Modeling Twitter Hashtag Trends

May 7, 2013

Ezra Kebrab and Tristan Renaud

Twitter logo
(Credit Twitter)

Primary References

- The AWK Programming Language, Authors: Alfred V. Aho, Brian W. Kernighan, Peter J. Weinberger, 1988
- Numerical Computing, Author: Cleve Moler, 2004
- TWITTER CENSUS - CONVERSATION METRICS: ONE YEAR OF URLS, HASHTAGS, SMILEYS USAGE (BY HOUR)
- Characterizing, modeling, and generating workload spikes for stateful services, Authors: Peter Bodik, Armando Fox, Michael J. Franklin, Michael I. Jordan, David A. Patterson, 2010

Twitter logo

Additional References

- Kanye West Crashes VMA stage During Taylor Swift's Award Speech, September 13, 2009 By Jayson Rodriguez
- AIG finalizes plan to repay U.S. government, September 30, 2010 By Brady Dennis
- Rihanna Bloodied, Beaten, Bitten By Chris Brown: Reports (UPDATE), First Posted March 12, 2009, Updated May 25, 2011
- Exact Details of Michael Jackson's Death Still Unclear, June 27, 2009
- 2009 H1N1 Flu ("Swine Flu") and You, February 10, 2010
- Authorities: 'Balloon boy ' incident was a hoax, October 19, 2009 By Greg Morrison and Janet DiGiacomo

Twitter logo

Initial Question

Does hashtag popularity escalate and decline in relation to an event? Moreover, do hashtag mentions associated with an event plateau at a higher frequency than they do prior to the event?

Twitter logo

Notes: Twitter Trend Algorithm

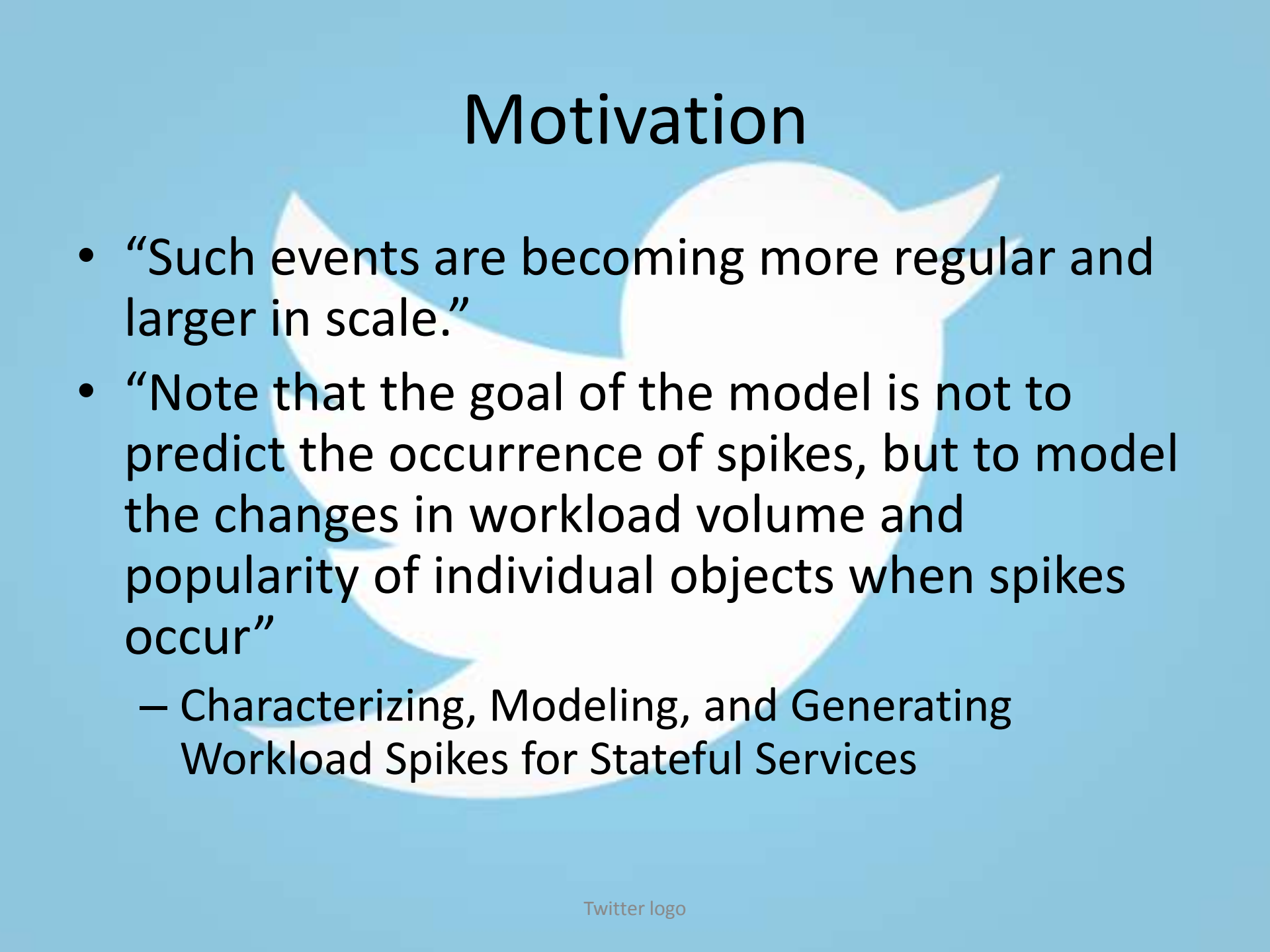
"Twitter Trends are automatically generated by an algorithm that attempts to identify topics that are being talked about more *right now* than they were previously. The Trends list is designed to help people discover the 'most breaking' breaking news from across the world, in real-time."

