UNIVERSITY OF AMSTERDAM

Life Mining

Maarten de Rijke

- Joint work with
  Krisztian Balog, Richard Berendsen, Matthias Breuss, Marc Bron, Katja Hofmann, Bouke Huurnink, Edgar Meij, Gilad Mishne, Andrei Oghina, Daan Odijk, Hendrike Peetz, Manos Tsagkias, Wouter Weerkamp, Shimon Whiteson

What?

- Search not as the target application, but as a core component
- What is the application?

Datamining – Wikipedia
nl.wikipedia.org/wiki/Datamining
data mining wikipedia
Datamining – Wikipedia
Niet aangemeld Overlegpagina IP-adres Bijdragen IP-adres Aanmelden Registreren
Artikel Overleg Lezen Bewerken Geschiedenis Zoeken
WIKIPEDIA
De vrije encyclopedie
Hoofdpagina
Vind een artikel
Vandaag
Etalage
Categorieën
Recente wijzigingen
Nieuwe artikelen
Willekeurige pagina
Informatie
Gebruikersportaal
Snelcursus
Hulp en contact
Donaties
Hulpmiddelen
Afdrukken/exporteren
In andere talen
العربية
Datamining
Datamining is het gericht zoeken naar (statistische) verbanden in grote verzamelingen gegevens voor wetenschappelijke of commerciële doeleinden. Zo'n verzameling gegevens kan gevormd worden door gebeurtenissen in een praktijksituatie te registreren (aankoopgedrag van consumenten, symptomen bij patiënten, et cetera) of door de resultaten van eerder uitgevoerde wetenschappelijke onderzoeken met elkaar te vergelijken en te herinterpreteren.
De naam komt voort uit de overeenkomsten tussen het zoeken naar statistische verbanden en het graven (mining) naar iets waardevols in een grote berg. Datamining helpt bedrijven en wetenschappers om de essentiële informatie uit gegevens te selecteren. Er kan een model mee gecreëerd worden dat het gedrag van mensen of systemen kan voorspellen.
Inhoud [weergeven]
Misbruik [bewerken]
Zie Misbruik van statistische gegevens voor het hoofdartikel over dit onderwerp.
Een valkuil die bij datamining op de loer ligt is de drogreden Cum hoc ergo propter hoc: als je maar genoeg gegevens analyseert zul je vroeg of laat ongetwijfeld een statistische correlatie tussen twee variabelen vinden, maar dat hoeft niet te betekenen dat er ook een oorzakelijk verband bestaat tussen de twee betreffende variabelen.

s a core

Datamining - Wikipedia

nl.wikipedia.org/wiki/Datamining

data mining wikipedia

Text mining - Wikipedia, the free encyclopedia

en.wikipedia.org/wiki/Text_mining

text mining wikipedia

Text mining - Wikipedia, the fre...

Log in Create account

Article Talk Read Edit View history Search

Text mining

From Wikipedia, the free encyclopedia

It has been suggested that this article or section be merged with *Text analytics*. (Discuss) *Proposed since June 2012.*

Text mining, sometimes alternately referred to as *text data mining*, roughly equivalent to *text analytics*, refers to the process of deriving high-quality information from text. High-quality information is typically derived through the devising of patterns and trends through means such as statistical pattern learning. Text mining usually involves the process of structuring the input text (usually parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent insertion into a database), deriving patterns within the structured data, and finally evaluation and interpretation of the output. 'High quality' in text mining usually refers to some combination of relevance, novelty, and interestingness. Typical text mining tasks include text categorization, text clustering, concept/entity extraction, production of granular taxonomies, sentiment analysis, document summarization, and entity relation modeling (*i.e.*, learning relations between named entities).

Contents [show]

History [edit]

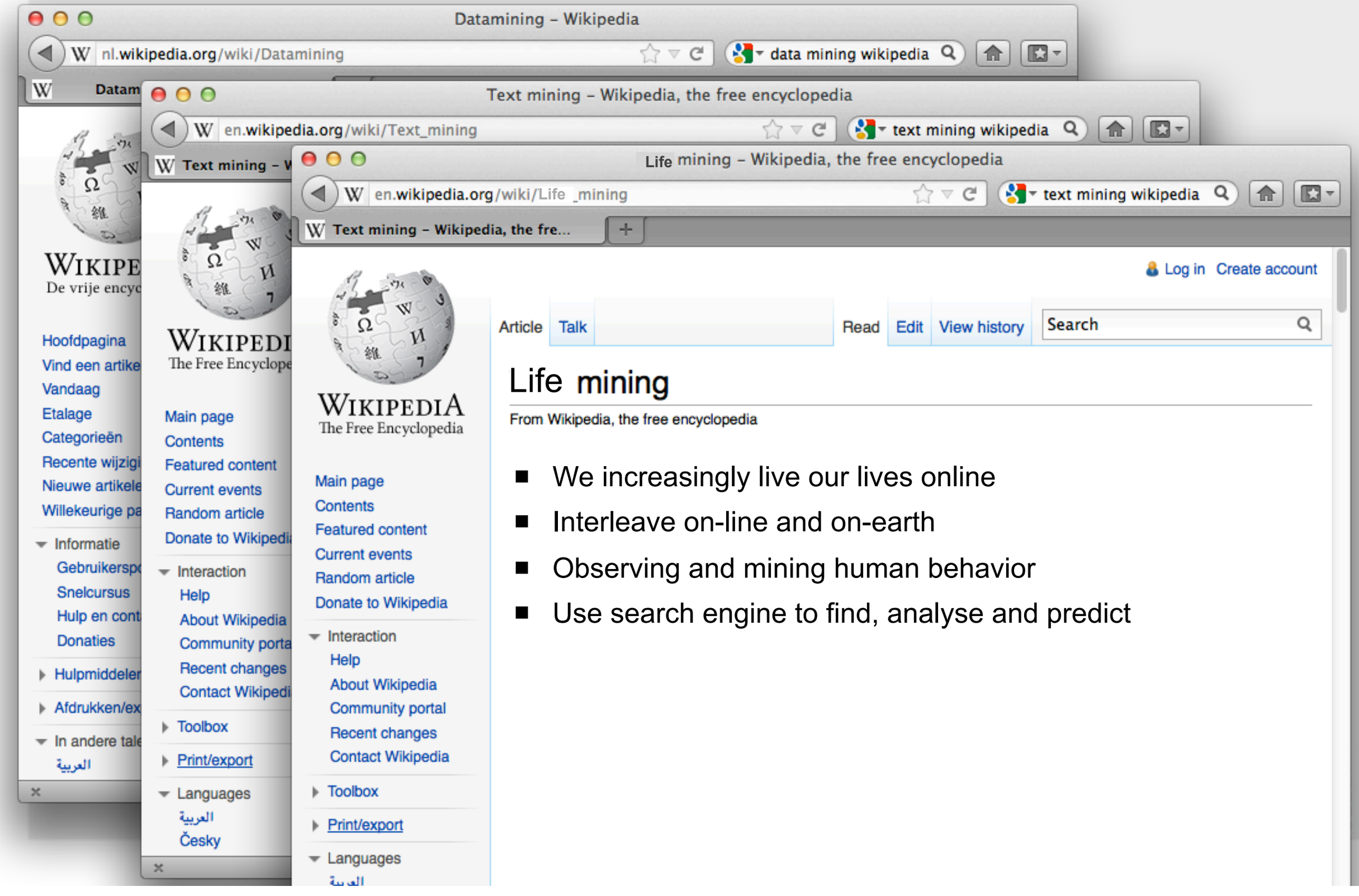
Labor-intensive manual text mining approaches first surfaced in the mid-1980s,[*citation needed*][*examples needed*] but technological advances have enabled the field to advance during the past decade. Text mining is an interdisciplinary field that draws on information retrieval, data mining, machine learning, statistics, and

Hoofdpagina
Vind een artike
Vandaag
Etalage
Categorieën
Recente wijzigi
Nieuwe artikele
Willekeurige pa
Informatie
Gebruikersp
Snelcursus
Hulp en cont
Donaties
Hulpmiddeler
Afdrukken/ex
In andere tal
العربية

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact Wikipedia
Toolbox
Print/export
Languages
العربية
Česky

re

4



- We increasingly live our lives online
- Interleave on-line and on-earth
- Observing and mining human behavior
- Use search engine to find, analyse and predict

Can we mine useful knowledge from digital traces of people's lives?

