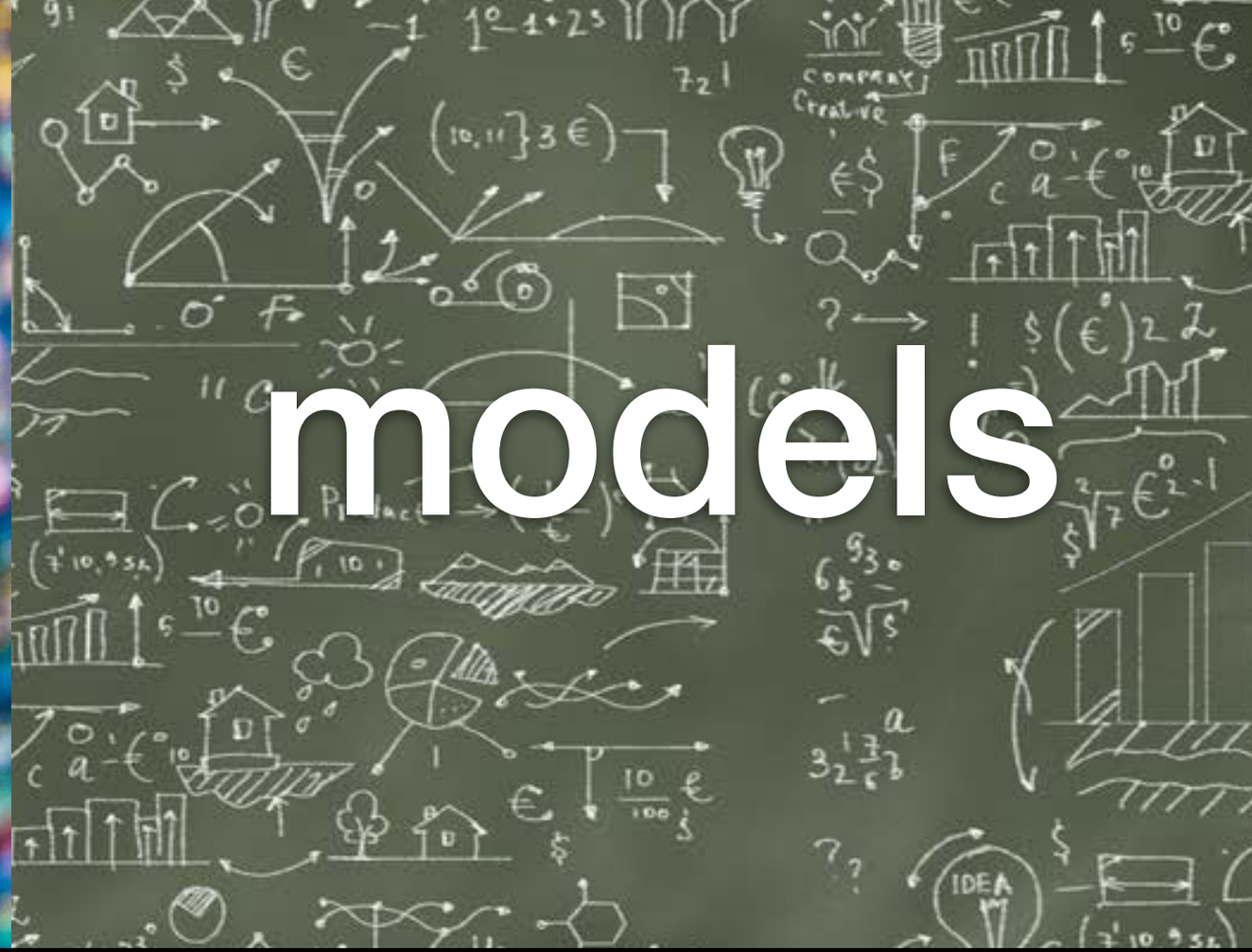
Data Science: emerging opportunities and challenges for regulatory decision making

Ciro Cattuto
University of Torino & ISI Foundation
@ciro

Veterinary Big Data Stakeholder Forum
European Medicines Agency
2 June 2021


(big) data


(big) data


models

(big) data

models

decisions

(big) data

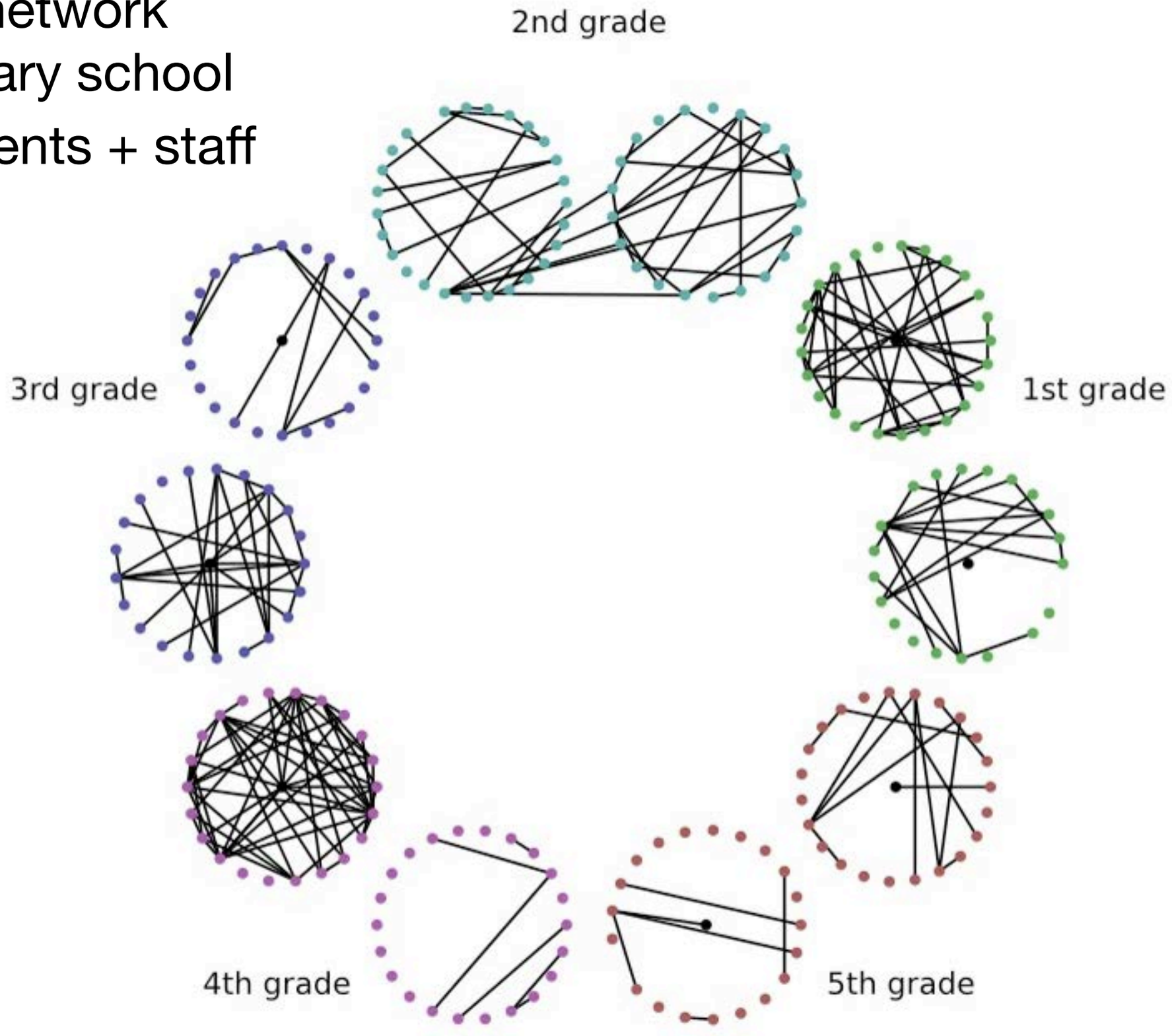
models

?

