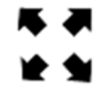
To view an archived recording of this presentation please click the following link:

http://pho.adobeconnect.com/pqulnvzhr0kf/

Please scroll down this file to view a copy of the slides from the session.

Helpful tips when viewing the recording:

- The default presentation format includes showing the "event index". To close the events index, please click on the following icon  and hit "close"
- If you prefer to view the presentation in full screen mode, please click on the following icon  in the top right hand corner of the share screen

Artificial Intelligence and Machine Learning for Public Health

Laura Rosella
Stacey Fisher
Melodie Song
October 29, 2020

Laura C. Rosella, PhD
Associate Professor and Canada Research Chair in Population Health Analytics,
Dalla Lana School of Public Health, University of Toronto
Site Director, ICES University of Toronto
Faculty Affiliate, Vector Institute

Stacey Fisher, PhD
CIHR Health System Impact Fellow in Equitable Artificial Intelligence
Post-doctoral Fellow, Public Health Ontario
Post-doctoral Fellow, Dalla Lana School of Public Health, University of Toronto

Melodie Song, PhD
CIHR Health System Impact Fellow in Equitable Artificial Intelligence
Post-doctoral Fellow, Public Health Ontario
VICTOIRE Post-doctoral Fellow, Dalla Lana School of Public Health, University of Toronto

Public Health Ontario | Santé publique Ontario

Population Health Analytics Laboratory

Dalla Lana School of Public Health

Supporting public health applications of artificial intelligence to improve **health equity** and **prevent chronic diseases** in the population

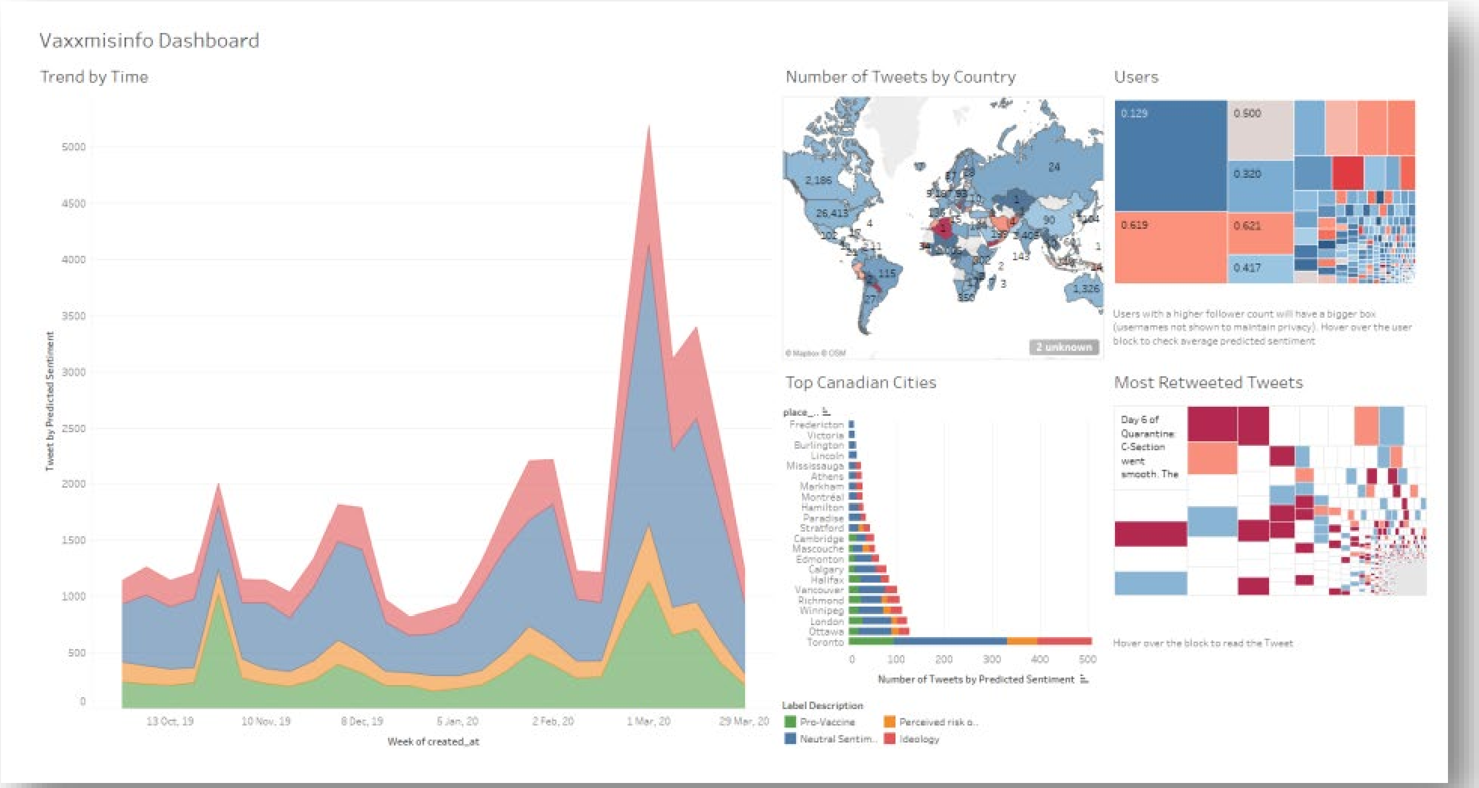
Inform the development of an AI strategy for PHO

Support public health units in building AI/ML capacity

Develop two new risk prediction models using AI/ML methods

HSIF Equitable AI for public health projects (2019-2020)

- Scoping review on AI use for immunization
- Deliberative expert panels
- Human-in-the-loop ML dashboard to detect vaccine misinformation on Twitter



DISCLOSURES

- None of the presenters at this session have received financial support or in-kind support from a commercial sponsor.

- None of the presenters have potential conflicts of interest to declare.

Poll Question

Poll – How much do you know about artificial intelligence and machine learning?

A) Pretty much nothing

B) A little bit, but I’ve never used it

C) I have some experience

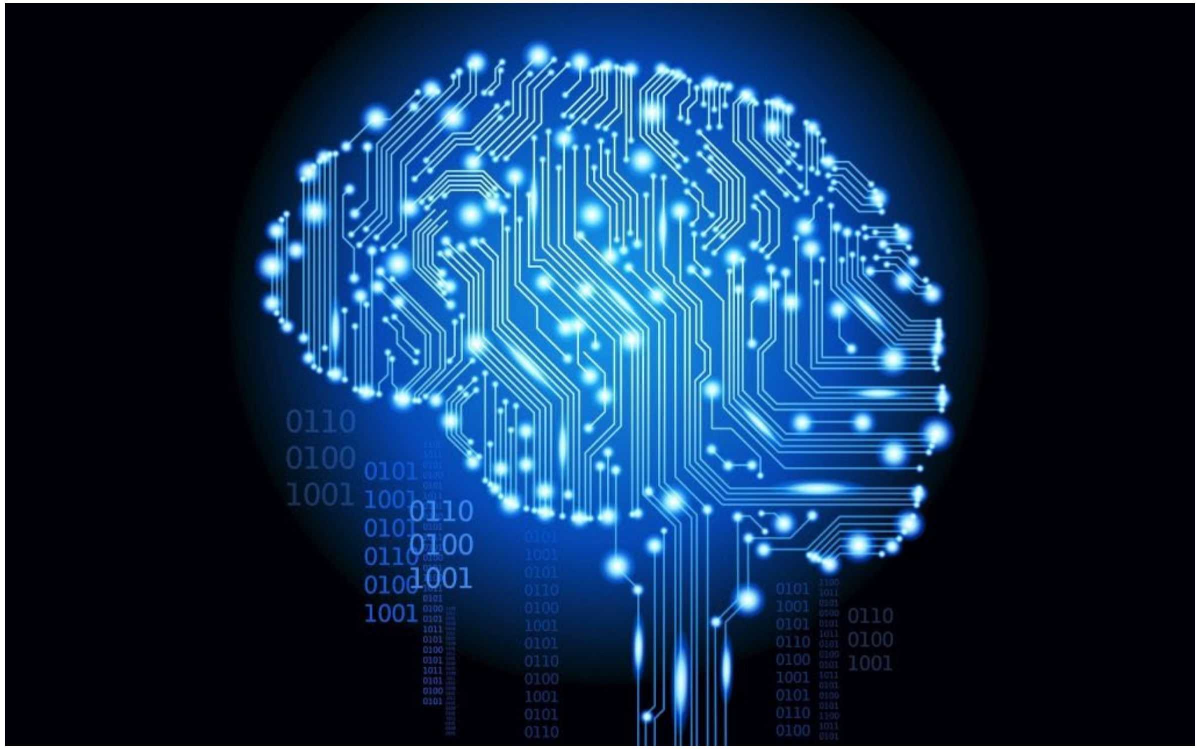
D) I have a lot of experience

What is Artificial Intelligence?

