Digital epidemiology

Lesson 14

Michele Tizzoni
Dipartimento di Sociologia e Ricerca Sociale
Via Verdi 26, Trento
Ufficio 6, 3 piano

UNIVERSITÀ DI TRENTO

FONDAZIONE BRUNO KESSLER

C2S2 Center for Computational Social Science and Human Dynamics

Digital public health surveillance

A quick step back to lecture 2

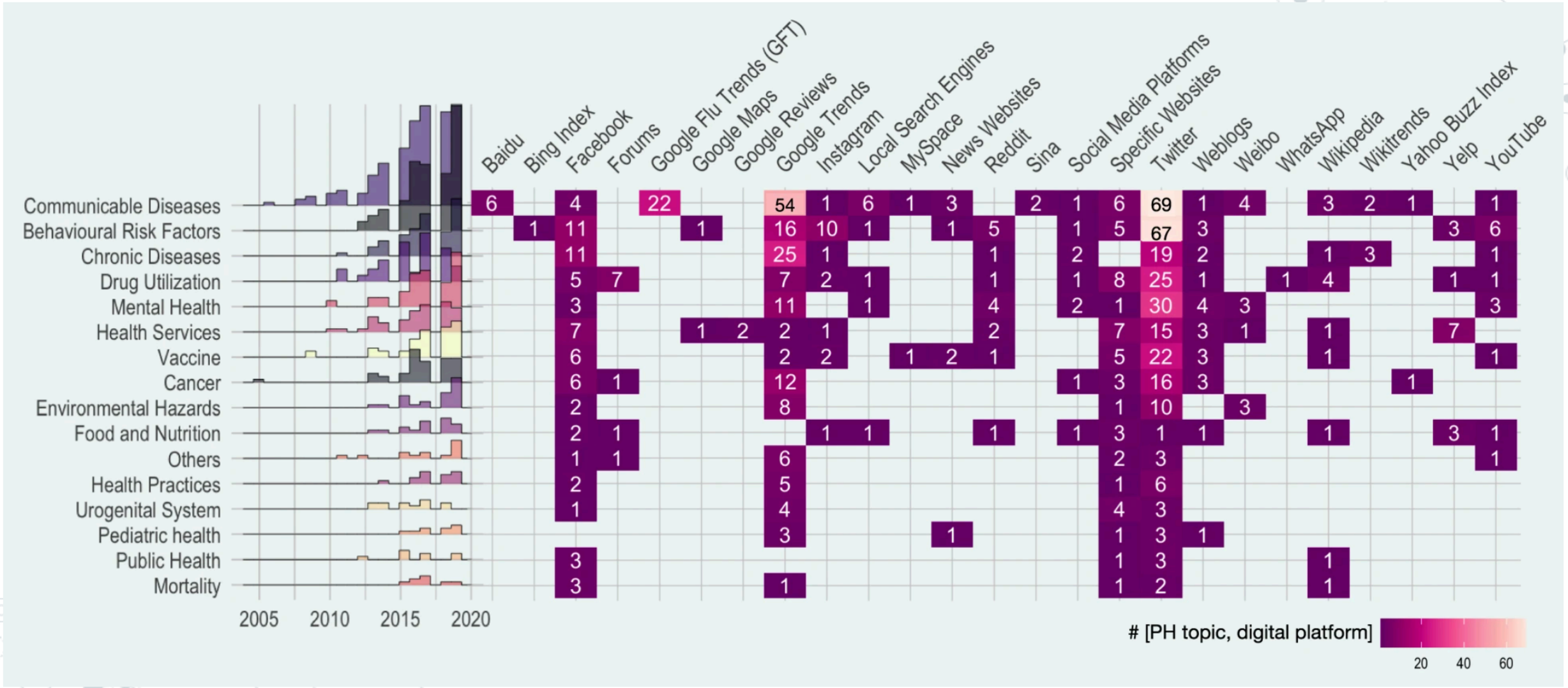
Public health surveillance

- **Population-based surveillance.** The public health systems surveys everyone in the population of interest. This is very resource-intensive.
- **Sentinel surveillance.** *Sentinels* (health facility, medical doctors) form a sentinel network and report about cases. Influenza surveillance in Europe and the US is done through sentinel doctors (1-5% of total).
  - Biases: not everyone seeks care, highly skewed towards older age groups
  - **Digital surveillance**: collect data through Web-based platforms (more in the next lectures)

Public health surveillance

- **Syndromic surveillance.** A case is identified by the symptoms the individual expresses (this is a classic example for influenza, where diagnosis is based on symptoms: influenza-like-illness, ILI).
- **Laboratory-confirmed surveillance.** A case is defined by a lab test. For instance, COVID-19 cases were generally confirmed by RT-PCR or antigen tests.
- **Digital surveillance** is typically syndromic (until now).

Digital surveillance



Hossein Abad et al. 2021

Early history

- ProMED (1994) gathers information on emerging and re-emerging disease outbreaks from across the world, and makes them accessible to anyone, most commonly through the use of email.
- Its message from December 30, 2019, titled “UNDIAGNOSED PNEUMONIA - CHINA (HUBEI): REQUEST FOR INFORMATION”, is generally regarded as the first public alert of the outbreak at the root of the COVID-19 pandemic.
- GPHIN (Global Public Health Intelligence Network) collects information from various online sources, which is then processed for epidemiological decision-making.

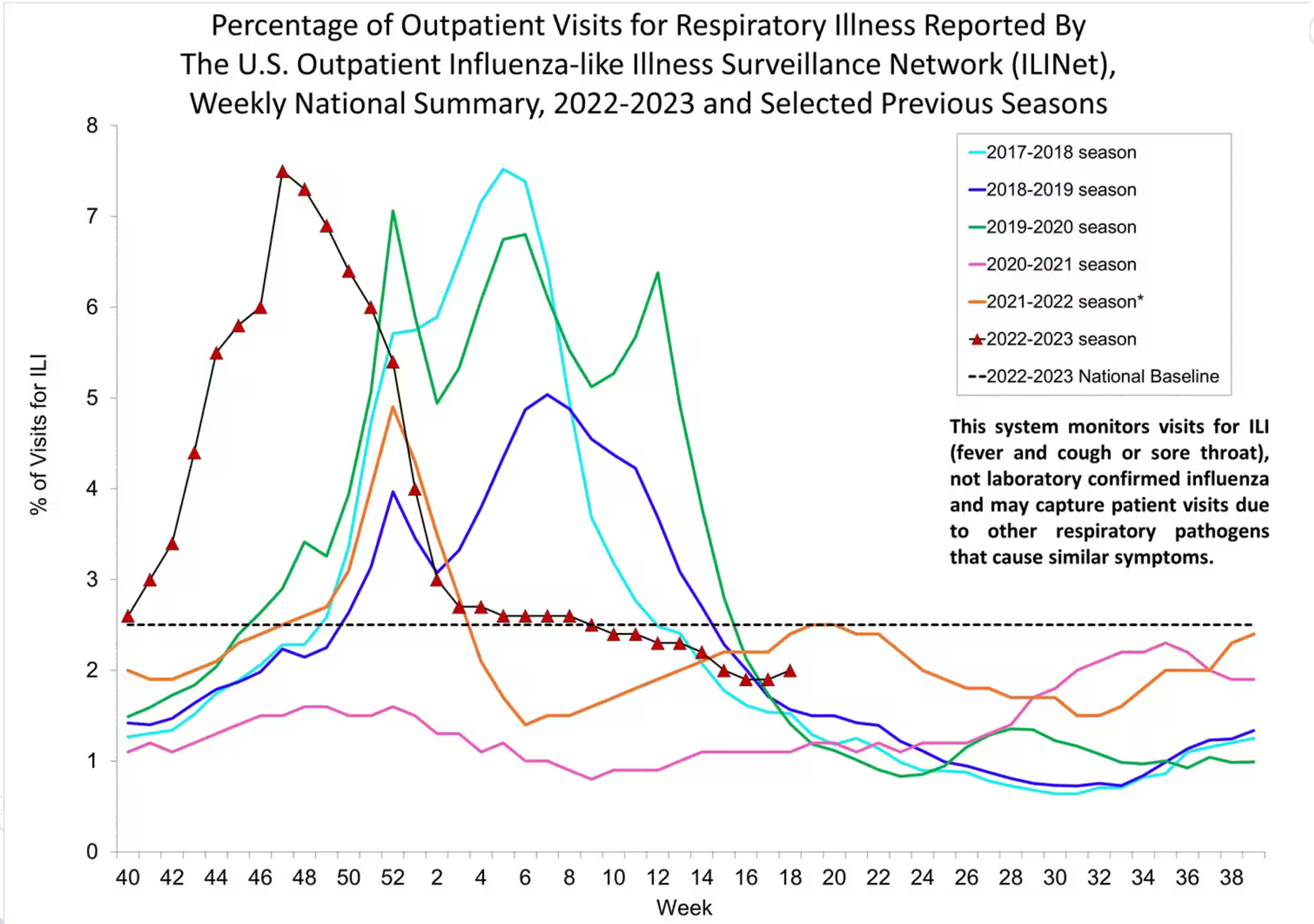
Healthmap


442 alerts for Alerts from past week
Outbreaks in Current Location
Sign up for email alerts
COVID-19 Outbreak News
COVID-19 Case Counts
174 Respiratory Alerts
COVID-19 (153), Influenza H1N1 (1), Whooping Cough (1), Hantavirus (1), SARS (4), Influenza (6), Legionnaires' (1), Pneumonia (1), Avian Influenza H5N1 (2), Diphtheria (1), Tuberculosis (1), Influenza H3N2 (1), Cold (1)
117 Animal Alerts
Canine Influenza (3), Brucellosis (2), Equine Herpes Virus (1), Avian Influenza (48),
healthmap.org
Google
Keyboard shortcuts
Map data ©2023
Terms of Us
HealthMap
About
Projects
Disease Daily
Log In

Data Sources

- **Search queries and access logs**
- Participatory surveillance
- Social media
- Mobile phones (subject of lesson 11)
- Wearable sensors
- Other data sources

Target: flu surveillance



Search queries

Google

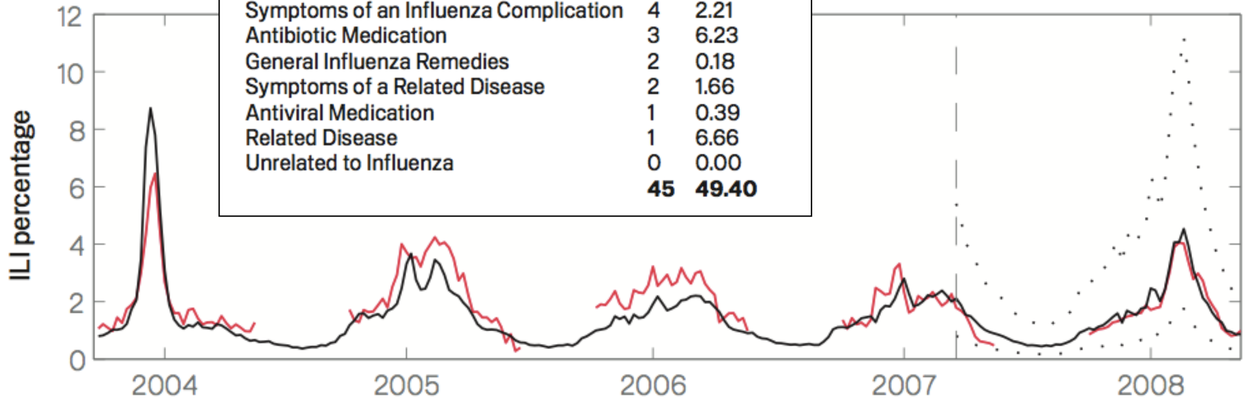
Detecting influenza epidemics using search engine query data

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

[1]Google Inc. [2]Centers for Disease Control and Prevention

Search Query Topic	Top 45 Queries N	Weighted
Influenza Complication	11	18.15
Cold/Flu Remedy	8	5.05
General Influenza Symptoms	5	2.60
Term for Influenza	4	3.74
Specific Influenza Symptom	4	2.54
Symptoms of an Influenza Complication	4	2.21
Antibiotic Medication	3	6.23
General Influenza Remedies	2	0.18
Symptoms of a Related Disease	2	1.66
Antiviral Medication	1	0.39
Related Disease	1	6.66
Unrelated to Influenza	0	0.00
	45	**49.40**

google.org/flutrends



J. Ginsberg *et al.*, Nature 457, 1012 (2009)

- Ideas proposed by Eysenbach et al. (2006) and Polgreen et al. (2008)
- Using search query data to predict flu activity in the USA
- Google Flu Trends is launched in 2009

http://www.google.org/flutrends

google.org Flu Trends

Language: English (United States)

Google.org home

Dengue Trends

Flu Trends

Home

Select country/region

How does this work?

