SOTON-WAIS @ SED2013

Sparse Features and incremental density based clustering

Sina Samangooei

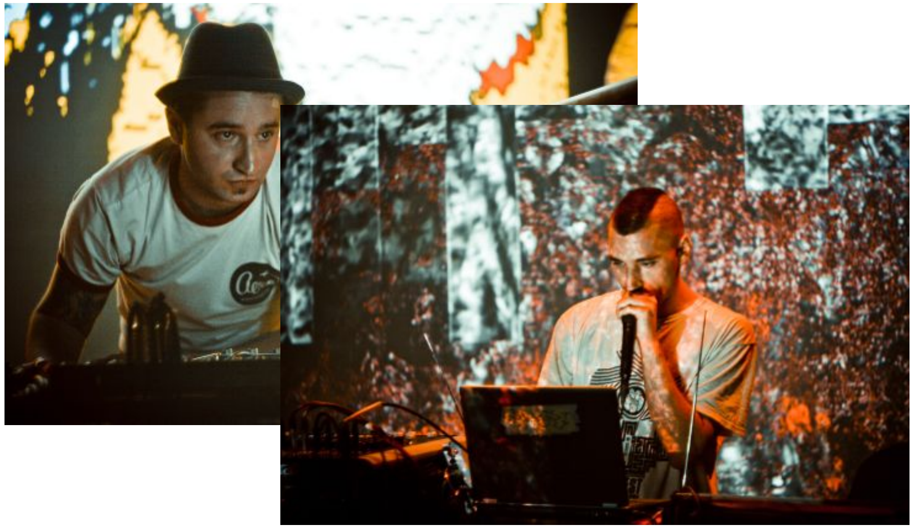
SOCIAL EVENT DETECTION @ MEDIAEVAL 2013

- Event clustering of multimodal social media streams
- Specifically:
  - Given 500k Flickr images with
    - image, tags, (some) geo, (some) time taken, time posted etc.
  - Cluster into "events"

SOCIAL EVENTS

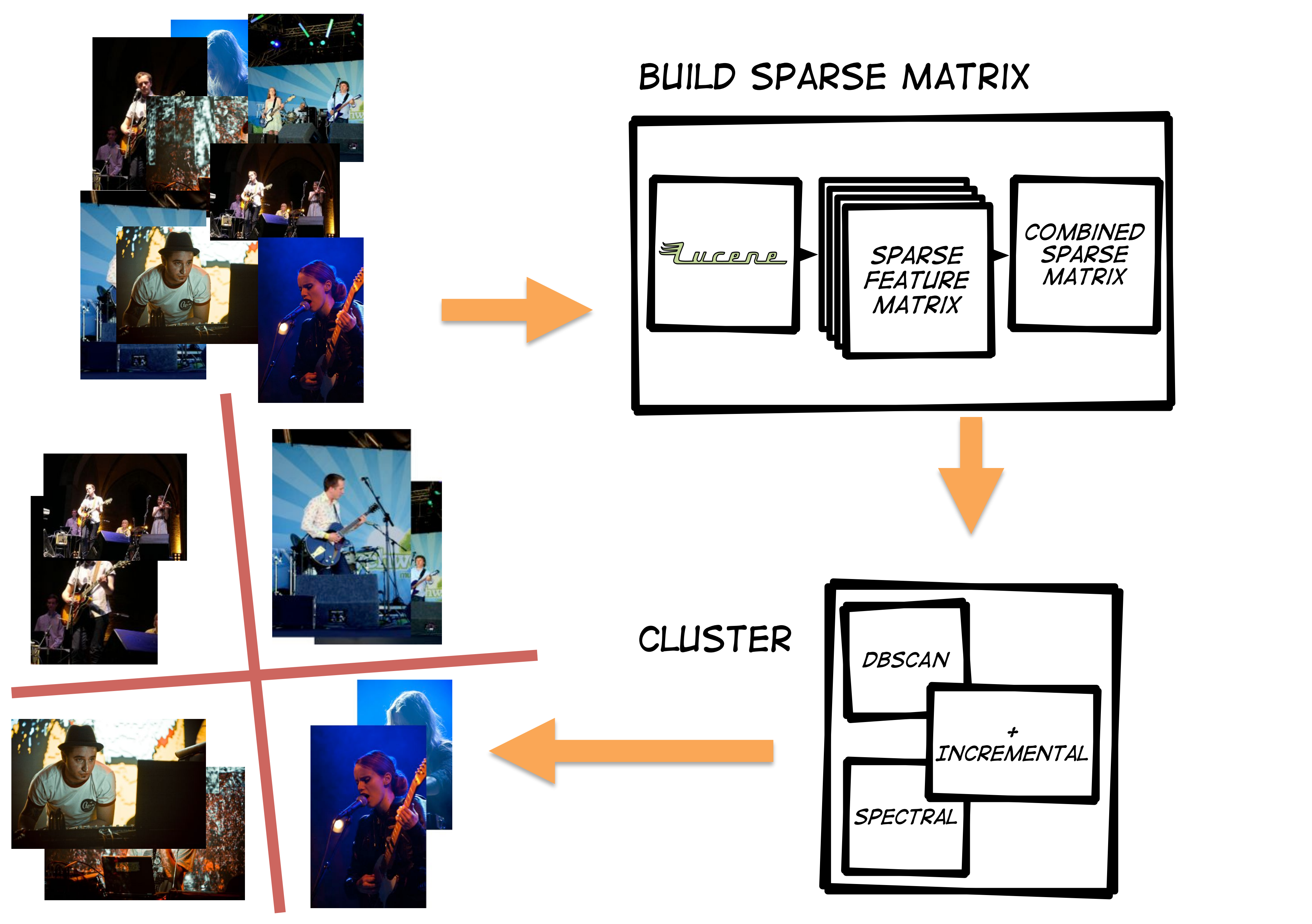
We define **social events** as events that are **planned** by people, **attended** by people and the media illustrating the events are **captured by people**

