NBS Data: Exploring 1

After the `load` and then `clean` of sample data, we now have a working (Pandas dataframe to explore. I'll poke around with the data here, and the once we know what type of analysis we are doing we can implement it in the `do` and `function` files.

The following fetches our data into a dataframe called `basicdf`:

In [1]:
```
from pandas import *

#Move to working directory
import os
os.chdir('C:\Users\Aman\Documents\Projects\NextBigSound\NBS-Data-Sample')

#Set data directory
datadir = 'data'
path = os.path.join(os.getcwd(), datadir)

#Retrieve dataframes from the hdf5 file
store = HDFStore(os.path.join(path, 'NBSData.h5'))
basicdf = store['basic']
store.close()
```

basicdf has the following data:

In [2]:
```
print basicdf.columns
print 'df index is: ', basicdf.index.names
```

```
array([Artist.ID, Day, Facebook.fans.d, Facebook.fans.t, Last.fm.plays.d,
       Last.fm.plays.t, MySpace.fans.d, MySpace.fans.t, MySpace.plays.d,
       MySpace.plays.t, Twitter.fans.d, Twitter.fans.t, Twitter.statuses.d,
       Twitter.statuses.t, YouTube.fans.d, YouTube.fans.t, YouTube.plays.d,
       YouTube.plays.t, Pandora.fans.d, Pandora.fans.t, Rdio.fans.d,
       Rdio.fans.t, Rdio.plays.d, Rdio.plays.t, SoundCloud.fans.d,
       SoundCloud.fans.t, SoundCloud.plays.d, SoundCloud.plays.t,
       iTunes.Album.Units.d, iTunes.Track.Units.d, Vevo.plays.d,
       Vevo.plays.t, SiteCatalyst.Visits.d, MediaGuide.Radio.Spins.d,
       Spotify.plays.d, Wikipedia.views.d], dtype=object)
df index is:  ['artist', 'date']
```

We can slice the data conveniently.

The following shows us all the data for an artist with id=1035.

In [3]:
```
basicdf.ix[1035,:]
```

Out[3]:
```
<class 'pandas.core.frame.DataFrame'>
Index: 536 entries, 2010-08-31 17:00:00 to 2012-02-17 16:00:00
Data columns:
Artist.ID                    536  non-null values
Day                          536  non-null values
Facebook.fans.d              514  non-null values
Facebook.fans.t              524  non-null values
Last.fm.plays.d              187  non-null values
Last.fm.plays.t              293  non-null values
MySpace.fans.d               519  non-null values
MySpace.fans.t               527  non-null values
MySpace.plays.d              512  non-null values
MySpace.plays.t              522  non-null values
Twitter.fans.d               522  non-null values
Twitter.fans.t               528  non-null values
Twitter.statuses.d          171  non-null values
Twitter.statuses.t          174  non-null values
YouTube.fans.d              460  non-null values
YouTube.fans.t              489  non-null values
YouTube.plays.d             403  non-null values
YouTube.plays.t             413  non-null values
Pandora.fans.d              176  non-null values
Pandora.fans.t              181  non-null values
Rdio.fans.d                 114  non-null values
Rdio.fans.t                 130  non-null values
Rdio.plays.d                209  non-null values
Rdio.plays.t                230  non-null values
SoundCloud.fans.d           212  non-null values
SoundCloud.fans.t           218  non-null values
SoundCloud.plays.d          75  non-null values
SoundCloud.plays.t          78  non-null values
iTunes.Album.Units.d        534  non-null values
iTunes.Track.Units.d        534  non-null values
Vevo.plays.d                0  non-null values
Vevo.plays.t                0  non-null values
SiteCatalyst.Visits.d       536  non-null values
MediaGuide.Radio.Spins.d    295  non-null values
Spotify.plays.d             0  non-null values
Wikipedia.views.d           532  non-null values
dtypes: float64(36)
```

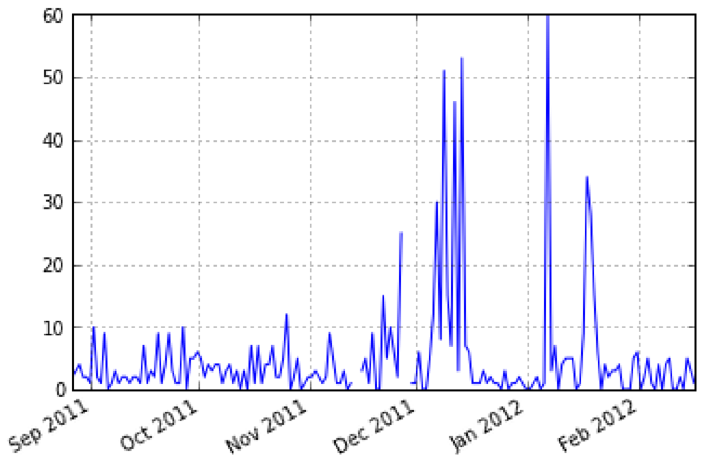
Apparently the sample dataset spans from (for this user, anyway) from the end of August last year to February 17th of this year 2012.

```
2010-08-31 17:00:00 to 2012-02-17 16:00:00
```

We've converted the data into a timeseries. This gives us a lot of flexibility.