AI is the science of making machines do things that would require **intelligence** if done by people. It is an umbrella term that includes:

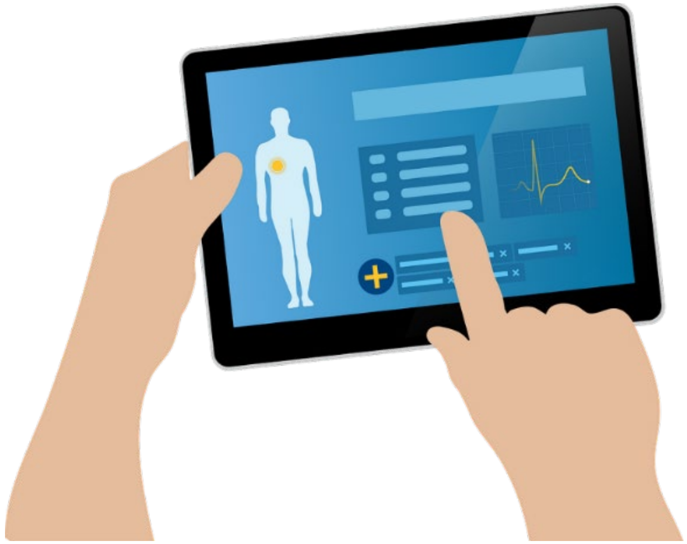
- Machine learning
- Natural language processing
- Deep learning
- Image processing
- Robotics
- Other things

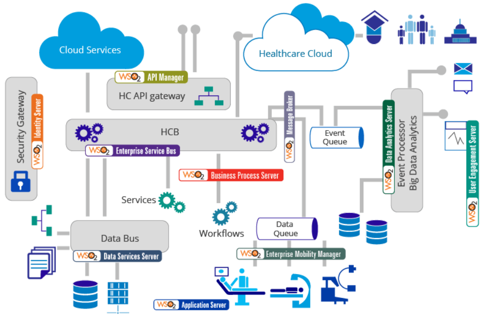
What is fueling the rapidly growing interest in Artificial Intelligence



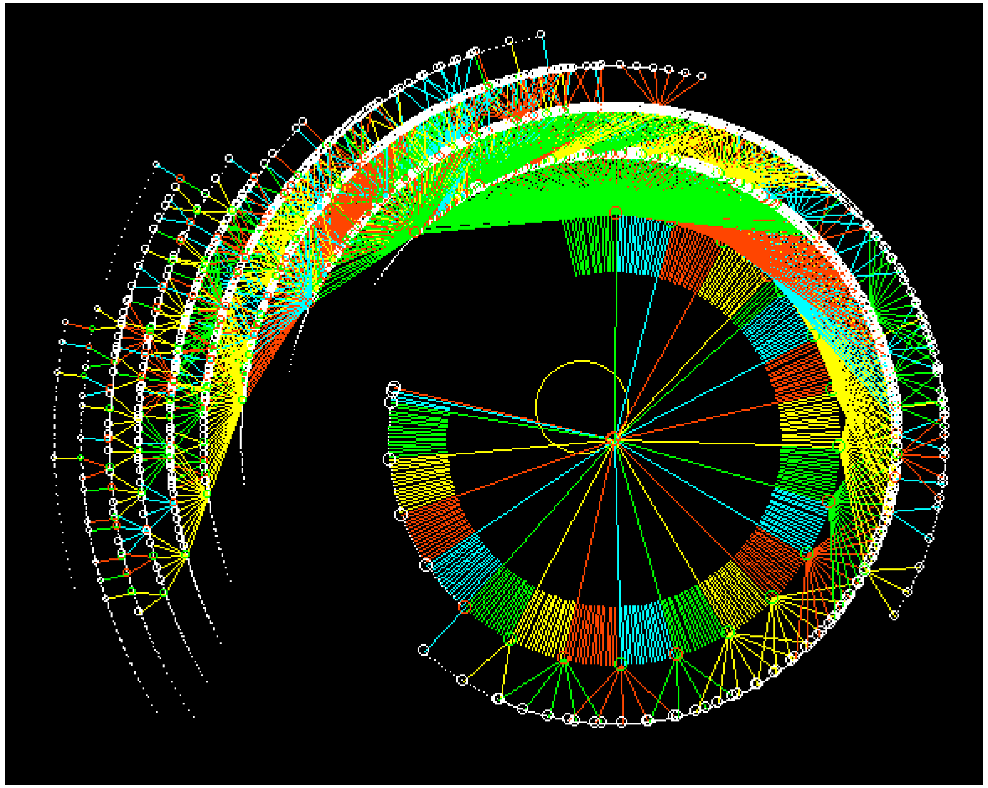
http://pophealthanalytics.com/

Rapidly evolving data environment

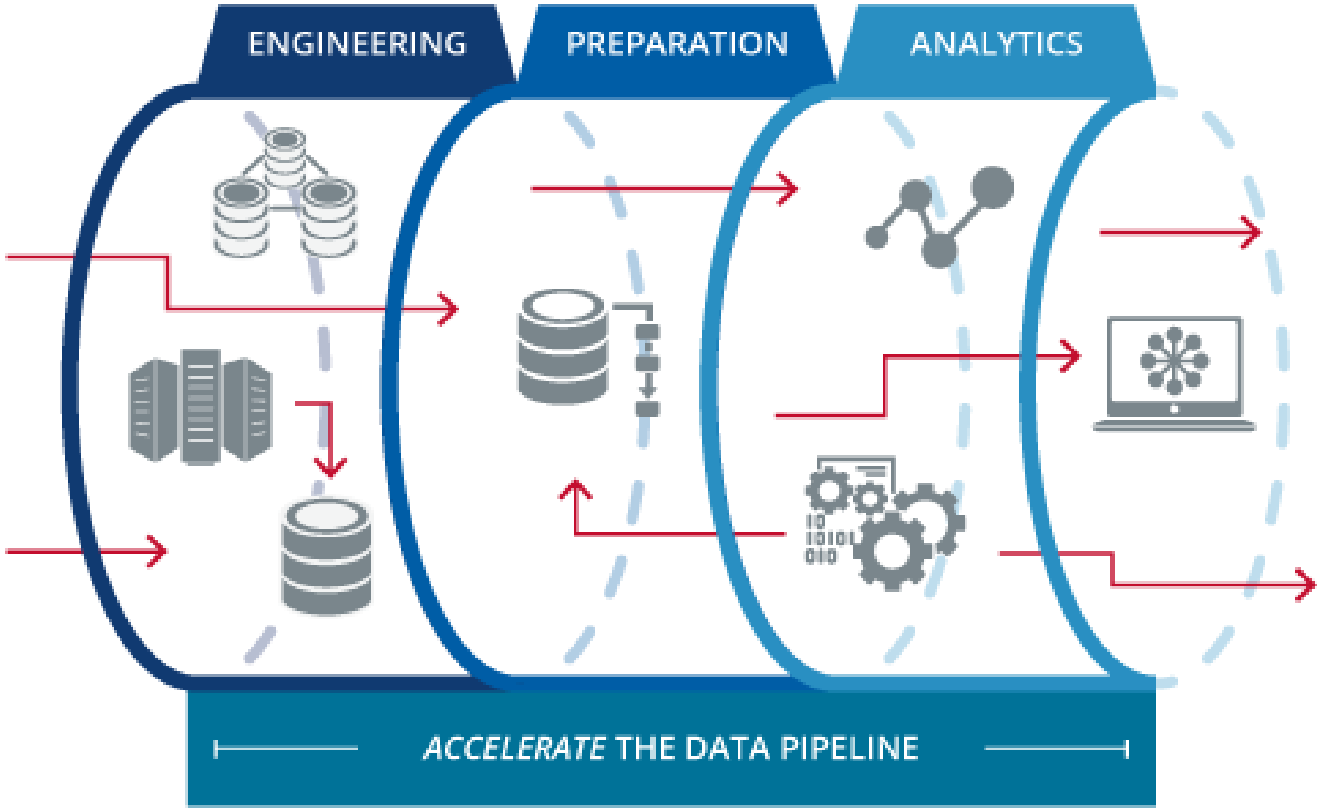




Increasing computational capacity



Improvements in data ingestion and processing



Greater demand for data-driven decisions

CONCETUAL CLARITY IN APPLICATIONS

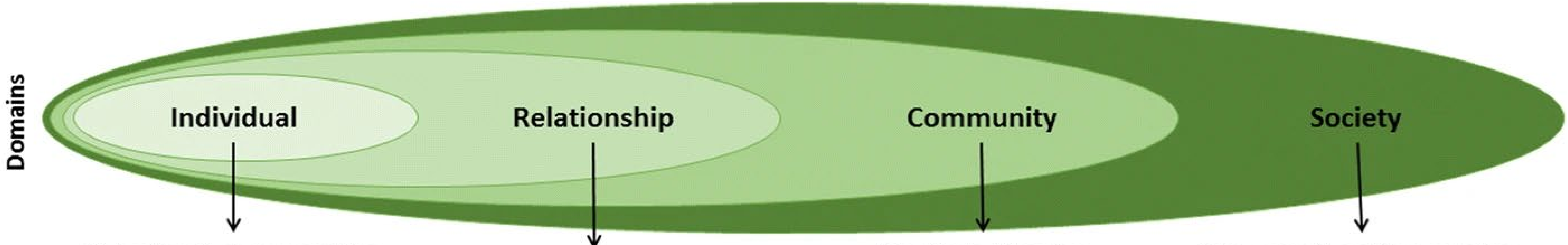
Description

Prediction

- Explain the model
- Operationalize the model

Causal Inference

Hernan et al. A second change to get causal inference right: A classification of data science tasks. CHANCE, 2019; 32:1.



	Behavioral, demographics, genetics, health, lifestyle...	*Family, school, sport, work...*	*Hospitals, libraries, neighborhood, parks...*	*City council, welfare services, state legislation...*
Data Types	Electronic health records Mobile apps Shopping receipts Social media posts Questionnaires Wearable techs	School records Social media networks Employment records	Area deprivation index, Crime rates, Food deserts, Green areas, Income, Pollution levels, Walkability	Gun control, Social security, Unemployment wage, Universal healthcare
Example Sources	Agency for Healthcare Research and Quality, National Health and Nutrition Examination Survey	Facebook, Instagram, National Center for Education Statistics, Twitter	American Community Survey, Department of Housing and Urban Development, Esri Demographics, Food Access Atlas, National Center for Environmental Protection, Terra Modis	CDC, CNN, FOX, White House

Prosperi et al 2018. Big data hurdles in precision medicine and precision public health. BMC Med Inform Decis Mak 2018; 18(1): 139.

