

The use of news alerts, Twitter microblogs, and Google searches as a complementary way to track disease activity

Case studies characterizing flu, dengue, Zika, and Ebola epidemics.
Fondation Mérieux, December 6th, 2017



Mauricio Santillana, PhD

Assistant Professor, Harvard Medical School
Faculty member, CHIP, Boston Children's Hospital Informatics Program
Associate, Harvard Institute for Computational and Applied Sciences



HARVARD
MEDICAL SCHOOL



HARVARD
School of Engineering
and Applied Sciences



**Boston
Children's
Hospital**

The use of news alerts, Twitter microblogs, and Google searches as a complementary way to track disease activity

Case studies characterizing flu, dengue, Zika, and Ebola epidemics.
Fondation Mérieux, December 6th, 2017



Collaborators: Sam C. Kou (Harvard Statistics), Shihao Yang (Harvard Statistics), Fred Lu (BCH), Nicholas Brooke (Break Dengue), Matt Biggerstaff (CDC), Julia Gunn (Boston Public Health Commission), Joe Conidi (Boston Public Health Commission), Michael Johansson (CDC), Nick Reich (Umass Amherst), Roni Rosenfeld (CMU), Kristin Baltrusaitis (Boston Univ), Alessandro Vespignani (Northeastern Univ), Nathan Kutz (Univ of Washington), Elaine Nsoesie (Univ of Washington), Rumi Chunara (NYU), John Brownstein (Harvard/BCH), Sarah McGough (Harvard), Leonardo C. Clemente (Inst. Politécnico Nacional, Mex), and many more



HARVARD
MEDICAL SCHOOL



HARVARD
School of Engineering
and Applied Sciences



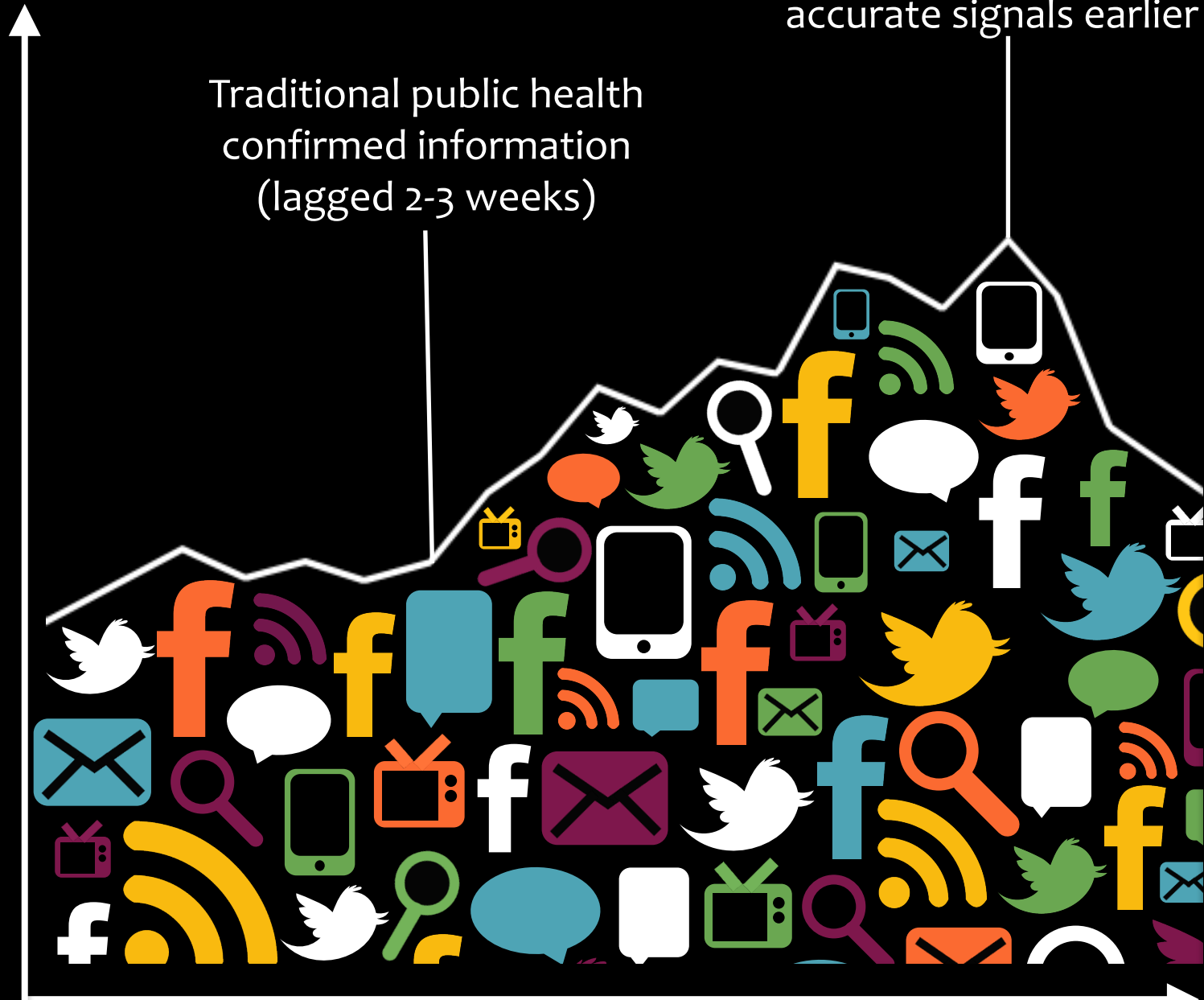
**Boston
Children's
Hospital**



Can Digital disease detection pick up accurate signals earlier?

Traditional public health confirmed information (lagged 2-3 weeks)

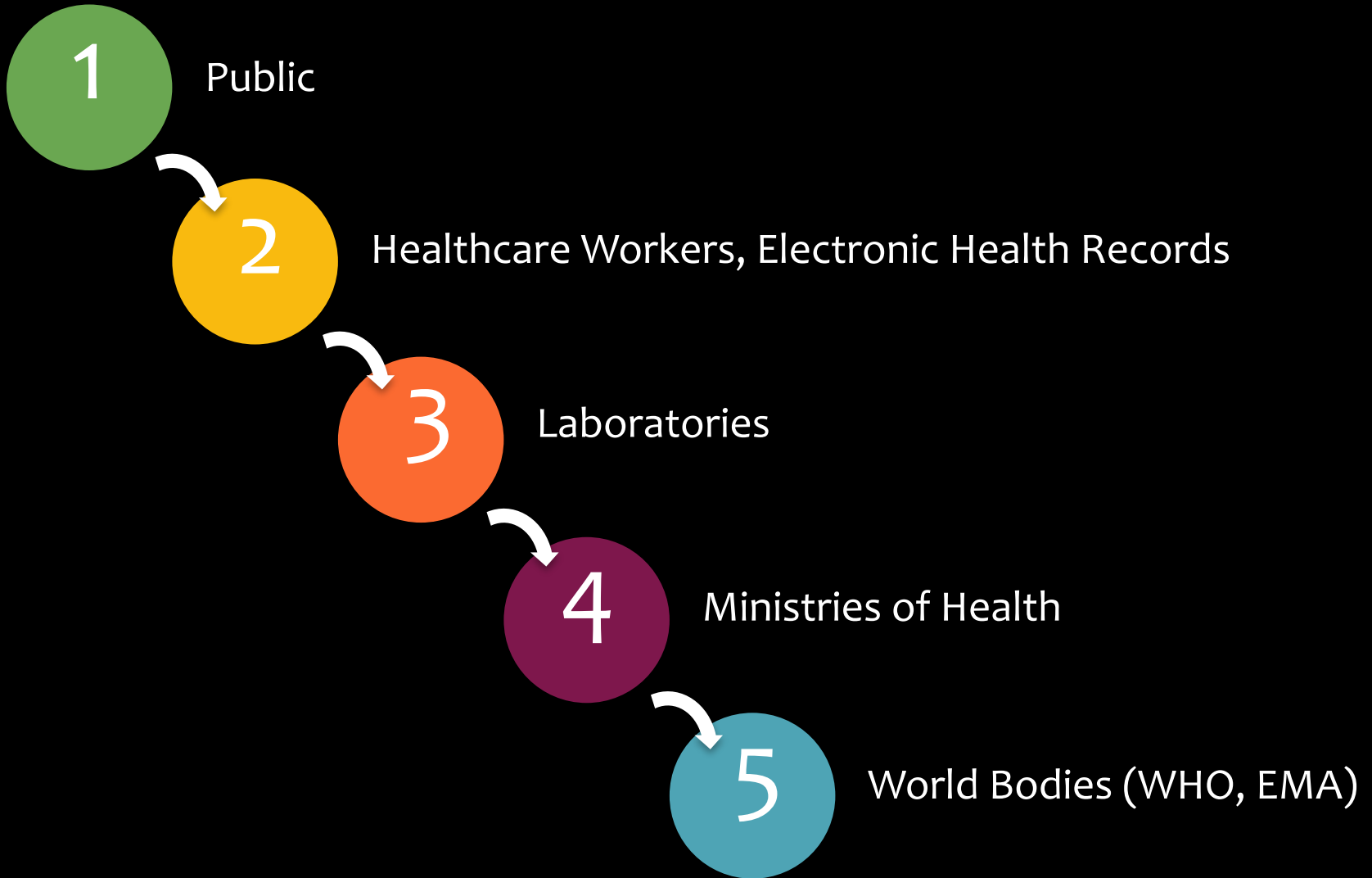
disease incidence

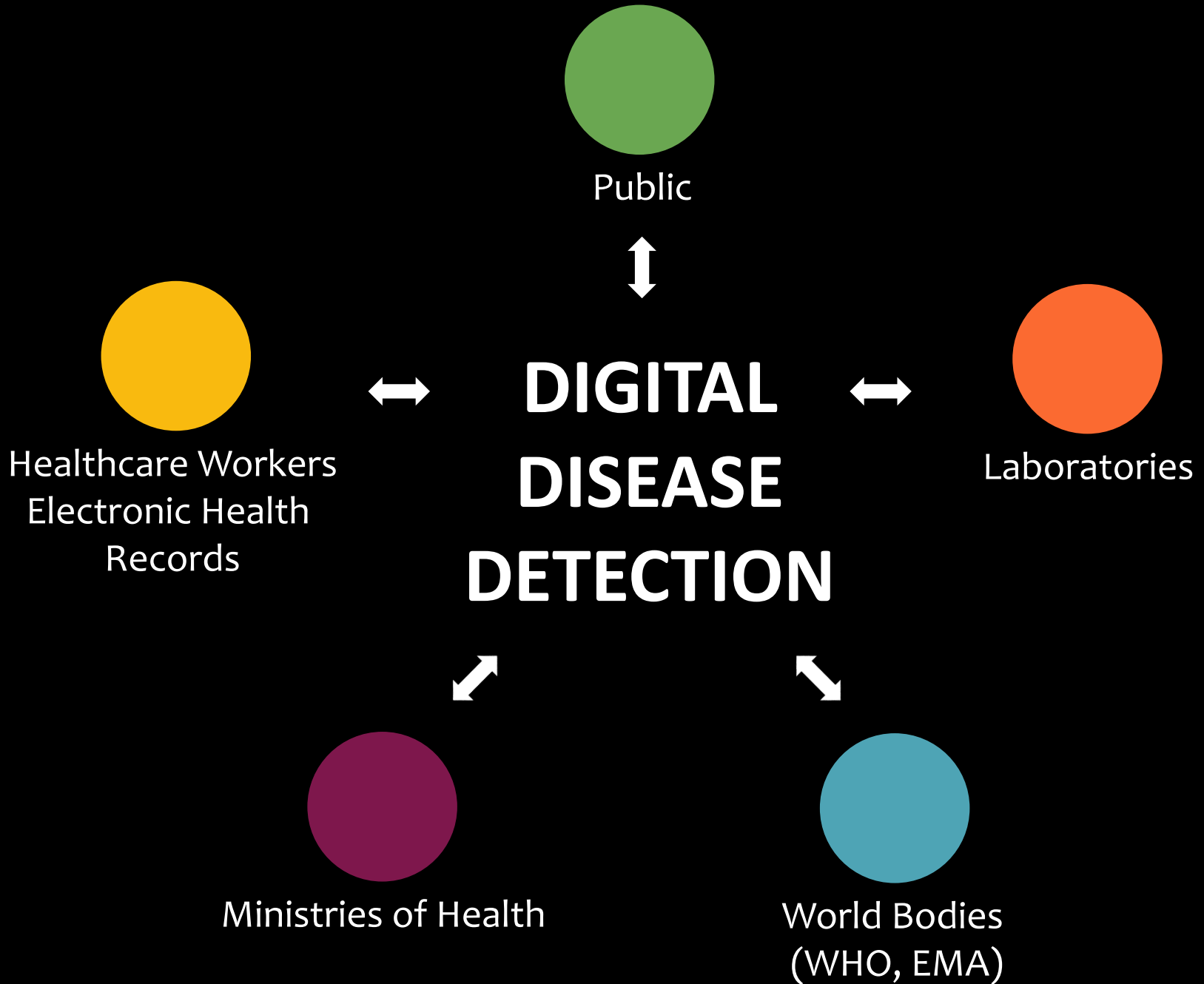


time

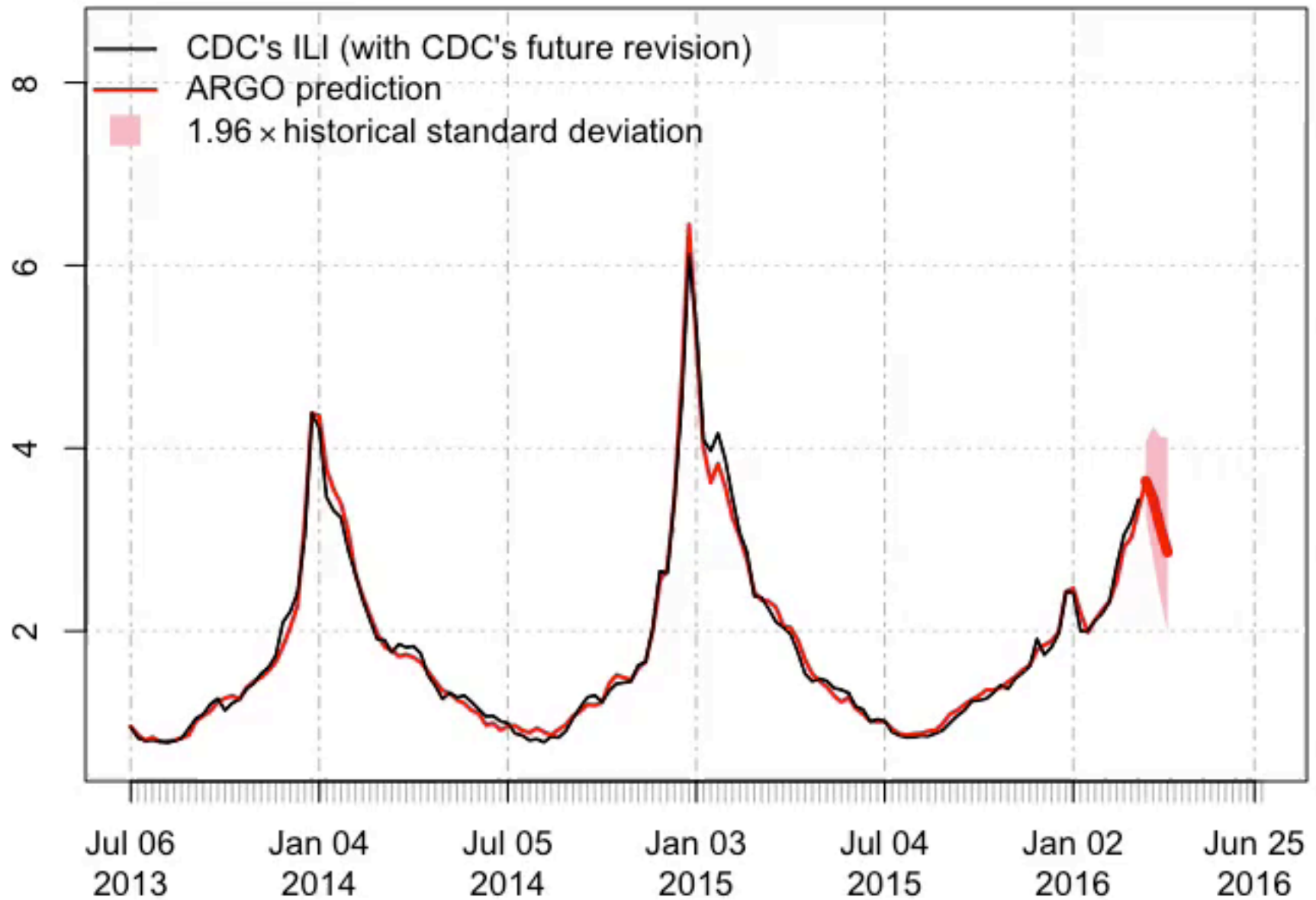


TRADITIONAL DISEASE REPORTING



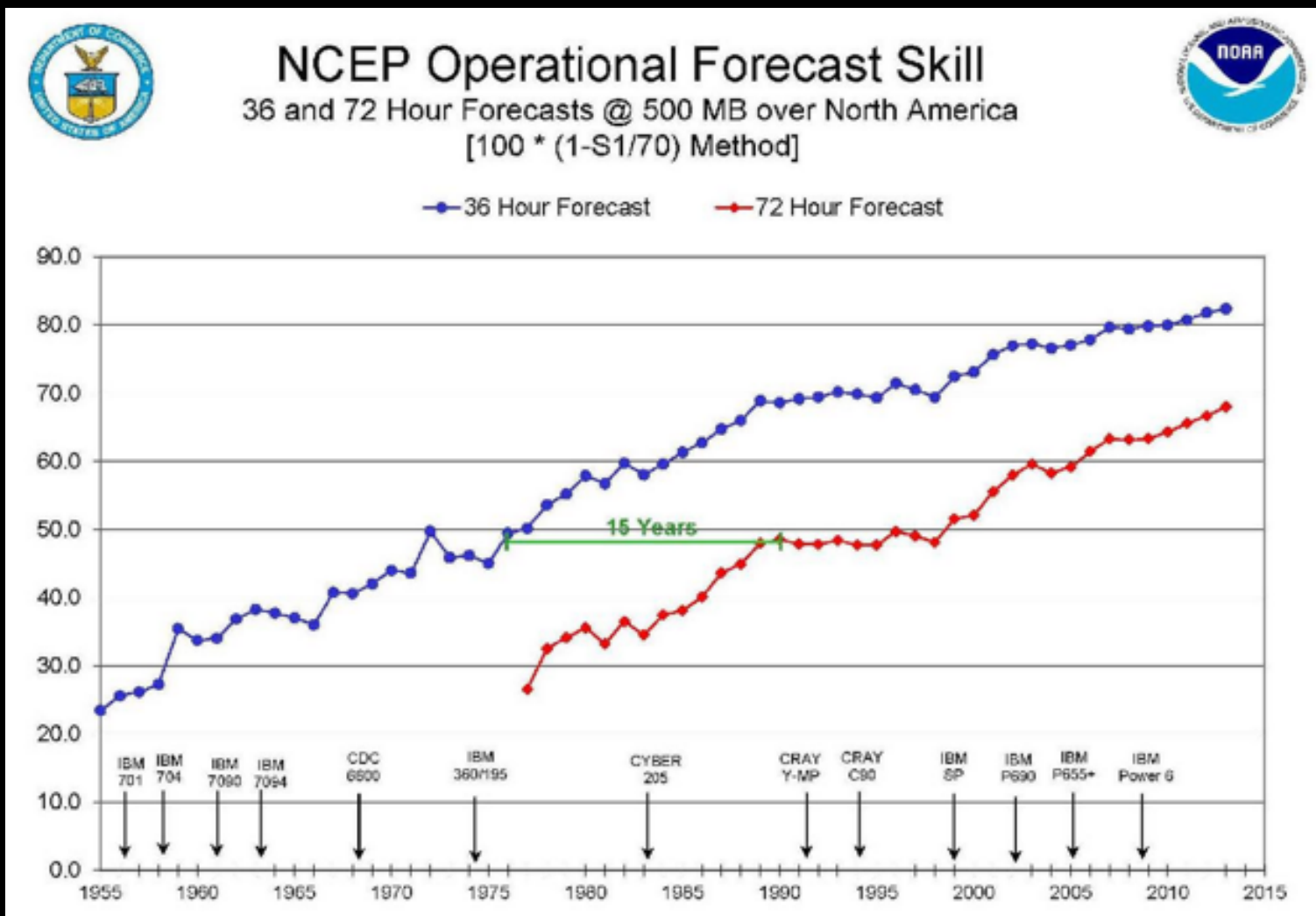


ARGO Prediction vs. CDC's ILI

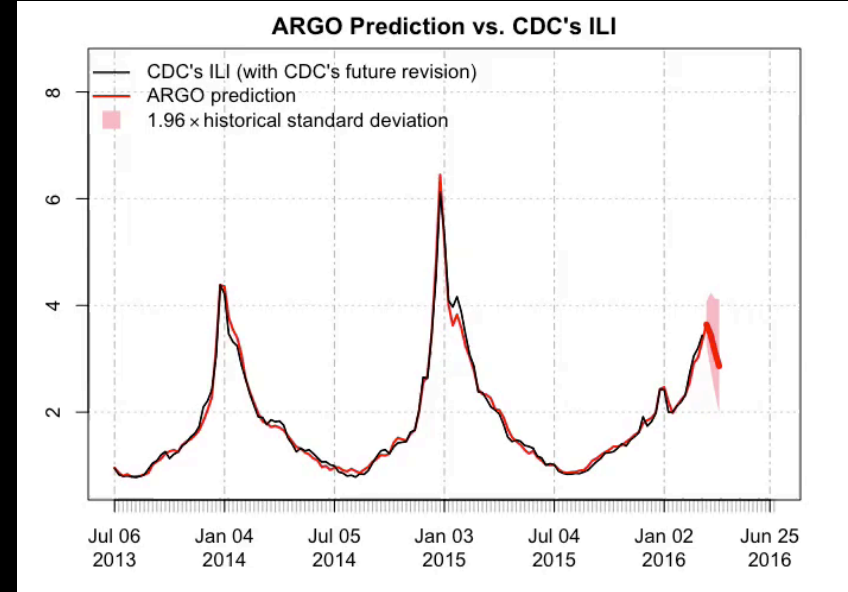
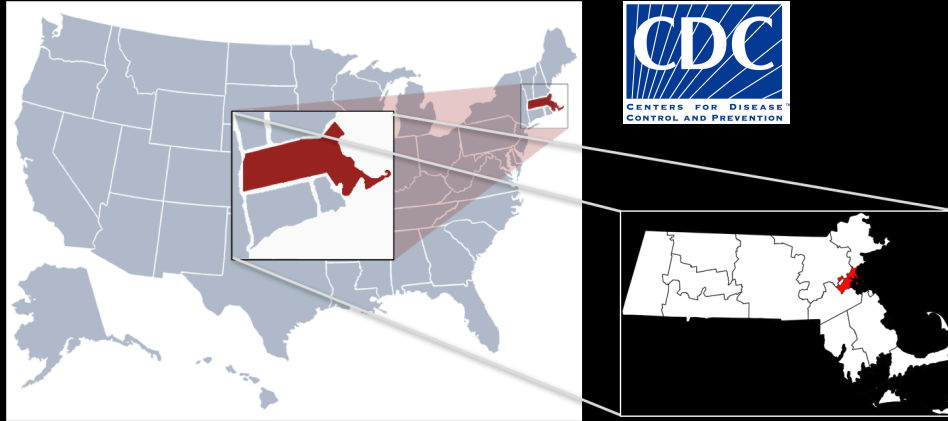


Real-time tracking vs predictions of disease incidence/risk

Similarities and differences with weather prediction



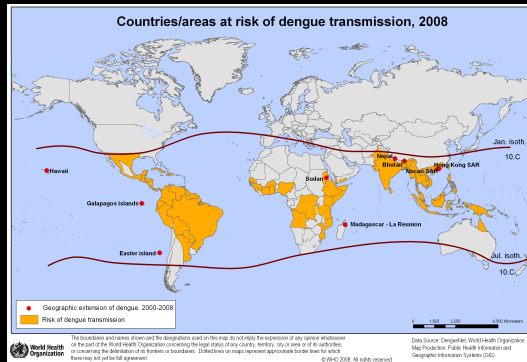
Part 1. Previous success stories in tracking and forecasting Influenza in data-rich high-income countries: USA



1. Multiple spatial resolutions: National, multi-state, state, city-level
2. Multiple data sources (hybrid systems): traditional healthcare-based, EHR, Google, Twitter, Crowd-sourced disease surveillance.

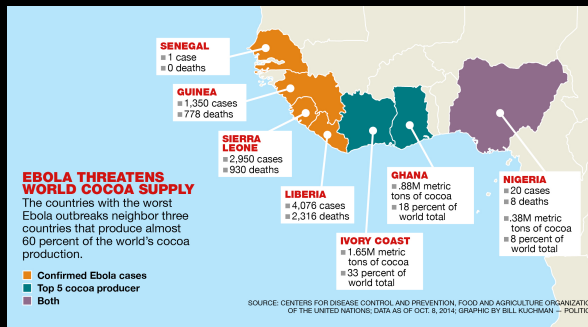
Part 2. Success stories in tracking and forecasting Flu, Zika, Dengue, Ebola in data-poor medium- to low-income countries.