In [4]:
```
i1035twitterstatus = basicdf.ix[1035,:]['Twitter.statuses.d']
#stored into a new var for easy access
i1035twitterstatus.plot()
i1035twitterstatus.describe()
```

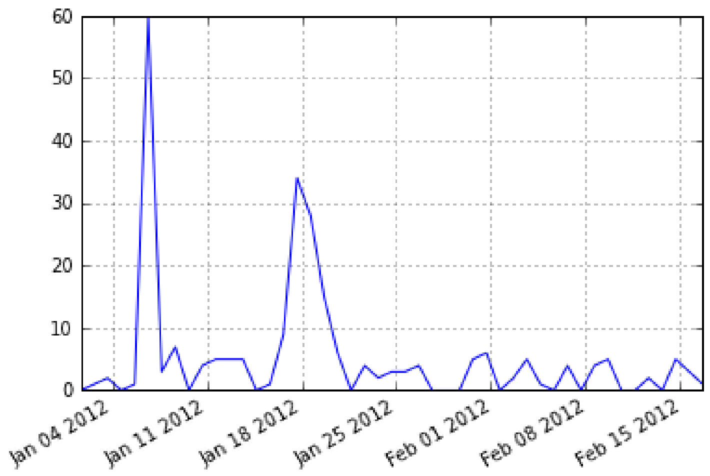
Out[4]:
```
count    171.000000
mean       4.929825
std        8.980095
min        0.000000
25%        1.000000
50%        2.000000
75%        5.000000
max       60.000000
```



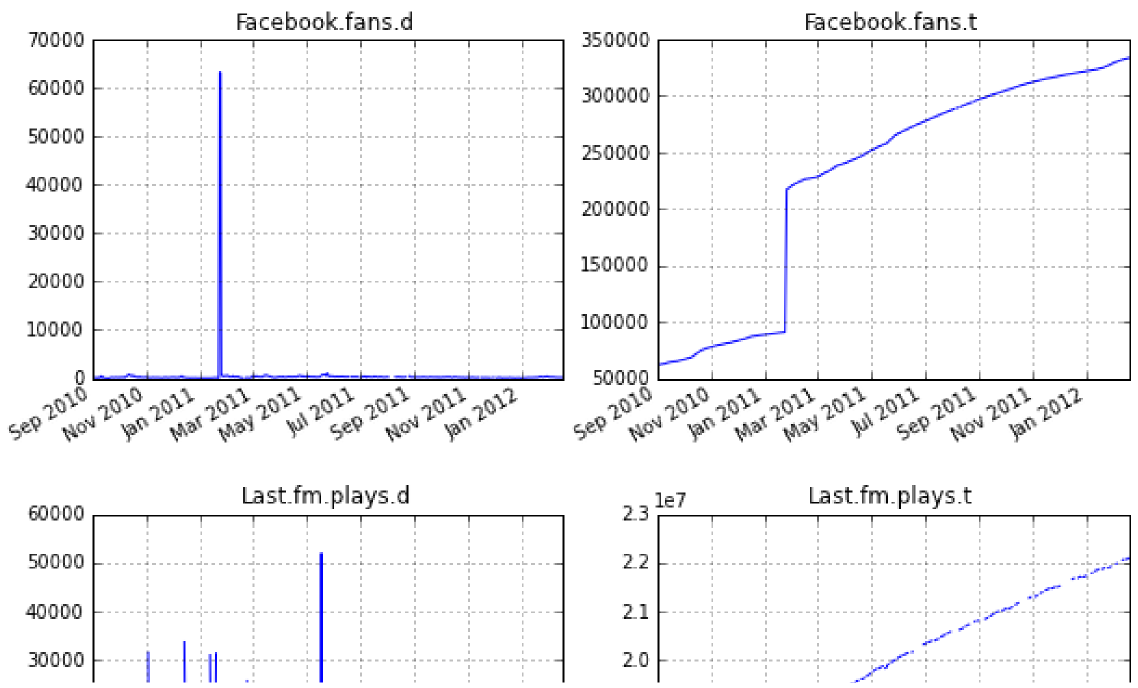
In [5]:
```
#zooming in on what happened just this year.
yrstart, yrend = datetime(2012, 01,01), datetime(2012, 2, 17)
```

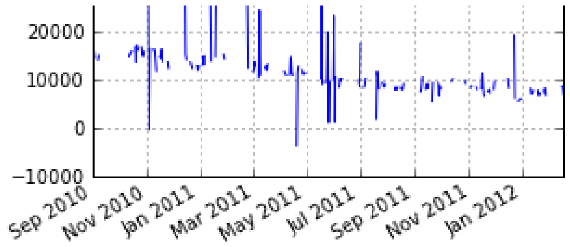
```
i1035twitterstatus[yrstart:yrend].plot()
```

