

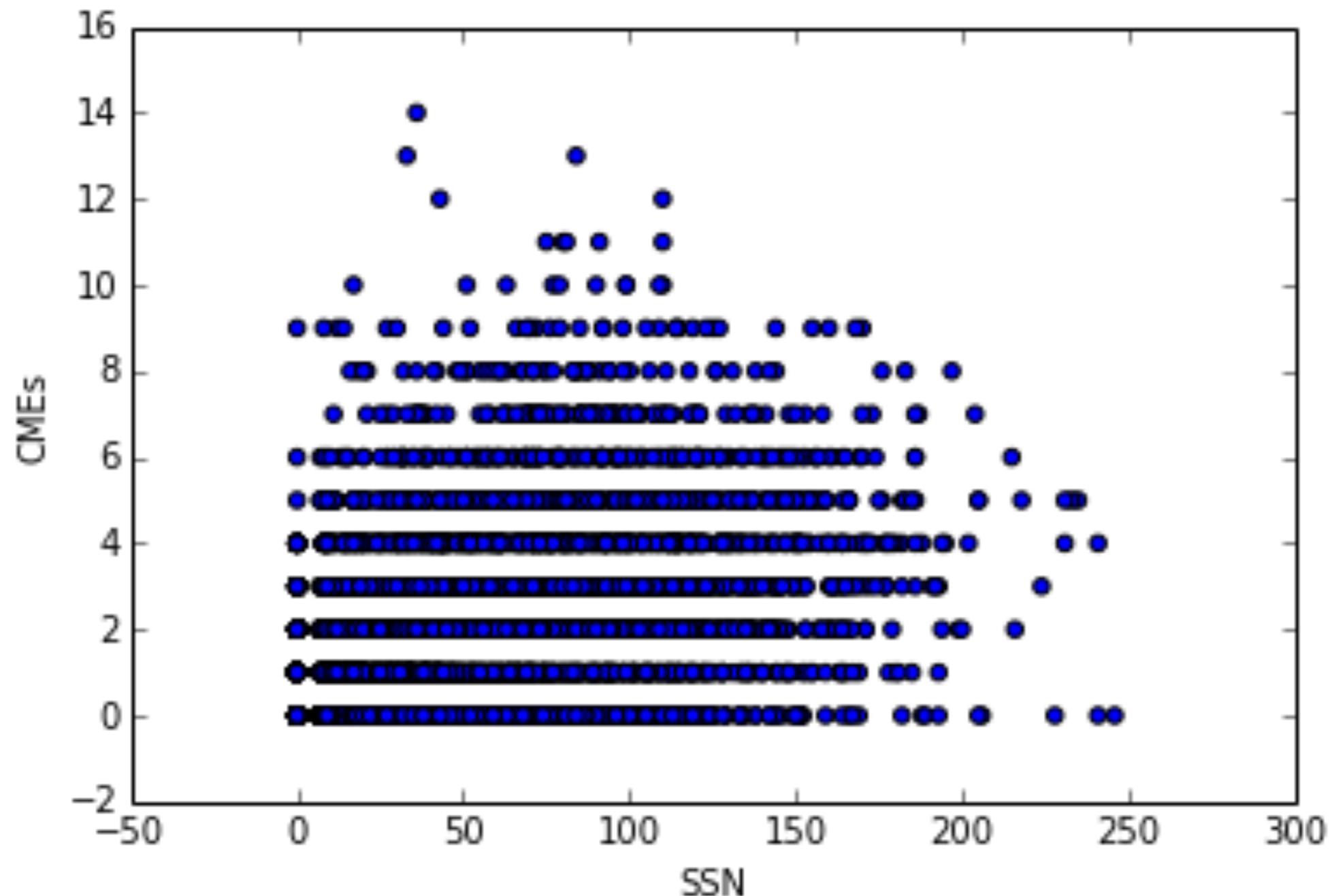
2000/07/14 08:12:10

Target: Predict a number of CMEs in a given time interval

- SOHO / LASCO CME Catalog 1997-2013:
http://cdaw.gsfc.nasa.gov/CME_list/
- Time of onset, speed, size, etc.
- 1-2 a day during solar minimum
- 5-10 a day during solar maximum

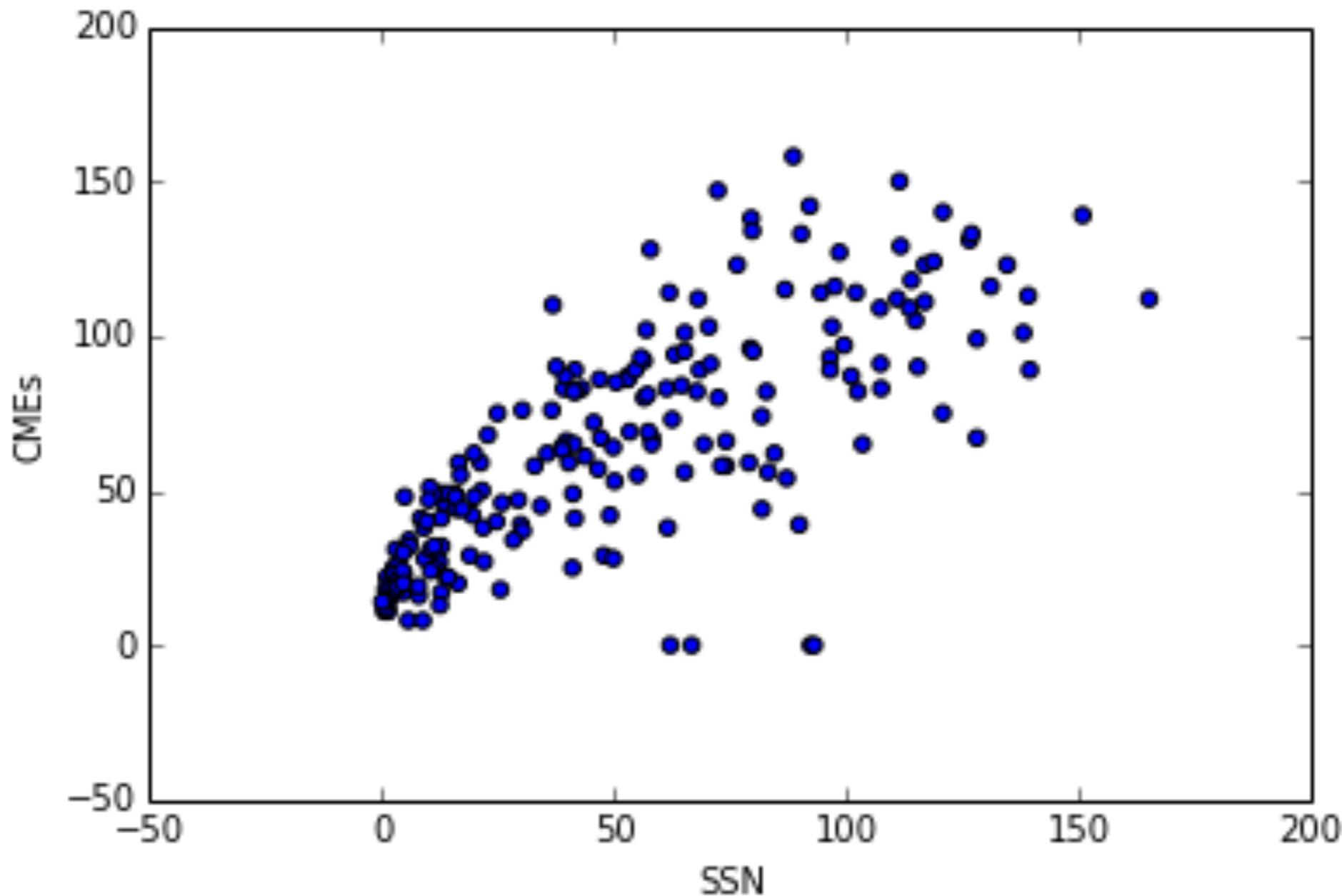
Part 1: Proxies for solar activity

- analysis and prediction



Part 1: Proxies for solar activity

- analysis and prediction



Part 1: Targets and Features

- Predict the number of CMEs tomorrow based on the information for today (categories)
- Predict the number of CMEs next month based on the information for this month - 1 month is approximately 1 solar cycle (regression)
- Solar irradiance, radio flux, total area of sunspots, previous CME information, sunspot groupings (txt files available for download)