decisions

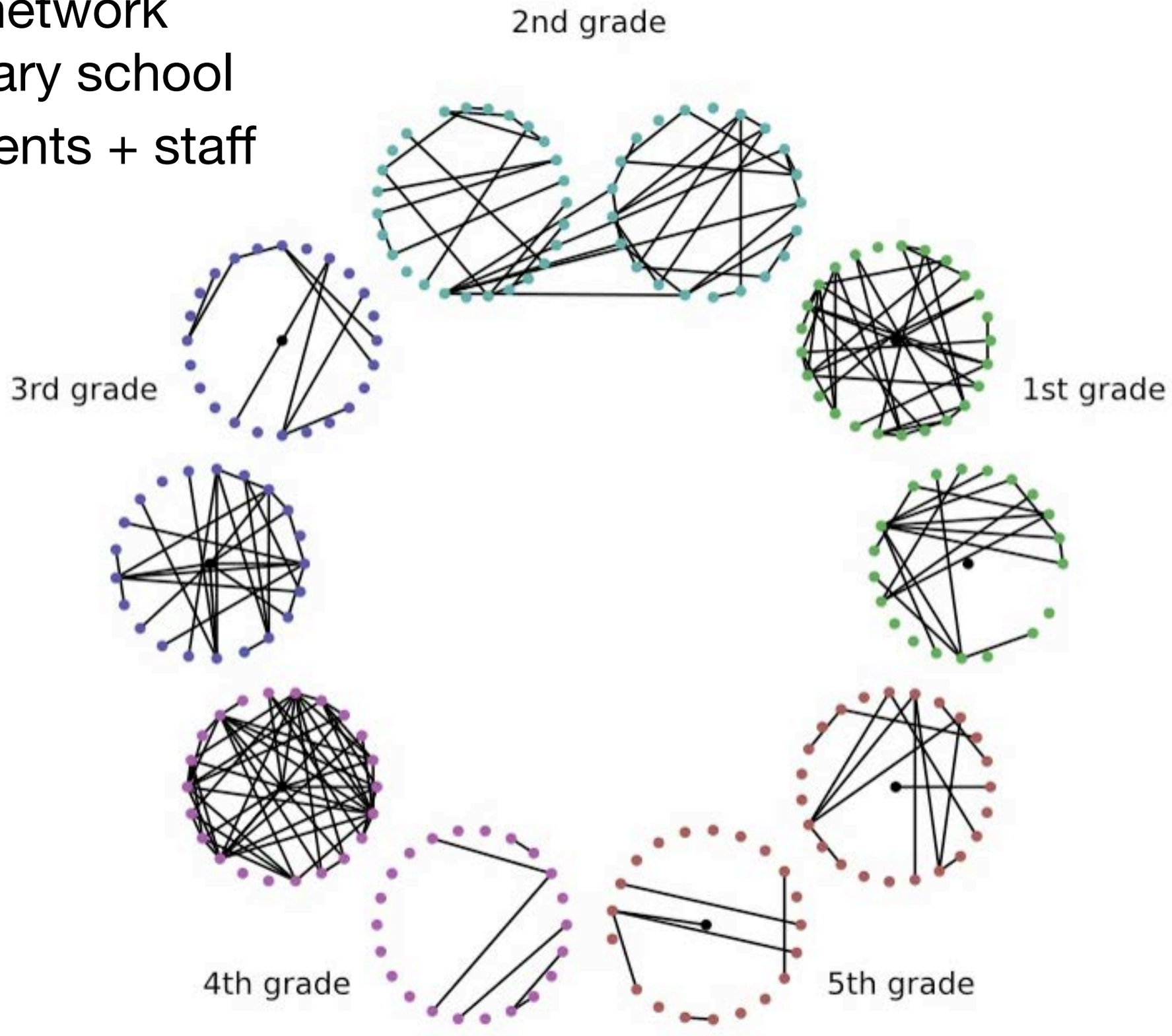

1.1 billion taxi pickup locations
6 years
toddwschneider.com

contact network
in a primary school
250 students + staff



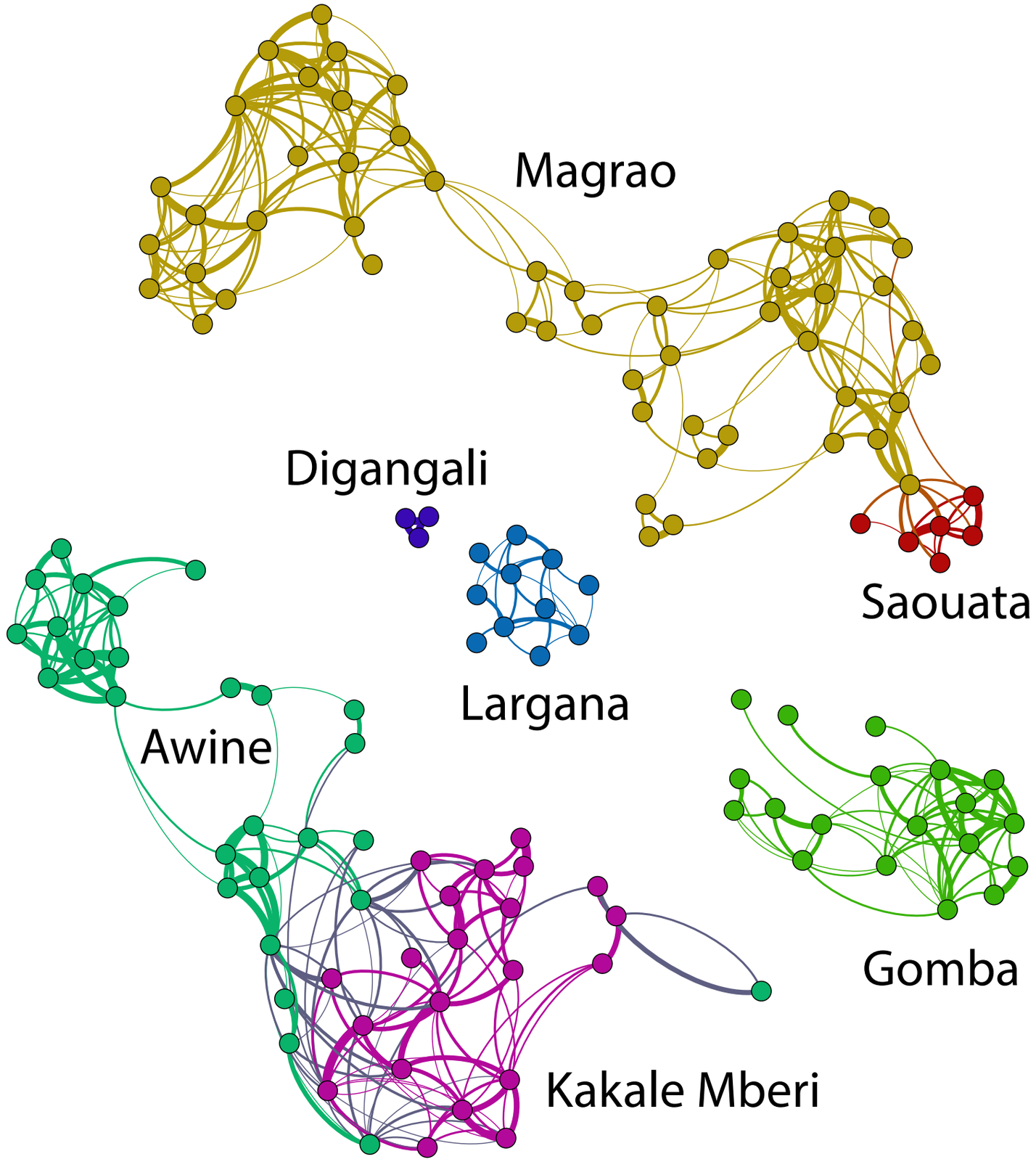
J. Stehlé *et al.*,
PLoS ONE 6(8), e23176 (2011)

