# Firehose Data Science
# Real-Time Analytics of Twitter Feeds

David Corliss is the founder and Director of Peace-Work, a volunteer cooperative of statisticians and data scientists applying statistical methods to issue-driven advocacy in poverty, education, social justice, and providing analytic support for charitable groups.

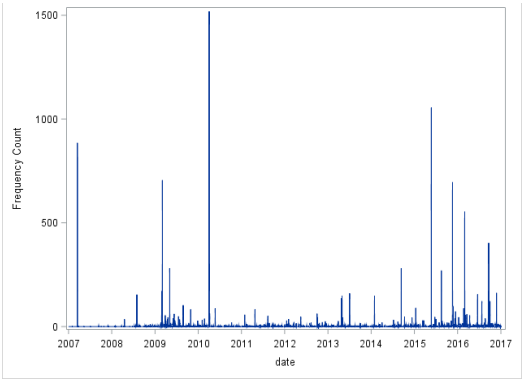
With a PhD in statistical astrophysics, Dr. Corliss works in Manufacturing Forecasting at Ford Motor Company. He is active in the American Statistical Association, includes writing a monthly column on Data For Good for *Amstat News*, serving on the Steering Committee of the Conference on Statistical Practice, and President of the Detroit chapter.

BIG DATA IGNITE

# Firehose Data Science
# Real-Time Analytics of Twitter Feeds

David J Corliss, PhD
Peace-work

# OUTLINE




## Social Media Data

## Creating a Twitter API

## Tweet Classification With ML

## Real Time Analysis

## Hate Tweets and Violence

## Summary

BIG DATA IGNITE

# Social Media Data

**Different Social Media channels have unique characteristics**

**Twitter: Short duration and highly reactive**

**Facebook: medium term, extended conversation threads**

**Google Trends: May be more predictive. The Social Media equivalent of Durable Goods.**

**YouTube – the #2 Social Media channel - and #3**
**Instagram: no effective means to mine images at this time**

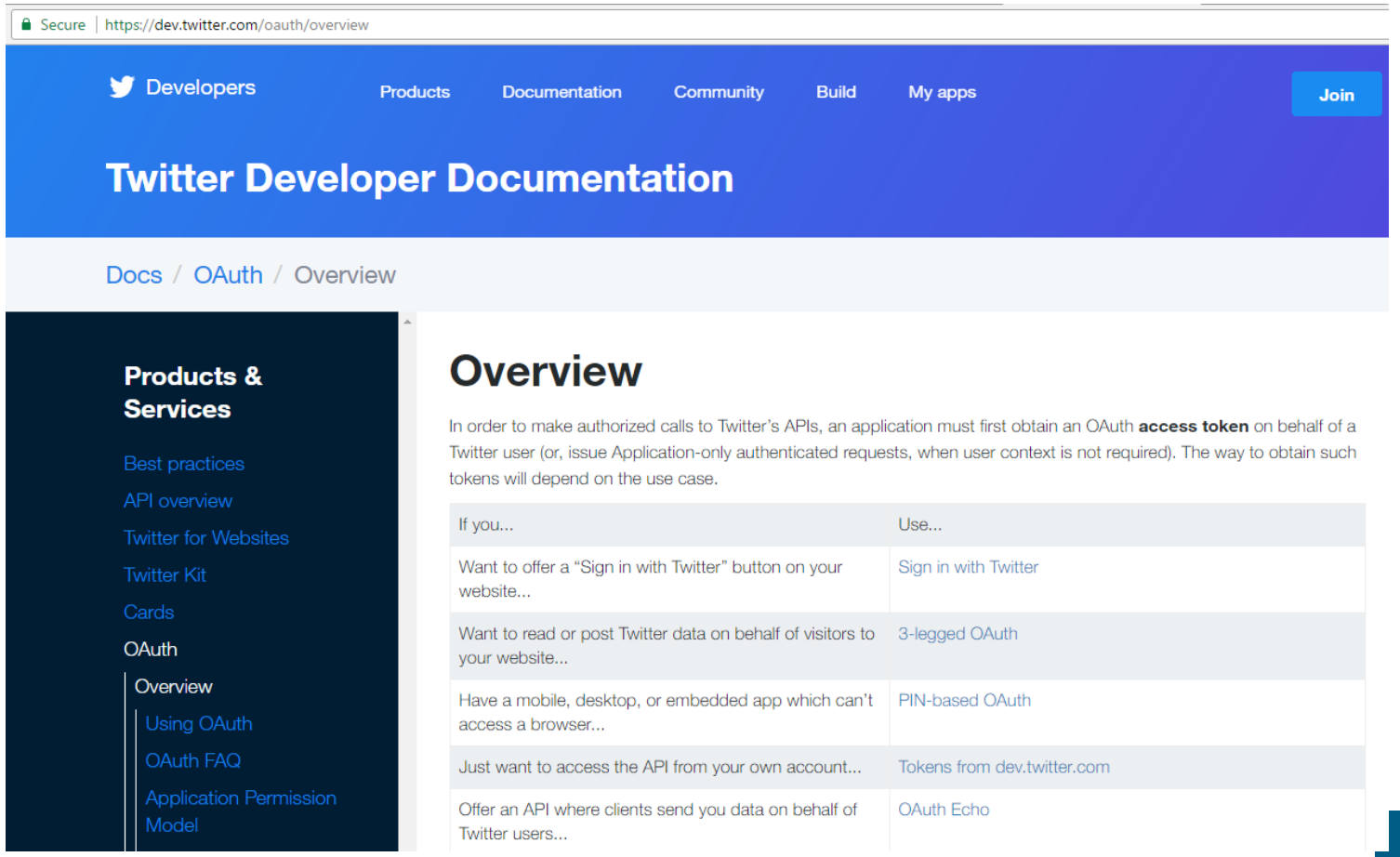
BIG DATA IGNITE

# Mining Twitter Feeds: A Process

1. **Register with Twitter as a developer to get access**
2. **API to deploy search terms and receive tweets**
3. **Parse plain text tweet stream into a dataset / data frame**
4. **Explore data for best search terms, train ML algorithms**
5. **High-volume search using results from #4**
6. **Write loop routine for real-time analysis**



# Step 1: Register With Twitter




https://dev.twitter.com/oauth/overview

# Step 2: Twitter API - Obtain the Tokens

1. **consumerKey: Twitter User ID**
2. **consumerSecret: Twitter login password**
3. **Bearer Token:  Code reads this to access tweets**



# Step 2: Twitter API – SAS Code

```
/* Code adapted from Isabel Litton and Rebecca Ottesen,
   "%GrabTweet: A SAS® Macro to Read JSON Formatted Tweets" */

%MACRO grab_tweet_100(search_term, type, target_amount);

/* Location of Token File */
filename auth "C:\dcorliss\SAS\Twitter\token.txt";

/* Location of Output File */
filename twtOut "C:\dcorliss\SAS\Twitter\Tweets.txt";

/* Set search parameters: COUNT = #, TYPE = Popular, Recent, or Mixed */

%IF &target_amount < 100 %THEN %LET
   num_tweet = %NRSTR(&count=)&target_amount;
%ELSE %LET num_tweet = %NRSTR(&count=100);
%LET type = %NRSTR(&result_type=&type);
```

BIG DATA IGNITE

# Step 2: Twitter API – SAS Code

```
/*Issues GET search/tweet URL with search term and number of tweets specified
by user to output JSON file*/

PROC HTTP
   HEADERIN = auth
   METHOD = "get"

   /* Send requested search terms to the Twitter search API */
   URL =
"https://api.twitter.com/1.1/search/tweets.json?q=&search_term&type&num_tweet"

   OUT = twtOut; /* Output text from tweets as one text string */;

RUN;
```

BIG DATA IGNITE

# Step 2: Twitter API – R Code

```
library(twitteR)

consumer_key <- "your_consumer_key"

consumer_secret <- "your_consumer_secret"

access_token <- "your_access_token"

access_secret <- "your_access_secret"

setup_twitter_oauth(consumer_key, consumer_secret, access_token, access_secret)
```

BIG DATA IGNITE

# Step 2: Twitter API – R Code

```
r_stats <- searchTwitter("#Rstats", n=1500, cainfo="cacert.pem")
# The cainfo parameter is only used in Windows environments
# This should return 1500

length(r_stats)
#[1] 1500

#Save the text output
r_stats_text <- sapply(r_stats, function(x) x$getText())
r_stats_text_corpus <- Corpus(VectorSource(r_stats_text))
```

