SG VIRTUAL CONFERENCE 7ma edición

Big Data: Revelando los Secretos de Twitter en México

22 de Octubre 2014

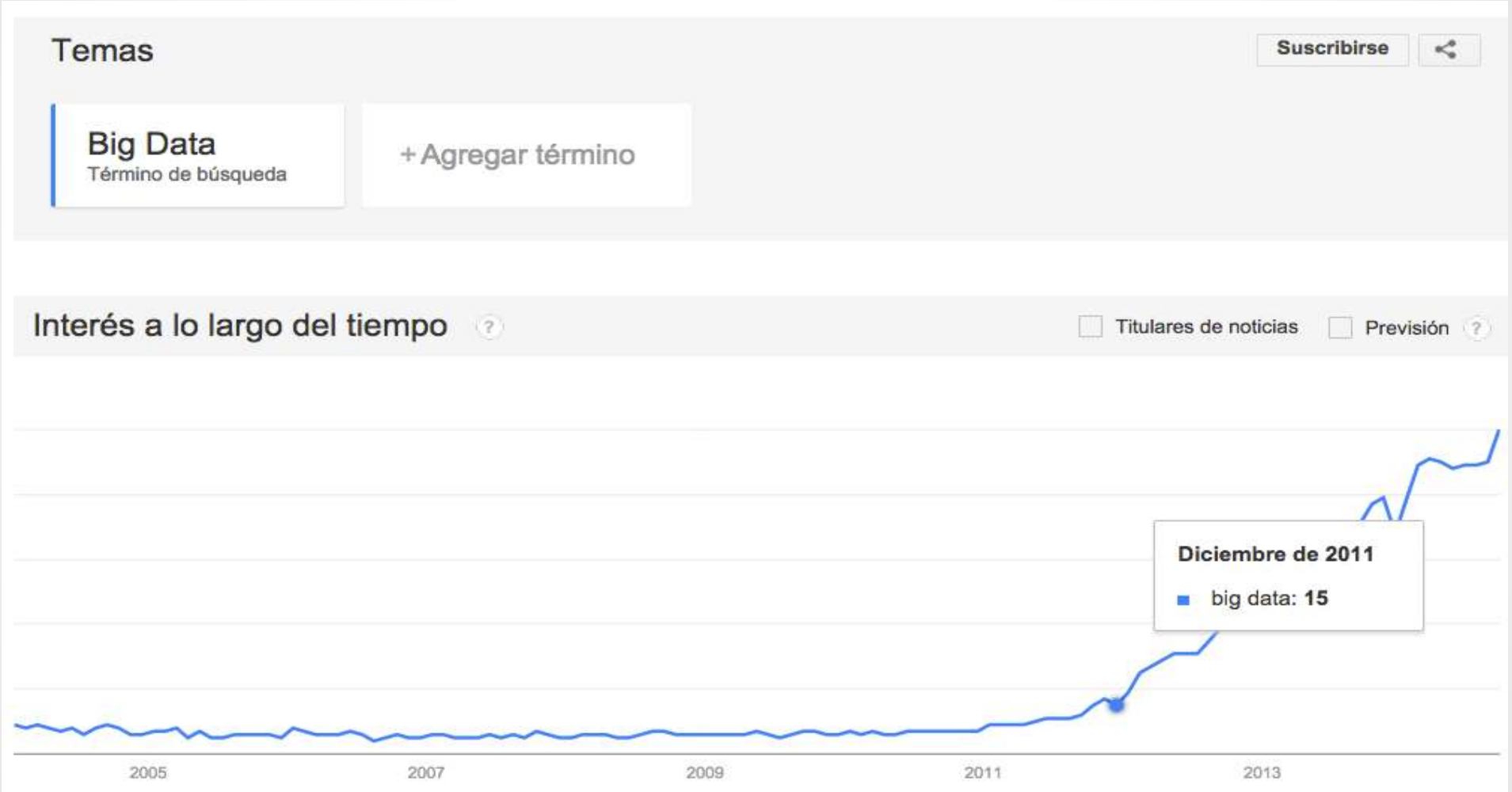
Presentado por:
Abel Alejandro Coronado Iruegas
@abxda

abel.coronado@inegi.org.mx

Objetivo

Inspirarlos para que le entren al mundo de Big Data.

Big Data



https://www.google.com.mx/trends/

https://twitter.com/abxda

¿Qué es Big Data?

http://datascience.berkeley.edu/what-is-big-data/

https://twitter.com/abxda

¿Qué es Big Data?

Ryan Swanstrom

Data Science Blogger, Data Science 101
Twitter: @swgoof

> *Big data used to mean data that a single machine was unable to handle.*
> ***Now big data has become a buzzword to mean anything related to data analytics or visualization.***

---

http://datascience.berkeley.edu/what-is-big-data/

¿Qué es Big Data?

Anna Smith

Analytics Engineer, Rent the Runway
Twitter: @OMGannaks

> ***Big data is when data grows to the point that the technology supporting the data has to change.*** *It also encompasses a variety of topics relating to how disparate data can be combined, processed into insights, and/or reworked into smart products.*

---

http://datascience.berkeley.edu/what-is-big-data/

https://twitter.com/ab xda

¿Qué es Big Data?

David Leonhardt

Editor, The Upshot, *The New York Times*
Twitter: @DLeonhardt

> ***Big Data is nothing more than a tool for capturing reality*** *— just as newspaper reporting, photography and long-form journalism are. But it's an exciting tool, because it holds the potential of capturing reality in some clearer and more accurate ways than we have been able to do in the past.*

---

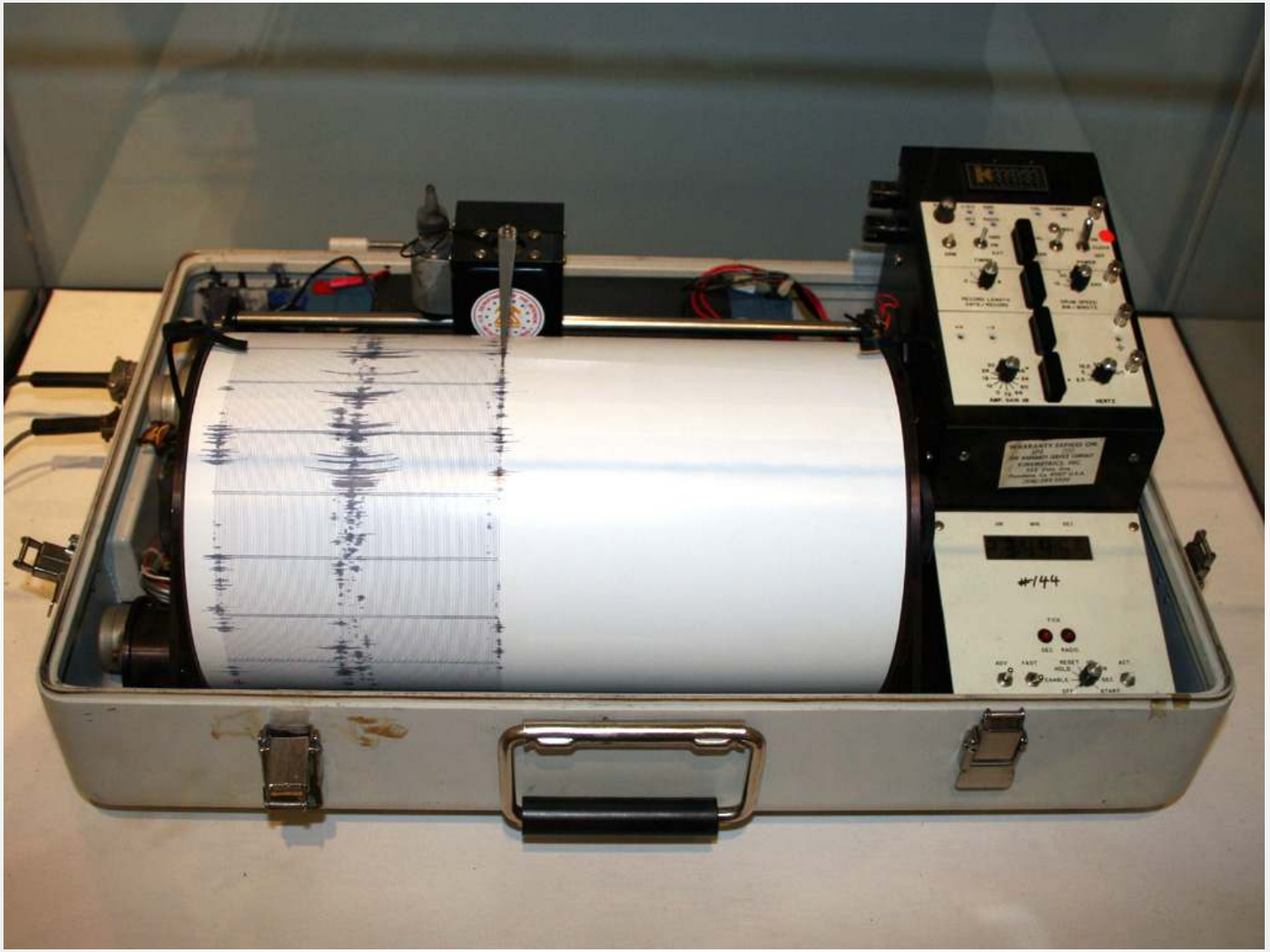
http://datascience.berkeley.edu/what-is-big-data/

https://twitter.com/abxda

Volumen

http://commons.wikimedia.org/wiki/Elephas#mediaviewer/File:Berlin_Landesvertretung_Niedersachsen_Elefant.jpg

Velocidad



http://upload.wikimedia.org/wikipedia/commons/0/0f/Kinemetrics_seismograph.jpg

Variedad

http://upload.wikimedia.org/wikipedia/commons/f/f6/Popular_Social_Networks%2C_Gavin_Llewellyn%2C_CC.jpg