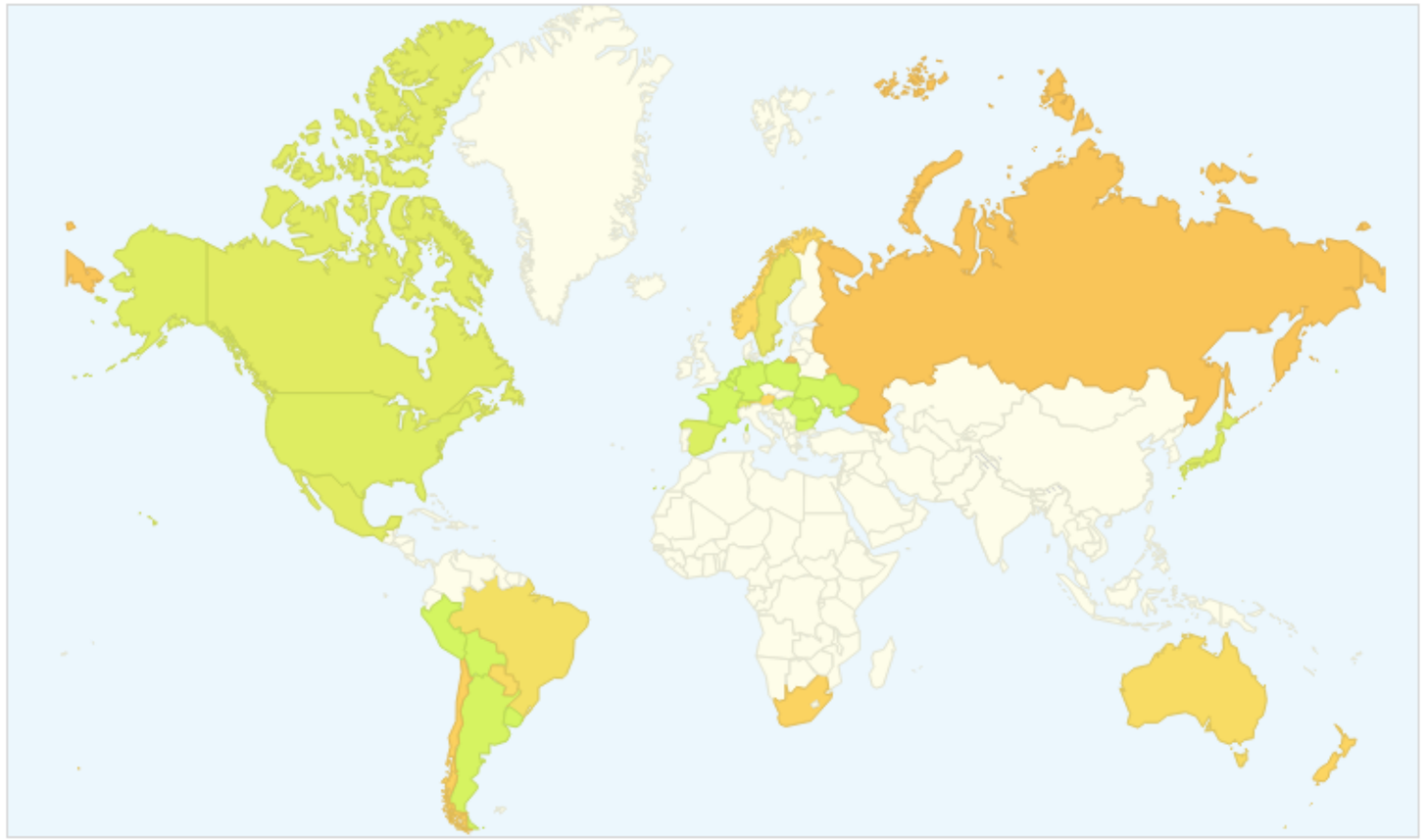
FAQ

Flu activity

- Intense
- High
- Moderate
- Low
- Minimal

Explore flu trends around the world

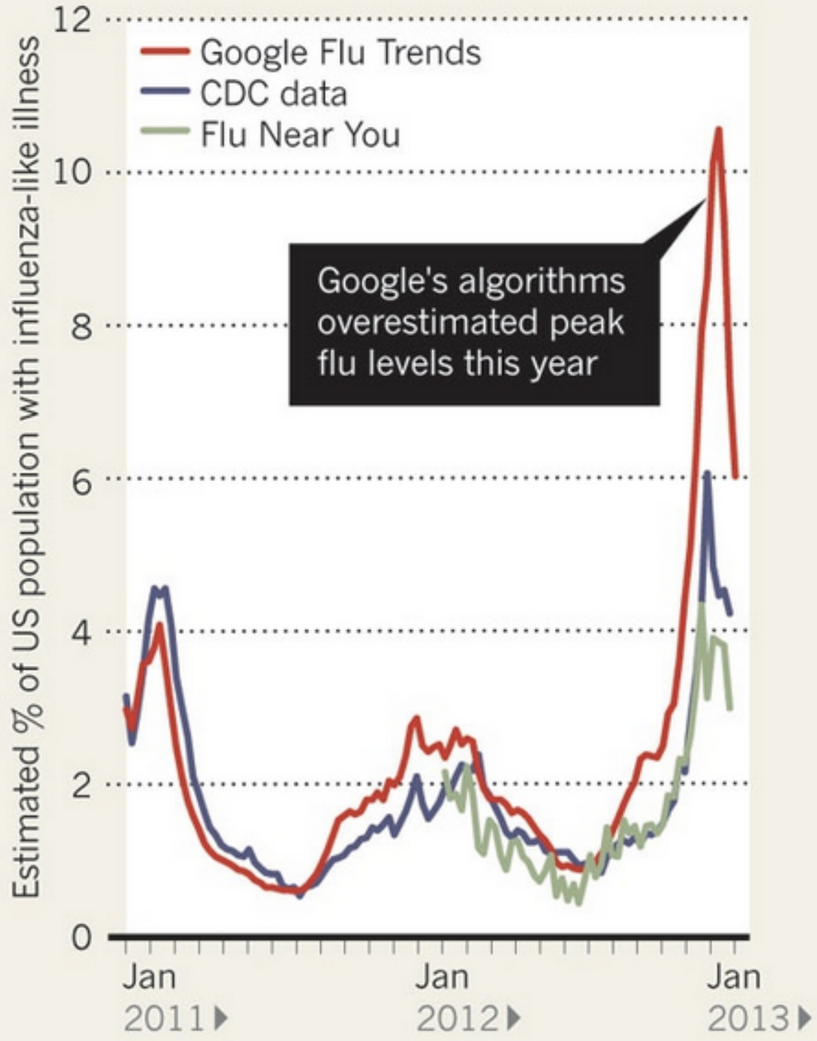
We've found that certain search terms are good indicators of flu activity. Google Flu Trends uses aggregated Google search data to estimate flu activity. Learn more »



Download world flu activity data - Animated flu trends for Google Earth - Compare flu trends across regions in Public Data Explorer

FEVER PEAKS

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



nature International weekly journal of science

NATURE | NEWS

عربي

When Google got flu wrong

US outbreak foxes a leading web-based method for tracking seasonal flu.

Declan Butler

Science 14 March 2014:
Vol. 343 no. 6176 pp. 1203–1205
DOI: 10.1126/science.1248506

‹ Prev |

POLICY FORUM

BIG DATA

The Parable of Google Flu: Traps in Big Data Analysis

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

± Author Affiliations

*Corresponding author. E-mail: d.lazer@neu.edu.

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

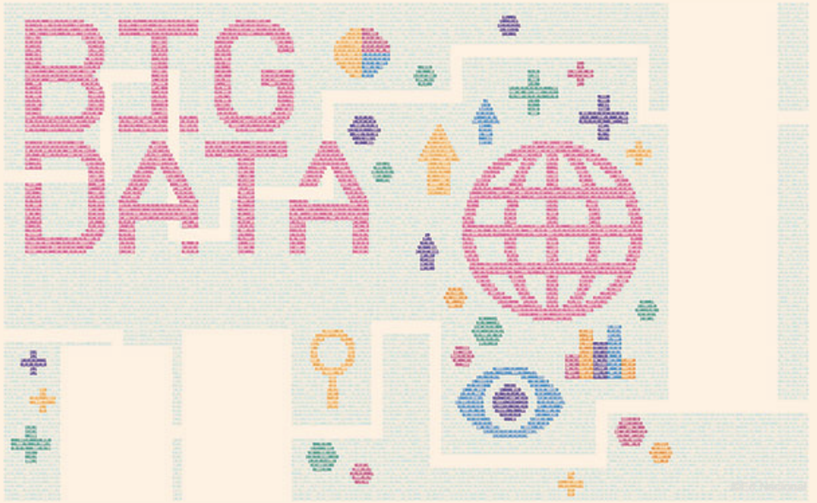
FT Magazine

March 28, 2014 11:38 am

Big data: are we making a big mistake?

By Tim Harford

Big data is a vague term for a massive phenomenon that has rapidly become an obsession with entrepreneurs, scientists, governments and the media



Wikipedia pageview data

OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

Wikipedia Usage Estimates Prevalence of Influenza-Like Illness in the United States in Near Real-Time

David J. McIver*, John S. Brownstein

Boston Children's Hospital, Harvard Medical School, Boston, Massachusetts, United States of America

Abstract

Circulating levels of both seasonal and pandemic influenza require constant surveillance to ensure the health and safety of the population. While up-to-date information is critical, traditional surveillance systems can have data availability lags of up to two weeks. We introduce a novel method of estimating, in near-real time, the level of influenza-like illness (ILI) in the United States (US) by monitoring the rate of particular Wikipedia article views on a daily basis. We calculated the number of times certain influenza- or health-related Wikipedia articles were accessed each day between December 2007 and August 2013 and compared these data to official ILI activity levels provided by the Centers for Disease Control and Prevention (CDC). We developed a Poisson model that accurately estimates the level of ILI activity in the American population, up to two weeks ahead of the CDC, with an absolute average difference between the two estimates of just 0.27% over 294 weeks of data. Wikipedia-derived ILI models performed well through both abnormally high media coverage events (such as during the 2009 H1N1 pandemic) as well as unusually severe influenza seasons (such as the 2012–2013 influenza season). Wikipedia usage accurately estimated the week of peak ILI activity 17% more often than Google Flu Trends data and was often more accurate in its measure of ILI intensity. With further study, this method could potentially be implemented for continuous monitoring of ILI activity in the US and to provide support for traditional influenza surveillance tools.

Citation: McIver DJ, Brownstein JS (2014) Wikipedia Usage Estimates Prevalence of Influenza-Like Illness in the United States in Near Real-Time. PLoS Comput Biol 10(4): e1003581. doi:10.1371/journal.pcbi.1003581

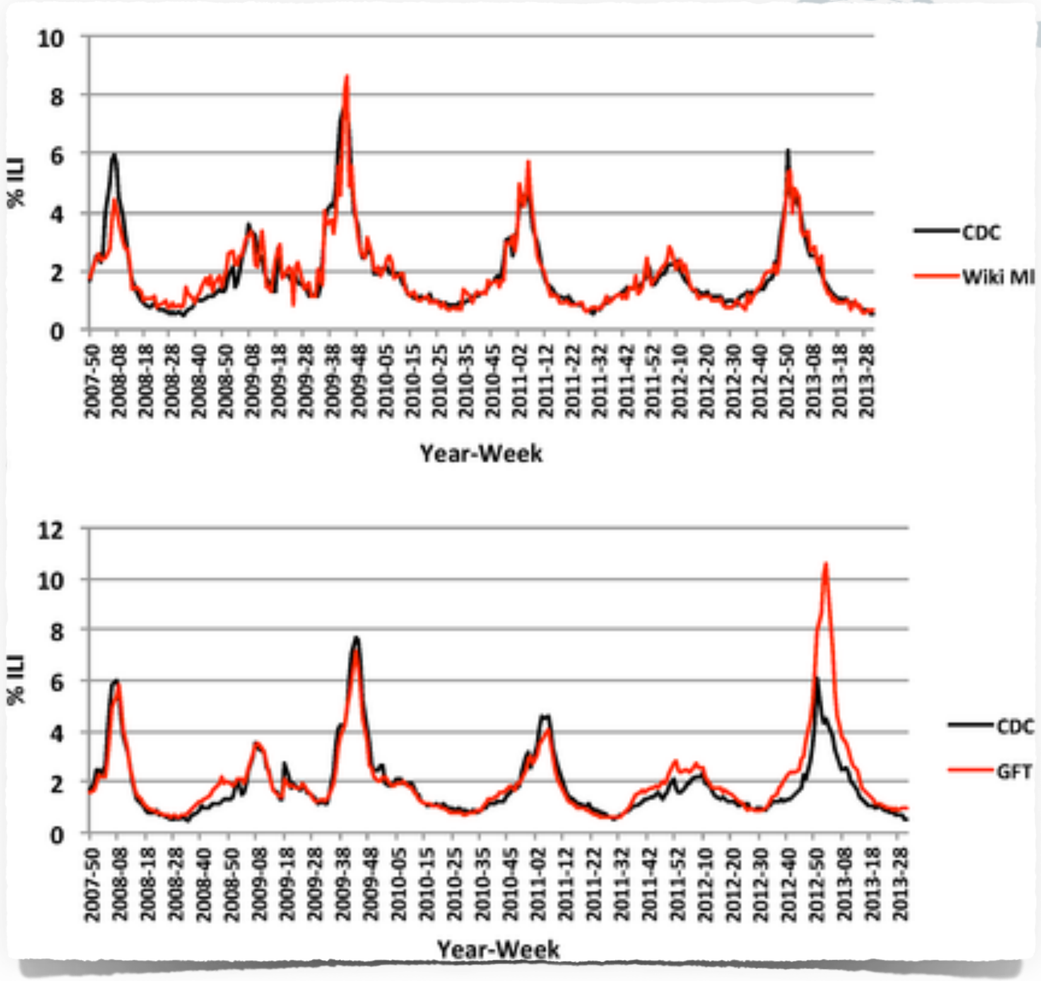
Editor: Marcel Salathé, Pennsylvania State University, United States of America

Received December 20, 2013; **Accepted** March 11, 2014; **Published** April 17, 2014

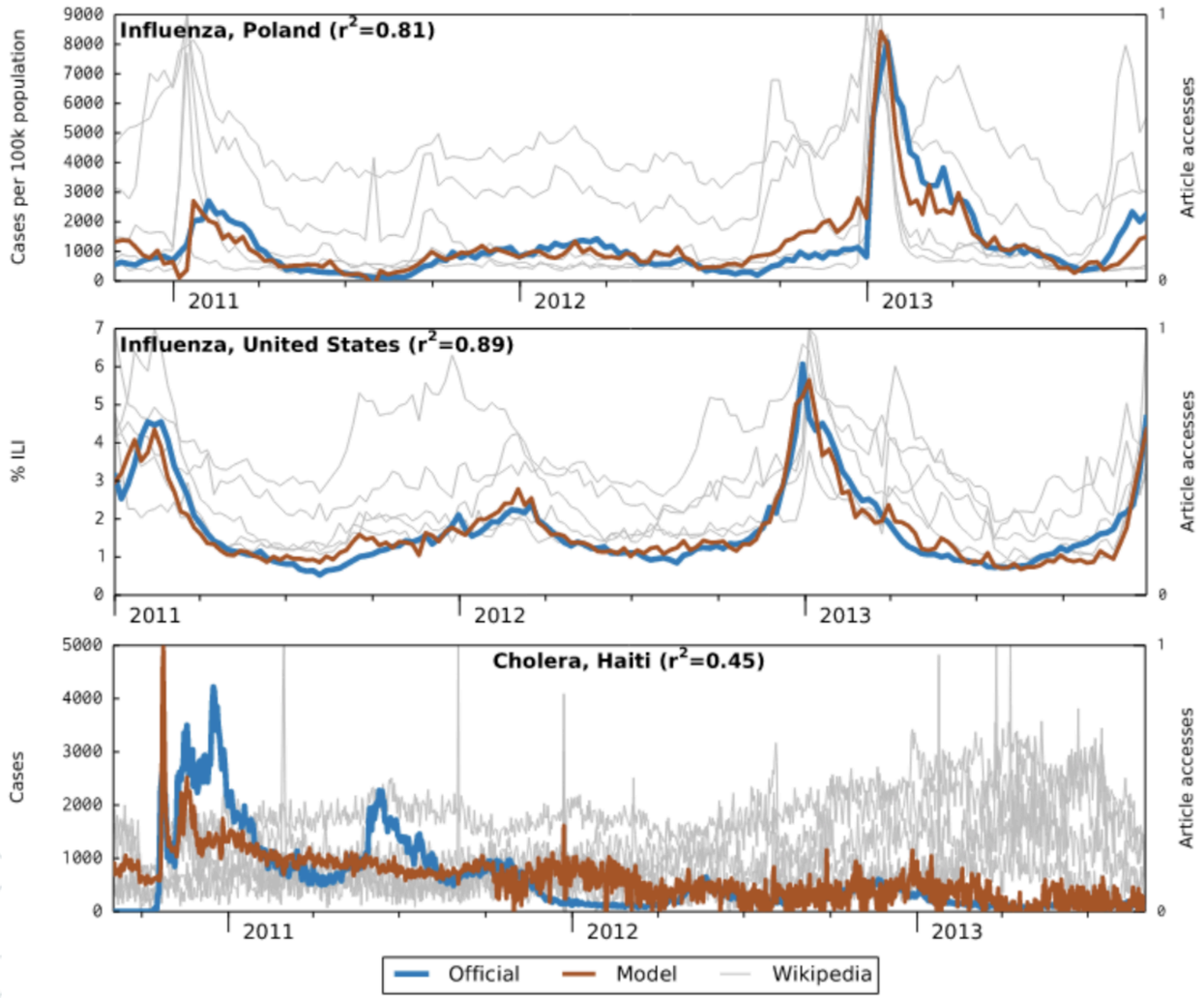
Funding: This work was funded by the National Institutes of Health and National Library of Medicine 1R01LM010812-03. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

