

Online | Mobile | Global

Bringing **Digital Epidemiology** to the Next Level

Marcel Salathé, Digital Epidemiology Lab, EPFL

@marcelsalathe

A brief history of the term

First mention on December 6, 2009, by John M. Smith in a blog post - but completely unrelated to health (about software)

A brief history of the term

First mention on December 6, 2009, by John M. Smith in a blog post - but completely unrelated to health (about software)

May 7, 2010: indeed.com job offer in a “EPSRC project 'Digital Epidemiology’" - RFID

A brief history of the term

First mention on December 6, 2009, by John M. Smith in a blog post - but completely unrelated to health (about software)

May 7, 2010: indeed.com job offer in a “EPSRC project 'Digital Epidemiology’" - RFID

November 2010: Economist mentions Digital Epidemiology in article on Nathan Wolfe.

A brief history of the term

First mention on December 6, 2009, by John M. Smith in a blog post - but completely unrelated to health (about software)

May 7, 2010: indeed.com job offer in a “EPSRC project 'Digital Epidemiology’" - RFID

November 2010: Economist mentions Digital Epidemiology in article on Nathan Wolfe.

2011: a few references to Economist article, and a Time article that references John Brownstein a digital epidemiologist

A brief history of the term

First mention on December 6, 2009, by John M. Smith in a blog post - but completely unrelated to health (about software)

May 7, 2010: indeed.com job offer in a “EPSRC project 'Digital Epidemiology’" - RFID

November 2010: Economist mentions Digital Epidemiology in article on Nathan Wolfe.

2011: a few references to Economist article, and a Time article that references John Brownstein a digital epidemiologist

2012: PLOS Computational Biology review article

A brief history of the term

First mention on December 6, 2009, by John M. Smith in a blog post - but completely unrelated to health (about software)

May 7, 2010: indeed.com job offer in a “EPSRC project 'Digital Epidemiology’" - RFID

November 2010: Economist mentions Digital Epidemiology in article on Nathan Wolfe.

2011: a few references to Economist article, and a Time article that references John Brownstein a digital epidemiologist

2012: PLOS Computational Biology review article

2013: NEJM perspective piece

A brief history of the term

2014: Healthmap tags with “digital epidemiology”, Swiss Re report

A brief history of the term

2014: Healthmap tags with “digital epidemiology”, Swiss Re report

2015: First conference DELSI - Digital Epidemiology and its Ethical, Legal, and Social Implications”

A brief history of the term

2014: Healthmap tags with “digital epidemiology”, Swiss Re report

2015: First conference DELSI - Digital Epidemiology and its Ethical, Legal, and Social Implications”

2016: Workshops, original research papers with “Digital Epidemiology” in the title (e.g. Bakker et al. PNAS)

A brief history of the term

2014: Healthmap tags with “digital epidemiology”, Swiss Re report

2015: First conference DELSI - Digital Epidemiology and its Ethical, Legal, and Social Implications”

2016: Workshops, original research papers with “Digital Epidemiology” in the title (e.g. Bakker et al. PNAS)

.

.

.

2020??

What’s the fuss about?

Digital Epidemiology - Computational Epidemiology by another name?

What’s the fuss about?

Digital Epidemiology - Computational Epidemiology by another name?

No.

What’s the fuss about?

Digital Epidemiology - Computational Epidemiology by another name?

No.

Computational is about **computation** (e.g. simulation models)

What’s the fuss about?

Digital Epidemiology - Computational Epidemiology by another name?

No.

Computational is about **computation** (e.g. simulation models)

Digital is about **digitized data** (e.g. data from social media, phones, etc.)

What’s the fuss about?

Digital Epidemiology - Computational Epidemiology by another name?

No.

Computational is about **computation** (e.g. simulation models)

Digital is about **digitized data** (e.g. data from social media, phones, etc.)

But ok, all data is now going digital, so what?

What's the fuss about?

Digital Epidemiology - Computational Epidemiology by another name?

No.

Computational is about **computation** (e.g. simulation models)

Digital is about **digitized data** (e.g. data from social media, phones, etc.)

But ok, all data is now going digital, so what?

To me, the exciting aspect about digital data is when it captures something that is **hard / impossible to capture otherwise**. In this context, the fact that most **digital epidemiology sources are not health systems** is pertinent.

Digital data

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

Digital data

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

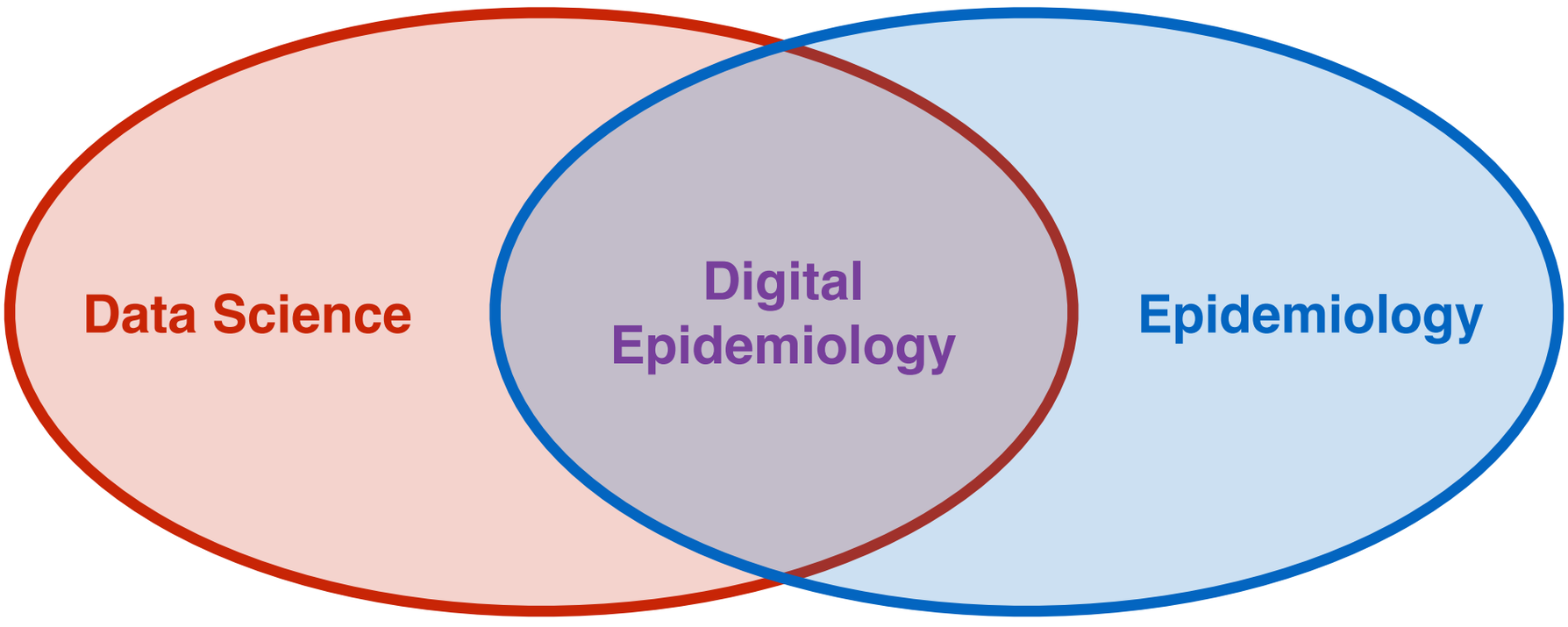
Digital Epidemiology: Epidemiology with digital data that captures states, events, processes, etc. that are difficult to capture otherwise*.

