



# Event detection from blogs and news sources via RSS

Mike Thelwall

Statistical Cybermetrics Research Group

University of Wolverhampton, UK

<http://cybermetrics.wlv.ac.uk/>



# Why blog analysis?



Wilde

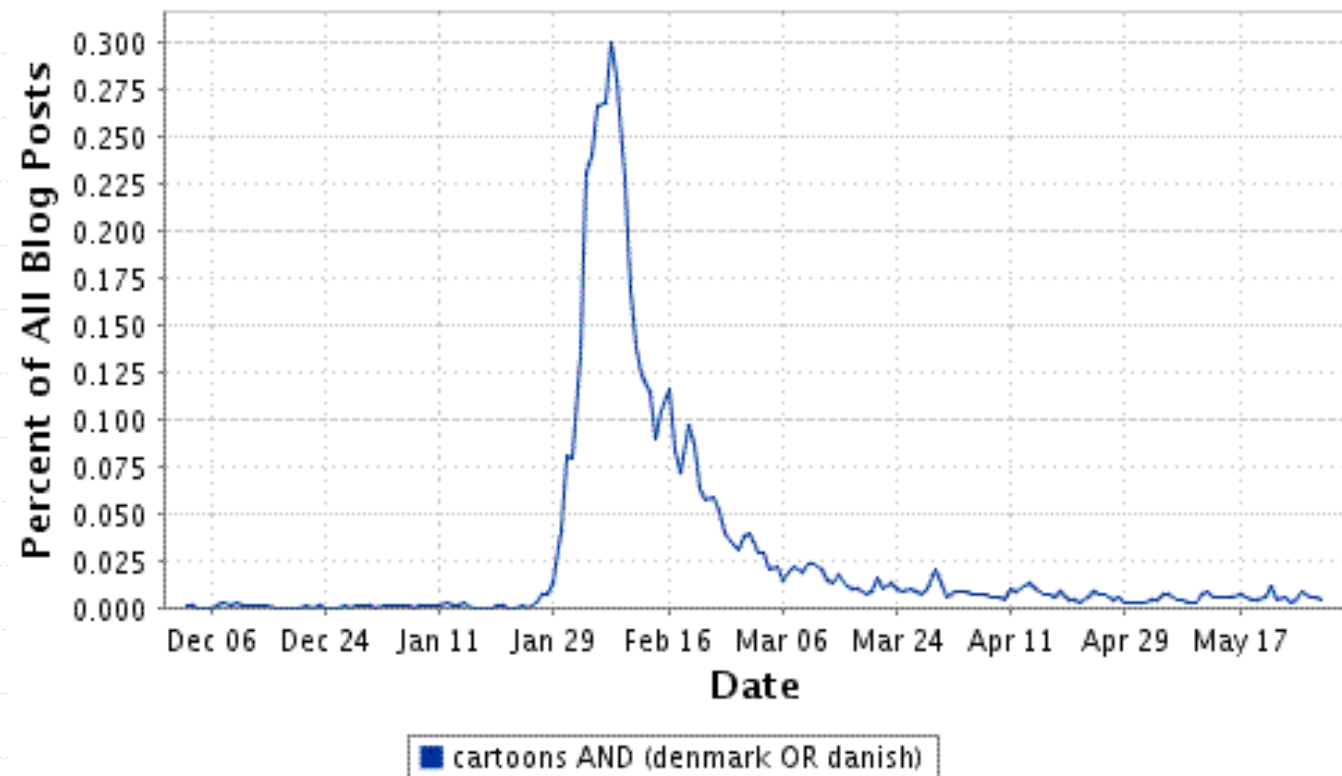
- ◆ There are hundreds of millions of bloggers
- ◆ Analysing blog post trends may give insights into public opinion
- ◆ Blog analysis
  - is simple, modern, highly flexible
  - can generate a wide variety of small self-contained interesting research projects with
  - can promote thought about key research ideas – such as method limitations

Hindsight fallacy

# Example of blog study

- ◆ Retrospective analysis of the genesis of the Danish Cartoons affair

Generated by BlogPulse Copyright 2006 Nielsen BuzzMetrics.



*A graph of the volume of blogging about the Danish Cartoons*



# Tracking Issues in Blogs

Using blogs as a public opinion repository

# Consumer monitoring and retrospective public opinion



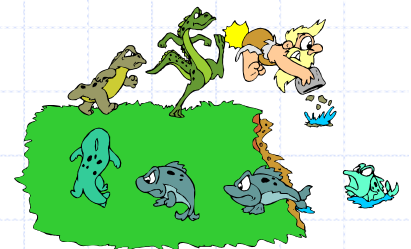
- ◆ IBM, Microsoft, [Nielsen Buzzmetrics](#), and [Market sentinel](#) etc. monitor consumer opinion as a market research service
  - Read consumers' minds about what they like/dislike – from their blog/forum postings
- ◆ Researchers can analyse public opinion retrospectively via simple tools such as [blogpulse.com](#)
  - Only source of general retrospective public opinion
  - But only works well for huge issues

# Blog searches

- ◆ Blog search engines allow blog-only searches for current opinion or *retrospective opinion*
  - [Blogpulse](#), Technorati , Blogdigger, IceRocket (allows date-specific searches)
- ◆ Blog searching allows genuine *date range* search
  - NB Google only allows “last updated” date range searches
  - Research idea: Pick a topic and run a content analysis of 100 blog posts from date 1, comparing it to a content analysis of 100 blog posts from date 2, looking for differences.

