Citizen-Sourced Data for Public Health Modeling

Rumi Chunara, PhD
Assistant Professor
NYU Computer Science & Engineering
NYU Global Public Health

NYU COLLEGE OF GLOBAL PUBLIC HEALTH

NYU TANDON SCHOOL OF ENGINEERING

Citizen-Sourced Data for Public Health Modeling

- Public health intro and data overview
- Crowdsourcing and knowledge generation in public health
- Learning spatio-temporal features
- Other data opportunities

Citizen-Sourced Data for Public Health Modeling

- Public health intro and data overview
- Crowdsourcing and knowledge generation in public health
- Learning spatio-temporal features
- Other data opportunities

Emerging and Re-emerging Infections, 1996-2010

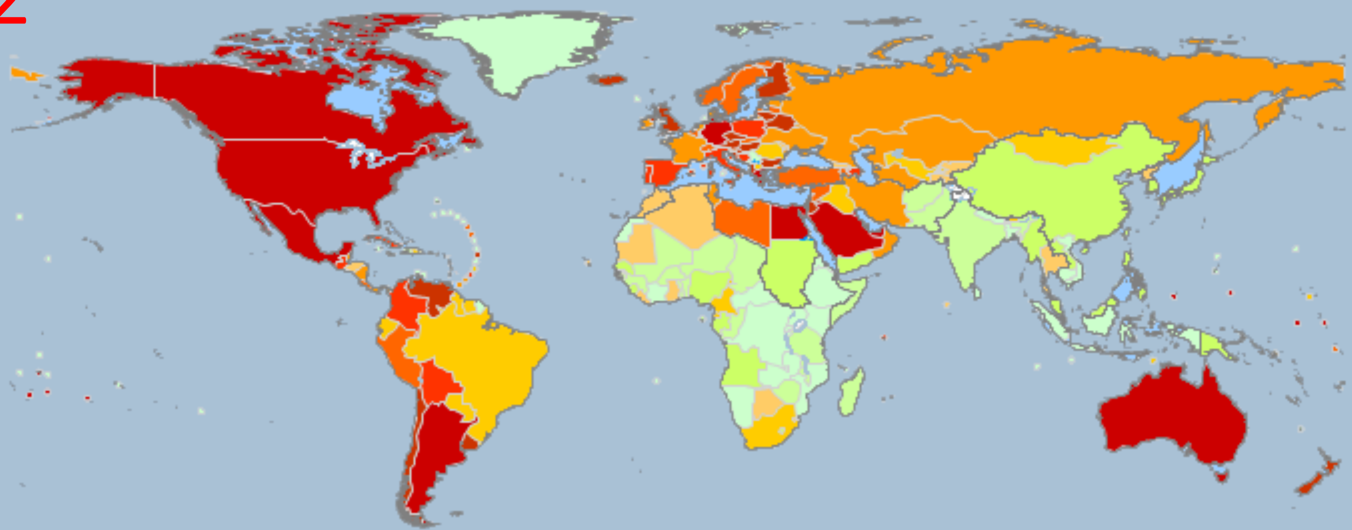
Cryptosporidiosis
Human Monkeypox
E.Coli O157
Venezuelan Equine Encephalitis
Dengue Haemorrhagic Fever
Ebola Haemorrhagic Fever
Marburg Haemorrhagic Fever
Ross River Virus
Hendra Virus
Reston Virus

West Nile Virus
Legionnaire’s Disease
Severe Acute Respiratory Syndrome (SARS)
Malaria
Typhoid
Cholera
BSE
Lassa Fever
Yellow Fever

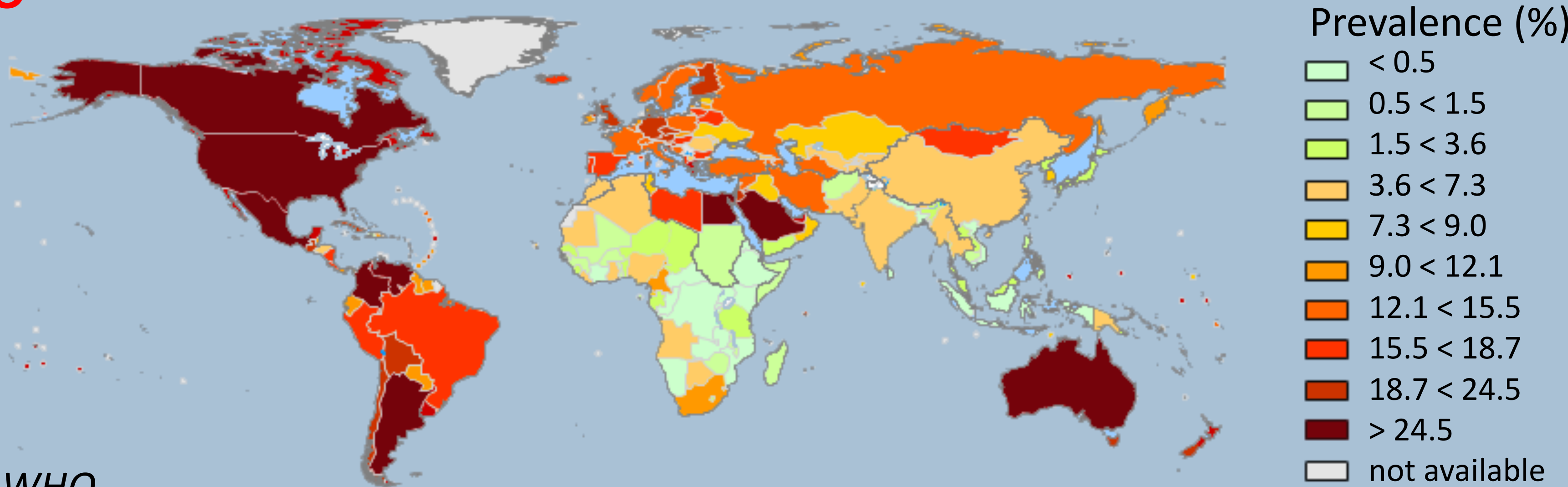
Lyme Borreliosis
Echinococcosis
Diphtheria
Influenza A (H5N1)
Nipah Virus
RVF/VHF
O’Nyong-Nyong Fever
Buruli Ulcer
Multidrug Resistant Salmonella
nvCJD