Dengue, Zika, and Flu



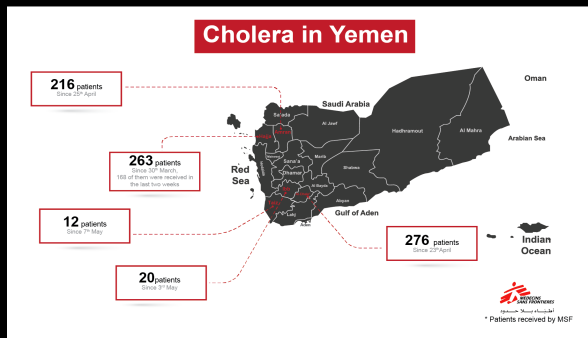
- Latin America (Flu, Zika, Dengue)
- South-east Asia (Dengue)

Ebola



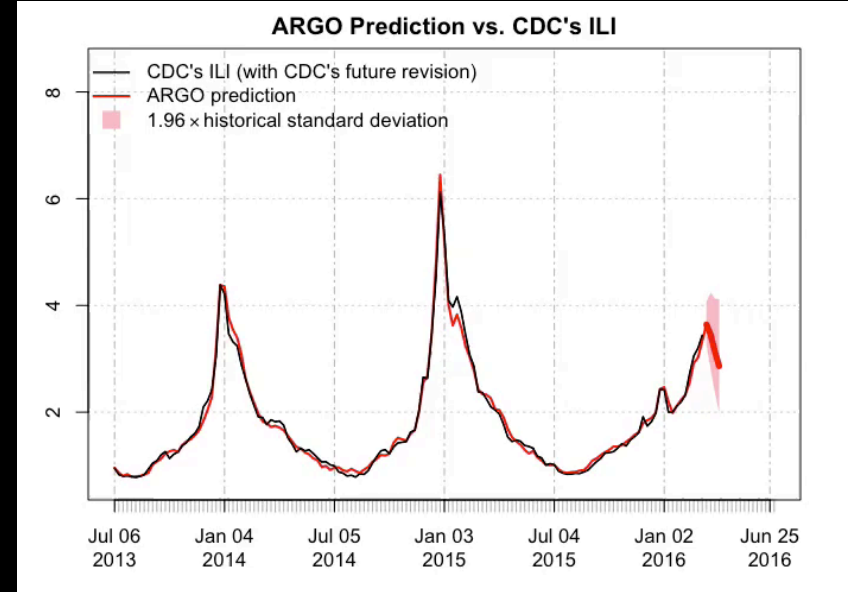
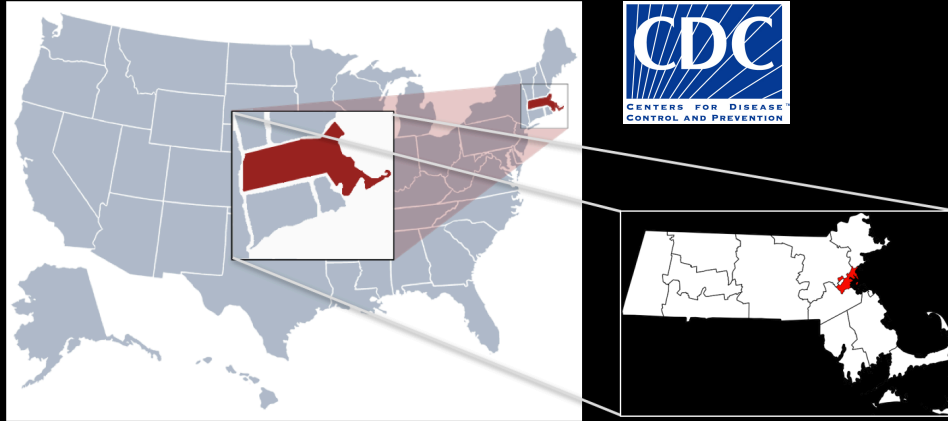
- West Africa

Cholera



- Middle East

Part 1. Previous success stories in tracking and forecasting Influenza in data-rich high-income countries: USA



1. Multiple spatial resolutions: National, multi-state, state, city-level
2. Multiple data sources (hybrid systems): traditional healthcare-based, EHR, Google, Twitter, Crowd-sourced disease surveillance.

Seminal work by Google

The promise of big data in public health

GOOGLE FLU TRENDS

Letter

Nature 457, 1012-1014 (19 February 2009) | doi:10.1038/nature07634; Received 14 August 2008; Accepted 13 November 2008; Published online 19 November 2008; [Corrected](#) 19 February 2009

Detecting influenza epidemics using search engine query data

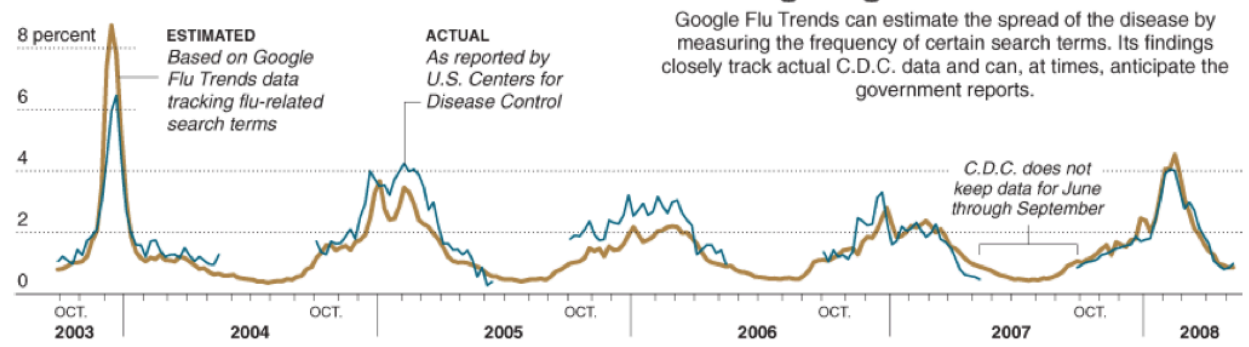
Jeremy Ginsberg¹, Matthew H. Mohebbi¹, Rajan S. Patel¹, Lynnette Brammer², Mark S. Smolinski¹ & Larry Brilliant¹

- 1. Google Inc., 1600 Amphitheatre Parkway, Mountain View, California 94043, USA
- 2. Centers for Disease Control and Prevention, 1600 Clifton Road, NE, Atlanta, Georgia 30333, USA

Correspondence to: Matthew H. Mohebbi¹ Correspondence and requests for materials should be addressed to J.G. or M.H.M. (Email: flutrends-support@google.com).

The New York Times

PERCENT OF HEALTH VISITS FOR FLU-LIKE SYMPTOMS Mid-Atlantic region



Sources: Google; Centers for Disease Control

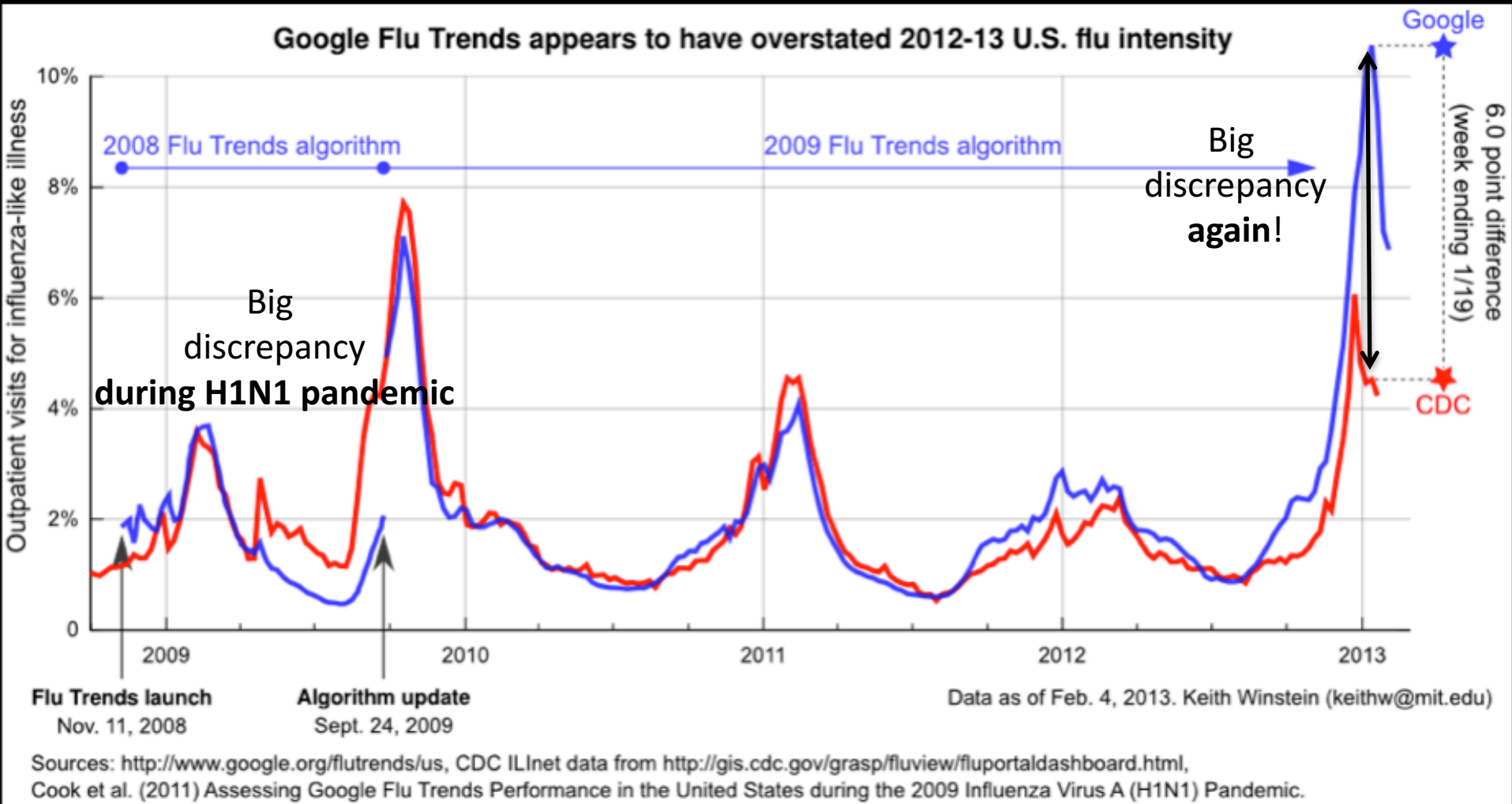
THE NEW YORK TIMES

Using Google to Monitor the Flu

Google Flu Trends can estimate the spread of the disease by measuring the frequency of certain search terms. Its findings closely track actual C.D.C. data and can, at times, anticipate the government reports.

What next? need to remove (not useful) terms.

Big discrepancies again!



Fixes were reported in: Cook et al. (2011) Assessing Google flu trends performance in the U.S. during the 2009 influenza virus A (H1N1) pandemic. PLoS One

Plot obtained from: <http://blog.keithw.org/2013/02/q-how-accurate-is-google-flu-trends.html>

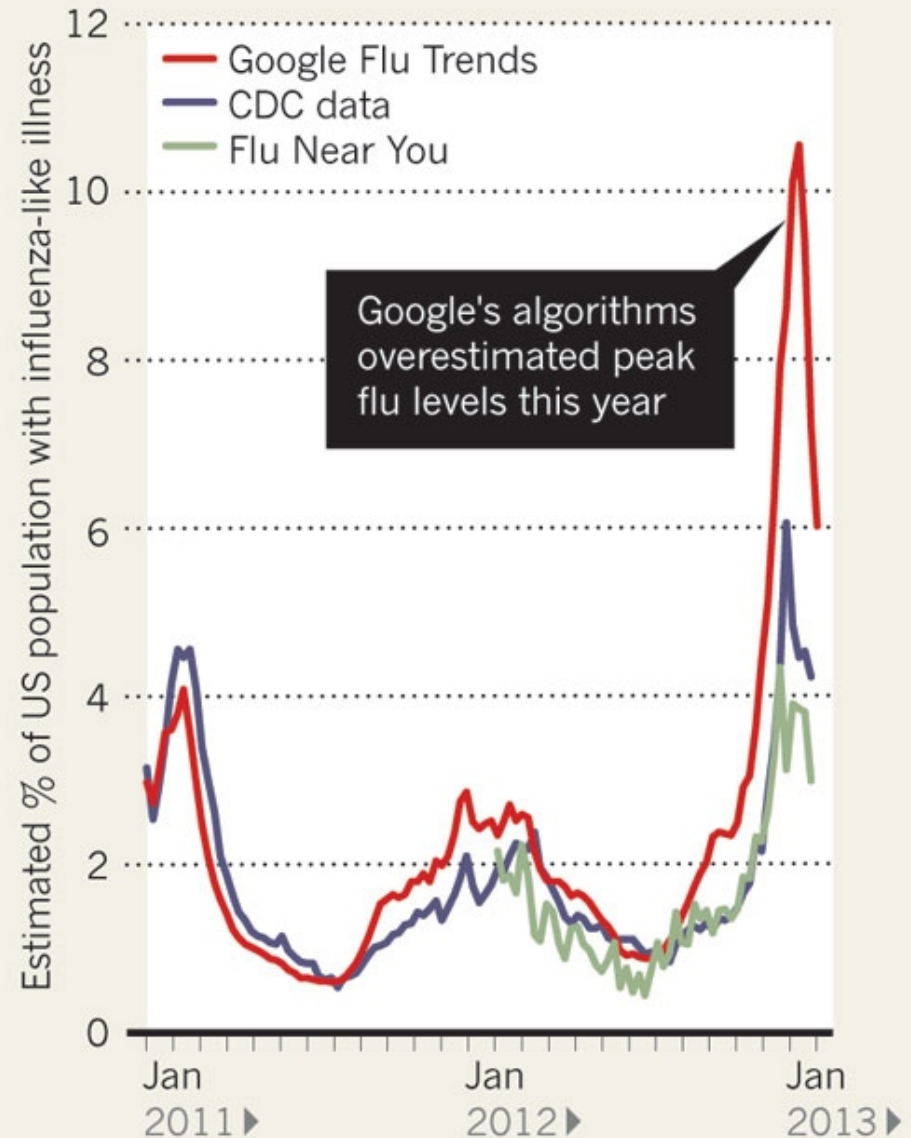


When Google got flu wrong.

nature.com/news/when-google-got-flu-wrong.

FEVER PEAKS

A comparison of three different methods of measuring the proportion of the US population with an influenza-like illness.



Snowden And The Challenge Of Intelligence: The Practical Case Against The NSA's Big Data

63

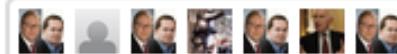
f Share

106

🐦 Tweet

61

in Share



12 comments, 7 called-out

+ Comment Now

+ Follow Comments

“ We should soon be able to keep track of most activities on the surface of the earth, day or night, in good weather or bad.”

s i l i c o n A N G L E

where computer science meets social science

{SILICON ANGLE}

CLOUD

MOBILE

SOCIAL

SERVICES

DEVOPS

RESEARCH

SiliconANGLE » Can Nate Silver's Data Culture Lead Us Out Of The NSA + Public Data Scare?

Can Nate Silver's Data Culture Lead Us Out of the NSA + Public Data Scare?

RYAN COX | SEPTEMBER 18TH

READ MORE

hive | A

il flu.

ancy

What Can Digital Disease Detection Learn from (an External Revision to) Google Flu Trends?

Mauricio Santillana, PhD, MS, D. Wendong Zhang, MA, Benjamin M. Althouse, PhD, ScM, John W. Ayers, PhD, MA

© 2014 Published by Elsevier Inc. on behalf of American Journal of Preventive Medicine Am J Prev Med 2014;47(3):341-347 341

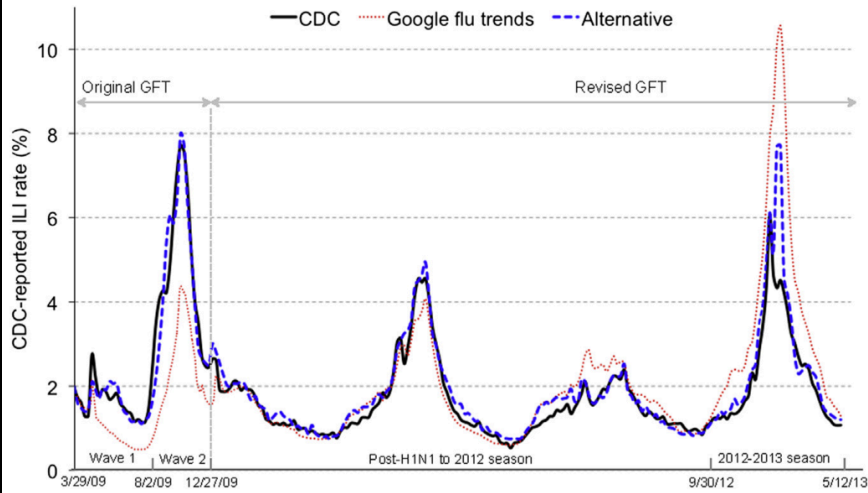


Figure 1. The alternative model outperforms Google Flu Trends

Google incorporated our proposed changes to GFT's engine in Oct 2014

We published a paper proposing changes to GFT's engine (2014)



Google Research Blog

The latest news from Research at Google

Google Flu Trends gets a brand new engine

Posted: Friday, October 31, 2014

222
 161
 104

Posted by Christian Stefansen, Senior Software Engineer


Each year the flu kills thousands of people and affects millions around the world. So it's important that public health officials and health professionals learn about outbreaks as quickly as possible. In 2008 we launched [Google Flu Trends](#) in the U.S., using aggregate web searches to indicate when and where influenza was striking in real time. These models [nicely complement](#) other survey systems—they're more fine-grained geographically, and they're typically more immediate, up to 1-2 weeks ahead of traditional methods such as the CDC's official reports. They can also be incredibly helpful for countries that don't have official flu tracking. Since launching, we've expanded Flu Trends to cover 29 countries, and launched [Dengue Trends](#) in 10 countries.

The original model performed surprisingly well despite its simplicity. It was retrained just once per year, and typically used only the 50 to 300 queries that produced the best estimates for prior seasons. We then left it to perform through the new season and evaluated it at the end. It didn't use the official CDC data for estimation during the season—only in the initial training.

SCIENTIFIC REPORTS PDF

Article | OPEN

Advances in nowcasting influenza-like illness rates using search query logs

Vasileios Lamos , Andrew C. Miller, Steve Crossan & Christian Stefansen

Scientific Reports **5**,
 Article number: 12760 (2015)
 doi:10.1038/srep12760


Received: 07 May 2015
 Accepted: 06 July 2015
 Published online: 03 August 2015

[Download Citation](#)

Applied mathematics
 Computer science Epidemiology
 Influenza virus

Google and collaborators published a paper improving our AJPM 2014 methodology in August 2015

We improved last effort by Google team and published our results in PNAS in September 2015

 CrossMark
 ← click for updates

Accurate estimation of influenza epidemics using Google search data via ARGO

Shihao Yang^a, Mauricio Santillana^{b,c,1}, and S. C. Kou^{a,1}

^aDepartment of Statistics, Harvard University, Cambridge, MA 02138; ^bSchool of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138; and ^cComputational Health Informatics Program, Boston Children's Hospital, Boston, MA 02115

Edited by Wing Hung Wong, Stanford University, Stanford, CA, and approved September 30, 2015 (received for review August 6, 2015)

Accurate real-time tracking of influenza outbreaks helps public health officials make timely and meaningful decisions that could save lives. We propose an influenza tracking model, ARGO (AutoRegression with Google search data), that uses publicly available online search data. In addition to having a rigorous statistical foundation, ARGO outperforms all previously available Google-search-based tracking models, including the latest version of Google Flu Trends, even though it uses only low-quality search data as input from publicly available Google Trends and Google Correlate websites. ARGO not only incorporates the seasonality in influenza epidemics but also captures changes in people's online search behavior over time. ARGO is also flexible, self-correcting, robust, and scalable, making it a potentially powerful tool that can be used for real-time tracking of other social events at multiple temporal and spatial resolutions.

CDC's ILI reports have a delay of 1–3wk due to the time for processing and aggregating clinical information. This time lag is far from optimal for decision-making purposes. To alleviate this information gap, multiple methods combining climate, demographic, and epidemiological data with mathematical models have been proposed for real-time estimation of flu activity (18, 21–25). In recent years, methods that harness Internet-based information have also been proposed, such as Google (1), Yahoo (2), and Baidu (3) Internet searches, Twitter posts (4), Wikipedia article views (5), clinicians' queries (6), and crowdsourced self-reporting mobile apps such as Influenzanet (Europe) (26), Flutracking (Australia) (27), and Flu Near You (United States) (28). Among them, GFT has received the most attention and has inspired subsequent digital disease detection systems (3, 8,

PNAS PNAS

APPLIED
HEMATICS

SCIENTIFIC REPORTS PDF

Article | OPEN

Advances in nowcasting influenza-like illness rates using search query logs

Vasileios Lamos, Andrew C. Miller, Steve Crossan & Christian Stefansen

Scientific Reports **5**,
 Article number: 12760 (2015)
 doi:10.1038/srep12760

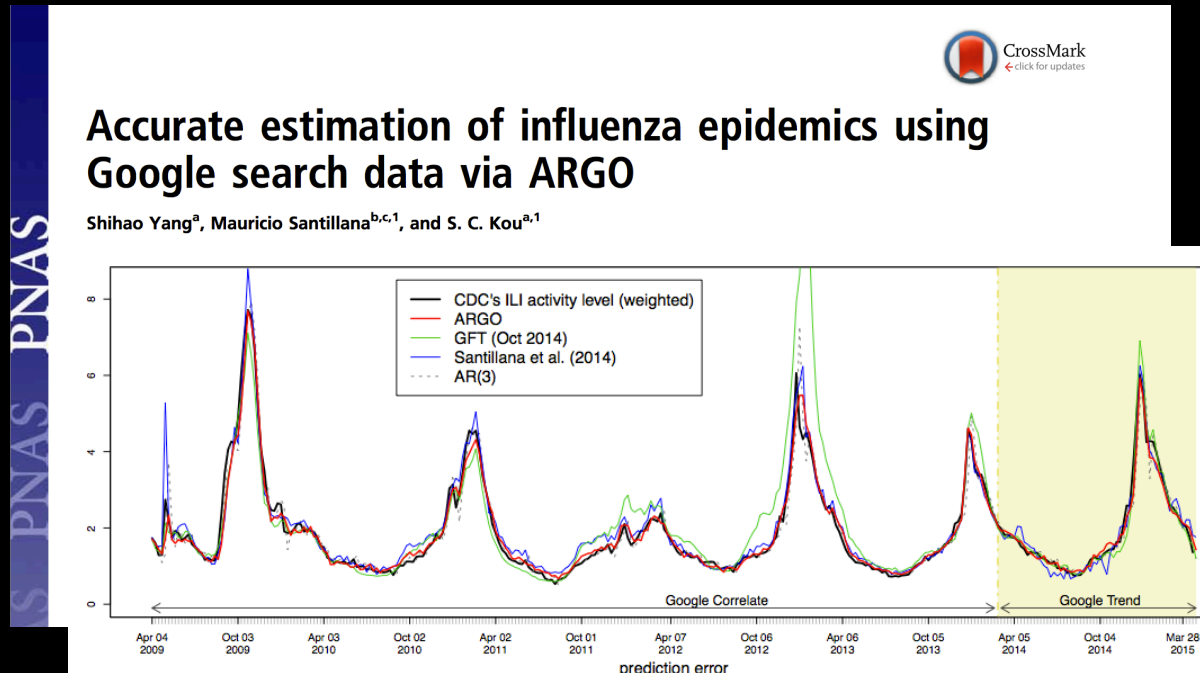
Received: 07 May 2015
 Accepted: 06 July 2015
 Published online: 03 August 2015

Download Citation

Applied mathematics
 Computer science Epidemiology
 Influenza virus

Google and collaborators published a paper improving our AJPM 2014 methodology in August 2015

We improved last effort by Google team and published our results in PNAS in September 2015



Google discontinues Flu Trends indefinitely!



Google Research Blog

The latest news from Research at Google

The Next Chapter for Flu Trends

Posted: Thursday, August 20, 2015

  7



Instead of maintaining our own website going forward, we're now going to empower institutions who specialize in infectious disease research to use the data to build their own models. Starting this season, we'll provide Flu and Dengue signal data directly to partners including [Columbia University's Mailman School of Public Health](#) (to update their [dashboard](#)), [Boston Children's Hospital/Harvard](#), and [Centers for Disease Control and Prevention \(CDC\) Influenza Division](#). We will also continue to make historical Flu and Dengue estimate data available for anyone to see and analyze.