-Twitter blog, December 8, 2010

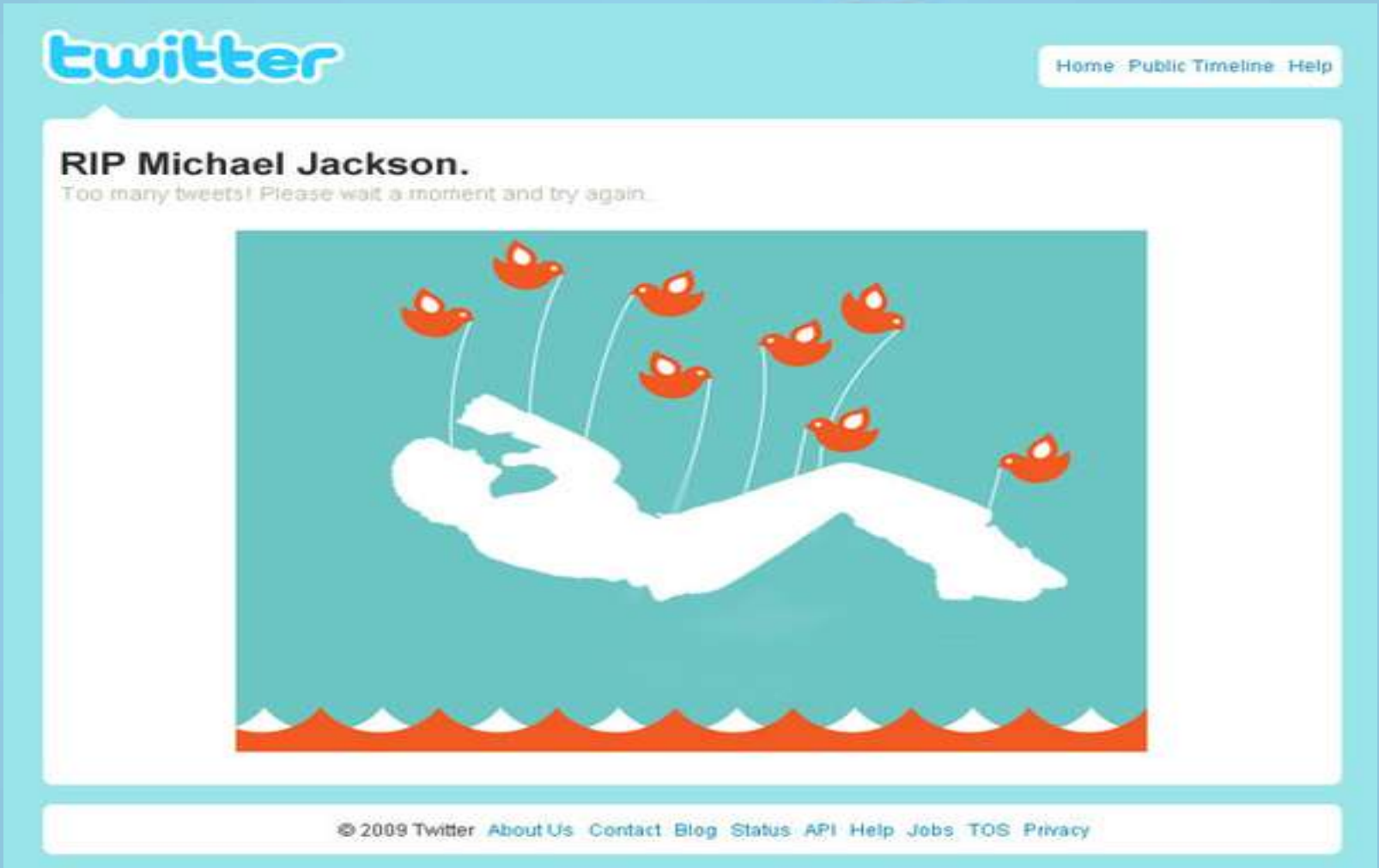
Twitter logo

Motivation

- “Such events are becoming more regular and larger in scale.”
- “Note that the goal of the model is not to predict the occurrence of spikes, but to model the changes in workload volume and popularity of individual objects when spikes occur”
  - Characterizing, Modeling, and Generating Workload Spikes for Stateful Services