Worldwide Obesity Prevalence

2002



2010



Source: WHO

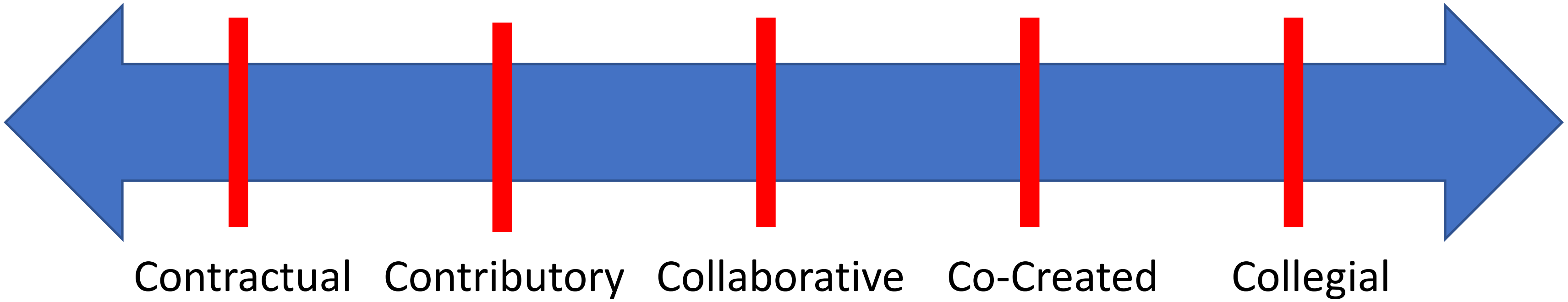
	Current Public Health Surveillance
Whole population	+
Specificity	+
Doctor/Nurse involvement	+
Speed	-
Sensitivity	-
Cost	-
Public engagement	-

	Current Public Health Surveillance	Crowdsourced Data
Whole population	+	-
Specificity	+	-
Doctor/Nurse involvement	+	-
Speed	-	+
Sensitivity	-	+
Cost	-	+
Public engagement	-	+

Citizen-Sourced Data for Public Health Modeling

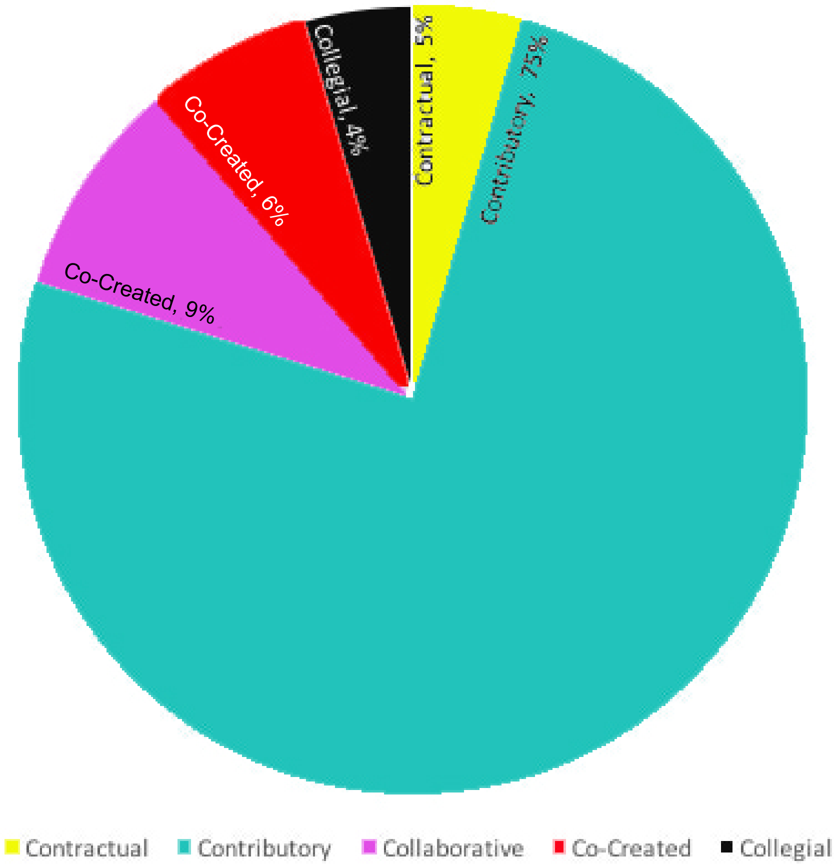
- Public health intro and data overview
- Crowdsourcing and knowledge generation in public health
- Learning spatio-temporal features
- Other data opportunities

Types of Crowdsourcing



Shirk et al. 2011

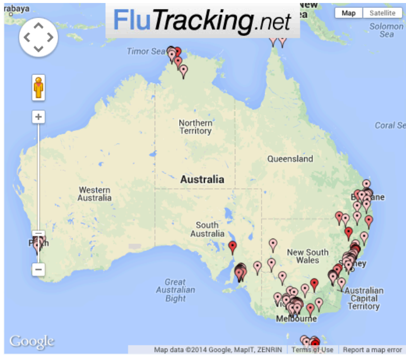
Types of Crowdsourcing



Quarooni et al. CHI 2016

80% of citizen projects only harness participation in a limited form, such as for completion of tasks

Crowdsourcing in Public Health



sickweather

gripenet

flu near you

do you have it in you?

