ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Online | Mobile | Global

From Social Networks to Machine Learning: Epidemiology is Going Digital

Marcel Salathé, Digital Epidemiology Lab, EPFL

@marcelsalathe

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

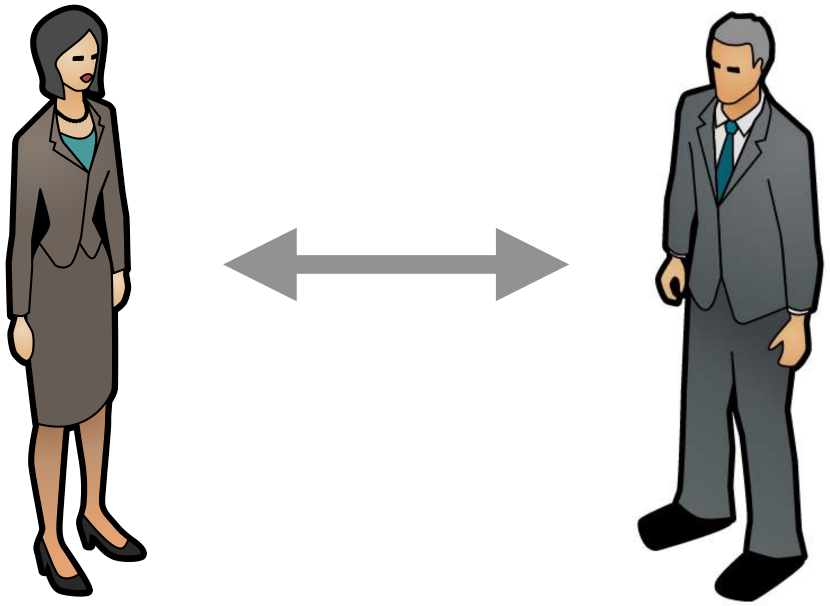
Online | Mobile | Global

Digital Epidemiology

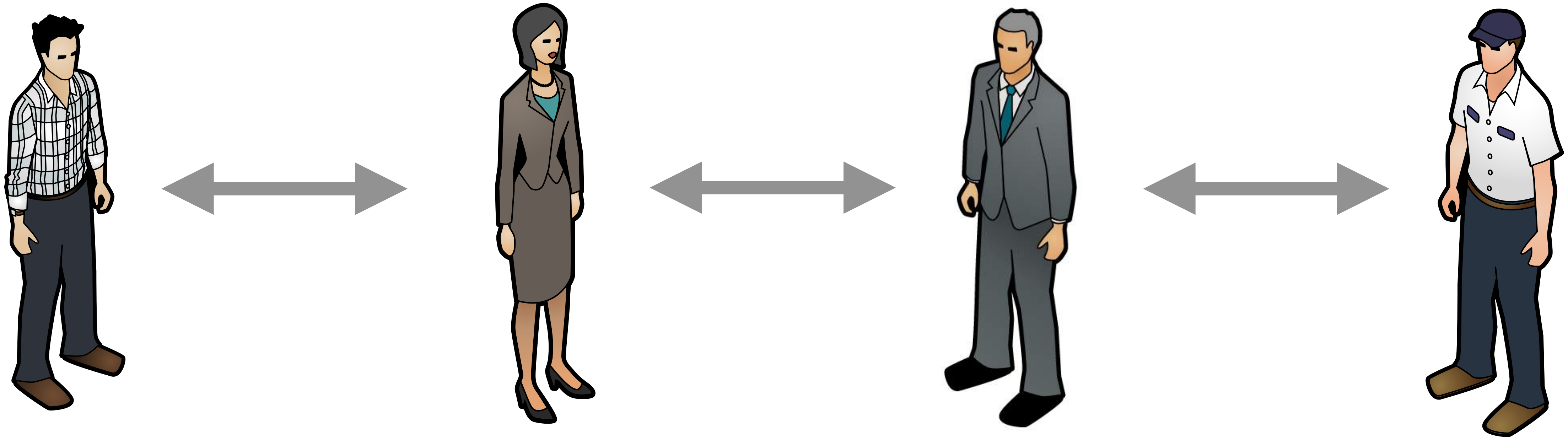
Marcel Salathé, Digital Epidemiology Lab, EPFL

@marcelsalathe

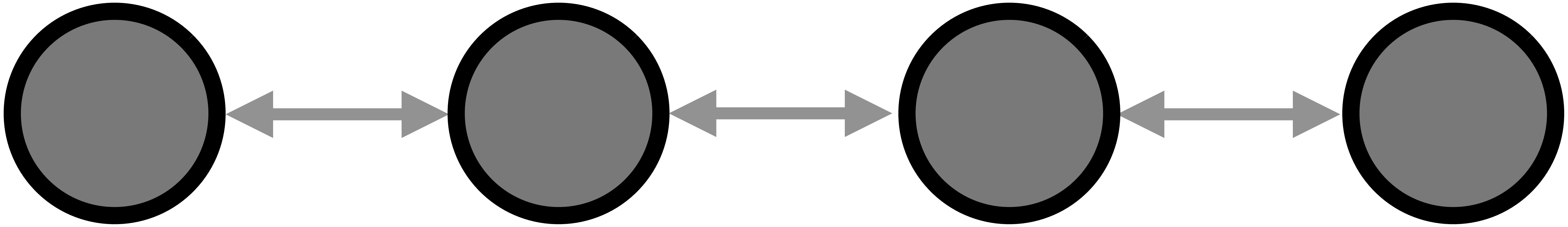
Social Networks

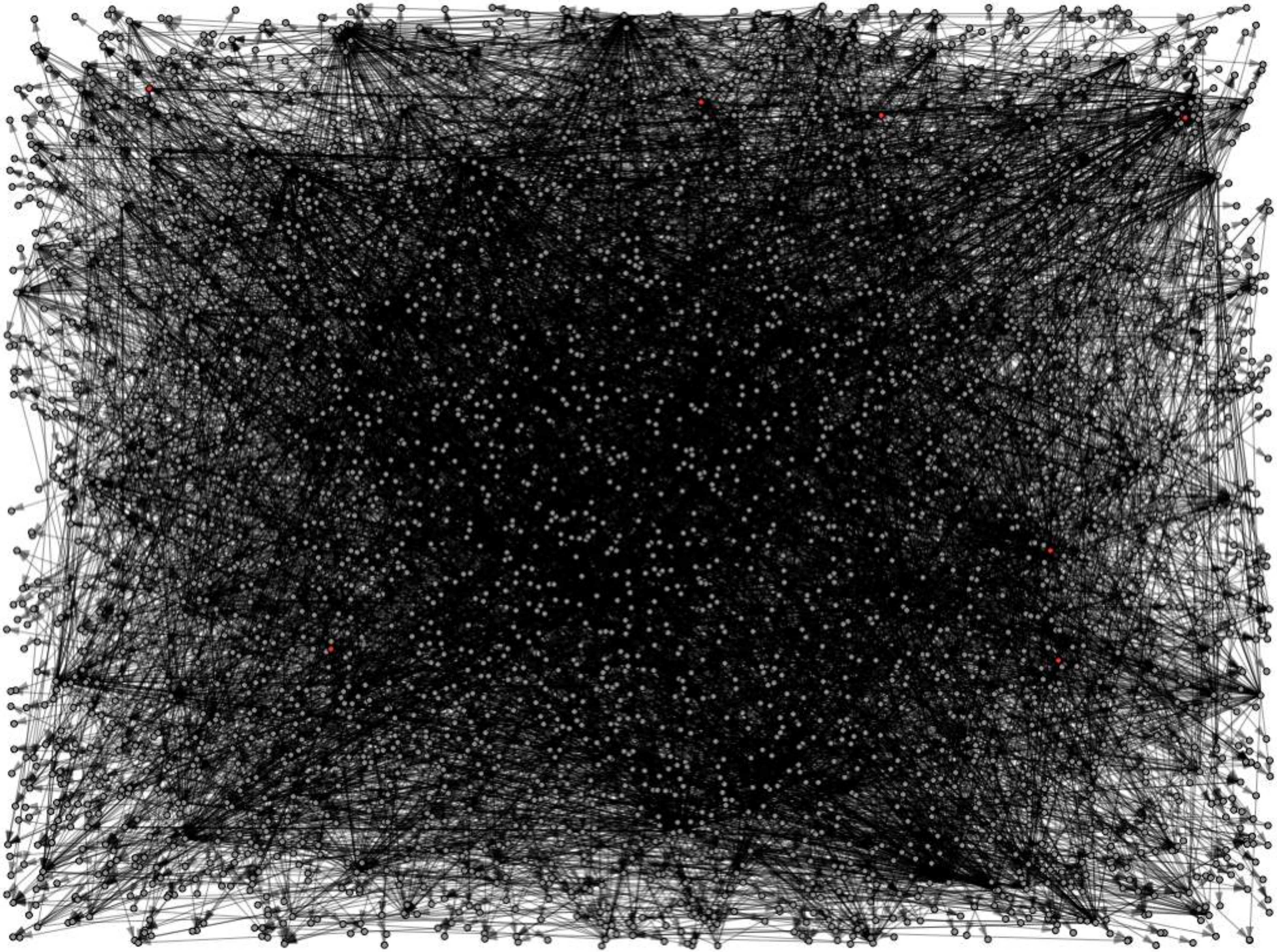


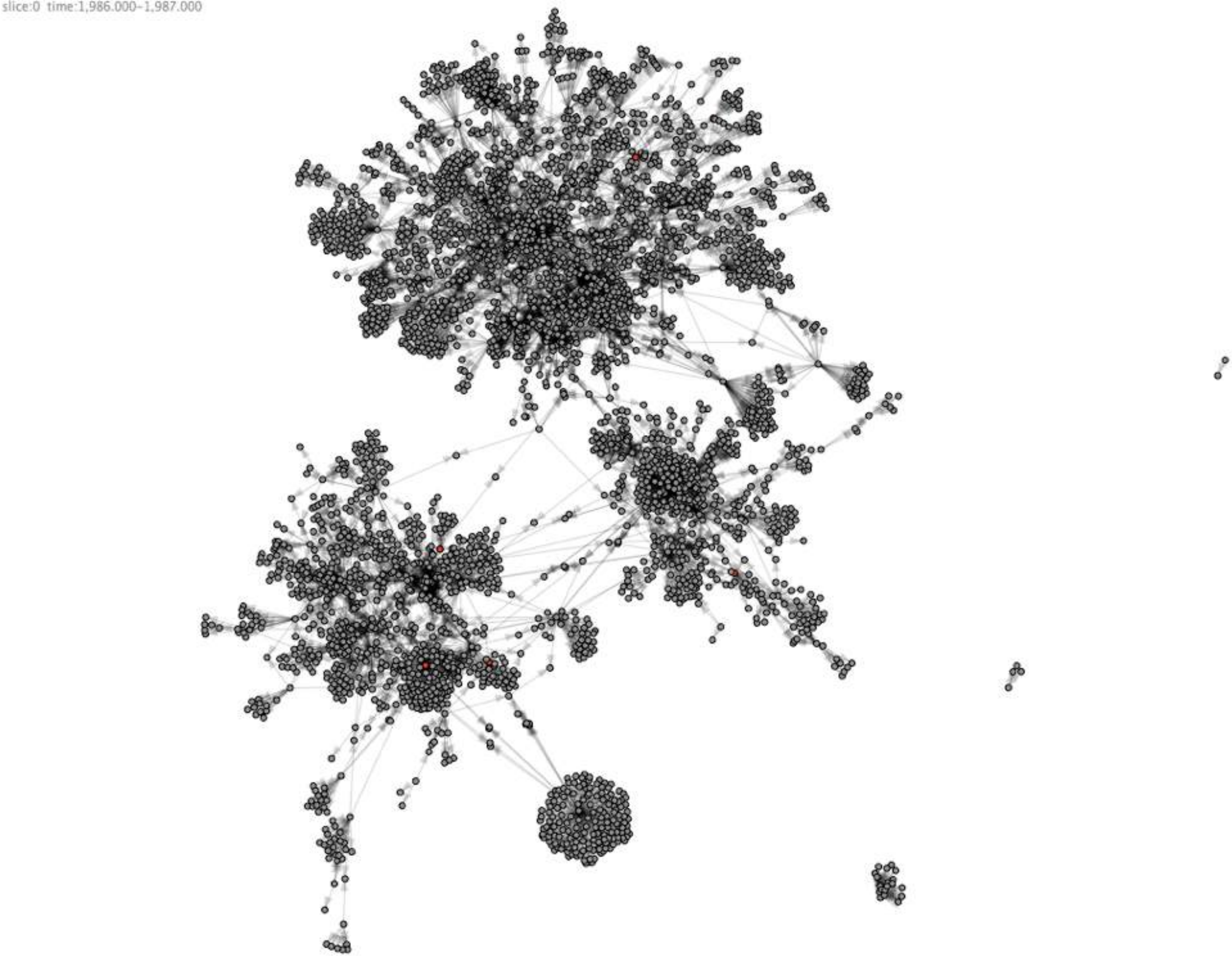
Social Networks



Social Networks

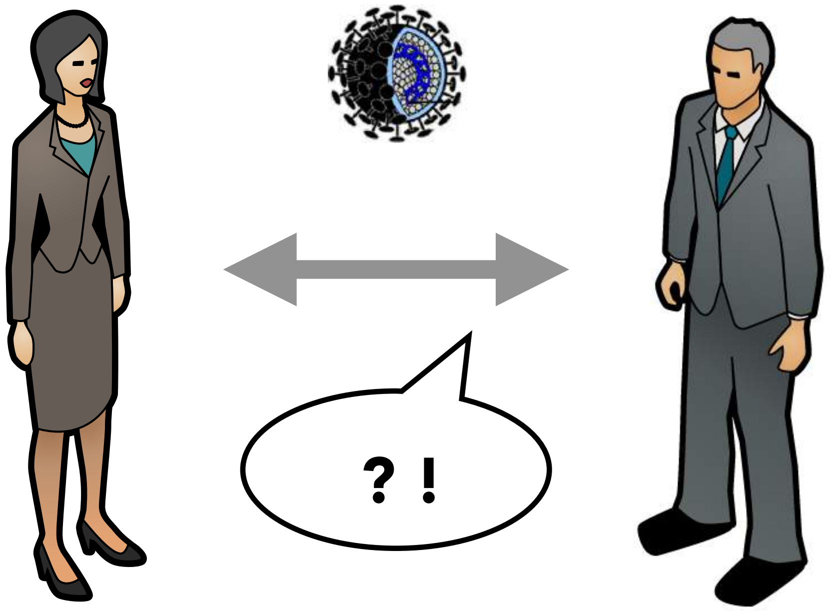





slice:0 time:1,986.000–1,987.000

Social Networks

Biological contagion



Social contagion

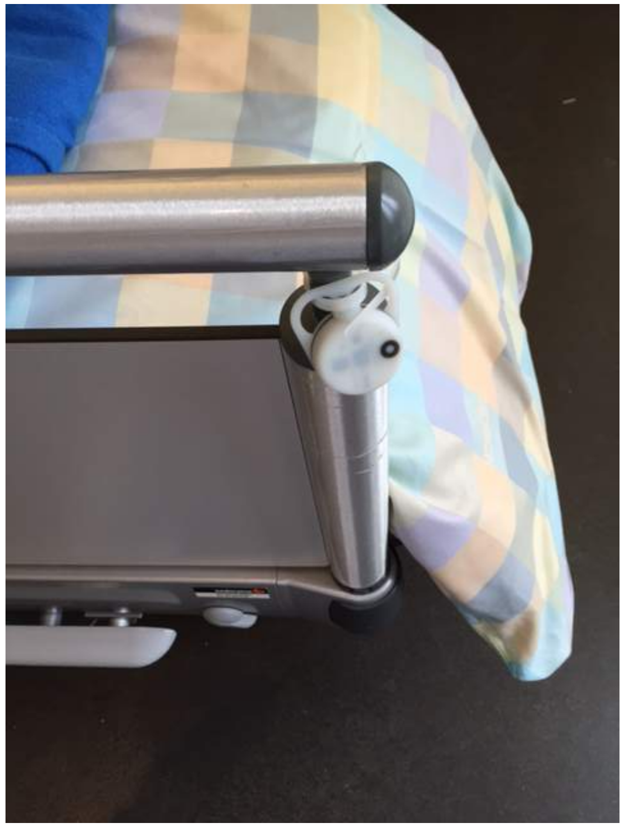
Measuring social networks

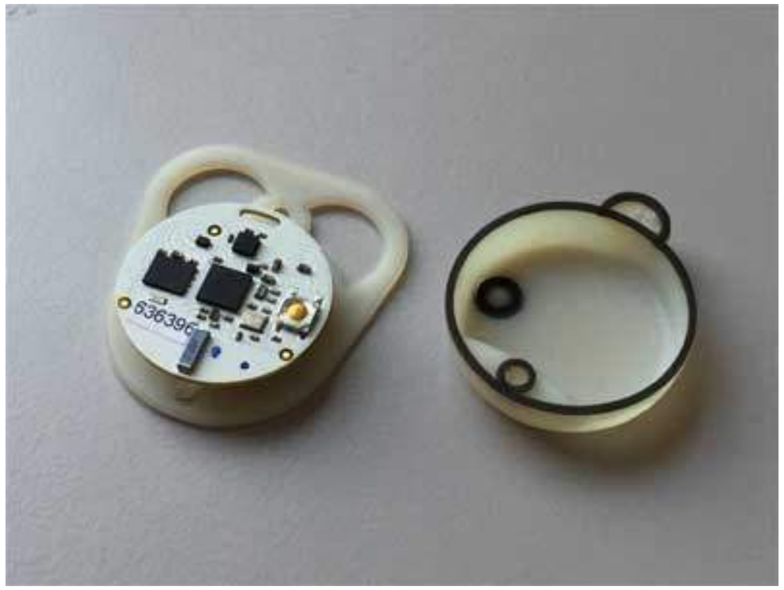
Surveys: widely used, but subjective and potentially inaccurate, high effort

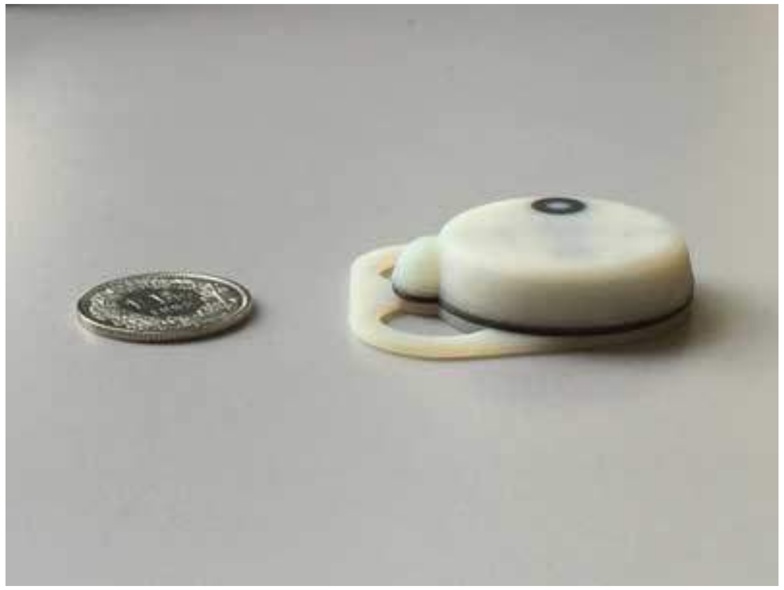
Wireless sensors (beacons): objective, high precision, low effort

http://postscapes.com/open-source-active-rfid-project-openbeacon

Measuring social networks

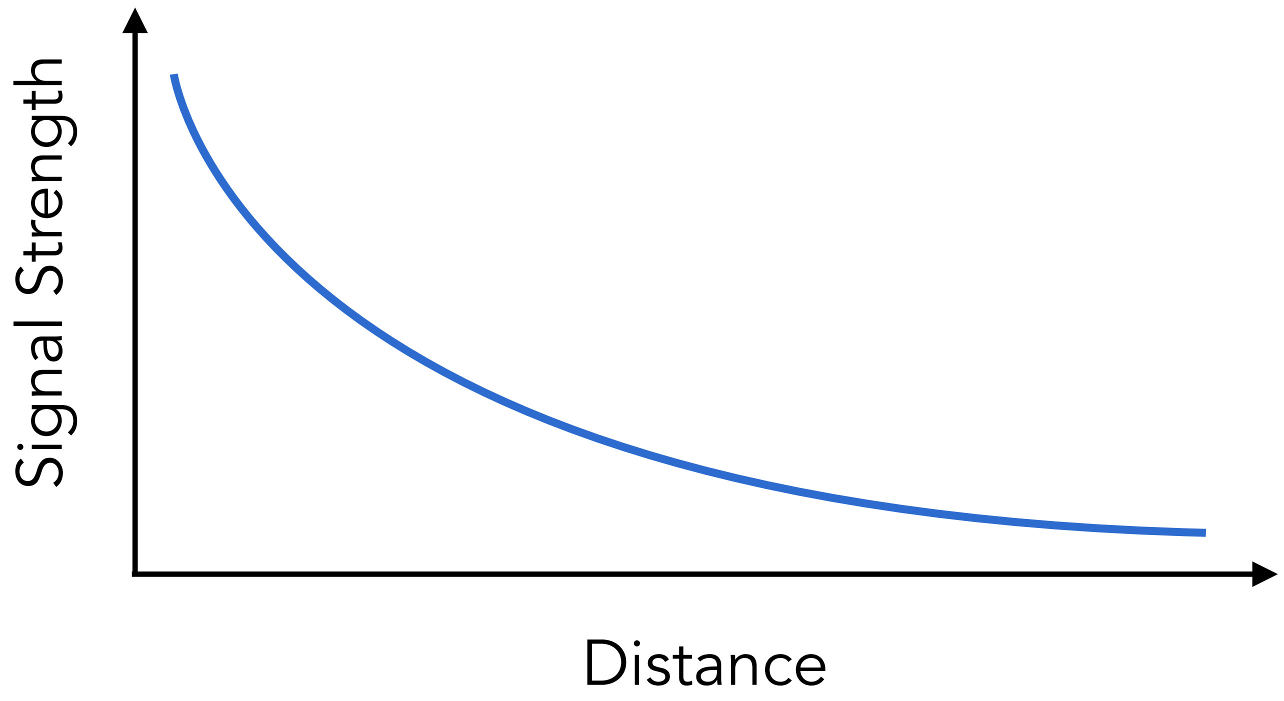




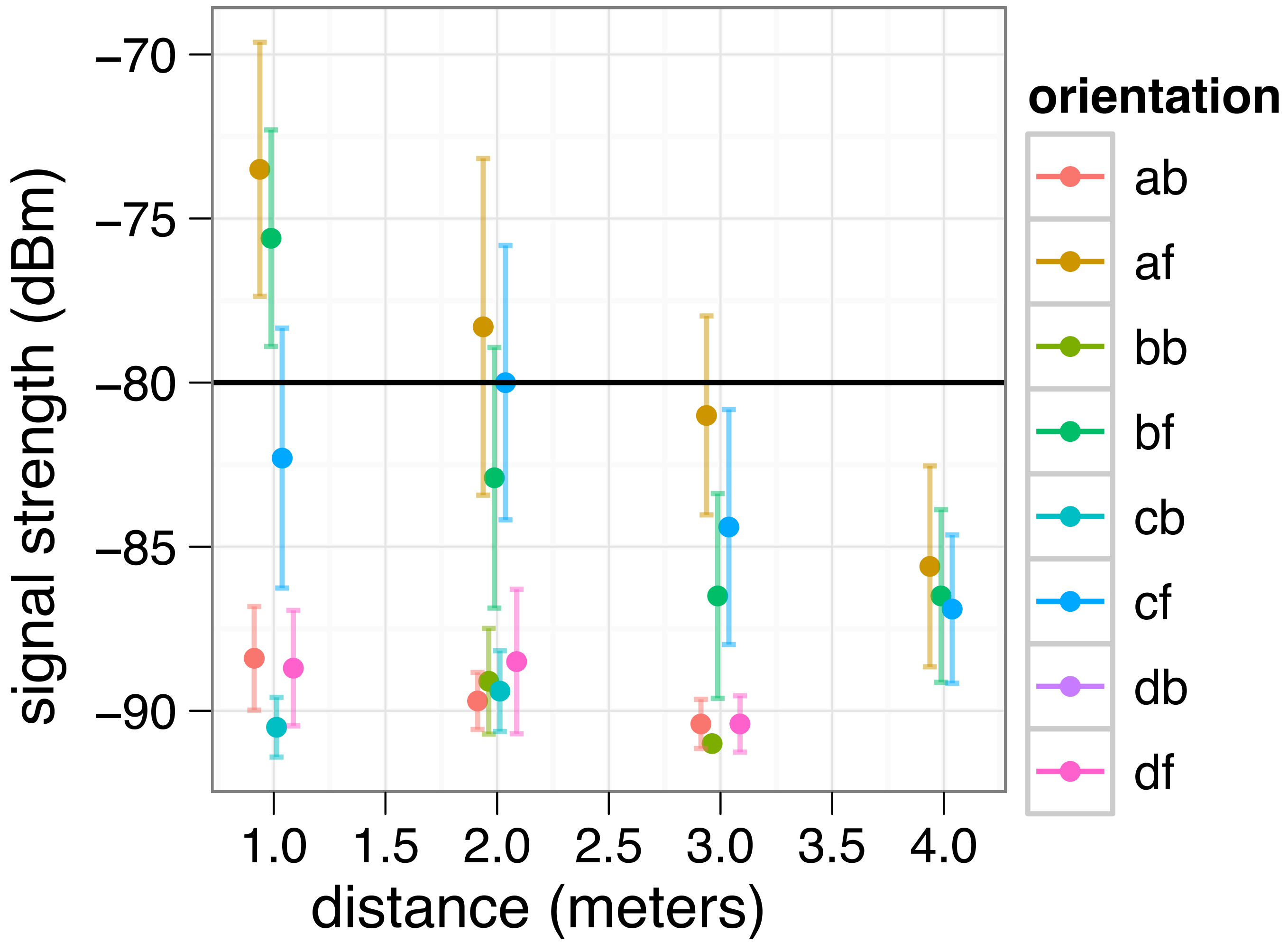


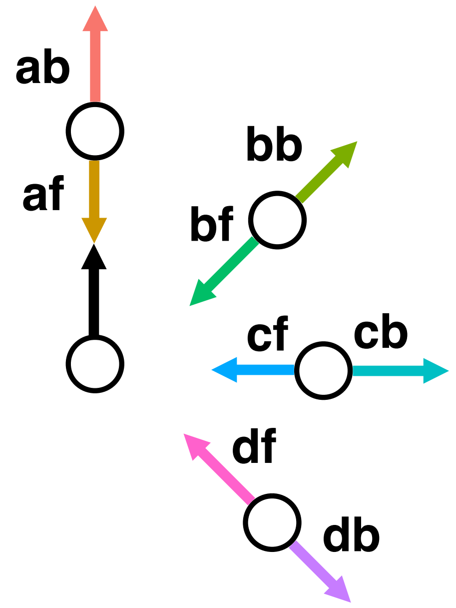
https://github.com/salathegroup/openbeacon-case

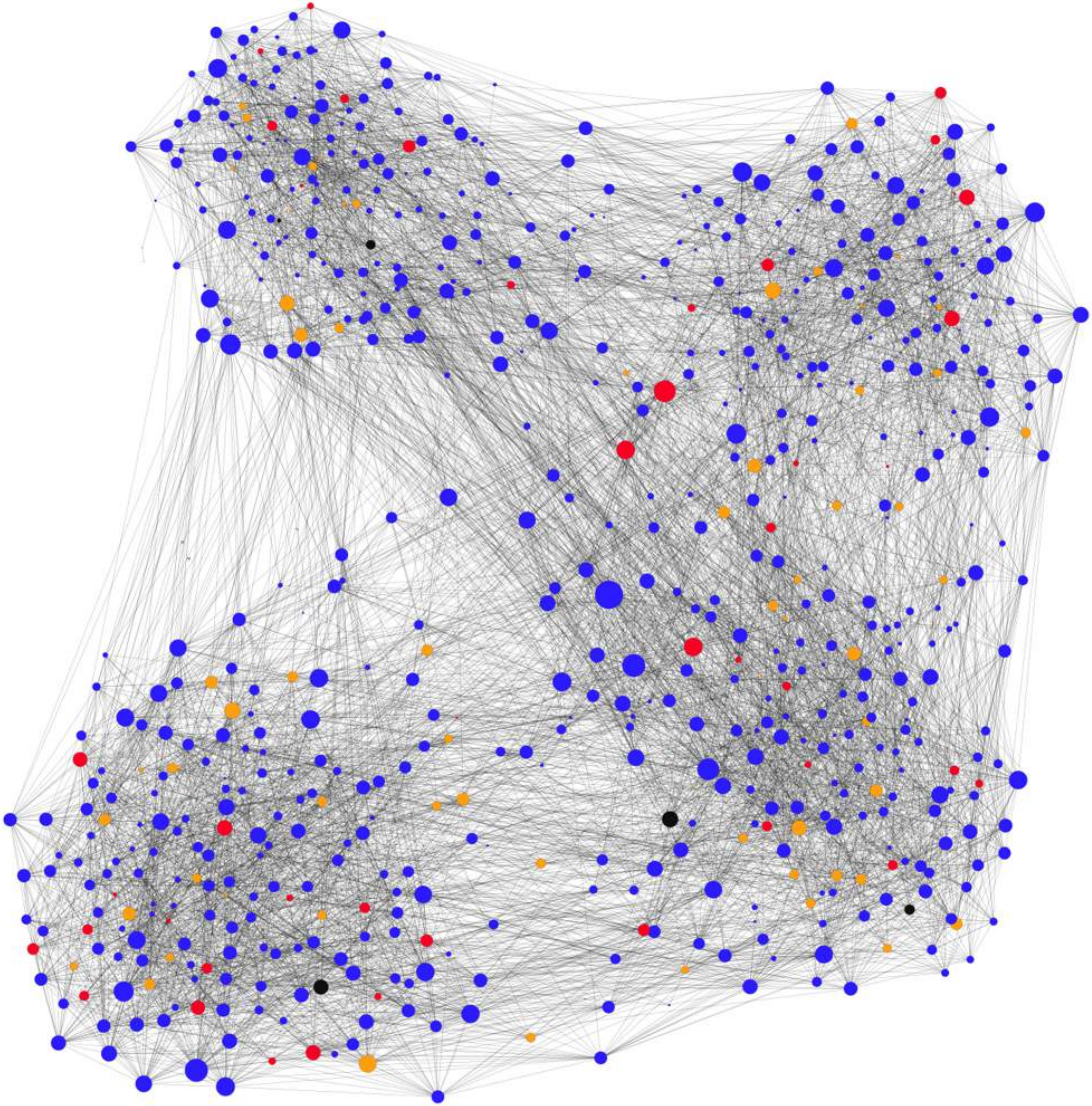
Measuring face to face interactions

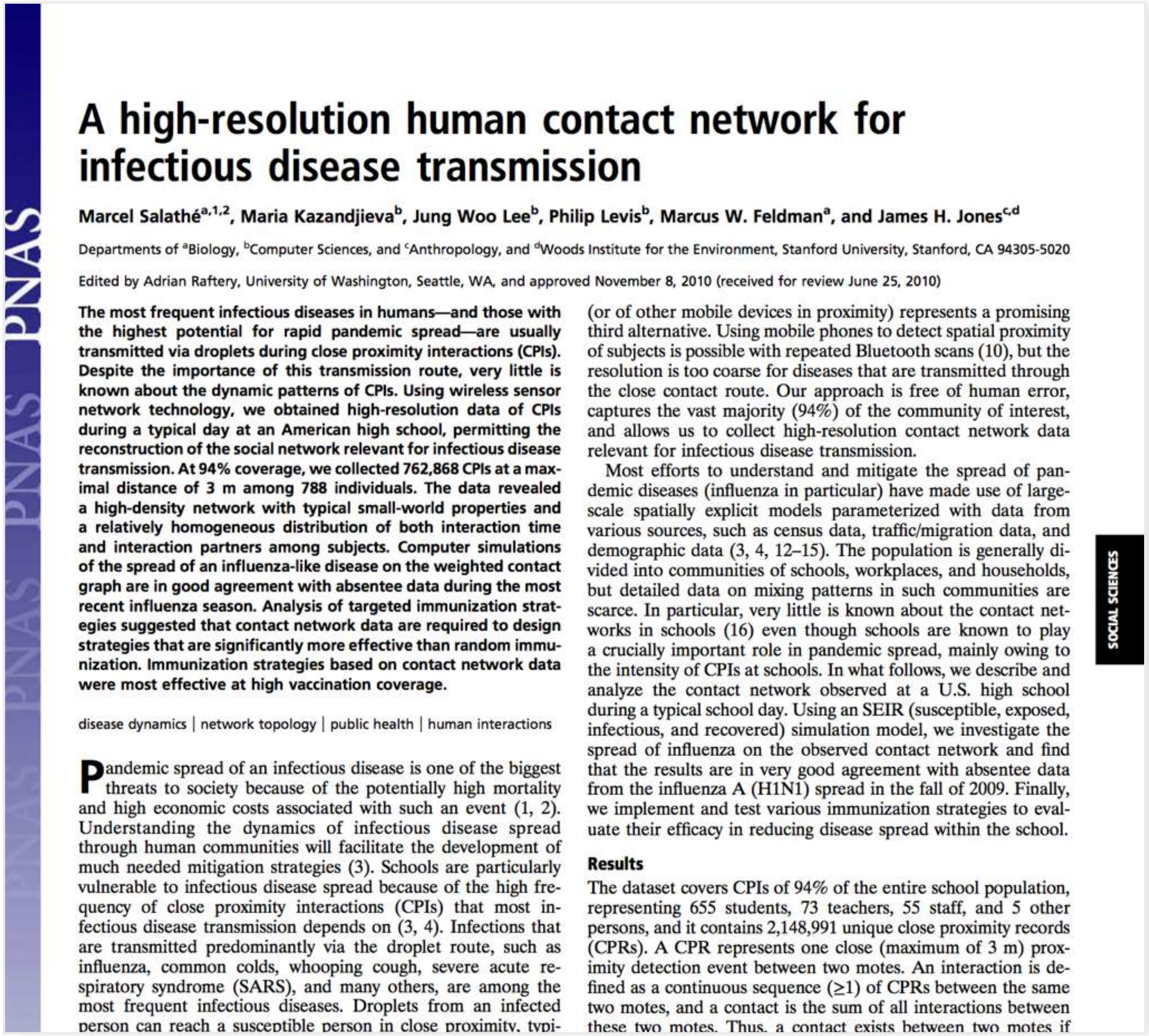


Measuring face to face interactions









A high-resolution human contact network for infectious disease transmission

Marcel Salathé[a,1,2], Maria Kazandjieva[b], Jung Woo Lee[b], Philip Levis[b], Marcus W. Feldman[a], and James H. Jones[c,d]

Departments of [a]Biology, [b]Computer Sciences, and [c]Anthropology, and [d]Woods Institute for the Environment, Stanford University, Stanford, CA 94305-5020

Edited by Adrian Raftery, University of Washington, Seattle, WA, and approved November 8, 2010 (received for review June 25, 2010)

The most frequent infectious diseases in humans—and those with the highest potential for rapid pandemic spread—are usually transmitted via droplets during close proximity interactions (CPIs). Despite the importance of this transmission route, very little is known about the dynamic patterns of CPIs. Using wireless sensor network technology, we obtained high-resolution data of CPIs during a typical day at an American high school, permitting the reconstruction of the social network relevant for infectious disease transmission. At 94% coverage, we collected 762,868 CPIs at a maximal distance of 3 m among 788 individuals. The data revealed a high-density network with typical small-world properties and a relatively homogeneous distribution of both interaction time and interaction partners among subjects. Computer simulations of the spread of an influenza-like disease on the weighted contact graph are in good agreement with absentee data during the most recent influenza season. Analysis of targeted immunization strategies suggested that contact network data are required to design strategies that are significantly more effective than random immunization. Immunization strategies based on contact network data were most effective at high vaccination coverage.

disease dynamics | network topology | public health | human interactions

Pandemic spread of an infectious disease is one of the biggest threats to society because of the potentially high mortality and high economic costs associated with such an event (1, 2). Understanding the dynamics of infectious disease spread through human communities will facilitate the development of much needed mitigation strategies (3). Schools are particularly vulnerable to infectious disease spread because of the high frequency of close proximity interactions (CPIs) that most infectious disease transmission depends on (3, 4). Infections that are transmitted predominantly via the droplet route, such as influenza, common colds, whooping cough, severe acute respiratory syndrome (SARS), and many others, are among the most frequent infectious diseases. Droplets from an infected person can reach a susceptible person in close proximity, typi-

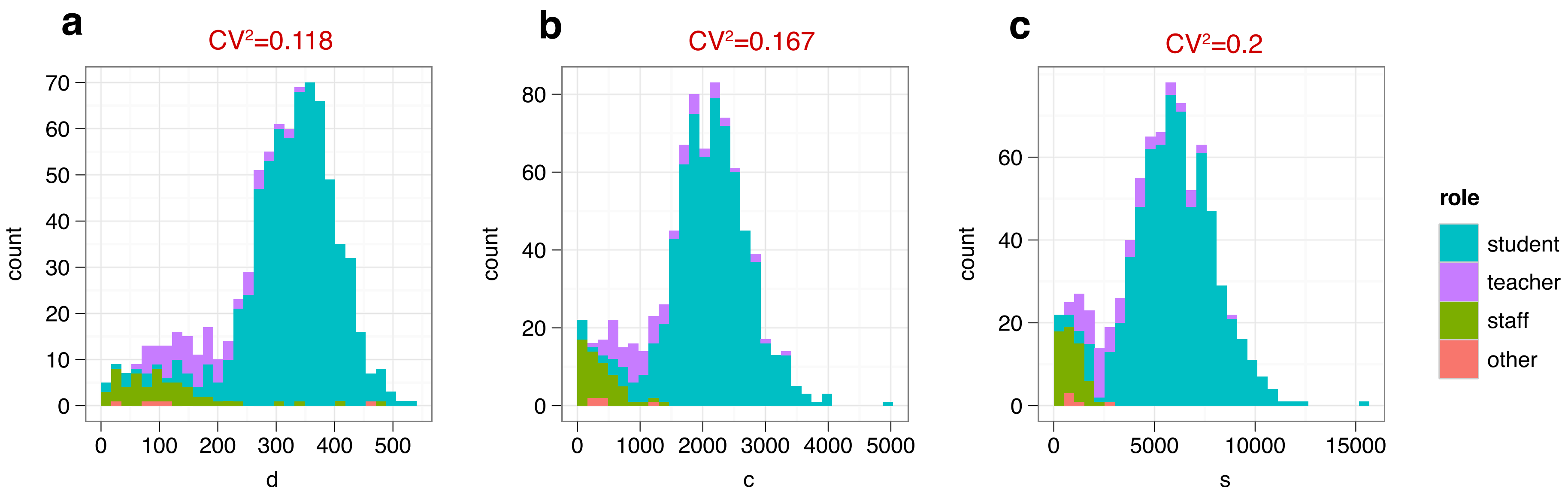
(or of other mobile devices in proximity) represents a promising third alternative. Using mobile phones to detect spatial proximity of subjects is possible with repeated Bluetooth scans (10), but the resolution is too coarse for diseases that are transmitted through the close contact route. Our approach is free of human error, captures the vast majority (94%) of the community of interest, and allows us to collect high-resolution contact network data relevant for infectious disease transmission.