What is Machine Learning?

Machine learning is a branch of computer science which enables computers to **learn** without being directly programmed

A term for techniques that fit models algorithmically by adapting to patterns in data

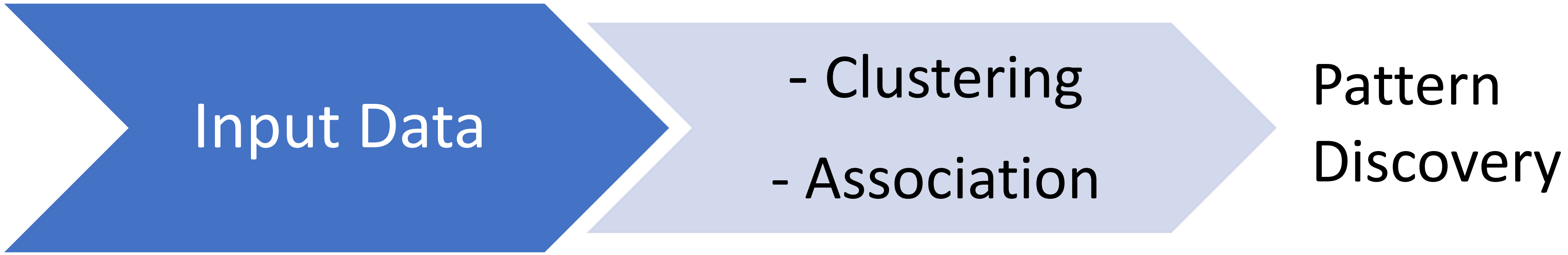
Machine Learning offers epidemiologists new tools to tackle problems for which **classical methods are not well-suited**

Computer Science Term	Epidemiology Term
Features	Independent variables
Labels	Outcomes
Noisy labels	Measurement error
Learning	Fitting
Data mining	Exploratory analyses
Classification algorithm	Algorithm with a categorical outcome
Regression algorithm	Algorithm with a continuous outcome
Precision	Positive predictive values
Recall	Sensitivity
Dimensionality	Number of covariates
Imbalanced data	Unequal outcome distribution
One-hot encoding	Creation of dummy variables

Qifang et al. What is machine learning? A primer for the epidemiologist. AJE, 2019. [Epub ahead of print]

Supervised Learning

Unsupervised Learning



Semi-supervised Learning

Table 3 Selected machine-learning approaches that have been applied to big data in public health

Approach	Learning type	Usage examples
K-means clustering	Unsupervised	Hot spot detection (4)
Retrospective event detection	Unsupervised	Case ascertainment (34)
Content analysis	Unsupervised	Public health surveillance (38)
K-nearest neighbors clustering	Supervised	Spatiotemporal hot spot detection (132); Clinical outcomes from genetic data; falls from wearable sensors
Naïve Bayes	Supervised	Acute gastrointestinal syndrome surveillance (51)
Neural networks	Supervised	Identifying microcalcification clusters in digital mammograms (100); predicting mortality in head trauma patients (31); predicting influenza vaccination outcome (126)
Support vector machines	Supervised	Diagnosis of diabetes mellitus (11); detection of depression through Twitter posts (27)
Decision trees	Supervised	Identifying infants at high risk for serious bacterial infections (8); comparing cost-effectiveness of different influenza treatments (115); and physical activity from wearable sensors (101)

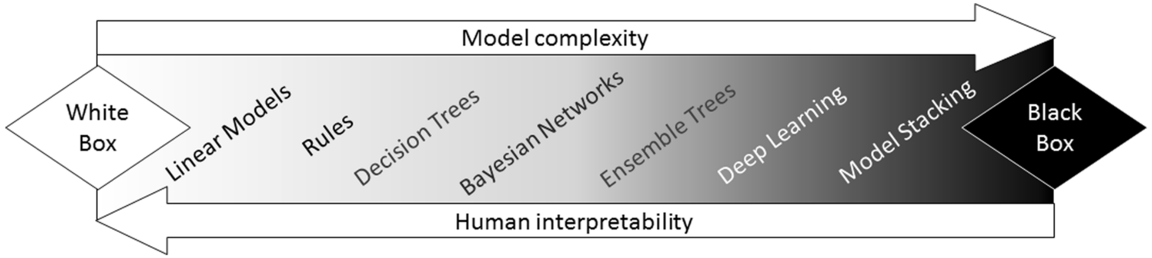
Mooney and Pejaver. Big Data in Public Health: Terminology, Machine Learning, and Privacy. Ann Rev Public Health 2018; 39:95–112

OPPORTUNITIES

- More *quickly* **identify emerging threats** (ex. COVID-19)
- More *detailed* and *up-to-date* understanding of <u>population disease and risk factor distributions</u> (ex. online disease surveillance tools; targeted lead inspections)
- **Forecasting** of disease incidence of population health planning
- Improved <u>targeting</u> of health promotion activities (ex. sentiment analysis; online tools/apps)
- And more (population health management; effects of policy change; causal inference)

CHALLENGES

- **Explainability**
- Bias
- Potential for *increased* **health inequities**
- **Privacy** concerns
- Data access and sharing
- *Outdated* data and analytic **infrastructure**
- *Lack* of AI education and skills within public health



A health care algorithm affecting millions is biased against black patients

A startling example of algorithmic bias

By Colin Lecher | @colinlecher | Oct 24, 2019, 2:00pm EDT

Prosperi et al 2018. Big data hurdles in precision medicine and precision public health. BMC Med Inform Decis Mak 2018; 18(1): 139.

Description

- Public health surveillance
- Association studies

*"Public health surveillance is the continuous, systematic collection, analysis and interpretation of health-related data needed for the planning, implementation, and evaluation of public health practice."**

*Definition from the World Health Organization

New Directions in Artificial Intelligence for Public Health Surveillance

Daniel B. Neill, *Event and Pattern Detection Laboratory, H.J. Heinz III College, Carnegie Mellon University*

Detecting previously unseen outbreaks with novel symptom patterns

Yandong Liu and Daniel B. Neill*

Event and Pattern Detection Laboratory, Carnegie Mellon University, Pittsburgh, PA, USA

- Use of free text from emergency department records to detect, localize, and characterize **newly emerging outbreaks** of disease

Neill. New directions in artificial intelligence for public health surveillance. IEEE Intelligent Systems, 2012; 27(1):56-59.
Liu and Neill. Detecting previously unseen outbreaks with novel symptom patterns. Emerging Health Threats Journal, 2011; 4: 11074.

Text-based spatial event-detection

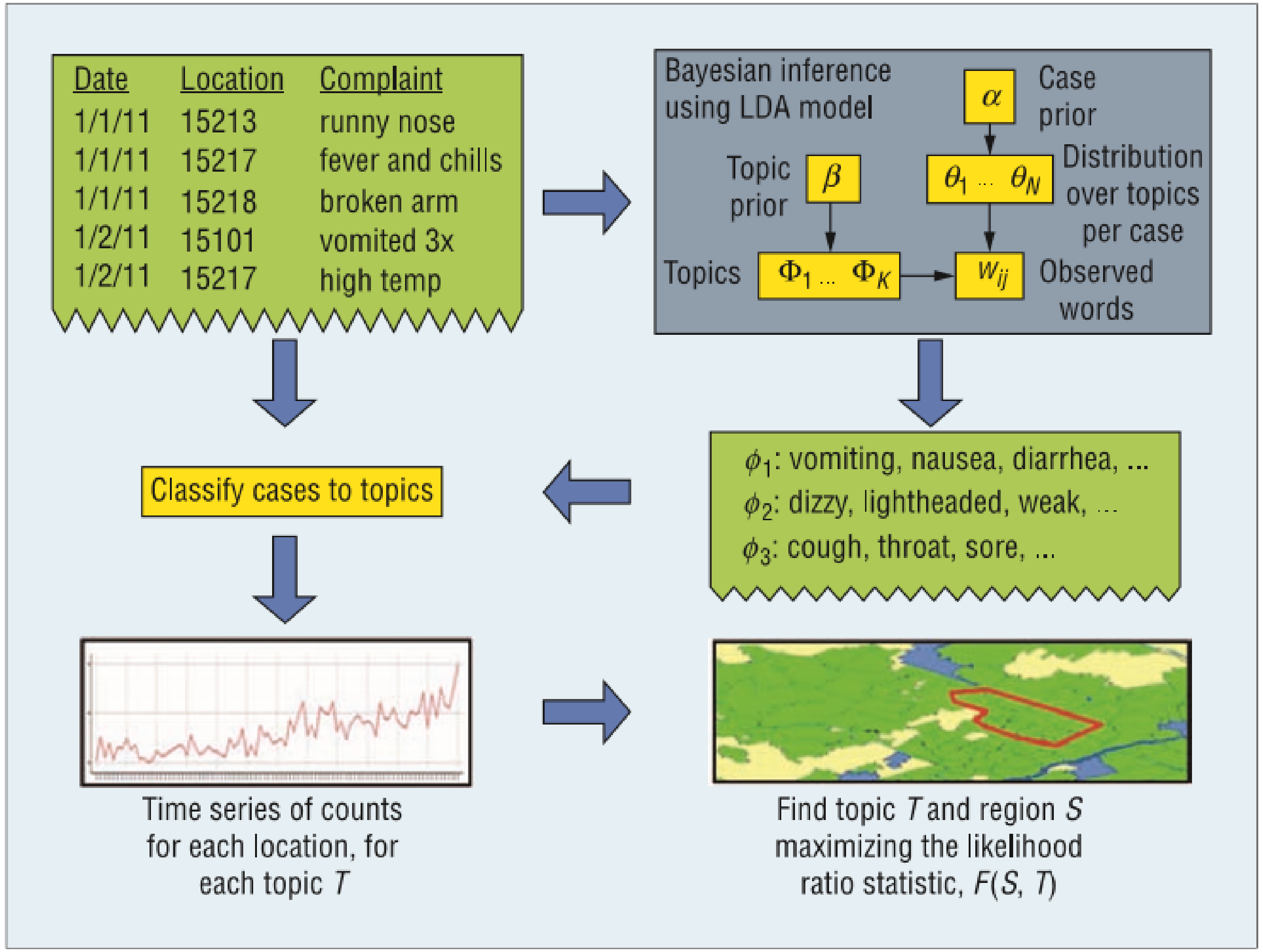


Figure 1. The semantic scan statistic learns a set of topics from the data using Latent Dirichlet Allocation, classifies each case into the most likely topic(s), and then maximizes a likelihood ratio statistic $F(S, T)$ over all topics T and all space-time regions S.