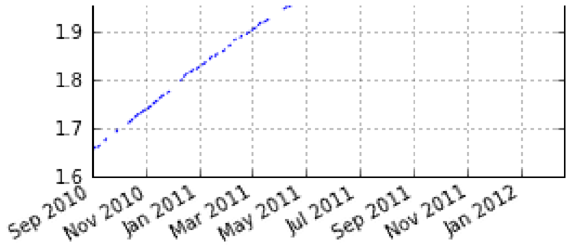
Out[5]: `<matplotlib.axes.AxesSubplot at 0x8c851f0>`

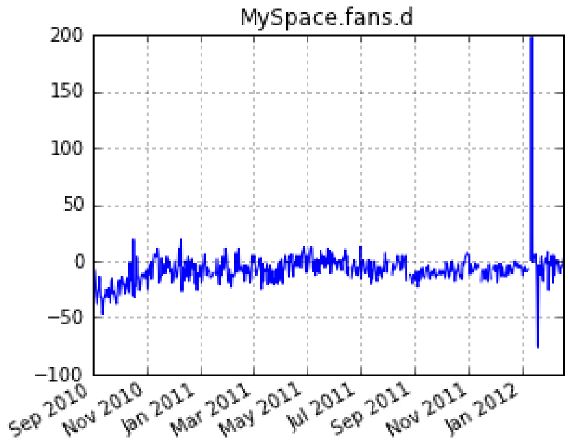


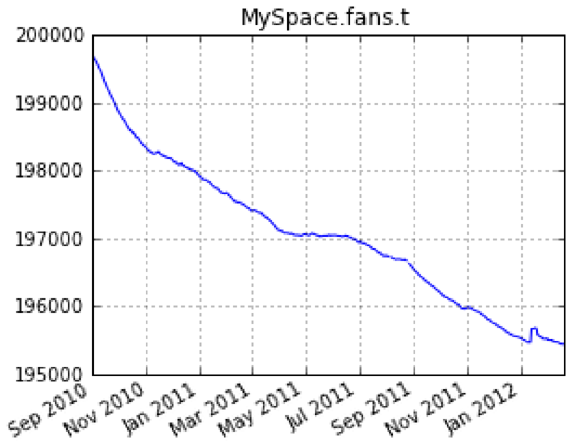
In [6]:
```
art1035 = basicdf.ix[1035,:]
del art1035['Artist.ID']
del art1035['Day']  #Don't need to plot this column

#Plotting parameters
ncols = len(art1035.columns)
plt.figure(figsize=(10,100))
subplots_adjust(hspace=0.4)

#Plot each column
for ii, col in enumerate(art1035):
    plt.subplot(20,2, ii+1)
    plt.title(str(col))
    if sum(art1035[col].notnull()) == 0:
        plt.text(0.5, 0.5, 'No Data points', fontsize=18, ha='center', va='top')
    else:
        art1035[col].plot()
```

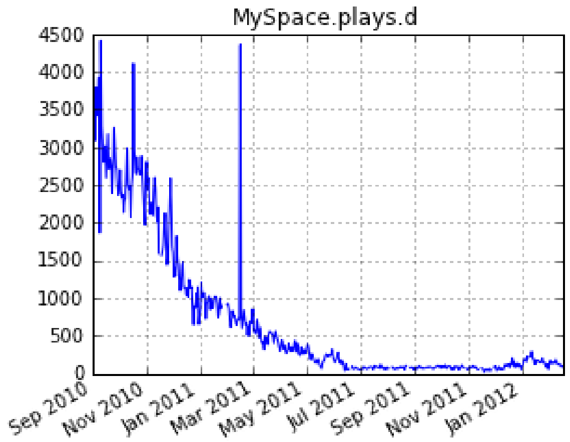


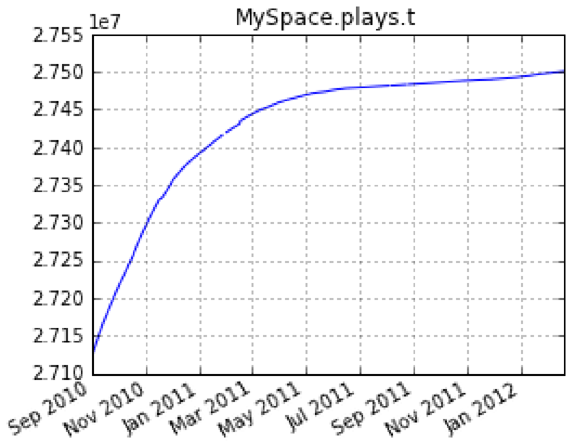

20000
10000
0
−10000
Sep 2010
Nov 2010
Jan 2011
Mar 2011
May 2011
Jul 2011
Sep 2011
Nov 2011
Jan 2012