Influenzanet

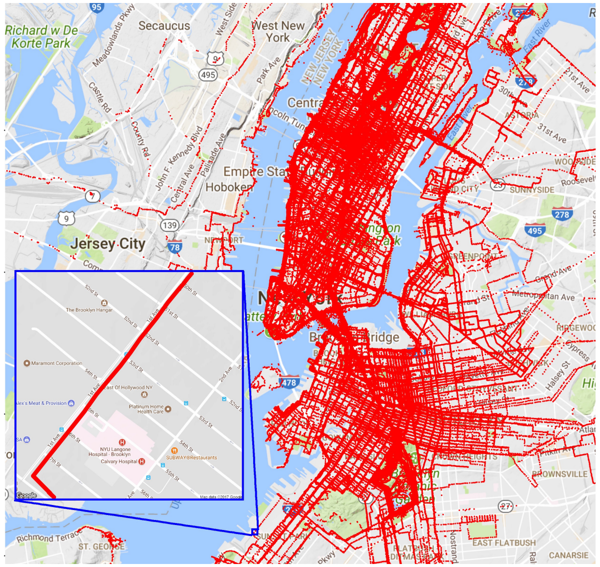
Influ

flusurvey

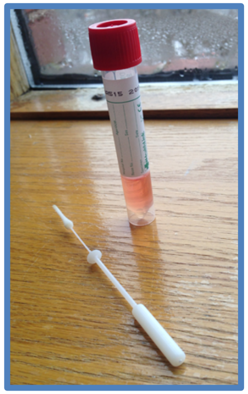
influweb

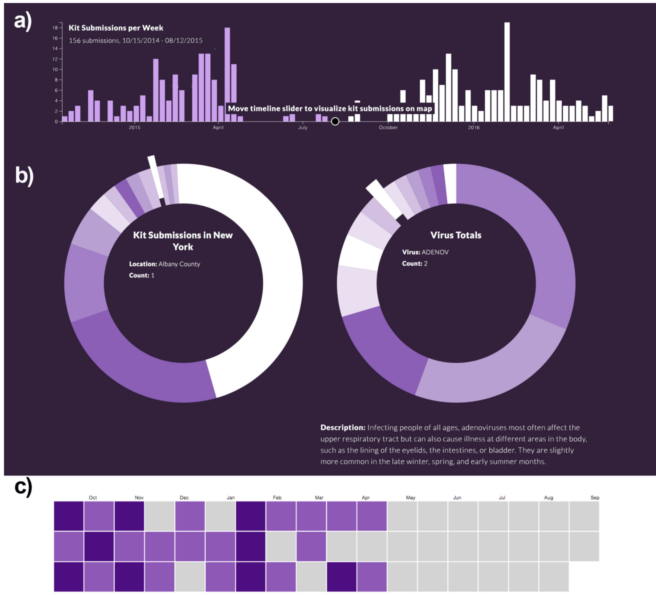
Keeping Pace

fitbit



goVIRAL





Goff et al. AMIA 2014, Plos Current 2015
Chunara (under submission) 2017

Findings from Our Efforts Generating Public Health Knowledge via Crowdsourcing

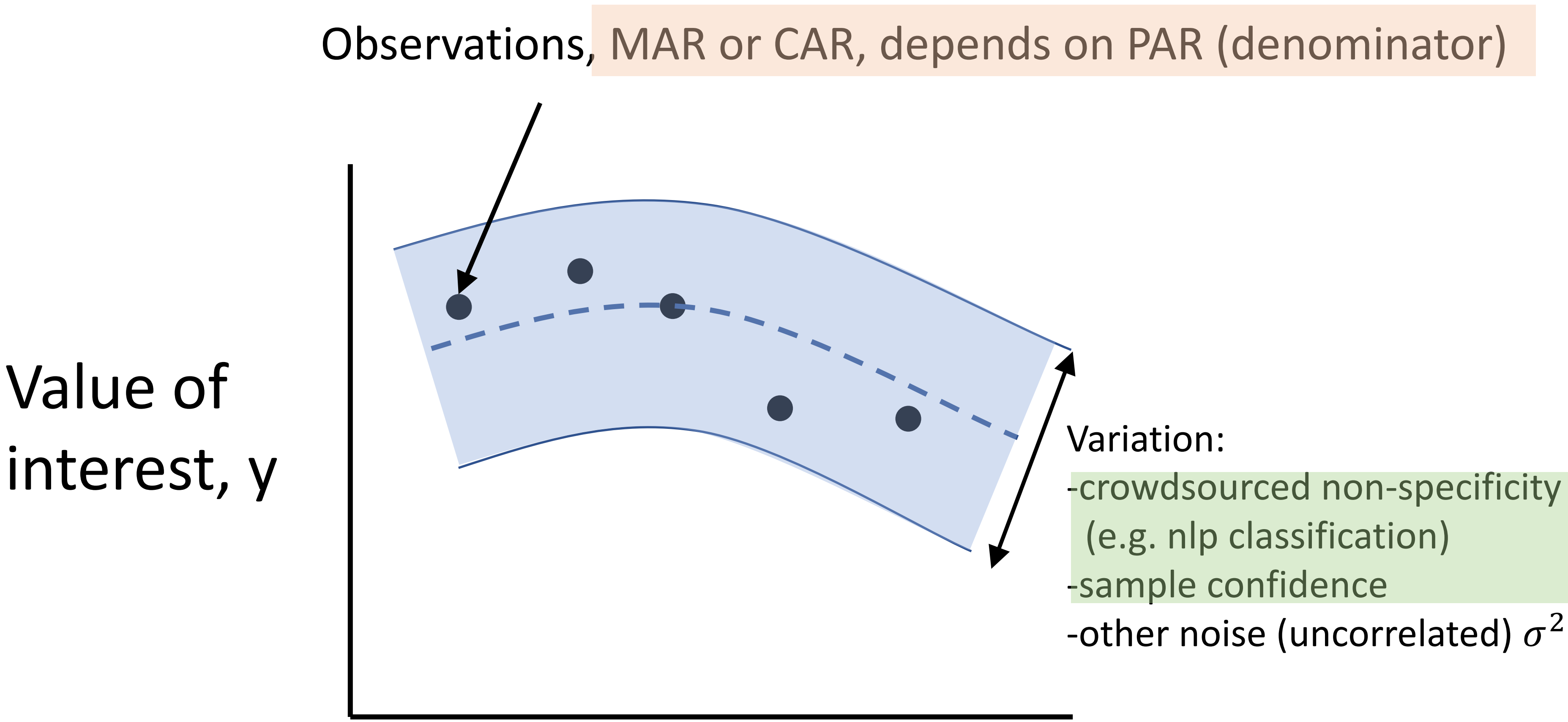
1. Unique incentives in public health: collective and intrinsic motivations are the most salient (aligns with Nov et al. 2011, Law et al. 2016)

2. Offline recruitment: important to improve external validity (Chunara et al. AJPM 2016)

3. Uniqueness of data generated: Spatio-temporal data Opportunities (Salathé et al. 2011, Relia et al., Rehman et al. 2017)