Most efforts to understand and mitigate the spread of pandemic diseases (influenza in particular) have made use of large-scale spatially explicit models parameterized with data from various sources, such as census data, traffic/migration data, and demographic data (3, 4, 12–15). The population is generally divided into communities of schools, workplaces, and households, but detailed data on mixing patterns in such communities are scarce. In particular, very little is known about the contact networks in schools (16) even though schools are known to play a crucially important role in pandemic spread, mainly owing to the intensity of CPIs at schools. In what follows, we describe and analyze the contact network observed at a U.S. high school during a typical school day. Using an SEIR (susceptible, exposed, infectious, and recovered) simulation model, we investigate the spread of influenza on the observed contact network and find that the results are in very good agreement with absentee data from the influenza A (H1N1) spread in the fall of 2009. Finally, we implement and test various immunization strategies to evaluate their efficacy in reducing disease spread within the school.

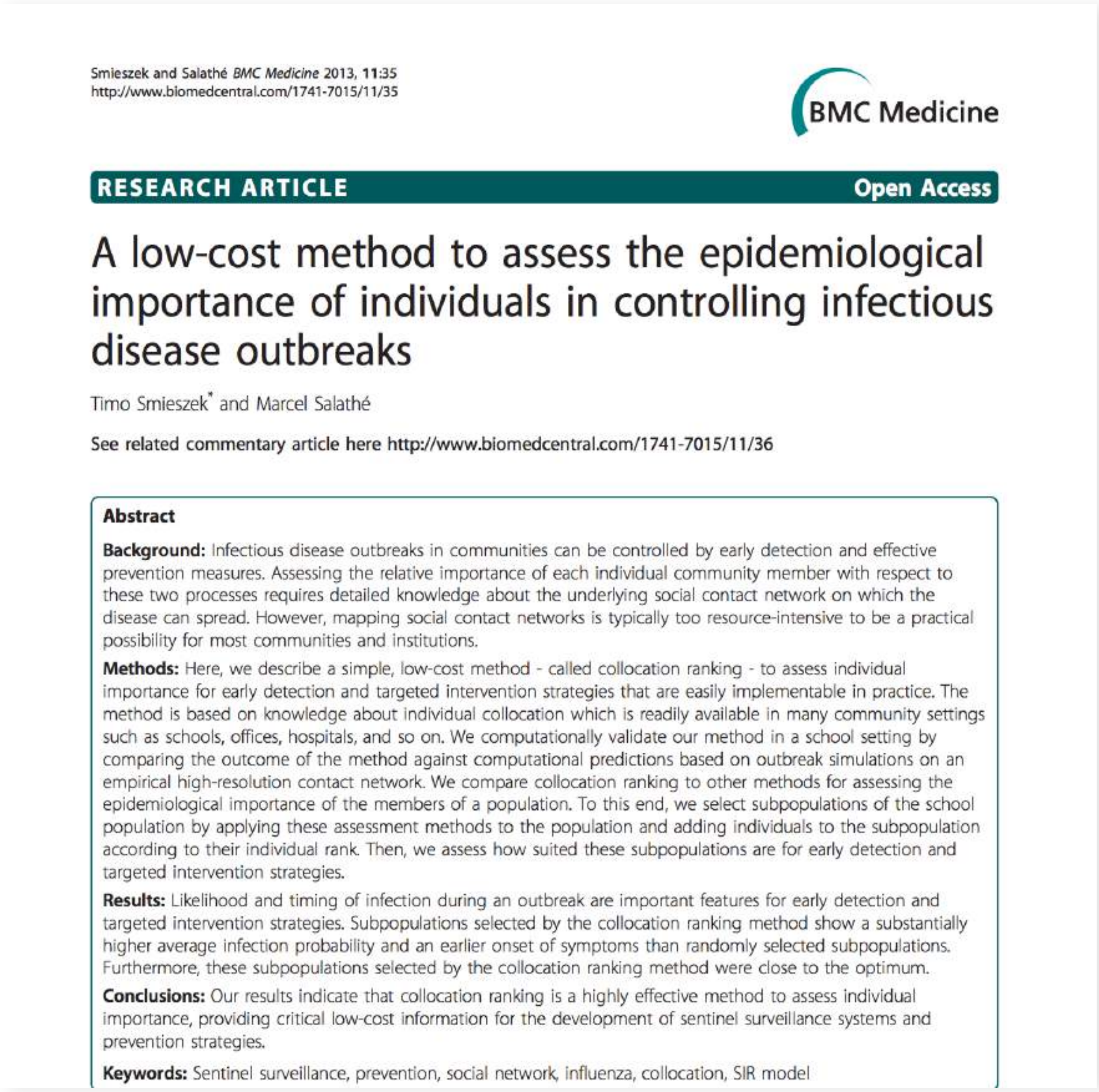
Results

The dataset covers CPIs of 94% of the entire school population, representing 655 students, 73 teachers, 55 staff, and 5 other persons, and it contains 2,148,991 unique close proximity records (CPRs). A CPR represents one close (maximum of 3 m) proximity detection event between two motes. An interaction is defined as a continuous sequence (≥1) of CPRs between the same two motes, and a contact is the sum of all interactions between these two motes. Thus, a contact exists between two motes if

First high-resolution data, statistical description & vaccine strategy modeling



Demonstrating collocation ranking based on schedules as an effective predictor of importance



Smieszek and Salathé *BMC Medicine* 2013, 11:35
http://www.biomedcentral.com/1741-7015/11/35

BMC Medicine

RESEARCH ARTICLE **Open Access**

A low-cost method to assess the epidemiological importance of individuals in controlling infectious disease outbreaks

Timo Smieszek* and Marcel Salathé

See related commentary article here http://www.biomedcentral.com/1741-7015/11/36

Abstract

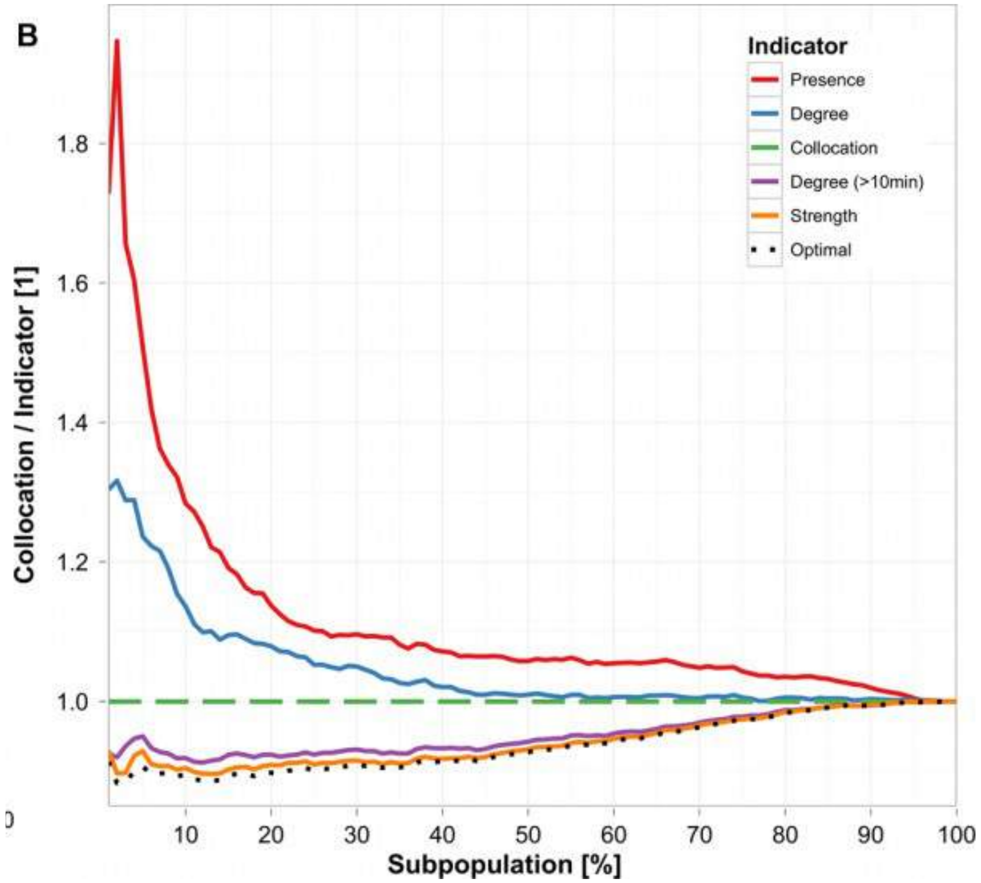
Background: Infectious disease outbreaks in communities can be controlled by early detection and effective prevention measures. Assessing the relative importance of each individual community member with respect to these two processes requires detailed knowledge about the underlying social contact network on which the disease can spread. However, mapping social contact networks is typically too resource-intensive to be a practical possibility for most communities and institutions.

Methods: Here, we describe a simple, low-cost method - called collocation ranking - to assess individual importance for early detection and targeted intervention strategies that are easily implementable in practice. The method is based on knowledge about individual collocation which is readily available in many community settings such as schools, offices, hospitals, and so on. We computationally validate our method in a school setting by comparing the outcome of the method against computational predictions based on outbreak simulations on an empirical high-resolution contact network. We compare collocation ranking to other methods for assessing the epidemiological importance of the members of a population. To this end, we select subpopulations of the school population by applying these assessment methods to the population and adding individuals to the subpopulation according to their individual rank. Then, we assess how suited these subpopulations are for early detection and targeted intervention strategies.

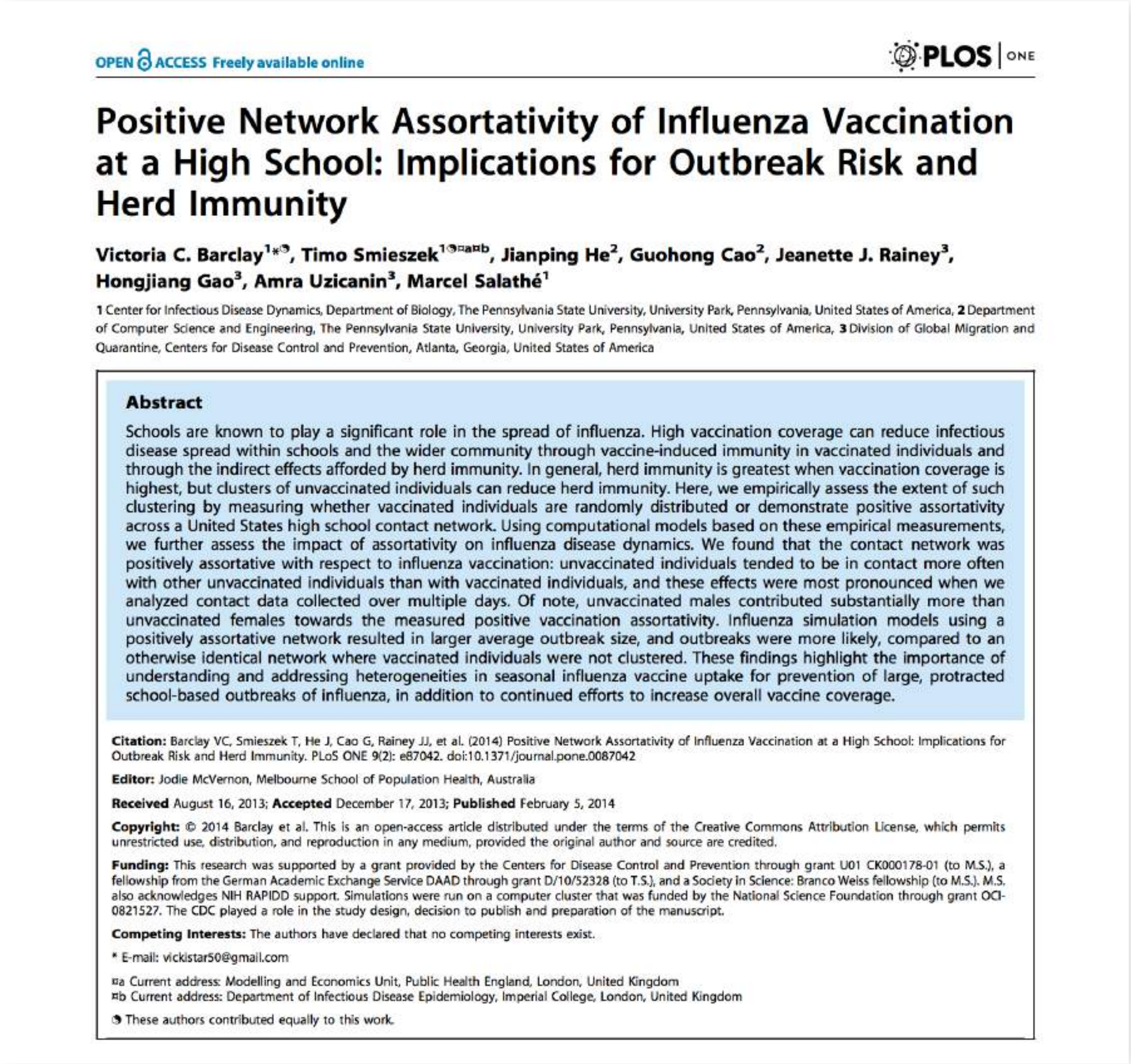
Results: Likelihood and timing of infection during an outbreak are important features for early detection and targeted intervention strategies. Subpopulations selected by the collocation ranking method show a substantially higher average infection probability and an earlier onset of symptoms than randomly selected subpopulations. Furthermore, these subpopulations selected by the collocation ranking method were close to the optimum.

Conclusions: Our results indicate that collocation ranking is a highly effective method to assess individual importance, providing critical low-cost information for the development of sentinel surveillance systems and prevention strategies.

Keywords: Sentinel surveillance, prevention, social network, influenza, collocation, SIR model



Measuring vaccination assortativity on empirical network



OPEN ACCESS Freely available online

PLOS | ONE

Positive Network Assortativity of Influenza Vaccination at a High School: Implications for Outbreak Risk and Herd Immunity

Victoria C. Barclay[1*◑], Timo Smieszek[1◑¤a¤b], Jianping He[2], Guohong Cao[2], Jeanette J. Rainey[3], Hongjiang Gao[3], Amra Uzicanin[3], Marcel Salathé[1]

1 Center for Infectious Disease Dynamics, Department of Biology, The Pennsylvania State University, University Park, Pennsylvania, United States of America, **2** Department of Computer Science and Engineering, The Pennsylvania State University, University Park, Pennsylvania, United States of America, **3** Division of Global Migration and Quarantine, Centers for Disease Control and Prevention, Atlanta, Georgia, United States of America

Abstract

Schools are known to play a significant role in the spread of influenza. High vaccination coverage can reduce infectious disease spread within schools and the wider community through vaccine-induced immunity in vaccinated individuals and through the indirect effects afforded by herd immunity. In general, herd immunity is greatest when vaccination coverage is highest, but clusters of unvaccinated individuals can reduce herd immunity. Here, we empirically assess the extent of such clustering by measuring whether vaccinated individuals are randomly distributed or demonstrate positive assortativity across a United States high school contact network. Using computational models based on these empirical measurements, we further assess the impact of assortativity on influenza disease dynamics. We found that the contact network was positively assortative with respect to influenza vaccination: unvaccinated individuals tended to be in contact more often with other unvaccinated individuals than with vaccinated individuals, and these effects were most pronounced when we analyzed contact data collected over multiple days. Of note, unvaccinated males contributed substantially more than unvaccinated females towards the measured positive vaccination assortativity. Influenza simulation models using a positively assortative network resulted in larger average outbreak size, and outbreaks were more likely, compared to an otherwise identical network where vaccinated individuals were not clustered. These findings highlight the importance of understanding and addressing heterogeneities in seasonal influenza vaccine uptake for prevention of large, protracted school-based outbreaks of influenza, in addition to continued efforts to increase overall vaccine coverage.

Citation: Barclay VC, Smieszek T, He J, Cao G, Rainey JJ, et al. (2014) Positive Network Assortativity of Influenza Vaccination at a High School: Implications for Outbreak Risk and Herd Immunity. PLoS ONE 9(2): e87042. doi:10.1371/journal.pone.0087042

Editor: Jodie McVernon, Melbourne School of Population Health, Australia

Received August 16, 2013; **Accepted** December 17, 2013; **Published** February 5, 2014

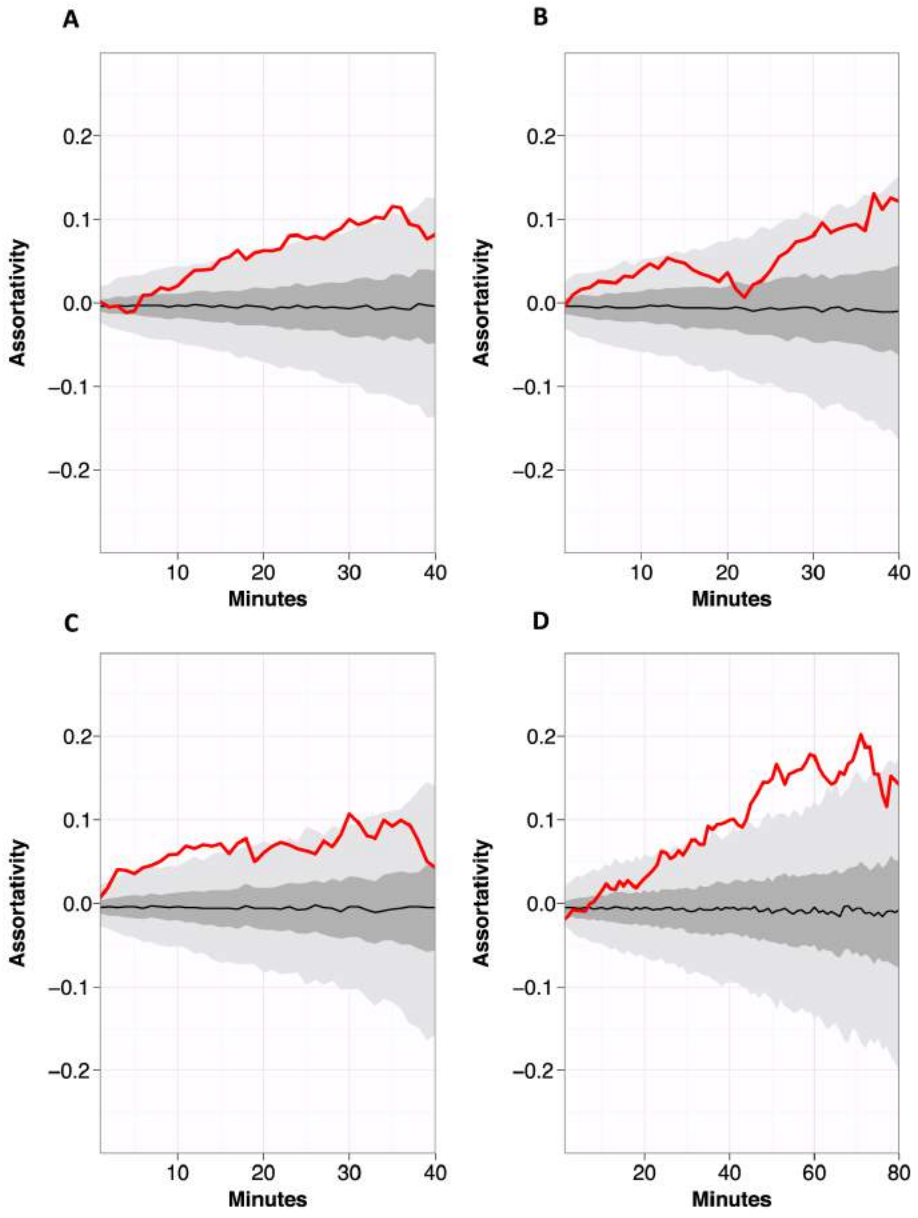
Funding: This research was supported by a grant provided by the Centers for Disease Control and Prevention through grant U01 CK000178-01 (to M.S.), a fellowship from the German Academic Exchange Service DAAD through grant D/10/52328 (to T.S.), and a Society in Science: Branco Weiss fellowship (to M.S.). M.S. also acknowledges NIH RAPIDD support. Simulations were run on a computer cluster that was funded by the National Science Foundation through grant OCI-0821527. The CDC played a role in the study design, decision to publish and preparation of the manuscript.

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

* E-mail: vickistar50@gmail.com

¤a Current address: Modelling and Economics Unit, Public Health England, London, United Kingdom
¤b Current address: Department of Infectious Disease Epidemiology, Imperial College, London, United Kingdom

◑ These authors contributed equally to this work.



Comparing surveys with sensor data and quantifying survey errors

Smieszek *et al. BMC Infectious Diseases* 2014, **14**:136
http://www.biomedcentral.com/1471-2334/14/136

BMC Infectious Diseases

RESEARCH ARTICLE **Open Access**

How should social mixing be measured: comparing web-based survey and sensor-based methods

Timo Smieszek[1,3*†], Victoria C Barclay[1†], Indulaxmi Seeni[1], Jeanette J Rainey[2], Hongjiang Gao[2], Amra Uzicanin[2] and Marcel Salathé[1]

Abstract

Background: Contact surveys and diaries have conventionally been used to measure contact networks in different settings for elucidating infectious disease transmission dynamics of respiratory infections. More recently, technological advances have permitted the use of wireless sensor devices, which can be worn by individuals interacting in a particular social context to record high resolution mixing patterns. To date, a direct comparison of these two different methods for collecting contact data has not been performed.

Methods: We studied the contact network at a United States high school in the spring of 2012. All school members (i.e., students, teachers, and other staff) were invited to wear wireless sensor devices for a single school day, and asked to remember and report the name and duration of all of their close proximity conversational contacts for that day in an online contact survey. We compared the two methods in terms of the resulting network densities, nodal degrees, and degree distributions. We also assessed the correspondence between the methods at the dyadic and individual levels.

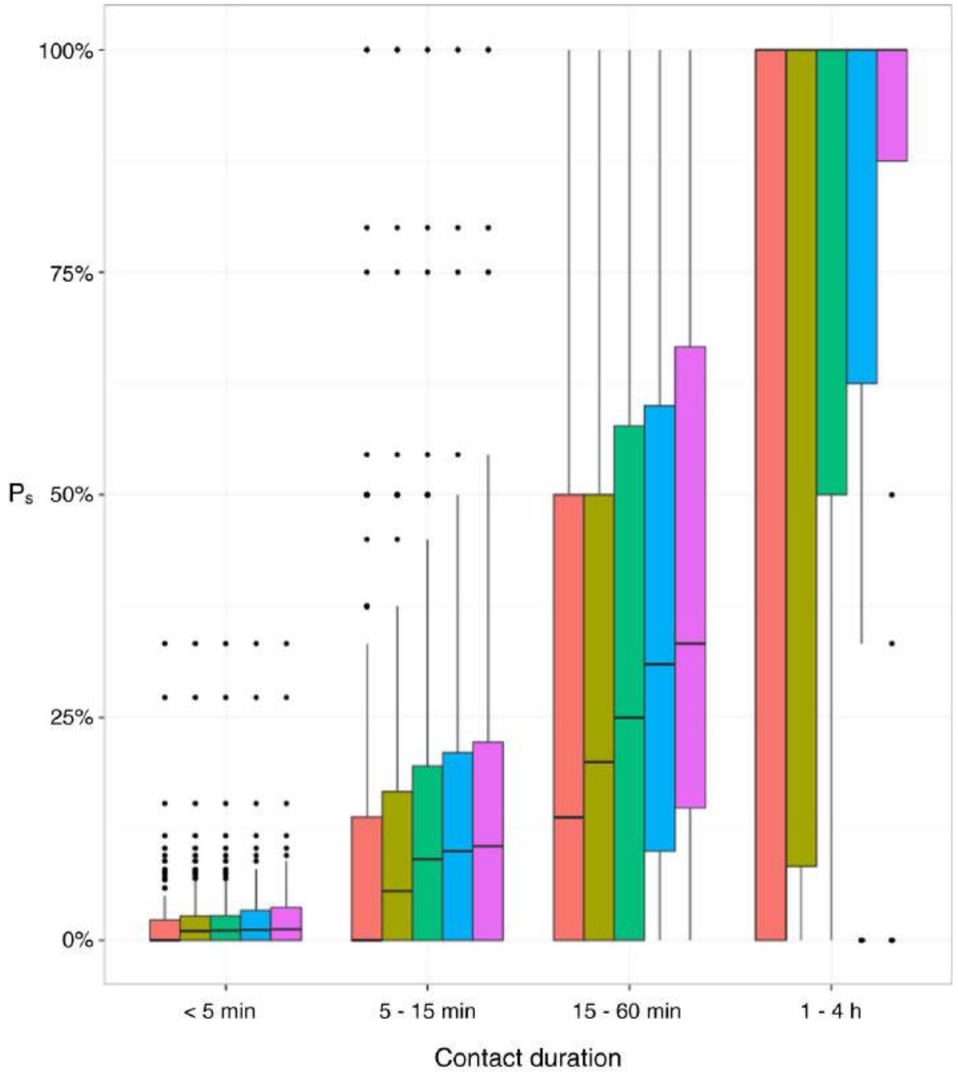
Results: We found limited congruence in recorded contact data between the online contact survey and wireless sensors. In particular, there was only negligible correlation between the two methods for nodal degree, and the degree distribution differed substantially between both methods. We found that survey underreporting was a significant source of the difference between the two methods, and that this difference could be improved by excluding individuals who reported only a few contact partners. Additionally, survey reporting was more accurate for contacts of longer duration, and very inaccurate for contacts of shorter duration. Finally, female participants tended to report more accurately than male participants.

Conclusions: Online contact surveys and wireless sensor devices collected incongruent network data from an identical setting. This finding suggests that these two methods cannot be used interchangeably for informing models of infectious disease dynamics.

Keywords: Contact networks, Social network, Proximity network, Droplet transmission, Contact survey, Wireless sensor network

Background

Contact network data are useful for parameterizing models of livestock and wild animals [11-15], are all better understood because of monitoring and modeling the interaction



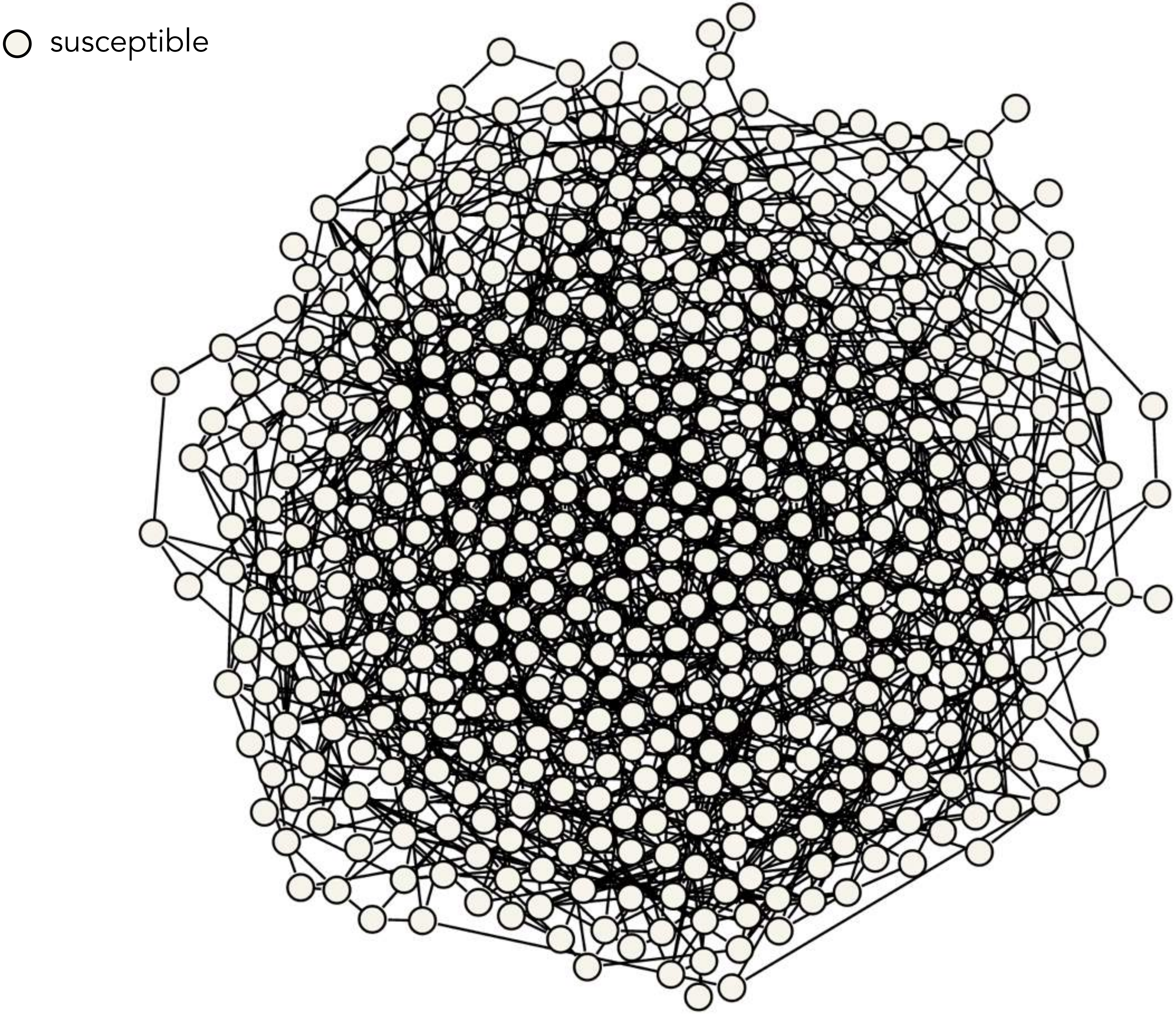
Vaccination

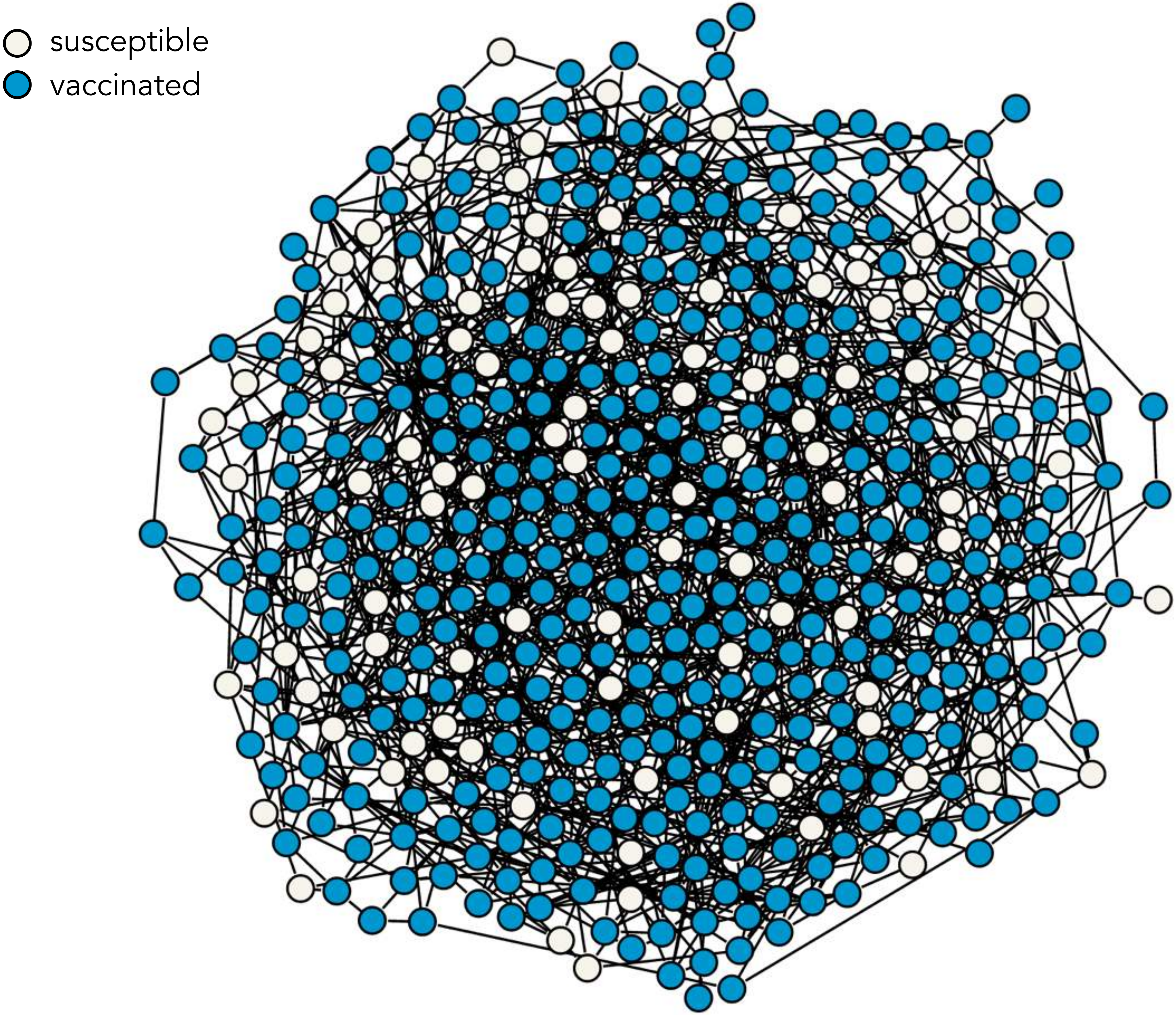
R_0 = number of secondary infections from the first case.

