Cátedra INEGI

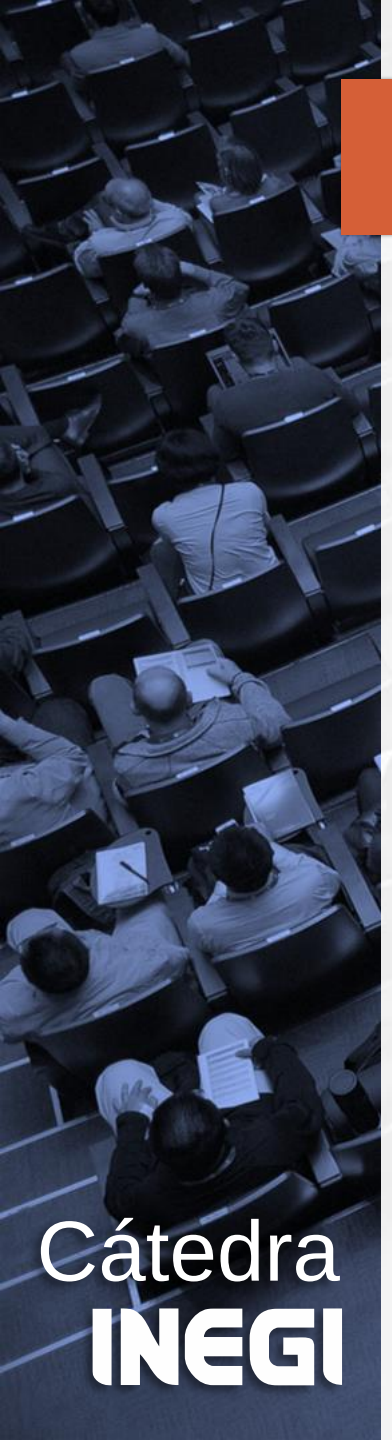
Estado de ánimo de los tuiteros en México.

INSTITUTO NACIONAL
DE ESTADÍSTICA Y GEOGRAFÍA

Universidad Veracruzana

Que es Big Data??

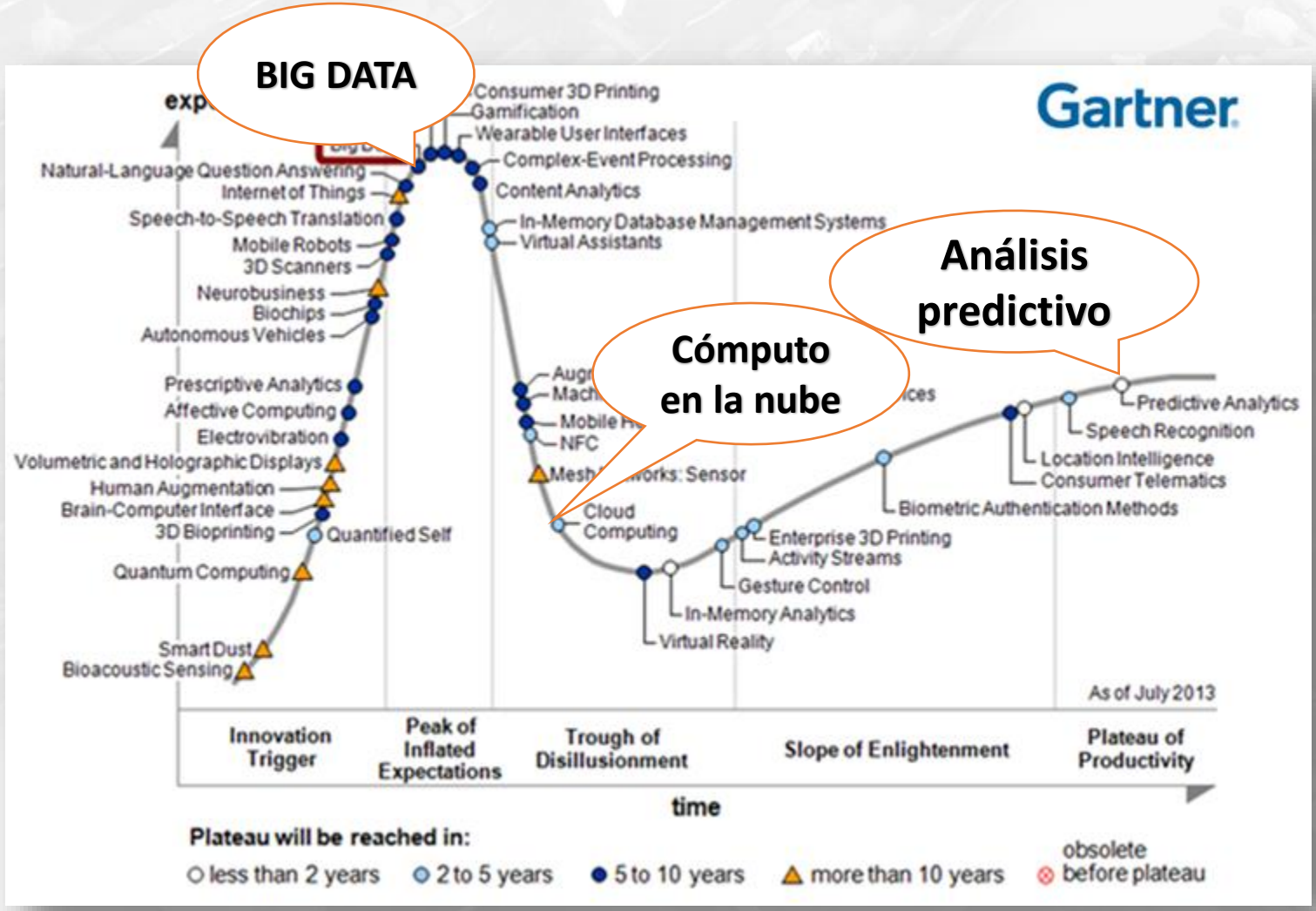
Cátedra INEGI



EL PARADIGMA BIG DATA

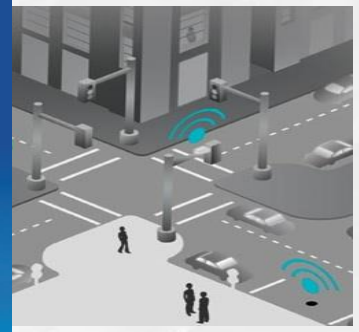
- Dejar que los datos hablen.
- N=todo (¿Mucho?).
- Granularidad constante.
- No hay diseño, ni pregunta, ni muestreo pues los datos no son escasos.
- Correlación más que causalidad.
- Menos exactitud y mucha paja o ruido.

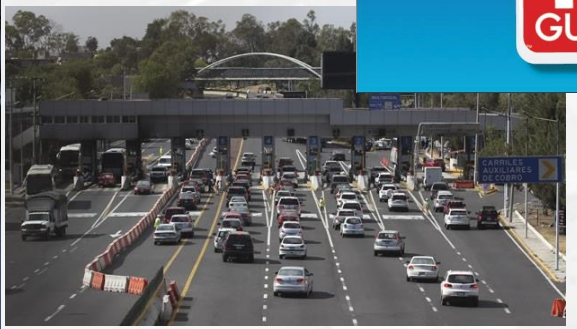
Posicionamiento de Big Data en el Ciclo de las Expectativas de Gartner



Cátedra INEGI

Las fuentes de información están creciendo y evolucionando…



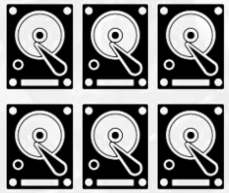


Cátedra
INEGI

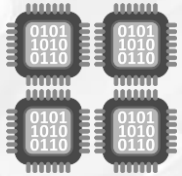
CONOCIMIENTO

Expertos marcan el rumbo.

Infraestructura Informática



Sistemas Distribuidos



Computo Paralelo y Concurrente

Estadística Matemática

Análisis de Datos

Statistical Learning

Minería de Datos

Estratificaciones

Y mucho más…

Muestreo

Análisis de Redes (Grafos)

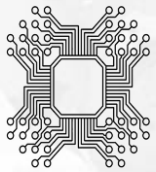
Machine Learning

Análisis de Regresión

BIG DATA:

Internet de las cosas

Sensores



Internet de las Personas

Redes Sociales

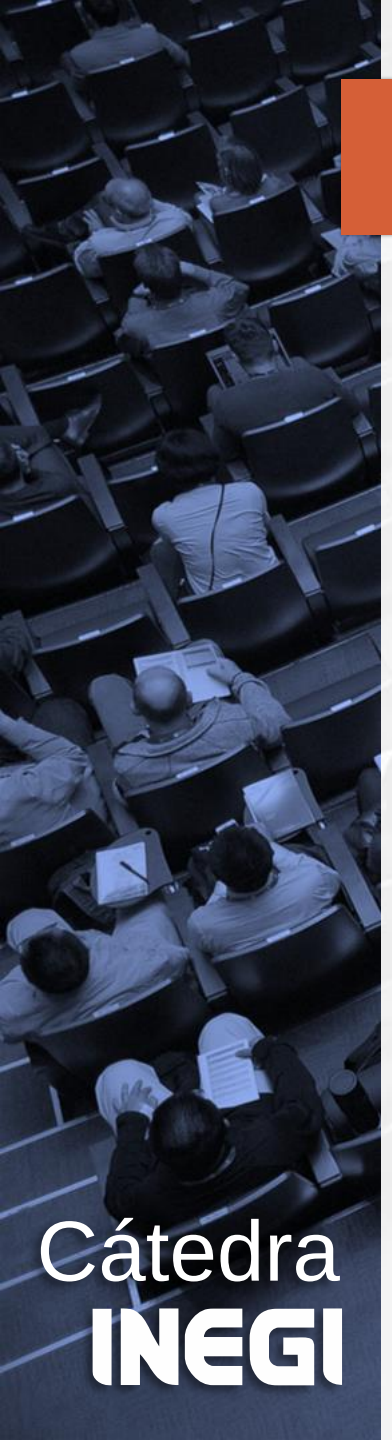
QUÉ SIGNIFICA “BIG DATA” PARA LA INFORMACIÓN OFICIAL?

Cátedra INEGI

INSTITUTO NACIONAL DE ESTADÍSTICA Y GEOGRAFÍA

BIG DATA Y ESTADÍSTICA OFICIAL

Cátedra INEGI

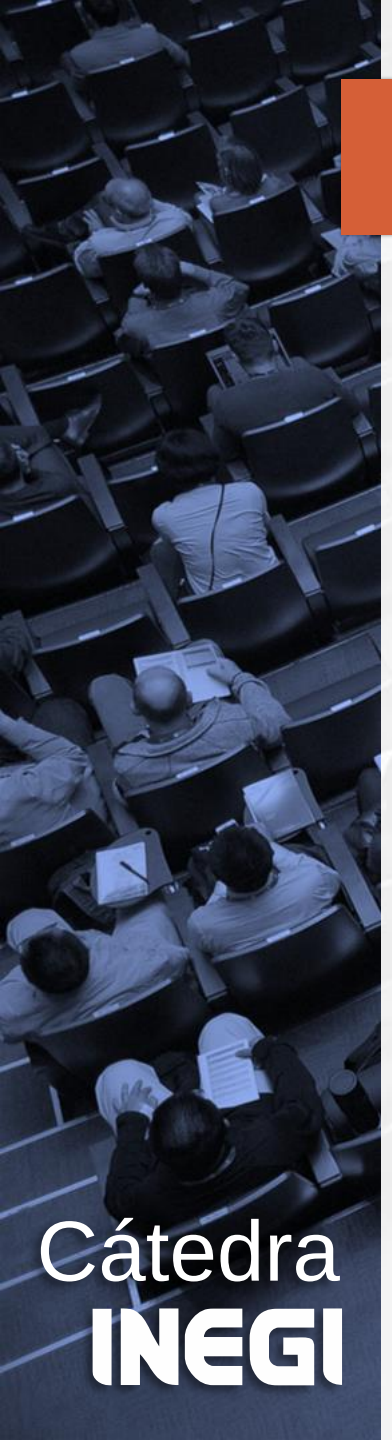


Posibles fuentes de “Big Data”

- Administrativas (gubernamentales o no);
- Comerciales o transaccionales;
- Sensores;
- De desplazamiento;
- De comportamiento;
- Opinión;

Retos

- Legislativos,
- Privacidad,
- Financieros,
- Administrativos,
- Metodológicos,
- Tecnológicos.



3 ámbitos abiertos a experimentación:

- Corto plazo: Combinación de Big data con estadísticas oficiales.
- Mediano plazo: Producción de información relativa a temas emergentes basada en Big Data (en particular, cuando no hay encuesta ni registro).
- Largo plazo: Remplazo de parte de la estadística oficial producida mediante Big Data.