NEWS

Google Flu Trends calls out sick, indefinitely

Google will pass along search queries related to the flu to health organizations so they can develop their own prediction models

By [Fred O'Connor](#) | [Follow](#)

IDG News Service | Aug 20, 2015 2:07 PM PT

MORE LIKE THIS ::

[Google Begins Tracking Swine Flu in Mexico](#)



[Google's Panicky Flu Estimates Were Dead Wrong](#)

BIG DATA

Google discontinues Flu Trends, starts offering data to researchers

JORDAN NOVET | AUGUST 20, 2015 12:17 PM

TAGS: [GOOGLE](#), [GOOGLE FLU TRENDS](#)

Our team at Boston Children's Hospital now has access to Google's search volumes, as one of the exclusive Google's partners.

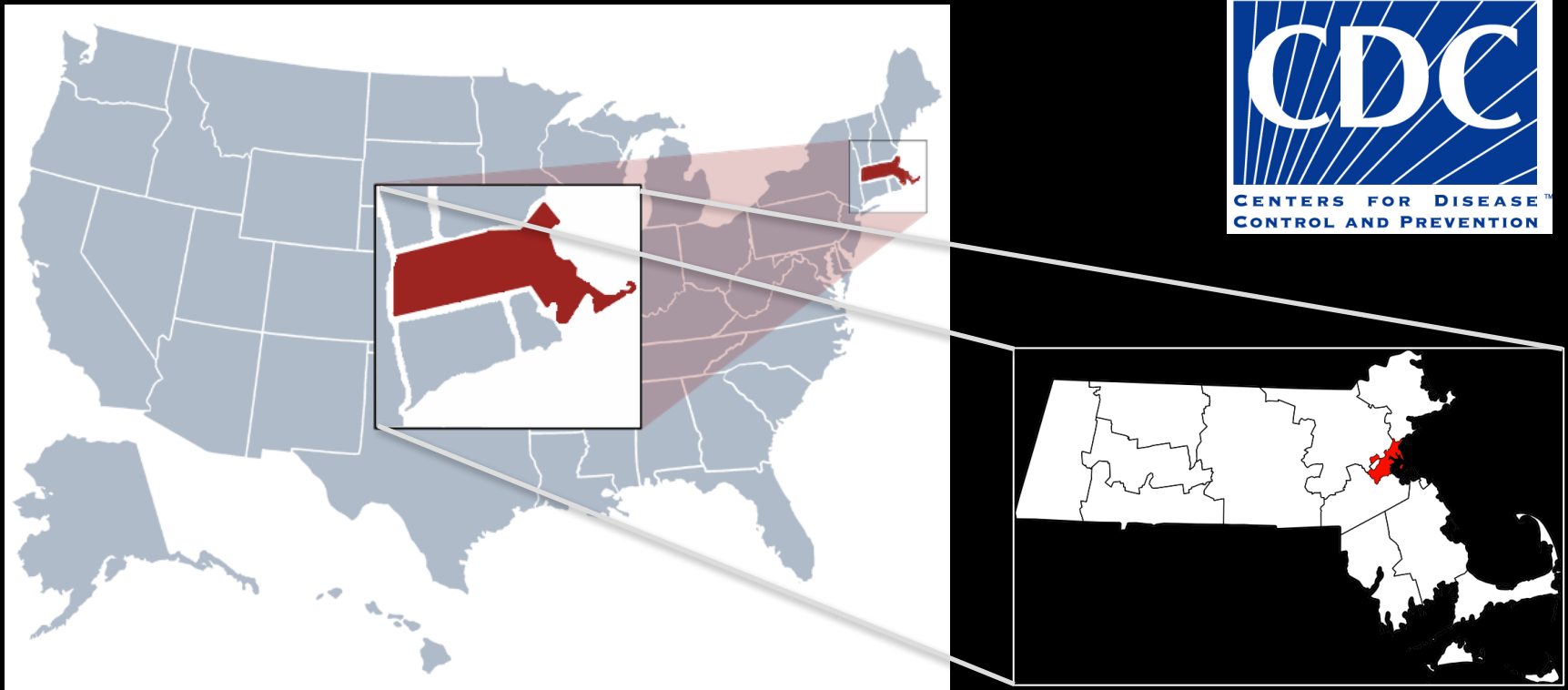
We are helping create a new improved disease forecasting platform funded by the Centers for Disease Control and Prevention



In collaboration with the CDC Influenza division, we are extending our work from National and Regional predictions, to state-level and city level (Boston as a pilot)

Grant: *Centers for Disease Control and Prevention's Cooperative Agreement PPHF 11797-998G-15*

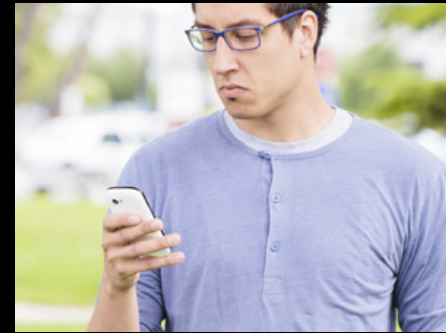
Team members: *Fred Lu, Leonardo C. Clemente*
CDC liaison and collaborator: *Matt Biggerstaff*



Beyond Google searches...



What are doctors searching for?

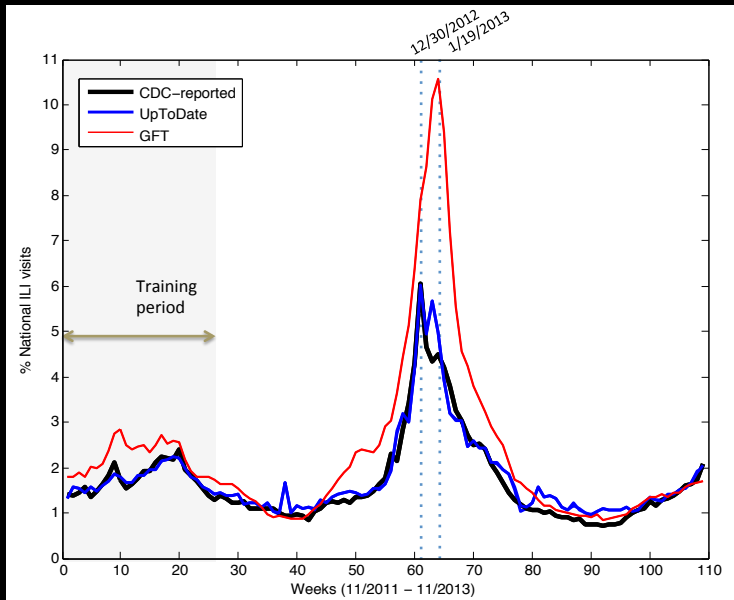


What are people tweeting? What are they reporting on crowd-sourced disease surveillance apps?

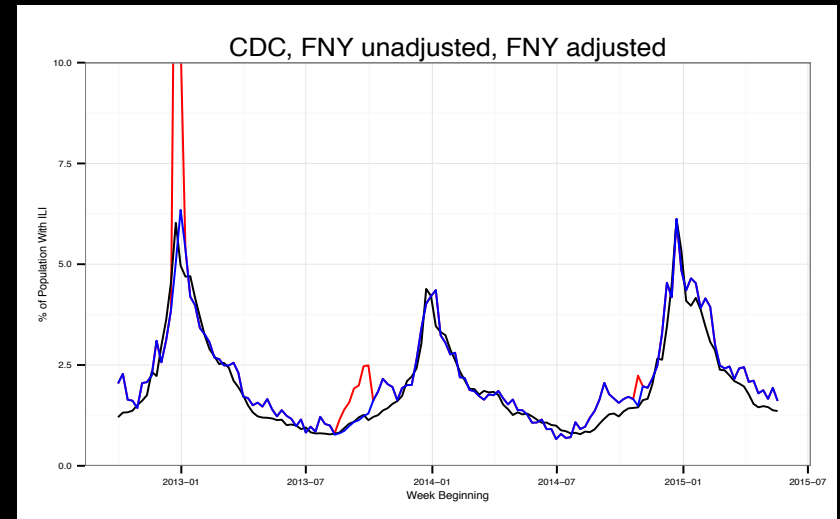


Can we use Electronic Health Records (EHR) to track disease incidence? What lab tests or medications are doctors prescribing?

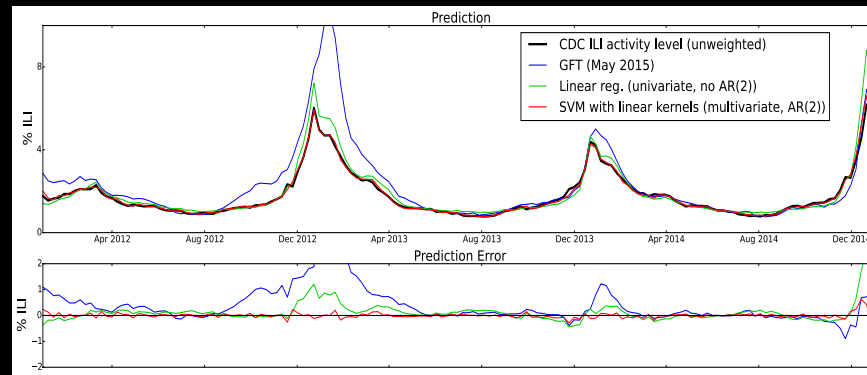
Beyond Google searches...



What are doctors searching for?

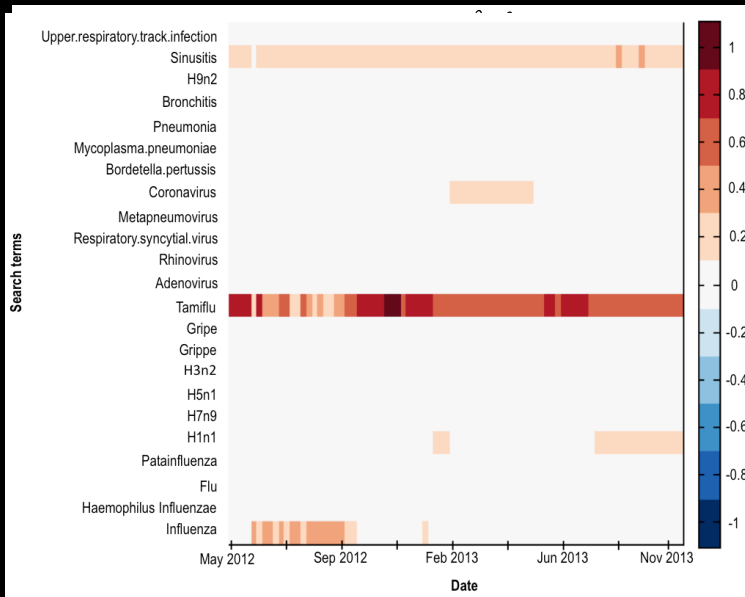


What are people tweeting? What are they reporting on crowd-sourced disease surveillance apps?

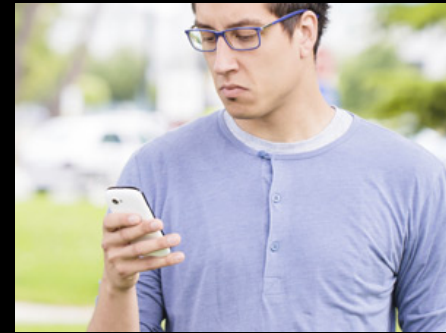


Can we use Electronic Health Records (EHR) to track disease incidence? What lab tests or medications are doctors prescribing?

Beyond Google searches...



What are doctors searching for?



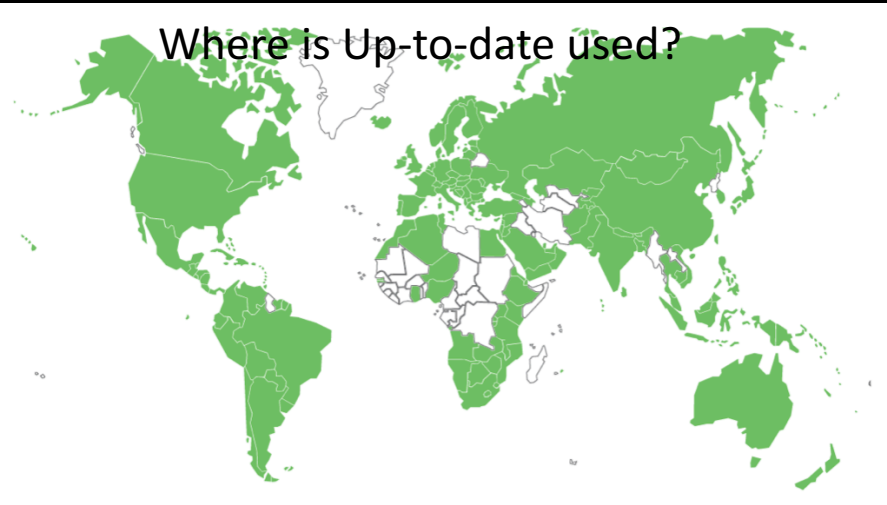
What are people tweeting? What are they reporting on crowd-sourced disease surveillance apps?



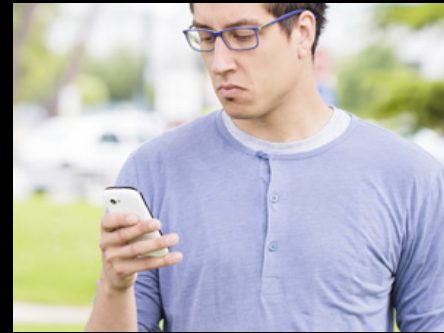
Can we use Electronic Health Records (EHR) to track disease incidence? What lab tests or medications are doctors prescribing?

Beyond Google searches...

Where is Up-to-date used?



What are doctors searching for?



What are people tweeting? What are they reporting on crowd-sourced disease surveillance apps?



Can we use Electronic Health Records (EHR) to track disease incidence? What lab tests or medications are doctors prescribing?

Beyond Google searches...

OXFORD JOURNALS

Clinical Infectious Diseases

Using Clinicians' Search Query Data to Monitor Influenza Epidemics

Mauricio Santillana,^{1,2} Elaine O. Nsoesie,^{2,3} Sumiko R. Mekaru,² David Scales,^{2,4} and John S. Brownstein^{1,5}

¹School of Engineering and Applied Sciences, Harvard University, Cambridge, ²Children's Hospital Informatics Program, Boston Children's Hospital, ³Department of Pediatrics, Harvard Medical School, Boston, and ⁴Department of Internal Medicine, Cambridge Health Alliance, Massachusetts; and ⁵Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, Quebec, Canada

Search query information from a clinician's database, UpToDate, is shown to predict influenza epidemics in the United States in a timely manner. Our results show that digital disease surveillance tools based on experts' databases may be able to provide an alternative, reliable, and stable signal for accurate predictions of influenza outbreaks.

Keywords. digital disease detection; Internet-based disease surveillance; prediction of influenza.

validated traditional surveillance systems and have the potential to provide timely epidemiologic intelligence to inform prevention messaging and healthcare facility staffing decisions.

The potential for the public's search activity to be influenced by anxiety, fears, and rumors raises concerns regarding reliability [10–13]. Although recent revisions to GFT have shown that these concerns can be partially mitigated [13–15], shifting Internet-based surveillance from the entire public to subject-matter experts may maintain timeliness while generating a more reliable and stable signal requiring much less data. A recent small retrospective study using data on queries to a Finnish primary care guidelines database demonstrated, for example, that disease-specific queries for Lyme disease, tularemia, and other infectious diseases correlated well with concurrent confirmed cases [16].

Here, we show that UpToDate (www.uptodate.com), a physician-authored clinical decision support Internet resource that is used by 700 000 clinicians in 158 countries and almost 90% of academic medical centers in the United States, can be used for syndromic surveillance of influenza. Specifically, we use UpToDate's search query activity related to ILI to design a timely sentinel of influenza incidence in the United States.

What are doctors searching for?

AJPM American Journal of Preventive Medicine

A Journal of the American College of Preventive Medicine and Association for Prevention Teaching and Research

Flu Near You: Crowdsourced Symptom Reporting Spanning 2 Influenza Seasons

Mark S. Smolinski, MD, MPH, Adam W. Crawley, MPH, Kristin Baltrusaitis, MA, Rumi Chunara, PhD, MS, Jennifer M. Olsen, DrPH, Oktavia Wijckij, PhD, Mauricio Santillana, PhD, MS, Andre Nguyen, and John S. Brownstein, PhD, MPH

Digital communications technologies have rapidly increased in use for public health disease surveillance. Mobile phones, tablets, digital pens, and satellites are making it possible for surveillance and rapid response teams in even remote areas of the globe to carry out an essential function of public health to protect against outbreaks of infectious disease. To date, public health surveillance has been limited by the capacity of public health authorities to conduct case and contact tracing and a reliance on data provided primarily by the medical system. The increased use of digital communications technology is now making it possible to enable the public to actively be part of the public health surveillance system.

Since 2003, participatory surveillance approaches have leveraged online survey technology with syndromic surveillance of human infectious diseases through volunteer symptom

Objectives. We summarized Flu Near You (FNY) data from the 2012–2013 and 2013–2014 influenza seasons in the United States.

Methods. FNY collects limited demographic characteristic information upon registration, and prompts users each Monday to report symptoms of influenza-like illness (ILI) experienced during the previous week. We calculated the descriptive statistics and rates of ILI for the 2012–2013 and 2013–2014 seasons. We compared raw and noise-filtered ILI rates with ILI rates from the Centers for Disease Control and Prevention ILINet surveillance system.

Results. More than 61 000 participants submitted at least 1 report during the 2012–2013 season, totaling 327 773 reports. Nearly 40 000 participants submitted at least 1 report during the 2013–2014 season, totaling 336 933 reports. Rates of ILI as reported by FNY tracked closely with ILINet in both timing and magnitude.

Conclusions. With increased participation, FNY has the potential to serve as a viable complement to existing outpatient, hospital-based, and laboratory surveillance systems. Although many established systems have the benefits of specificity and credibility, participatory systems offer advantages in the areas of speed, sensitivity, and scalability. (*Am J Public Health*. Published online ahead of print August 13, 2015; e1–e7. doi:10.2105/AJPH.2015.302696)

What are people tweeting? What are they reporting on crowd-sourced disease surveillance apps?

SCIENTIFIC REPORTS

OPEN

Cloud-based Electronic Health Records for Real-time, Region-specific Influenza Surveillance

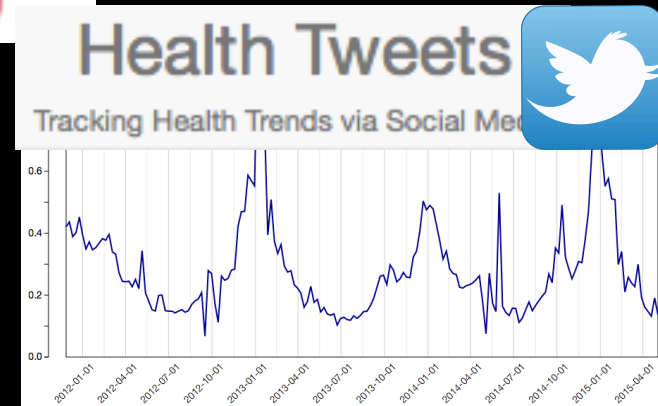
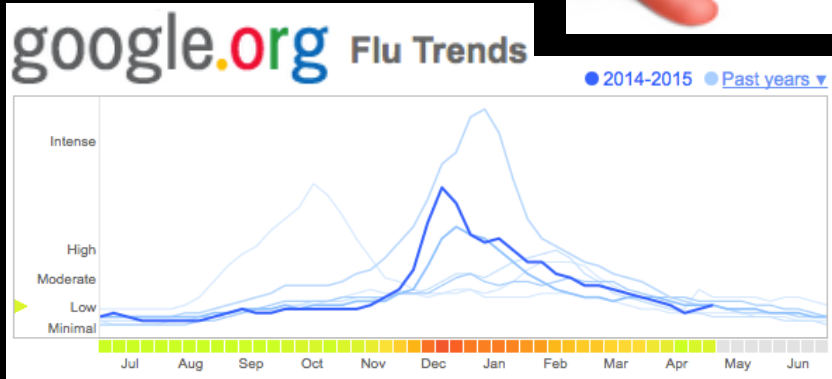
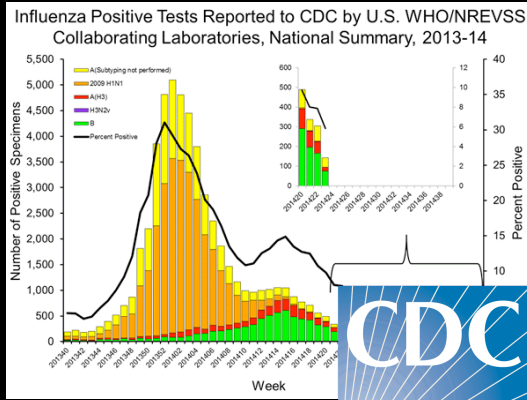
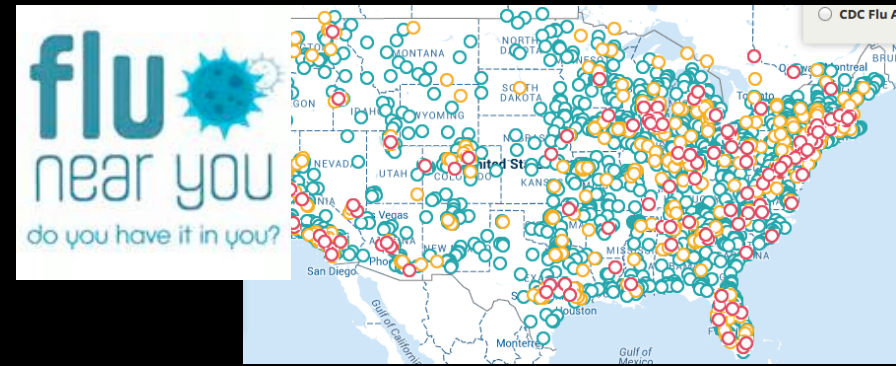
Received: 31 December 2015

Accepted: 20 April 2016

M. Santillana^{1,2,3}, A. T. Nguyen³, T. Louie⁴, A. Zink⁵, J. Gray⁵, I. Sung⁵ & J. S. Brownstein^{1,2}

Can we use Electronic Health Records (EHR) to track disease incidence? What lab tests or medications are doctors prescribing?