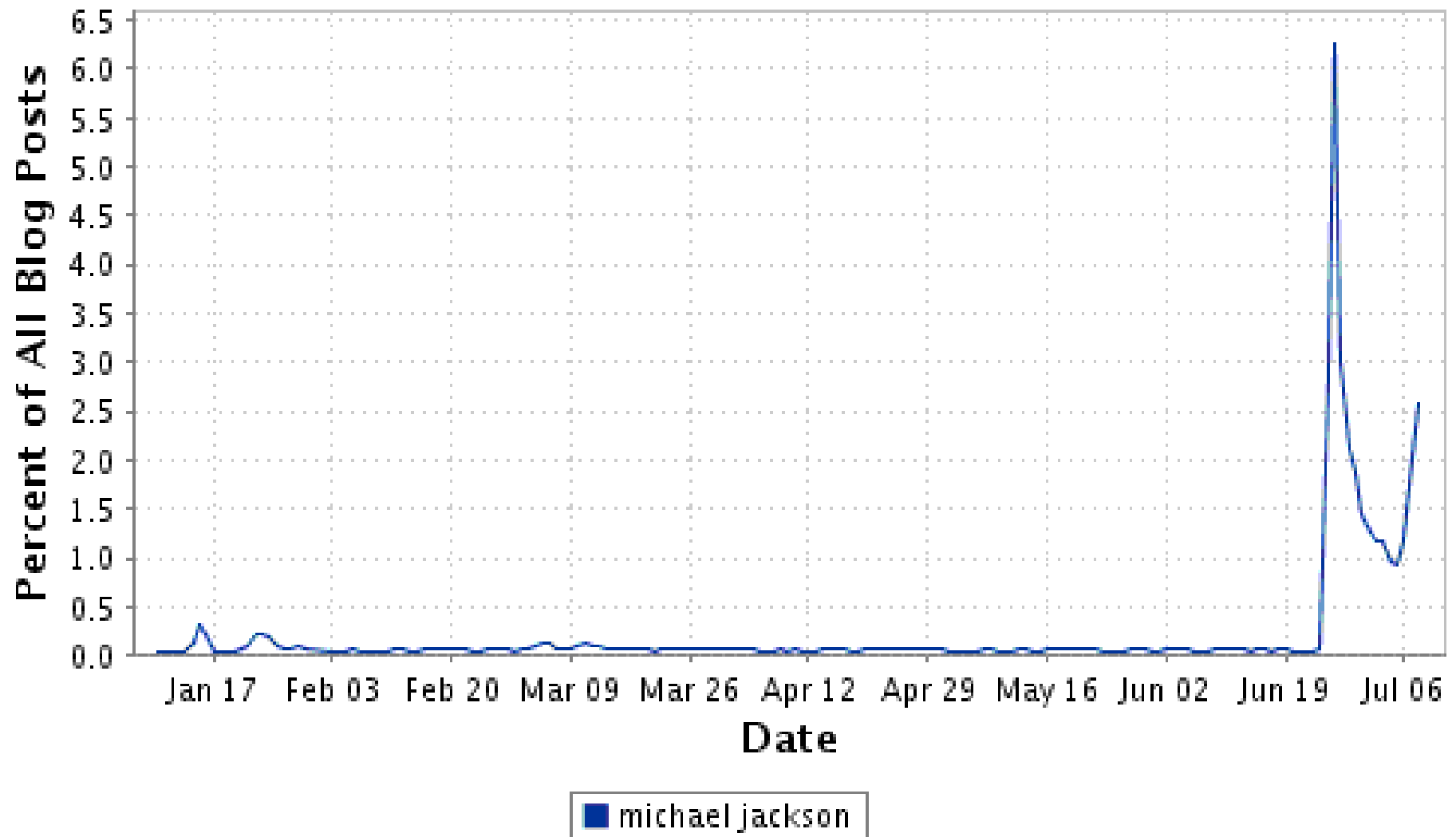
# Tracking evolution over time

- ◆ Is interest in a topic *changing*
  - E.g., stem cell research/GM food?
- ◆ Blogs can be used to generate *time series* reflecting changes in “public interest”
- ◆ [Blogpulse](#) and [Technorati](#) offer free time series graphs

