Department of Methodology Inaugural Lecture

LSE Department of Methodology

The Challenge of Big Data for the Social Sciences

Professor Kenneth Benoit
Professor of Quantitative Social Research Methods, LSE

Kenneth Cukier
Data Editor,
The Economist

Professor Simon Hix
Chair, LSE

LSE events

Suggested hashtag for Twitter users: **#LSEdata**

CPD CERTIFIED
The CPD Certification Service

The Challenge of Big Data for the Social Sciences

Kenneth Benoit
LSE Department of Methodology

Inaugural Lecture, 16 February 2015

LSE events @LSEpublicevents · 4h

16/2 @MethodologyLSE Inaugural Lecture by @kenbenoit on 'The Challenge of Big Data for the Social Sciences' #LSEdata buff.ly/1J9geQt

1 5

LSE events
@LSEpublicevents

Following

.@kncukier Data Editor at The Economist will respond to @kenbenoit's inaugural lecture on 16/2 chaired by @simonjhix lse.ac.uk/publicEvents/e...

RETWEETS	FAVORITES
2	2

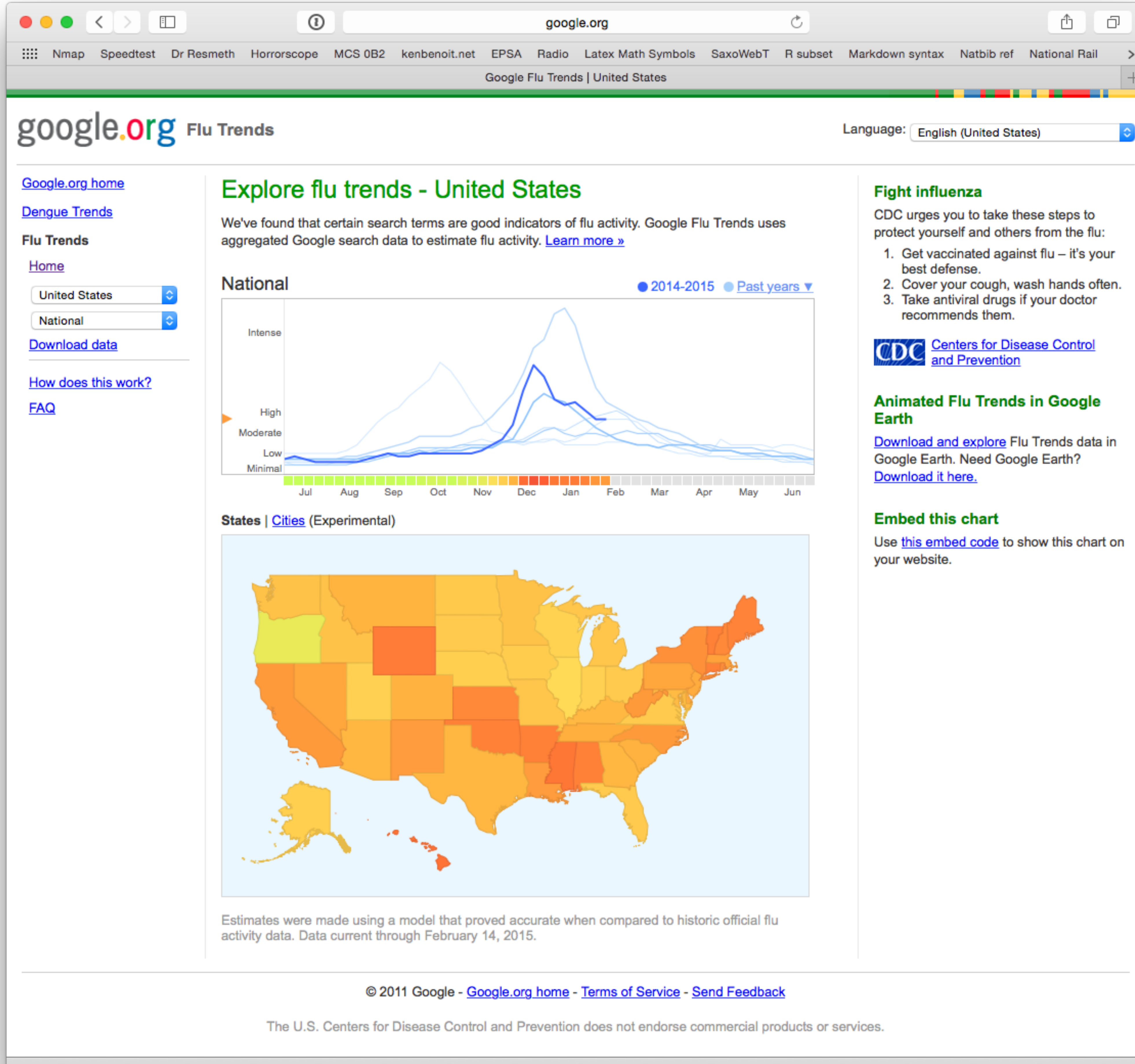
8:10 PM - 9 Feb 2015

exabyte = 1 x 10^{18}

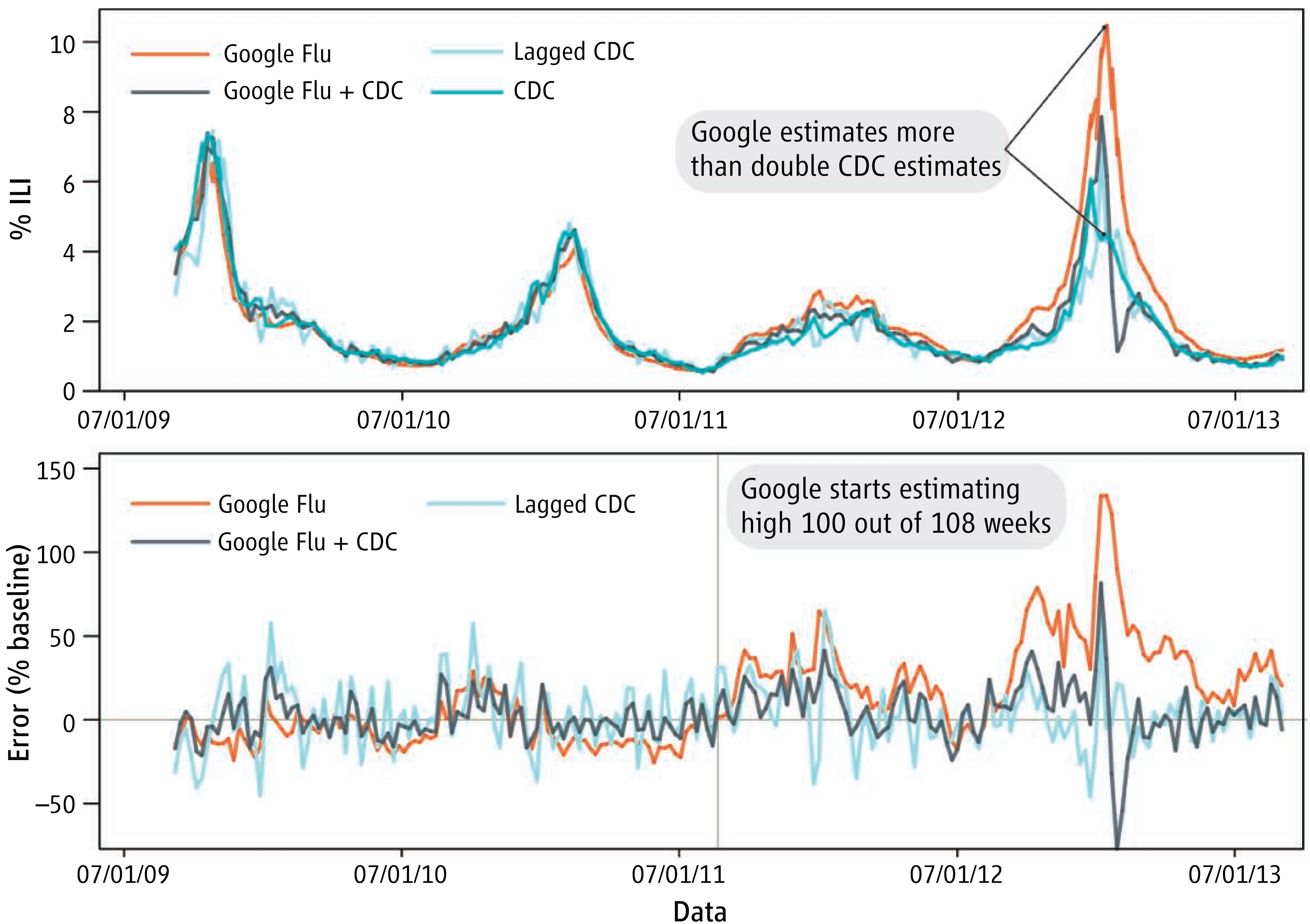
= 1,000,000,000,000,000,000 bytes

daily

Human Genome Project



google.org
Nmap Speedtest Dr Resmeth Horrorscope MCS 0B2 kenbenoit.net EPSA Radio Latex Math Symbols SaxoWebT R subset Markdown syntax Natbib ref National Rail
Google Flu Trends | United States
google.org Flu Trends
Language: English (United States)
Google.org home
Dengue Trends
Flu Trends
Home
United States
National
Download data
How does this work?
FAQ
Explore flu trends - United States
We've found that certain search terms are good indicators of flu activity. Google Flu Trends uses aggregated Google search data to estimate flu activity. Learn more »
National
2014-2015
Past years
Intense
High
Moderate
Low
Minimal
Jul Aug Sep Oct Nov Dec Jan Feb Mar Apr May Jun
States | Cities (Experimental)
Estimates were made using a model that proved accurate when compared to historic official flu activity data. Data current through February 14, 2015.
Fight influenza
CDC urges you to take these steps to protect yourself and others from the flu:
1. Get vaccinated against flu – it's your best defense.
2. Cover your cough, wash hands often.
3. Take antiviral drugs if your doctor recommends them.
CDC
Centers for Disease Control and Prevention
Animated Flu Trends in Google Earth
Download and explore Flu Trends data in Google Earth. Need Google Earth? Download it here.
Embed this chart
Use this embed code to show this chart on your website.
© 2011 Google - Google.org home - Terms of Service - Send Feedback
The U.S. Centers for Disease Control and Prevention does not endorse commercial products or services.