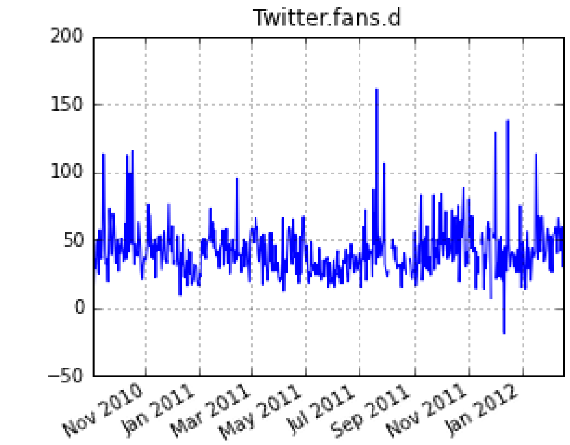

1.9
1.8
1.7
1.6
Sep 2010
Nov 2010
Jan 2011
Mar 2011
May 2011
Jul 2011
Sep 2011
Nov 2011
Jan 2012

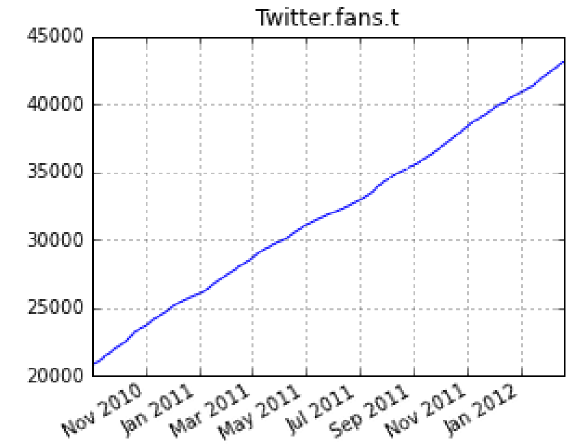

MySpace.fans.d
200
150
100
50
0
−50
−100
Sep 2010
Nov 2010
Jan 2011
Mar 2011
May 2011
Jul 2011
Sep 2011
Nov 2011
Jan 2012

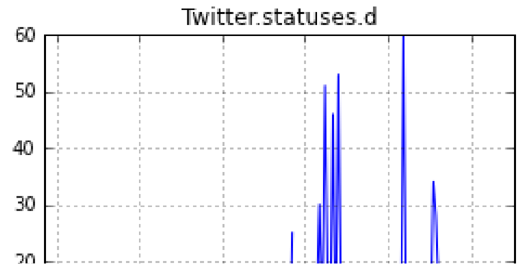

MySpace.fans.t
200000
199000
198000
197000
196000
195000
Sep 2010
Nov 2010
Jan 2011
Mar 2011
May 2011
Jul 2011
Sep 2011
Nov 2011
Jan 2012

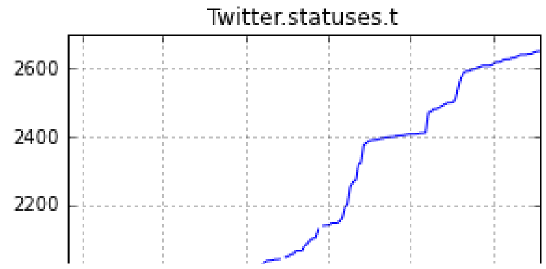

MySpace.plays.d
4500
4000
3500
3000
2500
2000
1500
1000
500
0
Sep 2010
Nov 2010
Jan 2011
Mar 2011
May 2011
Jul 2011
Sep 2011
Nov 2011
Jan 2012

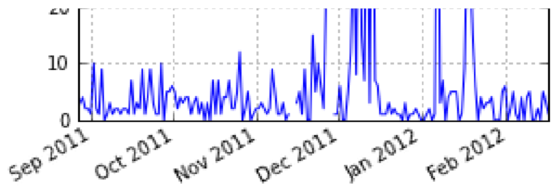

MySpace.plays.t
1e7
2.755
2.750
2.745
2.740
2.735
2.730
2.725
2.720
2.715
2.710
Sep 2010
Nov 2010
Jan 2011
Mar 2011
May 2011
Jul 2011
Sep 2011
Nov 2011
Jan 2012

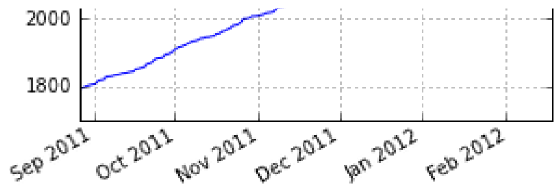

Twitter.fans.d
200
150
100
50
0
−50
Nov 2010
Jan 2011
Mar 2011
May 2011
Jul 2011
Sep 2011
Nov 2011
Jan 2012

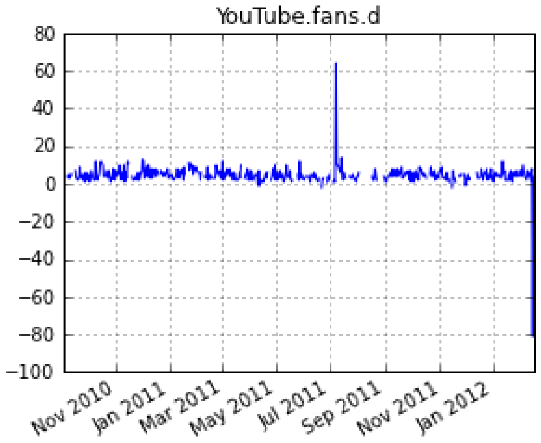

Twitter.fans.t
45000
40000
35000
30000
25000
20000
Nov 2010
Jan 2011
Mar 2011
May 2011
Jul 2011
Sep 2011
Nov 2011
Jan 2012