- Traditionally, ED visits are grouped in to “respiratory illness”, “gastrointestinal illness”, “influenza-like illness” etc.
- Instead, define the topics from the data
- For example, *coughing up blood* may be traditionally grouped as “respiratory illness”, **potentially diluting the outbreak signal** and **delaying detection** if there are many of these cases

Neill. New directions in artificial intelligence for public health surveillance. IEEE Intelligent Systems, 2012; 27(1):56-59.

Real-time processing of social media with SENTINEL: A syndromic surveillance system incorporating deep learning for health classification

Ovidiu Șerban[*,1,a], Nicholas Thapen[*,1,a], Brendan Maginnis[a], Chris Hankin[a], Virginia Foot[b]

[a] *Institute for Security Science and Technology, Imperial College London, South Kensington Campus, London SW7 2AZ, UK*
[b] *The Defence Science and Technology Laboratory (DSTL), Porton Down, Salisbury SP4 0JQ, UK*

- Syndromic surveillance using news, clinical and social media data using deep neural networks
- Real-time- processes 1.8 million tweets and 18,000 news articles/day
- NLP, deep neural networks, LASSO

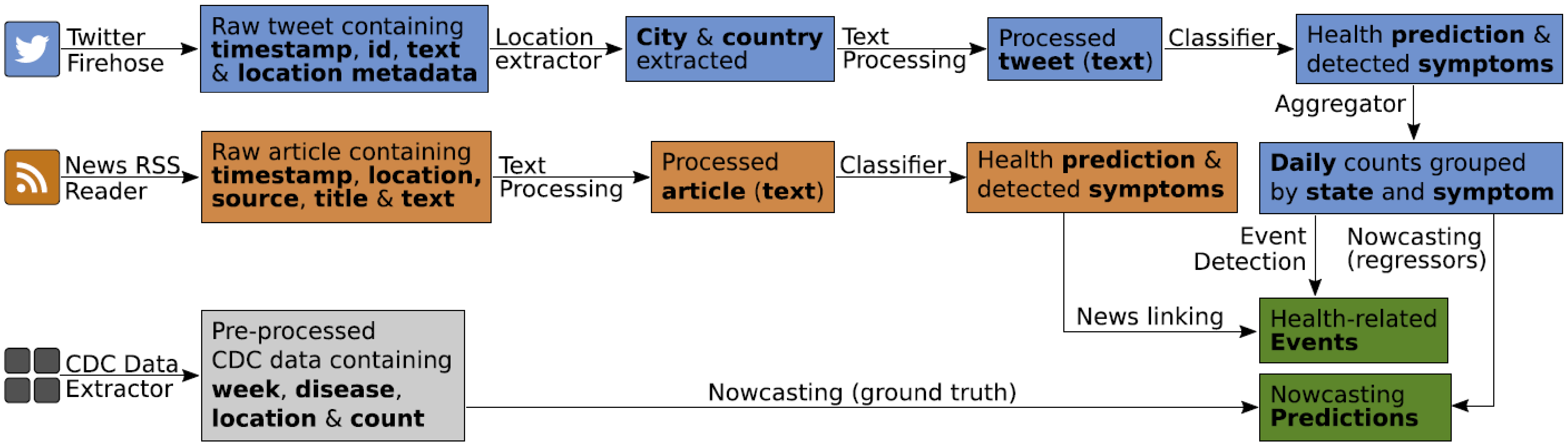
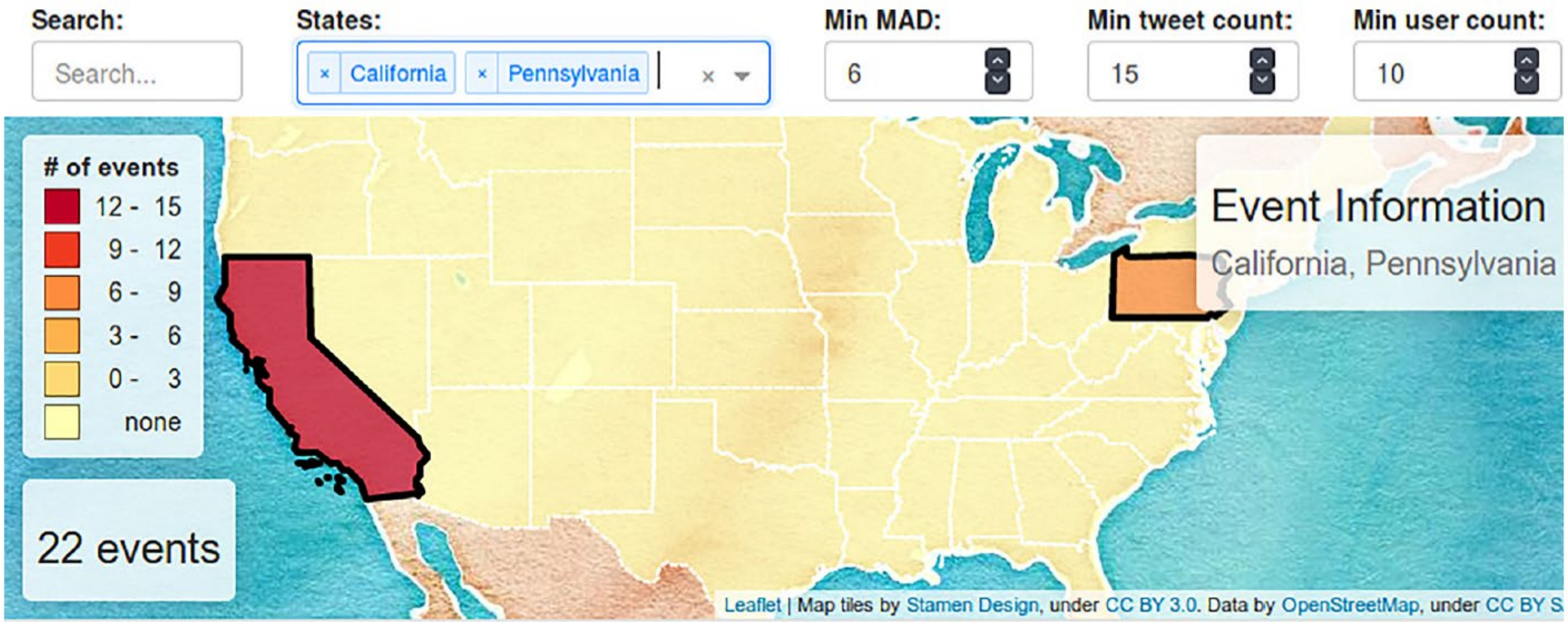


Fig. 1. A data integration diagram, showing the transformation process happening within SENTINEL.

Serban et al. Real-time processing of social media with SENTINEL: A syndromic surveillance system incorporating deep learning for health classification. Information Processing and Management, 2018; 56:1166-1184.



Event ID	Symptom	Country	State	Start Date	End Date	MAD	Tweet Count	User Cour
653	cough	United States	California	2016-10-03	2016-10-03	13.30	60	34
7058	apnea	United States	California	2017-03-02	2017-03-04	12.00	39	33
5647	arthritis	United States	California	2017-02-04	2017-02-05	12.00	32	26
5695	bleeding	United States	California	2017-02-05	2017-02-06	9.80	25	20
545	anxiety	United States	Pennsylvania	2016-10-05	2016-10-06	9.06	112	82

Fig. 3. The event list shown in the system.

Serban et al. Real-time processing of social media with SENTINEL: A syndromic surveillance system incorporating deep learning for health classification. Information Processing and Management, 2018; 56:1166-1184.