"The Parable of Google Flu", Lazer et al (2014)

BIG DATA

The Parable of Google Flu: Traps in Big Data Analysis

Large errors in flu prediction were largely avoidable, which offers lessons for the use of big data.

David Lazer,[1,2*] Ryan Kennedy,[1,3,4] Gary King,[3] Alessandro Vespignani[5,6,3]

In February 2013, Google Flu Trends (GFT) made headlines but not for a reason that Google executives or the creators of the flu tracking system would have hoped. *Nature* reported that GFT was predicting more than double the proportion of doctor visits for influenza-like illness (ILI) than the Centers for Disease Control and Prevention (CDC), which bases its estimates on surveillance reports from laboratories across the United States (*1*, *2*). This happened despite the fact that GFT was built to predict CDC reports. Given that GFT is often held up as an exemplary use of big data (*3*, *4*), what lessons can we draw from this error?

The problems we identify are not limited to GFT. Research on whether search or social media can predict x has become common-

run ever since, with a few changes announced in October 2013 (*10*, *15*).

Although not widely reported until 2013, the new GFT has been persistently overestimating flu prevalence for a much longer time. GFT also missed by a very large margin in the 2011–2012 flu season and has missed high for 100 out of 108 weeks starting with August 2011 (see the graph). These errors are not randomly distributed. For example, last week's errors predict this week's errors (temporal autocorrelation), and the direction and magnitude of error varies with the time of year (seasonality). These patterns mean that GFT overlooks considerable information that could be extracted by traditional statistical methods.

Even after GFT was updated

The Netflix Prize Rules

For a printable copy of these rules, go here.

Overview:

We're quite curious, really. To the tune of one million dollars.

Netflix is all about connecting people to the movies they love. To help customers find those movies, we've developed our world-class movie recommendation system: Cinematch[SM]. Its job is to predict whether someone will enjoy a movie based on how much they liked or disliked other movies. We use those predictions to make personal movie recommendations based on each customer's unique tastes. And while Cinematch is doing pretty well, it can always be made better.

Now there are a lot of interesting alternative approaches to how Cinematch works that we haven't tried. Some are described in the literature, some aren't. We're curious whether any of these can beat Cinematch by making better predictions. Because, frankly, if there is a much better approach it could make a big difference to our customers and our business.

So, we thought we'd make a contest out of finding the answer. It's "easy" really. We provide you with a lot of anonymous rating data, and a prediction accuracy bar that is 10% better than what Cinematch can do on the same training data set. (Accuracy is a measurement of how closely predicted ratings of movies match subsequent actual ratings.) If you develop a system that we judge most beats that bar on the qualifying test set we provide, you get serious money and the bragging rights. But (and you knew there would be a catch, right?) only if you share your method with us and describe to the world how you did it and why it works.

Serious money demands a serious bar. We suspect the 10% improvement is pretty tough, but we also think there is a good chance it can be achieved. It may take months; it might take years. So to keep things interesting, in addition to the Grand Prize, we're also offering a $50,000 Progress Prize each year the contest runs. It goes to the team whose system we judge shows the most improvement over the previous year's best accuracy bar on the same qualifying test set. No improvement, no prize. And like the Grand Prize, to win you'll need to share your method with us and describe it for the world.

There is no cost to enter, no purchase required, and you need not be a Netflix subscriber. So if you know (or want to learn) something about machine learning and recommendation systems, give it a shot. We could make it really worth your while.

C The 18 Predictors needed for 10%

In Figure 6 one can see that 18 predictors suffice to reach a 10% improvement on the quiz set. In Table 11 we provide an ordered list of these predictors. Most of them stem from nonlinear probe blends.

We found these predictors with backward selection. We started with all predictors and iteratively removed the predictor with the lowest contribution. Please note that this method is not guaranteed to find the smallest possible subset of predictors which achieve a 10% improvement. So there might be a smaller subset or an 18 predictor subset with a lower RMSE.

#	blend RMSE	Predictor
1	0.85834	PB-101, rmse=0.8584, Ensemble neural network blend, 1149 results from small nets with 4 neurons (29 trainable weights) in the hidden layer, 3-predictor random subsets. All results were combined with a 2-date-bins linear blender. The base set of predictors are {ALL-476, BC-Exact-340}.
2	0.85693	PB-115, rmse=0.8587, Neural network blend on tree blends, extended with SVD and RBM features, where the probe prediction of the trees is the out-of-bag estimate. The net has two hidden layers with 17 and 7 neurons. Training was stopped when the 4-fold CV error has reached the minimum (after 493 epochs). The input is: {PB-143 ... PB-163, SVD++10-cross-145, RBM-50-user}
3	0.85665	PB-107, rmse=0.8600, Ensemble neural network blend on a 168 predictor subset by Pragmatic Theory. 980 results were blended with a 4-frequency-bins linear blender. Each of the results were produced by a 3-predictor random subset blend of the 168. These subsets were blended by a 1HL net with 5 neurons. Input: {PT-168}
4	0.85662	PB-054, rmse=0.8658, Neural network blend on a forward selection by BigChaos' predictors. Here, top-15 are selected with a simple greedy forward selection algorithm. Blending is done with a 1HL net with 13 neurons. Training was stopped after 1334 epochs. Input: {BC-192-Top15}
5	0.85658	rmse=8713, A result by BellKor: last one at Sec. VII of [12]
6	0.85655	PB-112, rmse=0.8621, Kernel Ridge Regression blend on results by all 3 teams. The KRR algorithm is computational very expensive, therefore we draw 4000 random samples from the 1.4M probe set. A Gauss kernel with $\sigma = 100$ is used, the ridge constant is $\lambda = 1e - 6$. 8 results are combined linearly. Input: {ALL-476}

BigChaos solution to the Netflix Prize (2009)

#	effect	shrinkage	kernel width
0	Global time mean	α_0	σ_0
1	Movie time effect	α_1	σ_1
2	User time effect	α_2	σ_2
3	Movie time effect	α_3	σ_3
4	User time effect	α_4	σ_4
5	User time effect: user x sqrt(time(user))	α_5	σ_5
6	User time effect: user x sqrt(time(movie))	α_6	σ_6
7	Movie time effect: movie x sqrt(time(movie))	α_7	σ_7
8	Movie time effect: movie x sqrt(time(user))	α_8	σ_8
9	User time effect: user x average(movie)	α_9	σ_9
10	User time effect: user x votes(movie)	α_{10}	σ_{10}
11	Movie time effect: movie x average(user)	α_{11}	σ_{11}
12	Movie time effect: movie x votes(user)	α_{12}	σ_{12}
13	Movie time effect: movie x avgMovieProductionYear(user)	α_{13}	σ_{13}
14	User time effect: user x productionYear(movie)	α_{14}	σ_{14}
15	User time effect: user x std(movie)	α_{15}	σ_{15}
16	Movie time effect: movie x std(user)	α_{16}	σ_{16}
17	User time effect: user x average(movie) (from previous effect)	α_{17}	σ_{17}
18	Movie time effect: movie x average(user) (from previous effect)	α_{18}	σ_{18}
19	Movie time effect: movie x percentSingleVotes(user)	α_{19}	σ_{19}
20	Movie time effect: movie x avgStringlenTitle(user)	α_{20}	σ_{20}
21	Movie time effect: movie x ratingDateDensity(user)	α_{21}	σ_{21}
22	Movie time effect: movie x percentMovieWithNumberInTitle(user)	α_{22}	σ_{22}
23	User time effect: user x stringlengthFromTitle(movie)	α_{23}	σ_{23}
24	User time effect: user x ratingDateDensity(movie)	α_{24}	σ_{24}

Table 2: Overview on 24 Global Time Effects

BigChaos solution to the Netflix Prize (2009)

Computers in Human Behavior 36 (2014) 94–101

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: www.elsevier.com/locate/comphumbeh

Social network sites, marriage well-being and divorce: Survey and state-level evidence from the United States

Sebastián Valenzuela [a,*], Daniel Halpern [a], James E. Katz [b]

[a] *Pontificia Universidad Católica de Chile, School of Communications, Alameda 340, Santiago 8331150, Chile*
[b] *Boston University, College of Communication, Division of Emerging Media Studies, Boston, MA 02215, United States*

A R T I C L E I N F O

Article history:

Keywords:
Social network sites
Facebook
Marriage well-being
Divorce

A B S T R A C T

This study explores the relationship between using social networks sites (SNS), marriage satisfaction and divorce rates using survey data of married individuals and state-level data from the United States. Results show that using SNS is negatively correlated with marriage quality and happiness, and positively correlated with experiencing a troubled relationship and thinking about divorce. These correlations hold after a variety of economic, demographic, and psychological variables related to marriage well-being are taken into account. Further, the findings of this individual-level analysis are consistent with a state-level analysis of the most popular SNS to date: across the U.S., the diffusion of Facebook between 2008 and 2010 is positively correlated with increasing divorce rates during the same time period after controlling for all time-invariant factors of each state (fixed effects), and continues to hold when time-varying economic and socio-demographic factors that might affect divorce rates are also controlled. Possible explanations for these associations are discussed, particularly in the context of pro- and anti-social perspectives towards SNS and Facebook in particular.

