LA-UR-16-27157
Approved for public release; distribution is unlimited.

Title: Real-time Social Internet Data to Guide Forecasting Models

Author(s): Del Valle, Sara Y.

Intended for: Sponsor briefings

Issued: 2016-09-20

Real-time Social Internet Data to Guide Forecasting Models

Sara Del Valle, PhD

Team: Geoffrey Fairchild, Nick Generous, Kyle Hickmann, Dave Osthus, Reid Priedhorsky

Our goal

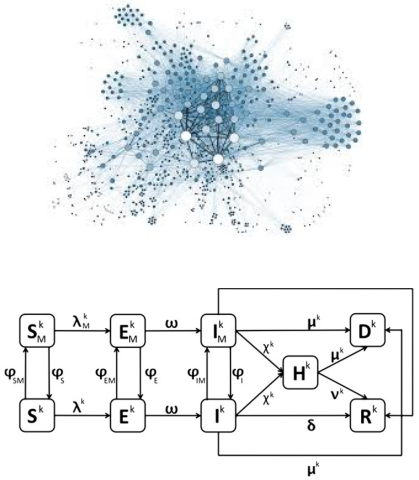
Improve decision support by monitoring and forecasting events using social media, mathematical models, and quantifying model uncertainty.

Our approach

Real-time, data-driven forecasts with quantified uncertainty:
Not just for weather anymore.

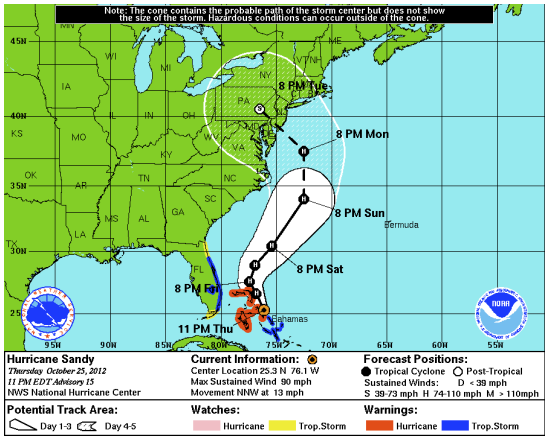
Real-time, voluminous, extremely noisy data

+



Mathematical models

=



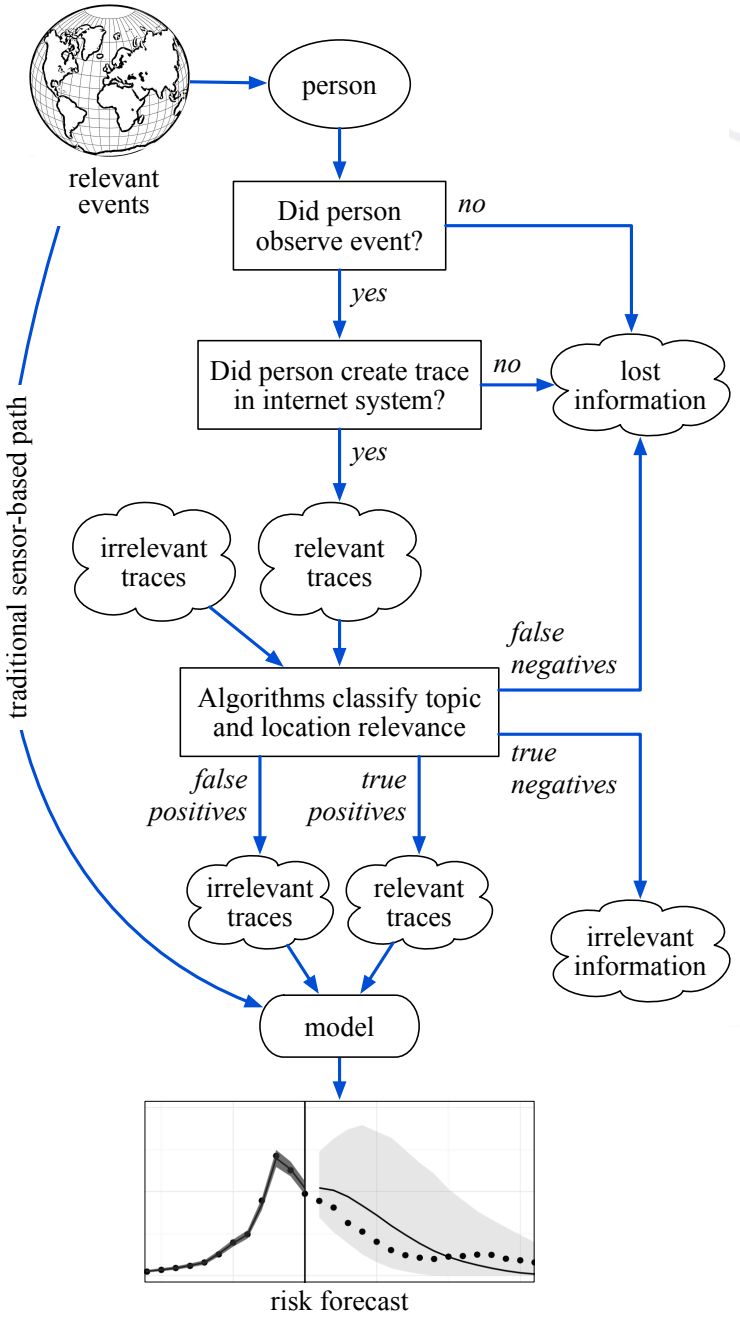
Forecasts with quantified uncertainty

1 + 1 = 3

The whole is greater than the sum of its parts! – Aristotle

Our approach

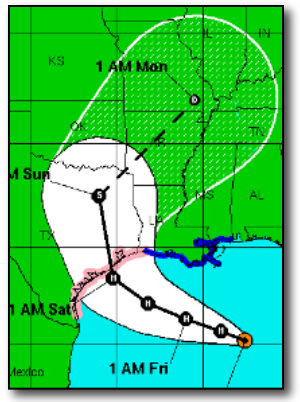
Information flow from human observations of events through an Internet system and classification algorithms to produce quantitatively uncertain forecast.



What do we want?