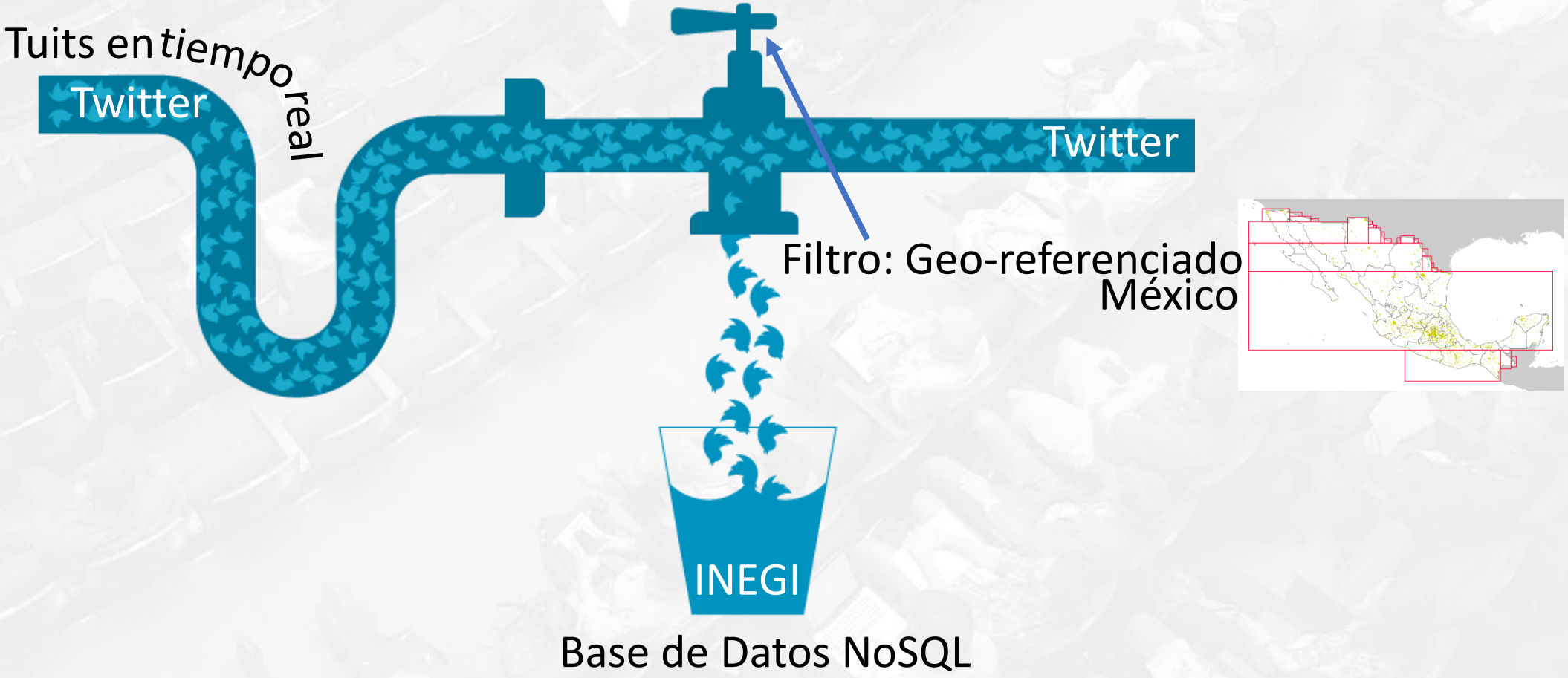
Cátedra INEGI

Qué estamos haciendo en INEGI?

PROYECTO PILOTO: EXPLOTACIÓN DE TWITTER

INSTITUTO NACIONAL DE ESTADÍSTICA Y GEOGRAFÍA

Twitter como fuente de datos



Cátedra INEGI

Por qué Twitter?

- Disponible fácilmente
- Hasta el 1% de los tweets mundiales sin costo
- Alrededor de 12 M de cuentas en México
- 700 mil cuentas con tuits georreferenciados en MX
- 100 millones de tuits descargados desde enero de 2014

Cátedra INEGI

Infraestructura para recolectar tuits

Cátedra INEGI

Prueba de concepto

Cluster (Hydra)

NoSql Database “Elasticsearch”

Unix “Red Hat”

Big Data Layers

Qué dicen los tuits….?

TEMAS DE ESTUDIO

- Bienestar Subjetivo a partir de Redes Sociales
- Flujos Turísticos
- Movilidad
- Elecciones 2018

TWITTER PARA BIENESTAR SUBJETIVO

Cátedra INEGI

INSTITUTO NACIONAL DE ESTADÍSTICA Y GEOGRAFÍA

Proceso de análisis de tuits:

- Se utiliza un **Método supervisado de clasificación**:
  - Humanos califican el sentimiento y clasifican el tema de un conjunto de tuits (conjunto de entrenamiento)
  - El conjunto de entrenamiento se utiliza para que la máquina “aprenda” a calificar y clasificar por similitudes nuevos tuits

INSTITUTO NACIONAL DE ESTADÍSTICA Y GEOGRAFÍA

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

API para desarrollar conjunto de entrenamiento

PIOANALISIS

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

Universidad Tec Milenio apoya con estudiantes en muchos estados del país para producir el conjunto de entrenamiento

Bienestar Subjetivo y las redes sociales como fuente de datos

7 de 20 - nivel 0

Que pague con diamantes su pecado

¿El tema del tuit es?

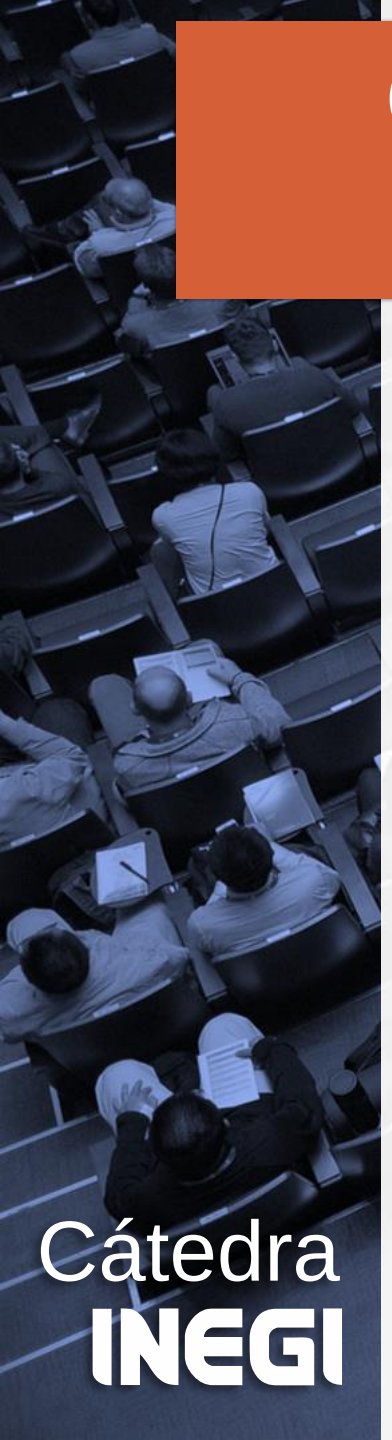
Política | Cultural / Entretenimiento | Deporte | Escolar / Laboral | Personal | Ni idea

¿El tuitero se sentia?

Negativo

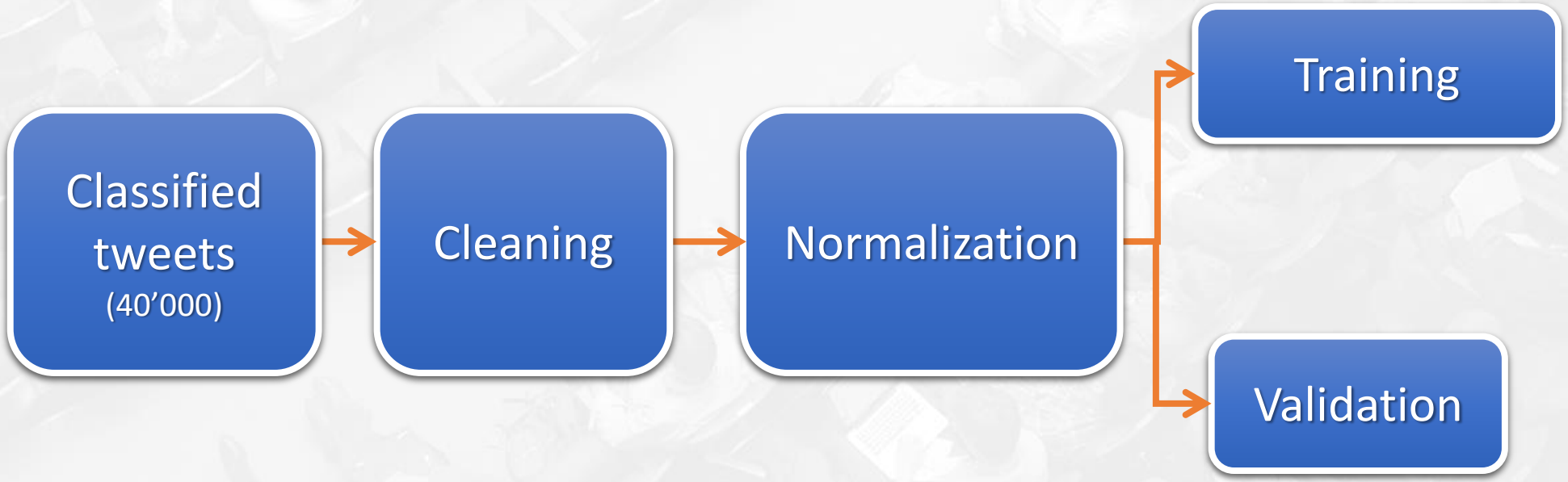
Cátedra INEGI

Colaboración con la academia para clasificar

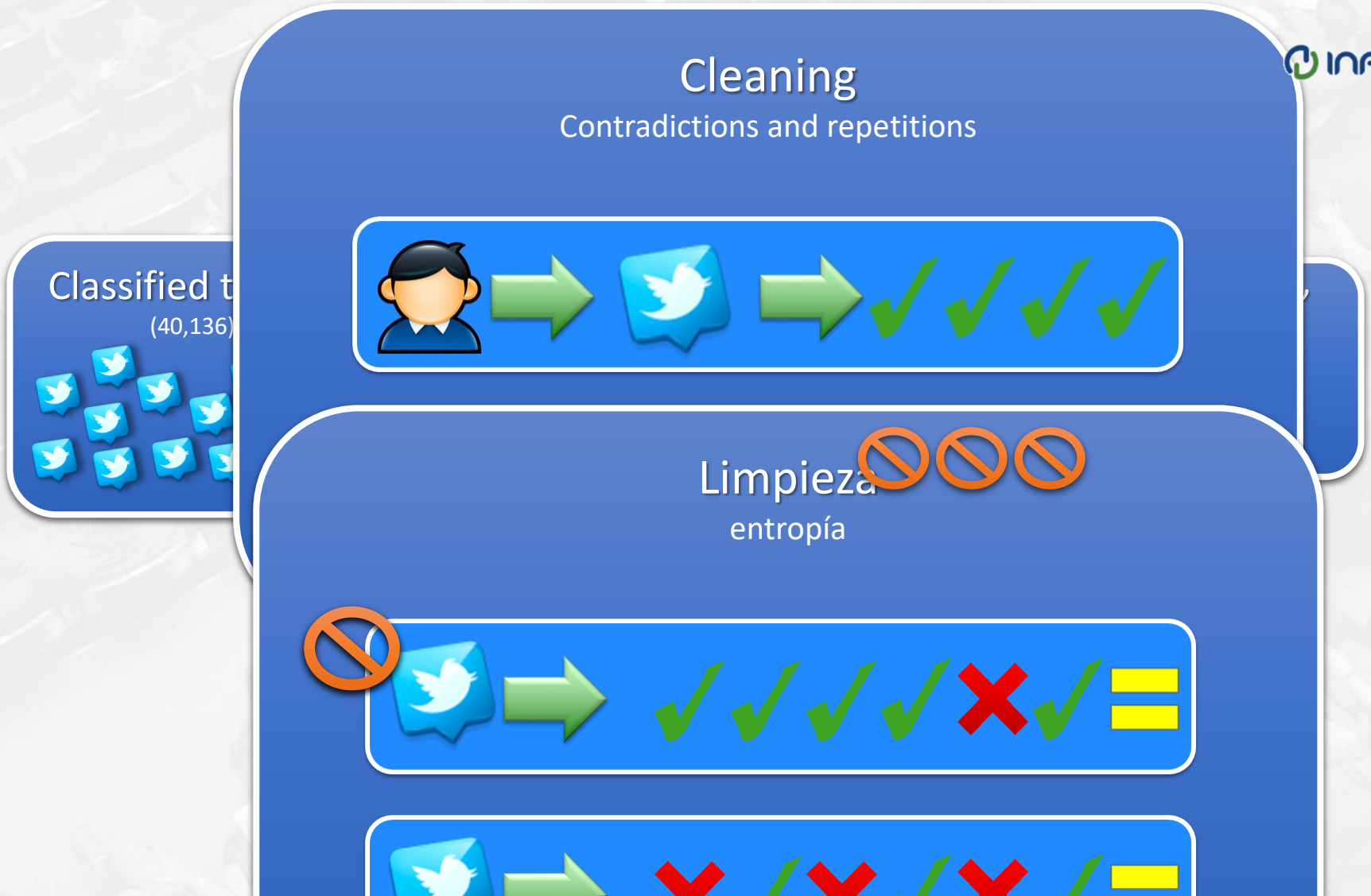


INFOTEC CENTROGEO

La información clasificada se divide en dos conjuntos como sigue:



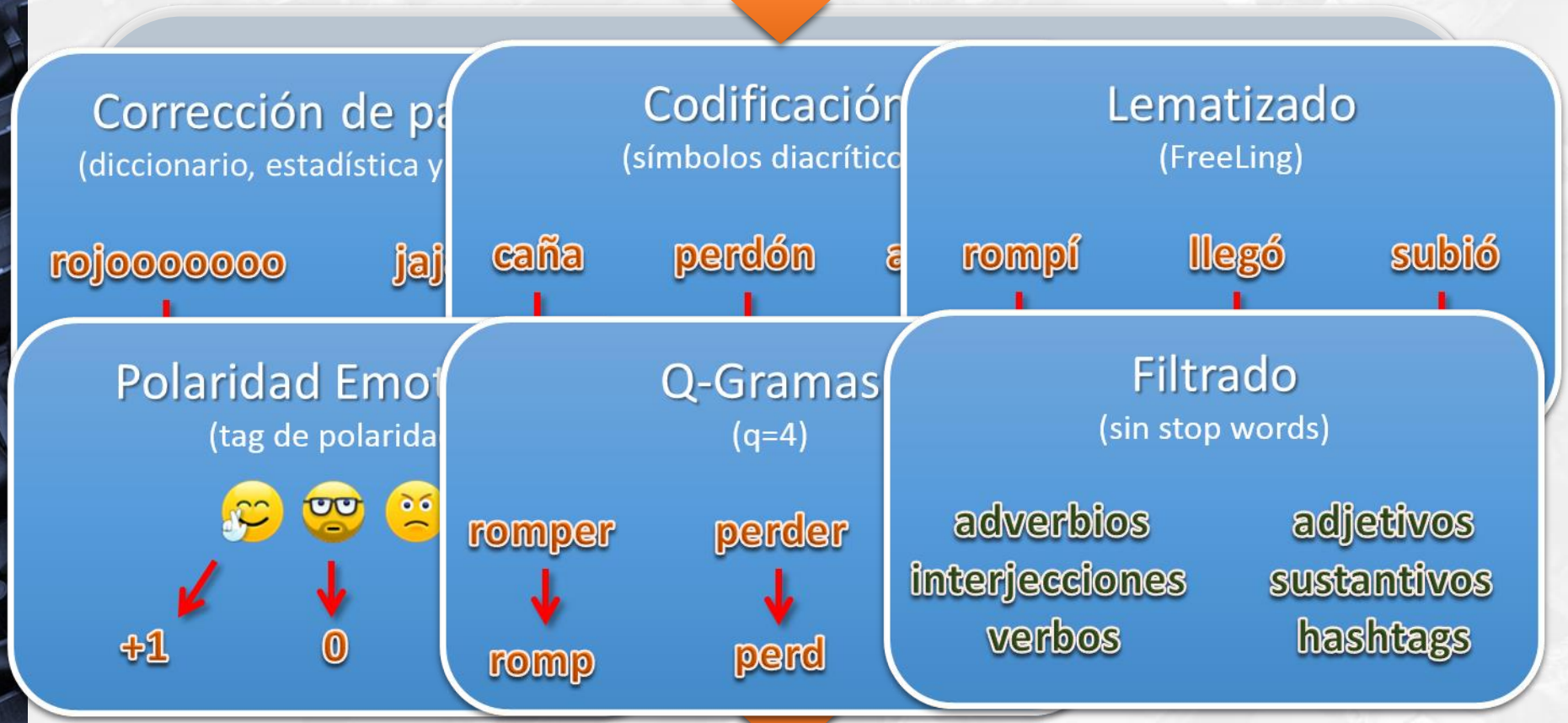
Depuración de cada conjunto clasificado (Limpieza)



Cátedra INEGI

Pre-procesamiento de cada conjunto clasificado (Normalización)

Tuits Limpios



Tuits Normalizados

Cátedra INEGI

Resultados de clasificación

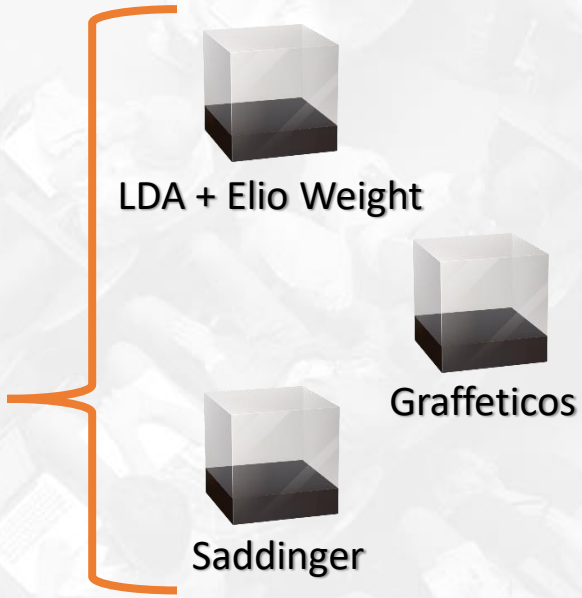
≈ 70% success

INFOTEC CENTROGEO

Tuits "normalizados" (19'163)

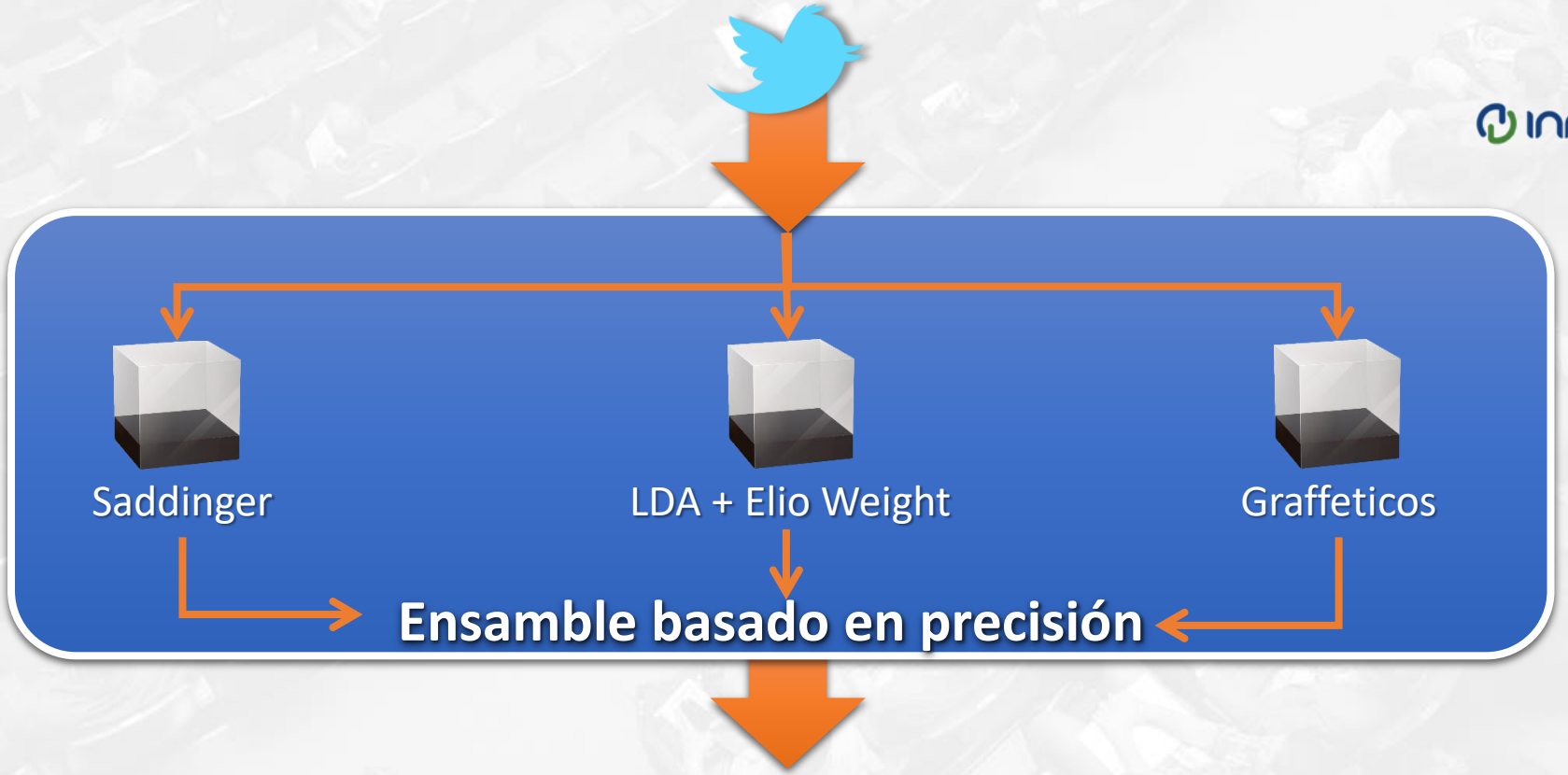
Tuits Entrenamiento (17'163)

Tuits Validación (2'000)