Part 1: Daily Results

Random Forest

R^2 train: 0.340327415264

Classification Report

	precision	recall	f1-score	support
--	-----------	--------	----------	---------

0.0	0.49	0.75	0.59	1079
1.0	0.27	0.23	0.25	959
2.0	0.26	0.31	0.28	762
3.0	0.22	0.35	0.27	583
4.0	0.00	0.00	0.00	418
5.0	0.00	0.00	0.00	246
6.0	0.00	0.00	0.00	152
7.0	0.00	0.00	0.00	68
8.0	0.00	0.00	0.00	32
9.0	0.00	0.00	0.00	24
10.0	0.00	0.00	0.00	6
11.0	0.00	0.00	0.00	4
12.0	0.00	0.00	0.00	2
13.0	0.00	0.00	0.00	1
14.0	0.00	0.00	0.00	1

avg / total	0.26	0.34	0.29	4337
-------------	------	------	------	------

R^2 test: 0.32688172043

Classification Report

	precision	recall	f1-score	support
--	-----------	--------	----------	---------

0.0	0.0	0.47	0.71	0.57	446
1.0	1.0	0.28	0.27	0.27	401
2.0	2.0	0.25	0.29	0.26	347
3.0	3.0	0.21	0.34	0.26	247
4.0	4.0	0.00	0.00	0.00	194
5.0	5.0	0.00	0.00	0.00	101
6.0	6.0	0.00	0.00	0.00	61
7.0	7.0	0.00	0.00	0.00	31
8.0	8.0	0.00	0.00	0.00	18
9.0	9.0	0.00	0.00	0.00	8
10.0	10.0	0.00	0.00	0.00	4
11.0	11.0	0.00	0.00	0.00	1
12.0	12.0	0.00	0.00	0.00	1
13.0	13.0	0.00	0.00	0.00	1

avg / total	0.25	0.33	0.28	1860
-------------	------	------	------	------

Part 1: Daily Results

Random Forest

Feature Importance

(0.10217336393637823, 'L_speed')

(0.093088329644439222, '2nd_o_speed')

(0.092635194989601835, '2nd_o_speed_20R')

(0.075360227964301738, 'F10.7')

(0.075162723733324707, 'hour')

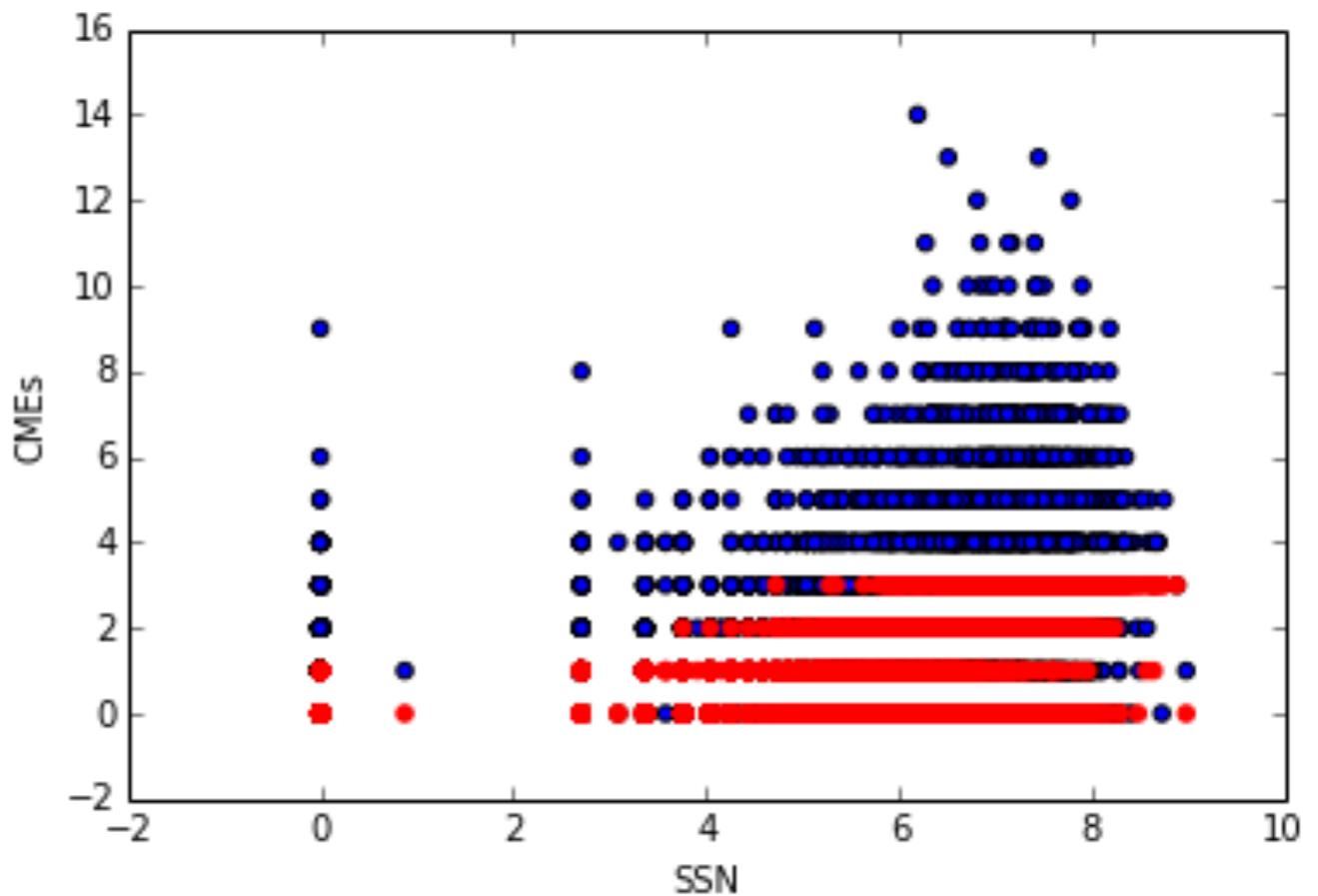
(0.060248208204694925, 'LymanAlpha')

(0.056812832819784748, 'Central_PA')

(0.052773346909765755, 'Width')

(0.042827630101142759, 'CME')

(0.04272818297068276, 'ssn')



