CCN

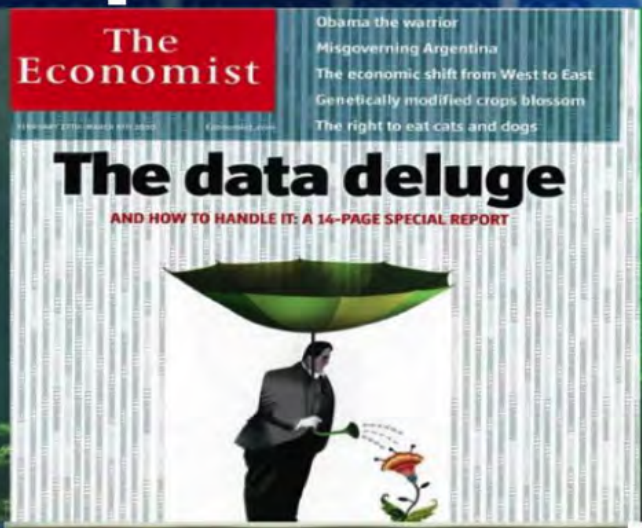
JG
INEGI

Junio, 2015

Eduardo Sojo Garza Aldape
Presidente de la Junta
de Gobierno del INEGI

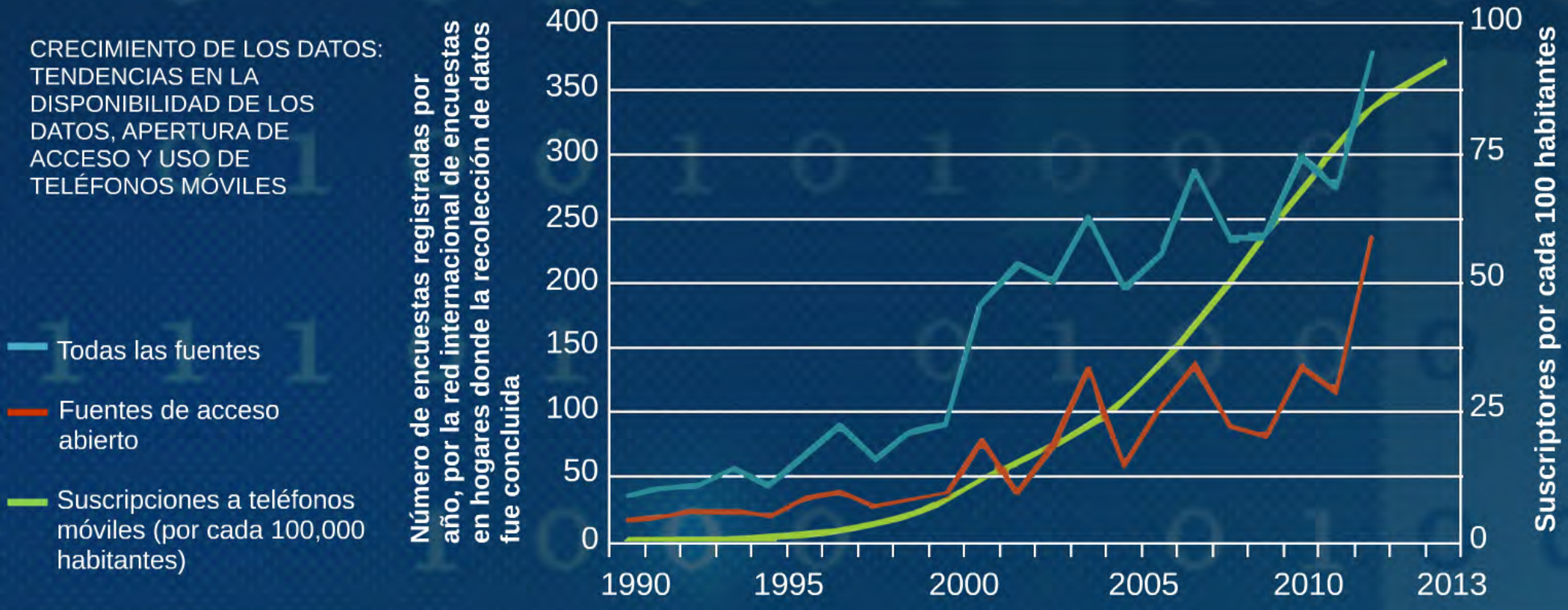
La revolución de los datos y el INEGI

Un tema del que se habla mucho



La información está aumentando

CRECIMIENTO DE LOS DATOS: TENDENCIAS EN LA DISPONIBILIDAD DE LOS DATOS, APERTURA DE ACCESO Y USO DE TELÉFONOS MÓVILES



Y lo seguirá haciendo

El mercado Big Data está estimado a crecer 45% anual hasta alcanzar 25 mil millones de dólares para 2015