SED2013 EXAMPLES



```
<photo id="4302746429" photo_url="http://farm5.staticflickr.com/4068/4302746429_f8cd7f2582.jpg" username="at the foot of the hill" dateTaken="2010-01-23 21:36:05.0" dateUploaded="2010-01-25 10:43:34.0">
<title>Poo, Dead Voices on Air @ A4</title>
<description>whole set ...</description>
<tags>
 <tag>2/58</tag>
 <tag>58mm</tag>
 <tag>concert</tag>
 <tag>f2</tag>
 <tag>Poo</tag>
</tags>
</photo>
```

```
<photo id="2280060852" photo_url="http://farm3.staticflickr.com/2110/2280060852_16539f5f0d.jpg" username="Pilot_10" dateTaken="2008-02-19 23:34:09.0" dateUploaded="2008-02-20 18:39:50.0">
<title>Jens Lekman @ Teatro Rasi, Ravenna</title>
<description>19 febbraio 2008</description>
<tags>
 <tag>All rights reserved</tag>
 <tag>concerti</tag>
 <tag>coolpix_4300</tag>
 <tag>emiliaromagna</tag>
</tags>
<location latitude="44.4153" longitude="12.2052"></location>
</photo>
```

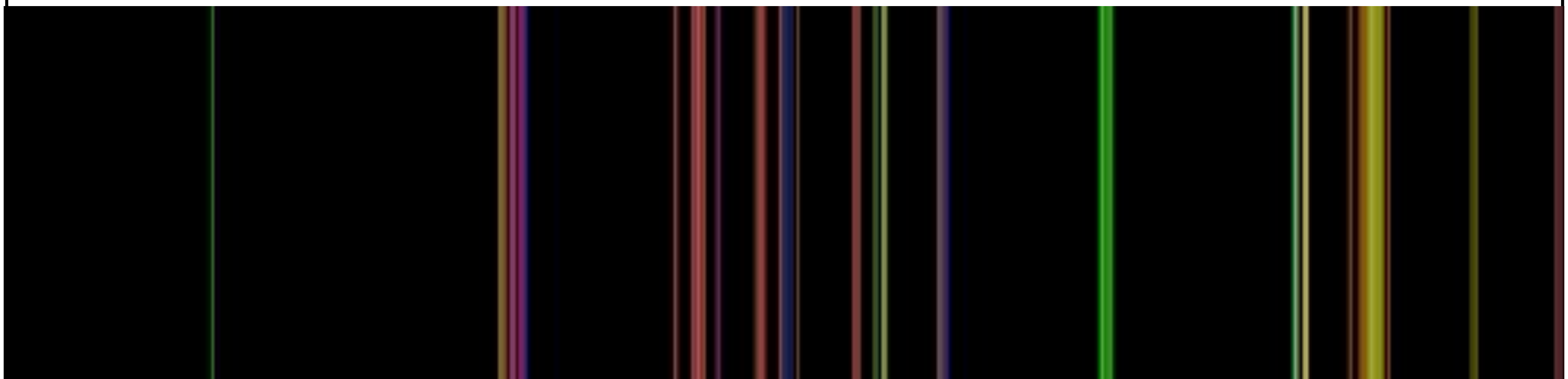

BUILD SPARSE MATRIX
SPARSE FEATURE MATRIX
COMBINED SPARSE MATRIX
CLUSTER
DBSCAN
+ INCREMENTAL
SPECTRAL

FEATURES