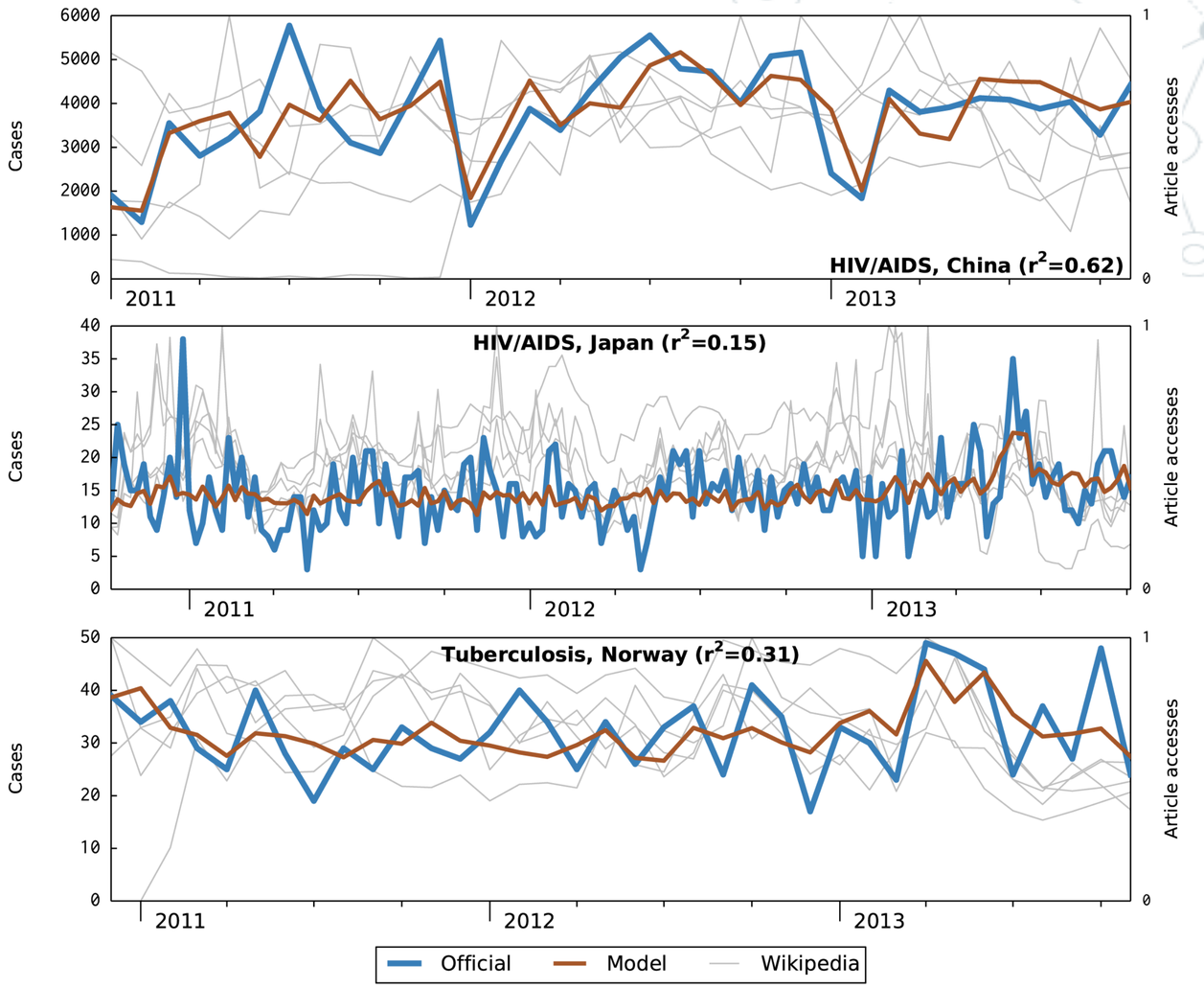
Competing Interests: The authors have declared that no competing interests exist.

* E-mail: david.mciver@childrens.harvard.edu



Wikipedia pageview data





Generous et al. PLOS Comp Bio 2014

Challenges

- The people getting the disease must be able to access these online resources.
- The search terms that made it into the original GFT model may have shown a strong seasonal pattern, coinciding with flu seasonality but unable to “see" unusual patterns such as the 2009 H1N1 pandemic.
- The relative search volumes themselves may be affected by the Google search algorithm and its results.
- Internet searches or Wikipedia page views can be driven by media interests more than actual disease prevalence
- A way to improve the models is to leverage additional, different data sources, for example from social media.

Media coverage and attention

RESEARCH ARTICLE

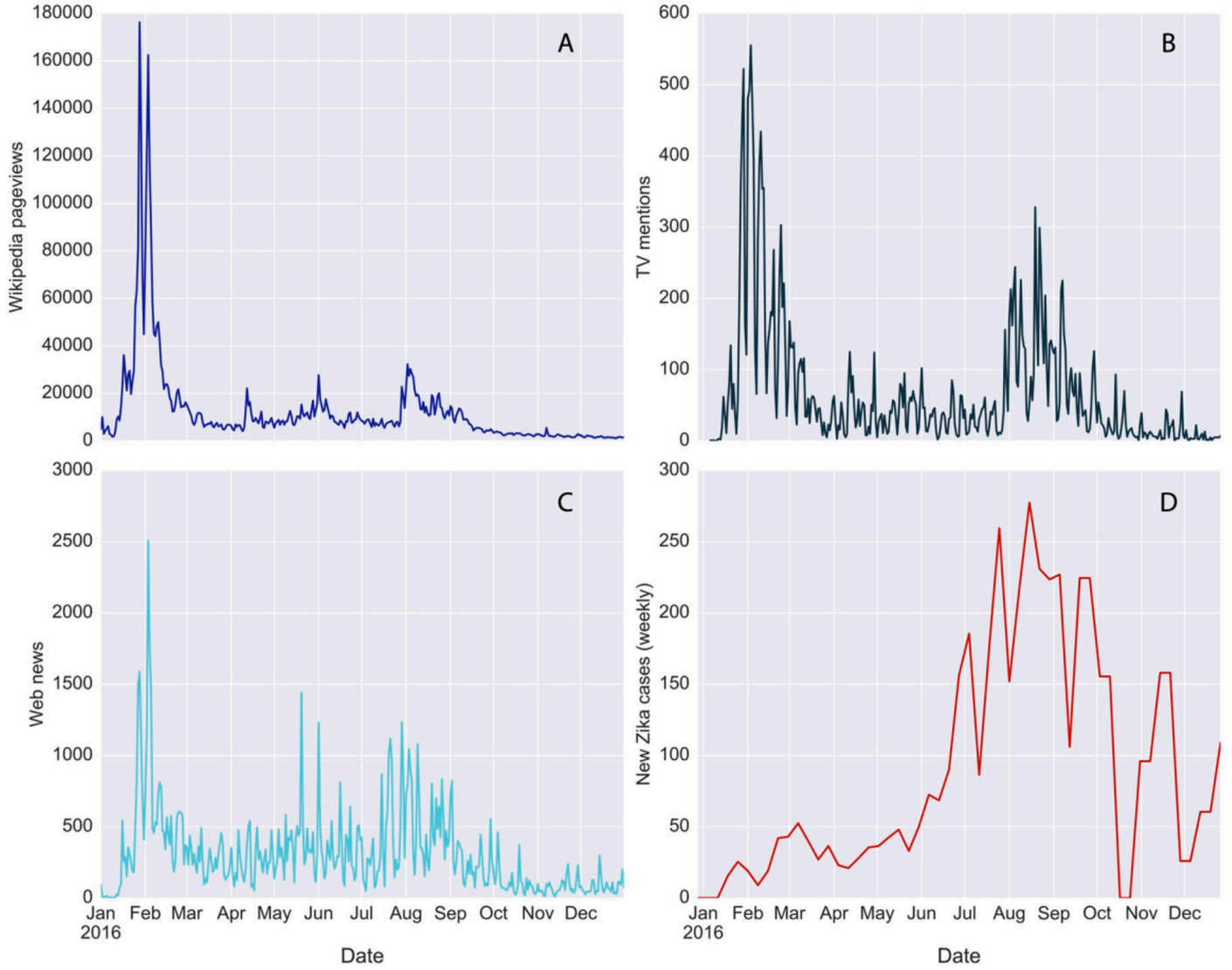
The impact of news exposure on collective attention in the United States during the 2016 Zika epidemic

Michele Tizzoni*, André Panisson, Daniela Paolotti, Ciro Cattuto

ISI Foundation, Turin, Italy

* michele.tizzoni@isi.it

- Collective attention during the 2016 Zika epidemic was mainly driven by media coverage
- **Attention hotspots** correspond to disease hotspots



Media coverage and attention

RESEARCH ARTICLE

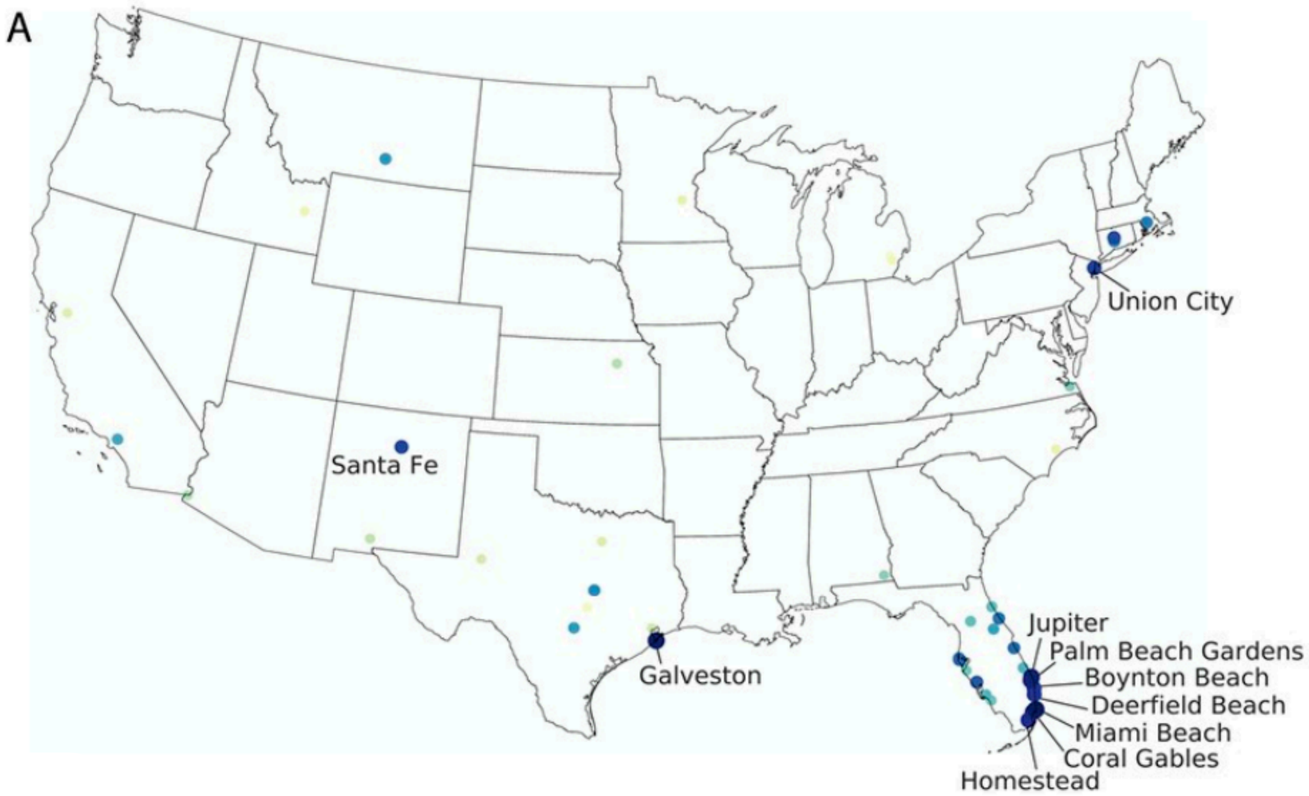
The impact of news exposure on collective attention in the United States during the 2016 Zika epidemic

Michele Tizzoni*, André Panisson, Daniela Paolotti, Ciro Cattuto

ISI Foundation, Turin, Italy

* michele.tizzoni@isi.it

- Collective attention during the 2016 Zika epidemic was mainly driven by media coverage
- **Attention hotspots** correspond to disease hotspots



COVID-19

- High correlation between search query volumes related to “smell”, “loss of smell” and the number of COVID-19 reported cases.
- Many other studies explored the connection between search queries logs and COVID-19 incidence.

ORIGINAL ARTICLE

Use of Google Trends to investigate loss-of-smell–related searches during the COVID-19 outbreak

Abigail Walker, MRSC, MSc, Claire Hopkins, FRCS(ORLHNS) and Pavol Surda, MD

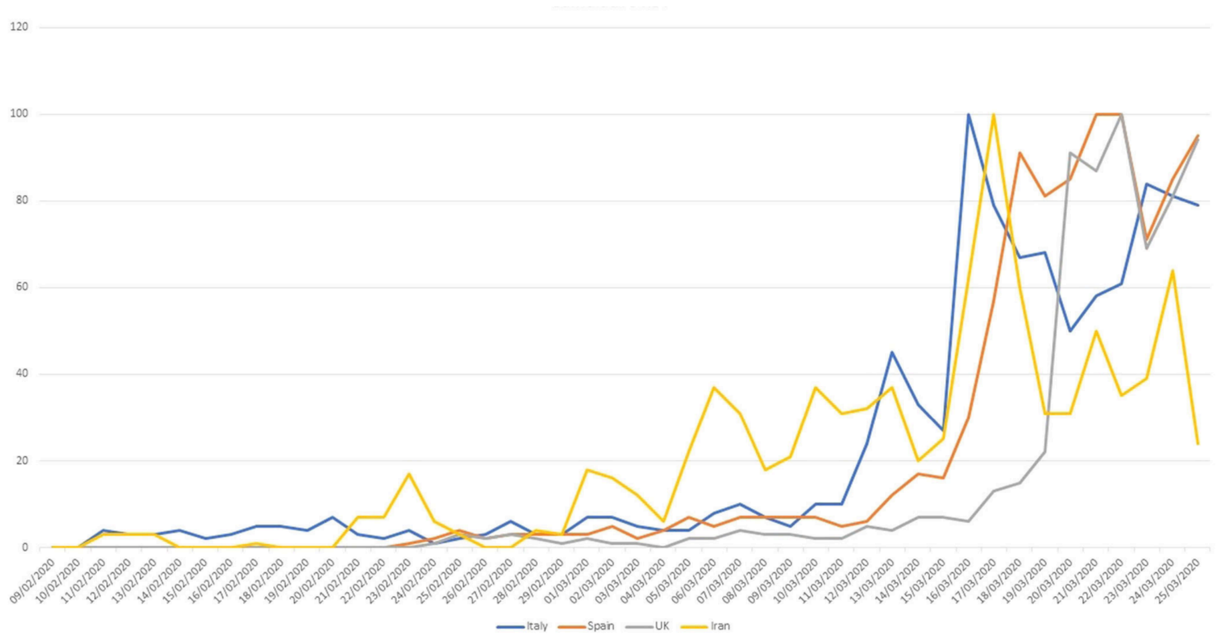


FIGURE 1. Cumulative trends of RSVs in Italy, Spain, and UK between February 3 and March 25, 2020. COVID-19 = coronavirus 2019; RSV = relative search volume.