BIG DATA IGNITE

# Step 3: Parse Plain Text Tweet Stream

## Raw Twitter Text Output

Tweets.txt - Notepad

File Edit Format View Help


{"statuses":[{"created_at":"Thu Jun 08 08:52:55 +0000 2017","id":872738196777426944,"id_str":"872738196777426944","text":"@lsarsour You hate America &amp
ars","url":null,"entities":{"description":{"urls":[]}},"protected":false,"followers_count":2013,"friends_count":2511,"listed_count":26,"created_at":"Wed
idebar_fill_color":"DDEEF6","profile_text_color":"333333","profile_use_background_image":true,"has_extended_profile":false,"default_profile":true,"defaul
"recent"},"source":"\u003ca href=\"http:\/\/twitter.com\/#!\/download\/ipad\" rel=\"nofollow\"\u003eTwitter for iPad\u003c\/a\u003e","in_reply_to_status_
ges\/themes\/theme1\/bg.png","profile_background_image_url_https":"https:\/\/abs.twimg.com\/images\/themes\/theme1\/bg.png","profile_background_tile":fal
guess what...America is on to you and your sharia law promotion.  Your Alinsky tactics are transparent.","truncated":false,"entities":{"hashtags":[],"sym
one":null,"geo_enabled":true,"verified":false,"statuses_count":1015,"lang":"en","contributors_enabled":false,"is_translator":false,"is_translation_enable
s":null,"translator_type":"none"},"geo":null,"coordinates":null,"place":null,"contributors":null,"is_quote_status":false,"retweet_count":1,"favorite_coun
_reply_to_user_id":null,"in_reply_to_user_id_str":null,"in_reply_to_screen_name":null,"user":{"id":1216957579,"id_str":"1216957579","name":"Karen D. Scio
g.com\/profile_background_images\/799702831\/608c64c78f543cfda598bc9311553460.png","profile_background_image_url_https":"https:\/\/pbs.twimg.com\/profile
at":"Wed Jun 07 14:39:45 +0000 2017","id":872463091677540353,"id_str":"872463091677540353","text":"Agree our schools are run by commies & traitors wh
n":"Libertarian Vice Presidential nominee, Best Selling Author, Youtube: https:\/\/t.co\/Cxln91HRMg WAR Now Radio: https:\/\/t.co\/WmcCUVjLrr","url":"htt
r":false,"is_translation_enabled":false,"profile_background_color":"222222","profile_background_image_url":"http:\/\/pbs.twimg.com\/profile_background_im
"place":null,"contributors":null,"is_quote_status":false,"retweet_count":36,"favorite_count":41,"favorited":false,"retweeted":false,"possibly_sensitive":
_reply_to_user_id_str":null,"in_reply_to_screen_name":null,"user":{"id":3290719405,"id_str":"3290719405","name":"Maine4Trumpism","screen_name":"Maine4Tru
s\/theme1\/bg.png","profile_background_image_url_https":"https:\/\/abs.twimg.com\/images\/themes\/theme1\/bg.png","profile_background_tile":false,"profil
k white men are most dangerous demographic, subscribe to primitive totalitaria\u2026 https:\/\/t.co\/r1CHYXsHsb","truncated":true,"entities":{"hashtags":
http:\/\/larrylockman.com","display_url":"larrylockman.com","indices":[0,23]}]},"description":{"urls":[]}},"protected":false,"followers_count":199,"frien
ink_color":"DD2E44","profile_sidebar_border_color":"000000","profile_sidebar_fill_color":"000000","profile_text_color":"000000","profile_use_background_i
2","indices":[95,107]},{"screen_name":"mainegop","name":"Maine GOP","id":22251316,"id_str":"22251316","indices":[120,129]},{"screen_name":"gehrig38","nam
te of Maine \ud83c\udf32 (DM US WITH TIPS) #mepolitics #mepol Recently Featured on @TuckerCarlson & @infowars","url":"https:\/\/t.co\/iuS68vcV1X","entiti
/7fqZZBA1_normal.jpg","profile_image_url_https":"https:\/\/pbs.twimg.com\/profile_images\/872561048095784961\/7fqZZBA1_normal.jpg","profile_banner_url":"
nt":0,"favorited":false,"retweeted":false,"lang":"en"},{"created_at":"Thu Jun 08 02:42:52 +0000 2017","id":872645070616489984,"id_str":"87264507061648998
coggin County GOP. Put Maine Citizens First! #mepolitics Award-Winner! https:\/\/t.co\/CItRSg8f8N","url":"https:\/\/t.co\/PjLJESlnmp","entities":{"url":{
file_background_tile":false,"profile_image_url":"http:\/\/pbs.twimg.com\/profile_images\/779527995501506560\/tqs6TWcE_normal.jpg","profile_image_url_http
,"truncated":true,"entities":{"hashtags":[],"symbols":[],"user_mentions":[],"urls":[{"url":"https:\/\/t.co\/r1CHYXsHsb","expanded_url":"https:\/\/twitter
ected":false,"followers_count":199,"friends_count":95,"listed_count":9,"created_at":"Wed Feb 19 03:09:53 +0000 2014","favourites_count":51,"utc_offset":-
color":"000000","profile_use_background_image":false,"has_extended_profile":false,"default_profile":false,"default_profile_image":false,"following":null,
[120,129]},{"screen_name":"gehrig38","name":"Curt Schilling","id":14945261,"id_str":"14945261","indices":[130,139]}],"urls":[{"url":"https:\/\/t.co\/NvM4
url":"https:\/\/t.co\/iuS68vcV1X","entities":{"url":{"urls":[{"url":"https:\/\/t.co\/iuS68vcV1X","expanded_url":"http:\/\/mainefirstmedia.wordpress.com",
qZZBA1_normal.jpg","profile_banner_url":"https:\/\/pbs.twimg.com\/profile_banners\/816402981851660288\/1496869729","profile_link_color":"ABB8C2","profile
9091027914753,"id_str":"872639091027914753","text":"RT @WayneRoot: Agree our schools are run by commies & traitors who hate America & Christianit


BIG DATA IGNITE

# Step 3: Parse Plain Text Tweet Stream

```
/* Code adapted from Isabel Litton and Rebecca Ottesen,
   "%GrabTweet: A SAS® Macro to Read JSON Formatted Tweets" */

   INPUT
      @'"created_at":' date_time_text
      @'"id":' tweet_id
      @'"text":' text
      @'"user":{"id":' user_id
      @'"name":' name
      @'"screen_name":' screen_name
      @'"location":' location
      @'"created_at":’ account_created
      @'"lang":' lang
      @'"contributors":null,' retweeted @@
      ;
```

BIG DATA IGNITE

# Step 4: Exploring Search Terms

```
**** Mining Twitter for Hate Speech ****;

%grab_tweet_100(Sharia America hate, recent, 100);
```

