Digital Epidemiology: designing machine learning approaches to combine Internet-based data sources

Case studies characterizing flu, dengue, Zika, and Ebola epidemics.

Mauricio Santillana, PhD
Assistant Professor, Harvard Medical School
Faculty member, Computational Health Informatics Program, Boston Children's Hospital
Associate faculty, Harvard Institute of Applied Computational Sciences

HARVARD MEDICAL SCHOOL

HARVARD School of Engineering and Applied Sciences

Boston Children's Hospital

Digital Epidemiology: designing machine learning approaches to combine Internet-based data sources

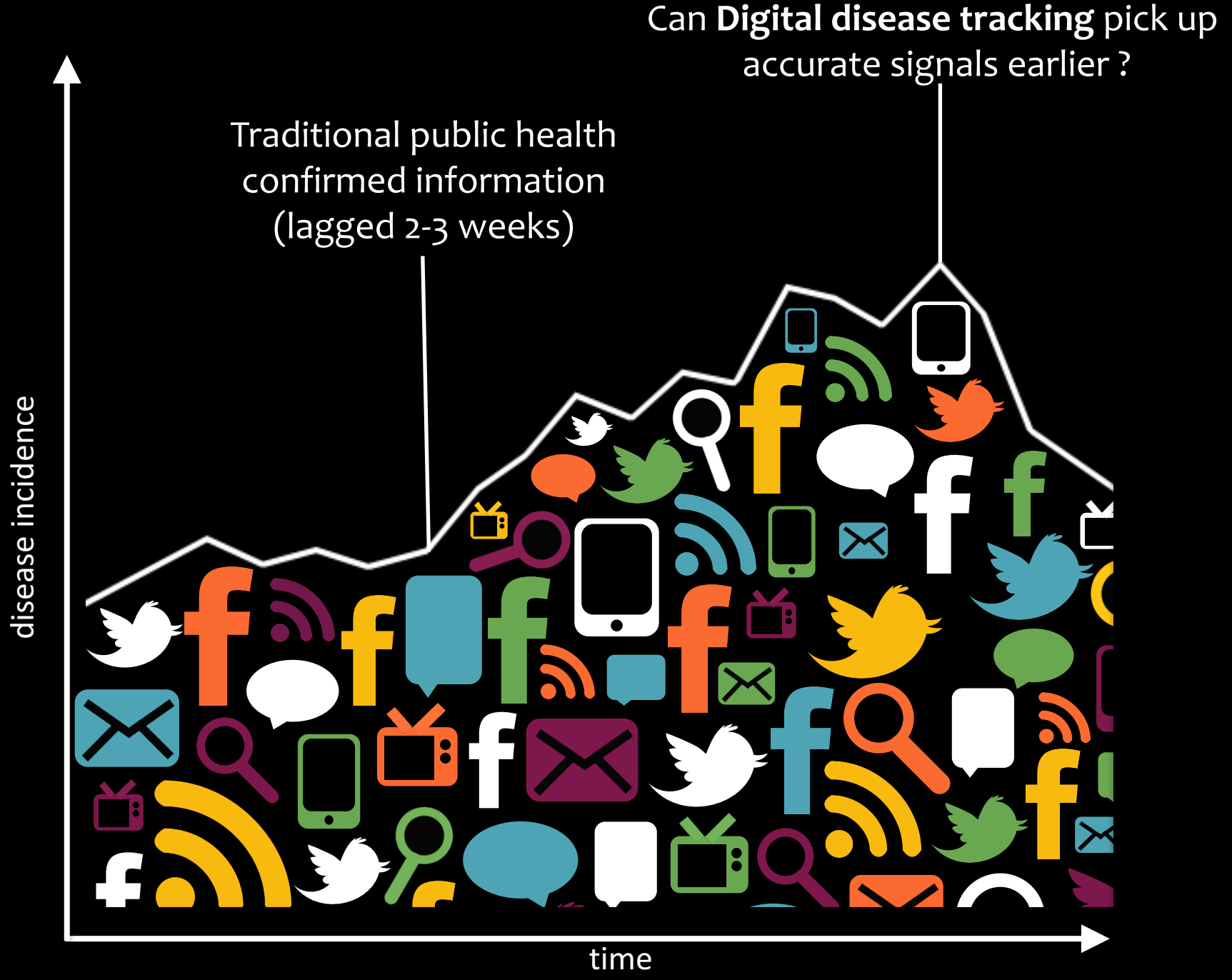
Case studies characterizing flu, dengue, Zika, and Ebola epidemics.

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), Sarah McGough (Harvard), Alessandro Vespignani (Northeastern Univ), Nathan Kutz (Univ of Washington), Elaine Nsoesie (Univ of Washington), Rumi Chunara (NYU), John Brownstein (Harvard/BCH), Leonardo C. Clemente (Inst. Politécnico Nacional, Mex), and many more

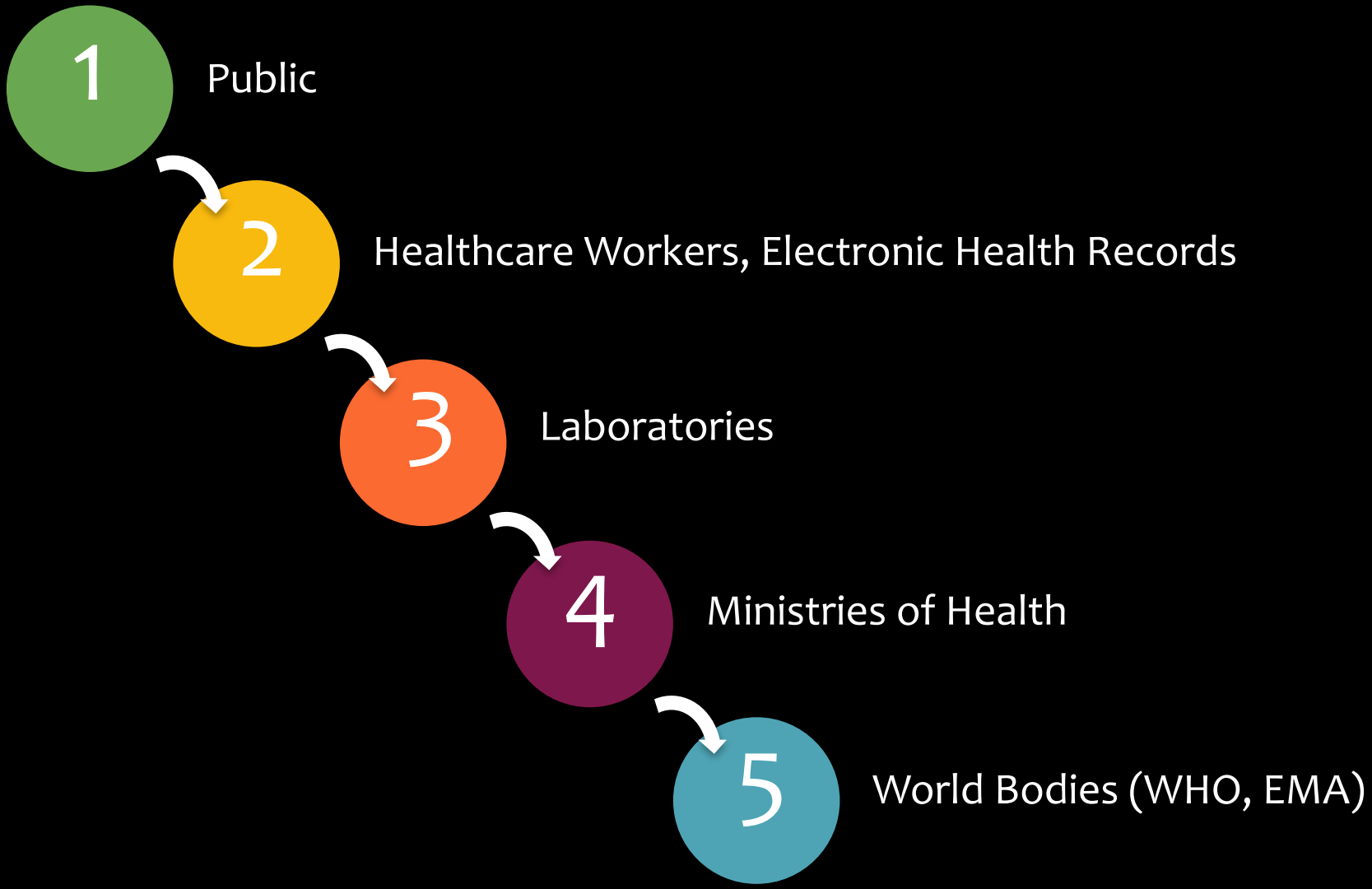
HARVARD MEDICAL SCHOOL

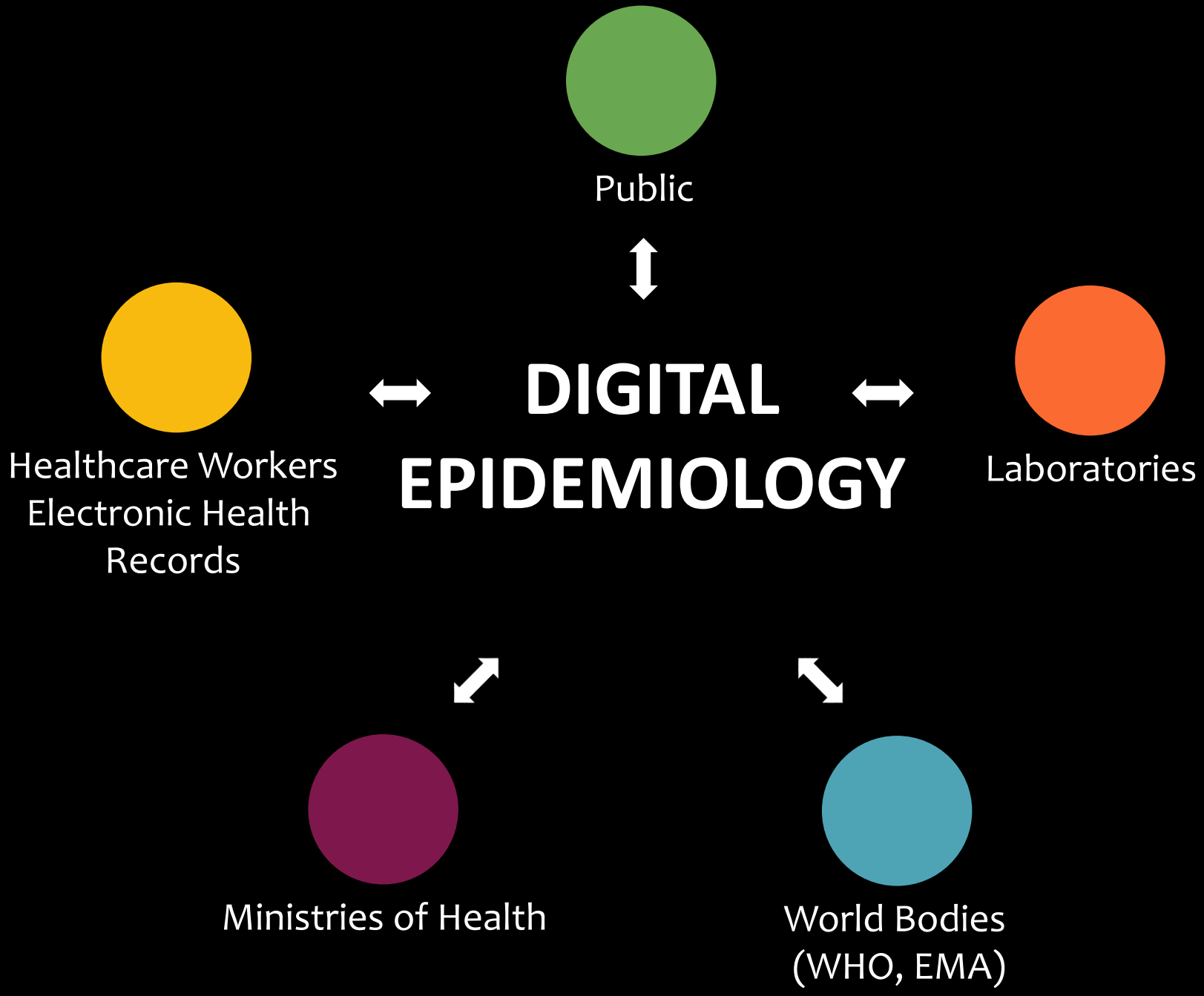
HARVARD School of Engineering and Applied Sciences

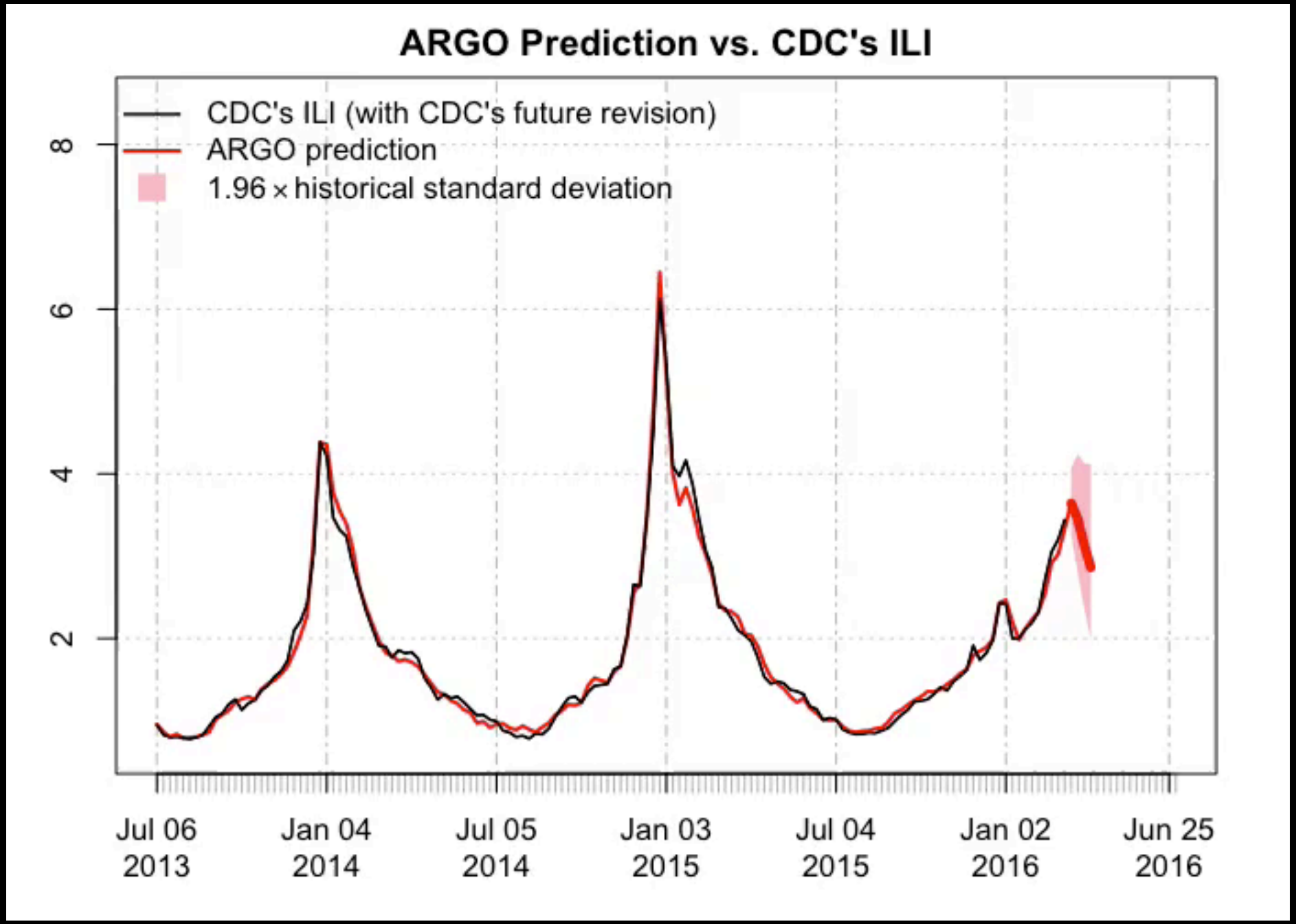
Boston Children's Hospital


Can **Digital disease tracking** pick up accurate signals earlier ?
Traditional public health confirmed information (lagged 2-3 weeks)
disease incidence
time

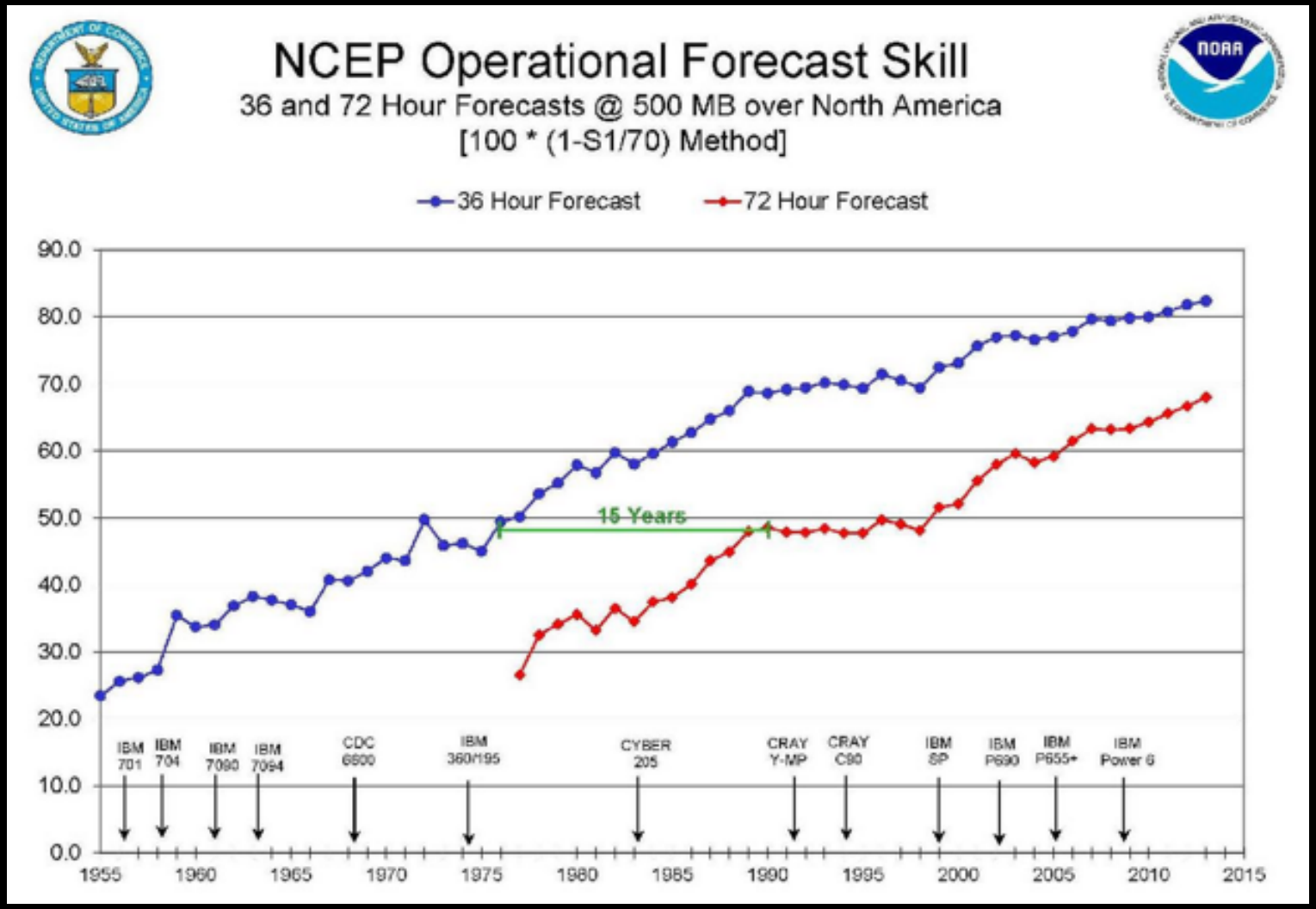
TRADITIONAL DISEASE REPORTING


1
Public
2
Healthcare Workers, Electronic Health Records
3
Laboratories
4
Ministries of Health
5
World Bodies (WHO, EMA)

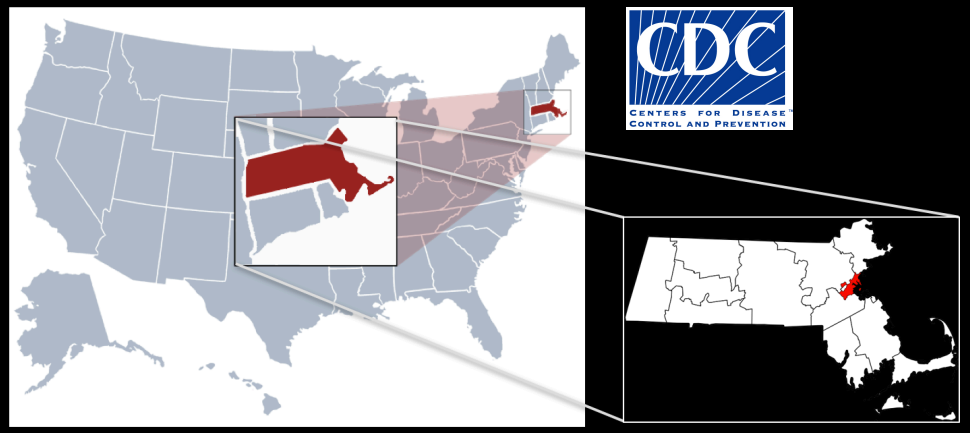

Public
DIGITAL EPIDEMIOLOGY
Healthcare Workers Electronic Health Records
Laboratories
Ministries of Health
World Bodies (WHO, EMA)

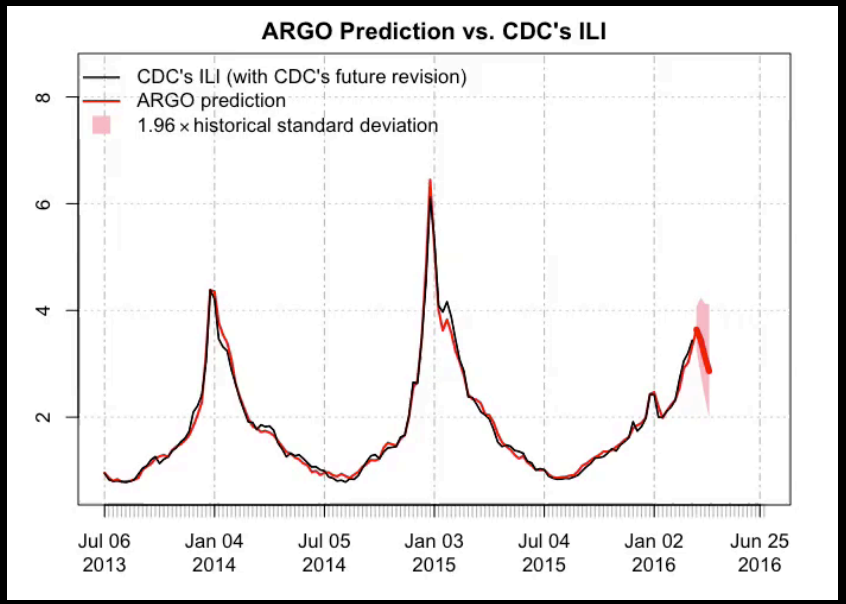

ARGO Prediction vs. CDC's ILI
CDC's ILI (with CDC's future revision)
ARGO prediction
1.96 × historical standard deviation
2
4
6
8
Jul 06 2013
Jan 04 2014
Jul 05 2014
Jan 03 2015
Jul 04 2015
Jan 02 2016
Jun 25 2016

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



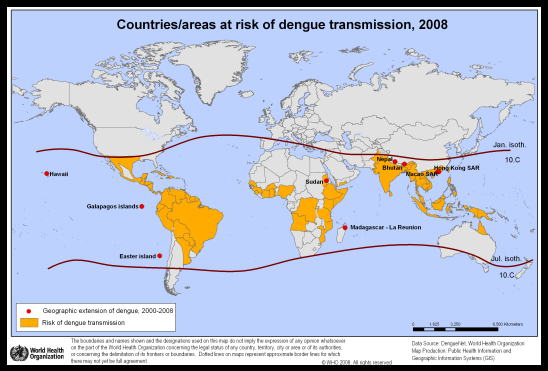


1. Multiple spatial resolutions: National, multi-state, state, city-level

2. Multiple data sources (hybrid systems): traditional healthcare-based, EHR, Google, Twitter, Crowd-sourced disease surveillance.

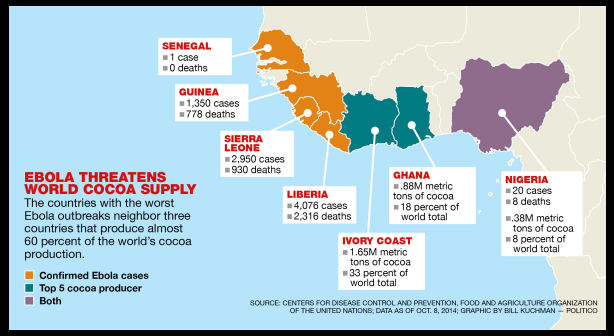
Part 2. Success stories in tracking and forecasting Flu, Zika, Dengue, Ebola in data-poor medium- to low-income countries.