In the past

AOL-user-ct-collection — less — 116×53

```
710766  wwwpeoplesearch.comwww.reviewplace.seardh       2006-05-30 22:10:13
710766  wwwpeoplesearch.comwww.reviewplace.seardh       2006-05-30 22:10:33
711391  can not sleep with snoring husband      2006-03-01 01:24:00
711391  cannot sleep with snoring husband       2006-03-01 01:24:07     9       http://www.wjla.com
711391  cannot sleep with snoring husband       2006-03-01 01:24:07     9       http://www.wjla.com
711391  cannot sleep with snoring husband       2006-03-01 01:33:06     1       http://www.epinions.com
711391  jackie zeaman nude      2006-03-01 15:26:27
711391  jackie zeman nude       2006-03-01 15:26:38
711391  strange cosmos  2006-03-01 16:07:15     1       http://www.strangecosmos.com
711391  mansfield first assembly        2006-03-01 16:09:20     1       http://www.mansfieldfirstassembly.org
711391  mansfield first assembly        2006-03-01 16:09:20     3       http://netministries.org
711391  reverend harry myers    2006-03-01 16:10:07
711391  reverend harry myers    2006-03-01 16:10:30
711391  national enquirer       2006-03-01 17:13:14     1       http://www.nationalenquirer.com
711391  how to kill mockingbirds        2006-03-01 17:18:11
711391  how to kill mockingbirds        2006-03-01 17:18:33
711391  how to kill annoying birds in your yards        2006-03-01 17:18:58
711391  how to kill annoying birds in your yards        2006-03-01 17:19:53     2       http://www.sortprice.com
711391  how to rid your yard of noisy annoying birds    2006-03-01 17:23:08     3       http://shopping.msn.com
711391  how to rid your yard of noisy annoying birds    2006-03-01 17:23:08     10      http://www.bergen.org
711391  how to rid your yard of noisy annoying birds    2006-03-01 17:24:35     15      http://www.saferbrand.com
711391  how do i get mocking birds out of my yard       2006-03-01 17:27:17
711391  how do i get mockingbirds out of my yard        2006-03-01 17:27:36     9       http://www.asri.org
711391  how do i get mockingbirds out of my yard        2006-03-01 17:30:14
711391  how to get rid of noisy loud birds      2006-03-01 17:30:52     3       http://www.bird-x.com
711391  how to get rid of noisy loud birds      2006-03-01 17:30:52     1       http://forums2.gardenweb.com
711391  how to get rid of noisy loud birds      2006-03-01 17:30:52     10      http://www.birding.com
711391  mansfield first assembly        2006-03-01 18:31:36     3       http://netministries.org
711391  beth moore      2006-03-01 19:42:41     1       http://www.lproof.org
711391  judy baker ministries   2006-03-01 19:49:03     2       http://www.embracinggrace.com
711391  god will fulfill your hearts desires    2006-03-01 19:59:06     10      http://www.pureintimacy.org
711391  online friendships can be very special  2006-03-01 23:09:37
711391  online friendships can be very special  2006-03-01 23:09:57
711391  online friendships      2006-03-01 23:10:24
711391  cypress fairbanks isd   2006-03-02 07:56:53     1       http://www.cfisd.net
711391  people are not always how they seem over the internet   2006-03-02 08:31:51
711391  friends online can be different in person       2006-03-02 08:32:42
711391  friends online can be different in person       2006-03-02 08:33:04     13      http://www.salon.com
711391  boston butts    2006-03-02 09:47:36
711391  community christian church houston tx   2006-03-02 16:07:53
```

```
710766  wwwpeoplesearch.comwww.reviewplace.seardh       2006-05-30 22:10:33
711391  can not sleep with snoring husband      2006-03-01 01:24:00
711391  cannot sleep with snoring husband       2006-03-01 01:24:07     9       http://www.wjla.com
711391  cannot sleep with snoring husband       2006-03-01 01:24:07     9       http://www.wjla.com
711391  cannot sleep with snoring husband       2006-03-01 01:33:06     1       http://www.epinions.com
711391  jackie zeaman nude      2006-03-01 15:26:27
711391  jackie zeman nude       2006-03-01 15:26:38
711391  strange cosmos  2006-03-01 16:07:15     1       http://www.strangecosmos.com
711391  mansfield first assembly        2006-03-01 16:09:20     1       http://www.mansfieldfirstassembly.org
711391  mansfield first assembly        2006-03-01 16:09:20     3       http://netministries.org
711391  reverend harry myers    2006-03-01 16:10:07
711391  reverend harry myers    2006-03-01 16:10:30
711391  national enquirer       2006-03-01 17:13:14     1       http://www.nationalenquirer.com
711391  how to kill mockingbirds        2006-03-01 17:18:11
711391  how to kill mockingbirds        2006-03-01 17:18:33
711391  how to kill annoying birds in your yards        2006-03-01 17:18:58
711391  how to kill annoying birds in your yards        2006-03-01 17:19:53     2       http://www.sortprice.com
711391  how to rid your yard of noisy annoying birds    2006-03-01 17:23:08     3       http://shopping.msn.com
711391  how to rid your yard of noisy annoying birds    2006-03-01 17:23:08     10      http://www.bergen.org
711391  how to rid your yard of noisy annoying birds    2006-03-01 17:24:35     15      http://www.saferbrand.com
711391  how do i get mocking birds out of my yard       2006-03-01 17:27:17
711391  how do i get mockingbirds out of my yard        2006-03-01 17:27:36     9       http://www.asri.org
711391  how do i get mockingbirds out of my yard        2006-03-01 17:30:14
711391  how to get rid of noisy loud birds      2006-03-01 17:30:52     3       http://www.bird-x.com
711391  how to get rid of noisy loud birds      2006-03-01 17:30:52     1       http://forums2.gardenweb.com
711391  how to get rid of noisy loud birds      2006-03-01 17:30:52     10      http://www.birding.com
711391  mansfield first assembly        2006-03-01 18:31:36     3       http://netministries.org
711391  beth moore      2006-03-01 19:42:41     1       http://www.lproof.org
711391  judy baker ministries   2006-03-01 19:49:03     2       http://www.embracinggrace.com
711391  god will fulfill your hearts desires    2006-03-01 19:59:06     10      http://www.pureintimacy.org
711391  online friendships can be very special  2006-03-01 23:09:37
711391  online friendships can be very special  2006-03-01 23:09:57
711391  online friendships      2006-03-01 23:10:24
711391  cypress fairbanks isd   2006-03-02 07:56:53     1       http://www.cfisd.net
711391  people are not always how they seem over the internet   2006-03-02 08:31:51
711391  friends online can be different in person       2006-03-02 08:32:42
711391  friends online can be different in person       2006-03-02 08:33:04     13      http://www.salon.com
711391  boston butts    2006-03-02 09:47:36
711391  community christian church houston tx   2006-03-02 16:07:53
711391  gay churches in houston tx      2006-03-02 16:08:23
711391  community gospel church in houston tx   2006-03-02 16:08:45     2       http://www.communitygospel.org
711391  houston tx is one hot place     2006-03-02 18:04:44
711391  houston tx is one hot place to live     2006-03-02 18:04:55     9       http://travel.yahoo.com
711391  houston tx is one hot place to live     2006-03-02 18:16:05     1       http://www.houston-texas-online.com
711391  texas hill country and sights around san antonio tx     2006-03-02 18:19:00     5       http://www.answers.c
om
711391  can liver problems cause you to loose your hair 2006-03-02 18:27:04
711391  can liver problems cause you to loose your hair 2006-03-02 18:27:30     1       http://www.askdoctrish.com
```

Open

Open

timeu.se: What Do People Do All Day?

timeu.se

timeu.se: What Do People Do All...

Google

timeu.se: What Do People Do All Day? [home | about | examples]

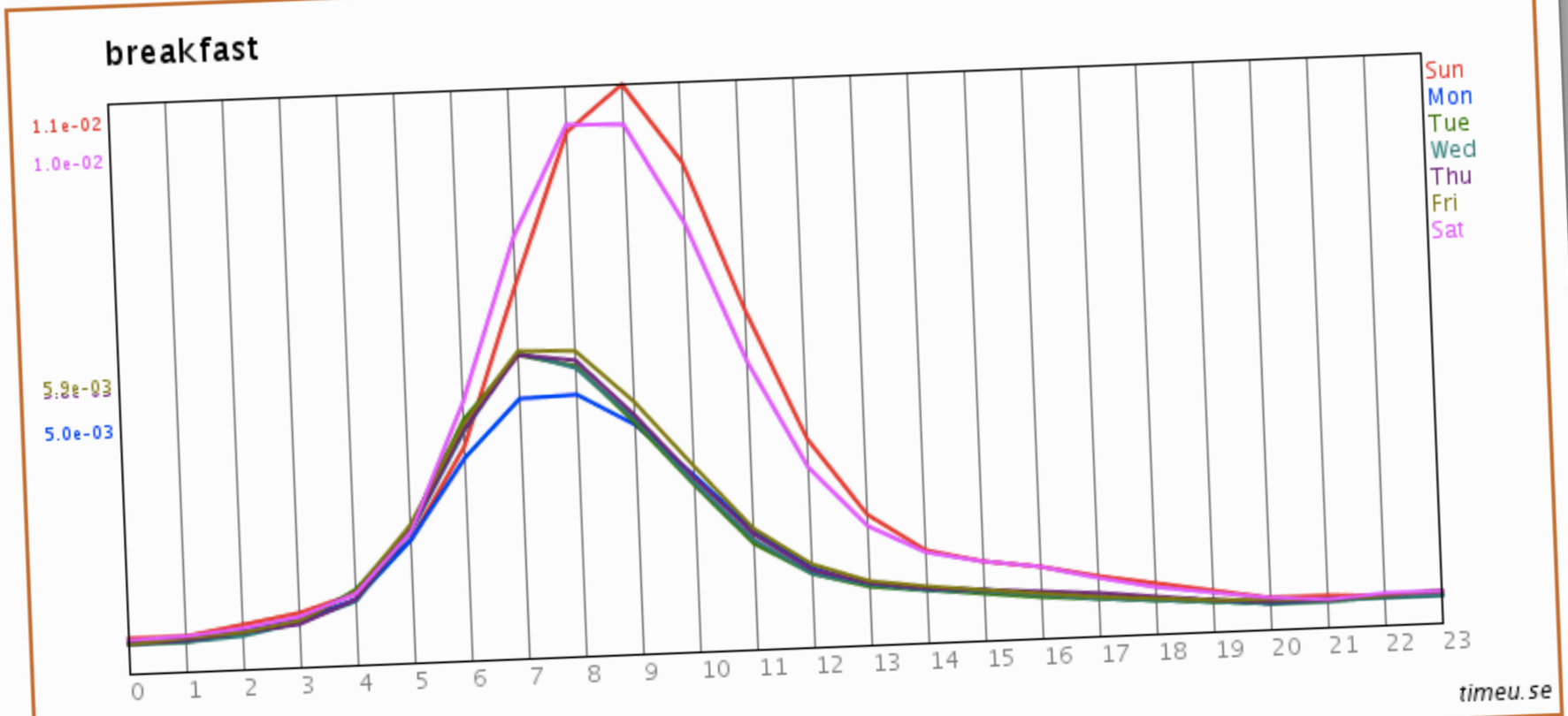
What Do People Do All Day? Let's find out! By collecting millions of messages from Twitter, we can explore in great detail exactly how people report spending their time. *learn more*