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

Digital data

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

What’s the problem?

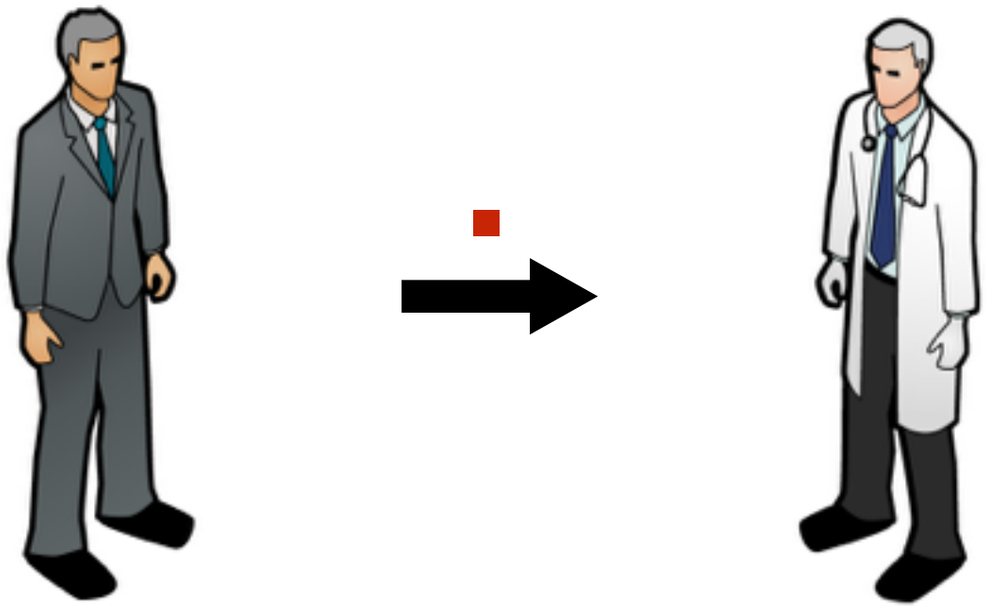
Is digital epidemiology a field in search of a problem?

Traditional Epidemiology

CDC et al.

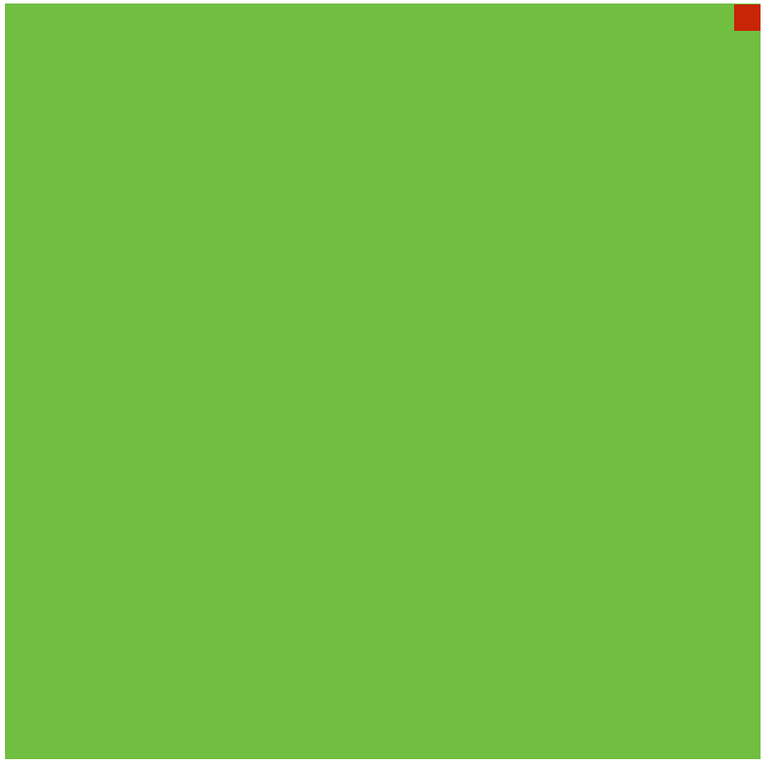
Academia

Traditional Epidemiology

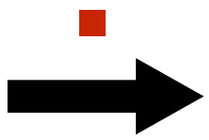


CDC et al.

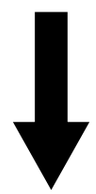
Academia



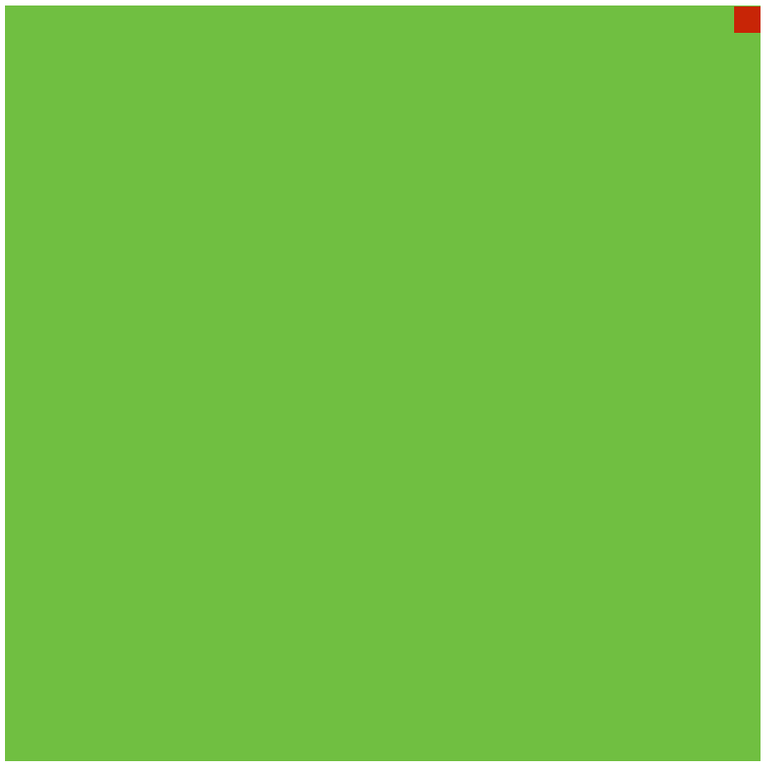
Traditional Epidemiology



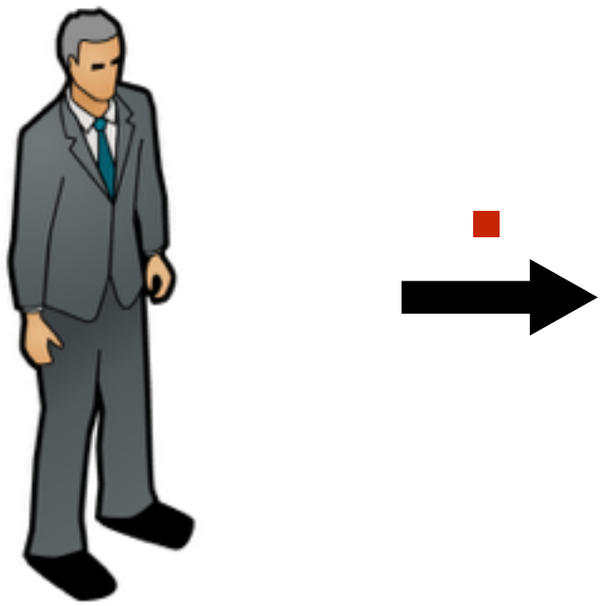
CDC et al.