Dengue, Zika, and Flu



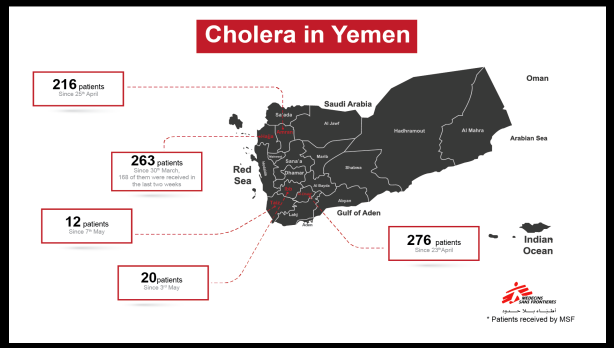
- Latin America (Flu, Zika, Dengue)
- South-east Asia (Dengue)

Ebola



- West Africa

Cholera



- Middle East

Seminal work by Google

The promise of big data in public health

GOOGLE FLU TRENDS

Google Flu Trends

nature International weekly journal of science

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

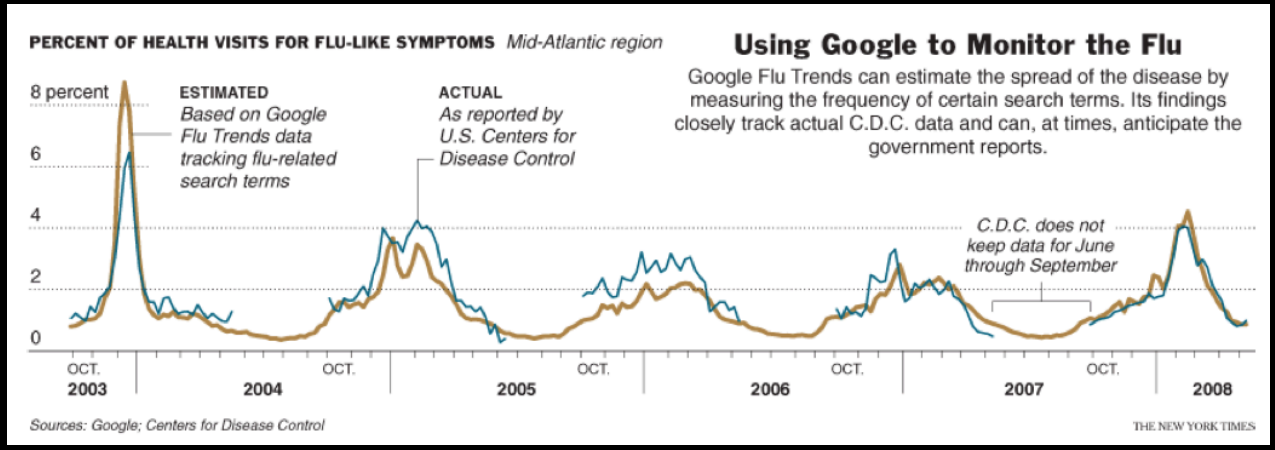
Detecting influenza epidemics using search engine query data

Jeremy Ginsberg[1], Matthew H. Mohebbi[1], Rajan S. Patel[1], Lynnette Brammer[2], Mark S. Smolinski[1] & Larry Brilliant[1]

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

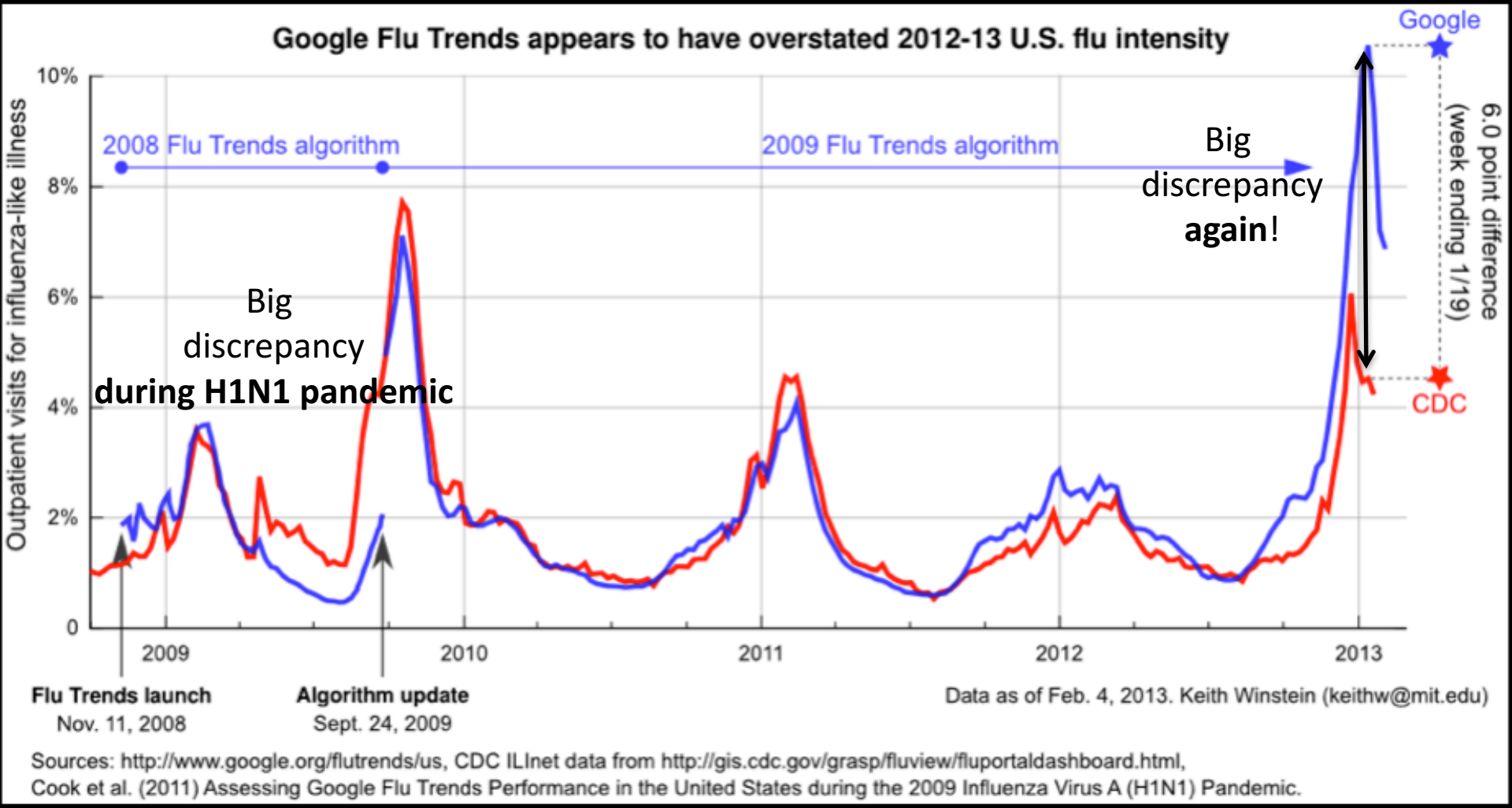
Correspondence to: Matthew H. Mohebbi[1] Correspondence and requests for materials should be addressed to J.G. or M.H.M. (Email: flutrends-support@google.com).

The New York Times



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

Big discrepancies again!



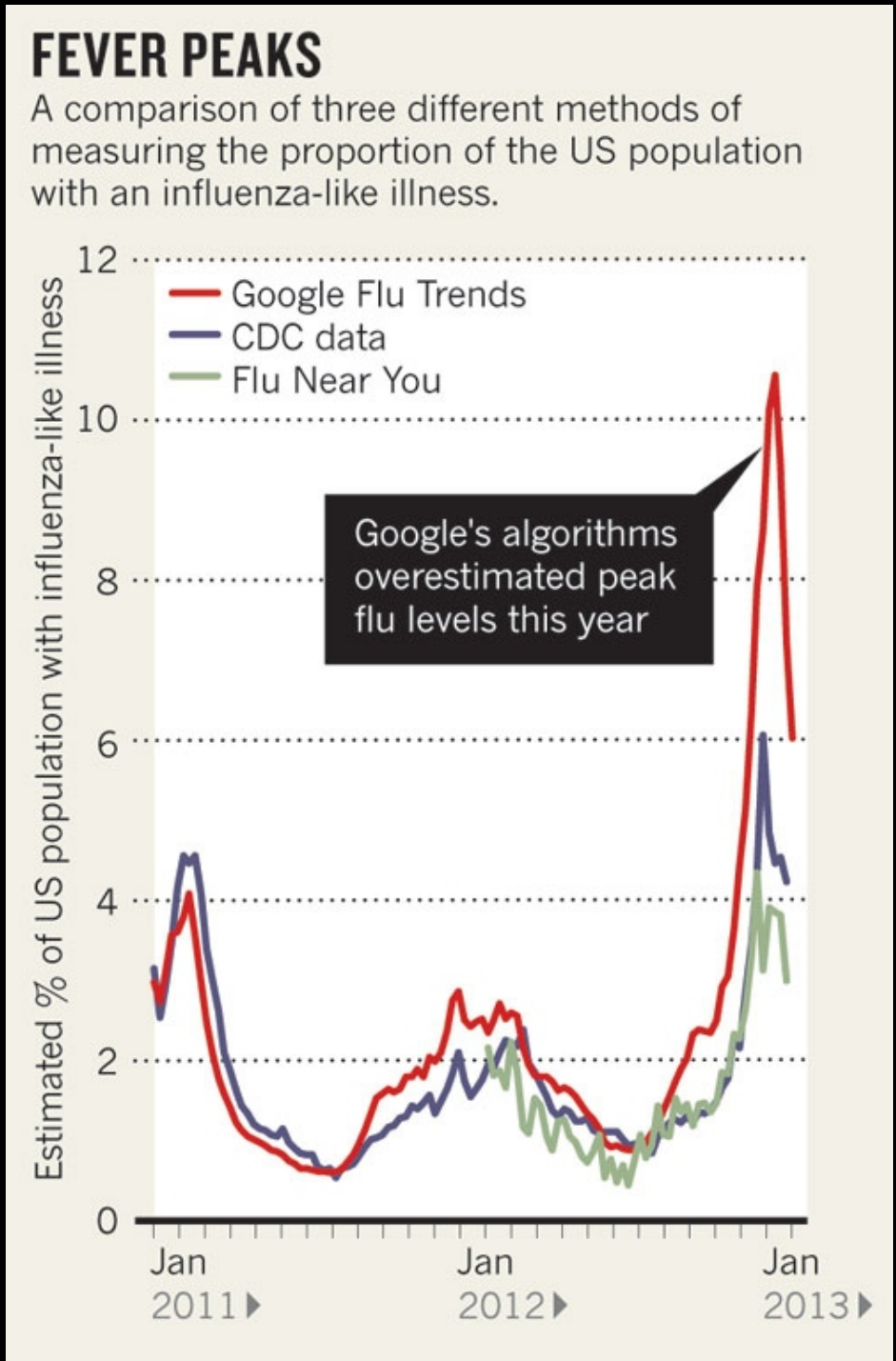
Fixes were reported in: Cook et al. (2011) Assessing Google flu trends performance in the U.S. during the 2009 influenza virus A (H1N1) pandemic. PLoS One

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

nature International weekly journal of science

When Google got flu wrong.

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



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

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

 Am J Prev Med 2014;47(3):341–347

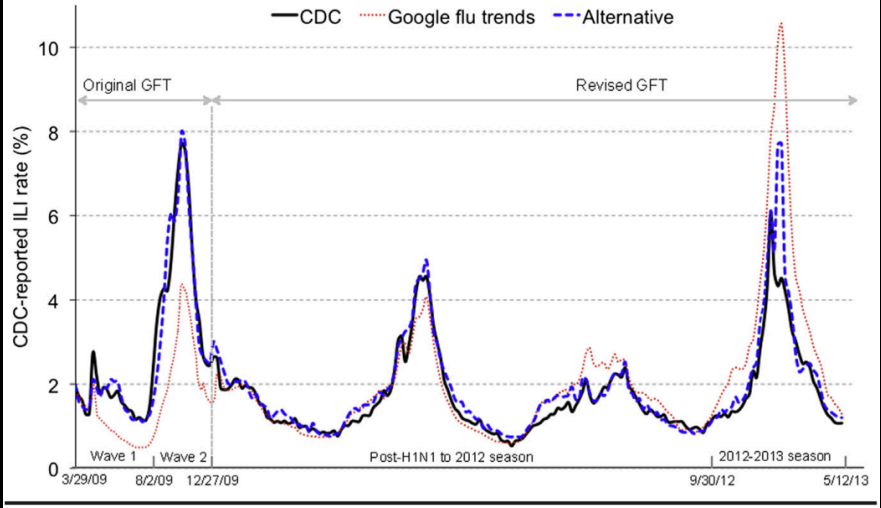


Figure 1. The alternative model outperforms Google Flu Trends

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

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

Google Research Blog

The latest news from Research at Google

Google Flu Trends gets a brand new engine

Posted: Friday, October 31, 2014

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 Lampos, Andrew C. Miller, Steve Crossan & Christian Stefansen

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

Applied mathematics
Computer science | Epidemiology
Influenza virus

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

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

PNAS

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]

[a]Department of Statistics, Harvard University, Cambridge, MA 02138; [b]School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138; and [c]Computational 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,

APPLIED MATHEMATICS

SCIENTIFIC REPORTS

PDF

Article | OPEN

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

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

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

Applied mathematics
Computer science | Epidemiology
Influenza virus

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

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

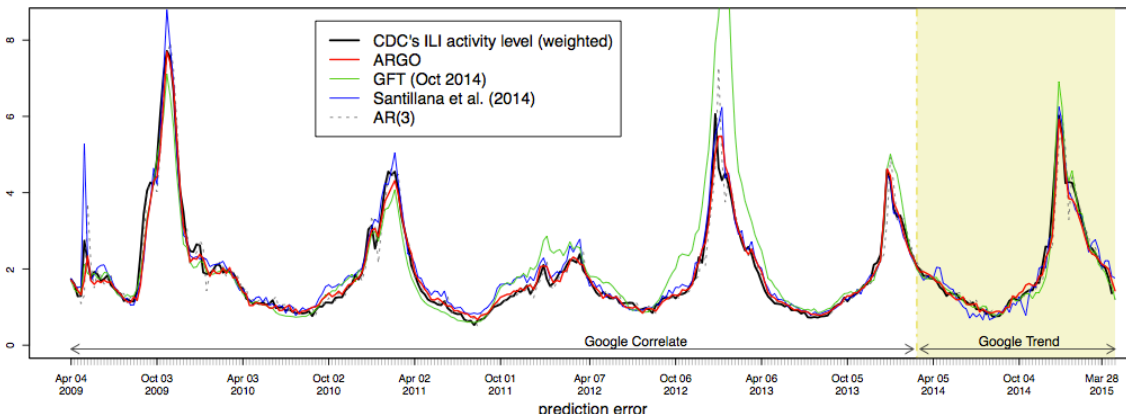
We improved last effort by Google team and published our results in PNAS in September 2015

PNAS

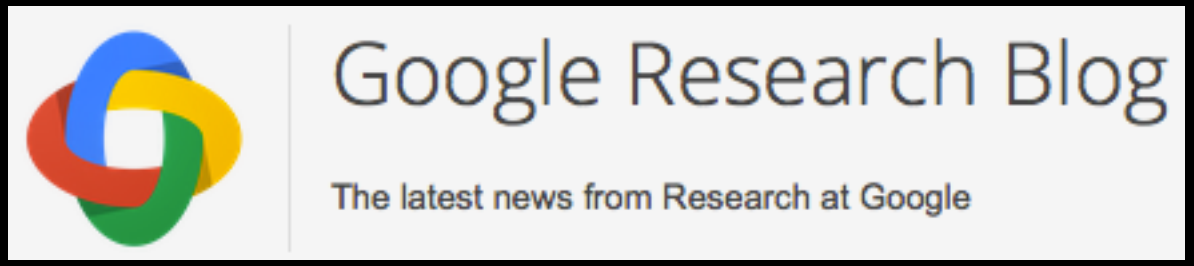
CrossMark click for updates

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]



Google discontinues Flu Trends indefinitely!



The Next Chapter for Flu Trends

Posted: Thursday, August 20, 2015

g+1 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

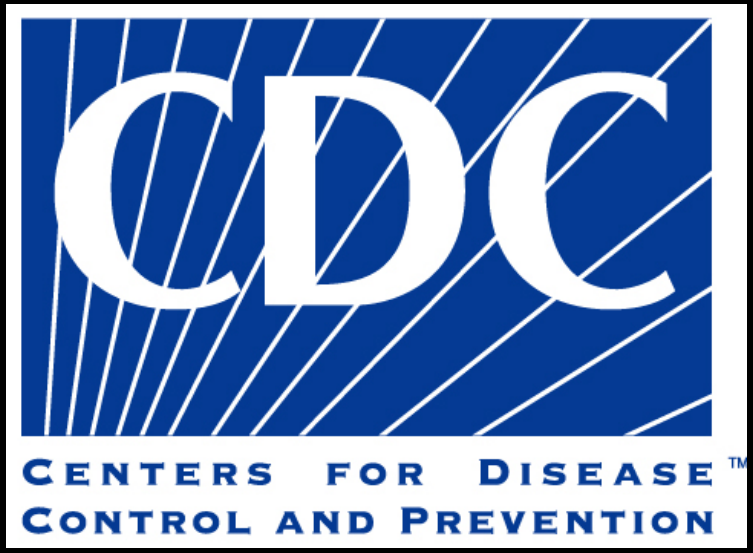
Google discontinues Flu Trends, starts offering data to researchers

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.

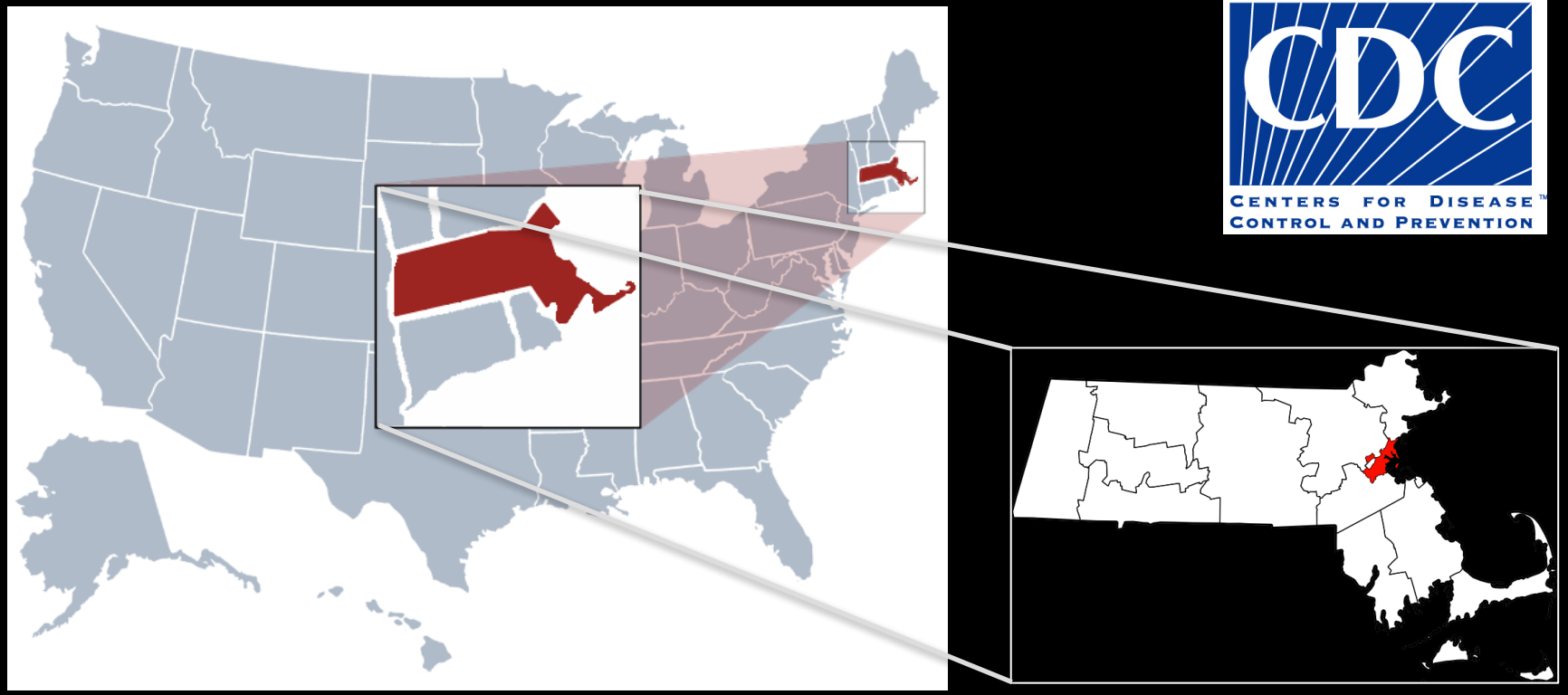
We are helping create a new improved disease forecasting platform supported partially by the Centers for Disease Control and Prevention



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*



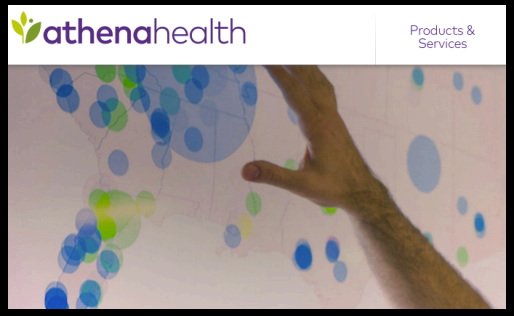
Prepared by Mauricio Santillana

Beyond Google searches…

What are doctors searching for?

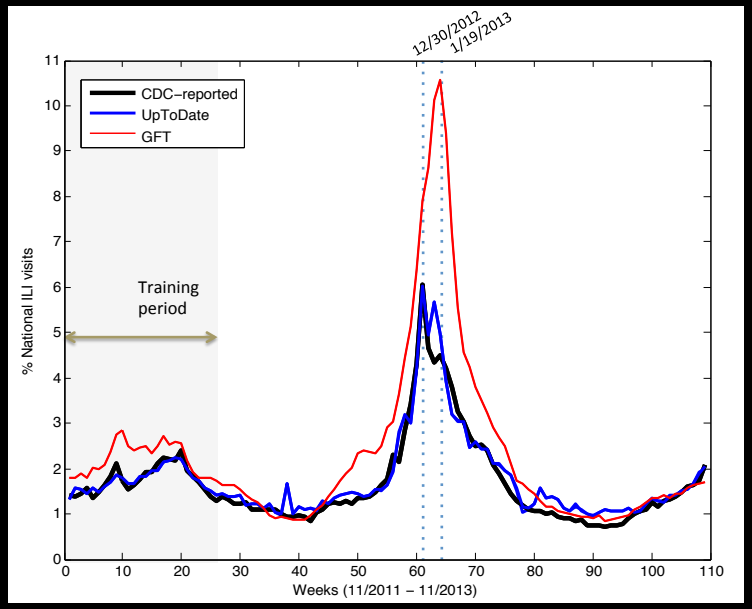
flu near you
do you have it in you?

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

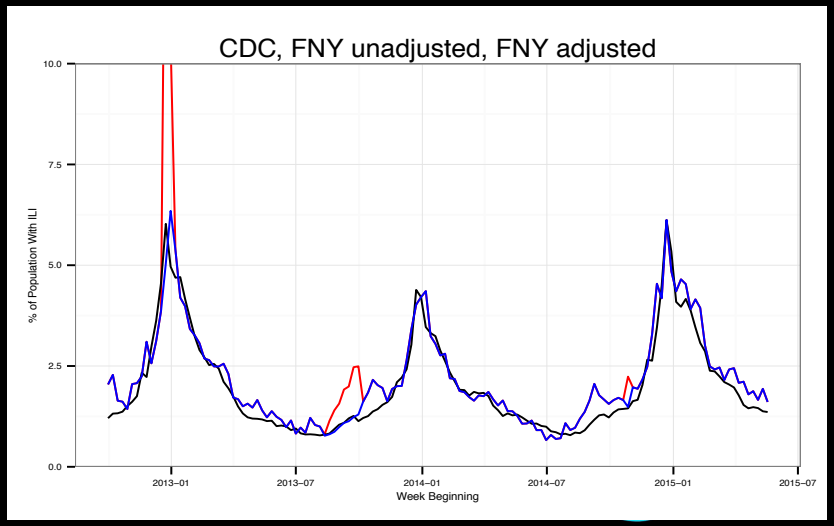


Can we use Electronic Health Records (EHR) to track disease incidence? What lab tests or medications are doctors prescribing?

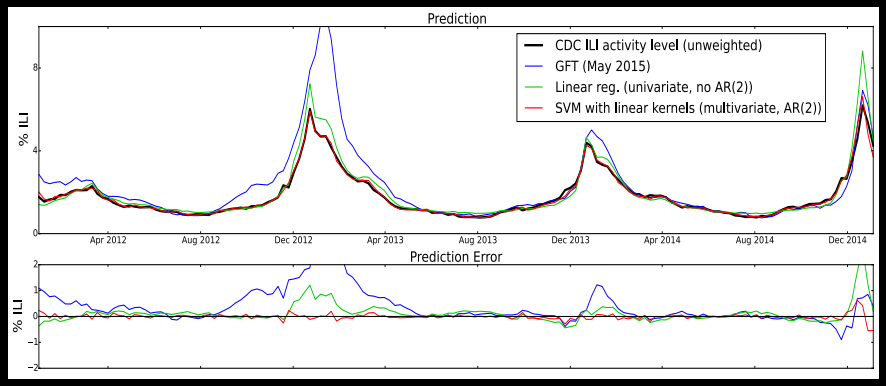
Beyond Google searches…



What are doctors searching for?

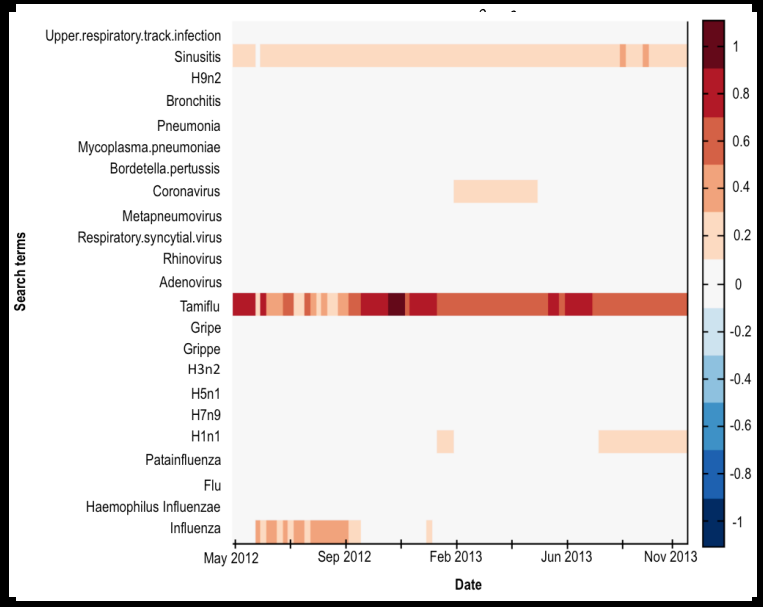


What are people tweeting? What are they reporting on crowd-sourced disease surveillance apps?



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

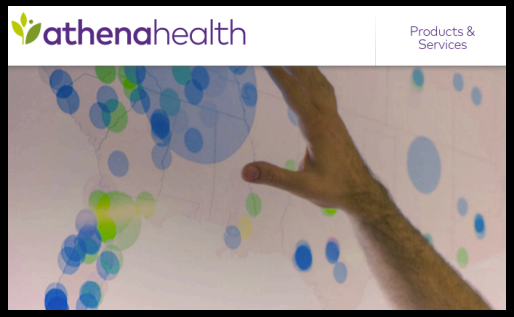
Beyond Google searches…



What are doctors searching for?