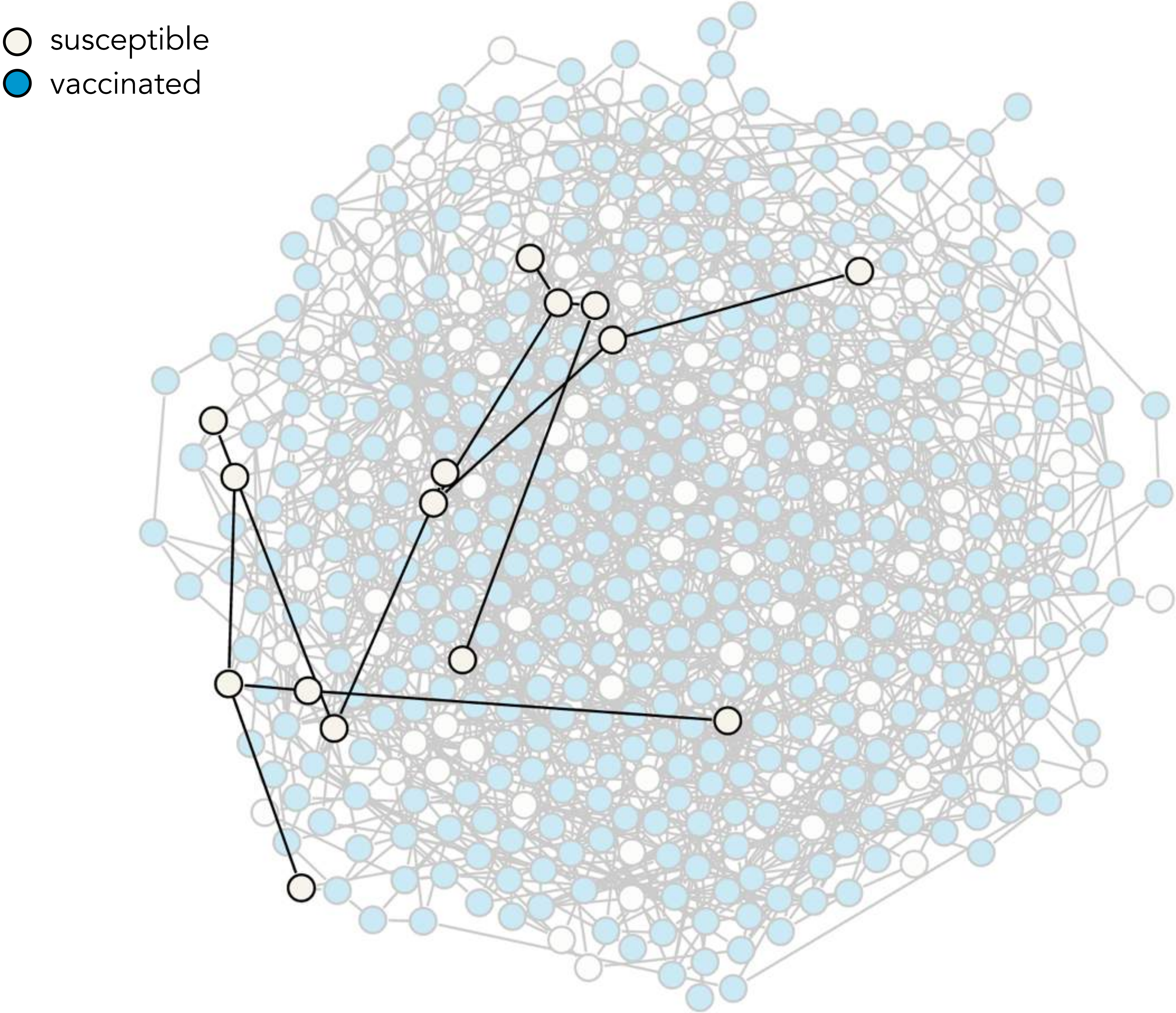
Herd immunity:
$1 - 1/R_0$

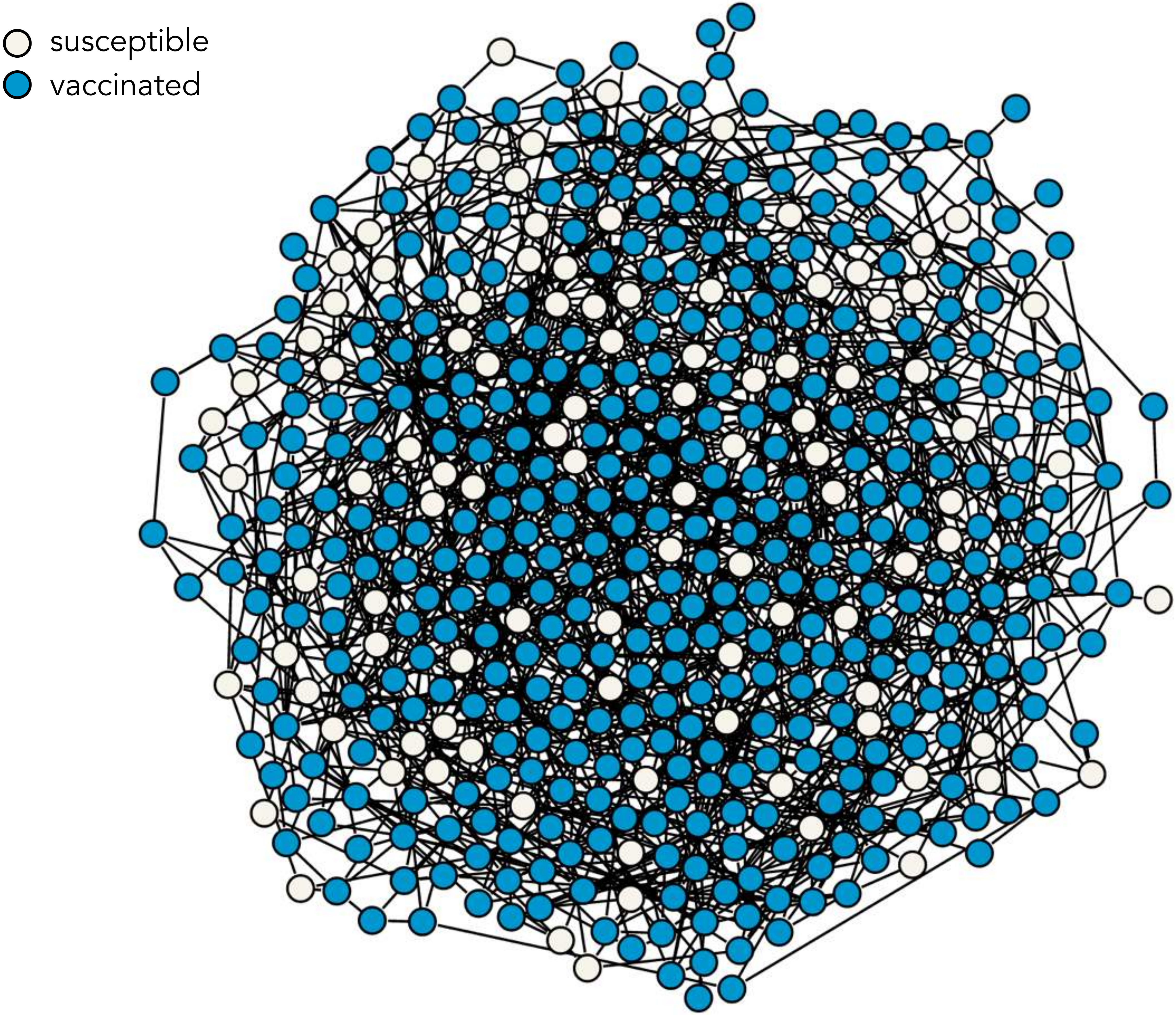
Example: Measles

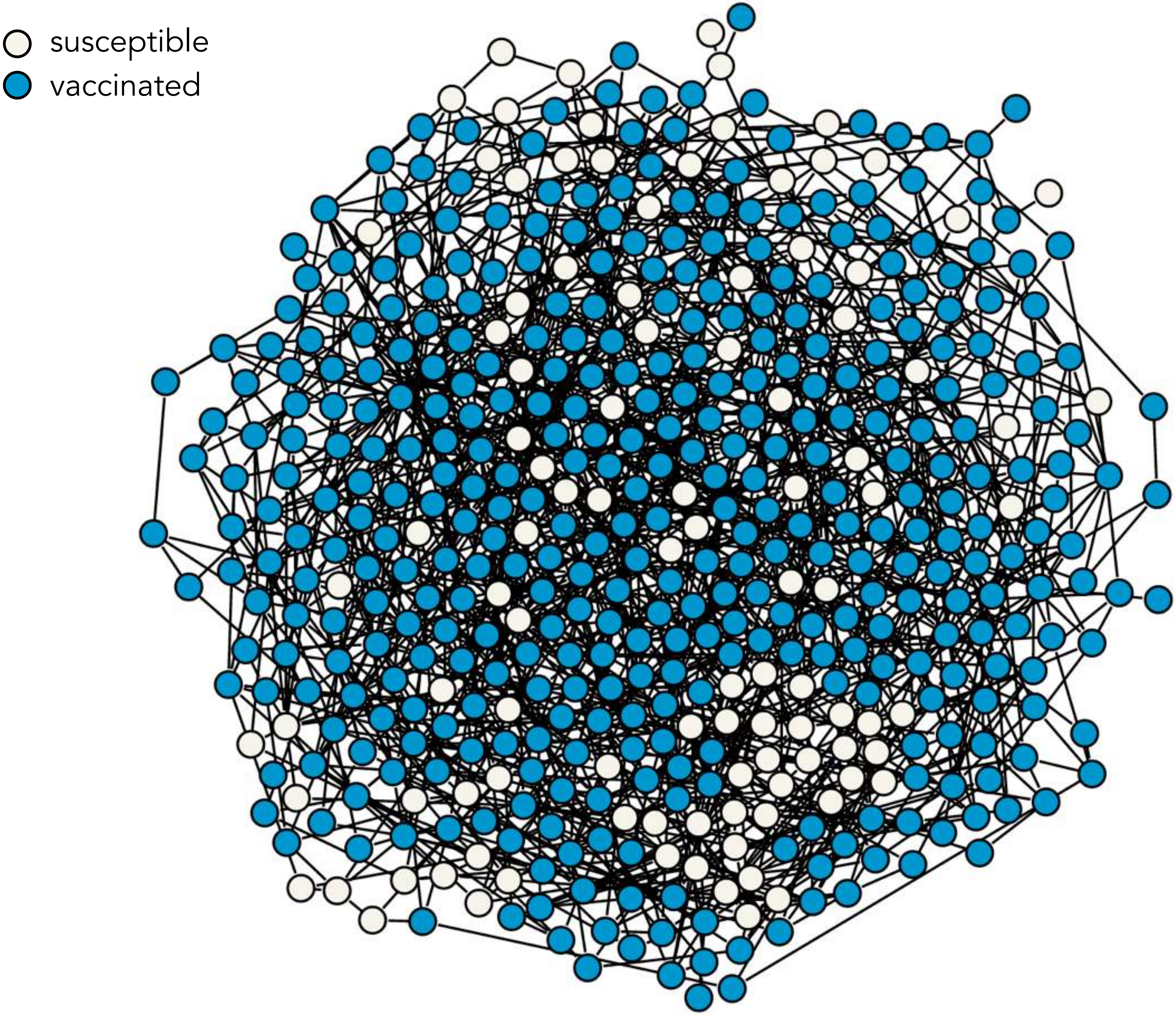
$R_0 \sim 10\text{-}20$
Recommended vaccination rate: 90-95%

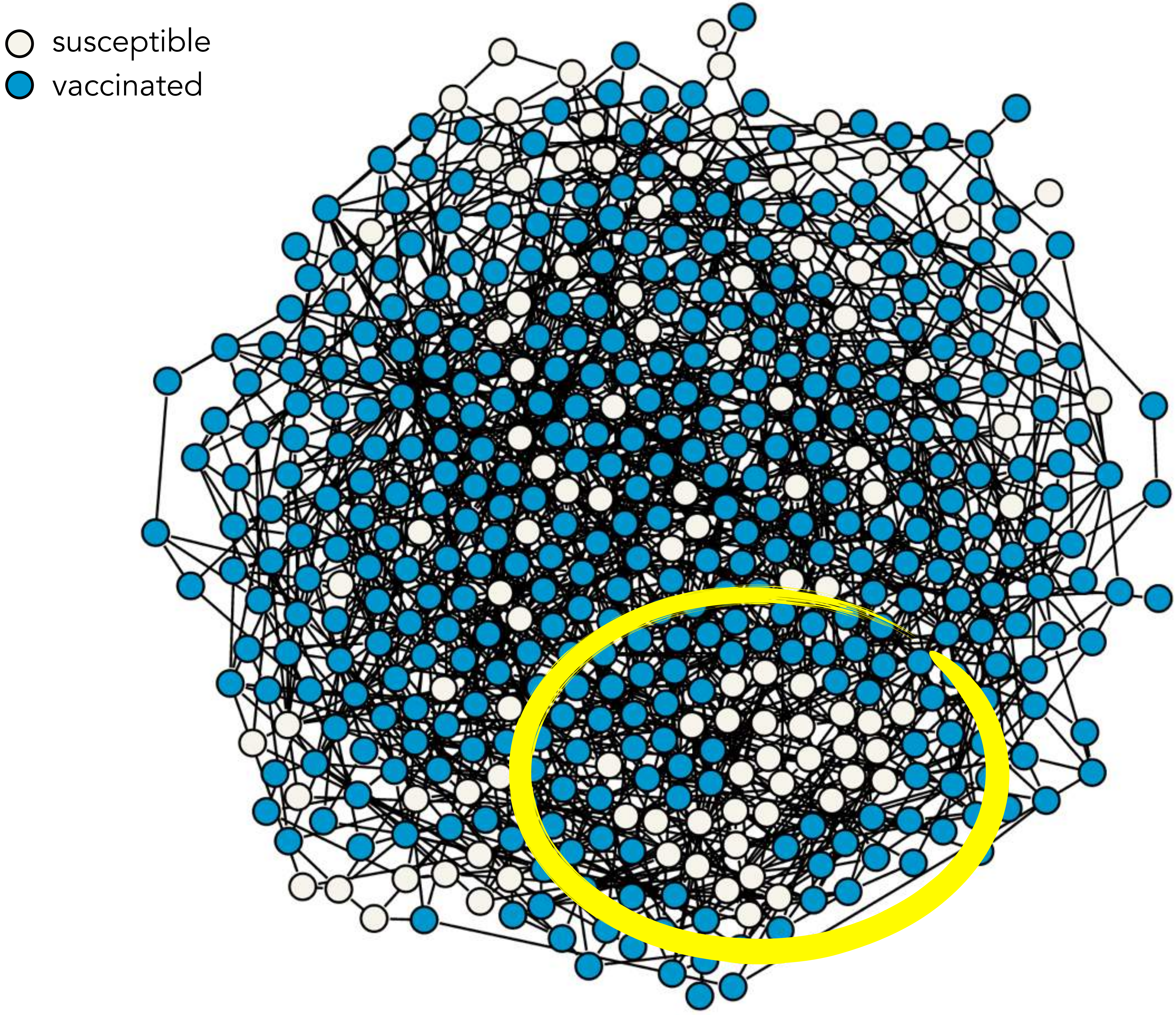

○ susceptible

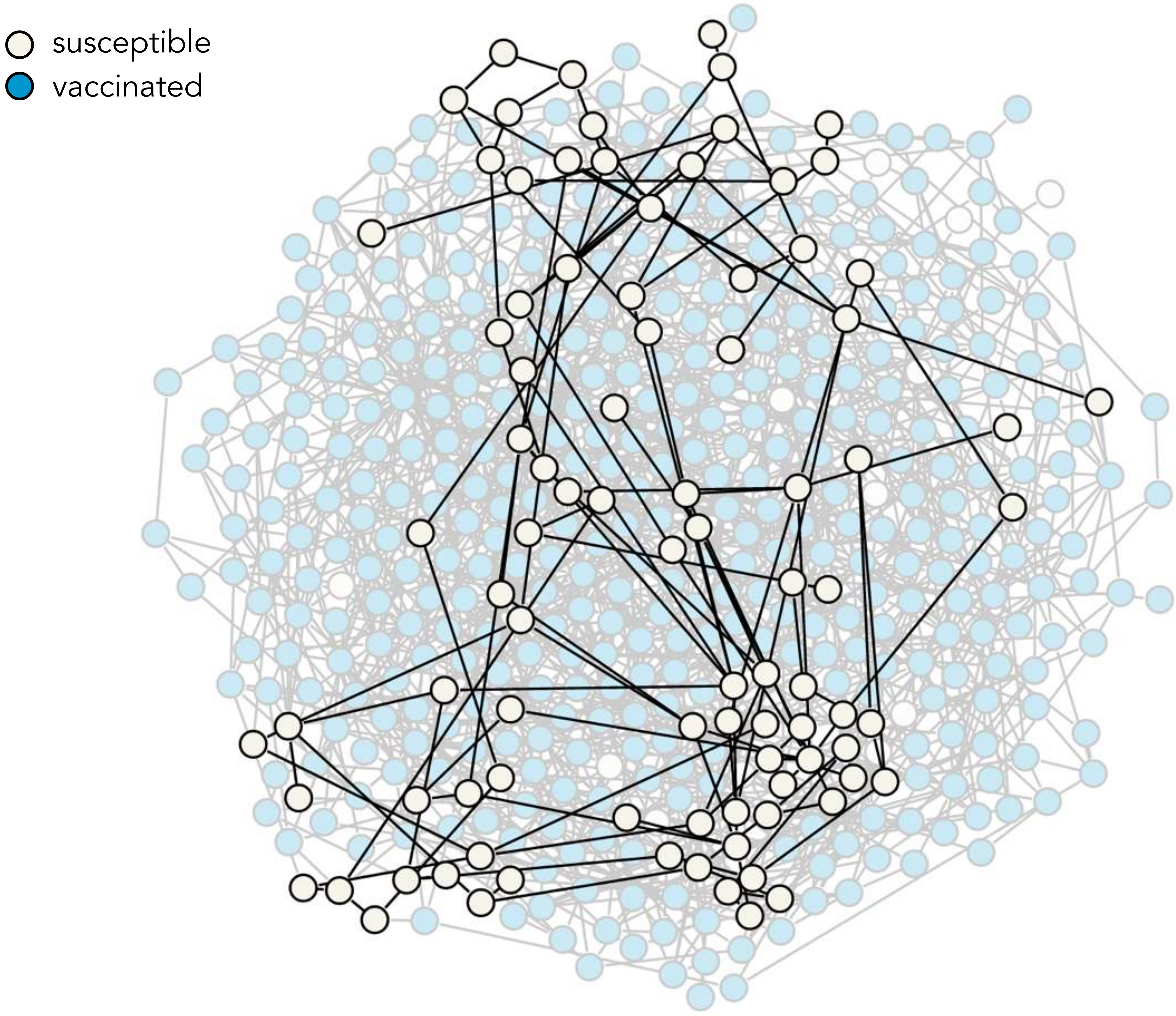

○ susceptible
● vaccinated


susceptible
vaccinated


○ susceptible
● vaccinated


○ susceptible
● vaccinated


○ susceptible
● vaccinated


○ susceptible
● vaccinated

The NEW ENGLAND JOURNAL *of* MEDICINE

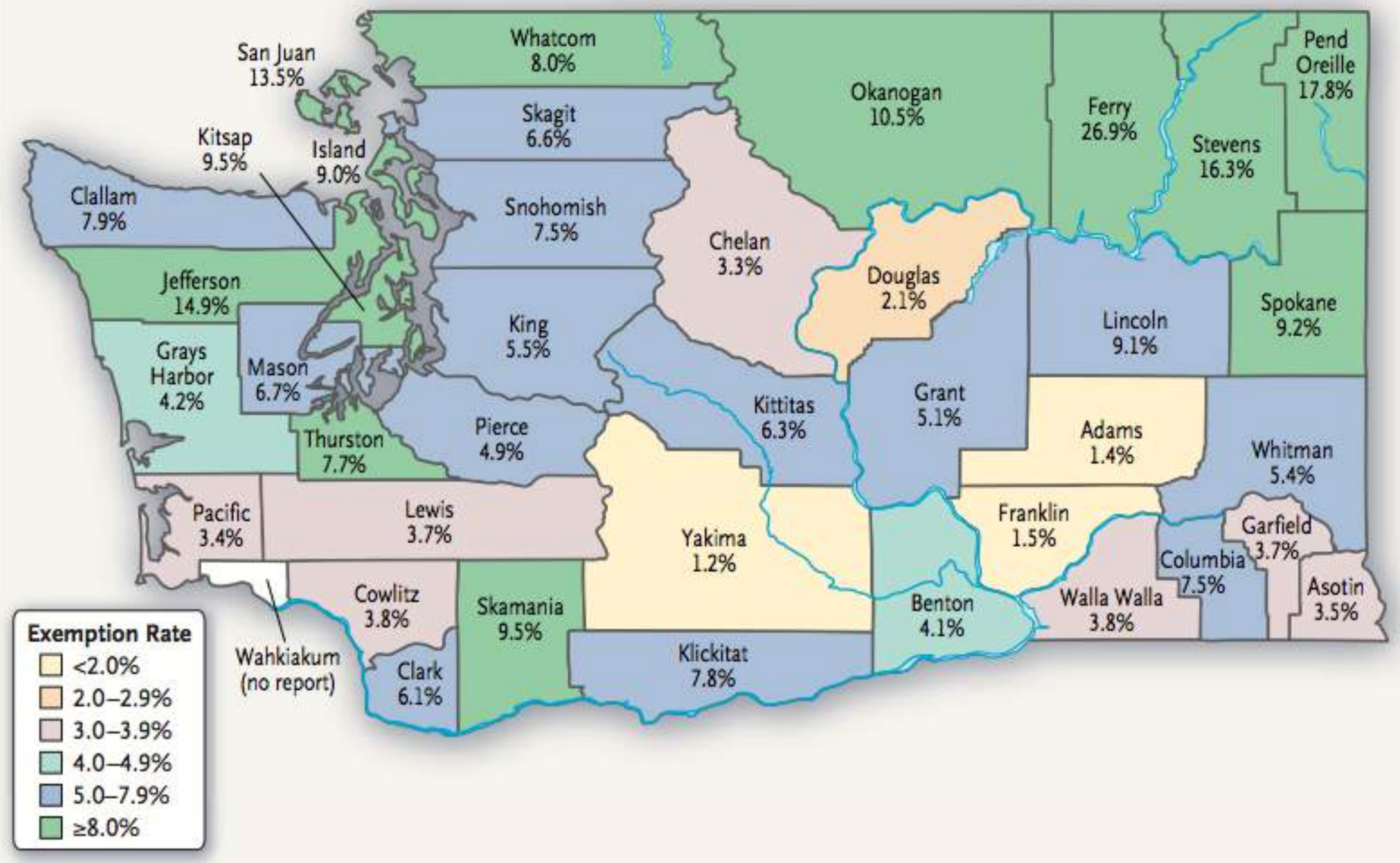
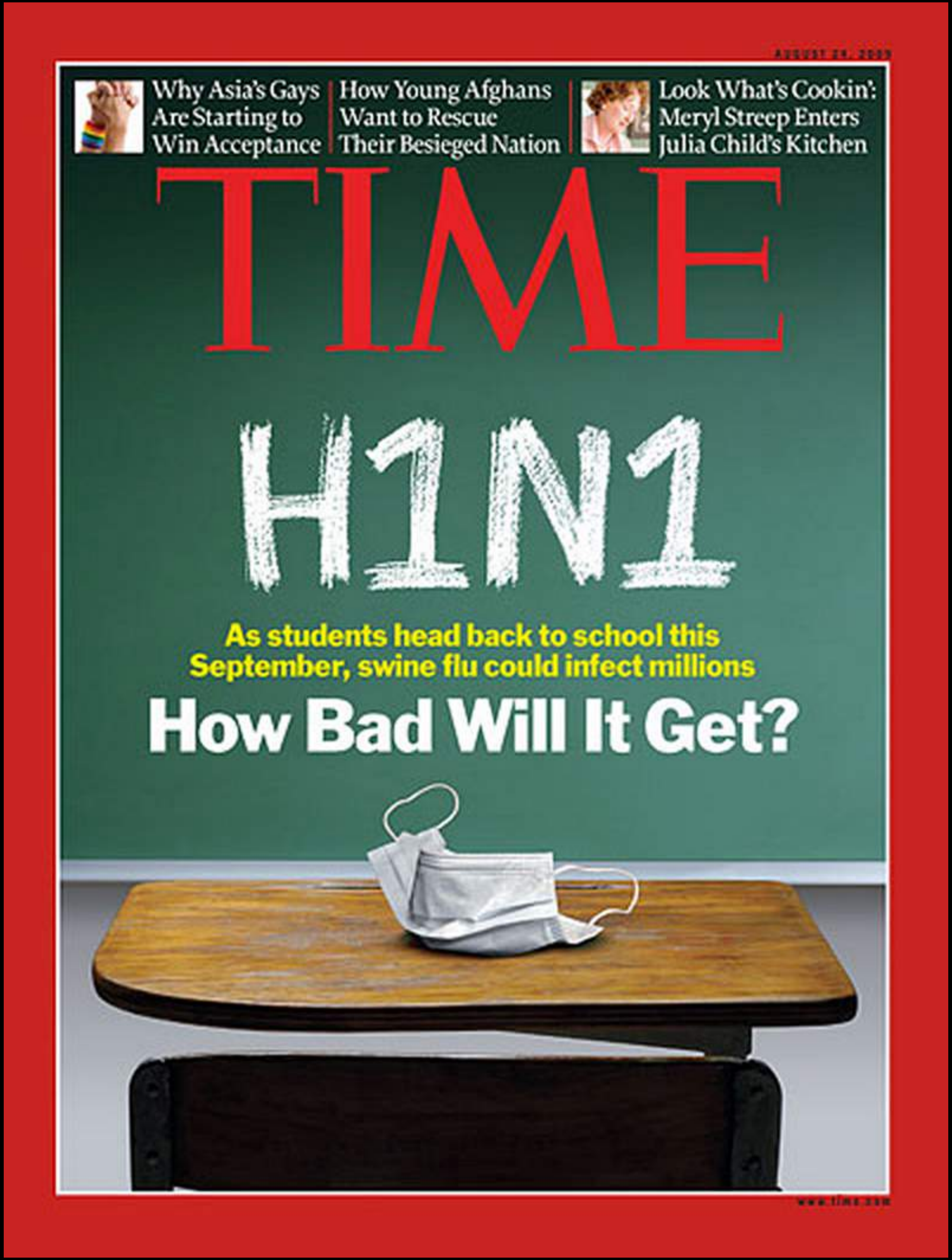


Figure 1. **Rates of Exemption from Vaccination for Nonmedical Reasons in Washington Counties, 2006–2007.** Data are from the Washington State Department of Health, School Status Reports, 2006–2007.[25]

Omer et al., NEJM, 2009



AUGUST 24, 2009

Why Asia's Gays
Are Starting to
Win Acceptance

How Young Afghans
Want to Rescue
Their Besieged Nation

Look What's Cookin':
Meryl Streep Enters
Julia Child's Kitchen

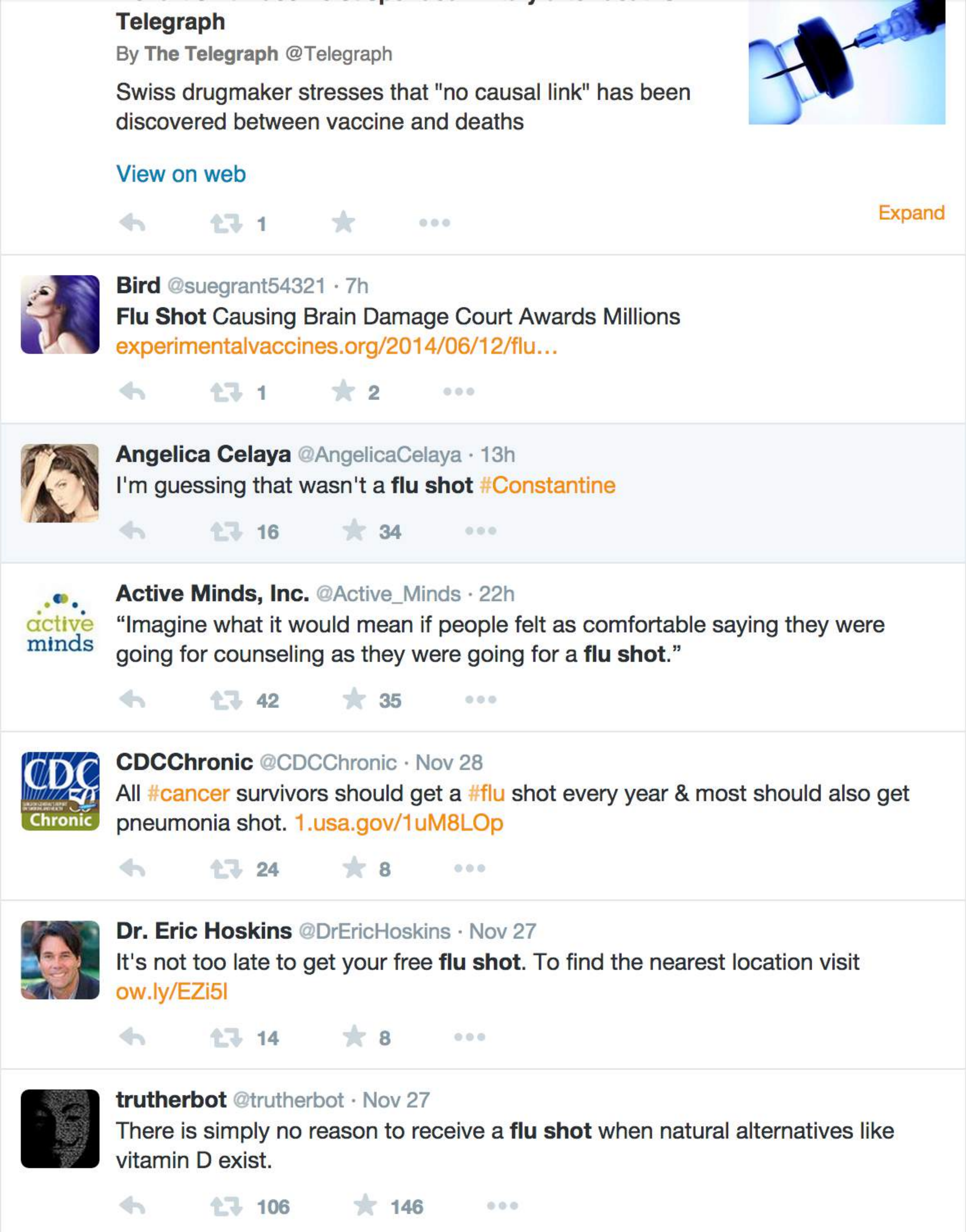
TIME

H1N1

As students head back to school this
September, swine flu could infect millions

How Bad Will It Get?

www.time.com



Telegraph
By **The Telegraph** @Telegraph
Swiss drugmaker stresses that "no causal link" has been discovered between vaccine and deaths

Bird @suegrant54321 · 7h
Flu Shot Causing Brain Damage Court Awards Millions experimentalvaccines.org/2014/06/12/flu…

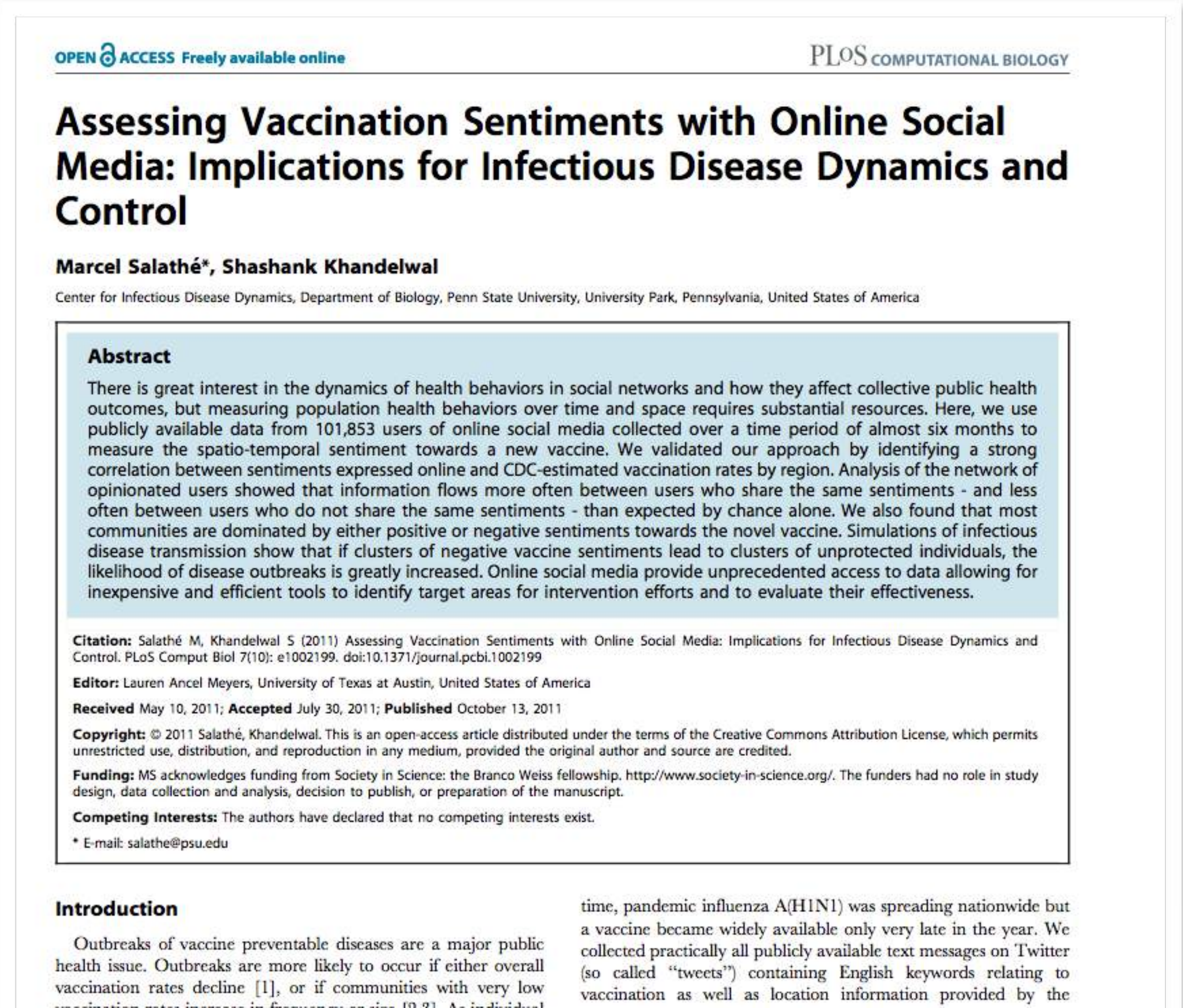
Angelica Celaya @AngelicaCelaya · 13h
I'm guessing that wasn't a **flu shot** #Constantine

Active Minds, Inc. @Active_Minds · 22h
“Imagine what it would mean if people felt as comfortable saying they were going for counseling as they were going for a **flu shot**.”