(1) Enter query words: breakfast *help*

(2) Choose a plot type:
Mouse-over each plot type for details.
more info

(3) Submit Query

Tweet This! *Email This!* **Permalink:** *http://timeu.se/a840741d*

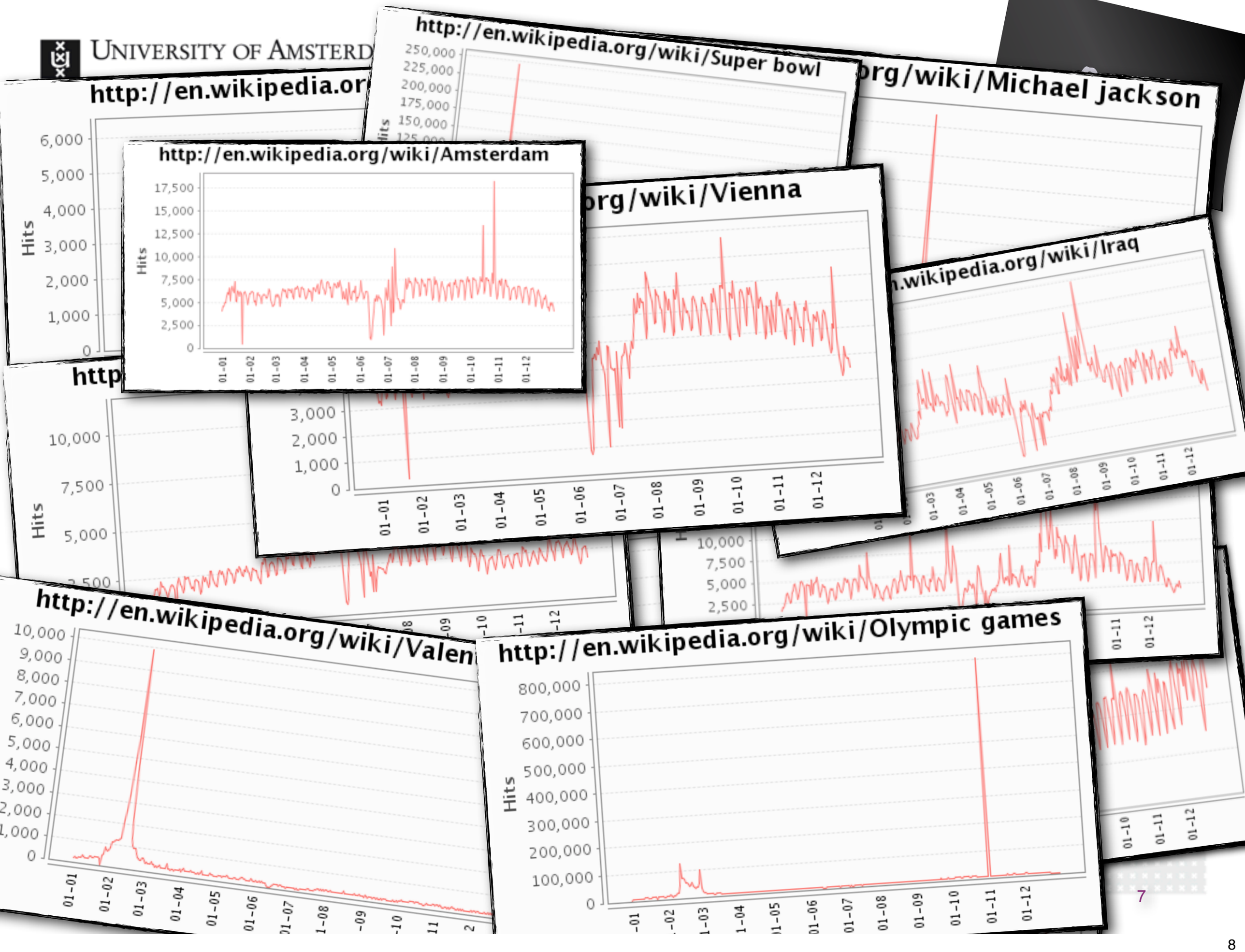


timeu.se *was created by Scott Golder, a graduate student in the Social Dynamics Lab in the Department of Sociology at Cornell University. On Twitter, he is @redlog.*

→ Like the website? ***Read the paper!*** *S. Golder, M. Macy. Diurnal and Seasonal Mood Vary with Work, Sleep and Daylength Across*

7

Open


UNIVERSITY OF AMSTERD
http://en.wikipedia.org/wiki/Super bowl
http://en.wikipedia.org/wiki/Amsterdam
http://en.wikipedia.org/wiki/Michael jackson
http://en.wikipedia.org/wiki/Vienna
http://en.wikipedia.org/wiki/Iraq
http://en.wikipedia.org/wiki/Olympic games
Hits
01-01
01-02
01-03
01-04
01-05
01-06
01-07
01-08
01-09
01-10
01-11
01-12

Long trails

Table II. Queries Correlated with "Mortgage" Over Time (These change dramatically as the query is farther away in time. The users' interests move from mortgage basics to property searching, to insurance and taxes, to furnishings, to pools and patios. Here we give the top 40 terms that did not show up in the previous time period).

Time Period				
0–30 min	1–7 days	7–30d	30–90d	90–365d
mortgage	realtors	llc	kohls	patio
mortage	owner	associates	bath	harbor
mortgage	homes	insurance	overstock	outdoor
calculator	mls	lowes	barn	replacement
mortgages	remax	notary	sears	pools
lenders	property	depot	linens	hampton
calculators	financial	savings	beyond	lawn
countrywide	appraisers	construction	kmart	enterprise
gmac	builders	condo	pottery	ymca
refinance	prudential	business	walmart	vehicle
rates	zillow	secretary	outlet	supply
interest	bankruptcy	furniture	costco	resorts
broker	real	allstate	target	lake
lending	keller	companies	pier	rv
lender	properties	contractors	bed	walgreens
payment	agreement	cost	grill	newport
loan	appraisals	reverse	kitchen	lumber
amro	residential	federal	shield	oak
emc	lease	sale	macys	authority
brokers	county	housing	vacations	concrete
abn	modular	assessors	southwest	vehicles
amortization	attorney	irs	chamber	chrysler

M. Richardson, Learning about the world through long-term query behavior. ACM TWeb 2(4), 2008

Long trails

Table II. Queries Correlated with "Mortgage" Over Time (These change dramatically as the query is farther away in time. The users' interests move from mortgage basics to property searching, to insurance and taxes, to furnishings, to pools and patios. Here we give the top 40 terms that did not show up in the previous time period).

Time Period				
0–30 min	1–7 days	7–30d	30–90d	90–365d
mortgage	realtors	llc	kohls	patio
mortage	owner	associates	bath	harbor
mortgage	homes	insurance	overstock	outdoor
calculator	mls	lowes	barn	replacement
mortgages	remax	notary	sears	pools
lenders	property	depot	linens	hampton
calculators	financial	savings	beyond	lawn
countrywide	appraisers	construction	kmart	enterprise
gmac	builders	condo	pottery	ymca
refinance	prudential	business	walmart	vehicle
rates	zillow	secretary	outlet	supply
interest	bankruptcy	furniture	costco	resorts
broker	real	allstate	target	lake
lending	keller	companies	pier	rv
lender	properties	contractors	bed	walgreens
payment	agreement	cost	grill	newport
loan	appraisals	reverse	kitchen	lumber
amro	residential	federal	shield	oak
emc	lease	sale	macys	authority
brokers	county	housing	vacations	concrete
abn	modular	assessors	southwest	vehicles
amortization	attorney	irs	chamber	chrysler

M. Richardson, Learning about the world through long-term query behavior. ACM TWeb 2(4), 2008

Long trails

Table II. Queries Correlated with "Mortgage" Over Time (These change dramatically as the query is farther away in time. The users' interests move from mortgage basics to property searching, to insurance and taxes, to furnishings, to pools and patios. Here we give the top 40 terms that did not show up in the previous time period).