contact network
in a primary school
250 students + staff

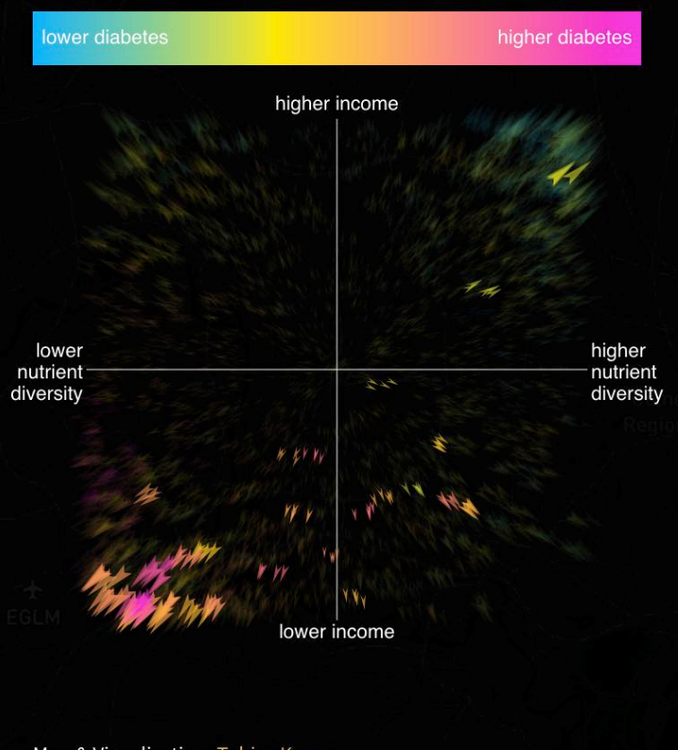


J. Stehlé *et al.*,
PLoS ONE 6(8), e23176 (2011)

free-roaming dogs in rural Chad

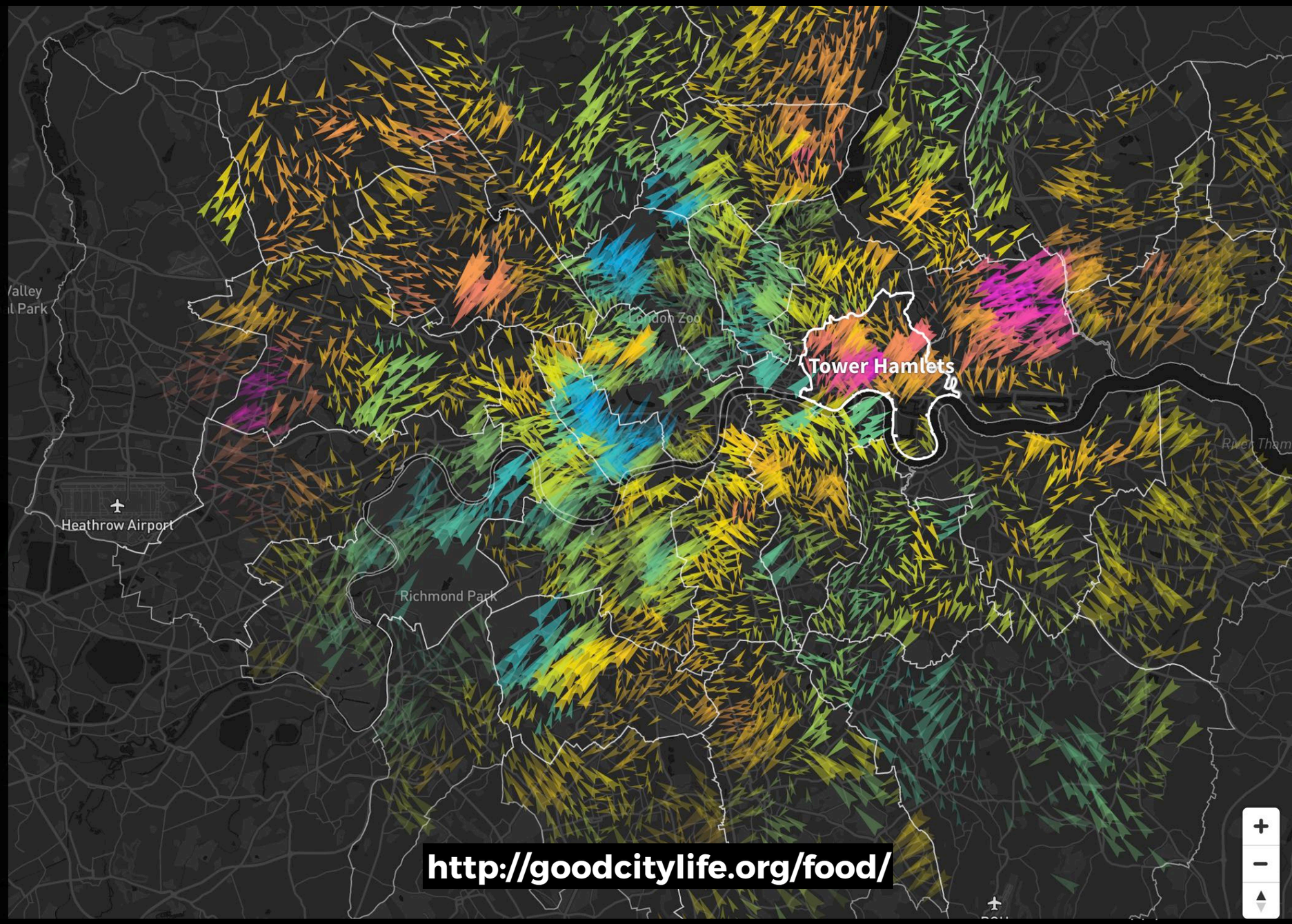


J. Wilson-Aggrawal *et al.*, PLOS Neg. Trop. Diseases 13(7), e0007565 (2019)



Map & Visualization: Tobias Kauer
Research & Data: Luca Maria Aiello, Rossano Schifanella, Daniele Quercia, Lucia Del Prete.

The map associates the prevalence of diabetes in London neighborhoods with the diversity of nutrients of 1.6B grocery purchases. In terms of diabetes prevalence, socio-economic conditions matter less than food habits: lower-income areas with healthy eating habits do not tend to be associated with diabetes. Hover on the quadrants to discover the areas with the healthiest eating habits. Hover on the map to see how different areas fare in terms of health and nutrition habits. Find out more.



slide courtesy of Rossano Schifanella

OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

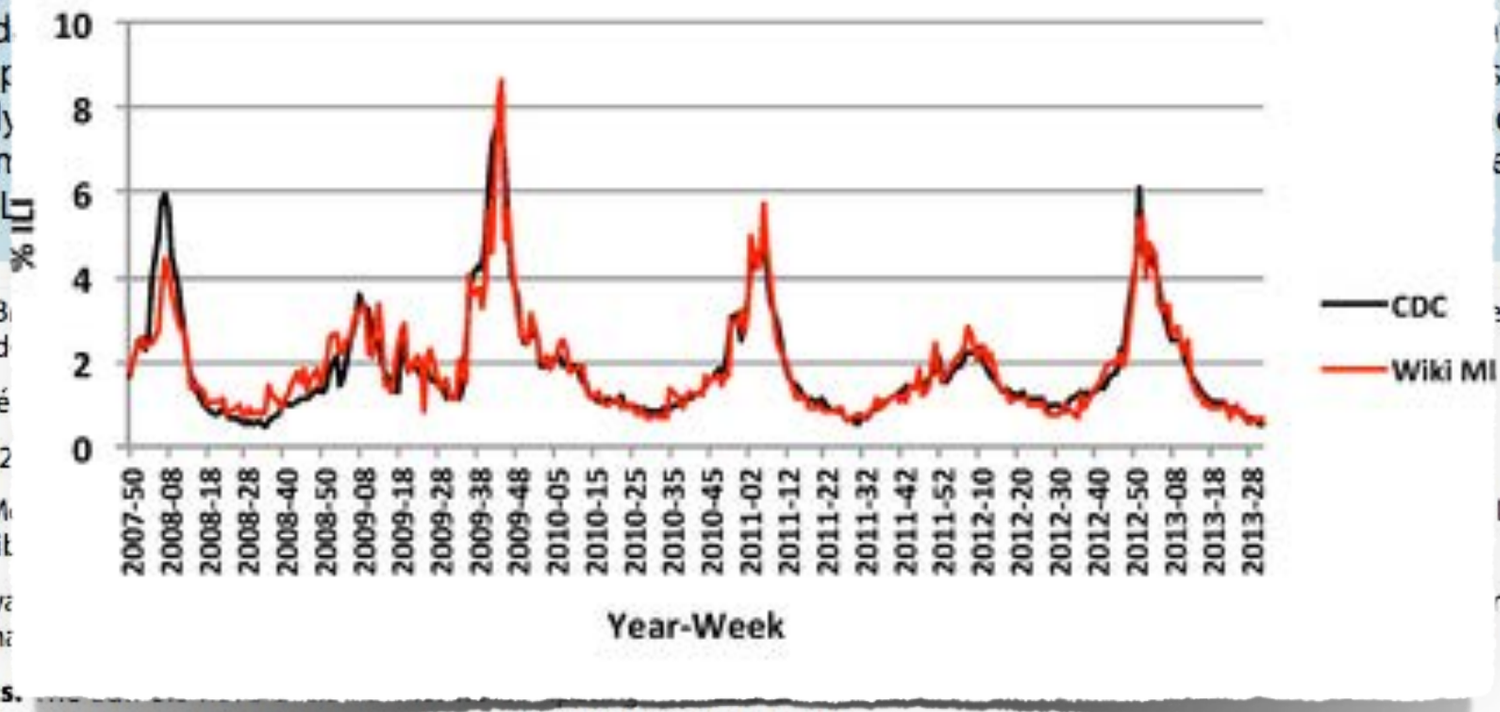
Wikipedia Usage Estimates Prevalence of Influenza-Like Illness in the United States in Near Real-Time

David J. McIver*, John S. Brownstein

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

Abstract

Circulating levels of both seasonal and pandemic influenza require constant surveillance to ensu
the population. While up-to-date information is critical, traditional surveillance systems can have d
to two weeks. We introduce a novel method of estimating, in near-real time, the level of influen
United States (US) by monitoring the rate of particular Wikipedia article views on a daily basis. We calc
times certain influenza- or health-related Wikipedia articles were accessed each day between Decembe
2013 and compared these data to official ILI activity levels provided by the Centers for Disease Control and
(CDC). We developed a Poisson model that accurately estimates the level of ILI activity in the American population, up to
two weeks ahea % over 294 weeks
of data. Wikiped ts (such as during
the 2009 H1N1 p eason). Wikipedia
usage accurately d was often more
accurate in its n ed for continuous
monitoring of IL