Beyond infectious diseases

How Search Engine Data Enhance the Understanding of Determinants of Suicide in India and Inform Prevention: Observational Study

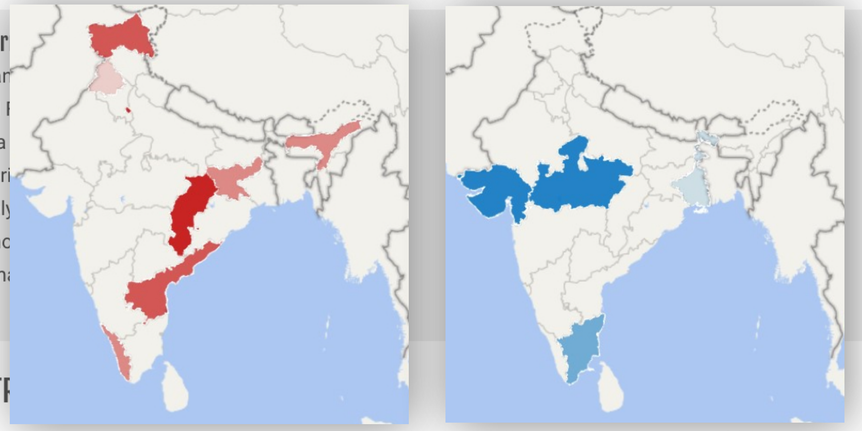
Natalia Adler[1], MA ; Ciro Cattuto[2], PhD ; Kyriaki Kalimeri[2], PhD ; Daniela Paolotti[2], PhD ; Michele Tizzoni[2], PhD ; Stefaan Verhulst[3], MA ; Elad Yom-Tov[4], PhD ; Andrew Young[3], MA

[1]United Nations International Children's Emergency Fund (UNICEF), New York, NY, United States

[2]ISI Foundation, Torino, Italy

[3]The Governance Lab, New York University, New York, NY, United States

[4]Microsoft Research, Herzeliya, Israel



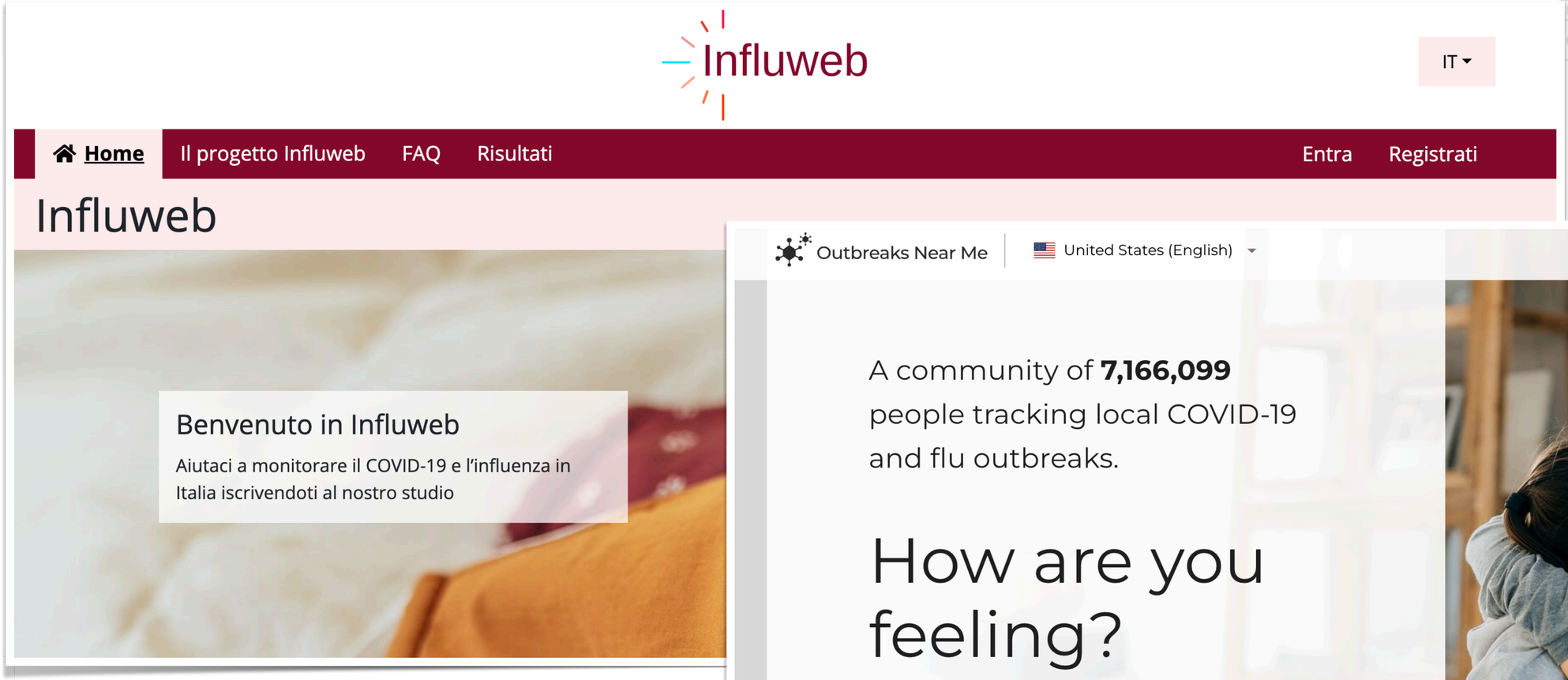
Citation
Please cite as:
Adler N, Cattuto C, Kalimeri K, Paolotti D, Tizzoni M, Verhulst S, Yom-Tov E, Young A
How Search Engine Data Enhance the Understanding of Determinants of Suicide in India and Inform Prevention: Observational Study
J Med Internet Res 2019;21(1):e10179
DOI: 10.2196/jmir.10179
PMID: 30609976

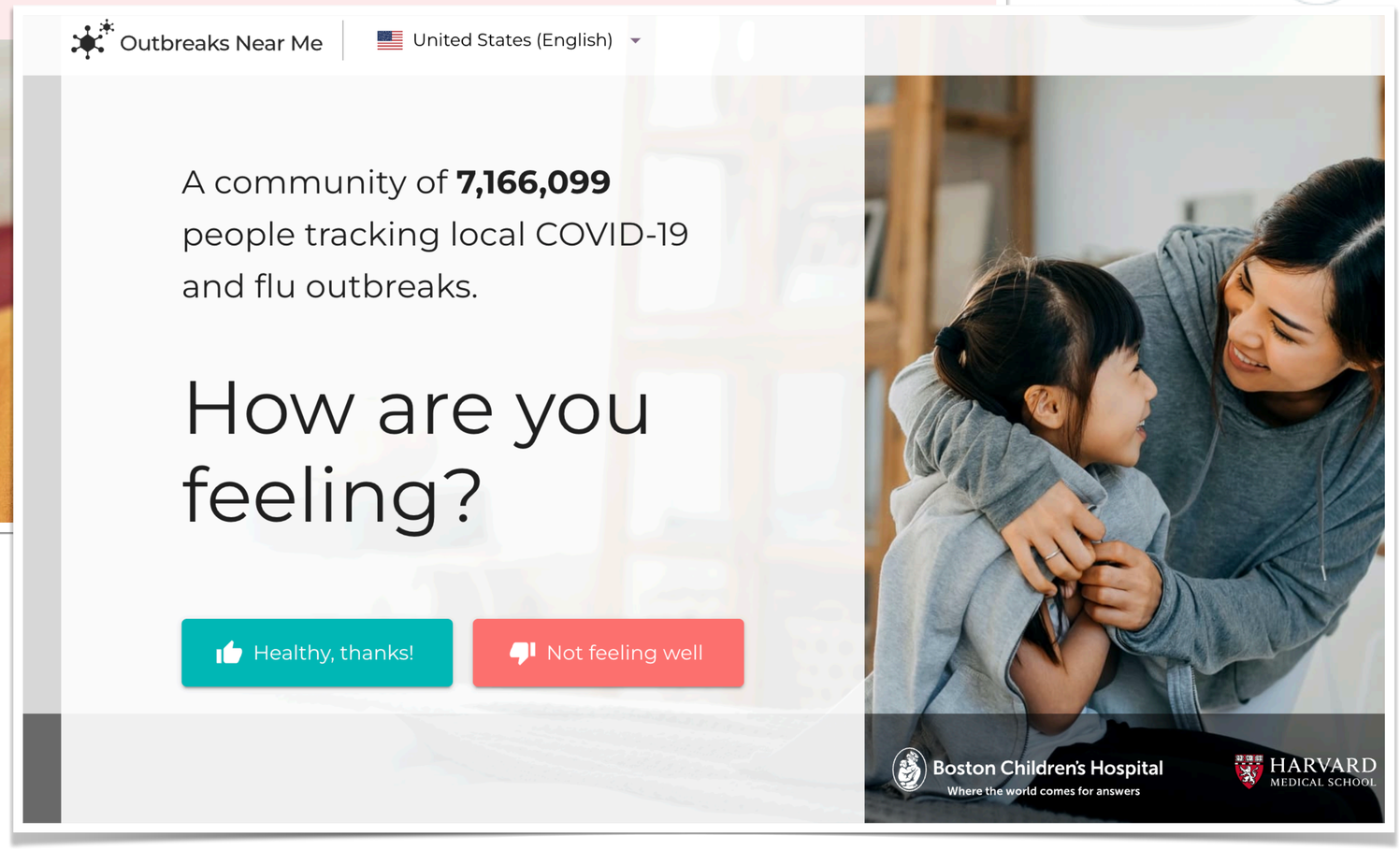
Background: India is home to 20% of the world's suicide deaths. Although statistics regarding suicide in India are distressingly high, data and cultural issues likely contribute to a widespread underreporting of the problem. Social stigma and only recent

Data Sources

- Search queries and access logs
- **Participatory surveillance**
- Social media
- Mobile phones
- Wearable sensors
- Other data sources

Participatory surveillance

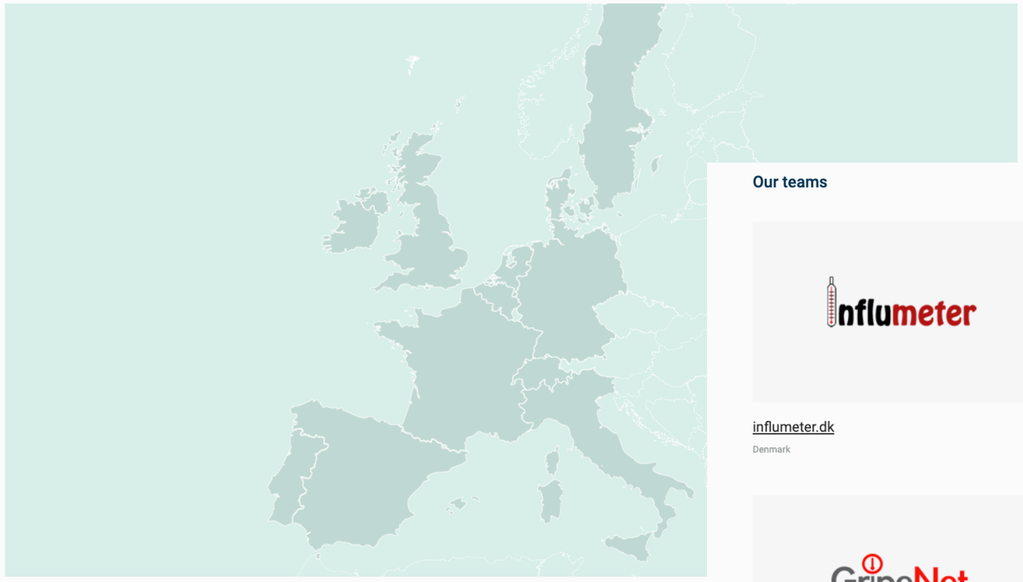




Participatory surveillance

influenzanet

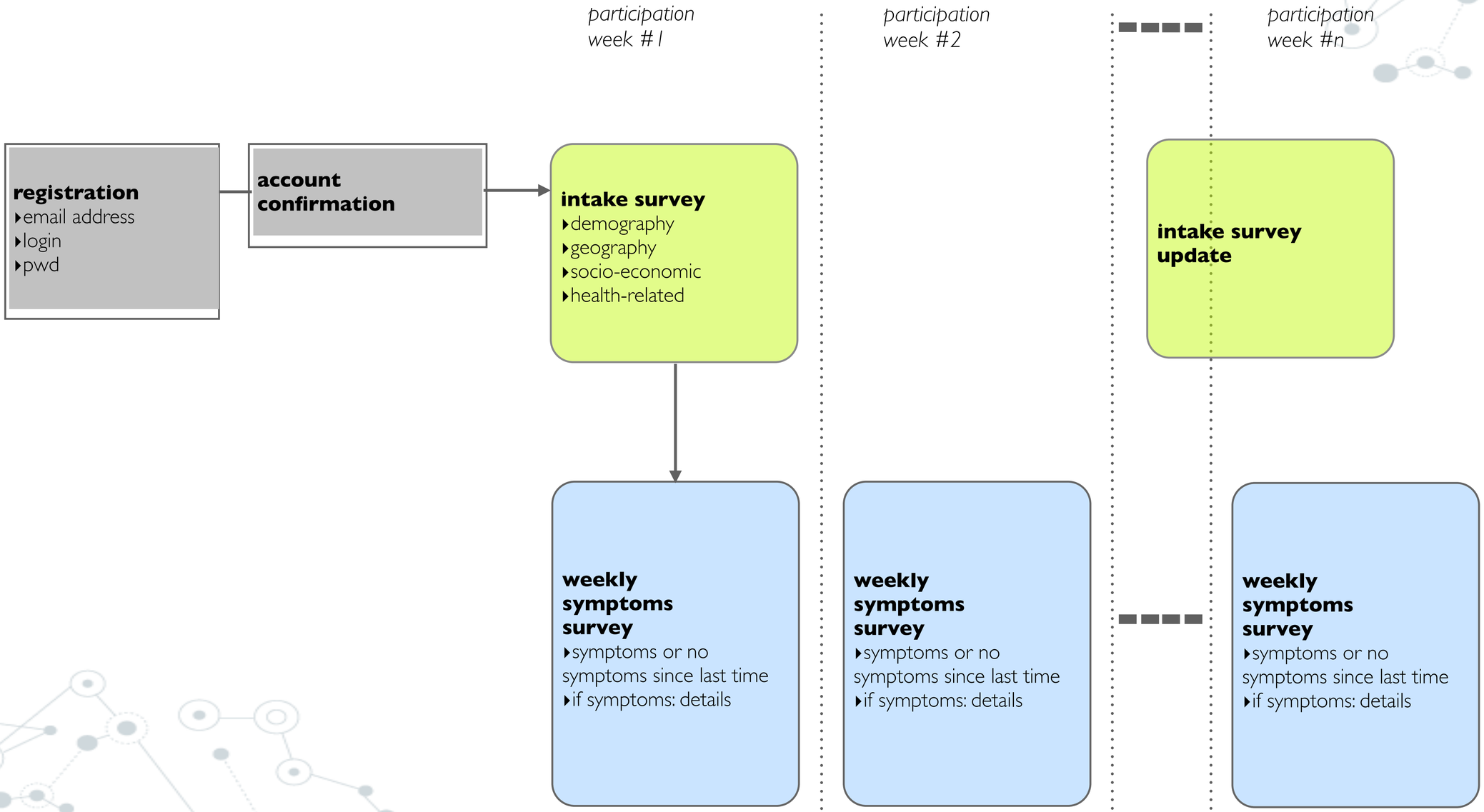
Explore our data | The project | About our data | Publications