CDCChronic @CDCChronic · Nov 28
All #cancer survivors should get a #flu shot every year & most should also get pneumonia shot. 1.usa.gov/1uM8LOp

Dr. Eric Hoskins @DrEricHoskins · Nov 27
It's not too late to get your free **flu shot**. To find the nearest location visit ow.ly/EZi5l

trutherbot @trutherbot · Nov 27
There is simply no reason to receive a **flu shot** when natural alternatives like vitamin D exist.



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.

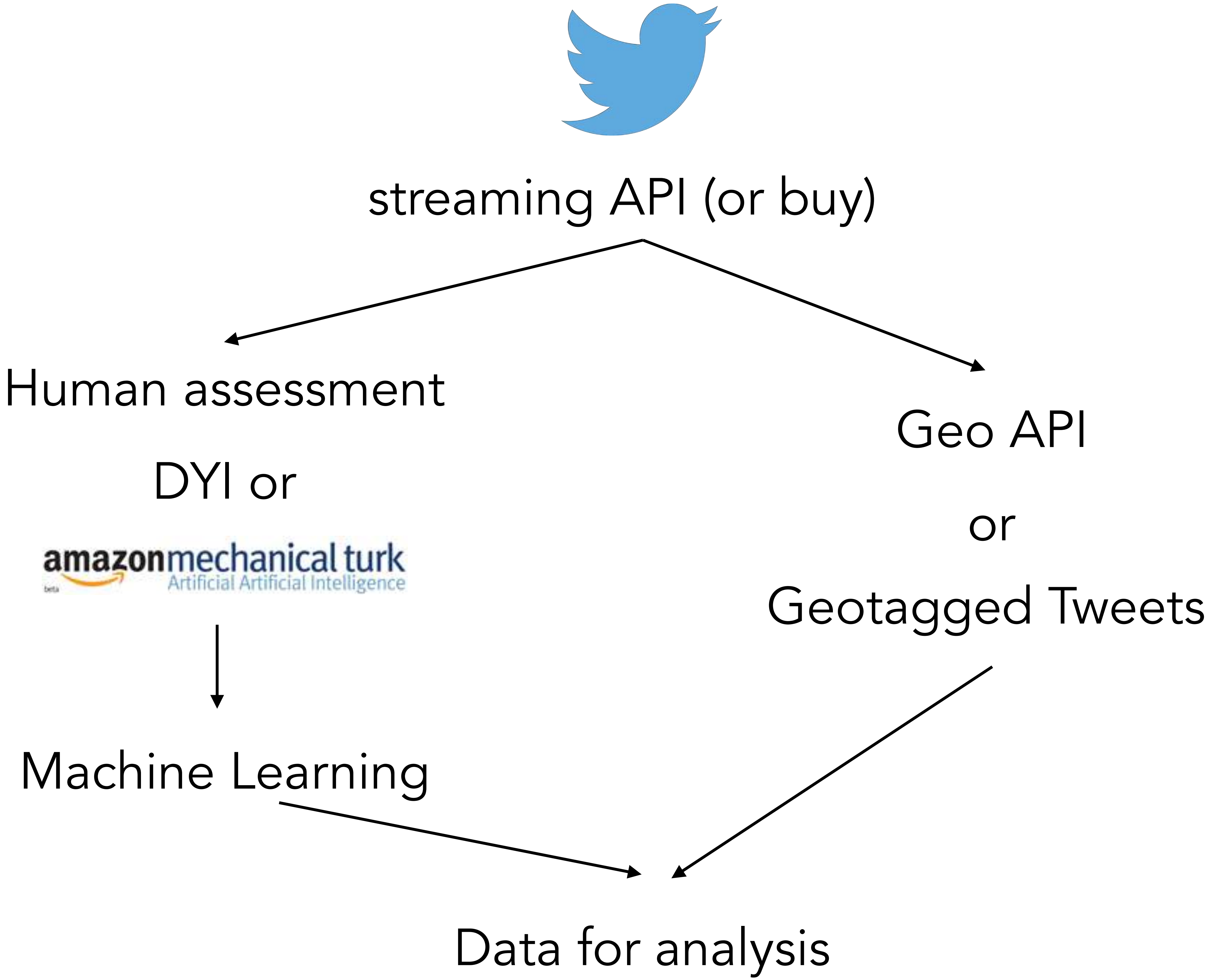
Competing Interests: The authors have declared that no competing interests exist.

* E-mail: salathe@psu.edu

Introduction

Outbreaks of vaccine preventable diseases are a major public health issue. Outbreaks are more likely to occur if either overall vaccination rates decline [1], or if communities with very low

time, pandemic influenza A(H1N1) was spreading nationwide but a vaccine became widely available only very late in the year. We collected practically all publicly available text messages on Twitter (so called "tweets") containing English keywords relating to vaccination as well as location information provided by the



streaming API (or buy)

Human assessment

DYI or

amazonmechanical turk
beta Artificial Artificial Intelligence

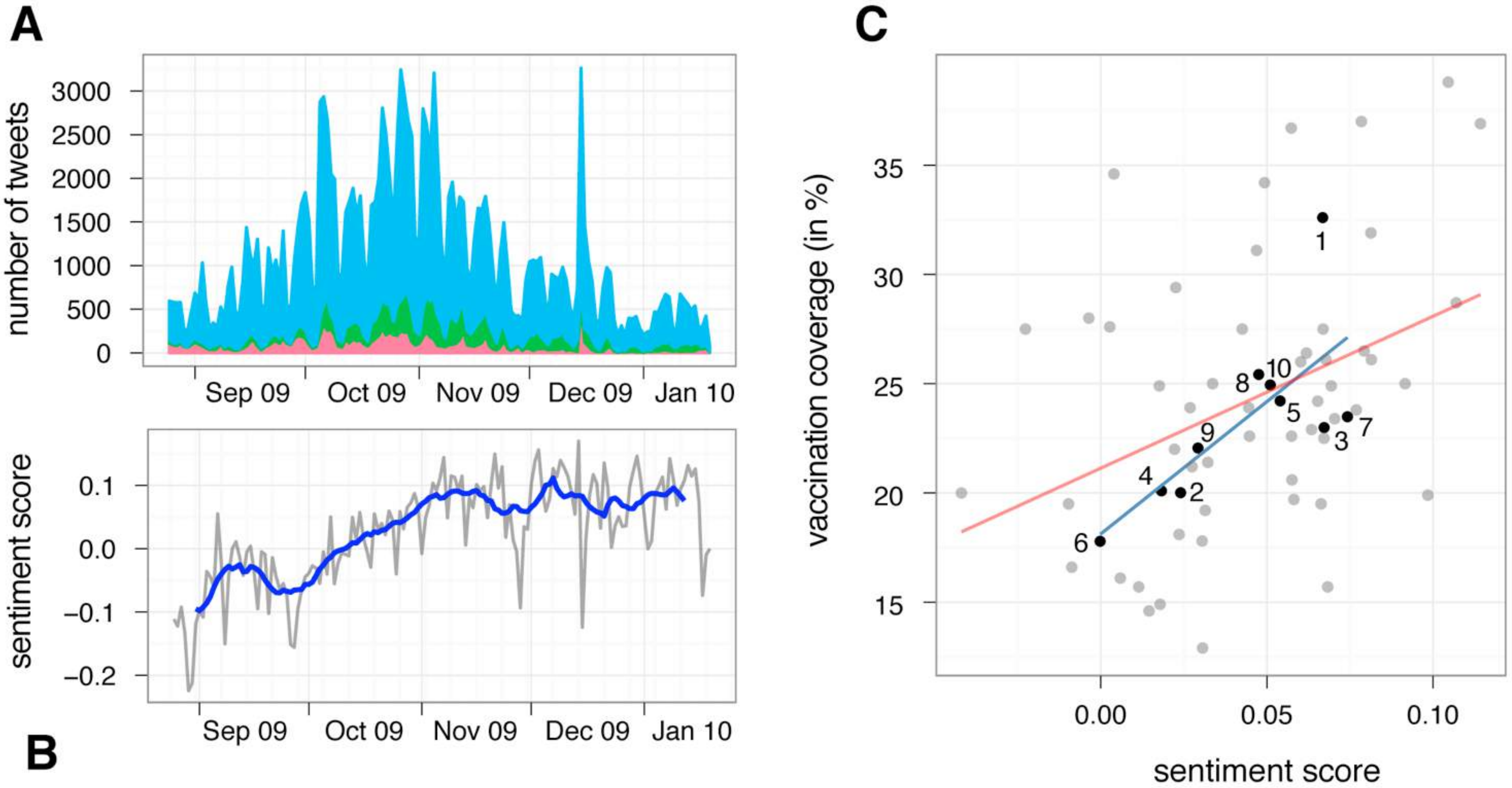
Machine Learning

Geo API

or

Geotagged Tweets

Data for analysis



(~10% Tweets were assessed by humans, rest by machine learning algorithms)

Salathé & Khandelwal, PLOS Computational Biology, 2011

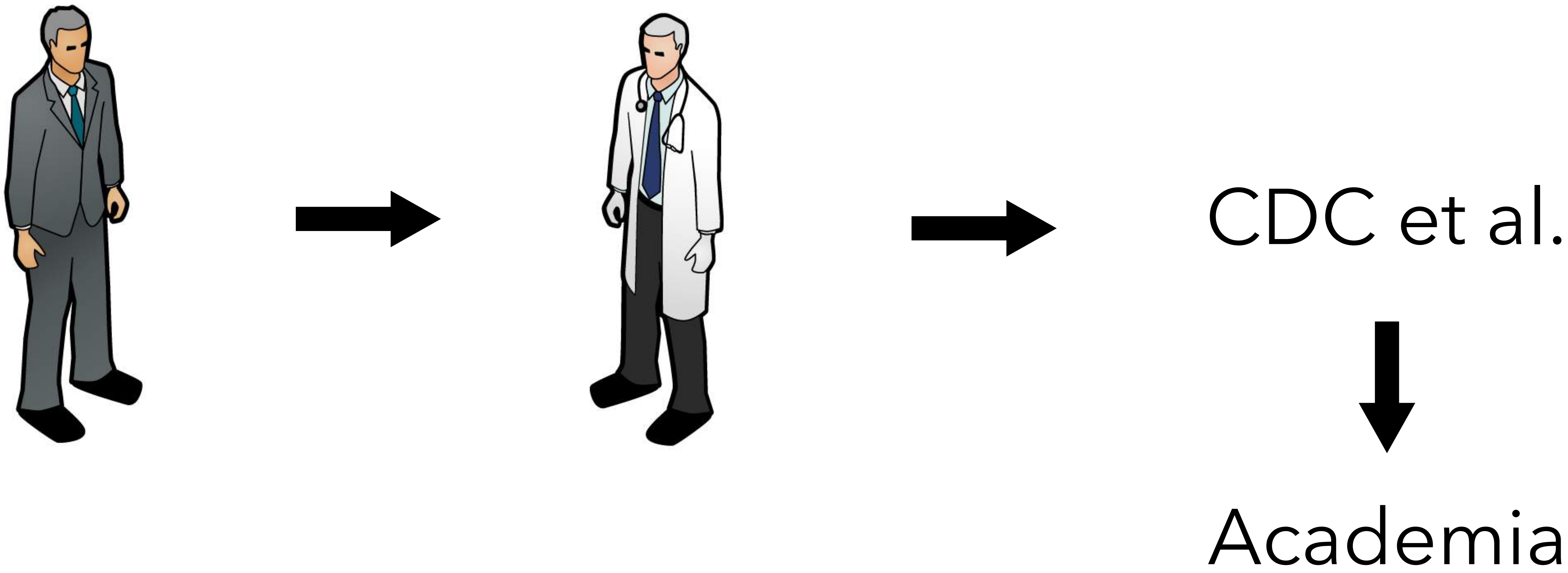
Digital Epidemiology

Epidemiology: Epidemiology is the study and analysis of the patterns, causes, and effects of health and disease conditions in defined populations.

Digital Epidemiology: Epidemiology with digital data that captures states, events, processes, etc. that are difficult to capture otherwise*.

*** what || where || when**

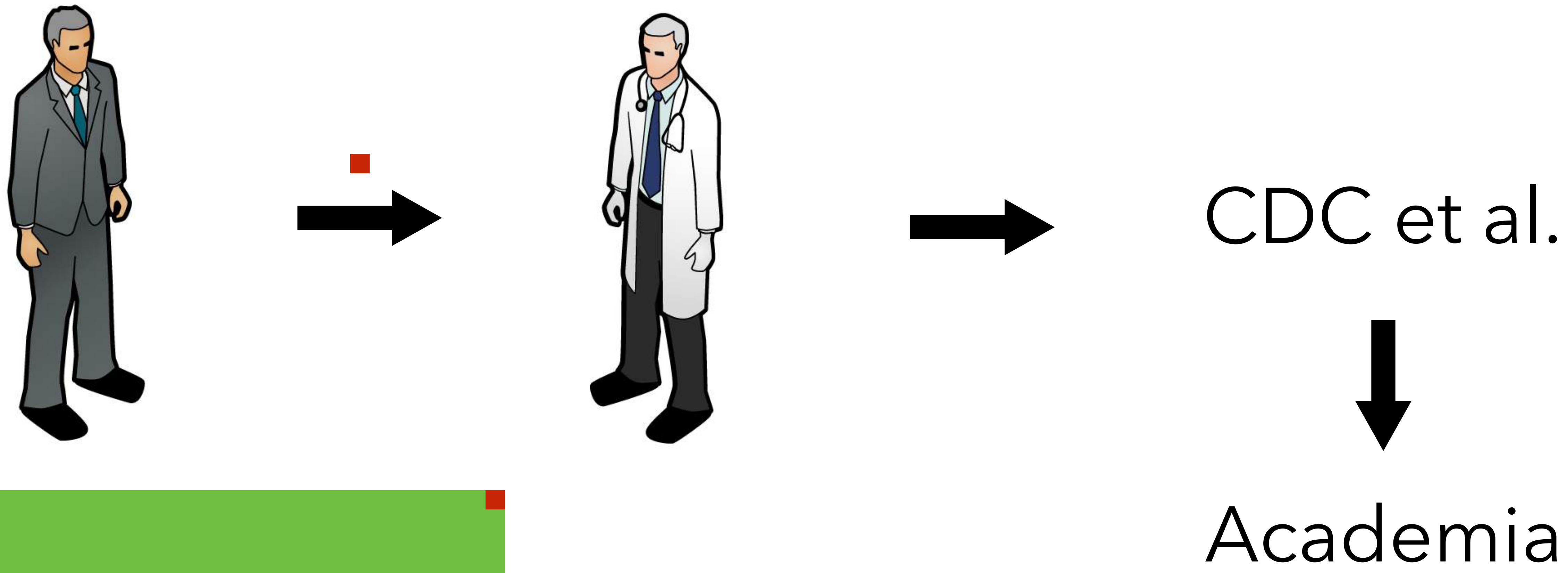
Traditional Epidemiology



CDC et al.

Academia

Traditional Epidemiology



CDC et al.

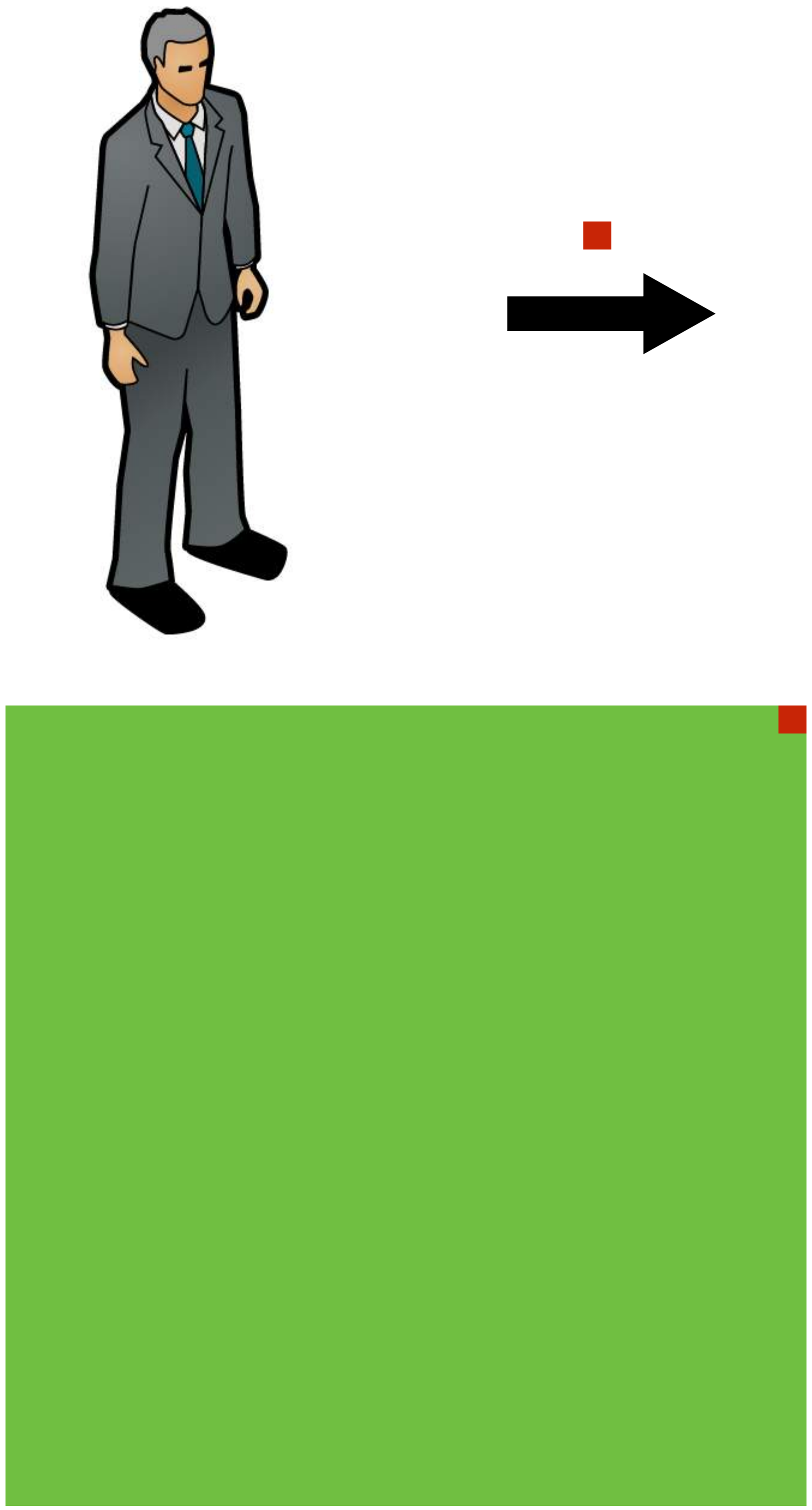
Academia

Traditional Epidemiology

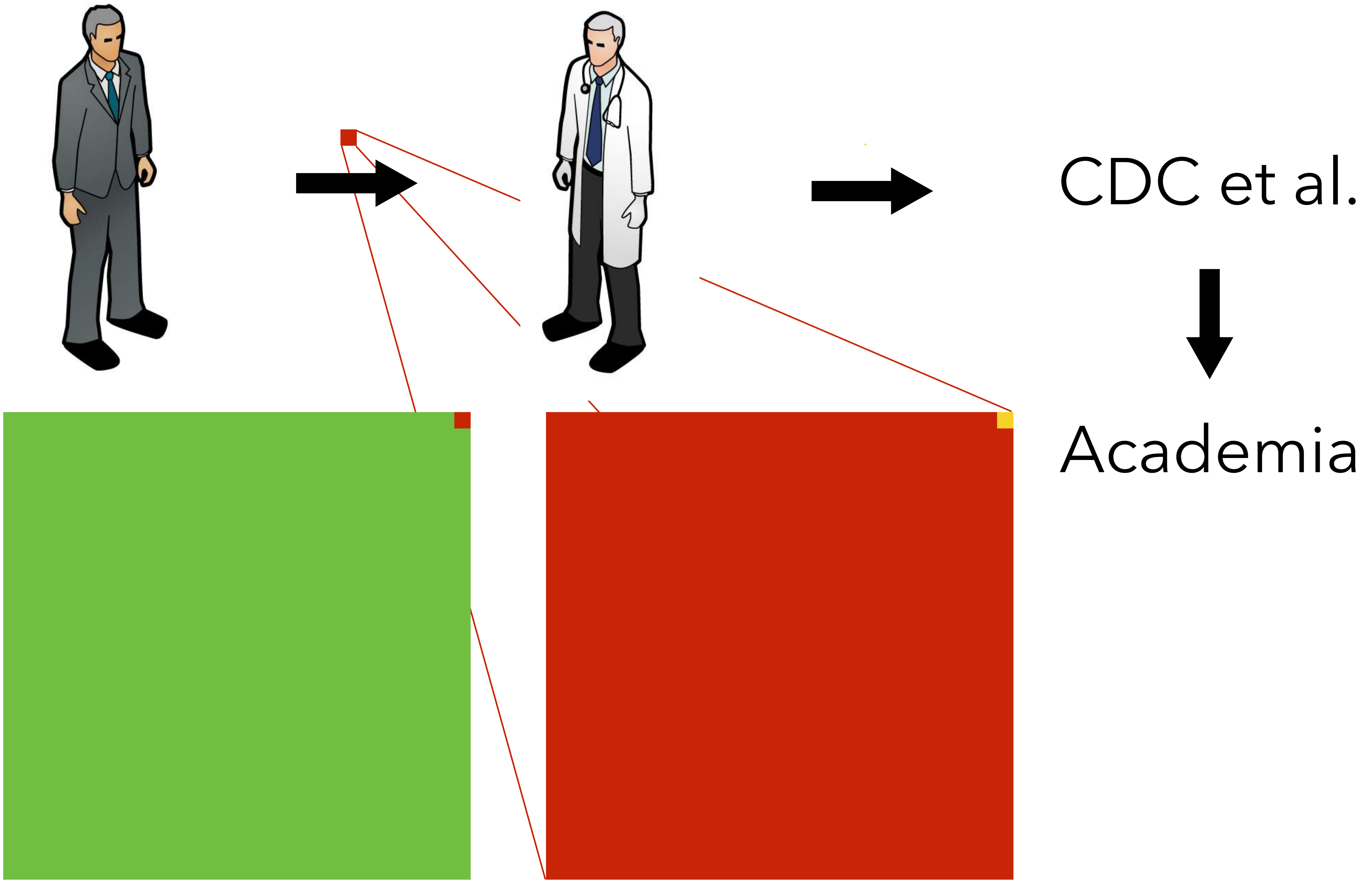
CDC et al.

Academia

Traditional Epidemiology



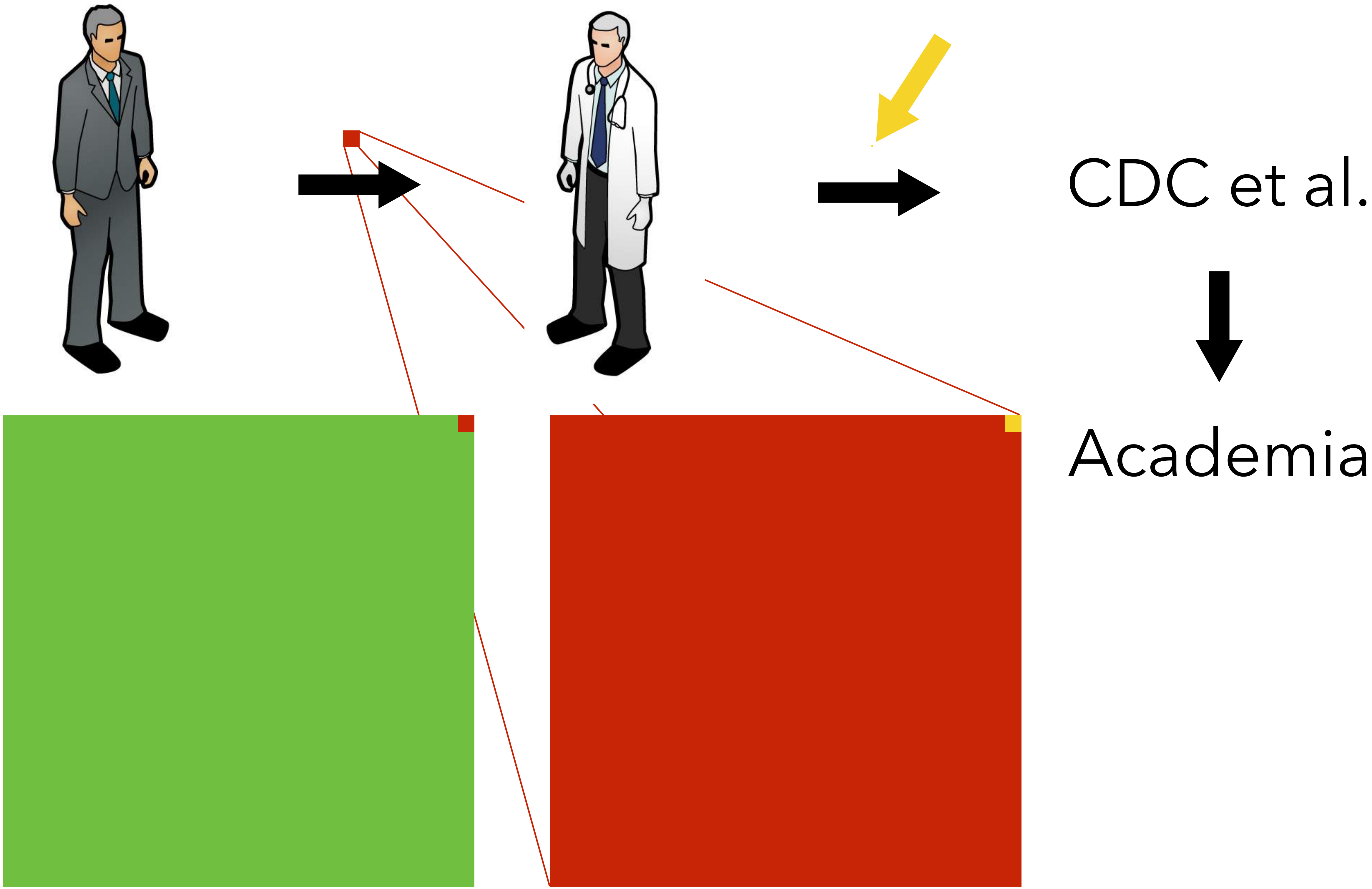
Traditional Epidemiology



CDC et al.

Academia

Traditional Epidemiology



CDC et al.

Academia

Digital Epidemiology: new data streams

"Got my flu shot this morning and now my throat is sore."

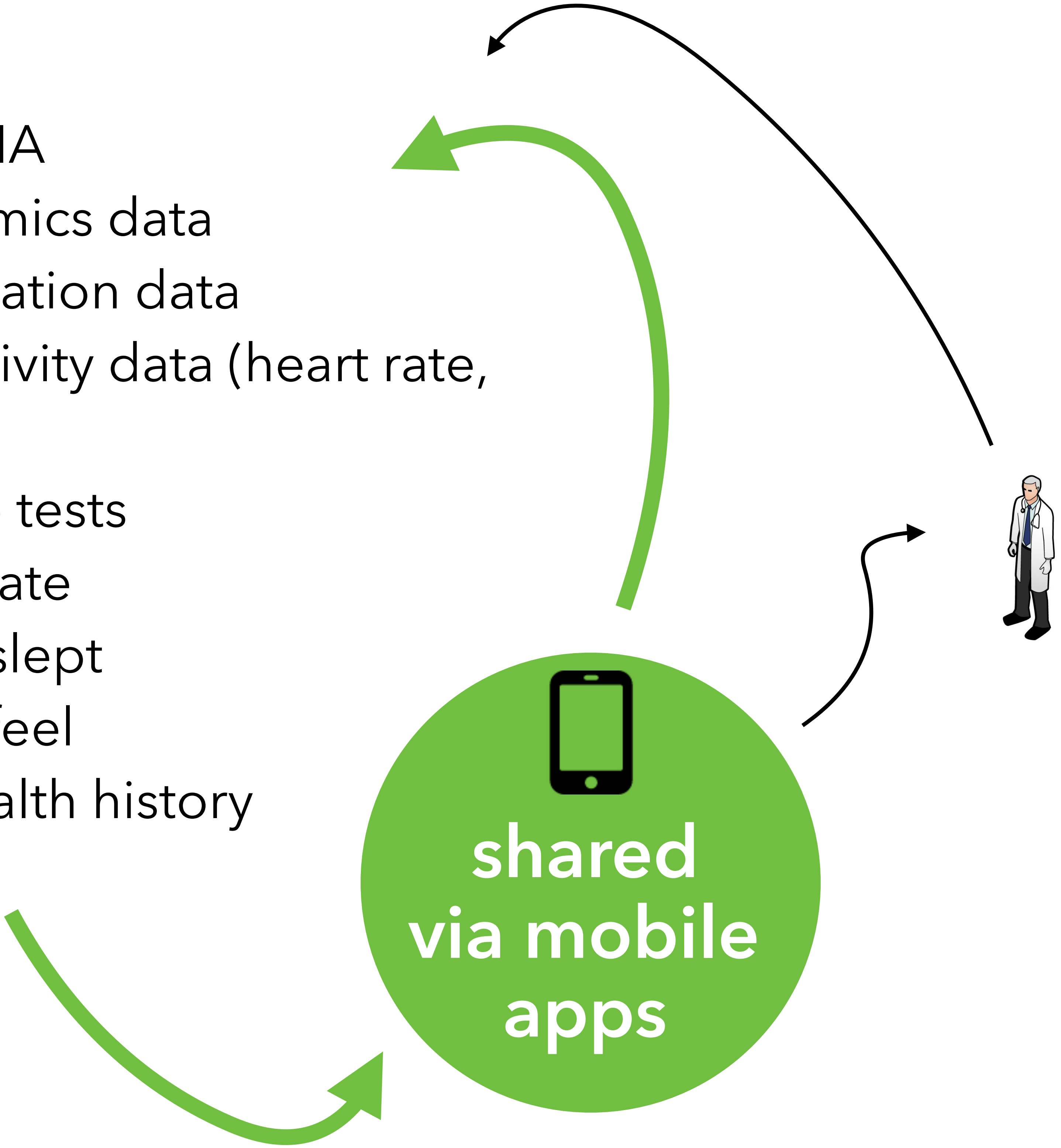
"Stomach flu & normal flu in the same month. I'm officially a germaphobe."

Text
Images
Videos
Sounds
Location
Biological data
etc.

"My weight: 170.1 lb. 10.1 lb to go. #raceweight @Withings scale auto-tweets my weight once a week http://withings.com"

"Such an upset stomach today. I hope it's just a bug and not the Truvada."