Citation: McIver DJ, B eal-Time. PLoS Comput
Biol 10(4): e1003581. d

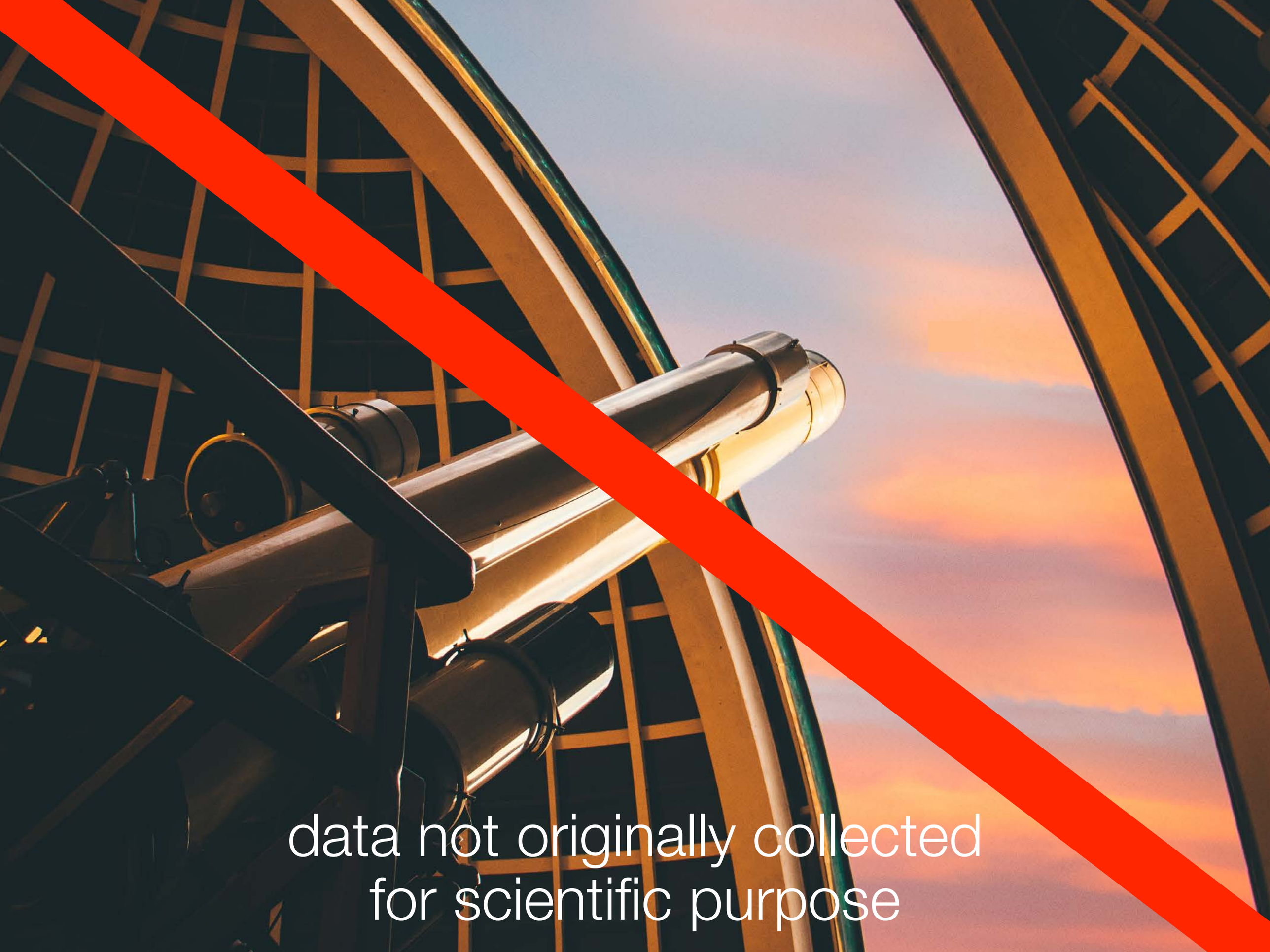
Editor: Marcel Salathé

Received December 2

Funding: This work wa no role in study design,
data collection and ana

Competing Interests.

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


data not originally collected
for scientific purpose

digital traces

digital traces

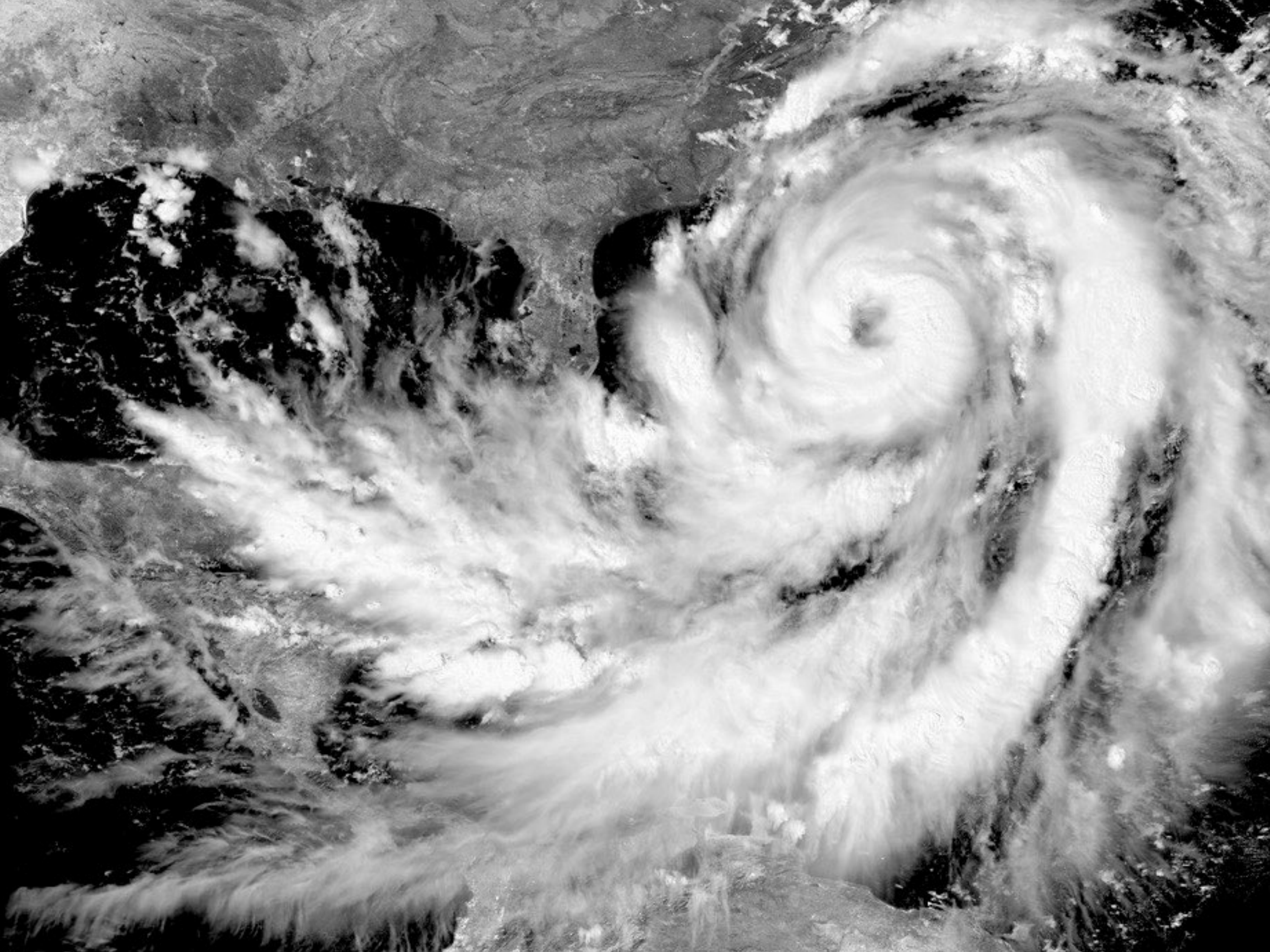
temporal horizon
limited reproducibility
limited context
data protection challenges
biases

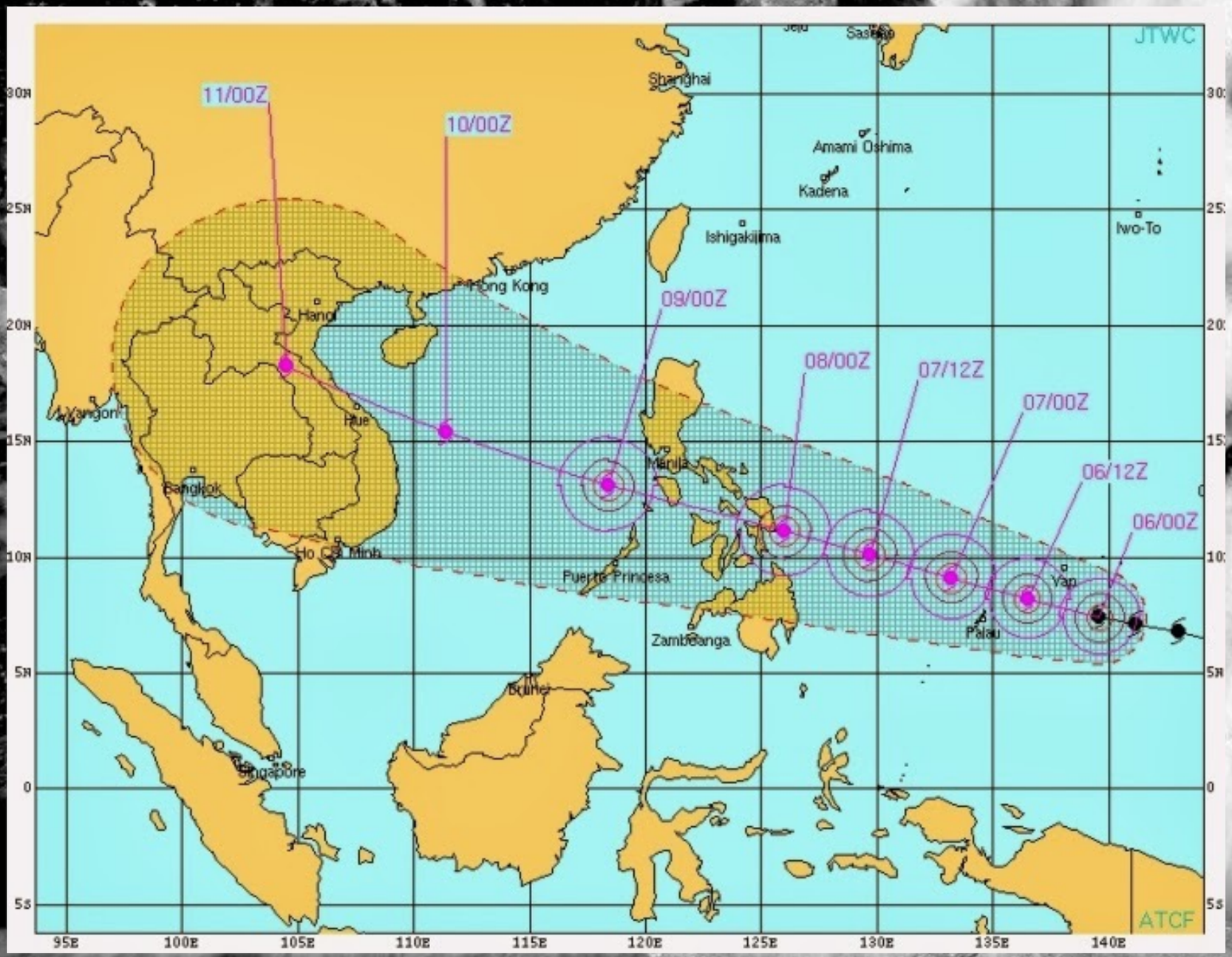
digital traces

temporal horizon
limited reproducibility
limited context
data protection challenges
biases