Part 1: Monthly Results

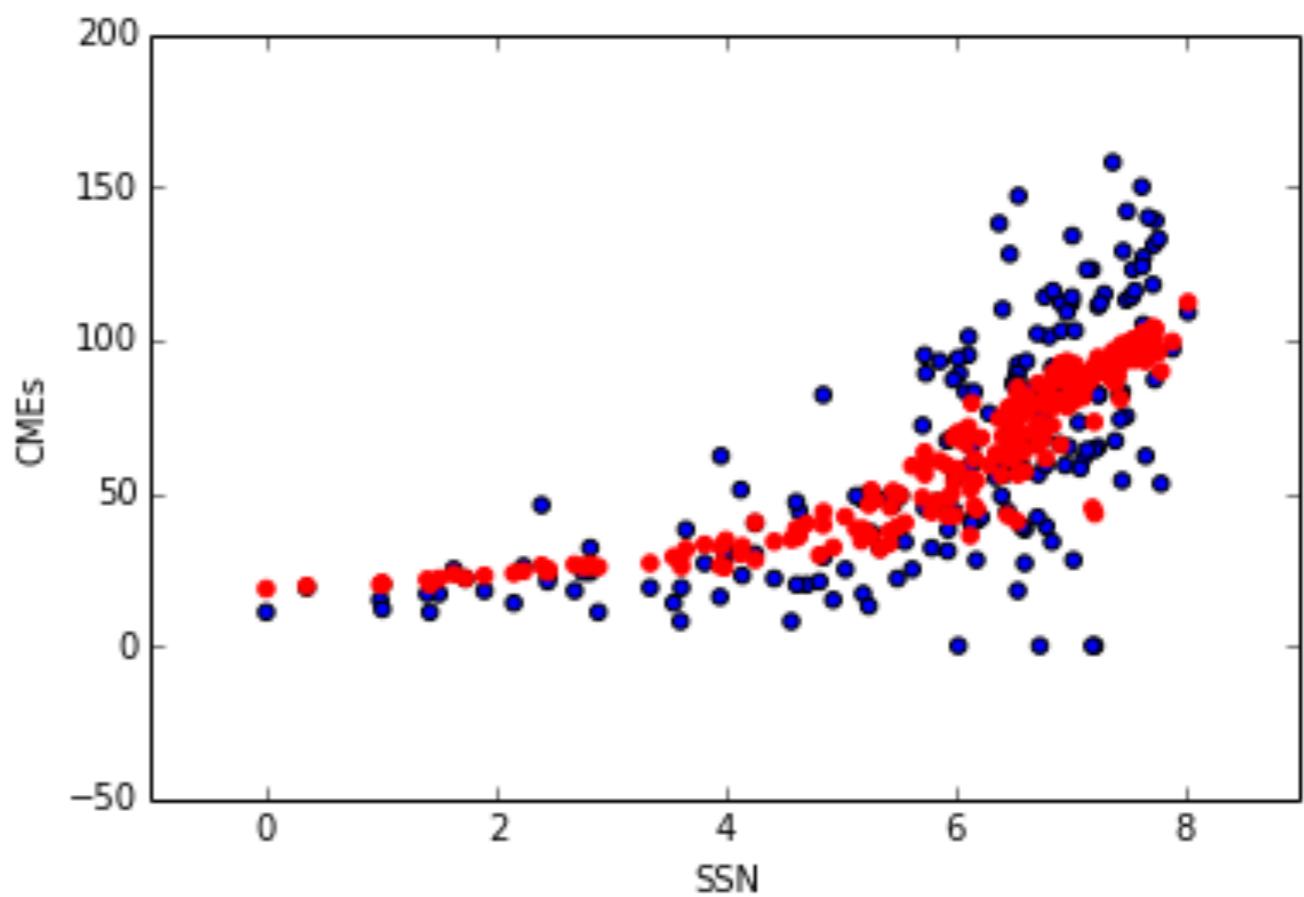
Ridge Regression

R² train: 0.653400997278

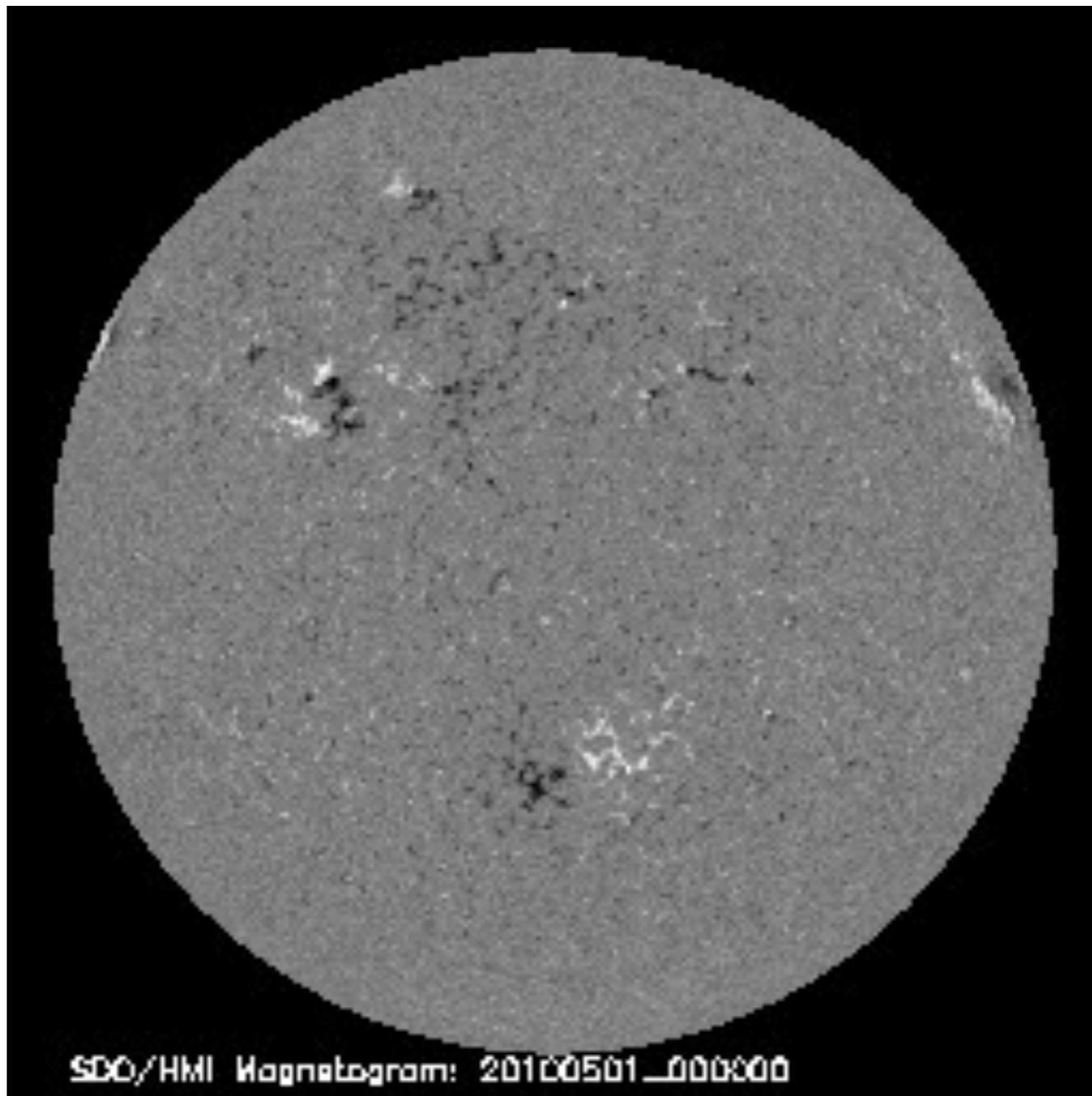
R² test: 0.507888792648

MAPE train: 129.544504755

MAPE test: 173.152010099



Part 2: Vector Magnetogram Analysis



Part 2: Getting data

```
wget -r -nd -N -np  
--tries=75 --no-  
clobber --accept  
'*_M_256.jpg' http://  
jsoc.stanford.edu/  
data/hmi/images
```

```
wget ... & wget ... &  
wget ...& ...
```

Index of /data/hmi/images/2010/05/01

Name	Last modified	Size	Description
Parent Directory		-	
20100501_000000_Ic_1k.jpg	22-Jun-2011 14:01	295K	
20100501_000000_Ic_4k.jpg	22-Jun-2011 14:01	4.5M	
20100501_000000_Ic_256.jpg	22-Jun-2011 14:01	19K	
20100501_000000_Ic_512.jpg	22-Jun-2011 14:01	65K	
20100501_000000_Ic_flat_1k.jpg	22-Jun-2011 14:01	329K	
20100501_000000_Ic_flat_4k.jpg	22-Jun-2011 14:01	5.0M	
20100501_000000_Ic_flat_256.jpg	22-Jun-2011 14:01	18K	
20100501_000000_Ic_flat_512.jpg	22-Jun-2011 14:01	69K	
20100501_000000_M_1k.jpg	22-Jun-2011 14:01	403K	
20100501_000000_M_4k.jpg	22-Jun-2011 14:01	7.0M	
20100501_000000_M_256.jpg	22-Jun-2011 14:01	28K	
20100501_000000_M_512.jpg	22-Jun-2011 14:01	101K	
20100501_001500_Ic_1k.jpg	22-Jun-2011 14:01	296K	
20100501_001500_Ic_4k.jpg	22-Jun-2011 14:01	4.5M	
20100501_001500_Ic_256.jpg	22-Jun-2011 14:01	19K	
20100501_001500_Ic_512.jpg	22-Jun-2011 14:01	65K	
20100501_001500_Ic_flat_1k.jpg	22-Jun-2011 14:01	329K	
20100501_001500_Ic_flat_4k.jpg	22-Jun-2011 14:01	5.0M	
20100501_001500_Ic_flat_256.jpg	22-Jun-2011 14:01	19K	
20100501_001500_Ic_flat_512.jpg	22-Jun-2011 14:01	69K	
20100501_001500_M_1k.jpg	22-Jun-2011 14:01	403K	
20100501_001500_M_4k.jpg	22-Jun-2011 14:01	7.0M	
20100501_001500_M_256.jpg	22-Jun-2011 14:01	28K	
20100501_001500_M_512.jpg	22-Jun-2011 14:01	100K	
20100501_003000_Ic_1k.jpg	22-Jun-2011 14:01	296K	
20100501_003000_Ic_4k.jpg	22-Jun-2011 14:01	4.5M	

Part 2: Getting data faster

```
import asyncio
import aiohttp
import urllib.request

def generate_url_list(base_url, year_start):
    .....

    @asyncio.coroutine
    def process_all():
        url_list, file_list = generate_url_list('http://jsoc.stanford.edu/data/hmi/images/',
                                                2010)
        url_chunks=[url_list[x:x+pool_size] for x in range(0, len(url_list), pool_size)]
        for urls, in url_chunks:
            yield from process_batch_of_urls(urls, file_names)
```