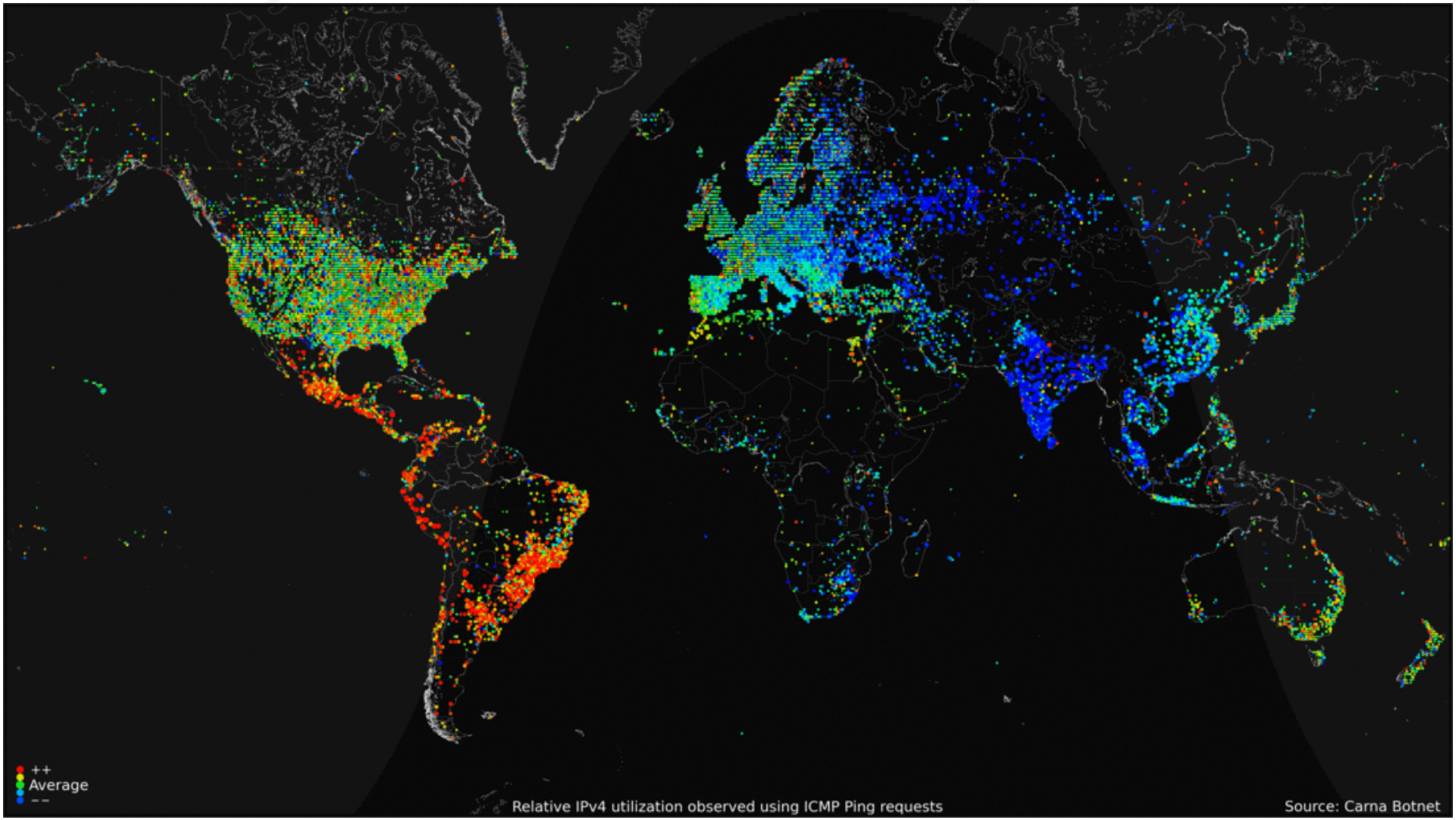
- **Holy grail**
  - Predict events before they occur/start

- **Realistic alternative**
  - Accurate, quantitatively uncertain **forecast**



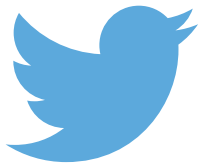
- **Our proposal**
  - Social Internet data can help