Chunara (under submission) 2017

Data Challenges



$$y \sim \mathcal{N}(\phi(t) + f(t) \; , \sigma^2)$$

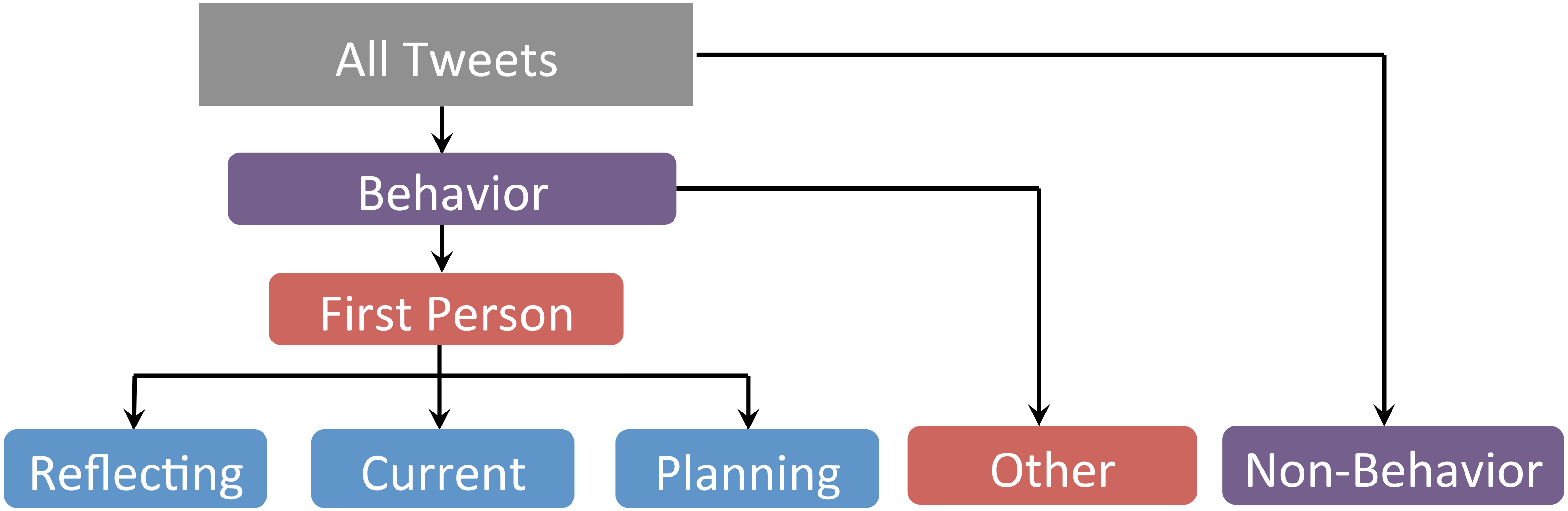
Citizen-Sourced Data for Public Health Modeling

- Public health intro and data overview
- Crowdsourcing and knowledge generation in public health
- Learning spatio-temporal features
- Other data opportunities

Learning Features – Twitter, alcohol behavior example

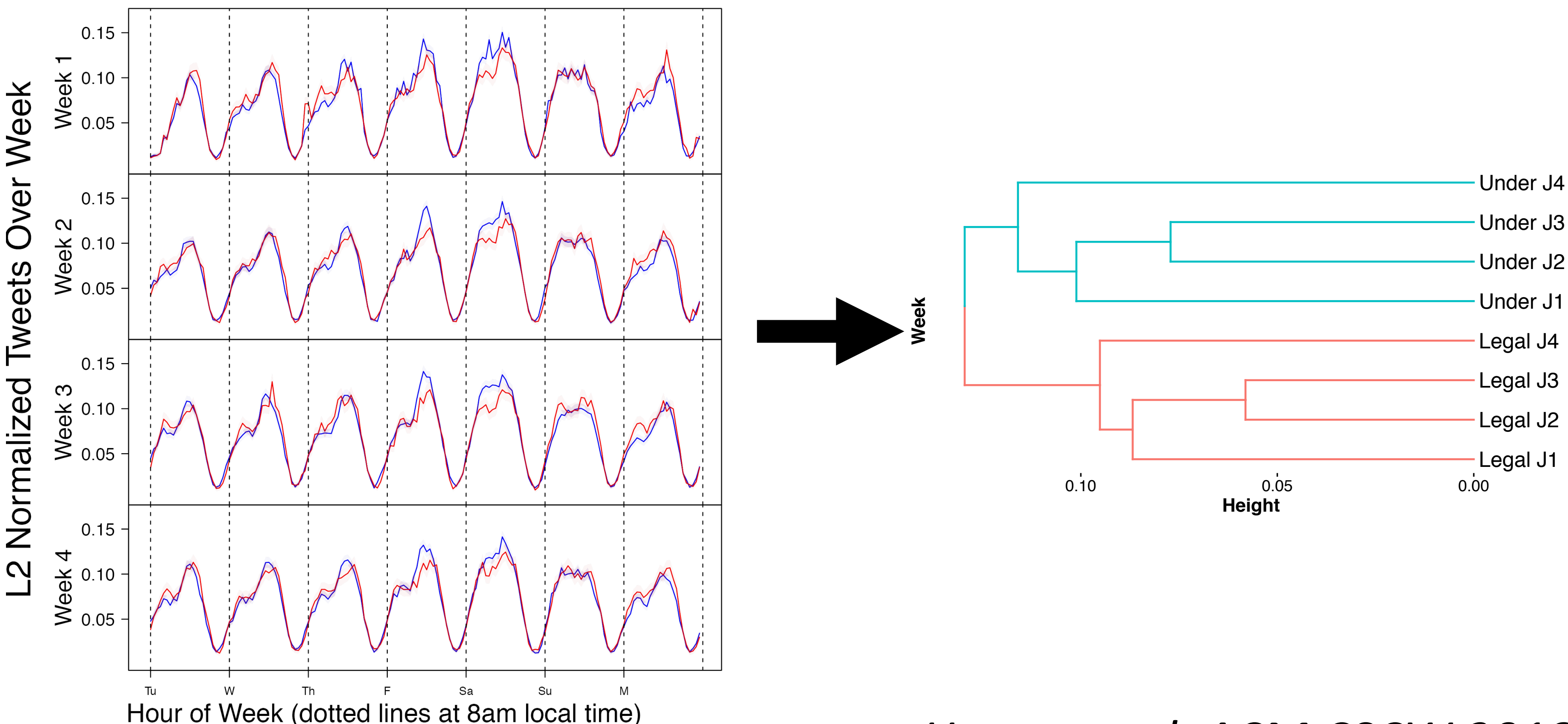
- Social media (e.g. Twitter) recognized as useful source to learn about incidence and behaviors related to infectious and non-communicable diseases
- Much work relates to time-series, mapping and prediction
- We developed an NLP pipeline to specifically isolate individuals engaged in a behavior, and then examines relevant temporal representations (features) (Liu et al CSCW 2017, Huang et al CSCW 2018)