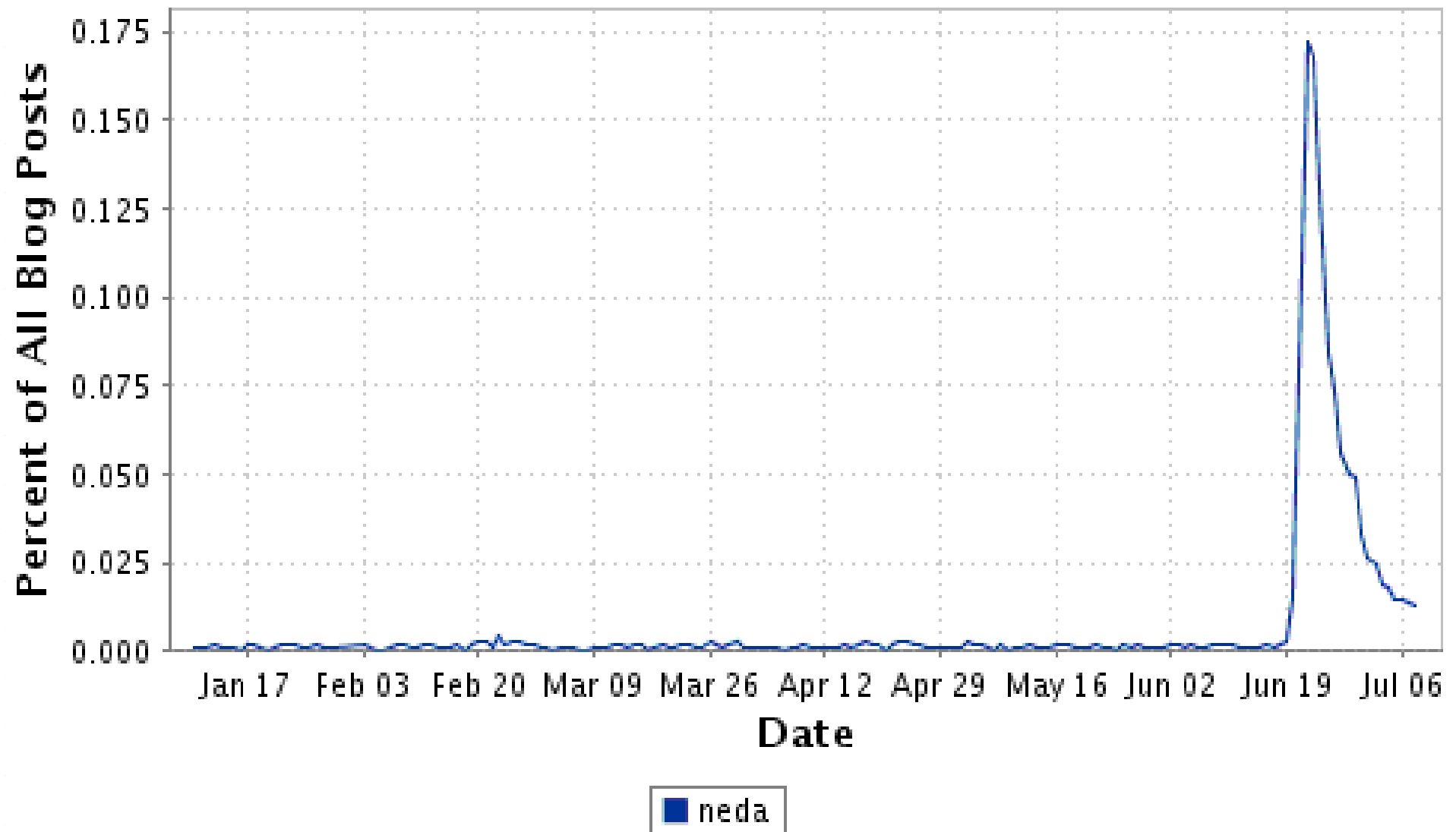


# Free debate evolution graphs

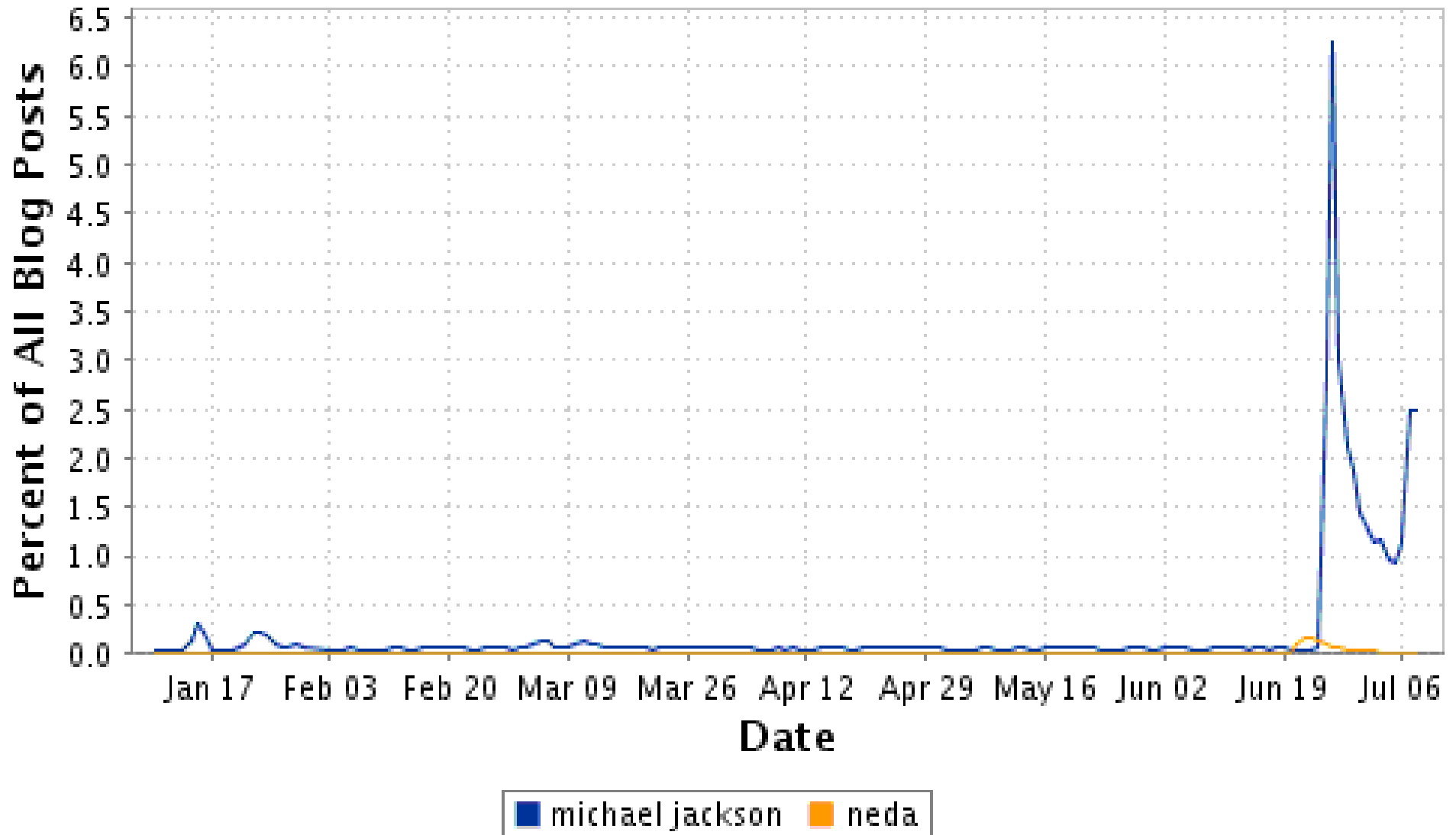
Generated by BlogPulse Copyright 2009 The Nielsen Company.