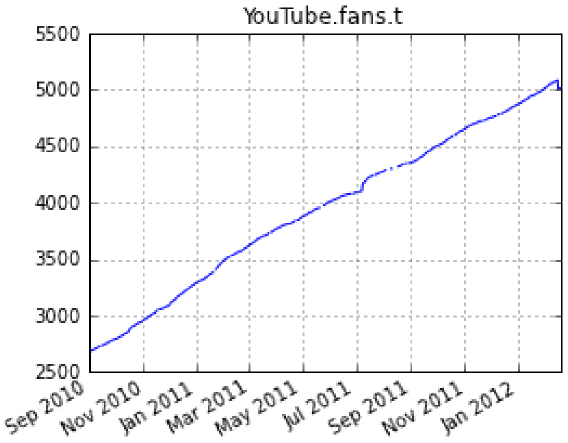

Twitter.statuses.d
60
50
40
30
20

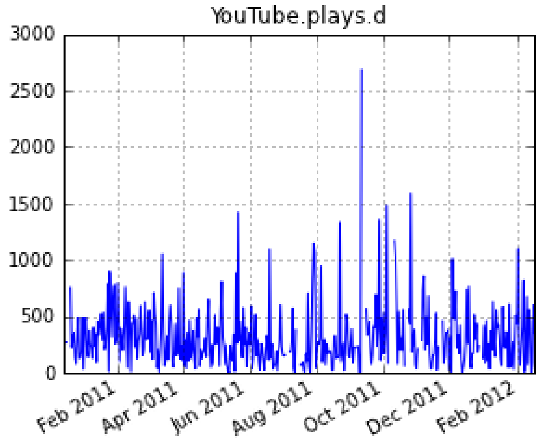

Twitter.statuses.t
2600
2400
2200

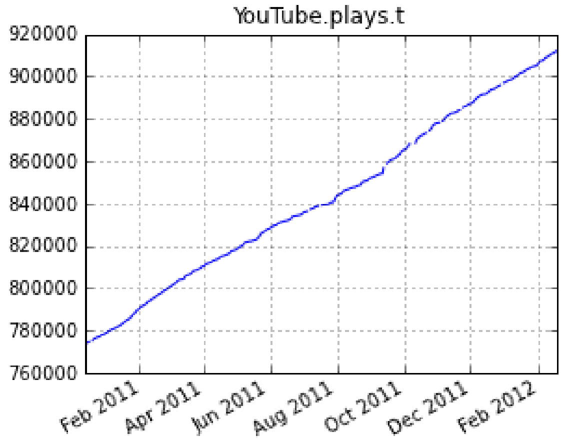

20
10
0
Sep 2011
Oct 2011
Nov 2011
Dec 2011
Jan 2012
Feb 2012

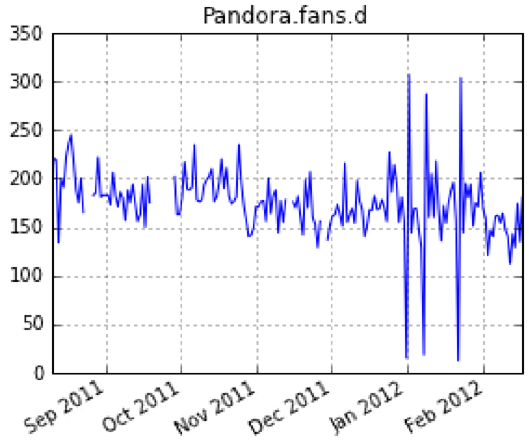

2000
1800
Sep 2011
Oct 2011
Nov 2011
Dec 2011
Jan 2012
Feb 2012

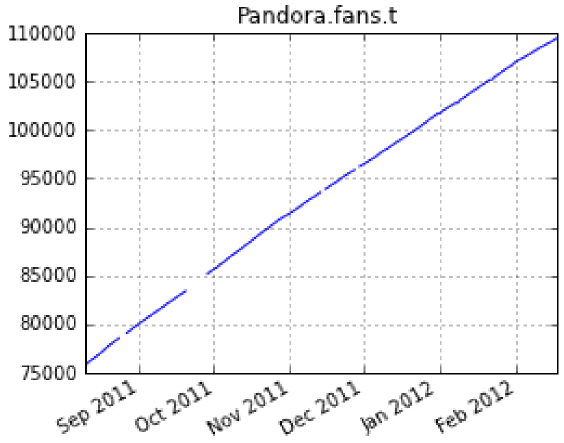

YouTube.fans.d
80
60
40
20
0
−20
−40
−60
−80
−100
Nov 2010
Jan 2011
Mar 2011
May 2011
Jul 2011
Sep 2011
Nov 2011
Jan 2012