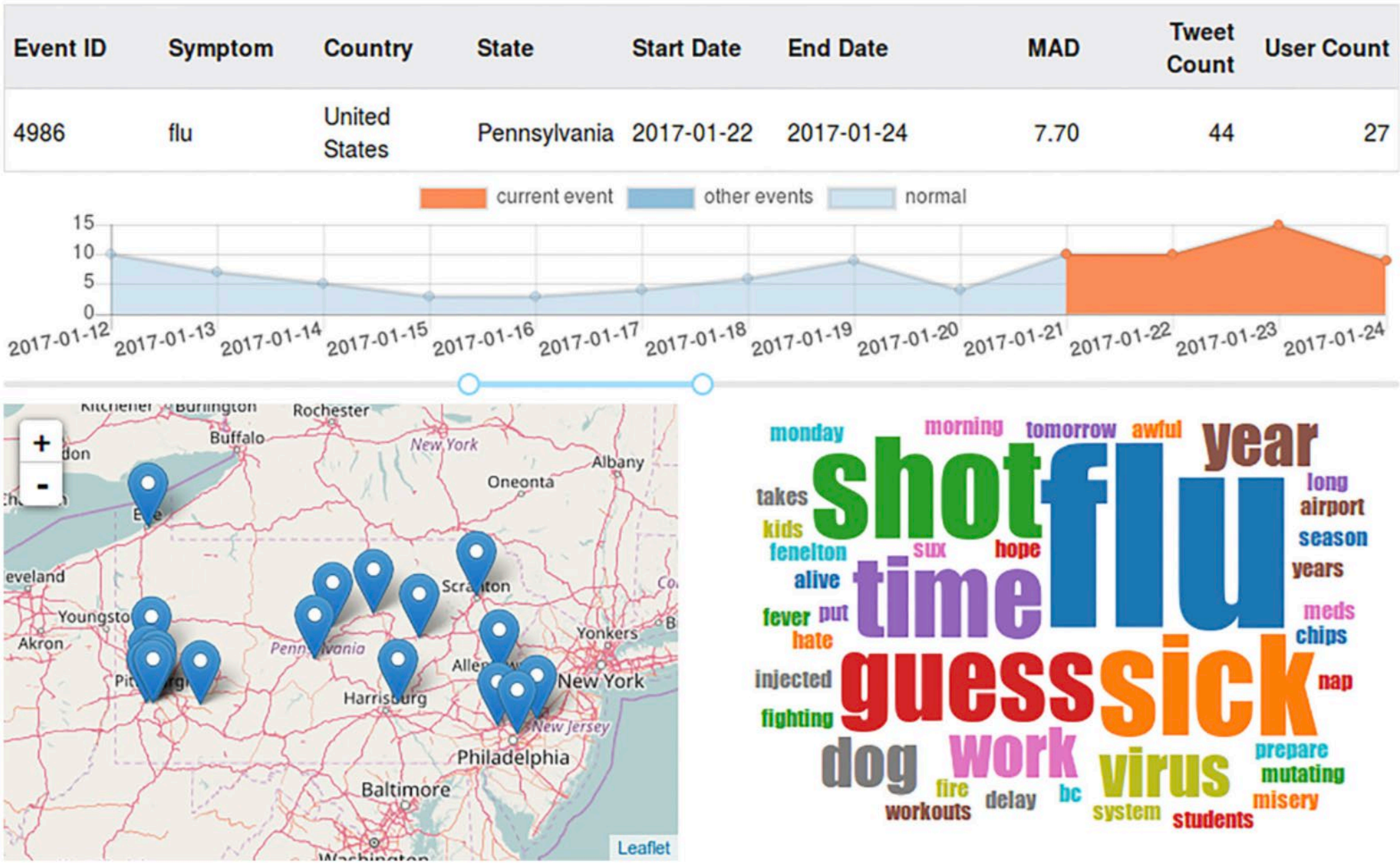


Fig. 4. The Situational Awareness page in the system - top half.

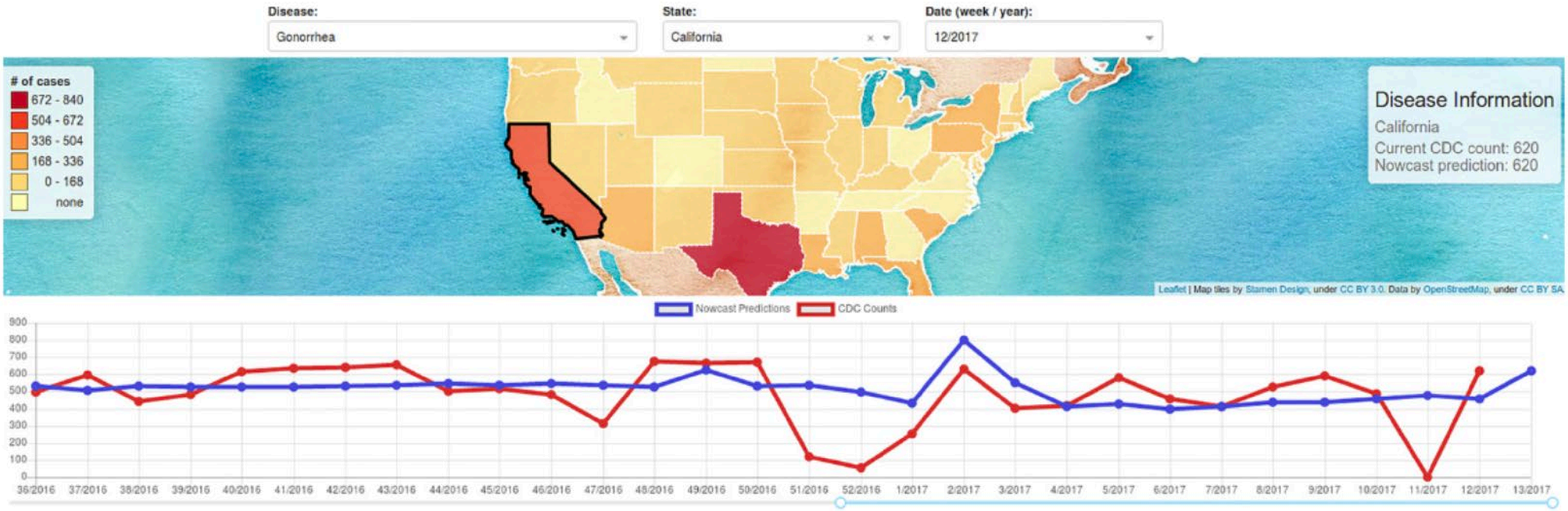


Fig. 6. The Nowcasting screen showing predictions for Gonorrhoea, in California on week 12 (2017).

Prediction

- Data-driven mapping of inputs to outputs; not able to reveal mechanisms or causality
- Disease diagnosis vs. disease incidence; Individual-level vs. population-level prediction
- Overfitting is a concern
- Requires careful model evaluation before deployment

JAMA Network Open

Original Investigation | Pediatrics

Validation of a Machine Learning Model to Predict Childhood Lead Poisoning

Eric Potash, PhD; Rayid Ghani, MS; Joe Walsh, PhD; Emile Jorgensen, MPH; Cortland Lohff, MD, MPH; Nik Prachand, MPH; Raed Mansour, MS

Random forest models derived using data from 1997 to 2012 (N =194,786); temporal validation on 2013 data (N = 6,182)

- Predicted lead exposure in individual children and homes
- Allows inspectors to **prioritize homes and identify children at highest risk**
- Compared to a <u>simple logistic regression model</u> (predictors included housing age, median income, ethnicity, fixed neighborhood effects)

Potash et al. Validation of a machine learning algorithm to predict childhood lead poisoning. JAMA Netw Open 2020; 3(9):e2012734.

DATA

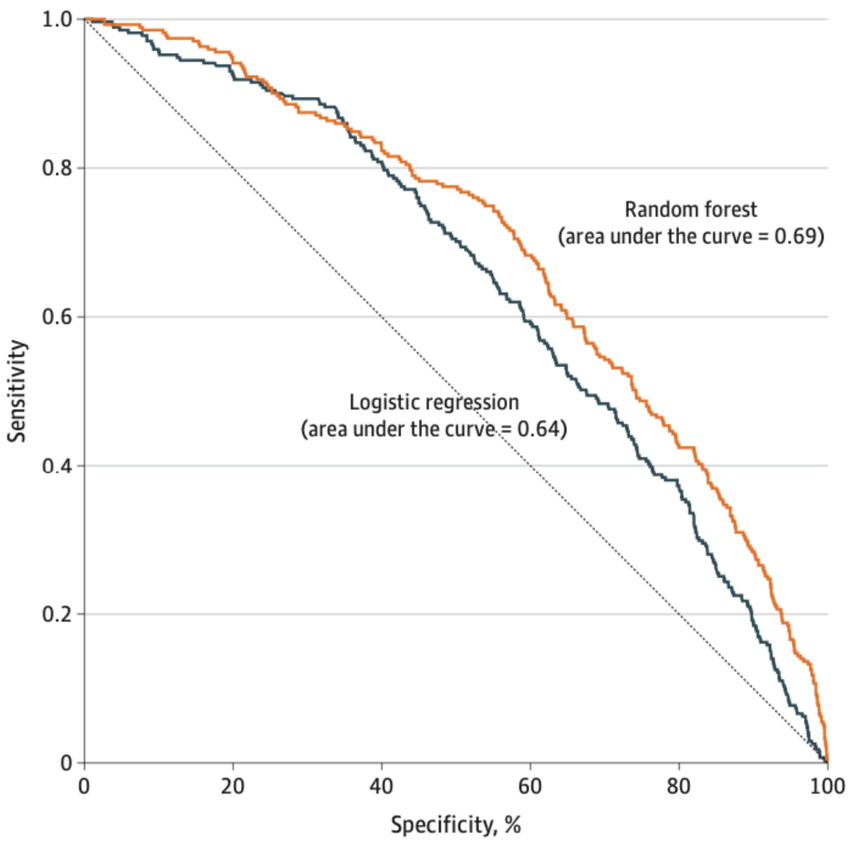
Derived using data from 1997 to 2012 (N =194,786); temporal validation on 2013 data (N = 6,182)

- **Spatial**:
  - Building information including year of construction, size, physical condition
- **Spatial-temporal:**
  - Blood lead level test results (2.5 million)
  - Lead home inspection records (70,000)
  - Chicago Department of Building permits and violations (2 million)
  - Sociodemographic variables at the census track level including education, health insurance, home ownership information

Potash et al. Validation of a machine learning algorithm to predict childhood lead poisoning. JAMA Netw Open 2020; 3(9):e2012734.