Time Period				
0–30 min	1–7 days	7–30d	30–90d	90–365d
mortgage	realtors	llc	kohls	patio
mortage	owner	associates	bath	harbor
mortgage	homes	insurance	overstock	outdoor
calculator	mls	lowes	barn	replacement
mortgages	remax	notary	sears	pools
lenders	property	depot	linens	hampton
calculators	financial	savings	beyond	lawn
countrywide	appraisers	construction	kmart	enterprise
gmac	builders	condo	pottery	ymca
refinance	prudential	business	walmart	vehicle
rates	zillow	secretary	outlet	supply
interest	bankruptcy	furniture	costco	resorts
broker	real	allstate	target	lake
lending	keller	companies	pier	rv
lender	properties	contractors	bed	walgreens
payment	agreement	cost	grill	newport
loan	appraisals	reverse	kitchen	lumber
amro	residential	federal	shield	oak
emc	lease	sale	macys	authority
brokers	county	housing	vacations	concrete
abn	modular	assessors	southwest	vehicles
amortization	attorney	irs	chamber	chrysler

M. Richardson, Learning about the world through long-term query behavior. ACM TWeb 2(4), 2008

Long trails

Table II. Queries Correlated with "Mortgage" Over Time (These change dramatically as the query is farther away in time. The users' interests move from mortgage basics to property searching, to insurance and taxes, to furnishings, to pools and patios. Here we give the top 40 terms that did not show up in the previous time period).

Time Period				
0–30 min	1–7 days	7–30d	30–90d	90–365d
mortgage	realtors	llc	kohls	patio
mortage	owner	associates	bath	harbor
mortgage	homes	insurance	overstock	outdoor
calculator	mls	lowes	barn	replacement
mortgages	remax	notary	sears	pools
lenders	property	depot	linens	hampton
calculators	financial	savings	beyond	lawn
countrywide	appraisers	construction	kmart	enterprise
gmac	builders	condo	pottery	ymca
refinance	prudential	business	walmart	vehicle
rates	zillow	secretary	outlet	supply
interest	bankruptcy	furniture	costco	resorts
broker	real	allstate	target	lake
lending	keller	companies	pier	rv
lender	properties	contractors	bed	walgreens
payment	agreement	cost	grill	newport
loan	appraisals	reverse	kitchen	lumber
amro	residential	federal	shield	oak
emc	lease	sale	macys	authority
brokers	county	housing	vacations	concrete
abn	modular	assessors	southwest	vehicles
amortization	attorney	irs	chamber	chrysler

M. Richardson, Learning about the world through long-term query behavior. ACM TWeb 2(4), 2008

Long trails

Table II. Queries Correlated with "Mortgage" Over Time (These change dramatically as the query is farther away in time. The users' interests move from mortgage basics to property searching, to insurance and taxes, to furnishings, to pools and patios. Here we give the top 40 terms that did not show up in the previous time period).

Time Period				
0–30 min	1–7 days	7–30d	30–90d	90–365d
mortgage	realtors	llc	kohls	patio
mortage	owner	associates	bath	harbor
mortgage	homes	insurance	overstock	outdoor
calculator	mls	lowes	barn	replacement
mortgages	remax	notary	sears	pools
lenders	property	depot	linens	hampton
calculators	financial	savings	beyond	lawn
countrywide	appraisers	construction	kmart	enterprise
gmac	builders	condo	pottery	ymca
refinance	prudential	business	walmart	vehicle
rates	zillow	secretary	outlet	supply
interest	bankruptcy	furniture	costco	resorts
broker	real	allstate	target	lake
lending	keller	companies	pier	rv
lender	properties	contractors	bed	walgreens
payment	agreement	cost	grill	newport
loan	appraisals	reverse	kitchen	lumber
amro	residential	federal	shield	oak
emc	lease	sale	macys	authority
brokers	county	housing	vacations	concrete
abn	modular	assessors	southwest	vehicles
amortization	attorney	irs	chamber	chrysler

M. Richardson, Learning about the world through long-term query behavior. ACM TWeb 2(4), 2008

Long trails

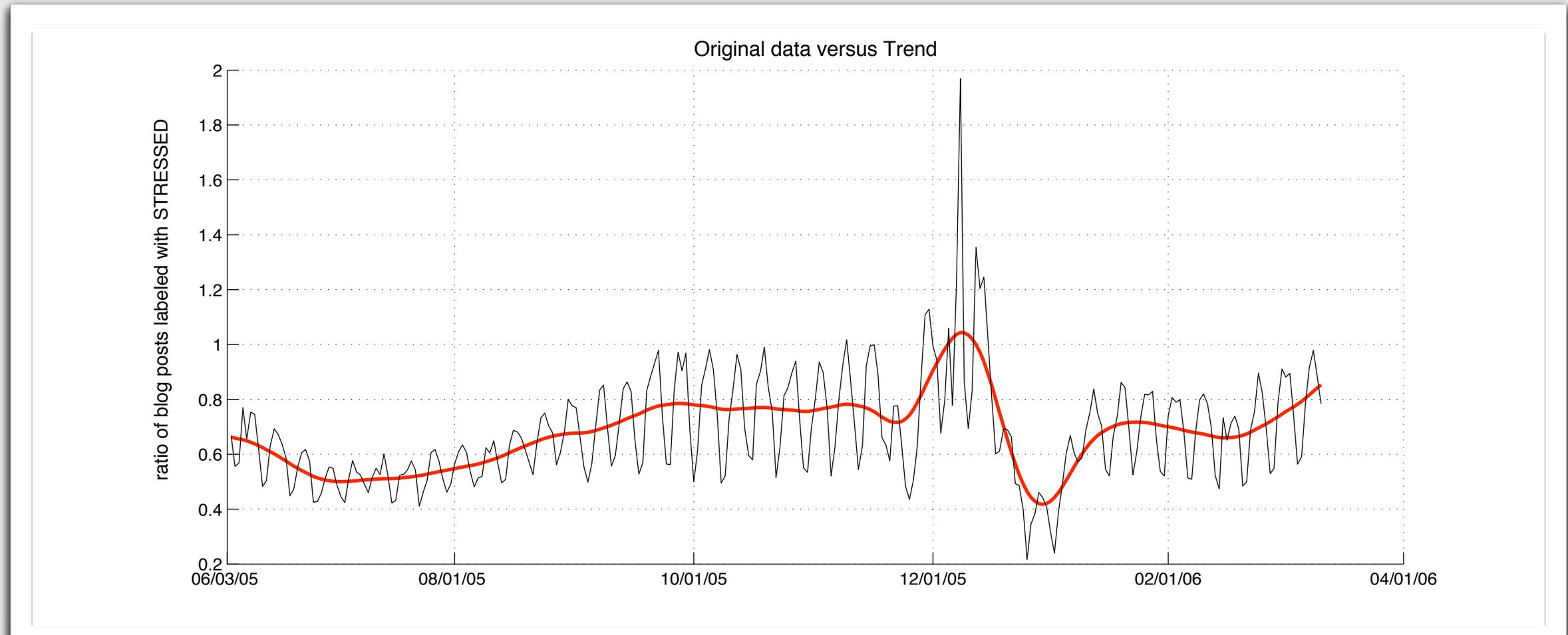
Table II. Queries Correlated with "Mortgage" Over Time (These change dramatically as the query is farther away in time. The users' interests move from mortgage basics to property searching, to insurance and taxes, to furnishings, to pools and patios. Here we give the top 40 terms that did not show up in the previous time period).

Time Period				
0–30 min	1–7 days	7–30d	30–90d	90–365d
mortgage	realtors	llc	kohls	patio
mortage	owner	associates	bath	harbor
mortgage	homes	insurance	overstock	outdoor
calculator	mls	lowes	barn	replacement
mortgages	remax	notary	sears	pools
lenders	property	depot	linens	hampton
calculators	financial	savings	beyond	lawn
countrywide	appraisers	construction	kmart	enterprise
gmac	builders	condo	pottery	ymca
refinance	prudential	business	walmart	vehicle
rates	zillow	secretary	outlet	supply
interest	bankruptcy	furniture	costco	resorts
broker	real	allstate	target	lake
lending	keller	companies	pier	rv
lender	properties	contractors	bed	walgreens
payment	agreement	cost	grill	newport
loan	appraisals	reverse	kitchen	lumber
amro	residential	federal	shield	oak
emc	lease	sale	macys	authority
brokers	county	housing	vacations	concrete
abn	modular	assessors	southwest	vehicles
amortization	attorney	irs	chamber	chrysler

M. Richardson, Learning about the world through long-term query behavior. ACM TWeb 2(4), 2008

Open
Long trails

Open
Long trails


Original data versus Trend
ratio of blog posts labeled with STRESSED
2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
06/03/05
08/01/05
10/01/05
12/01/05
02/01/06
04/01/06