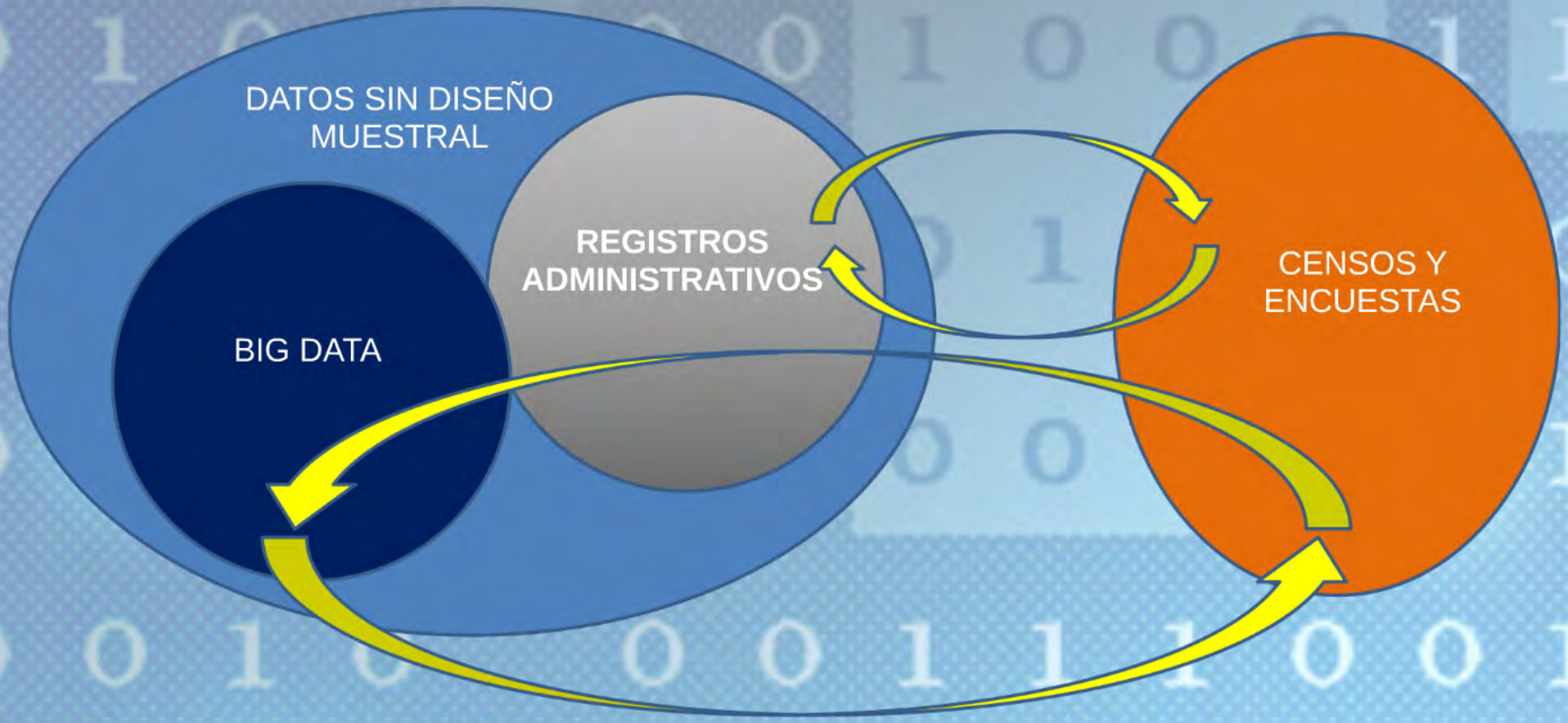
Diluvio de datos

- Para el año 2000, 25% del total de la información almacenada era digital; mientras que para 2015, cerca de 99% es digital.
- La cantidad de información análoga ha seguido creciendo.
- Más de 90% de la información en el mundo ha sido creada en los últimos dos años.
- El proceso se sigue acelerando.

La revolución de los datos

- Afecta a personas, los negocios, los gobiernos, las organizaciones sociales y el resto de las instituciones.
- Incide de manera importante en la manera como vivimos, trabajamos y pensamos.
- Para las estadísticas oficiales implica una transformación de fondo respecto del papel que les toca jugar en la sociedad y la economía: **más administradores y menos productores.**