RESULTS

Figure 2. Receiver Operating Characteristic Curves for Random Forest and Logistic Regression Models



Difference in the areas under the receiver operating characteristics curve was 0.05 (95% CI, 0.02-0.08).

Potash et al. Validation of a machine learning algorithm to predict childhood lead poisoning. JAMA Netw Open 2020; 3(9):e2012734.

RESULTS

Figure 2. Receiver Operating Characteristic Curves for Random Forest and Logistic Regression Models

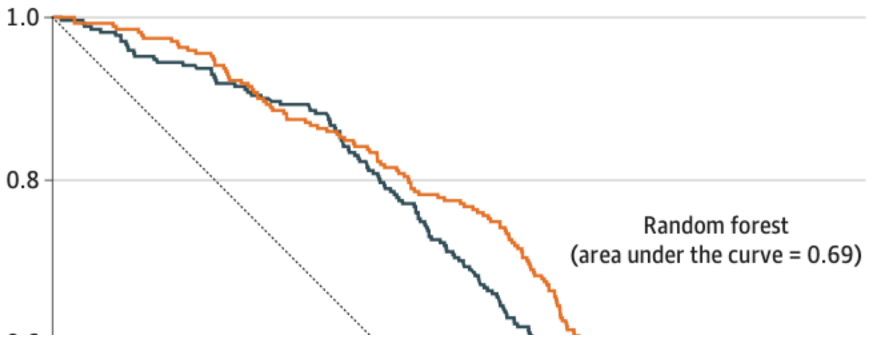
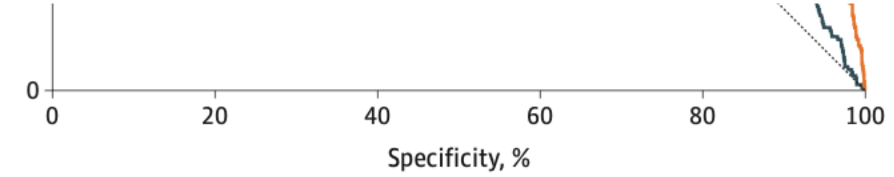


Table 2. Confusion Matrix Metrics for Random Forest and Logistic Regression Models

	Specificity, %			Sensitivity, %			PPV, %		
Population at highest risk, %[a]	Random forest	Logistic regression	Difference (95% CI)[b]	Random forest	Logistic regression	Difference (95% CI)[b]	Random forest	Logistic regression	Difference (95% CI)[b]
5	95.5	95.1	0.4 (0.0 to 0.7)	16.2	8.1	8.1 (3.9 to 11.7)	15.5	7.8	7.7 (3.7 to 11.3)
10	90.4	90.1	0.2 (−0.2 to 0.7)	27.3	19.9	7.4 (3.0 to 14.6)	12.7	9.4	3.3 (1.3 to 6.7)
20	80.3	79.9	0.3 (−0.1 to 1.4)	42.4	38.4	4.1 (−1.1 to 12.5)	9.9	8.9	1.0 (−0.1 to 3.0)



Difference in the areas under the receiver operating characteristics curve was 0.05 (95% CI, 0.02-0.08).

Potash et al. Validation of a machine learning algorithm to predict childhood lead poisoning. JAMA Netw Open 2020; 3(9):e2012734.

RESULTS

eTable 17. Confusion Matrix Metrics for the Random Forest Model by Race/Ethnicity

Specificity

	All	Race/Ethnicity			
		Hispanic	Non-Hispanic Black	Non-Hispanic White	Asian
Highest-risk %[a]					
5%	95.5%	97.1%	92.6%	97.4%	100.0%
10%	90.4%	93.3%	84.8%	97.0%	99.0%
20%	80.3%	84.3%	72.0%	94.0%	93.3%

Sensitivity

	All	Race/Ethnicity			
		Hispanic	Non-Hispanic Black	Non-Hispanic White	Asian
Highest-risk %[a]					
5%	16.2%	7.3%	26.0%	11.1%	0.0%
10%	27.3%	16.1%	38.9%	33.3%	0.0%
20%	42.4%	27.4%	58.8%	33.3%	14.3%

Positive Predictive Value

	All	Race/Ethnicity			
		Hispanic	Non-Hispanic Black	Non-Hispanic White	Asian
Highest-risk %[a]					
5%	15.5%	11.0%	17.5%	12.5%	NA
10%	12.7%	10.6%	13.5%	27.3%	0.0%
20%	9.9%	7.9%	11.3%	15.8%	4.5%

[a] Binary predictions are obtained from continuous risk scores by classifying this highest-risk percentage as positive.

Potash et al. Validation of a machine learning algorithm to predict childhood lead poisoning. JAMA Netw Open 2020; 3(9):e2012734.

IMPLICATIONS

- Home lead inspection prioritization
- Publication of risk scores
- EMR integration
- Landlord outreach

npj | Digital Medicine

www.nature.com/npjdigitalmed

ARTICLE OPEN

Machine-learned epidemiology: real-time detection of foodborne illness at scale

Adam Sadilek[1], Stephanie Caty[2], Lauren DiPrete[3], Raed Mansour[4], Tom Schenk Jr[5], Mark Bergtholdt[3], Ashish Jha[2,6], Prem Ramaswami[1] and Evgeniy Gabrilovich[1]

- Foodborne IllNess DEtector in Real time (FINDER): a machine-learned model for **real-time** detection of foodborne illness using *anonymous* and *aggregated* web search and location data

Sadilek et al. Machine-learned epidemiology: real-time detection of foodborne illness at scale. NPJ Digit Med 2018; 1:36.

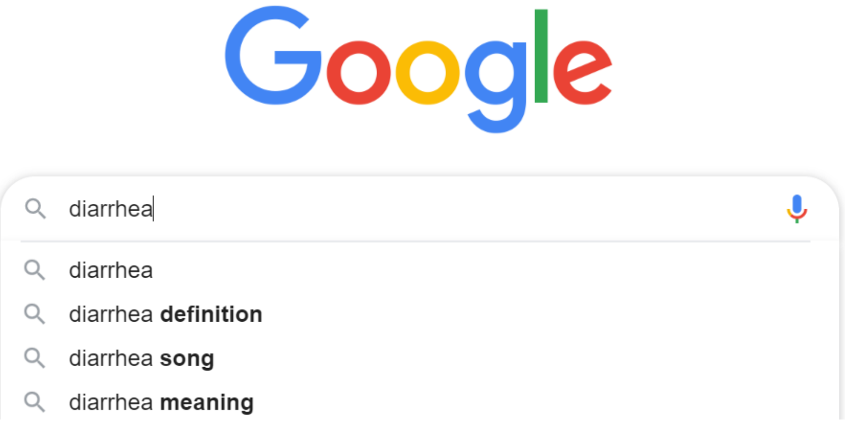
DATA

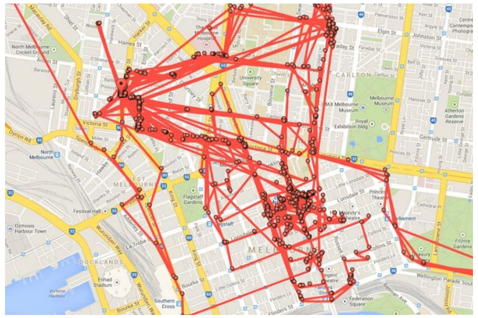
Anonymized web search data

- Using natural language processing and *web pages about foodborne illness*, a model was developed to identify web search queries that are about foodborne illness; queries are given a score between 0 and 1 (c-statistic = 0.85)

Anonymized location data

- *Ambient location data* three days prior to a web search about foodborne illness; this data is automatically collected by Google when opted into location sharing on a mobile device





Sadilek et al. Machine-learned epidemiology: real-time detection of foodborne illness at scale. NPJ Digit Med 2018; 1:36.

APPROACH

- Web search data is used to estimate the probability that a user has foodborne illness symptoms
- Location data is used to estimate the ***proportion* of visits to a restaurant** that were followed by a **web search query related to foodborne** illness, that is used to produce a list of high-risk restaurants
- FINDER was deployed into two local health departments, in **Chicago** and **Las Vegas** from May – August 2016

Sadilek et al. Machine-learned epidemiology: real-time detection of foodborne illness at scale. NPJ Digit Med 2018; 1:36.

RESULTS

Table 1. Number of inspections conducted during the experimental time period

	FINDER	BASELINE
Total	132	10,786
Las Vegas	61	4977
Chicago	71	5809
Complaint-driven	N/A	1291
Routine	N/A	4518
Risk level[a]		
High (% of total)	84 (63.6%)	5702 (52.9%)
Medium (%)	39 (29.6%)	2325 (21.6%)
Low (%)	9 (6.8%)	2759 (25.6%)

[a] *p* value for difference in risk distribution between FINDER and BASELINE <0.001, from X^2-test

Sadilek et al. Machine-learned epidemiology: real-time detection of foodborne illness at scale. NPJ Digit Med 2018; 1:36.