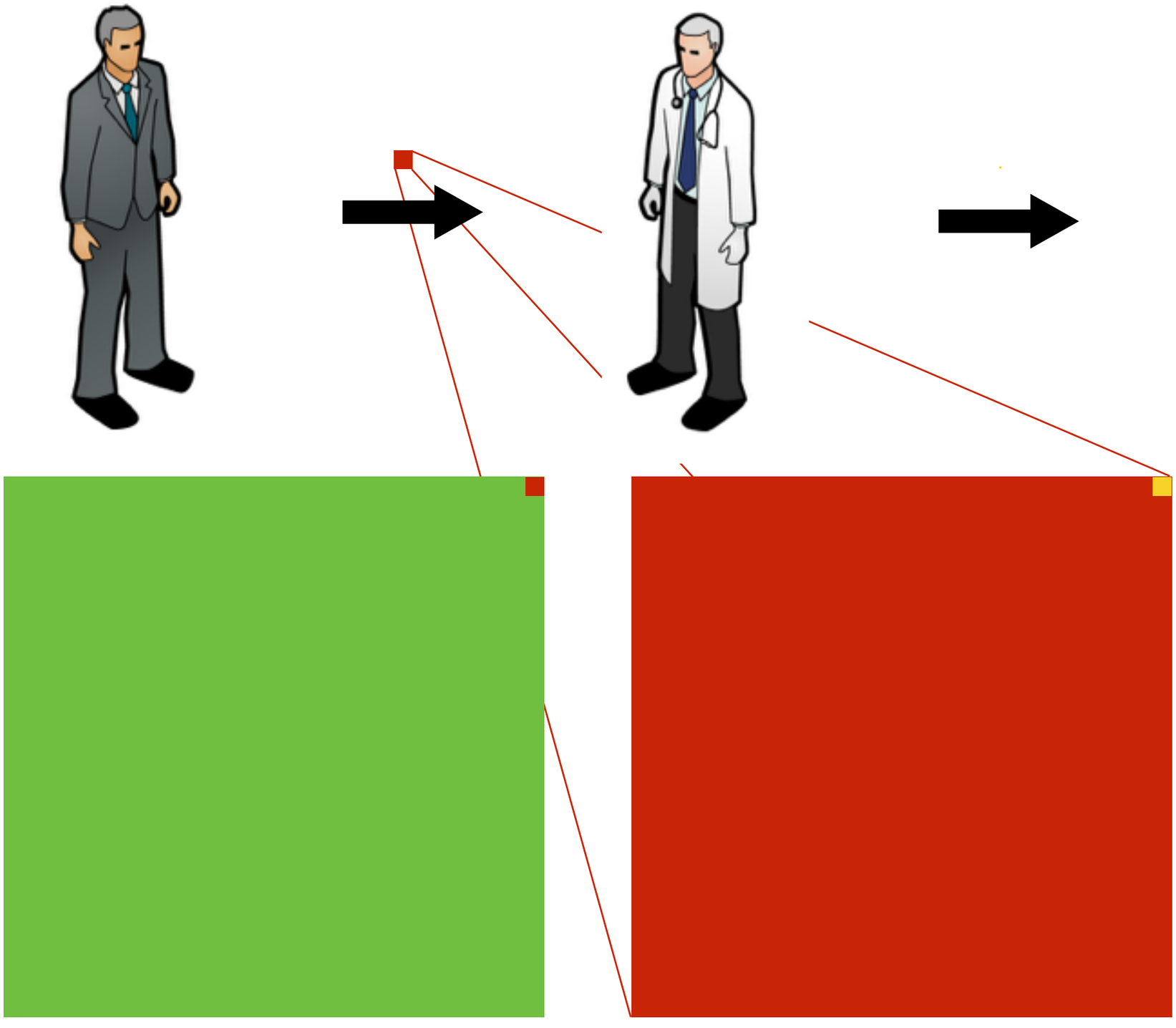
Academia



Traditional Epidemiology



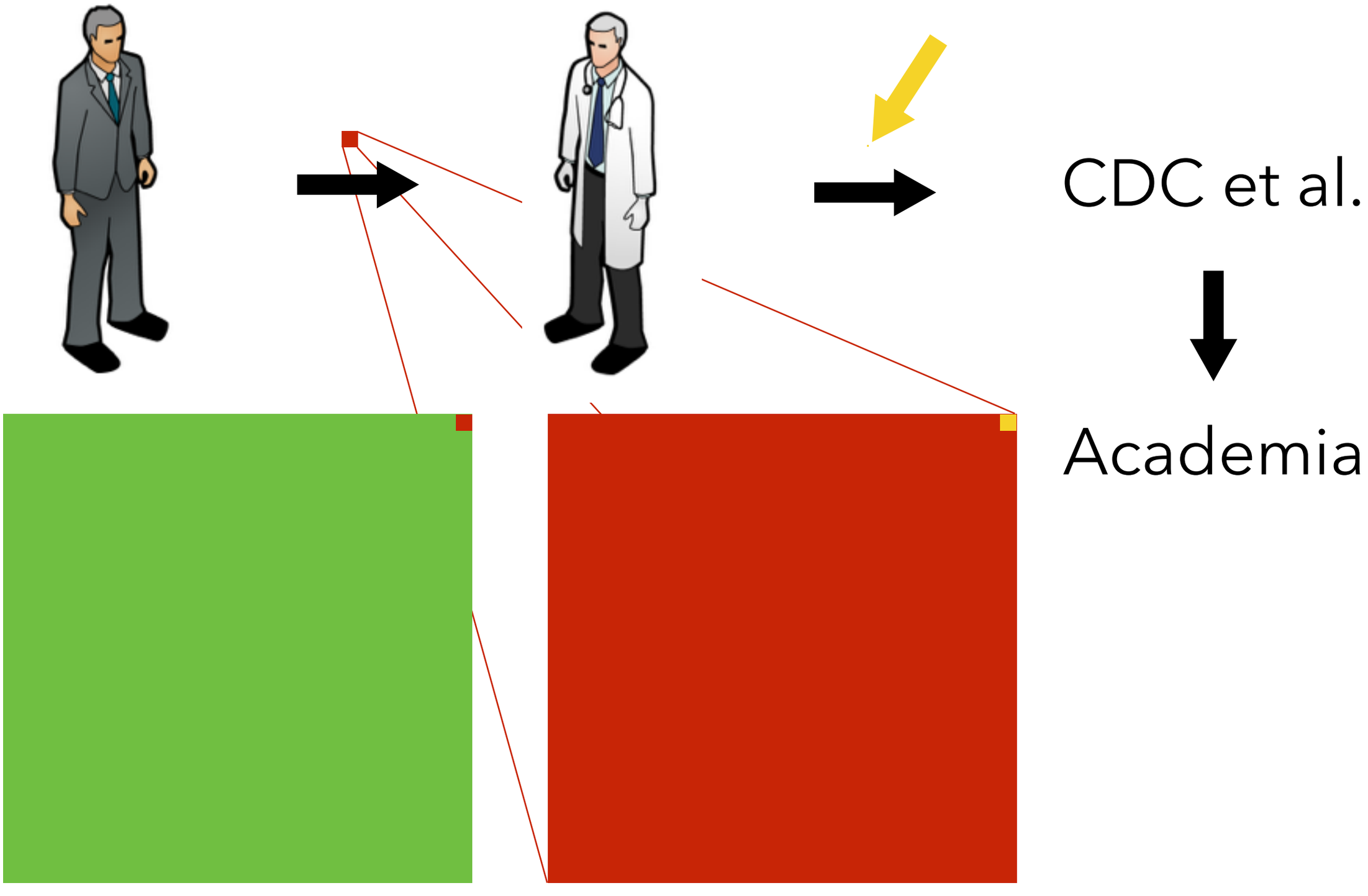
Traditional Epidemiology



CDC et al.