Ensemble approaches yield more accurate and more robust real-time and forecast flu estimates



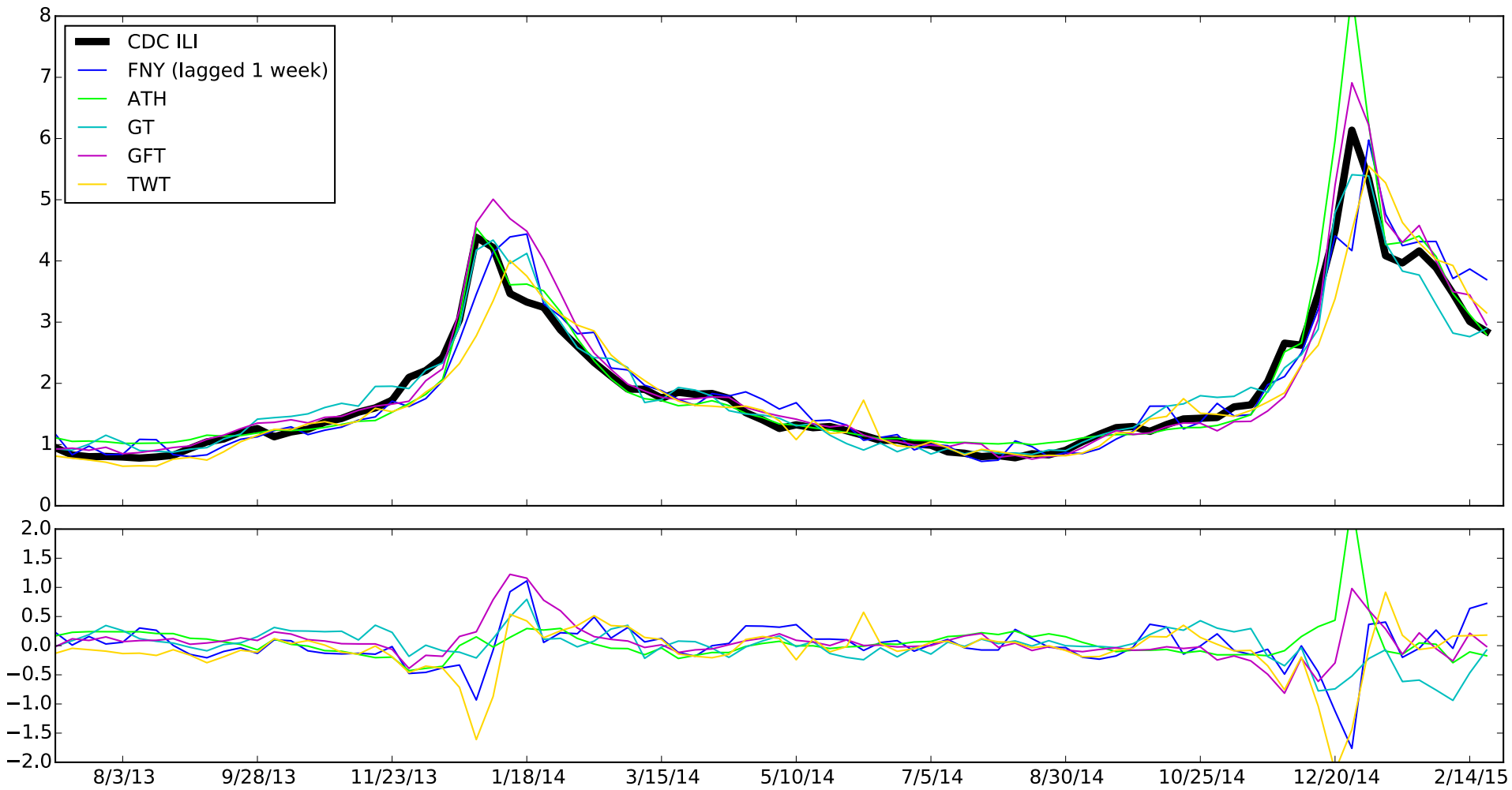
Performance of individual data sources

	CORR	RMSE (%ILI)	Rel RMSE (%)	RMAE (%)	Hit Rate
FNY	0.948	0.385	15.9	39.3	65.9
ATH	0.977	0.351	14.1	36.7	77.7
GT	0.978	0.245	13.3	42.9	65.9
GFT	0.980	0.333	12.3	35.3	75.3
TWT	0.937	0.414	15.1	50.1	62.4
CDC Baseline	0.930	0.501	18.2	46.7	68.2
CDC Virology	0.923	-	-	-	69.4

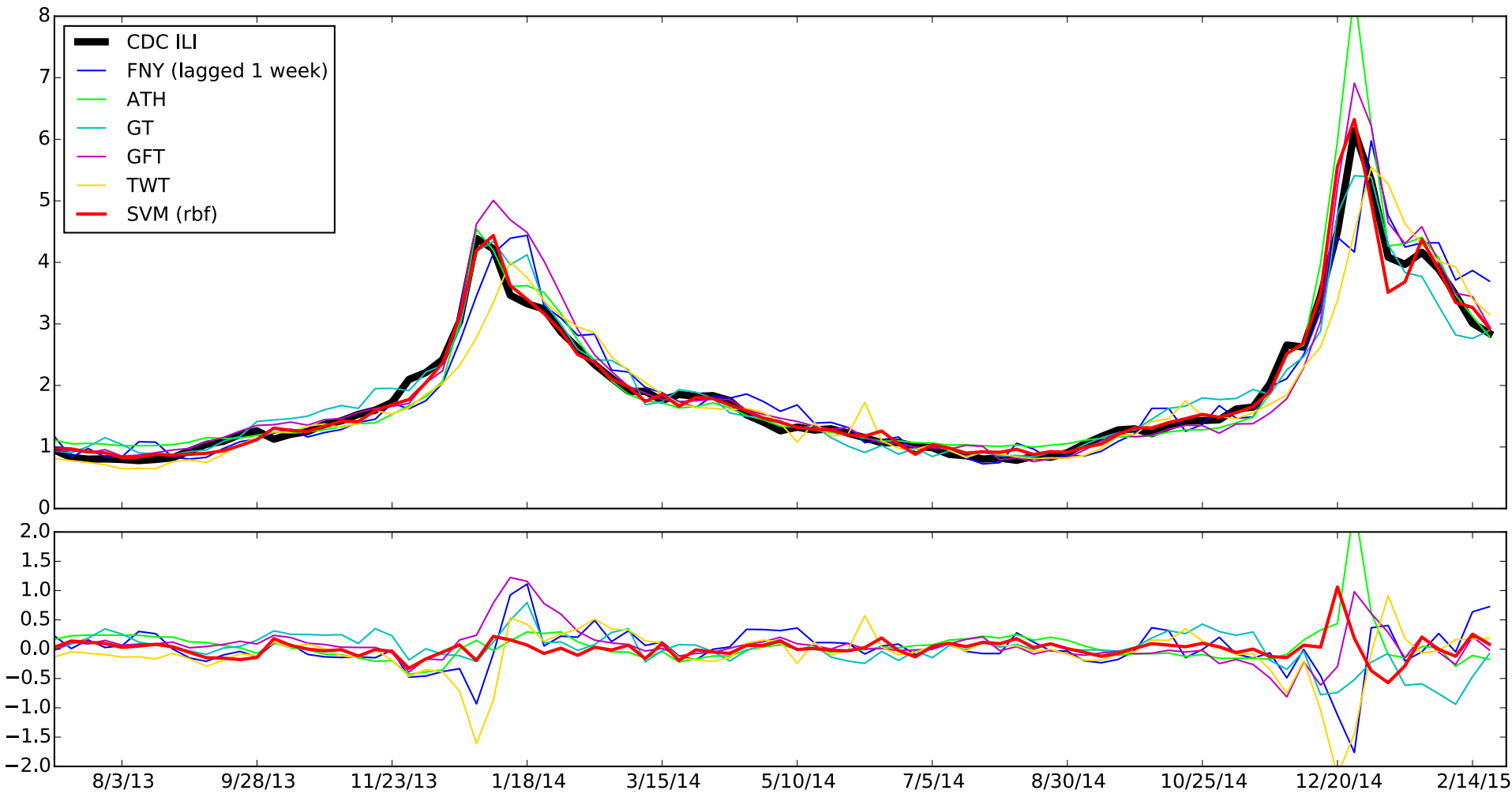
Performance ensemble

	CORR	RMSE (%ILI)	Rel RMSE (%)	RMAE (%)	Hit Rate
FNY	0.948	0.385	15.9	39.3	65.9
ATH	0.977	0.351	14.1	36.7	77.7
GT	0.978	0.245	13.3	42.9	65.9
GFT	0.980	0.333	12.3	35.3	75.3
TWT	0.937	0.414	15.1	50.1	62.4
CDC Baseline	0.930	0.501	18.2	46.7	68.2
CDC Virology	0.923	-	-	-	69.4
SVM (RBF)	0.989	0.176	8.27	23.6	69.4

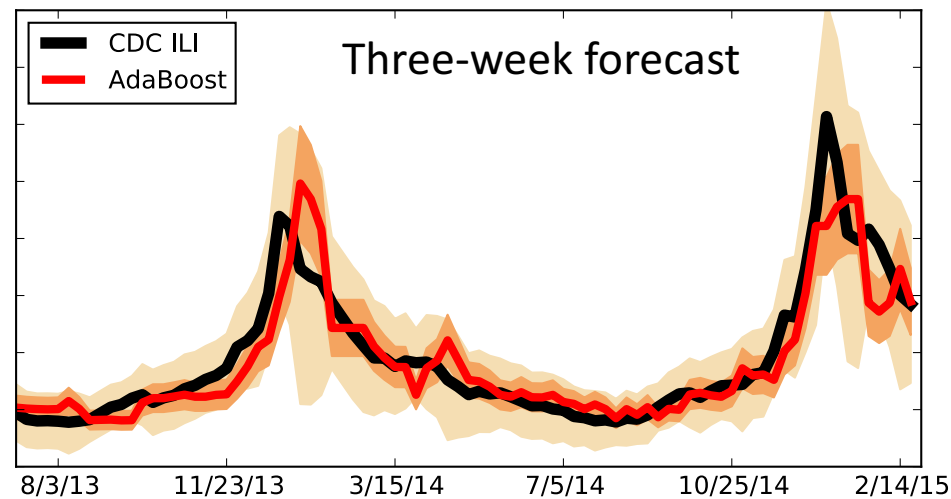
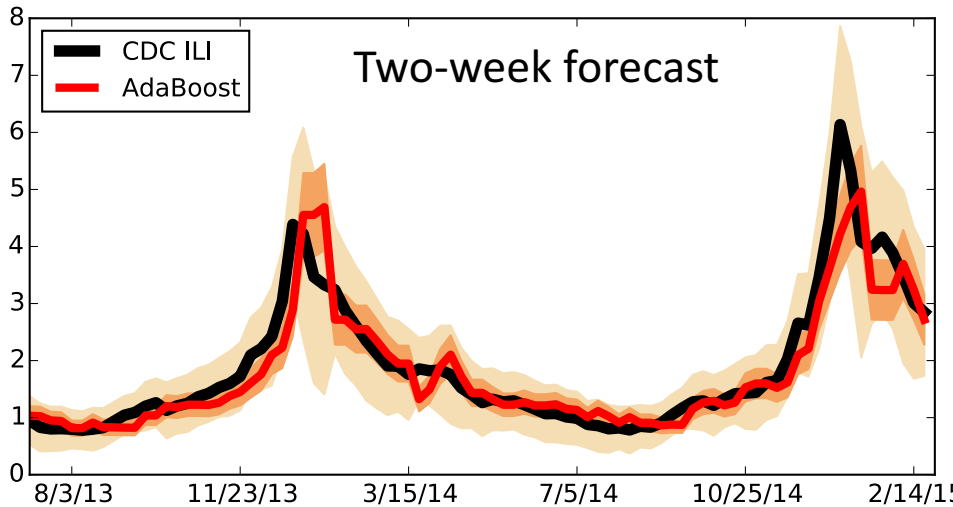
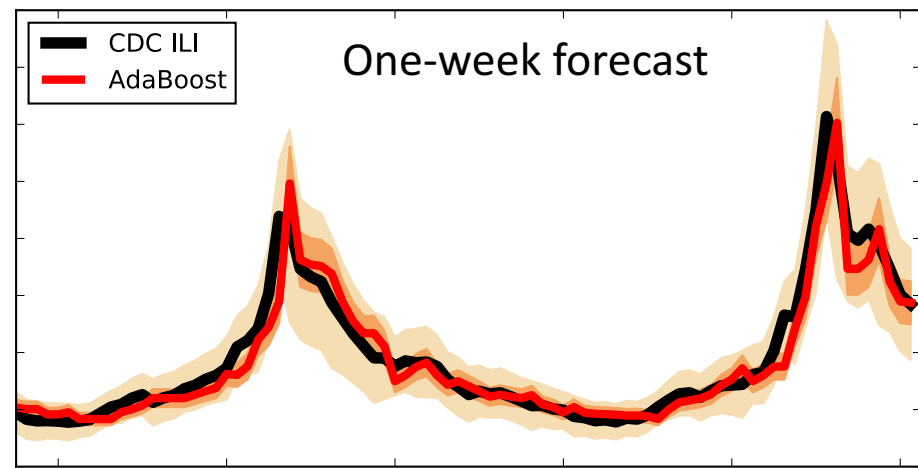
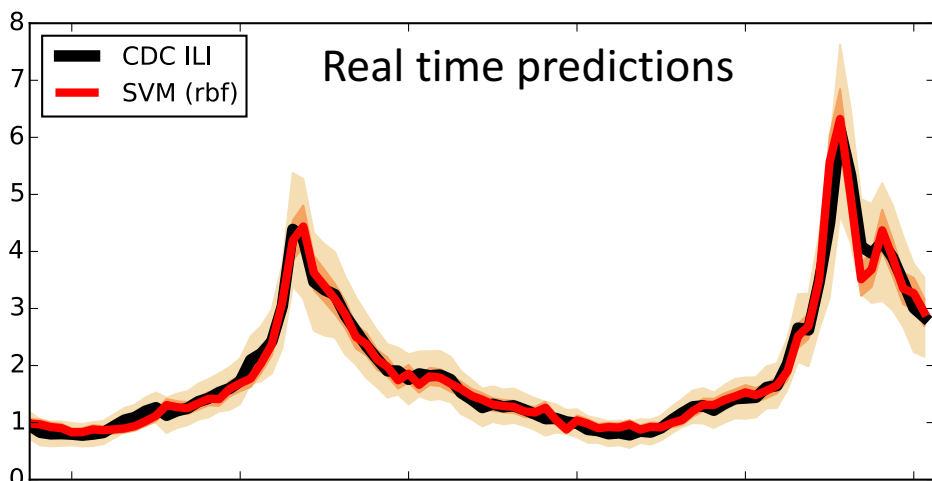
Performance of individual data sources



Performance ensemble



Real time predictions and Forecasts



Ensemble approaches yield more accurate and more robust real-time and forecast flu estimates

Yang et al. *BMC Infectious Diseases* (2017) 17:332
DOI 10.1186/s12879-017-2424-7

BMC Infectious Diseases

RESEARCH ARTICLE

Open Access



Using electronic health records and Internet search information for accurate influenza forecasting

Shihao Yang¹, Mauricio Santillana^{2,3*}, John S. Brownstein^{2,3}, Josh Gray⁴, Stewart Richardson⁴ and S. C. Kou^{1*}

Abstract

Background: Accurate influenza activity forecasting helps public health officials prepare and allocate resources for unusual influenza activity. Traditional flu surveillance systems, such as the Centers for Disease Control and Prevention's (CDC) influenza-like illnesses reports, lag behind real-time by one to 2 weeks, whereas information contained in cloud-based electronic health records (EHR) and in Internet users' search activity is typically available in near real-time. We present a method that combines the information from these two data sources with historical flu activity to produce national flu forecasts for the United States up to 4 weeks ahead of the publication of CDC's flu reports.

Methods: We extend a method originally designed to track flu using Google searches, named ARGO, to combine information from EHR and Internet searches with historical flu activities. Our regularized multivariate regression model dynamically selects the most appropriate variables for flu prediction every week. The model is assessed for the flu seasons within the time period 2013–2016 using multiple metrics including root mean squared error (RMSE).

Results: Our method reduces the RMSE of the publicly available alternative (Healthmap flutrends) method by 33, 20, 17 and 21%, for the four time horizons: real-time, one, two, and 3 weeks ahead, respectively. Such accuracy improvements are statistically significant at the 5% level. Our real-time estimates correctly identified the peak timing and magnitude of the studied flu seasons.

Conclusions: Our method significantly reduces the prediction error when compared to historical publicly available Internet-based prediction systems, demonstrating that: (1) the method to combine data sources is as important as data quality; (2) effectively extracting information from a cloud-based EHR and Internet search activity leads to accurate forecast of flu.

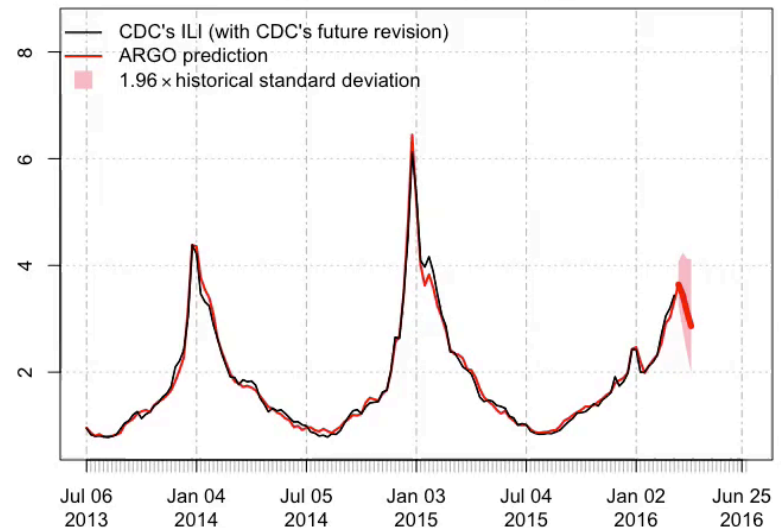
Keywords: Influenza-like illnesses reports, Digital disease detection, Dynamic error reduction, Validation test, Autoregression

RESEARCH ARTICLE

Combining Search, Social Media, and Traditional Data Sources to Improve Influenza Surveillance

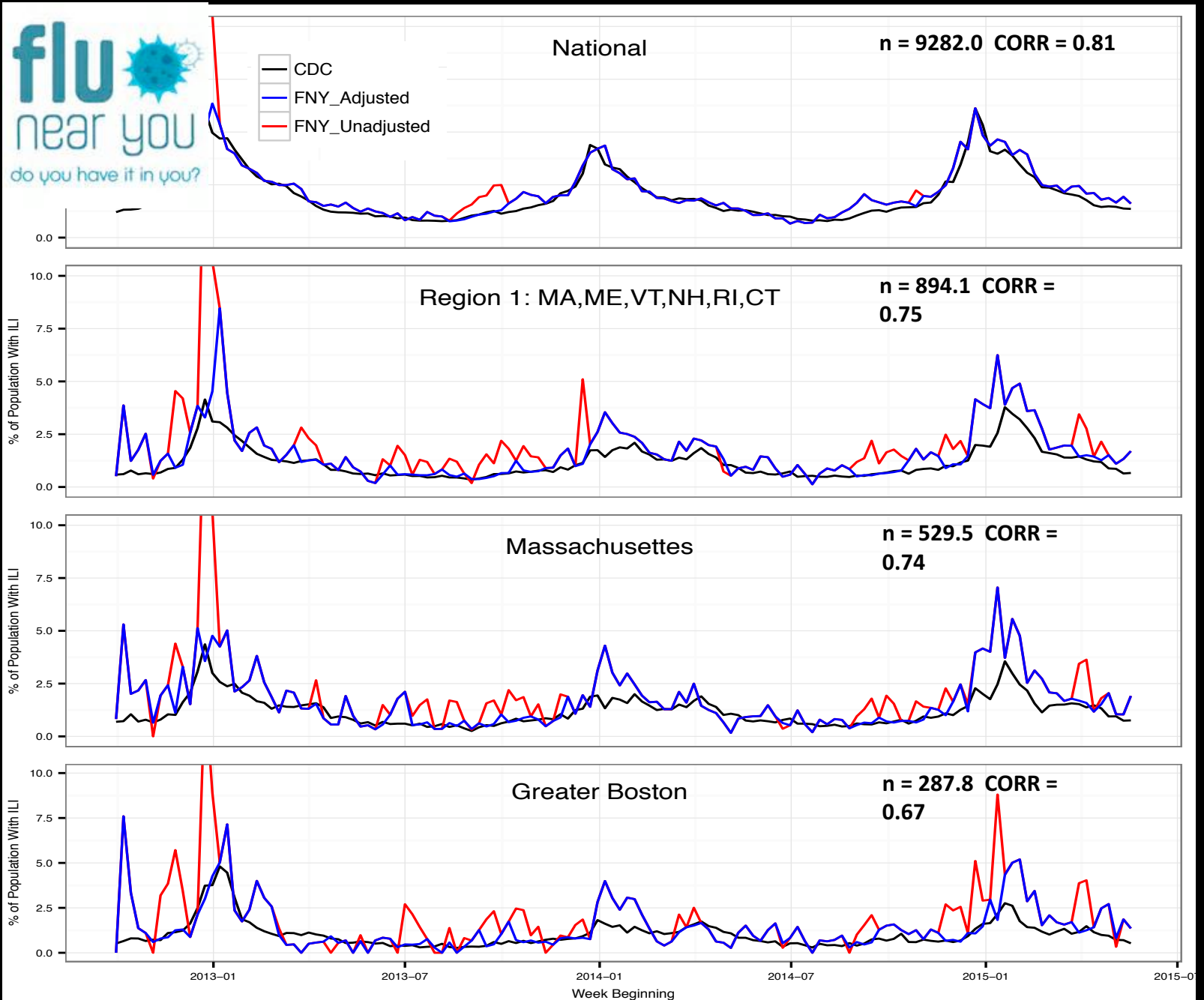
Mauricio Santillana^{1,2,3*}, André T. Nguyen¹, Mark Dredze⁴, Michael J. Paul⁵, Elaine O. Nsoesie^{6,7}, John S. Brownstein^{2,3}

ARGO Prediction vs. CDC's ILI

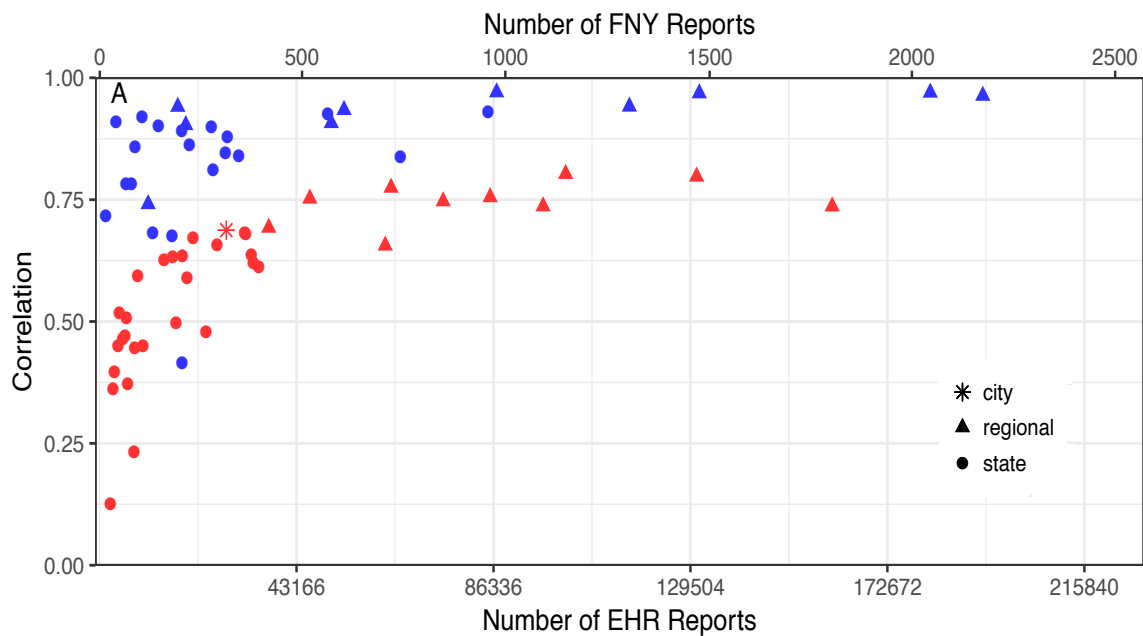


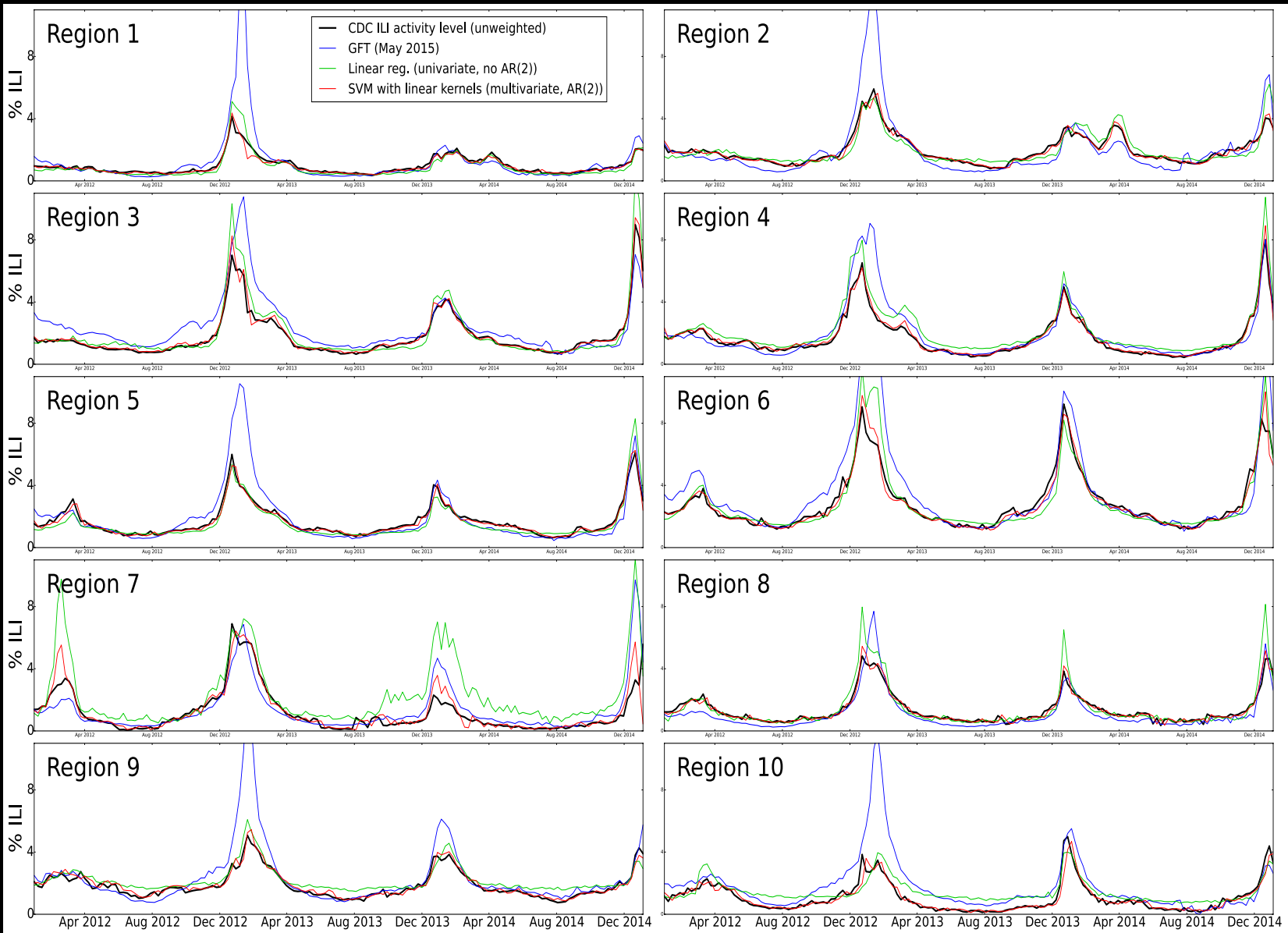
What about estimating flu in a regional level,
state-level, city-level?