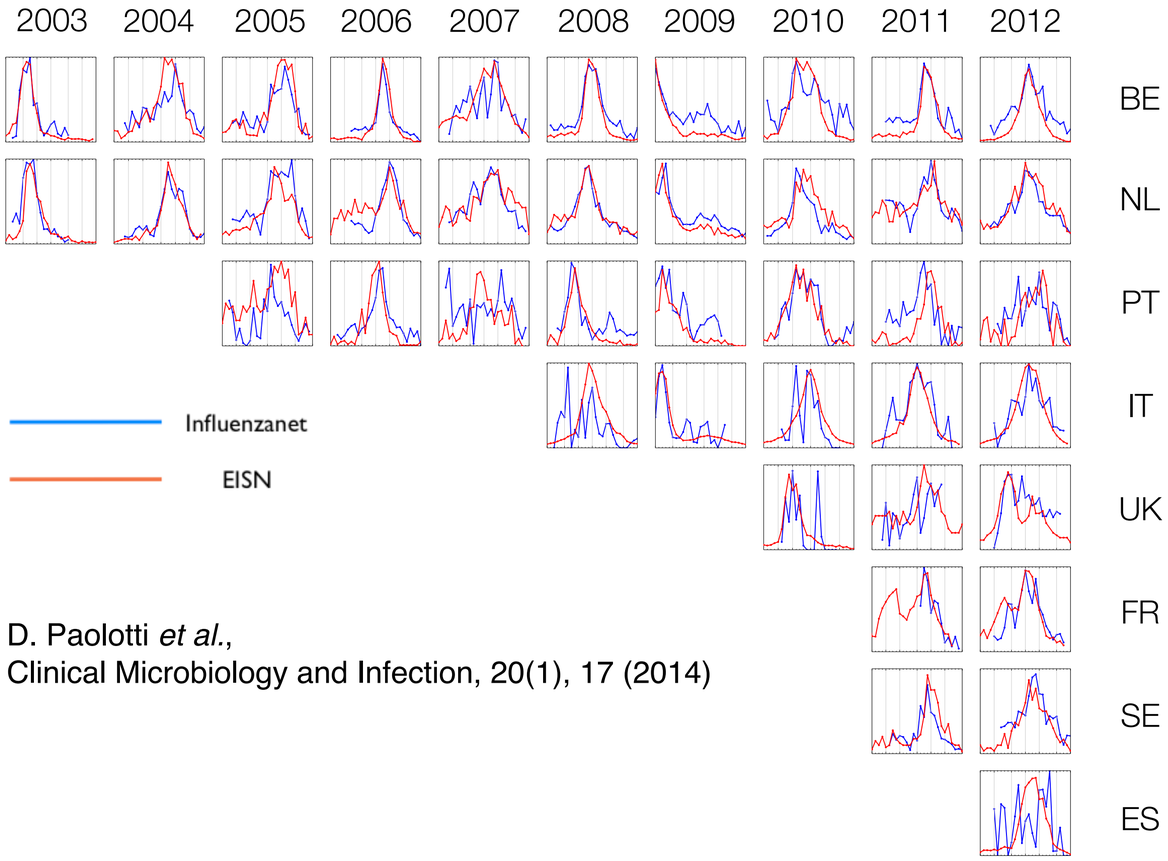
Our teams

influmeter.dk Denmark	covidmeter Denmark	flusurvey.net United Kingdom	Infectieradar Netherlands	GrippeWeb Germany
gripenet.pt Portugal	flusurvey.ie Ireland	gripenet.es Spain	grippenet.ch Switzerland	influweb.org Italy
grippenet.fr France	Infectieradar Belgium			

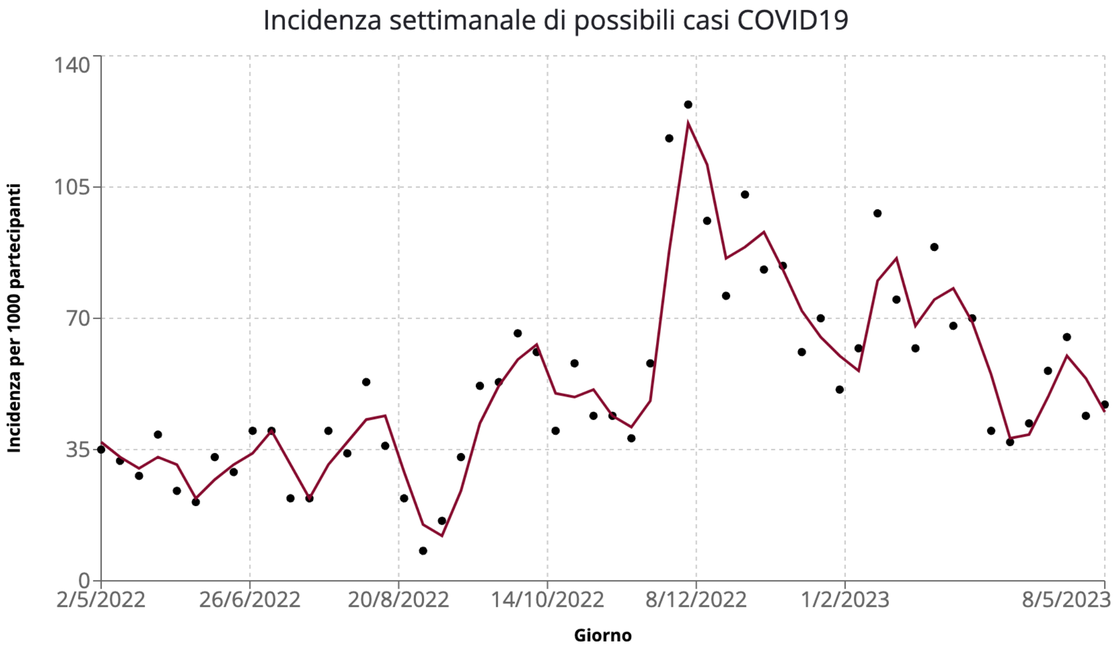
Participatory surveillance



Participatory surveillance

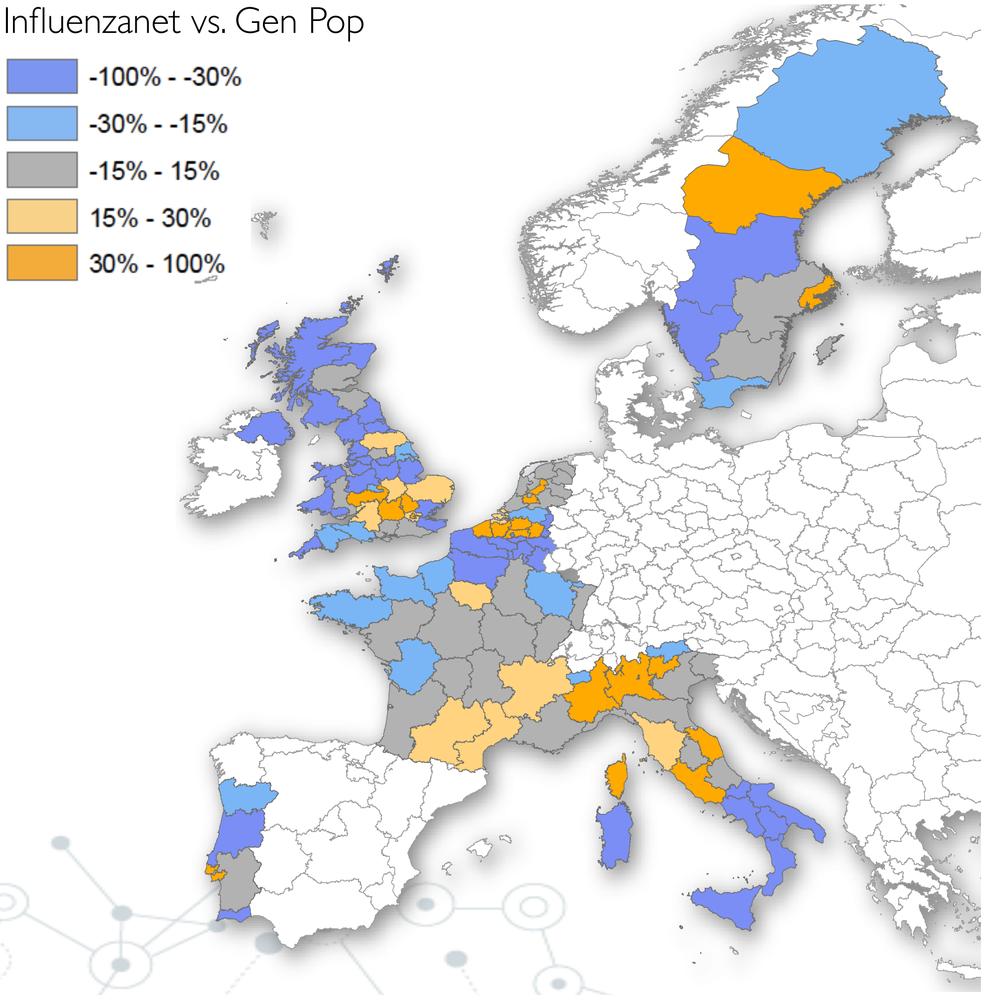


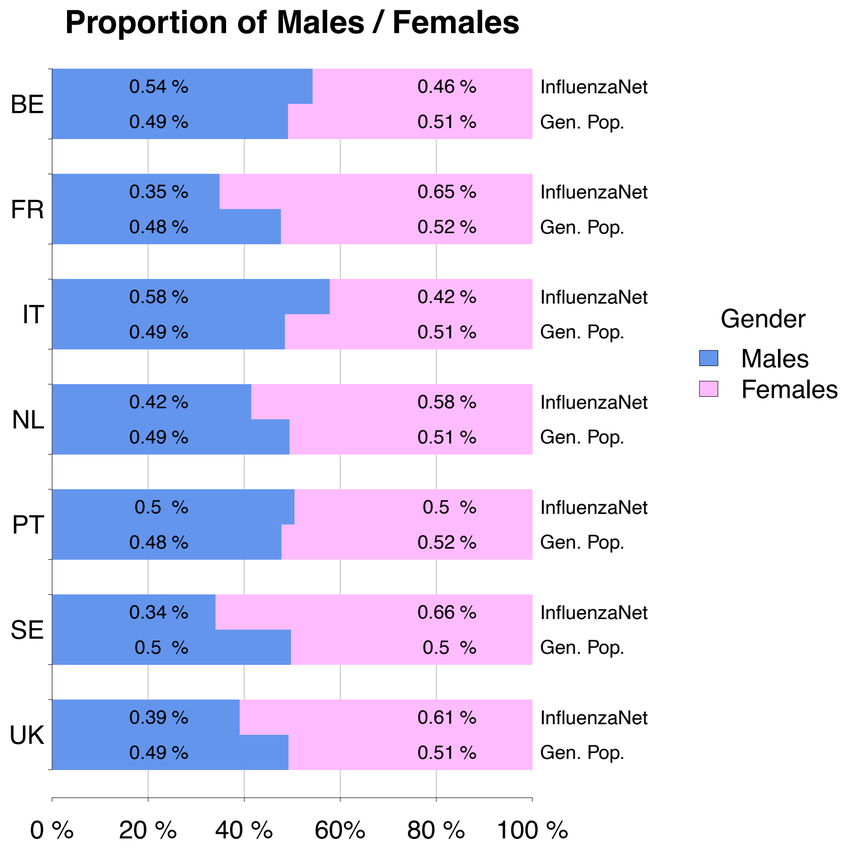
D. Paolotti *et al.*,
Clinical Microbiology and Infection, 20(1), 17 (2014)



Challenges

relative difference of
NUTS2 distribution,
Influenzanet vs. Gen Pop





Belgium
Italy
Portugal
Netherlands
UK
Sweden
France

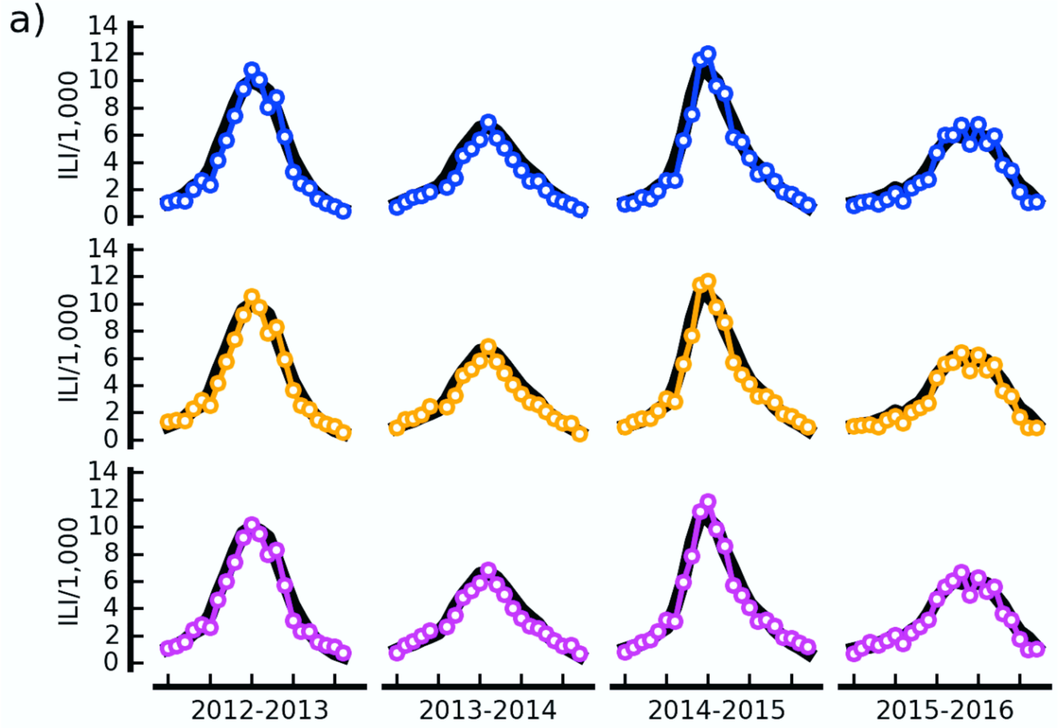
Participatory surveillance

Using Participatory Web-based Surveillance Data to Improve Seasonal Influenza Forecasting in Italy