NLP Hierarchical Pipeline



Liu et al. ACM CSCW 2017

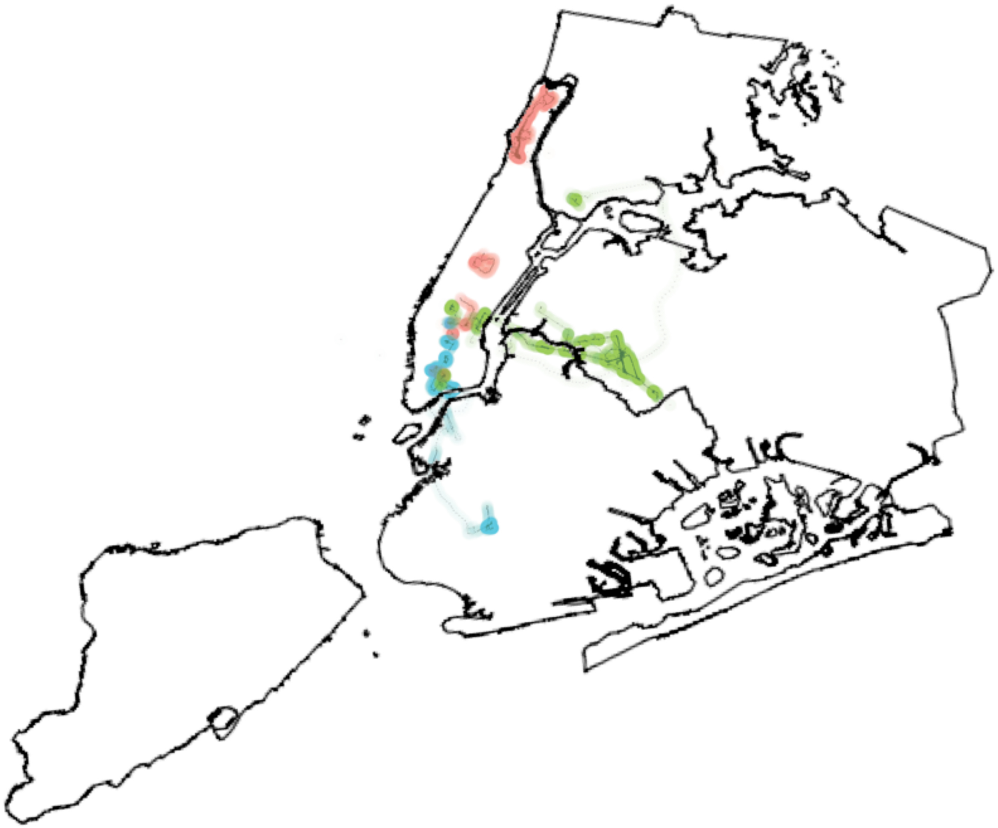
High-resolution Temporal Representations



red: overage, blue: underage

Huang et al. ACM CSCW 2018

Spatial Representations



Problem:

- Social attitudes like racism/homophobia can be predictors for health outcomes
- ZIP codes are defined to optimize mail delivery, need way to define exposure from a place based on the context

Approach:

- Develop a method for spatial representations based on SOMs and social media classification
- Use mobility data from a cohort of MSM to show that spatial representations matter

Learning Features – Twitter, social attitude example

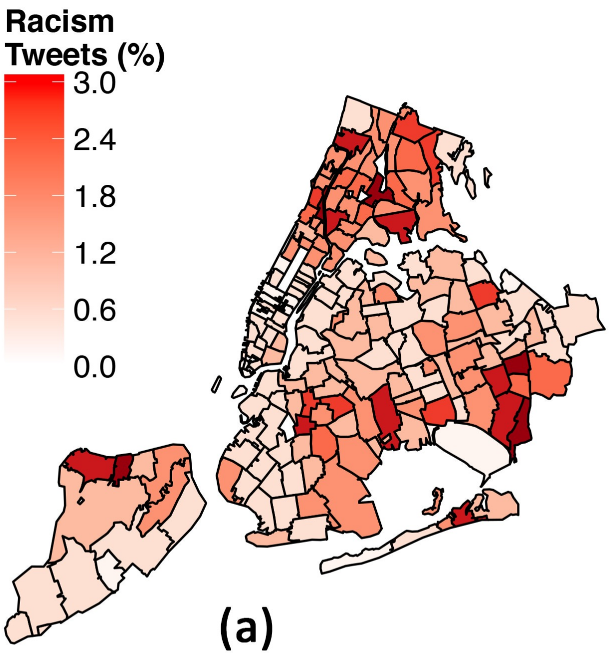
- Use SOM's a common neural network approach to learn embedded structure in the spatial distribution of classified Tweets
- Provides an interpretable output that can be mapped geographically

i-1,j-1 k	i-1,j k	i-1,j+1 s
i,j-1 k	i,j k	i,j+1 s
i+1,j-1 k	i+1,j s	i+1,j+1 s

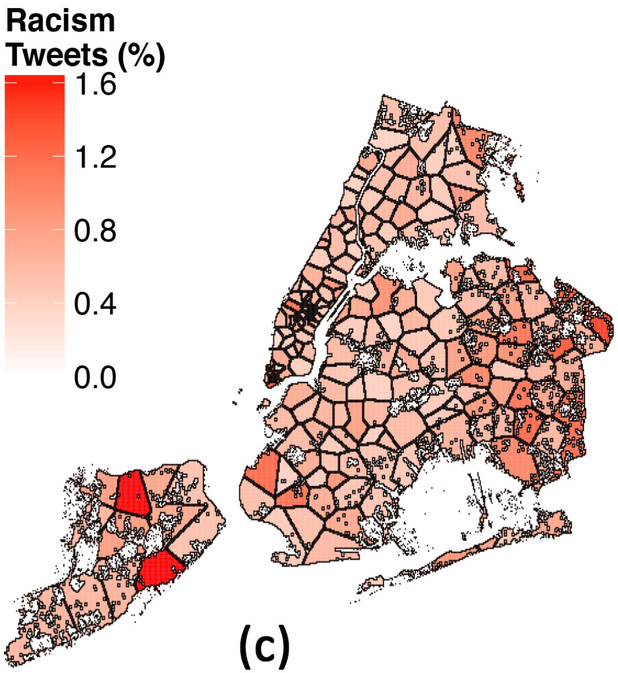
Illustration depicting grid cell (i,j) and its neighbors at threshold = 1 for formation of boundary between grid cells with different *weights* (value in second row)