Cátedra INEGI

Mejorar precisión ensamble de clasificadores innovadores



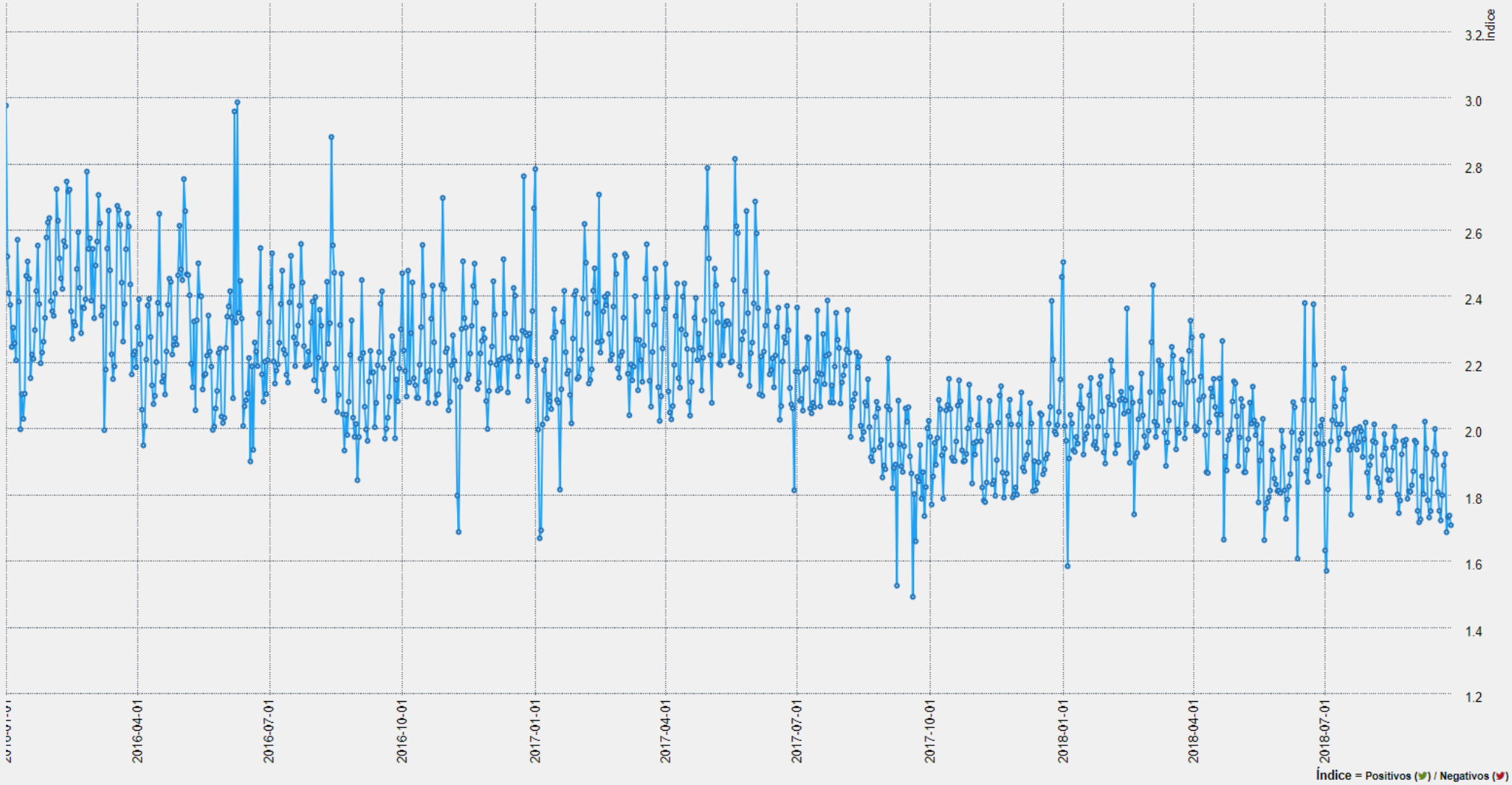
Positivo Negativo

≈ 80% de acierto

Cátedra INEGI

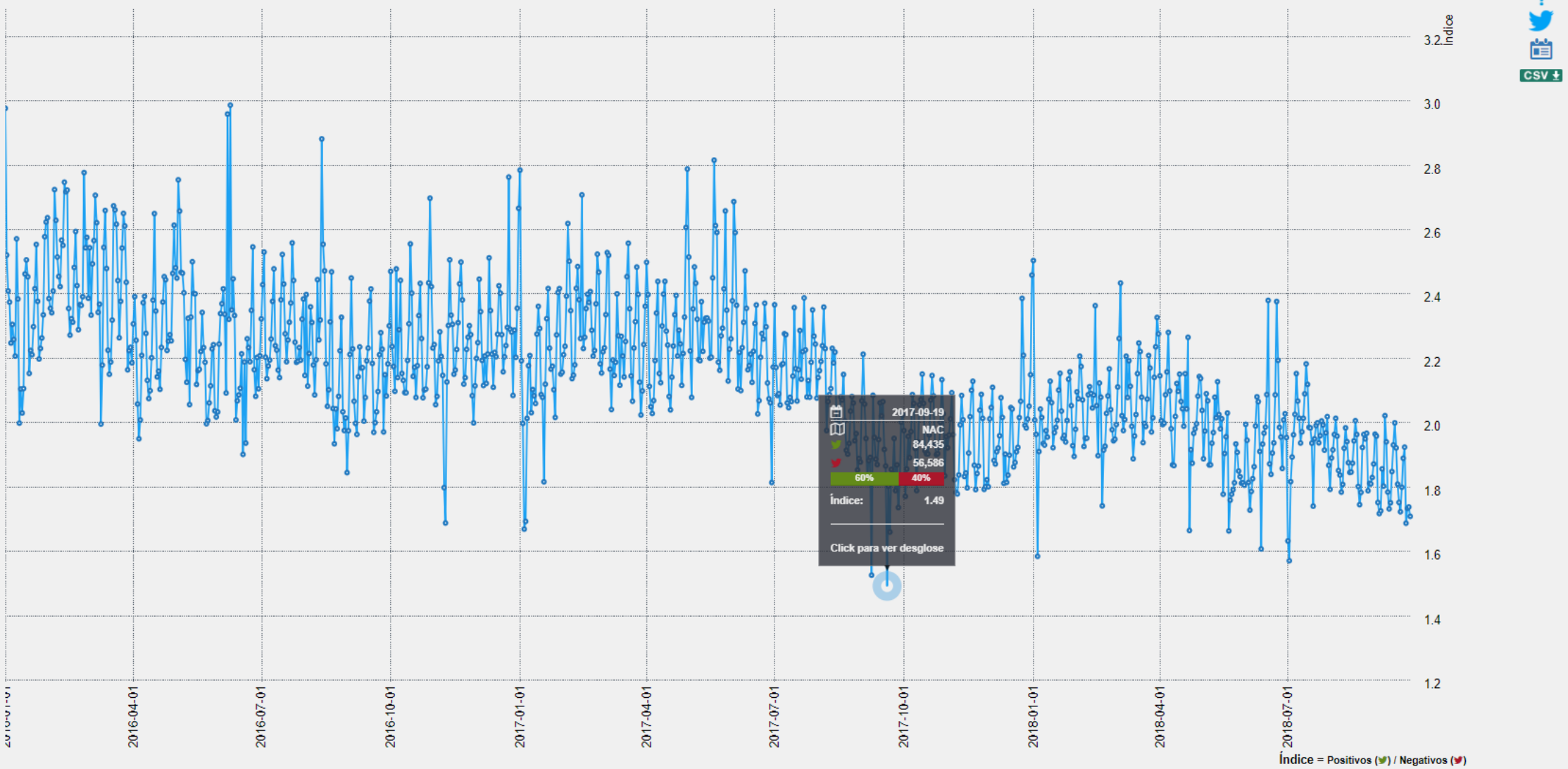
Estado de ánimo de los tuiteros en México

1 de enero de 2016 - 26 de septiembre de 2018



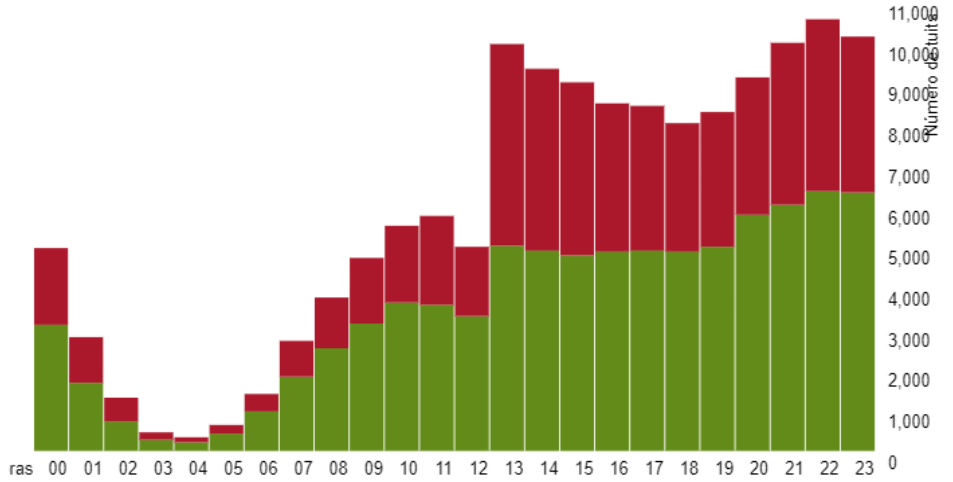
Estado de ánimo de los tuiteros en México

1 de enero de 2016 - 26 de septiembre de 2018



Clasificación por hora del día (NAC)

19 de septiembre de 2017



Clasificación | #Hashtags | Noticias

CSV

Índice = Positivos / Negativos

INEGI

Nube de #hashtags del día (NAC)

19 de septiembre de 2017

Clasificación | #Hashtags | Noticias

Noticias del día (NAC)

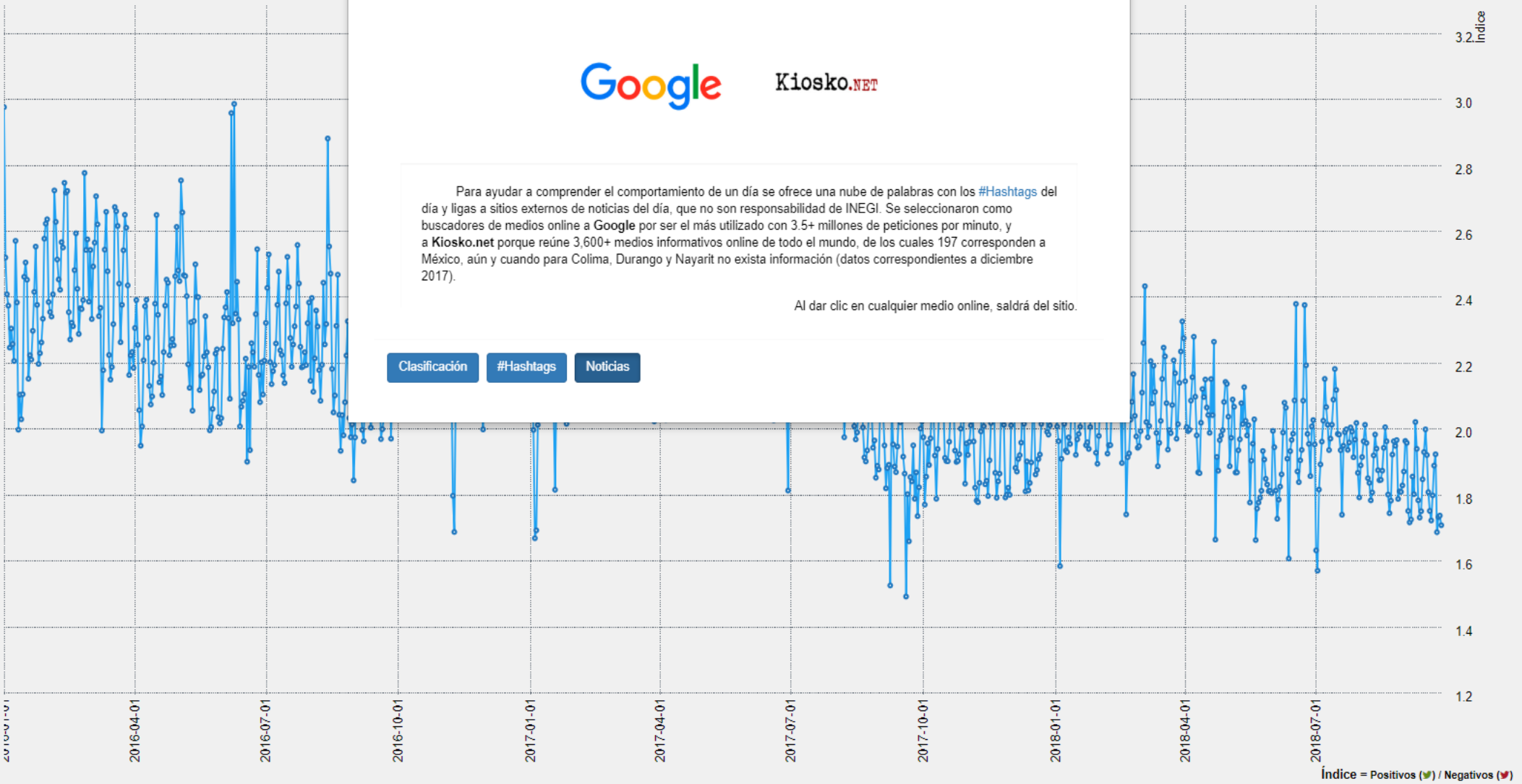
20 de septiembre de 2017

Google Kiosko.NET

Para ayudar a comprender el comportamiento de un día se ofrece una nube de palabras con los #Hashtags del día y ligas a sitios externos de noticias del día, que no son responsabilidad de INEGI. Se seleccionaron como buscadores de medios online a **Google** por ser el más utilizado con 3.5+ millones de peticiones por minuto, y a **Kiosko.net** porque reúne 3,600+ medios informativos online de todo el mundo, de los cuales 197 corresponden a México, aún y cuando para Colima, Durango y Nayarit no exista información (datos correspondientes a diciembre 2017).

Al dar clic en cualquier medio online, saldrá del sitio.

Clasificación #Hashtags Noticias



location:Nacional

Todo | Imágenes | Noticias | Shopping | Maps | Más | Preferencias | Herramientas

País: México | Todas las noticias | 19 de septiembre de 2017 | Ordenados por importancia | Borrar

Otro terremoto en 19 de septiembre
El Universal - 19 sep 2017
mario.dorantes@eluniversal.com.mx. Ni un día más, ni un día menos. Justo hoy 19 de septiembre de 2017 la Ciudad de México volvió a vivir las mismas ...