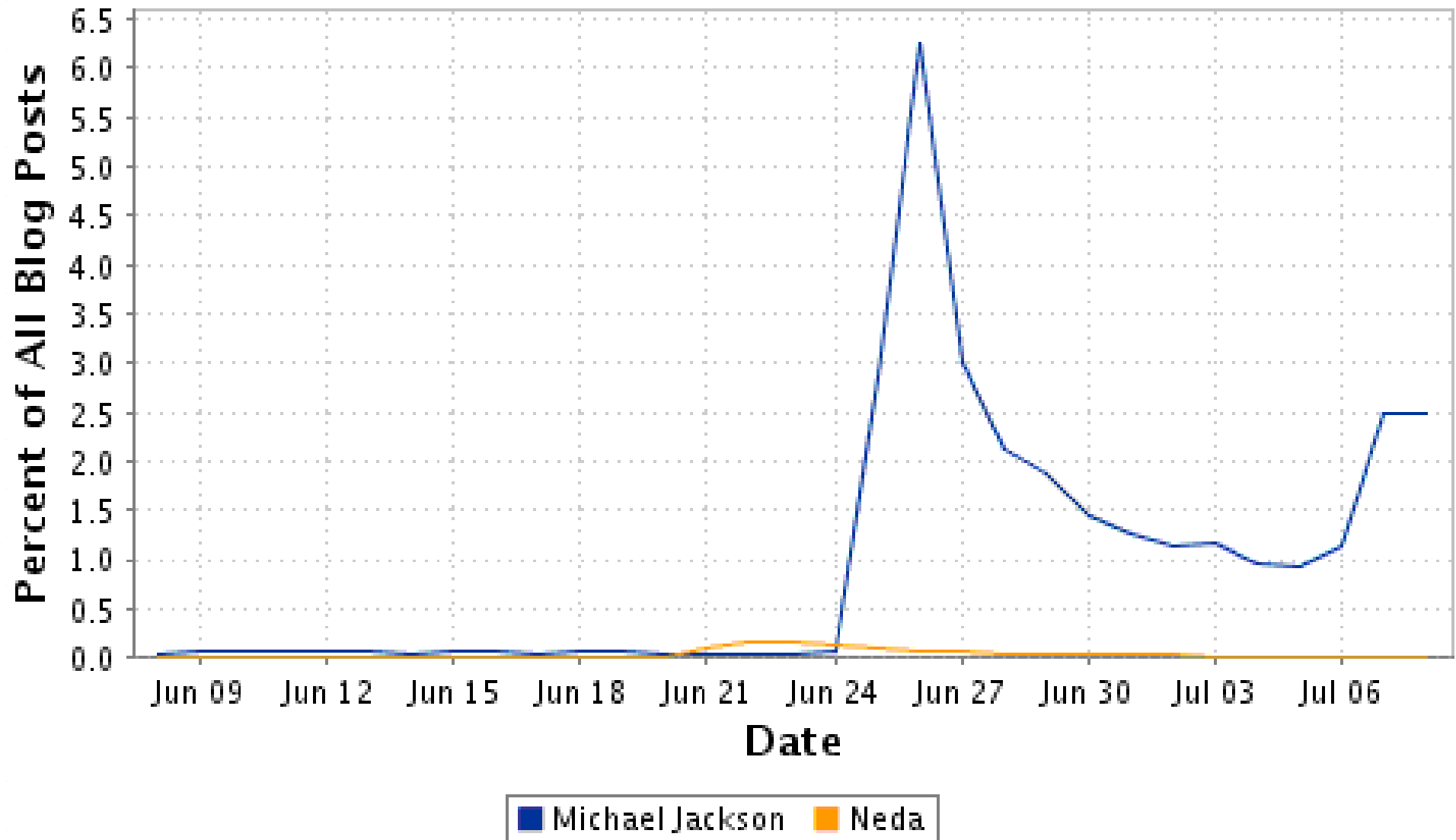
Generated by BlogPulse Copyright 2009 The Nielsen Company.



Generated by BlogPulse Copyright 2009 The Nielsen Company.




Generated by BlogPulse Copyright 2009 The Nielsen Company.