Global internet penetration



Internet data sources

- Information sharing



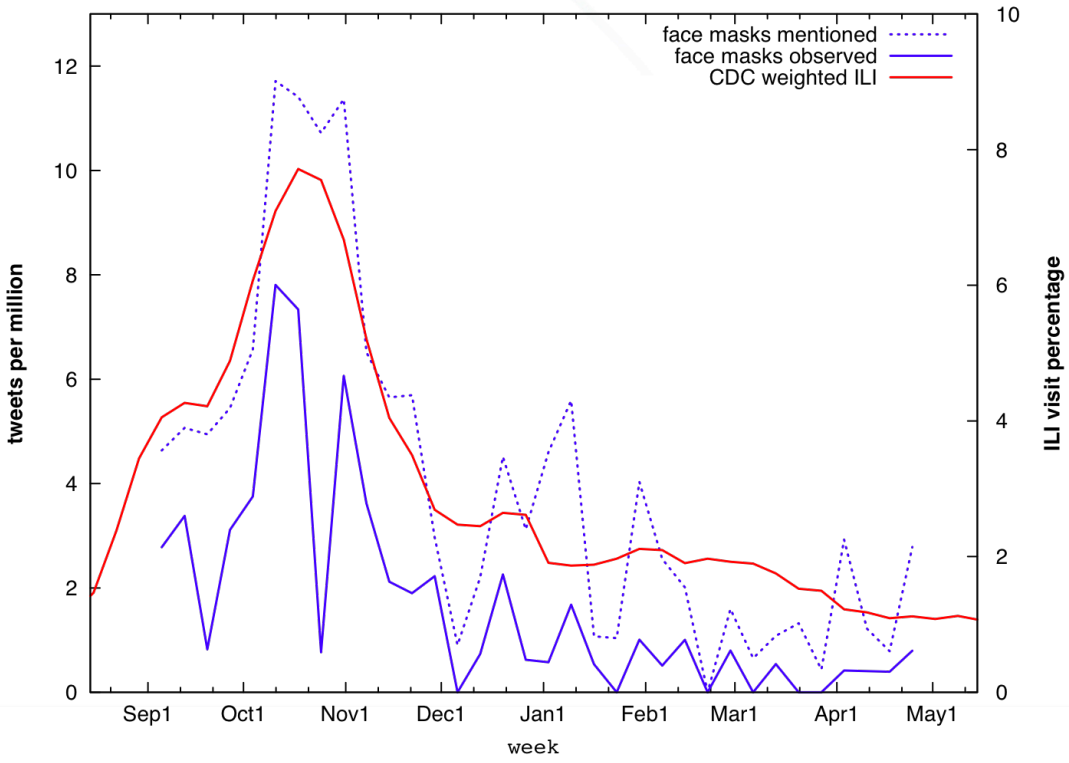
- Information seeking

WIKIPEDIA
The Free Encyclopedia

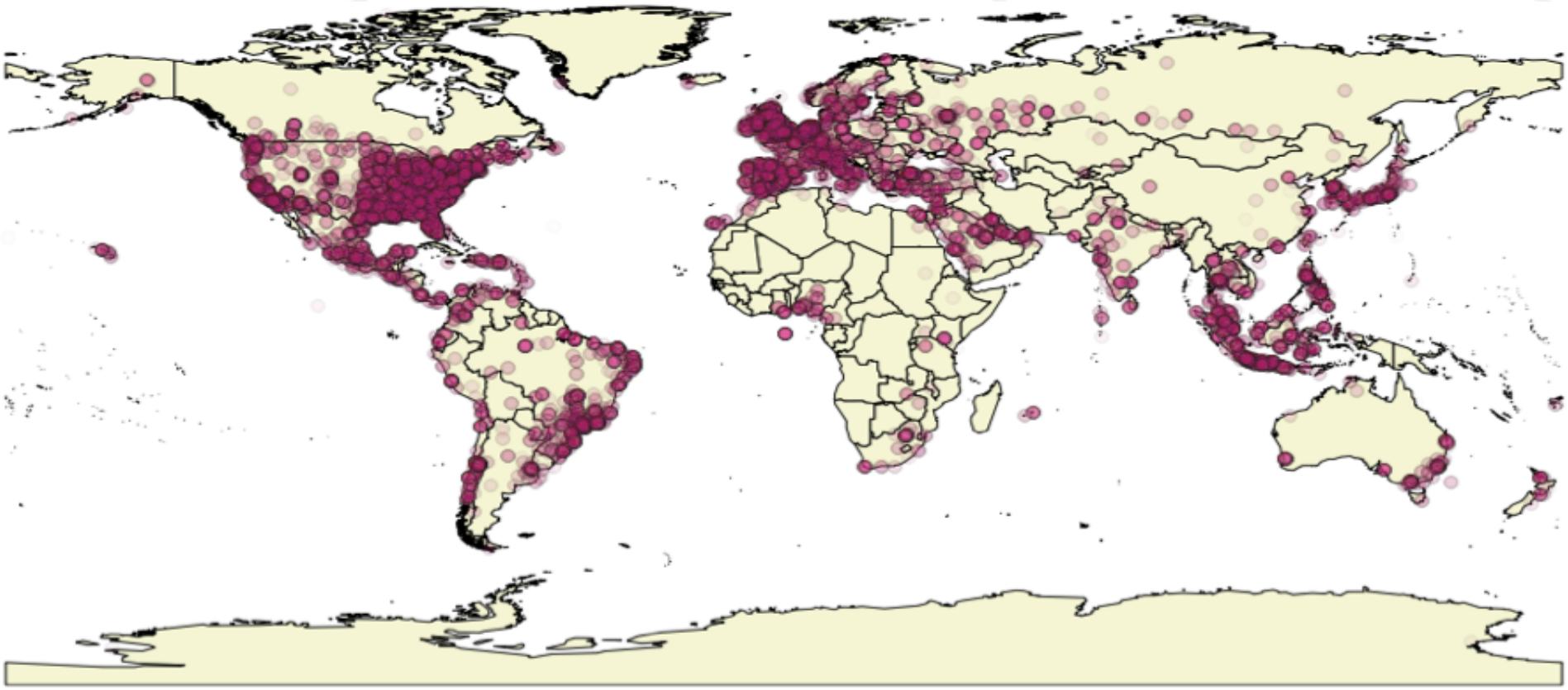
CENTERS FOR DISEASE
CONTROL AND PREVENTION

Using Twitter to Extract Behavior

- 10,000 Tweets/hour about “swine flu” in 2009
- Extract emergent behaviors such as facemask usage
- Use information to inform models and quantify impact



Where are people tweeting from?



- Challenges
  - Only 1-2% of tweets carry a geotag
  - Behavior is not demographically and **geographically** uniform

Locating a tweet

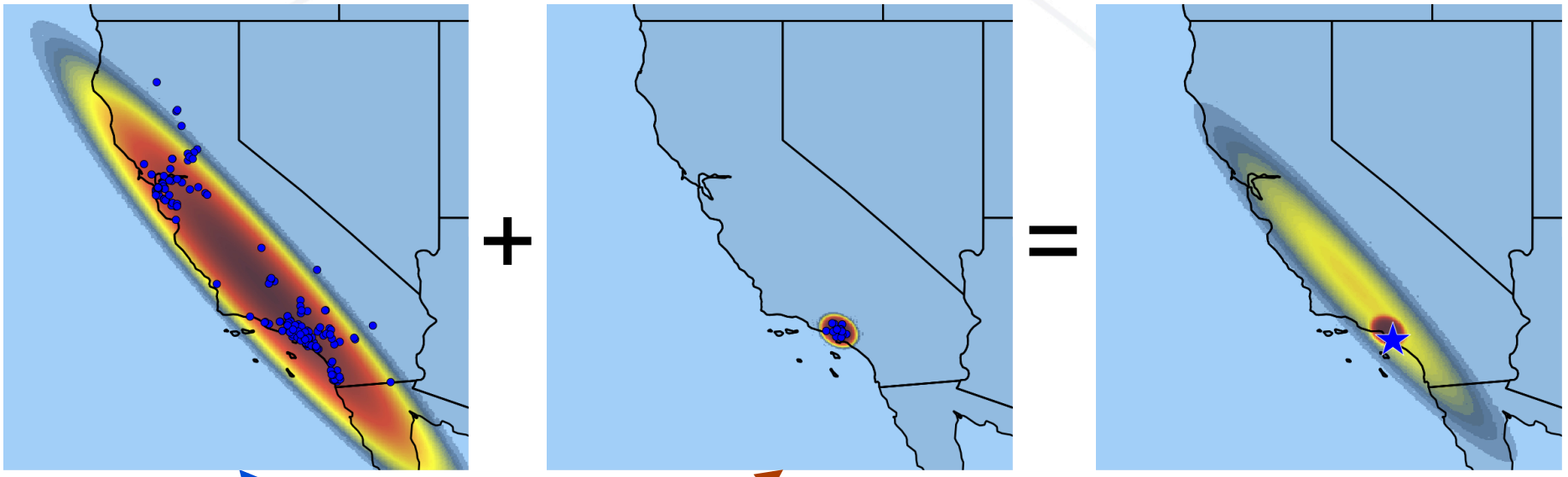
- Scalable, content-based approach – Gaussian mixture model
- Experiments on 13M global, multi-lingual tweets
- Only requires 30,000 tweets to train model

Fields: user location, user time zone, tweet text, user description, user language