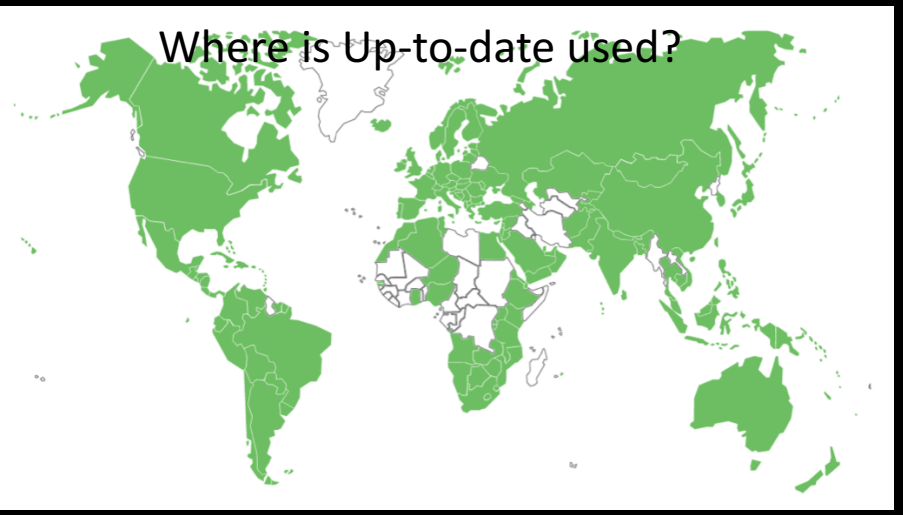
flu near you
do you have it in you?

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?

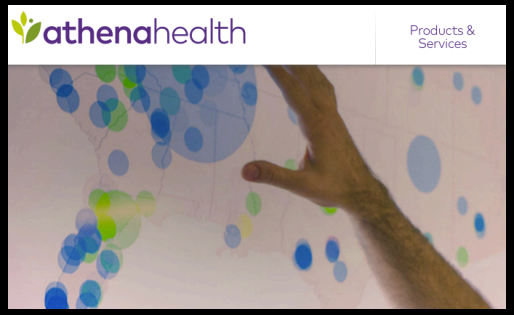
Beyond Google searches…



What are doctors searching for?

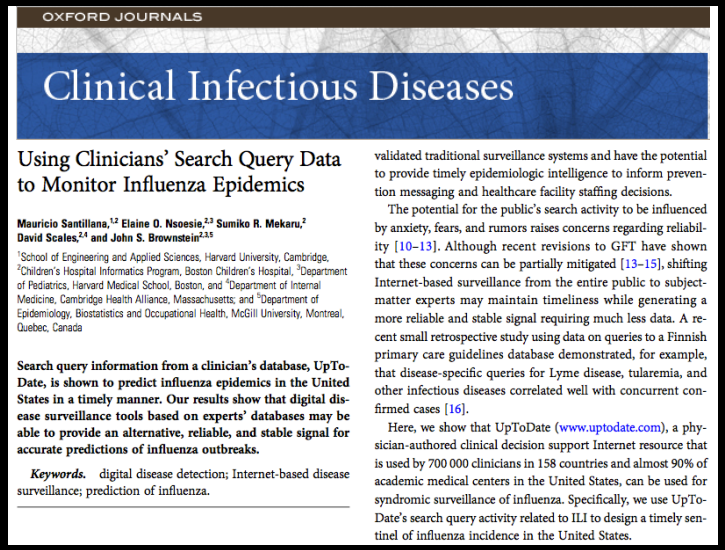
flu near you
do you have it in you?

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[2,5]

[1]School of Engineering and Applied Sciences, Harvard University, Cambridge, [2]Children's Hospital Informatics Program, Boston Children's Hospital, [3]Department of Pediatrics, Harvard Medical School, Boston, and [4]Department of Internal Medicine, Cambridge Health Alliance, Massachusetts; and [5]Department 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.

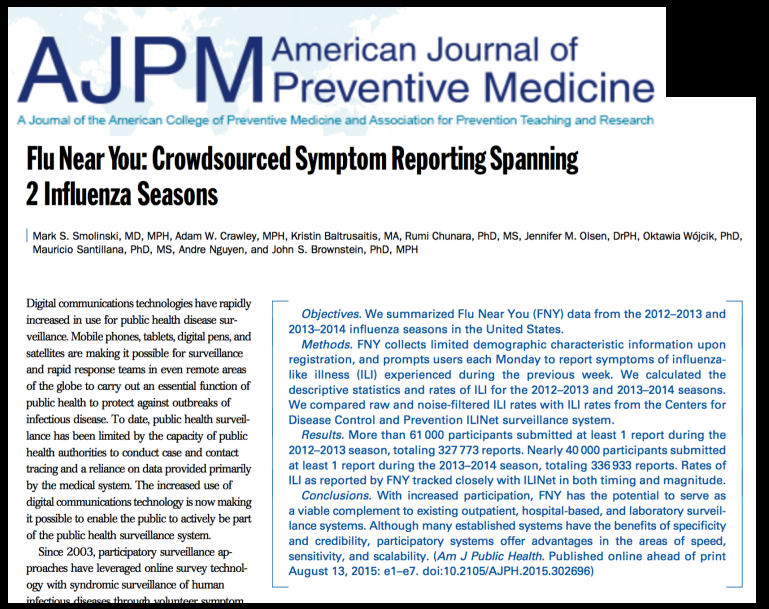
Keywords. digital disease detection; Internet-based disease surveillance; prediction of influenza.

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, Oktawia Wójcik, 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

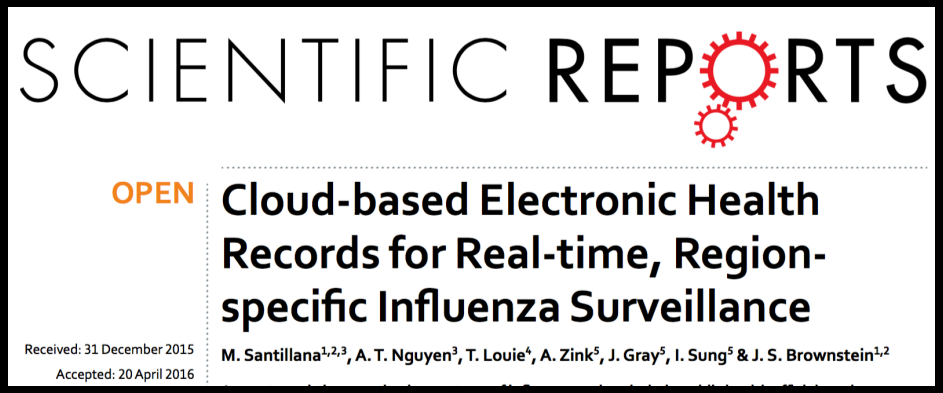
Objectives. We summarized Flu Near You (FNY) data from the 2012–2013 and 2013–2014 influenza seasons in the United States.

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

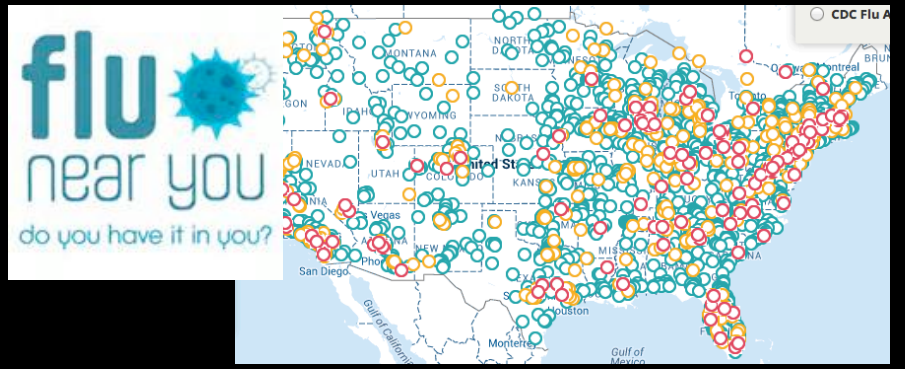
Cloud-based Electronic Health Records for Real-time, Region-specific Influenza Surveillance

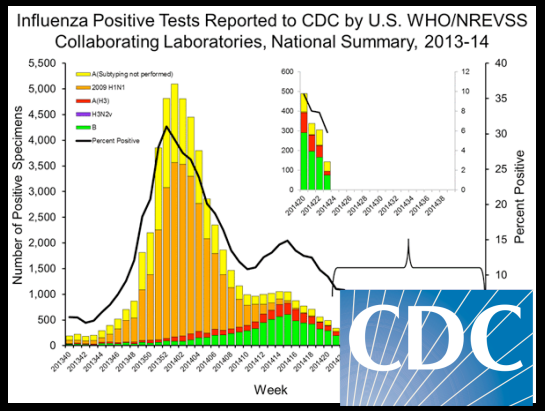
M. Santillana[1,2,3], A. T. Nguyen[3], T. Louie[4], A. Zink[5], J. Gray[5], I. Sung[5] & J. S. Brownstein[1,2]

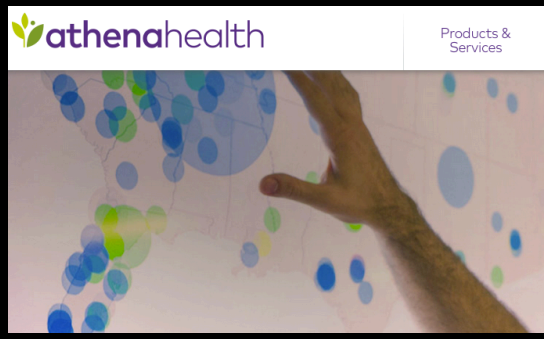
Received: 31 December 2015
Accepted: 20 April 2016

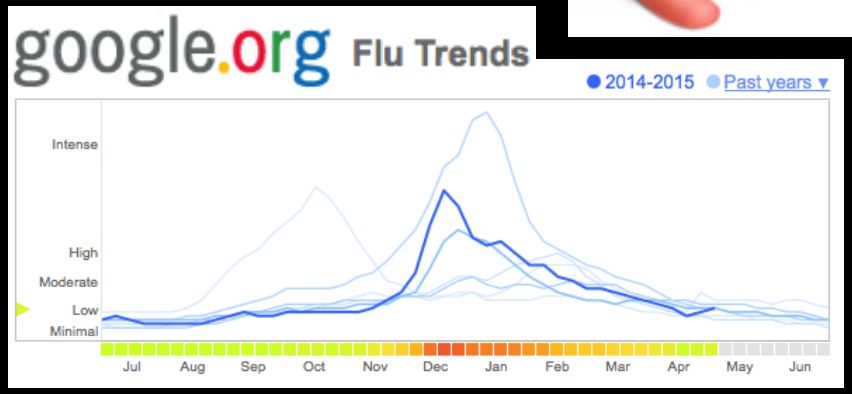
Can we use Electronic Health Records (EHR) to track disease incidence? What lab tests or medications are doctors prescribing?

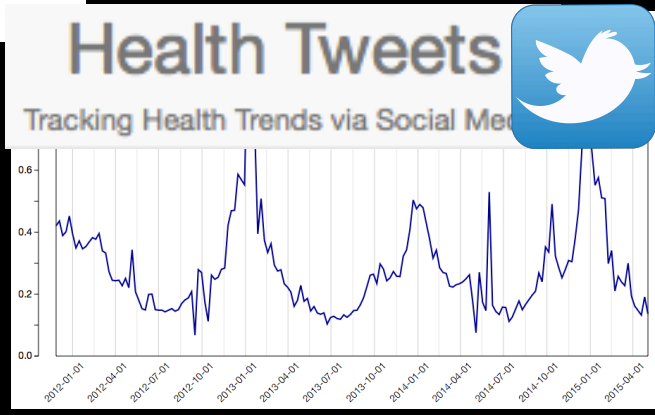
Ensemble approaches yield more accurate and more robust real-time and forecast flu estimates











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

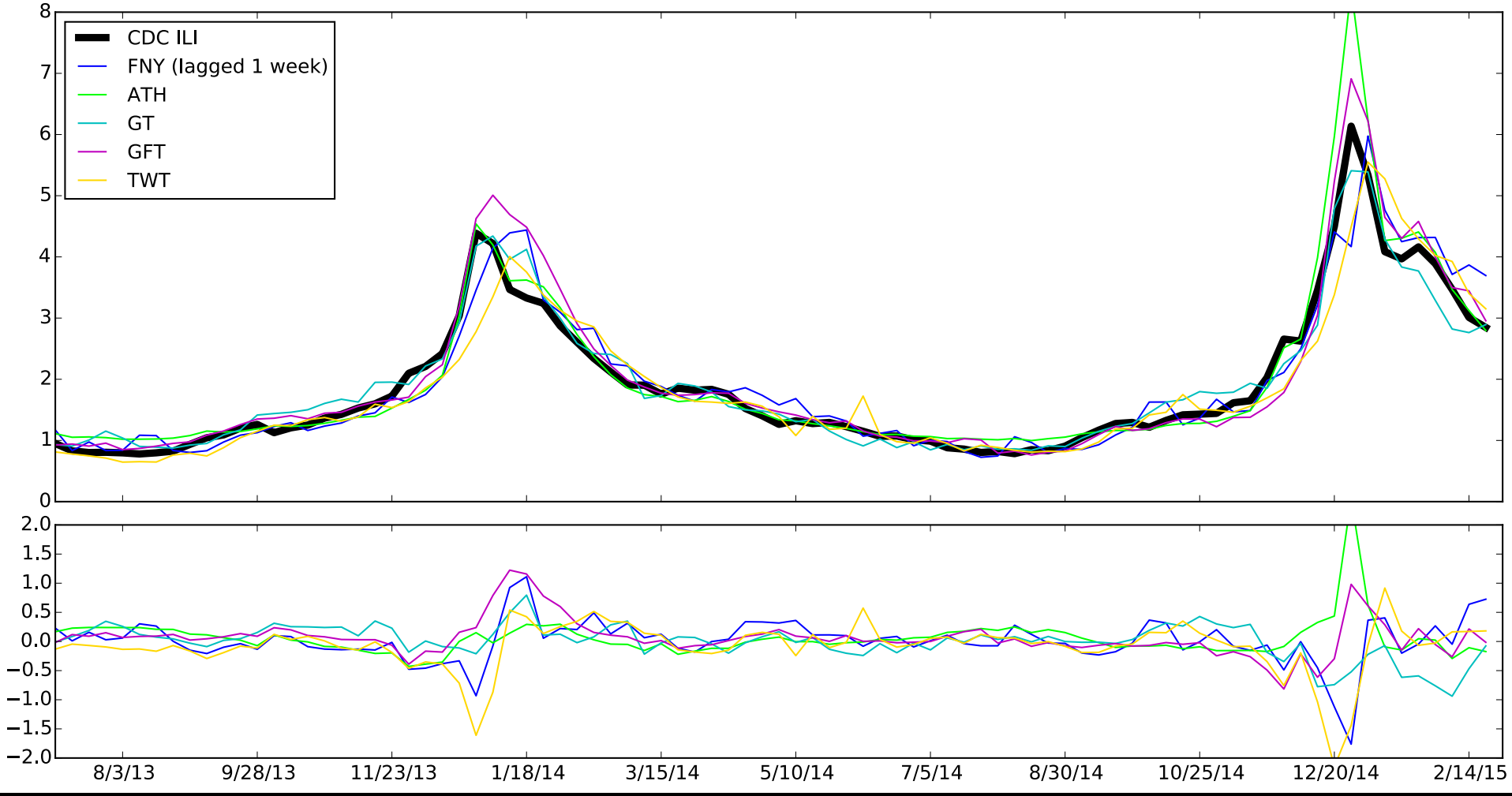
Santillana et al. PLoS Computational Biology, 2015

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

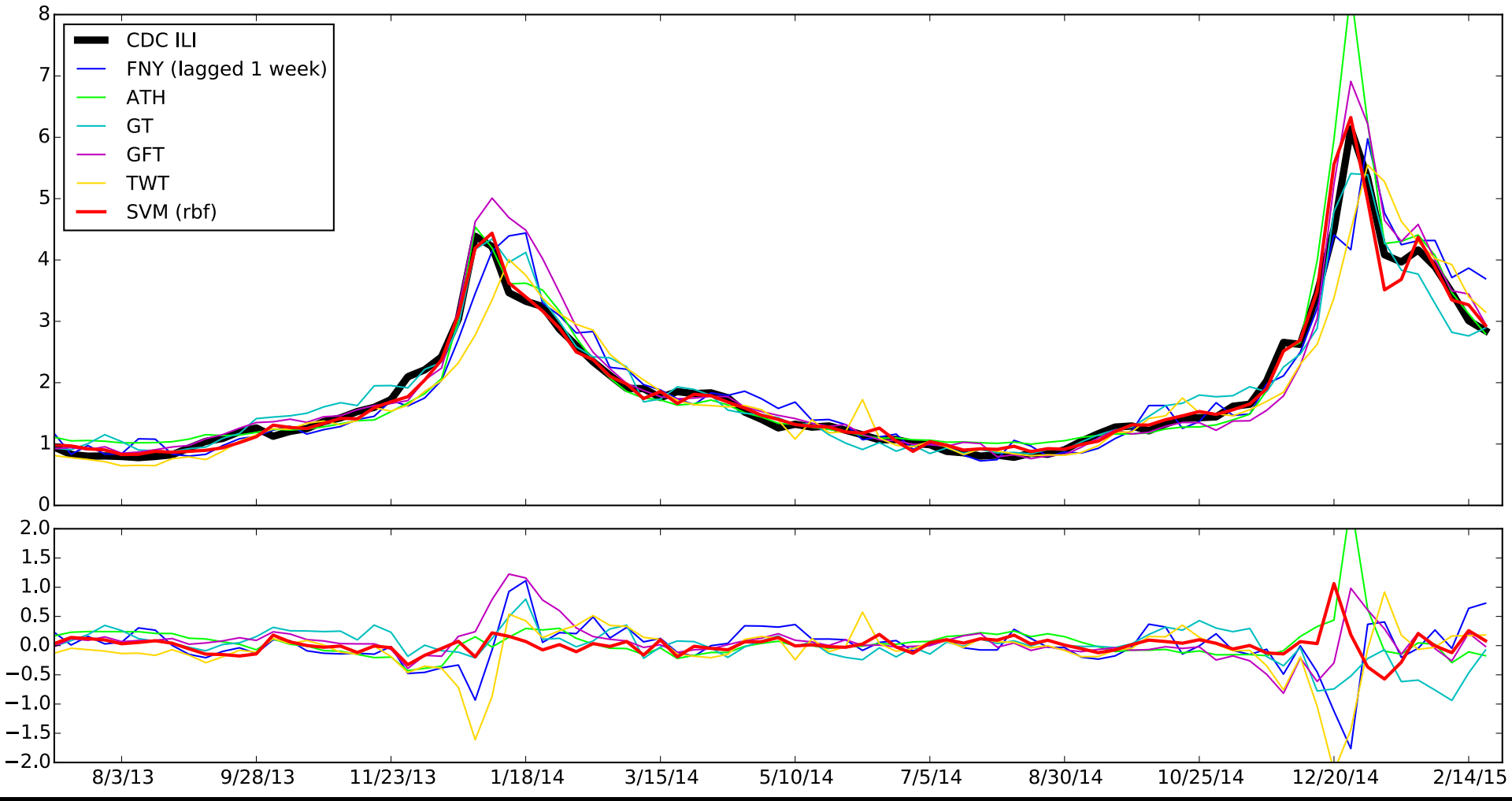
Santillana et al. PLoS Computational Biology, 2015

Performance of individual data sources



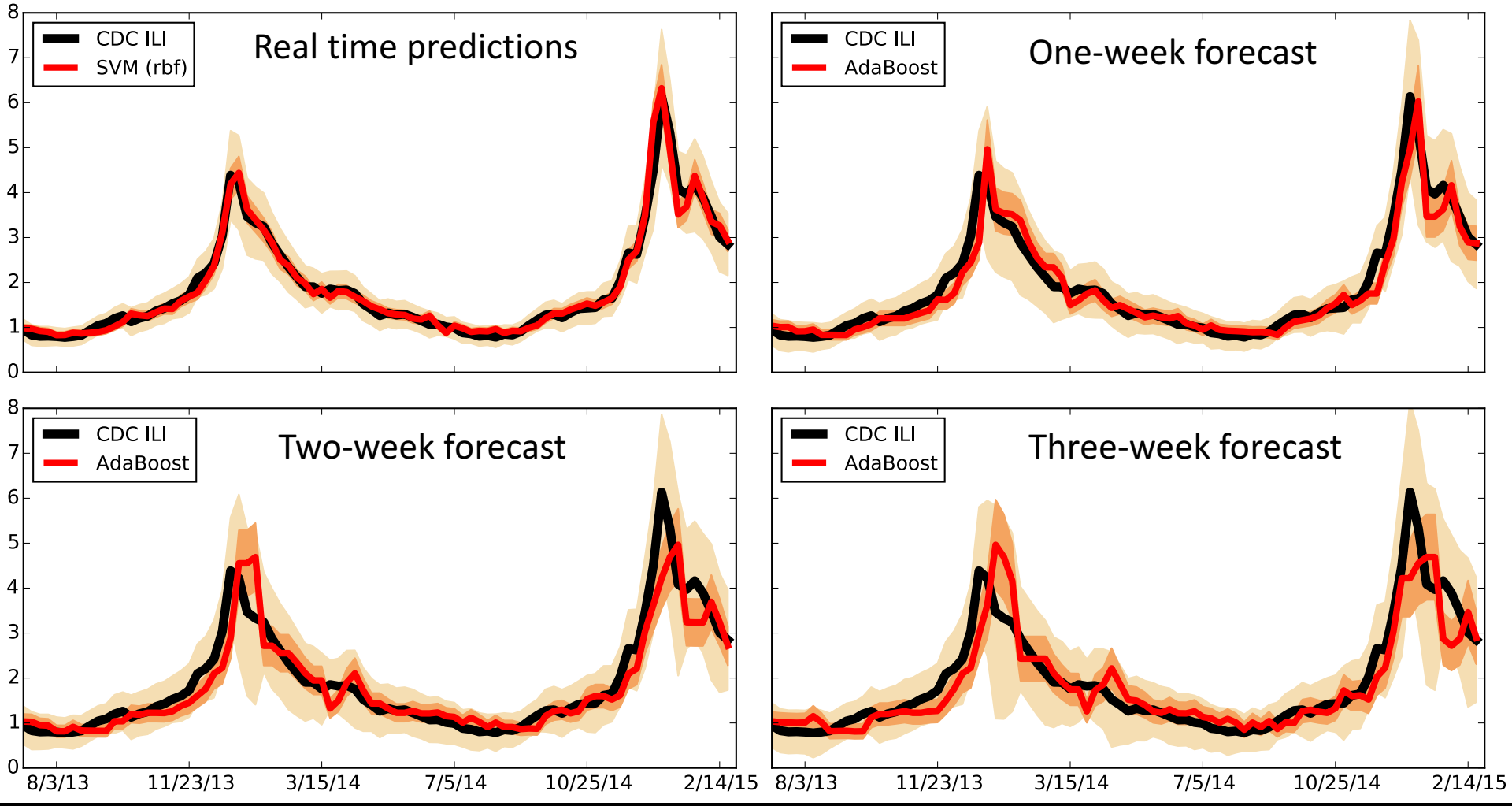
Santillana et al. PLoS Computational Biology, 2015

Performance ensemble



Santillana et al. PLoS Computational Biology, 2015

Real time predictions and Forecasts



Santillana et al. PLoS Computational Biology, 2015

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 Yang[1], Mauricio Santillana[2,3*], John S. Brownstein[2,3], Josh Gray[4], Stewart Richardson[4] 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.