# *Detecting* Emerging Issues from Blogs

Blogs can be used to retrospectively identify key events within a broad issue

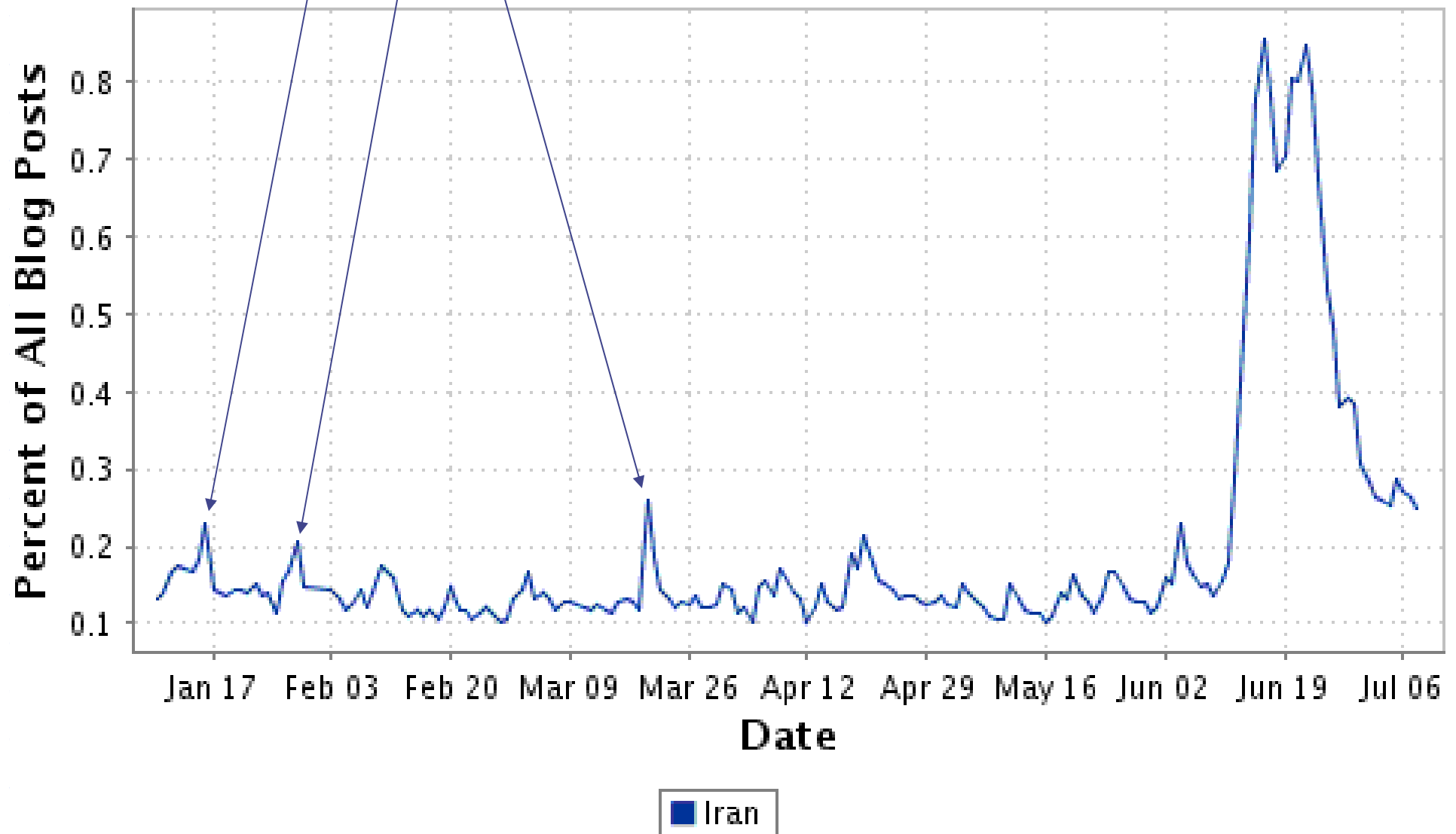


# How can key events be detected?

- ◆ Without blogs - Heuristic methods
  - E.g. Read papers, scan relevant blogs
- ◆ With blogs – look for spikes in a relevant search
  - i.e. look for sudden increase in relevant blogging

# Each spike may be an event

Generated by BlogPulse Copyright 2009 The Nielsen Company.

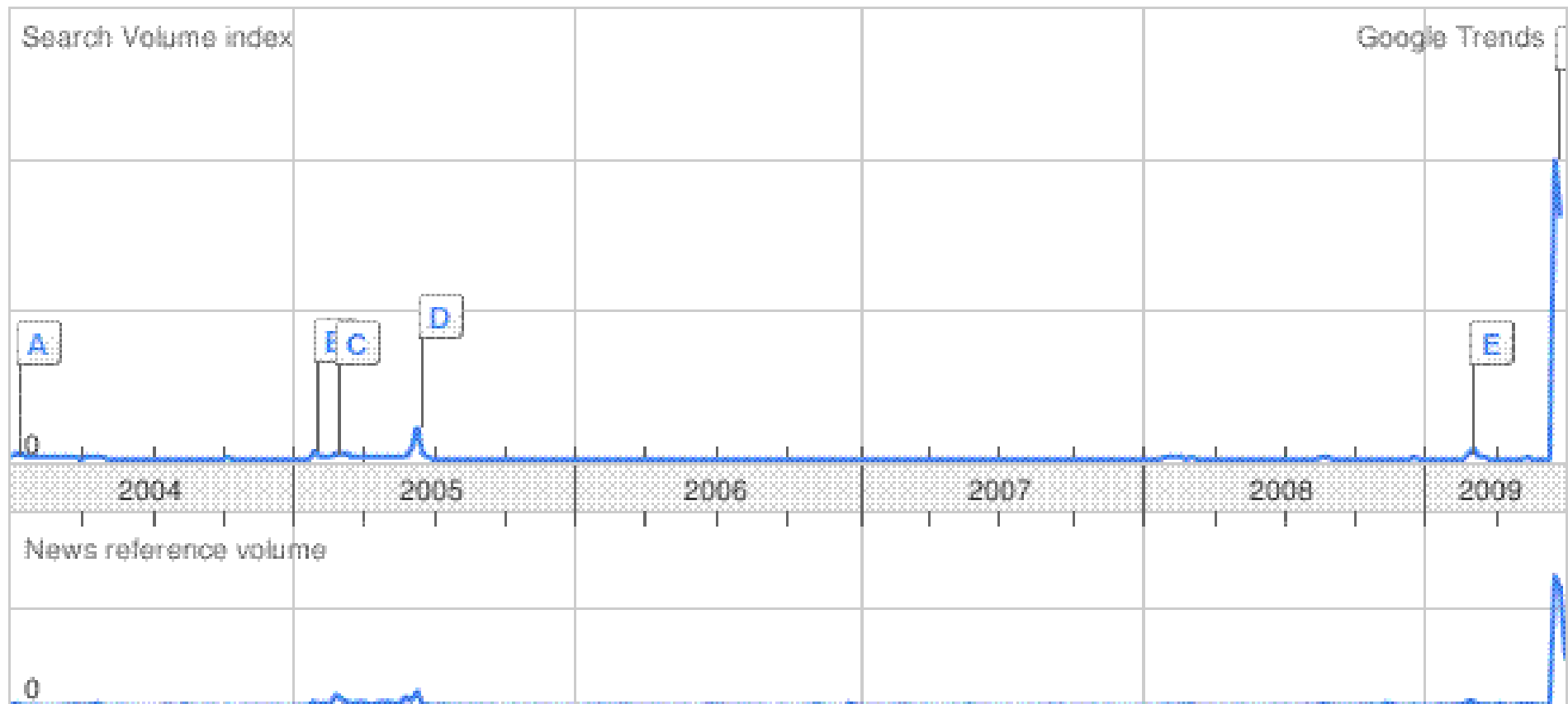


# Google Trends for triangulation

- ◆ Google trends tracks the popularity of Google searches
- ◆ Useful second opinion for blog trends
- ◆ Broader user base but can't do content analysis
- ◆ <http://www.google.com/trends>