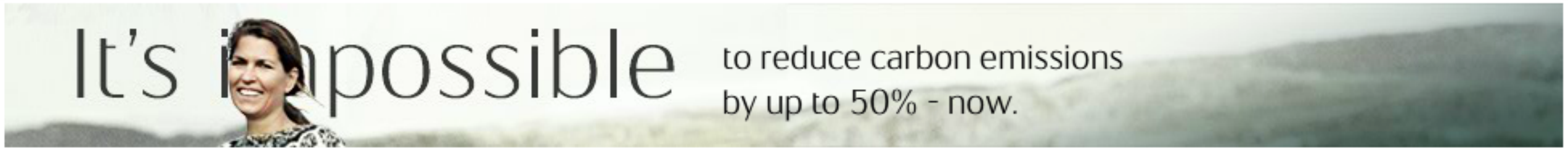
© 2014 Elsevier Ltd. All rights reserved.

“First, **excessive use of social media** has been associated with compulsive use, which **may create psychological, social, school and/or work difficulties in a person’s life.** These phenomena, in turn, may **trigger marriage unhappiness and, ultimately, divorce**.

Second, **Facebook in particular creates an environment with potential situations that may evoke feelings of jealousy between partners, harming the quality of their relationship.**

And third, we noted that services like Facebook have unique affordances that may help partners to **reduce searching costs for extra-matrimonial affairs** and consequently **may contribute to cheating.**”

Valenzuela et al (2014)



Social networking linked to divorce, marital unhappiness

584 SHARES

11 COMMENTS Join the Discussion

Everett Rosenfeld | @Ev_Rosenfeld
Tuesday, 8 Jul 2014 | 2:10 PM ET
CNBC

In what may be of little surprise to avid readers of FacebookCheating.com, a new study found a correlation between social media use and divorce rates in the United States.

The study, published in the journal Computers in Human Behavior by researchers from Pontificia Universidad Católica de Chile and Boston University, compared state-by-state divorce rates to per-capita Facebook accounts. In a separate analysis, they also used data from a 2011-2012 survey that asked individuals about marriage quality and social media use.

Facebook Users Are More Likely to Divorce, Study Finds

Comments 2 Share 1 g+1 5 Tweet 32 Like Share 186
Sign Up for Free eNewsletter ››
Like Share on Facebook Share on Twitter

BY MICHAEL GRYBOSKI , CHRISTIAN POST REPORTER

Public Opinion Quarterly, Vol. 72, No. 2, Summer 2008, pp. 216–240

ESTIMATING CAUSAL EFFECTS OF BALLOT ORDER FROM A RANDOMIZED NATURAL EXPERIMENT

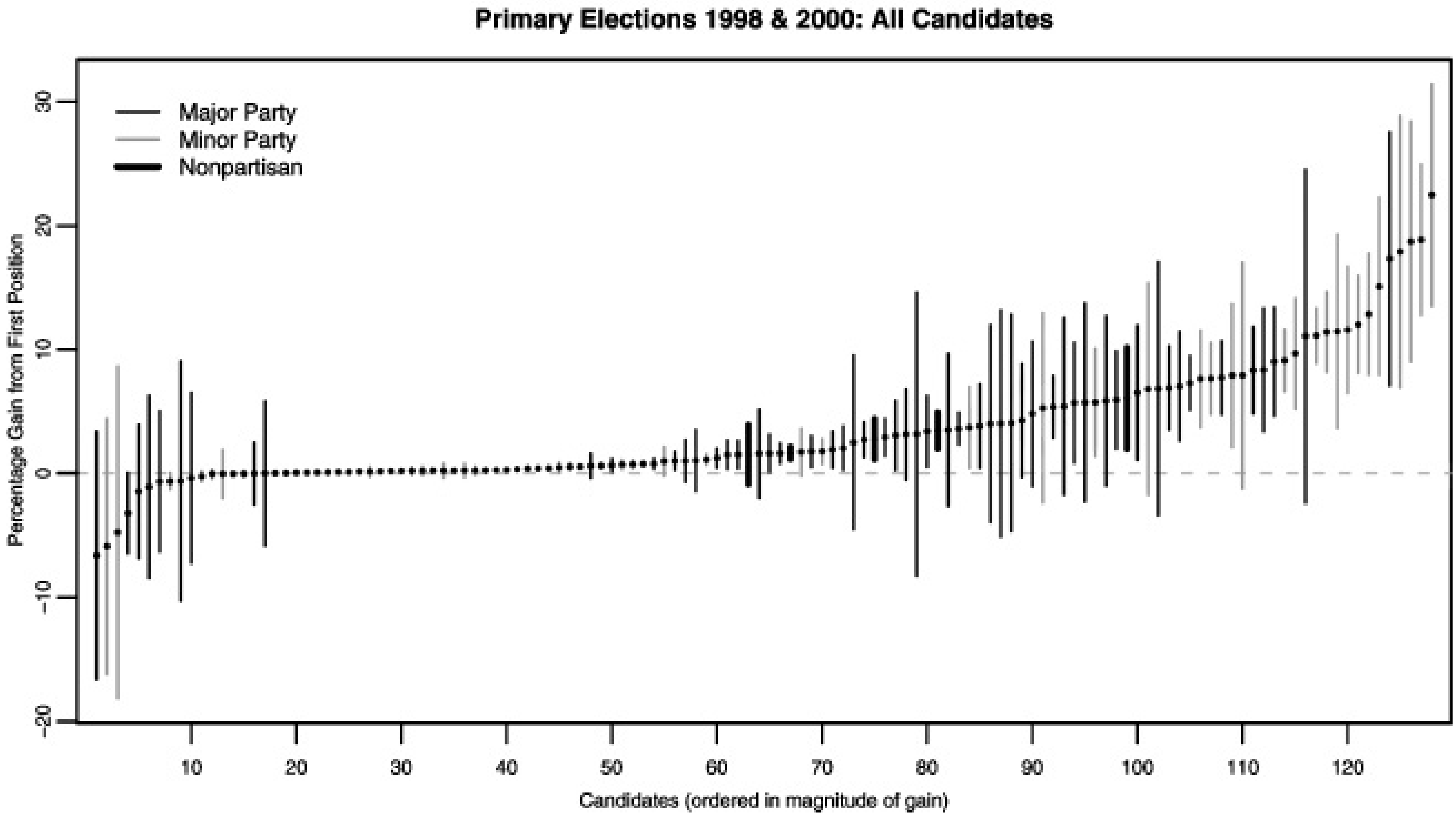
THE CALIFORNIA ALPHABET LOTTERY, 1978–2002

DANIEL E. HO
KOSUKE IMAI

Abstract Randomized natural experiments provide social scientists with rare opportunities to draw credible causal inferences in real-world settings. We capitalize on such a unique experiment to examine how the name order of candidates on ballots affects election outcomes. Since 1975, California has randomized the ballot order for statewide offices with a complex alphabet lottery. Adapting statistical techniques to this lottery and addressing methodological problems of conventional approaches, our analysis of statewide elections from 1978 to 2002 reveals that, in general elections, ballot order significantly impacts only minor party candidates, with no detectable effects on major party candidates. These results contradict previous research, finding large effects in general elections for major party candidates. In primaries, however, we show that