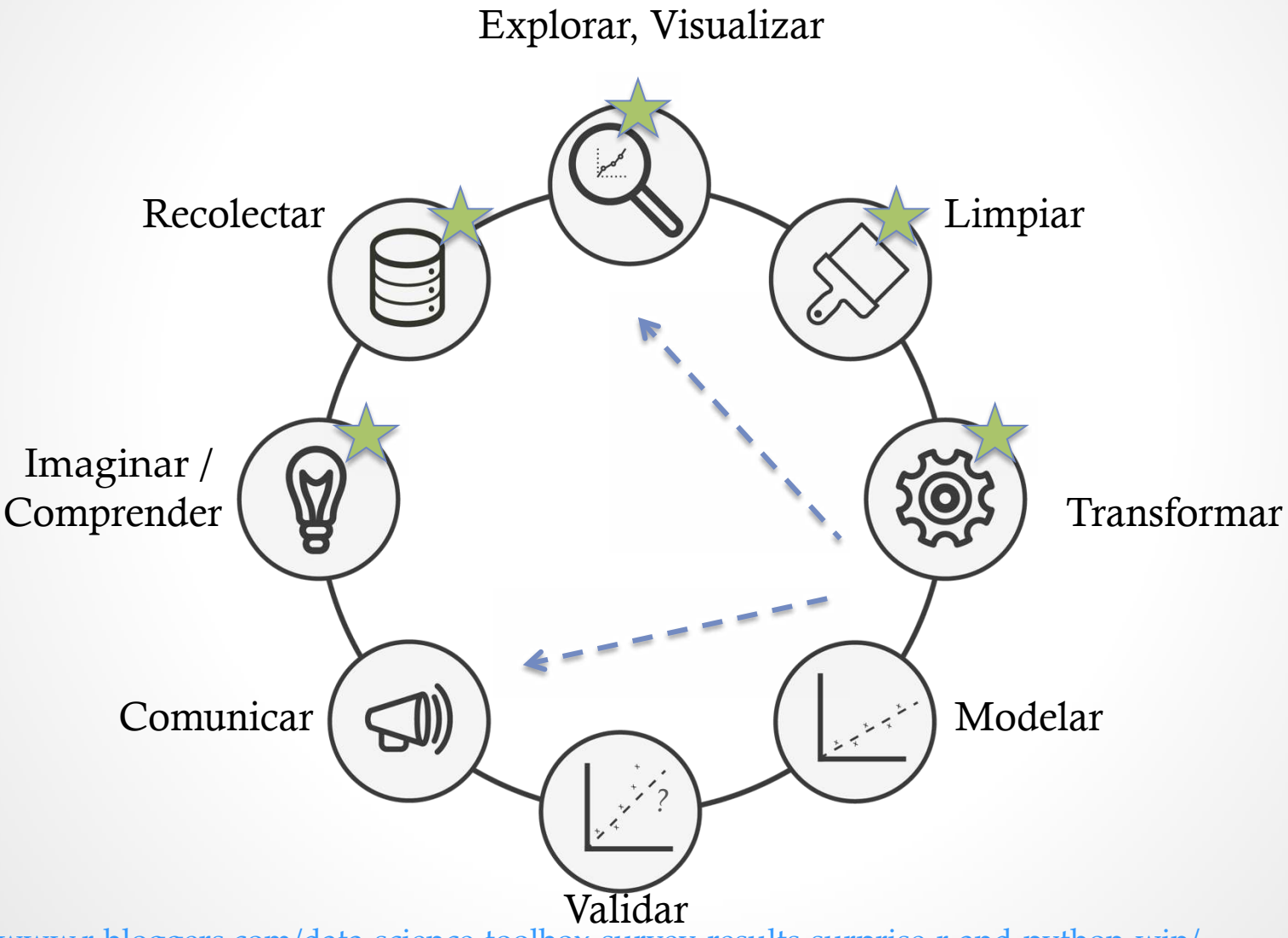
Tomar decisiones, actuar y crear valor

Ciencia de Datos

http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

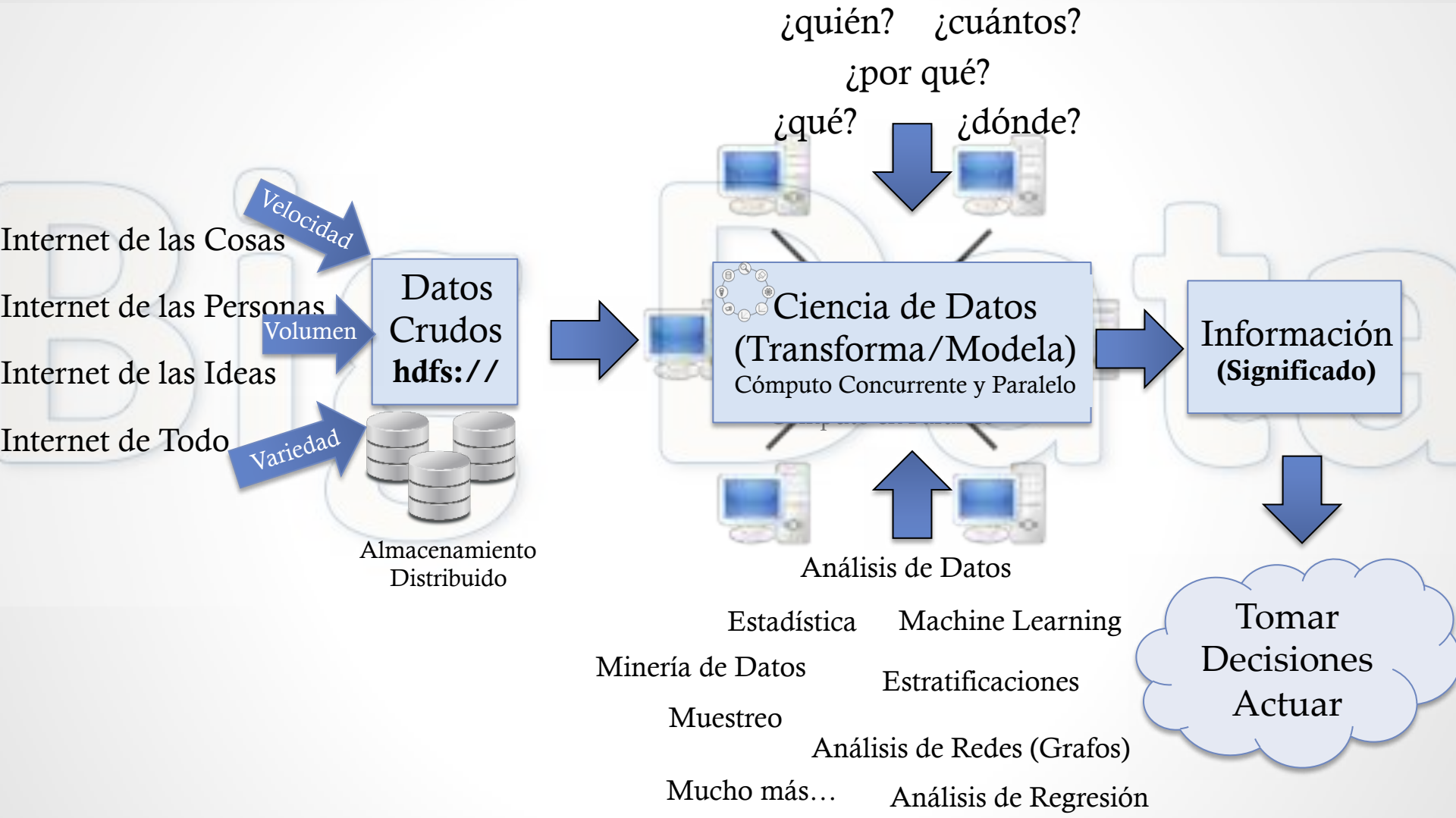
https://twitter.com/abxda

Ciencia de Datos



http://www.r-bloggers.com/data-science-toolbox-survey-results-surprise-r-and-python-win/

Ciencia de Datos y Big Data



Big Data en las Oficinas Nacionales de Estadística

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE 10 March 2013
CONFERENCE OF EUROPEAN STATISTICIANS

WHAT DOES "BIG DATA" MEAN FOR OFFICIAL STATISTICS?

http://www1.unece.org/stat/platform/download/attachments/58492100/Big+Data+HLG+Final.docx?version=1&modificationDate=1362939424184

Big Data en las Oficinas Nacionales de Estadística

- **It is clear that during the next two years there is a need to identify a few pilot projects that will serve as proof of concept.**
- **Statistical organisations are, therefore, encouraged to address formally Big data issues in their annual and multi-annual work programmes by undertaking research and pilot projects in selected areas and by allocating appropriate resources for that purpose.**

Big Data en las Oficinas Nacionales de Estadística

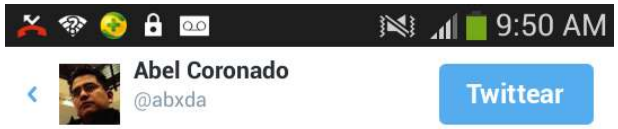
- **'new' exploration and analysis methods are required**: *Visualization methods*, *Text mining*, and *High Performance Computing.*
- To use Big data, **statisticians are needed with a different mind-set and new skills**. The processing of more and more data for official statistics requires statistically aware people with an analytical mind-set, an affinity for IT (e.g. programming skills)

Twitter como fuente de BigData

http://www.salesforcemarketingcloud.com/blog/2010/11/why-the-twitter-firehose-matters-to-you/

https://twitter.com/abxda

¿Cuántos caracteres?



Mañana 22 de Oct !!! Big Data: Revelando los Secretos de Twitter en México | SG: sg.com.mx/sgvirtual/7/se… vía @RevistaSG

https://twitter.com/abxda

140 ??

https://twitter.com/abxda

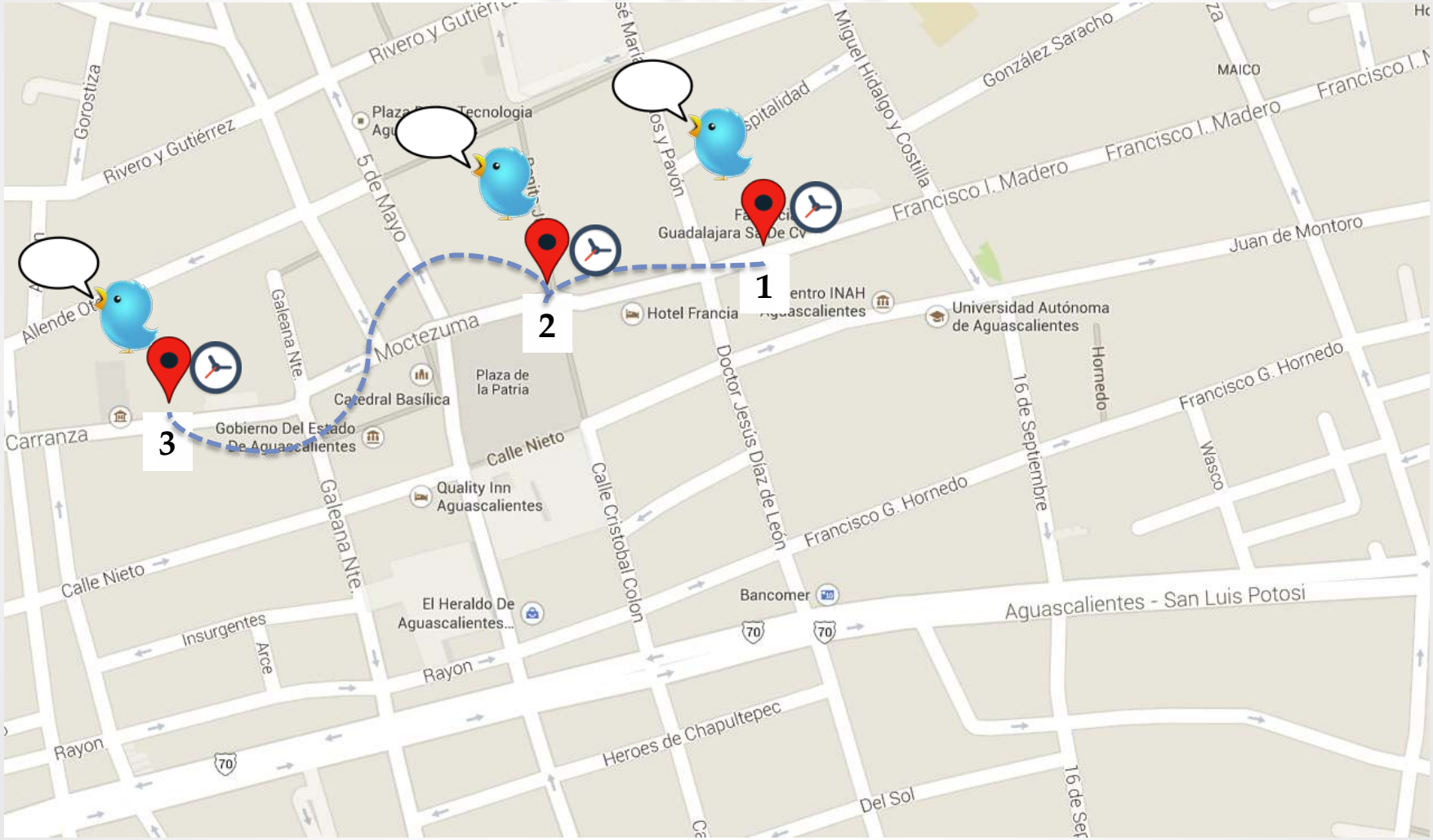
```
{
  "text": Mañana 22 de Oct !!! Big Data: Revelando los ...tos de Twitt...en Méx...co | S... sg.....mx...v
  "created_at": "2014-02-21T17:00:50.000Z",
  "source": "<a href=\"http://android.com\" rel=\"nofoll...\">Android</a>",
  "truncated": false,
  "mention": [],
  "retweet_count": 0,
  "hashtag": [],
  "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 Estadística Oficial by CIMAT,Emprendedor,
    "profile_image_url": "http://pbs.twimg.com/profile_images/378800000635411988/5c5b7a5754d65e3d1a2895
    "profile_image_url_https": "https://pbs.twimg.com/profile_images/378800000635411988/5c5b7a5754d65e3
  }
```