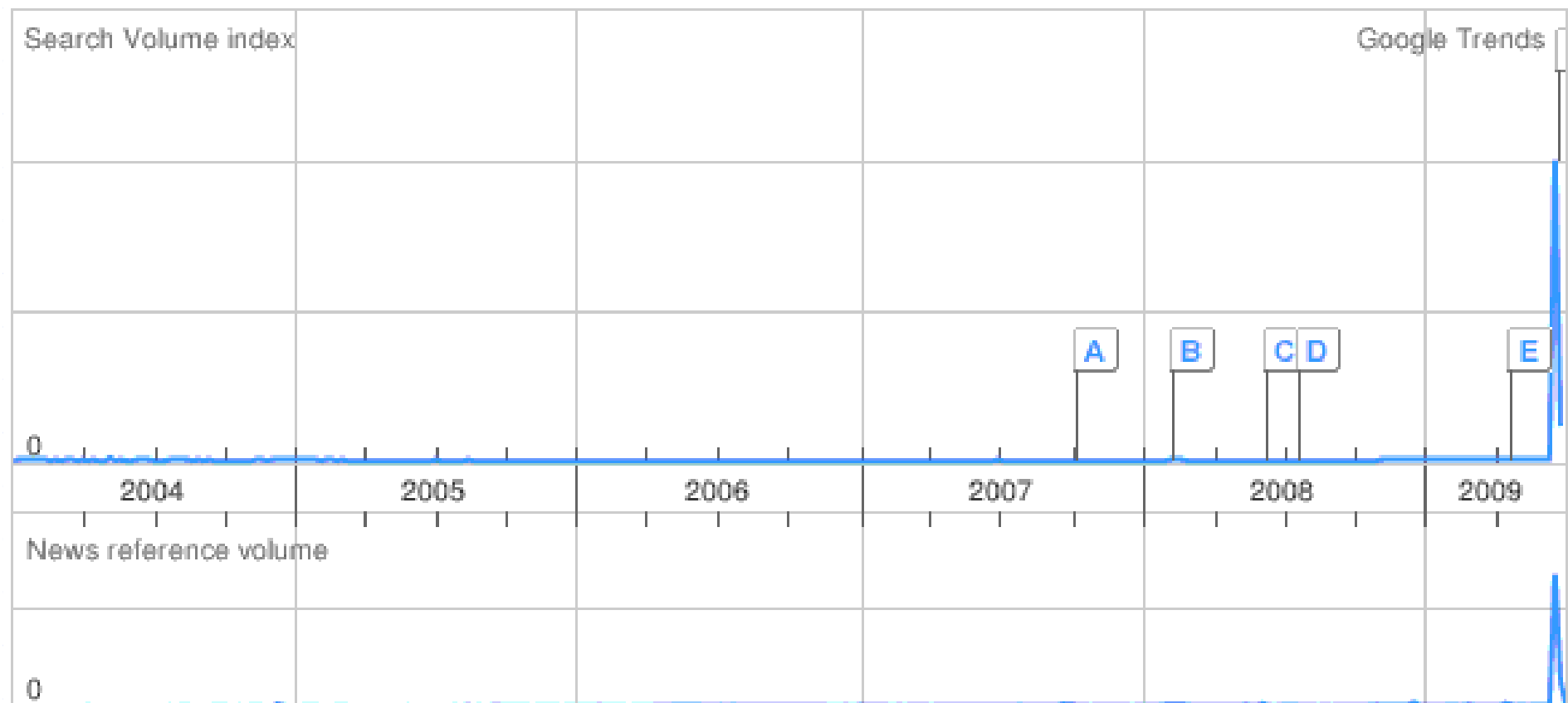
# Michael Jackson

● michael jackson



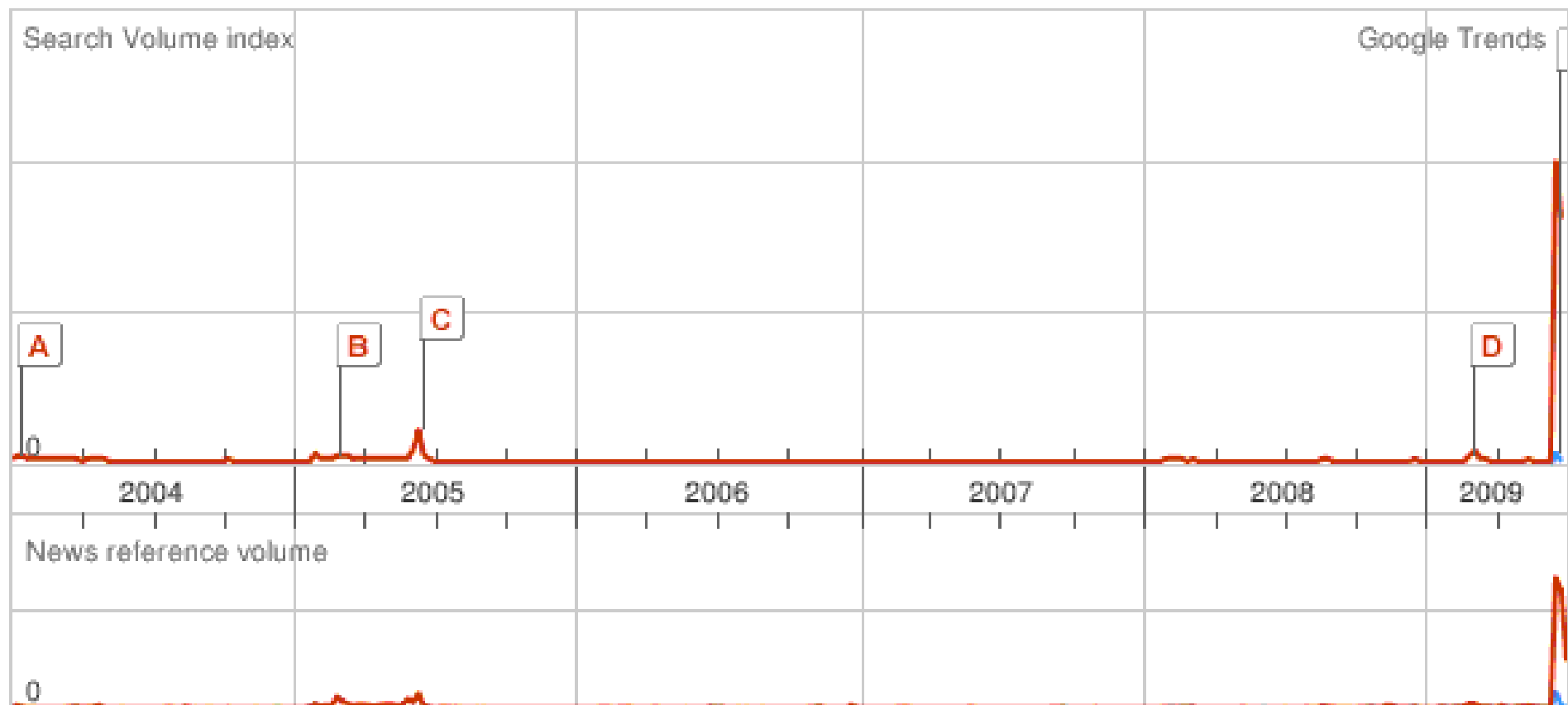
# Neda

● neda



# Michael Jackson vs. Neda

● neda ● michael jackson



# Limitations of trends

- ◆ Can't check the "meaning" of the searches –e.g., which Neda?
- ◆ Searchers may use other terms –e.g., *Neda Youtube video*

# Automatic detection

- ◆ Monitoring ~100k blogs is not too difficult via RSS feeds with their simple format
  - Not all blogs have RSS feeds
  - RSS feeds can be identified via filetype searches in search engines, via their APIs or via crawling
- ◆ Monitoring blogs rather than RSS feeds is much more labour-intensive
  - Need to write parser for each format
- ◆ Mozdeh RSS monitor
  - Generates sub-collections
  - Generates word time series
  - Allows keyword searches
  - Identifies hot topics

# RSS Feeds – very simple format

...

```
<item>
```

```
<title>UK interest rates remain on hold</title>
```

```
<description>The Bank of England holds interest rates  
at the record low of 0.5% for the sixth consecutive  
month.</description>
```

```
<link>http://news.bbc.co.uk/go/rss/-  
/1/hi/business/8247709.stm</link>
```

```
<pubDate>Thu, 10 Sep 2009 11:36:46  
GMT</pubDate>
```

```
<category>Business</category>
```

```
</item>
```

.....