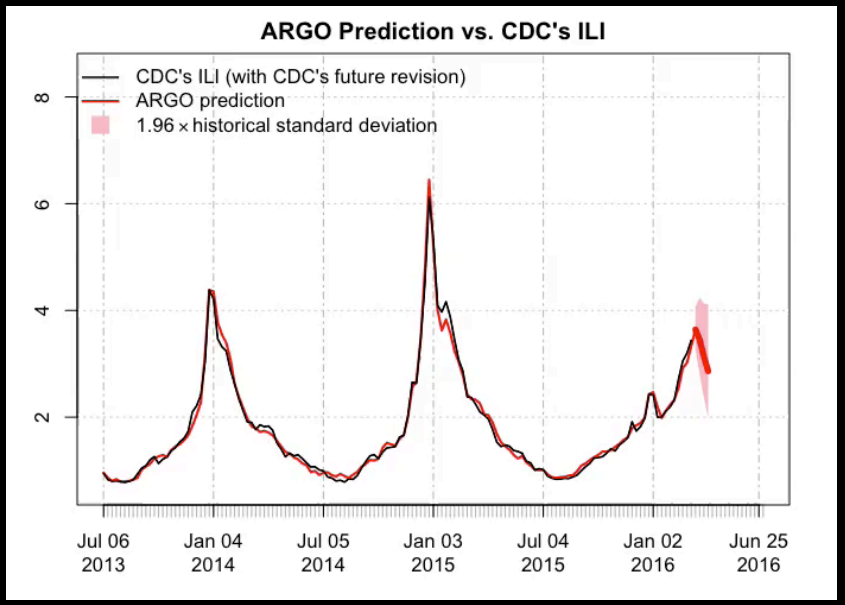
Keywords: Influenza-like illnesses reports, Digital disease detection, Dynamic error reduction, Validation test, Autoregression

PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

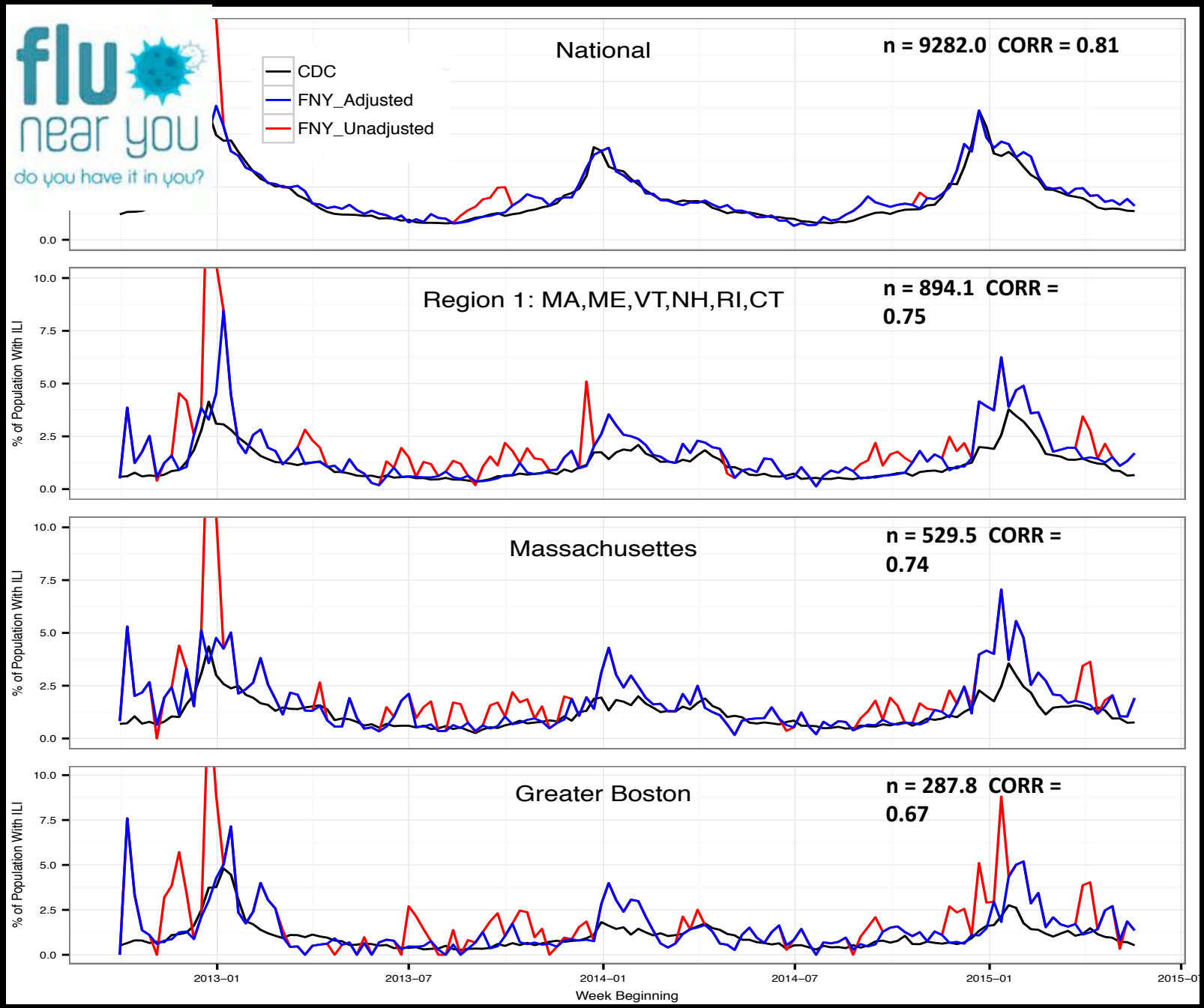
Combining Search, Social Media, and Traditional Data Sources to Improve Influenza Surveillance

Mauricio Santillana[1,2,3]*, André T. Nguyen[1], Mark Dredze[4], Michael J. Paul[5], Elaine O. Nsoesie[6,7], John S. Brownstein[2,3]

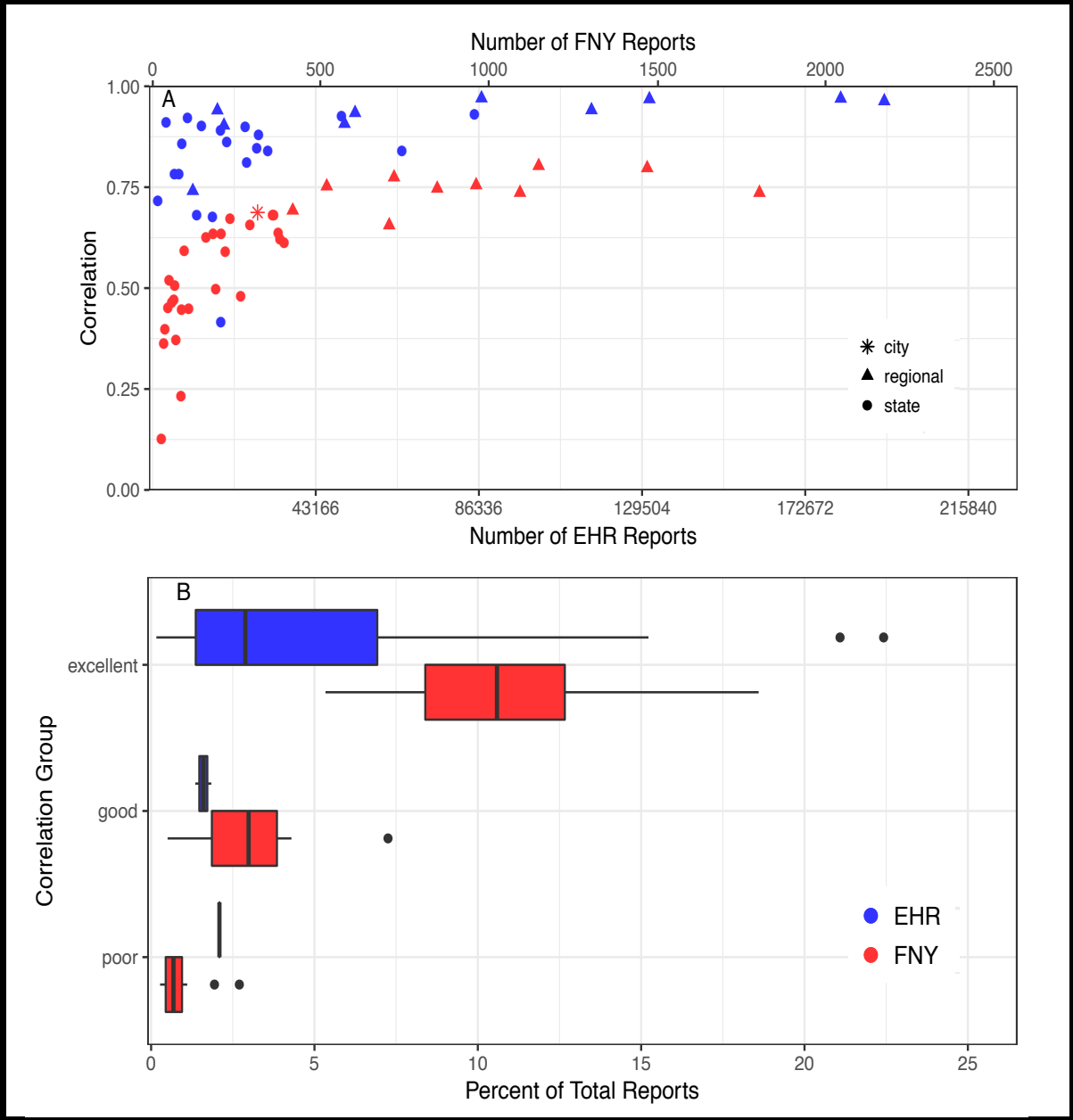


What about estimating flu in a regional level, state-level, city-level?

Correlation of FNY with CDC. Multiple Geographic Scales

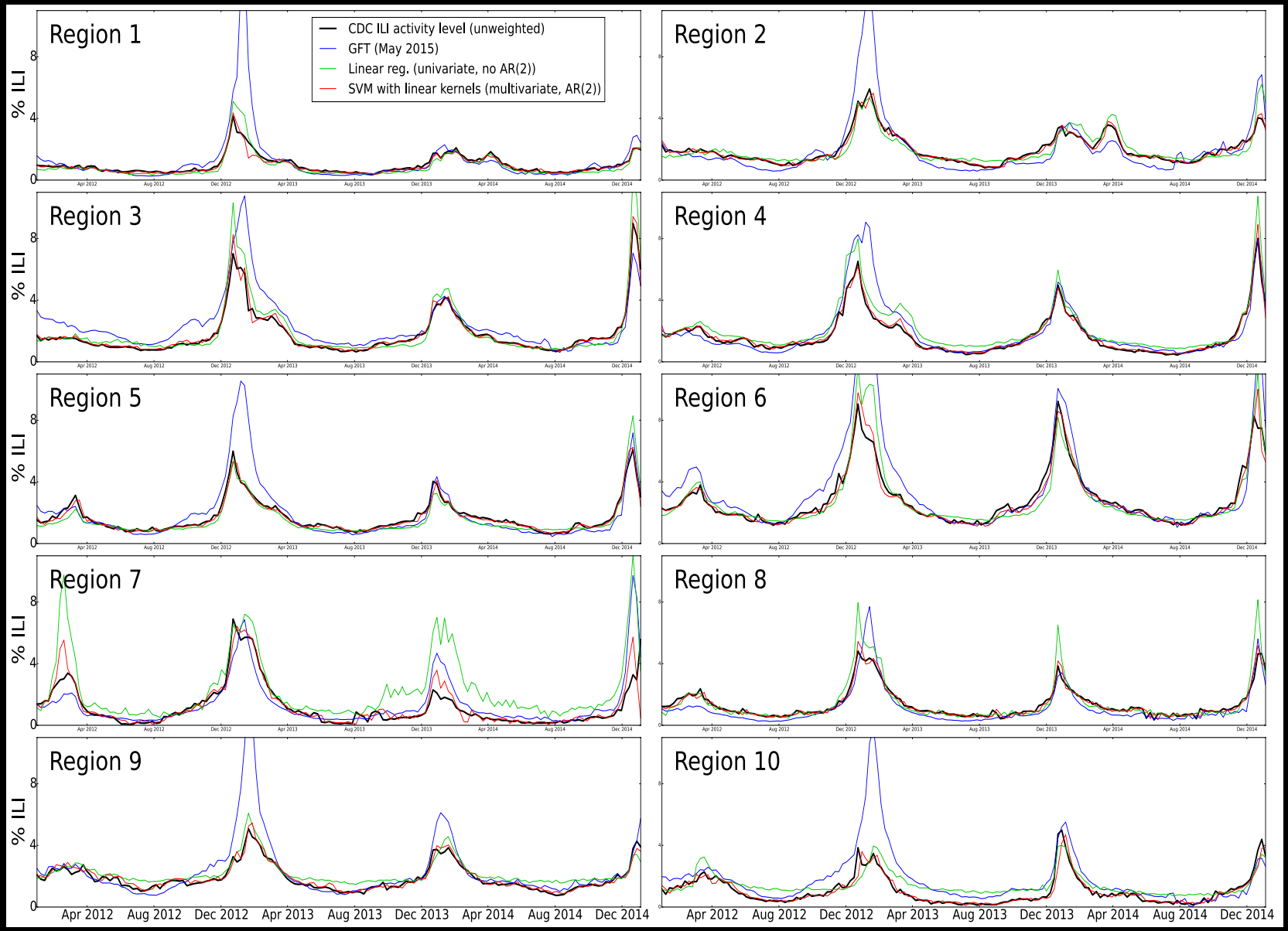


Comparisons between CDC ILINet and FNY and Athenahealth in multiple spatial scales



Baltrusaitis K, Brownstein JS, Scarpino SV, Bakota E, Crawley A, Biggerstaff M, Conidi G, Gunn J, Gray J, Zink A, ..., **Santillana M**. *Comparison of crowd-sourced, electronic health records based, and traditional health-care based influenza-tracking systems at multiple spatial resolutions in the United States of America*. In Press. BMC Infectious Diseases, 2018

athenahealth



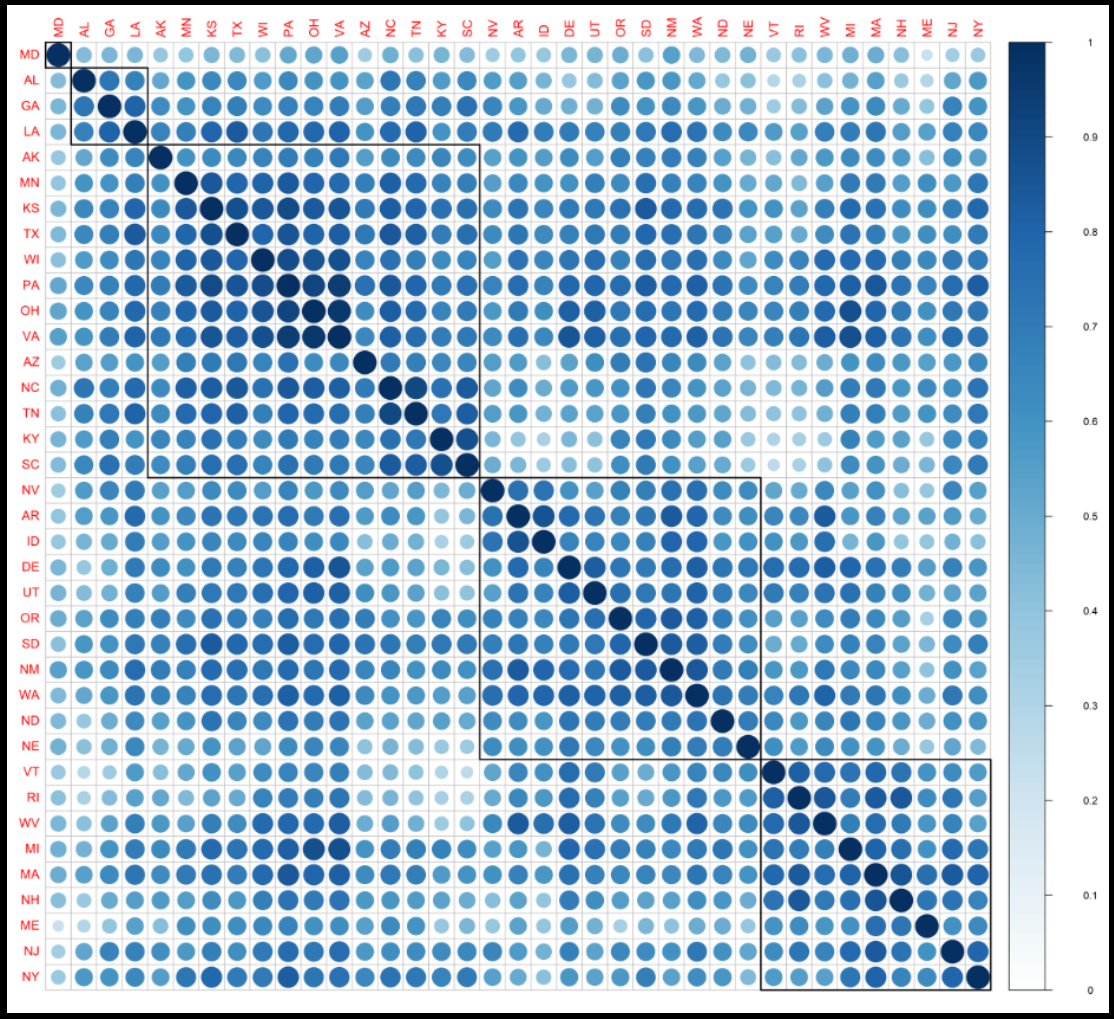
Santillana et al. 2016 Scientific Reports, 25732

How about state-level?

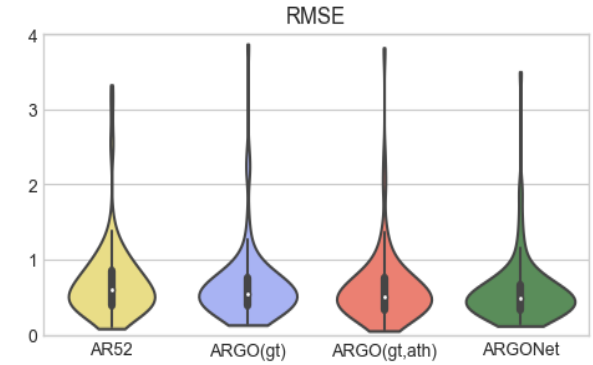
Spatial-temporal synchronicities

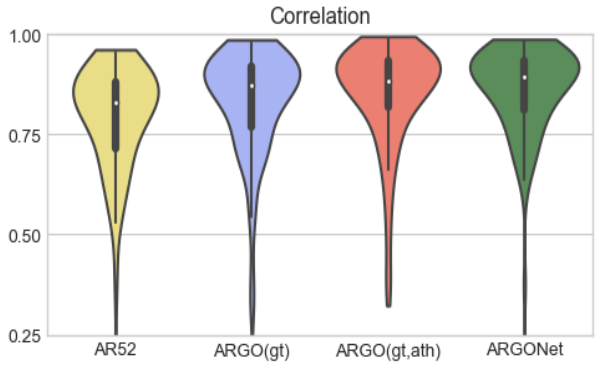
Flu-related Google search information

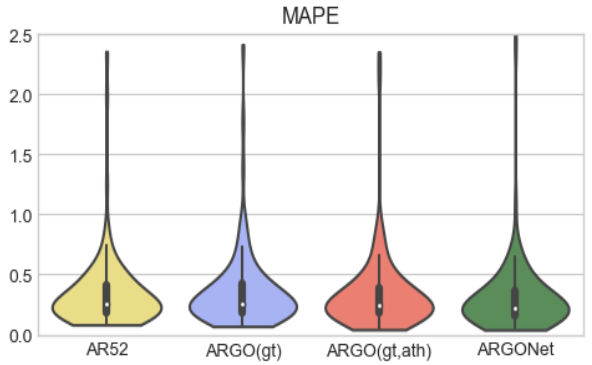
Heat map of pairwise %ILI **correlations** between all states.
Boxes denote clusters of highly correlated states.

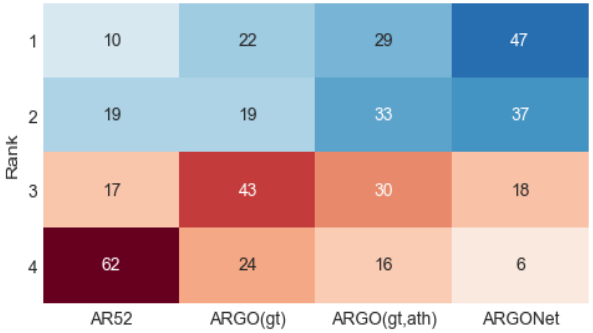


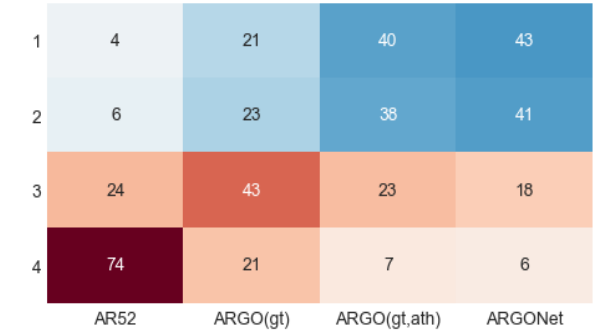
Clear improvements over previous methodologies

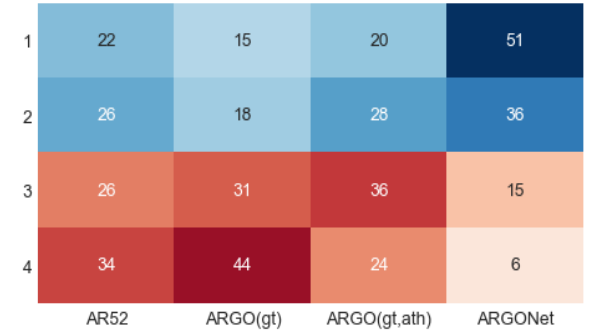


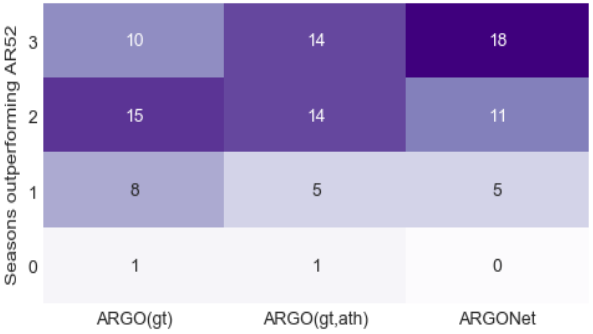


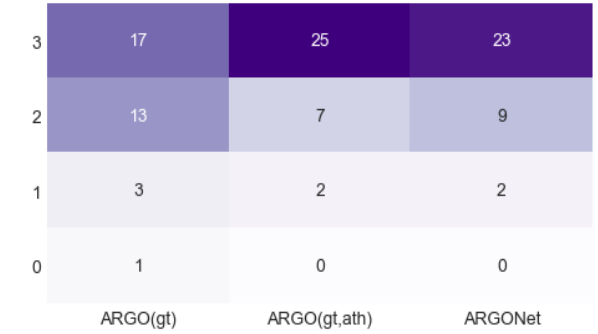


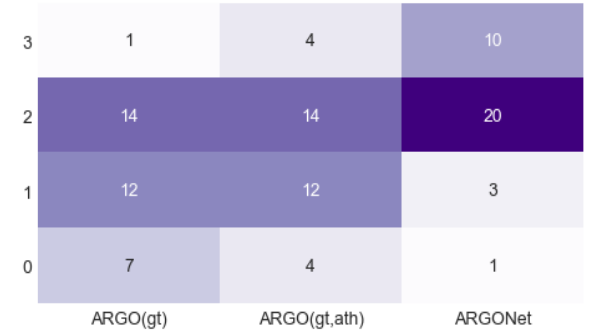












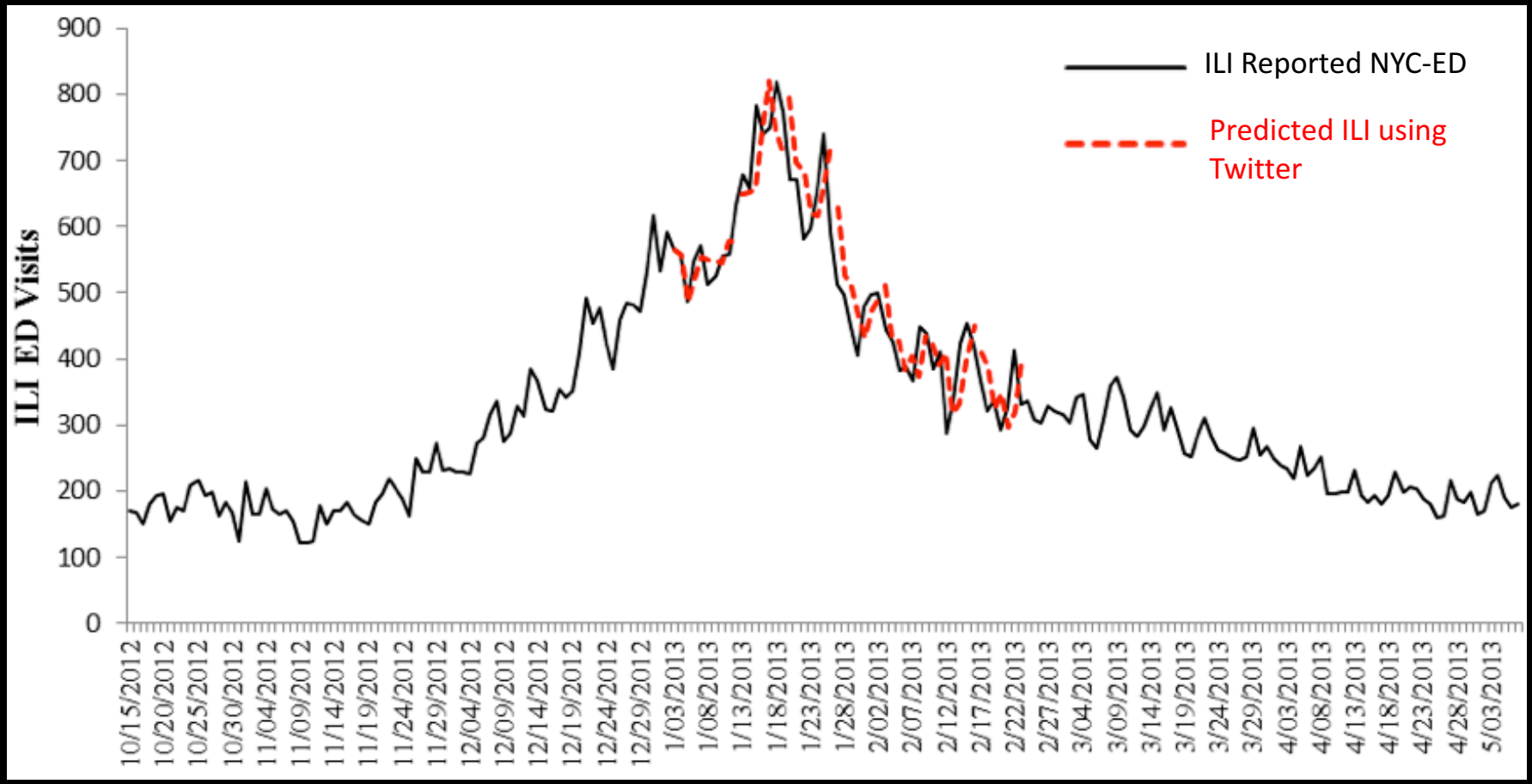
How about city-level?