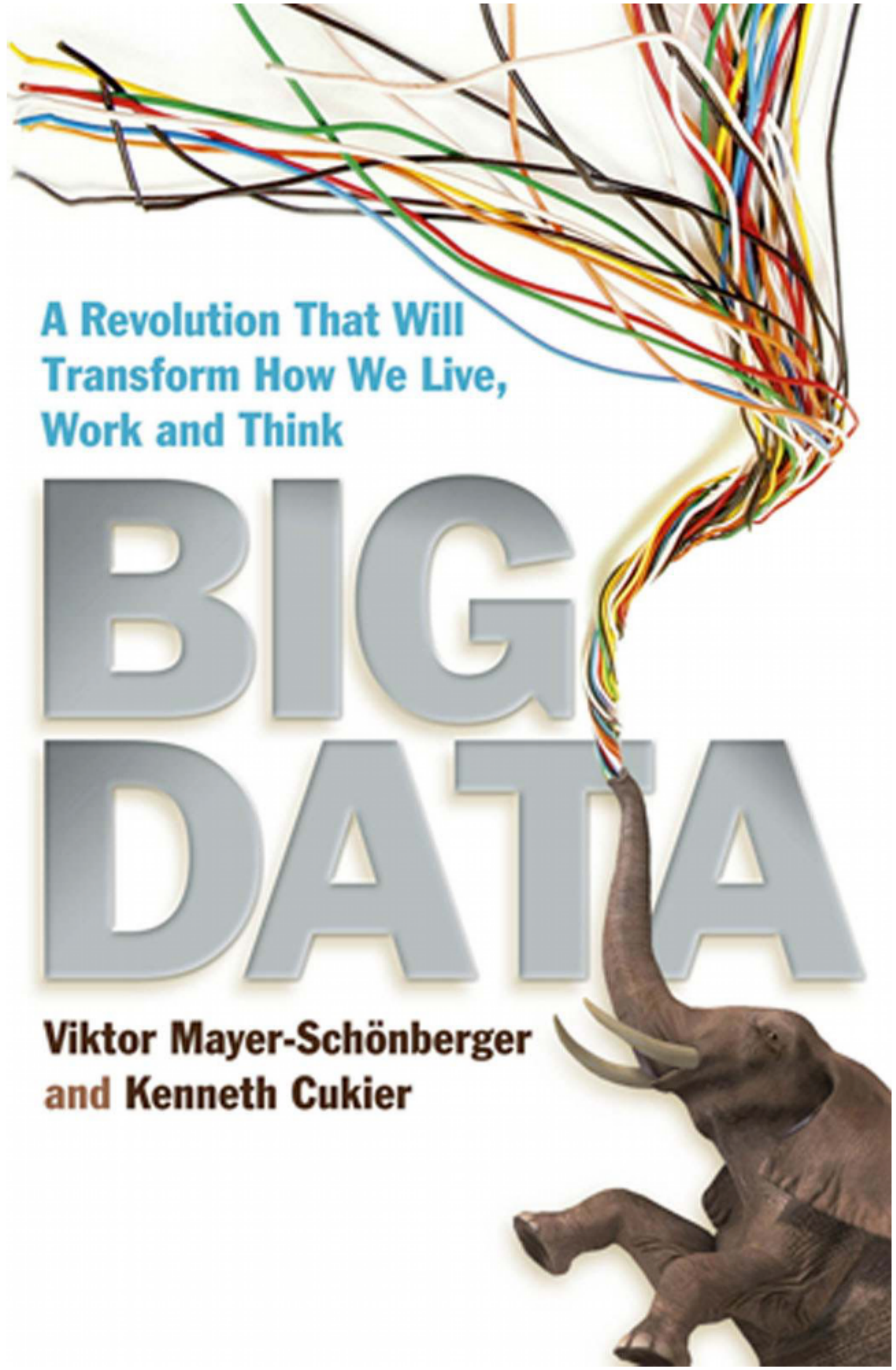
Table 1. Randomized Alphabets Used for the California Statewide Elections Since 1982

Year	Election	Randomized alphabet																									
1982	Primary	S	C	X	D	Q	G	W	R	V	Y	U	A	N	H	L	P	B	K	J	I	E	T	O	M	F	Z
	General	L	S	N	D	X	A	M	W	V	T	O	F	I	B	K	Y	U	P	E	Q	C	J	Z	H	R	G
1983	Consolidated	L	C	P	K	I	A	U	G	Z	O	N	B	X	D	W	H	E	M	F	V	R	S	T	Y	Q	J
1984	Primary	W	M	F	B	Q	Y	T	D	J	U	O	V	I	K	R	H	S	N	P	C	A	E	L	Z	G	X
	General	V	W	I	H	R	Q	G	J	O	M	T	S	Y	C	A	F	U	X	K	B	P	E	Z	N	D	L
1986	General	Q	N	H	U	B	J	E	G	M	V	L	W	X	C	K	O	F	D	Z	R	Y	I	T	S	P	A
1988	Primary	W	O	K	N	Q	A	V	T	H	J	F	Z	L	B	U	D	Y	M	I	R	G	C	E	S	X	P
	General	S	W	F	M	K	J	U	Y	A	T	V	G	O	N	Q	B	D	E	P	L	Z	C	I	X	R	H
1990	Primary	E	J	B	Y	Q	F	K	M	O	V	X	L	N	Z	C	W	A	P	R	D	G	T	H	I	S	U
	General	W	F	C	L	D	I	N	J	H	V	K	O	S	A	R	E	Q	B	T	M	Y	U	G	Z	X	P
1992	Primary	U	R	F	A	J	C	D	N	M	K	P	Z	Y	X	G	W	O	H	E	B	I	S	V	L	Q	T
	General	F	Y	U	A	J	S	B	Z	G	O	E	Q	R	L	I	M	H	V	N	T	P	D	K	X	C	W
1994	Primary	K	J	H	G	A	M	I	Q	U	N	C	Z	S	W	V	R	P	Y	B	L	O	T	D	F	E	X
	General	V	I	A	E	M	S	O	K	L	B	G	N	W	Y	D	P	U	F	Z	Q	J	X	C	R	H	T
1996	Primary	G	E	F	C	Y	P	D	B	Z	I	V	A	U	S	M	L	H	K	N	T	O	J	Q	R	X	W
	General	J	Y	E	P	A	U	S	Q	B	H	T	R	K	N	L	X	F	D	O	G	M	W	I	Z	C	V
1998	Primary	L	W	U	J	X	K	C	N	D	O	Q	A	P	T	Z	R	Y	F	E	V	B	H	G	I	M	S
	General	W	K	D	N	V	A	G	P	Y	C	Z	I	S	T	L	J	X	Q	O	F	H	R	B	U	M	E
2000	Primary	O	P	C	Y	I	H	X	Z	V	R	S	Q	E	K	L	G	D	W	J	U	T	M	B	F	A	N
	General	I	T	F	G	J	S	W	R	N	M	K	U	Y	L	D	C	Q	A	H	X	O	E	B	V	P	Z
2002	Primary	W	I	Z	C	O	M	A	Q	U	K	X	E	B	Y	N	P	T	R	L	V	S	J	H	D	F	G
	General	H	M	V	P	E	B	Q	U	G	N	D	K	X	Z	J	A	W	Y	C	O	S	F	I	T	R	L
2003	Recall	R	W	Q	O	J	M	V	A	H	B	S	G	Z	X	N	T	C	I	E	K	U	P	D	Y	F	L



Ho and Imai (2008)


1
10
12
2
4
6
7
8
11
3
5
9
2
10
8
5
Sample


A Revolution That Will Transform How We Live, Work and Think
BIG DATA
Viktor Mayer-Schönberger
and Kenneth Cukier

1936 U.S. Presidential election

v.

Predicted: Landon victory with 57% of the vote

Actual: Roosevelt victory with 61% of the vote

International Journal of Forecasting ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Forecasting elections with non-representative polls

Wei Wang[a,*], David Rothschild[b], Sharad Goel[b], Andrew Gelman[a,c]

[a] Department of Statistics, Columbia University, New York, NY, USA
[b] Microsoft Research, New York, NY, USA
[c] Department of Political Science, Columbia University, New York, NY, USA