# Science concern corpus

- ◆ EU project to automatically identify public fears about science
- ◆ An RSS collection of news, blog and social network postings containing a fear word AND a science word, matching the search:
  - (science OR scientist OR research OR researcher) AND (fear OR afraid OR worry)
- ◆ Trend detection used to identify hot “science fear” topics- words with sudden increases in usage
- ◆ Manual scanning of top words to identify genuine topics



# Information-theoretic metrics

- ◆ Experimented with different spike/burst measures for “Blogs per word per day”
- ◆ Absolute spike size worked best
- ◆ Also tried chi-sq, mutual information, relative spike size

Different time series scanning options

Word IDs to consider: 1 - 138396

#Items containing word: 2 - 1000000

Dates: 2005.2.18 - 2005.10.2

ReportTotalFeedsPerDayContainingEachWord  
 ReportProportionOfFeedsPerDayContainingEachWord

Do Spike or Burst (step) Test

Min spike/step size (proportion of posts in one day)  
 Min spike or step size / average size

Use Moving Average Instead of Full Time Series  days

Spike Test     Chi Square test (not done yet)

Burst Test  consecutive raised days is a burst

No Test

Make Top Stories Report

Words To Check     Maximum allowable word distance

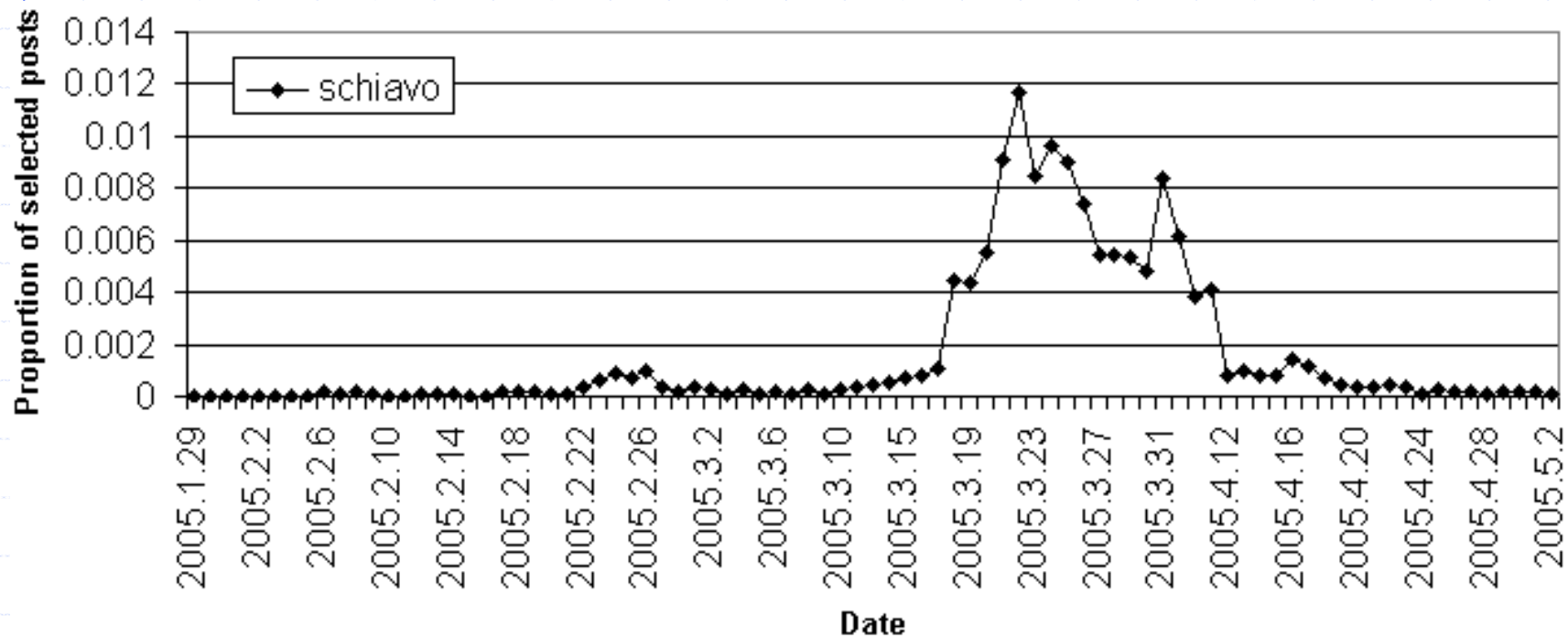
Proportion of cowords needed to Cluster two terms     Include Snippets