| | created_at1 | tweet_id | text | user_id | name | screen_name | location |
|---|---|---|---|---|---|---|---|
| 1 | Thu Jun 08 08:52:55 +0000 2017 | 872738196777426944 | @Isarsour You hate America & support Sharia here. Your time will come. | 1223623118 | Michael Erwin | michaeloferwins | California, USA |
| 2 | Thu Jun 08 03:52:00 +0000 2017 | 872662466714624001 | RT @LMARocks: @Isarsour Not a hate crime, and guess what...America is on to you and your sharia law promotion. Your Alinsky tactics are tr\ | 4223111532 | PorkChop Express | China_is_here | Wing Kong Exchange |
| 3 | Mon Feb 23 03:30:48 +0000 2009 | 872660605043449856 | RT @WayneRoot: Agree our schools are run by commies & traitors who hate America & Christianity? What exactly do you call this?\nhttp | 1216957579 | Karen D. Scioscia | KDScioscia | |
| 4 | Wed May 06 21:43:22 +0000 2009 | 872645595068067840 | RT @lockman4mehouse: Many \"new Mainers\" hate America | 3290719405 | Maine4Trumpism | Maine4Trump | #ME02 Moscow, Maine |
| 5 | Wed Feb 19 03:09:53 +0000 2014 | 872598556640780291 | #NewRelease \n\nMaine's Refugee Community: The Best And The Brightest \n\nhttps:\/\/t.co\/NvM4koRXXR\n\n #mepolitics @mainegop @ | 816402981851660 | Maine First Media | MaineFirstMedia | Waterville, ME |
| 6 | Thu Jun 08 02:42:52 +0000 2017 | 872645070616489984 | RT @lockman4mehouse: Many \"new Mainers\" hate America | 779527044287848 | Androscoggin GOP | AndroGOP | Maine, USA |
| 7 | Wed Feb 19 03:09:53 +0000 2014 | 872598556640780291 | #NewRelease \n\nMaine's Refugee Community: The Best And The Brightest \n\nhttps:\/\/t.co\/NvM4koRXXR\n\n #mepolitics @mainegop @ | 816402981851660 | Maine First Media | MaineFirstMedia | Waterville, ME |
| 8 | Thu Jun 08 02:19:06 +0000 2017 | 872639091027914753 | RT @WayneRoot: Agree our schools are run by commies & traitors who hate America & Christianity? What exactly do you call this?\nhttp | 2937854619 | Sammie Snickers | Sammie_Snickers | |
| 9 | Wed May 06 21:43:22 +0000 2009 | 872610362209837056 | RT @lockman4mehouse: Many \"new Mainers\" hate America | 816402981851660 | Maine First Media | MaineFirstMedia | Waterville, ME |
| 10 | Wed Feb 19 03:09:53 +0000 2014 | 872598556640780291 | #NewRelease \n\nMaine's Refugee Community: The Best And The Brightest \n\nhttps:\/\/t.co\/NvM4koRXXR\n\n #mepolitics @mainegop @ | 816402981851660 | Maine First Media | MaineFirstMedia | Waterville, ME |
| 11 | Thu Jun 08 00:24:21 +0000 2017 | 872610213991501828 | Many \"new Mainers\" hate America | 2351033034 | Larry Lockman | lockman4mehouse | |
| 12 | Wed Jun 07 23:38:02 +0000 2017 | 872598556640780291 | #NewRelease \n\nMaine's Refugee Community: The Best And The Brightest \n\nhttps:\/\/t.co\/NvM4koRXXR\n\n #mepolitics @mainegop @ | 816402981851660 | Maine First Media | MaineFirstMedia | Waterville, ME |

# Step 4: ML Classification w/ Decision Tree

```
proc hpsplit data=hate_tweets maxdepth=5;

   model hate_tweet(event='1') = hate_term slur expletive nationalist;

   prune costcomplexity;

   partition fraction(validate=0.3 seed=123);

   code file='hate_id.sas';

   rules file='hate_tweet_rules.txt';

run;
```

BIG DATA IGNITE

# Step 4: ML Classification w/ Decision Tree

```
# Use rpart package
library(rpart)

# Model
fit <- rpart(hate_tweet ~ hate_term + slur + expletive + nationalist,
   method="anova", data=hate_tweets)

# Develop visualizations
printcp(fit) # display the results
plotcp(fit) # visualize cross-validation results
summary(fit) # detailed summary of splits

# Plot results
plot(fit, uniform=TRUE, main="Regression Tree for Mileage ")
text(fit, use.n=TRUE, all=TRUE, cex=.8)
```

BIG DATA IGNITE

# Step 5: High-Volume Search

1. **Macro to extract to access Twitter-imposed limit of 100 tweets at a time**

2. **Macro wrapper to repeat until specified count, going backwards from end of previous iteration**