1482

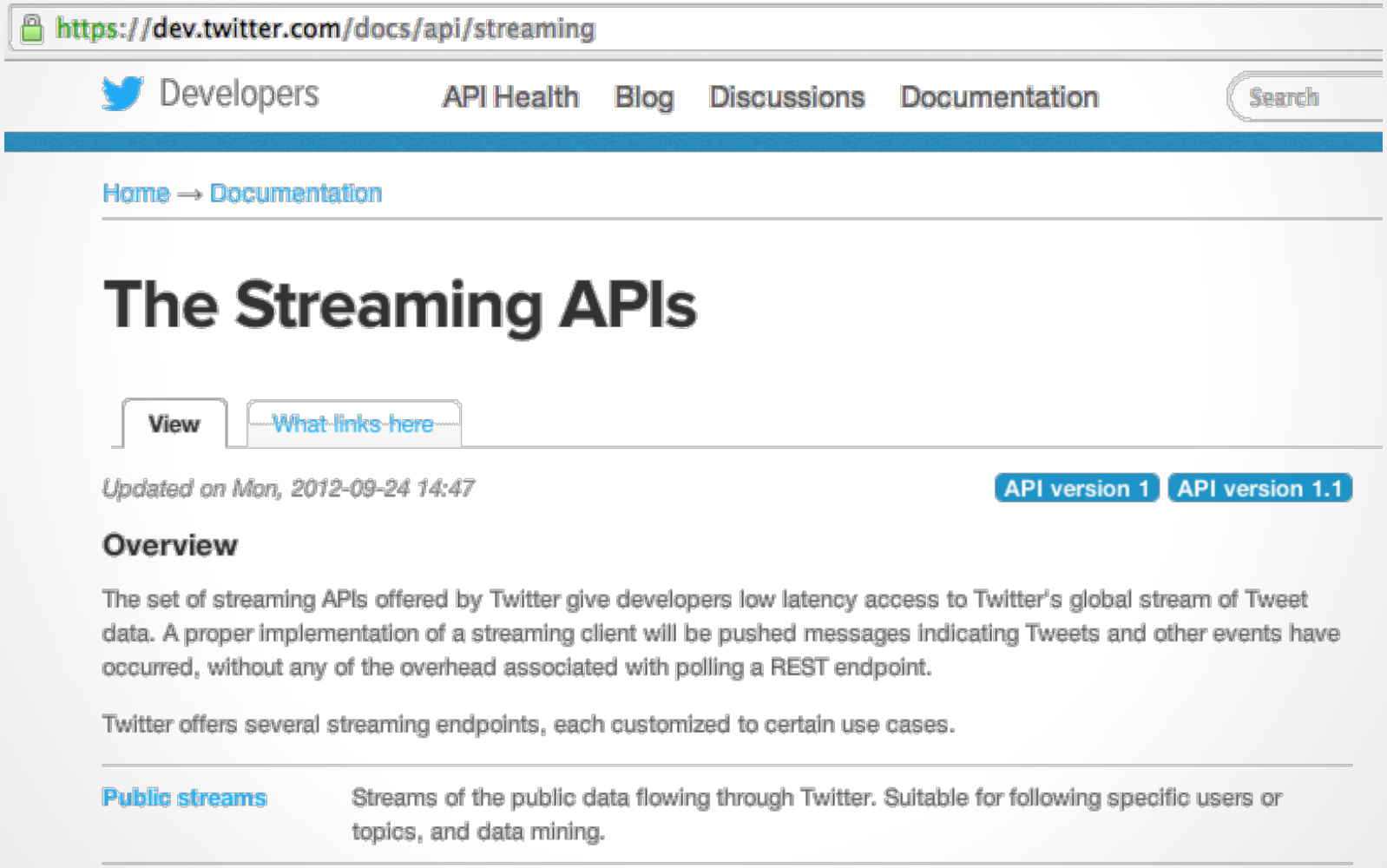
{JSON}

Json: Formato de Intercambio

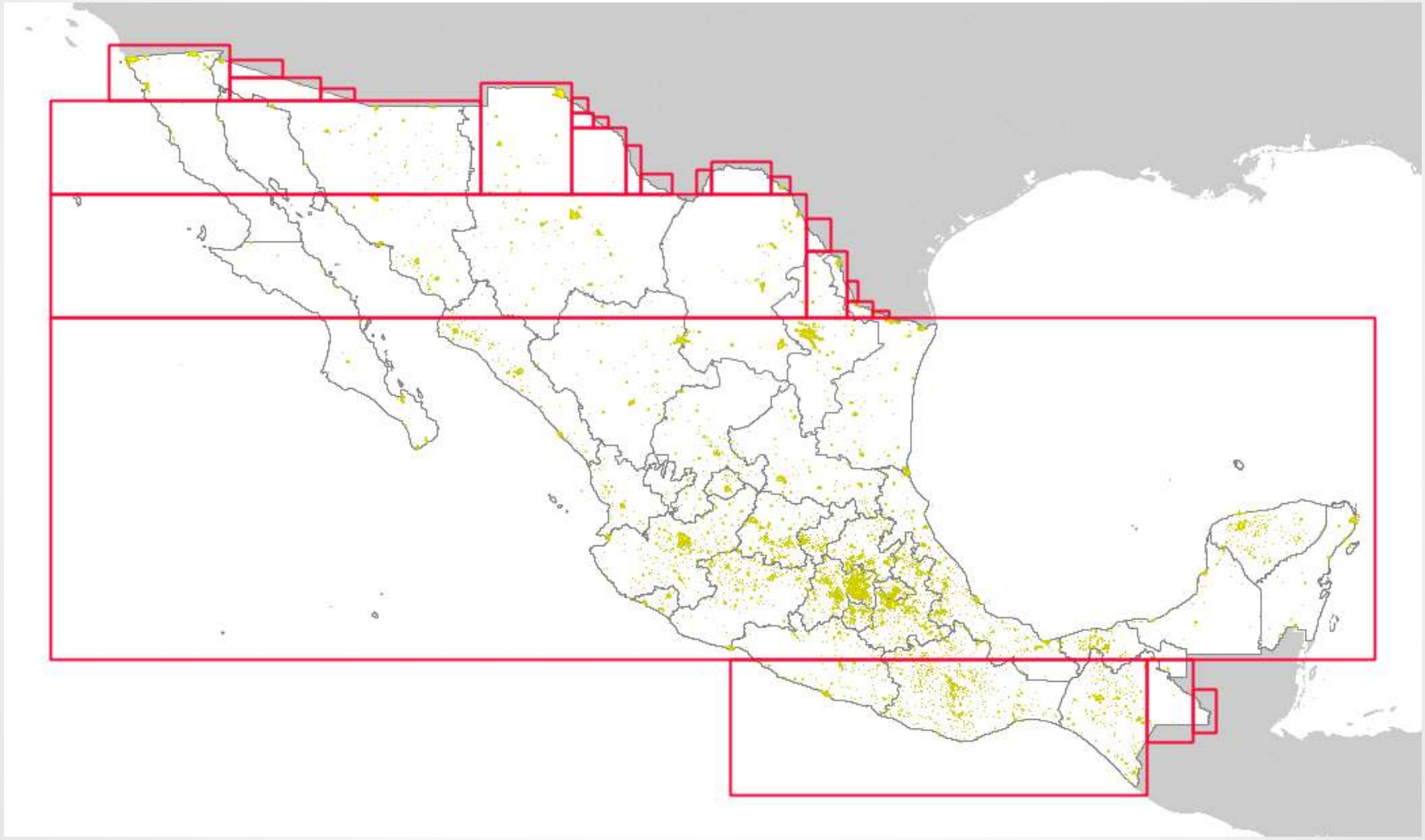
Nuestra huella en las Redes Sociales



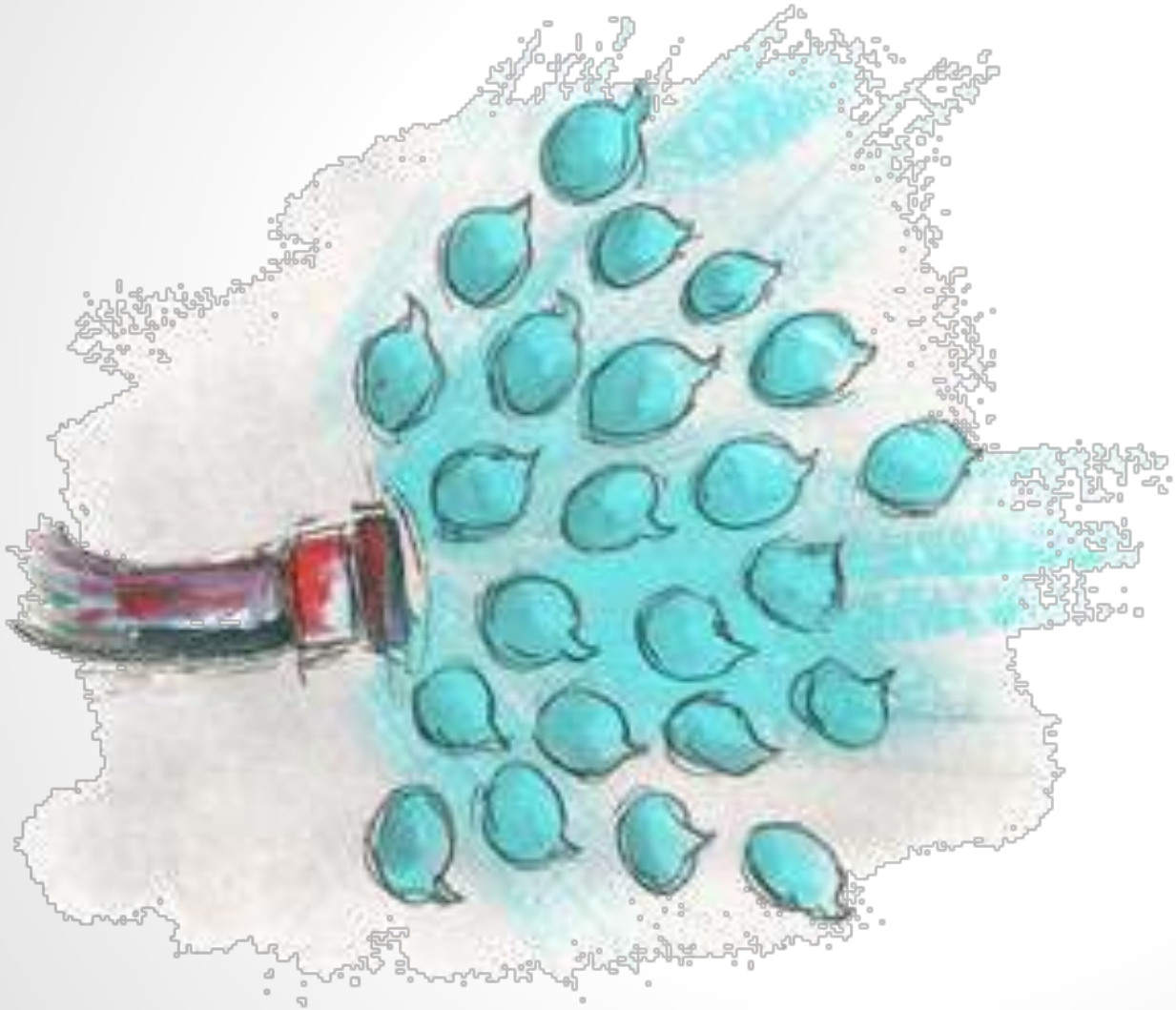
Todos los tuits están disponibles para su recolección en tiempo real.



Incluso permite consultas geográficas



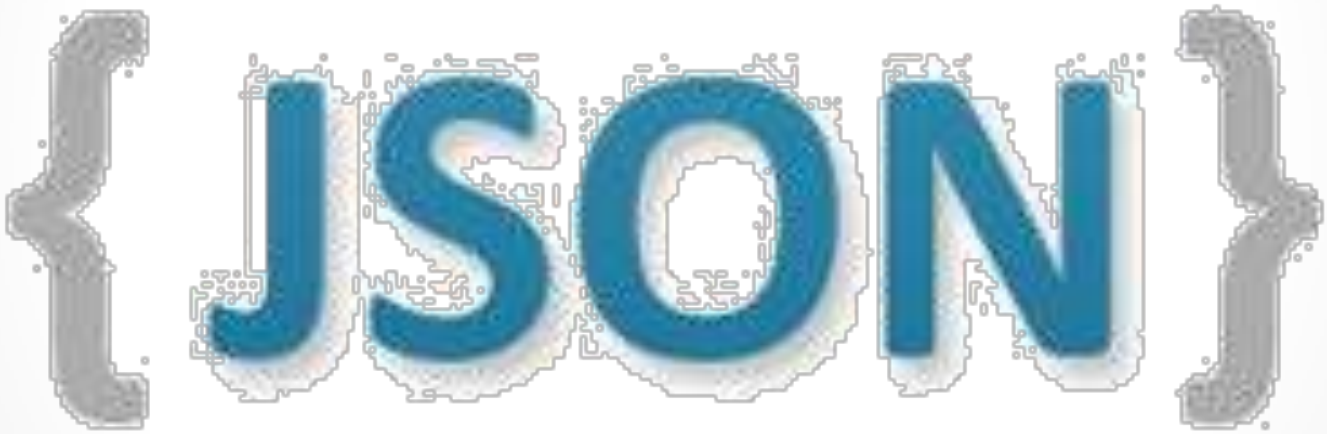
¿Dónde recolectar?



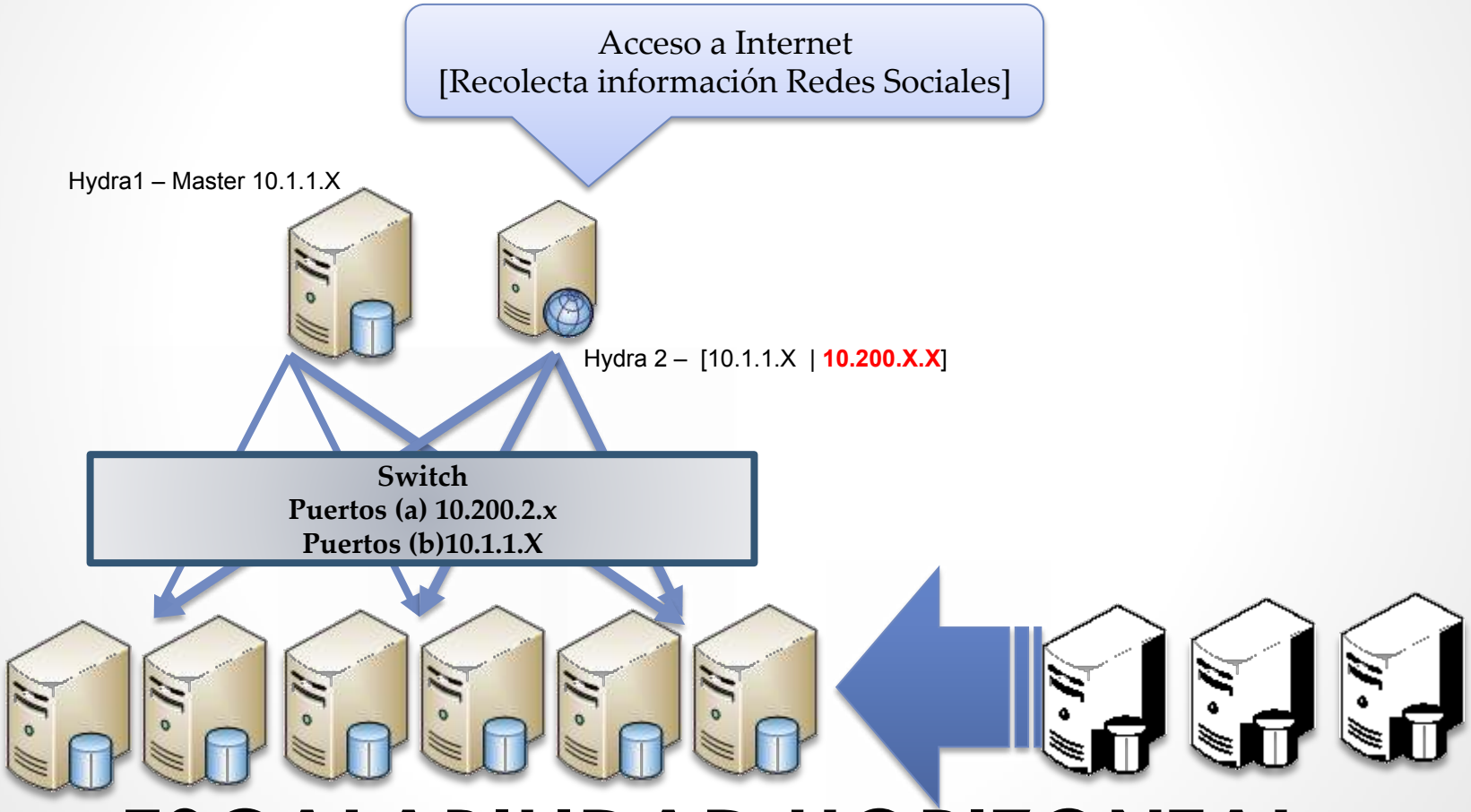
elasticsearch

http://www.elasticsearch.org/

¿Por qué ElasticSearch?



¿Por qué ElasticSearch?

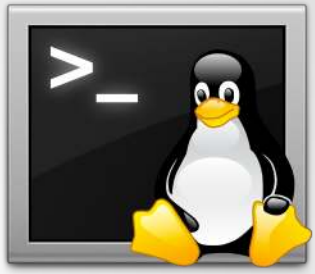


< ESCALABILIDAD HORIZONTAL >

Hydra

Hydra

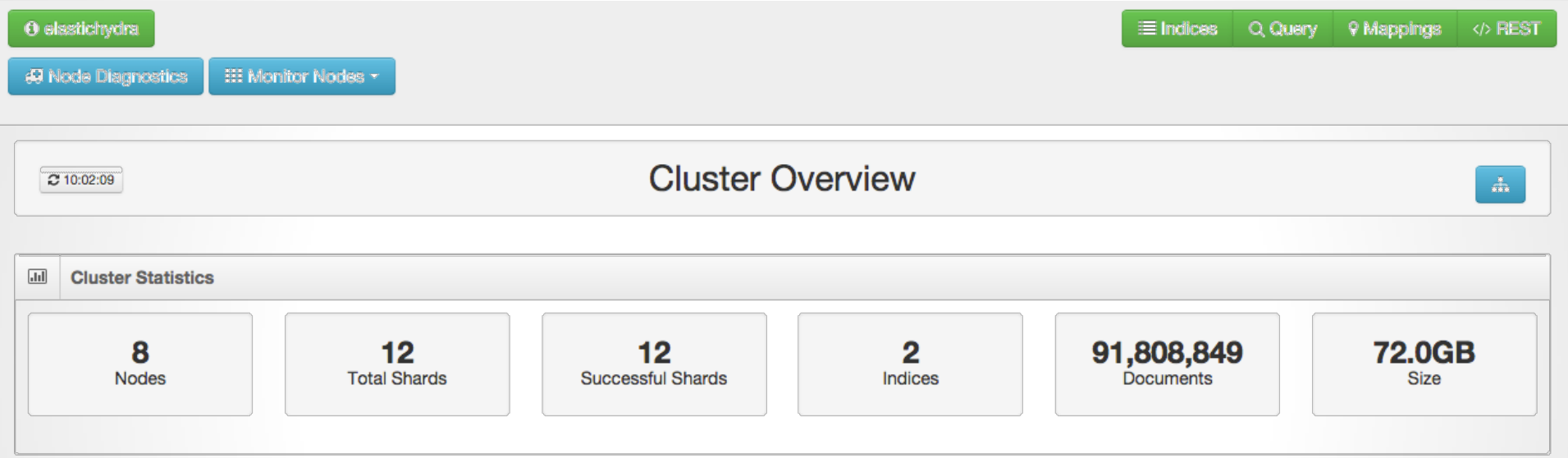
Twitter River



https://github.com/elasticsearch/elasticsearch-river-twitter

```
curl -XPUT localhost:9200/_river/my_twitter_river/_meta -d'
 {
 "type" : "twitter",
     "twitter" : {
         "oauth" : {
             "consumer_key" :”XXXxxXXxXxX",
             "consumer_secret" : "XXXxxXXxXxXXXXxxXXxXxXXXXxxXXxXxX",
             "access_token" : "XXXxxXXxXxXXXXxxXXxXxXXXXxxXXxXxX",
             "access_token_secret" : "XXXxxXXxXxXXXXxxXXxXxX"
         },
         "filter" : {
        "locations" :"-118.40764955,14.53209836,-86.71040527,32.71865357"
         }
     }
 }
 '
```

La recolección 2014



elasticsearch

https://twitter.com/abxda

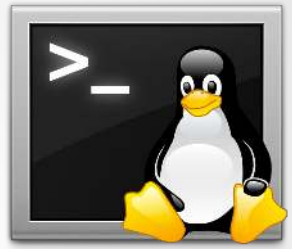
Extractor


```
es = Elasticsearch(['10.200.2.41:9200'])
rs = es.search(index=['my_twitter_river'],
scroll=duracion, search_type='scan', size=int(noTuits),
body={
    "query": {
    "range" : {
        "created_at" : {
            "gte": fechaInicio,
            "lte": fechaFin
        }
    }
}})
```

CSV

```
MacBook-Pro-de-Abel:DataBase abxda$ ls -alh tweets_f_s.csv
-rw-r--r--  1 abxda  staff    18G Sep 25 11:48 tweets_f_s.csv
MacBook-Pro-de-Abel:DataBase abxda$ head tweets_f_s.csv
```