Academia

Traditional Epidemiology



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.

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

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

From Personalized Health…

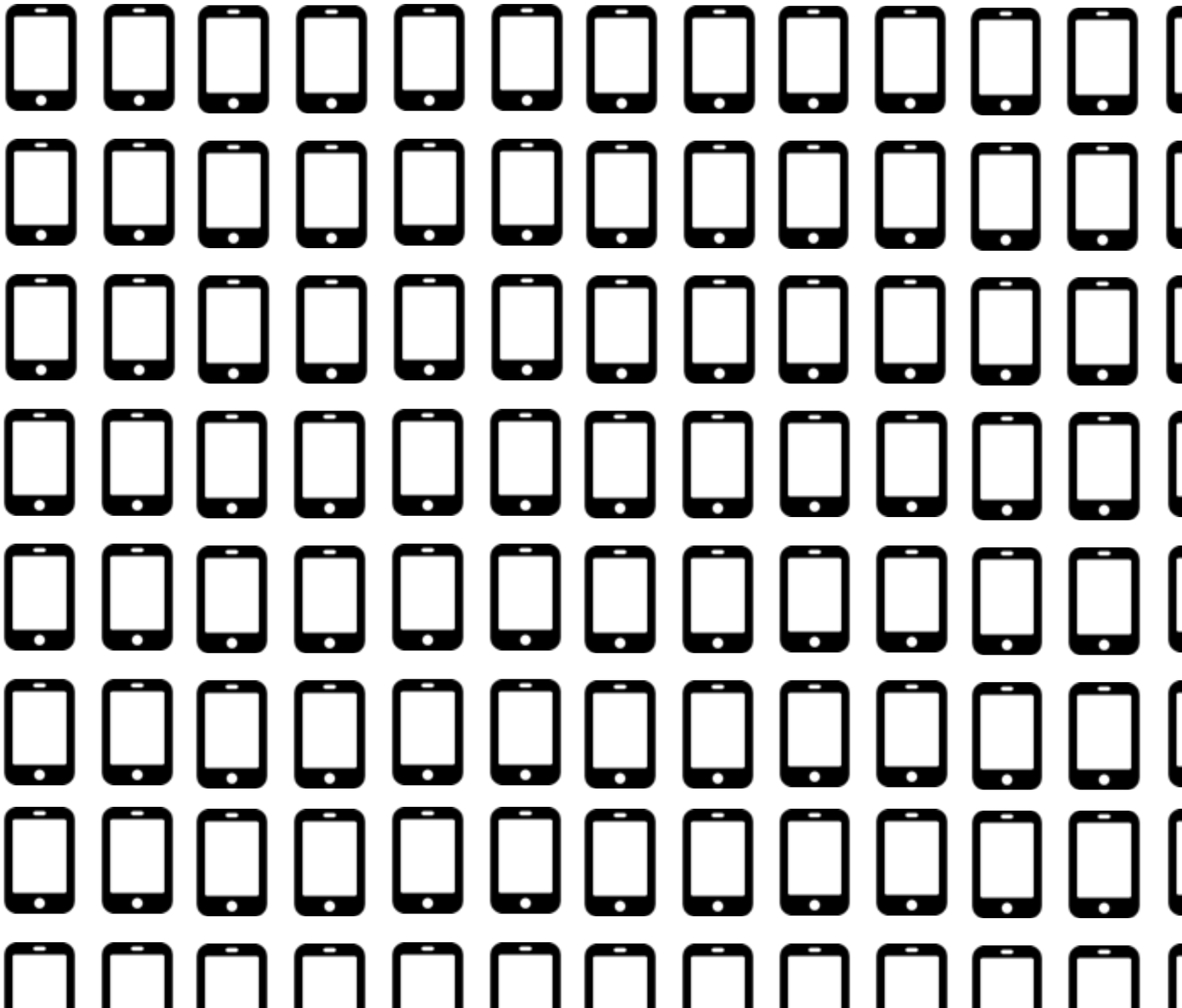
- my DNA
- my *omics data
- my location data
- my activity data (heart rate, etc.)
- my lab tests
- what I ate
- how I slept
- how I feel
- my health history
- etc.

shared via mobile apps

“The patient will see you now”