Explaining the unexpected

Long trails

Open

Connected signals

Explaining the unexpected

MoodViews — Moodgrapher

http://moodviews.com/Moodgrapher/index.cgi

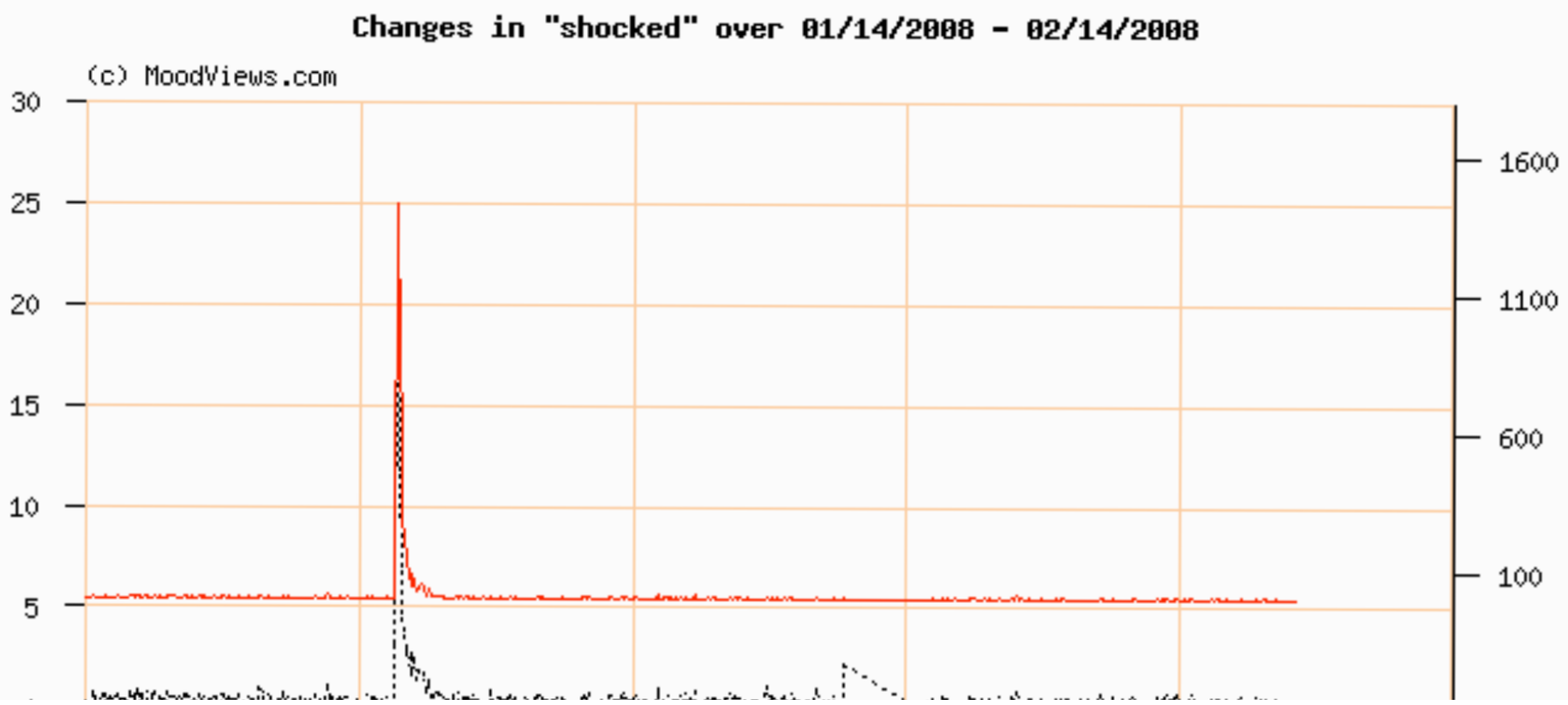
Google

MoodViews | **Moodgrapher** | Moodteller | Moodsignals | Moodspotter | Moodfeeds

About ... Cyclic events Global events

Moodgrapher

Moodgrapher plots the mood levels reported by LiveJournal users in their posts during the last few days, updated every 10 minutes. Moodgrapher tracks both the absolute counts and the rate of change. (what's this?)



MoodViews — Moodgrapher

http://moodviews.com/Moodgrapher/index.cgi

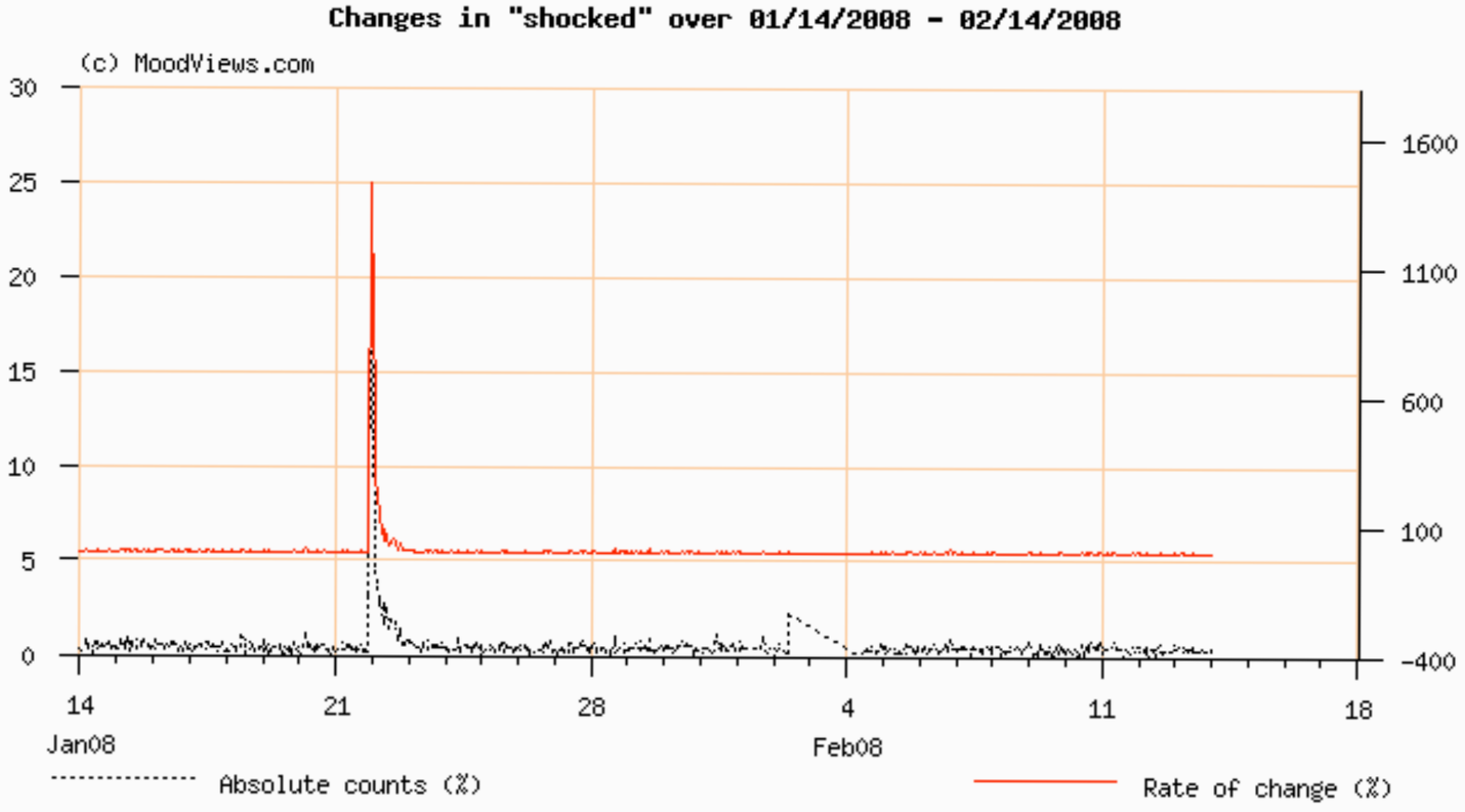
Google

MoodViews | **Moodgrapher** | Moodteller | Moodsignals | Moodspotter | Moodfeeds

About ... Cyclic events Global events

Moodgrapher

Moodgrapher plots the mood levels reported by LiveJournal users in their posts during the last few days, updated every 10 minutes. Moodgrapher tracks both the absolute counts and the rate of change. (what's this?)