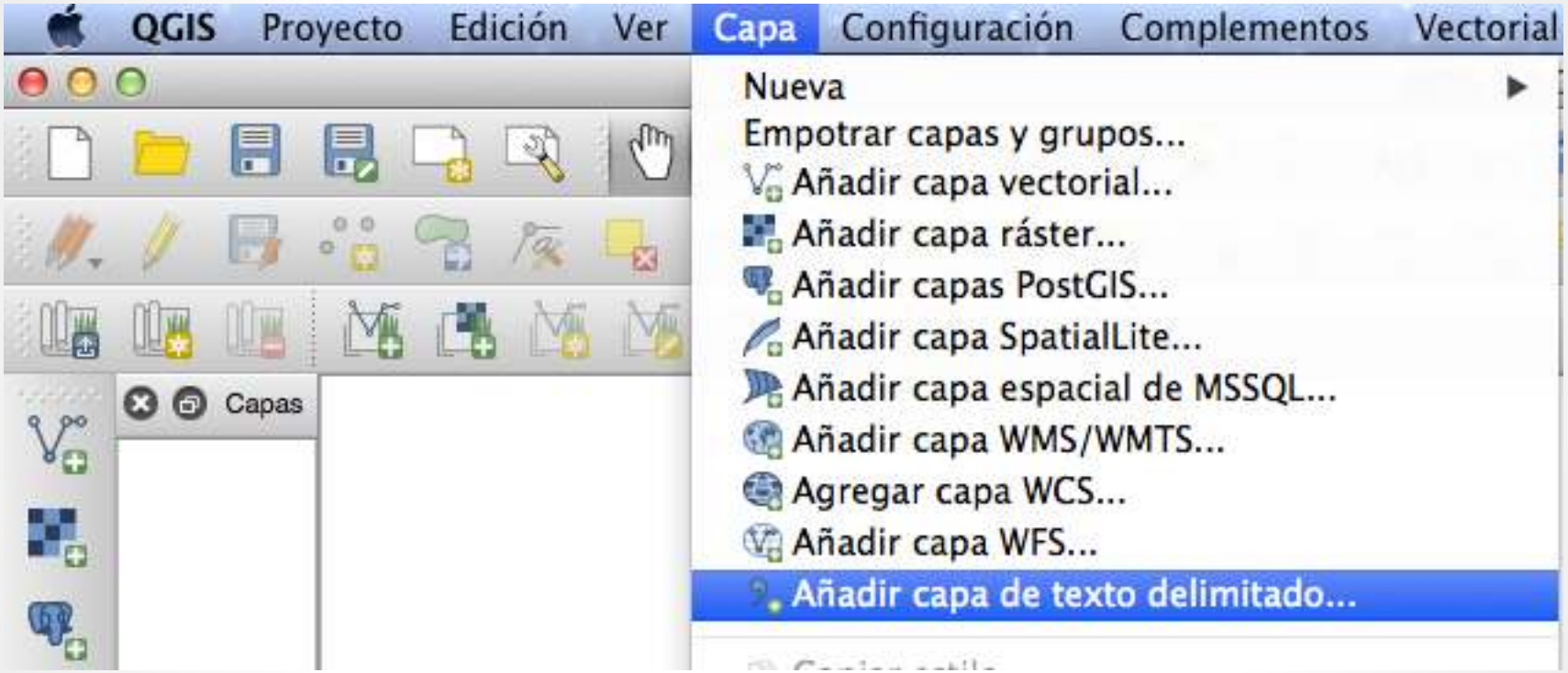
Se extraen los puntos del CSV



```
$cat tweets_feb_sep_ord_loc.csv | awk -F',' '{print $3 "," $4}'
```

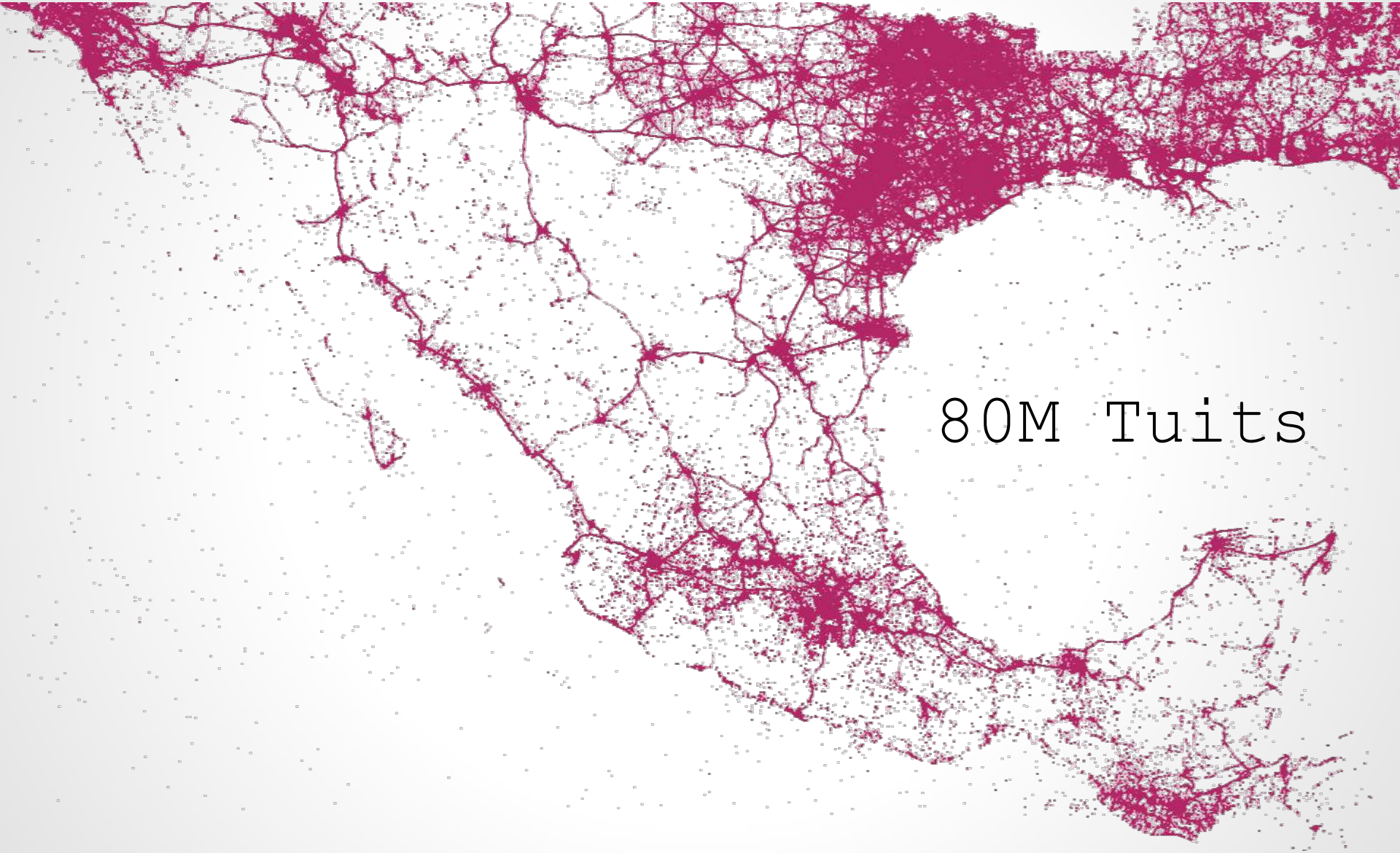
```
20.281523,-100.809407
20.281523,-100.809407
20.281667,-100.809311
20.281479,-100.809394
20.281526,-100.809377
20.281422,-100.809428
20.281478,-100.809406
20.281495,-100.809371
20.281521,-100.80937
25.767972,-103.274890
25.768021,-103.274900
25.768059,-103.274955
25.768019,-103.274900
25.768098,-103.274992
```

Quantum GIS

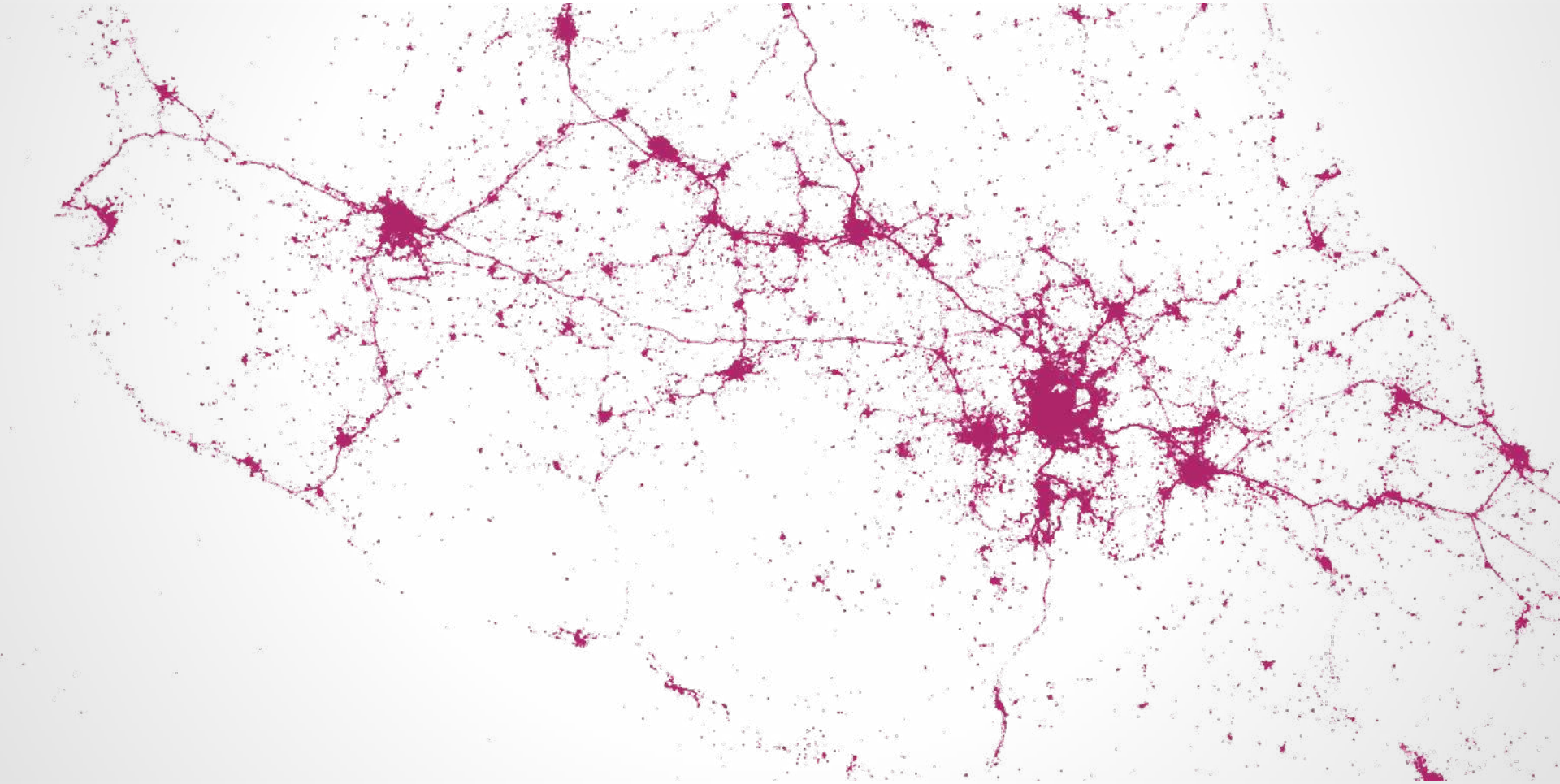


http://www.qgis.org/

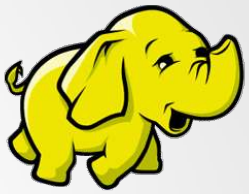
Resultado de la recolección



Un acercamiento



Hadoop Distributed File System
hdfs://



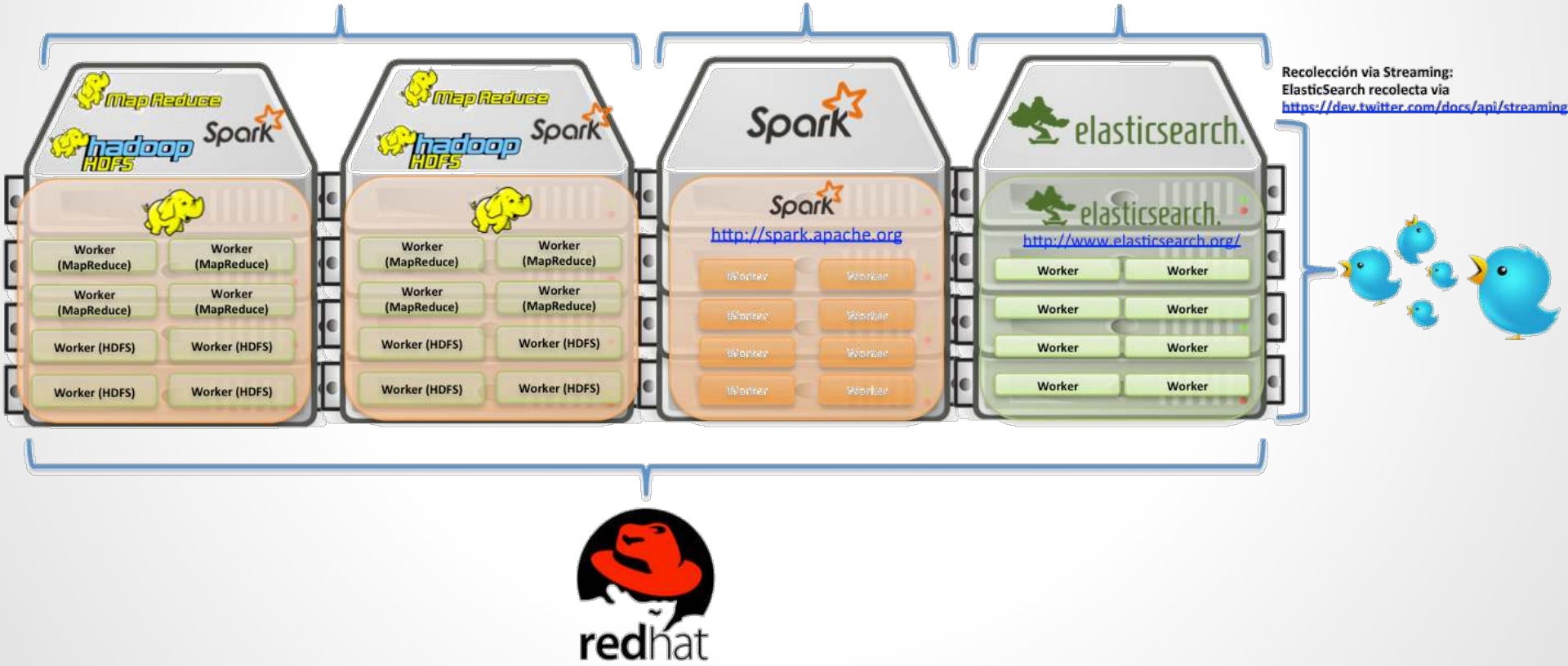
Type	Name	Size	User	Group	Permissions	Date
	.		acoronado	acoronado	drwxr-xr-x	October 06, 2014 02:53 PM
	..		acoronado	acoronado	drwxr-xr-x	September 25, 2014 06:18 AM
	2014-02_al_2014-09-23.csv	18.2 GB	acoronado	acoronado	-rw-r--r--	September 25, 2014 08:57 AM