From Personalized Health…

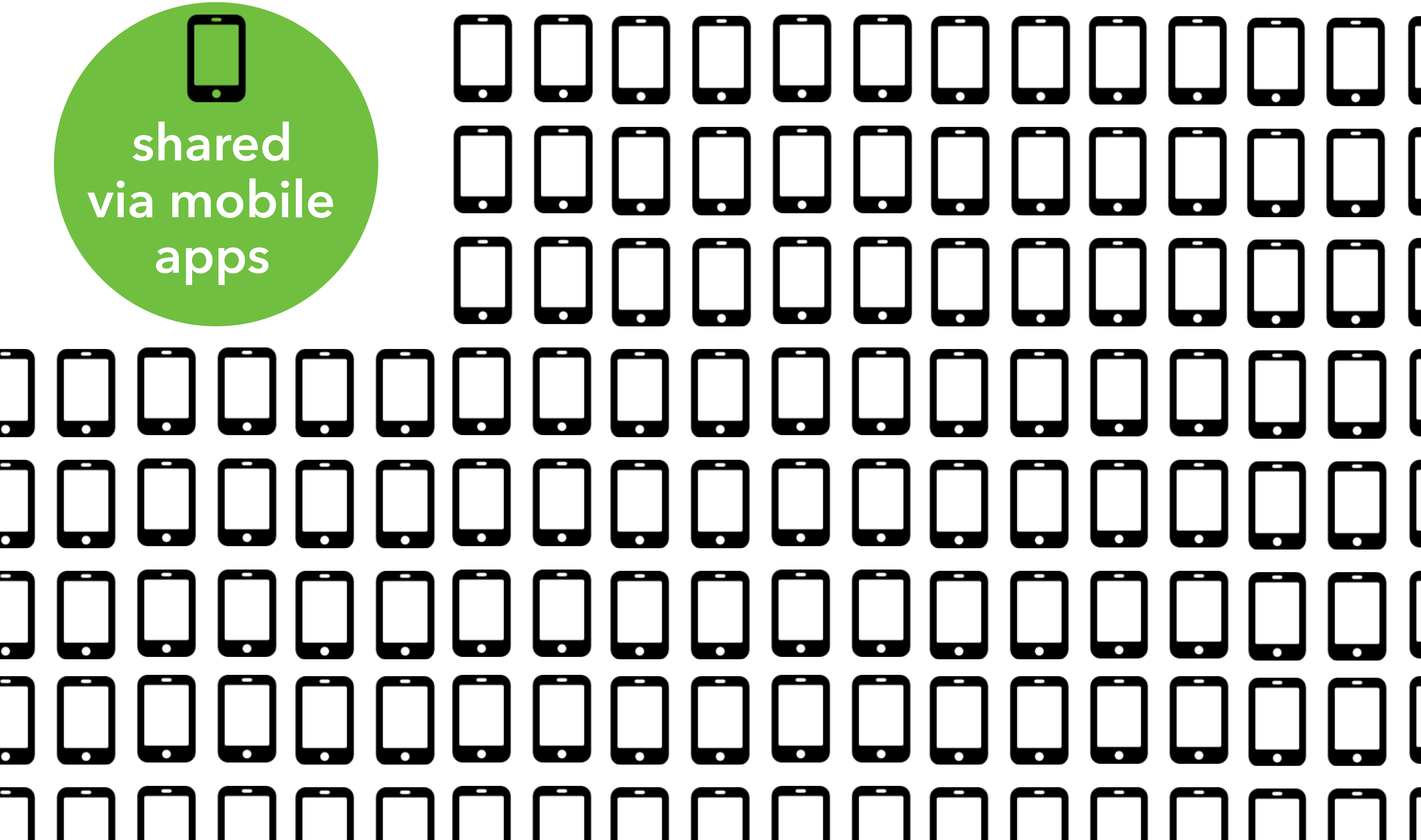


- my DNA
- my *omics data
- my location data
- my activity data (heart rate, etc.)
- my lab tests
- what I ate
- how I slept
- how I feel
- my health history
- etc.

shared via mobile apps

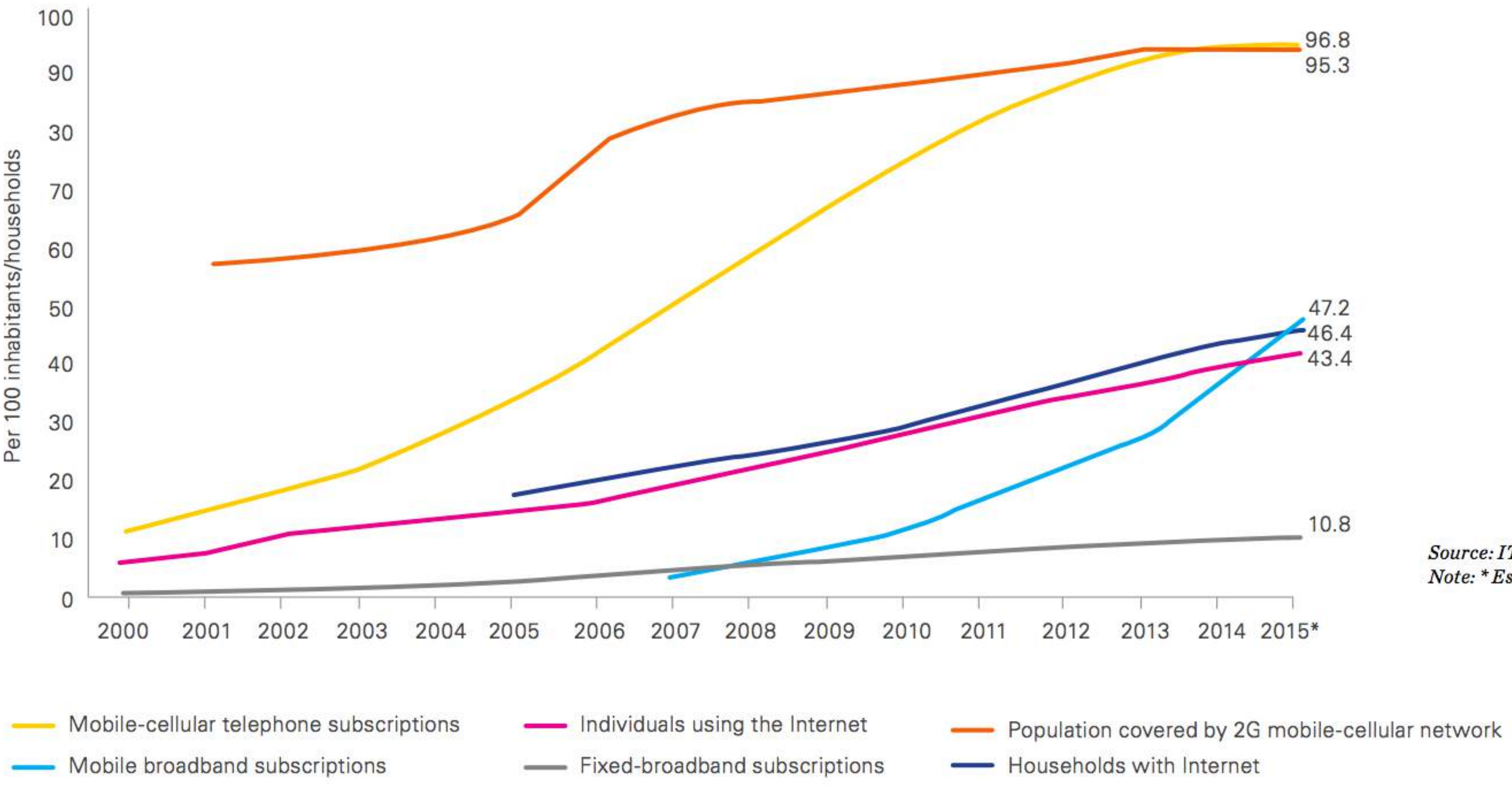
"The patient will see you now"

…To Truly Global Health



shared
via mobile
apps

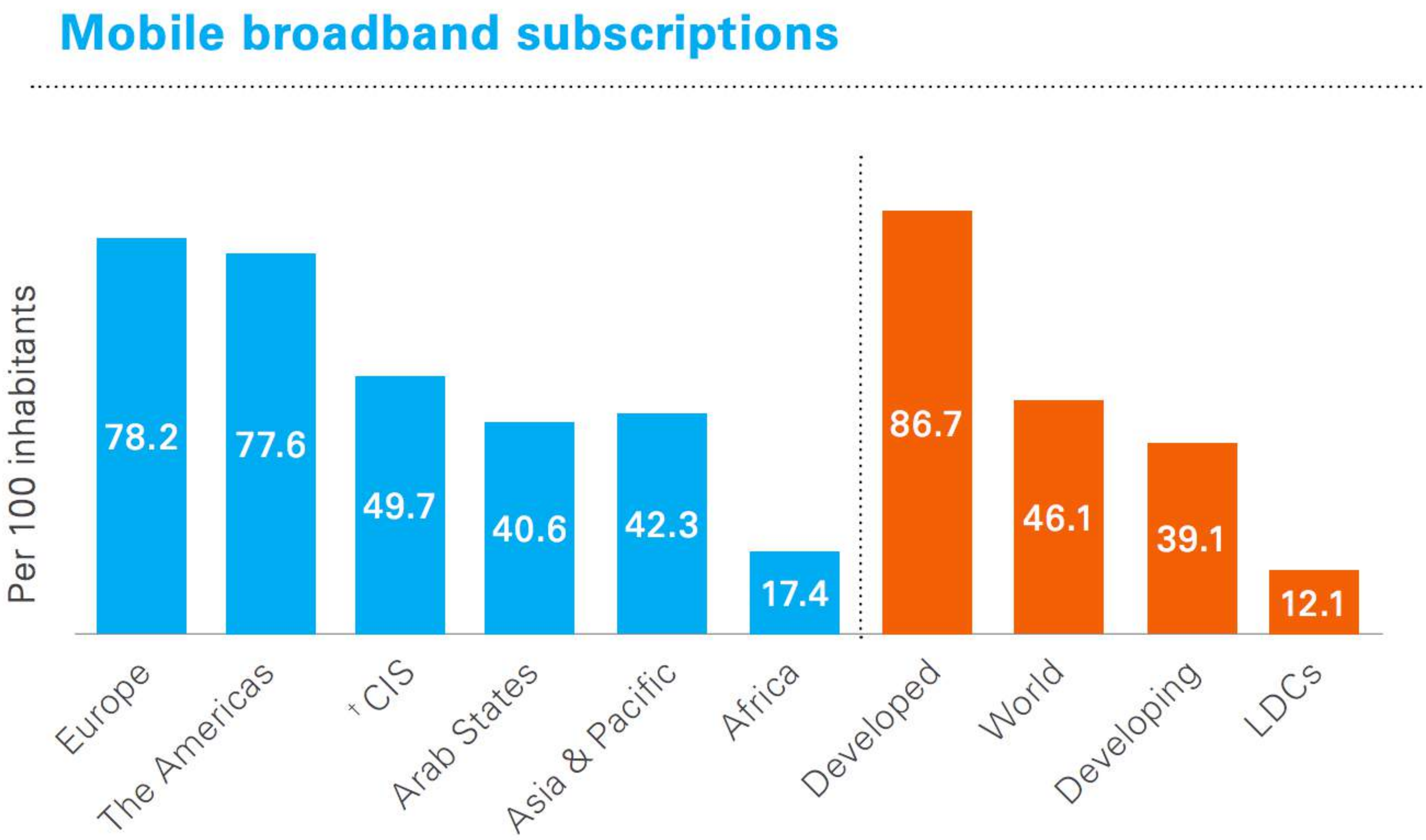
Mobile broadband



Source: ITU.
*Note: *Estimates.*

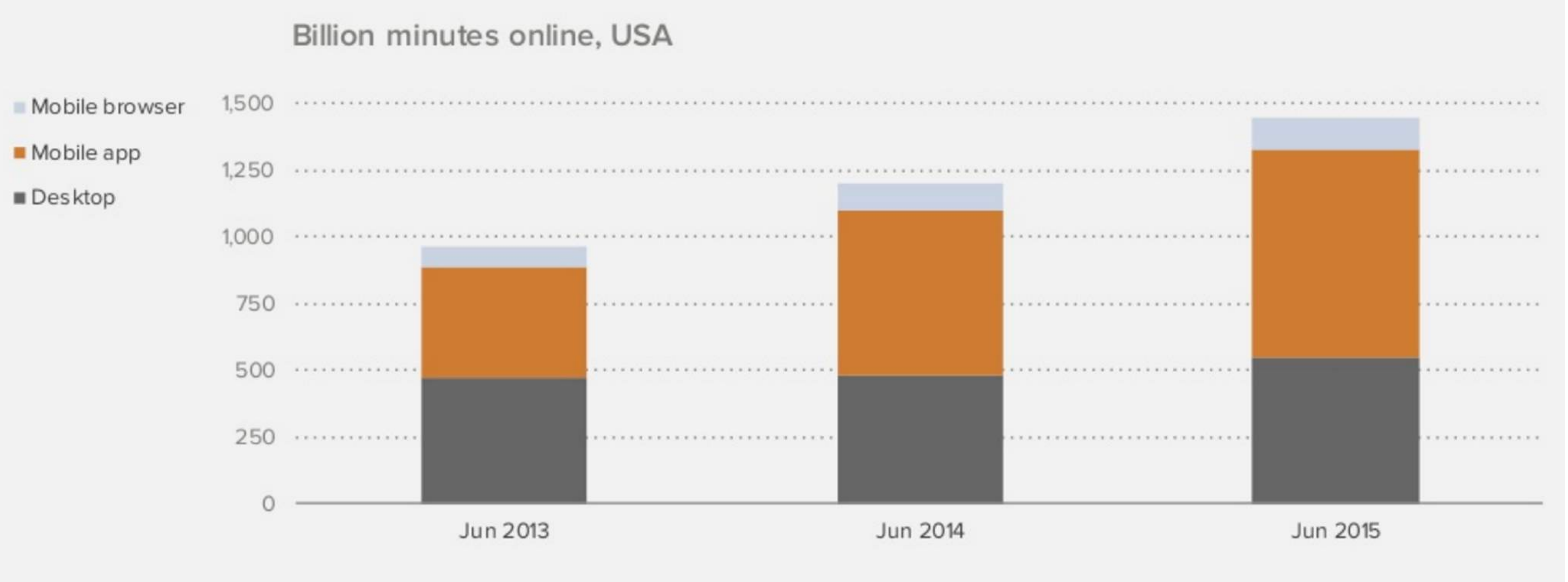
https://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2015.pdf

Mobile broadband



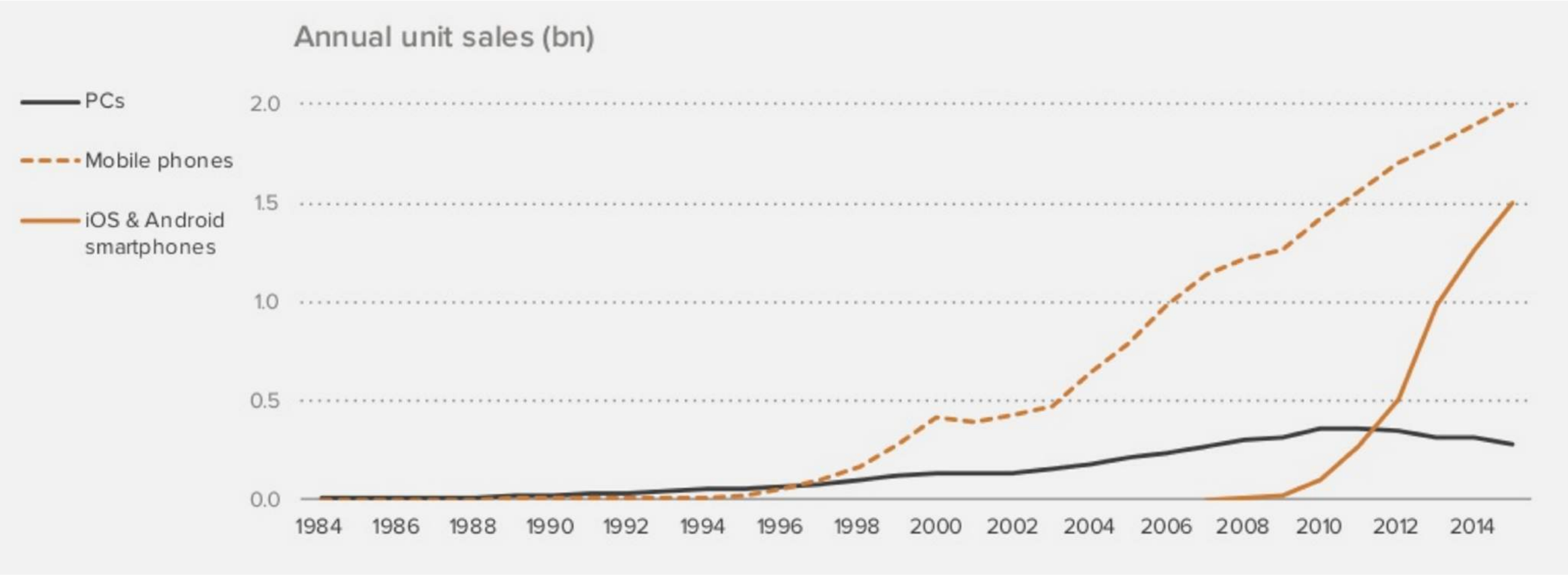
https://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2015.pdf

More time spent on <u>mobile</u> apps than all of web



http://ben-evans.com/benedictevans/?tag=Presentations

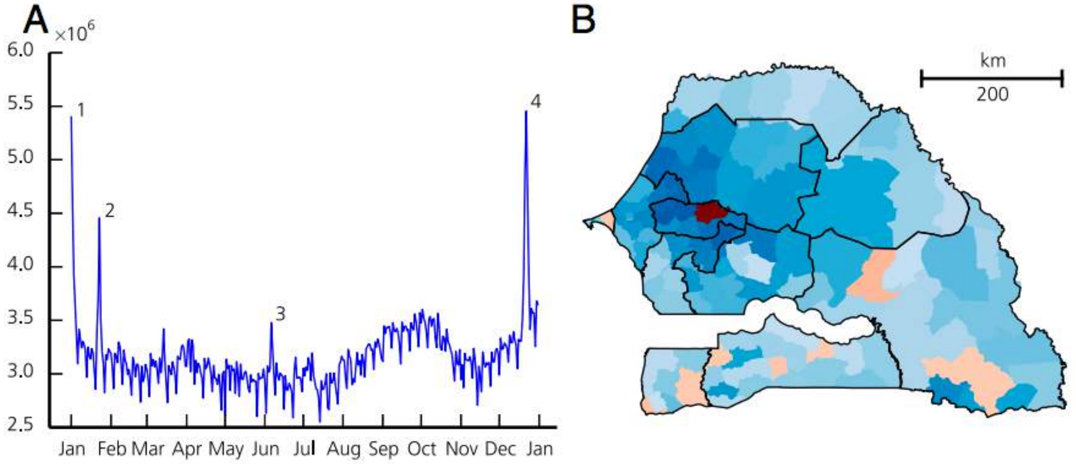
Mobile phone & smart phone revolutions dwarf the PC revolution



http://ben-evans.com/benedictevans/?tag=Presentations

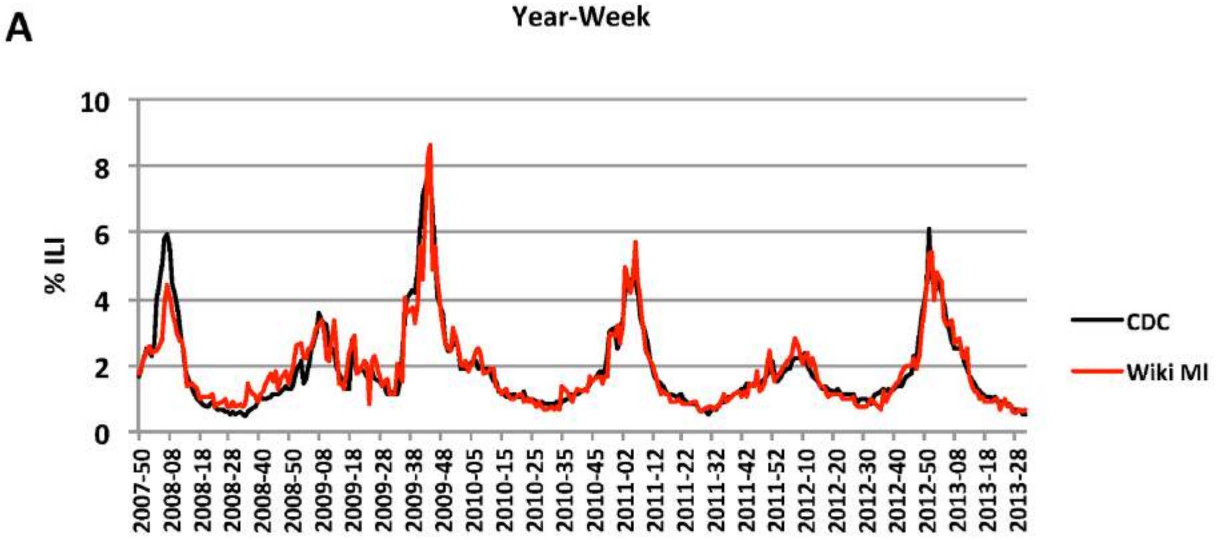
Mobile phones - Track disease outbreaks

Finger et al 2016



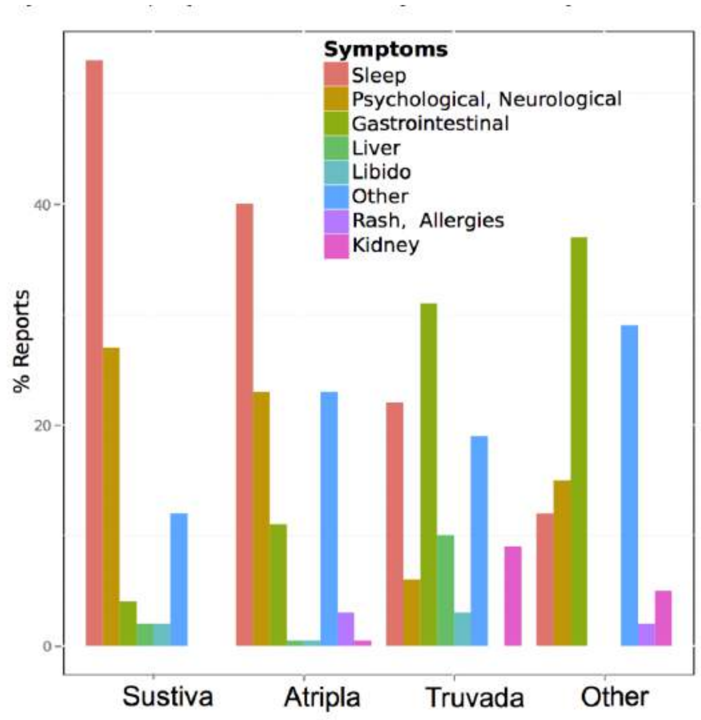
Wikipedia - Influenza forecasting

McIver & Brownstein 2014



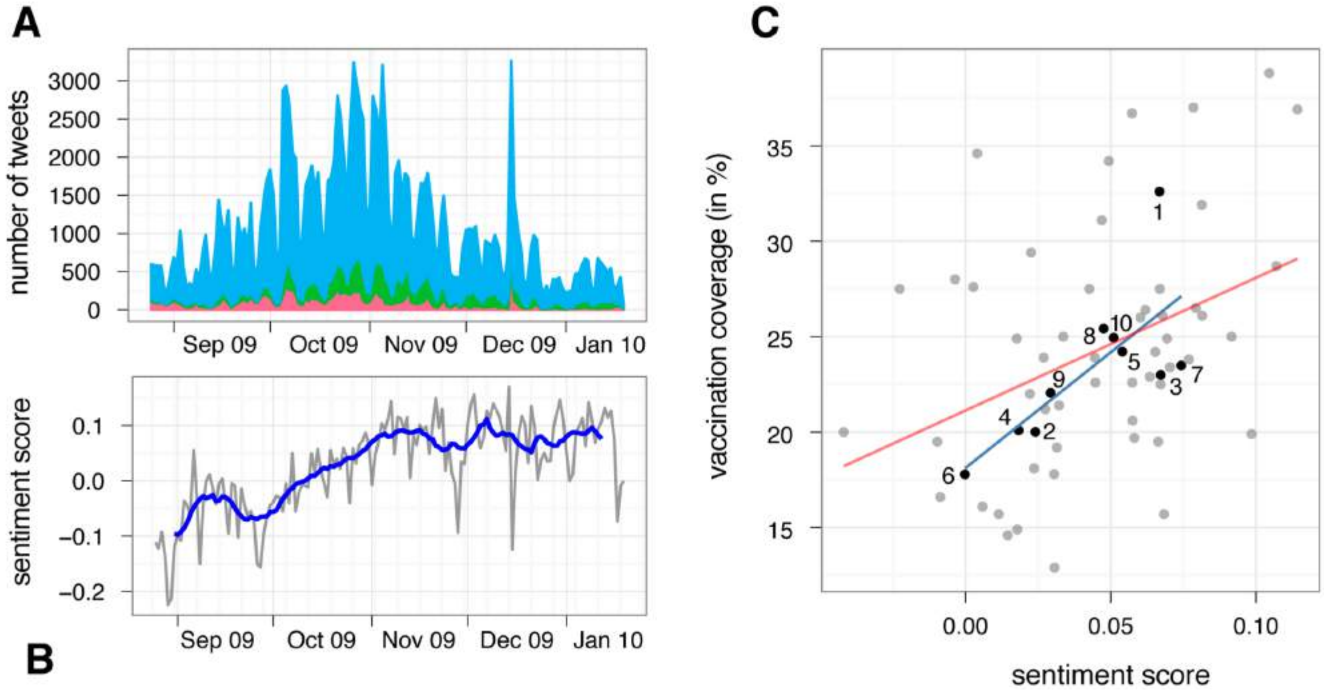
Twitter - Pharmacovigilance

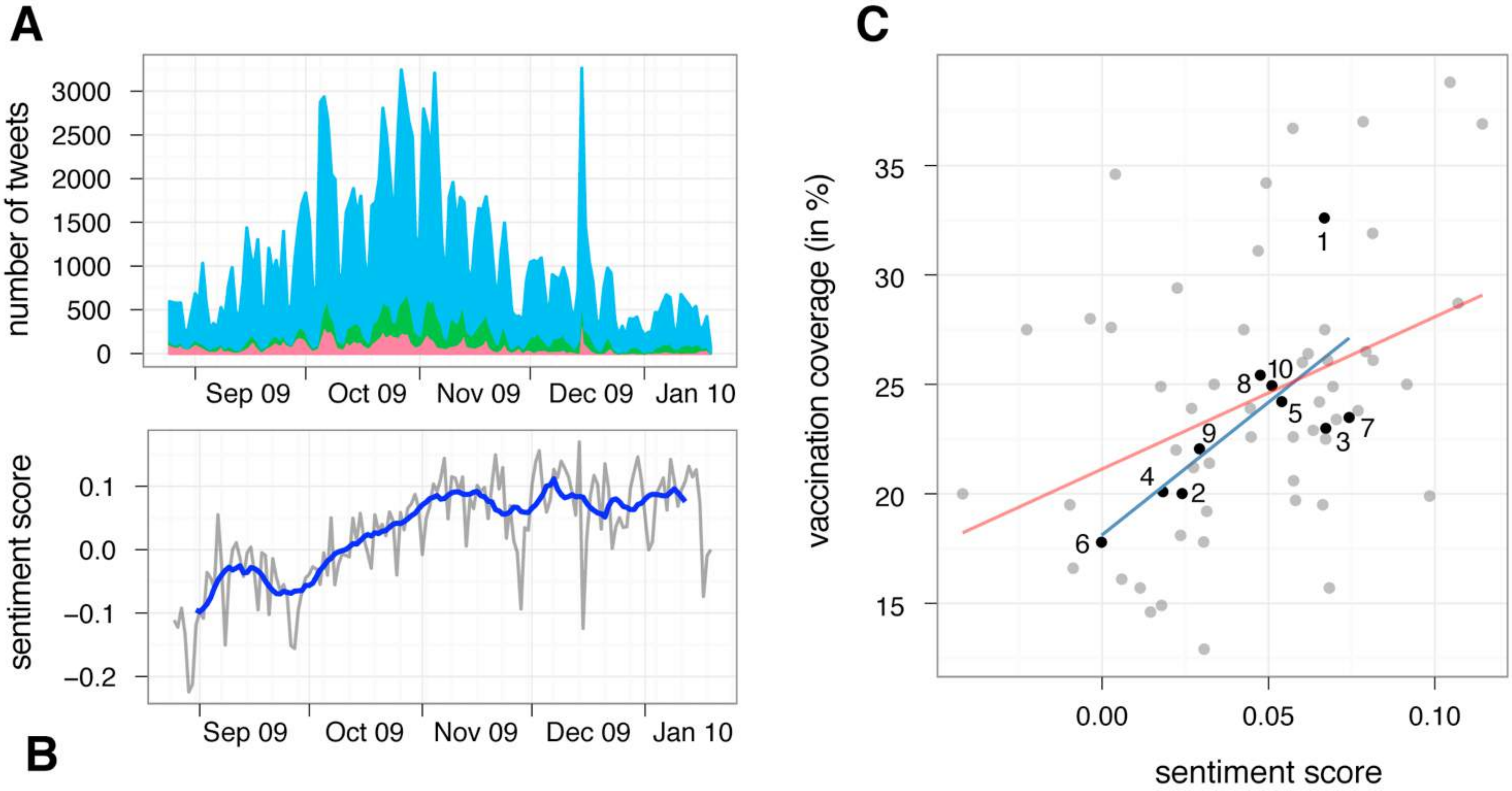
Adrover et al 2015



Twitter - Vaccine uptake

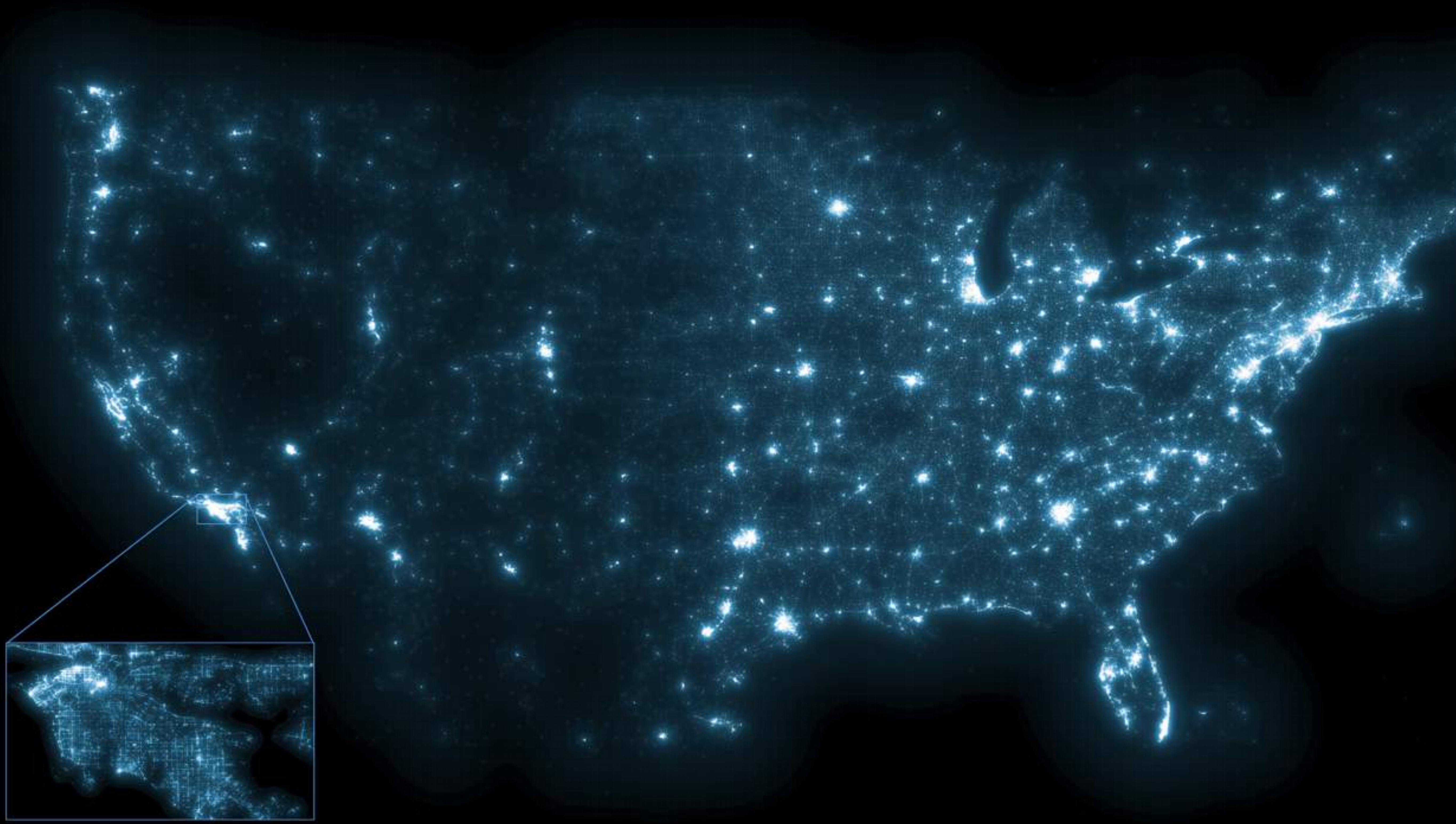
Salathé & Khandelwal, 2011



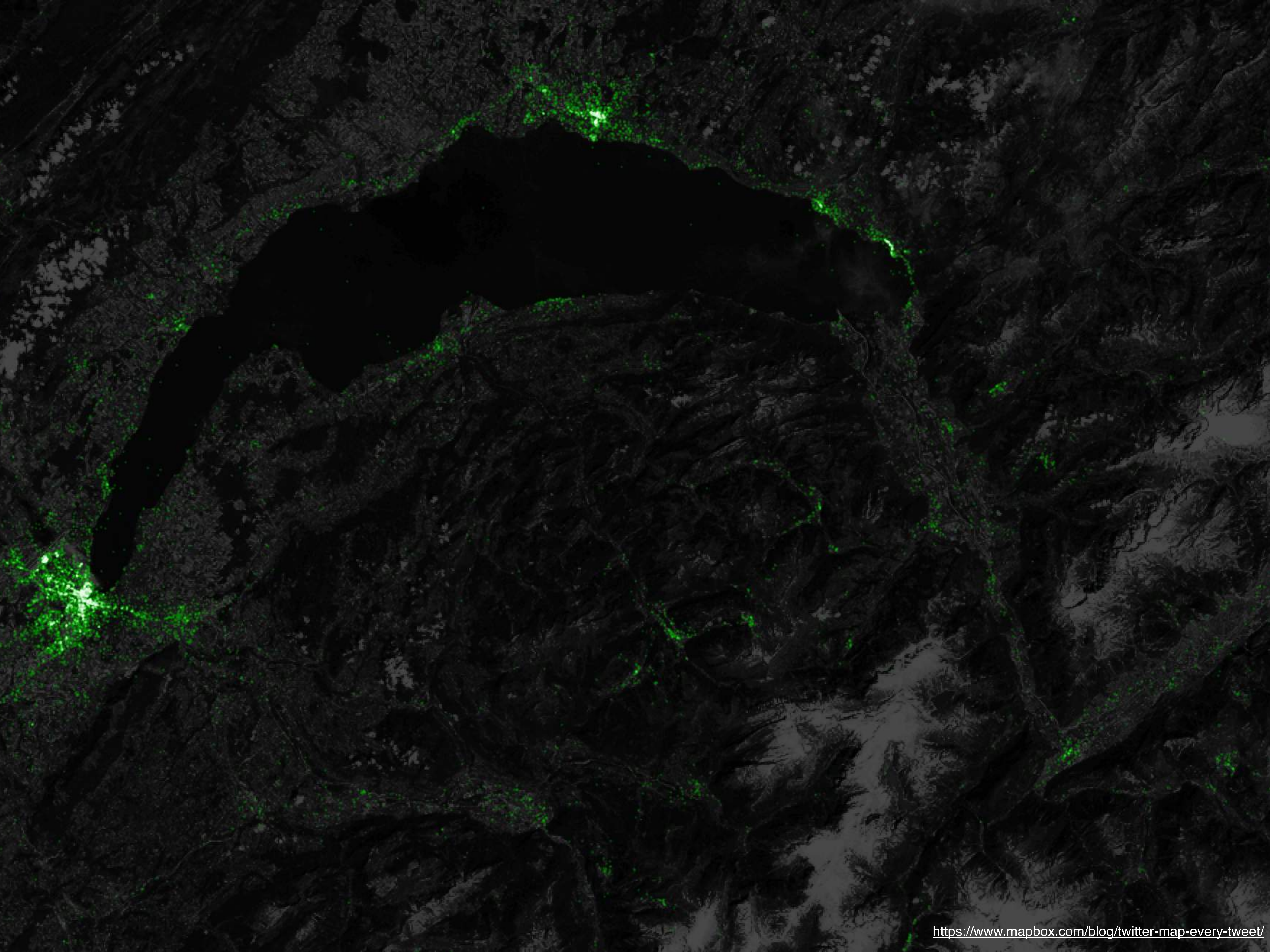


(~10% Tweets were assessed by humans, rest by machine learning algorithms)

Salathé & Khandelwal, PLOS Computational Biology, 2011



Salathé et al. PLOS Computational Biology 2012



https://www.mapbox.com/blog/twitter-map-every-tweet/

https://www.mapbox.com/blog/twitter-map-every-tweet/

PlantVillage: Machine Learning for Disease Recognition

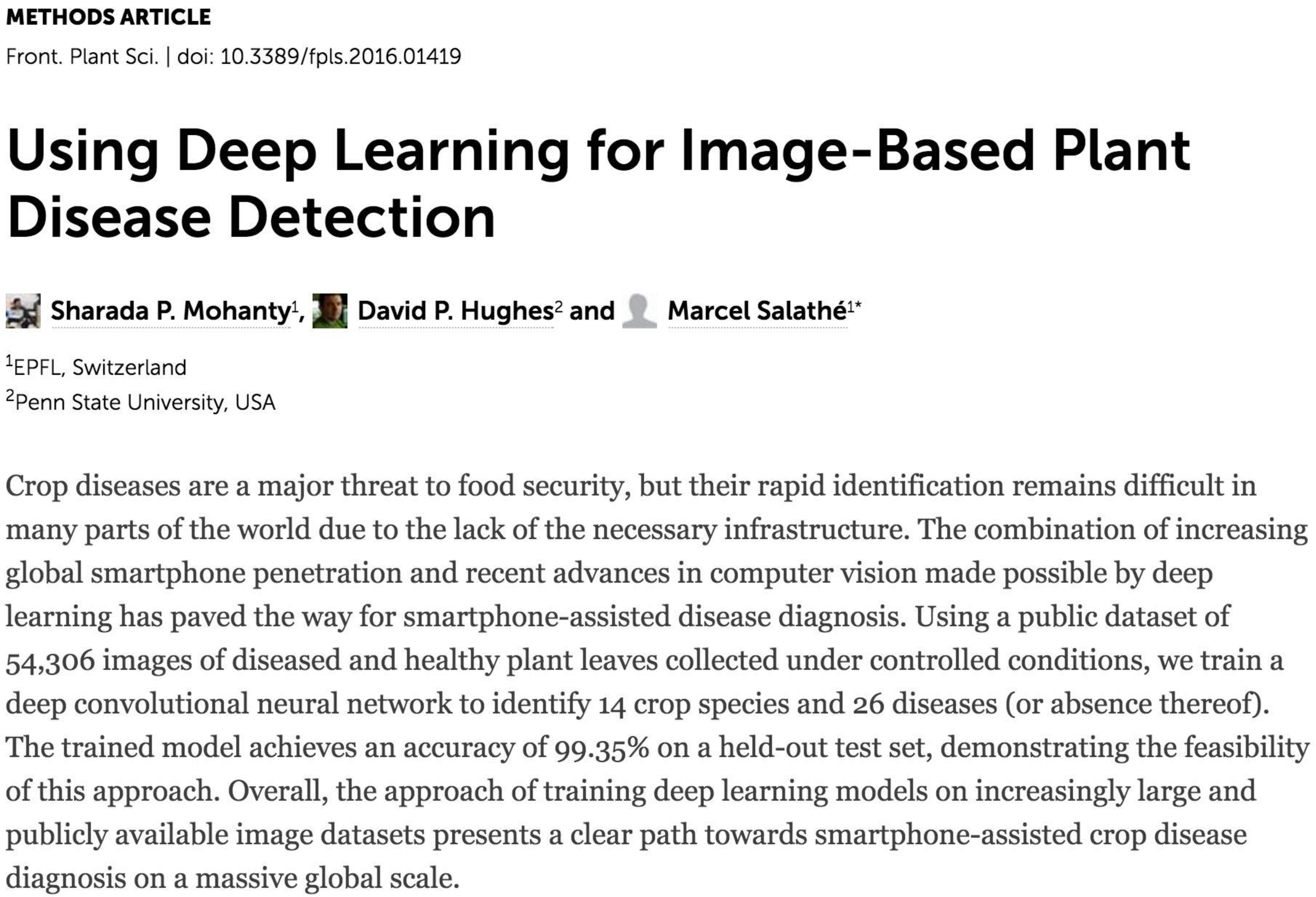
Image Data

Collecting 1M+ labelled images as training set for machine learning algorithm development