Part 2: Data preparation

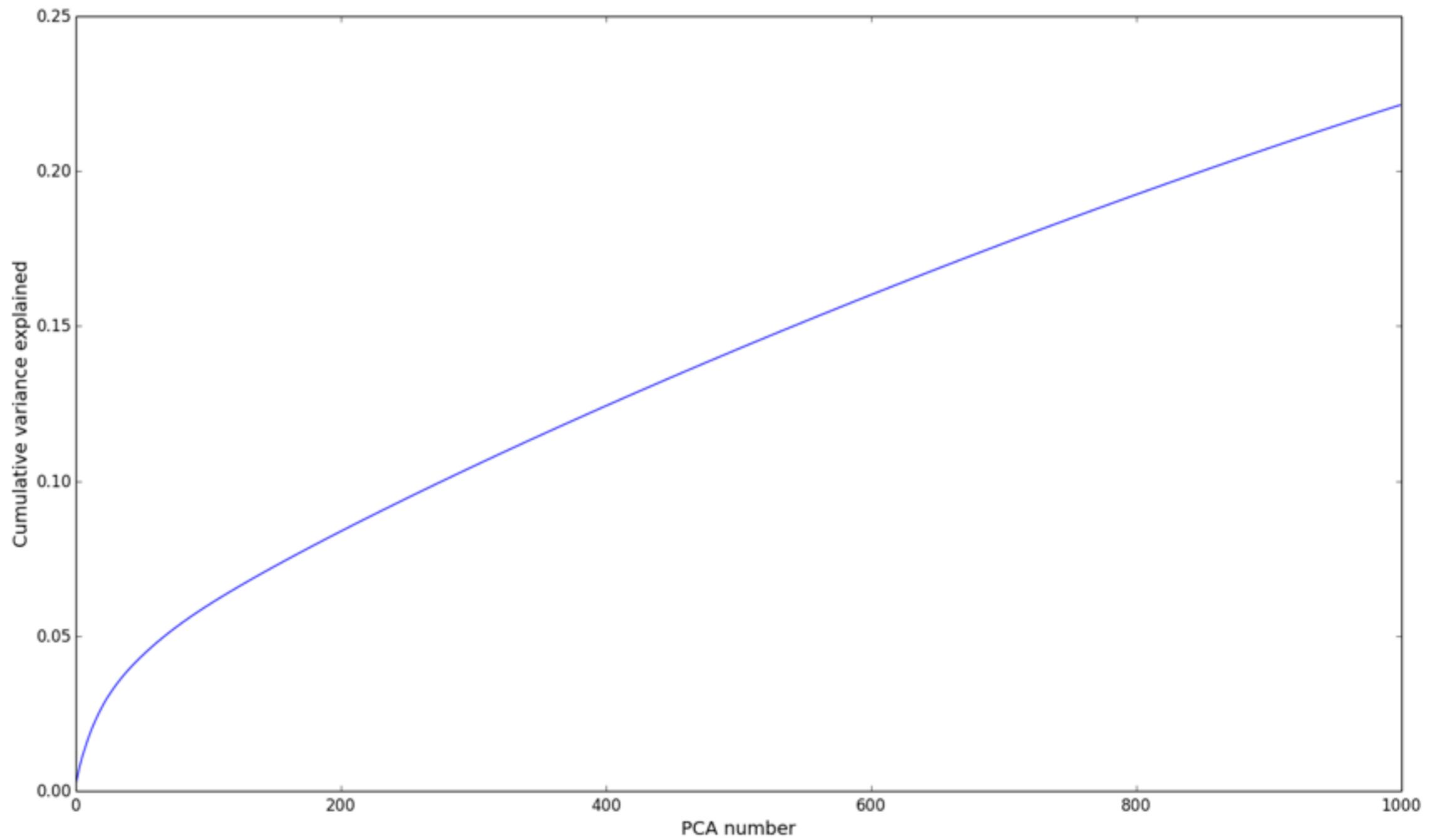
- Images are 256x256x256 arrays
- Clip to solar disc and flatten to become 42981 features
- Data from 2010 to 2013 every 15 mins = 126144 rows
- Find image just before CME event classified as ‘1’ otherwise ‘0’

```
LogisticRegression(penalty='l1').fit(x_train ,y_train)
```

Part 2: Results



Part 2: Dimensionality Reduction - PCA?



Part 2: Area under curve

- $.90-1 = \text{excellent (A)}$
- $.80-.90 = \text{good (B)}$
- $.70-.80 = \text{fair (C)}$
- $.60-.70 = \text{poor (D)}$
- $.50-.60 = \text{fail (F)}$
- Using MultinomialNB