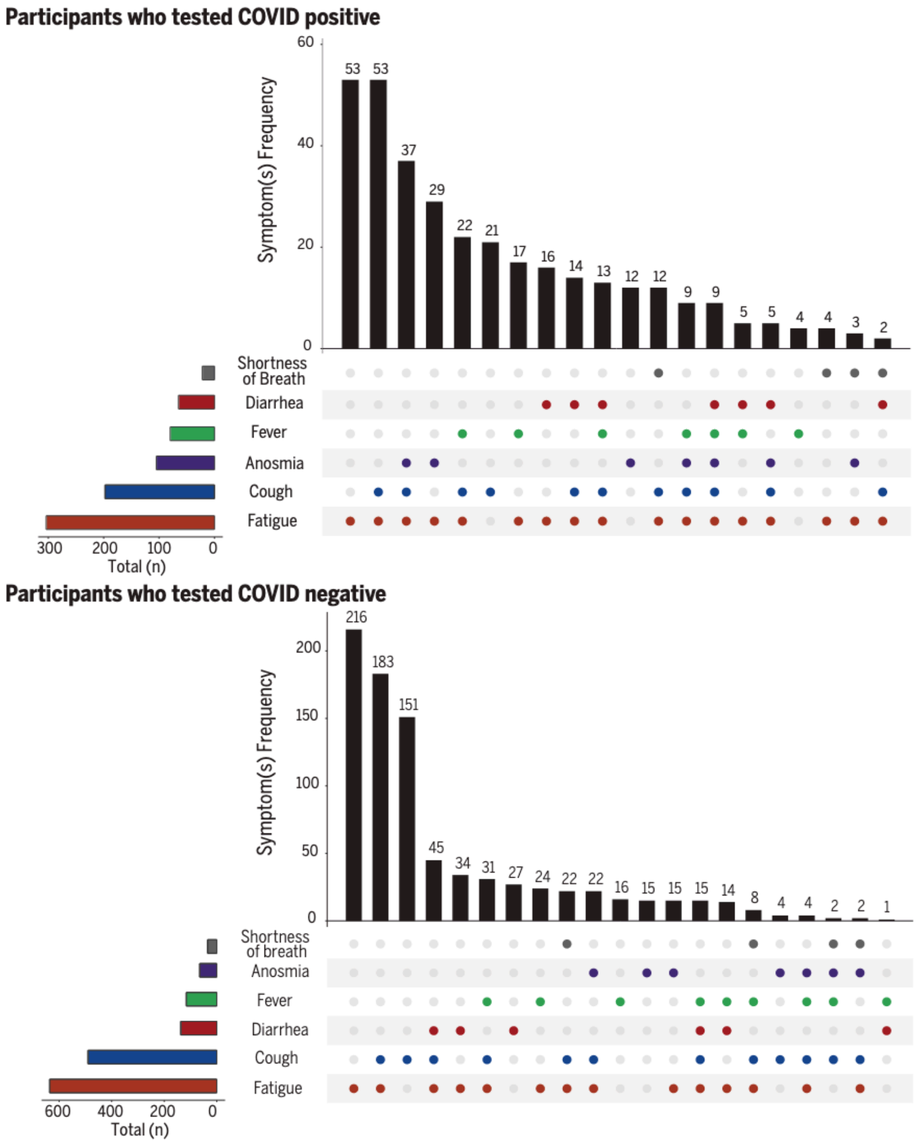
Daniela Perrotta
ISI Foundation
Turin, Italy
daniela.perrotta@isi.it

Michele Tizzoni
ISI Foundation
Turin, Italy
michele.tizzoni@isi.it

Daniela Paolotti
ISI Foundation
Turin, Italy
daniela.paolotti@isi.it



COVID-19



CORONAVIRUS

Rapid implementation of mobile technology for real-time epidemiology of COVID-19

David A. Drew[1]*, Long H. Nguyen[1]*, Claire J. Steves[2,3], Cristina Menni[2], Maxim Freydin[2], Thomas Varsavsky[4], Carole H. Sudre[4], M. Jorge Cardoso[4], Sebastien Ourselin[4], Jonathan Wolf[5], Tim D. Spector[2,5]†, Andrew T. Chan[1,6]†‡, COPE Consortium§

The rapid pace of the coronavirus disease 2019 (COVID-19) pandemic caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) presents challenges to the robust collection of population-scale data to address this global health crisis. We established the COronavirus Pandemic Epidemiology (COPE) Consortium to unite scientists with expertise in big data research and epidemiology to develop the COVID Symptom Study, previously known as the COVID Symptom Tracker, mobile application. This application—which offers data on risk factors, predictive symptoms, clinical outcomes, and geographical hotspots—was launched in the United Kingdom on 24 March 2020 and the United States on 29 March 2020 and has garnered more than 2.8 million users as of 2 May 2020. Our initiative offers a proof of concept for the repurposing of existing approaches to enable rapidly scalable epidemiologic data collection and analysis, which is critical for a data-driven response to this public health challenge.

Data Sources

- Search queries and access logs
- Participatory surveillance
- **Social media**
- Mobile phones
- Wearable sensors
- Other data sources

Social media

- Social media provides an important data source for digital public health surveillance.
- Twitter data has been relatively easy to acquire until recently, becoming a key data source for research.
- Early explorations into using Twitter as a digital public health surveillance data source focused on influenza-like illnesses (ILI),

Twitter

2010 2nd International Workshop on Cognitive Information Processing

Tracking the flu pandemic by monitoring the Social Web

Vasileios Lampos, Nello Cristianini

Intelligent Systems Laboratory
Faculty of Engineering
University of Bristol, UK
{*bill.lampos*, *nello.cristianini*}*@bristol.ac.uk*

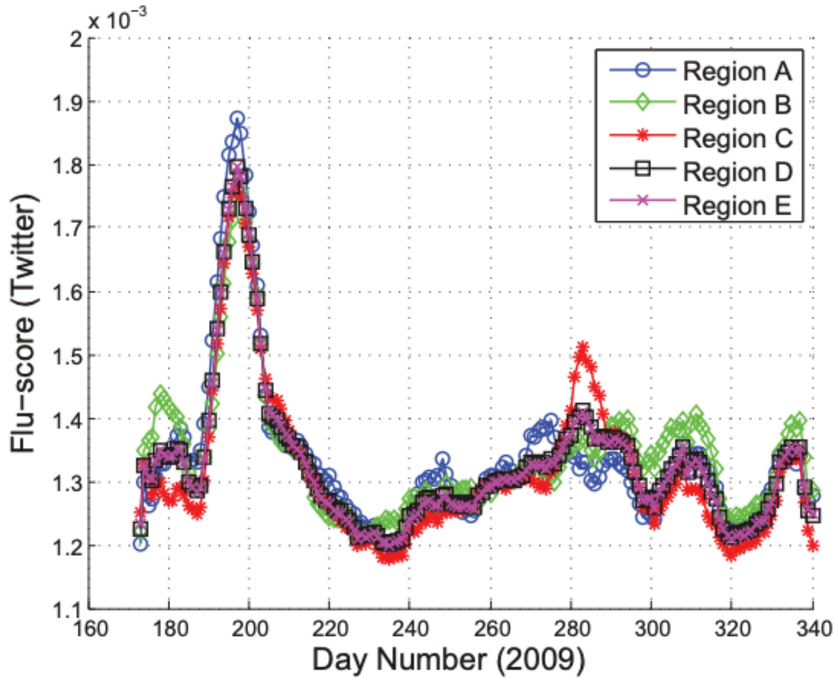
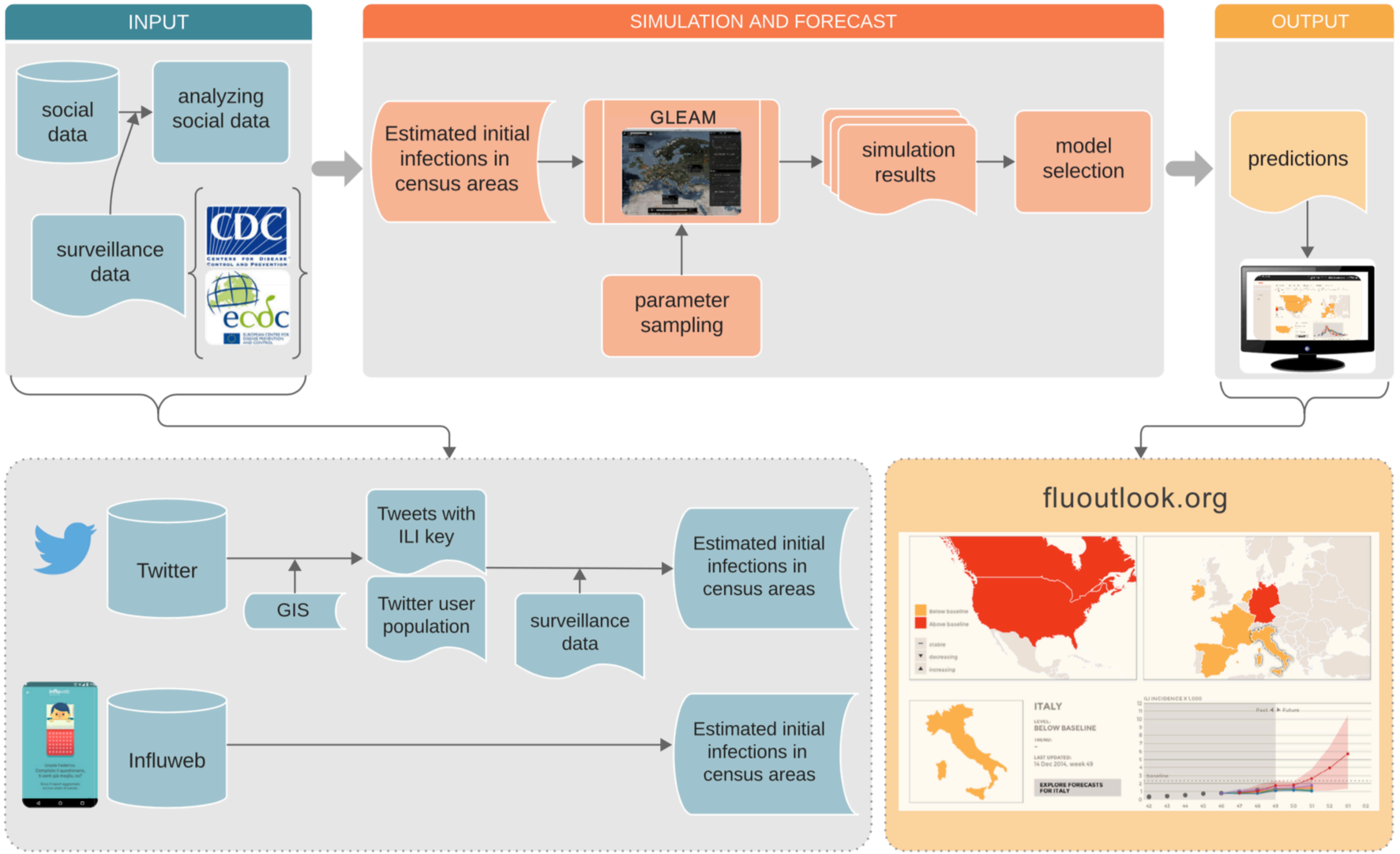


Fig. 2: Twitter's flu-scores based on our choice of markers for regions A-E (weeks 26-49, 2009). Smoothing with a 7-point moving average (the length of a week) has been applied.

User-generated content

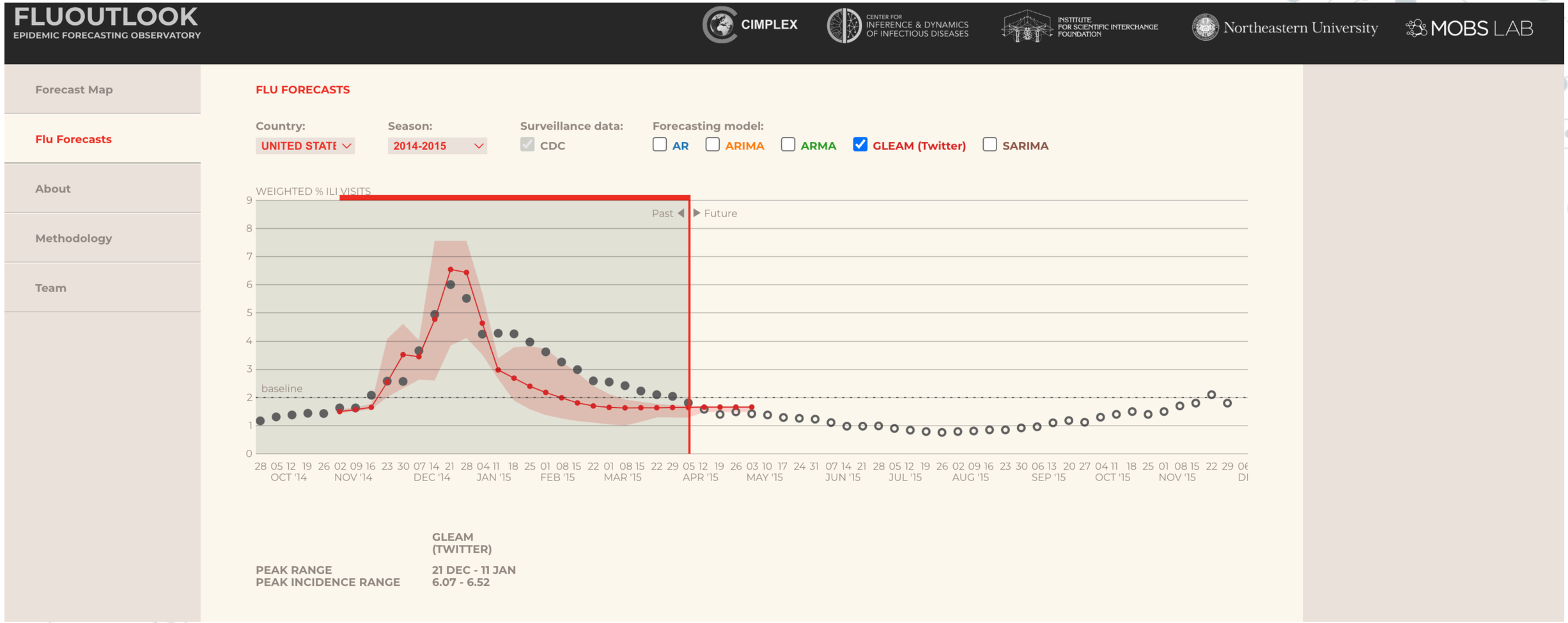
- Advantage: social media data is user-generated content created without any specific request from health researchers or practitioners allowing for **content analysis**
- Social media analyses also enable **network analyses**, for example, to understand the effect of social influences on health status or behaviors of interest.
- **Supervised approach**: a machine learning model is trained to classify tweets based on a labeled dataset.
- Back in the days: simple classifiers such as Naive Bayes and Maximum Entropy classifiers were commonly used. Nowadays: the current state-of-the-art approach leverages transformer models, or large language models like GPT.
- **Unsupervised approaches** attempt to categorize tweets but do so without knowing the categories in advance. A popular approach is **topic modeling,** which can identify topics by finding and clustering text patterns in documents.

Combining data streams



J. Brownstein *et al.*, JMIR Public Health and Surveillance 3, 4 (2017)

Combining data streams



fluoutlook.org

Attitudes and behaviors

OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

Assessing Vaccination Sentiments with Online Social Media: Implications for Infectious Disease Dynamics and Control

Marcel Salathé*, Shashank Khandelwal

Center for Infectious Disease Dynamics, Department of Biology, Penn State University, University Park, Pennsylvania, United States of America

Abstract

There is great interest in the dynamics of health behaviors in social networks and how they affect collective public health outcomes, but measuring population health behaviors over time and space requires substantial resources. Here, we use publicly available data from 101,853 users of online social media collected over a time period of almost six months to measure the spatio-temporal sentiment towards a new vaccine. We validated our approach by identifying a strong correlation between sentiments expressed online and CDC-estimated vaccination rates by region. Analysis of the network of opinionated users showed that information flows more often between users who share the same sentiments - and less often between users who do not share the same sentiments - than expected by chance alone. We also found that most communities are dominated by either positive or negative sentiments towards the novel vaccine. Simulations of infectious disease transmission show that if clusters of negative vaccine sentiments lead to clusters of unprotected individuals, the likelihood of disease outbreaks is greatly increased. Online social media provide unprecedented access to data allowing for inexpensive and efficient tools to identify target areas for intervention efforts and to evaluate their effectiveness.