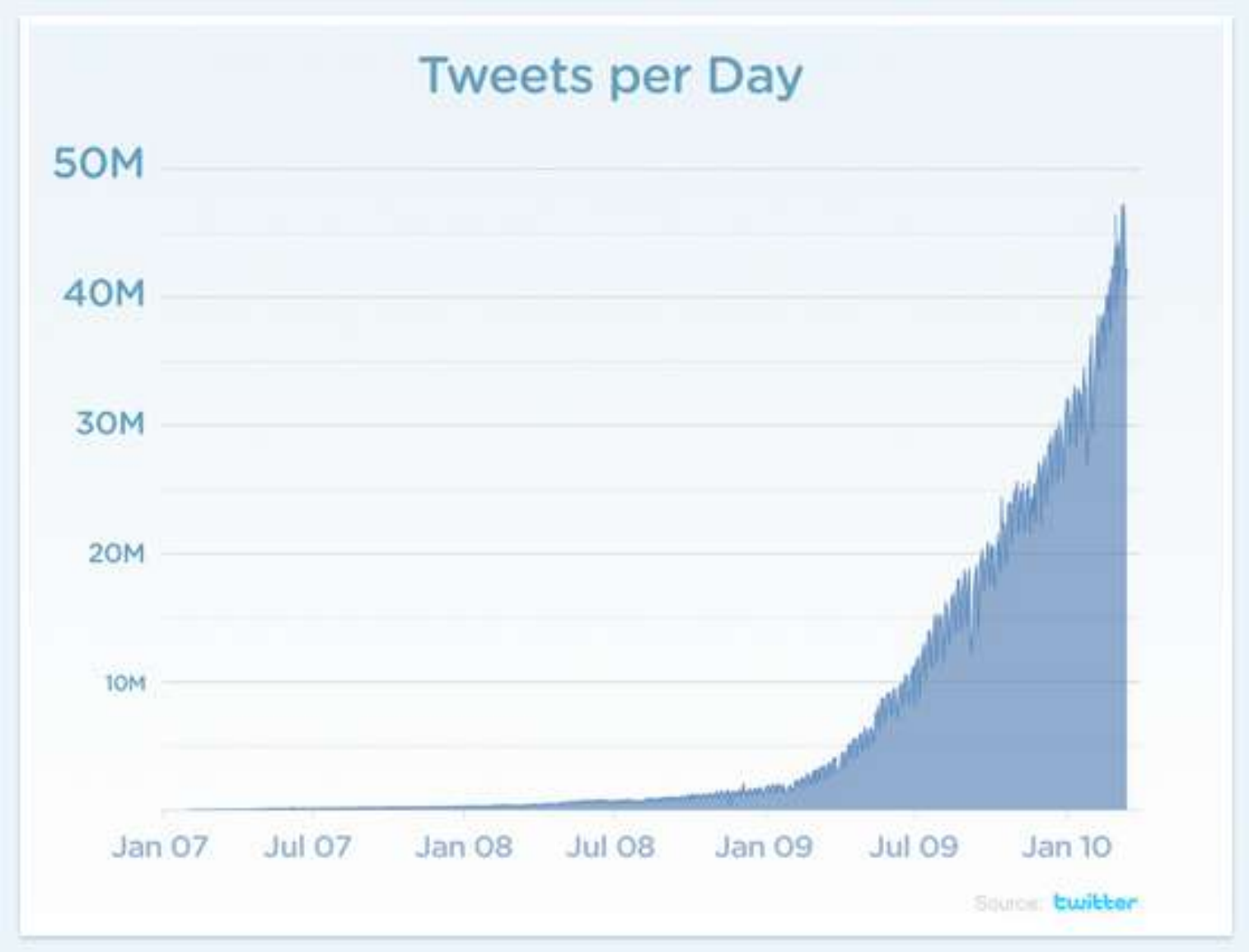


Michael Jackson's Death



http://s3-ec.buzzfed.com/static/imagebuzz/web04/2009/6/25/20/twitter-rip-michael-jackson-5253-1245974830-6.jpg

Notes: Twitter Usage



Revised Question

Is there is a general model that can fit Twitter spikes reasonably well to predict decline rates post data spikes?

Twitter logo

Events of Interest

- **Kanye West: "Ima let you finish"** - interrupts Taylor Swift during her acceptance speech for Best Female Video (September 13, 2009)
- **AIG Bailout** - US Goverment bailed out AIG w/$180BN to keep it from going bankrupt (Late 2008-Early 2009)
- **Rihanna Assault** - Chris Brown assaults Rihanna pre-VMAs (February 8, 2009)
- **Michael Jackson's Death** – dies unexpectedly from medicinal poisoning (June 25, 2009)
- **Ballon Boy Hoax** – boy thought to be in hot air balloon found hiding in attic (October 15, 2009)
- **Swine Flu** – Outbreak of a new influenza virus in the US (April 2009)

Twitter logo

Data

- Hourly hashtag counts for all hashtags
  - March 2006 – November 2009
- ~ 300 million rows
- ~ 14 GB
- AWK for data manipulation