BIG DATA Y ESTADÍSTICA OFICIAL

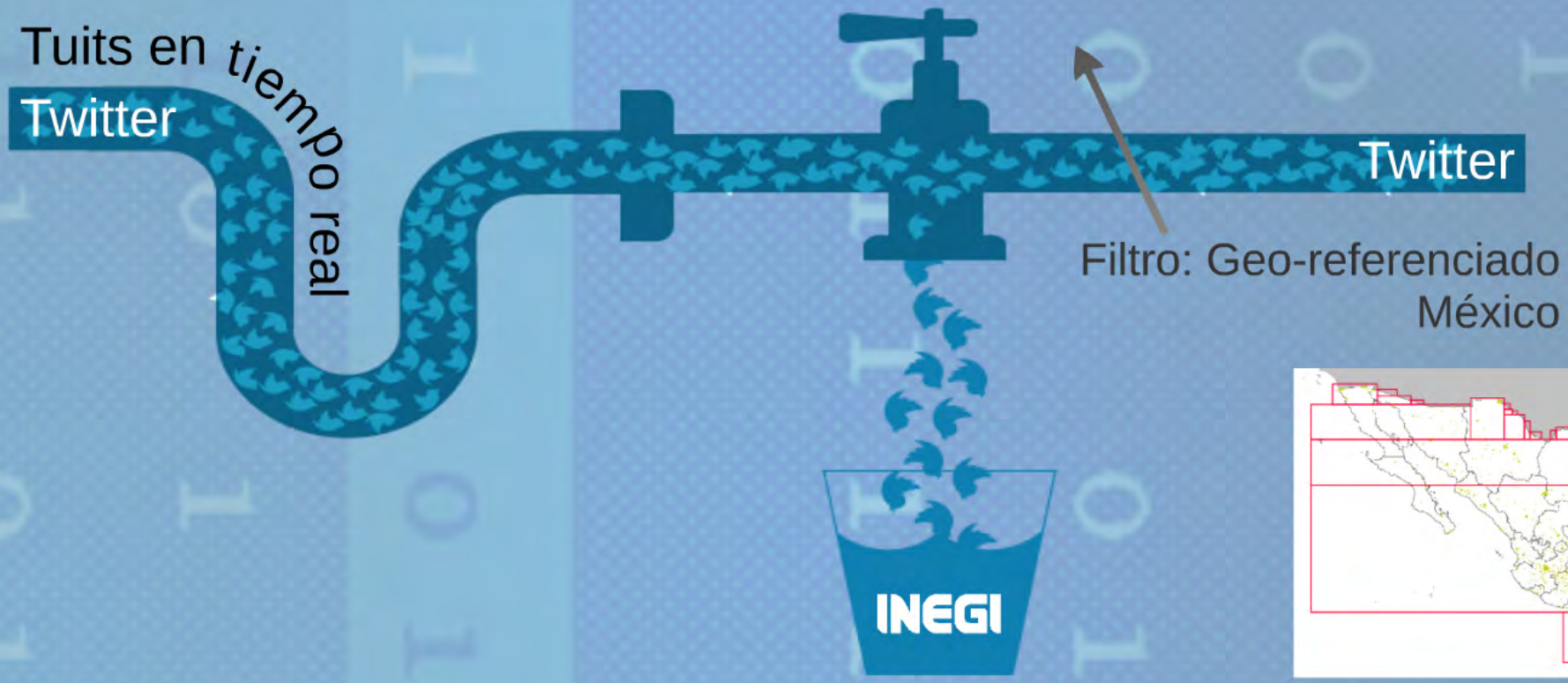


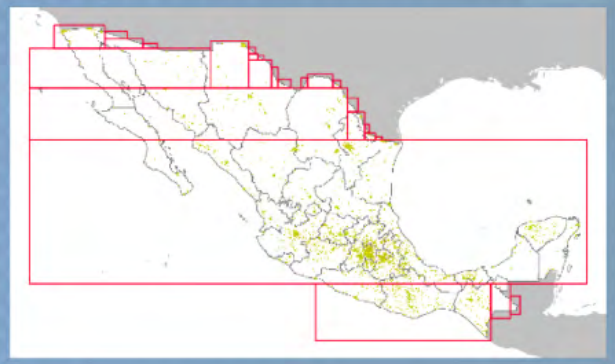
¿Qué estamos haciendo en el INEGI?

PROYECTO PILOTO:

EXPLOTACIÓN DEL TWITTER

Twitter como fuente de datos





Base de Datos NoSQL

TWITTER PARA BIENESTAR SUBJETIVO

En colaboración con:

UNIVERSIDAD TECMILENIO®

Innovación con propósito de vida.