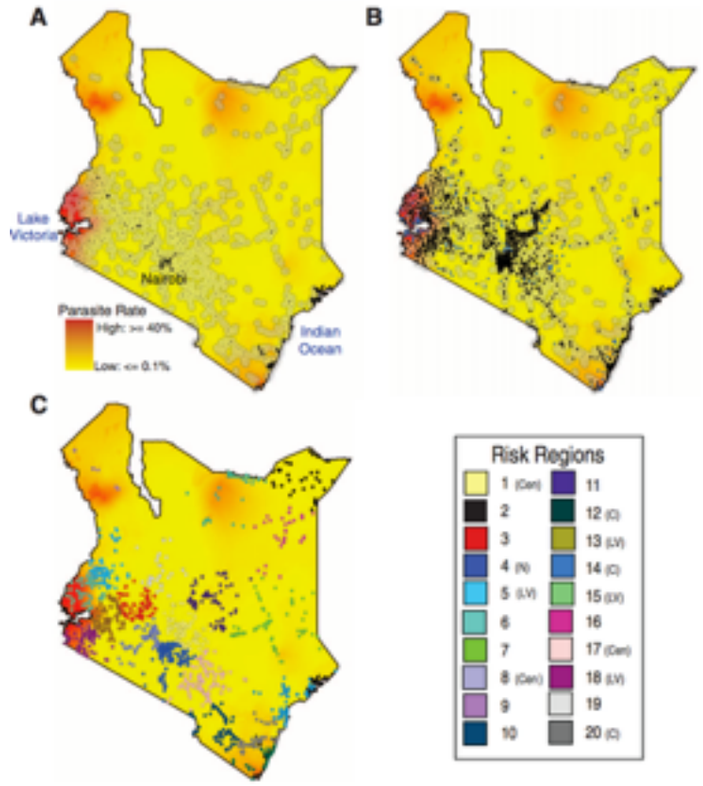
…To Truly Global Health

shared via mobile apps



Mobile phones - Malaria elimination

Wesolowski et al 2012

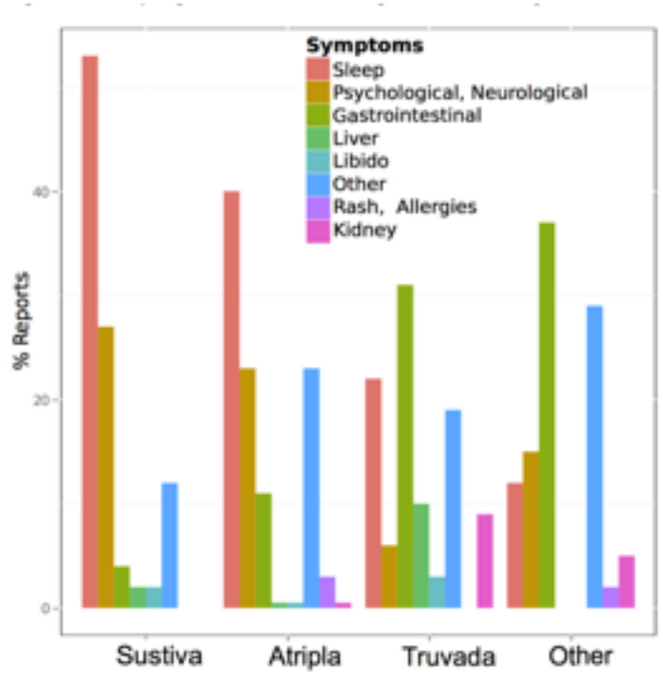


Wikipedia - Influenza forecasting

McIver & Brownstein 2014

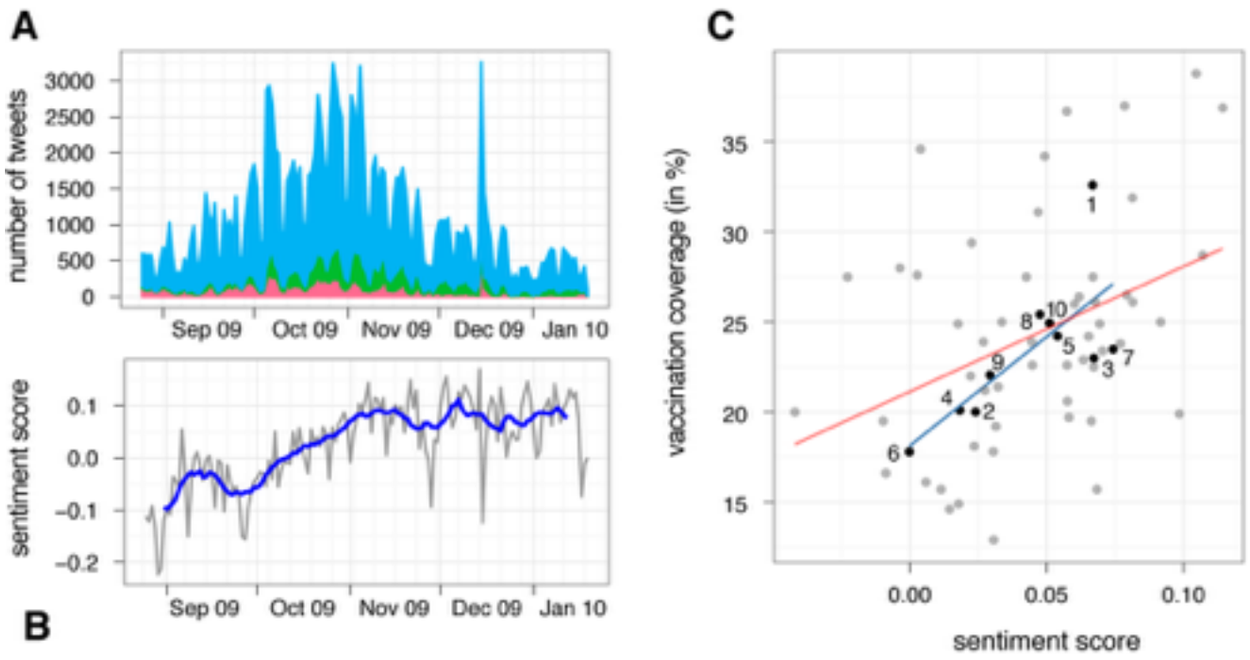
Twitter - Pharmacovigilance

Adrover et al 2015



Twitter - Vaccine uptake

Salathé & Khandelwal, 2011



Digital data

Data Science: Extracting knowledge from (un / structured) data using statstics, machine learning, data mining, etc.

Three major challenges:

1. **Access to data**

   “What happens to Digital Epidemiology when Twitter shuts down?”

Digital data

Data Science: Extracting knowledge from (un / structured) data using statstics, machine learning, data mining, etc.

Three major challenges:

1. **Access to data**

   “What happens to Digital Epidemiology when Twitter shuts down?”

2. **Validating the extracted “knowledge”**

   “You can force the data to tell you anything when you torture it enough!”

Digital data

Data Science: Extracting knowledge from (un / structured) data using statstics, machine learning, data mining, etc.

Three major challenges:

1. **Access to data**
   “What happens to Digital Epidemiology when Twitter shuts down?”

2. **Validating the extracted “knowledge”**
   “You can force the data to tell you anything when you torture it enough!”

3. **Methods from CS**
   “How many epidemiologists are proficient in running deep learning models on GPU clusters?”

1. Access to data

Use it while it lasts, and build your own data pipelines!