Twitter logo

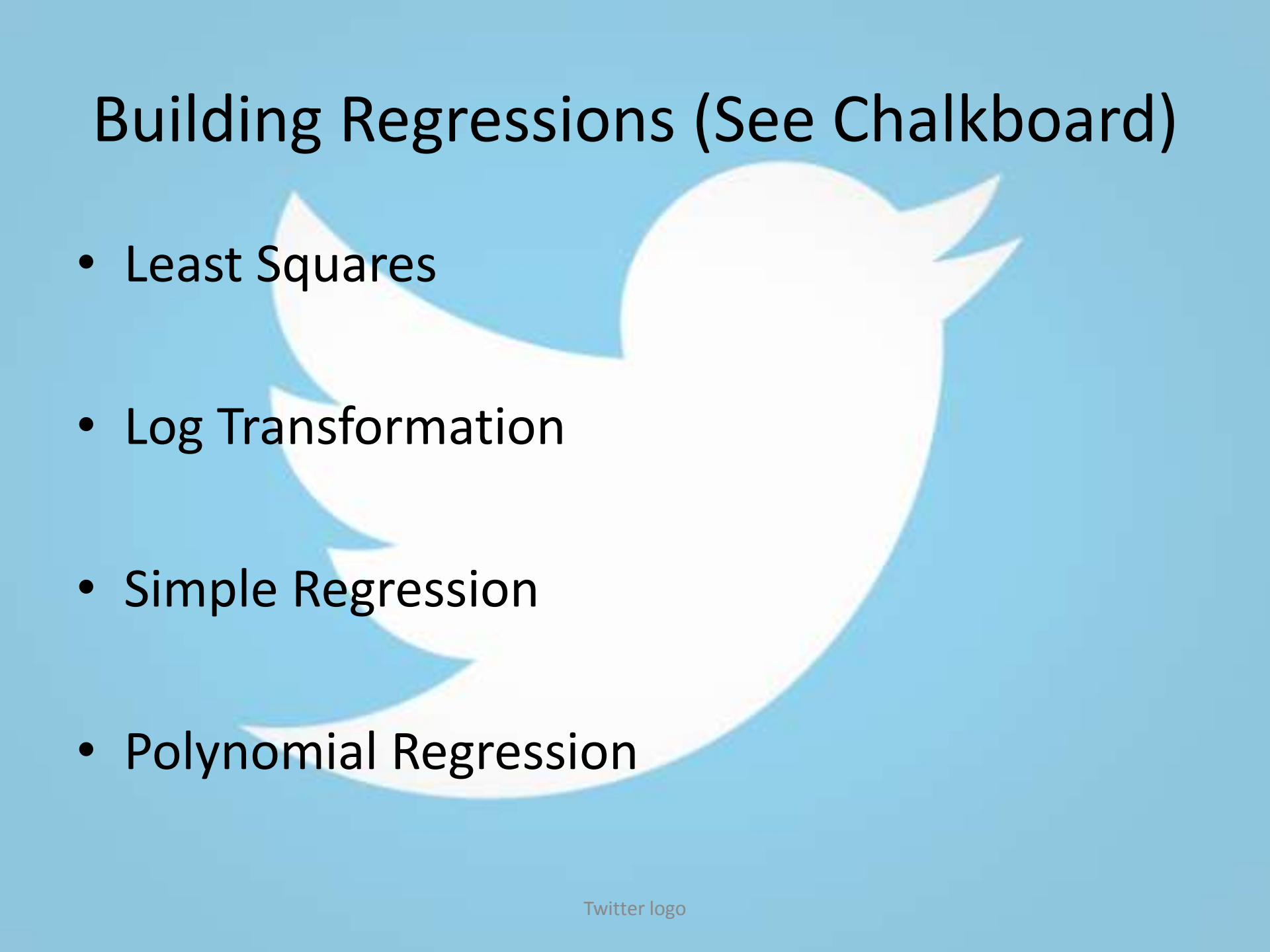
Regression Process

1. Download
2. Extract
3. Filter
4. Compile
5. Graph
6. Regression Analysis
7. Conclude

Twitter logo

Building Regressions (See Chalkboard)

- Least Squares
- Log Transformation
- Simple Regression
- Polynomial Regression



MATLAB Code (1/3)

```
1   function spikeplot(lin, pol, hourdata, t)
2
3   %Place estimated parameters into linear and trinomial models
4 -  lint = lin(2)*exp(lin(1)*t) + lin(3);
5 -  polt = pol(4)*exp(pol(3)*t + pol(2)*t.^2 + pol(1)*t.^3) + pol(5);
6
7   %Plot all together
8 -  plot(t, polt,'color','r');
9 -  hold on;
10 -  plot(t, lint,'color','g');
11 -  plot(t, hourdata,'color','b','linestyle','--');
12 -  hold off;
```

Twitter logo

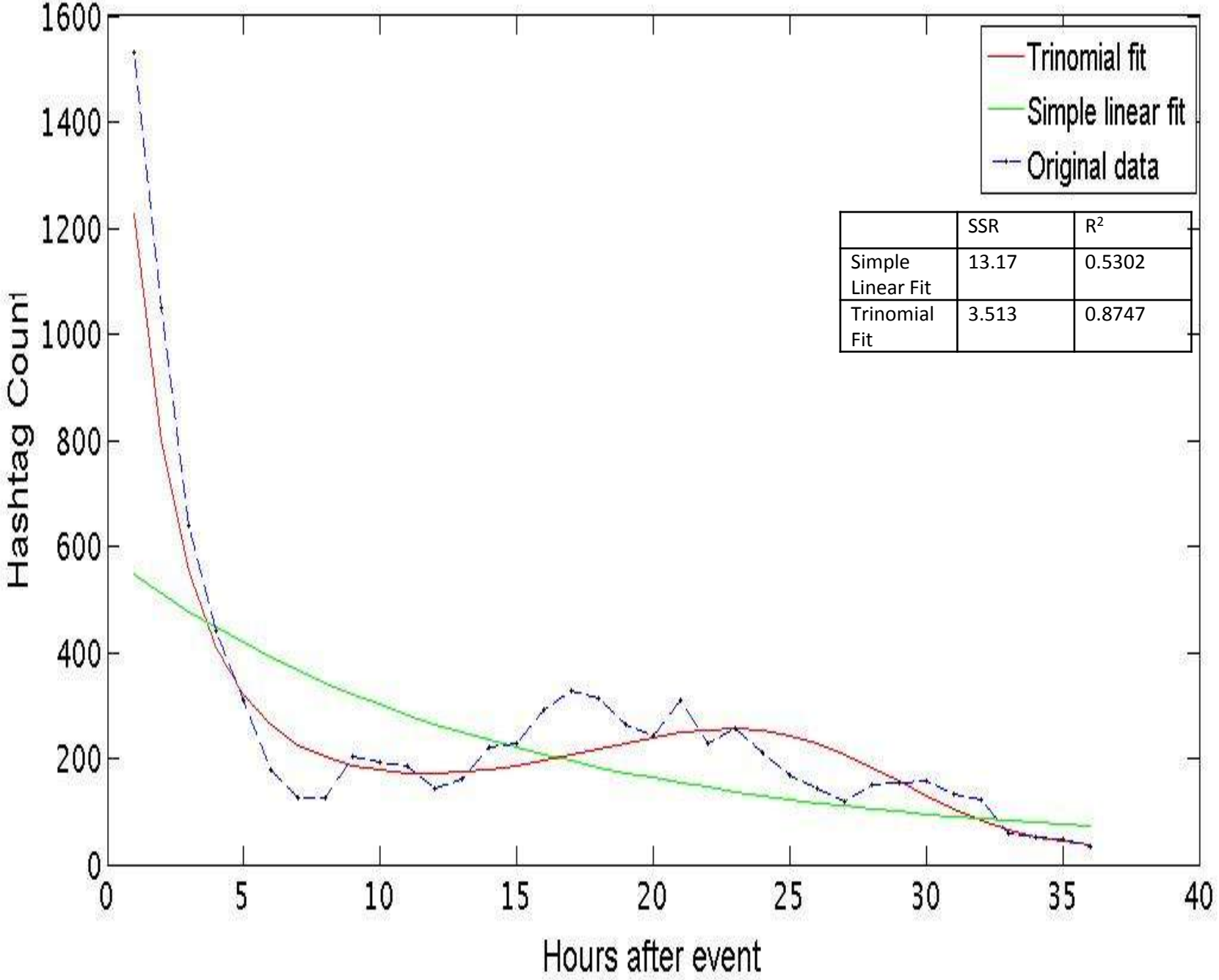
MATLAB Code (2/3)

```
function [lin pol resid t] = spikefit(hourdata, baseline)
%hourdata is a column of consecutive data points
%baseline is the observed tweet baseline

%lin model (coefficients in lin including)
% y = B2*exp(B1*t) + B3 (baseline)

%poly model (coefficients in poly including)
% y = B4*exp(B3*t + B2*t^2 + B1*t^3) + B5 (baseline)


%time array
[row column] = size(hourdata);
t = 1:row;
t = transpose(t);

%Baseline adjustment and log transformation
datafix = hourdata - baseline;
datalog = log(datafix);

%linear fit log y = log B0 + B1*t
linreg = polyfit(t, datalog, 1);
linfit = polyval(linreg, t);

%poly fit (n=3) log y = log B0 + B1*t + B2*t^2 + B3*t^3
polreg = polyfit(t, datalog, 3);
polfit = polyval(polreg, t);

%solve for original B0
linregfix = linreg;
linregfix(2) = exp(linreg(2));
polregfix = polreg;
polregfix(4) = exp(polreg(4));
```