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

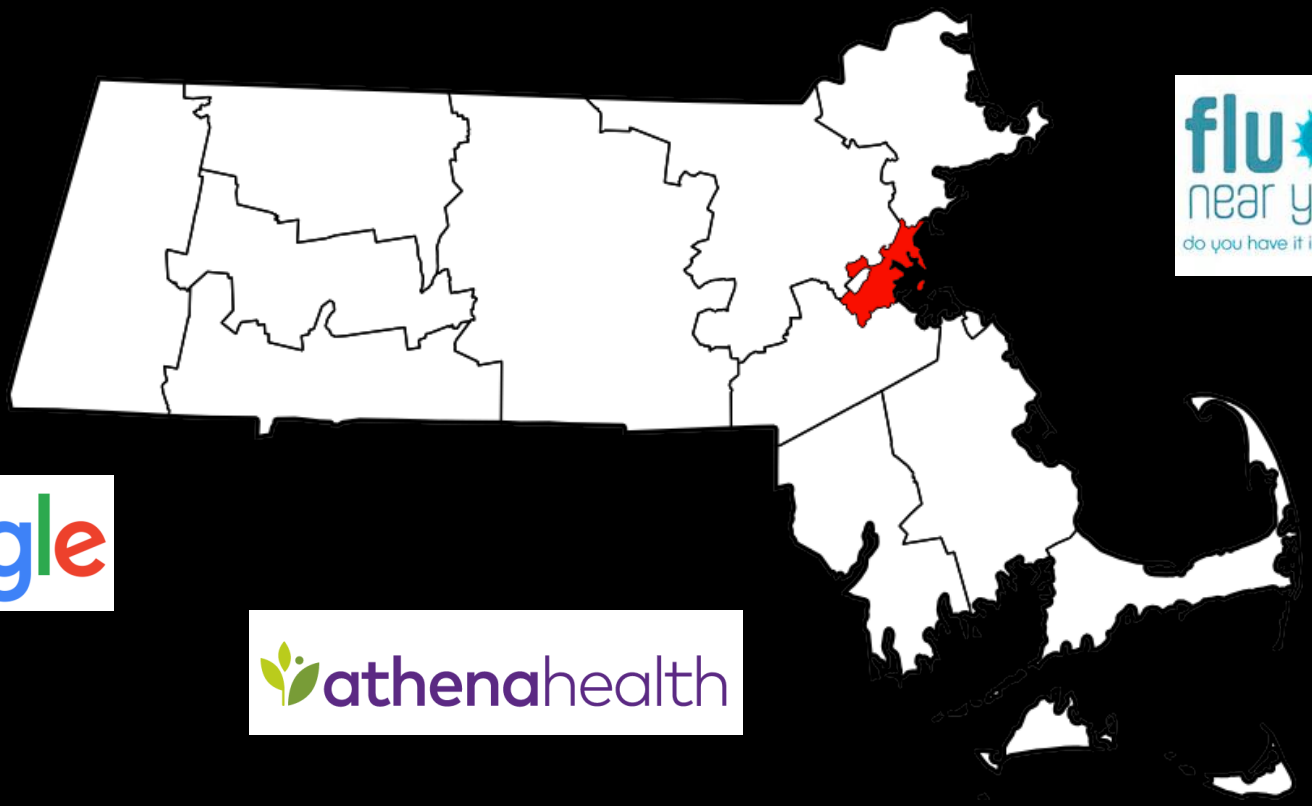
Daily ILI visits (as reported by the NYC emergency department) compared to predicted ILI using twitter data



Nagar et al. (2014) Journal of Medical Internet Research. In press

We will extend out methodology to finer spatial resolutions.
(Massachusetts and Boston)

Highlights: (a) dynamic-moving training window, (b) automatic feature selection, (c) ensemble approach



Lu F, Hou S, Baltrusaitis K, Shah M, Leskovec J, Sosic R, Hawkins J, Brownstein JS, Conidi G, Gunn J, ..., Santillana M. Accurate influenza monitoring and forecasting in the Boston metropolis using novel Internet data streams. Journal of Medical Internet Research. 2018;4 (1) :e4.7

Published on 09.01.18 in Vol 4, No 1 (2018): Jan-Mar

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

Infoveillance, Infodemiology and Digital Disease Surveillance Infodemiology and Infoveillance

Original Paper

Accurate Influenza Monitoring and Forecasting Using Novel Internet Data Streams: A Case Study in the Boston Metropolis

Fred Sun Lu[1], AB ; Suqin Hou[2], MS ; Kristin Baltrusaitis[3], MS ; Manan Shah[4] ; Jure Leskovec[4,5], PhD ; Rok Sosic[4], PhD ; Jared Hawkins[1,6], MMSc, PhD ; John Brownstein[1,6], PhD ; Giuseppe Conidi[7], MPH ; Julia Gunn[7], RN, MPH ; Josh Gray[8], MBA ; Anna Zink[8], BA ; Mauricio Santillana[1,6], MS, PhD

[1]Computational Health Informatics Program, Boston Children's Hospital, Boston, MA, United States

[2]Harvard Chan School of Public Health, Harvard University, Boston, MA, United States

[3]Department of Biostatistics, Boston University School of Public Health, Boston, MA, United States

[4]Computer Science Department, Stanford University, Stanford, CA, United States

[5]Chan Zuckerberg Biohub, San Francisco, CA, United States

[6]Department of Pediatrics, Harvard Medical School, Boston, MA, United States

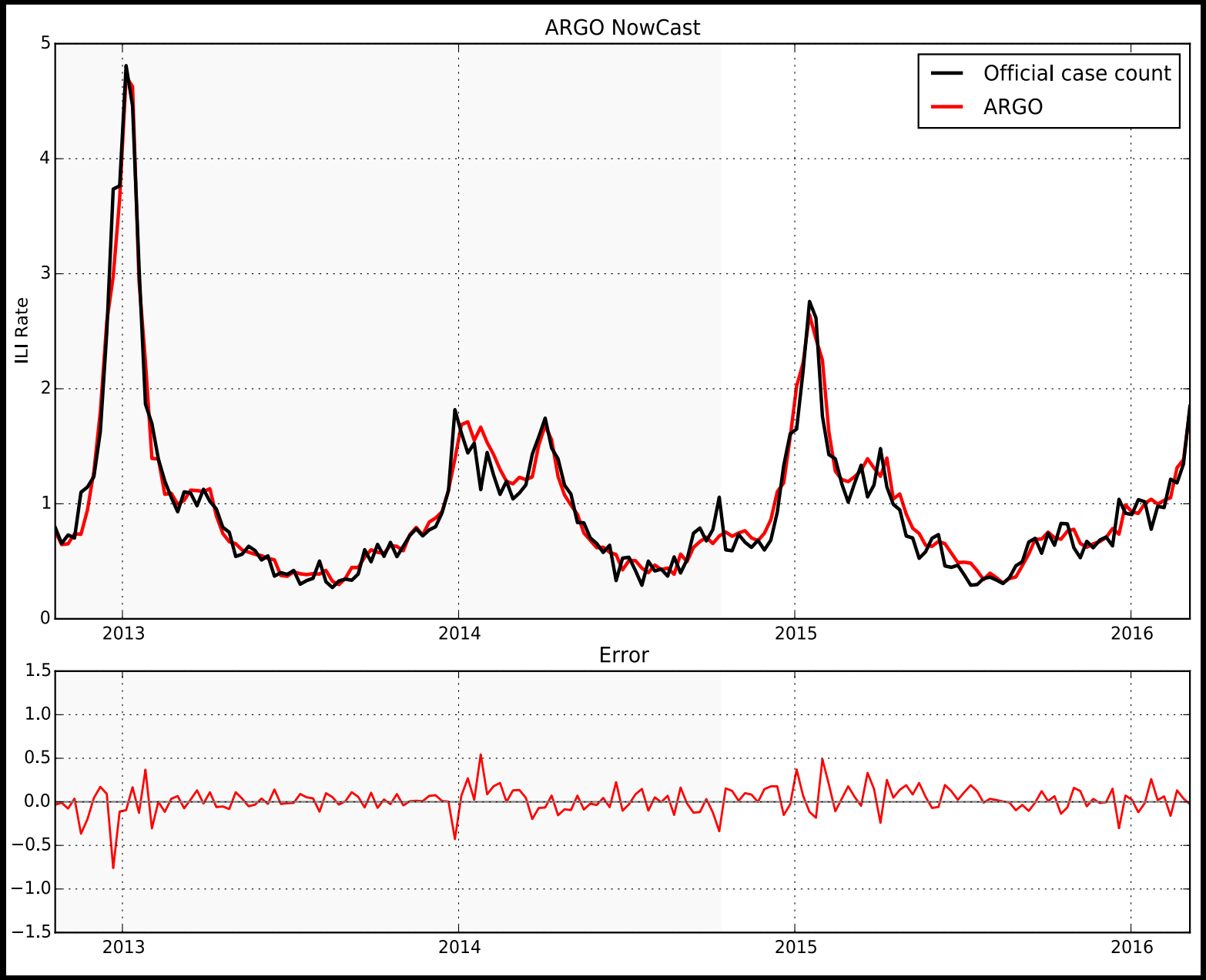
[7]Boston Public Health Commission, Boston, MA, United States

[8]athenaResearch, athenahealth, Watertown, MA, United States

Corresponding Author:

Mauricio Santillana, MS, PhD

Using multiple data sources to track flu in **Boston**



Plots produced by Fred Lu and Suqin Hou

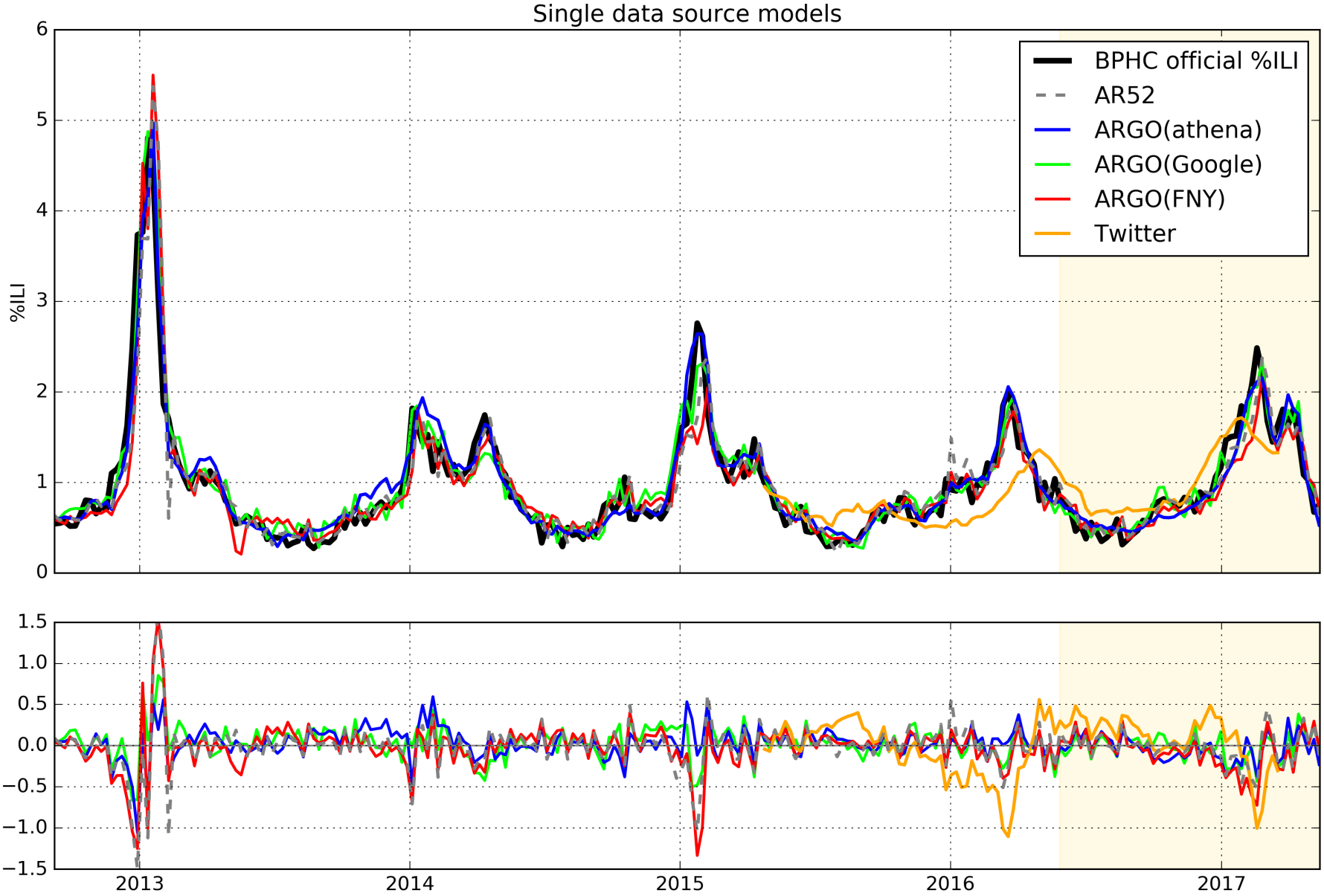
Using multiple data sources to **forecast** flu in **Boston**

Plots produced by Fred Lu and Suqin Hou

Boston – out of sample predictions one-week ahead

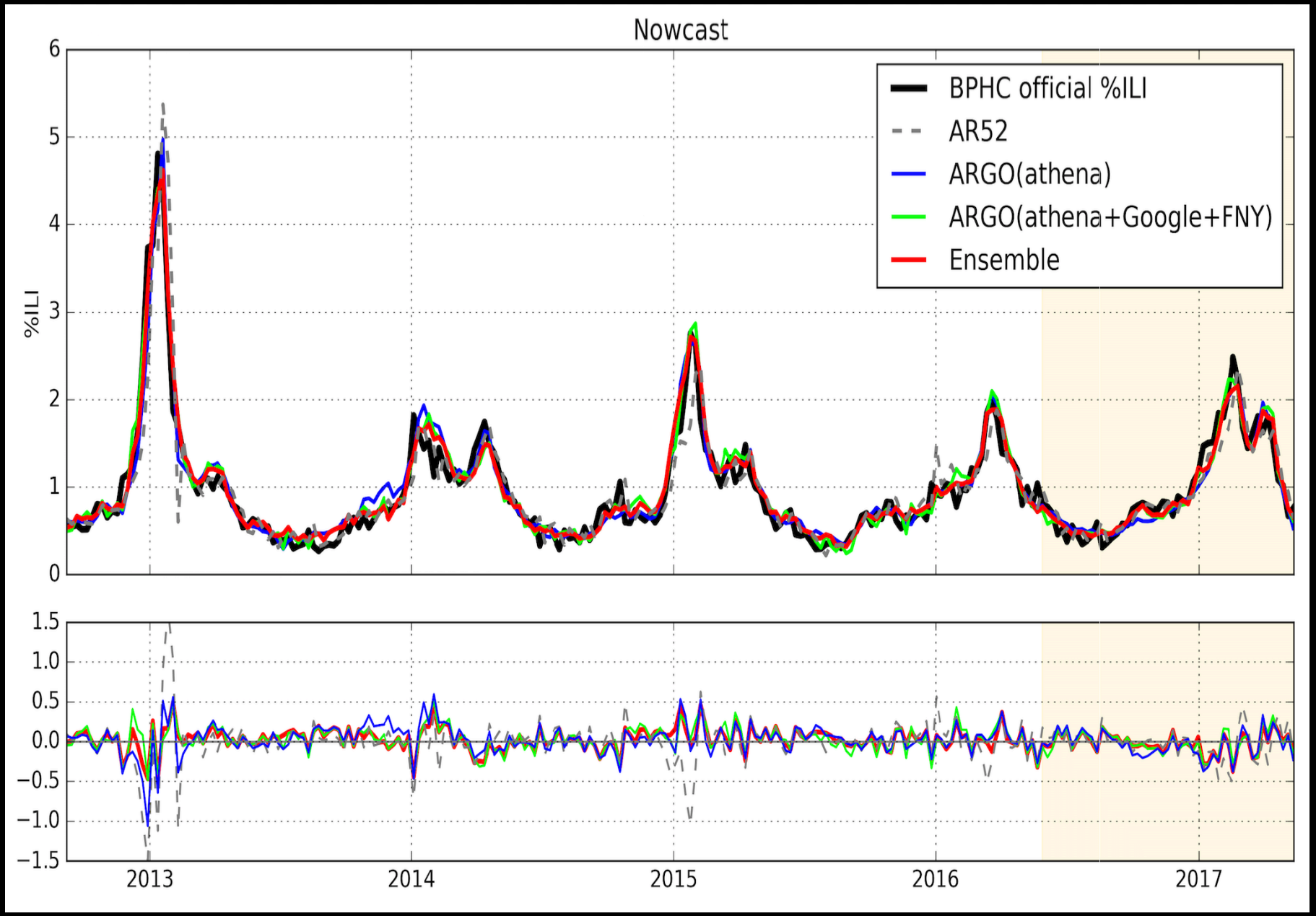
Model performance (individual data sources)

Model[a]	Whole period	Flu seasons				Holdout
	2012-16	2012-13	2013-14	2014-15	2015-16	2016-17
Root mean square error						
AR52	0.303	0.577	0.199	0.305	0.217	0.229
ARGO(athena)[b]	*0.195*	*0.306*	0.229	*0.192*	*0.133*	*0.182*
ARGO(Google)	0.206	0.312	*0.194*	0.247	0.161	0.188
ARGO(FNY)[c]	0.299	0.552	0.204	0.343	0.172	0.280
Twitter	–	–	–	0.162	0.427	0.351
GFT[d]	–	0.352	0.271	0.284	–	–
Naive	0.266	0.481	0.208	0.280	0.202	0.219


Single data source models
BPHC official %ILI
AR52
ARGO(athena)
ARGO(Google)
ARGO(FNY)
Twitter
%ILI
6
5
4
3
2
1
0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
2013
2014
2015
2016
2017

Ensemble models

Boston – out of sample predictions one-week ahead



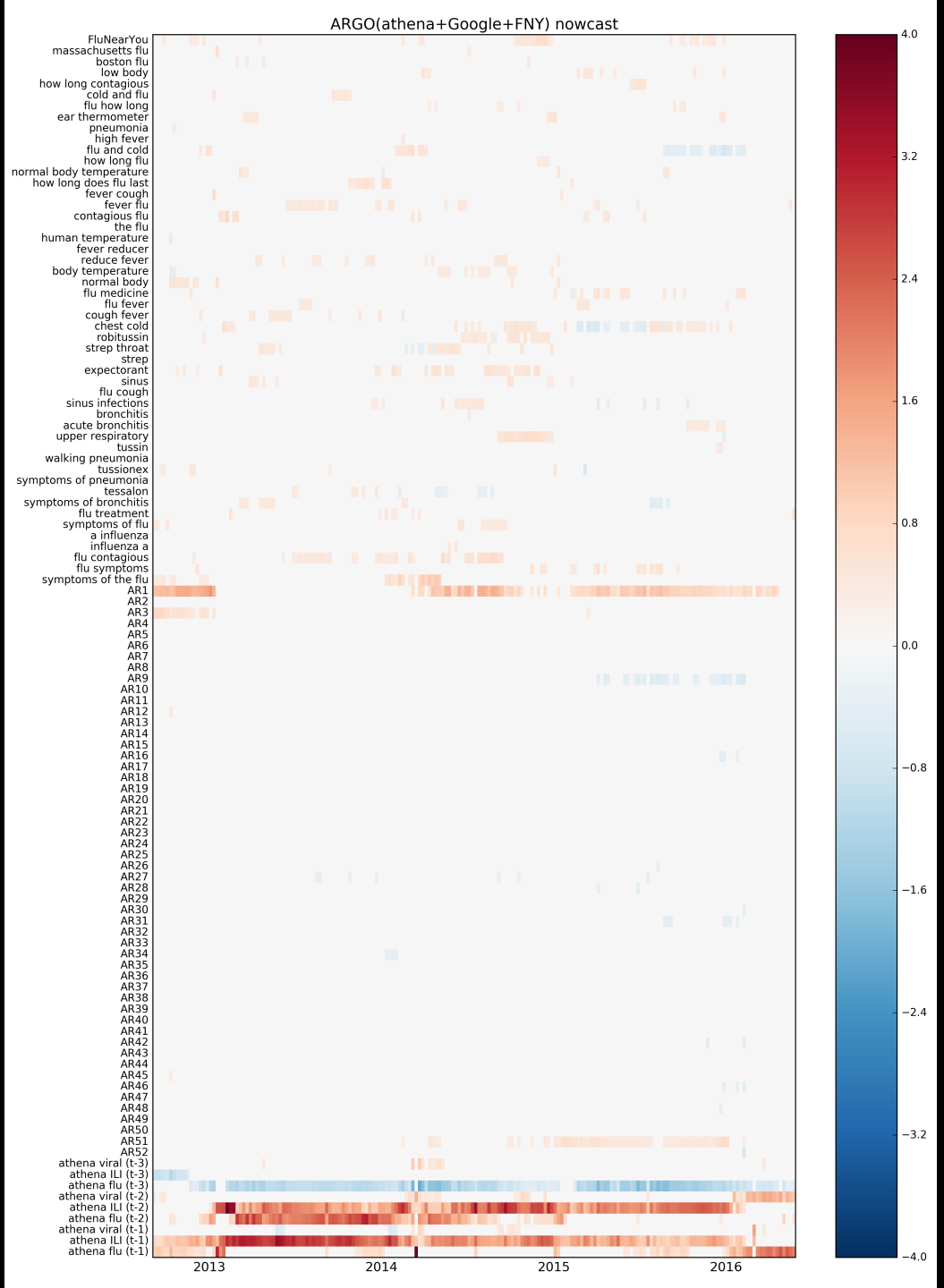
Model performance (individual data sources)

Model[a]	Whole period	Flu seasons				Holdout
	2012-16	2012-13	2013-14	2014-15	2015-16	2016-17
Root mean square error						
AR52	0.303	0.577	0.199	0.305	0.217	0.229
ARGO(athena)[b]	*0.195*	*0.306*	0.229	*0.192*	*0.133*	*0.182*
ARGO(Google)	0.206	0.312	*0.194*	0.247	0.161	0.188
ARGO(FNY)[c]	0.299	0.552	0.204	0.343	0.172	0.280
Twitter	—	—	—	0.162	0.427	0.351
GFT[d]	—	0.352	0.271	0.284	—	—
Naive	0.266	0.481	0.208	0.280	0.202	0.219

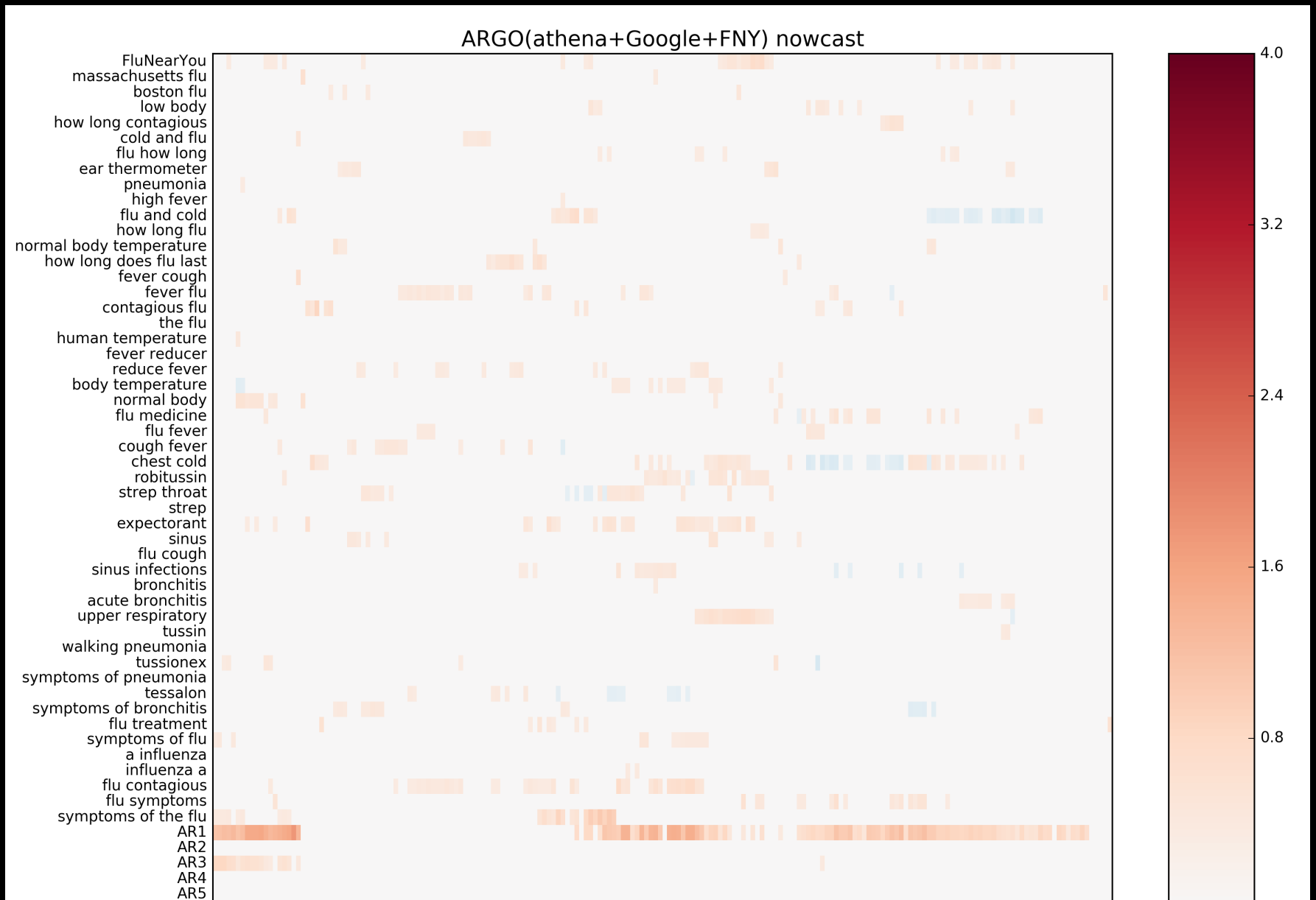
Model performance (ensemble approaches)

		Flu seasons				
	Whole period (2012-2016)	2012-2013	2013-2014	2014-2015	2015-2016	*Validation* 2016-2017
RMSE						
AR52	0.303	0.577	0.199	0.305	0.217	0.229
athena	0.208	0.377	**0.163**	0.219	0.144	0.169
Google	0.222	0.218	0.243	0.303	0.221	0.195
athena+Google	0.186	0.210	0.200	0.224	0.194	**0.145**
ARGO(athena)	0.195	0.306	0.229	0.192	**0.133**	0.182
ARGO(Google)	0.206	0.312	0.194	0.247	0.161	0.188
ARGO(athena+Google+FNY)	**0.165**	**0.199**	0.192	**0.189**	0.168	0.156
Ensemble	**0.151**	**0.193**	**0.170**	**0.176**	**0.139**	**0.150**

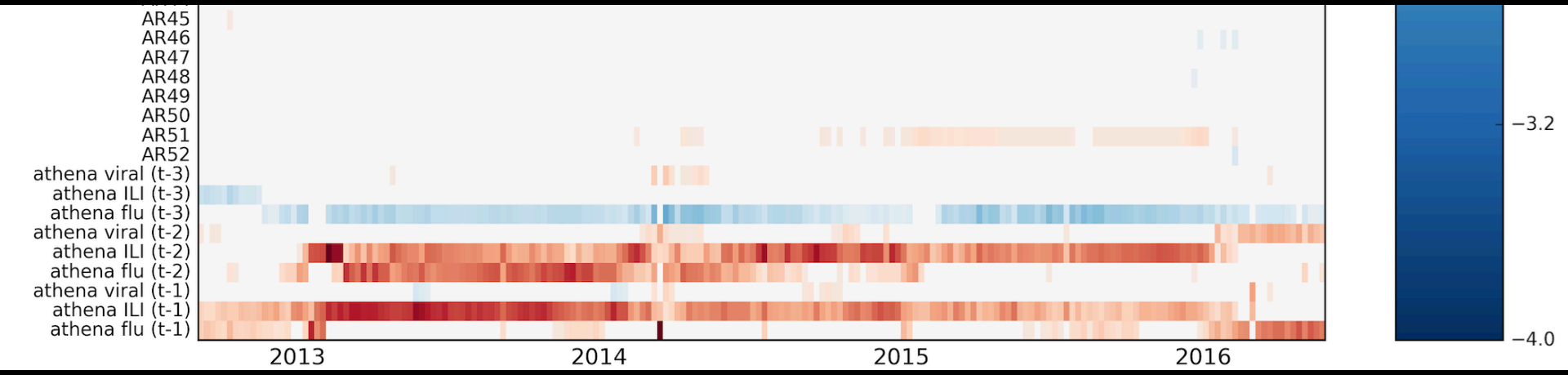
When combined, what are the strongest predictors?