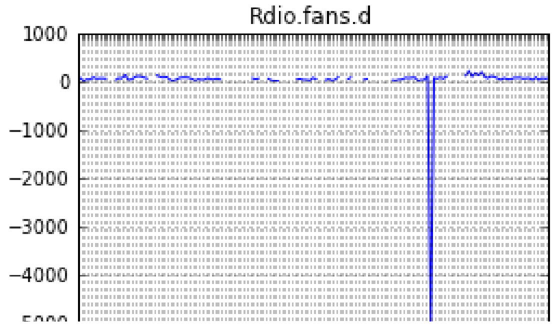

YouTube.fans.t
5500
5000
4500
4000
3500
3000
2500
Sep 2010
Nov 2010
Jan 2011
Mar 2011
May 2011
Jul 2011
Sep 2011
Nov 2011
Jan 2012

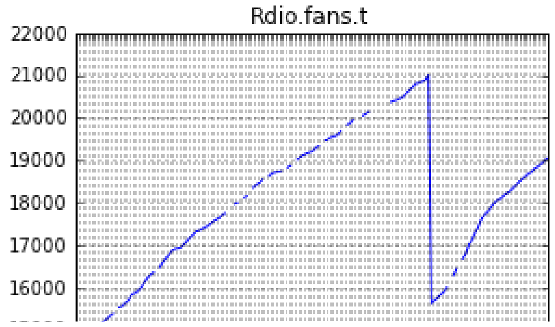

YouTube.plays.d
3000
2500
2000
1500
1000
500
0
Feb 2011
Apr 2011
Jun 2011
Aug 2011
Oct 2011
Dec 2011
Feb 2012

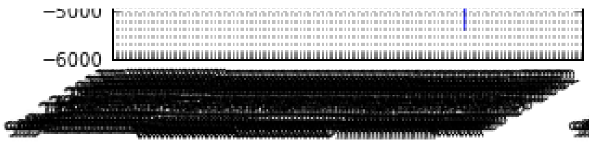

YouTube.plays.t
920000
900000
880000
860000
840000
820000
800000
780000
760000
Feb 2011
Apr 2011
Jun 2011
Aug 2011
Oct 2011
Dec 2011
Feb 2012

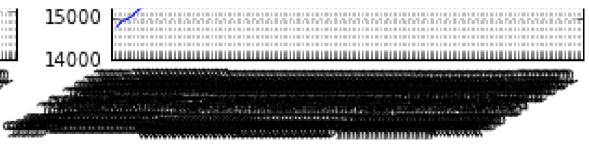

Pandora.fans.d
350
300
250
200
150
100
50
0
Sep 2011
Oct 2011
Nov 2011
Dec 2011
Jan 2012
Feb 2012

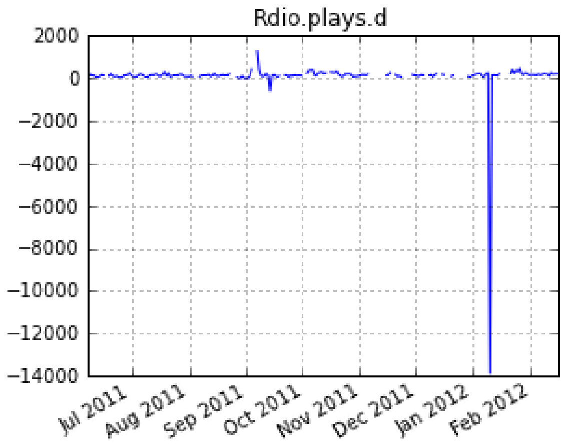

Pandora.fans.t
110000
105000
100000
95000
90000
85000
80000
75000
Sep 2011
Oct 2011
Nov 2011
Dec 2011
Jan 2012
Feb 2012