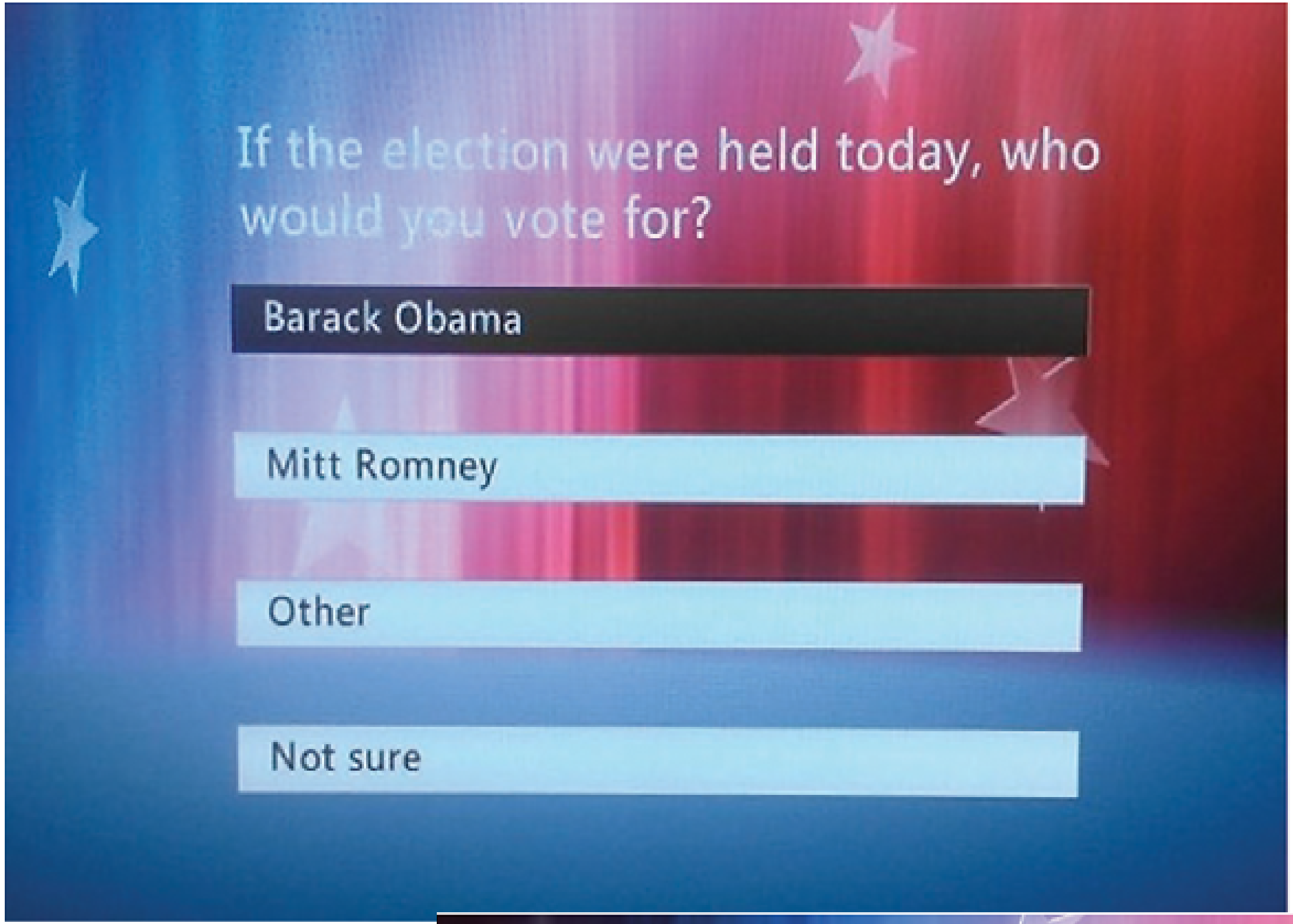
ARTICLE INFO

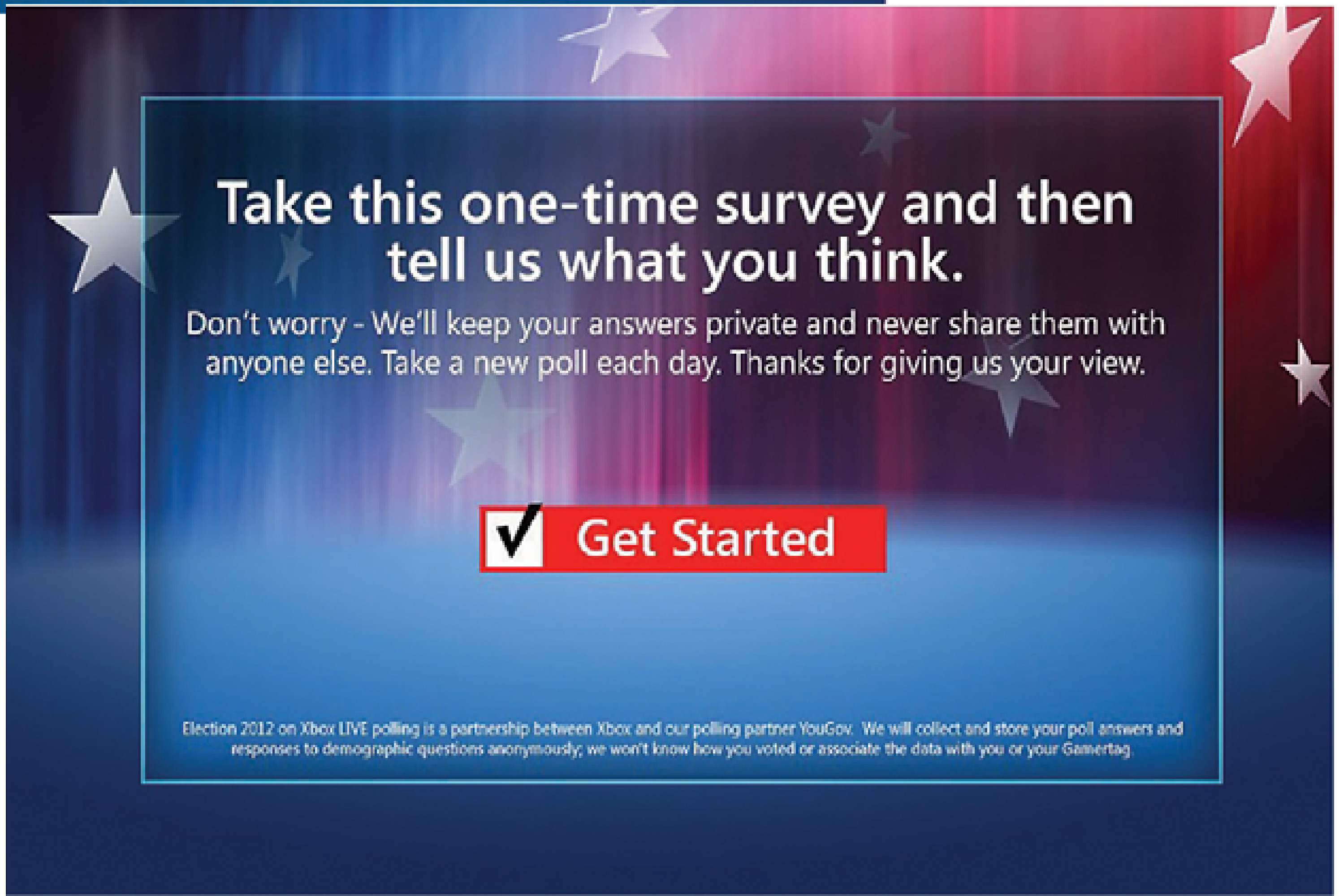
Keywords:
Non-representative polling
Multilevel regression and poststratification
Election forecasting

ABSTRACT

Election forecasts have traditionally been based on representative polls, in which randomly sampled individuals are asked who they intend to vote for. While representative polling has historically proven to be quite effective, it comes at considerable costs of time and money. Moreover, as response rates have declined over the past several decades, the statistical benefits of representative sampling have diminished. In this paper, we show that, with proper statistical adjustment, non-representative polls can be used to generate accurate election forecasts, and that this can often be achieved faster and at a lesser expense than traditional survey methods. We demonstrate this approach by creating forecasts from a novel and highly non-representative survey dataset: a series of daily voter intention polls for the 2012 presidential election conducted on the Xbox gaming platform. After adjusting the Xbox responses via multilevel regression and poststratification, we obtain estimates which are in line with the forecasts from leading poll analysts, which were based on aggregating hundreds of traditional polls conducted during the election cycle. We conclude by arguing that non-representative polling shows promise not only for election forecasting, but also for measuring public opinion on a broad range of social, economic and cultural issues.

© 2014 International Institute of Forecasters. Published by Elsevier B.V. All rights reserved.


If the election were held today, who would you vote for?
Barack Obama
Mitt Romney
Other
Not sure


Take this one-time survey and then tell us what you think.
Don't worry - We'll keep your answers private and never share them with anyone else. Take a new poll each day. Thanks for giving us your view.
Get Started
Election 2012 on Xbox LIVE polling is a partnership between Xbox and our polling partner YouGov. We will collect and store your poll answers and responses to demographic questions anonymously; we won't know how you voted or associate the data with you or your Gamertag.