3. **Data transformation required for some analyses e.g., time series dataset**

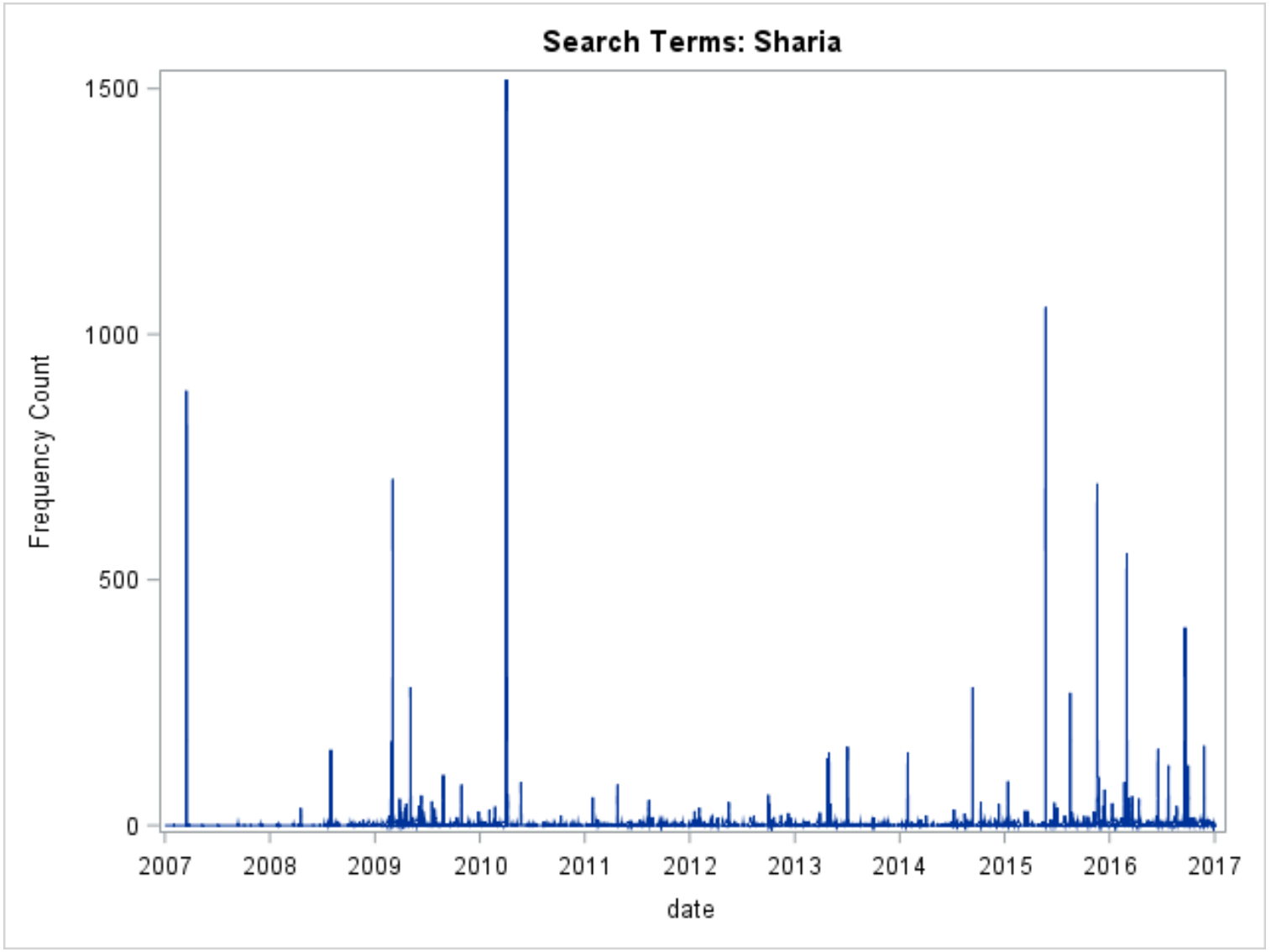
BIG DATA IGNITE

# Step 6: Real-Time Twitter Mining with ML

1. **Label data in a small, randomly-selected training set**
2. **Train an algorithm on the labelled data**
3. **Develop routine to capture tweets, classify, and output**
4. **Loop the routine to extract tweets and output in real-time**