- Events potentially separable using:
  - **Images**: should look similar?
  - **Time**: should be temporally close?
  - **Location**: should be geographically close?
  - **Text**: should be described similarly?
- Our social media stream contains:
  - Time taken (potentially inaccurate)
  - Time posted (accurate, though may be event agnostic)
  - Geo (often inaccurate, sparse)
  - Tags, title, description(multi-tag, spelling etc.)

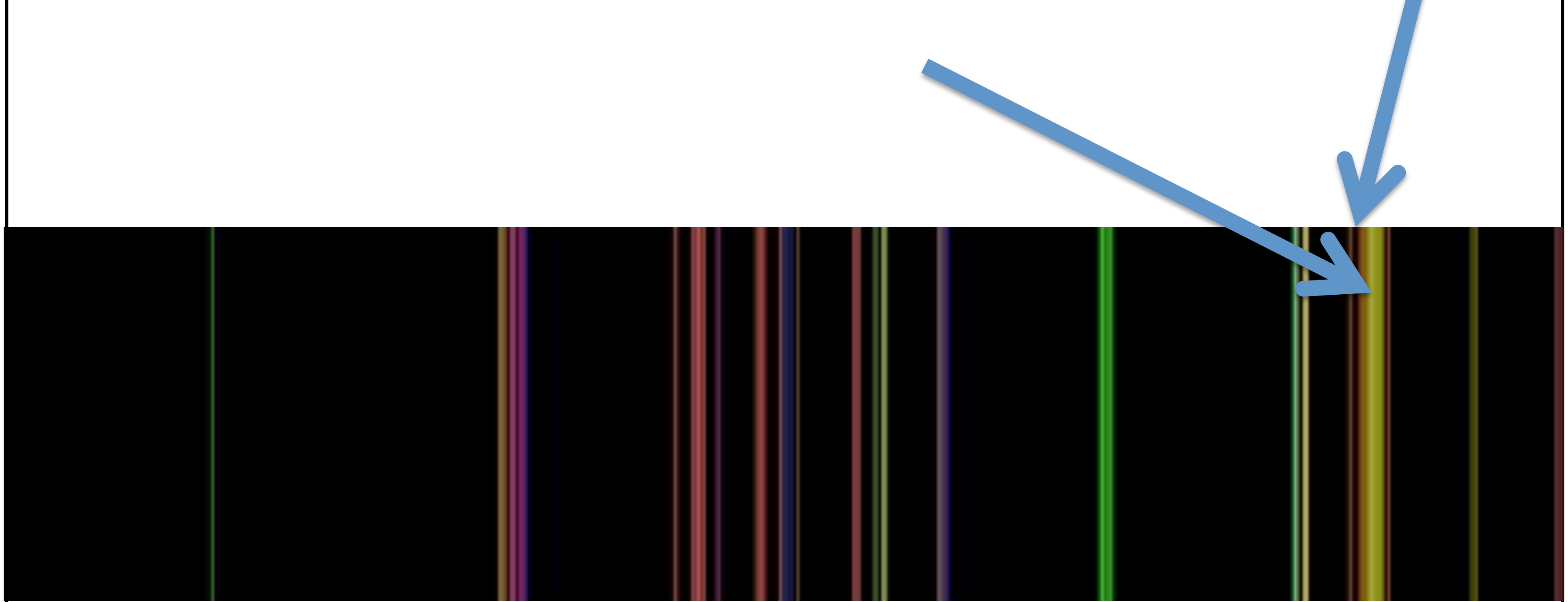
EVENT SEPARATION WITH FEATURES

- January 2007 until February 2007
- Random color assigned to clusters



EVENT SEPARATION WITH FEATURES

- Events like this could easily be confused



EVENT SEPARATION WITH FEATURES

- More may help separate events

FEATURE WEIGHTS

- The features **matter** for different reasons
- Some are more important than others
  - This is a feature fusion problem
- Experiments with feature weights against cluster quality
  - **Time taken** is apparently most important
  - **Time posted + geo** seem to hold the same information
  - Tags **beat** titles and descriptions