Hadoop / Apache Spark

Punto de Partida
[Hadoop]:
48 Cores >3 Ghz
128 Gb RAM
4 TB Almacenamiento Permanente

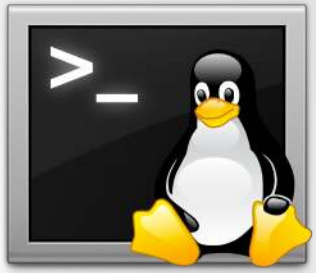
Punto de Partida
[Spark]:
24 Cores >3 Ghz
128 Gb RAM
1 TB Almacenamiento Volátil

Actualmente [ES]:
18 Cores 2.5 Ghz
68 Gb RAM
900 Gb Almacenamiento



Recorte Geográfico


```
object SimpleApp {
def main(args: Array[String]){
  …
  val csvPath = "hdfs://m01/user/acoronado/mov/2014-02_al_2014-09-23.csv"
  val csv = sc.textFile(csvPath)
  csv.cache()
  val clipPoints = csv.map({line: String =>
      val Array(usuario, lat, lon, date) = line.split(",").map(_.trim)
      val geometryFactory = JTSFactoryFinder.getGeometryFactory();
      val reader = new WKTReader(geometryFactory);
      val point = reader.read("POINT ("+lon+" "+ lat + ")" )
      val envelope = point.getEnvelopeInternal
      val internal = geoDataMun.get(envelope)
      val (cve_est, cve_mun) = internal match {
      case l => {
           val existe = l.find( f => f match { case (g:Geometry,e:String,m:String) => g.intersects(point)
                                               case _ => false} )
                                existe match {
                                                  case Some(t)  => t match {
                                                  case (g:Geometry,e:String,m:String) => (e,m)
                                                                  case _ => ("0","0")}
                                           case None => ("0", "0")
                                                       }
                                                  }
                  case _ => ("0", "0")
                                }
                   val time = …
                   line+","+time+","+cve_est+","+cve_mun
                   })
  clipPoints.coalesce(5,true).saveAsTextFile("hdfs://m01/user/acoronado/mov/resultados_movilidad_parts.csv")
  }
}
```

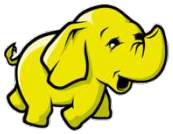


```
cat tweets_feb_sep.csv | awk -F',' '{print $1}'|sort| uniq | wc -l
```

Más de **700,000** tuiteros dentro del territorio Mexicano.

Calcular total de tuits por Hora

```
val csvPath ="hdfs://master/user/acoronado/tweets_feb_sep.csv"
csv.cache
val csv = sc.textFile(csvPath)


val hours =
csv.map({line:String =>
            val campos = line.split(",").map(_.trim)
            val d1 = new Date(campos(8).toLong)
            val format = new SimpleDateFormat("dd-MM-yyyy,HH")
            (format.format(d1),1)}).reduceByKey((a,b) => a+b)



hours.coalesce(1).saveAsTextFile("hdfs://…/days_hours_string.csv")
```



Home / user / acoronado / mov / days_hours_string.csv / **part-00000**

ACTIONS

View as binary

Edit file

Download

View file location

Refresh

INFO

Last modified
Sept. 29, 2014 7:41 a.m.

User
acoronado

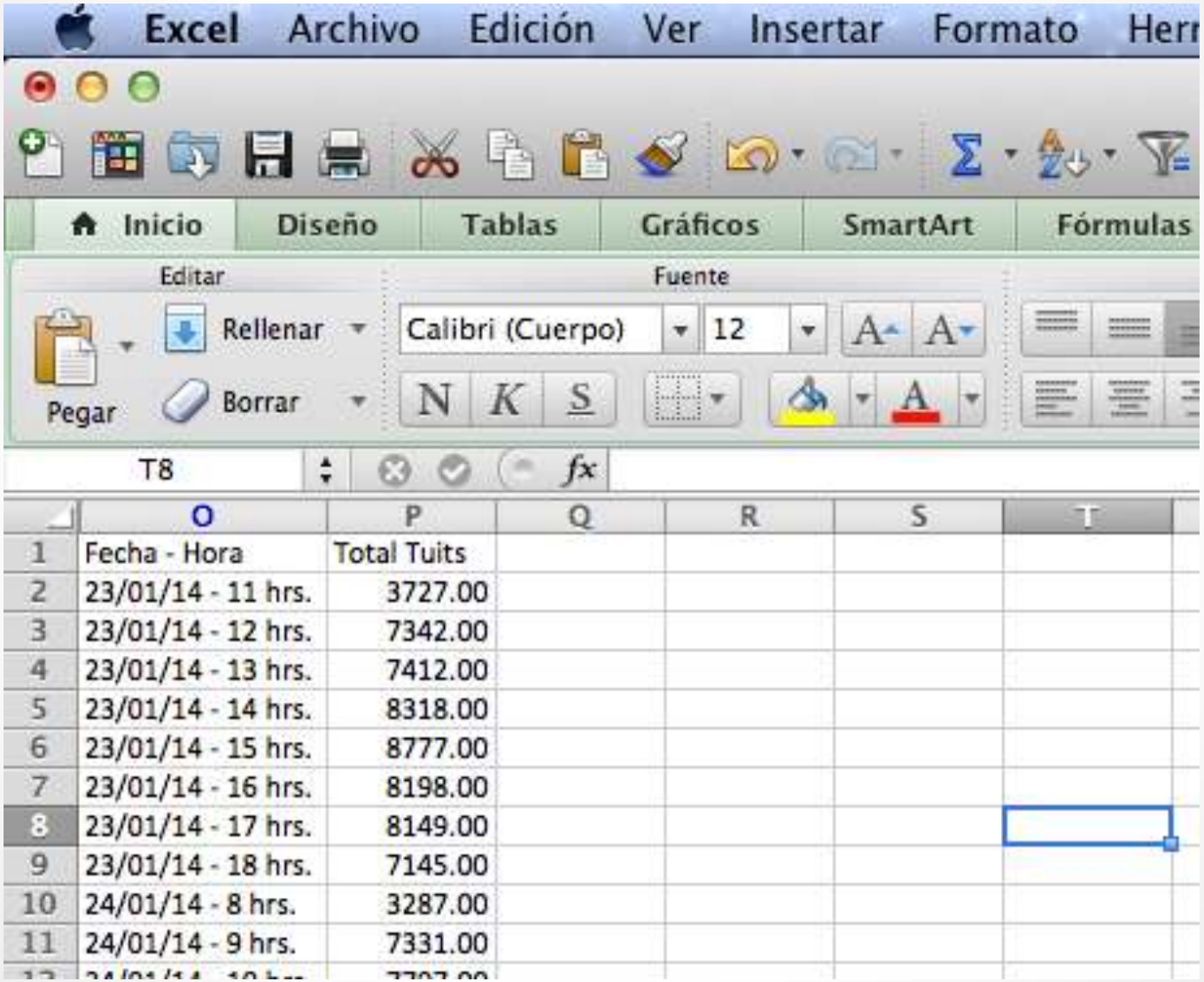
Group
acoronado

Size

First Block | Previous Block | Next Block | Last Block

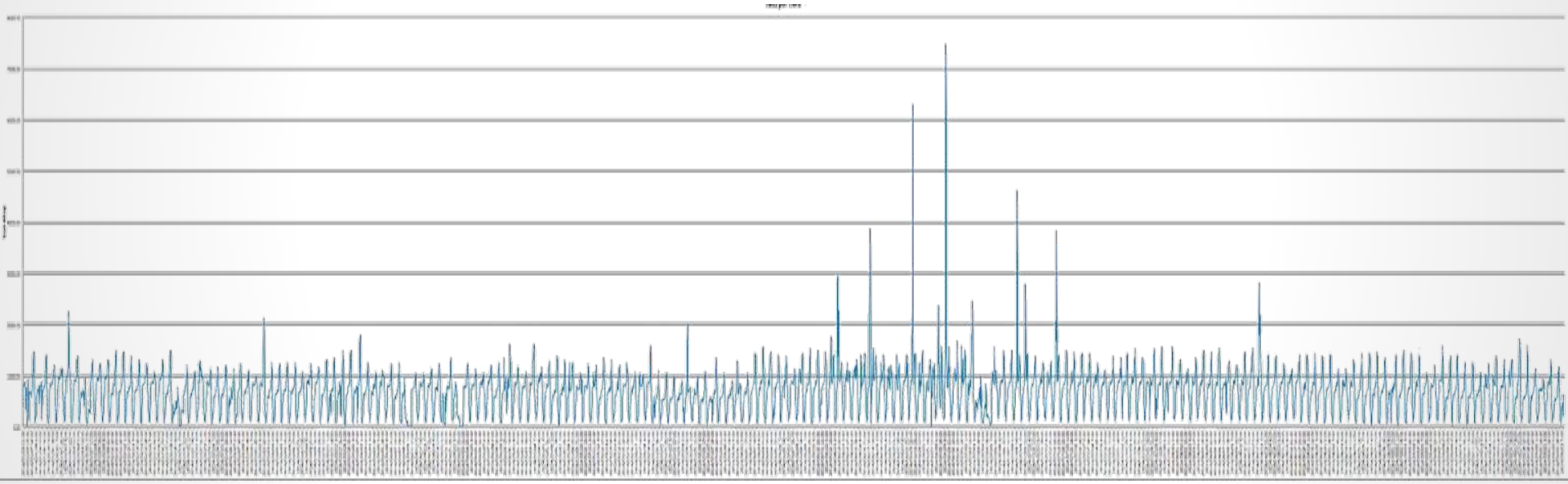
```
(28-08-2014,07,3883)
(05-05-2014,23,12930)
(25-08-2014,08,5085)
(09-06-2014,22,14460)
(06-06-2014,23,11730)
(14-02-2014,20,10515)
(01-07-2014,21,9643)
(22-08-2014,05,788)
(04-04-2014,23,10204)
(03-06-2014,20,12069)
(11-02-2014,21,13744)
(05-08-2014,20,10271)
(21-07-2014,09,5644)
(30-07-2014,03,3516)
(31-05-2014,06,923)
```

Generar la Gráfica

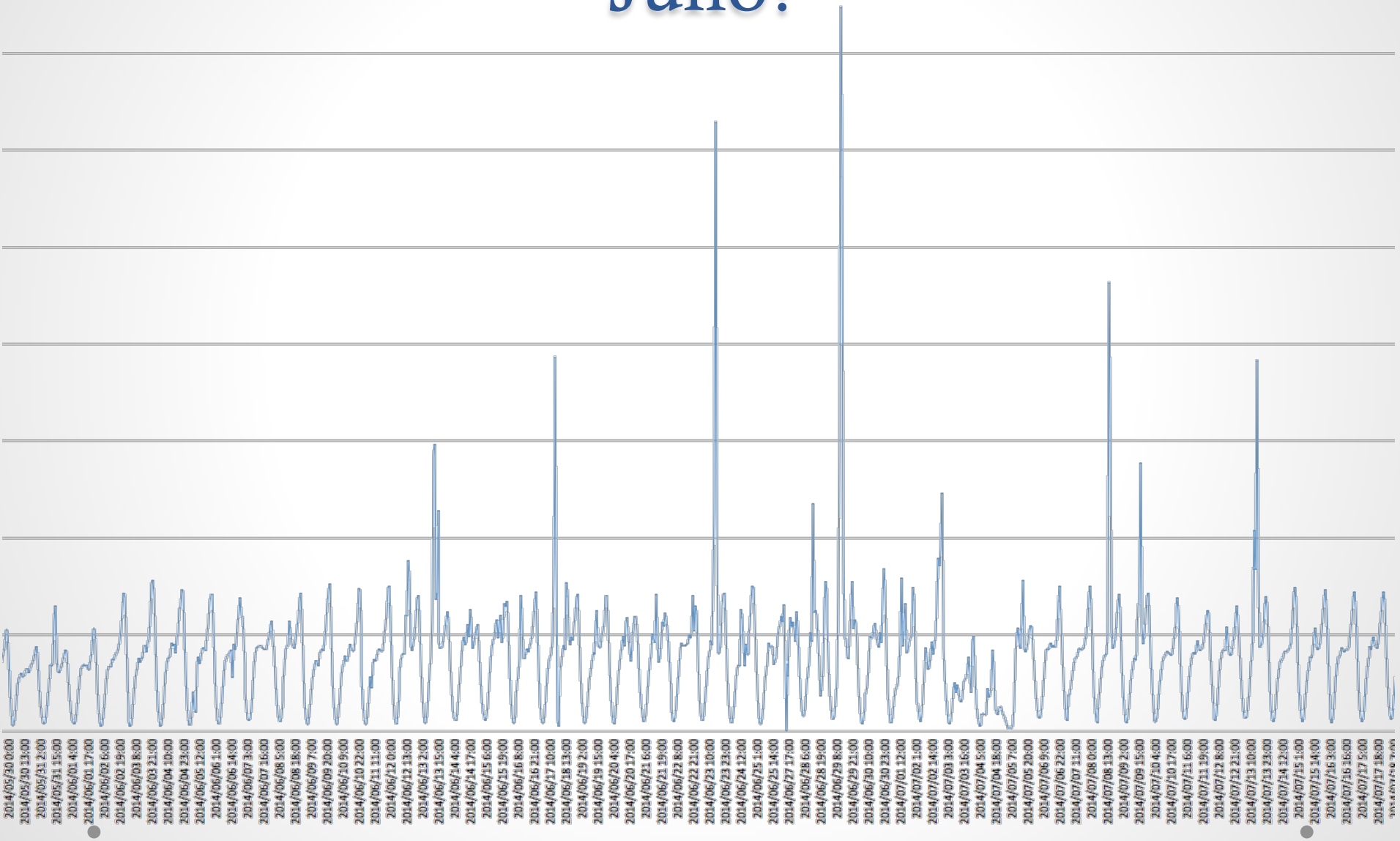


	O	P	Q	R	S	T
1	Fecha - Hora	Total Tuits				
2	23/01/14 - 11 hrs.	3727.00				
3	23/01/14 - 12 hrs.	7342.00				
4	23/01/14 - 13 hrs.	7412.00				
5	23/01/14 - 14 hrs.	8318.00				
6	23/01/14 - 15 hrs.	8777.00				
7	23/01/14 - 16 hrs.	8198.00				
8	23/01/14 - 17 hrs.	8149.00				
9	23/01/14 - 18 hrs.	7145.00				
10	24/01/14 - 8 hrs.	3287.00				
11	24/01/14 - 9 hrs.	7331.00				

A lo largo del tiempo



¿Qué pasó entre el 12 de Junio y el 13 de Julio?