Relia et al. (under review) 2017

Learning Features – Twitter, social attitude example



Distribution of Racism Exposure by ZIP



Distribution of Racism Exposure by SOM

Relia et al. (under review) 2017

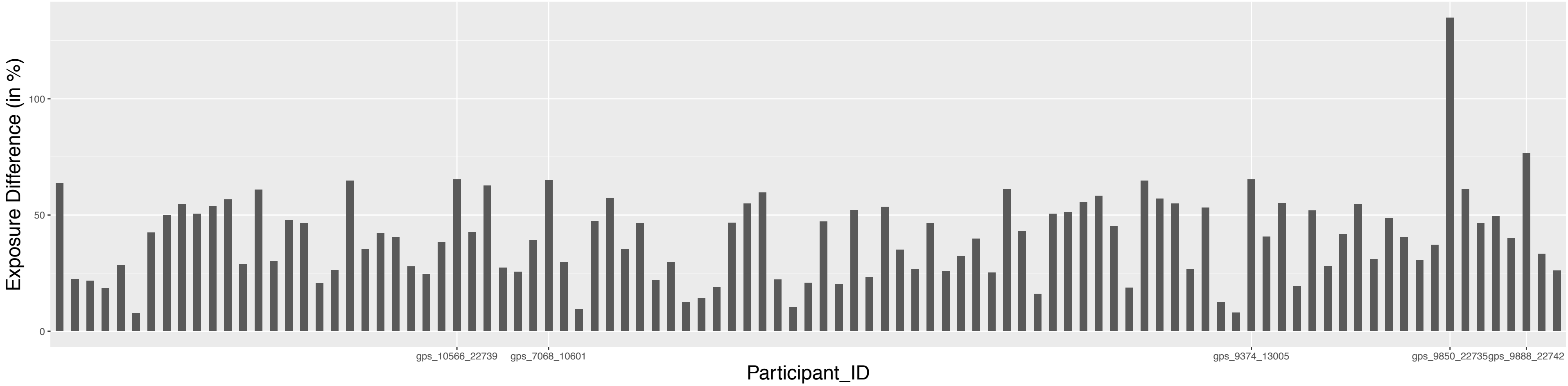
How Good Are These Representations?

- Evaluate using common cluster evaluation techniques:
  - ✓ Robustness to missing data
  - ✓ Lower mean variance
  - ✓ Entropy

Overall SOM provide a more consistent geographical compartmentalization of each attitude

What Difference Does This Make?

Mean racism exposure difference using SOM versus Zip Codes was **40.3%** (SD: 18.8%).



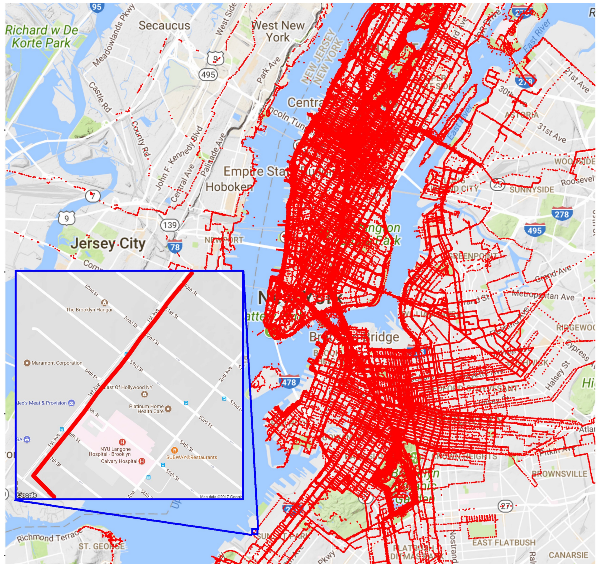
Citizen-Sourced Data for Public Health Modeling

- Public health intro and data overview
- Crowdsourcing and knowledge generation in public health
- Learning spatio-temporal features
- Other data opportunities

Data Opportunities

- Predicting individual-level mobility a growing problem of importance
- Existing data (CDRs, GPS trackers) temporally rich, though expensive

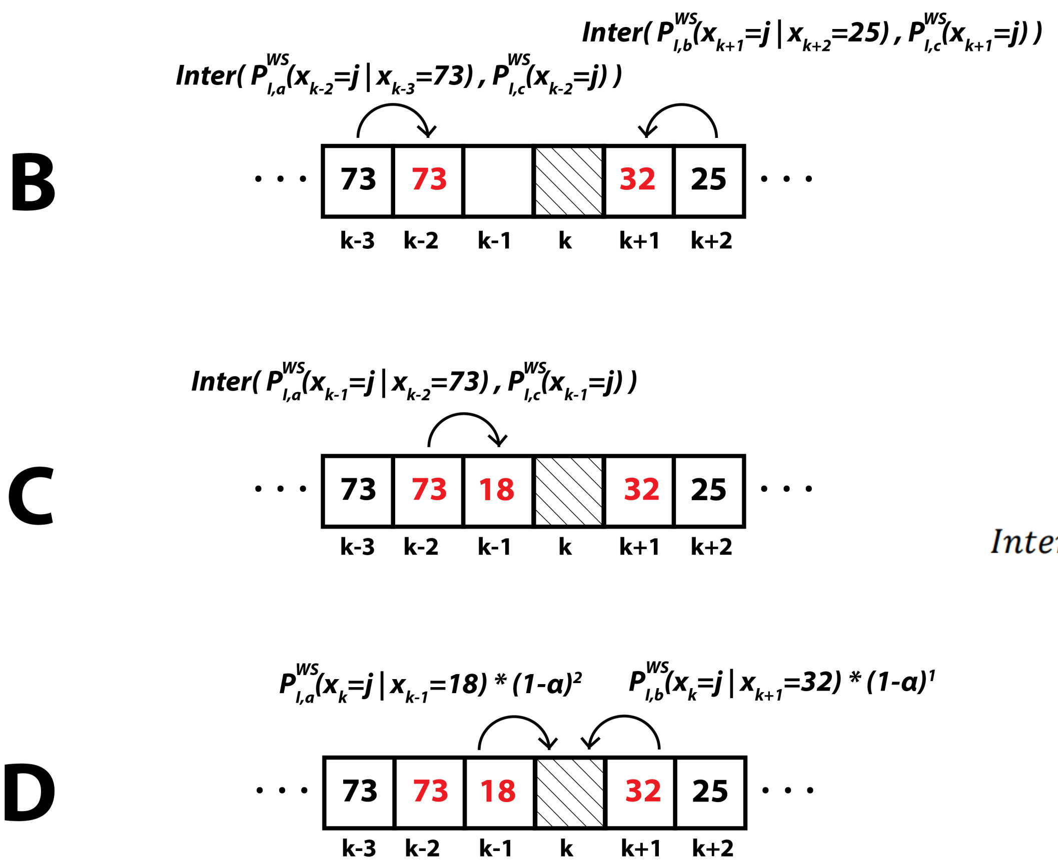
- High-resolution, and availability of social media
- Sparse nature brings challenges



Intermediate Location Computing: Pre-processing

- 6 months of Twitter data from the API
- Pre-processed data by mapping to grids (0.1, 0.5 and 1 mi resolution)
- Inferred stay and home location
- Removed non-personal accounts
- Included users must have at least 1 location value present for each h during daytime hours (irrespective of day and week)
- 29,491, 4,947 and 1,119 users (r_i = 1 hour) and 45,710, 8,083 and 2,395 users (r_i = 2 hours) included from NYC, DC and SF