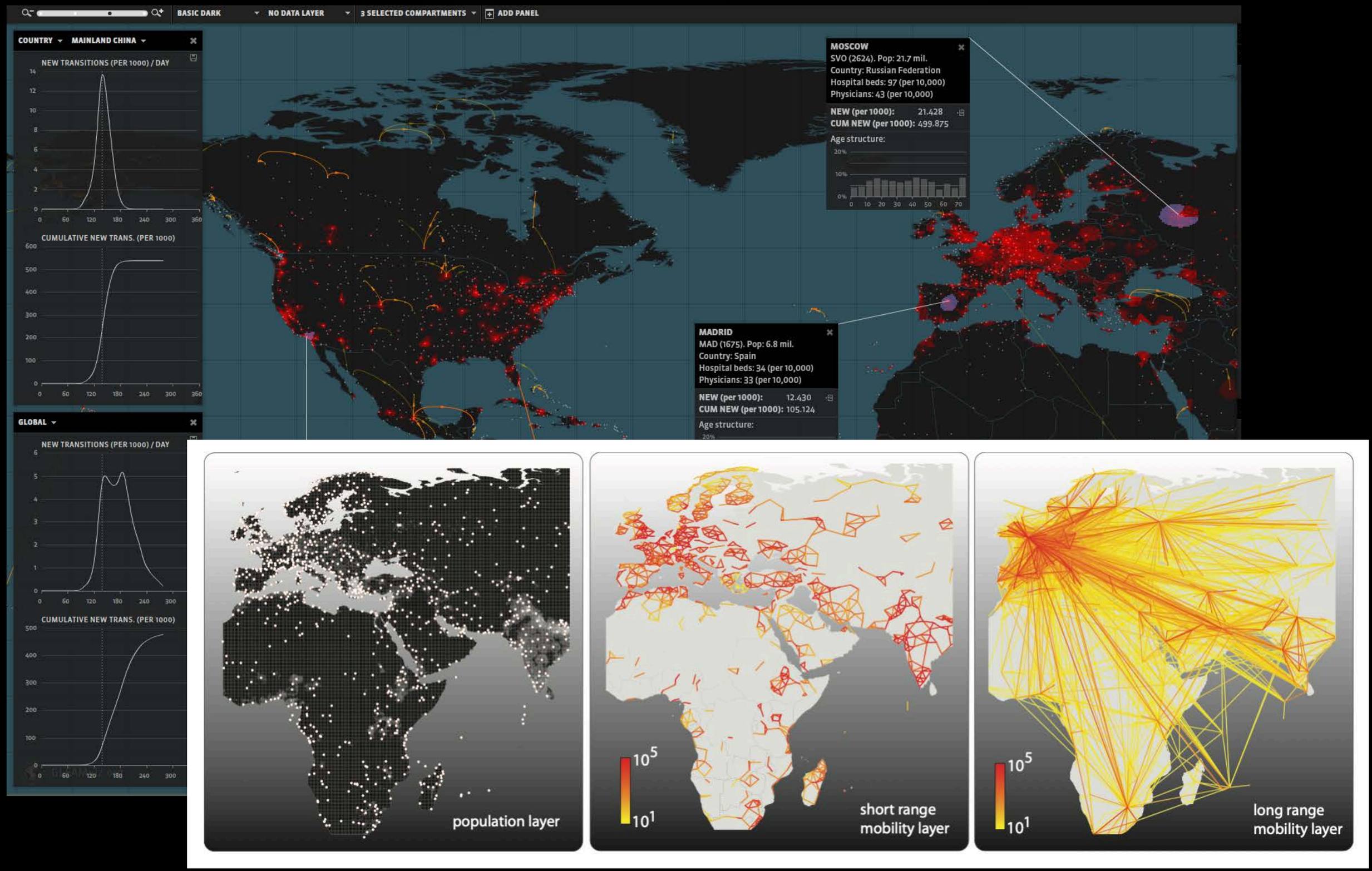
available as a side effect of many services
machine processable, pattern discovery
high coverage, can work at scale

models




JTWC
Shanghai
Amami Oshima
Kadena
Ishigakijima
Iwo-To
Hong Kong
Hanoi
Yangon
Hue
Bangkok
Ho Chi Minh
Manila
Puerto Princesa
Zamboanga
Palau
Yap
Brunei
Singapore
11/00Z
10/00Z
09/00Z
08/00Z
07/12Z
07/00Z
06/12Z
06/00Z
ATCF
30N
25N
20N
15N
10N
5N
0
5S
95E
100E
105E
110E
115E
120E
125E
130E
135E
140E