Correlation of FNY with CDC. Multiple Geographic Scales



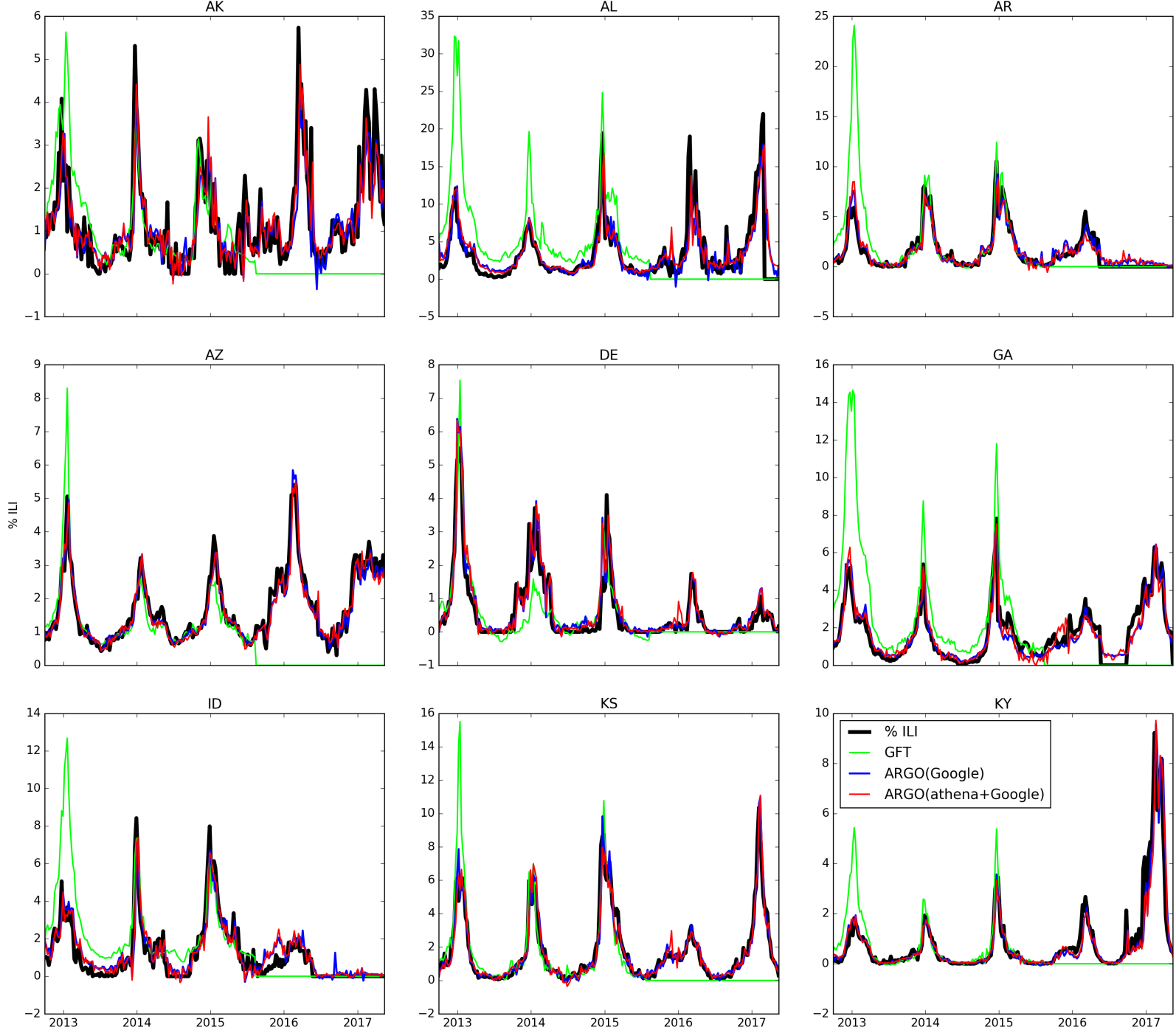
Comparisons between CDC ILINet and FNY and Athenahealth in multiple spatial scales





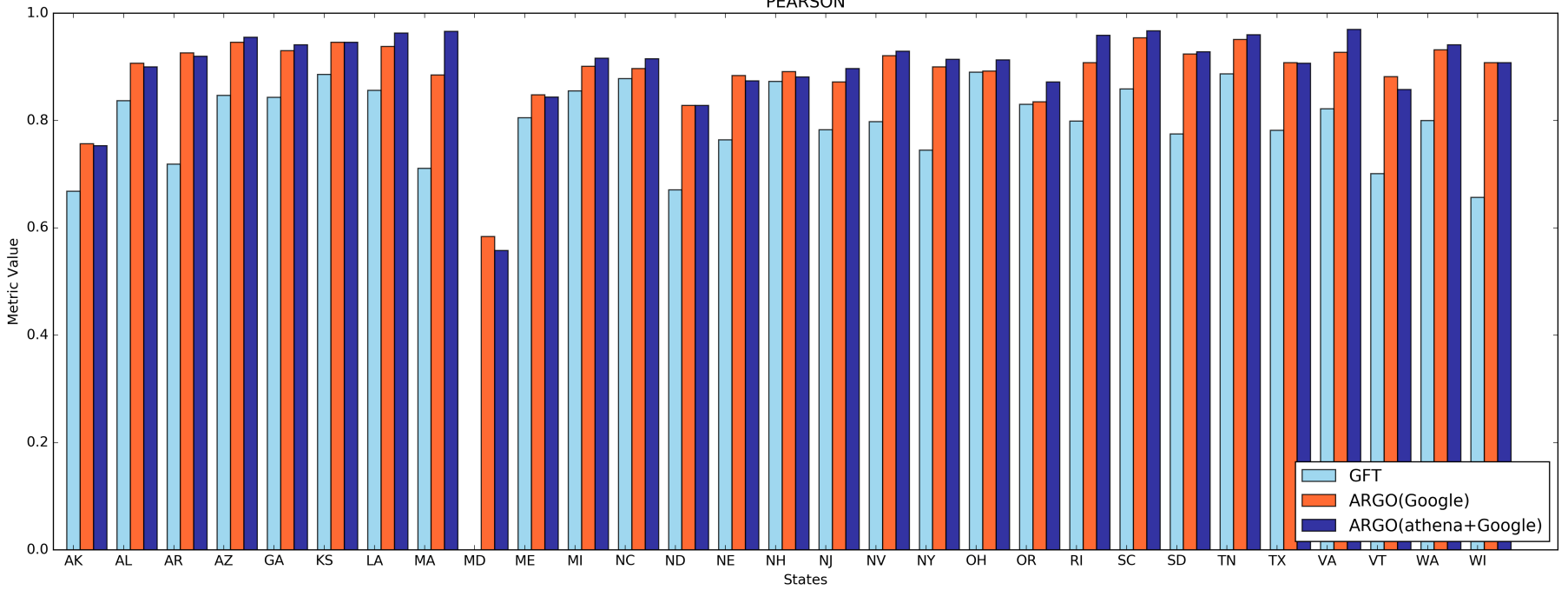
How about state-level?

Out of sample predictions



Metric	State	Model	2012-13	2013-14	2014-15	2015-16	2016-17	GFT Period	Whole Period
PEARSON	AK	GFT	0.673	0.932	0.768	--	--	0.669	0.669
		AR52	0.655	0.77	0.772	0.701	0.717	0.751	0.749
		ARGO(Google)	0.655	0.787	0.736	0.767	0.736	0.769	0.751
		ARGO(athena+Google)	0.682	0.828	0.73	0.814	0.801	0.795	0.766
	AL	GFT	0.915	0.867	0.939	--	--	0.837	0.837
		AR52	0.893	0.94	0.792	0.701	0.956	0.86	0.878
		ARGO(Google)	0.887	0.956	0.848	0.625	0.954	0.858	0.896
		ARGO(athena+Google)	0.962	0.958	0.85	0.564	0.964	0.859	0.905
	AR	GFT	0.913	0.949	0.968	--	--	0.719	0.719
		AR52	0.939	0.905	0.86	0.53	--	0.899	0.92
		ARGO(Google)	0.965	0.918	0.894	0.751	--	0.917	0.926
		ARGO(athena+Google)	0.966	0.923	0.893	0.758	--	0.905	0.916
	AZ	GFT	0.951	0.913	0.913	--	--	0.847	0.847
		AR52	0.89	0.902	0.955	0.91	0.898	0.942	0.933
		ARGO(Google)	0.949	0.944	0.953	0.938	0.892	0.952	0.958
		ARGO(athena+Google)	0.942	0.939	0.956	0.929	0.894	0.95	0.955
	DE	GFT	0.938	0.623	0.712	--	--	0.766	0.766
		AR52	0.927	0.702	0.81	0.623	0.692	0.887	0.875
		ARGO(Google)	0.96	0.732	0.778	0.912	0.83	0.907	0.891
		ARGO(athena+Google)	0.957	0.721	0.826	0.851	0.851	0.91	0.896
	GA	GFT	0.91	0.956	0.879	--	--	0.843	0.843
		AR52	0.95	0.895	0.815	0.676	0.847	0.913	0.907
		ARGO(Google)	0.969	0.953	0.896	0.87	0.915	0.932	0.936
		ARGO(athena+Google)	0.973	0.962	0.909	0.296	0.918	0.927	0.941
	ID	GFT	0.862	0.903	0.956	--	--	0.638	0.638
		AR52	0.812	0.802	0.897	0.368	--	0.862	0.871
		ARGO(Google)	0.857	0.827	0.915	0.431	--	0.873	0.884
		ARGO(athena+Google)	0.824	0.85	0.917	0.303	--	0.876	0.892
KS	GFT	0.88	0.954	0.957	--	--	0.886	0.886	
	AR52	0.918	0.897	0.949	0.588	0.909	0.933	0.946	
	ARGO(Google)	0.935	0.932	0.944	0.902	0.964	0.957	0.952	
	ARGO(athena+Google)	0.943	0.94	0.933	0.793	0.962	0.95	0.944	

PEARSON

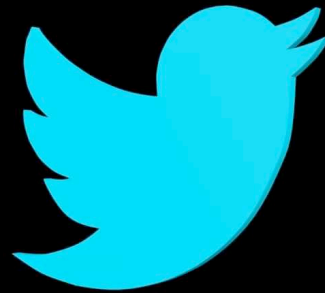


How about city-level?

Refining the spatial resolution...

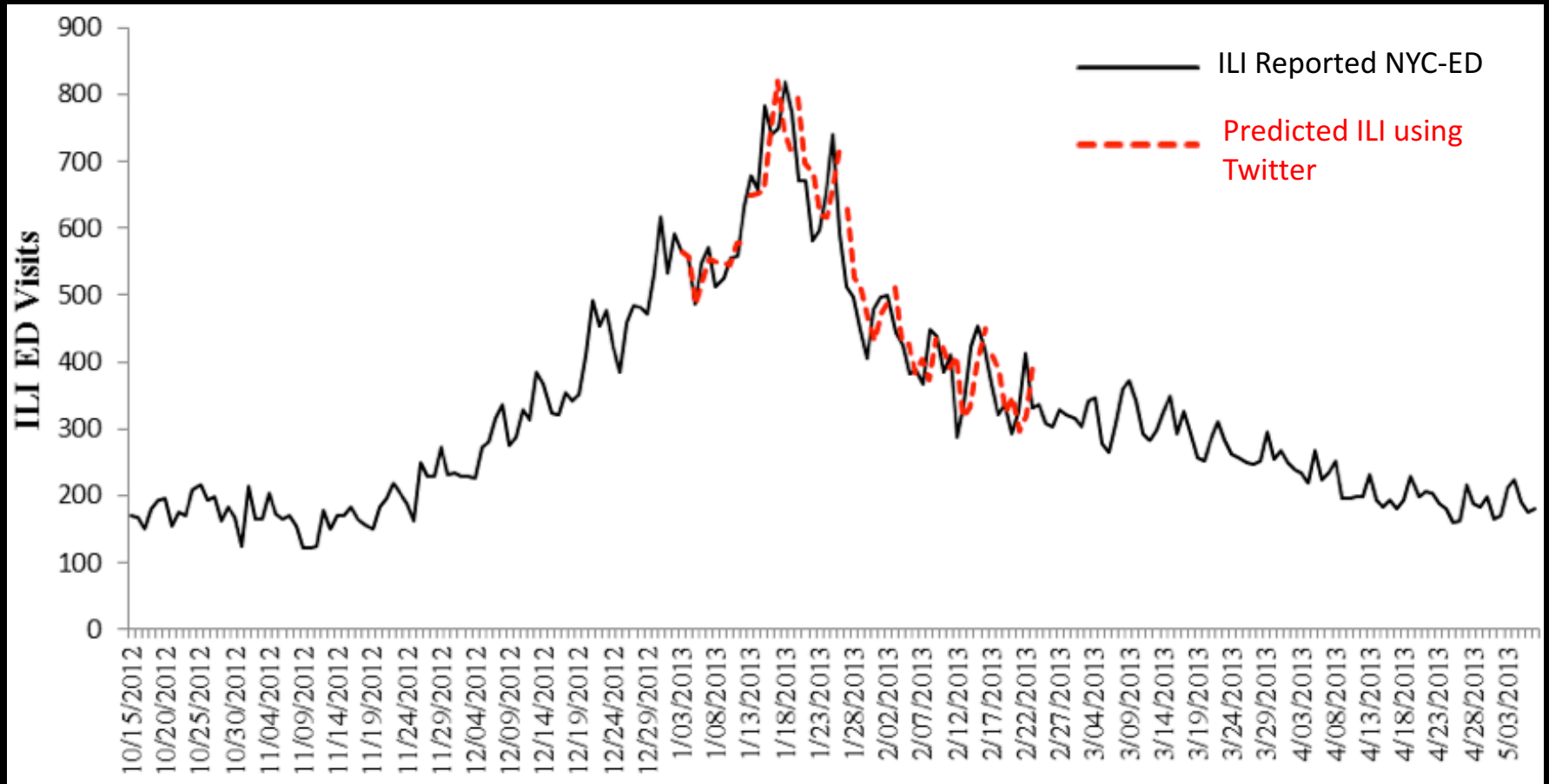


Tracking Flu using twitter
(Daily analysis in NYC)



Work with R. Nagar, Q. Yuan, C. Freifeld, A. Nojima, R. Chunara, and J. S. Brownstein

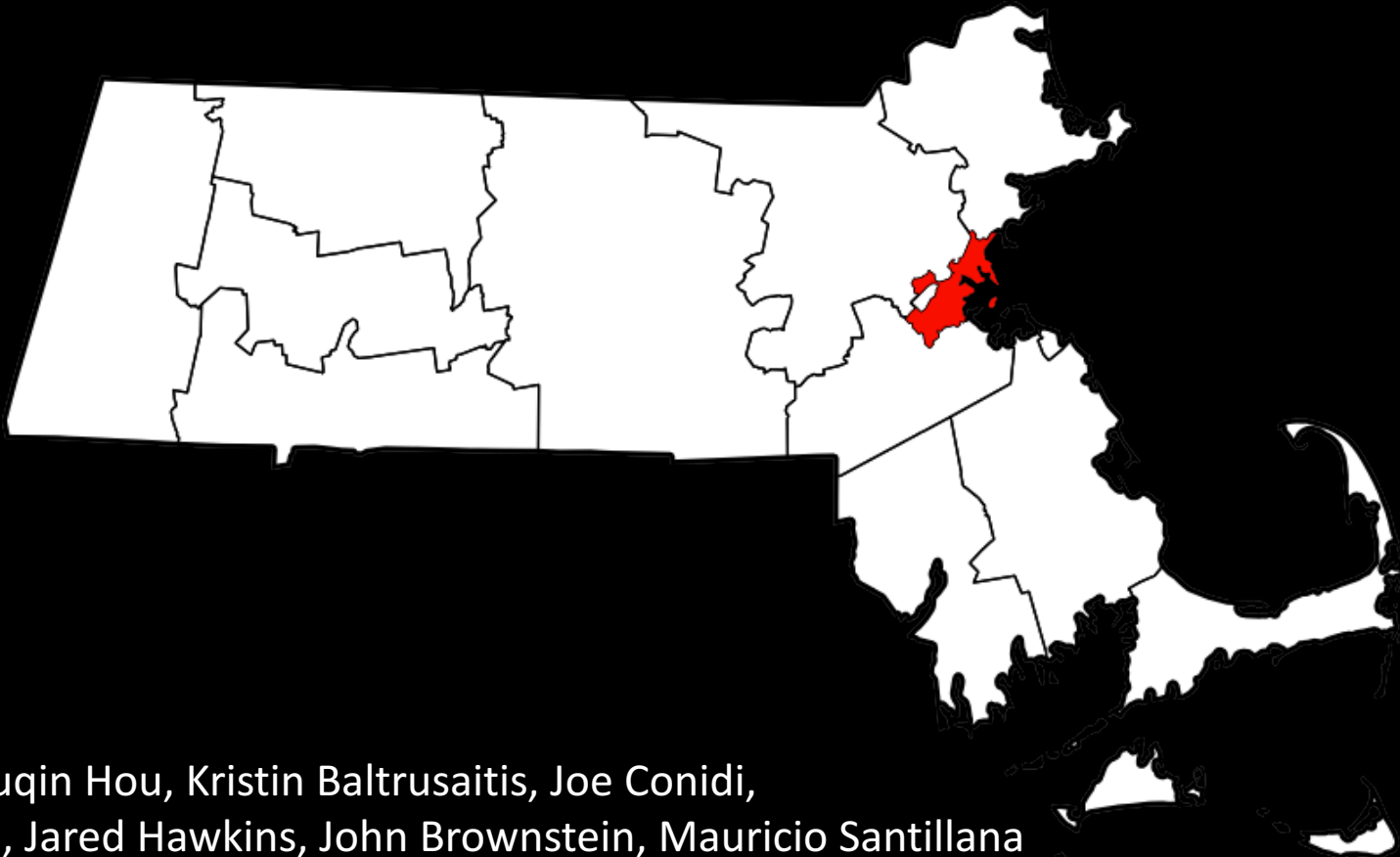
Daily ILI visits (as reported by the NYC emergency department) compared to predicted ILI using twitter data



We will extend out methodology to finer spatial resolutions. Pilot projects:

1. State level: Massachusetts

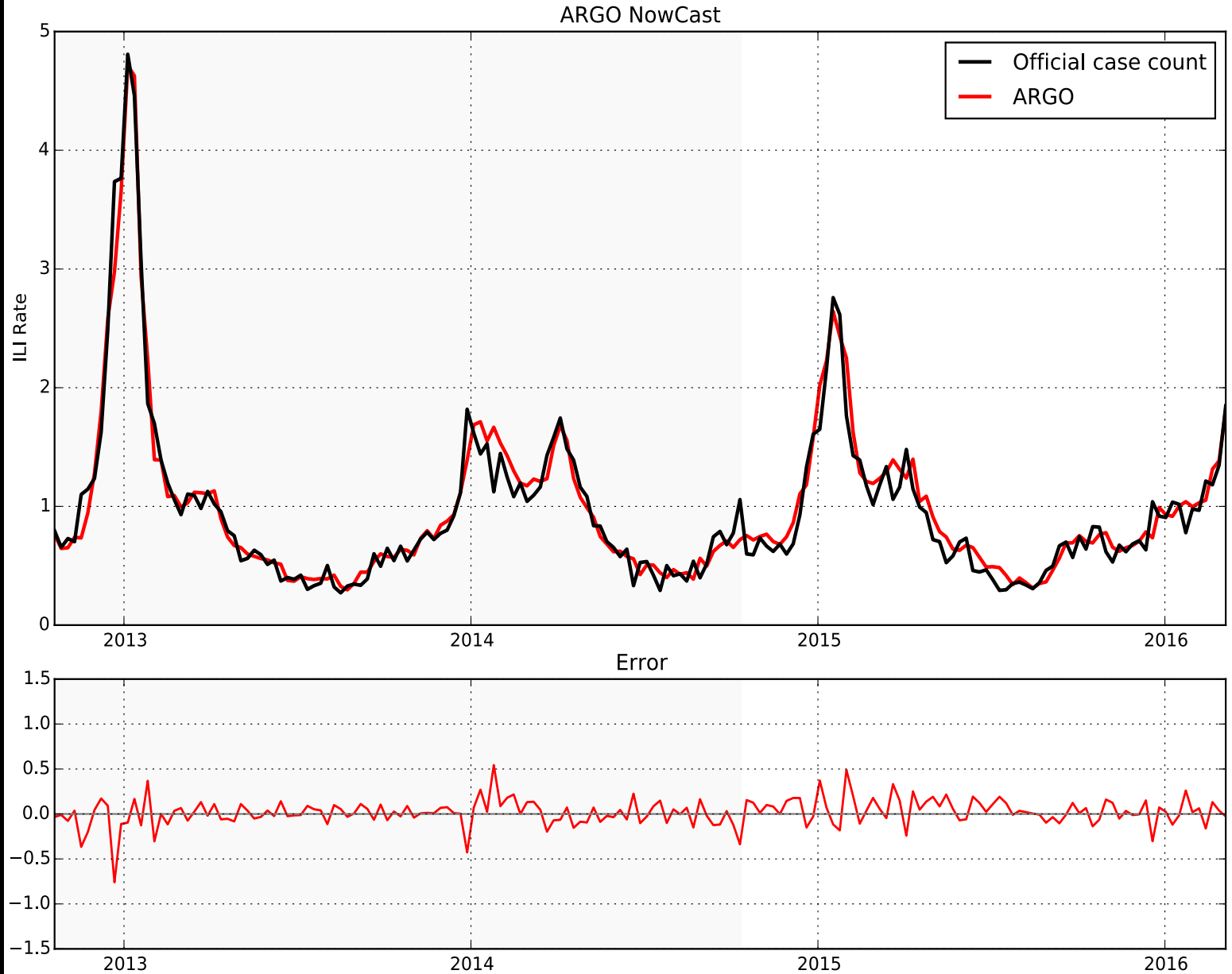
1. City level: Boston



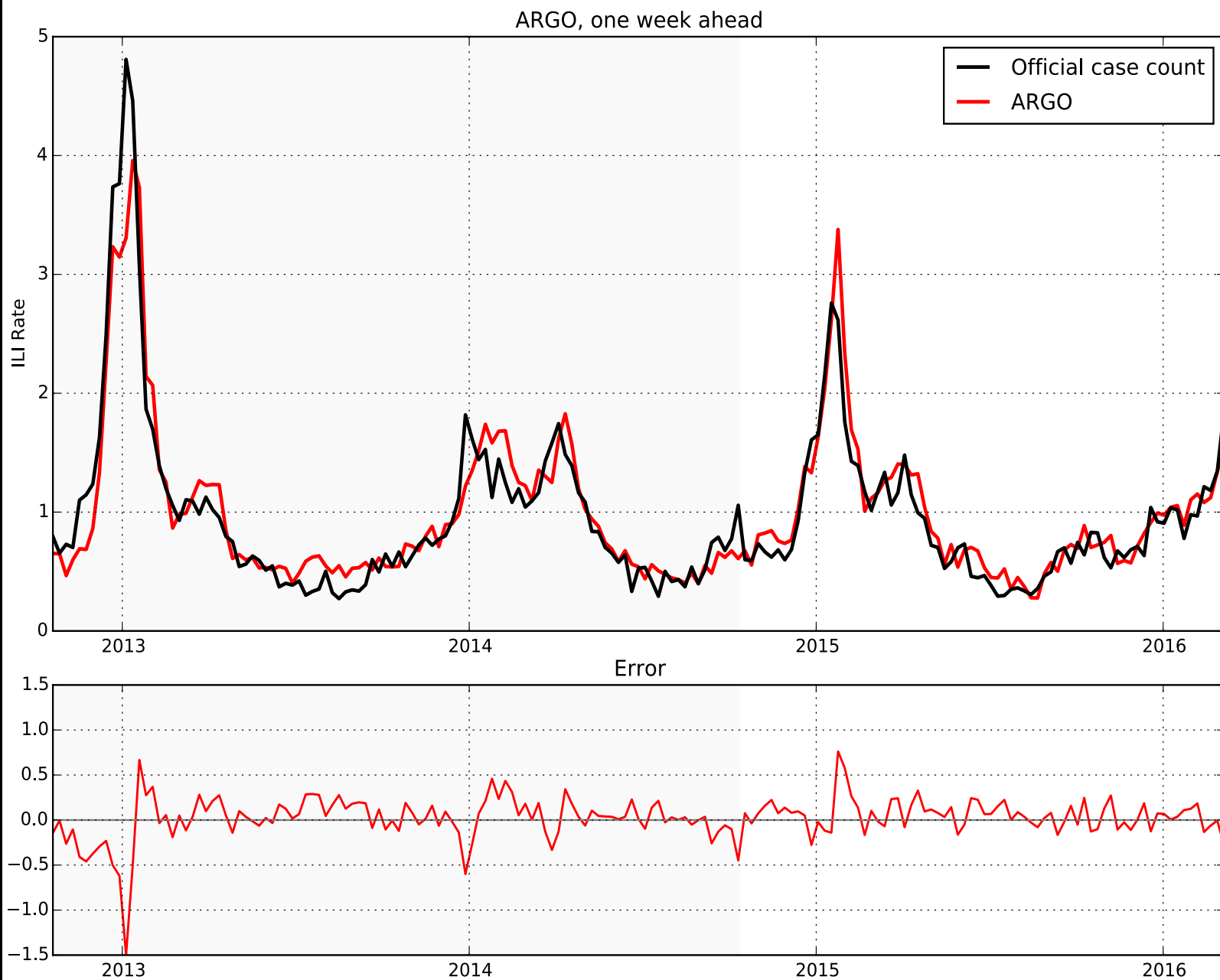
Team:

Fred Lu, Suqin Hou, Kristin Baltrusaitis, Joe Conidi,
Julia Gunn, Jared Hawkins, John Brownstein, Mauricio Santillana

Using multiple data sources to track flu in Boston

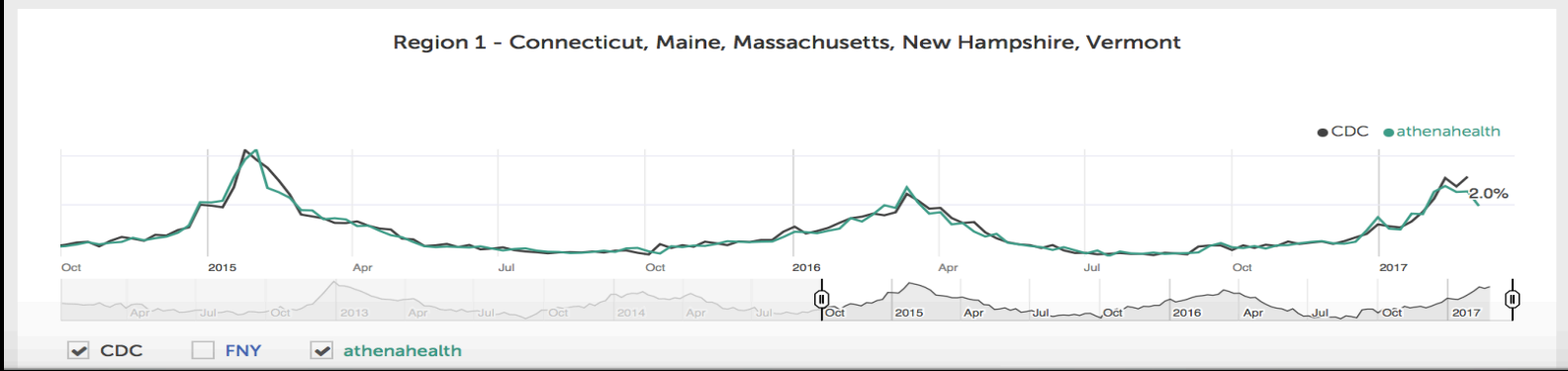
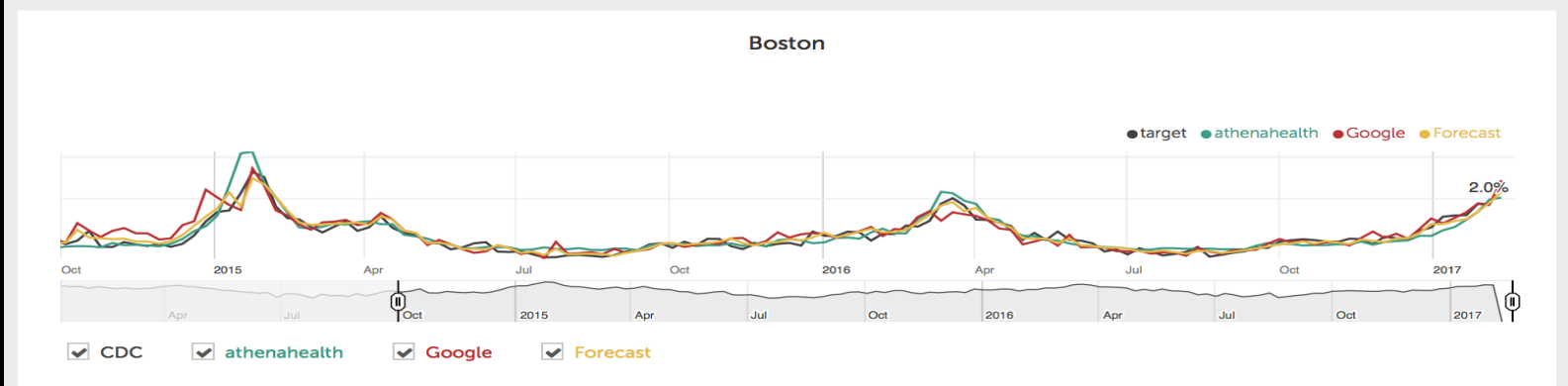
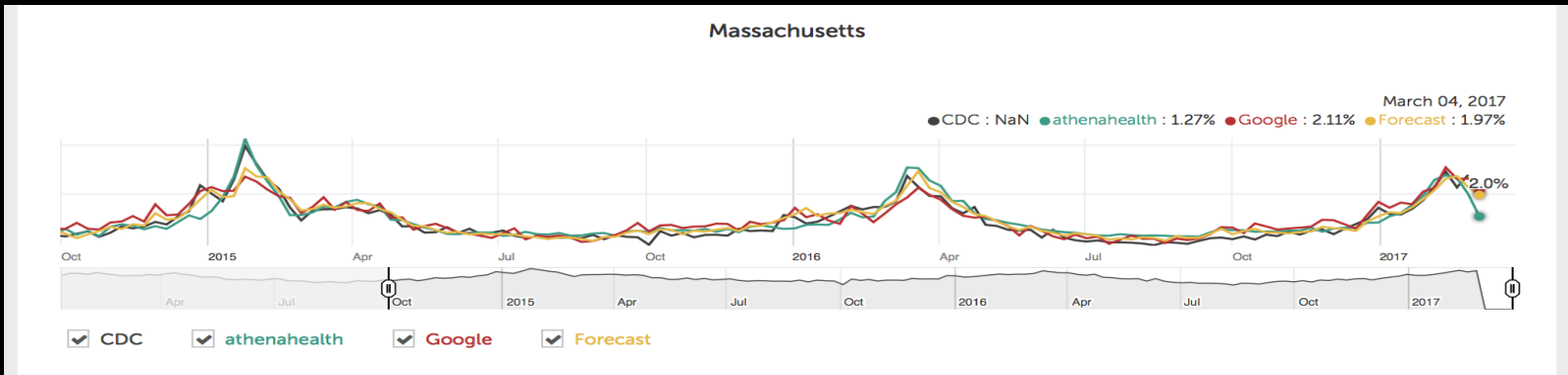


Using multiple data sources to forecast flu in Boston



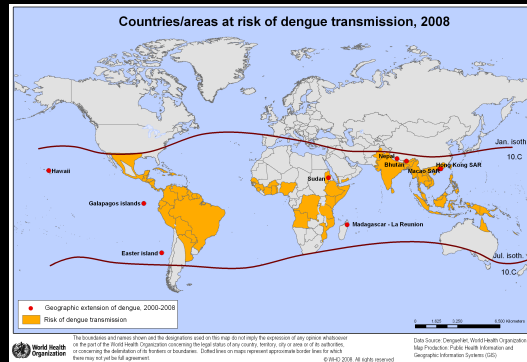
Aim is to display these predictions in a joint CDC-BCH website

Using multiple data sources to track flu at the state-level in the USA



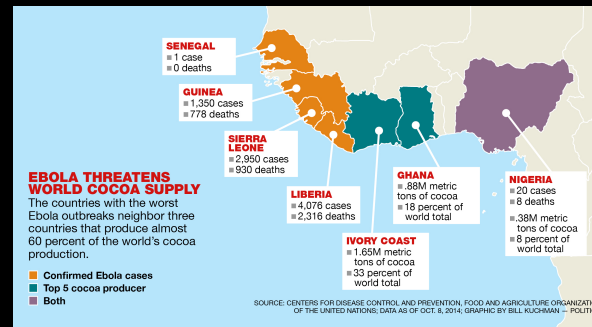
Part 2. Success stories in tracking and forecasting Flu, Zika, Dengue, Ebola in data-poor medium- to low-income countries.

Dengue, Zika, and Flu



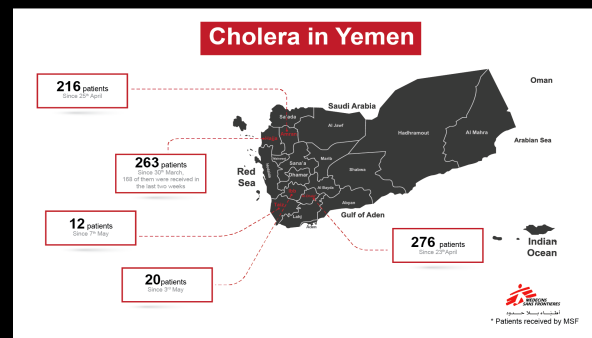
- Latin America (Flu, Zika, Dengue)
- South-east Asia (Dengue)

Ebola



- West Africa

Cholera



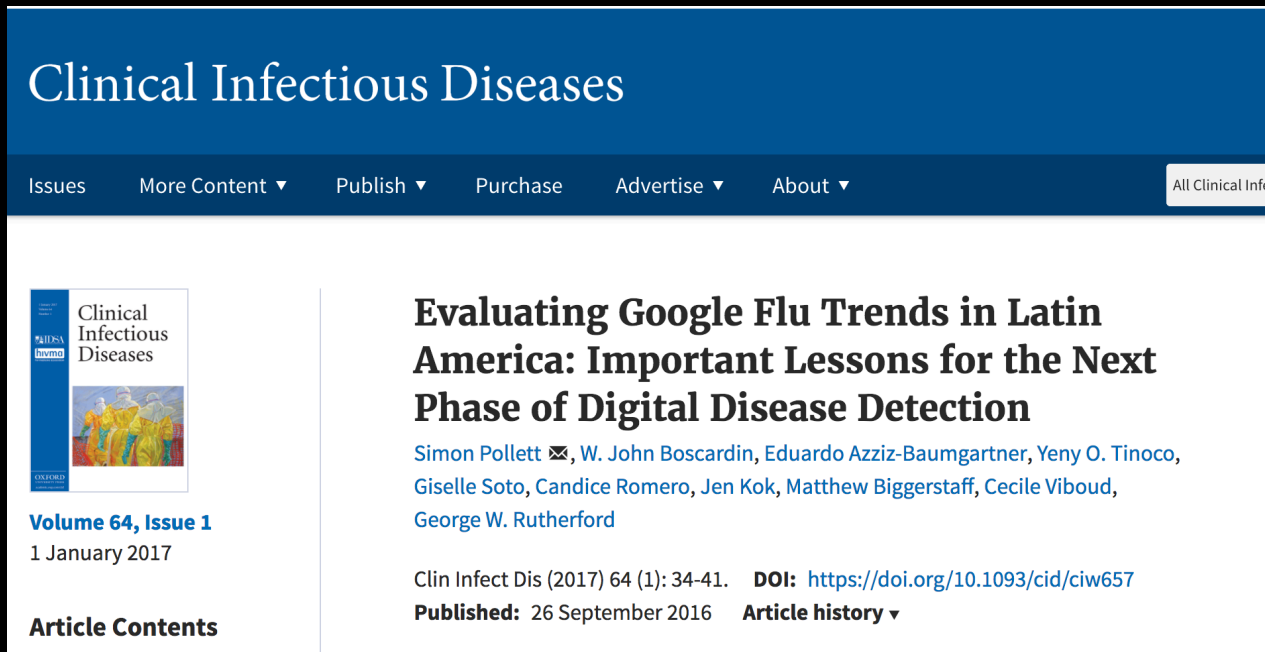
- Middle East

We are extending our tracking and forecasting systems to middle income countries

Now extending to Latin America

In response to a paper on *Clinical Infectious Diseases*, we have extended our methods to better predict flu in Latin America

Team: Leonardo C. Clemente and Fred Lu



The screenshot shows the journal's website interface. At the top, the journal title 'Clinical Infectious Diseases' is displayed in white on a blue background. Below this is a navigation bar with links for 'Issues', 'More Content', 'Publish', 'Purchase', 'Advertise', and 'About'. A search bar on the right contains the text 'All Clinical Infect'. The main content area features a cover image of the journal on the left, which includes the title 'Clinical Infectious Diseases', the volume and issue information 'Volume 64, Issue 1', and the date '1 January 2017'. To the right of the cover is the article title 'Evaluating Google Flu Trends in Latin America: Important Lessons for the Next Phase of Digital Disease Detection' in a large, bold font. Below the title, the authors are listed: Simon Pollett, W. John Boscardin, Eduardo Azziz-Baumgartner, Yeny O. Tinoco, Giselle Soto, Candice Romero, Jen Kok, Matthew Biggerstaff, Cecile Viboud, and George W. Rutherford. The article's publication details are provided at the bottom: 'Clin Infect Dis (2017) 64 (1): 34-41. DOI: <https://doi.org/10.1093/cid/ciw657> Published: 26 September 2016 Article history'.

Clinical Infectious Diseases

Issues More Content ▼ Publish ▼ Purchase Advertise ▼ About ▼ All Clinical Infect

Evaluating Google Flu Trends in Latin America: Important Lessons for the Next Phase of Digital Disease Detection

Simon Pollett ✉, W. John Boscardin, Eduardo Azziz-Baumgartner, Yeny O. Tinoco, Giselle Soto, Candice Romero, Jen Kok, Matthew Biggerstaff, Cecile Viboud, George W. Rutherford

Clin Infect Dis (2017) 64 (1): 34-41. DOI: <https://doi.org/10.1093/cid/ciw657>
Published: 26 September 2016 Article history ▼

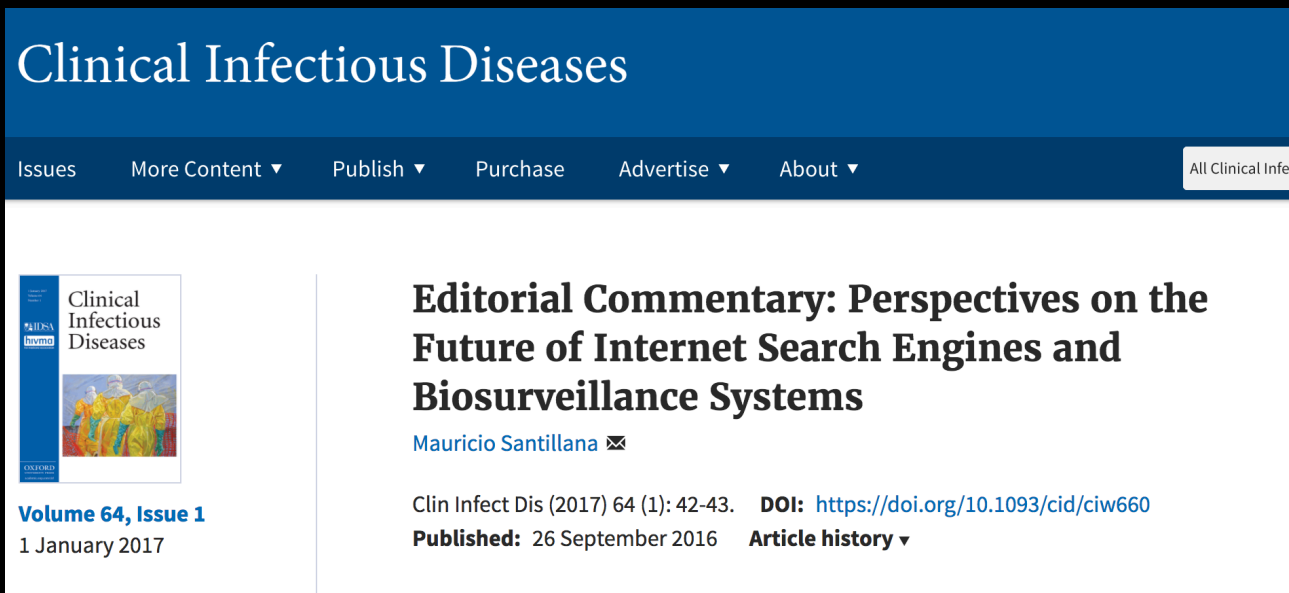
Volume 64, Issue 1
1 January 2017

Article Contents

Now extending to Latin America

In response to a paper on *Clinical Infectious Diseases*, we have extended our methods to better predict flu in Latin America

Team: Leonardo C. Clemente and Fred Lu



The screenshot shows the journal's website interface. At the top, the journal title "Clinical Infectious Diseases" is displayed in white on a dark blue background. Below this is a navigation bar with links for "Issues", "More Content", "Publish", "Purchase", "Advertise", and "About". A search box on the right contains the text "All Clinical Infe".

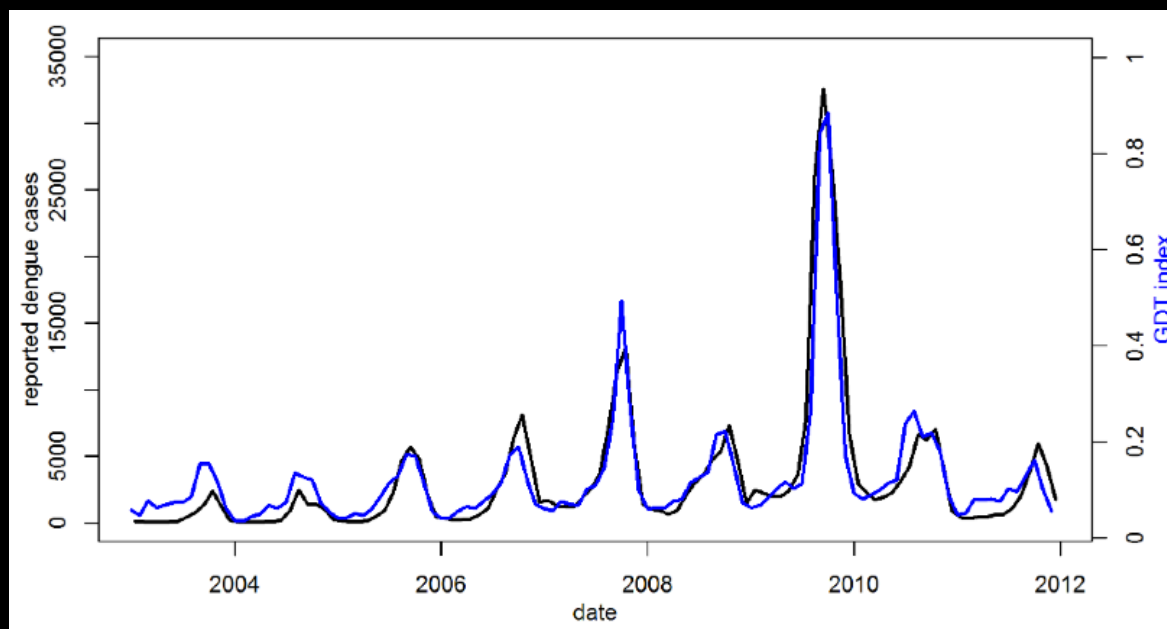
The main content area features a cover image of the journal on the left, which includes the text "Clinical Infectious Diseases" and "Volume 64, Issue 1, 1 January 2017". To the right of the cover is the article title: "Editorial Commentary: Perspectives on the Future of Internet Search Engines and Biosurveillance Systems" by "Mauricio Santillana". Below the title, the publication information is given as "Clin Infect Dis (2017) 64 (1): 42-43." with a DOI link: "https://doi.org/10.1093/cid/ciw660". The article was published on "26 September 2016" and has an "Article history" link.

Evaluation of Internet-Based Dengue Query Data: Google Dengue Trends

Rebecca Tave Gluskin^{1*}, Michael A. Johansson², Mauricio Santillana³, John S. Brownstein¹

¹ Children's Hospital Informatics Program, Children's Hospital Boston, Boston, Massachusetts, United States of America, ² Dengue Branch, Division of Vector-Borne Diseases, Centers for Disease Control and Prevention, San Juan, Puerto Rico, ³ School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts, United States of America

While Google Dengue Trends captures well the national incidence of disease

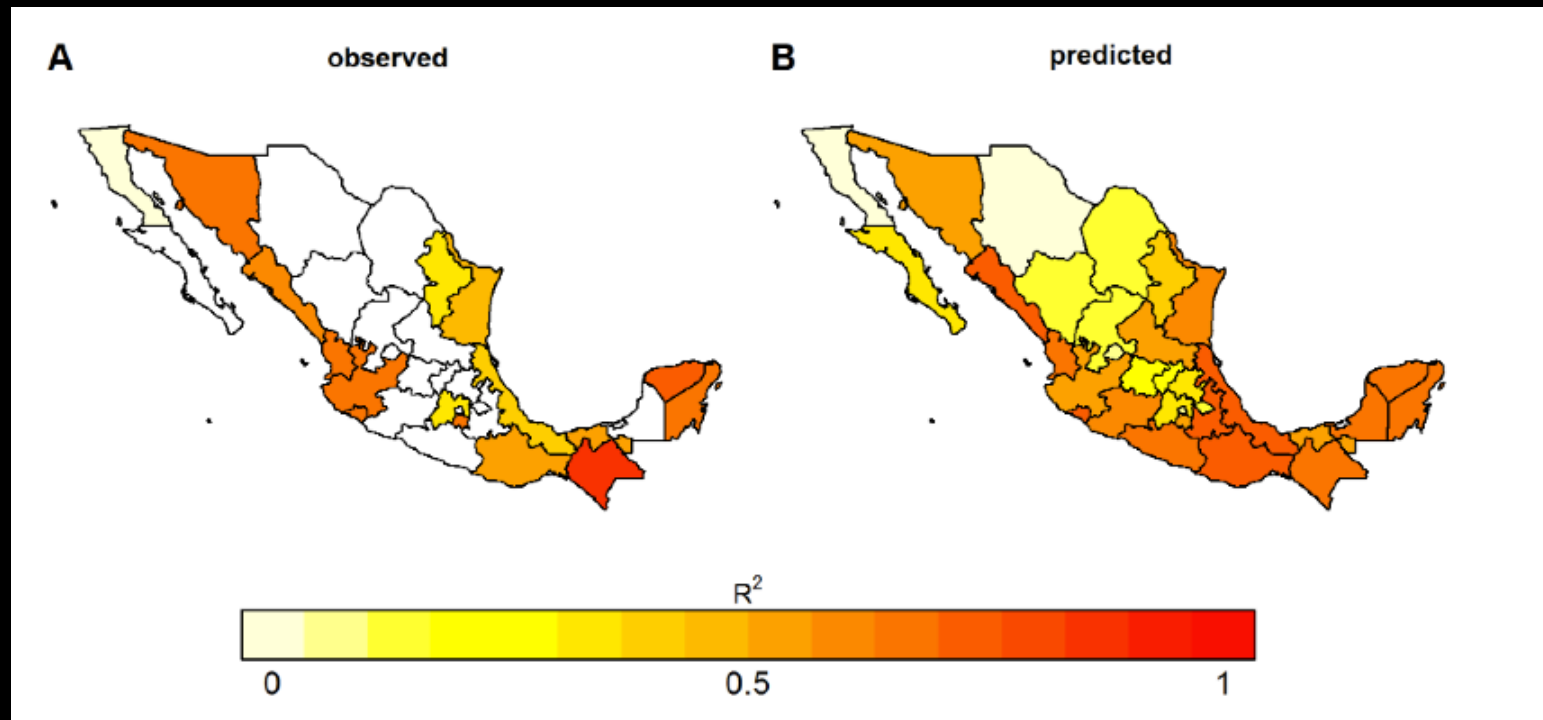


Evaluation of Internet-Based Dengue Query Data: Google Dengue Trends

Rebecca Tave Gluskin^{1*}, Michael A. Johansson², Mauricio Santillana³, John S. Brownstein¹