INFOTEC

data2x

partnering for a gender data revolution

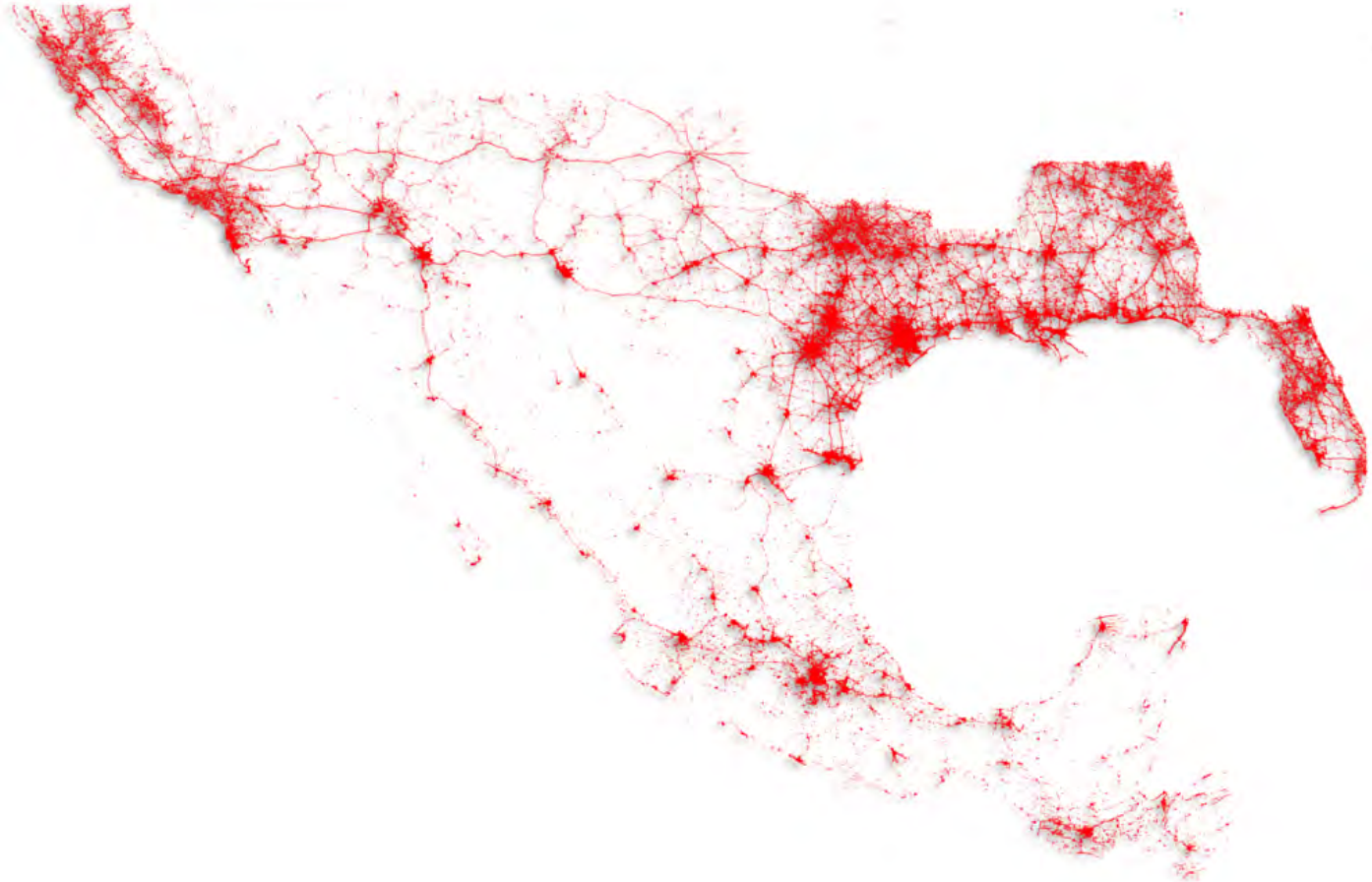
Penn

UNIVERSITY of PENNSYLVANIA

CENTROGEO

CIDE

30´000,000 de tuits georreferenciados



INEGI
INSTITUTO NACIONAL DE ESTADÍSTICA Y GEOGRAFÍA

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

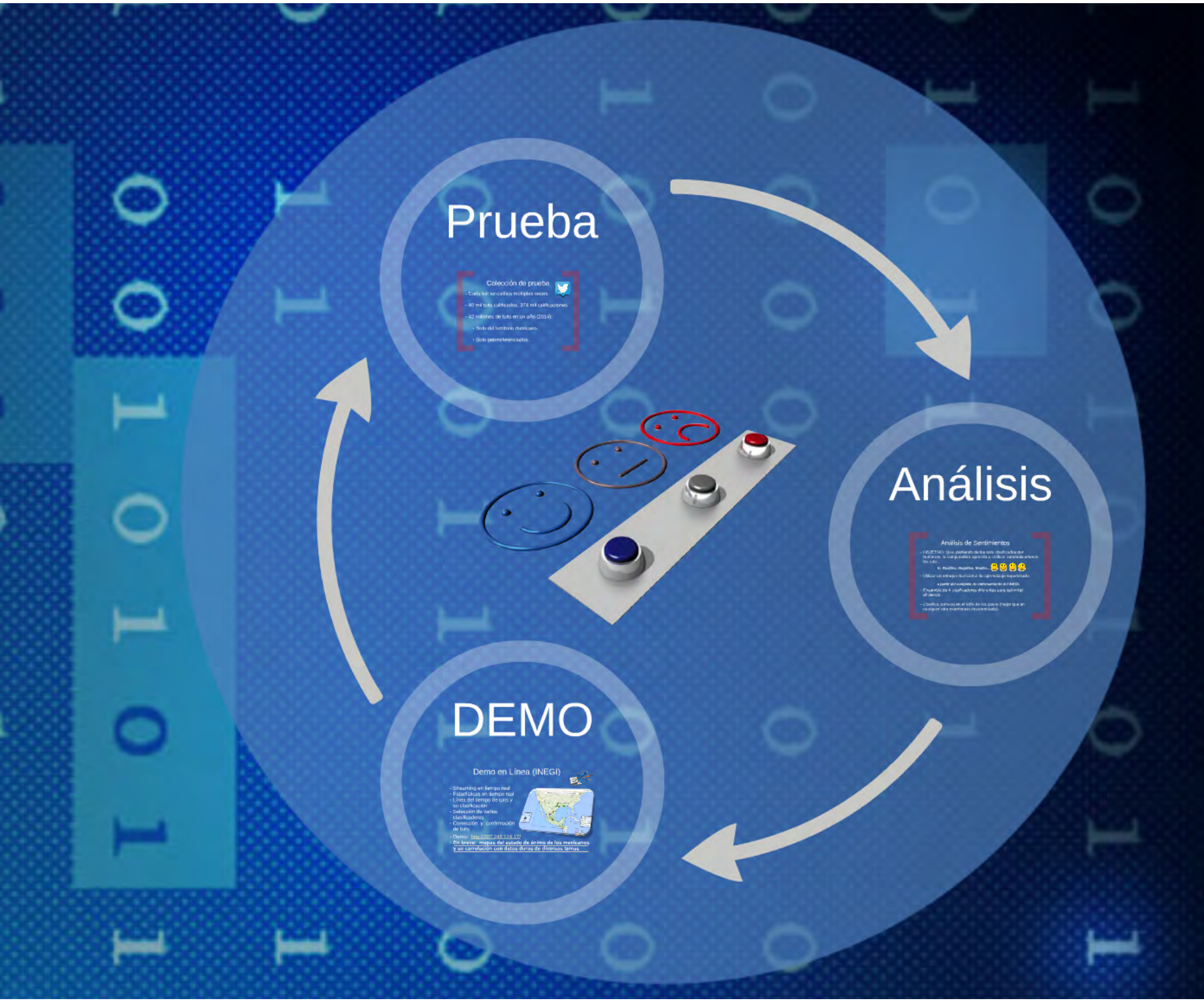
Para generar nuestro conjunto de entrenamiento se desarrolló una aplicación

Califica el sentimiento de los tuits en positivo, negativo o neutro, y clasifícalos en varios temas

PIOANALISIS

Universidad Tec Milenio apoya con estudiantes de sus Campus en muchos Estados del País para calificar el conjunto de entrenamiento

Bienestar Subjetivo y las redes sociales como fuente de datos

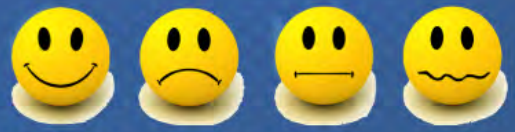

Prueba
Colección de prueba
- Cada tuit se califica múltiples veces.
- 40 mil tuits calificados, 374 mil calificaciones
- 42 millones de tuits en un año (2014):
• Solo del territorio mexicano.
• Solo georeferenciados.
Análisis
Análisis de Sentimientos
DEMO
Demo en Línea (INEGI)
- Streaming en tiempo real
- Estadísticas en tiempo real
- Línea del tiempo de tuits y su clasificación
- Selección de varias clasificaciones
- Corrección y confirmación de tuits
- Demo: http://207.248.124.17/
- En breve: mapas del estado de ánimo de los mexicanos y su correlación con datos duros de diversos temas

Colección de prueba

- Cada tuit se califica múltiples veces.
- 40 mil tuits calificados, 374 mil calificaciones.
- 42 millones de tuits en un año (2014):
  - Solo del territorio mexicano.
  - Solo georreferenciados.

Análisis de Sentimientos

- OBJETIVO: Que, partiendo de los tuits clasificados por humanos, la computadora aprenda a calificar automáticamente los tuits.