Rank	Fields					MCAE	success
1	lo	tz	tx		ln	1823	100.0%
2	lo	tz	tx	ds	ln	1826	100.0%
3	lo	tz				1862	87.7%
4	lo	tz	tx			1878	99.2%
5	lo	tz	tx	ds		1908	99.6%
6	lo	tz		ds		2013	94.1%
7	lo	tz		ds	ln	2121	100.0%
8	lo					2125	65.8%
9	lo		tx	ds	ln	2176	100.0%
10	lo	tz			ln	2207	100.0%
11	lo		tx	ds		2274	99.2%
12	lo		tx		ln	2310	100.0%
13	lo		tx			2383	98.0%
14		tz	tx	ds	ln	2492	100.0%
15	lo			ds		2585	88.3%
16		tz	tx	ds		2594	99.4%
17		tz	tx		ln	2617	100.0%
18		tz	tx			2691	98.7%
19	lo			ds	ln	2759	100.0%
20		tz				2945	76.1%
21		tz		ds		2991	91.8%
22		tz		ds	ln	3039	100.0%
23	lo				ln	3253	100.0%
24			tx	ds	ln	3267	100.0%
25			tx	ds		3426	98.8%
26		tz			ln	3496	100.0%
27			tx		ln	3685	100.0%
28			tx			3855	95.7%
29				ds		4482	79.7%
30				ds	ln	4484	100.0%
31					ln	6143	100.0%

MCAE: Mean Comprehensive Accuracy Error

Locating a tweet



0.12 ×

0.87 ×

(80th percentile)

text: Americans are optimistic about the economy & like what Obama is doing. What is he doing? Campaigning and playing golf. Ignorance is bliss.

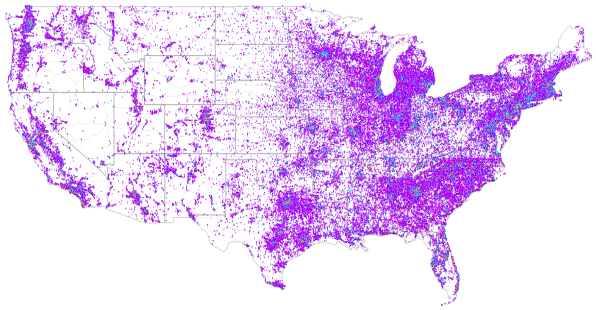
language: en

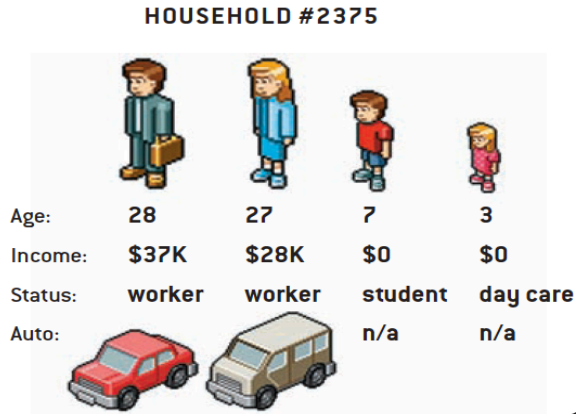
location: Los Angeles, CA

time zone: Pacific Time (US & Canada)

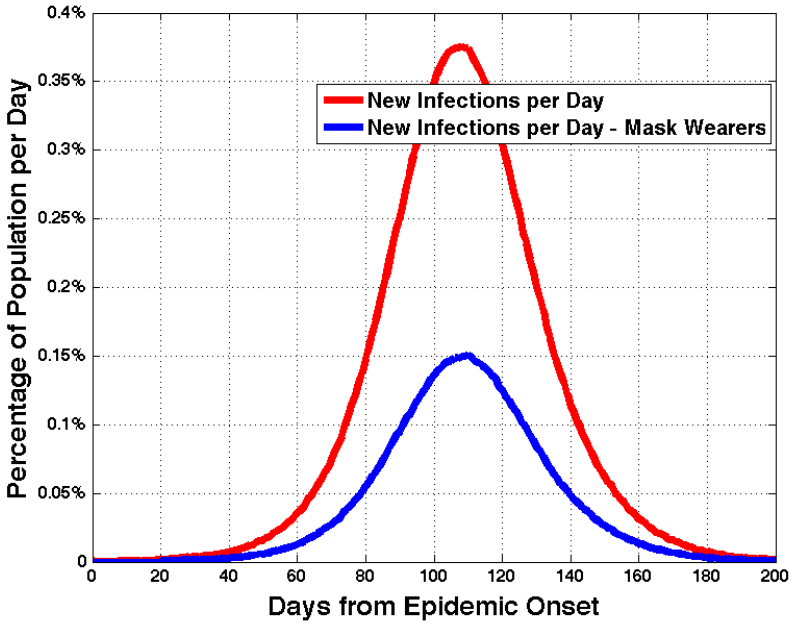
Agent-based Simulation

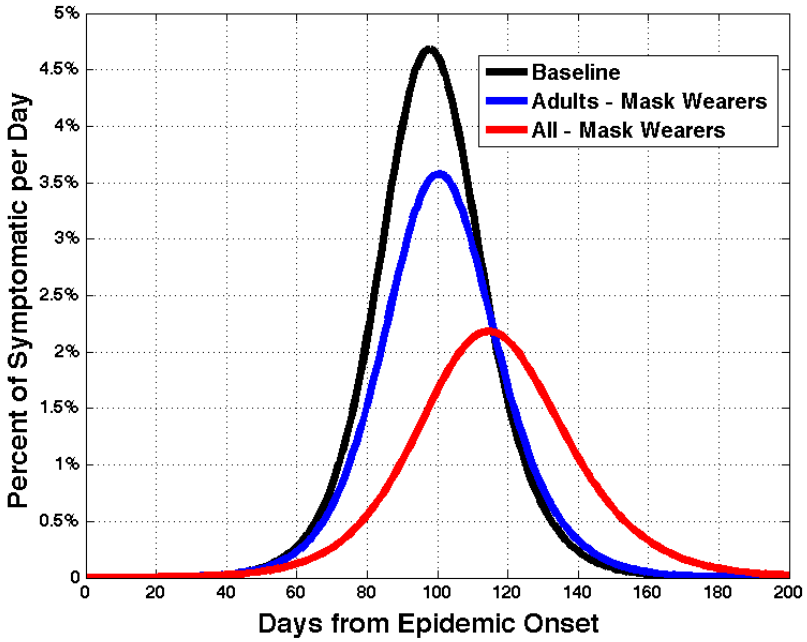
- Simulates the impact of disease spread and mitigation strategies within the U.S.
- Synthetic population with demographic and geospatial differentiation
- Workforce absenteeism by industry classification
- Explicit contact patterns and changes in human behavior