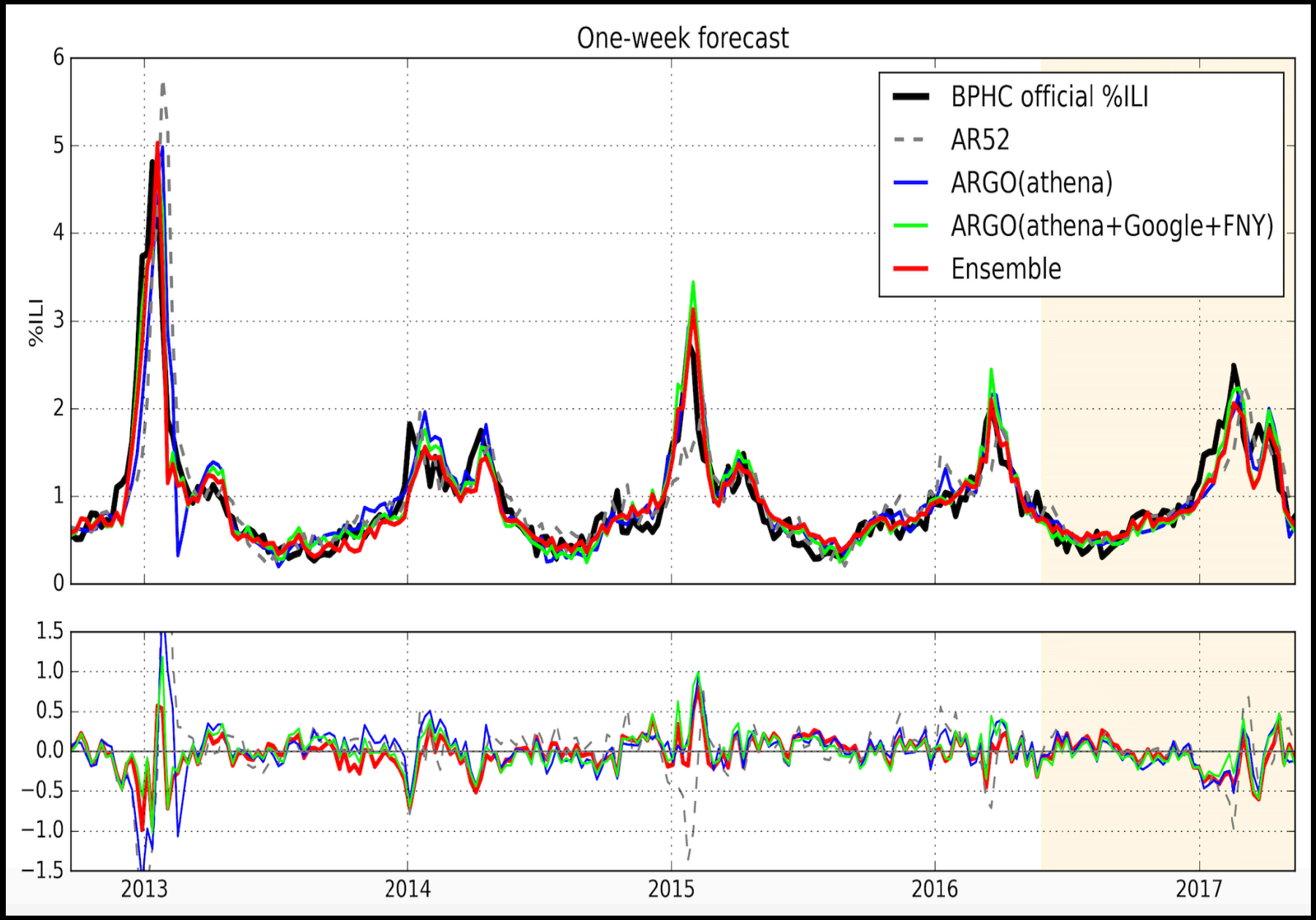


When combined, what are the strongest predictors?



When combined, what are the strongest predictors?




One-week forecast
BPHC official %ILI
AR52
ARGO(athena)
ARGO(athena+Google+FNY)
Ensemble
%ILI
6
5
4
3
2
1
0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
2013
2014
2015
2016
2017

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**

Hyper-local predictions

Can we predict daily emergency department visits in a hospital?

Boston Children's Hospital

Until every child is well℠

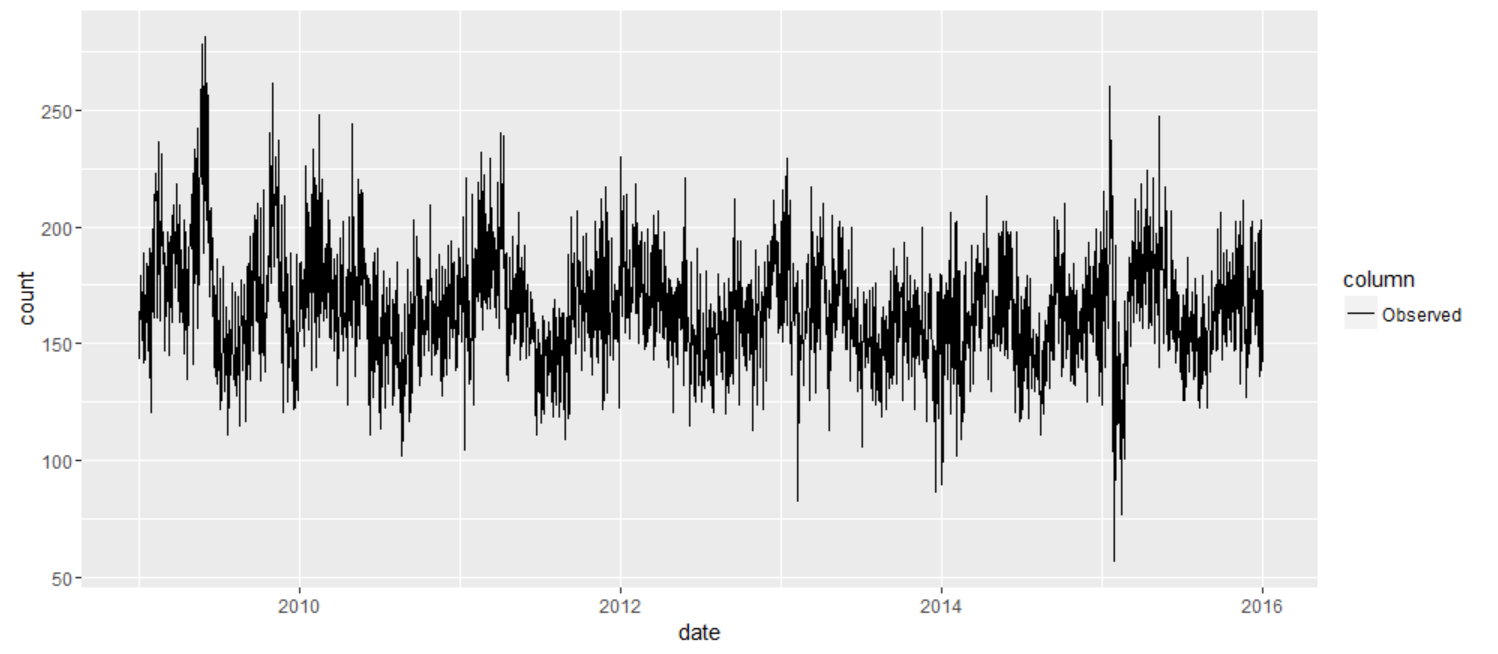
NOAA

NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION

U.S. DEPARTMENT OF COMMERCE

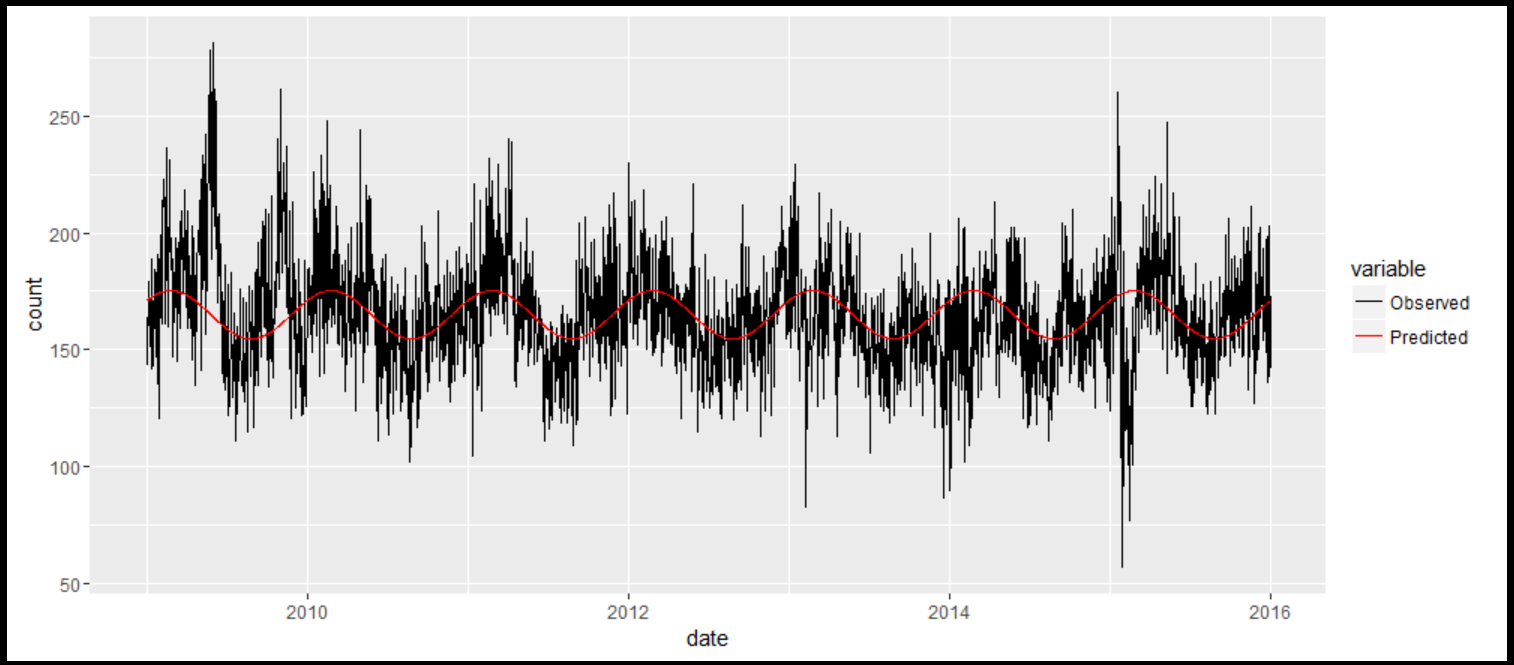
Google

Daily Visits 2009-2015



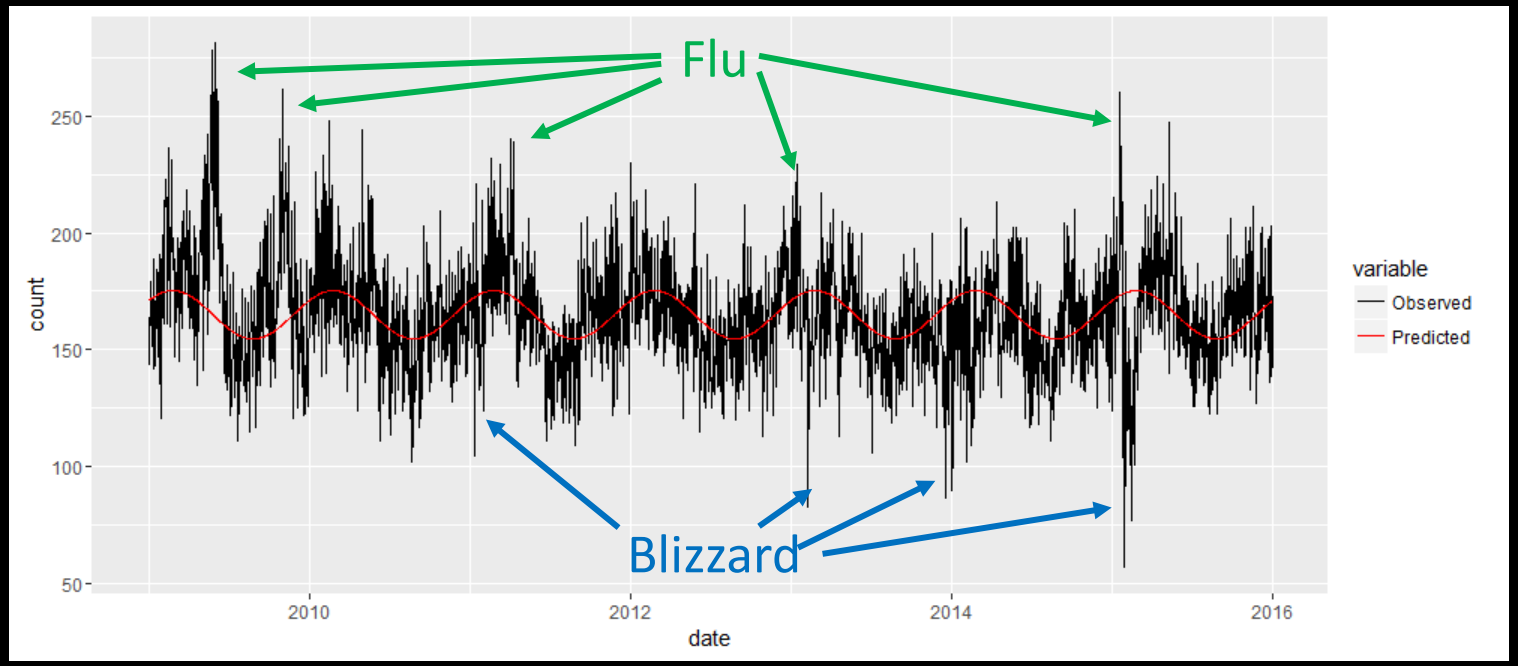
In collaboration with: Sam Tideman, Mauricio Santillana, Jon Bickel, and Ben Reis

Seasonal Trend



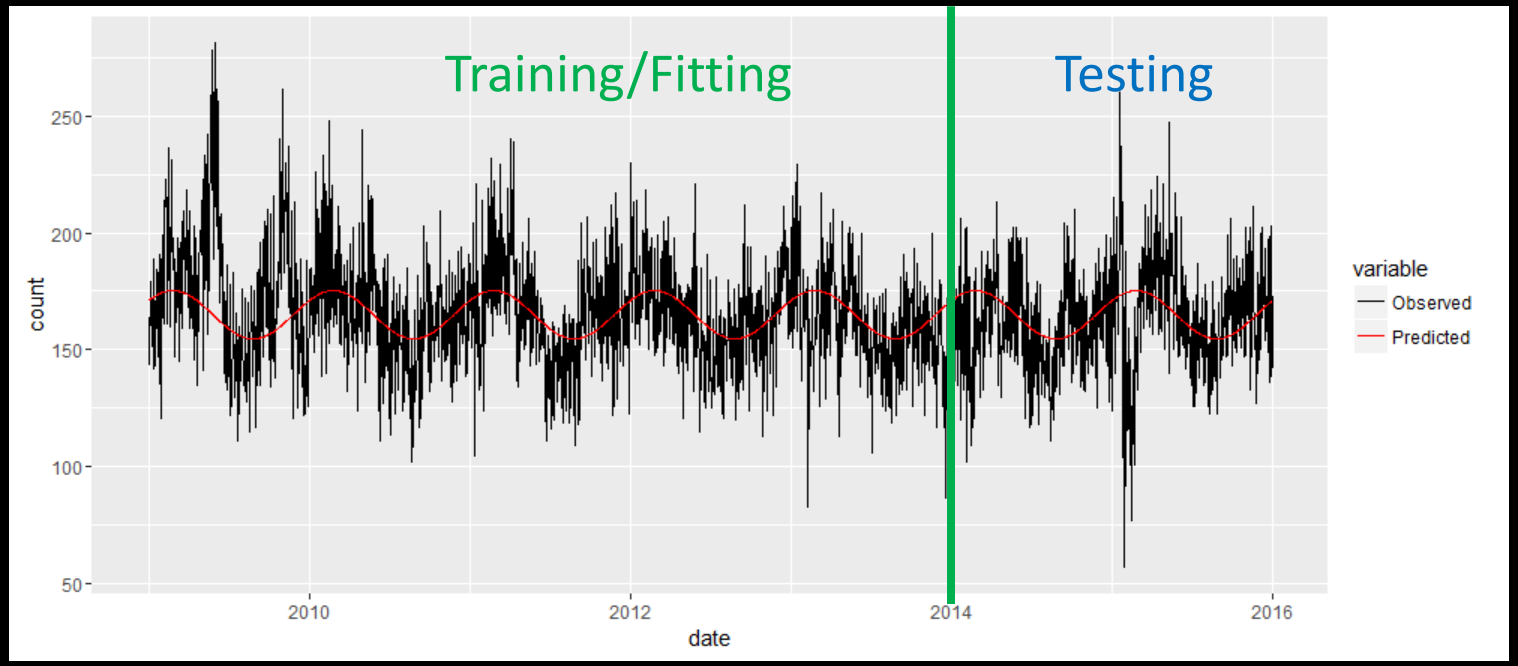
In collaboration with: Sam Tideman, Mauricio Santillana, Jon Bickel, and Ben Reis

Noticeable Events



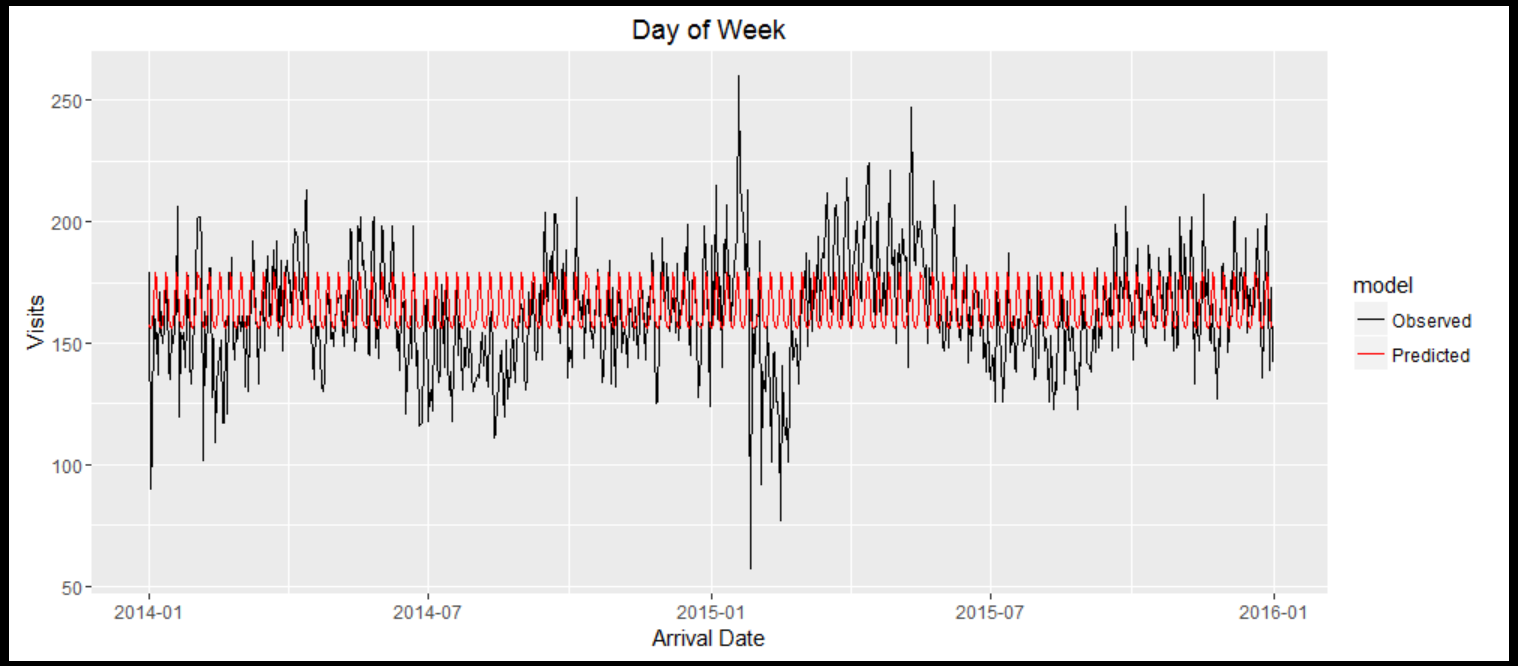
In collaboration with: Sam Tideman, Mauricio Santillana, Jon Bickel, and Ben Reis

Split data for modeling



In collaboration with: Sam Tideman, Mauricio Santillana, Jon Bickel, and Ben Reis

Current Staffing model = Day of Week

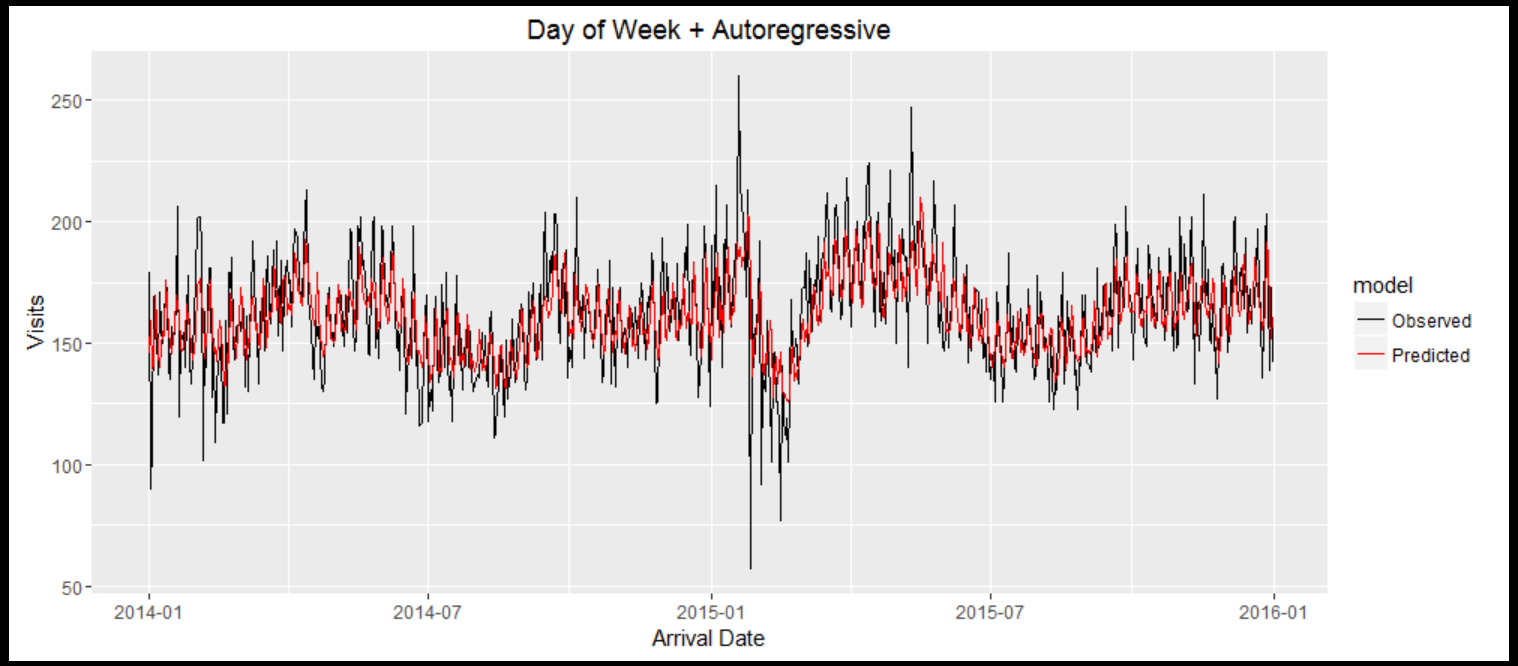


MAPE = 11.0%

Percent of days with bad staffing= 11.2%

In collaboration with: Sam Tideman, Mauricio Santillana, Jon Bickel, and Ben Reis

Add in Autoregression

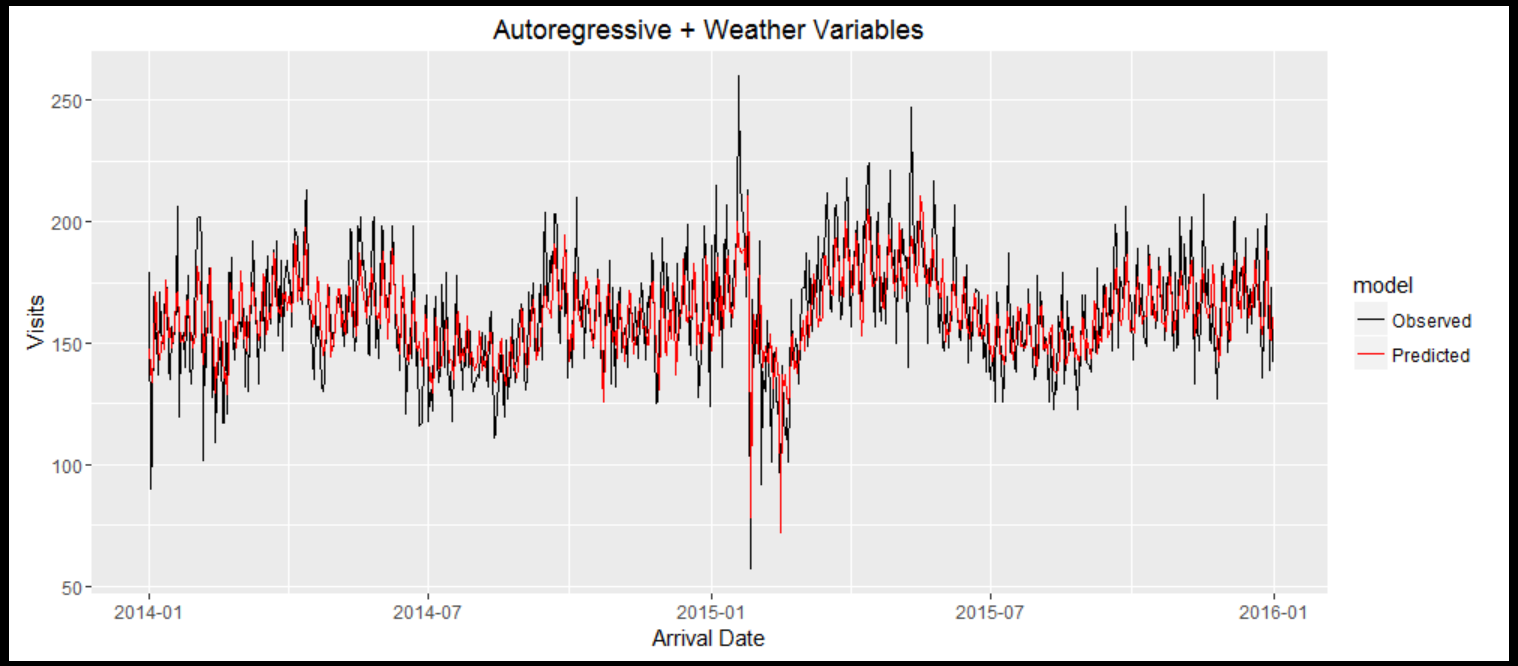


MAPE = 8.4%

Percent of days with bad staffing= 4.9%

In collaboration with: Sam Tideman, Mauricio Santillana, Jon Bickel, and Ben Reis