Pregúntale a Twitter

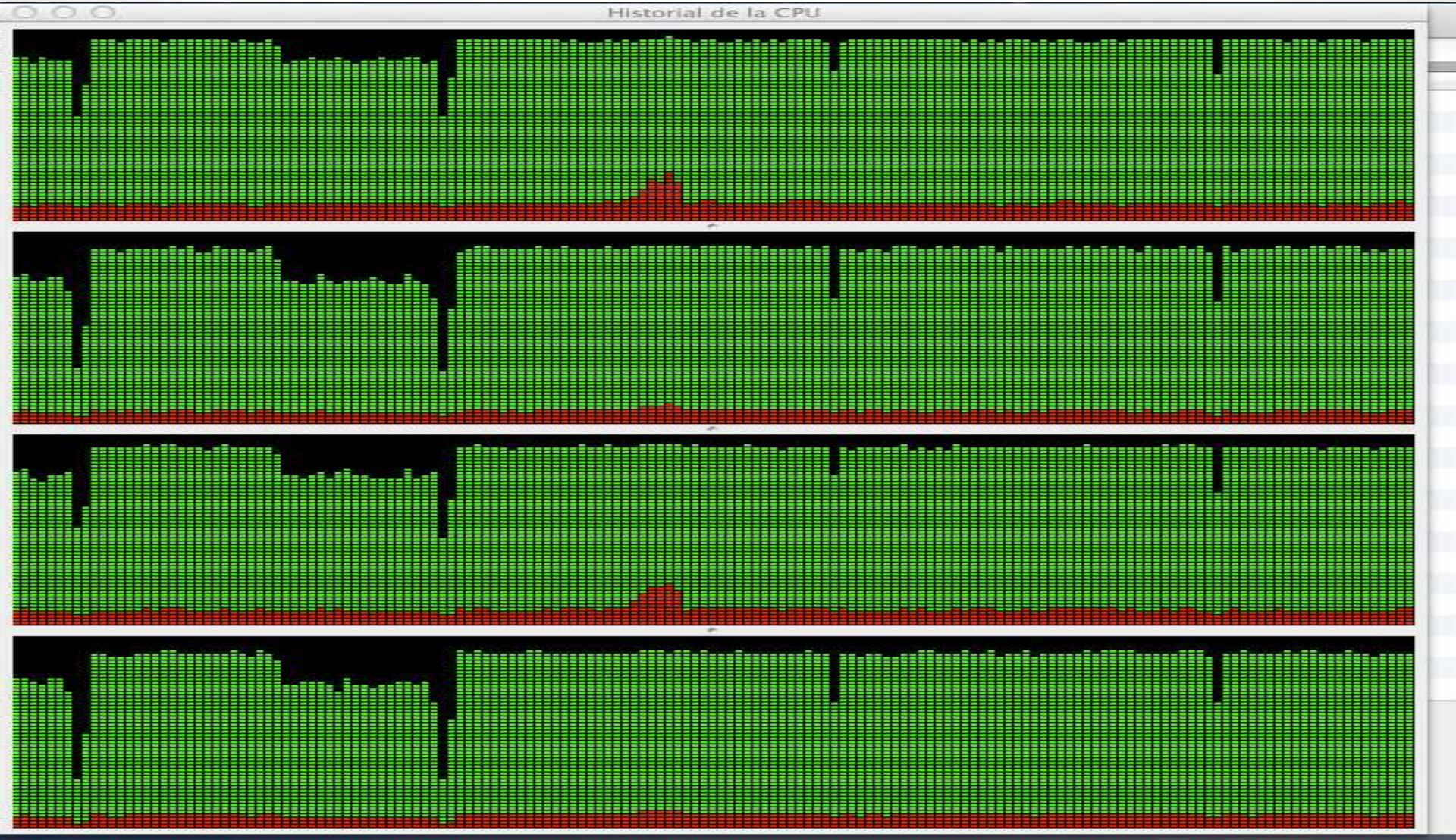
?

Busca tuits en la fecha especifica


```
object Main extends App {
val fecha1 = new SimpleDateFormat("yyyy-MM-dd'T'HH:mm:ss").parse("2014-06-12T00:00:00")
val fecha2 = new SimpleDateFormat("yyyy-MM-dd'T'HH:mm:ss").parse("2014-07-13T23:59:59")
          scala.io.Source.fromFile("/abxda/BigData/tweets_feb_sep_ord_loc.csv")
          .getLines()
          .grouped(250000)
          .flatMap { y=>
                              y.par.filter({line: String =>
                                                          val campos = line.split(",").map(_.trim)
                                                          val time = new Date(campos(8).toLong)
                                                          time.after(fecha1) && time.before(fecha2)
                                                })
                         }.foreach({ x: String =>
                                                println(x.toString)
                                                })
          }
```

Cómputo paralelo

```
y.par.filter
```

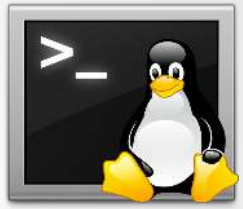


Encuentra Hashtags


```
# coding=utf-8
import codecs
import re
cnt = 0
with codecs.open('/abxda/BigData/Periodo.csv','r','utf-8') as f:
    for line in f:
        try:
            csv = line.split(',')
            text = csv[7]
            hashtags=re.findall(u"#([áéíóúÁÉÍÓÚñÑA-Za-z0-9_]+)",text,re.U)
            for ht in hashtags:
                print '#'+ht
        except Exception:
                pass
```

Prepara archivo para Wordle

http://www.wordle.net/



```
cat hashtagsMundial.txt | sort | uniq -c | sort -n | awk -F' '
'{print $2 ":" $1}' > wordleMun.txt
```

```
#NED:8313
#MundialBrasil2014:8777
#VamosMexico:8947
#BRA:10098
#CallMeCam:14531
#ARG:15663
#Brasil2014:16428
#GER:18030
#MEX:34035
```

¿Qué pasó entre el 12 de junio y el 13 de julio?

http://www.wordle.net/

https://twitter.com/abxda

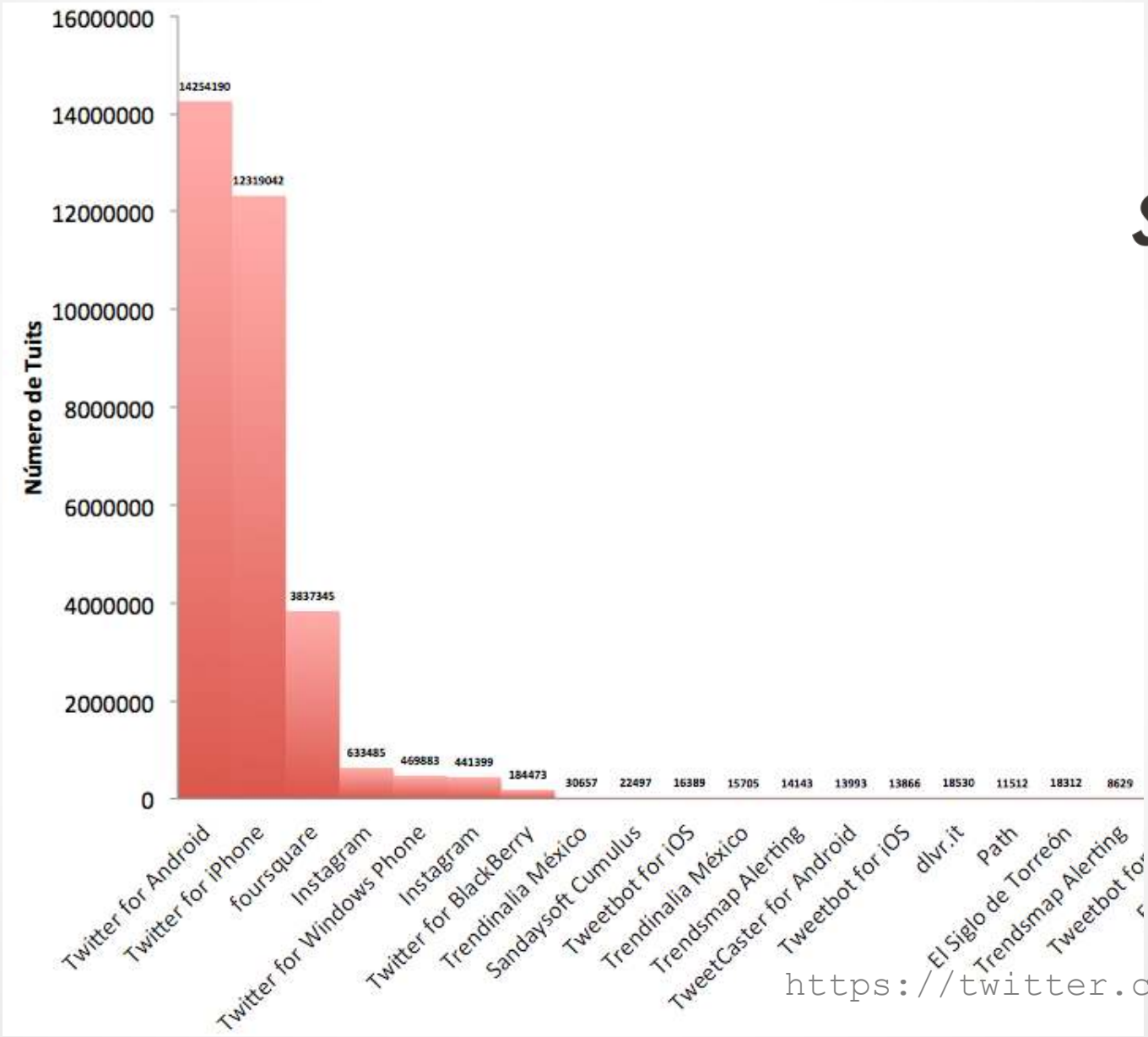
¿Qué pasó el 23 de junio?

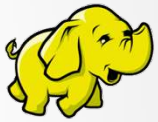
https://twitter.com/abxda

¿Qué pasó el 29 de junio?

https://twitter.com/abxda

¿Con qué tuiteamos?





https://twitter.com/abxda

¿A qué hora tuiteamos?

https://twitter.com/abxda

¿Qué tuiteamos?

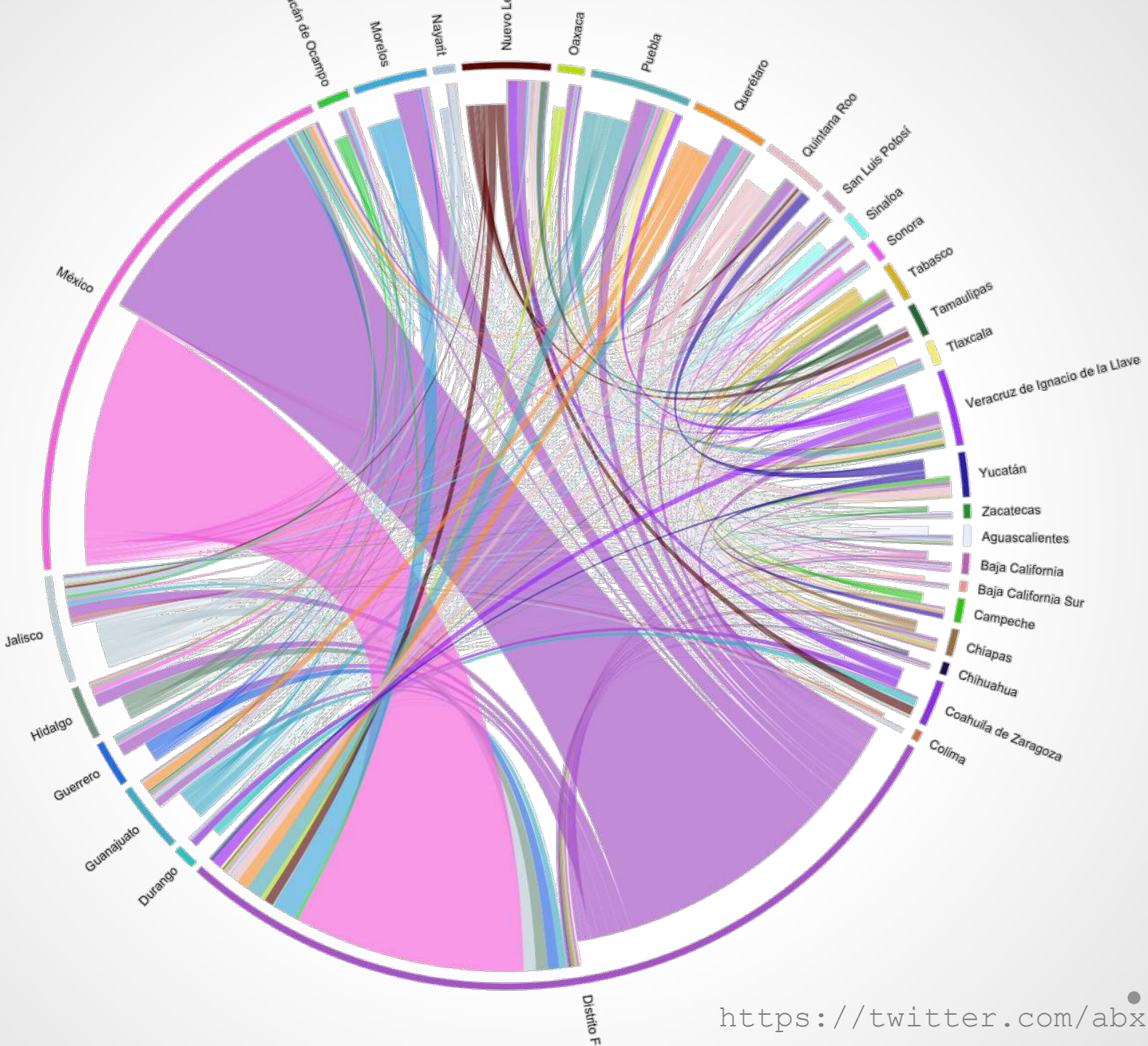
https://twitter.com/abxda

¿Cómo nos desplazamos mientras tuiteamos?