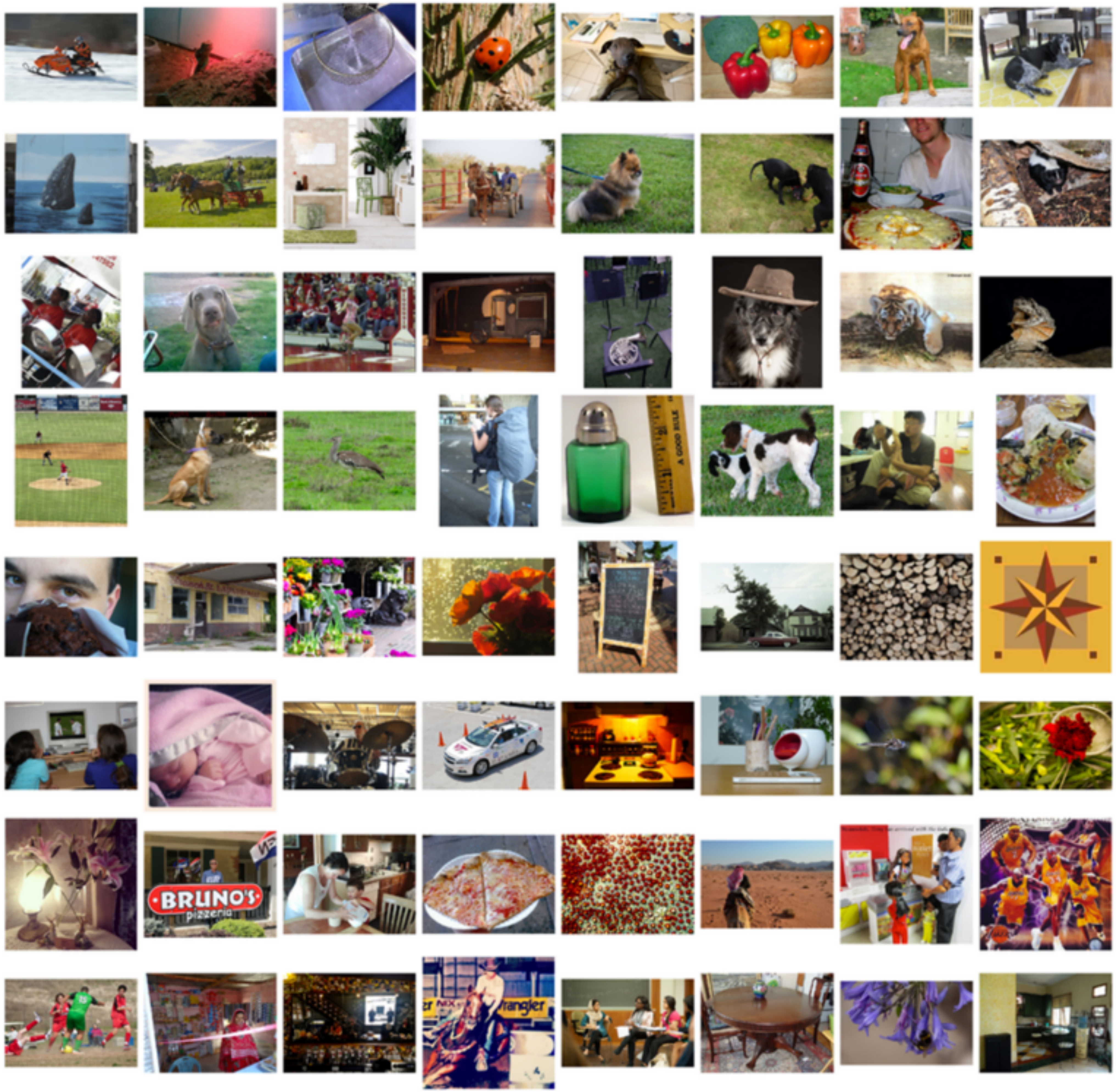
Data > Algorithms

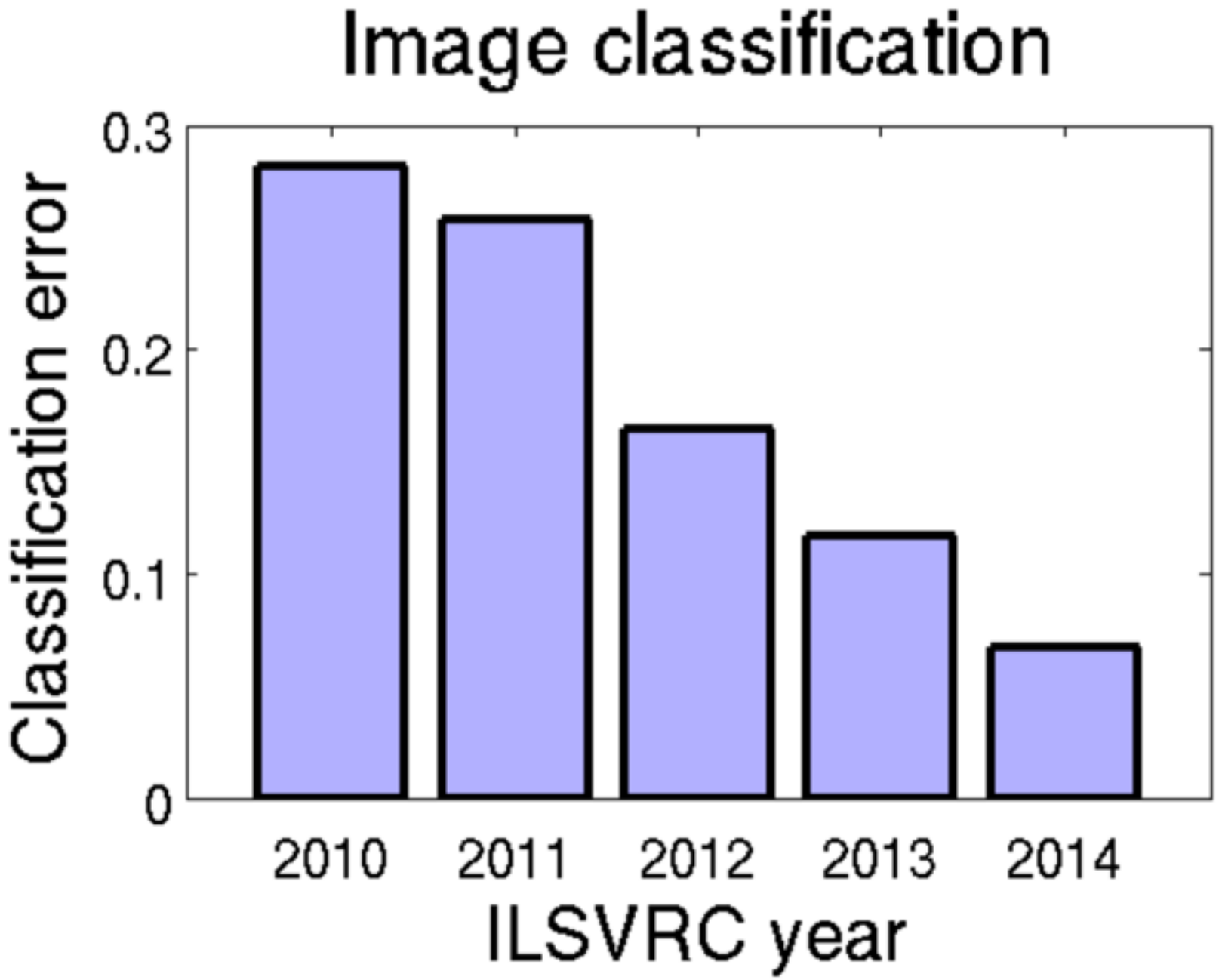
Current publication system doesn’t reflect this (yet)