Part 2: PCA Result

R² train: 0.976806342016

R² test: 0.976508825706

	precision	recall	f1-score	support
0	0.98	1.00	0.99	86350
1	0.02	0.00	0.00	1950

avg / total 0.96 0.98 0.97 88300

	precision	recall	f1-score	support
0	0.98	1.00	0.99	37000
1	0.08	0.00	0.01	844

avg / total 0.96 0.98 0.97 37844

Area Under Curve train: 0.499933781717

Area Under Curve test: 0.501707506084

Area Under Curve average: 0.500820643901



Part 2: PCA Result + time since last event

R² train: 0.659943374858

R² test: 0.663169855195

	precision	recall	f1-score	support
0	0.98	0.66	0.79	86353
1	0.03	0.48	0.06	1947

avg / total 0.96 0.66 0.78 88300

	precision	recall	f1-score	support
0	0.98	0.67	0.79	36997
1	0.03	0.46	0.06	847

avg / total 0.96 0.66 0.78 37844

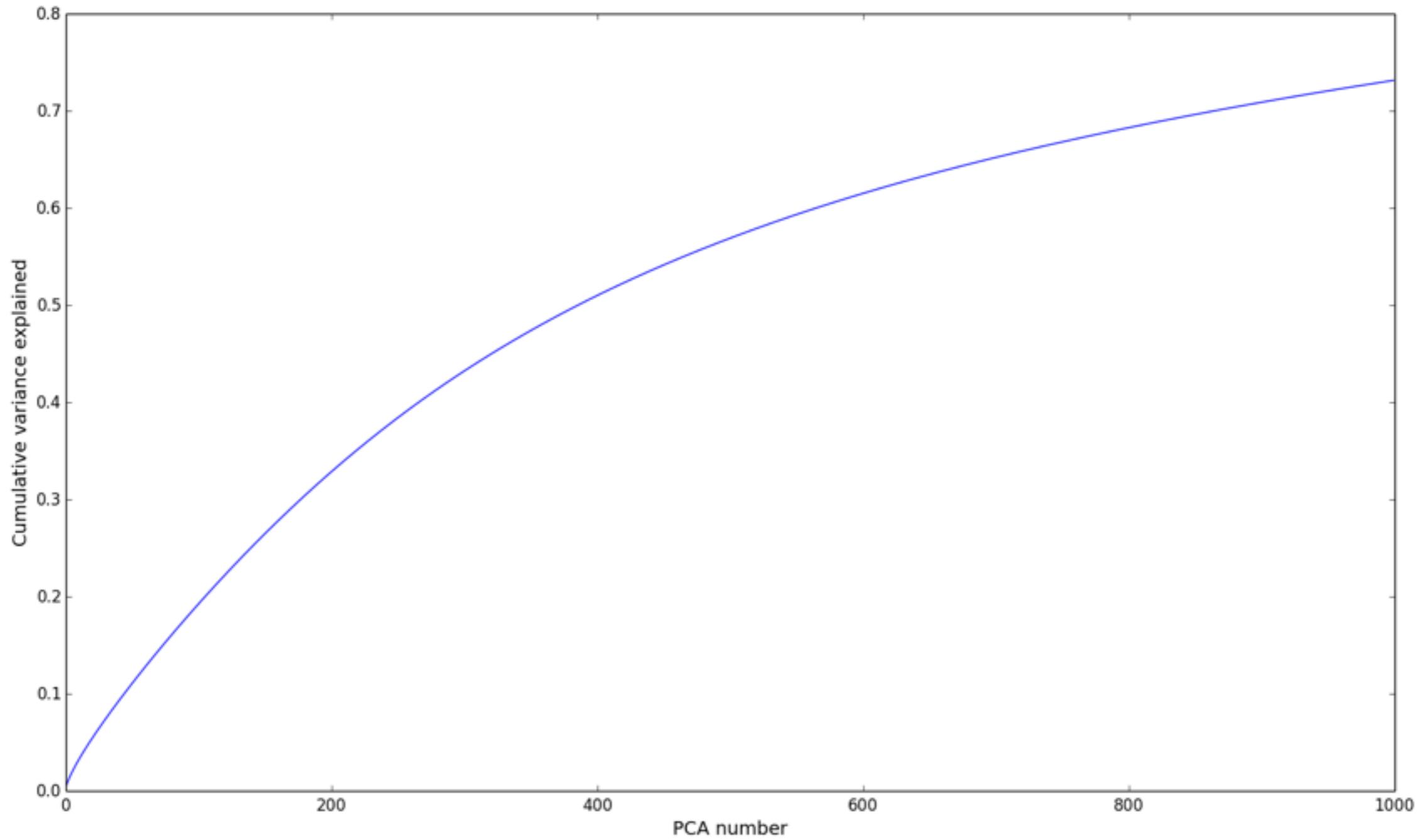
Area Under Curve train: 0.569851588204

Area Under Curve test: 0.561822556275

Area Under Curve average: 0.56583707224

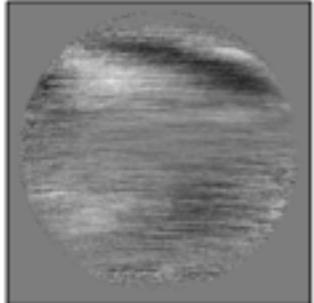


Part 2: PCA for one solar rotation

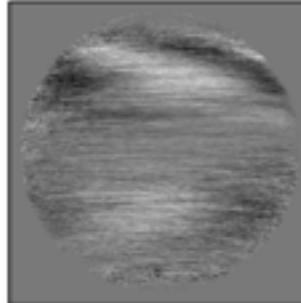


Part 2: PCA May 2010

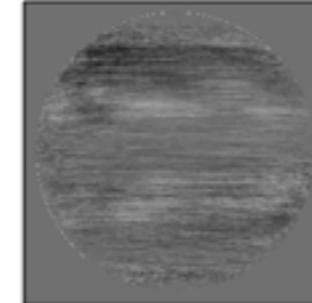
Eigensun 0 (evr=0.0039)



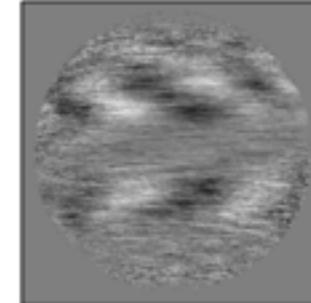
Eigensun 1 (evr=0.0038)



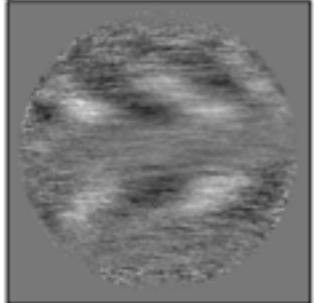
Eigensun 2 (evr=0.0032)



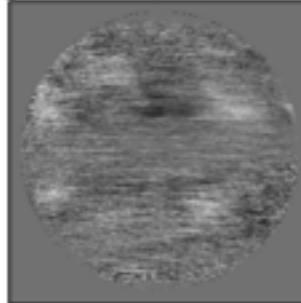
Eigensun 3 (evr=0.0031)



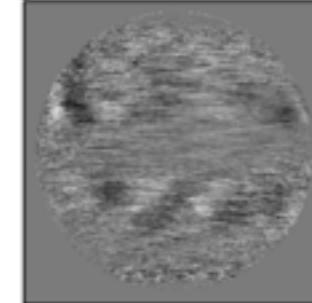
Eigensun 4 (evr=0.0030)



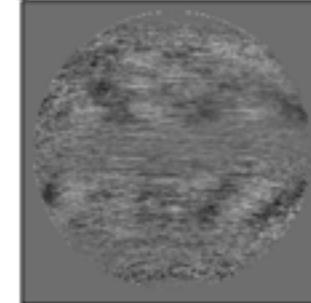
Eigensun 5 (evr=0.0024)



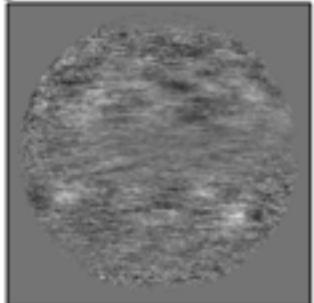
Eigensun 6 (evr=0.0023)



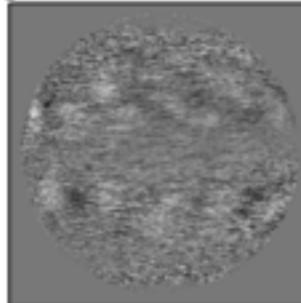
Eigensun 7 (evr=0.0022)



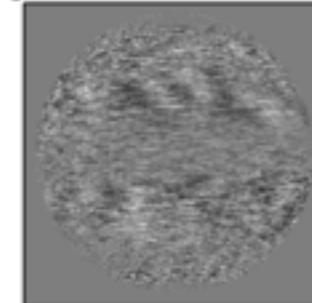
Eigensun 8 (evr=0.0021)



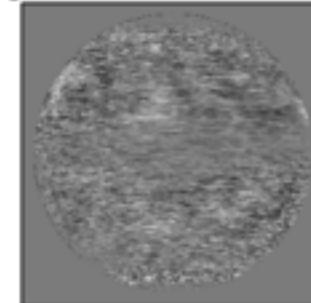
Eigensun 9 (evr=0.0020)



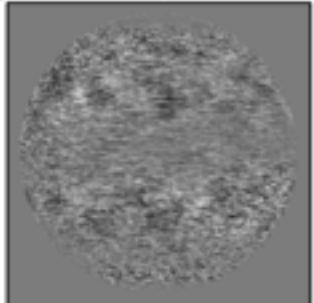
Eigensun 10 (evr=0.0020)



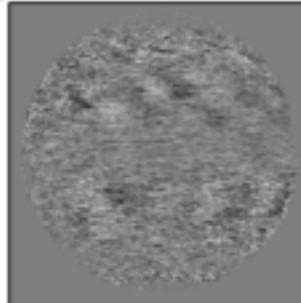
Eigensun 11 (evr=0.0019)



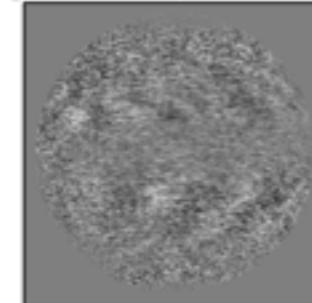
Eigensun 12 (evr=0.0019)



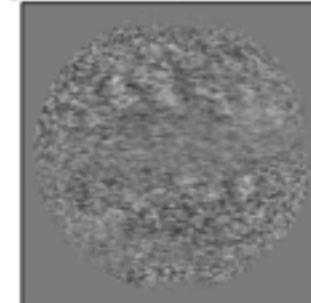
Eigensun 13 (evr=0.0019)



Eigensun 14 (evr=0.0018)

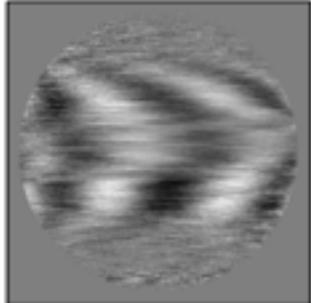


Eigensun 15 (evr=0.0018)

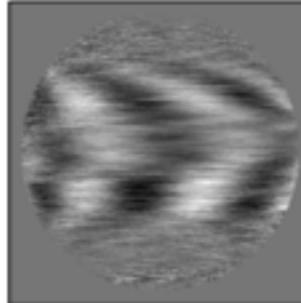


Part 2: PCA December 2013

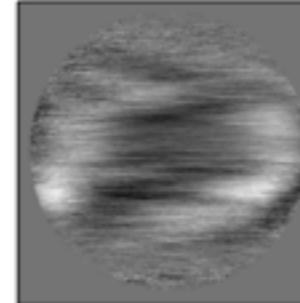
Eigensun 0 (evr=0.0081)



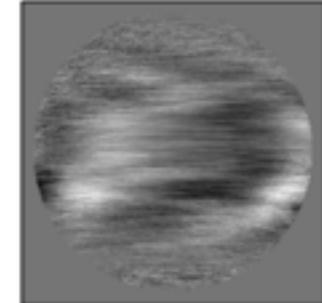
Eigensun 1 (evr=0.0081)