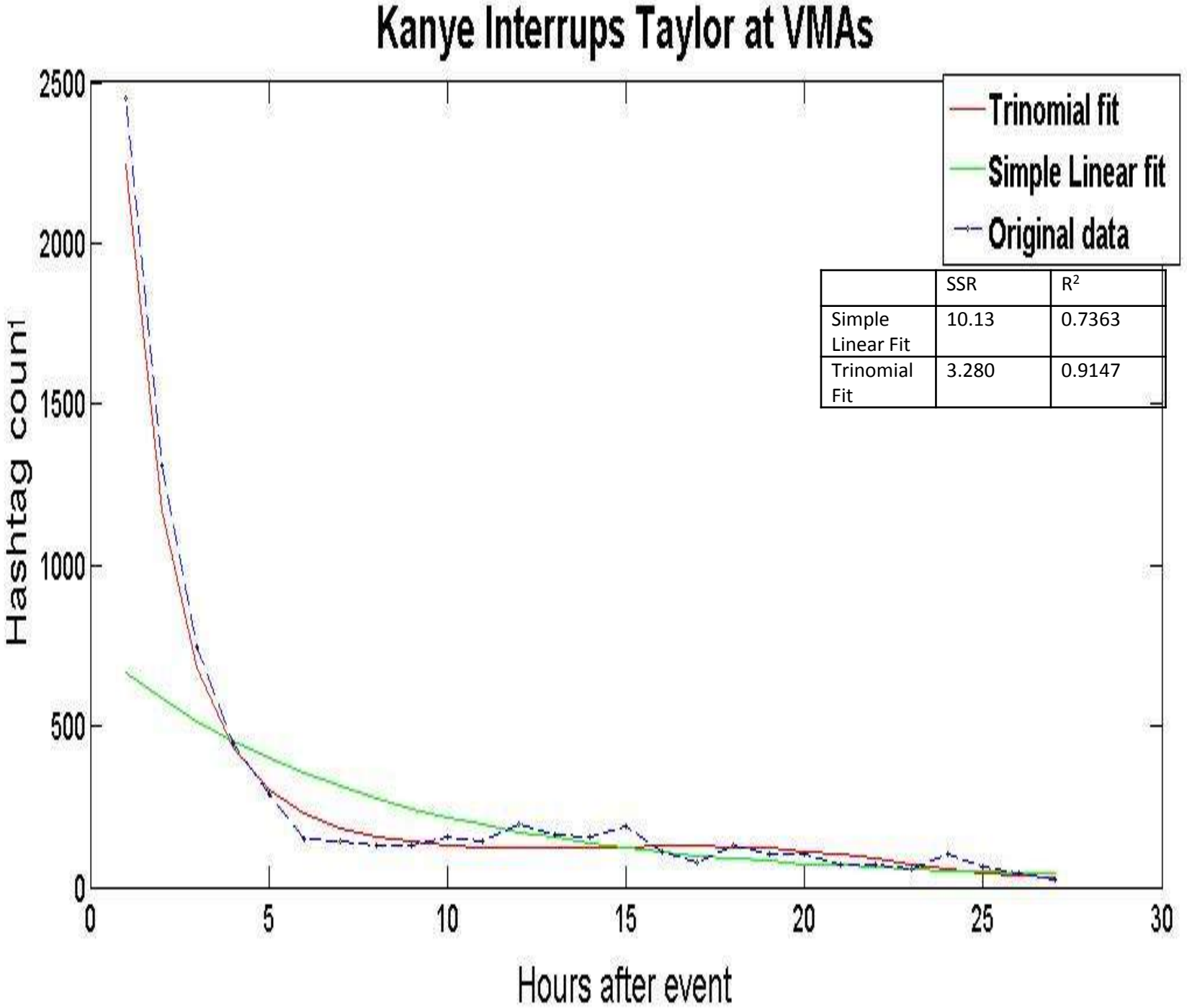
MATLAB Code (3/3)

```
35      %residual values, squared sum of residual and R^2
36 -    SStotal = (length(datalog)-1) * var(datalog);
37 -    linresid = datalog - linfit;
38 -    linSS = sum(linresid.^2);
39 -    linrsq = 1 - linSS/SStotal;
40 -    polresid = datalog - polfit;
41 -    polSS = sum(polresid.^2);
42 -    polrsq = 1 - polSS/SStotal;
43
44      %return error (SSR and R^2)
45 -    resid = ones(2,2);
46 -    resid(1,1) = linSS;
47 -    resid(2,1) = linrsq;
48 -    resid(1,2) = polSS;
49 -    resid(2,2) = polrsq;
50
51      %return parameters
52 -    lin(1:2) = linregfix(1:2);
53 -    lin(3) = baseline;
54 -    pol(1:4) = polregfix(1:4);
55 -    pol(5) = baseline;
```