Make Time Series File For All Words Matching Conditions     Also Report Gaps Between Words



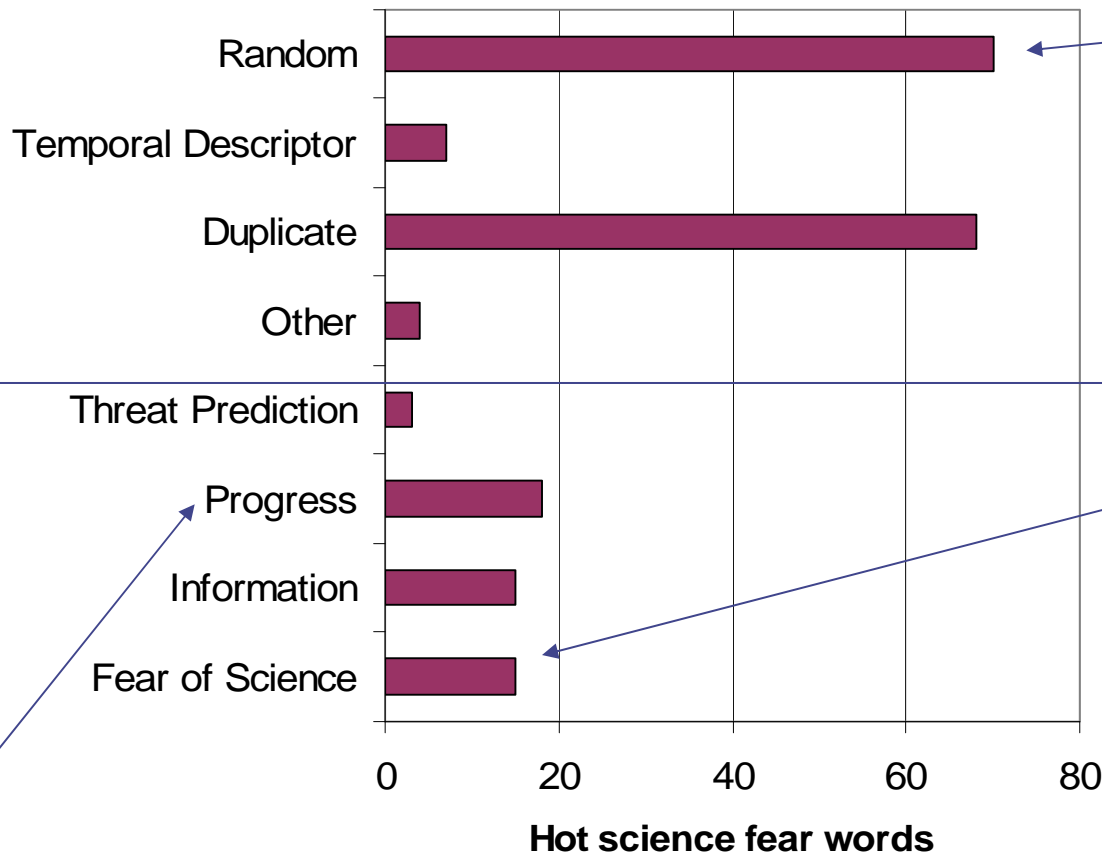
# Mozdeh – Visual output

- ◆ Research blog analysis project
- ◆ Gives control over the data source





# Top 200 words: Types



The words come from multiple stories

7.5% of top 200 Words Represent new public fears of Science stories

E.g. new medical cure



# Top science concern words

Word	Max. daily increase (feeds)	Classification
stem	19%	Science fear (stem cell research)
orlean	16%	Information (about hurricane)
CISCO	14%	Router security fears
Schiavo	12%	Life support machines

7.5% of the top 200 terms were new relevant debates

# Summary

- ◆ Event detection/analysis is easy with free online tools like Blogpulse
- ◆ Also possible with personal software for about 100k blogs – can be customised
- ◆ Works best on big news stories – otherwise too small blog footprint

# Reading

- ◆ Thelwall, M. & Prabowo, R. (2007). Identifying and characterising public science-related fears from RSS feeds. *Journal of the American Society for Information Science and Technology*, 58(3), 379-390.
- ◆ Thelwall, M. (2007). Blog searching: The first general-purpose source of retrospective public opinion in the social sciences? *Online Information Review*, 31(3), 277-289.