epidemic forecast



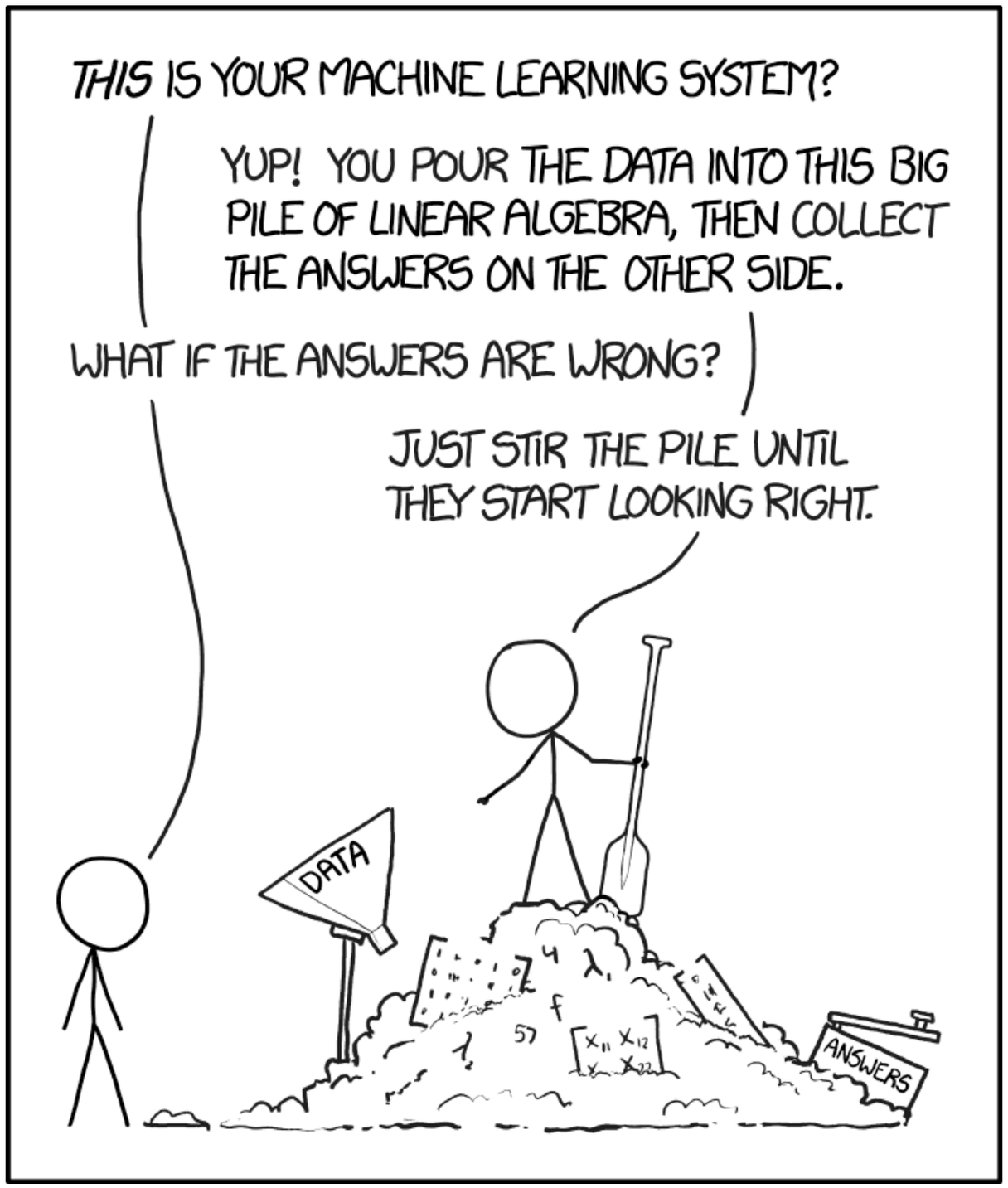
Northeastern

CENTER FOR INFERENCE & DYNAMICS OF INFECTIOUS DISEASES

Institute thématiques Inserm

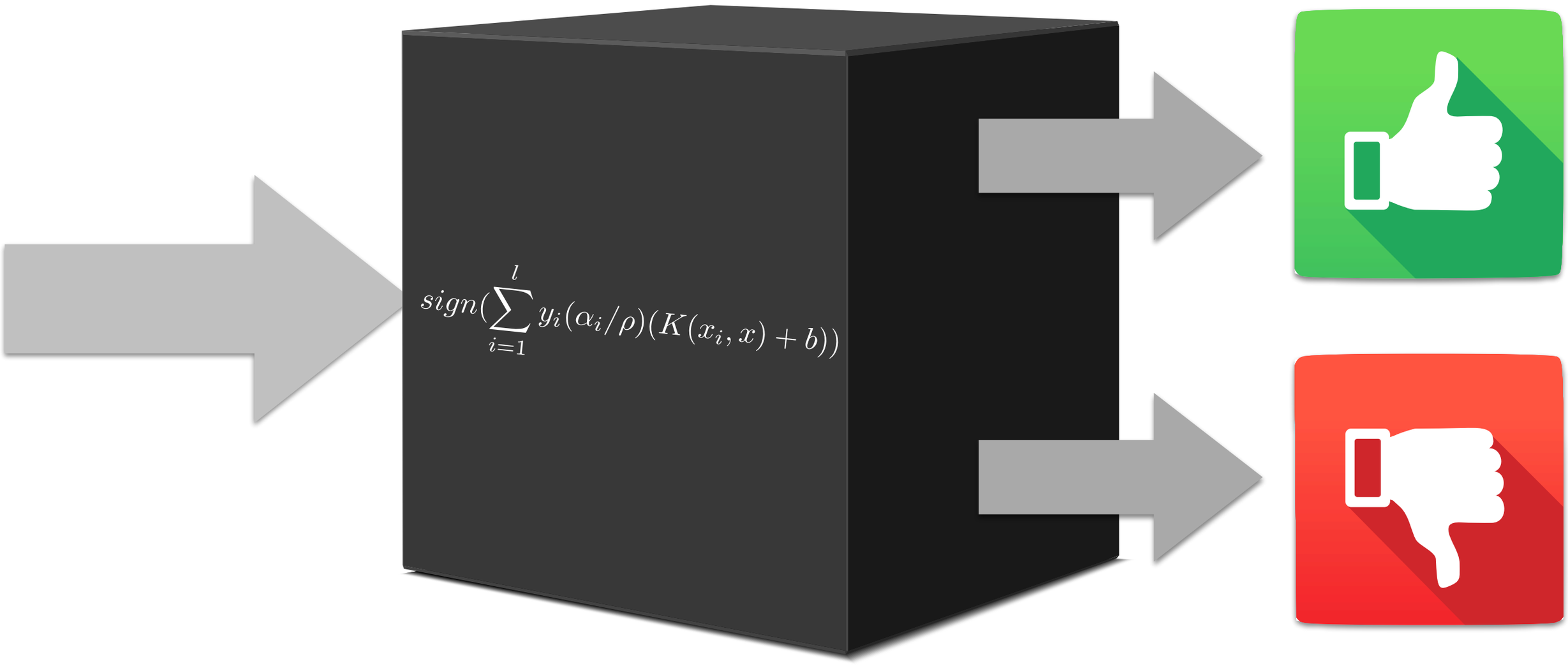
www.gleamviz.org

machine learning models



xkcd.com/1838

machine learning models



National Institutes of Health
Turning Discovery Into Health

ChestX-ray8

Home » News & Events » News Releases

NEWS RELEASES

Media Advisory Wednesday, September 27, 2017

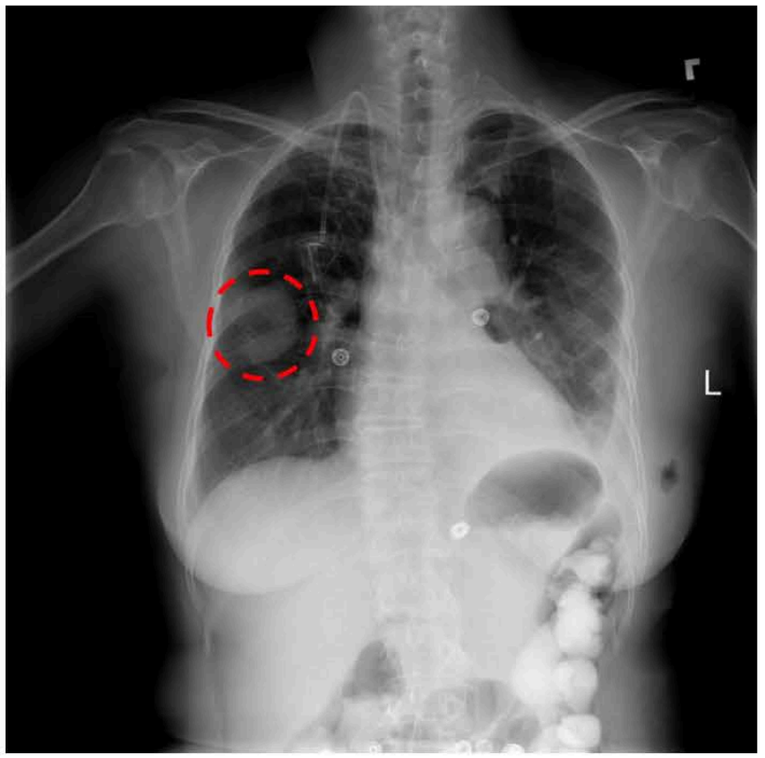
NIH Clinical Center provides one of the largest publicly available chest x-ray datasets to scientific community

The dataset of scans is from more than 30,000 patients, including many with advanced lung disease.

What

The NIH Clinical Center recently released over 100,000 anonymized chest x-ray images and their corresponding data to the scientific community. The release will allow researchers across the country and around the world to freely access the datasets and increase their ability to teach computers how to detect and diagnose disease. Ultimately, this artificial intelligence mechanism can lead to clinicians making better diagnostic decisions for patients.

NIH compiled the dataset of scans from more than 30,000 patients, including many with advanced lung disease. Patients at the NIH Clinical Center, the nation's largest hospital devoted entirely to clinical research, are partners in research and voluntarily enroll to participate in clinical trials. With patient privacy being paramount, the dataset was rigorously screened to remove all personally identifiable information before release.



A chest x-ray identifies a lung mass.

Institute/Center

Clinical Center (CC)

Contact

Molly Freimuth
301-549-5789

Connect with Us

Subscribe to news releases

RSS Feed

www.nih.gov/news-events/news-releases/nih-clinical-center-provides-one-largest-publicly-available-chest-x-ray-datasets-scientific-community

ChestX-ray8

National Institutes of Health
Turning Discovery Into Health

Home » News & Events » News Releases

NEWS RELEASES

Media Advisory Wednesday, September 27, 2017

NIH Clinical Center provides one of the largest publicly available chest x-ray datasets to scientific community

Institute/Center

Clinical Center (CC)

Contact

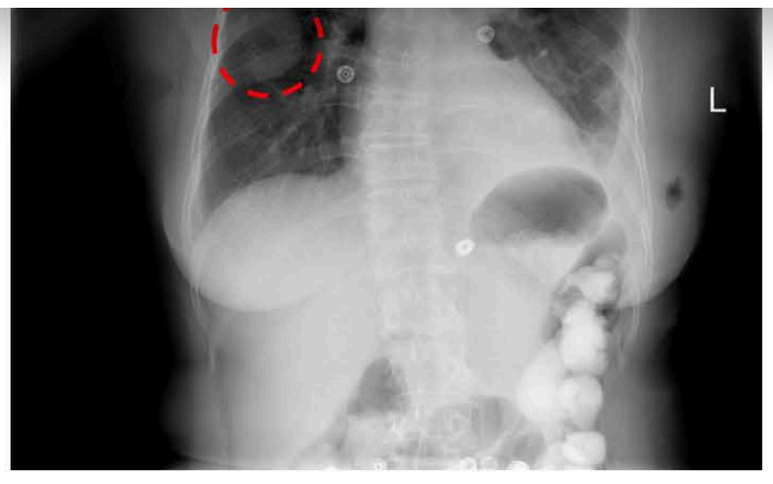
Molly Freimuth

This NIH Chest X-ray Dataset is comprised of 112,120 X-ray images with disease labels from 30,805 unique patients.

To create these labels, the authors used Natural Language Processing to text-mine disease classifications from the associated radiological reports.

to detect and diagnose disease. Ultimately, this artificial intelligence mechanism can lead to clinicians making better diagnostic decisions for patients.

NIH compiled the dataset of scans from more than 30,000 patients, including many with advanced lung disease. Patients at the NIH Clinical Center, the nation's largest hospital devoted entirely to clinical research, are partners in research and voluntarily enroll to participate in clinical trials. With patient privacy being paramount, the dataset was rigorously screened to remove all personally identifiable information before release.



A chest x-ray identifies a lung mass.

www.nih.gov/news-events/news-releases/nih-clinical-center-provides-one-largest-publicly-available-chest-x-ray-datasets-scientific-community

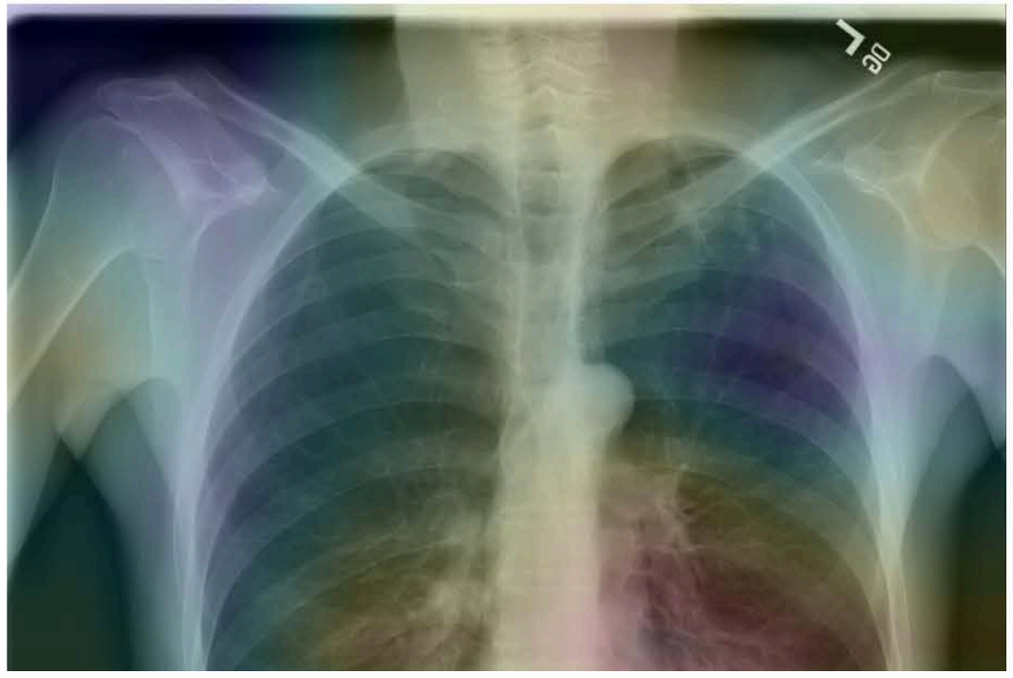
Stanford ML Group

CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning

Pranav Rajpurkar*, Jeremy Irvin*, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, Matthew P. Lungren, Andrew Y. Ng