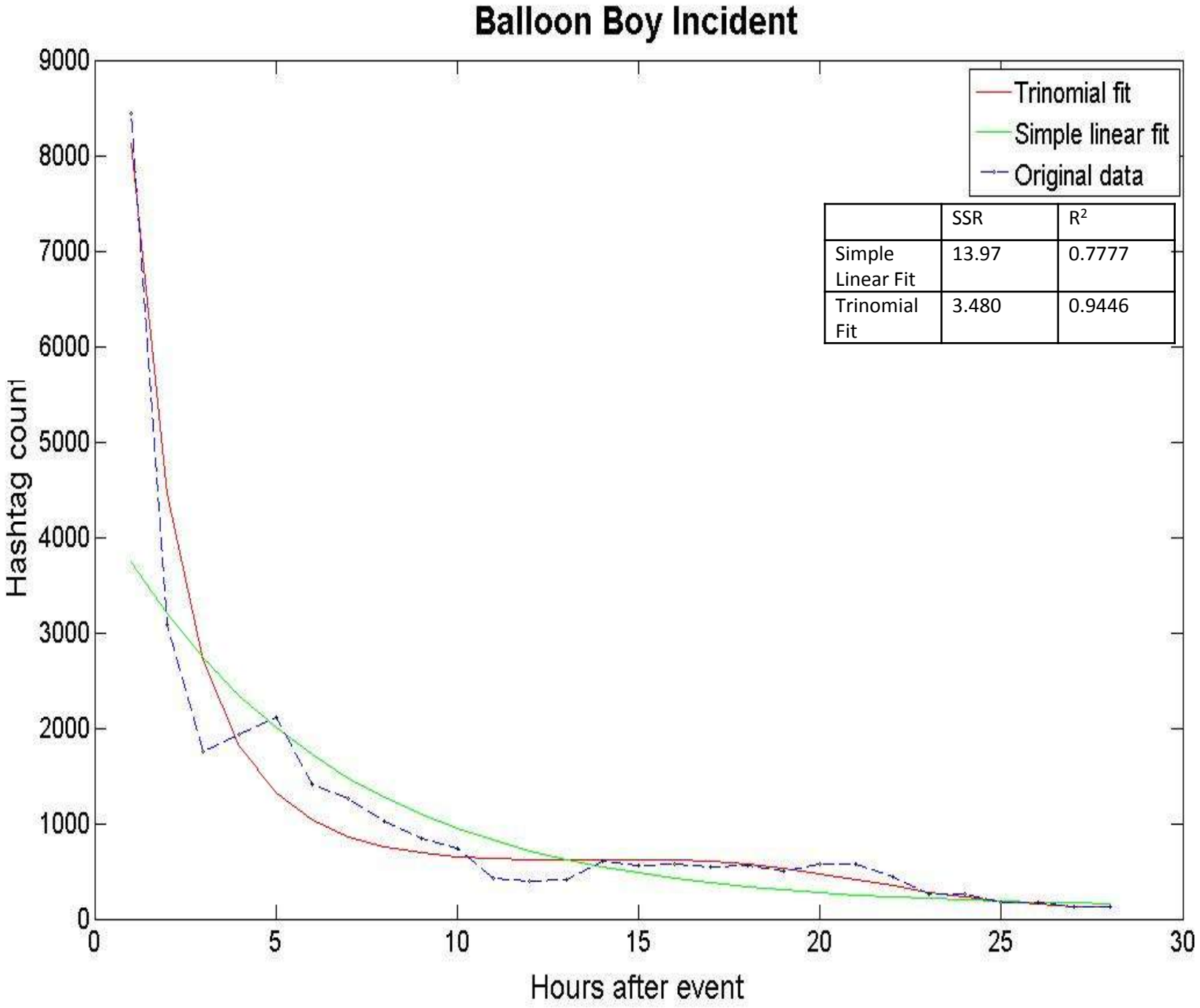
Micheal Jackson's Death



	SSR	R^2
Simple Linear Fit	13.17	0.5302
Trinomial Fit	3.513	0.8747



	SSR	R^2
Simple Linear Fit	10.13	0.7363
Trinomial Fit	3.280	0.9147

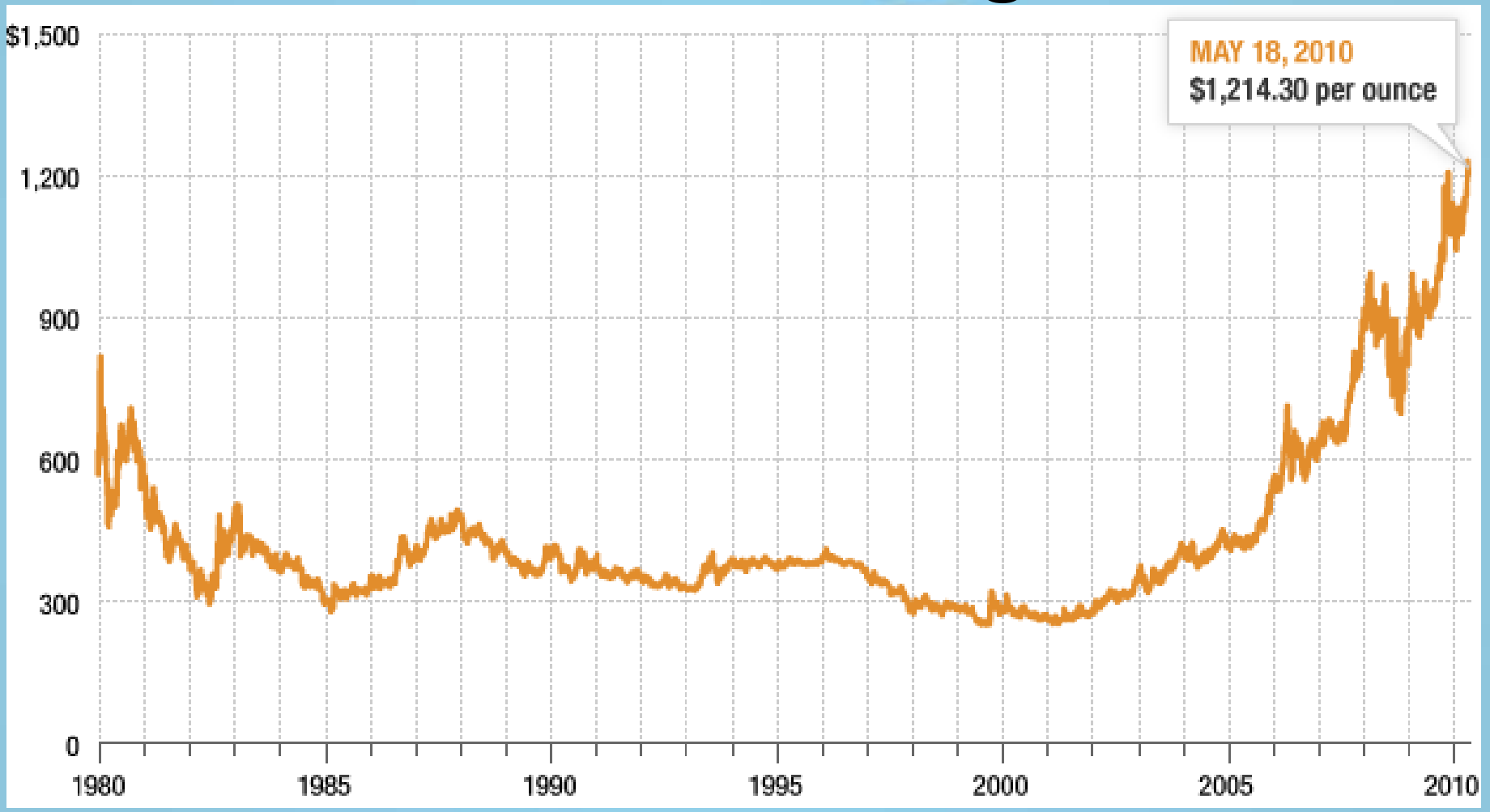


	SSR	R^2
Simple Linear Fit	13.97	0.7777
Trinomial Fit	3.480	0.9446

Applications

- Portfolio/Risk management
- Marketing
- Macroeconomic Predictions
- Server Management