Quantify impact of facemask usage

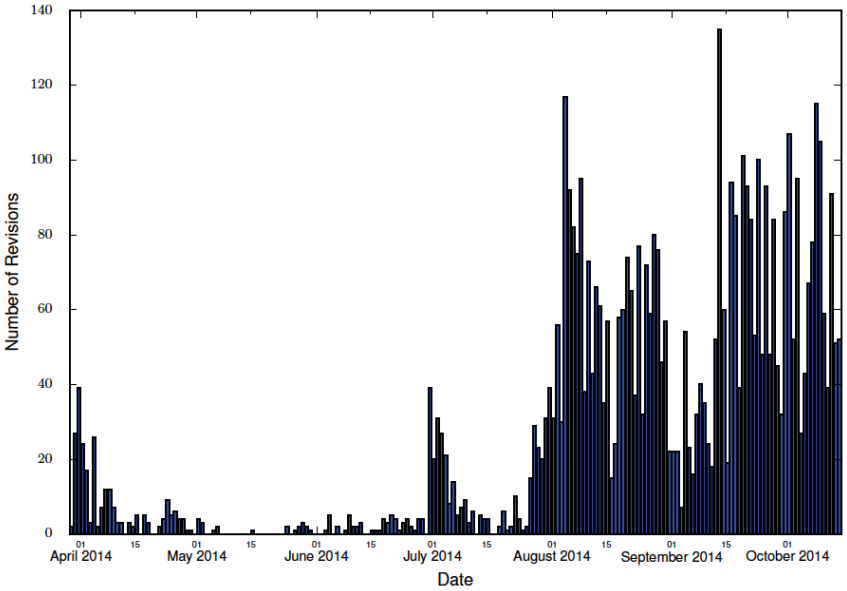




Compliance rates by demographic characteristics (age & gender)

Eliciting data from Wikipedia articles

- One stop shop for monitoring an event – Wikipedia article content
- Trained named-entity recognizer
- Wikipedia information closely align with ground truth data

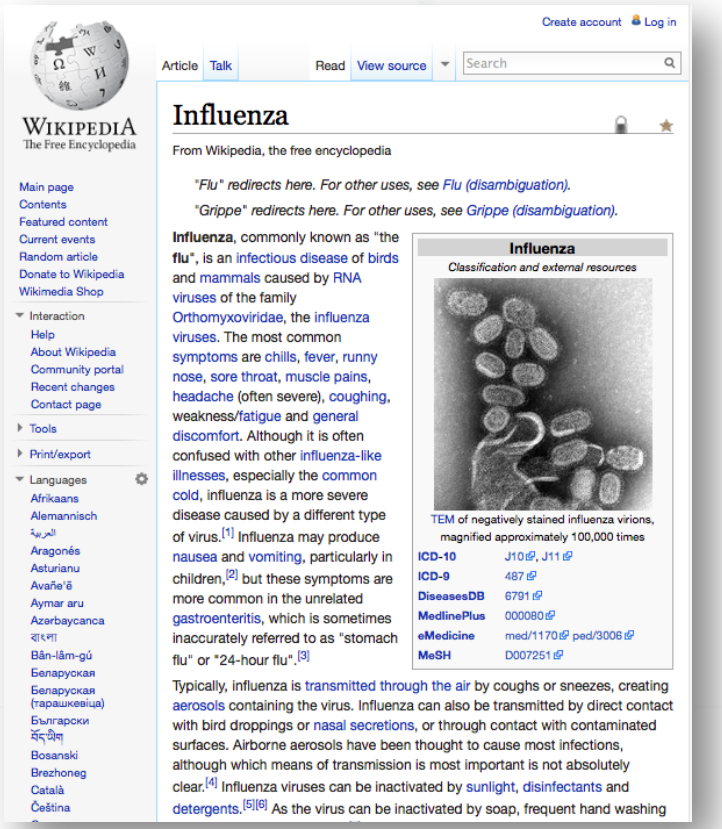


2014 Ebola Epidemic – cases, deaths, and hospitalizations

WIKIPEDIA
The Free Encyclopedia

Using Wikipedia to track incidence

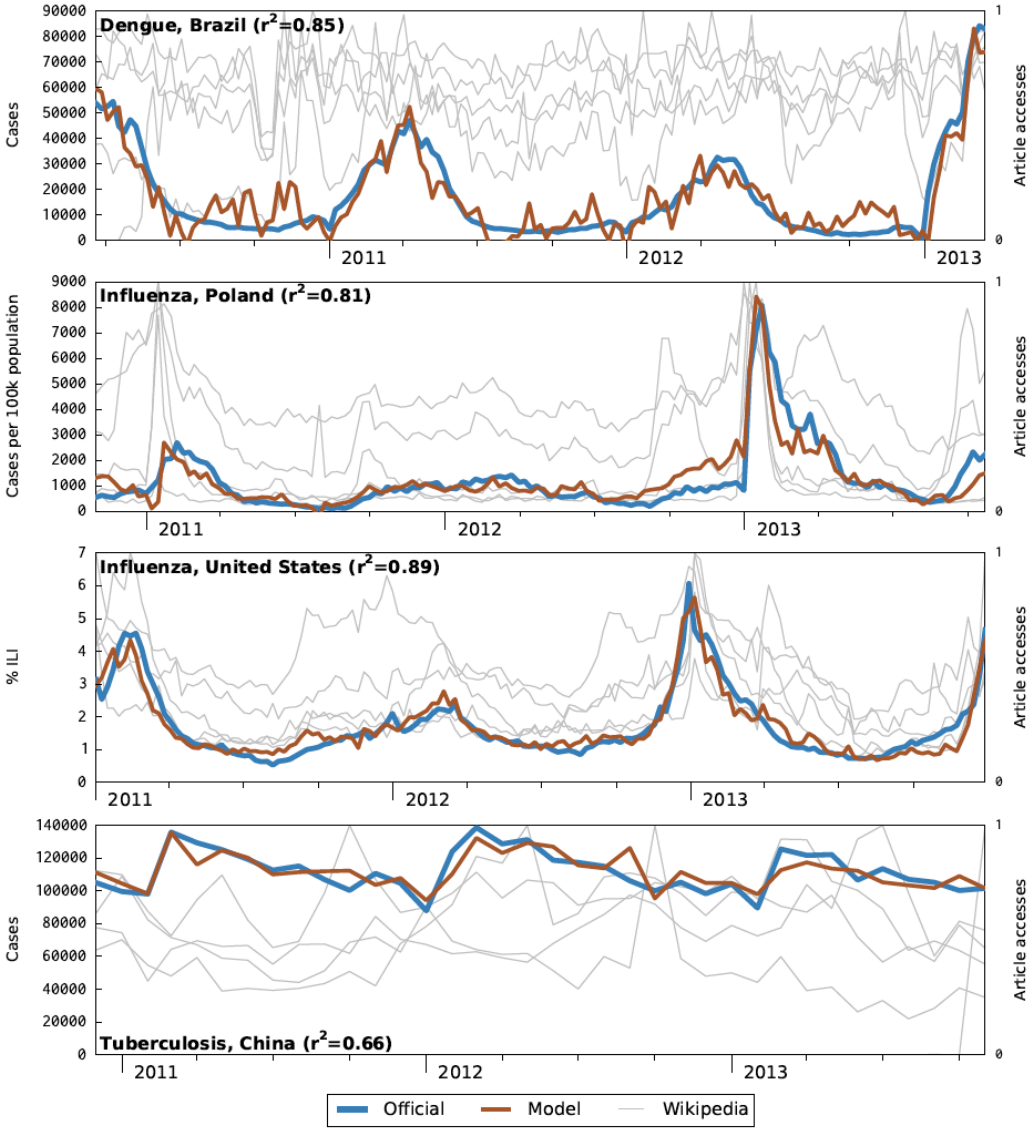
- 30 million articles in 287 languages
- 6th most popular website
- Article access logs available since 2007
- All data freely available on an hourly basis