Aumenta a 369 víctimas por sismo en México
El Financiero - 19 sep 2017
Al menos 369 personas fallecieron tras el sismo de 7.1 de magnitud que el 19 de septiembre sacudió el centro y sur del país. El coordinador nacional de ...

Noticieros Televisa

Fotos del sismo de la Ciudad de México que no pudimos transmitir en ...
Noticieros Televisa - 19 sep 2017
La impotencia ante la tragedia #Sismo #Puebla estas imágenes son muy fuertes pero hablan de la solidaridad pic.twitter.com/YBPTeXY2j3. — Maricela Luna ...

19 de septiembre, otra vez; coincidencias y diferencias entre sismos ...
El Universal - 19 sep 2017
19 de septiembre, otra vez. Fue el 19 de septiembre pero de 1985 cuando un terremoto de 8.1 grados sacudió a la Ciudad de México. Otro sismo este 19 de ...

Así se vivieron las primeras horas del terremoto de 1985
Milenio.com - 19 sep 2017
El 19 de septiembre de 1985 está marcado como una de las fechas más trágicas en la historia de México, pues ese día un terremoto de 8.1 grados Richter ...

Aeropuerto de México reanuda operaciones
El Diario de Juárez - 19 sep 2017
Ciudad de México– Tras el sismo de 7.1 grados, el Aeropuerto Internacional de la Ciudad de México (AICM) reactivó sus operaciones a las 16:00 horas.

INEGI

Kiosko.NET

Periódicos de México
Toda la prensa de hoy

México

Prensa de Información General

Prensa Deportiva

Prensa Económica

Mapa del estado de ánimo de los tuiteros en México

1 de enero de 2016 - 26 de septiembre de 2018

Índice promedio nacional: 2.15

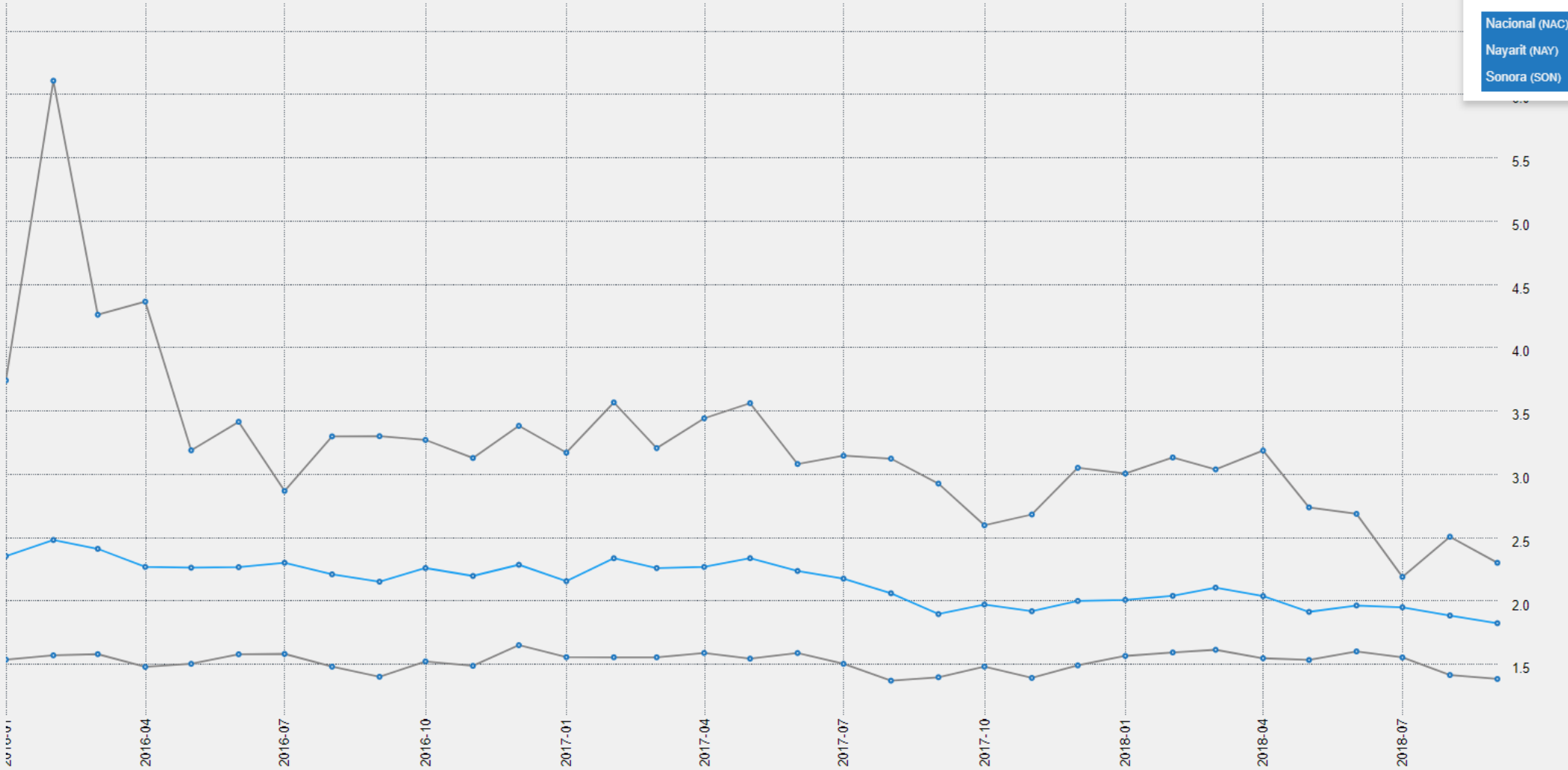
Rango del periodo: [-0.1 , 3.43]

Rango seleccionado: [0 , 3.43]

Índice = Positivos ( ) / Negativos ( )

Estado de ánimo de los tuiteros en México

1 de enero de 2016 - 26 de septiembre de 2018



Índice = Positivos / Negativos

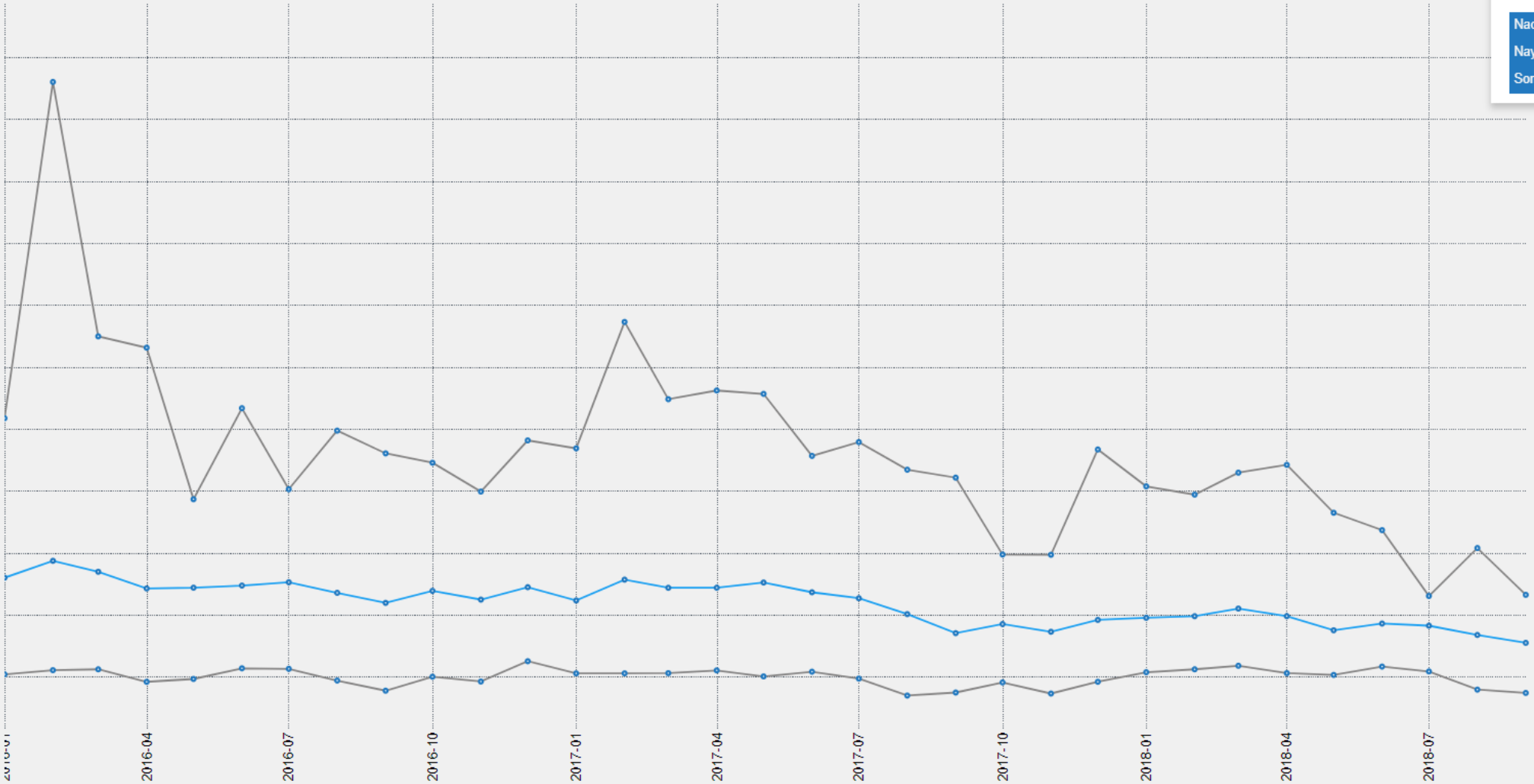
Entidad federativa

Buscar

Nacional (NAC)
Nayarit (NAY)
Sonora (SON)

Estado de ánimo de los tuiteros locales en México

1 de enero de 2016 - 26 de septiembre de 2018



Índice = Positivos / Negativos

Entidad federativa

Buscar

- Nacional (NAC)
- Nayarit (NAY)
- Sonora (SON)

Locales

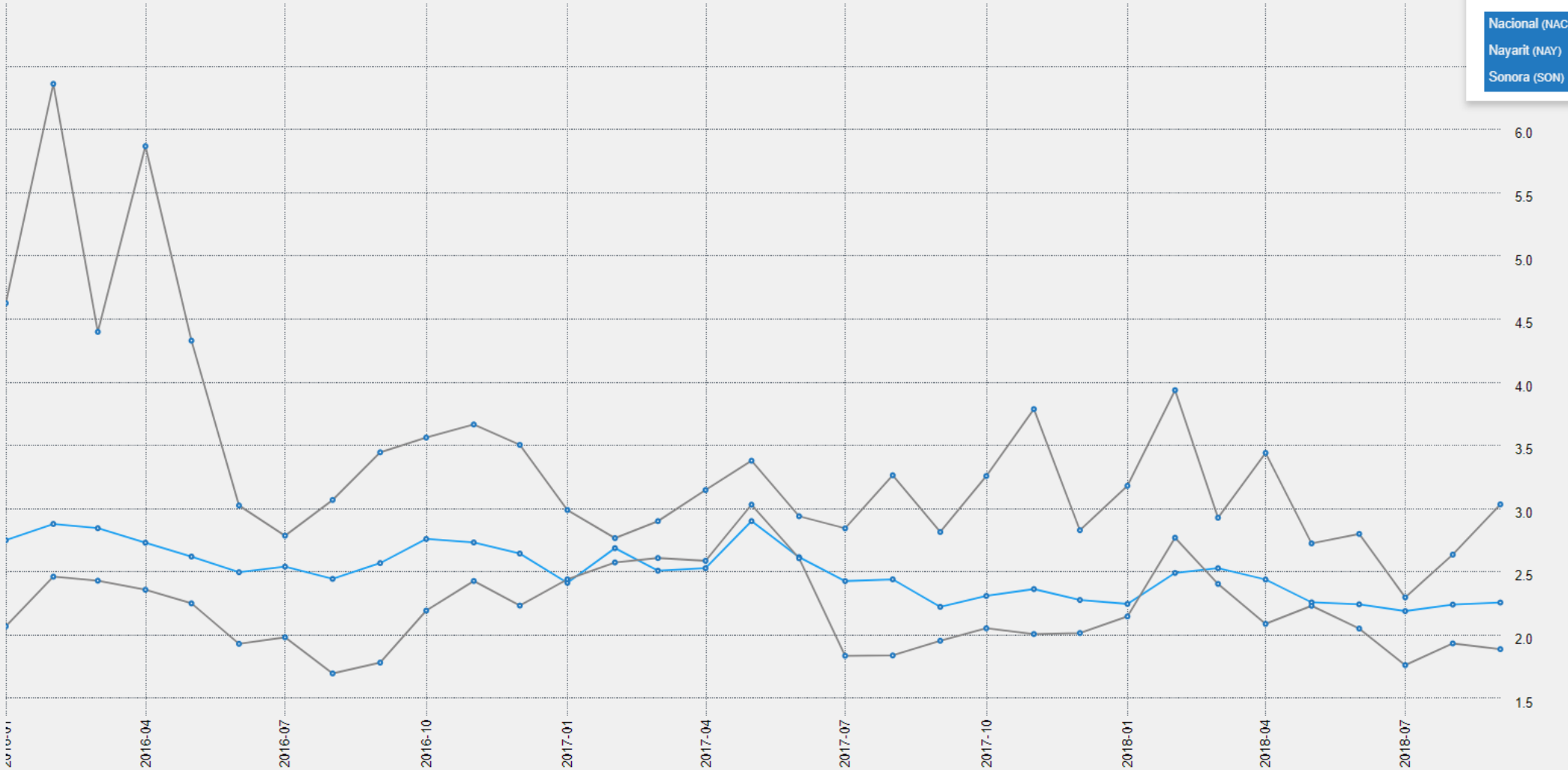
Estado de ánimo de los tuiteros visitantes en México