Intermediate Location Computing: Algorithm



x_k location at position k

$P_{I,a}$ Next location probability

$P_{I,b}$ Previous location probability

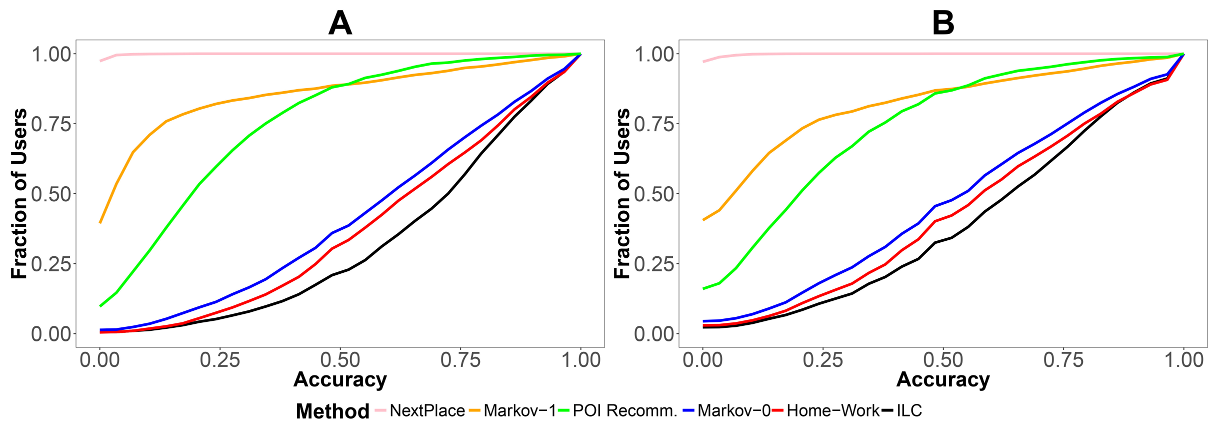
$P_{I,c}$ Community location probability

WS Week and hour specific

$$Inter(F1, F2) = \begin{cases} \max\_loc(F1) & if\ \max\_loc(F1)\ != NULL \\ \max\_loc(F2) & otherwise \end{cases}$$

Rehman et al. 2017

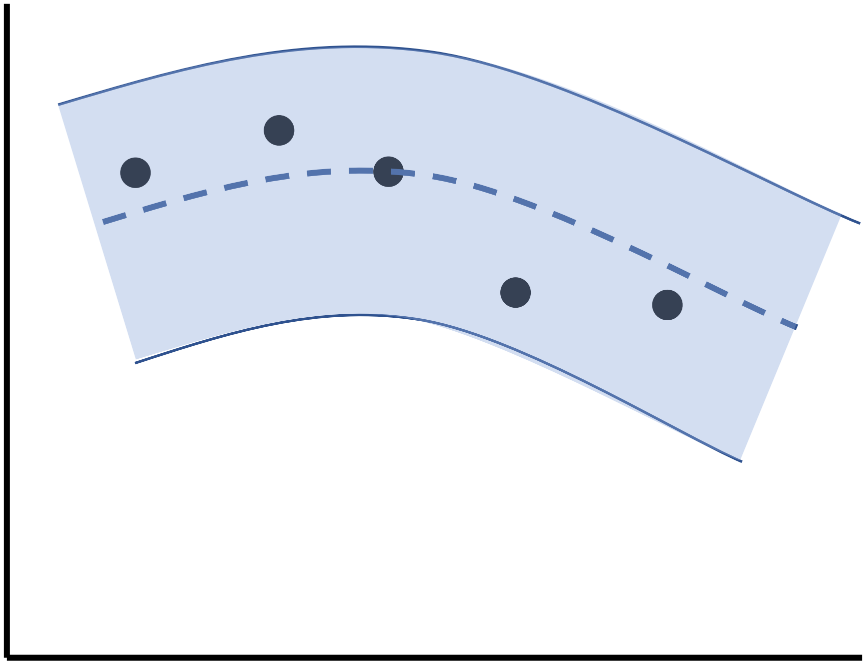
ILC: Accuracy versus baseline models



City	r_i	Top 1	Top 3	Home-Work	Markov O(0)	Markov O(1)	POI	NextPlace
New York City	r_i=1	72.69	82.35	65.54	64.65	26.39	15.59	0.17
	r_i=2	64.78	77.38	59.28	57.98	32.56	19.11	0.21
Washington, DC	r_i=1	75.08	83.61	66.91	65.76	27.75	31.27	0.11
	r_i=2	68.85	79.57	62.35	60.64	34.13	34.56	0.19
San Francisco	r_i=1	77.20	86.28	67.74	67.21	16.78	35.49	0.15
	r_i=2	70.78	82.06	63.66	62.91	19.52	32.69	0.22

Transfer Learning to Improve Specificity

- Point value specificity challenging in crowdsourced data
- Manifests in public health through syndromic data