Search "shocked" posts

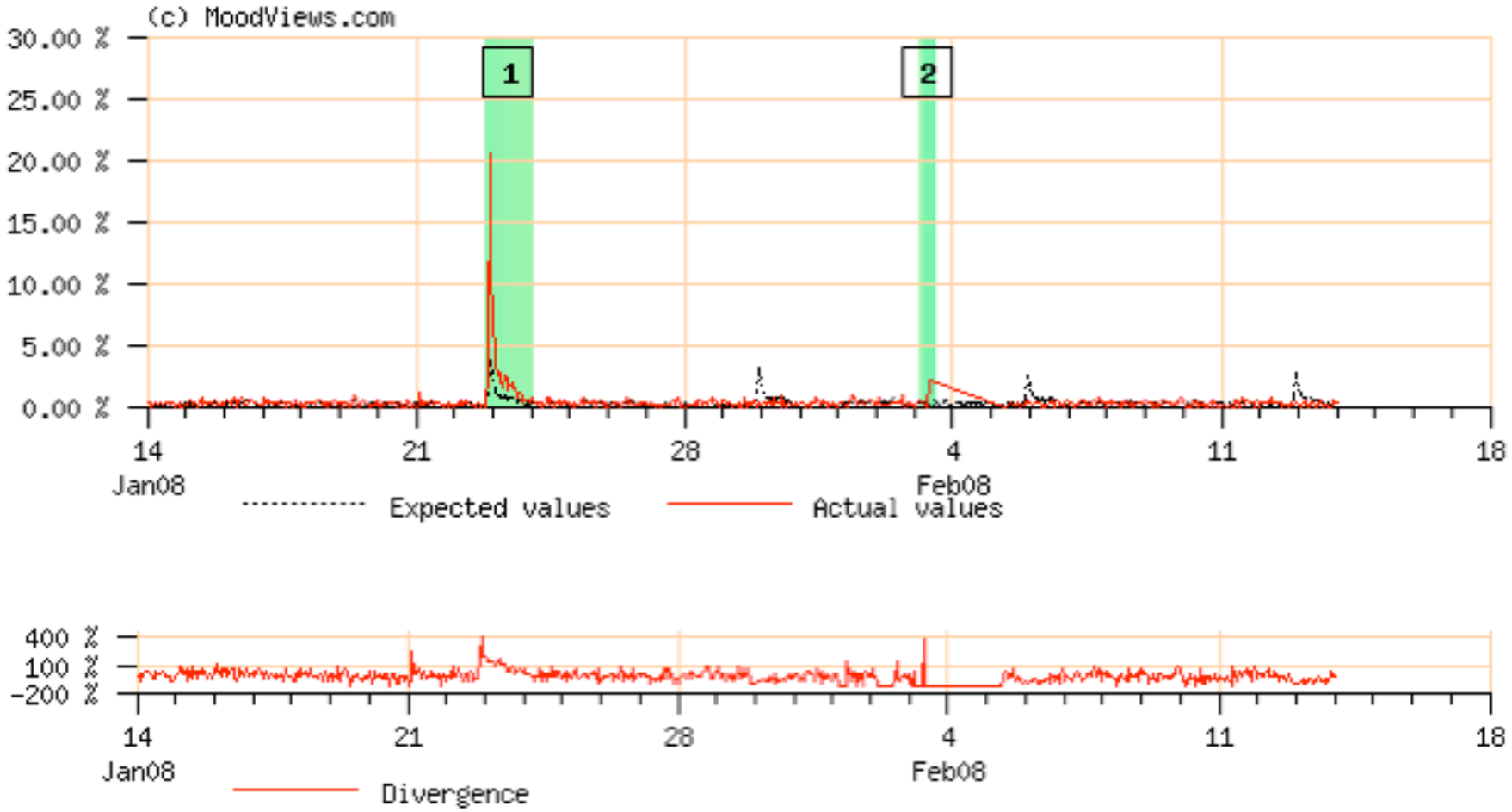
Show the mood level compare moods »

MoodViews — Moodgrapher

MoodViews — Moodsignals

http://moodviews.com/Moodsignals/index.cgi RSS Google

Changes in "shocked" over 01/14/2008 - 02/13/2008



Discover peaks

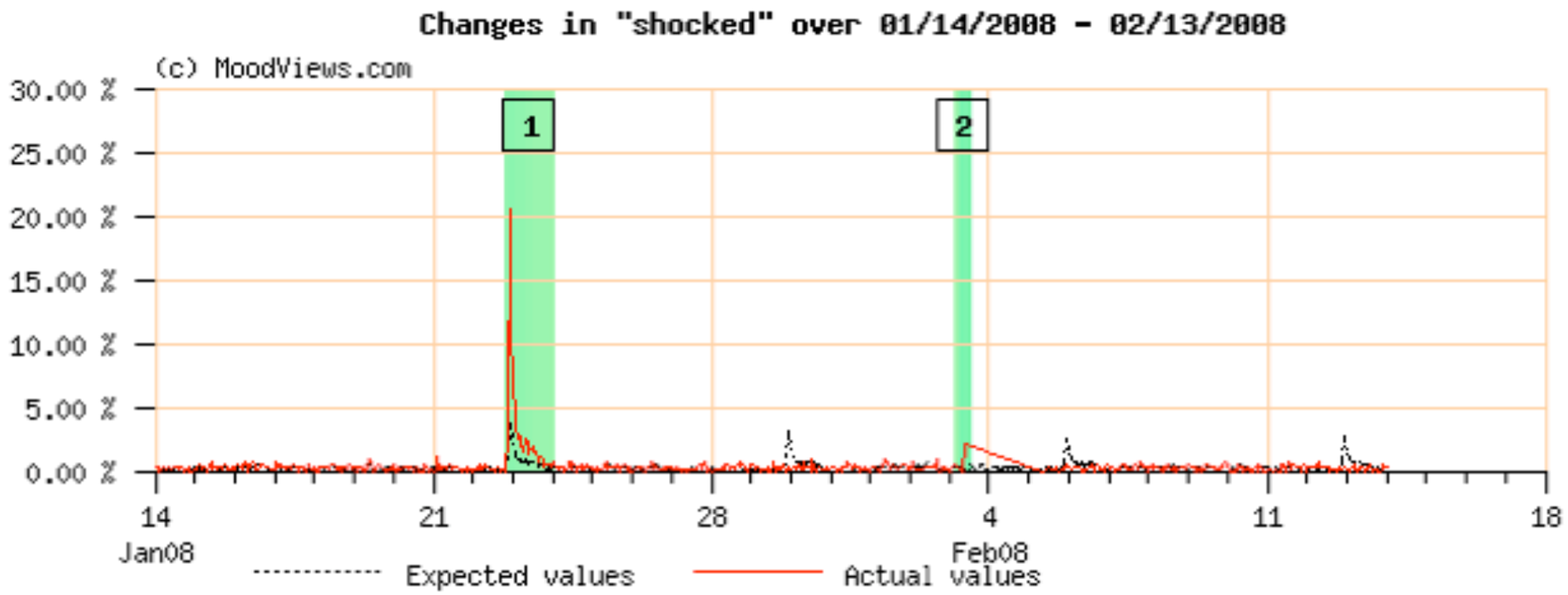
for this mood: shocked

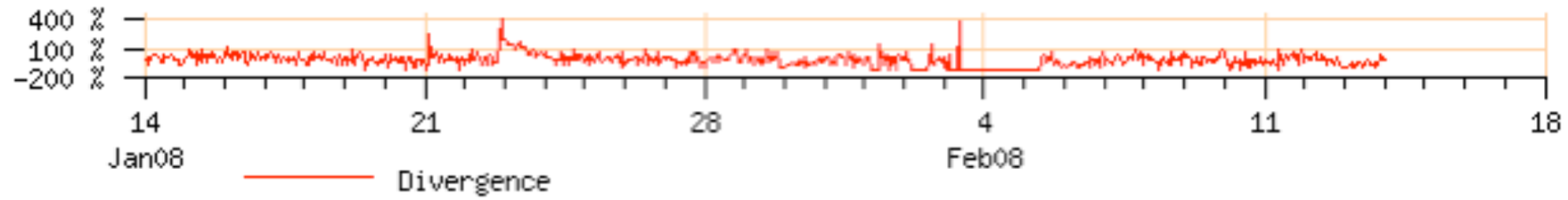
and for this date interval: 01/14/2008 02/13/2008

Graph

(Dates have to be provided in mm/dd/yyyy format.)

Peak explanation





Discover peaks

for this mood: shocked

and for this date interval: 01/14/2008 02/13/2008

Graph

(Dates have to be provided in mm/dd/yyyy format.)

Peak explanation

Peak (1) 01/22/2008 19h - 01/24/2008 01h

Overused terms during the peak period:

ledger, heath's, dead, actor, shock, joker, brokeback, knights, found, dying, renfro, overdose, news

More peaks in "shocked"

- Dec 27, 2007, 12:00
- Oct 31, 2007, 00:00
- Oct 20, 2007, 00:00
- Oct 19, 2007, 08:00

So what?

Search, recommend and predict

- Focus on user behavior
- Exploit user behavior
  - Recognize it, segment it, label it, interpret it, connect it
  - Predict it

Clicked URLs-behavior

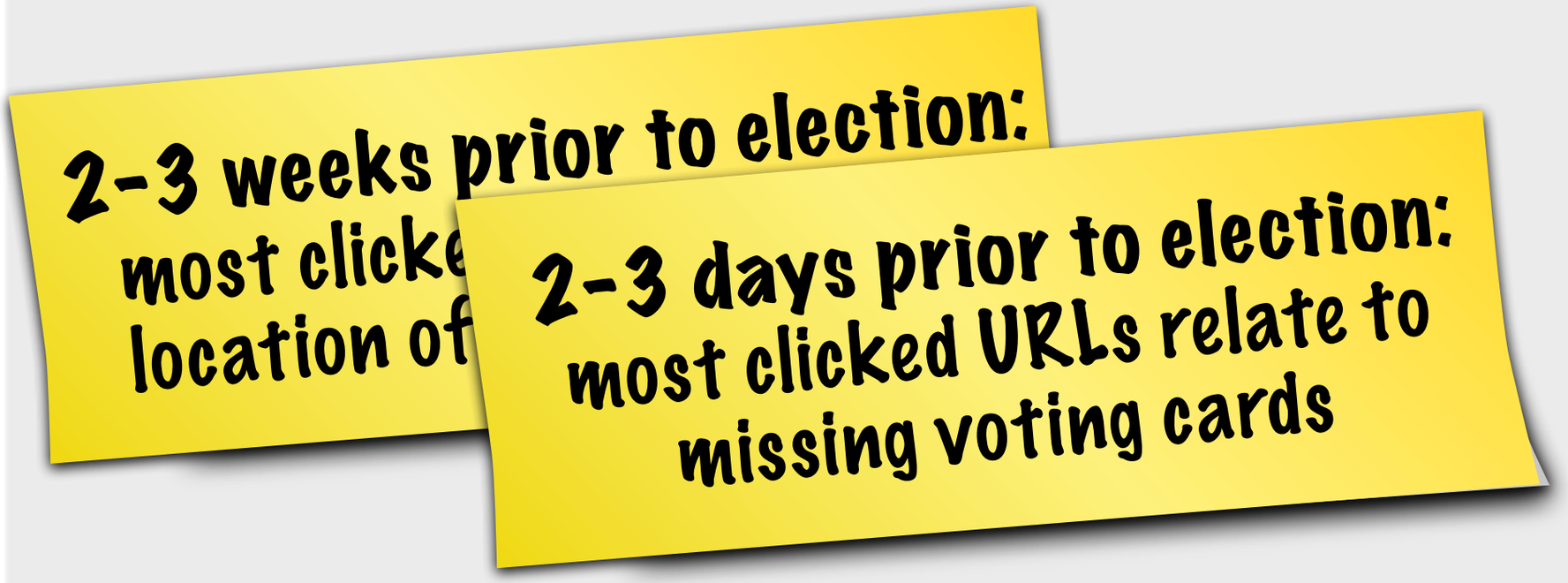
- Relevance is time-dependent
- Bring temporal user behavior to bear on ranking search results
- Adjust on the fly ≠ temporal priors