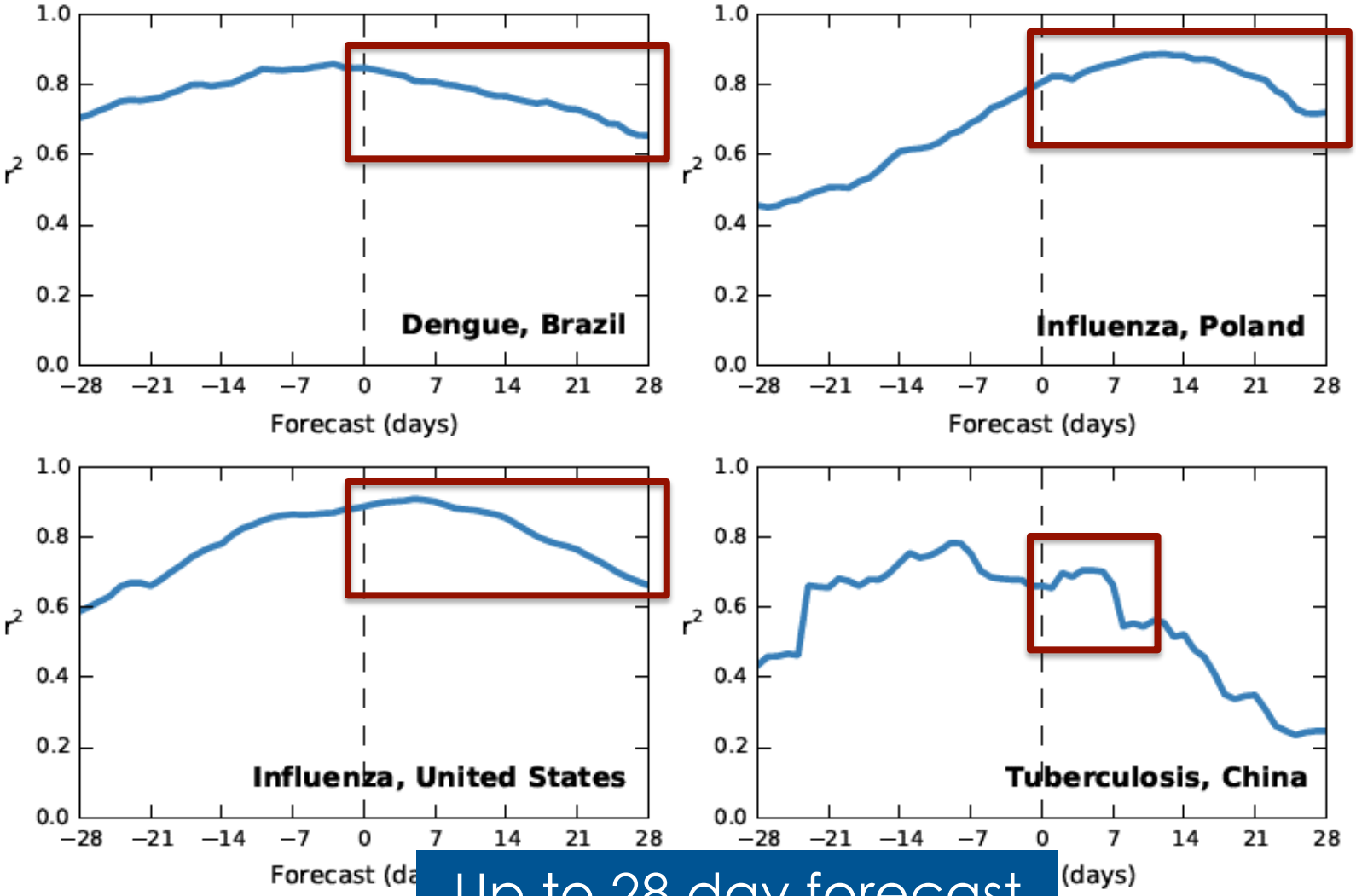
Global disease monitoring

- Wikipedia article traffic
- Linear models
- Monitor disease around the globe

Wikipedia traffic correlates with disease incidence



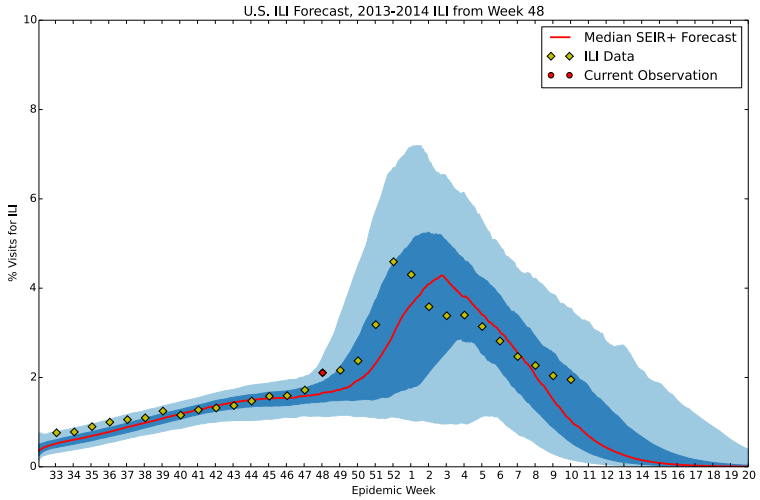
Forecasting

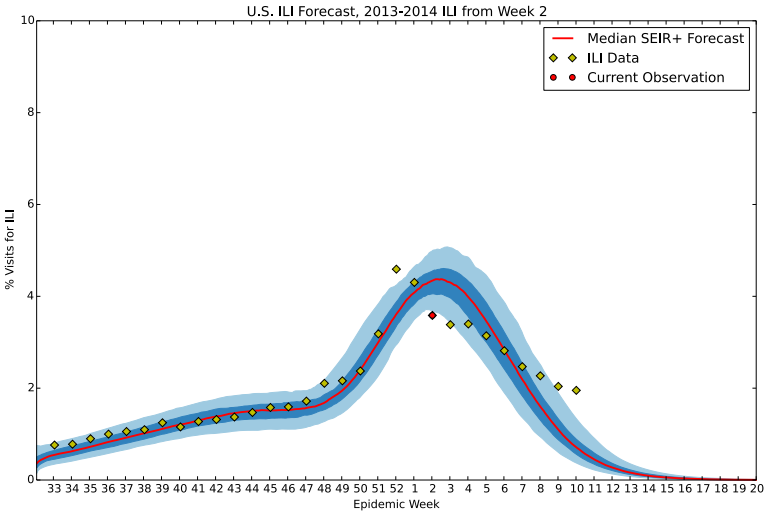


Up to 28 day forecast for some diseases.

2013-2014 Influenza Forecasts

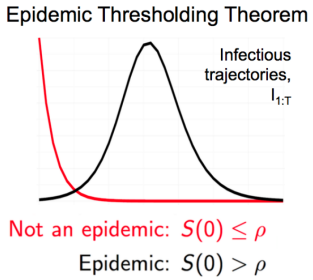
- Wikipedia traffic
- SEIR (susceptible – exposed – infectious – recovered) models
  - seasonality
  - heterogeneous mixing
  - Data assimilation
- Bi-weekly forecasts of the 2013-2014 flu season

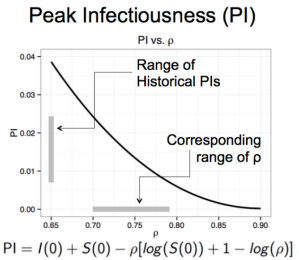


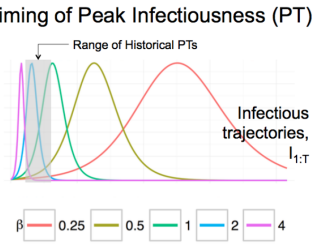


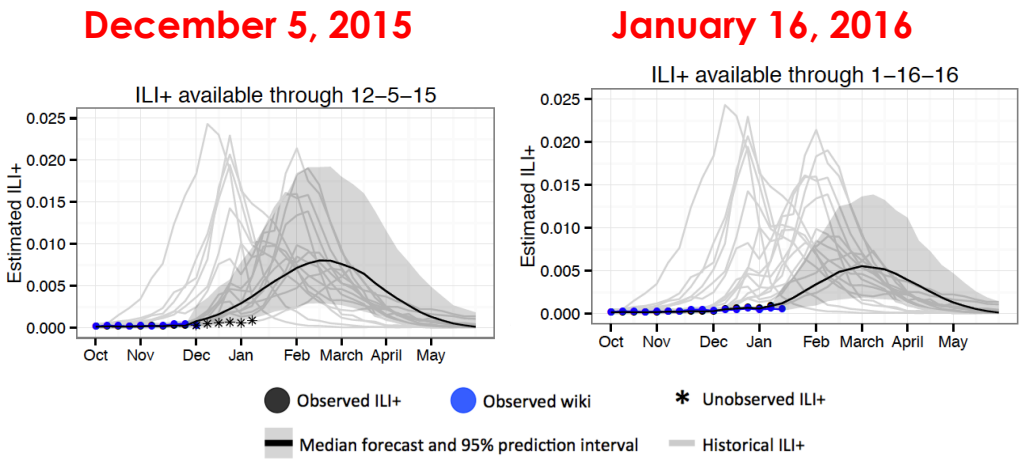
2015-2016 Influenza Forecasts

- Wikipedia traffic and clinical surveillance
- SIR (susceptible – infectious – recovered) model
  - SIR model structure
  - Historical flu seasons
  - Discrepancy