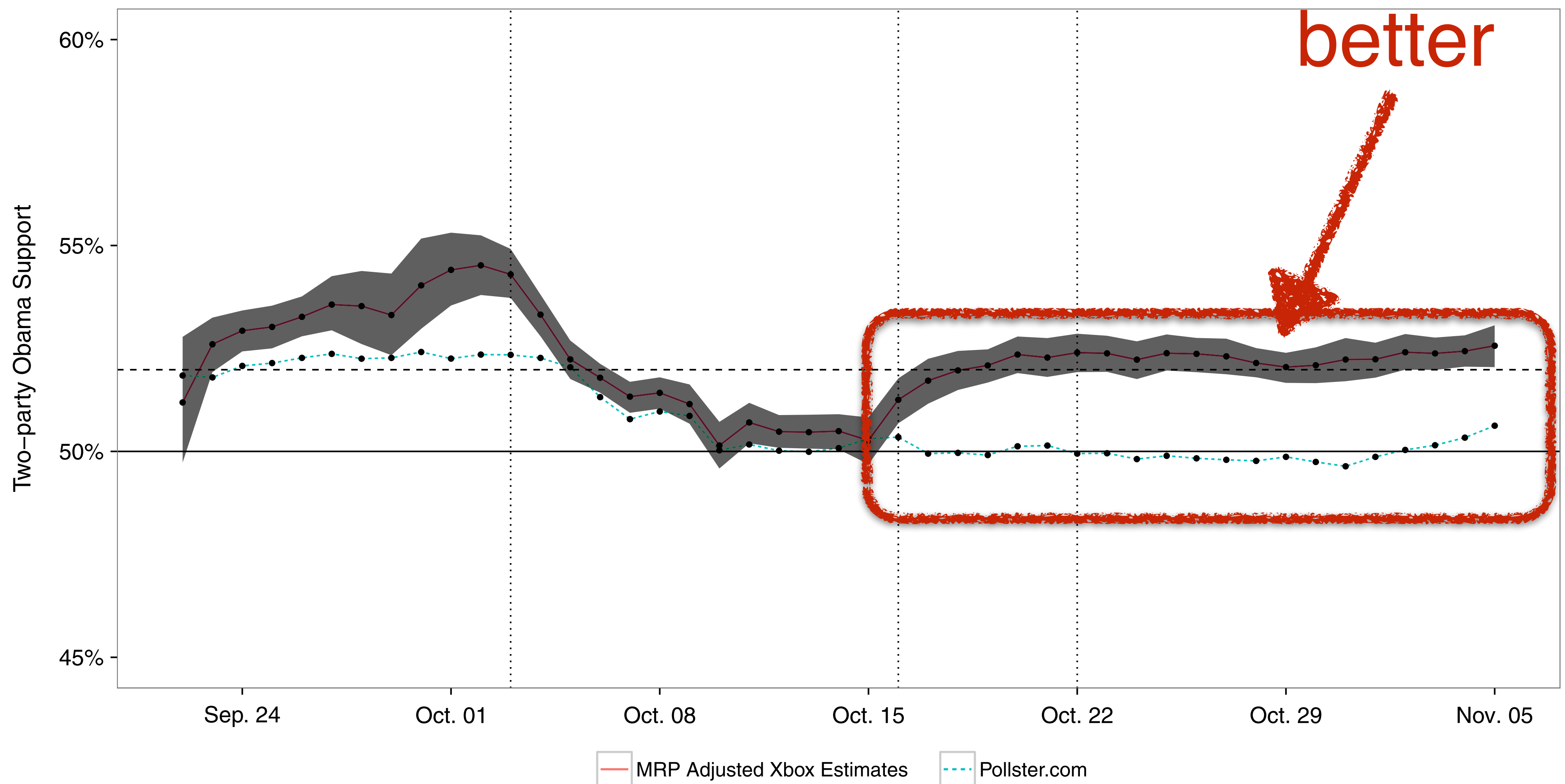
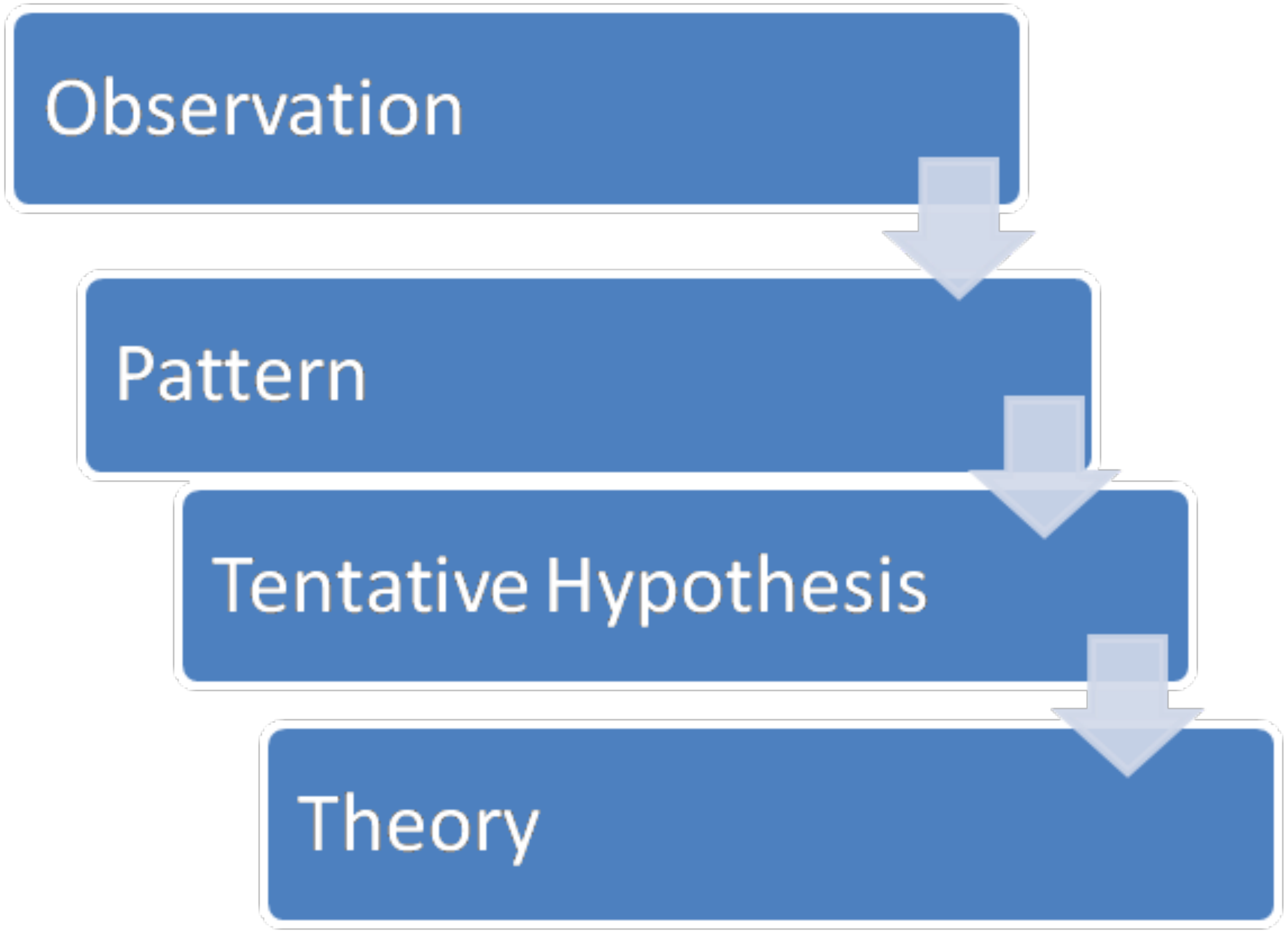


Fig. 3. National MRP-adjusted voter intent of two-party Obama support over the 45-day period, with the associated 95% confidence bands. The horizontal dashed line indicates the actual two-party Obama vote share. The three vertical dotted lines indicate the presidential debates. Compared with the raw responses in Fig. 2, the MRP-adjusted voter intent is much more reasonable, and the voter intent in the last few days is close to the actual outcome. On the other hand, the daily aggregated polling results from Pollster.com, shown by the blue dotted line, are further away from the actual vote share than the estimates generated from the Xbox data in the last few days. (For the interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)



Inductive research work flow

data science work flow?

Facebook Terms of Service:
“you grant us a non-exclusive, transferable, sub-licensable, royalty-free, worldwide **license to use any IP content that you post** on or in connection with Facebook (IP License).”

Apple EULA:
“You agree that Application Provider **may collect and use** technical data and related information, including but not limited to **technical information about Your device, system and application software, and peripherals, that is gathered periodically** to facilitate the provision of software updates, product support and other services to You (if any) related to the Licensed Application.”

PNAS PNAS PNAS PNAS

Experimental evidence of massive-scale emotional contagion through social networks

Adam D. I. Kramer[a,1], Jamie E. Guillory[b,2], and Jeffrey T. Hancock[b,c]

[a]Core Data Science Team, Facebook, Inc., Menlo Park, CA 94025; and Departments of [b]Communication and [c]Information Science, Cornell University, Ithaca, NY 14853

Edited by Susan T. Fiske, Princeton University, Princeton, NJ, and approved March 25, 2014 (received for review October 23, 2013)

Emotional states can be transferred to others via emotional contagion, leading people to experience the same emotions without their awareness. Emotional contagion is well established in laboratory experiments, with people transferring positive and negative emotions to others. Data from a large real-world social network, collected over a 20-y period suggests that longer-lasting moods (e.g., depression, happiness) can be transferred through networks [Fowler JH, Christakis NA (2008) *BMJ* 337:a2338], although the results are controversial. In an experiment with people who use Facebook, we test whether emotional contagion occurs outside of in-person interaction between individuals by reducing the amount of emotional content in the News Feed. When positive expressions were reduced, people produced fewer positive posts and more negative posts; when negative expressions were reduced, the opposite pattern occurred. These results indicate that emotions expressed by others on Facebook influence our own emotions, constituting experimental evidence for massive-scale contagion via social networks. This work also suggests that, in contrast to prevailing assumptions, in-person interaction and nonverbal cues are not strictly necessary for emotional contagion, and that the observation of others' positive experiences constitutes a positive experience for people.

computer-mediated communication | social media | big data

demonstrated that (*i*) emotional contagion occurs via text-based computer-mediated communication (7); (*ii*) contagion of psychological and physiological qualities has been suggested based on correlational data for social networks generally (7, 8); and (*iii*) people's emotional expressions on Facebook predict friends' emotional expressions, even days later (7) (although some shared experiences may in fact last several days). To date, however, there is no experimental evidence that emotions or moods are contagious in the absence of direct interaction between experiencer and target.

On Facebook, people frequently express emotions, which are later seen by their friends via Facebook's "News Feed" product (8). Because people's friends frequently produce much more content than one person can view, the News Feed filters posts, stories, and activities undertaken by friends. News Feed is the primary manner by which people see content that friends share. Which content is shown or omitted in the News Feed is determined via a ranking algorithm that Facebook continually develops and tests in the interest of showing viewers the content they will find most relevant and engaging. One such test is reported in this study: A test of whether posts with emotional content are more engaging.

The experiment manipulated the extent to which people (N = 689,003) were exposed to emotional expressions in their News Feed. This tested whether exposure to emotions led people to change their own posting behaviors, in particular whether ex-

BIOINFORMATICS ORIGINAL PAPER
Vol. 23 no. 12 2007, pages 1495–1502
doi:10.1093/bioinformatics/btm134

Gene expression

Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis

Hyunsoo Kim* and Haesun Park*

College of Computing, Georgia Institute of Technology, 266 Ferst Drive, Atlanta, GA 30332, USA

Received on November 2, 2006; revised on February 19, 2007; accepted on April 1, 2007

Advance Access publication May 5, 2007

Associate Editor: David Rocke

ABSTRACT

Motivation: Many practical pattern recognition problems require non-negativity constraints. For example, pixels in digital images and chemical concentrations in bioinformatics are non-negative. Sparse non-negative matrix factorizations (NMFs) are useful when the degree of sparseness in the non-negative basis matrix or the non-negative coefficient matrix in an NMF needs to be controlled in approximating high-dimensional data in a lower dimensional space.