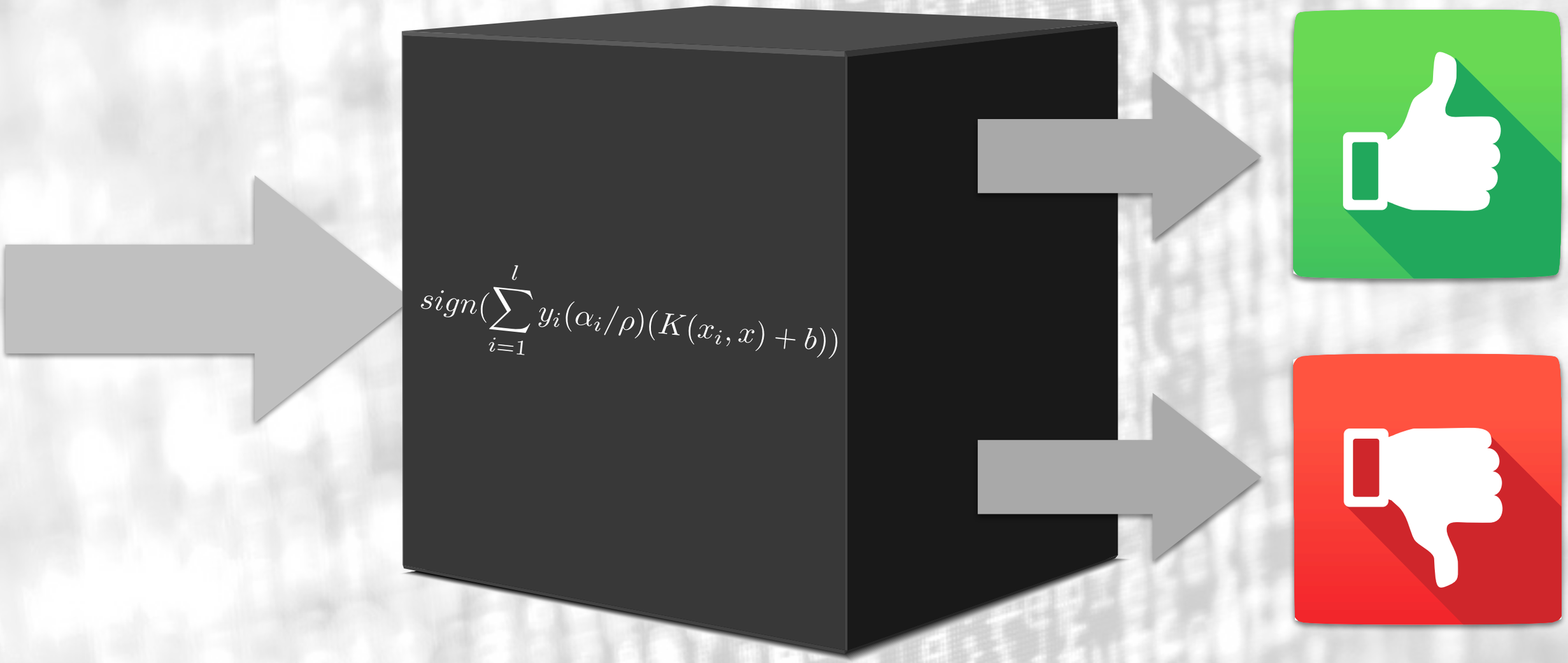
We develop an algorithm that can detect pneumonia from chest X-rays at a level exceeding practicing radiologists.

Chest X-rays are currently the best available method for diagnosing pneumonia, playing a crucial role in clinical care and epidemiological studies. Pneumonia is responsible for more than 1 million hospitalizations and 50,000 deaths per year in the US alone.



stanfordmlgroup.github.io/projects/chexnet

algorithmic decision



shift Obesity System Influence Diagram

Full Map
Clusters
Core Loop
Individual Psychology
Social Psychology
Individual Activity
Activity Environment
Food Consumption
Food Production
Individual Physiology
Physiology

Click-drag to scroll
Double-click to zoom in/out

Social Psychology
Individual Psychology
Individual Physical Activity
Physical Activity Environment
Food Production
Food Consumption
Individual Physiology
Physiology
ENERGY BALANCE

Negative Influence
Positive Influence

Media
Social
Psychological
Economic
Food
Activity
Infrastructure
Developmental
Biological
Medical

www.gov.uk/government/publications/reducing-obesity-obesity-system-map

algorithmic agricultural advice

PlantVillage

About Projects Team Blog Donate Contact us MORE

Profile
SIGN IN →

We help African smallholder farmers adapt to climate change

Climate change shocks coupled with pests and diseases prevent African farmers from becoming profitable enterprises

The PlantVillage solution

PlantVillage has developed a triple A model (Algorithmic Agricultural Advice) that works to increase the yield and profitability for millions of farmers. It is our goal to reach hundreds of millions in partnership with an ecosystem of farmer facing organizations and the farmers themselves. Our algorithms come

plantvillage.psu.edu

what is new ?

our digital image of the world
tracks the world more and more closely

what is new ?

our digital image of the world
tracks the world more and more closely

- **new data streams** providing a real-time, multi-modal, full-scale window on many systems and processes of interest

what is new ?

our digital image of the world
tracks the world more and more closely

- **new data streams** providing a real-time, multi-modal, full-scale window on many systems and processes of interest
- we can use computation to **discover patterns and understand systems** using tools from statistics, data mining, machine learning, network science, NLP, etc.

what is new ?

our digital image of the world
tracks the world more and more closely

- **new data streams** providing a real-time, multi-modal, full-scale window on many systems and processes of interest
- we can use computation to **discover patterns and understand systems** using tools from statistics, data mining, machine learning, network science, NLP, etc.
- we can use large-scale datasets to **train models of unprecedented complexity**

what is new ?

our digital image of the world
tracks the world more and more closely

- **new data streams** providing a real-time, multi-modal, full-scale window on many systems and processes of interest
- we can use computation to **discover patterns and understand systems** using tools from statistics, data mining, machine learning, network science, NLP, etc.
- we can use large-scale datasets to **train models of unprecedented complexity**
- **complex models** are **in the loop** of for more and more products, services, and decision-support systems

what is new for a regulator ?

what is new for a regulator ?

- **Complex models in the loop** of for more and more products, services, and decision-support systems drive new needs:
  - auditing AI-powered systems ?
  - transparency, fairness, accountability of ML systems ?

what is new for a regulator ?

- **Complex models in the loop** of for more and more products, services, and decision-support systems drive new needs:
  - auditing AI-powered systems ?
  - transparency, fairness, accountability of ML systems ?
- Potentially useful **data generated neither by the regulated subjects nor by the regulator**
  - new kinds of real world data (e.g., from digital therapeutics)
  - data access and governance ?
  - trusted data intermediaries ?