When Google Got Flu Wrong

nature International weekly journal of s

Home | News & Comment | Research | Careers & Jobs |

Archive > Volume 494 > Issue 7436 > News > Articl

NATURE | NEWS

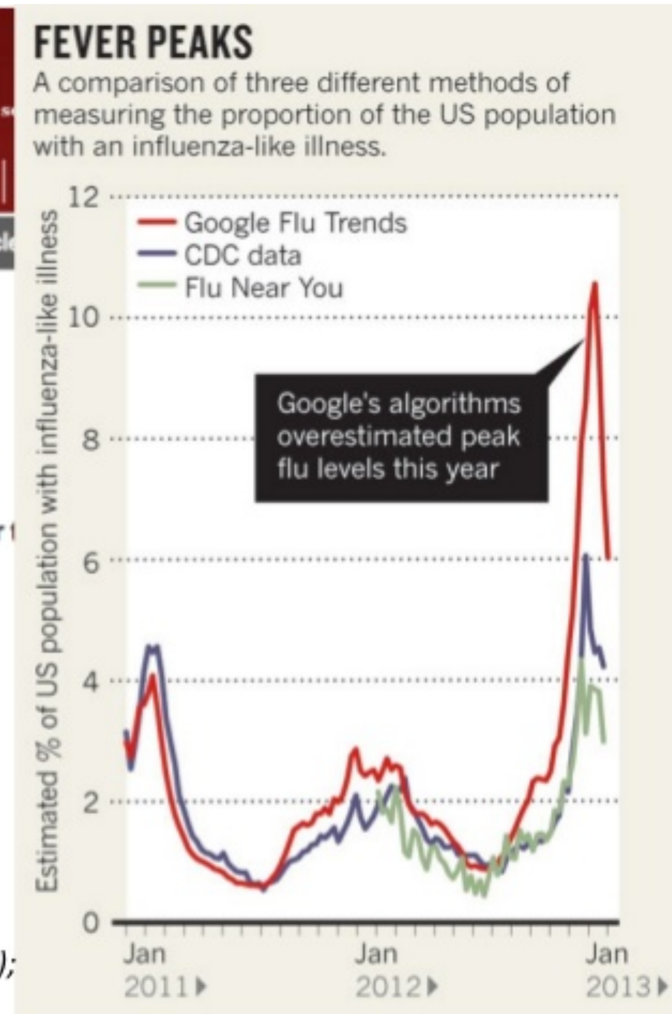
عربي

When Google got flu wrong

US outbreak foxes a leading web-based method for t

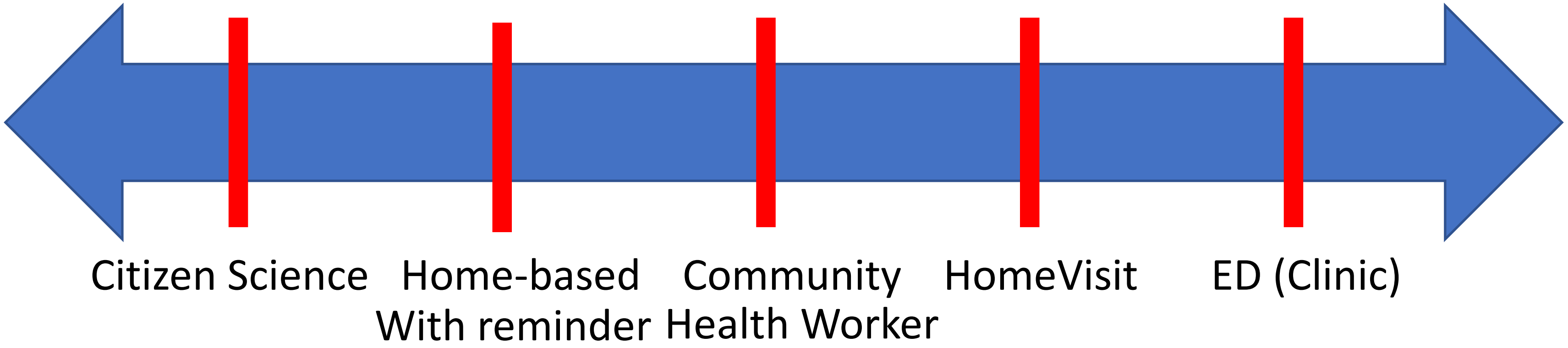
Declan Butler

13 February 2013



SOURCES: GOOGLE FLU TRENDS (WWW.GOOGLE.ORG/FLUTRENDS); CDC; FLU NEAR YOU

Types of Healthcare-facilitated Data Collection



Aliapoulios et al. (in prep) 2017

From the Crowd to the Clinic

Study	Location	Num. Observations (positive)	Symptoms Recorded	Design
NYUMC	New York	278 (23)	cough, diarrhea, fatigue, fever, headache, muscle, nausea, sorethroat, vomit	Clinical (emergency room)
GoViral	New York	899 (201)	body aches, chills, cough, diarrhea, fatigue, fever, leg pain, nausea, runnynose, shortness of breath, sorethroat, vomit	Citizen science
Nigeria	Ilorin, Nigeria	98 (23)	body ache, chills, cough, fever, nausea, runnynose, shortness of breath, sorethroat, vomit	Health-worker facilitated
Flu Watch	United Kingdom	2759 (844)	fever, cough, sorethropat, runnynose, blockednose, sneeze, diarrhea, muscle, headache, rash, earache, wheezy, chills, joint aches, loss of appetite, fatigue, vomit, nausea	Health-worker facilitated
Hong Kong	Hong Kong	4379 (917)	cough, fever, headache, muscle, phlegm, runnynose, sorethroat	Secondary infections recorded by a health worker
Hutterite	Canada	1897 (628)	blockednose, chills, cough, earache, fatigue, fever, headache, muscle, runnynose, sorethroat	Health-worker facilitated

Transfer Learning Paradigm

$$\mathcal{D} = \{(\mathbf{x_{j_i}}, y_{j_i}) \mid \mathbf{x_{j_i}} \in \mathcal{X}_j, y_j\}_{i=1}^{n_j}$$

$$y(\mathbf{x_i}) = (1 + \exp -(b_0 + \mathbf{w}\mathbf{x_i}))^{-1}$$

1. Blind transfer $\mathbf{y} = y_v, \mathcal{X} = \mathcal{X}_v$
2. Additive transfer $\mathbf{y} = [y_u; y_v], \mathcal{X} = [\mathcal{X}_u; \mathcal{X}_v]$
3. Projection on latent space (tbd)

Performance so far…

Study	Nigeria	Hong Kong	Hutterite	GoViral	FluWatch	NYUMC
Nigeria	0.56, 0.56	0.59*, 0.65	0.50, 0.56*	0.59, 0.65	0.50*, 0.56	0.50*, 0.50†
Hong Kong	0.68†, 0.81	0.82, 0.82	0.55, 0.67	0.79*, 0.77†	0.55, 0.61	0.50, 0.68
Hutterite	0.53*, 0.50†	0.54*, 0.52†	0.55, 0.55	0.53*, 0.47	0.51*, 0.51†	0.50*, 0.50†
GoViral	0.68, 0.75†	0.79*, 0.78†	0.53, 0.55	0.79, 0.79	0.53, 0.57	0.50, 0.57
Flu Watch	0.52*, 0.43	0.54*, 0.53†	0.51*, 0.56†	0.53*, 0.55†	0.56, 0.56	0.51*, 0.52†
NYUMC	0.50, 0.81†	0.68, 0.67	0.68, 0.68	0.57, 0.85†	0.63, 0.85†	0.86, 0.86

Aliapoulios et al. (in prep) 2017

Citizen-Sourced Data for Public Health Modeling

- Public health intro and data overview
- Crowdsourcing and knowledge generation in public health
- Learning spatio-temporal features
- Other data opportunities

Acknowledgements

- Jason Liu
- Tom Huang
- Nabeel Rehman
- Kunal Relia
- Anas Elghafari
- Vishwali Mhasawade
- Mohammad Akbari, PhD

NSF NIH USDA IGC International Growth Centre

Citizen-Sourced Data for Public Health Modeling

- Public health intro and data overview
- Crowdsourcing and knowledge generation in public health
- Learning spatio-temporal features
- Other data opportunities

Questions?

rumi.chunara@nyu.edu