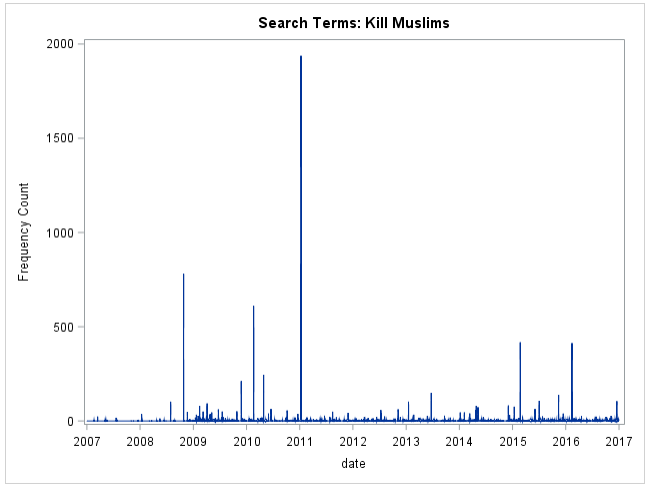
BIG DATA IGNITE

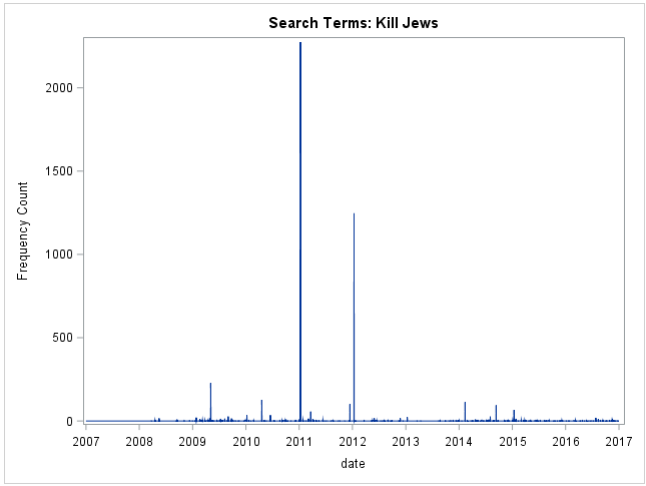
# Analytic Results

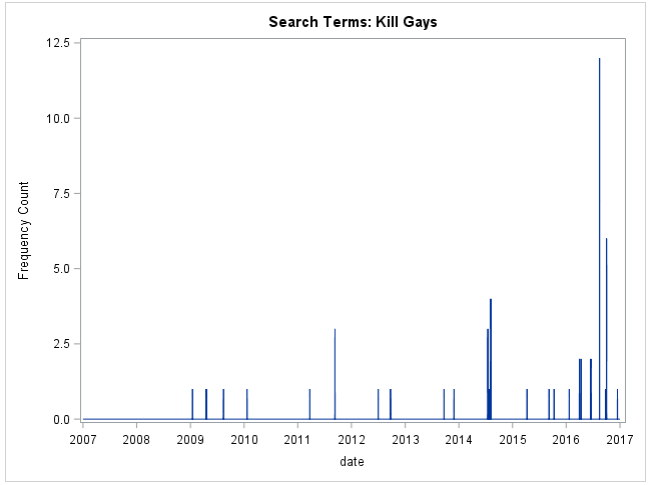


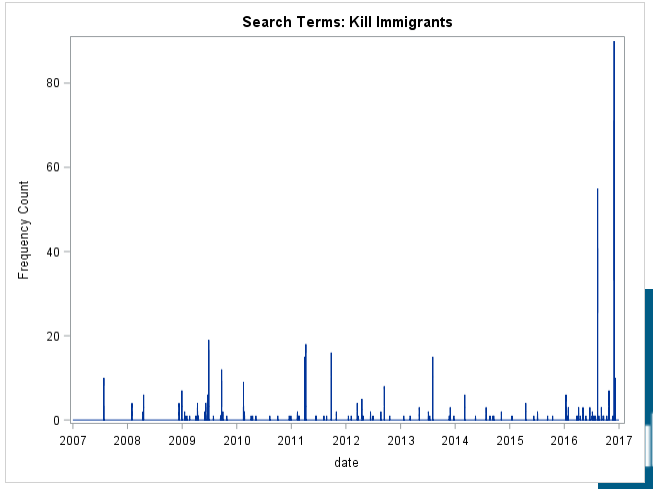

BIG DATA IGNITE

# Analytic Results



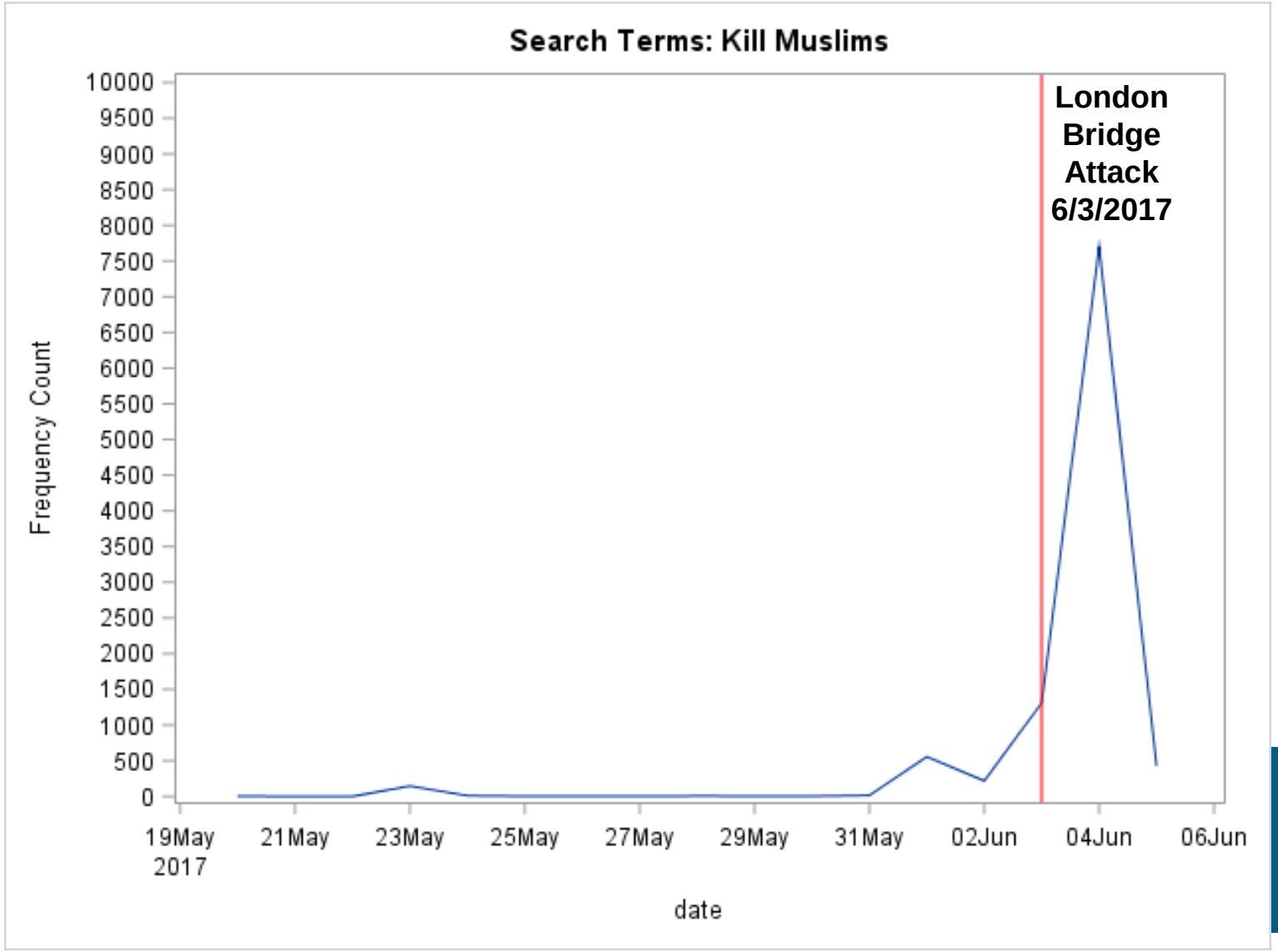




BIG DATA IGNITE

# Analytic Results




BIG DATA IGNITE

# Step 6: Real-Time Twitter Mining with ML

```
%macro timeloop(iterations);
%local i;
%let i=1;
%do i=1 %to &iterations;

%grab_tweet_100(Sharia,recent,100);

data work.tweet;
   set work.tweet;
   count = 1;

chardate=CATT(substr(dttext,9,2),substr(dttext,5,3),substr(dttext,27,4));
timestamp=dhms(input(chardate,date9.), substr(dttime_text,12,2),
substr(dttext,12,2), substr(dttext,18,2));
   date = datepart(timestamp);
   time = timepart(timestamp);
   ten_sec = round(time, 10);
   ten_sec_est = ten_sec - 18000;
```