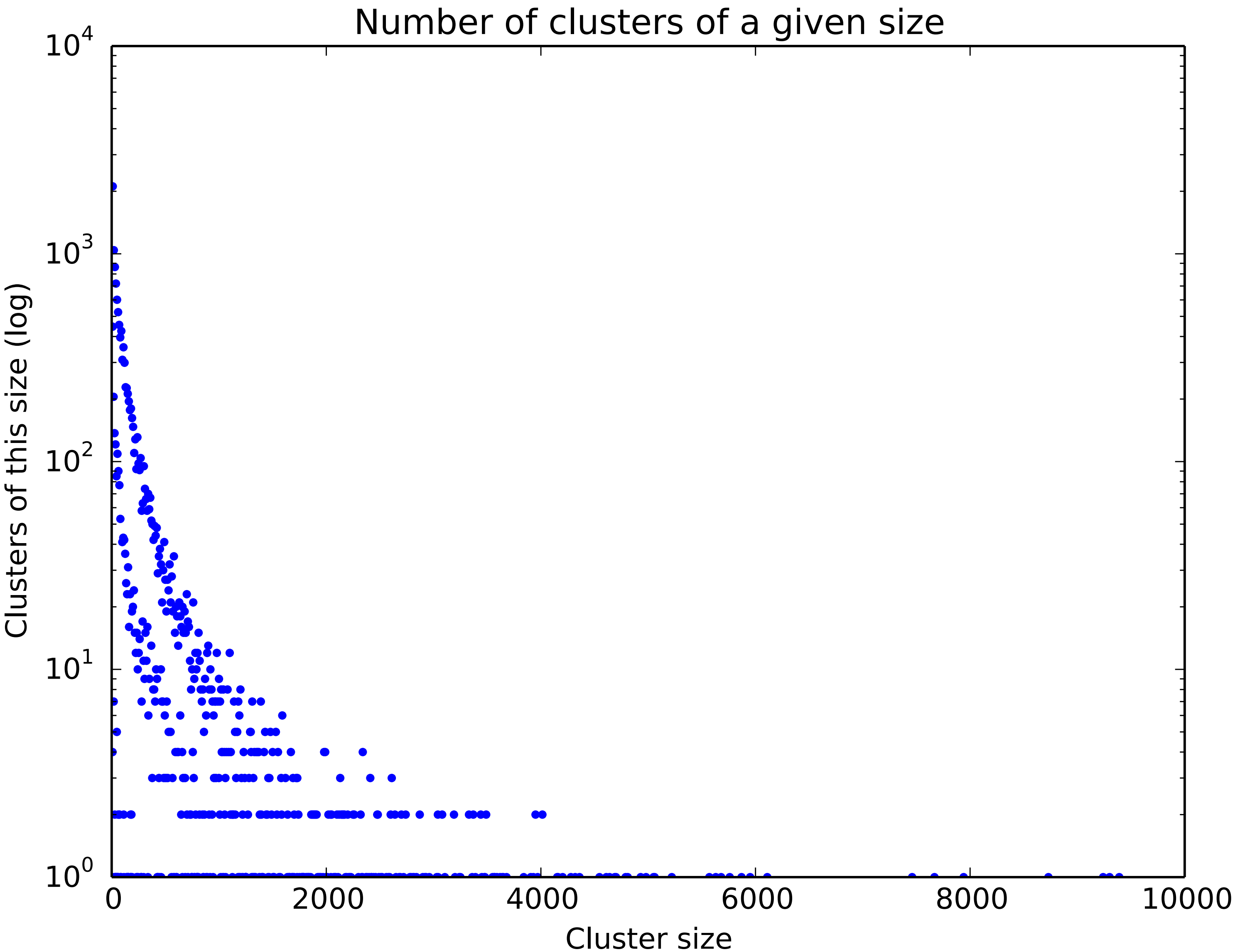
ENFORCE SPARSITY

- Use a query a **Lucene index** to get a subset of similar images
- Distance measures of Time and Geo calculated using a log decay function
  - force sparsity beyond certain threshold
- Distance measures of tags are inherently sparse
  - TFIDF

CLUSTERING – FINDING EVENTS

- Challenges
  - The baseline is it self noisy
    - Hard to know if we're doing well!
  - The task is ill posed (is Christmas an event? Are all Christmases an event? application specific)
  - Cluster number hard to estimate
    - A parameter in many clustering algorithms
  - Many noise points
    - 2% clusters with 1 member
  - Long tail of cluster membership