1 de enero de 2016 - 26 de septiembre de 2018

Entidad federativa

Buscar

- Nacional (NAC)
- Nayarit (NAY)
- Sonora (SON)

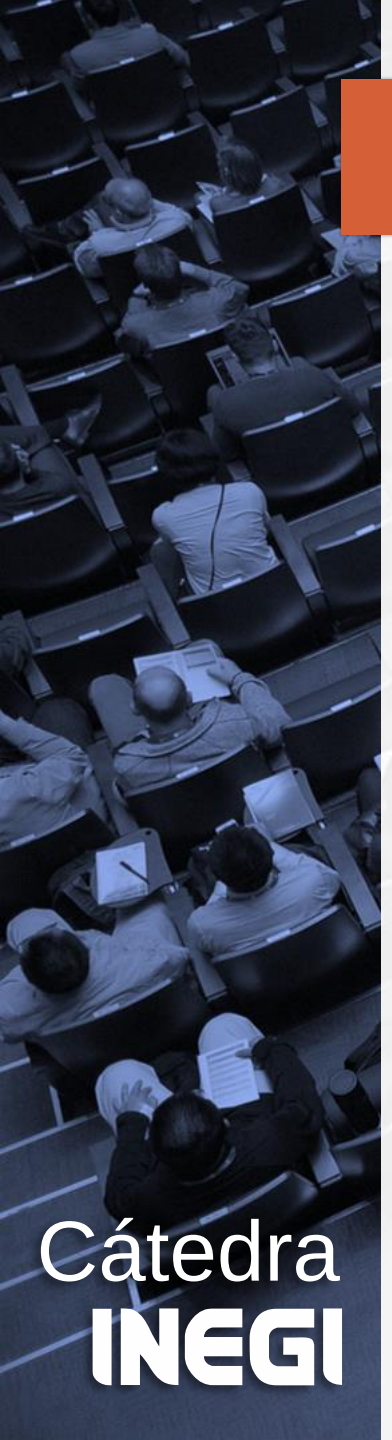


Índice = Positivos / Negativos

ELECCIONES, 2018

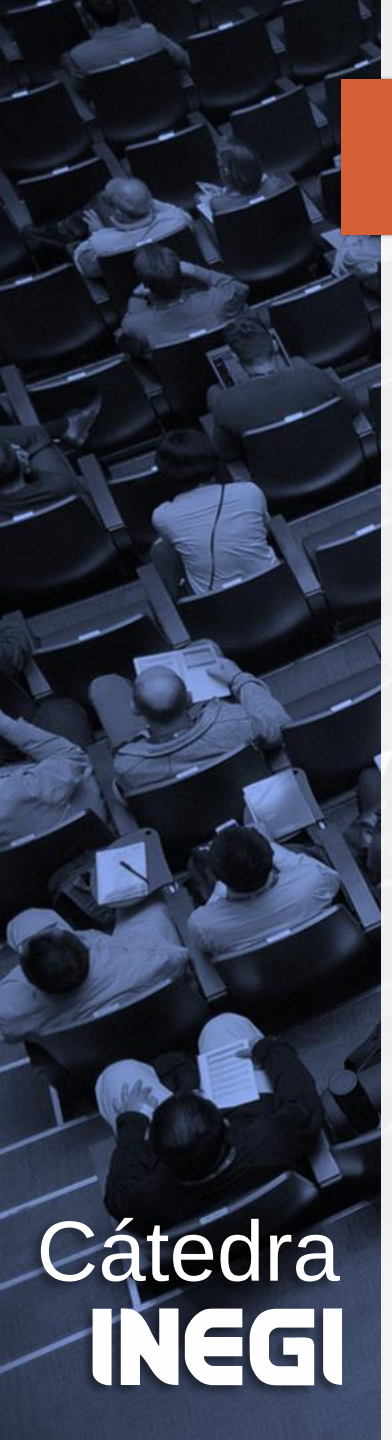
Cátedra INEGI

INSTITUTO NACIONAL DE ESTADÍSTICA Y GEOGRAFÍA



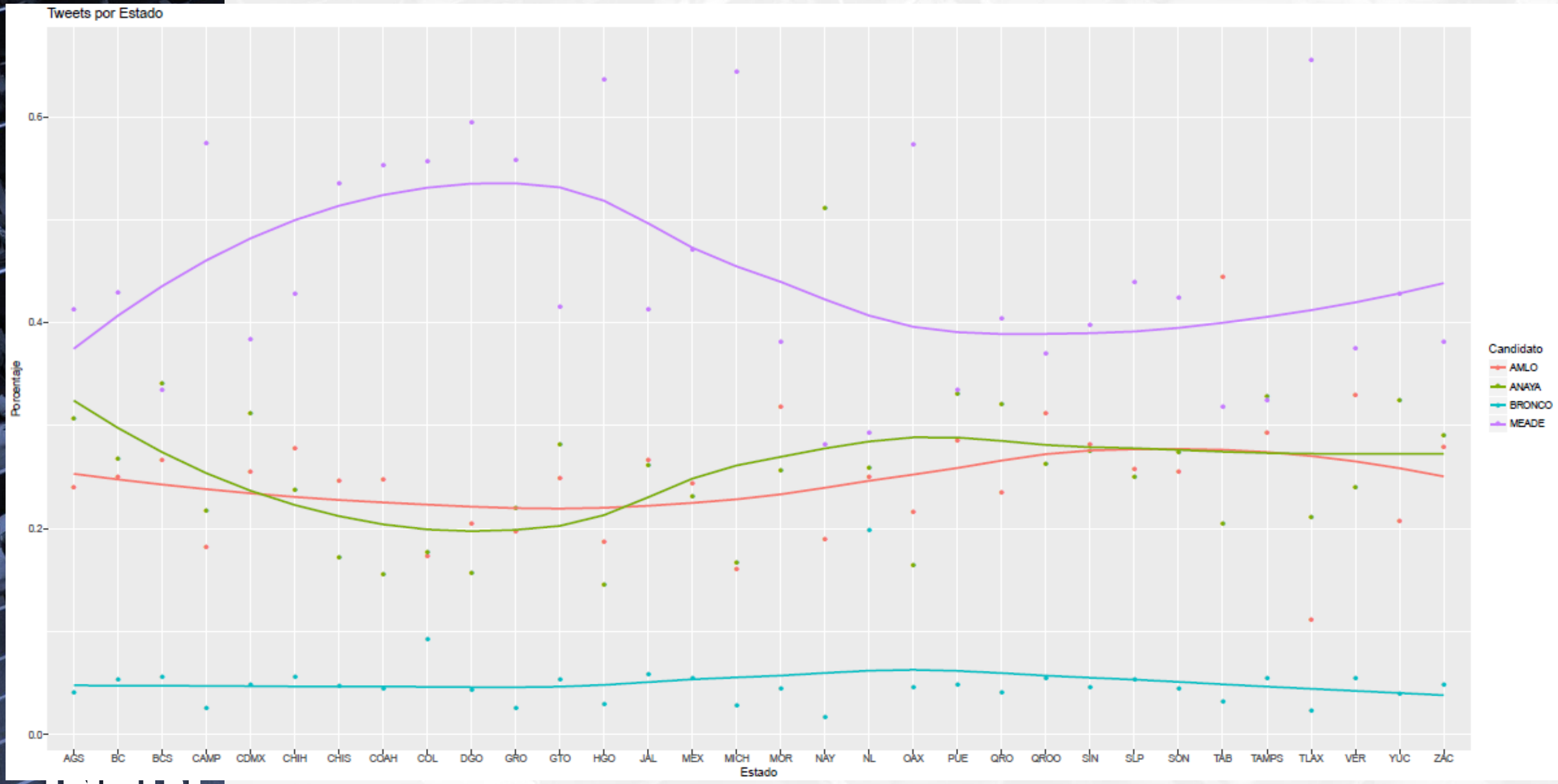
Datos

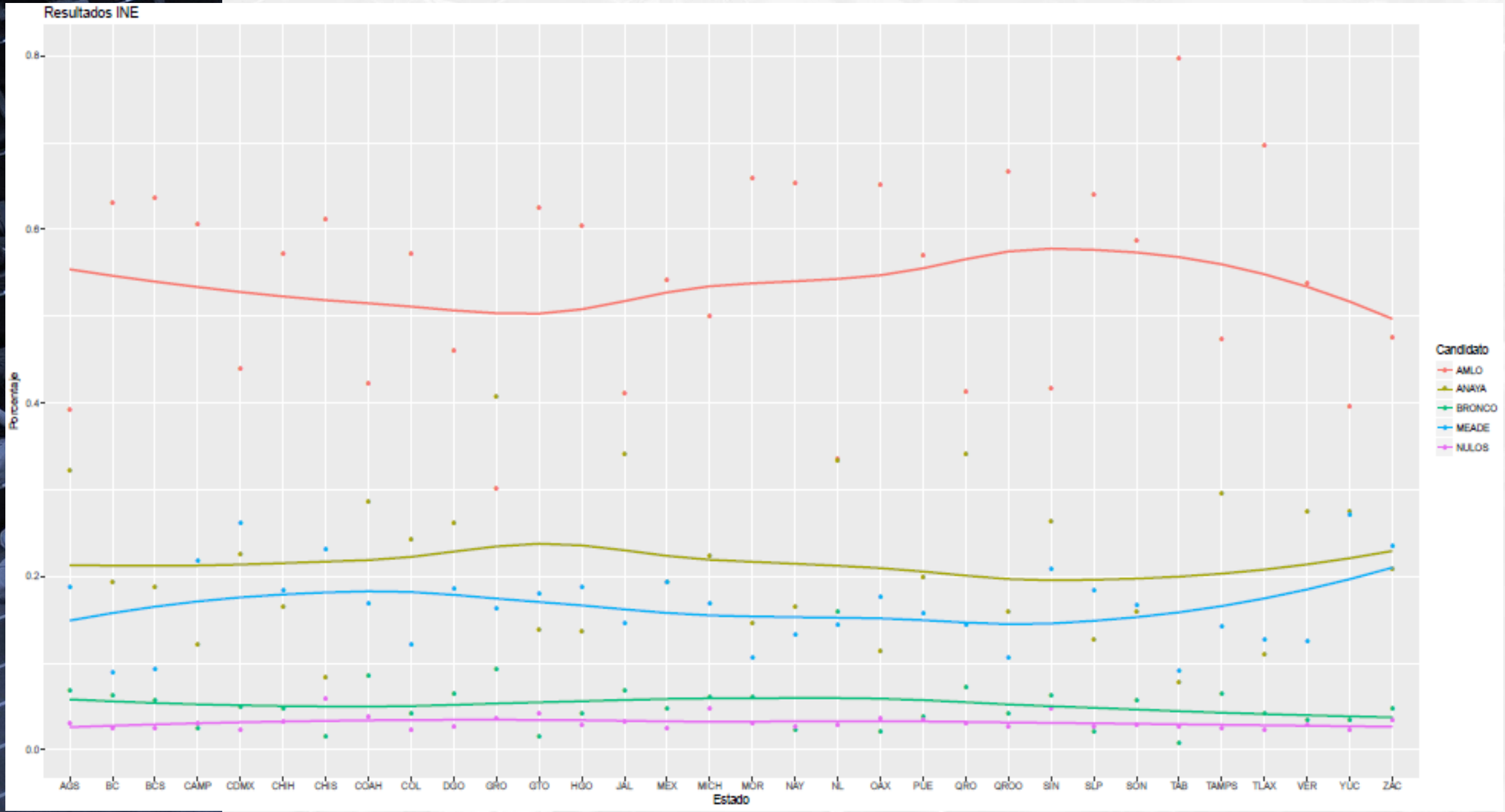
- Selección de tweets clasificados como “positivos”, filtrando por fecha y por texto.
- Texto es relacionado con cada uno de los cuatro candidatos, para generar una consulta que contenga información solo de ese candidato en un periodo de tiempo determinado.
- Las consultas fueron semanales a partir de Abril hasta Junio del presente año.