RESULTS

Table 2. Ability of FINDER to detect unsafe restaurants as compared to BASELINE rate and with subcategories of the baseline inspections, including complaint-based inspections that occurred in Chicago and routine inspections from both Chicago and Las Vegas

	FINDER $n = 132$	BASELINE $n = 10,786$	Odds ratio[a] [95% CI]	*p*-value
Overall, number unsafe (%)	69 (52.3%)	2662 (24.7%)	3.06 [2.14–4.35]	<0.001
Risk level				
High, number unsafe (%)	42 (50.0%)	1909 (33.5%)	1.98 [1.28–3.05]	0.002
Medium, number unsafe (%)	23 (59.0%)	536 (23.1%)	5.50 [2.83–10.72]	<0.001
Low, number unsafe (%)	4 (44.4%)	217 (7.9%)	7.35 [1.79–30.13]	0.006
Comparison of FINDER to complaint-based inspections				
	FINDER $n = 71$	**COMPLAINT** $n = 1291$		
Overall, number unsafe (%)	37 (52.1%)	508 (39.4%)	1.68 [1.04–2.71]	0.03
Risk level				
High, number unsafe (%)	27 (47.4%)	374 (39.4%)	1.38 [0.81–2.36]	0.24
Medium, number unsafe (%)	9 (75.0%)	115 (39.3%)	4.64 [1.23–17.51]	0.02
Low, number unsafe (%)	1 (50.0%)	19 (38.8%)	1.58 [0.09–26.78]	0.75

Sadilek et al. Machine-learned epidemiology: real-time detection of foodborne illness at scale. NPJ Digit Med 2018; 1:36.

IMPLICATIONS

- More efficient restaurant inspection and reduced foodborne illness
- More precise identification of
- However, small sample size and privacy concerns need to carefully considered

Sadilek et al. Machine-learned epidemiology: real-time detection of foodborne illness at scale. NPJ Digit Med 2018; 1:36.

Causal Inference

- Most AL/ML methods are being used for descriptive and predictive purposes, however, *we can apply the same tools for causal purposes*
- AI/ML as it is currently being used is not necessarily changing our conceptual understanding of causal paradigms – instead, it is bringing another <u>tool</u> to help us explore causality

The impact of social housing on mental health: longitudinal analyses using marginal structural models and machine learning-generated weights

Rebecca Bentley,[1,2*] Emma Baker,[2] Koen Simons,[3] Julie A Simpson[3†] and Tony Blakely[3,4†]

- Investigated the effect of **cumulative exposure to social housing and social housing transitions on mental health** using health survey data
- Used marginal structural models with IPTW generated using ensemble learning methods
    - Ensemble methods combine several machine learning models

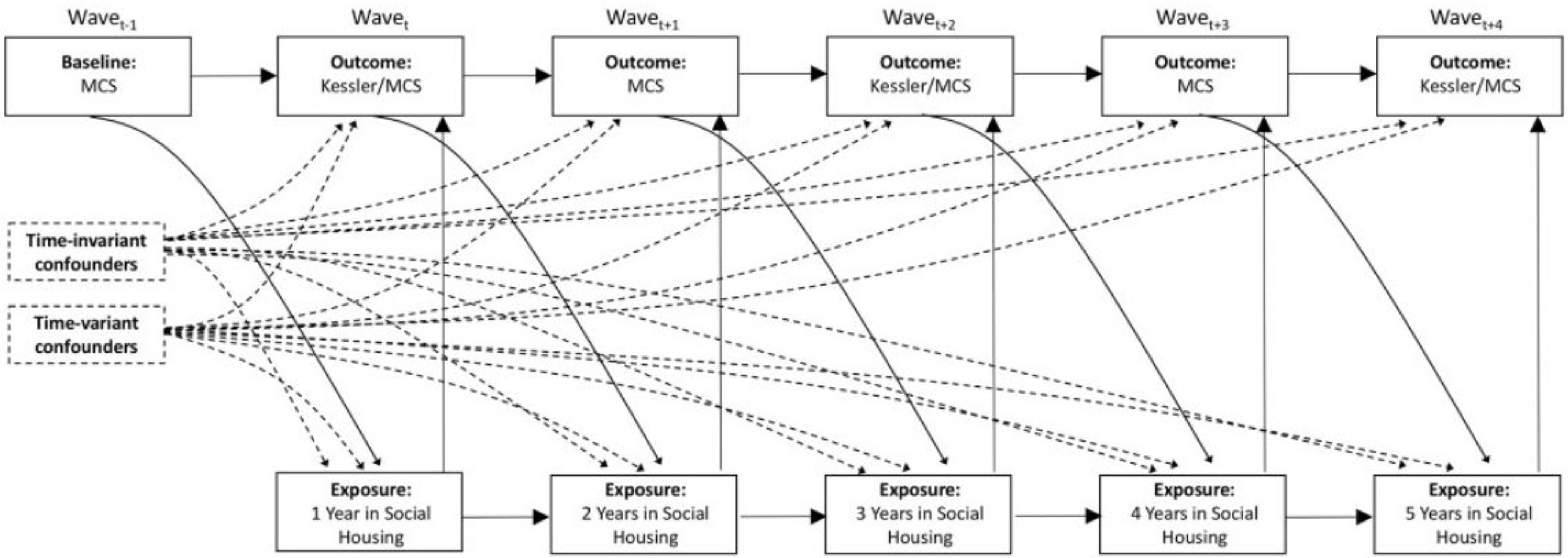


Figure 1. Directed acyclic graph (DAG).

- IPTW included many baseline and time-varying confounders, as well as mental health score from the previous year
- People with continuous exposure to social housing had worse mental health on average than people without

Bentley et al. The impact of social housing on mental health: longitudinal analyses using marginal structural models and machine learning-generated weights. 2018, IJE. 1414-1422.

5 types of big data

Source	Examples	Aspect of bigness[a]	Key technical issues	Typical uses
-omic/biological	Whole exome profiling, metabolomics	Wide	Lab effects, informatics pipeline	Etiologic research, screening
Geospatial	Neighborhood characteristics	Wide	Spatial autocorrelation	Etiologic research, surveillance
Electronic health records	Records of all patients with hypertension	Tall, often also wide	Data cleaning, natural language	Clinical research, surveillance
Personal monitoring	Daily GPS records, Fitbit readings	Tall	Redundancy, inference of intentions	Etiologic research, potentially clinical decision making
Ready-made data	Google search results, Reddit	Tall	Selection biases, natural language	Surveillance, screening, identification of hidden social networks

Annual Review of Public Health 39, no. 1 (April 2018): 95–112. https://doi.org/10.1146/annurev-publhealth-040617-014208.

Ready-made digital trace data

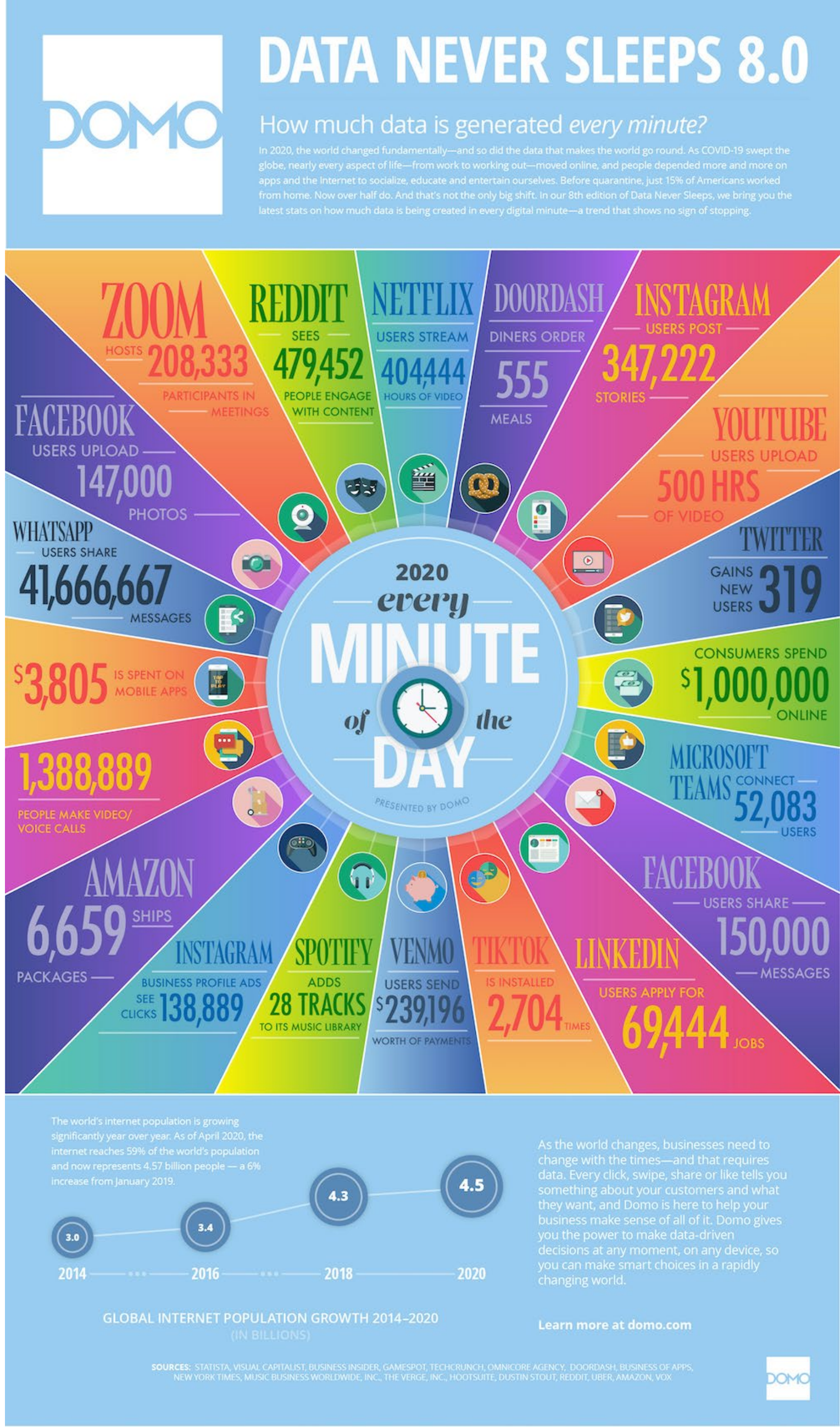
Advantages

- Big
- Always on
- Non-reactive
- Captures social relationships
- Low cost

Disadvantages

- Incomplete
- Inaccessible
- Non-representative
- Drifting
- Algorithmic confounding
- Dirty

Salganik, Matthew J. “Bit by Bit: Social Research in the Digital Age”. *Princeton & Oxford: Princeton University Press*. 2017.