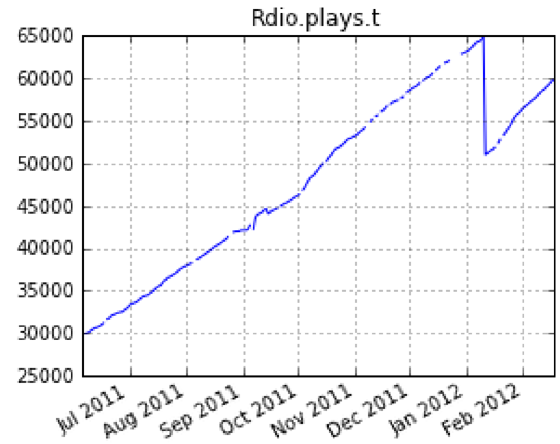

Rdio.fans.d
1000
0
−1000
−2000
−3000
−4000

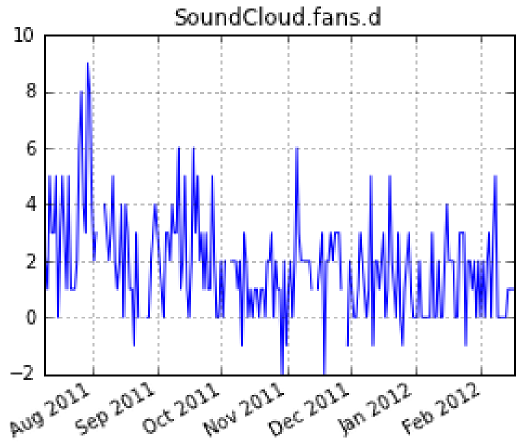

Rdio.fans.t
22000
21000
20000
19000
18000
17000
16000

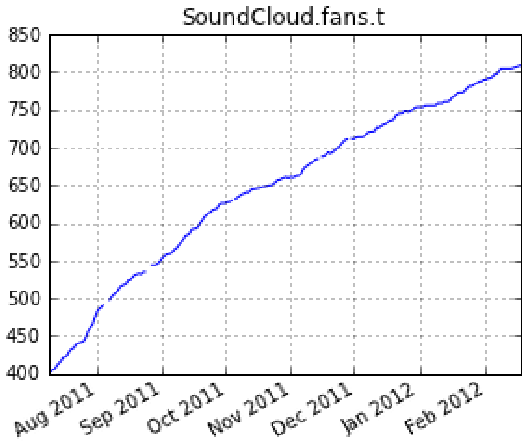

−5000
−6000

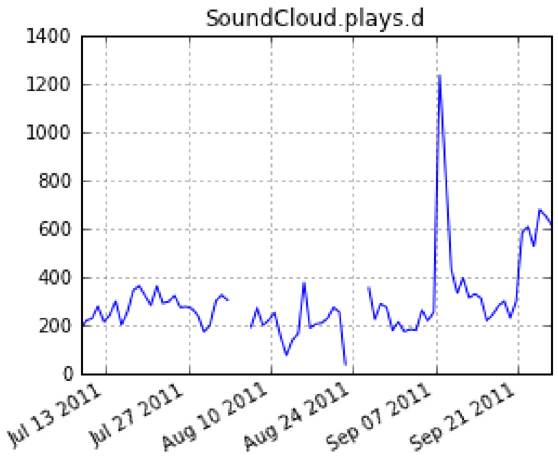

15000
14000

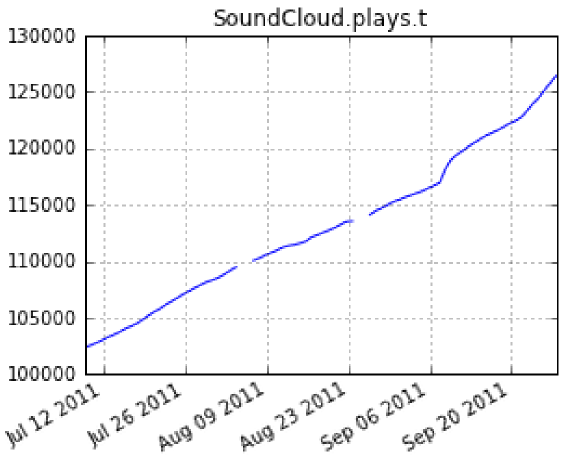

Rdio.plays.d
2000
0
−2000
−4000
−6000
−8000
−10000
−12000
−14000
Jul 2011
Aug 2011
Sep 2011
Oct 2011
Nov 2011
Dec 2011
Jan 2012
Feb 2012

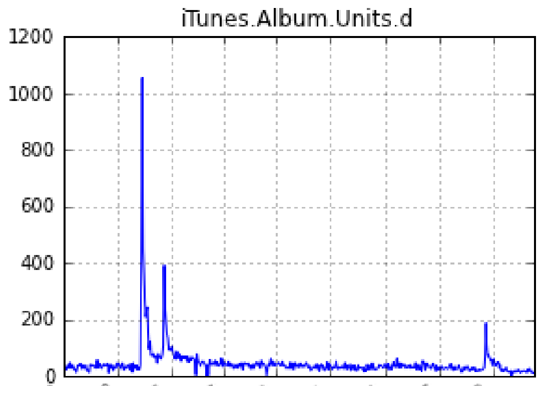

Rdio.plays.t
65000
60000
55000
50000
45000
40000
35000
30000
25000
Jul 2011
Aug 2011
Sep 2011
Oct 2011
Nov 2011
Dec 2011
Jan 2012
Feb 2012

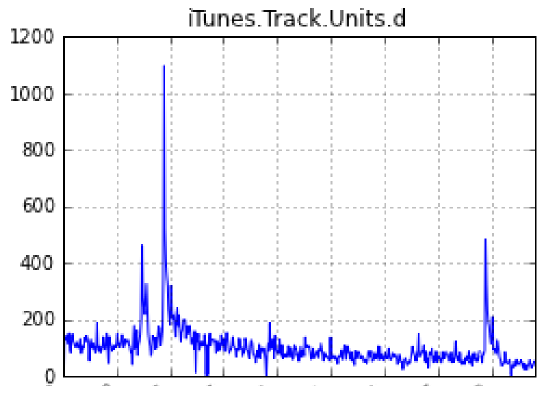

SoundCloud.fans.d
10
8
6
4
2
0
−2
Aug 2011
Sep 2011
Oct 2011
Nov 2011
Dec 2011
Jan 2012
Feb 2012