key ideas

→ 1

data are as important as algorithms and computational platforms

key ideas

→ 1

data are as important as algorithms and computational platforms

→ 2

secondary use of (often privately-held) (often personal) data

key ideas

→ 1

data are as important as algorithms and computational platforms

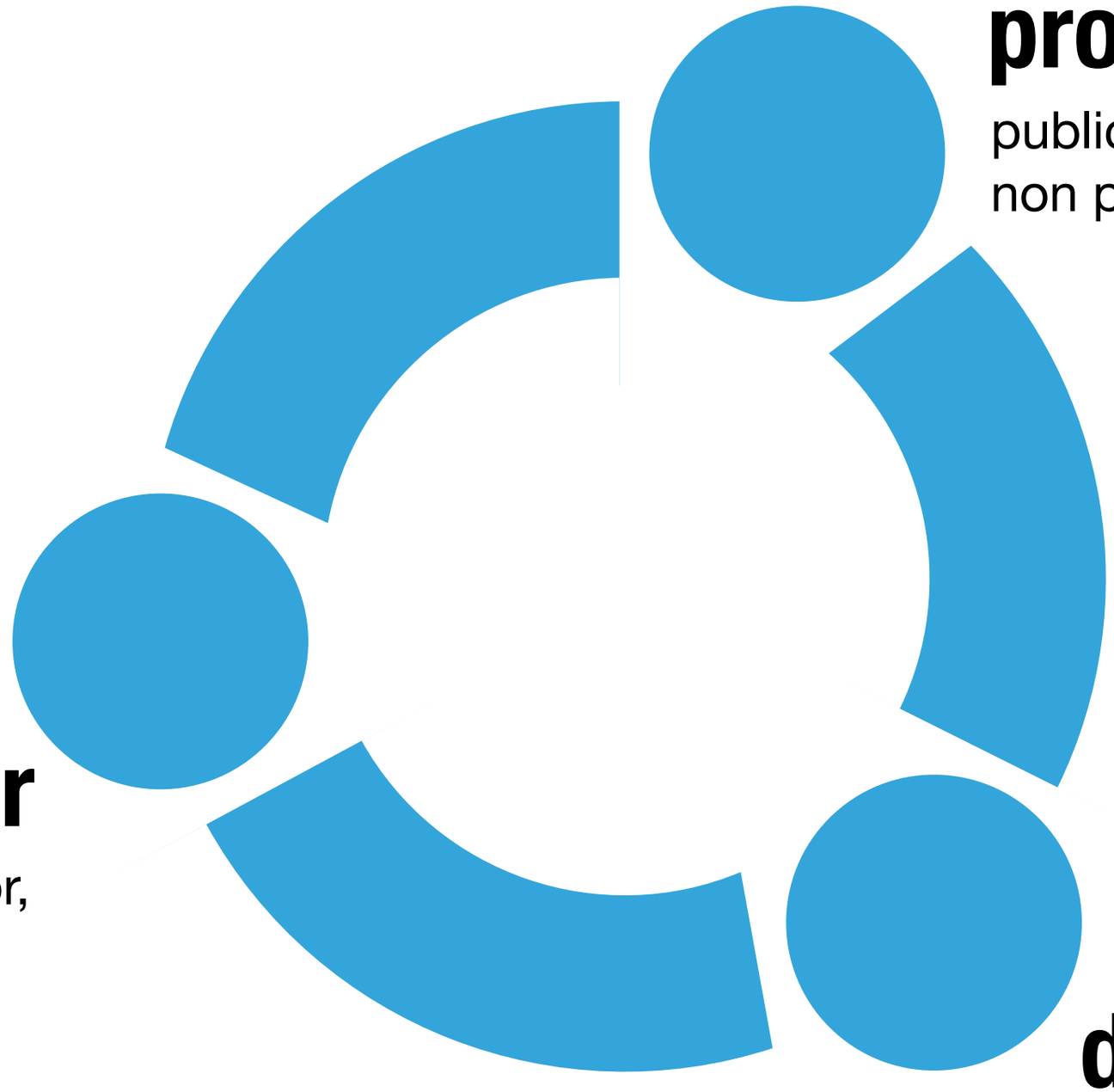
→ 2

secondary use of (often privately-held) (often personal) data

→ 3

data and AI plug into complex systems

a governance challenge



problem owner

public agency, municipality,
non profit, …

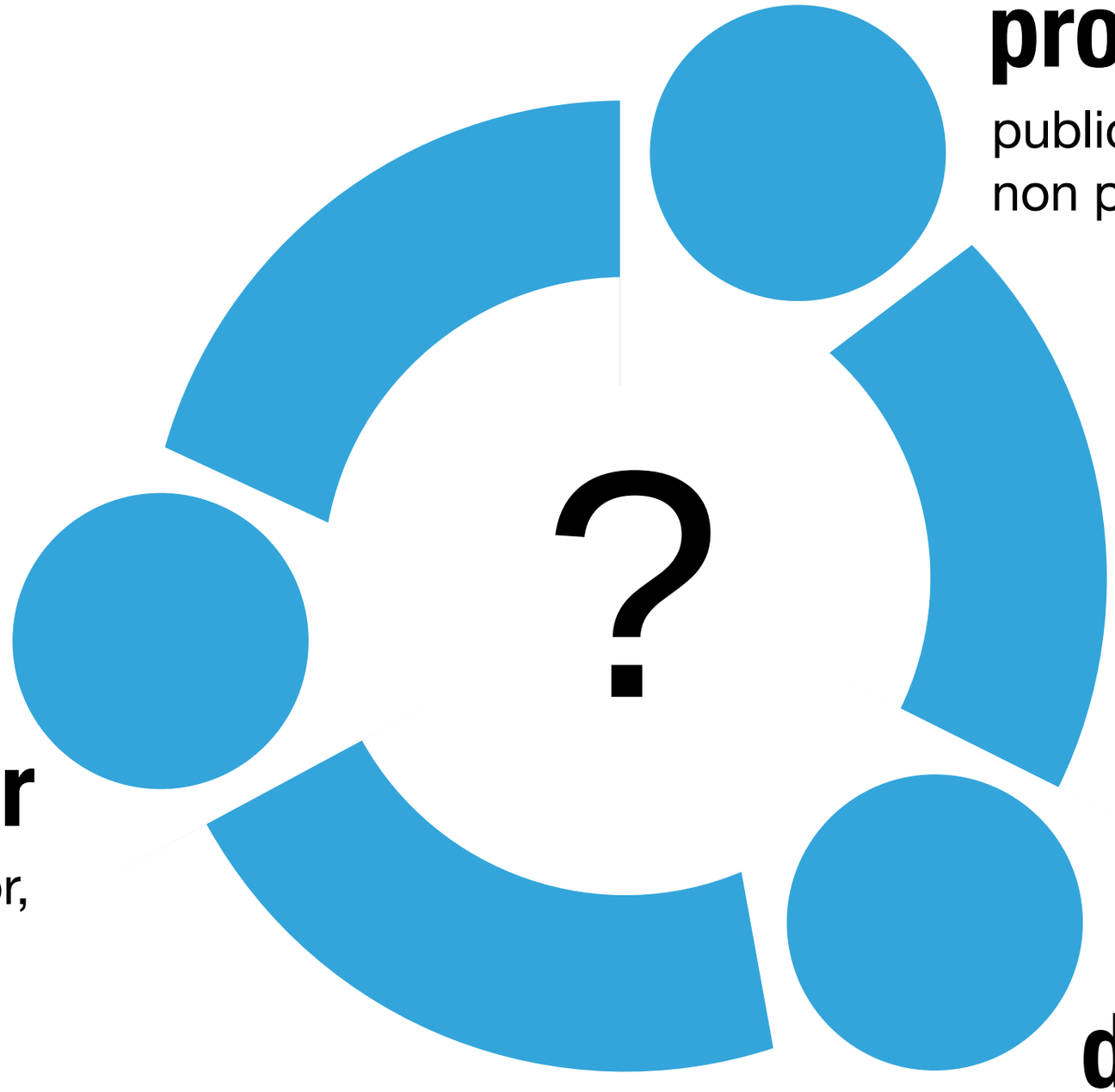
data owner

tech, private sector,
citizens, …

data expertise

research, academia,
private sector, …

a governance challenge



problem owner

public agency, municipality, non profit, …

data owner

tech, private sector, citizens, …

data expertise

research, academia, private sector, …

2017

Digital Single Market

EVENTS | 30/05/2017 | CENTRE CONFERENCES ALBERT BORSCHETTE (ROOM 4A), BRUSSELS

Workshop on access for public bodies to privately-held data of public interest

26/06/2017

This workshop responds to the mid-term review of the Digital Single Market strategy which calls for the need to further explore the issue of public bodies' access to privately held data which are of public interest.

digital-strategy.ec.europa.eu/en/events/workshop-access-public-bodies-privately-held-data-public-interest

2018

Digital Single Market

NEWS ARTICLE | 22 November 2018

Commission appoints Expert Group on Business-to-Government Data Sharing

Following an open selection process, the Commission has appointed 23 experts to a new Expert Group on Business-to-Government Data Sharing. The group is composed of independent experts with a high level of expertise in the field, covering a broad range of areas of interest and sectors.

digital-strategy.ec.europa.eu/en/news/commission-appoints-expert-group-business-government-data-sharing

2020

Towards a European strategy on **business-to-government data sharing for the public interest**

Final report

ec.europa.eu/newsroom/dae/document.cfm?doc_id=64954

- all Member States should have **governance structures** in place tasked with overseeing and giving advice on responsible B2G data-sharing practices

- all Member States should have **governance structures** in place tasked with overseeing and giving advice on responsible B2G data-sharing practices
- private, public and civil-society organisations should **create and promote the function of a data steward**. Furthermore, the European Commission should encourage the creation of a network of such data stewards, […]

- all Member States should have **governance structures** in place tasked with overseeing and giving advice on responsible B2G data-sharing practices
- private, public and civil-society organisations should **create and promote the function of a data steward**. Furthermore, the European Commission should encourage the creation of a network of such data stewards, […]
- The European Commission should explore the creation of an **EU regulatory framework to facilitate public-sector reuse of privately held data for the public interest**. This framework should include data sharing requirements, transparency requirements and safeguards […]

- all Member States should have **governance structures** in place tasked with overseeing and giving advice on responsible B2G data-sharing practices
- private, public and civil-society organisations should **create and promote the function of a data steward**. Furthermore, the European Commission should encourage the creation of a network of such data stewards, […]
- The European Commission should explore the creation of an **EU regulatory framework to facilitate public-sector reuse of privately held data for the public interest**. This framework should include data sharing requirements, transparency requirements and safeguards […]
- the European Commission should carry out **studies to obtain further empirical evidence of the macroeconomic and societal benefits of B2G data sharing** for the public interest.

2020

European Commission

Search

Shaping Europe's digital future

Home Policies News Library Funding Calendar Consultations

POLICY AND LEGISLATION

Proposal for a Regulation on European data governance (Data Governance Act)

The proposal is the first of a set of measures announced in the 2020 European strategy for data.
The instrument aims to foster the availability of data for use by increasing trust in data intermediaries and by strengthening data-sharing mechanisms across the EU.

digital-strategy.ec.europa.eu/en/library/proposal-regulation-european-data-governance-data-governance-act

2021

European Commission

Search

Shaping Europe's digital future

Home Policies News Library Funding Calendar Consultations

Home > Library > Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)

POLICY AND LEGISLATION

Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)

The Commission is proposing the first ever legal framework on AI, which addresses the risks of AI and positions Europe to play a leading role globally.

digital-strategy.ec.europa.eu/en/library/proposal-regulation-laying-down-harmonised-rules-artificial-intelligence-artificial-intelligence

thank you

Ciro Cattuto
University of Torino & ISI Foundation
www.cirocattuto.info

Veterinary Big Data Stakeholder Forum
European Medicines Agency
2 June 2021