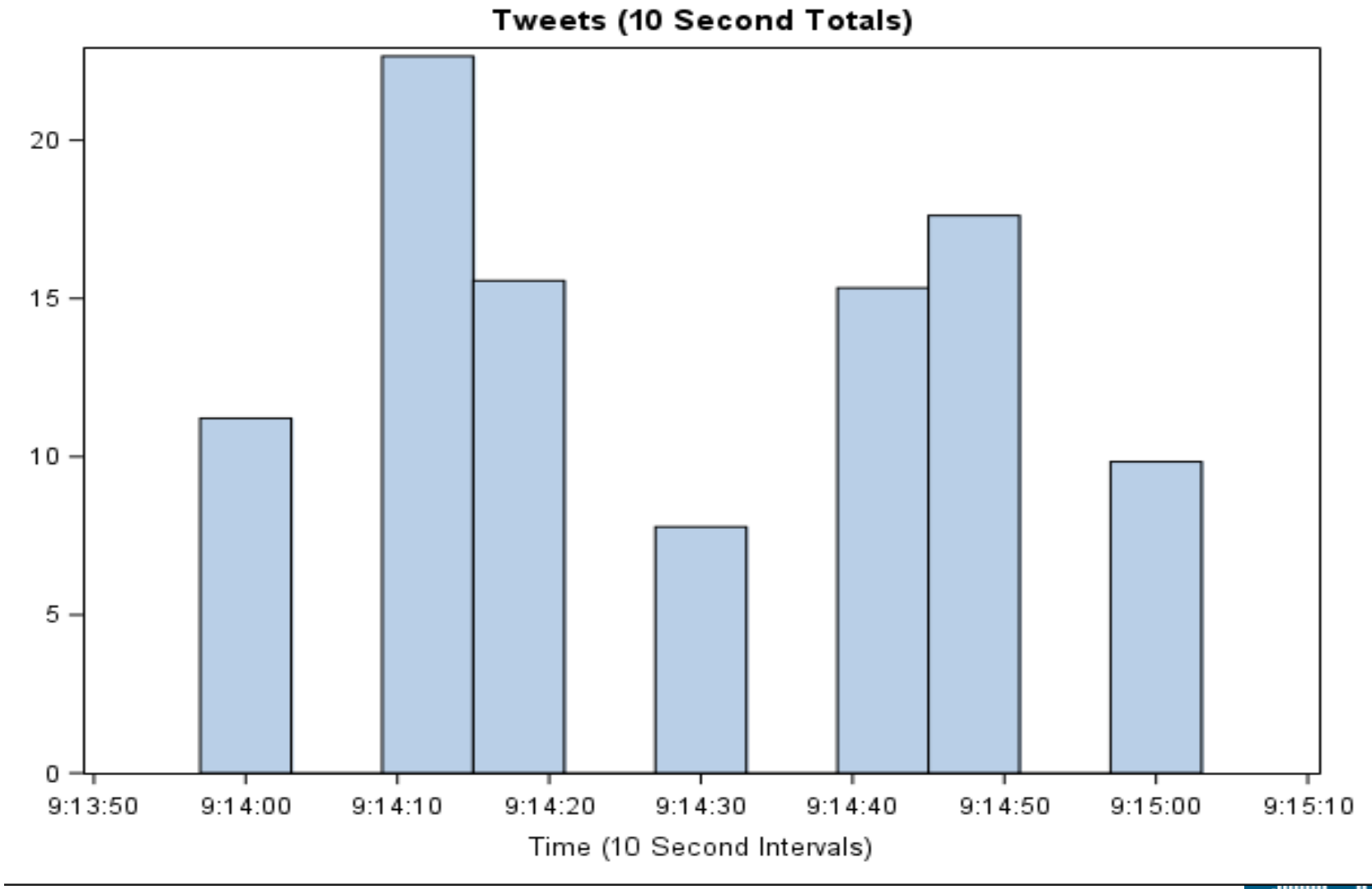
BIG DATA IGNITE

# Step 6: Real-Time Twitter Mining with ML

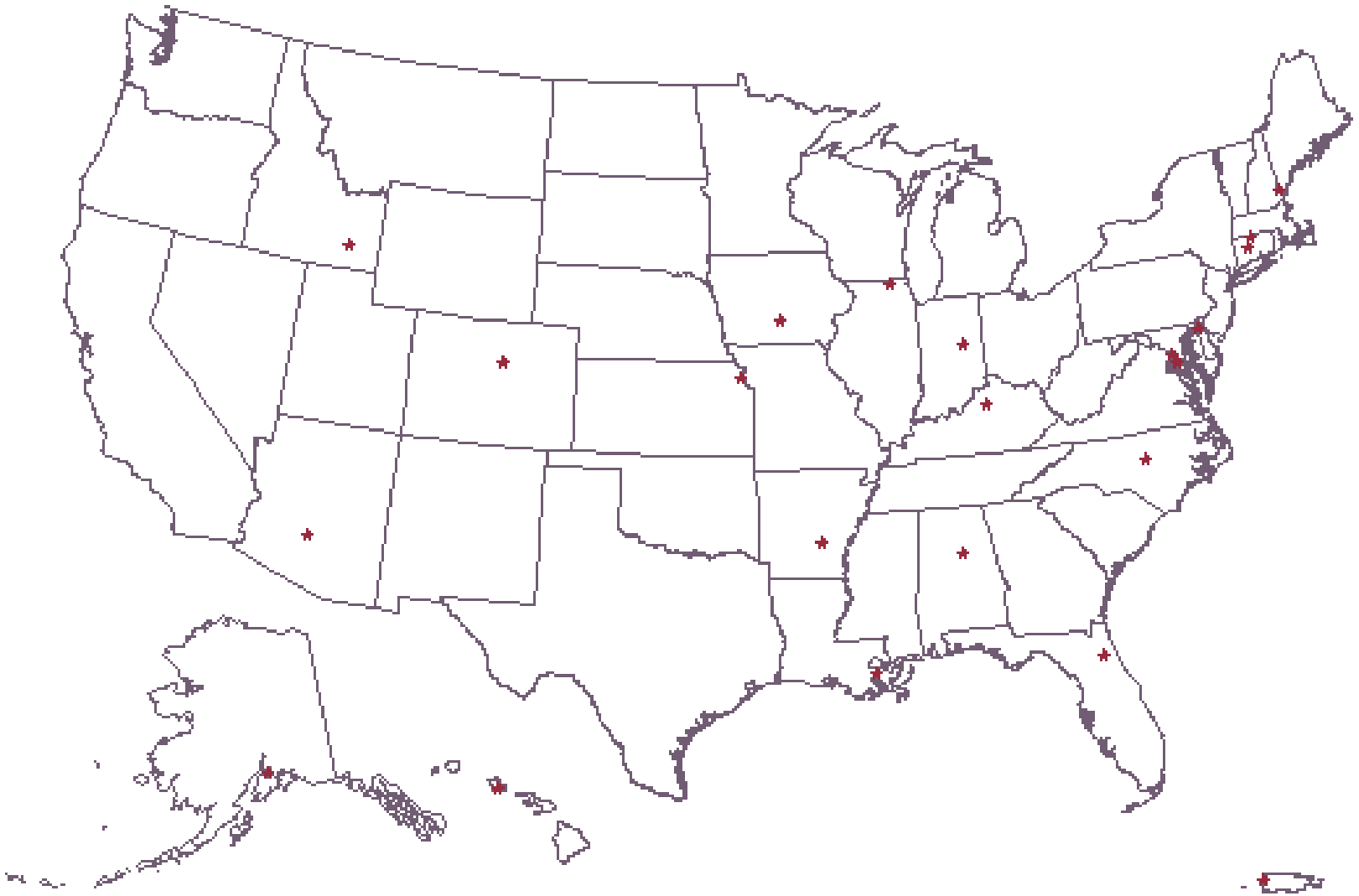
```
if upcase(text) contains 'SHARIA' and upcase(text) contains 'AMERICA' and
upcase(text) contains 'HATE'; run;

data work.graph_data; set work.graph_data work.tweet; run;

proc sort data=graph_data nodupkey; by time tweet_id; run;

ods listing close;
ods html path='C:\dcorliss\SAS\Twitter' file='test.html';

proc sgplot data=graph_data;
   vbar ten_sec_est; xaxis type=time interval=hour;
   format ten_sec_est hhmm.;
   title "Tweets (10 Second Totals)";
run;

ods html close;
ods listing;


%end;
%mend timeloop;
```

BIG DATA IGNITE

# Real-Time Updates - Charts



# Real-Time Updates - Maps



BIG DATA IGNITE

# Summary

**Social media data can be matched to event data for Time Series Analysis, including lead / lag modeling**

**Different social channels have different characteristics, including lead or lag times**

**Analytics methods optimized for big data methods may be necessary**

**Twitter data can be mined and analyzed at high volume**

**Analysis of Hate Crimes with social media data shows promise but better crime data is needed.**

BIG DATA IGNITE

# Questions

?

# Thank you!

David Corliss
davidjcorliss@gmail.com
davidjcorliss@peace-work.org
www.peace-work.org

BIG DATA IGNITE