Application for non-communicable diseases

Estimating geographic subjective well-being from Twitter: A comparison of dictionary and data-driven language methods

Kokil Jaidka[a,b,1], Salvatore Giorgi[c], H. Andrew Schwartz[d], Margaret L. Kern[e], Lyle H. Ungar[c], and Johannes C. Eichstaedt[f,g,1]

[a]Department of Communications and New Media, National University of Singapore, Singapore 117416; [b]Centre for Trusted Internet and Community, National University of Singapore, Singapore 117416; [c]Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA 19104; [d]Department of Computer Science, Stony Brook University, Stony Brook, NY 11794; [e]Melbourne Graduate School of Education, The University of Melbourne, Parkville, VIC 3010, Australia; [f]Department of Psychology, Stanford University, Stanford, CA 94305; and [g]Institute for Human-Centered Artificial Intelligence, Stanford University, Stanford, CA 94305

Jaidka, Girogi, Schwartz, Kern, Ungar, Eichstaedt. Estimating geographic subjective well-being from Twitter: A comparison of dictionary and data-driven language methods. PNAS, 2020; 117(19): 10165-10171.

Application for communicable diseases

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

Sarah F. McGough[1,2,3]*, **John S. Brownstein**[2,3,4], **Jared B. Hawkins**[2,3,4], **Mauricio Santillana**[2,3,4]*

1 Harvard T.H. Chan School of Public Health, Boston, Massachusetts, United States of America, **2** Computational Health Informatics Program, Boston Children's Hospital, Boston, Massachusetts, United States of America, **3** Computational Epidemiology Group, Division of Emergency Medicine, Boston Children's Hospital, Boston, Massachusetts, United States of America, **4** Department of Pediatrics, Harvard Medical School, Boston, Massachusetts, United States of America

McGough, Sarah F., et al. "Forecasting Zika incidence in the 2016 Latin America outbreak combining traditional disease surveillance with search, sociamedia, and news report data." *PLoS neglected tropical diseases* 11.1 (2017): e0005295. I

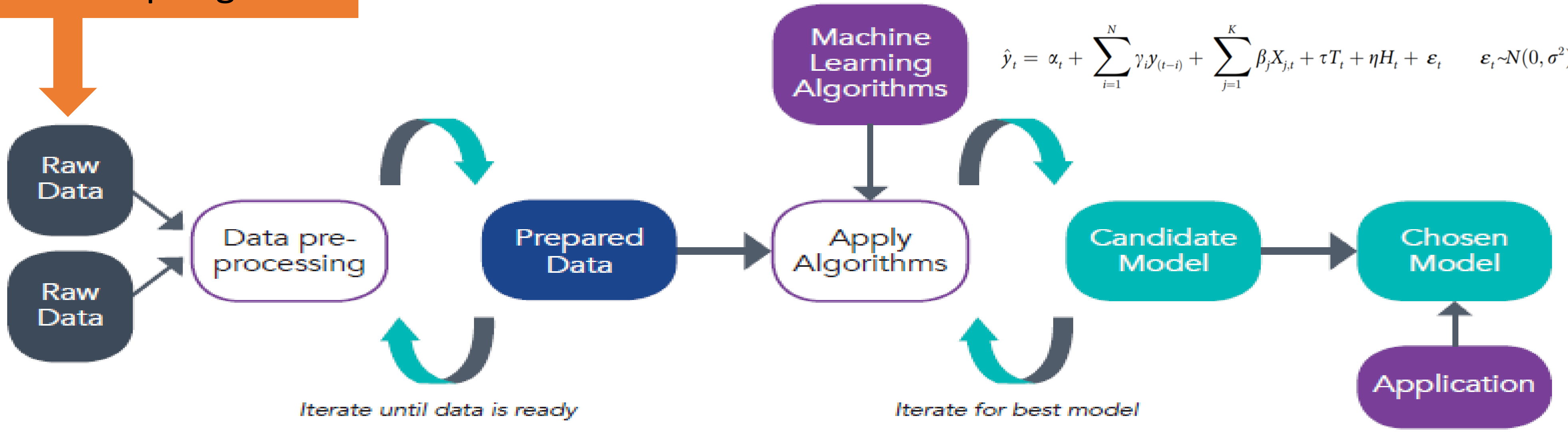
Forecasting Zika incidence using an ensemble ML approach in the Americas

Epidemiological data:
- PAHO
- National MOH data

Digital trace data:
- Twitter
- Google Correlate & Google Trends search
- Healthmap.org

Model	3 week		
	RMSE	rRMSE	ρ
AR	886.701	555.937	-0.903
G+T	**292.718**	**64.733**	**0.355**
ARGO+T	323.089	158.377	0.243
ARGO+TH	335.778	163.436	0.085

$$\hat{y}_t = \alpha_t + \sum_{i=1}^{N} \gamma_i y_{(t-i)} + \sum_{j=1}^{K} \beta_j X_{j,t} + \tau T_t + \eta H_t + \varepsilon_t \qquad \varepsilon_t \sim N(0, \sigma^2)$$



McGough, Sarah F., et al. "Forecasting Zika incidence in the 2016 Latin America outbreak combining traditional disease surveillance with search, social media, and news report data." *PLoS neglected tropical diseases* 11.1 (2017): e0005295.

Applications for mitigating harm from infodemic

Managing the COVID-19 infodemic: Promoting healthy behaviours and mitigating the harm from misinformation and disinformation

Joint statement by WHO, UN, UNICEF, UNDP, UNESCO, UNAIDS, ITU, UN Global Pulse, and IFRC

台灣事實查核中心
Taiwan FactCheck Center

Crowd Detecting, Real-Time Collecting, Automated Clustering

I predict that the next major outbreak – whether of a highly fatal strain of influenza or something else – will not be due to a lack of preventive technologies. Instead, emotional contagion, digitally enabled, could erode trust in vaccines so much as to render them moot. The deluge of conflicting information, misinformation and manipulated information on social media should be recognized as a global public-health threat.

Larson, Heidi J. “The Biggest Pandemic Risk? Viral Misinformation.” *Nature* 562, no. 7727 (October 16, 2018): 309–309.

Opportunities of using ready-made data for AI4PH

Opportunities:

- Open and accessible
- Fast turn-around
- Interdisciplinary

Challenges:

- Data acquisition
- Informed consent
- Privacy
- Making decisions in the face of uncertainty

Wrap-up

Defining the direction for public health

Tension between what we do for the individual versus what we do for the population

- *“There is no need to add the word “precision” to public health”*
- Focus is in the wrong place (not the social determinants of health)

-Chowkanyun, Bayer, Galea, NEJM 2018

- *“Precision public health offers a compelling opportunity to reinvigorate a discipline that has never been more important for advancing the health of our most vulnerable and excluded communities.”*
- -Horton, Lancet, 2018

Will AI/ML advance public health? Only if….

Fewer people developing diseases that are preventable

Promoting health in our environments and communities

Reduction in health inequities

Summary

- **A population health focus means we are interested in a broader set of data sources on the population**
  - AI equity = public health approach

- **The population health questions and focus are consistent; with emerging tools and new data sources there is potential**
  - Increasing linkages to a broad range of data sources that better reflect the determinants of health and new ways to work with these data
  - Increasingly we should be thinking of ways to integrate forward-looking planning tools at the population level
  - More strategic about matching action to population health needs and speeding up data-action cycle

- **AI/ML methods can be applied multiple ways and for different purposes - aim for clarity**
  - Clarity about what the problem is and why the data or problem fits best with the method
  - Descriptive/surveillance, prediction and causal inference

- **Issues related to bias continually need attention**
  - Measurement error/data quality
  - Selection bias

Poll Question

If you were in charge, what would you do to facilitate AI/ML use in your setting?

PRIORITIES TO FACILITATE THE USE OF AI BY PUBLIC HEALTH ORGANIZATIONS

1) Understand the *governance* context
2) *Modernized* **data and analytic infrastructure**
3) Use AI best practices, including explicit consideration of **EQUITY**
4) An **educated workforce**
5) Strategic collaborative partnerships

Inform the development of an AI strategy for PHO

Questions and Discussion

Learning Opportunities

Artificial Intelligence for Public Health (AI4PH) Summer Institute

- Postponed to Summer 2021
- www.ai4ph.ca

NeurIPS 2020 Workshop for Machine Learning in Public Health (virtual)

- Saturday December 12, 2020
- https://sites.google.com/nyu.edu/mlph2020

AI4PH

NEURAL INFORMATION PROCESSING SYSTEMS

Learning Resources

- Crash Course: Artificial Intelligence (youtube.com/crashcourse)
- Elements of AI (http://course.elementsofai.com)
- Coursera: Machine Learning (Stanford University) (https://www.coursera.org/learn/machine-learning)
- Social Media Research Toolkit (Ryerson University) (http://socialmediadata.org/social-media-research-toolkit/)

Thank you!

Stacey Fisher, PhD
stacey.fisher@utoronto.ca
@StaceyFisher_

Laura C. Rosella, PhD
laura.rosella@utoronto.ca
@LauraCRosella

Melodie Song, PhD
melodie.song@oahpp.ca
@MelodieYJSong

Public Health Ontario | Santé publique Ontario

Population Health Analytics Laboratory

Dalla Lana School of Public Health