- Weekly forecasts of the 2015-2016 flu season









https://bsvgateway.org/flu

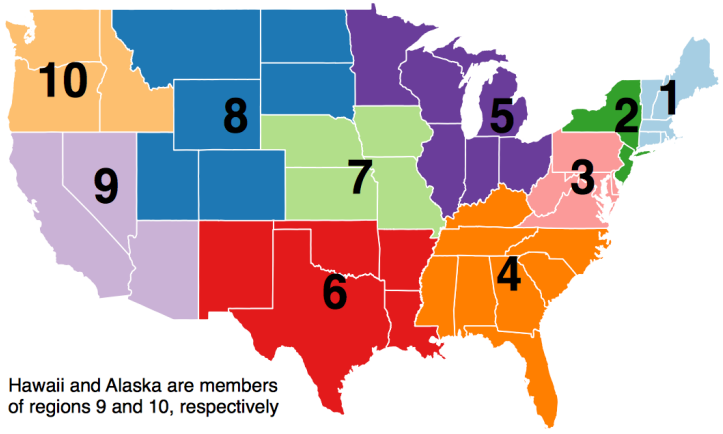
2015–2016 Influenza Forecasts

Dave Osthus, Reid Priedhorsky, Sara Del Valle

(based on ILI+ data through 2/13/16 and Wikipedia data through 2/18/16)

	Chance (%) flu season will start by ...			Chance (%) the flu season will peak ...					Chance (%) the flu season will be ...		
Region	2/7/16	2/14/16	2/21/16	before Feb.	in Feb.	in March	in April	after April	Mild	Moderate	Intense
National	38	62	76	<1	28	60	12	<1	99	1	<1
Region 1	8	21	36	2	13	57	27	1	>99	<1	<1
Region 2	45	61	68	<1	57	33	9	<1	97	3	<1
Region 3	2	11	25	<1	11	61	27	1	>99	<1	<1
Region 4	>99	>99	>99	<1	72	24	4	<1	>99	<1	<1
Region 5	6	25	42	<1	15	64	21	<1	98	2	<1
Region 6	7	29	45	<1	20	65	15	<1	>99	<1	<1
Region 7	5	19	34	<1	16	61	22	2	>99	<1	<1
Region 8	61	81	86	<1	26	59	15	<1	98	1	<1
Region 9	>99	>99	>99	<1	46	45	9	<1	96	3	<1
Region 10	>99	>99	>99	<1	13	65	22	<1	99	1	<1