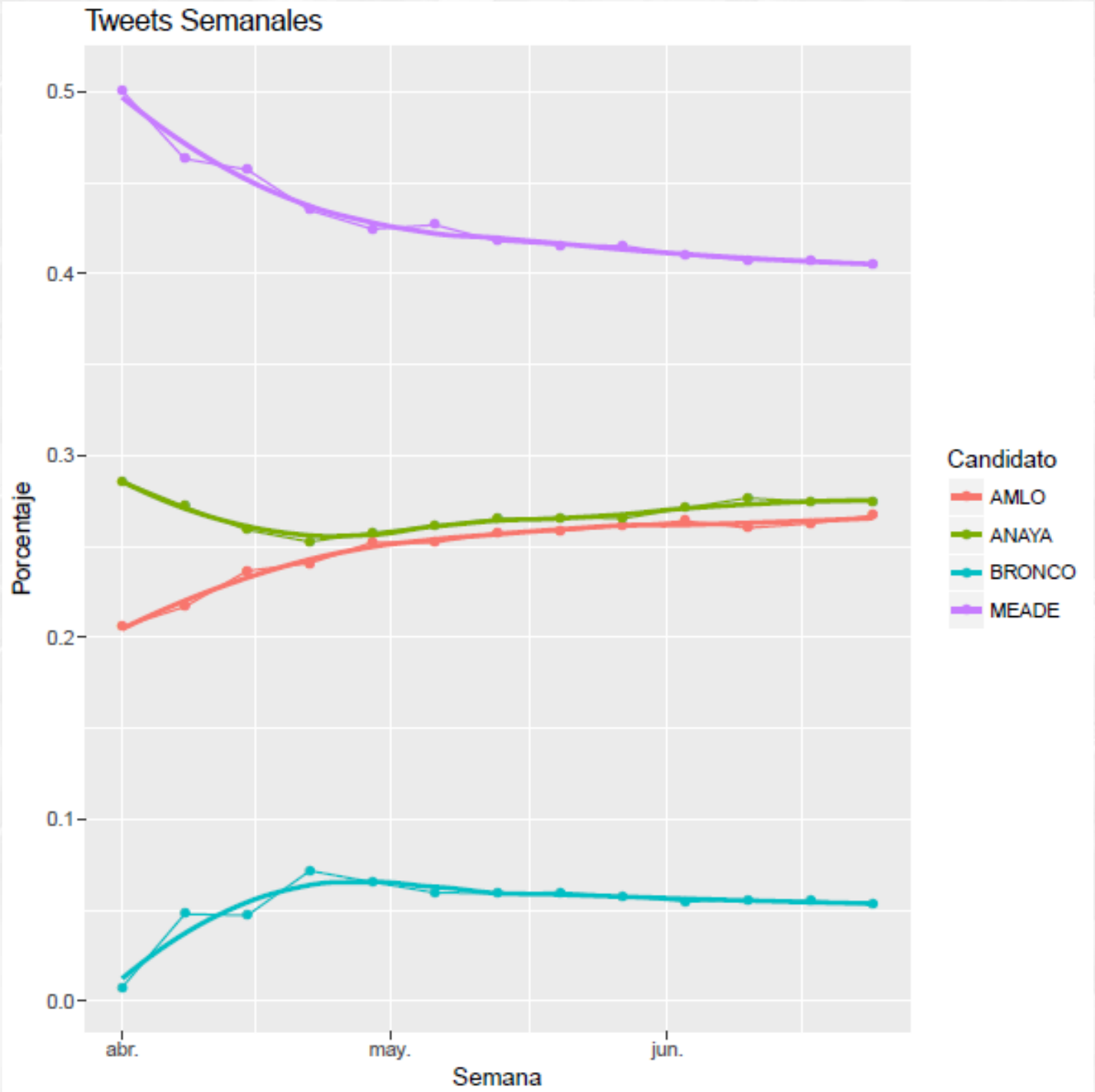
Datos

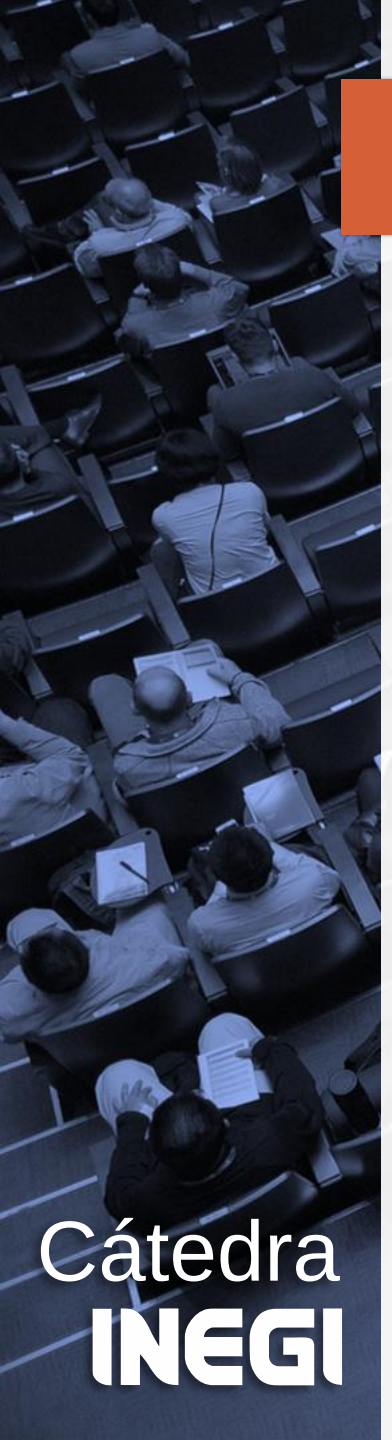
- Además, base de datos que contiene el conteo de votos por candidato, proporcionada por el INE.
- Con la base del INE se calculó el promedio de votos a favor de cada candidato por estado.
- En el caso de los tweets, se calculó el promedio de datos semanales y por estado, a favor de cada candidato.
- Dando como resultado las siguientes gráficas.


Tweets por Estado
Porcentaje
0.6
0.4
0.2
0.0
AGS BC BCS CAMP CDMX CHIH CHIS COAH COL DGO GRO GTO HGO JAL MEX MICH MOR NAY NL OAX PUE QRO QROO SIN SLP SON TAB TAMPS TLAX VER YUC ZAC
Estado
Candidato
AMLO
ANAYA
BRONCO
MEADE


Resultados INE
Porcentaje
0.8
0.6
0.4
0.2
0.0
AGS
BC
BCS
CAMP
CDMX
CHIH
CHIS
COAH
COL
DGO
GRO
GTO
HGO
JAL
MEX
MICH
MOR
NAY
NL
OAX
PUE
QRO
QROO
SIN
SLP
SON
TAB
TAMPS
TLAX
VER
YUC
ZAC
Estado
Candidato
AMLO
ANAYA
BRONCO
MEADE
NULOS