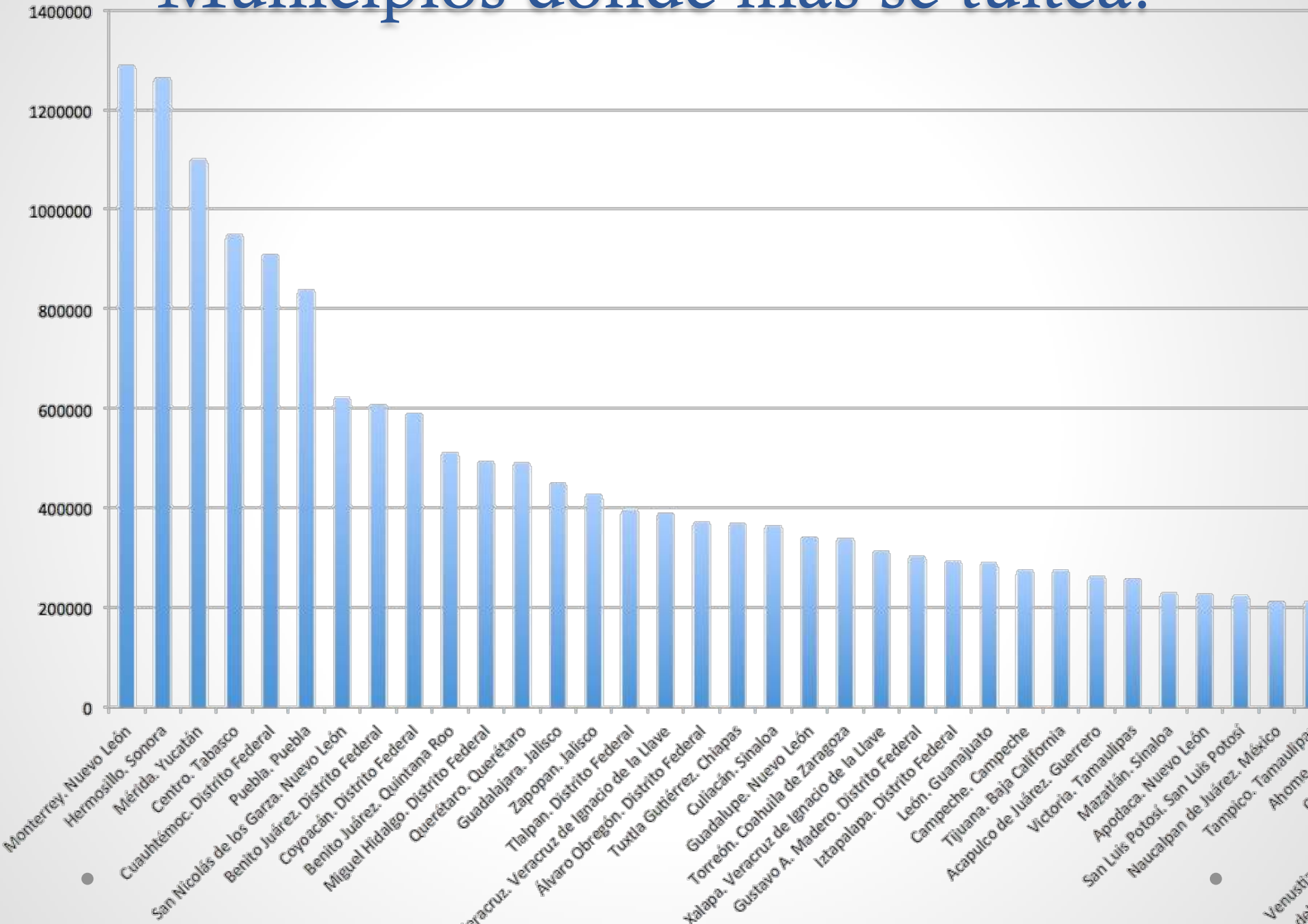
Gráfica de Movilidad

```
library(circlize)
testados = read.table("/abxda/TransladosConDFMexMUNICIPAL.csv", sep=";",
header=TRUE, stringsAsFactors = FALSE, quote = "" )

m = table(testados$estadoorigen, testados$estadodestino)
states = union(rownames(m), colnames(m))
circos.clear()
par(mar = c(1, 1, 1, 1))
chordDiagram(m, directional = TRUE, transparency = 0.3,annotationTrack = "grid",
             annotationTrackHeight = 0.01,
             preAllocateTracks = 1)

for(si in get.all.sector.index()) {
  xlim = get.cell.meta.data("xlim", sector.index = si, track.index = 1)
  ylim = get.cell.meta.data("ylim", sector.index = si, track.index = 1)
  circos.text(mean(xlim), ylim[1], si, facing = "clockwise", adj = c(0, 0.5),
              niceFacing = TRUE, cex = 0.9, col = "black", sector.index = si,
track.index = 1)
}
```

http://cran.r-project.org/web/packages/circlize/vignettes/circlize.pdf



Municipios donde más se tuitea.



Twitter-Bienestar Subjetivo.

Investigación para la utilización de Twitter:

Fuente de datos

- Estructura del tuit
- Disponibilidad
- aleatorización
- filtros georreferenciados

Estudio en otros países

"Análisis de sentimiento" Universidad de Pensilvania

"Mood of the Nation" de los Británicos

"Big Data and Official Statistics" de los Holandeses

"Taller de Análisis de Sentimiento 2013" de la SEPLN

Métodos de clasificadores

Naive Bayes, Support Vector Machines (SVM)

KNN

Word Count

Listas de Palabras y diccionarios utilizados en los ejercicios de análisis de sentimientos

Spanish Emotion Lexicon (SEL)KNN

AFINN

WordNet

ANEW

https://twitter.com/abxda

Twitter-Bienestar Subjetivo.

Proceso de análisis de tuits seleccionado en INEGI:

- No se utilizará el contabilizar y calificar palabras sueltas y tokenizadas (2 o 3 palabras juntas)
- Se probaran métodos supervisados de aprendizaje
- Manualmente se califica el sentimiento y se clasifica el tema de un conjunto de tuits (conjunto de entrenamiento)
- El conjunto de entrenamiento se utiliza para "enseñarle" al sistema a reconocerlos y a utilizarlos por similitudes para calificar y clasificar el resto de los tuits.

Twitter-Bienestar Subjetivo.

Para generar nuestro conjunto de entrenamiento se desarrolló una aplicación para calificar el sentimiento de los tuits en positivo, negativo o neutro, y clasificarlos en varios temas.

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

Acerca del proyecto

Bienvenido

Ayudanos a clasificar tuits

Completa el siguiente formulario para continuar...

Aguascalientes

Masculino

36

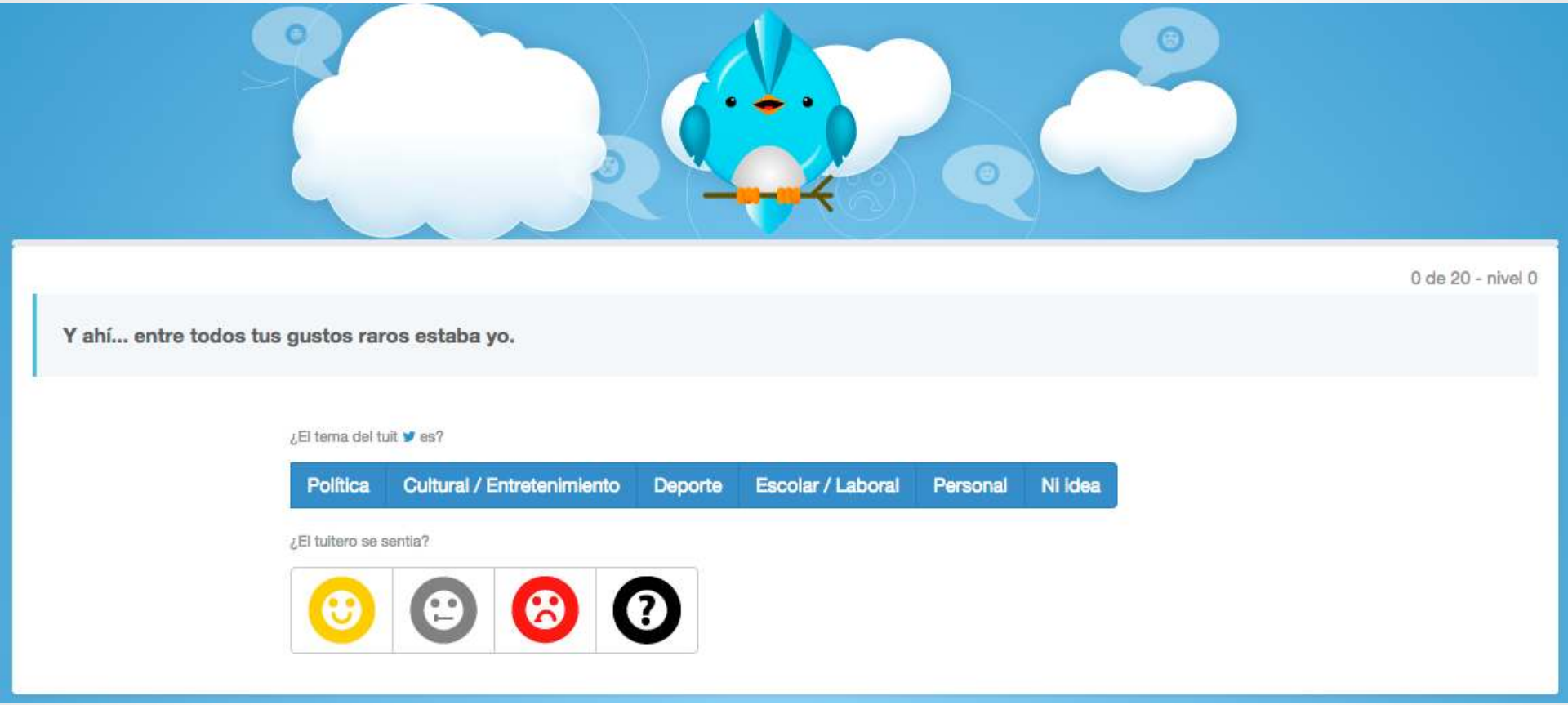
Maestría

881

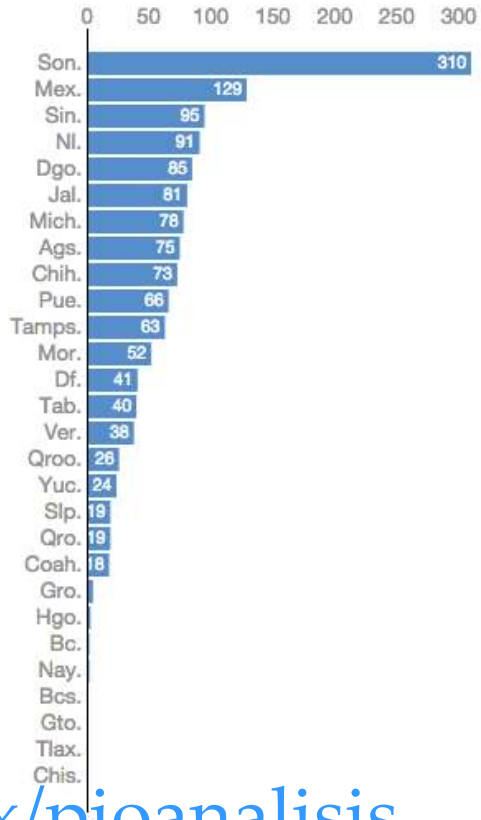
?

Comenzar

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



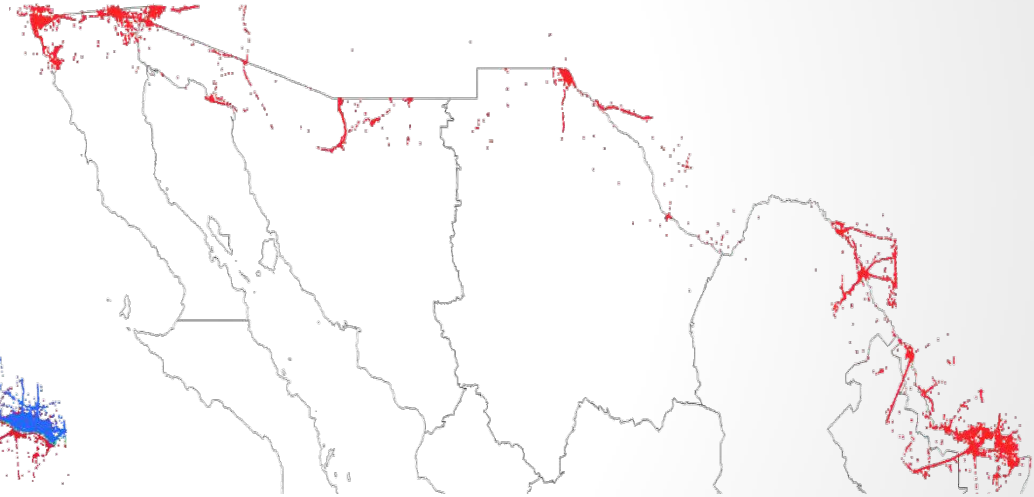
¿Cuántas veces han entrado a Pio Análisis, por estado?



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

Estudios de movilidad.

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



Actividad de los tuiteros en la frontera

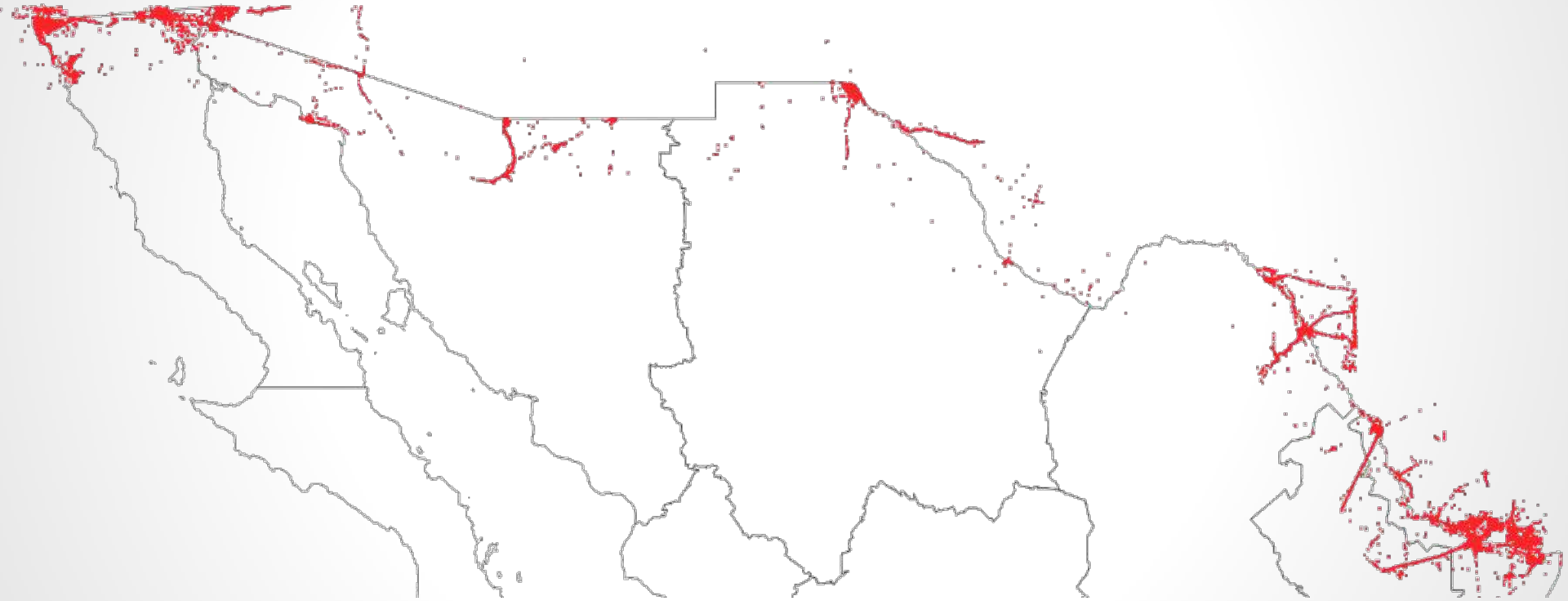
Azul =tuiteros de origen EUA

Rojo=tuiteros de origen MX.

