y camiones	983,615
Asociaciones y organizaciones religiosas	967,601
Cafeterías, fuentes de sodas, neverías, refresquerías y similares	901,126
Comercio al por menor de ropa, excepto de bebé y lencería	691,736
Consultorios dentales del sector privado	664,654
Panificación tradicional	612,145
Lavanderías y tintorerías	592,689
Servicios de acceso a computadoras	564,462
Elaboración de tortillas de maíz y molienda de nixtamal	556,633
Fabricación de productos de herrería	515,880
Restaurantes con servicio de preparaciónn de pizzas, hamburguesas, hot dogs y pollos r	509,583
Hoteles con otros servicios integrados	488,394
Restaurantes que preparan otro tipo de alimentos para llevar	476,261
Alquiler sin intermediación de salones para fiestas y convenciones	446,707
Banca múltiple	445,290
Comercio al por menor de cerveza	408,057
Comercio al por menor en ferreterías y tlapalerías	405,750
Escuelas de educación preescolar del sector privado	401,189
Administración pública en general	385,663
Centros de acondicionamiento físico del sector privado	366,161
Hojalatería y pintura de automóviles y camiones	357,171
Reparación y mantenimiento de otros artículos para el hogar y personales	353,976
Servicios de preparación de otros alimentos para consumo inmediato	351,555
Comercio al por menor de dulces y materias primas para repostería	338,091
Escuelas del sector privado que combinan diversos niveles de educación	329,189
Servicios de contabilidad y auditoría	327,615

BIGDATA
INEGI

abel.coronado@inegi.org.mx

@abxda