Results: In this article, we introduce a novel formulation of sparse NMF and show how the new formulation leads to a convergent sparse NMF algorithm via alternating non-negativity-constrained least squares. We apply our sparse NMF algorithm to cancer-class discovery and gene expression data analysis and offer biological analysis of the results obtained. Our experimental results illustrate that the proposed sparse NMF algorithm often achieves better clustering performance with shorter computing time compared to other existing NMF algorithms.

Availability: The software is available as supplementary material.

Contact: hskim@cc.gatech.edu, hpark@acc.gatech.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

require non-negativity constraints. For example, pixels in digital images and chemical concentrations in bioinformatics are non-negative. NMF is a useful technique in approximating these high-dimensional data.

Given a non-negative matrix A of size $m \times n$, where each column of A corresponds to a data point in the m-dimensional space, and a positive integer $k < \min\{m, n\}$, NMF finds two non-negative matrices $W \in \mathbb{R}^{m\times k}$ and $H \in \mathbb{R}^{k\times n}$ so that

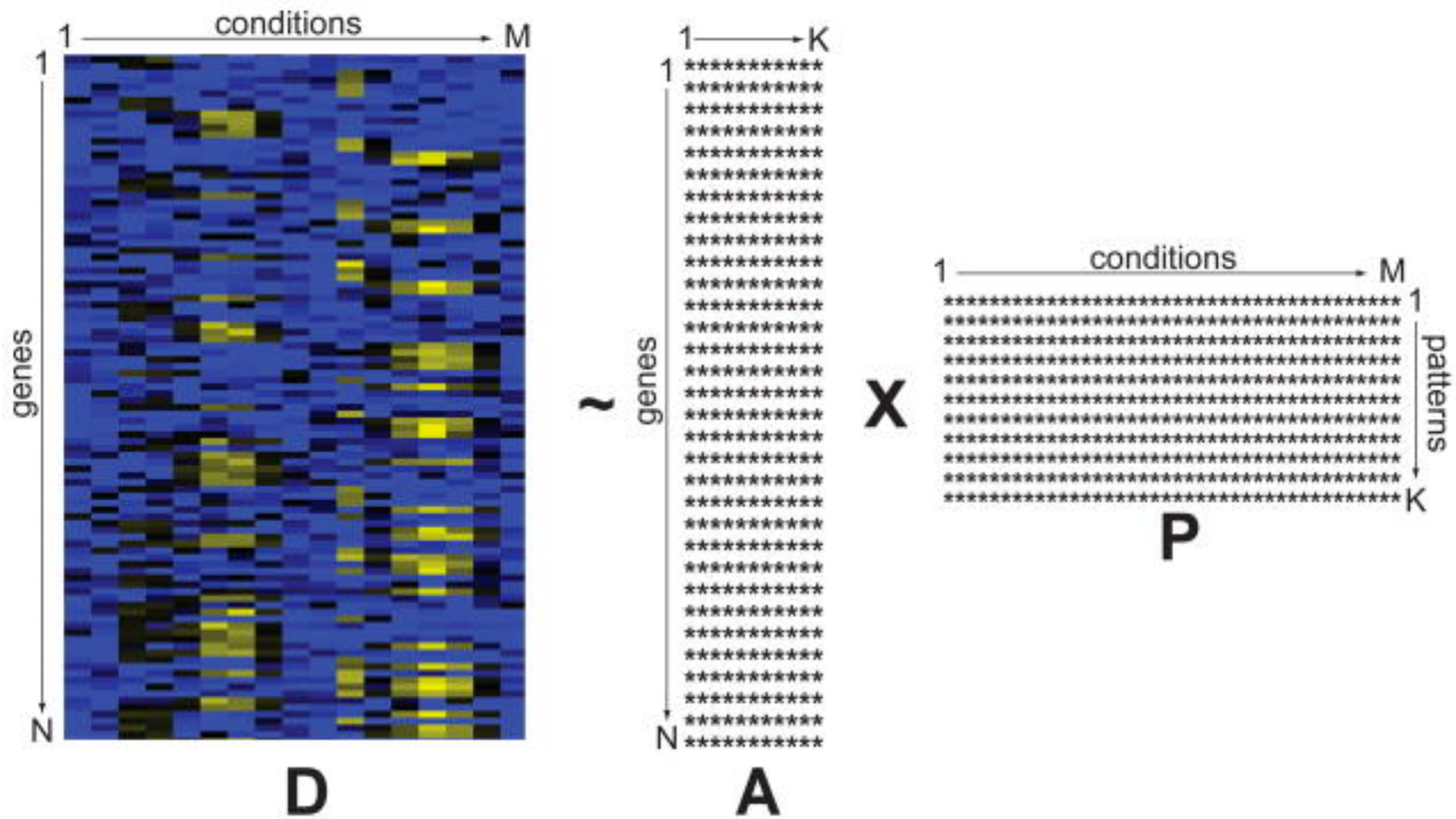
$$A \approx WH. \tag{1}$$

A solution to the NMF problem can be obtained by solving the following optimization problem:

$$\min_{W,H} f(W,H) \equiv \frac{1}{2}\|A - WH\|_F^2, \ s.t.\ W, H \geq 0, \tag{2}$$

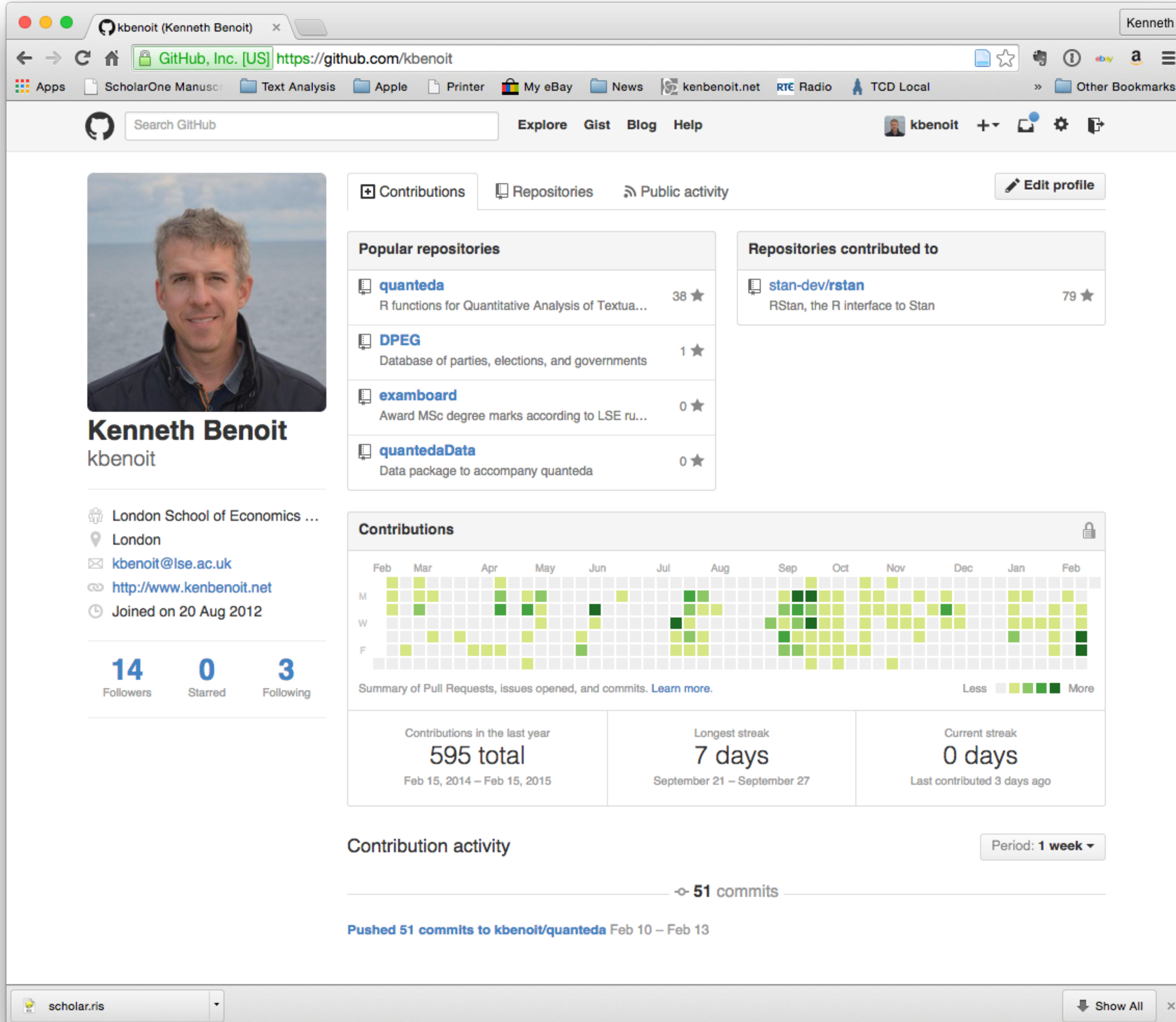
where $W \in \mathbb{R}^{m\times k}$ is a basis matrix, $H \in \mathbb{R}^{k\times n}$ is a coefficient matrix, $\|\cdot\|_F$ is the Frobenius norm and $W, H \geq 0$ means that all elements of W and H are non-negative. Due to $k < m$, dimension reduction is achieved and a lower dimensional representation of A in a k-dimensional space is given by H.