Machine Learning

Crowdsourced, open machine learning competitions based on open access images

[Collaboration with Penn State, Prof. David Hughes]



METHODS ARTICLE

Front. Plant Sci. | doi: 10.3389/fpls.2016.01419

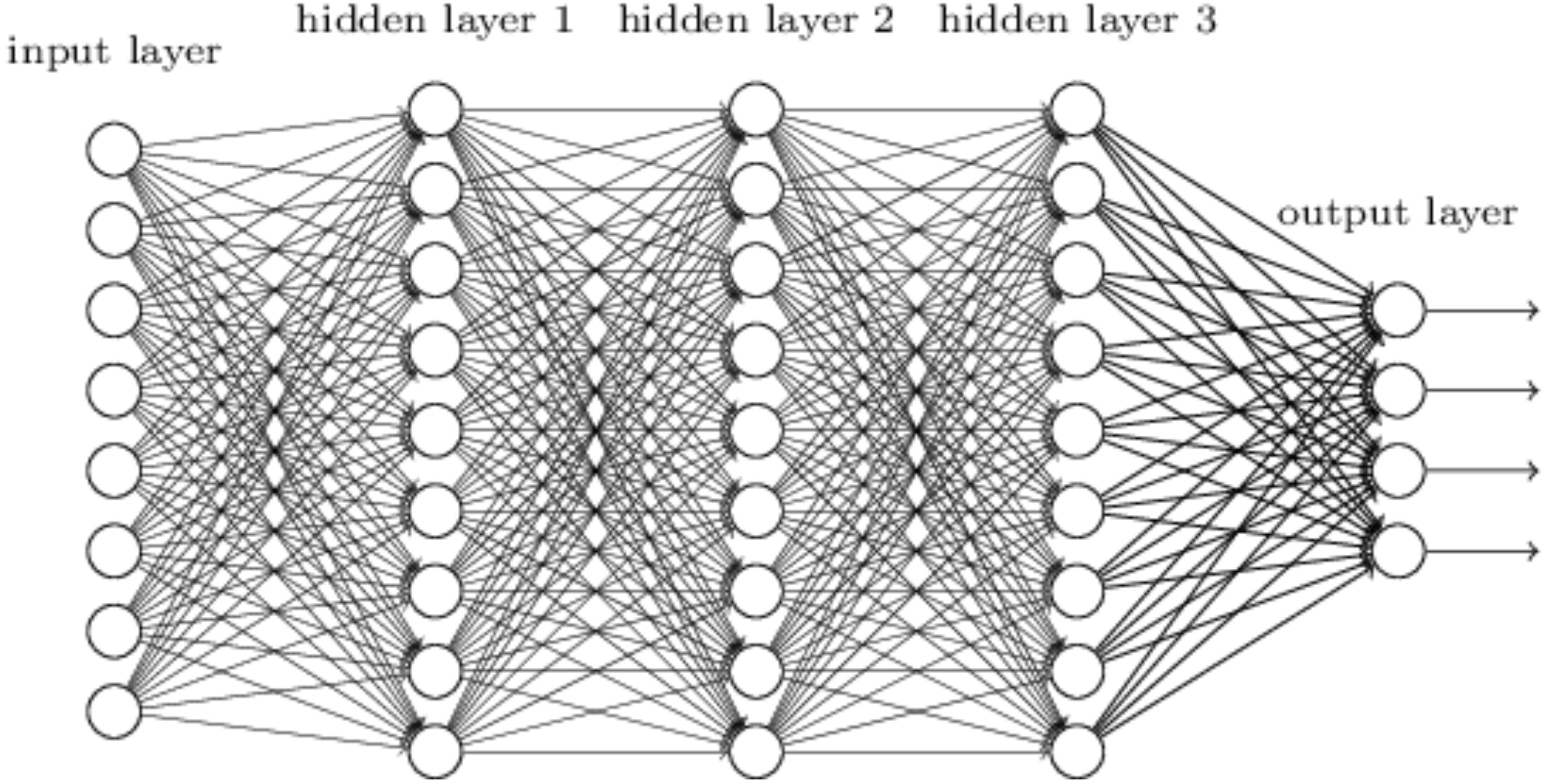
Using Deep Learning for Image-Based Plant Disease Detection

Sharada P. Mohanty[1], **David P. Hughes**[2] and **Marcel Salathé**[1*]

[1]EPFL, Switzerland
[2]Penn State University, USA

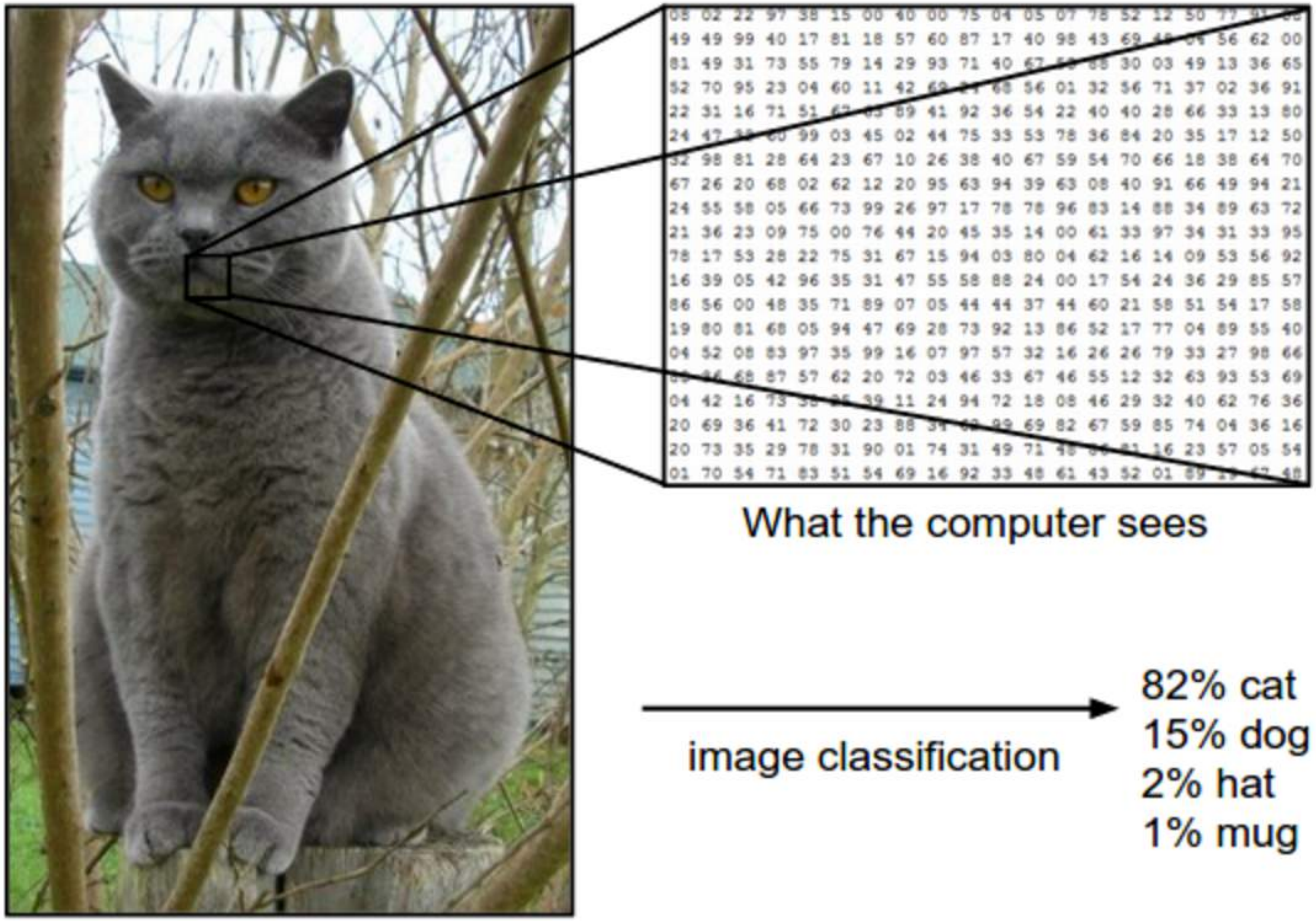
Crop diseases are a major threat to food security, but their rapid identification remains difficult in many parts of the world due to the lack of the necessary infrastructure. The combination of increasing global smartphone penetration and recent advances in computer vision made possible by deep learning has paved the way for smartphone-assisted disease diagnosis. Using a public dataset of 54,306 images of diseased and healthy plant leaves collected under controlled conditions, we train a deep convolutional neural network to identify 14 crop species and 26 diseases (or absence thereof). The trained model achieves an accuracy of 99.35% on a held-out test set, demonstrating the feasibility of this approach. Overall, the approach of training deep learning models on increasingly large and publicly available image datasets presents a clear path towards smartphone-assisted crop disease diagnosis on a massive global scale.

What is deep learning



What is deep learning

- The more data, the better
- Training is computationally very expensive - multiple hours or days on GPU clusters
- Once model is trained, it runs in ~1-2 seconds on CPU, no internet connection required.

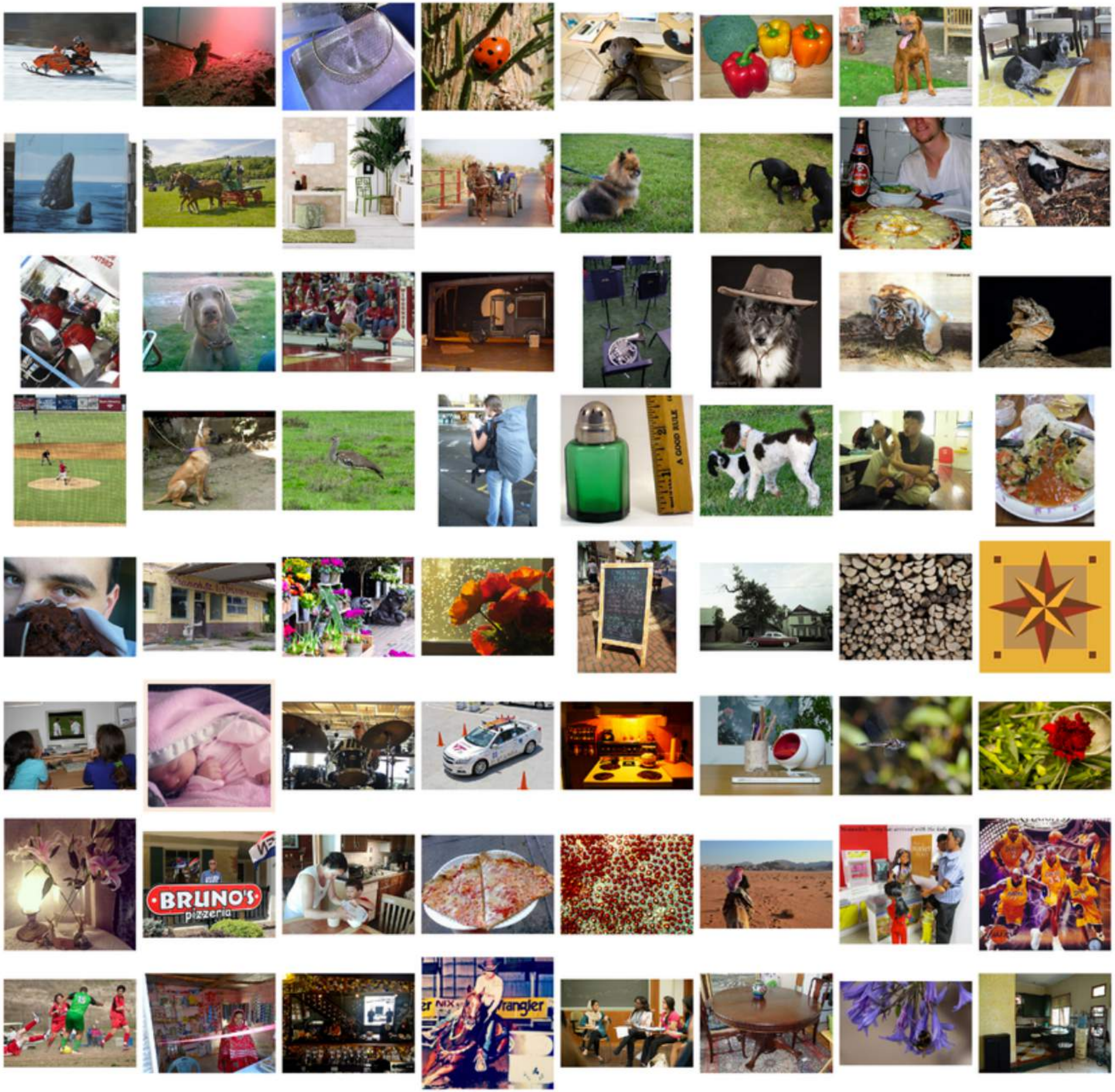


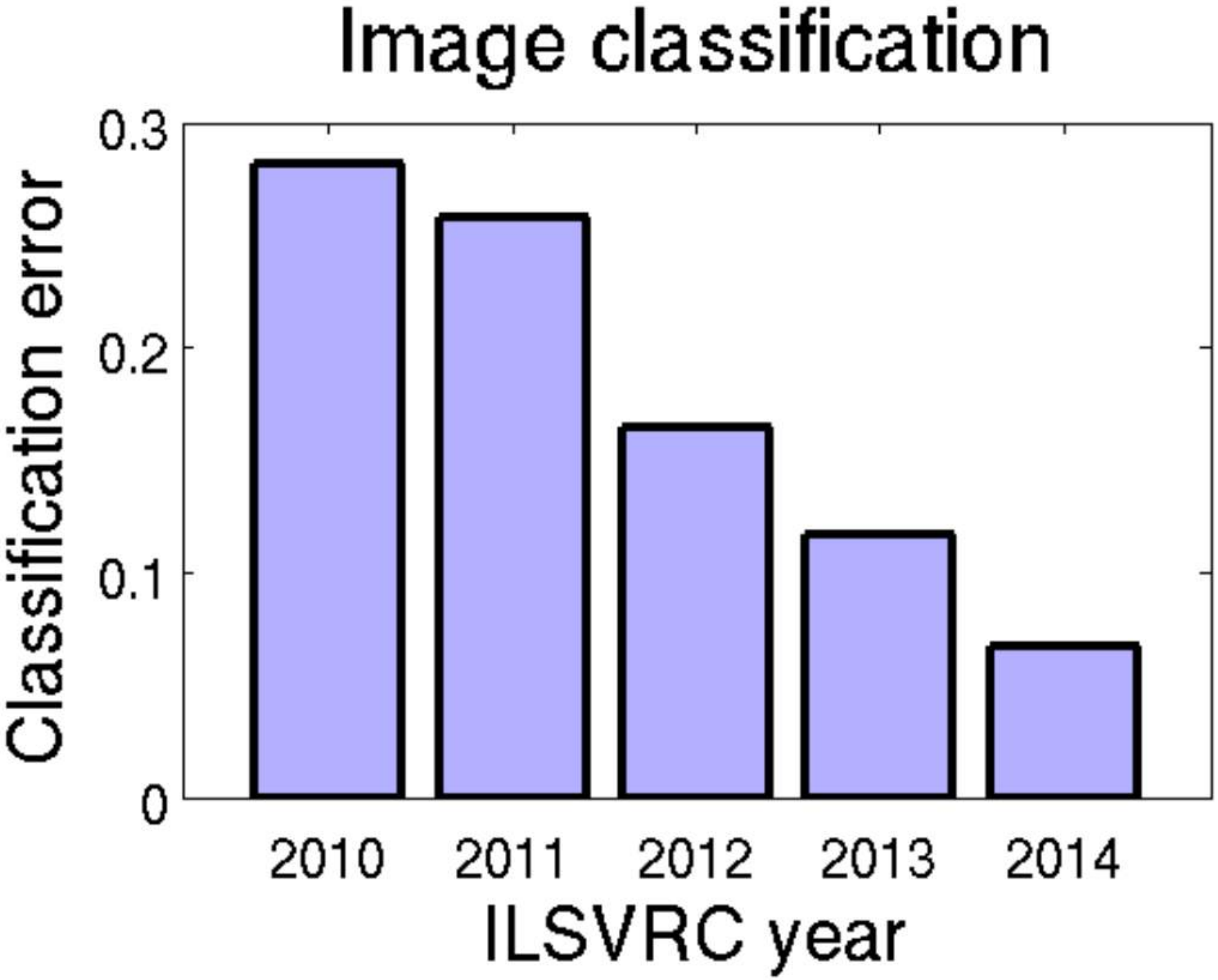
```
08 02 22 97 38 15 00 40 00 75 04 05 07 78 52 12 50 77 91 [illegible]
49 49 99 40 17 81 18 57 60 87 17 40 98 43 69 48 04 56 62 00
81 49 31 73 55 79 14 29 93 71 40 67 [illegible] 88 30 03 49 13 36 65
52 70 95 23 04 60 11 42 69 [illegible] 68 56 01 32 56 71 37 02 36 91
22 31 16 71 51 67 63 89 41 92 36 54 22 40 40 28 66 33 13 80
24 47 [illegible] 60 99 03 45 02 44 75 33 53 78 36 84 20 35 17 12 50
32 98 81 28 64 23 67 10 26 38 40 67 59 54 70 66 18 38 64 70
67 26 20 68 02 62 12 20 95 63 94 39 63 08 40 91 66 49 94 21
24 55 58 05 66 73 99 26 97 17 78 78 96 83 14 88 34 89 63 72
21 36 23 09 75 00 76 44 20 45 35 14 00 61 33 97 34 31 33 95
78 17 53 28 22 75 31 67 15 94 03 80 04 62 16 14 09 53 56 92
16 39 05 42 96 35 31 47 55 58 88 24 00 17 54 24 36 29 85 57
86 56 00 48 35 71 89 07 05 44 44 37 44 60 21 58 51 54 17 58
19 80 81 68 05 94 47 69 28 73 92 13 86 52 17 77 04 89 55 40
04 52 08 83 97 35 99 16 07 97 57 32 16 26 26 79 33 27 98 66
[illegible] 46 68 87 57 62 20 72 03 46 33 67 46 55 12 32 63 93 53 69
04 42 16 73 38 25 39 11 24 94 72 18 08 46 29 32 40 62 76 36
20 69 36 41 72 30 23 88 34 [illegible] 89 69 82 67 59 85 74 04 36 16
20 73 35 29 78 31 90 01 74 31 49 71 48 [illegible] 81 16 23 57 05 54
01 70 54 71 83 51 54 69 16 92 33 48 61 43 52 01 89 [illegible] 67 48
```

What the computer sees

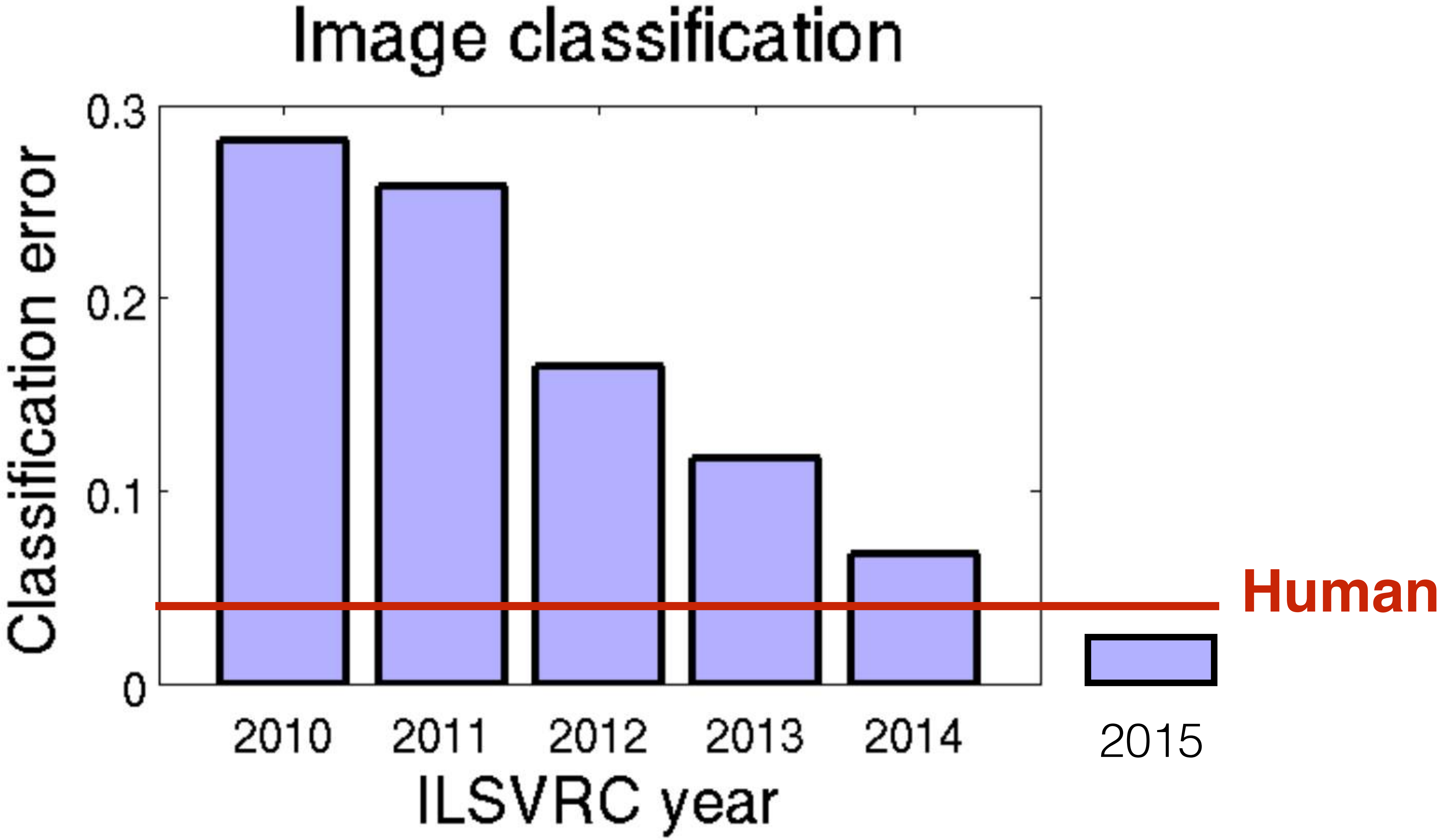
image classification

82% cat
15% dog
2% hat
1% mug


A GOOD RULE
BRUNO'S
pizzeria
Wrangler



https://arxiv.org/pdf/1409.0575v3



https://arxiv.org/pdf/1409.0575v3

Singh et al. 2016

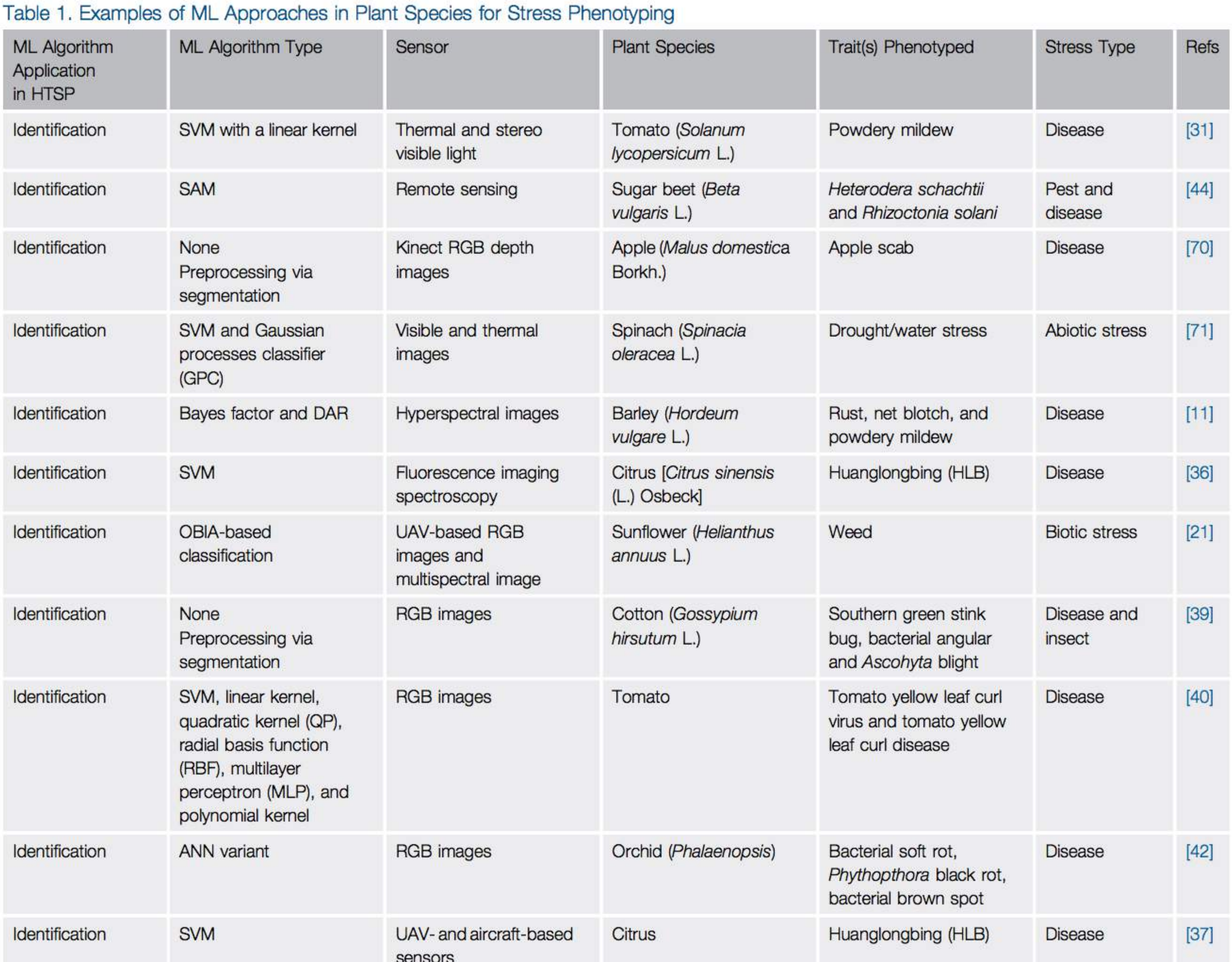
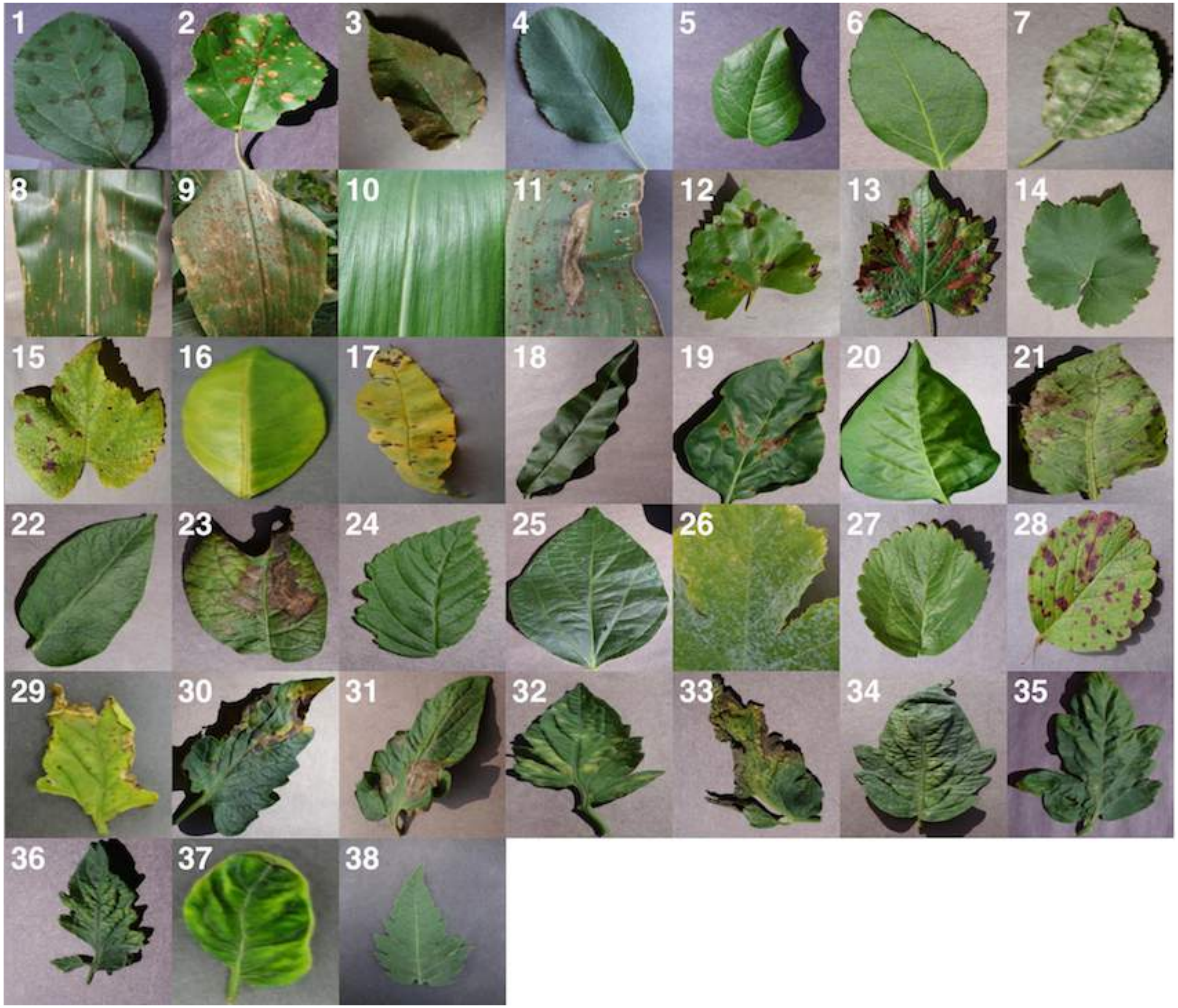


Table 1. Examples of ML Approaches in Plant Species for Stress Phenotyping

ML Algorithm Application in HTSP	ML Algorithm Type	Sensor	Plant Species	Trait(s) Phenotyped	Stress Type	Refs
Identification	SVM with a linear kernel	Thermal and stereo visible light	Tomato (*Solanum lycopersicum* L.)	Powdery mildew	Disease	[31]
Identification	SAM	Remote sensing	Sugar beet (*Beta vulgaris* L.)	*Heterodera schachtii* and *Rhizoctonia solani*	Pest and disease	[44]
Identification	None Preprocessing via segmentation	Kinect RGB depth images	Apple (*Malus domestica* Borkh.)	Apple scab	Disease	[70]
Identification	SVM and Gaussian processes classifier (GPC)	Visible and thermal images	Spinach (*Spinacia oleracea* L.)	Drought/water stress	Abiotic stress	[71]
Identification	Bayes factor and DAR	Hyperspectral images	Barley (*Hordeum vulgare* L.)	Rust, net blotch, and powdery mildew	Disease	[11]
Identification	SVM	Fluorescence imaging spectroscopy	Citrus [*Citrus sinensis* (L.) Osbeck]	Huanglongbing (HLB)	Disease	[36]
Identification	OBIA-based classification	UAV-based RGB images and multispectral image	Sunflower (*Helianthus annuus* L.)	Weed	Biotic stress	[21]
Identification	None Preprocessing via segmentation	RGB images	Cotton (*Gossypium hirsutum* L.)	Southern green stink bug, bacterial angular and *Ascohyta* blight	Disease and insect	[39]
Identification	SVM, linear kernel, quadratic kernel (QP), radial basis function (RBF), multilayer perceptron (MLP), and polynomial kernel	RGB images	Tomato	Tomato yellow leaf curl virus and tomato yellow leaf curl disease	Disease	[40]
Identification	ANN variant	RGB images	Orchid (*Phalaenopsis*)	Bacterial soft rot, *Phythopthora* black rot, bacterial brown spot	Disease	[42]
Identification	SVM	UAV- and aircraft-based sensors	Citrus	Huanglongbing (HLB)	Disease	[37]



Mohanty et al. 2016

54,306 images

14 crops

26 diseases

38 classes

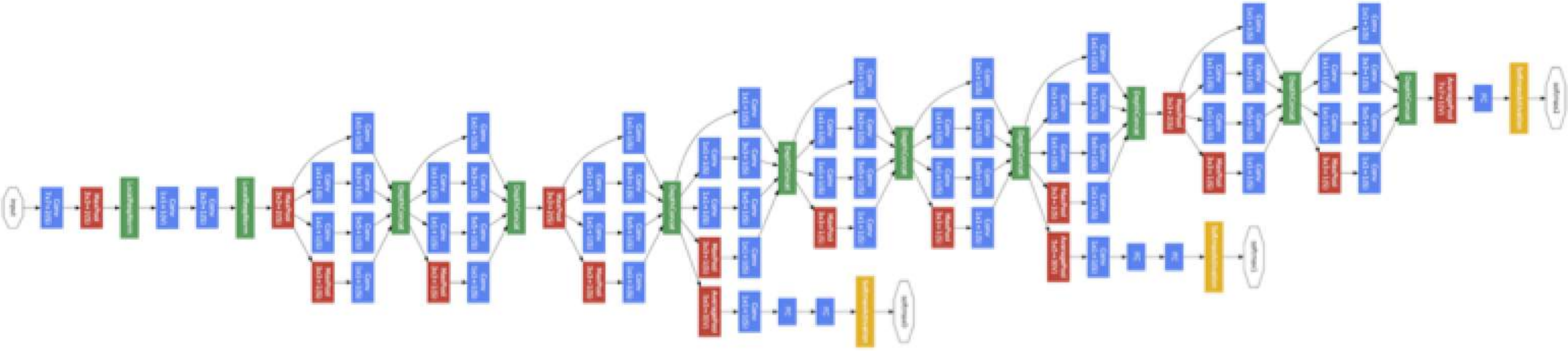
Now:

70,000+ images

17 crops

33 diseases

47 classes

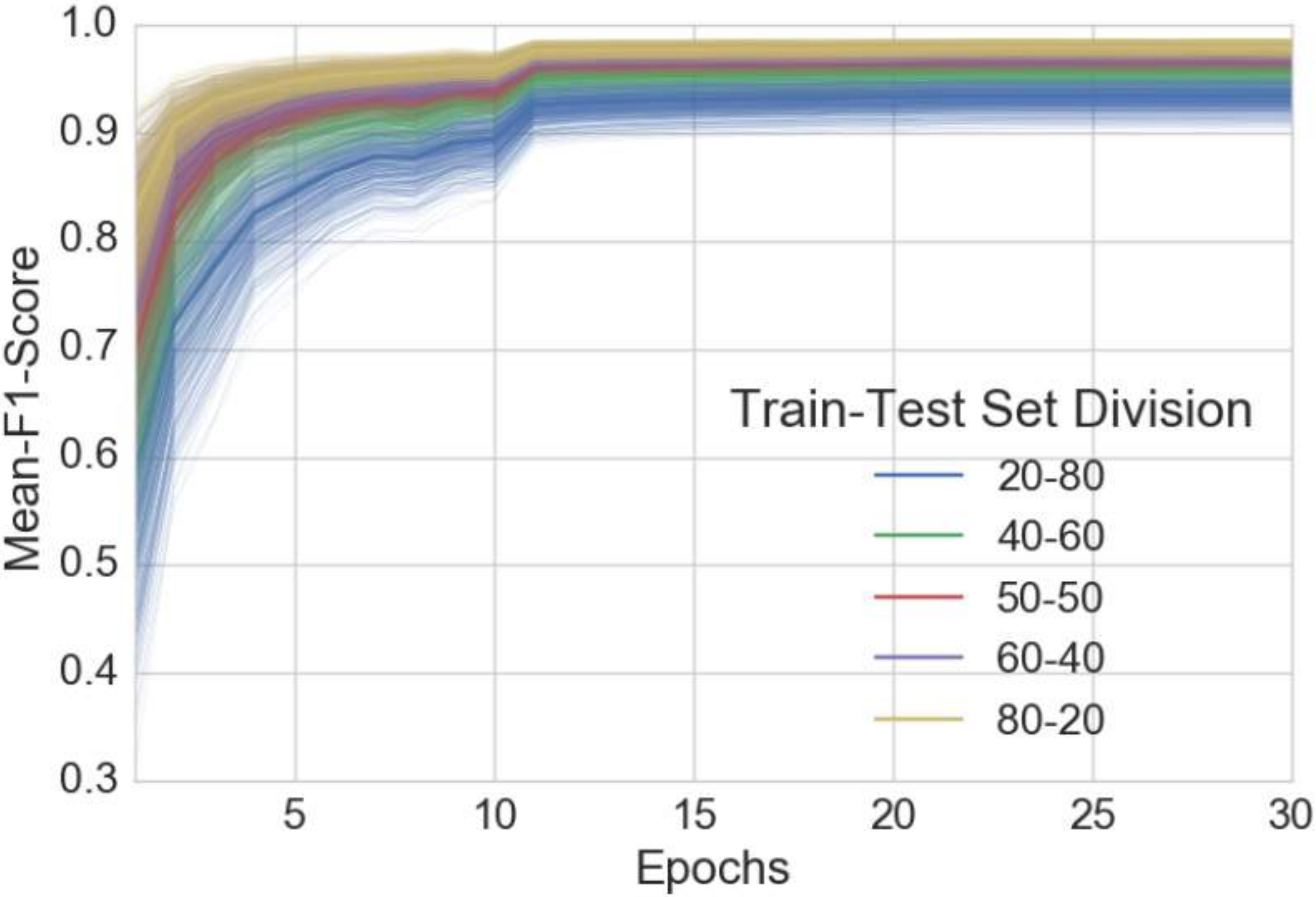


Convolution

Pooling

Softmax

Other



Mohanty et al. 2016

Traditional ML vs DL

Traditional ML	**DL**
one crop species	14 crop specices
mostly 2, at most 4 classes	38 classes
feature engineering	end-to-end ML
hard to extend	“just add data”

PlantVillage: Machine Learning for Disease Recognition

Image Recognition
(Machine Learning)

Diagnosis
Treatment Suggestions

crowdsourcing

doi:10.1038/nature21056

Dermatologist-level classification of skin cancer with deep neural networks

Andre Esteva[1]*, Brett Kuprel[1]*, Roberto A. Novoa[2,3], Justin Ko[2], Susan M. Swetter[2,4], Helen M. Blau[5] & Sebastian Thrun[6]

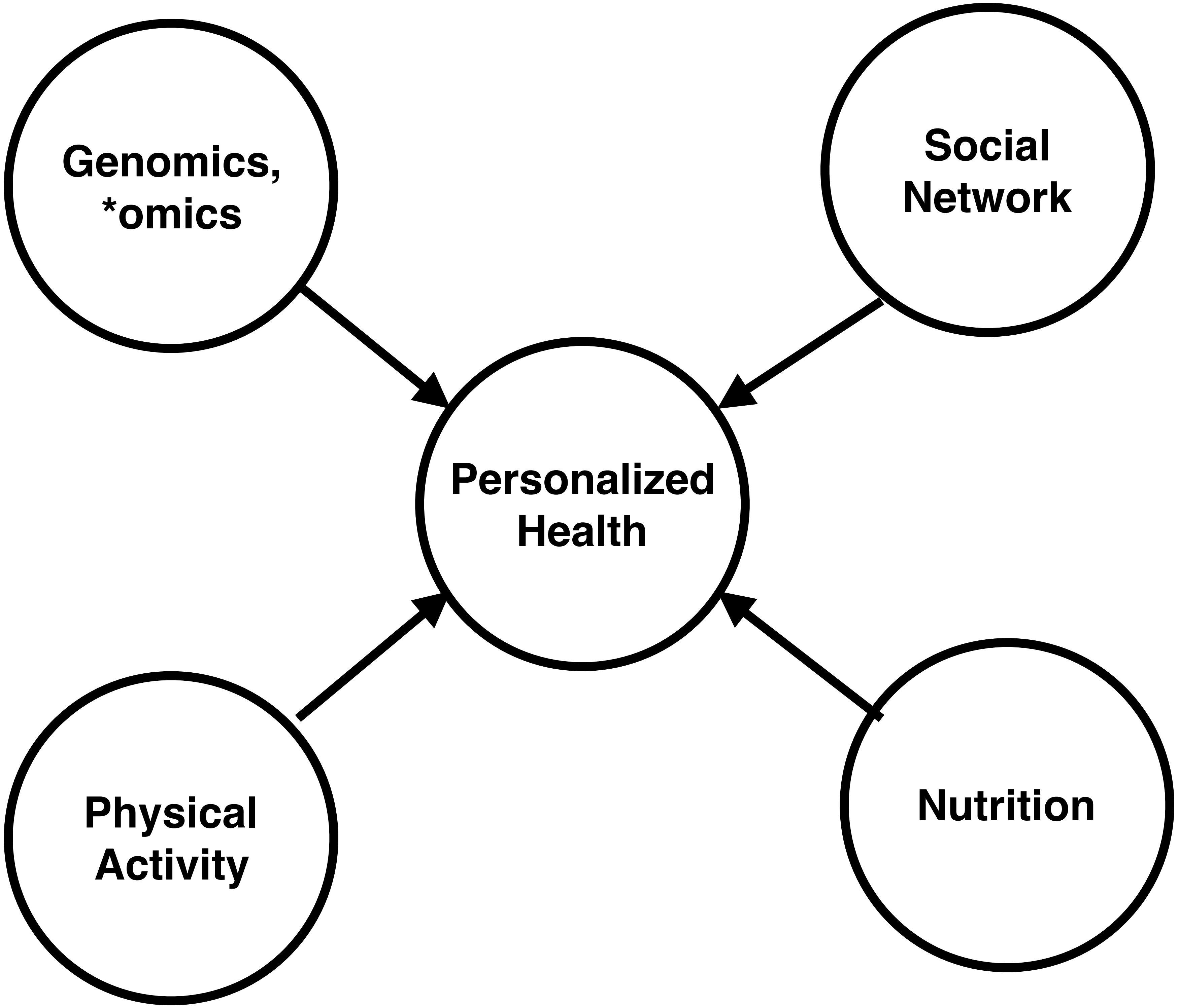
Skin cancer, the most common human malignancy[1–3], is primarily diagnosed visually, beginning with an initial clinical screening and followed potentially by dermoscopic analysis, a biopsy and histopathological examination. Automated classification of skin lesions using images is a challenging task owing to the fine-grained variability in the appearance of skin lesions. Deep convolutional neural networks (CNNs)[4,5] show potential for general and highly variable tasks across many fine-grained object categories[6–11]. Here we demonstrate classification of skin lesions using a single CNN, trained end-to-end from images directly, using only pixels and disease labels as inputs. We train a CNN using a dataset of 129,450 clinical images—two orders of magnitude larger than previous datasets[12]—consisting of 2,032 different diseases. We test its performance against 21 board-certified dermatologists on biopsy-proven clinical images with two critical binary classification use cases: keratinocyte carcinomas versus benign seborrheic keratoses; and malignant melanomas versus benign nevi. The first case represents the identification of the most common cancers, the second represents the identification of the deadliest skin cancer. The CNN achieves performance on par with all tested experts across both tasks, demonstrating an artificial intelligence capable of classifying skin cancer with a level of competence comparable to dermatologists. Outfitted with deep neural networks, mobile devices can potentially extend the reach of dermatologists outside of the clinic. It is projected that 6.3 billion smartphone subscriptions will exist by the year 2021 (ref. 13) and can therefore potentially provide low-cost universal access to vital diagnostic care.

There are 5.4 million new cases of skin cancer in the United States[2] every year. One in five Americans will be diagnosed with a cutaneous images (for example, smartphone images) exhibit variability in factors such as zoom, angle and lighting, making classification substantially more challenging[23,24]. We overcome this challenge by using a data-driven approach—1.41 million pre-training and training images make classification robust to photographic variability. Many previous techniques require extensive preprocessing, lesion segmentation and extraction of domain-specific visual features before classification. By contrast, our system requires no hand-crafted features; it is trained end-to-end directly from image labels and raw pixels, with a single network for both photographic and dermoscopic images. The existing body of work uses small datasets of typically less than a thousand images of skin lesions[16,18,19], which, as a result, do not generalize well to new images. We demonstrate generalizable classification with a new dermatologist-labelled dataset of 129,450 clinical images, including 3,374 dermoscopy images.

Deep learning algorithms, powered by advances in computation and very large datasets[25], have recently been shown to exceed human performance in visual tasks such as playing Atari games[26], strategic board games like Go[27] and object recognition[6]. In this paper we outline the development of a CNN that matches the performance of dermatologists at three key diagnostic tasks: melanoma classification, melanoma classification using dermoscopy and carcinoma classification. We restrict the comparisons to image-based classification.

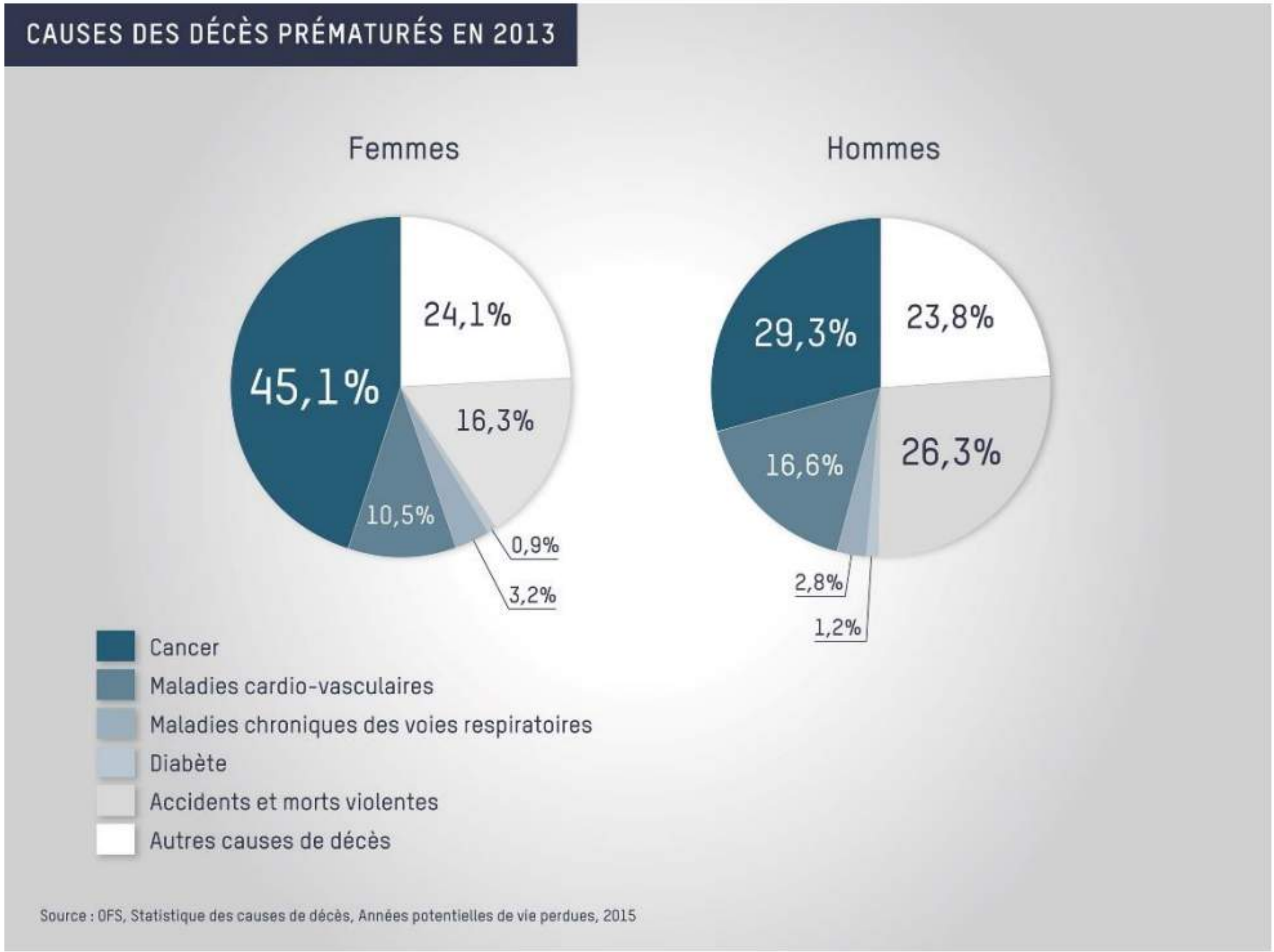
We utilize a GoogleNet Inception v3 CNN architecture[9] that was pre-trained on approximately 1.28 million images (1,000 object categories) from the 2014 ImageNet Large Scale Visual Recognition Challenge[6], and train it on our dataset using transfer learning[28]. Figure 1 shows the working system. The CNN is trained using 757 disease classes. Our dataset is composed of dermatologist-labelled images organized in a

Personalized health



Personalized health

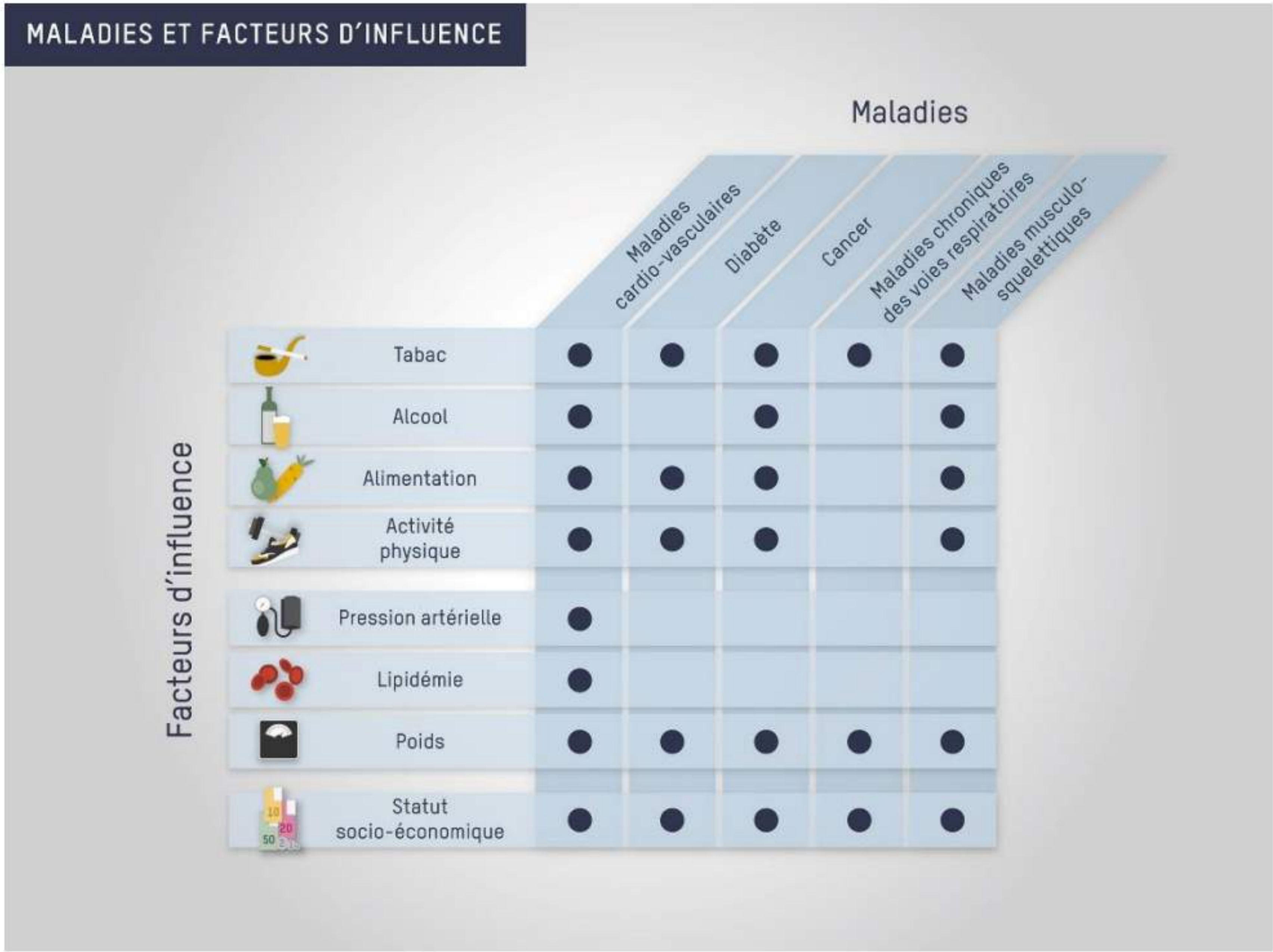
>80% of healthcare costs (65B) in Switzerland are due to NCD (non-communicable diseases).



Source : OFS, Statistique des causes de décès, Années potentielles de vie perdues, 2015

Personalized health

>80% of healthcare costs (65B) in Switzerland are due to NCD (non-communicable diseases).

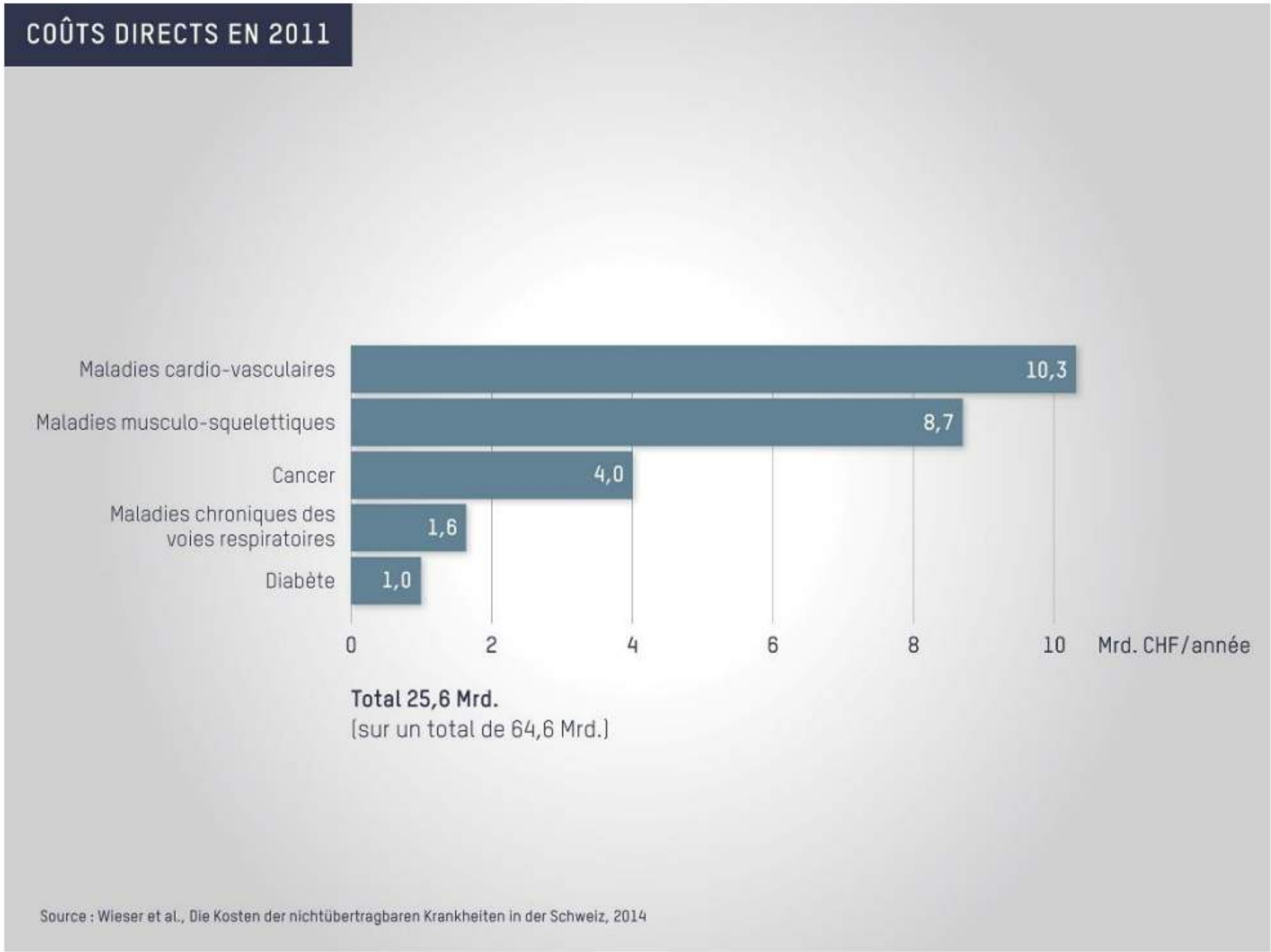


MALADIES ET FACTEURS D'INFLUENCE

Facteurs d'influence	Maladies cardio-vasculaires	Diabète	Cancer	Maladies chroniques des voies respiratoires	Maladies musculo-squelettiques
Tabac	●	●	●	●	●
Alcool	●		●		●
Alimentation	●	●	●		●
Activité physique	●	●	●		●
Pression artérielle	●				
Lipidémie	●				
Poids	●	●	●	●	●
Statut socio-économique	●	●	●	●	●

Personalized health

>80% of healthcare costs (65B) in Switzerland are due to NCD (non-communicable diseases).



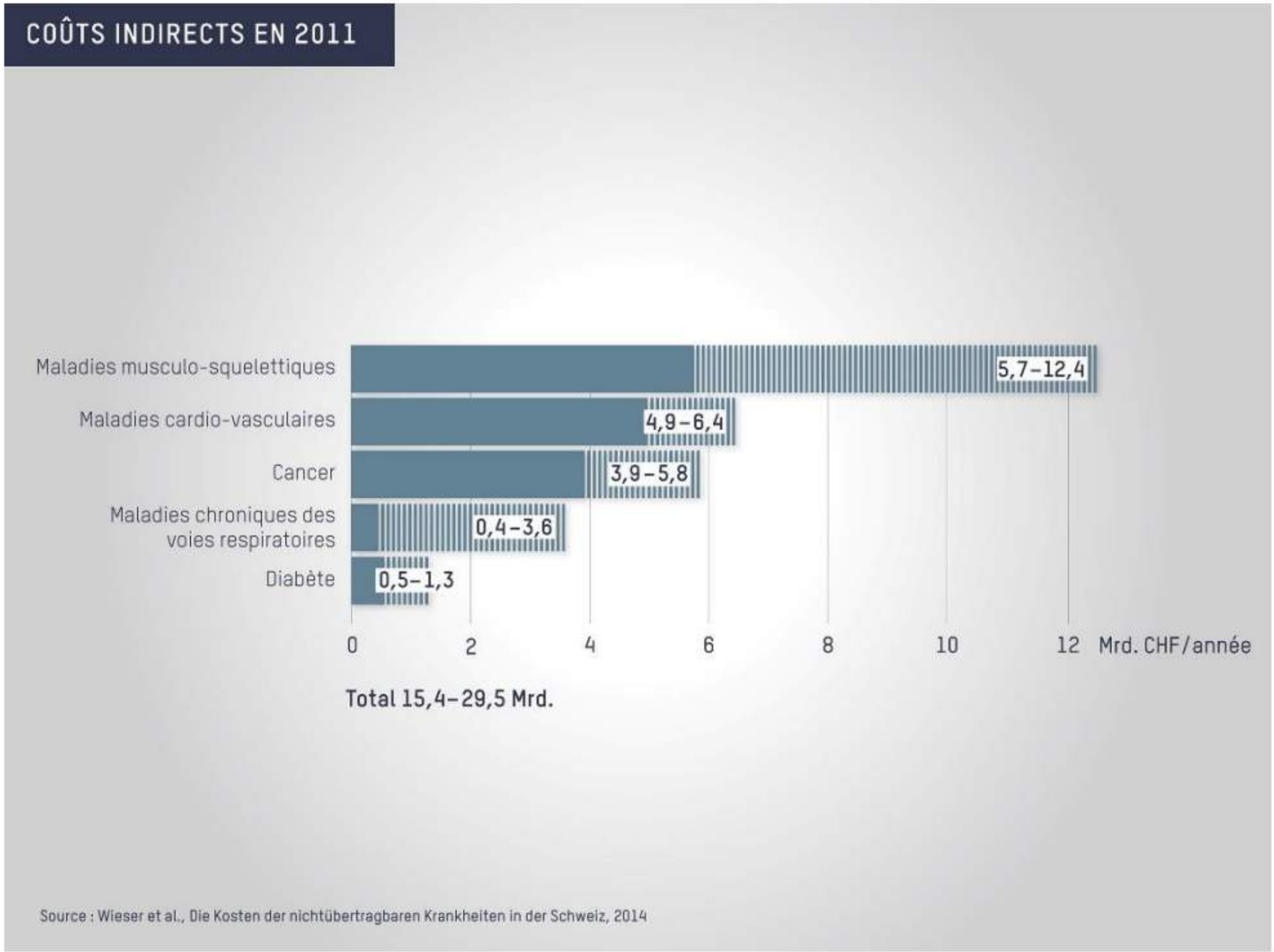
COÛTS DIRECTS EN 2011

Total 25,6 Mrd.
(sur un total de 64,6 Mrd.)

Source : Wieser et al., Die Kosten der nichtübertragbaren Krankheiten in der Schweiz, 2014

Personalized health

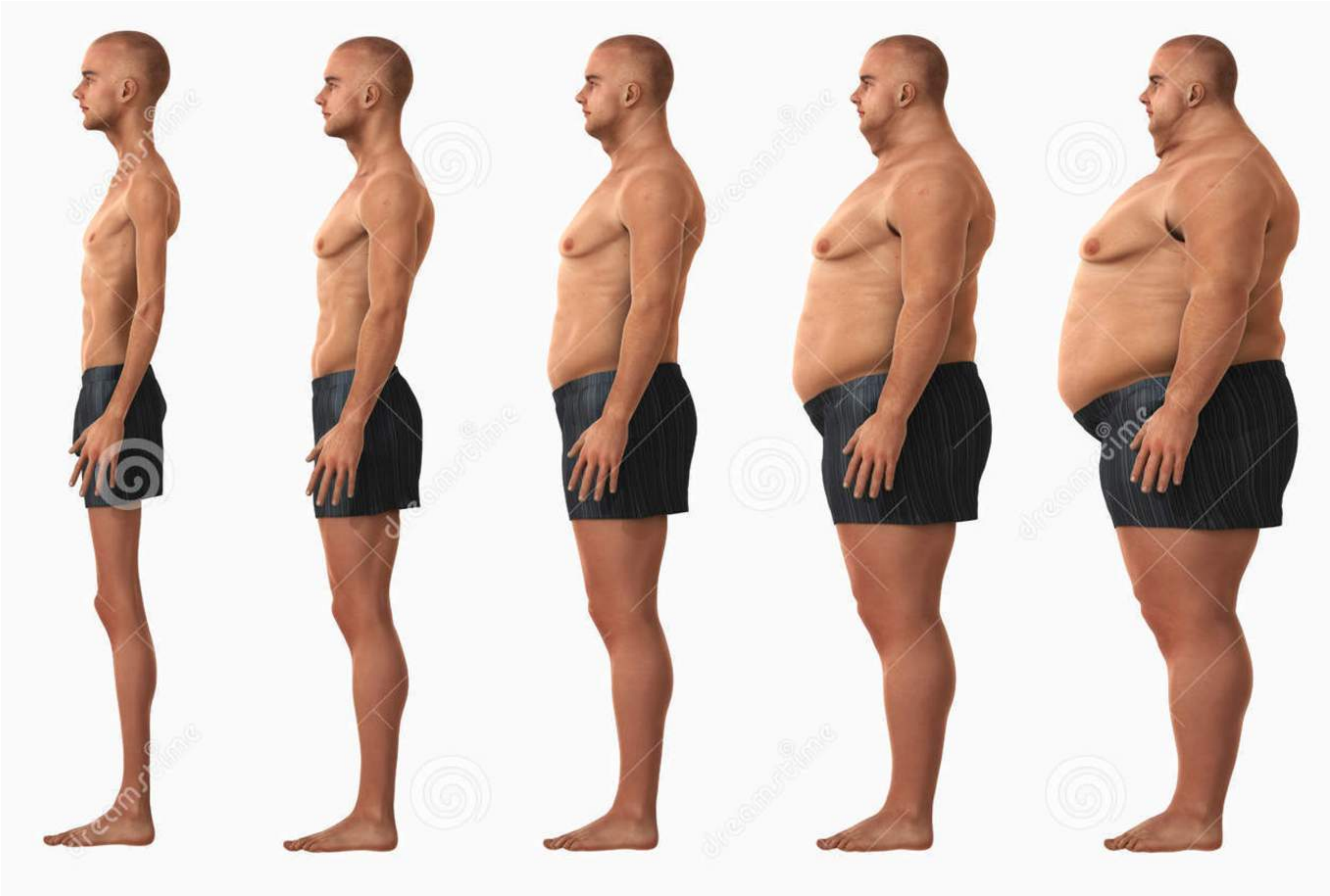
>80% of healthcare costs (65B) in Switzerland are due to NCD (non-communicable diseases).



Source : Wieser et al., Die Kosten der nichtübertragbaren Krankheiten in der Schweiz, 2014

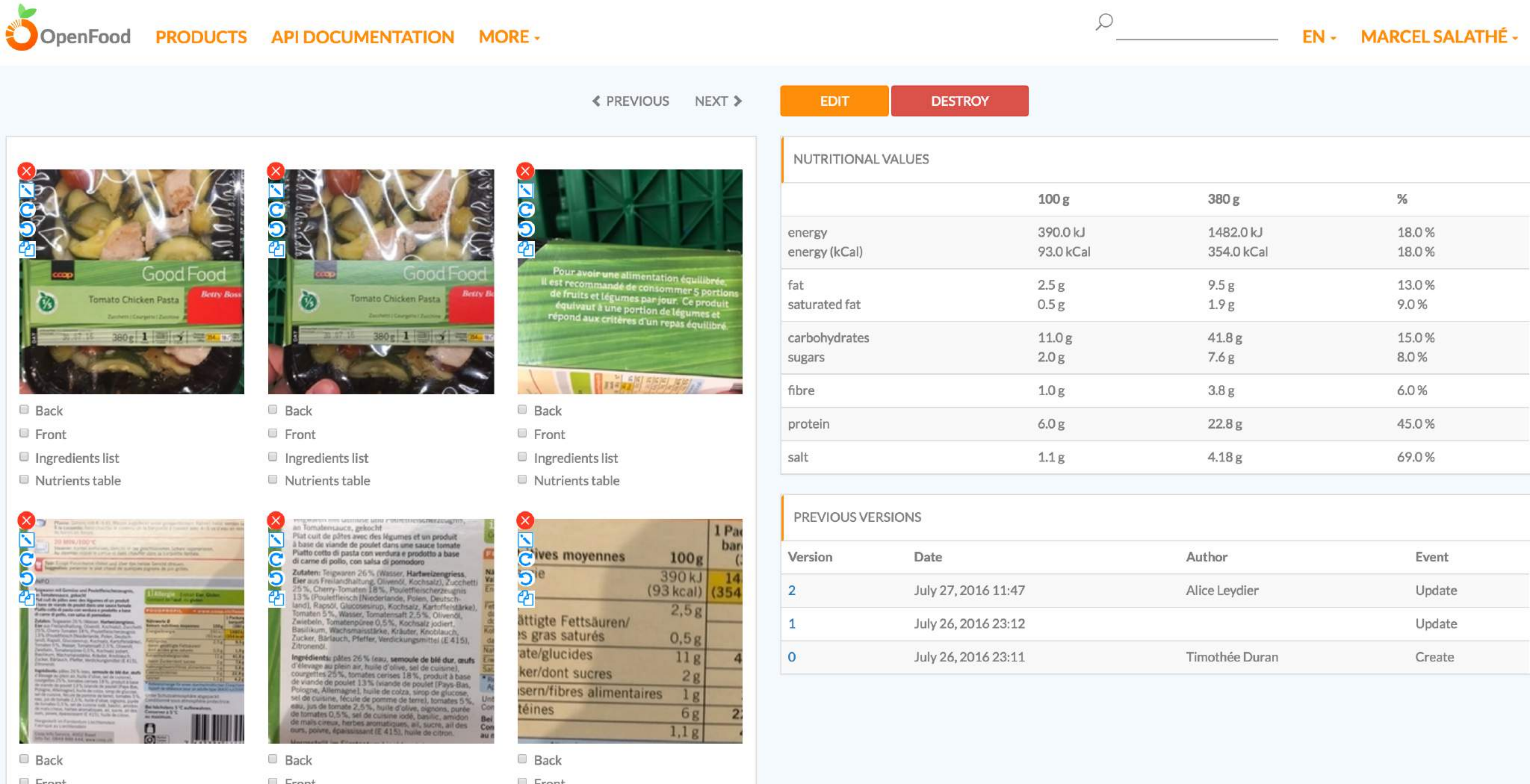
Personalized health

>80% of healthcare costs (65B) in Switzerland are due to NCD (non-communicable diseases).



www.openfood.ch

14,000+ food products in Switzerland via API



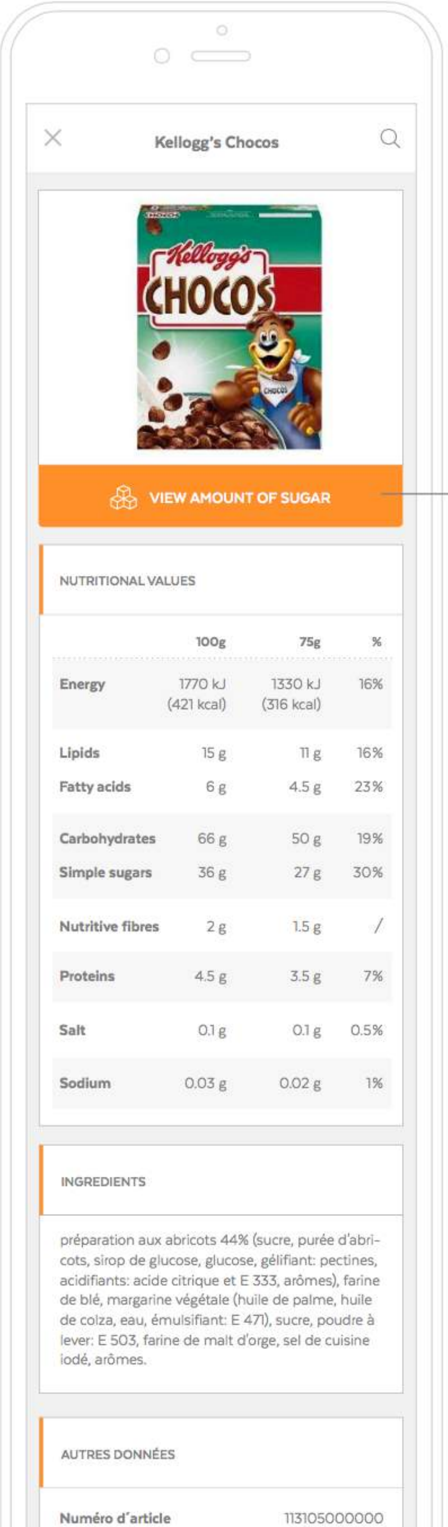
NUTRITIONAL VALUES

	100 g	380 g	%
energy energy (kCal)	390.0 kJ 93.0 kCal	1482.0 kJ 354.0 kCal	18.0 % 18.0 %
fat saturated fat	2.5 g 0.5 g	9.5 g 1.9 g	13.0 % 9.0 %
carbohydrates sugars	11.0 g 2.0 g	41.8 g 7.6 g	15.0 % 8.0 %
fibre	1.0 g	3.8 g	6.0 %
protein	6.0 g	22.8 g	45.0 %
salt	1.1 g	4.18 g	69.0 %

PREVIOUS VERSIONS

Version	Date	Author	Event
2	July 27, 2016 11:47	Alice Leydier	Update
1	July 26, 2016 23:12		Update
0	July 26, 2016 23:11	Timothée Duran	Create

openfood app



Kellogg's Chocos

NUTRITIONAL VALUES

	100g	75g	%
Energy	1770 kJ (421 kcal)	1330 kJ (316 kcal)	16%
Lipids	15 g	11 g	16%
Fatty acids	6 g	4.5 g	23%
Carbohydrates	66 g	50 g	19%
Simple sugars	36 g	27 g	30%
Nutritive fibres	2 g	1.5 g	/
Proteins	4.5 g	3.5 g	7%
Salt	0.1 g	0.1 g	0.5%
Sodium	0.03 g	0.02 g	1%

INGREDIENTS

préparation aux abricots 44% (sucre, purée d'abricots, sirop de glucose, glucose, gélifiant: pectines, acidifiants: acide citrique et E 333, arômes), farine de blé, margarine végétale (huile de palme, huile de colza, eau, émulsifiant: E 471), sucre, poudre à lever: E 503, farine de malt d'orge, sel de cuisine iodé, arômes.