Add in Weather Data

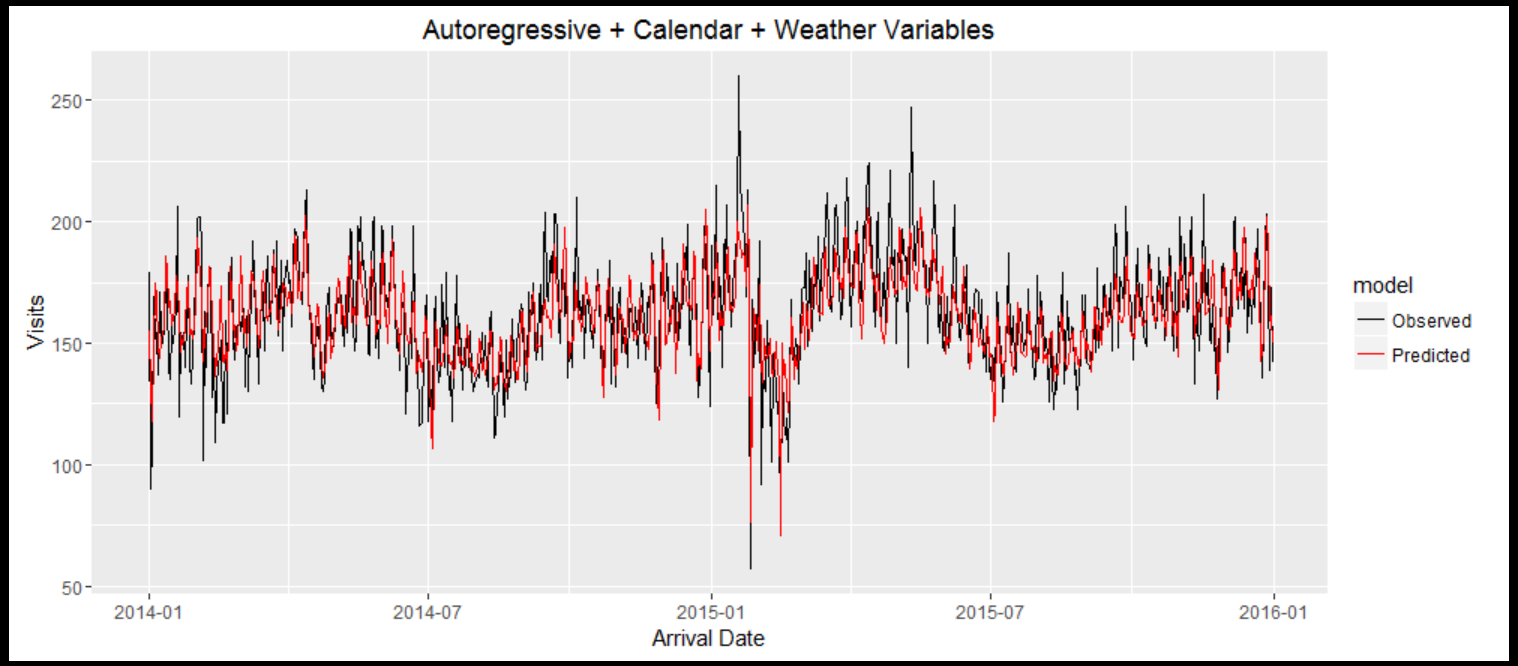


MAPE = 7.9%

Percent of days with bad staffing= 4.8%

In collaboration with: Sam Tideman, Mauricio Santillana, Jon Bickel, and Ben Reis

Add in Calendar Data

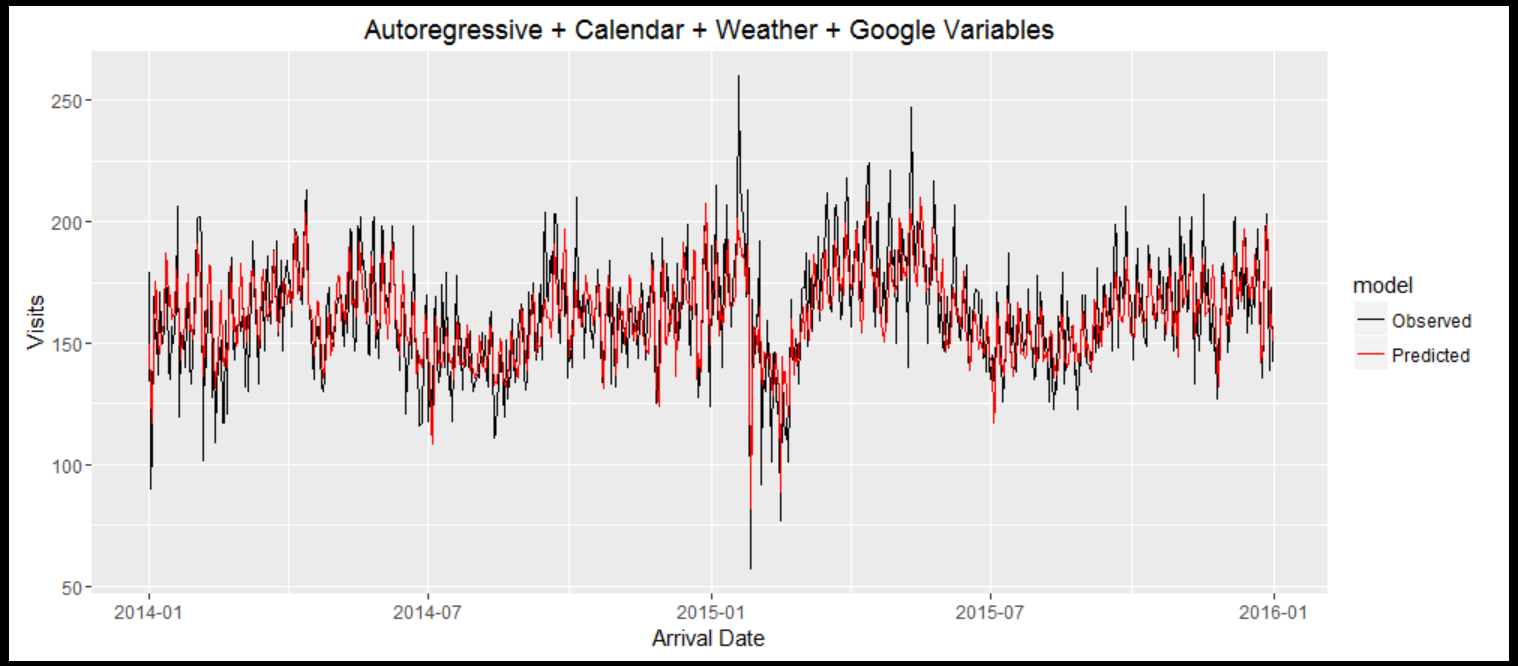


MAPE = 7.7%

Percent of days with bad staffing= 3.8%

In collaboration with: Sam Tideman, Mauricio Santillana, Jon Bickel, and Ben Reis

Add in Google Data



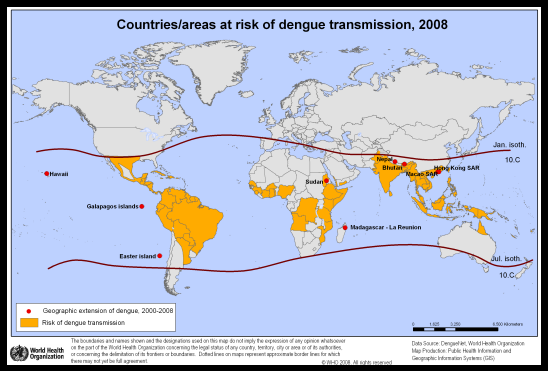
MAPE = 7.6%

Percent of days with bad staffing= 3.3%

In collaboration with: Sam Tideman, Mauricio Santillana, Jon Bickel, and Ben Reis

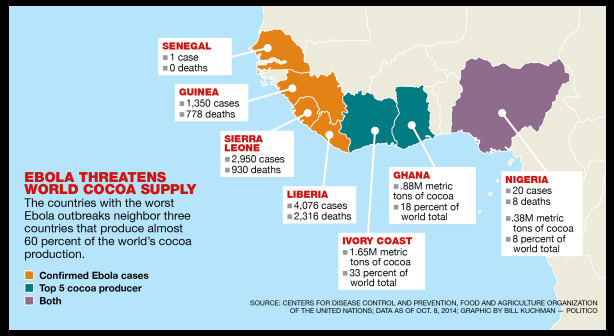
Part 2. Success stories in tracking and forecasting Flu, Zika, Dengue, Ebola in data-poor medium- to low-income countries.

Dengue, Zika, and Flu



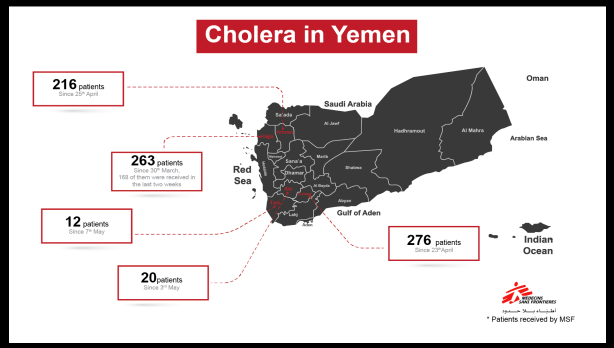
- Latin America (Flu, Zika, Dengue)
- South-east Asia (Dengue)

Ebola



- West Africa

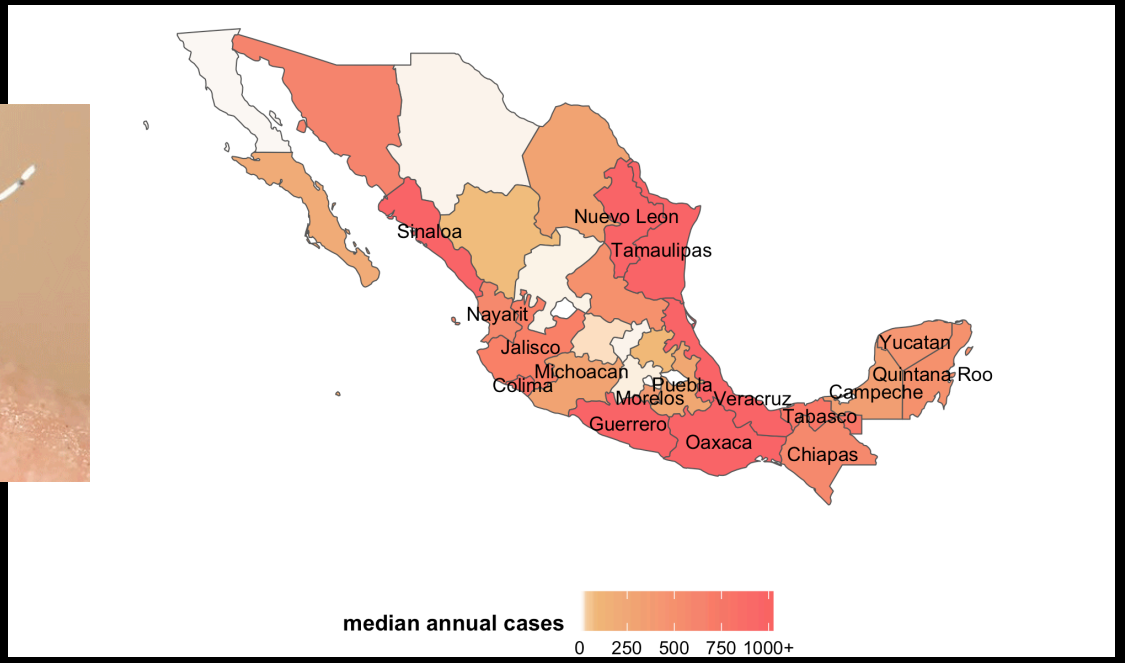
Cholera



- Middle East

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 Amherts)

SCIENTIFIC REPORTS

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

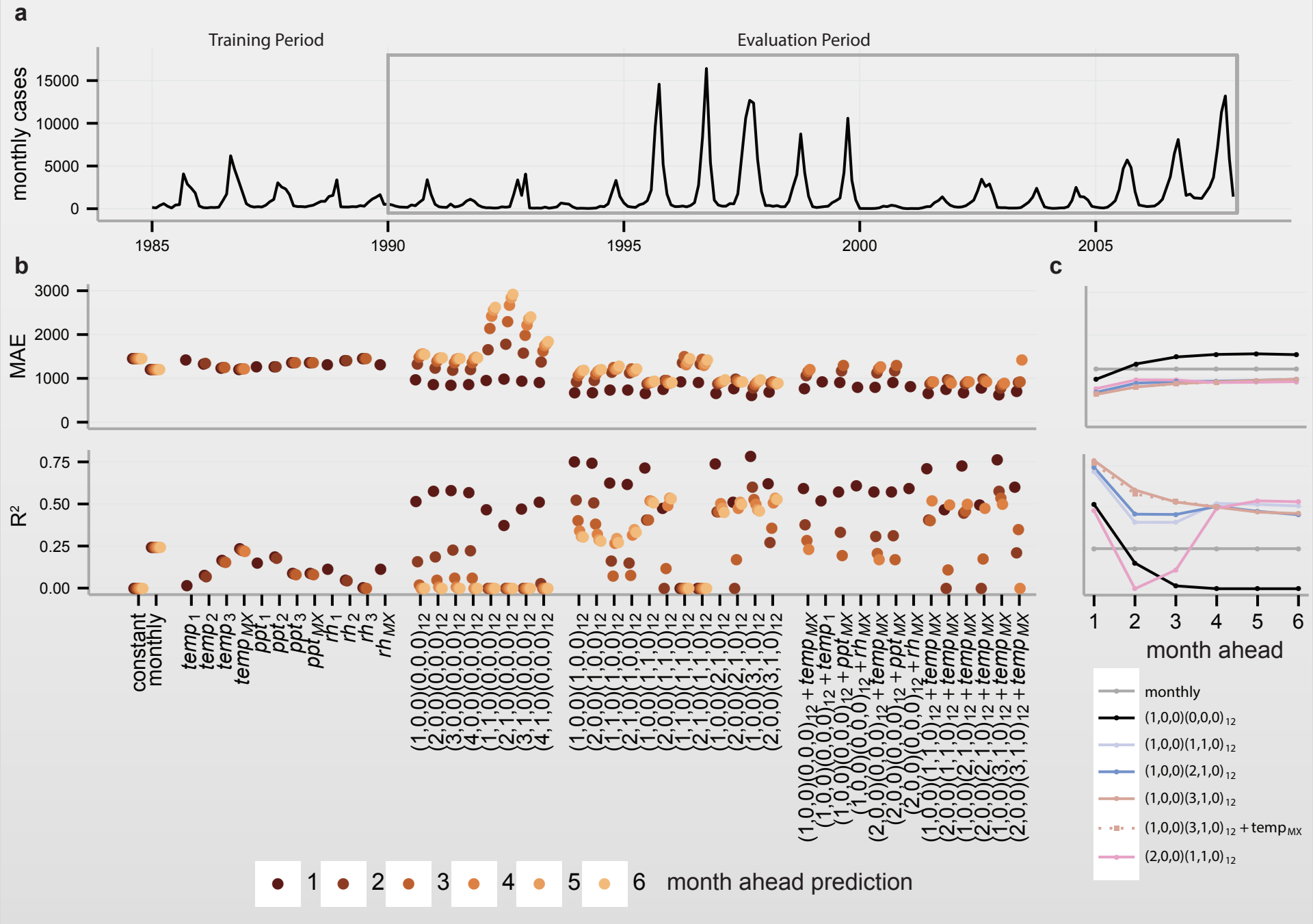
Download Citation

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

PLOS COMPUTATIONAL BIOLOGY

Browse Publish About

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

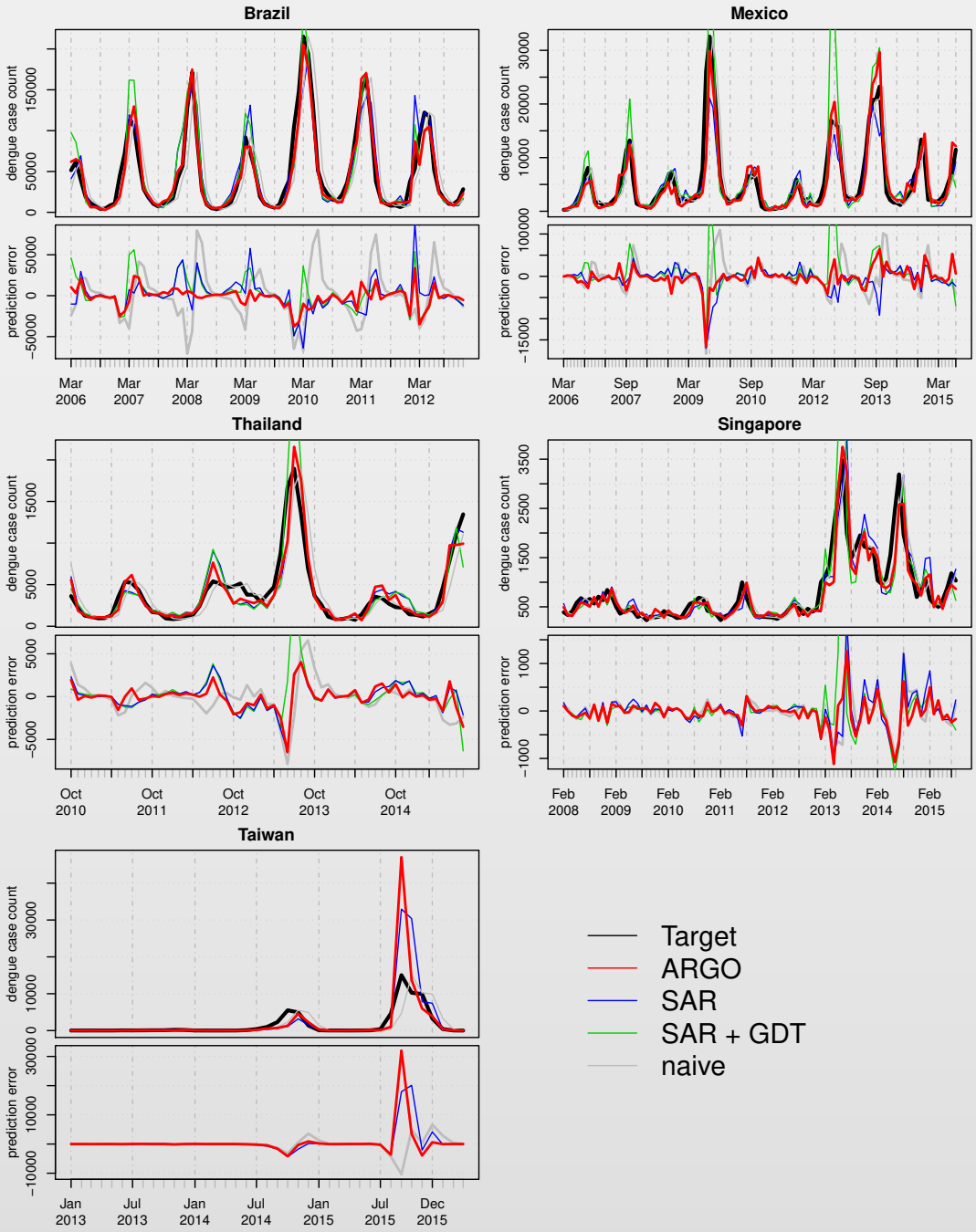
Advances in using Internet searches to track dengue

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

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


Brazil
dengue case count
0
50000
150000
prediction error
−50000
0
50000
Mar 2006
Mar 2007
Mar 2008
Mar 2009
Mar 2010
Mar 2011
Mar 2012
Mexico
dengue case count
0
10000
20000
30000
prediction error
−15000
0
100000
Mar 2006
Sep 2007
Mar 2009
Sep 2010
Mar 2012
Sep 2013
Mar 2015
Thailand
dengue case count
0
5000
15000
prediction error
−5000
0
5000
Oct 2010
Oct 2011
Oct 2012
Oct 2013
Oct 2014
Singapore
dengue case count
500
1500
2500
3500
prediction error
−1000
0
1000
Feb 2008
Feb 2009
Feb 2010
Feb 2011
Feb 2012
Feb 2013
Feb 2014
Feb 2015
Taiwan
dengue case count
0
10000
30000
prediction error
−10000
0
10000
30000
Jan 2013
Jul 2013
Jan 2014
Jul 2014
Jan 2015
Jul 2015
Dec 2015
Target
ARGO
SAR
SAR + GDT
naive

PLOS | CURRENTS OUTBREAKS

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

APRIL 28, 2015 · COMMENTARY

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.

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

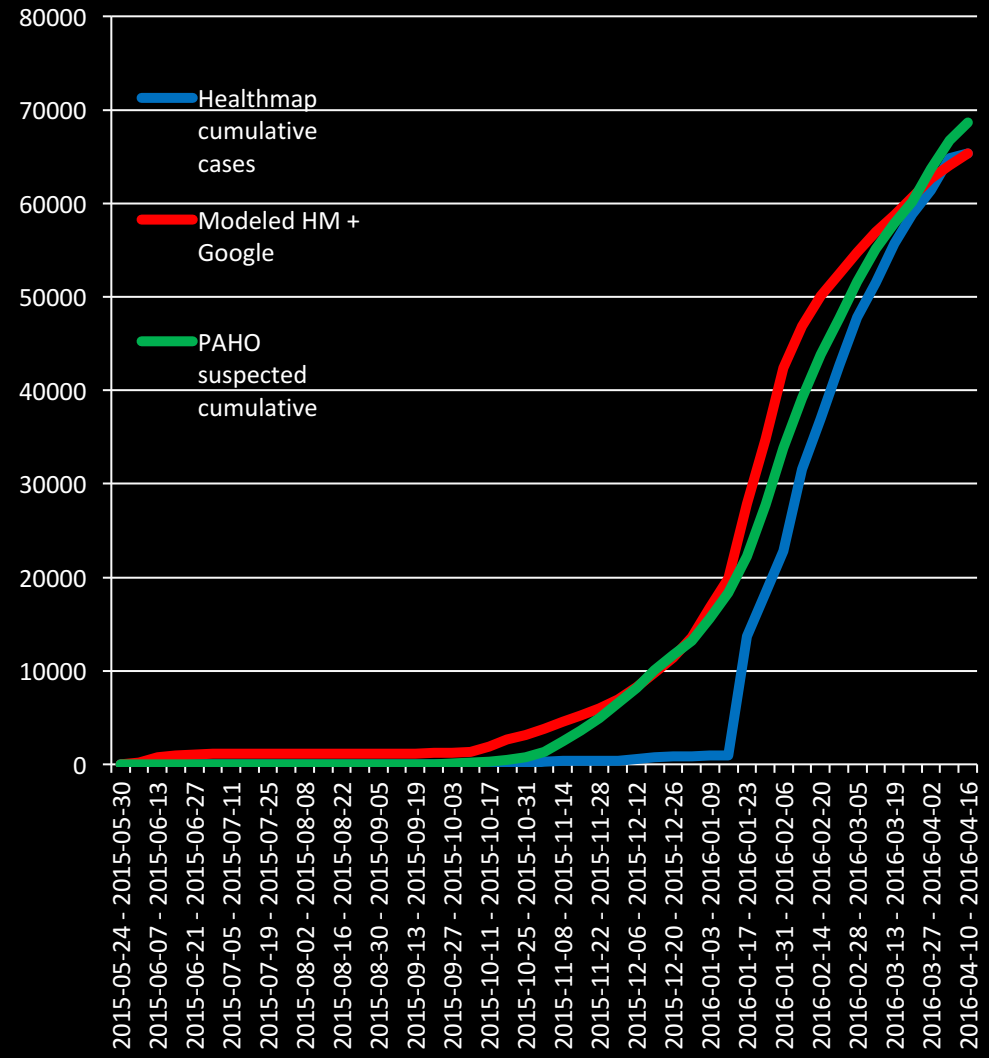
Original Paper

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 McGinnis[1], 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



HealthMap

+

Google Trends

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.