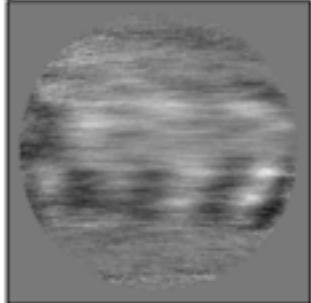
Eigensun 2 (evr=0.0072)



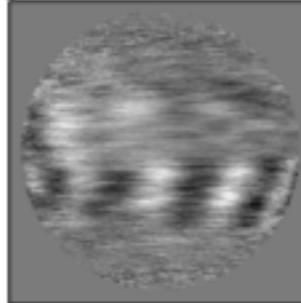
Eigensun 3 (evr=0.0071)



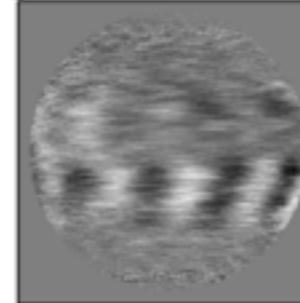
Eigensun 4 (evr=0.0053)



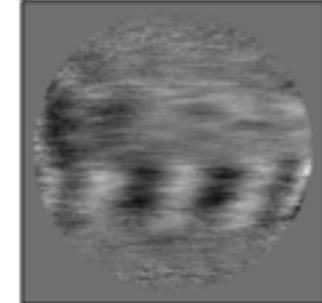
Eigensun 5 (evr=0.0052)



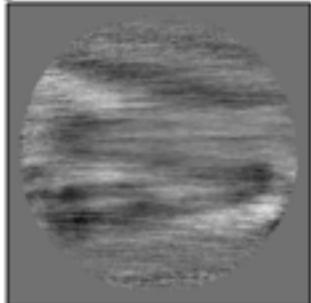
Eigensun 6 (evr=0.0051)



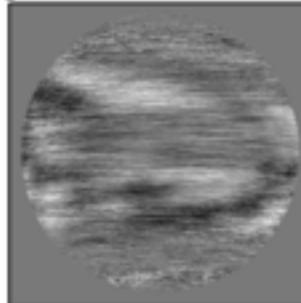
Eigensun 7 (evr=0.0049)



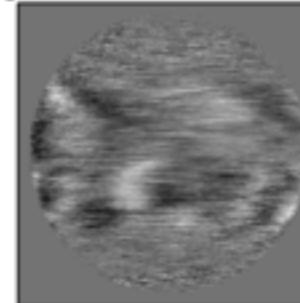
Eigensun 8 (evr=0.0042)



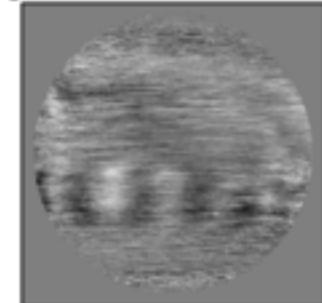
Eigensun 9 (evr=0.0040)



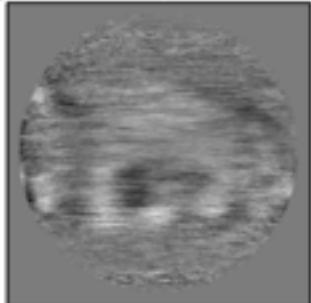
Eigensun 10 (evr=0.0036)



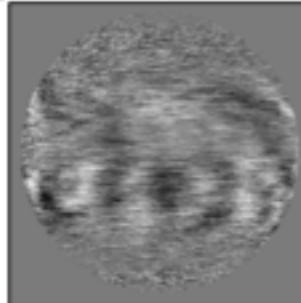
Eigensun 11 (evr=0.0034)



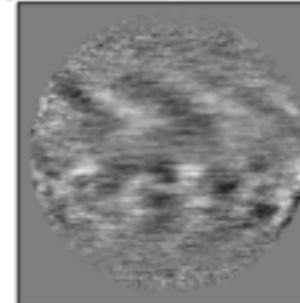
Eigensun 12 (evr=0.0034)



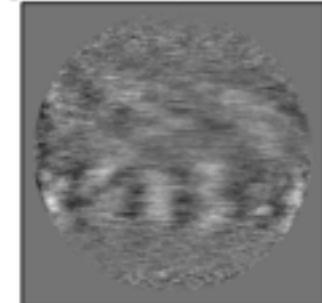
Eigensun 13 (evr=0.0031)



Eigensun 14 (evr=0.0030)

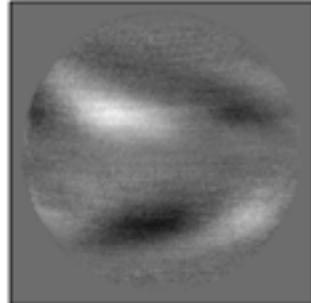


Eigensun 15 (evr=0.0029)

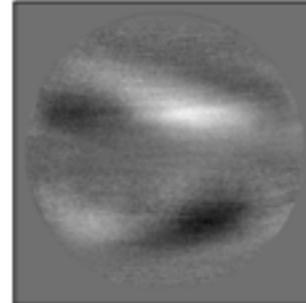


Part 2: Quiet Eigensuns

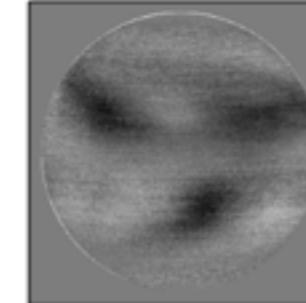
Eigensun 0 (evr=0.0027)



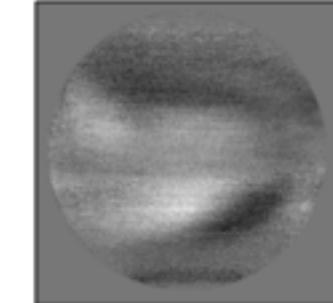
Eigensun 1 (evr=0.0027)



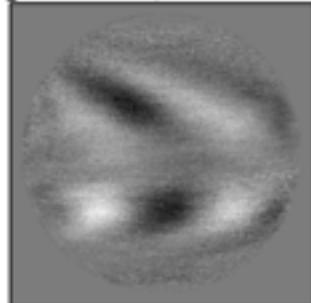
Eigensun 2 (evr=0.0024)



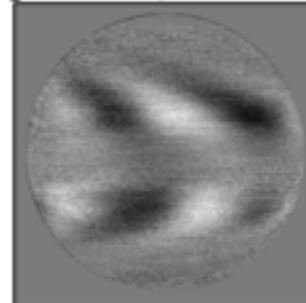
Eigensun 3 (evr=0.0022)



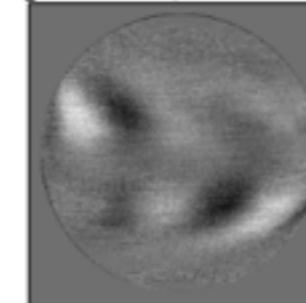
Eigensun 4 (evr=0.0022)



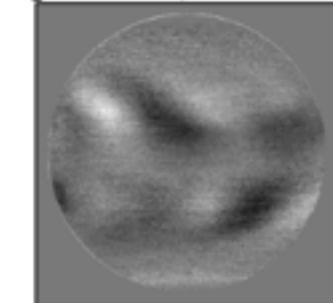
Eigensun 5 (evr=0.0021)



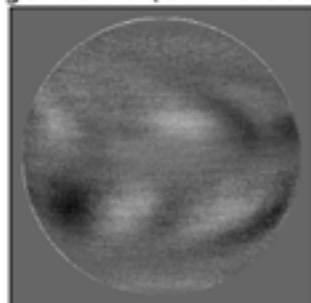
Eigensun 6 (evr=0.0019)



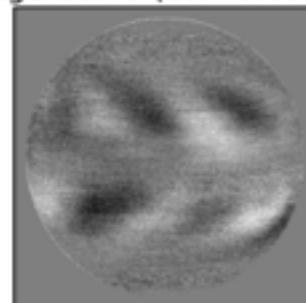
Eigensun 7 (evr=0.0019)



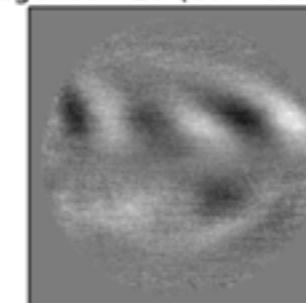
Eigensun 8 (evr=0.0017)