  Ej. ***Positivo, Negativo, Neutro...*** 

- Utilizar un enfoque tradicional de aprendizaje supervisado

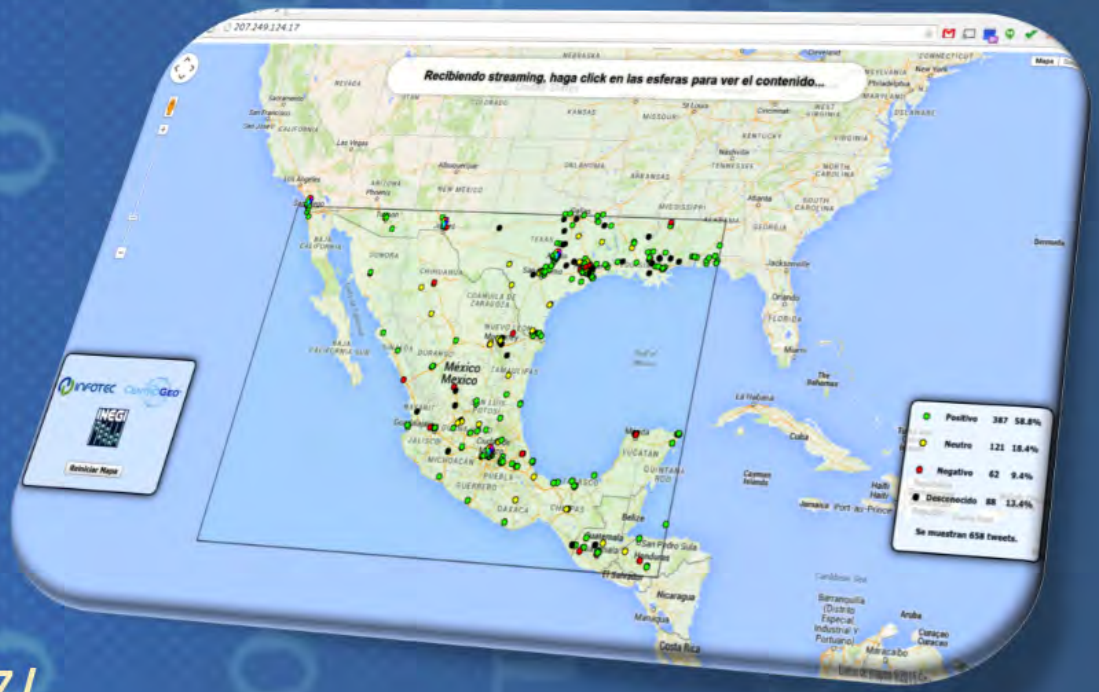
  a partir del conjunto de entrenamiento del INEGI.

- Ensamble de 4 clasificadores diferentes para optimizar eficiencia.

- Clasifica correcta en el 80% de los casos (mejor que en cualquier otra experiencia documentada).

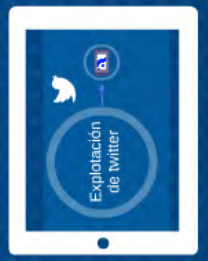
Demo en Línea (INEGI)

- Streaming en tiempo real
- Estadísticas en tiempo real
- Línea del tiempo de tuits y su clasificación
- Selección de varios clasificadores
- Corrección y confirmación de tuits



- Demo: http://207.249.124.17/
- **En breve: mapas del estado de ánimo de los mexicanos y su correlación con datos duros de diversos temas**

Twitter para Turismo



Twitter para Turismo

Antecedentes:

- En colaboración con SECTUR
- En México, el 95% de los usuarios de Twitter tienen entre 18 y 49 años.
- El 87% de los turistas en destinos de ciudad tienen entre 18 y 55 años de edad.