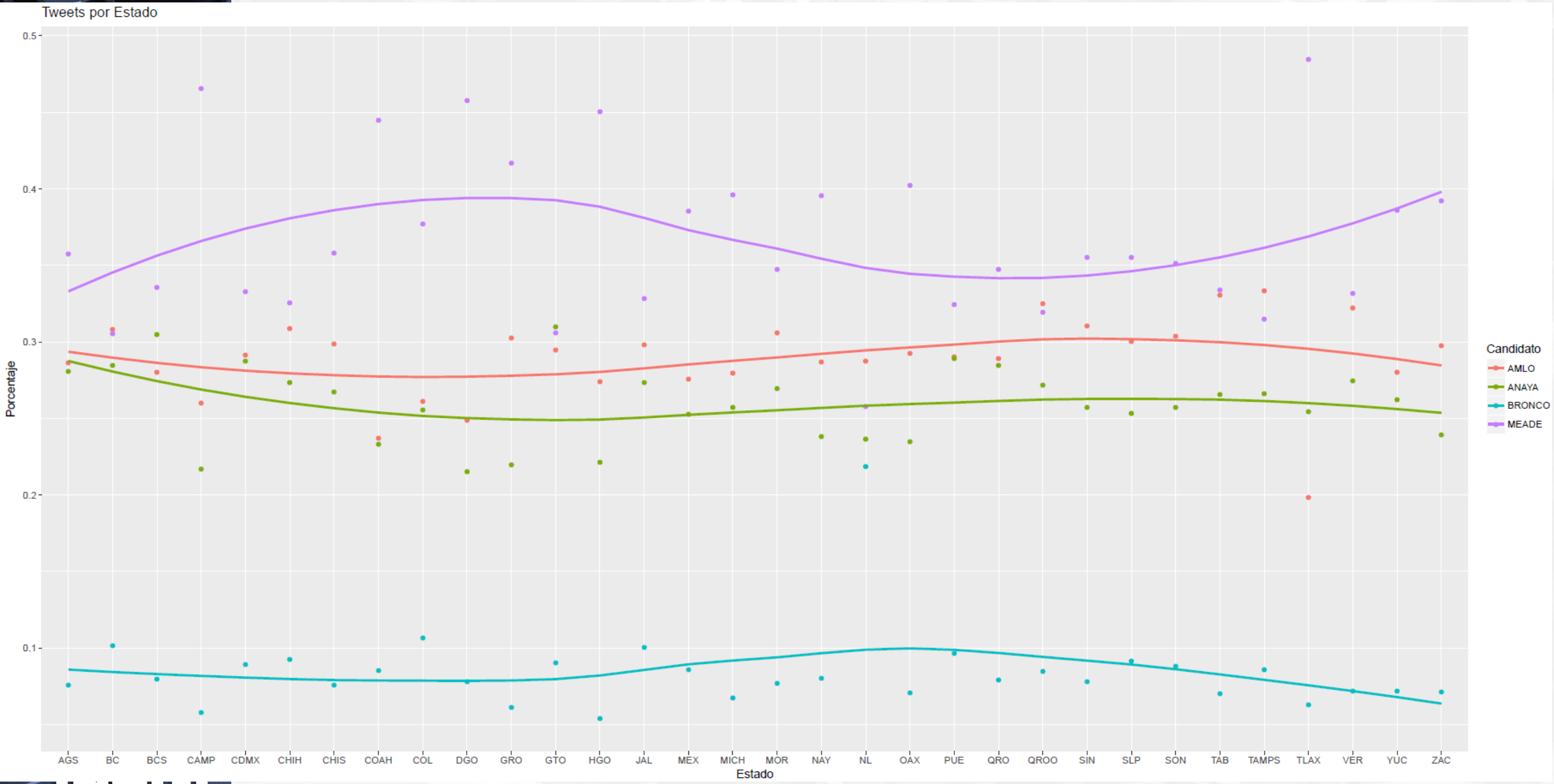

Cátedra
INEGI

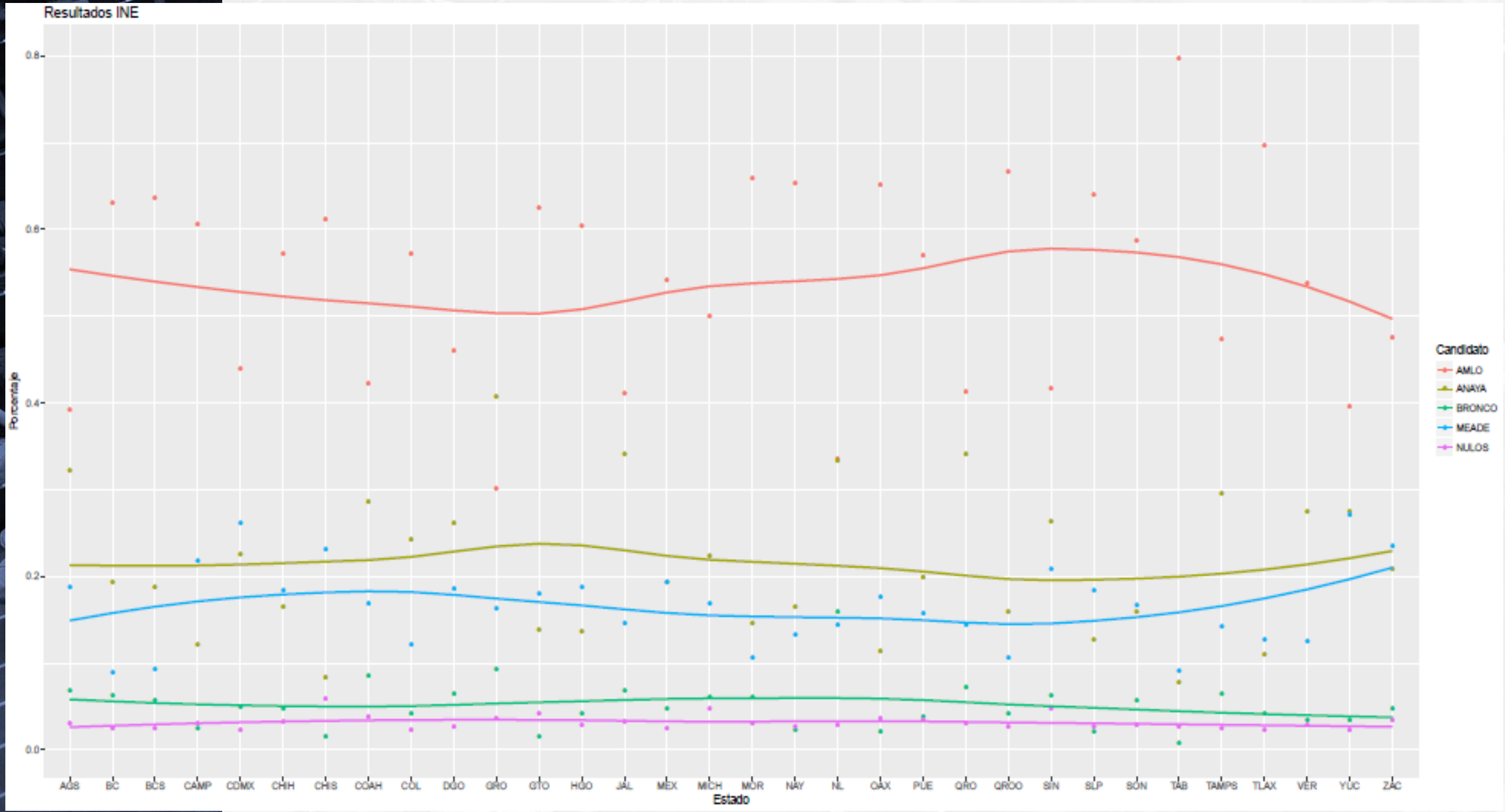

Tweets Semanales
0.5
0.4
0.3
0.2
0.1
0.0
Porcentaje
abr.
may.
jun.
Semana
Candidato
AMLO
ANAYA
BRONCO
MEADE



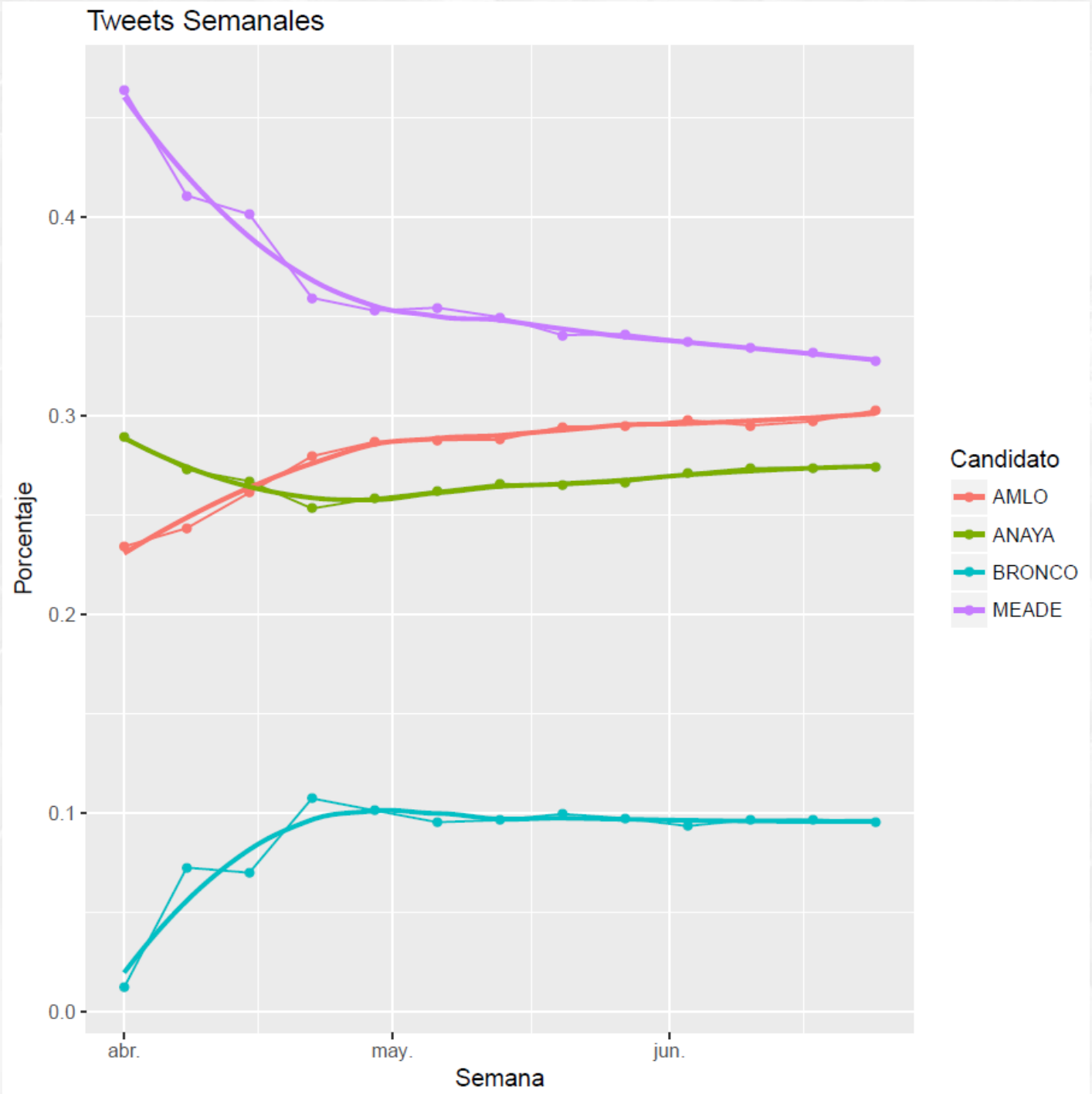
Segundo intento

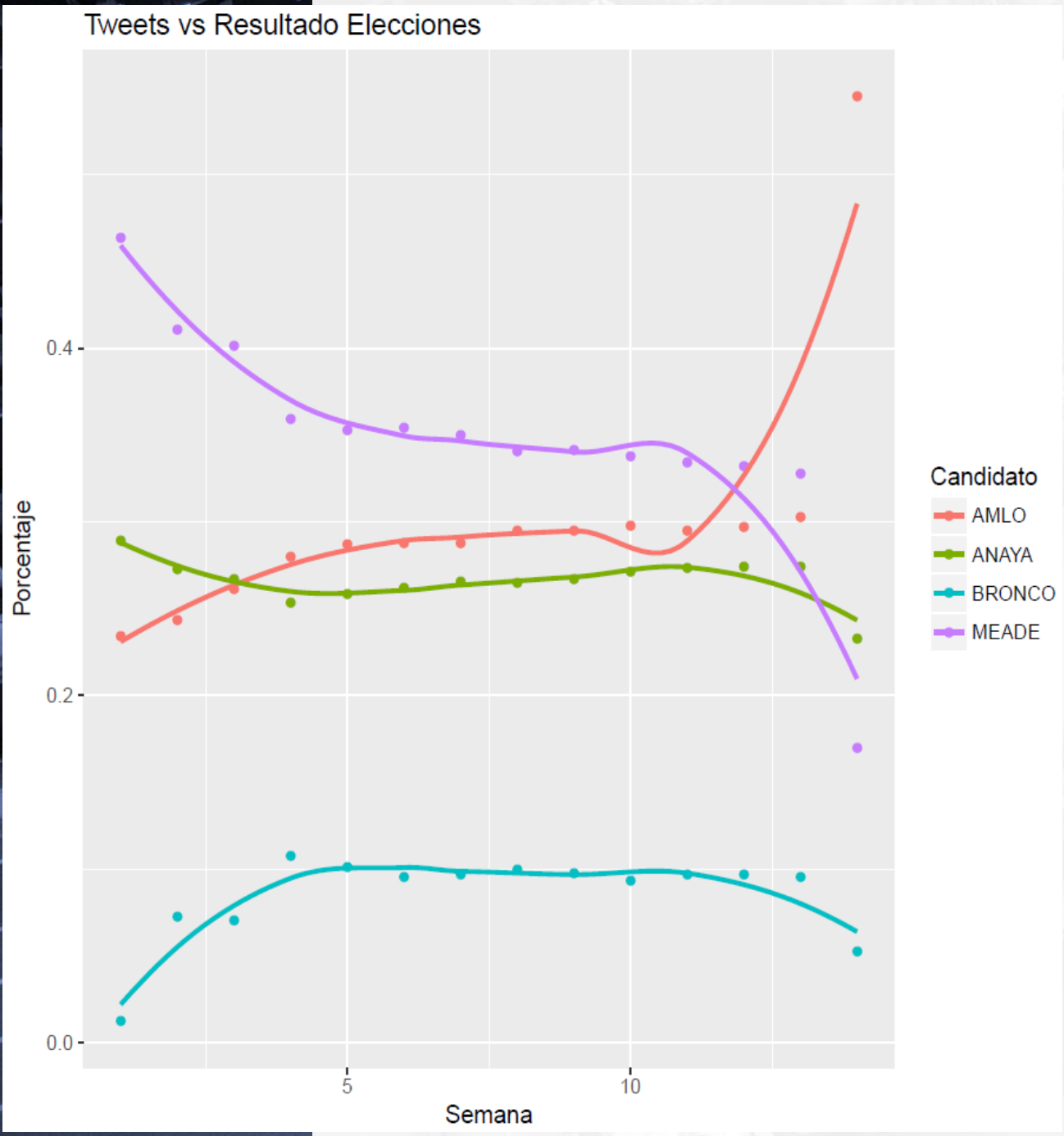
- La contabilización en el ejercicio anterior corresponde al número de tuits.
- Es decir, un tuitero puede votar más de una vez.
- Durante la elección ocurre algo diferente.
- En consecuencia, se decidió eliminar repeticiones del tuitero.
- En otras palabras, ahora contabilizamos tuiteros y no tuits
- Los resultados cambian.


Tweets por Estado
0.5
0.4
0.3
0.2
0.1
Porcentaje
AGS BC BCS CAMP CDMX CHIH CHIS COAH COL DGO GRO GTO HGO JAL MEX MICH MOR NAY NL OAX PUE QRO QROO SIN SLP SON TAB TAMPS TLAX VER YUC ZAC
Estado
Candidato
AMLO
ANAYA
BRONCO
MEADE


Resultados INE
Porcentaje
0.8
0.6
0.4
0.2
0.0
AGS
BC
BCS
CAMP
CDMX
CHIH
CHIS
COAH
COL
DGO
GRO
GTO
HGO
JAL
MEX
MICH
MOR
NAY
NL
OAX
PUE
QRO
QROO
SIN
SLP
SON
TAB
TAMPS
TLAX
VER
YUC
ZAC
Estado
Candidato
AMLO
ANAYA
BRONCO
MEADE
NULOS


Cátedra
INEGI


Tweets Semanales
Porcentaje
0.4
0.3
0.2
0.1
0.0
abr.
may.
jun.
Semana
Candidato
AMLO
ANAYA
BRONCO
MEADE



Agregando a la gráfica de los Twitteros semanales el resultado de las elecciones, se ve un salto no trivial en la línea.

Indica que el resultado de las elecciones no es predecible a partir de los datos encontrados en el análisis twittero.

El futuro de Big Data en la producción de información

- ¿Estamos dispuestos a integrarnos en un proceso de modernización acelerada?
- ¿Cuál será la aportación de valor de la información que produce el INEGI en el futuro?
- ¿Cómo nos adaptaremos a las nuevas necesidades de información de los ciudadanos?

Cátedra INEGI

Conociendo México

01 800 111 46 34
www.inegi.org.mx
atencion.usuarios@inegi.org.mx

INEGI Informa