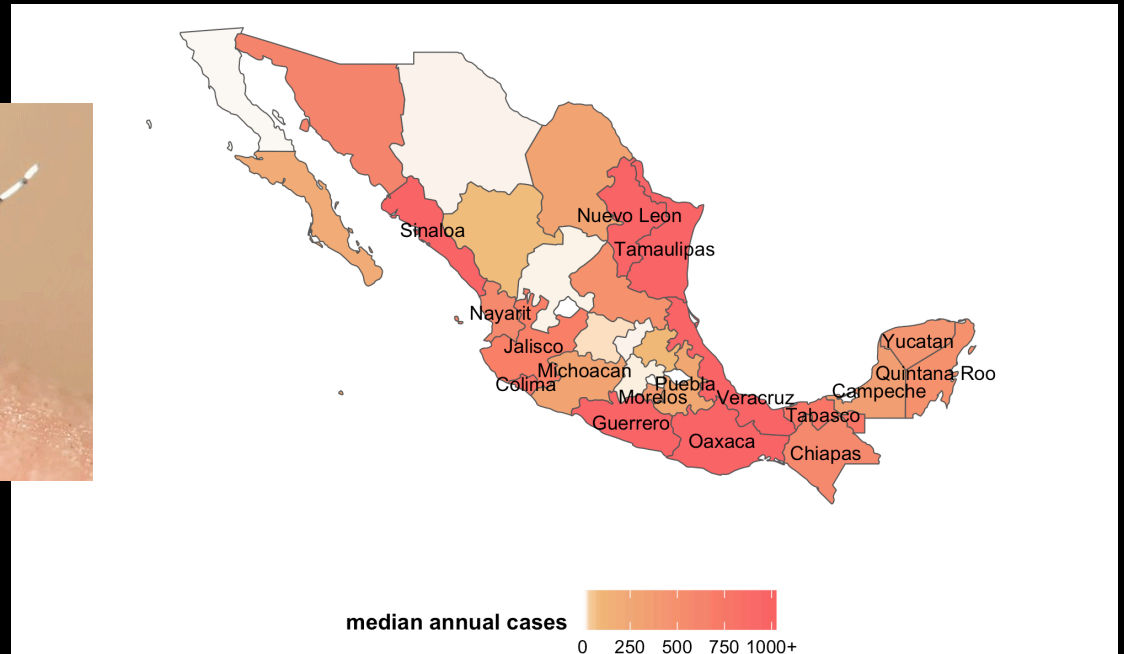
¹ Children's Hospital Informatics Program, Children's Hospital Boston, Boston, Massachusetts, United States of America, ² Dengue Branch, Division of Vector-Borne Diseases, Centers for Disease Control and Prevention, San Juan, Puerto Rico, ³ School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts, United States of America

It fails to capture the incidence of dengue at the state level in multiple cases



Forecasting Dengue Incidence in Mexico

Establishing a prediction baseline



Team:

Mauricio Santillana (BCH, Harvard),
Michael Johansson (CDC Puerto Rico),
Aditi Hota (Columbia Univ),
John Brownstein (BCH, Harvard),
Nick Reich (Umass Amherst)



Altmetric: 10 Citations: 4

[More detail >>](#)

Article | [OPEN](#)

Evaluating the performance of infectious disease forecasts: A comparison of climate-driven and seasonal dengue forecasts for Mexico

Michael A. Johansson , Nicholas G. Reich, Aditi Hota, John S. Brownstein & Mauricio Santillana 

Scientific Reports **6**, Article number: 33707
(2016)

doi:10.1038/srep33707

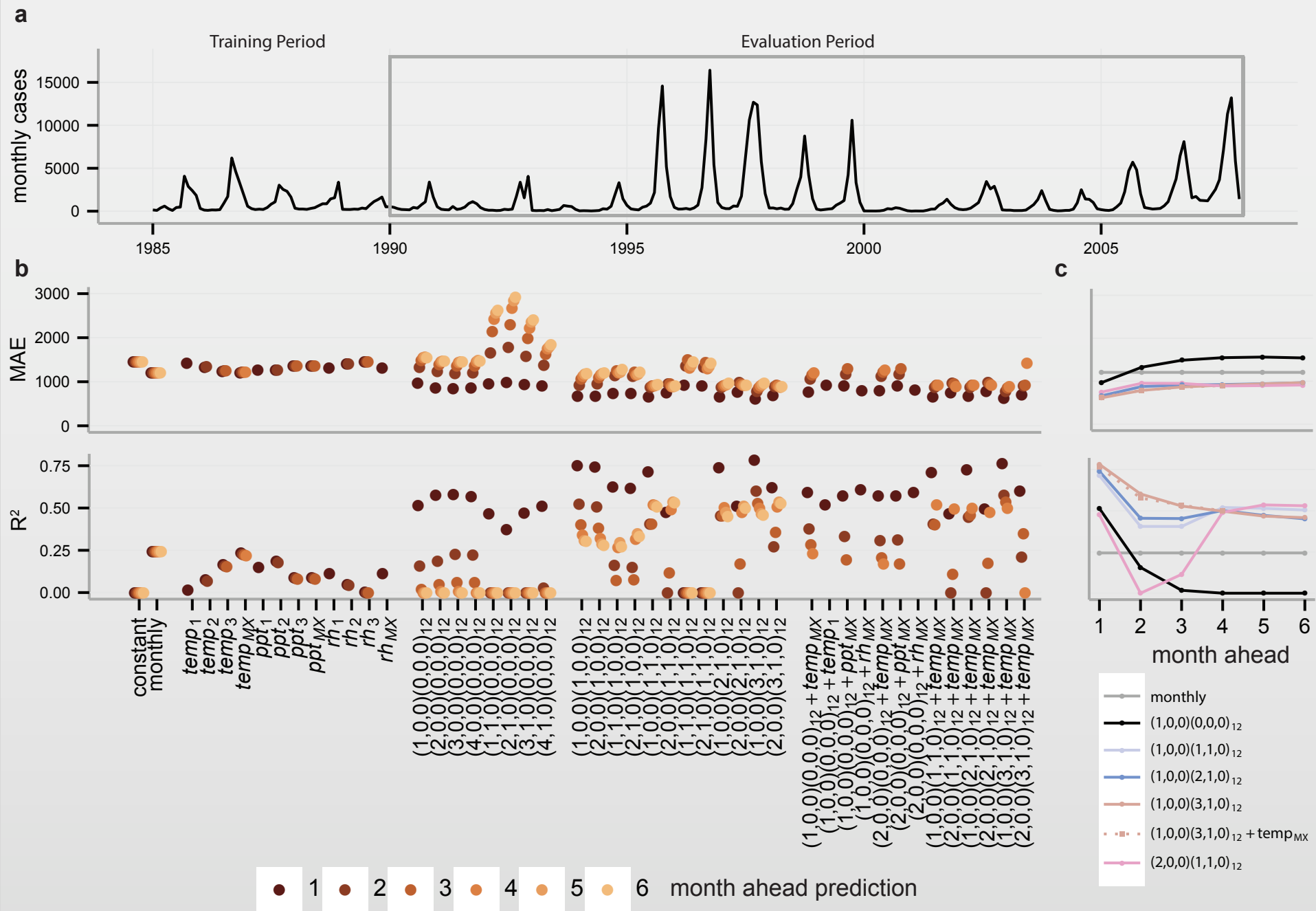
[Download Citation](#)

Received: 17 March 2016

Accepted: 24 August 2016

Published online: 26 September 2016

Mexico Dengue incidence (Country-level)





Extending use of Google searches to track Dengue in other countries:


Latin America: Mexico, Brazil

Southeast Asia: Thailand, Singapore, Taiwan

 OPEN ACCESS  PEER-REVIEWED

RESEARCH ARTICLE

Advances in using Internet searches to track dengue

Shihao Yang, Samuel C. Kou , Fred Lu, John S. Brownstein, Nicholas Brooke, Mauricio Santillana 

Published: July 20, 2017 • <https://doi.org/10.1371/journal.pcbi.1005607>

Article

Authors

Metrics

Comments

Related Content



Abstract

Author summary

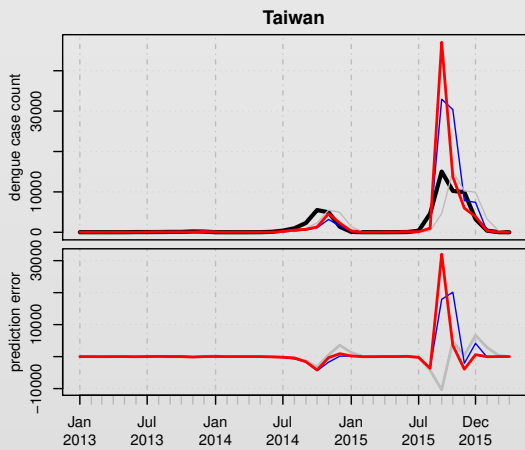
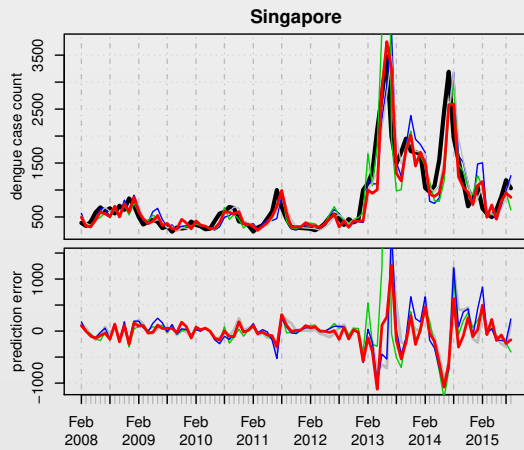
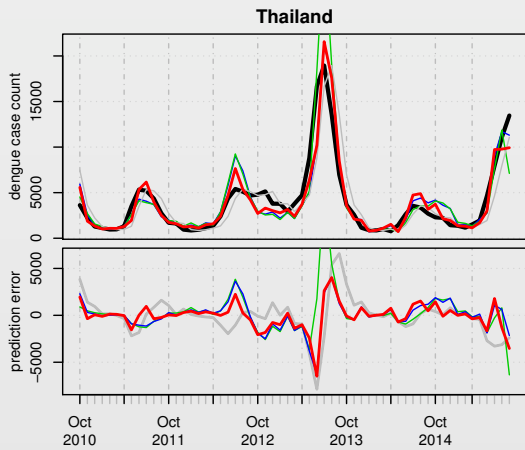
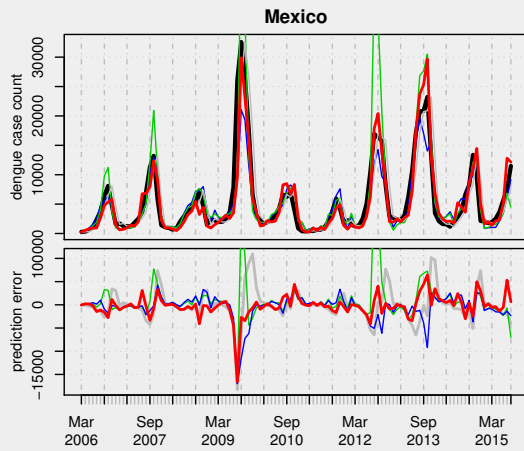
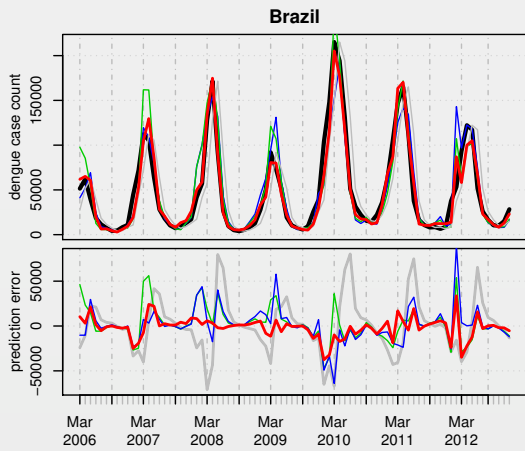
Introduction

Materials and methods

Results

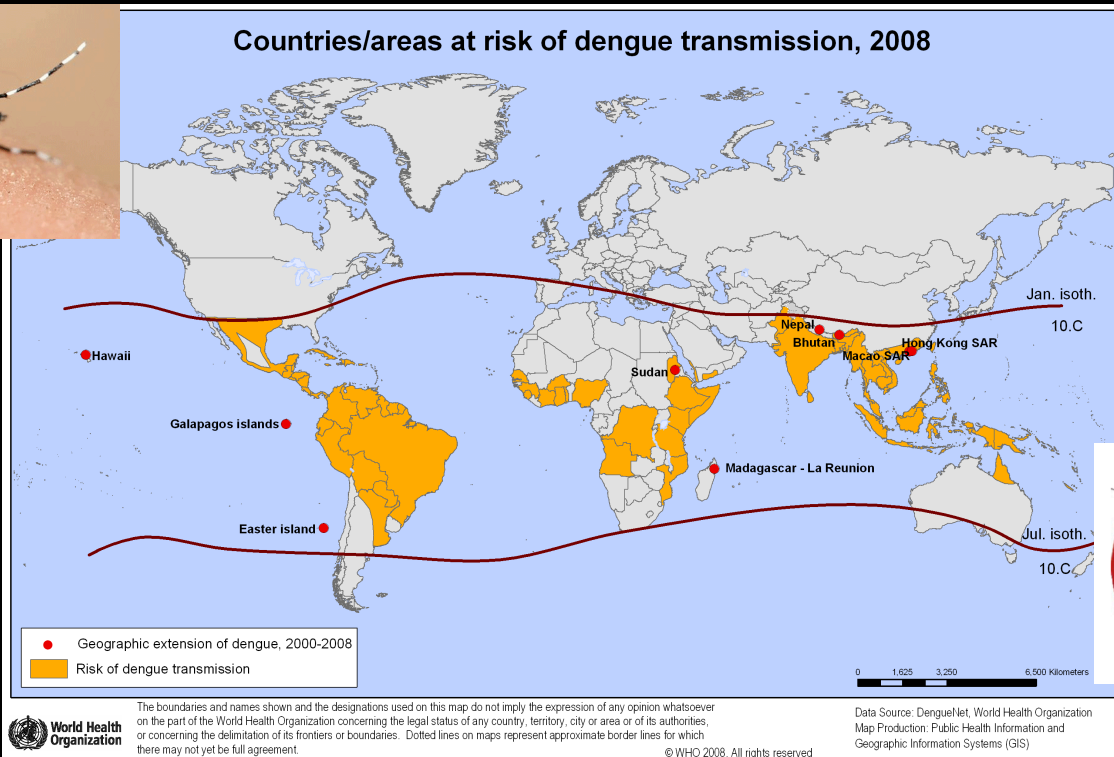
Abstract

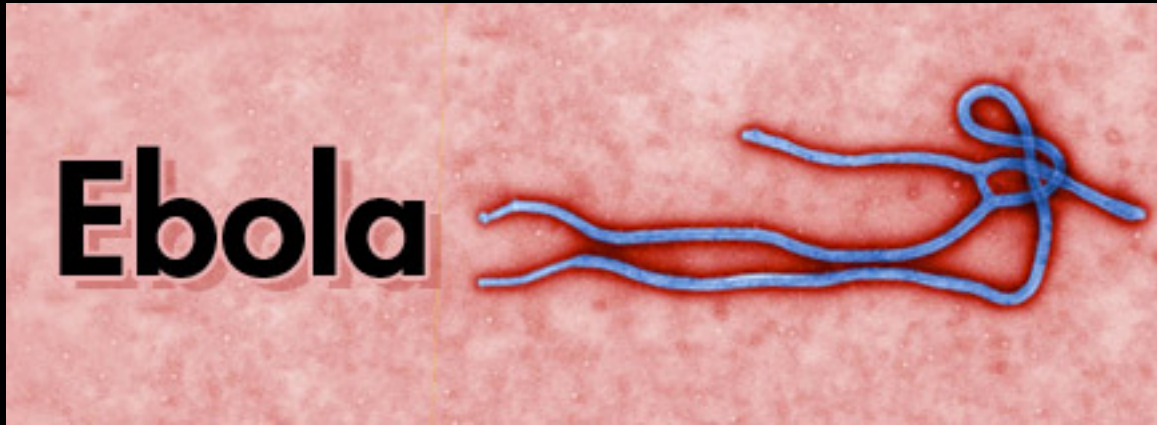
Dengue is a mosquito-borne disease that threatens over half of the world's population. Despite being endemic to more than 100 countries, government-led efforts and tools for timely identification and tracking of new infections are still lacking in many affected areas. Multiple methodologies that leverage the use of Internet-based data sources have been proposed as a



- Target
- ARGO
- SAR
- SAR + GDT
- naive

New alliance: Identifying hot spots for Dengue outbreaks



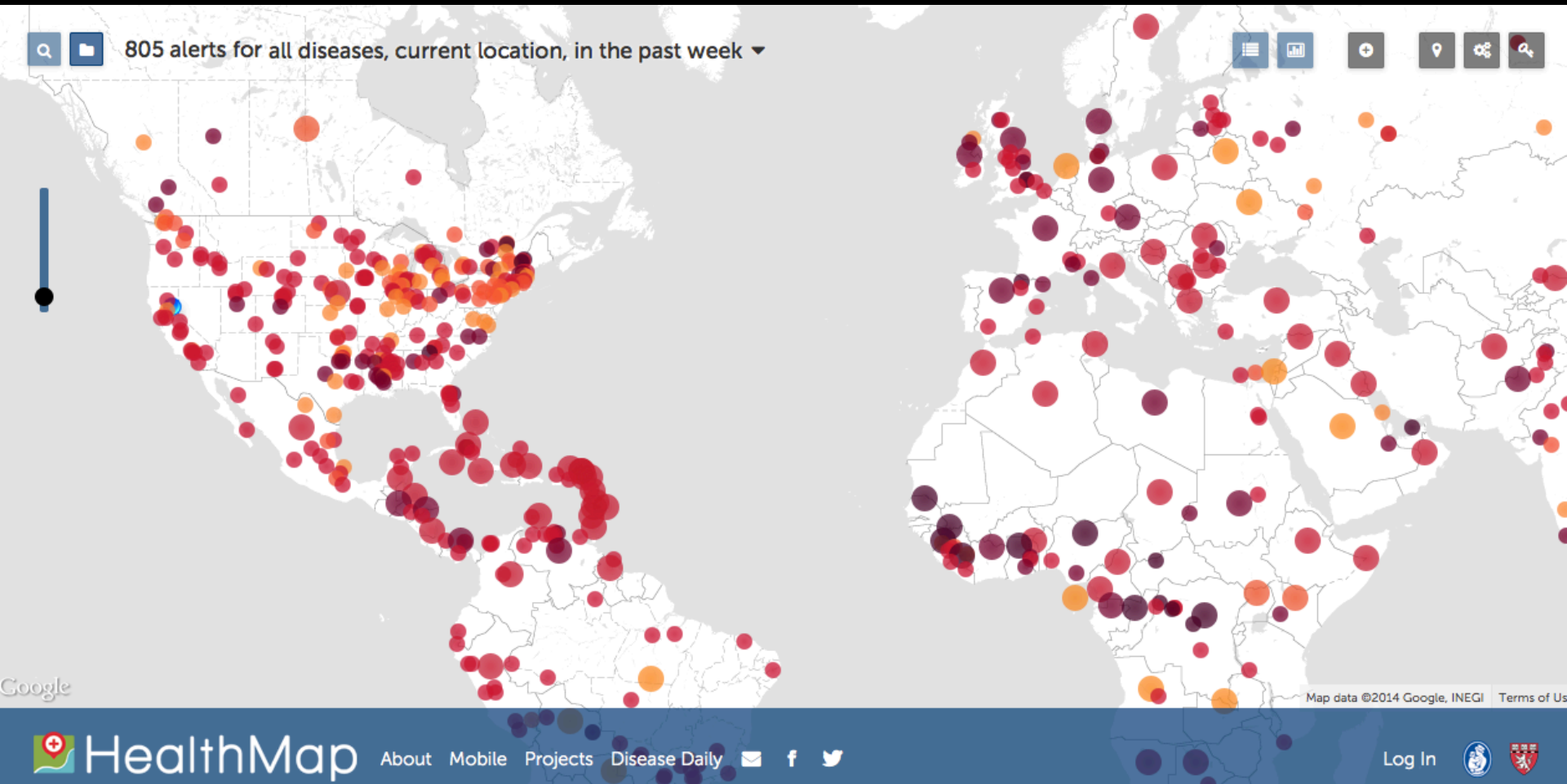


What if we could use **news reports** as a way to modulate predictions produced with models?

An example from the **Ebola** outbreak in 2015

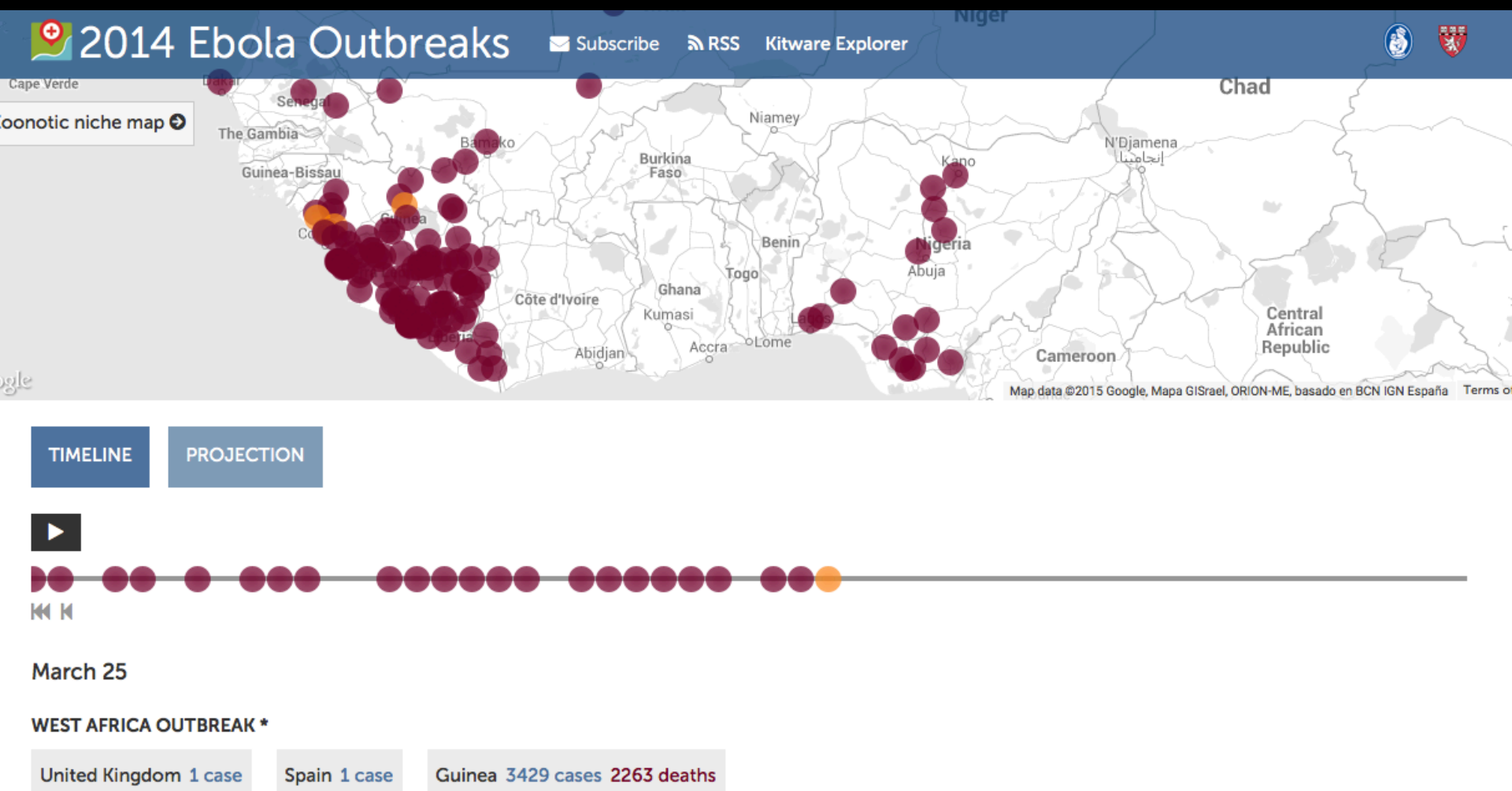
Healthmap.org

HealthMap brings together disparate data sources, including online news aggregators, eyewitness reports, expert-curated discussions and validated official reports, to achieve a unified and comprehensive view of the current global state of infectious diseases and their effect on human and animal health.