Number of clusters of a given size



SED2013 – DBSCAN

….the little baseline that could…

- DBSCAN is an old, well studied clustering algorithm
- Detects clusters and identify noise
- No knowledge of cluster count needed
- Requirements:
  - Neighborhood function (e.g. thresholded sparse similarity matrix)
  - Neighborhood density counts

SED2013 – SPECTRAL CLUSTERING

- Theoretically appealing non-parametric clustering algorithm
  - Rooted in graph theory
  - Potentially auto detects cluster count
- Basic premise is:

  *Use the smallest **(near zero)** eigenvalued eigenvectors of the **graph laplacian** of the **similarity matrix** of some data as a space within which to apply **another clustering algorithm***

SED2013 – INCREMENTAL CLUSTERING

- Practical restrictions of spectral clustering mean we can't apply it to the whole dataset
- Make an assumption about the data style
  - images likely to be clustered together will appear sequentially in terms of upload time
- Leverage this to cluster sub windows of data
  - Grow the window and see if a cluster changes
  - If not, tag those items as clustered, and remove them

RESULTS!

SETTING	TIME TAKEN	TIME POSTED	LOCATION	TEXT DESC	TEXT TITLE	TEXT TAGS
BEST	2	0	1	1	0	3
AVERAGE	2.1	1.8	1.4	0.7	0.3	1.7
WORST	0	1	1	3	3	0

	F1	NMI	F1(DIV)	RB F1	DIV F1
DBSCAN (BEST)	0.945	0.985	0.935	0.059	**0.887**
SPECTRAL (BEST)	0.911	0.977	0.882	0.058	0.853
DBSCAN (AVG)	**0.946**	**0.985**	**0.936**	0.060	0.886
SPECTRAL (AVG)	0.902	0.974	0.866	0.057	0.846
DBSCAN (WORST)	0.409		0.353	0.056	

RESULTS!

SETTING	TIME TAKEN	TIME POSTED	LOCATION	TEXT DESC	TEXT TITLE	TEXT TAGS
BEST	2	0	1	1	0	3
AVERAGE	2.1	1.8	1.4	0.7	0.3	1.7
WORST	0	1	1	3	3	0

All configurations using incremental clustering

	F1	NMI	F1(DIV)	RB F1	DIV F1
DBSCAN (BEST)	0.945	0.985	0.935	0.059	**0.887**
SPECTRAL (BEST)	0.911	0.977	0.882	0.058	0.853
DBSCAN (AVG)	**0.946**	**0.985**	**0.936**	0.060	0.886
SPECTRAL (AVG)	0.902	0.974	0.866	0.057	0.846
DBSCAN (WORST)	0.409		0.353	0.056	

RESULTS!

SETTING	TIME TAKEN	TIME POSTED	LOCATION	TEXT DESC	TEXT TITLE	TEXT TAGS
BEST	2	0	1	1	0	3
AVERAGE	2.1	1.8	1.4	0.7	0.3	1.7
WORST	0	1	1	3	3	0

Time Taken matters most!

	F1	NMI	F1(DIV)	RB F1	DIV F1
DBSCAN (BEST)	0.945	0.985	0.935	0.059	**0.887**
SPECTRAL (BEST)	0.911	0.977	0.882	0.058	0.853
DBSCAN (AVG)	**0.946**	**0.985**	**0.936**	0.060	0.886
SPECTRAL (AVG)	0.902	0.974	0.866	0.057	0.846
DBSCAN (WORST)	0.409		0.353	0.056	

WEIGHTS MATTER!

SETTING	TIME TAKEN	TIME POSTED	LOCATION	TEXT DESC	TEXT TITLE	TEXT TAGS
BEST	2	0	1	1	0	3
AVERAGE	2.1	1.8	1.4	0.7	0.3	1.7
WORST	0	1	1	3	3	0

They matter quite a lot

	F1	NMI	F1(DIV)	RB F1	DIV F1
DBSCAN (BEST)	0.945	0.985	0.935	0.059	**0.887**
SPECTRAL (BEST)	0.911	0.977	0.882	0.058	0.853
DBSCAN (AVG)	**0.946**	**0.985**	**0.936**	0.060	0.886
SPECTRAL (AVG)	0.902	0.974	0.866	0.057	0.846
DBSCAN (WORST)	0.409		0.353	0.056	

ANY QUESTIONS OR COMMENTS?