Twitter para Turismo

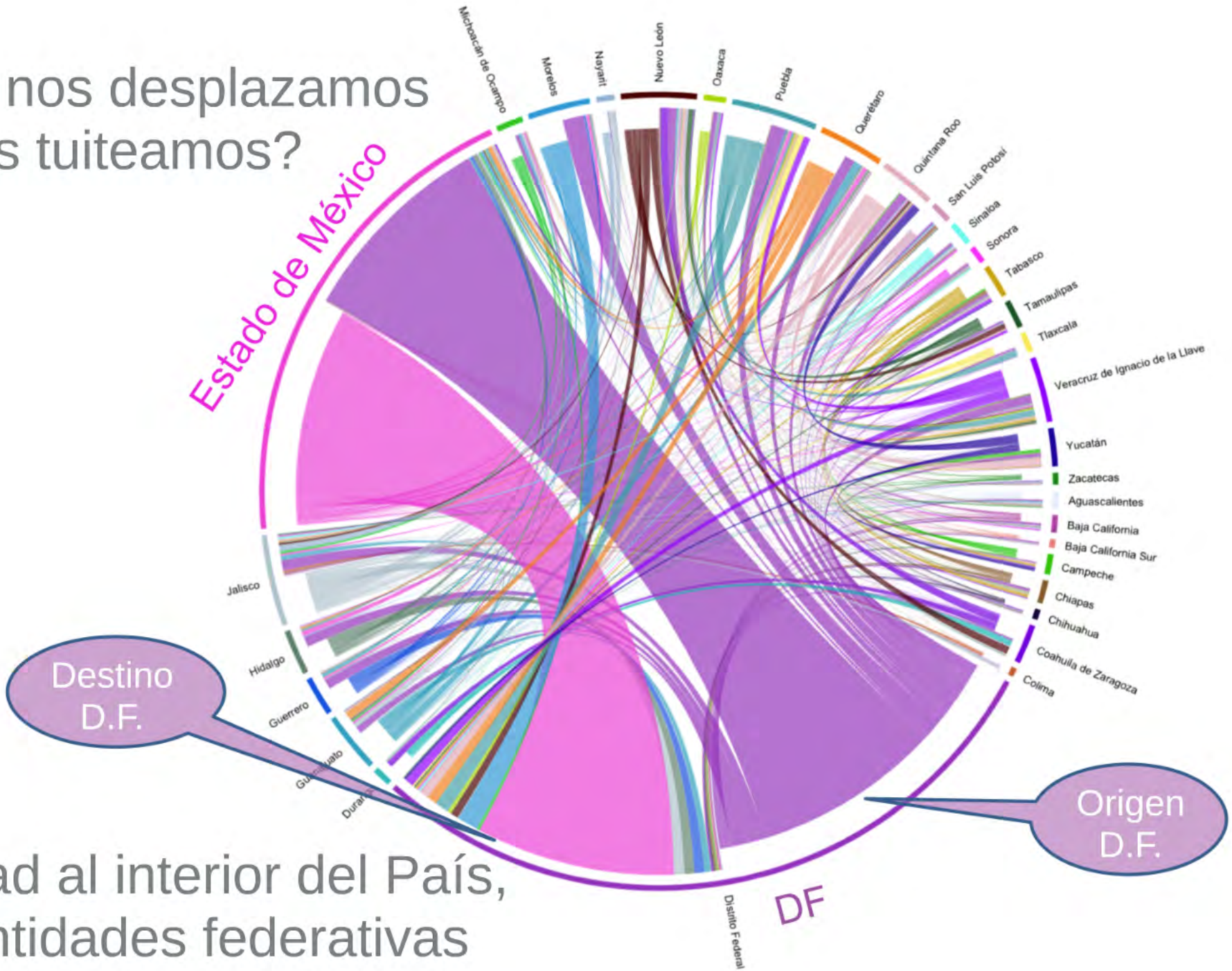
Generación de estadísticas turísticas con Twitter

- Se procesaron 60 millones de tuits (enero a julio 2014)
- Se seleccionaron los estados de Guanajuato y Puebla, los días 1, 2 y 3 de febrero
- Resultaron 7,955 tuiteros de dicha selección
- Se buscaron todos sus tuits durante los 6 meses
- Resultaron 827,424 tuits
- Se detectó cronológicamente desde qué entidad federativa habían tuiteado y los días de permanencia en cada estado
- Con los resultados se generaron un mapa para Guanajuato y otro para Puebla

Twitter para el turismo

TWITTER PARA MOVILIDAD

¿Cómo nos desplazamos mientras tuiteamos?