Datasets over Algorithms

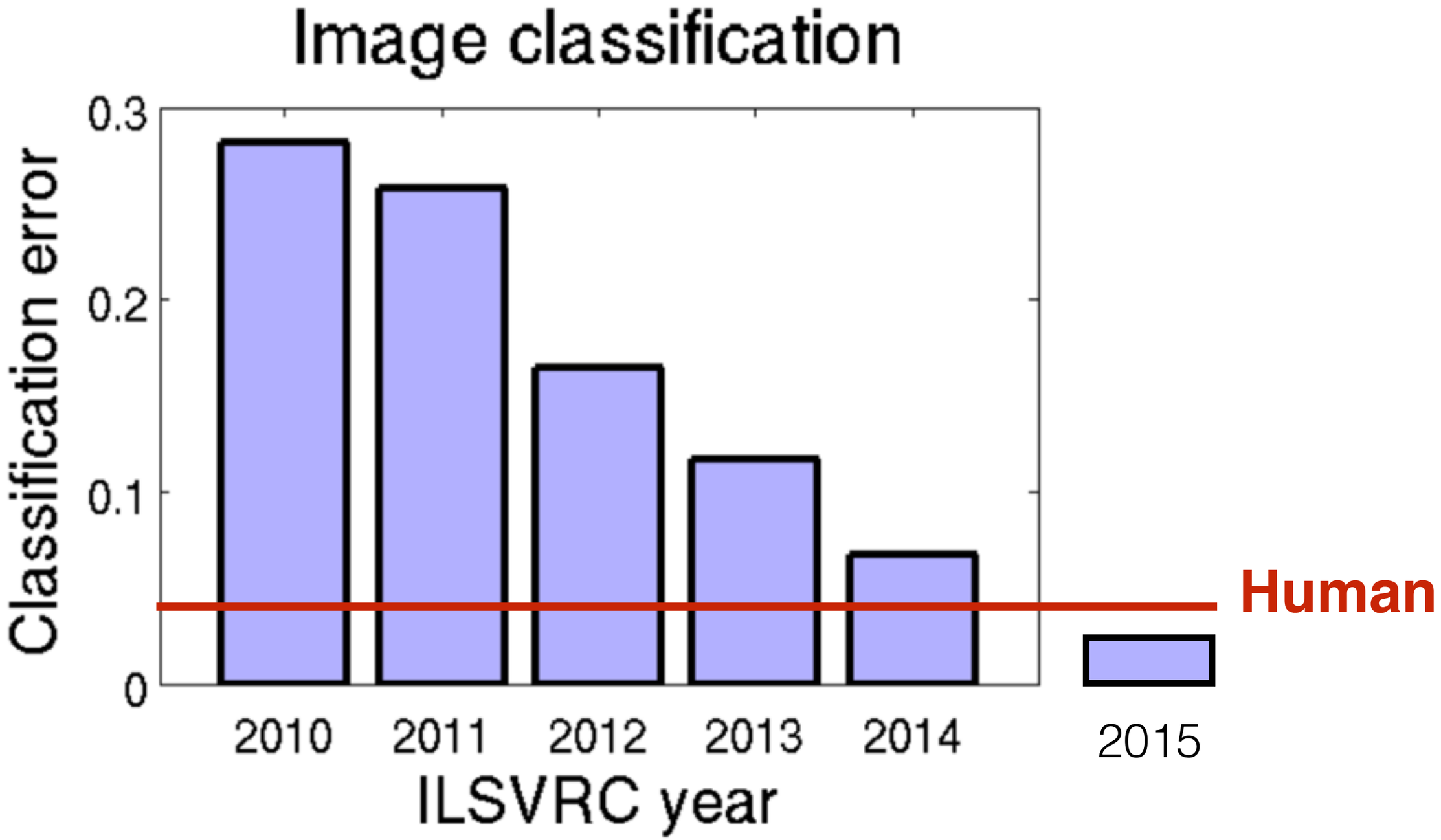
Year	Breakthroughs in AI	Datasets (First Available)	Algorithms (First Proposed)
1994	Human-level spontaneous speech recognition	Spoken Wall Street Journal articles and other texts (1991)	Hidden Markov Model (1984)
1997	IBM Deep Blue defeated Garry Kasparov	700,000 Grandmaster chess games, aka "The Extended Book" (1991)	Negascout planning algorithm (1983)
2005	Google's Arabic- and Chinese-to-English translation	1.8 trillion tokens from Google Web and News pages (collected in 2005)	Statistical machine translation algorithm (1988)
2011	IBM Watson became the world Jeopardy! champion	8.6 million documents from Wikipedia, Wiktionary, Wikiquote, and Project Gutenberg (updated in 2010)	Mixture-of-Experts algorithm (1991)
2014	Google's GoogLeNet object classification at near-human performance	ImageNet corpus of 1.5 million labeled images and 1,000 object categories (2010)	Convolution neural network algorithm (1989)
2015	Google's Deepmind achieved human parity in playing 29 Atari games by learning general control from video	Arcade Learning Environment dataset of over 50 Atari games (2013)	Q-learning algorithm (1992)
Average No. of Years to Breakthrough:		**3 years**	**18 years**

http://www.spacemachine.net/views/2016/3/datasets-over-algorithms


A GOOD RULE
BRUNO'S
pizzeria



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



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

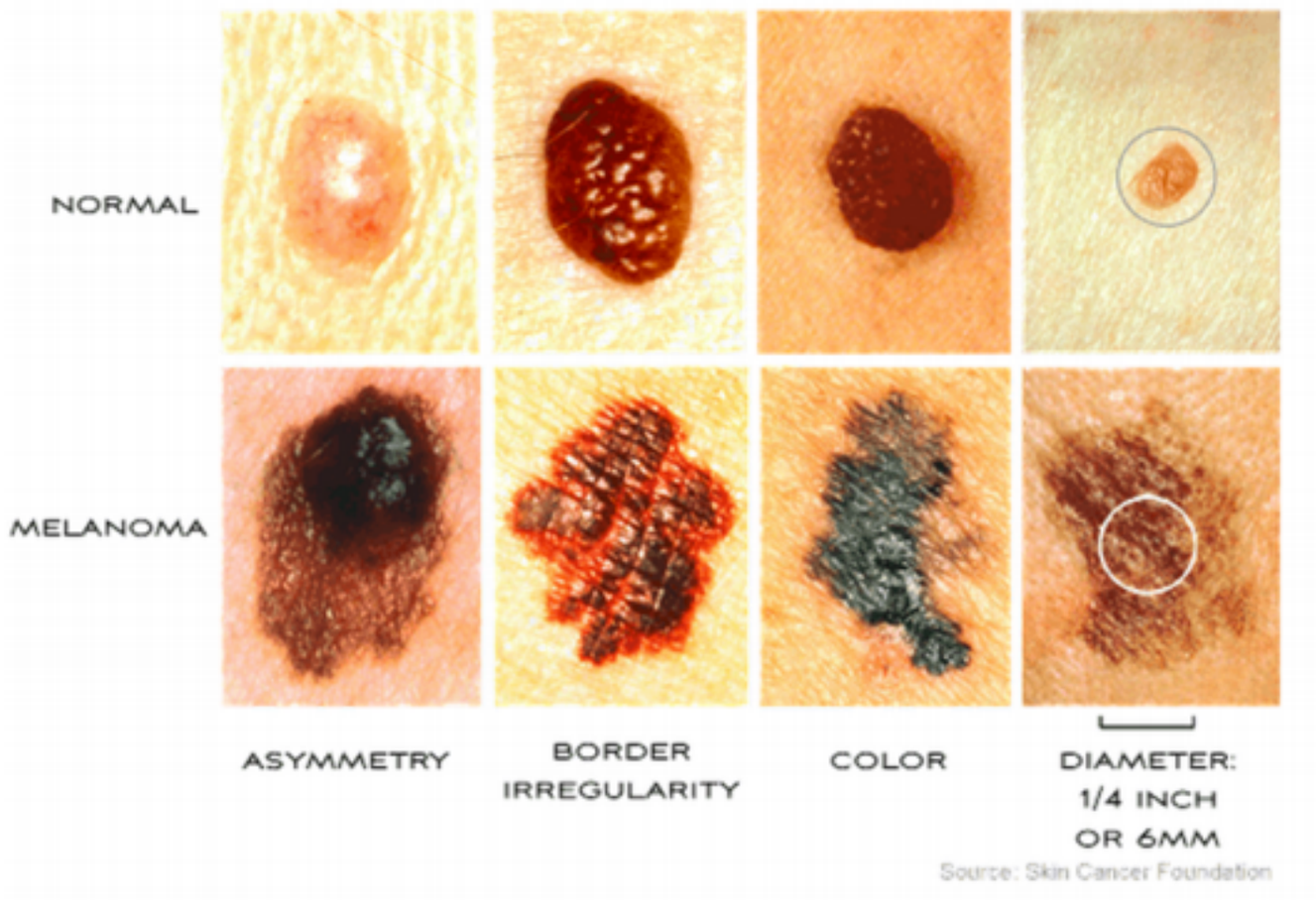


Figure 1: Skin cancers are visually distinct from benign marks, and can be screened for using the ABCDE features of skin cancer: A - asymmetry, B - border irregularity, C - multiple colors, D - large diameters, E - rapid evolution in time.