Clicked URLs-behavior

2-3 weeks prior to election: most clicked URLs relate to location of polling stations

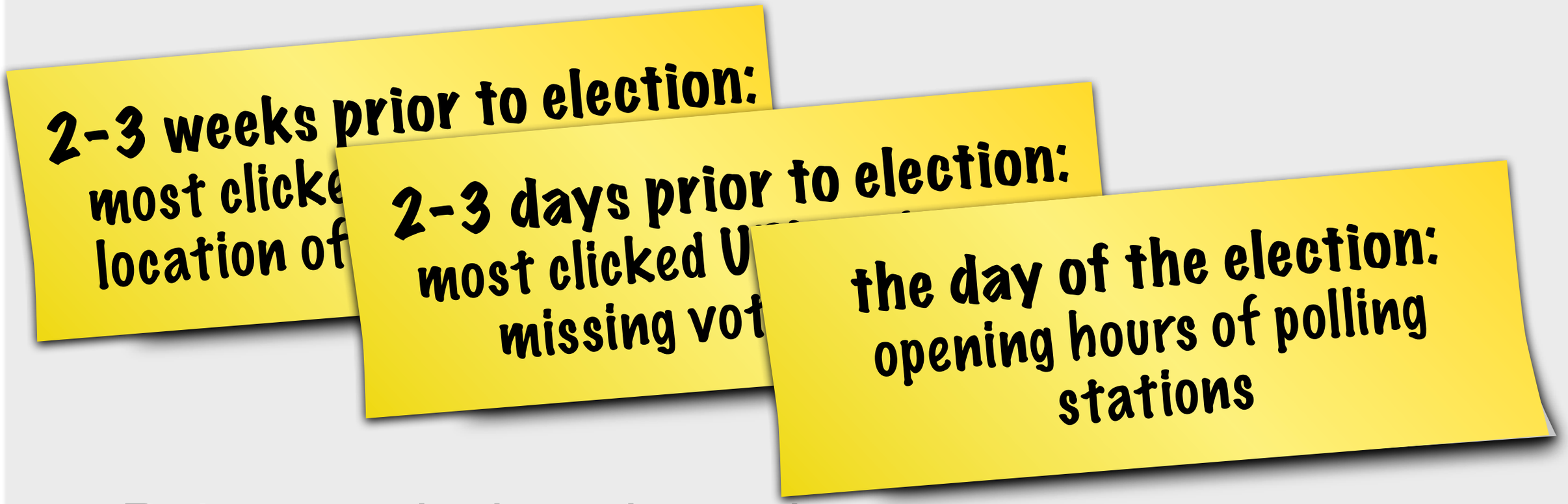
- Relevance is time-dependent
- Bring temporal user behavior to bear on ranking search results
- Adjust on the fly ≠ temporal priors

Clicked URLs-behavior



- Relevance is time-dependent
- Bring temporal user behavior to bear on ranking search results
- Adjust on the fly ≠ temporal priors

Clicked URLs-behavior



- Relevance is time-dependent
- Bring temporal user behavior to bear on ranking search results
- Adjust on the fly ≠ temporal priors

Prediction: The volume of comments

- How many (if any) comments will a news article get?
- "Impact prediction"
- Assumption: there are regularities in what makes people respond to a news article
- Core intuitions
  - Where, what, when, who?
  - What else is going on?

Long trails

Open

Connected signals

Prediction: The volume of comments

- How many (if any) comments will a news article get?
- "Impact prediction"
- Assumption: there are regularities in what makes people respond to a news article
- Core intuitions
  - Where, what, when, who?
  - What else is going on?

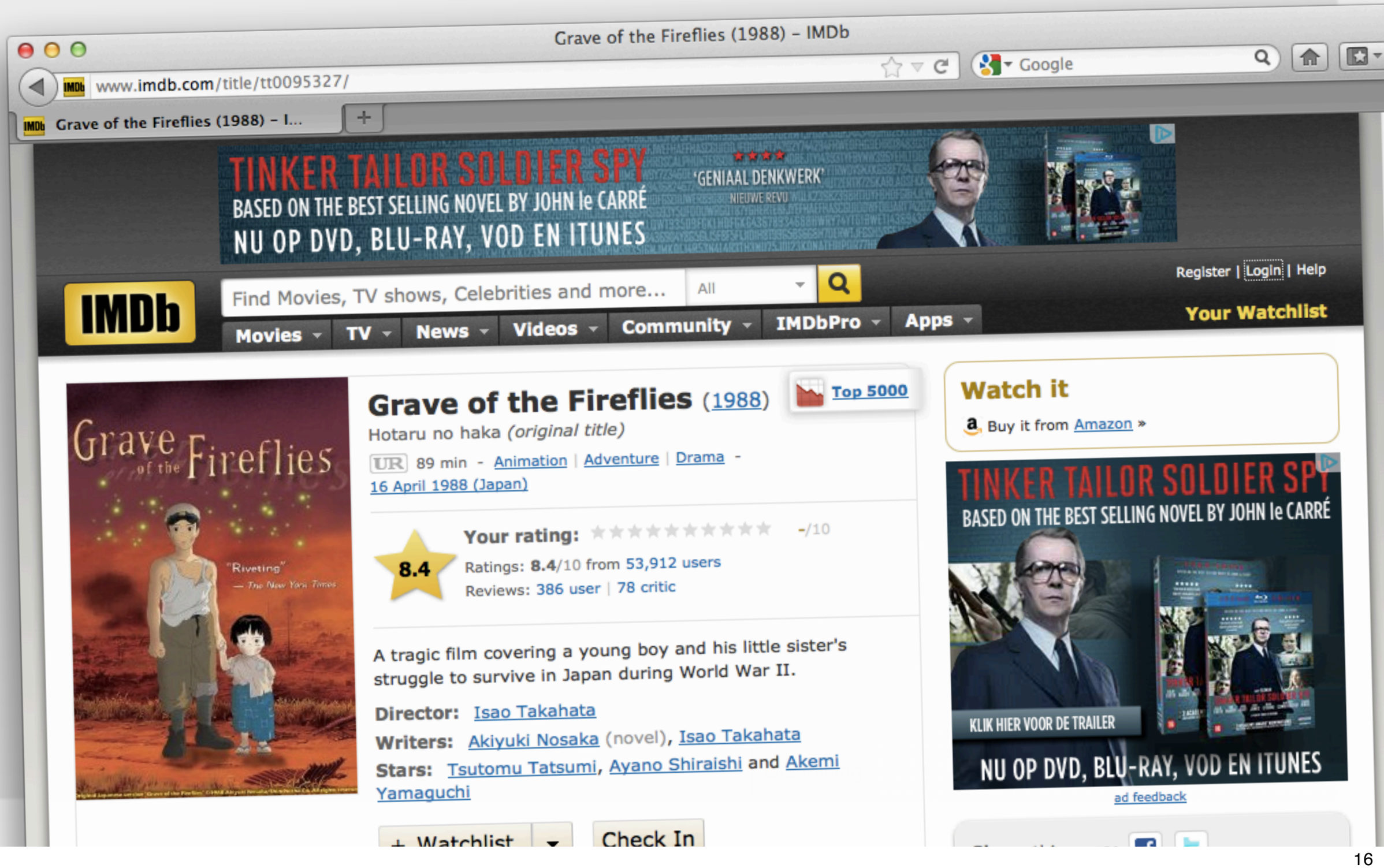
Feature	Description	Type
Surface features		
month	Month (1-12)	Nom
wom	Week of the month (1-4)	Nom
dow	Day of the week (1-7)	Nom
day	Day of the month (1-31)	Nom
hour	Hour of the day (0-23)	Nom
first_half_hour	Publication in the first 30 minutes of the hour	Nom
art_char_length	Article content length	Int
category_count	Number of categories it is published on	Int
has_summary	Article has summary	Int
has_content	Article has content (HTML incl.)	Int
has_content_clean	Article has content (only text)	Int
links_cnt	Number of out-links	Int
authors_cnt	Number of authors	Int
Cumulative features		
art_same_hr	Published articles in same hour for source	Int
dupes_int_cnt	Near-duplicates in same source	Int
dupes_ext_cnt	Near-duplicates in other sources	Int
Textual features		
	tf of top-100 terms ranked by their log-likelihood score for each source	Int
Semantic features		
ne_loc_cnt	Number of location-type entities	Int
ne_per_cnt	Number of person-type entities	Int
ne_org_cnt	Number of organisation-type entities	Int
ne_misc_cnt	Number of miscellaneous-type entities	Int
has_local	Any entities referring to the Netherlands	Int
	tf of top-50 entities from each entity type, ranked by their log-likelihood score for each source	Int
Real-word features		
temperature	Temperature in Celsius at publication time	Num