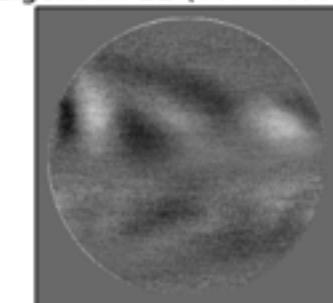
Eigensun 9 (evr=0.0015)



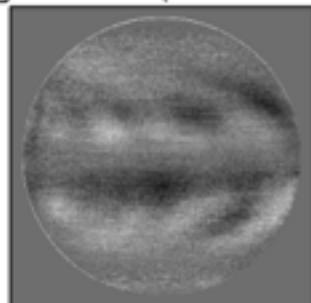
Eigensun 10 (evr=0.0015)



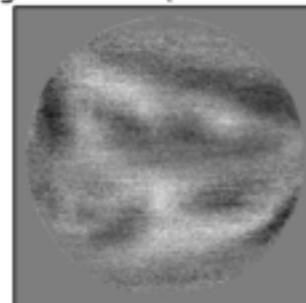
Eigensun 11 (evr=0.0015)



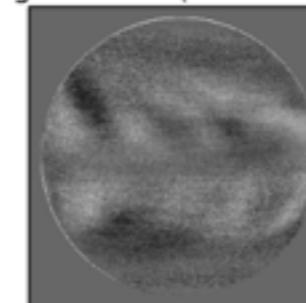
Eigensun 12 (evr=0.0014)



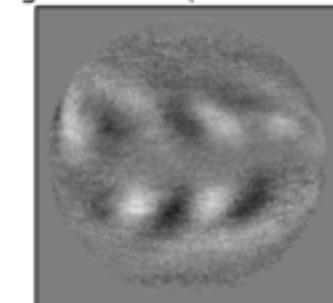
Eigensun 13 (evr=0.0013)



Eigensun 14 (evr=0.0013)

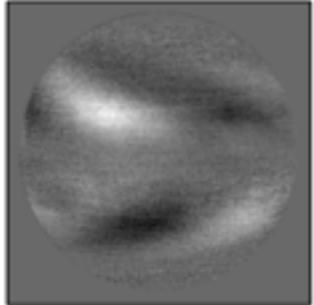


Eigensun 15 (evr=0.0012)

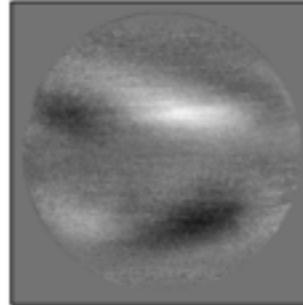


Part 2: CME Eigensuns

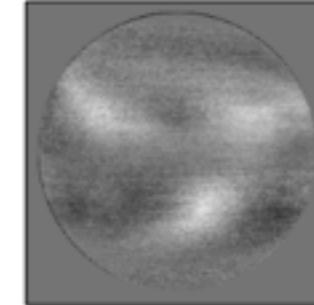
Eigensun 0 (evr=0.0027)



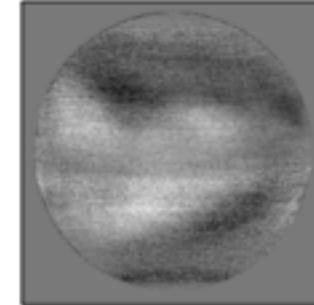
Eigensun 1 (evr=0.0027)



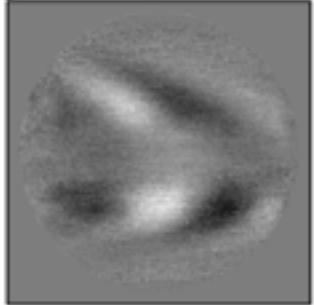
Eigensun 2 (evr=0.0024)



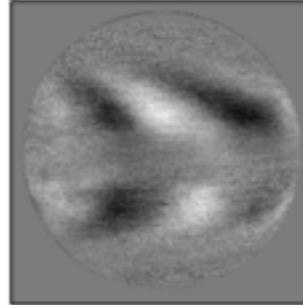
Eigensun 3 (evr=0.0022)



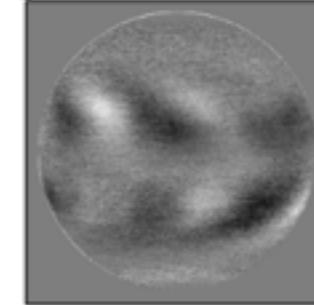
Eigensun 4 (evr=0.0022)



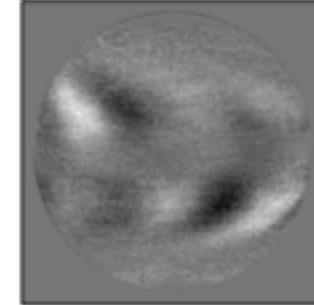
Eigensun 5 (evr=0.0021)



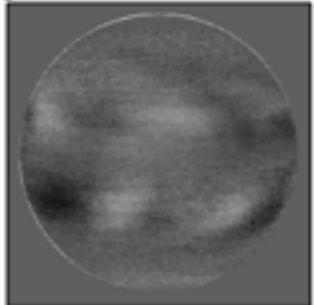
Eigensun 6 (evr=0.0019)



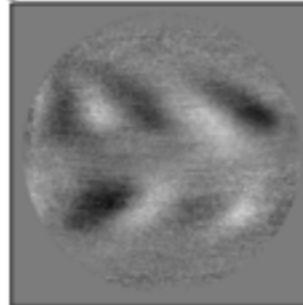
Eigensun 7 (evr=0.0019)



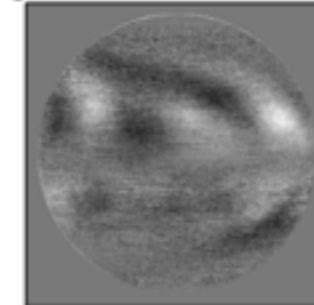
Eigensun 8 (evr=0.0017)



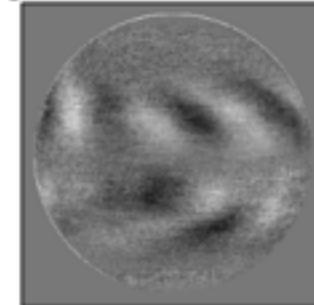
Eigensun 9 (evr=0.0016)



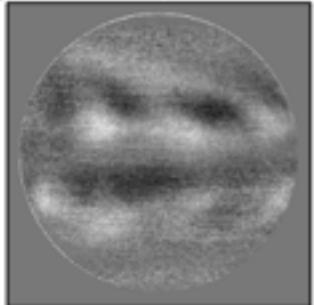
Eigensun 10 (evr=0.0016)



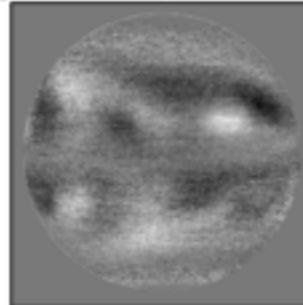
Eigensun 11 (evr=0.0015)



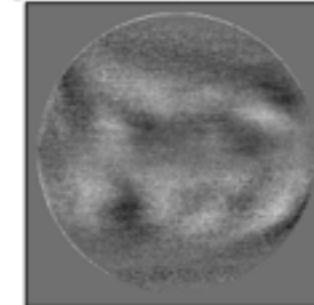
Eigensun 12 (evr=0.0014)



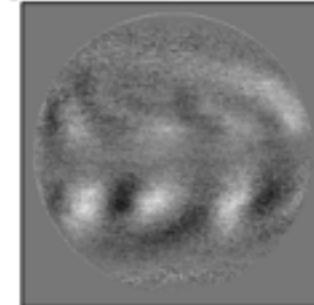
Eigensun 13 (evr=0.0014)



Eigensun 14 (evr=0.0013)



Eigensun 15 (evr=0.0012)



Part 2: PCA Result + time since last event + image similarity

R^2 train: 0.688482446206

R^2 test: 0.688748546665

	precision	recall	f1-score	support
0	0.98	0.69	0.81	86321
1	0.03	0.47	0.06	1979
avg / total	0.96	0.69	0.80	88300

	precision	recall	f1-score	support
0	0.98	0.69	0.81	37029
1	0.03	0.48	0.06	815
avg / total	0.96	0.69	0.80	37844