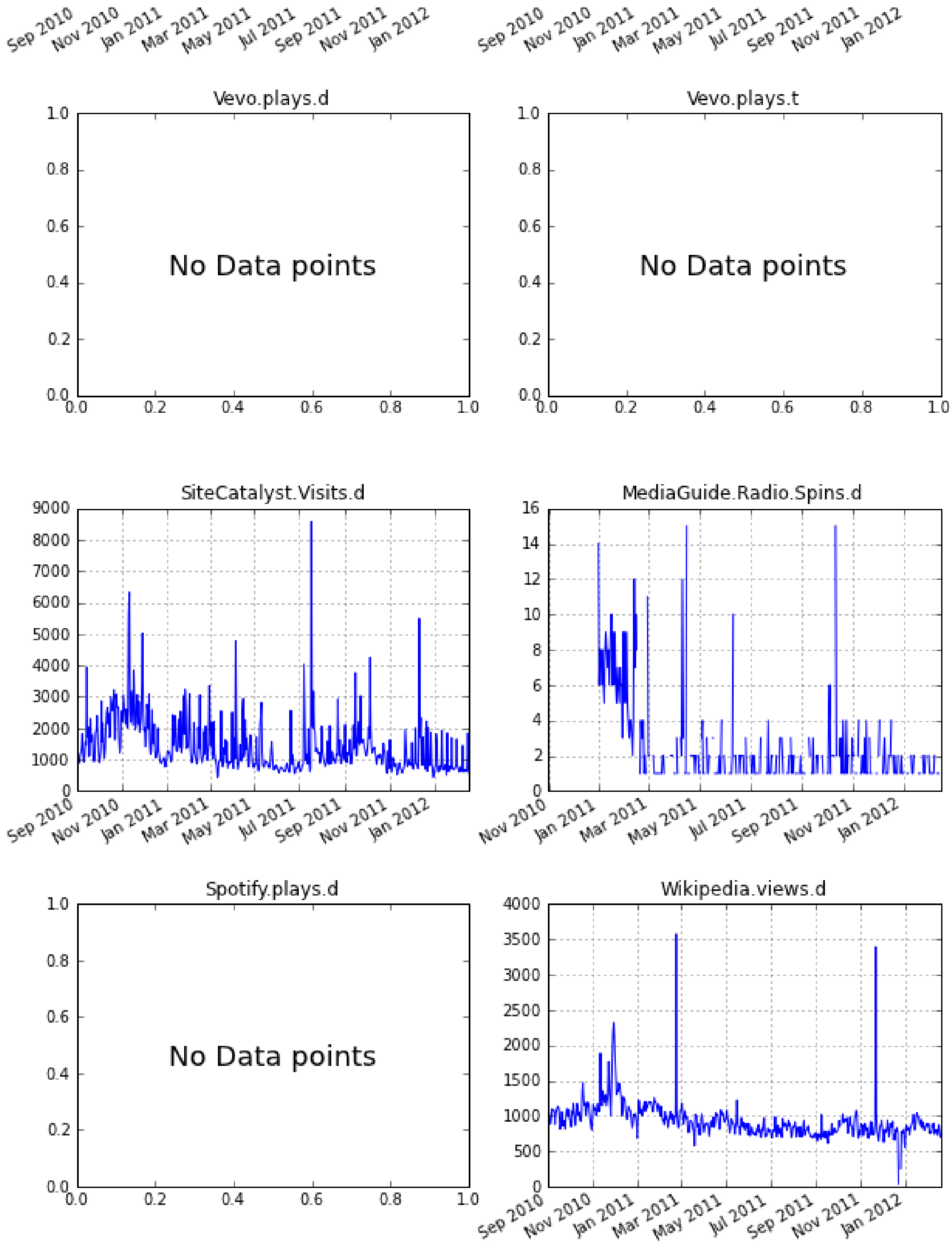

SoundCloud.fans.t
850
800
750
700
650
600
550
500
450
400
Aug 2011
Sep 2011
Oct 2011
Nov 2011
Dec 2011
Jan 2012
Feb 2012


SoundCloud.plays.d
1400
1200
1000
800
600
400
200
0
Jul 13 2011
Jul 27 2011
Aug 10 2011
Aug 24 2011
Sep 07 2011
Sep 21 2011


SoundCloud.plays.t
130000
125000
120000
115000
110000
105000
100000
Jul 12 2011
Jul 26 2011
Aug 09 2011
Aug 23 2011
Sep 06 2011
Sep 20 2011


iTunes.Album.Units.d
1200
1000
800
600
400
200
0


iTunes.Track.Units.d
1200
1000
800
600
400
200
0


Sep 2010
Nov 2010
Jan 2011
Mar 2011
May 2011
Jul 2011
Sep 2011
Nov 2011
Jan 2012
Vevo.plays.d
No Data points
Vevo.plays.t
No Data points
SiteCatalyst.Visits.d
MediaGuide.Radio.Spins.d
Spotify.plays.d
No Data points
Wikipedia.views.d

```
In [ ]:
```