Movilidad al interior del País, entre entidades federativas

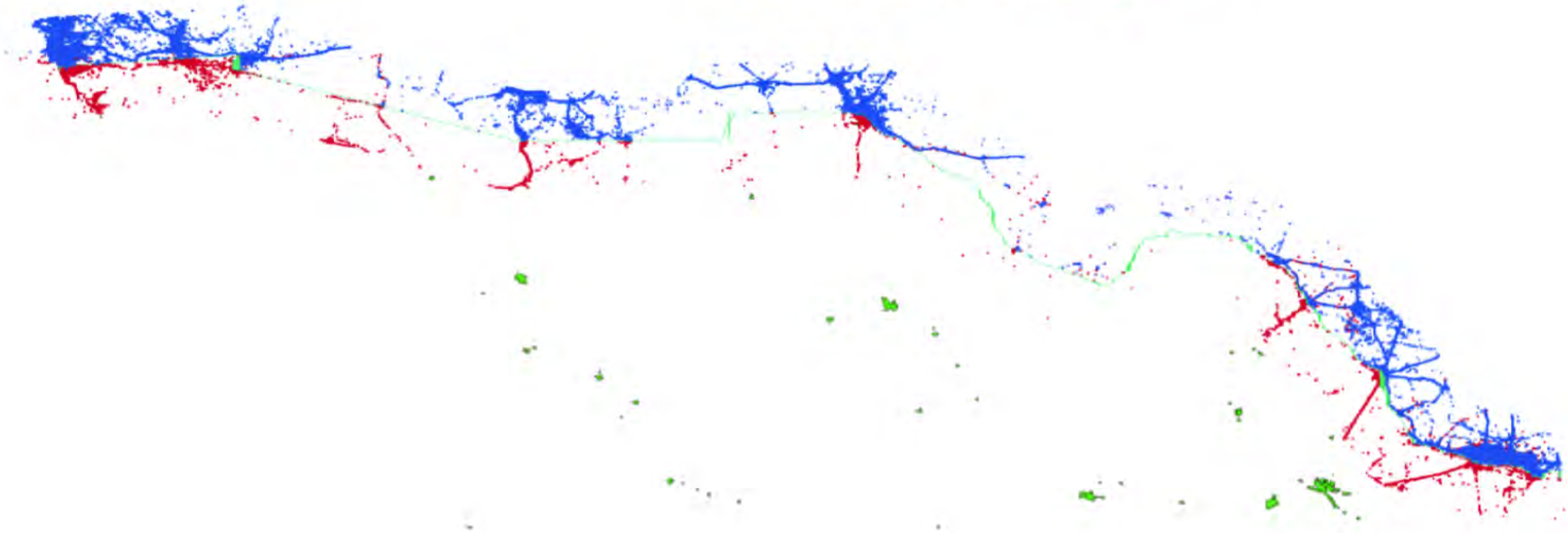
Twitter para Movilidad

¿Dónde tuiteamos?

Actividad de los tuiteros en la frontera

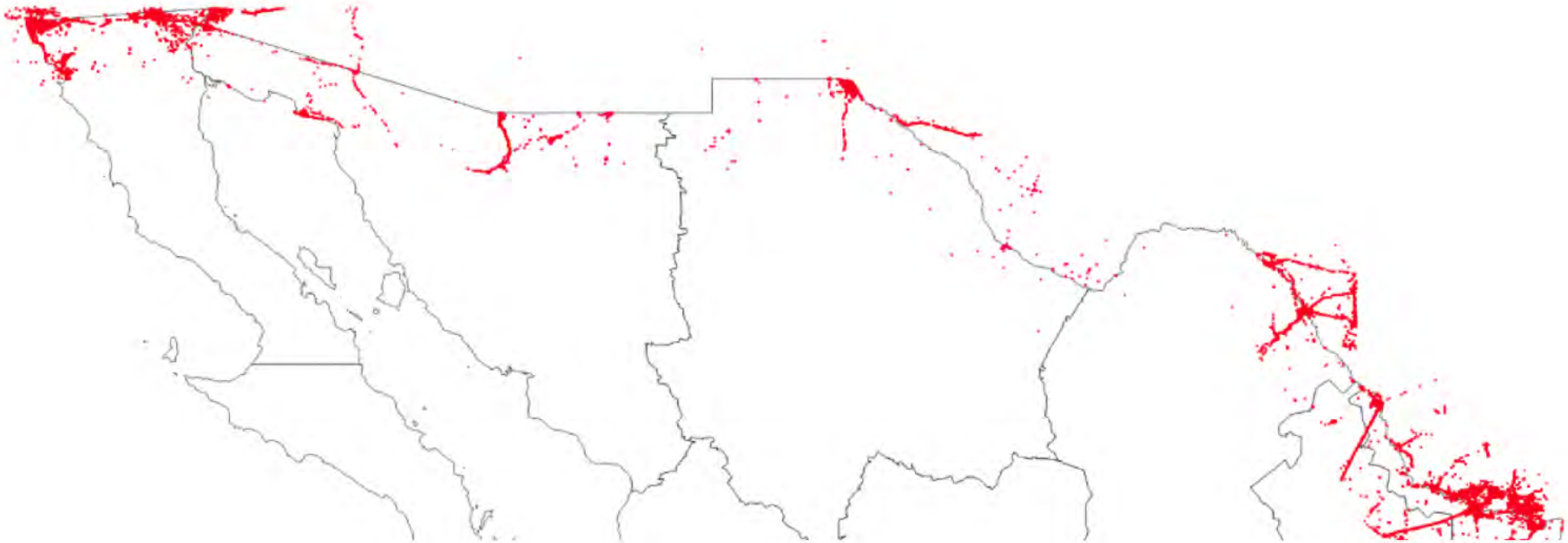
Azul= tuiteros de origen EUA

Rojo = tuiteros de origen MX



Twitter para Movilidad

Actividad solamente de tuiteros MX



Exploración para el desarrollo de una metodología de análisis para medir la movilidad transfronteriza con los tuits georreferenciados

Conclusiones

Conclusiones

- La revolución de los datos es un fenómeno en marcha, que se extiende exponencialmente alrededor del mundo, afectando muchas dimensiones de las vidas de las personas, de los negocios y de las instituciones.
- Toca en el tema central de la actividad de instituciones como el INEGI: **la información**.
- El reto es domar las fuerzas desatadas por esa revolución para incorporarlas a la generación de estadística oficial.
- Se trata de un reto a la vez formidable y apasionante.
- El INEGI ha buscado subirse a la cresta de la ola y comenzado a trabajar con BIG DATA, por lo pronto de manera experimental.
- En ello, se ha privilegiado el sentido práctico y la cooperación con otras instituciones de investigación nacionales y extranjeras, así como con organismos internacionales.

CCN

JG
INEGI

Junio, 2015

Eduardo Sojo Garza Aldape
Presidente de la Junta
de Gobierno del INEGI