Through an automated process, updating 24/7/365, the system monitors, organizes, integrates, filters, visualizes and disseminates online information about emerging diseases in nine languages, facilitating early detection of global public health threats

Recent success story: Ebola outbreak identification and tracking




<http://www.healthmap.org/ebola/#timeline>

2014 Ebola Outbreak: Media Events Track Changes in Observed Reproductive Number

APRIL 28, 2015 · COMMENTARY

 [Print or Save PDF](#)

 [Citation](#)

 [XML](#)

 [Email](#)

 [Tweet](#)

 [Like](#)

 10

■ AUTHORS

[Maimuna S. Majumder](#) [Sheryl Kluberg](#) [Mauricio Santillana](#) [Sumiko Mekaru](#) [John S. Brownstein](#)

■ ABSTRACT

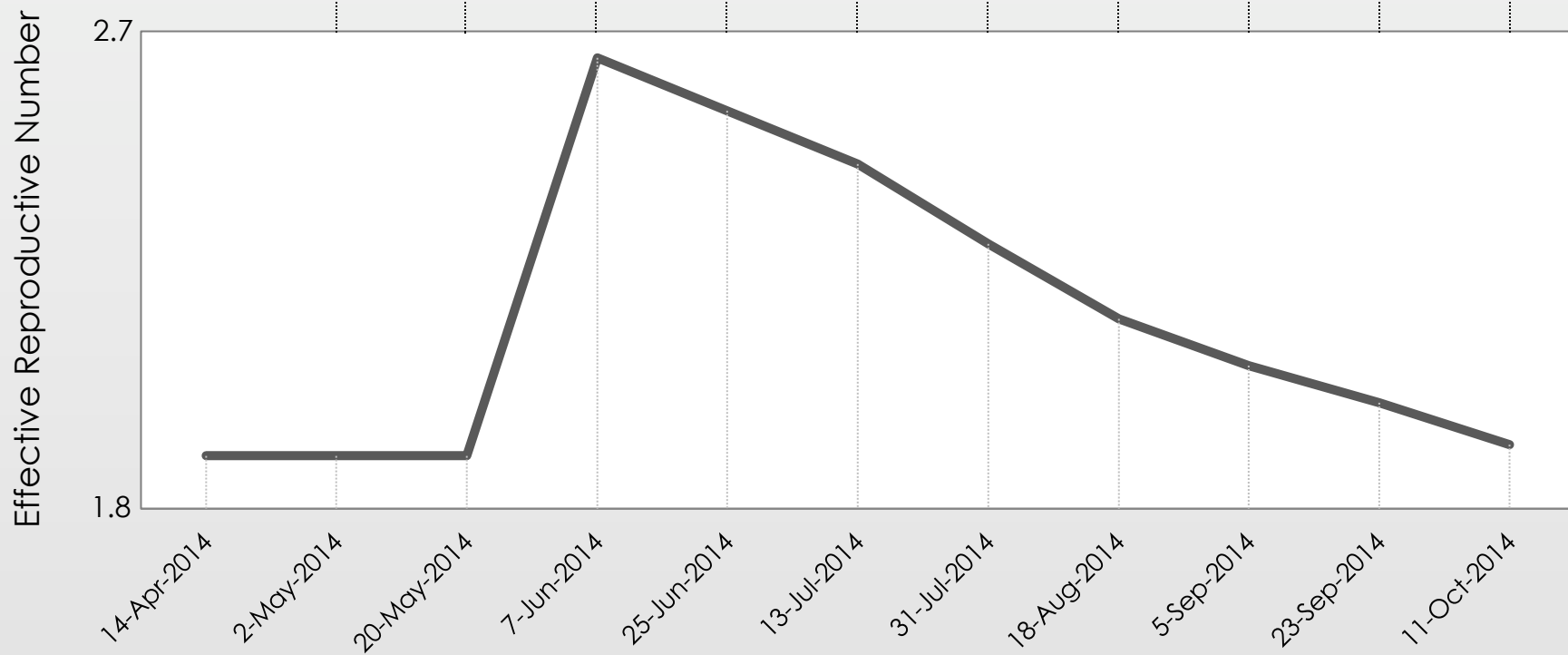
In this commentary, we consider the relationship between early outbreak changes in the observed reproductive number of Ebola in West Africa and various media reported interventions and aggravating events. We find that media reports of interventions that provided education, minimized contact, or strengthened healthcare were typically followed by sustained transmission reductions in both Sierra Leone and Liberia. Meanwhile, media reports of aggravating events generally preceded temporary transmission increases in both countries. Given these preliminary findings, we conclude that media reported events could potentially be incorporated into future epidemic modeling efforts to improve mid-outbreak case projections.

Strengthening Healthcare

Providing Education

Minimizing Contact

Aggravating Event



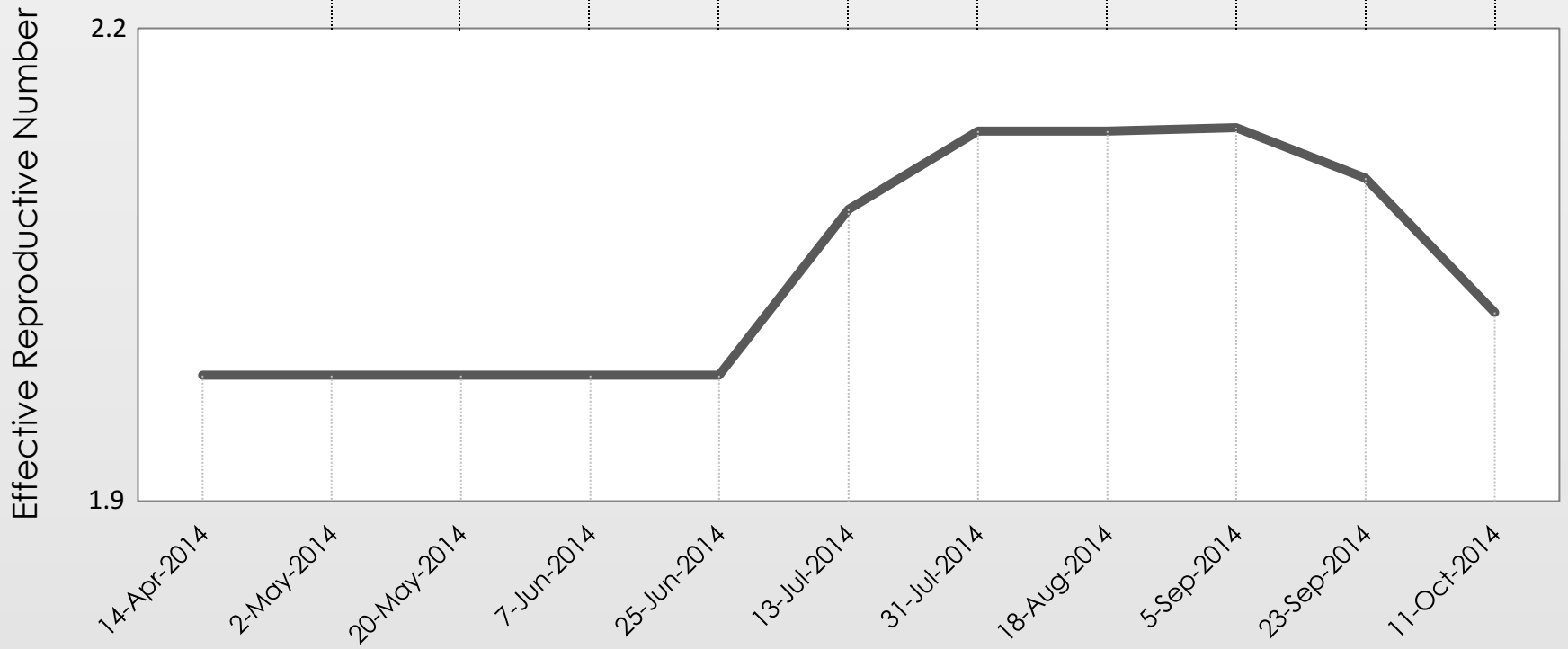
Sierra Leone

Strengthening Healthcare

Providing Education

Minimizing Contact

Aggravating Event



Liberia

A more recent contribution on the 2015 Latin American Zika outbreak



Data-poor environments (Zika)

A more recent contribution on the 2015 Latin American Zika outbreak



JMIR Publications



JMIR Public Health and Surveillance

Published on 01.06.16 in Vol 2, No 1 (2016): Jan-Jun

This paper is in the following e-collection/theme issue:

[Infoveillance, Infodemiology and Digital Disease Surveillance](#) [Infodemiology and Infoveillance](#)

Article

Cited By (0)

Tweetations (29)

Metrics

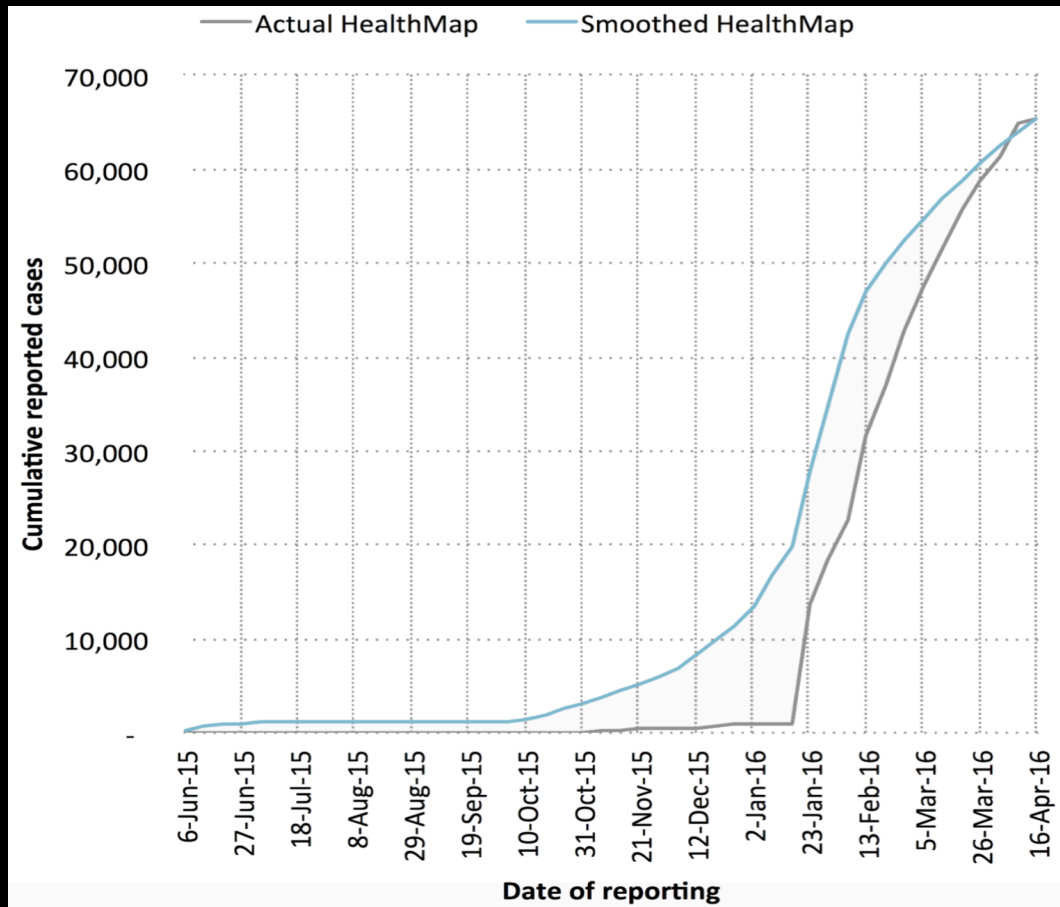
Original Paper

Utilizing Nontraditional Data Sources for Near Real-Time Estimation of Transmission Dynamics During the 2015-2016 Colombian Zika Virus Disease Outbreak

Maimuna S Majumder^{1,2}, MPH ; Mauricio Santillana^{1,3,4}, PhD ; Sumiko R Mekaru^{1,5}, PhD ; Denise P McGinnis¹, ScD ;
Kamran Khan^{6,7}, MD ; John S Brownstein^{1,4}, PhD

Data-poor environments (Zika)

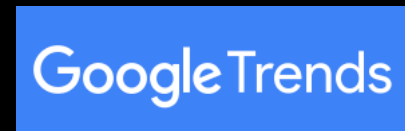
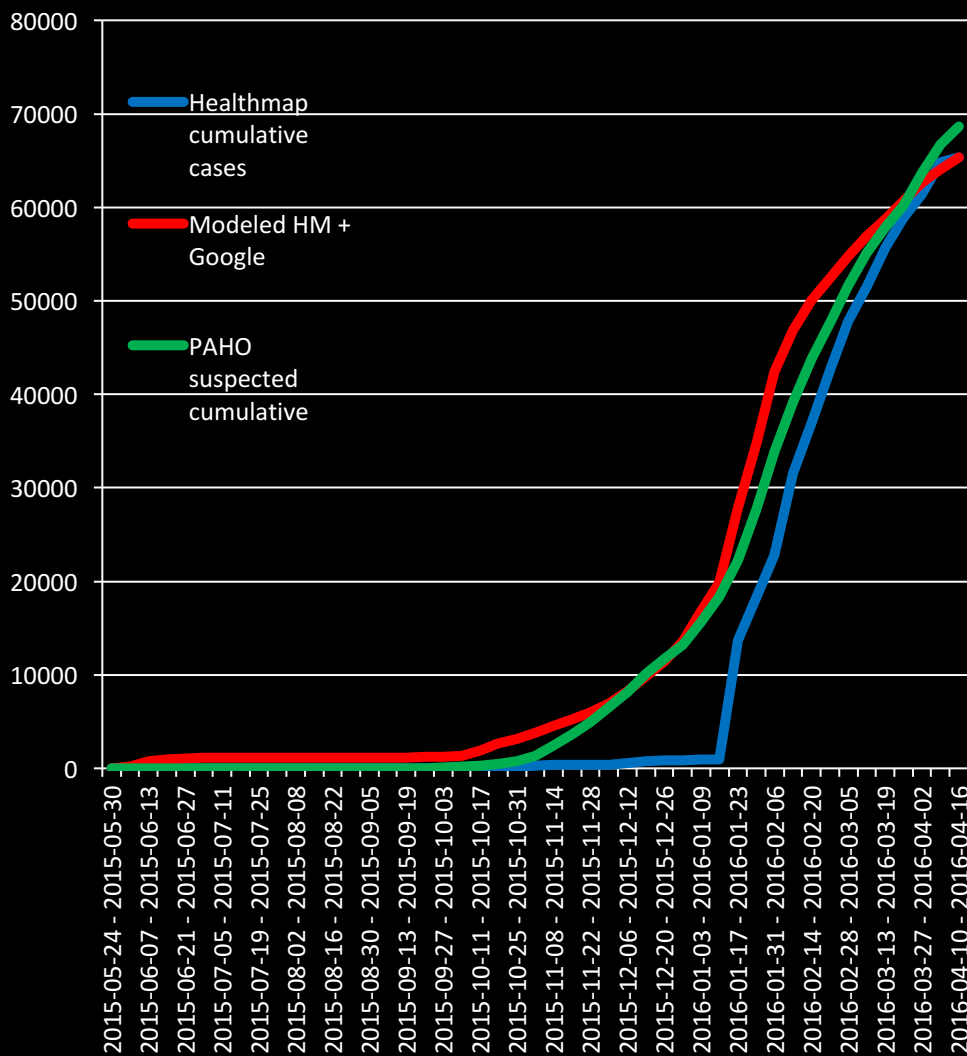
A more recent contribution on the 2015 Latin American Zika outbreak



With no access to traditional, government-lead disease surveillance information, we extracted the number of suspected cases as reported by new reports as a function of time. We then utilized the time behavior of Google searches of the word “zika” to smooth the news-reported incidence data.

Data-poor environments (Zika)

A more recent contribution on the 2015 Latin American Zika outbreak



When we gained access to government-lead disease surveillance information, we found great similarity with the curve we produced ahead of the publication of this information.

Data-poor environments (Cholera)

Extending work to characterize Cholera outbreak in Yemen

Cholera in Yemen

216 patients
Since 25th April

263 patients
Since 30th March,
168 of them were received in
the last two weeks

12 patients
Since 7th May

20 patients
Since 3rd May

276 patients
Since 23rd April




* Patients received by MSF

Forecasting Zika using Google searches and Twitter

 OPEN ACCESS  PEER-REVIEWED

RESEARCH ARTICLE

Forecasting Zika Incidence in the 2016 Latin America Outbreak Combining Traditional Disease Surveillance with Search, Social Media, and News Report Data

Sarah F. McGough , John S. Brownstein, Jared B. Hawkins, Mauricio Santillana 

Version 2  Published: January 13, 2017 • <http://dx.doi.org/10.1371/journal.pntd.0005295>

15 Save	0 Citation
3,819 View	17 Share

Article 	Authors	Metrics	Comments	Related Content
--	----------------	----------------	-----------------	------------------------

Download PDF 

Print **Share**

 Check for updates

- Abstract**
- Author Summary
- Introduction
- Methods
- Results
- Discussion
- Supporting Information

Abstract

Background

Over 400,000 people across the Americas are thought to have been infected with Zika virus as a consequence of the 2015–2016 Latin American outbreak. Official government-led case count data in Latin America are typically delayed by several weeks, making it difficult to track the disease in a timely manner. Thus, timely disease tracking systems are needed to design and

Included in the
Following Collection

[Zika](#)

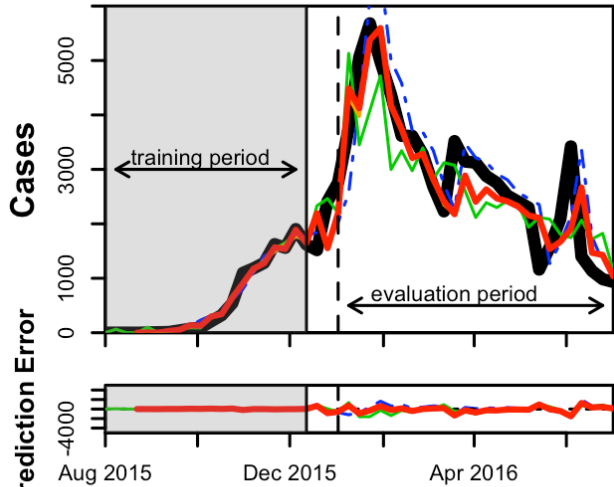
Subject Areas



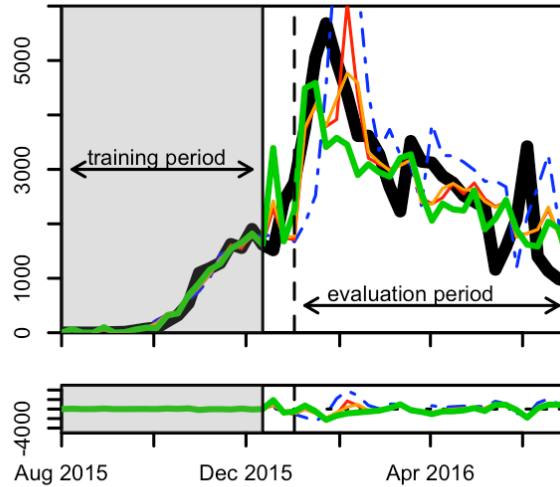
Forecasting Zika using Google searches and Twitter (with Sarah McGough)

(a) Colombia

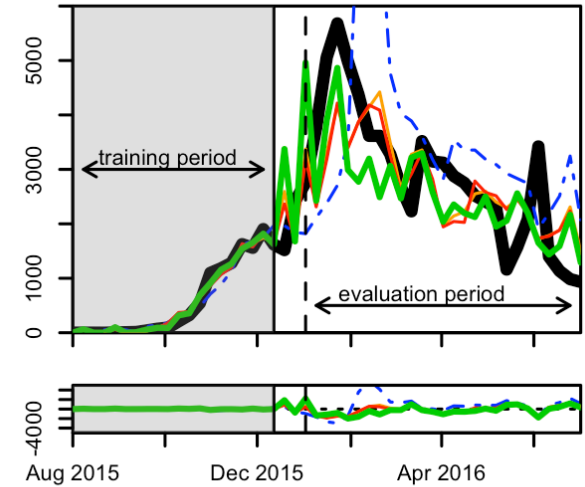
1 Week Ahead



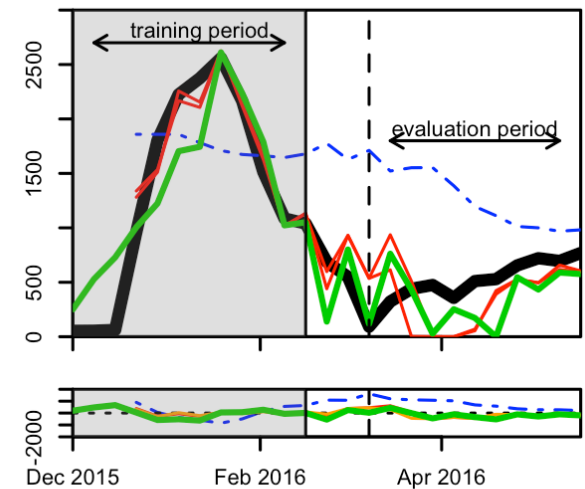
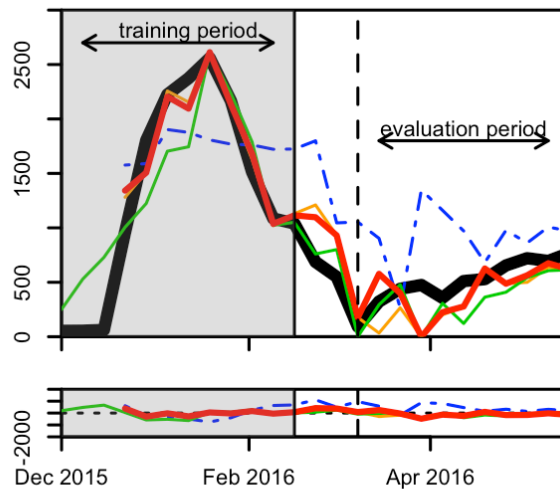
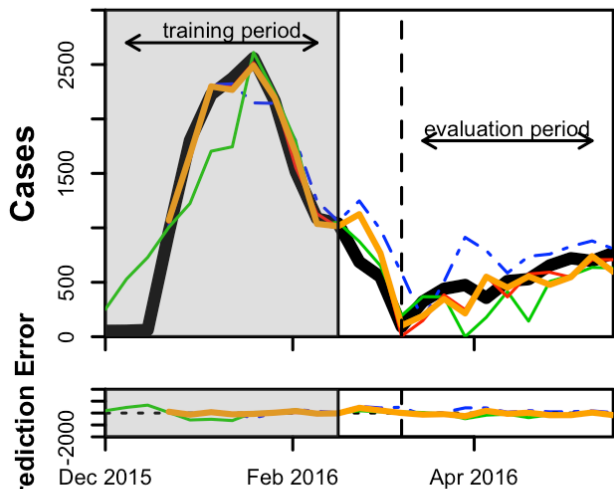
2 Weeks Ahead



3 Weeks Ahead



(b) Honduras

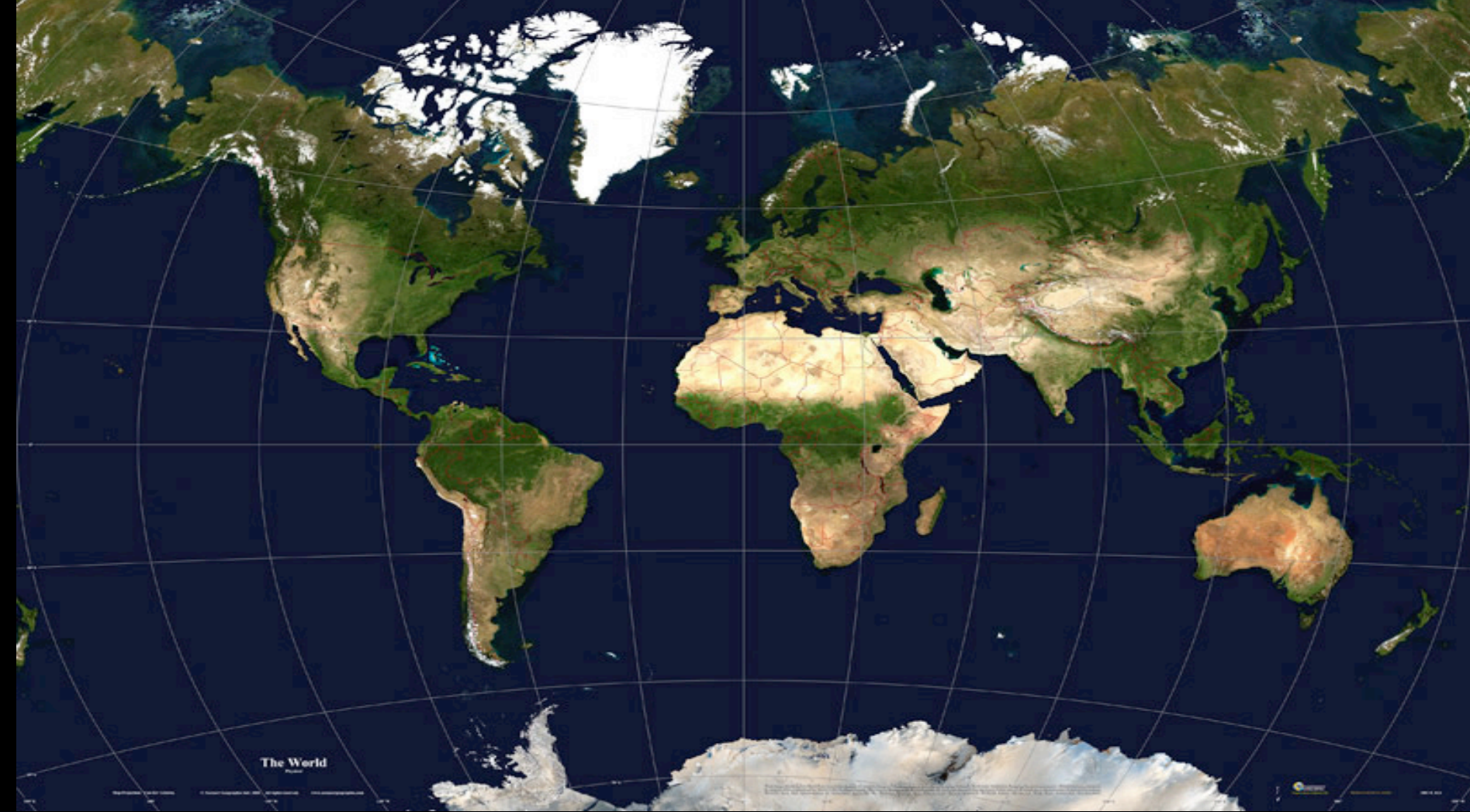


— Observed - - - AR

— G+T

— ARGO+T

— ARGO+TH



Thank you!

Contact: msantill@fas.harvard.edu