Since NMF may give us direct interpretation due to non-subtractive combinations of non-negative basis vectors, it has recently received much attention and it has been applied



Matrix Factorization of genetic microarray patterns
(Kossenkov and Ochs 2010)

```
Document-feature matrix of: 20 documents, 15 features.
20 x 15 sparse Matrix of class "dfmSparse"
                 features
docs              the  of and  to  in   a our that we  be is  it for  by have
  1789-Washington 116  71  48  48  31  14   1   18  1  23  7  11  12  20   12
  1793-Washington  13  11   2   5   3   0   0    1  0   2  0   2   1   2    1
  1797-Adams      163 140 130  72  47  51   6   22  3  31 14  34  21  30    7
  1801-Jefferson  130 104  81  61  24  21  24   24 10  15 15  13  10  16   10
  1805-Jefferson  143 101  93  83  35  20  24   37 13  22 21  21   7  22   24
  1809-Madison    104  69  43  61  34  19   9    9  2  10 10   6   5  11    8
  1813-Madison    100  65  44  42  21  25  22   10  4   6 11  18   5  13   15
  1817-Monroe     275 164 122 126  79  61  65   30 25  50 41  57  19  30   22
  1821-Monroe     360 197 141 146 136  76  60   59 18  64 33  64  30  58   43
  1825-Adams      304 245 116 101  62  28  36   36  8  14 23  19  16  38   36
  1829-Jackson     92  71  49  53  24  16  18   21  1  16 10  15  10  11    7
  1833-Jackson    101  76  53  46  23  15  19   12  5  11  6   9   8  10    8
  1837-VanBuren   252 198 150 139  76  59  60   60 30  33 16  42  27  17   33
  1841-Harrison   828 603 229 318 172 129  64  130 13 106 89 108  53 103   52
  1845-Polk       397 298 189 184  87  65 100   47  9  76 46  54  36  46   33
  1849-Taylor      99  62  52  61  20   8  15   11  3  16  6   8   6  17    3
  1853-Pierce     230 169 130 107  60  62  34   46 14  57 23  34  33  20   20
  1857-Buchanan   238 139  97 105  61  58  35   27 16  28 32  32  28  22   22
  1861-Lincoln    256 146 105 134  77  56  13   59 10  76 49  59  25  42   20
  1865-Lincoln     58  22  24  27   9   7   1   12  6   8  6  13   9   6    2
```


kbenoit (Kenneth Benoit)
Kenneth
GitHub, Inc. [US] https://github.com/kbenoit
Apps
ScholarOne Manusc
Text Analysis
Apple
Printer
My eBay
News
kenbenoit.net
Radio
TCD Local
Other Bookmarks
Search GitHub
Explore
Gist
Blog
Help
kbenoit
Contributions
Repositories
Public activity
Edit profile
Popular repositories
quanteda
R functions for Quantitative Analysis of Textua...
38
DPEG
Database of parties, elections, and governments
1
examboard
Award MSc degree marks according to LSE ru...
0
quantedaData
Data package to accompany quanteda
0
Repositories contributed to
stan-dev/rstan
RStan, the R interface to Stan
79
Kenneth Benoit
kbenoit
London School of Economics ...
London
kbenoit@lse.ac.uk
http://www.kenbenoit.net
Joined on 20 Aug 2012
14
Followers
0
Starred
3
Following
Contributions
Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec Jan Feb
M
W
F
Summary of Pull Requests, issues opened, and commits. Learn more.
Less
More
Contributions in the last year
595 total
Feb 15, 2014 – Feb 15, 2015
Longest streak
7 days
September 21 – September 27
Current streak
0 days
Last contributed 3 days ago
Contribution activity
Period: 1 week
51 commits
Pushed 51 commits to kbenoit/quanteda Feb 10 – Feb 13
scholar.ris
Show All

```r
segmentSentence <- function(x, delimiter="[.!?:;]", perl=FALSE) {
    # strip out CRs and LFs, tabs
    text <- gsub("\\n+|\\t+", " ", x)
    # remove trailing and leading spaces
    text <- gsub("^ +| +$", "", text)

    # remove . delimiter from common title abbreviations
    exceptions <- c("Mr", "Mrs", "Ms", "Dr", "Jr", "Prof", "Ph",
                    "M", "MM")
    findregex <- paste("\\b(", paste(exceptions, collapse="|"), ")\\.", sep="")
    text <- gsub(findregex, "\\1", text)

    # deal with i.e. e.g. pp. p. Cf. cf.
    text <- gsub("i\\.e\\.", "_IE_", text)
    text <- gsub("e\\.g\\.", "_EG_", text)
    text <- gsub("(\\b|\\()(p\\.)", "\\1_P_", text)
    text <- gsub("(\\b|\\()(pp\\.)", "\\1_PP_", text)
    text <- gsub("(\\b|\\()([cC]f\\.)", "\\1_CF_", text)

```

```
> tweets <- getTimeline(screen_name="kenbenoit", numResults=10,
+                       filename='~/Desktop/kentweets.json', key, cons_secret, token, access_secret)
tweets will be stored in JSON format in file: ~/Desktop/kentweets.json
authorizing...
Use a local file to cache OAuth access credentials between R sessions?
1: Yes
2: No

Selection: 2
10 tweets.
> tweets <- getTimeline(screen_name="LSEpublicevents", numResults=20,
+                       filename='~/Desktop/kentweets.json', key, cons_secret, token, access_secret)
tweets will be stored in JSON format in file: ~/Desktop/kentweets.json
authorizing...
20 tweets.
> |
```

```
{
  "created_at": [
    "Sun Feb 15 22:08:04 +0000 2015"
  ],
  "id": [
    5.6708294534182e+17
  ],
  "id_str": [
    "567082945341816832"
  ],
  "text": [
    "16\/2 @MethodologyLSE Inaugural Lecture by @kenbenoit on 'The Challenge of Big Data for the Social Sciences' #LSEdata http:\/\/t.co\/i
  ],
  "source": [
    "<a href=\"http:\/\/bufferapp.com\" rel=\"nofollow\">Buffer<\/a>"
  ],
  "truncated": [
    false
  ],
  "in_reply_to_status_id": null,
  "in_reply_to_status_id_str": null,
  "in_reply_to_user_id": null,
  "in_reply_to_user_id_str": null,
  "in_reply_to_screen_name": null,
  "user": {
    "id": [
      21643972
    ],
    "id_str": [
      "21643972"
    ],
    "name": [
      "LSE events"
    ],
    "screen_name": [
      "LSEpublicevents"
    ],
    "location": [
      "London"
    ],
    "profile_location": null,
    "description": [
      "Free public lectures and debates at LSE, with high profile speakers from government, politics, business, academia and civil society.
    ],
    "url": [
```

stackoverflow

Questions Tags Users Badges Unanswered Ask Question

Stack Overflow is a question and answer site for professional and enthusiast programmers. It's 100% free, no registration required.

Take the 2-minute tour ×

What is JSON and why would I use it?

asked 6 years ago

viewed 165031 times

active 3 months ago

277

190

I've looked on wikipedia and Googled it and read the official documentation, but I still haven't got to the point where I really understand what JSON is, and why I'd use it.

I have been building applications using PHP, MySQL and Javascript / HTML for a while, and if JSON can do something to make my life easier or my code better or my user interface better, then I'd like to know about it. Can someone give me a succinct explanation?

json

share improve this question

edited Apr 9 '14 at 11:02
lev
3 ● 2

asked Dec 20 '08 at 20:19
Ben
11.5k ● 19 ● 60 ● 87

1 What are you currently using for Ajax? – Todd Smith Dec 20 '08 at 20:45

JSON is a subset of YAML yaml.org – Brad Gilbert Dec 21 '08 at 2:15

8 copterlabs.com/blog/json-what-it-is-how-it-works-how-to-use-it this is a nice straightforward example of usage – Tom Jun 3 '12 at 19:46

add a comment

Jobs near you

Senior Java Developer who loves FOOD
FOODit
London, UK
java tdd

Full-Stack Developer
TransferWise
London, UK / relocation
html css

More jobs near Redhill...

Linked

-3 How to acces items in deeply nested JSON object?

15 Get user's name from Facebook Graph API

15 Actionresult vs JSONresult

0 What is JSON used for in web applications?

1 Convert Javascript string or array into JSON object

1 what is the use of json in .net

0 Why do we use JSON in

10 Answers

active oldest votes

384

JSON (JavaScript Object Notation) is a lightweight format that is used for data interchanging. It is also a subset of JavaScript's Object Notation (the way `objects` are built in JavaScript)

An example of where this is used is web services responses. In the 'old' days, web services used XML as their primary data format for transmitting back data, but since JSON appeared (*The JSON format is specified in* RFC 4627 *by Douglas Crockford*), it has been the preferred format because it is much more **lightweight**

The Challenge of Big Data for the Social Sciences

Kenneth Benoit 2015

Department of Methodology Inaugural Lecture

LSE Department of Methodology

The Challenge of Big Data for the Social Sciences

Professor Kenneth Benoit
Professor of Quantitative Social Research Methods, LSE

Kenneth Cukier
Data Editor,
The Economist

Professor Simon Hix
Chair, LSE

LSE events

Suggested hashtag for Twitter users: **#LSEdata**

CPD CERTIFIED
The CPD Certification Service