Area Under Curve train: 0.581466741011

Area Under Curve test: 0.587151522923

Area Under Curve average: 0.584309131967



Part 2: PCA Result + time since last event + image similarity + grey count

R^2 train: 0.672718006795

R^2 test: 0.671757742311

	precision	recall	f1-score	support
--	-----------	--------	----------	---------

0	0.98	0.68	0.80	86368
1	0.04	0.53	0.07	1932

avg / total	0.96	0.67	0.79	88300
-------------	------	------	------	-------

	precision	recall	f1-score	support
--	-----------	--------	----------	---------

0	0.98	0.67	0.80	36982
1	0.04	0.54	0.07	862

avg / total	0.96	0.67	0.78	37844
-------------	------	------	------	-------

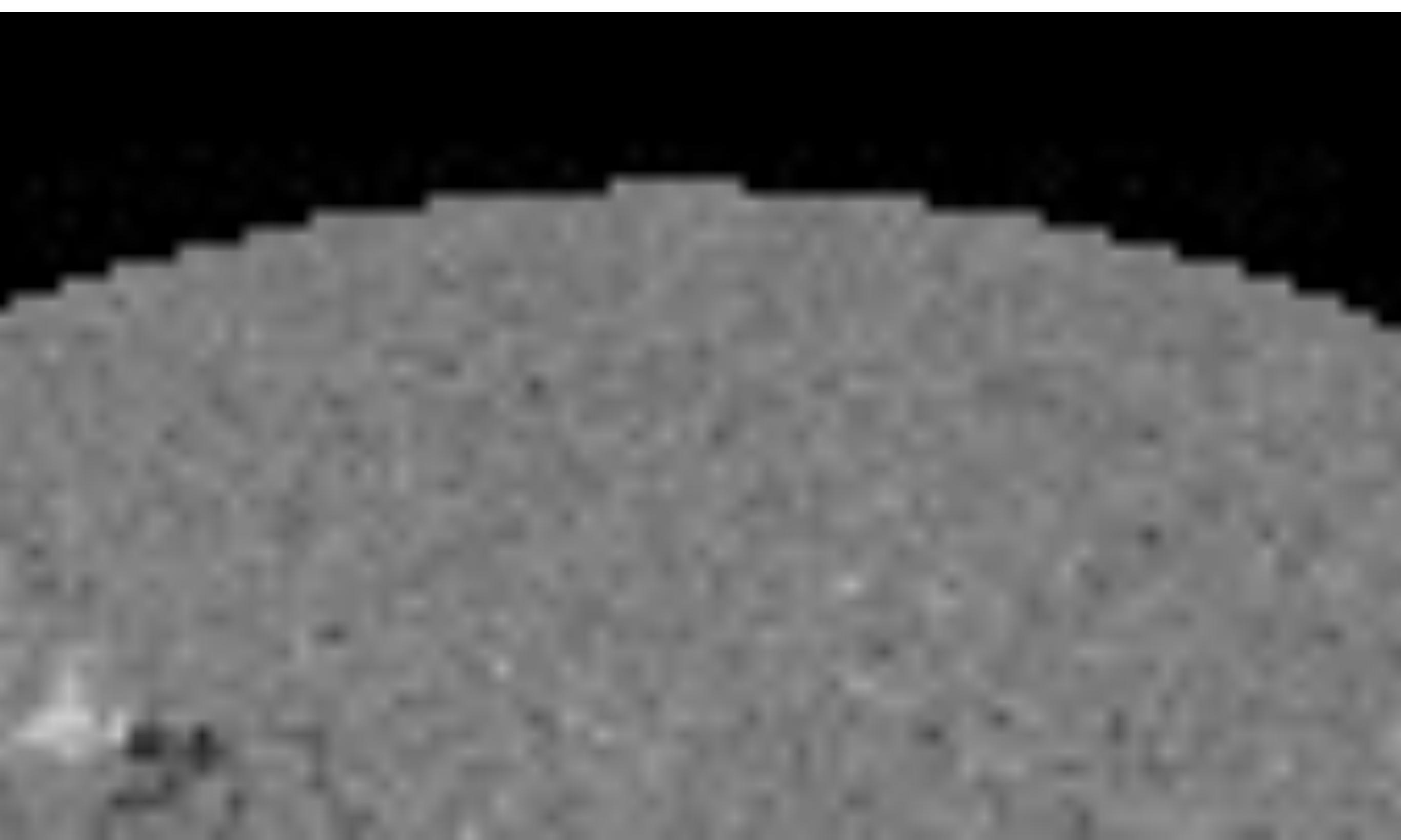
Area Under Curve train: 0.601700331654

Area Under Curve test: 0.60827553782

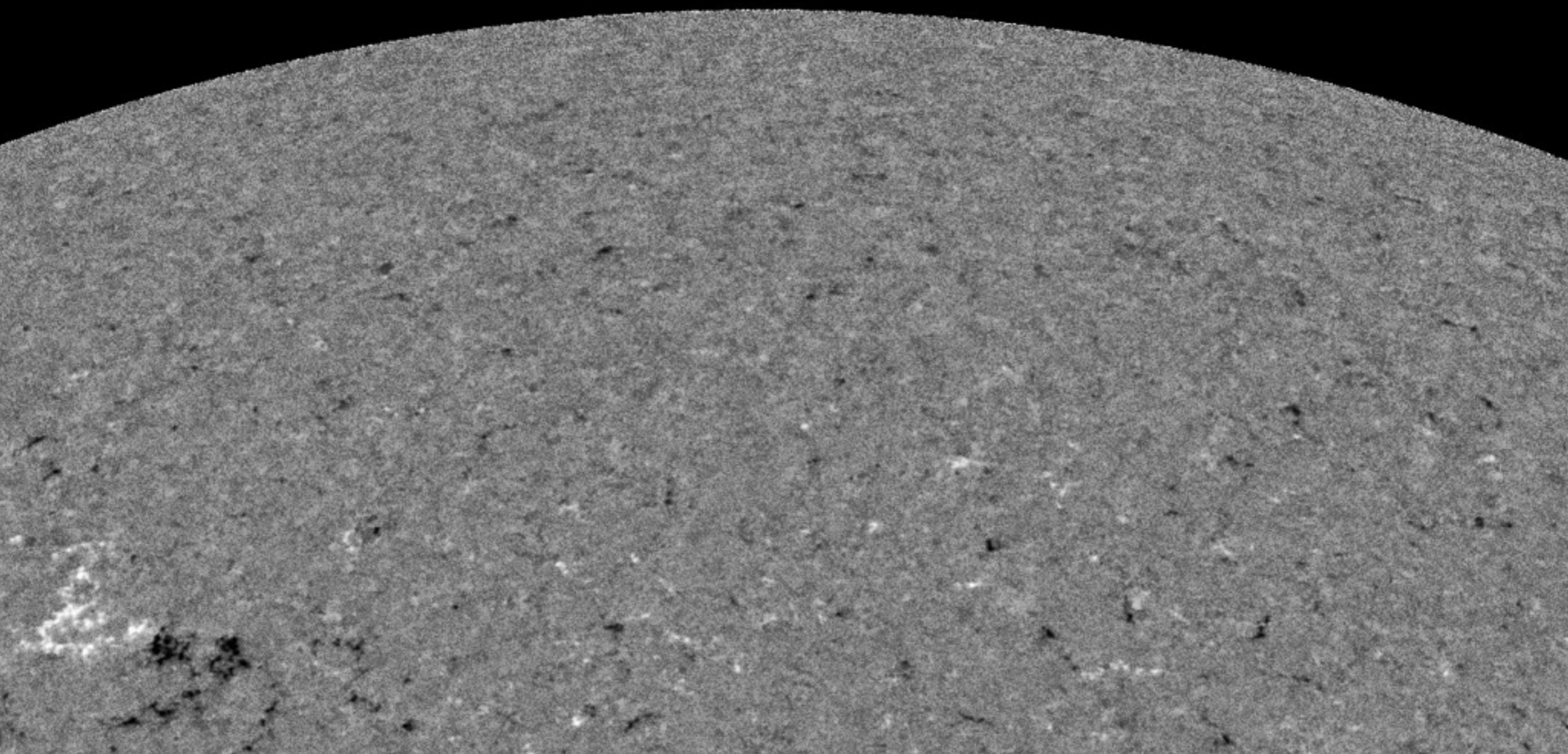
Area Under Curve average: 0.604987934737



Resolution!



Resolution!



Conclusions

- Very limited success - hints of the existence of predictive capability but likely at length / time scales smaller than resolution
- Most variance in longitudinal bands above equator but below poles
- PCA won't work well if the basis of what you are trying to model is transient
- Asyncio is cool
- Don't number crunch with an Air

Con 'grad' ulations!

- Thanks to Alfred, Patrick, Sean, and Ashley

https://github.com/dinob0t/ga_project_final