Health and Human Services Regions



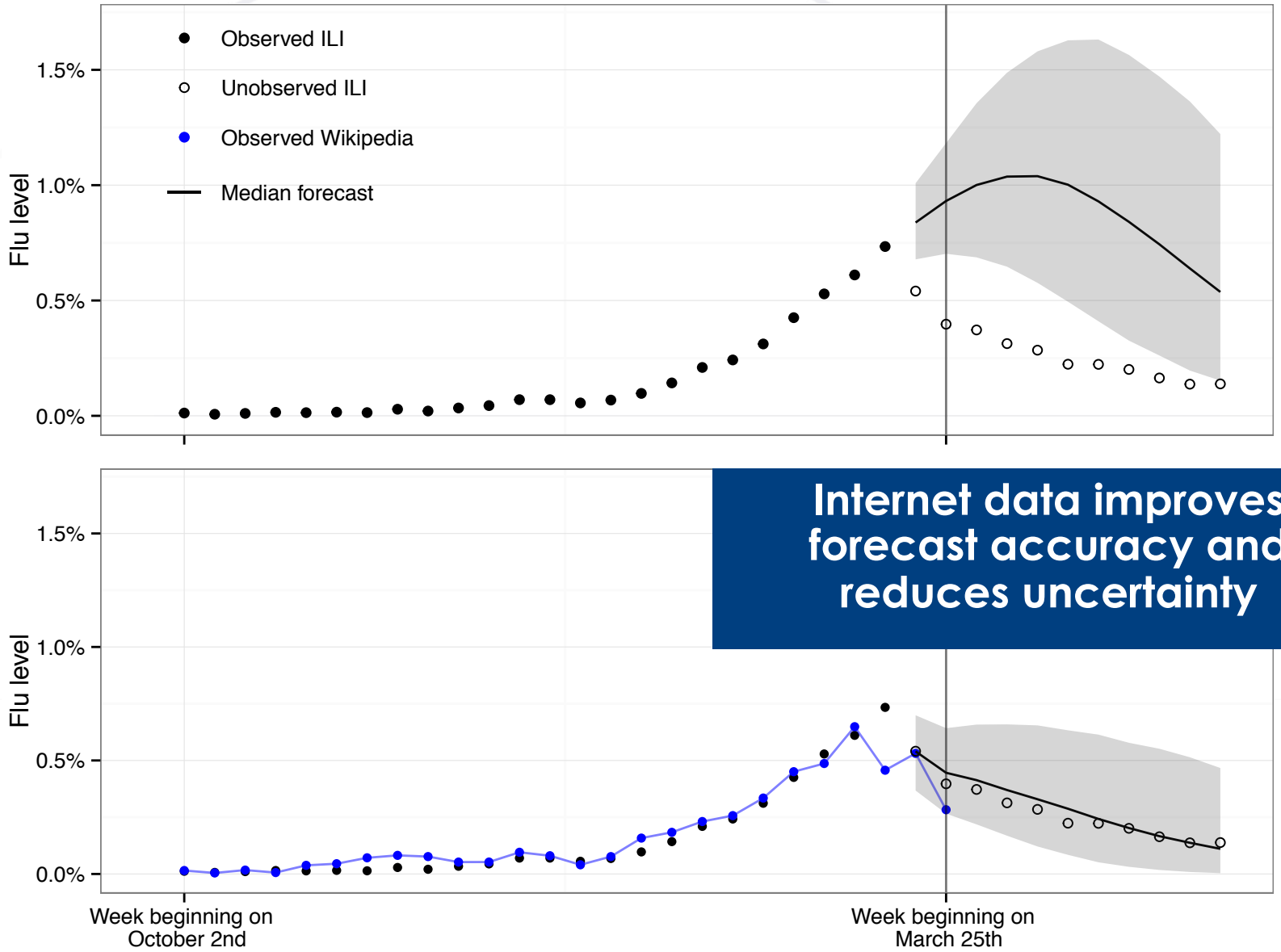
Hawaii and Alaska are members of regions 9 and 10, respectively

Previous forecasts can be accessed at
http://bsvgateway.org/flu/forecast-archives/

Model description: Our model produces probabilistic forecasts of the flu season, similar to how weather, presidential elections, and sporting event outcomes are forecasted. This means our model produces information such as, “there is an 80% chance the flu will peak after January” rather than “the flu will peak after January”. This approach explicitly acknowledges uncertainty in the data and the model. Forecasts are created at national and regional levels (see map). Our model combines three components: historical flu information, a mathematical representation of how flu spreads through a population, and data for the current flu season provided by the Centers for Disease Control and Prevention. When new data become available, the forecasts are updated. As a result, the model’s uncertainty about what may happen for the remainder of the flu season usually decreases with subsequent updates. Our forecasts will be updated every two weeks for the 2015–2016 flu season.

Special thanks to Geoffrey Fairchild, Jim Gattiker, Nicholas Generous, and Kyle Hickmann

The Value of Internet Data – 2014-2015



Quantitative Analysis of Chatter (CUAC)

- Software for acquiring, processing, and analyzing social Internet content.
  - Collects data from Twitter streaming API & Wikipedia access logs & content dumps
  - Estimates origin of locations of tweets with no geotag
  - Turns diverse kids of Internet data into hourly (daily) time series of event counts
  - Facilitates parallel access to the dataset & reasonable performance

https://github.com/reidpr/quac

Summary

- Develop new tools to extract useful information from Internet data streams
- Develop new approaches to assimilate real-time information into predictive models
- Validate approaches by forecasting events
- Our ultimate goal: develop an event forecasting system using mathematical approaches and heterogeneous data streams

Relevant papers

- Mniszewski, Del Valle, Priedhorsky, Hyman, Hickmann. Understanding the Impact of face Mask Usage through Epidemic Simulation of Large Social Networks. In Modeling and Simulation of Complex Social Systems. 2013.
- Priedhorsky, Culotta, Del Valle. Inferring the Origin Locations of Tweets with Quantitative Confidence. Proceedings of the 2014 Computer-Supported Cooperative Work Conference.
-
- Generous, Fairchild, Deshpande, Del Valle, Priedhorsky. Global Disease Monitoring and Forecasting with Wikpedia. PLoS Computational Biology 2014.
- Hickmann, Fairchild, Priedhorsky, Generous, Hyman, Deshpande, Del Valle. Forecasting the 2013-2014 Influenza Season Using Wikipedia. PLoS Computational Biology 2015.
- Fairchild, De Silva, Del Valle, Segre. Eliciting Disease Data from Wikipedia Articles. Proceedings of the 2015 ICWSM Workshop.