Long trails

Open

Connected signals

Prediction: Internet movie database ratings

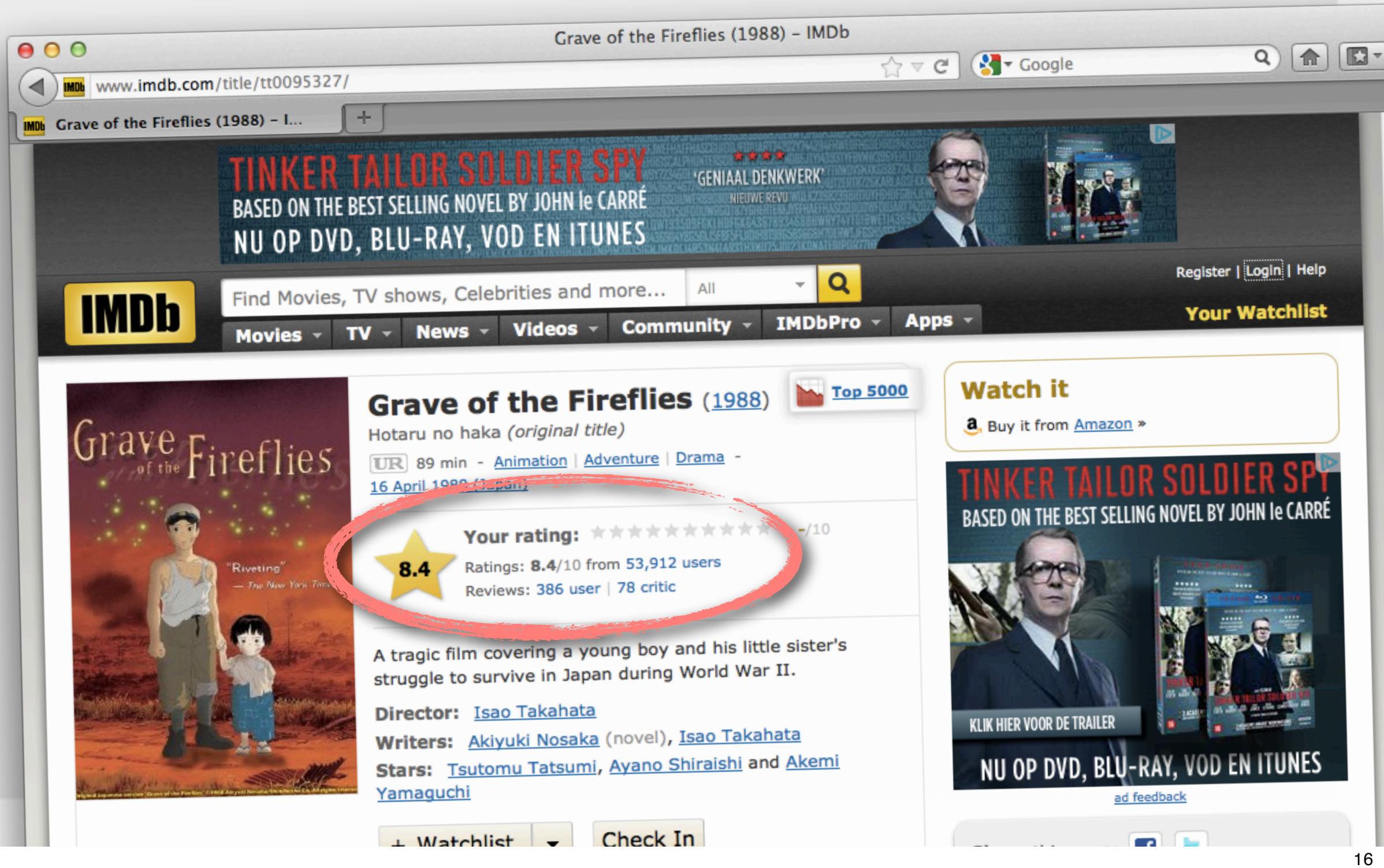


Long trails

Open

Connected signals

Prediction: Internet movie database ratings



Prediction: Internet movie database ratings

- Observe a “signal” in one channel, to predict a “signal” in another channel
  - Twitter & YouTube / IMDb
- Two classes of features
  - Quantitative indicators aim at capturing the amount of activity around a movie title
    - views, number of comments, number of favorites, number of likes, the number of dislikes, the fraction of likes over dislikes for each trailer clip on YouTube, the number of tweets on Twitter.
  - Qualitative indicators aim at capturing the meaning of the activity (i.e., what people say about a movie)
    - Textual features extracted by comparing the log-likelihood of a term in tweets and YouTube comments associated with the top-N and bottom-N movies, based on their IMDb ratings
      - Examples of extracted positive textual features amaz, perfect, awesom
      - Negative ones stupid, worst, terribl

Prediction: Radicalization

- Signs of online radicalization
- Not a simple classification task based on a static snapshot of data
- From “on earth” social psychology:
  - Two phases
    - One
      - Talking about a growing number of topics
      - Talking to an increasingly large group
    - Two
      - Restrict to a small number of topics
      - Limit audience to a small number of people
- Can be implemented by
  - Keeping track of language model
  - Keeping track of structure of social network
- Interesting false positives

What's next?

Life mining agenda

Interpret search trails across time

- Scattered across multiple sessions, status update, tweets
- Recognize and label activities

Life mining agenda

Interpret search trails across time

- Scattered across multiple sessions, status update, tweets
- Recognize and label activities

Link and interpret the trails

- Online-and-offline behavior
- Social signals, textual, multimedia, news, task setting, ...

Life mining agenda

Interpret search trails across time

- Scattered across multiple sessions, status update, tweets
- Recognize and label activities

Link and interpret the trails

- Online-and-offline behavior
- ...ask setting, ...

Self-tuning search and recommendation engines

- Use temporal user behavior for adjusting rankings

Life mining agenda

Interpret search trails across time

- Scattered across multiple sessions, status update, tweets
- Recognize and label activities

Link and interpret the trails

- Online-and-offline behavior
- ...ask setting, ...

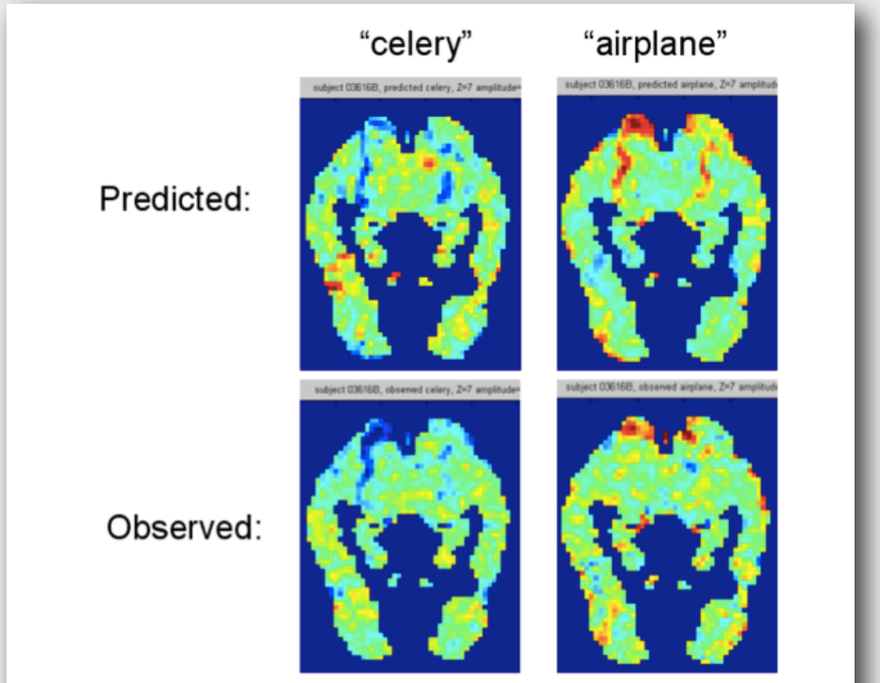
Self-tuning search and recommendation engines

- Use temporal user behavior for adjusting rankings

Develop predictive models

- Individual behavior, group behavior

And beyond that?



- Track even more signals to interpret and recommend …
  - Size and complexity of social network correlated with brain structure
  - Different spatial patterns of neural activation are associated with thinking about different semantic categories of words and pictures (e.g., tools, buildings, animals).
  - Brain structure of Googlers different from brain structure of "deep readers"