Citation: Salathé M, Khandelwal S (2011) Assessing Vaccination Sentiments with Online Social Media: Implications for Infectious Disease Dynamics and Control. PLoS Comput Biol 7(10): e1002199. doi:10.1371/journal.pcbi.1002199

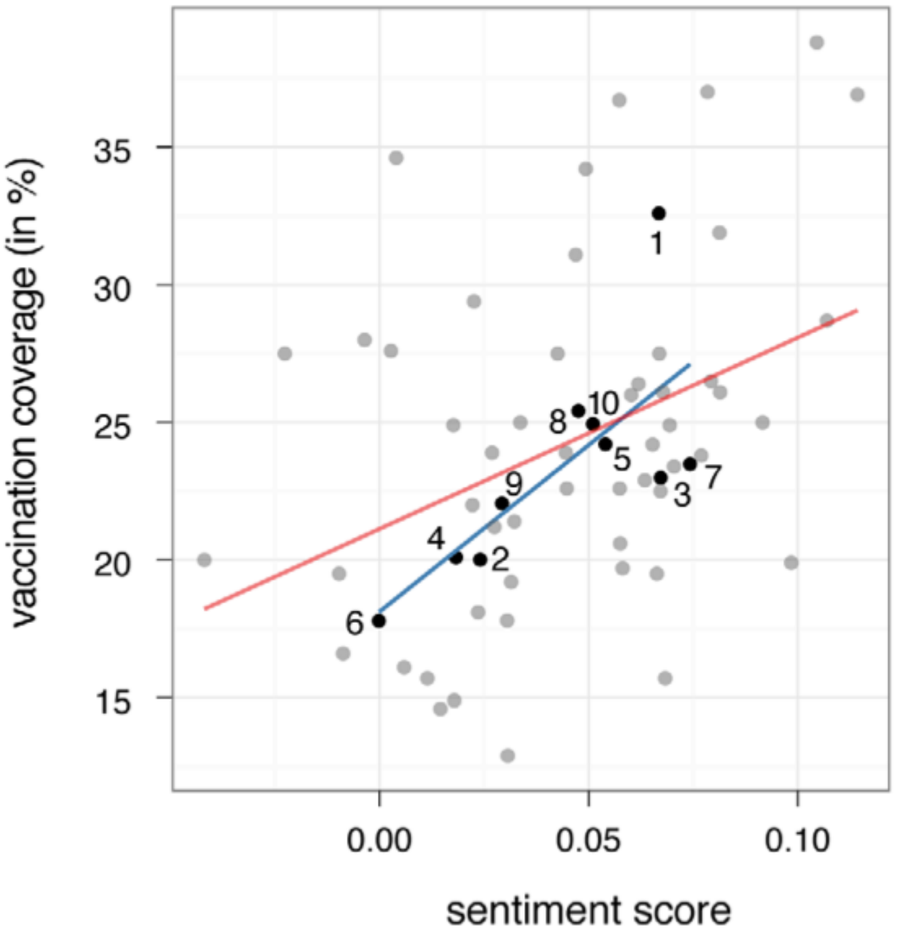
Editor: Lauren Ancel Meyers, University of Texas at Austin, United States of America

Received May 10, 2011; **Accepted** July 30, 2011; **Published** October 13, 2011

Funding: MS acknowledges funding from Society in Science: the Branco Weiss fellowship. http://www.society-in-science.org/. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: salathe@psu.edu



Instagram

Measuring and Characterizing Nutritional Information of Food and Ingestion Content in Instagram

Sanket S. Sharma
College of Computing
Georgia Institute of Technology
Atlanta, GA
sanket@gatech.edu

Munmun De Choudhury
School of Interactive Computing
Georgia Institute of Technology
Atlanta, GA
munmund@gatech.edu

Post tags	Canonical name	Calorie
miami, organic, garden, usa, food, fresh, creole, okra	okra	28.33
bykaila, postre, luneslight, cuantocomer, dessert, flan	flan	177.25
yam, instafood, hamburger, the-bird	yam, ham-burger	219.93
muesli, granola, easterngra-nola, localislovely	granola, muesli	330.87
cheesecake, breakfast, cheese-cakefactory, redvelvet	cheesecake	402

Table 1: Example posts and calorific content.

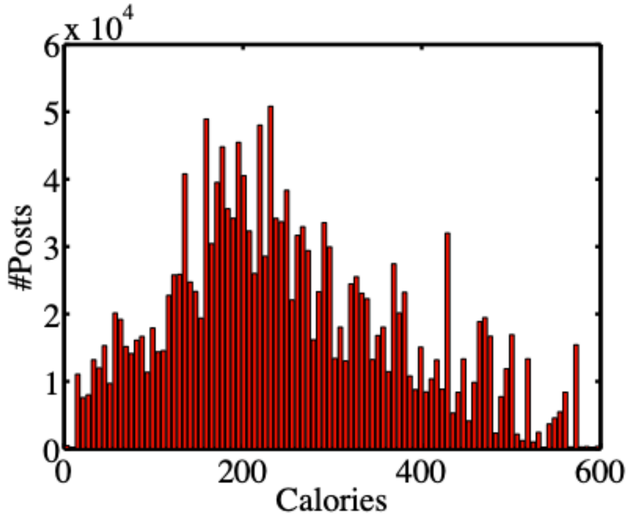
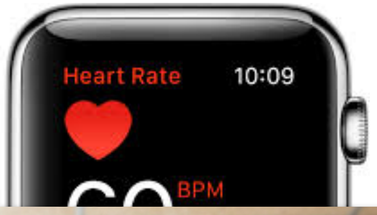


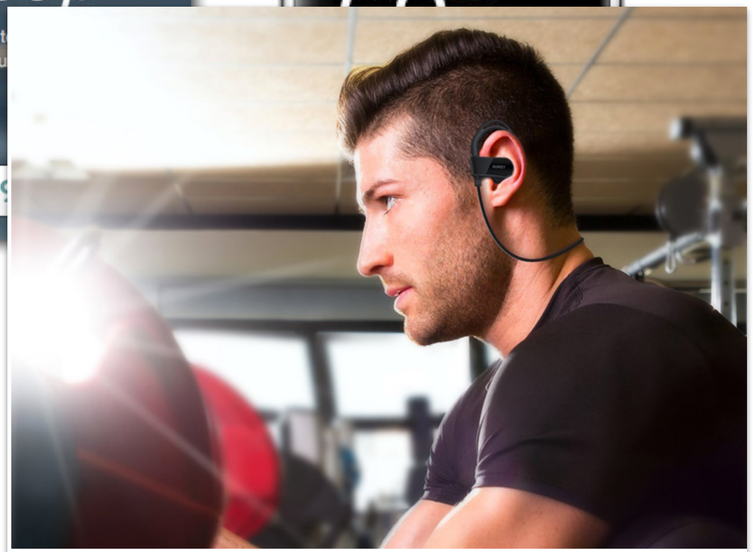
Figure 1: Distribution of number of posts over calorific values.

Data Sources

- Search queries and access logs
- Participatory surveillance
- Social media
- Mobile phones
- **Wearable sensors**
- Other data sources

Wearables

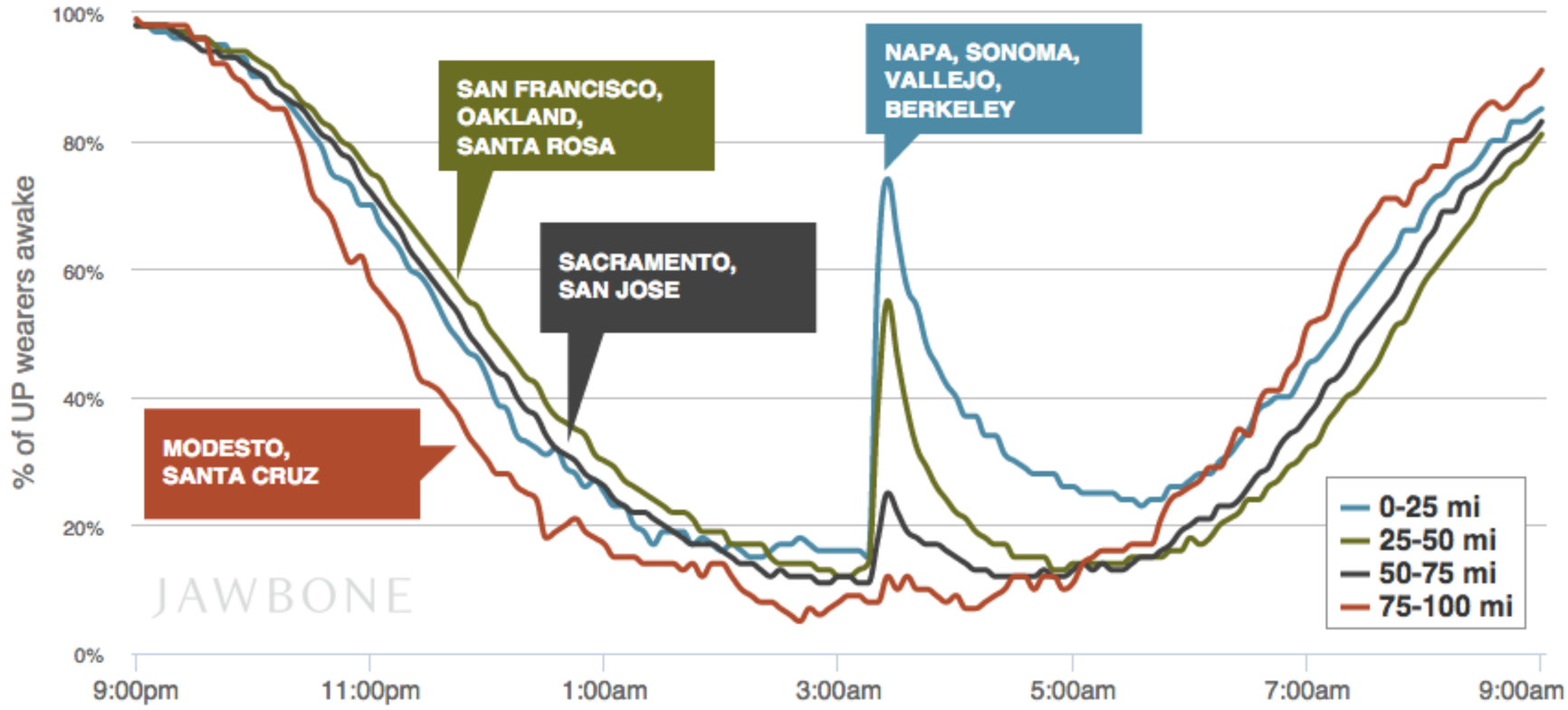




ACCESSORIES

These $20 Bluetooth earbuds have a built-in heart rate monitor

Wearables



jawbone.com/blog/napa-earthquake-effect-on-sleep

Wearables

ARTICLES
https://doi.org/10.1038/s41562-020-01046-9

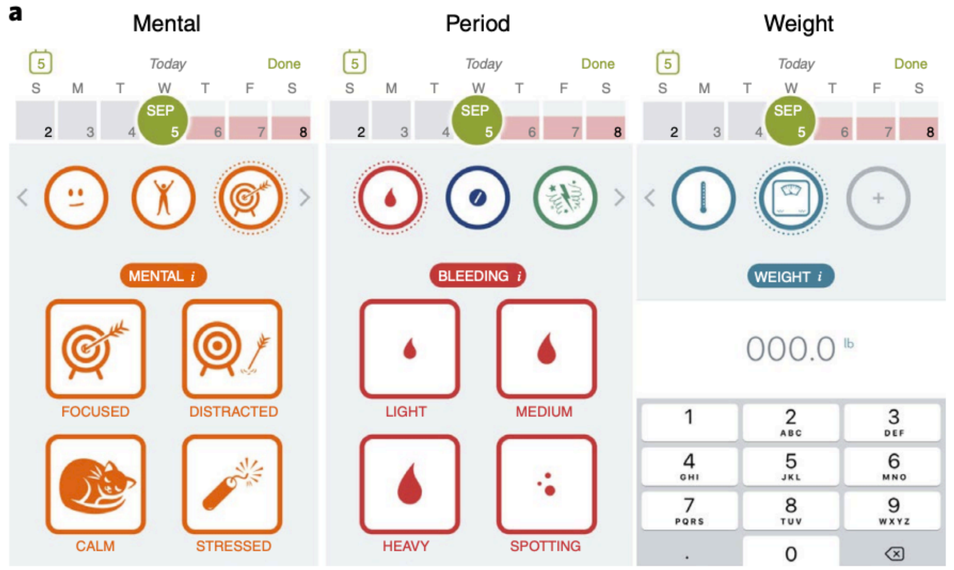
nature human behaviour

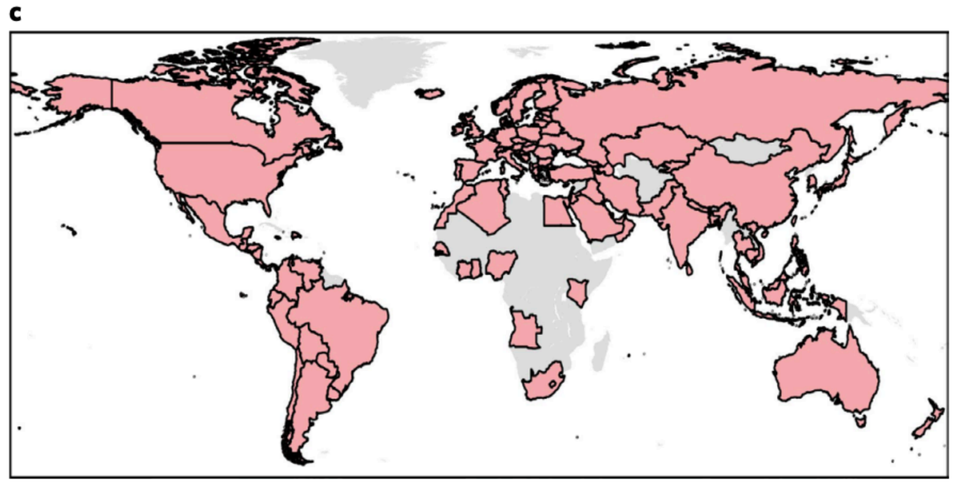
Check for updates

Daily, weekly, seasonal and menstrual cycles in women's mood, behaviour and vital signs

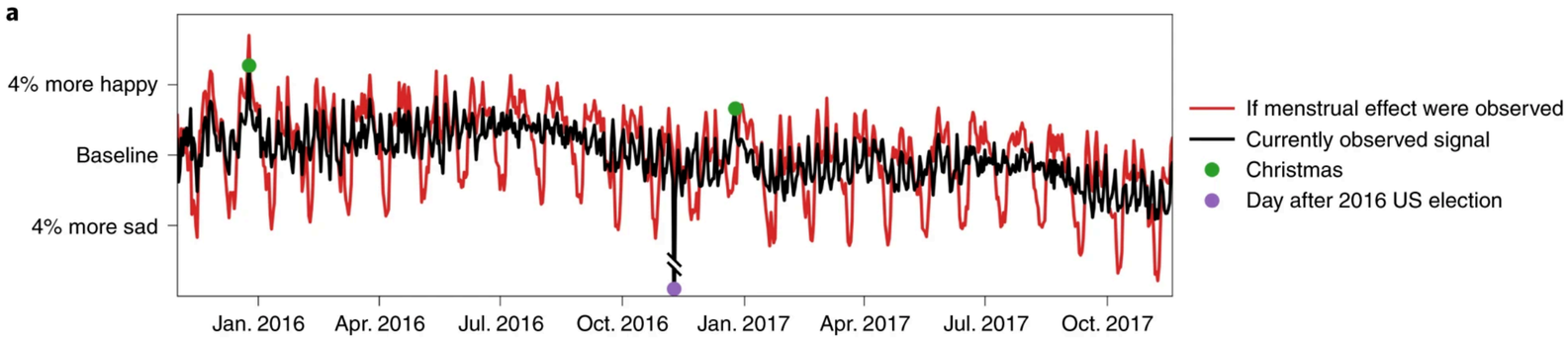
Emma Pierson[1,2], Tim Althoff[3], Daniel Thomas[4], Paula Hillard[5] and Jure Leskovec[1,6]