Accuracy in detecting correct cancer type

Dermatologist: 46%
Algorithm: 60%

Accuracy in detecting cancer

Algorithm: 90%

http://cs231n.stanford.edu/reports/esteva-paper.pdf

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.

PlantVillage: Machine Learning for Disease Recognition

Image Recognition
(Machine Learning)

Diagnosis
Treatment Suggestions

2. Validating the extracted knowledge

If it can be validated by traditional epidemiology, is it worth doing it with digital epidemiology? (yes if faster || broader)

2. Validating the extracted knowledge

If it can be validated by traditional epidemiology, is it worth doing it with digital epidemiology? (yes if faster || broader)

Why is the gold standard the gold standard? (e.g. flu)

2. Validating the extracted knowledge

If it can be validated by traditional epidemiology, is it worth doing it with digital epidemiology? (yes if faster || broader)

Why is the gold standard the gold standard? (e.g. flu)

Work on issues where traditional epidemiology has gaps

2. Validating the extracted knowledge

If it can be validated by traditional epidemiology, is it worth doing it with digital epidemiology? (yes if faster || broader)

Why is the gold standard the gold standard? (e.g. flu)

Work on issues where traditional epidemiology has gaps

Work as part of a lager epidemiology team

3. Methods from CS

Q: I’m interested in Public Health; I have a Bachelor / Masters in Epidemiology; I’d like to become a Digital Epidemiologist. How do I do it?

3. Methods from CS

Q: I’m interested in Public Health; I have a Bachelor / Masters in Epidemiology; I’d like to become a Digital Epidemiologist. How do I do it?

A: Learn Data Science. Even better, learn Computer Science too.

3. Methods from CS

Q: I'm interested in Public Health; I have a Bachelor / Masters in Epidemiology; I'd like to become a Digital Epidemiologist. How do I do it?

A: Learn Data Science. Even better, learn Computer Science too.

Think like an Epidemiologist, Work like a Data Sciencist

3. Methods from CS

Q: I'm interested in Public Health; I have a Bachelor / Masters in Epidemiology; I'd like to become a Digital Epidemiologist. How do I do it?

A: Learn Data Science. Even better, learn Computer Science too.

Think like an Epidemiologist, Work like a Data Sciencist

Digital Epidemiology = Epidemiology * Data Science

3. Methods from CS

Q: I'm interested in Public Health; I have a Bachelor / Masters in Epidemiology; I'd like to become a Digital Epidemiologist. How do I do it?

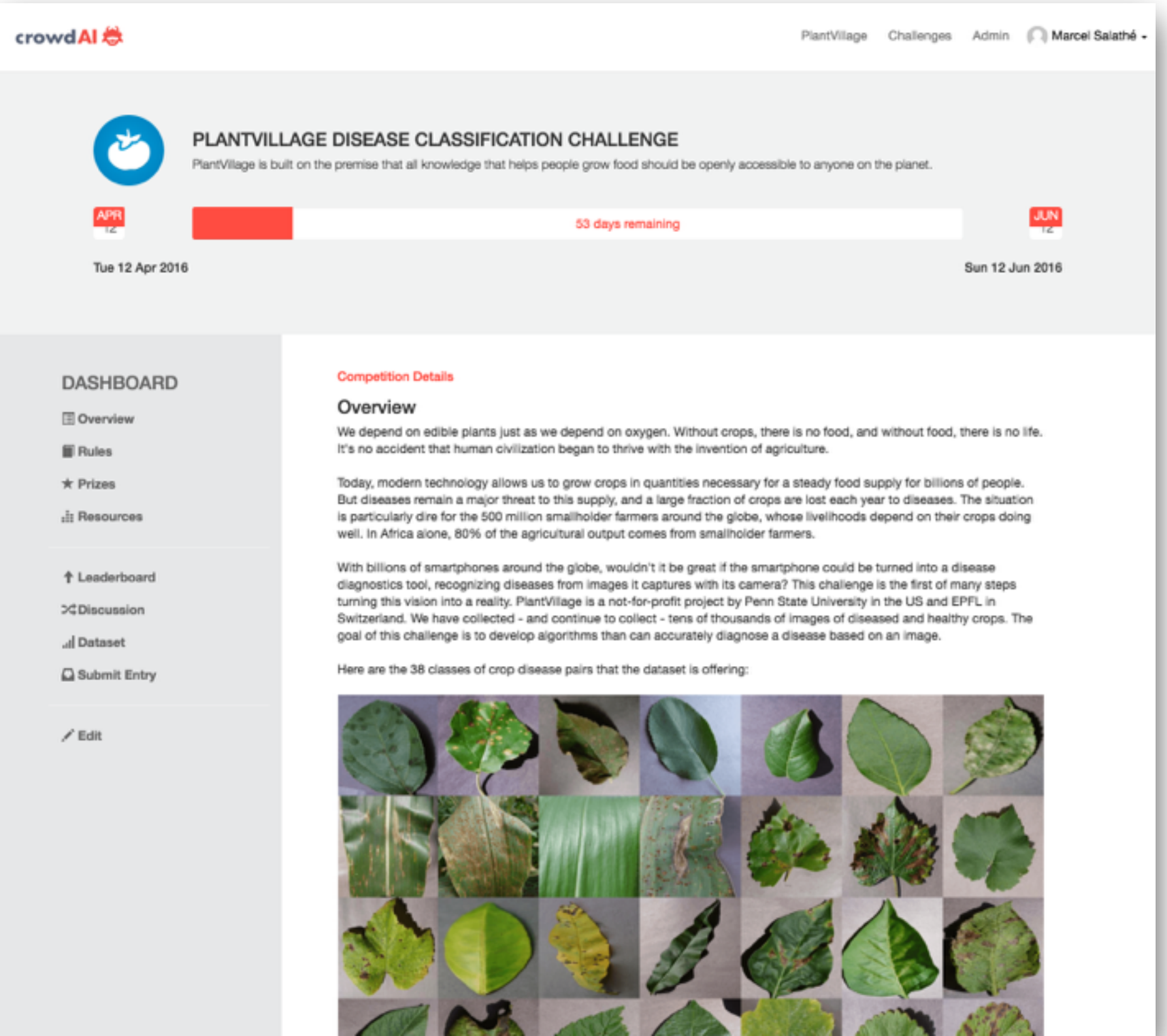
A: Learn Data Science. Even better, learn Computer Science too.

Think like an Epidemiologist, Work like a Data Sciencist

Digital Epidemiology = Epidemiology * Data Science

It takes much more effort to learn Data Science / CS than Epidemiology.

Crowdsourcing Machine Learning



From Big Data to Knowledge Generation, *fast*?

In science, increasingly through crowdsourcing.

Netflix Challenge, Kaggle challenges, ImageNet Challenge, Dream Challenges etc.

www.crowdai.org



Online | Mobile | Global

Bringing **Digital Epidemiology** to the Next Level

Marcel Salathé, Digital Epidemiology Lab, EPFL

@marcelsalathe