Forecasting Zika using Google searches and Twitter

PLOS | NEGLECTED TROPICAL DISEASES

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

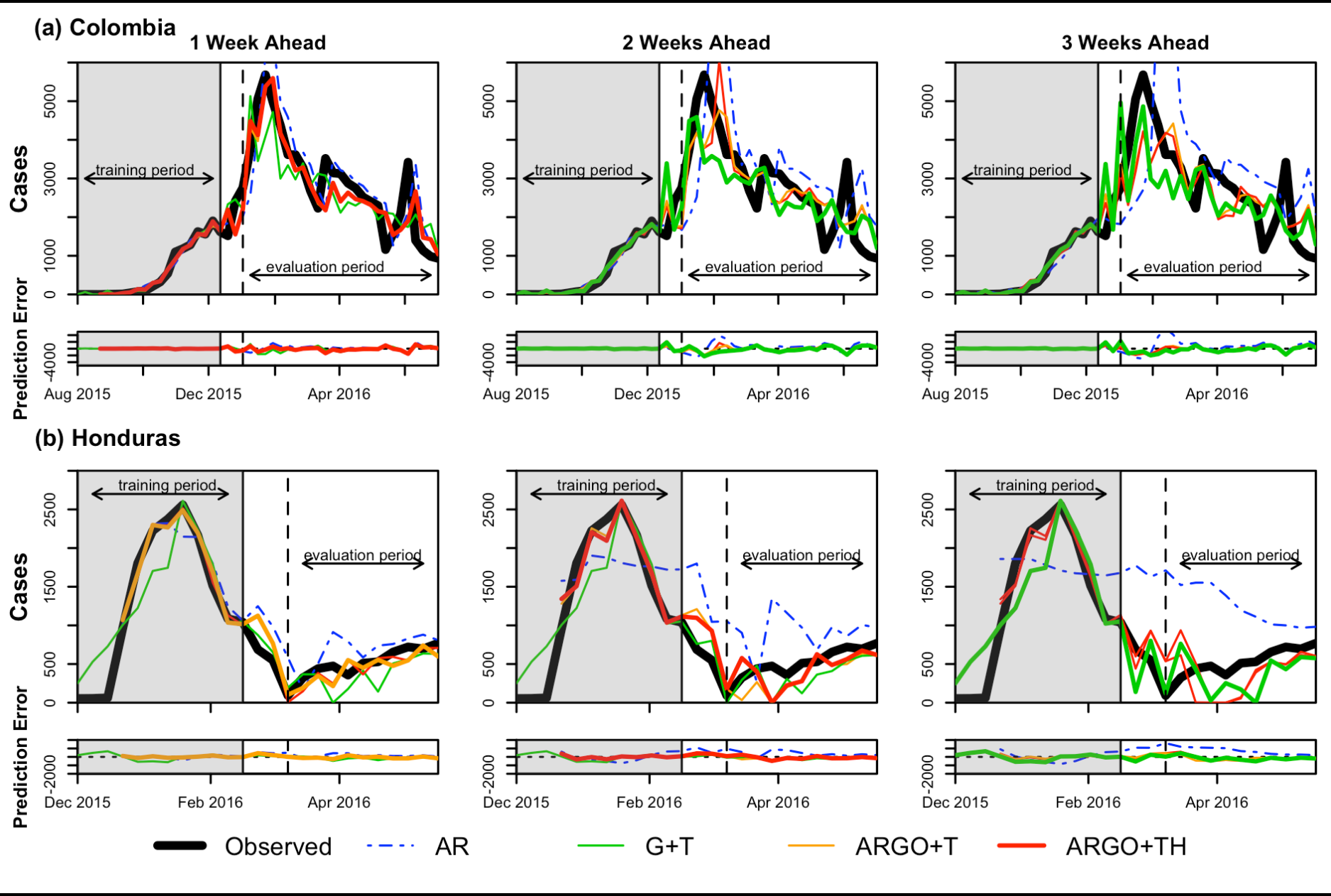
Abstract

Background

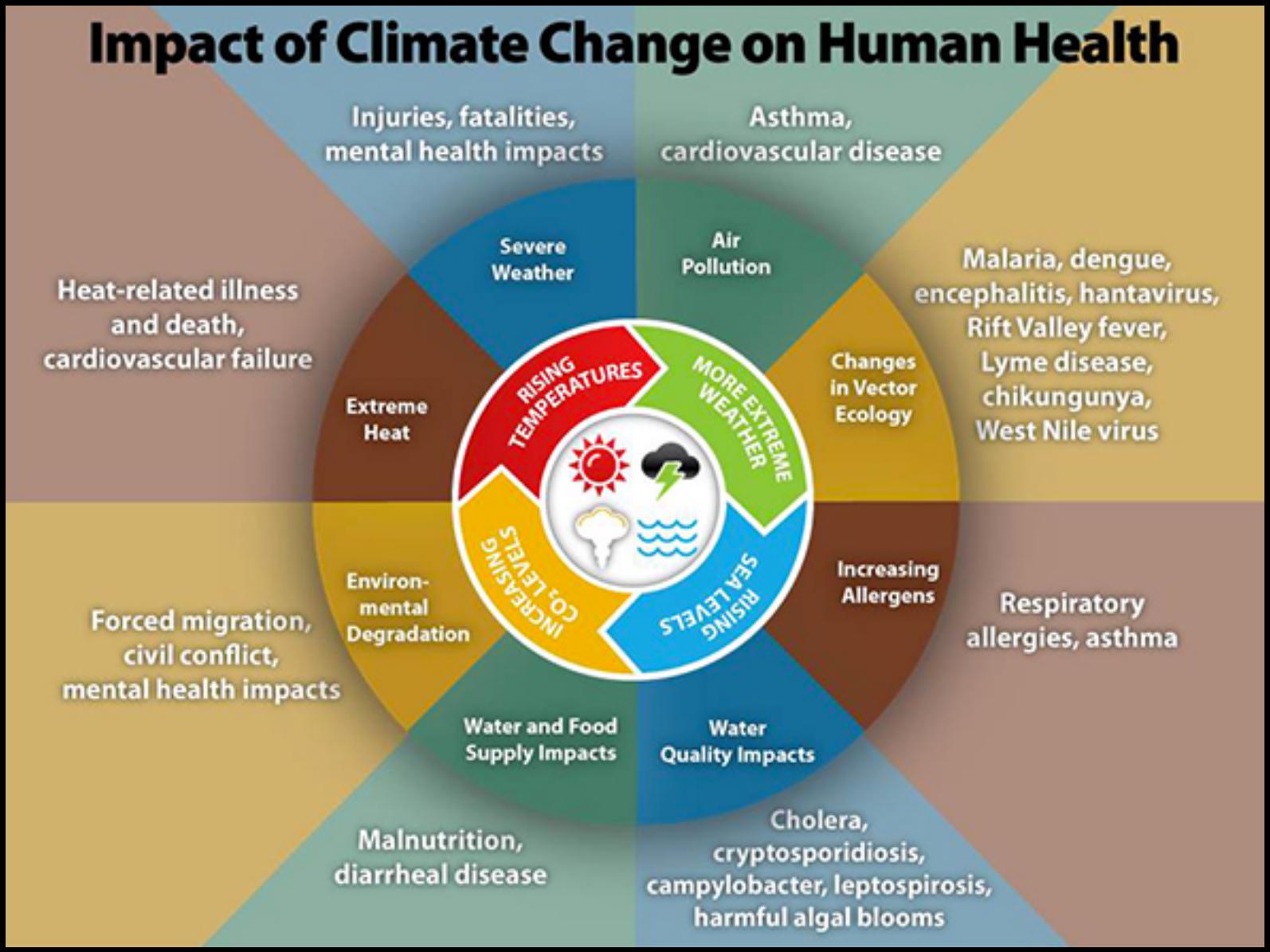
Over 400,000 people across the Americas are thought to have been infected with Zika virus as a consequence of the 2015–2016 Latin American outbreak. Official government-led case count data in Latin America are typically delayed by several weeks, making it difficult to track the disease in a timely manner.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

Forecasting Zika using Google searches and Twitter

(with Sarah McGough)

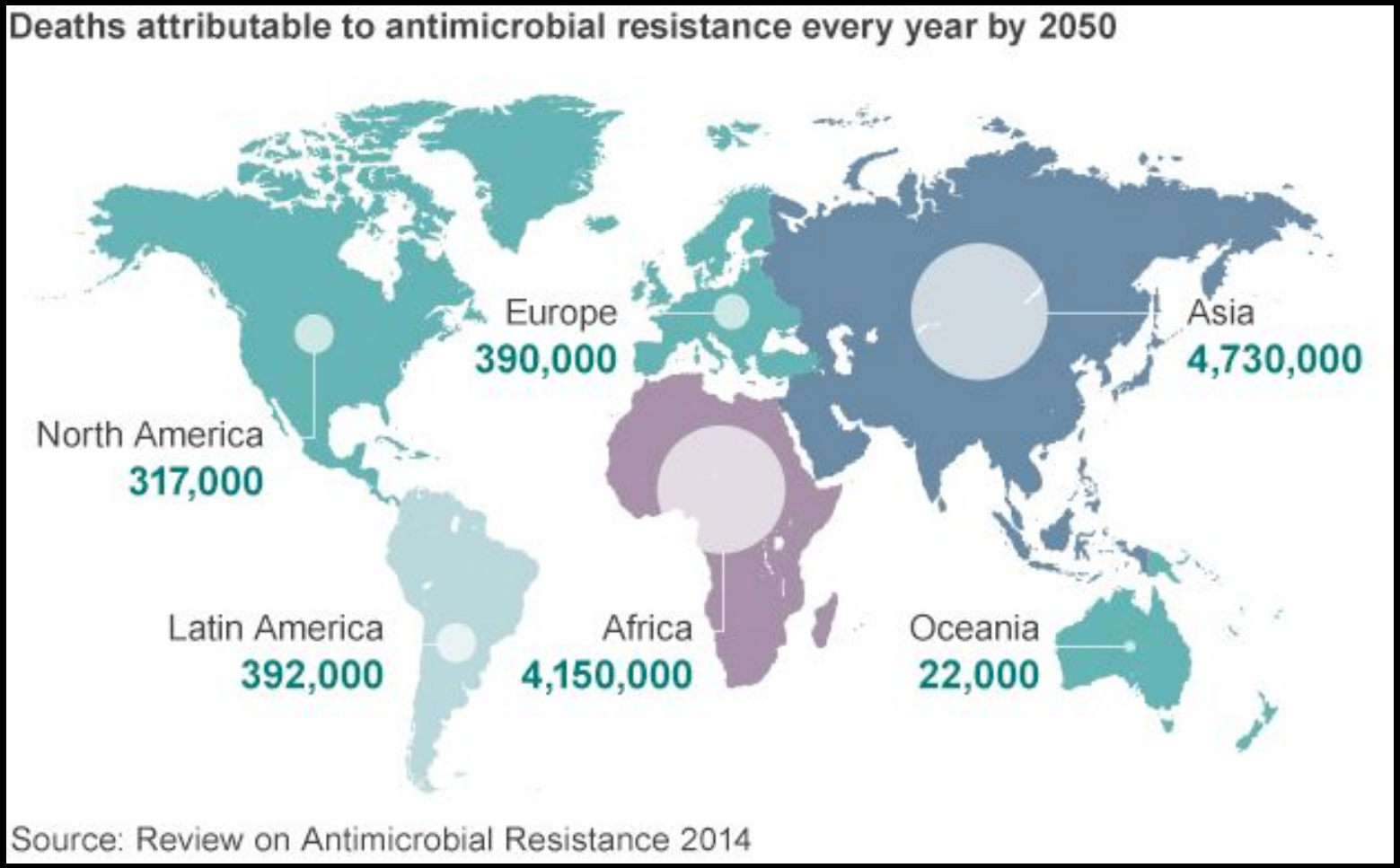


The influence of weather/climate on public health



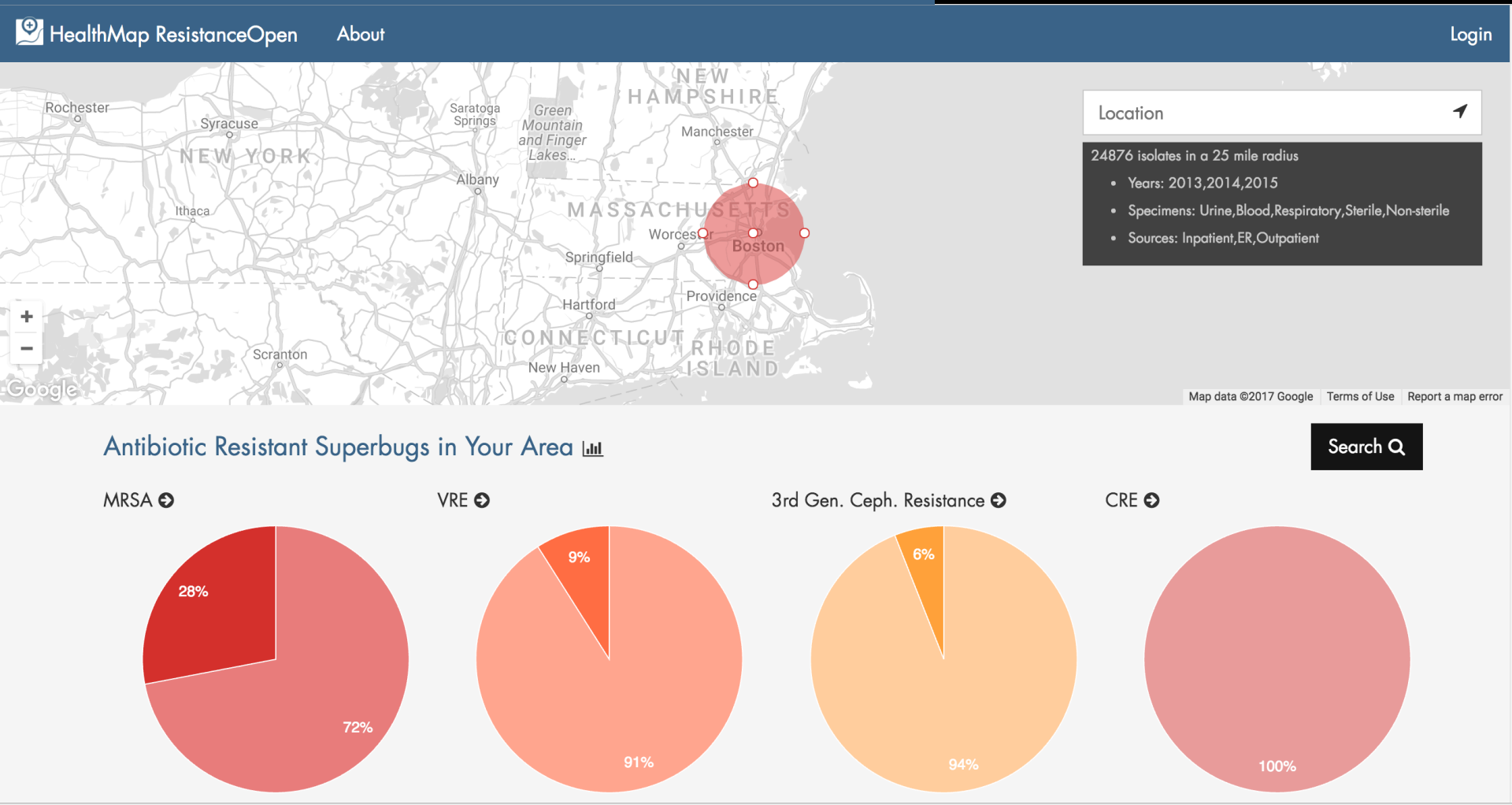
Source: Centers for Disease Control and Prevention

Antibiotic resistance is now recognized as one of the worlds greatest public health threats, with the potential to render existing antibiotics ineffective in the “not so far” future.



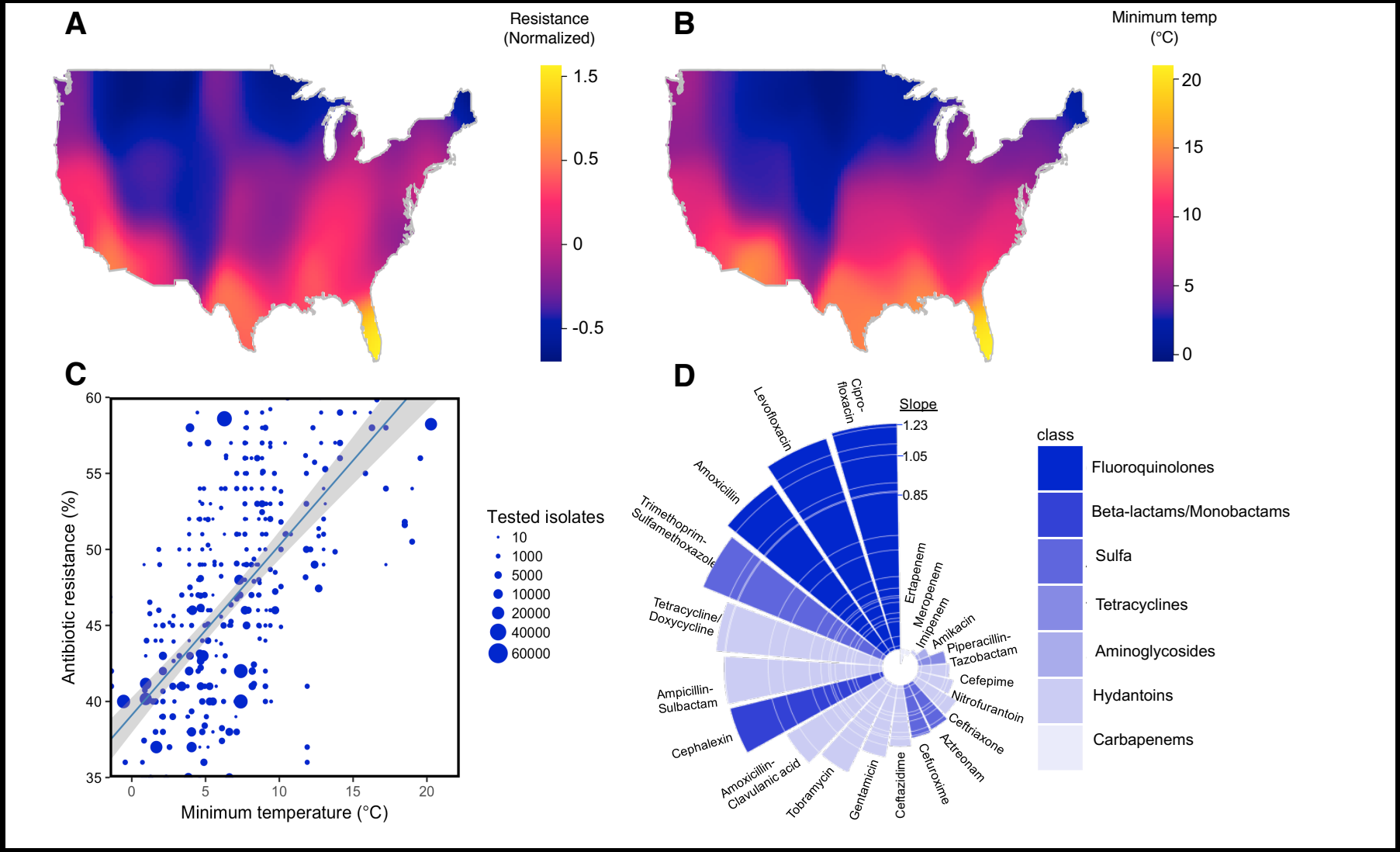
HealthMap ResistanceOpen

Tool conceived and implemented by *Derek R. MacFadden and John S. Brownstein*



Antibiotic Resistance Increases with Local Temperature

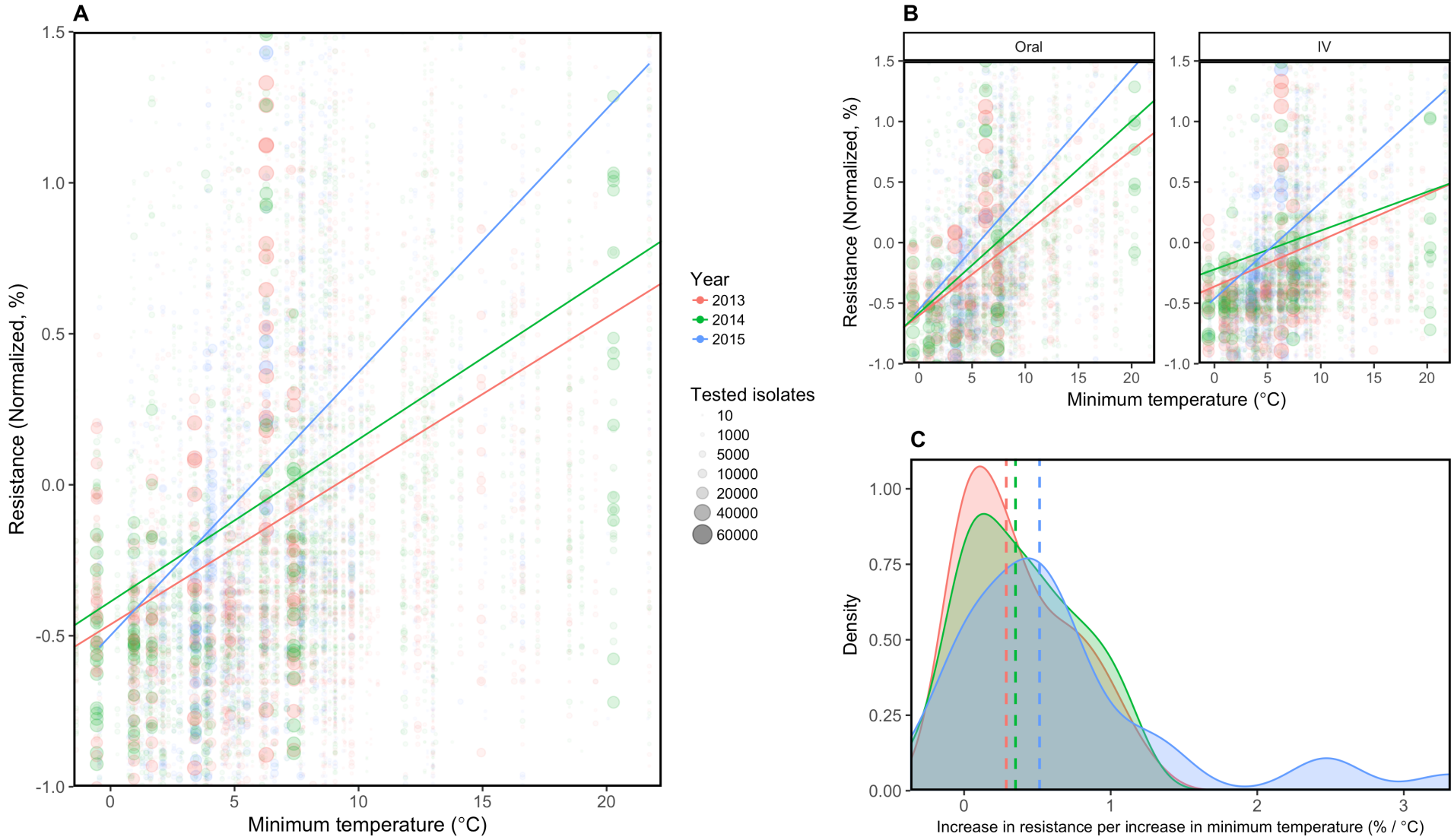
(Shown in figure *E. coli* resistance for all antibiotics)



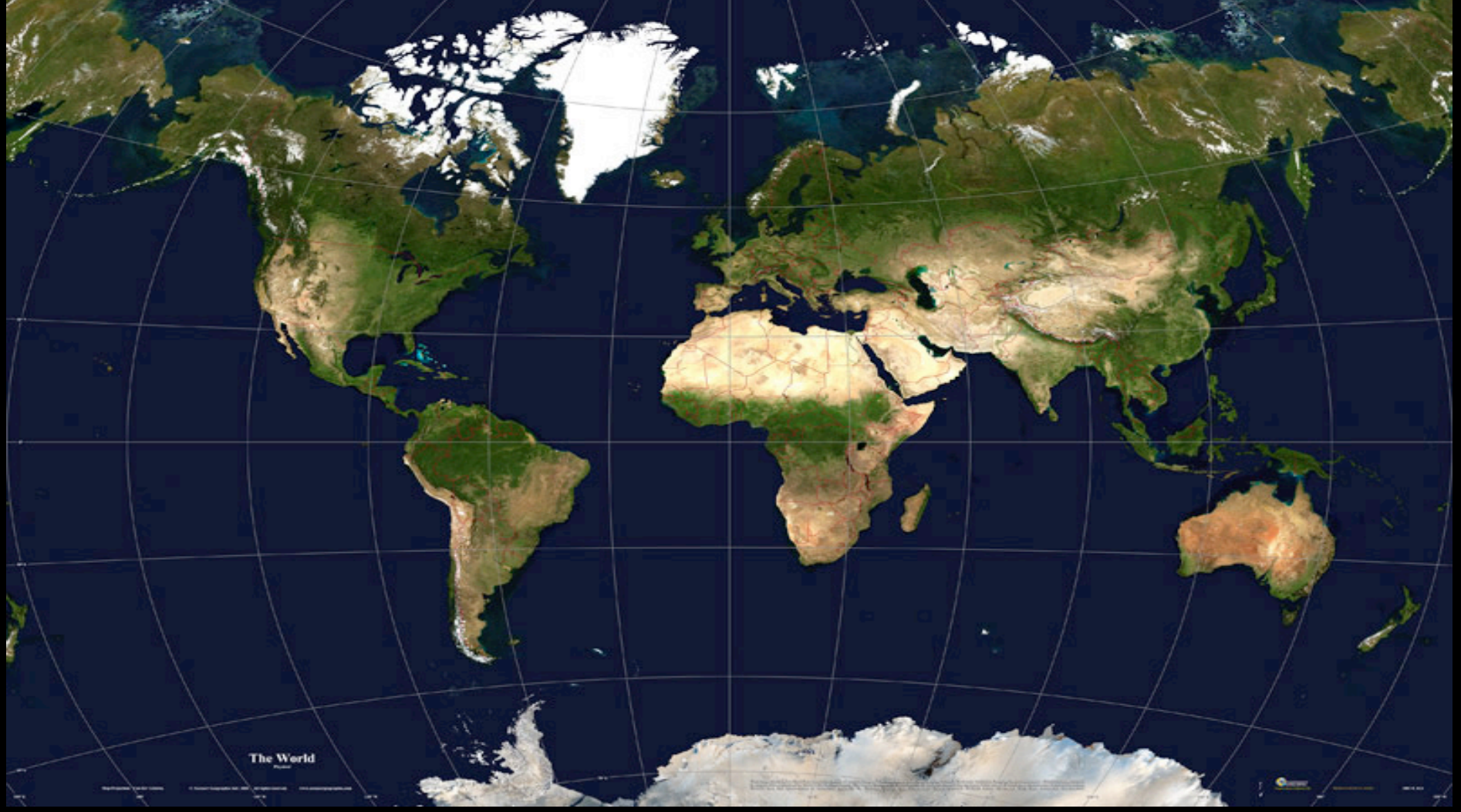
Team: Derek R. MacFadden, Sarah F. McGough, David Fisman, John S. Brownstein, and Mauricio Santillana. Nature Climate Change, May 2018

Antibiotic Resistance Increases with Local Temperature

(Shown in figure *E. coli* resistance for all antibiotics)



Team: Derek R. MacFadden, Sarah F. McGough, David Fisman, John S. Brownstein, and Mauricio Santillana. Nature Climate Change, May 2018



Thank you!

Contact: msantill@fas.harvard.edu