Portfolio/Risk Management

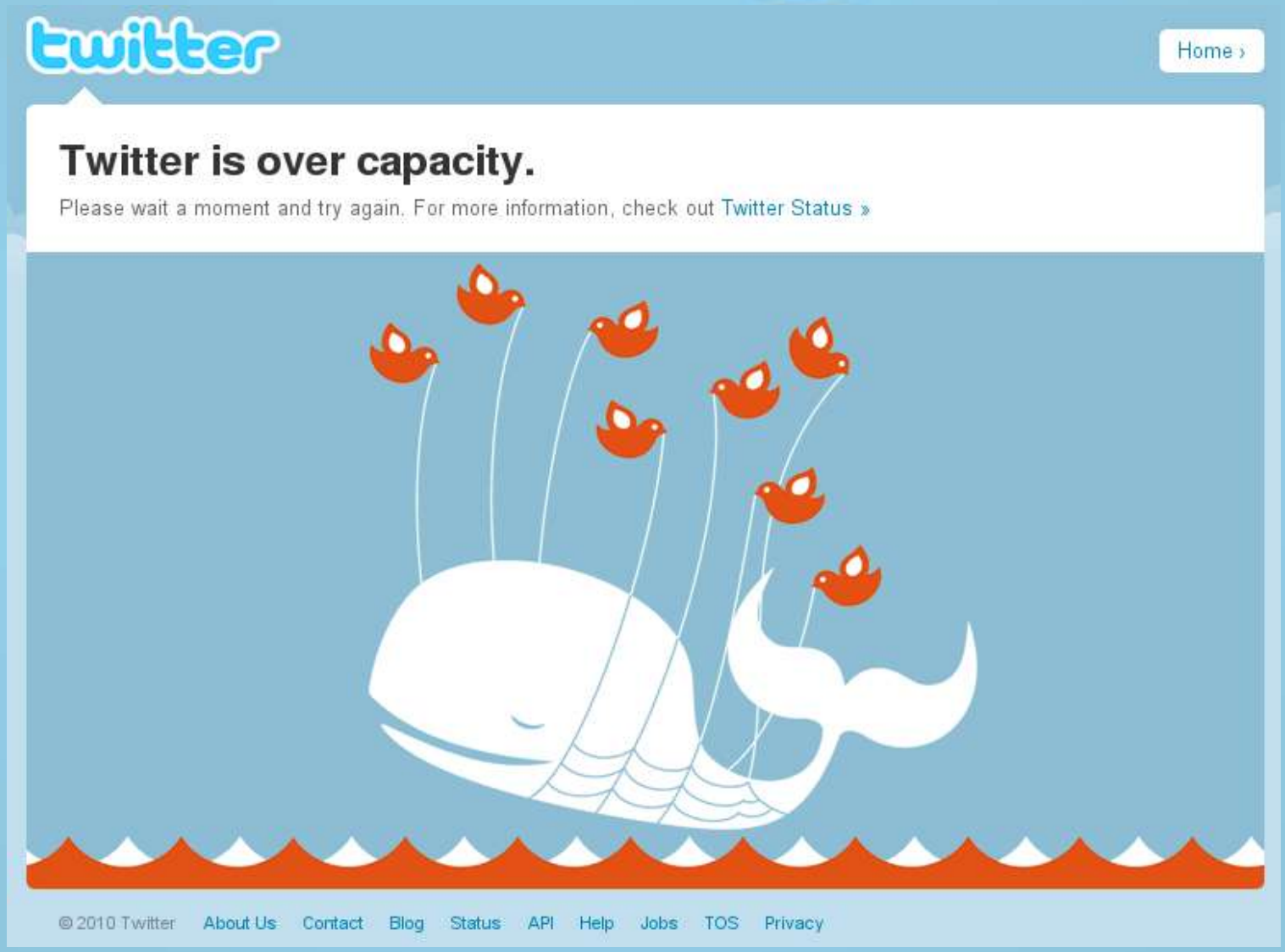


http://www.npr.org/news/graphics/2010/05/gr-gold-prices-624.gif

Marketing

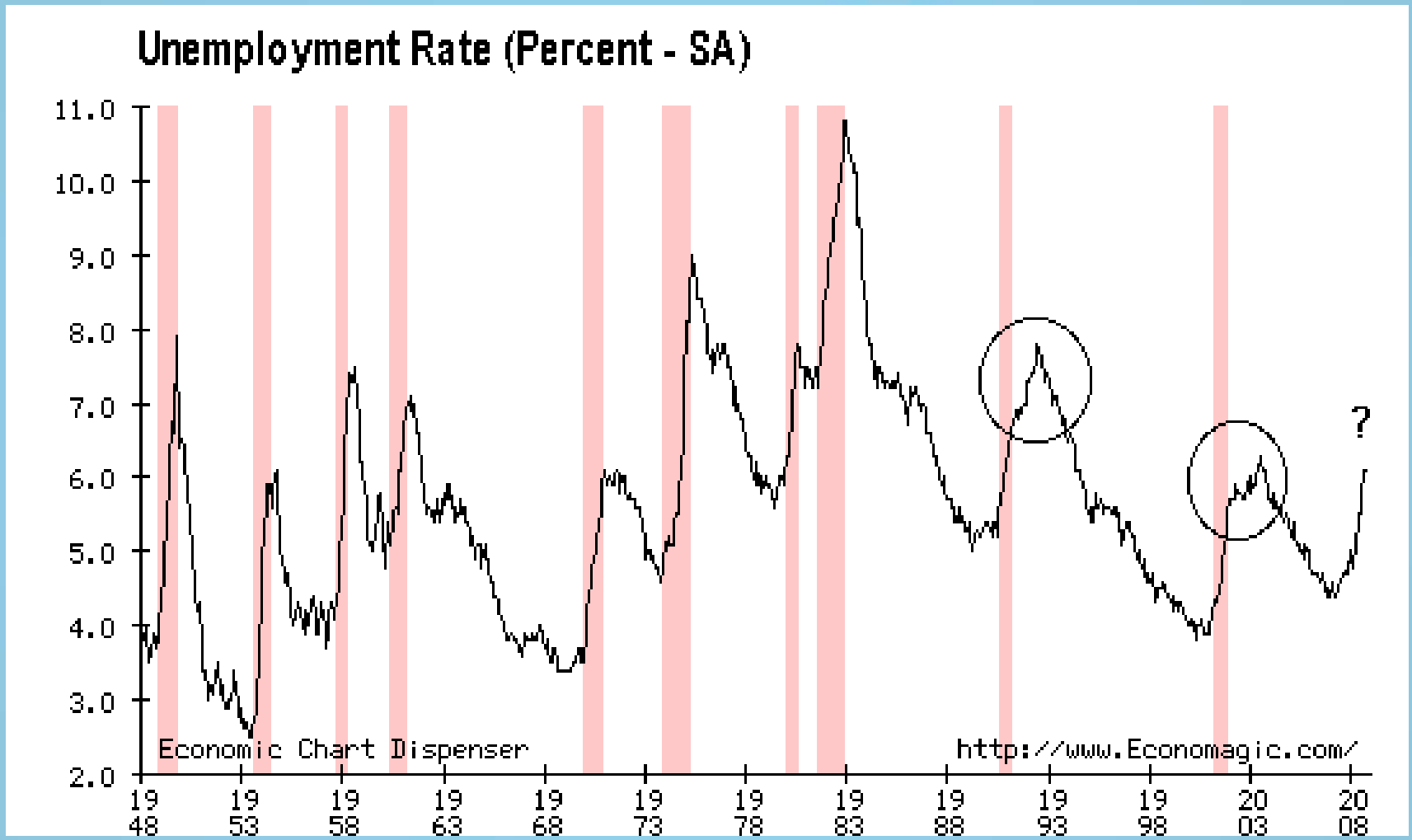
http://graphics8.nytimes.com/images/section/jobs/200703/clipart/marketing-product-jobs.jpg

Server Management



Twitter

Macroeconomic Predictions



http://www.marketoracle.co.uk/Article7189.html

Potential Errors

- Indirect sampling (hashtags only)
- Filters not related to event-related data spikes
- Per Hour data vs. Per Minute data
  - Accuracy vs. Noise
- Missing Filters, Typos, etc
- Scaling: Time, Volume

Twitter logo

Conclusion

- Findings
  - Model Effectiveness, Applications
- Restrictions
  - Costs, Equipment, Time
- Outlook
  - Future Possibilities

Twitter logo