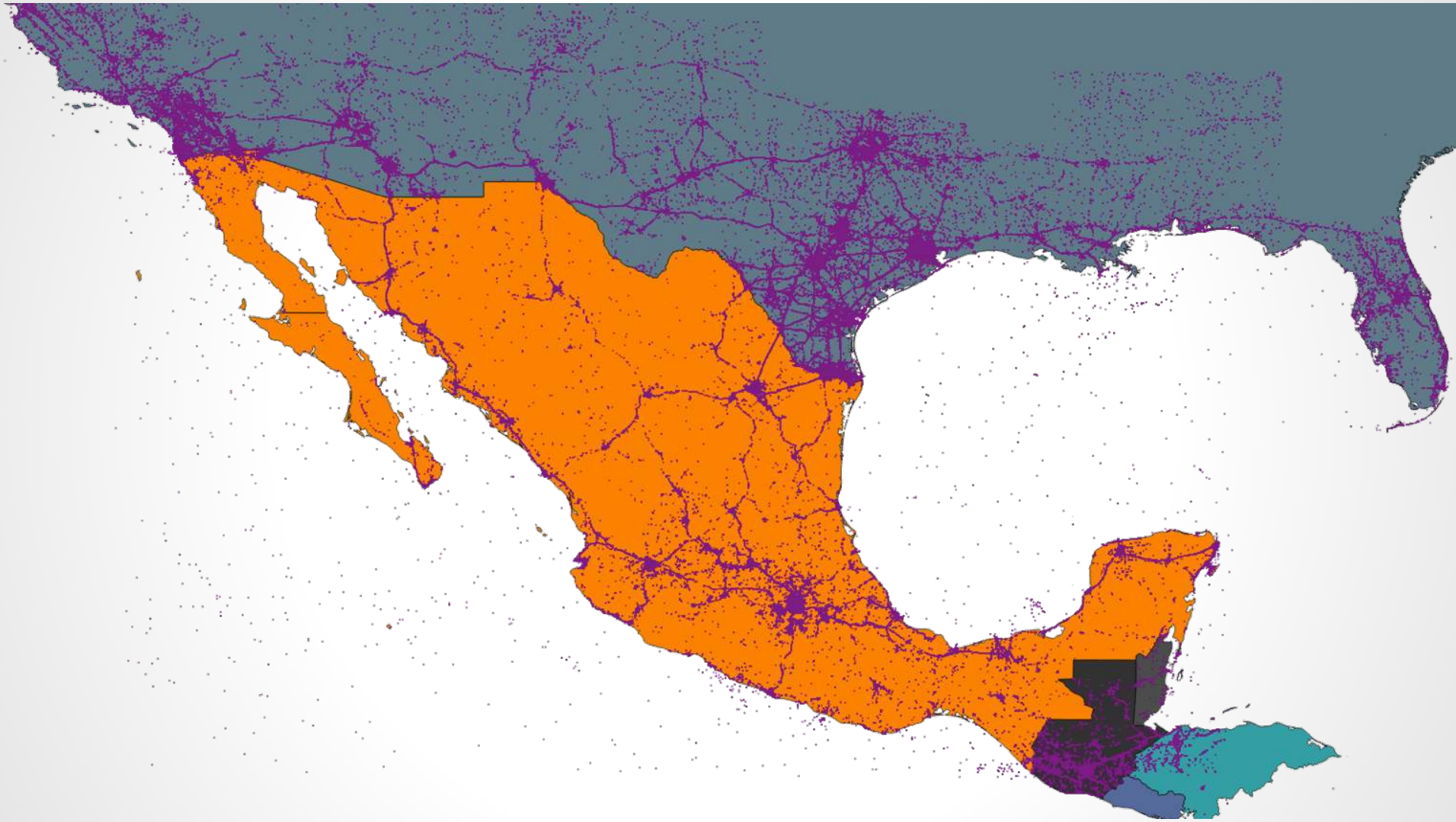
Actividad solamente de tuiteros MX

https://twitter.com/abxda

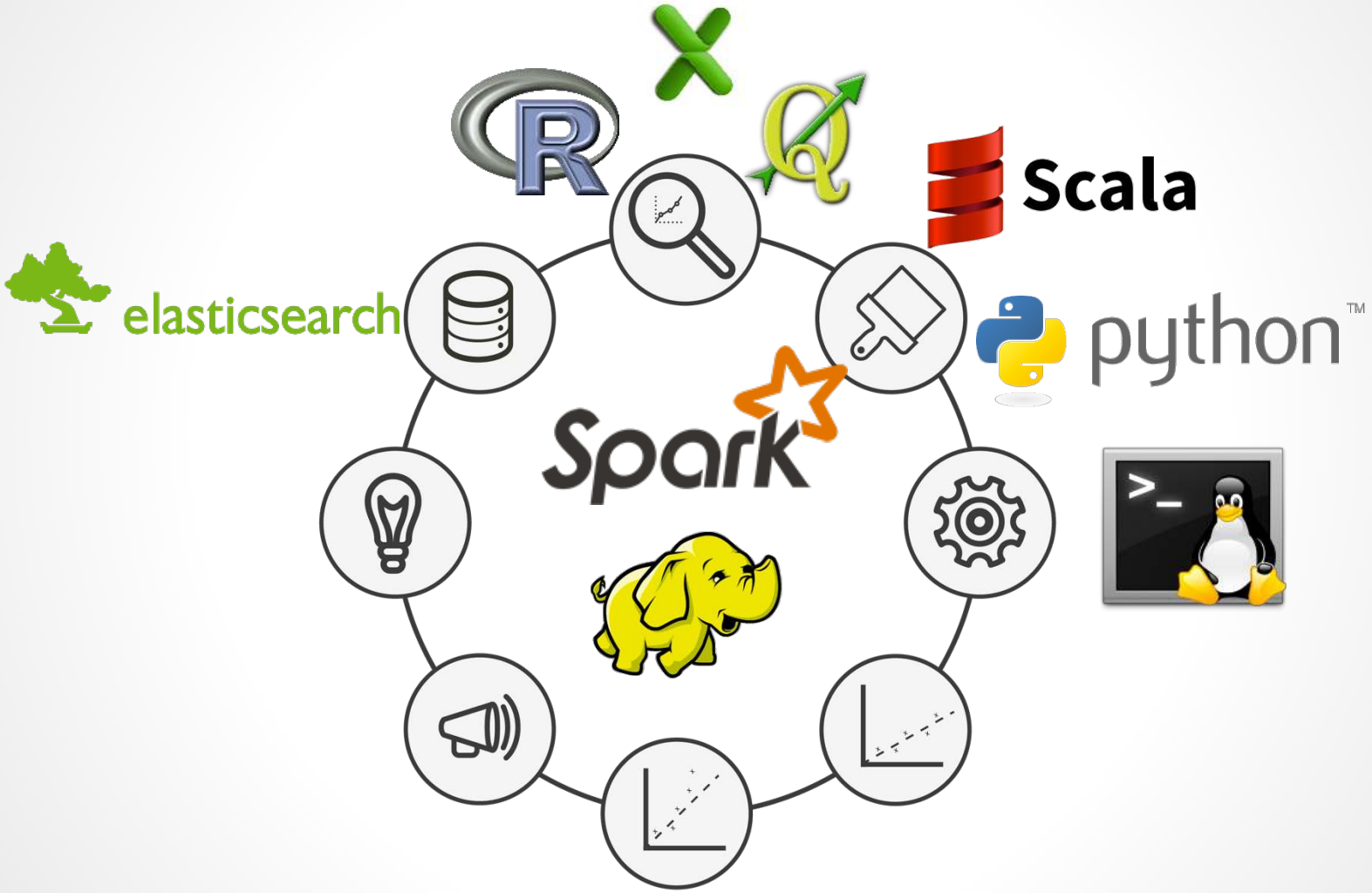
Actividad solamente de tuiteros MX



https://twitter.com/abxda

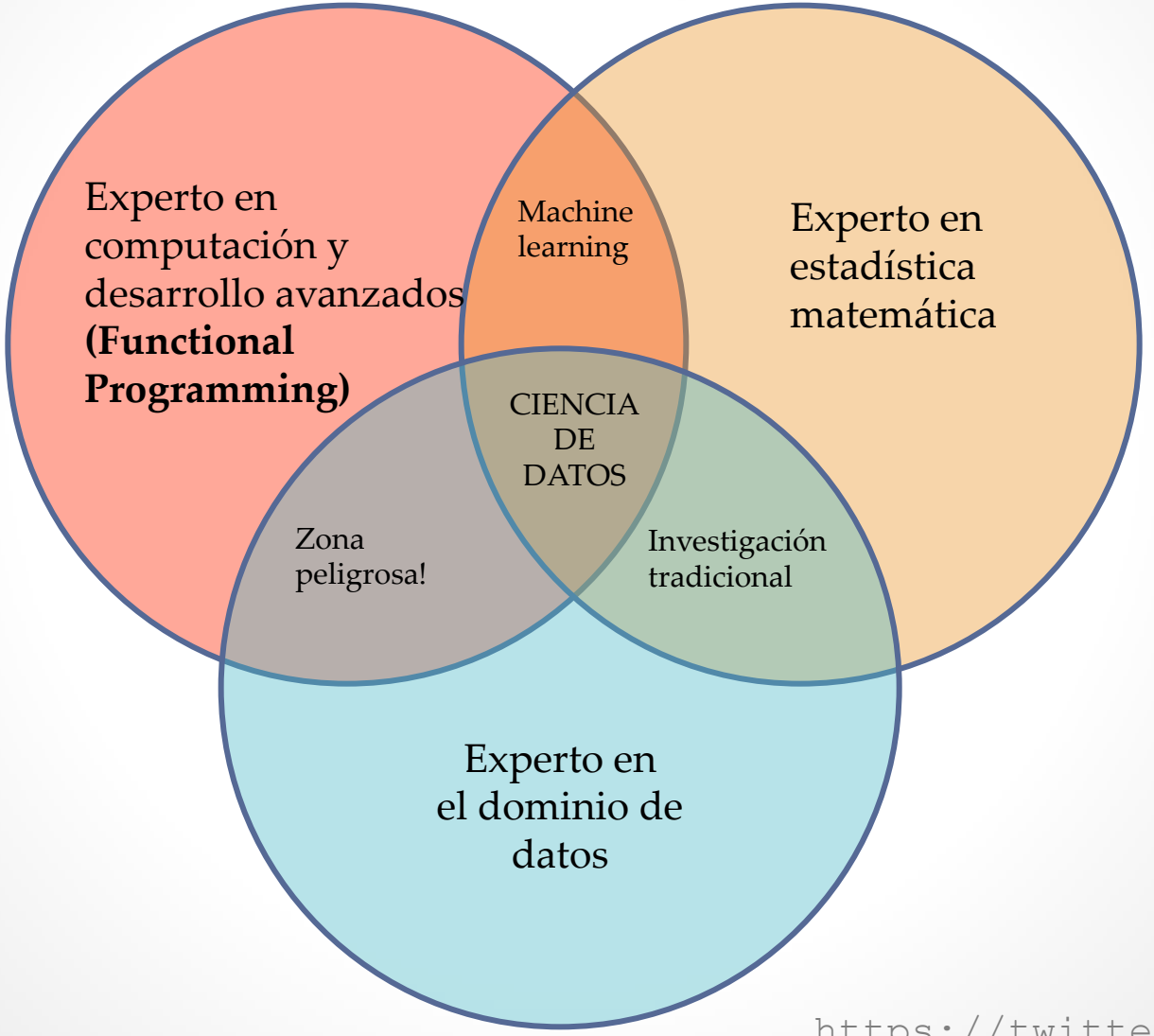


Herramientas



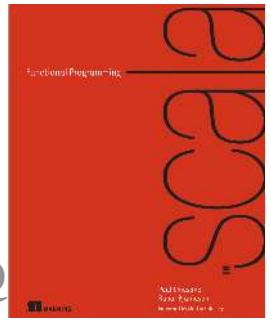
Los Retos:

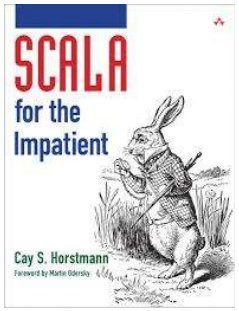
Infraestructura y Personal

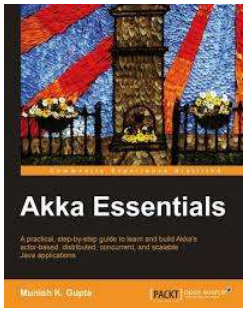


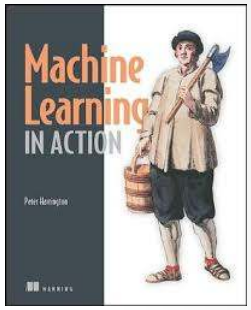
La tarea

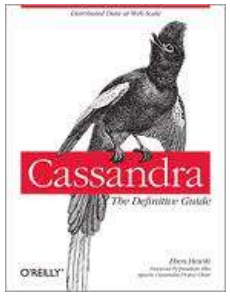
- Programación funcional
  - Scala
  - Akka
- Estadística
  - Probabilidad y Estadística
  - Muestreo
  - Machine Learning
  - R
- Almacenes de Datos NoSQ
  - Cassandra
  - MongoDB
  - Hbase
  - ElasticSearch
- Plataformas Big Data
  - Hadoop
  - Spark
- Visualización de Datos
  - D3.js

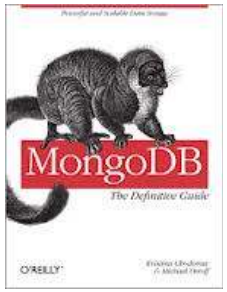


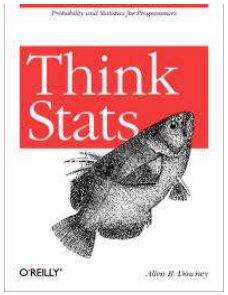


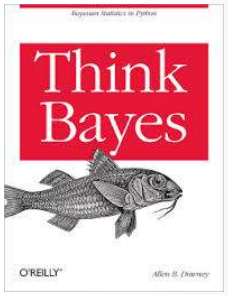


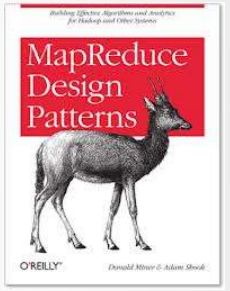


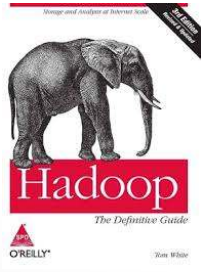


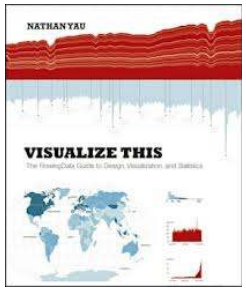


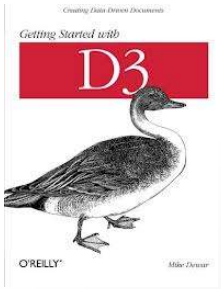


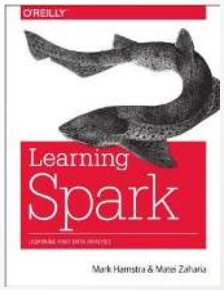


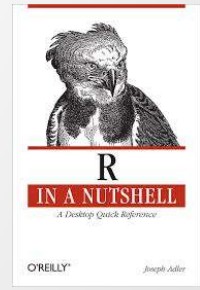












https://twitter.com/abxda

SG VIRTUAL CONFERENCE
7ma edición

Abel Alejandro Coronado Iruegas

@abxda