AUTRES DONNÉES

Numéro d'article	11310500000

Sugar in Kellogg's Chocos

26,6 g | 6 sugar cubes

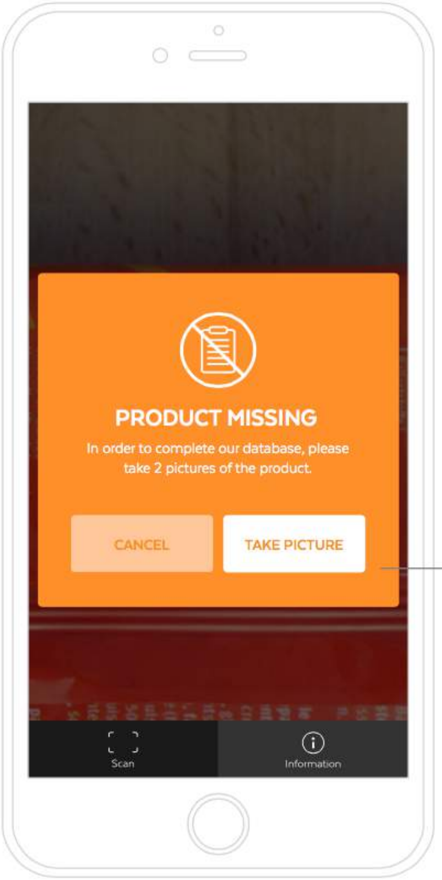
Sugar in Kellogg's Chocos

8 g | 2 sugar cubes

Sugar in Kellogg's Chocos

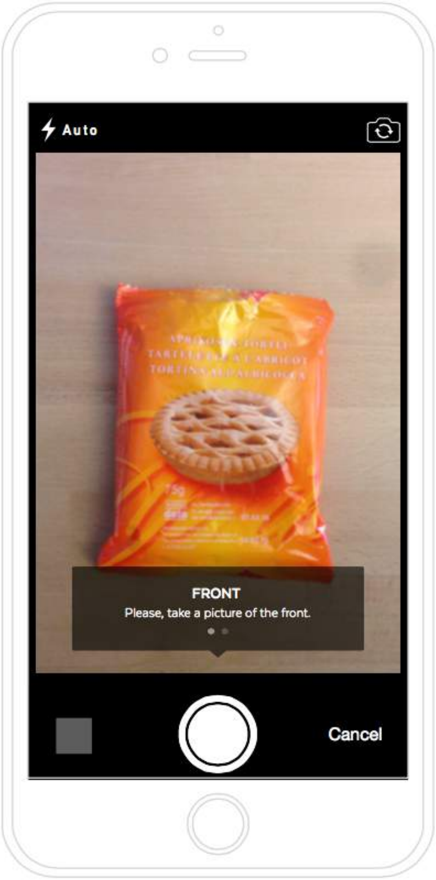
90 g 20 sugar cubes

openfood app



PRODUCT MISSING

In order to complete our database, please take 2 pictures of the product.

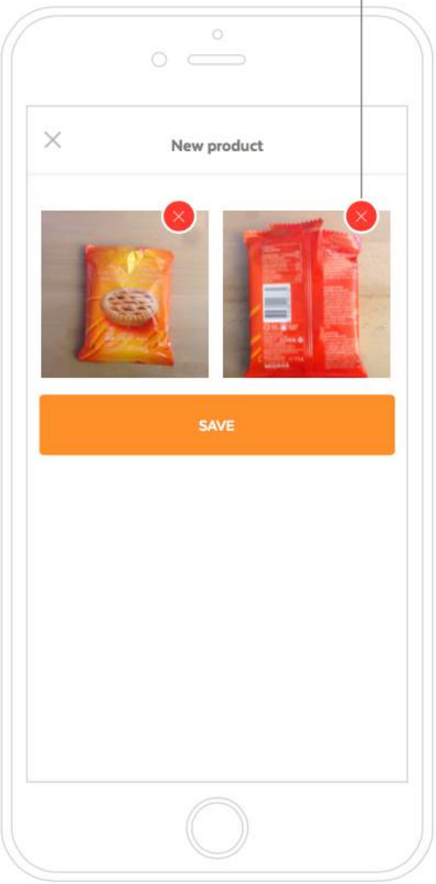


FRONT

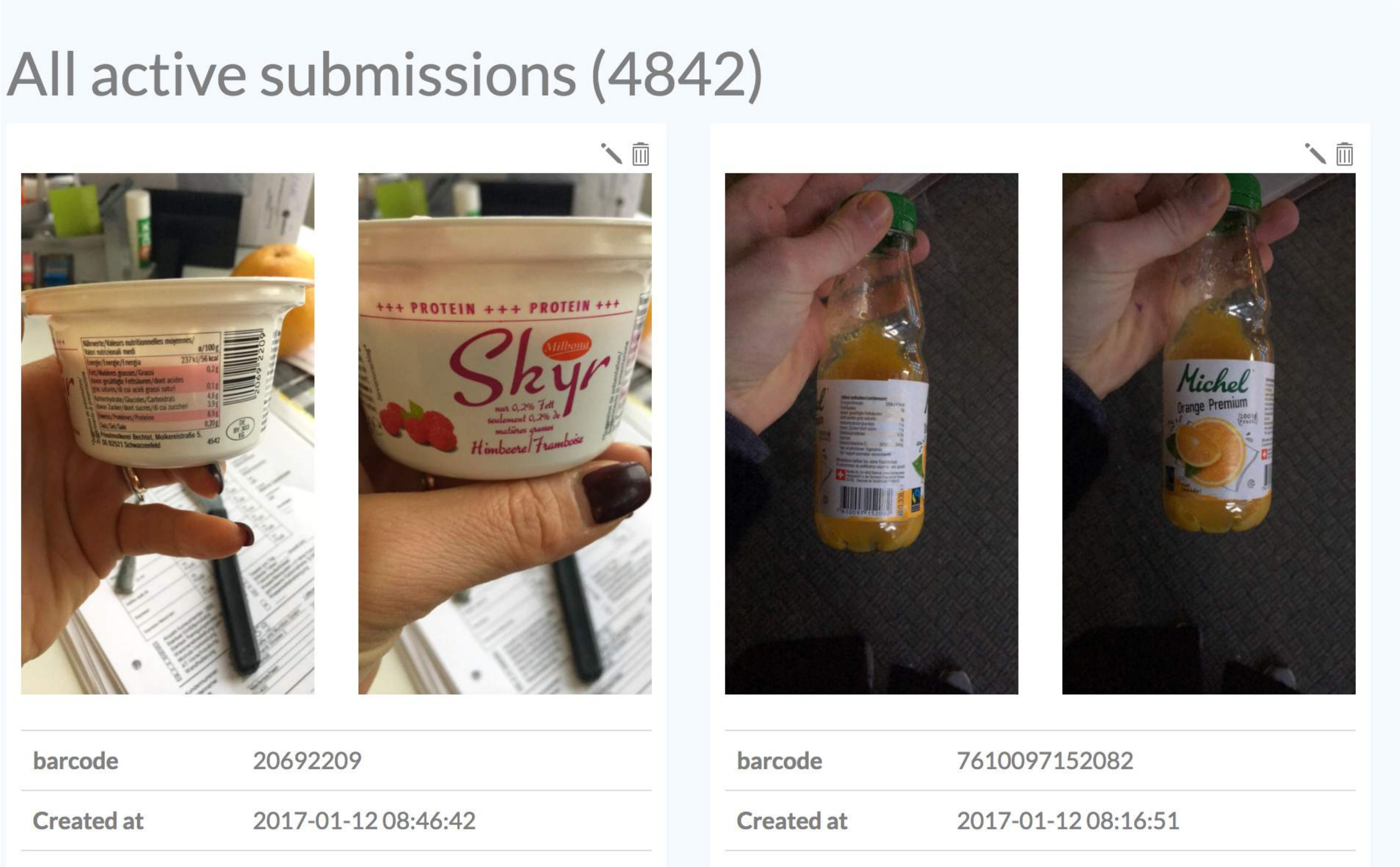
Please, take a picture of the front.

BACK

Thanks, please take a picture of the back



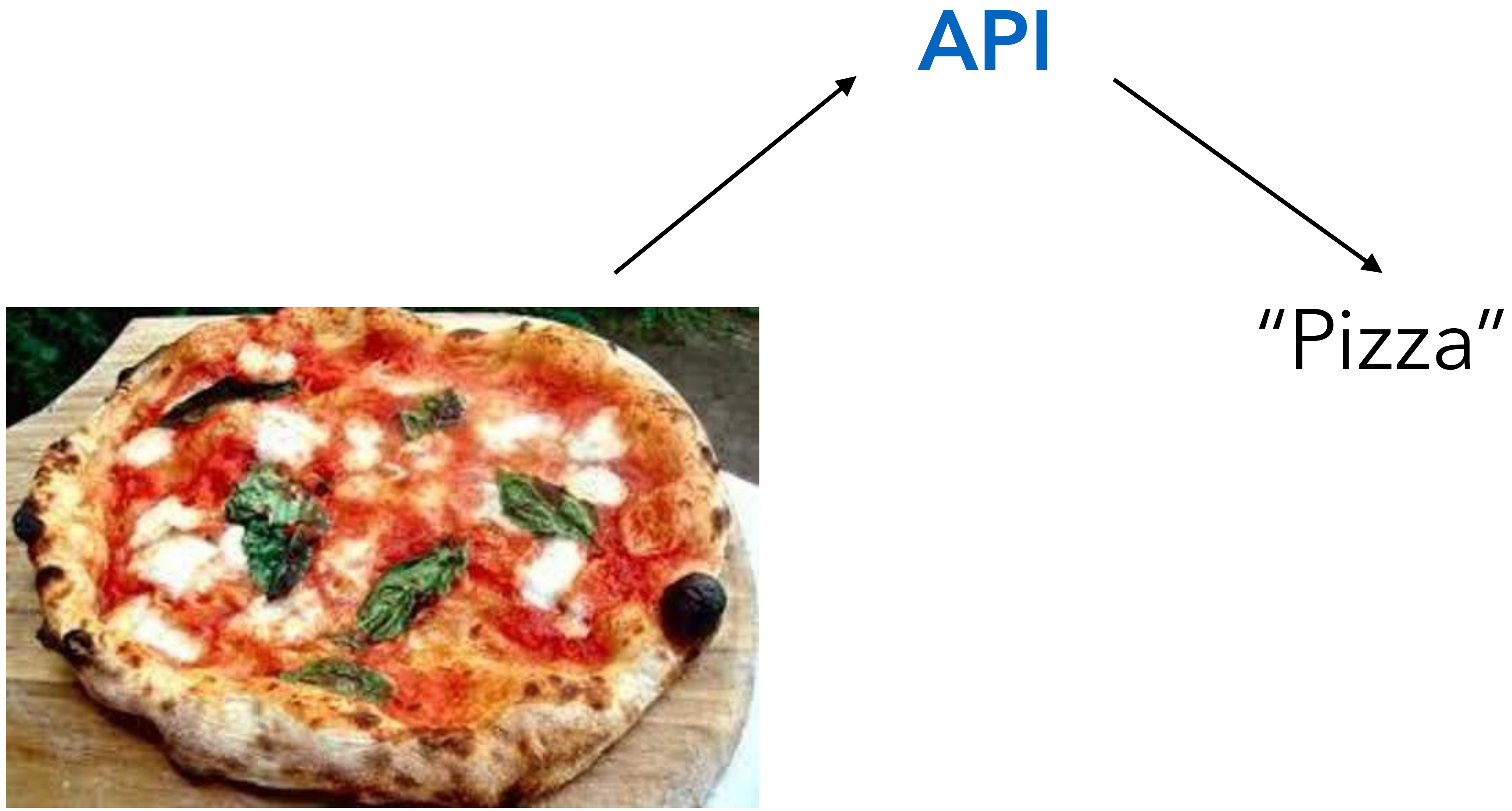
Crowdsourcing product collection



All active submissions (4842)

barcode	20692209
Created at	2017-01-12 08:46:42

barcode	7610097152082
Created at	2017-01-12 08:16:51

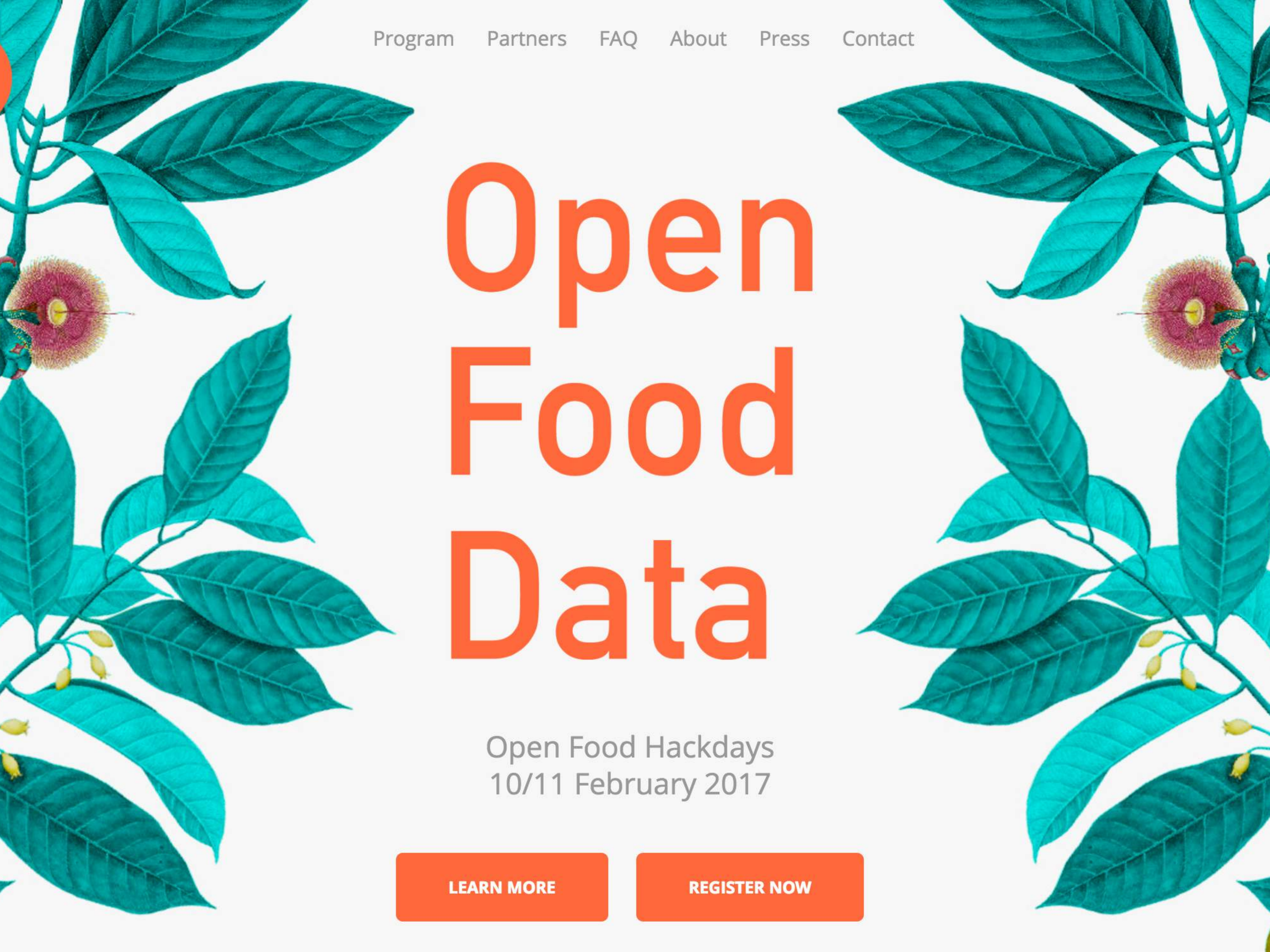


API

"Pizza"

API

"Pizza"
est. weight: 340g
est. calories: 890
est. carbs: 180g
est. fat: 20g
est. protein: 24g

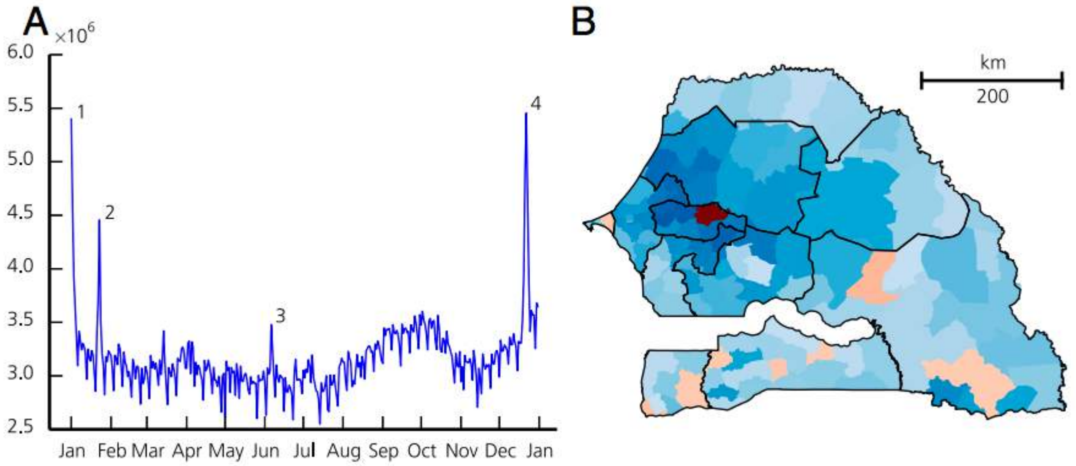


Open Food Data

Open Food Hackdays
10/11 February 2017

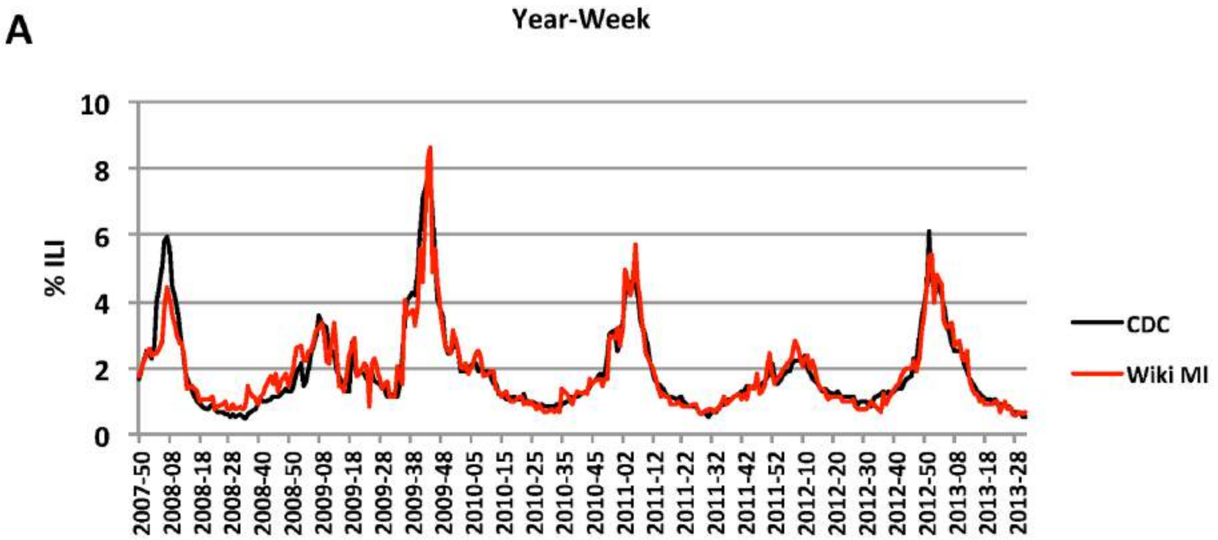
Mobile phones - Track disease outbreaks

Finger et al 2016



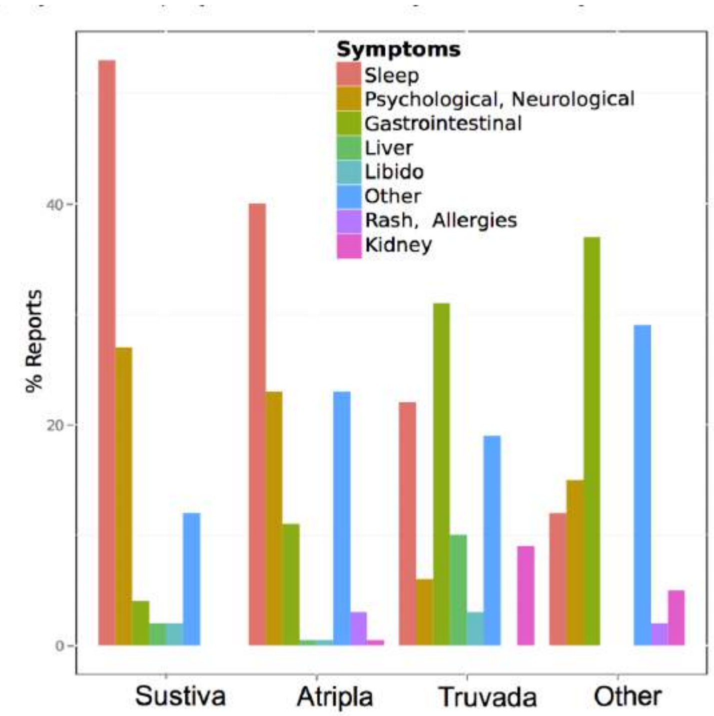
Wikipedia - Influenza forecasting

McIver & Brownstein 2014



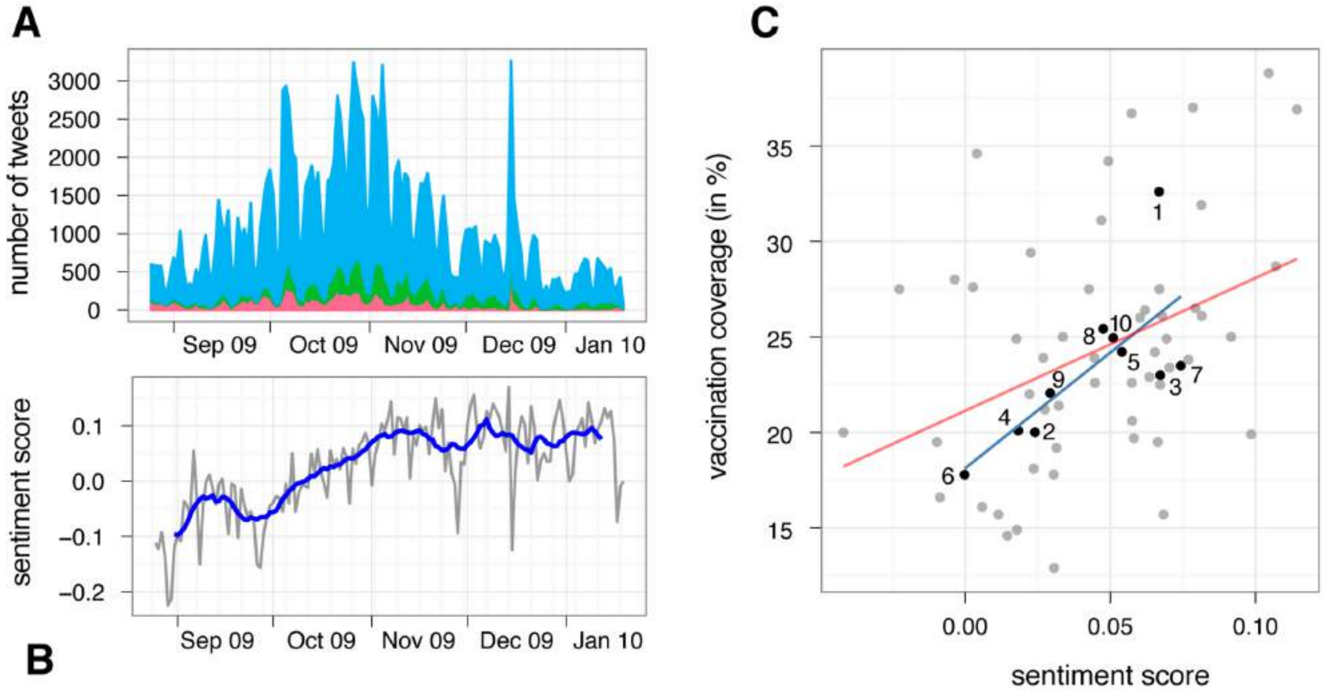
Twitter - Pharmacovigilance

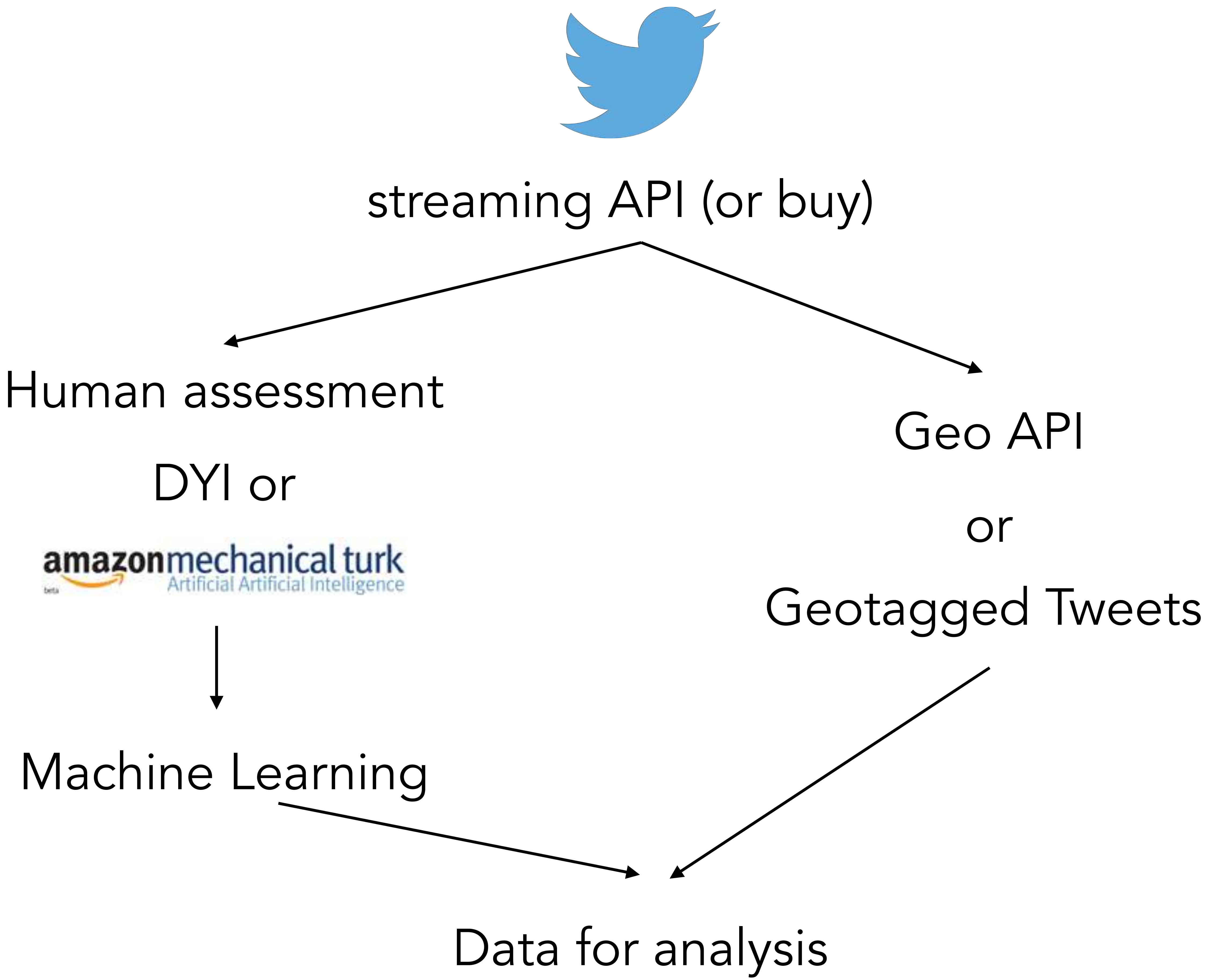
Adrover et al 2015



Twitter - Vaccine uptake

Salathé & Khandelwal, 2011





streaming API (or buy)

Human assessment

DYI or

amazonmechanical turk
beta
Artificial Artificial Intelligence

Machine Learning

Geo API

or

Geotagged Tweets

Data for analysis

Help us make crowdbreaks even better!

With a single click, you can help us improve our disease detection algorithm. Simply keep answering the questions below. [How does this work?]

Is this tweet about Influenza?

> Having the flu while traveling is probably one of the worst things to deal with on a rainy day...

Message content from Twitter

Help us make crowdbreaks even better!

With a single click, you can help us improve our disease detection algorithm. Simply keep answering the questions below. [How does this work?]

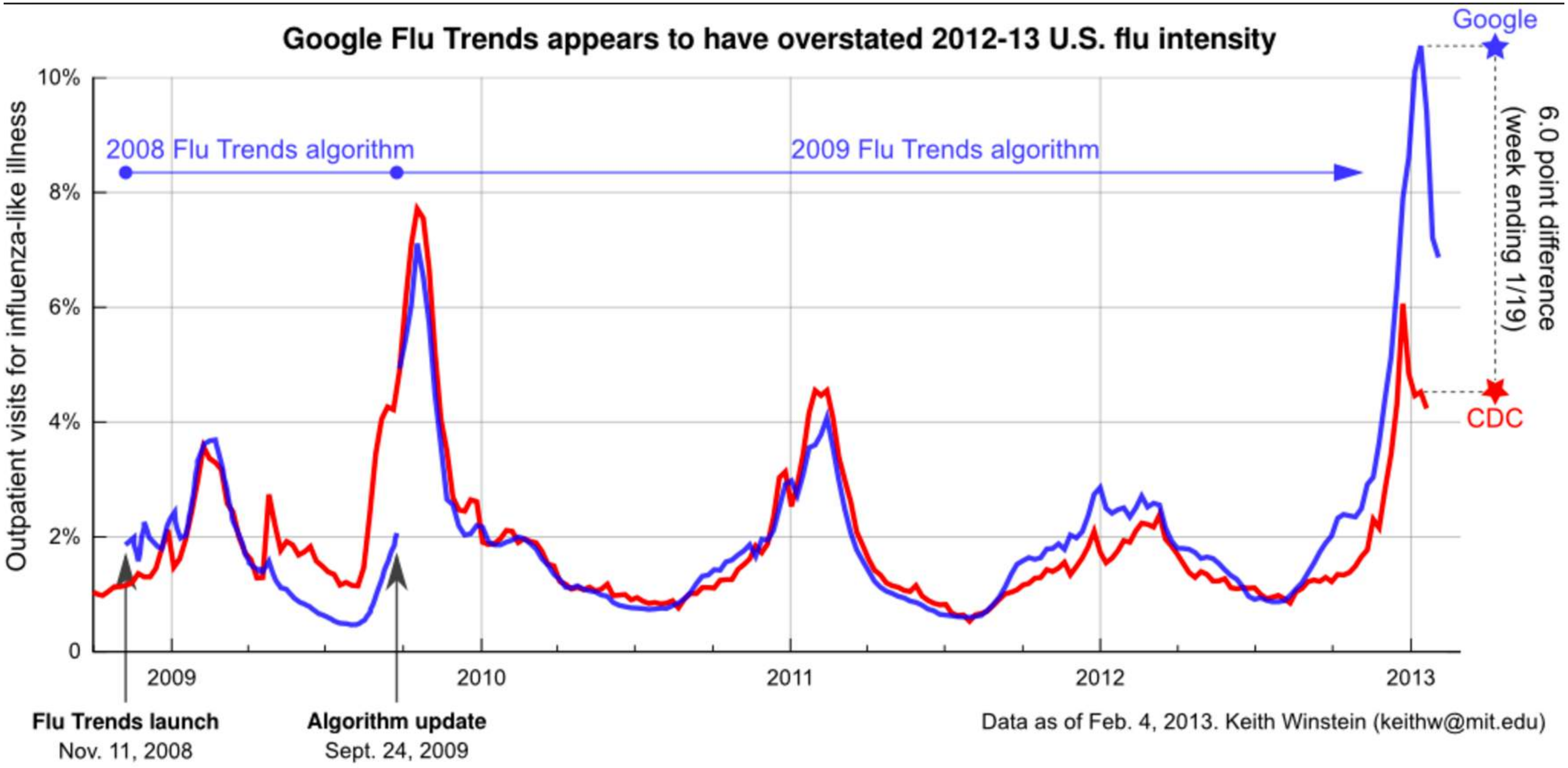
Is this tweet about Influenza?

> "My body is yearning for sleep so much that I got a crazy headache

Message content from Twitter

The crowdbreaks process

1. Define a Twitter query (or other data in the future)
2. Define a question tree
3. Go!
4. Model, validate, model, validate, model, etc.



Google Flu Trends appears to have overstated 2012-13 U.S. flu intensity

Data as of Feb. 4, 2013. Keith Winstein (keithw@mit.edu)

Sources: http://www.google.org/flutrends/us, CDC ILInet data from http://gis.cdc.gov/grasp/fluview/fluportaldashboard.html,
Cook et al. (2011) Assessing Google Flu Trends Performance in the United States during the 2009 Influenza Virus A (H1N1) Pandemic.

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Online | Mobile | Global

From Social Networks to Machine Learning: Epidemiology is Going Digital

Marcel Salathé, Digital Epidemiology Lab, EPFL

@marcelsalathe