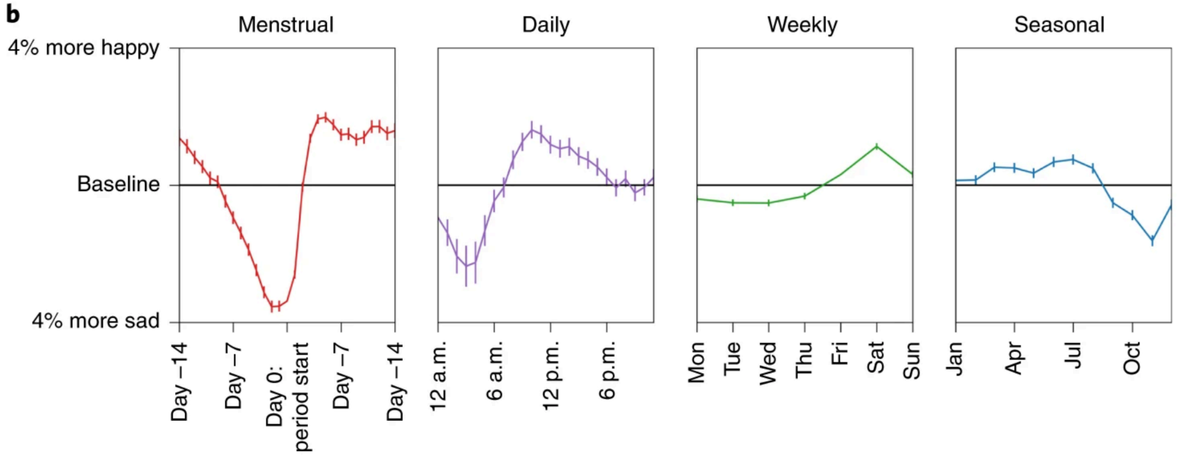
we analyse **241 million observations** from **3.3 million women across 109 countries**, tracking **15 dimensions** of mood, behaviour and vital signs using a women's health mobile app

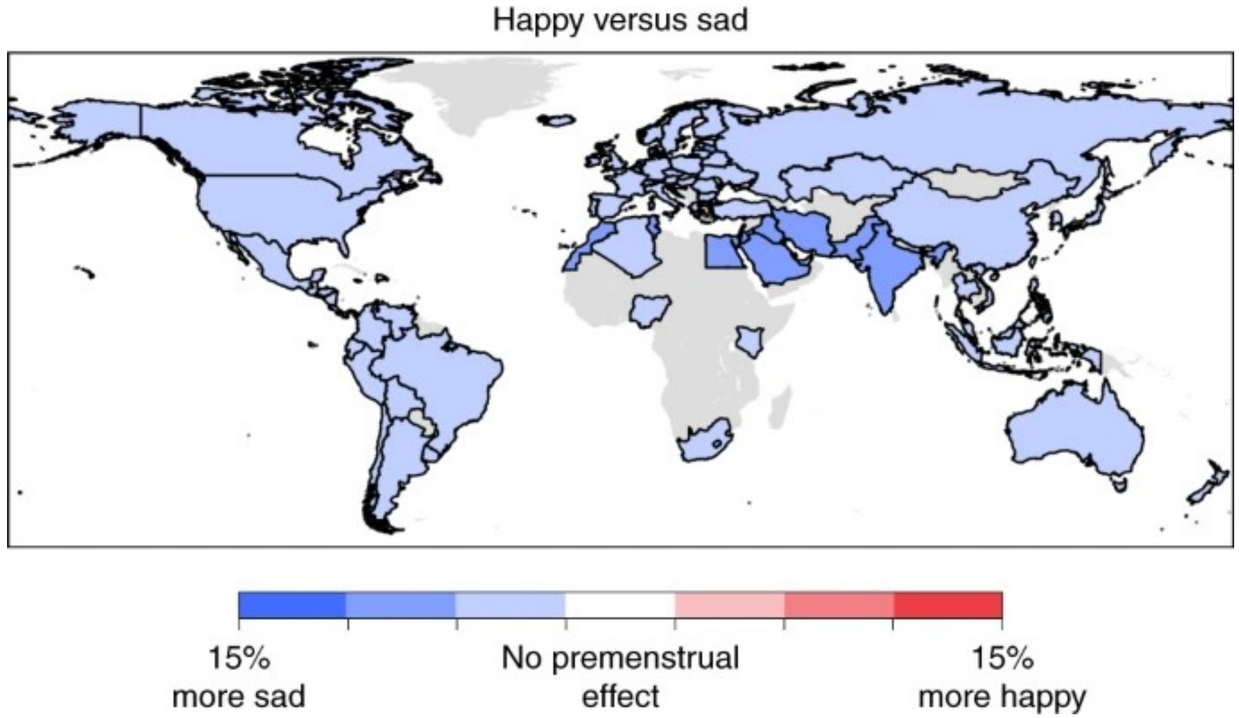




Wearables


a
4% more happy
Baseline
4% more sad
Jan. 2016
Apr. 2016
Jul. 2016
Oct. 2016
Jan. 2017
Apr. 2017
Jul. 2017
Oct. 2017
If menstrual effect were observed
Currently observed signal
Christmas
Day after 2016 US election


b
Menstrual
Daily
Weekly
Seasonal
4% more happy
Baseline
4% more sad
Day −14
Day −7
Day 0: period start
Day −7
Day −14
12 a.m.
6 a.m.
12 p.m.
6 p.m.
Mon
Tue
Wed
Thu
Fri
Sat
Sun
Jan
Apr
Jul
Oct


Happy versus sad
15% more sad
No premenstrual effect
15% more happy

Data Sources

- Search queries and access logs
- Participatory surveillance
- Social media
- Mobile phones
- Wearable sensors
- **Other data sources**

Purchase records

SCIENTIFIC DATA

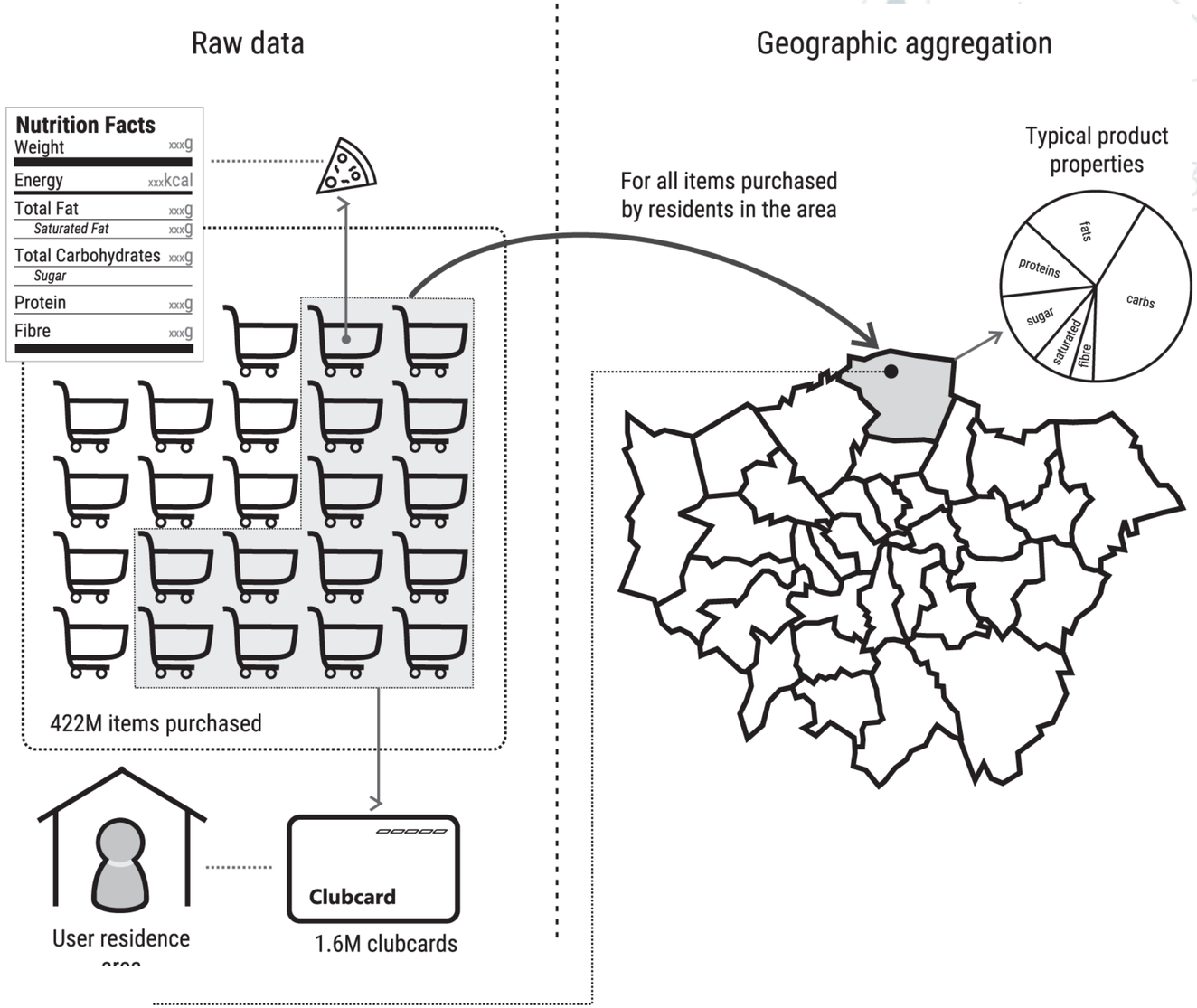
Check for updates

OPEN

DATA DESCRIPTOR

Tesco Grocery 1.0, a large-scale dataset of grocery purchases in London

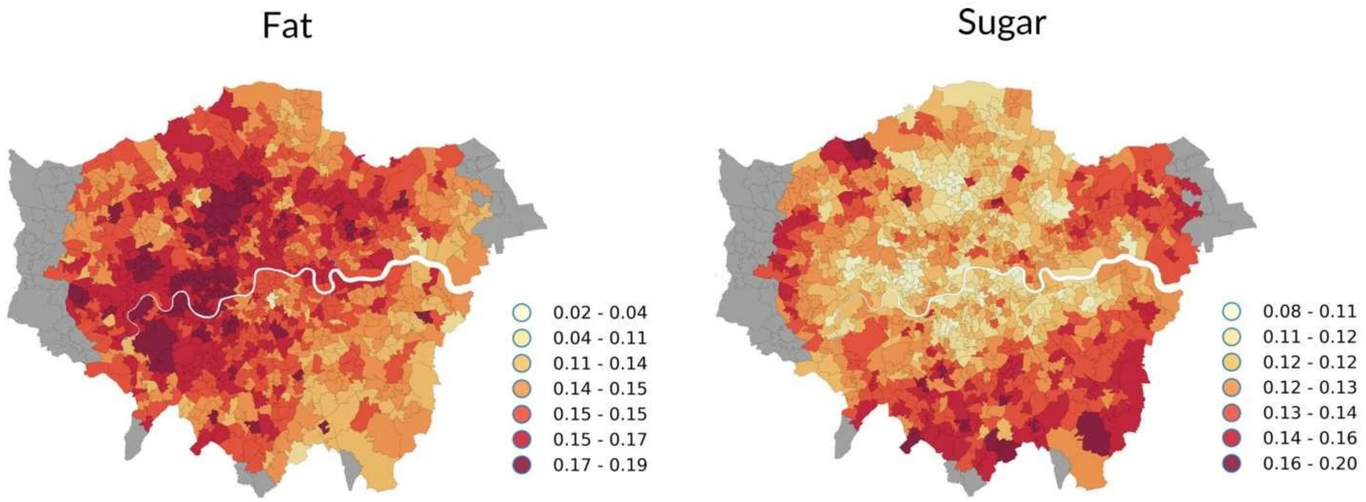
Luca Maria Aiello[1✉], Daniele Quercia[1,4], Rossano Schifanella[2,5] & Lucia Del Prete[3]

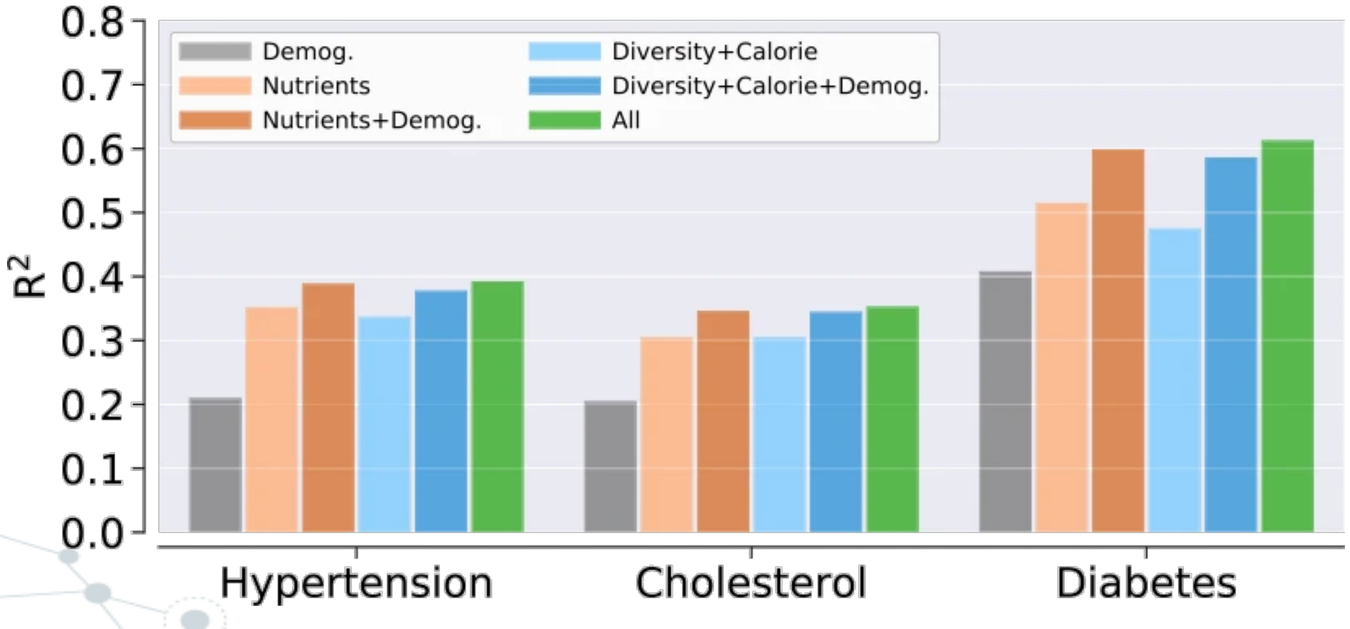


Fat

Sugar

Purchase records





Future challenges

Challenges

- **Availability - and ultimately ownership - of data.** Data sources are not persistent, see the case of Twitter.
- **Bias.** Data from digital services is generated by people who use those services. Moreover, our digital self is often quite different from our true self. What we share on online services is already highly selective.
- **Fake content.** According to Cloudfare, in 2022, about one-third of all internet traffic was generated by bots…
- **Models' training.** Digital epidemiology models trained on historical data from dynamic environments can over time deteriorate quite significantly.

Next… social contagion